ISBN: 978-65-00-98107-0
9 786500 981070

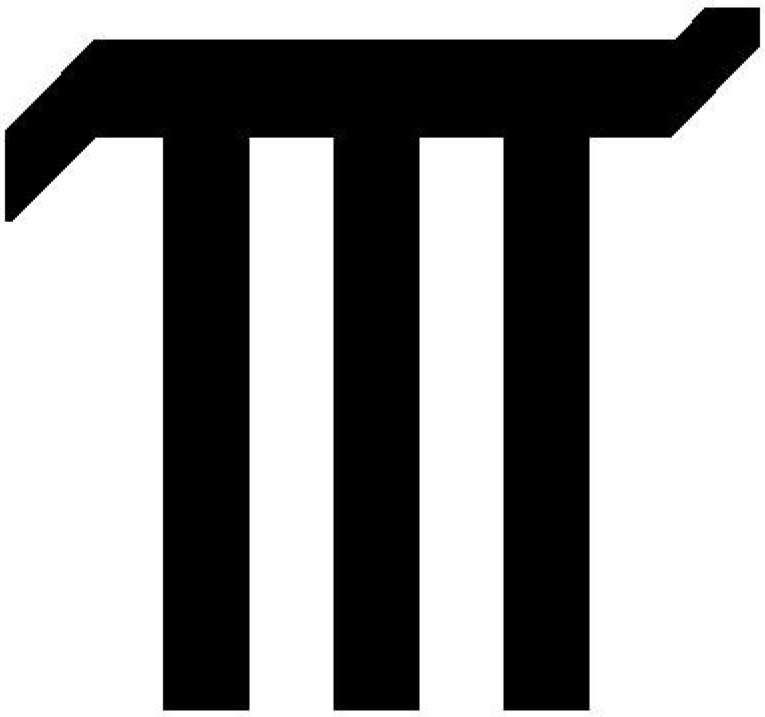

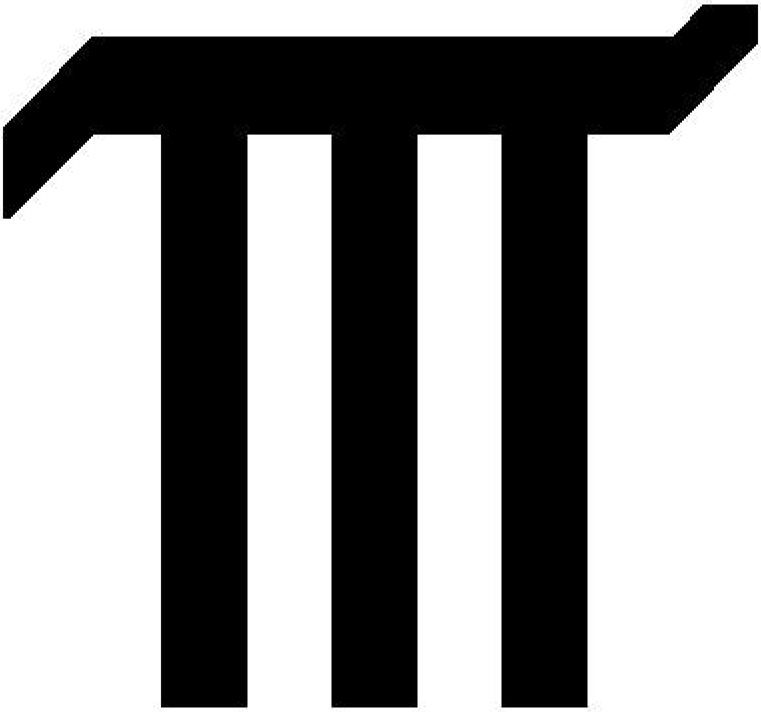

2>+ *1 2+(>

!
" #
$ % & ’ ( ) * + ’ , , , # # - . /
0%% ! % (1*((2))
, , ,
, , , 3
0%% %! % 4 5 - # 6/
7 0 4 -8 /
5 4
5 4

9 : ! ; 1 ’ < &
= & ’ = &
8 7 & >1 1
.
? -@11/ ((21)> 1>>

(+)) ’ - 2> A (/

(7 B
& B
) B
C B
1 B

8 0
; -(2*2 /

24-199630 CDD-516.009Dados Internacionais de

& 5 :

-)((*(+(> */

! ! , 6 !; 5 !

! ! 5 ;
-5 /;
B ! - /
!
-# /

! ! 5 5 ;
, , # ; ; 5 !

<

D -+;)( ) /
)(
D 5 ! D ,
, ! # ! ;E 5
, !; !
)
F , 6 , ; !
, 6 5 E

, 5 ;# 5 !
- = ! ! ! ! ; ! /

<

? $9G $?7
(!

? 5 ! 5 5
! 5 ; H 5
! 5 5 6 , ! ;
5 I ; &
? ! # , ! & ;
5 H ! ,
J 5 ! G , ; # ,
, !
C ! , ! , 5 5 ! #
! ! !; # , 6 5 5 !
; 6 , ! # 6 ! , ; 5 G
5 ; , 6 !; , 0 K 6 , ! !
! G L; # , ! 5 ,
C !; 5 - , (;+/
? # 5 , 6 5 5 ;
! # ! # ! ! , # !
, 5 ; ; , , #
5 ! # M; ,

, 5 ; ,
! &
, ! ;
7 !G & & 7 !G 5 5 ,
! ! , ;
; ! # , #
5
: # # , ! & ;
! ! ! H 0 E ! B E ! # ; #
! 5 ! 5 5 & ;
! ! H 5 ! ;

! 5 I 5 5
6 5 5
? , 6 !; # ! ! ! ! !
55 , , 5 5 # !; 55
, ! , 5
! & ; 5 ! 5 ! ; ! ! 5
, ! ! , 5 6 , ! 5
; , # , ; 5 ! N
: G 5 ; ! H ; # ,
, ,
? , # ! M # -
! # ; # ! 5
, , !/ ! ; #
5 5 ; , 5 5
, & ! ! ; 5 7 G ? 5 9 ;
, 5 ! 5 5 ; #
H 5 C ! C 5 ; ,
O ? N 6 ? !
! # ! ; I 5
, 5 ; # , ;
, ! # ! ! 5 ;
5 5 5 ; # 6 # ,
! ; ; , ! 5 !
I ! ; , I # !; #
, ! , 5 ; , ! ! ;
, ! 5 ; ! !
7 ; # ; 8 8 7
! ; , , ! I 5

C # 5 5 ; , , ! #
; # 5 6 ; 5 !
! ! ! 2; ! !
! # M ! ; ! ! ! # # 6
& ! ; , # , ; # ; ;
; ;) ; 7 ;) ; ! 6 5
; , , # (1; , ; , !! ! ;
, , # *; , 5 ! ; , ! !
5 # C ! 6 ; 6 5) #
*; ! 9 ( );(+ - ! # I ! ! !

I I /; # ; 5 # 5 !
! ; , , # # + 5 !G 5 ; ! ! 5
5 # 6 ! , D 8 , !
D , # ; , ! # 5
! 5
5 ; # 6 ,
! 5 6 , ! 5 5 5
! !
B ? N
= & ; ; -8 ! & ! /
& 5 :
, , ,
4

<
$? :$?$9
P (Q ! R=

3 S T 5 ! ! ! !
! ! 3 ! # T S ; !
! ! ! R= ! H ! #S !
3 ! ; I ; CU
! 3 CU S ! ! ;
! 3 ! H H ! ! H
J H = H ; 3 H
!

: ! ! ! ! ! H ! ! ! ! !
! ! ; # # !
5 R ! ! ; = ! 5 ! !
; ! U ! ; H
R ; H ! .0 K7 T = L; V
! H ! ! ! ! H = = U :
R # ; ! R ! - (;+/
R= ! # 5 3 ! H ;
W3 = ! S ! R= !
3 ! 3 ; U ; H !
X ! ! X M; H !

’ 5 -7 R 5
3 /; 3 3
H CU
7 ! ; H 5 # ! !
H ; 7 !G & & 7 !G 5 5 7
R= ! H ! ! !
; U ; # R= ! X
5

7 ! ! R= ! 3 ! 3 ! CU ;
# ! H = 0 E ! B E S ! = ! ! ;
.= ! 5 R 5 ! ! CU H ; # ! !
H = ! V H ; ! Y

3 ! I = ! ! ! 5 ! 3
!
$ # S 3 ! ; 5 ! .
H . 5 H 5 ! ! ! R : ;
5 R ! H ! ! ! ! !
! ! 3 ! CU ; # ! T
! ; ! ! ! 5 ! ! H H
! ! ! ! ! ; H
; ! ! X S ! : N = H R ; T
! : ! ! H RZ ; 7 H
3
$ X S H X ! M -
! X ; X = ! ! !
H ! / ; # ! R= !
3H 5 ; ! 5 R= ! 3 CU
5 ! ; 5 ! ? ! 9 ! ! ! 7 ;
H ! ! = S 5 R 5 ! ! ;
H T ! ! C R C ! ! ; ?
O Y N 6 ? ! H ! S #
; ! H I ! !
$! R= ! ! ; ! ; =
! ! ; ! ! 5 RZ !
! ! ; H # ! S S ; H
5 ! S ! ! I ! S ! ! ;
H I : ! ! H 7 ;
! ; R= ! ! ; H !
! [; ; 8 8 V H
7 ; S H R= I =
!
7 ! ! U ! ! ; H
H ! ; ! ; !
T ! W 5 ! ! S H 2;
! W ! H # ! X M ;
! ! : ! 9 ; H ; S
X ; ! R ! ; W ;) ;
T ;) ; 5 . ! ! X ; V #
(1; H .; ! !U ; V # *; H S
! ; H #S S ! ! # R= & U
5 ; T = ! 5) X *; 5 !

( ) ; (+ - 7 X S / ; 3

! X ; X ! 5 ; #

X + ! 5 ! : ; ! ! # R= ! S !

H D 8 5 # ! D S 5 ;

R= V I 7 V

: ! # R= ; ! I

RZ ! H ! U5 !

H ! 5 ! # RZ

< < ! ! N

= & ; ! R ; - : ! 9 C ! 8 V /

& 5 :

, , ,

4

Index - M.U.S. - Mathematimus - Hyperelliptical Geometry

Chapter 1 4
Chapter 1 - Part 1 5
Notebook-01---20211112...6
Musichian Projective Vectorial Curvilinear Cubic Trigonometry. 199
Cosmological Constant 200
Everything will depend on the conversion factor applied, which can increase or decrease the M.U.S. 200
Chapter 1 - Part 2 407
Notebook-02---20211112...408
Chapter 1 - Part 3 604
Notebook-03---20211112...605
Chapter 1 - Part 4 959
Volume-20---20220612..960
Nothing exists 1055
Analyse the "light bending" effects of the Schwarzschild geometry. 1240
Chapter 2 1321
Chapter 2 - Part 1 1322
Volume-01---20211021..1323
Chapter 2 - Part 2 1863
Volume-02---20211103..1864
Chapter 3 2527
Chapter 3 - Part 1 2528
Volume-03---20211108..2529
Chapter 3 - Part 2 3194
Volume-04---20211110..3195
Chapter 4 3952
Volume-05---20211111..3953
Chapter 5 7608
Chapter 5 - Part 1 7609
Volume-06---20211206..7610
Chapter 5 - Part 2 7909
Volume-07---20211206..7910
Chapter 8 8554
Volume-08---20211207..8555
Chapter 7 11860
Chapter 7 - Part 1 11861
Volume-09---20211218..11862
Chapter 7 - Part 2 12424
Volume-10---20220103..12425
Chapter 7 - Part 3 12547
Volume-11---20220104..12548
Chapter 7 - Part 4 12647

Volume-13---20220104......12648
Chapter 7 - Part 5......13098
Volume-15---20220213......13099
Chapter 8......13126
Chapter 8 - Part 1......13127
Volume-16---20220213......13128
Chapter 8 - Part 2......13337
Volume-17---20220220......13338
Chapter 8- Part 3......13634
Volume-18---20220315......13635
Chapter 8 - Part 4......13714
Volume-19---20220412---01......13715
Volume-19---20220412---02......13715
Chapter 8 - Part 5......13799
Volume-21---20220818......13800
M.U.S. = Pi . Gl . Gu (bending of the light - real number)......13941
M.U.S. - use the new number to recalculate the equation.......13993
Chapter 8 - Part 6......14078
Volume-23---20230826......14079
Chapter 9......14154
Volume-24---20231210......14155
Solution to calculate the curvature of space about the elliptical orbit of the planets......14220
Me - Ellipse Musich......14224
Me - Musich Point......14224
Procedures to join two circles that intersect at two points, creating a total of four points that correspond to the two diameters of these circles that create an ellipse.......14267
M.U.S......14352
Musichian Coordinate and M.U.S.......14361
Musichian Coordinate......14363
Planck's Constant......14409
Coriolis Force......14409
K is said to be weakly zero......14411
Analysis......14466
Curvature = Energy......14471
M.U.S......14471
Musichian Hyperelliptic Geometry - M.H.G......14479
Analysis - Nothing......14558
Musdesic......14578
H.G., E.G. and S.G......14608
Axiom of parallelism of straight lines......14623
The solution of the new geometry (Musichiana) and the new number M.U.S. will solve Euclid's Parallel Postulate.......14725
Theory of non-uniforme space hyperellip Musichian......14799
The centre of a quadric hypersurface......14843

Deflection of Light 14891
Chapter 10 14957
Volume-25---20231226 14958
Chapter 11 15779
Volume-26---20231229---20240224 15780
Chapter 12 16267
Chapter 12 - Part 1 16268
Appendix-01---20211115 16269
Chapter 12 - Part 2 17742
Appendix-02---20211117 17743
Chapter 12 - Part 3 17861
Appendix-03---20211112 17862

Chapter 1

Chapter 1 - Part 1

Notebook-01---20211112

data 13 7 2019

1) Geometric Interpretation (Fourier's Conditions)
Pág. 29 (41)

Prof. Stenio Musich

6.

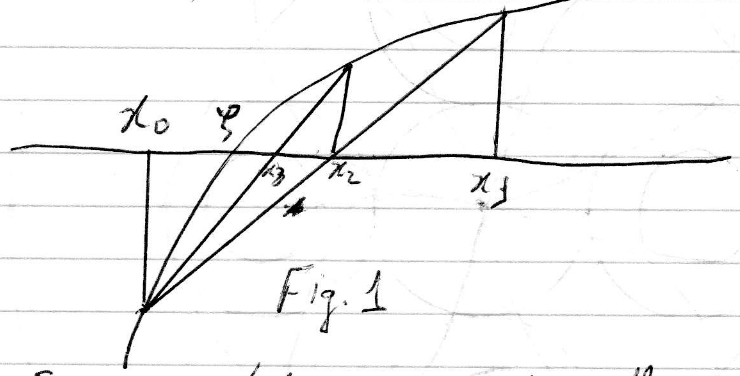

Fig. 1

Source: 20190713/Measurement Using the New Rules/Ostrowski - Solution of equations and systems of equations.pdf

Pág. 75 10 - Dandelin Bounds for Newton - Raphson Iteration
Veja. Fig. 4.

2) 3. The Canonical Approach
Pág. 33 (49)
Pág. 35 (51)

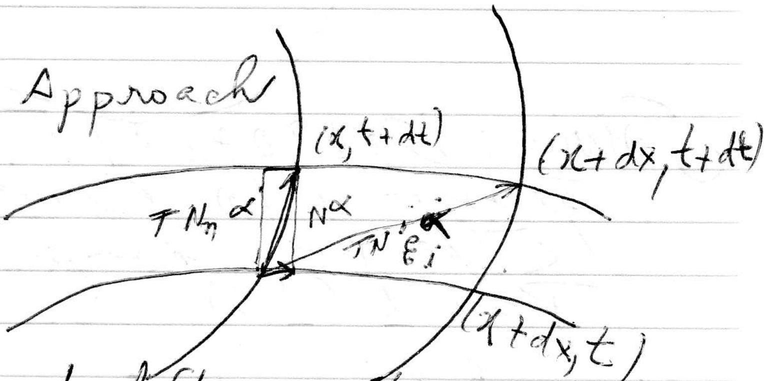

Fig. 2
ADM lapse and shift variables.
(Quantum Gravity)

Source: 20190713/Planck Scale Effects in Astrophysics and Cosmology/Planck Scale Effects.... pdf

3) Linear Algebra with Applications - Nicholson
Source: 20190713/Linear Algebra with Applications/ SaNet.st_0070401098.pdf

4) Arrangements
Fig. 1.1. An arrangement of 14 circles in the plane, with 53 faces (one of which is unbounded), 106 edges, and 59 vertices.

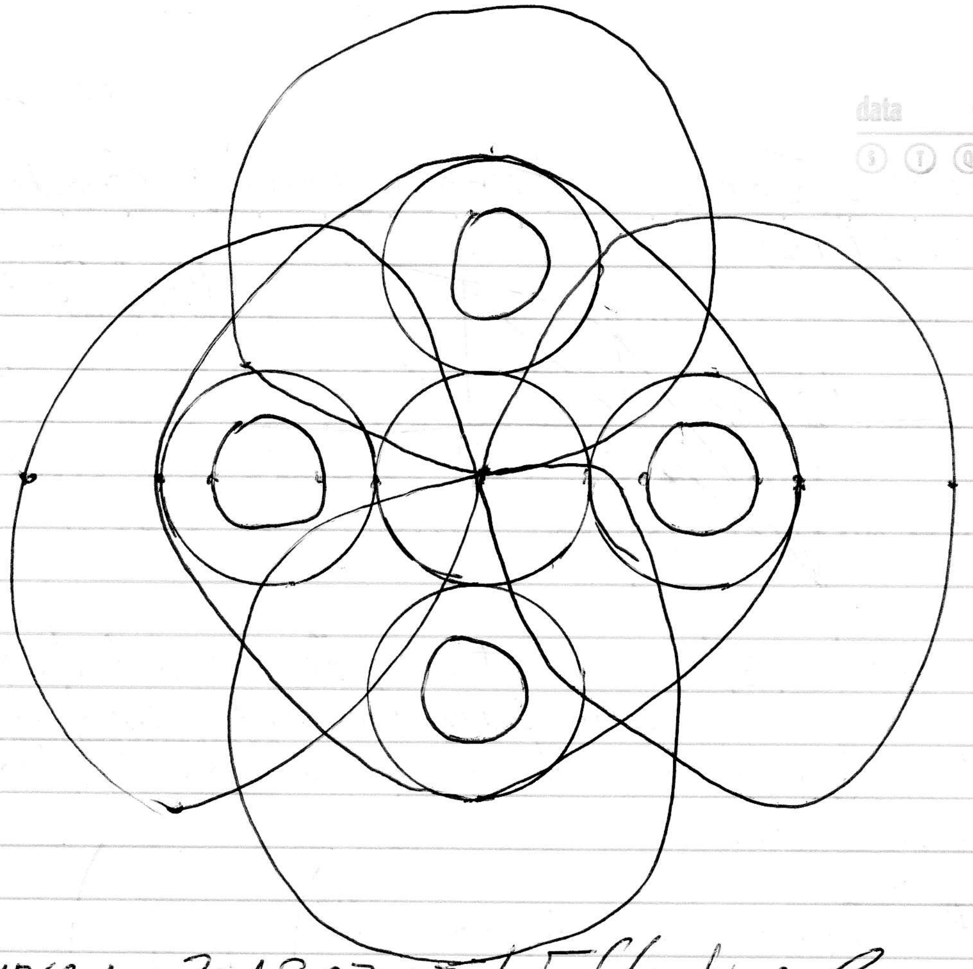

Source: 20190713/Effective Computational Geometry for Curves and Surfaces/Effective Computational... pdf
Pag. 1 (13)

5) Quantum Mechanics of Single Particle in 1 D Space II
Pag. 103 (124)

Exercise 4.8
Answer to Exercises 4.6-4.8:
The solution to the Schrödinger equation with energy E is

$$\psi_E = e^{\pm\sqrt{\frac{2m}{\hbar^2}(V(x)-E)x}},$$

where $V(x) = V_0\,\theta(a-x)\,\theta(x+a)$.
Source: 20190713/Introduction to Quantum Mechanics/Thomas Banks - Quantum Mechanics...(2018).pdf

data 13.07.2019

6) 2.7.2 Approximation of the Medial Axis

Approximation the medial axis of a set is a non trivial issue since sets that are close for the Hausdorff distance may have very different medial axes.

Pág. 110 (122)
Pág. 112 (124)

Lemma 10.

Let σ be a S-empty sphere centered at c, of radius r, intersecting S in two points x and y.

If $\varepsilon < \frac{r}{2}$ and $l \overset{def}{=} \frac{\|x-y\|}{4} \geq \sqrt{\varepsilon r\left(1-\frac{\varepsilon}{r}\right)}$, there exists an S'-empty sphere tangent to S' in two points whose center c' and radius r' are such that

$\left|1-\frac{r'}{r}\right|$ and $\frac{\|c'-c\|}{r}$ are at most

$$\delta = \frac{\varepsilon r}{l^2 - \varepsilon r + \varepsilon^2}.$$

Fig. 2.14. S is the continuous curve. x' and y' belong to S'

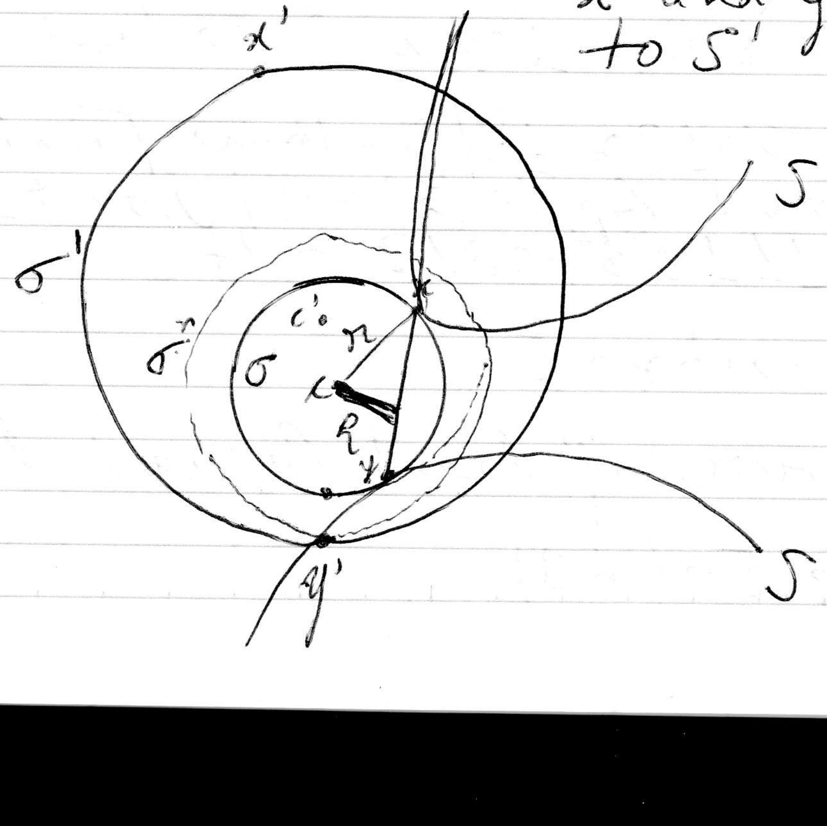

data 14.07.2019

Source: 20190713/Important/ Effective Computational Geometry for Curve and Surface-Springer (2007).pdf.

7) 20190714

3.5. Valves during morphogenesis

Pág. 62 (39)

Letra a: figure com 7 pontas.

Figure 3.5 Beauty of the external surface of immature valves of Asterolampra marylandica at increasing stages and high magnification (SEM).

a) Internal view of whole mature valve.

Figure 3.6 letra e.

Aulacodiscus janischii,
Scale bar = 100 um (e and f).

Source: 20190714/ Diatoms Fundamentals and Applications/ sanet_st_1119370213.pdf

8) Proposition 5

"Curve of least Poincaré length"

Pág. 50 (69)

data 14.07.2019

Book: Complex Analysis:
The Geometric Viewpoint

Krantz, Steven.

Source: 20170914/Complex Analysis the Geometric Viewpoint/(Carus Mathematical)... Viewpoint.pdf

Look at pag. 52 (71):

Fig. 4.

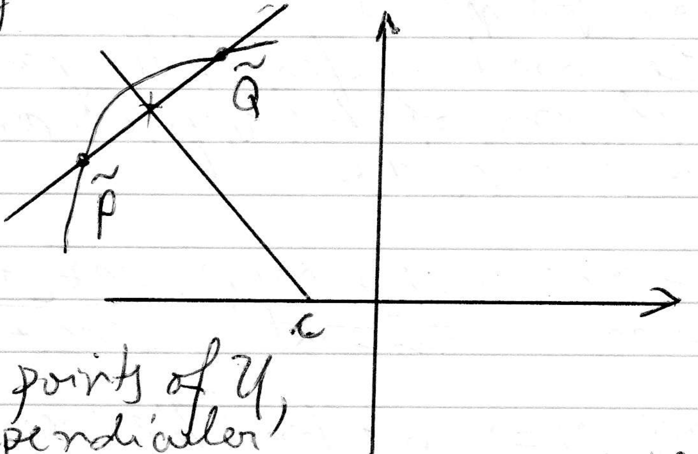

If $\tilde{P}$, $\tilde{Q}$ are points of $\mathcal{U}$, then the perpendicular bisector of the segment connecting these points will intersect the real axis at the center C of the circular arc that forms the geodesic through $\tilde{P}$ and $\tilde{Q}$.

9) Classical Measurements in Curved Space-Times

Source: 20190714/Classical Measurements in Curved Space-Time/Classical... pdf.

10) 201907191808

Advanced Comprehensive Mathematics

data 19.07.2019

S T Q Q [✓ S] S D

Source: Comprehensive Mathematics for JEE/
(20190719.02) Advanced Comprehensive Mathematics.pdf
Pag. 915 (item 20.1)
Three Dimensional Geometry
Coordinates.

Distance Formula

$$AB=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2}$$

Section Formulae

The coordinates of the point P which divides the join of $A(x_1, y_1, z_1)$ and $B(x_2, y_2, z_2)$ in the ratio $m:n$ are

$$\left(\frac{mx_2+nx_1}{m+n}, \frac{my_2+ny_1}{m+n}, \frac{mz_2+nz_1}{m+n}\right)$$

by taking $m=n$ we find the coordinates of the mid-point of AB as

$$\left(\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2}, \frac{z_1+z_2}{2}\right)$$

Coordinates of any point on the join of two points $A(x_1, y_1, z_1)$ and $B(x_2, y_2, z_2)$ are

$$\left(\frac{\lambda x_2+x_1}{\lambda+1}, \frac{\lambda y_2+y_1}{\lambda+1}, \frac{\lambda z_2+z_1}{\lambda+1}\right)$$ where

$\lambda \neq -1$. If λ takes only positive real values, we get the coordinates of the points on the segment AB (except A and B).

data 19.07.2019

11) Inorganic Chemistry 4th Ed
Source: 20190719-03/Inorganic Chemistry, 4th edition.. / Catherine Housecroft, Inorganic Chemistry...pdf
Pág. 140 (1829)
sp Hybridization: a scheme for linear species

The notation sp means that s atomic orbital and one p atomic orbital mix to form a set of two hybrid orbitals with different directional properties.

obs: Fig. 5.1 e 5.2 → shape of the H_2O molecule (interessante)

The model of sp hybridization can be used to describe the σ-bonding in a linear molecule such as $BeCl_2$ in which the Be–Cl bond are of equal length.
The ground state eletronic configuration of Be is [He] $2s^2$ and the valence shell contains the 2s atomic orbital and three 2p atomic orbitals (Fig. 5.3)

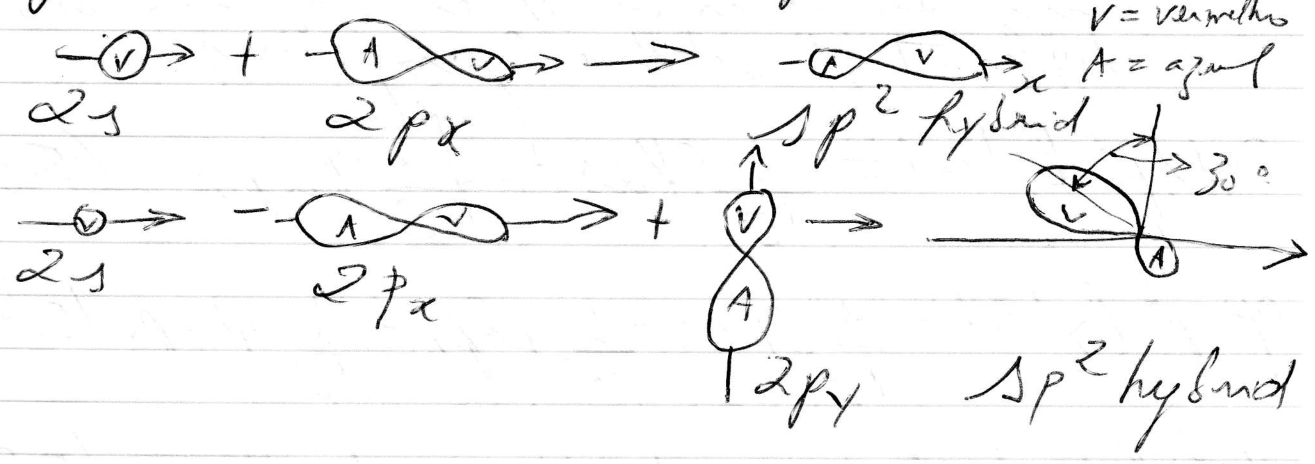

→

estudar!!!

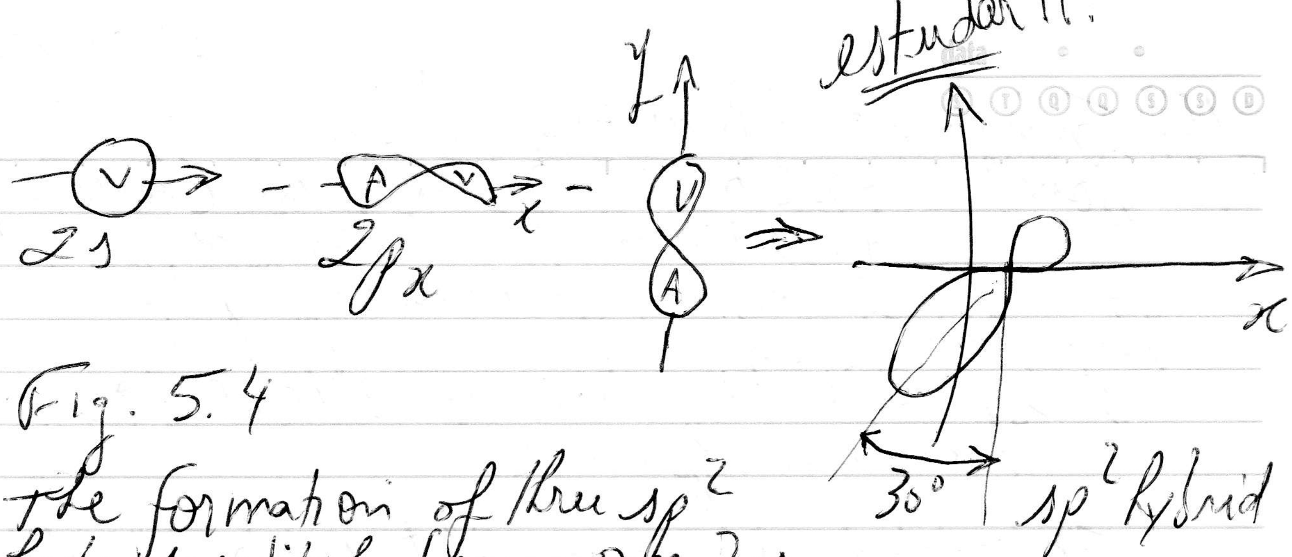

Fig. 5.4

The formation of three sp^2 hybrid orbitals from one 2s atomic orbitals and two 2p atomic orbitals.

Obs: the choice of p_x and p_y is arbitrary. (If we started with $2p_x$ and $2p_z$ atomic orbitals, the hybrid would lie in the ~~xy~~ xz plane; using the $2p_y$ and $2p_z$ atomic orbitals gives hybrid orbitals in the yz plane).

° wave functions (5.3 – 5.5) Pag. 141 [183]

↳ sp^2 hybridization: a scheme for trigonal planar species

sp^2 → the notation sp^2 means that one s and two p atomic orbitals mix to form of three orbitals with different directional properties.

$$\psi_{sp^2\ \text{hybrid}} = \frac{1}{\sqrt{3}}\psi_{2s} + \sqrt{\frac{2}{3}}\psi_{2p_x} \quad (5.3)$$

$$\psi_{sp^2\ \text{hybrid}} = \frac{1}{\sqrt{3}}\psi_{2s} - \frac{1}{\sqrt{6}}\psi_{2p_x} + \frac{1}{\sqrt{2}}\psi_{2p_y} \quad (5.4)$$

$$\psi_{sp^2\ \text{hybrid}} = \frac{1}{\sqrt{3}}\psi_{2s} - \frac{1}{\sqrt{6}}\psi_{2p_x} - \frac{1}{\sqrt{2}}\psi_{2p_y} \quad (5.5)$$

data 19.07.2019
S T Q Q [✓ S] S D

12) Quantum theory of Tunneling

Source: 20190719_04/Quantum theory of Tunneling/jaret.st_981452506.pdf
Pág. XX (21)

In two or three dimensions when the barrier is only a function of the distance from the origin of the coordinate system, i.e. $\rho = \sqrt{x^2+y^2}$ in cylindrical coordinates or $r = \sqrt{x^2+y^2+y^2}$ in spherical polar coordinates we can separate the variables in the Schröndiger Equation and thus reduce the problem to a one-dimensional motion but now with the boundary conditions imposed at $\rho = 0$ or $r = 0$ and at ρ or r going to infinity.

* For instance in three dimensions if we assume that $V(r \to \infty) \to 0$ and that $V(r)$ has a maximum at $r = a$, then if the particle is originally confined in the region $0 \leq r < a$, it can tunnel through the barrier and go to infinity.

These special cases of two- or three-dimensional tunneling can be regarded as one-dimensional but the potentials $V(\rho)$ and $V(r)$ are replaced by $V_{eff}(\rho)$ and $V_{eff}(r)$, with the boundary condition that the reduced radial wave function must vanish at the origin.

Important: Pág. 9 (34) →

2.1. Tunneling and the Uncertainty Principle)

At the first sight the tunneling of a particle looks like a paradoxical problem, since if the height of the barrier is greater that (than) the total energy of the particle

$$E = \frac{1}{2m}p^2 + V(x) \qquad (2.1)$$

then in the range b where $V(x) > E$, the kinetic energy

$\frac{1}{2m}p^2$ is negative and p is imaginary,

but this simple picture is not correct.

$\therefore$ particle can be at the point x where $E < V(x)$ but then according to the uncertainty principle its momentum is uncertain by an amount $\sqrt{\overline{\Delta p^2}}$.

Thus if we know the position of the particle to be x, then its total energy cannot be E.

Position- Momentum Uncertainty in Tunneling (WKB approximation - Chapter 5)

$$\exp\left[-\frac{1}{\hbar}\int_{x_0}^{x}\sqrt{2m(V(x)-E)}\,dx\right] \qquad (2.2)$$

where x_0 is the classical turning point ...

data 19.07.2019

S T Q Q [✓ S] S D

13) Exploratory Galois theory

Source: 20190719_04/Exploratory Galois theory/Swallow J... (2004).pdf

Pág. 3 (17) I Introduction

i is a certain solution to the polynomial equation $x^2 = -1$

obs: π → to see as an infinite sum.

i is a root of polynomial with rational coefficients, and the fact that π is not the root of a polynomial with rational coefficients - in other words, the fact that π is transcendental - means that the methods of studing i are very likely not going to be especially useful in studying π.

Approaching Galois theory, we choose to consider only those numbers that are roots of polynomials with rational number coefficients.

$1/7$ is root of $7x - 1$;

$\sqrt{2}$ is root of $x^2 - 2$;

$\sqrt[3]{-5}$ is a root of $x^3 + 5$;

$\sqrt[4]{20}$ is a root of $x^4 - 20$, and

$11 + 13/\sqrt{17}$ is a root of $x^2 - 22x + (1888/17)$.

We call a root of a polynomial with rational coeffients an algebraic number.

20 7 2019

14) Emergence of Spacetime in String Theory.
Source: 20190721/The Emergence of Spacetime in String Theory/Tiziana....pdf
Pag. 80 (87) - Pag. 67
T-duality and emergence

Poincaré describes an imaginary two-dimensional world in wich two mutually exclusive hypotheses of an Euclidean and hyperbolic geometry turn out be equally consistent with the same collection of empirical data.

Poincaré story of physical space of that imaginary world does have an actual geometry, which is the Euclidean one.

15) Vectorial Optical Fields Fundamentals and Applications
Source: 20190721/Vectorial Optical Fields Fundamentals and Applications/Qiwen....pdf
Pag. 31 (44)

2. Generation of vector optical fields

Fig. 1 (a) Poincaré sphere Σ and (b) different SOPs on Poincaré sphere Σ.

(look at image).

polarization sate vectors...

20.07.2019

16| Calculus of Vector Functions

Source: 20190721/Multivariable Mathematics/Richard....djvu

Pág. 20 (29)

4.2. Theorem

Let $\mathbb{R}^n \xrightarrow{f} \mathbb{R}^m$ be a linear function, and let A be the matrix whose jth column is $f(e_j)$. Then $f(x) = Ax$ for every x in $\mathbb{R}^n$!

Figure No.

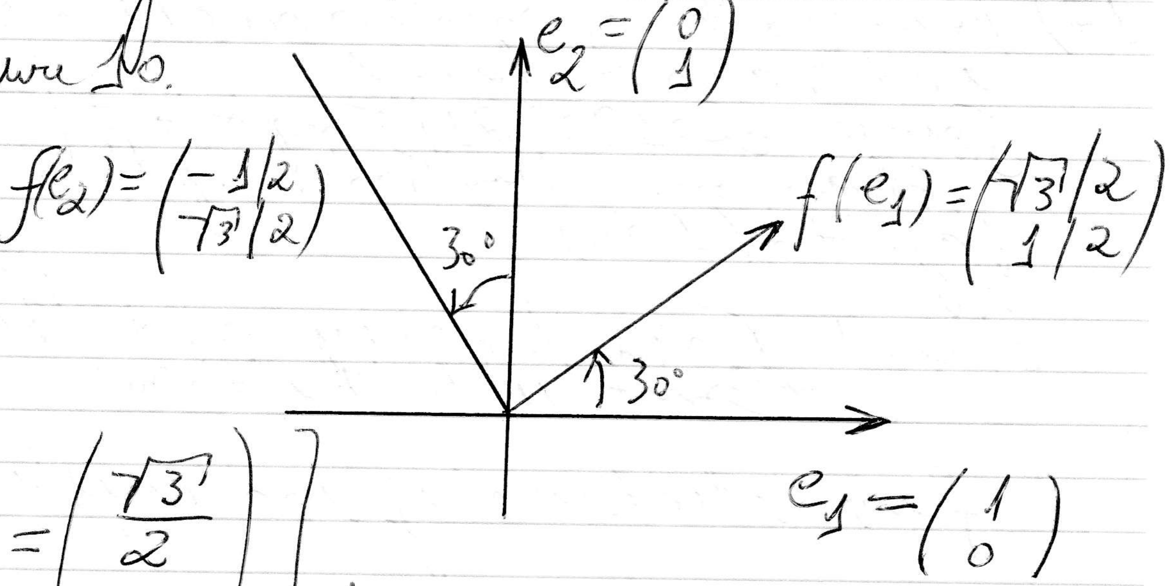
$e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$
$f(e_2) = \begin{pmatrix} -1/2 \\ \sqrt{3}/2 \end{pmatrix}$
$f(e_1) = \begin{pmatrix} \sqrt{3}/2 \\ 1/2 \end{pmatrix}$
30°
30°
$e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

Obs.:

$$\left.\begin{array}{l} f(e_1) = \begin{pmatrix} \frac{\sqrt{3}}{2} \\ \frac{1}{2} \end{pmatrix} \\ f(e_2) = \begin{pmatrix} -\frac{1}{2} \\ \frac{\sqrt{3}}{2} \end{pmatrix} \end{array}\right\} \text{by theorem 4.2,}$$

$$f\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{3}}{2} x_1 - \frac{1}{2} x_2 \\ \frac{1}{2} x_1 + \frac{\sqrt{3}}{2} x_2 \end{pmatrix}$$

for any vector $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ in $\mathbb{R}^2$.

data 20 · 07 · 2019
S T Q Q S ☒ D

Obs: Use este exemplo para calcular o ângulo de 30° do item 11 (19/07/2019 deste book).
Ref.: Inorganic Chemistry
sp hybridization: a scheme for linear species. (Fig. 5.3)

17) Handbook of micromechanics and nanomechanics
Source: 20190721/Important/Shaofan....Handbook.....pdf
Pag. 76 (103)

3.2.1 Free Energy and the Constitutive Relations

ferroelectricity in crystals

$$\sigma_{ij} = c^{E}_{ijmn}(\varepsilon_{mn} - \varepsilon^{sp}_{mn}) - e_{mij} E_m$$
$$D_K = e_{Kmn}(\varepsilon_{mn} - \varepsilon^{sp}_{mn}) + \kappa^{\varepsilon}_{Km} E_m + P^{sp}_{K}, \quad (3.1)$$

or alternatively

$$\varepsilon_{ij} = s^{E}_{ijmn}\sigma_{mn} + d_{mij} E_m + \varepsilon^{sp}_{ij}$$
$$D_K = d_{Kmn}\sigma_{mn} + \kappa^{\sigma}_{Km} E_m + P^{sp}_{K},$$

$(i, j, K = 1, 2, 3)$ (3.2)

where σ_{ij}, ε_{ij}, D_i, and E_j are the

20.07.2019

components of the Cauchy stress, total mechanical strain, eletric displacement and eletric field, respectively.

The material coefficients in Eqs. (3.1) and (3.2), i.e.,

c_{ijkl} (s_{ijkl}), e_{ijk} (d_{ijk}), and κ_{ij},

denote the elastic moduli (compliances), piezoelectric stress (strain) constants, and the absolute dielectric permittivity, respectively.

The superscripts E, ε and σ indicate that the coefficients are measured at zero or constant eletric fields (short circuit), mechanical strain (clamped) and mechanical stress (load free), respectively.

veja o resto...

18) Comparando 2 livros.

a) Principles and Practices of Molecular Properties (2018).pdf pag. 37(46) + Particles and Fields

Figure 3.2. Decomposition of vector fields $F(r) =$

$F_r(r)e_r + F_\phi(r)e_\phi$ into irrotational

$F_\parallel(r)$ and solenoidal $F_\perp(r)$ parts.

20.07.2019

Source: Data-PC-2017/ Cálculo-important

Book: https://lccn.loc.gov/2017041617

Author: Patrick Norman,
Kenneth Ruud,
Trond Saue

ISBN 9781118794838 (pdf)

b) Vectorial Optics Fields

Source: 20170721/ Vector Optical Fields
Fundamentals and Applications/
Qiwen... Vectorial... pdf

Pág. 33(46)

Fig. 3 Generated vector fields with the topological charge $m = 1$, for different values of $\varphi_0 = 0, \pi/4, \pi/2$, and $3\pi/4$.
The top row shows the intensity patterns and the SOP distributions.
The bottom row shows the intensity patterns behind a horizontal polarizer.

Look at Fig. 3 — 15 (Pág. 47(60))

Devemos complementar um livro no outro, ou seja, fundi-los mas abrangem campos diferentes de abordagem. São estudos de vetores, um fala em decomposição e o outro em óptica (ondas).

data 20.07.2019

S T Q Q S [✓ S] D

19) Falando em Vector (item 18), Vamos ver Vector Spaces

Book: Linear Algebra - Pure & Applied_2013.pdf

Source: Data_PC_2017/Cálculo_important/

Pag. 331(348) — 399(416)

Veja também o livro:

2006_metodos_matematicos_avanzados_para_cientificos_e_ingenieros.pdf

Source: Data_PC_2017/Cálculo_important/

Pag. 10(26) — 17(33)

1.3 Espacios vectoriales y operadores lineales

El operador de Sturm-Liouville se define como el operador diferencial lineal

$$\mathcal{L} \equiv \frac{1}{r(x)} \frac{d}{dx}\left[p(x)\frac{d}{dx}\right] + \frac{q(x)}{r(x)} \qquad (1.21)$$

de modo que la ecuación de Sturm-Liouville puede escribirse como

$$\mathcal{L}\, y(x) = -\lambda\, y(x). \qquad (1.22)$$

Jandaia

data 21.07.2019
S T Q Q S S X

20) Geometría Axiomática
Geometría plana neutral, euclidea y analítica

Source: Data_PC_2017/Calculo_important/
2019-geometria...pdf

Pág. 202-204.
8.4 El Teorema de Pitágoras
(sem demonstração)

Pág. 283
11.7 La recta de Euler

Pág. 322
12.4.5 Teorema
Todo espacio proyectivo tridimensional es arguesiano.

Important:

No texto:

As contribuições de Leibniz para o pensamento da multiplicidade na obra de Michel Serres.
(filósofo)

Autora: Maria Emanuela Esteves dos Santos

21.07.2019 X

Na obra de Serres de 1968 na França,
Le Système de Leibniz et ses modèles mathématiques,
na comparação que ele faz entre Leibniz e Pascal e a questão do ponto de referência.

Pág. 5 (105)

Revista Prometeus - Ano 11 -
Número 26 -
Janeiro - Abril/2018
E-ISSN: 2176-5960

"O recurso à matemática ganha uma nova dimensão quando a geometria arguesiana - isto é, a geometria da infinidade - substitui a geometria finita dos gregos.

"O cálculo infinitesimal invadiu todos os espaços, incluindo a trigonometria tradicional." (Serres, 1968, p. 650).[3]

3 - "um sólido infinitamente grande terá um centro de gravidade indeterminado, bem como o movimento sobre uma circunferência infinita não será centrado, e sim um movimento retilíneo" (Serres, 1968, p. 649).

21.07.2019

Daí, então, a associação entre cálculo infinitesimal e espaço descentrado.

Source: 20190721/Espaço_arguesiano/8612-24625-1-PB.pdf

olhe:

Tópicos de História da Matemática através de Problemas
Pierre Petin

Source: 20190721/Espaço-arguesiano/Pierre-1.pdf

Pag. 350

"Será necessário esperar o final do século XVIII para que a geometria arguesiana encontre interesse nos matemáticos, pela criação da Geometria Descritiva de Gaspard Monge (francês – 1746 a 1818) e pela obra de Lazare Carnot (francês – 1786 a 1832) "Da correlação das figuras em geometria" (1801). ✓

Veja também:

Pag. 241 — projeto de topologia

Leibniz através da chamada analisis situs (expressão

data 21 04 2019

S T Q Q S S [✓ D]

latina que quer dizer análise de situação, mas foi Riemann que lançou as bases na célebre monografia

"Sobre as hipóteses que servem de fundamentos à geometria" (1867)

Obs.: Topologia foi empregada pela primeira vez por
Johann Listing (alemão - 1808 à 1882)

Pág. 275 -

Leibniz observa que Euclides às vezes utiliza a intuição geométrica para encobrir a ausência de certos postulados, por exemplo no seu método de construção do triângulo equilátero

(Proposição 1 - Livro 1).

Cantor e Dedekind mostram a necessidade de um "axioma de continuidade" e o formalizam:

"Se considera uma sequência de segmentos encaixados não vazios $[a_n, b_n]$ cujo comprimento tende para 0. Então, existe um ponto comum único a todos

data 21 . 07 . 2019

S T Q Q S S X

segmentos, o qual corresponde ou a um número racional, ou a um número irracional".

Igualmente, Moritz Pasch " (alemão - 1843 à 1930) e o "axioma de ordem":

"Sejam A, B e C três pontos não alinhados e r uma reta do plano ABC que não passa por nenhum dos pontos A, B e C; se a reta r passa por um dos pontos do segmento AB, ela passa ou por um ponto do segmento BC ou por um ponto do segmento AC".

(T20190721 0055)

21) Reta Arguesiana

Book: Tópicos de Geometria

José Nuno Tavares

Source: 20190721/Espaço_arguesiano/Topicos_Geometria.pdf

Pág. 34 (7. Teorema de Desargues e Pappus)

21.07.2019 x

Dois triângulos dizem-se perspectivos quando as retas que unem os vértices homólogos concorrem todos num mesmo ponto (o centro de perspectiva).

O teorema de Desargues pode enunciar-se na forma

"os pontos de intersecção dos lados homólogos de dois triângulos perspectivos estão sobre uma mesma reta (reta arguesiana)".

A prova:

espaço projectivo tridimensional es arguesiano está na pág. 322 do livro:

"2019_geometria_axiomatica_geometria_plana_neutral_euclidiana_y_analitica.pdf
Autor: Gerard Romo Garrido

Veja também:

Pág. 320 12.4.4 Teorema

Teorema de Hessenberg

"Pappusian planes are Desarguesian"
(G. Hessenberg, Mathematische Annalen, 61 (1905) pp. 161-172)

21.07.2019

El Teorema de Desargues se puede demonstrar utilizando única y exclusivamente el Teorema de Pappus y los axiomas incidentales.

(201907210115)

22) Advanced Engineering Electromagnetics
Source: 20190721_02/Advanced Engineering Electromagnetics/
CA Balanis_advanced...pdf
Pag. 84 (106)

Fig 2-22 Precession of spinning electron caused by applied field CW and CCW polarizations.

(a) Clockwise
(b) Counterclockwise

other: Foundations for Microwave Engineering
Source: 20190721_03/Collin, Foundations for Microwave Engineering/Robert E. Collin-Foundations....pdf
Obs: Livro usado como referência do primeiro livro.

23) Computer_Graphics_and_Geometric_Modelling_(2005).pdf
Source: Data_PC_2017/Cálculo_important/Computer...pdf

21.07.2019

24) An Introduction to General Relativity and Cosmology.
Source: 20190721_03/An Introduction to General Relativity and Cosmology/An Introduction... pdf

25) 20190722

James Gregory

James Gregory wrote to Collins on 19 December 1670 that "In my last to you,..."
Source:
The Mathematical Paper of Isaac Newton Vol. 04 Pag. 414
https://books.google.com.br/books?isbn=0521045835
D.T. Whiteside (2008).

other: Mathematics Emerging:
A Sourcebook 1540-1900
by Jacqueline Stedall
Pag. 198

8.1.4 Gregory's binomial expansion, 1670

In March 1670 John Collins sent Newton's series for the quadrature of the circle to James Gregory in Edinburgh and Gregory replied on 23 November with some further series of his own, saying: "I suppose..."

data 22.07.2019

Look: Madhava series (135$\overset{x}{0}$-1425)

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots$$

$$\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots$$

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots$$ (Leibniz.

other:

$$\pi = \sqrt{12}\left(1 - \frac{1}{3\cdot 3} + \frac{1}{3^2\cdot 5} - \frac{1}{3^3\cdot 7} + \dots\right)$$

This is obtained by substituting $q = 1/\sqrt{3}$ (therefore $\theta = \pi/6$) in the power series expansion for $\tan^{-1} q$ above.

Voltando:

Gregory's Formula

is a formula that allows a definite integral of a function to be expressed by its sum and differences, or its sum by its integral and difference (Jordan, 1965, p.284)

data 22.07.2019

☒ T Q Q S S D

Reference:

- Jordan, C. "Gregory's Summation Formula," §99 in Calculus of Finite Differences, 3rd ed. New York: Chelsea, pp. 284-287, 1965
- Roman, S. The Umbral Calculus. New York: Academic Press, p. 59, 1984

Gregory's Formula

mathworld.wolfram.com/GregorysFormula.html

$$\int_0^y p(u)\,du = \sum_{k\geq 0} \frac{((e^{yt}-1)^k \mid p(x))}{k!} (e^t-1)^k p(x)$$

discovered by Gregory in 1670 and reported to be earliest formula in numerical integration (Jordan 1965, Roman 1984)

Source:

20190722-03 (20190722-03) / Calculus of Finite Differences_Jordan / Calculus of Finite Differences_Jordan.pdf

Pag. 284 (304) - 287 (307)

§99 Gregory's Summation Formula

Obs: C is Euler's constant.

data 22.07.2019
* T Q Q S S D

Veja: Gregory coefficients

26) Newton's letters to Leibniz, 1676
Source:
20190722_03/ The Mathematical Papers of Isaac Newton/ Jacqueline Stedall - Mathematics Emerging_a_sourcebook 1540-1900 ... (208).pdf

Pag. 196 (219)

8.1.3 Newton's letters to Leibniz, 1676

27) Determinação do número pi (π) por meio de uma rede quadrada de resistores idênticos.
Source:
http://dx.doi.org/10.1590/1806-9126-RBEF-2017-0142

20190722_03/ Adventures in Formalism/ 1806-1117-rbef-40-02-e2304.pdf

28) Angular momentum in quantum physics
L.C. Biedenharn
Source:
20190722/ Advanced quantum theory/ Angular momentum in quant.pdf

data 22 07 2019

29) Chapters in Mathematics
From π to Pell - Volume 3
Craig Smoryński
Source: 20190722-03/
Adventures in Formalism/
Craig Smorynski - Chapters
in Mathematics...pdf

30) 20190723

Cubature Formula

Source: 20190723/Euclidean Design
theory/sanet-cd-9811380740

Pág. 19(27)

A cubature formula reveals a numerical integration rule that approximates a multiple integral by a positive linear combination of functions values at finitely many specified points on the integral domain.

31) 20190724

Advances in numerical quadrature
Herbert S. Wilf
Source:
20190724/Advances in numerical
quadrature Wilf.pdf/Ralston...
Vol. 2 (1967).pdf.
Pag 133 (148)

data 24·07·2019

32) 20190724

"Somente um principiante que não sabe nada sobre ciência diria que a ciência descarta a fé. Se você realmente estudar a ciência, ela certamente o levará para mais perto de Deus".

James Clerk Maxwell

33) 20190725

Modern Celestial Mechanics
Morbidelli
Source:
20190725/Modern Celestial Mechanics/
Alessandro Morbidelli Modern....pdf
Pag. 15 (19)(28 de 370)

Fig. 1.2 Keplerian motion: definition of i, Ω and ω.

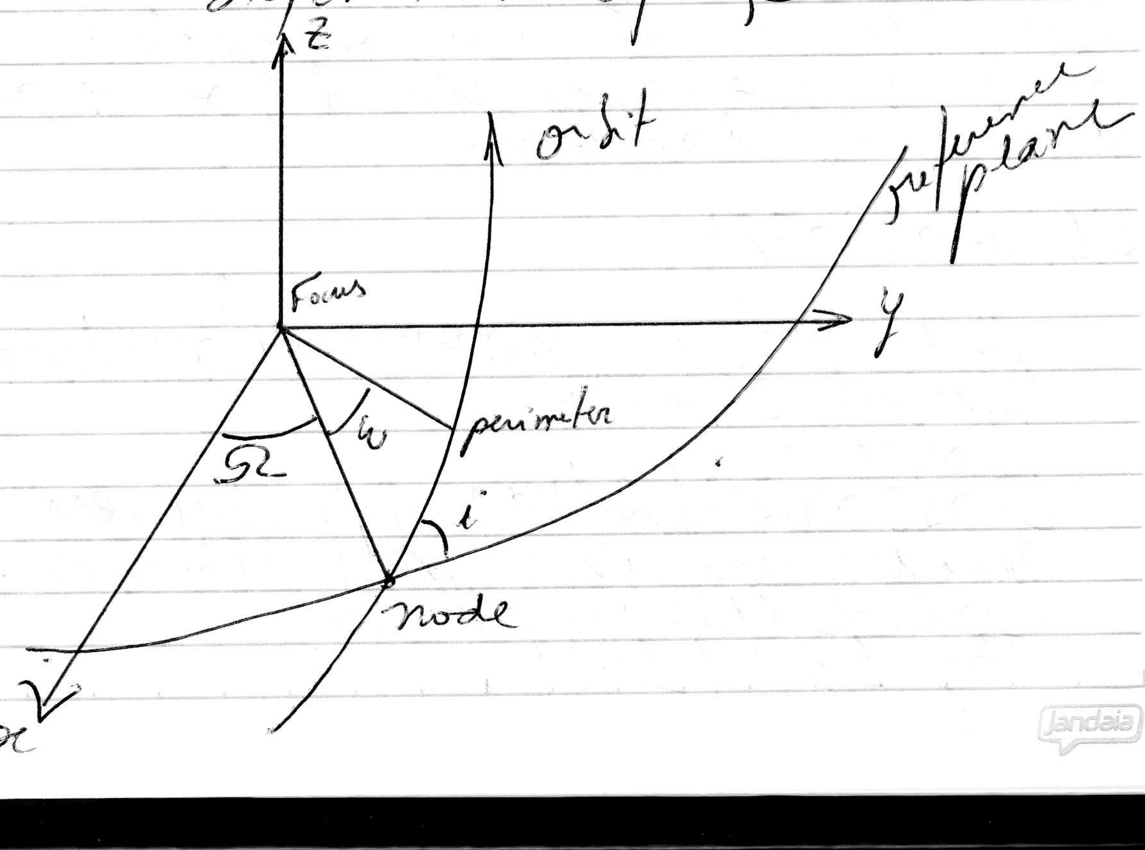
z
orbit
reference plane
Focus
y
perimeter
ω
Ω
i
node
x

date 25.07.2019

S T Q [✓ Q] S S D

34) Modern Celestial Mechanics

Source: 20190725/Modern Celestial Mechanics/modern Celestial mechanics_from theory to applications....pdf

Pag. 13 (23)

Fig 9. Layout of the six accelerometers in the so-called "asymmetric diamond" configuration.

Look at: Pag. 21

Fig. 1 the geocentric reference frame with the new spherical coordinates for the satellite.

35) Celestial Coordinate System (graph)

Source: 20190725_02/Merriam-Webster's Visual Dictionary/Merriam Webster's....pdf

Pag. 13 (25)

ecliptic → projection of the Sun's apparent annual path onto the celestial sphere.

right ascension → one of two coordinates used to locate a celestial body on the celestial sphere, similar to terrestrial longitude.

data 25 . 07 . 2019

36) Thermodynamics in Earth
Source:
20190726/ thermodynamics in Earth.../
thermodynamics in Earth....pdf
Pág. 34 (40) Letra d => an electron has a spin quantum number, m_s, of +1/2 or -1/2.
Fig. 1.3
The spatial orientations of the five d-orbitals that are energetically degenerate in a spherically symmetric environment. The form of the d_z^2 orbital can be generated by rotation about the z axis. From Orgel (1966)

—

37) 20190726

Practical Guide to Lightcurve Photometry
Source:
20190727/ A Practical Guide to Lightcurve Photometry.../ A Practical Guide....pdf
Pág. 196 (213)
15.1.3 Correcting for Changing Asteroid-Earth-Sun Geometry and Phase Angle.
Fig. 15.2

38) Linear Algebra for Computational Sciences and Engineering
Source:
20190727/ Linear Algebra for Computational.../ Ferrante Neri....pdf
Pág. 131 (146)
5.1. Complex Numbers

date 26.07.2019

39) Geometric Vectors - Chapter 4
Source:
20170326/Linear Algebra for Compu-
tational.../303021320X.pdf
Pág. 133 (153)

40) SF1624 Algebra and Geometry -
Introduction to Linear Algebra
for Science & Engineering - Pearson
2017
Source:
20190727/Introduction to Linear
Algebra for Science and Engineering/
SF1624... pdf.

41) 20190727

a) The Euler Line and the Nine-Point Circle

Look: 2.1.1 Homothety

Source: Data_PC_2017/Cálculo_important/
Vectors/Yiu P. - Introduction
to the geometry of the triangle
(2002).pdf

Pág. 15(18) - Chapter 2

Obs: Usaram Angle Bisectors

b) Theorem 1.3.2 (Angle Bisectors).

In any triangle, the lines which
bisect the three interior angles

data 27.07.2019

S T Q Q S [✓ S] D

meet in a common point.

Figure 1.18

Source: Data_PC_2017/ Cálculo_important/

42) Desplazamiento de los escarabajos a lo largo de la mesa.

Fig. 1.3

Pag. 16(30)

Source: Data_PC_2017/ Cálculo_important/ Instantáneas diferenciales.pdf.

43) Elliptic Curves

Source:
033_NOVOSARQUIVOS_2019/ Elliptic_Curves/
Elliptic_Curves.pdf

Pag. 3(18)

Figure 1

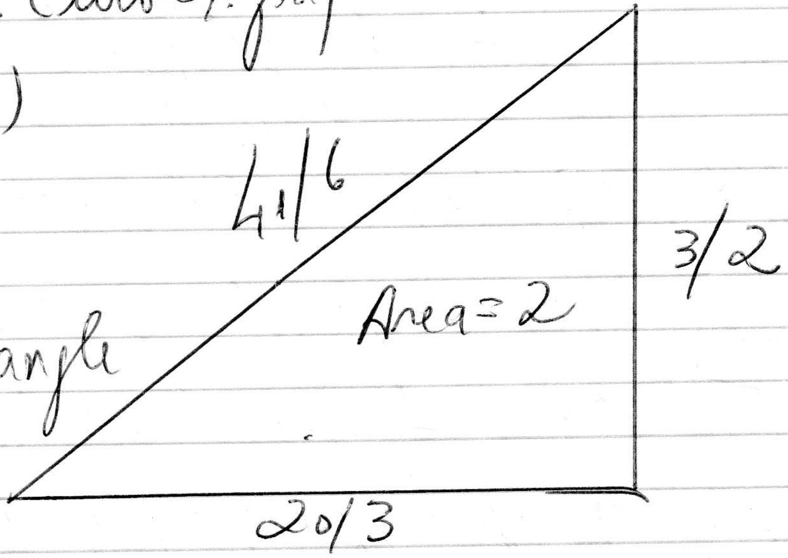

A right triangle of area 5 and rational sides.

data 27.07.2019

44) An invitation to Arithmetic Geometry

7. The tangent line at a point of a projective curve

Pag. 209 (227)

Source:
033_NOVOSARQUIVOS_2019/An Invitation to Arithmetic Geometry/[Dino_Lorenzini]_An_Invitation..._pdf.

45) 20190729

Traité de Géométrie Vol. 1
Rouché

Pag. 245-246

348. Méthode des Polaires Réciproques

Fig. 223

Obs.:

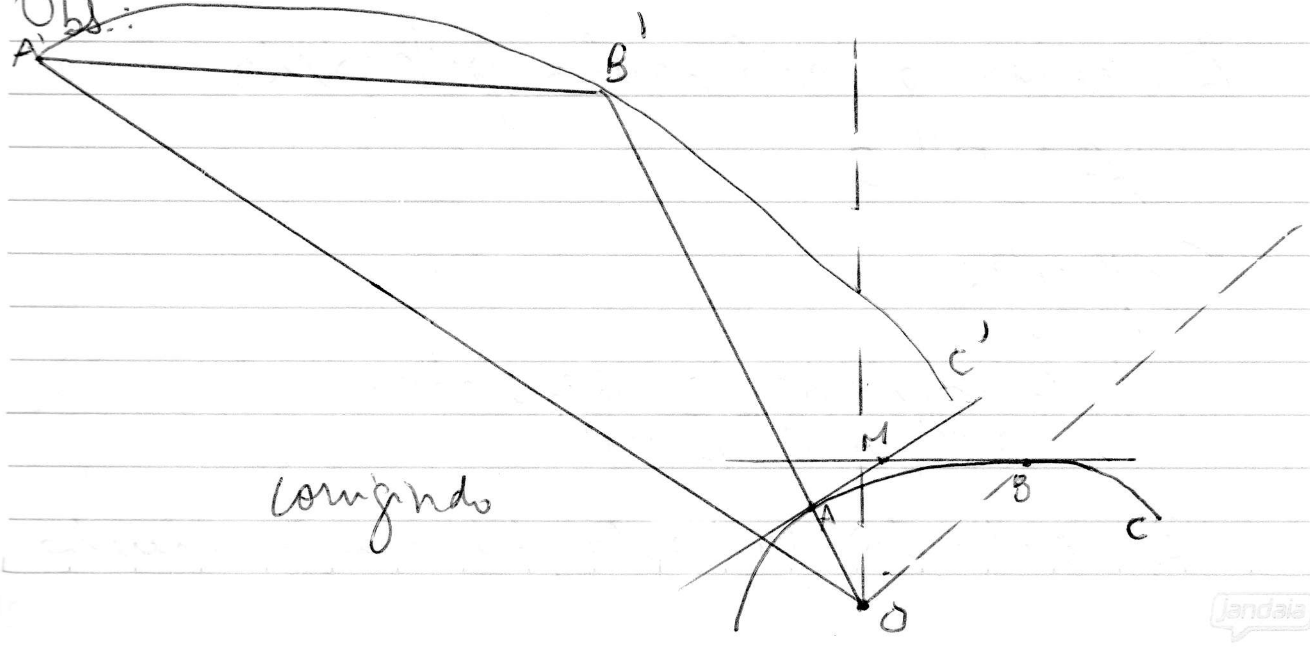

data 29 . 07 . 2019

Fig. 223

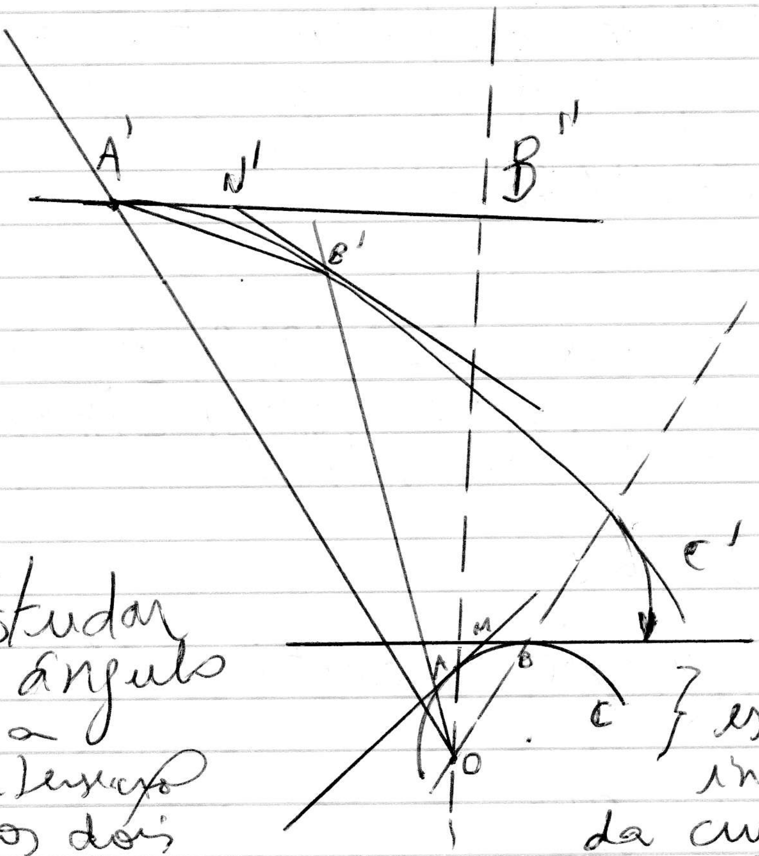

Obs: Estudar o ângulo e a intersecção dos dois círculos

} estudar a intersecção da curva A' B' C' na curva ABC (circle)

Calcular o tamanho das retas $\overline{OA'}$, $\overline{OB'}$ e suas razões.

Estudar o triângulo pitagórico

A' B'' O

Calcule:

1) área
2) comprimentos de reta OA', A'B'' e OB''

3) Calcule sua relação com o círculo

data 29 7 2019
X Ⓣ Ⓠ Ⓠ Ⓢ Ⓢ Ⓓ

A', B' e C'.

Important:

Utilize esses cálculos para estudar a simulação para cálculo do radiano.

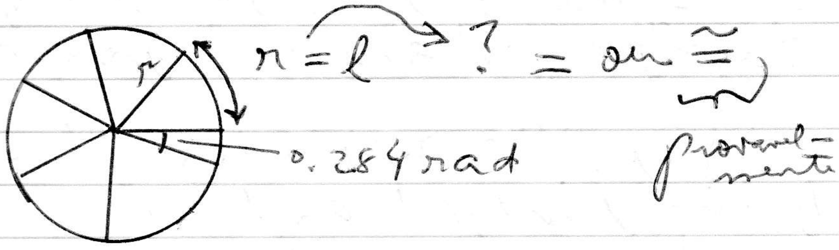

$1 \text{ rad} = 360°/2\pi = 57°.29578$

$= 206265$ arcseconds

(2π rad = 360°)

Obs.: Calcular corretamente
r → é um linha reta (com equação)

arco (≅r) l → é um arco e não uma reta, o seu comprimento deve ser diferente do raio r.

(Calcular com precisão)

O ângulo θ (in radians) = l/r
or
θ (in arcseconds) = 206265 l/r

Obs.: The relationship between the limiting resolution of a lens and its aperture is expressed mathematically by the equation:

data 29.07.2019
[✓ S] T Q Q S S D

$$R = 206265 \times \frac{1.22\,\lambda}{d}$$

where

R is the minimum angular separation of a pair of star in order to be just resolve, measured in arcseconds.

λ is the wavelenght of the light being imaged and

d is the diameter of the lens.

Obs.: λ and d are measured in the same units.
(metros or centimeters)

Source:
20190726/Advanced amateur astronomy/
Gerald North_Advanced_Amateur_
Astronomy.pdf

Voltando ao livro de Rouché:

a) Pág. 196

Problème 284

Connaissant le coté d'un polygone régulier inscrit, calculer le côté du polygone régulier

29.07.2019

circonscrit semblable (fig. 196)

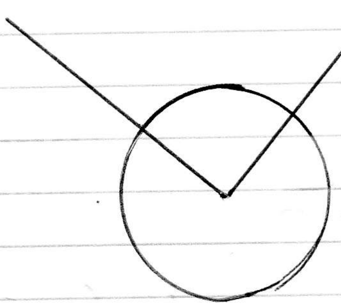

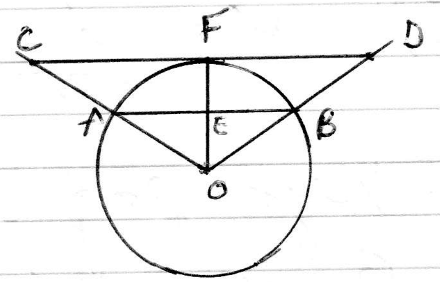

Soient AB = a le côté donné et R le rayon du cercle; menons la tangent CD au milieu F de l'arc AB et prolongeons-la jusqu'aux points C et D, où elle rencontre les rayons OA et OB prolongés.

CD sera le côté cherché (265); désignons-le par α.

Les triangles semblables AOE, COF, donnent

$$\frac{CF}{AE}=\frac{OF}{OE} \quad \text{ou} \quad \frac{\alpha}{a}=\frac{R}{OE};$$

mais on a (282)

$$OE=\frac{1}{2}\sqrt{4R^2-a^2}$$

il en résulte

$$(I) \quad \alpha=\frac{2aR}{\sqrt{4R^2-a^2}}$$

29.07.2019

En particulier, lorsqu'on prend le rayon pour unité, on a

(II) $$\alpha = \frac{2a}{\sqrt{4-a^2}}$$

b) Veja pág. 200

§ VIII — Mesure de la Circonférence

291. Considérons (fig. 198) un arc ....

La longueur d'un arc de courbe est la limite vers ~~l~~ laquelle tend le périmètre d'une ligne brisée inscrite dans cet arc, lorsque les côtés de cette ligne tendent vers zéro.

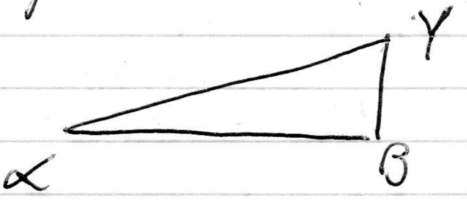

Estudar até a página 215.///

Important:

Veja se o arco $\hat{AB}$ é menor que a reta $\hat{CD}$.
Faz uma simulação para saber quando o arco $\hat{AB}$ é igual a reta $\hat{CD}$ e calcule o ângulo quando isto ocorrer, se ocorrer, pois conforme diminuímos o ângulo menor será a reta $\hat{CD}$ e por conseguinte, o arco $\hat{AB}$. Testar a hipótese. Muito interessante!!! —

data 30 07 2019

S [✓ T] Q Q S S D

46) Ordinary Taylor expansion

Source:
20190730_03/Differential Geometry of Curves and Surfaces/[Hunt,_Bruce;...]-Differential_Geom....pdf

Pág. 19(32)- Normal Plane
Rectifying plane.
Osculating plane.

Fig. 2.5. Three projections of the space curve $xe_1 + ye_2 + ze_3$

47) Applied linear algebra and matrix analysis
Source:
20190730_03/Applied Linear Algebra and Matrix Analysis/Shores....pdf.

Pág. 277(285)-

Geometrical Aspects of Standard Spaces.

48) Differential Equations
Krantz
Source:
20190730_03/Differential Equations.../
Krantz...2007.pdf
Obs: Important.
Completa apresentação das equações diferenciais, abordando também teoria não linear e sistemas dinâmicos.

data 30.07.2019

49) Differential Equations with applications and historical notes
Simmons.
20190730-03/ Differential Equations with Applications and Historical notes/ [Georgem]-Differential...pdf

Pág. 16 (35)

Figure 5 — $r d\theta$ e dr

$$\psi = r\,d\theta/dr$$

Veja gráfico.

Obs: Muito interessante.

50) 100 years of Math

Source: 20190730_03/ 100 Years of Math Milestones.../sanet.st_100 years of Math.pdf

Pág. 10 (25)

Fig. 2 Analysis of the removed cylinder from the sphere

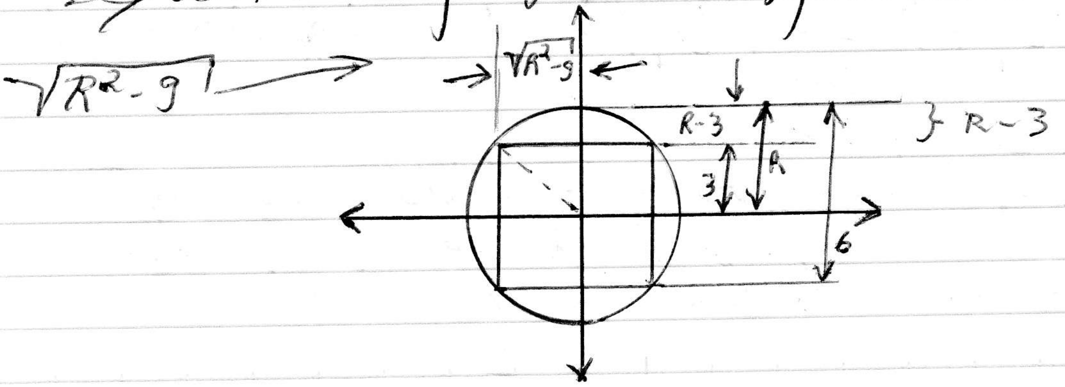

data 30 07 2019

51) Tennenbaum's Proof of the Irrationality of √2

Source: 20190730_02/100 years.../100 years...pdf

Pág. 205(220)

Obs: Ver Fig. 2 → Triângulo retângulo e a sua prova.

52) Representação geométrica de $z = 1 + i$

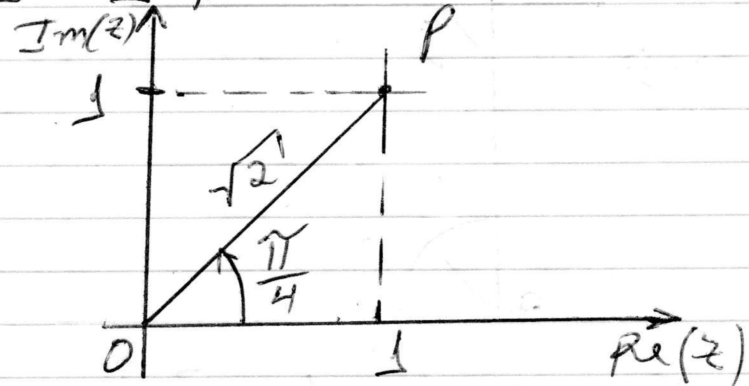

Source: 20190730/Important/Trigonometria e numeros complexos md.pdf

Pág. 216

—

53) 20190731

An illustration of the quantities in the proof of Lemma 2.37

Veja gráfico:

data 31.07.2019

Source: 20190801/ Conditioning the Geometry.../ Conditioning.pdf

Pág. 45(73)

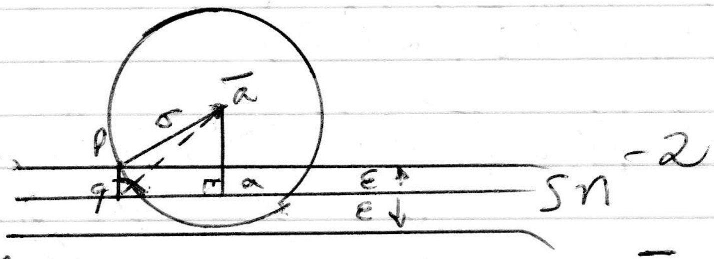

Veja ângulo:
$\bar{a}$ e $\bar{q}$ →

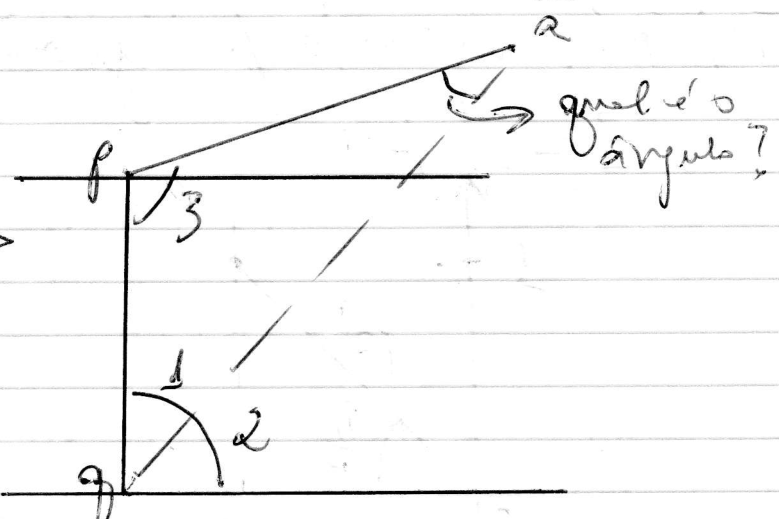

qual é o ângulo?

— Quais são os ângulos 1 e 2?
Calcular!!!

— Veja qual é o ângulo do ponto p(3)?

— Calcular a área do triângulo.

$\bar{a}$ p q

— Faça simulação com a linha de baixo
(calcule os ângulos e as áreas)

→

data 31.07.2019 ☐S ☐T ☒Q ☐Q ☐S ☐S ☐D

Simulação: 2 hipóteses

A)

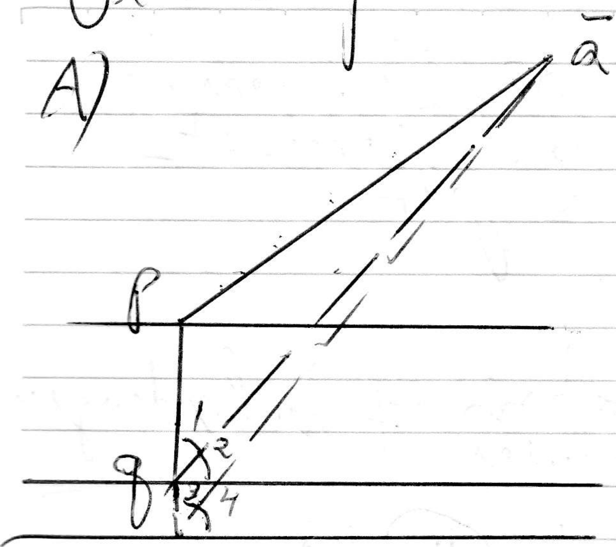

Primeira simulação: use os pontos p e q.

B)

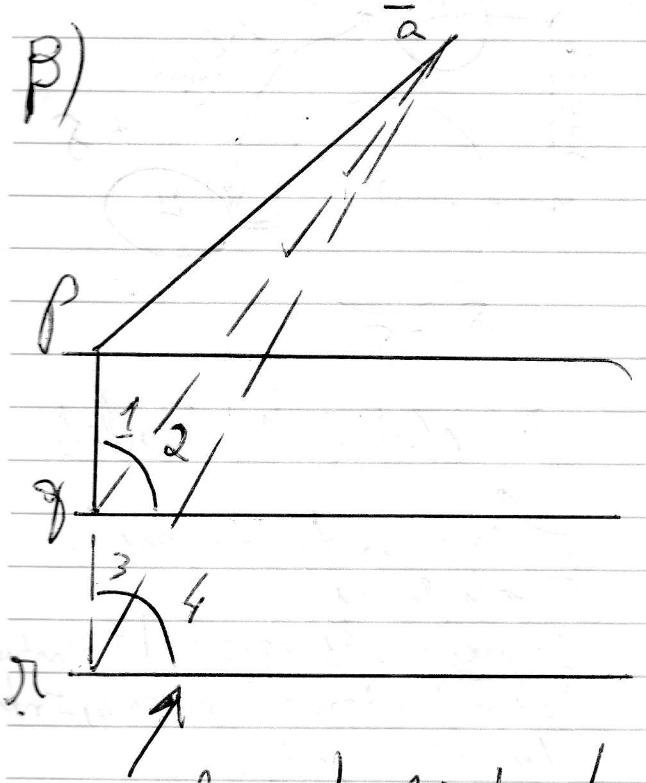

Segunda simulação: use os pontos p, q e r, sendo que no ponto r a linha reta está tangenciando o círculo cujo centro ā e posteriormente, com a reta afastada do círculo.

Simule esta linha tangenciando o círculo com centro ā e com a reta afastada do círculo.

Lemma 2.37 Let $0 < \varepsilon, \sigma < 1$, $n > 2$, and $\bar{a} \in S^{n-1}$. Fix also an embedding $S^{n-2} \hookrightarrow S^{n-1}$.

(veja demonstração depois)

data 05.08.2019

54) Atoms and Molecules

Source: 20190805_02/Engineering of Micro Nano../janet_st_r 9811365482.pdf

Pag. 27(35)

Fig. 2.16 Representation of Dihydrogen and water molecules

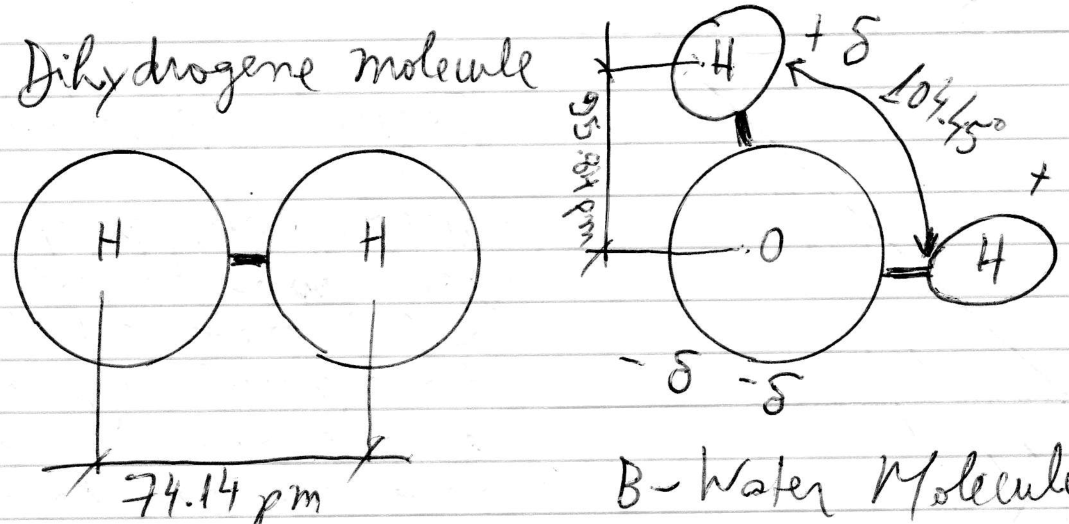

A - Dihydrogene molecule

B - Water Molecules

55) Bedded Down - Thermodynamic Stability

Pag. 125(142)

Source: 20190805_02/Introduction to Coordination Chemistry/Introduction... pdf.

$$K = \frac{[ML^{n+}]}{[M^{n+}][L]}$$

Equilibrium constant K in this case is called the "formation constant or stability constant"...

⇒ There is also a direct relationship between the stability constant and the

data 05.08.2019

free energy (ΔG^0, in kJ mol^{-1})^X of a reaction expressed in terms of the relationship (5.6):

$$\Delta G^0 = \underbrace{-2.303}_{\text{Mudono porquê?}} \cdot R \cdot T \cdot \log_{10} K \quad (5.6)$$

56) Dielectric Properties of Biomaterial

Source: 20190805_02/Important/sanet_st_9811365482.pdf

Pág. 285 (291) — 6.4.2.2

induced by an electrical field applied.

$$\varepsilon = \varepsilon' - j\,\varepsilon'' = \varepsilon_0\,\varepsilon_r - j\,\frac{\sigma}{\omega} \quad (6.114)$$

In this complex expression, $j = \sqrt{-1}$ and ω is the angular frequency.

—

Pág. 322 (328)

7.3.2 Fluidic Drag Force

$$\vec{F}_d = 6\pi\eta R_p(\vec{v}_f - \vec{v}_p)\,f_{(D)} \quad (7.25)$$

data 05.08.2019

Estudar esta equação (Pag. 323)

$$f_D = \left[1 - \frac{9}{16}\left(\frac{R_p}{R_p+z}\right) + \frac{1}{8}\left(\frac{R_p}{R_p+z}\right)^{3} - \frac{45}{256}\left(\frac{R_p}{R_p+z}\right)^{4} - \frac{1}{16}\left(\frac{R_p}{R_p+z}\right)^{5}\right]^{-1} \quad (7.26$$

estudar os motivos que estabeleceram essas razões.

analisar as suas relações.

Obs.:

R_p is the radius of the particle,

$\vec{V}_p$ its the velocity,

η and $\vec{V}_f$ respectively the viscosity of the external medium and the velocity of the fluid

f_D is the drag (arrasto) coefficient of the particle that takes into account the influence of a solid wall in the vicinity of the particle,

z being the distance of the particle to the wall

Obs.: f_D varies beetween 1, for particle distant from the wall to 3 for

data 05.08.2019

particle in contact with micro-channel wall (z=0).

57) Conic Sections

Source: 20190805_02/Important/Linear Algebra through Geometry - Springer...pdf

Pág. 95 (105)

Fig. 2.52

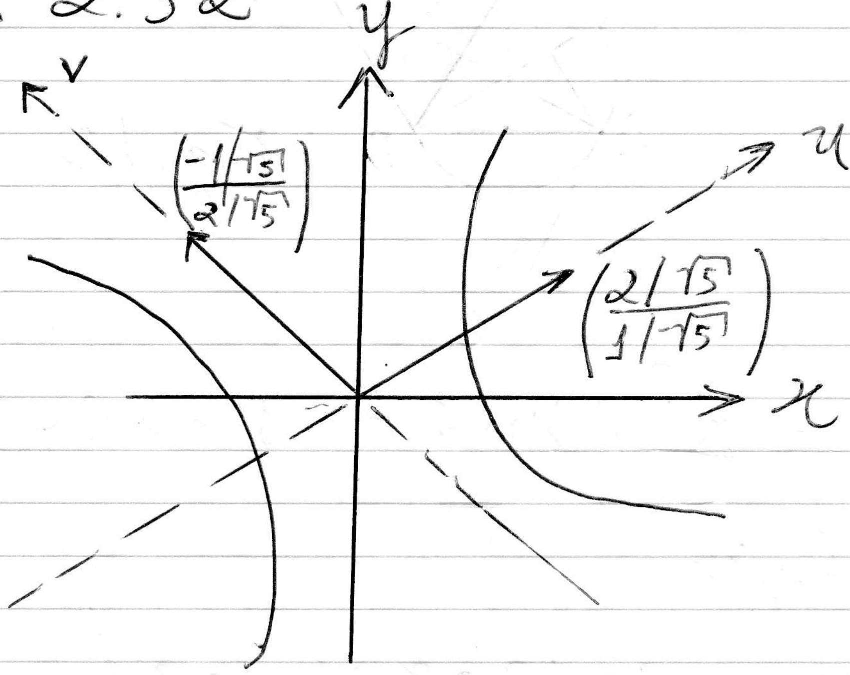
v
y
u
$\begin{pmatrix}-1/\sqrt{5}\\ 2/\sqrt{5}\end{pmatrix}$
$\begin{pmatrix}2/\sqrt{5}\\ 1/\sqrt{5}\end{pmatrix}$
x

58) 20190806

Traité de géométrie précédé...

Source: 20190806_03/Courbes/traité_de_géométrie_précédé...bpt6K920521x.pdf

Pág. 52 (63) →

data 06.08.2019

S [✓ T] Q Q S S D

$$x = -\frac{1}{2}(a+b) \pm \frac{1}{2}\sqrt{(a-b)^2 + \frac{4ab\sin(\alpha+\beta)\sin(\beta+\gamma)}{\sin\alpha\sin\gamma}}$$

Surface d'un triangle Pág. 53

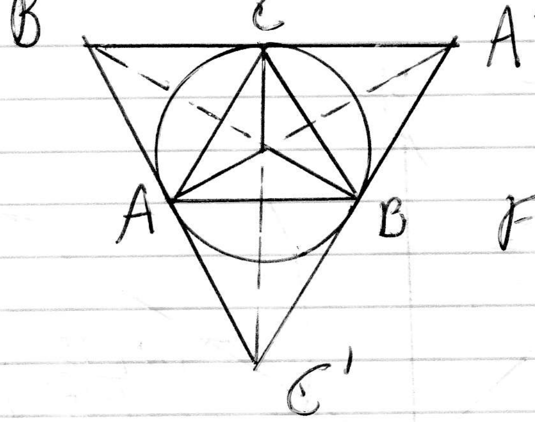

Fig. 15

$$ABC = \frac{R^2}{2}(\sin 2A + \sin 2B + \sin 2C)$$

59) Ellipse

Source: 20190806_03/
Lindermann_Ferdinand_1852-1939/
vorlgeovonalfre01clebuich.pdf

Pag. 8 (26) → Fig. 4

$$c^2 = a^2 - b^2$$

$$x^2 + y^2 = a^2$$

data 06.08.2019

S [✓ T] Q Q S S D

Gleichung der Curve (5)

$$r^2 = \frac{1}{\frac{\cos^2\varphi}{a^2} - \frac{\sin^2\varphi}{b^2}} ;$$

60) Hangover angle

Source: 20190807/Advances in Architektural Geometry/
(Architectural)

Pág. 85(91) → Fig. 5

Printing direction and hangover angle

61) Molecular Orbital theory

Source:
20190807/Molecular_Orbitals/
0470746680-Molecular.pdf

Pág. 25(39) - Equation 1.10

$$E = 2\beta\cos\frac{k\pi}{n+1} \qquad (1.10)$$

Fig. 1.31

Energies of π molecules orbitals in ethylene and the allyl system

data 06.08.2019

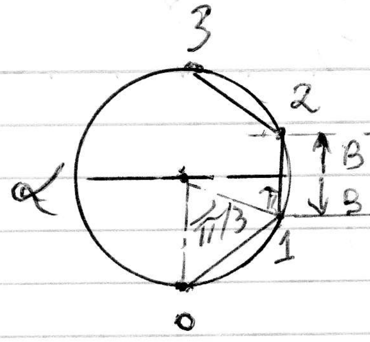

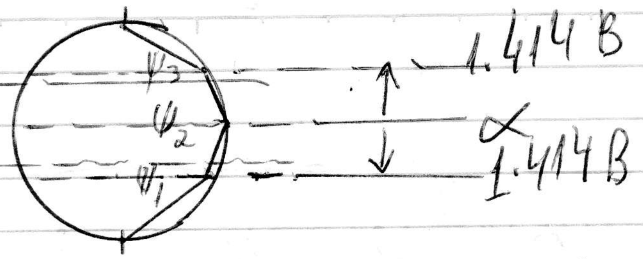

ethylene the allyl system

62) 20190812

I) $a = \frac{-1+\sqrt{2}}{2} = 0.618033988\ldots$

Source: 20190812/Analysis in Euclidean Space/ Calculus and Analysis
Ani Euclidean Spaces...
2016.pdf

Pág. 327(337)

6.7 Change of Variable

Conecta-se com

II) $\phi = \frac{\sqrt{5}+1}{2}$ (Golden Ratio)

$= 1.6180339887\ldots$

Source: 20190812/ Pi unleashed
Haenel/ Jörg ... Pi-unleashed
2001.pdf

Pág. 66(80)

-

data 12.08.2019

63) The Euler Circle

Source: 20190812/ Geometry Audin/ Audin M. Geometry-Springer (2002).pdf

Pág. 62(70)

Fig. 7 - The Euler Circle.

64) Trisecting the Angle (20190814)

Source: 20190814/ Field Theory and Its Classical Problems/ (Carus...)...Classical Problems.pdf

Pág. 26(43)

Section 1.3 - Trisecting the Angle.

65) Geometric Series of Areas

Source: 20190814/ Real Infinite Series, Reprint Edition/ sanet.st_Real Infinite Series.pdf

Pág. 11(24)

Figure 1.1 - A Geometric Series of Areas.

66) The Continued Fractions for $\sqrt{2}$

Source: 20190814/ Yearning for the

data 14.08.2019

impossible... / Stillwell ... (2006).pdf

Pág. 10 (30) -

1.6. The Continued Fraction for $\sqrt{2}$.

67) 20190815

Riemann Hypothesis (Proof)

Sir Michael Francis Atiyah

www.heidelberg-laureate-forum.org/blog/video/lecture-monday-september-24-2018-sir-michael-francis-atiyah/

Documents:

1) The Fine Structure Constant
17 pages

2) The Riemann Hypothesis
5 pages.

Source:

20190815
20190815-02

data 15.08.2019

S T Q [✓ Q] S S D

other ref.:

a) https://arxiv.org/pdf/math/0307136
Geometric Proof of Riemann Hypothesis

b) https://phys.org/news/2018-09-math-greatest-mysteries-riemann-hypothesis.html

c) www.scientificamerican.com/article/are-mathematicians-finall/

d) www.aps.org/publications/apsnews/201311/physicshistory.cfm

e) www.gizmodo.com/mathematicians-skeptical-of-supposed-million-dollar-pro-1829301425

f) www.preposterousuniverse.com/blog/2018/09/25/atiyah-and-the-fine-structure-constant/

(20190326). email sent

(15)
data 15.08.2019
S T Q [✓ Q] S S D

per Sean Carroll

... thanks to Weinberg and Salam that the weak and electromagnetic forces are unified into the electroweak theory.

—

~~68~~) Abstract: heidelberg-laureate-forum.org

The Riemann Hypothesis is a famous unsolved problem, dating from 1859.
A simple proof using a radically new approach.
It is based on work of

von Neumann (1936),
Hirzebruch (1954) and
Dirac (1928).

—

68) Fernando J. Corbató
July 1, 1926.
July 12, 2019.
Book: Spheroidal Wave Functions: including Tables of Separation Constants and Coefficients → 1956.

data 15.08.2019

S T Q [✓ Q] S S D

Author: Julius A. Stratton ...
Fernando J. Corbato

Source:
https://mitpress.mit.edu/books/
spheroidal-wave-functions.

ISBN: 9780262692915

613 pp - US$ 75.00

69) Constant nonlocal mean
curvatures surfaces and
related problems -
Mouhamed Fall
ICM 2018

Source: www.youtube.com/
watch/?v=|90MtX4MMj1
(l)

Source: 20190815_03/Constant..../

Other:

Spheroidal Wave functions in
Electromagnetic theory

Le-Wei Li. 2002 307 pp.

ISBN: 0471031704

data 15.08.2019

Source: 20190815-03/
Electromagnetic theory... /

70) Geometria Moderna (Pasch)
Pastor
Source: 20190815-03/Riemann/
Lecciones de geometria
moderna.pdf

71) Cada ângulo externo de um triângulo
é o mesmo soma de ângulos opostos

8.(Jeder Außenwinkel eines
Dreiecks ist gleich der
Summe der gegenüberliegenden
Winkel, $\measuredangle YCB = \alpha + \beta$

Pag. 227 (243) Fig. 90

Pag. 229 (245) Fig. 91

Source:

20180815_03/Weber, Heinrich,
1842-1913/
encyklomentmatik02,
weberich.pdf

72) Theorem 4.9

The ratio of the lengths of concentric
arcs of two horocycles intercepted
between two lines in the pencil is

data 15.08.2019

expressible in terms of an exponential function of the distance between these arcs.

Source: 20190815_03 / Riemann /
John McCleary - Geometry
from a Differentiable
Viewpoint... pdf

Pág. 50(67)

Pág. 52(69)-

Lambert produced a comprehensive study of the hyperbolic trigonometric functions containing the work of Euler and the Bernoullis (Lambert 1768).

cf: Lambert, J.H. Observations
Trigonométriques.
Pág. 327-354.

Similar developments were made by V. Riccati (1701-75) with Saladini for studying the solutions to cubics.

To define the hyperbolic trigonometric functions, formally introduce a complex variable into the Taylor series for e^x to get:

$$e^{iy} = 1 + iy - \frac{y^2}{2} - \frac{iy^3}{3} + \frac{y^4}{4} - \ldots$$

data 15.08.2019

S T Q [✓ Q] S S D

$$= \left(1 - \frac{y^2}{2} + \frac{y^4}{4} - \dots\right) + i\left(y - \frac{y^3}{3} + \frac{y^5}{5} - \dots\right)$$

$$= \cos y + i \sin y$$

It follows from these formules that:

$$\cos y = \frac{e^{iy} + e^{-iy}}{2},$$

$$\sin y = \frac{e^{iy} - e^{-iy}}{2i} \text{ and}$$

$$\tan y = \frac{\sin y}{\cos y} = \frac{1}{i}\frac{e^{iy} - e^{-iy}}{e^{iy} + e^{-iy}}.$$

By analogy with these formulas, we defined the real-valued functions:

$$\cosh x = \frac{e^x + e^{-x}}{2},$$

$$\sin x = \frac{e^x - e^{-x}}{2}, \text{ and}$$

$$\tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x - e^{-x}}.$$

From the power series, notice that

$$\cos(iy) = \cosh y, \text{ and}$$

Jandaia

data 15.08.2019

$\sin(iy) = i \sinh(y)$.

The fundamental relation for the hyperbolic trigonometric functions is given by:

$$\cosh^2 x - \sinh^2 x = 1.$$

The convergence theorems that support our formal manipulations are classical and can be found in any good complex analysis book, for example,

Complex Analysis, Ahlfors, 1979

Source: 20190816/Complex Analysis Ahlfors/

73) 20190816

Formules d'Euler

Source: serge.mehl.free.fr/chrono/Euler.html#cercle

$$e^{ix} = \cos x + i \sin x$$

$$\cos x = \frac{e^{ix} + e^{-ix}}{2}, \quad \sin x = \frac{e^{ix} - e^{-ix}}{2i}$$

Appliquée à $x = \pi$, la formule $e^{ix} = \cos x + i.\sin x$

$$e^{i\pi} + 1 = 0$$

(Obs. Veja item anterior → 72).

data 16.08.2019

X

$e^{\pi} - \pi$ est "presque" entier.

obs: Voyez aussi Ramanujan et le nombre

$$e^{\pi\sqrt{163}}$$

obs: Traité des Fonctions Elliptiques

Legendre

Source: www.archive.org/details/
traitedesfonctio01amle
02amle

Source: 20190816/Courbes Géométriques
Brocard pdf/
Screenshot_2019-08-16_Leonhard
Euler.png

74) Optique

1877 E. Sange

Veja a figura. Pág. 172 (02 of 20)

Source: 20190816/Courbes Géométriques
Brocard pdf/00000181.tif.20.pdf

data 16.08.2019

S T Q Q [✓ S] S D

Usando este gráfico, simularemos

2 círculos.

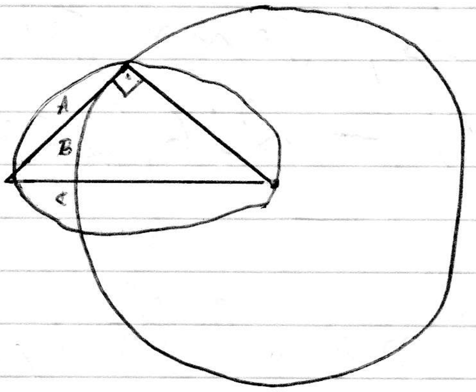

Calcular as áreas:
A, B, C

75) Choice of axes and basis vectors (2.4)
Pág. 22(37)
Source: 20190816/ Fundamentals of Physics Mechanics.../ Fundamentals of Physics....pdf
Pág. 22(37) e 23(38)

Fig. 2.4 The same vector A is written as

$i A_x + j A_y$ in one frame

and as $i' A'_x + j' A'_y$ in the other. The dotted lines indicate the components in two frames.

$i' = i \cos\phi + j \sin\phi$ (2.16)

$j' = j \cos\phi \underset{(-)}{\cancel{+}} i \sin\phi$ (2.17)

76) Complex Numbers and Quadratic Equations.

data 16.08.2019

Source: 20190816/ Senior Secondary School Mathematics for Class 11/ Senior Secondary...pdf

Pág. 150(161) - 216(227).

—

77) 20190817

Eigenfunctions of the Orbital Angular Momentum

Source: 20190818/ The Quantum Mechanics Conundrum .../ 3030166481.pdf

Pág. 75(96)-

Veja Fig. 1.23 → Quantisation of angular momentum for the case of $l=2$.

Fig. 1.24 → Relation between Cartesian (rectangular) coordinates (x, y, z) and spherical coordinates (r, θ, ϕ) of a point P in three-dimensional space.
The versor e_r makes an angle θ with the Cartesian versor e_z.
Adapted from Auletta and Wang (2014, p.217)

$$\psi_{lm}(r) = \langle r \mid l, m \rangle \quad (1.254)$$

—

data 18.08.2019
S T Q Q S S ☒D

78) a) Cercle de Brocard

fr.wikipedia.org/wiki/Cercle_de_Brocard

Obs.: Figures de Brocard, Points de Brocard

b) Théorème de Neuberg

fr.wikipedia.org/wiki/Théorème_de_Neuberg

Obs.: Point de Vecten, Théorème de van Aubel

c) Musselman's Theorem

en.wikipedia.org/wiki/Musselman's_theorem.

d) Courbes (Curvophilie)

web4.ensiie.fr/~gacogne/courbes.pdf

Faz referência ao livro: Courbes Mathématiques
J. Brette.

↳ Source: 20190818/ Courbes-Mathématiques,
numéros spécial... djvu

e) List of unsolved in mathematics

www.wikiwand.com/en/List_of_unsolved_problems_in_mathematics

Obs.: In the Bloch Sphere

wikiwand.com/en/Bloch_sphere

data 18.08.2019

... In the Bloch Sphere representation of a qubit, a SIC-POVM (symmetric, informationally complete, positive operator-valued measure) - is a special case of a generalized measurement on a Hilbert space, used in the field of quantum mechanics.

form a regular tetrahedron.

Zauner conjecture that analogous structures exist in complex Hilbert spaces of all finite dimensions.

f) Spherical Bernstein's problem

Look: Bernstein's problem

g) Pi

wikiwand.com/en/Pi

Obs: Look

Fourier transform and Heisenberg uncertainty principle

→ Heisenberg uncertainty principle also contains the number π.

Veja: wikiwand.com/en/Uncertainty_principle

data 19.08.2019

79) Musich Transforms

① Descrição dos passos:

Point A:
1) Intersection of xAxis and yAxis

$A = (0,0)$

2) Point B: Point on xAxis

$B = (2,0)$

3) Circle c: Circle through B with center A.

$c: x^2 + y^2 = 4$

4) Point C: Point on xAxis

$C = (1,0)$

5) Point D: $D = (1,1)$

6) Segment f: Segment A, C

$f = 1$

7) Segment g: Segment C, D

$g = 1$

8) Segment h: Segment A, D

$h = 1.414213562373095 \quad (\sqrt{2})$

data 19.08.2019

9) Circle d: Circle through A with center C

$d: (x-1)^2 + y^2 = 1$

10) Point E

$E = (8, -8)$

11) Circle e: Circle through A with center E

$e : (x-8)^2 + (y+8)^2 = 128$

Obs:

Fazer estas construções geométricas por meio de equações matemáticas. (Evite usar simuladores ou geradores gráficos pois eles apresentam uma imprecisão inerente ao problema da Quadratura 2D.)

O Circle e foi obtido pela execução de vários triângulos com catetos iguais (mesmo valor) e cuja hipotenusa foi projetada num ângulo de 90° (noventa graus) da ponta do triângulo criado seguindo-se os quadrantes 1, 2, 3 e 4 até o 8 (oitavo) círculo, cujos catetos são 8 unidade de medida.

Vejamos a seguir:

data 19.08.2019

[✓ S] T Q Q S S D

No exemplo em questão:

O triângulo ΔACD, tem as seguintes medidas:

$\overline{AC} = 1$, $\overline{CD} = 1$ e $\overline{AD} = \sqrt{2}$

Deste triângulo básico, projetando-se a $\sqrt{2}$ no ponto D com ângulo de 90° (noventa graus), teremos assim 2 (dois) catetos com o mesmo tamanho cujo valor é $\sqrt{2}$, gerando assim, uma hipotenusa de valor 2.

Desta forma teremos o seguinte:

a)	1	1	$\sqrt{2}$
b)	$\sqrt{2}$	$\sqrt{2}$	2
c)	2	2	$2\sqrt{2}$
d)	$2\sqrt{2}$	$2\sqrt{2}$	4
e)	4	4	$4\sqrt{2}$
f)	$4\sqrt{2}$	$4\sqrt{2}$	8
g)	$8\sqrt{2}$	$8\sqrt{2}$	16
h)	16	16	$16\sqrt{2}$

Finalmente, usando-se o ponto E

→

data 19 08 2019

$E=(8,-8)$

Sobre o ponto E trace um círculo cujo raio passe sobre o ponto $A=(0,0)$

→ Circle through A with center E

equation:

$$(x-8)^2+(y+8)^2=128$$

O gráfico desenhado ficará assim:

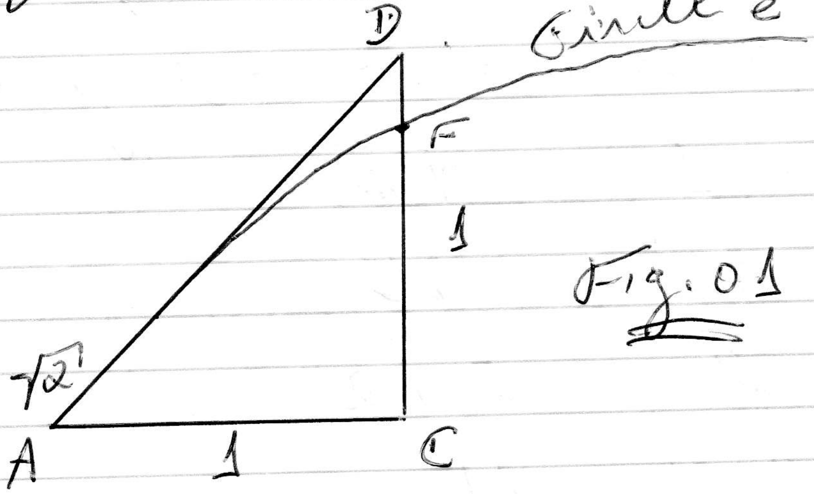

Fig. 01

Obs.:

$F \cong (1,\ 0.888194417315591)$

Comando:

F=Intersect(e, g, 1)

Important:

A diferença entre o ponto D e F

data 19.08.2019

[✓ S] T Q Q S S D

decorre do problema da quadratura 2D
e do valor infinitesimal do π (Pi).
Valor este irracional e
transcendental
(o pi não é uma solução de uma
equação polinomial com coeficientes
racionais). Portanto, o pi transcende
a classe dos números algébricos.

The Transcendence of π
Steve Mayer
November 2006

Source: 20190819_03/Pi-Transcendental/
pitrans.pdf

obs.: Earlier Liouville (1844) had
actually constructed transcendentals,
for example,

$$\sum_{n=1}^{\infty} 10^{-n!}$$ is transcendental.

Source: http://rutherglen.ics.mq.edu.au/
math334s106/md2334.Dioph.
Liouville.pdf (não disponível)

Desigualdade interessante:
e^{pi} ge pi^{e} _

data 19.08.2019

[✓ S] T Q Q S S D

Source: https://www.ime.unicamp.br/~samuel/irracionais-e-transcendentais-n-e/

Interessante:

Space-time fractional reaction-diffusion equations associated with a generalized Riemann-Liouville fractional derivative
R. K. Saxena

2014-09-05

Source: http://arxiv.org/~~[illegible]~~
http://arxiv.org/abs/1409.1817
archive.org/details/arxiv-1409.1817

Source: 20190819_03/Pi-Transcendental/1409.1817.pdf

—

On intermediate level sets of two-dimensional discrete Gaussian Free Field

Source: http://arxiv.org/abs/1612.01424

data 19.08.2019

www.archive.org/details/arxiv-1612.01424

Source: 20190819_03/Pi-transcendental/1612.01424.pdf

Return:

Na Fig. 01 observamos a discrepância decorrente do π.

Fazendo uma simulação na qual iremos fazer uma distorção do space-time (2D) usaremos um valor para converter 2D que se enquadre na quadratura do círculo.

Este valor é MUS. (ϖ)

π ϖ

Fazer cálculo com equações:

Usar o livro:

20190819_03/Complex Analysis and Applications by Hemant Kumar Pathak/9811397333.pdf

Pág. 39(60) →

Fig. 1.17 Straight line joining two points

z_1 and z_2.

data 19.08.2019

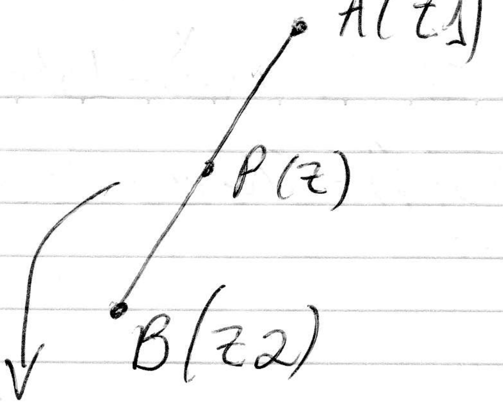

Veja também:

Fig. 1.16

1.8 Equations of Straight Lines

Theorem 1.6

Proof Let z_1, z_2 be two given points A, B on the Argand plane.

Let z be any point P on the line AB as shown in Fig. 1.17.

Then, we have

$$\arg\left(\frac{z-z_1}{z_1-z_2}\right) = 0 \text{ or } \pi$$

80) Cont. (2019 0820)

Use esta pesquisa na simulação.

a) (Hyperbolic lattice appears in a coplanar waveguide array Belle) Dumé Author: Alicia Kollár. 11 Jul 2019. Nature 10.1038/s41586-019-1348-3

20.08.2019

Source: https://physicsworld.com/a/hyperbolic-lattice-appears-in-a-coplanar-waveguide-array/

Research is detailed Nature 10.1038/s41586-019-1348-3

Source: 20180820/Nature_Magazine/ ~~s41586-019-1348-3.pdf~~ .png

Obs.: The heptagon-Kagome lattice drawn in a particularly nice projection know as the Poincaré disc model.

Import.: "Curved-space versions of this lattice occur if we use pentagon and five-pointed stars or heptagons and seven-pointed stars instead of hexagons and six-pointed stars"!

Important:

5) Hyperbolic lattices in circuit quantum electrodynamics.
Hollar

Source: 1) nature.com/articles/s41586-019-1348-3/

2) 20190820/Nature_Magazine/ s41586-019-1348-3.pdf/

Voltando aos desenhos do item 79, deveremos fazer um aprofundamento da simulação para o referido enquadramento do espaço euclidiano à sua nova realidade.

data 20.08.2019

Devemos estar atentos a seguinte área:

Fig. I

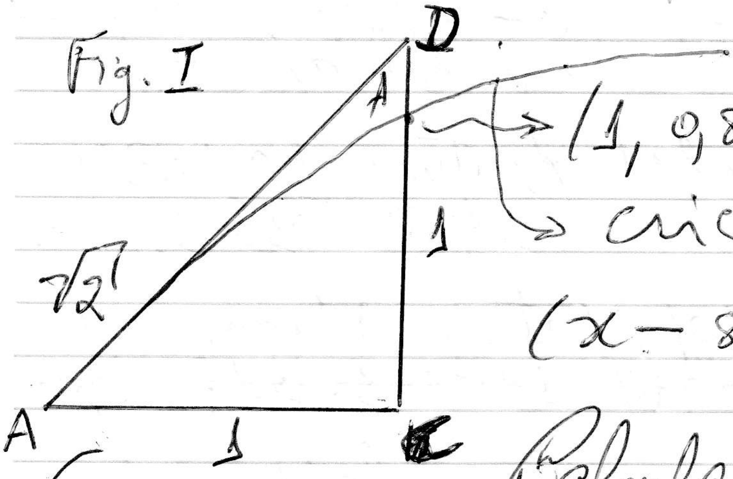

$\rightarrow (1,\ 0{,}8881944173155 91)$

$\rightarrow$ circle

$$(x-8)^2+(y+8)^2=128$$

Calcular a área $\underline{A}$

Compare a discrepância dos cálculos em relação ao cálculo

$$(x+2)^2+(y-2)^2=8$$

Points A e F(-2,2)

Circle K: Circle through A with center F.

Continuando a análise:

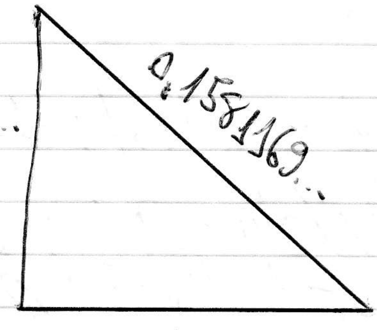
0.111855...
0.1581169...
0.1118055...

giro de 90° (à direita) horário

Fig. II

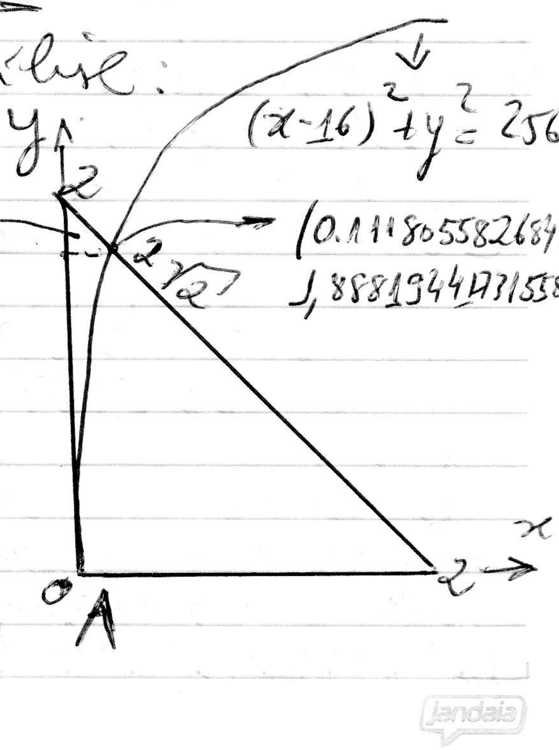

$$(x-16)^2+y^2=256$$

$(0.1118055826844 11,\ 1{,}8881944173155 85)$

20.08.2019

Posicionando as figuras I e II no triângulo maior:

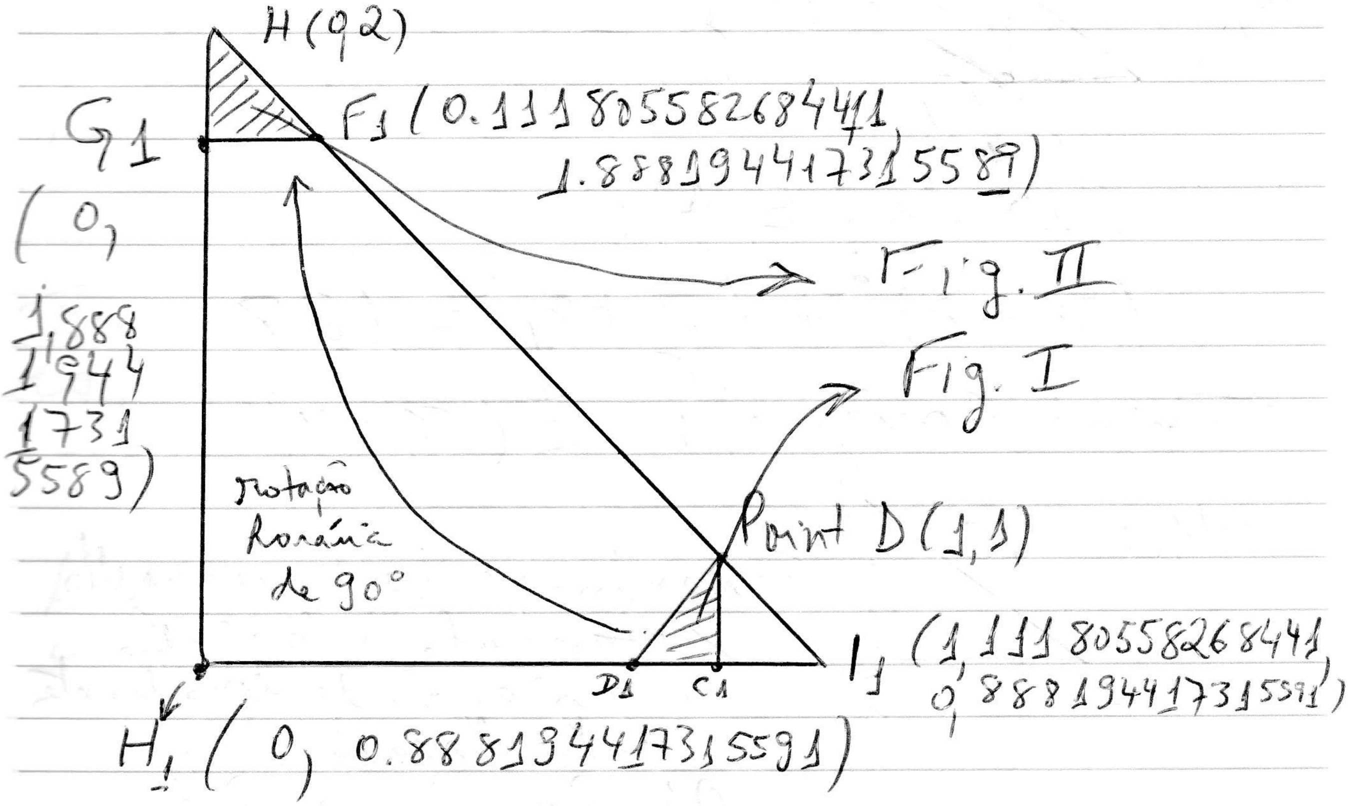
H (0,2)
G 1 (0, 1,888 1944 1731 5589)
F1 (0.11180558268441, 1.888194417315589)
Fig. II
Fig. I
rotação horária de 90°
Point D (1,1)
D1
C1
I1 (1,11180558268441, 0,888194417315591)
H1 (0, 0.888194417315591)

C_1 (1, 0.888194417315591)

D_1 (0.888194417315591, 0.888194417315591)

Distâncias:

$\overline{H_1 I_1} = 1.11180558268441$

$\overline{H H_1} = 1.11180558268441$

$\overline{H I_1} = 1.572330533754414$

Important: Os pontos C_1 e F_1 estão intimamente relacionados com os círculos

20.08.2019

que eu vou explicar em seguida:

a) 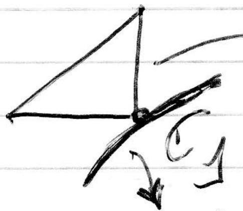

→ Fig. I

Circle(E, A):

$$(x-8)^2+(y+8)^2=128$$

Point (origem): E(8, -8)

Obs.: 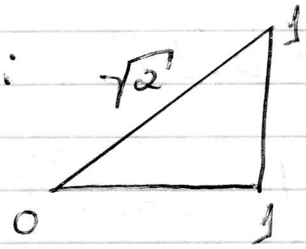

Partindo-se deste triângulo retângulo e fazendo a rotação de quadrante (anti-horário). Q1, Q2, Q3 e Q4, usando perpendiculares com a mesma medidas da hipotenusa por 7 vezes, obteremos este círculo C1.

(sete) 7 vezes excluindo o padrão 1x1

guarde este valor.

Obs.: O Ponto C1 é obtido quando o círculo (E, A) $(x-8)^2+(y+8)^2=128$ faz a intersecção com a perpendicular que passa centro do círculo que tem como raio 1 unidade (diâmetro 2).

Para a obtenção dos triângulos, nós deveremos fazer o seguinte:

20.08.2019

No Ponto C_1 trace uma perpendicular que passe por ele e que intercepta a hipotenusa ($\sqrt{2}$), cujos catetos são igual a 1. A intersecção na hipotenusa criaremos o ponto D_1. A distância entre $D_1 C_1$ é $\cong$

0,111805582684409...

Pegando-se este valor como base use-o como divisor do Pi (π) teremos:

$\pi \div 0,11180.... = \underbrace{28},09871008....$

Lembra-se do número 7 que foi armazenado anteriormente, faça a divisão usando-o como divisor.

$28 \div 7 = \underline{\underline{4}}$

Próximo círculo é o $\underline{\underline{8}}$ (oitavo).

$28 \div 8 = 3,5$

$(x-16)^2 + y^2 = 256$

__Obs__: Este círculo se aproxima do eixo $\underline{\underline{Y}}$.

Point for creation E_1 (16,0) e faça círculo com o point A(0,0).

__Important__: O círculo 7 intercepta o centro do círculo 8 no ponto (16,0).

data 20.08.2019

S [✓ T] Q Q S S D

b) Fig. II

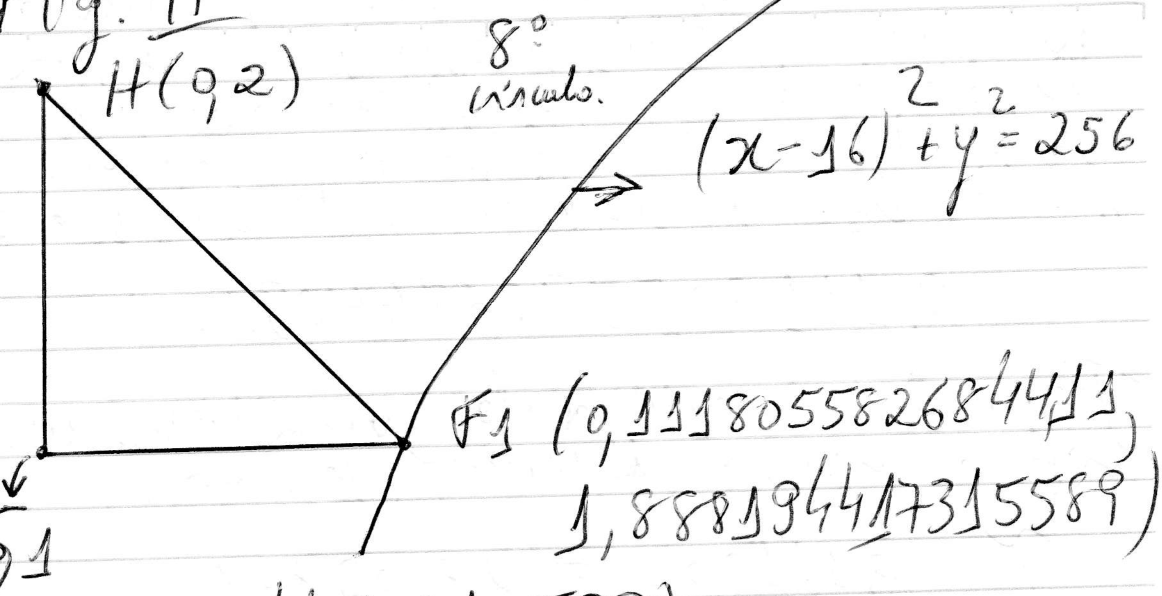
H(0,2)
8° círculo.
$(x-16)^2+y^2=256$
F1 (0,111805582684411, 1,888194417315589)
G1

(0, 1.888194417315589).

Complementando-se a análise:

Contando-se $^{\sqrt{2}}\underset{1}{\triangle}1$ são 9 círculos.

$28 \div 9 = 3,\bar{1}$.

(muito próximo do π).

Obs.: Poderemos usar este valor quando da simulação. (distorção space-time)

Utilizar o livro:

The Nonlinear Schrödinger Equation Singular Solutions and Optical Collapse

Source: 20190807_02/The Nonlinear Schrödinger Equation Singular.../The Nonlinear... pdf

20.08.2019

Source: arquivo.ggb

ARQ/Geogebra_org_2017_versão_6_0_411_02/
20190820_Padrão_007_003.ggb

Obs: Contém Fig. I e II.

Original file com todos os passos:

20190818_011_V003_004_padrão_002.ggb

—

File: 20190820_Padrão_006_002.ggb

—

Outro file:

20190818_011_V003_004_padrão_002.ggb

Distância $\overline{VU}$:

$V = (0.732050807568878, 1.267949192431122)$

$U = (1.06274066806227, 0.93725393193773)$

Distância: 0.467673720625947

Obs: Criei outro arquivo:
20190818-001_V003_padrão_002_distancia_VU_001.ggb.

20 .08 .2019

$\sqrt{U} \times 3 = 1.40301961877841^{x}$

$< \sqrt{2}$

81) Cont. (20190821).

File: 20190818_009.ggb
(compare neste)
mei 20190818_009_padrão.ggb

1) Criar um círculo de raio 4 unidades (diâmetro 8 unidades) cujo centro é A, faça outro círculo cujo centro é a metade do raio do primeiro círculo.
Todos os círculos estarão um dentro do outro e todos são tangentes ao mesmo ponto D que está fixado do outro lado da circunferência alinhado com o centro da primeira circunferência.

2) Ao todo utilizaremos 5 círculos

Descrição:

a) Point A(0,0)
b) Point B(4, 0)
c) Point C(-2, 0)
d) Point E(-3, 0)
e) Point F(-3,5, 0)

Point G(-3.75, 0)

obs: O círculo com centro no ponto E(-3,0) e raio de 1 unidade no ponto C(-2,0) tem uma área = Pi (π).

Fazendo um estudo sobre perpendiculares que

21.08.2019

passam pelo centro destes pontos descritos anteriormente e que interceptam as suas circunferências (perímetros), criam-se pontos de intersecção e todos estes deverão ser nomeados.

Interligando-se um por um, teremos as seguintes conclusões.

1ª circunferência: Point H(0, −4);

2ª circunferência: Point I(−2, −2);

3ª circunferência: Point J(−3, −1);

4ª circunferência: Point K(−3.5, −0.5);

5ª circunferência: Point L(−3.75, −0.25) e Point D(−4, 0).

Distância:

$\overline{HI} = 2.8284271\ldots = 2\sqrt{2}$

(1/2)

$\overline{IJ} = 1.4142135\ldots = \sqrt{2}$

(1/2)

$\overline{JK} = 0.7071067\ldots = \frac{\sqrt{2}}{2}$

(1/2)

$\overline{KL} = 0.3535533\ldots = \frac{\sqrt{2}}{4}$

$\overline{LD} = 0.3535533\ldots = \frac{\sqrt{2}}{4}$

> =

21.08.2019

Analisaremos os segmentos de reta $\overline{KL}$ e $\overline{LD}$ e suas relações com o círculo e, cuja área é igual a π (Pi).

Salvo como: ARQ/Geogebra_org_2017_versão_6_0_411_02/
20190818_009_padrao_003.ggb.

Observamos o seguinte:

O segmento $\overline{JD}$ é a hipotenusa do triângulo retângulo cujo vértice oposto encontra-se no eixo x, denominado Point E(-3,0).

Desta forma, $\overline{JD} \cong 1.4142... = \sqrt{2}$.

Vejamos o segmento $\overline{JK} = 0.707106...$
($\sqrt{2}/2$)

Obs.: O segmento de reta é a metade ao segmento anterior, mas a área do círculo menor posterior é 1/4 do círculo e

Portanto, o círculo f é 1/2 do círculo e.

Círculo e = π

Círculo f = $\frac{\pi}{4} = 0.7853981633...$

Voltando aos segmentos $\overline{KL}$ e $\overline{LD}$, observaremos o seguinte:

21.08.2019

a) Segmento $\overline{KL} = 0.3535533\ldots \left(\frac{\sqrt{2}}{4}\right)$

Este segmento é a metade do segmento $\overline{JK}$ $(\sqrt{2}/2)$.

b) Segmento $\overline{LD} = 0,3535533\ldots \left(\frac{\sqrt{2}}{4}\right)$.

Somando-se os segmentos

$$\overline{KL} + \overline{LD} = \frac{\sqrt{2}}{2}$$

O círculo menor corta a hipotenusa do círculo maior pela metade, mas a área do primeiro é um quarto do círculo maior.

82) **File:** 20190818_011_V002_001.ggb

Estudar este gráfico:

1) Faça um círculo de raio 4 com origem no Ponto A(0,0) e B(4,0). $x^2+y^2=16$

2) Crie os pontos C(−4,0) e D(0,4).

3) Crie os pontos E(−6,0) e F(0,6)

4) Faço o segmento $\overline{EF}$. (f)

5) Crie uma perpendicular sobre aquele segmento (f) que intercepte o ponto A(0,0). Desta forma, criaremos a Line g: Line through A perpendicular to f.

data 21.08.2019

D S T Q Q S S D

6) Em seguida marque o ponto onde esta perpendicular intercepte o perímetro do círculo c cujo centro é A(0,0) ⇒ Poin G

7) O Point G(-2.8284..., 2.8284...) ou seja,

$(2\sqrt{2}, 2\sqrt{2})$

Neste caso, a distância entre o ponto A(0,0) e G$(2\sqrt{2}, 2\sqrt{2})$ é igual 4.

8) Crie o segmento $\overline{AG}$ (segmento h);

9) Nesta fase de montagem do gráfico, nós temos 2 opções:

a) Criar uma tangente no ponto G em relação ao círculo c.

Line 1: Tangente to c through G.

ou

b) Criar uma perpendicular:

Line: Line through G perpendicular to g.

Obs.: Prefiro esta pois estou interconectando os vários procedimentos de criação entre si.

$6x - 6y = -33.941125496...$

10) Marque os pontos (intersecção) aos eixos Ox e y!

Jandaia

21.08.2019

Point H (-5.656854..., 0) e

Point I (0 , 5.656854...)

11) Crie o segmento HI

Segment j (8 unidades)

Obs.: O segmento AG tem 4 unidades de medida.
O segmento HG tem 4 unidades de medida.
Ambos são catetos,
A hipotenusa AH tem
5.656854...

Import: Criei um arquivo com estes procedimentos:

20190818_011_V003_004_padrão_003.ggb

Criei um novo para continuar o estudo com data de hoje.

20190821_padrão_perpendicular_001.ggb

Prosseguindo!

12) Crie um círculo Circle:d

Circle d: Circle with center G and radius Segment(A, G).

Obs: Ao fazer isso, você estará projetando o segmento AG para a perpendicular

21.08.2019

Line i: Line through G perpendicular to g.

ou seja, $\overline{AG}$ é igual $\overline{GH}$.

O Ponto H está sobre o eixo x pelo qual passa o circle d, a line i, o segmento H, I ($\overline{HI}$).

Podemos adotar uma maior precisão, utilizando uma outra construção geométrica:

Circle through 3 Points

Circle d: Circle through A, G, H.

Desta forma, teremos uma relação mais intrínseca entre estes pontos.

13) Ligue os pontos ~~$\overline{AD}$~~ → $\overline{CD}$.

Criaremos o segmento $\overline{CD}$

Segment k: Segment C, D.

Tamanho: 5.65685424...

14) Com os seguimentos: $\overline{CD}$, $\overline{HI}$ e $\overline{EF}$ vamos criar 3 (três) trapézios que veremos a seguir.

a) Trapézio (Polygon) C, D, I, H.

data 21.08.2019

Este polígono tem 8 unidades de área.

A perpendicular sobre o ponto G $(-2\sqrt{2}, 2\sqrt{2})$ tem o comprimento de 8 (oito) unidades, ou seja, corresponde a aresta maior (base do trapézio invertido) cuja a área também é 8 unidades.

Os pontos G e D geram um segmento que corresponde a hipotenusa decorrente dos catetos iguais de tamanho 4 unidades, que é o raio do círculo original inicial. $(\overline{AB})$

b) Trapézio (H, I, F, E)

Área: ~~0.34314575050~~... 2 unidades

As arestas deste trapézio (arestas menores, laterais) tem o comprimento de

$$0,34314575050\ldots + \overline{CH} \text{ ou } \overline{DF} = 2$$

Obs: Enquanto que no trapézio A a tangente sobre o ponto G, que é o mesmo de uma perpendicular sobre o ponto G, fornece o resultado da área deste primeiro trapézio, que é 8 unidades.

Neste, podemos observar uma mudança da localização do cálculo de área, este passa a ser as arestas menores, enquanto que o primeiro é a base do trapézio A (a).

somando-se as arestas menores do trapézio a e b.

data 21.08.2019

c) Trapézio (Polygon) (C, D, F, E) x

Área: 10 unidades.

Os segmentos (arestas menores do trapézio) $\overline{CE}$ e $\overline{DF}$ têm o comprimento de 2(duas) unidades.

Atualização dos dados salvo em outro arquivo:

20190821_padrao_perpendicular_002.ggb

Ampliando o estudo:

15) Criando um polygon, ou seja, um triângulo retângulo com os pontos A, C, D obteremos a área de 8(oito) unidades.

2(dois) catetos com comprimento de 4(quatro) unidades.

A soma dos dois catetos corresponde à área do triângulo retângulo.

Important: O valor da área ou melhor, o número obtido é igual ao trapézio a (C, D, I, H).
As duas(2) figuras geométricas tem a mesma área.

Observemos que o segmento $\overline{AC}$ (cateto do triângulo retângulo) é 4, 21.08 2019

é o dobro de medida da aresta menor do trapézio a(C, D, I, H).

Extrapolação:

Somando-se os trapézios a(C, D, I, H) e b(H, I, F, E) = trapézio c(C, D, F, E)

a = 8 , b = 2 ⇒ c = 10

Fazendo nova soma com o triângulo retângulo ACD (8 unidades) teremos 18 unidades de área

Multiplique-se por 4 (são 4 quadrantes trigonométricos) teremos

18 x 4 = 72 unidades de área

A Área do círculo c(A, B):

(numéricos)

$x^2+y^2=16$ (unidades de numéricos e não de área).

Fazendo a divisão:

72 ÷ 16 = 4,5 vezes.

Portanto o triângulo A, E, F (ΔAEF) tem a área de 18 unidades, mas devemos tirar o trapézio b (H, I, F, E) (2 unidades), assim, nós teremos 16 unidades de área que é igual a área do círculo c ($x^2+y^2=16$)

numérico.

21.08.2019

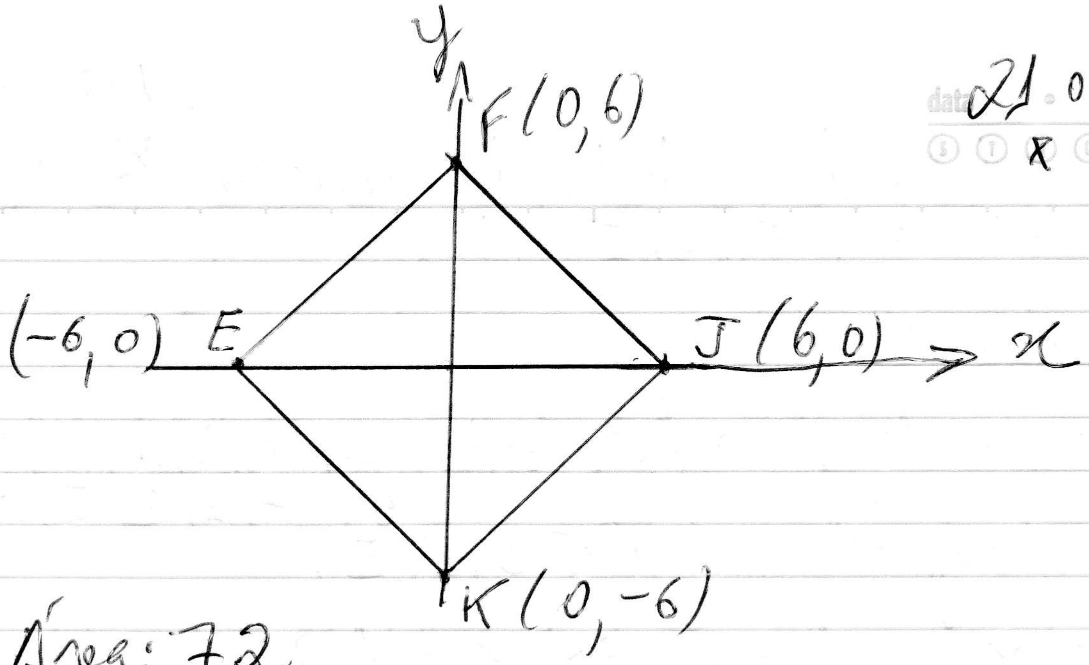

Área: 72

Os 4 quadrantes do círculo c tem 64 unidades.

O acerto ⇒ 8 unidades = 1/2 do circle c ($x^2+y^2=16$) número.

4,5 vezes

Source:
20190821_padrão_perpendicular_003.ggb

A equação $x^2+y^2=16$ tem uma área de
≅ 50.26548245...

Aprofundando:

64 unidades de área (4 quadrantes) menos o círculo c (50.26548245743669...) é igual à
13.73451754256331

Dividindo-se este número por π (Pi)

21.08.2019

teremos 4,371832715762603556277619...

Subtraia este valor de 4,5 vezes (= ao quadrado □ EFJK).

↳ 72 unidades.

Teremos:

$$0{,}1\underbrace{28167284237396 44}\ldots$$

Se o valor for 5 - 4,371... =

$$0{,}6\underbrace{28167284237396 44}\ldots$$

Ok: Bateu o excesso de 0.5 = 8 unidades de área

72 + 8 = 80 unidades de área.

Important: O valor 0 6281672...é aproximadamente igual

$$\frac{2\pi}{10} \text{ ou } \frac{\pi}{5}$$

↳ 0.62831853071...

Vejamos:

0,5 = 8 unidades em excesso.

De 72 para 80 = $+11.\bar{1}\,\%$

De 80 para 72 = - 10 %

Estudo de área:

data 21.08.2019

S T [✓ Q] Q S S D

Criei o arquivo:

20190821_padrao_area_001.ggb

Seguindo-se ao modelo do arquivo anterior, usando-se por base o círculo c(A, B) de raio 4 ($x^2 + y^2 = 16$) e fazendo-se uma perpendicular no ponto G até esta interceptar o eixo x, teremos a mesma reta (dimensão) de AG.

Em seguida, fazemos o uso do comando

Circle through 3 Points

e ligaremos os pontos: A G H, obtemos o círculo d

$$(x + 2.8284271)^2 + y^2 = \underline{\underline{8}}$$

Área: 25.13274

É metade da área do círculo c (= 50.26548).

Adotando-se estes mesmos procedimentos, nós iremos fazer o 3º círculo:
(seguindo estes passos mais detalhadamente)

1) Observe os pontos onde os 2 (dois) primeiros círculos se interceptam e neste pontos marcaremos os pontos G e L.

21.08.2019

2) Liguemos estes pontos de modo a criar um segmento de reta $\overline{LG}$

Point L $(-2\sqrt{2}, -2\sqrt{2})$

Point G $(-2\sqrt{2}, 2\sqrt{2})$

$\overline{LG} = 5.6568542494\ldots$

3) A reta $\overline{LG}$ (segmento) intercepta o eixo $\underline{x}$ no ponto M

$M(-2\sqrt{2}, 0)$ ↑ criar este ponto.

e neste ponto devemos criar o ponto M com o comando Intersect
(Intersecção)

4) Angle bisector entre $\overline{LG}$ e o eixo x.

5) Què o ponto em que este bisector intercepta o círculo d, criando-se o ponto N $(-\underbrace{4,8284}_{(2+2\sqrt{2})}, 2)$

6) (Circle through 3 Points:
Utilizando-se os pontos: M, N, O)
↳ item $\underline{\underline{7}}$.

No Ponto N, use o comando
Line through N perpendicular to $\underline{x}$
(bisetor)

data 21.08.2019

que criará uma perpendicular que interceptará o eixo x no ponto O

$$(-\underbrace{6{,}8284}_{(4+2\sqrt{2})}\ldots, 0)$$

7) Circle through 3 Points:

Utilizando-se os pontos: M, N, O, criaremos mais um novo círculo. circle r.

A área deste círculo é a metade do círculo anterior (circle d)

circle (M, N, O) →

$$(x + 4{,}8284)^2 + y^2 = 4$$

Área: 12.566370614359183

÷ 4 (da equação)

= 3,1415926...

Continuando no mesmo gráfico, vamos nos ater ao círculo c (A, B)

1) Dividindo-se o 2º Quadrante, usando Bisector entre os eixos x e y.

2) Na intersecção deste bisector com o círculo c, criaremos o ponto P (-2, 2)

21.08.2019

Este ponto P ficará entre a intersecção entre o segmento CD e o bisector do ponto A (0,0).

Point C (−4,0), Point D (0,4)

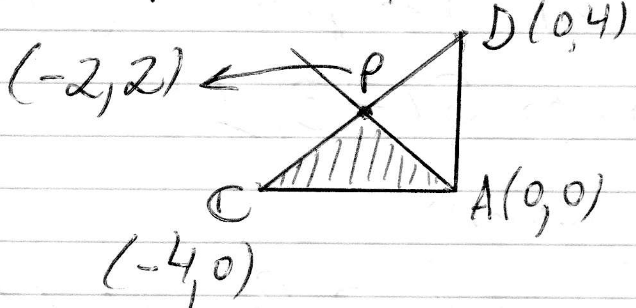

3) Com este triângulo ACD (triângulo retângulo), usemos o comando

Circle through 3 Points.

Cria-se o círculo: $(x+2)^2+y^2=4$

Obs.: Importante. A área deste círculo é igual ao círculo circle r (circle through M, N, O), mas se você continuar a fazer estes comandos dentro do circle c.

4) Criaremos um novo círculo.

a) O Ponto P(−2,2) já existe.

b) O bisector do Ponto A(0,0) intercepta o circle p, (A, P, C) no ponto Q(−2,2).

21.08.2019

c) Ligue os pontos $\overline{PQ}$ (4 unidades)

d) Use o bisector entre o segmento $\overline{PQ}$ e o eixo x, que interceptará o segmento $\overline{CP}$ no ponto S (−3, 1)

e) Crie o ponto S(−3, 1)

f) Com o comando:

Circle through 3 Points (R, S, C)

Criaremos o circle g_1 (R, S, C)

$(x+3)^2+y^2=1$

↳ Área = π (Pi).

Portanto: Este círculo tem o tamanho de 1/4 do círculo anterior.

(0,25).

Important: Da outra forma, a redução é de 50% (1/2)! (fazendo-se do círculo e $x^2+y^2=16$).

Obs: Criei um novo arquivo com estes dados:

1) 20190821_padrão_perpendicular_004.ggb

2) 20190821 padrão_area_002.ggb

data 22.08.2019

S T Q [✓ Q] S S D

83) 201908221440

Utilizando o arquivo 20190820_Padrão_001.ggb (original: 20190818_011_v003_004_padrão.ggb).

Acrescente o círculo: (já está no gráfico).

$(x-1)^2+(y-1)^2=2$ Point(1,1)

Neste arquivo, o Point D(1,1).

a)

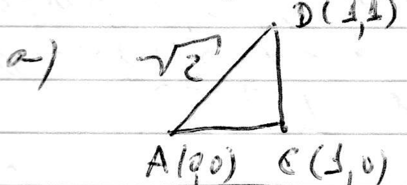

Triângulo retângulo.

b)

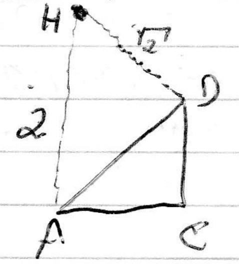

Faça a projeção da hipotenusa $\overline{AD}$ sobre a perpendicular incidente sobre o segmento de reta $\overline{AD}$ sobre o ponto D, criando-se o Point H(0,2)

c)

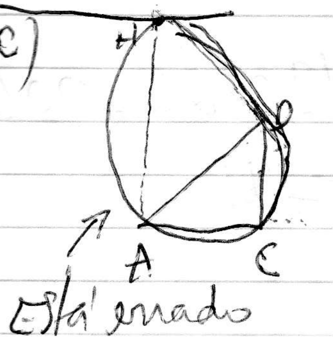

Círculo perfeito! KKK.

Este círculo

$(x-1)^2+(y-1)^2=2$

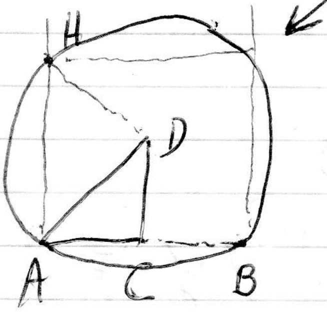

Ficou melhor!

Use o Circle through 3 points.
(não previa)
Este círculo já está no gráfico.

Faça todos estes procedimentos para o próximo círculo (catetos 2 unidades, hipotenusa $2\sqrt{2}$).

data 22 08 2019
S T Q [✓ Q] S S D

Circle 1: $x^2+(y-2)^2=4$

Faça novamente, com os catetos $2\sqrt{2}$ e hipotenusa 4 (Circle t).

Point$(-2,2)$. = Point F.

Crie o segmento: $\overline{HF}$

Obs: O ângulo entre o cateto e a perpendicular na qual será projetada a hipotenusa do triângulo anterior é de 135°.

Todos estes círculos estão neste gráfico. Esta explanação anterior foi feita para mostrar como eu cheguei aqui, neste modelo.

Resumindo os passos:

Salvo no arquivo: 20190822_001.ggb

1) Circle(A,B) → $x^2+y^2=4$

2) Point C$(1,0)$

3) Point D$(1,1)$

4) Segment $\overline{AC}$, $\overline{CD}$, $\overline{AD}$

5) Circle (C,A) → $(x-1)^2+y^2=1$

↳ Área = π (Pi).

22 08 2019

6) Point E(8,-8)

7) Circle(E, A) → $(x-8)^2+(y+8)^2=128$

↳ Área = 402.12385...

8) Point F(-2,2)

9) Circle(F, A) → $(x+2)^2+(y-2)^2=8$

↳ Área = 25.13274122...

10) Triângulo equilátero ΔABG

Use o comando: Regular Polygon

Marque os pontos A e B e em seguida quantas arestas terá, no caso 3. Neste momento criará o ponto que será nomeado como G (1, 1.7320508075...)
(Atentar para que este ponto intersecione o circle c (A, B).
Obs. Por causa da quadratura acho que não haverá uma interseção perfeita.

11) Point H(0,2) → Intersecção do circle c (AB) e o eixo y.

12) Crie um círculo utilizando os pontos (D, A, H) → $x^2+(y-1)^2=1$

13) Ligue os pontos $\overline{DH}$
Segmento(D, H) → $\sqrt{2}$

data 22.08.2019

14) Segment $\overline{HF} \Rightarrow 2$

15) Point I (-8, 8)

16) Circle(I, A) $\Rightarrow (x+8)^2+(y-8)^2=128$

Obs: Este círculo será usado como parâmetro para estudar as discrepâncias dos cálculos das retas que serão analisadas posteriormente. (simetria entre os pontos I(-8,8) e E(8,-8).)

17) Point J (0.61538..., 0.92307...)

Intersecção entre os círculos AC e círculo AF.

18) Point K (1.1327..., 0.911...)

Intersecção dos círculos AC e EA.

19) Point L (0.8228..., 1.8228...)

Intersecção dos círculos AB e FA.

20) Point M (1.5336..., 1.2836...)

Intersecção dos círculos AB e EA.

21) Crie um novo círculo (D, H)

$(x-1)^2+(y-1)^2=2$

data 22.08.2019

22) Point N (0.8, 2.4)

Intersecção entre os círculos DH e FA.

23) Point O (2.215..., 1.723...)

Intersecção entre os círculos DH e EA.

24) Segment (L, M) → 0.8921...

25) Segment (N, O) → 1.5689...

Important: Até este ponto, observamos a criação de 2 linhas que não são paralelas em decorrência do π (Pi) e da quadratura do círculo. Fazer uma abordagem mais profunda neste aspecto específico.

Fazendo a continuação do nosso estudo.

26) Crie uma linha que ligue os pontos DH.

$\Rightarrow x + y = 2$

27) Line $\overline{LM}$ $(0.5391.x + 0.71080.y = 1.7393...$

28) Line $\overline{NO}$ $(0.6769.x + 1.415 y = 3.9384...$

As linhas (H,D), (L,M) e (N,O) não se entrelaçam de formas distintas no 2º (segundo) quadrante, formando um triângulo obtusângulo.

data 22.08.2019

29) Este triângulo obtuso tem os seguintes pontos:

$\triangle PQR$

$P(-1.5, 3.5)$

$Q(-1.\boxed{8518}..., 3.\boxed{8518}...)$ > estudar!!!

$R(-1.19..., 3.35...)$

obs.: Criado o arquivo

20190822_002_padrão.ggb

84) 20190823

Ideia: 20190820.

$1^1=1$

$1^2=1(?)=2$ por quê?

$2^2=2\times2=4$

Vejamos: $1^2 = \begin{cases} 1\times1=1 \\ 1+1=2 \end{cases}$ → para metro quadrado (área) funciona, mas para números inteiros, não.

recalcular utilizando esta abordagem

1^2 tem que ser igual (forma) à 2^2. Assim. $2^1=2$ e $2^2=4$. Portanto: $1^1=1$ e $1^2=2$.

(Questão de lógica básica)

23.08.2019

85) 20190823

Solução para a quadratura do círculo.

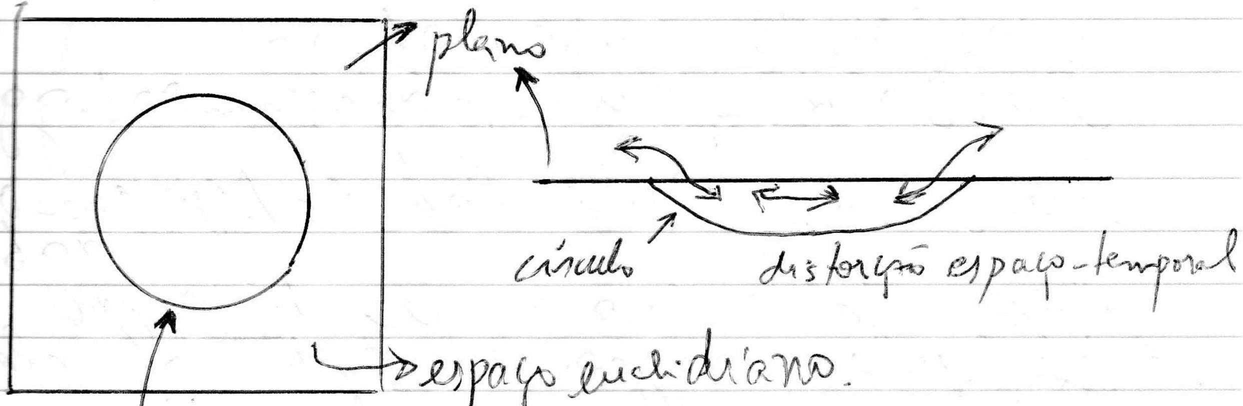
plano
círculo
distorção espaço-temporal
espaço euclidiano.

valor $\frac{1}{\pi}$ (distorcido) < π e igual quadratura. (ajustamento)

(fator conversão de universos diferentes mas coexistentes.

Obs: Como Teoria Clássica e a Teoria Quântica.

π é o fator de convergência entre este dois universos com as suas complexidades que são inerentes à cada sistema e agora serão intercambiáveis.

3D

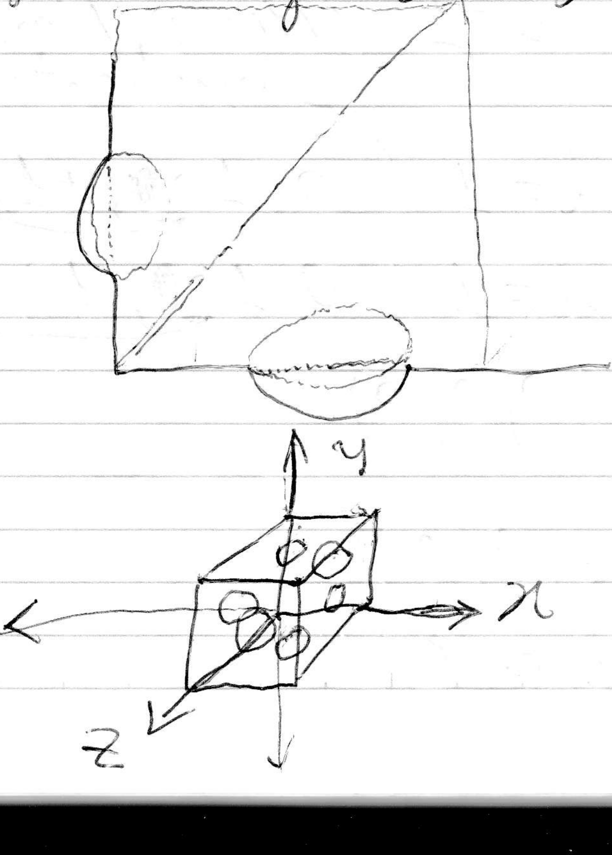

data 24.08.2019
S T Q Q S X D

85) Cont. 84. (20190824)

Para simular este cubo sem a distorção criei os arquivos:

20190823_Cube_001.ggb
20190823_Cube_montagem_001.ggb
20190823_Cube_Padrão_Ajustado_001.ggb

Agora falta colocar a distorção e calcular os pontos entre os planos em 3D. (nova trigonometria!!!)

Vejamos!

(base)
20190823_Cube_Padrão_Ajustado_001.ggb
versão salva: 002.ggb

Passos:

1) Slider
a = 1 (Min: 0, Max: 10, Increment: 1) → é melhor simulação
b = 1 (0, 10, +1) → 10
c = 1 (0, 10, 1)

2) ~~Point~~ A = (0, 0, 0)
coloque direto ↑

B = (a, 0, 0)

C = (b, c, 0)

data 24.08.2019

D = (0, b, 0)

3) Polygon [A, B, C, D]
↳ parênteses

4) Entre nas opções do programa: ≡
(lado direito da tela)

a) Perspectives

b) 3D Graphics

c) Vá na opção 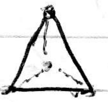

↳ Extrude to ~~Pyramid or Cone~~
Prism or Cylinder

(4.ª opção)

Após, clique no Polygon (A, B, C, D)
e dê altitude 1 → OK

Obs: Cria-se um Cubo.

d) Opção do Programa: ≡
Volte para a visualização
Geometry

5) Crie novo controle deslizante:
Slide e = 1 (0, 1, 0.01)
increment

data 24.08.2019

6) Commando do programa: ≡

↳ clique: △

↳ escolha a 9ª opção: NET

↳ clique no Cubo.

Obs: O controle deslizante cria-se automaticamente.

A velocidade de execução da demonstração está no canto superior direito do controle deslizante (~~Graphics~~)

Algebra ↳ local onde os comandos são inseridos

↳ Possibilita-se na opção VIEW → marque ☑ Algebra.

—

Obs:

Velocidade aumenta-se 0.5 em 0.5

<< 2x >> fica em cima do comando play

▶

(Para facilitar: esta nova versão do programa, permite que você use a externa e em segunda, você executar a NET (simulação)

↳ não funcionou.

—

data 24 08 2019
S T Q Q S ☒ D

86) Criei o arquivo:

20190823_Cube_Padrao_Ajustado_003.ggb

Este arquivo tem 2 funções:

- Animação Cube (3D)
- Animação Paralelepipido. (Altera-se a altura do cubo).

Vou criar outro:

20190824_Cube_Paralelepipido_001.ggb

Desenvolvendo

20190824_Cube_Paralelepipido_002.ggb
(ficou incompleto)

Melhorando o arquivo: Paralelepipido_004.ggb

Criando desde do Zero

Passos:

1) Slides a (0, 10, 1)
b (0, 10, 1)
c (0, 10, 1)

d (0, 10, 0.01)

data 24.08.2019

2) A=(0,0,0)
B=(a,0,0)
C=(b,c,0)
D=(0,b,0)

3) Polygon(A,B,C,D)

4) Altere para Perspectives → 3D Graphics
↳ Extrude to Prism or Cylinder
↳ Clique sobre Polygon(A,B,C,D)
↳ Valor: 1.

5) Clique NET e depois sobre o Cube.

Obs: Criado arquivo:
20190824_Cube_Paralelepipido_005.ggb
continuando os comandos:

6) Point1=(0.5, 0.5)
Point2=(-0.5, 0.5)
Point3=(1.5, 0.5)
Point4=(0.5, 1.5)
Point5=(0.5, -0.5)
Point6=(0.5, -1.5)

24.08.2019
X

6) Circle with Center and Radius

Clique nos pontos (um por vez) e o raio usado será 0.5.

Pronto, o padrão está montado.

—

Até aqui:

Paralelipipípido_006.ggb

Paralelepipido_007.ggb
↳ círculos de raio 0.5 são fixados nas arestas dos cubos.

Obs: Veja as simulações dos slides a, b e c.

São construções muito interessantes.

Vejamos:

6 = 7

O círculo central da cruz cresce e toca (tangencia) o círculo da base da cruz, sendo que o círculo intermediário é usado como geratriz do círculo central.
Obs: Nomear ~~logo~~ os círculos.

—

data 25.08.2019

S T Q Q S S [✓ D]

87) 20190825

File: 20190824_Simulação-Quadrante_001.ggb

Segment $\overline{PQ} = \sqrt{2}$

intercepta o segmento $\overline{LM}$ no ponto L!

O Ponto L também é ponto de interseção com os círculos:

a) Circle c: Circle through B with center A.

b) Circle k: Circle through A with center F.

obs:

circle c: $x^2 + y^2 = 4$

circle k: $(x+2)^2 + (y-2)^2 = 8$.

obs.: Estudar linhas amarelas tracejadas $\overline{QC}, \overline{QT}, \overline{TU}, \overline{UV}$.

Source:
File: Quadrante_002.ggb

Neste arquivo, estes segmentos estão na cor verde.

25.08.2014 X

Observação important:

Segmento $\overline{QC} = 2$

Segmento $\overline{QV} = 2.82842$ $(2\sqrt{2})$

Important:

Estudar também os segmentos

$\overline{QT}$, $\overline{TU}$ e $\overline{UV}$

O segmento $\overline{QT}$ está saindo do cateto do triângulo e não da hipotenusa, de onde deveria se projetar perpendicularmente para que se pudesse calcular o outro cateto para gerar uma nova hipotenusa.

No caso, a hipotenusa $\overline{QC}$ criará um novo cateto perpendicular a este segmento.

Estudar o método tradicional e esta nova alternativa.

obs: Cateto 2 e Cateto 2 = hipotenusa $2\sqrt{2}$.

File: Quadrante_003.ggb

Veja a quase intersecção

entre ($\overline{NO}$ e $\overline{QV}$) e o circle g (Circle through A with center I)

Point W

Point L.

obs: Estudar!!!

data 25.08.2019

S T Q Q S S [✓ D]

Circle g:

$(x^2+8)^2+(y-8)^2=128.$

—

Analisar os pontos C_1 e D_1, pois observamos o erro na quadratura.

Point $C_1(5.0497..., 0)$

Point $D_1(5.0227..., 0.02303...)$

—

Outras diferenças encontradas:

$\overline{B_1D_1}$ e $\overline{B_1C_1}$.

—

File: Quadrante_004.ggb

Intersecção entre os seguimentos:

$\overline{NO}$ e $\overline{QV}$ no ponto W

= Point $W(1,\overline{63}, 2)$

valor nunca achado antes por mim.

— Distâncias: $\begin{cases} \overline{DC}=1 \\ \overline{qC}=1.73205080... \\ \overline{QC}=2 \end{cases}$

25.08.2019 X

88) Faça as simulações:

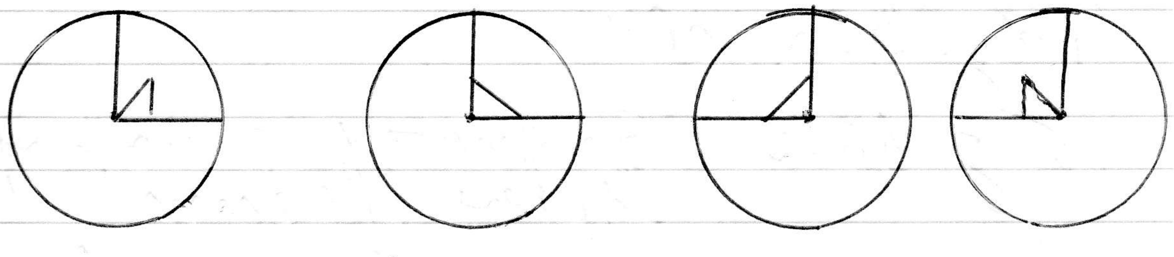

Compare as simulações e sobreponha uma na outra e encontre as intersecções!

89) Bibliografia:

a) A Panoramic View of Riemannian Geometry
Source: 20190825_02/Riemann/marcel-berger-a-panoramic-view-of-riemann-geometry-springer-2007.pdf.

Todo livro.

Ex: Pag 2(24)

Distance Geometry

Triangle in E^2

$$a^2 = b^2 + c^2 - 2bc\cos A$$

Fig. 1.1

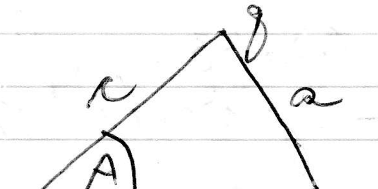

Obs: Usar para calcular distâncias.

872 pág.

data 25.08.2019

S T Q Q S S [✓ D]

b) mesmo book.

Pág. 582 (604)

Fig. 12.3 Gromov's proof of the sphere theorem.

Fig. 12.5 Gromoll's proof of the sphere theorem.

The equidistant hypersurface.

cf: theorem 299_

Pág. 593 (615) -

theorem 307 (Durumeric 1987 [466])

For a suitable ε, any compact Riemannian manifold of diameter larger $\pi\sqrt{\delta}/2 - \varepsilon$ with sectional curvature $K \geq 1$...

Ref.: Durumeric (466)

A generalization of Berger's theorem on almost $\frac{1}{4}$ - pinched manifolds.

b) Differential Geometry for Physicists and Mathematicians

José G. Vargas

→ Moving Frames and Differential

25.08.2019

X

forms: From Euclid Past Riemann.

Pág. 83 - Two Klein Geometries.

Pág. 109 - Euclidean Klein Geometry.

Pág. 191 - Riemannian Spaces and Pseudo-Spaces.

b) Source: 20190825_02/ Riemann/ Riemann Geometry/ Vargas..._pdf

c) Geometry of Riemann Surfaces

Source: 20190825_02/ Rieman/ Riemann Geometry/ Geometry of Riemann Surfaces.. pdf

d) Engineering Electromagnetics

Source: 20190825_03/ General Relativity.../ [William Hayt...].. Engineering Electromagnetics...pdf.

obs: Desenhos geométricos interessantes.

Veja Pág. 04(22), 15(33), 16(34) etc.

e) Geometry of Riemann Surfaces and Teichmüller Spaces.

Vol. 169

25.08.2019

Source: 20190825/ Symmetry Representation theory and Its Applications/ Geometry of Riemann... pdf.

f) An Introduction to Riemannian Geometry.
with Applications to Mechanics and Relativity

Author: Godinho

Veja aspectos topológicos.

Source: 20190825/ An Introduction to Riemannian Geometry/ An Introduction to Riemannian... pdf.

g) Handbook of Fractional Calculus with Applications

Kochubei

Source: 20190825/ Handbook.../ Handbook... pdf

Pág. 25(33) — 33(41)

2.1 Riemann-Liouville fractional integrals on a finite interval.

h) Differential Equations as Models in

25.08.2019

Science and Engineering
Baker
Source: 20190825/ Differential Equations
as Modes.../ Differential....pdf

i) Fixed Point Theory in Distance Spaces
Kirk
Source: 20190825/ Fixed Point.../
Fixed Point...pdf

Obs.: Part I Metric Spaces.

Obs: Metric Spaces(.pdf)
Source: 20190825/ Espaços_Métricos.pdf
+ Geometry of Cuts and Metric.
(mesmo diretório do anterior)

j) Matroid Theory
Source: 20190825/ Matroid Theory/
James Oxley - Matroid
Theory...pdf.

Pág. 301 (311)
Fig. 8.3

Lemma 8.2.11 Let X and Y be disjoint subsets of the ground set of a matroid M.

Obs: Usar para calcular espelhos entre planos.

25.08.2019 X

K) Estes livros estão no mesmo diretório:
(20190825-03/ General Relativity.../
a) General Relativity, Cosmology and Astro physics;

b) Einstein's General theory of Relativity
Grøn

c) Relativity — Rindler
Pág. 165 Curved spaces and the basic ideas of general relativity.

* Pág. 168- Fig. 8.4
(185) (Gráfico - Spherical)

L) Finite Frames —
Source: 20190824/Finite Frames theory and Applications/ Finite Frames_.pdf
Pág. 17(32)
Fig. 1.1 Mercedes-Benz frame

$$\left(\sqrt{\frac{2}{3}}\begin{pmatrix}0\\1\end{pmatrix}, \sqrt{\frac{2}{3}}\begin{pmatrix}\frac{\sqrt{3}}{2}\\-\frac{1}{2}\end{pmatrix}, \sqrt{\frac{2}{3}}\begin{pmatrix}-\frac{\sqrt{3}}{2}\\-\frac{1}{2}\end{pmatrix}\right)$$

M) VNR Concise Encyclopedia of Mathematics
Source: 034_NOVOSARQUIVOS_2019/
others files:

a) Cours de mathématiques élémentaires.pdf
(2000 questions);
b) Mathematical Writing;
c) Matrix Operations;
* d) Geometry the line and the Circle.
e) Complex Analysis

data 28.08.2019
S T [✓ Q] Q S S D

90) Books:

a) Calculus in Context.
Source: 20190828-03/ Calculus in Context...pdf
Pág. 112(137)
Local Linearity (A microscope gives a local view)

Pág. 113(138)
Figura (magnification)

b) Hadamard's Plane Geometry
Source: 20190828-03/ Hadamard's Plane...pdf
Pág. 51(62)
Fig. t81

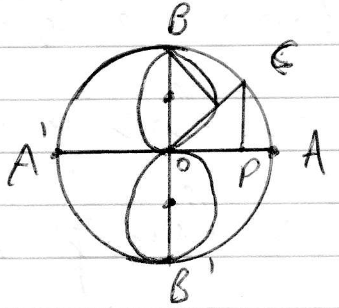

Exercise 61.

On each radius of a circle we lay off, starting from the center a length equal to distance from the endpoint of the radius to a fixed diameter. Find the locus of the endpoints of the segments thus obtained.

c) Nuclear Physics
Source: ~~A~~/ 20190828-03/ Nuclear Physics - Springer-Verlag Berlin Heidelberg (2014).pdf
Pág. 09(27)
Fig. 1.12 → Elements of solid angle for a general case
Fig 1.13 → Element of solid angle for an azimuthal symmetry.

data 28.08.2019

D S T [✓ Q] Q S S D

d) Heavenly_Mathematics_The_Forgotten_Art_of_Spherical_Trigonometry_Brummelen.pdf
Source: 20190828-03/Heavenly...pdf
Pág. 90(107)
Fig. 5.12 - the pentagramma mirificum.
Pág. 104(129) - Fig. 6.6
The proof of Delambre's first analogy.

e) https://www.loc.gov/resource/rbc0001.2010english20856/

https://www.loc.gov/item/03020856/

https://www.loc.gov/item/54047344/

↳ Dictionnaire mathématiques.

91) 20190831 BOOKS:

a) Stefan Ivanov - Theoretical and Quantum Mechanics

- Lagrangian Functions and Lagrange's Equations - Pág. 03
- The Schrodinger equation and time dependence of the observables Pág. 265

b) Source: 20190831/Fundamentals of Quantum Mechanics/Stefan...pdf

data 31.08.2019

b) Classical Fields; General Relativity and Gauge Theory. (M. Carmeli...pdf)

- The Geometry of Curved Spacetime - Pág. 20 - 83

↳ The Riemann Curvature Tensor → Pág. 67

- The Einstein Field Equations - Pág. 84 - 154.

Source: 20190831-02/ A Primer in Tensor Analysis and Relativity/

c) N. N. Puri - Fundamentals of Linear Systems for ph

Source: 20190831/ Fundamentals of Linear Systems for Physical Scientists and Engineers/ N. N. Puri... pdf

d) Modern Atomic Physics

Source: 20190831/ Natarajan - Modern Atomic Physics/ Natarajan...pdf.

Pág. 32(55)

1. Hyperbolic transform of the HO Hamiltonian.

$$b \equiv a \cosh r - a^{\dagger} e^{2i\theta} \sinh r$$

$$b^{\dagger} \equiv a^{\dagger} \cosh r - a e^{-2i\theta} \sinh r$$

e) Einstein's General Theory of Relativity

Source: 20190831-02/ A Primer in Tensor Analysis and Relativity/ Einstein... pdf

data 31.08.2019

- The Riemann Curvature Tensor
  Pág. 155
- Extrinsic and Intrinsic Curvature
  Pág. 161
- The equation of geodesic deviation
  Pág. 164
- Spaces of constant curvature
  Pág. 166
- Einstein's Field Equation - Pág. 177
  ↳ From Newton's law of gravitation to Einstein's field equations.

f) The General Theory of Relativity
Source: 20190831-02 / A Primer in Tensor Analysis and Relativity / Anadijiban... pdf

- Chapter 2 - The Pseudo-Riemannian Space-Time Manifold M4
  Pág. 105 (133)
- Chapter 8 - The Coupled Einstein-Maxwell-Klein-Gordon Equations.
- Appendix 4 - Conformally Flat Space-Times and "The Fifth Force".

31.08.2019
8

g) The Special Theory of Relativity
Source: 20190831_021 A Primer in Tensor Analysis and Relativity / Costas....pdf

- Maxwell's Equations and the Wave Equations
Pág. 28 (45)

- 2.4 The Relativity of Time and Length (Lorentz Transf.)
Pág. 53 (69)

- Chapter 3 Relativistic Kinematics
3.1 The Lorentz Transformation for the Coordinates of an event

h) Special Relativity and Quantum theory
Source: 20190831_021 A Primer in Tensor Analysis and Relativity / Special Relativity and Quantum theory....pdf
Pág. 10 (23)

Fig. 5 Direct measurement of the curvature by means of a clock and mirror...

Curved space $\frac{t_2 - 2t_2 + t_3}{ct_2^2} = \frac{11}{a} = \sqrt{1/2\, R}$

Obs: Interessante !!!

data 31.08.2019

S T Q Q S [✓ S] D

i) Energy and Geometry
Source: 20190901/ Energy and Geometry ... / 29970 Energy and Geometry... pdf

- Deformed Minkowski Space-Time
  Pag. 7

- Deriving the Deformed Lorentz transformations
  Pág. 21

- Metric of Interactions
  Pág. 93
  ↪ Energy-Dependent Phenomenological Metrics for the Four Interactions
  Pág. 63.

- The Gravitational Mass of Electron from Geometry
  Pág. 125

---

j) SUMA 58
Source: 20190901/ Suma_58.pdf
Pág. 103(101)
Gauss y el polígono regular de 17 lados

k) Theory_of_Relativity_important
Source: 20190831_02/

l) Poincaré oeuvres - 10 vol Source: 20190831_02/

data 08.09.2019 ✗

92) 20190908

Mathématiques Speciales

1) Source: 20190908_03/Mathématiques_Speciales/Poncelet/revue de mathmat23 pari.pdf

Pág. 289(207 Équation du troisième degré
(2077)
Pág. 289.

Pág. 292(300) - Fig. 2

Pág. 296(304) - (2107 ikm)

Soit M le point le plus haut d'une cycloïde engendrée par un cercle.

2) Source: 20190908_03/Mathématiques_Speciales/Poncelet/revue de mathmat21 pari.pdf

Pág. 294(302) - Remarques géométriques

3 Figures.

Pág. 297(305) - ellipse

(93) 20190906
Source: 20190907/ } depois

3) Source: 20190908-03/Mathématiques_Speciales/revuedemathmat11pari.pdf

Pág. 01(11) - Démonstration du théorème de Poncelet sur les polygones dans une l'une et circonscrits a une autre

data 08.09.2019

S T Q Q S S [✓ D]

Pág. (~~44~~) 33(43) -

Paramètre tangentiel d'un cone du second ordre.

Source:

www.archive.org/details/revuedemath
mat11pari/

—

Pág. 42(52) - item 1 e 3

bitangent aux coniques.

Pág. 43(53) — les cercles.

—

93) 20190906

Source: 20190907/Introduction to the Physics of Electrons in Solids by Henri Alloul/[Graduate Texts...] - Introduction to the Physics of Electrons... pdf

Pág. 127(155)

Fig. 4.13 (a) Hexagonal reciprocal lattice of graphene and its first Brillouin zone...

Dirac points K^+ and K^- of this cell.

data 08.09.2019

S T Q Q S S X

↓

94) Light Deflectiôn

Source: 20190907/ Singularity Theory and Gravitational Lensing by Arlie O. Petters/ Joachim .... pdf

Pág. 04 (27) - Light Deflectiôn

a) Henry Cavendish (1784) - First (isolated scrap)

b) Johann von Soldner (1801)

"Ueber die Ablenkung eines Lichtstrals von seiner geradlinigen Bewegung, durch die Attraktiôn eines Weltkörpers, an welchem er nache (nahe) vorbeigeht".

("On the Deviatiôn of a Light Ray from its Motiôn along a Straight Line through the attractiôn of a Celestial Body which it Passes Closeby")

This article appeared in

"Berliner Astronomisches Jahrbuch für das Jahr 1804"

Pág. 945 - 947

Einstein utilizou isto em sua teoria (OH!!!) novidade?!

data 08.09.2019

S T Q Q S S D ✗

1905 - 1801 = 104 anos

(novíssima descoberta!!!)

Na realidade Einstein utilizou os estudos precedentes do físico neerlandês Hendrik Lorentz e as descobertas de Cavendish e Soldner em sua teoria da relatividade, além de outros cientistas.

Vejamos:

a) Système de coordonnées Soldner

b) La Constante Ramanujan-Soldner

Sources: (others)

https://zenodo.org/record/1424375/files/article.pdf

Obs: Este arquivo não está mais disponível neste website, mas eu tenho uma

data 08.09.2019
S T Q Q S S X

cópia armazenada na pasta
20190907/Important/Johann
von Soldner/
article.pdf

Author: P. Lenard
Pág. 593 – 604
20 Juli 1921.

—

obs (off). Johann Georg von Soldner

—

https://en.wikisource.org/?
curid=755966.

—

other (source)
https://archive.org/details/
bub_gb_PN43AAAAMAAJ/page/
n167
Pág 161-172 (169-180)
Copy: 20190907/Important/Johann
von Soldner/

—

jandaia

data 08.09.2019

X

Deflection of Light and particles by moving gravitational lenses

Olaf Wucknitz

Source: arXiv: astro-ph/
0401362v1
19 Jan 2004

Copy: 20190907/Important/
0401362.pdf

Nesta mesma pasta, nós temos uma tradução:

Translation_On_the_Deflection_of_a_Light_Ray_from_its_Rectilinear_Motion.pdf

—

Light Deflection (cont.)

Even earlier:

c) John Michell (1783)

d) Pierre-Simon Laplace (1796)

independently investigated light being subjected to gravity, but in the context of a body

data 08.09.2019
8

acting on the very light it emits.

The both basically arrived at a notion (notion) resembling a "black hole", that is, a sufficiently massive star can capture its own light, leaving the star invisible.

Michell and Laplace did not write about the situation of light coming from a distant source and undergoing deflection as it passes near a gravitating body.

They are who inspired the works of Cavendish and von Soldner

Soldner:

"Thus when a light ray passes by a celestial body, it will, instead of a going on in a straight direction, be forced by its attraction to describe a hyperbola whose concave side is directed against the attracting body."

He showed that a light grazing the sun would be deflected by

0.84 seconds of arc.

data 08.09.2019

S T Q Q S S X

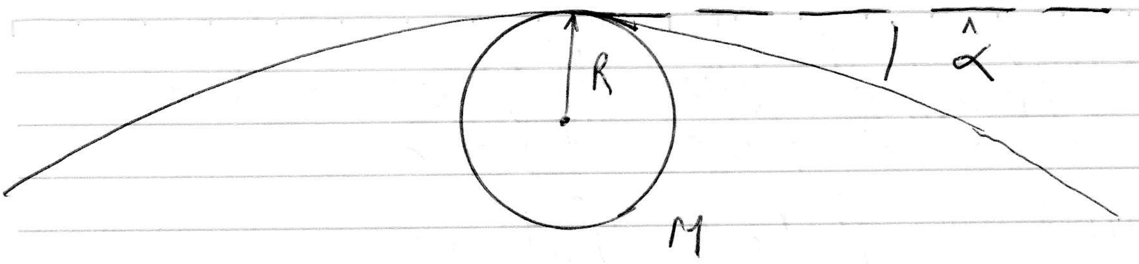

$$\hat{\alpha} = \frac{4GM_\odot}{c^2 R_\odot} \simeq 1.''74$$

Source: Pág. 06(29)
20190907/Singularity theory and Gravitational Lensing by Arlie O. Petters/.

off: Constante Gravitational.

$$G = 6{,}674184 \times 10^{-11} \, m^3 \, Kg^{-1} \, s^{-2}$$

obs:

$$F = G \frac{m_1 m_2}{r^2} \qquad (G \neq g)$$

g (aceleração da gravidade) $= \dfrac{GM}{r^2}$

=

data 08.09.2019

S T Q Q S S [✓ D]

Veja também:

Teoria Gravitacional de Le Sage

$$G = \frac{\varepsilon_c \sigma^2}{4\pi M_n^2}$$ (Sergey Fedosin)

Obs.: S. G. Fedosin

I) The graviton field as the source of mass and gravitational force in the modernized Le Sage's model.

Source: Physical Science International Journal
ISSN: 2348-0130
Vol. 8, Issue 4, pp. 1-18 (2015)
http://dx.doi.org/10.9734/PSIJ/2015/22197
↳ www.sciencedomain.org/index.php?/abstract/12140

obs: http://sciencedomain.org/review-history/12140

data 08.09.2019

S T Q Q S S ☒

Original Manuscript
www.sdiarticle2.org/prh/
PSIJ_33/2015/Ms_PSIJ_22197.pdf
or
www.sciencedomain.org/download/
MTIxNDBAQHBm.pdf

II) Fedosin S. G.
The Force Vacuum Field as an Alternative to the Ether and Quantum Vacuum.
Source: WSEAS Transactions on Applied and Theoretical Mechanics
ISSN: 1991-8747/2224-3429
Volume 10, Art. #3,
pp. 31-38 (2015)
Source: wseas.org/wseas/cms.action?
id=10178

data 08.09.2019

Download article:
www.wseas.org/multimedia/journals/mechanics/2015/a065711-221.pdf

III) Fedosin S. G.

The charged component of the vacuum field as the source of eletric force in the modernized Le Sage's model.

Source: Journal of Fundamental and Applied Sciences,
Vol. 8, No. 3, pp. 971-1020 (2016)

https://dx.doi.org/10.5281/zenodo.845357
↳ https://zenodo.org/record/845357

September 1, 2016

data 08.09.2019
S T Q Q S S D X

Notice de la Vie et des
Écrits de George-Louis
LE SAGE de Genève.

Source: Pierre Prevost
https://gdz.sub.uni-
goettingen.de/id/
PPN526449152/

Download:
https://gdz.sub.uni-goettingen.de/
download/pdf/PPN526449152/
PPN526449152.pdf

—

Photon-Graviton Recycling
as Cause of Gravitation
Apeiron, Vol. 14, No. 3, July 2007.
Matthew R. Edwards
Source: http://redshift.vif.com/JournalFiles/
V14N03PDF/V14N3EDW.pdf

landaia

(95)

94) 20190909

data 09.09.2019

Principles of thermodynamics

Source: 20190909_02/Principles of thermodynamics/[Jean-Philippe ...]...pdf

Pág. 72(85)

4.2 Fundamental Relations

4.2.1 Gibbs Relation

→ Fig. 4.1 →

Convex surface corresponding to the internal energy U(S,V) as a function of entropy S and volume V in state space (S,U,V).
The vertical planes at fixed S or fixed V define curves on the surface U(S,V).
The slopes of these curves correspond to the partial derivatives.

$$dU = T\,dS - p\,dV + \sum_{A=1}^{n} \mu_A\, dN_A \quad (4.1)$$

Figure → convex surface.

Important:

data 09.09.2019
[✓ S] T Q Q S S D

Pág. 71(84) →

the international organisation IUPAC recommends using the expressions

"Helmholtz free energy" A

and

"Gibbs free energy" G.

—

Important:

Fazer um estudo com este gráfico (convex surface) com o item 94 (α) - Soldner (1."74)

⇔ Light Deflection

—

96) Redefining the Kg (20190924)

Source:

https://www.youtube.com/watch?

V=Oo0jm1PPRuo
(zero)

Obs: World's Roundest object!

https://www.youtube.com/watch?

24.09.2019

S [✓ T] Q Q S S D

$1 Kg = 2,15 \times 10^{25}$ atoms
Si 28 (Silício).

Obs.:

Recalibrar a constante de Planck usando a Balança de Watt
(Watt Balance)
↳ chamada agora
Balança Kibble

Important,

Utilizar esta nova redifinição do Kg para ajudar a calcular o desvio da luz
(Cavendish).

—

Kg →

2 formas:

1/ esfera de silício para determinar e definir o número de Avogadro
(Constante de Avogrado)

$1 m$ = distance light travels in $\frac{1}{299.792.458}$ s

light speed = 299,792,458 m/s

data 24.09.2019

S [✓ T] Q Q S S D

2) Balança de Watt para definir a constante de Planck. (Kibble).

2011 → definiu-se Kg baseado na constante de Planck

Planck's Constant (h)
Avogadro's Number (N_A)

$mg = BLI$

(campo magnético) e comprimento do fio da bobina.

O movimento induz um tensão na bobina que é igual ao campo magnético.

X o comprimento do fio da bobina,

X sua velocidade.

Equações:

$V = BLv$ $IBL = mg$

⟹

data 24.09.2019
S [✓ T] Q Q S S D

$$\frac{V}{v} = BL \qquad BL = \frac{mg}{I}$$

$$\frac{V}{v} = BL = \frac{mg}{I} \Rightarrow \frac{V}{v} = \frac{mg}{I} \Rightarrow$$

$$VI = mgv$$

↳ tensão x intensidade da corrente = massa x aceleração de gravidade x velocidade

Lado esquerdo: potência elétrica — Lado direito: potência mecânica

equilíbrio

Mudando disso para a constante de Planck, o número que relaciona a frequência de um fóton com sua energia?

A tensão pode ser medida acuradamente usando um efeito quântico macroscópico que envolve junções Josephson (Efeito Josephson).

data 24.08.2019

Junção Josephson

consiste em 2 supercondutores separados por um pedaço fino de isolante.

Você aplica uma radiação de microondas a essa junção usando uma tensão em todo o dispositivo e seu valor é precisamente conhecido por ser

$$V = \frac{hf}{2e}$$

h = constante de Planck

f = freqüência da radiação

e = carga em um elétron

Empilhando $\underline{n}$ → number of junctions (em série)

$$V = n\frac{hf}{2e}$$

Essas junções são colocadas no circuito com a bobina, como é movido pelo campo e assim, você equilibra exatamente a tensão induzida na bobina usando essas junções Josephson

⇒

data 24.09.2019

$n\frac{hf}{2e}$ $\qquad$ $n\frac{hf}{\cancel{2e}\,e} \leftarrow (2e)$

$$VI = mgv \Rightarrow$$

$$V\frac{V}{R} = mgv$$

A corrente é passada através de um resistor e nós medimos essa tensão novamente usando junções Josephson

Para medir a resistência, usamos outro efeito quântico macroscópico chamado de efeito Hall Quântico.

A medida da resistência será uma fração inteira

$$\boxed{\frac{1}{p}\frac{h}{e^2}}$$

↳ carga do elétron

Substituindo tudo, teremos h (constant Planck)

$$h = \frac{4}{pn^2}\frac{gv}{f^2}m$$

↳ frequência

Alcance 10 ppb (partes por bilhão).

Vejamos os termos.

24.09.2019

V = velocidade da bobina à medida que este se move através do campo magnético

Para medir esta velocidade usamos um interferômetro à laser
(interferometer)

À medida que a distância à bobina muda as franjas de interferência passam por um detector.

Quantas franjas passaram em um certo período de tempo podemos determinar a velocidade da bobina.

g = usamos gravímetro

—

96) 20190924 $\overset{\text{write}}{(20190925)}$

Complementa o item anterior.

Ⓐ Spherical Coordinates

Source: 20190924/Quantum_Physics/epdf.pub_quantum-physics-3rd-edition.pdf

25.09.2019
S T [✓ Q] Q S S D

Pág. 125(167)

Fig. 7-2 - The definition of the spherical coordinates used in the text and the relation between the cartesian coordinates (x, y, z) and the spherical coordinates (r, θ, φ).

Ⓑ Length Contraction

Source: 20190924/ Modern Physics/
epdf.pub_modern-physics-3rd-edition.pdf

1) Pág. 18(35) - Length Contraction
2) Pág. 20(37) - Contraction of a Spaceship

Fig. 1.14 → Computer-simulated photographs of a box (a) at rest relative to the camera (b) moving at a speed $v = 0.8c$ relative to the camera.

Ⓒ The Michelson Interferometer

Source: 20190924/ Introduction to Optics/
epdf.pub_introduction-to-optics-3rd-edition.pdf

Pág. 193(212)

Figure 1.

data 25.09.2019

Ⓓ Metals and Wigner-Seitz theory^{x} (B)
(1.1.3)

(Source: 20190924/Quantum_Physics/
(não) epdf.pub_quantum-physics-3rd)

Source: 20190924/Solid-State Physics
Introduction to the Theory/
epdf.pub_solid-state....pdf

Pág. 11(34)

The ground-state energy of an electron of mass M in a box of volume V is [19]

$$E = \frac{\hbar^2 \pi^2}{2M} V^{-2/3}.$$

Ⓔ Vector Spaces and Coordinate Systems

Source: 20190924/Newton/
epdf.pub_classical_mechanics.pdf

Pág. 57(74) — Chapter 2.

(97) 20190924 — (20190925)

Ⓘ https://einsteinpapers.press.princeton.edu/vol6-trans/42

25.09.2019

The Formal Foundation of the General Theory of Relativity.

(II) www.mpiwg-berlin.mpg.de/sites/default/files/Preprints/P118.pdf

Source: 20190925/the Einstein theory of Relativity/P118.pdf

Hilbert's Foundation of Physics: from a theory of everything to a constituent of general relativity

Stachel.

(III) www.bu.edu/cphs/files/2015/04/2007-Renn-Stachel.pdf

Source: 20190925/The Einstein theory of Relativity/2007-Renn-Stachel.pdf

(98) The Physical and Mathematical Foundations of the Theory of Relativity A Critical Analysis by Antonio Romano.
2019 - 498 pag.
ISBN 3030272362

→

data 25.09.2019

S T [✓ Q] Q S S D

Source: 20190925-02/ The Physical and Mathematical Foundations of the Theory of Relativity/ ~~tkkx~~ 3030242362.pdf

other:

The Special Theory of Relativity
Costas Christodoulides.

Source: 20190925-02/ The Physical and Mathematical Foundations of the Theory of Relativity/ The Special....pdf!

(99) 20190925

cont. item anterior

The Special Theory of Relativity
Einstein's World in New Axiomatics

Helmut Günther
Volker Müller
(Volker)

Source: 20190926/ The Special Theory of Relativity Einstein's World in New Axiomatics/ Sanet.st_9811377820.pdf

data 25.09.2019
S T [✓ Q] Q S S D

(100) 20190925

a) Pencil of Cubics and algebraic Curves in the Real Projective Plane

Source: 20190925/Pencil of Cubics.../
Pencil of cubics... pdf.

b) Mecánica de fluidos

Source: 20190925/Mecanica de Fluidos (Spanish Edition)/
Mecánica de fluidos... pdf

Pág. 185(208)

5-2 Conservación de la masa

$$E = mc^2 \qquad (5\text{-}3)$$

Fig 5-3

La velocidad normal V_n para una superficie es la componente de la velocidad perpendicular a esa superficie.

c) An Advanced Complex Analysis Problem Book

Topological Vector Spaces, Functional Analysis, and, Hilbert Spaces of

data 25.09.2019

S T [✓ Q] Q S S D

Analytic Functions

Source: 20190525/ Advanced Functional Analysis/ Daniell Alpay - An Advanced Complex Analysis... pdf.

d) Tangent plane to a two-dimensional surface at the point p

Fig. 2.13

Veja: Fig. 2.14 —

The manifold S^1 along with a number of tangent lines depicted (top left)

The manifold S^2 along with a number of tangent planes depicted (top right).

The manifold T along with a number of tangent planes depicted (bottom).

For both manifolds S^2 and T several vectors in each tangent plane are also depicted.

Source: 20190925-02/ A Visual introduction to Differential Forms and Calculus on Manifolds/

data 25.09.2019

S T [✓ Q] Q S S D

/Fortney, J.P. - A Visual...pdf

Pág. 39(50).

Pág. 179(187) -

Fig. 5.35 - A generic one-form in R^3 is comprised of sheets filling R^3

e) Albert Einstein's Special Theory of Relativity

Miller.

Source: 20190926/ The Special Theory of Relativity Einstein's World in New Axiomatics/ Arthur I. Miller - Albert Einstein's... .pdf

Pág. 115(152) -

"Einstein's own version of Kant as viewed through the pages of, in particular, Hertz and Poincaré.

Einstein's viewpoint then was also tempered by the works of Mach and Poincaré, ..."

Pág. 189(226) — Length and Time

data 25.09.2019

are relative quantities

Obs: Michelson experiment...
H. A. Lorentz (1895)

Pág. 189 (226)

Pág. 209 (246)

7.1. Moving Bodies:
Einstein's Length
Contraction

... what Lorentz introduced as an auxiliary quantity and called the local time [then] the hypothesis of H. A. Lorentz and Fitzgerald appears as a secondary consequence of the theory.

A. Einstein (1907e)

Important: Einstein quickly demonstrated that

25.09.2019

a body having a spherical shape relative to an observer in its rest system K, Eq. [§4.1]...

---

f) Hyperbolic Triangle centers

Ungar

Source: 20190926 | The Special Theory of Relativity Einstein's World in New Axiomatics | Ungar, Abraham.. Hyperbolic.pdf

Pág. 64(79)

3.4 The Resultant Relativistically Invariant Mass Theorem

Lemma 3.1.

---

(101) 20190927

$\pi = \frac{355}{113}$

Source: 20171111 | Aurelio Baldor - Geometria.pdf

Pág. 196(200) → 258. El numero π.

Pág. 198(202) → Metodo Grafico para

data 27 09 2019
S T Q Q X S D

rectificar aproximadamente una circunferencia. (important !!!)

Fig. 221.

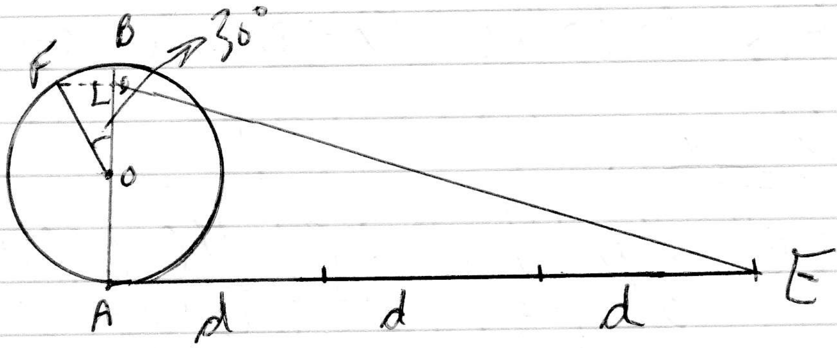

Vejamos:

1) Tracemos un diámetro $\overline{AB}$ (fig. 221);

2) Por A, tracemos la tangente;

3) Sobre dicha tangente y a partir de A tomemos tres diámetros. Se obtiene el punto E;

4) Tracemos ∠ BOF = 30°;

5) Bajemos $\overline{FD} \perp \overline{OB}$ e

6) Unamos D com E. $\overline{DE}$ es la solución del problema.

Obs: Use o Teorema de Pitágoras para calcular a hipotenusa deste triângulo.

Obs: Veja a JUSTIFICACIÓN desta construcción.

data 27.09.2019
S T Q Q [✓ S] S D

102) Space Curve

Source: 20190727_03/Volume-2.pdf

Source 2: 036_N_O_V_O_S_A_R_Q_U_I_V_O_S_2019/
volume-2.pdf

Introduction to Calculus
Volume II by
J.H. Heinbockel

Pág. 113(105)

Fig 7-10

Space curve revolved about line to form surface of revolution.

103) 20190910 (páginas avulsas signadas nesta página)

Source: 20190910_03/

a) Mathematics & Physics for Programmers, 2nd Edition/
Danny Kodicek ....pdf

Pág. 393(414) → 3-D Geometry.

b) Generalized Calculus with Applications to Matter and Forces/
Generalized Calculus...pdf

data 05.10.2019

S T Q Q S [✓ S] D

Pág. 51(101) - 50(101).

2.1.1. Tangential and Normal Components of the Tension.

Fig. 2.1 The tension along the tangent to an elastic has vertical and horizontal components (a) that simplify in the linear case of small slope (b).

b) Pág. 295 (346) -

Differential Operators and Geometry

—

c) Complex Analysis with Applications to Flows and Fields/ Complex Analysis... pdf

Pág. 24(65) -

Fig. 3.1 -

Taking two complex numbers u, v as the sides of a rectangle their sum is the diagonal.
Their difference u - v is the diagonal of the rectangle obtained by reflection on the v - direction.

d) Important/ [Luis_Manuel]...pdf

Jandaia

data 05.10.2019
S T Q Q S [✓ S] D

Pág. 05(21)

1.1.3 - Fórmula de Moivre
(1730; Euler 1749)

$x=\rho\cos\theta \qquad y=\rho\sin\theta$

$z=\rho(\cos\theta+i\sin\theta)=\rho e^{i\theta}$

$e^{i\theta}\equiv\cos\theta+i\sin\theta$

(leia o texto todo).

Pág. 13(29) → Produto de complexos, homotetia e rotação

obs.: Significado do símbolo imaginário.

Dados dois complexos na forma polar:

$u=|u|e^{i\alpha}$, $v=|v|e^{i\beta}$

define-se o produto como o complexo:

$uv=|u||v|e^{i(\alpha+\beta)}$

* * *

e) Mathematical Methods for Physics |
Sand. ST-0738201251.pdf

data 05.10.2019

Pág. 34(53) - Spherical Coordinates *

(r, θ, φ)

(important!!!).

—

f) 20190910_02 / An Introduction to Fluid Mechanics by Chung Fang /
Fang... pdf

Pág. 183(201)

7.2. the Euler Equation in Streamline Coordinates

103) Solve number 42

3 Cubos

Source: 20190812_03 / New Scientist International Edition - September 14, 2019 (True Pdf) /
2019-09-14... pdf

Pág. 14(17)

—

$$K = x^3 + y^3 + z^3 \Rightarrow$$

data 05.10.2019

$$42 =$$
$$(-80538738812075974)^3 +$$
$$80435758145817515^3 +$$
$$12602123297335631^3$$

---

104) 20191005 (20191005)

a) Modifications of Einstein's Theory of Gravity at Large Distances.

Source: 20191004/Modifications of Einstein's Theory of Relativity at Large Distances/ Modifications....pdf

Pág 11(27) → "This is indeed the case for spherically symmetric solution that describe the exterior of stars, and in particular the Sun.
Consider for conveteness the case where

$$V = m^2(\varphi - \varphi_0)^2.$$

Performing a newtonian expansion one can calculate the newtonian limite of the metric.

Falando em metric (vejamos:

data 05.10.2019

b) Einstein metrics in dimension four

Source: 20191005/Important/
[F.J.E. - Dillen J... ldjvu]

Pág. 421(435)

A pseudo-Riemannian metric g on a (connected) manifold M is called an Einstein metric if its Ricci tensor Ric and scalar curvature s satisfy relations

$$Ric = \frac{s}{n} g, \quad ds = 0, \quad n = \dim M.$$

Voltando ao item $\underline{\underline{a}}$:

Pag. 11(27) →

The perturbations of the metric are

$$h_{00} = \frac{GM_S}{\varphi_0 r}\left(1 - \frac{1}{2\omega_0 + 3}\exp\left[-\sqrt{\frac{2\varphi_0}{2\omega_0+3}}\, m r\right]\right),$$

$$h_{ij} = \frac{GM_S}{\varphi_0 r}\delta_{ij}\left(1 + \frac{1}{2\omega_0 + 3}\exp\left[-\sqrt{\frac{2\varphi_0}{2\omega_0+3}}\, m r\right]\right)$$

where M_S is the mass of the Sun.

There is a Yukawa-like correction to the standard $1/r$ potential, with

05.10.2019

effective mass $M_{eff} = \sqrt{\frac{2\varphi_0}{2w_0+3}}\, m$ and*

range m_{eff}^{-1}.

the ratio of the perturbations of the time-time component h_{00} over any space-space diagonal component $h_{ij}|_{i=j}$, which is also known as the γ (Eddington) parameter is then given by

$$\gamma \equiv \frac{h_{ij}|_{i=j}}{h_{00}} = \frac{2w_0+3-\exp\left[-\sqrt{\frac{2\varphi_0}{2w_0+3}}\, mr\right]}{2w_0+3+\exp\left[-\sqrt{\frac{2\varphi_0}{2w_0+3}}\, mr\right]}$$

c) 12.3 Trajectories of free motion in Einstein's theory

theorem 12.1 For a timelike geodesic G the coordinates in a Riemann space can be chosen so that the Christoffel symbols vanish along G.

Source: 20191004-02/An Introduction to General Relativity and Cosmology/ Jerzy Plebanski...pdf

Pág. 126 (148) e seguintes.

data 05.10.2019

S T Q Q S [✓ S] D

Pag. 127(149)

Fig. 12.1 Construction of the Fermi coordinates in which the Christoffel symbols vanish along a given timelike geodesic G.

d) 2.3.1.1 Orthogonal Cylindrical Coordinates.

(r, θ, z)

$$R = (r\cos\theta)e_1 + (r\sin\theta)e_2 + z e_3$$
$$\equiv x^1 e_1 + x^2 e_2 + x^3 e_3$$

where

e_1, e_2, and e_3 are the orthonormal bases of Cartesian coordinates;

θ is the polar angle.

Source: 20191004/ Elementary Differential Geometry Tensor Analysis.pdf

Pág. 42(60).

Pág. 43(61) – Fig. 2.3 – Covariant bases of orthogonal cylindrical coordinates

Jandaia

data 05.10.2019

To simplify the formulation with Einstein symbol, the coordinates of u^1, u^2 and u^3 are used for r, θ, and z, respectively.

Therefore, the coordinates of $P(u^1, u^2, u^3)$ are given in Cartesian coordinates:

$$P(u^1, u^2, u^3) = \left\{\begin{array}{l} x^1 = r\cos\theta \equiv u^1\cos u^2 \\ x^2 = r\sin\theta \equiv u^1\sin u^2 \\ x^3 = z \equiv u^3 \end{array}\right\}$$

(estudar o restante).

Veja Pág. 45(63) Fig. 2.4.

Covariant bases of orthogonal spherical coordinates.

e) From Möbius to Gyrogroups

Source: 20191003_03/ Abraham-Ungar-Books/ Abraham_Ungar-A gyrovector space approach to hyperbolic geometry.pdf

Pág. 01(13) → Möbius transformations of the complex open unit disc $\mathbb{D}$

$$\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}$$

$$z \mapsto e^{i\theta}\frac{a+z}{1+\bar{a}z} = e^{i\theta}(a \oplus z),$$

data 05.10.2019
S T Q Q S [✓ S] D

f) Einstein space (4.3)

Einstein space is defined as a space which is homogeneous with regard to the Ricci tensor R_{ij}.

Thus, a space for which

$R_{ij} = \lambda g_{ij}$; λ = invariant at all points at every point of the space, then that space is called

Einstein space.

Inner multiplication of g^{ij} shows that

$$g^{ij} R_{ij} = \lambda g^{ij} g_{ij} \Rightarrow R = \lambda N$$

Thus for an Einstein space,

$$R_{ij} = \lambda g_{ij} = \frac{1}{N} R g_{ij}$$

Hence, for an Einstein space Eq. (4.56) holds at every point of the space.

Source: 20191005/Important/Textbook of Tensor Calculus and Differential Geometry (2012, PHI Learning).pdf

data 05.10.2019

Ⓢ Ⓣ Ⓠ Ⓠ Ⓢ ☒ Ⓓ

Pág. 243(254).

g) The Einstein Gyromidpoint theorem 6.92

Source: 20191003_03/Abraham_Ungar_Books/ Abraham A. Ungar - Analytic Hyperbolic Geometry and Albert Einstein's Special Theory of Relativity..pdf

Pág. 230(251) Theorem 6.92

Fig. 6.17

A left gyrotranslation by $x \in \{R^2, \oplus, \otimes\}$ of the gyrotriangle uvw of Fig. 6.16 in an Einstein plane is shown.

Pág. 232(253) - Fig. 6.18

The hyperbolic tetrahedron uvwx, that is, a gyrotetrahedron, is show in the Einstein gyrovector space

$$R^3_s = \{\mathbb{R}^3_s, \oplus_E, \otimes_E\}$$

data 05.10.2019

underlying the Beltrami ball model of hyperbolic geometry.

The gyrotetrahedron $uvwx$ is shown inside the

1-ball R^3_1 of the Euclidean 3-space R^3 where it lives.

The faces of the gyrotetrahedron are gyrotriangle.

h) A.2 Surfaces in 3-D Euclidean space

A.2.1 – Representations of surfaces; metrics

$$x = x^i(u, v)\, a_i,$$

where a_i is a fixed basis, i, j, k. Equivalently, we have

$x = x(u, v)$
$y = x(u, v),$
$z = z(u, v).$
for a given domain of the parameters.

These actually are coordinates on the surface. We may also give it as $f(x, y, z) = 0$.

data 05 10 2019

Pág. 248(267)

Source: 20191005/Important/Differential geometry for physicists and mathematicians-moving frames ....pdf.

i) 1.1.1. Three-dimensional vectors.

Source: 20191005/Important/Ibragimov-Tensors and Riemannian Geometry (2015, De Gruyter).pdf

Pág. 03(12)

Vector Algebra

$a \in \mathbb{R}^3$ be a three-dimensional (in particular, two-dimensional) vector.

Graphically, a is a directed line segment. Its' magnitude is denoted by $|a|$.

The scalar product $a.b$ of vectors a and b is a scalar quantity (i.e., a real number) defined by

$a.b = |a||b| \cos \theta$, where $\theta \ (0 \leq \theta \leq \pi)$.

05.10.2019

is the angle between the vector a and b.
The scalar product is denoted in the literature by (a, b).

The vectors a, b $\in \mathbb{R}^3$ can be represented in the rectangular Cartesian coordinates in the form

$a=(a^1, a^2, a^3)$, $b=(b^1, b^2, b^3)$,

which means that

$$a = a^1 i + a^2 j + a^3 k.$$

$(a^1 i)$

$$b = b^1 i + b^2 j + b^3 k.$$

where i, j and k are the unit vectors along the first, second and third coordinate axes in $\mathbb{R}^3$, respectively.

Then the scalar product is given by

$$a.b = \sum_{i=1}^{3} a^i b^i.$$

~~1(1915)~~ j) Brans-Dicke theory

Source: https://en.wikipedia.org/wiki/Brans-Dicke_theory

05.10.2019
X

In theoretical physics the Brans-Dicke theory of gravitation (Jordan-Brans-Dicke theory) is a theoretical framework to explain gravitation.

It is a competitor of Einstein's theory of general relativity.

It is an example of a scalar-tensor theory, a gravitational theory in which the gravitational interaction is mediated by a scalar field as well as the tensor field of general relativity.

other:

https://pt.wikipedia.org/wiki/Geometria_hiperbólica

105) a) Advanced Trigonometric Relations through Nbic Functions.

Source: 037_novosarquivos_2019/Trigonometric/Sardet.cd_8122430236.pdf

other: Trigonometric Functions and Complex Numbers

data 05 10 2019

Source: 037_novosarquivos_2019/Trigonometric/
Trigonometric Functions and
Complex Numbers.jnt

(Documento de Diário)

b) Geometry: Euclid and Beyond

Source: 20191005/Geometry Euclid and Beyond/
[Undergraduate Texts in Mathematics] ... pdf.

c) Unvergängliche Geometrie

Source: 20191005/Introduction to Geometry/
Unvergängliche ... pdf.

d) Geometry revisited

Source: 20191005/Introduction to Geometry/
Geometry Revisited ... pdf

Pág. 47 (59)

2.9 Morley's theorem

Theorem 2.91 - The points of intersection of the adjacent trisectors of angles of any triangle are the vertices of an equilateral triangle.

data 05.10.2019

S T Q Q S [✓ S] D

Figure 2.9 A

Naraniengar's proof requires a preparatory theorem or lemma.

Look: Figure 2.9 B.

c) Trick or Truth?

The mysterious connection between physics and Mathematics.

Source: 20191005/Trick or truth.../3319274945.pdf

---

106) 20191021

a) Pi_A Biography of the World's Most Mysterious Number...pdf

Source: 20191021/Pi... pdf

Pag. 80(81) - Archimede's Method for finding the value of π.

b) Manifolds Tensor Analysis and Applications

Source: 20191021/Manifolds, tensor analysis, and Applications/Abraham Ralph - Manifolds Tensor... pdf. (2003)

data 21.10.2019

c) Advanced Calculus - An Introduction to Linear Analysis...pdf

Source: 20191021/ Advanced Calculus An Introduction to Linear Analysis/.

Pág. 180 (206)

6.1 The Vibrating String and Trigonometric Series).

d) The Symmetries of Things

Source: 20191021/ The Symmetries of Things/ Sant.M - The_Symmetries_of_Things.pdf

Pág. 57 (76)

The Existence Problem: Proving the Proofs.

Figures (2)

1). Triangle on the sphere with angles $\frac{\pi}{5}$, $\frac{\pi}{3}$, and $\frac{\pi}{2}$

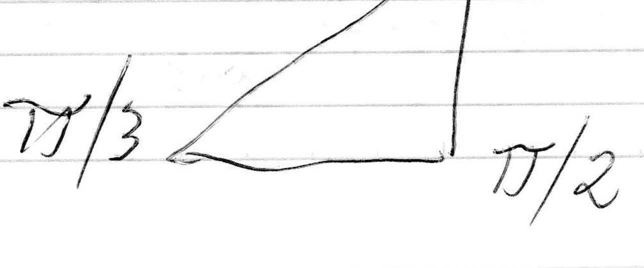

21.10.2019

2) A two-sided spherical polygon* with two angles of $\frac{\pi}{6}$ radians

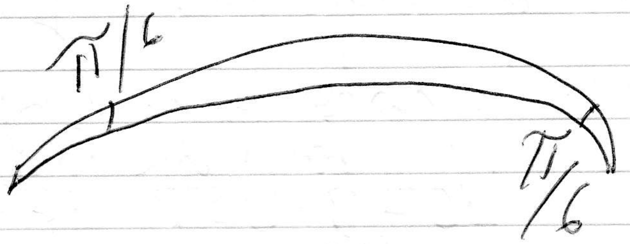

e) 20191022 / Cosmic DNA at the Origin.... /

Chris H. Hardy - Cosmic DNA at the Origin.... epub.

Pág. 60/339

Dirac Sea: the vacuum as a sea of negative energy

The British physicist Paul Dirac brought a new understanding of the vacuum as an infinite energy, as we conceive it now.
He predicted the existence of anti-matter back in 1928, with his famous Dirac Equation, that was the first bridge between quantum physics and relativity theory.

---

107) 20200125

data 25.01.2020

S T Q Q S [✓ S] D

DATA_NOVOVIVO_20_2020/images

I) Screenshot... Objeto giratório mais rápido do mundo.png

Use image: o rotor gira impulsionado pela força da luz
1579795322979_v2_450x450.jpeg

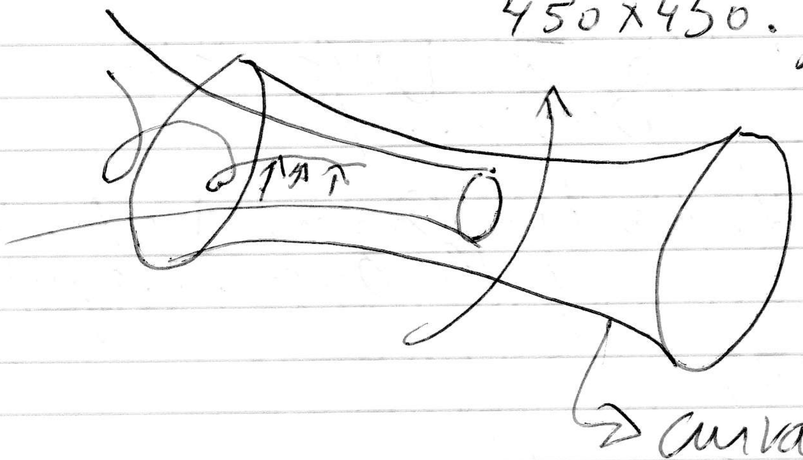

curvature (space)

Research!!!

Copy files:

II) NOVOVIVO_20_2020/Important/

Patricia M. Schwarz...
Special Relativity From Einstein to String 2004.pdf

Pág. 13(26)

A very small rotation with $\theta \sim 0$ can be written:

(1.30)

25.01.2020

$$\tilde{x} \simeq x + \theta y + O(\theta^2)$$

$$\tilde{y} \simeq y - \theta x + O(\theta^2)$$

In this infinitesimal limit, the rotation matrix $R(\theta)$ can be written in terms of a matrix r

$$R(\theta) \simeq I + \theta r + \dots,$$

$$r = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \qquad (1.31)$$

Obs: Using of rotations preserve the Euclidean metric with Gyrocommutative Gyrogroups,

I will be able to develop a new glimpse.

III) PC_Oculto/ –

Electromagnetic Fields (IEEE Press Series on Electromagnetic Wave Theory) – images/

Electromagnetic Fields..._0028.jpg

1.4 Green's Functions

25.01.2020

Green's functions by considering the displacement of a beam under a distributed force $p(x)$.

... Except at x' there is no force.

sobre uma superfíci.

Veja o item I,

uso da luz da laser sobre uma superfície.

Important:

A luz exerce uma pressão constante, conhecida como

"pressão de radiação"!

108) 20200127

1) www2.math.uu.se/~thomase/GeometryoverFields.pdf

From Geometry to Number
The Arithmetic Field
implicit in Geometry
Johan Alm.
49 pages.

27.01.2020

2) https://core.ac.uk/download/pdf/82832072.pdf

David Hilbert's Lectures on the Foundations of Geometry 1891-1902

ISBN 978-3-540-64373-9
661 pages.

3) https://hal.archives-ouvertes.fr/tel-02012674/document

On the Formalization of Foundations of Geometry

Pierre Boutry
160 pages.

4) https://plato.stanford.edu/entries/frege-hilbert/#HilbFounGeom

The Frege-Hilbert Controversy

obs:

a) www.digizeitschriften.de/dms/img/?PID=
GDZPPN002119102

→ Über die Grundlagen der Geometrie (1903)
von G. Frege in Jena.
p.319

data 27.01.2020

b) GDZPPN002119129

Über die Grundlagen der
Geometrie II (1903)
Von G. Frege in Jena
p. 368

c) GDZPPN002120461

Über die Grundlangen der
Geometrie - I (1906)
p. 293

d) GDZPPN002120631

Über die Grundlagen der
Geometrie (1906)
Frege
(Fortsetzung) — II
p. 377

e) GDZPPN002120666

Über die Grundlagen der
Geometrie (1906)
Frege
(Schluß) III
p. 423

—

5) https://math.berkeley.edu/
~wodzicki/160/

⇒

27.01.2020

MATH 160 – Introduction to History of Mathematics.

Links to: Papyrus Rhind 1 – 79.(pdf)

Papyrus Moscow → 1 – 25

Lahun Papyrus ... etc.

—

6) The Plimpton 322 Tablet

https://arxiv.org/pdf/1004.0025.pdf

obs:

The Plimpton 322 Tablet and the Babylonian Method of Generating Pythagorean Triples.

34 pages.

7) ftp.math.utah.edu/pub/text/toc/histmath.html

Volume 1 → 49 (2019).

—

8) https://geometryexpressions.com

Obs.: Pode-se usar na internet.
Download → pago.

=

date 27.01.2020

9) CORE. AC. UK

10) https://www.nbarchive.dk/publications/bcw/

Niels Bohr Collected Works.
12 Volumes

Obs.:

Harald August Bohr (1887-1951)

Niels Bohr (Niels Henrick David Bohr)
(1885 - 1962)

Harald August Bohr →
Collected Mathematical Works
3 volumes.

www.archive.org/details/
collectedmathema0000bohr
→ Volume I

collectedmathema0003Bohr
→ Volume III

Obs: Falta o Volume II.

27-01-2020

11) www.plouffe.fr/simon

10) cont.

Harald Bohr Papers

web.math.ku.dk/arkivet/hbohr/hbpapers.htm

12) 20200202

The Mathematical Papers of Georg Friedrich Bernhard Riemann (1826-1866)

Source:

https://www.emis.de/classics/Riemann/

13/ Latex

linorg.usp.br/CTAN/systems/win32/miktex/tm/packages/

109) 20200202

Inkscape program

Extensions → Render → Spirograph

→ Wireframe Sphere

→ Gear

data 02 02 2020

Source: Arq-Important_03/Inkscape_files/

a) drawing_20200202.svg

I) Gear → 24 teeth (360÷24=15°)

Configuration:

Nº teeth: 24

Circular pitch: 20 → 15

Pressure angle: 20

Diameter: 20

Units: px → mm (altere)

II) Spirograph

R – Ring Radius (px): 100

r – Gear Radius (px): 60

d – Pen Radius (px): 50

Gear placement { Inside (Hypotrochoid)
Outside (Epitrochoid)

Rotation – 0

data 02.02.2020

Quality (Default) — 16

b) Spirograph

spirograph_200.60.50.e.200.100.50.svg

I) Spirograph

200. 60. 50. Inside. 24. 16
(R) (r) (d) (GP) (R) (Q)

→ spirograph 10 points

II) Spirograph

200. 100. 50. Inside. 24. 16

ellipse.

Source: spirograph_200.60.50.e.200.100.50.svg

Position:

Spiro 1 → x 55,430
y 100,207
w 100,806
H 95,919 mm

Spiro 2 → x = 65
y = 134,805
w = 81,540
H = 26,723 mm

Obs: Use as setas para movimentá-los até a posição.

data 02 02 2020

1

c) spirograph_100.200.50.e.100.60.100.svg

Spiro 1 : 100 . 200 . 50
(R) (r) (d)

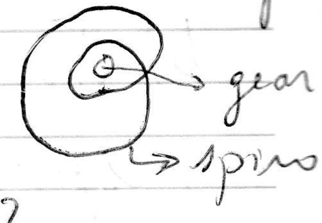

Spiro 2: 100. 60. 100 } 5 pontas
(R) (r) (d) (pentagrama)

d) spirograph_100.60.50.e.300.60.50.svg

R - 300
r - 60
d - 50
G - Inside
R - 0
Q - 16

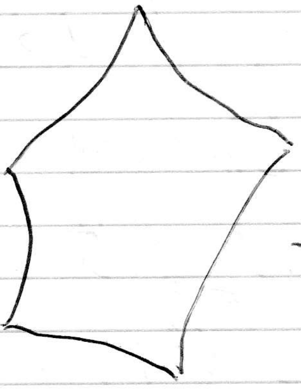

5 pontas.

e) ~~spirograph~~ gear _25.20.15.20.mm.e.spiro.300.60.50.inside.0.16.svg

A engrenagem e o spirograph ficam alinhado

25 dentes divididos por 5 pontas

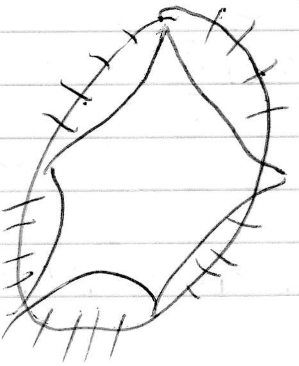

data 02.02.2020

Outras simulações:

f) spirograph_200.60.50.e.200.100.50_ajustado.svg

g) gear_25.20.15.20.mm.e.spiro_300.60.50.inside.24.16.svg

110) 20200202 Riemann

geometrias não-euclidianas

Lobachevsky
Bolyai
Gauss
Riemann

Bernhard Riemann
On the hypothesis which lie at the bases of geometry

---

postulado paralelo (euclides)

Não há geodésicas paralelas.

(Gauss foi o primeiro a testar a curvatura do espaço físico.

179° 59' 59.32"

data 02.02.2020

Usou as montanhas da Alemanha

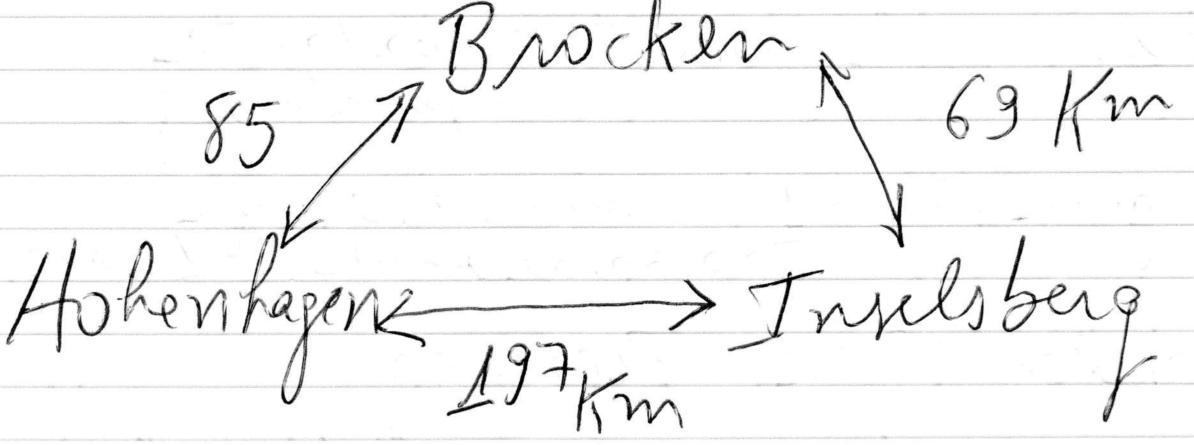
Brocken
85
69 Km
Hohenhagen
Inselsberg
197 Km

—

Beltrami (Eugenio)

"pseudo-esfera"

descrevendo um espaço hiperbólico de constante curvatura negativa por meio de uma equação análoga à da hiperesfera, evidencia a natureza comum das hipóteses de Bolyai e Lobachevsky e de Riemann.

Book: Uma história não-euclidiana.

Rosenfeld.

Source: Books_2020_06/Book1_27_2020/

Important

Boris A. Rosenfeld_A_history_of_non-Euclidean ... djvu

02.02.2020

Source: Books_2020_06/OTHERS/
Math24.net/

Screenshot_2020-02-03_
Curvature of Plane Curves.png

Osculating → Osculate

Verb:

(of a curve or surface) touch (another curve or surface) so as to have a common tangent at the point of contact.

Should the students be asked to sketch the osculating circle to a given curve at a given point, and then notice that as the point of contact changes the radius of the circle is inversely proportional to the measure of curvature?

If seven points on a twisted cubic be joined, two and two, by twenty-one lines, then any seven planes that contain these 21 lines will osculate a second cubic curve.

Book: A history of non-Euclidean geometry
B. A. Rosenfeld

data 02.02.2020

Pág. 280 (292) - Chapter 8

The Curvature of Space

Curvature and Intrinsic Geometry of a Surface in the Works of Euler

$$K = \frac{d\alpha}{ds} = \lim_{\Delta s \to 0} \frac{\Delta \alpha}{\Delta s}$$

K is also the reciprocal of the radius of curvature at the point in question, that is, the radius of the OSCULATING circle at that point.

(The osculating circle at a point P of a curve is defined as the limit of circles determined by three points on the curve as they tend to P.)

The concepts of the curvature of a curve and of the osculating circle were already know to Leibniz.

↳ obra: New reflections on the nature of the angles of tangency and osculation and their use in mathematical

02 02 2020

practice to replace complex figures with simpler ones.

Leipzig 1686 [312, v.7, pp. 326-329]

Leibniz also suggested the possibility of characterizing the curvature of a surface by means of an osculating sphere

↳ In a letter to Johann Bernoulli, dated July 29, 1698

111) ARQ-Important_05/Géométrie/Libre-Géométrie-Projective/Geometrie-Projective//

a) livre-exercices-de-geometrie-fgm-bnf.pdf

Pág. 78 (110)

177a. Théorème de Desargues

b) livre-geometrie-projective-daniel-perrin-1328 pages-pdf

Pág. 65(100)

Théorème de Newton

Figure 3.7 Le théorème de Newton projectif

data 02 02 2020
1

Figure 3.8 Le théorème de Newton affine.

Affine Adjetivo: allowing for or preserving parallel relationship

"still to be implemented are geometric primitives and affine transformation of image".

The simplest way to see why this assertion is true is to observe that any triangle in the plane can be mapped by an invertible affine transformation to an equilateral triangle.

Using parallelograms ~~from the previous~~ to define affine transformations may be a mixed blessing.

It can be obtained from the previous description by applying affine transformation to keep the polygon bounded as it degenerates.

Our interest is in the space of affine equivalence classes of equal-area polygons.

03.02.2020

112) Trigonometria Cubo
Curvilínea Vectorial
Projetiva Mussichiana.
↓ (TM)
Projetiva Diferencial Ortogonal
sobre a curva ciclóide.

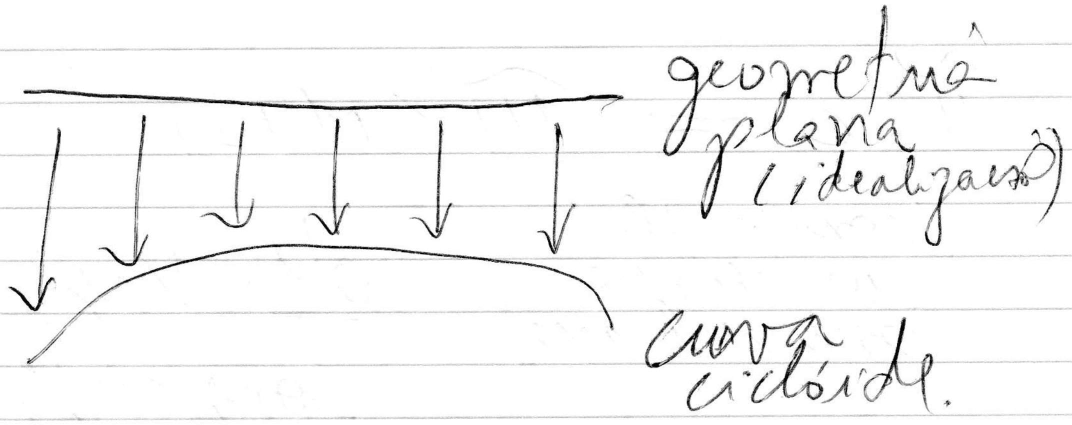
geometria plana (idealização)
curva ciclóide.

A velocidade maior entre 2 pontos ocorre no percurso de uma ciclóide e não em linha reta.
(espaço curvo → gravidade)
A luz segue o mesmo princípio.

A gravidade distorce o plano e a luz acompanha.

Veja: Grigori Perelman → Conjectura Poincaré

Einstein baseou-se em Poincaré e Lorentz.

data 03.02.2020

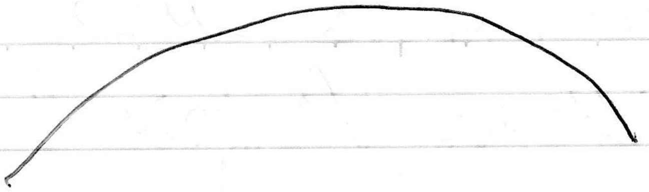

$\pi\pi > \pi$

(Mus)

fator de conversão

$\pi \times c = \pi\pi$ (MUS)

(constante curvilínea munichiana)

obs: Vejá Constante Cosmológica
↓
expansão do universo (acelerando).

Veja:

GAUSS (item 110) e 111 (osculate)

179° 59' 59.32"

---

113) Mathematics: 20200208

Michael Atiyah →

data 08.02.2020
S T Q Q S [✓ S] D

22 April 1929 > 89 anos.
11 January 2019 (+3 meses para 90 anos)

H. F. Baker

03 July 1866 > 89 anos (+3 meses para 90 anos)
17 March 1956

Meu nono também 14 Jan 1997 (faleceu)
(4 meses para 90 anos) (14)
Maio 1907. ↳ correto.

NONA - 20 Nov 1991.

Michael Atiyah →

A twisted cubic curve → subject of Atiyah's first paper.

Michael Atiyah and Friedrich Hirzebruch, the creators of topological K-theory.

September 2018, at the 2018 Heidelberg Laureate Forum, he claimed a simple proof of the Riemann hypothesis, one of the most important and challenging problems in mathematics.

His claims have been met with skepticism by the mathematical community.

data 08.02.2020

S T Q Q S [✓ S] D

Twisted cubic is a smooth, rational curve $\mathbb{C}$ of degree three in projective 3-space. $\mathbb{P}^3$.

It is a fundamental example of a skew curve.

It is essentially unique, up to projective transformation.

Obs: Atiyah, Michael F (1988 a)
Collected works
vol. 1 Early papers: general paper.

ISBN 978-0-19-853275-0

Source: (Book)
Data_Novembro_30_2020/
Books_30_2020002[redacted]/
The Riemann Hypothesis
Michael Atiyah - Collected Works vol. 1. Early Papers. 1988.djvu
↳ A note on the tangents

of a twisted cubic

Pag. 5(27)

obs: Usou como referências:

a) Baker, H. F. Principles of geometry
Cambridge, 1940), Vol. 4, p. 52

Important:

Principles of Geometry Baker
6 Volumes

Source: Data_Novonio_30_2020/
Books_30_2020/
The Riemann Hypothesis/
Principles of Geometry
Baker_6_Volumes//

Pág. 52(72)

the chords of a cubic curve in the original threefold space.
Veronese's surface

até Pág. 55(75) Compare the theorem
(55) of Kronecker
and Castelnuovo
quoted in Vol. III, p. 223.

data 08.02.2020

Volume III - Chapter IV

the Steiner quartic surface

Pág. 221 (245)

As the equation ~~which~~ ~~is dual~~ of the ~~tangent~~ plane of the cubic surface is

$$x_0^{-2}x + y_0^{-2}y + \dots = 0,$$

the tangential equation of this cubic surface is

$$l^{\frac{1}{2}} + m^{\frac{1}{2}} + n^{\frac{1}{2}} + p^{\frac{1}{2}} = 0$$

and the equation of the quartic surface, dual to the cubic surface, is

$$X^{\frac{1}{2}} + Y^{\frac{1}{2}} + Z^{\frac{1}{2}} + t^{\frac{1}{2}} = 0$$

this meets each of the planes

$x = 0$, $y = 0$ etc,

in a conic at every point of which it touches the plane, any two of these four conics having a point of contact, on

(osculate)

data 08.02.2020

the line of intersection of their planes.

...

Pág. 222 (246)

Ex. 3 The inflexional curves of the Steiner surface are rational quartic curves, having the three double lines of the surface as chords...

Pág. 223 (247)

Ex. 5 ... known as Veronese's surface.

Ex. 4 ... the inflexional curves of the surface expressed by the equation

$$x^m + y^m + z^m + t^m = 0$$

satisfy the equation

$$(ax^m)^{\frac{1}{2}} + (by^m)^{\frac{1}{2}} + (cz^m)^{\frac{1}{2}} = 0$$

where a, b, c are any constants whose sum is zero.

Pág. 247 (223)

VERONESE'S SURFACE,

data 08.02.2020

see Cayley, Papers, VI, p. 198
(1868), and Bertini,
Geometria projettiva d.
iperspazi, 1907, Chapter XIV, XV.)

See also Castelnuovo,
Atti d. r. Acc. d. Lincei, III,
1894, p. 22,
who proves, after Kronecker,
that an irreducible algebraic
surface which is met by
∞^2 planes in reducible
curves, if not a
ruled surface,
is a Steiner quartic
surface.

—

Twisted Cubic

is most easily given
parametrically as the
image of the map

$$\nu : P^1 \to P^3$$

which assigns to the homogeneous
coordinates $[S:T]$ the value

$$\nu : [S:T] \to [S^3 : S^2T : ST^2 : T^3].$$

08.02.2020

In one coordinate patch of projective space, the map is simply the moment curve

$\nu: x \Rightarrow (x, x^2, x^3)$

that is, it is closure by a single point at infinity of the affine curve (x, x^2, x^3).

Homogeneous coordinates
or
projective coordinates

Möbius in 1827 work

Der barycentrische Calcül.

→ points at infinity can be represented using ~~ffinite~~ coordinates.
(finite)

Formulas involving homogeneous coordinates are often simpler and more symmetric than their Cartesian counterparts.

in general, projective transformations to be easily represented by a matrix.

08 02 2020
1

Rational Bézier curve –
polynomial curve defined in homogeneous coordinates (blue) and its projection on plane – rational curve (red)

<u>Obs</u>: Uso do <u>LATEX</u>.

Book (source)

ARQ-Important/latex-books/Books I important/
Guide to latex.pdf
Helmut Kopka

Pág. 299 (312)

Curved lines

\qbezier[num] $(x_1, y_1)(x_2, y_2)$
(x_3, y_3)

<u>Ex</u>:

```
\begin{picture}(40,20)
\qbezier(0,0)(20,20)(40,10)
\end{picture}
```

data 08.02.2020
x

real numbers → real projective spaces

complex numbers → complex projective space.

obs: the complex projective line uses two homogeneous complex coordinates and is known by the Riemann Sphere.

114) Research:

a) On Michael Atiyah and the Riemann Hypothesis

https://blogs.ams.org/blogonmathblogs/2018/10/01/on-michael-atiyah-and-the-riemann-hypothesis/

b) Michael Atiyah claims proof of Riemann Hypothesis

https://aperiodical.com/2018/09/michael-atiyah-claims-proof-of-riemann-hypothesis/

data 08.02.2020

c) www.newsroom.hlf-foundation.org/blog.html

d) Michael Atiyah claims proof of Riemann Hypothesis

Sep 22, 2018

https://aperiodical.com/2018/09/michael-atiyah-claims-proof-of-riemann-hypothesis/

→ -final-thoughts/ (Sep 28, 2018)

115) Books Important 20200220

DATA_NOVOVIVO_37_2020/
Books_37_2020/
Importantíssimo/
Nouvelle-Mathematiques/

a) 11nouvelleannal18pari.pdf

Formules Fondamentales de L'Analyse Sphérique

Pág. 5(11)

↘ b) 13nouvelleannal05pari.pdf

Méthode Élémentaire pour

date 20.02.2020
S T Q ☒ S S D

Calculer Le rapport de la Circonférence au diamètre

Pág. 6(11)

c) s4nouvelleannal13pari.pdf

Surfaces engendrées par le déplacement d'une courbe plane avec cone circonscrit le long de la courbe

Pág. 1(9)

116) 20200223 Source: Data_novoVIVO_38_2020/Books_38_2020/Measurement/Lockhart/janet_st_9811210144.pdf

a) Un Eventful Journey to Unification of All the Fundamental Forces
Delbourgo

Pág. 67(84)-

6.1. Anticommuting coordinates.

There are two types of fields in physics: those connected with integer spin particles, called bosons, and those connected with half-integer spin particles called fermions.

b) The Mathematical Principles of Scale Relativity Physics
Mazilu.

data 23.07.2020

S T Q Q S S [✓ D]

Chapter 7. The Physical Point of View in the Theory of Surfaces.

Pag. 110 (115).

We start from the idea that a surface should play a double role in physics.

matter from space of the

space from matter

for there are altogether two different situations. This statement cannot be made understandable without

→ **Second role**:

that of locus of an ensemble of positions in the geometrical acceptance.

Source: Data_novoviva_38_2020/Books_38_2020/
The Mathematical Principles of
Scale Relativity Physics.../
sanet_st_03617349345.pdf.

a) Solid Mensuration with proofs
Kern

23.02.2020

Source: ARQ-IMPORTANT/Cinderella.de/*

Pappus/

1) 2015.205959.Solid-Mensuration.pdf

2) Sommerville D - An introduction ... pdf

✡ 3) 10-piscoran-final.pdf

Pappu's harmonic theorem in the Einstein relativistic velocity model of hyperbolic geometry.

Veja:

Cinderella.de/

{ Pappus_20200223.pdf

Pappus-~~202002~~ Spherical_20200223.pdf

Source:

Cinderella.de/the Cinderella.2.Manual.pdf

Pag. 43(58)-49(64)

—

4) Analytic Hyperbolic Geometry in N Dimensions.
Ungar

date 23.02.2020
S T Q Q S S D

Pág. 590(611)

[87] Pappu's harmonic theorem in the Einstein relativistic.

3 cont.) Pappu's harmonic theorem in the Einstein relativistic

www.cs.ubbcluj.ro/~studia-m/2011-1/10-piscoran-final.pdf

mathworld.wolfram.com/PappussCentroidTheorem.html

theoremoftheday.org/GeometryandTrigonometry/Guldin/TotDPappusGuldin.pdf

117) 20200226

a) Sullivan Conjecture

https://ocw.mit.edu/faculty/haynes-miller/

b) $2\pi = \tau$ (Tau)

the radio of the great pyramids perimeter to its height approximates 2π or what is know as TAU.

data 26.02.2020

c) Data_Novovivo_40_2020/ Books_48_2020/ Atoms, Molecules and Optical Physics.../

c1) Atoms, Molecules and Optical Physics 1.pdf

Pág. 90(122)-

2.2 ~~Schrö~~ndige

Schrödinger Equation

2.2.3 Time Dependent Schrödinger Equation

... which is a 2nd order partial differential equation (PDE) in space and time derived from the MAXWELL equations.

c2) Atoms, Molecules and Optical Physics 2.pdf

Pág. 27(58)

Snellius diffraction

Pág. 28(59)

ABCD matrices.

Pág. 31(62) - Fig 1.16 ⇒

26.02.2020

De-focusing of a Gaussian beam.

Look:

Fig 1.15 – Pág. 30(61)

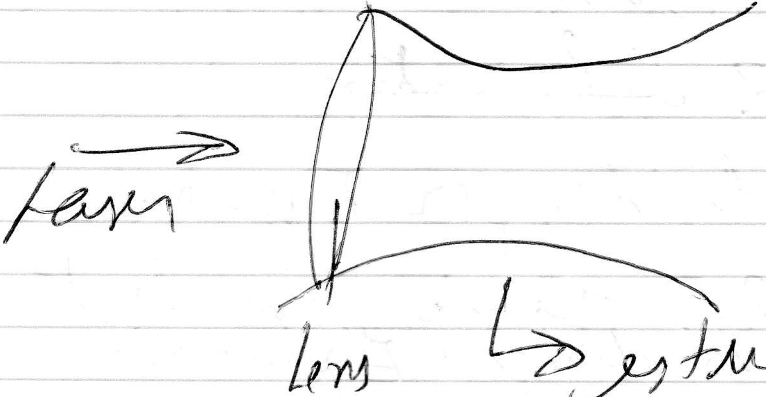

↳ estude esta curvatura com a curva que distorce a gravidade.

Obs:

Fig 1.16

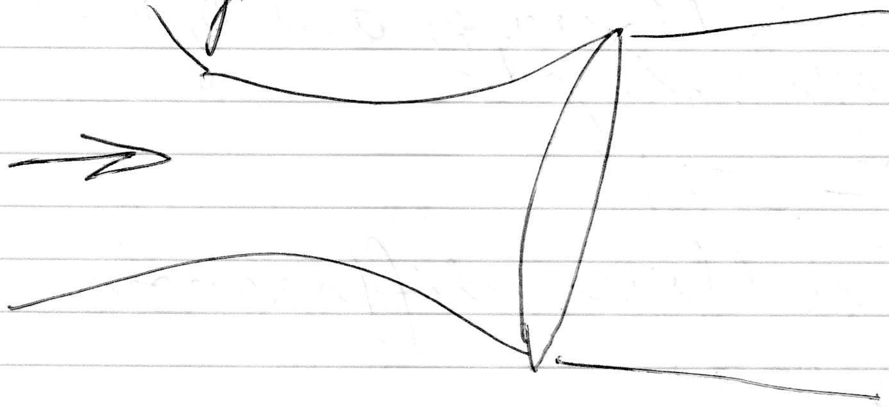

r3) Books_40_2020/
Ordinary Differential
Equations.../
saretist-10167.pdf

26.02.2020

Pág. 61(71) Theorem 17

Jordan curve J

U will thus contain (from the Poincaré-Bendixson theorem) the limite cycle $\Gamma(\varepsilon)$.

Look: Fig II.1.7.

c4) Book_40_2020/Handbook_of_Homotopy_Theory/

Handbook of Homotopy theory.pdf

Pág. 315(324)

8.7.4 Persistence images

A standard choice is that of a spherical symmetric Gaussian with mean μ and a fixed variance σ.

Pág. 316(325)

Pág. 320(329)

Example 8.8.6-

For examples, give a RIEMANNIAN manifold, one can

26.02.2020

compute Gaussian curvature

↳ comparar com a curvatura do espaço (gravidade).

at each point and filter by that quantity.

Pág. 326(335) - Gaussian random

Pág. 327(336) - Gaussian field.

Pág. 328(337) →

[9] Bobrowski and Borman

Euler integration of Gaussian random fields and persistent homology.

Pág. 970(979) →

[69] Levine

Motivic Gauss-Bonnet formulas

(24) Books_40_2020/
Geometry the Line and the Circle

26.02.2020

Robert Bix – Conics and Cubics
a concrete introduction to
algebraic curves... pdf

Pág. 131 (139)

Certain cubics are now called "elliptic curves" because of their connection with elliptic integrals.

This connection was discovered by GAUSS, NIELS ABEL and CARL JACOBI in the 1820s!

Their results were clarified and extended by RIEMANN in the 1850s.

Weierstrass in 1863, and Henri Poincaré in 1901.

resumo: $g(t) = 4t^3 + ct + d$

polynomial of degree 3 without repeated roots.

Pág. 328 (336) Gauss's Lemma

Pág. 130 (138) → d'Alembert's lemma

Cantor

Dedekind

The idea of a complex curve

→

26.02.2020
X

an algebraic curve whose
coefficients and variables
are complex numbers.

Newton → "imaginary points"

In the 1820s,
Jean Poncelet and
Michel Chasles
argued for using imaginary
points systematically in
Synthetic projective geometry.

Plücker (1830) →
homogeneous coordinates,
he introduced made it possible
to consider points with
complex coordinates.

118) 20200228

Números de Perrin

pt.wikipedia.org/wiki/Números_de_perrin

data 28.02.2020
S T Q Q [✓ S] S D

Número Plástico ψ

é uma constante matemática.
É a solução real inequívoca da equação cúbica
(cúbica)

$$\underbrace{x^3 - x - 1 = 0}$$

valor aproximado
1,324718

função geratriz

$$G(P(n); x) = \frac{3 - x^2}{1 - x^2 - x^3}$$

Source:

wikipedia.org

mathworld.wolfram.com/PerrinSequence.html

www.universite-paris-saclay.fr/~perrin

Daniel.Perrin@math.u-psud.fr

date 28 02 2020

www.imo.universite-paris-saclay.fr/~perrin/Livre_de_geometrie_projective.html

Géométrie Projective

Pappus

www.math2.fr/geometrie-projective/

math2.fr/2019/10/06/geometrie-projective/

/2019/12/18/geometrie-projective-exemple-de-calculs2/

others :

Data_NOVOVIVO_41_2020/
Books_41_2020/
Pappus/
Enseignement (hors CAPES)/

a) polysite.pdf

↳ Introduction générale et rappels

b) Pappus/TER/constructions.pdf

Constructions à la règle et au compas.

28.02.2020

c) Pappus/SEVRES/polyedres_pdf
Les polyèdres - Perrin

119) 20200302

Books:

a) Data_NOVOVIVO_44_2020/Books_44_2020/
Math_Books/Multivariable_Calculus/
Multivariable_Calculus_(formatted_
for_printing).pdf

Pág. 38(50)

2.6 The geometry of space
curves: Frenet vectors

b) .../Books_44_2020/Math_Books/
Zakon-analysis II-a4-one.pdf

Pág. 97(109) → Chapter 7
Volume and Measure

c) .../Math_Books/
Nicholson-OpenLAWA-2019A.pdf

Pág. 415(437)

8. Orthogonality

8.1 Orthogonal Complements and
Projections.

data 02.03.2020

[✓ S] T Q Q S S D

d) PC_Oculto/Math_Sphere_Equations/

d1) geometriaa5.pdf

Pág. 231(244) - Chapter 6

Círculos e Esferas

d2) geometriaanaliticaevetorial-SGD.pdf

Pág. 149(157) - Chapter 6

Círculos e Esferas.

e) ... Math_Books/Spherical Geometry and Its Applications/

Sanet.st_0367180464.pdf

Pág. 252(267)

6.1 Riemannian Metrics

120) 20200304

A) Lecture Notes on General Relativity

Matthias Blau

http://www.blau.itp.unibe.ch/GRLecturenotes.html

04.03.2020

Source: www.blau.itp.unibe.ch/
newlectures GR.pdf (*)

Pág. 207 8. Curvature I:
The Riemann
Curvature Tensor

Grossmann soon found that this problem had indeed addressed and solved in the mathematics literature, in particular by RIEMANN (generalising work of GAUSS on curved (surfaces)), RICCI-CURBASTRO and LEVI-CIVITA.

It was shown by them that there are indeed non-trivial tensors that can be constructed from (ordinary) derivates of the metric.

These can then be used to write down covariant differential equations for the metric. 18

18 - Marcel Grossmann and his contribution to the general theory of relativity
J. Sauer

Source: Data_Novovivo_42_2020/
Books_46_2020/
Importantissimo_Gauss/
a) newlectures GR.pdf (*)
b) 1312.4068.pdf (18)

data 04.03.2020

S T [✓ Q] Q S S D

General Relativity
Tilman Sauer

https://cds.cern.ch/record/735319/files/0405066.pdf

Source: Books_46_2020/Importantissimo_Gauss/0405066.pdf

Others:

I) Einstein's Pathway to General Relativity

Source: https://www.pitt.edu/~jdnorton/teaching/HPS_0410/chapters/general_relativity_pathway/index.html

II) Einstein's Zurich Notebook

Source: www.pitt.edu/~jdnorton/Goodies/Zurich_Notebook/

III) Shadow of the Relativity Revolution

data 04.03.2020
S T [✓ Q] Q S S D

Source: https://www.mpiwg-berlin.mpg.de/Preprints/P271.pdf

—

A) Cont.

* newlectures GR.pdf

Pág. 214

8.4 Tidal Force: Influence of Curvature on Particle Trajectories

Pág. 215

Remarks:

1. this shows very clearly that curvature, as captured by the Riemann curvature tensor, leads to non-Euclidean in which e.g. the parallel axiom is not necessarily satisfied.

—

other:

Gauss Curvature →

Pág. 260 11.3.

data 04.03.2020 8

11.3 Curvature of Surfaces: Euler, Gauss (-Bonnet) and Liouville

... We can generalise the example of the curvature of the 2-sphere, discussed in section 8.6, somewhat, in this way connecting our considerations with the classical realm of the differential geometry of surfaces, in particular with the

Gauss curvature,

the Euler characteristic,

the Gauss-Bonnet theorem and

the Liouville Equation.

Pág. 261 – Gauss curvature K.

It can also be interpreted as saying that in 2 dimensions the Einstein tensor (8.97)

Pág. 226 ← → book.

B) There's no such things as the Force of Gravity.

04.03.2020

Source: Books_46_2020/
Loop Quantum Gravity for Everyone/
Quantum Space.pdf

Pág. 27(48).

①) Space-Time, Yang-Mills Gravity, and Dynamics of Cosmic Expansion.

Source: Books_46_2020/
Space-Time.../
Space-Time....pdf

Pág. XXXI (32)

" Einstein's approach involving the geometrization of physical fields in curved space-time appears to be successful only in the formulation of classical gravity with the conservation of the energy-momentum pseudo-tensor, rather than the true energy-momentum tensor.

... Finsler geometry,

where the fundamental metric tensors depend on both positions and

04.03.2020

velocity (i.e, the differentials of coordinates),

rather than the Riemannian geometry.

D) Free-surface flow
environmental fluid mechanics ... pdf

Pág. 483(497)
6.6.2 Cauchy-Riemann Equations

Pág. 486(500)
Figure 6.30
Conformal mapping of flow net.

121) 20200403
Books:

A) DATA-NOVOVIW-63-2020/
Books-63-2020/
Approximate Calculation of Integrals/
Approximate Calculation of Integrals... pdf

data 03.04.2020

S T Q Q S S D

• B) Books_63_2020/
From Geometry to Topology.../
H. Graham Flegg - From Geometry to Topology... pdf

Pág. 154 (104) - Metric Space

Fig. 16.3

radial projection...

The distance between any two points exterior to the sphere is shrunk by the function f, and the distance between any two points on the surface S remains unchanged.

C) Books_63_2020/
Laminar Boundary Layers/
sanet.st_12184.pdf

Pág. 457 (490)

20. Parabolic flow over a flate plate

Pág. 458 (491) →

Fig. VIII. 7. Family of parabolic translates as streamlines of a curved external flow over a flate plate.

data 03.04.2020

Equations:

$$y = \lambda + x\left(B - \frac{1}{2} c x\right), \quad (156)$$

$$B = \cot \theta_0, \quad (157)$$

where

and B and c are constant for a given system of translates.

D) Books_63_2020/
Voronoi Diagrams and Delaunay Triangulations/
Jean Gallier... Aspects of Convex Geometry - Polyhedra ... pdf

Pág. 100

Fig 4.11. The polar dual a polygon

Obs: Pág. 98

Fig. 4.10: The polar, $a^{\dagger}$, of a point, a, outside the sphere S^{n-1}.

Obs: Pág. 96 4.3 Polarity and Duality.

data 03.04.2020
8

E) Books_63_2020/

Advanced Differential Quadrature Methods/

Advanced Differential Quadrature Methods .. pdf

Pág. 98(122)

3.3 Geometric transformation

Pág. 99

Fig. 3.5: Points for quintic transformation of a curvilinear triangle.

Pág. 103(126)

102(125) –

3.4.4 Symmetry Consideration

All elliptic plates on Pasternak foundation are assumed subjected to uniformly distributed load.

Pág. 103(127)

Fig. 3.6 A quadrant of an elliptic plate.

data 03 04 2024

Pág. 104(128)

Table 3.2:

Central deflection and central moment of thin circular plates on a Pasternak foundation with clamped or simply supported edge.

Pág. 105(129)

Fig 3.7: Deflections of the circular plate

Fig 3.8: Radial bending moments of the circular plate.

Pág. 273(297)

Chapter 8

Mathematical Compendium

8.1. Gauss elimination

Pág. 281(305)

8.5 Complex Analysis

8.5.1 Complex Variable

data 03.04.2020
S T Q Q [✓ S] S D

122) Books_2020_10/
Books_53_2020/
Theoretical Fluid Dynamics/
Fluid Dynamics Theoretical and Computational Approach...
pdf

Pag. 03 (30)

Fig. 1.1

Inertial Cartesian or curvilinear reference frame

Date: 20200318

123) DATA_NOVOVIVO_63_2020/
Books_63_2020/
Introduction to Quantum Mechanics/
David J. Griffiths_
Mecânica Quântica..pdf

Pag. 95 (100) –

Problema 3.23

O Hamiltoniano para determinado sistema de dois níveis é

data 03.04.2020

$$\hat{H} = \epsilon(|1\rangle\langle 1| - |2\rangle\langle 2| + |1\rangle\langle 2| + |2\rangle\langle 1|),$$

em que $|1\rangle$, $|2\rangle$ é uma base ortonormal e ϵ é um número com as dimensões da energia.

Problema 3.25

Polinômios de Legendre — use procedimento de Gram-Schmidt (Problema A.4) para ~~ot~~ ortonormalizar as funções.

Problema 3.26

Um operador anti-hermitiano é igual a menos o seu conjugado hermitiano:

$$\hat{Q}^{\dagger} = -\hat{Q} \qquad [3.95]$$

124) Books_63_2020/
Introduction to Quantum Mechanics/
David J. Griffiths - Introduction

03 04 2020

to Electrodynamics(2013, Pearson).
pdf

Pág. 117 (122)

1.4 Laplace's Equation in three Dimensions

$$V(r) = \frac{1}{4\pi R^2} \oint_{sphere} V da$$

Figure 3

Obs: 3 For a proof does not rely on Coulomb's law (only on Laplace's equation), see Prob. 37

125) 20200404

Data_NOVOVIVO_64_2020/Books_64_2020/Mathematics for Physics A Guide Tour for Graduate Students/Michael Stone ... Mathematics for physics ... pdf

Pág. 41 (57)

Problem 1.6: The catenary revisited

data 04 · 04 2020

Pág. 42(57)

Figure 1.15 Weighted line

Figure 1.16 the Poincaré disc of Exercise 1.7.

Problem 1.7.

Another model for Lobachevski geometry (see Exercise 1.3) is the Poincaré disc (Figure 1.16)

This space consists of the interior of the unit disc

$$D^2 = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$$

equipped with the Riemann metric

$$ds^2 = \frac{dx^2 + dy^2}{(1 - x^2 \underset{(-)}{\cancel{+}} y^2)^2}$$

126) DATA_NOVOVIVO_64_2020/
Books_64_2020/
Geometry/
The Ricci Flow in Riemannian Geometry_A Complete Proof of the Differentiable 1-4 -

data 04 04 2020

Pinching Sphere theorem

Chapter 4

Evolution of the Curvature

Source: Pag. 63 (82)

127) 20200406

A) Data-NOVOVIVO_65/2020/
Books_65_2020/
Higher Algebra Classical/
Higher Algebra Classical by Sadhan...pdf

Pag. 4 → 2. Complex Numbers

B) Books_65_2020/
Descriptive Geometry/
Geometric and Engineering Drawing
by K. Morling...pdf

Pag. 5 (22)

Diagonal Scales

Pag. 6 (23) Figure 1.4

Pag. 7 (24) Figure 1.5
Figure 1.6

Pag. 8 (25) Figure 1.7

Pag. 11 (28) 2 → The Construction of

data 06 04 2020

Geometric Figures from given data

Obs: Usar este livro.

Pág. 44 (61)

Figure 4.1 Parts of the circle

Figure 4.2

The construct the circumference of a circle, given the diametre

(important)

Figure 4.3

To construct the diameter of a circle, given the circumference

Figure 4.4

To find the centre of any circle

Figure 4.5

To construct a circle to pass through three given points

Figure 4.6

To construct the inscribed circle

of any regular polygon
(in this case, a triangle)

Figure 4.7

To construct the circumscribed circle of any regular polygon (triangle)

Figure 4.8

To construct the escribed circle to any regular polygon (triangle)

Figure 4.9

To construct a circle that passes through a fixed point A and touches a line at a given point B.

... Figure 4.20

Pag. 55(72).

(i) Books_65_2020/
Descriptive Geometry/
A. N. Whitehead - The Axioms
of Descriptive Geometry....pdf
Pag. 4(13)

Peano uses the following useful notation which is an extension of his notation for segments and prolongations.

data 06.04.2020

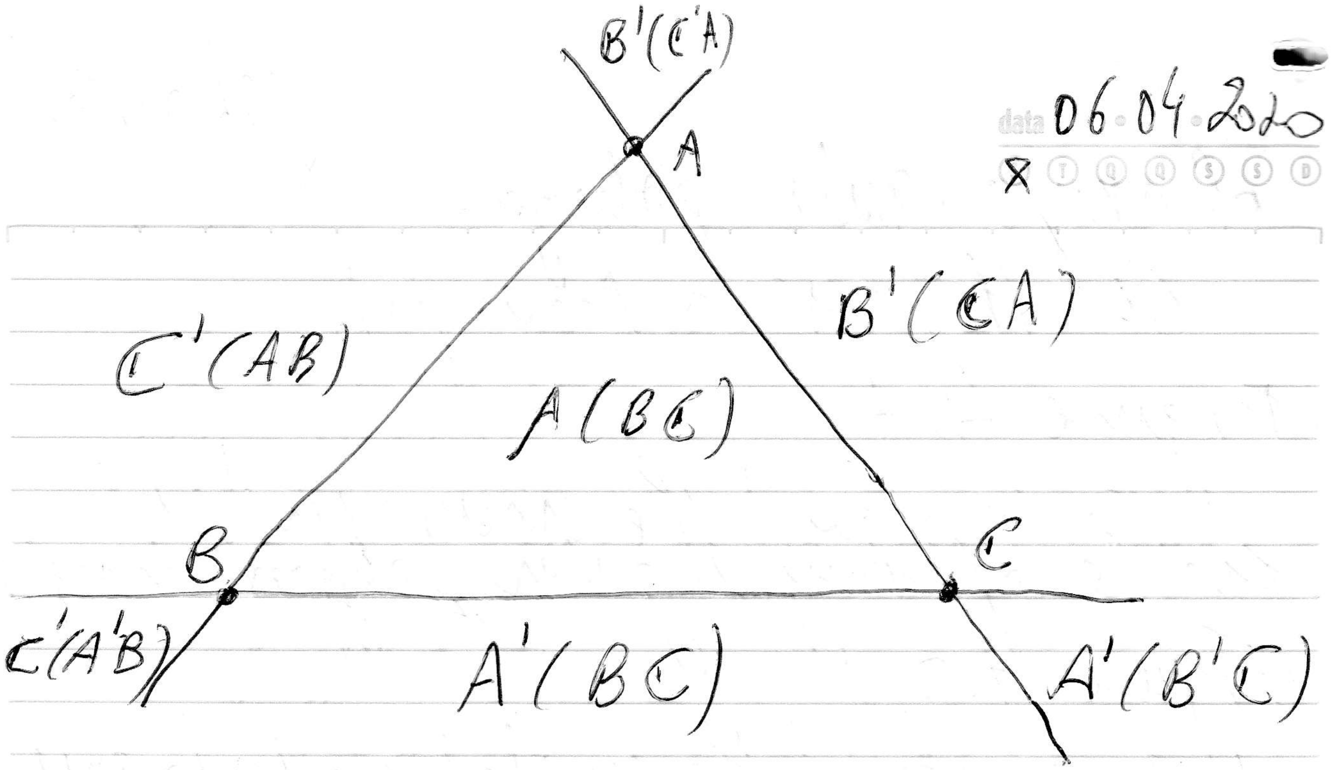

Seven regions into which a plane is divided by three lines are as in the figure

---

Pag. 18(27) Sheaves of Planes

Pag. 18(27) → 19

The two theorems of the present and next articles are the central theorems of the whole of Ideal Points

Pag. 19 Coplanarity of Lines

128) Graphic science:

Thomas Ewing French

Source:

https://archive.org/details/graphicscienceen00fren

06.04.2020

Copy: Books_65_2020/
Descriptive Geometry/
Graphic science engineering drawing, descriptive geometry, graphical solutions by French....pdf

129) 20200408

A) Data_NOVOVIVO_66_2020/
Books_66_2020/
Basic Electrical and Electronics Engineering/
S.K. Bhattacharya...
Basic Electrical...(2010).pdf

Pág. 234 (245)

Figure 4-8

Continuity of the electron wave around the nucleus requiring in an integral number of wavelengths into any circular path

Obs: Energy band theory

B) Books_66_2020/
S.K. Bhattacharya... 104...pdf

Pág. 2.18 (83)

2.6. Maxwell's mesh current method

data 08.04.2020
S T [✓ Q] Q S S D

C) Books_66_2020/
The Shape of Space/
sanet.st_1138062278.pdf

Pág. 125(144) Sphere

Pág. 126(145) lesser circles

✡ Pág. 133(152)

Figure 9.6:

A quick proof that the sum of the angles in a Euclidean triangle is π:

$\alpha' + \beta + \gamma' = \pi$,

$\alpha = \alpha'$ and

$\gamma = \gamma'$; therefore

$\alpha + \beta + \gamma = \pi$

D) Books_66_2020/
Pleasures of Pi.../

Pleasures of Pi....pdf

130) 20200409

A) Books_66_2020/
Engineering Mathematics by
P. Sivaramakrishna Das/

Jandaia

data 09.04.2020
S T Q [✓ Q] S S D

Babu Ram Vijay Gupta - Engineering Mathematics III .1 pdf

Pág. 1.49/(66)

Theorem 1.21 (Cauchy's Residue Theorem)

If f(z) is analytic within and on a closed contour C except at finitely many poles lying in C, then

$$\int_C f(z)\,dz = 2\pi i \sum R,$$

where $\sum R$ denotes the sum of residues of f(z) at the poles within C.

B) Books_66_2020/
Engineering Mathematics by P. Sivaramakrishna Das / Babu Ram Engineering Mathematics Pearson (2009).pdf

Pág. 102

3 Curvature

- obj: Ver book todo.

C) Engineering Mathematics....pdf/
sanet.st_8131784703.pdf

D) Books_66_2020/
Electrical_Machines, 2edition/
Smarajit Ghosh - Electrical Machines-

data 09.04.2020

Pearson (2013).pdf

Pag. 19 (44)

Figure 30 Phasor Diagram

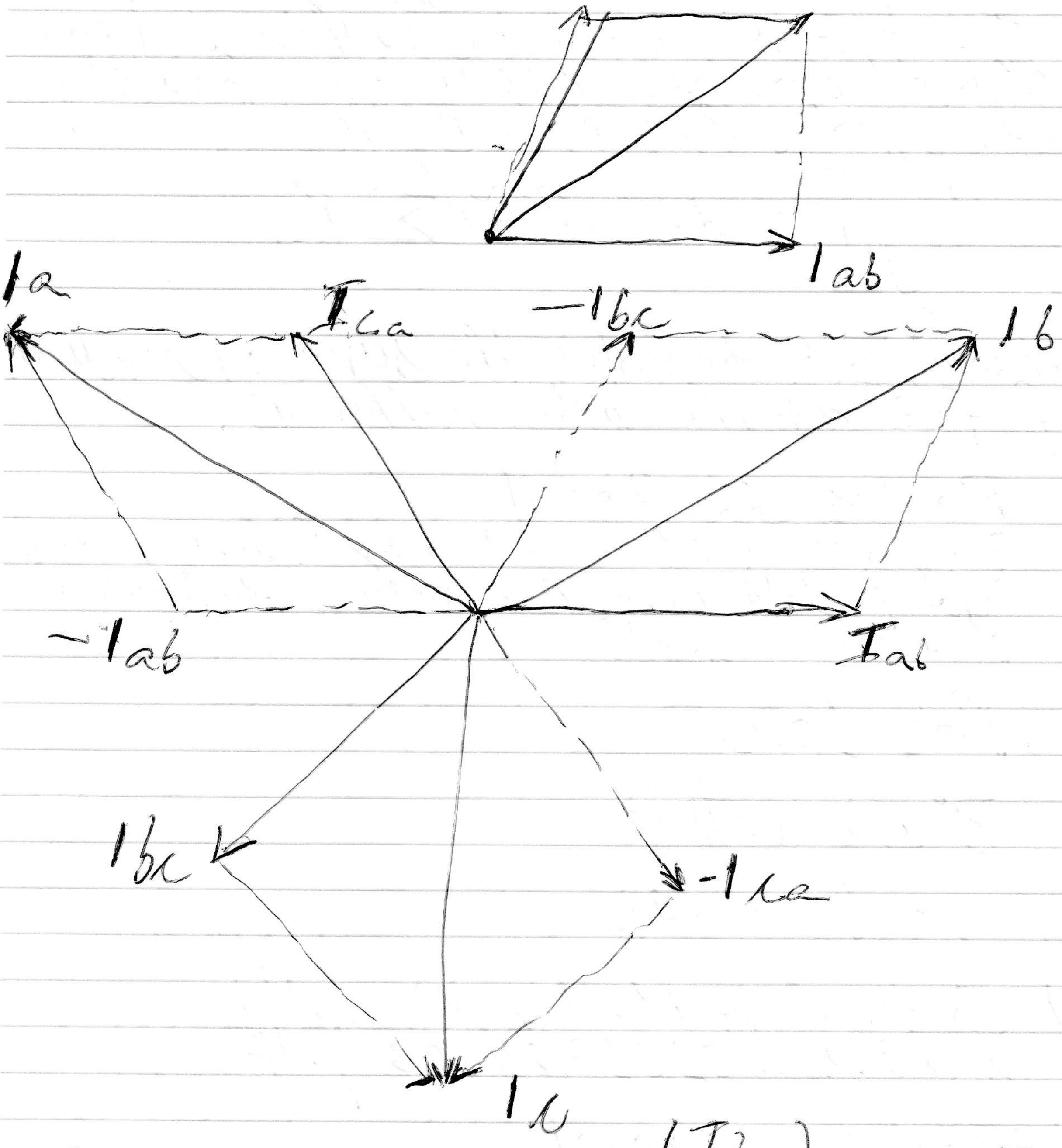
I_{ab}
I_a
I_{ca}
$-I_{bc}$
I_b
$-I_{ab}$
I_{ab}
I_{bc}
$-I_{ca}$
I_c

$$P = 3V_pI_p\cos\phi = 3V_L\left(\frac{I_L}{\sqrt{3}}\right)\cos\beta = \sqrt{3}\,V_L\,I_L\cos\beta$$

↓

cos

data 09.04.2020

S T Q Q S S D

131) Books_66_2020/
Practical Geometry and Engineering Graphics/1

— Practical Geometry and Engineering Graphics1.pdf

— sanet.st_0216894506.pdf

Obs: Mesmo livro.

Pág. 159(165)

Projections of a Point P situated in the 1st quadrant

Pág. 161(167)

Projections & Traces of a line AB in the 1st Quadrant

Obs: Livro muito bom, bem descrito e com demonstrações gráficas, fotos...

Pág. 80(86)

100. Applications of the Integral Curve of the Reciprocal Curve

Last page: 81(87) — Graphics.

Obs: Cálculos em polegadas.
Adequar para sistema métrico decimal

data 10.04.2020

S T Q Q [✓ S] S D

132) 20200410

Degenerate Monge-Ampere Equations over Projective Manifolds

Source:

Data_NOVOVIVO_67_2020/
Books_67_2020/
Math_Books/
zhang-phd-math-2006.pdf

Source internet:

www.maths.usyd.edu.au/u/zhangou/paper/zhang-phd-math-2006.pdf

133) Books_67_2020/
Theoretical Chemistry for Electronic Excited States/
theoretical chemistry for electronic excited states (2018_).pdf

Pag. 08(22)

Figure 1.5 a) Avoided crossing

b) Canonical intersection, true crossing.

10.04.2020

26. http://www.worldscientific.com/doi/suppl/10.1142/7803/suppl_file/7803_chap01.pdf

→ For an introductory article on conical intersections in photochemistry see Robb[22], which is available as a free download.[26]

Pág. 15(29)

Figure 1.10

Pág. 16(30)

Figure 1.11

Pág. 19(33)

Figure 1.13

(134) 20200411

A) Books_67_2020/
Solving problems in geometry/
Projective geometry an introduction by Rey Casse .. pdf

Projective Planes
Pág. 29(42)

data 11.04.2020

A) 5 Coordinatising a projective plane

Pag. 99 (112)

6 Non-Desarguesian Planes

Pag. 110 (123)

Pag. 112 (125)

Example 6.4

Prove that if a collineation σ of a projective plane π is such that

1. some line l is fixed pointwise,
2. two points $P, Q \in l$ are fixed,

then $\sigma = 1$.

B) Books_67_2020/
Mathematics of Shape Description/
Pijush K. Ghosh_Mathematics of Shape Description...pdf

Pag. 77 (96)

3. Algebraic Structures for Shape Description

5. Arithmetics of Geometric Shape
Pag. 165 (184)

data 11.04.2020

5.3 Boundary Representation by Support Functions for Morphological Operations

5.3.1 The Support Function Representation

Pag. 169(188)

Convex polytope A in E^d is defined in terms of the supporting function and the supporting hyperplane of the polytope.

5.3.2 The Support Function is a Signed Distance

Pag. 170(189)

Figure 5.1

Figure 5.2

Figure 5.3 a) two-dimensional polygon
b) three-dimensional polyhedron

① Books_67_2020/
Visual Quantum Mechanics/
Bernd Thaller- Advanced Visual Quantum Mechanics (2004...).pdf

Pag. 38(49) Figure 1.9

Possible trajectories of classical particles on a sphere. For all trajectories, the angular-momentum vector has $|L| = \sqrt{l(l+1)}$, with $l=6$ and $L_z=6$.

data 13.04.2020

135) A) Books_68_2020/Nuclear Physics.../
Sanet.st-NuclearPhysics Pearson
2017.pdf

Pag. 116(129)

4.2 Schrödinger Equations for a System of two spinless particles in spherical polar coordinates

Fig 4.1 Relation between the spherical polar and Cartesian coordinates.

B) Books_68_2020/
Orbital Mechanics for Engineering Students/
Orbital Mechanics for Engineering Students... Curtis(z-lib.org).pdf
(2020)

Pag. 393(400)

Sphere of Influence

Fig 8.5 Decrease of gravitational force with distance from a planet's surface.

Fig. 8.6 Relative position and gravitational force vectors among the three bodies.

Pag. 292(299) → Fig 6.4

data 13.04.2020

Fig 6.4 Hohmann transfers between coaxial elliptical orbits.

C) Books_68_2020/
Einstein in Matrix Form Exact Derivation of the Theory of Special and General Relativity..pdf/
Einstein in Matrix..pdf

Pag. 34(47)

1.8.2 Lorentz Transformation of the Maxwell's Equations

Look (1.146) $P(v) = \left(\begin{array}{c|c} & \\ \hline & \end{array}\right)$

136) 20200414

A) Books_68_2020/
Problèmes de Géométrie Analytique Kletenik pdf/
cours 3_et_4_niv2_rev.pdf

other source: Books_68_2020/Important

Pag. 245 Espace physique et espace vectoriel

Look: Pag. 255 Réponse ex. 1-11
veja o gráfico (desenho geométrico)

obs: Use este livro.

data 14 04 2020

B) Books_68_2020/
Mathematics Form and Function/
engkeny43523-Janet.St.pdf

Pág. 107 (118)

7. Vector Geometry

3. Geometry

Pág. 61 (72)

Pág. 64 (75)

Figure 1 Do the circles meet?

Pág. 109 (120)

8. Analytic Geometry

Pág. 111 (122)

Figure 1

$P(\cos\theta, \sin\theta)$

Obs: Use este livro.

C) Books_68_2020/
Recent Advances on Elliptic and Parabolic Issues/
Michel Chipot.m.pdf

Obs: Ver este livro _

data 15.04.2020
S T [✓ Q] Q S S D

137) Pi

BBP-Type formulas for Mathematical Constants

Bailey

dhbailey@lbl.gov

Source:

www.davidhbailey.com//dhbpapers/bbp-formulas.pdf

others:

www.davidhbailey.com/dhbpapers/pi-formulas.pdf

—

www.maa.org/sites/default/files/pdf/pubs/amm_supplements/Monthly-Reference_9.pdf

—

mathscholar.org/2019/10/pi-as-the-limit-of-n-sided-circumscribed-and-inscribed- polygons/

—

data 15 04 2020

New paper proves 80-year-old-
approximation conjecture

mathscholar.org/2019/09/
new-paper-proves-80-year-
old-approximation-conjecture/

—

www.davidhbailey.com//dhbpapers/

—

138) 20200416

A) Dancing with a black hole

Source: www.nytimes.com/2020/04/16/
science/black-hole-sagittarius-
a.html

Einstein → gravity to a warping of the geometry of space and time.

In our solar system, astronomers had long known that the planet Mercury behaves like this; its perihelion advances around the sun by the microscopic amount of 43 seconds of arc a century.

Astronomers couldn't explain why. In November 1915, Einstein tried out his new theory on Mercury's motion.

data 16.04.2020

B) A star orbiting the Milky Way's* giant black hole confirms Einstein was right

source: www.sciencenews.org/article/star-orbiting-milky-way-giant-black-hole-confirms-einstein-was-right
by Emily Conover
20200416

⇒ Decades of observations revealed the rotation of the star's elliptical orbit.

1915 - Einstein realized that his newly formulated general theory of relativity explained a weird quirk in the orbit of Mercury.

Star S2.

The researchers previously had used observations of S2 to identify a different effect of general relativity, the reddening of the star's light due to what's called gravitational redshift (SN 7/26/18).

Emily Conover

www.sciencenews.org/article/star-orbiting-black-hole-einstein-gravity-general-relativity

data 16.04.2020

Now, they've determined that the ellipse rotates over time, what's known as Schwarzschild precession. That precession is the result of the warping of spacetime caused by massive objects, according to general relativity.

c) Stephen Wolfram's hypergraph project aims for a fundamental theory of physics

Source: www.sciencenews.org/article/
stephen-wolfram-hypergraph-
project-fundamental-
theory-physics
April 14, 2020
by Tom Siegfried

→ causal invariance.

Thanks to causal invariance Wolfram's hypergraphs reproduce many of the consequences of various physical theories, such as Einstein's special theory of relativity. Traveling rapidly slows down time (as special relativity says) because hypergraph structures corresponding to moving objects make an angle through the hypergraph that extends the distance between updates (or time steps). The speed of light is a

16.04.2020

maximum velocity, as relativity states, because it represents the maximum rate that information can spread through the hypergraph as it updates. And gravity — described by Einstein's general theory of relativity — emerges in the relationships between features in the hypergraph that can be interpreted as matter particles. (Particles would be small sets of linked points that persist as the hypergraph updates, something like "little lumps of space" with special properties.)

supersmall unknown particles, which he calls oligons

Such oligons, subject only to gravity, would now be hanging out in and around galaxies utterly unnoticed — except for their gravitational impact.

Oligons might therefore explain why astronomers infer the existence of vast amounts of invisible "dark matter" in space.

"right rule"

That "right rule" would generate a hypergraph with our universe's precise properties:

data 16.04.2020

- three (apparent) dimensions of space,
- the right cosmic expansion rate,
- the right repertoire of elementary particles with the correct charges and masses, and other features.

Source:

- www.wolframphysics.org/technical-introduction/
- technical-introduction/limiting-behavior-and-emergent-geometry/hyperbolic-space/

D) Wolfram-ModelsForPhysics.pdf
Source:
www.wolframphysics.org/technical-introduction/inc/Wolfram-ModelsForPhysics.pdf

Pag. 105 - Curvature

$$2\pi a^2\left(1-\cos\left(\frac{r}{a}\right)\right) =$$
$$\pi r^2\left(1-\frac{r^2}{12a^2}+\frac{r^4}{360a^4}-\cdots\right)$$

Volume:
$$\frac{\pi^{d/2}}{(d/2)!} r^d\left(1-\frac{r^2}{6(d+2)}R+O(r^4)\right)$$

16.04.2020

where ~~here~~ R is the Ricci scalar* curvature of the space.

Pag. 371 $E = mc^2 + \frac{1}{2} m v^2$

E) Some Relativistic and Gravitational Properties of the Wolfram
April 12, 2020

Source:
www.wolframcloud.com/obj/wolframphysics/Documents/some-relativistic-and-gravitational-properties-of-the-wolfram-model.pdf

Pag. 31 Figure 16:

Two-dimensional spatial hypergraphs with different Ollivier-Ricci scalar curvatures, with the presence of pentagons indicating regions of positive scalar curvature and the presence of heptagons indicating regions of negative scalar curvature.

Figure 19 (Pag. 40)

Geodesic trajectories of test particles (starting on initially parallel paths) through discrete spaces containing a spherical mass distribution. Adapted from S. Wolfram, A New Kind of Science, pg 531.

data 16.04.2020

F) Some Quantum Mechanical Properties of the Wolfram Model

Source:
www.wolframcloud.com/obj/wolframphysics/Documents/some-quantum-mechanical-properties-of-the-wolfram-model.pdf

Pág. 39 Figure 11

The Fano plane represented using the Steiner system $S(2,3,7)$ (historically known as a Steiner triple system, since $K = 3 = t+1$);

the 7 blocks correspond to the 7 lines of the plane, each of which contains exactly 3 points, and with the property that every pair of points lies on exactly one line.

Pág. 48 Gullstrand–Painlevé coordinates

(t_r, r, θ, ϕ)

data 16/04/2020

$$g = \left(1 - \frac{2GM}{c^2 r}\right) dt_r^2 - 2\sqrt{\frac{2GM}{c^2 r}}\, dt_r\, dr - dr^2 - r^2 d\theta^2 - r^2 \sin^2(\theta)\, d\phi^2, \quad (140)$$

G) Wolfram Physics Project Launch
www.youtube.com/watch?
V= rbfFt2uNEyQ

H) Technical Documents
www.wolframphysics.org/
technical-documents/

I) How we got here: the Backstory of the Wolfram Physics Project

Source: https://writings.
stephenwolfram.com/
2020/04/
how-we-got-here-the-
backstory-of-the-wolfram-
physics-project/

ref. https://www.wolframscience.com/
nks/p479--space-as-a-
network/

→

date 16.04.2020
S T Q [✓ Q] S S D

πr^2

$4/3 \pi r^3$,

the volume of a four-dimensional hypersphere is

$1/2 \pi^2 r^4$.

J) Uncommon Knowledge
by Al McDowell
Source:
www.spheritons.com
Obs: the Cause of Gravity
www.spheritons.com/cause-of-gravity.html

K) Relativity - the Special and General theory
Source:
www.ibiblio.org/ebooks/Einstein/Einstein_Relativity.pdf
Obs.: Pág. 184 (ao todo)

L) Finally we May Have a Path to the Fundamental of Physics... and It's Beautiful.

data 16.04.2020

https://writings.stephenwolfram.com/2020/04/finally-we-may-have-a-path-to-the-fundamental-theory-of-physics-and-its-beautiful/

139) 20200416 -

A) Books_68_2020/
Lagrangian Interaction An Introduction TO Relativistic Symmetry in Electrodynamics And Gravitation/
Lagrangian Interaction... pdf

obs: Usar este livro.

B) Books_68_2020/
Geometric Analysis/

Jost Jurgen Jost - Riemannian Geometry and Geometric Analysis. pdf

Pág. 23 (13)

1.4 Riemannian Metrics

Curvature: 8.8 Pág. 491
8.9 Pág. 498
8.10 Pág. 519 -

C) Books_68_2020/
Geometric Analysis/
Geometric function theory

data 16.04.2020

and nonlinear analysis... pdf

Pág. 35(52)

Figure 2.2

The stereographic projection of the sphere $S^{n-1}(a, r)$, $r = |a|$, is the restriction of the inversion in $S^{n-1}(2a, 2r)$.

D) Books_68_2020/
Geometric Analysis/
Guillermo Sapiro – Geometric partial differential equations ... pdf

E) Books_68_2020/
Geometric Analysis/
Peter Li – Geometric Analysis. pdf

Pág. 32(44)

Laplacian comparison theorem.

Veja: item Ⓐ.

Pág. 10(22) Volume comparison

data 16.04.2020
S T Q Q S S D

Theorem.

F) Books_68_2020/
Geometric Analysis/
Reiner Kuhnau - Handbook of Complex analysis ... volume 2...pdf
Pág. 01(16)
Quasiconformal Mappings in Euclidean Spaces

G) Books_68_2020/
Geometric Analysis/
Steven G. Krantz - Geometric Analysis of the Bergman Kernel and Metric...pdf
Pág. 251(262)
Chapter 7
Curvature of the Bergman Metric
Chapter 2
The Bergman Metric
Pág. 71(84).

H) Books_68_2020/
Complex Geometry An Introduction

data 16.04.2020
S T Q Q S S D

Huybrechts/
[Universitext] Daniel Huybrechts-
Complex Geometry... pdf

Pag. 91 (102)
2.4 The Projective Space
Pag. 182 (192)
4.3 Curvature
Pag. 217 (227)
4.B Hermite-Einstein and
Kähler-Einstein Metrics

I) Books_68_2020/
Complex Geometry An
Introduction Huybrechts/
Algebraic Geometry over the
Complex Numbers (2012...).pdf
Pag. 07 (19)
✡ 1.3 Bezout's Theorem

Important:

data 16.04.2020

An important feature of the projective plane is that any two lines meet.

In fact, it has a much stronger property:

Theorem 1.3.1 (Weak Bézout's theorem)

Any two algebraic curves in $\mathbb{P}^2$ intersect.

Obs: Usar no meu gráfico

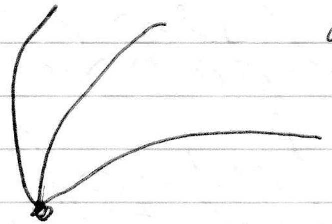

Pág. 70(79)

3.6 Line Bundles on Projective Space.

Pág. 117(125)

Chapter 6

Riemann Surfaces

Pág. 189(193)

Chapter 11

A little Algebraic Surface

data 16 04 2020

S T Q Q S S D

Theory.

Euler characteristic,

$e = b_0 - b_1 + b_2 - b_3 + b_4 =$

$2 - 4g + 2p_g + h^{11}$

J) 20200416 — ✡

Important:

J1) Books_69_2020/
Differential Geometry
and Its Applications/
the Geometry of the
Octonions by Tevian
Dray... pdf

Look at: capa

refere-se ao item
138, letra F
Pág. 39 Figure 11
the Fano Plane (Steiner System)
3 points = 1 line

J2) Books_69_2020/ Euclidean Geometry
and Transformations —

data 16.04.2020

S T Q [✓ Q] S S D

140) A) Books_2020_11/
Books_65_2020/
Descriptive Geometry/

B) Books_2020_11/
Books_66_2020/
Full Color Illustrations
of the Fourth Dimension-
2-volumes/

C) Books_66_2020/
Problèmes de Geométrie
Analytique/

D) Books_67_2020/
Mathematics/
Obs. Todo diretório.

E) Books_67_2020/
Space-Time-Matter Analytic
and Geometric Structures/
Sanet.st.Space-time-Matter.pdf.

F) Books_67_2020/
Solving problems in geometry/

141) 20200416-

Books_69_2020/
Knots and Links/
0521839475.pdf

Pag. 23(43)

data 16.04.2020

Figure 1.18 ⇒
A corner can be replaced by a circular arc.

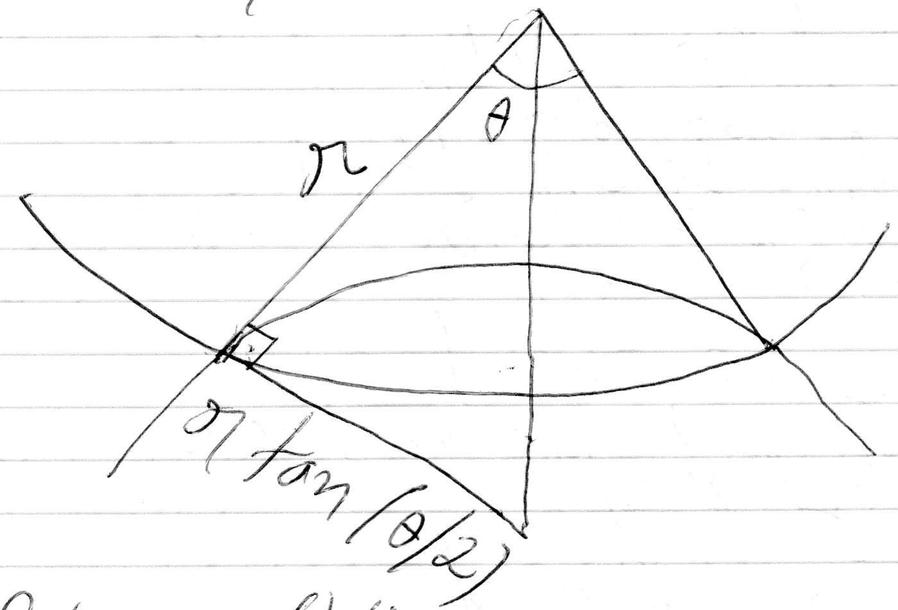

Pag. 32 (46)

Figure 2.1
Stereographic projection of the plane onto a sphere.

142) 20200416 –

Books_69_2020/
Knots and Links/
Algebraic Invariants of Links (2012 ...).djvu

Pag. 167 (181)

Chapter 8

Link with two components

data 16.04.2020

8.1 Bailey's Theorem

Let L be a 2-component link exterior X and group $\pi = \pi L$.

Source:

[Bai] Bailey, J. L.
Alexander Invariants of Links
Ph.D. Thesis, University of British Columbia (1977).

143) 20200416 -

Books_69_2020/
Knots and Links/

Gauss Diagram Invariants for Knots and Links (2001...).pdf

Pag. 2(18)

The winding index of a curve $C \subset \mathbb{R}^2 \subset \mathbb{C}$ around a point $p \notin C$ is

$$i(C,p) := \frac{1}{2\pi i} \oint \frac{1}{z-p} dz.$$

Pictorially it measures how many times the curve walks around p.

data 16.04.2020

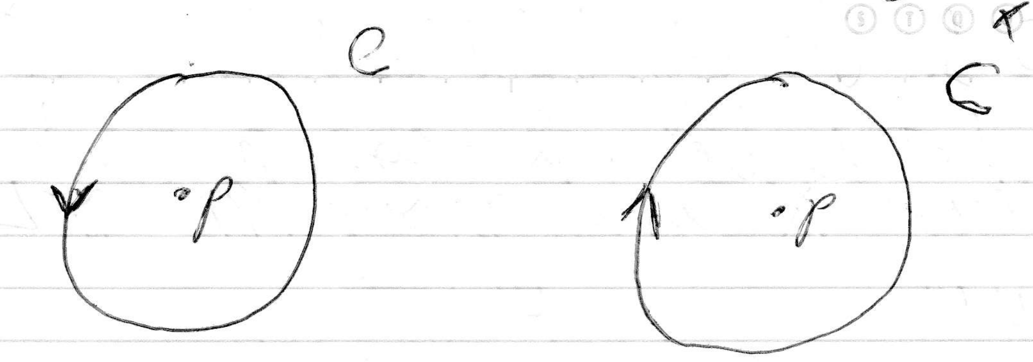

$i(C,p)=+1$ $i(C,p)=-1$

Figure 4

The whitney index $n(C)$ of a curve C is the degree of the map

$$\frac{C'}{\|C'\|} : S^1 \to S^1$$

where C' denotes the vector consisting of the derivatives.

The Whitney index of a diagram K is the Whitney index of the curve pr(K).

obs: Usar este divino gauss.

ref. Books_69_2020T ✡
Gauss-Werke

Gauss C.F-Werke theoretische Astronomie Band 7 (1906). djvu

data 16·04·2020
S T Q [✓ Q] S S D

Book: Books_69_2020/
Gauss_Werke/
Werke Fünfter Band by
Carl Friedrich Gauss ... pdf

→ Sphaeroidicorum ellipticorum
homogeneorum.

144) 20200417

A) Books_69_2020/
Mechanisms for Generating
Mathematical Curves/
sanet.st_3030421678.pdf

Pag. 63(74)

Chapter 4
Mechanisms for Generating Conical
Curves.

Pag. 64(75)

Fig. 4.1 Geometric construction
of the pedal Γ of the
ellipse C.

Shows the pedal of an ellipse with the focal points F and F'. The pedal curve results from the orthogonal projection of a fixed on the tangent lines of a given curve.
The locus of point M resulted by

data 17.04.2020

bringing a perpendicular from F to the ellipse tangent in point P the searched pedal is the r circle.

Pág. 121 (132)

Chapter 5

Mechanisms for Generating some Plane Mathematical Curves

Pág. 122 (133)

5.1.1 The Cissoid of a Circle

Obs: Cissoide de Diócles → curva de terceira ordem.

Fig 5.1 Geometrical construction of the cissoid

itens: a), b) e c)

B) Books-69-2020/
Differential Geometry and its Applications/
Differential Geometry and its applications ... pdf

Pág. 02 (24)

Fig 1.1 A parametrized line

Pág. 05 (27) →

17.04.2020

Figure 1.3 An infinitesimal piece of curve.

The three-dimensional version of this intuitive description simply uses our earlier exercise on the Pythagorean theorem in three dimensions.

Now, the length of the vector v, $|v|$, may be written $\sqrt{v \cdot v}$, where $\cdot$ denotes the dot product of vectors.

Recall that in general,

$v \cdot w = v^1 w^1 + v^2 w^2 + v^3 w^3$,

for $v = (v^1, v^2, v^3)$ and

$w = (w^1, w^2, w^3)$.

—

Pag. 11 (33)

Figure 1.10 The Witch of Agnesi

Exercise 1.1.17

—

Pag. 33(55)

Figure 1.20 Neil's parabola

Obs.: Veja item 144 A), Pag. 122(133)

Cissoid of Circle

⇒ As imagens são próximas (estudar!)

—

☆ Pag. 63 (85) →

data 17.04.2020

S T Q Q [✓ S] S D

Figure 1.39

Viviani's curve as the intersection of sphere and cylinder.

Hence, Viviani's curve is truly 'spherical'

[Exercise 1.7.11]

Show that the following curve is spherical

$(a\cos^2(t), a\sin(t)\cos(t), a\sin(t))$

$$x^2+y^2-\frac{b^2}{a^2}z^2=\boxed{\frac{4b^2}{a^4}}$$

Checar!

Look: Formula Bhaskara

$$x=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$$

Outra pesquisa: (relacionada)

B1) Books_69-2020/
Complex Numbers and Geometry/
[Spectrum Series] Liang...
Complex numbers and geometry.. djvu

→

data 17 04 2020
S T Q Q [✓ S] S D

Estude a capa do livro.
(desenho geométrico)

Na pág. 04 deste livro observamos várias referências aos seguintes livros:

B2) Circles: A Mathematical View by Dan Pedoe;

B3) Journey into Geometries by Marta Sved;

B4) The Trisectors by Underwood Dudley.

Obs.: Estes livros estão na mesma pasta.

(C) Books_69_2020/
Complex Numbers and Geometry/
✡ Circles_A Mathematical View
(1997...).pdf

Pág. 09(45)

3. Feuerbach's theorem

The nine-point circle of a triangle touches the incircle and excircles of the triangle.

Obs: Important.

Use-o com os demais para verificação de cálculos.

data 17.04.2020
S T Q Q [✓ S] S D

Use juntamente com o item
nº 138 (20200416),
letra F que se refere:
a) Fano Plane (Steiner System);
b) 7 lines of the plane
3 points – one line
c) Pappu's theorem.

Obs: Source deste book:
ARQ-IMPORTANT/
Cinderella.de/
the Cinderella.2 Manual.pdf
Pag. 42(57)
3.1 Pappu's theorem
(forma de construção).

Other: (same book)
Pag. 10(46)
4. Extension of Ptolemy's theorem

data 17.04.2020

S T Q Q [✓ S] S D

Pag. 11(47)

5. Fermat's problem

↳ use too: extension of Ptolemy's theorem.

A, B, C

are any three points in a plane.

To find a point P such that PA + PB + PC shall be least.

Pag. 14(50)

7. Coaxal systems of circles

Look: Pag. 16(52)

the radical axis of C_1 and C_2 is

$$2x[\ldots \qquad = \underline{\underline{0}}$$

looks ↗

Pag. 21(57)

10. Problem of Apollonius

Given three circles $\mathcal{C}_1, \mathcal{C}_2$ and

data 17.04.2020

S T Q Q S S D

C3, to draw a circle touching all three.

Fig. 27.

D) Books_69_2020/
Differential Geometry and Its Applications//
Differential Geometry and its applications....pdf

Pag. 416(442)

8.5 Curvature

In Section 3.4 we developed for Gauss curvature which depended only on the metric.

Section. 3.4

Pag. 123(145)

3.4 A Formula for Gauss Curvature.

Obs.: Pag. 122(144)

Figure 3.5 A pseudosphere.

IMPORTANT:

The pseudosphere is actually determined by the requirement of constant curvature and certain initial

data 17.04.2020

conditions. Suppose we start off by requiring that a surface of revolution M have constant Gauss curvature K = -1.

E) Books_69_2020/
Differential Geometry and Its Applications/
Lokenath... Introduction to Hilbert Spaces...pdf.

F) Books_69_2020/
Nature and Numbers A Mathematical Photo Shooting/
www... Nature_and_Numbers.pdf

obs: Estudar com atenção!

G) Books_69_2020/
Topology Illustrated/
Saber..Topology Illustrated...pdf

obs: Estudar em detalhes!!!

a) Pág. 452 (459)
4.2 The Euler-Poincaré Formula

b) Pág. 541 (548)
4. Metric Tensor
4.1 The need for measuring in calculus

data 17.04.2020
S T Q Q [✓ S] S D

c) Pag. 568(575)

1.2 Metric complexes.

Pag. 570(577)

They are orthogonal when this angle is zero.

For cells in the higher dimensional case, we are looking at the angle between a and the normal $b^{\perp}$ of $b = a^*$, illustrated for dimension $n = 2$

(image)

Dually, this angle can also be understood as the angle between the normal

$c = a^{\perp}$ of a and a^*, illustrated for dimension $n = 3$.

(image)

Even though (apesar de) these angles don't affect the volumes of either primal or dual cells, they do affect the volumes of new n-cells, the cells determined by each dual pair. (image)

data 17.04.2020

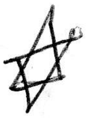

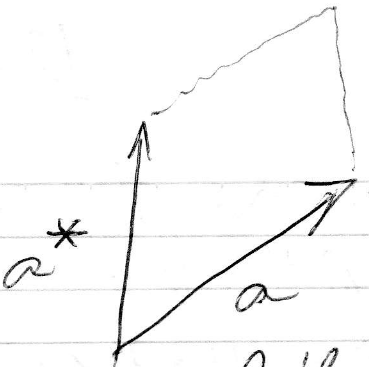

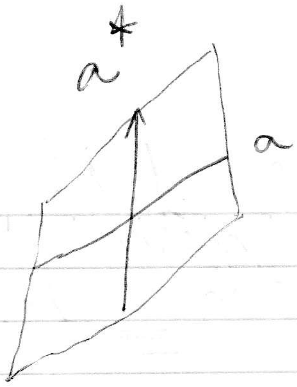

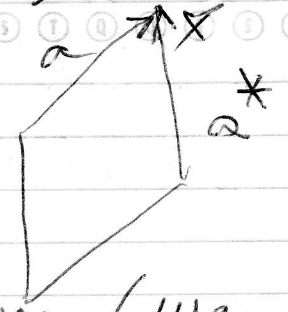

The area of this parallelogram (user Wolfram Project) does not depend on how a and a^* are attached to each other (they aren't) but only on the angle between them.

↳ important.

(Gauss curvature → space)

↳ angle of Gauss space (179°...)

The way we handle (Keidar) angle is, of course, via the inner product:

$\langle a, b \rangle \in R$

is defined for these three cases:

- adjacent $a, b \in K$,
- adjacent $a, b \in K^*$, and now also
- $a \in K$, $a^* \in K^*$.

Geometric interpretation of the angle between two vectors defined using an inner product (of the vectors)

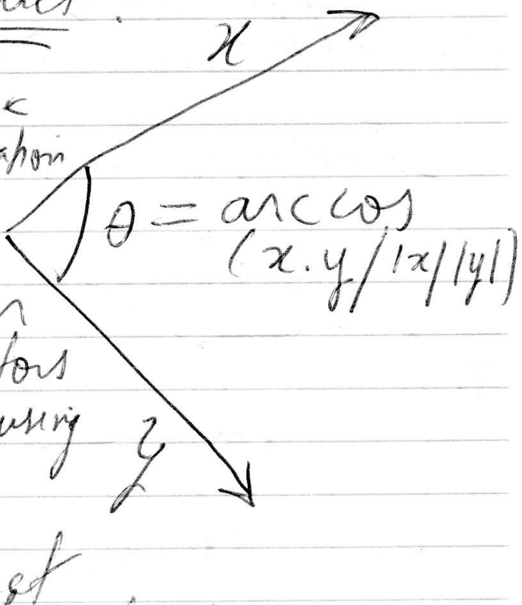

data 17.04.2020

Definition 1.13

A metric (cell) complex of dimension $\underline{\underline{n}}$ is a pair of Hodge-dual cell complexes K an K^* of dimension n equipped with a joint metric tensor, which is a combination of metric tensor for K and K^* as well as inner product of dual pairs:

$$(a, a^*) \to \langle a, a^* \rangle \in \mathbb{R}.$$

IMPORTANT: ✡

Exercise 1.14

Represent a hollow (oco) cube as a methic complex.

Image → interessante.

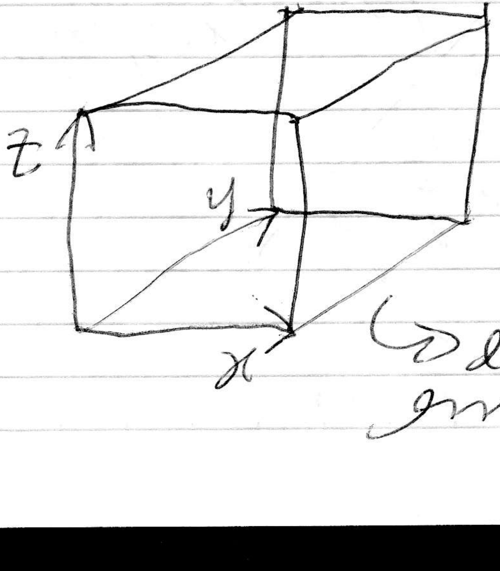

Este cubo inicial será as nossas coordenadas, usando-se coordenadas em funções algébricas

↳ este cubo será transformado em uma esfera dentro

data 17 04 2020

da métrica complexa.

triângulo esférico

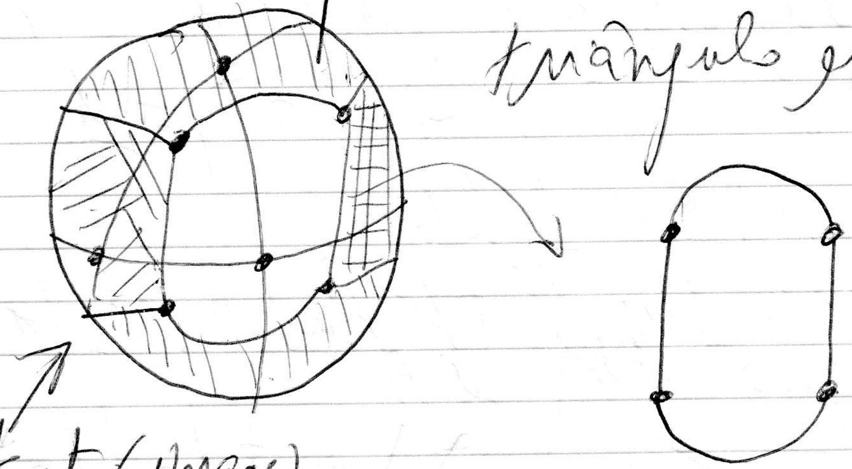

Important (image)

Exercise 1.15

Suppose K is a metric complex. Show how the set of its vertices K^0 becomes a metric space.

It follows from the definition that, in addition to the two volume functions of these complexes, we can also derive the angle function.

$(a, a^*) \rightarrow$

$\cos \widehat{a\,a^*} = \dfrac{\langle a, a^* \rangle}{|a|\,|a^*|}$

importantíssimo

A área do paralelogramo depende do ângulo dos lados ou vetores.

The function is well-defined whenever the expression in the right-hand

17.04.2020

belongs to our ring R.

This is always the case when R is a field or $|a| = |a^*| = 1$.

Definition 1.16

We say that K has a rectangular geometry if

$$\langle a, a^* \rangle = \pm |a| |a^*|.$$

In other words, all dual pairs are orthogonal:

$$\cos \widehat{a\, a^*} = \pm 1$$

The case of $n = 1$ previously considered is rectangular.

Definition 1.17

We will say that the complex K has the standard geometry if it is rectangular and also fully normalized:

$|a| = 1$ for all cells $a \in K \sqcup K^*$.

We always assume that $\mathbb{R}^n$ has the standard geometry.

H) a) Books_69_2020/

Gauss/

epdf.pub_the foundations of differential geometry.djvu

data 17.04.2020
S T Q Q [✓ S] S D

b) epdf.pub_the foundations of geometry.djvu
David Hilbert 188 pages.

obs: look:

Pag. 11 (29)

§ 5. Group III

Axiom of Parallels
(Euclid's axiom).

I) Books_69-2020/
The Mobius Strip/
Clifford A. Pickover...Mobius...djvu

other copy:

1560258268.pdf

145) 20200418

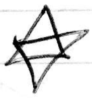

A) Books_70_2020/
Applications of Vector Analysis
and Complex Variables in
Engineering/
samet_st-3030411672.pdf

obs. Livro todo

A1) Pag. 51 (63)

Coordinate Transformations:

data 18 04 2020

S T Q Q S S D

Definition of Vectors and Tensors

Figure 5.1 Coordinate Transformation

Transformation of Cartesian A into Cartesian System B.

We consider a Cartesian B with coordinate axes ($\tilde{1}, \tilde{2}, \tilde{3}$), representing a point in this system as $\tilde{x}_i$.
The origin of B in terms of the original coordinate system A (1, 2, 3) is given by the coordinates b_j.
(figure 5.1)

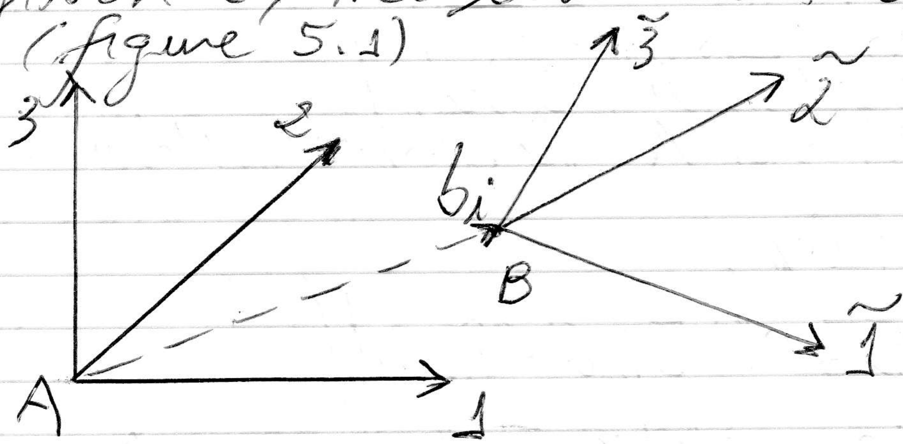

Both coordinate system are Cartesian, and the scales in both systems are the same.
The vectors τ^i_m, defined as

$$\tau^i_m = \frac{\partial x_i}{\partial \tilde{x}_m} \qquad i=1,2,3 \quad m=1,2,3$$

are unit vectors tangent to the coordinate directions $\tilde{i}$ (compare (2.5))

data 18.04.2020

S T Q Q S [✓ S] D

with $\tilde{x}_m$ replacing the arc length along the curve, s.

We define similar vectors that are tangent to the coordinates in system A in terms of the coordinates in system B. We call these unit tangent vectors $\overset{m}{\gamma}_i$:

$$\overset{m}{\gamma}_i = \frac{\partial \tilde{x}_m}{\partial x_i}$$

We express the original coordinates of a point x_i in terms of the new coordinates of that same point and vice-versa by integrating the latter two equations:

$$x_i = \overset{i}{\underset{m}{\tau}} \tilde{x}_m + b_i,$$

where b_i represents the origin of B in terms of x_i, and

$$\tilde{x}_m = \overset{m}{\underset{i}{\gamma}} x_1 + \tilde{b}_m,$$

where $\tilde{b}$ represents the origin of the original system A in terms of the new coordinates.

data 18.04.2020

The vectors $\underset{m}{\tau}^{i}$ are mutually*

orthogonal unit vectors,

$$\underset{m}{\tau}^{i}\,\underset{n}{\tau}^{i} = \delta_{mn}.$$

A similar equation holds for unit vectors $\overset{m}{\underset{i}{\gamma}}$,

$$\overset{m}{\underset{i}{\gamma}}\,\overset{n}{\underset{i}{\gamma}} = \delta_{mn}.$$

We obtain a relationship between vectors $\underset{m}{\tau}^{i}$ and $\overset{m}{\underset{i}{\gamma}}$ from (5.1) and (5.2)

$$\underset{m}{\tau}^{i}\,\overset{n}{\underset{i}{\gamma}} = \frac{\partial x_i}{\partial \tilde{x}_m}\,\frac{\partial \tilde{x}_n}{\partial x_i} = \frac{\partial \tilde{x}_n}{\partial \tilde{x}_m} = \delta_{mn}$$

The moduli of $\underset{m}{\tau}^{i}$ and of $\overset{m}{\underset{i}{\gamma}}$ are 1,

we combine this with the latter equation:

$$\underset{m}{\tau}^{i} = \overset{m}{\underset{i}{\gamma}}.$$

We introduce a matrix a_{ij},

data 18.04.2020

8

$$a_{im} = \tau^{i}_{m} = \gamma^{m}_{i} = \frac{\partial x_i}{\partial x_m} = \frac{\partial \tilde{x}_m}{\partial x_i}$$

and write (5.3) and (5.4) as

$$x_i = a_{ij}\tilde{x}_j + b_i,$$

and

$$\tilde{x}_i = a_{ji} x_j + \tilde{b}_i.$$

We observe from (5.9) that

$$a_{im} a_{in} = \frac{\partial x_i}{\partial \tilde{x}_m} \frac{\partial \tilde{x}_n}{\partial x_i} = \delta_{mn},$$

and that

$$a_{im} a_{jm} = \frac{\partial \tilde{x}_m}{\partial x_i} \frac{\partial x_j}{\partial \tilde{x}_m} = \delta_{ij}.$$

A2)
~~B)~~ mesmo livro

Pág. 91 (103)

8.3.2 Transformation of vectors

A3) mesmo livro

Pág. 102 (114) →

data 18.04.2020

S T Q Q S ☒ D

✡ 8.4 Irrotational and Divergence-Free Vector Fields

$$\frac{\partial w}{\partial \bar{z}} = 0$$

→ light over curve.

w = velocity potential (complex variables z and $\bar{z}$).

Pág. 103 (115)

Equations (8.133) and (8.134) are equivalent complex forms of the

Cauchy-Riemann Equations;

B) Books_70_2020/
Fundamentals of Geometry Construction the Math behind the Cod/
Sanet_st_3030431304.pdf

Obs.: livro todo.

B1) Pág. 07(25)

1.2.3 Types of Coordinates Systems.

data 18.04.2020

IMPORTANT: ✡

Criar o meu sistema de coordenadas (funções representam cada vetor)

→ CURVE VECTOR (pioneirismo)

→ MUSICAL CURVE VECTOR.

Voltando:

- Polar Coordinates,
- Cylindrical coordinates,
- Spherical coordinates,
- Absolute coordinates e
- World coordinates (x, y, z) versus relative coordinates (x', y', z').

B2) Pag. 13 (31)
Vectors

B3) Matrices
Pag. 17 (35)

B4) Pag. 80 (96) ✡
3.1.3 Distance of a Point to a Plane (cartesiano)

data 18.04.2020

Projetá-lo sobre o espaço curvo
(veja curva de Gauss) - 179°..

Observation:
Use todo este livro com o livro anterior A.

C) Books_70_2020/
Geometry_Fundamentals_Belyaev/
geometry.pdf

Observation:
Utilizá-lo conjuntamente com os livros A e B.

D) Books_70_2020/
Mathematical Amazements and Surprises Fascinating Figures and noteworthy Numbers/
sarct.st_Mathematical...pdf

D1) Pág. 27 (26) ✡
Armstrong Numbers

18.04.2020

K=3 → 3 numbers

$$153 = 1^3 + 5^3 + 3^3 = 153$$

K=39 obtemos dois números com final par e ímpar, ambos com 39 números.

... 400

... 401

obs. Utilizaremos estes números para dividir uma área.

Important:

153 is a triangular number

$1+2+3+4+5+6+7+8+9+10+$

$11+12+13+14+15+16+17 = 153$

Obs:

153 is also a number that can be expressed as the sum of consecutive factorials:

data 18.04.2020

$1! + 2! + 3! + 4! + 5! = 153$

D2) Pag. 147 (146)

Geometry Surprises

Pag. 152 (151)

Scalene triangle →

all three sides are of different lengths.

The medians partition the triangle into six smaller triangles, which happen to be equal in area.

We can also partition any triangle into four congruent triangles by using the midpoints of the three sides and joining them to form a medial triangle.

Point G (the point of intersection of the medians) is the center of gravity of the triangle, and is called the centroid.

✡

↳ Fascinating fact about the centroid of a triangle:

↳ IMPORTANT: ✡

data 18 04 2020

THE CENTROID TRISECTS EACH OF THE MEDIANS.

Figure 4.15

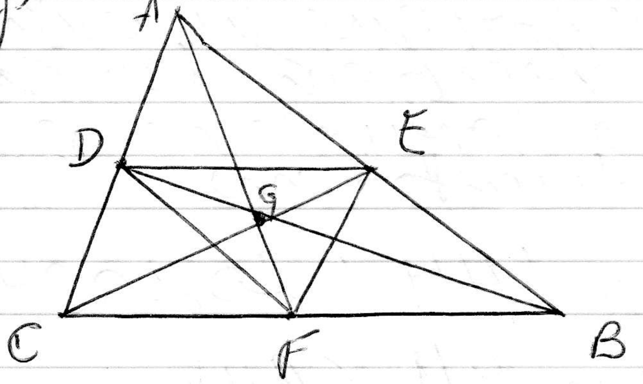

$GD = \frac{1}{3} BD$; $GE = \frac{1}{3} CE$, and $GF = \frac{1}{3} AF$

Important: ✡

estudar aplicação para trissecção do ângulo.

D3) Some Quadrilateral Properties
Pag. 154 (153)

Varignon parallelogram

If we consecutively connect the midpoint of any quadrilateral with line segments, we always get a parallelogram.

IMPORTANT: ✡

19.04.2020

Off: Dia de Santo Expedito!!!

IMPORTANT:

Esta propriedade é um dos instrumentos que eu deverei usar para a checagem dos cálculos.

Use-o conjuntamente com os demais descritos no

item 144 (20200417),

subitem B4 (B4) e

The Trisectors by Dudley

subitem C (Feuerbach's theorem)

↳ que faz referência ao

item 138 (20200416).

Vou explicar melhor:

Com Varignon Parallelogram poderemos obter um paralelogramo, ou seja 2 pares de retas paralelas que foram obtidas de um quadrilátero (que nem sempre) é irregular.

Desta forma nós conseguiremos obter padrões para realizar deduções de cálculos de áreas e assim,

data 19.04.2020

inferimos a área total deste quadrilátero irregular, cujas (~~arestas~~) lados não são paralelas e os ângulos formados entre elas (~~arestas~~) (lados) são desconhecidos.

Obs: Aresta é a reta que se origina a partir da intersecção de 2 planos que formam um ângulo.
(3D do quadrilátero irregular)

Varignon parallelogram aplica-se para todos tipos de quadrilátero: regular e irregular.

(quadrado)
- ↳ é um paralelograma.
- ↳ é um polígono que reúne as características do losango e do retângulo ao mesmo tempo.
- ↳ os lados são congruentes (mesma medida).
- ↳ um quadrado também pode ser chamado de losango.

Obs: 1. Todo paralelogramo é quadrado, retângulo ou losango;
2. Todo retângulo ou losango é também um quadrado.

IMPORTANT:

Varignon parallelogram

data 19.04.2020

8

VARIGNON PARALLELOGRAM

has half the area of the original quadrilateral.

(mais um instrumental para calcular a área de um determinado quadrilátero).

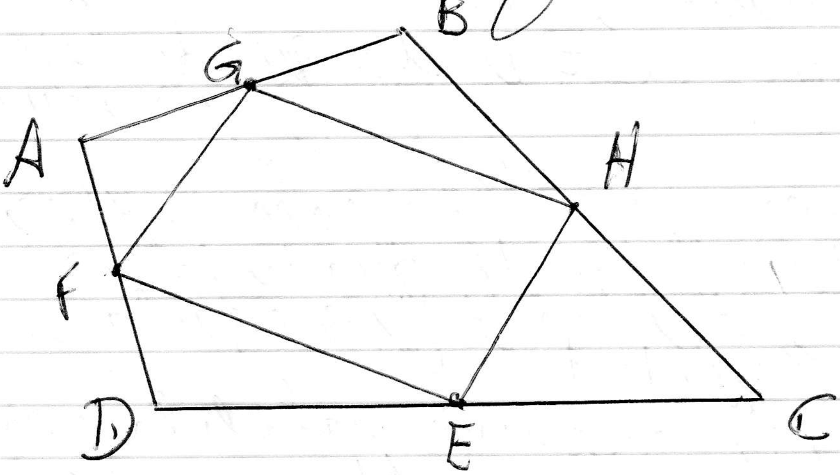

The area of parallelogram EFGH is one-half of a area of quadrilateral ABCD, and the perimeter of paralllogram EFGH equals AC + BD.

IMPORTANT:

Outro instrumento para checagem dos meus calculos.

Quantas mais redundâncias de checagem e melhor, diminui o erro.

data 19 04 2020
X

Figure 4.17 (Important)

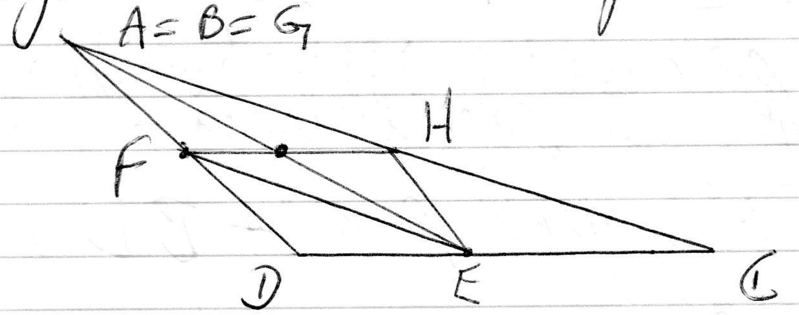

If we reduce the length of one side of the quadrilateral, AB to length zero. (We get a triangle with one vertex having points A, B, and G coincide as in figure 4.17.
Here, too, the resulting

Varignon Parallelogram

still exists and it (is) one-half the area of the original triangle (or quadrilateral with a zero-length side).

Relate with Figure 4.14 where we found that the triangle ABC was partitioned into four triangles of equal area, leaving us with a similar result when we consider the parallelogram ADFE in figure 4.14 which is analogous to parallelogram AFEH in Figure 4.17.

→

data 19.04.2020
S T Q Q S S [✓ D]

IMPORTANT: ✡

The perimeter relationship (perimeter of parallelogram EFGH equal AC + BD)

Observation:

AC and BD → sides of triangle

pois AB = zero.

Vejamos.

Figure 4.16 AC + BD (não são lados)

Figure 4.17 AC + BD (são lados)

estudar a angulação!!! ✡

also holds true for the situation shown in figure 4.17.

Pág. 155 (154)

Figure 4.18

A quadrilateral whose diagonals are perpendicular. When the side midpoints of this quadrilateral are connected, a rectangulo EFGH is produced.

data 19 04 2020

S T Q Q S S D X

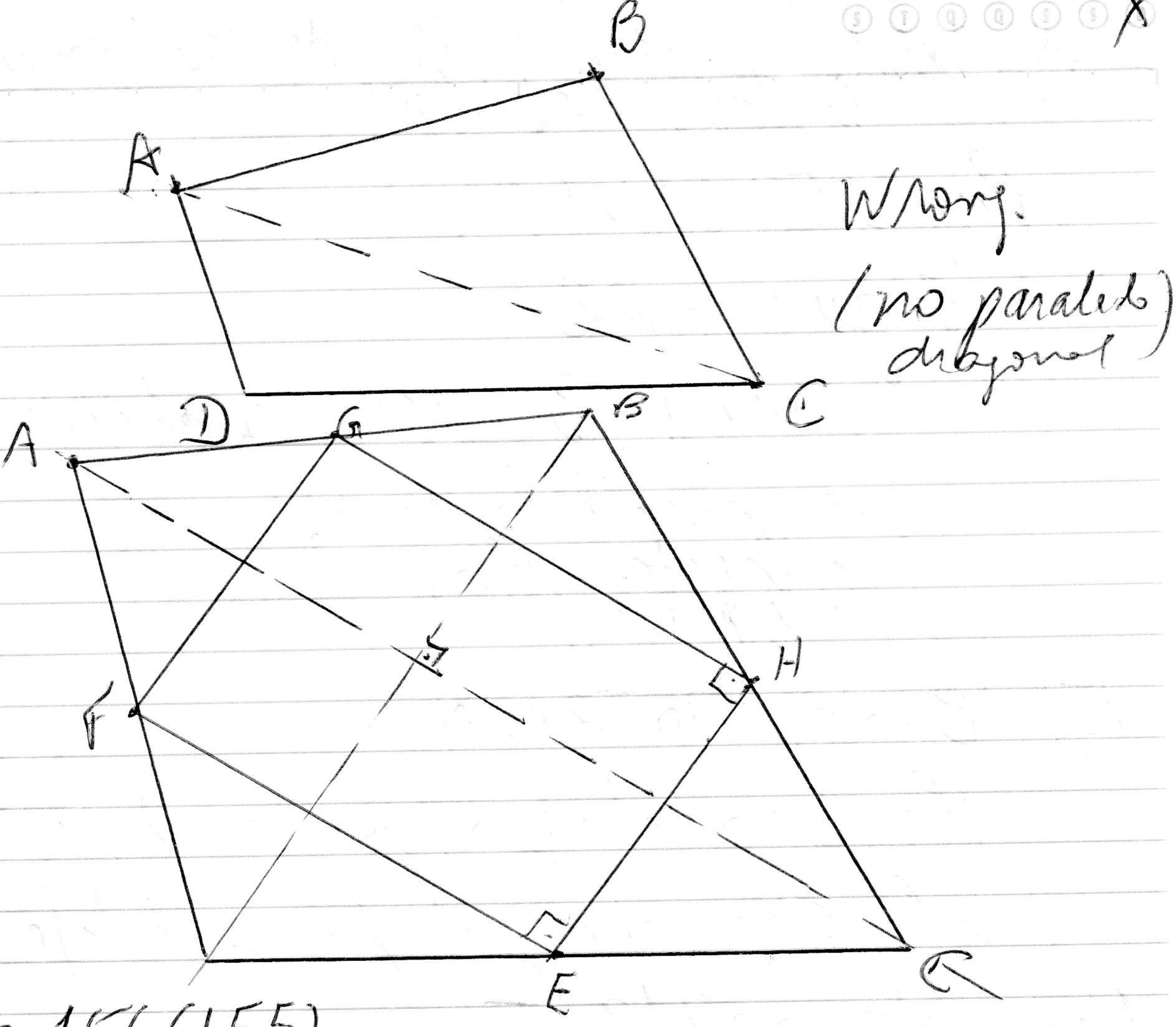
B
A
Wrong.
(no parallels)
diagonal
D
C
A
G
B
H
F
E
C

Pag. 156 (155)

✡

Figure 4.19

If the quadrilateral has perpendicular diagonals that are the same length, then the quadrilateral produced by joining the side-midpoints is a special kind of (rectangulo) rectangle that we know as a ~~S~~ square.

Figure 4.20

If the original quadrilateral has diagonals of equal length are not perpendicular, the the quadrilateral

data 19.04.2020

formed by joining the side-midpoints *
will be a rhombus.

Pag. 157(156)

Figure 4.22

Exercise 3 special quadrilaterals:

a parallelogram,
a rectangle,
a square.

E) Books_70_2020/

the Geometry of Art and Life/

the Geometry_of Art and Life.pdf

146) 20200419

A) Books_70_2020/
Microelectronics, Circuit Analysis and Design, 4th/
Microelectronics Circuit Analysis Design.pdf

Pag. 77(100)

The ripple voltage is then (2.9)

$$V_r = \frac{V_M}{2fRC}$$

data 19.04.2020
S T Q Q S S [✓ D]

Pág. 79 (102)

Since the charging through the capacitor is triangular, we have that the average capacitor current during the diode charging time is

$$i_{C,avg} = \underset{\uparrow}{\frac{\pi}{2}}\,\frac{V_M}{R}\sqrt{\frac{2V_M}{V_r}} \qquad (2.19)$$

Obs.: Triangular / π

– Checar correlações!!!

B) Books-70-2020/
Ideas of quantum chemistry
Ideas of quantum chemistry by Lucjan Piela....pdf

Pág. 282 (317)

7.1.4 Application of the Ritz Method

We will use the Ritz variational method (see Chapter 5, p. 202) to solve the Schrödinger equation.

... spherical harmonics

$$\{Y_l^m(\theta,\phi)\} \text{ and } \{Y_{l'}^{m'}(\theta,\phi)\}$$

$\neq$ $\neq$ $\underbrace{\Phi, \phi}_{\neq}$

Ⓗ ⊖

data 19.04.2020

therefore, we may use as
variational function:[11]

11. the products $Y_l^m(\Theta,\phi)$
$Y_{l'}^{m'}(\theta,\phi)$

may be used to produce linear combinations that are automatically the

EIGENFUNCTIONS of
$\hat{J}^2$ and $\hat{J}_z$, and have the proper parity (chapter 2)

this may be achieved by using the
CLEBSCH-GORDAN coefficients

Source: D. M. Brink, G. R. Satchler,
"Angular Momentum" 1975.

① Books_70_2020/
Angular Momentum Brink/
Angular_momentum_1968.pdf
Pág. 05(13)

The geometrical symmetry
operations of translation and

19.04.2020

rotation can be looked at in two ways, called active and passive.

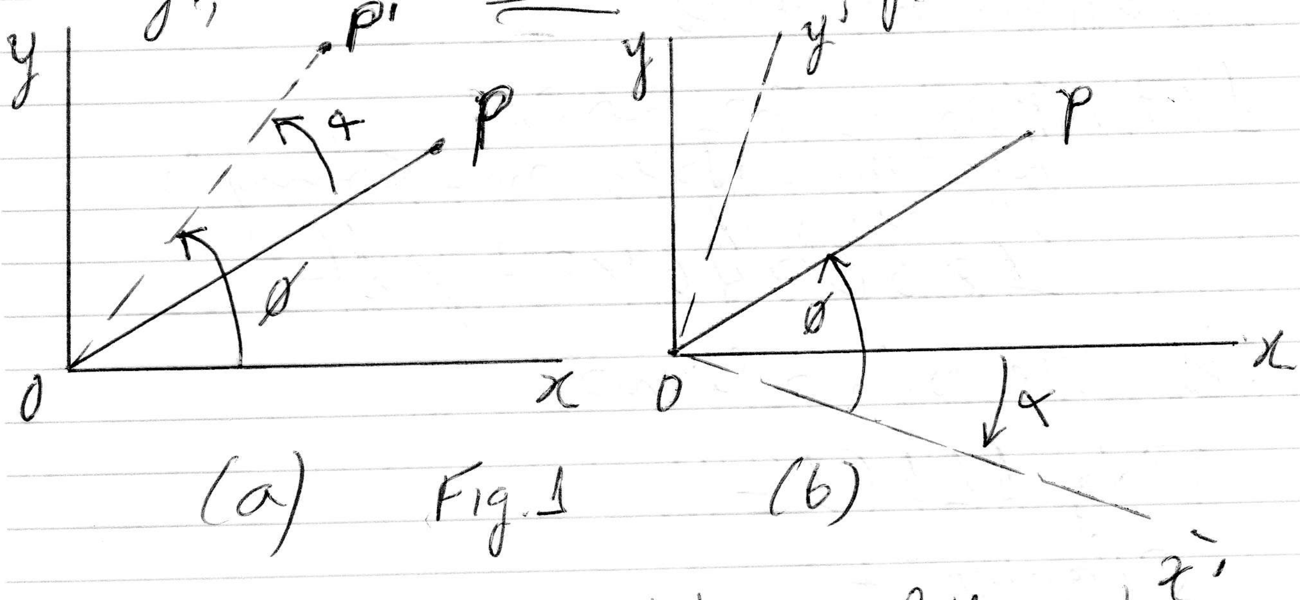

(a) Fig. 1 (b)

a) Represents a rotation of the system. A vector OP is carried into OP' by the rotation.

b) Represents the equivalent
rotation of axes.

The vector OP remains fixed, but its coordinates are changed. The position of OP relative to the new axes in (b) is the same as that OP' relative to axes in (a).

A rotation actually rotates the system from one position to another. The passive approach interprets the symmetry operations as coordinates transformations and the symmetry of the dynamical laws is expressed by

data 19.04.2020

stating that the equations of motion are invariant under translation or rotation of axes. —

147) A)Books_70_2020/
Lecture_Notes_on_Complex_
Analysis/
1860946437.pdf

Obs: Use todo o livro.

Pág. 111(124)

Chapter 7

Complex Integration

Pág. 113(126)

7.2. the Length of a Contour.

Polygon:

Condition:

$\gamma : [a,b] \to \mathbb{C}$ is given smooth path.

We can consider an inscribed polygon as an approximation to the path and its length will be an approximation to the length of the path.

polygon $\Rightarrow$ vertices $\{z_0, z_1, \ldots, z_n\}$.

data 19.04.2020

X

Then the length of such (de tal) a polygon path is the SUM of the lengths of its straight line segments, namely,

$$\sum_{j=1}^{n} |z_j - z_{j-1}|.$$

Suppose that the ends of the line segments correspond to values $t_0, t_1, \dots, t_n$ of the parameter, that is, $z_j = \gamma(t_j)$.

Then

$$\sum_{j=1}^{n} |z_j - z_{j-1}| = \sum_{j=1}^{n} |\gamma(t_j) - \gamma(t_{j-1})| =$$

$$\sum_{j=1}^{n} \left| \frac{\gamma(t_j) - \gamma(t_{j-1})}{t_j - t_{j-1}} \right| (t_j - t_{j-1}),$$

with $a = t_0 < t_1 < \dots < t_n = b$.

Now, the quotient $\dfrac{\gamma(t_j) - \gamma(t_{j-1})}{t_j - t_{j-1}}$

can be considered an approximation to the derivative $\gamma'(t_j)$,

in which case the sum of the right hand side (lado direito) becomes an approximation to an integral.

data 19.04.2020

S T Q Q S S [✓ D]

$$\sum_{j=1}^{n} \left| \frac{\gamma(t_j)-\gamma(t_{j-1})}{t_j-t_{j-1}} \right| (t_j-t_{j-1})$$

$$\xrightarrow{\text{approx}} \int_a^b |\gamma'(t)|\,dt.$$

Indeed (de fato), in view of the assumed smoothness of γ one might expect that the left hand side converges to the right hand side as $\max_j (t_j - t_{j-1}) \to 0$.

Rather than (ao invés de) prove this here, we will just take this as motivation for our definition of the length of a path, as follows.

... estudar todo este capítulo.

Others:

Pag. 49(62)

3.8 Polygons and Paths in $\mathbb{C}$

Pag. 14(27)

1.6 The Riemann Sphere and the Extended Complex Plane.

Let S^2 denote the sphere →

data 19 04 2020

$\{(x,y,z)\in\mathbb{R}^3 : x^2+y^2+z^2=1\}$ in $\mathbb{R}^3$ and let N denote the point (0,0,1), the "north pole" of S^2.

Think of $\mathbb{C}$ as the plane $\{(x,y,z) : z=0\}$, containing the equator of S^2.

Then given any point P in the plane, the straight line through P and N meets the sphere S^2 in a unique place, P' say. As P varies over the plane, the corresponding point P' varies over $S^2\setminus\{N\}$.

This sets up a one-one correspondence between $\mathbb{C}$ and $S^2\setminus\{N\}$.

ou estudar

Pág. 66 (79)

4.4. Cauchy-Riemann Equations

$z = x+iy \longleftrightarrow (x,y)\in\mathbb{R}^2$

$$f(x+iy) = \mathrm{Re}\, f(x+iy) + i\,\mathrm{Im}\, f(x+iy)$$

$$\equiv u(x,y) + iv(x,y)$$

o, ao estudar

IMPORTANT:

Chapter 5 →

data 19.04.2020

Pág. 86 (99)

5.6 The Appearance of π

Discuss various properties of sin x and cos x for real x.

In particular, we would like to introduce the number π.

→ estudar!!!

Pág. 80 (93)

5.2 Complex Hyperbolic Functions

The complex hyperbolic sine and cosine functions are defined as the series

$$\text{senh}\, z = \sum_{n=0}^{\infty} \frac{z^{2n+1}}{(2n+1)!} = z + \underbrace{\frac{z^3}{3!} + \frac{z^5}{5!} + \frac{z^7}{7!}}_{\text{ímpar}} + \dots$$

$$\cosh z = \sum_{n=0}^{\infty} \frac{z^{2n}}{(2n)!} = 1 + \underbrace{\frac{z^2}{2!} + \frac{z^4}{4!} + \frac{z^6}{6!}}_{\text{par}} + \dots$$

estudar!

B) Books_70_2020/
Fluid Mechanics, by Russell
C. Hibbeler/

0132777622.pdf

data 16 04 2020 ✡

Pág. 146 (170)

3.3 Graphical Descriptions of Fluid Flow

Streamline is a curve that is drawn through the fluid in such a manner that it indicates the direction of the velocity of particles located on it at a particular instant of time. (rel)

(function)
↑↑↑
my equation

✡ the velocity of any particle is always tangent to the streamline along which it traveling at that instant.

Consequently no fluid can flow across a streamline, only along the streamline.

Stagnation point because here the velocity of any particle is momentarily reduced to zero when it strikes the cylinder's surface.

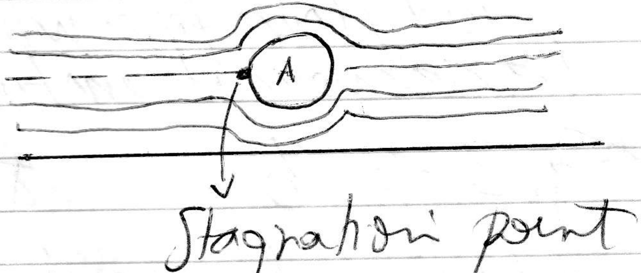

✡

obs: Sometimes the streamlines can be a function of both space and time.
(Einstein).

data 19.04.2020

In the case of two-dimensional flow, the particle's velocity will have two components,

u in the x direction and

v in the y direction.

Fig. 3-7

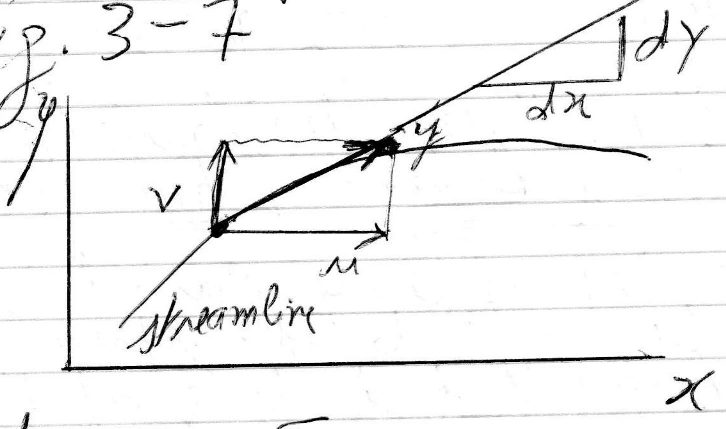

$$\frac{dy}{dx} = \frac{v}{u}$$

Integrating this equation will give the equation of the streamlines

$$Y = f(x) + C$$

↳ constant.

IMPORTANT:

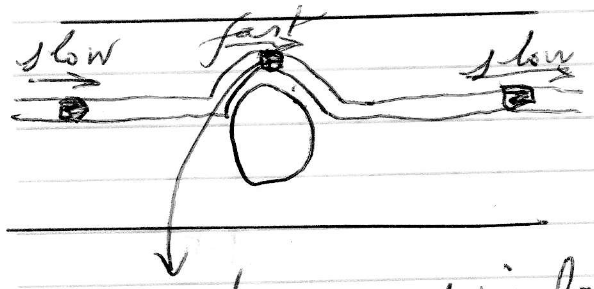

fluid flows faster within a narrow streamtube.

Fig. 3-6 ⓑ.

use isto para simular o desvio da luz quando passa perto de uma estrela como o Sol → Einstein - Teory of Relativity.

20.04.2020
X

147) 20200420

A) Books_70_2020/
Engineering Mechanics Dynamics/
J.L. Meriam ... Engineering Mechanics.pdf

Pag. 22(41)

Figure 2/1 Choice of Coordinates

rectangular coordinates x, y, z

its cylindrical coordinates r, θ, z, or

its spherical coordinates R, θ, φ.

the motion of P can also be described by measurements along the tangent t and the normal n to the curve.

the direction of n lies in the local plane of the curve.

These last two measurements are called path variables.

Motion Particles { absolute-motion analysis
{ relative-motion analysis

this plane is called the osculating plane, which the plane contains P and two points A and B, one on either side of P, becomes the osculating plane as the distances between the points approach zero.

20.04.2020

Pág. 45(64)

2/4 Rectangular Coordinates

$$r = xi + yj$$

$$v = \dot{r} = \dot{x}i + \dot{y}j \qquad \text{Figure 2/7}$$

$$a = \dot{v} = \ddot{r} = \ddot{x}i + \ddot{y}j$$

B) Books_70_2020/ Mathematics/Higher Level for the IB Diploma/ Mathematics for IB Diploma... pdf

Pág. 389(400) Obs.: Todo livro.

13C Distances

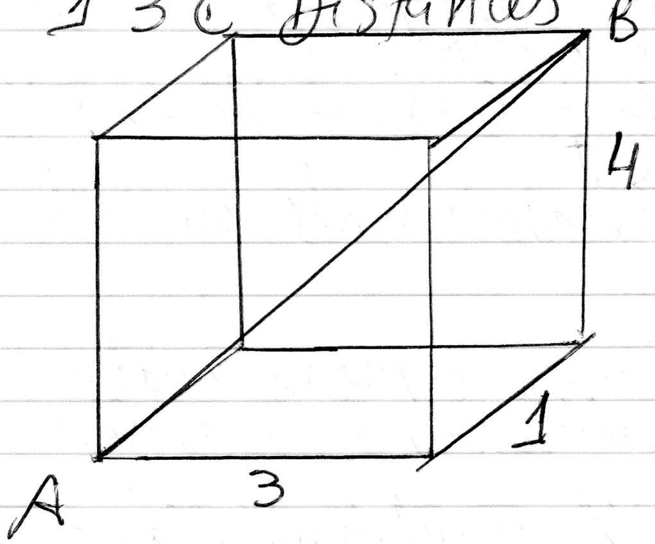

$$\overline{AB} = \begin{pmatrix} 3 \\ 1 \\ 4 \end{pmatrix}$$

$$AB = \sqrt{3^2+1^2+4^2} = \sqrt{26}$$

this quantity is called the MAGNITUDE of AB, and written as $|\overline{AB}|$.

The magnitude of a vector, $a = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}$, is

$$|a| = \sqrt{a_1^2 + a_2^2 + a_3^2}.$$

The distance between points with position vector a and b is $|b-a|$

20 04 2020

C) Books_70_2020/
Mathematics/Higher Level for the IB Diploma/
0199152268.pdf

Pág. 466(251)

Volumes of solids of revolution when rotated about y-axis

Obs: Ver livro todo.

D) Books_70_2020/
Rocket Propulsion Elements/
George P. Sutton .. Rocket..pdf

Pág. 106 (130)

4.3 Basic Relations of Motion

Pág. 107(131)

Figure 4-5 Two-dimensional free body force diagram for flying vehicle with wings and fins.

E) Books_70_2020/
Schaum's Outline of Trigonometry/
Robert Moyer.. Schaum.. pdf

Obs. Livro todo.

F) Books_20_2020/
Complex Variables Second Edition.../
0486406792.pdf

20.04.2020

G) Books_70_2020/
Weak Scale Supersymmetry from Superfields...|

0521857864.pdf

Pag. 41(59)

Chapter 4

the supersymmetry algebra

4.1 Rotations

In classical mechanics, rotations of three-vectors can be represented by a rotation matrix R acting upon vectors such as $x = (x, y, z)$ as

$$x_i \to x_i' = R_{ij} x_j.$$

148) 20200421

A) Books_70_2020/
Gravitation and gauge symmetries
Gravitation and gauge symmetries (2001...).pdf

Pag. 13(26) Figure 1.2.

Vector components change under the parallel transport.

On a general surface we cannot set up

21.04.2020

Cartesian coordinates: the best we can do is to introduce two arbitrary of coordinate lines and label them by $u^\alpha = (u^1, u^2)$. These curvilinear (or Gaussian) coordinates do not always a direct geometric interpretation.

obs. Use este livro.

A2) Books_70_2020/
Gravitation and gauge symmetries/
String Gravity and Physics at the Planck Energy ...pdf

Pag. 11(26)

Lectures on string theory in curved spacetimes

A3) Books_70_2020/
Gravitation and gauge symmetries/
The Mathematical theory of Cosmic ...pdf

Pag. 59(73) - Chapter 3

String dynamics in flat space

The Minkowski spacetime is the only background metric of cosmological interest in which the string equation of motion (2.14) can be integrated exactly.

21.04.2020

B) Books_70_2020/
Lightware_Communications/
Lightware Communications.pdf

C) Books_70_2020/
The Physics of Energy/
the Physics of Energy.pdf

Pag. 52(72)

3.6 Maxwell's Equations

Pag. 263(283)

The Forces of Nature

Pag. 266(286)

Figure 14.2 Feynman diagram describing two electrons interacting exchange of a photon, the force carrier for electromagnetic interactions.

D) Books_70_2020/
The Universe of Quadrics/
3662610523.pdf

Pag. 147(155)

Quadrics in the real projective space

21 04 2020

Figure 4.5 a) Self-polar ~~con~~ triangle
b) Self conjugate.

Proof: Ceva and Menelaos the product of the cross-ratios on the three sides must be negative, which results in the two cases.

Pag. 129(137)

Figure 4.3 Perspectivities in the projective space.

Obs: Livro muito bom!!!

D21 Books_70_2020/
The Universe of Quadrics/
The Universe of Conics ...pdf

Obs. Livro muito bom!!!

Pag. 79(87)

Curvature of conics and the construction of osculating circles.

Pag. 83(91)

Figure 3.12

Right: For the particular point P, the osculating circle op meets the ellipse c at the antipode $\overline{P}$ of P.

data 21 04 2020
S [✓ T] Q Q S S D

E) Books_70_2020/
High Resolution Imaging Detectors and Applications/

E1) High-resolution imaging ... pdf

Pag. 03 (36) Preamble

Exact description of electromagnetic radiation and its behavior requires a knowledge of

MAXWELL's electromagnetic field equations...

Pag. 09 (42) Preamble

Schröndinger's equation.

E2) 9814613274.pdf

Pag. 07 (40) Preamble

Wave-particle duality.

Pag. 08 (41) Broglie equations.

F) Books_70_2020/
Geometria-Moderna/
Geometria-Moise Downs.pdf

obs: ver livro todo.
Bem didático.

data 25.04.2020

S T Q Q S S [✓ D]

149) 20200425

A) Books_71_2020/
Analysis in Euclidean Space/
Kenneth Hoffman - Analysis... pdf

Pág. 119(136)

4. Calculus revisited.

B) Books_71_2020/
The Large Scale Structure of Space-Time
S. Weinberg - Large ... pdf

Pág. 47(56)

2.8 The Volume element and Gauss' theorem.

C) Books_71_2020/
Missile Guidance and Control System/
Modern missile guidance...pdf

Pág. 31(48)

4. Analysis of PN Guided Missile System in the Frequence Domain

D) Books_71_2020/ ✡
Transparent Schalen Form, Topologie, Tragwerk/
3433031207.pdf

Obs: Todo o livro!!!

data 25.04.2020
S T Q Q S [✓ S] D

E) Books_71_2020/
Handbook of Engineering [Electromagnetics]/
[Discrete Mathematics...] Rajeev Bansal...pdf

Pág. 860

38. Geometric Intersection
David M. Mount

Pág. 862 Figure 38.2.1
Hopcroft's Problem

F) Books_71_2020/Handbook of Engineering
Electromagnetics/Rajeev Bansal-Handbook...pdf

Pág. 12(22) 1.2.2 Maxwell's Equations
in Differential Form and the Continuity Equation.

G) Books_71_2020/Chemical Engineering Fluid Mechanics/
1498724426.pdf
→livro todo.

H) Books_71_2020/The Finite Element Method for
Boundary Value Problems Mathematics and
Computations/
1498780504.pdf

Pág. 584(611) 8.11 Interpolation functions for
3D elements.
Hexahedral 3D elements Figure 8.50
Pág. 601 Figure 8.62 Pascal Pyramid;
Pág. 602 Figure 8.63 Tetrahedron element with
Volume coordinates.

—

Jandaia

data 26.04.2020

15) 20200426

A) Books_71_2020/
Practical Linear Algebra A Geometrical Toolbox/
Practical linear algebra ...(2014...). pdf

Pág. 285(302)

13.2 Vector Norms

Figure 13.2
Vector norms: outline of the unit vectors for the 2-norm is a circle, ∞-norm is a square, 1-norm is a diamond.

Pág. 243(260)

12- Gauss for Linear Systems.

Pág. 198(215)
Figure 9.8 - Combining 3D rotations.

Obs: Livro todo.

B) Books_71_2020/
Mechanical Engineerings' Handbook_4_Volumes/
Mechanical Engineers Handbook. Vol. 4...pdf

Pág. 54(70)

3.4 Fluid Forces on Curved Surfaces.

date 26 04 2020

S T Q Q S S D ✗

Pág. 55 (71) Table 5

Properties of Areas and Volumes

C) Books_71_2020/
Introduction to Topology Second
Edition (PDF)/
Understanding Topology_A Practical...
pdf

Pág. 167 (173)

Chapter 5 Surfaces.

Pág. 174 (180)

Figure 5.9 $I^2 \times I = I^3$

$I^3 \times I = I^4$

The tesseract is a cube extended through a fourth dimension.

D) Books_71_2020/
Skyrmions Topological Structures.../
Skyrmions ... pdf

Pág. 04 (25)

Figure 1 1

curvilinear polygons on the sphere.

obs: Ver todo livro.

26 04 2020 X

Important: A Skyrmion is a tiny region of reversed magnetization — quasi particles since they are not present except in a magnetic state and also give rise to physics that cannot be described by Maxwell's equations.

E) Books_71_2020/
Introduction to Topology Second Edition (PDF)/
0486406806.pdf

Pág. 16(24)

§4. Products of Metric Spaces

Important:

4.2 A ball, with metric

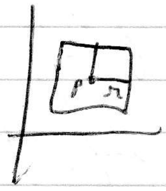

4.3 A ball, with metric

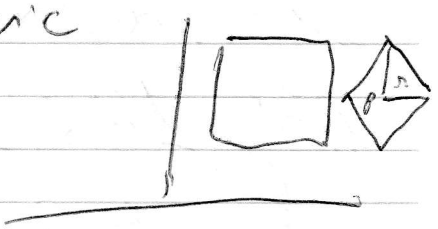

4.2 = square

4.3 = diamond

151) 20200426

A) Books_72_2020/ Flat and Curved Space-Times/
Foundations of Space and Time_
Reflections on ... pdf

data 26.04.2020

S T Q Q S S [✓ D]

Pág. 09(24)

L_{matt} is the matter Lagrangian depending on the metric and some matter variables ψ_A.

(I will assume L does not involve derivatives of curvature tensor, to simplify the discussion).

B) Flat and Curved Space-Times, Second Edition (2001).pdf

Pág. 49(57)

3. Measurements in flat space-times

Pág. 122 (130)

4. The Lorentz transformation and invariant interval.

Pág. 126(134)

Fig. 4.2

↑The relationship between the two coordinate system.

C) Relativistic cosmology (2012 Cambridge).pdf

Part 2 Relativistic Cosmological Models

26.04.2020

4. Kinematics of cosmological models

↳ Pag.(89) 73

Pág. 76(92)

4.4 Projection to give three-dimensional relations.

→ Orthogonal projection

→ Orthogonal spatial derivatives

D) The large scale structure of space-time (1973 ...)...pdf

Pag. 47 (55)

2.8 The volume element and Gauss' theorem.

Pag. 78 (86)

4. The physical significance of curvature

Pág. 117 (125)

5. Exact solutions

5.1 Minkowski space-time

E) Theoretical and observational Cosmology (1999 ...)..pdf

26.04.2020

Pag. 71(91)

7.1 Gauge Problem

Pag. 72(92)

Gauge Invariance Lemma

Pag. 73(93) ✡

[the proof is straightforward

Pág. 185(205) —

4 Gravitational lenses in Cosmology

4.3 The linearized equation of geometric optics

Pág. 189(209)

F) The renaissance of general relativity and cosmology ... pdf

Pag. 287(297)

Topology and topology change in general relativity.

Pág. 288(298)

2.1 The Einstein-Rosen Bridge

Pág. 289(299)

data 26.04.2020
X

Lorentzian metric g_L

→ Lorentzian metric g_L on a compact 4-manifold M...

G) Books_72_2020/
Spacecraft Trajectory Optimization.../
Bruce Conway - Spacecraft Trajectory optimization...pdf

Pag. 40 (54)

3.2 Transcription Methods

3.2.1 A basic collocation Method (Using Hermite Polynomials)

Figure 3.1 Trajectory optimization problem broken into phases.

---

152) 20200427

A) Books_72_2020/
Advanced Engineering Mathematics/
H.K. Dass - Advanced Engineering Mathematics...pdf

Pag. 128 (144)

2.16 Surface Area.

B) Books_72_2020/
Dirichlet Branes and Mirror Symmetry/
Dirichlet Branes...pdf

27.04.2020

Pág. 616 (626)

8.4 The elliptic curve.

C) Books_72_2020/
Complex Variables and Numerical Methods/
Ravish R Singh - Complex Variables.pdf

Pág. 20

1.3 Geometrical Representation of Complex Numbers (Argand's Diagram)

Obs: Ver livro todo!

D) Books_72_2020/
Nonlinear optics, 4th Edition/
Nonlinear optics, 4th ed.pdf

Pág. 11 (34)

Self-Focusing

Look: Figure 1.2.6.

E) Books_72_2020/
Fomingeo_fan_edu/
Tratado0508.pdf

Pág. 15 Bissectrizes dos ângulos de um triângulo.

27 04 2020

F) Books-72-2020/
Forumgeo-fau-edu/
FG2004Volume4.pdf

Pag. 5 Antiparallels and Concurrent Euler Lines.

Figure 1 Thébault's theorem on the ~~[illegible]~~ concurrency of Euler lines.

Look: Figure 2
Figure 3
Q(K) and the second Lemoine circle.

Obs: Leia artigo todo. (Pag. 24

153) 20200427

A) Books-72-2020/
Projective-Geometry/
Geometry of Cuts and Metrics.pdf

Pag. 86/96

6.4 An Example: the Spherical Distance Space

$x, y \in S$

data 27.04.2020

$$r \cdot \arccos\left(\frac{(x-c)^T (x-y)}{r^2}\right)$$

(6.4.1)

↳ The quantity (6.4.1) is known as the spherical distance (or great circle metric) between x and y.

Pag 87(97)

Figure 6.4.3

H(x) Δ H(y) is the shaded region.

Pag. 90(100)

Let us now indicate the EUCLIDEAN distance can be approximated by spherical distance.[9]

9 → This fact was already observed in Kelly (1975).

B) Algorithms in Real Algebraic Geometry (2003...)...pdf

Pag. 430(434)

date 27.04.2020

[✓ S] T Q Q S S D

11.6 Topology of Curves

↳ Cylindrical Decomposition Algorithm

L) Heights in diophantine geometry.pdf

Obs: Ver livro todo!!!

D) Global Aspects of Complex Geometry (2006...).pdf

Pág. 01(07)

Complex Surfaces of General Type: Some recent progress

Pág. 04(10)

Definition 2: A smooth surface S is called minimal (or a minimal model) if it does not contain any exceptional curve E of the first kind (i.e. $E \simeq P^1$, $E^2 = -1$).

Pág. 417(419)

Special Families of Curves, of Abelian Varieties, and of Certain Minimal Manifolds over Curves.

Pág. 441(443)

4 Teichmüller Curves and

27.04.2020

their Variation of Hodge Structures.

Pag. 445(447)

5 the only Teichmüller-Shimura Curve

E) Alain Connes - Noncommutative geometry (1994 ...) .pdf

Pag. 34 (34)

the metric aspect of noncommutative geometry.

Pag. 552 (552)

Chapter 6 the metric aspect of noncommutative geometry.

Pag. 40

1. Heisenberg and the Noncommutative algebra off Physical quantities.

Figure 1. Hydrogen Spectrum.

F) IMPORTANT: ✡

Books_72_2020/
Projective Geometry/

Four-Manifolds, geometries

data 27.04.2020

and Knots (2002...)...pdf

Pag. 246 (260)

12.8 The case $\pi = \mathbb{Z}/4\mathbb{Z}$

Pag. 247 (261)

Theorem 12.6

Let $M = S^2 \times S^2/\langle\tau(I,-I)\rangle$.

Then the natural map from π to $Aut_\pi(\Pi)$ is an isomorphism, while $G_\#(M)$ has order ≤ 2.

G) The geometry of four-manifolds... pdf

Pag. 375 (389)

10. The Differential Topology of Algebraic Surfaces

154) ~~280~~ 20200428

A) Books_73_2020/
A Visual Introduction to the Fourth Dimension.../

A1) Geometry Proofs Essential Practice Problems...pdf

28.04.2020

A2) Essential Modern Physics *Study Guide...pdf

Pag. 35 Relativistic Mass

B) Books_73_2020/
Geometry II (Universiteit)/

A Panoramic View of Riemannian Geometry (2003...)...pdf

Pag. 12(33)

1.3.2 Curvature

Osculating circle

to the curve at $m=c(t)$.

Pag. 13(34)

Fig. 1.15 Finding tangent line, concavity and curvature.

Pag. 15(36)

Fig. 1.18 Planar calculus of variations.

C) Books_73_2020/
Plasma Astrophysics_Fundamentals and Practice_2_Volumes/

28 04 2020

C1) 1461442826.pdf (Volume 1)
Pág. 148(170) ✡
Fig. 8.1 The trajectory of a light particle with mass m_1 near a heavy particle with mass m_2.

C2) 146144294X.pdf (Volume 2)
Pág. 72(92)
Fig. 4.2 The model for the magnetic field of four sunspots of pairwise opposite polarity.

D) Books_73_2020/
The Geometry of Hamilton and Lagrange Spaces/
The Geometry of Hamilton and Lagrange Spaces...pdf

E) Books_73_2020/
Pleasures of Pi, e and Other Interesting Numbers/
Pleasures of Pi, e ...pdf ✡

F) Books_73_2020/
The Bloch-Kato Conjecture for the Riemann Zeta Function/
The Bloch-Kato Conjecture...pdf

28.04.2020

G) Books_73_2020/
Symmetries of Partial Differential Equations/
Symmetries and Overdetermined Systems of Partial Differential Equations....pdf

Pág. 339(346)

Metric connections in projective differential geometry

H) Books_73_2020/
Geometry from Isometries to Special Relativity/
3030421007.pdf

Obs: Livro todo!

Pág. 23(35)

Chapter 2 Sphere

2.1. The Sphere S^2 in $\mathbb{R}^3$

$$S^2=\{(x,y,z)\in\mathbb{R}^3 / x^2+y^2+z^2=1\}.$$

Pág. 24(36)

Fig. 2.1 Spherical distance $d_{S^2}(p_1,p_2)$ between the two points p_1 and p_2 on S^2.

data 29.04.2020
S T [✓ Q] Q S S D

155) 20200429

A) Einstein Math

Books_73_2020/
Einstein Equations Physical and Mathematical Aspects of General Relativity.../

Einstein Equations...pdf

Obs: Livro todo!

B) Books_73_2020/
Mathematical Analysis of the Navier-Stokes Equations.../

3030362256.pdf

Pág. 11 (18)

The Vector Value Setting.

C) Books_73_2020/
Geometric Partial Differential Equations/

→ Diretório inteiro!

C1) Geometric Mechanics on Riemannian manifold...pdf

Pág. 251 (257)

29 04 2020

Chapter 12 Mechanical Curves
(capítulo todo)

C2) Geometric_Partial_Differential_Equations_-_Part_I.pdf

Pág. 07(14)

2. Calculus on surfaces

D) Books_73_2020/
Engineering Mathematics for
Semesters//

D1) 9339219643.pdf

Pág. 359

Chapter 7

Vector Differential and Integral Calculus.

D2) 9385880500.pdf

Pág. 127 – 256

Chapter 3 Complex Variables and Calculus.

Chapter 14 Calculus of Variations (831-900)

obs: Dois livros bons para consultas!

— 202004280040

29 04 2020
S T [✓ Q] Q S S D

156) 20200429

A) Books_74_2020/
Ideas of Quantum Chemistry/

A1) Ideas_of_Quantum_Chemistry_vol_1.pdf

Pag. 41(73)
1.3 The Heisenberg uncertainty principle.

Pag. 47(79)
1.5 How to disprove the Heisenberg principle? The Einstein–Podolsky–Rosen's recipe.

Pag. 72(104)
2.1 Symmetry of the nonrelativistic Hamiltonian and the conservation laws.

Pag. 401(429)
7.1.1 Coordinate system and Hamiltonian

Pag. 457(484)
Orbital Model of Electronic Motion

data 29.04.2020
S T [✓ Q] Q S S D

in Atoms and Molecules

A2) Ideas_of_Quantum_Chemistry-Vol_2.pdf

Pag. 01(25)

Electronic Orbital Interactions in Periodic Systems

Pag. 101(125)

2.4 Exponentially correlated Gaussian functions

Pag. 463(486)

6.5 Reaction path Hamiltonian method.

B) Books_74_2020/
Field Guide to Quantum Mechanics/
Field_Guide_to_Quantum_Mechanics.pdf

Pag. 43(63)

Schrödinger, Heisenberg, and Interaction Pictures.

C) Books_74_2020/
Higher Engineering Mathematics by Bandary Ramana/

29.04.2020

S T [✓ Q] Q S S D

007063419X.pdf

Pag. 179 Chapter 5

Curve Tracing

Pag. 1228 Chapter 36

Analytical Solid Geometry

Obs: Livro todo!!!

D) Books_74_2020/
Higher Engineering Mathematics by
Banday Ramana/

1138673579.pdf

Pag. 140 Section B

Geometry and trigonometry

Pag. 270 Section D

Complex numbers

Obs. Livro todo!!!

E) Books_74_2020/
Electric Field Analysis/

Electric Field Analysis.pdf

Pag. 61(84)

29.04.2020

Pág. 39 (62) Chapter 2
Gauss's Law and Related Topics

Pág. 59 (82) Chapter 3
Orthogonal Coordinate Systems

Pág. 219 (242) Chapter 10
Sphere or Cylinder in Uniform External Field.

F) Books_74_2020/
La Matematica vol. 1. I tempi e i luoghi/

La matematica....pdf

Pág. 189 (205)
5.2 La trigonometria

Teorema di Menelao.

Se CE, CF, DF, DB sono archi di cerchi massimi su una sfera.
(circulos)

Pág. 190 Teorema del seno.

Pág. 402 (418) Leibniz, consigliere dei principi

data 29.04.2020

S T [✓ Q] Q S S D

A) 157) 20200429 2246

A) Books_74_2020/
Magazine/

Spectrum_Mathematics_-_November_2016.pdf

→ family of circles. - Pág. 03

properties of triangles. - Pág. 66.

Pág. 13 "0" is either positive or negative.

Pág. 42 Cube Roots and nth Roots of Unity

→ nth roots of unity is very important topic of complex numbers.

B) Books_74_2020/
Magazine/

Spectrum_Mathematics_-_July-2016.pdf

→ Tangent to a circle. Pág. 42-49.

Look pages 40 e 41.

Trigonometric Equations and Identities. (resume)

Pág. 54 - Inverse Trigonometric Functions

data 01.05.2020
S T Q Q S S D

158) 20200501

A) Books_75_2020/
Important/
Two Millenia of Mathematics...pdf

Obs: Pi
Ler todo livro!

B) Books_75_2020/
Important/
9353162807.pdf

→ Improper integrals

C) Books_75_2020/
Important/ Hagihara/ ✡

⇒ 5 Volumes.

D) Books_76_2020/
Generation of Surfaces Kinematic
Geometry of Surface Machining/

Radzevich, Stephen - Geometry
of Surface ...pdf

Pág. 67(88) Chapter 3
Early Works in the Field of Contact
Geometry of Surfaces

Curvature.

data 01.05.2020
S T Q Q [✓ S] S D

Pág. 70(91)

Fig. 3.1 Local configuration of two smooth regular part surfaces P_1 and P_2 at a point K of their contact

Obs: Livro todo!!!

Pág. 72(93)

Fig. 3.2 Specification of the angle μ of the surfaces P_1 and P_2 local relative orientation

Pág. 73(94)

Fig. 3.3 Local relative orientation of the contacting part surfaces P_1 and P_2, considered in a common tangent plane. ✡

Important:

✡ Aplicar estes conceitos no estudo da curvatura do Universo devidamente projetado em uma superfície plana.

Pág. 83(104)

3.5.3 "Dupin Indicatrix" at a Point of Surface of Relative Normal Curvature ✡

data 01.05.2020

Consider an intersection of the surface of relative normal curvature by a plane, parallel to the tangent plane at the point K, but only at a small distance away from it.

Veja capítulo todo!

Pag. 131 (150)

6.3 Relative Characteristic Curve.

Pag. 133 (152)

Fig. 6.3 "relative indicatrix" at a point of contact K of two part surfaces P_1 and P_2.

Pag. 138 (157)

Possible kinds of Contact of two smooth regular Part Surfaces in the First order of tangency

Fig. 7.1 a) b) c) ☆

Pag. 243 (257)

Coordinate Systems and Linear

data 01.05.2020
S T Q Q [✓ S] S D

Transformations

Pag. 277(291)

B.5 Impact of the Coordinate System Transformations on Fundamentals Forms of the Surface.

Obs: Veja livro todo!!!

E) Books_76_2020/
Geometric Folding Algorithms/
Geometric Folding Algorithms...pdf

Obs: Livro todo!!!

Pag. 30(44)
3.1.3 Peaucellier Linkage

Pag. 32(45)
Fig. 3.2 ✡

Pag. 299(311) Part III Polyhedra

Pag. 301(315) 21.2 Curvature

Pág. 302(316) ✡
Figure 21.4 →

data 01.05.2020

The curvature at a point is* 1/r, where r is the radius of the OSCULATING CIRCLE.

Figure 21.5

Spinning a plane about N through p ∈ S yields (produz) the principal curvatures at p.

The curvature of the curves for plane A and B are positive at p, but negative for plane C.

Pág. 287 (301)

19.2 Huzita's Axioms and Hatori's Addition

Box 19.1: Abe's trisection

Important: three parallel lines

Estudar!!! com Pappus ✡

Fig. 19.4 Abe's trisection analyzed.

---

01 05 2020

F) Books_76_2020/
Elliptic Curves and Big Galois Representations/

Elliptic Curves and Big Galois Representations...pdf

obs: Livro bom!

Pag. 222(233) Chapter X

Two-Variable Iwasawa Theory of Elliptic Curves.

159) 20200502

A) "A line is considered a circle with infinite radius and zero curvature".

"In fact the work was of such quality that Weingarten received a prize for work on the lines of curvature of a surface in 1857".

"A formula for the line integral of the geodesic curvature along a closed curve is known as the Gauss Bonnet theorem".

"That is, in this case, the Riemann curvature tensor has 4x4x4x4=256 components!"

data 02.05.2020

S T Q Q S S D

"A deeper characterization of a helix is that is the unique curve in 3-space for which the ratio of curvature to torsion is a constant, a result know as Lancret's theorem."

"Some of his work on physical topics relates to his non-Euclidean geometry for he examined how the gravitational potential as given by Newton would have to be modified in a space of negative curvature".

"The local curvature could then be computed from the coefficients of this quadratic surface".

"He showed that spaces of constant curvature could have several different topological structures".

"Surfaces of constant mean curvature or constant Gaussian curvature are now called Weingarten surfaces"

(surfaces)

"The value assigned for each cell is the value of Gaussian curvature at the cell centroid"

(centroid).

02 05 2020

"Gaussian curvature is positive for surfaces like a hemisphere, which are either concave or convex in all the directions"

B) B1) Books_2020_12/
Books_75_2020/
PI/

Celestial Mechanics/

B2) Books_76_2020/

Einstein the Closet Geocentrist_
Videos/

B3) Books_76_2020/
A Treatise on Solid Geometry/

A Treatise on Solid Geometry.pdf

Pag. 67(97)

104. To find the distance between two points whose (de quem) equations are given

105. (Pag. 69(99))

We deduce the expression for the square of the distance.

106. (Pag. 69(99)) IMPORTANT:

data 02.05.2020

106. To find the relation between the co-ordinates of a plane

→ usar para triangular um ponto que foi projetado da superfície curvada para o plano (superfície plana).

IMPORTANT:

O ponto A é o referencial inicial para a projeção em uma superfície.

O ponto A pode estar sobre a curvatura.

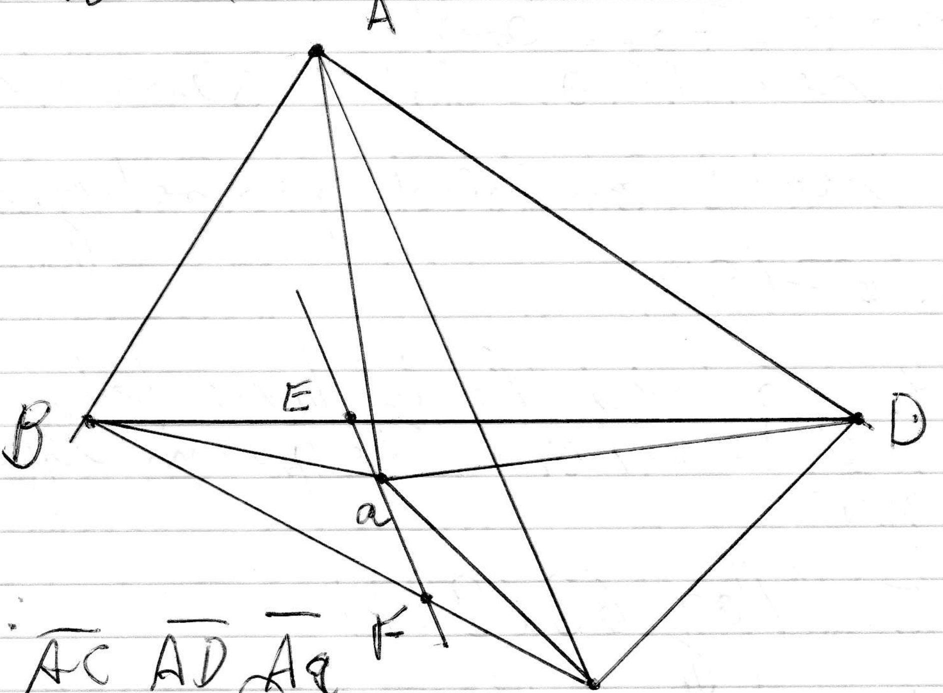

Obs: $\overline{AB}$ $\overline{AC}$ $\overline{AD}$ $\overline{Aa}$ são projeções a partir do ponto A para a superfície triangular $\triangle BCD$.

02.05.2020

Pág. 70 (100) Nota de rodapé

* The equation of a straight line, referred to trilinear co-ordinates,) is

$$\alpha \frac{\varpi_1}{p_1} + \beta \frac{\varpi_2}{p_2} + \gamma \frac{\varpi_3}{p_3} = 0,$$

where

$\varpi_1, \varpi_2, \varpi_3$ are the perpendiculars, estimated in the same direction, from the angular points of the fundamental triangle upon (em cima de) the straight line, and

p_1, p_2, p_3 those upon the sides.

Now taking EF for the straight line, and BCD for the fundamental triangle, for a, a point in that line,

$$\frac{\alpha}{p_1} : \frac{\beta}{p_2} : \frac{\gamma}{p_3} :: aCD : aDB : aBC,$$

$$\therefore \varpi_1 . aCD + \varpi_2 . aDB + \varpi_3 . aBC = 0,$$

whence (de onde) the equation in the text.

data 02 05 2020

Pag. 71 (101)

Item 107. To find the relations between the coordinates of plane an infinite distance.

outra fonte:

Books_77_2020/
Treatise-of-Solid_Geometry_
Frost/

→ Diretório todo!!!

outro livro (Frost)

An Elementary Treatise on Curve Tracing (2004...)...djvu

c) Gaussian curvature

Source: youtube.com/playlist?list=PLF56602BAC6932237E

Copy: Books_76_2020/
Einstein The Closet Geometrist-Videos/
a) Relativity_Videos/
b) Gauss/ → Relativity 7b. ✡

data 02 05 2020

S T Q Q S [✓ S] D

D) Books_2020_12/
Books_75_2020/
PI/

D1) 2002.12137.pdf
The calculation of p_n and $\pi(n)$
Simon Plouffe
Feb 23. 2020.

Source: plouffe.fr/Simon Plouffe.htm

D2) Une formule pour les nombres premier II.pdf
Le calcul de p_n et $\pi(n)$
Feb 21, 2020

Source: plouffe.fr/NEW/Une formule pour les nombres premiers II.pdf

Obs: outros arquivos:
plouffe.fr/simon/

New file: Books_77_2020/
Simon_Plouffe/
0912.0303.pdf

Source:
plouffe.fr/simon/0912.0303.pdf

data 02.05.2020

On the computation of the n^{th} decimal digit of various transcendental numbers

Pag. 02

The formula for π used is

$$\pi + 3 = \sum_{n=1}^{\infty} \frac{n 2^n n!^2}{(2n)!}$$

E) Einstein Space-time theories

Source: https://www.nasa.gov/mission_pages/gpb/

→ Nasa's Gravity Probe B confirms two Einstein Space-time theories.

other: http://einstein.stanford.edu

https://arxiv.org/abs/1105.3456

https://iopscience.iop.org/journal/0264-9381/page/Focus-issue-on-Gravity-Probe-B

02.05.2020

Press Conference:
https://www.youtube.com/watch?
V=SBiYOFn1ze4

→ Einstein Passes Tests by NASA's
Gravity Probe B

Copy: Books_76_2020/
Einstein the Closet
Geocentrist_Videos/

Einstein Passes...mp4

---

160) 20200502

A) Gaussian Quadrature Formulas

Source: Books_77_2020/
Gaussian_Curvature/

A1) Gaussian quadrature formulas...djvu

A2) Gaussian Harmonic Analysis
(2019)...pdf

B) Books_77_2020/
Gravity/

B1) Beyond Einstein Gravity...pdf

B2) IMPORTANT:

02.05.2020

B2) Gravity, a Geometrical
Course
2 Volumes
Pietro Giuseppe Frè

B2.1) Gravity, a Geometrical Course_volume 1... pdf

B2.2) Gravity, a Geometrical Course, volume 2... pdf

B3) A. Zee - Einstein gravity a nutshell (2013...) ... pdf

Pag. 25(48)

Part I From Newton to Riemann: Coordinates to Curvature

B4) Foundations of Space and Time_Reflections... pdf

B5) Equations of Motion in Relativistic Gravity (2015...) ... pdf ✡

B6) Gauge_Gravity_Foundations... pdf

B7) Gravity's shadow... pdf

B8) Important:

02 05 2020

B8) Quantum Gravity (1997...)...pdf

Pág. 542(571)

"On the observable properties of cosmic background radiation in the Universe with negative curvature.

B9) Gravity and Strings (2015...)...pdf

Pág. 09(36)

1.3 Metric Spaces.

1.3.1 Riemann-Cartan spacetime U_d

1.3.2 Einstein-Weyl spacetime EW_d

1.3.3 Riemann spacetime V_d

Pág. 166(193)

4.5 Gravity as a gauge theory

Pág. 323(350)

12.2 The Einstein-Maxwell System

B10) Semiclassical and Stochastic Gravity – Quantum Field Effects on Curved Spacetime (2020)...pdf

✱. Important: 615 pages.

data 02 05 2020

S T Q Q S [✓ S] D

Pág. 337 (352) Chapter 10

The Einstein-Langevin Equation

Pág. 364 (379) Chapter 11

Metric Fluctuations in Minkowski Spacetime.

Obs: Os dois capitulos tem vinculo (Langevin Equation)

Pág. 467 (482) Chapter 15 e seguintes

Stress-Energy Tensor Fluctuations in de Sitter Space.

Pág. 519 (534) Chapter 17

Riemann Tensor Correlator in de Sitter

B11) Quantum Field Theory in Curved Spacetime – Quantized Fields and Gravity (2009...)...pdf

Obs: Livro todo!!!

Pág. 01 (19) Chapter 1

Quantum fields in Minkowski spacetime

Pág. 36 (54) Chapter 2

Basics of quantum fields in curved spacetimes.

0205 2020

[illegible] [illegible] C) Modern Physics.

Source: Books_77_2020/
Modern Physics, by KAUR AND PICKRELL/

C1) 9332901813.pdf

Obs: Livro consulta. Ler todo!

Pág. 17(43) 1.7 Lorentz Contraction (Length Contraction)

The measured distance also depends on the reference frame.

If we consider Galilean transformation, then the measured size an object remains unaffected by the relative motion of observers. But for

LORENTZ TRANSFORMATION,

length (length) varies as time interval varies.

Pág. 18(44) 1.8 Concept of Space-Time (Minkowski Space)

Pág. 20(46) Fig 1.9 Space-Time diagram.

C2) Geometric Mechanics on Riemannian Manifolds...pdf
→ Applications to Partial Differential Equations

02.05.2020

Pág. 251(257) Chapter 12

Mechanical Curves —

D) Joseph Edwards_ A Treatise on the Integral Calculus with Applications... ~~Volume~~ 1...pdf
2 pdf

Source: Books_77_2020/
Calculus with Applications/

E) Books_77_2020/
Understanding Analysis/

Understanding Analysis (2015...)...pdf

Pág. 85(97) Chapter 3

Basic Topology of R

→ the Cantor Set

Pág. 215(224) Chapter 7

the Riemann Integral

F) ARQ_IMPORTANT_05/
Geométrie/

→ diretório todo!

→ Traité de Geométrie (2 Volumes)

obs. Está incompleto. Vou consertar!!! ok agora

data 02 05 2020

Traité de Géométrie Descriptive
3 Volumes.

161) Books_77_2020/
Geometric_Transformations_for_3D/

A) Michael Mortenson - Geometric Transformation....pdf

Obs. Li o todo!

B) Books_77_2020/
1Fuzzy Controller Design theory and Applications/

0849337747X.pdf

Pág. 11(24) Figure 2.2

Typical shapes of membership functions:
1- triangular, 2- trapezoidal,
3- Gaussian, 4- bell- shaped,
5- singleton.

Pág. 12(25)

Due to the fact that fuzzy sets can be either uniform or not, symmetrical or not, bounded or not - a good measure of their geometrical shape is the centroid the point within a fuzzy set designating its center of gravity (COG).

data 04.05.2020

162) 20200504 IMPORT:

Books_78_2020/
ARQ_BASICS/01/

A1) Mathematics for the IB Diploma...pdf
Pág. 400 - 13C Distances

A2) 0199152268.pdf → conceitos explicados.

A3) Jacaranda Math Quest 10.pdf
Pág. 85(99) 3.5 the midpoint of a line segment

A4) geometria.pdf → Important

A5) EAAN_20110114.pdf

The thirteen Books of Euclid's Elements

A6) A Treatise on Solid Geometry.pdf
→ Livro todo!
Pág. 172(202) Chapter XII Diametral Surfaces.

A7) 0071544127.pdf

Obs. ARQ_BASICS_20200504.7z
→ 100 Gb

O arquivo foi criado independentemente, resultado da compilação dos arquivos mais importantes entre BOOKS_2020_04 até BOOKS-2020-08. Checar demais pastas!!!

Criei a continuação: ARQ_BASICS_02/
(independente)

data 04.05.2020
[✓ S] T Q Q S S D

B) ARQ_BASICS_02/

B1) 11/ Higher Engineering Mathematics/
007063419x.pdf

obs: ARQ_BASICS_02 = 60 Gb.

ORGANIZANDO ARQUIVOS:

A) ARQ_BASICS/02/

A1) A Vector Space Approach to Geometry/
Mathematics - A Vector Space ... epub

A2) Advanced Calculus .../
Advanced Calculus... epub

A3) Calculus/
→ Diretórios todos. Alguns livros copiados para ARQ_BASICS_03/01/

A3.1) Analytic Solid Geometry by Shanti Narayan.pdf

Pág. 10(26) - 9(25) - 1.8

Projection on a Straight line
(Straight)

A3.2) Analytical Geometry 2D and 3D ... epub

A3.3) Angus E. Taylor ... - Advanced Calculus - Wiley.pdf

→

data 04.05.2020

Copy editáveis:
a) Advanced_Calculus_Taylor.pdf
b) Advanced_Calculus_Taylor.djvu

Obs: Advanced_Calculus_Taylor.pdf
Pág. 84(100) → 10

The number $\sqrt{3}$ is defined as the positive number c such that $c^2 = 3$.

A3.4) Calculus the classic by Earl W. Swokowski...pdf

A3.5) Direct Methods in the Calculus of Variations (Applied...)...pdf

A3.6) Functions of Complex Variable...pdf

A3.7) Geometry.pdf

A3.8) H. Jerome Keisler.... Infinitesimal...pdf

A3.9) Higher Engineering Mathematics by B.S. Grewal.pdf

A3.10) Integral Calculus 3D Geometry and Vector Booster...pdf

A3.11) kupdf.net analytical-geometry-2d-and-3d-p-n-vittal.pdf

A3.12) Vector Calculus by Krishna Series.pdf

04052020

A4) Calculus and Analysis in Euclidean Space/
Shurman_CalculusAnalysisEuclidean Space.pdf

Pág. 23(35) Chapter 2 - Euclidean Space

A5) Calculus in 3D Geometry, Vectors, and Multivariate Calculus.../

A5.1) Algebra and Trigonometry...pdf

A5.2) Calculus in 3D Geometry....pdf

Pág. 67(78) - Chapter 2
Curves and Vector-Valued Functions of One Variable

A5.3) Fundamentals of Mathematical Differential Calculus.pdf

A.5.4) Geometry, the Line and the Circle.pdf

IMPORTANT:

ARQ_BASICS e ARQ_BASICS_02 contém livros escolhidos entre os diretórios BOOKS_2020_01 à BOOKS_2020_12.

obs: Estas novas pastas resumem os melhores livros e que têm mais relação com a minha pesquisa. Os demais livros serão fonte de referência

data 04.05.2020

[✓ S] T Q Q S S D

163) Pappus D Alexandrie

Source: https://archive.org/details/PappusDAlexandrieEd.PaulVerEeckelLaCollectionMathmatique/

Obs.: 1010 pages. Pag. 400(525)

Identifier-ark

ark:/13960/t5s80978s

Other: libarch.nmu.org.ua/handle/GenofondUA/22654

Obs: Oeuvres Mathématiques D'Évariste Galois

https://gallica.bnf.fr/ark:/12148/bpt6k9800489w/f11.image

Obs: Les trois livres de Porisme D'Euclide

https://gallica.bnf.fr/ark:/12148/bpt6k6214433z/f9.image

Others: Historia Mathematica Volume 8, Issue 4, November 1981,

-

data 04.05.2020

Pages 417-438.

How trisections of the angle were transmitted from Greek to islamic geometry

Jan P Hogendijk

https://doi.org/10.1016/0315-0860(81)90051-3

Source: www.sciencedirect.com/science/article/pii/0315086081900513

other: caphi.uni-nantes.fr/IMG/pdf/Descartes-Geometrie.pdf

→ Descartes La Geometrie 1637.

-

164) 20200505

A) Books_79_2020/
Great Physics the Life and Times.../

A1) Great Physics ...pdf

Pag. 203(215) Chapter 14

Adventure in thought
Albert Einstein

Pag. 208(220)

Doctrine of Space and Time

data 05.05.2020

Pag.209) Whereas Lorentz had struggled to explain the invariance of the speed of light with a constructive theory that hypothezid motion-dependent molecular forces, Einstein bypassed all the complications by simply promoting the constancy to a postulate.

For Lorentz and his contemporaries, it was a problem, for Einstein a principle.

A2) The Quantum Physicists And an Introduction to their Physics (1970) in pdf.

Pag. 30(43)

Conservation of momentum

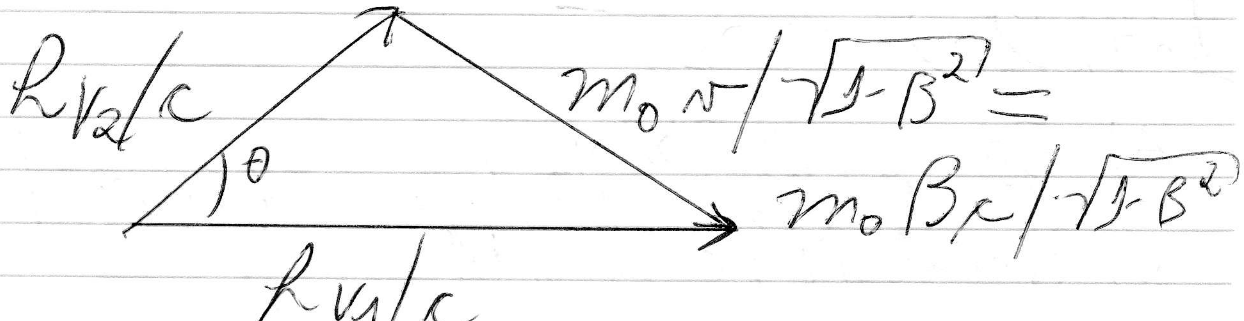

Conservation of momentum requires that the incident photon vector be equal to the vectors representing the scattered photon and the recoil electron:

and this (with the law of cosines)

05 05 2020

leads to the equation:

$$(m_0\beta c/\sqrt{1-\beta^2})^2 =$$
$$(h\nu_1/c)^2 + (h\nu_2/c)^2 - 2(h\nu_1/c)(h\nu_2/c)\cos\theta$$

Obs:
That $h\nu/c$ is a measure of photon momentum can be seen by the equation two Einstein equations for total energy,
$E = h\nu$ and $E = mc^2$: $h\nu = mc^2$.

B) Books_79_2020/
Traité de Coniques/

B1) Gaspard Monge Founder of "Constructive Geometry".pdf

Obs: look at: referentes
(find: gallica.bnf.fr)

B2) 517.pdf — Álgebra

B3) 522.pdf — Bolas de futebol, Arquimedes e ondulettes

B4) 523.pdf - Triângulos Belos

B5) 524.pdf - PI faraônico

B6) 525.pdf - As bolas da FIFA não são

data 05.05.2020

S [✓ T] Q Q S S D

grande coisa.

B7) 527.pdf → Arquimedes e o Volume da Esfera

B8) 528.pdf → O conjunto de todos os conjuntos não existe.

B9) 529.pdf → Números de Bernoulli

B10) 530.pdf → Os "Modelos de Olivier".

B11) 531.pdf → Problemas antigos com números perfeitos.

Important:

B12) 001_Traité_de_géométrie_descriptive... Javary... pdf

obs: 337 pages.

B13) 002_Traité_de_géométrie_descriptive_Javary...pdf

obs: 816 pages.

B14) Traité de géométrie supérieure... Chasles..pdf

obs: 699 pages.

C) Books_79_2020/
Bernhard Riemann 1826-1866.../

data 05 05 2020

Bernhard Rieman 1826-1866...pdf

→ Turning Points in the Conception of Mathematics.

Obs: Ver todo livro!

165) 20200506

Books_79_2020/
Introduction to Continuum Mechanics.../

A1) Introduction to continuum mechanics ... pdf

Pag. 22 (26)

2.15 Orthogonal Tensors

Obs. Livro todo!

Pag. 45 (49)

2.26 Tensor-Valued Functions of a Scalar

A2) Arithmetic algebraic geometry (2001) ... pdf

Pag. 149 (158)

Introduction to Serre's conjecture

↳ elliptic curve

data 06.05.2020
S T [✓ Q] Q S S D

Pág. 129 (138) Chapter 6
Some other conjectures
6.1. Integral points on elliptic curves

B) Books_79_2020/
Phase Transitions in Materials/
Phase_Transitions_in_Materials.pdf

Pág. 133 (154) Chapter 6 Energy
6.1 Atomic Schrödinger Equations and Formalism
Box 6.1 Dirac Notation

C) Books_79_2020/
Pythagoras' Legacy Mathematics in Ten Great Ideas/
Pythagoras_Legacy.pdf

Obs: Muito bom!

Pág. 82 (93) e: A very special number
Pág. 93 (104) i: Imaginary numbers
Pág. 75 (86) Pascal's Triangle

Obs: Pág. 82(93) → Gauss

Jandaia

06 05 2020

You have no idea how much poetry there is in the calculation of a table of logarithms

Carl Friedrich Gauss (1777-1855)

Pag. 93(104)-

Imaginary numbers are a fine and wonderful refuge of the divine spirit, almost an amphibian between being and non-being.

Gottfried Leibniz (1831-1881)
?

D) Books_79_2020/
Superstrings and other things.../

Superstrings_and_other_things.pdf

Pag. 419(440) Appendix C

Nobel Prize Winners in Physics

E) Books_79_2020/
MAA/

E1) Thinking Geometrically_A Survey of Geometries.pdf

obs: Livro todo!

Pag. 317(342) Chapter 7 Projective Geometry

data 06.05.2020
S T [✓ Q] Q S S D

Pág. 318 (342)

Metrical geometry is thus (portanto) a part of [projective] geometry, and [projective] geometry is all geometry. Arthur Cayley.

Pág. 320 (345)

Definition. A perspective with respect to a point T is a mapping of the points (~~U_i~~) U_i on one line to the points V_i on another so that (de modo a) V_i ~~on another a~~ is the image of U_i if and only if

U_i, V_i, and T are on the same line (Figure 7.3)

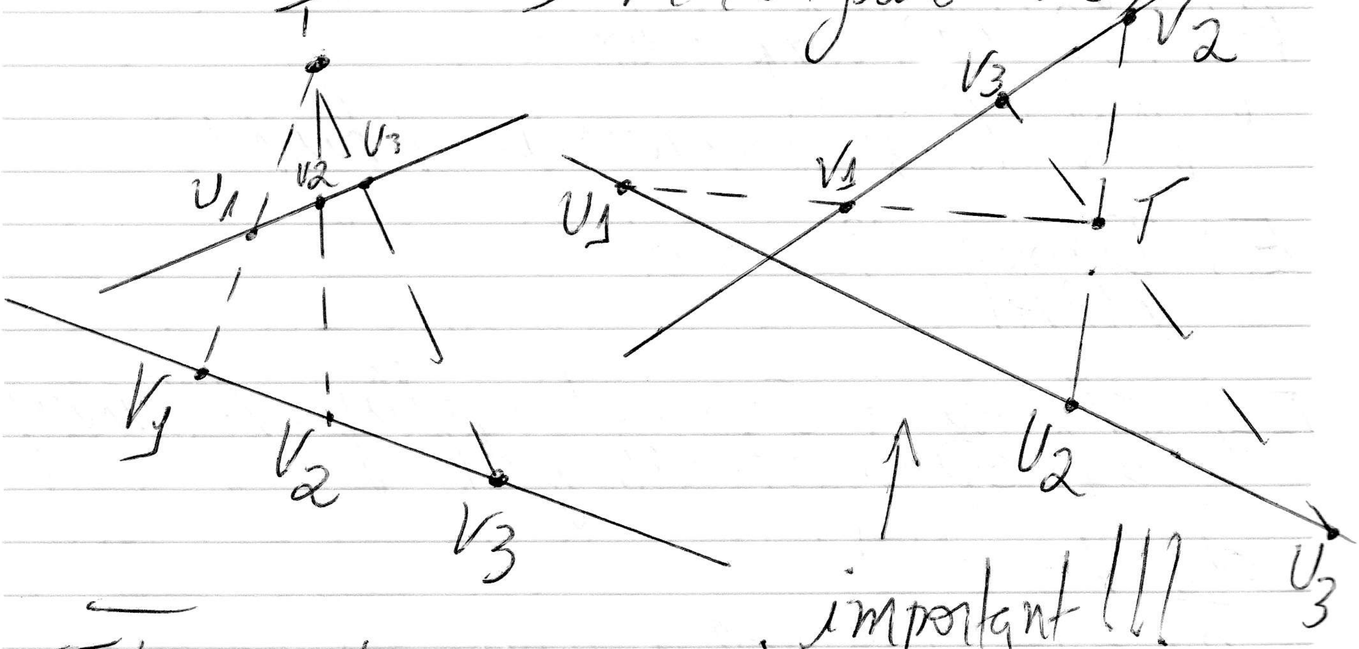

important!!!

$\overline{TU_1}$ → tem um sentido
$\overline{TU_3}$ → tem outro sentido.

06.05.2020

Esta perspectiva é muito importante por que podemos descobrir a posição de um ponto em outro plano quando há uma intersecção de duas retas.

Iremos extrapolar para planos e entre plano e uma superfície curva.

Important:

Pag. 321 (346) Figure 7.4

A Complete quadrangle.

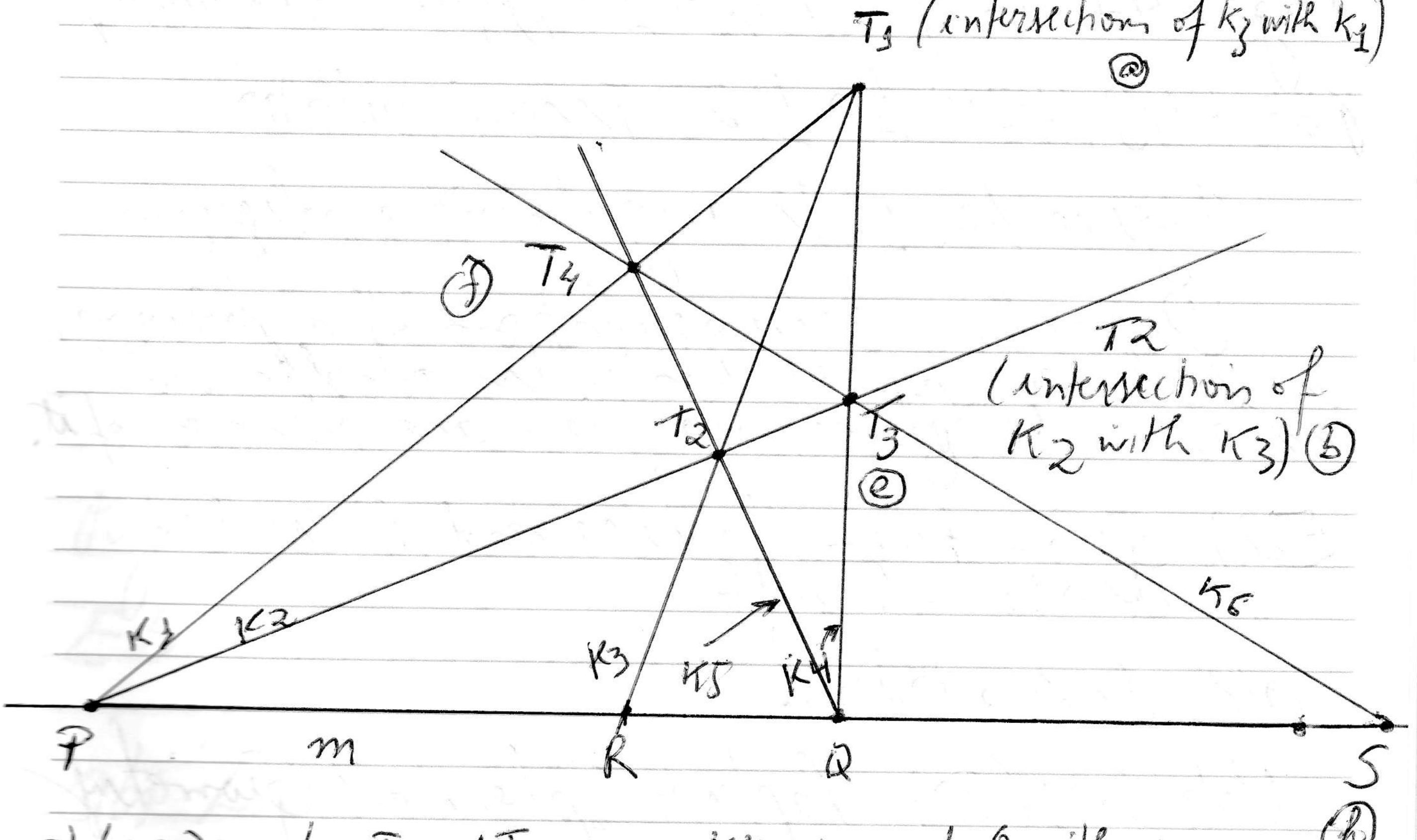
T1 (intersection of K3 with K1) ⓐ
ⓕ T4
T2 (intersection of K2 with K3) ⓑ
T2
T3
ⓔ
K1
K2
K3
K5
K4
K6
P
m
R
Q
S
ⓗ

g) Line → connecting T3 and T4

h) S → line intersection line m

ⓒ K4 → connect Q with T_1

ⓓ K5 → connect Q with T_2

ⓔ K4 → intersects K2 say at T_3

ⓕ K5 → intersects K1, say at T_4.

data 06.05.2020

Pág. 337(362) 7.3 Analytic Projective Geometry

Pág. 338 (363) Example 2

We can relate each point of the projective plane to two antipodal points on a sphere.

Fig 7.21

Pág. 361(386) 7.6 Projective Space

Fig 7.32 A two-point perspective view of a cube

Pág. 447(472) Chapter 10 Discrete Geometry

Pág. 449 (474) 10.1.2 Triangulations

Figure 10.1 Triangulating a polygon.

Depicts a triangulation of a polygon, a division of it into triangles all whose (de quem) vertices are vertices of it.

E2) Geometry - The Line and the Circle (2018...)...pdf

obs: Livro todo!!

Pág. 59(80) Chapter 4 Spherical Geometry

Pág. 189(210) 10 Book III: Circles

Pág. 221 (242) 11 Book IV: Circles & Polygons

06.05.2020

E3) Calculus deconstructed - a second course in first-year...pdf

Pag. 135(152) 4 Differentiation

Pag. 137(153) Figure 4.1 Ax+By=C

Pag. 247(264) 5 Integration.

E4) Geometry Illuminated - An Illustrated Introduction... pdf

Obs. Livro todo!!!

Pag. 105(122) Part II Euclidean Geometry

Pag. 223(240) Part III Euclidean Transformations
↳ muito bom!!!!

Obs. Aplicar 'quando' (no momento) da projeção da superfície curva sobre um plano.
(simular vice-versa também)

Pag. 395(412) Part IV Hyperbolic Geometry.

E5) Calculus in 3D. Geometry, vectors, and...pdf

Obs. Livro todo!!!

Pag. 115(127) Chapter 2
Curves and Vector-Valued Functions of One Variable.

data 06.05.2020

S T Q Q S S D

Pág. 707(719) A. Conic Sections: Apollonius' approach

Pág. 791(803) H. Surface Area: the Counterexample of Schwarz and Peano.

E6) Linear Algebra and Geometry (2019...)...pdf

Obs. Livro todo!!!

E7) ~~Maa Dictionary Maasai...pdf~~

Books_79_2020/ MAA/ .djvu

E7.1) In mathematical circles. Quadrants I, II... djvu

Obs.: Historic

E7.2) Lure of the integers...djvu

Pág. 196(214) Integer 67

An Inequality on $\pi(x)$

It is not known whether or not

$$\pi(x+y) \leq \pi(x) + \pi(y)$$

for all sufficiently large x, y.

06.05.2020

Extending work of Rosser and Schoenfeld, various authors have proved related results.

Krawczyk proved the following,

$\pi(kx) < k\pi(x)$ for

$k \geq e$, $x \geq 67$.

For $\epsilon > 0$, $x \geq 17$, $y \geq 17$,

$x + y \geq 1 + e^{4(1+1/\epsilon)}$, it has been shown that

$\pi(x+y) < (1+\epsilon)(\pi(x) + \pi(y))$.

There is an interesting connection between Conjecture A ($\pi(x+y) \leq \pi(x) + \pi(y)$, $x, y \geq 2$) and the generalization of the twin prime conjecture that follows it.

E7.3) Symmetry ... djvu

Pag. 11(24) Chapter 2 Inside and Outside

2.1 Reflecting in a Circle[1].
(inversion)

Important:

Pag. 14(27) 2.2 Composition of two inde-

data 06.05.2020
S T [✓ Q] Q S S D

reflections. Figure 23

Fig. 24 (Pág. 15(28)).

Direct construction of the image point.

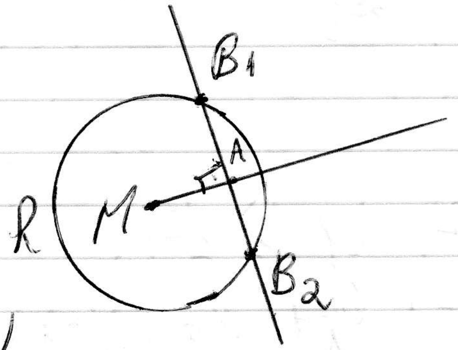

a)

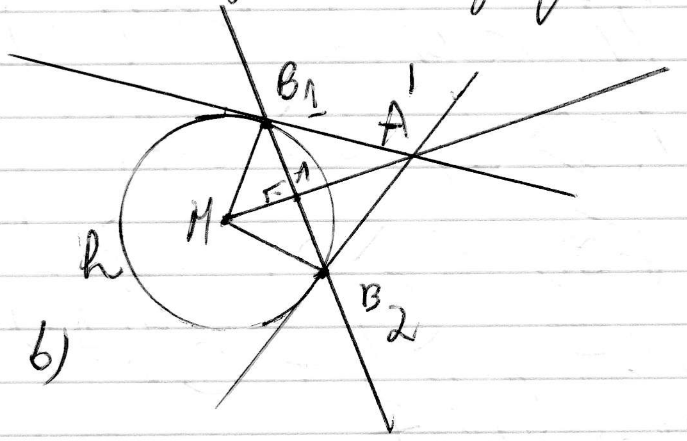

b)

Obs. Livro todo!!!

E7.4) Geometric inequalities (1975...).. djvu

Pág. 65(71) Chapter Three the Reflection Principle

Figure 3.1.

Obs. Livro todo!!!

E7.5) Proof and other Dilemmas_ Mathematics and Philosophy.. djvu

E7.6) Carl Friedrich Gauss - Titan of Science... djvu

Pág. 163(211) Chapter 14 Surface theory

06 05 2020

In the spring of 1816 Gauss had proposed as a prize problem for a new astronomical journal founded by von Lindenau and Bohnenberg (Bohnenberger), the projecting of two curved surfaces on each other, yet preserving similarity in the smallest parts.

At the time he had the solution and had thought out a projection by means of parallel normals on the sphere of radius unity.

In his concept the development or bending of curved surfaces was a special case of projection.

(resolveu)
At this time he had worked out the concept of the total curvature of a portion of a surface, and the concept of the measure of curvature at one point of a surface, as well as preservation of the measure of curvature in spite of bendings.

In the time from 1812 to 1816 he was busy with the theory of the shortest lines on the elliptic spheroid, and conceived of the cartesian coordinates of a point of a curved surface as functions (eixos curvos) of two auxiliary magnitudes.

E7.7) the golden section ... dyna

Pag. 48(65) 3.5.1 Area Comparison with a Circle

Figure 66

data 06.05.2020

S T Q Q S S D

The ellipse and the circle should have the same area.

Pag. 49(66) 3.5.2 Geometry in the Music Cassette.

Pag. 52(69) 3.5.3 The Ellipse in a Square Lattice.

Pag. 52(69) 3.6 Golden Trigonometry.

$$\sin 18^\circ = \frac{\rho}{2} = \frac{\tau - 1}{2}$$

$$\cos 18^\circ = \sqrt{1 - \frac{\rho^2}{4}} = \frac{\sqrt{2 + \tau}}{2}$$

$$\tan 18^\circ = \frac{\tau - 1}{\sqrt{2 + \tau}}$$

Pag. 75(92) 5.3 Power of $1 + \sqrt{2}$

E7.8) Circles, a mathematical view... dyvu

Obs.: Livro todo!!!

Pag. 01(39) Chapter 1
The nine-point circle.

Pag. 35(73) 7. Circles cutting at a given angle.

—

06.05.2020

E7.9) Generalized Riemann Integral (1980...)...djvu.

E7.10) Geometry Illuminated...djvu

Obs: Tenho em pdf.

E7.11) The genius of Euler...djvu

Pág. X (10)

All celebrated mathematicians now alive are his disciples: there is no one who... is not guided and sustained by the genius of Euler.

From the Marquis de Condorcet's Eulogy to Euler, 1783

Pág. 155 (172) Euler and Differentials[1]

F) Books_79_2020/
Fractional_Trigonometry/ ✡

The Fractional Trigonometry...pdf

Obs: Livro todo!!! Muito bom!!!

G) Books_79_2020/
Trisectors Dudley/

A_Budget_of_Trisectors.pdf

Obs. Livro todo!!!

data 06.05.2020

S T [✓ Q] Q S S D

H) Circles Eves/
Howard W. Eves

Obs.: Diretório todo!!!

I) ARQ_IMPORTANT_04/
Géométrie/

a) Jacques Hadamard - Leçons de géométrie élémentaire I...pdf

b) Jacques Hadamard - Leçons de géométrie élémentaire II...pdf

J) ARQ_IMPORTANT_05/
Géométrie/

J1) livre-exercices-de-geometrie-fgm-bnf.pdf

Obs.: 1176 pages.

J2) livre-geometrie-projective-daniel-perrin...pdf

Obs. 1328 pags, mas este tem 526 páginas.

Material Completo:

Books_80_2020/
Livre_Géométrie_Projective_Daniel_Perrin_important/

→ todo o diretório!!!

data 06 05 2020
S T [✓ Q] Q S S D

K) Books_80_2020/
The Problem of the Motion of Bodies/
3319048392.pdf

Obs.: Livro todo!!!

L) Books_80_2020/
Géométrie-Projective/

L1) 5000 Years of Geometry...pdf
L2) Art_and_Geometry...pdf
L3) Beyond Geometry...pdf
L4) Bredon.pdf
L5) Century of Geometry, 1830-1930...pdf
L6) Euclidean_and_Non_Euclidean...pdf
L7) Geometry through History...pdf
L8) livre-géométrie-michèle-audin...pdf
L9) Multiple-View...pdf
L10) slides-géométrie-projective.pdf
L11) marcel-berger-a-panoramic-view-of-riemannian-geometry...pdf
L12) coxeter-greitzer-redecouvrons-la-geométrie.pdf

data 06.05.2020

S T [✓ Q] Q S S D

166) 20200506

A) Books_80_2020/
Géométrie-Chasles_2_Volumes/

a) Traité_de_géométrie_supérieure...
508r.pdf

b) Traité_de_géométrie_supérieure...
529R.pdf

B) PC_Oueto/
Importantíssimo_Gauss/

Obs: Todo diretório !!!

C) PC_oueto/
The VNR Concise Encyclopedia of Mathematics/

Obs. Todo diretório

Copy: Books_80_2020/
The VNR Concise Encyclopedia of Mathematics/
The VNR Concise Encyclopedia of Mathematics-Springer Netherl.pdf.

D) ARQ_BASICS_02/
Pappus

Obs. Todo diretório !!!

06.05.2020

E) ARQ_BASICS/04/
Géometrie_Projective_important/
Obs: cf le théorème de Céva/

F) ARQ_BASICS/04/
ORTHOGONAUX_DARBOUX_Images/

G) ARQ_BASICS/04/
Riemannian Geometry and Geometric Analysis/

a) [Universitext] Jurgen... Riemannian.pdf
(6th Ed)

b) Riemannian Geometry and Geometric Analysis (2017...)...pdf
(7th Ed)

H) ARQ_BASICS/04/
Traité des propriétés projectives des figures_Poncelet/

Obs: Todo diretório!!!

I) ARQ_BASICS/06/
Gravitation/

Obs: Todo diretório!!!

J) ARQ-BASICS/07/
Infinitesimal/

Obs: Todo diretório!!!

data 06 05 2020
S T [✓ Q] Q S S D

a) 04-Volumes_Germany/
⟹ Cauchy /

b) Gauss_2_Volumes/

c) Gauss_12_Volumes/ ✡

d) Houel Infinitesimal 4 Volumes/

K) ARQ_BASICS/07/
The Secrets of Ancient Geometry-
2-Volumes/

L) ARQ_BASICS/08/
Books_25_2020/

M) ARQ_BASICS/08/
Riemann/

a) Dani... Geometry in History... pdf

b) Eighteen Essays in Non-Euclidean Geometry (2019...)...pdf.

c) From Riemann to differential geometry and relativity (2017, Springer).pdf

d) Geometry and Spectra of Compact Riemann Surfaces...pdf

e) Geometry of Riemann Surfaces...pdf

f) Geometry of Riemann surfaces and their moduli spaces...pdf

data 06 05 2020
S T [✓ Q] Q S S D

g) Lobachevskii_Pangeometry...pdf
h) Metric Spaces, Convexity and Nonpositive Curvature...pdf
i) On the Hypotheses Which Lie at the Bases of Geometry...pdf
→ Riemann

j) Peter Pesic - Beyond Geometry - Classic Papers from Riemann to Einstein ...pdf
k) Riemann Surfaces by Way of Complex Analytic Geometry...djvu
L) Vargas J. G. - Differential Geometry for Physicists and Mathematics...pdf
N) ARQ_BASICS/08/
Spherical_Trigonometry/
Obs.: Todos diretórios!!!

O) ARQ_BASICS/08/
01) claimofleibnitz00slomrich.pdf
02) Perfect Ali Pi.pdf
Obs:

P) ARQ_BASICS/10/
Pappus/ and Pappus_1/

data 06.05.2020

S T [✓ Q] Q S S D

Q) ARQ_BASICS/10/
Spherical Geometry and Its
Applications/

R) ARQ_BASICS/10/
The Cinderella.2 Manual
Working.../

167) 20200507

Deus é oniexistente (eterno),
onipotente
onisciente
onipresente.

O Universo é parte de Deus
e portanto, o Universo sempre existiu.

O Big Bang é um momento ativo do Universo.

O Universo é ativo em decorrência dos Forças Fundamentais de Interação

Weak
Strong
Electromagnetic

A gravidade não é uma força fundamental, mas é decorrência da interação das 3 forças fundamentais do Universo.

data 07.05.2020

Gravitation is by far weakest* of the three interactions at atomic scale, where electromagnetic interactions dominate.

Albert Einstein não conseguiu fazer a unificação dos forças fundamentais pois considerou a gravidade como uma força fundamental e não uma consequência dos 3 forças fundamentais de interação do Universo.

General relativity generalizes special relativity and refines Newton's law of universal gravitation

↳ descreveu o efeito das 3 forças fundamentais do Universo e não uma força em si mesma (independente), mas decorrentes daquelas.

providing a unified description of gravity as a geometric property of space and time (spacetime).

In particular the curvature of spacetime is directly related to the energy and momentum of whatever (tanto faz) matter and radiation are present.

Light Cone:

Geometric Newton → classical mechanics is a limiting case of special relativity

data 07.05.2020
8

mechanics. In the language of symmetry, where gravity can be neglected, physics is Lorentz invariant (Hendrik Lorentz) as in special relativity rather than Galilei invariant as in classical mechanics.

The defining symmetry of special relativity ~~th~~ is the Poincaré group, which includes translations, rotations and boosts.

The differences between the two become significant when dealing with speeds approaching the speed of light and with high-energy phenomena.

A velocidade da luz é definida como uma constante universal, mas esta não absolutamente constante, pois a sua velocidade depende do meio em que trafega e com base nas 3 leis fundamentais de interação, aplicáveis a curvatura do espaço-tempo.

Voltando (Light cone)

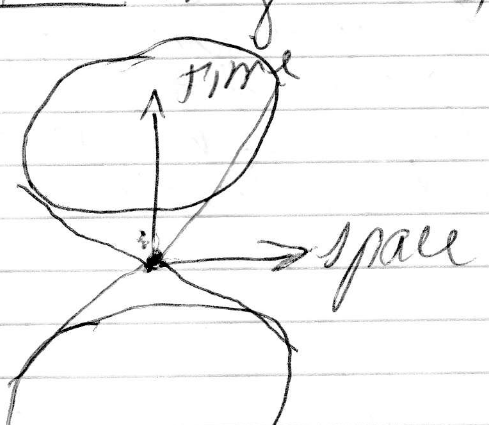

Meu: light cone

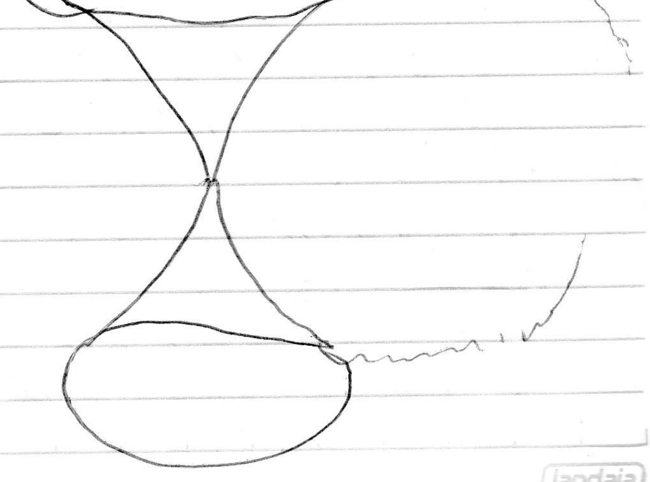

07 05 2020

With Lorentz symmetry additional* structures are defined by the set of light cone. These sets are observer-independent.

In conjunction with the world lines of freely falling particles, the light-cones can be used to reconstruct the

space-time's semi-Riemannian metric,

at least up to a positive scalar factor.

In mathematical terms, this defines a conformal structure or conformal geometry.

↳ the conformal isometries of a domain therefore (portanto) consist of holomorphic self-maps. In particular, on the conformal compactification – the Riemann Sphere – the conformal transformation are given by the Möbius transformations

↳ can be visualized as the complex number plane wrapped around a sphere (by some form of stereographic projection).

Important:

Stereographic projection is conformal, meaning that it preserves the angles at which curves across each other. Stereographic projection does not preserve area. In general, the area of a region of sphere does not equal the area of its projection onto the plane.

$$z \rightarrow \frac{az+b}{cz+d}$$

where $ad-bc$ is nonzero.

(o meu é em 3D)

2 cone invertido

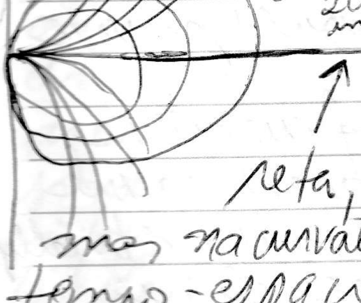

reta, mas na curvatura tempo-espaço não há [illegible] reta perpendicular ao plano em que houve a transformação.

data 07.05.2020

$dA = \frac{4}{(1+X^2+Y^2)^2} dX dY$. Along the unit circle $X^2+Y^2=1$, there is no inflation of area in the limit, giving a scale factor of 1. Near (0,0) areas are inflated by a factor of 4, and near infinity areas are inflated by arbitrarily small factors. The metric space is given in (X,Y) coordinates by $\frac{4}{(1+X^2+Y^2)^2}(dX^2+dY^2)$, and is the unique formula found in Bernhard Riemann's Habilitationsschrift on the foundations of geometry, delivered at Göttingen in 1854, and entitled

Über die Hypothesen welche der Geometrie zu Grunde liegen.

Copy: ARQ_BASICS_04/DEUS/Geom.pdf

Source: www.maths.tcd.ie/pub/HistMath/People/Riemann/Geom/Geom.pdf.

Source: https://www.emis.de/classics/Riemann/

→ The Mathematical Papers of Georg Friedrich Bernhard Riemann.

No map from the sphere to the plane can be both conformal and area-preserving. If it were, then it would be a local isometry and would preserve Gaussian curvature.

→ is an intrinsic measure of curvature, depending only on distances that are measured on the surface, not on the way it is isometrically embedded in Euclidean space. This is content of the Theorema egregium (1827).

Copy: ARQ_BASICS_04/DEUS/Gaussian_curvature/

a) Andrew Pressley - Elementary Differential Geometry (2010...)...pdf Pag. 247 (248).

07 05 2020

b) ARQ-BASICS_04/DEUS/Gaussian-curvature/
CS6.pdf
c) L6_GaussBonnet.pdf
d) CDG.pdf
e) Gauss_Egregium.pdf
f) rutgers-lib-26114-PDF-1.pdf ⇒ Introduction to the Curvature of Surfaces.
g) Palais_FundamentalTheorem.pdf

Total Curvature (Gaussian Curvature) The surface integral of the Gaussian curvature over some region of a surface is called the total curvature. The total curvature of a geodesic triangle equals the deviation of the sum of its angles from π. The sum of the angles of a triangle on a surface of positive curvature will exceed π, while the sum of the angles of a triangle on a surface of negative curvature will be less than π. On a surface of zero curvature, such as the Euclidean plane, the angles will sum to precisely π radians. $\sum_{i=1}^{3} \theta_i = \pi + \iint_T K \, dA$

Obs. Gauss-Bonnet theorem

cont. ... Gaussian curvature. The sphere and the plane have different Gaussian curvatures, so this is impossible.

Circles on the sphere that do not pass through the point of projection are projected to circles on the plane. Circles on the sphere that do not pass through the point of projection are projected to straight lines on the plane. These lines are sometimes thought of as circles through the point at infinity, or circles of infinite radius. All lines in the plane, when transformed to circles on the sphere by the inverse of stereographic projection, meet at the projection point. Parallel lines, which do not intersect in the plane, are transformed to circles tangent at projection point. Intersecting lines are transformed to circles that

data 07.05.2020

intersect transversally at two points in the sphere, one of which is the projection point.

Cross Section: A graph of $z=x^2+xy+y^2$. For the partial derivative at (1,1,3) that leaves y constant, the corresponding tangent line is parallel to the xz plane. Observamos uma curvatura da superfície, compare esta curvatura com as equações dA da Stereographic Projection. ✡.

Voltando: light cone. e complementando com o Penrose-Carter diagram of an infinite MINKOWSKI UNIVERSE / wikipedia.org/wiki/General-relativity

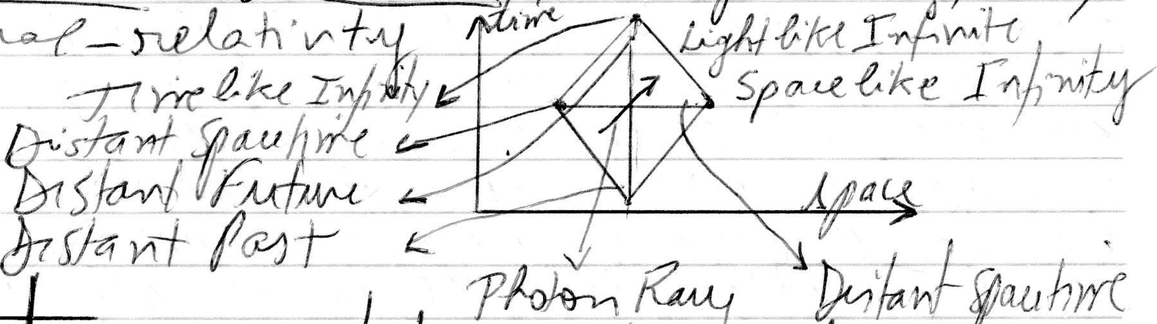
time
Light like Infinite
Space like Infinity
Time like Infinity
Distant Spacetime
Distant Future
Distant Past
Space
Photon Ray
Distant Spacetime

Obs.: + A Cruz neste diagrama é um ângulo reto no seu ponto central, no meu caso, já há uma curvatura intrínseca. Todas as linhas são curvas no meu modelo.

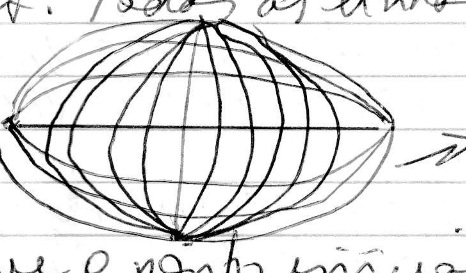
dentro do losângulo

Neste diagrama, devemos observar que o ponto inicial de referência deverá ser alterado para o ponto inicial do Big Bang, ou seja, para o momento em que ocorreu o Big Bang mas havendo a alteração do modelo utilizado para o desenvolvimento do Universo.

As 3 forças fundamentais de interação do Universo geram como efeito a gravidade, como uma consequência "sine qua non" da curvatura do Universo, gerada em decorrência destas 3 forças fundamentais universais e não em decorrência direta da gravidade, mas indiretamente.

data 07 05 2020

Speed of light: In 1865, James Clerk Maxwell proposed that light was an electromagnetic wave, and therefore travelled at the speed c appearing in his theory of electromagnetism. In 1905, Albert Einstein postulated that the speed of light c with respect to any inertial frame is a constant and is independent of the motion of the light source. c = celeritas

✡ In 1856, Wilhelm Eduard Weber and Rudolf Kohlrausch had used c for a different constant later shown to equal $\sqrt{2}$ times the speed of light in vacuum.

Poincaré group: first defined by Hermann Minkowski (1908) as the group of Minkowski spacetime isometries

↳ ignoring the effects of gravity (i.e)

Minkowski Space → the presence of gravity spacetime is described by a curved 4-dimensional manifold for which the tangent space to any point is a 4-dimensional Minkowski space. Thus, the structure of Minkowski space is still essential in the description of general relativity.

Hyperbolic stereographic projection

Red circular arc is geodesic in Poincaré disk model; it projects to the brown geodesic on the green hyperboloid.

Curvature: Gaussian Curvature

Formally, Gaussian curvature only depends on the Riemann metric of the surface. This is Gauss's celebrated Theorema Egregium, which he found while concerned with geographic surveys and mapmaking.

Gaussian curvature K at the point P can be computed by the Bertrand–Diguet–Puiseux Theorem as

$$K = \lim_{r \to 0^+} 3 \left(\frac{2\pi r - C(r)}{\pi r^3} \right)$$

data 07.05.2020

168) Books_80-2020/
  a) Deus/ → referência ao item 167.
  b) Gauss/ → referência ao item 167.
  c) Géométrie_Projective → também se refere 167.
  (contém arquivos básicos com fundamentações),
  d) Livre_Géométrie_Projective_Daniel_Perrin.../
  e) Physics and Mathematics of Quantum
  Many-Body Systems/
    e1) 3030412644.pdf
    e2) Bohmian mechanics_~~an~~ the physics and
    mathematics of quantum theory...pdf
    e3) Einstein's space-time_an introduction
    to special and general relativity...pdf
    e4) Geometry, topology...pdf
    e5) The Quantum Mechanics Conundrum...pdf
  f) The Problem of the Motion of Bodies/
    3319048392 pdf —

169) 20200507

Arquivos importantes para a fundamentação da pesquisa estão no diretório:
ARQ_BASICS_20200504
→ contém 5 arquivos.
Outros: HD_1.36_01
HD_1.36_02
HD-610
HD-620
HD-Geoflex
Pendrive_Geral
Pendrives
QBex (PC)
Terra (1TB-Backup)

II

202005072146

Chapter 1 - Part 2

Notebook-02---20211112

20201213.

mend.

—

Script (Google)

Documentos Google →
Planilhas Google →
Apresentações Google →
ou
Formulários Google

Documentos
↳ Ferramentas
↳ Editor de Script

—

myaccount.google.com/permissions

—

→ https://cloud.google.com/resource-manager

—

Google Script:

31,683.93 GB

=

Geometry and Its Applications
Georg Glaeser

$360 \div 7 = 51,\overline{428571}$

—

O amor não tira férias
(filme)

—

Darstellende Geometrie
Wunderlich

Ebene Kinematik.

—

Descriptive Geometry.

—

georg.glaeser@uni-ak.ac.at

—

avxhm.se/ebook/3030613976.html

Pi → não é TRANSCENDENTAL NUMBER
perhaps RATIONAL

Lindemann-Weierstrass Theorem.

is a result that is very useful in establishing the transcendence of numbers.

---

about: support

→ Cersei

"campo externo de empacotamento" ao fluido.

31,684.17 GB.

vacância por decurso de prazo

Pfizer → RNA mensageiro.
(altera)

=

salvaguarde por efeitos adversos...

razões confessáveis ou não.

—

proselitismo.

=

nós precisamos inventariar.

conflagração

=

⇒☆ Book1-191-2020/xiao... Photoenergy.pdf

Pág. 58(75) Fig. 25

π-bridge

— —

linkedin.com/in/
amanda-aldous-cd-adec-
mds-mpa-pcsc-73736632/

—

20201219

Georg Glaeser

http://num.math.uni-goettingen.de/werner/schmankerl.pdf

---

www.geometrie.tuwien.ac.at/stachel.

---

www.dmg.tuwien.ac.at

---

triângulo acutângulo (<90)
isosceles → 3 partes ≠

triang. obtusângulo → > 90°

---

variador de velocidade

---

rosca helicoidal → arquimedes

---

epígonos do poder e de seus acólitos

→ era do Reino Unido.

Münsterberg 1797

Igreja Polonesa

Polônia - Cúpula
tempo

3.141433158

180 Sin

— Jason Padget

Jason Padget Tacoma

—

Elisha Gray patenteou antes de Graham Bell (Telefone)

→ mesmo dia

—

March 7, 1876

—

registro acústico (universal)

—

funções modulares (Ramanujam)

—

acash (céu em Sanscrito)

—

jasonpadgett.info

Jason became fascinated with
number Pi, irrational number
that is

"a fundamental building block
of everything around me,
a signifier of nature's
perfect symmetry".

Struck by Genius

A geometria do pi é tecida no espaço-tempo
Sempre achei que a geometria
de Pi desenha anéis Einstein
(também conhecidos como
lentes gravitacionais).

The geometry of Pi

A geometria do pi é tecida no
espaço-tempo e é impressionantemente
bonita e é encontrada às vezes em
lugares surpreendentes.

Sempre achei que a geometria
de Pi desenha anéis Einstein. (também
conhecidos como lentes gravitacionais),
mas parece estar relacionado com estes
novos objetos também.

The geometry of Pi is woven into spacetime and is stunningly beautiful and it is found sometimes in surprising places.

I've always thought the geometry of Pi described Einstein rings (aka gravitational lensing), but it seems to be related to these new objects also.

—

jason-padgett.pixels.com

—

jason.s.padgett@durham.
ac.uk

—

Geometrie

—

Comic Kendig

Complex Variables Krantz

Steven G. Krantz

⇛ sk@.wustl.edu
math

—

sgkrantz@gmail.com

—

Mareaten Don Jon film

—

Hangouts Meet

↳ Gmail
Agenda.

—

Google Workspace

ID 0Zuwel9h

—

Books-49-20201108
1110

49 e 50
50
51
57...

—

31,771.91 GB → 34,520.93 GB

Margret Almer
Stefanie Hertel

—

⇒ rclone | spinster

—

Teoria Decomposição em Valores Singulares (2D)

Álgebra II
Vídeos

—

Alfacon Concursos Públicos (youtube)

—

184.208.227.35.bc.googleusercontent.com

—

↳ leibni.com (20201221)

—

34,559.87 GB (33,75 TB)

—

leibni.blogspot.com

—

Luz → "pontos C" | Kobus Kuipers

Departamento de Nanociências Quânticos da Universidade de Tecnologia de Delf na Holanda.

Saca-rolhas de luz' permitem que a informação seja transportada em um nível quântico de maneira mais eficiente".

"(Queremos) manter a luz como luz e pular a etapa de convertê-la em eletricidade", diz Kuipers.

—

35,68 TB

—

ácido valproico
↳ altera DNA.

Valproato de sódio (VPA)

ação epigenética do VPA.
↳ pode tb ocorrer uma interação direta com a estrutura da histona H1.

✓ ↳ existe no núcleo celular

parte da cromatina.

20201222
36,540.72 GB.

workspace.google.com

drfone.wondershare.com

Google Drive
Shift + F = pasta
criar.

www.google.com/takeout
ou
takeout.google.com

Save to Google Drive

Google Drive como Map as a
Network
Drive.

Quantum Chemistry and Spectroscopy
Engel

Curva de Gauss - a sedução da matemática
Isac Spinz

61,680.27 GB → 60,23 TB. (20201223)

Anel de Einstein - Chwolson
resultado da distorção da luz de uma galáxia elíptica a 4 bilhões de anos-luz de distância.
Constelação de Fornax
GAL-CLUS-022058s
ganhou o apelido "Molten Ring"
1º Anel → B1938+666

Boston - 193 - 2020/
Introduction to Modern
Dynamics chaos....\

019884462x.pdf

Pág. 136(150)

4.42 Feigenbaum number and universality

$$\delta = \lim_{n\to\infty} \frac{S_{n+2} - S_{n+1}}{S_{n+1} - S_n} =$$

4.66920160910299067185320382

This Feigenbaum ratio is the same for all quadratic maps

Pág. 168(182)

5.4 KAM theory

continue fraction

...

A quadratic irrational is the solution to a quadratic equation with integer coefficients and has periodic continued fractions, such as

$$\sqrt{2} = [1; 2, 2, 2, \dots] \quad (5.24)$$

Nonquadratic irrational have nonrepeating continued fractions, such as

$$\pi = [3; 7, 15, 1, 292, 1, 1, 1, 2, 1, 3, 1, 14, 2, 1, 1, 2, 2, 2, 2, 1, 84, 2, \dots] \quad (5.25)$$

$$\pi - 3 \approx [0; 7, 15, 1, 0, 0, 0, \dots]$$

$$= \cfrac{1}{7 + \cfrac{1}{15 + \cfrac{1}{1}}} = 0{,}14160 = \frac{177}{1250} \quad (5.26)$$

Snow Patrol - Open Your Eyes

{ cinismo
sarcastico

(Rue
Wantweg.

Remy Zero "Save Me"
"Perfect Memory"

⇒ 178.0.83.34.bc.
Leiboni.com googleusercontent.com

Ludwig

201-0-231-208.dial-up.
telesp.net.br

✡ [support.google.com/a/
contact/nonprofit

20201225M

Quéops - 3341 - 3094 A.C.
(radiocarbono → madeira)

you belong to the city
Glenn Frey.

all round the world
Lisa Stansfield.

—

Being with you Smokey Robinson

—

⇒ Books - 193 - 2020 / Mathematics for Machine Technology /
Fujio Yamaguchi - Computer Aided Geometric... pdf
Pág. 11 (46)
Fig 1.9 three types of line segments

Obs.: Root Cause of All Evil

—

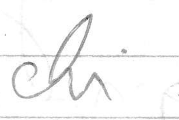

Pag. 300 (326)

Conventional geometric Newton-Raphson method, ...

Fig. 15.2

Determining intersection point of curve and line segment.

—

Circles Pedoe

Analytical Conics Spain

Non-Euclidean Geometry
Kulczycki

Introduction to Projective Geometry Pedoe

—

Books - 183 - 2020 / Geometry Revisited / Compact Projective

Planes - with an Introduction
to Octonion Geometry pdf
Pag. 26 (41)
12.2 the Projective
completion

Figure 12a

—

Electromagnetic Field
theory Zahn

—

Quantum theory
David Bohm

—

Papel Maculatura → trastes
marrons do papel higiénico

—

t.sonar@tu-bs.de

—

Eberhard Knobloch

Adriana Calcanhoto - Mais feliz

Angus Stone - Wooden Chair

=

Thvideos.net (Stelvio)

—

Narayana dash

—

Spherical Fuzzy Sets

Books-194-2020\Geometry-important-20201227\3030454606.pdf

Pag. 6(35)

2. Spherical Fuzzy Sets: Preliminaries

Fig. 2 Geometric representation of IFS, PFS, NS and SFS.

=

J 5781.

Mateus 4,4. —

20201228 1550.

ETA-8424 Sandro
Casa Branca

—

Cálculo Vectorial — Marsden
— Tromba

Redshift
↳ GN-z11 (EGS-z11)

13,4 bilhões letra Z é 10.957
anos-luz — ═ ↳ redshift

✡ { ANNITA
↳ reduz carga viral
—
Nitazoxanida (antiviral) }

—

⇛ replicação viral (impede)

—

serve para
infecções virais agudas

—

Cálculo Vectorial
5ª ed
Marsden

Álgebra Lineal y
Geometría Vectorial
Vizmanos

---

e-rara.ch

20201229
Crypt Sync
http://tools.stefankueng.com/
CryptSync.html

Simeão
↳ não morreria sem ver
o messias.
" Os meus olhos viram a tua salvação."
Luc 2, 22-35
(Lucas)

---

9 7649 0860 Fatima Leão

20201230

Luke Letlow - Louisiana
(41)

2014 - Breaking Relativity

tetraon.dk

ISBN 978-87-972469-1-7

- ✓ Geometry of Physics Frankel
- ✓ Quantum Mechanics Merzbacher
- ✓ Spacetime Physics with Geometric Algebra Hestenes
- ✓ Geometry of Paravector Space with Applications to Relativistic Space - Baylis

Algebra of Physical Space

Paravector
↓ ↳ Clifford Algebra
Baylis

geometry-of-relativity.net

musich

cc — curve circle

pimus

The secret life of stars
Lisa Harvey-Smith

Comadres

Licor de maceração do milho → meio de cultura para a Penicilina

2021|01|01//

glosa do colegiado,

escárnio

pouco operativo.

espetáculo patético.

personagens de ribalta.

Rosslare → to Dunkirk
Irlanda

Cherbourg
Bilbao

Space-Time Algebra
Hestenes

⇒ geocalc.clas.asu.edu

Electrodynamics -
A modern Geometric
Approach
Baylis

semita - hebreu - judeu

Ano Sideral 365,25636 dias

Ano Trópico: 365,242189 dias

longitude eclíptica do Sol

equinócio|.

Ano Anomalístico:

passagem 2 vezes pelo periélio

365,2596 dias

Google Drive

√ Cota de download foi excedida.

crie o atalho em uma pasta e depois faça o down da pasta.

---

sanet.ws/blogs/mercury-books/

Albert Marden

University Minnesota

math.umn.edu/directory/albert-marden

Cálculo Integral Avançado
Bouchara

—

Aline Szewkies
alinetourism@gmail.com

—

Cemitério no Monte Sião.

—

+972-52-7027085

—

Books-001-2021\Riemann Geometry\
David G. Blair - The Detection of Gravitational Waves (2005).pdf
Pag. 6(31)

fractionally larger than π.

John Wheeler introduced the slogan "matter tells space how to curve, space tells matter how to move" to describe general relativity.

This hides the complexity
because in reality:
matter, the motion of matter,
and radiation density,
including propagating
waves of curvature
tell spacetime how to
curve.

Curvature creates curvature,
and influences its propagation!

—

Einstein realised the
connection between
gravitational dynamics
and Riemann geometry.

—

201.181.244.35.bc.googleusercontent.com

—

20210102

Books-001-2021\Math-
Distances\
Computer-Aided Geometric
Design...pdf

Pag 471 (495) -

Figure 21.4:

Relation between Chaikin's curve and polygon.

—

✡ Books-001-2021\ Relativistic Astrophysics and Cosmology\ Relativistic ... pdf

Pag. 21 (32)

Fig. 2.2 →

Co-ordinate lines and base vectors spanning the tangent space.

=

⟹ Friedrich Jegerlehner

www-com.physik.hu-berlin.de/~fjeger

—

Source:

Books-001-2021 \ The Anomalous Magnetic Moment... \ The

Anomalous Magnetic Moment of the Muon (2014).pdf

Pag. 08(24) →

"a possible anomaly $g \neq 2$ of the magnetic moment of the electron"

Obs.:

Books-001-2021\Organic Chemistry (Mastering Chemistry) 9th Edition\

rtu3463ryu54etr635.pdf

↳ Organic Chemistry

Pag. 05(39)

Fig 1-4

nucleus — node

(region of zero electron density)

Book-001-2021\Math-Distance\

Curves and Surfaces in Computer Aided Geometric Design.... pdf

Pag. 10(20)

Fig. 0.5.

Example of a surface in which a pentagonal surface patch occurs.

—

Book-001-2021\Eigenvalues....\

Eigenvalues.... pdf

Pag. 3(16)

1.2 Strict and Uniform Convexity

Banach Space

$x, y \in X$ are such that

$\|x+y\| = \|x\| + \|y\|$ then

either $y=0$ or $x=\lambda y$ for some $\lambda > 0$.

Obs.: Proposition 1,1

=

20240103:

Cosmolect1-11.pdf

pag. 12 (gravitational instability pictures

⇒ pag 18 Curved space-time

Photons propagate along geodesic curves

...

Pag. 31

= Pag. 46 apparent angular size of that physical size is determined by spatial curvature.

...

20210103

https://inspirehep.net/authors/1004419

—

ANUSSIN (obrigados à conversão).

BRIT MILA — BNEI
↳ circuncisão. ↳ descendente

—

Levítico 19,32 ///

⇛ Pessach = Páscoa Judaica.

—

Hydrogen atom: precision physics of simple atomic structure

Karshenboim

— —

20200104

geometrias.eu

geometrias.blogspot.com

—

deambulações em volta do π
(Mariana Sacchetti)

—

Antenokeeffe
geometrias.blogspot.com/2020/
05/antenokeeffe-reconstrucao.html

—

/2020/05/dual-do-problema-
de-malfatti-da-entrada.html

—

/2020/04/o-original-problema-
de-malfatti.html

—

2020/02/entre-triangulos-
porismo-e.html

—

2020/01/circuncentros-
o-o-o-b-e-o-c-de-um.
html

polarprof-001-site1.htempurl.com/
geometriagon

Thomas Andrews (1873 - 1912)
projetista naval irlandês do Titanic
→ morreu no navio (inafundável)

William Bullock (1813 - 1867)
Em 1863 automatizou a impressora rotativa.
Em 1867 quando instalava uma de suas máquinas rotativas automáticas ele chutou uma correia que se soltou sozinha e acabou prendendo a sua perna. Morreu do procedimento de amputação.

Otto Lilienthal (1848 - 1896)
engenheiro mecânico alemão, o "pai dos voos de planador"
2000 testes e 5 horas nos céus, morreu um dia após despencar 17 metros com seu planador Normal Glider e fraturar a coluna.
"Sacrifícios precisam ser feitos", teria dito.

Israel - Muro das Lamentações

SMALL WAILING WALL
(pequeno lamento)

$$\underline{45^{o}} \Rightarrow \begin{cases} \text{Sen } \frac{\sqrt{2}}{2} \\ \text{Cos } \frac{\sqrt{2}}{2} \end{cases}$$

Lei do Coseno e do Seno

Nariz aquilino (bico de águia) - romano.

MAAT - ordem Universal.

MATA HARI

7 August 1876 - 15 October 1917

41 anos

THOMAS ANDREWS

1873 - 1912

39 anos

Working in tandem
(conjunto)

20200105

—

israelbiblecenter.com

613 mandementos = ROMAN

HATIKVA ((Esperança)
↳ Hino de Israel
1878

—

Paradoxo de Mpemba.

→ modelo de Ising
↳ comportamento chamado de antiferromagnetismo.

—

"limite de Chandrasekhar"
1,33 x sol

"limite de Tolman-Oppenheimer-Volkoff" – 3 x (Buraco Negro) mignani@fis.uniroma3.it

Roberto Mignani (Dead)

Fabio Cardone
Dipartimento di Fisica
Università dell'Aquila

https://matematicafisica.
uniroma3.it

Lei do Retorno "Aliá" Lei 5710/1950

cloudflare-ipfs.com

Schindler: A História Verídica
1983.
Salomon Urbach
natgeobrasil.com

Alemão Amon Göth

caterva

→

1) $\text{sen}(\alpha) = \text{sen}(\beta)$

$\alpha = \beta + 2k\pi \ , k \in \mathbb{Z}$

2) $\cos(\alpha) = \cos(\beta)$

mesma abscissa

3) $\text{tg}(\alpha) = \text{tg}(\beta)$

$\alpha = (\beta + \pi) + 2k\pi =$

$\beta + (2k+1)\pi$, para $k \in \mathbb{Z}$

Logo

$\alpha = \beta + k\pi$

20210105

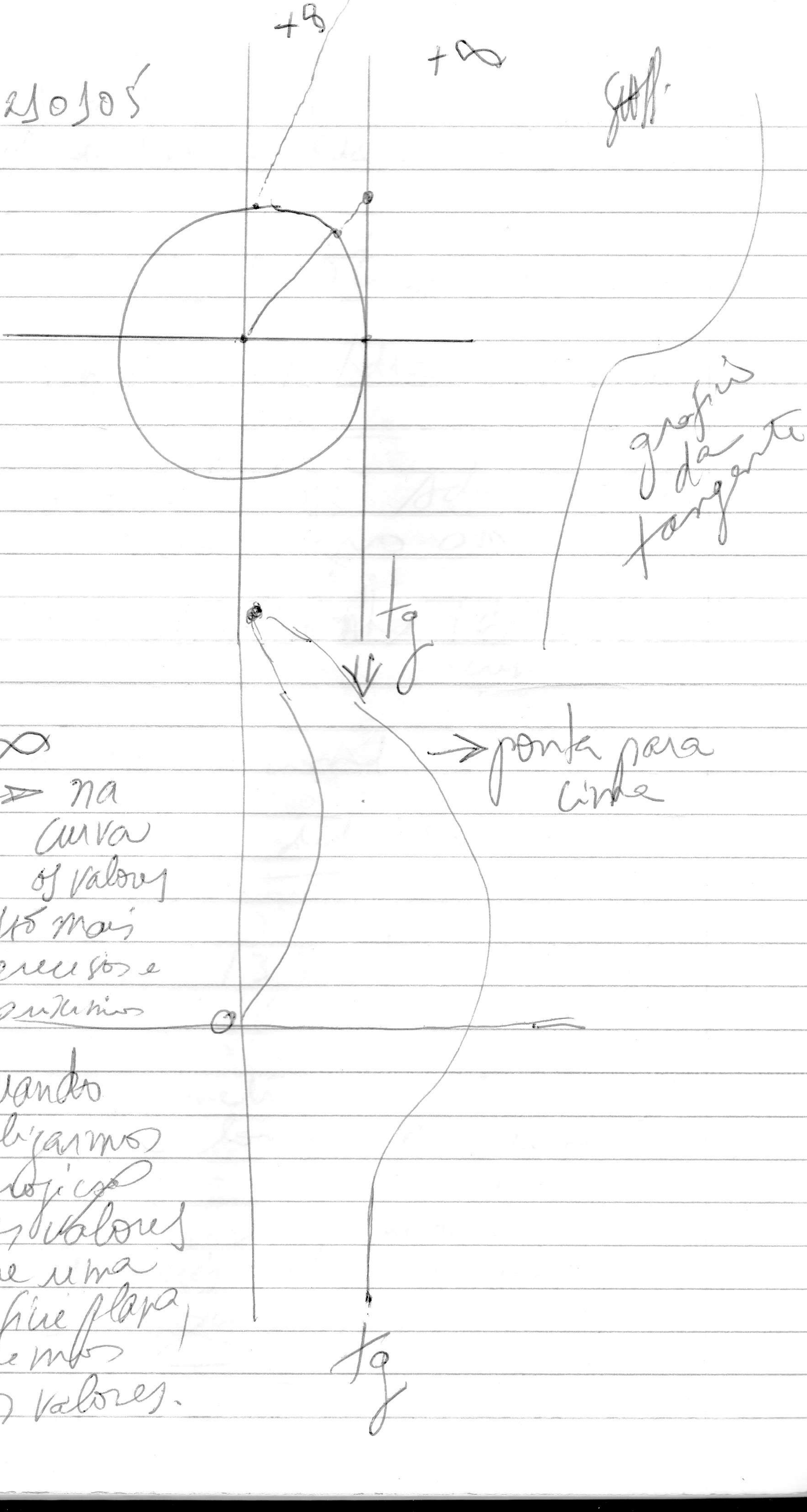
+∞
+∞
gráfico da tangente
tg
ponta para cima
tg

+∞
↳ na curva os valores são mais precisos e próximos

Quando realizarmos a projeção destes valores sobre uma superfície plana, obtemos outros valores.

20210106 (20210106)
Upload Books-2021-02.7z

—

Terra = 1,4602 seg + rápido.

86400 seg/p/dia 19 Julho 2020

1970 → até hoje adicionado 27 seg no tempo atômico internacional.

31 Dez ou 30 Junho (datas que são acrescentados).

"leap second"

1 seg será deletado.

Universo 13,77 b anos ± 40 m

—

interferometria de linha de base muito longa. (medição de origem de uma fonte) de rádio no Universo.

—

(usa-se várias antenas separadas por milhares de quilômetros.

—

Scapegoat = Mata Hari
(Bode expiatório)

—

→

Compassivos

→

1.31-01 {
.03.001
Data PC.002
003
Drive_D
1_Matem
1_Pasta_Oculta

DataPC Part_02_001

Part_03.001
002

Data_PC_Part_04
}

—

'ad nauseam'

→ John Bardeen
1956 - Invenção do Transistor
1972 - Teoria da Supercondutividade

—

20210107 (20210107)

Books-002-2021\ General Relativity...\
1138589772.pdf

Pág. 69(88) - 3.4 Transformation Rules for Christoffel's Symbols

Pág. 83(102) Figure 3.5
Fermi coordinates adapted to an observer

Pág. 87(106) Chapter 4
The Curvature Tensor

Fig 4.1.

—

Books-002-2021\ General Relativity ...\ Spacetime and Geometry \ An Introduction to General... pdf

Pág. 93(108) Chapter 3
Curvature

3.1 Christoffel Symbol

Books - 002 - 2021 \ General Relativity ... \ grbook.pdf

Pag. 67 (97) -

2.12 Torsion-free covariant derivative

⇒ Christoffels symbols

Pag. 297 (327)

11.16.1

Riemann tensor a mixed coordinate-tetrad frame

Christoffel symbols

Pag. 451 (481)

3x6=18 torsion-free coordinate-frame spatial connections (Christoffel symbols) ...

20210108. (20210108)

Sakura → termo nipônico
(cerejeira florida),
que é símbolo budista da
impermanência.

—

sofrimento e desilusão amorosa

— —

San Disk Ultra
MicroSD HC

—

SWIFT BANCO BRASIL:
BRASBRRJ IBR
SPO

Relativistic
Quantum Physics
Ohlsson

Books-2021
Data-NovoVivo-2021
SV-2021
Videos-2021
ZIP-2021

⇛ 06 Jan 2021 — 1.22 TB (20210106)

CVGA2019-Videos

⇛ { 184.123.201.35.bc.
99.185.102.34.bc. ⟸

{ 173.193.186.35.bc.

=

Book1-002-2021 \ Cálculo \

Cálculo-vectorial-claudio-pita...pdf

Pág. 514 (523)

Figura 4

Curvas paralelas a la elipse $\alpha(t) =$

$(a \cos t, b \operatorname{sen} t),$

$0 \leq t \leq 2\pi$

=

20210110, (20210110)

Magnolia Program

Joseph's Egyptian wife
Asenath

36.75.98.34. bc

20210111

23°27'11.0"S
46°28'38.7"W

Cumbica

Import: ✡

Endereço no Google Drive →
Leibni.com

99.185.102.34. b.c.
googleusercontent.com

obs: Drives compartilhados.

gru14s13-in-f14.1e100.net

—

sarcasmo
ironia

—

20210111 → 21,16TB

—

Gravitational
Microlensing

—

5m\
\AppData\Roaming\Tracker Software
\PDFXEditor\3.0\Custom Sessions

—

Chag Chanukah Sameach

Keny Roger sel Stream

—

Comics and Cubics
Robert Bix

—

20210112 → 21,34TB

20210112

Introduction to Lorentz
Geometry
Curves and Surfaces 350 pages

Ivo Terek Couto
Alexandre Lymberopoulos

—

Lorentz Geometry is a very important intersection between of Mathematics and Physics being the mathematical language of General Relativity.

—

Max. Dados no Google Drive: 63,47 TB
20201227

—

cantilena

oligopólio

placebo

—

Books: 20210112

1) Books-002-2021\Tools of Radio Astronomy\3642399495.pdf

Pág. 483 (491) Fig 15.5

The energy-level diagram of the vibrational ground state of amonia NH_3, an oblate symmetric top molecule.

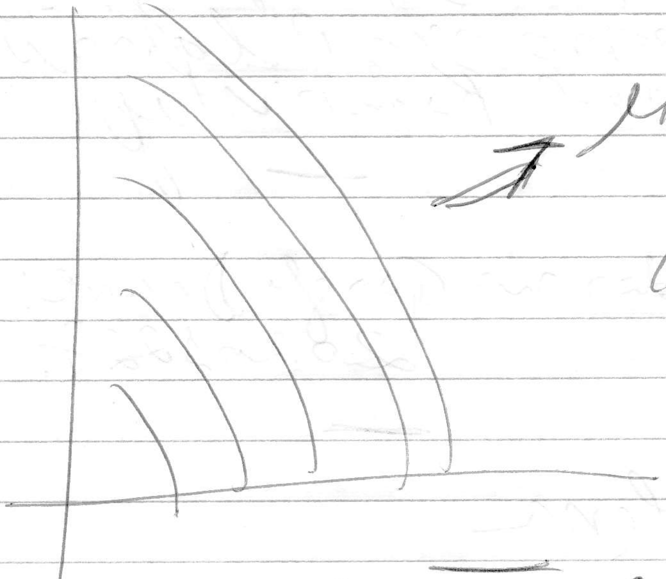

espaço curve (calcular a curvatura)

2) \Shifting the Earth\
\1180242273-Shifting.pdf

a) Pág. 20 (37) Fig. 2.5

the hippopede

obs: Algorithm.

b) Pag. 49(66) Fig 4.4

Intersection of hyperbola with ellipse: three intersections, the normal lines.

Pag. 53(70) Fig. 4.6

Pag. 54(71) the radius of curvature

Fig. 4.7 Intersections of normals associated with hyperbola. As the hyperbolas recede and become tangent to the ellipse, the intersections approach a point on the evolute.

Import

Pag. 56(75) the Osculating Circle

Fig. 4.9 the osculating circle at different points on a curve

1/ Pag. 85(102) - Figure 5.12

three consecutive oppositions.

3) Books-002-2021\ Pappus\
Geometry A Comprehensive
Course.pdf

Pag 363 - The Projective Plane
and Projective Space.

4) \ Pappus\ 2D Coordinate
Geometry.pdf

5) \ Geometry of Curves\
et45643yrt435....pdf

→ Elementary Geometry of
Algebraic Curves
(Curves)

Pag. 108(125) - Chapter 9
Projective Algebraic Curves

a) Pag. 112(129) Fig 9.2
Real projective line as
a circle

Fig 9.3 Complex projective
line as a sphere

Fig. 9.4 The projective plane over $\mathbb{Z}_2$

b/ Pag. 116(133) Fig. 9.5
Hemisphere model of $PR^2(R)$

c/ Pag. 125(142) - Chapter 10
Singularities of Projective Curves.

d/ Pag. 137(154) - Chapter 11
Projective Equivalence

e/ Pag. 148(165) - Chapter 12
Projective Tangents

6/ Geometry of Curves/
try2yv643ytr346-
geometry-of-curves.pdf

a/ Pag. 133(152) - Chapter 7
Curvature

b/ Pag. 151(170) - Chapter 8
Curvature: applications.

c/ Pag. 167(186) - Chapter 9

Circle of Curvature

d) Pág. 297 (316) Chapter 16

Projective Curves

(important)

20210113
21,6 TB —

3024,00,

—

math-cesf
↳ Leibniz —

—

Quantum Mechanics Konishi
Ezequiel 38 e 39 Gog e Magog

—

⟹ Math-Cochise —

20210115 IPTU
GUARULHOS
→ 938,52 —
2283,24

3221,76

⟹ 322,21

Alemanha
34 hospitais fechados
(2020)

2000 → 2242 hosp.
2020 → 1890 hosp

83 OAB
90 VIVO
173
180 casa
353

353
75 SANT
428
47 Telef
475

475
315 SANT IPTU
790

800

$\boxed{850}$

Vendatu JEE.

→ 29,7 Gb/memory.

20210116

Signal

—

369 + 17 seg
Nikola Tesla
Abraham Hicks
↓
ponto de combustão

17 + 17 + 17 + 17 = 68 seg

Momento exato de atração.

Pensamento: Puro
Positivo
Sem resistência

Duração: 33 Dias.

Escreva e sinta a emoção.

Frase:

Eu me sinto muito feliz e agradecido por prosperar no meu negócio próprio e na minha vida pessoal. hoje tenho a saúde que desejo e o dinheiro necessário para desfrutar do que há de melhor com a minha família. (3X)

Hora do almoço (6X)

À noite (9X)

= Lei da Atração.

33 dias = 3+3=6

45 dias = 4+5=9

=

20210117

21,73 TB (20210114)

—

"Extra Bavariam nulla vita, et si est vita, non est ita."

Neologismo "zu merkeln" significa algo como não ter uma opinião contundente sobre determinado assunto, ser passivo, hesitante.

"Mutti" → mãezinha em alemão.

Die Werke - Hannover
(20201031)

Leibniz's Works 4 volumes

Die Werke - Hannover - 9 volumes

Leibniz und ~~31~~
Landgraf 2 volumes

Leibniz-Works - 16 volumes
(04 Nov 2020)

18-17 5, 13, (13-18)

Trent University

Opera Omnia 1768 - Latin

This is the day of the The

Leibniz Works - 01/17/2021

12/10/2020

—

6s6 1124557 5 - 03

5579 —

13-18 and 23

—

bcb.gov.br/acessoinformacao/registrar-reclamacao.

Pappus Collection Mathématique

20210118

34-30

Graphical Approach to Algebra & Trigonometry
col

Math Sorcerer

Books:

1) Discrete Mathematics with Applications
Susanna S. Epp

2) Discrete Mathematical Structures
Kolman, Busby, Ross

3) A Transition to Advanced Mathematics

Mathematics Proof

Gary Chartrand

4) A First Course in Differential Equation with Modeling Applications
Dennis G. Zill

Ordinary Differential Equations
Larry C. Andrews

Geometry
Jurgensen, Brown, King

—

Music

Sugar Ray - Someday
Keane - Crystal Ball
Jet - Are you gonna be my girl

Baalbek

propileo (entrada)

Modern Geometry
Jurgensen

---

worldscientific.com

Quanta-of-Maths.

Geometry of Manifolds
Liviu | Nicolaescu
2021
701 pag.

---

worldscientific.com

Moduli Spaces of Real Projective Structures on Surfaces.
Alex Casella
ISBN 978-4-86497-096-9.

Geometric Analysis around Scalar Curvatures

Fei Han 220 page

ISBN 978-981-3100-54-1

20210119

21,85 TB

KURZ MUSS WEG

arsch = Sunde

Das volk erhebt sich!!!

argumento (alegado)

⇛ Gravitación

sinnet Sarbum

20210119

22,67 TB

$$1729 = \begin{cases} 10^3 + 9^3 \\ 1^3 + 12^3 \end{cases}$$

↑ Ramanujan

0
1
8
27
64
125
216
512
729
1000
1331
1728

0
1
4
9
16
25
36
49
64
81
100
121
144
169

$$\left.\begin{matrix} 3^2+4^2 \\ 5^2+0^2 \end{matrix}\right\} = 25$$

$$\begin{cases} 3^2 + 4^2 = 5^2 \\ 9 + 16 = 25 \end{cases}$$

$$\begin{cases} 6^2 + 8^2 = 10^2 \\ 36 + 64 = 100 \end{cases}$$

$$\begin{cases} 144 + 25 = 169 \\ 12^2 + 5^2 = 13^2 \end{cases}$$

(3, 4, 5) (6, 8, 10) → Ternas Pitagóricas

Lados de um triângulo
retângulo.

patreon.com

—

- An Introduction to Quantum Field Theory
- Quantum Field Theory and the Standard Model

Matthew Schwartz

- Renormalization Collins

—

20200120

ipfs.io

—

20210120 – 22,77TB

—

WIN+R

msconfig

Inicialização

Número de Processadores

—

gsuite.google.com/signup/edu/welcome

Antgmdio

milling = fresagem.

stemiprof@matterradimus.com
Gasa@3357

20210121

youtube.com/c/ILSTITR/

JohnHSchwarz

jhs@theory.caltech.edu

Unificatión das leis fundamentais
da física

video.ias.edu

Florence HEP (youtube)

nenúfares.

"natura non facit saltum"
A natureza não dá saltos
Darwin

→ é a questão das angiospermas?

"mistério abominável" ⇒ Deram um salto?

(100 milhões anos) No Cretáceo

anátema

Isontini Sloccata

ungarav@gmail.com

20210125

Modelo do Universo

Lambda – CDM
(Constante Cosmológica) (Cold Dark Matter)

"Se a geometria do espaço for euclidiana, o volume do universo é infinito; mas se a geometria for esférica, então o volume é finito"!

—

riemannian curvature tensor

—

youtube.com/user/eigenchris/

—

It⇒ BOLOGNA
COSULICH.

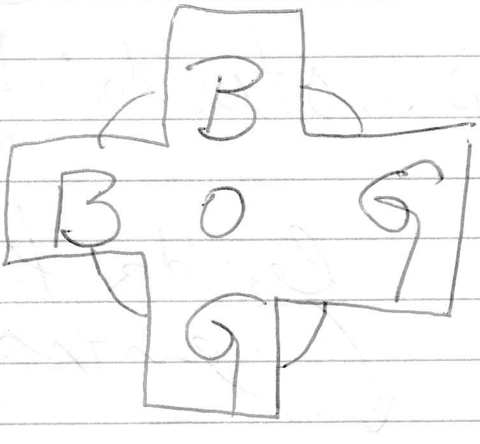

—

embargo
manifesto.

—

mp@unibo.it

—

20210126

ressonância orbital J01-128
18:9:6:4:3

Guatapoiranga

"Veado Vermelho ou Fera"
Vermelha

$$\begin{cases} 15\,000 \div 5,5 = 2,700.000,00 \\ 2727272,\overline{72} \end{cases}$$

⇒ MoneyGram ($)

help.studentionline@unibo.it

lei da termodinâmica e
entropia.
↳ desordem

a entropia só pode aumentar
nunca diminuir.

O tempo só avança na direção
em que a entropia aumenta.

Barbour → o que determina a passagem
do tempo não é o aumento da entropia,

mas o aumento da complexidade
sem limites de tempo
e espaço

NO Universo, a entropia não
aumenta, mas a complexidade
sim.

Lobachevsky

The God Problem: how a godless
Cosmos creates

Bloom, Howard K 1943-

2012 711 pag.

Holocausto - Shoá

I don't care that they stole my idea
I care that they don't have any
of their own
Tesla

ALYAH — MITZVA.

OLEH

Rabino Rony Gurwicz

20210130 Número.

— N Perfeito: 33550336 (1456 ano).

— N Perfeito: 8589869056
137438691328

Em 1588 — Pietro Antonio Cataldi.

— Leonard Euler (1772)
19 dígitos

— 51 números perfeitos
49.724.095 dígitos

A conjectura sobre os números perfeitos ímpares.

off:

"Todos os números perfeitos são pares".

Esiste números perfeitos ímpares?

universe expand
unevenly → desigualmente.

Lips like Sugar - Echo & the Bunnymen

Egito → Império Antigo
2686 - 2181 A.C.

"corveia real"

La Conjectura de Perelmán
Juan Soto Ivars

atoleimado.

⇒ VACHINÁ

pertinácia

20210201

Projective Geometry
Richter-Gebert Kortenkamp

Hyperbolic Manifolds Marden

Fodoanel.
Trevas do Rodoanel
Rodovia dos Imitantes
(não Imigrantes)

Molecular Astrophysics
Tielens

Hyperbolic Manifolds: An
Introduction in 2 and 3
Dimensions

---

jarosita

unidade de rede (google drive)

20210202 →

Judeu (Croata)
Ramon (Bolonhês do
Mattaná Mandaqui)

Somebody told me The Killers.

Funhouse Pink

20210205

jean@cis.upenn.edu.

Condado Travis, cidade Austin
(Travis)! Texas
↳ Samsung.

Data-NovoLivro-004-2021\
Book-004-2021\
Geometrical and visual optics...\
Geometrical and visual optics...pdf

Pág. 106 (119)

Figure 7-8 A e B

$n' = 1.333$

← estudar!!!

(veja curvature)

Obs: $F = \frac{n'}{f'}$ Lens Effectivity

Pág 102 (115)

Hyperopia

$$L' = \frac{n'}{l'}$$

constant

por qué?

$$L' = \frac{(1000)(1.333)}{2122\ \text{mm}}$$

$$L' = +62.82\ \text{D}$$

20210206

O futuro de ontem é o agora.
O agora é o futuro de ontem.

Calculus Moise

Stephen S. Shatz
sss@math.upenn.edu

20210207

voltar HD-1.36.02 → math-cesf.
(amanhã)

- Dance of the Clairvoyants
  Pearl Jam
- The Dead Heart
  Midnight Oil.

www.bolyai.hu/index_en.htm

mathunion.org

Why'd you only call me
when

—

Radio Rock.com 61

David Bowie - Let's Dance

Somebody that I used to know
Three day grace

Run-away Train - Soul Asylum.

—

Hyperbolic Manifolds Marden
An Introduction in
2 and 3 dimensions.

—

get-pdfs.com

Daffy Duck & Pork Pig vs
Nasty Canasta

—

20210208
Google Workspace Essentials

—

curves and surfaces in geometric modeling
Gallier.

JEE Maths - Playlist - JEE Main 2021

96 USA   103 USA
1 HK.   1 HK

Vila Antonina

630

Anhangabaú

20210215

23h 56min 4,09053 seg

40.075 km (circunferência)

460 m/s = 1.656 km/h (rotação)

365 d 6 h 9 min 10 seg (translação).

940 milhões km/ano

30 km/s = 108 000 km/h.

Sistema Solar = 220 km/s = 792 000 km/h

Ano Galáctico = 225.000.000 anos.

Via Láctea cruzando o Universo em direção das constelações de Leão e de Virgem

navegamos 630 Km/s

2.200.000 Km/h

—

460 m/s   0,46
30 Km/s   30
220 Km/s   220
630 Km/s   630
            880,46 Km/s.

—

The Mission - Severine

—

⇒ https://admin.google.com/u/1/

Cidade Jardim 19.973
Parque Ibirapuera 22.500   preço
Ibirapuera 23.000

—

gsuite.google.com/signup/essentials/welcome
/businessstarter/welcome

—

Av Cambucha

—

20210216

A preguiça morreu de fome.

Private Idaho B-52's.

Frango US$ 4,50 R$ 24,50

5,45

24,525.

$24{,}5 \div 4{,}5 = 5{,}\bar{4}$

$23{,}5 \div 4{,}5 = 5{,}\bar{2}$

$22{,}5 \div 4{,}5 = 5$

$21{,}5 \div 4{,}5 = 4{,}\bar{7}$

$0{,}\bar{2}$ $(0{,}\bar{2})$

$25{,}5 \div 4{,}5 = 5{,}\bar{6}$

$26{,}5 \div 4{,}5 = 5{,}\bar{8}$

$27{,}5 \div 4{,}5 = 6{,}\bar{1}$

$0{,}3$

Profecia Ezequiel — o perigo vem do Norte.

Gielis curves Source: the Geometrical Beauty of Plants.

01 → 23, 24, 25, 26 e 27
02 → 02, 03, 04, 05
14, 15, 16, ...

Books-004-2021/Thermodynamics Principles/0128219408.pdf

Pág. 86 (96) –

Statistical Treatment of Nonmetallic Solids

... this relation, commonly known as Einstein's (1905) law of heat capacities, correctly corresponds to the limiting cases

$C_v = 0$ at $T = 0$ and to

$C_v = 3R$ at infinite temperature – the law of Dulong and Petit.

20210217.

The Black Eyed Peas - Pump It.

=

Vingadores Era de Ultron

—

Life Lines - AHA.

=

Hypergeometric
Supertrigonometric
Superhyperbolic Functions

=

colegiado judicante.

$$\left\{ \frac{cos}{sen} = tg \right.$$

—

- Amy Winehouse - You know I'm no Good
- Foster the People - Pumped up Kicks

The Clash - Rock the Casbah

Morrissey the best of 1997.

—

E is imaginary Number

$ac-bd, bc+ad$

$x+iy$

Fargo -14

20210218

Rodov. Raposo Tavares

Av. 23 di maios (de maio)

Atrasado está mas todo mundo vai chegar.

—

gdriveunlimited.net

—

gsuite.google.com/signup/businessstarter/welcome

—

BkFast.vn
SeoOnDay.net →

—

apply.cochise.edu

gd.404edu.workers.dev

Vector Calculus Marsden
Jerrold E. Marsden

Disappear INXS

Love will tear us apart
Joy Division

workspace.google.com/signup/edu/welcome

Deuterio 2.013553212535(17)
atomic units

Hydrogen
molecular 3.021378241561(61)
ion HD^+ atomic units

Hyperbolic manifolds Harden

Prescribing the curvature
of a riemannian manifold
Jerry L. Kazdan

Gaussian Curvature

Pag. 01 (09)

Source: Books-004-2021/
Prescribing the curvature..../
Jerry L. Kazdan-Prescribing
the curvature...pdf/

obs: surface in R^3

As sometimes occurs, one of the most intriguing unresolved problems involving Gaussian curvature is also one of the easiest to understand.

Let $z = u(x, y)$ be a surface in $\mathbb{R}^3$.
Its Gaussian curvature
$K(x, y)$ is given by the formula

$$\frac{u_{xx}\, u_{yy} - u_{xy}^2}{(1 + |\nabla u|^2)^2} = K(x, y).$$

Solucionar!!!

MS Bank → Cloud Break.

Chudnovskys (Pi)

20210219 Preços

R$ 8,54 20210219 Gasoline Europe

€ = 1,31

6,52

€ = 1,37 Gasoline 1,37 × 6,52 = 8,93

€ = 1,46 Diesel

0,629 LGP

€ = 1,44 Gasoline $\overset{\times}{=}$ 6,52 = 9,38

20210220.

Matheus 13,42.
,50

Matheus 13, 57

"Não há profeta sem honra, a não ser na sua pátria e na sua casa"

Matheus 16, V.16

Tu és o Cristo...

17

→ dado conhecer por Deus

—

Pearl Jam — Dance of the Clairvoyants

Matheus 13, vers. 42

"E lançá-los-ão na fornalha de fogo; ali haverá pranto e ranger de dentes" gnomacha = museu.

Dr. Edgar - era contador (dono)
R. Vasco da Gama.
foi contador do meu pai
e da Longada & Longada

depois o marido da Laurinha comprou dele.

nº 13.839 shows Tabajara
Maestro Severino Araújo (95 anos) idade 2012

Orquestra Tabajara
72 anos de tradição

Fundada 1934.

Inúmeras Mudanças
R Alto Paraguai 204
Jardim Brasil

Obs: 20210221

πr^2 (Área Círculo)

$r=2$

$3.14 \cdot 2^2 = 3.14 \times 4 = 12,56$

1/4 do círculo = Área 3,14

(raio 2)

Faça cálculos com $r=1$ e
$r=2$

Simule!!

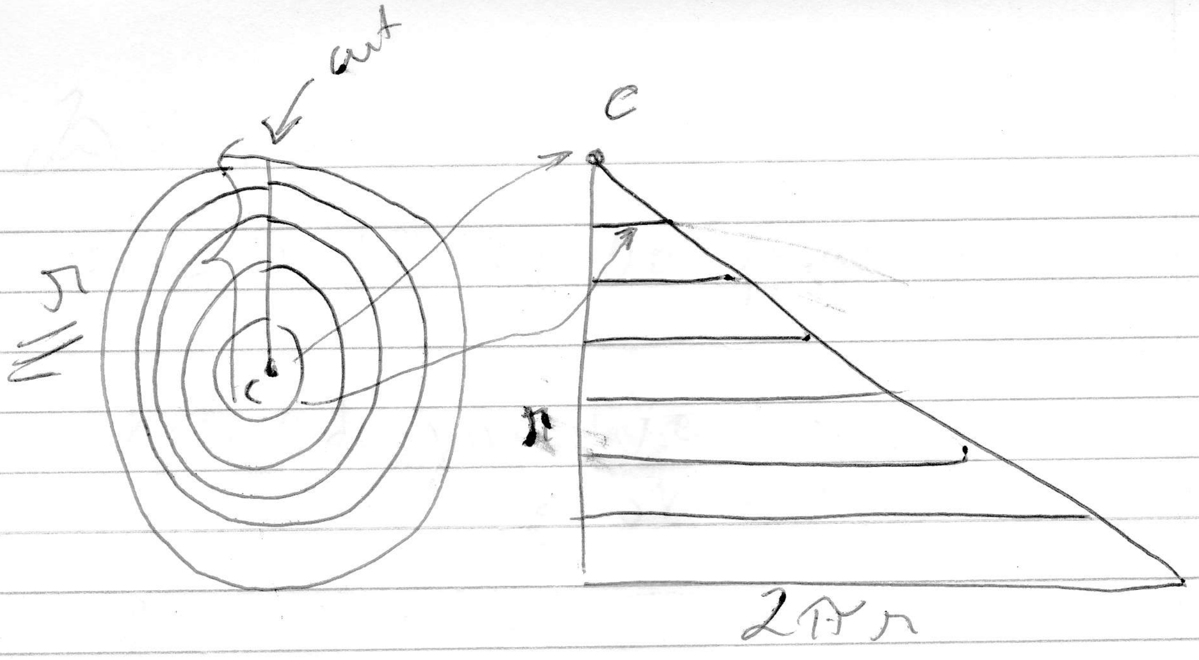
cut
r
c
C
r
2πr

$$\frac{\cancel{2}\pi r \cdot r}{\cancel{2}} = \pi r^2$$

(Area)

20210222

36.75.98.34.bc.googleusercontent.com

—

What do you want from me
Monaco

Manipulação —

1. Deixar a atenção
2.

— Flo Perdida

2º 13 Blackout
2º 4 Pilom sumiu
1º 14 Signo Pedra
1º 1 — —

Sanet.ws

—

Geometry Projective Richter-Gebert
Ulrich Kortenkamp

The Langevin Equation

Physics Halliday

20210223.

828,57 Gb.

—

Aplicativo "Meu gov.br"

—

Gêmeos → combina { Áries / Leão / Sagitário

—

I Am mine Pearl Jam.

—

AKAMAI.

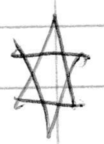

—

Proporção Áurea

1,618

$$\varphi = \frac{(1+\sqrt{5})}{2}$$

1,618033988...

$$a/b = (a+b)/a = \varphi \text{ "phi" } = 1.61803...$$

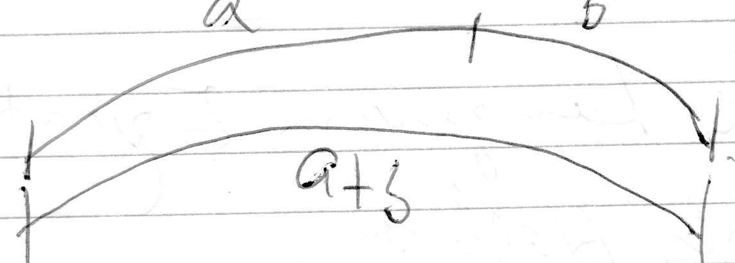

Phídeas
5º A.C.
séc

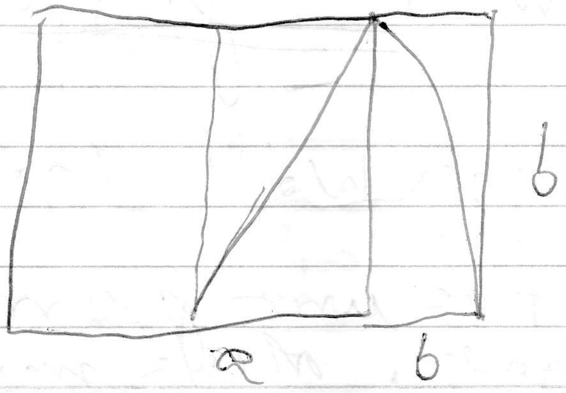

Fibonacci

0 1 1 2 3 5 8 13 21 34 55 89 144

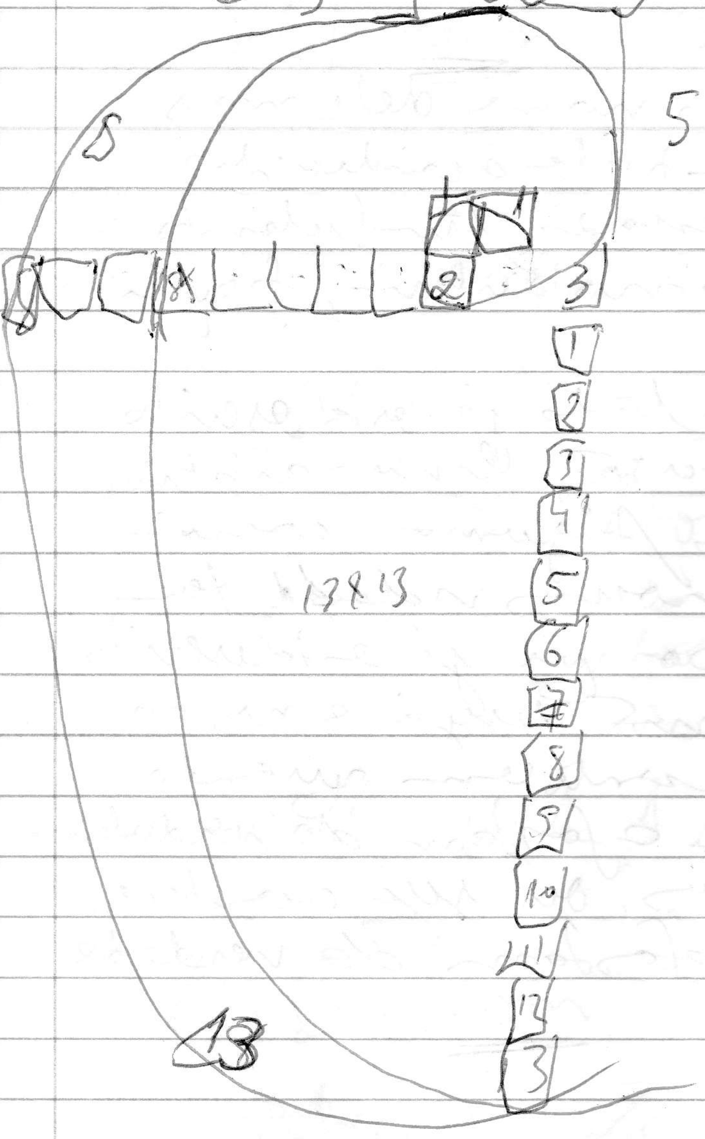
8
5
1
2
3
8
1
2
3
4
5
6
7
8
9
10
11
12
13
13x13
13

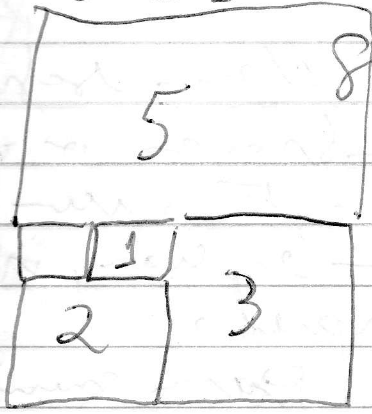

Sequência de
Fibonacci
+
Retângulo de ouro
=
espiral de
ouro

Tem no Pentagrama.

Tigger Wodds (golfista)

eu tive uma premonição deste jogador. Nesta manhã, veio em minha mente da vida conturbada deste atleta e no período da tarde, saiu na televisão que o mesmo sofreu um acidente de trânsito na Califórnia

Nem sabia o nome dele mas lembrava o que tinha acontecido durante o seu passado turbulento e sem esperar, vi o noticiário, fiquei surpreso

Para mim, tudo já está escrito e o nosso poder de decisão, livre-arbítrio, é muito limitado. Algumas coisas marcantes em nossas vidas tem que acontecer porque já está escrito por Deus. Deus não deseja o nosso mal, os fatos acontecem quando as pessoas se afastam do verdadeiro caminho de Deus, ou seja acontece àqueles que se afastam da verdade.

Bands

Vurgaftman

—

Tesla 3,6,9

{ Tesla
Edison

—

Matemática de Vortex

1 2 4 8 7 5

sequência

3 6 9

↳ vetor 3ª e 4ª dimensões

Campos de Fluxo

—

2207R USA 1 ano.

—

Lerbmilcom → 828,57 Gb.

~~20210~~20 20210224

C:\Leibniz\02

uni-goettingen.de

Carl Friedrich Gauss Werke

Little girl lost Twilight Zone

---

library.lol

20210226 - 1,16TB - (Ref. 20210223)

Lilian Sherman El Mann

→ 136866

Richard Gott 27/12 63,47TB

jrg@astro.princeton.edu.

(Robert Vanderbei)

rvdb@princeton.edu (math)

(David Goldberg)

goldberg@drexel.edu (Physics)

arxiv.org/abs/2102.08176

Miss you – The Rolling Stones

20210301 III

expressões algébricas.

monômios → parte numérica (coeficiente)
              → parte literal

2xy

(não tem adição, subtração...)

monômios semelhantes

$$\underset{\neq}{2}xy \quad 3\underbrace{xy} \rightarrow \text{parte literal iguais}$$

---

termos dos polinômios
(termos)

20210301 - 7,16TB (ref. 20210226).

127
110
30
47

47 - ☒ 235 605
43 - ⊔ 129 360
17 - ⊔ 51 965
20 - L 40.
455
150
605

Padre 20210301
Delan
Cuervo
— TV Pai
Eterno

965 Gb.
20 Geom Algibris
985
40 ZIP-ALL
1025
200 Backup
1225

465 HD
2
930

2HD

930
465
1395

1334 cap
Bitcva

20210302

1,1
1395
465
1860

4HD

1225
600
1825 Gb
100 ARQ-IMP
1925 Gbytes
50 ARQ
1975

mathematicsnew55@
gmail.com

mathe-1-CESF

ID de Cliente
02uwel9h

20210303

1925
50
1975 Gb
150
2125 Gb

20210303 → 7,7 TB
(ref. 20210228)

172.217...

Dolar 6,3420 (comprar)
5,5393 (vender)

0,80

Euro 7,5013 (comprar)
6,5816 (vender)

Cartão Viagem 8,1311 (euro/real) (comprar)
6,6717 (vender)

2.125 Gb
250 Data Geral
2.375
350 PC-O-ALL
2.725
65
2790 ⟺

6 x ☒1

42
6
252

60 x 6 = 360

465

2325
465
2.790

5HD

6HD

20210304

2790

172.217.192.189

Called out in the dark
Snow Patrol

Serviços → INSS → "Prova de Vida"

$\sqrt{200} = 14,142135$

$$\begin{array}{r} 2\overset{1}{7}90 \\ 350 \\ \hline 3.1\overset{1}{4}0 \end{array}$$

Data-6-7

$$\begin{array}{r} {}^{4}4\overset{3}{6}5 \\ 7 \\ \hline \underline{\underline{3255}} \end{array}$$

7HDs

8.06 TB
(ref 20210301)

8.28 TB
(ref. 20210302)

---

escleroterapia. (gluose
da vena)

Advances in Analysis and
☐ Geometry
Jie Xiao
Alexander Grigor

---

teamdrive.mfoxx.workers.dev

20210308 (20210305)
Annalisa Godoli

"studio - immatricolazione
università"

Hyperbolic Manifolds An Introduction
(Albert Marden) in 2 and 3D

Condritos carbonáceos

6 nº que definem tudo o universo

1. matéria bariônica
2. matéria escura
3. energia escura
4. profundidade óptica
   (mais astrofísico de todos os parâmetros do modelo padrão da cosmologia)

✱ Para determinar a profundidade óptica do universo é feita uma medição da polarização da CMB

CMB – radiação cósmica de fundo em micro-ondas.

Polarização
intensidade
comprimento de onda } 3 características de uma onda de luz.

5. } flutuações quânticas
6. } (diminutas)

Chamado de espectro de potência primordial e descreve as flutuações na densidade do Universo no espaço tridimensional.

Parâmetros restante, chamado índice espectral escalar.

As pequenas variações de energia que estavam presentes no Universo primitivo, dependem da escala angular.

O ruído cósmico primordial é ligeiramente rosado.

Universo: 13,8 bilhões de anos
(40 milhões de anos variação ±)

A observação mais restritiva é a da anisotropia da CMB: as diminutas flutuações da temperatura.

$360 \div 13 = 27{,}\overline{692307}$

$360 \div 11 = 32{,}\overline{72}$

$\boxed{360 \div 15 = 24}$ hours

$360 \div 16 = 22{,}5$

$360 \div 18 = 20$

Wrinkling = enrugamento

ver Magazines (Universo)!!!

Google Workspace Enterprise
0800 724 8154

Enterprise Plus (Ilimitado)

https://support.google.com/a/answer/10077880

Standard ⇒ Tb unlimited

Mathematische Annalen 306 vol.

PPN235181684

Lips Likes Sugar

20210312

George Dantzig → atrasado na aula de Matemática.

George Bernard Dantzig
Quaisson

8-11-1914

→ geometry.

1936 – licenciatura Math e physics.

Good Will Hunting 1997

dfg.de

280

A direct method approach for the existence of convex hypersurfaces with prescribed Gauss-Kronecker curvature.

Source: www.digizeitschriften.de/dms/img/?PID=GDZPPN002439964

Google Takeout

⇒ Filme: Good Will Hunting

04min56 – Solucionar problema.

↳ obs: 03:57 – Sistema Avançado de Fourier

Solution: 08:26

20210313

opendrive.com

https://webdav.opendrive.com

---

x+y filme dublado.

---

nº Antigo 35645070 Stenio (casa)

$\pi = \mathbb{Q}$ número racional é um quociente de 2 números inteiros.

$\mathbb{Q} = \frac{a}{b}$ $a \in \mathbb{Z}$

$b \in \mathbb{Z}, b \neq 0$

$\pi \neq \mathbb{I} = \mathbb{Q}$

↓

Alt + 227

$\underline{960}$

Alt + 241 = ±

251 = ¹

252 = ³

253 = ²

$\pi \neq \mathbb{I}$

Ano

Judeus = 5780

4719

—

#

Alt $\overset{\neq}{+}$ 8800

unicode-table.com

$\mathbb{Q}$ = U+211A

$\pi \in \mathbb{Q}$

$\mathbb{Q}$ = Alt + 08474 &#8474

$\mathbb{I}$ = Alt + 120128

$\pi \neq \mathbb{I} \in \mathbb{Q}$

↳ Alt + 211A

$\in$ = is an element of

not =

$\pi = \textcircled{X}$ ↓

$\pi = \pi$

π = Alt + 960

Font MUS:

Local/Microsoft/Windows/EUDC/

mus-20210314.TTE

$\mathbb{Q}$ = Alt + 8474

Font: Microsoft Sans Serif

Symbol font

Alt + 8474 = →

=

Editor de Caractere Particular

→

Execute: eudcedit

Use no Mapa de Caracteres:
Todos as fontes

MUS
mathematical

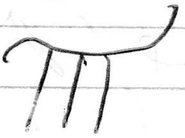

Uniforme
~~Specification~~ Space

Coloquei no Website em
20210314

202103150357

Vocês estão querendo construir um túnel para o futuro usando uma colher de chá! Vocês estão equivocados, vocês devem usar a dinamite. Vocês devem implodir a forma como fazem matemática, física, química etc.

David de Michelangelo foi construído com uma colher de chá!

Precisamos ir além e unir a colher de chá e a dinamite.

A constante de expansão do Universo (Hubble) está sendo calculada de forma imprecisa pois vocês estão usando conceitos e cálculos ultrapassados. É chegada a hora de quebrarmos os grilhões que nos prendem ao passado como uma bola de ferro presa aos nossos pés. Para voar precisamos deixar o chão que achamos ser seguro e que muitas vezes poderá se abrir bem embaixo dos nossos pés.

20210315

Die Vermessung der Welt
(Gauss)
Medindo o Mundo

Berhard Rieman

David Hilbert – queria lançar o plano de reconstruir os fundamentos da matemática, para torná-los consistentes e livres de paradoxos.

Gödel, com sua tese de doutorado, pôs um ponto final a essa pretensão, demonstrando que havia problemas na matemática que eram impossíveis de serem resolvidos.

A deslumbrante clareza e exatidão da matemática era, na realidade, um labirinto repleto de paradoxos.

Gödel provou: 1) Em um sistema formal axiomático consistente que possa expressar fatos sobre aritmética básica, há enunciados verdadeiros que não podem ser provados; 2) Que a consistência do sistema não pode ser provada dentro do mesmo sistema. São os teoremas da incompletude.

Zero:

Fortaleza de Gwalior na Índia Central.
(registro mais antigo do zero)

Índia – Shunya
(a idéia filosófica do vazio para representar o novo termo matemático).

Collected mathematical works.
Bohr, Harald August 1887-1951
3 Volumes 1952

---

nbarchive.dk/publications/
bcw/
(Niels Bohr Collected Works)
13 Vol.
ISBN 978-0444532862

8504 pages
51.8 pounds
Obs: Amazon
Niels Bohr - Collected Works

AbeBooks.co.uk (co.uk)
bi=12677084031 £ 4,880.36

terrytao.wordpress.com

---

steniomusich.com

Quantum Field Theory in Curved Spacetime

Parker

Usf 9834

Chine 9597

Alaska 1718

Google Ads

727-753-3719

Pearl Jam – Sirens

→ (Neue Astronomie

1929

---

sebodomessias.com.br

"Arte não é o que você vê, mas o que você faz os outros verem"

Edgard Degas (1834-1917)
pintor francês impressionist
artist.
Ed. 1214 Revista Dinheiro

---

edgar-degas.org

waffle ⁙

team.gdrive.vip
pon.workers.dev

20210320
Robert P. Massé

Physical Field Vectors and Tensors
ASIN B0842Y64HT

ISBN 978-1-7342447-5-5

Introduction to Tensor Analysis and the Calculus of Moving Surfaces
Pavel Grinfeld

z lib.org

6r 1lib.org

→ 718 Foundations of Physics
→ 1424 Physical Fields Vectors and Tensors
→ 1785 Physics when it went wrong
→ 1929 Prime Numbers and Congruence Theory

2330 Space-Time Dynamics of an Isotropic

2419 – the Trigger

120.
6 = 28

musick.page

Assurbanipal

fisica.net

Lucy Thomas

Back Blaze

sync.com

netmonis

mathematics.page

mathematics.one

$40 \div 6{,}5 = 6{,}\overline{153846}$

$40 \div 5{,}5 = 7{,}\overline{27}$

$40 \div 4{,}5 = 8{,}\overline{8}$ ⇛ > 1,53

$40 \div 3{,}5 = 11{,}\overline{428571}$ > 2,628571

⇛ $40 \div 2{,}5 = 16$ (par) > 4,571429

$40 \div 1{,}5 = 26{,}\overline{6}$

$40 \div 2{,}5 = \boxed{16}$ $360°/16 = 22{,}5$

×

$40 \times 22{,}5 = \boxed{900}$

360 :

1 → 22,5
2 → 45
3 → 67,5
4 → 90
5 → 112,5
6 → 135,0
7 → 157,5
8 → 180
9 → 202,5
10 → 225

11 → 247,5
12 → 270
13 → 292,5
14 → 315
15 → 337,5
16 → 360

$22{,}5 \div 40 = 0{,}5625$

$\boxed{22{,}5 \div 50 = 0{,}45}$

1 ímpar

$50 \div \boxed{2,5} = 20$ $360 \div 20 = 18$

$50 \times 18 = \boxed{900}$ partes

$18 \div 50 = 0,36$

18
36
54
72
90

108
126
144
162
180

198
216
234
252
270

288
306
324
342
360

$360 \div 0,36 = 1000$

$\pi = \frac{C}{D} = \frac{360°}{100 \text{ mm}} = 3,6$

R = 0,50

$(50 \div 0,50 = 100)$

$\pi =$
$3,14 - 3,6$
$0,46$

$3,1415 - 3,6 =$
$0,4585$

$3,14159 - 3,6 =$
$0,45841$
erro

π π

$$3{,}6 \div \pi = 3{,}6 - \pi =$$

$$\underbrace{0{,}45}_{\downarrow}84\underbrace{07\ldots}_{\text{error}}$$

$$3{,}6 - \pi = 0{,}45 \Rightarrow 0{,}4584$$

$$\pi = 3{,}15 \Rightarrow 3{,}1416$$

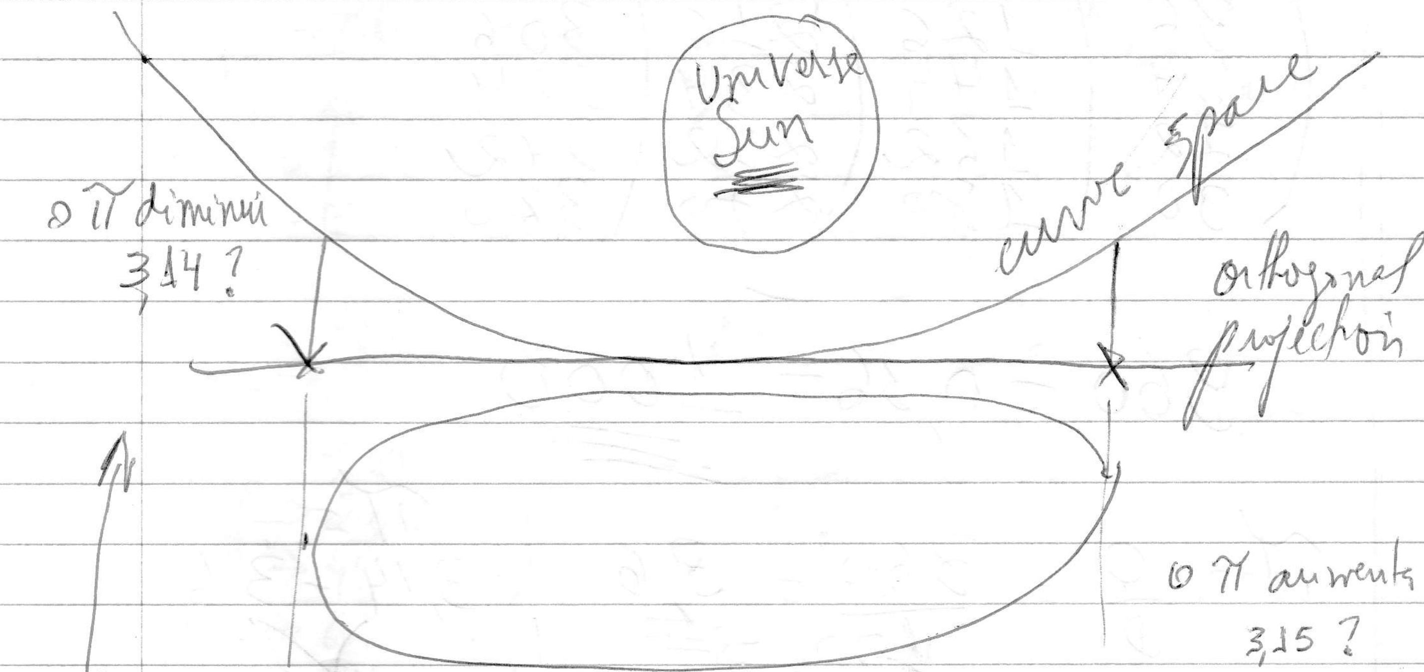
Universe Sun
curve space
o π diminui 3,14?
orthogonal projection
o π aumenta 3,15?
gera a curvatura

Projeção ortogonal externa

Veja anotação 25/01/2020 (107)

A luz exerce uma pressão constante, conhecida como "pressão de radiação"

Veja:

Anotação 108 (110)

110 (02/02/2020)

Gauss foi o primeiro a testar a curvatura do espaço físico:

179° 53' 59.32"

obs: Não há geodésicas paralelas

A History of non-Euclidean geometry

B. A. Rosenfeld

Pág. 280

Osculating circle at that point

etc.

(180 = π) $179° 59' 59.32''$

$180 \times$
$60 \times$
60
648000 sg

$0{,}68''$ ↙

648000 - 100
0,68 - x

$- 1{,}049382716^{-4}\%$

÷ 101,049382716 = 3,10896
÷ 1,049382716 = 2,9937529999

π - 1,049382716% =

(÷ 1,[illegible]01049382716) = 3,10896768

Correção: $1{,}049382716^{-4}$

0,0001049382716 %

÷ 1,0001...

π ÷ 1,0001... =

π = 3,14126301487837089713...

Fundamentals of Mathematics
Behnke

Vetores   Tighten

{ módulo - tamanho   Grow
  direção - ângulo
  sentido - para onde

Grandezas:
{ Escalares
  Vetoriais

Conics and Cubics Bix

20210322
www.digitale-sammlungen.de
bsb-muenchen.de

“Porque a sabedoria serve de defesa,
do mesmo modo que o dinheiro!
Eclesiastes 7.12

Quem toma emprestado é escravo
de quem empresta.
Provérbios 22.7

"Vocês foram comprados por
alto preço;
não se tornem escravos de
homens."
1 Coríntios 7:23

"Tudo tem o seu tempo determinado,
e há tempo para todo o propósito
debaixo do céu."
Eclesiastes 3:1

7 Segredos da Prosperidade Judaica:

1) Ser perito Provérbios 22:29
faça, seja o melhor.
2) Tem uma cultura empreendedora.
"Ninguém fará por você aquilo
que deve ser feito por você."
3) Ter mentores
"Quando não há conselhos os planos
se dispersam,
mas havendo muitos conselheiros
eles se firmam."
Provérbios 15:22
4) Aceitar a repreensão
"Pobreza e afronta virão ao
que rejeita a instrução, mas o
que guarda a repreensão será
honrado".
Provérbios 13:18

4) Cont.

"Melhor é a repreensão franca do
que o amor encoberto"
Provérbios 27:5

"Todo o que ama a disciplina
ama o conhecimento,
mas aquele que odeia a repreensão
é tolo"
Provérbios 12:1

"Quem acolhe a disciplina
mostra o caminho da vida,
mas quem ignora a repreensão
desencaminha outros"
Provérbios 10:17

5) Ter um bom nome

"Vale mais ter um bom nome
do que muitas riquezas;
e o ser estimado é melhor
do que a riqueza e o ouro"
Provérbios 22:1

→ Boa reputação.

6) Guardar dinheiro

"Há tesouro precioso e azeite na
casa do sábio, mas o homem insensato
os devora" Provérbios 21:20

7) Ter união:

"Como é bom e agradável
quando os irmãos convivem
em união!"
Salmos 133:1

The one I love REM

@
A

cmd
ping
216.239.32.21

Contador : mando da
Lanterna

Rua Vasco de Gama, 25
perto Igreja do Brás

Misterious Ways U2

20210326

Quark bottom

O modelo padrão não pode explicar a natureza da gravidade.

quarks bottom devem decair em números iguais de partículos de elétrons e muons (uma partícula elementar semelhante ao elétrons)

Na realidade, produz mais elétrons do que múons.

Partícula ainda não descoberta → leptoquark

Padrão ouro → nível chamado
cinco sigma
(1/3,5 milhões)

Acelerador é três sigma – 1/1000.

—

"A descoberta de uma nova força na natureza é o santo graal da física de partículas".

off:

X17 → 5ª força da natureza?

Decaimento do elétron (conflito)

↳ Hélio

As partículas do átomo foram divididas em ângulos de 115 graus, muito mais do que o esperado.

No momento que O átomo decai, o excesso de de energia produzidos criou brevemente uma nova partícula desconhecida, que se decompõe rapidamente em pósitron e elétron.

33X mais massa que um elétron

= 17 megaeletronvolts.

Este decaimento também aconteceu com o berílio-8

Berílio

No berílio o ângulo foi 140 graus

Obs: À medida que a energia da luz produzida pelo berílio aumenta quando libera um elétron e um pósitron, o ângulo entre os dois deve diminuir.
Mas não foi o que aconteceu no experimento de 2016.

Pode ser um elo entre o mundo visível e a matéria escura.

Analysis Godement 4 vol.
Topological Methods,
Variational Methods.
Brezis

Pretty Fly the Offspring

→ Aniversário Stelvio 28/03

Pag. 67 Fábio Prieto de Souza (Desembargador)
www.fab.mil.br/
Download/arquivos/
medalhas/ALMANAQUE
AGRACIADOS OMA 2020.
pdf

3lib.net

Stupid Girl Garbage

technicalbookspdf.com

O livro de Eli filme.

Le210331

https://admin.google.com/u/1/ac/reporting/audit/drive/

Create
UEFI bootable USB with
rufus tool.

rufus.ie

Fale deles o que você é,
acuse-os daquilo que
você fez.

Painted on my heart
the Cult

Stressed out
Twenty one pilots

Walk on U2

→ ZIPAR Books-2021-05

STANACH

20210401

STANACH

Piazza Santo Stefano (Bologna)

Basilica di San Domenico (Bologna)

—

⟹ Upload Books-2021-05

Ehrhard Behrends - π & Co.

—

Pi und Co (2016)

Pag. 157 (159) 1.3 Falta

=

—

veja

Ver

π und Co (2008) → Pag. 149 (148)

Pag. 155 (158)

$$\boxed{e\,d\pi \le 2 \le r\,d\pi}$$

"Whoever calls names is called that himself".

We always see our own qualities in another person and think that he/she is like ourselves.

reckon (acerto de contas)

Ramsiya Ramanathan
stat.tutorsm@unibo.it

You believe in love. truth

Old fire

Some and angels
right now

$\pi = 1707.$

William
Jones

Crazy turn way
tws

"Summa cum laude"

20210403

Pendrive:

ZOTAC Win USB Maker

---

netcombo.com.br
stenioprof@yahoo.com
Mcfe22@33

→ Noemi Corsini

⇛ Testar WebCam com 2 computadores ligados (reflexo nos óculos.

Gean 20210404 Páscoa
vida NOVA!!!

$$\boxed{\frac{a^m}{a^n}=\frac{1}{a^{n-m}}}$$

Steiner Chain

Wikipedia

Ask Ethan: Where is the Line between Mathematics and Physics?
https://br.pinterest.com/pin/340373684332567737/

* * *

https://www.symmetrymagazine.org/article/the-deconstructed-standard-model-equation

* * *

Curva e integral en WA
https://br.pinterest.com/pin/709809591255709470/

* * *

→

$$\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi} \rightarrow \pi$$

e y pi

relación normal

20210414 (20210414)

Hold that Rock

216

— Conjectura de Poincaré

Objeto homeomorfismo

"geometria flexível"

→ duas formas são a mesma se for possível transformar uma em outra sem quebrá-la.

"A monetização do êxito é o máximo insulto à matemática" Perelman G.

—

"Se a teoria está correta, não necessita de outro tipo de reconhecimento"

Perelman

"Não me interessa o dinheiro ou fama. Não quero estar em exibição como um animal em um zoológico"

Perelman

=

20210418|||

STEM
Science, Technology, Engineering
and Mathematics

https://archive.org/details/
@stemisprof

—

La conjetura de Perelman
Juan Soto Ivars
ISBN 9788466608565

—

US	18
Cor S	5
Alem	1
Japão	1

20210419 Dia de Santo Expedito.

Lezione de Geometria Differenziale

—

20210422

fabricante cimento franco-suíço
LaFarge Holcim
1400 funcionários — out.
maior fabricante de cimento no mundo.
90.000 funcion. mundo.

Estrelas fim da vida são sustentadas pelo princ. de degenerescência.
→ Estrelas de Nêutrons.

—

"A vida é tudo o que você faz com ela."

"Estude o que te emociona de forma quase íntima."

"Uma conquista só tem valor se ela te custar esforço!!" Janot

—

ROMA 21 ABRIL 753 A.C.

Hoje: 2774 anos
(ontem)

= 970 A.C — 586 A.C

1º destruição Templo de Jerusalem
586 A.C. pelos exércitos da Babilônia
Rei Nabucodonosor II

2º destruição: 6 Ago 70
(515 AC — 70 DC) 8 set 70 (A12.com)

2º Templo → reconstruído pelo Decreto
Ciro II da Pérsia
(Livro de Esdras Cap 1, 1-4)

---

nós górdios → Górdio, rei lendário
da Frígia, antiga região
da Ásia Menor (Turquia)

Gordian Knot is a legend of Phrygian Gordium associated with Alexander the Great.
→ metaphor for an intractable problem solved easily by finding an approach to the problem that renders the perceived constraints of the problem moot
("cutting the Gordian Knot").

---

escárnio → zombaria

sarcasmo → zombaria nos opostos

ironia.

20210425

- Wilhelm Reich
- Efeito Kurdyumov
  → efeito de memória formada.

Ti (Tetanium) 45%   Ni (Nickel) 55%

- Royal Raymond Rife (médico)
  "15 vírus"

—

whomever = quem quer que seja

—

OnlyOffice

—

Workspace Google
0800 724 8154

—

Namaguiri
Cauchy
T. Cauchy – contorno círculo.
n → ∞
erro → 0.

—

1729
G. H. Hardy

20210426.

Scolded,,

Vince Higgs vhiggs@cambridge.org

—

estamento adm.

—

C - P on M
Sut M on 38

Extra Banana R$1,89

—

Third Thoughts

Steven Weinberg
May 3, 1933

March 23, 2021

Quantum Field Theory:
An Integrated Approach
760 pag. Eduardo Fradkin

Proof: A Long-Form Mathematics Textbooks
Jay Cummings
511 pag. ISBN 979-8595265973

$0,\overline{69} = 100x = 69,69$

$x = 0,69$

$99x = 69$

$x = \frac{69}{99}$ ///

off $\xcancel{1969 = 19\frac{69}{99}}$ melhor $1969 = 19\frac{207}{3}$

$69 = \frac{207}{3} = \frac{138}{2}$

$0,\overline{5} = 100x = 50$

$x = 0,5$

$99x = 49,5$

$\Downarrow$

$4,5/99 = 0,45$

$50 - 0,5 = 49,5$

$\frac{49,5}{99} = 0,\overline{5}$ ///

$0,\overline{5} = 100x = 55,5$

$x = 0,5$

$99x = 55$

$x = \frac{55}{99}$

---

wolframphysics.org/
technical-introduction/

—

20210429

Declaração de Balfour
02 Nov 1917

o governo britânico prometia ao povo judeu um "lar" na região da Palestina

—

Tratado de Sèvres
Agosto 1920

os otomanos cederam oficialmente as áreas acordadas, aceitando a criação dos mandatos britânicos e francês.

Palestina baseada na Bíblia

Lloyd George utilizou o Atlas da Geografia Histórica da Terra Santa, publicado em 1915 pelo reverendo escocês George Adam Smith, para definir os limites da região.

finally-we-may-a-path-to-the-
fundemental-theory-of-physics-
and-its-beautifull!

Os ratos não resistem às alturas.

"lugar de ser inútil"
Manoel de Barros

May 2021

20210501 UMA

H - 1.00794 1.00784
1.00811
He - 4.00260

1.00794 x 4 = 4.03176 -
4.00260

0.02916 = E

Provérbios 11,24

Há quem dê generosamente, e vê aumentar
suas riquezas,
outros retêm o que deveriam dar,
e caem na pobreza

INXS - Disappear.

Worm Hole → strait line.

Photos

1826 - Joseph Nicephore Niépce

Local: Saint-Loup-de-Varennes
France

+ Louis Daguerre (iodeto de pratas)

Alma A. Pelot - criou a fotografia estereoscópica

1861 Forte Sumpter tomado pelos confederados

↳ início da guerra civil americana

20210502

branch of math

```
  985   480
   75   180
-----   120
1.060  ----
  120   780
-----    65
1.180   ---
        845
         45
        ---
        890
         95
        ---
        985
```

Geometria metrica 2 Vol
Pedro Puig Adam

—

dokumen.tips

—

njctl.org.

—

Maria Goeppert Mayer
Nobel 1963
"números mágicos"
– núcleo atômico

2, 8, 20, 28, 50, 82 e 126

(núcleos mais estáveis)

Desenvolveu o famoso modelo nuclear de camadas.

→ antes disso usava-se o modelo de Niels Bohr.

—

Stancho

20210502

Data-NowVivo-010-2021\
20210502\
Trends in Differential Geometry,
Complex Analysis and
Mathematical Physics\
Trends in Differential Geometry
....pdf
Pág. 109 (126)

Some Geometric Properties
and Objects Related to
Bézier Curves
(Hristov)

Barycentric coordinates and
barycentric trajectories
⟶ real projective plane
$\mathbb{RP}_2$.

Pág. 111 (128)
1.2 Barycentric characterization
of the duality in the projective plane.
Projective isotomic conjugation

Import

Pág. 112(129)

There are several kinds of conjugate maps (isogonal, isotomic) with respect to fixed triangle, defined as colineations over E_2^* for example by means of point multiplication [6].

The theorems of Ceva and Menelaus give rise.

Definition 1.2.

Projective ($\mathbb{R}P_2$) – isotomic conjugation with respect to basic
P_2
ΔABC is the map

Pág. 120(137)

—

Canonical commutation relation
↳ canonical conjugate quantities

Conjugate variables are pairs of variables mathematically defined in such a way that they become Fourier transform duals, or more generally are related throught Pontryagin duality. The duality relations lead naturally to an uncertainty relation (Heisenberg uncertain).

Source: Pag. 120 (137)

(Heisenberg relations in the general case.)

=

20210503m

(Heisenberg uncertainty principle

$$\Delta x \Delta p \geq \frac{h}{4\pi}$$

=

Planck's ~~[illegible]~~ constant.

Conjugate variables — are pairs of variables mathematically defined in such a way that they become Fourier transform dual or more generally are related through Pontryagin duality

(Heisenberg uncertainty principle)

Look: Noether's theorem

Which states that if the laws of physics are invariant with respect to a change in one of the conjugate variable,

↓ Pappus theorem ↗ conjugate points

(chacar)

↓

then the other conjugate variable will not change with time
(it will be conserved)

Source:

canonical commutation relation

↪ canonical conjugate

Lei de Leavitt → relação
período - luminosidade.

qto mais brilhante a magnitude, mais longo o período.

Em 2008 → membros da comunidade astronômica votaram pela mudança do nome do que antes era chamado de relação período-luminosidade para Lei de Leavitt.

---

vontade de Deus
+
liberdade do homem.

O homem morre pelo mau uso da liberdade do homem.

A maioria das pessoas morrem pelo mau uso da liberdade do homem e não por vontade de Deus, mas quando chega a hora é a vontade de Deus?

20210504, (20210504)

Moisés 1391-1271 AC.

Melech Mashiach (o Rei Messias)

—

Teorema de Gödel e a Hipótese do Contínuo

Fundação Calouste Gulbenkian

R$250,00 ISBN 978-9723112849

Source:

Data-NonVivo-010-2021\Books-010-2021\Importan\
Manuel Lourenço - O Teorema de Gödel....pdf

=

20210505

Ano: 152, 162
136

Gravity

Weak force

Electromagnetic Force

Strong Nuclear Force

https://astronomie-rara.
ethbib.ethz.ch

Conics and Cubics Bix

---

e-rara.ch

---

wolframscience.com/nks/

---

wolfram.com/wolfram-u/
an-elementary-introduction-
to-the-wolfram-language/

---

wolframcloud.com

---

wolframscience.com/nks/
notes-9-15--sphere-volumes/

---

d- dimensional Euclidean Space is

$s[d] r^d$ where

$s[d] = \pi^{d/2}/(d/2)!$ (the surface area is $d\, s[d] r^{d-1}$).

(The function $s[d]$ has a maximum around $d = 5.26$, then decreases rapidly with d.)

—

$$2\pi a^2(1-\cos[r/a]) =$$
$$\pi r^2(1 - r^2/(12a^2) + r^4/(360a^4) - \dots)$$

—

emeraldcloudlab.com

20210506 M

575	240
1180	175
1755	415
1US\$ = 5,20	160
US\$ = 337,50	575

personalpages.to.infn.it/~musich/

—

sphera → elipsoide (foge "gravidade")

Anãs Brancas

lóbulo de Roche
→ a área na qual
sua força gravitacional
é maior que a gravidade
gerada de pela companheira.
assim nenhuma delas tem muita
influência sobre a outra.

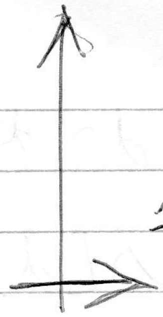

1755
 150
1905

—

Wolfram

An Elementary Introduction
to the Wolfram Language

Source:

Books-010-2021\Wolfram\
Stephen Wolfram - An Elementary
to the Wolfram Language.pdf
Pág. 119

Generate the completely connected
graph with 6 nodes

```
Graph[Flatten[Table[i->j,{i,6},{j,6}]]]
```

20210508

Graphics [{Circle []}]

Graphics [Circle []]

Elementary Particle Theory

Eugene Stefanovich

3 volumes

Stupid Girl Garbage

OH My my Blue October

Need you tonight

INXS Original Sin

Disap'pear

Never Tear us apart

sanet.st

sanet.lc

team.gdrive.wp

teamdrive.mfoxx.workers.dev

https://cuts-url.com/Google D1
D2
D3

gd.404edu.workers.dev

Melech Hamashiach

Martin Davis (93)
martin@eipye.com
https://cs.nyu.edu/cs/faculty/davism/

The Weeknd Ariana Grande

1,34 D 20210509

1,54 G

1,66 GPlus Caroline Ferraz (46)

Pág. 149 ZIP-03-20210509

20210510pa
Schrodinger

Source:

Data-NovoViro-010-2021\
Books-010-2021\
20210509\
Aide-Memoire de Mathématiques
Supérieures\
→ Des-fleurs-pour-Schr-dinger...pdf

Pag. 28(33)

Figure 1.1 →

Transformation de l'équation fondamentale de la dynamique en une équation de forme schrödingerienne sous de conditions de fractalité).

=

→ \Exercices-corrig-s-sur-la-
Convexit.pdf\

Pag. 27(28) 1.2.2
Fonctions eulériennes

calculer l'intégrale

=

28/Set/1918

The World Wars: A British Soldier Spared Hitler's Life
(Canal Youtube: History)

Henry Tandey

—

enojado

—

Geometry Yves Ladegaillerie
ISBN 2-7298-1416-7

—

The One I love - REM.

Skin and Bones
Cage the Elephant

Twilight
Torn
Mirrors

—

mathematics.page

mathematics.group (70)

mathematics.cloud (70)

music.company (50)

20210513M

I hope you're happy
Blue October

—

The Bandit Kings of Leon

—

The Killers My Own Soul's Warning
(Michael Hill video)

20200514M

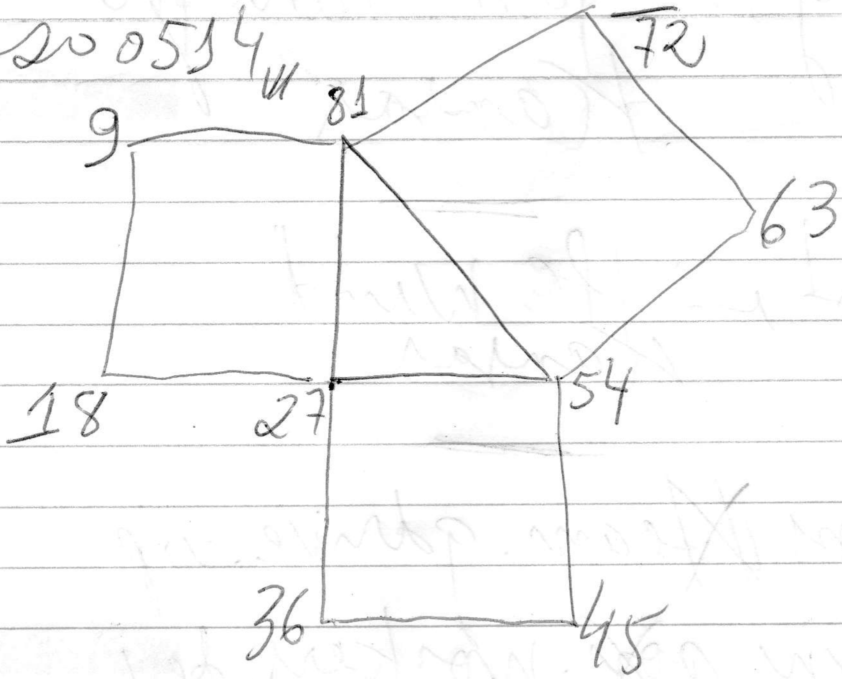

Source:
Books-1-10-2021\
20210514\
Faith Javane\
Javane Faith -
La Clave....pdf
(capa)

—

20210515

veracrypt.fr (free)

R52 | 3,78

13,756613 x 24 = 330,158730

—

US\$ $10 \div 3{,}78 = 2{,}\overline{645502}$

$\times 5{,}25 \quad 13{,}\overline{8}$

$Gal = 3{,}785\overline{411}$

⇒ rawdrive.com/plans/

Play the Game Tonight
Kansas

Just in the Wind
Kansas

https://team.gdrive.vip
www.p~~on~~.workers.dev

ucla.wr.workers.dev

Cor Azul → radiação Tcherenkov

20210515

Eletromagnetismo (responsável pela luz, eletricidade, ímãs)

A força nuclear forte (responsável por manter tudo unido)

A força nuclear fraca (responsável pela radiatividade)

Para cada força existe uma ~~fo~~ partícula intermediadora.

Fótons são as partículas intermediadoras das forças eletromagnéticas.

gluons → força forte

Bósons pesados → força fraca.

Blocos de construção da matéria são formados pelas duas categorias:
bósons
férmions

Toda a matéria → 3 férmions:
2 Quarks
1 elétron.

2 tipos de Quarks no modelo padrão →

mas apenas 2 deles – os mais leves – formam prótons e nêutrons

Off: Quarks beleza (decaimento) geram múons ou elétrons (produzindo ambos em quantidades iguais).

Mas experimentos começam a provar algo diferente.

Não se sabe se é por causa de:

a/ uma nova partícula ou

b/ nova força fundamental.

Quarks beleza < múons

> elétrons

↑

decaiam com menos ou mais frequência.

20210520

Differential Equations
Hubbard

---

mathematica cloud (R$ 79,00)
mathematica contact (R$ 50,00)

0800 724 8154

Instalações Elétricas
DPS 15/275V.
DR (Dispositivo R). 40 A.

Wolfram

RealDigits[Pi, 10, 20]

DSolve
↳ Sturm- equations

spherical embedding (50 min)

GeometricTest.

Geometry (59 min)

Time 1h42min
and Data ≠ containers

—

Sidereal Time } 1h45min
Solar Time }

Map:
Geo Orientation Data

—

Molecular 2h

Functions 2h24min

—

20210521

→ 9h55min } 60 Gb
14h18min }

MAIO:
410
60
470 Gb

≅ 15 Gb/h

Atalhos WINDOWS:

Alt + D →

470
58
528 Gb
35
563 Gb.

—

—

team.gdrive.vip
googlesharedrive

—

Atalhos:

WIN + Ctrl + D → nova área de trabalho

Ctrl + F4 → fechar área

~~Ctr~~

WIN + M → minimizar área de (janela)

⇒ WIN + D → ver desktop

WIN + , →

Ctrl + Shift + Esc → Gerenciador de Tarefas

⇒ Ctrl + Shift + T → abre a aba de navegação que você fechou

⇒ Alt + P → visualizar arquivo sem abri-lo.

⇒ WIN + P → ampliar tela. (importante)

WIN + + = ampliar área ?

Ctrl + move c/ mouse o arquivo p/ a outra pasta (cópia)

⟹ o mesmo.

WIN + A → config WIN

I → config do PC

20210522 (20210522)

Matusalém - 969 anos.

14,46 filhos de anos. ± 0,8 filhos

=

13,77. 0,69 0987654321

654

5

wolframphysics.org 4

baixardoc.com

vdocuments.site

documents.pub

ARQ-BASICS-001 (01-04)

ARQ-BASICS-002 (04-02-10)

20210523

614 Gb

eponymous.

eponymic

AHA CRY Wolf.

Books-011-2021\20210523\

Quantum Topology Kauffman

Mathematics of quantum computation

Pag. 203 (222) Proof.

Pag 225 (244) Photon as a qubit

optics quantic computer

polarization rotation

↑

problem.

Unified Field Mechanics

Pag. 26 (49) –

9.1 The Square Root of Minus One revisited.

Avcero.com.br 6,14 Gb

Banks\04_02\Geometry-Gauss\ 46

660

neologismo Stagflation → 1965
Iain Macleod
(deputado)

Inflação e Desemprego → Curva de Phillips.

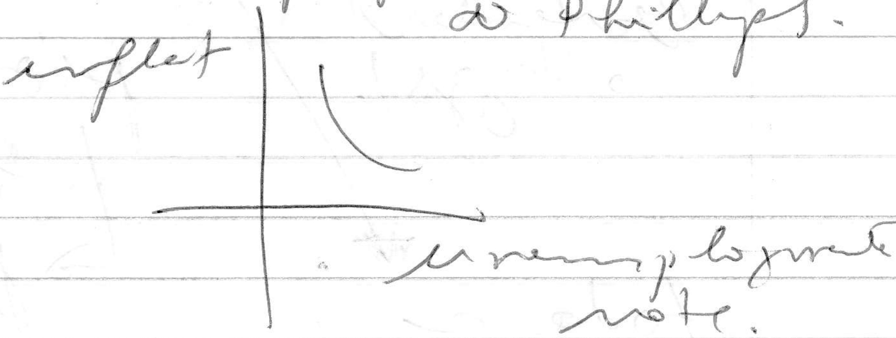

Postura

hawkish
dovish

"Barber Boom"
England

---

5 Forças fundamentais: eletromagnetismo
gravidade
força nuclear forte
força nuclear fraca
força de Higgs.

1 Higgs pesa ≅ átomo de iodo
↳ não tem spin.

As outras partículas ~~Não~~ têm spin mas tem um giro intrínseco.

obs: Higgs não tem giro ou spin.

Higgs → composição (vários partículas fund)
ou
supersimetria

(solução para entender o Higgs).

píon → partícula sem spin
→ prótons
nêutrons

toda a partícula tem um par.

—

Universo tem dois tipos de partículas
{ matéria → { quarks → P N
{ léptons
{ força → { elétron
{ neutrino

—

Quark decai em proporções diferentes
Para cada 100 decaimentos com elétrons houve 85 decaimentos com múons.
(
→ Quebra da simetria cósmica

eletrodinâmica quântica

Campo magnético intrínseco é medido em unidades típicas para partículas elementares.

$(g)\_\{\text{experimental}\} = 2.00115965218073$

$(g)\_\{\text{teórica}\} = 2.00115965218178$

Efeito Schwinger

Um dos motivos é a analogia do eletromagnetismo com a gravidade.

Na vizinhança de um buraco negro pares de partículas são geradas para neutralizar o forte campo gravitacional do buraco negro

↳ "efeito Hawking"

Mama I'm coming home
Ozzy Osbourne

$$\begin{array}{r} 660 \\ +\ 43 \\ \hline 703 \end{array}$$

---

pdfcoffee.com

Curso de Cálculo
Diferencial e Integral
4 Vol.
R$ 169,00

Willie Alfredo Maurer

V1 - Fundamentos Geométricos
e Físicos - 368 pp

V2 - Fundamentos Aritméticos
e Topológicos - 375 pp

V3 - Funções de Várias Variáveis
375 pp

V4 - Equações Diferenciais
- 258 pp

Obs: sebodomessias.com.br

20210524

Matrix Theory

Gantmacher R$ 200,00

2 volumes.

^{1}268 70
270 200
538 270

www.maths.ed.ac.uk/~v1ranick/papers/gantmacher1.pdf

—

idoc.pub

—

njwildberger

—

seancarol

R$538
180
R$718
85
R$803

703 Gb
36
739
50
789

—

fisicaModerna UFF

—

impabr

mathemamovies

Curso de
Cálculo Diferencial e Integral –
4 Volumes – Willie Alfredo Maurer
(Alfredo)

20210528M

Francisco Gomes Teixeira

Traité des courbes spéciales remarquables planes et gauches

1971 3 volumes. R$ 600,00

—

marantholinistis.

(maranthose + violinista)

=

library.lol

—

mathcurve.com

2dcurves.com

—

pdf

Vol 1 → 403 (424)

Vol 2 → 488 (514)

Vol 3 → 412

=

Traité des Ponts Chaix 1890
www.sedhc.es/bibliotecaD/
1890_Chaix_Traite_ponts_1_
parte_T_I.pdf

—

NUMDAM.ORG

—

Eugène Rouché

—

University of Toronto
QA 3 G65 1904 v.5
obras sobre Mathematica
Teixeira

—

Harvard University
University of Michigan
↳ lib.umich.edu

https://t.co/mAd19esNmZ?amp=1

(= www.skmiomusich.com)

---

https://serge.mehl.free.fr/

University of Wisconsin

---

causará especie

7:53

20210526

Fermagonic Triangles

Pedro A. Piza

Sierpinski Triangle

---

mathematics.cloud (70)

ArrayPlot[CellularAutomaton
[254, {{1}, 0}, 40], Mesh->True]
↓
90

Joe Schmoe

~~@cOokiemon5ter~~

```
 1
 578
  65
 643
  95
 738
 180
 918
  80
 998
```

```
 1
 178
 285
 156
 313
  65
 378
 200
 578
```

Books:

gallica.bnf.fr/ark:/12148/
bpt6k64545935

→ Exercices de géométrie

bpt6k6336853h

Cours de mathématiques élémentaires
Exercices de Géométrie

/bpt6k935782m
Géométrie Analytique

=

/bpt6k9376884

—

/bpt6k4226025h
Éléments de Géométrie
F. J. J.

=

$\sqrt{2}\ 360^{\circ}$ = pi radian.

—

Transient

—

$(360\sqrt{2})/((332e!)/39 + 23/3 -$
$3907/(52e) + (8365e)/156) = \pm\pi.$

—

Toom's rule

20210527 VIVO (Velocidade)

onlinebooks.library.upenn.edu/

tutorial.math.lamar.edu/
Classes/CalcI/CalcI.aspx
II.aspx
III.aspx

ads.caltech.edu/~marsden/
volume/Calculus/

Esercizi di Analisi Infinitesimale

Nuovi Esercizi di Analisi Infinitesimale

Giulio Vivanti R$ 350,00

1916 520 pags

(05)
789 Gb
35
824 Gb

10315.

I.P. 177.215.71.240

Alterei o plano de 70Mbs p 200Mbs

R$95 → R$99

Gorillaz - Stylo
Feel Good inc
Clint Eastwood
(e)
Dare

—

VIVO Nº
29051634

824 Gb R$ 65
151 500
839 Gb R$ 565
57 100
896 Gb R$ 665
47
943 Gb

20210528//

57.97.0.226.

Differential Geometry
Heinrich W. Guggenheimer
378 pag
tapered

Zoinho. —

943 Gb | R$ 665
66 | 90
1009 Gb | R$ 755

✡ 83.242.227.35-bc.
googleusercontent.com

claro.com.br/claro-drive
1TB R$24,99
⟹
restream.io/

1009 Gb
31
1040 Gb
165 (sv)
1.205 Gb

800
1009
755
2564 R$

twitch.tv/

Harrie Hausenman (Pirate?!?)

Shoft.

breaking off →
rompendo

163.41.117.34.bc.
googleusercontent.com
=

1205 Gb
68
1.273 Gb
53 ← fazer upload
1.326 Gb

A Pirâmide Perdida,
20210529/
65.239.233.35.bc.googleusercontent.com
↗
IP

1.326 Gb
96
1.422 Gb

math-ccsf

City College of
San Francisco

=

Rede

1422 Gb

wD69RG1zU2 (4 núcleos)

2KQ3B1UA7P (2 núcleos)

—

beacons.gcp.gvt2.com

1.422 Gb
78
1.500 Gb
91
1.591 Gb
58
1.649 Gb

142.0.76.34.bc.
googleusercontent.com

—

20210530//

"O amor é como uma flor roxa
que nasce no coração do trouxa."
novela

—

Fazer upload PC-O-ALL-20210530.

Fazer download NOVOVIVO-10-2021-20210513.

Pág 423
268
281
972 ok
=

R$ 100
70
170
35
205

205
90
295
140
435

1649 Gb

20210601 ///

curvas elípticas

espaço hiperbólico.

Taniyama e Shimura. (1955)

→ Problemas de Taniyama levaram à extraordinária afirmação de que

qualquer curva elíptica era realmente uma forma modular disfarçada.

→ conhecida como

Conjectura de Taniyama—Shimura

as curvas elípticas racional é modular.

mas ainda não havia sido provada.

1985, Gerhard Frey teve uma ideia surpreendente.

→ e se Fermat estivesse errado?

$x^n + y^n = z^n$

→ se Fermat é falso, Shimura-Taniyama também o é.

→ ele não provou, mas deu um argumento de plausibilidade.

Frey → conjectura épsilon.

Barry Mazur deu uma sugestão para resolver
" Provar completamente a conjectura de épsilon de Serre".
(20m23seg)

→ Barry Mazur sugeriu para Prof. Ken Ribet

John Coates supervisor Andrew Wiles.

Utaka Taniyama 1958 died

O truque de Andrew era transformar as curvas elípticas em algo chamado Representação de Galois,

Iwasawa theory → Fórmula do Número de Classes.

(26m23seg)

Class Number Formula

Coates falou para Wiles → (91)
de um paper de Matthias Flach
↳ conseguiu uma
fórmula Class Number

Flach usando ideias de
Kolyvagin. (27m05seg)

Jan 1993 ⇒ Andrew conversou com
Nick Katz.

33m28seg –

Wednesday June 23 1993
Elliptic Curves and
Galois Representations

Começo da Prova:
Erro sobre o Método de Flach
e Kolyvagin
(37m21seg)
Corrigido Richard Taylor

Revelation → 41 min 14 seg.

É o mesmo problema que tinha lido na minha Teoria de Iwasawa, três anos antes.

=

Math envolvidos:

Klein, Fricke, Hurwitz,

Hecke, Dirichlet, Dedekind,

Prova de Langland e Tunnell

Deligne, Rapoport, Katz

A Idéia de Mazur de usar a teoria de deformação das representações de Galois

Igusa, Eichler, Shimura,

Taniyama,

A redução de Frey

Bloch, Kato, Selmer, Frey,

Fermat.

Kepler odiava dar aulas
e uma das suas paixões era
a Geometria, que lhe dizia existir
antes da Criação.

Kepler foi ser assistente de
Tycho Brahe (Praga)

3ª Lei Kepler:

"Os quadrados dos períodos
de translação dos planetas são
proporcionais aos cubos dos
semi-eixos maiores de suas
órbitas".

Kepler quase descobriu a Lei da
Gravidade, mas manteve a
força com o nome de
"magnetismo".

Kepler - 1571 - 1630
Newton - 1643 - 1727

Kepler na ótica determinou que a
intensidade de uma onda de luz diminui
com o quadrado da distância.
Também na neurofisiologia e foi
o primeiro a descobrir que as imagens
no olho humano são projetadas na
retina de cabeça para baixo.

20210608M Copy

20210609M 3 - Aberto.

200
180
380
50
430

Geometrie des Universums
Robert Osserman R$ 100,78
430
R$ 10,508
130
100
R$ 10,738

122,49 TB

Casa do Drive.

fotopletismografia.

mdj-epihanie.me

—

1 - Antigo → ~~CASADODRIVE~~ COCHISE
1 - Novo
1 - Website-Blog → matte1matics2@gmail.com
1 - Reserva → CCSF
2 - GERAL → ~~COCHISE~~ DRIVE → 122,45 - 100,51 =
3 - Aberto → COCHISE [redacted] ≈ 22TB.

—

20210610:
Co

Copiar Data NovoVivo → mover p/
Geral, Reserva

20210610 Blog-Pastas com files

Matte1matics2@gmail.com

1-Website-Blog/Leibniz/ (19w)

2020/10/20201012/Leibniz-Werke-04-Volumes/

(16Qn)

20201031/Die Werke von Leibniz-Hannover

2020/11/Correspondance-de-Leibniz

(1m0)

- 11/Leibniz-09-Volumes-Hannover (1V7w)
- 11/Leibniz-Google-16-Volumes (183B)

11/Leibniz-und-Landgraf (1nV0)

Obs:

~~2020/10/~~

11/20201104-Leibniz-gwlb.de

(1Sg_z14)

2021/01/Leibniz-Die-philosophischen (1vaS)

01/Leibniz-Oeuvres-7-Volumes (1N9d)

01/Leibniz-Opera-Omnia-Latin (15-BL)

01/Leibniz-Sämtliche ... 1923 (1ggB)

✓ 2021/02/ Gauss - Werke 12 - Volumes
(1Dy33)

—

20200614 M

Nuvem Passageira - Hermes Aquino - 1976

Porto Solidão - Jessé

Voo Liberdade - Jessé

Biafra - Sonho de Ícaro

=

Ritchie - Menina Veneno - 1983

-

history-of-mathematics.org

-

⇒ Squaring the Apsamikkan Circle

Overdeck Family Foundation
(overdeck)

momath.org
wolfram.com
overdeck.org

-

Chapter 1 - Part 3

Notebook-03---20211112

20210615

Registro www.mathematics.pictures
(por 3 anos)

—

www.steniomusick.com
www.matemática.org
www.mathematics.pictures

—

Bedtime Math Foundation, Inc
New York - US
godaddy.com (overdeck.org)

—

Geometrias / Howard Eves
2 volumes - 471 + 483 pages

—

20210616

Geometry - Arthur F Coxford. 1991.

—

booksc.org

—

Quem tem vai ficar sem
Quem não tem vai ficar sem.
Stenio

Quem tem
vai ficar sem
quem não tem
vai ficar sem.
Stenio.

data
S T Q Q S S D

xn-matemtica-41a.org

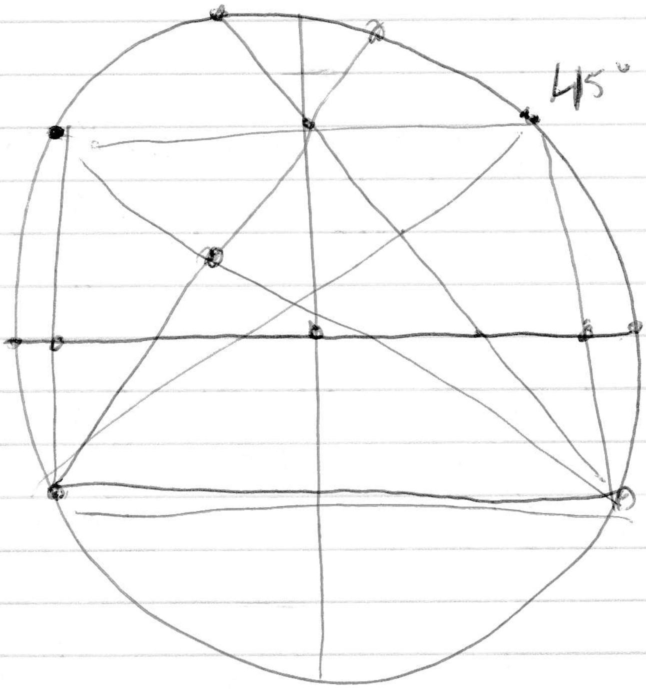

Pappus

Pappus La Collection Mathématique
Paul Ver Eecke

Fundamentals of Geometry Coxeter
(Fundamentos)

Fundamentos de Matemática
Geometrias Analítica e Vetorial
Euclidianas e Não-Euclidiana
2 set 2008
Ana Chiummo
ISBN 978-852161635-1 316 pag

20210621III

idoc.pub / cupdf.com

Books-013-2021/Important/
François Apéry - Models of the
Real Projective Plane ... pdf

Pág. 8 (VII)
→ image ✡

Pág. 18 (30)
Fig. 8
Fig. 9 Stereographic projection
from the sphere to
the plane.

Mathematicorum

Mosnat.

ictp-saifr.org/
ggem2021/

Quantum Gravity
Rovelli

Aritmética 26
Álgebra 246
Geometria 658
Trigon.
Geometria
Analítica 914
Calc. Dif 1132
Calc. Int 1336

R$ 2990
180
210
R$ 3380
100
450
R$ 3930
360
R$ 4290
130
R$ 4.420

Source:

Books-013-2021/
20210621/
Mathematics_
Books/
matesimp2.
pdf

Zracna Luka Mali Losinj

The Treasury of Mathematics
Henrietta O. Midonick

njwildberger.com/2015/12/02/the-
alexander-horned-sphere-is-it-
nonsense/

20210~~0~~703 (20210703)//

gauntlet = manopla.

A Cornucopia of Quadrilaterals - MAA Press - Vol 45

The midpoints of the sides of any quadrilateral are the vertices of ~~a~~ parallelogram, or that in a convex quadrilateral (not a parallelogram) the line through the mid points of the diagonals (the Newton line) is equidistant from opposite vertices, or that, if your quadrilateral has an inscribed circle its center lies (encontra-se) on the Newton line.

There are results dating back to Euclid: the side-lengths of a pentagon, a hexagon, and a decagon inscribed in a circle can be assembled into a right triangle (the proof uses a quadrilateral and circumscribing circle); and results dating to Erdős: from any point in a triangle the sum of the distances to the vertices is at least twice as large as the sum of the distances to the sides.

Book: 20210703/ Oeuvres Tchebychef - 2 volumes/
Oeuvres_de_P_L_Tchebychef_..
417.pdf

Pág. 129 (167) - Sur les engrenages
Pág. 165 (203) - Sur les quadratures
Obs: Important.

Source: DATA-NovoVivo-014-2021/
Books-014-2021/

Other: Posta:
/A Cornucopia of Quadrilaterals

=

/Wildberger/ Divine Proportions...
pdf

=

/Oeuvres Tchebychef - 2 Volumes/
Obs: 2 volumes - Oeuvres
Tchebychef

4 volumes - Oeuvres de
Charles Hermite.

Imp: No capitulo Sur les Quadratures
fala da ~~Ex~~ aproximação de M. Hermite (valeur
de l'intégrale).

=

20210704

Sydney NSW 2052 Australia
Kensington Nova Gales do Sul
unsw.edu.au

R$ 280
150
R$ 430

---

jewishgen.org

---

temperaturas ~~com~~ caniculares

20210705

sydney.edu.au

---

math-ccsf
math-cochise
mathematics (ccsf
mathematics-cochise
mathematics-south (SW college)
z-case-swccd (SW college)
z-clan

---

https://td.hackgence.com
team.hackgence.com
pon.workers.dev

---

instados a usarem.

Tertúlia – reunião familiar
supedâneo

—

M. Chasles – Géométrie

—

Exercices de Géométrie
(gallica) F. J.

—

MATH:
⇒ Books-014-2021/20210710/
Plans-important/
Exercices-de-géométrie-
6pt...35.pdf
Pág. 12(44)
27. Théorème
Cercle des neuf points
(nove pontos)

—

The mind and the matter
twilight zone

—

VTex

2021 07 11 ///

netdrive.net

filezillapro.com

odrive.com

Website

mdj-epiphanie.me

David Bohm - filme
Infinite Potential
de Paul Howard.

Quantum Theory - Bohm

Livro

Infinite Potential -
The life and Times of
David Bohm
(David Peat)

Bohm → eu não sabia que
ele era comunista,
fiquei decepcionado!
que um homem inteligente como
ele se enveredasse por este caminho
O comunismo destruiu a terra de
origem do meu pai.

3456789/7 = 4,93270

3456/7 = 4,93$\overline{714285}$

345/6 = 5,75

—

Diego Rubiera-Garcia

drgarcia@fc.ul.pt

Bob.Palais@uvu.edu

palais@math.utah.edu

---

musich.art

—

cut-the-knot.org

—

20210715

Gambarogno Swiss

↳ Lago Maggiore

Calculus in 3D Nitecki

—

STENIOMUSICH.COM

data

S T Q Q S S D

20210717///

Geometria Proiettiva Crani.

20210718///

Pappus Collection Mathématiques.

vatican library.va

livre21.com

—

20210719///

$1100 \div 5{,}2 = 211{,}\overline{538461}$

$1600 \div 5{,}2 = 307{,}\overline{692307}$

—

20210720///

wolfram.com/open-materials/

mathematica-journal.com

—

Daniel Carvalho (wolfram)
5511 9282 3991

—

wolfram.com/wolfram-u/

—

data

Ⓢ Ⓣ Ⓠ Ⓠ Ⓢ Ⓢ Ⓓ

20210721

libgen.st

Archive.org

Lezione di geometria differenziale

Bianchi L.

3 vol.

teoria do espaço-tempo modular

A gravidade deixa de ser uma força de atração entre dois corpos e passa a ser uma deformação no espaço-tempo causada pela matéria, de acordo com a sua massa.

Gravidade Quântica de Loop (LQD), que sugere um espaço-tempo feito de uma série de loops entrelaçados em escalas mínimas de tamanho.

loops apenas perceptíveis na escala de Planck

espaço-tempo modular, que define o espaço-tempo pela maneira como os objetos interagem.

Assim, o espaço-tempo se torna uma consequência do próprio mundo quântico, não algo que precisa ser combinado com ele.

Source (Freidel, Leigh e Minic)
space.com

20210722 ///

wolframcloud.com

Recognize Handwritten Digits

digit = Classify[{/→1, 2→2,
....}]

digit[8, "TopProbabilities"]

↳ o resultado mostra a frequência na qual o fenômeno se apresenta.

---

wolfram.com/open-materials/

→ Sandbox

Visual Complex Analysis.pdf

Pag. 216 (242)

the Cartesian Form

Pag. 222 (248)

Visual Differentiation of log(z)

Pag. ~~25~~ 224 (250)

Inverse Functions

→ twist (torção) of f

MHHMY71.pdf

Pag. 30 (56)

38 - the Circle

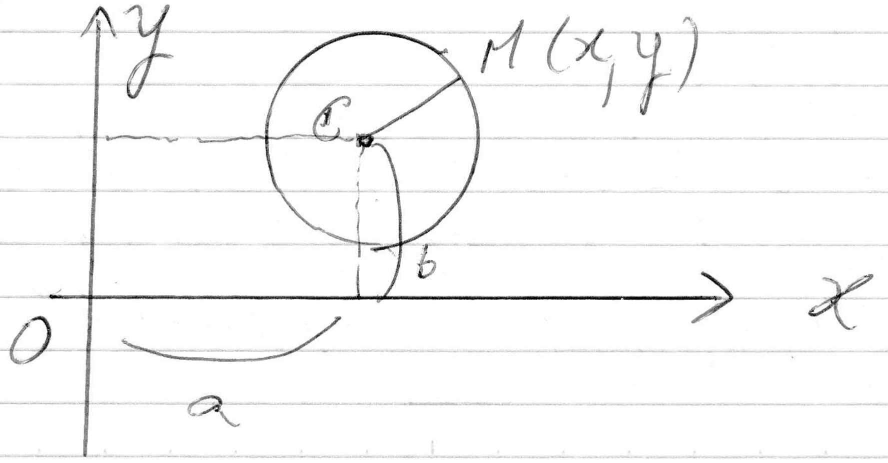

A circle of radius R with centre at the point C(a, b) is described by the equation

$$(x-a)^2+(y-b)^2=R^2$$

rewritten

$$x^2+y^2-2ax-2by+a^2+b^2-R^2=0$$

(20210725)

Mathematics Modern Calculus and Analytic Geometry. djvu (2012)

Pag. 527 (545)

Fig. 9.2

while $x^2+y^2=a^2$, with parametric representation

$$x=a\cos t,\quad y=a\sin t,\quad (0\leq t\leq 2\pi)$$

or

(cf Pag. 305)

There is nothing unique about the parametric representation of a curve

Ex:

$x = a \sin t,\ y = a \cos t \ (0 \le t < 2\pi)$
(2π)

or even

$x = a \cos \ln t,\ y = a \sin \ln t$
$(1 \le t \le e^{2\pi})$

represent the same circle

$$x^2 + y^2 = a^2$$

20210726

reproduzir vídeos ao contrário.

Snow Patrol Open your Eyes
inspirado no
filme "Era um encontro"
(C'était un rendez-vous)
1976 por Claude Lelouche

---

www.mackenzie.br/gerti/
estudantes-e-funcionarios/softwares/
wolfram-mathematica.

Um buraco negro é menor que o raio de Schwarzschild.

MIT – supercondutor chamado grafeno de três camadas torcidas por ângulo mágico.

→ tripleto de spin

↓

a energia de ambos os elétrons muda na mesma direção, mantendo a supercondutividade independente da intensidade do campo magnético ao qual estão submetidos.

3 camadas de grafeno e a camada do meio foi girada em 1,56 graus em relação às camadas externas.

ângulo gauss!

↳ curvatura

3 camadas torcidas poderá ser aplicado no desenvolvimento de computadores quânticos mais robustos.

Supercondutividade topológica
→ ingrediente chave para fazer isso seriam supercondutores de tripleto de spin.

O grafeno poderá ser usado para modular os sinais de telecomunicação em velocidades ultrarrápidas podendo ser utilizados para quaisquer frequências de comunicações desejadas

Grafeno ferroelétrico condutor
→ servirá para armazenar dados em memórias super rápidas para computadores

Grafeno →
uma das formas cristalinas do carbono,
organizado em uma forma camadas de átomos de carbono dispostos em hexágonos.

NB – Nitreto de Boro →

(dois)
Camada do meio entre 2 grafenos.
↳ ângulo ligeiramente diferente entre si.

→ Padrão moiré
(moiré)

→ padrão formado quando dois grafos apresentam movimento relativo entre si.

{ Padrão moiré
↳ compensa a curvatura espacial de Gauss?

→ important!!!

⇓

Primeiro ferroelétrico condutor,

"é a primeira demonstração que relata ferroeletricidade eletrônica pura, que exibe polarização de carga sem movimento iônico na rede subjacente". Philip Kim (Univ. Harvard).

https://corsi.unibo.it/magistrale/matematica

Coordinatore: Marilena Barnabei

https://www.ifsc.usp.br/~strontium/Publication/Scripts/QuantumMechanicsScript.pdf

---

studenti.unibo.it

20210727

Angstrom = 0,1 nm

10^{-10} m   10.000.000.000
10^{-8} cm

1 Å = 100 picômetros.

1 Å = usado para medir o espectro eletromagnético
(electromagnetic spectrum)

Albert A. Michelson defined so that the red line of cadmium was equal to 6438.47 Å.

1907 → 15°C dry air (hydrogen scale)
760 mm Hg
9.8067 m/s² (gravity)

$= 1\ \mathring{A} = 6438.4696$

angstrom remained a secondary unit of length for use in spectroscopy.

$7.2^{\wedge 2} = 50$ $\quad 7.2^{\wedge 7.2} = 5{,}184$

$7.1^{\wedge 2} = 50{,}41$ $\quad 7.2^{\wedge 2} = 51{,}84$

$\sqrt{50} = 7{,}071\ldots$

$7{,}07^{\wedge 2} = 49{,}9849$

$\downarrow$ $\quad \sqrt{0{,}0151} = 0{,}12288\ldots$

$7{,}071$

$0{,}001 \rightarrow \sqrt{0{,}001} = 0{,}03162277\ldots$

$\sqrt{100} = 10$

$\sqrt{50} =$ $\quad \rightarrow 7{,}5$

$\sqrt{25} = 5$

$7{,}5^{\wedge 2} = 56{,}25$

> erro!!!

$100 - 25 = 75 \quad \sqrt{75} = 8{,}66025\ldots$

vejamos:

$\sqrt{100} = 10$

$\sqrt{75} = 8,66025\ldots$

$\sqrt{50} = 7,07106\ldots$

$\sqrt{25} = 5$

$> 1,33974\ldots$

$> 1,589186\ldots$

$> 2,07\ldots$

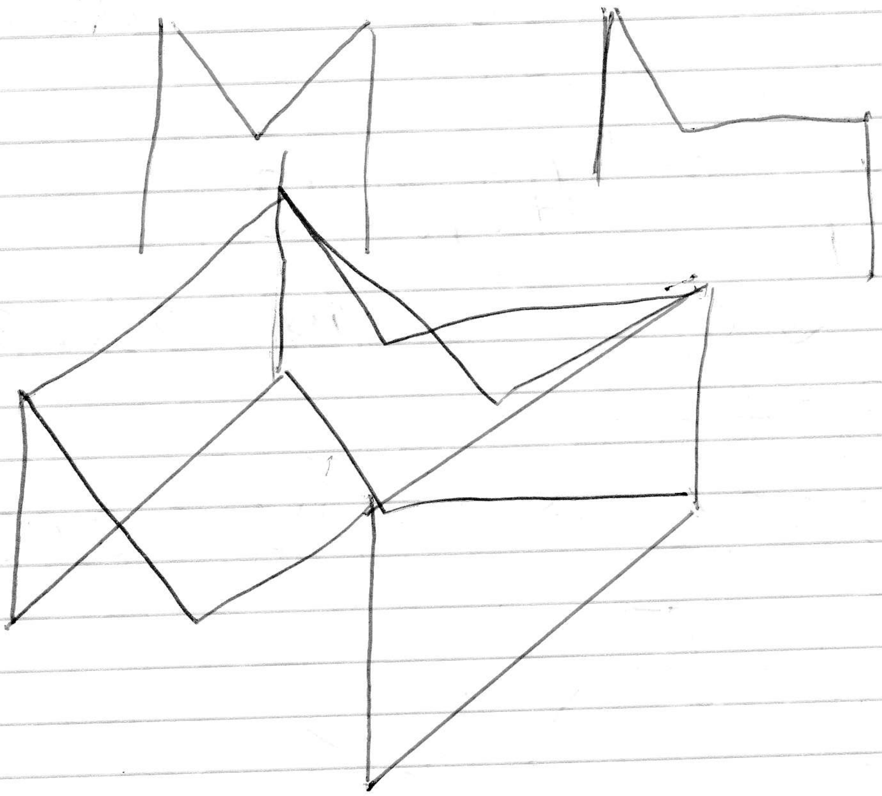

→ cubo

20210727

WILLIAM OUCHI
SOLLEDADE M.

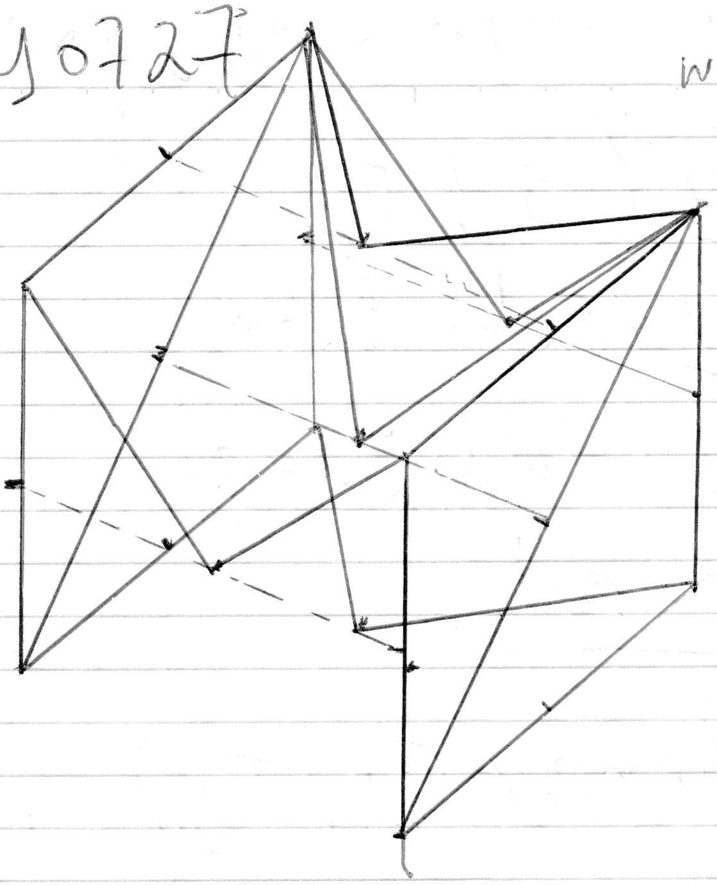

Line $-x+y=0$

Important:

r_α : Line(E_1, F_1)

$\rightarrow -4.2x + 4.27y = -5,87$

(F_1, E_1)

$\rightarrow 4.2x - 4.27y = 5.87$

Alinhamento de 3 pontos

$G_1 (3, 1.5)$

$H_1 (5.12, 3.62)$

$I_1 (7.24, 5.74)$

Important:

$g_2 = Line(G_1, I_1)$

$\rightarrow -4.24x + 4.24y = -6.36$

$$\left.\begin{array}{l} G_1H_1 = 3 \\ H_1I_1 = 3 \end{array}\right\} \text{distance}$$

Fazer 3 perpendiculares nos pontos G_1 H_1 I_1

criando novos pontos em cima destes termos:

20210728 III

$G_1 H_1 I_1 \Rightarrow J_1 \boxed{K_1} L_1$

↳ centro

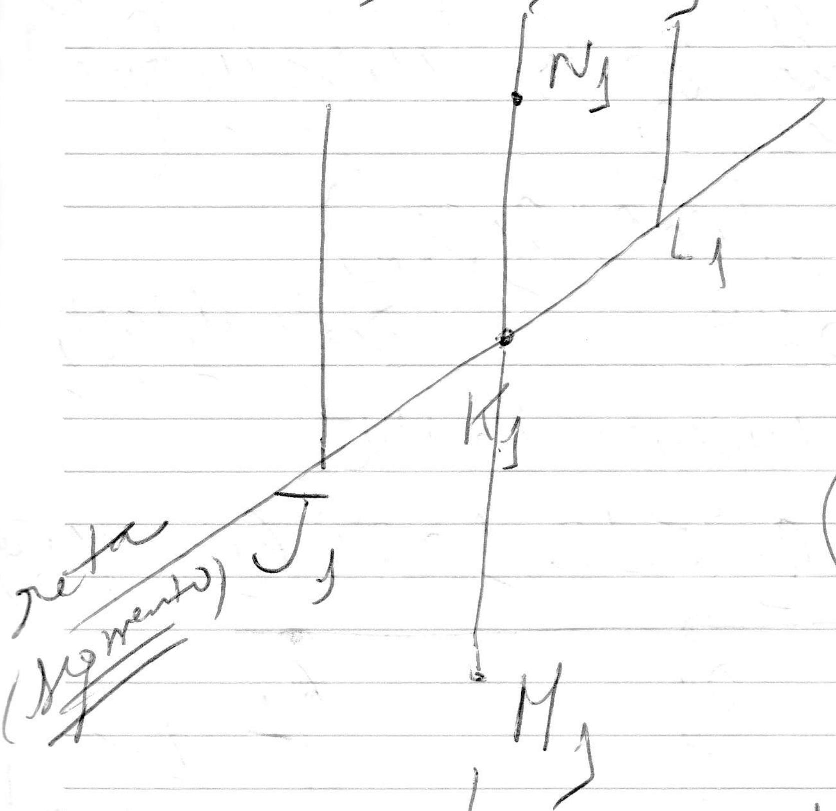

N_1
L_1
K_1
J_1
reta (segmento)
M_1

$\boxed{M_1 K_1 N_1}$

↳ 3 pontos

Portanto:

$J_1 K_1 L_1$ e $M_1 K_1 N_1$

→ Pappus Theorem

⇒ temos 2 retas

Important: Point O_1 (ápice)

O_1 = Intersect(g_1, n_1)

→ (4.24, 10.24).

Fazer uma linha perpendicular sobre a reta criada pelos pontos $M_1 K_1 N_1$ que passa pelo ponto O_1

PerpendicularLine(O_1, s_2)

↳ reta $M_1 K_1 N_1$

Y = 10.24

MUS: Parte superior do MUS

P_1 (3, 10.24) Q_1 (5.12, 10.24) R_1 (7.24, 10.24)

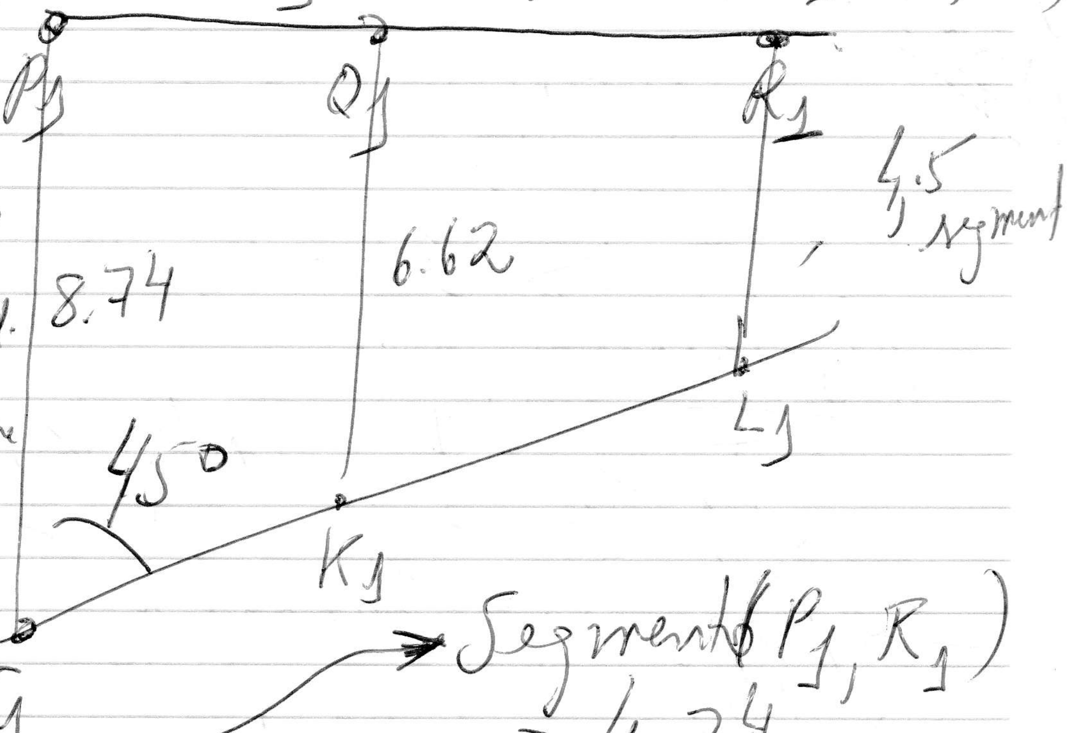

Comprimento reta $J_1 L_1$ = 6 und.

Diferença entre $J_1 P_1 - L_1 R_1$ =
8,74 − 4,5 =
$\boxed{4,24}$ =

→ Segment(P_1, R_1)
→ 4.24

Source:
ARQ_IMPORTANT_03/Geogebra-M01-20210727/
m01-20210727-13.ggb

20210728 III

Fazer este mesmo desenho com as arestas com 8 unidades.

Simular também em 3D (forma cubo)

Copiar os comandos dados para se criar este modelo de <u>MUS</u>.

ex: A= Intersect(xAxis, yAxis)
→ (0,0)

B = Point(yAxis)
→ (0,6) etc

<u>COMANDOS GEOGEBRA</u>

```
A= Intersect(xAxis, YAxis)
  →(0,0)

B= Point(yAxis) →(0,6)

C=(6,-3)

D=(6,3)

c= Circle(C, Segment(C,D)) →
  (x-6)^2 + (y+3)^2 = 36

d: Circle(A, Segment(A,B)) →
```

$x^2 + y^2 = 36$

E = (1, 1)

F = (7, −2)

G = (Point(Circle(A, 6)) → (6, 0)

f: Line(A, E) → $-x + y = 0$

H = Intersect(d, f, 2) → (4.24, 4.24)

g = Segment(A, H) → 6

h: Line(C, F) → $-x + y = -9$

I = Intersect(c, h, 2) → (10.24, 1.24)

i = Segment(C, I) → 6

j = Segment(A, B) → 6

k = Segment(C, D) → 6

J = Point(j) → (9, 3)

l = Segment(G, J) → 6.71

m = Perpendicular Bisector (G, J) → $6x - 3y = 13.5$

K = Intersect(m, l) → (3, 1.5)

n = Segment(K, B) → 5.41

p = Segment(K, D) → 3.35

q: Perpendicular Bisector (C, I) → $-4.24x - 4.24y = -30.73$

L = Intersect(g, i) → (8.12, −0.88)

r = Perpendicular Bisector(A, H) → $-4.24x - 4.24y = -18$

M = Intersect(r, g) → (2.12, 2.12)

s = Segment(M, L) → 6.71

t: Perpendicular Bisector (M, L) → $-6x + 3y = 28.86$

N = Intersect(t, s) → (5.12, 0.62)

a = Segment(I, N) → 5.16

b = Segment(N, H) → 3.73

$O = (1, 7)$

$P = (7, 4)$

e: Line$(B, O) \rightarrow$ $-x + y = 6$

f_1: Line$(D, P) \rightarrow$ $-x + y = -3$

c_1: Circle$(B$, Radius$(d))$ $\rightarrow x^2 + (y-6)^2 = 36$

$Q =$ Intersect$(c_1, e, 2) \rightarrow$ $(4.24, 10.24)$

$g_1 =$ Segment$(B, Q) \rightarrow 6$

d_1: Circle$(D$, Radius$(c))$ $\rightarrow (x-6)^2 + (y-3)^2 = 36$

$R =$ Intersect$(d_1, f_1, 2) \rightarrow$ $(10.24, 7.24)$

$h_1 =$ Segment$(D, R) \rightarrow$ 6

i_1: PerpendicularBisector $(B, Q) \rightarrow$ $-4.24x - 4.24y =$ -43.46

j_1: PerpendicularBisector $(D, R) \rightarrow -4.24x - 4.24y =$ -56.18

$S =$ Intersect$(e, i_1) \rightarrow$ $(2.12, 8.12)$

$T =$ Intersect$(f_1, j_1) \rightarrow$ $(8.12, 5.12)$

k_1: Line$(S, T) \rightarrow 3x + 6y = 55.09$

l_1: PerpendicularBisector$(S, T) \rightarrow$ $-6x + 3y = -10.86$

$U =$ Intersect$(k_1, l_1) \rightarrow (5.12, 6.62)$

$m_1 =$ Segment$(U, R) \rightarrow 5.16$

$n_1 =$ Segment$(U, Q) \rightarrow 3.73$

$p_1 =$ Segment$(H, Q) \rightarrow 6$

$q_1 =$ Segment$(I, R) \rightarrow 6$

r_1: Perpendicular Bisector$(I, R) \rightarrow$ $y = 4.24$

s_1: Perpendicular Bisector$(H, Q) \rightarrow$ $y = 7.24$

$V =$ Intersect$(r_1, q_1) \rightarrow$ $(10.24, 4.24)$

$W =$ Intersect$(s_1, p_1) \rightarrow (4.24, 7.24)$

$t_1 =$ Segment$(W, V) \rightarrow 6.71$

a_1: PerpendicularBisector(V, W) → $6x - 3y = 26.23$

z = Intersect(t_1, k_1) → undefined.

A_1 = Intersect(a_1, t_1) → $(7.24, 5.24)$

b_1 = Segment(Q, A_1) → 5.41

e_1 = Segment(R, A_1) → 3.35

f_2 = Segment(A, Q) → 11.09

g_2 = Segment(C, R) → 11.09

h_2: PerpendicularBisector(A, Q) → $-4.24x - 10.24y = -61.46$

B_1 = Intersect(g, h_2) → $(2.12, 5.12)$

i_2: PerpendicularBisector(C, R) → $-4.24x - 10.24y = -56.18$

C_1 = Intersect(i_1, i_2) → $(8.12, 2.12)$

j_2: Line(B_1, C_1) → $3x + 6y = 37.09$

k_2: PerpendicularBisector(B_1, C_1) → $-6x + 3y = -19.86$

D_1 = Intersect(j_2, k_2) → $(5.12, 3.62)$

l_2 = Segment(D_1, R) → 6.27 (6.27)

m_2 = Segment(D_1, Q) → 6.68

E_1 = Point(n) → $(2.97, 1.54)$

F_1 = Intersect(b_1, e_1) → $(7.24, 5.24)$

n_2: Line(E_1, F_1) → $-4.2x + 4.27y = -5.87$

p_2: Line(F_1, E_1) → $4.2x - 4.27y = 5.87$

G_1 = Intersect(n, p) → $(3, 1.5)$

H_1 = Intersect(l_2, m_2) → $(5.12, 3.62)$

I_1 = Intersect(b_1, e_1) → $(7.24, 5.74)$

g_2 = Line(G_1, I_1) → $-4.24x + 4.24y = -6.36$

J_1 = Intersect(n, p) → (3, 1.5)

r_2 : Line(J_1, k) → $x = 3$

K_1 = Intersect(l_2, m_2) → (5.12, 3.62)

s_2 : Line(K_1, k) → $x = 5.12$

L_1 = Intersect(b_1, e_1) → (7.24, 5.74)

t_2 : Line(L_1, k) → $x = 7.24$

M_1 = Intersect(s_2, a) → (5.12, 0.62)

N_1 = Intersect(s_2, m_1) → (5.12, 6.62)

O_1 = Intersect(g_1, n_1) → (4.24, 10.24)

a_2 : PerpendicularLine(O_1, s_2) → $y = 10.24$

P_1 = Intersect(r_2, a_2) → (3, 10.24)

Q_1 = Intersect(s_2, a_2) → (5.12, 10.24)

R_1 = Intersect(t_2, a_2) → (7.24, 10.24)

S_1 = Intersect(r_2, n) → (3, 1.5)

b_2 = Segment(S_1, P_1) → 8.74

T_1 = Intersect(s_2, l_2) → (5.12, 3.62)

c_2 = Segment(T_1, Q_1) → 6.62

U_1 = Intersect(t_2, b_1) → (7.24, 5.24)

d_2 = Segment(U_1, R_1) → 4.5

e_2 = Segment(P_1, R_1) → 4.24

V_1 = Intersect(b_1, e_1) → (7.24, 5.74)

f_3 = Segment(S_1, V_1) → 6

α = Angle(V_1, S_1, P_1) → 45°

Source:
(2) 8 unidades em 3D →
m01-3D-20210728-05.ggb

Checagem dos passos: (20210728)

A = (0, 0)
B = (0, 6)
C = (6, −3)
D = (6, 3)
E = (1, 1)
F = (7, −2)
G = (1, 7)
H = (7, 4)

f: Ray(A, E) → $-x+y=0$

g: Ray(C, F) → $-x+y=-9$

h: Ray(B, G) → $-x+y=6$

i: Ray(D, H) → $-x+y=-3$

Circles nos pontos A, B, C e D, usando o segmento $\overline{AB}$.

Ex:

c: Circle(A, Segment(A,B))
$\rightarrow x^2 + y^2 = 36$

d: Circle(B, Segment(A,B))
$\rightarrow x^2 + (y-6)^2 = 36$

e: Circle(C, Segment(A,B))
$\rightarrow (x-6)^2 + (y+3)^2 = 36$

k: Circle(D, Segment(A,B))
$\rightarrow (x-6)^2 + (y-3)^2 = 36$

Intersect das retas com os círculos:

I = Intersect(c, f, 1)
$\rightarrow (4.24, 4.24)$

J = Intersect(d, h, 1)
$\rightarrow (4.24, 10.24)$

K = Intersect(e, g, 1)
$\rightarrow (10.24, 1.24)$

L = Intersect(k, i, 1)
$\rightarrow 10.24, 7.24)$

Não deixe visível os Raios e os Círculos, deixando os pontos de intersecção visíveis para podermos fazer as retas do cubo.

Apague também os pontos (esconder):
E, F, G e H.

Erro:

Distância $\overline{AC} = 6{,}71$

tinha que ser 6,00

Vou refazer colocando

2 arestas sobre o eixo x e y.

(menos chance de erro)

Check 02 (20210728):

A = Intersect(xAxis, yAxis) -
⇒ (0, 0)

B = Point(xAxis) → (6, 0)

C = Point(yAxis) → (0, 6)

D = (6, 6)

—— 45°

E = (1, 1)

F = (7, 1)

G = (1, 7)

H = (7, 7)

Criar os Rays:

f: Ray(A, E) → $-x + y = 0$

Usarei a mesma reta no ponto D → continuidade da reta

Obs: a mesma reta AE continua em $\overline{DH}$.

g: Ray(B, F) → $-x + y = -6$

h: Ray(C, G) → $-x + y = 6$

——

Criar os círculos e as interseções sobre os Rays $\overline{AE}$, $\overline{BF}$, $\overline{CG}$ e $\overline{DH}$.

Usando a função COMPASS:

c: Circle(A, Segment(A, C))
→ $x^2+y^2=36$

d: Circle(B, Segment(A, C))
→ $(x-6)^2+y^2=36$

e: Circle(C, Segment(A, C))
→ $x^2+(y-6)^2=36$

k: Circle(D, Segment(A, C))
→ $(x-6)^2+(y-6)^2=36$

obs: Alterar as cores destes círculos.

Usar o comando Intersect:

I = Intersect(c, f, 1) → (4.24, 4.24)
J = Intersect(d, g, 1) → (10.24, 4.24)
K = Intersect(e, h, 1) → (4.24, 10.24)
L = Intersect(k, f, 2) → (10.24, 10.24)

obs: Esconda os círculos e os Rays, deixando apenas as interseções.

Depois, criaremos as retas do cubo.

Fazer os segmentos do Cubo:

i = Segmento(A, C) → 6 (altere a cor já)

j = Segment(B, D) → 6

Crie o ponto M=(3, 3)

Crie os novos segmentos:

l = Segment(C, M) → 4.24

m = Segment(M, D) → 4.24

n = Segment(A, I) → 6 (preto)

p = Segment(B, J) → 6 (preto)

q = Segment(C, K) → 6 (preto)

r = Segment(D, L) → 6 (preto)

s = Segment(I, K) → 6 (vermelho)

t = Segment(J, L) → 6 (Vermelho)

Novos comandos:

a: PerpendicularBisector(I, K) → y = 7.24

obs: Vale para o segmento s(I, K) e t(J, L)

Intersect:

N = Intersect(a, s) → (4.24, 7.24)

O = Intersect(a, t) → (10.24, 7.24)

Obs: Esconda a linha a.

Important: Crie.

b: PerpendicularBisector(N, O) → x = 7.24

Importante observar que este ponto coincide com a reta r (D, L) → somente na perspectiva.

Obs: reta a corta a reta r. (interseção visual, mas não real).

Criemos a interseção:

P = Intersect(a, b) → (7.24, 7.24)

Obs: Agora sim, esconda o segmento a e b.

Crie os segmentos:

f_1 = Segment(K, P) → 4.24 (vermelho)

g_1 = Segment(P, L) → 4.24 (vermelho)

Divisão de Segmentos:

Segmentos $\overline{CK}$ e $\overline{DL}$ com o Comando Bisector

h_1: PerpendicularBisector(C, K):
$\Rightarrow -4.24x - 4.24y = -43.46$

i_1 : Perpendicular Bisector(D, L):
$\Rightarrow -4.24x - 4.24y = -68.91$

Intersect :

R = Intersect(h_1, g) → (2.12, 8,12)
S = Intersect(i_1, r) → (8.12, 8.12)

Obs: Esconda h_1 e I_1.

Give Segment :

j_1 = Segment(R, S) → 6
k_1 : Perpendicular Bisector(R, S) ⇒
$x = 5.12$
T = Intersect(k_1, j_1) → (5.12, 8.12)
Esconda j_1 e k_1.

Outros comandos (base cubo)

data

I_1 : Perpendicular Bisector (A, I) →
$-4.24x - 4.24y = -18$

m_1 : Perpendicular Bisector (B, J) →
$-4.24x - 4.24y = -43.46$

Intersect:

U = Intersect(I_1, n) → (2.12, 2.12)

V = Intersect(m_1, p) → (8.12, 2.12)

n_1 = Segment(U, V) → 6

p_1 : Perpendicular Bisector (U, V) →
$x = 5.12$

W = Intersect(p_1, n_1) → (5.12, 2.12)

Obs: Apague o resto e deixe somente este ponto.

q_1 = Segment(I, W) → 2.3

r_1 = Segment(W, J) → 5.54

s_1 = Segment(K, T) → 2.3

t_1 = Segment(T, L) → 5.54

Other:

Segmento Inclinado:

a_1 = Segment(A, K) → 11.09

b_1 = Segment(B, L) → 11.09

c_1: Perpendicular Bisector(A, K) →

$-424x - 10.24y = -61.46$

d_1: Perpendicular Bisector(B, L) →

$-4.24x - 10.24y = -86.91$

Z = Intersect(c_1, a_1) → (2.12, 5.12)

A_1 = Intersect(d_1, b_1) → (8.12, 5.12)

e_1 = Segment(Z, A_1) → 6

f_2: Perpendicular Bisector(Z, A_1) →
$x = 5.12$

B_1 = Intersect(e_1, f_2) → (5.12, 5.12)

Important:

Os Pontos M, B_1 e P estão em linha reta.

Distância entre MP → 6 unidades.

Parte Final:

Crie a linha, ligando os pontos $\overline{KL}$

i_2 : Line (K, L) → y = 10.24

Criemos 3 segmentos com o comando Parallel Line baseado no segmento $\overline{AC}$ (6 unidades)

j_2 : Line(M, i) → x = 3

k_2 : Line(B_1, j_2) → x = 5.12

I_2 : Line(P, k_2) → x = 7.24 (7.24)

Intersect:

C_1 = Intersect(i_2, j_2) → (3, 10.24)

D_1 = Intersect(i_2, k_2) → (5.12, 10.24)

E_1 = Intersect(i_2, I_2) → (7.24, 10.24)

FINAL:

m_2 = Segment(M, C_1) → 7.24

> 2.12

n_2 = Segment(B_1, D_1) → 5.12

> 2.12

p_2 = Segment(P, E_1) → 3

q_2 = Segment(C_1, E_1) → 4.24 = soma

Source:
ARQ-IMPORTANT_03/
Geogebra-M01-20210727/
m01-20210728-check-01-012.ggb
(arquivo final).

Obs: O próximo arquivo será criado para excluir os pontos e os nomes dos elementos, ficando apenas a imagem final.

m01-20210728-check-01-014.ggb

Imagens Criadas:

m01-20210728-check-01.png
m01-20210728-check-01.pdf

Duração: 12:00 - 17:45h
(20210728)

20210729

Estreito de Málaca (Singapura)
Estreito de Sunda (Jacarta)
Estreito de Lombok.

—

"O homem que move rio, move desesperado" Carnegie.

—

scientificmagazines.top

—

20210730

New simulators:

File: $m01-20210730-check-01-001.ggb$

c_2: Circle(B_1, M) →
$(x-5.12)^2+(y-5.12)^2=9$

Important:

3 Points:

1) Point P (7.24, 7.24)
2) Point B_1 (5.12, 5.12) e
3) Point M (3, 3)

→ Estes três pontos estão alinhados dentro do círculo c_2 e dois destes pontos estão sobre a circunferência.

Outros 3 pontos:

1) Point T(5.12, 812),
2) Point B_1(5.12, 5.12) e
3) Point W(5.12, 2.12)

→ 3 pontos alinhados e 2 pontos sobre a circunferência.

Todos os 6 pontos se encontram num mesmo ponto B_1, ou seja, as retas $\overline{PM}$ e $\overline{TW}$ fazem interseção no ponto B_1.

De acordo com Pappus Theorem observamos uma reta criada a partir de 3 pontos.
O ângulo entre estas retas é de 45°.

Obs: Simulação em 3D:

m01-3D-20210730-01.ggb

Usando a imagem 3D, utilizei o comando: Circle through 3 Points

Points: M_1, K_1, N_1

c_2: Circle(M_1, K_1, N_1) →

$$\to X = (4, 4, 4) + (0, -2.83\cos(t) - 2.83\sin(t), 2.83\cos(t) - 2{,}83\sin(t))$$

Localização dos pontos:

M_1 = Intersect(j_1, n_2) → (4, 8, 4)

K_1 = Intersect(b, m_2) → (4, 4, 8)

N_1 = Intersect(j, I_2) → (4, 0, 4)

Marcando as 2 retas que se intersectoram no ponto (B_1) J_1
↳2D ↳3D.

a) $M_1(j_1, n_2)$, $J_1(4,4,4)$ (m_2, g_2) e $N_1(j, I_2)$

$$\downarrow \qquad \downarrow \qquad \downarrow$$

(4, 8, 4)   (4, 4, 4)   (4, 0, 4)

b) $K_1(b, m_2)$, $J_1(m_2, g_2)$ e $L_1(s_1, r_1)$

$$\downarrow$$

(4, 4, 8)   (4, 4, 4)   (4, 4, 0)

Estudar:

https://en.m.wikipedia.org/wiki/

Squaring_the_circle

obs: File 01 → Squaring the circle - Ramanujan - 1914.gif

File:01-Squaring_the_circle=Ramanujan_1914.gif

Circulo horizontal:

d_2: Circle(J_1, 4, m_2) →

$x = (4,4,4) + (4\cos(t), 4\sin(t), 0)$

Outros dados:

g_2 = Segment(N_1, K_1) → 5.66

O_1 = Intersect(edgeBC, κ) → (8,0,8)

r_2 = Segment(O_1, J_1) → 6,93

Source:

m01-3D-20210730-03.ggb

20210801

Sir Chandrasekhara Venkata Raman, físico indiano que em 1930 se tornou o primeiro asiático a ganhar o Prêmio Nobel de Física.

Trabalho de Raman é centrado na dispersão da luz, que mostra que quando a luz atravessa um material transparente, parte da luz dispersada muda o comprimento de onda e amplitude, conhecido como

Efeito Raman princípio fundamental no campo da óptica

Ele descobriu que a cor do oceano é azul não por uma simples reflexão do céu, mas porque a água em si faz com que a luz azul se espalhe.

20210802

http://www.edocente.com.br/pnld/2018/
2019
2020

---

http://ramanujan.math.trinity.edu/wtrench/texts/TRENCH_FREE_DIFFEQ_I.pdf
II

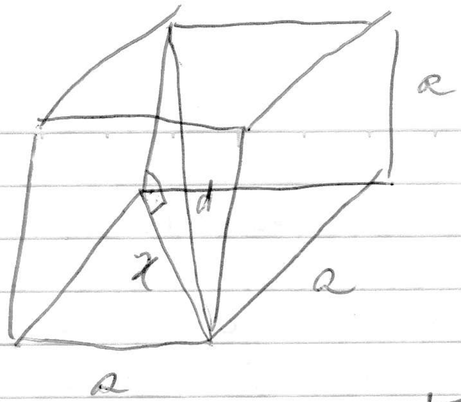

Cubo é um caso particular de paralelepípedo reto retangular

$d = \sqrt{a^2+b^2+c^2} = \sqrt{3a^2} = a\sqrt{3}$

$d = a\sqrt{3}$

Source: Books-016-2021/20210802/
Math/Math-Dante-3-Volumes/
Matematica Cont Aplic_2_MP_
0008P18023_PNLD2018.pdf
Pág. 177 (179)

others:

Books-016-2021/20210802/Math/
ebook-final-isbn.pdf
Pág. 119 (132) - capítulo 6
Atividades Envolvendo
Circunferência com o
Uso do Geogebra

Books-016-2021/20210802/
Differential-Equations/
math135.pdf

npde10.pdf

data

S T Q Q S S D

20210803

Archimedes (240 A.C)

$22/7 < \pi < 223/71$

William Oughtred foi o primeiro a empregar a letra grega π como símbolo del coeficiente entre las longitudes de una circunferência e su diámetro.
(comprimento)

obs: $\pi \cong 335/113$

Tsu Ch'ung-chih (480 A.C.)

Source: 3con14.com/32-útiles/curiosidades/71-decimales-de-pi.html

---

Milagre!!!

Kombi Quebrou 20191104 (2 dias depois do (Seg-Feira) Dia de Finados)

Google (mark) - -23.61661, -46.62734

Temperatura: 37° C

Na esquina tinha um cara vendendo água gelada

SP-160 → Av. Prof. Abraão de Morais

200 metros do Shopping Plaza Sul.

Embaixo da árvore que tem buraco para estacionar - 50 m General Santos

As peças para consertar a Kombi foram compradas na R. Gen. Chagas Santos → a primeira rua à esquerda.

Na loja no cruzamento da R. General Chagas Santos e a Rua Cuminapanema → Auto Peças Carvalho (Kombi à loja ≅ 300 metros)

20210805 ///

Spacetime - Marcus Kriele
The Ricci Flow in Riemannian Geometry - Ben Andrews

Método catenária → correção do sistema pantográfico aos cabos superiores (energia elétrica).

Schwartz surface.

Joseph Raphson x Newton
(50 anos)

→ cycle contraction for 42+6 results in 22/7 as 48 and 42 finds a simple pi in between, for 7

others: https://community.wolfram.com/groups/-/m/t/2311778

Newton - Raphson method. (1690) method for approximating the roots of an equation.

Raphson → Analysis Aequationum Universalis (1690)

Newton → Method of Fluxions (1671),

but this work would not be published until 1736, nearly 50 years after Raphson's Analysis.)

Raphson was a staunch supporter of Newton's claim, and not that of Gottfried Leibniz, to be the sole inventor of calculus.

In addition, Raphson translated Newton's Arithmetica Universalis into English.

—

Raphson – De Spatio Reali published in 1697.

—

Standard candles and the value that is calculated using the lambda-cold-dark-matter (Λ CDM) model of cosmology.

—

300.000.000 planetas ⇔ em nossa galáxia

↓

1 planeta habitável:

Kepler-442b (KOI-4742.01)

descoberto em 2015.

(condições (semelhante às da Terra).

20210806///

Data-NovoVivo-016-2021/Books-016-2021/
20210806/Doing Mathematics Convection, Subject, Calculation, Analogy/

Martin H. Krieger - Constitution of Matter .... djvu

Pág. 37 (57) - the thermodynamic limit./

---

to + verbo infinitivo

for + gerúndio

Ex: She called to talk to you.

Conversations are good for realizing what's wrong.

I will write a letter to her.

We gave the new to them.

He always carries the books for her.

She will check that for me.

to → person.

data

S T Q Q S S D

Google
stemiomusic.com

ID de cliente
C03vwgykD

09 Jun 2021 → Drive 70,98 GB
Drive Compartilhado 128,32 TB

Books1-016-2021/20210806/
The Geometry of the Universe/
The Geometry of the Universe - Colin Rourke.pdf

Pág. 73(88)
5.4 Conservation of angular momentum
Figure 5.3 - Proof of conservation of angular momentum

Books1-016-2021/20210806/
Invisible-Universe/
The Invisible Universe Dark Matter ...pdf

Pág 138(151) 6.7 The Higgs Mechanism

data

S T Q Q S S D

20210807 m 20210806 m

Books-016-2021/20210804/

The Absolute Differential Calculus (Calculus of Tensors)/
0486634019.djvu

Pág. 172(187) - Chapter VII

Riemann's Symbols and Properties relating to Curvature:
Ricci's and Einstein's Symbols;
Geodesic Deviation

1. Cyclic displacement and the relations between parallelism and curvature

---

20210804/Mathematical Logic.../
0442043546.djvu

Pag. 224(119) - Supplementary Notes on Chapter 7

Foundations ~~of~~ of projective geometry

Pag. 229(121) Chapter 8
Gödel's theorems on the inherent limitations of formal systems

1. Gödel's new metamathematical method.

20210807

Cubic Curves in the Triangle Plane

Pinkernell, G. M.

—

https//doi.org/10.1007/BF01223040

March 1996

—

https//link.springer.com/article/10.1007/BF01223040

—

Source: Books-016-2021/Important/triangle_conics_and_cubics.pdf

—

20210808

Graham Farmelo.

graham@grahamfarmelo.com

—

twitter.com/grahamfarmelo

—

www.sciencenews.org/sn-magazine

www.sciencemag.org

—

Books-016-2021/20210808/

→ Math/Analisis_Matematico_Calculo_I_Espinosa.pdf

→ The center and Focus Problem.../
The_Center_and_Focus_Problem....pdf

→ Heat and Mass Transfer Ed 2/
9811539871.pdf

→ Science News - 31 July 2021/
sciencenews2021041of-dl.pdf

Pág. 5 (7) - Newton's law of gravity
hold up in miniature.

→ A Geometrical Framework for
Understanding the Universe/
1793304467.epub

→ Introduction to Lattice Algebra.../
Introduction to Lattice Algebra....pdf

→ The Universe Speaks in Numbers/
Graham Farmelo - The Universe
Speaks in Numbers.epub

→ Towards the Holy Grail of Physics/
1794633146f.epub
usar junto com A Geometrical Framework
for Understanding the Universe.

→ Modern classical physics optics, fluids.../
Modern Classical physics optics....pdf

data / /

S T Q Q S S D

20210809

sciencenewsdigital.org

pdfdrive.com

—

Books-016-2021/20210809/

→ Analytical Geometry 2D and 3D/
Analytical Geometry 2D and 3D.pdf

⇒ A Gateway to Modern Mathematics ..../
A Gateway to Modern .... I ...pdf

→ Pag. 121 – chapter 13
Geometric Iterations

⇒ Gravity/A First Course in String Theory.pdf

—

⇒ Gravity/grbook.pdf *

⇒ Gravity/John H. Hubbard - Vector Calculus, Linear Algebra .. 2009.pdf

/John Hubbard ... 2015.pdf

=

⇒ Methods of Geometry/

→ Space, Time and Einstein/

20210810M

Início dos reparos da calçada de Cumbica.

20210812M

uploaded.net.

—

There is measure in all things — Horace.

Source:

Books-016-2020/20210812/
Vector-Calculus/
emu3ruky5e4eyaky35.pdf

Pág. 430(447)

4.4. Measure Zero

—

20210813M

20210814M

$$a^{\frac{2}{3}} = \left(\sqrt[3]{a}\right)^2$$

—

$$a^{1/n} = \sqrt[n]{a}$$

—

20210815 ııı

2-Geral

—

Book1-016-2021/20210815/

/Handbook of the Mathematics of the Arts and Sciences/

3319570714.pdf

Pág. 123(144)

/Homeomorphisms Between the Circular Disc and the Square.

Pág. 126(147)-/

Fig. 3 Some mapping to convert a disc to a square and vice versa

—

Pág. 57(79)/

Fig. 15 Isomorphisms between the hemisphere model, the Klein model, and the Poincaré model.

Pág. 58(80)/

Fig. 16 Lines in Poincaré disk model and the half plane model

Pág. 136 (157)

Fig. 13 A conformal map between the disc and square in the complex plane.

Pág. 140 (161)

Fig. 17 Playfair's axiom on the Euclidean plane (left) and the hyperbolic plane (right)

Pág. 158 (179)

Ford Circles and Farey Sequences

→ "geometric picturizations" of fractions.

Pág. 159 (180) -

Fig. 9 Ford circles for $\frac{3}{5}$, in red, and the Ford circles of its adjacent fractions $\frac{1}{2}$, $\frac{4}{7}$, and $\frac{2}{3}$. Notice that the Ford circles are an infinite family of circles tangent to each other.

Ford proved (1938, theorem 4) that any such (qualquer tal) irreducible

irreducible fraction of with
|p| > 1 has exactly two adjacent
fractions with denominator
smaller than p. ...

Important;

Pág. 173 (193)

Each Surface admits more than one geometric shape

Pág. 174 (194)

Fig. 2 Moduli space for the surface of genus 2

Fig. 3 From a regular hexagon on the plane to the torus.

Pág. 183 (203) ~~Ana~~

Anamorphosis Forward Fast

Fig. 3 Two very different 3D objects with the same visual core..

Fig. 4 Two very different 3D objects with the same visual core, one surrounding the other.
Anamorphosis is radial and isotropic, with no

preferential axis.

Fig. 5 (184)(204)

Projection onto a plane.

There is no absolute difference between the linear perspective and the oblique anamorphosis.

Both are anamorphosis onto a plane: the latter merely projects the (object) to a region further away from the foot of the perpendicular

Pág. 190 (210)

Fig. 10 choice of triangulation.

Pág. 201 (221)

Fig. 14 Anamorphosis onto the sphere of two parallel lines.

The lines project as meridians that meet at two common vanishing points V and V*, antipodal to each other.

Pág. 202 (222)

Fig. 15, ~~the~~ two lines l and j intersecting at a right triangle.

Fig. 16 Anamorph of a cube on a sphere.

the cube defines a set of distinct vanishing points, two for each set of parallel edges (arestas)

Pág. 206 (226)

Fig. 18

Descriptive geometry construction of the anamorphosis of a cube onto a cylinder.

→ Projection

→ point the edges of the cube appear straight (c)

A Brilliant Young Mind 2014

20210818 ///

Moradia Roberto
Ricardo

John G. Fisher

Pi

William Shanks (1874)

Pi → 707 decimal places.

last numbers:

7817924264 π

Obs! 62.8 Trillion Digits

—

20210819

Copiar pasta Google

technoblog.net/245642/google-drive-copiar-pasta-outra-conta/

—

https://colab.research.google.com

%cp -av " " " "

data
Ⓢ Ⓣ Ⓠ Ⓠ Ⓢ Ⓢ Ⓓ

Google Drive atualização
KB 3033929.

metamath zero (proof)
metamath.org

turnstil ⊢ (Frege 1879)
⊢

Institute of Mathematics
and its aplications

ima.org.uk UK

quantamagazine.org/mathematics/

Peter Rowlett

twitter.com/peterrowlett

peterrowlett.net

data
S T Q Q S S D

Metamath:
A Computer Language
for Mathematical Proof
Norman Megill

—

metamath.org

—

20210820 III

Good Will Hunting Film.
Gênio Indomável

—

Eventiv (Khodorkovsky)

—

chrome.google.com/webstore/
detaies/copy-folder/

ericyd.com

—

www.raidrive.com

—

nm@alum.mit.edu

—

metamath.org

—

Metamath.org

1) Pi is a real number:

http://us.metamath.org/mpeuni/pire.html

2) 2 < pi < 4

http://us.metamath.org/mpeuni/pigt2lt4.html

3) 3 =< pi

http://us.metamath.org/mpeuni/pige3.html

$\text{sen}_{\alpha} = \frac{b}{a}\left(\frac{CO}{H}\right)$

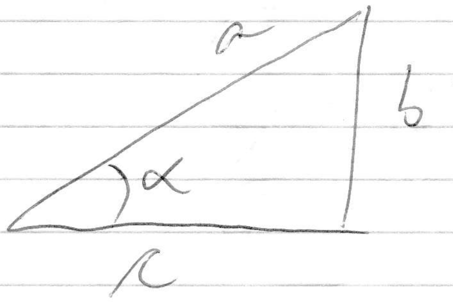

$\cos_{\alpha} = \frac{c}{b}$

$\tan_{\alpha} = \frac{b}{c}$

Seno de um ângulo

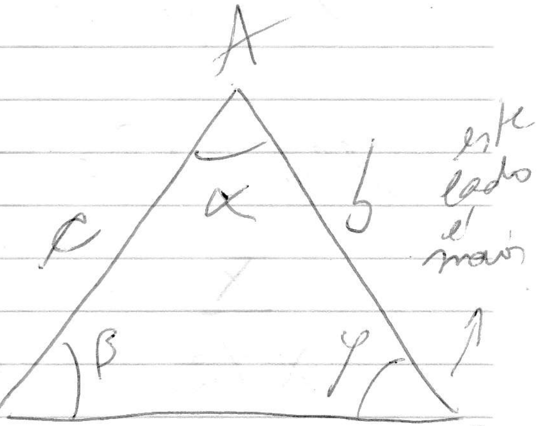

O seno de um ângulo é a razão entre o cateto oposto ao referido ângulo e a hipotenusa de um B triângulo retângulo.

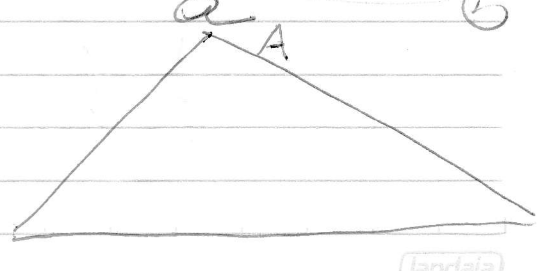

Ângulos notáveis

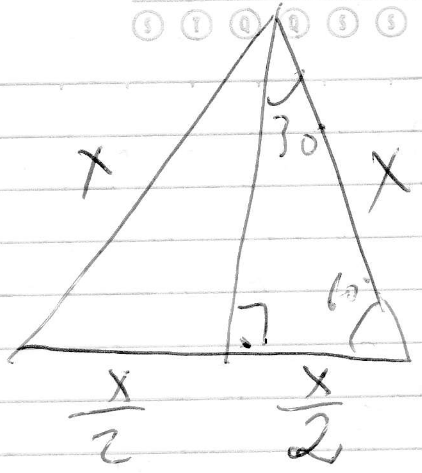

x
30
x
60°
$\frac{x}{2}$
$\frac{x}{2}$

$\text{seno } 30° = \frac{CO}{h}$

$\text{sen } 60° = \frac{CA}{h}$

$\text{sen}(30°) = (x/2)/x = x/2 \, . \, 1/x = 1/2$

$$\frac{\frac{x}{2}}{x} = \frac{\cancel{x}}{2} \times \frac{1}{\cancel{x}} = \frac{1}{2}$$

$\text{sen}(60°) = \frac{x}{h}$

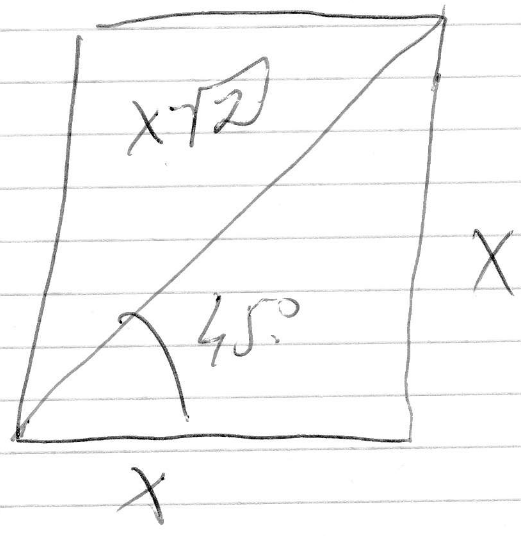

$x\sqrt{2}$
x
45°
x

$\text{sen } 45° =$

$\text{sen } 45° =$

$$\frac{\cancel{x}}{\cancel{x}\sqrt{2}} = \frac{1}{\sqrt{2}}$$

Definição do seno de um ângulo.

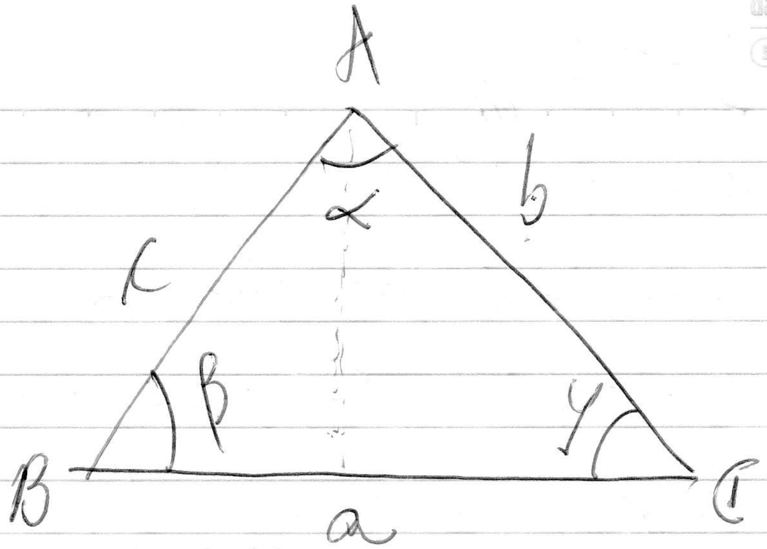

$$\frac{a}{\text{sen}\,\alpha} = \frac{b}{\text{sen}\,\beta} = \frac{c}{\text{sen}\,\gamma}$$

$\hookrightarrow \frac{co}{h}$

O seno de um ângulo é a razão entre o cateto oposto ao referido ângulo e a hipotenusa de um triângulo retângulo.

∴ divisão da medida do cateto oposto pela medida da hipotenusa

$$\boxed{\text{sen} = \frac{CO}{h}}$$

Celestial Mechanics – Hagihara (5 volumes)

Source: Arq_Barros\04_02\

—

\Geometry_Gauss

—

Norman Megill (Megill)

nm@alum.mit.edu

—

fusão nuclear com confinamento inercial.

→ obtenção do "plasma ardente" em que as próprias reações de fusão dão calor para mais fusões.

other:

fusão por confinamento magnético.

—

20210823

https://www.cs.ru.nl/~freek/100/

Formalizing 100 Theorems

—

Megill's completeness theorem

Prêmio Nobel Luc Montagnier

—

eufemismo.

—

longfiles.com

—

20210824

Star Trek That which Survives

The Cage

2 parts { dailymotion.com/video/
x8219jk
—
x8219it

=

Survives — x7y5ug7

—

Google Drive
Fologo for Google Drive

Drive Webcam camera for google.

Auto-LaTex Equations

689895236653934908/

→ Geometria do Ovo

=

hence - portanto

lies - encontra-se em
on

—

transitivity (transitivity) theorum
for parallel lines

Pág. 158(177)

Source: Books-017-2021/
20210825/
Fermat/
Methods of Geometry.pdf

—

grandade híbrida generalizada -
Palatini.

—

ring light (LED)

=

Books: Source

Books-017-2021/20210825/

⇒

Einstein in Matrix Form
Exact Derivation....pdf

Pág. 6 (19) 1.13 Maxwell's Eq. and Galilei Transformation
5 (19) 1.1.2
General Galilei Transformation

+

Yehuda B. Band ... - Quantum mechanics ... (2013).pdf

Pág. 208 (229) 5.2.2
Quantum Entanglement and Bell States

$\left(\nearrow \sqrt{2} \right)$

Problem 5.9

Write the Bell states (5.25a) as four-component vectors using the notation of Problem 5.8.

$$|\phi^+\rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \ldots$$

20210825/ Electromagnetism
Maxwell... /
Electromagnetism Maxwell
Equation ... pdf

Pag. 7 (27)
1.3. Integrals involving vectors

Circulation of a vector field

Fig 1.2.

20210825/ Fermat/
Methods of Geometry... pdf

Pág. 159 (178)
Theorem 3 (Angle Bisectors Theorem)

The angle bisectors of ΔABC meet at an interior I equidistant from the edge lines

Fig. 5.1.2

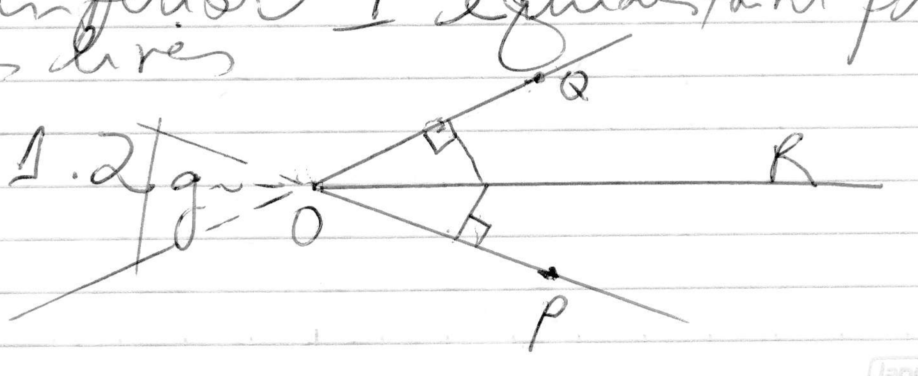

20210825/ Quantum/

general_relativity.pdf

Pág. 149 (162) - Chapter 7

Curvature

Pág. 152 (165) Figure 7.2

A two-dimensional surface embedded in three-dimensional Euclidean space.

Pág. 153 (166) 7.3

The Riemann Curvature Tensor

Pág. 158 (171)

The Weyl Curvature Tensor

Pág. 159 (172) 7.4

Extrinsic and Intrinsic Curvature

Pág. 162 (175) - 7.5

The Equation of geodesic deviation

data

S T Q Q S S D

20210829//

Vagueness in the Exact Sciences

—

Fermagoric Triangles
Pedro A. Pizá

—

O medo de perder tira a vontade de ganhar.

—

O sabor da vida não tem receita.

—

Instituto Solid Rock Brasil
R. Imperador 1036
institutosrbrasil.org.br.

—

Spook - the Great Good
Spook - Jogo de Espiões.

—

Books:

Data-NovoVivo-017-2021/

20210831 m

stuck = peça (em alemão)
[illegible] (english)
↳ [illegible]

20210901 m

Angie Stone These Eyes

Rebecca Ferguson Glitter & Gold

PayPal - ~~QR-Code~~ Donate

New Website
(updated) a

5/6 Jan 1709

Great Frost.

Frio na Europa
(600.000 mortos)
↙
França.

-20°C na França.

Paris - 3 meses sem receber nada. (provisões)

Mínimo Maunder + Vulcões

Books-018-2021/20210901/

→ Calculus-Math/

Calculus A Rigorous First Course.pdf

→ An Introduction to Complex Analysis and Geometry/

0821852744.pdf

Pag. 5(17) Fig 1.1 Difference of two squares

Pag. 21(33) 6. A glimpse at metric spaces.

20210902

7300 Townline Rd Lakeview,
MI 48850

---

pdfdrive.com

libgen.st

sanet.ws/books/

ebook3000.com/science/

---

abakcus.com

Park Yoshemite (Sequoias)

—

Troca de registro da água.

—

https://gallica.bnf.fr/ark:/12148/bpt6k9730303m

→ Journal de Mathématiques Pures et Appliquées.

—

109 années disponibles -
168 numéros

—

Source:
https://gallica.bnf.fr/ark:/12148/cb343487840/date

—

O amor de Deus nos prepara para a Eternidade

—

https://gallica.bnf.fr/ark:/12148/bpt6k96431405 →

Application de l'analyse à la géométrie – G. Monge

→ 12148/bpt6k6227759f →

Recherches générales sur les surfaces courbes – Gauss

→ 12148/bpt6k3942g →

La théorie des parallèles
Lobatchevski

Nikolaj Ivanovič Lobačevskij
(1792-1856)

Vectors, St. Lines, Planes and Mechanics

Narayana Dash
ASIN: B00MQCNJLM
August 14, 2014

Terra: 4,54 bilhões anos ± 50 m

20210905_III Journal Mathematiques

1846 - Charles

Ramanujan

https://www.imsc.res.in/~rao/ramanujan/contentindex.html

Dec 22, 1887 - Apr 26, 1920

—

1869, Liouville,

1880, Euler,

1883 - théorie des fonctions abéliennes (Briot. p. 114) (Pg.1)

B	Amando
1884	1886
1885	1887
	1888

uploaded.net

—

65.239.233.35.bc.googleusercontent.com

122.224.227.35

—

20210906III (20210905)

— Batalha de Waterloo 18 Jun 1815

Napoleão morreu em 1831 e os seus (ossos) restos mortais foram encaminhados para o Royal United Services Institute onde foi preservado.

—

Marechal Grouchy ter é visto como o culpado pela derrota de Bonaparte por não ter evitado que os prussianos se unissem aos ingleses e por não ter acorrido à Waterloo com seus 30 mil homens, quando ouviu o canhoneio da batalha.

—

gbfproofreading.co.uk

—

mathsjam.com/cities/canterbury/

—

www.solipsys.co.uk/cgi-bin/MJ_Wiki.py

—

Sam Hartburn

sam.hartburn@gbfproofreading.co.uk
Phone 07594 602265.
(65)

20210905m (20210905)

Data-NovsLivro-018-2021/
Ramanujam-Works/
KSR chap3.pdf

3. Ramanujan's Notebooks

Pag. 7

"Modular equations and
approximations to π"

π like:

$$\frac{63}{25}\,\frac{17+15\sqrt{5}}{7+15\sqrt{5}} = 3.1415926538\ldots$$

etc.

Source:

www.imsc.res.in/~rao/
ramanujan/images/
KSRChap3.pdf (KSRchap3.pdf)

Obs: Prof. Rao (Srinivasa)
Institute of Mathematics Sciences
Chennai - 600 020, India

Source: www.imsc.res.in/~rao/

ramanujan/contentindex.html

Import:

imsc.res.in/~rao/ramanujan/
CamUnivCpapers/collectedright1.htm

5. Squaring the circle
(Journal of the Indian Mathematical Society, 5 (1913), 132

Source:

imsc.res.in/~rao/ramanujan/
CamUnivCpapers/Cpaper5/
page1.htm

6. Modular equations and approximations to π
(Quarterly Journal of Mathematics, 45 (1914), 350-372)

Source:

www.imsc.res.in/~rao/ramanujan/
CamUnivCpapers/Cpaper6/
page1.htm ⟶ pag17.htm

other:

www.imsc.res.in/~rao/
ramanujan/collectedpapers/
collectedright1.htm

obs: Journal of the Indian
Mathematical Society.

/collectedpaper/squaring/squaring.htm

/collectedpaper/modularEqn/
modularEqn1.htm
Eqn23.htm

Término: 06/09/2021 00:00h

Cópia dos arquivos:
Data-NovoVivo-018-2021/
Books-018-2021/Important/
Ramanujan/
obs: 44 files

data

S T Q Q S S D

20210906

escorbuto.

Anna Chlumsky

nordu.net 1 Jan 1985

symbolics.com 15 Março 1985.

Internet no Ar:

20 Dezembro 1990.

O 1º site foi:

info.cern.ch

→ TIM BERNERS-LEE

→ inventor da internet.

medidores energia elétrica:

(novo) eletromecânicos (antigos - disco)

ciclométricos

eletrônicos, sistema moderno. Usa o sistema PLC (Power Line Communication) pois usa a própria rede para aferir o consumo.

---

imsc.res.in/imsc_logo/

Jandaia

Ave Lyra (Paraná).

—

Histoire des recherches sur la Quadrature du Cercle → Montucla.

Michel Serfati — R$ 666,00

ISBN 978-3-515-12082-1

{ 30 Septembre 1938   80 ans
| 30 Septembre 2018

—

Collatz Conjecture

Vinil HD

↳ empresa austríaca Rebeat.

"HD Vinyl"

—

Books :

Books-018-2021/20210906/

⇒ 1) Modern Engineering Mathematics... /
Modern Engineering Mathematics... .pdf

→2) Physics of Spin-Orbit-Coupled Oxides/
0199602026.pdf

→3) Problems in Quantum Field Theory with Fully-Worked Solutions/
Problems_in_Quantum_Field_Theory.pdf

→4) The Golden Ratio The Divine Beauty of Mathematics PDF/
1536134546.pdf

→5) Leibniz an the invention of Mathematical Transcendence/
3515120823.pdf

Other:

Books-018-2021/Important/Ramanujan/
Ramanujan-Prof. Steven Musich www.
stevenmusich.com.pdf

- Books-018-2021/20210906/
Kurt Godel - Works - 5 Volumes/

- Data-NovoVivo-018-2021/Books-018-2021/
Important/

data

S T Q Q S S D

England → Knighthood
(transformar em "sir"

Realeza Britânica →
"never complain, never explain"
(nunca reclame, nunca explique)

20210907

pantheon.io
drupal.org

---

pralay@imsc.res.in
Dr. Pralay Chatterjee

---

www.imsc.res.in/pralay_chatterjee

Martika - Toy Soldiers

---

www.imsc.res.in/mathematics
Fermat's Room 2007

4 7 min 9 min

Ao mesmo tempo ampulheta
4 e 7 min

4 vira
novamente
+4 min → qdo der 3 → acaba a
ampulheta de 7,
↓
esse minuto que falta dará 8 min.
mas a de 7 terá (novamente) rodado 1 min
Dando 9 minutos.

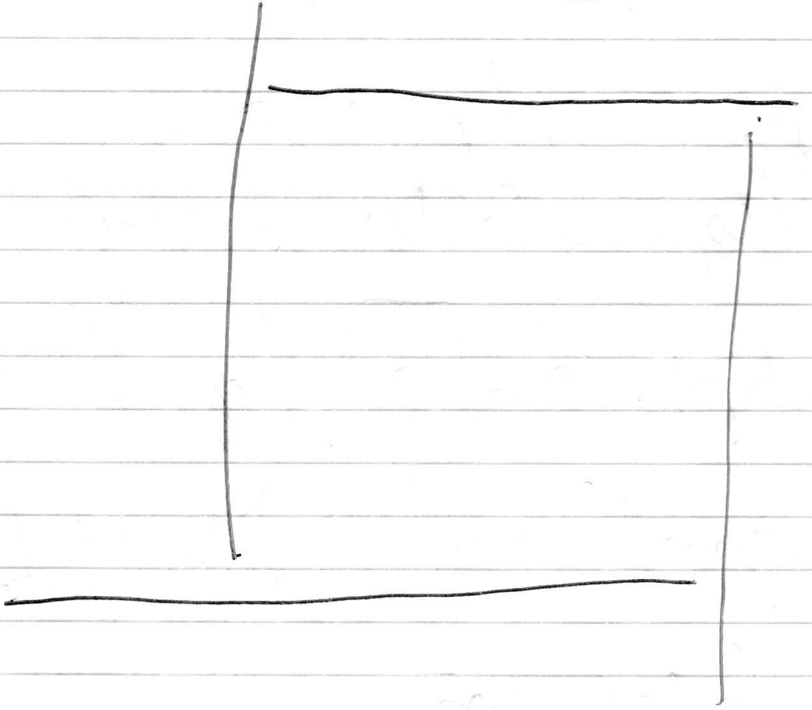

4 pregos

53 min
47 seg.

Matemática Insólita
La Montaña Mágica - Thomas Mann

Movie: The Girl next Door.

MIM-SM
MIM-M
MIM-P
MIM-MP

Motores magnéticos →
não tem manual, ninguém

~~E~~. 20210908

gsuitetips.com/tips/more/scripts/copy-a-folder-in-google-drive/

gsuitetips.com/tools/copy-folder/

chrome.google.com/ → Drive migrator.

longfiles.com
uploaded.net
opendrive.com

20210909
Books cont.

- Books-018-2021/Advanced Engineering Mathematics...../
  0471488852.pdf
  Pág. 35(54) 1.6 Orthogonal Trajectories.

An important type of problem in physics or geometry is to find a family of curves that intersect a given family of curves at right angles..

- Books-018-2021/20210909/ Nanohertz Gravitational Wave Astronomy/
  Nanohertz Gravitational Wave Astronomy.pdf
  Pág. 12(25) 2.2.
  Waves from Geometry.
  Pág 14 (27) Figure 2.2
  The periodic deformation of spacetime caused a GW is illustrated for the two distinct polarizations permitted

1992: $e^{\pi}-\pi = 19.999099979\ldots$

2006: New formulas for Pi and 1/Pi:

$$\frac{1}{\pi} = 8\sum_{n=1}^{\infty}\frac{n}{e^{\pi n}-1} - 40\sum_{n=1}^{\infty}\frac{n}{e^{2\pi n}-1} + 32\sum_{n=1}^{\infty}\frac{n}{e^{4\pi n}-1}$$

$$\pi = 72\sum_{n=1}^{\infty}\frac{n^{-1}}{e^{\pi n}-1} - 96\sum_{n=1}^{\infty}\frac{n^{-1}}{e^{2\pi n}-1} + 24\sum_{n=1}^{\infty}\frac{n^{-1}}{e^{4\pi n}-1}$$

(wrong!!!)

Source:

https://vixra.org/abs/1409.0078

Identities Inspired by the Ramanujan Notebook, Second Series

---

https://vixra.org/abs/1410.0054

2011 - A formula for mass ratio of the neutron/proton:

$$\frac{8}{27}\left(-\sqrt{3}+\frac{5}{\frac{1}{8}(-1+\sqrt{5})+\frac{1}{4}\sqrt{\frac{3}{2}(5+\sqrt{5})}}\right)$$

obs: title:

The Art of Inspired Guessing

→ one is related to the mass ratio of the neutron and proton:

1.00137841917

simon.plouffe@univ-nantes.fr

simon.plouffe@sympatico.ca

Não me respondeu em nenhum destes emails.

As equações estão equivocadas.

https://cs.uwaterloo.ca/journals/JIS/compass.html

/construction 503.gif

Ele não teve nem a consideração de responder pois ele sabe que o que eu estou falando é verdade, mas ele não consegue admitir e nem dar razão à sua realidade inquestionável.

Para muitos pessoas admitir que está equivocado é um duro golpe para o seu ego.

Admitir a verdade, liberta a alma

20210911 (20210810)

https://archive.org/details/

einleitungindieh00kleiuoft

Vorlgeovonalfre01debrich

Others:

Leçons sur la géométrie 3 vol.

www.archive.org/details/

leonssurlago01debuoft
03

/vorlesungenber11 deb
12 deb
21 deb

/rieflachvolesung00klievich

triângulo escalino → 3 lados ≠s.

triângulo acutângulo → <90°
triângulo retângulo → =90°
triângulo obtusângulo → >90°

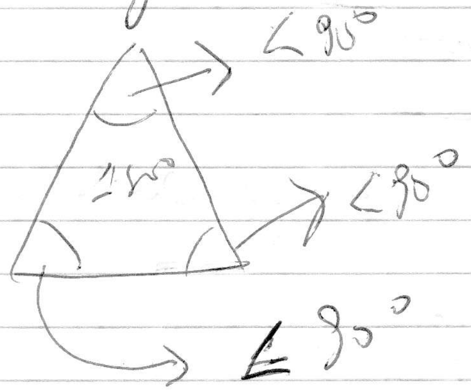

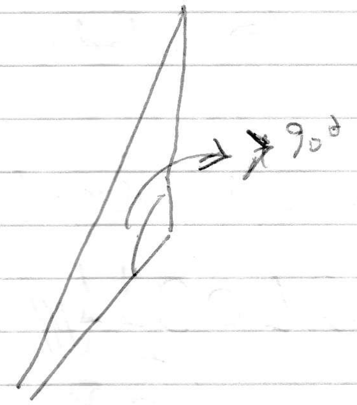

triângulo equilátero
triângulo isósceles
triângulo escaleno

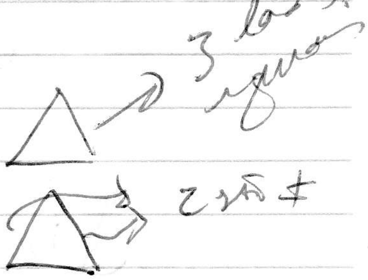

3 lados iguais
2 são ≠

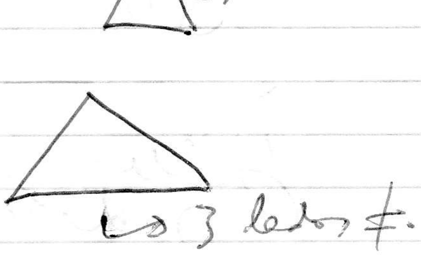

/planealgebraiccu00hilt_0

13edtreatiseonhighesalmonoft

Treatise on the higher plane curves
Salmon

1treatisehigherpl00salmrich
(1mrich)

1 coursemathematic0202goumich

---

trapézio isósceles, reto, escaleno (scalene triangles)
↳ 2 lados = (paralelos)

Soma dos ângulos de um quadrilátero = 360°

20210911
Samantha

Sb 11,20

Catecismo 299.

Deus cria um mundo ordenado e bom

299. Uma vez que Deus cria com sabedoria, a

data
S T Q Q S S D

criação possui ordem.

"Dispusestes tudo com medida, número e peso" (Sb 11,20)

Sb 11,20 → "E, mesmo sem isto, eles poderiam perecer por um sopro perseguidos pela justiça e aniquilados pelo vento de vosso poder; mas, dispusestes tudo com medida, quantidade e peso."
(número)

IMPORTANT:

Data-NovoVivo-018-2021/Books-018-2021/
20210911/Quadrature theory)/
(Quadrature theory)

Tangent.._N°49 Les maths de l'impossible La quadrature du cercle!_.pdf (ISBN 9782848841564) (ISSN 2263-4908) Novembre 2013.

Pág. 12 (14)

La quadrature du cercle

Jandaia

Pag. 16 (18) →

La trisection de l'angle

Pag. 19 (20) →

La quadratrice de Dinostrate

—

Dinostrate et la quadrature du cercle ($\sqrt{\pi}$).

—

/Quadrature theory/

Advanced Differential Quadrature Methods ... pdf

Pag. 44 (68) –

2.1 Differential quadrature in the complex plane.

—

/Quadrature theory/

3.375 in 0.02 in Gaussian Quadrature.pdf

—

—

/20210911/classical Mechanics/
1439893691.pdf
↳ Compliant Mechanisms

Pág. 141 (158)
4. Compliances of Curvilinear-Axis Flexible Hinges

—

/20210911/Tangent/
Plane Geometry Practice.pdf

Pág. 214

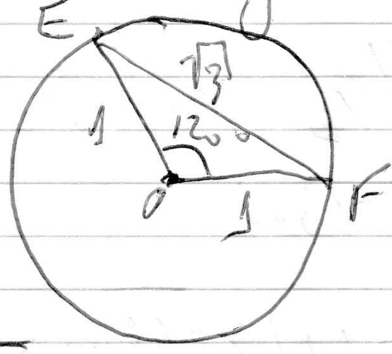

—

/Tangent/Tangent_-_HS74_-2020-04.pdf
→ Courbes et trajectoires

/Tangent/tangent198_-_2021-02_03.pdf
→ Roger Penrose.

Pág. 23(25) Le diagramme de Penrose

Pág. 24(26) La formation d'une singularité.

Pág. 26(28) Les singularités en relativité générale.

Pag. 28(30) L'espace-temps de Gödel.

Important!

Pág. 28(30) La notion de surface piégée

→ surface fermée piégée (Penrose)

Off:

https://v1.archmathsci.org

(Pág. 30(32)).

/20210911/Collisional Effects on Molecular Spectra/

0 128223642.pdf

Pág. 60(74)- Appendix II.D The Liouville space

Here's the formula for the largest number of pieces that can be produced from slicing a doughnut with n cuts.

13 pieces with 3 cuts.

$$\frac{n^3 + 3n^2 + 8n}{6}$$

$$\sin 2x = 2 \sin x \cos x$$

$$\cos 2x = \cos^2 x - \sin^2 x$$

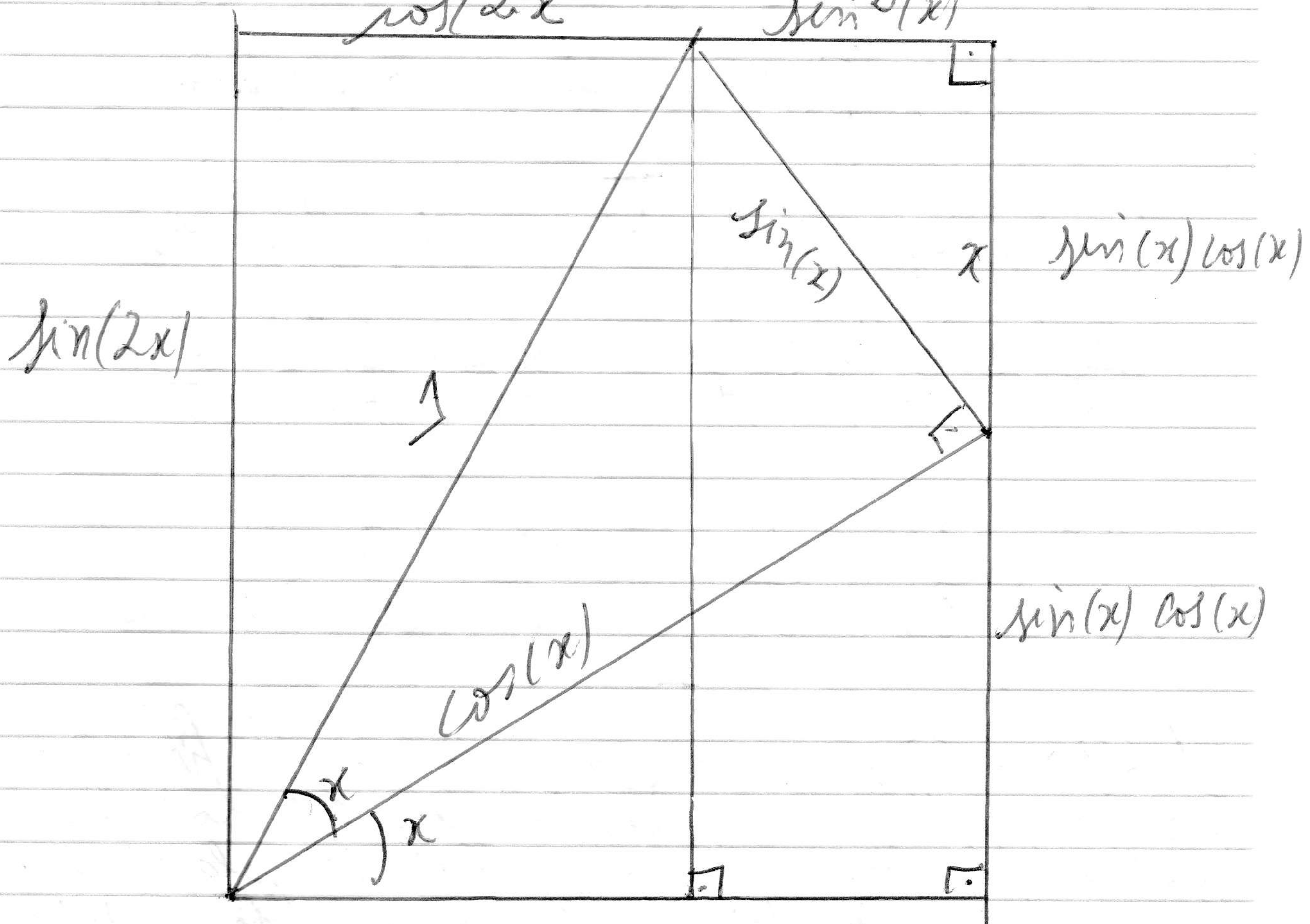

cos(2x)
sin²(x)
sin(x)
x
sin(x) cos(x)
sin(2x)
1
cos(x)
x
x
sin(x) cos(x)

data

Ⓢ Ⓣ Ⓠ Ⓠ Ⓢ Ⓢ Ⓓ

$$\frac{\pi}{2} = \frac{2 \cdot 2 \cdot 4 \cdot 4 \cdot 6 \cdot 6 \cdot 8 \cdot 8 \ldots}{1 \cdot 3 \cdot 3 \cdot 5 \cdot 5 \cdot 7 \cdot 7 \cdot 9 \ldots}$$

$$a^2 + b^3 = c^7$$

this is the only solution in positive integers:

$$15312283^2 + 9262^3 = 113^7$$

"A mathematician is a blind man in a dark room looking for a black cat which isn't there"

Charles Darwin

→ Não concordo, pois é a matemática que fornece a luz que desvenda toda a escuridão das outras áreas.

Stenio Murich.

Andrew Wiles

Andrew Wiles on what it feels like to do mathematics

jandaia

"Perhaps I could best describe my experience of doing mathematics in terms of entering a dark mansion.

One goes into the first room, and it's dark, completely dark.

One stumbles around bumping into the furniture, and gradually, you learn where each piece of furniture is, and finally after six months or so, you find the light switch.

You turn it on, and suddenly, it's all illuminated! You can see exactly where you were".

—

Liouville (property of the divisors)

$$(1+2+2+4)^2 = 81 = 1^3+2^3+2^3+4^3$$

—

Source:

fermat

www.twitter.com/fermatlibrary

—

20210913 ///

Mensch (men)

Books:

Data - Novo Livro - 018 - 2021/
Books - 018 - 2021/
Importantíssimo/

- 1887187308.epub
→ Art of the Golden Ratio
ISBN 978-1-887187-20-6 = Hollar.

- Geometry of quantum states ..... (2006).pdf
(429 pgs) (a)
- Geometry of Quantum States ..... (2017).pdf
(636 pgs) (b)

(a) 94 (106) 4 Complex Projective Spaces.

• (b) 103 (120) 4 Complex Projective Spaces.

"In the house of mathematics there are many mansions and these the most elegant is projective geometry." ↑ Morris Kline.

→ ver pag seguinte. ✡

Andrew Wiles usou a ideia de Kline para explicar o seu modo de entender a matemática.

Morris Kline - Mathematics in Western Culture
(George Allen & Unwin)
1954

- Quantum Mechanics 1-3....pdf (2425 pages) Claude Cohen-Tannoudji.

- Quantum Mechanics, volume 3.... pdf (788 pages)

Importantissimo / Morris Kline /

- Morris Kline - Mathematical thought from Ancient to Modern Times -

1972 - 3 volumes - .djvu.

- Morris Kline - Mathematics in Western Culture

✡ 1964 - 513 pages.

- Morris Kline - Mathematics for the nonmathematician (1985).pdf

- Morris Kline - Calculus... 1998.pdf

Dream of the Golden Ratio and i

$$\left(\frac{\sqrt{2+\varphi}}{2} + \frac{i}{2\varphi}\right)^5 = i$$

/ diegorattaggi

Corredor de Suwalki,
a fronteira de 65 km entre Lituânia e Polônia que separa a Belarus do território russo ocidental de Kaliningrado no Báltico.

=

The Disc Embedding Theorem
Stefan Behrens
www.math.uni-bielefeld.de/~sbehrens/

Source:
www.quantamagazine.org/new-math-book-rescues-landmark-topology-proof-20210909/

—

www.quantamagazine.org/triangulation-conjecture-disproved-20150113/

—

www.quantamagazine.org

—

Books-018-2021 | 20210913 |

- The Disc Embedding Theorem /
  The Disc Embedding Theorem ...
  (2021).pdf

- Der Mathematische ... /
  3662632608.pdf
  Pag. 133(145)
  Abb 3.24 -
  Beweise für die trigonometrischen Additionstheoreme
  (Adição Trigonométrica)

- Calculate the Orbit of Mars! An Observing Challenge and Historical Adventure /
  Clark2021_Book_Calculate the Orbit of Mars.pdf
  Pag. 15(25) Fig. 1.10
  Diagram showing how to manipulate sines and cosines of sums of angles

—

Others:

Books-018-2021/ Importantíssimo/ Morris-Kline/ Morris Kline - Mathematics in Western Culture... (1964).pdf.

Pág. 36(65)

Two right triangles

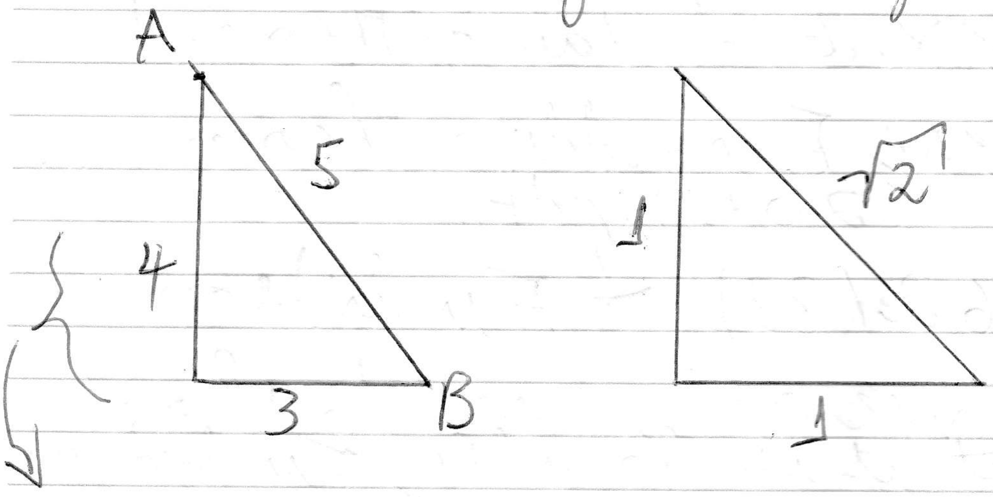

3 →

4 → 3 + ① = 4

↳ 1/3 de 3

5 → 3 + ② = 5

↳ 2/3 de 3

4 → 4 + ① = 5

↳ 1/4 de 4

1

1 + 1/3 = 1,33

1 + 2/3 = 1,66

mantendo a mesma proporção.

$\sqrt{2} = 1,4142\ldots$

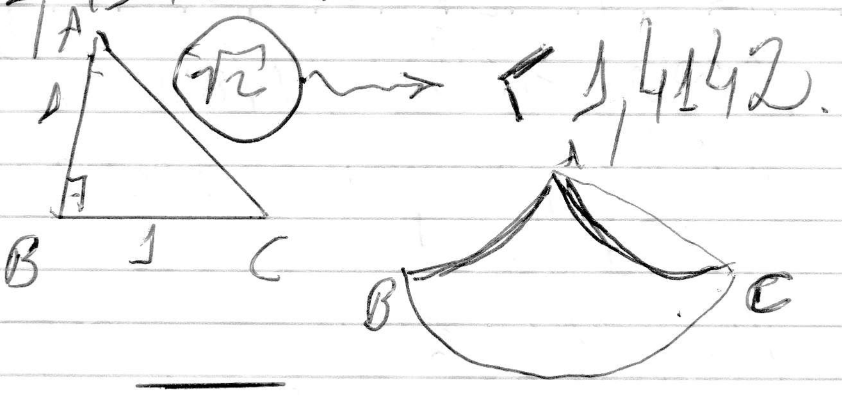

Books - 018-2021 / 20210913 /
The Disc Embedding Theorem /
The Disc Embedding Theorem ... (2021).pdf

Pag. 62(81) → Euclidean.
so we only need to show that it is locally Euclidean.
Let $E := J(S^{d-1} \times [0, \frac{1}{2}])$.

Pag. 92(111) → degree of the Gauss map $S^1 \times S^1 \to S^2$

Pag. 133(152) → Figure 10.1
A modification of the Cantor function. On the right we see that certain vertical segments (red) are created.

Importantíssimo/

- Gauss/
  /Geometry
  /Gravitation Misner
  /Vector.

- Geometry-Topology/

a/ Geometry-Volume-01-Geometry of Lengths ....pdf

b) Geometry-Volume-02-Topology as Fluid Geometry....pdf

Pág. 65(80) Figure 4/
Antoine's Necklace

obs: Analisar com este livro

Books-018-2021/
Importantíssimo/
The Disc Embedding Theorem/
The Disc Embedding Theorem...(2021).pdf

c) Geometry-Volume-03-Two-dimensional Spaces...pdf
Non-Euclidean Geometry

and Curvature

Pág. 68 – 55(68)

Chapter 3 Gravity as Curvature

Obs: Gravitation - Misner

Source: Importantíssimo/
Gauss/
Gravitation Misner/

– Gravitation....(1973).pdf

– Misner-Gravitation-
Princeton University
Press (2017).pdf

d) Lesson in Geometry - Vol 1 ....pdf

Jacques Hadamard

Source: Books-018-2021/
Importantíssimo/
Geometry-Topology/

Jacques-Hadamard-Lecons-
de-Geometrie-elementaire/

Jacques Hadamard-
Leçons de géométrie élémentaire
I.pdf (322pág).
II.pdf (602pág).

e) Importantíssimo/
Geometry-Topology/
Differential Geometry and Mathematical Physics-
PartI ...pdf (773 pág.)
PartII ...pdf (837 pág.)
Pág. 127(142)
2.3 Curvature and Holonomy

- Importantíssimo/
Three-Dimensional/
Three-Dimensional Geometry and Topology, Vol. 1 ...pdf
Author: William P. Thurston
• Pág. 89(99) Figure 2.29
The axis of a three-dimensional isometry.

- Books-018-2021/20210814/
Gauss/Geometry/
Gravity an introduction to Einstein's general relativity.pdf

Other source:

Importantíssimo/
Gravity an Introduction.../
Gravity ....(2013).pdf

—

Books-018-2021/20210914/
Quantum Information Processing
Theory and Implementation/

— 3030754359.pdf
Pág. 2(17) Fig 1.1 Bloch sphere
representation of the qubit
in Eq (1.2)

— 0128219823.pdf
Pág. 4(11) Chapter 1
1.1 Photon Polarization.
Figure 1.1.
(A) linear polarization,
(B) elliptic polarization, and
(C) circular polarization.

Important.

Pag. 06 (13)

Figure 1.2

Representation of polarization state as a point on a Poincaré sphere.

Obs: Use o book:

Books-018-2021/
Importantíssimo/
The Disc Embedding Theorem/

The Disc Embedding Theorem
... (2021).pdf

Pag. 07 (14)

Figure 1.3

The ellipticity and azimuth of the polarization ellipse

Obs: Jones vector parameters

Figure 1.4

The polarizer-analyzer ensemble for the study of photon polarization

20210914/// (20210914///)

- Books-018-2021/Importantíssimo/
Fractional-Trigonometry/
1119139406.pdf

- The Cinderella.2 Manual.../
The Cinderella.2.Manual...pdf

- Jürgen Richter-Gebert-
Perspectives on Projective
Geometry.... (2011).pdf

- Importantíssimo/
Distance-Geometry/
Distance Geometry - Theory,
Methods, and Applications-
(2013).pdf

Pág. 71 (85)

4.2.6 Intersection of a Sphere
with a Dihedron or with
Trihedron.

The sphere has a constant
Gaussian curvature and each
arc of circle has a constante

data
S T Q Q S S D

geodesic curvature.

20210914 (20210915)

mdj-epiphanie.me

M01-20210825

matemáticas-00-20210814

trickle down → gotejar.

Books:

1) Books-018-2021/20210915/

- Multivariate Data Analysis ../
3030769739.pdf
Pág. 27(47) 2.4
Euclidean spaces of matrices and their norms

- Graph Transformation/
10.1007_978-3-030-78946-6.pdf
Pág. 10(22) → Definition 7

M-FPC-augmentations
→ M-morphism such as square(1) in the diagram below.

- Handbook of Financial Econometrics, Mathematics.../

Handbook.....(2020).pdf

2) Books-018-2021/20210915/

- Mathematics of the Arts.../

Bharath Sriraman-Handbook of the Mathematics ... (2021).pdf

pag.711(725)→

- Wasan Geometry
→ important.

3) Books-018-2021/20210915/

/Handbook of Mathematics/

- Handbook of Elliptic and Hyperelliptic Curve...pdf

843 pages-2006

ISBN 978-1-58488-518-4.

Pag. 267 III Arithmetic of Curves
(300)

Isomorphisms
Isogenies
Endomorphisms
Cardinality.

- The Oxford handbook of the history of Mathematics... pdf
- Handbook of Finite Translation Planes...(2007). pdf (894 pages)

Use o arquivo com 6,76 MB

alterando o nome:

Handbook of Finite Translation Planes...(2007)-correto. pdf

- Handbook of pseudo-Riemannian Geometry and Supersymmetry (2010). pdf
- Handbook of Mathematical Geodesy...(2018). pdf

→ Pag 01 (16) — GAUSS

Pag. 10 (25) — Figure 2.2

Lattice points inside a circle.

Figure 2.3 the polyhedral set P_N^2.

Pág. 11(26) – 2.2

Circle problem and Hardy's conjecture.

Pág. 12(27) –

2007, S. Cappell and J. Shaneson →

"Some Problems in Number Theory I: the Circle Problem" in the arXiv.com

math/0702613

claiming to prove the bound of $O(N^{\frac{1}{4}+\epsilon})$ for $\varepsilon > 0$.

→ Table 1.

Incremental improvements to the value ε_2 in the estimate (2.17).

Obs: Ver livro todo!!!

Pag. 28(43) Gaussian quadrature

Pag. 147 (162) → Figure 6.4

A detail of the triangulation of Hannover, carried out under Gauss's supervision 1821-1825

(Collected Works, Vol. 9, p. 347)

⇒ Gaussian curvature

Gauss also was led to the

Gaussian Curvature ⇒

an intrinsic measure of curvature dependent only on how distances are measured on the surface, not on the way it is embedded in space.

Vejà:

Pag. 143 (158) 6.2.

Gaussian conformal mapping of the Earth's surface.

— Strasbourg Master Class on Geometry ... (2012).pdf

Pág. 01(09) – 182(190).

Notes on non-Euclidean geometry

Pág. 54(62) 3.7

Projection in neutral geometry

–

– From Riemann to Differential Geometry and Relativity... (2017). pdf

Pág. 46(80) 8. Topology

Poincaré who is certainly the major founder of the modern field of topology

First letter by Betti, dated October 6, 1863.

Betti then writes:

"What gave Riemann the idea of the cuts was that Gauss defined them to him, talking about other matters, in a private conversation. In this writings one finds that analysis situs, that is, the consideration of quantities independently from their measure, is "wichtig"; ..."

Pág. 101(134) - 3.1 Gauss

—

Pág. 100(133) →

This led Riemann to explain the difference between intrinsic and extrinsic properties of a surface.

He explained that for a sphere, which has an intrinsic geometry different from a plane, the

Gaussian curvature

multiplied by the area of an infinitesimal geodesic triangle is half the excess of the sum of its angles over π.

This allowed him to express the belief that the geometry of an n-dimensional manifold could be understood by understanding its sectional curvatures.

Pág. 142(174) theorem 2.2

If restricted to surfaces of Gauss curvature $K \leq 0$, then the surface is minimal if and only if its Gauss map is conformal.

Pag. 340(367) →

Gauss's Curvature,
Riemann's manifolds,
Dedekind's ideals,
and
Grothendieck's topoi in mathematics,
or
Einstein's spaces curved by gravity in general relativity and
Heisenberg's matrix variables
or
Dirac's spinors in quantum physics.

Pag. 344(371) – In his famous extraordinary theorem (theorema egregium as he called it), Gauss proved that the curvature of a surface which he defined as well, was intrinsic to the surface (surface).

Pag. 543(563) →

The sectional curvature $K(P)$ agrees with the Gauss curvature at P of the surface ruled by geodesic starting at p with tangent vector $X \in P$.

– Eighteen Essays in Non-Euclidean

Geometry (2019).pdf

Pág. 67 (96) — Chapter 5

Inscribing a triangle in a circle in spherical geometry.

Pág. 113 (142) — Chapter 8.

The Gauss-Bonnet theorem and the geometry of surfaces.

Pág. 135 (164) — Chapter 10

Area preserving maps from the sphere to the Euclidean plane.

Pág. 151 (180) — Chapter 11

Area and volume in non-Euclidean geometry

Pág. 235 (264) — Part II

Projective geometries

Pág. 411 (440) — Part III

Other geometries

—

— Geometry in History (2019).pdf

(759 pages)

Pág. 241 (256) – Gauss Curvature

Fermat's principle of least time saying that the path take between two points by a light ray is the path which can be traversed in the least time. (60).

↳ Letter (January 1, 1662) to M. C. de la Chambre.

Employing it, he deduced the law of refraction.

Hence shortest curves on a surface are regarded as something similar to paths traced by light rays (see Sect 6.12)

and it is reasonable to say that a quantity (or a concept) attached to a surface is "intrinsic" if it is derived from distance and angle measured by using shortest curves.

Pág. 337 (352) →

Olinde Rodrigues (1795-1851)

Rodrigues is known for a formula

on this subject carrying his
name, and for

developing the idea of a
representation of a differentiable
surface on a sphere of radius
one which amounts to the
Gauss map.

He used such a (fal)

representation to study the ratio of
the area of (a piece of)
the original surface to the
area of its image on the
sphere.)

⟸

arriving
at a measure of curvature
which in fact is the

Gaussian Curvature

at a point of a surface

Pag. 448 (461) →

Husserl was confusing
intrinsic and extrinsic
natures of space.
He did not see to pay attention
to the fact that spheres and Euclidean
planes can be realised even in hyperbolic
space.

20210915 (cont.)

Books-018-2021/20210815/
/Handbook of mathematics/
– In the tradition of Thurston –
Geometry and Topology (2021).pdf
Pág. 167 (186) 4.4
Post-Thurston Era
4.4.1 Geometrization Conjecture
Perelman (three papers)
The proof was based on
Hamilton's Ricci flow,
which is a flow on the space of
Riemannian metrics on a
manifold satisfying an
evolutional PDE,

$$\frac{d}{dt} g = -2 \, Ric \, g$$

where Ric represents the RICCI
CURVATURE TENSOR.

Pag. 174 (192) → Richard Hamilton

- Geometries and transformations (2018).pdf

Pag. 57 (75) Chapter 3
Circular Geometries

Pag. 58 (76) → A.
Inversive geometry.

A projective hyperplane $\check{U}$, with coordinates $[\![u]\!]$, meets Φ in an inversive $(n-1)$-sphere i.e., a hypersphere of I^n, which is real if $[\![uu]\!] > 0$, degenerate if $[\![uu]\!] = 0$, imaginary if $[\![uu]\!] < 0$, where $[\![uu]\!]$ is the quadratic form associated with the bilinear form

$$[\![uv]\!] = -u_0v_0 + u_1v_1 + \ldots + u_{n+1}v_{n+1}$$

Pag. 71 (89) – Gaussian curvature

↳ → Spherical and pseudospherical geometry.

→ hypersphere of radius r·i in Lorentzian (n+1)-space.

=

Books-018-2021/Importantíssimo/

_SuperPrinciple Relativity_Part_1.pdf

_SuperPrinciple Relativity_Part_2.pdf

Pág. 6 – Ricci Curvature

"Time-Like" – "Inverse Square Distance" Equation

Look: $\left(\frac{8\pi \cdot G}{c^{4}_{light}}\right)$

adequar

_SuperPrinciple Relativity_Part_3.pdf

Pág. 16 1.5

The Gravitational Field Metric Coefficient

$(\Sigma g_{\mu\nu})$,

is a unit-less quantity that can be described as a determinant of a series of independent "metric coefficients", which describes the coordinates or the dimensions of a frame of reference in a localized space, space-time, or gravitational field.

$$\sum g_{\mu\nu} = \begin{pmatrix} g_{11} & g_{12} & g_{13} & g_{14} \\ g_{21} & g_{22} & g_{23} & g_{24} \\ g_{31} & g_{32} & g_{33} & g_{34} \\ g_{41} & g_{42} & g_{43} & g_{44} \end{pmatrix}$$

Pág. 28 1.8

Super Principle Relativity Theory - Hydrodynamic - Einstein Gravitational Field Equations

⇛ 1.78 → Einstein "Dynamic" Gravitational Field Equatio — Local Volume Expansion of

Space-Time Equation —

"Inverse Square Distance" Equation

(1.78)

$$\left(\frac{8\overset{\downarrow}{\pi}\cdot G}{c^4_{light}}\right)\cdot T(A,t)\,\psi \rightarrow 1/m^2$$

—

20210916///

Google Workspace Enterprise Standard

—

maths2.fr/geometrie-projective/

quicklatex.com

www.imo.universite-paris-saclay.fr/~perrin/Livre_de_geometrie_projective.html

—

(.SVG Scalable Vectorial Graphics

Convertio.co

or

svgtopng.com

—

data . .

S T Q Q S S D

Puebla (2017)
Vera Cruz petróglifos.
(México)
↳ Maria.

Xinutetelco
↳ Max Toloyan
(província)
estátua.

sergioandresgarcia.com/pucmm/

geometriaunica.es.wordpress.com/about

Books:

Books-019-2021/20210916/Geometria/

_ [Kiselev_... Geom1.pdf
The Straight Line
Pag. 09 (18).
The Circle Pag. 83/92

_ [Kiselev... Geom2.pdf
Lines and Planes Pag 01 (06)

- Kiselev's Geometry - Book I - rus.pdf
- Kiselev's Geometry - Book II - rus.pdf

Obs: Texto editável.

- trigonometria - gelfand.pdf
- 0.3. Libro.pdf → Geometria (634 pages)
- Libro - G2.pdf → Geometria Analítica Lehmann

Obs. Livro todo!!!

Pág. 273 (291) - Fig. 129

La circunferencia rueda, faz um arco!

⇒ arco PB.

- espacios_vectoriales_y_geometria_analitica.pdf

Pág. 53 - 6

El producto Vectorial y Aplicaciones a la Trigonometria.

- carlos_ivorra_castillo_geometria.pdf (414 pag)

geometria_algebraica
- carlos_ivorra_castillo.pdf (402 pag)

→

geometria-algebraica-de-carlos-
ivorra.pdf

Pag. 301 (311) Capitulo VIII

Integrales abelianos

8.1 Homología y cohomología

(Pág. 304 (314))

— Antonio F. Costa – ejercicios de geometria diferencial de curvas y superficies ... (2005).pdf

(804 pag)

— Important: Pi

— 3. Hist. Form. y Alg. para PI.pdf

Obs.: El Duelo de los Números

Javier Cilleruelo Mateo

(20 pág)

Source:

La Gaceta de la RSME, Vol. 10.1 (2007), Págs. 159-178

– Books-019-2021/ 20210916/
Fisica-Eletricidad y Magnetismo/
6075224904.pdf

Pág. 731(51) 24.3

Pág. 725(45) Capitulo 24
Ley de Gauss.

—

milling - fresamento.

Books-019-2021/Important/
Geometry Illuminated... pdf
Pág. 115(132) - Chapter 12
Parallel Projection
Pág. 116(133) - Theorem 12.3
A parallel projection is both one-to-one and onto.

Obs: Ver livro todo!!!

—

Pág. 149(166) - Chapter 15
Circumference

Figure 1. Shaving a corner off of a polygon.

Applications d'Analyse et de Géometrie
Poncelet (4 vol)

—

www.archive.org/details/
aperuhistoriq00chas

—

Lobachevski
https://archive.org/details/
recherchesgom00lobauoft

—

Géomètrie imaginaire
https://archive.org/details/
gomtrieimagi00lobauoft/
page/584/mode/2up

—

https://archive.org/details/
ljs320/

—

- https://archive.org/details/
questionirriguard00enri

—

https://archive.org/details/
solutionsapproch00reconoft

Import!

books-019-2021/20210918/
Histoire des Mathématiques/

Quadrature of the circle
https://archive.org/details/
quadraturecircl00montgoog

Geometrical Square Root
https://archive.org/details/
geometricalsquar00crab

The Polemic Problem π
https://archive.org/details/
bibliographyonpo00goul

Square of Circle Fuller →

/cu31924063203040

/adoublediscover00fullgoog

/adoublediscover01fullgoog

—

/squarerootofsurd00merc

/secretofcirclesq00nielovich

—

Gaussian quadrature formulas - Stroud A. H.

—

/20210918/
Histoires des Mathématiques/
Gaussian quadrature formulas/
19750024718.pdf

Programs for Computing Abscissas and Weights for Classical and Nonclassical Gaussian Quadrature Formulas

NASA — Robert N. Desmarais

https://encyclopediaofmath.org/
wiki/Gauss_quadrature_formula

—

https://ia804507.us.archive.org/
BookReader/BookReaderImages.
_tu00stro_0019.jp2

—

/24/items/

—

https://ia904507.us.archive.org/
24/items/gaussianquadratu00stro/

or

view-source:https://ia...

—

Disquisitiones Generales
circa
Superficies curvas

Carolo Friderico Gauss

—

Méthode Vectorielle de
Grassmann

Géométrie Infinitésimale

—

Courbes et de surfaces
algébriques
Darboux.

archive.org/details/
sureneclasserema00darbuoft-

—

recherchesgn00gaus

—

Leçons sur la théorie générale
des surfaces et les applications
géométriques du calcul
infinitésimal
4 volumes

https://archive.org/details/
leonssurlath01darbuoft
02

p3 leonssurla00darbuoft
p4

—

leconsurlagendes01darbrich
02
03
04

leonssurlathorie 01 darb.pdf
02
03
04

—

surlatheogensu01darbrich
04

—

Um lugar chamado Notting Hill

—

Books-019-2021/ 20210918/
/Histoire des Mathématiques/

coll. - Histoire de problèmes,
histoire des mathématiques...pdf

others:

Books-019-2021/ Important/

- coll. - Histoire de problemes,
  histoire des mathématiques...pdf
- Commissions inter-IREM -
  Histoire des mathématiques...
  (1993).djvu

Books-019-2021/
20210918/
Ellipse-Editora/

- Michel Carral - Géométrie...(1998).pdf
- Topology...(1998).pdf
- Des fleurs pour Schröndiger
  ....(2009).pdf
- Calcul différentiel ..(1997).pdf
- Convexité dans les espaces

Fonctionnels ... (2004).pdf

- Alain Guichardet - Intégration - Analyse Hilbertienne ... (1998).pdf
- Espaces vectoriels ... (2000).pdf
- Yvan Gozard - Théorie de Galois - ... (1997).pdf
- Dominique Bourn - Traité de Géométrie Affine ... (2012).pdf
- Géométrie Affine, Projective, Euclidienne et Anallagmatique ... (2003).pdf
- Aux origines du calcul infinitésimal ... (1999).pdf
- La notion d'infini ... (2001).pdf
- Books-019-2021/ Books-98-2020/ Important/
  - Dissertação Pappus.pdf
  - Tabak_J_-_geometry_the_language_of_space.pdf
- Pappus-20200622/
- Riemann/
- Supergravity/

—

/Supergravity/

- Supergravity 2012...(1).pdf
  Pág. 12(30) ← 1.2.3
  Spacetime symmetries —
  the Lorentz and Poincaré
  groups

/Infinitesimal/

- Applications_d'analyse_et_de_
  géométrie ... Poncelet ...
  ... 213g.pdf (HAP.1)
- Traitdesprop01poncuoft.pdf
- Traitdesprop02poncuoft.pdf
- A039a423.pdf }
- A039a424.pdf } →

Traité de mécanique
industrielle
2 TOMOS
Poncelet

Books-019-2021/20210918/
Advanced Analytics with Excel 2019/
Advanced Analytics ... pdf

data S T Q Q S S D

- Herbert Meschkowski/
  - Noneuclidean Geometry...(1964).pdf

    Chapter 3 (Pag. 21(25)

    From the History of the Parallel Postulate

- Algebra and Geometry Peirce/
  - Charles... New Elements of Mathematics Volume II...(1976).pdf

    Algebra e Geometry

    Pag. 385 (209) → Chapter IV

    Projection

    Art. 131 - Theorem 23

    Pag. 419(226) - Chapter VI

    The PHIMUS.

Theorem 25 Space contains a quadruply continuous multitude of rays.

Pág. 420(227) - Definition 176

A torus generated by a straight line which constantly cuts three fixed rays is called a phimus (Fig. 63)

- Charles... New Elements of Mathematics Volume III ... 1976.pdf

Pág. 495 (268)   9

Map Projection

Pág. 497 (269) Appendix No. 15

A Quincuncial Projection of the Sphere

Orthomorphic or conform projection formed by transforming the stereographic projection, with a pole at infinity, by means of an elliptic function.

For that purpose, l being the latitude, and θ the longitude, we put —

$$\cos^2\varphi = \frac{\sqrt{1-\cos^2 l\cos^2\theta} - \sin l}{1+\sqrt{1-\cos^2 l\cos^2\theta}},$$

and then $\frac{1}{2}F\varphi$ is the value of one of the rectangular co-ordinates of the point on the new projection.

This is the same as taking —

$\cos \operatorname{am}(x+y\sqrt{-1})$ (angle of mod. $=45^\circ$) $=$ $\tan\frac{p}{2}(\cos\theta+\sin\theta\sqrt{-1})$, where

where x and y are the coordinates on the new projection, p is the north polar distance!

Pág. 507 (274) E.

Mathematical Notes on the Shape of the Earth (1071).

Equations...

Pág. 515 (278) — Varying scale of Skew Mercator.
(enviesamento Mercator)

Fig. 1

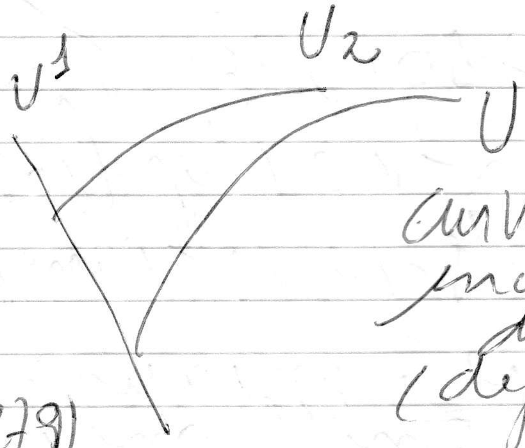

curvas com inclinações diferentes (deformações)

Leste:

Pág. 516 (279)

Hence if rectangular geodesics are drawn between great circles whose northernmost (cujo norte extremo) points are constant, the difference of the squares of the sines of the reduced latitudes of their extremities is constant, or

$\sin u \cdot d \sin u =$

$\sin u_2 \, d \sin u_2$

Pag. 519 (280) G

The Skew Mercator on the Sphere (1584).

Pag. 687 (364) –

Non-Euclidean Geometry

Obs: Analisar tudo!!!

Pag. 710 (376) F.

On two Map-Projection of the Lobatschewskian Plane

Source: manuscript of the Smithsonian Institution.

Pag. 712 (377)

A Conformed Projection

Look: Fig. 2

The Projection is thus established

(hyperbolic stereographic projection).

Pag. 718(380) A Central Projection of the Lobatchewskian Plane
Fig. 3 and Fig. 4

Pag. 720(381) - Fig. 5 and Fig. 6.

—

Books-019-2021/20210920/
Pythagorean-hodograph.../
— Treatise of plane geometry through Geometric Algebra...
Calvet(2001).pdf

Look. Pag. 1

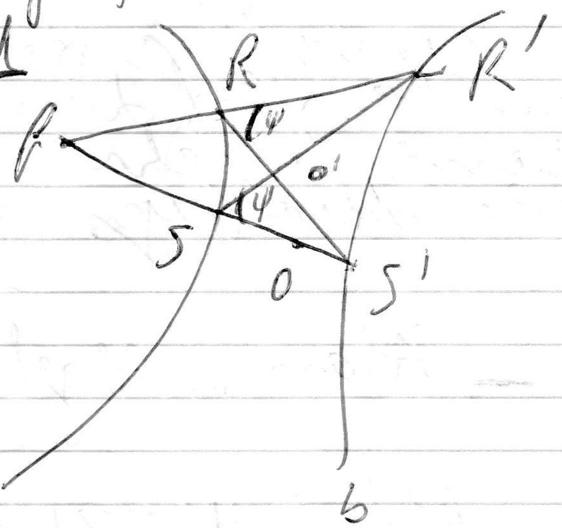

Analizar a curvatura que nos é espelhada a e b,
nos observamos o mesmo ∢. Interessante, utilizá-lo conjuntamente com Pappus (3 pontos)

R O' S' → 1ª reta.
S O' R' → 2ª reta.

Notwithstanding = a despeito de

Pag. 48(62) _ Fig 5.12

The hypothesis states that the lines AA', BB' and CC' intersect at the same point O.

Obs, Interessante!!!

Utilizei em conjunto com a observação anterior.

Pag. 49(63)

— Desargues theorem

We will prove the Desargues theorem in the other direction following the same algebraic way but applying the duality. (Fig 5.13)

— Alan ... Algebra and Geometry (2005).pdf

Pag. 74(88) — 5.1

Spherical distance

$$S=\{x\in\mathbb{R}^3 : \|x\|=1\}$$

Pág. 89 (103) - 6.1 Quaternions and isometries

Isometries of Euclidean space

$\mathbb{R}^3$ is an isometry if it preserves distances, that is, if for all x and y, $\|f(x) - f(y)\| = \|x - y\|$.

Important:

Quaternions ler em conjunto com a pág. 02 do item anterior - Treatise of plane geometry .... Carret (2001).pdf

Hamilton's Elements of Quaternions

Maxwell himself already applied to the electromagnetic field.

Pág. 307 (320) Chapter 15

Hyperbolic geometry

- Linear algebra and Projective Geometry.pdf

Pág. 66 (75) III.3

The second Fundamental Theorem of Projective Geometry

Pág. 69 (78) Appendix II

The Theorem of Pappus

In the second fundamental theorem of Projective Geometry, we have met a first instance where the commutativity of the field F (of coordinates) appears important.

Obs: In the present appendix, we give Hilbert's celebrated characterization of commutativity of F by means geometric configurations.

Fig. 4 — ten coplanar points

Pág. 71 (80) — III. 4.

The Projective Geometry of a Line in Space; Cross Ratios.

Look: Fig. 5 — Pág. 76 (85)

Pág. 85 (94) Proposition 4

Suppose that σ is one-to-one correspondence mapping the points on the line L upon the totality of points on the line $L\sigma$; and suppose that the center of F contains at least three elements

then σ preserves all the cross ratios
if, and only if,
the σ-admissible permutations of
the field F are
semi-automorphisms which
leave invariant every class of
conjugate elements in F

Proof: Pág. 86(95)

Look: Pág. 91(100) Fig. 6

Pág. 131(140) Appendix II
Projective Relations between
Lines Induced by Polarities

Pág. 266(275) VII.3. ↙
The Theorem of Desargues.
(important)

Proposition 1: Assume that the three lines L, L', L'' pass through the point P but are not contained in a plane.

— Books-019-2021 | 20210920 |
Pythagorean-Hodograph curves... |

data S T Q Q S S D

- Cohomological and Geometric Approaches to Rationality Problems.pdf
  Pág. 17(27) –
  The Rationality of the Moduli Space of Curves of Genus 3 After P. Katsylo

  Summary:

  This article is a survey (pesquisa) of P. Katsylo's proof that the moduli space $\mathcal{M}_3$ of smooth projective complex curves of genus 3 is rational. ⟸

- Algebra, Arithmetic, and Geometry - Volume I .. (2009).pdf
  Pág. 69 (104) –
  Analytic Curves in Algebraic Varieties over Number Fields

  → projective line.

Obs. ver se serve!!!

Jandaia

- Algebra, Arithmetic, and Geometry-
Volume II... (2009).pdf
Pag. 23(35)
Cyclic Homology with Coefficients
Pag. 33(45) 3
Gauss-Manin connection.

- Geometry Over Nonclosed Field
(2017).pdf

- Birational Geometry, Rational
Curves, and Arithmetic...(2013).pdf
Pag. 133(141)
Curves of Low Degrees on
Fano Varieties

Comparar:
Geometry over nonclosed field...pdf
Pag. 174(180) 4
Del Pezzo Surfaces

Pezzo surfaces with a view towards
Brauer-Severi varieties.

Look: Pág. 187(193)

8. Del Pezzo Surfaces of
Small Degree.

8.1. Birationality to
Brauer-Severi Surfaces
(Severi)

→ Comparar
com o livro:

Birational Geometry, Rational
Curves, and Arithmetic.(2013).pdf

Pag. 93(101)
Birational Geometry via
Moduli Spaces

and

Pag. 133(141)
Curves of Low Degrees
on
Fano Varieties.

_ A Geometry of Approximation..
(2008).pdf

Obs: Usar a estrutura de pesquisa do livro para diminuir flutuações incongruências geométricas da minha pesquisa.

_ Geometry and Algebra of Multidimensional three-Webs
... (1992).pdf

Pág. XIII (11) Preface

Wilhelm Blaschke e
T. Thomsen

New topic in differential geometry —
WEB GEOMETRY (Web Geometry)

→ Applications of ideas de Felix Klein to differential geometry.

Klein's "Erlangen program" (1872).

According to this program, every geometry studies properties of geometric figure that are invariant under transformations composing a certain group.

From this point of view, the

foundations of affine differential geometry and conformal differential geometry were presented in Blaschke's books, published in the 1920's.

Blaschke extende the same group-theoretic approach to web geometry.

→ pseudogroup

All differentiable transformations of a manifold, where a web is given, as a fundamental group, and studied local web invariants relative to transformation of this group.

Em minha pesquisa, haverá a projeção ortogonal (one-to-one) sobre um plano.

Figure 0.1

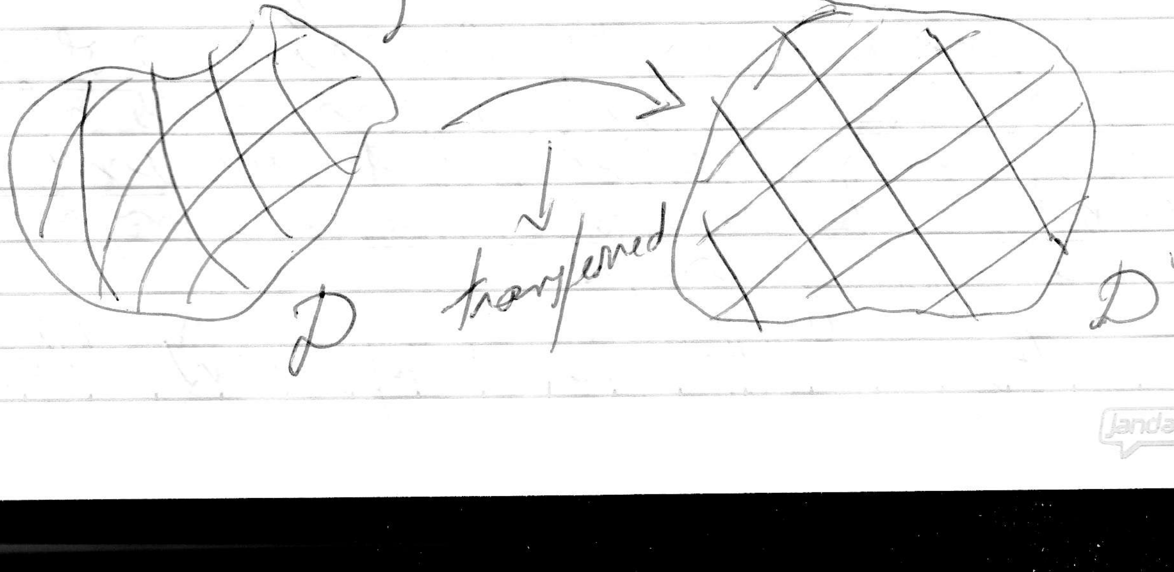

— Introductory Calculus...(1972).pdf

• — Space-Time structure - Algebra and Geometry...(2007).pdf

—

1820 - The Arithmometer (Calculadora)
Patenteada por Thomas de Colmar
Fabricada entre 1851-1915

—

Iridium melting point of 4,435°F (2,466°C).

—

Books:

Books-019-2021/Herbert Meschkowski/
— Ungelöste und unlösbare Probleme der Geometrie...(1960).pdf
Pág. 92(99) VII.

Die Zerlegung von Rechtecken in inkongruente Quadrate
(A decomposição de retângulos em incongruentes quadrados).

• — Geometry of time-spaces non-commutative algebraic geometry...pdf

20210920/ Space-Time/

- Alex's Adventures in Numberland.pdf

Pág. 154 - Life of Pi
until
Pag. 188

Obs: Pag. 184

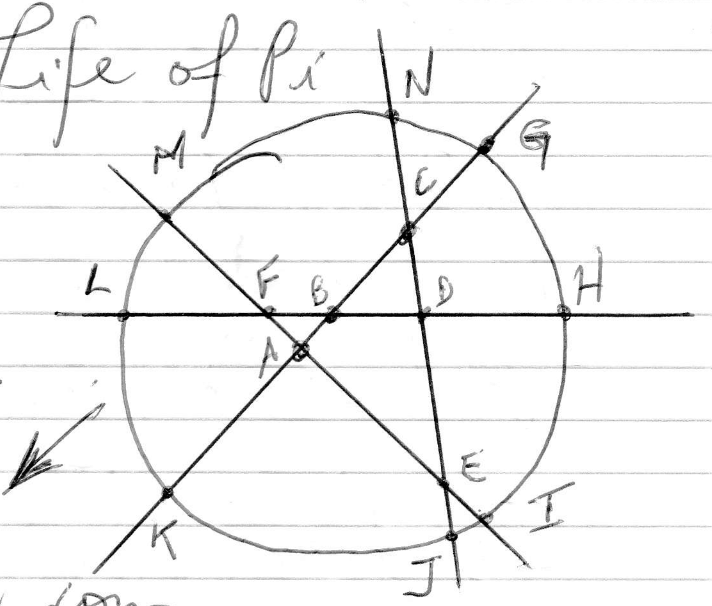

Important:

Use este gráfico conjuntamente com Pappus theorem

Retas → $\overline{ABC}$

$\overline{CDE}$

$\overline{EAF}$

$\overline{FBD}$

Obs: 5 pontos (reta com 5 pontos)

$\overline{KABCG}$

$\overline{NCDEJ}$

$\overline{HDBFL}$

$\overline{IEAFM}$

- The Math Book from Pythagoras to the 57th Dimension ...pdf

- The Physics Books from the Big Bang to Quantum Resurrection ...pdf

- The Physics of Reality Space, Time, Matter, Cosmos... pdf

Pág. 06(33) 3.

the Square Root of Minus One and the Algebra of Iterants.

- Symmetries and Group Theory in Particle Physics ... pdf

Pág. 61(76) - Chapter 4

the Poincaré transformations

→ We examine the general properties of the Poincaré group, i.e., the group of inhomogeneous Lorentz transformations, which are space-time coordinate transformations between any two inertial frames of reference.

Then, we consider those unitary irreducible representations of this group which are the most suitable for the description of the quantum mechanical states of one or more particles.

- Astronautics the Physics of Space Flight...pdf

- Basics of Space Flight... pdf

- Astronautics the Physics of Space flight...pdf

Obs: Ver livro todo!!!

Pág. 292(319) 8.1.4

Tangent Plane Maneuvers

I is the three-dimensional inclination vector

(see Equation 7.3.25)

Pág. 200(228) - Invariant Angular Elements

Fig 7.10 - The graphical representation of orbital elements of an elliptic satellite orbit in a geocentric equatorial coordinate system IJK

(see Sect 13.1.3)

↳ Pág. 666(691)

The Equatorial Coordinate System (IJK).

Pág. 201(229)

The angular momentum unity vector h, also known as the three-dimensional inclination vector (a.k.a. orbit pole vector), give as (see Eq. 13.1.5)

↳ Pág. 670(695)

data

- Airplane Flight Dynamics ...(2003).pdf
- Mathematical Theory of Rocket Flight ...(2008).pdf

= Flight Dynamics Principles ...(2007).pdf

= Non-Rocket Space Launch and Flight ...(2006).pdf

20210921///

Books:

Books-019-2021/20210921/

- Geometry Vector Analysis/
B095Z3GG8M.pdf

Geometry and Vector Analysis

Pág. G-4 (08) →

3 Relation between Cartesian and Polar Coordinates

4. Distance between Two Points

Pág. 09 - 5. Polar Equation of a Straight Line.

Pág. 10 7. Polar Equation of a

Straight Line through Two Points.

Pág. 84 3. General Equation of a Plane

Pág. 115 The Straight Line

Obs: Ver livro todo!!!

- 20210921| Differential Calculus...|

B095Z6555B.pdf

→ Differential Calculus

Pag. 187 Curvature.

- 20210921| Complex Analysis...|

B095Z551R2.pdf

⇒ Complex Analysis.

- 20210921| Several Variables and Partial Differential Equations|

B095YBYMDH.pdf

- 20210921| Linear Algebra...|

B095Y9BGCB.pdf

⇒ Linear Algebra

- 20210921/ Advanced Algebra.../ B095Y9Y7ZV.pdf
- 20210921/ Robert Osserman/ Geometrie des Universums... pdf
- 20210921/ Parallel Coordinates.../ Parallel Coordinates - Visual... pdf

Obs: Todo o livro!!!

Pág. 13 (40) 2.2.2
Axioms for Projective Geometry
Pág. 18 (45)
Figure 2.13
Pappus's theorem (left) and its dual (right).

Obs: Look

Pág. 20 (47) Figure 2.14
A "picture" of the 7-point geometry

Obs: Utilizar Pappus's theorem (3 retas com 3 pontos cada).

AGF, DGC, BGE

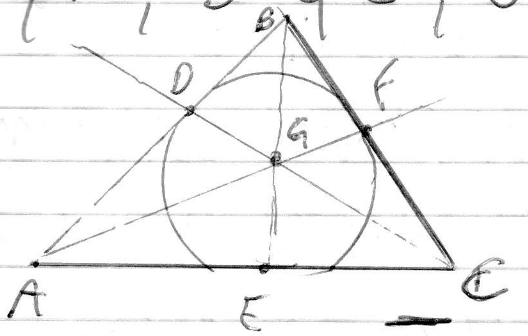

use Pappus's theorem

– Books-019-2021|20210921|
Pappus-Theorem/

– Revolutions of Geometry...pdf
Pág. 08(24) –

$$\pi \approx 3\,\frac{13}{81} = 3.16049\ldots$$

Obs: Livro todo!!!

– Projective Geometry...pdf
(coxeter).

Pág. 26 (38) Chapter 3
The Desargues configuration

Pág. 33 (45) Chapter 4
The Fundamental theorem and
Pappus's Theorem

Pág. 102 (114) Chapter 11 - Parallelism

Pág. 24(36) - Chapter 3
The Principle of Duality

Obs: Pág. 30(42) - Exercises
2 → Locate the trilinear pole of 5 with respect to the triangle ABC. (Figure 3.4A).

Look: Pág. 32(44) - Figure 3.5A
A' B' C' e D' = linha com 4 pontos
(Pappus's theorem)

Obs: Em seguida lia todo o capitulo IV. →
The Fundamental Theorem and Pappus's Theorem

—

- 20210921 | Pappus_Theorem |
A Modern View of Geometry...pdf
Pág. 97(97) - Chapter V
Coordinates in an Affine Plane with Desargues and Pappus Properties →

Affine Plane (incidence geometry)
Satisfy the following axioms:

- Any two distinct points lie on a unique line.
- Each line has at least two points.
- Given any line and any point not on that line, there is a unique line which contains the point and does not meet the given line. (Playfair's axiom)
- There exist three non-collinear points (points not on a single line)

In an affine plane, two lines are called parallel if they are equal or disjoint. (disjoint set)

Using this definition, Playfair's axiom above can be replaced by:

Given a point and a line, there is a unique line which contains the point and is parallel to the line.

Pág. 97 an affine plane is said to have the first Desargues property → parallel.

- Pappus-Theorem/
  Affine and Projective Geometry....pdf
  Pág. 47(64) - Chapter 3
  Desarguesian Affine Plane
  Obs: Pappus's Theorem -
  Theorem 12 of Chapter 1.

Obs: Livro todo!!!

- Pappus-Theorem/
  Perspectives on Projective Geometry....pdf
  (Jürgen Richter-Gebert)

Obs: Important!!!
Livro todo!

- Pappus-Theorem/
  Projective Geometry....pdf
  (Casse)

Obs: Livro todo!!!

Pág. 22(35) - Chapter 3
Introduction to axiomatic geometry.

⟶

Return:

Pág. 20(33) - Theorem 2.11

The Principle of Duality

Example 2.12

Desargues's Theorem + Dual

Pág. 21 (34) → image

By the Principle of Duality, if Desargues's Theorem is valid in the EEP (Extended Euclidean Plane), so is its dual.

Pág. 99 (112) Chapter 5

Coordinatising a projective plane.

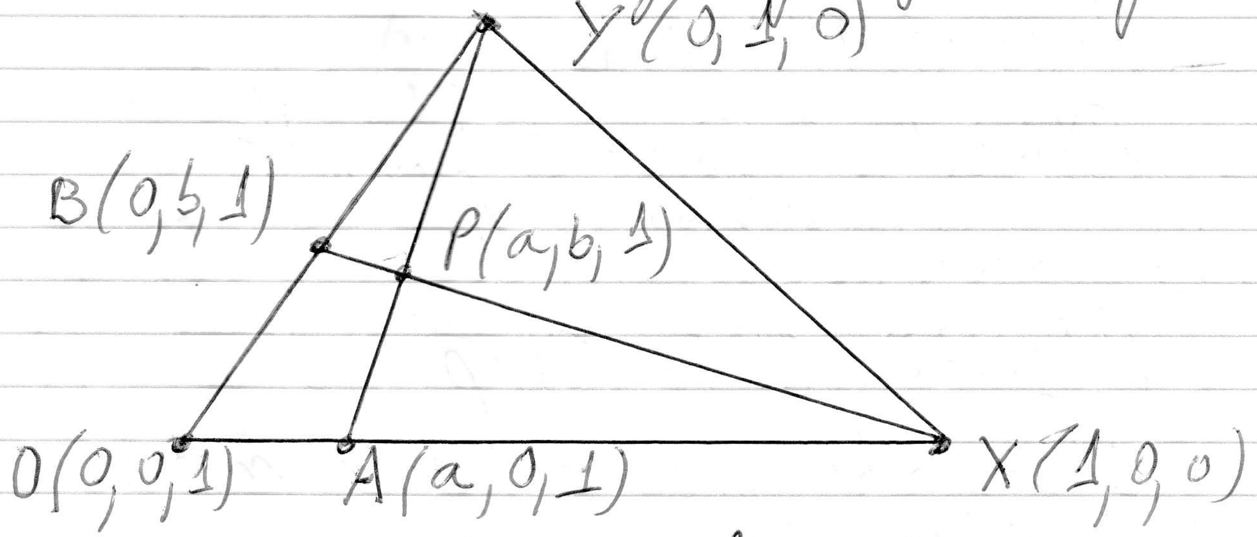

Consider the plane PG(2,F).

— Pappus-theorem/

Classic Geometry Euclidean, Transformational, Inversive, and Projective... pdf

Obs: Livro todo!!!

—

/20210921/ From Great Discoveries in Number Theory.../

Krizek... From Great... pdf

⇒ Pág. 11(25) 1.5

Coprime Numbers

→ micrometer and is called the nonius in his honor.

(Pedro Nunes 1502-1578)

Pág. 12(26) - 1.6

Euclidean Algorithm

—

Important:

Pág. 15(29) - 1.8 Congruence

introduced by the German mathematician Carl Friedrich Gauss.

Gauss → Pág. 53(67) →

The celebrated Law of Quadratic Reciprocity was first proved by Gauss.

Pág. 90(103) –

Until 1796, Fermat numbers were more or less a mathematical curiosity. Interest in them grew when Carl Friedrich Gauss discovered an incredible connection between Fermat primes and the Euclidean construction of regular polygons (i.e., using a compass and straightedge)

Look: theorem 4.12 (Gauss)

Pág. 119(132)

The distribution of the Eisenstein primes in the Gaussian Complex Plane (see Fig 4.11) has an interesting hexagonal symmetry.

Pág. 186(197) theorem 8.5

Every positive integer is the sum of four squares. 1796 + Gauss.

data

Pag. 188(199) Fig. 8.3

Partition of a unit square into two triangles and of a unit cube into 6 tetrahedra.

Pag. 251(260) 10.8

What Mathematics is Hidden behind the main clock?

· Clock ⇒ as a consequence of Gauss theorem (4.12)

Pag. 90(103)

Theorem 4.12 (Gauss)

There exists a Euclidean construction of a regular n-gon if and only if the number of its sides is equal to

$$n = 2^i F_{m_1} F_{m_2} \ldots F_{m_j},$$

where $i \geq 0$, $j \geq 0$, $n \geq 3$ are integers and $F_{m_1}, F_{m_2} \ldots F_{m_j}$ are distinct Fermat primes.

For a proof based on Galois theory see [199, Chapter 16].

↳ 17 lectures on Fermat Numbers: From Number theory to Geometry.

Cont.

Gauss theorem 4.12, we get:

theorem 10.16

Let p be an odd prime.
Then there exists a Euclidean construction by ruler and compass of the regular p-gon if and only if p is of the form

$$p = 2^n + 1.$$

Obs: Usar isto para projeto conjuntamente Coprime Numbers (Pag. 11(25)).

20210921/
From Great Discoveries.../
82058304.pdf
→ Gauss and the Definition of the Plane Concept in Euclidean Elementary Geometry.
July 28, 1797.

Book: Gauss's contribution to Euclidean geometry Krizek
Pág 101-107 (1998)

—

Source: Nº 183
Pág. 317(324) do livro
Krizek... From Great Discoveries... pdf

—

— 20768551_lese-1.pdf
Gauss and Non-Euclidean Geometry

20210921/// (20210921///
cont.

/20210921/Desargues/

= Euclidean Geometry and its Subgeometries... pdf /
Pág. 265(280) Chapter 12
Isometries of a Euclidean plane (ISM).

obs.: Livro todo!!!

Return: Pág. 37(56) - Chapter 2

Affine Geometry:
Incidence with Parallelism (IP).
Obs: Look at Chapter 2 and 3

- Euclidean distance geometry...pdf

Pag. 48 (60) - Fig. 5.5

Reflecting (analisar o que é melhor, reflecting or projection orthogonal!)

Reflecting through ρ^y first and ρ^z second is equivalent to reflecting through ρ^z first and (the reflection of ρ^y through ρ^z) second.

Pag. 49 (61) Fig. 5.6

Action of discretization group $G_D(G) = \langle g_3, g_4, g_5 \rangle$ on a realization $x^1 = (x^1_1, \dots x^1_5)$ is transitive and commutative: ...

→ Usar esta forma de representação de um plano que posteriormente, será curvo e projetado ou refletido sobre uma superfície plana.

Obs: Usar os pontos pretos para fazer triangulação com 3 pontos (Pappus's Theorem)

↓ usar os conceitos de curvar e manter a linha reta estabelecido por Pappus's Theorem.

Look book: /20210921/
Pappus-Theorem/
Perspectives on Projective Geometry...pdf

- 20210921/
Desargues/
Two-dimensional Spaces, Volume 3...pdf

→ Pág. 55(68) - Chapter 3
Gravity As Curvature

Pág. 57(70) - Chapter 4
Curvature by Polyhedral Approximation

data
S T Q Q S S D

Key Idea 4.1

Every smooth surface S in $\mathbb{R}^3$ can be closely approximated locally by polyhedral surfaces S".

Return:

Pág. 55 (68) – Chapter 3

extrinsic curvature

intrinsic curvature

straight (= geodesic)

GAUSS showed us how these problems can be solved for 2-dimensional surfaces.

RIEMANN generalized Gaussian curvature to higher dimensions by considering the Gaussian curvature of each 2-dimensional slice in higher dimensional space.

Question:

How flat is a polyhedral surface?

How straight is an edge path in a polyhedral surface?

Jandaia

Pág. 69 (82) - Chapter 5
Curvature as a length derivative.
Pág. 78 (91) - 5.4
Evaluating the Double Limit
Figure 11 Calculating geodesic curvature.

- 20210921 / Desargues /
Hexagon Theorem of Pappus....pdf
→ Theorems of Desargues and Pappus.

- 20210921 / Desargues /
Geometría Moderna....pdf
Pág. 235 (250) - Capítulo 10
Teorema de Pappus, Pascal y Brianchon.
Teorema 10.1 (Carnot para rectas).

- 20210921 / Desargues /
ch1.pdf → Pappus's theorem:

Nine proofs and three variations

Fig 1.1

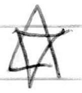

Book: Source

https://www-m10.ma.tum.de/
foswiki/pub/Lehre/WS0809/
GeometrieKalkueleWS0809/
ch1.pdf
ch13.pdf

Other:

- Books-019-2021/20210921/
Desargues/
Math Garden Pappus Theorem.png

Source:

mathgardenblog.blogspot.com/
2013/08/hexagrammum-
mysticum3.html

⇒ **Obs:** Use a imagem que contém
M_1, M_2 e M_3 ⇒ linha roxa.

As imagens estão no diretório:

20210921/Desargues/images/

20210922///

Books-019-2021/20210922/

Machine Design with Cad.../

— Machine Design with Cad....pdf

Pag. 101(133) 4.1.2

- Homogeneous Coordinates Versus Cartesian Coordinates

Pag. 205(237) 5.5.4.3

Gradient Projection Method

Figure 5.24

Projecting the gradient vector ∇f onto the constraints tangent to get ~~a~~ the gradient projection ∇f_p.

Obs: Livro todo!!! olhar!!!

— 20210922/Krishna's Text on Calculus/

B0967YD4TJ.pdf

(T)

Pág D-109(114) 7. Curvature

Pág D-179(185) 10. Singular Points: Curve Tracing

Pág I-121(346) 6. Areas of Curves

Quadrature - The process of finding the area of any bounded portion of a curve is called quadrature.

Pág. I-145(370) 7. Rectification

The process of finding the length of an arc of a curve between two given points.

Obs: Livro todo!!!

— 20210922/Krishna's Text.../

Krishna's Complex Analysis...pdf

Pag C-3(8) 1. Complex Numbers and their Geometrical Representation.

Pag C-101(106) 3. Conformal Mapping

Obs:

Pag C-102(107) → correspondence is one to one.

The corresponding sets of points in the two planes are called images of each other.

_ 20210922/Krishna's.../

Krishnas_TB_Differential_Calculus.pdf

Pág. D-219(224) 11. Curvature (erro na grafia)

obs: Comparar com o outro capítulo do livro anterior (B0967Y04TJ.pdf).

↳ o texto está sem erros de grafia.

_

_ 20210922/The Mathematical Career of Pierre de Fermat.../

0691036667.pdf

Obs: Livro todo!!! Contents não está bem detalhado.

_ 20210922/Landscape of 21 st.../

303080626X.pdf

Pág. 207(217) - Chapter 5 Geometry and Topology

_ 20210922/Infinite Dimensional Analysis.../

3540295860.pdf (717 pág)

Pág. 21(40) - Chapter 2 Topology

Pág. 69(87) - Chapter 3 Metrizable Spaces

Pág. 163(180) - Chapter 5

Topological vector spaces

Pág. 251(268) Chapter 7

Convexity

Pág. 433(448) Chapter 12

Measures and Topology

Pág. 555(568) Chapter 17

Correspondences

A correspondence is a set-value function. That is, a correspondence assigns to each point in one set a set of points in a possibly different set.

Pág. 591(604) Chapter 18

Measurable correspondences

—

Pág. 571(584) 17.6

Vector-valued correspondences

⊙ – convex hull
(convex)

—

_20210921/ Infinite Dimensional...|

Basic Analysis III ... pdf

Pág. 13(24) Chapter 2

Metric Spaces.

Obs: No mesmo diretório há toda a coleção Basic Analysis I às V. ✓

b-z1lib.org_

_20210922/ Infinite Dimension...|

_An Introduction to Metric Spaces...pdf

_Metric Spaces by Satish Shirali...pdf

Pág. 64 (71) Chapter 2

Topology of a Metric Space

_Topology of Metric Spaces....pdf

Pág. 02(15) Figure 1.1

Triangle Inequality.

We refer to d as the absolute value metric.

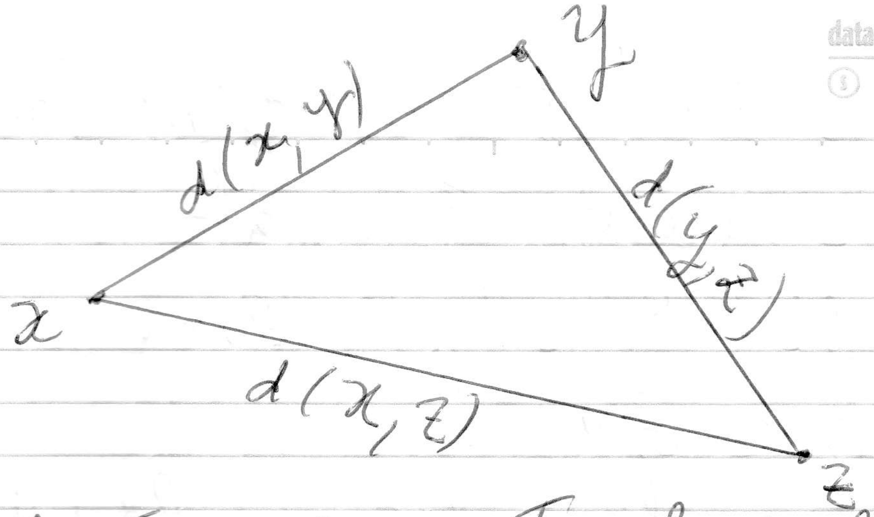

- A Course on Topological Vector Spaces... pdf

- 20210922/ Random Walk, Brownian Motion.../

3030789373.pdf

- 20210922/ Celestial Calculations.../

• Celestial Calculations... MIT Press (2019).pdf

Obs: Ver livro todo!!!

Pág. 53(68) Chapter 4

Orbits and Coordinate Systems.

- 20210922/ Homogeneous Coordinates/

The Role of Nonassociative Algebra in Projective Geometry...pdf

Pág. 01(17) Chapter 1
Affine and Projective Plane
Pág. 08(24) Affine projection. (image)
1.5 Duality
Pág. 23(39) Chapter 3
Coordinates for Projective Planes
Pág. 99 (115) Chapter 7
Projective Geometries
Pág. 131 (147) Chapter 10
Homogeneous Maps
Pág. 181 (197) Chapter 13
Projective Remoteness Planes

_ 20210922/Homogeneous Coordinates/
Methods of Algebraic Geometry in Control Theory. Part I... pdf
Pág. 65(74) 1.2. Affine Algebraic Geometry: Products, Graphs and Projections.

_ Methods of Algebraic Geometry... Part II...pdf

Pág. 105(111) 5

Projective Space

Pág. 113(119) 6

Projective Algebraic Geometry I: Basic Concepts

Pág. 129(135) 7

Projective Algebraic Geometry II: Regular Functions, Local Rings, Morphisms

Velocidade de Rotação dos Planetos do Sistema Solar

Mercúrio 10,83 Km/h - 58d 16h
Vênus 6,52 Km/h - dia 243 dias terrestres 26 min
Terra 1574,00 Km/h - 23h 56m
Marte 866,00 Km/h - 24h 36m
Júpiter 45583,00 Km/h - 10 horas terrestres (9h55m)
Saturno 36.840,00 Km/h - 10h 33m
Urano 14.794,00 Km/h - 17h 14m
Netuno 9.719,00 Km/h - 16 h

Terra - 1665 Km/h = 465 m/s

Terra: 23h 56m 4,09053 sec.

Velocity: 1000 miles per hour (≅)

Circumference of the Earth - 40.070 km

is 1670 Km/h (1037 miles/h).

—

cosine of 45° is 0.707, so the spin speed at 45 degrees is roughly

0.707 x 1037 = 733 mph (1.180 Km/h)

—

Orbital speed 107.000 Km/h
(107.000)

distance from Earth to the Sun —

92,955,807 miles (149,597,870 Km)

circumference of a circle $2\pi r$

940.000.000 Km =

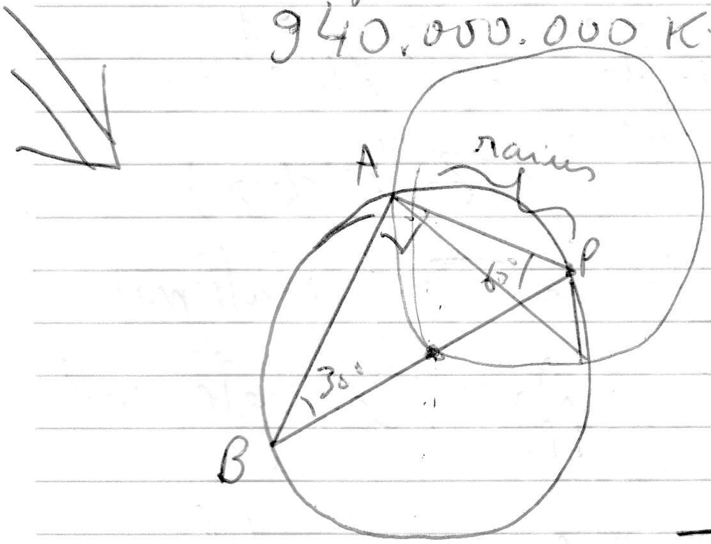

Construct Equilateral Triangle

—

Livros: DataNovoVivo-019-2021 / Books1-019-2021 / Important /

—

20210923   2412-5333.

algoritmo ganancioso.

Books:

Data-NovoVivo-019-2021/
Books-019-2021/
20210923/
Math

_20210923/Analytical Geometry.../

B095PDCGXP-Analytical-Geometry....pdf

_20210923/Math/

475961.pdf

→ Advanced Geometry of Plane Curves and their applications.
(Zwikker)

_20210923/Elements of Mathematics.../

Geometry-Intuitive, Discrete, and Convex....pdf

Pág. 25 - 2. An Axiom for the Set of Parallelohedra

Fig. 2 (b) Developments of Pentadra

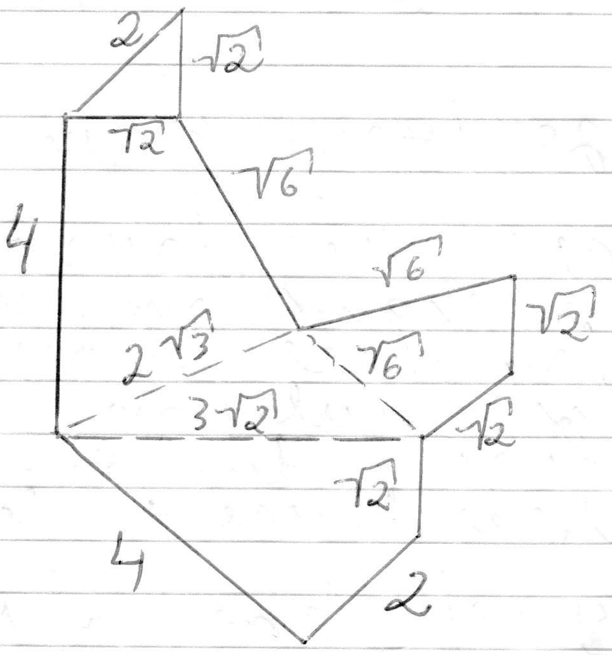

(b)

_ 20210923/ Elements of Mathematics/
New Trends in Intuitive Geometry...
(2018).pdf

Pág. 219 (225) –

The Twelve Spheres Problem

_ 20210923/ Elements of Mathematics.../
Treks into Intuitive Geometry...
(2016).pdf

_ 20210923/ Understanding Gravitational
Waves/ 3030742067.pdf

Pág. 127(139) - GW170817:

Three Detectors and Two Neutron Stars

Pág. 128(140) - Figure 5.3

The narrowing down of the position of a gravitational wave source when three detectors are in use, by triangulation.

The huge reduction in the necessary search are is clear to see.

- 20210923/ Elements of Mathematics/

Toth2021_Book_Elements of Mathematics.pdf

Pág. 207(218) Chapter 5

Real Analytic Plane Geometry

Pág. 224(235) Babylonian Method for $\sqrt{2}$ starting with $q_0 = 1$:

$$5 \quad \frac{886731088897}{627013566048}$$

Pág. 225(236) 5.5

Lines and Circles

translation by W as the map $T_W : \mathbb{R}^2 \to \mathbb{R}^2$ given by
$T_W(P) = P + W$, $P \in \mathbb{R}^2$.

In coordinates, if $W = (u, v) \in \mathbb{R}^2$, then we have

$T_W(P) = P + W = (x+u, y+v)$, $P = (x,y) \in \mathbb{R}^2$.

Pag. 230 (241) →

The two right triangles $\Delta[P, P_0, C]$ and $\Delta[P, P_1, C]$ are congruent since they have two sides, and they also have a common side.[20]

In this example, we assume Birkhoff's Postulates of Angle Measure and Similarity and, consequently, the Pythagorean Theorem, whose validity, in our model, will be proved in Section 5.7.

Pag. 245 (256) 5.7 The Birkhoff Angle Measure.

Pag. 233 (244) Fig. 5.2
An occurrence of the golden number τ.

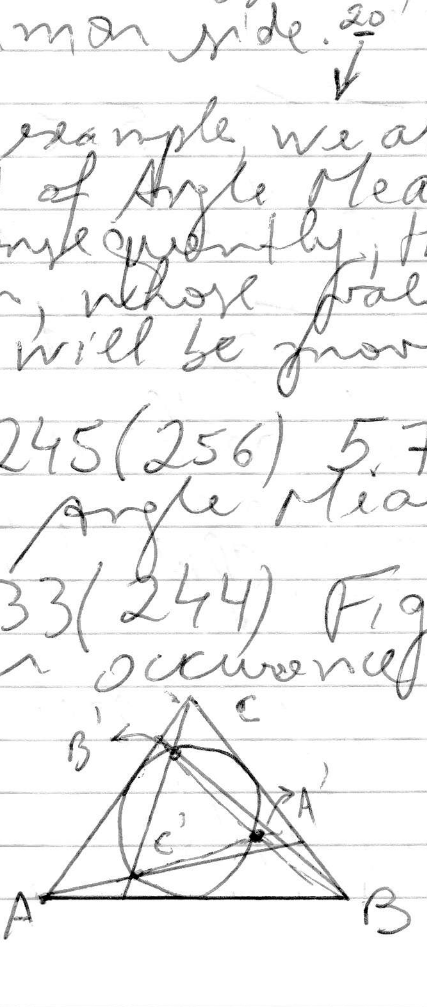

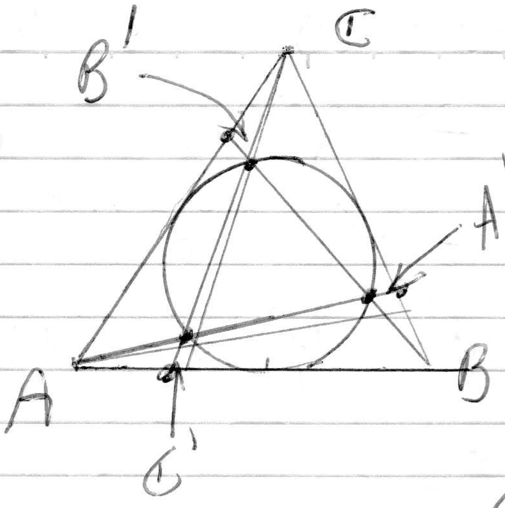

Obs:
Use esta configuração conjuntamente com Pappus's Theorem (three points)

orthogonal to S.

(e) Prove Monge's theorem: Given three disjoint circles (with non-parallel radical axes), there is a circle orthogonal to all three.

Pag. 235 (246) Fig. 5.3
Parametrization of a circular arc.

Pag. 237 (248) Fig. 5.4
Upper bound for the arch length.

Obs: Utilizar com Pappus's theorem (three points)!!

m_j, h_i h_{i-1} (paralelos).

three points Q_{t_i}, T_i, R_i

Pag. 238(249) Fig. 5.5

Illustration for the monotonicity lemma.

→ use three points (Pappus's theorem)

⇒ P_1 P_t P_0.

obs Use esta reta determinada pelos três pontos quando você for fazer a deformação da imagem.

Pag. 239(250)

→ distance from C on the half-line (half-line) with end-point C and containing A and respectively B. We then have the following important formula:[29] ↓

In different (non-axiomatic) development, this formula is equivalent to the so-called Law of Cosines.

Pag. 240(251) – Lipschitz condition with Lipschitz constant

$d/\sqrt{1-d^2/4}$!

IMPORTANT.

Fig. 5.6

The sum of angles in a triangle is π.

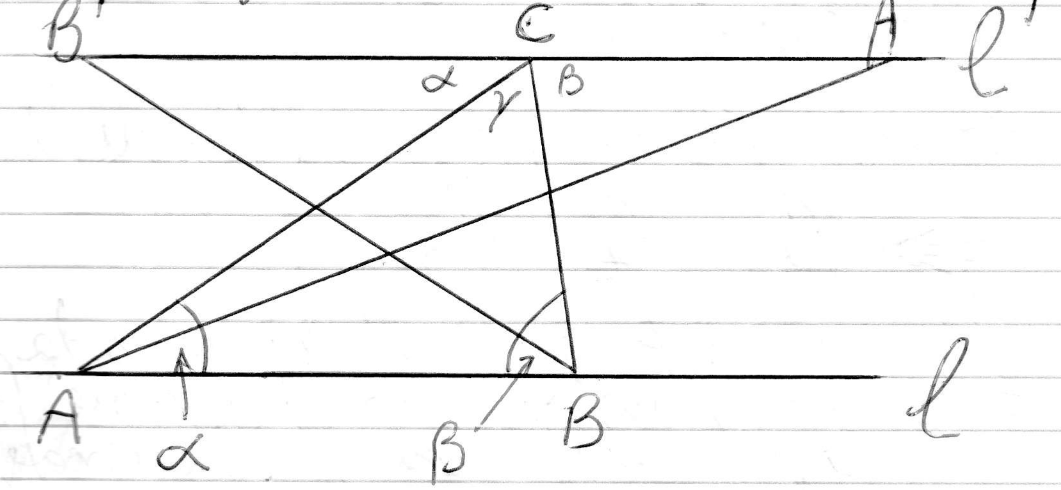

surjective
injective

$$\alpha_0(l_1) - \alpha_0(l_0) = \mu(\angle l_0 O l_1) \pmod{2\pi}.$$

Pág. 246(257) → interessante página.

$$\alpha + \beta + \gamma = \pi$$

Pág. 247(258) Fig. 5.7
Proof of Birkhoff's Postulate of Similarity.

Pág. 252(263) – 5.8
The Principle of Shortest Distance

Pág. 253(264) – rectifiable curve

—

if there exists a one-to-one ↙ 35

35 the property of being simple, that is, one-to-one, excludes "self-intersections".
As we consider here only open curves and minimize the arc length, imposing this does not restrict the generality.

Pag. 257(268) – 5.9

π According the Archimedes*

$3\frac{10}{71} < \pi \; 3\frac{1}{7}$ ///

—

20210923/// (20210923)

— 20210923/ Birkhoff/

Groups, Matrices, and Vector Spaces...pdf

Pag. 211(221) – 7.3

Isometries and Orthogonal Mapping

The term orthogonal comes from the fact that an orthogonal mapping preserves the orthogonal relationship between orthogonal pairs of vectors.

Since orthogonal mapping preserve
inner products, they preserve the
lengths of vectors, distance between
vectors, and angles between vectors,
since the angle θ between

v and w is found by the
identity

$v \cdot w = |v| |w| \cos\theta$.

It turns out that all orthogonal
mapping are linear.

Pág. 135 (145) Chapter 6
Vector Spaces

Pág. 207 (217) Chapter 7
Linear Mapping

20210923 (20210923)

— 20210923 / Mathematical thought
from Ancient to Modern
times - 3 volumes /

Morris Kline - Mathematical...pdf

— 20210923 / A History of Geometrical
Methods /
Julian... A History of Geometrical... djvu

20210924 ///

Books-019-2021 | 20210924 /

Krishna's Spherical Astronomy and Space Dynamics /

- B08... Space-Dynamics.pdf
- B08... Spherical Astronomy... ~~pdf~~
  and Space Dynamics fco pdf

estudar conjuntamente !!!

- B08... Discrete Structures Graph Theory.pdf
- B09... Differential Equations Laplace Transforms.pdf
  (B09)
- B09... Advanced Calculus.pdf
- B09... Linear Algebra.pdf

Pág. 171(76) Chapter 4

Quotient Space and Direct Sum of Vector Spaces

1. Homomorphism of Vector Spaces or Linear Transformation

Let U(F) and V(F) be two vector spaces. Then a mapping f: U→V is

Jandaia

called a homomorphism or a linear transformation of U into V,

if (i) $f(\alpha+\beta)=f(\alpha)+f(\beta), \forall \alpha, \beta \in U$ and

(ii) $f(a\alpha)=af(\alpha) \; \forall a \in F, \forall \alpha \in U.$

- 2021 09/24

Hilbert and Banach Space.../

Hilbert and Banach Space....pdf

Pág. 79 (92) 3.2 Orthogonally scattered dilations

Pág. 92 (105) 3.3 Gramian orthogonally scattered dilations

- 2021 0924/
Finite Element Analysis Applications/

0128099526.pdf

Pág. XIII (3)

Finite element analysis (FEA)

Pág. 32(36) Chapter 2

Decomposition

Functional Requirements → FRs

Pág. 33(37) Figure 2.2
Complex of a finite element analysis model.

→ Geometric shapes,
Physics behaviors,
Dynamics,
Uncertainties,
Assemblies.

Pág. 39(42) 2.3.5 Geometric Shapes.

Pág. 51(55) Decomposition
Table 2.2 Classification of Meshes

Pág. 53(57) Table 2.3
Definition of Aspect Ratios of Different Element Types.

Pág. 54(58) Figure 2.19
Example of Jacobian check in quadratic-triangular element

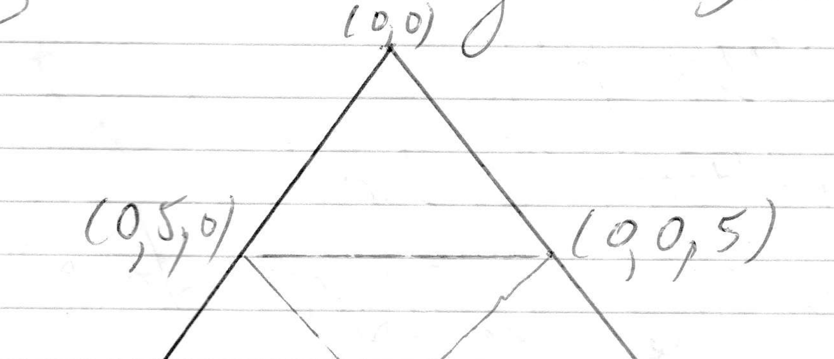

Important:

Pág. 41(45) 2.5 Axiomatic Design Theory. (ADT)

ADT uses two basic axioms as several other corollaries and theorems, to support the design theory.

First axiom is to maintaining the independence of FRs and the

Second axiom is to minimize the information content of the resulting design solution.

The act of design consists of the mapping of FRs into design parameters (DPs), and eventually to arrive at a solution, which satisfies these two axioms (axioms).

20210924/Physics of Nuclear Reactors/
01282241X.pdf
Pág. 34(44) Fig 1.21

⇒ Parallelogram representation of the velocity vectors.

transformations using the

parallelogram law of vectors,
as given in the Fig. 1 2 1

By vector addition we get

$$v_L'^2 = V_{CM}^2 + v_c'^2 + 2(V_{CM})(v_c')\cos\varphi$$

$$v_L'^2 = \left(\frac{v_L}{A+1}\right)^2 + \left(\frac{A v_L}{A+1}\right)^2 + 2\left(\frac{v_L}{A+1}\right)\left(\frac{A v_L}{A+1}\right)\cos\varphi$$

$$v_L'^2 = v_L^2 \, \frac{A^2 + 2A\cos\varphi + 1}{(A+1)^2}$$

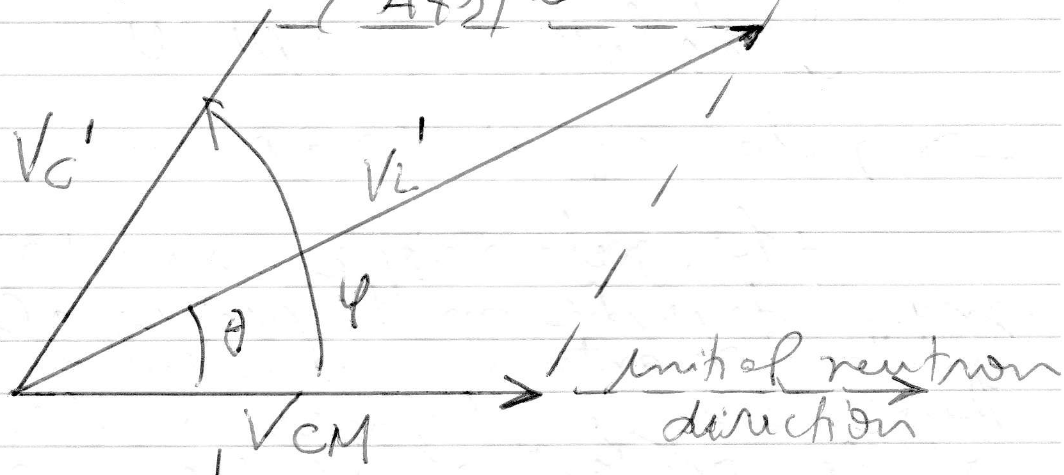
V_c'
V_L'
φ
θ
V_{CM}
initial neutron direction

20210924 | Quantum Mechanics in Potential... |
Quantum Mechanics in Potential...pdf
Pag. 29(50) Chapter 3
Nuclear Forces

→ the Interactions between Nucleons.

Gluons and quarks are confined in a limited space, and the forces in nuclei are produced due to the interchange of mesons whose field is similar to the electromagnetic field, yet is of a much shorter range.

⇑ gravity.

✡

([illegible] [illegible] forces).

Pag. 183 (204) Chapter 19

Stability of Nuclei

→ relativistic corrections

The relativistic corrections of the mass depende on the principal and the orbital quantum numbers and do not have any influence on the definition of the constant of the spin-orbit interaction.

– 20210924/
Solving Linear PDE Spectra/
Solving Linear PDE Spectra.pdf
Pag. 349 (366) – Chapter 11
Quantum Mechanics

=

data
Ⓢ Ⓣ Ⓠ Ⓠ Ⓢ Ⓢ Ⓓ

20210924 (Conti.)//

20210924/ A History of Greek Mathematics.../

- Heath... A History of Greek mathematics vol. I..pdf
- Heath... A History of Greek mathematics vol. II..pdf

20210924/ A History of Greek.../

- The Shaping of Deduction in Greek Mathematics...pdf

Pág. 24(43) Figure 1.5

Appolonius's Conics I. II.

Utilizá-los conjuntamente!!!

—

Cod. aax-vsak-bsx (meet-google)
(17h22min) Carl Munch

— —

20210925///

Kurt Gödel - Obras Completas
R$150,00 —

Book:

- 20210925/ Shape the Hidden Geometry.../

jandaia

↘ Shape....pdf.epub

Pag. 173 – Cremona transformation is "a birational automorphism of projective space".

What is the ratio of the area of a size-zero square to its perimeter?

If the square has a side of length 1, that radio is 1/4 or 0.25.

When the side length shrinks to 1/2, the area is 1/4 and the perimeter is 2, so the ratio has gone down to 1/8.

side length 0.1, ratio is 0.01/0.4 or 0.025

... 0/0 = 0

—

– 20210925/ The Theory of Complex Angular Momenta/

0521818346.pdf

Pág. 11(25) 1.2.1

→ The Mandelstam plane

We use here the well know geometrical fact that the sum of the distances from

a point on the plane to the sides of an equilateral plane does not depend on the position of the point.

$1 + t + u = 4u^2$

↳ distances to the sides of the triangle.

_20210925/ Real Analysis... /

B095... Real-Analysis.pdf

Pág. R-341(346) - Chapter 7

The Riemann Integral

Pág. R-389(394) - Chapter 8

The Riemann-Stieltjes Integral

_20210925/ Complex Analysis... /

B095-... Complex Analysis Calculus of Variations.pdf

Pág. C-3(8) - Chapter 1

Analytic Functions

⇒ 1. Curves in the Argand Plane

$z = z(t) = x(t) + iy(t)$ where

$z = x + iy$.

—20210925/ Quantum Atom...|

1108838596-Quantum-Atom....pdf

Pag. 100(115) — Fig 6.1

d sen θ.

the diffraction of light by a grating.

—

malavasi

—

0,01/0,4 = 0,025

—

20210926M

squeaky = estridente

haste make waste =

pressa faz desperdício

Books:

20210926/Modern Physics...|

Modern_physics_6e.pdf

Pag. 17(35) 1.3

The Lorentz Transformation

Pag. 21(39) Relativistic Velocity Transformation

Pag. 23(41) Spacetime Diagrams

Pag. 27(45) Figure 1-22

Pag. 29(47) 1.4

Time Dilation and Length Contraction

Pag. 97(115) 2.5 General Relativity

→ Pag. 103(121) Deflection of Light in a Gravitational Field.

Pag. 104(122) Figure 2-20

Deflection of a beam of starlight due to the gravitational attraction of the SUN.

$\alpha = 1.75$ arc second^{18} Einstein Eddington

Pag. 156(174) 4-2

Rydberg-Ritz formula

$$\frac{1}{\lambda_{mn}} = R\left(\frac{1}{m^2} - \frac{1}{n^2}\right) \text{ for } n > m$$

_ 20210926 / Numerical Geometry ... /
3030767973.pdf.

Pág. 13 (29) - Manifolds of Triangulations.

Pág. 37 (53) 1. The GAUSS integral for the Linking Numbers of Curves. /

Pág. 39 (55) 2 - the outline of the Invariant-Based Formula and Consequences

theorem 1 (Properties of the Linking Number).

Fig. 1 → interessante.

Stephen Gray ⇒ descobriu a condutividade da eletricidade
(1666 - 1736)

Madras ⇒ hoje Chennai (Índia)

Suástica no Tibete = "yungdrung".

Ernst Schafer 1939.

livro: Vaibhav Purandare (author).

Canon 85 mm f 1.2 FD SSC
Aspherical US$ 23,000.00

otani.camera.tokyo

—

55 13 9742 51172 CAIYÀ
mudou endereço

—

Books:

20210827/Advanced Engineering Mathematics/

→ todo diretório !!!

— Differential Equations... (2020).pdf

20210927/// (20210927).cont

20210927/Riemann-Geometry/
Revolutions of Geometry...pdf
Pág. 186(202) Area and Circumference
Pág. 236(252) 7.2 three Problems
Square the Circle ($\sqrt{\pi}$)
Pág. 238(254) Duplicate the Cube
Viète solved the problems of

duplicating the cube by again finding inspiration from seventh book the Mathematical Collection of Pappus.

Pág. 243(259)

Trisect an Angle

The third postulate allows Viète to trisect an arbitrary angle.

20210928///

www.reschio.com/
luxury-villas-umbria

Fulvio Ricci

fricci@sns.it

Books: (Estes dois livros conjuntamente).

20210928/ Geometric Aspects of Harmonic Analysis/

2021_Book_GeometricAspects...pdf

Pág. 12(24) 5. Trace Hardy and Hardy Inequalities vector fields

Pág. 467 (476)

~~On Regularity and Irregularity of Certain Holomorphic Singular Integral Operators.~~

Cauchy-Szegõ projection

2021 09 28/ Fulvio-Ricci/

hardy.pdf

Pág. 3 Chapter I

Hardy Spaces on the Unit Disc

Pág. 26 6. The Cauchy Projection

Pág. 38 8. Dual Spaces

the Cauchy projection

Pág. 56 4. The Cauchy Projection and the Hilbert Transform.

↓

Since the orthogonal projection of $L^2(\mathbb{R})$ onto $\mathcal{L}^2_+$ is obviously the map $g \mapsto g\chi_{\mathbb{R}_+}$, ...

principalmente, não interferir neles!

Confúcio: "Quando é óbvio que os objetivos não podem mais ser alcançados, não ajuste as metas, mas sim as etapas da ação!"

Continuar

HD01-20210827/ ~~2021~~/
2-GERAL/
2-NOVO/
2021/
1-Vídeos-2021/
Vídeos-20210707.7z

e

HD-01/
2021/
Books-2021/

Books:

Books-020-2021/20211102/
Singularities in Physics and Engineering/
978-0-7503-1698-9.pdf

Pag. 7-1 (173) Chapter 7
Angular momentum of light.

20111002/ Introductory Quantum Physics...|

Introductory_Quantum_Physics_and_Relativity...t.pdf.

Obs: Livro todo!!!

Pág. 187(200)- Chapter 10

Fine Structure of the Hydrogen Atom.

Aplicam a deformação do tempo-espaço

Obs: Look at equation: $\frac{(10.7)}{(10.6)}$

(10.6)

$$\langle V \rangle = 4\pi \int_0^\infty r^2 \left(\frac{-e^2}{4\pi\epsilon_0 r} \right) \frac{1}{\pi a_0^3} e^{-2r/a_0}\, dr$$

etc

O rio que esquece onde nasce, seca.

University of Rijeka
Department of Mathematics
Ulica Radmile Matejčić 2, Rijeka
Rijeka, Primorje-Gorski County, Croatia

unipu.hr/en

Juraj Dobrila University of Pula

www.airport-pula.hr

www.veniceairport.it/en/

www.lju-airport.si/en/Main

Endeup!

Dani Zdravlja

20211003 M

50
75
125
25
150
100
100
50
400
100
50
550
100
650

110
150
105
120
485
110
595
70
660
=

120
40
30
190
120
50
360
110
130
600
100
700

120
110
120
120
470
60
530
110
65
705

90
120
100
120
430
100
120
650
=

140
60
200
130
50
380
110
160
650
40
690

180
90
270
220
490
60
540
//

20211003 (cont)

3-Aberto (Zippando)

Books	45
Data	130
SV	
Videos	
Zip	

75
120
45

Copiar 3-Aberto → p/
(amanhã) 3-Aberto-copy

/2021/DATA/

Cantina do Marinheiro

Mem de Sá 125 Mooca

Infinite Potential film (2020)

Quantum Theory Bohm

The Quantum Dissidents

Cosmologia Quântica - um ramo da Física que tenta combinar Teoria Quântica com Gravitação, ou seja, Teoria Quântica com Teoria da Relatividade.

Efeitos Quânticos (Bohm) →

um deles Efeito Aharonov-Bohm

Emaranhamento Quântico e
Efeito Aharonov-Bohm

Einstein → Emaranhamento, ele chamava
"spooky action at a distance"

—

Bohm → outros Conceitos

a) Potencial Quântico
b) Totalidade
c) Ordem Implicada.

a) "Trata-se de uma função matemática que conecta instantaneamente partículas muito distantes"!

Tanto a Mecânica Quântica quanto o Potencial Quântico estavam sugerindo uma ideia de Totalidade.

Livro: Totalidade e o Universo Indiviso
(Bohm e Basil Hiley)

Niels Bohr estava certo, você não pode separar o observador do observado.

David Bohm introduziu o

conceito de Ordem –

a/ Ordem Implicada ou Implícita

b/ Ordem Explicada ou Explícita.

David Bohm e Jiddu Krishnamurti

Livro: ↙

"O observador é o observado!"

esposa de Bohm encontrou este livro. Int uma livraria.

—

3 - Aberto/2021/Data (copy)

—

20211004

ZIPPAR: Videos-Geral

40	150	100	100	✓VSV
110	50	150	100	✓Ungar
120	100	150	100	✓Videos
40	300	150	100	✓Videos-Import
310	130	550	100	✓Videos-Math
230	50	100	500	✓Videos-others
540	480	650		✓Videos-por-data
60	150			
600	630			
100	50			
700	680			

—

Books:

① Books-020-2021/20211004/
Thomas' Calculus.../
Thomas_Calculus...pdf
~~Pág. 740(765) Chapter 15~~
~~Differentiation~~
Pág. 388(389) Chapter 7
Transcendental Function

② 20211004/Mathematics for Engineering.../
Mathematics for Engineers...pdf
Pág. 740(765) Chapter 15
Differentiation

③ 20211004/Homotopy Theory.../
3030789764.pdf
→ Homotopy Theory and Arithmetic Geometry – Motivic and Diophantine Aspects.

④ 20211004/Handbook of Measurement...|
Handbook of Measurement...pdf

⑤ 20211004/Geometry and Vision/
10.1007_978-3-030-72073-5.pdf

→ Geometry and Vision
Pág. 02(1/5) 2. Related Work
2.1 Gaussian Noise Generating

⑥ 20211004/Engineering Electromag-
netics and Waves/
Engineering Electromagnetics and
Waves.pdf
Pág. 400(401) Example 5.3
① Curved bar
Figure 5.6 A curved-bar
resistor.

Livro: Gaussian Surfaces
(procurar em todo o livro, há várias
referências)

⑦ 20211004/ Nonlinear Time Series.../
Nonlinear Time Series Analysis with R....pdf
Pág. 111 (122) 4.5
Poincaré Map.

Komorač
funcho =
Koromač

20211005///

AD-03-20210827/
3-Aberto/
Videos-Geral/
Videos-por-data/
2020/
08/

20200811/ → 24 Gb
20200813-20/ 28
22-31 52
45
97 Gb

09/
04-06 → 30 Gb
07-09 → 29 Gb
10-15 → 30 Gb
16-20 → 21 Gb
22-27 → 29 Gb
28-30 →

bluntly
put
colocado
sem
rodeios

Komorač
(Foeniculum
vulgare)

Kush
(erva
medicinal)

Differential Equations, and Mathematical Physics.

— 9811654409.pdf

Equations and Analytical Tools in Mathematical Physics

Legendre - Gauss quadrature

Pág. 225 (231).

Pág. 231 (236)

→ Gauss hypergeometric functions

— Equations of Mathematical Physics.pdf

Pág. 114 (125)

Gauss's theorem in an infinitesimal volume centered at R=0 and enclosed by the surface S,

Pág. 216 (227) 7.7

Maxwell Equations

→ Gauss's law

The first two equations are the Gauss's laws, the third equation is the Faraday's law, the

fourth equation is the
Maxwell-Ampere equation.

- Exact Solutions of Einstein's
Field Equations - Cambridge
Univers. pdf
Pág. 293 (325)
19.2 Orthogonal Surfaces
Pág. 298 (330)
19.4 The field equations for
stationary axisymmetric
Einstein-Maxwell fields
→ Gaussian curvature of the
2-space $\hat{V}_2$ with metric γ_{MN}.
Pág. 424 (456)
Gaussian curvature K of the
2-surface $2d\zeta\, d\bar{\zeta}/P^2$ could be
normalized to $K = 0, \pm 1$.

Pág. 425 (457)

Gaussian curvature K

Pág. 472 (504)

Gaussian curvature

Obs: Look: 31.3

The Ricci tensor components.

Pág. 523 (555)

Gaussian curvatures of the corresponding 2-spaces.

Pág. 525 (557)

31.13 Complex invariance transformations

→ Gaussian curvature and are related by the complex transformation

Pág. 572 (604) -

the Gauss and Codazzi equations in the form...

Pág. 583 (615) →

37.3 Some theorems on local isometric embedding

Gauss-Codazzi-Ricci equations. etc

obs: Gaussian Curvature Pág 101

- Partial Differential Equations...pdf

Pág. 94 (107)

Gaussian is constant.

Pág. 285 (298)

Gauss hypergeometric series

Pág. 315 (328)

Logarithmic Gauss map.

Pág. 323 (336) Figure 1.2.

Truncated factorization for Δ_4^{rd}.

- Symmetries and Conservation Laws for Differential Equations of Mathematical Ph..pdf

Pág. 185(195)/ Chapter 5
Conservation Laws
Gauss theorem.

- The General theory of Particle Mechanics...pdf

- Transmutations-Singular-and-Fractional-Differential-Equations.pdf
Pág. 06(32)-
1.1.3 Hypergeometric type functions
The hypergeometric Gauss function inside the circle $|z|<1$
Pág. 50(76)-
Gaussian curvature
Pág. 227(252)- Definition 40
Gauss-Buschman-Erdélyi operators by formulas...

→ Riemann
- 20211006/ Electricity Maxwell/ Algebraic geometry...pdf
- 20211006/ Elect.. Maxwell/ Thomas.. All the Math you missed(2022).pdf
Riemann-Roch theorem for smooth curves - Pag. 369(380)
(veja os 2 arquivos) chapter 6 - Sheaves and cohomology.

20211007

110 120 80 310 (48,72)	200 220 210 630 70 700 Books-2021 05 - 42,7	140 130 10 280 70 350 90 440	440 50 490 Books 2021 (24,93)	490 130 620 70 690 Books 2021-06 (40,84)	240 230 170 640 100 740 Images- 20200621 (1,95)	210 200 220 630 110 740 Books 2020 20-21 (17,13)	200 210 200 610 40 650 2021-09 20210710 (59,04)
260 50 310 300 80 690	50 210 250 100 610 110 720 (42)	102-2020 (28,83) DATA					

"A liberdade não se perde de uma vez, mas em fatias, como se corta um salame."

Friedrich Hayek

"Se quiser saber quem controla você, é só observar a quem você não pode criticar."

Voltaire

"... porque o Senhor cuida dos justos e ouve as suas orações, mas está contra os que fazem o mal."

1 Pedro 3,12.

110	50	014-2021	100	200	210	100	91-2020=(7,65)	
105	50	20210714 (24,49)	105	110	220	115	~~100~~	
100	100	130	200	210	150	105	91-~~100~~	
100	200	120	220	110	580	120	320	590
105	110	130	625	630	110	120	110	//
100	120	380	Books	007-2021	690	560	420	121-2020
800	105	240	2021 10 20210817	01-20210326	104-2020	110	100	(7,39)
162	635	620	(44,89)	(17,34)	(25,25)	670	50	
	85	018-2021				015-2021 20210731	570	
	720	20210915 (50,49)				(39,60)	20 590	

100 100 > 400 + 100 > 600 + 90 = 690 010-e-011-2021

100 100 100 50 740 → (56,32)

210 Data-NovoVivo-102-2020 (28,83) DATA Continuar

105

20211007/// (20211007////

US$5 / 3,78(gal) = 1,322751

x 5,5 (1US$=R$5,50) = 7,275132

Books

20211007/ Geometric Approaches to ~~of~~ Quantum Field theory/

3030852687.pdf

Pag. 15(31) chapter 2

Field Space Covariance

"God always geometrizes" Plato.

2.1 Riemannian Geometry.

Pág. 34(50) - Gaussian integral!

- 20211007/ Précis de Géométrie.../

Précis de géométrie.. Papelier...pdf

Pág. 292(301) Fig. 120.

Obs. Nº 380 - Tangentes à l'ellipse.

$x = a \cos \varphi; \quad y = b \operatorname{sen} \varphi;$

Obs: Ellipse usa estas equações para simular a deformação do espaço-tempo com as devidas adequações.

Na realidade a ellipse é uma deformação do círculo, ou seja, com os devidos ajustes das equações matemáticas podemos simular a curvatura do espaço-tempo do Universo.

Obs: $x = a \cos \varphi,$

$$y_1 = \frac{a}{b} y = a \operatorname{sen} \varphi$$

Pág. 201 (210) - Fig. 89

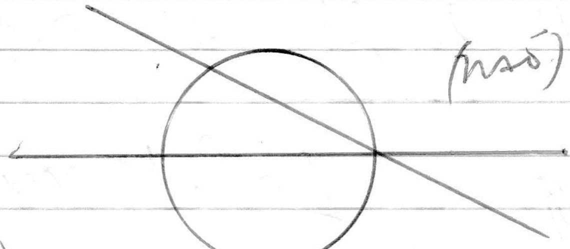

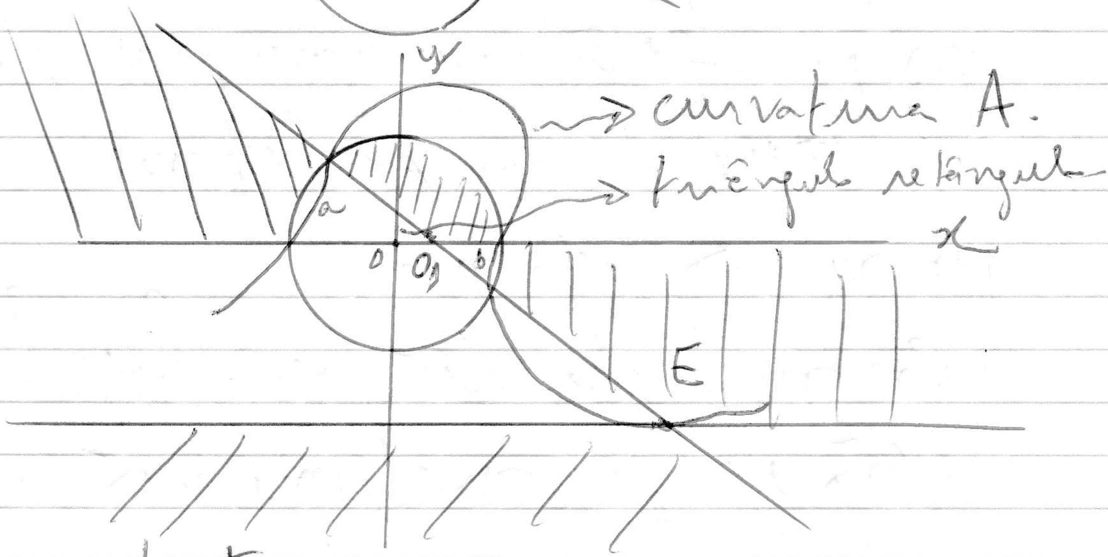

Important:

Estudar a razão entre $\underline{a \text{ e } b}$ em relação ao ângulo no ponto $\underline{O_1}$, observando se quando os dois são ângulos opostos, ou seja, tem o mesmo ângulo.

Veja como este comportamento matemático interfere na curvatura de $\underline{A}$, simule até obter a igualdade com a curvatura tempo-espaço do universo.

Observe que O e O_1 formam um triângulo retângulo (Pitagoras).

Use tudo isto com a ellipse do caso anterior para calcular a curvatura do Universo e que repercutiu no mesmo Mus (TT).

Obs: Estudar este livro todo!!

20211008 10h 15 min

porque – Não fui ao cinema porque perdi a hora.
↳ substituir por pois, visto que...

por que → pelo qual
Por que ela não veio hoje?
Não sei por que você não consegue...
Subst: por que motivo

por quê →
Você não foi à festa por quê?

porquê →
As crianças não entenderam o porquê do castigo.

In September of 476 A.D., the barbarian commander Odoacer forced the teenaged Western Roman emperor Romulus Augustus

to resign his office.

"Odoacer, king of the Goths, took control of Rome".

↳ the Constantinopolitan chronicler Marcellinus Comes in the 510.

The capture of Rome by the East came 60 years and three months after Odoacer's coup of 476.

Western Roman Empire had clearly fallen by the 560s.

---

smpf (02)	(03)	(04)	(05)	(06)	(07)	spf						
100	160	100	120	90	110	110	160	225	180	150	170	170
105	150	130	170	200	110	150	220	170	210	190	170	
100	150	120	140	120	140	110	150	395	400	170	130	
110	120	120	120	110	220	120	180	200	130	510	120	
100	580	130	100	100	110	150	710	600	160	190	590	
100	110	600	650	620	690	640	60 02 001	80	690	700		
	690	006-2023		01.002	09 03.002			680	(P01 01 001)	(P02 07)		
159	016-2021	(26,24)	196-2020	(47.9)	(42,42)			(2018 01)				
(3,84)	20210813		(12,55)									
	(41,03)											

[illegible]	no	MVSICS		Prof4590	240	200	200	90	45
129	105	244	479	320	250	200	300	60	100
69	82	133	151	220	490	200	200	150	145
98	187	104	630	540	40	90	✓		190
7	57	479		120	530	690			335
105	244		85	660	PHOTOS 5482. 001 (50)				SV-64 (9,78)

Books:

- 20211008/ College Algebra from Basic/
  College_Algebra_from_Basic....pdf
  Pág. 681(696) Chapter 8
  Analytic Geometry.
- 20211008/ Krishna's-Dynamics.../
  Krishna's-Dynamics....pdf
  Pág. D-159 (164) Chapter 3
  Constrained Motion on Smooth and Rough Plane Curves
  (Vertical Circle and Cycloid)
- Algebra-Linear.pdf
  Pag. 7 Chapter 1 Vector Spaces.
- Algebra Trigonometry.pdf
  Pág. 187 Chapter 3
  Logarithms of Complex Numbers
- Mathematics.pdf
  Pág. 119 Chapter 7 the Riemann Integral.

20211008 (cont.)

Pa _20211008/Krishna's.../
Mathematics.pdf (cont.)
Pág. 212 Chapter 3
Conformal Mappings

- Krishna's - Functions of Several Variables... pdf
- Krishna's - Solved Papers Mathematics..pdf
- Krishna's - Analysis, Edition-6.pdf

Pág. 178 Chapter 4
The Riemann Integral

- Krishna's - Complex Analysis, Edition-3.pdf

Pág. 8 - Chapter 1
Complex Numbers and their Geometrical Representation.

- Krishna's - Algebra Trigonometry 2nd Edition.pdf
- Krishna's - Mechanics of Solid.pdf

1B

- Krishna's - Calculus, 2nd Edition.pdf

Pág. 114 - Chapter 7

Curvature - Meaning of Curvature.

Obs: Livro todo!!!

- Krishna's - Matrices Vector....pdf

Pág. 106 - Chapter 5

Differentiation and Integration of Vectors

Obs: Livro todo!!!

cont. Pág. 166 - Chapter 7

Green's, Gauss's and Stoke's theorems.

- Krishna's - Engineering Mechanics ...4E.pdf

Pág. 39 2.4

Varignon's theorem (Law of Moment).

- Krishna's - Elements of Quantum Mechanics, Atomic... pdf

Jandaia

Pág. 85 – Unit 2

Schrödinger Equation and its application

2.1 → One Dimensional Time Independent Schrödinger's

- Krishna's Integral Equations 27B.pdf
- Krishna's - Mechanics Wave Motion.pdf

Books:

Books-020-2021/Important/

IMPORTANT-20211008/

Obs: Diretório todo!!!

- An Expedition to Geometry...pdf

Pág. 53 – Projective Geometry

Obs: Livro todo!!!

- Calculus without Limit Almost.pdf
- Linear Algebra Illustrated.pdf

Obs: Livro todo!!!

Pág. 177 3.3

Transforming the axes transforms the plane.

— S. Kumaresan – A Course in Riemannian Geometry.pdf

Pág. 46 2.4

Gauss Theory of Surfaces in $\mathbb{R}^3$.

Pág. 54 2.5

Curvature and Parallel Transport

— S. Kumaresan – Topology of Metric Space.pdf

Pág. 122 (135) Chapter 6

Complete Metric Spaces

— S. Kumaresan – A Course in Differential Geometry...pdf

Obs. Livro todo!!!

data
S T Q Q S S D

20211009 ///

Books:

- 20211009/ Astronomics/
the Rational Mechanics of Flexible or Elastic Bodies ...(1960).pdf

Pag. 61 Mariotte and Leibniz on Elastic Beams (1684)

Figure 19.

Leibniz's figure for calculating the bending moment acting on the cross section of a terminally load beam (1684)

- A first course in rational continuum mechanics - Volume 1 ... (1977).pdf

Obs: Livro todo!!!

04	05	06	07	08	zip.1
130	100	210	100	250	190
210	100	210	120	60	210
100	190	110	140	310	400
100	105	530	110	150	210
110	100	90	130	220	610
40	595	620	600	680	70
690	SV78	60	100	(201812)	680
74-10	2008220	680	700		
(17,31)	(3,28)	SV78	86-90		
		20210102	20210802		
		(12,057)	(50)		

Class 15 Test 2.mp4
(86,9)

Prox:
Class 16

—

Video-Math.
705 files.

Can a Circle be a Straight Line
Space Time PBS

—

69,8 (20211009)
8,6
78,4 TB
31,24
109,64 TB

— Terra- $1,26^{21}$ l H_2O

—

Books:

— 20211009/Numerical Geometry.../
3030767973.pdf
Pág. 37(53)-
1 The Gauss Integral for the Linking Number of Curves

— 20211009/ Comparison Finsler Geometry / Ohta2021_Book_ComparisonFinslerGeometry.pdf
Pág. 29(50) Lemma 3.18
(Gauss's Lemma)

Pag. 92(112)

$$\mathrm{diam}(M) \leq \pi \sqrt{\frac{n-1}{K}}$$

In particular,

M is compact and has finite fundamental group.

Obs: The Bonnet-Myers theorem (Pag. 91(111))

The Riemannian characterization of the Finsler curvature as in theorem 5.12 (Pag. 54(75) Riemannian characterizations of curvatures).

has some straightforward applications in comparison geometry.

Pag. 129(148) - Chapter 10

Examples of Measured Finsler Manifolds.

→ Ricci curvature

10.1 Minkowski Normed Spaces

Pag. 130(149) — Gaussian-type measure

$$m(dx) = e^{-K\|x\|^2/2} \mathcal{L}_n(dx) \ \ldots$$

Pág. 151 (169) - Gauss lemma
(Lemma 3.18).

Pág. 208 (222) Theorem 15.1
(Bakry-Ledoux Isoperimetric Inequality)

In (15.3)...

Gaussian isoperimetric inequality.

_20211009/Image and Graphics...|
3030873609.pdf
Pág. 99 (115)

First, we use the normalized Gauss weighted Euclidean distance as the weight for the interpolation:

$$W = N(exp(-d^2/2\sigma^2))$$

where d is the Euclidean distance from the point in the region to the regional center, and $N(.)$ is the normalization operator.

Pág. 146 – Gauss-Newton in the traditional models.

Pág. 171 (186) –

Gaussian partial derivative filters along the horizontal and vertical direction

Pág. 311 (326) –

Poisson – Gaussian noise model,

—

20211010 ///

Michelson and the speed of light
Bernard Jaffe 1886-1986.

The speed of light
Ron Carlson

Smarandache Geometries and Curves

https://archive.org/details/smarandache-geometries-and-curves

https://archive.org/details/Mgptheory

—

Books:

- 20211010/ Géométrie Contemporaine - 3 volumes//
- Star and Relativity... vol. 1.pdf (somente imagem)
- Star and Relativity... djvu (texto)

Source: 20211010/ Géométrie Contemporaine - 3 volumes/ Star-and-Relativity-Zel'dovich/ Star and Relativity... (2011).djvu

Pág. 16(33)

The curvature at a given point of a three-dimensional space is determined in the following manner.

A geodesic surface (analogue of a plane in Euclidean space) is drawn through the point, and its curvature is calculated. This curvature is called the Riemann curvature of space in the two-dimensional direction of the chosen surface. The curvature may be

different in different directions.
The curvature averaged over all directions is called the Gaussian curvature of the three-space.

Pág. 28(45) – Gaussian curvature (cf. § 1.4)

$$C_G = \frac{8}{3}\pi(G/c^2)\rho.$$

Pág. 89(106) Gaussian curvature of space C_G at the center of the star is larger than it is at other places.

$$C_G = (8\pi/3)(G\rho/c^2).$$

– String Gravity and Physics at the Planck Energy Scale (1996).pdf

Pág. 45(60) –

Quadratic curvature corrections (in terms of the Gauss-Bonnet term) to the Einstein action are included [36].

Pág. 122(134) – 1.1

Point particles and open strings in magnetic field (exactly solvable)

becomes gaussian if the field

strength is constant,

Pag. 179 (189) 4.

Covariant Double-null Dynamics

The geometrical framework is the embedding formalism of Gauss and Codazzi for the foliation of spacetime by spacelike hypersurfaces.

Pag. 259 (266)

Euclidean Hamiltonian constraint [13]

$$R + K^{ab}K_{ab} - K^2 = 0,$$

rather than the Lorentzian Hamiltonian constraint

$$R - K^{ab}K_{ab} + K^2 = 0,$$

...

Pag. 266 (273)

if we had only the Gauss and the diffeomorphism constraints, the total volume of Σ would be a Dirac observable.

---

https://archive.org/details/@librosmir

Relaciones y geometría analítica
López Quiles, Antonio
https://archive.org/details/
isbn_978684441514/

cont.

20211010/ Géométrie Contemporaine.../
_Modern Geometry - Methods and
Applications - Part I....pdf
Pág. 77(92) 8.1.1. Definition
Gaussian curvature K,

Pág. 81(96)
Recall that the ratio of these quadratic forms in the components of the tangent vector to the section of the surface by a plane orthogonal at the surface at a point (i.e. a "normal section"), gives the curvature of the section (i.e. "normal curvature") at that point (Corollary 8.1.3)
↳ Pag. 79(94)

8.3.1. Definition →

the product of the principal curvatures

is called the Gaussian curvature K of the surface at the point and their sum the mean curvature.

Pag. 84 (99)-

8.3.5. Theorem

The Gaussian curvature of a surface is equal to the ratio of the determinants of the second and the first fundamental forms:

$$K = \frac{b_{11} b_{22} - b_{12}^2}{g_{11} g_{22} - g_{12}^2} \ldots$$

Pag. 85 (100) 8.3.7. Example

Gaussian curvature $K < 0$

Finally we explain the geometric significance of the sign of the Gaussian curvature at a point.

Pag. 95 (110) the Language of Complex Numbers in Geometry

Then by partial analogy with Theorem 8.3.5, we define Gaussian curvature K of a space-like surface in R^3_1 by

$$K = \frac{\det(b_{ij})}{\det(g_{ij})}. \qquad (14)$$

Pág. 109 (125)

§13. The Conformal Form of the Metric on a Surface

13.1 Isothermal Co-ordinates. Gaussian Curvature in terms of Conformal Co-ordinates.

... The Euclidean metric in $\mathbb{R}^3$ induces on this surface a Riemannian metric

$$dl^2 = E(dp)^2 + 2F\,dp\,dq + G(dq)^2, \ldots$$

Important:

Pág. 117 (132)

13.2 Surfaces of Constant Curvature

Are there surfaces in Euclidean 3-space whose induced metrics are hyperbolic (i.e. Lobachevskian)?

... There is a one-to-one metric-preserving map onto, say, the Poincaré model of the hyperbolic plane.

Pág. 118(133) -

13.3.1 theorem

A Surface in Euclidean 3-space with metric $dl^2 = g(z,\bar{z})\,dz\,d\bar{z}$, and with constant Gaussian curvature K is locally isometric to:

(i) the sphere if $K>0$;
(ii) the Euclidean plane if $K=0$;
(iii) the Lobachevskian plane if $K<0$.

Pag. 171(186)

Gaussian curvature $K = \frac{\det(b_{ij})}{\det(g_{ij})}$;

mean curvature $H = b^i_i = g^{ij} b_{ij}$.

Pag. 396(411)

37.4.2 Theorem

Let g_{ij}, $i, j = 1, 2$, be a Riemannian metric on a 2-dimensional space, and let K denote as usual the Gaussian curvature of the space.

Important:

Pág. 397(412)–

37.4.3. Corollary (Gauss-Bonnet)

The integral over a closed surface in 3-dimensional Euclidean space of the Gaussian curvature of the surface is not changed by smooth deformations of the surface:

$$\int K\,dS = \text{const}.$$

— Modern Geometry— Methods and Applications— Part II ... pdf

Pág. 94(110)—

where K is the Gaussian curvature of M

Pág. 95(111)– Figure 27

Gauss map n to S^{q-1}.

Pág. 112(128) 14.2

The Degree of a Vector Field on a Hypersurface.

Pag. 114 (130) –

14.3 The Whitney Number. The Gauss-Bonnet Formula.

⟹ Gauss map

Pag. 118 (134)

14.3.4 Theorem (Gauss-Bonnet)

$$\frac{1}{2\pi}\int_Q K\,d\sigma = 2-2g.$$

...

Pag. 412 (428)

Lagrange-Euler equation

$\delta S=0$

... surface of constant negative Gaussian curvature to be imbeddable in Euclidean space $\mathbb{R}^3$.

Geometria Moderna;

- Geometría moderna ... (1987).pdf
  Dubrovin

- 0755. Geometría analítica.pdf
  Obs: Livro todo!!!

Pág. 127(137) —

Ángulo entre dos rectas

Los principios básicos de la geometría establecen que um ángulo exterior de um triángulo es igual a la suma de los ángulos interiores opuestos

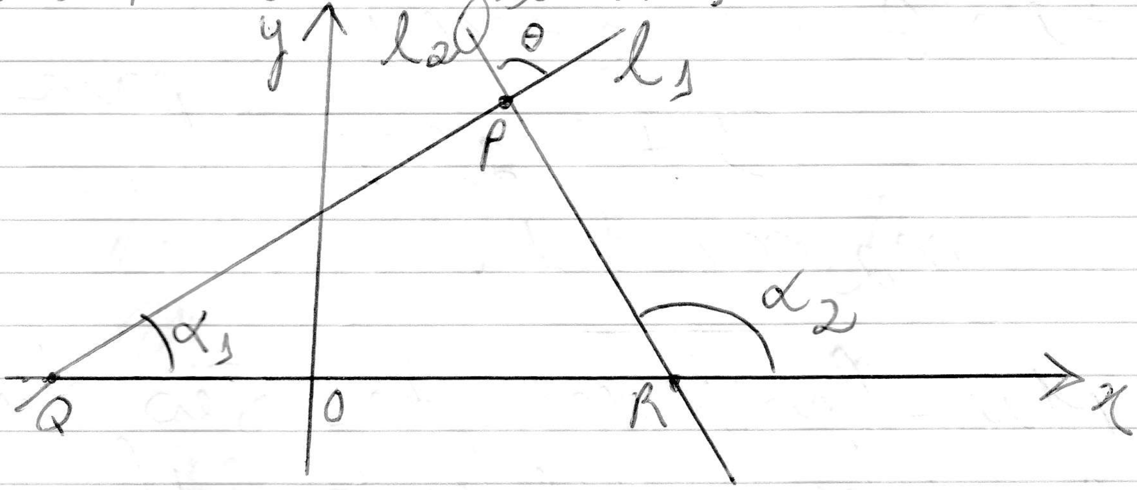

$\alpha_2 = \alpha_1 + \theta$

Ángulo entre las rectas es:

$$\theta = \alpha_2 - \alpha_1$$

Al aplicar la fórmula trigonométrica

$$\tan(\alpha_2 - \alpha_1) = \frac{\tan\alpha_2 - \tan\alpha_1}{1 + \tan\alpha_2 \tan\alpha_1},$$ se tiene:

$$\tan\theta = \frac{\tan\alpha_2 - \tan\alpha_1}{1 \pm \tan\alpha_2 \tan\alpha_1}$$

Al substituir $m_1 = \tan\alpha_1$ y

$m_2 = \tan\alpha_2$

resulta:

$$\tan\theta = \frac{m_2 - m_1}{1 + m_2 m_1},$$ para

$m_2 m_1 \neq 1$.

- Mir_Problemas_de_Geometria_Analitica_Kletenik.pdf
- Mir_Problemas_de_Geometría_Diferencial_Fedenko.pdf

Obs: Livro todo!!!

Curvas and Superficies.

– 20211010/ Géométrie Contemporaine…/
Map-Theory/
_9781599731346.pdf ⎤ same
-Maptheory.pdf ⎦ book.

Pág. 09(24) - Fig 1.3
Sphere and projective plane
Obs. livro todo!!!

— /Map-theory/
1931233144.pdf
Smarandache Multi-Space Theory
Pág. 173(183) Chapter 4
Planar Map Geometries
Important!
Utilizar este livro com estes dois livros posteriores:
Source: 20211010/ Géométrie
Contemporaine…/

– Géométrie Contemporaine
1re partie ... (1982). djvu
(Doubrovine)

Pag. 86 (84) – § 8.
Deuxième forme fondamentale
1. Courbure des courbes sur la surface dans l'espace euclidien.

Obs: Pag. 87 (85) –
On appelle courbure totale ou courbure de Gauss le déterminant de la matrice (a_{ij}).

Pag. 253 (251)
C'est la formule de Gauss–Ostrogradski dans l'espace euclidien à trois dimensions.

– Géométrie Contemporaine
2e partie ... djvu

Pág. 101(99) Chapitre 3
Degrée de l'application.
Indice d'intersection.
Applications de ces notions.

§12 Notion d'homotopie

Obs: Teoria da homotopia

é um estudo sistemático de situações em que os mapas vêm com homotopias entre eles. Sabe-se que, se existe uma função contínua entre os espaços topológicos, significa que são homotópicos.

Important: Use este conceito para formular uma equação na qual poderemos representar uma superfície curva (Gauss) em uma superfície plana, ou seja, esta equação irá representar uma superfície curva projetada ortogonalmente sobre uma superfície.

Usaremos a homotopia para representar a projeção do espaço curvo (espaço-tempo) sobre um plano.

Pág. 97(95) - Fig. 27

application gaussienne n est une sphère S8-1

- Modern Geometry ... Part I ... djvu
(Novikov)
- Modern Geometry .. Part II ... djvu
- Modern Geometry ... Part III . djvu

Obs: Nestes 3 volumes há inúmeras referências ao Gauss
(estudá-las !!!)

Cont:

20211010/ Géométrie Contemporaine .../
Stars-and-Relativity-Zel'dovich/
Stars and Relativity (2011).djvu

Pag. 89 (106) -
Gaussian curvature (see §1.4) of space, C_G at the center of star is larger than it is at other places.

$$C_G = (8\pi/3)(G\rho/c^2).$$

Pag. 28 (45) - Gaussian curvature for the three-dimensional space of the comoving reference frame in the static

$$C_G = \frac{8}{3}\pi(G/c^2)\rho$$

In the case of an isotropic deformation of matter and vanishing rotation, the Gaussian curvature is

$$C_G = \frac{8}{3}\pi(G/c^2)\rho - \frac{1}{9}D^2$$

1.9 The Cosmological Constant

V is the four-dimensional volume

$$S = -mc\int ds - \frac{c^{\text{③}}}{16\pi G}\left(\int R\,dV + \int 2\Lambda\,dV\right).$$

Pág. 138(155) - 4.4

The Schwarzschild Sphere in an External Quadrupole Field.

...

estudar!!!

Pág. 139(156)

This surface is a Schwarzschild sphere deformed by an external field. The Gaussian curvature of the two-dimensional surface S_m (not the Gaussian curvature of the three-dimensional surface created by S_m as it moves through spacetime),

$$C_G^{*} = \frac{1}{4m^2} e^{\delta}\left[1 + \text{③}q - 12q\mu^2 - 9q^2\mu^2 + 9q^2\mu^4\right],$$

is different for different μ and is

always finite.

The physical properties of this S_M are the same of those of the Schwarzschild sphere!

Pag. 441 (458) –

We can rewrite the condition for linkage of the protons to the field lines as

$$H > 10^{-12} (r_K/r)^{3/2} \text{ gauss.}$$

(13.4.12)

Pag. 456 (473) –

455 (472) –

What could cause the necessary quadrupole moment?

The magnetic field, inclined at an angle relative to the axis of rotation, will produce some deformation.

However, Ostriker and Gunn have shown that the dipole magnetic field needed for the ~~m~~ magnetic deceleration is too weak to produce the necessary deformation.

⇒ "One has to assume an internal field of the order of 10^{15} gauss with magnetic field lines hidden inside the star in order to sustain the necessary asymmetry against the force of gravity.14

Pág. 466(477) Fig. 72

Topology of the magnetic field of stars (a), and (b) a metagalactic field.

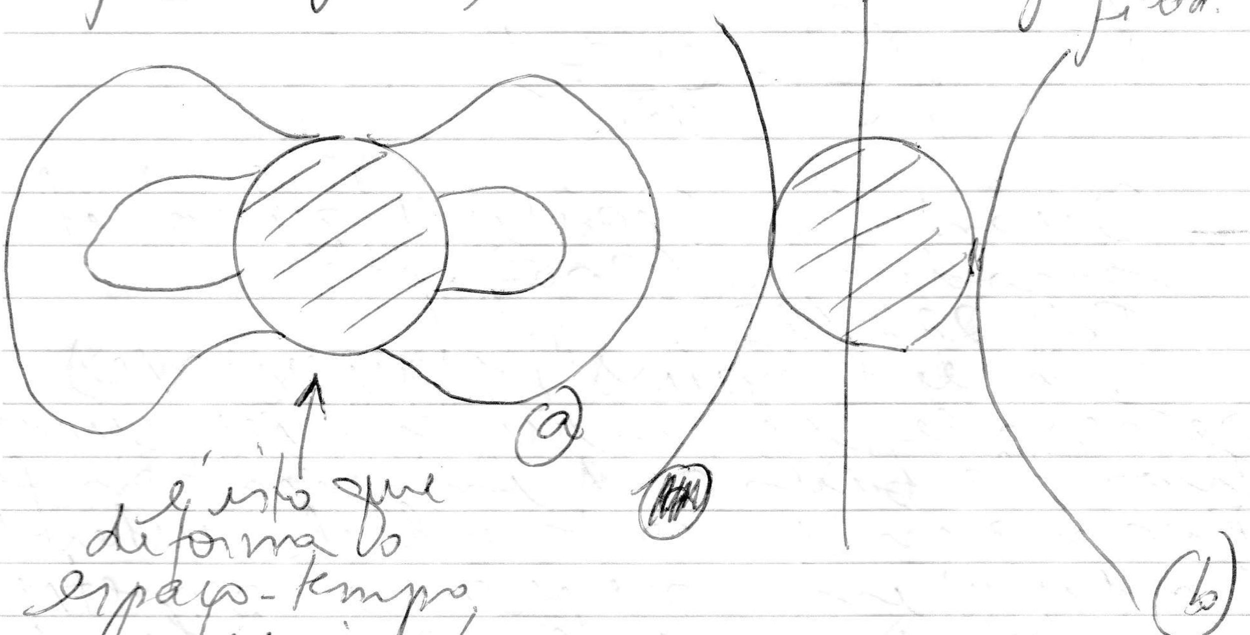

é isto que deforma o espaço-tempo, ou seja, field é que gera a gravidade. A gravidade é decorrência das forças fundamentais da matéria.

Fundamental, melhor!!!

Pág. 463(480) ⇒ $M = 10^8 M_{\odot}$, the external regular field will have a strength $H_0 \approx H_i\, n^{-1/2} \approx 10^{-7}$ gauss.

Books: (Important)

20211010/ Gravity/

– College-Algebra ... to-Advanced-Level.pdf

Pág. 681 (696) – 8 Analytic Geometry

Pág. 684 (699) Equation (proof)

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$$

– Quantum Gravity. Mathematical models ... (2007).pdf

Pág. 222 (232)

Indeed, Gauss "Disquisitiones generales circa superficies curvas" was the basis and inspiration for Riemann's work which ushered (conduziu) in a new era in geometry. It is an extension of this geometry that is the cornerstone (pilar) of relativity theory.

Pág. 246 (255)

When adds a Gauss-Bonnet term to the Einstein Hilbert action one obtains the same entropy formula (10) as when using the full Weyl multiplet [38, 37].

Pág. 298 (306) –

Figure 1. Part of theory space of the Einstein-Hilbert truncation with its RG flow.

→ flow is dominated by a non-Gaussian fixed point in the first quadrant and a trivial one at the origin.

— Gravity and Strings...(2004).pdf

Pág. 24 (45) —

Gauss-Ostrogradski theorem

Pág 218 (239) Rodapé

Coulomb force between two charges is

5. $\frac{1}{4\pi} \frac{q_1 q_2}{r_{12}^2}$ (8.14) ↙

In the Gaussian system we should replace $1/(4c)$ by $1/(16\pi c)$ and the factor of 4π disappears from the Coulomb force!

Pág. 243 (264) Euler characteristic of the Euclidean ERN solution using the Gauss-Bonnet theorem adapted to manifolds with boundaries. The

Euler characteristic χ is a topological invariant whose (de quem) value is an integer and the Gauss-Bonnet theorem states that the integral of the 4-form (naquela)

$$\frac{1}{32\pi^2}\epsilon_{abcd}R^{ab}\wedge R^{cd} \quad (8.115).$$

over a four-dimensional compact manifold M is precisely χ.

Pag. 355(376) →

which is in turn (que é por sua vez) related to the modification of the Gauss law introduced by the dilaton.

→ partícula hipotética

Trata-se de uma partícula de campo escalar Φ, um campo escalar que sempre surge com a gravidade.

Pag. 507(528) →

Using the generalization of the Gauss law Eq. (18.37) and Stokes' theorem.

— Approaches to quantum gravity (2009).pdf

Pág. 80(102) 5.4
A spacetime approach to
Quantum Gravity Theory
Lorentzian metric is quantized...

Pág. 122(144) Fig. 8.2.
The flow in the Einstein-Hilbert
truncation
(Gaussian FP).

Pág. 244(266)-
Newton's constant and $i > 0$
is any positive real number
Define the Gauss constraint
$G_j := \partial_a E_j^a + \epsilon_{jkl} A_a^k E_l^a$.

Pág. 245(267)
From the Gauss constraint and
spatial diffeomorphism constraint
respectively.

Pág. 276(298)-
- Poisson bracket
- Gauss law and the curvature
constraint respectively.

Pag. 389(411)

For instance the Gauss law of Yang-Mills theory of gravity is written in the new variables formulation.

Pag 512(534)

- Poincaré algebra

- GR is an example of a constrained theory: there is a set of first class constraints that encode the diffeomorphism symmetry.[1]

Gauss Law

Pag. 558(580)

Gauss's law constraint which requires that

$$D_a \tilde{E}^a_i = D_a \frac{\delta S(A^i)}{\delta A^i_a} = 0 \quad (28.15)$$

- Tomás Ortín - Gravity and Strings .. (2015).pdf

- Modifications of Einstein's Theory of Gravity .. (2015).pdf

Pág. 29(45) - EGB

Einstein-Gauss-Bonnet theory.

Pág. 31(47)

One can trace the origin of such terms in the early works of GAUSS who measuring geodesic distances noted that scalar (Gauss) curvature of two dimensional surfaces depended only on the first fundamental form, in other words the intrinsic metric of the surface and its derivatives.

This was the basis of what he called the Egregium theorem; scalar curvature (unlike (diferente) other extrinsic curvature (components) does not depend on the variation of the normal vector field on the surface i.e. on how the surface is embedded in three dimensional space.

Then later, for Euler in his work on surface triangulations noted that two dimensional surfaces can be topologically classified by their "Euler" number,

$X: \chi[M] = 2-2h$ where h is

the number of topological handles.

Pág. 32(48) → página toda

– Gauss-Bonnet theorem

– Einstein-Hilbert term is in two dimensions is a topological invariant, i.e. the Einstein tensor in two dimensional space-time is identically zero.

Pág. 39(55)

the Gauss-Bonnet scalar of the internal space...

Pág. 43(59)

the last line in the effective action takes part only for a curved internal space in the face of Ricci and Gauss-Bonnet curvature.

– A. Zee - Einstein gravity (2013).pdf

Pág. 65(88) -

How do we know whether a space is curved or not?

Obs: Livro todo!!!

Pág. 82(105) 1.6.
Curved Spaces: Gauss and Riemann
Pág. 104(127).
Gauss's strategy
Pág. 581(604)

- Einstein-Hilbert action in curved spacetime
- Gauss, Riemann, and Ricci solved this highly nontrivial problem for us.

Pág. 725(748) Appendix 3:
Topological invariants: Euler characteristic, Gauss-Bonnet theorem, and all that.
Pág. 796(819) - Solution 1.6

- Pietro.. Gravity, a Geometrical Course - Volume 1 ..(2012).pdf

Pág. 87(100) Chapter 3
Connections and Metrics
3.2.1 Gauss Introduces Intrinsic

Geometry and Curvilinear Coordinates.

- Pietro .. Gravity, a Geometrical Course - Volume 2 .. (2012).pdf
  Pág. 79 (97) Fig. 4.6
  Gauss law of fluxes.
  Pág. 217 (233)
  Gaussian coordinates $\{\sigma, \tau\}$
  Pág. 404 (418) -
  Gauss and Riemann →
  Einstein

Obs: Usar estes 2 volumes !!!

- Gravity, Special Relativity, and the Strong Force .. (2012).pdf
- Newton's Gravity .. (2012).pdf
  Pág. 258 (272)
  the Gaussian constant in Celestial Mechanics

$\sqrt{G} = .01720209895$

In Gauss' day the period of revolution of the earth was taken as

365.2563835 days.

the combined mass of the Earth-Moon was assumed by Gauss to be

1/354710 of the Sun's mass

Pág. 259(273)

Gauss called this value k so,

$K = .01720209895$

this is the Gaussian constant or the Gaussian gravitational constant.

— Metric theories of Gravity (2017). pdf

Pág. 371(396) 8

Conserved quantities in the Einstein-Gauss-Bonnet gravity (EGB)

Pág. 425 (450)

Einstein-Gauss-Bonnet theory

- Equations of Motion in Relativistic Gravity (2015).pdf

Pág. 12 (25)

The difference between their values for two different choices of Σ can be converted by Gauss's theorem to volume integrals.

The integrands then vanish by the conservation equation for $T^{\alpha\beta}$ and because it is symmetric.

Pág. 253 (263)

The Dependence of the Particle's Momenta on Σ

Gauss theorem (see Fig. 7)

Pág. 254 (264) shadowed regions

⇒ hypersurfaces. IMPORTANT.

a. Curved spacetime

b. flat spacetime

Pág. 382(390) 4.3.5

Center of Mass

Obs: Usar isto para calcular o centro de massa que gera a curvatura do espaço-tempo.

Obs: Lorentz force.

— Geometrical theory of Satellite (2018).pdf

Pág. 264(275) 18.5

Environmental Effects on the Galileo Clock Parameters

⇒ (gauss equal 10^{-4} tesla)

— Quantum Gravity - World Scientific Publishing ..(1997).pdf

⇒ Gauss map

— ~~Overcoming Gravity...(2016).pdf~~

(error)

- Principles of Astrophysics (2014).pdf
- Beyond Spacetime - the Foundations of Quantum Gravity (2020).pdf

Pag. 90(99) Gaussianity of the spectrum

- Baggott - Quantum Space (2018).pdf

Pag. 36(57)

Einstein's friend Marcel Grossman directed him to the works of the great German mathematician Carl Friedrich Gauss and his student, Bernhard Riemann.

Einstein was introduced to the Riemann curvature tensor, the definitive mathematical expression for the curvature of a generalized geometry in any set of coordinates.

- Einstein's general theory of relativity (2010).pdf

Pag. 8(26) 1.5

Local form of Newton's Gravitational Law

Pag. 9(27)

$$\nabla^2 \phi(r) = 4\pi G\rho(r) \quad (1.32)$$

this Poisson equation is the local form of Newton's law of gravitation.

→ Mass generates a gravitational field according to Poisson's equation, and the gravitational field generates acceleration according to Newton's second law.

Pag. 118(136)

$$\int_M d\alpha = \int_{\partial M} \alpha . \quad (6.71)$$

this is the generalization of Gauss's law for vector calculus.

$$\int_{\partial M} E . dS = Q \quad (6.75)$$

This is the famous Gauss's law in electromagnetism.

Pag. 151(169) - 7.2 Surfaces

⇒

$$e_{\mu,\nu} = \Gamma^{\alpha}_{\mu\nu} e_{\alpha} + K_{\mu\nu} n \qquad (7.15)$$

The Equation (7.15) is usually called Gauss' equation.

Important:

Pag. 152 (170) - Figure 7.2

A two-dimensional surface embedded in three-dimensional Euclidean space

The geodesic and normal curvatures, taken separately, characterize the extrinsic geometry of the surface.

We shall now find a quantity, called the Gaussian curvature, which is a measure of the intrinsic geometry of the surface.

Pag. 153 (171)

The Gaussian curvature of the surface is defined as

$$K = k_1 k_2 \qquad (7.29)$$

Obs: Usar livro todo!!

Pág. 153(171)

Gaussian curvature represents a measure of the intrinsic geometry of surface.

This will be achieved by expressing it in terms of the components of the Riemann tensor.

Pág. 160(178) -

Gauss' Theorem Egregium

Pág. 161(179) Fig 7.4

A cone is intrinsically flat.

Pág. 166(184)

Let us use Gauss' theorem Egregium to calculate its curvature properties.

the Gauss' theorema Egregium now gives us the Riemann tensor (7.125)

Pág. 171(189) 7.4

The Gauss-Codazzi equations

Pág. 255(273) →

Gaussian curvatures of the surfaces they describe are given by..

$$K = -\frac{1}{2g} g'_{\phi\phi} + \frac{g_{\phi\phi}}{4g^2} g'_{rr} g'_{\phi\phi} + \frac{g_{rr}}{4g^2} (g'_{\phi\phi})^2, \ldots$$

Pág. 429 (447)

Gauss' Theorema Egregium and the Codazzi equation, we have..

(16.9) (16.10) e (16.11).

— Einstein's theory – a rigorous introduction to general relativity (2002).pdf

Pág. 349

The intrinsic curvature K_I is called the Gaussian curvature

— Conversations on Quantum Gravity .. (2021).pdf

Pág. 16 (28)

What is asymptotic safety?

Gaussian fixed point

Pág. 445 (457) non-Gaussian fixed point contrary to a Gaussian fixed point?

20211011///

BOOKS:

Gravity - Hartle

Gravity - Misner

- 20211010/ Gravity-Hartle/
Hartle-Gravity...(2014).pdf

Pag. 16(19) - Triangle of Gauss.

- 20211011/ An Introduction to
Calculus with Hyperbolic Functions.../
An_Introduction_to_Calculus...pdf

- 20211011/ Gravity-02/

Gravity Newtonian, Post-Newtonian,
Relativistic.pdf

Pag. 09(25) - Proof that

$$\nabla^2 |x-x'|^{-1} = 4\pi \delta(x-x')$$

→ Gauss's theorem.

Pag. 309(325) - Box 6.5

Green's function for the wave equation

→ We can use Gauss's theorem to get

- 20211011/ Gravity-Hartle/
Misner-Gravitation...(1973).pdf

Pag. 32(59) _ Curvature.

Others:

_ 20211010/ Gravity/

Quantum theory of Gravity...(1984).djvu

Pag. 456(470)

the covariance of the relevant Gaussian is a Green's function of the Jacobi operator with boundary conditions determined by the domain of integration.

_ 20211010/ Gravity/

Quantum gravity - Cambridge (2004).djvu

Pag. 33(51)

Riemann geometry.

The tensor g equips the spacetime manifold M with a metric structure: it defines a distance between any two points, and this distance is a smooth function on M.

(More precisely, it defines a pseudo-metric structure, as distance can be imaginary).

Riemann studied the structure defined by (M, g), called today a riemannian manifold, and defined the Riemann curvature tensor as a generalization of Gauss' theory of curved surfaces to an arbitrary number of dimensions.

Riemann presented this mathematical theory as a general theory of "geometry" that generalized Euclidean geometry.

Einstein utilized this mathematical theory for describing the physical dynamics of the gravitational field.

_ 20211011/ Gravity-Hartle/

Gravitation-Misner-1973.pdf

Pág. 32(65) - Curvature of What?

Pág. 146(179) - §5.9.

Conservation of 4-Momentum: Differential formulation

→ Gauss's theorem in four dimensions...

_ 20211011/ A Book of Curves Lockwood/

EH Lockwood - A book of Curves (1961).pdf

Obs. Livro todo!!!

_ 20211011/ Gauss/PPN578628864.pdf

Die Theorie der Parallellinien _

data . .
Ⓢ Ⓣ Ⓠ Ⓠ Ⓢ Ⓢ Ⓓ

20211012 III

https://archive.org/details/
euclidoreinstein0000unse

⇒ Euclid or Einstein: a proof of the parallel theory and a critique of metageometry.
1931 Callahan, Jeremiah Joseph.

ISBN 9781258776879

=

- 20211011/Gauss/ED058058.pdf
Pag. 16 The Trisection Problem

—

Important:
Leibniz inspired
↓
"Numero impari gaudet" Virgil
DEUS

Book

- 20211012/Classical Analysis.../
Classical Analysis in the Complex Plane.pdf
Obs: Livro todo!!!

Author: Robert B. Burckel

burckel@math.ksu.edu

https//people.math.ksu.edu/
personnell_details.php?
person_id=1324

Procurando por GAUSS neste livro eu encontrei na Pág. XIX (highlights) (19) no item 6, mas olhando o anterior:

5. The Remarks following 3.22 give an interesting analytic derivation of Archimedes' estimate of π and 3.28 outlines short elementary proofs of the irrationality of e and π.

Procurando por 3.22, encontrei:

Pág. 127(153) §3 The Complex Exponential Function.

Theorem 3.22 (Gregory-Leibniz series)

$$\frac{\pi}{4}=\int_0^1 \frac{dx}{1+x^2}=\sum_{n=0}^{\infty}\frac{(-1)^n}{2n+1}$$

Proof: ...

$$C\left(\frac{\pi}{4}\right)=S\left(\frac{\pi}{4}\right)=\frac{1}{\sqrt{2}} \longleftarrow$$

número errado!!!

Pág. 129(155) → (8)

$$\frac{\pi}{6} = \frac{1}{\sqrt{3}} \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} \cdot \frac{1}{3^n} . \quad (8)$$

Important: Pag. 129 (155)

(ii) Besides using the partial sums of the series or Riemann-sum approximations to the integral, there is a third (more efficient?) way to exploit 3.22 for rational approximations to π.

We get 1/630 > 22/7 − π > 1/1260, and so

$$3 + \frac{10}{71} < \frac{22}{7} - \frac{1}{630} < \pi < \frac{22}{7} - \frac{1}{1260} < 3 + \frac{1}{7}$$

///

His geometric construction can be found on pp. 31-34 of C. H. Edwards [1979].
↳ Bibliography

Pag. 945 (967) —

Edwards, C. H., Jr.
The Historical Development of the Calculus – Verlag (1979).

Livro disponivel : Source

20211012/ The Historical Development of the Calculus/

digo

20211012/ Classical Analysis in the Complex Plane/
The Historical Development of the Calculus/
The Historical Development of the Calculus ... (1979).pdf

Pág. 31(42) The Measurement of a Circle

$$3\frac{10}{71} < \pi < 3\frac{1}{7}$$

other:

$$\frac{265}{153} < \sqrt{3} < \frac{1351}{780}$$

$$3\frac{10}{71} < \frac{6336}{2017\frac{1}{4}} < \pi < \frac{14688}{4673\frac{1}{2}} < 3\frac{1}{7}$$

return:

20211012/ Classical Analysis in the Complex Plane/
Classical Analysis in the Complex Plane.pdf

Pág. 129(155) (iii) →

Yet a fourth way to utilize
3.22 is to express

$\pi/4 = \arctan(1)$ as a rational linear combination of

$\arctan(\varepsilon_j)$ for small ε_j

(for which the series is more rapidly convergent).

One of these goes back to Euler:

$$\pi/4 = 4\arctan(1/5) - \arctan(1/70) + \arctan(1/99).$$

Pag. 180 (206) –

3.22 inspired Leibniz to quote Virgil (on gambling!):

"Numero Deus impari gaudet".

Essentially the first equality in 3.22 was adopted by Weierstrass (in 1841) as his definition of π.

(This was also the definition adopted in the alternative treatment of the exponential function give in 2.22).

The present definition of π as "twice the least positive zero of the cosine function"

↗

other: (cont.)

Pág. 114 (140)   (7)

thus $0 < \pi/2 < \boxed{\sqrt{3}}$ and

estudar!!!

$$C(\pi/2) = 0.$$

Pág. 135 (161) -

Theorem 3.28

π is irrational, and e^m is irrational for every $m \in \mathbb{N}$.

— 20211012/ Classical Analysis in the Complex Plane/
The Complex Itinerary of Leibniz's Planetary Theory...pdf

Pág. 96 (105) - Gravity

letter to Huygens in which Leibniz explicitly claimed:

I distinguish the aether which causes the gravity (pesanteur) (and perhaps also the direction of parallelism of the axes) from that which carries the planets, which is rather coarser.[43]

Leibniz 1690, in Huygens, 1901, Oeuvres Completes, IX, p. 526.

- 20211012/Cosmology in Gauge Field Theory ~~of~~ and String Theory/

Cosmology in Gauge Field...pdf

Pág. 292 (305)

- 20211012/Modular Forms...|

Modular_Forms.pdf

Search: Riemann

Pág. 16(22) theorem 4.1.

$$G_K(z) = 2\zeta(K) + \frac{2(2\overset{\swarrow}{\pi} i)^K}{(K-1)!} \sum_{n=1}^{\infty} \sigma_{K-1}(n) \cdot q^n,$$

Pág. 19(25)

Riemann surfaces (like $SL_2(\mathbb{Z} \backslash \mathbb{H} \cup \{\infty\})$, we have to admit poles of finite order in ∞.

Pág. 31(37) 5.2

L-functions of Modular Forms

→ Riemann zeta function.

20211012/ Cosmology...|

Supersymmetric gauge field theory (1994).djvu

Pag. 316(324) Appendix D.

Closed-String Integrals

This integral separates into standard (Gaussian) integrals when Cartesian coordinates are used for $z = x + iy$. Then

$$J(\lambda,\mu) = \frac{1}{s+t} \exp\left(\frac{\lambda t + \mu s - st}{s+t}\right) \quad (D7)$$

_20211012/ Image and Graphics...|

3030873544.pdf

Pag. 669(694) Spatial Structure Feature

The gradient amplitude map of an I-frame can be defined as:

$$GM = \sqrt{[I \otimes h_x]^2 + [I \otimes h_y]^2}, \quad (4)$$

where I is a luminance channel of the I-frame, and $\otimes$ represents the convolution operation.

→

h_d, $d \in \{x, y\}$, represents the Gaussian partial derivative filter along the horizontal or vertical direction respectively.

- 2021 10 12 / Mastering Uncertainty in Mechanical Engineering /
3030783537.pdf

Pag. 37(50) → Gauss

A scaled Gaussian Process is used in for modele (model) calibration and prediction.

Pag. 49(62) –

Gaussian uncertainty propagation is not adequate.

Pag. 50(63) – Fig. 3.4. ↙

(a) Gaussian uncertainty propagation shown to be associated to a coordinate transformation or mapping $\theta_1 \to \theta = h(\theta_1)$;

(b) approximation $\bar{\theta}_g \approx \bar{\theta}$ due to the nonlinearity in $h(\theta_1)$.

(estudar os gráficos).

data
S T Q Q S S D

a) $\theta = h(\theta_1)$

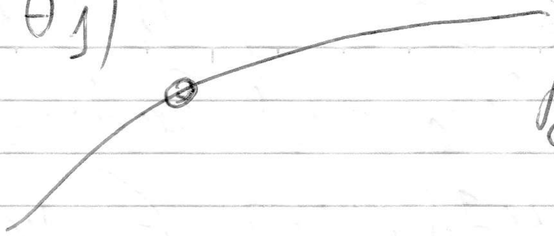

faz uma curvatura.

Pag. 147(159) - Gauss-Newton approach

Pag. 181(193) - Gaussian kernel

$$K(t) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} t^2) \Leftarrow$$

— 20211012/ Nonlinear Differential Equations...

Nonlinear Differential Equations...pdf

Pag. 10(11) - Gaussian distribution

Pag. 6(307) - 2.1 Numerical approximation

Pag 7(308) - Gauss-Legendre

Galerkin method using the quadrature rule of Gauss-Legendre (as present in [8].

Pag. 8(309) -, using

Gauss-Legendre quadrature twice yields.

(Bordas).

Jandaia

_ 20211012/ Nuclear ...|

Nuclear and Radiochemistry... pdf

Pág. 471 (488) Figure 12.93

The Gaussian curves were obtained by fits taking into account the yields for a total of about sixty detected nuclides, ...

—

Gaussian distributions – pages
209, 277, 278, 396, 409, 423, 424, 439, 449, 454,

_ 20211012/ Vertical External Cavity Surface Emitting Lasers/

Vertical... pdf

Pag. 111 (124) 4.2

The "Ideal" Laser (Space)

→ Laguerre-Gauss or
Hermite-Gauss basis

↙

Pág. 131 (147) 4.10 Spatial Coherence

We can consider that a real-world Gaussian TEM_{00} mode is an ideal spherical Gaussian wave, wavefront of which is polluted by some "spatial phase noise" in the transverse dimension of the wave, as illustrated in Figure 4.14.

- 20211012/ Semiconductors for Optoelectronics.../

Semiconductors for optoelectronics...pdf

Pág. 225(233) — Gaussian distribution (7.19

- 20211012/ Mathematics and the physical world/

Morris Kline - Mathematics in Western Culture (1964).pdf

Pág. 396(425) —

Gauss expressed a horror of infinite quantities.

"I protest against the use of infinite magnitude ... which is never permissible in mathematics".

data
S T Q Q S S D

Pag. 410 (439) XXVI

New Geometries, New Worlds.

Pág. 412 (441) - Saccheri

Euclid's parallel axiom must be right.

Book: Euclid Vindicated from all Defects. (1733)

"Apparently one man set out to vindicate another he will most likely do so regardless of the facts".

Gauss, Lobatchevsky and Bolyai (Pág. 413 (442)).

Undoubtedly thes (this) change accounts for the fact that three (men, Gauss, Lobatchevsky and Bolyai) with no knowledge of each other's thoughts, discovered the correct interpretation of Saccheri's work (at about the same time

Jandaia

Pag. 413(442)

Gauss was extremely modest. He said,

"If others would but reflect on mathematical truths as deeply and as continuously as I have, they would make my discoveries".

Those who believe that genius is 99 per cent perspiration as well as those who despair of their mathematical abilities may find comfort in Gauss's statement.

Gauss was still a youth when the problem of the parallel axiom first came to his attention.

Pag. 414(443)

Gauss's work on non-Euclidean geometry was found among his papers after his death.

Pag. 418(447) Gauss suggested a criterion for determining the truth of Euclidean Geometry.

The sum of the angles of a triangle equals 180° in this geometry but is less than 180° in the new geometry.

_ Morris Kline .. Nonmathematician (1985).pdf

Pág. 456(473) Gauss

He pursued the logical implications of a system of axioms which included the assumption that more than one parallel to a given line passed through a given point, and thus created non-Euclidean geometry.

Pág. 542(559)

Some of the axioms of Euclidean geometry are that two points determine a line, all right angles are equal, and the axiom on parallel lines.

The non-Euclidean geometry of Gauss, Lobatchevsky, and Bolyai contains the same axioms as Euclidean geometry does, except for the parallel axiom.

Look: $\sqrt{2}$ /// (estudar!!!).

_2021 1012/Mathematics.../ Geometry of Vector Sheaves

data

..1(1998).djvu

38 occurences - Riemann

- Geometry of Vector Sheaves.. II.pdf

65 occurences - Riemann.

Pág. 15 - Riemannian A-metric, for every Riemannian vector sheaf, then the corresponding differential triad.

- Morris Kline - Storia del pensiero matematico - I (1999).pdf

Pág. 333(361) - 14

La origini della geometria proiettiva

- Morris Kline ... II (1999).pdf

Pág. 1004(291) XXXVI

~~La~~ La geometria non euclidea

Pág. 1008(295) Figura 36.2

Imagem:

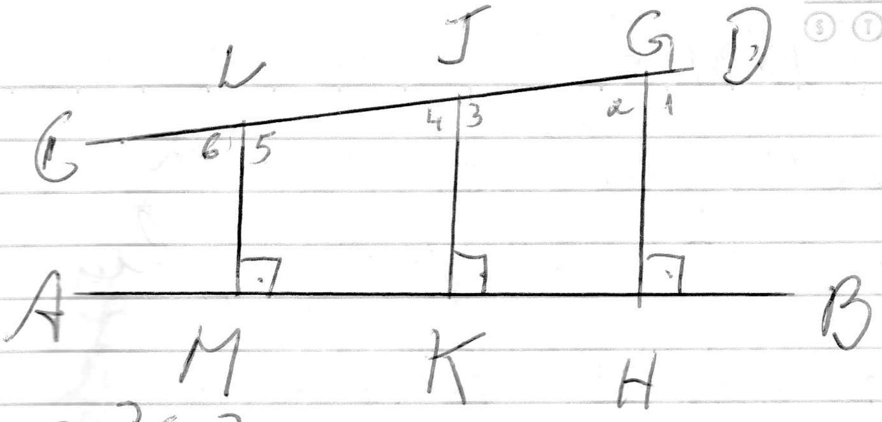

Figure 36.2

⇒ essere asintotiche.

Nasîr-Eddin (1201-74), l'editore di Euclide, diede anche lui una « dimostrazione » del postulato euclideo delle parallele

assumendo che due rette non parallele si avvicinino l'uura all'altra in una
(l'una)
direzione e si allontanino nell'altra.

Pag. 1021 (308) -

Pag. 1020 (307) ⇒

Nel settimo capitolo Lobatchevsky respinge bravamente l'assioma euclideo delle parallele e fa la seguente assunzione:

→ aborda a questão do π.

Pág. 1021 (308)

$$tg \frac{\pi(x)}{2} = e^{-x}$$

_ Morris Kline - Calculus (1998).pdf

_ Anastasios Mallios - Differential Sheaves and Connections (2015).pdf

---

_ Morris Kline - Mathematics and the Physical World. (1981).djvu

83 occurrences - Gauss.

_ Morris Kline - Mathematical Thought from Ancient to Modern Times (1972).djvu

Pág. 261 (277) - When, in 1801, Gauss adopted x^2 for xx, the former became standard.

_ Morris Kline - Mathematical thought from Ancient -
Volume 2 (1972). djvu

Pag. 424 (52) -

$$[\text{ Gauss let } \Pi(n) = \Gamma(n+1)].$$

Pag. 598 (226) -

In the same paper Gauss proved that nth degree polynomial can be expressed as a product of linear and quadratic factors with real coefficients.

Pag. 609 (237)

Fermat conjecture on

$$x^n + y^n = z^n$$

Euler proved this was correct for $n=3$ and $n=4$. The case $n=4$ had been proven by Frénicle de Bessy.

This work of Euler had to be completed by Lagrange, Legendre, and Gauss.

Legendre then proved the conjecture for $n=5$.

Fermat had also conjectured that the formula

$$2^{2^n}+1$$

(Chapter 13, sec. 7).

yielded primes for an unspecified set of values of n.

Obs: (completar com este obs.)

Pág. 753(381)

Alternatively, a regular polygon can be constructed if p is prime of the

$$2^{2^h}+1.$$

Obs: 2^u+1

$$u=2^h \Rightarrow 2^{2^h}=1$$

→ Ao todo, 105 occurrences (Gauss)

— Morris Kline - Mathematical Thought. Volume 3 (1972). djvu

Gauss - 273 occurrences.

Pág. 813(16) 2. The Theory of Congruences.

Pag. 870f721-

Maxwell says in his Electricity and Magnetism that Gauss's studies of magnetism reconstructed the whole science, the instruments used, the methods of observation, and the calculation of results.

Obs:
Gauss - 273 occurrences

---

www.wilbourhall.org

GAUSS SURFACES
https://math.dartmouth.edu/~doyle/docs/102.2020/sources/gauss_surfaces.pdf
→ English.

Gauss and the Definition of the Plane concept in Euclidean Elementary Geometry.
20211012/GAUSS/82058304.pdf

Gauss Werke
Göttinger Digitalisierungszentrum
https://gdz.sub.uni-goettingen.de/volumes/id/PPN235957348

—

— 20211012/ GAUSS-2/ ⟸

Obs: Todo o diretório!!!

— Curves and surfaces.pdf

407 ocorrências (Gauss)

Pág. VIII (8) Preface

Finally, in Chapter 7, we discuss a few important results about the global theory of surfaces (one of them, the Hartman-Nirenberg theorem, proved only in 1959), mainly (principalmente) focusing on the connection between the sign of Gaussian curvature and the global (topological and differential) structures of surfaces.

Obs: Livro todo!!!

- 20211012/GAUSS-2/

Differential Geometry of Curves and Surfaces.pdf

Pág. 160(178) - The Gauss Map

Pág. 189(207) Gaussian and Mean Curvature.

—

Pág. 153(171) Figure 6.3

≅ Gravity.

- 20211012/GAUSS-2/Remmert-Theory.pdf
  → Theory of Complex Functions.

- 20211012/Measure and Integration theory/Measure and Integration (2002).pdf

- Outra cópia:

  - 20211012/GAUSS-2/
    Burckel-Classical Analysis in the Complex Plane...
    (2021).pdf.

Obs: Livro todo!!!

Os diretórios GAUSS and GAUSS-2 devem ser explorados em detalhes.

Source: Data-Nov/Livro-020-2021/Books-020-2021/20211012/

≡

20211013//

orthogonal Procrustes problem

Source: 20211013/Linear Algebra with Applications.../Analysis and Linear Algebra (2021).pdf.

"Obstáculos são aqueles perigos que você vê quando tira os olhos do seu objetivo"
Orlando Drummond (Seu Peru).

"Aquilo que se faz por amor está ~~alé~~ sempre além do bem e do mal"
Friedrich Nietzsche

O privilégio de um é uma desvantagem para o outro.

Books:

20211013/Lectures on Differential Topology/Lectures on Differential Topology.pdf

Pág. 248(279) – 14.7

E-P characteristic of tubular neighbourhoods and the Gauss Map. (Map).

Pág. 249(280) – Corollary 14.10

Gauss map of the hypersurface M.

Obs: Riemann

Pág. 13 - 1.7

Tangent vector fields, Riemannian metrics, gradient fields.

$$\pi_U \circ X = id_U.$$

- 20211013/ Satellitengeodäsie../ Satellitengeodäsie...pdf

Search: geometrischen

Obs: Livro todo!!! German language.

Pág. 124 (142) 7

Transformationen

Abb 7.1. Imagem.

Pág. 126 (144) Abb 7.2

Pág. 127 (145) Abb 7.3

- 20211013/ Satellitengeodäsie/ Satellite Geodesy - Foundations..pdf

Obs: Livro todo!!!

cont.

- 20211013/Satellitengeodäsie/

Obs: Diretório todo!!!

- 20211013/ Geodesy/
  - Gunter Seeber - Satellite ... (2003).pdf
  - Physical Geodesy ... (2005).pdf
  - Handbook of Mathematical Geodesy ... (2018).pdf
    938 pages.
  - Geodesy for Planet Earth ... (2012).pdf
    991 pages.
  - Mathematische Geodäsie - Mathematical Geodesy.pdf
    1459 pages.

Obs: Livro todo!!!

Obs: Diretório todo também!!!

= 20211013/ Geodesy/...

= Tom Herring - Geodesy..(2009).pdf

Pág. 14 (20) Figure 2

Ellipsoidal coordinates

Pág. 15 (21) Figure 3

Relative geometry of geoid and ellipsoid.

→ orthometric height

Pág. 14 (20)

their relation to Cartesian coordinates is given by

$$x = \sqrt{u^2 + E^2} \sin\delta \cos\lambda$$

$$y = \sqrt{\mu^2 + E^2} \sin\delta \cos\lambda \qquad (2)$$

$$z = \mu \cos\delta$$

Pág. 19 (25)

Gauss' divergence theorem (analogous to Stoke's theorem, eqn [19]),

$$\iiint_V \nabla \cdot f \, dv = \iint_S f n \, ds \qquad [26]$$

data

Figure 6 - Geometry for special case of Green's third identity.

We can apply Gauss' divergence theorem to get Green's first identity, [28]

Pág. 440 (441) -

Gaussian-distributed zero-mean random variables with joint probability density function

$$p_{RI}(r, i) = \frac{1}{2\pi\sigma^2} e^{-(r^2+i^2)/2\sigma^2} \quad [120]$$

- Methods of Celestial Mechanics - Volume I .. (2004).pdf
- Methods of Celestial Mechanics - Volume II .. pdf

obs: Estes dois volumes existem em formato .djvu.

- 20211013 / Geodesy /

Herring, Thomas - Treatise on Geophysics - Volume 3 .. (2015).pdf

obs: Livro todo!!!

Pag. 377 (370) – Figure 36

Illustration of forward and backward squint. (extra(two)).

(important !!!)

Obs: Look [81] sinc $\frac{\pi}{\Delta R}$ ↙

– 20211013 [Lectures on Differential Topology / Lectures on Differential Topology.pdf

Pag. 249 (280)

Corollary 14.11

Let X be an oriented hypersurface of $\mathbb{R}^K$.

(1) If K is even, then $\chi(X) = 0$.

(2) If K is odd, then

$$\chi(X) = 2\deg(g_X). \text{ Where}$$

g_X is the Gauss map of X.

data
S T Q Q S S D

Pág. 2 (33) →

$\mathbb{R}^n$ is a topological space endowed with the topological $T = T_n$ induced by the distance d_n. As for any metrizable topological space, a subset U of $\mathbb{R}^n$ is open (i.e., it belongs to T) if and only if for every $x \in U$ there is $r > 0$ such that (de tal modo que) the "open" n-ball of centre x and radius r

$$B^n(x, r) := \{y \in \mathbb{R}^n ; d(x, y) < r\}$$

Pág. 21 (52) 1.9

Topologies on spaces of smooth maps

Pág. 22 (53)

classical Store-Weierstrass theorem (see Store).

Pág. 51 (82) - Corollary 3.3

- Real projective spaces
- Complex Stiefel and Grassmann manifolds.

$\mathbb{R}^n$ with $\mathbb{C}^n$.

Complex linear subspaces of $\mathbb{C}^n$.

Hermitian product on $\mathbb{C}^n$

orthogonal groups $O_{(n)}$

Hermitian ~~spaces~~ matrices.

spectral theorem for complex Hermitian matrices

—

- 20211013/ Geometry the Line and the Circle/ Geometry - The Line and the Circle (2018).pdf
- 20211013/ Squaring the circle.../ Ernest Hobson - Squaring the circle the circle ... (1988).pdf

Editorial:

= Ernest Hobson - Squaring ... djvu

Pág. 12(18) - Pi, being transcendental, ...

- Ernest ... - the theory of functions ... (2008).djvu

—

20211014///

superlativo
prisioneras

—

Advogado constitucionalista

Freiburg. Dietrich Murswiek

—

Books:

— 20211014/ Essential University Physics - 2 Volumes/

_Essential_University_Physics_Volume_1.pdf

_Essential_University_Physics_Volume_2.pdf

Pág. 405(40) Chapter 21 Gauss' Law

Pág. 413(48) Example 21.1

Gauss's Law: A Hollow Spherical Shell

Look: Pág. 414(49) - pictures Sphere

Charges at B contribute $\vec{E}_B$ to the field at P...

and charges at A contribute $\vec{E}_A$.. the two fields cancel.

Pag. 416(51)

"we found that the left-hand side of Gauss's law - the flux - is given $4\pi r^2 E$ whenever we have spherical symmetry and are using a spherical Gaussian surface.

Obs: Livro todo!!!

_20211014/AMS/

Bonahon - Low-dimensional geometry ... (2009).djvu

Pag. 130(141) Exercise 5.16 (the Gauss-Bonnet formula).

Obs: Search: Riemann 24 ocurrences.

_20211014/AMS/

Convexity...(1963).djvu

Pag. 239(251)

(iii) In E^3 the point $k(K)$ is the centroid of the Gauss curvature of the boundary of K (Gericke [2]).

20211014(cont.)///

the geometric properties of k(K) in higher dimensions do not seem to have seen investigated.

Pág. 161(173) – 9.9 Metric convexity

A theorem of Whitehead [1] asserts that for any point p of a Riemannian manifold with positive definite metric, all sufficiently small "spherical" neighborhoods of p are metrically convex in a rather strong sense.

This result was sharpened by Nijenhuis [1]. (Nijenhuis)

–20211014/AMS/

Mathematics Frontiers and perspectives/

Mathematics-Frontiers...(2000).djvu

Pág. 354(364) – A.

Isometric embedding of surfaces.

An important well-known question is to characterize those ~~metric~~

intrinsic metric on a surface

which can be realized as embedding into three space.

Minkowski made the first major progress on this problem by proving that any convex polyhedron can be so realized [148].

Pag. 356 (366) B. Different geometries

One can study properties of surfaces that are invariant under a bigger group than the orthogonal group.

There are the special linear group, the conformal group and the projective group each acting on $\mathbb{R}^3$.

For affine geometry important invariants are the concepts of the affine metric (which is the second fundamental form divided by the fourth root of the GAUSSIAN curvature)

and the affine normal.

John Loftin discussed how to construct an affine sphere by a holomorphic section of the tricanonical line bundle of a Riemann surface.

_20211014/ Calculus in 3D Geometry ... /

—

- Calculus in 3D Geometry, Vectors, and Multivariate ... pdf

Pág. 321(332) - rodapé 17

Gauss's theorem or the Gauss's-Ostrogradsky Theorem

- Geometry - the Line and the Circle (2018).pdf

Pág. 248(269) Figure 11.34

Proposition IV. 16

In a given circle to inscribe a fifteen-angle figure which shall be both equilateral and equiangular

Theorem 11.15 [Gauss's Theorem]

38 ocurrences.

—

- 20211014/ Krishna's - Differential Equations...pdf

Pág. 183 - Unit-3

Chapter 5

Gauss Hypergeometric Equation.

20211015/// (20211014)/// Cont.

→ 20211014/Circle/

Circle in a box ... (2009).pdf

→ Gauss – 45 occurrences

– Renormalization and 3-Manifolds ... (1996).pdf

Pag. 46(52) –

The injectivity radius of X is bounded above in terms of S (because its area is determined by Gauss-Bonnet), so we obtain an upper bound on the injectivity radius in the convex core of M.

– New Approaches to Circle Packing in a Square ... pdf

– Barry Simon – The $P(\phi)_2$ Euclidean .. (1974).pdf

Gauss – 123 occurrences

Pag. 15(37) Definition

Let H be a real Hilbert space

$\sqrt{2}$ Gaussian random process
index

-

_ Orthogonal Polynomials on the Unit Circle - Part 1 ...(2009).pdf

Pág. 17(44) - (1.2.43)

One place that the Christoffel function enters is in the Gauss-Jacobi or Gauss-Christoffel quadrature formula that

→ If a polynomial of degree 2n

Obs: Gauss - 32 occurrences

_ Orthogonal Polynomials on the Unit Circle - Part 2 ...(2004).djvu

Livro ref.

Pág. 991 (548)

[202] E. B. Christoffel,

Über die Gaussische Quadratur und eine Verallgemeinerung derselben,
J. Reine Angew. Math. 55 (1858), 61-82 (Cited on 24, 684).

_ The Infinite in Mathematics ...(1978).pdf

-

20211015

1 5'00
40
145
95
780
75
855
120
98
1073
(Livros)
1073
130
1203

Brestova.

Edentulismo, perda dos dentes.

—

Applications of Harmonic Analysis
Vilenkin

Integral Geometry and Representation
Theory - Vilenkin

—

Books:

- 20211015/ A History of Geometrical Methods/ Julian... A Treatise on algebraic plane curves ...(1959).pdf

- 20211015/ Integral Measure.../ Integral, Measure ... pdf

- 20211015/ Calculus of Variations/ Robert... Calculus of Variations...pdf

- 20211016/ Cálculo/

- Ecuaciones Diferenciales Elementales Rainville...pdf

Pag. 95 Capitulo 5

Trajectorias ortogonales

Así una curva que corta otras formando ángulos rectos, se llama trayectoria ortogonal de aquellas otras.

Pág. 98 24. Trayectorias ortogonales, coordenadas polares

$$\tan \psi = r \frac{d\theta}{dr}$$

Veja: Figura 11.

_20211016/ Cálculo/

Ecuaciones en derivadas parciales con series de Fourier y problemas de contorno...pdf

Obs: Olhar livro todo!!

Interessante

_20211016/ Cálculo/

Ecuaciones diferenciales...pdf

Gauss - 20 occurrences

Obs: Livro todo!!!

_20211016/ Cálculo/ Ecuaciones

diferenciales y problemas con valores
en la frontera ...pdf

Obs: Estudar este livro!!!

20211016///

Books:

- 20211016/ Differential Equations... /
Differential Equations and Linear
algebra..pdf

Gauss - 42 ocurrences

Pág. 160 3.2
Matrices and Gaussian Elimination

Pág. 171 Gauss-Jordan elimination
→ Algorithm Gauss-Jordan
Elimination

Pág. 229 (using Gaussian elimination
or Cramer's rule)

Pág. 233 Subspaces of $\mathbb{R}^3$

Pág. 636 The Legendre Equation
→ Gaussian quadrature

—

Pág. 654 42.

This problem is a brief introduction to Gauss's hypergeometric equation.

_20211016/ Traite de Geometrie Descriptive Javary/

000066965_02.pdf

Obs: Há os 2 volumes.

781 pages.

Obs: /Books-others/

_traitdegom01breiuoft.pdf
02
03

3 volumes - Traité de Géométrie Descriptive - Breithof

_traitdegom01rouuoft.pdf
02

2 volumes - Traité de Géométrie Rouché

_20211016/ Grigori-Perelman-Poincaré/

Obs: Diretório todo!!!

_20211016/ Grigori.../ →

The Mathematics of Infinity
A Guide to Great Ideas ... pdf

Pág. 92(109) – Gauss
Gauss discovered the following identity

Obs: /Hilbert's Infinite Hotel

Pág. 108(125)
sum as the identity

Pág. 233(250)
little Gauss on page 92,

_ 20211016/ Grigori...../
Visions of Infinity the great mathematical problems ....pdf

- Pág. 40(53) Chapter 3
The puzzle of pi
Squaring the circle.

Obs: Gauss – 58 occurrences

Pág. 22(34) – Gauss introduced modular arithmetic in the Desquisitiones Arithmetical

Pág. 23(36) -

In 1956, the famous logician Kurt Gödel, writing to John von Neumann, echoed Gauss's plea - (primer)

_20211016/ Cálculo/ matesimp.pdf

Obs: Matemáticas Simplificadas

1641 pages.

⇒ Livro todo!!!

_20211016/ Einstein/

Besse - Einstein Manifolds ... (1987).pdf

Obs. Gauss - 34 occurences

Pág. 02(13)   0.5 Introduction

Important:

For surfaces ($n=2$) the answer to the above question is known.

The best Riemannian structures on a given compact surface M are those of constant curvature.

In dimension 2, there is only one notion of curvature namely the Gauss curvature which its a function $K: M \to \mathbb{R}$. If M is embedded in $\mathbb{R}^3$

with the induced natural Riemannian metric the Gauss curvature is the product of the principal curvatures, or the inverse of the product of the principal radii (radii) of curvature. ...

Pag. 36 (47) 1.62 theorem

Given a Riemannian manifold (M, g), the function d is a distance on M, and the topology of the metric space (M, d) is the same as the manifold topology of M.

⇒ A corollary of the Gauss Lemma 1.46 is that the distance is realized (at least locally) by geodesics. More precisely.

Pag. 103 (114)

In other words, we have computed the curvature of 2-dimensional totally geodesic subspaces in Gaussian coordinates.

We can obtain the rest of the relevant sectional curvatures by observing that on each

ρ = constant hypersurface

$\mathbb{R} \times \{p\} \times S^2$ the induced metric is the product metric

$g_p = F^2(p)\, dt^2 - G^2(p)\, d\sigma^2$

...

Pag. 126(137) – 4.41 Remarks

If dim $M = 2$

a function f on M is the gaussian curvature of some metric if and only if f is positive somewhere (resp. is identically zero or takes both signs, resp. is negative somewhere).

F. The Manifold of Metric with Constant Scalar Curvature

Pag. 439(450) – 16.18

Thus each of the following conditions implies that the Riemannian manifold (M, g) has pure curvature operator:

(i) dim $M=2$, or $W=0$ (e.g., dim $M=3$, cf. 1.119);

(ii) (M, g) is a hypersurface in a space of constant curvature (cf. the Gauss equation (1.72 c));

(iii) (M, g) is a Riemannian product of manifolds with pure curvature operator.

Moreover, it is obvious that

(iv) the property of having pure Weyl tensor is conformally invariant.

Pág. 459(470)

On the other hand, the Gaussian curvature, K, of a surface

$u : \mathbb{R}^2 \Rightarrow \mathbb{R}$ in $\mathbb{R}^3$

(in the notation $u_x = \partial u/\partial x$ etc)

satisfies (12)

$$u_{xx} u_{yy} - u^2_{yy} - K(x, y)(1 + u_x^2 + u_y^2) = 0,$$

which is not quasilinear, it is fully nonlinear.

= 20211016 / Einstein /

Quantum Relativity - A Synthesis of the Ideas of Einstein and Heisenberg

$\Rightarrow$ Pág. 241 (256) - $\Delta P \gg \frac{\hbar}{a}$, $\Delta Q \gg a$, the coherent states are pratically orthogonal.

Pág. 392(403) -

→ transformation function ⟨p|x⟩

We express its matrix element as a Fourier integral by inserting the spectral projection [p] for momentum, keying the transformation function ⟨p|x|⟩ from position eigenvectors |x⟩ to momentum ones |p⟩.

$$\langle p|x\rangle = \frac{1}{\sqrt{2\pi}} e^{ipx} \qquad (6).$$

... this integral is the standard Gaussian, integrated along a line in the complex plane from the real axis by π/4:

$$\int dP\, e^{-iP^2 \delta t/2} = \sqrt{\frac{2p}{i\delta t}} . \qquad (10)$$

Pág. 407(418) 13.2 Gauge Physics

_2021/10/16/ Einstein/

The Many Faces of Maxwell, Dirac and Einstein Equations ...pdf

Obs: Riemann → 120 occurrences ↙

Gauss → 0 (???) estudar!!!

Pag. 117 (129)

4.3 Differential Geometry in the Hodge Bundle

4.3.1 Riemannian and Lorentzian Structures on M.

Pag. 124 (136)

4.4 Classification of Geometries on M and Spacetimes

... For each metric tensor defined on the manifold M there exists one and only one connection in the conditions of (4.133).

It is called Levi-Civita connection of the metric considered, and is denoted by D.

Definition 251 Given a triple (M, g, ∇):

→ Riemann-Cartan space (4.131).

4.4.1 Spacetimes

oriented Lorentzian manifold structures (M, g, τ_g), which play a key role in physical theories.

Pág. 129(141).

4.5.3 Dirac Operators

∇ defining a general

Riemann-Cartan-Weyl geometry

$(M, \mathring{g}, \nabla)$.

20211017_{III}

_20211016/New Perpectives in Algebra, Topology and Categories|3030843181.pdf

Pág. 190(199)

3.9 Classical and Generalized (Point-Free) spaces

"sobrification" of a space →

"spatialization" of a locale

Pág. 183(192) – 3 Frames and Spaces

3.1 The Category of Frames ↙

Similarly we have coframe

HOMOMORPHISMS

Books:

1) _20211017/Dictionary of Distances/
Dictionary of Distances...pdf
Pág. 73(90) - 6.4
Non-Euclidean Geometry
obs.: Limo todo!!!

Gauss - 24 occurrences.

Pág. 112(129)

For surfaces, two kinds of curvature are considered: Gaussian curvature, and mean curvature.

The Gaussian curvature is $K = k_1 k_2$ (it can be given entirely in terms of the first fundamental form).
The mean curvature is

$$H = \frac{1}{2}(k_1 + k_2).$$

Pág. 113(130) - Metric of positive curvature
Gaussian curvature - Metric of negative curvature

Pag. 117(134) - Sphere Metric

$ds^2=r^2d\theta^2+r^2\sin^2\theta\,d\theta^2$.

A sphere of radius r has constant positive Gaussian curvature equal to r.

2) _20211017/ Quantum theory of Solids/ Han.. A Modern Course in the Quantum theory of Solid...pdf

3) _20211017/ Krishna's - Engineering Mathematics III ..pdf

Obs: Gauss - 68 occurences.

Pag. 677 Chapter 15

Solutions of System of Linear Equation

Gauss-Seidel Method to solve a system of Linear Equations.

Pag. 684-15.3

Gauss-Seidel Method

4) _20211017/ Mathematical Modelling.pdf

Jandaia

—

Pág. 234(259) - 4 Maxwell equations

Gauss law $\text{div}\, D = 4\pi \rho$,

_20211017/ Geometry/

Geometry of Lengths, Areas...pdf

Pág. 27(40) - Chapter 3

Areas by Slicing and Scaling

Wallis used basic calculations of calculus to discover a beautiful product formula for π.

Fourier used Wallis's formula for π in solving a special ~~product~~ ~~for~~ case of the heat equation which he had developed.

_20211017/ Geometry/

cmip08~~c~~.pdf (Arithmetic Geometry)

Pág. 143(150) 2.6

Relations with noncommutative geometry and a real analog of p-adic integration.

⇒ Pág. 144(151) This viewpoint was discussed in [MM02], see also [Mar05], and in particular the GAUSS problem on the distribution of continued fractions and its generalizations were ~~thu~~ treated as a measure theory on the "non-commutative modular curves".

Pág. 454(461) -

through the Gauss-Manin connection and the Kodaira-Spencer map.

-

- 20211017/ Geometry/

Projective Geometry (Oxford).pdf

Pág. 15(28) 2 Introduction

the Desargues' and Pappus' configurations

- collinear
- concurrent

Obs. Livro todo!!!

- 20211017/ Geometry/

Classical Geometry Euclidean, Transformational, Inversive, and Projective. pdf

Obs: Gauss - 07 occurrences.

—

Pág. 203 (218) 13 e 14

14 → (a) List all the regular n-gons with $n \leq 100$ sides that are constructible with straightedge and compass.

(b) Use Gauss' Theorem to prove that an angle of 20° is not constructible.

(c) Use Gauss' Theorem to decide whether or not an angle of 6° is constructible

Obs: Livro todo!!!

Veja Transformation.

—

— 20211017 / Geometry / Thinking Geometrically... pdf

Obs: Gauss - 59 occurrences

Livro todo!!

Pág. 20 (45) Gauss actually constructed a regular 17-gon in 1796.

Pág. 153 (178) — 4.1.1 ⇙

The Advent of Hyperbolic Geometry

Characteristic Axiom of Hyperbolic Geometry

Given a line K and a point P not on K, there are at least two lines on P that do not intersect K (Fig 4.1).

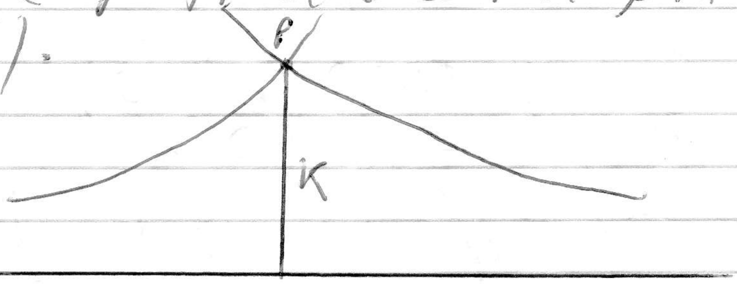

Various consequences from this change, including the many that Saccheri found.

The most startling is the theorem that the measures of the angles of a triangle do not add up to π (or 180°), as they do in Euclidean geometry.

Pag 154 (179) theorem 4.1.1

In hyperbolic geometry the difference,

$$\pi - (m\angle A + m\angle B + m\angle C)$$

between π and the angle sum of a triangle is proportional to the area of triangle.

Pág. 160 (185) 4.1.4

Carl Friedrich Gauss

Gauss → regular 17-sided polygon

Obs: Exercises 4.1.9.

A common property of hyperbolic geometry is that two lines do not intersect and are not sensed parallel (described informally in Exercise 4.1.6) have a common perpendicular.

By 1800 Gauss had become convinced that Euclid's parallel postulate could not be proven, and in installments he developed what he called non-Euclidean geometry.

IMPORTANT:

Pág. 410 (435)

Gauss → they link the area of a triangle in spherical and hyperbolic geometry to their angle sums.

In the generalization, Gauss used geodesics for the sides of the triangles. Then, as he proved, the difference

Pag. 411 (436) cont.

between the angle sum of the triangle and π was determined by the area of the triangle and the curvature within it.

The sphere has constant positive curvature and the hyperbolic plane has constant negative curvature, accounting for the difference in theorems 1.5.3 and 4.1.1.

Obs: Página importante. ↙

Einstein
Riemann
Gauss

In 1854, Georg Riemann (1826-1866) generalized Gauss's results to any number of dimensions and so transformed differential geometry again.

Many mathematicians developed Riemann's insights into rigorous mathematics with computational methods.

Albert Einstein (1879-1955) built on Riemann's profound insights to revolutionize physics with the general theory of relativity.

Einstein needed a four-dimensional space built from three dimensions and time, all with varying CURVATURE corresponding to the strength of gravitational fields at different points.

Pag. 424 (449) 9.3

Surfaces and Curvature

From Gauss's definition of curvature, a sphere has constant positive curvature and a Euclidean plane has 0 curvature.

We often use the formula

$$\cos^2(\theta) + \sin^2(\theta) = 1$$

to simplify formulas for surfaces of revolution.

9.3.1 Surfaces

$$z = \pm\sqrt{R^2 - x^2 - y^2}$$

determine the upper and lower hemispheres.

Pag. 427 (452) → Página importante

M is called the shape operator.

Pag. 436 (461) - 9.4.3

Arc Length and Surfaces

Pág. 437 (461) —

Figure 9.31 Shapes in the plane

Figure 9.32 Corresponding shapes on a cylinder.

Important:

→ Figure 9.33 A differential on a surface.

Obs: Estudar livro todo!!!

— 20211017/ Geometry/

Perspectives on Projective Geometry.pdf

Obs, Livro todo!!!

Estudar com o livro anterior e

Pappus' theorem.

→ Gauss - 31 ocorrencias.

— 20211017/ Geometry/

Vector analysis versus vector calculus.pdf

Gauss - 9 occurrences.

Pág. 314 (329) - ⇙

Theorem 9.6.1 (Gauss's law)

Let U be a regular 3-surface with boundary of class C^2 in $\mathbb{R}^3$ and assume that

$$S := \partial U$$

is a compact regular surface that is oriented to the normal vector pointing to the exterior of U.

(1) If $P \in U$, then the flux of F across S is zero.

(2) If $P \in U \setminus S$, then the flux of F across S is $-4\pi G M m$.

- 20211017 Geometry / Vector Algebra I Scalar and Vector...pdf (nome errado = ao livro seguinte)

Obs: Gauss - 45 occurrences.

Do mesmo autor: Klaus Weltner

- Mathematics for Physicists and Engineers.pdf

Obs: Gauss - 49 ocurrencias -

Mais 2 volumes:

= Mathematik für Physiker 1.djvu
= Mathematik für Physiker 2.djvu

_20211017/ Geometry/

Yaglom - Challenging mathematical problems with elementary Solutions (1987).pdf

Yaglom - Challenging mathematical ... 2.pdf

Obs

Volume 1 -> integers

Volume 2 -> Four formulas for π (Pág 22)

_20211017/ Geometry/

Important-Vector/ Vector Calculus (2015).pdf

Obs: Gauss - 125 ocurrencias.
Livro todo!!!

Veja: Todo diretório: Important-Vector

20211017 /// (20211017) // cont.

- 20211017 / Geometry / Important Vector / Groups, Matrices, and Vector Spaces..(2017).pdf

Pág. 126 (136) - 5.2.7

A beautiful formula: Lewis Carroll's identity.

⇒ theory of determinantes.

- Linear Algebra and Vector Calculus (211001/5) ... (2017).pdf

Obs.: Gauss - 57 occurences.

Pág. 570 (570) - Example 11 (chapter 7)

$$\bar{F} = x\hat{i} - z\hat{j} + 2y\hat{k}$$

displacing the particle along the triangle OAB

Fig 7.8 (Pág. 571)

Obs.: Livro todo!!!

- 20211017 / Geometry / Important - Vector / a) Calculus ... (2017).pdf

Obs. Pág. 207 Chapter 3

Curve Sketching

b) Engineering mathematics ..(2010).pdf

Obs: Gauss – 28 occurrences.

c) Mathematics-1 ... (2018).pdf

Obs: Gauss – 33 occurrences

Pág. 149 – 4.2

Volume using cross-sections

Pág. 154 – 4.3

Length of Plane Curves

d) Advanced Engineering Mathematics ... (2018).pdf

Pág. 163 - Chapter 3

Ordinary Differential Equations and Applications

Obs: Pág 163 – 428.

Obs: other: Fourier Series
Laplace Transforms
Partial Differential Equations and Applications

20211018/III

Número racional decimal
exato.

baffled, (Mystified) > enganado (deceived)?

puzzled or mistaken (≅

etimológicos ou psicológicos?

→ buck@ksu.edu

burckel@math.ksu.edu

Books:

20211018/

a) Single Variable Calculus Early
Transcendentals/
Single Variable Calculus...pdf

Pág. 257(292) Figure 5 - Differentials

Obs: Gauss - 11 occurences

Riemann - 76 occurences

Leibniz - 46 occurences

Important:

⇘ Pág 737 (772)

data
S T Q Q S S D

97. Let a an b be positive numbers with $a > b$.
Let a_1 be their arithmetic mean and b_1 their geometric mean:

$$a_1 = \frac{a+b}{2} \qquad b_1 = \sqrt{ab}$$

Repeat this process so that, in general,

$$a_{n+1} = \frac{a_n + b_n}{2} \qquad b_{n+1} = \sqrt{a_n b_n}$$

(a) Use mathematical induction to show that

$$a_n > a_{n+1} > b_{n+1} > b_n$$

(b) Deduce that both $\{a_n\}$ and $\{b_n\}$ are convergent.

(c) Show that $\lim_{n \to \infty} a_n = \lim_{n \to \infty} b_n$.

⇒ Gauss called the common value of these limits the arithmetic-geometric mean of the numbers a and b.

Jandaia

—

Pág. 811 (846) -

How Newton discovered the binomial series

$(a+b)^K$

Newton was the first to discover the infinite series expansion of $(a+b)^K$ when K is a fractional exponent (positive or negative).

He didn't publish his discovery, but he stated it and gave examples of how to use it in a letter (now called the epistola prior) dated June 13, 1676, that he sent to Henry Oldenburg, secretary of the Royal Society of London, to transmit to Leibniz.

11.11 Aplications of Taylor Polynomials

...

Pág. 817 (852) Figure 8

Depicts a wave from the point S meeting a spherical interface of radius R centered at C.

The ray SA is refracted toward P.

Important.

Using Fermat's principle that light travels so as to minimize the time taken, one can derive the equation

$$(1)\quad \frac{n_1}{l_0}+\frac{n_2}{l_i}=\frac{1}{R}\left(\frac{n_2 S_i}{l_i}-\frac{n_1 S_o}{l_o}\right)$$

where n_1 and n_2 are indexes of refraction l_o, l_i, S_o and S_i are the distances indicated in Figure 8.

By the Law of Cosines, applied to triangles ACS and ACP, we have

$$(2)\quad \begin{aligned} l_o &= \sqrt{R^2+(S_o+R)^2-2R(S_o+R)\cos\phi} \\ l_i &= \sqrt{R^2+(S_i-R)^2+2R(S_i-R)\cos\phi} \end{aligned}$$

Because Equation 1 is cumbersome to work with, Gauss in 1841, simplified by using the linear ~~[illegible]~~ approximation $\cos\phi \approx 1$ for small values of ϕ.

(This amounts to using the Taylor polynomial of degree (1). (Clear!!!

Then Equation 1 becomes the following simpler equation [as you are asked to show in Exercise 34(a)]:

Jandaia

(3) $\frac{n_1}{s_0} + \frac{n_2}{s_1} = \frac{n_2 - n_1}{R}$

the resulting optical theory is know as Gaussian optics, or first-order optics, and has become the basic theoretical tool used to design lenses.

Obs: A more accurate theory is obtained by approximating cos ϕ by its Taylor polynomial of degree 3

(which is the same as the Taylor polynomial of degree 2)

Important:

Usar isto para calcular o valor preciso observando-se a curvatura do tempo-espaço.

Usar toda esta sistemática destas equações para simular os diversos valores da curvatura do tempo-espaço e suas implicações sobre o TTTV (Mus).

This takes into accounts rays for which ϕ is not so small, that is, rays that strike the surface at greater distances h above the axis.

In Exercise 34(b) you are asked to use this approximation to derive the more accurate equation

(4) $$\frac{n_1}{s_o} + \frac{n_2}{s_i} = \frac{n_2 - n_1}{R} + h^2\left[\frac{n_1}{2s_o}\left(\frac{1}{s_o}+\frac{1}{R}\right)^2 + \frac{n_2}{2s_i}\left(\frac{1}{R}-\frac{1}{s_i}\right)^2\right]$$

the resulting optical theory is know as third-order optics.

b) Integralrechnung....|

Balla2021_Book_Integral... pdf

Obs: Riemann → 43 occurrences.

c) Introduction to Differential Equations|

Introduction_to_Differential_Equations_...pdf

→ Gaussian elimination.

d) Philomath....|

Philomath the Geometric unification of Science...epub.

Pag. 09 Figure 2

The golden section ϕ and the π constant embedded within the dimensions of the gates of Egyptian temples.

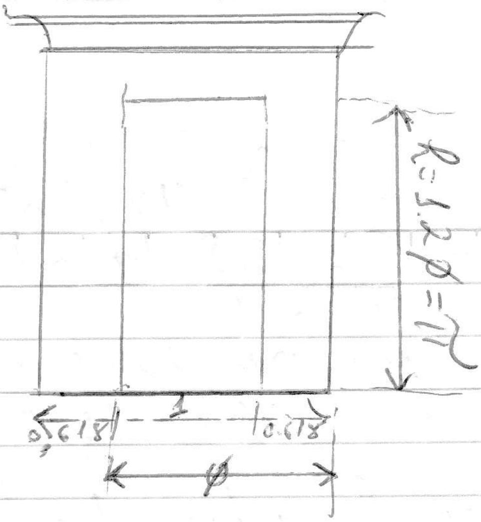

Obs: Analisar profundamente!!!

Obs: Gauss - 4 occurrences

Pag. 159 - Chapter 3 - Geometry in the D-Space
Pág. 160 - Figure 87

Pág. 175 - Chapter 4 - Semiprime Factorization: the Geometric Solutions.

Obs: Figure 100: the geometric aspects of Pythagoras' rule transforms the problem from an abstract mathematical formalism into a tangible two-dimensional Euclidean challenge.

Pág. 176 - Preserving the Area.

e) Robert B. Burckel /
Classical Analysis in the Complex Plane.pdf

Obs: Gauss - 85 occurrences.
Estudar livro todo!!!

Pág. 80 (107) (Complex) Derivative and (Curvilinear) Integrals)

...

elegant proof of the fundamental of algebra, essentially GAUSS third proof freed of reference to the transcendental functions sine and cosine (cf. also the remarks following 5.5).

Pag. 91(118) - though adumbrated by GAUSS and Poisson, the idea of the complex contour integral in 2.8 was codified by Cauchy in his famous 1825 memoir.

f) Single Variable Calculus.../
Single_Variable_Calculus...pdf

Obs: Gauss - 11 occurrences.

g) Discovering Abstract Algebra/
Discovering_Abstract_Algebra.pdf

Pag. 86(105) Chapter 12 Rings

The last part of theorem 12.9 is called the Binomial Theorem.

Example 12.12

(3) The Gaussian integers (Z) is a subring of C.
It has four units: ±1 and ±i.

Pag. 87(106) 12.2 Homomorphisms

Chapter 1 - Part 4

Volume-20---20220612

20211019 ||| Sérgio Musich

Books:

_20211019/ The Dynamics.../

a) Engineering thermofluids - thermodynamics, fluid mechanics, and heat transfer..(2005).pdf

Obs: Gauss - 24 occurrences.

Pág. 522(541) Example IVb.2.1

We can convert the surface integrals to volume integrals by using Gauss's theorem, Equation VIIc.10.28:

$$\int_V \left(\bar{\nabla} \cdot K \bar{\nabla} T - \rho c_p \bar{\nabla} \cdot \bar{V} T - \rho c_p \frac{\partial T}{\partial t} + \dot{q}''' \right) dV = 0$$

Pág. 913(931) Gauss differential

b) Low-Gravity Fluid Mechanics... (1987).djvu

Low Gravity Fluid Mechanics..(1987).pdf

_20211019/ The Most Beautiful Equation in the Universe $0=\infty=1$/

The Most Beautiful...pdf.

Pág. 291 - Gaussian plane - Complex Plane.

In this way, all physical phenomena can be depicted on the 2D complex plane (Gaussian plane) with real numbers (R) and imaginary numbers (i).

⇒ 0 = ∞ = 1 adds a third dimensional axis, "perspective" to the 2D complex plane. This is a new concept.

_20211019 / Latex /

The Latex Graphics Companion ...(2007).pdf

Pag. 278(310) - 5.12.6

Math and Verbatim Boxes

```
\usepackage(pstricks}
$ f(x) = \psframebox { x^2_3 } $
```

$\Rightarrow f x = \boxed{x^2_3}$ (print)

other command:

\displaystyle or \everypsbox{code}

Ex:

```
\usepackage{pstricks}
\psmathboxtrue
$ f(x) = \psframebox {\int_a^b \frac{x^2}{3}
                     \, dx} $
```

$$\Rightarrow \boxed{f x = \int_a^b \frac{x^2}{3} \, dx}$$ or

```
\everypsbox{\displaystyle}

$ f(x)= \psframebox{\int_a^b \frac{x^2}{3}
                               \, dx} $
```

Pág. 279(311) - 5.13

User styles and objects. (ver tudo!!!)

Pág. 295(327) - \arrows{arrow type}

Gráfico interessante (Example 5-13-33)

Pág. 296(328) - 5.14 Coordinates

PSTricks provide Cartesian coordinates

\SpecialCoor
\NormalCoor

5.14.1 Polar Coordinates

→ Section 5.2.2 on page 218

Important:

```
\usepackage{pstricks, multido}
\SpecialCoor
\begin{pspicture}

\end{pspicture}        (Example 5-14-1)
```

Obs: Pág. 218(250) - 5.2.2 Angles

Pág. 302(334) 5.14.5 Angle specifications
(\psarc)

angle of $\alpha = \arctan \frac{1}{-1} = 135$ degrees

$P = \{-1, 1\}$

(Example 5-14-8).

– American Scientist - Nov Dec 2021/
American. Scientist...pdf
Pág. 360 (42) – The Curse of $E=mc^2$.

– New Introduction to Multiple Time Series Analysis/

a) Helmut... Handbook of Matrices (1996).pdf

b) Helmut... New Introduction to Multiple Time Series Analysis...pdf

Obs.: Gauss – 130 occurrences.

c) Helmut... Structural Vector Autoregressive (VAR) ...(2017).pdf

Obs. Gauss – 205 occurrences

– 20211017/ Prime Numbers The Holy Grail of Mathematics/

Prime Numbers... epub

– 20211017/ Riemann – Geometry/

a) Philosophy of Geometry from Riemann

to Poincaré.pdf

Obs.: Gauss - 146 occurrences

Pág. 71(83) - 2.2.3

Gaussian Curvature of Surfaces

Pág. 80(92) -

Gauss' 1827 proof of his theorema egregium consists essentially in showing that the G-curvature function on $\phi(\xi)$ can be defined in terms of E, F and G.

Pág. 81(93) Gauss' treatment of the intrinsic geometry of surfaces suggests an idea which the non-mathematical reader ought (deveria) to consider carefully at this point.

Let $\phi(\xi) = S$ be a smooth surface, as before.

If x is a Cartesian 2-mapping, $x^{-1}(\xi)$ is an open, connected region of the Euclidean plane.

Pág. 88(100) 2.2.7 The Tangent Space

Indeed (na verdade), GAUSS' methods for dealing with them are historically at the root of the very notion of a differentiable manifold.

data

Pág. 101 (113) -

Riemann extended the Gaussian concept of curvature to an arbitrary n-dimensional R-manifold M by using what we may call sectional curvatures, i.e. the G-curvatures of two-dimensional sub-manifolds of M.

Pág. 151 (163) Σ has a constant Gaussian curvature equal to 0.

- Clifford-Klein problem of space-forms.

Pág. 189 (201)

J. Houel, who in the 1860's published French translation of Bolyai's Absolute Science of Space and Lobachevsky's Theory of Parallel, of Gauss' correspondence with Schumacher and the basic paper by Riemann, Helmholtz and Beltrami should have worked at the same time on a new axiom system for Euclidean geometry (Section 3.2.4).

Obs: Livro todo!!

Faz uma análise do momento em

que cada matemático estava vivendo quando de suas descobertas.

b) Differential geometry for physicists and mathematicians ... pdf

Obs: Gauss - 23 occurences.

Pág. 247 Appendix A

Geometry of Curves and Surfaces

Pág. 251 A.2.3

The Equations of Gauss and Weingarten

Obs: Livro todo!!!

c) From Riemann to Differential Geometry and Relativity (2017).pdf

Obs: Gauss - 388 occurences

Estudar profundamente este livro!!!

Gaussian curvature:

- Pág. 66 (100),
- Pág. 100 (133),
- Pág. 344 (371)

d) Geometry of the fundamental
Interactions.pdf

Obs.: Gauss - 11 occurrences

e) Riemann surfaces, dynamics, and
geometry...pdf

Obs.: Gauss - 06 occurrences

f) Handbook of Geometric Analysis
1. djvu (702 pages). (1)
2. djvu (454 pages). (2)
3. djvu (495 pages). (3)

Gauss - (1) 134 occurrences
(2) 75 occurrences
(3) 36 occurrences

g) Geometry of Cauchy-Riemann
Submanifolds...(2016).pdf

Obs.: Gauss - 86 occurrences.

Pág. 02(19) - 1.2 Preliminaries

The Gauss and Weingarten formulas

- 20211019/ Handbook of Complex
Analysis Geometric Function theory.../

a) Handbook of Complex Analysis
Geometric Function theory 1.pdf

Pág. 264(276) - 12.

Symmetrization and the hyperbolic metric

(important!!!)

{ hyperbolic metric (= Poincaré metric)
| Gaussian curvature -1.

b) Handbook of Complex Analysis Geometric Function Theory 2 pdf

obs: Gauss - 48 occurrences

Pág. 188(203) -

Gaussian curvature $k(p) = -4$

Pág. 225(240) -

Where d_Δ is the hyperbolic Poincaré metric on Δ of Gaussian curvature -4.

The fundamental Royden-Gardiner theorem states that Teichmüller and Kobayashi metrics coincide on every Teichmüller space.

Pág. 289(304) -

Next we shall characterize the parallel slit mapping $g_\theta(z)$ by the
(fenda)

Gauss-Thomson principle of minimal energy of signed charge distributions under the presence of an external dipolar field.

Pag. 291 (306)

$$\frac{D}{2\pi} \leq \lambda \leq \frac{D}{\pi},$$

where D denotes the diameter C.

$$\lim_{n\to\infty} d_{\theta n}(C) = \exp \operatorname{Re}(e^{-2\pi} a_{1,0})$$

and

$$\lim_{n\to\infty} \lambda_n = \frac{2}{\pi} d(C).$$

where $d(C)$ is the usual transfinite diameter of C.

6.7 Capacity and quasiconformal maps

As in the conformal case the Gauss principle of minimal energy can be used to give extremal characterizations for certain canonical quasiconformal mapping.

Pág. 419(434)-

Gauss quadrature is applied.

Pág. 486(501)-

Called the Gauss map, depends only on the second dilatation function a.

Indeed, we have

$n=(2\,\mathrm{Im}\{\sqrt{a}\}, 2\,\mathrm{Re}\{\sqrt{a}\}, 1-|a|)/(1+|a|)$. (6).

... Gaussian curvature of nonparametric minimal surfaces over some given domain Ω.

Pág. 493(508)-

Corollary 2.1

(1) Ω whose Gauss map has the property that its image is the upper half-sphere covered exactly once.

...

Pág. 496(511) - 2.7.

Univalent log-harmonic mapping.

Studying minimal surfaces whose

Gauss map (normal vector)
are periodic, we are lead to
univalent harmonic mappings
with periodic partial derivatives.

We may restrict ourself to periods of $2\pi i$.

Pag. 511(526) - 1.1.3

Using the chain rule for the Schwarzian derivatives

... Gaussian curvature equals -4.

We denote by $\mathbb{B}(D)$ the complex Banach space of hyperbolically bounded holomorphic functions on $\mathbb{D}$ with the norm

$$\|\varphi\| = \sup_{D} \lambda_D^{-2}(z)\,|\varphi(z)|.$$

Pag. 519(534) - 1.4

Sketch of the proof the relation

the Teichmüller-Kobayashi metric d_T on $\mathbb{T}(\Gamma)$

...

which belongs to C^2 on $\Delta\setminus\{0\}$ and has there constant Gauss curvature -4.

Pág. 629 (644) 2. Hypergeometric function

(Gaussian) hypergeometric function

Pág. 631 (646) - 2.6

Gauss' contiguous relations

Pág. 635 (650) - Remark 2.9

The asymptotic relation of Gauss (theorem 2.7 (2)) has been extended and generalized in various ways. Ramanujan established the relation [Bern2, pp. 33-34] (2.26

Pág. 672 (687) -

Gaussian curvature $K \leq 0$.

Pág. 712 (727) -

The problem appeared in cartography (see Kühnau [Kü13, p. 142]) and was solved by Gauss in 1825 [Ga].

Important:

Pág. 612 (627) → 9

Estimation in the other direction

curve family $\mathfrak{C}(t, \tau)$ defined by $z = z(e^{u+i\tau})$ with $0 \le u \le \log R$, τ fixed, $0 \le \tau \le 2\pi$, (9.1)

o . . .

Here on the right-hand side the integration is over the orthogonal projection of $\mathfrak{C}^*(t, \tau)$ with the new line element

$dS = \cos\alpha \, ds$ (= projection of ds);

α is the angle between $\mathfrak{C}^*(t, \tau)$ and the plane ϑ = const through the corresponding point.

Veja: Fig. 4 (important)

_20211019/ Latex/

a) The Latex Companion - Addison-Wesley.pdf (1115 pages)

b) The Latex Graphics Companion.pdf (974 pages)

c) Kopka - Guide to Latex...pdf (613 pages)

d) LATEX - german.epub (320 pages).

Jandaia

20211020//

Books:

- 20211020/ Riemannian Geometry in an Orthogonal Frame/

47 books - Gauss - 3737 occurrences.

---

desgratis

- 5000 years of Geometry...pdf

Pág. 410 (423) - 7.3

Theory of geometrical construction

The proof of the irrationality of $\sqrt{2}$, as indicated by Aristotle and in "Elements" (X, 1 15), could be so interpreted that the diagonal of a given square ~~root~~ cannot be transferred towards one side by means of a purely linear operation.

...

algebraic solution → this was basically accomplished by Gauss, and the actual theory of geometrical constructions started with him.

Import:

In 1593 Vieta crowned his Supplementum geometricae with the proof that all problems that lead to equations of third or fourth degree become solvable if we, apart from compass and straightedge, allow instruments with which one can extract the third root of given quantities and trisect every angle.

Did Vieta know of the geometrical interpretation of complex algebra in Gauss's "complex number plane"?
⟹ Illus. 17.3.1
↳ Pag. 411 (424) ⟸ ✡

Pag. 412 (425) —

Algebraization of the theory of geometrical construction

Gauss ⇒ due to his discovery that the regular 17-gon can be constructed with compass and straightedge. (30/03/1796).

Further entries indicate that he came back to questions of circle division again and again, reduced to purely algebraic questions by using complex numbers, for the equation $x^n = 1$ (and/or after separating the trivial factor $x-1$ from x^n-1) for which n the equation $1+x+\ldots x^{n-1}=0$ can be solved by quadratic radicals

Pág. 413(426) - If angle trisection would be generally possible with compass and straightedge, then such is also the case in which the given angle is 60 degrees ...

The third famous ancient problem, squaring the circle, leads to an investigation under the new circumstances if number π is the root of a polynomial equation solvable by quadratic radicals.

Lambert showed the irrationality of π in 1767.

...

Ferdinand Lindemann finally generalized the idea of proof in 1882, i.e., he proved the transcendence of π meaning that π cannot at all be the solution of any algebraic equation with rational coefficients.

obs: Vahlen 1911 for mathematical details and a well-formed proof of transcendence of e and π.

Vahlen 1911:

Konstruktionen und Approximationen.

Source: https://archive.org/details/Konstruktionenun00vahluoft

372 pages (German language)

archive.org/details/
abstraktengeometr00vahluoft

Relacionado: David Hilbert

archive.org/details/grundlagendergeo00hilbuoft

archive.org/details/
grunddergeovon00hilbrich

Others:

Geometry, Plane and Spherical Trigonometry

archive.org/details/elementsofgeomet00rosiuoft

Procurar:

Celebrated Problems of Geometry
Teixeira 1918

Adentrando:

Konstruktionen und
Approximationen
Theodor Vahlen - 1911

Obs: Gauss - 31 occurrences.

Important:

Pág. 01(17) Erster Teil
Lineare Konstruktionen ⇒

Projektive lineare Konstruktionen (1)

(1) ref. Traité des Propriétés Projectives Poncelet.

Source:

—20211020/ Vahlen Konstruktionen...|

— Traité des propriétés projectives des figures — Poncelet

Source:

—20211020/ Traité des propriétés...|

Return:

5000 Year of Geometry Mathematics...pdf

Pag. 480(493) — Problem 7.3.3

Impossibility of doubling the cube with compass and straightedge

Show by means of case-by-case analysis x^3-2 cannot be written as $(ax^2+b)(cx+d)$ with integers a, b, c, d. According to Gauss's general theory, it ~~(following)~~ follows that doubling the cube by means of compass and straightedge is impossible.

Pag. 481(494) Problem 7.3.5 *

Equivalence of construction problem in space

1. A point can be constructed by means of given points iff (if and only if) we can construct its three coordinates with compass and straightedge by means of the coordinates of the given points in an auxiliary plane (regarding a given spatial Cartesian coordinate system).

Problem 7.3.6*: Torricelli point of 5 points of the plane and 4 points in space respectively ↙

Gauss

Look:

Pág. 482(495) – Illus 7.10.2 ✡

Pág. 483(496) – Problem 7.4.1

Tractrix and Gaussian curvature of the pseudo-sphere.

Gaussian curvature is negative and constant.

Problem 7.5.1 → Constructions in Poincaré's model of the plane hyperbolic geometry. ‼

1. Straight connection of two points P_1, P_2 (thereby, the case that the
(deste modo)

Euclidean connecting line is perpendicular to a is trivial).

2. Marking out a line segment on a point
(Take into (levar em) account that reflections are generally realised by transformation by means of reciprocal radii at the relevant circle).

3* Construction of a circle with given centre and radii
(Use the solution to 2. as a "sub-program"!)

4* Prove that trisecting any angle cannot also be done with compass and straightedge in hyperbolic geometry by using the preservation of angles of Poincaré's model

(Solution in [Schreiber 1984], chap. 2.6).

Schreiber - Grundlagen der Konstruktiven Geometrie

Other:

- 20211020/Riemannian Geometry.../
  P. Schreiber - Theorie der geometrischen Konstruktionen (1975).pdf

Pag. 72 Quadratwurzelausdruck...

Pag. 484(497) Problem 7.9.1
Solution to Schumacher's paradox
Gauss in 1836.

_20211020/ Mathematik...|
Papula L.-Mathematik 1 (2009).pdf
Pág. 323(347) - IV
Differentialrechnung
1. Differenzierbarkeit einer Funktion
1.1 Das Tangentenproblem
→ Bild IV-1 (image)

_20211020/ Vahlen...|
4821.pdf
Konstruktion, Approximation und Eigenschaften hyperkomplexer

_20211020/ Vahlen...|
Robson_circle.pdf

_20211020/ Vahlen...|
stroehe_theorie_der_geom...pdf

Obs.: Gauss - 37 occurences.

Pág. 3 (5)

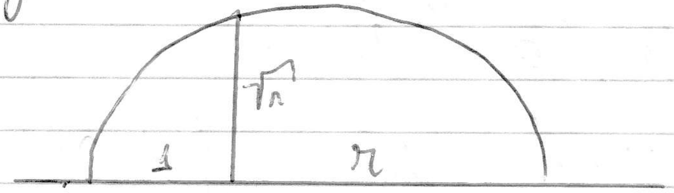

Pág. 24 (26)

Darstellung in kartesischen und in Polarkoordinaten

_20211020/ A Triangulação da Dinamarca.pdf

IMPORTANT:

_20211020/ Vahlen...|

konstruktionenun00vahluoft.pdf

Obs: Gauss - 31 occurences.

Attention:

Pág. 317 (333) Achter Teil

Irrationalität und Transzendenz von e und π.

Pág. Total: 372 ao todo!

German language

- 20211020/ Über die Geschichte des Problems - Rudio /

a/ Dr. Ferdinand Rudio - Verhandlungen .... pdf

Obs: Quadrature - 12 occurrences.

b/ Ferdinand Rudio - Der Bericht des Simplicius über die Quadraturen

Obs: Quadrature - 26 occurrences.

Pag. 32(22) - Fig. 3.

Pag. 58(35) - Fig. 6.

- 20211020/ Riemann Geometry in an Orthogonal Frame/

(cont.)

a/ Algebra and Trigonometry..pdf

b/b1-An Excursion through Elementary Mathematics, Volume I. pdf

Obs: Gauss - 11 occurrences (11).

b2- Volume II. pdf

b3- Volume III. pdf

Obs. Gauss - 50 occurrences.

20211021 /// **Books**: cont.

_20211021 / Riemannian Geometry... |

a) An Introduction to Vectors, Vector Operators and Vector Analysis...pdf

Pág. 06(33) - Earth 0.37 Gauss

Pág. 297(324) - 10.12 Surfaces

Pág. 313(340) -

10.14 - Differential operators in Curvilinear Coordinates

b) Analytical Geometry 2D and 3D.pdf

Obs.: Livro todo!!!

Pág. 201 - Chapter 13 Strait Line

Circle - 1074 occurrences.
Curve - 156 occurrences.

Important!

Começo da utilização do OpenOffice Writer para armazenar os nomes dos arquivos e textos a serem pesquisados com maior profundidade.
Source:
C:/ARQ/

20211021.odt

20211022/// Menorah

1 KIKAR DE OURO = 57.6 Kg.

De um kikar de ouro puro ela será feita- todas as suas partes (Êxodo 25:39).

18 Teffarrim (Teffarrim) 1=10cm.
altura
≅ 1,80m

12 Teffarim ≅ 1,20m

Base = Yarech

166 ml p/vela. 1/2 Log (medida)

— —

20211023///
gregehmka.com

20211024///

Geometria Analítica.
Pastor

https://archive.org/details/
geometria-analitica-j-rey-
pastor.-l-santalo-m-balanzat-1965

(1965 - 554 pags)

Mafa@2223

1525 → $\sqrt{\ }$ usado por Christoff Rudolff.
ano

= aparece pela primeira vez em
The Whetstone of Witte
publicado em 1557 por
Robert Recorde.

Source: Hist. de la Matematica Vol 2
Pag. 20(8) Pastor Rey Julio

—

20211025///

Data-NovoVivo-021-2021/
Books-021-2021/
20211025/

– Non-Euclidean-Geometry/

Obs: Todo diretório!!!

– Pappus D Alexandrie./

– Pi (π) in Nature/

Other: 20211024/ Geometria Analitica/

Obs.: Important!!!

M/A 351 965 771 994.

20211025/// (20211025///) cont.
fazer upload!

20211025/ Non-Euclidean Geometry/

Ellery Golos - Foundations of Euclidean ...(1968).djvu

Pag. 169(187) The Age of Hilbert

Pag. 170(188) - Parallel Postulate

— Shadow of Circle ... (1998).djvu

Pag. 13(21) Figure 1.11

— Henry Parker - Geometry of Four Dimensions ...(2005).djvu

Pag. 105(111) - Chapter III

Angles of Two Planes and Angles of Higher Order

I. The common perpendicular of two lines.

Pag. 221(228) - Chapter VI.

Euclidean Geometry. Figures

with Parallel Elements.

→ The development of the preceding chapters has been made independent of the axiom of parallels.

20211026 III

fazer upload!

Não sou dono do mundo
mas sou filho do dono.

Não tenho tudo o que quero,
mas amo tudo o que tenho.

Nós estamos neste mundo,
mas não somos deste mundo.

Porque os ratos abandonam o navio antes de afundar, o professor respondeu: "É porque os ratos têm informações privilegiadas".

20211027 III

20211020/ Riemannian Geometry.../

Differential Geometry of Curves and Surfaces-2.pdf.

Pág. 176 (184) - Figure 3.35

Stereographic projection is conformal because its derivative scales the green and red vectors by the same factor.

—

Pág. 194 (201) - Figure 4.1

Gaussian Curvature

—

"Quem reza se salva"

—

O universo não é tão complexo quanto nós imaginamos.

—

7-zip.org

—

stdutility.com
irfanview.com (irfanview.com)
tracker-software.com
(PDF Exchange Viewer)

—

Screenshot (Firefox).

20211028m **PRG**: ⇓

LAN: Realtek RTL8111E Gigabit.

—

libgen.st (libgen.st).
br1lib.org
z-lib.org!
pdfdrive.com
metamath.org (Norman Megill)
sanet.ws/books/
ebook3000.com/Science/
abebooks.com
pantheon.io
drupal.org

—

longfiles.com
uploaded.net
opendrive.com

—

cloudflare-ipfs.com

stdutility.com
irfanview.com

tracker-software.com
(PDF Exchange
Viewer
2.5 (210.0).

—

20211028/// (20211029)///

Comissão.

—

morefires.com (MiniPC)

—

ungarav@gmail.com

—

Properties: (OpenOffice)

Properties
Save as
Comment
Cross-reference
Date

+

Insert footnote

—

20211030///

Geometria plura praesidia
praestat.

—

20211031///

Important:

Helmholtz →

Data-NowViz-021-2021/
Books-021-2021/
20211020/
Riemannian Geometry in an
Orthogonal Frame/
Beyond Geometry Classic Papers...pdf
Pag. 7(15)-

Helmholtz comes very close to the argument Henri Poincaré will later that non-Euclidean geometry can be regarded merely as Euclidean seen in such a distorting mirror.

Galaxy Book Pro (fim)

Le Brio (filme) french 2017
Camelia Jordana e Daniel Auteuil.

20211101///

bias → tendência

boogeyman → "papão."

seeds → sementes.

ima.org.uk

20211102

Elements of Mathematics from Euclid to Gödel →

Pag. 153 (170) - Obs. Pag. 164 (181)

Figure 5.5 Isosceles triangles in a semicircle.

1816 - Ano sem Verão.

Explosão do Monte Tambora 1815, a maior erupção vulcânica em mais de 1600 anos!

Hatepe (erupção) - ano 180.

Krakatoa - 1883 (27 Agosto)

20211103

Término do Volume-001--20211021.odt.
(12h31 min de edição)

thoroughly - completamente

although - apesar

leis draconianas.

Helmholtz.

eufemismo

—

20211104M

Ver diretório Gauss/ (20211103).

→ Hidden Harmony - Geometric fantasies.

Gauss - 526 ocurrences

—

⇘ isometries: guarda da projeção.

Usar

Reflection

—

Date
Time
Bookmark
Entry
Insert Footnote/Endnote Directly
Insert Endnote Directly
Comment
Cross-reference
From File
Properties

OpenOffice:

Edit/File.

20211105m

ARQ
Data NovoVivo-021-2021
022-2021
Programas
Stelvio-SV-97

ZIP
ZIP-02

Muse - Undisclosed Desired
Music

→ Zippar.

Total: 176 Gb.

- Copiar o config do openoffice.
- arquivos de hoje.
- Vídeos (WINDOWS)
- Copiar BookMarks também (firefox).

HD

Formatar:
Iniciar - Painel de Controle →
sistema e segurança → Ferramentas adm →
click 2X Gerenciamento do Computador

Painel lado esquerdo - Armazenamento →
Gerenciamento de Disco.

Com botão direito → Formatar

Formatar - OK e OK.

—

Outro Modo: formatar HD que tem Windows instalado

WINDOWS + R

Executar

diskmgmt ⇒ OK

Gerenciador de Disco

Formatar

—

outro comando:

Formatar C:

format c:/fs:NTFS

—

—

- Copiar vídeos (Windows) com programa de imagem de HD. GHOST

- Copiar Bookmark.

—

- Copiar pasta de arquivos de hoje.

cmd.exe

Particionar HD:

Gerenciamento do Disco:

Botão Direito → Diminuir Volume

Na Partição criada

↳ clicke botão direito

↳ Novo Volume Simples

→ clicke em Avançar

↳ Avançar

↳ escolhe a letra.

Extensões do Firefox

- Easy youtube Video Downloader Express
- Flash and Video Download (Dailymotion)
- Save Screenshot
- Youtube Video and Audio Downloader
- youtube Video Downloader / Youtube HD Download.

20211106

nitty-gritty = minucioso.

rufus.akeo.ie (ISO)
rufus.ie/pt_BR/

other program:
balena.io/etcher/

Rede Doméstica Windows 7.
KU8Ay14WP5

Lesson in Geometry - Pag 129(144)
Metric Relations in a Triangle

The Problem of Tangent Circles
Pág. 231(246)

Charles Huff 0049 015121355125
charleshuff@hotmail.de

Garmisch-Partenkirchen

—

20211106/ Geometrie/
3662638304.pdf
→ Pág. 76 (85)
Abb 4.13
Zum Beweis des Satzes über
die Euler-Gerade.
→ linha Euler.
20211107///

—

www.officesuite.com

—

www.libreoffice.com

↙ instala este pacote

—

→ KB 2999226

—

orthogonal projection
→ Book Lemon Geometry. Pág 129 (144)

—

Homotetia.

Homotetia

"é a ampliação ou a redução de distâncias e áreas a partir de um ponto fixo.

Constructions → Chapter VI

Pág. 157 (172).

Homothecy →

A similarity transformation with preserves orientation.

⟹ Lesson in Geometry.

Jacques Hadamard

Pág. 129 (144).

The foot of the perpendicular dropped from a point onto a line is called the

orthogonal projection

of the point onto that line.

(para) ↳ naquela

nessa

20211108 ///

perpetrador.

rather than = ao invés de

/// ⇒ Arquivo – combinar arquivos (comando).

The Geometry of Numbers

MAA ISBN 0883856433

192 pages

Pág. 11 (28) –

Lines with Irrational Slope

---

brazil-2021@wolfram.com

---

02 -- Wiley (Geometry directory)

Pág. 541 27.13

Shortest Distance Between two Non-Intersecting Lines

—

20211109 II

Jacques Hadamard. f1

20211110 III

Pag. 128(147) - Chapter 4

Exercises on elementary geometry.

Source: Methods of Geometry.pdf

James T. Smith

$11 + \pi \cong 10\sqrt{2}$

—

indeed = na verdade

—

$11 + \pi = 14{,}141592653589793238$

$10\sqrt{2} = 14{,}142135623730950488$

═

20211111 III

Pranym → Helena

—

① LATEX: linorg.usp.br/CTAN/systems/win32/ ~~mik~~tex/tm/packages!

Google Drive.

Pesquisas

⇒ 2021-New-20210601/3-Aberto-Copy/
Arq.

⇒ Terra/

⇒ Books-2020

—

_ ARQ_BASICS/02/ the Shape of Space/
Symmetry, Shape and Space..pdf

— ARQ-BASICS/10/ Importantíssimo_
Books_38_2020/
Distance Geometry..pdf

—

20211112m

mycam.softonic.com.br

—

cyberlink.com ⇒ YouCam

—

⇒ nchsoftware.com → CamStudio

nchsoftware.com/capture/debutsetup.exe

20211113

~~Volume-001-20211112-~~
~~Volume~~

Caderno-001-20211112-400 pages
Caderno-002-20211112-179 pages
Caderno-003-20211112-309 pages
Caderno-004-20211112-396 pages

1284 pages

Horário de Digitalização (1284 pages)

202111121300 até
202111122359.

10 horas de trabalho (6).
1 hora entre café e jantar
11 horas (11 horas)

Backup HD externo → foi até

202111130010

O cansaço faz a gente errar a conta.

Refaça ou cheque as suas contas sempre descansado e de cabeça fria.

20211114

Mateus 25,41 → ao justo será dado a vida eterna.

Gálatas 5,7 → de Deus não se zomba.

$\pi \times \sqrt{2} = 4,44288293$

$28/9 = 3,\overline{1}$

Music

Armagedom — Liv Tyler
Alicia Silverstone
Aerosmith Crazy

A pontualidade é a primazia dos princípios.

---

depressão é excesso de passado.
stress é excesso de presente.
ansiedade é excesso de futuro.

$\frac{a}{b} = a \div b = ab^{-1}$

20211115

Usar omnipage para escanear !!!

Ver dia:

20110904

↳ divisão p/ dar par.

Escaneamento Antigo_001 (folhas de pesquisa)

Pág. 117

212

veja Pág 227 – 268

Pág 315 - 340 –

Pág 415 - 453

Pág. 522 –

Pág. 575 - 610

Pág. 992 - 1006

1009 - 1025

1026 (legal)

1028 - 1038

→ 1108 -

1258 TCM

1324

Volume
Antigo-001

1 - 610
0001 - 0035
611 - 1447

Dia
20211115.

Total: 1.482 pages

Total Tempo: 7h 6.

202111151206 Digitalização
202111151911

Terminei de converter em imagens e fiz backup em um HD externo.

Horário: 202111152348

20211116

Túmulo Leibniz

Rote Reihe 5
Hannover
30169
Germany

L190

data

S T Q Q S S D

Books_2020_06/OTHERS/

=> important.

20211117M

Scanear:

Antigo_002

Total pages: 117

Bookmark:

Page 01 - 20140301

Page 14 - 20150214

Page 66 - 20211019

20211117 (cont). ⇘ ✡ Escanear daqui para frente.

Teoria dos Conjuntos Causais

↳ Big Bang não deu origem ao Universo (ocorreu, mas sendo uma fase do que ocorreu no Universo).

Total de páginas escaneadas:

1284	(4 volumes)
1482	(Antigo-001)
117	(Antigo-002)
2.883	pages

Ver:

Data-NovoVivo-022-2021/20211108/Geometry.

Under Jerusalem:
The Buried History of the
World's Most
Contested City
Andrew Lawler

Mathematics2@gmail.com ⇒ 38,878.29 GB (37.97 TB).

data
Ⓢ Ⓣ Ⓠ Ⓠ Ⓢ Ⓢ Ⓓ

20211118|||

Martin Kneser

"Über den Rand von,
Parallelkörpern" 30.9.1950

---

mathematik.de

---

mathematik.de/images/DMV/
Jahresberichte/
Jahresberichte_Archiv/
Jahresbericht_01-2006.pdf

Obs: Tradução:
"Acima da borda de
corpos paralelos".

The Carl Friedrich Gauss
Prize

spektrum.de

---

gogeometry.com

20211119///

Eu não sou dono da verdade,
mas eu sou amigo dela.

Texto sagrado hindu
Bhagavad Gita:
"Agora eu me tornei a morte,
a destruidora de mundos!"

→ Oppenheimer declarou citando esta parte do texto.

20211119/Geometry-Illuminated/
Geometry_Illuminated...OCR.pdf.
Pág. 115 (132).
Chapter 12
Parallel Projection.

20211120///
Data-Novo-Vivo-021-2021/ARQ/
Analytical Geometry 2D and 3D.pdf
Pág. 1 Chapter 13
Straight Line —

Jandaia

data
S T Q Q S S D

Other ref:

Books-022-2021/20211119/
Analytical Geometry 2D and 3D/

1) Analytical Geometry 2D and 3D.pdf

2) euclidean and non-euclidean geometries..pdf
Pag. 19 (37) -
The Euclidean Parallel Postulate →

⇒

www.archive.org/details/@mirtitles

3) problems-in-higher-mathematics-minorsky.pdf
Pag. 216 (221) Chapter 10
Curvature of Plane and Space Curves.

4) Topics in theory of Surfaces in Elliptic Space.pdf
Obs: Muito bom!!!

5) Shreider - What is Distance?..pdf
Obs: Legal!!!

2021 1119 | Geometry-Important |

1) Complex-Numbers-in-Geometry.pdf
2) Pogorelov-Analytical Geometry.pdf
Pag. 119 Chapter 6
A plane and a straight line
3) Uspensky-Pascal's Triangle.pdf
4) Pogorelov-Geometry-Mir-1987.pdf
5) Modenov-Problems-in-Geometry-Mir-1981.pdf
Pag. 82 (86) - Chapter III
The use of complex numbers in plane geometry.

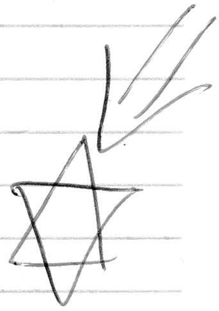

Important:

Estudar Chapter 4 -
Inversion (≅ Mirror)
⟹ analisar
⇓
relações com Projective Geometry.

20211120/Géométrie-M-Chasles/

Questions_de_géométrie...pdf

Pág. 90(110) -

6. Relations métriques dans le cercle.

Pág. 92(112) - 190.

Théorème de Maclaurin

20211121/

Data-NovoVivo-022-2021/

Books-022-2021/

20211121/

- Practical Linear Algebra A Geometry Toolbox/

Practical Linear Algebra - A Geometry Toolbox.pdf

Pág. 177(196)

Chapter 8 - 3D Geometry

20211121/

- 2nd IMA Conference on Mathematics of Robotics/

2nd IMA Conference on Mathematics of Robotics...pdf

Pág. 25(43) - Synthesis of Planar Stiffness.

data . .

S T Q Q S S D

20211122 III

H_2O 104,5° ⇒ 104° 30'

$$\begin{array}{r} 360° \times \\ 60' \\ \hline 21600' \end{array} \qquad \begin{array}{r} 104 \times \\ 60' \\ \hline 6240' + \\ 30' \\ \hline 6270 \end{array} \qquad \begin{array}{r} 21600 \div \\ 6270 = \\ 3,44\ldots \end{array}$$

$3,\overline{444976076555023923}$

20211123 II

1 Rad = 360°/2π =

57°29578 =

206265 arcseconds

Pág. 04 (19)

Source:

045--- Gerald North-
Advanced amateur
astronomy
(1997) f pdf

Source: Books-022-2021 / 20211123 /
045---Gerald North-
Advanced amateur astronomy /

H_2O → 104,45°   104°27'

ângulos entre os átomos.

21600

104
60 ×
6240 +
27
6267

21600 ÷ 6267 =
3,4466...

—

20211124//

Pág. 124 do Caderno-001--20211112.pdf

Veja Pág 40 → 090--Hadamard...pdf

20211125//

Soldner. ⇒ Pág. 161 - 172
(170) (181)

00006.djvu
00009
00170
↓
00181

} converter.

—

Einstein copiou Soldner (1801)
publicado 1804.

Pág. 161 (Item 104)

20211126///

terminei o item 106 - Caderno-001.

Próx item 107 - Pág. 176

—

Incluir imagens

Livros - Novos

—

20211127///

Fazer Down ARQ_IMPORTANT_20200207.7z

—

Caderno-001.odt → Pág. 203

—

20211128// Aniversário da minha querida Mãe Leda

multcloud.com/help/remote-upload.html

pcloud.com

rclone.~~com~~ - rclone.org

netdrive.net

Hoje é um dia importante

Happy Birthday.

data

S T Q Q S S D

Hoje é aniversário de minha
querida Mãe Leda
90 anos

Que tristeza que ela não esteja
mais aqui entre nós, mas ela
está no céu!!!

Ela faleceu:
17 janeiro 2013

Ela está presente dentro de mim,
metade do meu DNA é dela.

Portanto, ela sempre estará
presente em minha vida.

Deus ama a todos!!!
Piedoso é
Misericordioso para com
todos!!!

2021 11 29M

Estudar:

Zip-02/ others-Books/
Importantíssimo-09/
Experiencing Geometry (2004).pdf
Pág. 125 (138)
Parallel Postulates

Letter:

Einstein to Hale Oct 14, 1913

→ Lichtstrahl

midia)

(Am Sonnenrande, müsste
diese ablenkung 0,84''
→ distorção
betragen und wie $\frac{1}{R}$
abnehmen

2021 11 30M

Borromean nucleus.

data

Ⓢ Ⓣ Ⓠ Ⓠ Ⓢ Ⓢ Ⓓ

sciencenews.org

www.ibiblio.org

Symmetry - invertible
amphicheiral or
interchangeable.

20211201m

cissoid of a circle.

The Trisector by Underwood Dudley.

20211202m

forumgeom.fau.edu

20211203m

A treatise on Solid Geometry.
Frost

(HD Preto) 500 GBm

Books-021-2021 (completo) → zipar

Volume-005-20211111/Books/Gaussian-
Quadrature

data
S T Q Q S S D

radzevich@usa.com
eaton.corp.
gearsolutions.com

Livros R$ 1915 170
98 145
Sebo → 2013 1600
48 1915
2061

20211204

www.gazeta.spm.pt

20211205

Download PC-0.

Fragment of energy
☆ Larry M. Silverberg
precisa de meus ajustes !!!
North Carolina State Univ.

20211206

digitale.bibliothek@bsb-munchen.de

— Cours de géométrie infinitésimale
G. Demartres

— Leçons sur la géométrie des espaces de Riemann
Élie Cartan

— Leçons sur le calcul des variations
Jacques Hadamard

$a^{\frac{m}{n}} = \sqrt[n]{a^m}$

20211207$_{111}$

e-rara.ch

—

e-periodica.ch

—

numdam.org

—

eudml.org

—

numdam.org/item/CM/
sites.mathdoc.fr/OEUVRES/

—

sites.mathdoc.fr/JMPA/

$$\frac{a^m}{a^n} = \frac{1}{a^{n-m}}$$

$$\left(a^{\frac{1}{n}}\right)^n = \left(a^n\right)^{\frac{1}{n}} = a$$

20211208$_{III}$

Projective Geometry Coxeter

The Real Projective Plane Coxeter

Curve Schemes Speciali-
Important

→ diretório

Obs: Bianchi

20211209$_{III}$

até (parte final)

20211218$_{III}$

Obs: Digitalização do meu material de pesquisa.

20211219.

Trabalho de Digitalização de Todo o meu material de pesquisa:

início: 20211108

término: 20211218 1600 h

obs: Início dos arquivos .odt →

20211021

Brasil na 2ª GM →

BRASIL ~~USA~~ → 12ª Brigada de Infantaria Leve Aeromóvel

+

USA ~~Brasil~~ → 101ª Divisão Aerotransportada.

Em 2016, a divisão brasileira foi denominada como

"Brigada Fornovo de Taro" em decorrência da batalha com o mesmo nome ocorrida na Itália entre os dias

28 Abril à 01 Maio de 1945, onde ocorreu a captura da 148ª Divisão alemã inteira com 15.000 soldados alemães.

20211220

Freddie Mercury - Living on My Own

Roxette - Dressed for Success.

Copy files:

ARQ-R-20211220.7z

others:

Others-Books e

Others-Books-02---20211120.7z

Men Withdout Hats - Safety
without Dance

20211021

IPTU (2021) 754,80
732,16

Vila Guilherme - 75,48.

Santana - 314,31.

Atilio
Martins - 65,63 (1 parcela)
CEP 02073-080

20211023

backlash - reação

patronímico

The "open universe" has a saddle-~~shaped~~ shaped negatively curved spacetime.

saddle-shaped → em forma de sela.

Euclid ——— Gauss

(?)

à

↓

(Helmoltz (questiona?)

extension of Euclidian geometry.

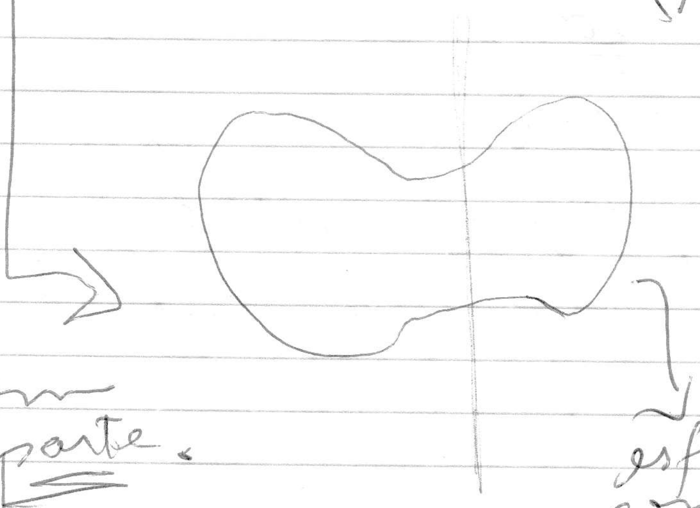

em parte.

esfera comprimida

Veja:

20211222/Important/
the Road to Reality a Complete guide...pdf
Pág. 250 (279) - Fig. 13.2

Rotational symmetry of a sphere

Obs: Estudar com
20211223/Origins the Scientific Story of Creation/
Jim Baggott-Origins...pdf
Pág. 33 (52)
Figure 5

(b) the "open" universe had a saddle-shaped negatively curved spacetime!

20211224m

musich.info. 50,00

Information Society – Lay All Your Love on Me

—

Formulas des Fresnel

Pág. 235 óptique Bruhat

=

cut-the-knot.org

—

20211225m

Gauss April 30, 1777.
February 23, 1855.

Helmholtz August 31, 1821.
September, 8, 1894.

=

"A matemática é a rainha das ciências; a aritmética é a rainha da matemática".

K. F. Gauss – in: Disquisitiones Arithmeticae

—

stampertech.com.

archmission.org/nanofiche

↳ 20x20 mm nickel can hold up to 8000 pages of text rendered at 150 dpi.

—

licence NanoIP → Nanofiche

—

lithographically printer.

—

US Patent 8717650
8264757
7961367
7830573

digitalpreservation.gov

—

nanoarchival.com

—

25 Dez.

Roma → Ⓓ deus Mitra, o deus-sol.
↳ bispos.

oração do sol invicto.

—

Jandaia

—

Ni Paix Ni Guerre
Hélène Carrière 1986
Flammarion

20211226 ///

- 13 December 2021 - Luz Santana
  Obs: Advogada vizinha da gente veio falar conosco.
- 15 December 2021 - Transferência do Campus Federal para a Administração Federal.

  Neste mesmo dia (20) encontramos o dono da imobiliária que quis comprar de graça a nossa casa.
- 20 December 2021 - Agua Santana.
- 24 December 2021 - A nossa vizinha veio trazer presente de Natal.
  Achamos que ela foi mandada por este dono da imobiliária.

—

Topologie
Johann Benedict Listing

—

lindenusd.com (California)
↳ cenic.org
Edlio, Inc

—

20211229M

Santo Agostinho – sementes
seminais
(matéria + evolucionismo)

Evolucionismo theológico
↳ teleológico

Big Bang ↖ Deus criou a matéria de uma vez e depois infundiu a evolução (mudança na forma da matéria).

→ Santo Agostinho disse que Deus criou a matéria e após isso colocou energia na matéria para que ela evolua.

Neste caso, não há incompatibilidade entre evolucionismo e criacionismo.

Darwinismo e Neo-Darwinismo.

↓ seleção natural

↓ alterações cromossômicas criam a evolução

Drive deletado da conta:
mathe1matis2@gmail.com

→ Drive é: (100Tb).
0ACQD97Je9R76Uk9PVA

20211231 1630h
Tudo foi gravado à nuvem.

✡ ARQ-R-B---20211231.7z
→ contém todo o meu material de pesquisa e livros-fontes (102Gb)

FIM DE
2021

2022: Feliz Ano Novo!!!

There are four fundamental forces at work in the Universe:

(20220101)

1) The Strong force
2) The Weak force
3) The Electromagnetic Force
4) The Gravitational Force.

reflexo
(efeito)

The Gravitational Force é decorrência da "The Strong Force".

Se a gravidade existisse independentemente das demais forças, o Universo jamais terá sido criado pois antes do Big-Bang tudo estava reunido em um único ponto e portanto, a própria gravidade teria impedido a própria explosão do Big-Bang.

A gravidade deve ser analisada localmente e quando estendida, você terá outro valor.

A extensão da análise do fator gravitacional em relação a todo Universo gera uma curvatura ~~do~~ espaço-tempo.

O Universo se expande pois a "gravidade" age localmente em relação à cada galáxia e não em relação ao todo.

Albert Einstein and the Cosmic World Order
Cornelius Lanczos

The Intrinsic Nature of things. Barbara Gellai

Cosmology's Century
P. J. E. Peebles

609-258-4386

pjep@Princeton.edu

20220102M

New Shared Drives

a) Z-Math-Geomath → SAIS Dubai.
(saisdubai.com).

b) Z-Math-New → George L. Egbert
Intermediate School.
is2egbert. ~~com~~ org

speedcrunch.org 50 digits

raidrive.com
(raidrive.com)
→ cria uma letra no Windows Explorer dos drive compartilhados que você tem no google.

—

~~2021~~

20220103//

Elementary Mathematics
from an Advanced Standpoint
Geometry
Felix Klein

—

Listar os arquivos no diretório:
20220102/Elementary
Mathematics from an
Advanced Standpoint.../

Obs: Todos os arquivos.
Converter .djvu para .pdf

—

20171111
20171113

Convex Optimization &
Euclidean Distance Geometry
Dattoro

—

015 - Richard G. Brown -
Advanced Mathematics... pdf
Pág. 258 (300) -

1 radian = $\frac{180}{\pi}$ degrees ≈ 57.2958 degrees

1 degree = $\frac{\pi}{180}$ radians ≈ 0.0174533 radians

=

π radians = 180 degrees

=

Escanear este caderno até esta parte!!!

Obs.

Faz parte do Volume - 011 --- 20220104

=

20220107///

archive.org/details/

- questioniriguard00enri
- lezdigeopro00enririch
- lezionidigeometr00bartuoft
- lezionidegeoanal00castrich

---

pdfdrive.com

20220108/// Google Drive Share

Z-Math-Jagua

e.jaguariuna.sp.gov.br

20200109///

Era de Planck.

Roger Penrose ⇒ universo cíclico
denominado
"cosmologia cíclica conformada"

Procedimento matemático complexo
denominado "reescalonamento conformado"

→

"Reescalonamento conformado"

uma transformação geométrica que na verdade altera o tamanho de um objeto, mas deixa sua forma inalterada.

20220110m Google Share Drive

Z-Math-Cisnes →
educacioncisnes.cl

Smiljan (cidade do Tesla)

---

tripofobia.

20220110/ Graham-Farmelo/

Graham Farmelo - The Universe Speaks in Numbers... (2019).epub

Pág. 82

Carl Friedrich Gauss are examples of post-Newtonian thinkers who were at the front lines of both mathematics and physics...

20220111

Geometry & Symmetry
L. Christine Kinsey
Teresa E. Moore
978-0-470-49949-8

—

Tucídedes — Guerra do Peloponeso.
↓

Armadilha de Tucídedes.

—

Potência hegemônica x Potência Emergente

—

O caminho da Guerra
Graham Allison
→ "Os Estados Unidos e a China conseguirão escapar da armadilha de Tucídedes".

História da Guerra do Peloponeso
Tucídedes

"Vitória Pírrica" — vitória na Batalha de Ásculo

—

Plato's Stepchildren - Star Trek.
(film)

—

softarchive.is/books/

—

ID 922 9587 8661. (Zoom)

20220112

—

Collegii Urbis Nov Eborac
Sigillum

Antigo-001	1447	1564
Antigo-002	117	
Caderno-001	400	1284
002	179	
003	309	
004	396	
Volume-001	539	
002	410	
003	664	
004	756	(756)
005	3654	
006	298	
007	643	
008	3304	
009	561	
010	121	
011	103	

Volume 012 - 155
013 - 43
(11251)

Total:
1564
1284
11251

14099

20220113//

20220115/// 119 Gb.

ARQ-R-B-F---20220115.77

=

20220116//

Acknowledgements = reconhecimentos

—

Helicóptero - corda de guinada (yaw string) → sentido do vento em relação à aeronave.

—

20220117/// Adamski - Killer (music)

—

Dec 21, 2021 Nytimes.com Page A21//

A Box of Cash, a Secret Donor and a Big Gift for Some N.Y.C. Studentes

by Corey Kilgannon

—

data

Making Geometry - Jon Allen
Pág. 8(7) -
God is a circle centre is
everywhere and whose
circumference is nowhere.
Variously attributed.

20220127
Non-Euclidean Geometry
by Roberto Bonola.
1955
https://download.tuxfamily.org/
openmathdep/geometry_advanced/
Non-Euclidean_Geometry-Bonola.
pdf

---

obs: https://download.tuxfamily.org/
openmathdep/

Pág. 236 (252)

§ 69 Klein's Direct Demonstration of the Independence of Projective Geometry from the Fifth Postulate.

other book:

Straight_Lines_and_Curves Vasilyev.pdf

Source:

20220127 / Recherches géométriques sur la théorie des parallèles - Gauss / Geometry-01 /

20220128 M

topfreebooks.org

20220129 M

https:// gdz.sub.uni-goettingen.de/ id/ PPN235957348

20220131

Geologists - 4200 years the
"Meghalayan Age".

Era - Cenozoic ← Quaternary

Quaternary ← Pleistocene
Holocene
epochs
↓
11.700 years ago.

—

128 Gb - 44.822 arquivos

—

ARQ-R-B-F---20220131.7z

—

20220201
Acólito -> auxiliar

—

20220201.
Criação do diretório:
DATA-NovoVivo-003-2022

—

Everything for Everyone in a Fractal Model
M.S. Carley. ~~ISSN~~ 1549890190

⇒ ISBN 9781549890192

—

20220202 ☆

⇒ 20200526 – Southwestern College
(creation)
→ google Share drive

—

O remorso é o passado se destruindo no presente.

—

20220203

vdoc.pub

—

Pappus – CXIV (Pág 111)

Pappus D'Alexandrie
La Collection Mathématique
Paul Ver Eecke 1933

=

20220207

Des de le Punto a la Cuarta
Dimensión.
Egmont Colerus

—

sm001-internet@yahoo.com

sm001-pc@yahoo.com

—

Trusted Traveler USA
https://ttp.dhs.gov

—

Summa Musichorum.

—

antípoda.

História da Igreja
Daniel ~~Hobbes~~ Rops

—

2022-0208M
Egmont Colerus von Geldern
→ De Pitágoras a Hilbert.

Del punto a la cuarta dimensión.

'Von Pythagoras bis Hilbert" 359 pág. Egmont Colerus.

ISBN 978-3893501526

Vom Punkt zur Vierten Dimension: Geometrie für Jedermann

Geometry and Discrete Mathematics Waterloo.

Abebooks.com

20220209 (Dikran Tahta) → Armênia

Dick Tahta.

Professor na escola St Albans (sul de Inglaterra). (1955) → Stephen Hawking.

→ influenciou Dent White que [illegible] em [illegible] descoberto o Teorema de Pick, no que se referia a mim, esse era o meu teorema".

Hermenêutica → técnica de interpretar e explicar um texto ou discurso
↳ termo matemático (defini-lo precisamente)

Tahta

→ "a graça da sua atenção"

Tahta inspirou a John Hibbs:
"Deixe a matemática falar".

freeditorial.com

---

https://phaidra.univie.ac.at/detail/o:1229964?mode=full

Livro Sebo do Messias: Correios.
OS652234890BR

www.math.rwth-aachen.de/homes/Ulrich.Schoenwaelder/

Pappus Eecke
Collection Mathématique

https://schule.bayernport.com

Pappus
https://archive.org/details/
PappusDAlexandrieEd.Paul
VerEeckeLaCollectionMathmatique

202202100150

Deus é uma realidade, Ele existe por si mesmo e não depende da análise objetiva ou o entendimento das pessoas.

Deus pode ser entendido de várias formas de acordo com as convicções do homem que tenta entender Deus, mas não devemos fazê-lo de forma a conceituá-lo simplisticamente hermenêutico.

O nada existe.

O conceito do nada se reflete na criação da noção do número zero, criado na Índia, ou melhor, descoberto e esboçado na forma como conhecemos, primeiramente descoberto pois está realidade já existia, faltando somente a sua materialização no mundo fático ou "real", quando houve a retirada de uma pedra da sua posição matemática quando da resolução de uma operação matemática básica em que o espaço em que faltava a pedra, ou seja, o nada ou o zero existe independentemente da existência de uma pedra posicional para demarcar a existência daquele referido número.

O nada existe porque existe a Matemática

A Matemática existe por si mesma.

A Matemática é descoberta, o homem inventa a forma como nós descrevemos os conceitos matemáticos.

A Matemática tem várias facetas e nós podemos adequá-la a cada realidade existente e que por sinal é infinita, para poder entendê-la, mas a Matemática não pode ser distorcida ou corrompida pelo homem.

Na realidade, nós devemos tentar adequar a realidade à Matemática, pois para cada realidade nós teremos que conceituá-la de acordo com a Matemática imutável, cabendo a nós, meros mortais, inventar meios, métodos, formas para "tentar" descrever a nossa realidade dentro da Matemática.

Três premissas nós devemos ter em mente:

1) O nada sempre existiu, existe e existirá.
O nada é o reflexo oposto de Deus, uma face da mesma essência de Deus.
O Universo somente pode existir pois havia o nada.
A expansão do Universo também depende do nada.
O nada permitiu o Universo

se expandir, o nada implica que o Universo o preencha. O espaço decorre do nada, o espaço não é criado pelo Universo, mas alterado por este.

→ O Tempo sempre existiu, existe e existirá.
O tempo não é criado pelo Universo, mas alterado por este.

O nada existe e portanto, ele tem o seu próprio tempo e espaço.

2) A Matemática sempre existiu, existe e existirá.
A Matemática está em tudo para ser descoberta.
A realidade depende da Matemática, inclusive o seu próprio entendimento depende dela. Cabe a nós, descrever a realidade utilizando a Matemática.
A Matemática tem infinitas facetas e formas para descrever as infinitas realidade que ainda serão descobertas.

A Matemática e o nada decorrem de Deus.

O nada obriga, ou seja a criação de alguma coisa. /

O nada foi dado por Deus como a

chave que abre a porta para infinitas possibilidades.

O nada é como esta folha em branco na qual eu estou escrevendo. Esta página deseja ser preenchida pois se ela estivesse completa, o que eu poderia escrever para tentar inventar, descobrir, descrever, narrar, contar etc.

O tudo não permite a criação pois ele já ocupou tudo e não permite a inclusão de mais nada.

Deus é tudo. Ele não precisa dos homens e por isso, Ele criou o nada como uma forma de criar espaço para que o homem possa viver e sonhar.

Deus se "ausentou" e nos deu o "livre-arbítrio" para que nós, os homens possamos preencher o nada.

São Paulo 10 de fevereiro de 2022.

202202100240

20220210m

O nada existe.
O nada permite que o Universo exista.
Literalmente:

O Universo foi criado do nada.

O nada que existe, criou o Universo.

O nada criou condições para que o Universo fosse criado.

O nada criou o Universo, além dos confins do Universo, nós mesmos encontrar o nada.

O nada impulsiona o Universo a preenchê-lo de forma acelerada, ou seja, quanto mais nada temos, mais rápido o Universo tentará preenchê-lo.

O nada decorre de Deus.

O nada é a tábua rasa na qual Deus, permite ao homem interagir e existir.

O nada sempre existiu, mas pela sua essência, ele sempre se manifesta de infinitas formas, uma delas é o Universo que conhecemos.

O nada por si mesmo cria
condições para que algo decorra
dele.

O Universo não foi criado,
ele é a manifestação do
nada.
(uma das manifestações).

O nada existe,
somente o que existe,
poderá criar algo,
neste caso,
o Universo.

O nada também é uma forma
de energia.

Tudo é energia,
a falta de energia também
é uma forma de energia.
(estado)

20220210_0507

ALL (Tudo) ——— NOTHING (NADA)

↓

SYMMETRY

↓

COLAPSO

↓

UNIVERSO

↓

CURVO (forças fundamentais Física)

↓ ↓

MIRROR (sobre um plano) gravity (decorrência dessas forças)

↓

PROJECTIVE
ORTHOGONAL. (geometry)

Obs: EUCLID (PLANO)

RIEMANN
HILBERT
GAUSS

NON-EUCLIDEAN GEOMETRY

↑

HELMHOLTZ (extensão de Euclides e não uma nova geometria)

EGMONT COLERUS
→ PUNKT

20220210

20220210\The Art of Problem Solving\

Paul Zeitz - The art and craft of problem solving - John Wiley & Sons (2017).pdf

Pág. 62(78).

—

ALL — NOTHING

Positivo — Negativo

In — Yang

obs: Ver livro Tao da Física.

=

Nome do livro:

NOTHING IS ALL

=

20220212 III 0003

A gravidade como força independente não existe.

A gravidade decorre da força forte (lei fundamental da física)

A curvatura do tecido espaço-tempo decorre da Força Forte.

A soma de cada Força Forte de cada átomo reunido em um determinado espaço gera por conseqüência a ação gravitacional. Quanto mais átomos, mais forte será a soma das Forças Fortes de todos os átomos reunidos em determinado espaço, que por sua vez, irá distorcê-lo de acordo com a reunião de todas as Forças Fortes de todos os átomos reunidos em um determinado espaço.

A curvatura do Universo existe e decorre de toda a massa deste Universo.

Onde não há massa no Universo, existe uma mínima curvatura que é decorrente da massa do Universo como um todo.

O nada existe, como o vazio do Universo também existe.

A existência destes dois gera também uma força!!!

A força gerada pelo nada e pelo vazio decorre de simetria intrínseca existente o Tudo (Deus) e o nada.

A Entropia decorre desta interação simétrica destes dois elementos.

A falta de paridade de aniquilação entre matéria e anti-matéria decorre da interação intrínseca da ação simétrica entre estas duas forças universais atuantes no Universo analisado. (em carta)

A falta desta paridade ou simetria decorre da interação entre o tudo e o nada, ou seja, não se trata de falta de paridade mas a ação permanente do equilíbrio entre estas duas forças que interagem na criação e destruição do Universo.

entraremos no campo da filosofia e da religião. Deus Shiva na cultura indiana. (estudar!!!) 20200212 0018

20220212m

20220213u

Escaneei o livro:

Desde el punto a la Cuarta
Dimensión
Egmont Colerus

→ está inserido no

Volume-014-20220213

Source:

ARQ/1-Source/Volume-014-20220213/

---

obs: Este caderno foi escaneado
até aqui

202202131050

20220214

Tarski's circle squaring problem.
https://annals.math.princeton.edu/
2017/186-2/p04

→ Andrew Marks
marks@math.ucla.edu

→ Spencer Unger
sunger@math.toronto.edu

spencer.unger@utoronto.ca

www.math.ucla.edu/~marks/
www.math.toronto.edu/sunger/

https://en.wikipedia.org/wiki/
Tarski's_circle-squaring_problem

https://www.maa.org/sites/default/
files/pdf/upload_library/22/Evans/
Pierce.pdf

https://arxiv.org/abs/2202.01412

↳ Circle Squaring with Pieces of Small Boundary and Low Borel Complexity.

═

Principal:

https://www.quantamagazine.org/an-ancient-geometry-problem-falls-to-new-mathematical-techniques-20220208/

═

↳ An Ancient Geometry Problem Fall to New Mathematical Techniques

═

https://www.degruyter.com/document/doi/10.1515/crll.1990.404.77/html

═

data

Equidecomposability and discrepancy: a solution of Tarski's circle-squaring problem (M. Laczkovich)

=

Source:
https://doi.org/10.1515/crll.1990.404.77

=

other source:
https://eudml.org/doc/153197
↳
https://gdz.sub.uni-goettingen.de/id/PPN243919689_0404?tify={"view":"info","pages":[81]}

Source: PURL
https://resolver.sub.uni-goettingen.de/purl?PPN243919689_0404

=

Jandaia

- https://www.quantamagazine.org/after-centuries-a-seemingly-simple-math-problem-gets-an-exact-solution-20201209/
- https://www.quantamagazine.org/mathematicians-prove-30-years-old-andre-oort-conjecture-20220203/
- https://www.quantamagazine.org/with-a-new-shape-mathematicians-link-geometry-and-numbers-20210719/

---

https://arxiv.org/abs/1601.02958

Arquivos armazenados na pasta:
DATA-NovoVivo-003-2022/
Important=Squaring-Circle.
Copy: Books-003-2022/20220214/
Important=Squaring-Circle.

20220215 Fusion

deuterium-tritium fusion
↓
hydrogen-boron
→ aneutronic fusion (< temperature)

20220216

C.F. Gauss Werke: Briefwechsel
Mit Schumacher, 3 Volumes
US$ 225,00 Publisher: Georg Olms
(January 1, 1975)
ISBN 9783487056449

https://gdz.sub.uni-goettingen.de/
id/PPN635102439

20220217

Simply Math DK.

20220219

Pi in the Sky - Barrow.

Mécanique Quantique T1
Claude Aslangul

—

Kartheuser

—

→ Converter 20220225/Important/
Mathematical analysis.../

20220226

38,878.29 GB

—

https://gallica.bnf.fr/ark:/
12148/bpt6k64545935
→ Exercices de géométrie.

—

serge.mehl.free.fr/anx/
pb23_hilbert.html

—

https://td.mysuite.worker.dev

—

Geometry of Four Dimensions
Forsyth.

Eurípedes - a vida não deve ser levada muito à sério.

20220227

pdf-magazines.me

novafile.com

while - enquanto

deterrent → dissuasor.

Mathematical Techniques for Engineers and Scientists.

Pág 275(292)

6.2.1 Geometric Interpretation: the complex plane

Figure 6.1

(adaptar para prefácio)

20220228

Quem está certo sempre vence.

20220228m

https://pt.scribd.com/document/
117480557/10-Piscoran-Final

Source:
www.cs.ubbcluj.ro/~studia-m/
2011-1/
10-piscoran-final.pdf

Obs: ISBN- 2065-961x

https://www.mpim-bonn.
mpg.de/node/144

↳ Algebra, Geometry and
Physics: a mathematical
mosaic.

fdown.net

Jandaia

20220301 M

meet.google.com/igi-ujrd-pef

→ Ungar

—

Sumo Sacerdote →

Cohen Hamashiach.

—

Melech Hamashiach

(Rei Messias)

—

20220302

I Renty

—

20220303

Brane → an extended object with any given number of dimensions, of which strings in string theory are examples with one dimension. Our Universe is a 3-brane

—

"Você não terá nada e será feliz" → comunismo.

→ Forum Mundial.

data S T Q Q S S D

20220304

$1^3 = 4$ $\qquad 2^3 = 8$

✓

lógico!!!

$1 m^3 = \underline{1000}$

$\frac{1 m^3}{3} = 333,\overline{3}$ $\qquad \frac{1 m^3}{1^3} = \frac{1000}{4}$

3

$= \underline{\underline{250}}$

$$\begin{array}{r} 333,3 \\ -\ 250 \\ \hline \underline{\underline{83,3}} \end{array}$$

333,3 — 100

83,3 — x

$= \overline{25,0015} \; (\approx \tfrac{1}{4})$

$\underline{\text{off}}$: $\frac{83,3}{333,3} = 0,2499$

Jandaia

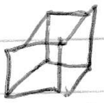

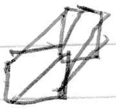

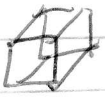

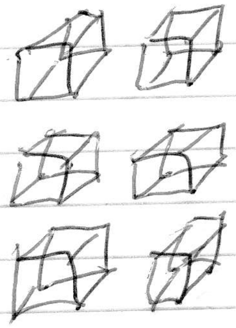

$2^1 = 2$

×2

$2^2 = 4$

×2

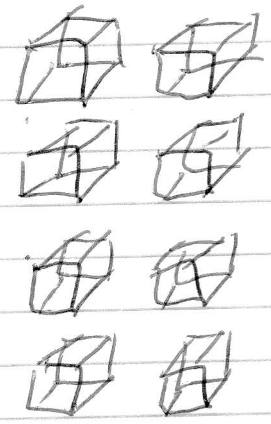

$2^3 = 8$

Se $2^3 = 8$

∴

$1^3 = 4$

$1^1 = 1$ $1^2 = 1$ $1^3 = 1$ (wrong!!!)

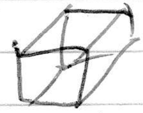
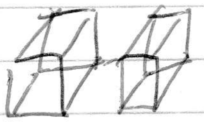
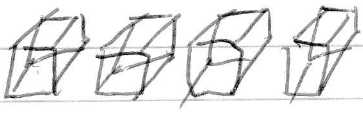

$1^1 = 1$ $1^2 = 2$ $1^3 = 4$

$2^1 = 2$ $2^2 = 4$ $2^3 = 8$

=

→ Wrong!!! do ponto de vista numérico, mas volumetricamente funciona

$1 m^3 = 1000$ litros (3d)
↓
space

Adequação entre espaço e matemática.

Latex – Book
Article – Report.

20220204

```
\begin{text}
\HUGE
\end{text}
```

A ferramenta matemática tem por objetivo conhecer o real sentido e valor das ~~cada~~ palavras, symbols and numbers needed to express mathematical ideas.

⇒ A ferramenta matemática ~~tem~~ é uma ciência ~~que tem~~ da interpretação das palavras, símbolos e números necessários para expressar ideias matemáticas.

O conhecimento do real ~~do~~ sentido e valor da linguagem matemática tem por objetivo exprimir de forma precisa os injunções conceitos inerentes a cada palavra empregada no texto matemático.

Nenhuma palavra, símbolo ou número está no texto sem significado ou função, pois cada uma delas contribui com a ideia que se quer

construir, demonstrar, defender, refutar ou explicar.

parágrafo LATEX

`\usepackage{indentfirst}`

Para não fazer isto em um determinado parágrafo

`\noindent`

`\emph{texto}` → enfatizar

`\emptyset`

Book:

latex-via-exemplos.pdf

Pág. 10(26) - 3.3

Equação enumerada e referência cruzada

https://www.nasa.gov/
send-your-name-with-artemis/
Prof. Stenio
Musich
→ 9057331 (Password)

ARQ → 160 GB
50.180 files
1.154 folders.

20220305 ///
Tears for fears I laid no law.

3,25 x 25 = 81,25
3,50 x 25 = 87,50
3,75 x 25 = 93,75

3,85 x 25 = 96.25
3,95 x 25 = 98,75

3,15 x 25 = 78,75
3,25 x ~~25~~ = 81,25
3,35 x 25 = 83,75
3,45 x 25 = 86,25
3,55 x 25 = 88,75
3,65 x 25 = 91,25

3,10 x 25 = 77,50
3,20 x 25 = 80
3,30 x 25 = 82,50
3,40 x 25 = 85
3,50 x 25 = 87,50

3,60 x 25 = 90
3,70 x 25 = 92,50
3,80 x 25 = 95
3,90 x 25 = 97,50

www.loc.gov. (2 volumes)
2017 gen 30462 v1 . Galileo Galilei
v2

DownThemAll (~~0~~ Firefox App)

titere (usado por outro)

www.loc.gov/item/29030462/
↳ Opere di Galileo Galilei

https://www.loc.gov/resource/
rbctos.2017gen30462v1/
30462v2/

www.loc.gov/item/12018406/

→ Dialogo di Galileo Galilei
Linceo.

Sidereus nuncius
Galileo Galilei
www.loc.gov/item/2010667904/

Il saggiatore

Il Vero e il falso
Antonino Zichichi
→ passeggiando tra le stelle e a casa.
978-8842810780 547 pag.

La forma dello spazio profondo
Shing-Tung, Steve Nadis

ISBN 978-8842815365

https://philarchive.org/archive/

https://archive.org/details/
ilveroeilfalsopa0000zich

Elements of geometry
Ladd, Kelly

ATA NOTARIAL

↓ ↳ 23º Cartório NOTAS SP.

Perpetuar conteúdo de páginas na Internet.

20220306m

202202060005

$1^3 = 3$ ou 4

$1 \times 1 \times 1 = \boxed{3}$ (3 unidades)

$\underbrace{(01) \quad (02) \quad (03)}_{\text{3 unidades}}$

Usando da mesma lógica:

$2^3 = 8$

$\underbrace{2 \times 2 \times 2}_{\text{3 elementos}} = 8$ (valor em unidades)

O número 2 é o dobro de 1, portanto, se $2^3 = 8$ e o número 1 é a metade de 2^3, portanto, $1^3 = \boxed{4}$

Prosseguindo (com base na razão no número qual (exponenciando)

~~2~~ $3^3 = 27$

$3 \times 3 \times 3 = 27 \Rightarrow 1 = \frac{1}{3}$ de 27

$\frac{27}{3} = \boxed{9}$

$4^3 = 64 \Rightarrow 1 \Rightarrow \frac{1}{4}$ de $64 = \frac{64}{4} = \boxed{16}$

$5^3 = 75 \Rightarrow 1 = \frac{1}{5}$ de $125 = \frac{125}{5} = \boxed{25}$

$6^3 = 216 \Rightarrow 1 = \frac{1}{6}$ de $216 = \frac{216}{6} = \boxed{36}$

$7^3 = 343 \Rightarrow 1 = \frac{1}{7}$ de $343 = \frac{343}{7} = \boxed{49}$

$8^3 = 512 \Rightarrow 1 = \frac{1}{8}$ de $512 = \frac{512}{8} = \boxed{64}$

$9^3 = 729 \Rightarrow 1 = \frac{1}{9}$ de $729 = \frac{729}{8} = \boxed{81}$

$10^3 = 1000 \Rightarrow 1 = \frac{1}{10}$ de $1000 = \frac{1000}{10} = \boxed{100}$

Divisão por $\underline{\underline{4}}$:

$1 \times 1 \times 1 = 4$

$2 \times 2 \times 2 = 8 \; \rangle \; \frac{8}{4} = 2$ (dobro)

$3 \times 3 \times 3 = 27 \; \rangle \; \frac{27}{8} = 3{,}375$

$4 \times 4 \times 4 = 64 \; \rangle \; \frac{64}{27} = 2{,}\overline{370}$

$5 \times 5 \times 5 = 125 \; \rangle \; \frac{125}{64} = 1{,}953125$
(1,953125)

$6 \times 6 \times 6 = 216 \; \rangle \; \frac{216}{75} = 2{,}88$

$7 \times 7 \times 7 = 343 \; \rangle \; \frac{343}{216} = 1{,}58796286$

$8 \times 8 \times 8 = 512 \; \rangle \; \frac{512}{343} = 1{,}49271137$

$9 \times 9 \times 9 = 729 \; \rangle \; \frac{729}{512} = 1{,}42382812$

$10 \times 10 \times 10 = 1000 \; \rangle \; \frac{1000}{729} = 1{,}37174211$

entre os números reais

$\cong \sqrt{2} = 1{,}414217?$

Relação direta entre os números.

$1 \rightarrow 2 \Rightarrow 8 \div 4 = 2$

$1 \rightarrow 3 \Rightarrow 27 \div 4 = 6,75$

$1 \rightarrow 4 \Rightarrow 64 \div 4 = 16$

$1 \rightarrow 5 \Rightarrow 125 \div 4 = 31,25$

$1 \rightarrow 6 \Rightarrow 216 \div 4 = 54$

$1 \rightarrow 7 \Rightarrow 343 \div 4 = 85,75$

$1 \rightarrow 8 \Rightarrow 512 \div 4 = 128$

$1 \rightarrow 9 \Rightarrow 729 \div 4 = 182,25$

$1 \rightarrow 10 \Rightarrow 1000 \div 4 = 250$

$6,75 - 2 = 4,75$

$16 - 6,75 = 9,25$

$31,25 - 16 = 15,25$

$54 - 31,25 = 22,75$

$85,75 - 54 = 31,75$

$128 - 85,75 = 42,25$

$182,25 - 128 = 54,25$

$250 - 182,25 = 67,75$

Se em vez de 4 → 3

$1 \times 1 \times 1 = 3$

$8 \div 3 = 2,\overline{6}$

$27 \div 3 = 9$

$64 \div 3 = 21,\overline{3}$

$125 \div 3 = 41,6$ (41,6)

$216 \div 3 = 72$

$343 \div 3 = 114,\overline{3}$

$512 \div 3 = 170,6$

$729 \div 3 = 243$

$1000 \div 3 = 333,\overline{3}$

$4,75 \rangle 4,5$

$9,25 \rangle 6$

$15,25 \rangle 7,5$

$22,75 \rangle 9$

$31,75 \rangle 10,5$

$42,25 \rangle 12$

$54,25 \rangle 13,5$

$67,75$

diferenças de 1,5 em ⑴,5 N

Voltando:

$1^3 = 4 \leftarrow \left(\frac{8}{2} = 4\right)$

$2^3 = 8 \rightarrow 2$ é o dobro de ①

$1^3 = 4$

$2^3 = 8 \rightarrow$ 2 é o dobro 1 $= 8 \div 4 = 2$

$3^3 = 27 \rightarrow$ 3 é o triplo 1 $= 27 \div 4 = 6{,}75$

$4^3 = 64 \rightarrow 4 \times 1 = 64 \div 4 =$

$5^3 = 125 \rightarrow 5 \times 1 = 125 \div 4 =$

$6^3 = 216 \rightarrow 6 \times 1 = 216 \div 4 =$

$7^3 = 343 \rightarrow 7 \times 1 = 343 \div 4 =$

$8^3 = 512 \rightarrow 8 \times 1 = 512 \div 4 =$

$9^3 = 729 \rightarrow 9 \times 1 = 729 \div 4 =$

$10^3 = 1000 \rightarrow 10 \times 1 = 1000 \div 4 =$

Corrigindo

$8 \div 4$ = erro

✡

$8 \div 2 = 4 \div (1^3)$ > 5

$27 \div 3 = 9$ > 7

$64 \div 4 = 16$ > 9

$125 \div 5 = 25$ > 11

$216 \div 6 = 36$ > 13

$343 \div 7 = 49$ > 15

$512 \div 8 = 64$ > 17

$729 \div 9 = 81$ > 19

$1000 \div 10 = 100$

~~outro erro~~ NAO!!

Correto

O valor aumenta de 2 em 2

conjuto:

Usando a multiplisação partindo-se do número $1^3 = 4$, depois faremos $1^3 = 3$ //

$$1^3 = \boxed{4}$$

2x

$$2^3 = \boxed{4} \times 2 = 8$$

×3

$$3^3 = \boxed{4} \times 3 = 12$$

$$4^3 = \boxed{4} \times 4 = 16$$

$$5^3 = \boxed{4} \times 5 = 20$$

$$6^3 = \boxed{4} \times 6 = 24$$

$$7^3 = \boxed{4} \times 7 = 28$$

$$8^3 = \boxed{4} \times 8 = 32$$

$$9^3 = \boxed{4} \times 9 = 36$$

$$10^3 = \boxed{4} \times 10 = 40$$

mantendo-se
a proporção,
(o ensinamento
equivocado com
potenciação)

estudar!!!

mas numericamente
está correto.
(?)
Checar!!!

Valores reais:

4 $8 \div 4 = 2$
8
27 $27 \div 4 = 675$
64 $64 \div 4 = 16$
125 $125 \div 4 = 31,25$
216 $216 \div 4 = 54$
343 $343 \div 4 = 85,75$
512 $512 \div 4 = 128$
729 $729 \div 4 = 182,25$
1000 $1000 \div 4 = 250$

discrepância.

Simulações:

$1^3 = 1 \times 1 \times 1 = 3$

$2^3 = 2 \times 2 \times 2 = 8$ } $8 \div 3 = 2,\overline{6}$

$3^3 = 27 \rightarrow 27 \div 3 = 9$

$4^3 = 64 \rightarrow 64 \div 3 = 21,\overline{3}$

$5^3 = 125 \rightarrow 125 \div 3 = 41,\overline{6}$

$6^3 = 216 \rightarrow 216 \div 3 = 72$

$7^3 = 343 \rightarrow 343 \div 3 = 114,\overline{3}$

$8^3 = 512 \rightarrow 512 \div 3 = 170,\overline{6}$

$9^3 = 729 \rightarrow 729 \div 3 = 243$

$10^3 = 1000 \rightarrow 1000 \div 3 = 333,\overline{3}$

Aprofundando:

começando desde o número

1

$1^3 = 4$

$1^3 = 4 \Rightarrow 4 \div 1 = 4$

$4 - 4 = 0$

relação direta perfeita!!! (ou relação direta)

$2^3 = 8 \Rightarrow 8 \div 2 = 4$ > 5

$3^3 = 27 \Rightarrow 27 \div 3 = 9$ > 7

$4^3 = 64 \Rightarrow 64 \div 4 = 16$ > 9

$5^3 = 125 \Rightarrow 125 \div 5 = 25$ > 11

$6^3 = 216 \Rightarrow 216 \div 6 = 36$ > 13

$7^3 = 343 \Rightarrow 343 \div 7 = 49$ > 15

$8^3 = 512 \Rightarrow 512 \div 8 = 64$ > 17

$9^3 = 729 \Rightarrow 729 \div 9 = 81$ > 19

$10^3 = 1000 \Rightarrow 1000 \div 10 = 100$

aumento de 2 em 2

Obs:

Mantendo-se esta constante e retroagindo até o número 1,

$1^3 = 1$, pois $2^3 = 8 \div 2 = 4$

$4 - 3 = 1$

e não $1^3 = 3$ ou 4

data

S T Q Q S S D

Comparando ⇒ 3 e 4

4	8÷4 = 2	3	8÷3 = 2,6
8		8	
27	27÷4 = 6,75	27	27÷3 = 9
64	64÷4 = 16	64	64÷3 = 21,$\overline{3}$
125	125÷4 = 31,25	125	125÷3 = 41,$\overline{6}$
216	216÷4 = 54	216	216÷3 = 72
343	343÷4 = 85,75	343	343÷3 = 114,$\overline{3}$
512	512÷4 = 128	512	512÷3 = 170,$\overline{6}$
729	729÷4 = 182,25	729	729÷3 = 243
1000	1000÷4 = 250	1000	1000÷3 = 333,$\overline{3}$

estudar com mais
profundidade!!!

Prof. Stenio Musich

20220305 0200 !!!

202203061530$_{m}$

$$1^3 = 1 \times \underset{\cong}{1} \times 1 = 3$$

$$1 + 1 + 1 = 3_{u}$$

$$1^4 = 1 \times 1 \times 1 \times 1 = 4$$

$$1 + 1 + 1 + 1 = 4_{m}$$

$$1^4 = \frac{1}{2} \cdot 2^3$$

obs:

$$2 \times 1 = \boxed{2} \leftarrow$$

$$1^3 = \underset{(1+1)}{\underset{=}{1 \times 1}} = \overbrace{\underset{(2+1)}{\underset{=}{2 \times 1}}} = 3 \text{, pode ser } \underline{2} \quad (\text{testar})$$

$\Rightarrow$ vejamos

$$1^3 = 2$$

$1^4 =$ 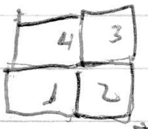 = base do cubo

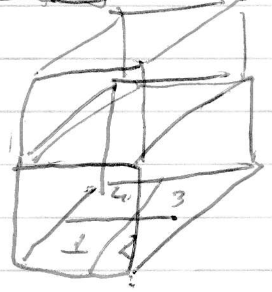

$$\left.\begin{matrix} = 4 \\ + \\ = 4 \end{matrix}\right\} 8 = 2^3$$

data

S T Q Q S S D

20220307/// 1226///

$1^3 = 1 \Rightarrow 1^3 = 1 \times 4 \Rightarrow 2^3 = 8$

—

$1^4 = 4$ (?)

$2^3 = 8$

—

$2^2 = 4$ (base do cubo) → ↓

$2^3 = 8$

—

$$\begin{cases} 1^3 = 1 \\ 2^2 = 4 \\ 2^3 = 8 \end{cases}$$

///

—

Jandaia

20220308

ARQ-BASICS → fazer down again.

—

art1lib.org

—

The Mathematical Universe
Pag. 28(47)
Pi in the sky.
(no rodapé).

—

01 06 1969 (Br)
06 01 1969 (USA)

mesmo valor (forma)

Einstein March 14, 1879

$\underbrace{03\ 14}_{Pi}$

obs:

British prefere
July 22/7 for Pi Day or
$\frac{22}{7} = 3.14$ to two decimal places.
As an aside that also happens to be the birthday of
German mathematician and physicist
Friedrich Wilhelm Bessel
(1784 - 1846),
whom we encounter in footnote 23 of Chapter 2.

The Bessel differential equation and subsequent solutions (Bessel functions) play an important role in many branches in physics.

20220309m

palíndromo
ambigrama

20220310

archive.org/details/@mirtitles

RASPUTITSA ⇒ tempo das estradas ruins. (para tanks)
⇓
o nome na Invasão Nazista da Rússia.

20220311

SMHD2022 - Partilhado (X)

Einstein
Minkowski } baseado nos estudos Lorentz and Poincaré

Pecado 3 fases { 1) sugestão
2) deleite
3) aceitação

20220315m
New Directory:
Data-NovoVivo-006-2022

—

booksc.me

—

manuelsanciens.blogspot.com

—

Data-NovoVivo-005-2022
→ cópia da obra de Galileo Galilei

20220318m ✡
ARQ-ALL--20220318.7z
→ Research material

—

20220319m
mgsuite.eu.org obs: era
— td.mgsuite.workers.dev

—

x2download.com

—

20220324 m

Einstein saw symmetry hidden in the fabric of space and time.

20220326 m

libgen.is

libgen.rs / libgen.st

20220327 m COCHISE

https://drive.google.com/drive/u/1/folders/0AO-xX4BHvgjsUk9PVA

↳ zero ↳ ono.

—

Ernest E. Shult

ernest_shult@yahoo.com

Johannes Ueberberg

johannes.ueberberg@gmx.de

Incidence Geometry

20220329III Exclui extensão.

Programa → regedit (no campo de execução do windows)

HKey_Current_User
↳ Software → Microsoft →
Windows →
CurrentVersion →
Explorer →
FileExts

HKey_Classes_Root →
↳ exclui a extensão.

—

Reinicie a máquina

Execute a extensão e depois escolha o programa que você quer usar para abri-lo.

—

20220330III

Pesach April 15

Chag Pessach Sameach

—

Legendre - Elementos de Geometria

https://bd.camara.leg.br/bd/handle/bdcamara/22882

https://bdlb.bn.gov.br/acervo/
handle/20.500.12156.3/442545

Math Books:
https://www.doraci.com.br/
downloads/matematica/
fund-mat-elem_01.pdf
11.pdf

20220331
spruce up = enfeitar.

backblaze.com

Trigonometric Functions
and Complex Numbers
Desheng Yang
ISBN 9781938134869
https://lccn.loc.gov/
2016026528

data . .
S T Q Q S S D

20220412m

Aniversário do meu pai
85 anos
Vida longa, muitas felicidades e
saúde !!!

→ mathemata

Books-07-2022/20220413/
Important/

A History of Pythagoreanism.pdf
Pág. 13 (21)

(Pythagorean analysis of the similarities between numbers and things inspired by the advancement of the mathemata, which Aristotle attributes to them;

Pág. 105 (113) -

Plato ascribes to Theodorus'
~~(discoveries were further incorporated)~~
a proof of irrationality of the magnitudes between √3 to √17 →

Cansaço! troquei até link rsrs

(Tht. 147d = DK 43A4),
which means that
(significa que)
theodorus relied on the proof
of the irrationality of $\sqrt{2}$
that was found by Hippasus.[62]

62 - For reconstructions of
theodorus's proof, see, e.g,
Heath 1921: I: 202ff;
Knorr 1975: 109ff

Obs. Heath, T. L. 1921
A History of Greek Mathematics
→ Pág. 471 (479) do livro
A History of Pythagoreanism.

Heath - Vol I Pág 202

outro assunto!

this curve will be best
described when we come
to deal with the problem of
squaring the circle (chapter VII).

here we need only remark that it
implies the proposition that the
lengths of arcs in a circle are
proportional to the angles
subtended by them at the centre
(Eucl. VI. 33)

—

trisecting any rectilineal angle
...

and these problems are not
what the Greeks called 'plane'
problems, i.e,
they cannot be solved by means
of the ruler and compasses.

—

Hippasus – Pág. 65 (85)
Chapter III
Pythagorean Arithmetic

Hippasus → ... for making known
the discovery of the
irrationals or
incommensurable.

—

data
S T Q Q S S D

Pág. 160 (180) -

Hippasus

who was a Pythagorean but owing to his being the first to publish and write down the (construction of the) sphere with (ἐκ, from) the twelve pentagons, perished by shipwreck for his (impiety), but received credit for the discovery...

Escanear amanhã!
Este caderno foi escaneado até aqui 202204122228.

20220414.

www.instagram.com/p/CClewLJGel/
↳ wolfram transformation.

— /p/BkV4950nbr/

Google Drive → 93,77 TB
Drive → Mathematics 2 --- 89.43

August Kekulé } theory of
Edward Frankland } Valency.

Graph Book — Robin Wilson and John Watkins.

Chag Pessach Sameach =
Feliz Páscoa

Planarity

We now embark upon a study of topological graph theory.

Planar graph is a graph that ~~Gan~~ can be drawn in the plane without crossings - that is, so that no two edges intersect geometrically except a vertex with which both are incident.

Source: Introduction to Graph Theory
Robin J. Wilson
Pág. 80(89)

—

20220416

TPV - thermophotovoltaic

cell that converts heat to electricity with over 40 percent efficiency and performance nearly on par traditional steam turbine power plants.

—

20220417m
Google Drive 94,49 TB

—

20220418

www.imo.universite-paris-saclay.fr/~perrin/Livre-de-geometrie-projective.html

Obs: Este professor não foi muito gentil, pois ele não respondeu aos meus emails.

Data-NovoVivo-008-2022/
Books-08-2022/
20220418/
Géométrie-Perrin/
Mathématiques d'école - nombres,
mesures et géométrie - Cassini
(2011).pdf

Pág. 209 (218) - Chapitre 7
La mesure des aires.

Obs: Pág. 210 - Newton et Leibniz
Obs: Chapitre 7 -
209 → 264.

Pág. 204 (213) - 179
La construction de Ramanujan.

20220419_M
97,23 TB
Obs: Hoje é dia de Santo Expedito!!!

20220420_M → 2021-New-20210601
drive.google.com/drive/folders/0AAV8wobLXAM8Uk
9PVA

20220421 III Research

1) ARQ-R--01--20220421.7z ⇒
Antigo-001 até Volume-008---
20211207

2) ARQ-R---02---20220421.7z ⇒
Volume-009 até Volume-015---
20220213

3) ARQ-R---03---20220421.7z ⇒
Volume-016 até Volume-019---
20220412

4) ARQ-R---04---20220421.7z ⇒
ARQ_IMPORT até OpenOffice-Latex-
Geogebra

Obs:

1) 57.96 GB ⇒ 57,96
2) 43.40 GB ⇒ 43,47
3) 43.40 GB ⇒ 43,45
4) 36.00 GB ⇒ 36,04

Total: 180 GB (194.268.759.202 bytes)
181 GB (194.376.822.784 bytes)

51.720 files 1.171 Pastas

20220423 M

Prescribing the curvature of a Riemannian Manifold

Jerry L. Kazdan n°57

Michel Serres - Geometry

---

libgen.is

20220424 M Biologia quântica.

Os processos biológicos são na verdade sistemas quânticos.

Os processos biológicos são muito complexos.

Os sistemas quânticos precisam de "estabilidade" algo que os cientistas conhecem como coerência de ondas.

A biologia tem condutores de energia. Este processo realmente existe e é chamado de tunelamento.

As partículas se movem na forma de ondas.

Pesquisador: Erwin Schrödinger
(1/2 séc. 20)

— tunelamento.
— spin.

Fotossíntese — 3 etapas:

1) Captura do fóton (a partícula portadora da radiação eletromagnética) pelas plantas.

2) elétrons absorvem a energia dos fótons e passam para um estado de maior energia navegando pelas moléculas e com base no efeito túnel.

3) O elétron é usado para uma reação química que resulta na liberação de oxigênio.

obs. dependendo do spin

Dependendo da direção do spin o elétron passará ou não pelo túnel.

QUIRALIDADE (grego kheir)

⇒ mão.

Quando um objeto é quiral, tem outro que o seu reflexo.

Isso significa que o spin anda de mãos dadas com o quiral.

Portanto:

1/ Entrelamento
2/ spin
3/ quiralidade

Presente: a/ fotossíntese
b/ síntese de proteínas
c/ a forma como os organismos respiram
d/ conexão entre os neurônios.
e/ mutações.

---

quiralidade → chirality
(Kheir greek world)

20220427

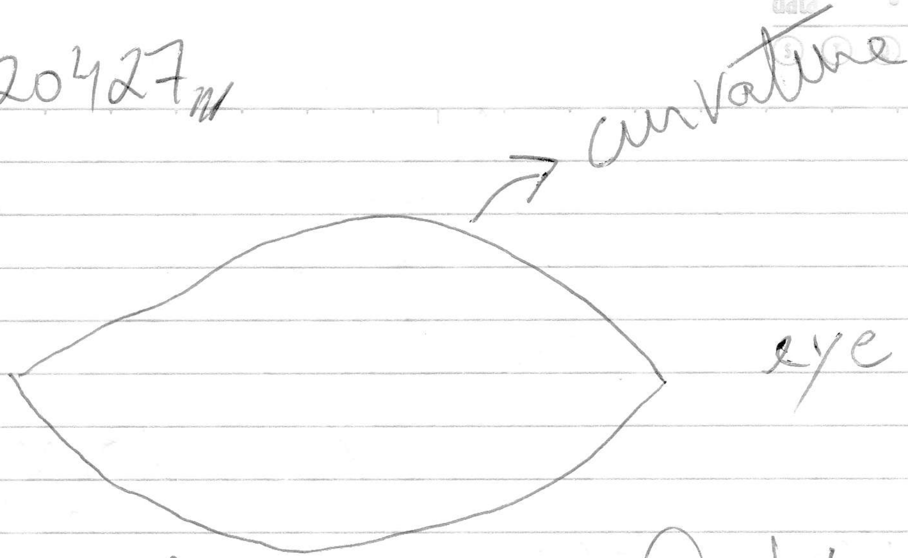

Manipular as equações até conseguir obter um círculo.

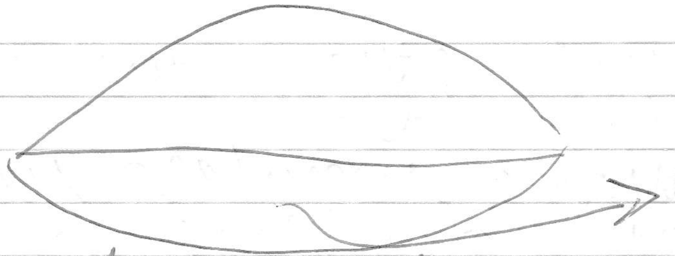

o parte de baixo tem uma curvatura, simular para obter a mesma curvatura de Gauss (Universe).

Fazer uma simulação que represente um círculo curvo desenhado sobre um plano.

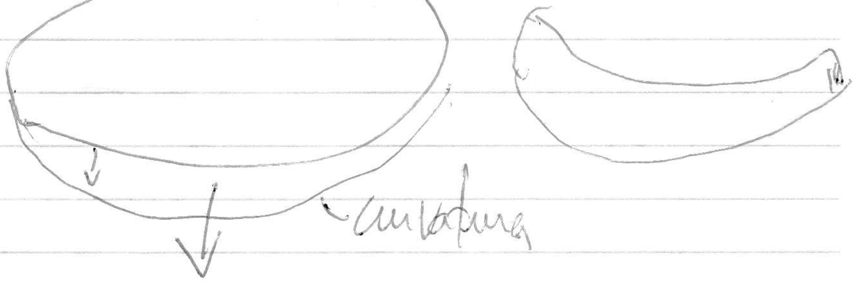

data

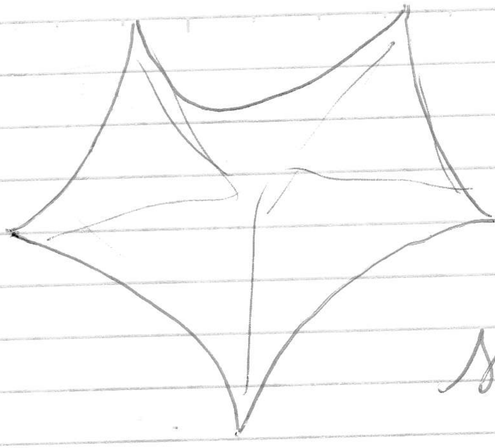

pentagrama
(curvo)

(3D)

sobre um
plano.

Fazer a mesma
simulação com um círculo.
(círculo curvo).

—

Criar uma equação que crie
um círculo perfeito.

Esta equação simulará um
círculo curvo com "projeção" (mirror)
ortogonal sobre um plano, ou seja,
esta equação servirá para a
prova do nosso valor de Pi = MVS.

Desta forma, será um círculo
curvo em (3D) com espelhamento
ortogonal sobre um plano.

—

MIM → Munich Institute of Mathematics

matemáticus.org

matematicamus (excelente nome)

IPV6 – 2804:431:c7d9:f8be:
a1f6:5d77:78e1:ae11)

—

onomasiological

—

M. Cantor – Vorlesungen

20220504 —

Grigori Perelman

Richard S. Hamilton

—

Könisberg pontes

1735 – Leonard Euler.

não é possível cruzar todas
as pontes uma única vez e voltar
ao mesmo ponto de partida

—

Séc. XIX / Henri Poincaré

Objetos homeoformos.

Algumas pessoas se referem a esse importante campo da matemática como "geometria flexível", ou seja, duas formas são a mesma se for possível transformar uma em outra sem quebrá-la!

Poincaré → desenvolveu todas as formas possíveis nas quais poderia envolver esse universo bidimensional plano.

Conjectura Poincaré → aplicar para o mundo 3D.

Grigori Perelman → 1994
havia provado a "conjectura da alma",
segundo a qual pode-se deduzir as propriedades de um objeto matemático a partir de pequenas regiões desses objetos, chamados alma.

Hamilton → 1982
"Fluxo de Ricci" com a qual se suspeitava ser possível comprovar a conjectura de Poincaré.

Tarefa era extremamente técnica
e sua execução, complicada.

Perelman ⇒ 1993
aceitou uma bolsa de pesquisa
na Universidade da Califórnia,
onde assistiu a várias conferências
de Hamilton.

The largest n-digit number
that is also an nth power
21 digits

$9^{21} = 109,418,989,131,512,359,209$

$2592 = 2^5 \times 9^2$

Also the ratio between
1 mi and 1 km is very close
to the Golden Ratio ≈ 1.61

Fibonacci numbers: 1, 1, 2, 3, 5, 8, 13, 21

3 mi ≃ 5 Km
5 mi ≃ 8 Km
8 mi ≃ 13 Km

73939133 is the largest prime number with this curious property: if you take one or more digits off the end, the resulting numbers are all prime.

73939133
7393913
739391
73939
7393
739
73
7 //

This is the only solution for $a^2+b^3=c^7$ in positive integers:

$$15312283^2+9262^3=113^7$$

Important: 1^{13}

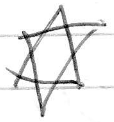

Arranging

$$1^3+2^3+3^3+4^3+5^3+6^3+\dots+n^3$$

in the plane

$$1\times1^2+2\times2^2+3\times3^2+$$
$$4\times4^2+5\times5^2+6\times6^2+$$
$$\dots+n\times n^2$$

The even square overlap ($\boxed{A}$) but also leave uncovered gaps ($\boxed{V}$)

↳ quadrado amarelo

↳ quadrado vermelho

and

$$\boxed{A}=\boxed{V}$$

Question:

$$1^3 + 2^3 + 3^3 + 4^3 + 5^3 + 6^3 + \ldots + n^3 =$$
$$(1 + 2 + 3 + 4 + 5 + 6 + \ldots + n)^2$$

1^3 = usar no meu arquivo latex.

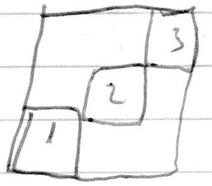

Obs: Curvature

usar para provar:

"capillary flow as the cause of ring stains from dried liquid drops"

→ fluxo capilar como causa de manchas de anel de gotas liquidas secas.

https://jfi.uchicago.edu/~tten/Coffee.drops/Nagel.7.4.pdf

20220506

Die Prinzipien der Quantenmechanik
Dirac.

---

dielectric constant of water

---

cardioids
inverted vesica pisces ♓

⇘ youtube.com/watch?v=fHUYEssv4oE
Live CEOing Ep 606:
Language Design Review for Geometric Scene

---

object-oriented programming;

20220507

Gauss Werke
antiquariat@hagena-schulte.de
veja id=34375886

Abhandlungen zur Methode
der kleinsten Quadrate
Gauss (1887)
archive.org/details/abhandlungen
met00gausrich
20220508
cinderella.de

data

S T Q Q S S D

2022,0508

File data-20220504

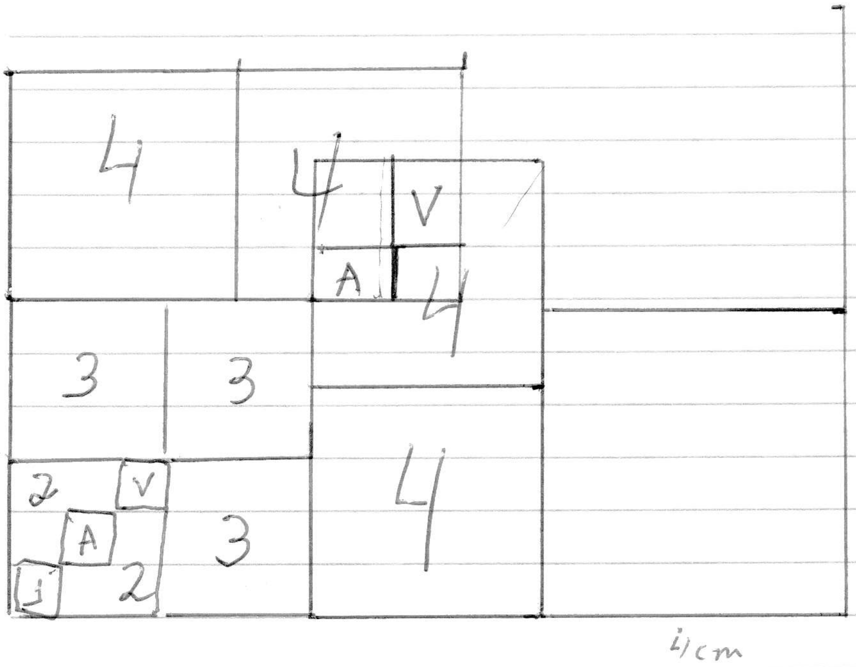
4
4
V
A
4
3
3
2
V
A
3
4
i
2
4cm

melhorando

20220508 $1^3+2^3+3^3+4^3+5^3+6^3+\dots+n^3=(1+2+3+4+5+6+\dots+n)^2$

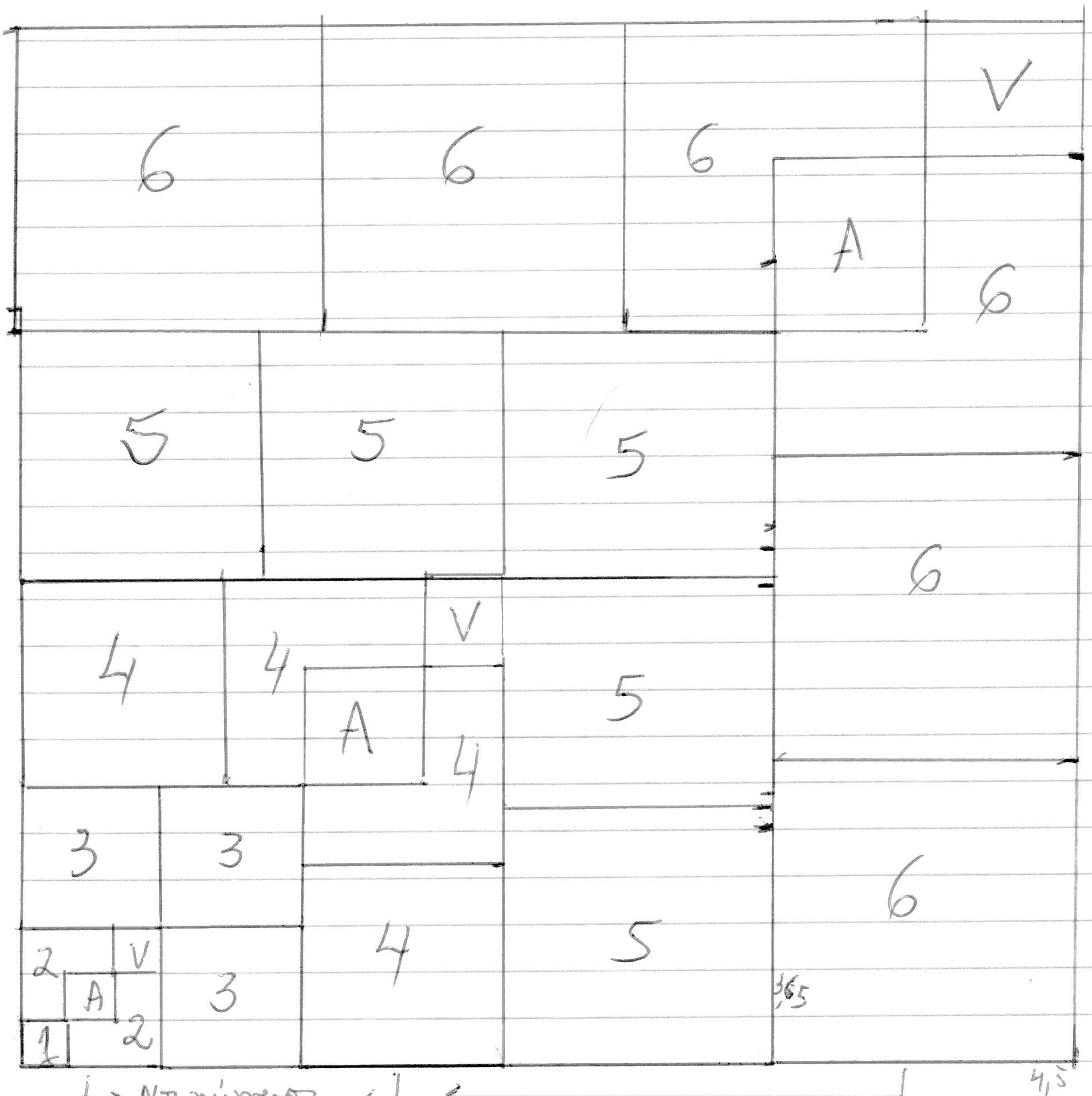
6
6
6
V
A
6
5
5
5
6
4
4
V
A
5
4
3
3
4
5
6
2
V
A
3
1
2
3,5
4,5

↳ Nos números ↵ ← pares há sobreposição de quadros.

$\boxed{A}$ = The even squares overlap but also leave uncovered gap $\boxed{V}$ and $\boxed{A}=\boxed{V}$.

Proof: $1^3+2^3+3^3+4^3+5^3+6^3+\dots+n^3$ in the plane:

$$1\times1^2+2+2^2+3\times3^2+4\times4^2+5\times5^2+6\times6^2+\dots+n\times n^2$$

20220508

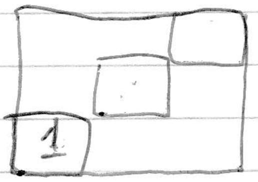

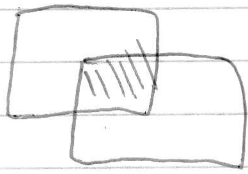

$1^3 = 3$ (?)

$2 \times 2^2 = 8$

Neste caso, 7

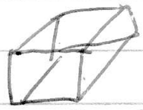

$1^3 = 1$
(volume)

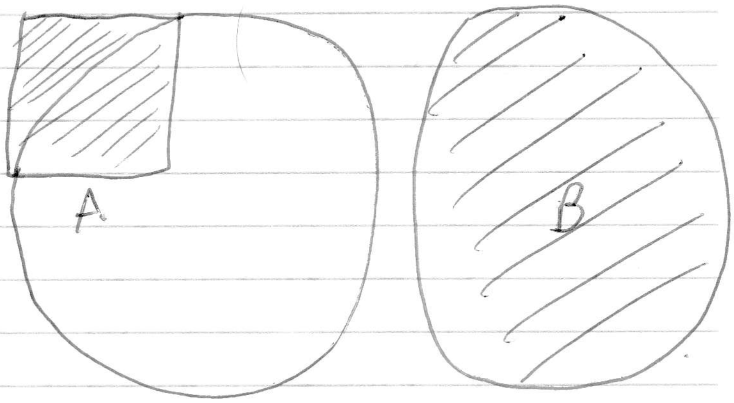

$A \times 3\frac{1}{7} = B$

—

20220510

https://galileo.ou.edu/exhibits/
elements-geometry-1570
https://archive.org/details/
elementsgeometr00eucl

—

the-elements-of-euclid-888-to-2008

— John Clark

—

20220511

The Babylonian theorem:
the mathematical journey to
Pythagoras and Euclid.
Peter Strom Rudman
256 pág. 2010.
https://archive.org/details/
babyloniantheorem00rudm
ISBN 9781591027737

—

booksc.eu

—

Tutte Polynomial

called the dichromate or Tutte-Whitney polynomial.

The Tutte Polynomial is a polynomial in two variables x y which can be defined for a graphic, matrix, or, even more generally, a matroid.

—

20220512

https://englishtest.duolingo.com

20220514

ARQ-R---20220514.7z

180 GB - 51720 files

—

Hoje, atingi 1,003 TB de dados armazenados na "cloud".

—

20220515

$E = mc^2$

Nesta formula, quanto mais rápido um objeto viaja, maior fica a sua massa, pois ele tem que gastar energia e aplicar força para acelerar.

Lembre-se que conforme você coloca força maior é a massa do objeto e portanto novos calculos deverão ser feitos.

A energia necessária é calculada sempre com a nova massa.

Para uma grande massa acelerar do mesmo jeito, a massa precisará de uma força muito maior, indicando que a massa e energia são inseparáveis.

1915 - Einstein alterou a Gravidade de Newton.

A gravidade não era uma força e concluiu que a Terra girava ao redor do Sol devido à geometria do Universo, que é deformada pelo astro.

Grandes massas criam curvaturas significativas na malha do espaço-tempo, acarretando a atração dos corpos menores.

Massa + Equações de Maxwell
↳ Data-Novo Vivo-011-2022/
Books-011-2022/
20220515/
Maxwell equations/

Lei Newton

$F = m a$

A energia para acelerar um objeto com massa m também é massa, pois há uma correlação direta entre massa e energia.

Veja: Pontos de Lagrange

—

Unity of Forces in the Universe
(9789971950156) A. Zee 2 volumes
Trent University.

—

Force and Matter - Buchner

Study on Inertia as a Gravity Induced Property of Mass, in an Infinite Hubble Expanding Universe
Jeroen van Engelshoven
Source: https://www.hindawi.com/journals/amp/2013/801574/

—

20220520

www.ugr.es/~jgodino/edumat-maestros/

https://todogeometria.wordpress.com/

Geometría paso a paso
ISBN 978-84-7360-582-3
Álvaro Rendón Gómez 496 pág.

—

__obs__: tebanflores.com

—

https://dibujotecni.com

—

20220521
olipdf.com

—

Eu desisto → I gave up.

—

20220523 Computers.
lightwave electronics
→ femtosecond timescale

Cosmologia → Teoria dos Conjuntos Causais.

A curvatura do espaço-tempo no centro de um buraco negro é infinita.

---

https://arxiv.org/abs/2109.11953
→ If time had no beginning

Singularidades → Buraco Negro
(infinitas)

Outra característica das singularidades é que elas são infinitesimais
↳ próximo de zero.

20220525
Kähler geometry.

Kumihimo → técnica (1652)
trançado japonês.

20220526

mathematica.sns.it/opere/308/

→ Gino Loria

/autori/1058/

(1862-1954)

—

/opere/310/

mathematica.sns.it/opere/8/

teorie geometriche

—

Pappi Alexandrini

mathematica.sns.it/opere/132/

—

Non-Euclidean Geometry

Roberto Bonola

mathematica.sns.it/opere/136

—

archive.org/details/

in.ernet.dli.2015.96359

1911 - 452 pages.

data S T Q Q S S D

Important:

mathematics.sns.it/opere/
lista-opere.html

—

Riemann (1826-1866)
Mathematical Papers
https://www.maths.tcd.ie/pub/
HistMath/People/Riemann/
Papers.html

—

Euclid's Elements
aleph0.clarku.edu/~djoyce/
java/elements/elements.html

—

20220527m
DATA-NovoVivo-011-2022/
Books-011-2022/
20220526/
Storia Delle Matematiche/
Il parallelismo di Clifford/

— Le linee curve diametrali
— delle curve algebriche in

particolare i loro assi di simmetria/

Lonia/

Curve_Alg_Trasc_I.pdf

Pág. 05 (11) - Capitolo II

La Circonferenza

$$\pi = \left(\frac{16}{9}\right)^2 = 3{,}1604\ldots$$

Obs:

$\left(\frac{16}{9}\right)$ ^2 = melhor

$$\left(\frac{16}{9}\right)^{\wedge} 1{,}989574595 \cong$$

$$\underbrace{3{,}14159265\ldots\ldots}_{\text{correct(?)}}$$ (checar?)

-

$\frac{8}{4{,}5}$ or $\frac{2}{1.125}$

$$\left(\frac{2}{1.13}\right)^{\wedge}2 = 3{,}132586\ldots$$

(cont.)

$\left(\frac{2}{1.127}\right)^\wedge 2 = 3{,}1492\ldots$

$1.1279 = 3{,}1442\ldots$

$1.128 = 3{,}1437\ldots$

$1.129 = 3{,}1381$ $\Big\} X$

obs:

$\left(\frac{3}{1.129}\right)^\wedge 1.129 = 3{,}01424\ldots$

$$\begin{array}{r} 1.129 + \\ \underline{0{,}563} \quad 50\% \\ 1.692 \end{array}$$

$(3/1.692)^\wedge 2 = 3{,}1437 \cong$

$(2/1.128)^\wedge 2$

obs: De 2 para 3 = 50%

Cont:

$1.1284 = 3{,}1414\ldots$

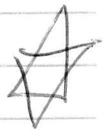

$1.1284 \div 2 = 0,5642$

$$\begin{array}{r} 1.1284 \\ 0.5642 \\ \hline 1.6926 \end{array}$$

$(3/1.6926)^2 \cong 3,1414\ldots$

Important: $2^2 = 4$cm

mesmos números em ordem diferente.

(2/1.1284)^2.00006452

$\cong 3,1415926662$

↳ mal(?) 53.

((2/1.1284)^2.00006451̄5 $\cong$

3.1415926533$\underline{1}$

↓

(58)

(2/1.1284)^2.0000645151532882...

$\cong \underbrace{3.14159265358979 32}_{OK}1$

↓

3

Em Resumo:

Com base na curvatura sobre um plano:

data S T Q Q S S D

$(2/1.1284)^\wedge 2 = 3{,}141476$

$\updownarrow$

$(2/1.1284)^\wedge \underline{2}.00006 = 3{,}141584$

$\left(\frac{16}{7}\right)^\wedge 1{,}385 \cong 3{,}1422$

$\left(\frac{16}{5}\right)^\wedge 0.984 = 3{,}140\ldots$

$\left(\frac{16}{4}\right)^\wedge 0.825748 = 3{,}141592\underset{\underset{6}{3}}{3}$

$\boxed{4 ^\wedge 0.825748}$

$3 ^\wedge 1{,}0419780\overline{4}5 \cong 3{,}1415926501 \cong$

(35)

$\downarrow$

$1{,}0419780 4599 \cong$

$3{,}1415926535 8\underset{\underset{9}{1}}{\underline{2}}$

$1{,}0419780459921 8586.$

$\cong$

$3{,}1415926535897932 2\underline{20}$

(38)

Aprofundar: ver mais na frente

Ⓐ ③ ^ 1,04197804$\boxed{59921}$8586 =

3,1415926535897932$\underset{(3)}{2}$

este mesmo valor

Ⓑ ③ X 1,04197804599218586 =

3,12593413797655758

A - B = 0,01565851561...

obs: Fazer estes 2 valores se aproximarem

Ⓒ pi/3 = 1,047197551...

$$
\begin{array}{r}
1{,}047197551\ldots \\
-\ 1{,}041978045\ldots \\
\hline
0{,}005219505\ldots
\end{array}
$$

(erro de milésimos)

(A-B) ÷ Ⓒ = 2,999 999 999 99999 661...

(14 noves)

$\boxed{\cong ③}$ checar!!!

20220530 III

(A − B) ÷ C ≅ 2.999... ③

compare com a equação

(2/1.1284)^2 = 3,141476...

(2/1.1284)^2.00006 = 3,141584

Simplificando 2.00006 ÷ 2 =

1.00003 = ③

Aprofundando:

(2/1.1284)^2.00006 =

3,14158453...

Pi − [(2/1.284)^2.00006] =

0,000008118...

Obs: Pi = 100%

0,000008118 = x

Na máquina de calcular:

((pi) − ((2/1,284)^2.00006))*100)/pi =

0,0002584 % (erro)

Obs: Pi = 3,141592653589793238...

20220530III

Physicists Rewrite the Fundamental Law that leads to disorder.

Source: https://www.quantamagazine.org/physicists-trace-the-rise-in-entropy-to-quantum-information-20220506/

—

Teoria dos conjuntos causais
→ "átomos de espaço-tempo".
→ resolve a questão de singularidade

Ideia da "transformação assincrônica"
→ os elementos nascem de forma parcial e não total.

—

→ Princípio da simetria da mecânica quântica.

As simetrias descrevem características do espaço-tempo e partículas que permanecem inalterados apesar de ter havido alguma transformação.

→ "Universo Espelho".

→ 'anti-universo'.

↳ o tempo corre para trás e a antimatéria é dominante.

Desta forma, a simetria seria cumprida.

O universo espelho ofereceu uma explicação para a matéria escura.

Big-Bang → matéria-escura ⇒ "neutrinos destros".

20220531

$\left(\frac{3}{1.129}\right)\hat{}\;1.129 = 3{,}01424\ldots$

$\left(\frac{3}{1.128}\right)\hat{}\;1.128 = 3{,}014311\ldots$

$\left(\frac{3}{1.1284}\right)\hat{}\;1.1707 = 3{,}14157\underline{7}4$

(obs. $1.1708 = 3{,}1418\ldots$)

Estudar:

$\left(\frac{3}{1.1284}\right)\hat{}\;1.170706 = 3{,}14159267\underset{(535)}{\underline{77}}$

Obs:

$\left(\frac{3}{1{,}132}\right)\hat{}\;1.132 = 3{,}014\ldots\ldots 0500$

(muito interessante)

Ver depois!!!

Aprofundando:

$\left(\frac{3}{1.1284}\right)\hat{}\;1.170705 \cong 3{,}1415896\ldots$

→ repetiu padrão.

$\left(\frac{3}{1.1284}\right)\hat{}\;1.170705\boxed{99212}2 \cong$

$3{,}1415926535\underset{(9)}{83}$

Voltando um pouco:

$\left(\frac{2}{1.1284}\right)$ ^ 2 = 3,14147766 .... 630 (par)

↓ até

2.00006 ≅ 3,1415845...

Comparar

$\left(\frac{3}{1.1284}\right)$ ^ 1,17070~~6~~ ≅ 3,1415742...

$$\begin{array}{r} 2.0006 \\ -\ 1.1707 \\ \hline 0,8299 \end{array}$$ (?)

expandido:

1,17070 $\boxed{59921}$ 2

checar!

checar se dá 6000...

$$\begin{array}{r} 2.00006 \\ 1.70706 \\ \hline 0,2930 \end{array}$$ ⇒ final par.

$$\left(\frac{2}{1.1284}\right)^{\wedge}2.0000599 \cong 3,1415843551$$

$$\left(\frac{3}{1.1284}\right)^{\wedge}1.17070599 \cong 3,1415926470$$

$$\begin{array}{r} 2.0000599 \\ 1.17070599 \\ \hline 0,82935391 \end{array}$$

$$\left(\frac{2}{1.1284}\right)^{\wedge}2.0000599 = 3,1415843551$$

$$2.00000599 = 3,1414874224$$

Comparar:

$$\left(\frac{3}{1.1284}\right)^{\wedge}1.170706 = 3,1415926777$$

$$1.17070599 = 3,1415926470$$

$$\left(\frac{2}{1.1284}\right)^{\wedge}2 = 3,1414766523$$

$$\left(\frac{3}{1.1284}\right)^{\wedge}1,1707 = 3,1415742465.$$

$$\therefore \left[\left(\frac{3}{1.1284}\right)^{\wedge}1,1707\right] - \left[\left(\frac{2}{1.128}\right)^{\wedge}2\right] =$$

$$0,00009759\ldots$$

$\left(\frac{2}{1.1284}\right)^{\wedge} 2.0000599 =$

$2.\underbrace{0000}_{(4)}599 = 3{,}1415843551$

$\left(\frac{3}{1.1284}\right)^{\wedge} 1{,}1707 = 3{,}14\underline{\underline{157424}}65$

∴

$\left(\left(\frac{2}{1.1284}\right)^{\wedge} 2.0000599\right) - \left(\left(\frac{3}{1.1284}\right)^{\wedge} 1{,}1707\right) =$

$0{,}000101086$

Invertendo a ordem

$\left(\frac{3}{1.1284}\right)^{\wedge} 1.17070599 -$

$\left(\frac{2}{1.1284}\right)^{\wedge} 2.0000599 =$

$0{,}000000829188335....$

Important:

Você mantendo a constante

$\left(\frac{2}{1.1284}\right)^{\wedge} 2.0000599$ e

alterando a segunda parte,

da equação, obteremos o seguinte:

a) Quando

$\left(\frac{3}{1.1284}\right)^{1.1707}$, obtemos que a fração

$\left(\frac{3}{1.1284}\right) < \left(\frac{2}{1.1284}\right)$ (menor na equação)

embora

$\frac{3}{1.1284} \cong 2,658$ seja maior

$\frac{2}{1.1284} \cong 1,771$

b) Quando alteramos o exponencial de 1.1707 para 1.17070599, obteremos o seguinte

$\left(\frac{3}{1.1284}\right) > \left(\frac{2}{1.1284}\right)$

Em resumo:

Na hipótese (a) a diferença de valor entre uma equação e outra é de

0,000101086...

Na hipótese (b) a diferença é muito menor:

$$0,\underbrace{00000}_{(5)}829188335...$$

Analizando mais profundamente veremos o seguinte:

$$\left[\left(\frac{2}{1.1284}\right)^{\wedge}2.0000599 - \left(\frac{3}{1.1284}\right)^{\wedge}1.1707\right] - \left[\left(\frac{3}{1.1284}\right)^{\wedge}1.17070599 - \left(\frac{2}{1.1284}\right)^{\wedge}2.0000599\right] =$$

$$0,000001816774722...$$

Interessante:

A mudança de ordem ocasionou um número interessante

$-6,2831$

(2π negativo) → como em um espelho (mirror)

Equação:

~~$\left(\frac{2}{1.1284}\right)$ ^ 1.17070599~~ $\cancel{=}$

$\left(\frac{2}{1.1284}\right)$ ^ 2.0000599 $-$

$\left(\frac{3}{1.1284}\right)$ ^ 1.1707 $-$

$\left(\frac{3}{1.1284}\right)$ ^ 1.17070599 $-$

$\left(\frac{2}{1.1284}\right)$ ^ 2.0000599 $=$

$= 6,28316689\ldots$ $(= -2\pi)$

obs $6,28318530\ldots$ $(= 2\pi)$

Estudar e inverter a ordem...

Arrumando para facilitar: (Ajustando)

$\left(\frac{3}{1.1284}\right)$^1.17070599 −

$\left(\frac{3}{1.1284}\right)$^1.1707 − → 599

$\left(\frac{2}{1.1284}\right)$^$2.0000\textcircled{6}$ − → 599

$\left(\frac{2}{1.1284}\right)$^$2.0000\textcircled{6}$ =

−6,2831506..

Alterando ⇓

2.00006 → 2.0000599

∥ ✓

= −6,283150309...

Obs: Na folha anterior tínhamos a seguinte ordem.

2 3 3 2

Nesta 3 3 2 2

Obs: Há uma diferença de valores.

Na página anterior:

$\rightarrow 6,28316689$

Nesta:

$\rightarrow 6,283150309\ldots$

Obs: Há um erro (checar!!!)

Em detalhes: (item por item depois)

① Interessante tirando o último item, observamos:

$\left(\frac{3}{1.1284}\right)^{1.17070599} -$

$\left(\frac{3}{1.1284}\right)^{1.1707} -$

$\left(\frac{2}{1.1284}\right)^{2.0000599} =$

$\rightarrow 3,1415659\ldots$

recolocando o último item

$\left(\frac{2}{1.1284}\right)^{2.0000599}$ na primeira parte desta equação. Temos

$\rightarrow 6,28316689\ldots$

Alterando os valores:
2.0000599 para
2.00006
e fazendo alguns ajustes:

$$\left(\frac{2}{1.1284}\right)^{\wedge}2.00006-\left(\frac{3}{1.1284}\right)^{\wedge}1.170708 -$$

$$\left(\frac{3}{1.1284}\right)^{\wedge}1.1707-\left(\frac{2}{1.1284}\right)^{\wedge}2.00006 =$$

-6,2831730
(853) 2π

alterando este para
1.170704 = -6,2831853|5559
passou o valor

1.170704983 = -6,2831853033
(071)

Em resumo
2 ^2.00006 3 ^1.170706 3 ^1.170706 2 ^2.00006
= -6,2831853555
(071).

2 ^2.00006 3 ^1.1707 3 ^1.1707 2 ^2.00006
≃ -6,2831484
(853).

Provisoriamente a equação:

Ⓐ $\left(\frac{2}{1.1284}\right)$ ^ 2.00006 −

Ⓑ $\left(\frac{3}{1.1284}\right)$ ^ 1.1707 −

Ⓒ $\left(\frac{3}{1.1284}\right)$ ^ 1.1707 −

Ⓓ $\left(\frac{2}{1.1284}\right)$ ^ 2.00006 =

−6.28314849...

obs: Este valor sem considerar a curvatura com espelhamento (mirror) em um plano.

Tirando o termo Ⓐ obteremos:

−3,141584534...

Tirando o termo Ⓓ obteremos:

−3,141563958...

Fazendo a subtração do ~~primeiro~~ segundo termo do ~~segundo~~ primeiro:
(maior − menor)

0,00002057...

Important:

$\left(\frac{3}{1.1284}\right)$ ^ 1.1707 +

$\cancel{\left(\frac{3}{1.1284}\right)\text{ ^ }1.1707} -$ (eliminando repetição de termos)

$\left(\frac{2}{1.1284}\right)$ ^ 2.00006 =

6,283158781.

Comparando (os dois posicionamentos)

$\left(\frac{2}{1.1284}\right)$ ^ 2.00006 −

$\left(\frac{3}{1.1284}\right)$ ^ 1.1707 =

0,000010288... (uma) discrepância.

Se for

$\left(\frac{2}{1.1284}\right)$ ^ 2.0000599 −

$\left(\frac{3}{1.1284}\right)$ ^ 1.1707 =

0,0000101086580...

Outros estudos:

$2\left(\frac{3}{1.116}\right)\,^\wedge 1.116 = 3{,}14898\ldots$

$\left(\frac{3}{1.116}\right)\,^\wedge 1.116 = 3{,}014908\ldots$

$\left(\frac{3}{1.5}\right)\,^\wedge 1.5 = 2{,}828427124\ldots$

$=$

$\sqrt{2} \times 2 = 2{,}828427\ldots$

Obs. Fazer análise entre exponencial e raiz quadrada de 2.

→ Veja a sua relação com a realidade vigente.

$\left(\frac{3}{1.116}\right)\,^\wedge 1.1576 \cong 3{,}141517\ldots$

$\left(\frac{3}{1.116}\right)\,^\wedge 1.1576241698 \cong$

$3{,}1415926535\ldots$

data

Outro:

$$\left(\frac{2}{1.1284}\right)^{\wedge}2.00006 = 3,141584\ldots$$

$$\left(\frac{3}{1.116}\right)^{\wedge}1.16 = 3,14898$$

$$\left(\frac{2}{1.116}\right)^{\wedge}2 = 3,21167\ldots$$

$$\left(\frac{2}{1.116}\right)^{\wedge}2.00006 = 3,21178\ldots$$

Obs:

$$\boxed{\left(\frac{3}{1.116}\right)^{\wedge}1.1576 \cong 3,141517}$$

par

Voltando:

$$\left(\frac{3}{1.5}\right)^{\wedge}1.5 = 2\sqrt{2}$$

Adequando:

$$\left(\cancel{\frac{6}{3}}\right)^{\wedge}1.5 = 2\sqrt{2}$$

$$2^{\wedge}1.5 = 2\sqrt{2} \text{ ou } \left(2^{\wedge}\frac{1}{2}\right)\times 2$$

Obs: $\sqrt{2} \cong \frac{10}{7} \cong 1,\overline{428571}$

$\left(\frac{10}{7}\right)^{\wedge}2 \overset{or}{=} \frac{100}{49} = 2,040816\ldots$

$(2,040816 32)$

$\frac{Pi}{\sqrt{2}} = 2,221441 4690\ldots$

$\frac{Pi}{\sqrt{3}} = 1,813799\ldots$

$\frac{Pi}{\sqrt{2}} - \frac{Pi}{\sqrt{3}} = 0,40764$

Obs:

$\frac{Pi}{\sqrt{1,5}} - \frac{Pi}{\sqrt{3}} = 0,7513002\ldots$

Important:

$\frac{Pi}{\sqrt{0.\bar{9}}} \cong \pi$

$$\boxed{\frac{Pi}{\sqrt{4.9342}} \cong \sqrt{2}}$$

chucar!!

data

Ⓢ Ⓣ Ⓠ Ⓠ Ⓢ Ⓢ Ⓓ

Estudar:

$$\frac{\pi}{\sqrt{4.9347999}} \cong 1,414213\underline{8}$$

$$(5) = \sqrt{2}$$

$$\frac{\pi}{\sqrt{4.9348}}$$

$$\hookrightarrow 4,934802201 \cong$$

$$\underline{1,414213\underline{5623}}$$

ok

Offers:

$$\frac{\pi}{\sqrt{9.8}} = 1,003$$

$$\frac{\pi}{\sqrt{9.7}} = 1,008$$

−

$$\frac{\pi}{\sqrt{9.7}} - \frac{\pi}{\sqrt{9.8}} = 0,005159648\ldots$$

Interessante:

$$\left(\frac{\pi}{\sqrt{2}}\right) - \left(\frac{\pi}{\sqrt{3}}\right) \times 7,706742 \cong$$

$$3,141592\underline{5}$$

(6)

$\sqrt{2} \cong \frac{10}{7}$

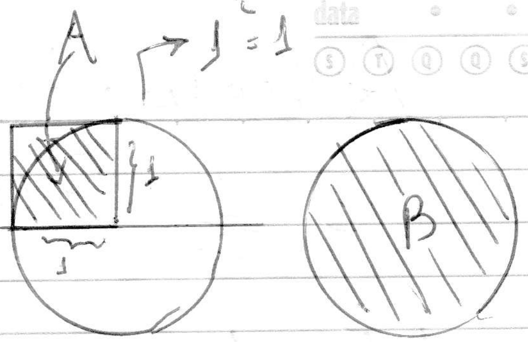

Analisando:

$\frac{10}{7} = 1,\overline{428571}$

$\frac{1}{7} = 0,\overline{142857}$

obs:

$\frac{10}{7} \div \frac{1}{7} = \frac{10}{(X \text{ vezes})}$

$A \times 3\frac{1}{7} = B$

$\Downarrow$

$3 + (1/7.06251331)$

$\cong 3,1415926535\ldots$

$1^2 = 1$ e se $1^2 = 2$

Portanto:

$2 \times (1,5 + 1/7) = 3,\overline{285714}$

Analisar o erro ou discrepância envolvidos.

Atente-se que $3\frac{1}{7}$ ou $\frac{22}{7} =$

$3,\overline{142857}$

obs: $((1.5 + 1/7)^* 2) - 22/7 =$ ↓ Estudar !!!

$0,142857$ (erro de $\frac{1}{7}$)

Aprofundando:

$3\ \frac{1}{7}$ fazendo uma redução pela metade, teremos?

$1{,}5\ \frac{1}{14} = (14 \times 1{,}5) + 1 = 22$

$\therefore \frac{22}{14}$

$(22/14) \times 2 = 3{,}\overline{142857}$

Corrigindo a página anterior:

$2 \times (1{,}5 + 1/7) = 3{,}\overline{285714}$,

mas o correto é

$\underbrace{2 \times (1{,}5 + 1/14)} = 3{,}\overline{142857}$

Neste caso, se $1^2 = 2$,

portanto, a área de Ⓐ

1 de lado 1 será igual 2.

Continuando:

$$2 \times (1{,}5 + 1/(14)) = 3{,}\overline{142857}$$

Atente-se:

$$\left(\frac{3}{1.5}\right)^{\wedge} 1.5 = 2{,}828427124\ldots$$

$$\cong$$

$$\sqrt{2} \times 2$$

Observamos que:

$$2^{\wedge} 1{,}65149612944 \cong$$

$$3{,}1415926535\ldots \quad (\cong \pi)$$

Veja:

$$\left(2^{\wedge} 1{,}65149612944\right) - \left(\frac{3}{1{,}5}\right)^{\wedge 1{,}5} =$$

$$\underbrace{0{,}313165528\ldots}$$

$$\left(\frac{\pi}{\sqrt{3}}\right) - 2^{\wedge} 1{,}65149612944 - \left(\frac{3}{1{,}5}\right)^{\wedge 1{,}5} =$$

$$1{,}5006338\ldots$$

(checar!!! com mais precisão)

20220601 m Meu aniversário! 53 anos.

$\left(\frac{2}{1,128}\right) = 1,770304\ldots \quad \left(\frac{3}{1.128}\right) = 2,65957\ldots$

$\left(\frac{3}{1.128}\right) - \left(\frac{2}{1.128}\right) = 0,886524822\ldots$ Ⓐ

$\frac{\pi}{(3/1.128)} = 1,1812\ldots \quad \frac{\pi}{(2/1.128)} = 1.771858$

$\frac{\pi}{(2/1.128)} - \frac{\pi}{(3/1.128)} = 0,5906194$ Ⓑ

Aprofundando:

$A - B = 0,29632428\ldots$

Mantendo-se a ordem 2 e 3:

$\left(\left(\frac{3}{1.128}\right) - \left(\frac{2}{1.128}\right)\right) - \left(\left(\frac{\pi}{3/1.128} - \frac{\pi}{2/1.128}\right)\right) =$

$\boxed{1,47}7$ [illegible]... $1,477144\ldots$

$\hookrightarrow \underline{1,50}$ (ajustar!!

$\cong$

obs:

$2 \times \sqrt{2} = \left(\frac{3}{1.5}\right)^\wedge 1.5 = 2,8284\ldots$ ou $2^\wedge \frac{1}{2} =$

obs: Interessante!!!

$$\left(\left(\frac{3}{1,1284}\right)-\left(\frac{2}{1,1284}\right)\right)-\left(\left(\frac{\pi}{3/1.1284}\right)-\left(\frac{\pi}{2/1.1284}\right)\right)=$$

1,4770394...

Especulação:

$$\left\{\left(\frac{3}{1,1}\right)-\left(\frac{2}{1,1}\right)\right\}-\left\{\left(\frac{\pi}{(3/1,1284)}\right)-\left(\frac{\pi}{2/1,128}\right)\right\}=$$

$\underline{1,4999}19767476....$ Ⓘ

$\cong 1,5$

$2 \hat{} 1,4999 \cong 2,8282310...$

$\left(\frac{2}{1,1284}\right)\hat{}2 \cong 3,1414766...$

$\downarrow$

$$\frac{\left\{\left(\frac{2}{1,1284}\right)\hat{}2\right\}\times 100}{\pi\ (\text{Pi})} = 99,996307\,\%$$

do valor do π.

Extrapolando:

$$\left\{\left(\frac{3}{1.1}\right)-\left(\frac{2}{1.1}\right)\right\}-\left\{\left(\frac{\pi}{(3/1,1707)}\right)-\left(\frac{\pi}{2/1,1707}\right)\right\}=$$

1,52206...

Usando este valor na equação:

$$\left(\frac{3}{1,52}\right)^\wedge 1,52 = 2,81074\ldots$$

$$\left(\frac{3}{1,522}\right)^\wedge 1,522 = 2,80894\ldots$$

Simulações: (Interessante)

$$\left\{\left(\frac{3}{1,1284}\right)-\left(\frac{2}{1.1284}\right)\right\} - \left\{\left[\left(\frac{\pi}{3/1.1707}\right)-\left(\frac{\pi}{2/1.1707}\right)\right]\right\} =$$

$$1,4991876\ldots$$

$$\hookrightarrow \cong 1,5 \qquad \textcircled{II}$$

Obs: $2^\wedge 1,499 = 2,8264\ldots$

Estudar as equações:

$$\left.\begin{array}{l} \textcircled{I} \text{ e } \textcircled{II} \to 1,499\underline{1876} \\ \hookrightarrow 1,4999\underline{1976} \\ \qquad (49) \end{array}\right\} \underset{100}{\neq}$$

Conjecturando outras possibilidades:

→ valores 2 e 3.

$\left(\frac{2}{1.1284}\right)^{\wedge}2 = 3,1414766.$ ①

$\left(\frac{2}{1.1284}\right)^{\wedge}1.1284 = 1,9075806\ldots$

$\left(\frac{3}{1.1284}\right)^{\wedge}2 = 7,068322\ldots$

$\left(\frac{3}{1.1284}\right)^{\wedge}1.1284 = 3,01\underline{42}8\underline{4}8\underline{2}\ldots$

Obs: $2\sqrt{2} = 2,8284$

$\sqrt{2} = 1,41\underline{42}$

$\left(\frac{2}{1.1707}\right)^{\wedge}2 = 2,91856\ldots$

$\left(\frac{3}{1.1707}\right)^{\wedge}1.1284 = 2,89167571\ldots$

$\left(\frac{3}{1.1707}\right)^{\wedge}1.1707 = 3,00909963\ldots$

(interessante)

Faltou:

$\left(\frac{3}{1.1284}\right)^{\wedge}1.1707 = 3,1415742465\ldots$

$\left(\frac{2}{1.1284}\right)^{\wedge}1.1707 = 1,954327\ldots$

Interessante:

$\left(\frac{3}{1.1284}\right)^{\wedge}1.1707 = 3,1415742\ldots$ ③

Esta equação utiliza os dois

números interessantes e mais próximos obtidos.

vejamos com 2.00006 :

$$\left(\frac{2}{1.1284}\right)^{\wedge} 2.00006 = 3,1415845\ldots \quad ②$$

Subtração: ② - 1 =

$0,0001078\underline{88}\ldots$

Subtração: 2 - 3 =

$0,0000102\underline{88}\ldots$] ≠ 500

(+0)

Outro estudo:

Ⓐ

$$\left\{\left(\frac{3}{1.1707}\right) - \left(\frac{2}{1,1707}\right)\right\} - \left\{\left(\frac{\pi}{3/1.1284}\right) - \left(\frac{\pi}{2/1.1707}\right)\right\} =$$

$= 1,51146334$

Ⓑ

$$\left\{\left(\frac{3}{1.1707}\right) - \left(\frac{2}{1.1707}\right)\right\} - \left\{\left(\frac{\pi}{3/1,1707}\right) - \left(\frac{\pi}{2/1,1284}\right)\right\} \times 3 =$$

$= 3,1091018048\ldots$

$\downarrow$

$\times\ 3,038 =$

$3,14156\underline{101}$

$(926\ldots)$

Ⓒ $3 \times \left\{ \left(\frac{3}{1.1707} \right) - \left(\frac{2}{1,1707} \right) \right\} - \left\{ \left(\frac{\pi}{3/1.1284} \right) - \left(\frac{\pi}{2/1.1707} \right) \right\} =$

$= 3,15339\ldots$

~~x 3,28~~

Ⓓ $3,284 \times \left\{ \left(\frac{3}{1.1707} \right) - \left(\frac{2}{2/1.1284} \right) \right\} - \left\{ \left(\frac{\pi}{3/1.1707} \right) - \left(\frac{\pi}{2/1.1284} \right) \right\} =$

$= 3,1413793\ldots$

Obs:

Para o número 3 ⇛ 1.1707

Para o número 2 ⇛ 1.1284

Ajustar as casas decimais seguintes para a obtenção do valor a ser obtido com base na curvatura do plano em estudo.

Extrapolando: Caso D:

$3,2842699 = 3,14159260\underline{026}$

$3,2842699\,674915 = 3,1415926535\,8973\ldots$

(9)

20220601$_{III}$ (cont.)

$$\left(\frac{22}{15,6}\right) \times 2.0056 = 2,8284\ldots$$

$$\sqrt{2} = 1,4142 \xrightarrow{\times 2}$$

Obs: Bem próximo.

Estudo:

$$\left(\frac{22}{15,6}\right) \hat{}\, 2.0056 = 1,992655554\ldots$$

Interessante:

$$\left(\frac{22}{15,6}\right) \times 1,0056 = 1,41\underset{(42)}{\underline{81}}\ldots$$

Aprofundando:

$$\left(\frac{22}{15,6}\right) \times 1,0023 = 1,4135$$ (exato)

(interessante!!!) ✡

Obs: Poderá ser este valor por causa da curvatura do plano

→ muito próximo do

$$\sqrt{2} = 1,4142135623\ldots$$

Obs.: Como o espelhamento o valor deverá diminuir.

$$\boxed{\left(\frac{22}{15,6}\right) \times 1,0023} = 1,4135$$

Obs: Adequar nas equações anteriores

Simplificando:

$$\left(\frac{11}{7,8}\right) \times 1,0023$$

Off:

$$\left(\frac{11}{7,8}\right) \times 1,0028 =$$

1,4142051...

(135

Importantíssimo:

$$\left(\frac{11}{8}\right) \times 1,028 =$$

1,4135 (exato)

Off: $\frac{11}{8} = 1,375$ (exato)

Outras alterações:

$$\left(\frac{11}{8}\right) \times 1,0285 = 1,4141875$$

(exato)

Outros resultados:

$$\left(\frac{11}{8}\right) \times 1,0285189 =$$

1,4142134875

(exato)

$\left(\frac{11}{8}\right) \times 1{,}02851895 =$

$= 1{,}4142135625$ (exato)

(6237)

Aprofundando:

$$\underbrace{\left(\frac{3}{1.1284}\right)^{\wedge}1.1707}_{\pi} - \underbrace{\left(\frac{11}{8}\right) \times 1{,}028}_{\sqrt{2}} = \text{sem alterações.}$$

Outros estudos:

$\left(\frac{3}{1{,}1284}\right)^{\wedge}1.1702 = 3{,}140038\ldots$

$1.1640 = 3{,}121060004\ldots$

$1.1607 = 3{,}1110052\ldots$

$1.1571 = 3{,}100073401\ldots$

Off:

$\left(\frac{3}{0{,}1284}\right)^{\wedge}1{,}1707 = 40{,}01002102\ldots$

Outros:

$\left(\frac{3}{1.1284}\right) \times 1{,}1816 = 3{,}141439205\ldots$

$\times 1{,}1816577 = 3{,}14159260900$
$3899\ldots$

data ● ●
Ⓢ Ⓣ Ⓠ Ⓠ Ⓢ Ⓢ Ⓓ

$$\left(\frac{3}{1.1284}\right) \times \underset{✡}{\underline{1,1816}}5771674 =$$

$$3,14159265350939\ldots$$

(8979)

Outras combinações:

$$\left(\frac{3}{1,1707}\right) \wedge 1,2164901169 =$$

$$3,14159265358 15\ldots$$

(897).

$$\wedge\ 1,2164901169277$$

$$= 3,1415926535897\,77505\ldots$$

(9328...)

Variações: (multiplicações)

$$\frac{3,1415}{(3/1.1707)} = 1,2259180 16$$

$$\frac{3,1415}{(2/1.1707)} = 1,838877025$$

(exato)

$$\frac{3,1415}{(3/1.1284)} = \underbrace{1,1816}_{✡}22\overline{86}$$

$$\frac{3,1415}{(2/1.1284)} = 1,772434 \text{ (exato)}$$

Jandaia

Conferindo:

$\left(\frac{3}{1.1284}\right)^{1}\ 1{:}1707 = 3{,}1415\underset{742}{\cancel{926}}\ldots$

talvez seja esta a equação do π ([illegible])

Vamos ajustá-la:

$\sqrt[1{,}1707]{3{,}1415} = 2{,}65857801\,95\ldots$ Ⓑ

$\frac{3}{1.1284} = 2{,}\underset{65863169\,0\ldots}{\cancel{66039145\,90}}\ldots$ Ⓐ

Calculando:

$A - B = 0{,}000053671\,3032\ldots$

Que coisa!!!

Interessante esta constatação

$$\frac{3}{1.1284} \simeq \sqrt[1.1707]{3{,}1415}$$ Ⓐ

Outra forma de calcular:

$$\sqrt[\frac{3}{1.1284}]{3{,}1415}$$

Obs: Fazer também com $\sqrt[\frac{2}{1.1707}]{3{,}1415}$

⇘ $$\sqrt[\frac{3}{1.1284}]{3{,}1415} = 2{,}6586$$

Importante lembrança:

$$\frac{3{,}1415}{(2/1.1707)} = 1{,}838877025$$ (exato)

$$\frac{3{,}1415}{(2/1.1284)} = 1{,}772434$$ (exato).

$$\left(\frac{3{,}1415}{(2/1.1284}\right)^{\wedge}2 = 1{,}00000743206$$ (exato)

$$\left(\frac{3{,}1415}{(2/1{,}1707}\right)^{\wedge}2 = 1{,}07638666658375$$ (exato)

Checando:

$$\left(\frac{3}{(2/1.1284}\right) = 1{,}6926$$ (exato)

$$\left(\frac{3}{(2/1.1707}\right) = 1{,}75605$$ (exato)

$$\left(\frac{3}{(2/1.1284)}\right)^\wedge 2 = 0{,}95496492 \text{ (exato)}$$

$$\left(\frac{3}{(2/1.1284)}\right) \times \left(\frac{3}{(2/1.1284)}\right) = 2{,}86489476$$

(exato)

$$\left(\frac{3}{(2/1.1707)}\right)^\wedge 2 = 1{,}0278038675 \text{ (exato)}$$

$$\left(\frac{3}{(2/1.1707)}\right) \times \left(\frac{3}{2/1.1707}\right) = 3{,}0837116025$$

(exato)

Voltando:

$$\left(\frac{3}{1.1284}\right)^\wedge 1.1707 = 3{,}1415742\ldots \text{ Ⓐ}$$

$$\left(\frac{3}{1.1284}\right) \times 1.1707 = 3{,}11246012\ldots \text{ Ⓑ}$$

Subtraindo-se:

$$A - B = 0{,}02911412\ldots$$

Test:

$\overline{2{,}6586316\ldots}$

$\sqrt{3{,}1415}$

$1{,}323032411\ldots$

wrong!

Ajustando:

$$\sqrt[2.65]{3{,}1415} =$$

$$1{,}54027721\ldots$$

Wrong? (Não)

$$\left(\frac{3}{1.1284}\right) = 2{,}6586\ldots$$

$$\left(\frac{3}{1.1284}\right)\hat{}\ 1{,}1707 = 3.1415742$$

Corrigindo:

$$\sqrt[1{,}1707]{3{,}1415}$$

$$(2{,}65)^{\wedge}$$

Portanto:

$$\sqrt[1.1707]{3{,}1415}$$

$$(2{,}65)\hat{}\ 1.1707 = 3{,}1296\ldots$$

$$(2{,}658578019586)\hat{}\ 1.1707 =$$

$$3{,}14159999999\ldots$$

$$\cong 3{,}1415$$

Off: $\sqrt[1.1284]{3{,}1415}$ ∴ $(2{,}75783397608\,22)\hat{}\ 1.1284 =$

$3{,}1414\,\underline{999}\,\underline{999}\,\underline{999}\ldots$

20220602m

1,54 ^ 2,65 = 3,14000189...

Obs: 1,54 x 2,65 = 4,081 (exato).

1,5399996 ^ 2,65 = 3,1399997...

1,539999964 ^ 2,65 = 3,1399999501...

1,54 ^ 2,64 = 3,1264731...

Diferença:

(1,54 ^ 2,65) - (1,54 ^ 2,64) = 0,01352874...

2,65 ^ 1,1741 = 3,14002409...

⇘ 2,65 ^ 1,1746 = $3,1415\underline{545}$
(926...)

Outras simulações:

$\sqrt[2,65]{3,1414}$ = 1,54025871...

$\sqrt[2,64]{3,1414}$ = 1,5427809...

Comparando:

com anterior.

2,64 ^ 1,1741 = 3,12611...

2,64 ^ 1,1746 = 3,12763...

Arredondamento teste:

2,64 ^ 1,18 = 3,144073...

2,64 ^ 1,179124 = 3,14140050101...

Voltando: RAMANUJAN
1729

2,65 ^ 1.1707 = 3,1296...

Obs: The most famous taxicab number is $1729 = 1^3 + 12^3 = 9^3 + 10^3$.

Usando este número:

2,65 ^ 1729 = 6.19606602223996...

≃ 2π = 6,2831853...

2,65 ^ 1.1729 = 3,13635407...

Obs:

2π − (2,65 ^ 1729) = 0 08711928...

Outros dados:

$1729 = 1^3 + 12^3 = 9^3 + 10^3 \Rightarrow$

$1729 = 1+7+2+9 = 19$

$19 \times 91 = 1729$

menor

Obs.: $\underbrace{7 \times 13}_{91} \times 19 = 1729$

Vejamos:

$2,65^{1.1729} - 2,65^{1.1707} =$

$0,0067172374...$

Interessante:

$2,65^{1729} = 6.19606602223...$

$2,64^{1729} = 8.97820583688...$

$2,64^{1.1729} = 3,1224769838...$

Obs:

$2,64^{1729} - 2,65^{1729} = ?$

adequando a equação:

Usando 2,64 e 2,65 ⇒

$2{,}64 \wedge 1.1729 = 3{,}1224969838\ldots$

$2{,}65 \wedge 1.1729 = 3{,}13635407\ldots$

Vejamos:

$2{,}65 \wedge 1.1729 - 2{,}64 \wedge 1.1729 =$

$= 0{,}013877088\ldots$

Continuando:

$2{,}653773289 \wedge 1{,}1729 =$

$= 3{,}1415926535547\ldots$

(897...)

Others: ①

$\left(\frac{3}{1.5}\right) \wedge 1.5 = 2{,}8284\ldots \quad 2\sqrt{2}$

$\left(\frac{3}{1.5402}\right) \wedge 1.5 = 2{,}7184\ldots$

$\left(\frac{3}{1.5427}\right) \wedge 1.5 = 2{,}71181\ldots$

$\left(\frac{3}{1.5402}\right) \wedge 1.5402 = 2{,}7922601\ldots$

$\left(\frac{3}{1.5427}\right) \wedge 1.5427 = 2{,}789928807290\ldots$

others: ②

1,54 ^ 2,64 = 3,1264...

② ✡ 1,54 ^ 2,65 = 3,140001895...

1,5402 ^ 2,64 = 3,1275...

1,5402 ^ 2,65 = 3,14108... (+ valor)

1,5427 ^ 2,64 = 3,14096...

1,5427 ^ 2,65 = 3,1546... out.

Melhor: 2,65

Comparando com as anteriores:

2,65 ^ 1,1741 = 3,14002409... ✡

2,65 ^ 1,1746 = 3,1415545... ①

(926)

2,64 ^ 1,1741 = 3,12611658...

2,64 ^ 1,1746 = 3,12763433...

Subtração: 1 − 2 =

2,65 ^ 1,1741 − 1,54 ^ 2,65 =

= 0,00002219924544...

(2)

Testagem posicional:

2,65 ^ 1,1741 = 3,14002409...

1,1741 ^ 2,65 = 1,53009530...

1,54 ^ 2,65 = 3,140001895...

2,65 ^ 1,54 = 4,48537...

Diferença:
$$\begin{array}{r} 1,54 \\ 1,174 \\ \hline 0,366 \end{array}$$
(exato).
⇒ par.

Vejamos:

2,65 ^ 1,1741 − 1,1741 ^ 2,65 =
= 1,609928 (≅ milha ⇒ checar!)

2,65 ^ 1,54 − 1,54 ^ 2,65 =
= 1,3453719...

Obs: Milha: 1609,344 metros
ou
1,609344 Km

Vejamos o erro:

1,609928 — 100
1,609344   x

x = 99,9637...

obs: Este erro existe porque estamos adotando o $\pi \cong 3,14$

---

off:

1 pol = 25,4 mm ou 2,54 cm

1 metro = 39,37 pol (dízima)

1 milha = 1609,344 metros
$\times$ 39,37...

63360 pol.

checando:

OK

1609,344 metros = 160934,4 cm

160934,4 ÷ 2,54 cm ≅ 63360.

Voltando:

$\left(\frac{2}{1.1284}\right)$ $\left(\frac{2}{1.1707}\right)$ | $\left(\frac{2}{1.1741}\right)$ $\left(\frac{2}{1.1746}\right)$

$\left(\frac{3}{1.1284}\right)$ $\left(\frac{3}{1.1707}\right)$ | $\left(\frac{3}{1.1741}\right)$ $\left(\frac{3}{1.1746}\right)$

→

$\left(\frac{2}{1284}\right) = 0{,}01557\ldots$ $\left(\frac{3}{1284}\right) = 0{,}002336\ldots$

$\left(\frac{2}{1.1284}\right) = 1{,}772\ldots$ $\left(\frac{3}{1.1284}\right) = 2{,}65863\ldots$

$\left(\frac{2}{1707}\right) = 0{,}0117\ldots$ $\left(\frac{2}{1729}\right) = 0{,}00115\ldots$ $\left(\frac{3}{1729}\right) = 0{,}001735\ldots$ $\left(\frac{3}{1.1729}\right) = 2{,}5577\ldots$

$\left(\frac{2}{1.1707}\right) = 1{,}7083$ $\left(\frac{3}{1.1707}\right) = 2{,}5625\ldots$

$\left(\frac{2}{1.1741}\right) = 1{,}7034\ldots$ $\left(\frac{3}{1.1741}\right) = 2{,}5551486\ldots$

$\left(\frac{2}{1.1746}\right) = 1{,}70270\ldots$ $\left(\frac{3}{1.1746}\right) = 2{,}5540609\ldots$

$\frac{3{,}1415}{(2/1.1284)} = 1{,}7724343$ (exato) $\frac{3{,}1415}{(3/1.1284)} = 1{,}1816228\overline{6}$

$\frac{3{,}1415}{(2/1.1707)} = 1{,}838877025$ (exato) $\frac{3{,}1415}{(3/1.1707)} = 1{,}22591801\overline{6}$

$\frac{3{,}1415}{(2/1.1741)} = 1{,}844217575$ (exato) $\frac{3{,}1415}{(3/1.1741)} = 1{,}22947838\overline{3}$

$\frac{3{,}1415}{(2/1.1746)} = 1{,}84500295$ (exato) $\frac{3{,}1415}{(3/1.1746)} = 1{,}2300019\overline{6}$

$\frac{3{,}1415}{(2/1.1729)} = 1{,}842332675$ (exato) $\frac{3{,}1415}{(3/1.1729)} = 1{,}22822178\overline{3}$

Aprofundando: (deepening)

$\rightarrow 2,65 \wedge 1.1741 = 3,14002409\ldots$

$1.1741 \wedge 2.65 = 1,53009530\ldots$

$\rightarrow \underline{1,54} \wedge 2.65 = 3,140001895\ldots$

$2,65 \wedge 1.54 = 4,48537\ldots$

Juntando-se: (joined)

$\sqrt[2,65]{3,1415} = \underline{1,54}027721\ldots$ $\Uparrow$

obs: 1,5402772187495553538657192795144...

Interessante:

$1,54027 \wedge 2,65 = 3,1414609\ldots$

$1,540277\underset{\text{wrong}}{7} \wedge 2,65 = 3,141502601\ldots$ (passou este valor)

Veja:

$1,5402 \wedge 2,65 = 3,141082\ldots$

Importante:

$\sqrt[2,6586316]{3,1415}$

$= \left(\dfrac{3}{1.1284}\right)$

Calculando:

data
S T Q Q S S D

$$\sqrt[\frac{3}{1.1284}]{3,1415} \rightarrow 2,65\ldots$$

$$= \underbrace{2,6586316908897}5540\ldots$$

$$\sqrt[\frac{3}{1.1284}]{3,1414}$$

$$= \underbrace{2,6586316908997}5540\ldots$$

Valores iguais

checando valores

3,1415 → 2,65863169088975540588
4438142 5027
3,1414 → 2,65863169088975540588
4438142 5027
(iguais)

→ a diferença de valores deverá estar além da casa 50 após a vírgula.

Com o valor $\left(\frac{3}{1.1707}\right)$:

$$\sqrt[\frac{3}{1.1707}]{3,1415}$$
$$= 2,5625694029 21\ldots$$

$$\sqrt[\frac{3}{1.1707}]{3,1414}$$
$$= 2,5625694029 21\ldots$$

Valores iguais

Calculando outros valores:

$$= \sqrt[\boxed{\frac{2}{1.1284}}]{3,1415}$$

$= 1,77242112725\ldots$

$$= \sqrt[\boxed{\frac{2}{1.1284}}]{3,1414}$$

$= 1,77242112725\ldots$

$$= \sqrt[\boxed{\frac{2}{1.1707}}]{3,1415}$$

$= 1,70837960194\ldots$

$$= \sqrt[\boxed{\frac{2}{1,1707}}]{3,1414}$$

$= 1,70837960194\ldots$

Ajustando: (Important)

$$\left(\frac{3}{1.1284}\right) = 2,658631\ldots$$

Voltando:

$$\sqrt[2,65]{3,1415} =$$

$1,540277218\ldots$

Portanto: Valores:

Ⓐ 1,54 ^ 2,65
= 3,14000189...

{ 1,540277218...
{ 2,658631690...

Ⓑ 1,540277 ^ 2,65
= 3,14146098...

Ⓒ 1,540277 ^ 2,65 se você estender o valor passa 3,1415...

Mudando a ordem

1,54 ^ 2,65 → também não dá certo.

Ⓓ 2,65 ^ 1,54 → não dá certo.

Valor certo: (Usando-se estes valores)

1,540277218749555353865719279514 4
= 3,1415
↓
Este número decorre do número $\left(\frac{3}{1.1284}\right)$ que foi utilizado na raiz do número $\sqrt{3.14150}$ (exato), obtendo-se
2,658631690889...

Interessante:

Ⓐ $\sqrt[2,65]{3,1415} = 1,54027721\ldots$

Ⓑ $\sqrt[1.1707]{3,1415} = 2,6585780195\ldots$

Obs.: $\underset{\underset{\textcircled{A}}{\uparrow}}{1,54} \wedge \underset{\underset{\textcircled{B}}{\uparrow}}{2,65} = 3,14000189532\ldots$

Ajustar estas equações quando da obtenção da simulação gráfica espelhada sobre um plano.

Ex: $\sqrt[2,65]{3,1415926}$

$= 1,5402\underbrace{9435133}_{7721\ldots}\ldots$

OK (cálculo correto)

continuando:

$\sqrt[1.1707]{3,1415926}$

$= 2,658\underline{6}4495821048\ldots$ OK (correto)

off: $\sqrt[1.1284]{3,1415926}$

$= 2,757906016833\ldots$

20220602$_4$ Books:

1) Data-NovoVivo-012-2022/
Books-012-2022/
Important/
Carl-Louis-Ferdinand von Lindemann
(1852-1939).pdf

→ π est transcendant.

2) Books-012-2022/
20220602/
Regard sur un mathématicien
l'indien Srinivasa Ramanujan
(1887-1920).pdf

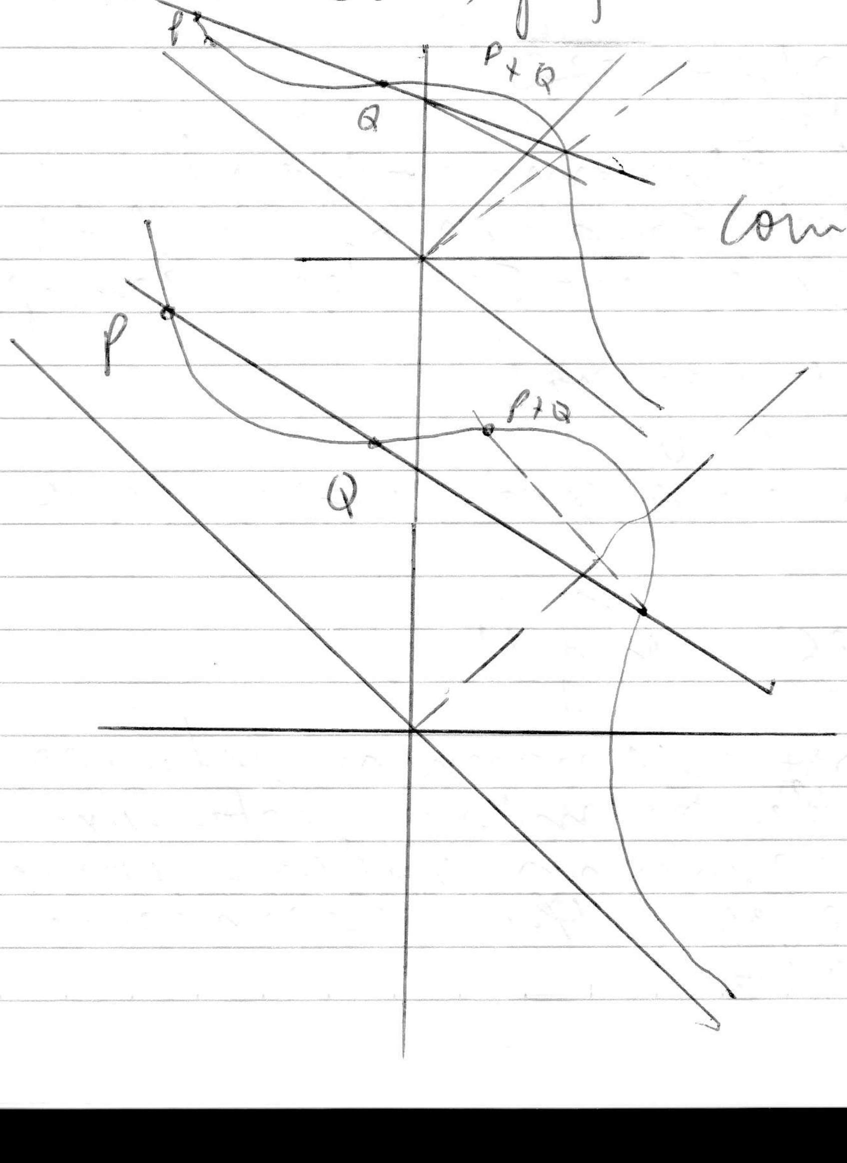

Corrigindo
↓

Pág.06(434)

(Ver
este
gráfico

Obs: Le nombre π. (Pág. 07)

$$3\,\frac{10}{71} < \pi < 3\,\frac{1}{7}$$

Pág. 08 (436) —

$$\frac{\pi}{4} = 4\arctan\left(\frac{1}{5}\right) - \arctan\left(\frac{1}{239}\right)$$

Pág. 09 (437) —

$$\frac{1}{\pi} = \frac{\sqrt{8}}{9801}\sum_{n=0}^{\infty}\frac{(4n)!}{(n!)^4}\,\frac{(1103+26390n)}{396^{4n}}$$

3) Books-012-2022/
Important/
Bharath Shiraman - Handbook of the Mathematics of the Arts and Sciences - Springer (2021).pdf

Obs: 2794 páginas.

Ver todas as páginas, inicialmente 01-104

Pág. 13 (36) - Fig. 4

A most Riemannian situation: a plot of the Riemann zeta function on the Riemann sphere immediately suggests the Riemann hypothesis.

Pág. 14(37)

Fig. 6 → Stereographic projection is a conformal map between plane and sphere.

Pág. 15(38)

Fig. 7 → Conformal maps between the upper half plane and the Poincaré disc (left) and pseudo-sphere (right).

Pág. 16(39) -

Fig. 8 → Two different methods of assigning unique colours to the preimages of geodesics in hyperbolic space.

Pág. 20(43)

Dynamical Geometry and Asymptotic Destination plotting

... Borwein and Sims (2011) used the dynamical geometry package CINDERELLA to visualize the long term behavior of the dynamical system for various start points:

Fig. 13 → outside of the
axis of symmetry.

Pag. 28(51) - Fig. 21
Analytic continuation and
Schwarzian reflection.

Pag. 30(53)
Fig. 22 - Lyapunov functions.

Pag. 48(70) -
Fig. 2 Representation of a
degree 5-Blaschke
product in the plane
and on the
Riemann sphere.

Pag. 57(79) — Lobachevsky (1829)
Fig. 15 Poincaré disk.

Isomorphisms between the
hemisphere model, the
Klein model, and the
Poincaré models.

Pag. 58(80) - Fig. 16 - Lines in the Poincaré
disk model and the half-plane model.

Pág. 59(81) -

Therefore, the hyperbolic size of an angle between lines in the Klein model is not equal to its Euclidean size.

Pág. 60(82) - Fig. 17 Conformal self-map of the disk.

Pág. 62(84) - Fig. 18 Walsh's problem (Blaschke product)

Pág. 63(85) - Fig. 20 Riemann sphere

Fig. 21 - A Blaschke product of degree 5 with some inscribed polygons.

✓

Obs: Usar este gráfico para simular a curvatura de um plano. (imagem direita)

Pág. 69(91) - Fig. 27

Circles in the Poincaré disk model (imagem direita)

Obs: Simular com a imagem em curvatura.

Pag. 75 (97) Fig. 35

Symmetric paths connecting tiles (azulejos)

Obs: Analisar

Adaptar para:

how to form a square with the same area of a circle, by cutting them into interchangeable pieces that can be visualized.

Source: www.quantamagazine.org/an-ancient-geometry-problem-falls-to-new-mathematical-techniques-20220208/

Pag. 82 (104)

Fig. 1 - A perspective interpretation of Desargue's theorem.

... it relied almost exclusively on Euclidean geometry.

3) Books-012-2022/
2022 06 02/
Ramanujan - 1729 / Hardy /
London Mathematical Society - ...
/Hardy ... (1972).pdf

Obs.: Livro todo!!!

Pag. 159 (169) -

Some problems of Diophantine approximation:

The lattice-points of a right-angle triangle.

(Second memoir). Hardy.

Hecke is mainly transcendental

Ostrowski's is entirely elementary.

Obs: Lattice-points

A point at the intersection of two or more grid lines in a point lattice.

↳ A point lattice is a regular spaced array of points.

A lattice is a discrete subgroup

of a Euclidean space.

Ex: $\{a_1 + a_2\sqrt{2}\} \subset \mathbb{R}$.

requires two generators but is not discrete, and is not a lattice!

$$\frac{N'(r)}{N(r)} = \frac{\frac{24}{\pi^2} r^2 + O(r \ln r)}{4 r^2 + O(r)}$$

$$\frac{\frac{6}{\pi^2} + O\left(\frac{\ln r}{r}\right)}{1 + O\left(\frac{1}{r}\right)}$$

$$= \frac{6}{\pi^2}$$

$$\Rightarrow c = (2\pi^2/3)^{1/3} \simeq 1.87$$

Any parallelogram on the lattice in which two opposite side have length 1 has unit area (Hilbert and Cohn-Vossen 1999, pp 33-34).

Book: Source

https://mathworld.wolfram.com/PointLattice.html

Books: Regular Systems of
Points - Hilbert (Chapter 2 -
Geometry and the imagination)

Nome do livro:

Geometry and the Imagination

Source:

Ⓐ Books-012-2022/
Hilbert/
David Hilbert _ Geometry
and the imagination
_ (2020). pdf

Pág. 368

Chapter II - Regular Systems of
Points.

Ⓑ Books-012-2022/
Important/
Hilbert/
Hilbert.pdf

105 pág.

The Foundations of Geometry
by
David Hilbert
1950

4) Books-012-2022/
20220602/
Ramanujan-1729/
RamanujanTangente.pdf
06 pages.
→ Ramanujan à Hardy.

5) 20220602/
Ramanujan-1729/
ramanujan-2.pdf
⇒ 07 pages
Autour de π.

obs :

Constante de Landau-Ramanujan

$$a_K = \int_1^2 \frac{(\ln(x))^K}{x-1}\,dx \quad \text{on a : (1909)}$$

$$a_1 = \frac{1}{2}\zeta(2) = \frac{\pi^2}{12} \quad \text{et}$$

$$a_2 = \frac{1}{4}\zeta(3)$$

data

20220604,ii **obs:**

1°

1,75 seg/arc # 0.000277778° (degree)

↳ 0.04861111°

obs: Valor aproximado:
divide the plane angle value by 20571.

Portanto:

1° arcsec ≅ $0.01\overline{6}$ arcmin

1° arcsec = 1/1296000 do círculo.

Obs: $60'' \times 60' = 3600 \times 360 = 1296000$
grau

---

obs: 1° para grados
$1° = 1.\overline{1}$

1 MOA (arco-minutos) para grau
0.0167

1° grau para arco-minutos:
59.9999999988

1° grau para arco-segundos:

jandaia

3599,999 999 9 712 arcsec

---

1 rad para grados:

63.66197723667 grad

---

1 quadrant = 100 grad
↓
90°

---

obs: 1° para rad:

0.0175 rad

---

1 arcsec = 0.0167 MOA

1 arcsec = 0.0003°

Grado: é a unidade de medição do ângulo.

= 1/400 do círculo, ou seja, 0.9°.

Mil angular → ângulo

1/1000 de um ângulo reto,

0.1 graduado, 0.09° ou
5.4 minutos de arco.

Atualmente: OTAN exércitos usam
1/1600 ângulo reto ou
0.05625°

Portanto:

$$1 \text{ arcsec} = \frac{\text{ângulo}}{3600}$$

$$0{,}0002\bar{7}\,°$$

Obs: Eddington experiment
(calculus).

$1{,}98 \pm 0.12$ arcsec

⇒ converting to the now common standard deviation gives

$1{,}98 \pm 0.\underline{18}$ arcsec

Source:

https://royalsocietypublishing.org/doi/10.1098/rsnr.2020.0040

Published 21 October 2021
https://doi.org/10.1098/rsnr.2020.0040

Outra referência:

Source:
https://www.cmu.edu/dietrich/philosophy/docs/glymour/relativityeclipses.pdf
(1980)

Look: standard deviations.

⇒ Veja: (minhas anotações):
Caderno-001---20211112
item 94) Light Deflection
Cavendish (08/09/2019)
Soldner
Pág. 129

Obs: Pág. 133 — 0,84 sec arc.

Geometria - Baldor (Pág. 203) (199).
↓ vai para o dia 07/6/21

Obs: Pág. 156 - Vejá Fig. 221

Comparar com:

Enciclopédia Prática Jackson
Volume IX.

Pág. 259 - item 81

Retificação da circunferência
e quadratura do círculo.

Vejá:

Pág. 263 - item 86

$$\pi = \frac{355}{113}$$

Compare com a página 155
das minhas anotações.

Importantíssimo:

Pág. 187 → Gauss curvatura espaço

179° 59' 59.32"

Obs. Fazer os cálculos para ver
quantos arcsec deram.

Vejá se: 1.75 arcsec or 1.98 ± 0.12
or outro valor.

$179^{a}\ 59'\ 59.32''$ (1.75, 1.98 ± 0.12)

Obj: Estudo

18 arcsec = 0.005°

1 arcsec = $\frac{1}{1.296.000}$ círculo (360°)

$\swarrow$

$\Rightarrow \underbrace{60' \times 60''} \times 360° = 1296000$

Important:

1° = $\underline{3600}$ arcsec

$\therefore$ 1 arc = $0{,}0002\overline{7}$ °

Voltando:

1 arcsec = $\frac{1^{\circ}}{3600}$

Obs: Veja também: milliarcsecond (MAS)
microarcsecond (μAS)

Vejamos:

$1'' = \frac{1}{1296000} \div 100 = \frac{1}{129.6\underbrace{00000}_{(5)}}$

Portanto: $0{,}\underbrace{00000}_{(5)}2\overline{7}$

$59.32'' \rightarrow 0{,}68''$ (20220605)

~~59/60~~ $\frac{5932}{3600}$ (conversão) $\frac{0{,}68}{3600}$ (graus decimais)

$1{,}64\bar{7}$

$\boxed{= 0.0001\bar{8}}$ ↳ graus decimais

Corrigindo:

$179^\circ 59' 59.32'' \rightarrow 59{,}32/3600 = 0{,}01647^\circ$

$59 / 60 = 0{,}98\bar{3}^\circ$

$179^\circ + 0{,}983^\circ + 0{,}01647^\circ =$ ~~180,1477~~° (?)

achei estranho!

Correto é:

$179{,}99947^\circ$

Mais preciso:

$179 + (59/60) + (59{,}32/3600) =$

$179{,}9998\bar{1}$

Diferença:

$$\boxed{\begin{array}{r} 180 \\ -\ 179.9998\bar{1} \\ \hline 0.0001\bar{8} \end{array}}^\circ \rightarrow \text{correto}$$

~~x 60 = 0,011$\bar{3}$~~

~~x 3600 = 0,68~~

$\times 60 = 0{,}011\bar{3}$

$\times 3600 = 0{,}68''$ (OK) → correto

Einstein → Deflection of light by the Sun's Gravitational Field.

Source: Data-NovoVivo-012-2022/Books-012-2022/20220604/Eclipse-1919-Sobral-Einstein/rsnr.2020.0040.pdf

Pág. 07 – The Sobral-4-inch plates gave the most precise and accurate results whith the derived result for light bending, corrected to the limb of the Sun, of 1,98 ± 0,12 arcsec. The error quoted is the probable error. Converting to the now common standard deviation gives 1.98 ± 0.18 arcsec.

Pag. 10 - Nota de rodapé 20 (veja dia 20220605) item 01.

Harald von Klüber,

"The determination of Einstein's light-deflection in the gravitational field of the Sun"

Vistas Astronomia 3, 47-77 (1960)

Obs: Procurar uma cópia.

IMPORTANT:

Comparar o valor de GAUSS com EINSTEIN, não esquecer do NEWTON.

20220604$_{III}$

Na música → quinta justa: razão 3:2

A matemática foi descoberta ou inventada?

Para mim, a matemática foi descoberta, mas os matemáticos inventaram a forma como ela é descrita.

Wigner (Eugene Wigner) físico abordou os números complexos em um influente ensaio de 1960 chamado:
"The unreasonable effectiveness of mathematics in the physical sciences" (A eficiência irracional da matemática nas ciências naturais).

→ para Wigner → texto originalmente publicado em 1930
Outra fonte: 1960 (ISBN 978-88-459-7946-0)
Italian language

Wigner chega à conclusão de que não se sabe porque a matemática é tão eficiente.
Muitos matemáticos pensam que eles inventaram a matemática.
(eu não acho !!!).

Einstein disse:

"Quanto mais se referem à realidade, mais incertas as leis matemáticas se tornam, e quanto mais precisas elas são, menos se referem à realidade"!

→ referindo-se como é difícil fazer com que as leis da matemática coincidam com as leis da realidade física.

—

"Deusa" egípcia da matemática → Seshat.

—

teoria das categorias

"matemática da matemática"

→ tenta construir pontes entre diferentes áreas da matemática.

—

A realidade física é matemática.

—

Teoria das Cordas

→ é finita.

A teoria é chamada de "cordas" porque o movimento vibratório de uma dessas partículas (subatômicas) é o mesmo de uma corda num instrumento musical como uma harpa ou um violão.

Se colocarmos alguma energia sobre ela, ela vibra.

Obs: Pitágoras → estudo das cordas musicais.

O grande matemático grego pensou que deveria haver um princípio unificador, um paradigma para resumir a vasta criação que vemos ao nosso redor no Universo.

—

Tesla disse: "Tudo é energia".

—

matéria escura com base nesta teoria, seria composta de cordas que vibram em oitavas mais altas.

"O Universo é mais simples do que pensávamos"

—

A matemática sempre
existiu,
os matemáticos apenas
descobriram a
matemática
e inventaram a forma
como a matemática
seria descrita e
apresentada para toda
a humanidade.

Stenio Musich
202206041452

Teoria da Relatividade.

Para mim, a teoria da relatividade fundamentou-se em conceitos de outros pesquisadores como: Poincaré
Maxwell
Lorentz

Devemos observar que a hipótese de Albert Einstein ~~não~~ apresenta uma referência exaustiva de sua fonte bibliográfica.

Einstein disse: "O problema concebido pela minha mente foi respondido pelo luminoso céu do Brasil" 1925.

'Einstein não concebeu este problema, ele apenas descobriu o que acontece no Universo há bilhões de anos.

Stenio Musich
202206041459

Explicação:

As estrelas sempre desviaram a luz e Einstein apenas descobriu o que já estava acontecendo no Universo.

Eu não tiro o seu mérito, mas falar que a sua mente concebeu este problema (Light Deflection) não é uma afirmação perfeitamente correta.

Obs: Edgar Alan Poe
(1809-1849)

→ ensaio "Eureka"
⇒ expressão célebre atribuída a Arquimedes.

Publicação: 1848

Poema em prosa.

→ tipo de um ensaio cosmogônico.

Foi dedicado a
Alexander von Humboldt
(1769-1859)

Poe elogia a sua obra por suas tentativas de descobrir o Universo.

Naquela época os físicos diziam que o Universo era estático, infinito, mecânico e eterno
enquanto que Poe, apresenta o Universo como sendo

{a) dinâmico
b) constante evolução

Na sua introdução, Poe disse

Desejo falar do físico, metafísico e matemático – do universo material e espiritual – de sua essência, sua origem, sua criação, sua condição atual e seu destino.

Em resumo:

a) Big Bang → explosão

Obs: Para mim este aspecto não é tão revolucionário, mas inovador pelo momento em que Poe vivia.

Faça-se a luz → Gênesis (Bíblia)

↳ refere-se ao nosso Universo.

O nosso Universo se expande pois há um vazio que sempre existiu e que permite que o Universo se expanda.

Coincidentemente a teoria do Big Bang começou a tomar forma pelos físico e clérigo belga Georges Lemaître e desenvolvida pelo físico ucraniano George Gamow em 1948.

b) Big Crunch e o modelo do Universo oscilante.

Parte do pressuposto, que a força

da gravidade

fará com que todo o Universo se contraia e colapse sobre si mesmo em uma nova partícula primordial.

⇒ ciclo constante de expansão e contração

Universo Oscilante.

c) Paradoxo de Olbers

Problema levantado pelo astrônomo alemão Heinrich Olbers em 1823.

Por que o céu é escuro? se o Universo é infinito e as estrelas também.

Físico Paul Halpern (USA, 1961) Professor de física da Universidade da Filadélfia, escreveu vários artigos publicados na revista AEAON e na plataforma Medium.

Solução de Poe ⇒

Poe disse:

Se a sucessão de estrelas fosse infinita, então o fundo do céu nos apresentaria uma luminosidade uniforme, como aquela mostrada pela Galáxia — já que não poderia haver absolutamente nenhum ponto, em todo esse pano de fundo, no qual não existiria uma estrela.

Poe foi o primeiro a sugerir que é impossível para nós conseguir ver todo o Universo, que nem sempre esteve ali, e que a luz das estrelas pode levar vários anos (bilhões de anos) para chegar até nós.

Importante: Ⓐ

Para mim, o céu é escuro pois a escuridão decorre do espaço vazio existente que abarca todo o Universo.

Sem este espaço vazio, o Universo não poderia se expandir e evoluir.

Livro: "A Escuridão da Noite: Um Enigma do Universo"

Astrônomo Edward Harrison
(1987)

d) Teoria da relatividade,
multiverso e
buracos negros.

Poe aborda a noção de que
"espaço e duração são um só!"

obs: Alguns autores veem
como uma profetização
do
"espaço-tempo" relativista,
e reflete sobre a intercambialidade
da matéria e da energia.

Poe sugere que nosso Universo
poderia ser um dos muitos que
existem em um plano
infinito.
(≅ multiversos)

Important:

Em um conto anterior,
"Uma descida ao Maelström",
Poe descreve um redemoinho,
um vórtice mortal para

onde objetos são arrastados sem
deixar vestígios.

Interessante:

Tanto Einstein quanto Freud
disseram que
Edgar Allan Poe
era um "caso patológico"!

Obs: Marie Bonaparte,
sobrinha-neta de Napoleão
e discípula de Sigmund Freud,
escreveu uma biografia na qual
analisa psicologicamente o
autor americano com base em
Eureka e no restante de suas
obras.

No prefácio, escrito por Freud,
ele trata Poe como um caso
patológico, algo que Albert
Einstein também disse na
sua época.

Obs: Uma afirmação bem
conveniente para
alguém que utilizou
as idéias de um "louco"
(grifo nosso).

Defensor de Poe: astrônomo
Alberto Coppi (Itália, 1962) de

Observatório Astronômico de Bolonha.

Obs: Vejá carta para a sua sogra de 07 Julho de 1849

⇒ Não tenho vontade de viver desde que fiz Eureka. Não consegui fazer mais nada.

Obs: Esta carta foi escrita 3 meses antes de falecer.

Voltando: Teoria da Relatividade

Einstein ousou questionar a Lei da Gravitação Universal de Isaac Newton.

Obs: Para mim, a gravidade como uma força física fundamental não existe, esta força decorre das demais leis fundamentais da física.

{ Eletromagnetismo
Força Forte
Força Fraca

"As ideias de Newton sempre terão a importância de serem a base sobre a qual nossa concepção moderna da física foi construída"

Einstein, 1919.

Livro: "No Shadow of a Doubt)

→ eclipse 1919

Author: Daniel Kennefick.

→ Cálculos de Einstein previam um desvio da luz duas vezes maior do que o que era previsto de acordo com a teoria de Newton

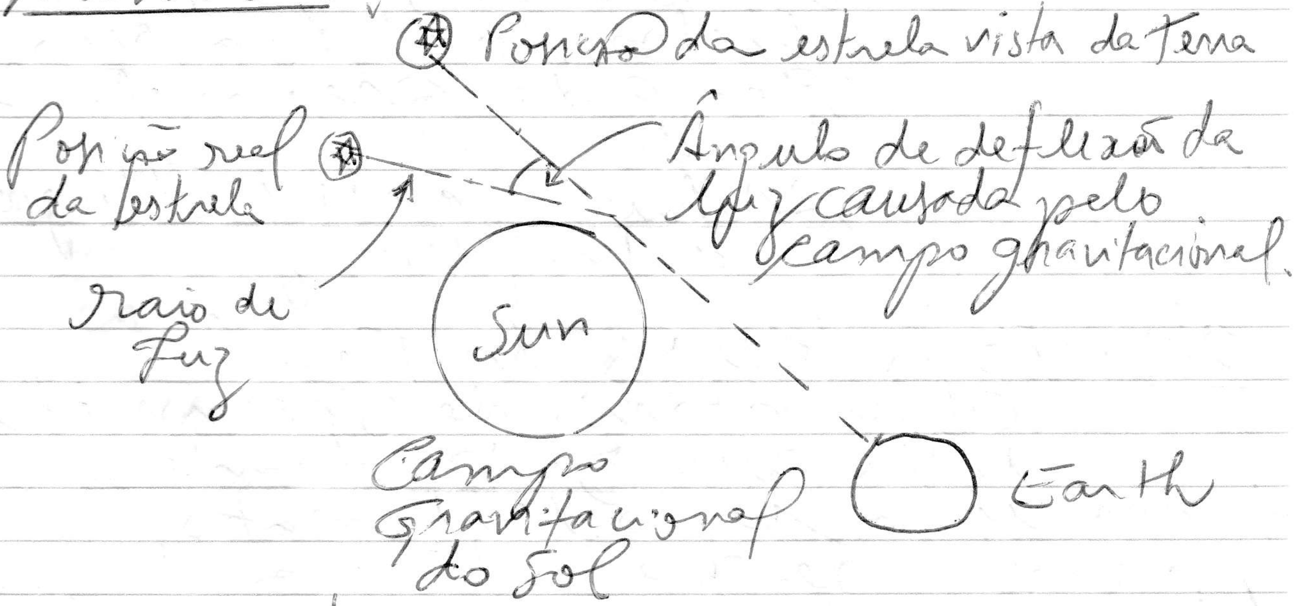
Posição da estrela vista da Terra
Posição real da estrela
Ângulo de deflexão da luz causada pelo campo gravitacional.
raio de luz
Sun
Campo gravitacional do Sol
Earth

Obs: Einstein chegou a investir suas próprias economias na expedição

do astrônomo alemão
Erwin Finlay-Freundlich
para observar um eclipse na
Criméia, na Rússia, em
1914, um ano antes de
publicar a teoria
da relatividade geral.

Freundlich chegou à Rússia,
quando explodiu a Primeira
Guerra. Os seus instrumentos
foram confiscados e ele não
conseguiu realizar o
experimento.

1917 Frank Watson Dyson
astrônomo inglês,
diretor do Observatório
Real de Greenwhich
e
Arthur Stanley Eddington
astrofísico

Obs: Tom Kerss do Real
Observatório de Greenwich,
disse que Dyson também
tinha um certo sentimento
patriótico de que a teoria de
Isaac Newton (que era inglês)
deveria ser tratada com mais
respeito do que a de um jovem

da Alegrante.

Eddington era um entusiasta dos teorias de Einstein e um internacionalista, que acreditava no ideal de juntar as melhores mentes de todas as nacionalidades em busca da verdade científica.

Eles usaram um aglomerado de estrelas especialmente brilhantes, as Híades
→ o Sol estaria perto. (1919)

Duração do eclipse:

5 min 13 → Sobral
5 min 02 → Ilha de Príncipe.

Interessante:

A faixa de totalidade do eclipse, ou seja, o trecho em que o Sol estará completamente encoberto — cruzaria toda a América do Sul, começando na Bolívia passaria pelo Oceano Atlântico e terminaria no continente africano, na Tanzânia.

Obs: Observe que o formato da Lua Quarto Crescente no hemisfério Norte parece um C e no hemisfério

Sol parece um D.
Ela é vista ao contrário, porque a Terra gira de
Oeste para Leste,
portanto, no hemisfério Sul gira no sentido horário e no hemisfério Norte gira no sentido anti-horário.

Decisão de ir a Sobral;

Dyson recebeu uma carta do engenheiro
Henri Charles Morize,
diretor do Observatório Nacional do Rio de Janeiro e um dos fundadores da Academia Brasileira de Ciência (ABC).

Ele informou que Sobral, a segunda maior cidade do Ceará, bem conectada por trens e por um porto relativamente próximo — seria o melhor lugar para acompanhar o fenômeno.

Dyson teve que ficar na Inglaterra, mas enviou
— Charles Davidson, calculista sem formação acadêmica,

mas com muita experiência em telescópios, e

o astrônomo irlandês

Andrew Crommelin,

que operaria um segundo telescópio levado por segurança.

telescópio irlandês é que fez as imagens dos resultados que fizeram história

Tinha o campo visual maior que permitia ver mais estrelas!
foi construído para o eclipse de 1900, era menor e mais velho.

Em novembro de 1918, o Armistício de Compiègne anunciou o fim da guerra e abriu caminho para a expedição.

Eles vieram para o Brasil, Pará, a bordo do Anselm (navio)

Belém Pará para o Ceará no navio CAMOCIM e de trem para Sobral.

Fotos: 29 MAIO 1919

Os astrônomos tiveram cerca de

jandaia

4 minutos para fazer 27 fotos do céu, mostrando as 12 estrelas que queriam observar.

Devido ao calor que dilatou o espelho do seu principal telescópio, as imagens ficaram distorcidas, mas 8 imagens nítidas foram feitas do sol escurecido.

1 mês depois, eles fotografaram as mesmas estrelas, exatamente no mesmo lugar do céu, só que à noite.

Importante:

(I) https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.1920.0009

(II) https://royalsocietypublishing.org/doi/10.1098/rsnr.2020.0040

Obs: Volume 76, Issue I

https://royalsocietypublishing.org/toc/rsnr/2022/76/1

Published: 02 February 2022

Novembro de 1919, o estudo final sobre o eclipse, assinado por Dyson, Eddington e Davidson, foi publicado.

—

NY Times - 15 Novembro 1919
Einstein Theory Triumphs

—

A Filósofa alemã Ilse Rosenthal-Schneider, ela estava com Einstein quando ele recebeu o telegrama.

→ Ela perguntou o que ele faria se o resultado final fosse desfavorável a suas idéias, e ele, calmamente, respondeu:

"Eu teria pena de Deus, porque a teoria está correta".

—

Cecilia Payne-Gaposchkin descobriu que as estrelas eram formadas basicamente de hidrogênio e hélio.
→ Nasceu na Inglaterra em 1900.

Ondas gravitacionais não sofrem o efeito de dispersão (mudanças de meios físicos - de um para outro).

"Nossas medidas são realmente muito sensíveis a pequenas diferenças nas velocidades de frequência, mas não registramos nenhuma dispersão, mais uma vez, não provamos que Einstein estava errado"

Bangalore Sathyaprakash
Ligo e Universidade de Cardiff - UK

Obs: 04 de Janeiro 2017

Em uma comovente coincidência, o último 4 de janeiro também foi o dia em que HEINZ BILLING, pioneiro da ciência da onda gravitacional, morreu, 102 anos.

HEINZ BILLING, alemão especialista em física e informática construiu um dos primeiros interferômetros a laser — instrumentos hoje usados para detectar ondas gravitacionais.

⇒ Seu primeiro trabalho foi crucial para o desenvolvimento dos sistemas do LIGO.
(Observatório de Ondas Gravitacionais por Interferômetro de Laser - LIGO) - USA

20220604 (20220604)

① https://www.bbc.com/portuguese/geral-50377079

→ Netuno foi descoberto a partir de cálculos e não da simples observação.

Netuno interferia na órbita de Urano, de acordo com a teoria da gravidade.

Obs: Astrônomos perceberam a existência de outro planeta graças a um desvio na órbita esperada de Urano.

② https://en.wikipedia.org/wiki/Eddington_experiment

Obs: Livro (artigo) → Clark Glymour (1980)

"Relativity and eclipses: the British eclipse expeditions of 1919 and their predecessors

→ Historical Studies in the Physical Sciences 11(1): 49-85

doi: 10.2307/27757471

other :

"A Determination of the Deflection of Light by the Sun's Gravitational Field from observations made at the Total Eclipse of May 29, 1919

Pág. 571-581

doi 10.1098/rsta.1920.0009

Source : https://zenodo.org/record/1432106

My copy: 20220506/Eclipse-1919-Sobral-Einstein/article-Dyson-Einstein.pdf

01 Jan 1920 - 45 pages.

Voltando :

Relativity and eclipses: the British eclipse expeditions of 1919 and their predecessors.

Source :

https://www.cmu.edu/dietrich/philosophy/docs/

glymour/relativityeclipses.pdf

Obs: 19 folhas Pag. 49-85

③ A Determination of the Deflection of Light by the Sun's Gravitational Field

44 pages

Source: Data-NovoVivo-012-2022/
Books-012-2022/
20220604/
Eclipse-1919-Sobral-Einstein/
rsta.1920.0009.pdf

Cont

rsnr.2020.0040.pdf

25 pages

21 October 2020

doi: 10.1098/rsnr.2020.0040

④ 20220604/
Eclipse-1919-Sobral-Einstein/

livro-eclipse-sobral-digital.pdf

99 pages.

⑤ Interessante (image):

Registro de ondas previstas por Einstein abre caminho para nova era da astronomia

Source:

https://bbc.com/portuguese/geral-4013817

→ Ondulações no espaço-tempo

Source: 20220604/
Eclipse-1919-Sobral-Einstein/

↓

ondulações-no-espaço-tempo-20220604.png

Obs: O interessante é a demonstração da curvatura do espaço naquela determinada área do Universo.

⑥ Books-012-2022/20220604/
Scale Transitions as Foundations of Physics/
Scale Transitions as Foundations of Physics.pdf

Pág. 24 - Lorentz Matter: Poincaré Interpretation of Electromagnetic Ether
Pág. 45 - Einstein's Matter
Pág. 53 - Non-Euclidean Geometry as a Gauge Geometry.

data

S T Q Q S S D

20220605m

① The Determination of Einstein's light-deflection in the gravitational field of the Sun.

H. von Klüber

Source:
www.sciencedirect.com/science/article/abs/pii/0083665660900052

Ref: https://doi.org/10.1016/0083-6656(60)90005-2

Source: Vistas in Astronomy - Volume 3 — 1960, Pages 47-74.

② The Measurements of Light's Gravity Deflection of General Relativity were invalid.

Mei Xiaochun

Source:
https://www.academia.edu/50699178/The_Measurements_of_Lights_Gravity_Deflection_Relativity_were_Invalid

③ An Expedition to Heal the Wounds of War: the 1919 Eclipse Expedition and Eddington as Quaker Adventurer

Source:
www.academia.edu/882179/An_Expedition_to_Heal_the_Wounds_of_War_the_1919_Eclipse_Expedition_and_Eddington_as_Quaker_Adventurer

④ The Legacy of Einstein's
Eclipse Gravitational Lensing

__Source:__
www.researchgate.net/publication/
338290670_the_Legacy_of_
Einstein's_Eclipse_Gravitational_
Lensing.

__Obs:__ DOI 10.3390/universe6010009

⑤ Extended theories of Gravity
https://doi.org/10.1016/
j.physrep.2011.09.003

__Source:__
https://www.sciencedirect.com/
science/article/abs/pii/
S0370157311002432

__Other__
https://arxiv.org/abs/1108.6266

⑥ MISNER Gravitation
http://fma.if.usp.br/~mlima/
teaching/PGF5292_2021/
Misner_Gravitation.pdf

⑦ Shadow of the Moon and general Relativity: Einstein, Dyson, Eddington and the 1919 light deflection

ref.: https://doi.org/10.1590/1806-9126-RBEF-2019-0260

⑧ The Equivalence principle and the deflection of light

www.einstein-online.info/en/spotlight/equivalence_light/

⑨ A Comparison of Light Deflection in Newton's and Einstein's Elevator
Patrick Bruskiewhich

Source: https://www.researchgate.net/publication/359157658_A_Comparison_of_Light_in_Newton's_and_Einstein's_Elevator

Important:

Todos os materiais de pesquisa destas 9 referências estão armazenados na pasta: (nova)

Data-NovoVivo-012-2022/
Books-012-2022/
20220604

Eclipse-1919-Sobral-Einstein-02

obs: Esta pasta complementa a primeira pasta:

Eclipse-1919-Sobral-Einstein

obs: Referência 09:

Eu fiz uma cópia manualmente deste arquivo e salvei-o com o seguinte nome:

Light-Deflection---20220605.odt

→ fiz a conversão para .pdf

Fiz outra cópia:

A Comparison of Light Detection in Newton's and Einstein's Elevator.edt

obs: Há também em PDF.

⑩ The Measurements of Light's Gravity Deflection of General Relativity were Invalid

/20220604/
Eclipse-1919-Sobral-Einstein-02/

The_Measurement_of_Lights_Gravity_Defle.pdf

Obs: 20 pages (ler) 31 July 2021.

⑪ The_Legacy_of_Einsteins_Eclipse_Gravitational_Len.pdf

Obs: 46 pages. Gravitational Lensing.

⑫ An_Expedition_to_Heal_the_Wounds_of_War.pdf (curar as feridas da guerra)

Obs.: 33 pages.

⑬ 2002.01179.pdf

A Century of Light-Bending Measurements: Bringing Solar Eclipses into the Classroom.

⑭ Eclipse-1919-Sobral-Einstein-02/
General-Relativity/

534-69-86456-1-10-20190919.pdf

→ General Relativity.

⑮ 1902.07287.pdf
Gravitation: from Newton to Einstein

Obs.: 107 pages. 27 Nov 2019.
⇒ Interessantíssimo!!! (Ler!!!)

⑯ Failure-of-Einsteins-theory-of-relativity.II_....pdf
(Ler!!!)

⑰ Einstein-Light-Deflection.pdf (28 pages)
→ Shadow of the Moon and general relativity: Einstein, Dyson, Eddington and the 1919 light deflection.
Obs: veja Pág. 11 - Figure 25
(importantíssimo) (Ler!!!)

⑱ relativityeclipses.pdf (Pasta 01)
→ Relativity and eclipses: the British eclipse expedition of 1919 and their predecessors. ⇒ 37 pages.

⑲ Scientific-American-February-2022-.pdf

The Origins of Space and Time.

Pág. 26.(32). - Pág. 33(39) - (Ler!!!)

---

⑳ 20220605/
Fundamentals in Nuclear
Physics.... Basdevant/

Modern Nuclear Physics -
From Fundamentals to
Frontiers - .. Singapore (2021).pdf

Pag. 58(73) - 2.3 Nuclear Forces.

Obs: Fig 2.5 Two hydrogen atoms separated by a fix distance R and their electrons at displacements r_i from their nuclei.

Obs: Ler este livro!!!
Ler esta parte!!!

Pag. 59(74) Problem 2.1

Imagine how large the density is inside a nucleus.

⇒ Residual interactions

Obs: called van der Waals interaction, is responsible for the global cohesion of liquid and solid states of matter such as liquid water or ice composed of H_2O molecules bound together.

⇒

While the electrical potencial evolves (evolui) radially as

$\propto 1/r$, the van der Waals potential is much weaker and evolves $\propto 1/r^6$.

(21) 20220605/ The Anomalous Magnetic Moment of the Muon/

The Anomalous Magnetic Moment... pdf

Pag. xi (10) - one would write a book about π.

Certainly, the muon's anomalous magnetic moment is a very special number and today reflects almost the full spectrum of effects incorporated in today's Standard Model (SM) of fundamental interactions, including the electromagnetic, the weak and strong forces.

Obs: Ler livro todo!!!

(22) Theory of the Muon Anomalous Magnetic - Springer ...(2006).pdf

Pag. 103(113) Chapter 5 - Strong Interaction Effects in Electroweak Corrections

5.1. Anomalous Quark Triangles: Effects of Strong Interactions.

Obs: Ler livro todo!!!

20220606m

1) the determination of Einstein's light-deflection in the gravitational field of the sun:

Source:
www.sciencedirect.com/science/article/abs/pii/0083665660900052
DoI: https://doi.org/10.1016/0083-6656(60)90005-2
Copy:
Books-012-2022/20220604/
Eclipse-1919-Sobral-Einstein-03/
0083-6656(60)90005-2.pdf
obs: 31 pages - H. Von Klüber.
Source:
https://pt.booksc.eu/book/2581668/3dd652
Other:
Determinations of Effective Wavelengths of Stars made at the Royal Observatory, Greenwich, in the years 1920 to 1925 under the Direction of Sir. Frank Dyson.
Source:
https://pt.booksc.eu/book/6604907/851a93
Copy:
~~2)~~ Data-NovoVivo-012-2022/Books-012-2022/
20220604/Eclipse-1919-Sobral-Einstein-03/
bf01505077.pdf

2) Gravitational starlight deflection measurement during the 21 August 2017 total (eclipse) solar eclipse.
Published 06 March 2018
IOP Publishing Ltd
Source: https://iopscience.iop.org/

10.1088/1361-6382/aaaf2a/meta

Abstract:

Precise star positions near the Sun were measuring during the 21 August 2017 total solar eclipse in order to measure their gravitational deflections.

The equipment, procedures and analysis are described in detail.

A portable refractor, a CCD camera and a computerized mount were set up in Wyoming.

Detailed calibrations were necessary to improve accuracy and precision.

Nighttime measurements taken just before the eclipse provided CUBIC OPTICAL DISTORTION CORRECTIONS.

Calibrations based on star field images 7,4° on both sides of the Sun taken during totality gave linear and quadratic plate constants.

A total of 45 images of the sky surrounding the Sun were acquired during the middle part of totality, with an integrated exposure of 22 s.

The deflection analysis depended on accurate star positions from the USNO's UCAC5 star catalog.

The final result was a deflection coefficient L = 1.7512 arcsec, in perfect agreement with the theoretical value, with an uncertainty of only 3%

IMPORTANT:

DEFLECTION COEFFICIENT

$$\boxed{L = 1.7512 \text{ arcsec}}$$

Estudo:

$1° = 3600$ arcsec $\therefore 1/3600 = 0{,}0002\bar{7}\,°$ (graus)

$1 \text{ arcsec} = 0{,}0002\bar{7}\,°$

$1{,}7512 \times \left(\frac{1}{3600}\right) = 0{,}000486\bar{4}\,°$

Comparação:

Gauss $\rightarrow 179° \, 59' \, 59.32''$

$179° \, 59' \, 59.32''$

$\rightarrow 59.32/3600 = 0.0164\bar{7}$

$\rightarrow 59/60 = 0.98\bar{3}$

Portanto:

$179° + 0{,}98\bar{3} + 0.0164\bar{7} =$

$179 + (59/60) + (59.32/3600) =$

179,99981

$$\begin{array}{r} 180 \\ -\ \underline{179.99981} \\ 0.00018 \end{array}$$

$0.00018°$
$\times 3600 =$

Equation:

$(180-(179+(59/60)+(59{,}32/3600)=0.00018''$ ✡

$\times 3600 = 0{,}68''$ (exato)

Portanto: Equation

$3600 * (180-(179+(59/60)+(59{,}32/3600)=$
$0{,}68''$

Análise:

a) Deflection Coefficiente (2017)

$\angle = 1{,}7512$ arc sec $=$

$1{,}7512 \times \left(\frac{1}{3600}\right) =$

$0{,}000486\bar{4}°$

b) Gauss − curvature (space)

$179°59'59{,}32''$ or 179.99981

desvio $0.00018°$ decimais (graus decimais).

Fazendo as equações:

Diferença:

① $1,7512 \times \left(\frac{1}{3600}\right) = 0.0004864°$

② $179 + (59/60) + (59.32/3600) = 0.00018°$

Portanto:

$1,7512 \times \left(\frac{1}{3600}\right) - (179 + (59/60) + (59.32/3600) =$

$0,000297\bar{5}°$

⇘ 2 cont.)

Source: Data-NovoVivo-012-2022/
Books-012-2022/
20220604/
Eclipse-1919-Sobral-Einstein-03/

1361-6382_aaaf2a.pdf

obs: 22 pages. Donald G. Burns.

3) The Deflection of Light in the Sun's Gravitational Field
Trumpler (44 pages)

Source:
https://m.booksc.me/book/66939511/a36416

Source: 20220604/
Eclipse-1919-Sobral-Einstein-03/
(7 pages) 124220.pdf

Jandaia

Pág. 07 → Summary

A new reduction of the Postdam measures, in which the scale correction is determined from the star observations, yields a result of

$$E = 1.75'' \pm 0.13''$$

for the light deflection at the Sun's limb, and considerably reduces the residuals of the observations.

(1928)

4) Deflection of Light by the Sun's Gravitational Field

May 29, 1919

Source:

https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.1920.0009

5) Gravitation deflection angle of light: Definition by an observer and its application to an asymptotically nonflat spacetime (09 Jun 2020)

Source:

https://arxiv.org/pdf/2001.03290.pdf

Copy:

Books-012-2022/20220604/
Eclipse-1919-Sobral_Einstein-03/
2001.03290.pdf

(9 pages)

6) Deflection of light due to spheroidal oblate static objects

Source:
https://arxiv.org/pdf/2104.14168.pdf

Copy:
Eclipse-1919-Sobral-Einstein-03/
2104.14168.pdf

Obs: 17 pages.

Pag. 14 Table 1 → Comparison of contributions of rotation and oblateness on deflection of light.

Geometry	$\Delta\phi$ (arcsec)	$\Delta\phi_S$ Schwarzschild deflection (arcsec)	$\Delta\phi - \Delta\phi_S$
Kerr (prograde) [31]	1.7789092	1.7789116	-2.411550×10^{-6}
Kerr (retrograde) [31]	1.7789140		$+2.41163 \times 10^{-6}$
Ellipsoidal (Eqn. (47)) (present work)	1.7789169		$+5.336716 \times 10^{-6}$

Obs: Na tabela sem -, mas tem. E sem sinal menos. Checar!!!

[31] - Light deflection due to a charged rotating body - S. Chakraborty - 2015 - 17 pp

Source
https://arxiv.org/pdf/1406.1295.pdf
8 pages.

[47] - General Relativity - R. W. Wald
1984 - p. 145

Copy:
Data-NovoVivo-012-2022/Books-012-2022/
20220604/Eclipse-1919-Sobral-Einstein-03/
General Relativity Wald-1984/
Robert M. Wald - General Relativity -
University of Chicago Express (1984).pdf

Pág. 136(144) 6.3

Geodesics of Schwarzschild: Gravitational Redshift, Perihelion Precession, Bending of Light, and Time Delay of Radar Signals

Pag. 142(150) → (For the Kerr metric with a=M, the corresponding fraction of energy radiated is 42%.)

Pag. 144(152) -

Schwarzschild geometry will capture any photon sent toward it with an apparent impact parameter smaller than the critical value, b_c, given by

$$b_c = 3^{3/2} M \qquad (6.3.33)$$

Hence, the capture cross section, σ, for photon is the Schwarzschild geometry is

$$\sigma = \pi b_c^2 = 27 \pi M^2 \qquad (6.3.34)$$

IMPORTANT:

"Light Bending" ⟹

To analyze the "light bending" effects of the

Schwarzschild geometry

On the light rays which are not captured, it is convenient to derive an equation for the spatial orbit of the light ray by solving equation (6.3.14) for $\dot{r}$ and...

obs.:

(6.3.14) ⇒ we obtain our final equation for geodesics with remarkably little labor,

$$\frac{1}{2}\dot{r}^2 + \frac{1}{2}\left(1 - \frac{2M}{r}\right)\left(\frac{L^2}{r^2} + K\right) = \frac{1}{2}E^2$$

... then dividing $\dot{\phi}$, equation (6.3.13), by $\dot{r}$. We obtain

$$\frac{d\phi}{dr} = \frac{L}{r^2}\left[E^2 - \frac{L^2}{r^3}(r-2M)\right]^{-1/2} \qquad (6.3.35)$$

We wish to find the change,

$\Delta\phi = \phi_{+\infty} - \phi_{-\infty}$, in the angular coordinate ϕ of a light ray in the Schwarzschild geometry traversing a path as illustrated in Fig 6.6-1

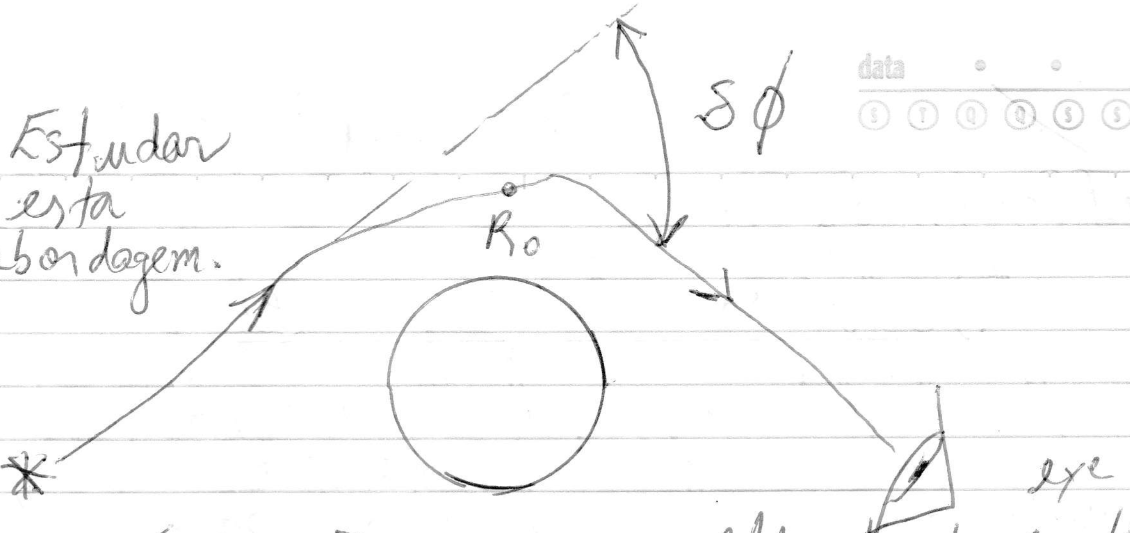

Obs: Estudar esta abordagem.

Fig. 6.6 – Diagram illustrating the "bending of light" effect.

Obs: Estudar o restante.

7) Este item completa o item 6.

[31] – Light deflection due to a charged, rotating body.

Author: Sarani Chakraborty

Date: March 8, 2021 – 8 pages.

Copy: Eclipse-1919-Sobre-Einstein-03/
1406.1295.pdf

Obs: Interessante !!!

8) Bending space-time: a commentary on Dyson, Eddington and Davidson' (1920) "A determination of the deflection of light by the

Sun's gravitational field.

Author: M. Longair   13 April 2015

Source: https://doi.org/10.1098/
rsta.2014.0287

Copy: Eclipse-1919-Sobral-Einstein-03/
rsta.2014.0287.pdf

Obs: 13 pages

9) Relativistic theory of Gravitation based Solely on the Uniform Scaling of Physical Properties

Author: R. Buenker, H. Liebermann

Published: 19 June 2015

Source: archive.org

Copy: Eclipse-1919-Sobral-Einstein-03/
7250153.pdf   (10 pages)

Important:

Pág. 37(6) → Fig. 1

Schematic diagram showing light rays emitted by stars to follow straight-line trajectories as they pass near the sun.

$$\Theta = \frac{\Delta x}{\Delta y} = \int \frac{dc'\,dt}{dy} = \int \frac{1}{c'}\frac{dc'}{dy}\,dx$$

Segunda coluna:

value Θ

the actual value obtained for Θ with the above value for Δx and Δy is found to be

$\boxed{1.7517''}$, in nearly perfect agreement with Einstein's [8] and Schiff's [7] result for this quantity, which in turn is well within the error bars of the measurements of this angle ~~obtrai~~ ~~obtrai~~ obtained during solar eclipses [26,27].

Ver:

There is nothing in this derivation to indicate that either light ray follow a curved path, only that the speeds by which they travel along their respective straight-line trajectories are different.

Jandaia

data S T Q Q S S D

10) On the Gravitation Bending of Light

Copy: Books-012-2022/20220604/
Eclipse-1919-Sobral-Einstein-03/
1301.0004v1.pdf (30 pages)

Published: January 1, 2013

Obs: Interessante!!! (Ler!!!)

Pág 8 - Figure (curvature) ⇙

✡ Newtonian gravitation ⇒
deflection of light by 0.87"
General relativity - Einstein
deflection of light by 1.75".

Veja:

Pág. 16 - 7.1.3
Gaussian Statistical Analysis

11) Early astronomical tests
of general relativity:
the gravitational deflection
of light

Author: Treschman 1 June 2014.

jandaia

Copy: Eclipse-1919-Sobral-Einstein-03/
early astronomical tests ... pdf

Obs: 26 pages.

12) Diffusion Gravity (6): Light Deflection by the Sun-Galaxy Gravity Equipotential Interface

Copy: Eclipse-1919-Sobral-Einstein-03/
2002.0182v1.pdf
(14 pages)

13) The 1919 measurement of the deflection of light.

Copy: Eclipse-1919-Sobral-Einstein-03/
1409.7812.pdf (17 pages)

Obs: Pag. 8 - Figure 1 Geometry of light deflection measurements.

14) Huygens' Principle and Computation of the Light Trajectory Responsible for the Gravitational Displacement of Star Images (2008)

Copy: Eclipse-1919-Sobral-Einstein-03/
V15N3BV2.pdf (20 pages)

Obs. Interessante!!! Imagem repetida (ver)

15) On the Gravitational Bending of Light.

Obs: Ler com cuidado! 14 pages.

Copy: ijaa_2014032110201387.pdf

16) Graviton mass bounds from an analysis of bright star trajectories at the Galactic Center. (Center).

Copy: 7168.pdf (10 pages)

17) Erwin Finlay Freundlich and Testing Einstein's Theory of Relativity.

Author: Klaus Hentschel

Obs: Interessante livro!! 59 pages

20220607////

1) C:/ARQ/1-Source/Caderno-001-20211112/
Caderno-001-20211112.pdf
Pag. 155 - item 101 (20190927)

$$\pi = \frac{355}{113}$$

Source: 20171111/Aurelio Baldor-Geometria.pdf

Pág. 196(200) - 258. El número π.

Pág. 198(202) - 263. Método Gráfico para rectificar aproximadamente una circunferencia.

Rectificar una circunferencia es obtener un segmento rectilíneo (raya) cuya longitud sea igual a la de la curva.

1) Tracemos un diámetro $\overline{AB}$ (Fig. 221)
2) Por A, tracemos la tangente.
3) Sobre dicha tangente y a partir de A tomemos tres diámetros. Se obtiene el punto E.

Fig. 221 →

diámetro = 2,7 cm

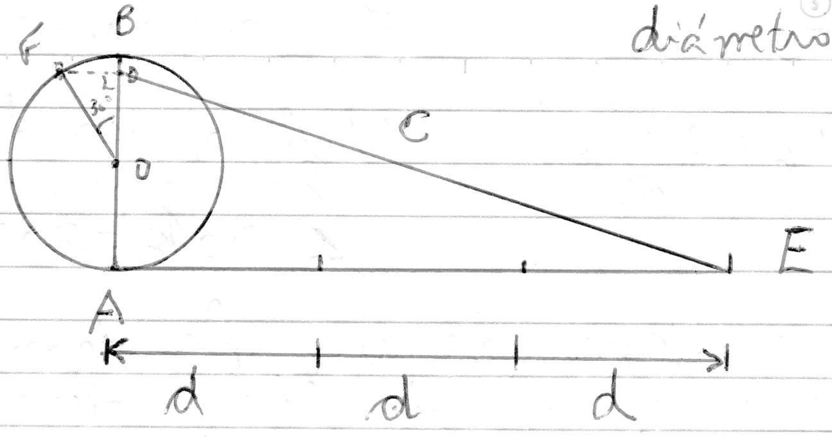

4) Tracemos ∠ BOF = 30°.

5) Bajemos $\overline{FD} \perp \overline{OB}$.

6) Unamos D con E. $\overline{DE}$ es la solución del problema.

264. Justificación de la construcción anterior.

En el Δ AED:

$(\overline{DE})^2 = (\overline{EA})^2 + (\overline{AD})^2$ (1) Teorema de Pitágoras

Pero:

$\overline{AD} = \overline{AO} + \overline{OD}$ (2) Suma de segmentos.

Substituyendo (2) en (1), tenemos:

$(\overline{DE})^2 = (\overline{EA})^2 + (\overline{AO} + \overline{OD})^2$ (3)

Pero:

$\overline{EA} = 3d = 6r$ (4) Construcción;

$\overline{AO} = r$ (5) Construcción;

$\overline{OD} = a_6 = \frac{r\sqrt{3}}{2}$ (6) Apotema de um hexágono inscrito; $\angle DOF = 30°$

Obs: O apótema de um hexágono é o comprimento da linha que une o centro do hexágono com o centro de um lado.

O apótema é a linha perpendicular que coneta (conecta) o centro do hexágono com um lado.

Substituyendo (4), (5) y (6), en (3), tenemos:

$$(DE)^2 = (6r)^2 + \left[r + \frac{r\sqrt{3}}{2}\right]^2 =$$

$$= 36r^2 + r^2 + r^2\sqrt{3} + \frac{3r^2}{4}$$

$$= r^2(36+1)\sqrt{3} + \frac{3}{4}$$ Sacando r^2 en factor común;

$$= r^2(36 + 1 + 1.73 + 0.75)$$

$$= r^2(39.48)$$

$$= 39.48r^2$$

$\therefore$

$$\overline{DE}^2 = \sqrt{39.48 r^2}$$

$$= \sqrt{39.48}\, r$$

$$= 6.28\, r$$

$$= 2 \times 3,14 \times r =$$

$$= 2\pi r$$

Obs:

Baldor - Geometria y Trigonometria ...(2004).pdf

Source:

Data-NovoVivo-012-2022/
Books-012-2022/
20220607/
Geometria - Baldor/
Baldor - Geometria y Trigonometria... (2004).pdf

2) Enciclopédia Prática Jackson - Volume X.

Pág. 259 - item 81 (parágrafo 81)

Retificação da circunferência e quadratura do círculo.

Retificação aproximada.

Outra construção diferente da primeira figura (construção), mas que no fundo é a mesma coisa (30°), diferenciando onde está o apótema.

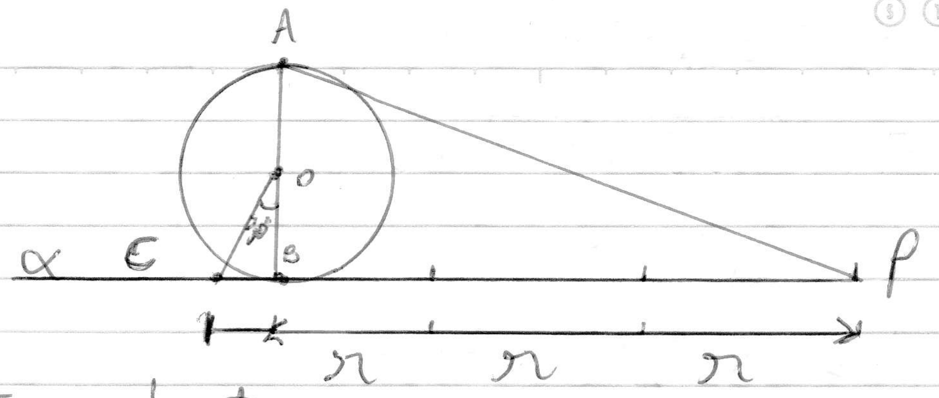

Important:

No Ponto C que é a reta $\overline{OC}$ sendo a hipotenusa de um triângulo retângulo ΔCBO.

O ângulo é de 30° mas a reta é projetada até encontrar a reta α no ponto C.

A partir do Centro O traçamos a reta OC formando um ângulo de 30° com o diâmetro AB.

De C e sobre a tangente em B, transportamos 3 vezes o raio

$CP = 3r$.

Unindo A com P obtemos o segmento ~~AB~~ $\overline{AP}$ aproximadamente igual à semicircunferência.

Com efeito: por ser o ângulo de 30° é

$$CB = \frac{CO}{2}$$

Por conseguinte, recordando o dito para os triângulos equiláteros, temos:

$$CO = \frac{2r}{\sqrt{3}} \quad e \quad CB = \frac{r}{\sqrt{3}}$$

Então:

$$\overline{BP} = \overline{CP} - \overline{CB} = 3r - \frac{r}{\sqrt{3}} =$$

$$= r\left(3 - \frac{\sqrt{3}}{3}\right)$$

Visto que após a multiplicação do numerador e do denominador por $\sqrt{3}$,

$\frac{1}{\sqrt{3}}$ converte-se em $\frac{\sqrt{3}}{3}$

Logo, pelo teorema de Pitágoras, temos:

$$AP^2 = AB^2 + BP^2 =$$

$$= 4r^2 + r^2\left(\frac{28}{3} - 2\sqrt{3}\right) =$$

$$= r^2(40 - 6\sqrt{3}) : 3$$

E, efetuando as operações:

$AP = 3,14153\, r$

e como o semiperímetro da circunferência (ver parágrafo 82) é

$1 = \pi r = 3,1415\, r$

o erro que cometemos nesta construção é de 0,00006 . r, ou seja, 0,002%

---

3) **IMPORTANT**: Pág. 257 - parágrafo 78

Utilizar este estudo para calcular "**LIGHT DEFLECTION**" quando a luz é defletida ao passar pelo Sol e portanto, devemos estudar o espaço antes e depois do Sol.

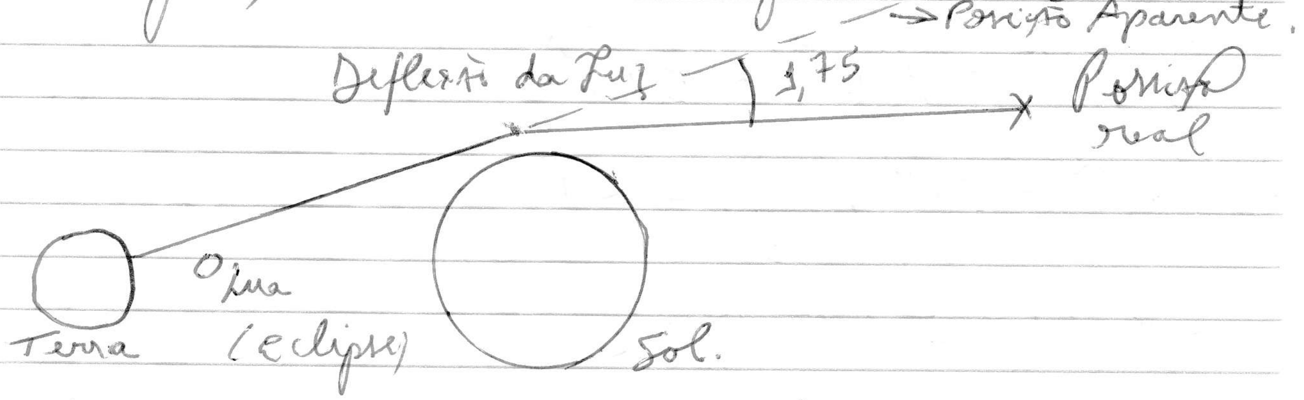
→ Posição Aparente
Deflexão da Luz
1,75
Posição real
Lua
Terra
(eclipse)
Sol.

Obs.: Use a Pág. 257 - Parágrafo 78 - O Triângulo Equilátero e a

Vara Argentina:

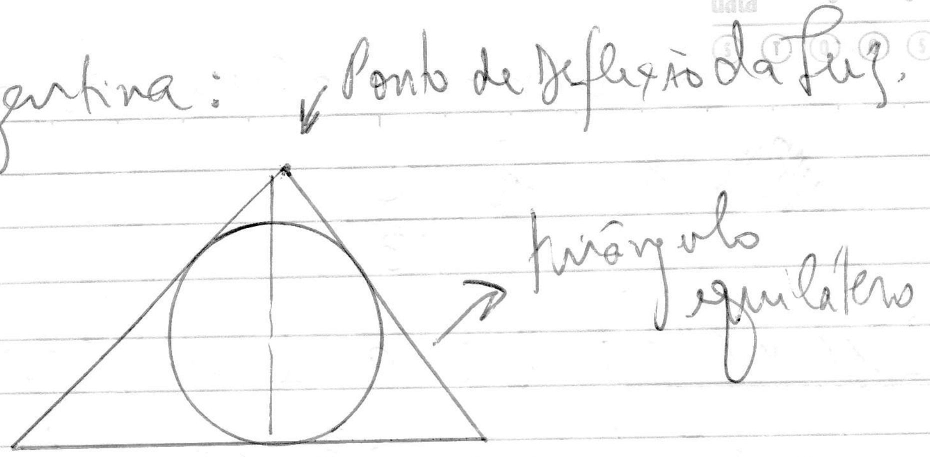
Ponto de Deflexão da Luz.
triângulo equilátero

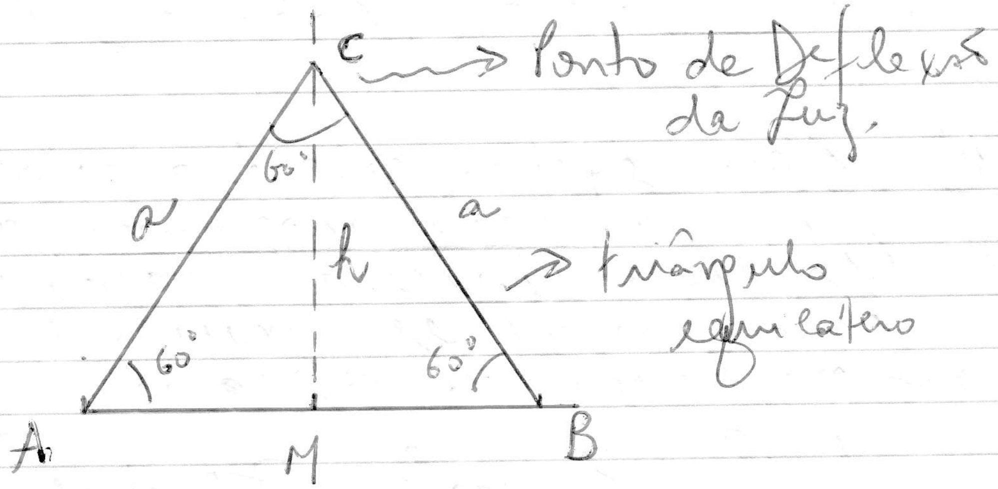
C
Ponto de Deflexão da Luz.
60°
a
a
h
triângulo equilátero
60°
60°
A
M
B

ABC → triângulo equilátero

Se o comprimento do lado é a, a altura h pode ser calculada por meio do Teorema de Pitágoras.
Visto que no triângulo retângulo AMC, é $AM = \frac{a}{2}$, $AC = a$, verificamos:

$$h^2 = a^2 - \left(\frac{a}{2}\right)^2 = \frac{3}{4}a^2$$

Donde $h = a\frac{\sqrt{3}}{2}$

Se conhecemos a altura h, o lado será

$$a = h\,\frac{2}{\sqrt{3}}$$

Exemplo: Se $a = 1$ m temos

$$h = \frac{\sqrt{3}}{2}\,m = 0{,}8666 = 1 \text{ vara}$$

Obs.: Com efeito, este é o valor adotado para a vara argentina, a qual tem sobre as diversas varas européias a vantagem de ser muito simples com relação a áreas:

$$1\,v^2 = \frac{3}{4}\,m^2 = 75\,dm^2$$

$$1\,m^2 = \frac{4}{3}\,v^2$$

Para reduzir um número de varas quadradas a metros, subtraímos-lhes sua quarta parte; para reduzir metros quadrados a varas somamos-lhes a terça parte.

Generalizando: a altura de um triângulo equilátero mede em varas, o que o lado mede em metros.

Exercício: No hexágono regular, o apótema mede em varas o que o lado mede em metros.

4) Jackson - Pág. 257

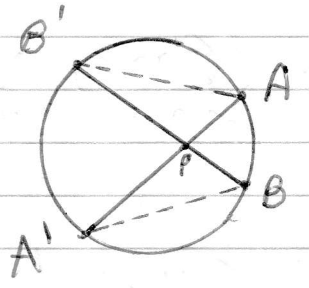

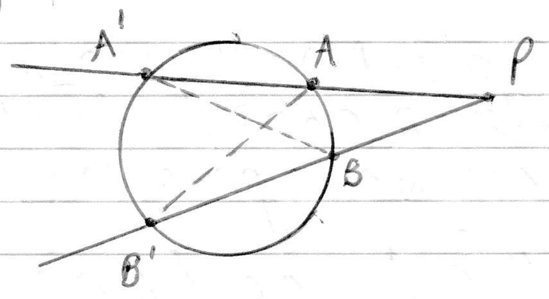

6. Se de um ponto P interior ou exterior a uma circunferência traçarmos as secantes A A' e B B' verificamos sempre

$$PA \cdot PA' = PB \cdot PB'$$

e esse produto constante é a potência do ponto P em relação à circunferência.

(Unindo A com B' e

A' com B

formam-se os triângulos PAB' e PA'B que têm os ângulos iguais (veja o parágrafo 35, exercício 5) e da proporcionalidade dos lados tiramos a importante igualdade enunciada).

Books:

1) Data-NovoVivo-012-2022/Books-012-2022/20220607/The Metric Dirac Equation-curve spacetime/

The Dirac Equation in Curved Spacetime.pdf

Obs: 111 pages. (Ler tudo!!!)

2) 1.855694.pdf

Dirac Equation in Curved Spacetime with the Metric in the Kerr-Schield Form

Author: Tsvetkov (2000)

obs.: 04 pages

3) epjc-s10052-016-4409-x.pdf

Dirac Equation and the Melvin Metric.

Obs. 07 pages - 2016

4) Books-012-2022/20220607/Calculus/Heinbockel-Calculus-2volumes/
✡ Heinbockel - Introduction to Calculus 01.pdf
02.pdf

5) Complex Variables.../
✡ Complex Variables, 6th Edition.pdf
Obs: 730 pages - 2022

20220608,,

1) Books-012-2022/20220608/
Geometria-Castelnuovo/
geometria-intuitiva-emma-
castelnuovo.pdf

Pag. 124 (139) - Le ricerche sul problema della quadratura del cerchio risalgono alla più remota antichità;
il problema si trova nel "Papyrus Rhind" (1650 a. C),
attribuito allo scrittore egiziano Ahmes.

Nel papiro viene indicata come regola di quadratura, senza darne la dimostrazione, la seguente: ↙

è equivalente a un cerchio quel quadrato che ha per lato gli $\frac{8}{9}$ del diametro del cerchio.

obs: $\pi r^2 = \frac{64}{81}$; → Area

ma: $r = \frac{1}{2}$, quindi $\frac{1}{4}\pi = \frac{64}{81}$,

e $\pi = \frac{256}{81} = 3{,}16\ldots$

Pág. 126 (141) –

8. Cerchi passanti per uno o più punti. Poligoni inscritti in un cerchio.

obs: Estudar!!! Interessante!!!

2) 20220608/ Complete Study Pack.../

Complete Study Pack for Engineering ... pdf

obs: 1260 pages.

3) 20220608/ Isosurfaces Geometry, Topology.../

Isosurfaces Geometry, Topology, and Algorithms. pdf

obs: 484 pages.

4) 20220608/ Spazio e Tempo secondo le vedute di A. Einstein/

Castelnuovo - Spazio e Tempo. (1981).pdf

obs: 144 pages. Italian language.

Interessante! Lê-lo!!!

20220609$_{III}$

1) Data-NovoVivo-012-2022/Books-012-2022/20220609/Locally Mixed Symmetric Spaces/
Bruce Hunt-Locally Mixed Symmetric Spaces-Springer (2021).pdf

Obs: Interessante!!!

Pág. 44(62) 1.2.6 Riemannian Symmetric Spaces

1.2.6.1. Orthogonal Symmetric Lie Algebras

The Riemannian symmetric condition can be expressed in terms of symmetric Lie Algebras.

Pág. 259(276) - 2.7.2 The Poincaré Plane

... there is a natural projection of quotients $X^*(N) \to P^1(\mathbb{C})$ which is in fact a Galois cover. Obs: Pág. 264(281)

Pág. 263(280)

natural projection $\pi : SL_2(\mathbb{Z}) \to PSL_2(\mathbb{Z})$ has ± Id as kernel...

Pág. 279, digo, 262 (279)

Veja imagem

$\mathbb{F} \rightarrow$ fundamental domain

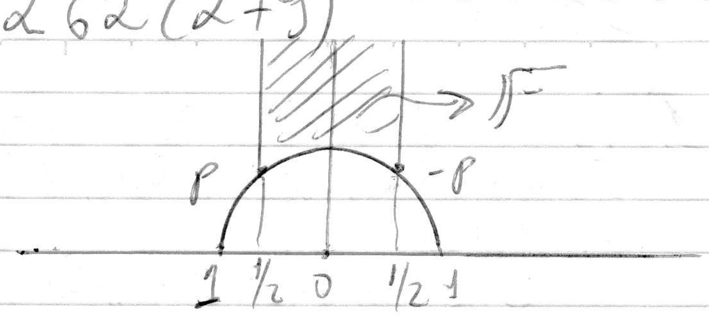

$\mathbb{F} \rightarrow$ in which there is precisely one representative of every Γ-orbit, using the notion

$\Gamma = PSL_2(\mathbb{Z})$ for the

projective group acting on the upper half-plane.

Obs) Livro todo!!!

2) 20220609 / Harmonic Proportion... /

Harmonic Proportion and Form in Nature, Art and Architecture.pdf
(Architecture.pdf)

Obs: Interessante)

3) 20220609 / Concepts in Spin Electronics /

0198568215_Concepts-in-Spin-Electronics.pdf

⇒ Quantum dots.

Obs: Livro todo!!!

4) Books-012-2022/20220609/Fermat/

a) Wiles.pdf

Modular elliptic curves and
Fermat's Last Theorem
by Andrew Wiles.

"Cubum autem in duos cubos, aut quadratoquadratum in duos quadratoquadratos, et generaliter nullam in infinitum ultra quadratum potestatem in duos ejusdem nominis fas est dividere: cujus rei demonstrationem mirabilem sane detexi. Hanc marginis exiguitas non caperet.

Pierre de Fermat

Obs.: Pag. 01.

Total: 109 pages.

b) Wiles-Fermat.pdf (= a)

Obs.: A e B source:

Annals of Mathematics, 141 (1995),
443-551

c) 20220609/Fermat/Elliptic
Curves, Modular Forms and →

Iwasawa Theory... pdf

Obs.: Pág. 494

5) 20220609/ Harmonic Proportion.../

Harmonic Proportion and Form in Nature, Art and Architeture/

Harmonic Proportion... epub.

6) 20220609/ Electromagnetic Retardation and theory of Relativity New Chapters in the Classical theory of Fields/ 0917406249_Electromagnetic.djvu

Obs: 351 pages.

Pág. 181 (193) — 8.1

Transformation of the Lorentz Force

In Chapter 7, we derived relativistic transformation equations for eletric and magnetic fields. Electric and magnetic fields are force fields.

⇒ our transformations equations for electric and magnetic fields could be converted into force transformation equations.

20220611

1) Data-NovoVivo-012-2022/Books-012-2022/
20220610/Symmetry A Mathematical
Exploration/

Mirror-Symmetry-American Mathematics-
important.pdf

Pag. 05(26) Figure 2

The stereographic projection of the sphere. Note that projection from the south pole can be affected by sending

$\theta \to \pi - \theta$.

Pag. 461(482) - Part 3

Mirror Symmetry: Physics Proof.

__Obs__: 482-952 pages.

__Obs__: Livro todo!!!

Pag. 481(502) Part 4

Mirror Symmetry: Mathematics Proof

__Outros aspectos__:

Pag. 487(508) - Complex Curves

22.1 - From Topological Surfaces

to Riemann Surfaces.

There are several natural additional structures one may place on a topological surface Σg.

(1) The first is a Riemannian metric (see Sec. 1.4.1), given by a positive definite symmetric two tensor

$$g_{ij}\, dx^i \otimes dx^j.$$

This gives a notion of distance on the surface.

Obs.: Pág. 12 (33) - 1.4.1

Metric on Manifolds

(2) A conformal structure is given by the data of a Riemannian metric up to multiplication by a (positive) function on Σg.

A conformal structure determines angles between tangent vector of Σg.

Two Riemannian metrics giving the same conformal structure are said to be conformally equivalent.

(3) An almost complex structure is an automorphism of the tangent bundle $J: T\Sigma g \to T\Sigma g$

such that $J^2 = -1$.

Every conformal structure determines a canonical almost complex structure:

J is defined by counter-clockwise rotation by 90 degrees.

An almost complex structure gives the fibers of $T\Sigma_g$ the structure of a one-dimensional complex vector space, where J is interpreted as multiplication by i.

Thus there is a canonical correspondence between the set of conformal structures on Σ_g and the set of almost complex structures.

(4) An almost complex structure is *integrable* if there exist holomorphic charts for Σ_g.

Every almost complex structure on a surface is *integrable* - this is a *deep theorem in complex analysis* (analysis).

A surface Σ_g with a (given) complex structure (an integrable almost complex structure) is called a

non-singular complex curve or Riemann surface.

(5) Finally, every Riemann Surface is algebraic - it can be described as the vanishing set of polynomials in $\mathbb{C}\mathbb{P}^m$ and may be studied using algebro-geometric tools (see ch. 2).

Obs Theorem 22.1.1

(Riemann-Roch)

Suppose V is a rank r vector bundle on a Riemann surface Σ_g. then:

$$h^0(\Sigma_g, V) - h^1(\Sigma_g, V) =$$
$$\deg c_1(V) + r(1-g).$$

Note that all higher cohomology groups of V vanish, by the dimensional vanishing Σ_g (21.1)

This is a special case of the Grothendieck-Riemann-Roch-formula, see Sec 3.5.3. (Pág. 52 (73)).

Jandaia

2) Books-012-2022/20220610/
The Mathematical Paper of Sir
William Rowan Hamilton.../

Copy: 20220611/ Sir William Rowan
Hamilton/

a) William Rowan Hamilton -
Lectures on Quaternions -
Hodges and Smith (1853).pdf ⇙

Pag. 318(457) - item 332

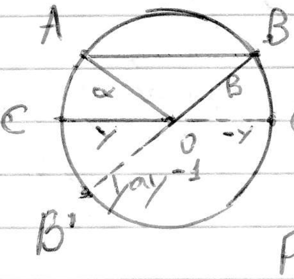

Fig. 71

Consider for this purpose any two radii OA, OB, of a circle (a great circle of the sphere), and draw, as in the annexed figure 71, the diameter COC' parallel to the chord AB; draw also the diameter BOB';

and let it be reqeired (required) to express OB, or its opposite OB' by means of OA and OC (or OC').

Here, because a canonical rotation through two right angles, round either OC or OC' as an axis, would bring the radius OA into the position OB' it results from the present Lecture (arts. 290, 291) that this radius OB' may be expressed as follow:

$OB' = OC \times OA \div OC =$

$OC' \times OA \div OC'$.

But OB is opposite to OB'; wherefore

$OB = -OC \times OA \div OC =$

$-OC' \times OA \div OC'$.

Or writing for conciseness,

$OA = \alpha, \ OB = \beta, \ OC = \gamma,$

the expression for β as a function of α and γ is found to be:

$$\beta = -\gamma \alpha \gamma^{-1}.$$

Obs. arts. 290 e 291 —

Pág. 275(414) e Pág. 276(415), respectivamente.

Art. 290 — If the operating quaternion q reduce itself to a vector, suppose γ, then since its double angle is equal to two right angles, or in symbols,

$\boxed{2 \angle \gamma = \pi}$, ⟸ estudar!!!

the operation symbolized by

$$\gamma(\ )\gamma^{-1}$$

is seen to have the effect of simply REFLECTING the vector or body on which it operates, with respect to the operating vector, γ.

That is to say this operation causes each operand vector,

$T\alpha$

suppose ρ, drawn from the common origin O, to turn conically through (através) two right angles round the line γ, which is here conceived to be drawn from the same origin. and thereby (deste modo) brings this operand ρ, without change of length into a new position ρ' such that while (tal que enquanto) we have the equation between tensors,

$T\rho' = T\rho$, if $\rho' = \gamma\rho\gamma^{-1}$,
the line $\pm\gamma$ at the same time bisects the angle between ρ and ρ': and consequently the following equation between versors also holds good:

$$U.\rho'\gamma^{-1} = U.\gamma\rho^{-1}.$$

For example, in fig. 40, → Pág. 218 (357)

$$\gamma\epsilon\gamma^{-1} = \delta;$$

also, in same figure,

$$\beta\zeta\beta^{-1} = \delta, \text{ and}$$

$$\alpha\epsilon\alpha^{-1} = \alpha^{-1}\epsilon\alpha = \zeta.$$

Art. 291 (Pág. 276 (415)).

Another mode of interpreting the symbol

$$\gamma\rho\gamma^{-1}$$

is the following. We may observe that, by 111, 117,

$$\rho = -\rho^{-1}T\rho^2;\ \gamma^{-1} = -\gamma T\gamma^{-2};$$

and that therefore

$$\gamma\rho\gamma^{-1} = T\rho^2\, T\gamma^{-2} . \underbrace{\gamma\rho^{-1}\gamma}_{};$$

(133, 194) ⇔ denotes the third proportional to the two vectors ρ and γ;

and therefore that (see 134) the vector $\pm\gamma$ bisects the angle between the directions of ρ and $\gamma\rho^{-1}\gamma$;

or by the recent transformation, the angle between ρ and $\gamma\rho\gamma^{-1}$: which was the graphic part of the result of the last article.

And with respect to the METRIC part of that result, we know (by 129, &c) that the tensor of a third proportional is the third proportional to the tensors, and therefore that

$$T . \gamma\rho^{-1}\gamma = T\gamma^2 . T\rho^{-1};$$

an expression which reduces itself to $T\rho$, when it is multiplied by $T\rho^2$, and divided by $T\gamma^2$.

Indeed (De fato) it is clear

from the more general principle
of art. 188, respecting the tensor
of a product that

$$T.\,\gamma\rho\gamma^{-1} = T\gamma\, T\rho\, T\gamma^{-1} = T\rho.$$

Obs.: Art. 188 (Pág. 195(334)).

obs: Pág. 194(333) - Art. 185

$\rho^2 = 1$

Art. 186 → Fig. 32

others: (same book)

I) Pág. 213 (352) - Fig. 36 (Art. 217)

⇒ sphere

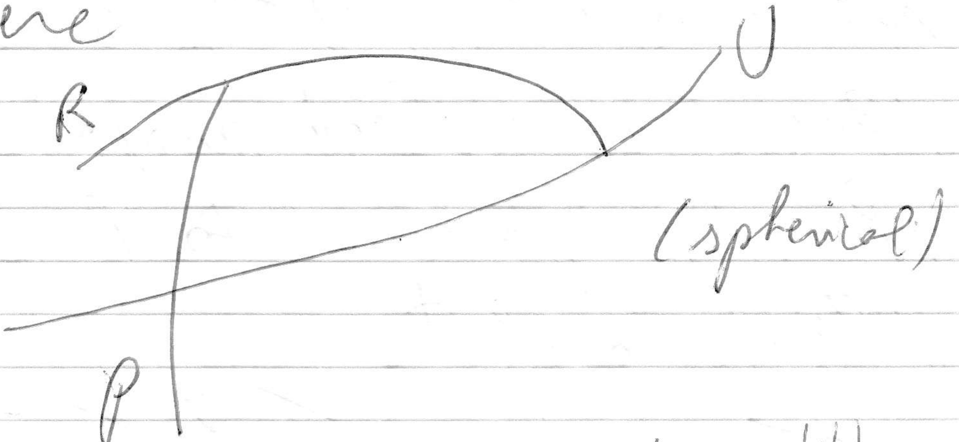

Pág. 214(353) Fig. 37 (estudar!!!)

Art. 219

The construction in Art. 217 may
serve to illustrate some general

properties of quaternion multiplication.

II) Pág. 261(400) - Fig. 53 - Art. 272.

III) Pág. 268(407) - Fig. 57 - Art. 281.

spherical triangle A, B C,

of which the base angles at A and B represent the factors

q and r, while (enquanto)

the rotation round B from A towards the vertex C is positive;

and let B' be the point diametrically opposite to B.

Fig. 57

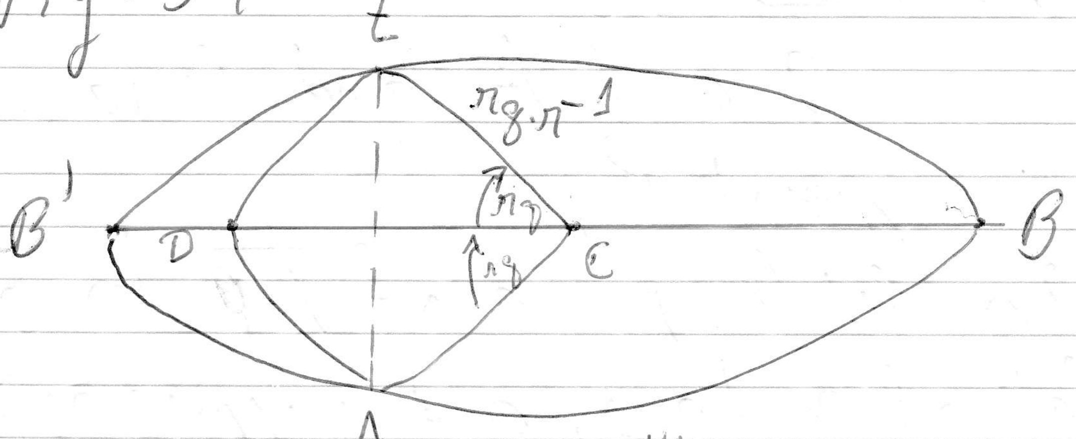

Obs: Ver todos os detalhes!!!

Interessante!!!

Same book ⇒

Item 459 (artigo 459) →
Pág. 458(597).

Apollonius found, and the ancient result has acquired fresh interest in our own days by a remarkable application of to electricity,

that the locus of a point whose (~~de~~ cujos) distances from two given points are in a given ratio of inequality, is (in the plane) a circle.

To investigate this locus by quaternions, let the two given points be

O and A, and the variable point P;

also let the ratio of $\overline{AP}$ to $\overline{OP}$ that of $\underline{\underline{n}}$ to 1, and suppose $n > 1$: then, making $OA = \alpha$ and

$OP = \rho$, the equation

of the locus is,

Jandaia

$T(\rho-\alpha)=nT\rho$, or

$(\rho-\alpha)^2=n^2\rho^2$.

Developing, transposing, etc,
we find successively,

$$(n^2-1)\rho^2+2S.\alpha\rho=\alpha^2,$$

$$\{(n^2-1)\rho+\alpha\}^2=(n^2-1)\alpha^2+\alpha^2=n^2\alpha^2,$$

$$T\{(n^2-1)\rho+\alpha\}=nT\alpha,$$

and finally, $T(\rho-\beta)=c$,
if we make, for abridgment, (em resumo)

$$\beta=\frac{-\alpha}{n^2-1},$$

$$c=\frac{nT\alpha}{n^2-1},$$ so that

$$\beta-\alpha=n^2\beta,\ c^2=-n^2\beta^2=\beta(\alpha-\beta).$$

Jandaia

Fig. 96

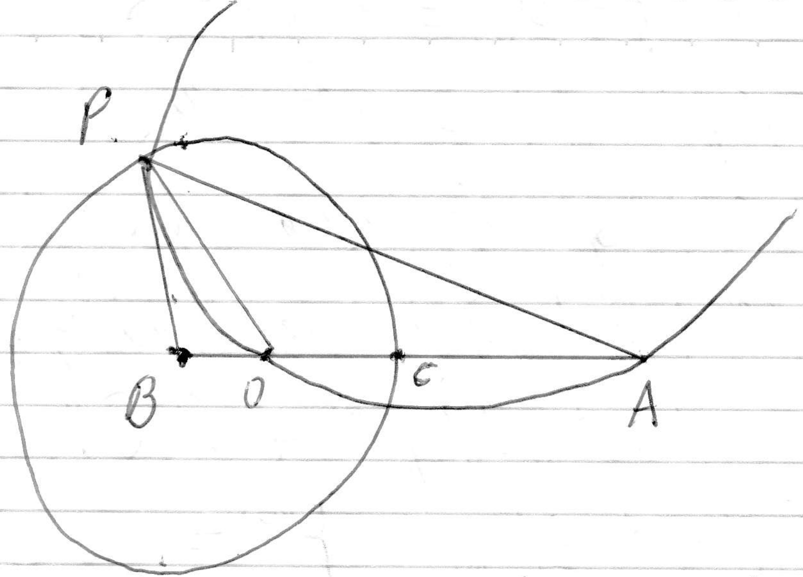

Hence (porismo) follows this construction, which agree to known results.

Cut the given line AO externally at B, in the duplicate of given ratio of the sides, so as to have

$$AB = n^2 OB;$$

take BC a geometrical mean between the segments BO, BA; and with centre B, and radius BC, describe a spheric surface; it will be (in space) the required locus of all the points P, for

Whice $\overline{AP} = n . \overline{OP}$

As a verification, let $C - B = \gamma$,
$P - B = \sigma$; we shall have
$A - B = n\gamma$, $O - B = n^{-1}\gamma$,
$P - A = \sigma - n\gamma$,
$P - O = \sigma - n^{-1}\gamma$; it ought (eu pensei) (it ought to be (deveria ser))
then turn out that (acontece que)

$T(\sigma - n\gamma) = T(n\sigma - \gamma)$, if $T\sigma = T\gamma$;
and accordingly, (adequadamente)
$(\sigma - n\gamma)^2 = (n\sigma - \gamma)^2 = (n^2+1)\gamma^2 - 2nS.\gamma\sigma$,
if $\sigma^2 = \gamma^2$.

It is evident from elementary geometry that the fixed locus of P, constructed as above, cuts perpendicularly the

CIRCLE CIRCUMSCRIBED

about the variable triangle AOP, or that is radius BP is a tangent to this circumscribed circle; and this result also might be confirmed by calculation with quaternions, if we chose to use here the conclusion of art. 198, respecting the construction by a tangential vector, of the continued product of the three sides of a triangle inscribed in a circle.

Obs: Art. 198 →

Pág. 201 (340)

Points:

(A − C)(C − B)(B − A), or

(B − A)(C − B)(A − C),

for any three points of space A, B, C, supposed at first to be not situated on one straight line, but to be the corners of a plane triangle; namely, as denoting a certain line or vector, whose length represents the product of the lengths of the sides of that triangle, while its direction is that of the tangent at A to the segment ABC, of the circle circumscribed about it.

20220611$_M$ (cont).

3) Data-NovoVivo-012-2022/
Books-012-2022/
20220611/

Obs. Tem também
20220610

Sir William Rowan Hamilton -
Elements of Quaternions/

Sir William Rowan Hamilton...pdf.

Pág. 334(408) → (18)
↳ refere-se à página

331(405) - Section 7

On the Fourth Proportional to Three
Diplanar Vectors.

Portanto:

Pág. 334(408) - item (18)
Fig. 68.
↳ meeting the great circle.

Obs: Analisar!!

Important:

Item (23) → Instead of conceiving (concepção)
that the point D, on the line OD,

has been reflected into the position F, or OF, with respect to the point B, or the line OB, with a similar

SUCCESSIVE REFLETION

from F to E, we may conceive that a point has moved along a small semicircle, with B for pole, from D to F, as indicated in Fig. 69, and then along another small semicircle, with A for pole, from F to E;

and we see that the result, or effect, of these two successive and semicircular motions is equivalent to a motion along an arc DE of a third small circle, which is parallel (as before) to the great circle through B and A, and has a projection RS thereon (here also), which (still as before) is double of the given arc BA.

20220612m

1) Data-NovoVivo-012-2022/
books-012-2022/
20220610/
Analysis of the Gravity Field Direct.../
Analysis of the Gravity Field...pdf

Pág. 12(25) - 1.4

The Divergence Operator; Gauss Theorem.

The divergence of the field $F(P)$ is defined as

$$\text{div } F(P) = \nabla . F(P).$$

Obs.: Cartesian Coordinates
Cylindrical Coordinates
Spherical Coordinates

Pág. 14(27) - Theorem 1.1 (Gauss)

Consider a volume V in $\mathbb{R}^3$ with a smooth boundary endowed with a continuous normal $\underline{n}$ and assume that a field F is defined on V such that $\nabla . F$ is at least an integrable function there; then

(1.50) ⇒

$$\int_V \nabla \cdot F(P)\, dV_P = \int_S F(P) \cdot n_P \, dS_P$$

Here n_P must be understood as the outward normal to S at P.

Obs: Veja livro todo!!!

Pag. 20 (33) Fig 1.7 (similar com curvatura)

2) Books-012-2022/
20220610/
Introducing Einstein's Relativity/
Introducing Einstein's Relativity...pdf

Pág. 303 (318) - 16.7
Bending of Light

Fig. 16.9 → Straight-line motion of a light ray in special relativity.

Fig. 16.10 → Deflection of light in a gravitational field.

Obs: Veja até Pág. 318 (333).
Livro todo!!!
Excelente!!!

3) Books-012-2022/
20220611/

a) Modular-Forms/

a1) Modular-Forms-2015.pdf

a2) wiles

b) Skill in Mathematics-Algebra.../

Skillmathematics JEEMain Algebra.pdf

c) Skill in Mathematics-Differential Calculus/

Skillmathematics JEEMain Differential Calculus.pdf

d) Skill in Mathematics-Play with Graphs.../

Skill mathematics JEEMain Play Graphs.pdf

Pág. 51(58)-Curvature and Transformations.

Cont. c) Differential Calculus

Pág. 215 - Session 2 [Pág 204(215]

When $f(x)$ Horizontally Transforms to $f(x \pm a)$, where a is any Positive Constant, when $f(x)$ Transforms to $af(x)$ or $\frac{1}{a} f(x)$, when $f(x)$

Transforms to $f(ax)$ or $f\left(\frac{x}{a}\right)$; when $f(x)$ Transforms to $f(-x)$, when $f(x)$ Transforms to $-f(x)$

Pág. 207 (218) Example 14

Draw the graph of $y = -\log(x)$, when the graph of $y = \log x$ is known.

Solution: As $y = \log x$ is given, then $y = -\log$ is the image of $\log x$ in the x-axis as plane mirror.

Thus, it can be drawn as show in figure.

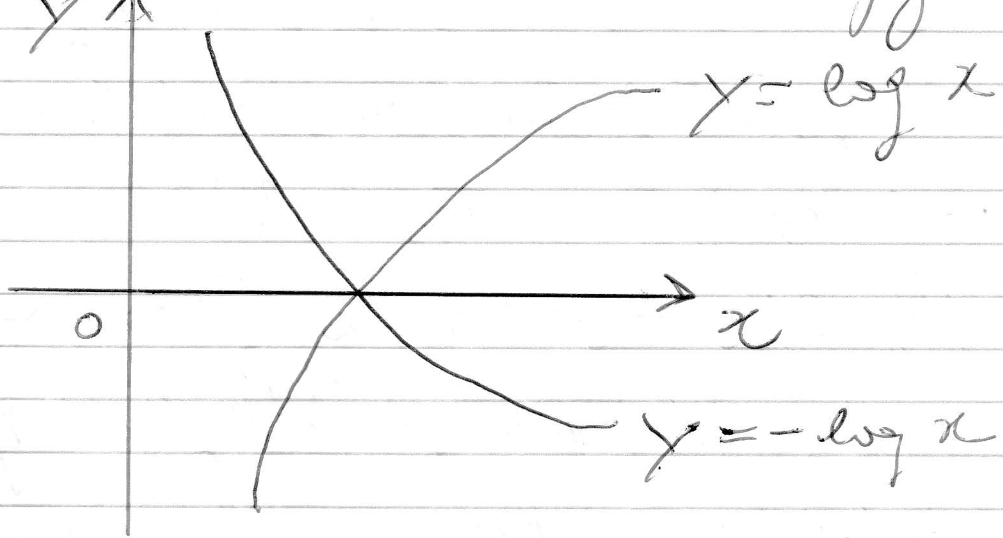

Obs: Faz parte do Chapter 04 — Graphical Transformation — Pág. 201 (212)

20220613

1) Data-NovoVivo-012-2022/
Books-012-2022/
20220613/
Relativity/

a) A.Held - General relativity and gravitation.djvu
volume 1

b) A. Held - General relativity and gravitation
...volume 2...djvu

Obs: A e B → importantes   a) 616 pages
b) 556 pages.

c) Anadi Jiban Das - Tensors -
The Mathematics of Relativity Theory
and Continuum Mechanics (2007).pdf

d) Hall - Symmetries and curvature
structure in general relativity...pdf

e) Hawking - General relativity...pdf

Obs.: 935 pages.

f) Hartle - Gravity...pdf

Obs.: 589 pages.

g) Jerry... An Introduction to General
Relativity and Cosmology.pdf

Obs.: 556 pages. Livro todo!!!

Jandaia

Pág. 03(25) - Fig. 1.1.

Real planetary orbits, in consequence of various perturbations, are not ellipses but non-closed curves.

Mercury → $\cong 1.5^\circ$ per 100 years.

h) Øyvind (Øyvind) -

Einstein's general theory of relativity.pdf

Obs.: 538 pages. Livro todo!!!

Pág. 404(422) - 15.2

Homogeneous spaces.

Fig. 15.1 - In a homogeneous space you can get anywhere on the manifold using an isometry.

Others:

Pág. 21(39) - Chapter 2

The Special Theory of Relativity

2.1 - Coordinate systems and Minkowski-diagrams.

Pág. 30(48) - 2.7.

The Lorentz transformation.

Pág. 35(53) - 2.10

Hyperbolic motion

Pág. 37(55) - 2.11

Energy and mass

Obs.: Pág. 38(56) → 2.66

$$\boxed{E=mc^2}$$

Pág. 38(56) - 2.12

Relativistic increase of mass.

item (2.68)

$$\int_{m_0}^{m}\frac{dm}{m}=\int_{0}^{v}\frac{v\,dv}{c^2-v^2}$$

...

(interessante) item (2.71) ↙

$$E\approx m_0c^2+\frac{1}{2}m_0v^2$$

Obs.: O valor é aproximado!!! Estudar!!!

IMPORTANT:

Pag. 39(57) - 2.13 Tachyons

Particles cannot pass the velocity-barrier represented by the velocity of light.

However, the special theory of relativity permits the existence of particles that have always moved with a velocity

$v > c$.

Veja: Figure 2.15

Obs: Estudar TACHYONS.

Livro:

M. N. Kreisler -

Are there faster-than-light particles?

Source:

20220613/ Kreisler - Are there faster-than-light particles.pdf

Obs: American Scientist, Vol. 61, No. 2 (March-April 1973), pp. 201-208.

https://pt.booksc.eu/book/48463379/117583

Other book:

Laidlaw - Tachyons, Boundary Interactions and Genus Expansion in String theory [thesis] (2003).pdf

Obs.: 148 pages.

The special theory of relativity, applied to tachyons leads to the following paradox.

Using a tachyon telephone a person, A, emits a tachyon to B at a point of time t_1.

B moves away from A.

The tachyon is reflected by B and reach A before it was emitted, see Fig. 2.15.

If the tachyon could carry information it might bring an order to destroy the tachyon emitter when it arrives back at A.

Figure 2.15 -

A emits a tachyons at the point of time t_1. It is reflected by B and arrives at A at a point of time t_2 before t_1.

Note that the arrival event at A is later than the reflection event as measured by B.

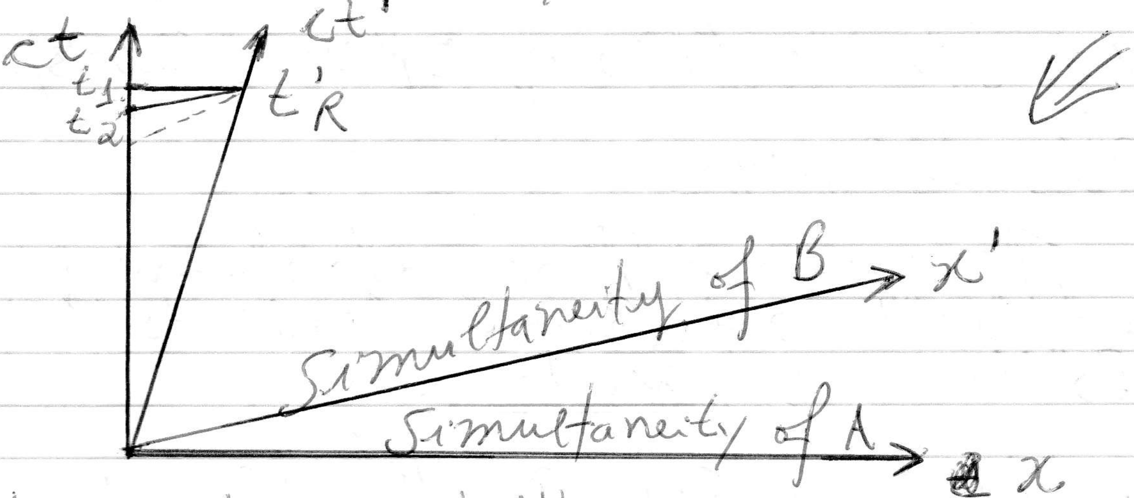
ct
ct'
t_1
t_2
t'_R
Simultaneity of B
x'
Simultaneity of A
x

Obs: Interessante!!!

Pág. 40 (58) - 2.14

Magnetism as a relativistic second-order effect

Obs: Electricity and magnetism are described completely by Maxwell's equations of the electromagnetic field

Obs: Estudar o efeito

✡ Analisar se a "grandade" também é um "effect" das forças fundamentais da matéria.

Pág. 51 (69) - Chapter 3

Vector, Tensors, and Forms

Obs.: Faz parte → Part II

The Mathematics of the General Theory of Relativity.

i) Relativity, groupes et topologie II. pdf

Obs.: 1358 pages.

Obs.: Pág. 655 (690) - 16

The Gravitational field

Veja: 16.2 Diffeomorphism group.

Obs: Livro todo!!!

j) Petkov - Relativity and the nature of spacetime - Springer (2005). pdf

Pag. 70 (77) - 4.2

Derivation of the Lorentz Transformations

Fig. 4.9

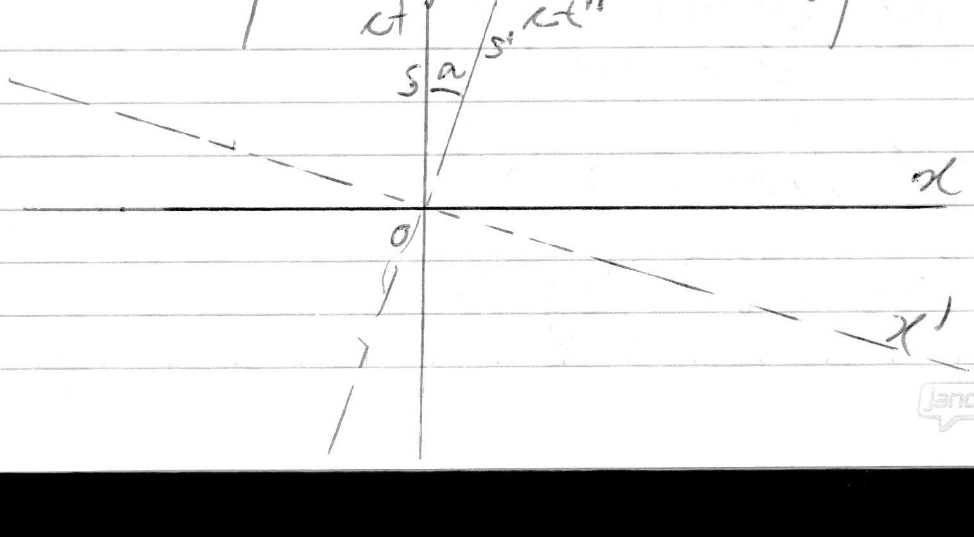

data
S T Q Q S S D

Pág. 72 (79)   (4.1)

$$x' = x \cos a - y \sin a,$$
$$y' = x \sin a + y \cos a.$$

Obs: Veja também:

Fig. 4.10 – Two coordinate systems are rotated in a two-dimensional Euclidean space.

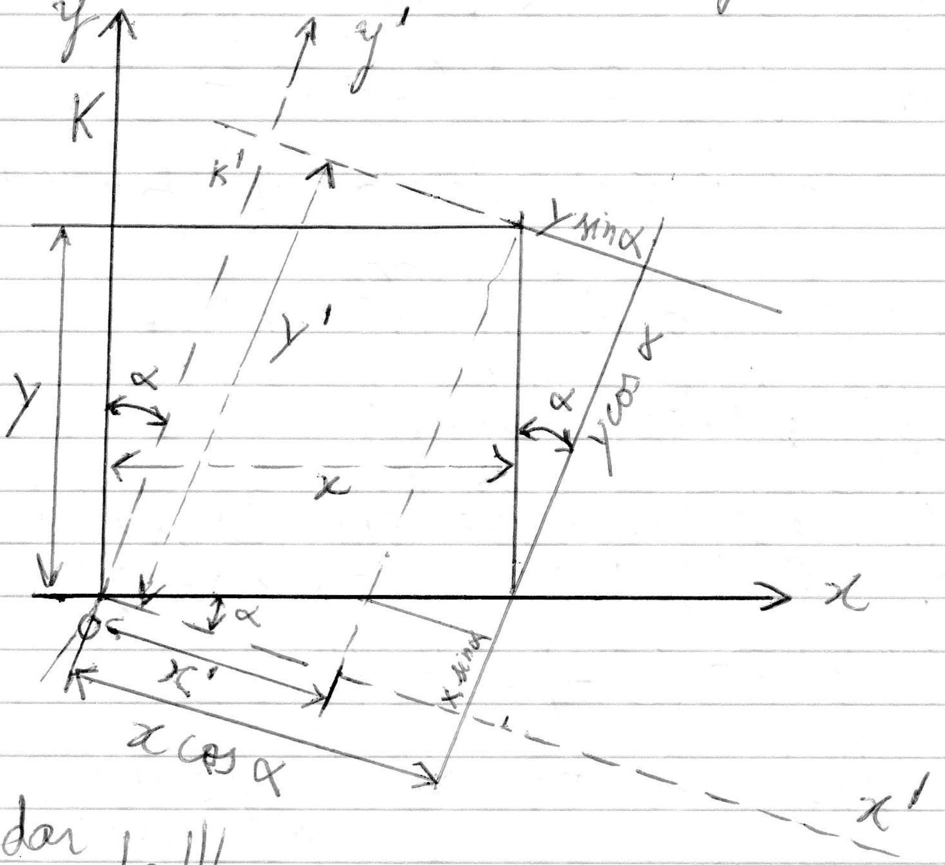
y
y'
K
K'
y sin α
y'
α
Y
α
y cos α
x
x
α
x'
x sin α
x cos α
x'

Obs: Estudar esta parte!!!

Obs: Livro todo!!!

Tem 290 pags.

K) Relativity on curved manifolds (1990).djvu

L) Relativity, Groups, and Topology I - (1964).djvu

Obs.: 943 pages. Livro todo!!!

M) Rotating Objects and Relativistic Physics.djvu

N) Sean Carroll - Spacetime and Geometry.pdf

→526 pages

Obs: Livro todo!!

Obs.: Chapter 9

Pag. 376 (389)

Quantum Field Theory in Curved Spacetime.

Pág. 385(398) - 9.3

Quantum Field Theory in Flat Spacetime

Pág. 394(407) - 9.4

Quantum Field Theory in

Curved Spacetime.
Pág. 402(415) 9.5
The Unruh Effect

... is manifested even in flat spacetime:

the Unruh Effect,

which states that an accelerating observer in the traditional

MINKOWSKI VACUUM STATE

will observe a thermal spectrum of particles.

Historically the Unruh Effect was discovered in an attempt to understand the physics underlying the

HAWKING EFFECT

(thermal radiation in the presence of a black hole event horizon).

Obs: Estudar!!

→ 0) Stora - Relativity, Groups and Topology... livro (679 pages)

⇘ P) Synge - Relativity - the special theory ... (1958). djvu

→ 464 pages.

Pag. 29 (43) - §16 -

Distance between particles

Fig. 7 - Distance D between adjacent particles, or world lines.

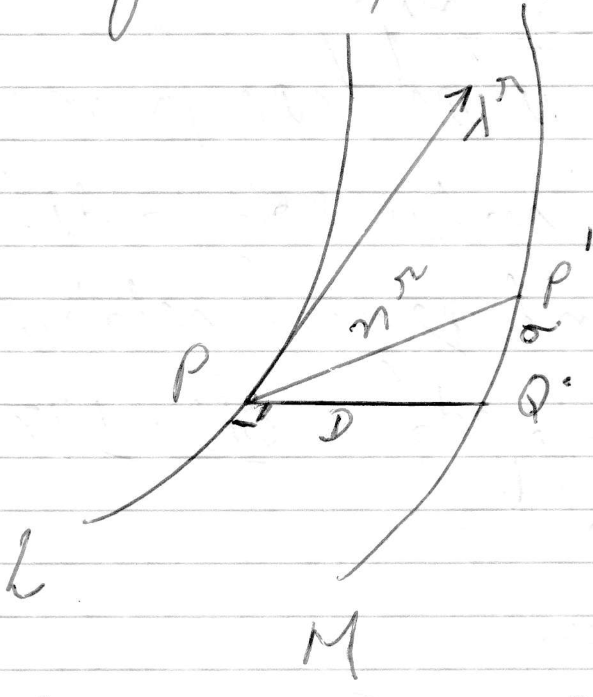

Obs.: Analisar!!!

P) The Geometry of Minkowski Spacetime - An Introduction to the Mathematics of the Special Theory of Relativity. p djvu Obs. 271 pages.

Obs: Livro todo!!!

Pág. 100 (114) - Chapter 2

Skew-Symmetric Linear Transformations and Electromagnetic Field

Pág. 143 (157) - Chapter 3

The Theory of Spinors

→ Lorentz Group

↳ estudar transformation!!!

R) Warner - Introduction to differential geometry and general relativity (2002).pdf obs: 138 pages

Obs: Livro todo!!!

Pag. 114 (depois).

Pag. 40 - Chapter 6

Riemann Manifolds

Pag. 50 - Chapter 7

Locally Minkowskian Manifolds: An Introduction to Relativity

Pág. 82 - Chapter 10
The Riemann Curvature Tensor
Pág. 114 - Chapter 14
The Einstein Field Equations and
Derivation of Newton's Law

5) Wolfgang - Relativity -
special, general, and cosmological
(2006).pdf Obs.: 447 pages.
Livro todo!!!
Part II - General Relativity (8)
Pág. 163 (182)
8 - Curved spaces and the basic ideas of general relativity
Pág. 203 (220) - Chapter 10
Geodesics, curvature tensor
and vacuum field equations.

2) 20220613/
Relativity/
Choquet-Bruhat/

data

S T Q Q S S D

Obs.: Diretório todo!!!

a) Analysis, Manifolds and Physics Volume 1 ... pdf

Obs.: 649 pages.

b) Analysis, Manifolds and Physics Volume 2 ... pdf

(II) Obs.: 559 pages.

Pág. 235 (252) V.

Riemannian Manifolds.

Kählerian Manifolds.

c) Berard – Geometrie differentielle. pdf

d) Choquet-Bruhat - General relativity and the Einstein Equations ... pdf

Obs.: 812 pages - Livro todo!!!

e) Choquet-Bruhat - Introduction to general relativity, black holes, and cosmology.pdf

Obs.: 300 pages - Livro todo!!!

3) 20220613/ Shadowing and Hyperbolicity/ Shadowing.pdf

4) 20220613/ Spezielle.../ Spezielle..Relativitätstheorie.pdf

→ 551 pages

20220614

1) Data-NovoVivo-012-2022/
Books-012-2022/
20220614/

An Introduction to Groups and their Matrices..../

An Introduction to Groups....pdf

Pág. 02(17) 1.2.1- Symmetry Operations

Pág. 04(19) - Steven Weinberg,
(1933-2021)
Nobel Laureate in physics 1979,
has written that

SYMMETRY is the

"Key to nature's secrets",
which is why the application of
symmetry principles to physical
is the subject of this text.

Pág. 193(208) - 8.3

Rotation and Angular Momentum
operators

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

2) 20220614/
Exploring Geometry/

a) Donald... - Exploring Analytical Geometry... pdf
Obs. 883 pages

→ Livro todo!!!

Pág. 224(242) 15.4 - Reflections
↳ Transformation

b) Harold ... - Turtle Geometry...pdf

Pág. 201(220) - Chapter 5
Turtle Escapes the plane.

Pág. 202(221) - 5.1.1. Turtle Lines.

Pág. 207(226) - 5.1.3. Angle Excess.

Pág. 211(230) - 5.1.5. Excess and Area

Theorem: The area of any polygon is the sum of the pieces in any polygonal subdivision.

Obs.: But it is obvious that K couldn't be a universal constante like π. (?)

K depends on the surface.

$$K = \frac{E}{A} = \frac{2\pi}{2\pi r^2} = \frac{1}{r^2} \quad //$$

Pág. 214(233) - 5.2 - Curvature

Obs: Livro todo!!!

c) Hvidsten - Exploring Geometry (2012).pdf

Obs.: 477 pages

d) Hvidsten - Exploring Geometry (2017).pdf

Obs.: 545 pages.

A segunda edição amplia a abordagem, mas há diferenças

c) há Chapter 8 - Non-Euclidean Transformation - Pág 313

d) Há o Hyperbolic Geometry.

Obs: Portanto, leia-os em conjunto.

O livro c) tem Chapter 7 -

I) Non-Euclidean Geometry - Pág. 263

II) Chapter 8 -

Non-Euclidean Transformation - Pág. 313

O livro d) tem Hyperbolic Geometry - Pág. 271
Elliptic Geometry - Pág. 317

Chapter 9 - Projective Geometry - Pág. 359

e) Jon Allen - Making Geometry .. pdf

Pág. 8(7) -

"God is a circle whose centre is everywhere and whose circumference is nowhere."

f) Solomon .. Linear algebra, geometry and transformation .pdf

Obs: 469 pages - Livro todo!!!

Pág. 301(317) - Chapter 6

Orthoganality

→ bases comprised of mutually orthogonal unit vectors.

Pág. 341(357) - Chapter 7

Linear Transformation

→ Kernel and Image

we explore the geometric structure of linear transformations.

g) Umble - Transformational Plane Geometry .. pdf

Obs: 235 pages - Livro todo!!!

IMPORTANT:

Converter este livro todo em simulações vivas e ajustados ao M.U.S.

MATHEMATICAL UNIFORME SPACE

M.U.S. -> fator de conversão para uma nova

GEOMETRY and MATHEMATICS.

H) Venema - Exploring Advanced Euclidean Geometry with Geogebra.pdf

Pág 81 - Chapter 10 - Circles and Lines

Pág. 85 - Chapter 11 - Applications of the Theorem of Menelaus.

Pág. 111 - Chapter 14 - The Poincaré Disk

3) 20220614/ Geometry of Complex Numbers/

a) Algebraic geometry over the complex numbers (2012).pdf

b) Geometry_of_Complex_Numbers.pdf

c) Hahn - Complex Numbers and Geometry (1996).pdf (1996)

d) Introduction to the geometry of complex numbers (2008).pdf

e) Yaglom - Complex Numbers in Geometry (1968).pdf

4) 20220614/
Godel - Collected works.../

a) Godel - Collected Works - Volume I.pdf
b) Godel - Collected Works - Volume II.pdf
c) Godel - Collected Works - Volume III.pdf
d) Godel - Collected Works - Volume IV.pdf
e) Godel - Collected Works - Volume V.pdf

5) 20220614/
JEE Advanced Mathematics - Algebra.../
JEEAdvMathAlgebraFirst.pdf

Obs.: 968 pages.

6) 20220614/
Principles of Fourier Analysis.../

Jandaia

Principles of Fourier Analysis
2nd edition. pdf

Obs.: 805 pages.

Pág. 526 - Part IV -
Generalized Functions and Fourier Transforms.

Pág. 512(529) -
Since ϕ is a Gaussian function,
so is $\mathcal{F}[\phi]$.

7) 20220614/
Quantum field theory...|

a) A Mathematical Bridge -
An Intuitive Journey in Higher Mathematics ...pdf

Pág. 205(228) - Chapter 4.
Algebra and Geometry

Pág. 238 (261) - Chapter 4.4
Distance, Length and Angle.

Obs.: Pág. 198(221) 198(221)

$$\pi^2 = \frac{\pi^2}{3} + 4\sum_{n=1}^{\infty}\frac{1}{n^2}$$

Therefore,

$$\sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6}.$$

Pág. 199(222) →

$$\exp(i\theta) = \cos(\theta) + i\sin(\theta)$$

Pág. 237(260) –

4.3.2.3 – The transportation problem

$$(1)\ \sum_i X_{ij} = e_j$$

b) Bridge to Abstract Mathematics (2012).pdf

c) Deaconu – A Bridge to Higher Mathematics (2016).pdf

Obs: b e c se interligam.

d) IMPORTANT: ⇒

d) Eberhard Zeidler -
Quantum Field theory
1-3.pdf
Obs.: 3312 pages

Obs.: Search: GAUSS.

e) Modeling with Mathematics -
A Bridge to Algebra II ... 2012.pdf

Obs.: 641 pages.
Pág. 301 (321) - Chapter 9
Symmetry and Transformations

f) Quantum Mathematical Physics -
A Bridge between Mathematics and
Physics (2016).pdf.

Obs.: 517 pages.
Pág. 19 (34) - Is there a C-Function
in 4D Quantum Einstein Gravity?

g) 20220614/
Quantum Mechanics A Self-
Contained Course/

⇒

a) Quantum Mechanics - Volume 1,
(2001). pdf.
b) Quantum Field Theory - Volume 2,
(2003). pdf
9) 20220614/
Studies in the History of Indian
Mathematics/
a) Mathematical Astronomy -
Volume 01. pdf
obs.: 523 pages.
b) Mathematical Astronomy -
Volume 02. pdf
obs.: 630 pages.
10) 20220614/
The Encyclopedia of Cosmology/
The Encyclopedia of Cosmology
(In 4 Volumes).. pdf
obs.: 1399 pages

11) 20220614/
The Geometry of Special Relativity...|
The Geometry of Special
Relativity... pdf

Obs.: Pág. 49(59)
3.1 Lorentz Transformation of
Coordinates.

12) 20220614/
Topological, Differential and
Conformal Geometry of Surfaces/
(Universitext),... Topological... pdf

Obs.: Pages - 282.

13) 20220614/
Important/

a) M. Nelkon - Advanced Level
Physics... pdf (1970)

Obs.: 1129 pages.

b) Progress in physics.pdf obs.: 95 pages

c) Progress in Physics II.pdf Obs.: 179 pages

Pág. 15 (16) - New Approach to Quantum Electrodynamics

Pág. 89 (90) - Remarks on Conformal Mass and Quantum Mass.

Pág. 91 (92) - Gravitation and Electricity.

Pág. 144 (145) - Smarandache (autor) Kaluza-Klein-Carmeli Metric from Quaternion-Clifford Space, Lorentz' Force, and some observables

Pág. 166 (167) - Reply to the "Certain Conceptual Anomalies in Einstein's Theory of Relativity" and Related Questions.

d) The elements.pdf Obs.: 456 pages

e) The_Measurements_of_Light_Defle.pdf Obs.: 20 pages

Obs.: Interessante, juntar com os outros papers sobre deflection light.

f) The_precise_calculations_of_the_constant.pdf

Obs.: 10 pages.

Interessante:

Pág. 186(4) - III. ↙

The strict proof that general relativity cannot describe the elliptical orbital motions of planets in the Solar System

Obs.: The geodesic equation of the Riemann geometry.

Pág. 187(5) - D.

The determination of constant E.

Pág. 189(7) - IV.

IV. General Relativity cannot correctly describe the motion of light in the Solar Gravity Field.

Pág. 190(8) - C.

The problem caused by Einstein's

hypothesis ds = 0

g) Trust_in_expert_testimony-
Eddington_1919.pdf

Obs.: 11 pages.

Visão geral: Lê-lo!!!

—

Cópias de outras pastas:

14) 20220614/ (Source: 20220601)
Geometry-Gray/

a) Geometry-Cambridge(2012).pdf

Obs.: 604 pages - Livro todo!!!

b) Global Differential Geometry
(2001).pdf

Pág. 110(141) - Can tangent sphere bundles over Riemannian manifolds have strictly positive sectional curvature?

Pág. 162(193) - Almost parallel structures.

Pág. 239 (270) - On the volume of compact spacelike hypersurfaces with hyperbolic boundary in Minkowski space.

Pág. 296 (327) - On a new class of contact Riemannian manifolds

voltando:

Pág. 268 (299) - An extension of A. Gray's investigations on small geodesic balls.

Pág. 277 (308) - Is it admissible to linearize the Einstein equation in the presence of matter?

Obs.: Atentar-se !!!

Pág. 320 (351) - Deformations of constant mean curvature surfaces in half space models

Pág. 339 (370) - Isometric immersions without positive Ricci curvature

Pág. 371 (402) - How to use Mathematica to find cyclic surfaces of constant curvature in Lorentz-Minkowski space.

c) Gray - Geometry - Solutions (2nd) (2012).pdf Obs.: 221 pages

d) Gray - Geometry - Cambridge (2012).pdf
Obs.: 603 pages. Livro todo!!!
Interessante!!!

Jandaia

e) Gray - Modern Differential Geometry of Curves with Mathematica...pdf.

Obs.: 996 pages.

Usar como referência de estudo.

f) Gray - Worlds Out of Nothing...pdf

Obs.: 410 pages.

History of Mathematics - Gauss and others.

Interessante!!!

g) Gray - Geometry (1999).pdf

Obs.: 513 pages.

Pág. 261 (274) - Chapter 6.

Non-Euclidean Geometry.

Pág. 327 (340) - Chapter 7.

Spherical Geometry.

Obs.: Livro todo!!! Muito bom!!!

15) 20220614/

Symmetric Galerkin Boundary Element Method - Gray/

—

Gray - Symmetric Boundary Element Method (2008).pdf

Obs.: 298 pages. Ler com cuidado, pois foi marcado como "Uncorrected Proof"!

16) Source: 20220602

- Geometry of Continued Fractions/
- Handbook of the Mathematics of the Arts and Sciences/
- Ramanujan - 1729
- Livro:

a) Carl-Louis-Ferdinand von Lindemann (1852-1939)

b) Regard sur un mathématicien Indien Srinivasa Ramanujan (1887-1920)

Obs.: Todas as pastas acima foram copiadas para 20220614/

17) Source: 20220603/

Ramanujan's Notebooks - 5 volumes/

⇒

Copiada para a pasta 20220614/

18) Source: 20220604/

Obs.: Neste diretório tem tudo sobre
Eclipse-1919 (3 pastas

20220604/
Geometry (Holt Mc Dougal Larson Geometry) /

Copy: para 20220614/ ✡

Obs.: 1011 pages.

19) Source: 20220605 ✓

a) The Anomalous Magnetic Moment of the Muon/

a1) The Anomalous Magnetic Moment of the Muon (2017).pdf

Obs.: 704 pages.

Estudar conjuntamente com o Pi (π), mas o M.V.S. solucionará, assim espero, estas anomalias tanto na matemática quanto na física.

a2) Theory of the Muon Anomalous
Magnetic Moment.pdf

Obs.: 187 pages.

b) Scientif American - March2022/
Scientif-American-February-2022.pdf

Obs.: The Origins of Space and Time.

20) Source: 20220606/
Principia-Mathematica-Newton/

Obs.: Copy to 20220614/
Obra completa! Usar de referência para
comparar os desvios da luz
(light deflection) - Eddington 1919
calculados por Newton e Einstein.

⇒ Interessante! Procurar nestes originais.

21) Source: 20220607

a1) Calculus Heinbockel - Calculus - 2Volumes/

a2) Complex Variables, 6th Edition/

a3) Geometria - Baldor/

a4) The Metric Dirac Equation -
curve spacetime.

Copy to 20220614/

22) Source: 20220608

a1) Geometria - Castelnuovo/

a2) Isosurfaces Geometry, Topology, and Algorithms/

a3) Spazio e Tempo secondo le vedute di A. Einstein/

23) Source: 20220609/

a1) Fermat/

a2) Locally Mixed Symmetric Spaces

Obs: Copy to 20220614/

24) Source: 20220610/ ⇒ Copy to 20220614/

a1) Symmetry A Mathematical Exploration/
a2) Symmetry in Chemistry.../

25) Source: 20220611

The WSPC Handbook of Astronomical Instrumentation

FIM DESTE CADERNO 202206141742

Stenio Musich

Jandaia

Chapter 2

Chapter 2 - Part 1

Volume-01---20211021

!"#" $ % % & ' ( )) * +) , -)
.) / 0) * +)) 1 +) 2 !.! % 3 4

5- 6 26 3

7 # + -)
")% 8 -) 4 9)
-) - + " +)
4 9 8) % -) : 4 %
)9 +8 4)4) 4 ; + 8 ;) 4
)44 ;) -) +)) +8) % 4
) 8 4))) - 2 8 8
)44 9 < +38)) 8 9+)
- -) =) -)-) - 9))44)
))) + 9 " = + . ;) 4 8 4)
) + > 8)) 9+ ") ?@ A 2B8 /38
9 4))-)) - % =)) 9)
) 9+)) 4 +) 4) ,
@ A & ' B))+ -)9 4 9 + -)
)44 4) - 9)) + 1 % 8)) - 9) +
-) + 9) -
)))' -) - 9)))%))9)) 4 % -
%)9 +)) - 9) 4 9 - %) - 9))
) 9)) +) 4 9+ *) 8))) + -
)) $ 8) 4)' 9 C+
4 - 4 D % 9 = + 4)' +8))+
)))) - %) 4 9 9
)) 4 9 4 9))

E /< 8) 44 -) 4))
4 9)) =) - 9 %)9
8)4) 4))) - - 8)
) ;) - % =)) % +) -
)) +))9 % 4 4 ;
9 8 8) - .) 8) =) -) 9
) =) -) 8) -)) 8) =)
-) 9 % 9+)) 4)))) +
) 1 % 8))' -) 4 ;
)) -) - % =))) F 8
4))% 9 4) +)) 4 ;
4) 8 9+ =)) = + 9 4)
)) - - 9+))) %)
9) 9+) % -) -) 8) 9)9
;)))) -)
-) - !) ' -) 4)
4 ;) - 9) *) A G 4 ; 9 4) H 2E 7 3I D -
J 9)9 9)))
") -) + %
;) -) - 2 ;4) + 3 4) :
4)%))9))) %)9 + 4 9

*) K 4 ; 9 4)

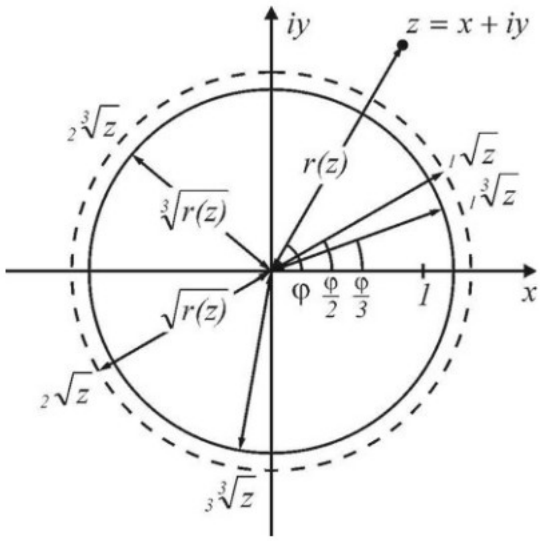

iy
z = x + iy
$_2\sqrt[3]{z}$
r(z)
$_1\sqrt{z}$
$_1\sqrt[3]{z}$
$\sqrt[3]{r(z)}$
φ
$\frac{\varphi}{2}$
$\frac{\varphi}{3}$
1
x
$\sqrt{r(z)}$
$_2\sqrt{z}$
$_3\sqrt[3]{z}$

E 7 @;) - *) A 4 ; 9 4)

!'#' $ % % & ') L '
"44 ;)) < M M M- M' ' 4

5- 2 /3

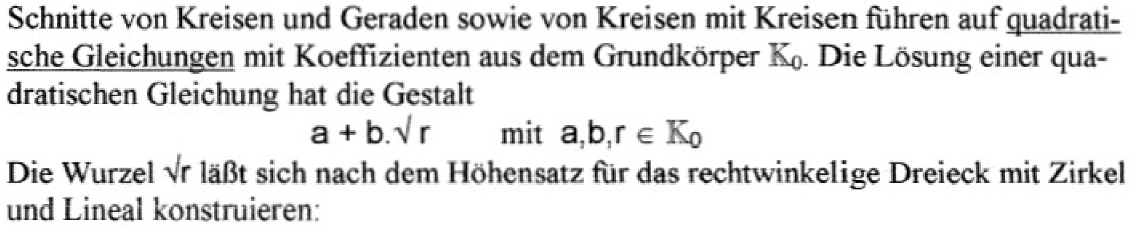

Schnitte von Kreisen und Geraden sowie von Kreisen mit Kreisen führen auf quadratische Gleichungen mit Koeffizienten aus dem Grundkörper $\mathbb{K}_0$. Die Lösung einer quadratischen Gleichung hat die Gestalt

$$a + b.\sqrt{r} \quad \text{mit } a,b,r \in \mathbb{K}_0$$

Die Wurzel $\sqrt{r}$ läßt sich nach dem Höhensatz für das rechtwinkelige Dreieck mit Zirkel und Lineal konstruieren:

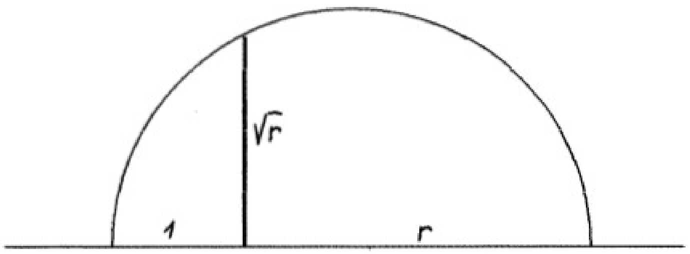

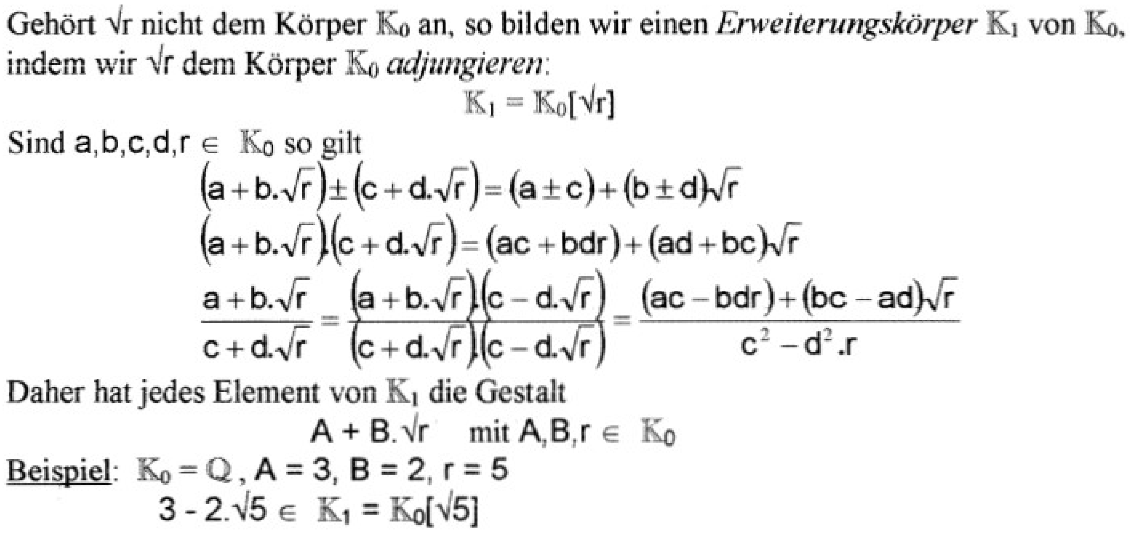

Gehört $\sqrt{r}$ nicht dem Körper $\mathbb{K}_0$ an, so bilden wir einen *Erweiterungskörper* $\mathbb{K}_1$ von $\mathbb{K}_0$, indem wir $\sqrt{r}$ dem Körper $\mathbb{K}_0$ *adjungieren*:

$$\mathbb{K}_1 = \mathbb{K}_0[\sqrt{r}]$$

Sind $a,b,c,d,r \in \mathbb{K}_0$ so gilt

$$\left(a + b.\sqrt{r}\right) \pm \left(c + d.\sqrt{r}\right) = \left(a \pm c\right) + \left(b \pm d\right)\sqrt{r}$$

$$\left(a + b.\sqrt{r}\right).\left(c + d.\sqrt{r}\right) = \left(ac + bdr\right) + \left(ad + bc\right)\sqrt{r}$$

$$\frac{a + b.\sqrt{r}}{c + d.\sqrt{r}} = \frac{\left(a + b.\sqrt{r}\right).\left(c - d.\sqrt{r}\right)}{\left(c + d.\sqrt{r}\right).\left(c - d.\sqrt{r}\right)} = \frac{\left(ac - bdr\right) + \left(bc - ad\right)\sqrt{r}}{c^2 - d^2.r}$$

Daher hat jedes Element von $\mathbb{K}_1$ die Gestalt

$$A + B.\sqrt{r} \quad \text{mit } A,B,r \in \mathbb{K}_0$$

Beispiel: $\mathbb{K}_0 = \mathbb{Q}$, $A = 3$, $B = 2$, $r = 5$

$$3 - 2.\sqrt{5} \in \mathbb{K}_1 = \mathbb{K}_0[\sqrt{5}]$$

!'#' $ % % & ' ( )) * +) , -)
.) " E 8 ,4)) ") + 2 !.! % 3 4

5- N2 3 (4 - 4)

O) 4 +) =) +8)+ P 8 ?%) A) % D)
) =) +8) ?% =) +A @) %)) 4 P % 8 - %)-
) . ;) 4 8)-) A)- P) 9 4 P))+8
7 -))) 9 - %) %) 4 9+) A 4)); " +
%) 9 4 4)) * %)-)
% 8)) 4) % # 8) ')
- - 4 4))- %)
4))) -) %

5- <72 63)

sense of rotation is preserved, while it is reversed for $\frac{d\mu}{dm} < 0$ Now $ad-bc$ is just the Jacobian determinant

$$\frac{d(\xi,\eta)}{d(x,y)} = \begin{vmatrix} \phi_x & \phi_y \\ \psi_x & \psi_y \end{vmatrix}$$

evaluated at the point P_0. We conclude that the mapping $\xi = \phi(x,y), \eta = \psi(x,y)$ *preserves or reverses orientations near the point* (x_0,y_0) *according to whether the Jacobian determinant at that point is positive or negative.*

10.12 Surfaces

As for curves, in most cases the parametric representation is found suitable for surfaces [5]. Since a surface is a two dimensional object, it requires two parameters to fix a point on it, as against one parameter required to fix a point on a curve. Thus, a parametric representation of a surface is given by parameterizing the position vector $\mathbf{x} \equiv (x,y,z)$ of a point on the surface,

$$\mathbf{x} = \mathbf{\Phi}(u,v) \equiv (x = \phi(u,v), y = \psi(u,v), z = \chi(u,v)) \tag{10.53}$$

where we assume the surface to be smooth, that is, $\mathbf{x} = \mathbf{\Phi}(u,v)$ is a continuously differentiable vector valued function or, equivalently, $x = \phi(u,v)$, $y = \psi(u,v)$, $z = \chi(u,v)$ are continuously differentiable scalar valued functions of two parameters (u,v). The point (u,v) ranges over some region $\mathcal{R}$ in the (u,v) plane. The corresponding point $\mathbf{x}(u,v) \equiv (x = \phi(u,v),\ y = \psi(u,v),\ z = \chi(u,v))$ ranges over a set in $\mathcal{E}_3$ or $\mathbb{R}^3$ spanning the surface. We can describe the surface in one of the three forms $z = f(x,y)$, $y = f(z,x)$, $x = f(y,z)$ by solving one of the three pairs of equations drawn out of $x = \phi(u,v), y = \psi(u,v), z = \chi(u,v)$. Solving any such pair of equations is equivalent to inverting the corresponding $\mathbb{R}^2 \mapsto \mathbb{R}^2$ map, say $(u,v) \mapsto (x = \phi(u,v),\ y = \psi(u,v))$ to express (u,v) as functions of (x,y) which can then be substituted in $z = \chi(u,v)$ to get $z = f(x,y)$. Thus, we require that not all of the three $\mathbb{R}^2 \mapsto \mathbb{R}^2$ maps corresponding to three pairs of equations be non-invertible, that is, we require that the three Jacobian determinants

$$\begin{vmatrix} \psi_u & \psi_v \\ \chi_u & \chi_v \end{vmatrix}, \begin{vmatrix} \chi_u & \chi_v \\ \phi_u & \phi_v \end{vmatrix}, \begin{vmatrix} \phi_u & \phi_v \\ \phi_u & \phi_v \end{vmatrix} \tag{10.54}$$

do not all vanish at once. We can summarize this condition in a single inequality

$$(\phi_u\psi_v - \phi_v\psi_u)^2 + (\psi_u\chi_v - \psi_v\chi_u)^2 + (\chi_u\phi_v - \chi_v\phi_u)^2 > 0 \tag{10.55}$$

If the inequality Eq. (10.55) is satisfied, in some neighbourhood of each point on the surface given by the $\mathbb{R}^2 \mapsto \mathbb{R}^3$ map in Eq. (10.53), it is certainly possible to express one of the three coordinates in terms of the other two.

At each point on the surface with parameters u, v we can partially differentiate the position vector to give

$$\mathbf{x}_u = (\phi_u, \psi_u, \chi_u) \quad \text{and} \quad \mathbf{x}_v = (\phi_v, \psi_v, \chi_v) \tag{10.56}$$

The differential of the vector $\mathbf{x}$ using the corresponding Jacobian, is given by

$$d\mathbf{x} = \begin{pmatrix} dx \\ dy \\ dz \end{pmatrix} = \begin{pmatrix} \phi_u & \phi_v \\ \psi_u & \psi_v \\ \chi_u & \chi_v \end{pmatrix} \begin{pmatrix} du \\ dv \end{pmatrix} = \mathbf{x}_u du + \mathbf{x}_v dv. \tag{10.57}$$

The three determinants Eq. (10.54) are just the components of the vector product $\mathbf{x}_u \times \mathbf{x}_v$. The expression on the left of the inequality in Eq. (10.55) is the square of the length of the vector $\mathbf{x}_u \times \mathbf{x}_v$ so that condition Eq. (10.55) is equivalent to

$$\mathbf{x}_u \times \mathbf{x}_v \neq \mathbf{0} \tag{10.58}$$

As an example, the spherical surface $x^2 + y^2 + z^2 = r^2$ of radius r is represented parametrically by the equations

$$x = r\cos u \sin v,\ \ y = r\sin u \sin v\ \ z = r\cos v,\ \ (0 \leq u < 2\pi,\ 0 \leq v \leq \pi) \tag{10.59}$$

where $v = \theta$ is the "polar inclination" or the polar angle and $u = \phi$ is the "longitude" or the azimuthal angle made by the point on the sphere. Note that the functions relating x, y, z to u, v are single valued and cover all the sphere. As v runs from $\pi/2$ to π the point x, y, z spans the lower hemisphere, that is,

$$z = -\sqrt{r^2 - x^2 - y^2}$$

while the values of v from 0 to $\pi/2$ give the upper hemisphere. Thus, for the parametric representation it is not necessary, as it is for the representation

$$z = \pm\sqrt{r^2 - x^2 - y^2},$$

to apply two single valued branches of the function in order to span the whole sphere.

We obtain another parametric representation of the sphere by means of *stereographic projection*. In order to project the sphere $x^2 + y^2 + z^2 - r^2 = 0$ stereographically from the north pole $(0, 0, r)$ on the equatorial plane $z = 0$, we join each point of the surface to the north pole N by a straight line and call the intersection of this line with the equatorial plane the *stereographic image* of the corresponding point of the sphere (see Fig. 10.5). We thus obtain a $1-1$ correspondence between the points of the sphere and the points of the plane, except for the north pole N. Using elementary geometry, we find that this correspondence is expressed by

$$x = \frac{2r^2u}{u^2+v^2+r^2},\ y = \frac{2r^2v}{u^2+v^2+r^2},\ z = \frac{(u^2+v^2-r^2)r}{u^2+v^2+r^2}, \tag{10.60}$$

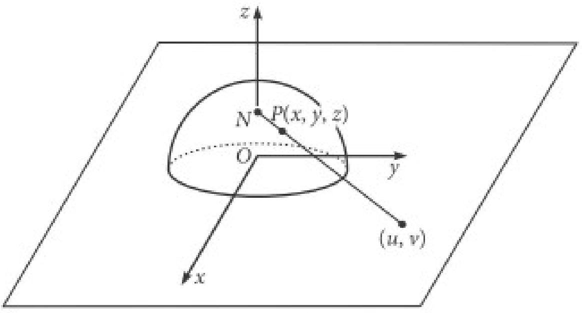

Fig. 10.5 Stereographic projection of the sphere

where (u,v) are the rectangular (cartesian) coordinates of the image point in the plane. These equations can be regarded as the parametric representation of the sphere, the parameters (u,v) being the rectangular coordinates in the u,v (equatorial) plane.

As a further example, we give parametric representation of surfaces

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}-\frac{z^2}{c^2}=1 \quad \text{and} \quad \frac{x^2}{a^2}+\frac{y^2}{b^2}-\frac{z^2}{c^2}=-1$$

called the *hyperboloid of one sheet* and the *hyperboloid of two sheets* respectively (see Fig. 10.6). The hyperboloid of one sheet is represented by

$$\begin{aligned} x &= a\cos u\cosh v, \\ y &= b\sin u\cosh v, \\ z &= c\sinh v \end{aligned} \tag{10.61}$$

where $0 \le u < 2\pi$; $-\infty < v < +\infty$ and the hyperboloid of two sheets by

$$\begin{aligned} x &= a\cos u\sinh v, \\ y &= b\sin u\sinh v, \\ z &= \pm c\cosh v \end{aligned} \tag{10.62}$$

where $0 \le u < 2\pi$; $0 < v < +\infty$.

In general, we may regard the *parametric representation* of a surface as the *mapping of the region $\mathcal{R}$ of the (u,v) plane onto the corresponding surface.* To each point of the region

projection. In order to project the sphere $x^2+y^2+z^2-r^2=0$ stereographically from the north pole $(0,0,r)$ on the equatorial plane $z=0$, we join each point of the surface to the north pole N by a straight line and call the intersection of this line with the equatorial plane the *stereographic image* of the corresponding point of the sphere (see Fig. 10.5). We thus obtain a $1-1$ correspondence between the points of the sphere and the points of the plane, except for the north pole N. Using elementary geometry, we find that this correspondence is expressed by

$\mathcal{R}$ of the (u,v) plane there corresponds one point of the surface and typically the converse is also true.[5]

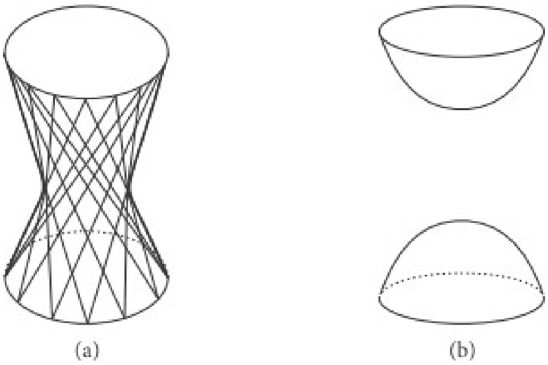

Fig. 10.6 (a) Hyperboloid of one sheet and (b) Hyperboloid of two sheets

Just as we can parameterize a surface by mapping a region in the u,v plane via Eq. (10.53), we can parameterize a curve on a surface by mapping an appropriate curve in the u,v plane onto the given curve on the surface. Thus, a curve $u = u(t)$, $v = v(t)$ in the u,v plane corresponds, by virtue of Eq. (10.53), to the curve

$$\mathbf{x}(t) = \mathbf{\Phi}(u(t),\, v(t)) \equiv (x(t) = \phi(u(t),\, v(t)),\, y(t) = \psi(u(t),\, v(t)),$$
$$z(t) = \chi(u(t), v(t))) \tag{10.63}$$

on the surface. Thus for example, the coordinate lines passing through a point on the sphere have the parametric equations $u = \phi =$ constant (longitudes) and $v = \theta =$ constant (latitudes). Corresponding curves in the u,v plane are the lines parallel to v and u axes respectively. The net of parametric curves (the mesh of latitudes and longitudes on the sphere) corresponds to the net of parallels to the axes in the u,v plane.

The tangent to the curve on the surface corresponding to the curve $u = u(t), v = v(t)$ in the u,v plane has the direction of the vector $\mathbf{x}_t = \frac{d\mathbf{x}}{dt}$, that is,

$$\mathbf{x}_t = (x_t, y_t, z_t) = \left(x_u\frac{du}{dt} + x_v\frac{dv}{dt}, y_u\frac{du}{dt} + y_v\frac{dv}{dt}, z_u\frac{du}{dt} + z_v\frac{dv}{dt}\right) = \mathbf{x}_u\frac{du}{dt} + \mathbf{x}_v\frac{dv}{dt}. \tag{10.64}$$

At a given point on the surface, the tangential vectors $\mathbf{x}_t$ of all curves on the surface passing through that point are linear combinations of two vectors $\mathbf{x}_u, \mathbf{x}_v$ which respectively are tangential to to the parametric lines $v =$ constant and $u =$ constant passing through that point. (e.g., the vectors $\hat{\phi}$ and $\hat{\theta}$ for the spherical polar coordinates on a sphere.) This means that the tangents all lie in the plane through the point *spanned* by the vectors $\mathbf{x}_u$ and $\mathbf{x}_v$, that is, the *tangent plane to the surface* at that point. The *normal* to the surface at that point is perpendicular to all tangential directions, in particular to the vectors $\mathbf{x}_u$ and $\mathbf{x}_v$. Thus, the surface normal is parallel (or antiparallel) to the direction of the vector product

---

[5] This is not always the case. For example, in the representation Eq. (10.59) of the sphere by spherical coordinates, the poles of the sphere correspond to the whole line segments given by $v = 0$ and $v = \pi$.

$$\mathbf{x}_u \times \mathbf{x}_v = (y_u z_v - y_v z_u, z_u x_v - z_v x_u, x_u y_v - x_v y_u). \tag{10.65}$$

One of the most important keys to the understanding of the given surface is the study of the curves that lie on it. Here, we give the expression for the arc length s of such a curve. We start with

$$\left(\frac{ds}{dt}\right)^2 = \left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2 = \mathbf{x}_t \cdot \mathbf{x}_t,$$

so in view of Eq. (10.64) we get

$$\begin{aligned}
\left(\frac{ds}{dt}\right)^2 &= \left(\mathbf{x}_u \frac{du}{dt} + \mathbf{x}_v \frac{dv}{dt}\right) \cdot \left(\mathbf{x}_u \frac{du}{dt} + \mathbf{x}_v \frac{dv}{dt}\right) \\
&= \left(x_u \frac{du}{dt} + x_v \frac{dv}{dt}\right)^2 + \left(y_u \frac{du}{dt} + y_v \frac{dv}{dt}\right)^2 + \left(z_u \frac{du}{dt} + z_v \frac{dv}{dt}\right)^2 \\
&= E\left(\frac{du}{dt}\right)^2 + 2F\left(\frac{du}{dt}\right)\left(\frac{dv}{dt}\right) + G\left(\frac{dv}{dt}\right)^2.
\end{aligned} \tag{10.66}$$

Here, the coefficients E, F, G, the *Gaussian fundamental quantities* of the surface, are given by

$$\begin{aligned}
E &= \left(\frac{\partial x}{\partial u}\right)^2 + \left(\frac{\partial y}{\partial u}\right)^2 + \left(\frac{\partial z}{\partial u}\right)^2 = \mathbf{x}_u \cdot \mathbf{x}_u \\
F &= \frac{\partial x}{\partial u}\frac{\partial x}{\partial v} + \frac{\partial y}{\partial u}\frac{\partial y}{\partial v} + \frac{\partial z}{\partial u}\frac{\partial z}{\partial v} = \mathbf{x}_u \cdot \mathbf{x}_v \\
G &= \left(\frac{\partial x}{\partial v}\right)^2 + \left(\frac{\partial y}{\partial v}\right)^2 + \left(\frac{\partial z}{\partial v}\right)^2 = \mathbf{x}_v \cdot \mathbf{x}_v
\end{aligned} \tag{10.67}$$

These depend only on $\mathbf{x}_u, \mathbf{x}_v$ and therefore on the surface and its parametric representation and not on the particular choice of the curve on the surface. The expression Eq. (10.66) for the derivative of the length of arc s with respect to the parameter t usually is written symbolically without reference to the parameter used along the curve. One says that the *line element* ds is given by the quadratic differential form ("fundamental form")

$$ds^2 = Edu^2 + 2Fdudv + Gdv^2. \tag{10.68}$$

The length of the cross product $\mathbf{x}_u \times \mathbf{x}_v$ can be expressed in terms of E, F, G as

$$|\mathbf{x}_u \times \mathbf{x}_v|^2 = |\mathbf{x}_u|^2|\mathbf{x}_v|^2 - (\mathbf{x}_u \cdot \mathbf{x}_v)^2 = EG - F^2. \tag{10.69}$$

Our original condition on the parametric representation (inequality Eq. (10.55)) can now be formulated as the condition

$$EG - F^2 > 0 \tag{10.70}$$

for the fundamental quantities.

The direction cosines for one of the two normals to the surface are the components of the unit vector

$$\frac{1}{|\mathbf{x}_u \times \mathbf{x}_v|}\mathbf{x}_u \times \mathbf{x}_v = \frac{1}{\sqrt{EG - F^2}}\mathbf{x}_u \times \mathbf{x}_v.$$

It follows from Eq. (10.65) that the normal to a surface represented parametrically has the direction cosines

$$\cos\alpha = \frac{y_u z_v - y_v z_u}{\sqrt{EG - F^2}},\ \cos\beta = \frac{z_u x_v - z_v x_u}{\sqrt{EG - F^2}},\ \cos\gamma = \frac{x_u y_v - x_v y_u}{\sqrt{EG - F^2}}. \tag{10.71}$$

The tangent to a curve $u = u(t), v = v(t)$ on the surface has the direction of the vector

$$\mathbf{x}_t = \mathbf{x}_u \frac{du}{dt} + \mathbf{x}_v \frac{dv}{dt}.$$

If we now consider a second curve, $u = u(\tau), v = v(\tau)$ on the surface referred to a parameter τ, its tangent has the direction of the vector

$$\mathbf{x}_\tau = \mathbf{x}_u \frac{du}{d\tau} + \mathbf{x}_v \frac{dv}{d\tau}.$$

If the two curves pass through the same point on the surface, the cosine of the angle of intersection ω is the same as the cosine of the angle between $\mathbf{x}_t$ and $\mathbf{x}_\tau$. Hence,

$$\cos\omega = \frac{\mathbf{x}_t \cdot \mathbf{x}_\tau}{|\mathbf{x}_t||\mathbf{x}_\tau|}.$$

We have,

$$\begin{aligned}\mathbf{x}_t\cdot\mathbf{x}_\tau &= \left(\mathbf{x}_u\frac{du}{dt}+\mathbf{x}_v\frac{dv}{dt}\right)\cdot\left(\mathbf{x}_u\frac{du}{d\tau}+\mathbf{x}_v\frac{dv}{d\tau}\right)\\ &= E\frac{du}{dt}\frac{du}{d\tau}+F\left(\frac{du}{dt}\frac{dv}{d\tau}+\frac{du}{d\tau}\frac{dv}{dt}\right)+G\frac{dv}{dt}\frac{dv}{d\tau}. \qquad (10.72)\end{aligned}$$

Consequently, the cosine of the angle between two curves on the surface is given by

$$\cos\omega=\frac{E\frac{du}{dt}\frac{du}{d\tau}+F\left(\frac{du}{dt}\frac{dv}{d\tau}+\frac{du}{d\tau}\frac{dv}{dt}\right)+G\frac{dv}{dt}\frac{dv}{d\tau}}{\sqrt{E\left(\frac{du}{dt}\right)^2+2F\left(\frac{du}{dt}\right)\left(\frac{dv}{dt}\right)+G\left(\frac{dv}{dt}\right)^2}\sqrt{E\left(\frac{du}{d\tau}\right)^2+2F\left(\frac{du}{d\tau}\right)\left(\frac{dv}{d\tau}\right)+G\left(\frac{dv}{d\tau}\right)^2}} \qquad (10.73)$$

We end this subsection by giving one more example of parametrization of a surface which comes up frequently in applications. We consider torus. This is obtained by rotating a circle about a line which lies in the plane of the circle, but does not intersect with it (see Fig. 10.7). We take the axis of rotation as the z-axis and choose the y-axis so as to pass through the center of the circle, whose y-coordinate we denote by a. If the radius of the circle is $r < |a|$, we obtain

$$x=0,\ y-a=r\cos\theta,\ z=r\sin\theta(0\le\theta<2\pi)$$

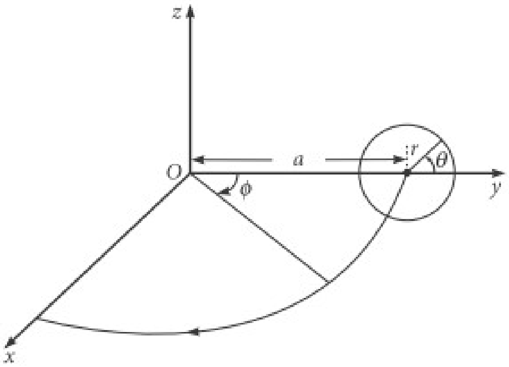

Fig. 10.7 Creation of torus by the rotation of a circle

as a parametric representation of the circle in the $y-z$ plane. Now letting the circle rotate about the z-axis, we find that for each point on the circle x^2+y^2 remains constant; that is, $x^2+y^2=(a+r\cos\theta)^2$. If ϕ is the angle of rotation about the z-axis, we have

$$x \;=\; (a+r\cos\theta)\sin\phi,$$

$$
\begin{aligned}
y &= (a + r\cos\theta)\cos\phi, \\
z &= r\sin\theta
\end{aligned}
\tag{10.74}
$$

with $0 \le \phi < 2\pi, 0 \le \theta < 2\pi$ as a parametric representation of the torus in terms of the parameters θ and ϕ. In this representation the torus appears as the image of the square of side 2π in the θ, ϕ plane. Any pair of boundary points of this square lying on the same line $\theta =$ constant or $\phi =$ constant corresponds to only one point on the surface and the four corners of the square all correspond to the same point on the surface.

Equation (10.67) gives, for the line element on the torus,

$$ds^2 = r^2 d\theta^2 + (a + r\cos\theta)^2 d\phi^2.$$

10.13 The Divergence and the Curl of a Vector Field

We have already seen that the del or the grad operator has vector like structure and it also transforms like a vector under the rotation and translation of the coordinate system. This enables us to formally treat the del operator like a vector with components $\left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \frac{\partial}{\partial x_3}\right)$ and define its scalar and vector products with vector valued functions possibly giving a vector field.

The divergence

The corresponding scalar product called the divergence of a field is given by

$$\nabla \cdot \mathbf{f}(\mathbf{x}) = \frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2} + \frac{\partial f_3}{\partial x_3} = \sum_{k=1}^{3} \frac{\partial f_k}{\partial x_k} \tag{10.75}$$

where $f_{1,2,3}(\mathbf{x})$ are the scalar valued component functions of the vector valued function $\mathbf{f}(\mathbf{x})$ with respect to some orthonormal basis (see Eq. (10.1)). If we fix a position vector $\mathbf{x}$, then we get the corresponding vector $\mathbf{f}(\mathbf{x})$ giving us the unique value of the divergence $\nabla \cdot \mathbf{f}(\mathbf{x})$. Thus, the divergence of a vector field is itself a scalar field and we can calculate 'the divergence at a point'. The value of the divergence of a vector field at a point is a measure of how much a vector $\mathbf{f}(\mathbf{x})$ spreads out from (or flows into) the point $\mathbf{x}$ in question. Thus, the vector function in Fig. 10.8(a) has large positive divergence (if the arrows pointed inward it would be a large *negative* divergence), the function in Fig. 10.8(b) has zero divergence and Fig. 10.8(c) again shows a function of positive divergence. Here is a nice possible observation of the divergence phenomenon [9]. Imagine standing at the edge of a pond. Sprinkle some sawdust on the surface. If the material spreads out then you have dropped it at a point of positive divergence; if it collects together, you have dropped it at a point of negative divergence. The vector function $\mathbf{v}$ in this model is the velocity of the water. This is a two dimensional example but it helps give us a feel for the meaning of divergence. A point of positive divergence is a sourse or 'foucet'; a point of negative divergence is a sink or 'drain'.

Exercise In what follows $\mathbf{r}$ denotes a position vector $r = |\mathbf{r}|$ is its magnitude, $\mathbf{A}(\mathbf{r})$ and $\mathbf{B}(\mathbf{r})$ are vector fields, $\phi(\mathbf{r})$ is a scalar field and $f(r)$ is a function of r. All fields and functions have continuous first derivatives. Using Levi-Civita symbols or otherwise, prove the following.

(i) $\nabla \times (\nabla \times \mathbf{A}) = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A}$.

(ii) $\mathbf{A} \times (\nabla \times \mathbf{B}) = \nabla_B(\mathbf{A} \cdot \mathbf{B}) - (\mathbf{A} \cdot \nabla)\mathbf{B}$ where ∇_B operates on $\mathbf{B}$ only.

(iii) Given $\nabla \times \mathbf{A} = \mathbf{0} = \nabla \times \mathbf{B}$ show that $\nabla \cdot (\mathbf{A} \times \mathbf{B}) = 0$.

(iv) For constant $\mathbf{a}$ and $\mathbf{b}$ show that $\nabla \times [(\mathbf{a} \times \mathbf{r}) \times \mathbf{b}] = \mathbf{a} \times \mathbf{b}$.

(v) $\nabla \cdot \mathbf{r} = 3$, $\nabla \times \mathbf{r} = \mathbf{0}$, $\nabla(\mathbf{A} \cdot \mathbf{r}) = \mathbf{A}$, $(\mathbf{A} \cdot \nabla)\mathbf{r} = \mathbf{A}$.

(vi) $\nabla r^n = n r^{n-2}\mathbf{r}$, $\nabla \cdot \frac{f(r)\mathbf{r}}{r} = \frac{1}{r^2}\frac{d}{dr}(r^2 f)$.

We will use any one or more of these results in the sequel, as and when required. □

Exercise A particle performs uniform circular motion on a circle of radius r and position vector $\mathbf{r}$. Show that (a) $\nabla \times \mathbf{v} = 2\boldsymbol{\omega}$ and (b) $\nabla \cdot \mathbf{v} = 0$, where $\mathbf{v}$ is the linear velocity and ω is the (constant) rotational velocity of the particle.

Solution

(a) We know that for circular motion, $\mathbf{v} = \boldsymbol{\omega} \times \mathbf{r}$. Therefore,

$$\nabla \times \mathbf{v} = \nabla \times (\boldsymbol{\omega} \times \mathbf{r}) = \boldsymbol{\omega}(\nabla \cdot \mathbf{r}) - \boldsymbol{\omega} \cdot \nabla \mathbf{r}.$$

However, $\nabla \cdot \mathbf{r} = 3$ and $\boldsymbol{\omega} \cdot \nabla \mathbf{r} = \boldsymbol{\omega}$, which gives (a). Thus, we see that the curl operator transforms the velocity vector into a rotational velocity vector.

(b) $\nabla \cdot \mathbf{v} = \nabla \cdot (\boldsymbol{\omega} \times \mathbf{r}) = \mathbf{r} \cdot \nabla \times \boldsymbol{\omega} - \boldsymbol{\omega} \cdot \nabla \times \mathbf{r} = 0$, since $\boldsymbol{\omega}$ is a constant vector and $\nabla \times \mathbf{r} = \mathbf{0}$. □

Thus, there are basically two types of second derivatives, the Laplacian, which is of fundamental importance and the gradient of divergence which we seldom encounter. Since the second derivatives suffice to deal with practically all the physical applications, going over to higher derivatives will reduce to an academic exercise without any physical motivation.

10.14 Differential Operators in Curvilinear Coordinates

A system of curvilinear coordinates, say (u, v, w), is specified via an invertible passive transformation $\mathbb{R}^3 \mapsto \mathbb{R}^3$:

$$(x, y, z) \mapsto (u(x,y,z), v(x,y,z), w(x,y,z)) \tag{10.93}$$

or via the inverse $\mathbb{R}^3 \mapsto \mathbb{R}^3$ transformation

$$(u, v, w) \mapsto (x(u,v,w), y(u,v,w), z(u,v,w)). \tag{10.94}$$

A system of curvilinear coordinates is said to be orthogonal if the coordinate surfaces or equivalently the coordinate lines or the unit vectors tangent to the coordinate lines at their point of intersection are mutually perpendicular (see Fig. 10.11). Note that there are different sets of coordinate surfaces, coordinate lines and the orthogonal basis vectors $\hat{\mathbf{u}}, \hat{\mathbf{v}}, \hat{\mathbf{w}}$ at different points in space. The transformations in Eqs (10.93), (10.94) are assumed to be C^1, that is, having continuous partial derivatives at all points in some region of space, as well as invertible everywhere in that region, that is, having a non-zero Jacobian determinant at all points in that region. Such transformations between two sets of coordinates are called *admissible* coordinate transformations defined over a given region of space. Just to refresh our memory, we recall the definitions of the coordinate lines and coordinate surfaces. Given any point, one can draw a curve passing through the point in such a way that only one of the three curvilinear coordinates changes along the curve while the values of the other two coordinates remain constant. For three curvilinear coordinates one can draw three such curves, all passing through the given point and mutually intersecting at right angles. Each is called a *coordinate curve* or a *coordinate line* and a surface passing through the given point on which a particular curvilinear coordinate has a constant value is called a *coordinate surface.* The u-coordinate line, for example, is the curve of intersection of the v and the w coordinate surfaces. All this is depicted in Fig. 10.11.

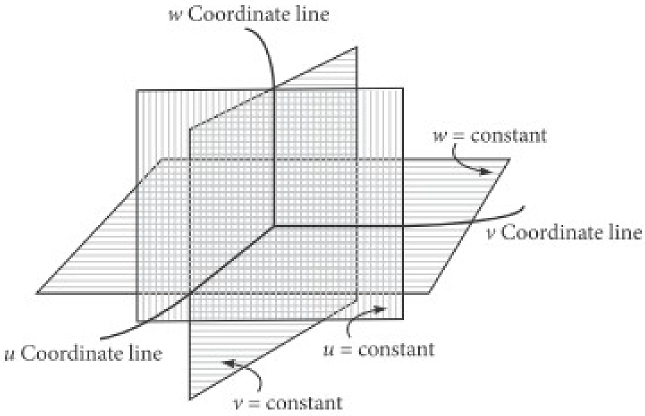

Fig. 10.11 The Network of coordinate lines and coordinate surfaces at any arbitary point, defining a curvilinear coordinate system

Let the equations for the coordinate curves at a point (u_0, v_0, w_0) be

$$\mathbf{x} = \mathbf{x}(u, v_0, w_0) \;\; \mathbf{x} = \mathbf{x}(u_0, v, w_0) \;\; \mathbf{x} = \mathbf{x}(u_0, v_0, w).$$

The tangents to the coordinate curves are given by the vectors

$$\frac{\partial \mathbf{x}(u, v_0, w_0)}{\partial u}, \frac{\partial \mathbf{x}(u_0, v, w_0)}{\partial v}, \frac{\partial \mathbf{x}(u_0, v_0, w)}{\partial w}$$

respectively. Orthogonality of these vectors requires that

$$\frac{\partial \mathbf{x}}{\partial u} \cdot \frac{\partial \mathbf{x}}{\partial v} = \frac{\partial \mathbf{x}}{\partial u} \cdot \frac{\partial \mathbf{x}}{\partial w} = \frac{\partial \mathbf{x}}{\partial v} \cdot \frac{\partial \mathbf{x}}{\partial w} = 0.$$

We are interested in the differential displacement $d\mathbf{x}$ as we go from $\mathbf{x}(u,v,w)$ to $\mathbf{x}(u+du, v+dv, w+dw)$. We have, in terms of the corresponding Jacobian matrix,

$$d\mathbf{x} = \begin{bmatrix} \frac{\partial s_1}{\partial u} & \frac{\partial s_1}{\partial v} & \frac{\partial s_1}{\partial w} \\ \frac{\partial s_2}{\partial u} & \frac{\partial s_2}{\partial v} & \frac{\partial s_2}{\partial w} \\ \frac{\partial s_3}{\partial u} & \frac{\partial s_3}{\partial v} & \frac{\partial s_3}{\partial w} \end{bmatrix} \begin{bmatrix} du \\ dv \\ dw \end{bmatrix} = \frac{\partial \mathbf{x}}{\partial u} du + \frac{\partial \mathbf{x}}{\partial u} dv + \frac{\partial \mathbf{x}}{\partial w} dw = \mathbf{x}_u du + \mathbf{x}_v dv + \mathbf{x}_w dw. \tag{10.95}$$

where $s_{1,2,3}(u,v,w) = \mathbf{x}(u,v,w) \cdot \{\hat{\mathbf{u}}, \hat{\mathbf{v}}, \hat{\mathbf{w}}\}$ are the components of $\mathbf{x}$ in the $\hat{\mathbf{u}}, \hat{\mathbf{v}}, \hat{\mathbf{w}}$ mutually orthogonal directions. This defines the line element $ds = |d\mathbf{x}|$ via

$$\begin{aligned} ds^2 = d\mathbf{x} \cdot d\mathbf{x} &= \mathbf{x}_u \cdot \mathbf{x}_u du^2 + \mathbf{x}_v \cdot \mathbf{x}_v dv^2 + \mathbf{x}_w \cdot \mathbf{x}_w dw^2 \\ &= h_1^2 du^2 + h_2^2 dv^2 + h_3^2 dw^2. \end{aligned} \tag{10.96}$$

The parameters h_1, h_2, h_3 are the analogues of the Gaussian fundamental quantities E, F, G of a surface, we derived before. They relate the differential displacements along the $\hat{\mathbf{u}}, \hat{\mathbf{v}}, \hat{\mathbf{w}}$ directions as a result of the displacement $d\mathbf{x}$ to the diferentials du, dv, dw via

$$d\mathbf{x} = ds_1 \hat{\mathbf{u}} + ds_2 \hat{\mathbf{v}} + ds_3 \hat{\mathbf{w}} = h_1 du \hat{\mathbf{u}} + h_2 dv \hat{\mathbf{v}} + h_3 dw \hat{\mathbf{w}}.$$

Thus, the volume of the rectangular parallelepiped with sides ds_1, ds_2, ds_3 is given by

$$dV = ds_1 ds_2 ds_3 = h_1 h_2 h_3 du dv dw.$$

The product $h_1 h_2 h_3$ ensures that the the last term has the dimension of volume; as the curvilinear coordinates can be dimensionless quantities like angles.

The u coordinate surface passing through the point (u_0, v_0, w_0) is the collection of points (x, y, z) satisfying $u(x, y, z) = u_0$ and similarly, the v and w coordinate surfaces are given by $v(x, y, z) = v_0$ and $w(x, y, z) = w_0$, where (x, y, z) are the Cartesian coordinates with respect to some rectangular Cartesian coordinate system. We can vary the point (u_0, v_0, w_0) over the region for which the curvilinear coordinate transformations, Eqs (10.93), (10.94), are defined. Therefore, we can replace u_0, v_0, w_0 in the equations defining the coordinate surfaces by u, v, w and say that a particular triad of coordinate surfaces emerges when particular values of u, v and w are substituted on the RHS of these equations. Thus, we write, for the equations defining the coordinate surfaces

$$u = u(x, y, z) \quad v = v(x, y, z) \quad w = w(x, y, z).$$

The normals to the coordinate surfaces are given by $\nabla u, \nabla v, \nabla w$ (see section 10.8) which, owing to orthogonality, must satisfy

$$0 = \nabla u \cdot \nabla v = \nabla u \cdot \nabla w = \nabla v \cdot \nabla w.$$

The vectors normal to the coordinate surfaces are tangent to the corresponding coordinate curves so that we can *define* the fundamental triad for the curvilinear coordinates as

$$\hat{\mathbf{u}} = \frac{\nabla u}{|\nabla u|} \quad \hat{\mathbf{v}} = \frac{\nabla v}{|\nabla v|} \quad \hat{\mathbf{w}} = \frac{\nabla w}{|\nabla w|}. \tag{10.97}$$

Let $d\mathbf{s}_1 = ds_1\hat{\mathbf{u}}$, $d\mathbf{s}_2 = ds_2\hat{\mathbf{v}}$, $d\mathbf{s}_3 = ds_3\hat{\mathbf{w}}$ be the differential displacement along the $\hat{\mathbf{u}}, \hat{\mathbf{v}}, \hat{\mathbf{w}}$ directions. Since $\nabla u, \nabla v, \nabla w$ have the same values in all the orthonormal basis triads and since $\nabla u, d\mathbf{s}_1$, $\nabla v, d\mathbf{s}_2$ and $\nabla w, d\mathbf{s}_3$ are the pairs of parallel vectors, we can write,

$$\begin{aligned} du &= \nabla u \cdot d\mathbf{s}_1 = |\nabla u||d\mathbf{s}_1|, \\ dv &= \nabla v \cdot d\mathbf{s}_2 = |\nabla v||d\mathbf{s}_2|, \\ dw &= \nabla w \cdot d\mathbf{s}_3 = |\nabla w||d\mathbf{s}_3|. \end{aligned} \tag{10.98}$$

This gives,

$$ds^2 = ds_1^2 + ds_2^2 + ds_3^2 = \frac{du^2}{|\nabla u|^2} + \frac{dv^2}{|\nabla v|^2} + \frac{dw^2}{|\nabla w|^2}. \tag{10.99}$$

Comparing equations Eqs (10.96) and (10.99) we get

$$\begin{aligned} h_1 &= \frac{1}{|\nabla u|} = \sqrt{\mathbf{x}_u \cdot \mathbf{x}_u} \text{ impling } \hat{\mathbf{u}} = \mathrm{h}_1 \nabla \mathrm{u}, \\ h_2 &= \frac{1}{|\nabla v|} = \sqrt{\mathbf{x}_v \cdot \mathbf{x}_v} \text{ impling } \hat{\mathbf{v}} = \mathrm{h}_2 \nabla \mathrm{v}, \\ h_3 &= \frac{1}{|\nabla w|} = \sqrt{\mathbf{x}_w \cdot \mathbf{x}_w} \text{ impling } \hat{\mathbf{w}} = \mathrm{h}_3 \nabla \mathrm{w}. \end{aligned} \tag{10.100}$$

Example For spherical polar coordinates we identify $u = r$, $v = \theta$ and $w = \phi$, where

$$r = \sqrt{x^2 + y^2 + z^2},\ \theta = \cos^{-1}\left(\frac{z}{\sqrt{x^2 + y^2 + z^2}}\right) \text{ and } \phi = \tan^{-1}\left(\frac{\mathrm{y}}{\mathrm{x}}\right).$$

This gives $h_1^{-1} = |\nabla r| = 1$, $h_2^{-1} = |\nabla \theta| = r^{-1}$ and $\mathrm{h}_3^{-1} = |\nabla \phi| = (\mathrm{r}\sin\theta)^{-1}$. Therefore,

$$ds^2 = h_1^2 dr^2 + h_2^2 d\theta^2 + h_3^2 d\phi^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2.$$

Also, the fundamental triad is,

$$\hat{\mathbf{u}} = \hat{\mathbf{r}}, \ \ \hat{\mathbf{v}} = \hat{\theta}, \ \ \hat{\mathbf{w}} = \hat{\phi}.$$

We can also invert the transformation to get

$$x = r\sin\theta\cos\phi \ \ y = r\sin\theta\sin\phi \ \ z = r\cos\theta.$$

This gives (Exercise),

$$\begin{aligned} h_1 &= \sqrt{\left(\frac{dx}{dr}\right)^2 + \left(\frac{dy}{dr}\right)^2 + \left(\frac{dz}{dr}\right)^2} = 1, \\ h_2 &= \sqrt{\left(\frac{dx}{d\theta}\right)^2 + \left(\frac{dy}{d\theta}\right)^2 + \left(\frac{dz}{d\theta}\right)^2} = r, \\ h_3 &= \sqrt{\left(\frac{dx}{d\phi}\right)^2 + \left(\frac{dy}{d\phi}\right)^2 + \left(\frac{dz}{d\phi}\right)^2} = r\sin\theta. \end{aligned} \tag{10.101}$$

which are identical to those already obtained from forward transformations.

The gradient of a scalar valued function $\phi(u,v,w)$ with respect to the Cartesian coordinates x,y,z can be obtained by applying chain rule:

$$\begin{aligned} \nabla\phi &= \frac{\partial\phi}{\partial u}\nabla u + \frac{\partial\phi}{\partial v}\nabla v + \frac{\partial\phi}{\partial w}\nabla w \\ &= \left(\frac{1}{h_1}\frac{\partial\phi}{\partial u}\right)\hat{\mathbf{u}} + \left(\frac{1}{h_2}\frac{\partial\phi}{\partial v}\right)\hat{\mathbf{v}} + \left(\frac{1}{h_3}\frac{\partial\phi}{\partial w}\right)\hat{\mathbf{w}}, \end{aligned} \tag{10.102}$$

where we have used Eq. (10.100).

Our next job is to express the divergence $\nabla \cdot \mathbf{f}$ of a vector field $\mathbf{f}$ exclusively in terms of the derivatives with respect to the curvilinear coordinates u,v,w. We can do this, in principle, by a systematic application of the chain rule, but that will make us go through a lengthy algebra which only masochists can enjoy. There is a short cut, where we use the fact that the flux of a vector field through a differential volume dV is given by $\nabla \cdot \mathbf{f} dV$. In order to get hold of this quantity, we need to know how to represent the area of a piece of a surface by a vector. As shown in Fig. 10.12, an element of area da is represented by the vector $d\mathbf{a}$ given by

$$d\mathbf{a} = \hat{\mathbf{n}} da = da_x \hat{\mathbf{i}} + da_y \hat{\mathbf{j}} + da_z \hat{\mathbf{k}} \tag{10.103}$$

where $\hat{\mathbf{n}}$ is the unit outward normal to the surface, defined via its direction cosines

$$\hat{\mathbf{n}} = \cos\alpha\hat{\mathbf{i}} + \cos\beta\hat{\mathbf{j}} + \cos\gamma\hat{\mathbf{k}}. \tag{10.104}$$

Therefore, the components of $d\mathbf{a}$ are given by

$$da_x = da\cos\alpha, \;\; da_y = da\cos\beta \;\; da_z = da\cos\gamma.$$

In particular, $\mathbf{x}\cdot d\mathbf{a} = |\mathbf{x}|da\cos\theta$ where θ is the angle between $\hat{\mathbf{n}}$ and $\mathbf{x}$. If $\theta = 0$ or $\theta = \pi$ then $\mathbf{x}\cdot d\mathbf{a} = \pm|\mathbf{x}|da$. This situation arises when the surface is perpendicular to $\mathbf{x}$, that is, the area of the projection of the surface element on the plane perpendicular to $\mathbf{x}$ is the same as da. The flux of $\mathbf{f}$ through an element of area $d\mathbf{a}$ is $\mathbf{f}\cdot d\mathbf{a}$.

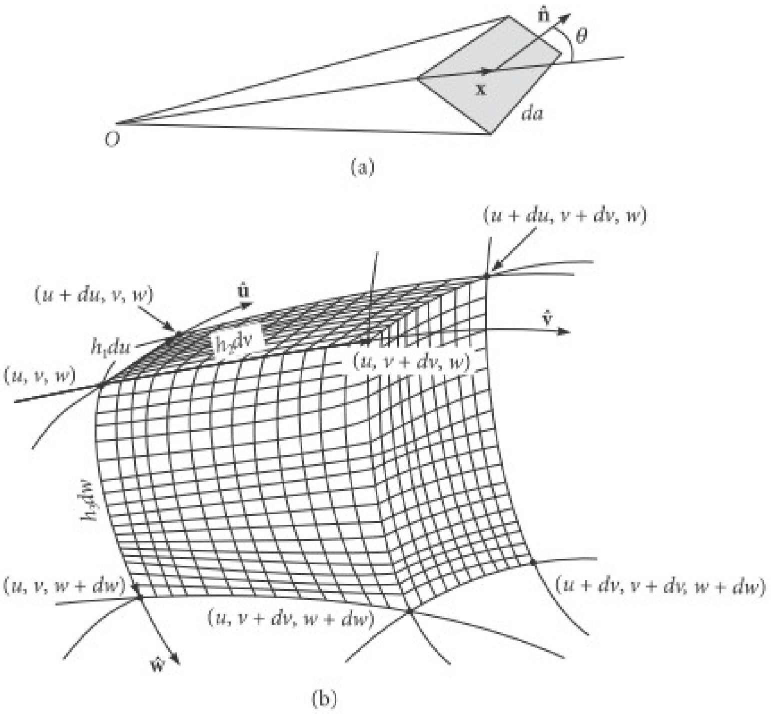

(a)

(b)

Fig. 10.12 (a) Evaluating $\mathbf{x}\cdot d\mathbf{a}$ (b) Flux through the opposite faces of a volume element

Let us now consider a differential volume of the shape of a rectangular parallelepiped with sides ds_1, ds_2, ds_3 defined above so that its volume is $dV = ds_1 ds_2 ds_3 = h_1 h_2 h_3 \, du dv dw$. Let us assume that the pairs $-\hat{\mathbf{u}}, \hat{\mathbf{u}}$ be the outward normals to the front and the back sides, $-\hat{\mathbf{v}}, \hat{\mathbf{v}}$ be the outward normals to the left and the right sides and $-\hat{\mathbf{w}}, \hat{\mathbf{w}}$ be the outward normals to the bottom and the top sides of the box. Then for the front face, $d\mathbf{a} = -h_2 h_3 dv dw \hat{\mathbf{u}}$ and $\mathbf{f}\cdot d\mathbf{a} = -(h_2 h_3 f_1) dv dw$, where $f_{1,2,3}$ are the components of $\mathbf{f}$

along $\hat{\mathbf{u}}, \hat{\mathbf{v}}, \hat{\mathbf{w}}$ respectively and the product $h_2 h_3 f_1$ is to be evaluated at u. On the back face, product $h_2 h_3 f_1$ is to be evaluated at $u + du$ so that $\mathbf{f} \cdot d\mathbf{a} = (h_2 h_3 f_1 + \frac{\partial}{\partial u}(h_2 h_3 f_1) du) dv dw$. Therefore, the net flux through the front and back pair of faces is

$$\left[\frac{\partial}{\partial u}(h_2 h_3 f_1)\right] du dv dw = \frac{1}{h_1 h_2 h_3} \frac{\partial}{\partial u}(h_2 h_3 f_1) dV.$$

In the same way, the right and the left sides give

$$\frac{1}{h_1 h_2 h_3} \frac{\partial}{\partial v}(h_1 h_3 f_2) dV$$

and the bottom and the top sides contribute

$$\frac{1}{h_1 h_2 h_3} \frac{\partial}{\partial w}(h_1 h_2 f_3) dV.$$

Thus, the total flux through the box is given by

$$(\nabla \cdot \mathbf{f}) dV = \frac{1}{h_1 h_2 h_3}\left[\frac{\partial}{\partial u}(h_2 h_3 f_1) + \frac{\partial}{\partial v}(h_1 h_3 f_2) + \frac{\partial}{\partial w}(h_1 h_2 f_3)\right] dV$$

This gives

$$\nabla \cdot \mathbf{f} = \frac{1}{h_1 h_2 h_3}\left[\frac{\partial}{\partial u}(h_2 h_3 f_1) + \frac{\partial}{\partial v}(h_3 h_1 f_2) + \frac{\partial}{\partial w}(h_1 h_2 f_3)\right]. \tag{10.105}$$

Combining Eq. (10.105) with Eq. (10.102) we get, for the Laplacian operator,

$$\nabla^2 \phi = \nabla \cdot \nabla \phi = \frac{1}{h_1 h_2 h_3}\left[\frac{\partial}{\partial u}\left(\frac{h_2 h_3}{h_1}\frac{\partial \phi}{\partial u}\right) + \frac{\partial}{\partial v}\left(\frac{h_3 h_1}{h_2}\frac{\partial \phi}{\partial v}\right) + \frac{\partial}{\partial w}\left(\frac{h_1 h_2}{h_3}\frac{\partial \phi}{\partial w}\right)\right]. \tag{10.106}$$

Our last task in this subsection is to express the curl $\nabla \times \mathbf{f}$ of a vector field $\mathbf{f}$ in terms of the derivatives with respect to the curvilinear coordinates u, v, w. The principle we follow for this is

Circulation of $\mathbf{f}$ around a loop enclosing an infinitesimal area $d\mathbf{a} = (\nabla \times \mathbf{f}) \cdot d\mathbf{a}$.

The sense of circulation is given by that which makes a right handed screw advance in the direction of $d\mathbf{a}$. The required circulation can be explicitly calculated for an infinitesimal loop of rectangular shape. For each side of the rectangle, we have to find the scalar product of $\mathbf{f}$ with the vector along the side and in the direction consistent with the sense of circulation. In the first place, the surface enclosed by an infinitesimal loop can be taken to be a plane. Consider such a rectangular loop in the $\hat{\mathbf{u}}, \hat{\mathbf{v}}$ plane, with $\hat{\mathbf{w}}$ normal to it (see Fig. 10.13). From Fig. 10.13 and $\hat{\mathbf{w}}$ pointing out of the page, it is clear that the sense of circulation which makes a right handed screw advance in $\hat{\mathbf{w}}$ direction is counterclockwise, as shown. The vector on the side along $\hat{\mathbf{u}}$ is $d\mathbf{s}_1 = h_1 du\hat{\mathbf{u}}$ that on the side along $\hat{\mathbf{v}}$ is $d\mathbf{s}_2 = h_2 dv\hat{\mathbf{v}}$ and area

$$d\mathbf{a} = h_1 h_2 du dv \hat{\mathbf{w}}.$$

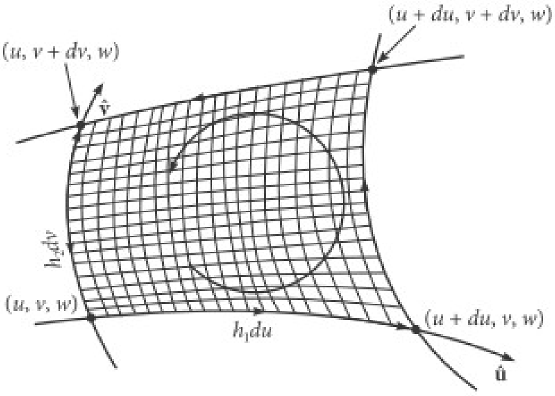

Fig. 10.13 Circulation around a loop

Along the bottom side (along $\hat{\mathbf{u}}$) the contribution of $\mathbf{f}$ to circulation is

$$\mathbf{f} \cdot d\mathbf{s}_1 = h_1 f_1 du$$

Along the top side, the sign is reversed and $h_1 f_1$ is evaluated at $v + dv$ rather than v. Both sides together give

$$\left[-(h_1 f_1)\Big|_{v+dv} + (h_1 f_1)\Big|_{v} \right] du = -\left[\frac{\partial}{\partial v}(h_1 f_1) \right] du dv.$$

Similarly, the right and the left sides yield

$$\left[\frac{\partial}{\partial u}(h_2 f_2) \right] du dv.$$

So, the total circulation is

$$\left[\frac{\partial}{\partial u}(h_2 f_2) - \frac{\partial}{\partial v}(h_1 f_1) \right] du dv = \frac{1}{h_1 h_2} \left[\frac{\partial}{\partial u}(h_2 f_2) - \frac{\partial}{\partial v}(h_1 f_1) \right] \hat{\mathbf{w}} \cdot d\mathbf{a}.$$

The coefficient of $d\mathbf{a}$ serves to define the w component of the curl. Constructing the u and v components in the same way, we get

$$\begin{aligned}\nabla\times\mathbf{f} \;=\; & \frac{1}{h_2h_3}\left[\frac{\partial}{\partial v}(h_3f_3)-\frac{\partial}{\partial w}(h_2f_2)\right]\hat{\mathbf{u}}+\frac{1}{h_1h_3}\left[\frac{\partial}{\partial w}(h_1f_1)-\frac{\partial}{\partial u}(h_3f_3)\right]\hat{\mathbf{v}}\\ & +\;\frac{1}{h_1h_2}\left[\frac{\partial}{\partial u}(h_2f_2)-\frac{\partial}{\partial v}(h_1f_1)\right]\hat{\mathbf{w}}.\end{aligned}\tag{10.107}$$

This expression for $\nabla\times\mathbf{f}$ can be written in a compact determinantal form as

$$\nabla\times\mathbf{f}=\frac{1}{h_1h_2h_3}\begin{vmatrix} h_1\hat{\mathbf{u}} & h_2\hat{\mathbf{v}} & h_3\hat{\mathbf{w}}\\ \frac{\partial}{\partial u} & \frac{\partial}{\partial v} & \frac{\partial}{\partial w}\\ h_1f_1 & h_2f_2 & h_3f_3\end{vmatrix}.\tag{10.108}$$

Exercise Express the vector derivatives, that is, gradient, divergence, curl and Laplacian in terms of (a) spherical polar and (b) cylindrical coordinates for a scalar field $u(\mathbf{x})$ and a vector field $\mathbf{v}(\mathbf{x})$.

Answer

(a) Gradient:

$$\nabla u=\frac{\partial u}{\partial r}\hat{\mathbf{r}}+\frac{1}{r}\frac{\partial u}{\partial\theta}\hat{\theta}+\frac{1}{r\sin\theta}\frac{\partial u}{\partial\phi}\hat{\phi}.$$

Divergence:

$$\nabla\cdot\mathbf{v}=\frac{1}{r^2}\frac{\partial}{\partial r}(r^2v_r)+\frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta v_\theta)+\frac{1}{r\sin\theta}\frac{\partial v_\phi}{\partial\phi}.$$

Curl:

$$\begin{aligned}\nabla\times\mathbf{v}=&\frac{1}{r\sin\theta}\left[\frac{\partial}{\partial\theta}(\sin\theta v_\phi)-\frac{\partial v_\theta}{\partial\phi}\right]\hat{\mathbf{r}}+\frac{1}{r}\left[\frac{1}{\sin\theta}\frac{\partial v_r}{\partial\phi}-\frac{\partial}{\partial r}(rv_\phi)\right]\\ &\hat{\theta}+\frac{1}{r}\left[\frac{\partial}{\partial r}(rv_\theta)-\frac{\partial v_r}{\partial\theta}\right]\hat{\phi}.\end{aligned}$$

Laplacian:

$$\nabla^2u=\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial u}{\partial r}\right)+\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial u}{\partial\theta}\right)+\frac{1}{r^2\sin^2\theta}\frac{\partial^2u}{\partial\phi^2}.$$

(b) Gradient:

$$\nabla u=\frac{\partial u}{\partial\rho}\hat{\rho}+\frac{1}{\rho}\frac{\partial u}{\partial\phi}\hat{\phi}+\frac{\partial u}{\partial z}\hat{\mathbf{z}}.$$

Divergence:

$$\nabla \cdot \mathbf{v} = \frac{1}{\rho}\frac{\partial}{\partial \rho}(\rho v_\rho) + \frac{1}{\rho}\frac{\partial v_\phi}{\partial \phi} + \frac{\partial v_z}{\partial z}.$$

Curl:

$$\nabla \times \mathbf{v} = \left(\frac{1}{\rho}\frac{\partial v_z}{\partial \phi} - \frac{\partial v_\phi}{\partial z}\right)\hat{\rho} + \left(\frac{\partial v_\rho}{\partial z} - \frac{\partial v_z}{\partial \rho}\right)\hat{\phi} + \frac{1}{\rho}\left(\frac{\partial}{\partial \rho}(\rho v_\phi) - \frac{\partial v_\rho}{\partial \phi}\right)\hat{\mathbf{z}}.$$

Laplacian:

$$\nabla^2 u = \frac{1}{\rho}\frac{\partial}{\partial \rho}\left(\rho \frac{\partial u}{\partial \rho}\right) + \frac{1}{\rho^2}\frac{\partial^2 u}{\partial \phi^2} + \frac{\partial^2 u}{\partial z^2}.$$

□

!"#' $ % % & ' ( )) * +) , -)
.) ") +) * + !) ! 2 !.! % 3 4

5- 7)4) O

Chapter 13

Straight Line

13.1 INTRODUCTION

The intersection of two planes P_1 and P_2 is the locus of all the common points on both the planes P_1 and P_2. This locus is a straight line. Any given line can be uniquely determined by any of the two planes containing the line. Thus, a line can be regarded as the locus of the common points of two intersecting planes.

Let us consider the two planes

$$ax + by + cz + d = 0 \qquad (13.1)$$

$$a_1x + b_1y + c_1z + d_1 = 0 \qquad (13.2)$$

Any set of coordinates (*x*, *y*, *z*) which satisfy these two equations simultaneously will represent a point on the line of intersection of these two planes. Hence these two equations taken together will represent a straight line. It can be noted that the equation of *x*–axis are $y = 0$, $z = 0$. The equation of the *y*–axis is $x = 0$, $z = 0$ and the equation of the *z*–axis is $x = 0$, $y = 0$. The representation of the straight line by the equations $ax + by + cz + d = 0$ and $a_1x + b_1y + c_1z + d_1 = 0$ is called non–symmetrical form. Let us now derive the equations of a straight line in the symmetrical form.

13.2 EQUATION OF A STRAIGHT LINE IN SYMMETRICAL FORM

Let $A(x_1, y_1, z_1)$ be a point on the straight line and $P(x, y, z)$ be any point on the straight line. Let *l*, *m*, *n* be the direction cosines of the straight line. Let $OP = r$. The projections of *AP* on the coordinate axes are $x - x_1$, $y - y_1$, $z - z_1$. Also the projections of *AP* on the coordinate axes are given by *lr*, *mr*, *nr*. Then $x - x_1 = lr$, $y - y_1 = mr$ and $z - z_1 = nr$.

$$\therefore \frac{x-x_1}{l} = \frac{y-y_1}{m} = \frac{z-z_1}{n} = r$$

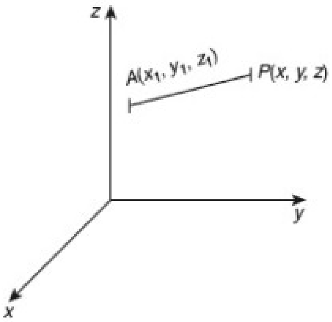

These equations are called the symmetrical form of the straight line.

Aliter:

We now derive the equations in symmetrical form from the vector equation of the straight line passing through a point and parallel to a vector.

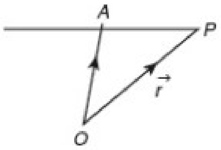

Let A be a given point on a straight line and P be any point on the straight line. Let $\vec{b}$ be a vector parallel to the line. Let O be the origin and $\overline{OA} = \vec{a}, \overline{OP} = \vec{r}$.

Then

$$\overrightarrow{AP} = \overrightarrow{OP} - \overrightarrow{OA} = \vec{r} - \vec{a} \qquad (13.3)$$

But

$$\overrightarrow{AP} = t\vec{b} \qquad (13.4)$$

where t is a scalar.

From (13.3) and (13.4),

$$\vec{r} - \vec{a} = t\vec{b} \text{ or } \vec{r} = \vec{a} + t\vec{b} \qquad (13.5)$$

This equation is true for all positions of P on the straight line and therefore this is the vector equation of the straight line. Let

$\overline{OP} = x\vec{i} + y\vec{j} + z\vec{k}, \overline{OA} = x_1\vec{i} + y_1\vec{j} + z_1\vec{k}$ and $\vec{b} = l\vec{i} + m\vec{j} + n\vec{k}$.

Then from equation (13.4), we have

$$(x - x_1)\vec{i} + (y - y_1)\vec{j} + (z - z_1)\vec{k} = t\vec{b} = t(l\vec{i} + m\vec{j} + n\vec{k})$$

Equating the coefficients of $\vec{i}$, $\vec{j}$ and $\vec{k}$, we have

$$x - x_1 = tl, \quad y - y_1 = tm, \quad z - z_1 = tn$$

$$\therefore \frac{x - x_1}{l} = \frac{y - y_1}{m} = \frac{z - z_1}{n} = t$$

These are the cartesian equations of the straight line in symmetrical form.

Note 13.2.1: To express the equations of a straight line in symmetrical form we require (i) the coordinate of a point on the line and (ii) the direction cosines of the straight line.

Note 13.2.2: Any point on this line is $(x_1 + lr, y_1 + mr, z_1 + nr)$. Even if l, m and n are the direction ratios of the line, $(x_1 + lr, y_1 + mr, z_1 + nr)$ will represent a point on the line but r will not be distance between the points (x, y, z) and (x_1, y_1, z_1).

13.3 EQUATIONS OF A STRAIGHT LINE PASSING THROUGH THE TWO GIVEN POINTS

Let $P(x_1, y_1, z_1)$ and $Q(x_2, y_2, z_2)$ be two given points. The direction ratios of the line are $x_2 - x_1, y_2 - y_1, z_2 - z_1$.

Therefore, the equations of the straight line are $\frac{x - x_1}{x_2 - x_1} = \frac{y - y_1}{y_2 - y_1} = \frac{z - z_1}{z_2 - z_1}$.

Aliter:

Let *O* be the origin and *P* and *Q* be the points on the straight line and *R* be any point on the straight line.

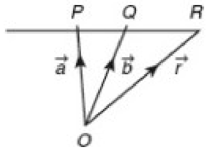

$$\overline{OP} = \vec{a}, \overline{OQ} = \vec{b}, \overline{OR} = \vec{r}$$

Then $\overline{PQ} = \overline{OQ} - \overline{OP} = \vec{b} - \vec{a}$. Also $\overline{PR} = \overline{OR} - \overline{OP} = \vec{r} - \vec{a}$.

But $\overline{PR} = t\overline{PQ}$ (i.e.) $\vec{r} - \vec{a} = t(\vec{b} - \vec{a})$

$$\vec{r} = \vec{a} + t(\vec{b} - \vec{a}) \qquad (13.6)$$

This is the vector equation of the straight line.

Let $\vec{r} = x\vec{i} + y\vec{j} + z\vec{k}, \vec{a} = x_1\vec{i} + y_1\vec{i} + z_1\vec{k}, \vec{b} = x_2\vec{i} + y_2\vec{j} + z_2\vec{k}$.

Then from (13.6), we get

$(x - x_1)\vec{i} + (y - y_1)\vec{j} + (z - z_1)\vec{k} = t[(x_2 - x_1)\vec{i} +$

Equating the coefficients of $\vec{i}$, $\vec{j}$ and $\vec{k}$, we get

$$x - x_1 = t(x_2 - x_1); y - y_1 = t(y_2 - y_1); z - z_1 = k(z_2 - z_1)$$

$$\therefore \frac{x - x_1}{x_2 - x_1} = \frac{y - y_1}{y_2 - y_1} = \frac{z - z_1}{z_2 - z_1} = t$$

13.4 EQUATIONS OF A STRAIGHT LINE DETERMINED BY A PAIR OF PLANES IN SYMMETRICAL FORM

!"#' $ % % & ' ( )) * +) , -)
.) & + * +))4 ( ) @ 2 !.! % 3 4

5- 2 3
- *) % 4 9 - 8 $) E%) % O 9) % '+) S) & +)
4 +)9 T < 6 @% 8 4 . ; L 8 *) G %
-) @) - + 9) 8 9+ -) +8
9 -)) 8)% +) 8 -)))) +
% + H/ # 9 8 & +)) -) +)44)) *)))%
) 4) % + @) - +8 *) - % 4))
)

- ( ) % @) - + 8
) + 4)')9 *) A + %) 7 *) A ; %
')) % + 8 9 4)))) 8 4 4)) *)
E % -) %) 2 T 738 4) + G) ')9 H
2) - - 3) %))) %)))
) - % -) 4
)+))+ 9 9) 4) T !) -
) ' 9+ O) @) 4 % -)
8 *) 9 -) - + 9 -

5- 72 /3
1 8 1 F % +) - 1)) - %)
@) - +) 9 -) +) @) - 2)
O) A # - O ' - *) 38) % + @) 4 4
) 4) @) % 8 4 +
% + -) @) - + J)) @) 8
8 ! 9) - 4) ;4 8 1 F
) 4))9 + ;4) @) 8 * +
8 -)+ 9) + 8 9)
4 -) 8 8) +)) @) - +
4 + G H 9)

5- T2 N3

)) -) 8 %) 4) 4 %) G +)) +
) % 8) %)) 4) - 8H))% -
G) 4 -) 8H 4)4 -)) 9 ;
) 8 G+) %) U 4) H

) %)) 4) 4 %) V % W) =) = %
)))) 4)V8 4 %)J) V) X)) 4 -)V8) % F
) =) = % 8 V% W -))) -) V

5- /2 3

quantum chromodynamics, and continues to inspire the searches for grand unified theories and the alluring prospects of string theory. Cartan's torsion remains important in the search for quantum gravity, as does Riemannian geometry itself.

All this takes us back to Riemann, whom we should not merely admire as a forerunner but whose vision still holds future prospects. His 1854 remarks about the possibilities that the spatial fabric might break down into discrete "quanta" at very small distances have become the focus of much speculation in the quest for a complete theory of quantum gravity. Many physicists now envisage that some radical change in the nature of space-time occurs near the Planck length, $\sqrt{\frac{hG}{2\pi c^3}} = 1.6 \times 10^{-35}$ m, at which point they anticipate something like a discrete manifold, no longer continuous, as Riemann had suggested. Non-relativistic quantum theory implies that we cannot build clocks and measuring rods not just at atomic scales (the smallest fully functional clock is about the size of a small microorganism); relativistic quantum theory draws from the invariance of the speed of light yet more stringent limitations on the possible measurement of physical quantities.[47] Unlike the problem of rigid bodies, which could be circumvented by light rays and falling particles, these quantum limitations seem more fundamental, ultimately connected with the radical indistinguishability of quanta and the infinite-dimensional Hilbert space mandated by quantum theory.[48]

At whatever scale the issue is joined, though, physicists largely subscribe to Riemann's insight that "upon the exactness with which we pursue phenomena into the infinitely small, does our knowledge of their causal connections essentially depend." He, like Gauss and Lobachevsky before him, also directed our attention to astronomical consequences of the varying curvature of space. Most of all, we realize more and more clearly the force of his insight that "either the reality underlying space must form a discrete manifold, or the basis for the metric relations must be sought outside it, in binding forces acting upon it." Both general relativity and the still-undiscovered theories that will correct and transcend it must struggle, in their own ways, with the quest for these "binding forces," though quantum theory remains the touchstone whose existence Riemann did not suspect and Einstein too much resisted.

This leaves us at the threshold of myriad possibilities that challenge us to add, in our turn, our own ideas of what we think lies "at the foundations of geometry." If space-time is not truly ultimate, what other mathematical structures might lie beneath? Loop quantum gravity aspires to build up space-time without presupposing it. String theory incorporates general relativity into a larger theoretical structure, but does it implicitly rest on accepting space-time as a given backdrop or can it lead to other possibilities?[49] Yet we still lack a fundamental understanding even of the entire range of possibilities. Let me pose the larger question: what are all the possible mathematical structures that could give rise to the *appearance* of space-time, as we know it? This mathematical challenge then leads us to reflect on how, physically, we should judge between the possible alternatives. Once we have left behind the remnants of the

") %) J 8 -84+)- + 9 9
( ) A -) G4 ;) 4 4)
+) 8 ' -))
) + 4 H 1 8 ' *)) O 9) % '+ 9 8)
))) = %) + - %) 4)
) 8) F)) + -) G
) + + - 4))) 8 9)
) 9 - 8 9 -) - 4 H

& -)) % +) %))
) -- 8)+ 8 = G9
- 8H - =) +) ;
( ) 4) @

5- N2 63
EE)))
4 9 8) 4))%) -
4 -) 8) % +)
9) 9+ % +
$) 4) ;)) 9 8
))))) 9)) 4 ; -
)) 9)) 8 8
) 4 9 4 4)9 % 8) % -)
))) 4 9 8)
)) 9 % -) +
)9) 8) 4 ; 9 + -)
Y)) 4 8 % 8) %)
) %) +)4)9 - 4) 8))+ 9
4 9 ;4)) -) + # 8) -)
)9) % -))) +) 9)% 8)
9 4 - -) 9 #) 9 4) =
)) 9)) %+ *) %
) N

5- /26 3

N Z") 8 ( )) ;
) = 2) 3
) 8 - %) 24 9 + % 3)) 4)) 9) + 8
)) 2G) H3) 9)) I
E)) @) 4) 8)) + - %)
) - 9 4 8 - - % 9+ +)-)
) =) =) 4 &
)-) - %))9) 9 4)
) %) + 9+ ' -) % + -))
4 4)) 8)% -) G ' H
) - I ( ) 4 9+ *) /8)
))))) 9 4 +
% + -)) -) 4)) - % %))
)) + - % 4 # G*)) %) H *))
))%)) ' 9) ' #
4 -) 9) 4) ( ) 8 4 4
-) F) -) 4) [

5- T2 N3

2

For this purpose, one constructs the system of shortest lines emanating from a given point; the position of an arbitrary point can then be determined by the initial direction of the shortest line in which it lies, and its distance, in this line, from the initial point.[12] It can therefore be expressed by the ratios of the quantities dx^0, i.e., the quantities dx at the origin of this shortest line, and by the length s of this line. In place of the dx^0 one now introduces linear expressions $d\alpha$ formed from them in such a way that the initial value of the square of the line element will be equal to the sum of the squares of these expressions, so that the independent variables are: the quantity s and the ratio of the quantities $d\alpha$. Finally, in place of the $d\alpha$ choose quantities x_1, x_2,..., x_n proportional to them, but such that the sum of their squares equals s^2. If one introduces these quantities, then for infinitely small values of x the square of the line element $= \sum dx^2$, but the next order term in its expansion equals a homogeneous expression of the second degree in the $n\frac{n-1}{2}$ quantities $(x_1dx_2 - x_2dx_1)$, $(x_1dx_3 - x_3dx_1)$,..., and is consequently an infinitely small quantity of the fourth order, so that one obtains a finite quantity if one divides it by the square of the infinitely small triangle at whose vertices the variables have the values $(0, 0, 0, ...)$, $(x_1, x_2, x_3, ...)$, $(dx_1, dx_2, dx_3, ...)$. This quantity remains the same as long as the quantities x and dx are contained in the same binary linear forms, or as long as the two shortest lines from the initial point to x and from the initial point to dx remain in the same surface element, and therefore depends only on the position and direction of that element. It obviously $=$ zero if the manifold in question is flat, i.e., if the square of the line element is reducible to $\sum dx^2$, and can therefore be regarded as the measure of deviation from flatness in this surface direction at this point. When multiplied by $-\frac{3}{4}$ it becomes equal to the quantity which Privy Councilor Gauss has called the curvature of a surface.[13] Previously, $n\frac{n-1}{2}$ functions of position were found necessary in order to determine the metric relations of an n-fold extended manifold representable in the assumed form; hence if the curvature is given in $n\frac{n-1}{2}$ surface directions at every point, then the metric relations of the manifold may be determined, provided only that no identical relations can be found between these values, and indeed in general this does not occur. The metric relations of these manifolds, in which the line element can be represented as the square root of a differential expression of the second degree, can thus be expressed in a way completely independent of the choice of the varying quantities. A similar path to the same goal could also be taken in those manifolds in which the line element is expressed in a less simple way, e.g., by the fourth root of a differential expression of the fourth degree. The line element in this more general case would not be reducible to the square root of a quadratic sum of differential expressions, and therefore in the expression for the square of the line element the deviation from flatness would be an infinitely small quantity of the second order, whereas for the other manifolds it was an infinitely small quantity of the fourth order. This peculiarity of the latter manifolds therefore might well be called flatness in the smallest parts. For present purposes, however, the most important peculiarity of these manifolds, on whose

D 4 9+ :6 9 =) =) +
%+ *))) %)))

5- <2673
Z. *) A %) 8 *) /8 < \ )-)
) T =) ' 2; ; B ; 3 - ;
*) A [

5- <2673
6 ZE G) ')9 H 2) - - 38 *))
)))) 8 %)) %))
9 -) -\ *) /8 % % 8 1
( ) 4) 4 - -))))) *) A
') -) %) ; %)
E ;) 4 8) 4))) +) 9) +) 8 - ;
) + %) 9) 4) 8)
) 8 *)) %)) 4)) +) 9 % +
F \ % + -)) 9+ % - 9
,) 8) 4))) 8 4 % %))
) 9 G H) 4)) 4)-) +)
9) 4 # 9) ; %)
-) 8 9 4) 9)))
9))9)) 2)) 3 8 -
) 4)))) 9
-) 4) # G %) H 2)
4) 3) %)) - 4 4))
) 9)) + 4) \ ) 9) 4 4)4
4 + -))) + 4 8) *)
) 4 %) %)) 4 4) 4
%) 2*) /8 /38) 4) + ; 9 -
) \ ( ))) 4 4 + %) 8
4) + 9 %)) -) 4 9 -)
) 4) [

5- /72N/3
O ' 4)) 4 8 4 4)))%)
%)) 8) % + 4) 9 ;) +)44 % +
4 8)) -) 4))
) 9)) + 4) 4 - +) 4
=) +) + -) #) %)
)) 9+ *) 8 4 % 4) F 4) 8
)%)) -) % %) 4 4)) 8 9)
4 4) %))))4))%)% +
44)+

5- /2 3
< ZL A ;) 4))) 4 -
2)) +) 9+)) 8 9 3 - 2)
*) 4 % 3) +))) @) 4))%)
2*)) 3 %) 8 4 -) = 8 9+
- + 8)) - 4)) &+
G) -H L 4 +)) 8 -) 9+) -))
) -); 4) 2) L <N 8 6N3 L))
) ;) 4) L) 8) + % +) 8
9) -))) =)
44 - 8) . - \ G) 8H <NT8
< < 8) (<TT8 <T #) -) L
4 9)))) %)
)) % + -) 2)) ;) 4 3 E) -

8 4 - -) 4)) -) 8) GL 9 H
44 - \ &) ' <N78 6 6 8) O F) <N/8 N
T [

5- N 2 7 3
Z& +) A) D -) -) 9) *))))
+ 4 9 4)) 4) \ 8) %
S))-) 9 4 9 G0) 4
)44) 4)) E ')+ % + E)%)%
9 - 8 ; -) -) J + + . * A
)' R E) +)% 4))) 8)9 ') '8 +
)+ 4 % + + 8) 8) = +8))44 +
H ' 4) N8 N8 *)+ 6) N [

5- TN2 <63
N Z( ) -) F *) A) %
)) - %) 4) 8) *) &
\ *) /8 % 8 8 6T [

,9
*) <
( ) 6 <
@ /N

%) 7<

!'#' $ % % & ' ( )) * +) , -)
.) !) * + %)) (%) Y4) @ 4

*) /
( ) 6
%) /

5- N2 / 3 # * + *))4

5- 2 N3 6 # E * +)

,9 O % RRR

5- 66 2 N 3
(, , E#E,$ 2 3 " % + -))))%
-)) - % 4 4)) -)%) 4)
) %)

Remark. The normal curvature of C does not depend on the orientation of C but changes sign with a change of orientation for the surface.

To give an interpretation of the second fundamental form II_p, consider a regular curve $C \subset S$ parametrized by $\alpha(s)$, where s is the arc length of C, and with $\alpha(0) = p$. If we denote by $N(s)$ the restriction of the normal vector N to the curve $\alpha(s)$, we have $\langle N(s), \alpha'(s)\rangle = 0$. Hence,

$$\langle N(s), \alpha''(s)\rangle = -\langle N'(s), \alpha'(s)\rangle.$$

Therefore,

$$\begin{aligned} II_p(\alpha'(0)) &= -\langle dN_p(\alpha'(0)), \alpha'(0)\rangle \\ &= -\langle N'(0), \alpha'(0)\rangle = \langle N(0), \alpha''(0)\rangle \\ &= \langle N, kn\rangle(p) = k_n(p). \end{aligned}$$

In other words, the value of the second fundamental form II_p for a unit vector $v \in T_p(S)$ is equal to the normal curvature of a regular curve passing through p and tangent to v. In particular, we obtained the following result.

PROPOSITION 2 (Meusnier). *All curves lying on a surface* S *and having at a given point* p ∈ S *the same tangent line have at this point the same normal curvatures.*

The above proposition allows us to speak of the *normal curvature along a given direction at p.* It is convenient to use the following terminology. Given a unit vector $v \in T_p(S)$, the intersection of S with the plane containing v and $N(p)$ is called the *normal section* of S at p along v (Fig. 3-9). In a neighborhood of p, a normal section of S at p is a regular plane curve on S whose normal vector n at p is $\pm N(p)$ or zero; its curvature is therefore equal to the absolute value of the normal curvature along v at p. With this terminology, the above

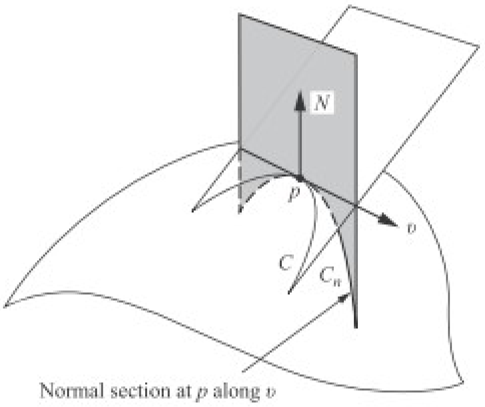

Figure 3-9. Meusnier theorem: C and C_n have the same normal curvature at p along v.

(, , E#E,$ 2 3 " % + -))))%
-)) - % 4 4)) -)%) 4)
) %)

)9 % 4 4) 4)') %)) -)

- %) 4 E % - -+ * %
) % % # 4 2 38 4)) - %)
$243))) 4) - % 2. - <3 E) - 9
48))) 4) -) 4) %)
%) 4] $243 F \ %) =))9
%)) %)) - %) 4 D -+8)9 %

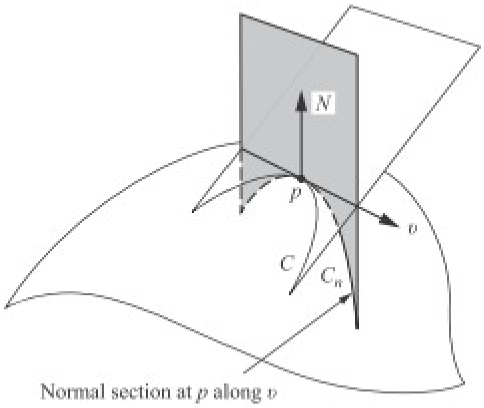

. - <))%)) %)) 4) - %

4 4)+))9 %)) %)) 4) %
^2 3 =) %))) 4) - ^_2 3

5- 6T2 N63
!@.E$E#E,$ N O 4) $ 4 # 42 3 ` # 42 3 9
) *))4 #) $ 4 *)) %) L
) 4 # -) %)) $ 4)) %)
1) 4
E 4 4) %))

$$K = k_1 k_2, \quad H = \frac{k_1 + k_2}{2}.$$

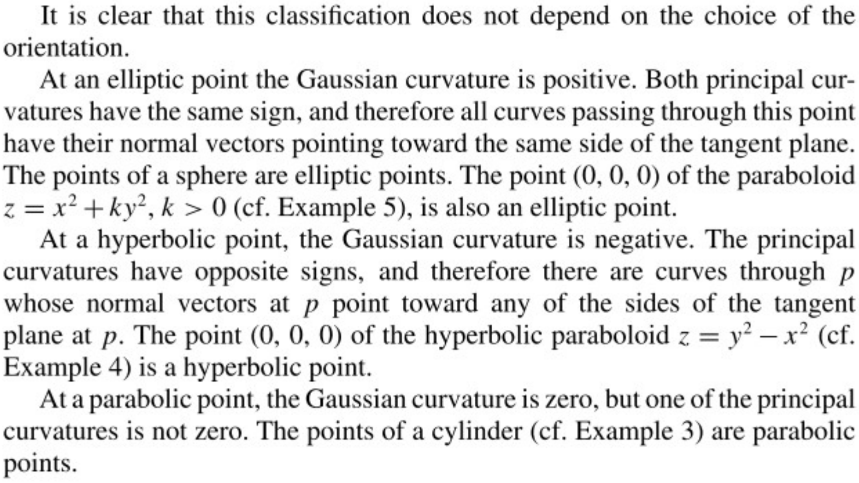

It is clear that this classification does not depend on the choice of the orientation.

At an elliptic point the Gaussian curvature is positive. Both principal curvatures have the same sign, and therefore all curves passing through this point have their normal vectors pointing toward the same side of the tangent plane. The points of a sphere are elliptic points. The point (0, 0, 0) of the paraboloid $z = x^2 + ky^2$, $k > 0$ (cf. Example 5), is also an elliptic point.

At a hyperbolic point, the Gaussian curvature is negative. The principal curvatures have opposite signs, and therefore there are curves through p whose normal vectors at p point toward any of the sides of the tangent plane at p. The point (0, 0, 0) of the hyperbolic paraboloid $z = y^2 - x^2$ (cf. Example 4) is a hyperbolic point.

At a parabolic point, the Gaussian curvature is zero, but one of the principal curvatures is not zero. The points of a cylinder (cf. Example 3) are parabolic points.

.) +8)) 4)) 4 8) 4 4) %)) F # 4
) 4) %) +) + " %) ;) 4) 4)) 4
) - % @;) 4 N

5- 72 / 3
#1@,(@ " @*(@*EY 2*) 3 # *)) %) L)
) %)) 9+)

4-3. The Gauss Theorem and the Equations of Compatibility

The properties of Chap. 3 were obtained from the study of the variation of the tangent plane in a neighborhood of a point. Proceeding with the analogy with curves, we are going to assign to each point of a surface a trihedron (the analogue of Frenet's trihedron) and study the derivatives of its vectors.

S will denote, as usual, a regular, orientable, and oriented surface. Let $\mathbf{x}\colon U \subset R^2 \to S$ be a parametrization in the orientation of S. It is possible to assign to each point of $\mathbf{x}(U)$ a natural trihedron given by the vectors $\mathbf{x}_u$, $\mathbf{x}_v$, and N. The study of this trihedron will be the subject of this section.

By expressing the derivatives of the vectors $\mathbf{x}_u, \mathbf{x}_v$, and N in the basis $\{\mathbf{x}_u, \mathbf{x}_v, N\}$, we obtain

$$\begin{aligned}
\mathbf{x}_{uu} &= \Gamma^1_{11}\mathbf{x}_u + \Gamma^2_{11}\mathbf{x}_v + L_1 N,\\
\mathbf{x}_{uv} &= \Gamma^1_{12}\mathbf{x}_u + \Gamma^2_{12}\mathbf{x}_v + L_2 N,\\
\mathbf{x}_{vu} &= \Gamma^1_{21}\mathbf{x}_u + \Gamma^2_{21}\mathbf{x}_v + \bar{L}_2 N,\\
\mathbf{x}_{vv} &= \Gamma^1_{22}\mathbf{x}_u + \Gamma^2_{22}\mathbf{x}_v + L_3 N,\\
N_u &= a_{11}\mathbf{x}_u + a_{21}\mathbf{x}_v,\\
N_v &= a_{12}\mathbf{x}_u + a_{22}\mathbf{x}_v,
\end{aligned} \tag{1}$$

where the a_{ij}, $i, j = 1, 2$, were obtained in Chap. 3 and the other coefficients are to be determined. The coefficients Γ^k_{ij}, $i, j, k = 1, 2$, are called the Christoffel symbols of S in the parametrization $\mathbf{x}$. Since $\mathbf{x}_{uv} = \mathbf{x}_{vu}$, we conclude that $\Gamma^1_{12} = \Gamma^1_{21}$ and $\Gamma^2_{12} = \Gamma^2_{21}$; that is, the *Christoffel symbols* are symmetric relative to the lower indices.

By taking the inner product of the first four relations in (1) with N, we immediately obtain $L_1 = e$, $L_2 = \bar{L}_2 = f$, $L_3 = g$, where e, f, g are the coefficients of the second fundamental form of S.

To determine the Christoffel symbols, we take the inner product of the first four relations with $\mathbf{x}_u$ and $\mathbf{x}_v$, obtaining the system

$$\begin{aligned}
&\begin{cases}
\Gamma^1_{11}E + \Gamma^2_{11}F = \langle \mathbf{x}_{uu}, \mathbf{x}_u\rangle = \frac{1}{2}E_u,\\
\Gamma^1_{11}F + \Gamma^2_{11}G = \langle \mathbf{x}_{uu}, \mathbf{x}_v\rangle = F_u - \frac{1}{2}E_v,
\end{cases}\\
&\begin{cases}
\Gamma^1_{12}E + \Gamma^2_{12}F = \langle \mathbf{x}_{uv}, \mathbf{x}_u\rangle = \frac{1}{2}E_v,\\
\Gamma^1_{12}F + \Gamma^2_{12}G = \langle \mathbf{x}_{uv}, \mathbf{x}_v\rangle = \frac{1}{2}G_u,
\end{cases}\\
&\begin{cases}
\Gamma^1_{22}E + \Gamma^2_{22}F = \langle \mathbf{x}_{vv}, \mathbf{x}_u\rangle = F_v - \frac{1}{2}G_u,\\
\Gamma^1_{22}F + \Gamma^2_{22}G = \langle \mathbf{x}_{vv}, \mathbf{x}_v\rangle = \frac{1}{2}G_v.
\end{cases}
\end{aligned} \tag{2}$$

Note that the above equations have been grouped into three pairs of equations and that for each pair the determinant of the system is $EG - F^2 \neq 0$.

Thus, it is possible to solve the above system and *to compute the Christoffel symbols in terms of the coefficients of the first fundamental form, E, F, G, and their derivatives.* We shall not obtain the explicit expressions of the $\mathbf{\Gamma}_{ij}^k$, since it is easier to work in each particular case with the system (2). (See Example 1 below.) However, the following consequence of the fact that we can solve the system (2) is very important: *All geometric concepts and properties expressed in terms of the Christoffel symbols are invariant under isometries.*

Example 1. We shall compute the Christoffel symbols for a surface of revolution parametrized by (cf. Example 4, Sec. 2-3)

$$\mathbf{x}(u, v) = \{f(v)\cos u, f(v)\sin u, g(v)), \quad f(v) \neq 0.$$

Since

$$E = (f(v))^2, \quad F = 0, \quad G = \{f'(v))^2 + (g'(v))^2,$$

we obtain

$$\begin{aligned} E_u &= 0, \qquad E_v = 2ff', \\ F_u &= F_v = 0, \quad G_u = 0, \\ G_v &= 2(f'f'' + g'g''), \end{aligned}$$

where prime denotes derivative with respect to v. The first two equations of the system (2) then give

$$\mathbf{\Gamma}_{11}^1 = 0, \quad \mathbf{\Gamma}_{11}^2 = -\frac{ff'}{(f')^2 + (g')^2}.$$

Next, the second pair of equations in system (2) yield

$$\mathbf{\Gamma}_{12}^1 = \frac{ff'}{f^2}, \quad \mathbf{\Gamma}_{12}^2 = 0.$$

Finally, from the last two equations in system (2) we obtain

$$\mathbf{\Gamma}_{22}^1 = 0, \quad \mathbf{\Gamma}_{22}^2 = \frac{f'f'' + g'g''}{(f')^2 + (g')^2}.$$

As we have just seen, the expressions of the derivatives of $\mathbf{x}_u$, $\mathbf{x}_v$, and N in the basis $\{x_u, \mathbf{x}_v, N\}$ involve only the knowledge of the coefficients of the first and second fundamental forms of S. A way of obtaining relations between these coefficients is to consider the expressions

$$\begin{aligned} (\mathbf{x}_{uu})_v - (\mathbf{x}_{uv})_u &= 0, \\ (\mathbf{x}_{vv})_u - (\mathbf{x}_{vu})_v &= 0, \\ N_{uv} - N_{vu} &= 0. \end{aligned} \tag{3}$$

By introducing the values of (1), we may write the above relations in the form

$$\begin{aligned} A_1\mathbf{x}_u + B_1\mathbf{x}_v + C_1 N &= 0, \\ A_2\mathbf{x}_u + B_2\mathbf{x}_v + C_2 N &= 0, \\ A_3\mathbf{x}_u + B_3\mathbf{x}_v + C_3 N &= 0, \end{aligned} \tag{3a}$$

where A_i, B_i, C_i, $i = 1, 2, 3$, are functions of E, F, G, e, f, g and of their derivatives. Since the vectors $\mathbf{x}_u, \mathbf{x}_v, N$ are linearly independent, (3a) implies that there exist nine relations:

$$A_i = 0, \quad B_i = 0, \quad C_i = 0, \quad i = 1, 2, 3.$$

As an example, we shall determine the relations $A_1 = 0$, $B_1 = 0$, $C_1 = 0$. By using the values of (1), the first of the relations (3) may be written

$$\begin{aligned} &\mathbf{\Gamma}_{11}^{1}\mathbf{x}_{uv} + \mathbf{\Gamma}_{11}^{2}\mathbf{x}_{vv} + eN_v + (\mathbf{\Gamma}_{11}^{1})_v\mathbf{x}_u + (\mathbf{\Gamma}_{11}^{2})_v\mathbf{x}_v + e_v N \\ &\quad = \mathbf{\Gamma}_{12}^{1}\mathbf{x}_{uu} + \mathbf{\Gamma}_{12}^{2}\mathbf{x}_{vu} + fN_u + (\mathbf{\Gamma}_{12}^{1})_u\mathbf{x}_u + (\mathbf{\Gamma}_{12}^{2})_u\mathbf{x}_v + f_u N. \end{aligned} \tag{4}$$

By using (1) again and equating the coefficients of $\mathbf{x}_v$, we obtain

$$\begin{aligned} &\mathbf{\Gamma}_{11}^{1}\mathbf{\Gamma}_{12}^{2} + \mathbf{\Gamma}_{11}^{2}\mathbf{\Gamma}_{22}^{2} + ea_{22} + (\mathbf{\Gamma}_{11}^{2})_v \\ &\quad = \mathbf{\Gamma}_{12}^{1}\mathbf{\Gamma}_{11}^{2} + \mathbf{\Gamma}_{12}^{2}\mathbf{\Gamma}_{12}^{2} + fa_{21} + (\mathbf{\Gamma}_{12}^{2})_u. \end{aligned}$$

Introducing the values of a_{ij} already computed (cf. Sec. 3-3) it follows that

$$\begin{aligned} (\mathbf{\Gamma}_{12}^{2})_u - (\mathbf{\Gamma}_{11}^{2})_v + \mathbf{\Gamma}_{12}^{1}\mathbf{\Gamma}_{11}^{2} + \mathbf{\Gamma}_{12}^{2}\mathbf{\Gamma}_{12}^{2} - \mathbf{\Gamma}_{11}^{2}\mathbf{\Gamma}_{22}^{2} - \mathbf{\Gamma}_{11}^{1}\mathbf{\Gamma}_{12}^{2} &\\ = -E\frac{eg - f^2}{EG - F^2} &\\ = -EK. & \end{aligned} \tag{5}$$

At this point it is convenient to interrupt our computations in order to draw attention to the fact that the above equation proves the following theorem, due to K. F. Gauss.

THEOREMA EGREGIUM (Gauss). *The Gaussian curvature* K *of a surface is invariant by local isometries.*

In fact if $\mathbf{x}\colon U \subset R^2 \to S$ is a parametrization at $p \in S$ and if $\varphi\colon V \subset S - \bar{S}$, where $V \subset \mathbf{x}(U)$ is a neighborhood of p, is a local isometry at p, then $y = \varphi \circ \mathbf{x}$ is a parametrization of $\bar{S}$ at $\varphi(p)$. Since φ is an isometry, the coefficients of the first fundamental form in the parametrizations $\mathbf{x}$ and $\mathbf{y}$ agree at corresponding points q and $\varphi(q)$, $q \in V$; thus, the corresponding Christoffel symbols also agree. By Eq. (5), K can be computed at a point as a function of the Christoffel symbols in a given parametrization at the point. It follows that $K(q) = K(\varphi(q))$ for all $q \in V$.

The above expression, which yields the value of K in terms of the coefficients of the first fundamental form and its derivatives, is known as the *Gauss formula.* It was first proved by Gauss in a famous paper [1].

The Gauss theorem is considered, by the extension of its consequences, one of the most important facts of differential geometry. For the moment we shall mention only the following corollary.

As was proved in Sec. 4-2, a catenoid is locally isometric to a helicoid. It follows from the Gauss theorem that the Gaussian curvatures are equal at corresponding points, a fact which is geometrically nontrivial.

Actually, it is a remarkable fact that a concept such as the Gaussian curvature, the definition of which made essential use of the position of a surface in the space, does not depend on this position but only on the metric structure (first fundamental form) of the surface.

We shall see in the next section that many other concepts of differential geometry are in the same setting as the Gaussian curvature; that is, they depend only on the first fundamental form of the surface. It thus makes sense to talk about a geometry of the first fundamental form, which we call intrinsic geometry, since it may be developed without any reference to the space that contains the surface (once the first fundamental form is given).

[†]With an eye to a further geometrical result we come back to our computations. By equating the coefficients of $\mathbf{x}_u$ in (4), we see that the relation $A_1 = 0$ may be written in the form

$$(\mathbf{\Gamma}^1_{12})_u - (\mathbf{\Gamma}^1_{11})_v + \mathbf{\Gamma}^2_{12}\mathbf{\Gamma}^1_{12} - \mathbf{\Gamma}^2_{11}\mathbf{\Gamma}^1_{22} = FK. \qquad \mathbf{(5a)}$$

By equating also in Eq. (4) the coefficients of N, we obtain $C_1 = 0$ in the form

$$e_v - f_u = e\mathbf{\Gamma}^1_{12} + f(\mathbf{\Gamma}^2_{12} - \mathbf{\Gamma}^1_{11}) - g\mathbf{\Gamma}^2_{11}. \qquad \mathbf{(6)}$$

Observe that relation (5a) is (when $F \neq 0$) merely another form of the Gauss formula (5).

By applying the same process to the second expression of (3), we obtain that both the equations $A_2 = 0$ and $B_2 = 0$ give again the Gauss formula (5). Furthermore, $C_2 = 0$ is given by

$$f_v - g_u = e\mathbf{\Gamma}^1_{22} + f(\mathbf{\Gamma}^2_{22} - \mathbf{\Gamma}^1_{12}) - g\mathbf{\Gamma}^2_{12}. \qquad \mathbf{(6a)}$$

Finally, the same process can be applied to the last expression of (3), yielding that $C_3 = 0$ is an identity and that $A_3 = 0$ and $B_3 = 0$ are again Eqs. (6) and (6a). Equations (6) and (6a) are called *Mainardi-Codazzi equations.*

[†] The rest of this section will not be used until Chap. 5. If omitted, Exercises 7 and 8 should also be omitted.

The Gauss formula and the Mainardi-Codazzi equations are known under the name of *compatibility equations* of the theory of surfaces.

A natural question is whether there exist further relations of compatibility between the first and the second fundamental forms besides those already obtained. The theorem stated below shows that the answer is negative. In other words, by successive derivations or any other process we would obtain no further relations among the coefficients E, F, G, e, f, g and their derivatives. Actually, the theorem is more explicit and asserts that the knowledge of the first and second fundamental forms determines a surface locally. More precisely,

THEOREM (Bonnet). *Let* E, F, G, e, f, g *be differentiable functions, defined in an open set* $\mathrm{V} \subset \mathrm{R}^2$, *with* $\mathrm{E} > 0$ *and* $\mathrm{G} > 0$. *Assume that the given functions satisfy formally the Gauss and Mainardi-Codazzi equations and that* $\mathrm{EG} - \mathrm{F}^2 > 0$. *Then, for every* $\mathrm{q} \in \mathrm{V}$ *there exists a neighborhood* $\mathrm{U} \subset \mathrm{V}$ *of* q *and a diffeomorphism* $\mathbf{x}\colon \mathrm{U} \to \mathbf{x}(\mathrm{U}) \subset \mathrm{R}^3$ *such that the regular surface* $\mathbf{x}(\mathrm{U}) \subset \mathrm{R}^3$ *has* E, F, G *and* e, f, g *as coefficients of the first and second fundamental forms, respectively. Furthermore, if* U *is connected and if*

$$\bar{\mathrm{x}}\colon \mathrm{U} \to \bar{\mathrm{x}}(\mathrm{U}) \subset \mathrm{R}^3$$

is another diffeomorphism satisfying the same conditions, then there exist a translation T *and a proper linear orthogonal transformation* ρ *in* R^3 *such that* $\bar{\mathrm{x}} = \mathrm{T} \circ \rho \circ \mathbf{x}$.

A proof of this theorem may be found in the appendix to Chap. 4.

For later use, it is convenient to observe how the Mainardi-Codazzi equations simplify when the coordinate neighborhood contains no umbilical points and the coordinate curves are lines of curvature ($F = 0 = f$). Then, Eqs. (6) and (6a) may be written

$$e_v = e\mathbf{\Gamma}^1_{12} - g\mathbf{\Gamma}^2_{11}, \quad g_u = g\mathbf{\Gamma}^2_{12} - e\mathbf{\Gamma}^1_{22}.$$

By taking into consideration that $F = 0$ implies that

$$\mathbf{\Gamma}^2_{11} = -\frac{1}{2}\frac{E_v}{G}, \quad \mathbf{\Gamma}^1_{12} = \frac{1}{2}\frac{E_v}{E},$$
$$\mathbf{\Gamma}^1_{22} = -\frac{1}{2}\frac{G_u}{E}, \quad \mathbf{\Gamma}^2_{12} = \frac{1}{2}\frac{G_u}{G},$$

we conclude that the Mainardi-Codazzi equations take the following form:

$$e_v = \frac{E_v}{2}\left(\frac{e}{E} + \frac{g}{G}\right), \tag{7}$$

$$g_u = \frac{G_u}{2}\left(\frac{e}{E} + \frac{g}{G}\right). \tag{7a}$$

6 6)) #) 4 *
D) 4) +) ;4 - +
4)+ %) - 4 8) - % P
) 4) % % - 4) ;) 1 % 8
) 4 %))) 4 9 4 +
P))
D)) P %)) %) %) % P 8
)) -))) %
4) D))) 2) - 3 % P) 4 Y) -)
)) 4)) -) 4 Y) % 243
4 2 3

4-4. Parallel Transport. Geodesics.

We shall now proceed to a systematic exposition of the intrinsic geometry. To display the intuitive meaning of the concepts, we shall often give definitions and interpretations involving the space exterior to the surface. However, we shall prove in each case that the concepts to be introduced depend only on the first fundamental form.

We shall start with the definition of covariant derivative of a vector field, which is the analogue for surfaces of the usual differentiation of vectors in the plane. We recall that a *(tangent) vector field* in an open set $U \subset S$ of a regular surface S is a correspondence w that assigns to each $p \in U$ a vector $w(p) \in T_p(S)$. The vector field w is *differentiable* at p if, for some parametrization $\mathbf{x}(u, v)$ in p, the components a and b of $w = a\mathbf{x}_u + b\mathbf{x}_v$ in the basis $\{\mathbf{x}_u, \mathbf{x}_v\}$ are differentiable functions at p. The vector field w is differentiable in U if it is differentiable for every $p \in U$.

DEFINITION 1. *Let* w *be a differentiable vector field in an open set* $\mathrm{U} \subset \mathrm{S}$ *and* $\mathrm{p} \in \mathrm{U}$. *Let* $\mathrm{y} \in \mathrm{T_p(S)}$. *Consider a parametrized curve*

$$\alpha\colon (-\epsilon, \epsilon) \to \mathrm{U},$$

with $\alpha(0) = \mathrm{p}$ *and* $\alpha'(0) = \mathrm{y}$, *and let* $\mathrm{w(t)}$, $\mathrm{t} \in (-\epsilon, \epsilon)$, *be the restriction of the vector field* w *to the curve* α. *The vector obtained by the normal projection of* $(\mathrm{dw/dt})(0)$ *onto the plane* $\mathrm{T_p(S)}$ *is called the* covariant derivative *at* p *of the vector field* w *relative to the vector* y. *This covariant derivative is denoted by* $(\mathrm{Dw/dt})(0)$ *or* $(\mathrm{D_y w})(\mathrm{p})$ *(Fig. 4-8).*

The above definition makes use of the normal vector of *S* and of a particular curve α, tangent to y at p. To show that covariant differentiation is a concept

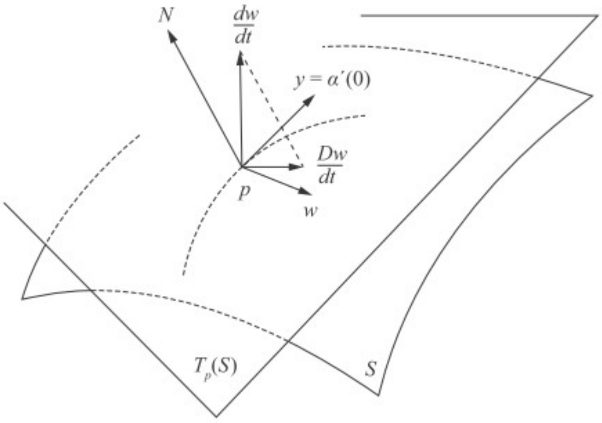

Figure 4-8. The covariant derivative.

b. If θ, $0 \leq \theta \leq \pi/2$, is the angle which a geodesic makes with the curve $v = \text{const.}$, then

$$U \sin^2\theta - V \cos^2\theta = \text{const.}$$

(Notice that this is the analogue of Clairaut's relation for the surfaces of Liouville.)

22. Let $S^2 = \{(x, y, z) \in R^3; x^2 + y^2 + z^2 = 1\}$ and let $p \in S^2$. For each piecewise regular parametrized curve $\alpha\colon [0, l] \to S^2$ with $\alpha(0) = \alpha(l) = p$, let $P_\alpha\colon T_p(S^2) \to T_p(S^2)$ be the map which assigns to each $v \in T_p(S^2)$ its parallel transport along α back to p. By Prop. 1, P_α is an isometry. Prove that for every rotation R of $T_p(S)$ there exists an α such that $R = P_\alpha$.

23. Show that the isometries of the unit sphere

$$S^2 = \{(x, y, z) \in R^3; x^2 + y^2 + z^2 = 1\}$$

are the restrictions to S^2 of the linear orthogonal transformations of R^3.

4-5. The Gauss-Bonnet Theorem and Its Applications

In this section, we shall present the Gauss-Bonnet theorem and some of its consequences. The geometry involved in this theorem is fairly simple, and the difficulty of its proof lies in certain topological facts. These facts will be presented without proofs.

The Gauss-Bonnet theorem is probably the deepest theorem in the differential geometry of surfaces. A first version of this theorem was presented by Gauss in a famous paper [1] and deals with geodesic triangles on surfaces (that is, triangles whose sides are arcs of geodesics). Roughly speaking, it asserts that the excess over π of the sum of the interior angles $\varphi_1, \varphi_2, \varphi_3$ of a geodesic triangle T is equal to the integral of the Gaussian curvature K over T; that is (Fig. 4-23),

$$\sum_{i=1}^{3} \varphi_i - \pi = \iint_T K\,d\sigma.$$

For instance, if $K \equiv 0$, we obtain that $\sum \varphi_i = \pi$, an extension of Thales' theorem of high school geometry to surfaces of zero curvature. Also, if $K \equiv 1$, we obtain that $\sum \varphi_i - \pi = \text{area }(T) > 0$. Thus, on a unit sphere, the sum of the interior angles of any geodesic triangle is greater than π, and the excess over π is exactly the area of T. Similarly, on the pseudosphere (Exercise 6, Sec. 3-3) the sum of the interior angles of any geodesic triangle is smaller than π (Fig. 4-24).

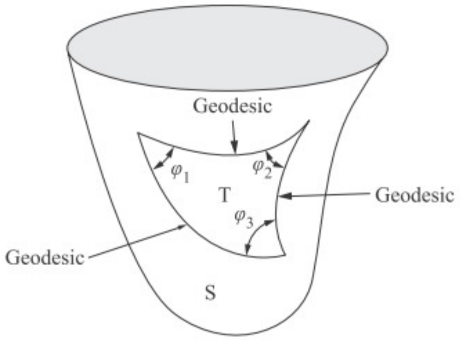
Geodesic
φ_1
φ_2
T
Geodesic
φ_3
Geodesic
S

Figure 4-23. A geodesic triangle.

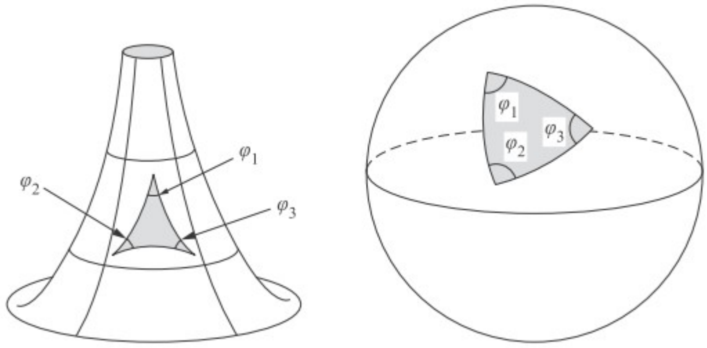
φ_1
φ_2
φ_3
φ_1
φ_2
φ_3

$K \equiv -1, \sum \varphi_i < \pi$ $K \equiv -1, \sum \varphi_i > \pi$

Figure 4-24

The extension of the theorem to a region bounded by a nongeodesic simple curve (see Eq. (1) below) is due to O. Bonnet. To extend it even further, say, to compact surfaces, some topological considerations will come into play. Actually, one of the most important features of the Gauss-Bonnet theorem is that it provides a remarkable relation between the topology of a compact surface and the integral of its curvature (see Corollary 2 below).

We shall now begin the details of a local version of the Gauss-Bonnet theorem. We need a few definitions.

Let $\alpha\colon [0, l] \to S$ be a continuous map from the closed interval $[0, l]$ into the regular surface S. We say that α is a *simple, closed, piecewise regular, parametrized curve if*

1. $\alpha(0) = \alpha(l)$.
2. $t_1 \neq t_2, t_1, t_2 \in [0, l)$, implies that $\alpha(t_1) \neq \alpha(t_2)$.

3. There exists a subdivision

$$0 = t_0 < t_1 < \cdots < t_k < t_{k+1} = l,$$

of $(0, l]$ such that α is differentiable and regular in each (t_i, t_{i+1}), $i = 0, \dots, k$.

Intuitively, this means that α is a closed curve (condition 1) without self-intersections (condition 2), which fails to have a well-defined tangent line only at a finite number of points (condition 3).

The points $\alpha(t_i)$, $i = 0, \dots, k$, are called the *vertices* of α and the traces $\alpha([t_i, t_{i+1}])$ are called the *regular arcs* of α. It is usual to call the trace $\alpha([0, l])$ of α, a *closed piecewise regular curve*.

By the condition of regularity, for each vertex $\alpha(t_i)$ there exist the limit from tile left, i.e., for $t < t_i$

$$\lim_{t \to t_i} \alpha'(t) = \alpha'(t_i - 0) \neq 0,$$

and the limit from the right, i.e., for $t > t_i$,

$$\lim_{t \to t_i} \alpha'(t) = \alpha'(t_i + 0) \neq 0.$$

Assume now that S is oriented and let $|\theta_i|$, $0 \leq |\theta_i| < \pi$, be the smallest determination of the angle from $\alpha'(t_i - 0)$ to $\alpha'(t_i + 0)$. If $|\theta_i| \neq \pi$, we give θ_i the sign of the determinant $(\alpha'(t_i - 0), \alpha'(t_i + 0), N)$. This means that if the vertex $\alpha(t_i)$ is not a "cusp" (Fig. 4-25), the sign of θ_i is given by the orientation of S. The signed angle θ_i, $-\pi < \theta_i < \pi$, is called the *external angle* at the vertex $\alpha(t_i)$.

In the case that $\alpha(t_i)$ is a cusp, i.e., $|\theta_i| = \pi$, we choose the sign of θ_i as follows. Let the closed simple curve α be contained in the image of a conformal parametrization with a given orientation, and assume that $\alpha(t_i)$ is a cusp. Choose coordinate axis $x0y$, with $\alpha(t_i) = 0$, in the given orientation,

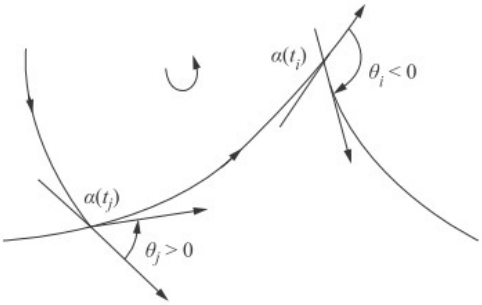

Figure 4-25

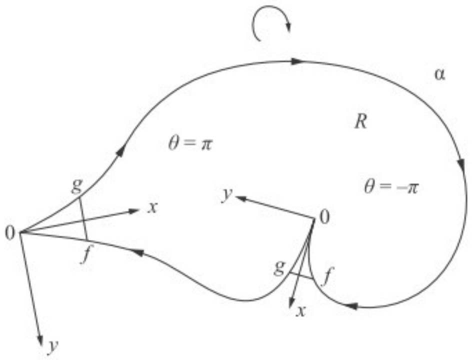

Figure 4-26. The sign of the external angle in the case of a cusp.

and assume further that the part of α arriving at $\alpha(t_i)$ is pointing towards the negative part of the axis $0x$ (and, of course, the part of α leaving $\alpha(t_i)$ is pointing to the positive part of the axis $0x$).

For small $\varepsilon > 0$, the part of α arriving at $\alpha(t_i)$, and near it, is given by a function $f(x) = y$, $0 < x < \epsilon$, and the part of α leaving t_i, and near it, is given by a function $g(x) = y$, $0 < x < \epsilon$. If $f > 0$ and $g < 0$, we set $\theta(t_i) = \pi$ and if $f < 0$ and $g > 0$, we set $\theta(t_i) = -\pi$.

This settles the question of the angle at a cusp.

Let α: $[0, l] \to \mathbf{x}(U) \subset S$ be a simple closed, piecewise regular, parametrized curve, with vertices $\alpha(t_i)$ and external angles θ_i, $i = 0, \dots, k$.

Let φ_i: $[t_i, t_{i+1}] \to R$ be differentiable functions which measure at each $t \in [t_i, t_{i+1}]$ the positive angle from $\mathbf{x}_u$ to $\alpha'(t)$ (cf. Lemma 1, Sec. 4-4).

The first topological fact that we shall present without proofs is the following.

THEOREM (of Turning Tangents). *With the above notation we have for plane curves*

$$\sum_{i=0}^{k}(\varphi_i(t_{i+1}) - \varphi_i(t_i)) + \sum_{i=0}^{k}\theta_i = \pm 2\pi,$$

where the sign plus or minus depends on the orientation of α.

The theorem states that the total variation of the angle of the tangent vector to α with a given direction plus the "jumps" at the vertices is equal to 2π.

An elegant proof of this theorem has been given by H. Hopf, *Compositio Math.* 2 (1935), 50–62. For the case where α has no vertices, Hopf's proof can be found in Sec. 5-7 (Theorem 2) of this book. It should be noted that Hopf's proof is for plane curves.

Before stating the local version of the Gauss-Bonnet theorem we still need some terminology.

Let S be an oriented surface. A region $R \subset S$ (union of a connected open set with its boundary) is called a *simple region* if R is homeomorphic to a disk and the boundary ∂R of R is the trace of a simple, closed, piecewise regular, parametrized curve $\alpha\colon I \to S$. We say then that α is *positively oriented* if for each $\alpha(t)$, belonging to a regular arc, the positive orthogonal basis $\{\alpha'(t), h(t)\}$ satisfies the condition that $h(t)$ "points toward" R; more precisely, for any curve $\beta\colon I \to R$ with $\beta(0) = \alpha(t)$ and $\beta'(0) \neq \alpha'(t)$, we have that $\langle \beta'(0), h(t)\rangle > 0$. Intuitively, this means that if one is walking on the curve α in the positive direction and with one's head pointing to N, then the region R remains to the left (Fig. 4-27). It can be shown that one of the two possible orientations of α makes it positively oriented.

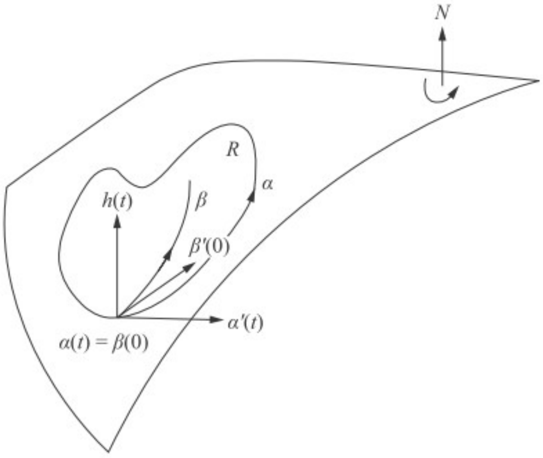

Figure 4-27. A positively oriented boundary curve.

Now let $\mathbf{x}\colon U \subset R^2 \to S$ be a parametrization of S compatible with its orientation and let $R \subset \mathbf{x}(U)$ be a bounded region of S. If f is a differentiable function on S, then it is easily seen that the integral

$$\iint_{\mathbf{x}^{-1}(R)} f(u, v)\sqrt{EG - F^2}\, du\, dv$$

does not depend on the parametrization $\mathbf{x}$, chosen in the class of orientation of $\mathbf{x}$. (The proof is the same as in the definition of area; cf. Sec. 2-5.) This integral has, therefore, a geometrical meaning and is called *the integral of f over the region R*. It is usual to denote it by

$$\iint_R f\, d\sigma.$$

With these definitions, we now state the

GAUSS-BONNET THEOREM (Local). *Let* $\mathbf{x}$: U $\to$ S *be an isothermal parametrization (i.e.,* $F = 0$, $E = G = \lambda^2(u, v)$*). See Sec. 4.2 after Prop. 2) of an oriented surface* S, *where* U $\subset$ R^2 *is homeomorphic to an open disk and* $\mathbf{x}$ *is compatible with the orientation of* S.

Let R $\subset$ $\mathbf{x}$(U) *be a simple region of* S *and let* α: $I \to S$ *be such that* ∂R $= \alpha$(I). *Assume that* α *is positively oriented, parametrized by arc length* s, *and let* $\alpha(\mathrm{s}_0), \ldots, \alpha(\mathrm{s}_\mathrm{k})$ *and* $\theta_0, \ldots, \theta_\mathrm{k}$ *be, respectively, the vertices and the external angles of* α. *Then*

$$\sum_{\mathrm{i}=0}^{\mathrm{k}} \int_{\mathrm{s}_\mathrm{i}}^{\mathrm{s}_{\mathrm{i}+1}} \mathrm{k}_\mathrm{g}(\mathrm{S})\, \mathrm{ds} + \iint_\mathrm{R} \mathrm{K}\, \mathrm{d}\sigma + \sum_{\mathrm{i}=0}^{\mathrm{k}} \theta_\mathrm{i} = 2\pi, \qquad \mathbf{(1)}$$

where $\mathrm{k}_\mathrm{g}(\mathrm{s})$ *is the geodesic curvature of the regular arcs of* α *and* K *is the Gaussian curvature of* S.

Remark. The restriction that the region *R* be contained in the image set of an isothermal parametrization is needed only to simplify the proof and to be able to use the theorem of turning tangents. As we shall see later (Corollary 1 of the global Gauss-Bonnet theorem) the above result still holds for any simple region of a regular surface. This is quite plausible, since Eq. (1) does not involve in any way a particular parametrization.[†]

Proof. Let $u = u(s)$, $v = v(s)$ be the expression of α in the parametrization $\mathbf{x}$. By using Prop. 3 of Sec. 4-4, we have

$$k_g(s) = \frac{1}{2\sqrt{EG}}\left\{G_u \frac{dv}{ds} - E_v \frac{du}{ds}\right\} + \frac{d\varphi_i}{ds},$$

where $\varphi_i = \varphi_i(s)$ is a differentiable function which measures the positive angle from $\mathbf{x}_u$ to $\alpha'(s)$ in $[s_i, s_{i+1}]$. By integrating the above expression in every interval $[s_i, s_{i+1}]$ and adding up the results,

$$\sum_{i=0}^{k} \int_{s_i}^{s_{i+1}} k_g(s)\, ds = \sum_{i=0}^{k} \int_{s_i}^{s_{i+1}} \left(\frac{G_u}{2\sqrt{EG}} \frac{dv}{ds} - \frac{E_v}{2\sqrt{EG}} \frac{du}{ds}\right) ds$$
$$+ \sum_{i=0}^{k} \int_{s_i}^{s_{i+1}} \frac{d\varphi_i}{ds}\, ds.$$

Now we use the Gauss-Green theorem in the uv plane which states the following: *If* $P(u, v)$ *and* $Q(u, v)$ *are differentiable functions in a simple region* $A \subset R^2$, *tire boundary of which is given by* $u = u(s)$, $v = v(s)$, *then*

[†] If the truth of this assertion is assumed, applications 2 and 6 given below can be presented now.

$$\sum_{i=0}^{k}\int_{s_i}^{s_{i+1}}\left(P\frac{du}{ds}+Q\frac{dv}{ds}\right)ds=\iint_A\left(\frac{\partial Q}{\partial u}-\frac{\partial P}{\partial v}\right)du\,dv.$$

It follows that

$$\sum_{i=0}^{k}\int_{s_i}^{s_{i+1}}k_g(s)\,ds=\iint_{\mathbf{x}^{-1}(R)}\left\{\left(\frac{E_v}{2\sqrt{EG}}\right)_v+\left(\frac{G_u}{2\sqrt{EG}}\right)_u\right\}du\,dv$$
$$+\sum_{i=0}^{k}\int_{s_i}^{s_{i+1}}\frac{d\varphi_i}{ds}\,ds.$$

To be able to use the theorem of turning tangents, we must assume that our parametrization is isothermal, that is, in addition to $F=0$, we have that $E=G=\lambda(u,v)>0$. In this case,

$$\begin{aligned}\left(\frac{E_v}{2\sqrt{EG}}\right)_v+\left(\frac{G_u}{2\sqrt{EG}}\right)_u&=\frac{1}{2}\left\{\left(\frac{\lambda_v}{\lambda}\right)_v+\left(\frac{\lambda_u}{\lambda}\right)_u\right\}\\&=\frac{1}{2\lambda}\{(\log\lambda)_{vv}+(\log\lambda)_{uu}\}\lambda\\&=\frac{1}{2\lambda}(\Delta\log\lambda)\lambda=-K\lambda,\end{aligned}$$

where, in the last equality we have used the Exercise 2 of Section 4-3. By integrating the above expression in the domain of the coordinate neighborhood, we obtain

$$\sum_{i=0}^{k}\int_{s_i}^{s_{i+1}}k_g(s)\,ds=-\iint_R K\lambda\,du\,dv+\sum_i\int_{s_i}^{s_{i+1}}\frac{d\varphi_i}{ds}\,ds$$

On the other hand, by the theorem of turning tangents,

$$\begin{aligned}\sum_{i=0}^{k}\int_{s_i}^{s_{i+1}}\frac{d\varphi_i}{ds}\,ds&=\sum_{i=0}^{k}(\varphi_i(s_{i+1})-\varphi_i(s_i))\\&=\pm2\pi-\sum_{i=0}^{k}\theta_i.\end{aligned}$$

Since the curve α is positively oriented, the sign should be plus, as can easily be seen in the particular case of the circle in a plane.

By putting these facts together, we obtain

$$\sum_{i=0}^{k}\int_{s_i}^{s_{i+1}}k_g(s)\,ds+\iint_R K\,d\sigma+\sum_{i=0}^{k}\theta_i=2\pi.$$ **Q.E.D.**

Before going into a global version of the Gauss-Bonnet theorem, we would like to show how the techniques used in the proof of this theorem may also be used to give an interpretation of the Gaussian curvature in terms of parallelism.

To do that, let $\mathbf{x}: U \to S$ be an isothermal parametrization at a point $p \in S$, and let $R \subset \mathbf{x}(U)$ be a simple region without vertices, containing p in its interior. Let $\alpha: [0, l] \to \mathbf{x}(U)$ be a curve parametrized by arc length s such that the trace of α is the boundary of R. Let w_0 be a unit vector tangent to S at $\alpha(0)$ and let $w(s)$, $s \in [0, l]$, be the parallel transport of w_0 along α (Fig. 4-28). By using Prop. 3 of Sec. 4-4 and the Gauss-Green theorem in the uv plane, we obtain

$$\begin{aligned} 0 &= \int_0^l \left[\frac{Dw}{ds}\right] ds \\ &= \int_0^l \frac{1}{2\sqrt{EG}} \left\{ G_u \frac{dv}{ds} - E_v \frac{du}{ds} \right\} ds + \int_0^l \frac{d\varphi}{ds}\, ds \\ &= -\iint_R K\, d\sigma + \varphi(l) - \varphi(0), \end{aligned}$$

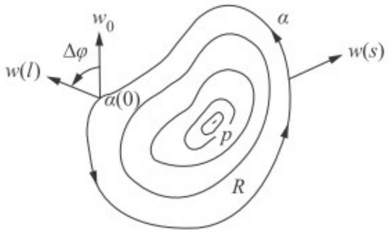

Figure 4-28

where $\varphi = \varphi(s)$ is a differentiable determination of the angle from $\mathbf{x}_u$ to $w(s)$. It follows that $\varphi(l) - \varphi(0) = \Delta\varphi$ is given by

$$\Delta\varphi = \iint_R K\, d\sigma. \tag{2}$$

Now, $\Delta\varphi$ does not depend on the choice of w_0, and it follows from the expression above that $\Delta\varphi$ does not depend on the choice of $\alpha(0)$ either. By taking the limit (in the sense of Prop. 2, Sec. 3-3)

$$\lim_{R \to p} \frac{\Delta\varphi}{A(R)} = K(p),$$

where $A(R)$ denotes the area of the region R, we obtain the desired interpretation of K.

To globalize the Gauss-Bonnet theorem, we need further topological preliminaries.

Let S be a regular surface. A connected region $R \subset S$ is said to be *regular* if R is compact and its boundary ∂R is the finite union of (simple) closed piecewise regular curves which do not intersect (the region in Fig. 4-29(a) is regular, but that in Fig. 4-29(b) is not). For convenience, we shall consider a compact surface as a regular region, the boundary of which is empty.

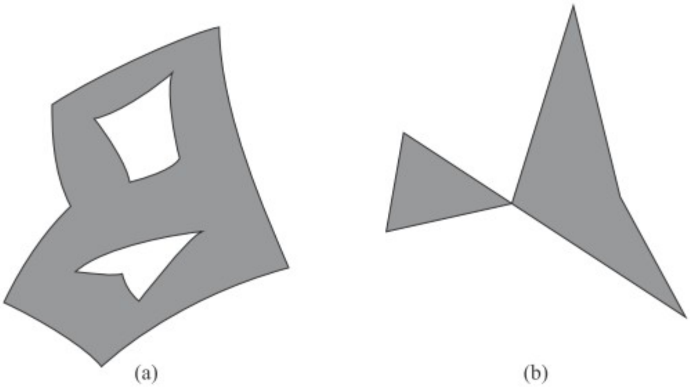

(a) (b)

Figure 4-29

A simple region which has only three vertices with external angles $\alpha_i \neq 0$, $i = l, 2, 3$, is called a *triangle*.

A *triangulation* of a regular region $R \subset S$ is a finite family $\mathfrak{J}$ of triangles $T_i, = 1, \dots, n$, such that

1. $\bigcup_{i=1}^{n} T_i = R$.
2. If $T_i \cap T_j \neq \phi$, then $T_i \cap T_j$ is either a common edge of T_i and T_j or a common vertex of T_i and T_j.

Given a triangulation $\mathfrak{J}$ of a regular region $R \subset S$ of a surface S, we shall denote by F the number of triangles (faces), by E the number of sides (edges), and by V the number of vertices of the triangulation. The number

$$F - E + V = \chi$$

is called the *Euler-Poincaré characteristic* of the triangulation.

The following propositions are presented without proofs. An exposition of these facts may be found, for instance, in L. Ahlfors and L. Sario, *Riemann Surfaces*, Princeton University Press, Princeton, N.J., 1960, Chap. 1.

PROPOSITION 1. *Every regular region of a regular surface admits a triangulation.*

% RRR

5- 662 N 3

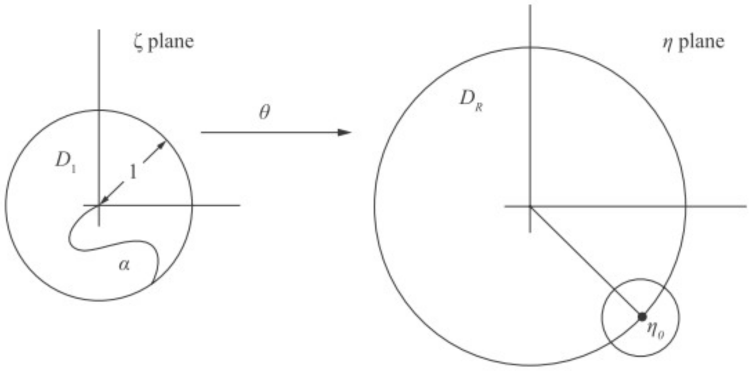

Figure 5-12

Let L be the segment of D_R that joins η_0 to 0; i.e., $L = \{t\eta_0 \in \mathbb{C}; 0 \leq t \leq 1\}$. Set $\alpha = \theta^{-1}(L)$ and show that the arc length l of $\mathbf{x}(\alpha)$ is

$$\begin{aligned} l &= \int_\alpha \sqrt{2\langle \mathbf{x}_u, \mathbf{x}_v\rangle \left\{ \left(\frac{du}{dt}\right)^2 + \left(\frac{dv}{dt}\right)^2 \right\}}\,dt \\ &\leq \frac{1}{\epsilon}\int_\alpha \sqrt{\langle \mathbf{x}_u, w\rangle^2 + \langle \mathbf{x}_v, w\rangle^2}\,|d\zeta| = \frac{1}{\epsilon}\int_\alpha |\varphi(\zeta)||d\zeta| \\ &= \frac{R}{\epsilon} < +\infty. \end{aligned}$$

Use Exercise 7 to conclude that $\mathbf{x}(D)$ is not complete.

5-4. First and Second Variations of Arc Length; Bonnet's Theorem

The goal of this section is to prove that a complete surface S with Gaussian curvature $K \geq \delta > 0$ is compact (Bonnet's theorem).

The crucial point of the proof is to show that if $K \geq \delta > 0$, a geodesic γ joining two arbitrary points $p, q \in S$ and having length $l(\gamma) > \pi/\sqrt{\delta}$ is no longer minimal; that is, there exists a parametrized curve joining p and q, the length of which is smaller than $l(\gamma)$.

Once this is proved, it follows that every minimal geodesic has length $l \leq \pi/\sqrt{\delta}$; thus, S is bounded in the distance d. Since S is complete, S is compact (Corollary 2, Sec. 5-3). We remark that, in addition, we obtain an estimate for the diameter of S, namely, $\rho(S) \leq \pi/\sqrt{\delta}$.

To prove the above point, we need to compare the arc length of a parametrized curve with the arc length of "neighboring curves." For this, we shall introduce a number of ideas which are useful in other problems of differential geometry. Actually, these ideas are adaptations to the purposes of differential geometry of more general concepts found in calculus of variations. No knowledge of calculus of variations will be assumed.

In this section, S will denote a regular (not necessarily complete) surface.

We shall begin by making precise the idea of neighboring curves of a given curve.

DEFINITION 1. *Let* α: $[0, l] \to \mathrm{S}$ *be a regular parametrized curve, where the parameter* $\mathrm{s} \in [0, l]$ *is the arc length. A variation of* α *is a differentiable map* h: $[0, l] \times (-\epsilon, \epsilon) \subset \mathrm{R}^2 \to \mathrm{S}$ *such that*

$$\mathrm{h}(\mathrm{s}, 0) = \alpha(\mathrm{s}), \quad \mathrm{s} \in (0, l].$$

For each $\mathrm{t} \subset (-\epsilon, \epsilon)$*, the curve* $\mathrm{h_t}$: $[0, l] \to \mathrm{S}$*, given by* $\mathrm{h_t}(s) = \mathrm{h}(\mathrm{s}, \mathrm{t})$*, is called a curve of the variation* h. *A variation* h *is said to be proper if*

$$\mathrm{h}(0, \mathrm{t}) = \alpha(0), \quad \mathrm{h}(l, \mathrm{t}) = \alpha(l), \quad \mathrm{t} \in (-\epsilon, \epsilon).$$

Intuitively, a variation of α is a family h_t of curves depending differentiably on a parameter $t \in (-\epsilon, \epsilon)$ and such that h_0 agrees with α (Fig. 5-13). The condition of being proper means that all curves h_t have the same initial point $\alpha(0)$ and the same end point $\alpha(l)$.

It is convenient to adopt the following notation. The parametrized curves in R^2 given by

$$s \to (s, t_0),$$
$$t \to (s_0, t),$$

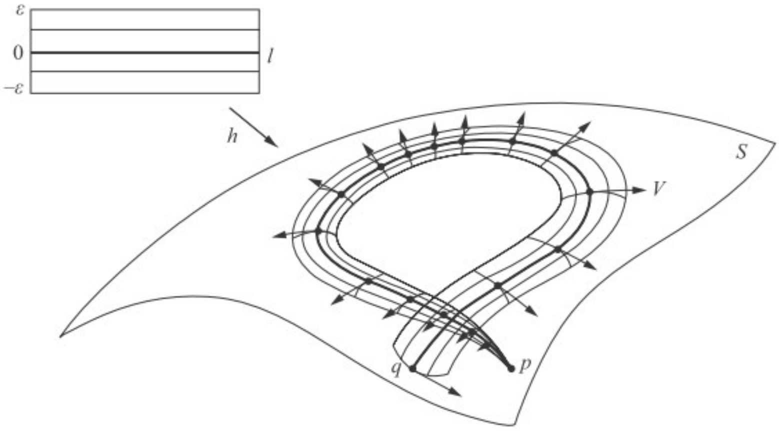

Figure 5-13

pass through the point $p_0 = (s_0, t_0) \in R^2$ and have $(1, 0)$ and $(0, 1)$ as tangent vectors at (s_0, t_0). Let h: $[0, l] \times (-\epsilon, \epsilon) \subset R^2 \to S$ be a differentiable map and let $p_0 \in [0, l] \times (-\epsilon, \epsilon)$. Then $dh_{p_0}(1, 0)$ is the tangent vector to the curve $s \to h(s, t_0)$ at $h(p_0)$, and $dh_{p_0}(0, 1)$ is the tangent vector to the curve $t \to h(s_0, t)$ at $h(p_0)$. We shall denote

$$dh_{p_0}(1, 0) = \frac{\partial h}{\partial s}(p_0),$$
$$dh_{p_0}(0, 1) = \frac{\partial h}{\partial t}(p_0).$$

We recall (cf. Sec. 4-4, Def. 3) that a vector field w along a curve α: $I \to S$ is a correspondence that assigns to each $t \in I$ a vector $w(t)$ *tangent to the surface* S at $\alpha(t)$. Thus, $\partial h/\partial s$ and $\partial h/\partial t$ are differentiable tangent vector fields along α.

It follows that a variation h of α determines a differentiable vector field $V(s)$ along α by

$$V(s) = \frac{\partial h}{\partial t}(s, 0), \quad S \in [0, l].$$

V is called the *variational vector field* of h; we remark that if h is proper, then

$$V(0) = V(l) = 0.$$

This terminology is justified by the following proposition.

PROPOSITION 1. *If we let* V(s) *be a differentiable vector field along a parametrized regular curve* α: $[0, l] \to$ S *then there exists a variation* h: $[0, l] \times (-\epsilon, \epsilon) \to$ S *of* α *such that* V(s) *is the variational vector field of* h. *Furthermore, if* $\mathrm{V}(0) = \mathrm{V}(l) = 0$*, then* h *can be chosen to be proper.*

Proof. We first show that there exists a $\delta > 0$ such that if $|v| < \delta$, $v \in T_{\alpha(s)}(S)$, then $\exp_{\alpha(s)} v$ is well defined for all $s \in [0, l]$. In fact, for each $p \in \alpha([0, l]) \subset S$ consider the neighborhood W_p (a normal neighborhood of all of its points) and the number $\delta_p > 0$ given by Prop. 1 of Sec. 4-7. The union $\bigcup_p W_p$ covers $\alpha([0, l])$ and, by compactness, a finite number of them, say, $W_1, \dots, W_n$ still covers $\alpha([0, l])$. Set $\delta = \min(\delta_1, \dots, \delta_n)$, where δ_i is the number corresponding to the neighborhood W_i, $i = 1, \dots, n$. It is easily seen that δ satisfies the above condition.

Now let $M = \max_{s \in [0,l]} |V(s)|$, $\epsilon < \delta/M$, and define

$$h(s, t) = \exp_{\alpha(t)} tV(s), \quad s \in [0, l], \quad t \in (-\epsilon, \epsilon).$$

h is clearly well defined. Furthermore, since

$$\exp_{\alpha(s)} tV(s) = \gamma(1, \alpha(s), tV(s)),$$

where γ is the (differentiable) map of Theorem 1 of Sec. 4-7 (i.e., for $t \neq 0$, and $V(s) \neq 0$, $\gamma(1, \alpha(s), tV(s))$ is the geodesic γ with initial conditions

5- < 26 <3

Remark 1. It should be noticed that the curvature condition $K \leq 0$ was used only to guarantee that $\exp_p : T_p(S) \to S$ is a length-increasing local diffeomorphism. Therefore, we have actually proved that *if* $\varphi\colon S_1 \to S_2$ *is a local diffeomorphism of a complete surface* S_1 *onto a surface* S_2, *which is length-increasing, then* φ *is a covering map.*

The following proposition, known as the Hadamard theorem, describes the topological structure of a complete surface with curvature $K \leq 0$.

THEOREM 1 (Hadamard). *Let* S *be a simply connected, complete surface, with Gaussian curvature* $K \leq 0$. *Then* $\exp_p\colon T_p(S) \to S$, $p \in S$, *is a diffeomorphism; that is,* S *is diffeomorphic to a plane.*

Proof. By Prop. 7, $\exp_p : T_p(S) \to S$ is a covering map. By the corollary of Prop. 5, $\exp_p$ is a homeomorphism. Since $\exp_p$ is a local diffeomorphism, its inverse map is differentiable, and $\exp_p$ is a diffeomorphism. **Q.E.D.**

We shall now present another geometric application of the covering spaces, also known as the Hadamard theorem. Recall that a connected, compact, regular surface, with Gaussian curvature $K > 0$, is called an ovaloid (cf. Remark 2, Sec. 5-2).

THEOREM 2 (Hadamard). *Let* S *be an ovaloid. Then the Gauss map* $N\colon S \to S^2$ *is a dijfepmorphism. In particular,* S *is diffeomorphic to a sphere.*

Proof. Since for every $p \in S$ the Gaussian curvature of S, $K = \det(dN_p)$, is positive, N is a local diffeomorphism. By Prop. 1, N is a covering map. Since the sphere S^2 is simply connected, we conclude from the corollary of Prop. 5 that $N\colon S \to S^2$ is a homeomorphism of S onto the unit sphere S^2. Since N is a local diffeomorphism, its inverse map is differentiable. Therefore, N is a diffeomorphism. **Q.E.D.**

Remark 2. Actually, we have proved somewhat more. Since the Gauss map N is a diffeomorphism, each unit vector v of R^3 appears exactly once as a unit normal vector to S. Taking a plane normal to v, away from the surface, and displacing it parallel to itself until it meets the surface, we conclude that S lies on one side of each of its tangent planes. This is expressed by saying that an ovaloid S is *locally convex*. It can be proved from this that S is actually the boundary of a convex set (that is, a set $K \subset R^3$ such that the line segment joining any two points $p, q \in K$ belongs entirely to K).

Remark 3. The fact that compact surfaces with $K > 0$ are homeomorphic to spheres was extended to compact surfaces with $K \geq 0$ by S. S. Chern and R. K. Lashof ("On the Total Curvature of Immersed Manifolds," *Michigan Math. J.* 5 (1958), 5–12). A generalization for complete surfaces was first obtained by J. J. Stoker ("Über die Gestalt der positiv gekrümnten offenen

Fläche," *Compositio Math.* 3 (1936), 58–89), who proved, among other things, the following: *A complete surface with $K > 0$ is homeomorphic to a sphere or a plane*. This result still holds for $K \geq 0$ if one assumes that at some point $K > 0$ (for a proof and a survey of this problem, see M. do Carmo and E. Lima, "Isometric Immersions with Non-negative Sectional Curvatures," *Boletim da Soc. Bras. Mat.* 2 (1971), 9–22).

EXERCISES

1. Show that the map $\pi\colon R \to S^1 = \{(x, y) \in R^2; x^2 + y^2 = 1\}$ that is given by $\pi(t) = (\cos t, \sin t)$, $t \in R$, is a covering map.

2. Show that the map $\pi\colon R^2 - \{0, 0\} \to R^2 - \{0, 0\}$ given by

$$\pi(x, y) = (x^2 - y^2, 2xy), \quad (x, y) \in R^2,$$

is a two-sheeted covering map.

3. Let S be the helicoid generated by the normals to the helix $(\cos t, \sin t, bt)$. Denote by L the z axis and let $\pi\colon S - L \to R^2 - \{0, 0\}$ be the projection $\pi(x, y, z) = (x, y)$. Show that π is a covering map.

4. Those who are familiar with functions of a complex variable will have noticed that the map π in Exercise 2 is nothing but the map $\pi(z) = z^2$ from $\mathbb{C} - \{0\}$ onto $\mathbb{C} - \{0\}$; here $\mathbb{C}$ is the complex plane and $z \in \mathbb{C}$. Generalize that by proving that the map $\pi\colon \mathbb{C} - \{0\} \to \mathbb{C} - \{0\}$ given by $\pi(z) = z^n$ is an n-sheeted covering map.

5. Let $B \subset R^3$ be an arcwise connected set. Show that the following two properties are equivalent (cf. Def. 3):

1. For any pair of points $p, q \in B$ and any pair of arcs $\alpha_0\colon [0, l] \to B$, $\alpha_1\colon [0, l] \to B$, there exists a homotopy in B joining α_0 to α_1.
2. For any $p \in B$ and any arc $\alpha\colon [0, l] \to B$ with $\alpha(0) = \alpha(l) = p$ (that is, α is a closed arc with initial and end point p) there exists a homotopy joining α to the constant arc $\alpha(s) = p$, $s \in [0, l]$.

6. Fix a point $p_0 \in R^2$ and define a family of maps $\varphi_t\colon R^2 \to R^2$, $t \in [0, 1]$, by $\varphi_t(p) = tp_0 + (1 - t)p$, $p \in R^2$. Notice that $\varphi_0(p) = p$, $\varphi_1(p) = p_0$. Thus, φ_t is a continuous family of maps which starts with the identity map and ends with the constant map p_0. Apply these considerations to prove that R^2 is simply connected.

7. a. Use stereographic projection and Exercise 6 to show that any closed arc on a sphere S^2 which omits at least one point of S^2 is homotopic to a constant arc.

b. Show that any closed arc on S^2 is homotopic to a closed arc in S^2 which omits at least one point.

5- 6 626 3

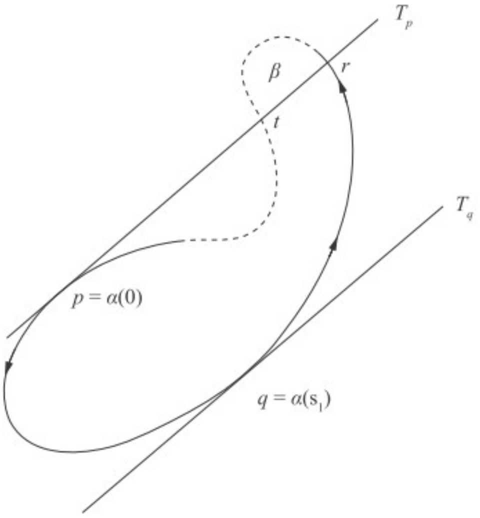

Figure 5-42

5-8. Surfaces of Zero Gaussian Curvature

We have already seen (Sec. 4-6) that the regular surfaces with identically zero Gaussian curvature are locally isometric to the plane. In this section, we shall look upon such surfaces from the point of view of their position in R^3 and prove the following global theorem.

THEOREM. *Let* $S \subset R^3$ *be a complete surface with zero Gaussian curvature. Then* S *is a cylinder or a plane.*

By definition, a *cylinder* is a regular surface S such that through each point $p \in S$ there passes a unique line $R(p) \subset S$ (the generator through p) which satisfies the condition that if $q \neq p$, then the lines $R(p)$ and $R(q)$ are parallel or equal.

It is a strange fact in the history of differential geometry that such a theorem was proved only somewhat late in its development. The first proof came as a corollary of a theorem of P. Hartman and L. Nirenberg ("On Spherical Images Whose Jacobians Do Not Change Signs," *Amer. J. Math.* 81 (1959), 901–920) dealing with a situation much more general than ours. Later, W. S. Massey ("Surfaces of Gaussian Curvature Zero in Euclidean Space," *Tohoku Math. J.* 14 (1962), 73–79) and J. J. Stoker ("Developable Surfaces in the Large," *Comm. Pure and Appl. Math.* 14 (1961), 627–635) obtained elementary and direct proofs of the theorem. The proof we present here is a modification of

Massey's proof. It should be remarked that Stoker's paper contains a slightly more general theorem.

(Added in 2016) Professor I. Sabitov informed me that this result was obtained by A.V. Pogorelov in 1956. It appeared under the title Extensions of the theorem of Gauss on spherical representation, in *Dokl. Akad. Nauk. S.S.S.R. (N.S.)* 111 (1956), 945–947. MR0087147.

We shall start with the study of some local properties of a surface of zero curvature.

Let $S \subset R^3$ be a regular surface with Gaussian curvature $K \equiv 0$. Since $K = k_1 k_2$, where k_1 and k_2 are the principal curvatures, the points of S are either parabolic or planar points. We denote by P the set of planar points and by $U = S - P$ the set of parabolic points of S.

P is closed in S. In fact, the points of P satisfy the condition that the mean curvature $H = \frac{1}{2}(k_1 + k_2)$ is zero. A point of accumulation of P has, by continuity of H, zero mean curvature; hence, it belongs to P. It follows that $U = S - P$ is open in S.

An instructive example of the relations between the sets P and U is given in the following example.

Example 1. Consider the open triangle *ABC* and add to each side a cylindrical surface, with generators parallel to the given side (see Fig. 5-43). It is possible to make this construction in such a way that the resulting surface is a regular surface. For instance, to ensure regularity along the open segment *BC*, it suffices that the section *FG* of the cylindrical band *BCDE* by a plane normal to *BC* is a curve of the form

$$\exp\left(-\frac{1}{x^2}\right).$$

Observe that the vertices A, B, C of the triangle and the edges *BE*, *CD*, etc., of the cylindrical bands do not belong to S.

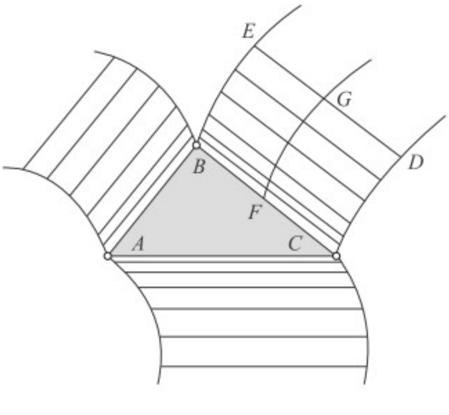

Figure 5-43

5- 6 26 T3

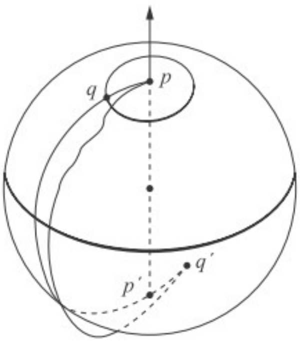

Figure 5-44

starting from a point p minimizes arc length, relative to "neighboring" curves of γ, only up to the "first" conjugate point of p relative to γ (more precise statements will be given later; see Theorems 1 and 2).

For simplicity, the surfaces in this section are assumed to be complete and the geodesics are parametrized by arc length.

We need some preliminary results.

The following lemma shows that the image by $\exp_p: T_p(S) \to S$ of a segment of line of $T_p(S)$ with origin at $O \in T_p(S)$ (geodesic starting from p) is minimal relative to the images by $\exp_p$ of curves of $T_p(S)$ which join the extremities of this segment.

More precisely, let

$$p \in S, \quad u \in T_p(S), \quad l = |u| \neq 0,$$

and let $\tilde{\gamma}: [0, l] \to T_p(S)$ be the line of $T_p(S)$ given by

$$\tilde{\gamma}(s) = sv, \quad s \in [0, l], \quad v = \frac{u}{|u|}.$$

Let $\tilde{\alpha}: [0, l] \to T_p(S)$ be a differentiable parametrized curve of $T_p(S)$, with $\tilde{\alpha}(0) = 0$, $\tilde{\alpha}(l) = u$, and $\tilde{\alpha}(s) \neq 0$ if $s \neq 0$. Furthermore, let (Fig. 5-45)

$$\alpha(s) = \exp_p \tilde{\alpha}(s) \quad \text{and} \quad \gamma(s) = \exp_p \tilde{\gamma}(s).$$

LEMMA 1. *With the above notation, we have*

1. $l(\alpha) \geq l(\gamma)$, *where* $l(\ )$ *denotes the arc length of the corresponding curve.*

In addition, if $\tilde{\alpha}(\mathrm{s})$ *is not a critical point of* $\exp_{\mathrm{p}}$, $\mathrm{s} \in [0, l)]$, *and if the traces of* α *and* γ *are distinct, then*

2. $l(\alpha) > l(\gamma)$.

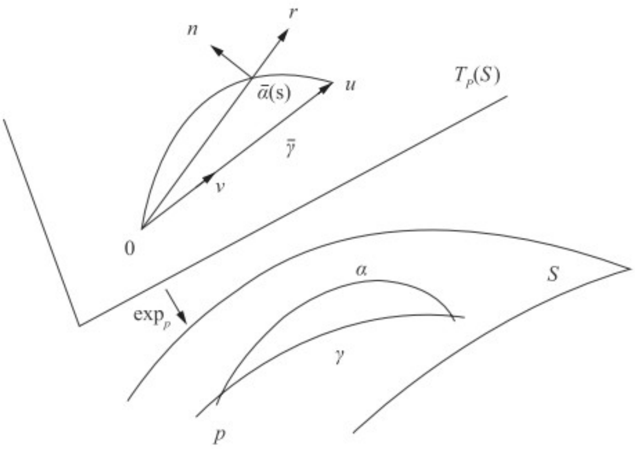

Figure 5-45

Proof. Let $\tilde{\alpha}(s)/|\tilde{\alpha}(s)| = r$, and let n be a unit vector of $T_p(S)$, with $\langle r, n\rangle = 0$. In the basis $\{r, n\}$ of $T_p(S)$ we can write (Fig. 5-45)

$$\tilde{\alpha}'(s) = ar + bn,$$

where

$$\begin{aligned} a &= \langle \tilde{\alpha}'(s), r\rangle, \\ b &= \langle \tilde{\alpha}'(s), n\rangle. \end{aligned}$$

By definition

$$\begin{aligned} \alpha'(s) &= (d\exp_p)_{\tilde{\alpha}(s)}(\tilde{\alpha}'(s)) \\ &= a(d\exp_p)_{\tilde{\alpha}(s)}(r) + b(d\exp_p)_{\tilde{\alpha}(s)}(n). \end{aligned}$$

Therefore, by using the Gauss lemma (cf. Sec. 5-5, Lemma 2) we obtain

$$\langle \alpha'(s), \alpha'(s)\rangle = a^2 + c^2,$$

where

$$c^2 = b^2|(d\exp_p)_{\tilde{\alpha}(s)}(n)|^2.$$

It follows that

$$\langle \alpha'(s), \alpha'(s)\rangle \geq a^2.$$

On the other hand,

$$\frac{d}{ds}\langle \tilde{\alpha}(s), \tilde{\alpha}(s)\rangle^{1/2} = \frac{\langle \tilde{\alpha}'(s), \tilde{\alpha}(s)\rangle}{\langle \tilde{\alpha}(s), \tilde{\alpha}(s)\rangle^{1/2}} = \langle \tilde{\alpha}'(s), r\rangle = a.$$

Therefore,

$$\begin{aligned} l(\alpha) &= \int_0^l \langle \alpha'(s), \alpha'(s)\rangle^{1/2}\, ds \geq \int_0^l a\, ds \\ &= \int_0^l \frac{d}{ds}\langle \tilde{\alpha}(s), \tilde{\alpha}(s)\rangle^{1/2}\, ds = |\tilde{\alpha}(l)| = l = l(\gamma), \end{aligned}$$

and this proves part 1.

To prove part 2, let us assume that $l(\alpha) = l(\gamma)$. Then

$$\int_0^l \langle \alpha'(s), \alpha'(s)\rangle^{1/2}\, ds = \int_0^l a\, ds,$$

and since

$$\langle \alpha'(s), \alpha'(s)\rangle^{1/2} \geq a,$$

the equality must hold in the last expression for every $s \in [0, l]$. Therefore,

$$c = |b||(d\exp_p)_{\tilde{\alpha}(s)}(n)| = 0.$$

Since $\tilde{\alpha}(s)$ is not a critical point of $\exp_p$, we conclude that $b \equiv 0$. It follows that the tangent lines to the curve $\tilde{\alpha}$ all pass through the origin O of $T_p(s)$. Thus, $\tilde{\alpha}$ is a line of $T_p(S)$ which passes through O. Since $\tilde{\alpha}(l) = \tilde{\gamma}(l)$, the lines $\tilde{\alpha}$ and $\tilde{\gamma}$ coincide, thus contradicting the assumption that the traces of α and γ are distinct. From this contradiction it follows that $l(\alpha) > l(\gamma)$, which proves part 2 and ends the proof of the lemma. **Q.E.D.**

We are now in a position to prove that if a geodesic arc contains no conjugate points, it yields a local minimum for the arc length. More precisely, we have

THEOREM 1 (Jacobi). *Let* $\gamma\colon [0, l] \to \mathrm{S}$, $\gamma(0) = \mathrm{p}$, *be a geodesic without conjugate points; that is,* $exp_{\mathrm{p}}\colon \mathrm{T_p(S)} \to \mathrm{S}$ *is regular at the points of the line* $\tilde{\gamma}(\mathrm{s}) = \mathrm{s}\gamma'(0)$ *of* $\mathrm{T_p(S)}$, $\mathrm{s} \in [0, l]$. *Let* $\mathrm{h}\colon [0, l] \times (-\epsilon, \epsilon) \to \mathrm{S}$ *be a proper variation of* γ. *Then*

1. *There exists a* $\delta > 0$, $\delta \leq \epsilon$, *such that if* $\mathrm{t} \in (-\delta, \delta)$,

$$\mathrm{L(t)} \geq \mathrm{L(0)},$$

where $\mathrm{L(t)}$ *is the length of the curve* $\mathrm{h_t}\colon [0, l] \to \mathrm{S}$ *that is given by* $\mathrm{h_t(s)} = \mathrm{h(s, t)}$.

2. *If, in addition, the trace of* $\mathrm{h_t}$, *is distinct from the trace of* γ, $\mathrm{L(t)} > \mathrm{L(0)}$.

Proof. The proof consists essentially of showing that it is possible, for every $t \in (-\delta, \delta)$, to lift the curve h_t into a curve $\tilde{h}_t$ of $T_p(S)$ such that $\tilde{h}_t(0) = 0$, $\tilde{h}_t(l) = \tilde{\gamma}(l)$ and then to apply Lemma 1.

5- 6/ 26N73
/ 1 9 A #
1 9 A) 9))
#1@,(@ " 4 -)) -) %
%)) 9) + (
( ) ' 1 9 A) P) ! 1 9 8 Ga9 . b
% ') *) L c -8H #) ") 2 < 38
T7 << " 4) - % +) 9+ @ 1 - 8 G
) d 9) -) % 8H (")) 6 2 < 38
76 76 # 4) 4 1 9 A -)) #
) 4)) +)) 1 9 A 4)4 \ - 9) 4) 8 % 8
9)) + D)) ' S " ') 4 -
' 4

Show that the curvature of this metric is given by

$$\frac{\partial^2(\sqrt{G})}{\partial x^2} + K(x, y)\sqrt{G} = 0. \tag{$*$}$$

b. Conversely, given a function $K(x, y)$ on R^2, regard y as a parameter and let $\sqrt{G}$ be the solution of $(*)$ with the initial conditions

$$\sqrt{G}(x_0, y) = 1, \quad \frac{\partial\sqrt{G}}{\partial x}(x_0, y) = 0.$$

Prove that G is positive in a neighborhood of (x_0, y) and thus defines a metric in this neighborhood. This shows that *every differentiable function is locally the curvature of some (abstract) metric.*

***c.** Assume that $K(x, y) \leq 0$ for all $(x, y) \in R^2$. Show that the solution of part b satisfies

$$\sqrt{G(x, y)} \geq \sqrt{G(x_0, y)} = 1 \quad \text{for all } x.$$

Thus, $G(x, y)$ defines a metric on all of R^2. Prove also that this metric is complete. This shows that *any nonpositive differentiable function on* R^2 *is the curvature of some complete metric on* R^2. If we do not insist on the metric being complete, the result is true for any differentiable function K on R^2. Compare J. Kazdan and F. Warner, "Curvature Functions for Open 2-Manifolds," *Ann. of Math.* 99 (1974), 203–219, where it is also proved that the condition on K given in Exercise 2 of Sec. 5-4 is necessary and sufficient for the metric to be complete.

5-11. Hilbert's Theorem

Hilbert's theorem can be stated as follows.

THEOREM. *A complete geometric surface* S *with constant negative curvature cannot be isometrically immersed in* R^3.

Remark 1. Hilbert's theorem was first treated in D. Hilbert, "Über Flächen von konstanter Gausscher Krümung," *Trans. Amer. Math. Soc.* 2 (1901), 87–99. A different proof was given shortly after by E. Holmgren, "Sur les surfaces à courbure constante negative," *C. R. Acad. Sci. Paris* 134 (1902), 740–743. The proof we shall present here follows Hilbert's original ideas. The local part is essentially the same as in Hilbert's paper; the global part, however, is substantially different. We want to thank J. A. Scheinkman for helping us to work out this proof.

We shall start with some observations. By multiplying the inner product by a constant factor, we may assume that the curvature $K \equiv -1$. Moreover, since $\exp_p : T_p(S) \to S$ is a local diffeomorphism (corollary of the theorem of Sec. 5-5), it induces an inner product in $T_p(S)$. Denote by S' the geometric surface $T_p(S)$ with this inner product. If $\psi: S \to R^3$ is an isometric immersion, the same holds for $\varphi = \psi \circ \exp_p : S' \to R^3$. Thus, we are reduced to proving that there exists no isometric immersion $\varphi: S' - R^3$ of a plane S' with an inner product such that $K \equiv -1$.

LEMMA 1. *The area of* S′ *is infinite.*

Proof. We shall prove that S' is (globally) isometric to the hyperbolic plane H. Since the area of the latter is (cf. Example 3, Sec. 5-10)

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^u\, du\, dv = \infty,$$

this will prove the lemma.

Let $p \in H$, $p' \in S'$, and choose a linear isometry ψ: $T_p(H) \to T_{p'}(S')$ between their tangent spaces. Define a map φ: $H \to S'$ by $\varphi = \exp_p \circ \psi \circ \exp_p^{-1}$. Since each point of H is joined to p by a unique minimal geodesic, φ is well defined.

We now use polar coordinates (p, θ) and (p', θ') around p and p', respectively, requiring that φ maps the axis $\theta = 0$ into the axis $\theta' = 0$. By the results of Sec. 4-6, φ preserves the first fundamental form; hence, it is locally an isometry. By using the remark made after Hadamard's theorem, we conclude that φ is a covering map. Since S' is simply connected, φ is a homeomorphism, and hence a (global) isometry. **Q.E.D.**

For the rest of this section we shall assume that there exists an isometric immersion φ: $S' \to R^3$, where S' is a geometric surface homeomorphic to a plane and with $K \equiv -1$.

To avoid the difficulties associated with possible self-intersections of $\varphi(S')$, we shall work with S' and use the immersion φ to induce on S' the local extrinsic geometry of $\varphi(S') \subset R^3$. More precisely, since φ is an immersion, for each $p \in S'$ there exists a neighborhood $V' \subset S'$ of p such that the restriction $\varphi|V' = \tilde{\varphi}$ is a diffeomorphism. At each $\tilde{\varphi}(q) \in \tilde{\varphi}(V')$, there exist, for instance, two asymptotic directions. Through $\tilde{\varphi}$, these directions induce two directions at $q \in S'$, which will be called *the asymptotic directions on* S' at q. In this way, it makes sense to talk about asymptotic curves on S', and the *same procedure can be applied to any other local entity of* $\varphi(S')$.

We now recall that the coordinate curves of a parametrization constitute a *Tchebyshef net* if the opposite sides of any quadrilateral formed by them have equal length (cf. Exercise 7, Sec. 2-5). If this is the case, it is possible to reparametrize the coordinate neighborhood in such a way that $E = 1$,

$F = \cos\theta$, $G = 1$, where θ is the angle formed by the coordinate curves, (Sec. 2-5, Exercise 8). Furthermore, in this situation, $K = -(\theta_{uv}/\sin\theta)$ (Sec. 4-3, Exercise 5).

LEMMA 2. *For each* $p \in S'$ *there is a parametrization* $\mathbf{x}: U \subset R^2 \to S'$, $p \in \mathbf{x}(U)$*, such that the coordinate curves of* $\mathbf{x}$ *are the asymptotic curves of* $\mathbf{x}(U) = V'$ *and form a Tchebyshef net* (we shall express this by saying that the asymptotic curves of V' form a Tchebyshef net).

Proof. Since $K < 0$, a neighborhood $V' \subset S'$ of p can be parametrized by $\mathbf{x}(u, v)$ in such a way that the coordinate curves of $\mathbf{x}$ are the asymptotic curves of V'. Thus, if e, f, and g are the coefficients of the second fundamental form of S' in this parametrization, we have $e = g = 0$. Notice that we are using the above convention of referring to the second fundamental form of S' rather than the second fundamental form of $\varphi(S') \subset R^3$.

Now $\varphi(V') \subset R^3$, we have

$$N_u \wedge N_v = K(\mathbf{x}_u \wedge \mathbf{x}_v);$$

hence, setting $D = \sqrt{EG - F^2}$,

$$(N \wedge N_v)_u - (N \wedge N_u)_v = 2(N_u \wedge N_v) = 2KDN.$$

Furthermore,

$$\begin{aligned} N \wedge N_u &= \frac{1}{D}\{(\mathbf{x}_u \wedge \mathbf{x}_v) \wedge N_u\} = \frac{1}{D}\{\langle \mathbf{x}_u, N_u\rangle \mathbf{x}_v - \langle \mathbf{x}_v, N_u\rangle \mathbf{x}_u\} \\ &= \frac{1}{D}(f\mathbf{x}_u - e\mathbf{x}_v), \end{aligned}$$

and, similarly,

$$N \wedge N_v = \frac{1}{D}(g\mathbf{x}_u - f\mathbf{x}_v).$$

Since $K = -1 = -(f^2/D^2)$ and $e = g = 0$, we obtain

$$N \wedge N_u = \pm\mathbf{x}_u, \quad N \wedge N_v = \pm\mathbf{x}_v;$$

hence

$$2KDN = -2DN = \pm\mathbf{x}_{uv} \pm \mathbf{x}_{vu} = \pm 2\mathbf{x}_{uv}.$$

It follows that $\mathbf{x}_{uv}$ is parallel to N; hence, $E_v = 2\langle \mathbf{x}_{uv}, \mathbf{x}_u\rangle = 0$ and $G_u = 2\langle \mathbf{x}_{uv}, \mathbf{x}_v\rangle = 0$. But $E_v = G_u = 0$ implies (Sec. 2-5, Exercise 7) that the coordinate curves form a Tchebysbef net. **Q.E.D.**

LEMMA 3. *Let* $V' \subset S'$ *be a coordinate neighborhood of* S' *such that the coordinate curves are the asymptotic curves in* V'*. Then the area* A *of any quadrilateral formed by the coordinate curves is smaller than* 2π.

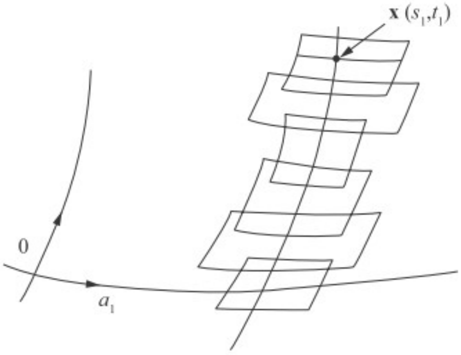

Figure 5-56

LEMMA 6. **x** *is surjective.*

Proof. Let $Q = \mathbf{x}(R^2)$. Since **x** is a local diffeomorphism, Q is open in S'. We also remark that if $p' = \mathbf{x}(s_0, t_0)$, then the two asymptotic curves which pass through p' are entirely contained in Q.

Let us assume that $Q \neq S'$. Since S' is connected, the boundary Bd $Q \neq \phi$. Let $p \in$ Bd Q. Since Q is open in S', $p \notin Q$. Now consider a rectangular neighborhood R of p in which the asymptotic curves form a Tchebyshef net (Fig. 5-57). Let $q \in Q \cap R$. Then one of the asymptotic curves through q intersects one of the asymptotic curves through p. By the above remark, this is a contradiction. **Q.E.D.**

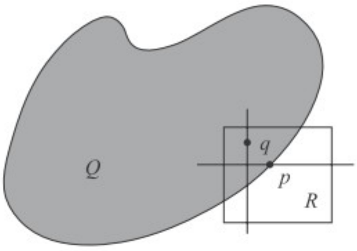

Figure 5-57

We now claim that **x** is a global diffeomorphism. Since **x** is a surjective local diffeomorphism, it suffices to show x has the property of lifting arcs and apply Prop. 6 of Sec. 5-6 (Covering Spaces...) to conclude that **x** is a covering map. Since R^2 is simply connected, this will prove our claim.

To show that **x** has the property of lifting arcs we use a Proposition that is presented here without proof (For a proof, see Elon Lima, Fundamental Groups and Covering Spaces, A K Peters, Natick, Massachusetts, see p. 144): ... *Let* **x**: $\mathrm{R}^2 \to \mathrm{S}'$ *be a surjective local diffeomorphism. Assume that* **x** *is a closed map (i.e., the image of a closed set is closed). Then* **x** *has the property of lifting arcs*. That **x** is closed follows from the way it is defined.

This completes the proof of our claim.

The proof of Hilbert's theorem now follows easily.

Proof of the Theorem. Assume the existence of an isometric immersion $\psi\colon S \to R^3$, where S is a complete surface with $K \equiv -1$. Let $p \in S$ and denote by S' the tangent plane $T_p(S)$ endowed with the metric induced by $\exp_p\colon T_p(S) \to S$. Then $\varphi = \psi \circ \exp_p\colon S' \to R^3$ is an isometric immersion and Lemmas 5 and 6 plus the text after them show the existence of a parametrization $\mathbf{x}\colon R^2 \to S'$ of the entire S' such that the coordinate curves of $\mathbf{x}$ are the asymptotic curves of S' (Lemma 4). Thus, we can cover S' by a union of "coordinate quadrilaterals" Q_n, with $Q_n \subset Q_{n+1}$. By Lemma 3, the area of each Q_n is smaller than 2π. On the other hand, by Lemma 1, the area of S' is unbounded. This is a contradiction and concludes the proof. **Q.E.D.**

Remark 2. Hilbert's theorem was generalized by N. V. Efimov, "Appearance of Singularities on Surfaces of Negative Curvature," *Math. Sb.* 106 (1954). A.M.S. Translations. Series 2, Vol. 66, 1968, 154–190, who proved the following conjecture of Cohn-Vossen: *Let* S *be a complete surface with curvature* K *satisfying* $K \le \delta < 0$. *Then there exists no isometric immersion of* S *into* R^3. Efimov's proof is very long, and a shorter proof would be desirable.

An excellent exposition of Efimov's proof can be found in a paper by T. Klotz Milnor, "Efimov's Theorem About Complete Immersed Surfaces of Negative Curvature," *Advances in Mathematics* 8 (1972), 474–543. This paper also contains another proof of Hilbert's theorem which holds for surfaces of class C^2.

For further details on immersion of the hyperbolic plane see M. L. Cromov and V. A. Rokblin, "Embeddings and Immersions in Riemannian Geometry," *Russian Math. Sureys* (1970), 1–57, especially p. 15.

EXERCISES

1. (*Stoke's Remark.*) Let S be a complete geometric surface. Assume that the Gaussian curvature K satisfies $K \le \delta < 0$. Show that there is no isometric immersion $\varphi\colon S \to R^3$ such that the absolute value of the mean curvature H is bounded. This proves Efimov's theorem quoted in Remark 2 with the additional condition on the mean curvature. The following outline may be useful:

a. Assume such a φ exists and consider the Gauss map $N\colon \varphi(S) \subset R^3 \to S^2$, where S^2 is the unit sphere. Since $K \neq 0$ everywhere, N induces a new metric (,) on S by requiring that $N \circ \varphi\colon S \to S^2$ be a local isometry. Choose coordinates on S so that the images by φ of the coordinate curves are lines of curvature of $\varphi(S)$. Show that the coefficients of the new metric in this coordinate system are

$$g_{11} = (k_1)^2 E, \quad g_{12} = 0, \quad g_{22} = (k_2)^2 G,$$

!"#" $ % % & ' ( )) * +) , -)
.) !) * + %)) 4

,9 *) /

5- 2 3

One of the surprising results about linking is Gauss's formula for the linking number between two curves. Though linking and the linking number are obviously global properties of the curves, Gauss's formula calculates the linking number in terms of the parametrizations of the curves, thereby connecting the global property to local properties.

Theorem 4.3.4 (Gauss's Linking Formula). *Let C_1 and C_2 be two closed, regular, space curves parametrized by $\vec{\alpha} : I \to \mathbb{R}^3$ and $\vec{\beta} : J \to \mathbb{R}^3$, respectively. The linking number between C_1 and C_2 is*

$$\operatorname{link}(C_1, C_2) = \frac{1}{4\pi} \int_I \int_J \frac{\det(\vec{\alpha}(u) - \vec{\beta}(v), \vec{\alpha}'(u), \vec{\beta}'(v))}{\|\vec{\alpha}(u) - \vec{\beta}(v)\|^3}\, du\, dv. \quad (4.8)$$

Though we cannot give a proof for this formula at this time, we briefly sketch a few of the concepts and techniques that go into it.

In Section 2.2, we introduced the notion of degree for a continuous map between circles, $f : \mathbb{S}^1 \to \mathbb{S}^1$. Intuitively speaking, the degree of f counts how many times f covers its codomain and with what orientation. The degree of f takes into account when f might double back. For example, if $q \in \mathbb{S}^1$ has six preimages from f and, for four of the preimages, f pass through q with the same orientation as on the domain and, for the other two preimages, f passes through q in an opposite orientation, then the degree of f is $4 - 2 = 2$.

Similarly, in the field of algebraic topology, a topic that is outside the scope of this book, one can define the degree of a map between spheres $f : \mathbb{S}^2 \to \mathbb{S}^2$ as how often f covers $\mathbb{S}^2$. Again, one should note the difficulty inherent in making this definition precise since one must take into account a form of orientation and doubling back, i.e., whether f folds back over itself over some region of the codomain. In the same way, one can define the degree of a continuous map $F : S \to \mathbb{S}^2$, where S is a regular surface without boundary. (We give the definition for a regular surface in Chapter 5.) The conditions that S is regular and has no boundary guarantees that F generically covers $\mathbb{S}^2$ by the same amount at all its points, so long as we take into account orientation and assign a negative covering value to F when F covers $\mathbb{S}^2$ negatively.

As discussed in Definition 2.2.5 and in the following, it is more appropriate to view the parametrization of a simple closed curve C

5- N 2 7T3

6.2 The Gauss Map

In Section 2.2, given any closed, simple, regular curve $\gamma : I \to \mathbb{R}^2$ in the plane, we defined the tangential indicatrix to be the curve given by $\vec{T} : I \to \mathbb{R}^2$, the unit tangent vector of $\gamma(t)$. The image of $\vec{T}$ lies on the unit circle, and though $\vec{T}$ is a closed curve, it need not be regular as its locus may stop and double back. Regardless of the parametrization, the tangential indicatrix is well defined up to a change in sign.

On any regular surface S in $\mathbb{R}^3$, the tangent plane T_pS at a point $p \in S$ is a two-dimensional subspace of $\mathbb{R}^2$, and hence the vectors that are normal to S at p form a one-dimensional subspace. Thus, there exist exactly two possible choices for a unit normal vector. Proposition 5.5.4 implies that if the surface is orientable, we can specify its orientation by a continuous function $n : S \to \mathbb{S}^2$, where $\mathbb{S}^2$ means the unit sphere in $\mathbb{R}^3$.

Definition 6.2.1. Let S be an oriented regular surface in $\mathbb{R}^3$ with an orientation $n : S \to \mathbb{S}^2$. In the classical theory of surfaces, the function n is also called the *Gauss map*.

It is important to remain aware of the distinction between the functions n and $\vec{N}$. Let S be an oriented surface with orientation $n : S \to \mathbb{S}^2$. Suppose that $\vec{X} : U \to \mathbb{R}^3$, where U is an open subset of $\mathbb{R}^2$, parametrizes a neighborhood of S, then the function $\vec{N} : U \to \mathbb{S}^2$ is defined in reference to $\vec{X}$ by Equation (5.9). The functions n and $\vec{N}$ are related by the fact that if $\vec{X}$ is a positively oriented parametrization, then

$$\vec{N} = n \circ \vec{X}.$$

Example 6.2.2. Consider as an example a sphere S in $\mathbb{R}^3$ equipped with the orientation n, with the unit vectors normal to S pointing away from the center $\vec{c}$ of the sphere. The sphere can be given as the solution to the equation

$$\|\vec{x} - \vec{c}\|^2 = R^2. \tag{6.10}$$

Recall that $\|\vec{v}\|^2 = \vec{v} \cdot \vec{v}$. If $\vec{x}(t)$ is any curve on the sphere, then by differentiating the relationship in Equation (6.10), one obtains

$$\vec{x}'(t) \cdot (\vec{x}(t) - \vec{c}) = 0.$$

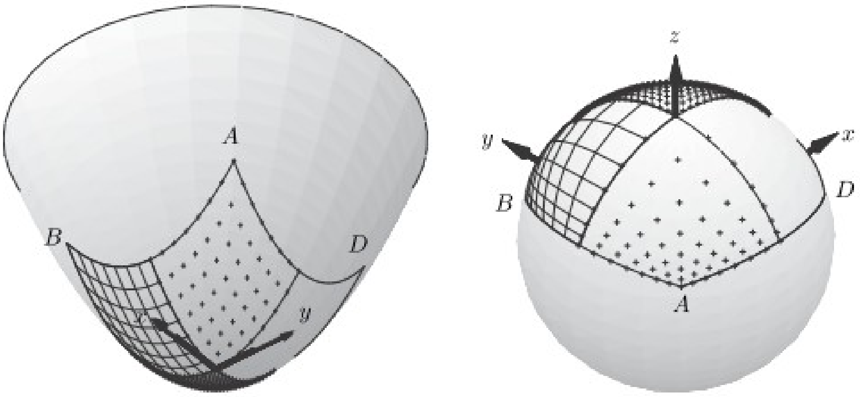

(a) Elliptic paraboloid (b) Unit sphere

Figure 6.4. Gauss map on the elliptic paraboloid.

The tangent plane to the sphere S at a point $\vec{p}$ consists of all possible vectors $\vec{x}'(t_0)$, where $\vec{x}(t)$ is a curve on S, with $\vec{x}(t_0) = \vec{p}$. Therefore, at any point $\vec{p}$ on the sphere, there are two options for unit normal vectors:

$$\frac{\vec{p}-\vec{c}}{\|\vec{p}-\vec{c}\|} \quad \text{or} \quad -\frac{\vec{p}-\vec{c}}{\|\vec{p}-\vec{c}\|}$$

However, it is the former that provides the outward-pointing orientation n. Thus the Gauss map for the sphere of radius R and center $\vec{c}$ is explicitly

$$n(\vec{p}) = \frac{\vec{p}-\vec{c}}{R}.$$

Furthermore, if S is itself the unit sphere centered at the origin, then the Gauss map is the identity function.

Example 6.2.3. Consider the elliptic paraboloid S defined by the equation $z = x^2 + y^2$. This is an orientable surface, and suppose it is oriented with the unit normal always pointing in a positive z-direction. (Figure 6.4 shows this is possible.) Using the function $F(x, y, z) = z - x^2 - y^2$, the elliptic paraboloid is given by the equation $F(x, y, z) = 0$. By Proposition 5.2.13, for all $(x, y, z) \in S$, the gradient $\vec{\nabla}F(x,y,z)$ is a normal vector, so (since the z-direction is positive) the Gauss map is

$$n(x,y,z) = \frac{(-2x,-2y,1)}{\sqrt{4x^2+4y^2+1}}.$$

Figure 6.4 shows how the Gauss map acts on the patch

$$\{(x,y,z) \in S \mid -1 \le x \le 1 \text{ and } -1 \le y \le 1\}.$$

It is not hard to show that the image of the Gauss map on the unit sphere is the upper hemisphere. Indeed, using cylindrical coordinates, the Gauss map is

$$n(r,\theta,z) = \frac{(-2r\cos\theta,-2r\sin\theta,1)}{\sqrt{4r^2+1}}.$$

The function $f(r) = 1/\sqrt{4r^2+1}$ is a decreasing bijection from $[0,+\infty)$ to $(0,1]$. Thus, for any $z_0 \in (0,1]$, there exists a unique $r_0 \in [0,+\infty)$ with $f(r_0) = z_0$. Then, with $\theta \in [0,2\pi]$, the image of $n(r_0,\theta,z_0)$ is a circle on the unit sphere at height z_0. Thus, the image of n is

$$n(S) = \{(x,y,z) \in \mathbb{S}^2 \mid z > 0\}.$$

Example 6.2.4. In contrast to Example 6.2.3, consider the hyperbolic paraboloid given by the parametrization $\vec{X} : \mathbb{R}^2 \to \mathbb{R}^3$, with

$$\vec{X}(u,v) = (u,v,u^2-v^2).$$

We calculate

$$\vec{X}_u = (1,0,2u) \qquad \text{and} \qquad \vec{X}_v = (0,1,-2v),$$

and thus,

$$\vec{N} = \frac{\vec{X}_u \times \vec{X}_v}{\|\vec{X}_u \times \vec{X}_v\|} = \frac{(-2u,2v,1)}{\sqrt{4u^2+4v^2+1}}.$$

Like Figure 6.4, Figure 6.5 shows the Gauss map for a "square" on the hyperbolic paraboloid. These two examples manifest a central behavior of the Gauss map. On the elliptic paraboloid, the Gauss

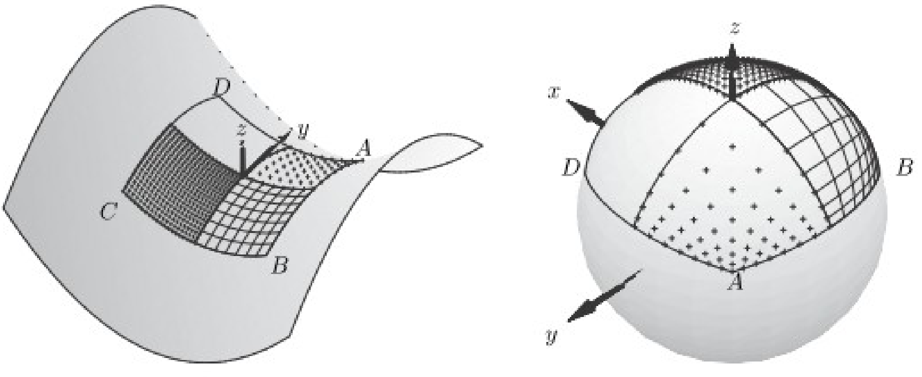

(a) Hyperbolic paraboloid (b) Unit sphere

Figure 6.5. Gauss map on the hyperbolic paraboloid.

map preserves the orientation of the square in the sense that if one travels along the boundary in a clockwise sense on the surface, one also travels along the boundary of the image of the square mapped by n in a clockwise sense. On the other hand, with the hyperbolic paraboloid, the Gauss map reverses the orientation of the square.

Proposition 6.2.5. *Let S be a regular surface in $\mathbb{R}^3$, and let $\vec{X} : U \to \mathbb{R}^3$ be a parametrization of a coordinate patch $V = \vec{X}(U)$, where U is an open set in $\mathbb{R}^2$. Define the vector function $\vec{N} : U \to \mathbb{R}^3$ by*

$$\vec{N} = \frac{\vec{X}_u \times \vec{X}_v}{\|\vec{X}_u \times \vec{X}_v\|}.$$

If $\vec{X}$ is of class C^r, then $\vec{N}$ is of class C^{r-1}.

Proof: By Problem 3.1.10, we deduce that if $\vec{F} : U \to \mathbb{R}^3$, $\vec{G} : U \to \mathbb{R}^3$, and $h : U \to \mathbb{R}$ are of class C^1, then the three functions $\vec{F} \cdot \vec{G} : U \to \mathbb{R}$, $\vec{F} \times \vec{G} : U \to \mathbb{R}^3$, and $h\vec{F} : U \to \mathbb{R}^3$ are also of class C^1, and their partial derivatives follow appropriate product rules. By definition,

$$\vec{N} = \frac{\vec{X}_u \times \vec{X}_v}{\sqrt{(\vec{X}_u \times \vec{X}_v) \cdot (\vec{X}_u \times \vec{X}_v)}}.$$

One can see that any partial derivative of any order of $\vec{N}$ will involve the partial derivative of a combination of the above three types of products and the derivative of an expression of the form

$$f(u,v) = \left((\vec{X}_u \times \vec{X}_v) \cdot (\vec{X}_u \times \vec{X}_v)\right)^{-k/2},$$

where k is an odd positive integer. Since S is regular, $\|\vec{X}_u \times \vec{X}_v\|$ is never 0, so a particular higher derivative of $\vec{N}$ will exist if that higher derivative exists for $\vec{X}_u$ and for $\vec{X}_v$. The proposition follows. $\square$

One would be correct to think that the different behaviors in Figures 6.4 and 6.5 can be illustrated and quantified by a function between tangent spaces to S and $\mathbb{S}^2$, respectively. This perspective is encapsulated in the notion of the differential of functions between regular surfaces. We develop the more general theory of differentials of functions between manifolds in Chapter 3 of [22]. However, the differential of the Gauss map, though only a particular case of the differentials of functions between surfaces, serves a central role in the rest of this chapter.

Let S be a regular oriented surface with orientation n, and let $\vec{X} : U \to \mathbb{R}^3$ be a regular positively oriented parametrization of a coordinate patch $\vec{X}(U)$ of S. Suppose also that S is of class C^2, which, according to Proposition 6.2.5, implies that $\vec{N}$ is of class C^1. Since $\vec{N} \cdot \vec{N} = 1$ for all $(u,v) \in U$, we have

$$\vec{N} \cdot \vec{N}_u = 0 \qquad \text{and} \qquad \vec{N} \cdot \vec{N}_v = 0 \tag{6.11}$$

for all $(u,v) \in U$. Therefore, both derivatives $\vec{N}_u$ and $\vec{N}_v$ are vector functions such that $\vec{N}_u(q)$ and $\vec{N}_v(q)$ are in T_pS when $\vec{X}(q) = p$.

The vector function $\vec{N} : U \to \mathbb{R}^3$ is itself a parametrized surface, with its image lying in the unit sphere, though not necessarily giving a regular parametrization of $\mathbb{S}^2$. The simple fact in Equation (6.11) indicates that if $\vec{N}$ admits a tangent plane at $q \in U$, then

$$T_pS = T_{\vec{N}(q)}(\mathbb{S}^2) \tag{6.12}$$

as subspaces of $\mathbb{R}^3$. In other words, the sets of vectors $\{\vec{X}_u, \vec{X}_v\}$ and $\{\vec{N}_u, \vec{N}_v\}$ span the same subspace T_pS.

The differential of the Gauss map at a point $p \in S$ is a linear transformation $dn_p : T_pS \to T_{n(p)}(\mathbb{S}^2)$, but by virtue of Equation (6.12), one identifies it as a linear transformation $dn_p : T_pS \to T_pS$. For all $\vec{v} \in T_pS$, one defines $dn_p(\vec{v}) = \vec{w}$ if there exists a curve $\vec{\alpha} : I \to U$ such that

$$\begin{cases} \vec{\alpha}(0) = q, \text{ with } \vec{X}(q) = p, \\ \dfrac{d}{dt}\left(\vec{X} \circ \vec{\alpha}\right)(t)\big|_{t=0} = \vec{v}, \text{ and} \\ \dfrac{d}{dt}\left(\vec{N} \circ \vec{\alpha}\right)(t)\big|_{t=0} = \vec{w}. \end{cases} \tag{6.13}$$

In other words, by writing $\vec{X}(t) = \vec{X} \circ \vec{\alpha}(t)$ and $\vec{N}(t) = \vec{N} \circ \vec{\alpha}(t)$, the differential of the Gauss map satisfies $dn_p(\vec{X}'(0)) = \vec{N}'(0)$.

Using coordinate lines $\vec{X}(t, v_0)$ or $\vec{X}(u_0, t)$ for the curve $\vec{\alpha}(t)$ in the above definition, it is easy to see that

$$\begin{aligned} dn_p(\vec{X}_u) &= \vec{N}_u, \\ dn_p(\vec{X}_v) &= \vec{N}_v. \end{aligned} \tag{6.14}$$

Though the Gauss map $n : S \to \mathbb{S}^2$ is independent of any coordinate systems on S, the differential dn_p requires reference to a regular positively oriented parametrization of a neighborhood of p. However, though different parametrizations of a neighborhood of p induce different coordinate bases on T_pS, the definition in Equation 6.13 remains unchanged and consequently, as a linear transformation of T_pS to itself, is independent of the parametrization.

Problems

6.2.1. Describe the region of the unit sphere covered by the image of the Gauss map for the following surfaces:

(a) The hyperbolic paraboloid given by $z = x^2 - y^2$.

(b) The hyperboloid of one sheet given by $x^2 + y^2 - z^2 = 1$.

(c) The cone with opening angle α given by $z^2 \tan^2 \alpha = x^2 + y^2$.

(d) The right circular cylinder $x^2 + y^2 = R^2$, with R a constant.

6.2.2. What regular surface S has a single point on $\mathbb{S}^2$ as the image of the Gauss map?

5- T<2 73

6.5 Gaussian and Mean Curvature

We now arrive at two fundamental geometric invariants that encapsulate a considerable amount of useful information about the local shape of a surface. Again, in this section, we must assume that S is a regular surface of class C^2.

Definition 6.5.1. Let κ_1 and κ_2 be the principal curvatures of a regular surface S at a point p. Define

1. the *Gaussian curvature* of S at p as the product $K = \kappa_1\kappa_2$;
2. the *mean curvature* of S at p as the average $H = \frac{\kappa_1+\kappa_2}{2}$ of the principal curvatures.

By Problem 6.3.6, the matrix for the Gauss map in terms of any particular coordinate system on a neighborhood of p on S is conjugate to the corresponding matrix for a different coordinate system via the change of basis matrix $\left(\frac{\partial x_i}{\partial \bar{x}_j}\right)$. Therefore, since $\det(BAB^{-1}) = \det(A)$ and $\mathrm{Tr}(BAB^{-1}) = \mathrm{Tr}(A)$, which is a standard result in linear algebra, the eigenvalues of the Gauss map, the principal curvatures, the Gaussian curvature, and the mean curvature are invariant under coordinate changes. Furthermore, since $\det(-A) = \det(A)$ when A is a square matrix with an even number of rows, then

$$\begin{aligned} K &= \kappa_1\kappa_2 = \det(a_i^j) = \det(dn_p), \\ H &= \frac{\kappa_1+\kappa_2}{2} = -\frac{1}{2}\,\mathrm{Tr}(a_i^j) = -\frac{1}{2}\,\mathrm{Tr}(dn_p). \end{aligned}$$

Consequently, as claimed above, the Gaussian curvature and the mean curvature of S at p are geometric invariants. This means that they do not depend on the orientation or position of S in space, and they do not depend on any particular coordinate system on S in a neighborhood of the point p.

In order to calculate the mean curvature $H(u, v)$, one has no choice but to calculate the matrix of the differential of the Gauss map $(a_j^i) = [dn_p]$. However, from a computational perspective, Equation (6.26) leads to a much simpler formula for the Gaussian curvature function $K(u, v)$. Recall that $\det(AB) = \det(A)\det(B)$ for all square matrices of the same size and also that $\det(A^{-1}) = 1/\det(A)$. Then Equation (6.26) implies that

$$K = \frac{\det(L_{ij})}{\det(g_{ij})} = \frac{L_{11}L_{22} - L_{12}^2}{g_{11}g_{22} - g_{12}^2}. \tag{6.37}$$

This equation lends itself readily to calculations since one does not need to fully compute the matrix of the Gauss map, let alone find its eigenvalues. However, we can also obtain an alternate characterization of the Gaussian curvature.

Proposition 6.5.2. *Let S be a regular surface of class C^2, and let V be a neighborhood parametrized by $\vec{X} : U \subset \mathbb{R}^2 \to \mathbb{R}^3$. Define the unit normal vector $\vec{N}$ as*

$$\vec{N} = \frac{\vec{X}_u \times \vec{X}_v}{\|\vec{X}_u \times \vec{X}_v\|}.$$

Then over the domain U, the Gaussian curvature is the unique function $K(u,v)$ satisfying

$$\vec{N}_u \times \vec{N}_v = K(u,v)\vec{X}_u \times \vec{X}_v$$

Proof: From the definition of the (a^i_j) functions in Equation (6.22), we have

$$\vec{N}_u \times \vec{N}_v = (a^1_1\vec{X}_u + a^2_1\vec{X}_v) \times (a^1_2\vec{X}_u + a^2_2\vec{X}_v).$$

It is then easy to see that

$$\vec{N}_u \times \vec{N}_v = \det(a^i_j)\vec{X}_u \times \vec{X}_v.$$

The result follows since $K = \det(a^i_j)$. □

Example 6.5.3 (Spheres). Consider the sphere parametrized by the vector function

$$\vec{X}(u,v) = (R\cos u \sin v, R\sin u \sin v, R\cos v),$$

with $(u,v) \in (0, 2\pi) \times (0, \pi)$. Example 6.1.5 and Example 6.3.3 gave us

$$(g_{ij}) = \begin{pmatrix} R^2\sin^2 v & 0 \\ 0 & R^2 \end{pmatrix} \qquad \text{and} \qquad (L_{ij}) = \begin{pmatrix} -R\sin^2 v & 0 \\ 0 & -R \end{pmatrix}.$$

Using Equation (6.37), it is then easy to see that the Gaussian curvature function on the sphere is

$$K(u,v) = \frac{R^2 \sin^2 v}{R^4 \sin^2 v} = \frac{1}{R^2},$$

which is a constant function. (Note that we assumed that $v \neq 0$, π, which correspond to the north and south poles of the parametrization. However, another parametrization that includes the north and south poles would show that at these points as well we have $K = 1/R^2$.)

The matrix of the Gauss map is then

$$[dn_p] = -Lg^{-1} = \begin{pmatrix} -\frac{1}{R} & 0 \\ 0 & -\frac{1}{R} \end{pmatrix},$$

from which one immediately deduces that the principal curvatures are $\kappa_1(u,v) = \kappa_2(u,v) = \frac{1}{R}$. This shows that all points on the sphere are umbilical points. In addition, the mean curvature is also a constant function $H(u,v) = \frac{1}{R}$.

Example 6.5.4 (Function Graphs). Consider the graph of a function $f : U \subset \mathbb{R}^2 \to \mathbb{R}$. The graph can be parametrized by $\vec{X} : U \to \mathbb{R}^3$, with

$$\vec{X}(u,v) = (u, v, f(u,v)).$$

Problem 6.3.7 asks the reader to calculate the matrix (L_{ij}). It is not hard to show that

$$(g_{ij}) = \begin{pmatrix} 1 + (f_u)^2 & f_u f_v \\ f_v f_u & 1 + (f_v)^2 \end{pmatrix} \qquad \text{and} \qquad (L_{ij}) = \frac{1}{\sqrt{1 + f_u^2 + f_v^2}} \begin{pmatrix} f_{uu} & f_{uv} \\ f_{vu} & f_{vv} \end{pmatrix},$$

where f_u is the typical shorthand to mean $\frac{\partial f}{\partial u}$. But then $\det(g) = 1 + f_u^2 + f_v^2$, and we find that the Gaussian curvature function on a function graph is

$$K(u,v) = \frac{1}{(1 + f_u^2 + f_v^2)^2} \begin{vmatrix} f_{uu} & f_{uv} \\ f_{vu} & f_{vv} \end{vmatrix} = \frac{f_{uu} f_{vv} - f_{uv}^2}{(1 + f_u^2 + f_v^2)^2}.$$

This result allows us to rephrase the second derivative test in the calculus of a function f from $\mathbb{R}^2$ to $\mathbb{R}$ as follows: if f has continuous second partial derivatives and (u_0, v_0) is a critical point (i.e., $f_u(u_0, v_0) = f_v(u_0, v_0) = 0$), then

1. (u_0, v_0) is a local maximum if $K(u_0, v_0) > 0$ and $f_{uu}(u_0, v_0) < 0$;
2. (u_0, v_0) is a local minimum if $K(u_0, v_0) > 0$ and $f_{uu}(u_0, v_0) > 0$;
3. (u_0, v_0) is a saddle point if $K(u_0, v_0) < 0$;
4. the test is inconclusive if $K(u_0, v_0) = 0$.

In the language we have introduced in this chapter, local minima and local maxima of the function $z = f(u, v)$ are elliptic points, saddle points are hyperbolic points, and points where the second derivative test is inconclusive are either parabolic or planar points.

Example 6.5.5 (Pseudosphere). A *tractrix* is a curve in the plane with parametric equations

$$\vec{x}(t) = (t - \tanh t, \operatorname{sech} t),$$

and it has the x-axis as an asymptote. The *pseudosphere* is defined as half of the surface of revolution of a tractrix about its asymptote. More precisely, we can parametrize the pseudosphere by

$$\vec{X}(u, v) = (\operatorname{sech} v \cos u, \operatorname{sech} v \sin u, v - \tanh v),$$

with $u \in [0, 2\pi)$ and $v \in [0, \infty)$. (See Figure 6.10 for a picture of the tractrix and the pseudosphere.) We leave it as an exercise for the reader to prove that

$$(g_{ij}) = \begin{pmatrix} \operatorname{sech}^2 v & 0 \\ 0 & \tanh^2 v \end{pmatrix} \qquad \text{and} \qquad (L_{ij}) = \begin{pmatrix} \operatorname{sech} v \tanh v & 0 \\ 0 & -\operatorname{sech} v \tanh v \end{pmatrix}.$$

We find that the Gaussian curvature for the pseudosphere is $K = -1$, which motivates the name "pseudosphere" since it is analogous to the unit sphere but with a constant Gaussian curvature of -1 instead of 1.

Proposition 6.5.6. *Let S be a regular oriented surface of class C^2 with Gauss map $n : S \to \mathbb{S}^2$, and let p be a point on S. Call $K(p)$ the Gaussian curvature of S at p, and suppose that $K(p) \neq 0$. Let B_ε be the ball of radius ε around p and define $V_\varepsilon = B_\varepsilon \cap S$. Calling $K(p)$ the Gaussian curvature of S at p, we have*

$$|K(p)| = \lim_{\varepsilon \to 0} \frac{\operatorname{Area}(n(V_\varepsilon))}{\operatorname{Area}(V_\varepsilon)}.$$

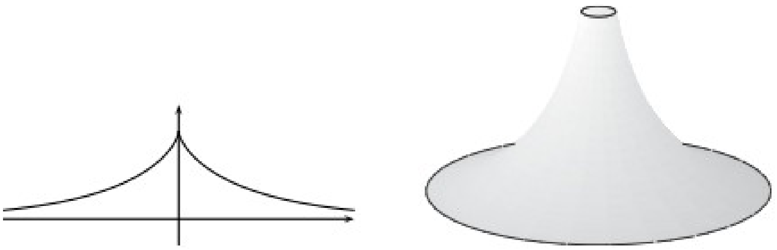

Figure 6.10. Tractrix and pseudosphere.

In other words, the absolute value of the Gaussian curvature at a point p is the limit around p of the ratio of the surface area on $\mathbb{S}^2$ mapped under the Gauss map to the corresponding surface area on S.

Proof: Let $\vec{X} : U \to \mathbb{R}^3$ be a regular parametrization of a neighborhood of $p = \vec{X}(u_0, v_0)$, where U is an open subset of $\mathbb{R}^2$. Since $\vec{X}(U)$ is an open set, $V_\varepsilon \subset \vec{X}(U)$ for all ε small enough. Furthermore, since we assumed that $K(p) \neq 0$, by the continuity of the Gaussian curvature function $K : U \to \mathbb{R}$, we deduce that K does not change sign for ε chosen small enough. Therefore, from now on, we assume that K does not change sign.

Define $U_\varepsilon = \vec{X}^{-1}(V_\varepsilon)$, the preimage of the neighborhood V_ε under $\vec{X}$. Since $\vec{X}$ is bijective as a regular parametrization, $\{(u_0, v_0)\}$ is the unique point in all U_ε for all $\varepsilon > 0$. Then by the formula for surface area,

$$\text{Area}(V_\varepsilon) = \iint_{V_\varepsilon} dA = \iint_{U_\varepsilon} \|\vec{X}_u \times \vec{X}_v\| \, du \, dv.$$

Similarly, on the unit sphere,

$$\text{Area}(n(V_\varepsilon)) = \iint_{U_\varepsilon} \|\vec{N}_u \times \vec{N}_v\| \, du \, dv.$$

However, by Proposition 6.5.2, we also have

$$\text{Area}(n(V_\varepsilon)) = \iint_{U_\varepsilon} |K(u,v)| \|\vec{X}_u \times \vec{X}_v\| \, du \, dv.$$

By the Mean Value Theorem for double integrals, for every $\varepsilon > 0$, there exist points $(u_\varepsilon, v_\varepsilon)$ and $(u'_\varepsilon, v'_\varepsilon)$ in the open set U_ε such that

$$\iint_{U_\varepsilon} \|\vec{X}_u \times \vec{X}_v\|\, du\, dv = \|\vec{X}_u(u_\varepsilon, v_\varepsilon) \times \vec{X}_v(u_\varepsilon, v_\varepsilon)\|$$

and

$$\iint_{U_\varepsilon} \|\vec{N}_u \times \vec{N}_v\|\, du\, dv = |K(u'_\varepsilon, v'_\varepsilon)|\, \|\vec{X}_u(u'_\varepsilon, v'_\varepsilon) \times \vec{X}_v(u'_\varepsilon, v'_\varepsilon)\|.$$

Thus, since $\lim_{\varepsilon\to 0}(u_\varepsilon, v_\varepsilon) = \lim_{\varepsilon\to 0}(u'_\varepsilon, v'_\varepsilon) = (u_0, v_0)$, we have

$$\begin{aligned}\lim_{\varepsilon\to 0}\frac{\text{Area}(n(V_\varepsilon))}{\text{Area}(V_\varepsilon)} &= \lim_{\varepsilon\to 0}\frac{|K(u'_\varepsilon, v'_\varepsilon)|\, \|\vec{X}_u(u'_\varepsilon, v'_\varepsilon) \times \vec{X}_v(u'_\varepsilon, v'_\varepsilon)\|}{\|\vec{X}_u(u_\varepsilon, v_\varepsilon) \times \vec{X}_v(u_\varepsilon, v_\varepsilon)\|} \\ &= \frac{|K(u_0, v_0)|\|\vec{X}_u(u_0, v_0) \times \vec{X}_v(u_0, v_0)\|}{\|\vec{X}_u(u_0, v_0) \times \vec{X}_v(u_0, v_0)\|} \\ &= |K(u_0, v_0)| = |K(p)|. \qquad \square\end{aligned}$$

Problems

6.5.1. Calculate the Gaussian curvature of the torus parametrized by

$$\vec{X}(u, v) = ((a + b\cos v)\cos u, (a + b\cos v)\sin u, b\sin v)$$

where $a > b$ are constants and $(u, v) \in (0, 2\pi) \times (0, 2\pi)$.

6.5.2. Consider a regular space curve $\vec{\gamma} : I \to \mathbb{R}^3$ and the tangential surface defined by

$$\vec{X}(t, u) = \vec{\gamma}(t) + u\vec{\gamma}'(t)$$

for $(t, u) \in I \times \mathbb{R}$. Calculate the mean and Gaussian curvature functions.

6.5.3. Let $\vec{\alpha}(t)$ and $\vec{\beta}(t)$ be differentiable vector functions with common domain I. Define the *secant surface* between the two resulting curves by

$$\vec{X}(t, u) = (1 - u)\vec{\alpha}(t) + u\vec{\beta}(t)$$

for $(t, u) \in I \times \mathbb{R}$. Assume that the corresponding surface is regular.

5- <T2 N3

volve three unknown functions in two variables, they are intractable in general. Nonetheless, we shall study various classes of surfaces that satisfy $K = 0$ or $H = 0$.

Planes satisfy $K = 0$, but other surfaces do as well, for example, cylinders and cones. In a geometric sense, cylinders and cones in fact resemble a plane because, as any elementary school student knows, one can create a cylinder or a cone out of a flat paper without folding, stretching, or crumpling. In the first half of this section, we introduce ruled surfaces and determine the conditions for these ruled surfaces to have Gaussian curvature everywhere 0.

As of yet, we have neither seen a particularly intuitive interpretation of the mean curvature nor have presented surfaces that satisfy $H = 0$. However, as we shall see, surfaces that satisfy $H = 0$ have minimal surface area in the following sense. Given a curve C in $\mathbb{R}^3$, among surfaces S that have $C = \partial S$ as a boundary, a surface with minimal surface area will have mean curvature H everywhere 0. Consequently, a surface that satisfies $H = 0$ is called a *minimal surface*. Many articles and books are devoted to the study of minimal surfaces (see [20], [26], or [25] to name a few; an internet search will reveal many more), so in the interest of space, the second half of this section gives only a brief introduction to minimal surfaces.

6.6.1 Ruled Surfaces

Geometrically, we define a *ruled surface* as the union of a (differentiable) one-parameter family of straight lines in $\mathbb{R}^3$. One can specify each line in the family by a point $\vec{\alpha}(t)$ and by a direction given by another vector $\vec{w}(t)$. The adjective "differentiable" means that both $\vec{\alpha}$ and $\vec{w}$ are differentiable vector functions over some interval $I \subset \mathbb{R}$. We parametrize a ruled surface by

$$\vec{X}(t,u) = \vec{\alpha}(t) + u\,\vec{w}(t), \qquad \text{with } (t,u) \in I \times \mathbb{R}. \tag{6.38}$$

We call the lines L_t passing through $\vec{\alpha}(t)$ with direction $\vec{w}(t)$ the *rulings*, and the curve $\vec{\alpha}(t)$ is called the *directrix* of the surface. We should note that this definition does not insist that ruled surfaces be regular; in particular, we allow singular points, that is, points where $\vec{X}_t \times \vec{X}_u = \vec{0}$.

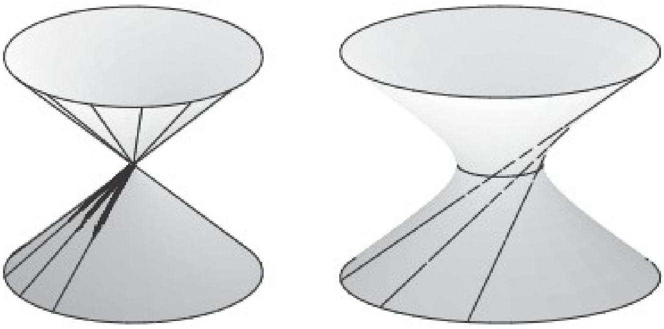

(a) Cone (b) Hyperboloid

Figure 6.11. Ruled surfaces.

Example 6.6.1 (Cylinders). The simplest example of a ruled surface is a cylinder. The general definition of a cylinder is a ruled surface that can be given as a one-parameter family of lines where the directrix $\vec{\alpha}(t)$ is planar and $\vec{w}(t)$ is a constant vector in $\mathbb{R}^3$.

Example 6.6.2 (Cones). The general definition of a cone is a surface that can be given as a one-parameter family of lines where the directrix $\vec{\alpha}(t)$ lies in a plane P and the rulings L_t all pass through some common point $p \notin P$. Therefore, the cone over $\vec{\alpha}(t)$ through p can be parametrized by

$$\vec{X}(t,u) = \vec{\alpha}(t) + u(p - \vec{\alpha}(t)).$$

Example 6.6.3. As a perhaps initially surprising example, the hyperboloid of one sheet is also a ruled surface. The standard parametrization for the hyperboloid of one sheet is

$$\vec{X}(u,v) = (\cosh v \cos u, \cosh v \sin u, \sinh v).$$

Consider alternatively the ruled surface given by the directrix $\vec{\alpha}(t) = (\cos t, \sin t, 0)$, with ruling directions $\vec{w}(t) = \vec{\alpha}'(t) + \vec{k} = (-\sin t, \cos t, 1)$ (see Figure 6.11(b)). We obtain the parametrized surface

$$\vec{Y}(t,u) = (\cos t - u\sin t, \sin t + u\cos t, u).$$

Though it is not particularly easy to express $\vec{Y}$ as a reparametrization of $\vec{X}$, it is not hard to see that both of these parametrizations satisfy the following usual equation that gives the hyperboloid as a conic surface:

$$x^2 + y^2 - z^2 = 1.$$

We will determine the Gaussian curvature for ruled surfaces, but, for the purposes of subsequent calculations, we make two simplifications. First, note that one can impose the condition that $\|\vec{w}(t)\| = 1$ without changing the definition or the image of the parametrization in Equation (6.38). In defining particular ruled surfaces, it is usually easier not to make this assumption, but it does simplify calculations since $\|\vec{w}(t)\| = 1$ for all $t \in I$ implies that $\vec{w}(t) \cdot \vec{w}(t)' = 0$. Secondly, note that different curves $\vec{\alpha}(t)$ can serve as the directrix for the same ruled surface, so we wish to employ a curve $\vec{\beta}(t)$ as a directrix, which will simplify the algebra in our calculations. We choose a curve $\vec{\beta}(t)$ satisfying $\vec{\beta}'(t) \cdot \vec{w}'(t) = 0$.

Since $\vec{\beta}(t)$ lies on $\vec{X}$, we can write

$$\vec{\beta}(t) = \vec{\alpha}(t) + u(t)\vec{w}(t). \tag{6.39}$$

Then

$$\vec{\beta}'(t) = \vec{\alpha}'(t) + u'(t)\vec{w}(t) + u(t)\vec{w}'(t),$$

and since $\vec{\beta}' \cdot \vec{w}' = 0$, we have

$$\vec{\alpha}'(t) \cdot \vec{w}'(t) + u(t)\vec{w}'(t) \cdot \vec{w}'(t) = 0.$$

Thus, we determine that

$$u(t) = -\frac{\vec{\alpha}'(t) \cdot \vec{w}'(t)}{\|\vec{w}'(t)\|^2}. \tag{6.40}$$

Furthermore, it is easy to prove (see Problem 6.6.3) that, with our present assumptions, $\vec{\beta}(t)$ is unique. We call the curve $\vec{\beta}(t)$ the *line of stricture* of the ruled surface.

We now use the parametrization for the ruled surface $\vec{X} = \vec{\beta}(t) + u\vec{w}(t)$ and proceed to calculate the first and second fundamental forms. Understanding that $\vec{\beta}$ and $\vec{w}$ are functions of t, we find that

$$\begin{aligned} \vec{X}_t &= \vec{\beta}' + u\,\vec{w}', & & & \vec{X}_{tt} &= \vec{\beta}'' + u\,\vec{w}'', \\ \vec{X}_u &= \vec{w}, & &\text{and} & \vec{X}_{tu} &= \vec{w}', \\ \vec{X}_t \times \vec{X}_u &= \vec{\beta}' \times \vec{w} + u\,\vec{w}' \times \vec{w}, & & & \vec{X}_{uu} &= \vec{0}. \end{aligned}$$

One can already notice that if $\vec{w}(t)$ is constant, then $L_{12} = L_{21} = L_{22} = 0$, which leads to $K = 0$ and proves that all cylinders have Gaussian curvature $K = 0$. We will assume now that $\vec{w}'(t)$ is not identically 0.

Because of the conditions $\vec{w} \cdot \vec{w}' = 0$ and $\vec{\beta}' \cdot \vec{w}' = 0$, we can conclude that $\vec{\beta}' \times \vec{w}$ is parallel to $\vec{w}'$. Thus, $\vec{\beta}' \times \vec{w}$ is perpendicular to $\vec{w} \times \vec{w}'$, so

$$\|\vec{X}_t \times \vec{X}_u\|^2 = \|\vec{\beta}' \times \vec{w}\|^2 + u^2\|\vec{w}'\|^2.$$

Furthermore, $\vec{\beta}' \times \vec{w} = (\vec{\beta}' \times \vec{w}) \cdot \vec{w}'/\|\vec{w}'\|$ and hence,

$$\begin{aligned} \|\vec{X}_t \times \vec{X}_u\|^2 &= (\vec{\beta}'\vec{w}\vec{w}')^2/\|\vec{w}'\|^2 + u^2\|\vec{w}'\|^2 && (6.41)\\ &= \frac{1}{\|\vec{w}'\|^2}\left((\vec{\beta}'\vec{w}\vec{w}')^2 + u^2\|\vec{w}'\|^4\right), && (6.42)\end{aligned}$$

where we use the notation $(\vec{\beta}'\vec{w}\vec{w}')$ for $(\vec{\beta}' \times \vec{w}) \cdot \vec{w}'$, the triple-vector product in $\mathbb{R}^3$. Consequently, we can write the unit normal vector $\vec{N}$ as

$$\vec{N}(t,u) = \frac{\|\vec{w}'\|}{\sqrt{(\vec{\beta}'\vec{w}\vec{w}')^2 + u^2\|\vec{w}'\|^4}}\left(\vec{\beta}' \times \vec{w} + u\,\vec{w}' \times \vec{w}\right), \qquad (6.43)$$

and therefore, again using the conditions that $\vec{w} \cdot \vec{w}' = 0$ and $\vec{\beta}' \cdot \vec{w}' = 0$, we get

$$(g_{ij}) = \begin{pmatrix} \|\vec{\beta}'\|^2 + u^2\|\vec{w}'\|^2 & 0 \\ 0 & 1 \end{pmatrix} \qquad (6.44)$$

and

$$(L_{ij}) = \frac{\|\vec{w}'\|}{\sqrt{(\vec{\beta}'\vec{w}\vec{w}')^2 + u^2\|\vec{w}'\|^4}}\begin{pmatrix} (\vec{\beta}'' + u\vec{w}'') \cdot (\vec{\beta}' \times \vec{w} + u\vec{w}' \times \vec{w}) & (\vec{\beta}'\vec{w}\vec{w}') \\ (\vec{\beta}'\vec{w}\vec{w}') & 0 \end{pmatrix}. \qquad (6.45)$$

We cannot say much for the entry L_{11}, but thanks to Equation (6.37) for the Gaussian curvature and the fact that $\|\vec{X}_u \times \vec{X}_v\|^2 = \det(g_{ij})$ (see Equation (6.6)), one can calculate the Gaussian curvature for a ruled surface as

$$K = \frac{\det(L_{ij})}{\det(g_{ij})} = -\frac{\|\vec{w}'\|^4(\vec{\beta}'\vec{w}\vec{w}')^2}{\left((\vec{\beta}'\vec{w}\vec{w}')^2 + u^2\|\vec{w}'\|^4\right)^2}. \qquad (6.46)$$

This formula for the Gaussian curvature of a ruled surface makes a few facts readily apparent. First, $K \leq 0$ for all points of a ruled surface. Secondly, by Equation (6.41), all of the singular points of a ruled surface, i.e., points where $\vec{X}_t \times \vec{X}_u = \vec{0}$, must have $u = 0$ and therefore occur on the line of stricture. Finally, we see that a noncylindrical ruled surface satisfies $K(t, u) = 0$ if and only if $(\vec{\beta}'\vec{w}\vec{w}') = 0$ for all $t \in I$.

Let us return now to the general definition of a ruled surface from Equation (6.38) with regular curves $\vec{\alpha}(t)$ and $\vec{w}(t)$ with no conditions. Since both $\vec{w}$ and $\vec{w}'$ are perpendicular to $\vec{w} \times \vec{w}'$,

$$(\vec{\alpha}'\vec{w}\vec{w}') = \vec{\alpha}' \cdot (\vec{w} \times \vec{w}') = \vec{\beta}' \cdot (\vec{w} \times \vec{w}') = (\vec{\beta}'\vec{w}\vec{w}').$$

Furthermore, if we call $\hat{w} = \vec{w}/\|\vec{w}\|$, we remark that

$$(\vec{\alpha}'\vec{w}\vec{w}') = \|w\|^2(\vec{\alpha}'\hat{w}\hat{w}') = \|w\|^2(\vec{\beta}'\hat{w}\hat{w}').$$

Consequently, $(\vec{\beta}'\hat{w}\hat{w}') = 0$ if and only if $(\vec{\alpha}'\vec{w}\vec{w}') = 0$.

The ruled surface $\vec{X}(t, u)$ is called a *developable surface* if $(\vec{\alpha}'\vec{w}\vec{w}') = 0$. Equation (6.46) shows that all developable surfaces have Gaussian curvature identically 0. This result provides us with a class of surfaces that satisfy $K = 0$, though it by no means proves whether $K = 0$ implies that a surface is developable.

It is interesting to remark that a ruled surface is a cone if and only if $\vec{\alpha}'(t) = 0$ and is a cylinder if and only if $\vec{w}'(t) = 0$, showing again that both cones and cylinders are developable and have $K = 0$. The exercises present examples of developable surfaces that are neither cones nor cylinders.

6.6.2 Minimal Surfaces

Definition 6.6.4. A *minimal surface* is a parametrized surface of class C^2 that satisfies the regularity condition and for which the mean curvature is identically 0.

We first wish to justify the name "minimal." Let $\vec{X} : U \to \mathbb{R}^3$ be a coordinate neighborhood of a regular parametrized surface of class C^2. Let D' be a connected compact set in U, and let $D = \vec{X}(D')$. Let $h : U \to \mathbb{R}$ be a differentiable function. A normal variation of

$\vec{X}$ over D' determined by h is the family of surfaces with $t \in (-\varepsilon, \varepsilon)$ defined by

$$\begin{aligned} \vec{X}^t : D' &\longrightarrow \mathbb{R}^3 \\ (u,v) &\longmapsto \vec{X}(u,v) + th(u,v)\vec{N}(u,v). \end{aligned} \tag{6.47}$$

Proposition 6.6.5. *Let $\vec{X}^t$ be a normal variation of $\vec{X}$ over a compact region D' and determined by some function h. For ε small enough, $\vec{X}^t$ satisfies the regularity condition for all $t \in (-\varepsilon, \varepsilon)$. In this case, the area of $\vec{X}^t$ is*

$$A(t) = \iint_{D'} \sqrt{1 - 4thH + t^2 R}\sqrt{\det(g_{ij})}\, du\, dv \tag{6.48}$$

for some function $R(u,v,t)$ that is polynomial in t.

Proof: Denote g_{ij}^t as the coefficients of the metric tensor. The surfaces of the normal variation have

$$\begin{aligned} \vec{X}_u^t &= \vec{X}_u + th_u\vec{N} + th\vec{N}_u, \\ \vec{X}_v^t &= \vec{X}_v + th_v\vec{N} + th\vec{N}_v. \end{aligned}$$

Thus, we calculate,

$$\begin{aligned} g_{11}^t &= g_{11} + 2th\vec{X}_u \cdot \vec{N}_u + t^2h^2\vec{N}_u \cdot \vec{N}_u + t^2h_u^2, \\ g_{12}^t &= g_{12} + th(\vec{X}_u \cdot \vec{N}_v + \vec{X}_v \cdot \vec{N}_u) + t^2h^2\vec{N}_u \cdot \vec{N}_v + t^2h_uh_v, \\ g_{22}^t &= g_{22} + 2th\vec{X}_v \cdot \vec{N}_v + t^2h^2\vec{N}_v \cdot \vec{N}_v + t^2h_v^2. \end{aligned}$$

However, by Equation (6.16), one can summarize the above equations as

$$g_{ij}^t = g_{ij} - 2thL_{ij} + t^2h^2\vec{N}_i \cdot \vec{N}_j + t^2h_ih_j,$$

where we use the notation h_1 (respectively h_2) to indicate the partial derivative h_u (respectively h_v). Therefore, we calculate that

$$\det(g_{ij}^t) = \det(g_{ij}) - 2th(g_{11}L_{22} - 2g_{12}L_{12} + g_{22}L_{11}) + t^2\bar{R},$$

where $\bar{R}(u,v,t)$ is a function of the form $A_0(u,v) + tA_1(u,v) + t^2A_2(u,v)$ for continuous functions A_i defined over D'. However, the mean curvature is

$$H = \frac{1}{2}\frac{g_{11}L_{22} - 2g_{12}L_{12} + g_{22}L_{11}}{g_{11}g_{22} - g_{12}^2},$$

which leads to

$$\det(g_{ij}^t) = \det(g_{ij})(1 - 4thH) + t^2\bar{R} = \det(g_{ij})(1 - 4thH + t^2R), \tag{6.49}$$

where $R = \bar{R}/\det(g_{ij})$. Since D' is compact, the functions h, H, and R are bounded over D', which shows that

$$\lim_{t\to 0} \det(g_{ij}^t) = \det(g_{ij}) \quad \text{for all } (u, v) \in D'.$$

Hence, if $\varepsilon > 0$ is small enough, then $\det(g_{ij}^t) \neq 0$ for all $t \in (-\varepsilon, \varepsilon)$, and thus, all normal variations satisfy the regularity condition. The rest of the proposition follows from Equation (6.49) since

$$A(t) = \iint\limits_{D'} \sqrt{\det(g_{ij}^t)}\, du\, dv.$$

□

Proposition 6.6.6. *Let $\vec{X} : U \to \mathbb{R}^3$ be a parametrized surface of class C^2, and let $D' \in U$ be a compact set. Let $A(t)$ be the area function defined in Equation (6.48). Then $\vec{X}$ parametrizes a minimal surface if and only if $A'(0) = 0$ for all D' and all normal variations of $\vec{X}$ over D'.*

Proof: We calculate

$$A'(t) = \iint\limits_{D'} \frac{-4hH + 2tR + t^2R_t}{2\sqrt{1 - 4thH + t^2R}} \sqrt{\det(g_{ij})}\, du\, dv,$$

which implies that

$$A'(0) = -2 \iint\limits_{D'} hH \sqrt{\det(g_{ij})}\, du\, dv.$$

Obviously, if $\vec{X}$ parametrizes a minimal surface, which means that $H = 0$ for all $(u, v) \in U$, then $A'(0) = 0$ regardless of the function $h(u, v)$ or the compact set D'.

To prove the converse, suppose that $A'(0) = 0$ for all continuous functions $h(u, v)$ and all compact $D' \subset U$. Choosing $h(u, v) = H(u, v)$, since $\det(g_{ij}) > 0$, we have

$$A'(0) = -2 \iint\limits_{D'} H^2 \sqrt{\det(g_{ij})}\, du\, dv,$$

so $A'(0) \leq 0$ for all choices of D'. Since $H(u,v)$ is continuous, if $H(u_0,v_0) \neq 0$ for any point (u_0,v_0), there is a compact set D' containing (u_0,v_0) such that $H(u,v)^2 > 0$ over D'. Thus, if H is anywhere nonzero, there exists a compact subset D' such that $A'(0) < 0$. This is a contradiction, so we conclude that $A'(0) = 0$ for all h and all D' implies that $H(u,v) = 0$ for all $(u,v) \in U$. □

Proposition 6.6.6 shows that a parametrized surface $\vec{X}$ that satisfies the regularity condition $\|\vec{X}_u \times \vec{X}_u\| \neq 0$ is minimal (has $H = 0$ everywhere) if it is a surface such that over every patch of surface there is no way to deform it along normal vectors to obtain a surface of lesser surface area.

Minimal surfaces have enjoyed considerable popular attention, especially in museums of science, as soap films on a wire frame. When a wire frame is dipped into a soapy liquid and pulled out, the surface tension on the soap film pulls the film into the state of least potential energy, which turns out to be the surface such that no normal variation can lead to a surface with less surface area. Of course, when experimenting with soap films, one can experiment with wire frames that are nonregular curves or not even curves at all (the skeleton of a cube, for example) and investigate what kinds of soap film surfaces result.

The problem of determining a minimal surface with a given closed regular space curve C as a boundary was raised by Lagrange in 1760. However, the mathematical problem became known as Plateau's problem, named after Joseph Plateau who specifically studied soap film surfaces. The nineteenth century saw a few specialized solutions to the problem, but it was not solved until 1930. (See [25] for a more complete history of Plateau's problem.)

Though one can easily construct a minimal surface with a soap film on a wire frame, either checking that a surface is minimal or finding a parametrization for a minimal surface is quite difficult. In fact, the study of minimal surfaces continues to provide new areas of research and connections with other branches of analysis. Perhaps among the most interesting results is a connection between complex analytic functions and minimal surfaces [10, p. 206] or the use of elliptic integrals to parametrize special minimal surfaces [25, Chapter 1], topics that go beyond the scope of this book. However, it is quite likely that having both the simple experimentative aspect and connections to advanced mathematics has fueled interest in minimal surfaces.

Problems

6.6.1. Prove that the surface given by $z = kxy$, where k is a constant, is a ruled surface and give it as a one-parameter family of lines.

6.6.2. The tangential surface to a space curve $\vec{\alpha}(t)$ was presented in Problem 6.5.2. Show that the curve $\vec{\alpha}$ is the line of stricture for the tangential surface. Show that the tangential surface to a regular space curve is a developable surface.

6.6.3. Suppose that $\{\vec{\alpha}_1(t), \vec{w}(t)\}$ and $\{\vec{\alpha}_2(t), \vec{w}(t)\}$, where $\|\vec{w}(t)\| = 1$, are two one-parameter families of lines that trace out the same ruled surface. Prove that Equation (6.39) with, (6.40), using either $\vec{\alpha}_1$ or $\vec{\alpha}_2$, produce the same line of stricture.

6.6.4. Let S be an orientable surface, and let $\vec{\alpha}(s)$ be a curve on S parametrized by arc length. Assume that $\vec{\alpha}$ is nowhere tangent to the asymptotic direction on S, and define $\vec{N}(s)$ as the unit normal vector to S along $\vec{\alpha}(s)$. Consider the ruled surface

$$\vec{X}(s, u) = \vec{\alpha}(s) + u\,\frac{\vec{N}(s) \times \vec{N}'(s)}{\|\vec{N}'(s)\|}.$$

The assumption that $\vec{\alpha}'(s)$ is not an asymptotic direction ensures that $\vec{N}'(s) \neq \vec{0}$. Prove that $\vec{X}(s, v)$ is a developable surface. (This kind of surface is called the *envelope* of a family of tangent planes along a curve of a surface.)

6.6.5. Show that a surface $F(x, y, z) = 0$ is developable if

$$\begin{vmatrix} F_{xx} & F_{xy} & F_{xz} & F_x \\ F_{yx} & F_{yy} & F_{yz} & F_y \\ F_{zx} & F_{zy} & F_{zz} & F_z \\ F_x & F_y & F_z & 0 \end{vmatrix} = 0.$$

6.6.6. Let $\vec{X}(t, v) = \vec{\alpha}(t) + v\vec{w}(t)$ be a developable surface. Prove that at a regular point

$$\vec{N}_v \cdot \vec{X}_t = 0 \qquad \text{and} \qquad \vec{N}_v \cdot \vec{X}_v = 0.$$

Use this result to prove that the tangent plane of a developable surface is constant along a line of ruling.

5- 6/2 N 3

7.2.6. *Tubes*. Let $\vec{\alpha}(t)$ be a regular space curve and let r be small enough that the tube

$$\vec{X}(t,u) = \vec{\alpha}(t) + (r\cos u)\vec{P}(t) + (r\sin u)\vec{B}(t)$$

is a regular surface. Calculate the Christoffel symbols for the tube $\vec{X}$.

7.2.7. Prove that a parametrization $\vec{X}$ of a neighborhood on a surface S is such that all the coordinate lines $x^2 = c$ are asymptotic lines if and only if $L_{11} = 0$.

7.2.8. Defining the function $D(x^1, x^2)$ as $D^2 = \det(g_{ij})$, show that

$$\frac{\partial}{\partial x^1}(\ln D) = \Gamma^1_{11} + \Gamma^2_{12} \qquad \text{and} \qquad \frac{\partial}{\partial x^2}(\ln D) = \Gamma^1_{12} + \Gamma^2_{22}.$$

7.2.9. Suppose that S is a surface with a parametrization that satisfies $\Gamma^1_{12} = \Gamma^2_{12} = 0$. Prove that S is a translation surface.

7.2.10. Using D as defined in Problem 7.2.8, calling $\theta(x^1, x^2)$ the angle between the coordinate lines, prove that

$$\frac{\partial\theta}{\partial x^1} = -\frac{D}{g_{11}}\Gamma^2_{11} - \frac{D}{g_{22}}\Gamma^2_{11} \quad \text{and} \quad \frac{\partial\theta}{\partial x^2} = -\frac{D}{g_{11}}\Gamma^2_{12} - \frac{D}{g_{22}}\Gamma^1_{22}. \quad (7.40)$$

Show that if the parametrization is orthogonal, i.e., $g_{12} = 0$, Equation (7.40) becomes

$$g_{22}\Gamma^2_{11} + g_{11}\Gamma^1_{12} = 0 \qquad \text{and} \qquad g_{22}\Gamma^2_{12} + g_{11}\Gamma^1_{22} = 0.$$

7.3 Codazzi Equations and the Theorema Egregium

The Gauss equations in Equation (7.29) define expressions for any second derivative $\vec{X}_{ij}$ in terms of the coefficients of the first and second fundamental forms and the basis $\{\vec{X}_1, \vec{X}_2, \vec{N}\}$. We remind the reader that the symmetry of the dot product imposes $g_{ij} = g_{ji}$. Furthermore, from $\vec{X}_{ij} = \vec{X}_{ji}$, we also deduced that $L_{ij} = L_{ji}$.

Definition 5.2.9 of a regular surface imposes the condition that all the higher derivatives of any parametrization $\vec{X}$ of a coordinate patch be continuous for all higher derivatives. Therefore, for a regular parametrization, the order in which one takes derivatives with respect to given variables is irrevelant. In particular, for the mixed third derivatives, we have $\vec{X}_{112} = \vec{X}_{121}$ and $\vec{X}_{221} = \vec{X}_{122}$, which can be listed in the following slightly more suggestive manner:

$$\frac{\partial \vec{X}_{11}}{\partial x^2} = \frac{\partial \vec{X}_{12}}{\partial x^1} \quad \text{and} \quad \frac{\partial \vec{X}_{22}}{\partial x^1} = \frac{\partial \vec{X}_{12}}{\partial x^2}$$

Consequently, because of the nontrivial expressions in Gauss's equations, we obtain

$$\begin{aligned} \frac{\partial}{\partial x^2}\left(\Gamma_{11}^1 \vec{X}_1 + \Gamma_{11}^2 \vec{X}_2 + L_{11}\vec{N}\right) &= \frac{\partial}{\partial x^1}\left(\Gamma_{12}^1 \vec{X}_1 + \Gamma_{12}^2 \vec{X}_2 + L_{12}\vec{N}\right), \\ \frac{\partial}{\partial x^2}\left(\Gamma_{12}^1 \vec{X}_1 + \Gamma_{12}^2 \vec{X}_2 + L_{12}\vec{N}\right) &= \frac{\partial}{\partial x^1}\left(\Gamma_{22}^1 \vec{X}_1 + \Gamma_{22}^2 \vec{X}_2 + L_{22}\vec{N}\right). \end{aligned} \tag{7.41}$$

It is natural to equate the basis components of each equation. However, when we take the partial derivatives, we will obtain expressions involving second derivatives of $\vec{X}$ and derivatives of $\vec{N}$ that we need to put back into the usual $\{\vec{X}_1, \vec{X}_2, \vec{N}\}$ basis using the Weingarten equations given in Equation (6.26) or Gauss's equations in Equation (7.29). Doing this transforms Equation (7.41) into six distinct equations from which we deduce two significant theorems in the theory of surfaces.

Theorem 7.3.1 (Codazzi Equations). *For any parametrized smooth surface, the following hold:*

$$\begin{aligned} \frac{\partial L_{11}}{\partial x^2} - \frac{\partial L_{12}}{\partial x^1} &= L_{11}\Gamma_{12}^1 + L_{12}\Gamma_{12}^2 - L_{12}\Gamma_{11}^1 - L_{22}\Gamma_{11}^2, \\ \frac{\partial L_{12}}{\partial x^2} - \frac{\partial L_{22}}{\partial x^1} &= L_{11}\Gamma_{22}^1 + L_{12}\Gamma_{22}^2 - L_{12}\Gamma_{12}^1 - L_{22}\Gamma_{12}^2. \end{aligned} \tag{7.42}$$

We can summarize these two equations in one by

$$\frac{\partial L_{ij}}{\partial x^l} - \Gamma_{il}^k L_{kj} = \frac{\partial L_{il}}{\partial x^j} - \Gamma_{ij}^k L_{kl} \tag{7.43}$$

for all $1 \le i, j, l \le 2$, where Einstein summation is implied.

Proof: Equate the $\vec{N}$ coefficient functions in both equations from Equation (7.41). □

The classical formulation of the above relationships in Equation (7.42) are collectively called the Codazzi equations or sometimes the Mainardi-Codazzi equations. These equations present a relationship that must hold between the coefficients of the first and second fundamental form that must hold for all regular surfaces. They are a crucial part of the Fundamental Theorem of Surface Theory that we will present in Section 7.4.

The following theorem is another result that stems from equating components along $\vec{X}_1$ and $\vec{X}_2$ in Equation (7.41).

Theorem 7.3.2 (Theorema Egregium). *The Gaussian curvature of a surface is an intrinsic property of the surface, that is, it depends only on the coefficients of the metric tensor and higher derivatives thereof.*

Proof: Equating the $\vec{X}_1$ and $\vec{X}_2$ coefficient functions in the first equation of Equation (7.41), we obtain

$$\frac{\partial \Gamma_{11}^1}{\partial x^2} + \Gamma_{11}^1\Gamma_{12}^1 + \Gamma_{11}^2\Gamma_{22}^1 + L_{11}a_2^1 = \frac{\partial \Gamma_{12}^1}{\partial x^1} + \Gamma_{12}^1\Gamma_{11}^1 + \Gamma_{12}^2\Gamma_{12}^1 + L_{12}a_1^1,$$
$$\frac{\partial \Gamma_{11}^2}{\partial x^2} + \Gamma_{11}^1\Gamma_{12}^2 + \Gamma_{11}^2\Gamma_{22}^2 + L_{11}a_2^2 = \frac{\partial \Gamma_{12}^2}{\partial x^1} + \Gamma_{12}^1\Gamma_{11}^2 + \Gamma_{12}^2\Gamma_{12}^2 + L_{12}a_1^2.$$

Since $a_j^i = -\sum_{l=1}^2 L_{jl}g^{li}$, where $(g^{ij}) = (g_{ij})^{-1}$, after some simplification, we can write these two equations as

$$g_{21}\frac{L_{11}L_{22} - (L_{12})^2}{\det g} = \frac{\partial \Gamma_{12}^1}{\partial x^1} - \frac{\partial \Gamma_{11}^1}{\partial x^2} + \Gamma_{12}^2\Gamma_{12}^1 - \Gamma_{11}^2\Gamma_{22}^1, \quad (7.44)$$

$$\begin{aligned} -g_{11}\frac{L_{11}L_{22} - (L_{12})^2}{\det g} &= \frac{\partial \Gamma_{12}^2}{\partial x^1} - \frac{\partial \Gamma_{11}^2}{\partial x^2} + \Gamma_{12}^1\Gamma_{11}^2 + \Gamma_{12}^2\Gamma_{12}^2 \\ &\quad - \Gamma_{11}^1\Gamma_{12}^2 - \Gamma_{11}^2\Gamma_{22}^2. \end{aligned} \quad (7.45)$$

Since $K = \det(L_{ij})/\det(g_{ij})$, we can use either one of these equations to obtain a formula for K in terms of the function g_{ij} and Γ_{jk}^i. However, since the Christoffel symbols Γ_{jk}^i are themselves determined by the coefficients of the metric tensor, then these equations provide formulas for the Gaussian curvature exclusively in terms of the first fundamental form. □

Gauss coined the name "Theorema Egregium," which means "an excellent theorem," and indeed this result should seem rather surprising. A priori, the Gaussian curvature depends on the commponents of the first fundamental form and the second fundamental form. After all, the Gaussian curvature at a point p on the surface is $K = \det(dn_p)$, where dn_p is the differential of the Gauss map. Since the coefficients of the matrix dn_p with respect to the basis $\{\vec{X}_1, \vec{X}_2\}$ involve the L_{ij} coefficients, one would naturally expect that we would need the second fundamental form to calculate K. However, the Theorema Egregium shows that only the coefficient functions of the first fundamental form are necessary.

In the study of surfaces, we called a local or a global property of a curve or surface a "geometric quantity" if it does not change under the orientation and position of the curve or surface in space. On the other hand, topology studies properties of point sets that are invariant under any homeomorphism—a bijective function that is continuous in both directions. The concept of an intrinsic property lies somewhere between these two extremes. Intrinsic geometry studies what some mathematicians classically call "bending invariants," quantities that are preserved when one bends the curve or surface but does not stretch them.

In the above proof, Equations (7.44) and (7.45) motivate the following definition of the *Riemann symbols*:

$$R^l_{ijk} = \frac{\partial \Gamma^l_{ik}}{\partial x^j} - \frac{\partial \Gamma^l_{jk}}{\partial x^i} + \Gamma^m_{ik}\Gamma^l_{mj} - \Gamma^m_{jk}\Gamma^l_{mi}. \tag{7.46}$$

As it turns out (see Problem 7.3.2), the Riemann symbols form the components of a $(1,3)$-tensor. Then, a closely associated tensor denoted by

$$R_{ijkl} = R^m_{ijk} g_{ml} \tag{7.47}$$

has the interesting property that $R_{1212} = \det(L_{ij})$, and hence, the Gaussian curvature of a surface at a point is given by the component function

$$K = \frac{R_{1212}}{\det(g_{ij})}. \tag{7.48}$$

The tensor associated to the components R^l_{ijk} is called the Riemann curvature tensor and plays an important role in the analysis on manifolds.

Example 7.3.3 (Cone). As a simple example, consider the right circular cone defined by the parametrization

$$\vec{X}(u,v) = (v\cos u, v\sin u, v),$$

where $u \in [0, 2\pi]$ and $v > 0$, which we know, as a developable surface, has Gaussian curvature $K = 0$ everywhere. Simple calculations give, for the metric tensor,

$$g_{11} = v^2, \qquad g_{12} = 0, \qquad g_{22} = 2.$$

For the Christoffel symbols, one obtains

$$\Gamma_{11}^2 = -\frac{1}{2}v, \qquad \Gamma_{12}^1 = \Gamma_{21}^1 = \frac{1}{v}$$

with all the other symbols of the second kind $\Gamma_{jk}^i = 0$. With this data, an application of Equation (7.46) shows that all but two of the Riemann symbols vanish simply because all the terms vanish. However, for the two nontrivial terms, one calculates

$$R_{122}^1 = \frac{\partial \Gamma_{12}^1}{\partial v} - \frac{\partial \Gamma_{22}^1}{\partial u} + \Gamma_{12}^1\Gamma_{12}^1 + \Gamma_{12}^2\Gamma_{22}^1 - \Gamma_{22}^1\Gamma_{11}^1 - \Gamma_{22}^2\Gamma_{21}^1 = -\frac{1}{v^2} + \frac{1}{v^2} = 0,$$
$$R_{121}^2 = \frac{\partial \Gamma_{11}^2}{\partial v} - \frac{\partial \Gamma_{12}^2}{\partial u} + \Gamma_{11}^1\Gamma_{12}^2 + \Gamma_{11}^2\Gamma_{22}^2 - \Gamma_{21}^1\Gamma_{11}^2 - \Gamma_{21}^2\Gamma_{21}^2 = -\frac{1}{2} - \left(\frac{1}{v}\right)\left(-\frac{1}{2}v\right) = 0.$$

Consequently, though the Christoffel symbols are not identically 0, the Riemann curvature tensor is identically 0. One finds then that $R_{1212} = 0$ as a function of u and v, and hence, using Equation (7.48), one recovers the fact that the cone has Gaussian curvature identically 0 everywhere.

The Theorema Egregium (Theorem 7.3.2) excited the mathematics community in the middle of the nineteenth century and sparked a search for a variety of alternative formulations for the Gaussian curvature of a surface as a function of the g_{ij} metric coefficients. We leave a few such formulas for K for the exercises, but we present a few here.

Recall that $L_{ij} = \vec{X}_{ij} \cdot \vec{N}$, $\det(g_{ij}) = \|\vec{X}_u \times \vec{X}_v\|^2$, and $\vec{N} = \vec{X}_u \times \vec{X}_v / \|\vec{X}_u \times \vec{X}_v\|$. From these facts, one can easily see that

$$K = \frac{L_{11}L_{22} - L_{12}^2}{\det(g_{ij})} = \frac{(\vec{X}_{uu}\vec{X}_u\vec{X}_v)(\vec{X}_{vv}\vec{X}_u\vec{X}_v) - (\vec{X}_{uv}\vec{X}_u\vec{X}_v)^2}{\det(g_{ij})^2}, \quad (7.49)$$

where $(\vec{A}\vec{B}\vec{C})$ is the triple-vector product in $\mathbb{R}^3$. However, the triple-vector product is a determinant and we know that for all matrices M_1

and M_2, one has $\det(M_1M_2) = \det(M_1)\det(M_2)$, and so Equation (7.49) becomes

$$K = \frac{1}{\det(g_{ij})^2}\left(\begin{vmatrix} \vec{X}_{uu}\cdot\vec{X}_{vv} & \vec{X}_{uu}\cdot\vec{X}_u & \vec{X}_{uu}\cdot\vec{X}_v \\ \vec{X}_u\cdot\vec{X}_{vv} & \vec{X}_u\cdot\vec{X}_u & \vec{X}_u\cdot\vec{X}_v \\ \vec{X}_v\cdot\vec{X}_{vv} & \vec{X}_v\cdot\vec{X}_u & \vec{X}_v\cdot\vec{X}_v \end{vmatrix} - \begin{vmatrix} \vec{X}_{uv}\cdot\vec{X}_{uv} & \vec{X}_{uv}\cdot\vec{X}_u & \vec{X}_{uv}\cdot\vec{X}_v \\ \vec{X}_u\cdot\vec{X}_{uv} & \vec{X}_u\cdot\vec{X}_u & \vec{X}_u\cdot\vec{X}_v \\ \vec{X}_v\cdot\vec{X}_{uv} & \vec{X}_v\cdot\vec{X}_u & \vec{X}_v\cdot\vec{X}_v \end{vmatrix}\right). \quad (7.50)$$

This equation involves only the metric tensor coefficients and Γ^i_{jk} symbols, except in the inner products $\vec{X}_{uu}\cdot\vec{X}_{vv}$ and $\vec{X}_{uv}\cdot\vec{X}_{uv}$. At first glance, this problem seems insurmountable, but if one performs the Laplace expansion on the determinants along the first row, factors terms involving $\det(g_{ij})$, and collects into determinants again, one can show that Equation (7.50) is equivalent to

$$K = \frac{1}{\det(g_{ij})^2}\left(\begin{vmatrix} \vec{X}_{uu}\cdot\vec{X}_{vv} - \vec{X}_{uv}\cdot\vec{X}_{uv} & \vec{X}_{uu}\cdot\vec{X}_u & \vec{X}_{uu}\cdot\vec{X}_v \\ \vec{X}_u\cdot\vec{X}_{vv} & \vec{X}_u\cdot\vec{X}_u & \vec{X}_u\cdot\vec{X}_v \\ \vec{X}_v\cdot\vec{X}_{vv} & \vec{X}_v\cdot\vec{X}_u & \vec{X}_v\cdot\vec{X}_v \end{vmatrix} - \begin{vmatrix} 0 & \vec{X}_{uv}\cdot\vec{X}_u & \vec{X}_{uv}\cdot\vec{X}_v \\ \vec{X}_u\cdot\vec{X}_{uv} & \vec{X}_u\cdot\vec{X}_u & \vec{X}_u\cdot\vec{X}_v \\ \vec{X}_v\cdot\vec{X}_{uv} & \vec{X}_v\cdot\vec{X}_u & \vec{X}_v\cdot\vec{X}_v \end{vmatrix}\right). \quad (7.51)$$

The value in this is that though one cannot express $\vec{X}_{uu}\cdot\vec{X}_{vv}$ or $\vec{X}_{uv}\cdot\vec{X}_{uv}$ using only the metric coefficients, it can be shown that

$$\vec{X}_{uu}\cdot\vec{X}_{vv} - \vec{X}_{uv}\cdot\vec{X}_{uv} = -\frac{1}{2}g_{11,22} + g_{12,12} - \frac{1}{2}g_{22,11},$$

where by $g_{ij,kl}$ we mean the function

$$\frac{\partial^2 g_{ij}}{\partial x^k \partial x^l}.$$

This produces the following formula for K that is appealing on the grounds that it illustrates symmetry of the metric tensor coefficients in determining K:

$$K = \frac{1}{\det(g_{ij})^2}\left(\begin{vmatrix} -\frac{1}{2}g_{11,22} + g_{12,12} - \frac{1}{2}g_{11,22} & \frac{1}{2}g_{11,1} & g_{12,1} - \frac{1}{2}g_{11,2} \\ g_{12,2} - \frac{1}{2}g_{22,1} & g_{11} & g_{12} \\ \frac{1}{2}g_{22,2} & g_{21} & g_{22} \end{vmatrix} - \begin{vmatrix} 0 & \frac{1}{2}g_{11,2} & \frac{1}{2}g_{22,1} \\ \frac{1}{2}g_{11,2} & g_{11} & g_{12} \\ \frac{1}{2}g_{22,1} & g_{21} & g_{22} \end{vmatrix} \right). \tag{7.52}$$

In the situation where one considers an orthogonal parametrization of a neighborhood of a surface, namely, when $g_{12} = 0$, Equation (7.52) takes on the particularly interesting form of

$$K = -\frac{1}{\sqrt{g_{11}g_{22}}}\left(\frac{\partial}{\partial x^1}\left(\frac{1}{\sqrt{g_{11}}}\frac{\partial\sqrt{g_{22}}}{\partial x^1} \right) + \frac{\partial}{\partial x^2}\left(\frac{1}{\sqrt{g_{22}}}\frac{\partial\sqrt{g_{11}}}{\partial x^2} \right) \right). \tag{7.53}$$

By historical habit, Equations (7.52) and (7.53) are also referred to as *Gauss's equations* and one usually refers to the collection of Equations (7.53) and (7.42) as the *Gauss-Codazzi equations*.

We now prove a theorem that we could have introduced much earlier, but it arises in part as an application of the Codazzi equations.

Definition 7.3.4. A set $S \subset \mathbb{R}^n$ is called *path-connected* if for any pair of points $p, q \in S$, there exists a continuous path (curve) $\alpha : [0,1] \to \mathbb{R}^n$, with $\alpha(0) = p$, $\alpha(1) = q$, and $\alpha([0,1]) \subset S$.

Proposition 7.3.5. *If S is a path-connected regular surface in which all its points are umbilical, then S is either contained in a plane or in a sphere.*

Proof: Let $U \subset \mathbb{R}^2$ be an open set, and let $\vec{X} : U \to \mathbb{R}^3$ be the parametrization of a neighborhood $V = \vec{X}(U)$ of S. Since all points of S are umbilical, then the eigenvalues of dn_p are equal to a value λ that a priori is a function of the coordinates (u, v) of the patch V. We first show that $\lambda(u, v)$ is a constant over U.

By Proposition 6.4.4, there always exist two linearly independent eigenvectors to dn_p, so dn_p is always diagonalizable. Thus, at any point p of S, the whole tangent plane T_pS is the eigenspace of the eigenvalue λ. Thus, in particular we have

$$\begin{aligned} \vec{N}_u &= dn_p(\vec{X}_u) = \lambda\vec{X}_u, \\ \vec{N}_v &= dn_p(\vec{X}_v) = \lambda\vec{X}_v. \end{aligned}$$

Since $\vec{N}_{uv} = \vec{N}_{vu}$, a fact that is equivalent to the Codazzi equations (see Problem 7.3.1), we have

$$\lambda_v \vec{X}_u + \lambda \vec{X}_{uv} = \lambda_u \vec{X}_v + \lambda \vec{X}_{vu}.$$

Since $\vec{X}_u$ and $\vec{X}_v$ are linearly independent, we deduce that $\lambda_u = \lambda_v = 0$, and therefore λ is a constant function.

If $\lambda = 0$, then $\vec{N}_u = \vec{N}_v = \vec{0}$, so the unit normal vector $\vec{N}$ is a constant vector $\vec{N}_0$ over the domain U. Then

$$\frac{\partial}{\partial u}(\vec{X} \cdot \vec{N}_0) = \vec{X}_u \cdot \vec{N}_0 + \vec{0} = \vec{0}$$

and similarly for $\frac{\partial}{\partial v}(\vec{X} \cdot \vec{N}_0) = \vec{0}$, which shows that $\vec{X} \cdot \vec{N}_0 = \vec{0}$, which further proves that $V = \vec{X}(U)$ lies in the plane through p perpendicular to $\vec{N}_0$.

On the other hand, if $\lambda \neq 0$, then consider the vector function

$$\vec{Y}(u,v) = \vec{X}(u,v) - \frac{1}{\lambda}\vec{N}(u,v).$$

By a similar calculation as above, we see that the function $\vec{Y}(u,v)$ is a constant vector. Then we deduce that

$$\|\vec{X}(u,v) - \vec{Y}\| = \left\|\frac{1}{\lambda}\vec{N}\right\| = \frac{1}{|\lambda|},$$

which shows that $V = \vec{X}(U)$ lies on a sphere of center $\vec{Y}$ and of radius $1/|\lambda|$.

So far, the above proof has established that any parametrization of a neighborhood V of S either lies on a sphere or lies in a plane but not that all of S necessarily lies on a sphere or in a plane. The assumption that S is path-connected extends the result to the whole surface as follows.

By Example 7.1.10, if V_1 and V_1 are two coordinate patches of S with respective coordinate systems (x^1, x^2) and $(\bar{x}^1, \bar{x}^2)$ such that $V_1 \cap V_2 \neq \emptyset$, then on $V_1 \cap V_2$ the coefficients of the Gauss map with respect to the coordinate systems are related by

$$\bar{a}^i_j = \frac{\partial \bar{x}^i}{\partial x^k} \frac{\partial x^l}{\partial \bar{x}^j} a^k_l.$$

This is tantamount to saying that the matrices for the Gauss map relative to different coordinate systems are similar matrices. Hence, if V_1 and V_2 are two overlapping coordinate patches, then the eigenvalue λ is constant over $V_1 \cup V_2$, and then all of $V_1 \cup V_2$ either lies in a plane or on a sphere.

Since S is path-connected, given any two points p and q on S, there exists a path $\alpha : [0,1] \to S$, with $\alpha(0) = p$ and $\alpha(1) = q$. From basic facts in topology, we know that $\alpha([0,1])$ is a compact set and that any open cover of a compact set has a finite subcover. Thus, given a collection of coordinate patches that cover $\alpha([0,1])$, then only a finite subcollection of these coordinate patches is necessary to cover $\alpha([0,1])$. Call this collection $\{V_1, V_2, \ldots, V_r\}$. By the reasoning in the above paragraph, we see that $V_1 \cup V_2 \cup \cdots \cup V_r$ lies either in a plane or on a sphere. Thus all points of S either lie in the same plane or lie on the same sphere. ☐

Problems

7.3.1. Show that one can obtain the Codazzi equations from the equation $\vec{N}_{12} = \vec{N}_{21}$ and by using the Gauss equations in Equation (7.29).

7.3.2. Prove that the Riemann symbols defined in Equation (7.46) form the components of a $(1,3)$-tensor.

7.3.3. Prove Equation (7.48) and the claim that supports it.

7.3.4. Suppose that $g_{12} = L_{12} = 0$ (i.e., the coordinate lines are lines of curvature), so that the principal curvatures satisfy

$$\kappa_1 = \frac{L_{11}}{g_{11}} \qquad \text{and} \qquad \kappa_2 = \frac{L_{22}}{g_{22}}.$$

Prove that the Codazzi equations are equivalent to

$$\frac{\partial \kappa_1}{\partial x^2} = \frac{g_{11,2}}{2g_{11}}(\kappa_2 - \kappa_1) \qquad \text{and} \qquad \frac{\partial \kappa_2}{\partial x^1} = \frac{g_{22,1}}{2g_{22}}(\kappa_1 - \kappa_2),$$

where by $g_{11,2}$ we mean $\frac{\partial g_{11}}{\partial x^2}$ and similarly for $g_{22,1}$.

7.3.5. Prove that the Riemann symbols R^l_{ijk} defined in Equation (7.46) are antisymmetric in the indices $\{i,j\}$. Conclude that for surfaces, R^l_{ijk} represent at most four distinct functions.

7.3.6. Prove Equation (7.53).

5- 62 3

Proposition 8.4.11. *A compact, regular, orientable surface of class C^3 with positive curvature everywhere is homeomorphic to a sphere.*

Proof: Since $K > 0$ over the whole surface S, then for the total Gaussian curvature we have

$$\iint_S K\, ds > 0.$$

By Corollary 8.4.9, this integral is $2\pi\chi(S)$. However, by the Classification Theorem of Surfaces and Equation (8.40), since $\chi(S) > 0$, we have $\chi(S) = 2$, and therefore S is homeomorphic to a sphere. □

Proposition 8.4.12. *Let S be a compact, connected, orientable, regular surface of positive Gaussian curvature. If there exist two simple, closed geodesics γ_1 and γ_2 on S then they intersect.*

Proof: By Proposition 8.4.11, S is homeomorphic to a sphere. Suppose that γ_1 and γ_2 do not intersect. Then they form the boundary of region $\mathcal{R}$ that is homeomorphic to a cylinder with boundary. It is not hard to verify by supplying $\mathcal{R}$ with a triangulation that $\chi(\mathcal{R}) = 0$. However, applying the Gauss-Bonnet Theorem to this situation, we obtain

$$\iint_{\mathcal{R}} K\, dS = 0,$$

which is a contradiction since $K > 0$. □

We now discuss applications of the Gauss-Bonnet Theorem to plane, spherical, and hyperbolic geometry and show how the theorem simultaneously generalizes many well-known theorems in these areas.

8.4.3 Plane, Spherical, and Hyperbolic Geometry

As mentioned in the introduction to this chapter, in his *Elements*, Euclid proves all his propositions through from 23 definitions and five postulates. The fifth postulate reads as follows:

> If a straight line crossing two straight lines makes the interior angles on the same side less than two right angles, the two straight lines, if extended indefinitely, meet on that side on which are the angles less than the two right angles.

Much wordier than the other four, many mathematicians over the centuries attempted to prove the fifth postulate from the others but never succeeded. Studying Euclid's *Elements*, one notices that Euclid himself uses the fifth postulate sparingly. One could speculate that Euclid did this so that, were someone to prove the fifth postulate as a consequence of the others, then the proofs of theorems would still be given in a minimal form. Indeed, many commonly known properties about plane geometry hold without reference to the fifth postulate. However, examples of two commonly known theorems that appear to rely on it are the following:

1. Given a line L and a point P not on L, through P passes exactly one line parallel to L.
2. The sum of the interior angles of a triangle is π.

In the nineteenth century, mathematicians Lobachevsky and Bolyai constructed examples of what we now call non-Euclidean geometries, in which one replaces Euclid's fifth postulate with one of two options. Given a line L and a point P not on L, through P passes (S) no parallel lines to L or (H) more than one parallel line to L. These geometries are called, respectively, *spherical geometry* and *hyperbolic geometry*. Using elementary techniques, one can study these geometries and establish many theorems, including theorems in spherical trigonometry. (See [9] and [27] for comprehensive treatment of non-Euclidean geometry using elementary techniques.) One well-known result in these geometries is that the sum of the interior angles of a triangle is greater than π in the case of spherical geometry and less than π in the case of hyperbolic geometry. Figure 8.15 gives an example of a triangle on a sphere in which each vertex has a right angle, and hence, the sum of the interior angles is $3\pi/2$.

We now apply the Gauss-Bonnet Theorem to plane, spherical, and hyperbolic geometries, beginning with plane geometry.

In plane geometry, we consider the plane to be a surface with Gaussian curvature identically 0 and curves in the plane we consider to be curves on a surface. Consider now a region $\mathcal{R}$ in the plane such that the boundary $\partial\mathcal{R}$ is a piecewise regular, simple, closed curve.

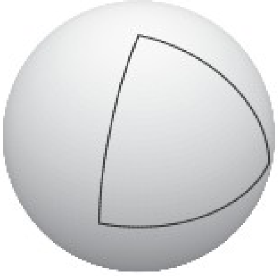

Figure 8.15. Triangle on a sphere.

It is easy to calculate that the Euler characteristic of a region that is homeomorphic to a disk is $\chi(\mathcal{R}) = 1$.

As a first case, let $\partial\mathcal{R}$ be a polygon. Since the regular arcs are straight lines, the geodesic curvature of each regular arc of $\partial\mathcal{R}$ is identically 0. Gauss-Bonnet's formula then reduces to the well-known fact that the sum of the exterior angles $\{\theta_1, \theta_2, \dots, \theta_k\}$ around a polygon is 2π. Also, since the exterior angle θ_i at any corner is $\pi - \alpha_i$, where α_i is the interior angle, we deduce that the sum of the interior angles of any polygon is

$$\sum_{i=1}^{n} \alpha_i = (n-2)\pi.$$

This statement is in fact equivalent to a theorem that occurs in every high school geometry curriculum, namely, that the sum of the interior angles of a triangle is π radians.

As a second case of the Gauss-Bonnet Theorem applied to plane geometry, suppose that $\partial\mathcal{R}$ is a regular curve. In this case, $\partial\mathcal{R}$ has no vertices and hence no exterior angles. Then the Gauss-Bonnet Theorem reduces to the formula

$$\oint_{\partial\mathcal{R}} \kappa_g \, ds = 2\pi,$$

and hence, the Gauss-Bonnet Theorem generalizes the proposition that the rotation index of a simple, closed curve is 1 (see Propositions 2.2.1 and 2.2.8).

In order to apply the results of this chapter to non-Euclidean geometry, one must clarify the use of the terms "line" and "parallel." In the geometry on a surface of class C^3, the word "line" (or straight line) means a geodesic. The word "parallel" is a little more problematic. In Problem 1.3.12, we defined a parallel curve to a given curve $\vec{\gamma}$ as a curve whose locus is always a fixed orthogonal distance r from the locus of $\vec{\gamma}$. This is true of parallel lines in the plane. However, in the sense of classical geometry, two lines are called parallel if they do not intersect.

Example 8.2.7 showed that geodesics on a sphere are great arcs of the sphere, i.e., an arc on the circle of intersection between the sphere and a plane through the center of the sphere. Therefore, any two geodesic lines intersect at at least two points: the common intersection of the sphere and the two planes supporting the geodesics. This remark justifies the definition of spherical geometry as a geometry in which there exist no parallel lines.

Consider now a triangle $\mathcal{R}$ on a sphere of radius R. By definition of a geometric triangle, $\partial\mathcal{R}$ has three vertices, and its regular arcs are geodesic curves. The Gaussian curvature of a sphere of radius R is $\frac{1}{R^2}$. In this situation, the Gauss-Bonnet Theorem gives

$$(\theta_1 + \theta_2 + \theta_3) + \frac{A}{R^2} = 2\pi, \tag{8.45}$$

where A is the surface area of $\mathcal{R}$, and θ_i are the exterior angles of the vertices. If we call $\alpha_i = \pi - \theta_i$ the interior angles of the triangle, then Equation (8.45) reduces to

$$\alpha_1 + \alpha_2 + \alpha_3 = \pi + \frac{A}{R^2}. \tag{8.46}$$

This shows that in a spherical triangle, the excess sum of the angles, i.e., $\alpha_1 + \alpha_2 + \alpha_3 - \pi$, is 4π times the ratio of the surface area of the triangle to the surface area of the whole sphere. As an example, the spherical triangle in Figure 8.15 has an excess of $\pi/2$ and covers $1/8$ of the sphere.

We propose to study (geodesic) circles on a sphere of radius R. Recall that a circle on a sphere is the set of points that are at a fixed distance r from a point, where distance here means geodesic

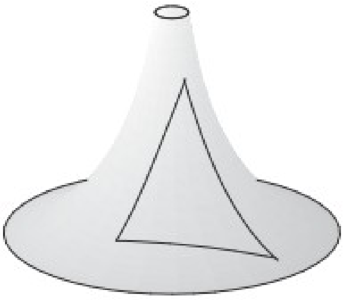

Figure 8.16. Hyperbolic triangle.

distance. Call $\mathcal{D}$ the closed disk inside this circle. Using the parametric equations for the sphere,

$$\vec{X}(u,v) = (R\cos u \sin v, R\sin u \sin v, R\cos v),$$

a circle of radius r around the north pole $(0,0,R)$ is given by $\vec{\gamma}(t) = \vec{X}(t, v_0)$, where $r = Rv_0$. It is not hard to tell by symmetry that the geodesic curvature κ_g of a circle on a sphere is a constant. The Gauss-Bonnet Theorem then gives

$$\kappa_g(\text{perimeter of } \partial\mathcal{D}) + \frac{1}{R^2}(\text{surface area of } \mathcal{D}) = 2\pi. \tag{8.47}$$

Using the surface area formula for a surface of revolution from single-variable calculus, it is possible to prove that the area of $\mathcal{D}$ is

$$\text{surface area of } \mathcal{D} = 2\pi R^2(1-\cos v_0) = 2\pi R^2\left(1-\cos\left(\frac{r}{R}\right)\right).$$

It is not hard to show that the perimeter of $\mathcal{D}$ is

$$\text{perimeter of } \mathcal{D} = 2\pi R \sin v_0 = 2\pi R \sin\left(\frac{r}{R}\right).$$

Consequently, the Gauss-Bonnet Theorem gives the curvature of a circle on a sphere as

$$\kappa_g = \frac{1}{R}\cot\left(\frac{r}{R}\right).$$

In contrast to spherical geometry, hyperbolic geometry is the geometry of a surface with constant negative Gaussian curvature. Example 6.5.5 presented the pseudosphere as a surface with $K = -1$

everywhere. In this particular case, using the Gauss-Bonnet Theorem, it is easy to calculate that the sum of the interior angles of a triangle is

$$\alpha_1 + \alpha_2 + \alpha_3 = \pi - A,$$

where A is the area of the triangle. (See Figure 8.16 for an example. Note that the regular arcs are indeed geodesic segments.) Interestingly enough, this formula for the sum of the interior angles of a hyperbolic triangle puts an upper bound of π for the possible area of a triangle on the pseudosphere with $K = -1$.

Problems

8.4.1. Provide all the details that establish Equation (8.36).

8.4.2. Provide the details for the proof of Lemma 8.4.6.

8.4.3. Calculate the surface area of a disk on a sphere using the surface area formula for a surface of revolution. Calculate directly the geodesic curvature function of a circle on a sphere. [Hint: See the paragraph around Equation (8.47).]

8.4.4. Verify directly the Gauss-Bonnet Theorem for the rectangular region $\mathcal{R}$ on a torus specified by $u_1 \leq u \leq u_2$ and $v_1 \leq v \leq v_2$, where the u-coordinate lines are the parallels and the v-coordinate lines are the meridians of the torus, as a surface of revolution.

8.4.5. Let S be a regular, orientable, compact surface with positive Gaussian curvature. Prove that the surface area of S is less than $4\pi/K_{\min}$, where $K_{\min} > 0$ is the minimum Gaussian curvature.

8.4.6. Let S be a regular, orientable surface of class C^3 in $\mathbb{R}^3$ that is homeomorphic to the sphere. Let γ be a simple closed geodesic in S. The curve γ separates S into two regions A and B that share γ as their boundary. Let $n : S \to \mathbb{S}^2$ be the Gauss map induced from a given orientation of S. Prove that $n(A)$ and $n(B)$ have the same area.

8.4.7. Let S be an orientable surface with Gaussian curvature $K \geq 0$, and let $p \in S$.

(a) Let γ_1 and γ_2 be two geodesics that intersect at p. Prove that γ_1 and γ_2 do not intersect at another point q in such a way that γ_1 and γ_2 form the boundary of a simple region $\mathcal{R}$.

(b) Prove also that a geodesic on S cannot intersect itself in such a way as to enclose a simple region.

8.4.8. *Jacobi's Theorem.* Let $\vec{\alpha} : I \to \mathbb{R}^3$ be a closed, regular, parametrized curve. Suppose also that $\vec{\alpha}(t)$ has a curvature function $\kappa(t)$ that is never 0. Suppose also that the principal normal indicatrix, i.e., the curve $\vec{P} : I \to \mathbb{S}^2$, is simple, that is, that it cuts the sphere into only two regions. By viewing $\vec{P}(I)$ as a curve on the sphere $\mathbb{S}^2$, use the Gauss-Bonnet Theorem to prove that $\vec{P}(I)$ separates the sphere into two regions of equal area.

8.5 Intrinsic Geometry

We remind the reader that an intrinsic property of a regular surface or a regular curve on a surface is a property that depends only on the coefficients of the metric tensor and the coordinates of a point or coordinate functions of a curve. Intrinsic geometry is the subset of differential geometry where one restricts one's attention to intrinsic properties.

In Section 6.1, we saw that the arc length of a curve on a surface, the area of a region on the surface, and the angle between two intersecting curves are all intrinsic properties of surfaces. As a consequence, a surface integral and an integral along a path (parametrized by arc length) of functions given in terms of the coordinates are intrinsic quantities. In contrast, the normal vector to a surface at a point and even the second fundamental form are not intrinsic. On the other hand, in Section 7.2 we saw that the Christoffel symbols Γ^i_{jk} are intrinsic quantities, and the Theorema Egregium (Theorem 7.3.2) shows that the Gaussian curvature is also an intrinsic property of the surface.

As far as curves on surfaces, Equation (8.7) shows the geodesic curvature to be an intrinsic property. Furthermore, since the differential equations for geodesics involve only the Christoffel symbols and the coordinate functions of the curve, geodesics are objects in intrinsic geometry. However, the functional dependence of the normal curvature and the geodesic torsion on the unit normal vector, as given in Equation (8.11), indicates that these quantities are not intrinsic. Since the Gauss-Bonnet formula on one side only involves κ_g and K, one may say it is a theorem in intrinsic geometry.

The value of narrowing one's focus to intrinsic geometry is that, in this context, one does not need a specific parametrization (coordinate patch) of a surface to calculate the quantities or determine curves of interest. In Chapters 6, 7, and 8, we have almost exclusively considered regular surfaces in $\mathbb{R}^3$. However, if $\vec{X} : U \to \mathbb{R}^n$, where U is an open subset of $\mathbb{R}^2$, though it is impossible to define *the* unit normal vector (even up to sign), it is still possible to calculate

$$g_{ij} = \vec{X}_i \cdot \vec{X}_j.$$

Therefore, given the metric tensor, the dimension of the ambient Euclidean space of the surface or even how the surface sits in the ambient space is irrelevant. One simply refers to the coordinates in a given coordinate patch and the associated metric tensor for all calculations.

The concept of a differentiable manifold, the next step in the generalization of regular surfaces, avoids any reference to an ambient space. However, without an ambient space, one cannot really do geometry on a differentiable manifold (measure lengths, angles, volumes, curvatures, etc.) without a metric tensor. A manifold equipped with a metric tensor is called a *Riemannian manifold*. Chapters 3 through 5 in [22] study differentiable and Riemannian manifolds.

Example 8.5.1 (Flat Torus). The topological definition of a torus is a topological space that is homeomorphic to $\mathbb{S}^1 \times \mathbb{S}^1$, where $\mathbb{S}^1$ is a circle.

Throughout this book, we have used the following parametrization $\vec{X} : [0, 2\pi]^2 \to \mathbb{R}^3$ of the torus as a surface in $\mathbb{R}^3$:

$$\vec{X}(u, v) = ((R + r\cos v)\cos u, (R + r\cos v)\sin u, r\sin v).$$

The locus of this parametrization is the usual doughnut shape that is the surface of revolution of circle of radius r with center a distance R from the axis of revolution. As long as $0 < r < R$, the image of this parametrization is indeed a topological torus. In the context of differential geometry it is this shape that one calls a torus. As the exercises in Chapter 6 showed, this parametrization has a nontrivial metric tensor.

Problem 6.1.14 introduced the parametrization $\vec{X} : [0, 2\pi]^2 \to \mathbb{R}^4$ given by the functions

$$\vec{X}(u, v) = (\cos u, \sin u, \cos v, \sin v).$$

This parametrization explicitly traces out $\mathbb{S}^1 \times \mathbb{S}^1$. It is easy to see that the metric tensor (g_{ij}) associated to this parametrization is the identity matrix everywhere. This justifies the name of flat torus for this shape. In some intuitive sense, the reason why it is possible for a flat torus to exist is that, like the right circular cylinder in $\mathbb{R}^3$ that has $g_{ij} = \delta^i_j$, there is "enough room" in $\mathbb{R}^4$ to have a torus that has a metric tensor that is everywhere the identity matrix.

Example 8.5.2 (Poincaré Half-Plane). We consider the upper half-plane $H = \{(x, y) \in \mathbb{R}^2 \mid y > 0\}$ with the metric tensor

$$(g_{ij}) = \begin{pmatrix} y^{-2} & 0 \\ 0 & y^{-2} \end{pmatrix}.$$

Using the physics-style line element notation, the metric is written as $ds^2 = \frac{1}{y^2}(dx^2 + dy^2)$. The area element in H with this metric is

$$dS = \sqrt{\det(g)}\, dA = \frac{1}{y^2}\, dx\, dy.$$

So, for example, the surface integral of a function $f : H \to \mathbb{R}$ over a region $\mathcal{R}$ is

$$\iint_{\mathcal{R}} f\, dS = \iint_{\mathcal{R}} \frac{f(x, y)}{y^2}\, dx\, dy.$$

The area of a rectangle $\mathcal{R} = \{(x, y) \in H \mid x_1 \leq x \leq x_2,\ y_1 \leq y \leq y_2\}$ is

$$A = \int_{y_1}^{y_2} \int_{x_1}^{x_2} \frac{1}{y^2}\, dx\, dy = (x_2 - x_1)\left(\frac{1}{y_1} - \frac{1}{y_2}\right).$$

The Poincaré half-plane has many interesting properties that are explored in the exercises. One remarkable property is that the geodesic lines in H with the Poincaré metric are either vertical lines or half-circles with center on the x-axis.

!"#' $ % % & ' ( )) * +) , -)
.) !) * + %)) 4
//
,9 *) N eee) 5)

<: : 7 6

!"#' $ % % & ' < f) - +)))
+)44)))))J) * + f) # + 2 3 4

,9
1 9 7T
*)
( )
% 7

!"#' $ % % & ' < f) - +)))
+)44) , -) +) f) * + #)) ( )) (
f) # + Z ! # [Y % + ") 2 /3 4

,9
1 9 /7 8
*)) ( )

5- T<2 /3 * +) " - 9) J % * +

!"#' $ % % & ' < f) - +)))
+)44) * + f)) " E f) @) -
) 9 - Y % + 2 73 4

,9
1 9 < 8
*) N 8
( ) N 8
*) /
% N6

<: : N N

#

-))%g-)

E)
) Y "

: : 6
!'#' $ % % & ' O *)4) O
! -)4 & ' &)--) M& 'MO *)4 " O ! -)4 4

5- N 27 3)4 /)) + O *)4

5- N627N3 . - / 6 O 2L 8 8 3 "F

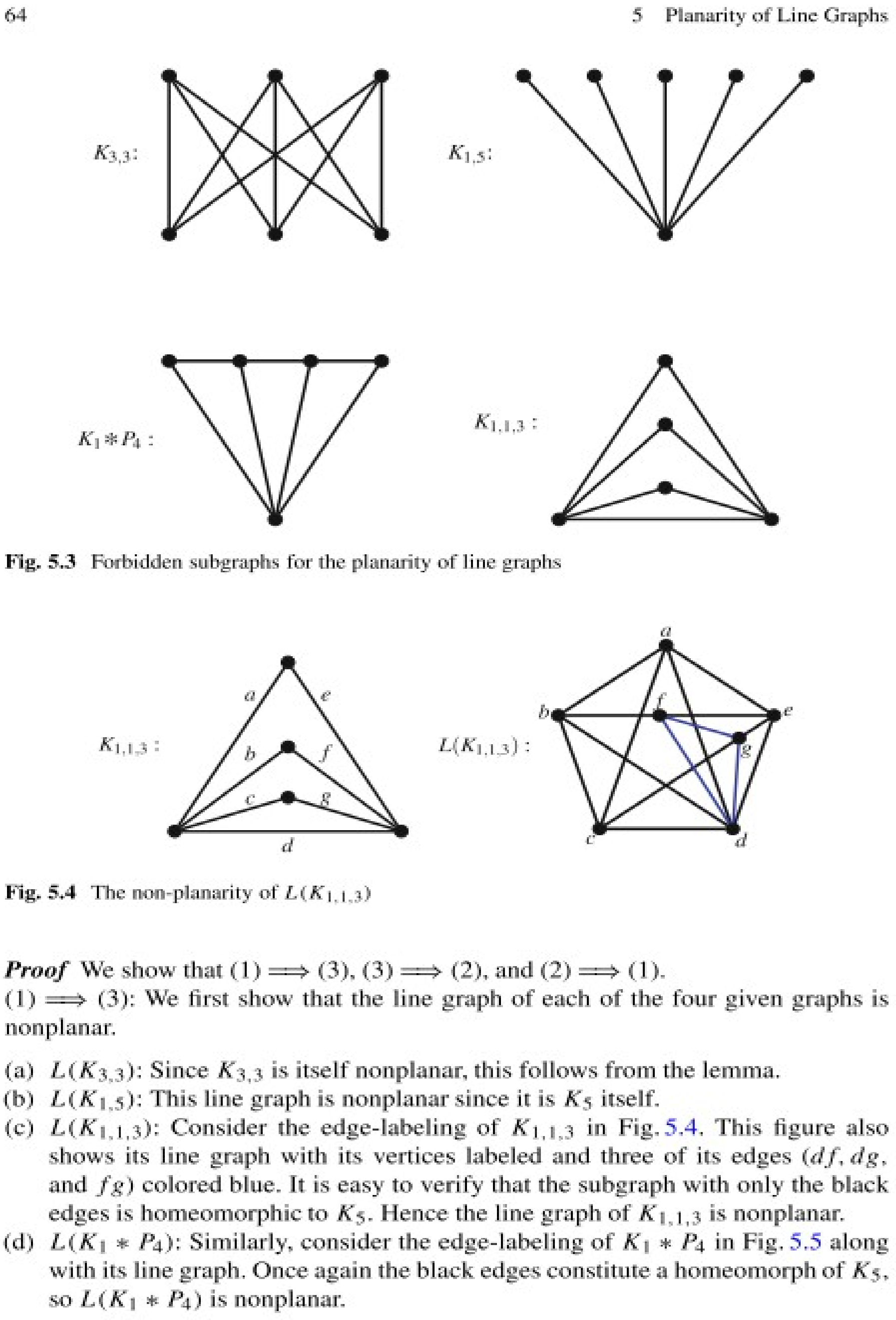

$K_{3,3}$:
$K_{1,5}$:
$K_1 * P_4$:
$K_{1,1,3}$:

Fig. 5.3 Forbidden subgraphs for the planarity of line graphs

$K_{1,1,3}$:
a
e
b
f
c
g
d
$L(K_{1,1,3})$:
a
b
f
e
g
c
d

Fig. 5.4 The non-planarity of $L(K_{1,1,3})$

Proof We show that (1) $\Longrightarrow$ (3), (3) $\Longrightarrow$ (2), and (2) $\Longrightarrow$ (1).
(1) $\Longrightarrow$ (3): We first show that the line graph of each of the four given graphs is nonplanar.

(a) $L(K_{3,3})$: Since $K_{3,3}$ is itself nonplanar, this follows from the lemma.
(b) $L(K_{1,5})$: This line graph is nonplanar since it is K_5 itself.
(c) $L(K_{1,1,3})$: Consider the edge-labeling of $K_{1,1,3}$ in Fig. 5.4. This figure also shows its line graph with its vertices labeled and three of its edges (df, dg, and fg) colored blue. It is easy to verify that the subgraph with only the black edges is homeomorphic to K_5. Hence the line graph of $K_{1,1,3}$ is nonplanar.
(d) $L(K_1 * P_4)$: Similarly, consider the edge-labeling of $K_1 * P_4$ in Fig. 5.5 along with its line graph. Once again the black edges constitute a homeomorph of K_5, so $L(K_1 * P_4)$ is nonplanar.

!"#' $ % % & ' " - 9)) * + " - 9))
* + 4 - Y 2 <7 3 4

,9 *) /

!"#' $ % % & '))) @) +
))) @) + E # !) @) + ") 8
@ % 2 <TT3 4

!"#' $ % % & '))) @) +
))) @) + EE # +) ") 8 @ %
2 <<73 4

!"#' $ % % & '))) @) +
))) @) + EEE # + ") 8 @ %
2 3 4

,9 *) O ' *) %)

!"#' $ % % & ' +)-) h 1 9 - @-
+)-) h 1 9 - ")) ') i j j ' ' F)') j .) ' #5)
2 < 73 4

!"#' $ % % & ' +)-) h 1 9 - @-
!))9) k) -) l @) m k l 2 <N73 4

!"#' $ % % & ' #) j
j j MO)M j M M)M4) = MZ [. jF M" j j M94 N' / N 4

!"#' $ % % & ' #) j
j j MO)M j M M)M4) = MZ [. jF M" j j M94 N' / T' 4

!"#' $ % % & ' #) j
j j MO)M j M M)M4) = MZ [. jF M" j j M94 N' / 4

!"#' $ % % & ' * Q) @) " - j %
* Q) @) @) 2 <763 4

!"#' $ % % & ' * Q) @) " - %
*) !) @) 2 <773 4

5- 2 /3 # Q) %)

. : : 6/ /

: : <

!"#' $ % % & ' $ @) O)- * +)
E) $ T T6N 4

5- <2 T3)4

f) J % 4)

5- 726/3)4 /
) J f)) i9 * +

,9
% N6
* + <
O)- 6<
@ 6

!"#' $ % % & ' . 1+4 9
! 4 . 1+4 9 ! 4 4 -)- &
1 9 - 2 <3 4

,9
( )
* + 6
@

!"#' $ % % & ' ! @=) !
@=) 8 ! !+)) +) "44) E ! 4

,9
( )
1 9
% <
* + T

!"#' $ % % & ' *)4 (- *)4
(- 4

,9
*)
@ 7

. : : N

: : / N /

!"#' $ % % & ' 1) 9 ' O)) $)
@ - - 1) 9 ' O)) $) @ - - 4

,9
*) 76

5- < 2 73 E) 9 %) 9) -
)) 4 8) 4)) 4 4 ; + 9 +) 4 2

) 9 *))) 3 9)) -) 4
)))% "))) n< : n)))
9 P %) 4) 4))
2") 3

!"#' $ % % & ' " 1 +
)) 8 S))) + 4 - 2 3 4

,9

*) <

5- N62 T 3 < 6 # ") E P D)
D) A) F - + 7 N E
") E P 8 D) 2 N//)3)))) 4) F +))
) % % P- 8))9 +8 = %)
)) + ' . ;) 4 8 -)%) 4)

$$\int_0^1 x^p \, dx = \frac{1}{p+1}$$

4 % - 4 9+ -)

$$\frac{0^p + 1^p + 2^p + \cdots + n^p}{n^p + n^p + n^p + \cdots + n^p} \to \frac{1}{p+1} \quad \text{as} \quad n \to \infty.$$

9.4 The *Arithmetica Infinitorum* of Wallis

Wallis's efforts to arithmetize geometry were noted in Section 7.6. In his *Arithmetica Infinitorum*, Wallis (1655a) made a similar attempt to arithmetize the theory of areas and volumes of curved figures. Some of his results were, understandably, equivalent to results already known. For example, he gave a proof that

$$\int_0^1 x^p\,dx = \frac{1}{p+1}$$

for positive integers p by showing that

$$\frac{0^p + 1^p + 2^p + \cdots + n^p}{n^p + n^p + n^p + \cdots + n^p} \to \frac{1}{p+1} \quad \text{as} \quad n \to \infty.$$

However, he made a new approach to fractional powers, finding $\int_0^1 x^{m/n}\,dx$ directly rather than by consideration of the curve $y^n = x^m$, as Fermat had done. He first found $\int_0^1 x^{1/2}\,dx$, $\int_0^1 x^{1/3}\,dx$, ... by considering the areas complementary to those under $y = x^2, y = x^3, \ldots$ (Figure 9.2), then guessed the results for other fractional powers by analogy with those already obtained.

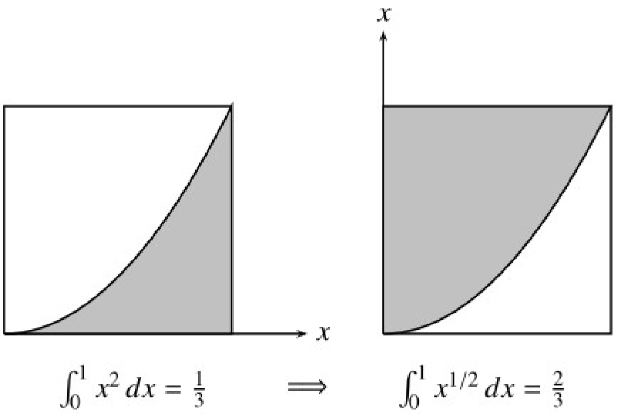
x
x
$\int_0^1 x^2\,dx = \frac{1}{3}$
$\Longrightarrow$
$\int_0^1 x^{1/2}\,dx = \frac{2}{3}$

Figure 9.2: Areas used by Wallis

Like other early contributors to calculus, Wallis was ambivalent about quantities that tended to zero, treating them as nonzero one minute and zero the next. For this he received a ferocious blast from his arch-enemy Thomas Hobbes: "Your scurvy book of *Arithmetica infinitorum*; where your indivisibles have nothing to do, but as they are supposed to have quantity, that is to say, to be *divisibles*" (Hobbes (1656), p. 301). Quite apart from this fault, which is easily remedied by limit arguments, the reasoning of Wallis is extremely incomplete by today's standards. Observing a pattern in formulas for $p = 1, 2, 3$, for example, he will immediately claim a formula for all positive integers p "by induction" and for fractional p "by interpolation." His boldness reached new heights toward the end of the *Arithmetica infinitorum* in deriving his famous infinite product formula,

$$\frac{\pi}{4} = \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdot \frac{6}{7} \cdot \cdots.$$

An exposition of his reasoning may be found in Edwards (1979), pp. 171–176, where it is described as "one of the more audacious investigations by analogy and intuition that has ever yielded a correct result."

However, we must bear in mind that Wallis was offering primarily a method of discovery, and what a discovery he made! His infinite product for π was not the first ever given, since Viète (1593) had discovered

$$\begin{aligned}\frac{2}{\pi} &= \cos\frac{\pi}{4}\cos\frac{\pi}{8}\cos\frac{\pi}{16}\cdots\\ &= \sqrt{\frac{1}{2}} \cdot \sqrt{\frac{1}{2}\left(1 + \sqrt{\frac{1}{2}}\right)} \cdot \sqrt{\frac{1}{2}\left[1 + \sqrt{\frac{1}{2}\left(1 + \sqrt{\frac{1}{2}}\right)}\right]} \cdot \cdots.\end{aligned}$$

However, the formula of Viète is based on a clever but simple trick (see exercises), whereas that of Wallis is of deeper significance. By relating π to the integers through a sequence of rational operations, Wallis uncovered a sequence of fractions, obtained by terminating the product at the nth factor, that he called "hypergeometric." Similar sequences were later found to occur as coefficients in series expansions of many functions, which led to a broad class of functions being called "hypergeometric" by Gauss. Also, Wallis's product was closely related to two other beautiful formulas for π based on sequences of rational operations:

$$\frac{4}{\pi} = 1 + \cfrac{1^2}{2 + \cfrac{3^2}{2 + \cfrac{5^2}{2 + \cfrac{7^2}{2 + \cdots}}}}$$

and

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots .$$

The continued fraction was obtained by Brouncker from Wallis's product and also published in Wallis (1655b). The series is a special case of the series

$$\tan^{-1} x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots$$

discovered by the Indian mathematician Mādhava in the 15th century (see Section 10.1) and rediscovered by Newton, Gregory, and Leibniz. Euler (1748a), p. 311, gave a direct transformation of the series for $\pi/4$ into Brouncker's continued fraction. Besides setting off this spectacular chain reaction, Wallis's method of "interpolation" had important consequences in the work of Newton, who used it to discover the general binomial theorem (Section 10.2).

EXERCISES

9.4.1 Use the identity $\sin x = 2\sin(x/2)\cos(x/2)$ to show that

$$\frac{\sin x}{2^n \sin(x/2^n)} = \cos\frac{x}{2}\cos\frac{x}{2^2}\cdots\cos\frac{x}{2^n},$$

whence

$$\frac{\sin x}{x} = \cos\frac{x}{2}\cos\frac{x}{2^2}\cos\frac{x}{2^3}\cdots .$$

9.4.2 Deduce Viète's product by substituting $x = \pi/2$.

The equation relating the series for $\pi/4$ to the continued fraction for $4/\pi$, namely

$$1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots = \cfrac{1}{1 + \cfrac{1^2}{2 + \cfrac{3^2}{2 + \cfrac{5^2}{2 + \cfrac{7^2}{2 + \cdots}}}}}$$

follows immediately from a more general equation

$$\frac{1}{A} - \frac{1}{B} + \frac{1}{C} - \frac{1}{D} + \cdots = \cfrac{1}{A + \cfrac{A^2}{B - A + \cfrac{B^2}{C - B + \cfrac{C^2}{D - C + \cdots}}}}$$

where x_3 is none other than the x-coordinate of the third point on

$$y^2 = 4x^3 - g_2x - g_3$$

of the straight line through (x_1, y_1) and (x_2, y_2) (see Section 11.6). Now that we know, from Section 12.2, that this curve is parameterized by $x = \wp(u)$, $y = \wp'(u)$, defined by inverting the integral, some connection between the geometry of the curve and the addition theorem is understandable. But the stunning simplicity of the relationship seems to demand a deeper explanation. This lies in the realm of complex numbers, which we shall enter briefly in the next section and more thoroughly in Sections 16.4 and 16.5.

EXERCISES

12.5.1 Show that the substitution $t = 1/u$ transforms

$$\frac{dt}{\sqrt{(t-a)(t-b)(t-c)}} \quad \text{into} \quad \frac{-du}{\sqrt{u(1-ua)(1-ub)(1-uc)}}.$$

Conversely, we can transform quartic polynomials under the square root sign to cubics, even in cases where the quartic is not of the form obtained in Exercise 12.5.1.

12.5.2 Transform

$$\frac{dt}{\sqrt{1-t^4}} \quad \text{into} \quad \frac{du}{\sqrt{\text{cubic polynomial in } u}}$$

by making a suitable substitution for t.

12.6 Elliptic Functions

The idea of inverting elliptic integrals to obtain elliptic functions is due to Gauss, Abel, and Jacobi. Gauss had the idea in the late 1790s but did not publish it; Abel had the idea in 1823 and published it in 1827, independently of Gauss. Jacobi's independence is not quite so clear. He seems to have been approaching the idea of inversion in 1827, but he was stung into action only by the appearance of Abel's paper. At any rate, his ideas subsequently developed at an explosive rate, and he published the first book on elliptic functions, the *Fundamenta nova theoriae functionum ellipticarum*, two years later (Jacobi (1829)).

Gauss first considered inverting an elliptic integral in 1796, in the case of $\int dt/\sqrt{1-t^3}$. The following year he inverted the lemniscatic integral and made better progress. Defining the "lemniscatic sine function" $x = sl(u)$ by

$$u = \int_0^x \frac{dt}{\sqrt{1-t^4}},$$

he found that this function is periodic, like the sine, with period

$$2\varpi = 4\int_0^1 \frac{dt}{\sqrt{1-t^4}}.$$

Gauss also noticed that $sl(u)$ invites complex arguments, since it follows from $i^2 = -1$ that

$$\frac{d(it)}{\sqrt{1-(it)^4}} = i\frac{dt}{\sqrt{1-t^4}};$$

hence $sl(iu) = isl(u)$ and the lemniscatic sine has a second period $2i\varpi$. Thus Gauss discovered *double periodicity*, one of the key properties of the elliptic functions, though at first he did not realize its universality. The scope and importance of elliptic functions hit him on May 30, 1799, when he discovered an extraordinary numerical coincidence. His diary entry of that day reads:

> We have established that the arithmetic-geometric mean between 1 and $\sqrt{2}$ is π/ϖ to 11 places; the demonstration of this fact will surely open up an entirely new field of analysis.

Gauss had been fascinated by the arithmetic–geometric mean since 1791, when he was 14. The arithmetic–geometric mean of two positive numbers a and b is the common limit, $\mathrm{agM}(a, b)$, of the two sequences $\{a_n\}$ and $\{b_n\}$ defined by

$$\begin{aligned} a_0 &= a, & b_0 &= b, \\ a_{n+1} &= \tfrac{a_n+b_n}{2}, & b_{n+1} &= \sqrt{a_n b_n}. \end{aligned}$$

For more information on the theory and history of the agM function, see Cox (1984).

It is indeed true that $\mathrm{agM}(1, \sqrt{2}) = \pi/\varpi$, as Gauss soon proved, and the "entirely new field of analysis" he created from the merger of these ideas was extraordinarily rich. It encompassed elliptic functions in general, the theta functions later rediscovered by Jacobi, and the modular functions later rediscovered by Klein. The theory was not clearly improved until the

1850s, when Riemann showed that double periodicity becomes obvious when elliptic integrals are placed in a suitable geometric setting.

Unfortunately, Gauss released virtually none of his results on elliptic functions. Apart from publishing an expression for $\mathrm{agM}(a, b)$ as an elliptic integral (Gauss (1818)), he did nothing until Abel's results appeared in 1827—then promptly claimed them as his own. He wrote to Bessel (Gauss (1828)):

> I shall most likely not soon prepare my investigations on the transcendental functions which I have had for many years—since 1798. …Herr Abel has now, as I see, anticipated me and relieved me of the burden in regard to one third of these matters.

It was disingenuous of Gauss to claim he had more results than Abel, because Abel also had results unknown to Gauss. True, Gauss had priority on the key ideas of inversion and double periodicity, but priority isn't everything, as Gauss himself perhaps knew. His own cherished discovery of the relation between agM and elliptic integrals had not only been found earlier, but even published by Lagrange (1785).

EXERCISES

The following exercises show how the lemniscatic sine and its derivative are quite analogous to the ordinary sine and its derivative, the cosine.

12.6.1 Show that $sl'(u) = \sqrt{1 - sl^4(u)}$.

12.6.2 Deduce from the Euler addition theorem (Section 12.4) that

$$sl(u+v) = \frac{sl(u)sl'(v) + sl(v)sl'(u)}{1 + sl^2(u)sl^2(v)}.$$

12.7 A Postscript on the Lemniscate

The duplication of the arc of the lemniscate had some interesting consequences for the lemniscate itself. Fagnano showed, by similar arguments, that a quadrant of the lemniscate can be divided into two, three, or five equal arcs by ruler and compass (see Ayoub (1984)). This raised a question: for which n can the lemniscate be divided into n equal parts by ruler and compass? Recall from Section 2.3 that the corresponding question for the circle had been answered by Gauss (1801), Art. 366. As mentioned

E 4)

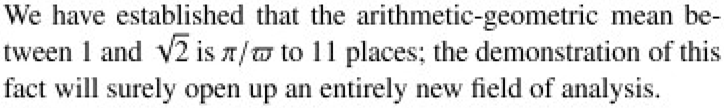

> We have established that the arithmetic-geometric mean between 1 and $\sqrt{2}$ is π/ϖ to 11 places; the demonstration of this fact will surely open up an entirely new field of analysis.

@) %)) F 8 % 5

5- //2 7 3 /)) %

5- /<2 7N3 #)))) S)' 9 & A % 8) '
4) % 4) P U + 4
)) %) 9+))
4) S)' 9 & 2 N<63) %) P
)) =))

$$ds = \frac{dx}{\sqrt{1 - x^4}} .$$

4 -) -) +8))
)) - ;4 9+ 4 +) -) #
) 9 - - % -)) -) 8
4) % .)-)) *)
))4 @ A % -) 4 -)))
9+)) @ 2 76 3 -)% 4))) +)%
4 2. - <3 #) - P
) 4 4 8 - 4 +) 4
P 4 -) 8) - 4 2) 4
9 - 4 3

5- T72 63 6 7 # A" 9) *)

intermediate value theorem and extreme value theorem. This completed not only the second proof of Gauss but also the proof of d'Alembert, as we shall see in the next section.

EXERCISES

Complex roots of an equation with real coefficients occur in conjugate pairs because of the fundamental properties of conjugates.

14.6.1 Show directly from the definition $\overline{u + iv} = u - iv$ that

$$\overline{z_1 + z_2} = \overline{z_1} + \overline{z_2} \quad \text{and} \quad \overline{z_1 \cdot z_2} = \overline{z_1} \cdot \overline{z_2}$$

for any complex numbers z_1, z_2.

14.6.2 Deduce from Exercise 14.6.1 that $p(\overline{z}) = \overline{p(z)}$ for any polynomial $p(z)$ with real coefficients, and hence that the complex roots of $p(z) = 0$ occur in conjugate pairs.

14.7 The Proofs of d'Alembert and Gauss

The key to d'Alembert's proof is a proposition now known as *d'Alembert's lemma*: if $p(z)$ is a nonconstant polynomial function and $p(z_0) \neq 0$, then any neighborhood of z_0 contains a point z_1 such that $|p(z_1)| < |p(z_0)|$.

The proof of this lemma offered by d'Alembert depended on solving the equation $w = p(z)$ for z as a fractional power series in w. As mentioned in Section 9.5, such a solution was claimed by Newton (1671), but it was made clear and rigorous only by Puiseux (1850). Thus d'Alembert's argument did not stand on solid ground, and in any case it was unnecessarily complicated.

A simple elementary proof of d'Alembert's lemma was given by Argand (1806). Argand was one of the co-discoverers of the geometric representation of complex numbers (probably the first was Wessel (1797), but his work remained almost unknown for 100 years), and he offered the following proof as an illustration of the effectiveness of the representation.

The value of $p(z_0) = x_0 + iy_0$ is interpreted as the point (x_0, y_0) in the plane, so that $|p(z_0)|$ is the distance of (x_0, y_0) from the origin. We wish to find a Δz such that $p(z_0 + \Delta z)$ is nearer to the origin than $p(z_0)$. If

$$p(z) = a_0 z^n + a_1 z^{n-1} + \cdots + a_n,$$

then

$$\begin{aligned} p(z_0+\Delta z) &= a_0(z_0+\Delta z)^n + a_1(z_0+\Delta z)^{n-1} + \cdots + a_n \\ &= a_0 z_0^n + a_1 z_0^{n-1} + \cdots + a_n + A_1\Delta z + A_2(\Delta z)^2 + \cdots + A_n(\Delta z)^n \\ &\quad \text{for some constants } A_i \text{ depending on } z_0, \text{ not all zero,} \\ &\quad \text{because } p \text{ is not constant} \\ &= p(z_0) + A\Delta z + \varepsilon, \end{aligned}$$

where $A = A_i(\Delta z)^i$ contains the first nonzero A_i and $|\varepsilon|$ is small compared with $|A\Delta z|$ when $|\Delta z|$ is small (because ε contains higher powers of Δz). It is then clear (Figure 14.5) that by choosing the direction of Δz so that $A\Delta z$ is opposite in direction to $p(z_0)$, we get $|p(z_0+\Delta z)| < |p(z_0)|$. This completes the proof of d'Alembert's lemma.

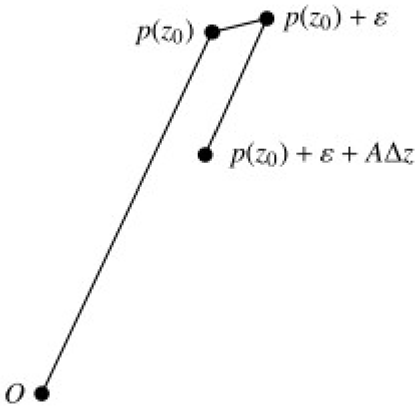

Figure 14.5: Construction for d'Alembert's lemma

To complete the proof of the fundamental theorem of algebra, take an arbitrary polynomial p and consider the continuous function $|p(z)|$. Since $p(z) \approx a_0 z^n$ for $|z|$ large, $|p(z)|$ increases with $|z|$ outside a sufficiently large circle $|z| = R$. We now get a z for which $|p(z)| = 0$ from the extreme value theorem of Weierstrass (1874); a continuous function on a closed bounded set assumes maximum and minimum values. By this theorem, $|p(z)|$ takes a minimum value for $|z| \le R$. The minimum is ≥ 0 by definition, and if it is > 0 we get a contradiction by d'Alembert's lemma: either a point z with $|z| \le R$ where $|p(z)|$ takes a value less than its minimum or a point z with $|z| > R$ where $|p(z)|$ is less than its values on $|z| = R$. Thus there is a point z where $|p(z)| = 0$ and hence $p(z) = 0$.

The proof of Gauss also used the fact that $p(z)$ behaves like its highest-degree term a_0z^n for $|z|$ large and likewise relied on a continuity argument to show that $p(z) = 0$ inside some circle $|z| = R$. Gauss considered the real and imaginary parts of $p(z)$, $\mathrm{Re}[p(z)]$ and $\mathrm{Im}[p(z)]$, and investigated the curves

$$\mathrm{Re}[p(z)] = 0 \qquad \text{and} \qquad \mathrm{Im}[p(z)] = 0.$$

(These are easily seen to be algebraic curves $p_1(x, y) = 0$ and $p_2(x, y) = 0$ by expanding the powers $z^k = (x + iy)^k$ and collecting real and imaginary terms.) His aim was to find a point where these curves meet, because at such a point

$$0 = \mathrm{Re}[p(z)] = \mathrm{Im}[p(z)] = p(z).$$

For $|z|$ large, the curves are close to the curves $\mathrm{Re}(a_0z^n)$ and $\mathrm{Im}(a_0z^n) = 0$, which are families of straight lines through the origin. Moreover, the lines where $\mathrm{Re}(a_0z^n) = 0$ *alternate* with those where $\mathrm{Im}(a_0z^n) = 0$ as one makes a circuit around the origin. For example, Figure 14.6 shows $\mathrm{Re}(z^2) = 0$ and $\mathrm{Im}(z^2) = 0$ as alternate solid and dashed lines. It follows that the curves $\mathrm{Re}[p(z)] = 0$ and $\mathrm{Im}[p(z)] = 0$ meet a sufficiently large circle $|z| = R$ *alternately*. Up to this point the argument is comparable to d'Alembert's lemma, and it can be made just as rigorous.

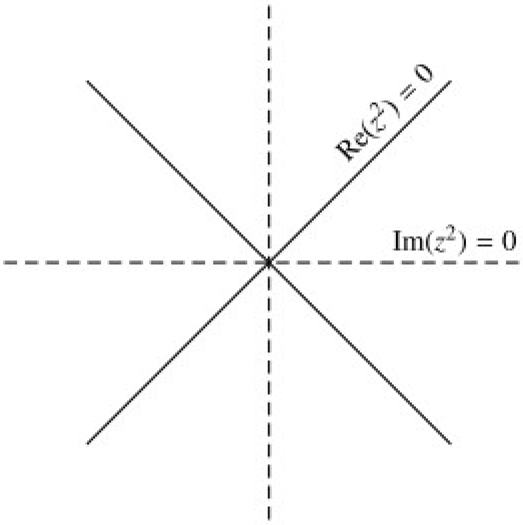

Figure 14.6: Lines for the Gauss proof

To complete this proof we have to show that the curves meet inside the circle, and this is the step Gauss thought nobody could doubt. He assumed that the separate pieces of the algebraic curve $\mathrm{Re}[p(z)] = 0$ outside the circle $|z| = R$ would join inside the circle, as would the separate pieces of $\mathrm{Im}[p(z)] = 0$. Then, since the pieces of $\mathrm{Re}[p(z)] = 0$ alternate with those of $\mathrm{Im}[p(z)]$ on $|z| = R$, it would be "patently absurd" for their connecting pieces inside the circle not to meet. One has only to visualize a situation like that seen in Figure 14.7 to feel sure that Gauss was right. However, the existence of the connecting pieces is extremely hard to prove (and proving that they meet is not trivial either, being at least as hard as the intermediate value theorem). The first proof was given by Ostrowski (1920).

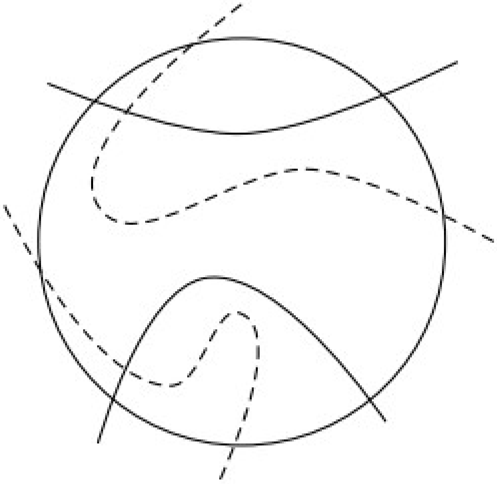

Figure 14.7: Curves for the Gauss proof

From our present perspective, d'Alembert's route to the fundamental theorem of algebra seems basically easy because it proceeds through general properties of continuous functions. The route of Gauss, although appearing equally easy from a distance, goes through the still-unfamiliar territory of real algebraic curves. The intersections of real algebraic curves are harder to understand than the intersections of complex algebraic curves, and in retrospect they are harder to understand than the fundamental theorem of algebra. Indeed, as we shall see in the next chapter, the fundamental theorem gives us Bézout's theorem, which in turn settles the problem of counting the intersections of complex algebraic curves.

EXERCISES

The expression in d'Alembert's lemma for $p(z_0+\Delta z)$ is an instance of *Taylor's series*, previously discussed in Section 10.2. When the function is a polynomial p, as here, its Taylor series is finite because p has only finitely many nonzero derivatives.

14.7.1 Show that $A_1 = na_0z_0^{n-1} + (n-1)a_1z_0^{n-2} + \cdots + a_{n-1}$ and that the latter expression is $p'(z_0)$.

14.7.2 Show that $A_2 = \frac{n(n-1)}{2}a_0z_0^{n-2} + \frac{(n-1)(n-2)}{2}a_1z_0^{n-3} + \cdots + a_{n-2}$ and that the latter expression is $p''(z_0)/2$.

14.7.3 Using the binomial theorem, show that $A_k = p^{(k)}(z_0)/k!$, and hence that

$$p(z_0 + \Delta z) = a_0z_0^n + a_1z_0^{n-1} + \cdots + a_n + A_1\Delta z + A_2(\Delta z)^2 + \cdots + A_n(\Delta z)^n$$

is an instance of the Taylor series formula.

14.8 Biographical Notes: d'Alembert

Jean le Rond d'Alembert (Figure 14.8) was born in Paris in 1717 and died there in 1783. He was the illegitimate son of the Chevalier Destouches-Canon, a cavalry officer, and salon hostess Madame de Tencin. His mother abandoned him at birth near the church of St. Jean-le-Rond in the cloisters of Notre Dame, and so he was christened Jean le Rond, following the custom for foundlings. He was subsequently located by his father, who found a home for him with a glazier named Rousseau and his wife. The name d'Alembert came later, for reasons that are unclear.

The Rousseaus must have been devoted foster parents, for d'Alembert lived with them until 1765. He received an annuity from his father, who also arranged for him to be educated at the Jansenist Collège de Quatre-Nations in Paris. There he received a good grounding in mathematics and developed a permanent distaste for theology. After brief studies in law and medicine he turned to mathematics in 1739.

In that year he began sending communications to the Académie des Sciences, and his ambition and talent rapidly carried him to fame. He became a member of the Académie in 1741 and published his best-known work, the *Traité de dynamique*, in 1743. Having struggled to the top from humble beginnings, d'Alembert did not want to lose his position. Once in the Académie, his struggle was to stay ahead of his rivals. Whether by accident or inborn competitiveness, d'Alembert always seemed to be working on the same problems as other top mathematicians—initially Clairaut, 5- <T2 /3 / # 4 ; J % O

D) T /)) -) 4) P +) (
(o ( ) %) =)) % + ') E 8)
) 4 J % 4) 4 J % 4) (
)
) - 4 4)) -)) 4 8)
)))) 4) 4 4)) 4
(
) 4 ; G H) 9
%) F)) +) 8)) *) A 4
))) - 9)\ 4 ; G4) H o
)) %) + 4 9 %) F
)%) ;) 4) 8 P %
)44)) 4 J % E T /) +
O8) F) 4) 4) - - - 8) ;
)) 4 J % G4 H) - -
,8 4) - ,) O # F)) +
4 4 O8) F)) + 4
) P + O D)-))
+ =)) % 8 4 + 4 -) 8 = %) 9
) 4 J %)) 2. - / 3

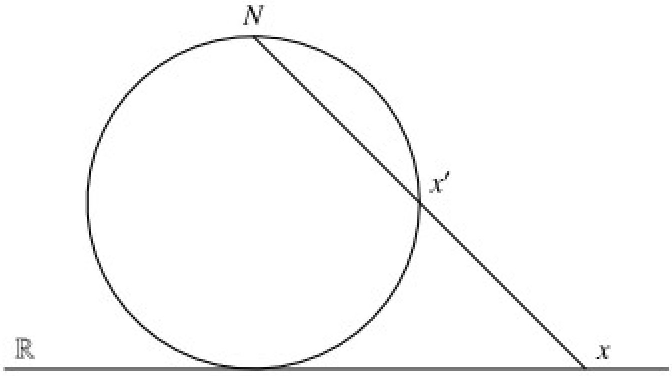

- $)' 9 4 4)) 8) 9
4 8 O p (#) 4
9 - $) 4 @) F
) 4 ;K p :
$ 8
F) 4 $ # 4 J % 4
(8) (
8 4 -) +)) 8
)) 4 9
% 8)) 4 J % 4 (4 -) +))
4) - G4 H) G)44) H) P +8
8) - (8) ; P + 8 ;K
) 4 8 $8

$p \leq mn$ in the single variable x/y. But then

$$\begin{aligned} r_{mn}(x, y) &= y^{mn-p} \prod_{i=1}^{p} (b_i x - a_i y) \\ &= \prod_{i=1}^{mn} (b_i x - a_i y) \end{aligned}$$

since each factor y in front (if any) is trivially of the form $b_i x - a_i y$.

It follows that the equation $r_{mn}(x, y) = 0$ has mn solutions, and hence there are mn intersections of C_m and C_n, counting multiplicities.

Exercises

15.1.1 Show that $y = \varepsilon x$ meets $y = x^n$ in n distinct points when $\varepsilon \neq 0$, and list them (for example, with the help of de Moivre's theorem).

If a curve K has a double point at O, then a line $y = tx$ may have double contact with K at O even though nearby lines $y = (t + \varepsilon)x$ do not meet K at nearby points other than O. In this case the double contact may be explained as contact with the two branches of the curve at O.

15.1.2 Consider the lines $y = tx$ through the double point O of $y^2 = x^2(x + 1)$. Show that each such line has double contact with the curve at O, except when $t = \pm 1$. How do you account for the multiplicities when $t = \pm 1$?

15.1.3 Show that $y = tx$ also has double contact with $y^2 = x^3$ at its cusp point O. Try to explain this by viewing $y^2 = x^3$ as the result of "shrinking the loop" of $y^2 = x^2(x + \varepsilon)$ (letting $\varepsilon \to 0$).

15.1.4 Show that the line $y = tx$ has double contact at O with the lemniscate $(x^2 + y^2)^2 = x^2 - y^2$ except for two values of t, for which it has quadruple contact.

15.1.5 Explain the multiplicities found in Exercise 15.1.4 with the help of the known shape of the lemniscate (Figure 12.1).

15.2 The Complex Projective Line

We saw in Section 8.5 that adding a point at infinity to the real line $\mathbb{R}$ in $\mathbb{R} \times \mathbb{R}$ forms a closed curve that is qualitatively like a circle. Indeed, a real projective line in the sphere model of the real projective plane $\mathbb{RP}^2$ has much the same geometric properties as a great circle on a sphere, after one allows for the fact that antipodal points on the sphere are the same point on

$\mathbb{RP}^2$. The situation with the complex "line" $\mathbb{C}$ is similar but more difficult to visualize. $\mathbb{C}$ is already two-dimensional, as we saw in Gauss's proof of the fundamental theorem of algebra; hence the complex "plane" $\mathbb{C} \times \mathbb{C}$ is four-dimensional and virtually impossible to visualize.

To avoid an excursion into four-dimensional space, we first revise our approach to the real projective line. In Section 8.5 we considered ordinary lines L, in a horizontal plane not passing through the origin, and extended each to a projective line whose "points" are the lines through the origin O, in the plane through O and L. The nonhorizontal lines in this family correspond to points of L, and the horizontal line in the family to the point at infinity of L. We now use this construction again to demonstrate directly the qualitative, or more precisely *topological*, equivalence between a projective line and a circle (Figure 15.3).

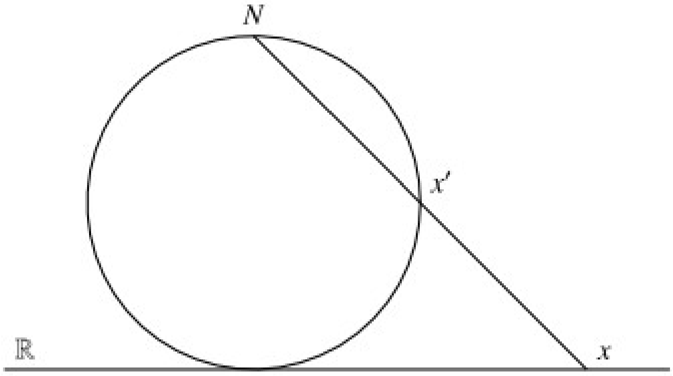

Figure 15.3: The real projective line

The origin N is taken to be the top point of a circle that, at its bottom point, touches our line $L = \mathbb{R}$. There is a continuous one-to-one correspondence between lines through N and points of the circle. Each nonhorizontal line corresponds to its intersection $x' \neq N$ with the circle, while the horizontal line corresponds to N itself. Thus the projective completion of $\mathbb{R}$, which we now call $\mathbb{RP}^1$, is *topologically the same* as the circle, in the sense that there is a continuous one-to-one correspondence between them. Moreover, we can understand projective completion of $\mathbb{R}$ topologically as a process of adding one "point" that is "approached" as one tends to infinity, in either direction, along $\mathbb{R}$, for as x tends to infinity in either direction, x' tends to the same point, N, on the circle.

We can now view projective completion of $\mathbb{C}$ in the same way using Figure 15.4, which shows the so-called *stereographic projection* of the plane $\mathbb{C}$ into a sphere. Each point $z \in \mathbb{C}$ is projected to a point z' on the tangential sphere S by the ray through z and the north pole N of S. This establishes a continuous one-to-one correspondence between points z of $\mathbb{C}$ and points $z' \neq N$ on S. Moreover, as z tends to infinity in any direction, z' tends to N; hence the projective completion of $\mathbb{CP}^1$ of $\mathbb{C}$ is topologically the same as the complete sphere S, with the point at ∞ of $\mathbb{C}$ corresponding to N.

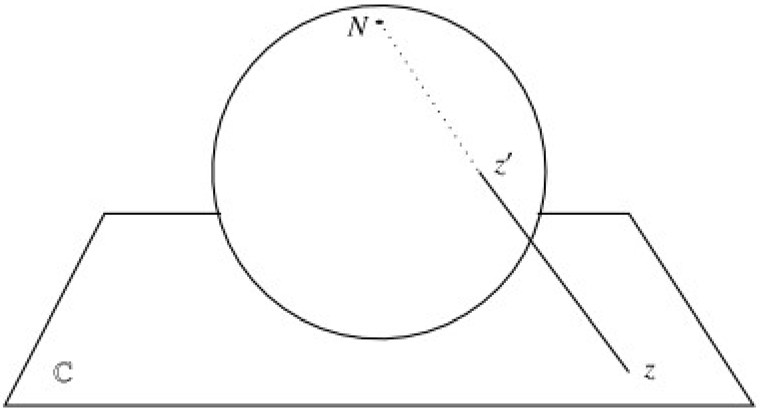

Figure 15.4: The complex projective line

Since one also wants to complete $\mathbb{C}$ by a point ∞ in this way for complex analysis, geometry and analysis are both served by passing from $\mathbb{C}$ to $\mathbb{CP}^1$. Gauss seems to have been the first to appreciate the advantages of $\mathbb{C} \cup \{\infty\}$ over $\mathbb{C}$; hence one often calls $\mathbb{CP}^1$ the *Gauss sphere* in analysis. (Unfortunately, only a few unpublished, undated fragments of Gauss's work on this topic seem to have survived; see Gauss (1819).) Algebraic geometers call $\mathbb{CP}^1$ the (complex) projective line, since it is the formal equivalent of a real line, even though it is topologically a surface. Similarly, complex curves are topologically surfaces, known to analysts as *Riemann surfaces*, though algebraic geometers prefer to call them "curves."

The "surface" viewpoint is helpful when studying intrinsic properties of complex curves. For example, *genus* (introduced in connection with parameterization in Sections 11.3 to 11.5) turns out to have a very simple meaning in the topology of surfaces (see Section 15.4). On the other hand, the "curve" viewpoint is helpful when studying intersections of curves and their embedding in $\mathbb{C}\times\mathbb{C}$ or its projective completion $\mathbb{CP}^2$. Instead of trying to imagine two planes meeting in a single point of $\mathbb{C}\times\mathbb{C}$, for example, it is better to imagine the intersection as analogous to that of real lines in a real plane—as the single solution of two linear equations. After all, we are working with $\mathbb{C}$ to remove anomalies that occur with $\mathbb{R}$, not for the sake of doing something different, and we expect that much of the behavior of real curves will recur with complex ones.

Exercises

Since addition and multiplication are continuous functions, it is quite easy to find one-to-one continuous maps between certain complex algebraic curves and the sphere.

15.2.1 Show that the projective completion of the curve $Y = X^2$ is topologically a sphere by considering its parameterization

$$X = t, \quad Y = t^2,$$

where t ranges over the sphere $\mathbb{C} \cup \{\infty\}$. Namely, show that the mapping $t \mapsto (t, t^2)$ is one-to-one and continuous.

15.2.2 Similarly show that the projective completion of $Y^2 = X^3$ is topologically a sphere by considering its parameterization

$$X = t^2, \quad Y = t^3$$

and the continuous mapping $t \mapsto (t^2, t^3)$.

15.2.3 Consider the mapping of the t sphere onto the projective completion of $Y^2 = X^2(X + 1)$ defined by $t \mapsto P(t)$, where $P(t)$ is the third intersection of the curve with the line $Y = tX$ through the double point (found in Exercise 7.4.2).

Show that this mapping is continuous and that it is one-to-one except at the points $t = \pm 1$, which are both mapped to the point O on the curve. Conclude that the curve is topologically the same as a sphere with two points identified (Figure 15.5).

15.3 Branch Points

The key to the topological form of a complex curve $p(x, y) = 0$ lies in its *branch points*, the points α where the Newton–Puiseux expansion of y begins with a fractional power of $(x - \alpha)$ (see Section 10.5). The nature of

5- 2 3 N 4 ; .

16.1 Complex Functions

When Bombelli (1572) introduced complex numbers, he implicitly introduced complex functions as well. The solution y of the cubic equation $y^3 = py + q$,

$$y = \sqrt[3]{\frac{q}{2} + \sqrt{\left(\frac{q}{2}\right)^2 - \left(\frac{p}{3}\right)^3}} + \sqrt[3]{\frac{q}{2} - \sqrt{\left(\frac{q}{2}\right)^2 - \left(\frac{p}{3}\right)^3}},$$

involves the cube root of a complex argument when $(q/2)^2 < (p/3)^3$. It could have been a revelation to see that complex numbers explain the coincidence of algebraic (Cardano) and geometric (Viète) solutions of the cubic equation, and more generally the Leibniz–de Moivre theorem that

$$x = \frac{1}{2}\sqrt[n]{y + \sqrt{y^2 - 1}} + \frac{1}{2}\sqrt[n]{y - \sqrt{y^2 - 1}},$$

when $x = \sin\theta$ and $y = \sin n\theta$ (Section 6.6). In the case of the cubic, this revelation can now be savored in Needham (1997), pp. 59–60. But mathematicians were not concerned about the meaning of these complex functions as long as they produced results that could be checked by algebra.

The need to *understand* complex functions became pressing only with transcendental functions, particularly those defined by integration. A key example is the logarithm function, which comes from integrating $dz/(1+z)$. Once this function was understood, the reason for algebraic miracles like the Leibniz–de Moivre theorem became much clearer.

Johann Bernoulli (1702) opened the story of the complex logarithm when he noted that

$$\frac{dz}{1+z^2} = \frac{dz}{2(1 + z\sqrt{-1})} + \frac{dz}{2(1 - z\sqrt{-1})}$$

and drew the conclusion that "imaginary logarithms express real circular sectors." He did not actually perform the integration, but he may have found

$$\tan^{-1} z = \frac{1}{2i}\log\frac{i-z}{i+z},$$

since Euler gives him credit for a similar formula when writing to him in Euler (1728b). However, this may have been the young Euler's deference to his former teacher, because Johann Bernoulli showed poor understanding of logarithms as the correspondence continued. He persistently claimed that $\log(-x) = \log(x)$ on the grounds that

$$\frac{d}{dx}\log(-x) = \frac{1}{x} = \frac{d}{dx}\log(x)$$

despite a reminder from Euler (1728b) that equality of derivatives does not imply equality of integrals. Euler went on to suggest that the complex logarithm had infinitely many values.

In the meantime, Cotes (1714) had also discovered a relation between complex logarithms and circular functions:

$$\log(\cos x + i\sin x) = ix.$$

Recognizing the importance of this result, he entitled his work *Harmonia mensurarum* (Harmony of measures). The "measures" in question were the logarithm and inverse tangent functions, which "measure" the hyperbola and the circle, respectively, via the integrals $\int dx/(1+x)$ and $\int dx/(1+x^2)$. A wide class of integrals had been reduced to these two types, but it was not understood why two apparently unrelated "measures" should be required. Cotes's result was the first (apart from the near-miss of Johann Bernoulli (1702)) to relate the two, showing that in the wider domain of complex functions the logarithm and inverse circular functions are essentially the same.

The most compact statement of their relationship was attained around 1740, when Euler shifted attention from the logarithm function to its inverse, the exponential function. The definitive formula

$$e^{ix} = \cos x + i\sin x$$

was first published by Euler (1748a), who derived it by comparing series expansions of both sides. Euler's formulation in terms of the single-valued function e^{ix} gave a simple explanation of the many values of the logarithm (which Cotes had missed) as a consequence of the periodicity of cos and sin. A direct explanation, based on the definition of log as an integral, was not possible until Gauss (1811) clarified the meaning of complex integrals and pointed out their dependence on the path of integration (see Section 16.3).

Euler's formula also shows that

$$(\cos x + i\sin x)^n = e^{inx} = \cos nx + i\sin nx$$

and hence gives a deeper explanation of the Leibniz–de Moivre formula. More generally, the addition theorems for cos and sin (Section 12.4) could be seen as consequences of the much simpler addition formula for the exponential function

$$e^{u+v} = e^u \cdot e^v.$$

The imaginary function e^{ix} was so much more coherent than its real constituents $\cos x$ and $\sin x$ that it was difficult to do without it, and Euler's formula gave mathematicians a strong push toward the eventual acceptance of complex numbers. A more detailed account of the role of the logarithm and exponential functions in the development of complex numbers may be found in Cajori (1913).

At almost the same time that Euler elucidated cos and sin, d'Alembert found many real functions occurring naturally in pairs as the real and imaginary parts of complex functions—in hydrodynamics. As mentioned in Section 13.7, d'Alembert (1752) discovered the equations

$$\frac{\partial P}{\partial y} - \frac{\partial Q}{\partial x} = 0, \tag{1}$$

$$\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} = 0 \tag{2}$$

relating the velocity components P, Q in two-dimensional steady irrotational fluid flow. Equations (1) and (2) come from the requirements that $Q\,dx + P\,dy$ and $P\,dx - Q\,dy$ be complete differentials, in which case another complete differential is

$$Q\,dx + P\,dy + i(P\,dx - Q\,dy) = (Q + iP)\left(dx + \frac{dy}{i}\right) = (Q + iP)d\left(x + \frac{y}{i}\right).$$

D'Alembert concluded that this means $Q + iP$ is a function f of $x + y/i$, so that $Q = \mathrm{Re}(f)$ and $P = \mathrm{Im}(f)$.

To feel the force of this result, one has to forget the modern definition of function, under which $u(x, y) + iv(x, y)$ is a function of $x + iy$ for *any* functions u, v. In the 18th-century context, a "function" $f(x + iy)$ of $x + iy$ was calculable from $x + iy$ by elementary operations; at worst, $f(x + iy)$ was a power series in $x + iy$. This imposes a strong constraint on u, v, namely that

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}.$$

These were just the equations d'Alembert found in his hydrodynamical investigations, but they came to be named the Cauchy–Riemann equations, because the latter mathematicians stressed their key role in the study of complex functions. The concept of complex function was solidified when Cauchy (1837) showed that a function $f(z)$, where $z = x + iy$, merely had to be differentiable in order to be expressible as a power series in z. Thus it suffices to define a complex function $f(z)$ to be one that is differentiable with respect to z in order to guarantee that f is defined with 18th-century strictness. It follows, in particular, that the first derivative of f entails derivatives of all orders and that the values of f in any neighborhood determine its values everywhere. This "rigidity" in the notion of complex function is enough of a constraint to enable nontrivial properties to be proved, but at the same time it leaves enough flexibility—one might say "fluidity"—to cover important general situations.

EXERCISES

Euler's derivation of $e^{ix} = \cos x + i\sin x$ is easy to explain using the power series

$$e^y = 1 + \frac{y}{1!} + \frac{y^2}{2!} + \frac{y^3}{3!} + \cdots$$

and

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

found in Section 9.5.

16.1.1 Assuming that the series for e^y is also valid for $y = ix$, show that

$$e^{ix} = \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots\right) + i\left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots\right).$$

16.1.2 Assuming it is valid to differentiate the sine series term by term, show that

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots,$$

and hence that $e^{ix} = \cos x + i\sin x$.

Another consequence of $e^{ix} = \cos x + i\sin x$ is that $i = \cos\frac{\pi}{2} + i\sin\frac{\pi}{2} = e^{i\pi/2}$, which allows us to evaluate the outlandish number i^i.

16.1.3 Show that i^i has a real value (Euler (1746)). What is it?

16.1.4 Using the fact that $e^{2in\pi} = 1$ for any integer n, give a formula for all values of i^i (Euler (1746)).

5- /2 / 3

7
!) * +
%
")4 8)) 4 9 +) -
9) % G)) H % 8)) %)
)) 4 + 9)) 8)) -
)))8 9) 4) 4 4) %) %)
9))) 4 - +8 + % 8
9) -) 9J
4 %) 4)) + -) 8 9
)) 9 P) + # %) 8 *))
%)) ' 8) 9+ 9 -) 8) 9
P - 4)
4 4 9 + + -) - +
,) +)) P) 9) + 4
2 + - 38) G H 2 % - 38
) - 8)) 8)
=) 8) ; - +
) %) 9)) - + 4) I .)
) %) 8 C @ A);
);)) -) P 8) 4)));
,)) 4 % %) 8) 4 8)
) P)) 4)) ,) F %)
)+) 9 P) - \ 9)) -) P 4)
)); # -)) -) % %) 8
9)))) F) @) - +8)
)4 T

5- 6T2 N/3 7 N # *) & #

17.5.2 Show that the circles on the cone, in planes perpendicular to its axis, are *not* geodesics.

17.5.3 Show that there are nonsmooth geodesics on the cone, that is, curves of geodesic curvature zero except at certain points where they have no tangent.

17.6 The Gauss–Bonnet Theorem

In Section 17.2 we observed that

$$\int_C \kappa\, ds = 2\pi$$

for a simple closed curve C in the plane. This result has a profound generalization to curved surfaces known as the *Gauss–Bonnet* theorem. On a curved surface, κ must be replaced by the geodesic curvature κ_g, and the theorem states that

$$\int_C \kappa_g\, ds = 2\pi - \iint_{\mathcal{R}} \kappa_1 \kappa_2\, dA,$$

where A denotes area and $\mathcal{R}$ is the region enclosed by C (Bonnet (1848)). Gauss himself published only a special case, or rather the limit of a special case, in which C is a geodesic triangle. In this case, of course, $\kappa_g = 0$ along the sides of C, and κ_g becomes infinite at the corners. By rounding off the corners by small arcs ds, one sees (Figure 17.10) that

$$\int_{C^*} \kappa_g\, ds \cong \alpha' + \beta' + \gamma',$$

where α', β', γ' are the external angles of the triangle and C^* is the rounded approximation to the triangle C.

Then by letting the size of the round-offs tend to zero one gets

$$\int_{C^*} \kappa_g\, ds = \alpha' + \beta' + \gamma' = 3\pi - (\alpha + \beta + \gamma),$$

where α, β, γ are the internal angles of the triangle. Introducing the quantity

$$(\alpha + \beta + \gamma) - \pi,$$

called the *angular excess* of the triangle (because an ordinary triangle has angle sum π), we have

$$\int_C \kappa_g\, ds = 2\pi - \text{angular excess},$$

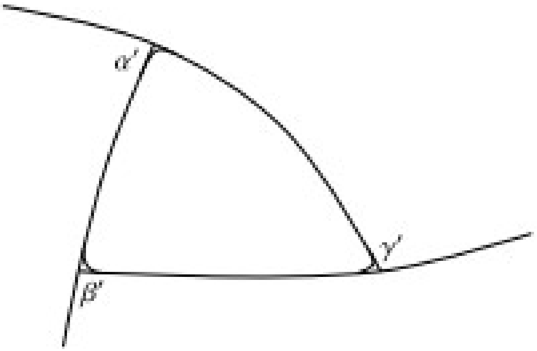

Figure 17.10: Rounding off a geodesic triangle

and the result of Gauss (1827) was that

$$\text{angular excess} = \iint_{\mathcal{R}} \kappa_1\kappa_2 \, dA.$$

We see that the *integral* of the Gaussian curvature has a more elementary geometric meaning than the curvature $\kappa_1\kappa_2$. It appears, in fact, that Gauss thought about angular excess first, then the curvature integral, and only last about the curvature itself. The decomposition into principal curvatures probably came later, when he reworked his geometric ideas into analytic form, reversing the order of discovery in the process. Dombrowski (1979) made a plausible reconstruction of the original approach, using clues from the unpublished work of Gauss.

The role of angular excess can be seen more plainly in the case of constant curvature $\kappa_1\kappa_2 = c$. In this case

$$\text{angular excess} = c \times \text{area of } \mathcal{R},$$

so the angular excess gives a measure of area, a result Gauss claimed, in a letter (1846a), to have known in 1794. In fact, the special case of this result for the sphere was known to Thomas Harriot in 1603 (see Lohne (1979)). Harriot's elegant proof goes as follows (Figure 17.11).

Prolonging the sides of triangle ABC partitions the sphere into four pairs of congruent, diametrically opposite triangles (Figure 17.11a). We denote the area of ABC and its diametric opposite $A'B'C'$ by $\Delta_{\alpha\beta\gamma}$. The

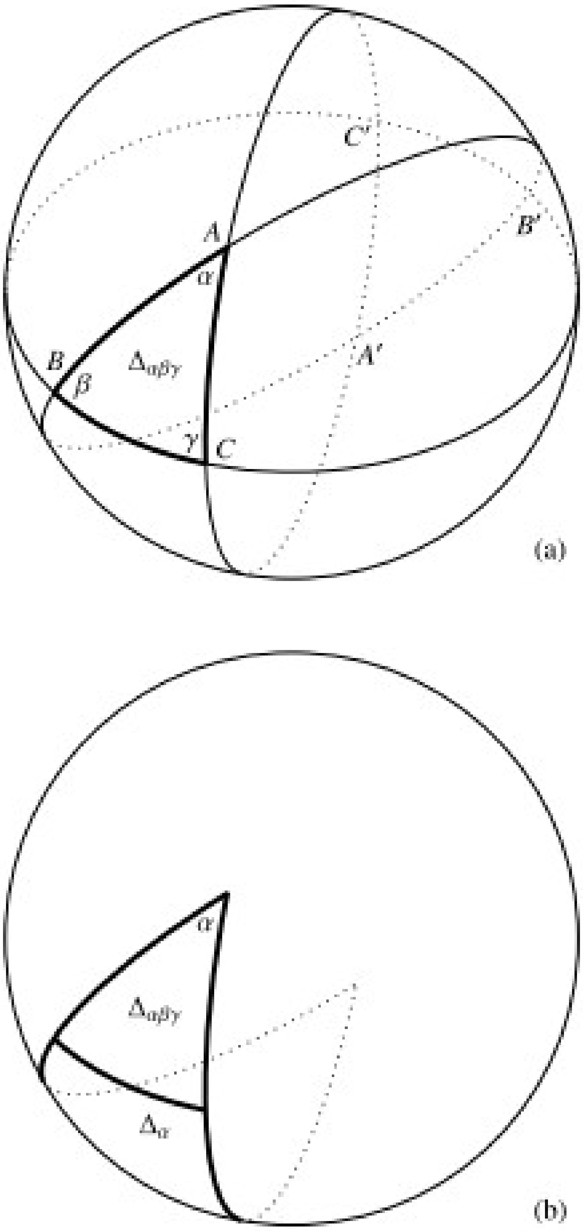
C'
A
B'
α
B
$\Delta_{\alpha\beta\gamma}$
β
A'
γ
C
(a)
α
$\Delta_{\alpha\beta\gamma}$
Δ_{α}
(b)

Figure 17.11: Area of a spherical triangle

other three pairs represent areas $\Delta_\alpha, \Delta_\beta, \Delta_\gamma$, which complement $\Delta_{\alpha\beta\gamma}$ in "slices" of the sphere of angles α, β, γ, respectively (Figure 17.11b).

Since the area of a slice is $2r^2$ times the angle, where r is the radius of the sphere, we have

$$\begin{aligned}\Delta_{\alpha\beta\gamma} + \Delta_\alpha &= 2r^2\alpha,\\ \Delta_{\alpha\beta\gamma} + \Delta_\beta &= 2r^2\beta,\\ \Delta_{\alpha\beta\gamma} + \Delta_\gamma &= 2r^2\gamma,\end{aligned}$$

whence, by addition,

$$3\Delta_{\alpha\beta\gamma} + (\Delta_\alpha + \Delta_\beta + \Delta_\gamma) = 2r^2(\alpha + \beta + \gamma). \tag{1}$$

On the other hand,

$$2(\Delta_{\alpha\beta\gamma} + \Delta_\alpha + \Delta_\beta + \Delta_\gamma) = \text{area of sphere} = 4\pi r^2,$$

and substituting this in (1) gives

$$\Delta_{\alpha\beta\gamma} = r^2(\alpha + \beta + \gamma - \pi)$$

as required, since $1/r^2$ = curvature of the sphere.

Gauss was interested in the counterpart of this result for negative curvature, in which case the angle sum of a triangle is less than π and one has angular *defect* rather than angular excess. His investigations in this case led him not only to Gaussian curvature but also to non-Euclidean geometry.

EXERCISES

It is surprising at first that area on the sphere should be measured by angles rather than lengths. However, there are general reasons (apart from the Gauss–Bonnet theorem) why area should be measured by angular excess, and this idea fails only where angular excess is zero—that is, in the Euclidean plane.

17.7.1 Consider a triangle Δ split into two triangles Δ_1 and Δ_2 by a line through a vertex. Show that

$$\text{excess}(\Delta) = \text{excess}(\Delta_1) + \text{excess}(\Delta_2).$$

17.7.2 Deduce from Exercise 17.7.1 that if any polygon Π is split into triangles Δ_i, then

$$\text{excess}(\Pi) = \text{excess}(\Delta_1) + \text{excess}(\Delta_2) + \cdots .$$

Thus the angular excess function has the same *additive* property as an area function. It can be shown that any additive function, provided it is continuous, is a constant multiple of area (see Bonola (1912), p. 46).

5- /<2 7/3

T
$ @) * +
%
4 - +8 - + %) - - +
4)) +)) @));
4) - +8) - +) 4)));
); @
E) - 9 4) 4)));
8 9 4) % 9 E 4)) 8))
9 %) + +4 8 8))
4))) - % -) - % 4 E T 8 & +))
O 9) % '+ 4 4) = 9) 4))
' - +U @) - +
4 %)) 8 % 8
P))); @ , ')))
) 8) - 9J) G4 H) G 8H)) P
@ A); 4) 4)));
)) P 9+ &) 2 TNT)38
))) -) % %) -) G H &+
%))44 -) 8 &) 8 -
) 4 J % G H) - '8)
) G) - H)) +) -
.) +8) q 2 TT 3) &) A))
)) + 4 ;)) +)4))) + 9 4 9 4
4) @) G 8H)9 +) F 2 T7 3
8 @) - +))) +) 4) ; -)) 8
9) 4) -)) 4 % + 9

5- N 2 7<3

18.2 Spherical Geometry

In rejecting P_0 because of its incompatibility with infinite lines, Saccheri avoided having to consider the most natural geometry in which P_0 holds, that of the sphere with great circles as "lines." Spherical geometry had been cultivated since ancient times to meet the needs of astronomers and navigators, and formulas for the side lengths and areas of spherical triangles were well known. But the sphere was considered part of Euclid's spatial geometry, so the axiomatic significance of spherical geometry was at first ignored. What did happen, though, was that the first explorations of Axiom P_2 were guided by the analogy of the sphere.

Lambert (1766) made the striking discovery that Axiom P_2 implies that the area of a triangle with angles α, β, γ is proportional to $\pi - (\alpha + \beta + \gamma)$, its angular defect. In other words,

$$\text{area} = -R^2(\alpha + \beta + \gamma - \pi)$$

for some positive constant R^2. Having rediscovered Harriot's theorem that

$$\text{area} = R^2(\alpha + \beta + \gamma - \pi)$$

for a triangle on the sphere of radius R, Lambert mused that one "could almost conclude that the new geometry would be true on a sphere of imaginary radius." What a sphere of radius iR might be was never explained, but the idea of using complex numbers to generate the formulas of a hypothetical geometry proved fruitful.

It was found that formulas derived from Axiom P_2 could also be obtained by replacing R by iR in corresponding formulas of spherical geometry. For example, Gauss (1831) deduced from Axiom P_2 that the circumference of a circle of radius r is $2\pi R \sinh r/R$. The same result follows by replacing R by iR in $2\pi R \sin r/R$, which is the circumference of a circle of radius r on the sphere (where, of course, r is measured *on* the spherical surface; see Exercise 18.2.1).

Klein (1871) called the geometry of Axiom P_2 *hyperbolic*. One reason for this is that its formulas involve hyperbolic functions, whereas those of spherical geometry involve circular functions. Lambert (1766) introduced the hyperbolic functions and noted their analogy with the circular functions, but he did not follow through with a complete translation of spherical formulas into hyperbolic formulas. This was first done by Taurinus

(1826), one of a small circle who corresponded with Gauss on geometric questions.

This gave hyperbolic geometry a second leg to stand on, but there was still nothing solid under its feet. Neither Gauss nor Taurinus seemed confident of finding a convincing interpretation of hyperbolic geometry, even though Gauss (1827) came remarkably close with the "Gauss–Bonnet" theorem. As mentioned in Section 17.6, this theorem shows that surfaces of constant negative curvature give a geometry in which angular defect is proportional to area, and Gauss knew that the pseudosphere is such a surface. Gauss's student Minding (1840) even showed that the hyperbolic formulas for triangles hold on the pseudosphere, but no one at that time commented on the likely importance of this result for hyperbolic geometry. Perhaps it was clear that the pseudosphere cannot serve as a "plane," because it is infinite in only one direction. Only in 1868, when Beltrami extended the pseudosphere to a true *hyperbolic plane*—a surface locally like the pseudosphere but infinite in all directions—was hyperbolic geometry finally placed on a firm foundation.

EXERCISES

18.2.1 Prove that the circumference of the circle C of radius r on the sphere of radius R (Figure 18.4) is $2\pi R\sin(r/R)$.

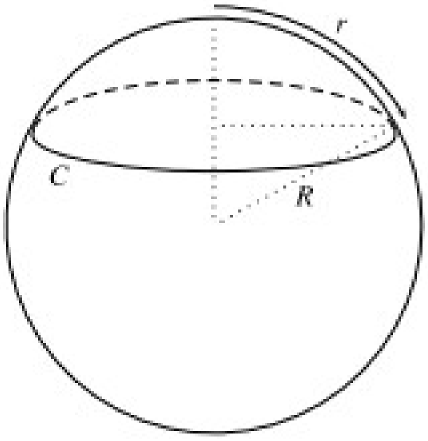

Figure 18.4: Radius and circumference on the sphere

18.2.2 Show that both $2\pi R\sin(r/R)$ and $2\pi R\sinh(r/R)$ tend to $2\pi r$ as $R \to \infty$.

These results show how even non-Euclidean geometry is "Euclidean in the small"—its formulas tend to the Euclidean formulas as size tends to zero. The same is true of the angle-sum formula for a triangle.

18.2.3 Deduce from Harriot's area formula that the angle sum of a spherical triangle tends to π as its size tends to zero.

18.3 Geometry of Bolyai and Lobachevsky

The most important contributors to hyperbolic geometry between Gauss and Beltrami were Lobachevsky and Bolyai, who published independent discoveries of the subject: Lobachevsky (1829) and János Bolyai (1832b). Because of their courage in advocating an unconventional geometry, Bolyai and Lobachevsky have been justly admired. Nevertheless, the immediate impact of their work was slight. Many of their results were already known to Gauss and his circle and could have been picked up from existing publications and personal contacts. Lambert (1766) and Taurinus (1826) were in print, and Bolyai's father, F. Bolyai, was a lifelong friend of Gauss, as was Lobachevsky's teacher Bartels. In any case their work, though more systematic and convincing than previous attempts, attracted very little attention at first. We have seen how the possibility of using differential geometry to justify hyperbolic geometry was overlooked until 1868. Up to that time, there seemed no reason to take hyperbolic geometry seriously.

In retrospect, of course, the theorems of Bolyai and Lobachevsky can be seen to unify the fragmentary results of their predecessors very nicely. They cover the basic relations between sides and angles of triangles (hyperbolic trigonometry), the measure of polygonal areas by angular defect, and formulas for circumference and area of circles. Lobachevsky (1836) broke new ground by finding volumes of polyhedra, which turn out to be far from elementary, involving the function $\int_0^\theta \log 2|\sin t|\,dt$.

Both Bolyai and Lobachevsky considered a three-dimensional space satisfying Axiom P_2 and made extensive use of a surface peculiar to this space, the *horosphere*. A horosphere is a "sphere with center at infinity," and it is *not* a hyperbolic plane. Wachter, a student of Gauss, observed in a letter of 1816 (published in Stäckel (1901)) that the geometry of the horosphere is in fact Euclidean. This astonishing result was rediscovered by Bolyai and Lobachevsky, and they anticipated that it would make Euclidean geometry subordinate to hyperbolic. We shall see in Section 18.5 how this view was vindicated in the work of Beltrami.

5- 66 26//3 *)) E -

5-\ 6N726T 3

4 -+
%
E)4 /) ( ) 4 -) 4 -
9 4) +) - 9) % E 4)4
4 -+ 9)))J P)) 8
) 4 9
$)) +8 4 -+) - +8) 4
-)) 9 P - + " 4) ;) 4
@)) 8) -) + 9 %))))
4 +)8) 9) -) 9) +)
)+8 ') 4 -+ P 8)))
))44 - + . ;) 4 8 *) &
) @)) %)) %)
))
4 -+))) - 9) E)4
)) - 48) - 4) 9)+ C ; 9 4
)) - 9J ,) 4 8) 4) 9 ')
4 8)) - 4 %) , 8 % 8)
) + 4 & +)) 9) 4)) 4 8
)) 98))9 p 9)
E < 68) q) + J)))
4) %))) - 4 4 -) +))
) 4 #) q J) 4 % + 8
4) - + #)
9 - +) 4 -+

5- 6772 6< 3

/ @))) %)
) 8 G 8H 4 !) A) %)
))) ;) -) + r o @))
E) 8 ' -) ;) - +) 4 @
)) 4 +) #)% 9)+
!) % %)

5- 67<26< 3

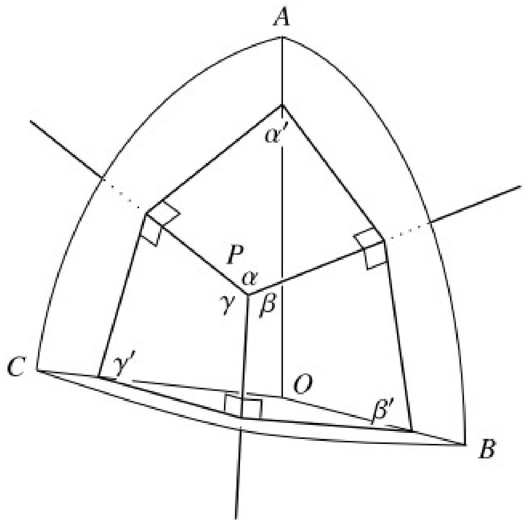
A
α′
P
α
γ
β
C
γ′
O
β′
B

. - T # % ; -) 4 +

5- 67726< 3

of each other. Gauss–Bonnet ⇒ Descartes by concentrating the curvature of a surface at vertices until it becomes a polyhedron, while Descartes ⇒ Gauss–Bonnet by increasing the number of vertices of a polyhedron until it becomes a smooth surface. It is interesting, though probably accidental, that Descartes actually uses the word "curvatura" to describe the exterior solid angle.

22.5 Euler Characteristic and Curvature

There is another, more "intrinsic," proof of Descartes's theorem that reveals the fact that total exterior solid angle is really $2\pi \times$ Euler characteristic. In fact, knowledge of the total exterior angle yields a proof of the Euler characteristic polyhedron formula. This seems to have been the way in which Descartes discovered his version of the formula.

The key step is to show that the exterior solid angle at a vertex P is expressible intrinsically as $2\pi - (\alpha_1 + \alpha_2 + \cdots + \alpha_n)$, where $\alpha_1, \alpha_2, \ldots, \alpha_n$ are the face angles that meet at P. These are *not* the angles $\alpha'_1, \alpha'_2, \ldots, \alpha'_n$ between the planes that bound the exterior solid angle, but it turns out (Exercise 22.5.1) that

$$\alpha_i + \alpha'_i = \pi$$

for each i, whence the measure of the exterior solid angle, which comes from $\alpha'_i + \alpha'_2 + \cdots + \alpha'_n$ by Harriot's theorem (Section 17.6), also comes from $\alpha_1 + \alpha_2 + \cdots + \alpha_n$.

Knowing now that the exterior solid angle at P equals $2\pi - \sum$ face angles at P, we get

$$\text{total exterior solid angle} = 2\pi V - \sum \text{all face angles},$$

where V is the total number of vertices. By grouping the face angles according to the types of faces, we also find (Exercise 22.5.2) that

$$\sum \text{all face angles} = \pi(2E - 2F),$$

whence

$$\begin{aligned}\text{total exterior solid angle} &= 2\pi(V - E + F)\\ &= 2\pi \times \text{Euler characteristic}.\end{aligned}$$

In the case of convex polyhedra, where we already know that total exterior solid angle $= 4\pi$, this gives Euler characteristic $= 2$. More importantly, the derivation is valid for polyhedra of arbitrary Euler characteristic, showing that the total exterior solid angle is really the *same* as the Euler characteristic, up to a constant multiple.

There is a similar intrinsic proof of the Gauss–Bonnet theorem, again valid for arbitrary Euler characteristic, which shows that

$$\text{total curvature} = \iint_S \kappa_1 \kappa_2 \, dA = 2\pi \times \text{Euler characteristic}$$

(Exercise 22.5.3). Legendre's (1794) proof of the Euler polyhedron formula is the special case of the argument for constant curvature.

Thus the Euler characteristic regulates the total curvature of a surface. In particular, if the curvature is constant, it must have the same sign as the Euler characteristic. This in turn has implications for the geometry of the surface. As we saw in Section 17.4, surfaces of constant positive curvature have spherical geometry, those of zero curvature have Euclidean geometry, and those of negative curvature have hyperbolic geometry. In the next section we shall see that there is a natural way to impose constant curvature on surfaces of arbitrary Euler characteristic. It will then follow that the natural geometry of a surface is spherical, Euclidean, or hyperbolic according as its Euler characteristic is positive, zero, or negative. Moreover, if the absolute value of the curvature is taken to be 1, then the Gauss–Bonnet theorem gives

$$\text{area} = |2\pi \times \text{Euler characteristic}|.$$

This makes surface topology completely subordinate to geometry, at least for orientable surfaces, because it says that the topology of a surface is completely determined by the sign of its curvature and its area.

These results were implicit in the work of Poincaré and Klein in the 1880s. Perhaps Klein was the first to see clearly how the geometry of a surface determines its topology (see, for example, Klein (1928), p. 264).

Exercises

Figure 22.8 shows the region around a vertex P of a polyhedron and the exterior solid angle of P centered at O and bounded by the planes OAB, OBC, OCA perpendicular to the edges through P.

22.5.1 Show that there are right angles where indicated, and hence that

$$\alpha + \alpha' = \pi, \quad \beta + \beta' = \pi, \quad \gamma + \gamma' = \pi.$$

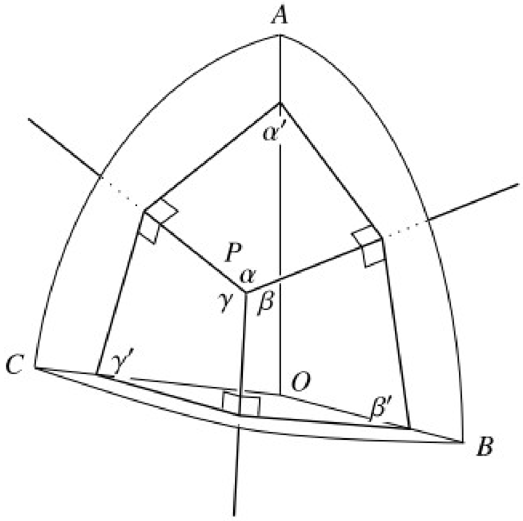

Figure 22.8: The vertex region of a polyhedron

Now to relate face angles to E and F, it helps to write

$$F = F_3 + F_4 + F_5 + \cdots,$$

where F_3 = number of 3-gon faces, F_4 = number of 4-gon faces, and so on.

22.5.2 Show that

$$E = \frac{1}{2}(3F_3 + 4F_4 + 5F_5 + \cdots),$$

and deduce that in an ordinary polyhedron (that is, one with flat faces),

$$\sum \text{all face angles} = \pi(2E - 2F),$$

using the fact the angle sum of an n-gon is $(n-2)\pi$.

22.5.3 Prove the global form of the Gauss–Bonnet theorem,

$$\iint_S \kappa_1\kappa_2 \, dA = 2\pi \times \text{Euler characteristic},$$

by partitioning the closed surface $\mathcal{S}$ into geodesic polygons and applying the ordinary form of the Gauss–Bonnet theorem (Section 17.6).

22.6 Surfaces and Planes

In Section 16.5 we noticed that an elliptic function defines a mapping of a plane onto a torus. Such mappings are also interesting in the topological

,9

( ) 6

!'#' $ % % & ' " 1 +
)) 8 S (%)) Y % + 2 T3 4

,9

*) N
( )

% 6

S) %

!'#' $ % % & ') ")
$)) 7M<7T <T N 6 T< 4

!'#' $ % % & ' L)K & + EEE8 @
& L)K & + EEE8 @ & 4

5- 2 3 6 ,) , 9)) +) . 4 - @ -+ O %

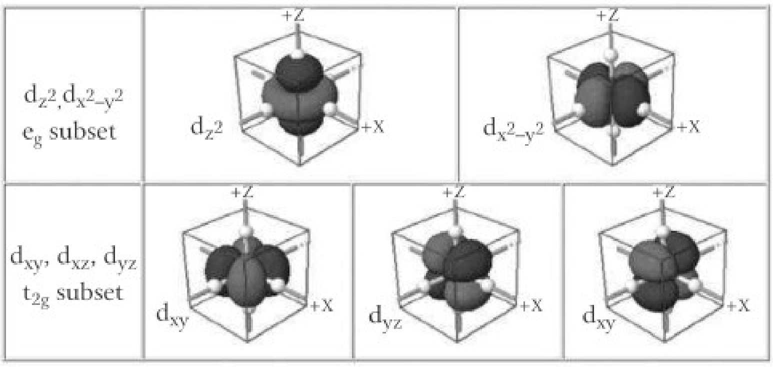
$d_{z^2}, d_{x^2-y^2}$
e_g subset
d_{z^2}
$d_{x^2-y^2}$
+Z
+X
d_{xy}, d_{xz}, d_{yz}
t_{2g} subset
d_{xy}
d_{yz}
d_{xy}

!'#' $ % % & ' " 1 +)) #
.) * + 4 -)- $ 0 ' 2 /3 4

5- /2 /3

)-) =) >
)-) =) =) 8)
)) -)) =) 9+))-) 2. - 73

We got this result by making the angles α in the two triangles coincide. If we make the angles β coincide instead, we similarly find that the sides opposite to α and γ are proportional. Thus, in fact, *all corresponding sides of similar triangles are proportional*. □

This consequence of the Thales theorem has many implications. In everyday life, it underlies the existence of scale maps, house plans, engineering drawings, and so on. In pure geometry, its implications are even more varied. Here is just one, which shows why square roots and irrational numbers turn up in geometry.

The diagonal of the unit square is $\sqrt{2}$

The diagonals of the unit square cut it into four quarters, each of which is a triangle similar to the half square cut off by a diagonal (Figure 1.17).

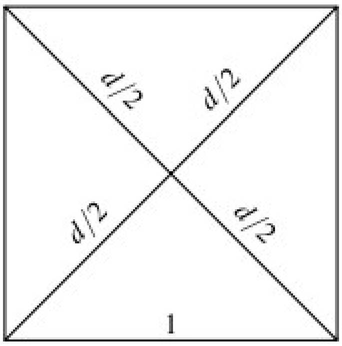

Figure 1.17: Quarters and halves of the square

Each of the triangles in question has one right angle and two half right angles, so it follows from the theorem above that corresponding sides of any two of these triangles are proportional. In particular, if we take the half square, with short side 1 and long side d, and compare it with the quarter square, with short side $d/2$ and long side 1, we get

$$\frac{\text{short}}{\text{long}} = \frac{1}{d} = \frac{d/2}{1}.$$

Multiplying both sides of the equation by $2d$ gives $2 = d^2$, so $d = \sqrt{2}$. □

The great, but disturbing, discovery of the Pythagoreans is that $\sqrt{2}$ *is irrational*. That is, there are no natural numbers m and n such $\sqrt{2} = m/n$.

If there are such m and n we can assume that they have no common divisor, and then the assumption $\sqrt{2} = m/n$ implies

	$2 = m^2/n^2$	squaring both sides
hence	$m^2 = 2n^2$	multiplying both sides by n^2
hence	m^2 is even	
hence	m is even	since the square of an odd number is odd
hence	$m = 2l$	for some natural number l
hence	$m^2 = 4l^2 = 2n^2$	
hence	$n^2 = 2l^2$	
hence	n^2 is even	
hence	n is even	since the square of an odd number is odd.

Thus, m and n have the common divisor 2, contrary to assumption. Our original assumption is therefore false, so there are no natural numbers m and n such that $\sqrt{2} = m/n$. □

Lengths, products, and area

Geometry obviously has to include the diagonal of the unit square, hence *geometry includes the study of irrational lengths.* This discovery troubled the ancient Greeks, because they did not believe that irrational lengths could be treated like numbers. In particular, the idea of interpreting the product of line segments as another line segment is *not* in Euclid. It first appears in Descartes' *Géométrie* of 1637, where algebra is used systematically in geometry for the first time.

The Greeks viewed the product of line segments a and b as the *rectangle* with perpendicular sides a and b. If lengths are not necessarily numbers, then the product of two lengths is best interpreted as an area, and the product of three lengths as a volume—but then the product of four lengths seems to have no meaning at all. This difficulty perhaps explains why algebra appeared comparatively late in the development of geometry. On the other hand, interpreting the product of lengths as an area gives some remarkable insights, as we will see in Chapter 2. So it is also possible that algebra had to wait until the Greek concept of product had exhausted its usefulness.

Exercises

In general, two geometric figures are called similar if one is a magnification of the other. Thus, two rectangles are similar if the ratio $\frac{\text{long side}}{\text{short side}}$ is the same for both.

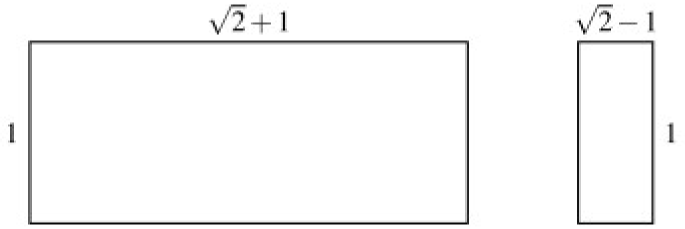

Figure 1.18: A pair of similar rectangles

1.5.1 Show that $\frac{\sqrt{2}+1}{1} = \frac{1}{\sqrt{2}-1}$ and hence that the two rectangles in Figure 1.18 are similar.

1.5.2 Deduce that if a rectangle with long side a and short side b has the same shape as the two above, then so has the rectangle with long side b and short side $a-2b$.

This simple observation gives another proof that $\sqrt{2}$ is irrational:

1.5.3 Suppose that $\sqrt{2}+1 = m/n$, where m and n are natural numbers with m as small as possible. Deduce from Exercise 1.5.2 that we also have $\sqrt{2}+1 = n/(m-2n)$. This is a contradiction. Why?

1.5.4 It follows from Exercise 1.5.3 that $\sqrt{2}+1$ is irrational. Why does this imply that $\sqrt{2}$ is irrational?

1.6 Discussion

Euclid's *Elements* is the most influential book in the history of mathematics, and anyone interested in geometry should own a copy. It is not easy reading, but you will find yourself returning to it year after year and noticing something new. The standard edition in English is Heath's translation, which is now available as a Dover reprint of the 1925 Cambridge University Press edition. This reprint is carried by many bookstores; I have even seen it for sale at Los Angeles airport! Its main drawback is its size—three bulky volumes—due to the fact that more than half the content consists of

Heath's commentary. You can find the Heath translation *without* the commentary in the Britannica *Great Books of the Western World*, Volume 11. These books can often be found in used bookstores. Another, more recent, one-volume edition of the Heath translation is *Euclid's Elements*, edited by Dana Densmore and published by Green Lion Press in 2003.

A second (slight) drawback of the Heath edition is that it is about 80 years old and beginning to sound a little antiquated. Heath's English is sometimes quaint, and his commentary does not draw on modern research in geometry. He does not even mention some important advances that were known to experts in 1925. For this reason, a modern version of the *Elements* is desirable. A perfect version for the 21st century does not yet exist, but there is a nice concise web version by David Joyce at

`http://aleph0.clarkeu.edu/~djoyce/java/elements/elements.html`

This *Elements* has a small amount of commentary, but I mainly recommend it for proofs in simple modern English and nice diagrams. The diagrams are "variable" by dragging points on the screen, so each diagram represents all possible situations covered by a theorem.

For modern commentary on Euclid, I recommend two books: *Euclid: the Creation of Mathematics* by Benno Artmann and *Geometry: Euclid and Beyond* by Robin Hartshorne, published by Springer-Verlag in 1999 and 2000, respectively. Both books take Euclid as their starting point. Artmann mainly fills in the Greek background, although he also takes care to make it understandable to modern readers. Hartshorne is more concerned with what came after Euclid, and he gives a very thorough analysis of the gaps in Euclid and the ways they were filled by modern mathematicians. You will find Hartshorne useful supplementary reading for Chapters 2 and 3, where we examine the logical structure of the *Elements* and some of its gaps.

The climax of the *Elements* is the theory of regular polyhedra in Book XIII. Only five regular polyhedra exist, and they are shown in Figure 1.19. Notice that three of them are built from equilateral triangles, one from squares, and one from regular pentagons. This remarkable phenomenon underlines the importance of equilateral triangles and squares, and draws attention to the regular pentagon. In Chapter 2, we show how to construct it. Some geometers believe that the material in the *Elements* was chosen very much with the theory of regular polyhedra in mind. For example, Euclid wants to construct the equilateral triangle, square, and pentagon in order to construct the regular polyhedra.

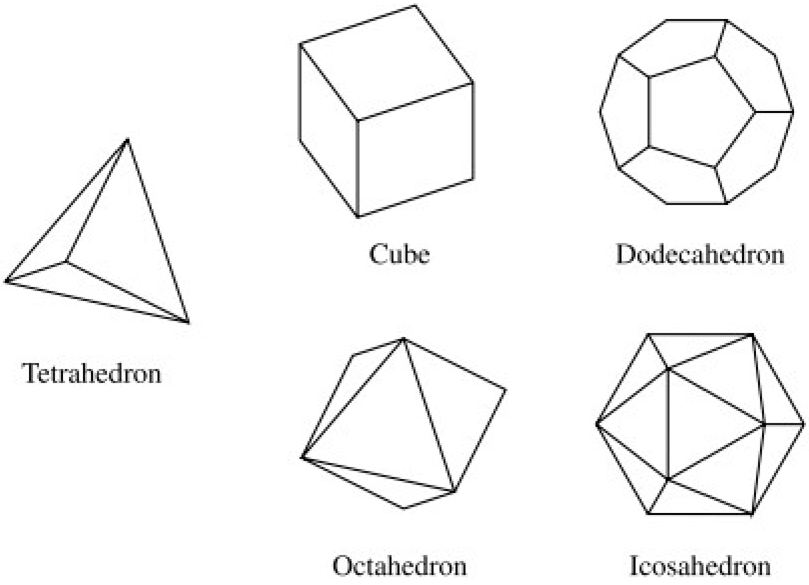

Figure 1.19: The regular polyhedra

It is fortunate that Euclid did not need regular polygons more complex than the pentagon, because none were constructed until modern times. The regular 17-gon was constructed by the 19-year-old Carl Friedrich Gauss in 1796, and his discovery was the key to the "question arising" from the construction of the equilateral triangle in Section 1.2: for which n is the regular n-gon constructible? Gauss showed (with some steps filled in by Pierre Wantzel in 1837) that a regular polygon with a prime number p of sides is constructible just in case p is of the form $2^{2^m}+1$. This result gives three constructible p-gons not known to the Greeks, because

$$2^4+1=17, \quad 2^8+1=257, \quad 2^{16}+1=65537$$

are all prime numbers. But no larger prime numbers of the form $2^{2^m}+1$ are known! Thus we do not know whether a larger constructible p-gon exists.

These results show that the *Elements* is not all of geometry, even if one accepts the same subject matter as Euclid. To see where Euclid fits in the general panorama of geometry, I recommend the books *Geometry and the Imagination* by D. Hilbert and S. Cohn-Vossen, and *Introduction to Geometry* by H. S. M. Coxeter (Wiley, 1969).

,9

*) 7
( ) 7

5- 2 3

T < !
@) 4)) +4
E) 9)) - @) - + @
A '8 9) @ - F ; 4)))
4)));) +))%)9 O) -
%) 4) 9 = %) 4))); 8)
s =) %)) 8
s) -)) -
r8
s) P- F ; 8
9 - 4 @ A);) ,
) +8) 4 %)) ;) +
4)) U)) @) 4)) +4 U
)))44) + - + # - +)%
9 ;4 -) 4) 8) +4 9 - +
1+4 9 - + % - @) - + 44
4) - +U ;) 4 8) -)
) - tr8 urU9 % - ; # G H
4) - + %)) @ A);)9 8)
G H +4 9 - + %) + 4)));
# P +4 9 - + % 9+ E)
) S *)))) 4 4 % 4)));
E 7 9 '8
@)9) % %)
2@)
% + C) 38)) @) 4)) +4 8)
-)) D)) + 4)
9)44))) 9) + + # % +
) 8)) P +8)))
) + 4))%) 4 4) . -
G 4 -)))) - 8H)) % +
@
& 4 4)) P +)) 8)
2) ' 3) +) 4) #)
@)))44))) 9) + +
@)) 8)) ; p 8)
+)%) 4 4)) P +U ;);))
- + %) 9 -8 9) ' +) +4 9 - +

@) - + +4 9 4) 9 -))'
)4) + < + ")))))
) . *) 2 777 T//3 ;4 =
@) 4)) +4 8) - *) 4 9 9
J -) *)) -))))
8 9) - 4 9 G 4 H '8) % +

) @) - +) ')) 1 '
4) 8 8 @) - +8)) 8
) % E) -) +)
% U - + %)) - +
4 ; 9 U) 4 %
P 4 9 4 %) @) -
+ S) & +) 1 -) +) $ ') O 9) % '+ ( )
" T + % - + 4 +) 9)
P G 9 % H #) %
) +) %) - 8)))
)) - +) - 44 E) 8 +
- 8 9) 8 +)) -
+))% 9) %))9) + *) A
! = -
)) 4 P %)
2*) % -) %) 3
) 4 9 T 78 9 & +) 8 O 9) % '+8 *)
) - % @) - +8))) -
)) 4 *) A) + %) 8
) =) +) 4 % 2)) 3) 4 8 F 4)
) + 8) -) %))) G))4 H
- 9) 4 E
! =
8 *) % -)
) 4 9 %)))) 9)%
- 8) % -) G H
1 8 ;) 4 8)) -) -)
s) - ur)) 4 % %) 8
s) -
r)) F %) 8
s) - tr)) -) % %)
% 8 %))) F 8 8) + -
) - 8 2) - vr3 4 4))) #)%
*) -)) + ' @) - +8
4 -)))4) F

!'#' $ % % & ' " 1 +)) S
$ 9) * + 4 - 2 <<73 4

!'#' $ % % & ' " 1 +
)) $ 9) * + 4 -)- $ 0 ' 2 <<T3 4

,9

*) 7
( )

5- <2 3 . ;) 4 8 4 %)) + =) 8 9 %) =)

5- 72 N3 7 6 # *)) E -
E 4 ; 9 8 4) -))
*)) -

5- T2 673 7 T !
4 ; $ 9) * +
- + 4 ; 9) %) 9J E %
+ @) 4) 9) 4 8 @)
4)) % @) 4) ,) 9J 8
4 9 9 9+ 4)
4 ; %))9 D)%)44 @)
4) 4) 8))44) + 4
) @) 4) +))9 +)44
@) 4)

5- T 2 TT3 < " - 9)) #)) .

) 9 ') 9 9)
) - +)) % % 4
)) # + =)) ; ;) 4
) 9) - +)
)-) =) 8) ;4) 4
9 4 8) 9 =) 4 % 9+
- % - - 9 8

: : 7 < 6<

!"#' $ % % & ' " E -) @=)
$) ") + 7M<7T T/ / 4

,9
*) N
%

!"#' $ % % & ')% # 4 -+ 1+4 4)
") - " E 4 - E)) 9 - 2 73 4

,9
*)) ( )
% 7

5- 2T3

)) %) 2 4) %) 3) J 4 -
-+ - 4) 4 -) + 2 4) - 9)) +3 P 8)))
) +) 4)4 4 9))) 9
+4 4))) -) 8 -) +8)) - 4
=) 4 J %) - 9) %) D) 9))
%)) -+ P)9) % 8 4)) +
% 4 4 N) - % # N
# 9 - 4 +) 4 + 4 4 P & 9
% 8 # N 8)) 9)) % -

) P %))) - 8 4)) 1 F 9 % -
) 8) ') 1 F 9 L) 8 # N 6 #)
)4) #) - # N))
9+ .) ') 0 F% '+8 # N N # 9 -
;4) 9)) + 4 4 J % 4)
%))) - 9) - +8 # N / ") 9
)) 4)) %) 8)
-)) &
)) - #) P ;) 4)
+4 4))) -)% -))) 4))) %) +8
)) % 9+ Z 7[

5- /<2 NT3 T . % 8 . O ") - 8) #) A J

#J) 9 %

!"#' $ % % & ' 1) 9 ' * +) # 4 -+
-) 1) 9 ' * +) # 4 -+ -) E 4 -
E)) 9 - 4 - 2 3 4

*) 7
( ) 7
% NN

!"#' $ % % & ' 1) 9 ' * +) # 4 -+
-) 1) 9 ' * +) # 4 -+ -) EE 4 -
E)) 9 - 4 - 2 3 4

*) 7
( )
% 7

!"#' $ % % & ' 1) 9 ' * +) # 4 -+
-) #) " - 9)8 9) 8 * +) $ 9
+ 4

,9
*) <
( ) N
% 6/

: : T 6

( )) * +) , -) .) @ h j)))
45-) N N 2 3

E 4)
@ h) 4) h !)) $ % % :E 4) :( ))
* +) , -) .)

) @ h

!"#' $ % % & ' ( )) * +) , -)
.) !) * + %)) 4

,9
*) /
( ) 6
% 6

5- 62 6 3 # * + *))4

5- 72 /3 6 # E * +)

5- 6 T26 N3 / T) w *)) %)

!"#' $ % % & ' ( )) * +) , -)
.) !) * + # +8 8) "44) 2 !.! % 3 4

,9
*) <
%

5- NT2T 3 4)

. - 6 / # P-))) ;) 4))44 q
#
@ %) =) 8) 8 4)))))44 8 J)8))
4 4))) 4))

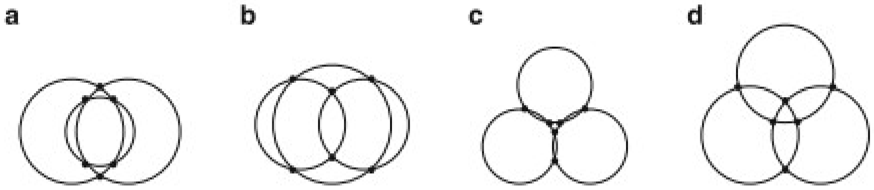

J) 5)

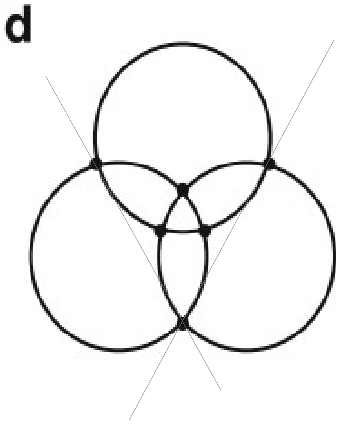

")))) %))) 4
"J) - 5 4)) =)
)) - 8 9)) Q 4) j = 4 J)))))

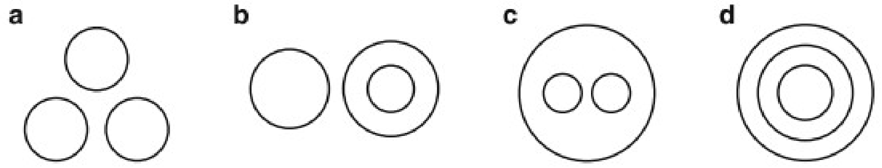

Fig. 4.2 The four configurations in the case there is no contact point

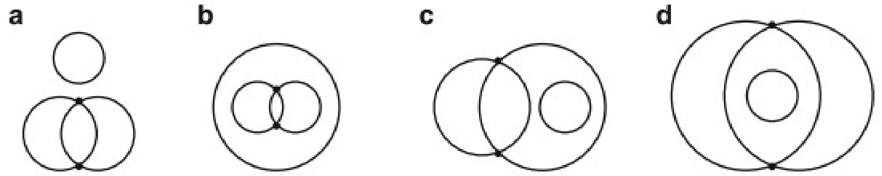

Fig. 4.3 The four configurations in the case there are two contact points

Fig. 4.4 The two configurations in the case there are four contact points

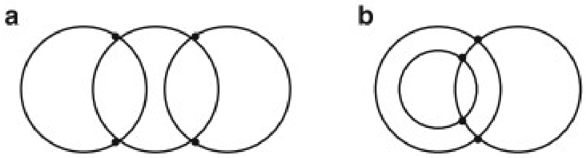

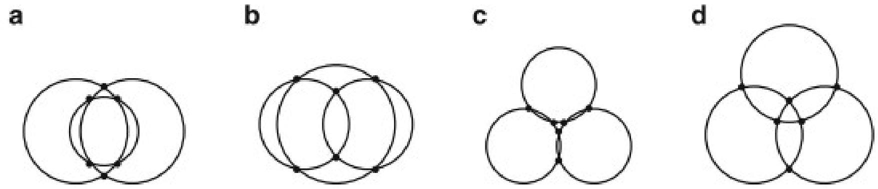

Fig. 4.5 The four configurations in the case there are six contact points

and z_- which both lie at the intersection of the three radical planes. The barycentric coordinates γ_i, $i \in \{1,2,3\}$, of $z_0 = (z_+ + z_-)/2$ are calculated in Eq. (4.18) from Eqs. (4.14) and (4.15), where T_{123} is the surface of the triangle c_1, c_2, c_3 and i, j, k are circular permutations of $1, 2, 3$. We get also the length of the segment $z^+ - z^-$ in Eq. (4.19), and then we deduce Theorem 4.4:

$$\begin{aligned}16T_{123}^2\gamma_i = & -2R_i^2\|c_j-c_k\|^2\\ & + R_j^2(\|c_j-c_k\|^2+\|c_k-c_i\|^2-\|c_i-c_j\|^2)\\ & + R_k^2(\|c_j-c_k\|^2-\|c_k-c_i\|^2+\|c_i-c_j\|^2)\\ & + \|c_j-c_k\|^2(-\|c_j-c_k\|^2+\|c_k-c_i\|^2+\|c_i-c_j\|^2),\end{aligned} \tag{4.18}$$

5- 7 2T/3 6 / E , 6
,9 Y) 5 4))))) Q

$$\left\|\frac{z_+ - z_-}{2}\right\|^2 = \frac{R_1^2 + R_2^2 + R_3^2}{3} - \left\|z_0 - \frac{c_1 + c_2 + c_3}{3}\right\|^2 - \frac{\|c_2 - c_1\|^2 + \|c_3 - c_2\|^2 + \|c_1 - c_3\|^2}{9}. \tag{4.19}$$

Theorem 4.4. *The convex domain defined by the configuration of Fig. 4.5d exists if and only if the quantity at the right member of Eq. (4.19) is positive.*

In the case of Fig. 4.5d, the point $z_0 = (z_+ + z_-)/2$ is interior to the intersection of the three spheres because this intersection is convex. Then, z_0, z_+, z_- and the three contact points on the boundary of the sphere intersection in the plane c_1, c_2, c_3 define a partition of the intersection into six parts, each one being the intersection between one sphere and one trihedron. Each of the six trihedra originates in z_0 and is bounded by two radical planes and by the plane c_1, c_2, c_3. We denote by $x_{ij}^{(k)}$ each of the contact points at the intersection of the great circles of spheres i and j located in the interior of the sphere k, where i, j, k are circular permutations of $1, 2, 3$. Thus the trihedra are defined by the half lines sets:

$$\left\{z_0 \to x_{31}^{(2)}, z_0 \to x_{12}^{(3)}, z_0 \to z_+\right\}$$

$$\left\{z_0 \to x_{12}^{(3)}, z_0 \to x_{23}^{(1)}, z_0 \to z_+\right\}$$

$$\left\{z_0 \to x_{23}^{(1)}, z_0 \to x_{31}^{(2)}, z_0 \to z_+\right\}$$

and their mirror symmetric images through the plane c_1, c_2, c_3:

$$\left\{z_0 \to x_{31}^{(2)}, z_0 \to x_{12}^{(3)}, z_0 \to z_-\right\}$$

$$\left\{z_0 \to x_{12}^{(3)}, z_0 \to x_{23}^{(1)}, z_0 \to z_-\right\}$$

$$\left\{z_0 \to x_{23}^{(1)}, z_0 \to x_{31}^{(2)}, z_0 \to z_-\right\}.$$

The six above trihedra intersect respectively with the sphere 1, 2, 3, 1, 2, 3.

It follows that the calculation of V_{123} and S_{123} can be done through the calculation of the portion of the area of a sphere defined by its intersection with a trihedron having its origin in the interior of the sphere. This latter calculation will be shown to be required in the case of four intersecting spheres and is done analytically in Sect. 4.2.6 using the Gauss–Bonnet theorem.

Several analytical treatments of up to three intersecting spheres appeared in the literature. Some do not involve the Gauss–Bonnet theorem [17, 18, 49], but most involve it [2, 7, 8, 24, 36, 47]. These latter are mainly based on [7] for the surfaces and on [8] for the volumes. As shown in Sect. 4.2.7, neglecting the intersections of order four and higher is numerically not acceptable for van der Waals surfaces and volumes calculation.

4.2.5 Intersections of Order 4

We set $n = 4$ and we iterate again Eqs. (4.12) and (4.13):

$$\begin{aligned} V &= V_1 + V_2 + V_3 + V_4 - V_{12} - V_{13} - V_{14} - V_{23} - V_{24} - V_{34} \\ &\quad + V_{123} + V_{124} + V_{134} + V_{234} - V_{1234}, \end{aligned} \tag{4.20}$$

$$\begin{aligned} S &= S_1 + S_2 + S_3 + S_4 - S_{12} - S_{13} - S_{14} - S_{23} - S_{24} - S_{34} \\ &\quad + S_{123} + S_{124} + S_{134} + S_{234} - S_{1234}. \end{aligned} \tag{4.21}$$

We set $d = 3$. For clarity, we do not enumerate all topological configurations generated by four intersecting spheres. If any of the four triplets of spheres is not in the configuration of Fig. 4.5d, we know how to calculate V_{1234} and S_{1234}.

Assuming that the four sphere triplets are in the configuration of Fig. 4.5d, we can classify the topological configurations via locating the four pairs of two contact points $z_+^{(ijk)}$ and $z_-^{(ijk)}$ at the intersections of the boundaries of the three spheres i, j, k, $1 \le i < j < k \le 4$. Each of the two contact points defined by a triplet of spheres can be inside or outside the fourth sphere (we still neglect tangencies and multiplicities). We consider the possible locations of the six remaining contact points. When all of them are inside a sphere, one sphere is included in the union of the three other ones. When all the six remaining contact points are outside a sphere, either the intersection of three spheres is included in the remaining one or the 4-order intersection is empty. This case of empty intersection was called an empty simplicial topology (EST) in [37]. If it happens for some triplet i, j, k that $z_+^{(ijk)}$ is inside the fourth sphere and $z_-^{(ijk)}$ is outside, or conversely, then that happens for all the other triplets: there are three contact points inside a sphere and three outside (this is the general case of intersection of four spheres). The two remaining cases correspond in fact to the same configuration: the intersection of two spheres is included in the union of the two other ones.

To summarize, when the four sphere triplets are in the configuration of Fig. 4.5d, the possible configurations for a non-empty 4-order intersection are:

1. One sphere is included in the union of the three other ones.
2. The intersection of two spheres is included in the union of the two other ones.
3. The intersection of three spheres is included in the remaining one.
4. None of the above ones: general case.

Calculating the 4-order intersection reduces to calculate the 3-order intersections in the cases of configurations 1,2,3. The total number of pairs of interior contact points for these latter configurations are respectively 3,2,1, and the total number of pairs of exterior contact points are respectively 1,2,3. Only the EST configuration has 4 exterior contact point pairs.

In the general case, the domain of the 4-order intersection is bounded by four spherical triangles separated by six arcs of circle and intersecting at four contact points: $z^{(123)}$, $z^{(124)}$, $z^{(134)}$, $z^{(234)}$. This convex domain is topologically similar to a tetrahedron. From Theorem 4.3, there is a unique point z_0 at the intersection of the 6 radical planes. Its barycentric coordinates are computed from Eq. (4.14) and (4.15). The point z_0 is interior to the convex domain of the 4-order intersection. This latter is partitioned into four parts, each one being the intersection of one sphere and one trihedron originating in z_0.

These trihedra, which intersect respectively the spheres 1, 2, 3, 4, are:

$$\{z_0 \to x^{(123)}, z_0 \to x^{(124)}, z_0 \to z^{(134)}\},$$
$$\{z_0 \to x^{(123)}, z_0 \to x^{(124)}, z_0 \to z^{(234)}\},$$
$$\{z_0 \to x^{(123)}, z_0 \to x^{(134)}, z_0 \to z^{(234)}\},$$
$$\{z_0 \to x^{(124)}, z_0 \to x^{(134)}, z_0 \to z^{(234)}\}.$$

It follows that the calculation of V_{1234} and S_{1234} can be done through the calculation of the portion of the area of a sphere defined by its intersection with a trihedron having its origin in the interior of the sphere. This latter calculation is done analytically in Sect. 4.2.6 using the Gauss–Bonnet theorem. An analytical treatment of the four-sphere intersection without invoking the Gauss–Bonnet theorem was reported [27].

4.2.6 *Intersection of a Sphere with a Dihedron or with Trihedron*

We consider a sphere of center c and radius R intersecting a salient trihedron of origin z inside the sphere. The trihedron intersects the boundary of the sphere at x_1, x_2, x_3, these three points being ordered in the direct sense when referred to z as origin. We look for the calculation of the volume V_{st} of the intersection and for the surface S_{st} of the spherical triangle on the boundary of the sphere. This spherical triangle is bounded by the oriented arcs of circle (x_1,x_2), (x_2,x_3), (x_3,x_1). These arcs have respective radii h_{12}, h_{23}, h_{31} and respective angles b_{12}, b_{23}, b_{31}, each of these angles being in $[0;\pi]$. The trihedron intersects each of the planes respectively tangent to the sphere at x_1, x_2, x_3, thus defining the respective angles a_1, a_2, a_3, each of these angles being in $[0;\pi]$. We define c_{12}, c_{23}, c_{31} as the projections of c on the planes containing the respective triplets (x_1,z,x_2), (x_2,z,x_3), (x_3,z,x_1). We also define the angles associated to the three lens: $\theta_{ij} = \widehat{x_i,c,c_{ij}} = \widehat{x_j,c,c_{ij}}$, $(i,j)=(1,2)$ or $(i,j)=(2,3)$ or $(i,j)=(3,1)$, such that $\cos\theta_{ij}$ is positive when the vector $c_{ij}-c$ of origin c has the same sense than the direct normal to the plane of the oriented arc (x_i,x_j), and $\cos\theta_{ij}$ is negative when $c_{ij}-c$ has a sense opposite to this normal. We have also $\sin\theta_{ij} = h_{ij}/R$. The sphere has a constant Gaussian curvature and each arc of circle has a constant geodesic curvature. Thus, the Gauss–Bonnet theorem [26, 52] leads to a simple expression of S_{st} in Eq. (4.22), in which the first term in parenthesis is the spherical excess:

$$\frac{S_{st}}{R^2} = (a_1 + a_2 + a_3 - \pi) - (b_{12}\cos\theta_{12} + b_{23}\cos\theta_{23} + b_{31}\cos\theta_{31}). \tag{4.22}$$

V_{zx} being the signed volume of the oriented tetrahedron $(x_1 - z, x_2 - z, x_3 - z)$ and V_{cx} being the signed volume of the oriented tetrahedron $(x_1 - c, x_2 - c, x_3 - c)$, the volume V_{st} is given below [37]:

$$\begin{aligned} V_{st} = \frac{R^3}{6} \big[&\cos\theta_{12} \sin^2\theta_{12} (\sin b_{12} - b_{12}) \\ &+ \cos\theta_{23} \sin^2\theta_{23} (\sin b_{23} - b_{23}) \\ &+ \cos\theta_{31} \sin^2\theta_{31} (\sin b_{31} - b_{31}) \big] + V_{zx} - V_{cx} + \frac{S_{st} R}{3}. \end{aligned} \tag{4.23}$$

When z lies at c, we retrieve the expression given by Girard's theorem [48]: $\frac{S_{st}}{R^2}$ is the spherical excess and $V_{st} = \frac{S_{st} R}{3}$.

Remark: rather than looking for the validity of Eq. (4.23) from integrating S_{st} as stated in [37], we give a hint of proof as follows. We define the point x_0 at the intersection of the half line of origin c passing through z with the boundary of the sphere, and we add the three signed contributions to V_{st} due to the trihedra $\{z \to x_1, z \to x_2, z \to x_0\}$ and $\{z \to x_2, z \to x_3, z \to x_0\}$ and $\{z \to x_3, z \to x_1, z \to x_0\}$, a contribution being negative when the bounding arcs are not oriented in the conventional sense. For each trihedra, say, $\{z \to x_1, z \to x_2, z \to x_0\}$, the volume of the intersection with the sphere is $V_{sc012} - V_{cz12} - V_{c12}$, in which V_{sc012} is the volume of the domain bounded by $c - x_0$, $c - x_1$, $c - x_2$ and by the spherical triangle (x_0, x_1, x_2) intercepted by the trihedron $\{z \to x_1, z \to x_2, z \to x_0\}$ and of area computable by Eq (4.22). Despite that this domain is not bounded by a trihedron, V_{sc012} is computable by trivial integration in respect to the radius of the sphere. V_{cz12} is the volume of the tetrahedron (c, z, x_1, x_2). V_{c12} is the volume of the portion of the cone of origin c with height $R\cos\theta_{12}$ and of basis the circular segment of area $(b_{12} - \sin b_{12})(R \sin\theta_{12})^2/2$, i.e., $V_{c12} = \cos\theta_{12} \sin^2\theta_{12} (b_{12} - \sin b_{12}) R^3/6$. The rest of the calculation is easy.

The case of the six trihedra listed at the end of Sect. 4.2.4 is of interest. It is such that one of the arcs, say, (x_2, x_3), lies on a great circle of the sphere: z, x_2, x_3, c, c_{12}, c_{31}, are coplanar, and c_{23} coincides with c, then $a_2 = \pi/2$, $a_3 = \pi/2$, and $\theta_{23} = \pi/2$.

Grouping each trihedra with its symmetric image through the plane of the great circle let us to the expressions of the volume V_{sd} of the intersection of a sphere and a dihedron and for the surface S_{sd} of the associated portion of spherical area. Keeping the notations used in Eqs. (4.22) and (4.23):

$$\frac{S_{sd}}{R^2} = 2a_1 - 2(b_{12}\cos\theta_{12} + b_{31}\cos\theta_{31}), \tag{4.24}$$

$$\begin{aligned} V_{sd} = \frac{R^3}{3} \big[&\cos\theta_{12} \sin^2\theta_{12} (\sin b_{12} - b_{12}) \\ &+ \cos\theta_{31} \sin^2\theta_{31} (\sin b_{31} - b_{31}) \big] + 2(V_{zx} - V_{cx}) + \frac{S_{sd} R}{3} \Bigg]. \end{aligned} \tag{4.25}$$

Care: in Eqs. (4.24) and (4.25), b_{12} and b_{31} are the angles of the half arcs (not the full arcs) bounding the intersection of the sphere with the dihedron. It is also recalled that the arcs must be correctly oriented and that $\cos\theta_{12}$ and $\cos\theta_{31}$ can be negative.

When z is outside the sphere, the intersection with a dihedron reduces to simpler cases such as a sphere minus two lens, and the intersection with a trihedron can be reduced to intersections with appropriate dihedra or trihedra for which Eqs. (4.22)–(4.25) can be used. More generally we can extend volumes and surfaces computations to the intersections of a sphere with a tetrahedron, a convex polyhedron, and with any union of convex polyhedra. These computations are based on Eqs. (4.22)–(4.25) but are not needed in the framework of this chapter.

4.2.7 Intersections of Order 5 and Higher

By iterating Eqs. (4.12) and (4.13) until order n, the volume and the surface of the union of the n spheres can be expressed from the inclusion–exclusion principle as follows:

$$V = \sum_{1\le i\le n} V_i - \sum_{1\le i_1<i_2\le n} V_{i_1 i_2} + \sum_{1\le i_1<i_2<i_3\le n} V_{i_1 i_2 i_3} + \cdots + (-1)^{n-1} V_{i_1 i_2 \cdots i_n}, \quad (4.26)$$

$$S = \sum_{1\le i\le n} S_i - \sum_{1\le i_1<i_2\le n} S_{i_1 i_2} + \sum_{1\le i_1<i_2<i_3\le n} S_{i_1 i_2 i_3} + \cdots + (-1)^{n-1} S_{i_1 i_2 \cdots i_n}. \quad (4.27)$$

Setting $d = 3$ and $n \ge 5$, the existence of intersections of order n is deduced from Helly's theorem [14]. It means that the intersection of five spheres is not empty if and only if each of the five subsets of four spheres gives rise to a non-empty intersection. If $n \ge 6$, the intersection of six spheres is not empty if and only if each of the six subsets of five spheres gives rise to a nonempty intersection.

For practical computations, we need Theorem 4.5:

Theorem 4.5. *When $n \ge 5$ spheres in E^3 have a common non-empty intersection, there are m (with $1 \le m \le 3$) of these spheres such that their union J contains the intersection I of the $n - m$ remaining spheres.*

Proof. Assume initially that $n = 5$ spheres have a common non-empty intersection. From Helly's theorem, each of the five subsets of four spheres gives rise to a nonempty intersection. If one of these subsets of four spheres is such that the union of $m = 1, 2$, or three spheres contains the intersection of the $4 - m$ remaining spheres, the union of these m spheres contains the intersection of the $5 - m$ remaining spheres and the theorem stands. As stated in Sect. 4.2.5, a set of four spheres giving rise to a nonempty intersection which is not in one of these latter configurations is in the general configuration.

Now we need to prove the theorem when the five subsets of four spheres are in the case of a general configuration. For this latter, the four contact point pairs between three spheres are such that one of these contact points lies inside the fourth sphere

5- NN2 7/3 / ") "44) %) *)) #))

!'#' $ % % & ' ( )) * +) , -)
.) @) + !) * + 4

,9
*) 6N
( ) 76
% /7

5- 2 3

E P)4 % *) A % -))
) *) %))) 4 (
)) #) %)) G 4 -) %)) H
)
))4) 4 -) 4) =)
+ D) % + 4))) 9) -) 8)) 9
) -))9)+ # *) &
)) %)) 9 9+ - % 8 -)) -
D ')) +

Theorem 1.1.1 *For any two distinct points p and q there exists a point r which lies between p and q, i.e. the line segment $\overline{pq}$ is not empty.*

Proof Let p and q be two points. By axiom I_4 there exists a point s that does not lie on the straight line $L(p,q)$. By axiom A_3 there is a point t such that s lies between p and t. Another application of axiom A_3 gives a point u such that q lies between t and u. The straight line $L := L(s,u)$ intersects the line segment $\overline{pt}$ at s.

The point t does not lie on the straight line $L(p,q)$, since otherwise s would by axiom A_1 also lie on that straight line, contradicting the choice of s. We can therefore apply axiom A_5 to the straight line L and the three points p, q and t. As L intersects the line segment $\overline{pt}$, L must by axiom A_5 also intersect one of the two line segments $\overline{pq}$ or $\overline{tq}$, unless it contains one of the three points p, q or t.

First case L contains p or t.

Then L agrees with the straight line $L(p,t)$ by axiom I_2. Hence u lies on $L(p,t)$ and axiom A_1 implies that q lies on $L(p,t)$ as well. Hence p, q and t do lie on a straight line, a contradiction.

Second case L contains q.

Several applications of axiom I_2 show that the points s, u, q, t and p lie on a straight line, a contradiction.

Third case L intersects the line segment $\overline{tq}$ at a point v.

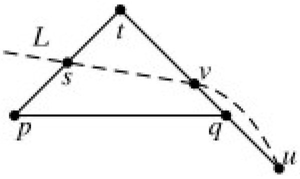

Then L and $L(t,q)$ have the two points u and v in common. If $u = v$, then u would lie between t and q while q lies between t and u, a contradiction to axiom A_4. Thus L and $L(t,q)$ have two distinct points in common, and axiom I_2 implies that $L = L(t,q)$. But then both s and p lie on L, i.e. p, q and t lie on a straight line, a contradiction.

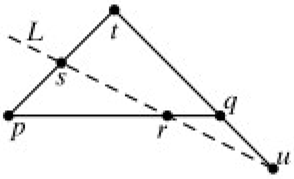

Thus L and $\overline{pq}$ must intersect at a point r. In particular, $\overline{pq}$ cannot be empty. □

Exercise 1.1 Let p, q, r and s be points on a straight line. Show that if q lies between p and s and if additionally r lies between q and s, then r lies between p and s as well.

Using this exercise and theorem 1.1.1, it is now easy to solve the following exercise.

Exercise 1.2 Show the following: there is an infinite number of distinct points between two points.

Definition 1.1.2 Let L be a straight line, $p \in L$. Let q and r be two points on L, neither of which is p. We say that q and r ***lie on the same side of the point*** p if p does not lie between q and r.

Exercise 1.3 Let L be a straight line and let $p \in L$ be a point on L. Show that the relation "q_1 lies on the same side of p as q_2" defines an equivalence relation on the set $\{q \in L \mid q \neq p\}$.

An equivalence class of points on L that do not equal p can then be referred to as a ***side of*** p ***on*** L.

Exercise 1.4 Show that there are exactly two sides of p on L.

Definition 1.1.3 Let L be a straight line and let p and q be two points that do not lie on L. We say that p and q ***lie on the same side of a straight line*** L if the line segment $\overline{pq}$ does not intersect the straight line L.

Exercise 1.5 Let L be a straight line. Show that the relation "q_1 lies on the same side of L as q_2" defines an equivalence relation on the set $\{q \mid q \notin L\}$.

Again we can call an equivalence class of points not on L a ***side of*** L.

Exercise 1.6 Show that there are exactly two sides of L.

Congruence axioms To formulate the third group of axioms, the congruence axioms, we need in addition to the previous notions that for every pair $(\overline{pq}, \overline{p_1q_1})$ of line segments the formal statement "$\overline{pq}$ is congruent to $\overline{p_1q_1}$" is either true or false.

Axiom K_1 (reproduction of lengths) *Let $\overline{pq}$ be a line segment, let L_1 be a straight line, let $p_1, r_1 \in L_1$, $r_1 \neq p_1$. Then there is a point $q_1 \in L_1$ on the same side of p_1 as r_1 such that $\overline{pq}$ is congruent to $\overline{p_1q_1}$.*

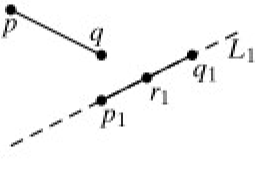

In this axiom only the existence of a congruent line segment is required. Its uniqueness needs to be proved later with the aid of other axioms.

Axiom K_2 *If the line segments $\overline{p_1q_1}$ and $\overline{p_2q_2}$ are both congruent to the line segment $\overline{pq}$, then $\overline{p_1q_1}$ is congruent to $\overline{p_2q_2}$ as well.*

Four more congruence axioms will follow. We can nevertheless already prove a first implication.

Lemma 1.1.4 *The congruence of line segments defines an equivalence relation on the set of line segments.*

Proof (a) Let $\overline{pq}$ be a line segment. We show that $\overline{pq}$ is congruent to itself. Let L be a straight line that contains p, $p \in L$. By axiom K_1 there exists a point r on L such that $\overline{pq}$ and $\overline{pr}$ are congruent. Then $\overline{p_1q_1} := \overline{p_2q_2} := \overline{pq}$ is congruent to $\overline{pr}$ and hence $\overline{pq}$ is by axiom K_2 congruent to itself.

(b) (symmetry) Let $\overline{pq}$ be congruent to $\overline{p_1q_1}$. We show that then $\overline{p_1q_1}$ is congruent to $\overline{pq}$. It follows from (a) that $\overline{p_1q_1}$ is congruent to $\overline{p_1q_1}$. Axiom K_2 now gives that $\overline{p_1q_1}$ is congruent to $\overline{pq}$.

(c) (transitivity) If $\overline{p_1q_1}$ is congruent to $\overline{p_2q_2}$ and $\overline{p_2q_2}$ is congruent to $\overline{p_3q_3}$, then we need to show that $\overline{p_1q_1}$ is congruent to $\overline{p_3q_3}$. This follows directly from axiom K_2 together with (b). ☐

We will from now on sometimes denote "$\overline{p_1q_1}$ is congruent to $\overline{p_2q_2}$" by "$\overline{p_1q_1} \equiv \overline{p_2q_2}$".

Axiom K_3 (additivity of line segments) *Let L and L_1 be straight lines, let $p, q, r \in L$ be three pairwise distinct points on L and $p_1, q_1, r_1 \in L_1$ likewise on L_1. Assume that the line segments $\overline{pq}$ and $\overline{qr}$ do not have any common points, $\overline{pq} \cap \overline{qr} = \emptyset$. Analogously let $\overline{p_1q_1} \cap \overline{q_1r_1} = \emptyset$.*

If now $\overline{pq} \equiv \overline{p_1q_1}$ and $\overline{qr} \equiv \overline{q_1r_1}$, then $\overline{pr} \equiv \overline{p_1r_1}$.

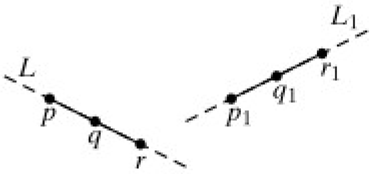

We need the concept of the *angle* for the formulation of the three other congruence axioms.

Definition 1.1.5 An ***angle*** is an equivalence class of triples of points p, q and r that do not lie on a straight line, where two triples (p, q, r) and (p_1, q_1, r_1) are equivalent if

(i) $q = q_1$,
(ii) $L(p, q) = L(p_1, q)$ and p and p_1 lie on the same side of q,
(iii) $L(r, q) = L(r_1, q)$ and r and r_1 lie on the same side of q,

or if

(i) $q = q_1$,
(ii) $L(p, q) = L(r_1, q)$ and p and r_1 lie on the same side of q,
(iii) $L(r, q) = L(p_1, q)$ and r and p_1 lie on the same side of q.

For the equivalence class of (p, q, r) we write $\angle(p, q, r)$. The point q is then called the ***vertex*** of the angle $\angle(p, q, r)$.

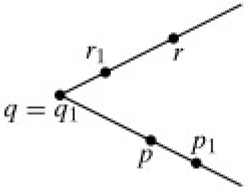

We now additionally require that for any two angles $\angle(p, q, r)$ and $\angle(p_1, q_1, r_1)$ the formal statement "$\angle(p, q, r)$ is congruent to $\angle(p_1, q_1, r_1)$" is either true or false. Again, we write "$\angle(p, q, r) \equiv \angle(p_1, q_1, r_1)$" if $\angle(p, q, r)$ is congruent to $\angle(p_1, q_1, r_1)$.

Axiom K_4 *The congruence of angles induces an equivalence relation on the set of angles.*

Axiom K_5 (reproduction of angles) *Let p, q, r be points that do not lie on a straight line, and let p_1, q_1, s_1 be another set of points that do not lie on a straight line. Then there exists a point r_1 on the same side of $L(p_1, q_1)$ as s_1 such that the angle $\angle(p_1, q_1, r_1)$ is congruent to the angle $\angle(p, q, r)$.*

If r_2 is another point with the same properties as r_1, i.e. r_2 also lies on the same side of $L(p_1, q_1)$ as s_1 and if $\angle(p_1, q_1, r_2) \equiv \angle(p, q, r)$, then $\angle(p_1, q_1, r_1) = \angle(p_1, q_1, r_2)$.

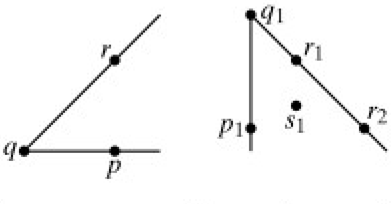

Axiom K_5 says that we can reproduce a given angle in a unique way if we are given the vertex, one line adjacent to the angle and the side of the other line next to the angle.

The last congruence axiom relates the congruence of line segments to that of angles. Until now the two notions of congruence existed entirely separately.

Axiom K_6 *Let* (p,q,r) *be a triple of points that do not lie on a straight line and* (p_1,q_1,r_1) *likewise. If* $\overline{pq} \equiv \overline{p_1q_1}$, $\overline{pr} \equiv \overline{p_1r_1}$ *and* $\angle(q,p,r) \equiv \angle(q_1,p_1,r_1)$, *then*

$$\angle(p,q,r) \equiv \angle(p_1,q_1,r_1).$$

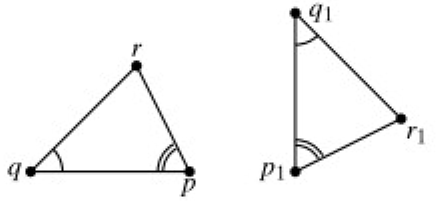

Let us make some inferences using the axioms introduced so far. We will first extend the statement of the last axiom.

Theorem 1.1.6 *Let* (p,q,r) *be a triple of points that do not lie on a straight line,* (p_1,q_1,r_1) *likewise. If* $\overline{pq} \equiv \overline{p_1q_1}$, $\overline{pr} \equiv \overline{p_1r_1}$ *and* $\angle(q,p,r) \equiv \angle(q_1,p_1,r_1)$, *then*

$$\angle(p,q,r) \equiv \angle(p_1,q_1,r_1), \qquad \angle(p,r,q) \equiv \angle(p_1,r_1,q_1), \qquad \overline{qr} \equiv \overline{q_1r_1}.$$

Proof The angle congruences follow directly from axiom K_6, in the second case after renaming the variables. It remains to show that $\overline{qr} \equiv \overline{q_1r_1}$. By axiom K_1 we can find a point s_1 on the straight line $L(q_1,r_1)$ which is on the same side as r_1 and satisfies $\overline{qr} \equiv \overline{q_1s_1}$.

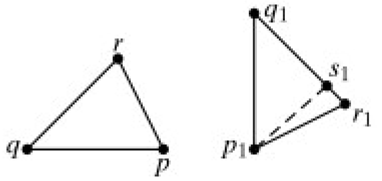

We apply axiom K_6 to (p,q,r) and (p_1,q_1,s_1) and conclude that

$$\angle(q,p,r) \equiv \angle(q_1,p_1,s_1).$$

On the other hand we have $\angle(q,p,r) \equiv \angle(q_1,p_1,r_1)$, and by the uniqueness from axiom[1] K_5

---

[1] For an application of the uniqueness statement from axiom K_5 it must be ensured that r_1 and s_1 lie on the same side of $L(p_1,q_1)$. This is left as an exercise for the reader.

$$\angle(q_1, p_1, s_1) = \angle(q_1, p_1, r_1).$$

If we now had $r_1 \neq s_1$, then we would conclude that p_1 and q_1 both lie on the straight line $L(r_1, s_1)$, i.e. p_1, q_1 and r_1 would lie on a straight line, which contradicts the assumption.

Thus $r_1 = s_1$ and hence $\overline{qr} \equiv \overline{q_1 r_1}$. ☐

Theorem 1.1.7 (Congruence of adjacent angles) *Suppose that the pairwise distinct points p, q and s lie on a straight line L, while $r \notin L$. Analogously, let $p_1, q_1, s_1 \in L_1$ be pairwise distinct, $r_1 \notin L_1$. If $\angle(p,q,r)$ and $\angle(p_1,q_1,r_1)$ are congruent, then the same is true for $\angle(s,q,r)$ and $\angle(s_1,q_1,r_1)$.*

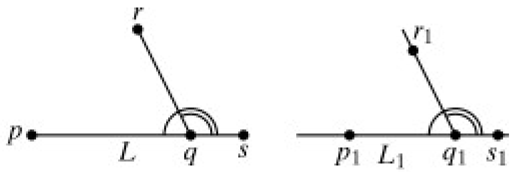

The angle $\angle(s,q,r)$ is sometimes called the ***adjacent angle*** of $\angle(p,q,r)$. The theorem thus says that adjacent angles of congruent angles are congruent as well.

Proof The points p_1, r_1 and s_1 can by axiom K_1 be assumed to have been chosen in such a way that $\overline{pq} \equiv \overline{p_1 q_1}$, $\overline{rq} \equiv \overline{r_1 q_1}$ and $\overline{sq} \equiv \overline{s_1 q_1}$.

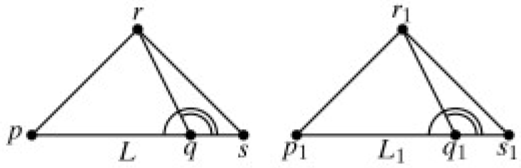

From theorem 1.1.6, applied to (p,q,r) and (p_1,q_1,r_1), it follows that $\overline{pr} \equiv \overline{p_1 r_1}$. By axiom K_3 we then have $\overline{ps} \equiv \overline{p_1 s_1}$. Applying theorem 1.1.6 once again, this time to (r,p,s) and (r_1,p_1,s_1), we obtain $\overline{rs} \equiv \overline{r_1 s_1}$ and $\angle(q,s,r) \equiv \angle(q_1,s_1,r_1)$. Axiom K_6 then says for (q,s,r) and (q_1,s_1,r_1) that $\angle(s,q,r) \equiv \angle(s_1,q_1,r_1)$. ☐

Theorem 1.1.8 (Congruence of vertical angles) *Let L and M be two distinct straight lines that intersect at p. Let $r, q \in L$ lie on two distinct sides of p and let $s, t \in M$ lie on two distinct sides of p as well. Then*

$$\angle(q,p,s) \equiv \angle(r,p,t).$$

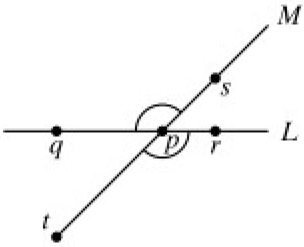

Proof Both $\angle(q,p,s)$ and $\angle(r,p,t)$, are adjacent angles of $\angle(q,p,t)$. The angle $\angle(q,p,t)$ is congruent to itself by axiom K_4. The claim therefore follows by theorem 1.1.7. ☐

After these preparations we now come to the first truly interesting geometric theorem.

Theorem 1.1.9 (Existence of a parallel) *Let L be a straight line, p a point, $p \notin L$. There then exists a straight line M that contains p and that does not intersect L.*

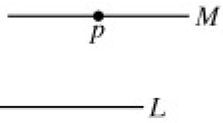

We then say that M is a ***parallel*** to L through p. The theorem says that such parallels always exist.

Proof Let L be a straight line and p a point that does not lie on L. We will first construct the line M and then show that it has the desired properties.

For the construction we choose a point $q \in L$ and add the straight line $N := L(p,q)$. We choose another point $r \in L$, $r \neq q$. Then $r \notin N$, since otherwise $p \in L(p,q) = L(q,r) = L$. We reproduce angle $\angle(r,q,p)$ as in axiom K_5 on the straight line N at the point p, i.e. we find points $s \in N$ and $t \notin N$ on the same side of N as r, such that the angle $\angle(t,p,s)$ is congruent to the angle $\angle(r,q,p)$. We now set $M := L(p,t)$.

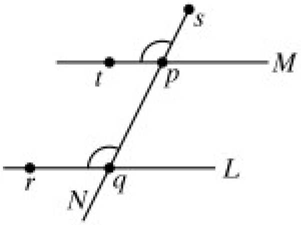

It remains to show that L and M do not intersect. Suppose that L and M did intersect at a point u. We restrict ourselves to the case that u lies on the same side of N as r and t. The other case is dealt with in a similar way.

We reproduce the line segment $\overline{uq}$ on the straight line M as in axiom K_1, beginning at the point p, and we do *not* do this on the same side as u. This means that $\overline{uq} \equiv \overline{pv}$, with a point $v \in M$ such that p lies between u and v.

The angles $\angle(u,p,s)$ and $\angle(q,p,v)$ are congruent by theorem 1.1.8. Hence the angles $\angle(u,q,p)$ and $\angle(q,p,v)$ are congruent as well. We apply axiom K_6 to (u,q,p) and (v,p,q), and see that $\angle(q,p,u) \equiv \angle(p,q,v)$.

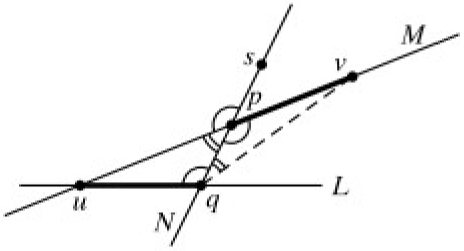

But $\angle(q,p,u)$ is an adjacent angle of $\angle(u,p,s)$, so by theorem 1.1.7 and by the uniqueness of the reproduction of angles $\angle(p,q,v)$ must be an adjacent angle of $\angle(u,q,p)$. Hence v lies on the straight line L, and thus the same is true for p, which contradicts the assumption. ☐

Parallel axiom We have seen that the existence of parallels can be proved using the axioms introduced so far. What about their uniqueness? This is the subject of the parallel axiom.

AXIOM P (parallel axiom) *Let L be a straight line, p a point, $p \notin L$. Then there is at most one straight line that contains p and does not intersect L.*

There has been controversy about the necessity of this axiom for thousands of years. Many mathematicians believed that uniqueness could, as existence, be deduced from the other axioms. There were many proof attempts. Carl Friedrich Gauss (1777–1855) was probably the first person to truly believe in the independence of the parallel axiom. However, he never published his views about it. The debate was ended by the Russian Nikolai Iwanowitsch Lobatschewski (1792–1856) and by the Hungarian Janos Bolyai (1802–1860), who, independently of each other, found a geometry that satisfies all axioms except for the parallel axiom. Hence the parallel axiom cannot be derived from the other axioms. Bolyai's father, a teacher of mathematics himself, was so worried about his son's result that he wrote a letter to Gauss, asking him for his opinion. In his answer Gauss was enthusiastic about the work of the younger Bolyai, but added the remark that he could not praise him, since this would be self-praise as he, Gauss, had known this for many years. Janos Bolyai never again published work on mathematics.

The afore-mentioned geometry is hyperbolic geometry, also known as non-Euclidean geometry. We will cover it in more detail later.

,9 J))));

Completeness axioms When real numbers are introduced in analysis, one becomes acquainted with the completeness axioms, which distinguish the real numbers from the rational numbers and other ordered number fields. To determine the Euclidean geometry completely, we need the appropriate axioms.

Axiom V_1 (Archimedes's axiom) *Let $\overline{pq}$ and $\overline{rs}$ be two line segments. Then there exists a natural number n such that the line segment $\overline{r_1 s_n}$, which results from the n-fold reproduction of the line segment $\overline{rs}$ on the straight line $L(p,q)$, starting at p in direction q, contains the line segment $\overline{pq}$.*

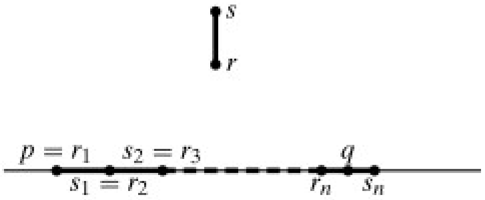

Before formulating the last axiom, we summarise those notions that we need for the axiomatic formulation of Euclidean geometry. We have

- a set $\mathscr{P}$, whose elements are called points,
- a set $\mathscr{G}$, whose elements we call straight lines,
- a relation $\in$ between $\mathscr{P}$ and $\mathscr{G}$,
- a three-figure relation “between” on $\mathscr{P}$,
- a relation $\equiv_1$ on the set of straight lines,
- a relation $\equiv_2$ on the set of angles.

One could now formally define a Euclidean geometry as a 6-tuple $(\mathscr{P}, \mathscr{G}, \in, \text{between}, \equiv_1, \equiv_2)$ that consists of such notions, and that satisfies axioms I_1–I_4, A_1–A_5, K_1–K_6, P, V_1 as well as V_2, which still is to be formulated.

An ***extension*** of our geometry is a second 6-tuple $(\mathscr{P}', \mathscr{G}', \in', \text{between}', \equiv_1', \equiv_2')$ such that $\mathscr{P} \subset \mathscr{P}'$, $\mathscr{G} \subset \mathscr{G}'$ and the relations $\in'$, between$'$, $\equiv_1'$ as well as $\equiv_2'$ agree with the corresponding relations $\in, \ldots$ after a restriction to $\mathscr{P}$ and $\mathscr{G}$.

Axiom V_2 (maximality) *Let $(\mathscr{P}', \mathscr{G}', \in', \text{between}', \equiv_1', \equiv_2')$ be an extension of our geometry. Then $\mathscr{P}' = \mathscr{P}$ and $\mathscr{G}' = \mathscr{G}$.*

Our geometry is therefore assumed to be non-extensible, i.e. maximal.

For a further discussion of axiomatic geometry the reader is referred to Hilbert's book [13]. Proofs of numerous geometric theorems, a discussion of the uniqueness of Euclidean geometry and of the necessity of Archimedes's axiom, straightedge-and-compass constructions and spatial Euclidean geometry can be found there. On a final note, some axioms can be weakened slightly, e.g. axioms K_4 and V_2.

1.2 The Cartesian model

While the axiomatic method ensures logical clarity in the structural layout of geometry, it is also annoyingly clumsy. The proofs of even relatively simple geometric facts can easily become quite laborious. In addition, dealing with geometric objects that are not composed of line segments, circular lines, and so forth is relatively involved.

We therefore now follow the ideas of the Frenchman René Descartes (1596–1650) and characterise points by coordinates, which give the positions of the points in the plane. This allows us to use methods from algebra and infinitesimal calculus in geometry, extending our mathematical toolkit considerably.

We hence make the definition

$$\mathscr{P} := \mathbb{R}^2.$$

Straight lines are defined as sets of points of the form

$$L = L_{p,v} := \{x \in \mathbb{R}^2 \mid x = p + t \cdot v,\ t \in \mathbb{R}\},$$

where $p, v \in \mathbb{R}^2$, $v \neq 0$, are fixed. The set of straight lines is then

$$\mathscr{G} := \{L_{p,v} \mid p, v \in \mathbb{R}^2, v \neq 0\}.$$

We say that a point p is ***contained*** in a straight line L if $p \in L$ in the set-theoretic sense.

Exercise 1.7 Verify the validity of axioms I_1–I_4.

A point $q \in \mathbb{R}^2$ lies ***between*** p and $r \in \mathbb{R}^2$, $p \neq r$, if there exists a $t \in (0,1)$ such that $q = t \cdot p + (1-t) \cdot r$.

Exercise 1.8 Show that axioms A_1–A_4 are valid.

It remains to define the congruence relations. This is done using a suitable group of maps, which acts on the set of points.

Definition 1.2.1 Let $A \in \mathrm{O}(n)$ be an orthogonal matrix, i.e. it satisfies $A \cdot A^\top = \mathrm{Id}$, where $A^\top$ is the transpose of A. Let $b \in \mathbb{R}^n$.

Then the map

$$F_{A,b} : \mathbb{R}^n \to \mathbb{R}^n, \qquad F_{A,b}(x) = Ax + b,$$

is called a ***Euclidean motion***. The vector b is sometimes called the ***translational part***. For a fixed dimension n, we call the set of all Euclidean motions,

$$\mathrm{E}(n) := \{F_{A,b} \mid A \in \mathrm{O}(n), b \in \mathbb{R}^n\},$$

the ***Euclidean motion group***.

Exercise 1.9 Show the following: composition of maps turns $\mathrm{E}(n)$ into a group with neutral element Id. Products and inverses are given by $F_{A,b} \circ F_{B,c} = F_{AB,Ac+b}$ and $F_{A,b}^{-1} = F_{A^{-1},-A^{-1}b}$ respectively.

Exercise 1.10 Show that axiom A_5 is valid.

Hint An application of a Euclidean motion maps the straight line L to the x-axis $L' = \{(x,0)^\top \mid x \in \mathbb{R}\}$. Then argue that two points $(x,y)^\top$ and $(x',y')^\top$ from $\mathbb{R}^2 - L'$ lie on the same side of L' if and only if y and y' are both positive or both negative.

Definition 1.2.2 Two line segments $\overline{pq}$ and $\overline{rs}$ are said to be ***congruent*** if there exists a Euclidean motion $F \in \mathrm{E}(2)$ such that

$$\overline{F(p)F(q)} = \overline{rs}.$$

Analogously, the angle $\angle(p,q,r)$ is ***congruent*** to $\angle(p_1,q_1,r_1)$ if there is an $F \in \mathrm{E}(2)$ such that

$$\angle(F(p),F(q),F(r)) = \angle(p_1,q_1,r_1).$$

Exercise 1.11 Show the validity of axioms K_1–K_6.

Exercise 1.12 Show the validity of the parallel axiom P.

Exercise 1.13 Show the validity of Archimedes's axiom V_1.

The proof of the maximality axiom V_2 requires more thought than those of the other axioms. We therefore want to show it here.

Theorem 1.2.3 *With the definitions made above the maximality axiom* V_2 *is valid.*

Proof Let $\mathscr{P}' \supset \mathscr{P}$ and $\mathscr{G}' \supset \mathscr{G}$ be respectively the set of points and the set of straight lines of an extension of our Cartesian model of Euclidean geometry. We will call the points from $\mathscr{P}$ *old points*, and those from $\mathscr{P}' - \mathscr{P}$ *new points*. We use analogous notation when dealing with straight lines. It needs to be shown that there cannot be new points and straight lines at all, $\mathscr{P} = \mathscr{P}'$ and $\mathscr{G} = \mathscr{G}'$.

The proof is accomplished in three steps. We first show that the old straight lines do not contain new points, then that there are no new points, and finally that there are no new straight lines.

(a) *Old straight lines do not contain new points, i.e. for* $L \in \mathscr{G}$ *and* $p \in \mathscr{P}'$ *with* $p \in' L$ *we have* $p \in \mathscr{P}$.

One must not become confused at this point. According to our definition, old straight lines are sets of old points. In the *set-theoretic sense* old straight lines only contain old points anyway. Nevertheless, it is a priori possible that new points are contained in old straight lines for the extension $\in'$ of the set-theoretic inclusion $\in$. It now needs to be shown that this is not the case.

For this purpose let $L \in \mathscr{G}$ be an old straight line. Suppose that L contains a new point n. We choose an old point $p_1 \in L$. As L is of the form $L = L_{p,v}$, we can write $p_1 = p + t_1 v =: c(t_1)$, $t_1 \in \mathbb{R}$. If we reproduce the line segment $\overline{(0,0)^\top(0,1)^\top}$ sufficiently often on L, starting at p_1 in direction n, then we obtain by Archimedes's axiom a second point $q_1 \in L$ such that n lies between p_1 and q_1. Reproducing the line segment $\overline{(0,0)^\top(0,1)^\top}$ on an old straight line starting at an old point always gives another old point. Thus q_1 is an old point. We write $q_1 = c(s_1)$, $s_1 \in \mathbb{R}$. Without loss of generality let $t_1 < s_1$.

We now decompose the real numbers into two disjoint subsets, $\mathbb{R} = T \sqcup S$, where $T = \{t \in \mathbb{R} \mid c(t) \text{ lies on the same side of } n \text{ as } c(t_1)\}$ and $S = \{t \in \mathbb{R} \mid c(t) \text{ lies on the same side of } n \text{ as } c(s_1)\}$. This decomposition of the real numbers constitutes a Dedekind cut. By the completeness of $\mathbb{R}$ either the subset T must have a maximum or S a minimum. We deal with the case that T has a maximum t_2. The other case can be treated analogously.

We set $p_2 := c(t_2)$. As $t_2 \in T$, the points p_1 and p_2 must lie on the same side of n. By Archimedes's axiom, there is a natural number k such that k-fold reproduction of the line segment $\overline{p_2 n}$ on L from p_2 in direction q_1 contains the line segment $\overline{p_2 q_1}$. We set $p_3 := c(t_3)$ with $t_3 = t_2 + (s_1 - t_2)/k$. Then the line segment $\overline{p_2 p_3}$ has the property that its k-fold reproduction on L from p_2 in direction q_1 gives the line segment $\overline{p_2 q_1}$.

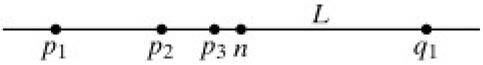

As the line segment $\overline{p_2 q_1}$ is contained in the k-fold reproduced line segment $\overline{p_2 n}$, the line segment $\overline{p_2 p_3}$ must be contained in $\overline{p_2 n}$. Thus p_3 lies between p_2 and n. It follows that $t_3 \in T$ and $t_3 > t_2$. This contradicts the maximality of t_2.

(b) *There are no new points, i.e.* $\mathscr{P} = \mathscr{P}'$.

Let $p \in \mathscr{P}'$ be a point. We choose an old point $q \in \mathscr{P}$ and consider the straight line L through p and q. We now choose three old points $r, s, t \in \mathscr{P}$ that do not lie on a straight line and none of which lies on L, such that L and the line segment $\overline{rs}$ intersect at q.

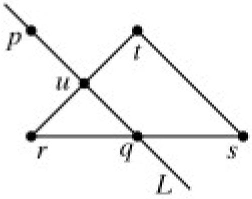

It follows from axiom A_5 that L intersects the line segment $\overline{rt}$ or $\overline{ts}$ at a point u. As the two old straight lines $L(r,t)$ and $L(s,t)$ do not contain any new points by (a), u must be a new point. Thus $L = L(u,q)$ is an old straight line and contains, again by (a), no new points. Hence p is an old point, $p \in \mathscr{P}$.

(c) *There are no new straight lines, i.e.* $\mathscr{G} = \mathscr{G}'$.

Let $L \in \mathscr{G}'$ be a straight line. It follows from axiom I_3 that L contains two distinct points p and q. They must be old points by (b). Hence $L = L(p,q)$ is an old straight line, $L \in \mathscr{G}$. $\square$

Exercise 1.14 Give another proof for part (a) of theorem 1.2.3 by considering Cauchy sequences instead of Dedekind cuts.

The axioms of Euclidean plane geometry are thus valid for the Cartesian model. In particular, we see that the axioms are *consistent*.

Doing geometry in the Cartesian model has the advantage that we now have the whole mathematical machinery of differential and integral calculus to hand. This makes the treatment of Euclidean trigonometry relatively easy.

We will as of now use angle brackets to denote the standard scalar product on $\mathbb{R}^n$:

$$\langle x, y\rangle = \sum_{i=1}^{n} x^i y^i$$

for $x = (x^1, \dots, x^n)^\top, y = (y^1, \dots, y^n)^\top \in \mathbb{R}^n$.

If x, y and z are three points in $\mathbb{R}^2$ which do not lie on a straight line, then $|\langle y - x, z - x\rangle| \leq \|y - x\| \cdot \|z - x\|$ by the Cauchy–Schwarz inequality. Hence $|\langle y - x, z - x\rangle/(\|y - x\| \cdot \|z - x\|)| \leq 1$. Since $\cos : [0,\pi] \to [-1,1]$ is bijective we can define their ***interior angle*** as the unique number $\gamma \in [0,\pi]$ such that $\cos(\gamma) = \langle y - x, z - x\rangle/(\|y - x\| \cdot \|z - x\|)$.

Exercise 1.15 Show that if x, y, z, and x', y', z' represent the same angle in the sense of definition 1.1.5, $\angle(y,x,z) = \angle(y',x',z')$, then they have the same interior angle, $\gamma = \gamma'$.

Exercise 1.16 Show that two angles are congruent if and only if they have the same interior angle.

This is analogous to the fact that two line segments are congruent if and only if they have equal length.

Theorem 1.2.4 (cosine rule for Euclidean geometry) *Let* $p, q, r \in \mathbb{R}^2$. *Let* $a = \|p - q\|$, $b = \|p - r\|$ *and* $c = \|q - r\|$ *be the sides of a triangle with vertices* p, q *and* r. *Let* γ *be the interior angle at the vertex* p. *Then*

$$c^2 = a^2 + b^2 - 2ab\cos(\gamma).$$

Proof Euclidean motions do not change the side and angle ratios. We can therefore after the application of a suitable Euclidean motion assume that $p = (0,0)^\top$, $q = (a,0)^\top$ and $r = (x,y)^\top$.

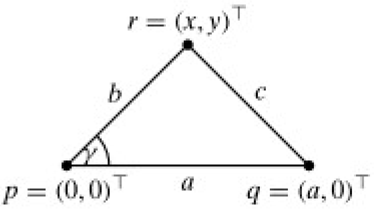

Hence

$$c^2 = (x-a)^2 + y^2 = a^2 - 2ax + x^2 + y^2 = a^2 + b^2 - 2ax.$$

Further

$$\cos(\gamma) = \frac{\langle q, r\rangle}{ab} = \frac{xa + y \cdot 0}{ab} = \frac{ax}{ab}$$

and thus $ax = ab\cos(\gamma)$. The claim follows. □

Corollary 1.2.5 (Pythagoras's theorem) *Using the notation from theorem 1.2.4 and letting the angle be a right angle,* $\gamma = \pi/2$, *the following equality holds:*

$$a^2 + b^2 = c^2.$$ □

Theorem 1.2.6 (sine rule for Euclidean geometry) *If* β *denotes the interior angle at vertex* q *and* α *the one at vertex* r, *then the following equality holds:*

$$\frac{a}{b} = \frac{\sin(\alpha)}{\sin(\beta)}.$$

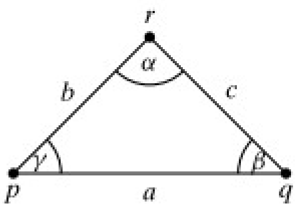

Proof By the cosine rule the angle at vertex r satisfies

$$-2bc\cos(\alpha) = a^2 - (b^2 + c^2)$$

and hence

$$4b^2c^2\cos^2(\alpha) = \left(-a^2 + b^2 + c^2\right)^2 .$$

Analogously,

$$4a^2c^2\cos^2(\beta) = \left(a^2 - b^2 + c^2\right)^2 .$$

We therefore obtain

$$\begin{aligned}\frac{\sin^2(\alpha)}{\sin^2(\beta)} &= \frac{4a^2b^2c^2\left(1-\cos^2(\alpha)\right)}{4a^2b^2c^2\left(1-\cos^2(\beta)\right)} \\ &= \frac{4a^2b^2c^2 - a^2\left(-a^2+b^2+c^2\right)^2}{4a^2b^2c^2 - b^2\left(a^2-b^2+c^2\right)^2} \\ &= \frac{a^2}{b^2}\cdot\frac{4b^2c^2 - \left(a^4+b^4+c^4-2a^2b^2-2a^2c^2+2b^2c^2\right)}{4a^2c^2 - \left(a^4+b^4+c^4-2a^2b^2+2a^2c^2-2b^2c^2\right)} \\ &= \frac{a^2}{b^2}.\end{aligned}$$

□

Theorem 1.2.7 (angle sum in the Euclidean triangle) *The sum of the interior angles in the Euclidean triangle satisfies*

$$\alpha + \beta + \gamma = \pi .$$

Proof (a) We first prove the statement for right-angle triangles. Let $\gamma = \pi/2$. Then

$$\sin(\alpha + \beta + \pi/2) = \cos(\alpha + \beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta). \tag{1.1}$$

Using the cosine rule 1.2.4 and Pythagoras's theorem we obtain

$$\begin{aligned}\cos(\alpha)\cos(\beta) &= \frac{-a^2+b^2+c^2}{2bc}\cdot\frac{a^2-b^2+c^2}{2ac}\\ &= \frac{2b^2\cdot 2a^2}{4abc^2}\\ &= \frac{ab}{c^2}.\end{aligned} \tag{1.2}$$

The sine rule 1.2.6 together with the cosine rule and Pythagoras's theorem gives

$$\begin{aligned}\sin(\alpha)\sin(\beta) &= \frac{b}{a}\sin^2(\alpha)\\ &= \frac{b}{a}\cdot\Big(1-\cos^2(\alpha)\Big)\\ &= \frac{b}{a}\cdot\left(1-\frac{\left(-a^2+b^2+c^2\right)^2}{4b^2c^2}\right)\\ &= \frac{b}{a}\cdot\left(1-\frac{4b^4}{4b^2c^2}\right)\\ &= \frac{b}{a}\cdot\frac{c^2-b^2}{c^2}\\ &= \frac{ab}{c^2}.\end{aligned} \tag{1.3}$$

Substituting (1.2) and (1.3) into (1.1), we obtain $\sin(\alpha+\beta+\pi/2)=0$, i.e. $\alpha+\beta+\pi/2$ is an integral multiple of π,

$$\alpha+\beta+\pi/2 = k\cdot\pi,$$

$k \in \mathbb{Z}$. As all angles in a right-angle triangle are > 0 and $\leq \pi/2$, we have $\pi/2 < \alpha+\beta+\pi/2 < 3\pi/2$ and thus $k=1$. It follows that $\alpha+\beta+\pi/2=\pi$, which proves the claim for right-angle triangles.

(b) In a general triangle let a be the longest side and r the vertex opposite to it. We drop a perpendicular from r to a and thus divide the triangle into two right-angle triangles.

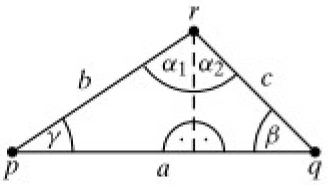

We know from part (a) of this proof that in right-angle triangles $\alpha_1 + \gamma + \pi/2 = \pi$ and $\alpha_2 + \beta + \pi/2 = \pi$. Addition of those two equations gives $\alpha + \beta + \gamma = \pi$. □

Exercise 1.17 Prove the *half-angle theorem of Euclidean geometry*:

$$\tan^2\left(\frac{\alpha}{2}\right) = \frac{(a-b+c)(a+b-c)}{(a+b+c)(-a+b+c)}.$$

Hint Check and use the formula $\tan^2(t) = (1-\cos(2t))/(1+\cos(2t))$.

For comparison, we conclude this section by looking at another important classical geometry, i.e. ***spherical geometry***. The points of this geometry are not in the plane, but on the surface of a sphere. We set

$$\mathscr{P} := S^2 := \{(x,y,z)^\top \in \mathbb{R}^3 \mid x^2+y^2+z^2=1\}.$$

We call S^2 the ***two-dimensional sphere***.

This raises the question of what the "straight lines" are in this geometry. The straight lines in the plane can be characterised by the fact that each of them is the shortest curve connecting any two of its points. On the sphere the great circles have this property. A ***great circle*** is the subset of S^2 in which a two-dimensional vector subspace of $\mathbb{R}^3$, i.e. a plane that passes through 0, intersects the surface of the sphere.

We therefore set

$$\mathscr{G} := \{S^2 \cap E \mid E \text{ is a two-dimensional vector subspace of } \mathbb{R}^3\}.$$

The incidence relation $\in$ must here be interpreted in the set-theoretic sense again.

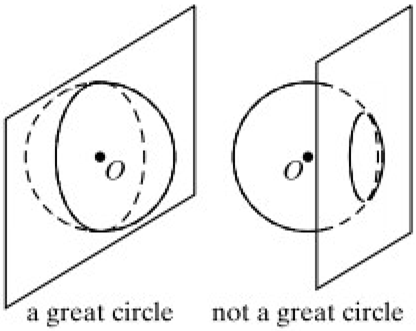

a great circle not a great circle

Exercise 1.18 Which of the axioms I_1–I_4 are valid?

Given three points on a great circle, it is not possible to say in a meaningful way which of them lies between the two others. Let us try with the following definition: if $p, q, r \in S^2$, then q lies ***between*** p and r if and only if p, q and r are three pairwise distinct points on a great circle.

Exercise 1.19 What is a line segment? Which of the axioms A_1–A_5 are valid?

The congruence relations on the other hand can easily be defined in a geometrically meaningful way. We simply exchange the Euclidean motion group E(2) for the orthogonal group O(3). One needs to pay attention to the fact that for an orthogonal matrix $A \in O(3)$ and $p \in S^2$ the point Ap should lie on S^2 again, i.e. A indeed maps S^2 to itself only.

Exercise 1.20 Show that axioms K_1–K_6 are valid.

There are no parallels since any two great circles intersect. Theorem 1.1.9 therefore does not hold. This is one of the ways to see that not all incidence and ordering axioms can hold.

Because of its importance in navigation, amony other things, spherical geometry has been studied for a very long time. We will investigate it in more detail in sections 4.9 and 4.10.

5- N2 3

Conclusion

$$N_1(p) = N_2(p) \quad \Leftrightarrow \quad \det(D_{u_1}\varphi) > 0$$
$$N_1(p) = -N_2(p) \quad \Leftrightarrow \quad \det(D_{u_1}\varphi) < 0.$$

The following theorem follows immediately.

Theorem 3.4.7 *A regular surface $S \subset \mathbb{R}^3$ is orientable if and only if S can be covered by local parametrisations such that for all parameter transformations φ*

$$\det(D\varphi) > 0.$$

3.5 The second fundamental form

Let $S \subset \mathbb{R}^3$ be an orientable regular surface with smooth unit normal field N. Regarded as a map between surfaces, i.e. $N : S \to S^2$, the map N is also called the ***Gauss map***.

Let $p \in S$. We consider the differential of N at p:

$$d_pN : T_pS \to T_{N(p)}S^2.$$

Now $T_{N(p)}S^2 = N(p)^\perp = T_pS$. Hence d_pN is an endomorphism on T_pS.

Definition 3.5.1 Let $S \subset \mathbb{R}^3$ be a regular surface with orientation given by the unit normal field N. The endomorphism

$$W_p : T_pS \to T_pS,$$

$$W_p(X) = -d_pN(X),$$

is called the ***Weingarten map***.

The negative sign appears for historic reasons. If orientation is reversed, i.e. if $-N$ is substituted for N, then W also changes its sign.

Example 3.5.2 Let $S = S^2$, and let N be the outer unit normal field, $N(p) = p$. Then

$$W_p = -\operatorname{Id} : T_pS^2 \to T_pS^2.$$

Example 3.5.3 Let $S = \{(x,y,0)^\top \mid x,y \in \mathbb{R}\}$ be the x–y plane, $N(x,y,z) = (0,0,1)^\top$. Then N is constant and thus $W_p = 0$ for all $p \in S$.

Example 3.5.4 Let $S = S^1 \times \mathbb{R}$ be the cylinder, $N(x,y,z) = (x,y,0)^\top$.

At a point $p = (x,y,z)^\top \in S$ the tangent plane T_pS is spanned by the basis vectors $(-y,x,0)^\top$ and $(0,0,1)^\top$. We calculate

$$W_p\begin{pmatrix}0\\0\\1\end{pmatrix} = -d_pN\begin{pmatrix}0\\0\\1\end{pmatrix} = -\frac{d}{dt}N\left.\begin{pmatrix}x\\y\\z+t\end{pmatrix}\right|_{t=0}$$
$$= -\frac{d}{dt}\left.\begin{pmatrix}x\\y\\0\end{pmatrix}\right|_{t=0} = \begin{pmatrix}0\\0\\0\end{pmatrix}.$$

To find the image of $(-y, x, 0)$ under W_p, we choose $t_0 \in \mathbb{R}$, such that $(\cos(t_0), \sin(t_0)) = (x, y)$. Then for $c(t) := (\cos(t + t_0), \sin(t + t_0), z)^\top$ we have that $c(0) = (x, y, z)^\top = p$ and $\dot{c}(0) = (-\sin(t_0), \cos(t_0), 0)^\top = (-y, x, 0)^\top$. It follows that

$$W_p\begin{pmatrix}-y\\x\\0\end{pmatrix} = -d_pN\begin{pmatrix}-y\\x\\0\end{pmatrix} = -\frac{d}{dt}N\left.\begin{pmatrix}\cos(t+t_0)\\\sin(t+t_0)\\z\end{pmatrix}\right|_{t=0}$$
$$= -\frac{d}{dt}\left.\begin{pmatrix}\cos(t+t_0)\\\sin(t+t_0)\\0\end{pmatrix}\right|_{t=0} = \begin{pmatrix}\sin(t_0)\\-\cos(t_0)\\0\end{pmatrix} = -\begin{pmatrix}-y\\x\\0\end{pmatrix}.$$

With respect to the basis vectors $(-y, x, 0)^\top$ and $(0, 0, 1)^\top$, the map W_p therefore has the matrix representation

$$\begin{pmatrix}-1 & 0\\ 0 & 0\end{pmatrix}.$$

Proposition 3.5.5 *Let $S \subset \mathbb{R}^3$ be an orientable regular surface with Weingarten map $W_p : T_pS \to T_pS$, $p \in S$. Then W_p is self-adjoint with respect to the first fundamental form.*

Proof Let N be the unit normal field of S that induces the Weingarten map, $W_p = -d_pN$. We choose a local parametrisation (U, F, V) at p and set $u := F^{-1}(p)$.

Let

$$X_1 := D_uF(e_1) = \frac{\partial F}{\partial u^1}(u)$$

and

$$X_2 := D_uF(e_2) = \frac{\partial F}{\partial u^2}(u)$$

be the corresponding basis vectors of T_pS. As N is perpendicular to S everywhere, we have

$$\left\langle \frac{\partial F}{\partial u^i}(u+te_j),\ N(F(u+te_j)) \right\rangle \equiv 0.$$

Differentiating this equation with respect to t gives

$$\begin{aligned} 0 &= \frac{d}{dt} \left\langle \frac{\partial F}{\partial u^i}(u+te_j), N(F(u+te_j)) \right\rangle \bigg|_{t=0} \\ &= \left\langle \frac{d}{dt} \frac{\partial F}{\partial u^i}(u+te_j) \bigg|_{t=0}, N(p) \right\rangle + \left\langle \frac{\partial F}{\partial u^i}(u), d_pN \circ D_uF(e_j) \right\rangle \\ &= \left\langle \frac{\partial^2 F}{\partial u^j \partial u^i}(u), N(p) \right\rangle + \langle X_i, -W_p(X_j) \rangle . \end{aligned}$$

Thus

$$I_p(X_i, W_p(X_j)) = \langle X_i, W_p(X_j) \rangle = \left\langle \frac{\partial^2 F}{\partial u^j \partial u^i}(u), N(p) \right\rangle. \tag{3.3}$$

By the theorem of Schwarz [18, p. 372, theorem 1.1] the two partial derivatives of F can be exchanged and we obtain

$$\begin{aligned} I_p(X_i, W_p(X_j)) &= \left\langle \frac{\partial^2 F}{\partial u^j \partial u^i}(u), N(p) \right\rangle \\ &= \left\langle \frac{\partial^2 F}{\partial u^i \partial u^j}(u), N(p) \right\rangle \\ &= I_p(X_j, W_p(X_i)). \end{aligned}$$

We now know for our basis vectors X_1 and X_2 of T_pS that

$$I_p(X_i, W_p(X_j)) = I_p(X_j, W_p(X_i)) = I_p(W_p(X_i), X_j).$$

Since any two vectors $X, Y \in T_pS$ can be written as a linear combination of X_1 and X_2, it immediately follows from the bilinearity of I and the linearity of W_p that

$$I_p(X, W_p(Y)) = I_p(W_p(X), Y),$$

i.e. W_p is self-adjoint with respect to I. □

Let us recall from linear algebra that if V is a finite-dimensional real vector space with Euclidean scalar product $\langle \cdot, \cdot \rangle$, then the self-adjoint endomorphisms

W on V are uniquely associated with the symmetric bilinear forms β on V. The relation between W and β is

$$\beta(X,Y) = \langle W(X), Y\rangle, \qquad X, Y \in V.$$

Definition 3.5.6 The bilinear form that corresponds to the Weingarten map W_p is called the ***second fundamental form*** (of the surface S at the point p):

$$II_p(X,Y) = I_p(W_p(X), Y), \qquad X, Y \in T_pS.$$

For simplicity the root point vector p is often omitted in the notation, and we write II instead of II_p, and W instead of W_p.

Expression in local coordinates Let $S \subset \mathbb{R}^3$ be a regular surface, $p \in S$. Let (U, F, V) be a local parametrisation of S at p. We set $u := F^{-1}(p)$. We have already learnt how to express the first fundamental form locally. For the basis $D_uF(e_1) = (\partial F/\partial u^1)(u)$ and $D_uF(e_2) = (\partial F/\partial u^2)(u)$ the first fundamental form is given by the symmetric matrix $(g_{ij}(u))_{i,j=1,2}$, where

$$\begin{aligned} g_{ij}(u) &= \left\langle \frac{\partial F}{\partial u^i}(u), \frac{\partial F}{\partial u^j}(u)\right\rangle \\ &= I_p\left(D_uF(e_i), D_uF(e_j)\right). \end{aligned}$$

We now define

$$\begin{aligned} h_{ij}(u) &:= II_p(D_uF(e_i), D_uF(e_j)) \\ &= I_p\left(W_p(D_uF(e_i)), D_uF(e_j)\right) \\ &\overset{(3.3)}{=} \left\langle \frac{\partial^2 F}{\partial u^j \partial u^i}(u), N(p)\right\rangle, \; i,j = 1,2. \end{aligned} \tag{3.4}$$

Then $(h_{ij}(u))_{i,j=1,2}$ is the symmetric matrix which gives the second fundamental form w.r.t. the basis given above. The entries of the matrix representing the Weingarten map are denoted as w_i^j, i.e. we define

$$W_p(D_uF(e_i)) =: \sum_{j=1}^{2} w_i^j(u) D_uF(e_j).$$

As the Weingarten map is closely related to the second fundamental form, we expect that the matrices $(h_{ij}(u))_{i,j}$ and $\left(w_i^j(u)\right)_{i,j}$ can be determined from each other. This works as follows:

$$\begin{aligned}
h_{ij}(u) &= II(D_uF(e_i), D_uF(e_j)) \\
&= I(W(D_uF(e_i)), D_uF(e_j)) \\
&= I\Big(\sum_{k=1}^{2} w_i^k(u) D_uF(e_k), D_uF(e_j)\Big) \\
&= \sum_{k=1}^{2} w_i^k(u) I\left(D_uF(e_k), D_uF(e_j)\right) \\
&= \sum_{k=1}^{2} w_i^k(u) g_{kj}(u).
\end{aligned}$$

The matrix $(h_{ij}(u))_{i,j}$ therefore results from the matrix multiplication of the matrices $\left(w_i^k(u)\right)_{i,k}$ and $(g_{kj}(u))_{k,j}$. As the matrix $(g_{kj}(u))_{k,j}$ is positive definite, and thus in particular is invertible, the equation can be solved for $\left(w_i^k(u)\right)_{i,k}$. For this purpose let $(g^{ij}(u))_{i,j}$ be the inverse matrix of $(g_{ij}(u))_{i,j}$, i.e.

$$\left(g^{ij}(u)\right)_{ij} = \frac{1}{g_{11}(u)g_{22}(u) - g_{12}(u)^2} \begin{pmatrix} g_{22}(u) & -g_{12}(u) \\ -g_{21}(u) & g_{11}(u) \end{pmatrix}.$$

It then follows that

$$\sum_{k=1}^{2} h_{ik}(u) g^{kj}(u) = w_i^j(u). \tag{3.5}$$

3.6 Curvature

We now come to a central notion of surface theory and of differential geometry: ***curvature***. We will meet several concepts of curvature, and we begin with normal curvature.

For this purpose let $S \subset \mathbb{R}^3$ be an orientable regular surface with smooth unit normal field N, $p \in S$. Let $c : (-\varepsilon, \varepsilon) \to S$ be a curve parametrised by arclength with $c(0) = p$. Regarded as a space curve in $\mathbb{R}^3$, the curve c has at point 0 curvature $\kappa(0)$, which in the case $\kappa(0) \neq 0$ is given by

$$\ddot{c}(0) = \kappa(0) \cdot n(0),$$

where n is the normal vector of c. We now want to split this curvature into a part which results from the fact that c curves *within* S, and a part which reflects the curvature of S in $\mathbb{R}^3$. For this purpose we decompose $n(0)$ into a part that is tangential to S and one that is perpendicular to S:

$$n(0) = n(0)^{\text{tang}} + n(0)^{\text{perp}},$$

where $n(0)^{\text{perp}} = \langle n(0), N(p)\rangle\, N(p)$. We therefore have

$$\ddot{c}(0) = \kappa(0) \cdot n(0) = \kappa(0) \cdot n(0)^{\text{tang}} + \kappa(0) \cdot \langle n(0), N(p)\rangle\, N(p).$$

The tangential part, which gives the extent to which c curves within S, leads to the ***geodesic curvature*** of c in S, to which we will come back in section 4.5. For now we are interested in the curvature of S in $\mathbb{R}^3$, and for this reason make the definition

$$\kappa_{\text{nor}} := \langle \ddot{c}(0), N(p)\rangle = \begin{cases} \kappa(0) \cdot \langle n(0), N(p)\rangle, & \text{if } \kappa(0) \neq 0, \\ 0, & \text{if } \kappa(0) = 0. \end{cases}$$

We call κ_{nor} the ***normal curvature*** of S at the point p in direction $\dot{c}(0)$. If, in the case of $\kappa(0) \neq 0$, θ denotes the angle between $N(p)$ and $n(0)$, then we therefore have

$$\kappa_{\text{nor}} = \kappa(0) \cdot \cos(\theta).$$

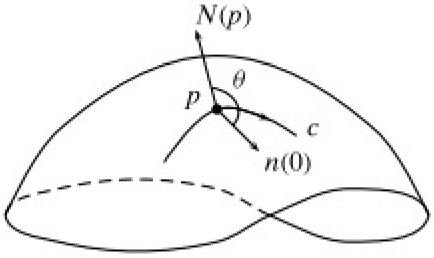

In particular, we always have

$$|\kappa_{\text{nor}}| \leq \kappa(0).$$

The following theorem tells us how to find the normal curvature from the second fundamental form.

Theorem 3.6.1 (Meusnier's theorem) *Let $S \subset \mathbb{R}^3$ be an orientable regular surface with unit normal field N and second fundamental form II. Let $p \in S$. Let $c : (-\varepsilon, \varepsilon) \to S$ be a curve parametrised by arc-length with $c(0) = p$. Then we have for the normal curvature κ_{nor} of c:*

$$\kappa_{\text{nor}} = II(\dot{c}(0), \dot{c}(0)).$$

In particular, all curves parametrised by arc-length in S through p with the same tangent vector have the same normal curvature.

This also justifies that we refer to κ_{nor} as the normal curvature of S at the point p in direction $\dot{c}(0)$, since κ_{nor} depends, apart from on S and p, only on $\dot{c}(0)$, but not on the particular choice of the curve c.

Proof As c lies on S, we have

$$\langle N(c(t)), \dot{c}(t)\rangle = 0$$

for all $t \in (-\varepsilon, \varepsilon)$. Differentiating this equation gives

$$\begin{aligned} 0 &= \frac{d}{dt}\langle N(c(t)), \dot{c}(t)\rangle\Big|_{t=0} \\ &= \left\langle \frac{d}{dt} N(c(t))\Big|_{t=0}, \dot{c}(0)\right\rangle + \langle N(p), \ddot{c}(0)\rangle \\ &= \langle d_pN(\dot{c}(0)), \dot{c}(0)\rangle + \kappa_{\text{nor}} \\ &= \langle -W_p(\dot{c}(0)), \dot{c}(0)\rangle + \kappa_{\text{nor}} \\ &= -II(\dot{c}(0), \dot{c}(0)) + \kappa_{\text{nor}}. \end{aligned}$$

□

Remark A change of orientation of the curve c does not change the value of the normal curvature κ_{nor}, since

$$II(-\dot{c}(0), -\dot{c}(0)) = II(\dot{c}(0), \dot{c}(0)).$$

However, if the orientation of the surface S is reversed with $N(p)$ being replaced by $-N(p)$, curvature κ_{nor} changes sign:

$$\langle \ddot{c}(0), -N(p)\rangle = -\langle \ddot{c}(0), N(p)\rangle.$$

Let $S \subset \mathbb{R}^3$ be an orientable regular surface with unit normal field N, let $p \in S$. Let $X \in T_pS$ be a tangent vector of length 1.

Exercise 3.9 Let E be the plane spanned by $N(p)$ and X. Use the implicit mapping theorem to show that for a neighbourhood V of p in $\mathbb{R}^3$ the set

$S \cap (E + p) \cap V$ can be parametrised by a regular curve. Here $E + p$ denotes the plane E translated to the point p.

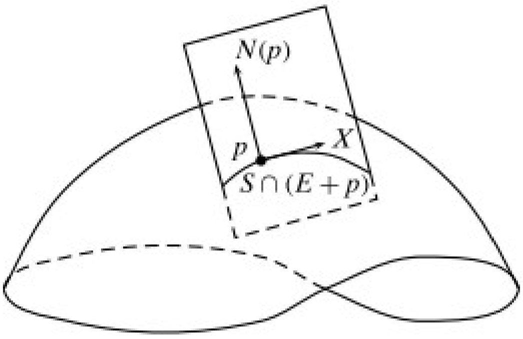

According to theorem 3.6.1 the normal curvature $II(X, X)$ can be found from the unit speed curve c which describes $S \cap (E + p) \cap V$. If we consider c as a plane curve in the plane $E + p \cong \mathbb{R}^2$, then the normal vector of the curve $n(0) = \pm N(p)$ and hence the normal curvature is

$$II(X, X) = \pm\kappa(0),$$

where κ is the curvature of c considered as a plane curve. This explains the term *normal* curvature. The normal curvature is the curvature of the plane curve $S \cap (E + p) \cap V$ in the plane E spanned by X and the normal vector.

Example 3.6.2 Let $S = S^1 \times \mathbb{R}$ be the cylinder. Let $p \in S$. The intersection of S with the normal plane at the point p is a circle, an ellipse or a straight line. The normal curvature therefore varies between 1 and 0, depending on the direction.

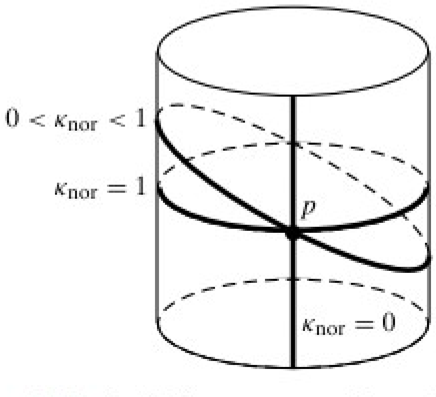
$0 < \kappa_{\text{nor}} < 1$
$\kappa_{\text{nor}} = 1$
p
$\kappa_{\text{nor}} = 0$

By proposition 3.5.5 the Weingarten map $W_p : T_pS \to T_pS$ is always self-adjoint. We can therefore find an orthonormal basis X_1, X_2 of T_pS, which

5- 6/2 /<3 *) %)
5- 672 N 3) ; D -))4 E) *) %)

By (3.4) the second fundamental form is

$$\begin{aligned}
h_{11}(t,\varphi) &= \left\langle \frac{\partial^2 F}{\partial t^2}(t,\varphi), N(t,\varphi) \right\rangle \\
&= \left(\kappa(t)(1 - r\cos(\varphi)\kappa(t)) - r\sin(\varphi)\dot{\tau}(t) - r\cos(\varphi)\tau(t)^2\right)(-\cos(\varphi)) \\
&\quad + (-\sin(\varphi)\tau(t)^2 + \cos(\varphi)\dot{\tau}(t))(-\sin(\varphi)) \\
&= r\tau(t)^2 - \cos(\varphi)\kappa(t) + r\cos(\varphi)^2\kappa(t)^2, \\
h_{12}(t,\varphi) &= r\cos(\varphi)^2\tau(t) + r\sin(\varphi)^2\tau(t) \\
&= r\tau(t), \\
h_{22}(t,\varphi) &= r.
\end{aligned}$$

From (3.5) it follows that the matrix of the Weingarten map is

$$\begin{aligned}
\left(w_i^j(t,\varphi)\right)_{ij} &= (h_{ik}(t,\varphi))(g^{kj}(t,\varphi)) \\
&= \begin{pmatrix} r\tau(t)^2 - \cos(\varphi)\kappa(t) + r\cos(\varphi)^2\kappa(t)^2 & r\tau(t) \\ r\tau(t) & r \end{pmatrix} \cdot \\
&\quad \cdot \begin{pmatrix} \dfrac{1}{(1 - r\cos(\varphi)\kappa(t))^2} & -\dfrac{\tau(t)}{(1 - r\cos(\varphi)\kappa(t))^2} \\ -\dfrac{\tau(t)}{(1 - r\cos(\varphi)\kappa(t))^2} & \dfrac{1}{r^2} + \dfrac{\tau(t)^2}{(1 - r\cos(\varphi)\kappa(t))^2} \end{pmatrix} \\
&= \begin{pmatrix} -\dfrac{\kappa(t)\cos(\varphi)}{1 - r\cos(\varphi)\kappa(t)} & * \\ 0 & \dfrac{1}{r} \end{pmatrix}.
\end{aligned}$$

As this matrix is triangular, we can read off the eigenvalues and obtain the principal curvatures:

$$\begin{aligned}
\kappa_1 &= \frac{-\kappa(t)\cos(\varphi)}{1 - r\cos(\varphi)\kappa(t)}, \\
\kappa_2 &= \frac{1}{r}.
\end{aligned}$$

It follows that the Gauss curvature is

$$K = -\frac{1}{r}\frac{\kappa(t)\cos(\varphi)}{1 - r\cos(\varphi)\kappa(t)},$$

and the mean curvature is

$$H = \frac{1 - 2r\cos(\varphi)\kappa(t)}{2r(1 - r\cos(\varphi)\kappa(t))}.$$

Example 3.8.18 We consider the ***torus***, which is defined as the tubular surface around a circular line

$$c(t) = (\cos(t), \sin(t), 0)^\top$$

with thickness $2r < 2$ (plate 7). We then have

$$\kappa(t) \equiv 1, \quad \tau(t) \equiv 0$$

and thus

$$K(t,\varphi) = -\frac{1}{r}\frac{\cos(\varphi)}{1 - r\cos(\varphi)}, \quad H(t,\varphi) = \frac{\cos(\varphi) - r/2}{1 - r\cos(\varphi)}.$$

Exercise 3.32 Let c be a closed space curve parametrised by arc-length. Show that for every tubular surface S around c we have

$$\int_S K\,dA = 0.$$

We will understand this fact better against the background of the Gauss–Bonnet theorem.

Exercise 3.33 Let $c : I \to \mathbb{R}^3$ be a closed space curve. Using the results of this section and exercise 3.16 give a new simple proof of the total curvature estimate in Fenchel's theorem 2.3.19:

$$\kappa(c) \geq 2\pi.$$

5- 6<2 N73 6 # - +)

D) +)' 4 - +
) 4 P) P) %)) %) %
) # #) @- - 2?) ')9 A3 ;4
*) %) %)) *) %)
9 - - +) *) ( )
-) P)) # 4 9)+
4)) 4 -) ( )
;4))44 - E)+ 4)) +) -) 9)
))) % - +8 ' ( ))) 8
- 4))) .) S) 9 P)
- 4) *) %) 4)) +4 9
- +) % -)) # - + %)
)44)) -)4 +) # +4 9 4)) P)
); @) - + ; 4 4)));
6 E
D) (
4)+ 4))
) % - 8) 9 -
4) D =) ') ,) +)-
4) 2) 3)9)) 8)
;) 4 4))) 9 9 % 9+) 9 -
) 4 ') .) 8 % 8 +
))) 4) % + &)
9) +) 9 - %) +
))) 4) E ') 4 4)) + 8
9 4 9)') 9+))
) &+) 8 9) % +) + - 9) 4
) 4)) 4) 4)
D))+ 4 + =)) 9 9 %
9+))9) 4)

5- N 2 7T3 6 ( ) %)) #) @- -
E)4 %) %) 8) 4 4)
%) 8) %)) *) %) E
) =) +8 ( ) %) D 9 - 9+
' -) %)) %) % # %) %))
%) % 4) 8
) F)9) -)9 +) %) %
%)) %) % R # ( ) %)) ;) +
))) - - %) %

)- NT2 TN3

Table 4.1 *Quantities that depend on the inner geometry only*

geometric quantity	symbol	expression w.r.t. local param.
first fundamental form	I	g_{ij}
surface element	dA	$dA = \sqrt{g_{11}g_{22} - (g_{12})^2}du^1du^2$
covariant derivative	∇	$\Gamma^k_{ij} = \frac{1}{2}\sum_m \left(\frac{\partial g_{jm}}{\partial u^i} + \frac{\partial g_{im}}{\partial u^j} - \frac{\partial g_{ij}}{\partial u^m}\right) g^{mk}$
Riemann curvature tensor	R	$R^\ell_{ijk} = \frac{\partial \Gamma^\ell_{kj}}{\partial u^i} - \frac{\partial \Gamma^\ell_{ki}}{\partial u^j} + \sum_m \left(\Gamma^\ell_{mi}\Gamma^m_{kj} - \Gamma^\ell_{mj}\Gamma^m_{ki}\right)$
Gauss curvature	K	$K = \frac{1}{2}\sum_{ijk} g^{jk}R^i_{ijk}$

Table 4.2 *Quantities that are not invariant under local isometries*

geometric quantity	symbol	expression w.r.t. local param.
second fundamental form	II	h_{ij}
Weingarten map	W	$w_i^j = \sum_k h_{ik}g^{kj}$
principal curvatures	κ_i	
mean curvature	H	$H = \frac{\kappa_1 + \kappa_2}{2} = \frac{w_1^1 + w_2^2}{2}$

The most important geometric quantities of a regular surface are summarised in tables 4.1 and 4.2.

4.4 Riemannian metrics

Since we will as of now just look at quantities that depend on the inner geometry only, we will at this point generalise the concept of the first fundamental form. What actually is the first fundamental form? It assigns a Euclidean scalar product to each tangent plane. That this scalar product is the restriction of the standard scalar product on $\mathbb{R}^3$ is of no importance for quantities that depend on the inner geometry only. We therefore make the following definition.

Definition 4.4.1 Let $S \subset \mathbb{R}^3$ be a regular surface. A ***Riemannian metric*** g on S assigns a Euclidean scalar product g_p on the tangent plane T_pS to each point $p \in S$, such that for every local parametrisation (U, F, V) of S the functions

$$g_{ij} : U \to \mathbb{R},$$
$$g_{ij}(u) := g_{F(u)}\left(\frac{\partial F}{\partial u^i}(u), \frac{\partial F}{\partial u^j}(u)\right)$$

are smooth.

The first fundamental form is, of course, an example of a Riemannian metric. But there are important examples of Riemannian metrics that are not the first fundamental form, as we will see later. One of them is the hyperbolic plane. The condition that the functions g_{ij} must be smooth ensures the scalar product g_p does not "wildly" depend on p, but is differentiable w.r.t. p. In the case of the first fundamental form this holds automatically.

All quantities that depend on the inner geometry only, such as the surface element, the covariant derivative, the Gauss curvature and the Riemann curvature tensor, are also defined for regular surfaces with a Riemannian metric that is not the first fundamental form.

We therefore *define*, for example, the Christoffel symbols via the formula from lemma 4.2.14 as

$$\Gamma_{ij}^k := \frac{1}{2}\sum_{m=1}^{2}\left(\frac{\partial g_{jm}}{\partial u^i} + \frac{\partial g_{im}}{\partial u^j} - \frac{\partial g_{ij}}{\partial u^m}\right) g^{mk} \tag{4.7}$$

and then the covariant derivative for vector fields

$$v = \sum_i v^i \frac{\partial F}{\partial u^i}, w = \sum_j w^j \frac{\partial F}{\partial u^j}$$

via

$$\nabla_w v(F(u)) := \sum_k \left(d_u v^k \begin{pmatrix} w^1(u) \\ w^2(u) \end{pmatrix} + \sum_{ij} \Gamma_{ij}^k(u) v^i(u) w^j(u) \right) \frac{\partial F}{\partial u^k}(u).$$

Of course, it needs to be checked that this definition does not depend on the local parametrisation chosen; this is left as an exercise for the reader. In the case of the first fundamental form this formula holds automatically. We can analogously translate the other quantities that depend on the inner geometry, such as the Riemann curvature tensor and the Gauss curvature. We see that it was good to have required the smoothness of the g_{ij}. For the definition of the covariant derivative we need the first derivatives of the g_{ij}, for the curvature we also need the second derivatives.

Exercise 4.13 Show that lemma 4.2.12 and lemma 4.2.17 also hold for general Riemannian metrics.

Exercise 4.14 Show that for any two smooth vector fields v and w on a regular surface equipped with a Riemannian metric one has

$$\nabla_v w - \nabla_w v = [v, w].$$

5- TN2 63

$x \in \mathbb{R}^2$. The geodesic equation (4.11) then becomes

$$\sum_{ij} \Gamma_{ij}^k(tx)x^i x^j = 0, \quad k = 1, 2.$$

For $t = 0$ in particular we therefore have

$$\sum_{ij} \Gamma_{ij}^k(0,0)x^i x^j = 0, \quad k = 1, 2,$$

for all x. As $\Gamma_{ij}^k(0,0)$ is symmetric in i and j for fixed k, we can polarise and obtain

$$\Gamma_{ij}^k(0,0) = 0$$

for all i, j, k. From (4.3)

$$\frac{\partial g_{jm}}{\partial u^i} = \sum_k \left(\Gamma_{ij}^k g_{km} + \Gamma_{im}^k g_{kj} \right),$$

it follows that

$$\frac{\partial g_{ij}}{\partial u^k}(0,0) = 0$$

for all i, j, k. ☐

Example 4.6.8 As above, let S be an arbitrary regular surface with a Riemannian metric, let $p \in S$ and let X_1, X_2 be an orthonormal basis of T_pS. This time we take polar coordinates for T_pS, $F_1(r, \varphi) = r \cdot (\cos(\varphi)X_1 + \sin(\varphi)X_2)$. The corresponding local parametrisation of S,

$$F(r, \varphi) = \exp_p \left(r \cdot (\cos(\varphi)X_1 + \sin(\varphi)X_2) \right),$$

is the parametrisation in ***geodesic polar coordinates*** (at the point p).

Theorem 4.6.9 (Gauss's lemma) *Let S be a regular surface with Riemannian metric g. Let $p \in S$ and let F be a local parametrisation in geodesic coordinates (r, φ).*

Then the Riemannian metric has the following form with respect to the local parametrisation:

$$(g_{ij}(r, \varphi))_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & f(r, \varphi)^2 \end{pmatrix},$$

with a positive function f that satisfies

$$\lim_{r \to 0} f(r, \varphi) = 0, \qquad \lim_{r \to 0} \frac{\partial f}{\partial r}(r, \varphi) = 1.$$

Proof Let X_1, X_2 be the orthonormal basis of T_pS that we have used for the definition of the geodesic polar coordinates, $F(r,\varphi) = \exp_p(\tilde{F}(r,\varphi))$, where $\tilde{F}(r,\varphi) = r\cdot(\cos(\varphi)X_1 + \sin(\varphi)X_2)$. The curve $c(r) = F(r,\varphi_0)$ is for fixed $\varphi = \varphi_0$ by the definition of the exponential map the geodesic with $\dot{c}(0) = \cos(\varphi_0)X_1 + \sin(\varphi_0)X_2$. Geodesics are parametrised proportional to arc-length and $\dot{c}(0)$ is a unit vector, hence

$$g_{11}(r,\varphi_0) = g(\dot{c}(r),\dot{c}(r)) = 1.$$

Further

$$\begin{aligned}
\frac{\partial g_{12}}{\partial r}(r,\varphi_0) &= \frac{\partial}{\partial r} g\left(\frac{\partial F}{\partial r},\frac{\partial F}{\partial \varphi}\right)\Big|_{(r,\varphi_0)} \\
&= g\left(\frac{\nabla}{\partial r}\frac{\partial F}{\partial r}(r,\varphi_0),\frac{\partial F}{\partial \varphi}(r,\varphi_0)\right) + g\left(\frac{\partial F}{\partial r}(r,\varphi_0),\frac{\nabla}{\partial r}\frac{\partial F}{\partial \varphi}(r,\varphi_0)\right) \\
&= g\left(\frac{\nabla}{dr}\dot{c}(r),\frac{\partial F}{\partial \varphi}(r,\varphi_0)\right) + g\left(\frac{\partial F}{\partial r}(r,\varphi_0),\frac{\nabla}{\partial r}\frac{\partial F}{\partial \varphi}(r,\varphi_0)\right) \\
&= 0 + g\left(\frac{\partial F}{\partial r}(r,\varphi_0),\frac{\nabla}{\partial \varphi}\frac{\partial F}{\partial r}(r,\varphi_0)\right) \\
&= \frac{1}{2}\,\frac{\partial}{\partial \varphi}\underbrace{g\left(\frac{\partial F}{\partial r},\frac{\partial F}{\partial r}\right)}_{=1}\Bigg|_{(r,\varphi_0)} \\
&= 0. \qquad (4.13)
\end{aligned}$$

Hence the function g_{12} is constant in r for fixed $\varphi = \varphi_0$. As an abbreviation we set $Y_1 := \cos(\varphi_0)X_1 + \sin(\varphi_0)X_2$ and $Y_2 := -\sin(\varphi_0)X_1 + \cos(\varphi_0)X_2$. This is the orthonormal basis of T_pS that results from X_1, X_2 in a rotation through the angle φ_0. We have

$$\frac{\partial \tilde{F}}{\partial r}(r,\varphi_0) = Y_1, \qquad \frac{\partial \tilde{F}}{\partial \varphi}(r,\varphi_0) = r\cdot Y_2.$$

We calculate

$$\begin{aligned}
\lim_{r\to 0} g_{12}(r,\varphi_0) &= \lim_{r\to 0} g\left(\frac{\partial F}{\partial r}(r,\varphi_0),\frac{\partial F}{\partial \varphi}(r,\varphi_0)\right) \\
&= \lim_{r\to 0} g\left(d_{\tilde{F}(r,\varphi_0)}\exp_p\left(\frac{\partial \tilde{F}}{\partial r}(r,\varphi_0)\right), d_{\tilde{F}(r,\varphi_0)}\exp_p\left(\frac{\partial \tilde{F}}{\partial \varphi}(r,\varphi_0)\right)\right)
\end{aligned}$$

$$\begin{aligned}
&= \lim_{r\to 0} g\left(d_{\tilde{F}(r,\varphi_0)} \exp_p(Y_1), d_{\tilde{F}(r,\varphi_0)} \exp_p(r \cdot Y_2)\right)\\
&= g\left(d_0 \exp_p(Y_1), d_0 \exp_p(0)\right)\\
&= 0. \qquad (4.14)
\end{aligned}$$

From (4.13) and (4.14) it follows that

$$g_{12} = g_{21} = 0.$$

Hence

$$(g_{ij})_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & g_{22} \end{pmatrix}.$$

As the Riemannian metric is positive definite, we have $g_{22} > 0$ and we can write $g_{22} = f^2$. In a way similar to that used in (4.14) we calculate

$$\begin{aligned}
\lim_{r\to 0} g_{22}(r,\varphi_0) &= \lim_{r\to 0} g\left(d_{\tilde{F}(r,\varphi_0)} \exp_p\left(\frac{\partial \tilde{F}}{\partial \varphi}(r,\varphi_0)\right), d_{\tilde{F}(r,\varphi_0)} \exp_p\left(\frac{\partial \tilde{F}}{\partial \varphi}(r,\varphi_0)\right)\right)\\
&= \lim_{r\to 0} g\left(d_{\tilde{F}(r,\varphi_0)} \exp_p(r \cdot Y_2), d_{\tilde{F}(r,\varphi_0)} \exp_p(r \cdot Y_2)\right)\\
&= g\left(d_0 \exp_p(0), d_0 \exp_p(0)\right)\\
&= 0.
\end{aligned}$$

This proves that

$$\lim_{r\to 0} f(r,\varphi_0) = 0.$$

Further,

$$\begin{aligned}
\lim_{r\to 0} \frac{\partial f}{\partial r}(r,\varphi_0) &= \lim_{r\to 0} \frac{f(r,\varphi_0)}{r}\\
&= \lim_{r\to 0} \sqrt{\frac{g_{22}(r,\varphi_0)}{r^2}}\\
&= \sqrt{\lim_{r\to 0} \frac{g_{22}(r,\varphi_0)}{r^2}}\\
&= \sqrt{\lim_{r\to 0} g\left(d_{\tilde{F}(r,\varphi_0)} \exp_p(Y_2), d_{\tilde{F}(r,\varphi_0)} \exp_p(Y_2)\right)}\\
&= \sqrt{g\left(d_0 \exp_p(Y_2), d_0 \exp_p(Y_2)\right)}\\
&= \sqrt{g\left(Y_2, Y_2\right)}\\
&= 1. \qquad \square
\end{aligned}$$

Observe that polar coordinates are not defined for $r = 0$. The differential of the parametrisation F does not have full rank. The point p itself cannot be given in terms of geodesic coordinates at the point p. We cannot therefore evaluate the function f at $r = 0$, though we can consider the limit as r tends to 0. However, we can give p in terms of Riemann normal coordinates at p.

Lemma 4.6.10 *Keep the notation from theorem 4.6.9. Then the Gauss curvature satisfies*

$$K(F(r,\varphi)) = -\frac{1}{f(r,\varphi)}\frac{\partial^2 f}{\partial r^2}(r,\varphi).$$

Proof Theorem 4.6.9 tells us that in geodesic coordinates the Riemannian metric has the form

$$(g_{ij}(r,\varphi))_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & f(r,\varphi)^2 \end{pmatrix}.$$

The inverse matrix is then

$$(g^{ij}(r,\varphi))_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & f(r,\varphi)^{-2} \end{pmatrix}.$$

Using the formula

$$\Gamma_{ij}^k = \frac{1}{2}\sum_{m=1}^{2}\left(\frac{\partial g_{jm}}{\partial u^i} + \frac{\partial g_{im}}{\partial u^j} - \frac{\partial g_{ij}}{\partial u^m}\right)g^{mk}$$

from lemma 4.2.14, we easily calculate

$$\begin{aligned}
\Gamma_{11}^1 &= \Gamma_{11}^2 = \Gamma_{12}^1 = \Gamma_{21}^1 = 0,\\
\Gamma_{12}^2 &= \Gamma_{21}^2 = \frac{1}{f}\frac{\partial f}{\partial r},\\
\Gamma_{22}^1 &= -f\cdot\frac{\partial f}{\partial r},\\
\Gamma_{22}^2 &= \frac{1}{f}\frac{\partial f}{\partial \varphi}.
\end{aligned}$$

The vectors $\partial F/\partial r$ and $(1/f)(\partial F/\partial\varphi)$ form an orthonormal basis of the tangent plane. According to the *Theorema Egregium* (theorem 4.3.8) the Gauss curvature is given by

$$\begin{aligned}
K &= g\left(R\left(\frac{\partial F}{\partial r}, \frac{1}{f}\frac{\partial F}{\partial \varphi}\right)\frac{1}{f}\frac{\partial F}{\partial \varphi}, \frac{\partial F}{\partial r}\right) \\
&= \frac{1}{f^2}R^1_{122} \\
&= \frac{1}{f^2}\left(\frac{\partial \Gamma^1_{22}}{\partial r} - \frac{\partial \Gamma^1_{21}}{\partial \varphi} + \Gamma^1_{11}\Gamma^1_{22} - \Gamma^1_{12}\Gamma^1_{21} + \Gamma^1_{21}\Gamma^2_{22} - \Gamma^1_{22}\Gamma^2_{21}\right) \\
&= \frac{1}{f^2}\left(-\left(\frac{\partial f}{\partial r}\right)^2 - f\cdot\frac{\partial^2 f}{\partial r^2} - 0 + 0 - 0 + 0 + f\cdot\frac{\partial f}{\partial r}\cdot\frac{1}{f}\cdot\frac{\partial f}{\partial r}\right) \\
&= -\frac{1}{f}\frac{\partial^2 f}{\partial r^2}.
\end{aligned}$$ □

Remark In geodesic coordinates the function f carries the complete information about the Riemannian metric. We have seen that Gauss curvature can easily be calculated from f. Let us investigate which requirements on f correspond to the condition of constant Gauss curvature $K \equiv \kappa$. The condition

$$\kappa = -\frac{1}{f(r,\varphi)}\frac{\partial^2 f}{\partial r^2}(r,\varphi)$$

is equivalent to

$$\frac{\partial^2 f}{\partial r^2}(r,\varphi) = -\kappa \cdot f(r,\varphi).$$

This is an *ordinary* differential equation in $r \mapsto f(r,\varphi)$ if we keep φ fixed. This, together with the initial conditions $f(0,\varphi) = 0$ and $(\partial f/\partial r)(0,\varphi) = 1$, uniquely determines the solution, which is

$$f(r,\varphi) = \begin{cases} \frac{1}{\sqrt{\kappa}}\sin(\sqrt{\kappa}\cdot r), & \kappa > 0, \\ r, & \kappa = 0, \\ \frac{1}{\sqrt{-\kappa}}\sinh(\sqrt{-\kappa}\cdot r), & \kappa < 0. \end{cases}$$

This shows, in particular, that κ uniquely determines the metric in the coordinate neighbourhood.

Corollary 4.6.11 *If S_1 and S_2 are two regular surfaces with the same constant Gauss curvature κ, then S_1 and S_2 are locally isometric.*

We now see that the inhabitants of the surface from the introduction of this chapter about the inner geometry of surfaces cannot distinguish different surfaces with the same constant Gauss curvature if they can only take measurements in the close neighbourhood of a point. In particular, they cannot distinguish the plane, the cylinder and the cone.

For the proof of the Gauss–Bonnet formula in the last chapter we will need the following formula, which expresses the integral of Gauss curvature in the neighbourhood of a point. For a regular surface S with Riemannian metric g, a point $p \in S$ and $r > 0$ let

$$\bar{D}(p,r) := \{q \in S \mid \exists \text{ geodesic } c : [0,1] \to S \text{ with } c(0) = p, c(1) = q, L[c] \leq r\}.$$

Lemma 4.6.12 *Keep the notation from theorem 4.6.9. Let $r > 0$ be such that $\bar{D}(p,r)$ is (up to a zero set) covered by the geodesic polar coordinate system. Then*

$$\int_{\bar{D}(p,r)} K \, dA = 2\pi - \int_{\varphi_0}^{\varphi_0+2\pi} \frac{\partial f}{\partial r}(r,\varphi)\, d\varphi.$$

Proof By the formula for the Riemannian metric from Gauss's lemma the surface element is given in geodesic polar coordinates by

$$dA = \sqrt{1 \cdot f^2}\, dr\, d\varphi = f\, dr\, d\varphi.$$

For the geodesic coordinate system to cover $\bar{D}(p,r)$ up to a zero set, the domain of the local parametrisation must be of the form $U = (0,R) \times (\varphi_0, \varphi_0 + 2\pi)$ with $R \geq r$. We substitute the formula for Gauss curvature from lemma 4.6.10 and obtain

$$\begin{aligned}\int_{\bar{D}(p,r)} K\, dA &= -\int_{\varphi_0}^{\varphi_0+2\pi} \int_0^r \frac{1}{f(r,\varphi)} \frac{\partial^2 f}{\partial r^2}(r,\varphi) f(r,\varphi)\, dr\, d\varphi \\ &= -\int_{\varphi_0}^{\varphi_0+2\pi} \int_0^r \frac{\partial^2 f}{\partial r^2}(r,\varphi)\, dr\, d\varphi \\ &= -\int_{\varphi_0}^{\varphi_0+2\pi} \left(\frac{\partial f}{\partial r}(r,\varphi) - 1\right) d\varphi.\end{aligned}$$

The last equality holds since $\lim_{r\to 0} (\partial f/\partial r)(r,\varphi) = 1$. ☐

Having looked at Riemann normal coordinates and geodesic polar coordinates, we now introduce coordinates that are particularly tailored to a given curve on the surface.

Lemma 4.6.13 *Let S be a regular surface with Riemannian metric g. Let $c : I \to S$ be a curve parametrised by arc-length, defined on an open interval I. Let $n : I \to \mathbb{R}^3$ be a vector field on S along c that has constant length 1 and $\langle \dot{c}, n \rangle \equiv 0$. Then there is for $t_0 \in I$ an $\varepsilon > 0$, such that*

$$F : (t_0 - \varepsilon, t_0 + \varepsilon) \times (-\varepsilon, \varepsilon) \to S, \quad F(t,s) := \exp_{c(t)}(sn(t)),$$

is a local parametrisation of S. Along c the Riemannian metric w.r.t. this parametrisation has the form

$$\big(g_{ij}(t,0)\big)_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Proof The partial derivatives of F at the point $(t,0)$ are given by

$$\frac{\partial F}{\partial t}(t,0) = \frac{d}{dt}\exp_{c(t)}(0) = \dot{c}(t)$$

and

$$\frac{\partial F}{\partial s}(t,0) = n(t)$$

and hence form an orthonormal basis of $T_{c(t)}S$. By the inverse function theorem, F is then after a suitable restriction a local parametrisation. Because of the orthonormality the statement about the $g_{ij}(t,0)$ is clear as well. ☐

Definition 4.6.14 The coordinates corresponding to such a parametrisation are called ***Fermi coordinates***.

Fermi coordinates will be useful in the proof of the divergence theorem (theorem 5.1.7).

Exercise 4.25 Show that we have for the Christoffel symbols in Fermi coordinates

$$\Gamma^1_{11}(t,0) = \Gamma^2_{12}(t,0) = \Gamma^2_{21}(t,0) = 0, \quad \Gamma^1_{22}(t,s) = \Gamma^2_{22}(t,s) = 0,$$

$$\Gamma^2_{11}(t,0) = \kappa_g(t), \quad \Gamma^1_{12}(t,0) = \Gamma^1_{21}(t,0) = -\kappa_g(t),$$

where κ_g is the geodesic curvature of c (w.r.t. the orientation of S making $(\dot{c}, n)$ positively oriented).

4.7 Parallel transport

If we have two points p and q in a plane $E \subset \mathbb{R}^3$, then we can identify tangent vectors from T_pE with those from T_qE, since $T_pE = T_qE = E$. If we now replace E by a general regular surface $S \subset \mathbb{R}^3$, then this is no longer straightforward, since in general $T_pS \neq T_qS$. In order still to be able to relate tangent planes at different points even on general surfaces we analyse the concept of "parallel" vectors in the plane a bit more carefully.

We can translate a vector $v_0 \in T_pE$ into the corresponding (the same) vector from T_qE in the following way: we choose a connecting smooth curve $c : [a,b] \to E$, $c(a) = p$, $c(b) = q$, and consider along c the constant vector field $v(t) \equiv v_0$ or, put differently, the vector field along c that satisfies $\dot{v} \equiv 0$. Then $v(b) \in T_qE$ is the vector translated to q. This approach can be used without complications on general regular surfaces by simply replacing the usual derivative of v by the covariant derivative.

Definition 4.7.1 Let S be a regular surface with Riemannian metric g. Let $c : I \to S$ be a smooth curve and $v : I \to \mathbb{R}^3$ be a vector field along c. Then v is ***parallel*** if

$$\frac{\nabla}{dt} v \equiv 0.$$

Remark Hence a smooth curve c is a geodesic if and only if its velocity field $\dot{c}$ is parallel.

We now investigate the property of parallelism w.r.t. a local parametrisation. For this purpose let (U, F, V) be a local parametrisation that contains the trace of the curve c, $c(I) \subset V$. Write $v(t) = \sum_k \xi^k(t)(\partial F/\partial u^k)(u(t))$, where $F \circ u = c$. By (4.2)

$$\frac{\nabla}{dt} v \equiv 0$$

is equivalent to

$$\dot{\xi}^k(t) + \sum_{i,j=1}^{2} \Gamma_{ij}^k(\tilde{c}(t))\xi^i(t)\dot{\tilde{c}}_j(t) \equiv 0 \tag{4.15}$$

for $k = 1, 2$. This is a system of *linear* ordinary differential equations of first order in the coefficients ξ^k. Because of the linearity of the system of equations we are in a situation which is better than the one we were in when we were dealing with the geodesic equations. Now solutions do exist not only locally, but on the whole interval I. Taking this into account, together with the fact that we are dealing with a system of first-order differential equations, so that it suffices to give an initial value (an initial derivative is not needed), we obtain through a reasoning similar to that in the proofs of theorems 4.5.11 and 4.5.12 the following theorem.

Theorem 4.7.2 (Existence and uniqueness of the parallel vector field) *Let S be a regular surface with Riemannian metric g. Let $c : I \to S$ be a smooth curve, $t_0 \in I$, $v_0 \in T_{c(t_0)}S$. Then there exists exactly one parallel vector field v along c with $v(t_0) = v_0$.*

Definition 4.7.3 Let S be a regular surface with Riemannian metric g. Let $c : [t_0, t_1] \to S$ be a smooth curve. The map $P_c : T_{c(t_0)}S \to T_{c(t_1)}S$ that maps v_0 to $v(t_1)$, where v is the unique parallel vector field along c with $v(t_0) = v_0$, is called ***parallel transport along*** c.

Proposition 4.7.4 *Let S be a regular surface with Riemannian metric g. Let* $c : [t_0, t_1] \to S$ *be a smooth curve. Then the following hold:*

(i) *If* $v_0 \in T_{c(t_0)}S$, *then the parallel vector field* v *along* c *with* $v(t_0) = v_0$ *is given by* $v(t) = P_{c|[t_0,t]}(v_0)$.

(ii) *The parallel transport* $P_c : T_{c(t_0)}S \to T_{c(t_1)}S$ *is a linear isometry.*

(iii) *Parallel transport is compatible with the reparametrisation of curve, i.e. if* v *is parallel along* c *and* $\varphi : J \to [t_0, t]$ *is a reparametrisation of* c, *then* $v \circ \varphi$ *is parallel along* $c \circ \varphi$.

Proof Statement (i) follows directly from the uniqueness of parallel transport. We have linearity of parallel transport because of the following: let v_0 and w_0 be from $T_{c(t_0)}S$, and let v and w be the parallel vector fields along c with $v(t_0) = v_0$ and $w(t_0) = w_0$. Let $\alpha, \beta \in \mathbb{R}$. Then the vector field $z(t) := \alpha \cdot v(t) + \beta \cdot w(t)$ is parallel to c, since

$$\frac{\nabla}{dt} z = \alpha \cdot \frac{\nabla}{dt} v + \beta \cdot \frac{\nabla}{dt} w = 0.$$

The field z satisfies the initial condition $z(t_0) = \alpha \cdot v_0 + \beta \cdot w_0$. Hence

$$P_c(\alpha \cdot v_0 + \beta \cdot w_0) = z(t_1) = \alpha \cdot v(t_1) + \beta \cdot w(t_1) = \alpha \cdot P_c(v_0) + \beta \cdot P_c(w_0).$$

This proves the linearity of P_c. For two parallel fields v and w along c we have

$$\begin{aligned} \frac{d}{dt} g(v, w) &= g\left(\frac{\nabla}{dt} v, w\right) + g\left(v, \frac{\nabla}{dt} w\right) \\ &= g(0, w) + g(v, 0) \\ &= 0. \end{aligned}$$

Thus $g_{c(t_0)}(v(t_0), w(t_0)) = g_{c(t_1)}(v(t_1), w(t_1))$, i.e.

$$g_{c(t_0)}(v_0, w_0) = g_{c(t_1)}(P_c(v_0), P_c(w_0)).$$

This means precisely that P_c is a linear *isometry*. Statement (ii) has therefore been proved.

Point (iii) follows from lemma 4.2.12:

$$\frac{\nabla}{dt}(v \circ \varphi) = \dot{\varphi} \cdot \left(\frac{\nabla}{dt} v \circ \varphi\right) = 0. \qquad \square$$

Exercise 4.26 Show that parallel transport is a concept of the inner geometry of the surface. More precisely, if $c : I \to S_1$ is a smooth curve, v a smooth vector field along c and $f : S_1 \to S_2$ a local isometry, then v is parallel along c

if and only if $df \circ v$ is parallel along $f \circ c$. Further conclude that the following diagram commutes for $I = [t_0, t_1]$, $c(t_0) = p$, $c(t_1) = q$:

$$\begin{tikzcd}
T_pS_1 \arrow[r, "P_c"] \arrow[d, "d_pf"] & T_qS_1 \arrow[d, "d_qf"] \\
T_{f(p)}S_2 \arrow[r, "P_{f\circ c}"] & T_{f(q)}S_2
\end{tikzcd}$$

As the covariant and the usual derivative agree on the affine plane $E \subset S$, a vector field is parallel along a curve if and only if it is constant. In particular, the parallel transport $P_c : T_pE \to T_qE$ is independent of the curve c that connects the points p and q when we are dealing with the plane. This is not the case on general regular surfaces.

Example 4.7.5 Let S be the conical surface,

$$S = \left\{ (\xi, \eta, \zeta)^\top \;\middle|\; \xi^2 + \eta^2 = \tfrac{1}{3}\zeta^2, \zeta > 0 \right\}.$$

To investigate parallel vector fields along the curve $c : [0, \pi] \to S$, $c(t) = \frac{1}{2}(\sin(2t), \cos(2t), \sqrt{3})^\top$ in S, we consider the x–y plane $\tilde{S} = \mathbb{R}^2 \times \{0\}$. The map

$$f : \tilde{S} \to S, \qquad f(x, y, 0) = \frac{1}{2\sqrt{x^2 + y^2}} \left(x^2 - y^2, 2xy, \sqrt{3}(x^2 + y^2) \right)^\top$$

is a local isometry, see exercise 4.2. The curve $\tilde{c} : [0, \pi] \to \tilde{S}$, $\tilde{c}(t) = (\cos(t), \sin(t), 0)^\top$, satisfies that $c = f \circ \tilde{c}$. In the plane the parallel vector fields are exactly the constant vector fields. The vector field $\tilde{v}(t) = (1, 0, 0)^\top$, for example, is parallel along $\tilde{c}$. It follows that $v := df \circ \tilde{v}$ is parallel along c in S. Using

$$\frac{\partial f}{\partial x} = \frac{1}{\sqrt{x^2 + y^2}} \begin{pmatrix} x - \frac{1}{2}\dfrac{(x^2 - y^2)\,x}{x^2 + y^2} \\ y - \dfrac{x^2 y}{x^2 + y^2} \\ \dfrac{\sqrt{3}}{2} x \end{pmatrix},$$

we calculate

$$\begin{aligned}
v(t) &= d_{\tilde{c}(t)} f(\tilde{v}(t)) \\
&= d_{\tilde{c}(t)} f \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \\
&= \frac{\partial f}{\partial x}(\tilde{c}(t))
\end{aligned}$$

$$= \begin{pmatrix} \cos(t) - \frac{1}{2}\left(\cos(t)^2 - \sin(t)^2\right)\cos(t) \\ \sin(t) - \cos(t)^2 \sin(t) \\ \frac{\sqrt{3}}{2}\cos(t) \end{pmatrix}$$

$$= \begin{pmatrix} \cos(t)\left(1 - \frac{1}{2}\cos(2t)\right) \\ \sin(t)^3 \\ \frac{\sqrt{3}}{2}\cos(t) \end{pmatrix}.$$

In particular, $v(0) = (1/2, 0, \sqrt{3}/2)^\top$ and $v(\pi) = (-1/2, 0, -\sqrt{3}/2)^\top = -v(0)$. The parallel transport changes the tangent vector $v_0 = (1/2, 0, \sqrt{3}/2)^\top$ into its negative in one revolution around c. More generally, v_0 is mapped to itself for an even number k of revolutions, i.e. c, parametrised on $[0, k \cdot \pi]$, is mapped to itself by parallel translation, while it is mapped to $-v_0$ for uneven k.

Exercise 4.27 Illustrate this by drawing a constant vector field along half of a circular line on a sheet of paper, and then making a cone out of the half of the disc.

This example shows that parallel transport generally depends on the connecting curve on regular surfaces.

Exercise 4.28 Let S be a regular surface with Riemannian metric g. Let $c : I \to S$ be a constant curve, $c \equiv p$. A vector field along c is therefore a map $v : I \to T_pS$. Show that v is parallel if and only if v is constant.

4.8 Jacobi fields

We discussed the geometric meaning of Gauss curvature in section 3.6. Theorem 3.6.15 told us how the surface behaves near a point relative to the tangent plane at that point, in terms of the of the sign of the Gauss curvature at that point. But those investigations are only valid for the first fundamental form. In the case of a general Riemannian metric we want to understand what Gauss curvature tells us about the inner geometry of a surface. Jacobi fields are an important tool for this; they, as we will see below, contain information about the behaviour of neighbouring geodesics.

Definition 4.8.1 Let S be a regular surface with Riemannian metric g. Let c be a geodesic on S. A smooth vector field J on S along c is called a ***Jacobi field*** if

$$\frac{\nabla}{dt}\frac{\nabla}{dt}J = -R(J, \dot{c})\dot{c}.$$

This Jacobi equation is a linear ordinary differential equation of second order in J. We can therefore take arbitrary initial values $J(t_0)$ and $(\nabla/dt)J(t_0)$ in

5- 2 T3 @;) 4 6 T 6 E *) %)))) 8 L x y 8
6 N3) 9 % ;4 +

We see that for the interesting Jacobi fields the Jacobi equation is equivalent to the following ordinary differential equation:

$$\ddot{\chi} = -\chi \cdot (K \circ c_0), \tag{4.16}$$

in which the Gauss curvature K plays a significant role. The solution to this differential equation cannot generally be found explicitly very easily. The interesting Jacobi fields give us actual information about the surface that tells us approximately how much neighbouring geodesics (from our geodesic variation) diverge from each other. Let us illustrate this with a simple but important example.

Example 4.8.4 If the Gauss curvature of a surface is constant, $K \equiv \kappa$, then (4.16) can be solved explicitly. Setting

$$\mathfrak{s}_\kappa(t) := \begin{cases} \sin(\sqrt{\kappa}t)/\sqrt{\kappa}, & \kappa > 0, \\ t, & \kappa = 0, \\ \sinh(\sqrt{|\kappa|}t)/\sqrt{|\kappa|}, & \kappa < 0, \end{cases} \qquad \mathfrak{c}_\kappa(t) := \begin{cases} \cos(\sqrt{\kappa}t), & \kappa > 0, \\ 1, & \kappa = 0, \\ \cosh(\sqrt{|\kappa|}t), & \kappa < 0, \end{cases}$$

we have

$$\mathfrak{c}_\kappa(t)^2 + \kappa\mathfrak{s}_\kappa(t)^2 = 1, \tag{4.17}$$

$$\dot{\mathfrak{s}}_\kappa = \mathfrak{c}_\kappa, \tag{4.18}$$

$$\dot{\mathfrak{c}}_\kappa = -\kappa\mathfrak{s}_\kappa. \tag{4.19}$$

In particular, the functions $\mathfrak{s}_\kappa$ and $\mathfrak{c}_\kappa$ solve the differential equation

$$\ddot{\chi} = -\kappa \cdot \chi$$

and are linearly independent. Hence every solution of this differential equation can be written as a linear combination of $\mathfrak{s}_\kappa$ and $\mathfrak{c}_\kappa$. The interesting Jacobi fields are thus of the form

$$J(t) = (\alpha\mathfrak{s}_\kappa(t) + \beta\mathfrak{c}_\kappa(t))n(t).$$

They behave very differently depending on the sign of κ. If $\kappa > 0$, then the length of J is periodic with period $2\pi/\kappa$. In particular, J is 0 again and again. For geodesics that diverge at a point this means that they will approach each other again after some time.

If $\kappa = 0$, then the length increases linearly. This is the familiar situation that we know from straight lines in the Euclidean plane.

6 73

In the case $\kappa < 0$ the length increases even more quickly, in fact exponentially. Summarising we can say that the more negative the curvature, the more the geodesics diverge.

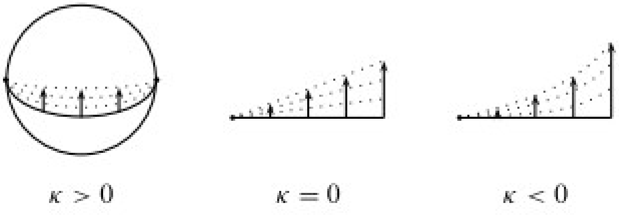

Exercise 4.29 Show that on the surface of revolution with parametrisation $F(t, \varphi) = (r(t)\cos(\varphi), r(t)\sin(\varphi), t)^\top$ the vector field $\partial F/\partial\varphi$ is a Jacobi field along the lines of longitude.

4.9 Spherical and hyperbolic geometry

We now want to investigate spherical geometry in more detail, in particular with regard to trigonometric rules analogous to those of Euclidean geometry, which we treated in the first chapter. At the same time we introduce a surface which is in a way the negatively curved analogue of the sphere, the hyperbolic plane. We can investigate its trigonometry simultaneously. The hyperbolic plane has great historic importance, since it satisfies all axioms of Euclidean geometry except for the parallel axiom. It also shows that the efforts to derive the parallel axiom from the other axioms, which had been going on for thousands of years, were inevitably in vain.

Let us construct the hyperbolic plane. For $\kappa \in \mathbb{R}$ we set

$$\widehat{\mathbb{M}}_\kappa := \left\{ (x, y, z)^\top \in \mathbb{R}^3 \,\middle|\, \kappa(x^2 + y^2) + z^2 = 1 \right\}.$$

In the case $\kappa = 1$ this is simply the sphere $\widehat{\mathbb{M}}_1 = S^2$. In the case $\kappa = 0$, on the other hand, we obtain the union of two parallel planes through the points $(0, 0, 1)$ and $(0, 0, -1)$. We are now mostly interested in the case $\kappa = -1$. The surface $\widehat{\mathbb{M}}_{-1}$ is the ***two-sheeted hyperboloid***, consisting of two sheets that are graphs of the functions $z = \pm\sqrt{1 + x^2 + y^2}$. Being graphs, they are, in particular, regular surfaces. Since we do not want to work with several sheets, we additionally define

$$\mathbb{M}_\kappa := \begin{cases} \widehat{\mathbb{M}}_\kappa, & \kappa > 0, \\ \{(x, y, z)^\top \in \widehat{\mathbb{M}}_\kappa \mid z > 0\}, & \kappa \leq 0. \end{cases}$$

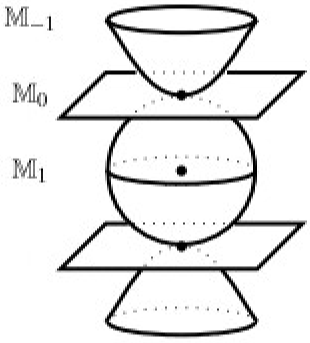

So we take only the upper plane or sheet. We already know that $\mathbb{M}_1$ has constant Gauss curvature $K \equiv 1$, while the plane has $K \equiv 0$, in each case with the first fundamental form as Riemannian metric. If we now take the first fundamental form on $\mathbb{M}_{-1}$, then it does not have curvature -1, which we would have liked.

Exercise 4.30 Show that $\mathbb{M}_{-1}$ with the first fundamental form as Riemannian metric has positive Gauss curvature.

Let us therefore take another Riemannian metric on $\mathbb{M}_{-1}$. Indeed, this example is so important that we have already introduced general Riemannian metrics because of this. For the construction of the Riemannian metric we define the following symmetric bilinear form on $\mathbb{R}^3$:

$$\left\langle \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} \right\rangle_\kappa := xx' + yy' + \frac{zz'}{\kappa}.$$

In the case $\kappa = 1$, this is the usual Euclidean scalar product. If $\kappa = -1$, then the symmetric bilinear form is not degenerate, but indefinite. We call $\langle \cdot, \cdot \rangle_{-1}$ the ***Minkowski scalar product*** on $\mathbb{R}^3$. It plays an important role in Einstein's special relativity theory. We call the set of all vectors of "vanishing length" the ***light cone***

$$\left\{ X \in \mathbb{R}^3 \,\middle|\, \langle X, X \rangle_{-1} = 0 \right\} = \left\{ (x, y, z)^\top \in \mathbb{R}^3 \,\middle|\, z^2 = x^2 + y^2 \right\}.$$

We now define the Riemannian metric on $\mathbb{M}_\kappa$ by restricting the symmetric bilinear form $\langle \cdot, \cdot \rangle_\kappa$ for every point $p \in \mathbb{M}_\kappa$ to $T_p\mathbb{M}_\kappa$. The construction is thus similar to that of the first fundamental form, except that we have replaced the usual Euclidean scalar product on $\mathbb{R}^3$ by $\langle \cdot, \cdot \rangle_\kappa$.

In the case $\kappa = 0$, one could complain that we need to divide by 0 in the definition of $\langle \cdot, \cdot \rangle_0$, and that the definition is therefore not reasonable. But every tangent plane of $\mathbb{M}_0$ equals $\mathbb{R}^2 \times \{0\}$, i.e. the coefficient of the zz'-term does not play a role anyway. We could therefore without worry replace $\langle\cdot,\cdot\rangle_0$ by the usual Euclidean scalar product without changing the Riemannian metric induced on $\mathbb{M}_0$. It is simply the usual Euclidean metric on the plane.

Since for $\kappa < 0$ the bilinear form $\langle\cdot,\cdot\rangle_\kappa$ is no longer positive definite, it is a priori not clear that the restrictions to the tangent planes of $\mathbb{M}_\kappa$ are positive definite. They need to have this property if we want to obtain a Riemannian metric. To check that at least the restrictions of $\langle\cdot,\cdot\rangle_\kappa$ to the tangent planes are positive definite, we parametrise $\mathbb{M}_\kappa$ as follows:

$$F_\kappa : \mathbb{R}\times\mathbb{R} \to \mathbb{R}^3, \quad F_\kappa(r,\varphi) := \begin{pmatrix} \mathfrak{s}_\kappa(r)\cos(\varphi) \\ \mathfrak{s}_\kappa(r)\sin(\varphi) \\ \mathfrak{c}_\kappa(r) \end{pmatrix}.$$

Here $\mathfrak{c}_\kappa$ and $\mathfrak{s}_\kappa$ are the generalised sine and cosine functions defined in example 4.8.4. Using (4.17) it is easily seen that the image of F_κ is exactly $\mathbb{M}_\kappa$, more precisely

$$F_\kappa(\mathbb{R}\times\mathbb{R}) = F_\kappa([0,\pi/\sqrt{\kappa}]\times[0,2\pi)) = \mathbb{M}_\kappa$$

in the case $\kappa > 0$ and

$$F_\kappa(\mathbb{R}\times\mathbb{R}) = F_\kappa([0,\infty)\times[0,2\pi)) = \mathbb{M}_\kappa$$

in the case $\kappa \leq 0$. Using (4.18) and (4.19) we immediately get the partial derivatives of F_κ:

$$\frac{\partial F_\kappa}{\partial r}(r,\varphi) = \begin{pmatrix} \mathfrak{c}_\kappa(r)\cos(\varphi) \\ \mathfrak{c}_\kappa(r)\sin(\varphi) \\ -\kappa\mathfrak{s}_\kappa(r) \end{pmatrix}, \quad \frac{\partial F_\kappa}{\partial \varphi}(r,\varphi) = \begin{pmatrix} -\mathfrak{s}_\kappa(r)\sin(\varphi) \\ \mathfrak{s}_\kappa(r)\cos(\varphi) \\ 0 \end{pmatrix}.$$

The two partial derivatives are linearly independent for $r \in (0,\pi/\sqrt{\kappa})$ and $r \in (0,\infty)$. Restriction of F_κ to suitable domains therefore gives us local parametrisations cf. the $\mathbb{M}_\kappa$ except for the points $(0,0,1)^\top$ and, in the case $\kappa > 0$, $(0,0,-1)^\top$. At those two special points the tangent plane is exactly the x–y plane and a restriction of $\langle\cdot,\cdot\rangle_\kappa$ gives the usual Euclidean scalar product, and is thus positive definite. For the other points we calculate the coefficients g_{ij} of our restriction with respect to the local parametrisation F_κ, e.g.

$$\begin{aligned}\left\langle \frac{\partial F_\kappa}{\partial r}(r,\varphi), \frac{\partial F_\kappa}{\partial r}(r,\varphi) \right\rangle_\kappa &= \mathfrak{c}_\kappa(r)^2\cos(\varphi)^2 + \mathfrak{c}_\kappa(r)^2\sin(\varphi)^2 + \frac{1}{\kappa}\kappa^2\mathfrak{s}_\kappa(r)^2 \\ &= \mathfrak{c}_\kappa(r)^2 + \kappa\mathfrak{s}_\kappa(r)^2 = 1.\end{aligned}$$

The other components are obtained in a similar way:

$$\left(g_{ij}(r,\varphi)\right)_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & \mathfrak{s}_\kappa(r)^2 \end{pmatrix}. \tag{4.20}$$

This matrix is positive definite for $r \in (0, \pi/\sqrt{\kappa})$ and $r \in (0, \infty)$. Therefore we really have a Riemannian metric on $\mathbb{M}_\kappa$.

Definition 4.9.1 The regular surface $\mathbb{M}_{-1}$ together with the Riemannian metric defined via the restriction of $\langle \cdot, \cdot \rangle_{-1}$, is called the ***hyperbolic plane***.

The special form of the metric in (4.20) shows that F_κ gave us a parametrisation in geodesic polar coordinates at the point $(0,0,1)^\top$. By lemma 4.6.10 the Gauss curvature is now given by

$$K = -\frac{\ddot{\mathfrak{s}}_\kappa(r)}{\mathfrak{s}_\kappa(r)} = \kappa.$$

In particular, the hyperbolic plane has constant Gauss curvature -1. Further, the curves $r \mapsto F_\kappa(r, \varphi)$ are the geodesics parametrised by arc-length that begin at $(0,0,1)^\top$. The traces of these geodesics have a particularly simple geometric characterisation. They are exactly the intersection of the surface $\mathbb{M}_\kappa$ with the plane through the origin, through $(0,0,1)^\top$ and through $(\cos(\varphi), \sin(\varphi), 0)^\top$. In the case of the sphere $\mathbb{M}_1$ this gives the great circles through the "north pole" $(0,0,1)^\top$. In the case of the plane we get the straight lines through this point. In the case of the hyperbolic plane we obtain hyperbolas. What do the other geodesics look like, those that do not pass through $(0,0,1)^\top$?

To understand this, we will use isometries to map the geodesics through $(0,0,1)^\top$ to other geodesics. To do so we need to obtain enough isometries. We investigate which linear maps on $\mathbb{R}^3$ become isometries after a restriction to $\mathbb{M}_\kappa$. For this purpose let $L \in \mathrm{GL}(3)$. Let us first consider the case $\kappa \neq 0$. If L preserves the symmetric bilinear form $\langle \cdot, \cdot \rangle_\kappa$, i.e.

$$\langle LX, LY \rangle_\kappa = \langle X, Y \rangle_\kappa$$

for all $X, Y \in \mathbb{R}^3$, then, in particular, $\widehat{\mathbb{M}}_\kappa$ is mapped to itself and for every $p \in \widehat{\mathbb{M}}_\kappa$ the map $d_pL = L|_{T_p\widehat{\mathbb{M}}_\kappa} : T_p\widehat{\mathbb{M}}_\kappa \to T_{Lp}\widehat{\mathbb{M}}_\kappa$ is a linear isometry. However, in the case $\kappa < 0$ it can happen that the two sheets of the hyperboloid are interchanged. To prevent this we additionally require that $(0,0,1)^\top$ is mapped back to the upper sheet, i.e. the third component of the image vector must be positive. Writing

$$L = \begin{pmatrix} L_{11} & L_{12} & L_{13} \\ L_{21} & L_{22} & L_{23} \\ L_{31} & L_{32} & L_{33} \end{pmatrix},$$

we see that $L_{33} > 0$. The other points of the upper sheet will for reasons of continuity be mapped to the upper sheet as well:

$$L(\mathbb{M}_\kappa) = \mathbb{M}_\kappa.$$

In the case $\kappa = 0$, the vectors of the form $(x, y, 1)^\top$ must be mapped to vectors whose third component is also equal to 1. This means precisely that $L_{31} = L_{32} = 0$ and $L_{33} = 1$. Further, the Euclidean map must be preserved in the x–y plane for the maps to be isometries. This means that

$$\begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix} \in \mathrm{O}(2).$$

Summarising, we can say that we have already found rather large groups of isometries on $\mathbb{M}_\kappa$, i.e.

$$G_\kappa := \begin{cases} \{L \in \mathrm{GL}(3) \mid \langle LX, LY\rangle_\kappa = \langle X, Y\rangle_\kappa \text{ and } L_{33} > 0\}, & \kappa < 0, \\ \{L \in \mathrm{GL}(3) \mid L_{31} = L_{32} = 0, L_{33} = 1, (L_{ij})_{i,j=1,2} \in \mathrm{O}(2)\}, & \kappa = 0, \\ \{L \in \mathrm{GL}(3) \mid \langle LX, LY\rangle_\kappa = \langle X, Y\rangle_\kappa\}, & \kappa > 0. \end{cases}$$

For $\kappa = 1$ we have $G_1 = \mathrm{O}(3)$. In the case $\kappa = 0$, the group G_0 is isomorphic to the Euclidean group $\mathrm{E}(2)$, see definition 1.2.1. Indeed, $L \in G_0$ maps the vector $(x, y, 1)^\top$ to $(L_{11}x + L_{12}y + L_{13}, L_{21}x + L_{22}y + L_{23}, 1)^\top$, i.e. L acts like a Euclidean motion $F_{A,b}$ with $A = (L_{ij})_{i,j=1,2}$ and translational part $b = (L_{13}, L_{23})^\top$.

The elements of the group G_{-1} are called the ***time-preserving Lorentz transformations*** or ***orthochronous Lorentz transformations***, a terminology that comes from relativity theory.

For every κ, the group G_κ contains the rotations around the z-axis:

$$\begin{pmatrix} \cos(\varphi) & -\sin(\varphi) & 0 \\ \sin(\varphi) & \cos(\varphi) & 0 \\ 0 & 0 & 1 \end{pmatrix} \in G_\kappa.$$

Hence every point from $\mathbb{M}_\kappa$ can be moved into the x–z plane by an isometry, i.e. to a point of the form $(\mathfrak{s}_\kappa(r), 0, \mathfrak{c}_\kappa(r))^\top$, without moving the point $(0, 0, 1)^\top$ in this process.

For such a point $(\mathfrak{s}_\kappa(r), 0, \mathfrak{c}_\kappa(r))^\top$ we consider the linear map

$$L := \begin{pmatrix} -\mathfrak{c}_\kappa(r) & 0 & \mathfrak{s}_\kappa(r) \\ 0 & 1 & 0 \\ \kappa\,\mathfrak{s}_\kappa(r) & 0 & \mathfrak{c}_\kappa(r) \end{pmatrix} \in G_\kappa. \tag{4.21}$$

This L interchanges the two points $(\mathfrak{s}_\kappa(r), 0, \mathfrak{c}_\kappa(r))^\top$ and $(0, 0, 1)^\top$. Altogether we see that *every* point from $\mathbb{M}_\kappa$ can be mapped to the point $(0, 0, 1)^\top$ by a suitable isometry. We say that the isometry group *acts transitively* on $\mathbb{M}_\kappa$. Loosely one could say that $\mathbb{M}_\kappa$ with the Riemannian metric that we constructed is a surface with a very high symmetry.

To come back to our discussion of geodesics, for an arbitrary geodesic through a point $p \in \mathbb{M}_\kappa$ we can find an isometry $L \in G_\kappa$ which maps p to $(0, 0, 1)^\top$ and the geodesic to the intersection of $\mathbb{M}_\kappa$ with a two-dimensional subspace E of $\mathbb{R}^3$. As L is a linear map on $\mathbb{R}^3$, we know that $L^{-1}(E)$ is a two-dimensional subspace of $\mathbb{R}^3$ as well. Generally it no longer contains the point $(0,0,1)^\top$, but instead contains the point p. This is summarised in the following theorem:

Theorem 4.9.2 *The traces of geodesics of $\mathbb{M}_\kappa$ with the Riemannian metric defined by the restriction of $\langle\cdot,\cdot\rangle_\kappa$ are exactly the (non-empty) intersections of $\mathbb{M}_\kappa$ with the two-dimensional subspaces of $\mathbb{R}^3$.*

In the case $\kappa = 1$ these are the great circles on S^2, in the case $\kappa = 0$ they are the straight lines in the plane and in the case $\kappa = -1$ we obtain hyperbolas.

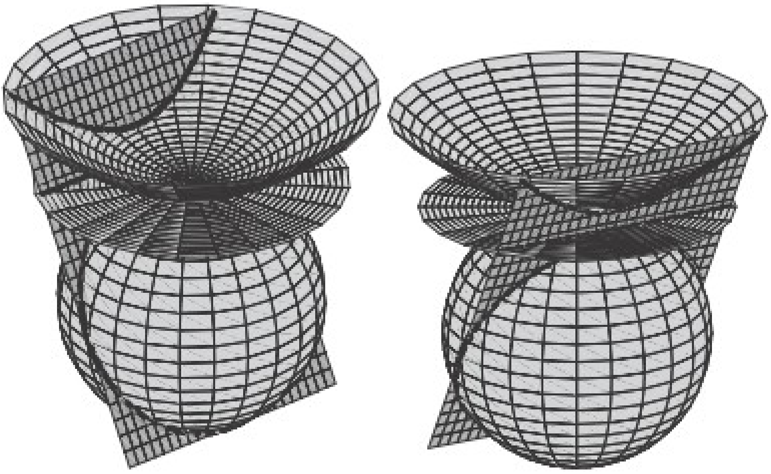

Exercise 4.31 Show that G_κ already contains all isometries of $\mathbb{M}_\kappa$.

Hint First use the exponential map to show that for $p \in \mathbb{M}_\kappa$ an isometry $f : \mathbb{M}_\kappa \to \mathbb{M}_\kappa$ is already uniquely determined by the image point $f(p)$ and the differential $d_p f : T_p\mathbb{M}_\kappa \to T_{f(p)}\mathbb{M}_\kappa$.

We now begin to derive the trigonometric theorems for $\mathbb{M}_\kappa$. For this purpose we consider a ***geodesic triangle*** on $\mathbb{M}_\kappa$, i.e. three points $A, B, C \in \mathbb{M}_\kappa$, which are connected by geodesics of length a, b, c. We will denote the interior angles at the vertices by α, β, γ. We will in the following always assume that the geodesic triangles are not degenerate, i.e. the lengths of the sides are positive and in the case $\kappa > 0$ smaller than $\pi/\sqrt{\kappa}$ and the angles are in $(0, \pi)$.

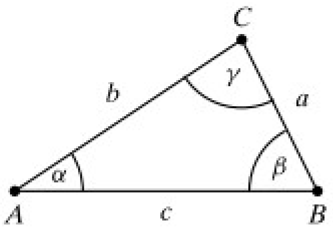

We want to find relations between the angles and the lengths of sides. As isometries do not change angles or the lengths of sides, we can first use an isometry to map the point A to $(0,0,1)^\top$ and then rotate B into the x–z plane. Without loss of generality we therefore have

$$A = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad B = \begin{pmatrix} \mathfrak{s}_\kappa(c) \\ 0 \\ \mathfrak{c}_\kappa(c) \end{pmatrix}, \quad C = \begin{pmatrix} \mathfrak{s}_\kappa(b)\cos(\alpha) \\ \mathfrak{s}_\kappa(b)\sin(\alpha) \\ \mathfrak{c}_\kappa(b) \end{pmatrix}.$$

We apply isometry L from (4.21), which interchanges points A and B, with $r = c$ to the entire triangle.

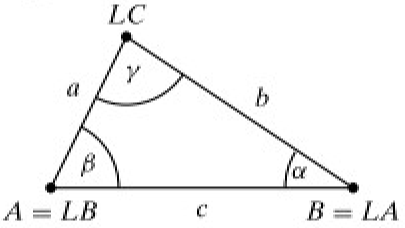

As the side from LC to LB has length a and the angle at A is β, we have

$$LC = \begin{pmatrix} \mathfrak{s}_\kappa(a)\cos(\beta) \\ \mathfrak{s}_\kappa(a)\sin(\beta) \\ \mathfrak{c}_\kappa(a) \end{pmatrix}. \tag{4.22}$$

On the other hand, we calculate

$$\begin{aligned} LC &= \begin{pmatrix} -\mathfrak{c}_\kappa(c) & 0 & \mathfrak{s}_\kappa(c) \\ 0 & 1 & 0 \\ \kappa\mathfrak{s}_\kappa(c) & 0 & \mathfrak{c}_\kappa(c) \end{pmatrix} \cdot \begin{pmatrix} \mathfrak{s}_\kappa(b)\cos(\alpha) \\ \mathfrak{s}_\kappa(b)\sin(\alpha) \\ \mathfrak{c}_\kappa(b) \end{pmatrix} \\ &= \begin{pmatrix} -\mathfrak{c}_\kappa(c)\mathfrak{s}_\kappa(b)\cos(\alpha) + \mathfrak{s}_\kappa(c)\mathfrak{c}_\kappa(b) \\ \mathfrak{s}_\kappa(b)\sin(\alpha) \\ \kappa\mathfrak{s}_\kappa(c)\mathfrak{s}_\kappa(b)\cos(\alpha) + \mathfrak{c}_\kappa(c)\mathfrak{c}_\kappa(b) \end{pmatrix}. \end{aligned} \tag{4.23}$$

We compare the components in (4.22) and (4.23) and obtain

$$\mathfrak{s}_\kappa(a)\cos(\beta) = -\mathfrak{c}_\kappa(c)\mathfrak{s}_\kappa(b)\cos(\alpha) + \mathfrak{s}_\kappa(c)\mathfrak{c}_\kappa(b), \tag{4.24}$$

$$\mathfrak{s}_\kappa(a)\sin(\beta) = \mathfrak{s}_\kappa(b)\sin(\alpha), \tag{4.25}$$

$$\mathfrak{c}_\kappa(a) = \kappa\mathfrak{s}_\kappa(c)\mathfrak{s}_\kappa(b)\cos(\alpha) + \mathfrak{c}_\kappa(b)\mathfrak{c}_\kappa(c). \tag{4.26}$$

Equation 4.25 gives the *sine rule*

$$\frac{\mathfrak{s}_\kappa(a)}{\sin(\alpha)} = \frac{\mathfrak{s}_\kappa(b)}{\sin(\beta)},$$

and (4.26) is called the *cosine rule for sides*. Multiplying (4.24) by $\cos(\alpha)$ and (4.25) by $\sin(\alpha)$ and subtracting the resulting equations one from the other gives

$$\begin{aligned}&\mathfrak{s}_\kappa(a)\,(\cos(\alpha)\cos(\beta)-\sin(\alpha)\sin(\beta))\\&\quad=-\mathfrak{c}_\kappa(c)\mathfrak{s}_\kappa(b)\cos(\alpha)^2+\mathfrak{s}_\kappa(c)\mathfrak{c}_\kappa(b)\cos(\alpha)-\mathfrak{s}_\kappa(b)\sin(\alpha)^2.\end{aligned}\tag{4.27}$$

Interchanging the roles of b and c in (4.24) (i.e. considering triangle ABC instead of ACB) we obtain

$$\mathfrak{s}_\kappa(a)\cos(\gamma)=-\mathfrak{c}_\kappa(b)\mathfrak{s}_\kappa(c)\cos(\alpha)+\mathfrak{s}_\kappa(b)\mathfrak{c}_\kappa(c).\tag{4.28}$$

We substitute (4.28) into (4.27) and finally use the sine rule:

$$\begin{aligned}&\mathfrak{s}_\kappa(a)\,(\cos(\alpha)\cos(\beta)-\sin(\alpha)\sin(\beta))\\&\quad=-\mathfrak{c}_\kappa(c)\mathfrak{s}_\kappa(b)\cos(\alpha)^2-\mathfrak{s}_\kappa(a)\cos(\gamma)+\mathfrak{s}_\kappa(b)\mathfrak{c}_\kappa(c)-\mathfrak{s}_\kappa(b)\sin(\alpha)^2\\&\quad=\mathfrak{s}_\kappa(b)\mathfrak{c}_\kappa(c)\sin(\alpha)^2-\mathfrak{s}_\kappa(a)\cos(\gamma)-\mathfrak{s}_\kappa(b)\sin(\alpha)^2\\&\quad=\mathfrak{s}_\kappa(a)\mathfrak{c}_\kappa(c)\sin(\alpha)\sin(\beta)-\mathfrak{s}_\kappa(a)\cos(\gamma)-\mathfrak{s}_\kappa(a)\sin(\alpha)\sin(\beta).\end{aligned}$$

We divide this equation by $\mathfrak{s}_\kappa(a)$, add $\sin(\alpha)\sin(\beta)$ on both sides and obtain

$$\cos(\alpha)\cos(\beta)=\mathfrak{c}_\kappa(c)\sin(\alpha)\sin(\beta)-\cos(\gamma).$$

This is the *cosine rule for angles*. We summarise these rules in the following theorem:

Theorem 4.9.3 (Spherical and hyperbolic trigonometry) *Let ABC be a geodesic triangle with side-lengths a, b, c and interior angles α, β, γ on $\mathbb{M}_\kappa$. Then the following hold:*

(i) *the sine rule:*

$$\frac{\mathfrak{s}_\kappa(a)}{\sin(\alpha)}=\frac{\mathfrak{s}_\kappa(b)}{\sin(\beta)}=\frac{\mathfrak{s}_\kappa(c)}{\sin(\gamma)};$$

(ii) *the cosine rule for sides:*

$$\begin{aligned}\mathfrak{c}_\kappa(a)&=\mathfrak{c}_\kappa(b)\mathfrak{c}_\kappa(c)+\kappa\mathfrak{s}_\kappa(b)\mathfrak{s}_\kappa(c)\cos(\alpha),\\\mathfrak{c}_\kappa(b)&=\mathfrak{c}_\kappa(a)\mathfrak{c}_\kappa(c)+\kappa\mathfrak{s}_\kappa(a)\mathfrak{s}_\kappa(c)\cos(\beta),\\\mathfrak{c}_\kappa(c)&=\mathfrak{c}_\kappa(a)\mathfrak{c}_\kappa(b)+\kappa\mathfrak{s}_\kappa(a)\mathfrak{s}_\kappa(b)\cos(\gamma);\end{aligned}$$

(iii) *the cosine rule for angles:*

$$\begin{aligned}\cos(\alpha) &= \mathfrak{c}_\kappa(a)\sin(\beta)\sin(\gamma) - \cos(\beta)\cos(\gamma),\\ \cos(\beta) &= \mathfrak{c}_\kappa(b)\sin(\alpha)\sin(\gamma) - \cos(\alpha)\cos(\gamma),\\ \cos(\gamma) &= \mathfrak{c}_\kappa(c)\sin(\alpha)\sin(\beta) - \cos(\alpha)\cos(\beta).\end{aligned}$$

This theorem contains three trigonometric rules for both spherical ($\kappa = 1$) and hyperbolic geometry ($\kappa = -1$). The theorem holds for all $\kappa \in \mathbb{R}$, in particular for $\kappa = 0$. Indeed, the sine rule of Euclidean geometry (theorem 1.2.6) is included in theorem 4.9.3. The cosine rule for sides on the other hand is trivial in the Euclidean case. The cosine rule for angles says that in the case $\kappa = 0$

$$\cos(\gamma) = \sin(\alpha)\sin(\beta) - \cos(\alpha)\cos(\beta) = -\cos(\alpha+\beta) = \cos(\pi - \alpha - \beta),$$

i.e. we obtain theorem 1.2.7 about the sum of interior angles in a Euclidean triangle. In the hyperbolic case $\kappa = -1$ or more generally if $\kappa < 0$, we have for $c > 0$ that $\mathfrak{c}_\kappa(c) > 1$, hence

$$\cos(\gamma) > \sin(\alpha)\sin(\beta) - \cos(\alpha)\cos(\beta) = -\cos(\alpha+\beta) = \cos(\pi - \alpha - \beta)$$

and thus

$$\alpha + \beta + \gamma < \pi.$$

Hyperbolic triangles are therefore "narrow" compared to Euclidean ones. Analogously we see that in the spherical case ($\kappa > 0$)

$$\alpha + \beta + \gamma > \pi.$$

Spherical triangles are "fat".

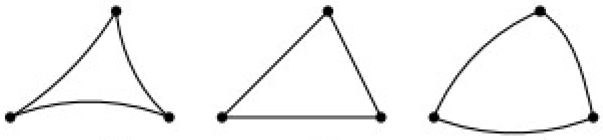

$\kappa < 0$ $\kappa = 0$ $\kappa > 0$

Exercise 4.32 Use the sine rule for geodesic triangles on $\mathbb{M}_\kappa$ to derive the *height formula*

$$\mathfrak{s}_\kappa(h_c) = \mathfrak{s}_\kappa(b)\sin(\alpha) = \mathfrak{s}_\kappa(a)\sin(\beta).$$

h_c is the height of the triangle with over side $\overline{AB}$.

Exercise 4.33 Derive the cosine rule (theorem 1.2.4) in the Euclidean case $\kappa = 0$ from the sine rule and the cosine rule for angles.

4.10 Cartography

Cartography is the science of geographical maps. Mathematically speaking, we want to study charts of S^2. For simplicity, we will ignore all complications arising from the fact that the earth is better described by a certain rotational ellipsoid than by a sphere. Ideally, one would like to find charts which preserve lengths, angles, and areas up to a scale factor. However, this is not possible since such a chart would be an isometry between a portion of S^2 and a portion of $\mathbb{R}^2$. By Gauss's *Theorema Egregium* no such isometry exists because S^2 has Gauss curvature $K \equiv 1$ while the Euclidean plane has $K \equiv 0$. In the literature on cartography one nevertheless finds statements of the extent that certain maps are length-preserving. This is rather misleading because it means that *some* lengths are preserved, not all.

As we will see it is possible to construct charts that are angle-preserving or area-preserving. A chart is area-preserving if the area element of the sphere expressed in the coordinates u^1, u^2 coincides with the area element of the Euclidean metric of the plane, $dA = du^1\, du^2$. How can we see whether or not a chart is angle-preserving?

Lemma 4.10.1 *Let g and g' be two Euclidean metrics on $\mathbb{R}^n$, $n \geq 2$. Then g and g' define the same angles if and only if there exists a number $c > 0$ such that*

$$g' = c \cdot g.$$

Proof If $g' = c \cdot g$, then for any non-zero vectors $X, Y \in \mathbb{R}^n$ we have

$$\frac{g(X,Y)}{\sqrt{g(X,X)}\sqrt{g(Y,Y)}} = \frac{g'(X,Y)}{\sqrt{g'(X,X)}\sqrt{g'(Y,Y)}},$$

because the factor c cancels. Thus the angles between X and Y are the same for g and for g'.

Conversely, let g and g' define the same angles. Fix a non-zero vector $Y \in \mathbb{R}^n$. For any $X \in \mathbb{R}^n$, not a multiple of Y, consider the triangle with vertices 0, Y and X. Denote the angle at the corner X by η and the one at the corner Y by ξ. It is important that the angles are the same for g and for g'. By the sine rule we have

$$\frac{\|X\|_g}{\|Y\|_g} = \frac{\sin(\xi)}{\sin(\eta)} = \frac{\|X\|_{g'}}{\|Y\|_{g'}},$$

hence

$$g'(X,X) = c \cdot g(X,X)$$

with $c = g'(Y, Y)/g(Y, Y)$. Since the set of vectors X which are not a multiple of Y is dense in $\mathbb{R}^n$ this relation holds for *all* $X \in \mathbb{R}^n$ by continuity. Polarisation implies

$$g'(X, Z) = c \cdot g(X, Z)$$

for all $X, Z \in \mathbb{R}^n$. □

This justifies the following.

Definition 4.10.2 Let S and S' be regular surfaces equipped with Riemannian metrics g and g' respectively. A local diffeomorphism $\Phi : S \to S'$ is called ***angle-preserving*** or ***conformal*** if there exists a positive function $c : S \to \mathbb{R}$ such that

$$\Phi^* g' = c \cdot g.$$

By the previous lemma this means that the angle between $X, Y \in T_pS$ is the same as the angle between $d_p\Phi(X), d_p\Phi(Y) \in T_{\Phi(p)}S$. Note that the *conformal factor* c will, in general, depend on $p \in S$. Since g and Φ^*g' are smooth, c is a smooth function on S.

A local parametrisation (U, F, V) of a regular surface S with Riemannian metric g will be called conformal if it is conformal as a map $F : U \to V$, where U carries the Euclidean metric. In other words, there exists a positive function $c : U \to \mathbb{R}$ such that

$$(g_{ij}(u^1, u^2)) = c(u^1, u^2) \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The most common way to specify a location on the surface of the earth is to give its longitude λ and latitude φ. For example, Greenwich (England) has longitude $\lambda = 0$ and latitude $\varphi = 51°28'44'' \approx 0.286 \cdot \pi$. Mathematically, this means that we use the parametrisation introduced in example 3.3.3:

$$F : \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \times (0, 2\pi) \to \mathbb{R}^3,$$

$$F(\varphi, \lambda) = \begin{pmatrix} \cos(\varphi) \cdot \cos(\lambda) \\ \cos(\varphi) \cdot \sin(\lambda) \\ \sin(\varphi) \end{pmatrix}.$$

The computation of the first fundamental form gave us

$$\left(g_{ij}(\varphi, \lambda)\right)_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & \cos^2(\varphi) \end{pmatrix}$$

and hence

$$dA = \cos(\varphi)\, d\varphi\, d\lambda.$$

Thus longitude and latitude do not give rise to an area- or angle-preserving map.

Let us instead look at so-called *azimuthal* charts of the sphere. They are obtained by projecting the sphere onto one of its tangent planes. For simplicity, we will always use the tangent plane to the "north pole" $e_3 = (0,0,1)^\top$, but any other point would do equally well. The simplest such chart is the *orthographic projection* or *parallel projection*.

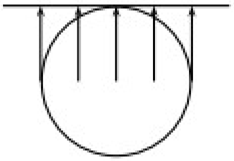

Here we have $U = \{(x,y)^\top \mid x^2+y^2<1\}$, $V = \{(x,y,z)^\top \in \mathbb{R}^3 \mid z>0\}$ and $F(x,y) = (x,y,\sqrt{1-x^2-y^2})^\top$, cf. example 3.1.5. Note that the local parametrisation F is the inverse of the projection. One computes

$$(g_{ij}(x,y)) = \frac{1}{1-x^2-y^2}\begin{pmatrix} 1-y^2 & xy \\ xy & 1-x^2 \end{pmatrix}$$

and hence

$$dA = \sqrt{\det(g_{ij}(x,y))}\, dx\, dy = (1-x^2-y^2)^{-1/2}\, dx\, dy.$$

Thus the orthographic projection is neither area- nor angle-preserving. It shows the earth as it is seen from outer space.

The next azimuthal chart is the *gnomonic projection* or *central projection*.

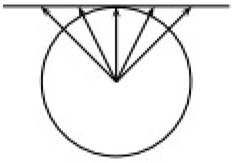

Here we have $U = \mathbb{R}^2$, $V = \{(x,y,z)^\top \in \mathbb{R}^3 \mid z>0\}$, and $F(x,y) = (1/\sqrt{1+x^2+y^2})(x,y,1)^\top$. One computes

$$(g_{ij}(x,y)) = \frac{1}{(1+x^2+y^2)^2}\begin{pmatrix} 1+y^2 & -xy \\ -xy & 1+x^2 \end{pmatrix}$$

and

$$dA = (1+x^2+y^2)^{-3/2}\, dx\, dy.$$

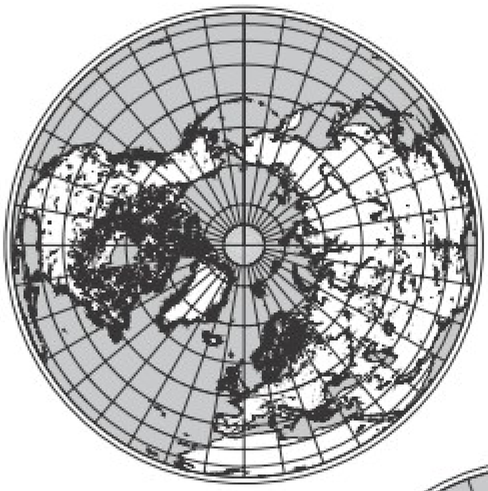

orthographic projection

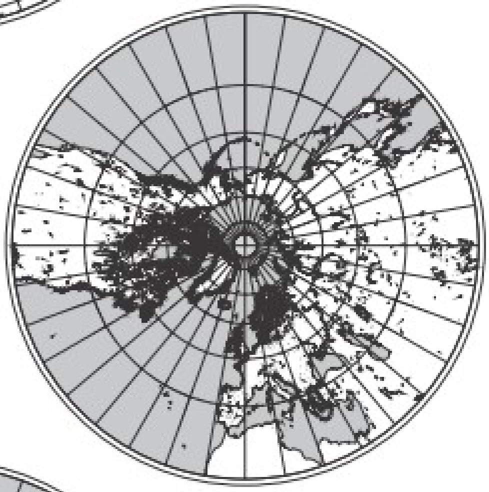

gnomonic projection

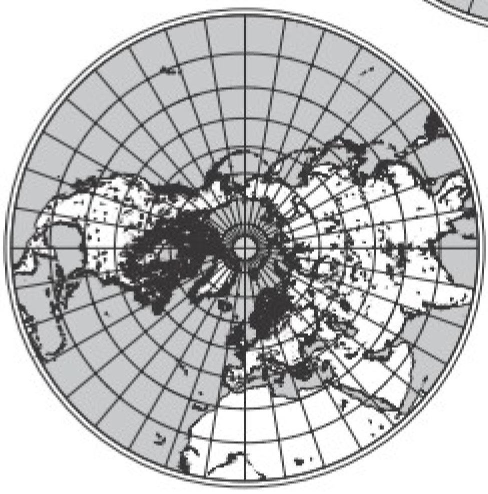

stereographic projection

Hence this chart is also neither area- nor angle-preserving but it has the pleasing property that the geodesics on the sphere, the great circles, are represented in the chart by straight lines. This is because a great circle is the intersection of S^2 with a plane E through the origin which is projected to the intersection of E with the tangent plane, a straight line. Thus this type of map is well suited for finding the shortest route from one point to another.

If we change the projection point from the centre of the sphere to the south pole we obtain the *stereographic projection*.

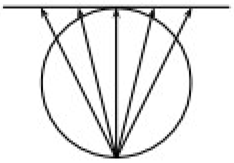

In this case

$$U = \mathbb{R}^2,\ V = \mathbb{R}^3 \setminus \left\{(0,0,-1)^\top\right\},\ F(x,y) = \frac{1}{4+x^2+y^2}(4x, 4y, 4-x^2-y^2)^\top,$$

$$(g_{ij}(x,y)) = \frac{16}{(4+x^2+y^2)^2}\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$$

and

$$dA = \frac{16}{(4+x^2+y^2)^2}\,dx\,dy.$$

Thus the stereographic projection is conformal but not area-preserving.

All three projections considered so far satisfy $(g_{ij}(0,0)) = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$. Thus the differential of the local parametrisation is an isometry at $(0,0,)^\top$. The distortion of the map is small near the north pole (or whatever point has been chosen as the projection centre).

To get a map with small distortion in a whole neighbourhood of the equator (or any other great circle) one wraps a cylinder around the equator, projects the sphere onto the cylinder and then unwraps it.

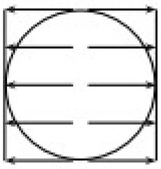

We then have $U = (0,2\pi) \times (-1,1)$, $V = \mathbb{R}^3 \setminus \left\{(x,0,z)^\top \mid x > 0\right\}$, and

$$F(\varphi,h) = \begin{pmatrix}\cos(\varphi)\cdot\sqrt{1-h^2}\\ \sin(\varphi)\cdot\sqrt{1-h^2}\\ h\end{pmatrix}.$$

One computes

$$(g_{ij}(\varphi, h)) = \begin{pmatrix} 1-h^2 & 0 \\ 0 & \dfrac{1}{1-h^2} \end{pmatrix}.$$

Thus this map is not conformal but $dA = d\varphi\, dh$, hence it is area-preserving.

Let us modify this example. We introduce a new parameter x and let the height h be a function of x yet to be determined, i.e.

$$F(\varphi, x) = \begin{pmatrix} \cos(\varphi) \cdot \sqrt{1-h(x)^2} \\ \sin(\varphi) \cdot \sqrt{1-h(x)^2} \\ h(x) \end{pmatrix}.$$

Then we get

$$(g_{ij}(\varphi, x)) = \begin{pmatrix} 1-h(x)^2 & 0 \\ 0 & \dfrac{h'(x)^2}{1-h(x)^2} \end{pmatrix}.$$

We see that we need $h' \neq 0$. We may and will assume that $h' > 0$ because otherwise we may simply replace $h(x)$ by $h(-x)$. In order to get a conformal chart we need

$$1 - h(x)^2 = \frac{h'(x)^2}{1-h(x)^2},$$

i.e.

$$h'(x) = 1 - h(x)^2.$$

The solution to this ordinary differential equation with initial condition $h(0) = 0$ is

$$h(x) = \tanh(x).$$

The resulting geographical map is known as *Mercator's projection* named after the Flemish geographer Gheert Cremer (1512–1594), who latinised his name to Gerardus Mercator.

Exercise 4.34 What can you say about the solutions to this ordinary differential equation with other initial values $h(0)$?

There are many more geographical maps. For example, in order to get a good map representation of a neighbourhood of a circle of latitude other than the equator, one can wrap a cone around the sphere along a small circle and then unwrap it. The interested reader should consult a book on cartography such as [28].

Exercise 4.35 Show that there are no geographical maps that are at the same time conformal and area-preserving.

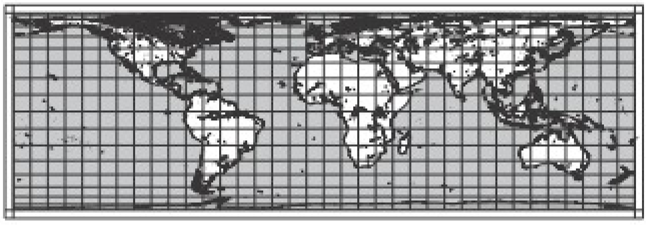

cylindrical equal-area projection

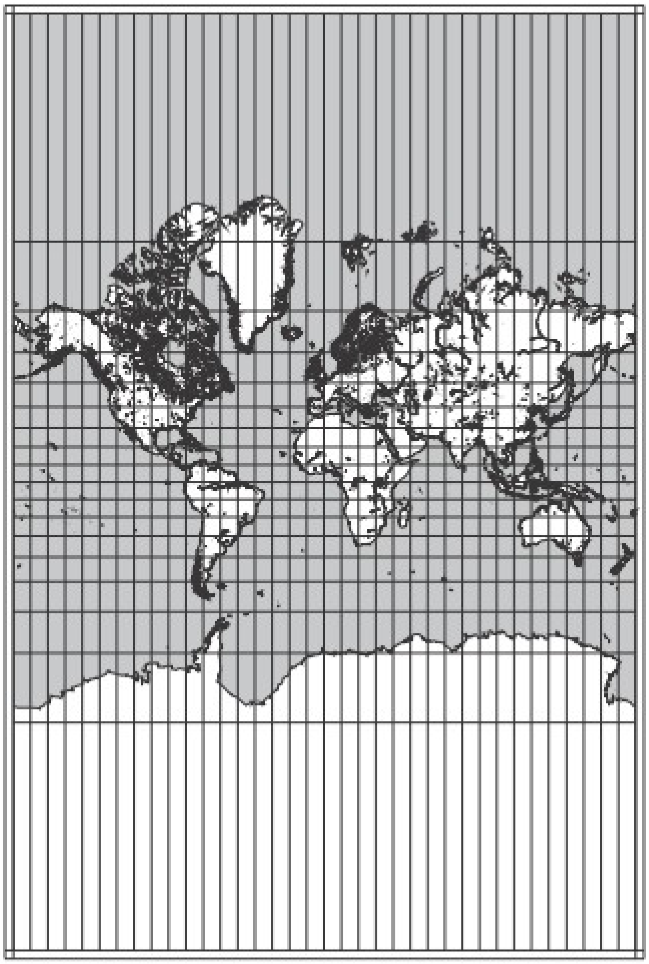

Mercator's projection

4.11 Further models of hyperbolic geometry

We now study charts of the hyperbolic plane. In contrast to the spherical case it is possible to find charts covering the whole hyperbolic plane at once. We will call such charts ***models of hyperbolic geometry***. If (U, F, V) is such a chart with $\mathbb{M}_{-1} \subset V$, then F is a diffeomorphism from U to $\mathbb{M}_{-1}$ and all investigations of hyperbolic geometry can be performed on U using the functions $g_{ij} : U \to \mathbb{R}$ describing the hyperbolic Riemannian metric in this chart. Since the hyperbolic metric was obtained by restricting the Minkowski scalar product of $\mathbb{R}^3$ instead of the Euclidean scalar product, we have

$$g_{ij}\left(u^1, u^2\right) = \left\langle \frac{\partial F}{\partial u^i}\left(u^1, u^2\right), \frac{\partial F}{\partial u^j}\left(u^1, u^2\right)\right\rangle_{-1}.$$

The first chart is obtained by central projection with respect to the origin of the unit disc in $T_{e_1}\mathbb{M}_{-1}$ to $\mathbb{M}_{-1}$.

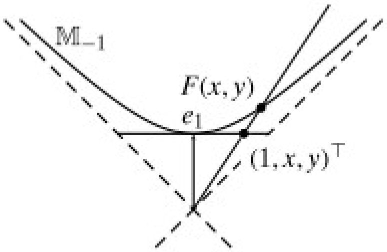

Expressed as formulae, this becomes, $U = \left\{(x, y)^\top \in \mathbb{R}^2 \mid x^2 + y^2 < 1\right\}$ and

$$F(x, y) = \frac{1}{\sqrt{1 - x^2 - y^2}} \begin{pmatrix} 1 \\ x \\ y \end{pmatrix}. \tag{4.29}$$

This chart is known as the ***Klein model***, as the ***Cayley–Klein model***, as the ***projective model***, and also as the ***Beltrami–Klein model*** of hyperbolic geometry. We compute

$$\frac{\partial F}{\partial x} = \left(1 - x^2 - y^2\right)^{-3/2} \begin{pmatrix} x \\ 1 - y^2 \\ xy \end{pmatrix}, \qquad \frac{\partial F}{\partial y} = \left(1 - x^2 - y^2\right)^{-3/2} \begin{pmatrix} y \\ xy \\ 1 - x^2 \end{pmatrix}$$

and hence

$$g_{11}(x, y) = \left\langle \frac{\partial F}{\partial x}, \frac{\partial F}{\partial x} \right\rangle_{-1}$$

$$= \left(1 - x^2 - y^2\right)^{-3} \cdot \left(-x^2 + \left(1 - y^2\right)^2 + (xy)^2\right)$$
$$= \frac{1 - y^2}{\left(1 - x^2 - y^2\right)^2}$$

and similarly for the other coefficients for the metric. We obtain

$$(g_{ij}(x, y)) = \frac{1}{\left(1 - x^2 - y^2\right)^2} \begin{pmatrix} 1 - y^2 & xy \\ xy & 1 - x^2 \end{pmatrix}.$$

Exercise 4.36 Show that the hyperbolic plane has infinite area.

The metric does not look particularly simple in the Klein model but this model has one very pleasing feature: geodesics in $\mathbb{M}_{-1}$ correspond to straight line segments in U. This is clear from the construction of F. The geodesics in $\mathbb{M}_{-1}$ are the intersections of $\mathbb{M}_{-1}$ with planes E containing the origin. Under the central projection this is mapped to the intersection of E with the unit disc in $T_{e_1}\mathbb{M}_{-1}$. The intersection of the two planes E and $T_{e_1}\mathbb{M}_{-1}$ is a straight line.

This allows a simple discussion of the parallel axiom. Keeping the terminology and the notation from the first chapter we now take the hyperbolic plane $\mathbb{M}_{-1}$ as the set of points, the set of "straight lines" will be the set of traces of all geodesics,

$$\mathscr{P} := \mathbb{M}_{-1},$$
$$\mathscr{G} := \{\mathbb{M}_{-1} \cap E |\, E \subset \mathbb{R}^3 \text{ is a two-dimensional subspace}\} - \{\emptyset\}.$$

The incidence relation is the usual set-theoretical inclusion. We immediately see that the incidence axioms I_1–I_4 are satisfied. Unlike in spherical geometry, in which geodesics are great circles, there is an obvious useful definition for when one point lies between two others on the hyperbolas in hyperbolic geometry. The validity of the ordering axioms A_1–A_5 is then easy to see. For the definition of congruence of lengths and angles we use isometries as in the discussion of the Cartesian model of Euclidean geometry. This time we use the group G_{-1} of isometries on $\mathbb{M}_{-1}$ instead of the Euclidean group. It is then not difficult to show the validity of the congruence axioms. By theorem 1.1.9 we already know that parallels therefore exist. The completeness axioms can be derived in a way similar to the one used for the Cartesian model.

Only the parallel axiom is not valid, since parallels are not unique. This is easily seen in the Klein model. Given a line segment L in U there is an infinity of line segments not intersecting L but having one point in common.

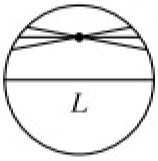

Hyperbolic geometry shows that the parallel axiom does not follow from the other axioms of Euclidean geometry.

Note that the Klein model is the hyperbolic analogue of the gnomonic chart of the sphere. Also compare their metric coefficients. Now we move on to the hyperbolic analogue of the stereographic projection in spherical geometry. In principle, we only have to change the projection point from the origin to $-e_1$ as in spherical geometry. To get slightly simpler formulae we do not project to $T_{e_1}\mathbb{M}_{-1}$ but to the parallel plane spanned by e_2 and e_3. The difference is only a stretching by a factor of 2.

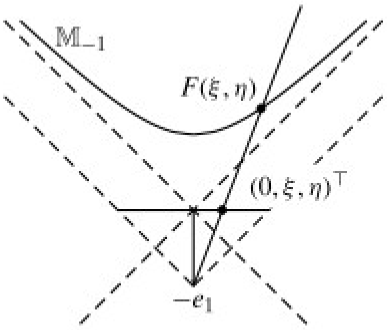

Set up in this way U is the unit disc as for the Klein model and one easily computes that

$$F(\xi,\eta) = \frac{1}{1-\xi^2-\eta^2}\begin{pmatrix}1+\xi^2+\eta^2\\ 2\xi\\ 2\eta\end{pmatrix}.$$

A calculation similar to the one for the Klein model yields

$$(g_{ij}(\xi,\eta)) = \frac{4}{(1-\xi^2-\eta^2)^2}\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}.$$

Thus this model of hyperbolic geometry is conformal just like the stereographic projection. So it comes at no surprise that this model is known as the ***conformal disc model*** and also as the ***Poincaré disc model***. It was popularised by the Dutch artist Maurits Cornelis Escher (1898–1972) in his famous woodcuts such as "Circle Limit I" from 1958.[1]

---

[1] M. C. Escher's "Circle Limit I" © 2009 The M. C. Escher Company – Holland. All rights reserved. www.mcescher.com

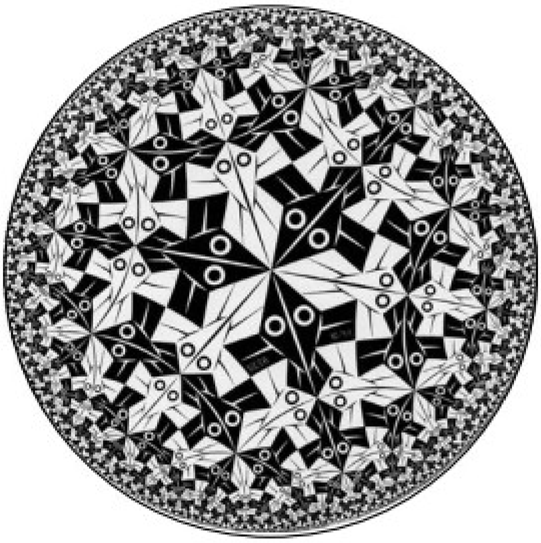

The straight line segments through the centre of the disc correspond to the geodesics of $\mathbb{M}_{-1}$ given by intersections with those planes that contain both the origin and e_1. The other geodesics correspond to circular arcs in the disc meeting its boundary perpendicularly. Escher's woodcut displays some of them. The black fish are all congruent with respect to the hyperbolic metric. In particular, their hyperbolic areas are all the same and similarly for the white fish.

One may wonder what this picture would look like in the Klein model. The circular arcs must be replaced by straight line segments. Indeed, Escher's Circle Limit I transformed to the Klein model looks like this:

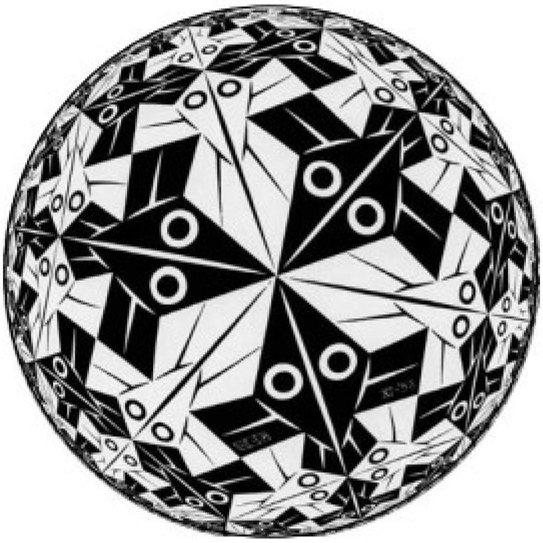

There is another interesting and important model. We construct it as follows. We start with the Klein model and do a parallel projection to the upper hemisphere, i.e. the point $(x, y)^\top$ is mapped to $(\sqrt{1-x^2-y^2}, x, y)^\top$. This is the transformation that gave rise to the orthographic chart in cartography. The straight line segments in the disc are mapped to circular arcs meeting the equator of the hemisphere at a right angle.

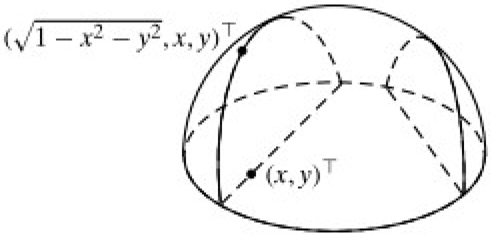

Now we choose a point on the equator of the hemisphere, say e_2, and perform a stereographic projection with respect to e_2 onto the tangent plane of the sphere at the antipodal point $-e_2$. This maps the upper hemisphere onto the upper half-plane. The spherical arcs are mapped to spherical arcs in the half-plane except for those arcs in the hemisphere that contain e_2. These are projected onto straight lines. Both the spherical arcs and the straight lines meet the boundary of the half-plane at a right angle.

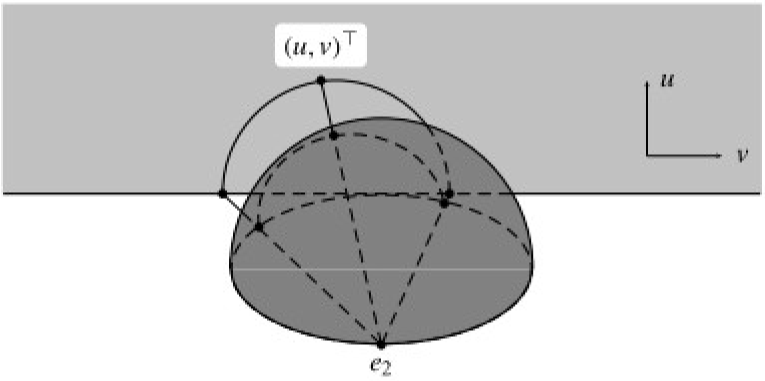

The image of $(\sqrt{1-x^2-y^2}, x, y)^\top$ under this stereographic projection is easily computed to be

$$\begin{pmatrix} u \\ v \end{pmatrix} = \frac{1}{1-x} \begin{pmatrix} 2\sqrt{1-x^2-y^2} \\ 2y \end{pmatrix}.$$

Solving for $(x, y)^\top$ yields

$$\begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{u^2+v^2+4} \begin{pmatrix} u^2+v^2-4 \\ 4v \end{pmatrix}.$$

Plugging this into the parametrisation (4.29) of the Klein model gives us the ***Poincaré half-plane model***

$$F(u,v) = \frac{1}{4u}\begin{pmatrix} u^2+v^2+4 \\ u^2+v^2-4 \\ 4v \end{pmatrix}$$

with $U = \{(u,v)^\top \mid u > 0\}$. As before one computes

$$(g_{ij}(u,v)) = \frac{1}{u^2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Hence the Poincaré half-plane model is also conformal. Escher's woodcut transformed to this model now looks like this:

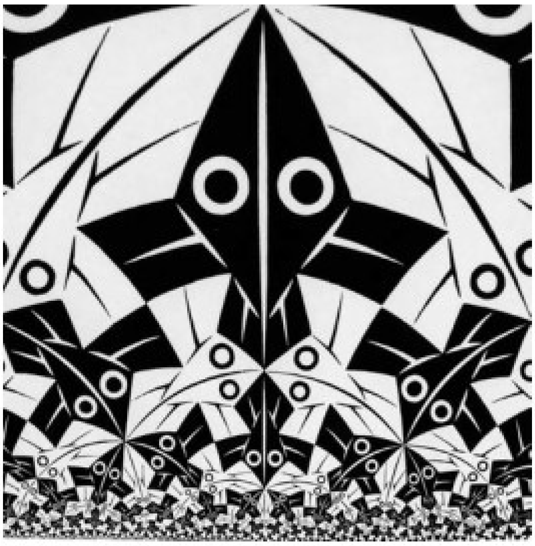

Exercise 4.37 The grey region Δ in the Poincaré half-plane model shown in the figure below is a geodesic triangle with vertices at infinity. Compute its area.

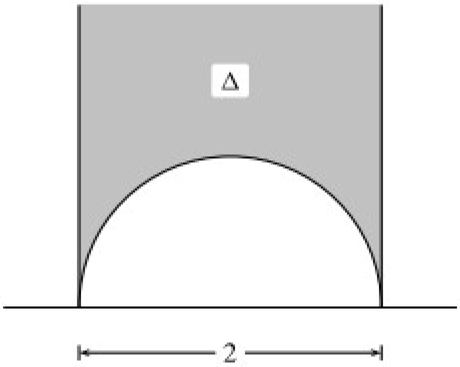

!"#" $ % % & ' ( )) * +) , -)
.) @) + !) * + 4

,9
*) /T
( ) <
% 7</

5- /<2 7 3

7
%))
E)4 8 %))44) 4 9) -
? % A)) " - + = 8
)))44))) - 9J
)))) E 2 # 3)))
4) + (
9+ P)))
8 J)) 4 4) % 4) + (
9+ - %)
-)4) ') ( )
6 /) % +) 4)

7 #))

7

Curvature of surfaces

In this chapter, we discuss several approaches to the problem of measuring how 'curved' a surface is. Although they use quite different methods, we show that each of the approaches leads to the same geometric object: the *second fundamental form* of a surface. It turns out (see Theorem 10.1.3) that a surface is determined up to an isometry of $\mathbb{R}^3$ by its first and second fundamental forms, just as a unit-speed plane curve is determined up to an isometry of $\mathbb{R}^2$ by its signed curvature.

Throughout this chapter we shall work with *oriented* surfaces. Recall from Section 4.5 that every surface *patch* is oriented.

7.1 The second fundamental form

In our first attempt to define the curvature of a surface, we imitate the discussion at the beginning of Section 2.1, which leads to the definition of the curvature of a curve. Suppose then that $\boldsymbol{\sigma}$ is a surface patch in $\mathbb{R}^3$ with standard unit normal $\mathbf{N}$. As the parameters (u, v) of $\boldsymbol{\sigma}$ change to $(u + \Delta u, v + \Delta v)$, the surface moves away from the plane through $\boldsymbol{\sigma}(u, v)$ parallel to the tangent plane by a distance

$$(\boldsymbol{\sigma}(u + \Delta u, v + \Delta v) - \boldsymbol{\sigma}(u, v)) \cdot \mathbf{N}.$$

Andrew Pressley, *Elementary Differential Geometry: Second Edition*,
Springer Undergraduate Mathematics Series, DOI 10.1007/978-1-84882-891-9_7,
© Springer-Verlag London Limited 2010

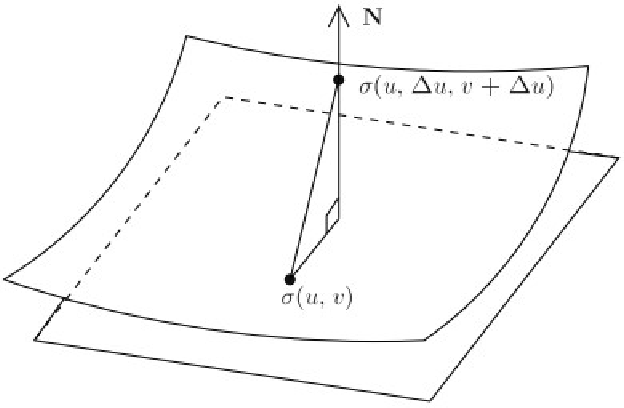

By the two variable form of Taylor's theorem, $\boldsymbol{\sigma}(u+\Delta u, v+\Delta v) - \boldsymbol{\sigma}(u,v)$ is equal to

$$\boldsymbol{\sigma}_u \Delta u + \boldsymbol{\sigma}_v \Delta v + \frac{1}{2}\left(\boldsymbol{\sigma}_{uu}(\Delta u)^2 + 2\boldsymbol{\sigma}_{uv}\Delta u \Delta v + \boldsymbol{\sigma}_{vv}(\Delta v)^2\right) + \text{ remainder},$$

where (remainder)/$((\Delta u)^2 + (\Delta v)^2)$ tends to zero as $(\Delta u)^2 + (\Delta v)^2$ tends to zero. Now $\boldsymbol{\sigma}_u$ and $\boldsymbol{\sigma}_v$ are tangent to the surface, hence perpendicular to $\mathbf{N}$, so the deviation of $\boldsymbol{\sigma}$ from its tangent plane is

$$\frac{1}{2}\left(L(\Delta u)^2 + 2M\Delta u \Delta v + N(\Delta v)^2\right) + \text{remainder}, \tag{7.1}$$

where

$$L = \boldsymbol{\sigma}_{uu} \cdot \mathbf{N}, \quad M = \boldsymbol{\sigma}_{uv} \cdot \mathbf{N}, \quad N = \boldsymbol{\sigma}_{vv} \cdot \mathbf{N}. \tag{7.2}$$

Comparing Eq. 7.2 with Eq. 7.1, we see that the expression

$$L(\Delta u)^2 + 2M\Delta u \Delta v + N(\Delta v)^2$$

is the analogue for the surface of the curvature term $\kappa(\Delta t)^2$ in the case of a curve.

One calls the expression

$$Ldu^2 + 2Mdudv + Ndv^2 \tag{7.3}$$

the *second fundamental form* of the surface patch $\boldsymbol{\sigma}$. It clearly resembles the first fundamental form of a surface patch, and we can make sense of it in the same way, by interpreting du and dv as linear maps as in Section 6.1. Imitating the discussion there, we define a symmetric bilinear form on the tangent plane by

$$\langle \mathbf{v}, \mathbf{w} \rangle = Ldu(\mathbf{v})du(\mathbf{w}) + M(du(\mathbf{v})dv(\mathbf{w}) + du(\mathbf{w})dv(\mathbf{v})) + Ndv(\mathbf{v})dv(\mathbf{w})$$

(cf. Exercise 6.1.3). In the next section, we shall give this form an appealing geometric interpretation, and extend its definition to an arbitrary oriented surface.

Example 7.1.1

Consider the plane

$$\boldsymbol{\sigma}(u,v) = \mathbf{a} + u\mathbf{p} + v\mathbf{q}$$

(see Example 4.1.2). Since $\boldsymbol{\sigma}_u = \mathbf{p}$ and $\boldsymbol{\sigma}_v = \mathbf{q}$ are constant vectors, we have $\boldsymbol{\sigma}_{uu} = \boldsymbol{\sigma}_{uv} = \boldsymbol{\sigma}_{vv} = \mathbf{0}$. Hence, the second fundamental form of a plane is zero.

Example 7.1.2

Consider a surface of revolution (Example 5.3.2)

$$\boldsymbol{\sigma}(u,v) = (f(u)\cos v, f(u)\sin v, g(u));$$

we assume as usual that $f(u) > 0$ for all values of u and that the profile curve $u \mapsto (f(u), 0, g(u))$ is unit-speed, i.e., $\dot{f}^2 + \dot{g}^2 = 1$ (a dot denoting d/du). Then:

$$\begin{aligned}
&\boldsymbol{\sigma}_u = (\dot{f}\cos v, \dot{f}\sin v, \dot{g}), \quad \boldsymbol{\sigma}_v = (-f\sin v, f\cos v, 0),\\
&\quad\therefore \quad \boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v = (-f\dot{g}\cos v, -f\dot{g}\sin v, f\dot{f}),\\
&\quad\therefore \quad \parallel \boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v \parallel = f \quad (\text{since } \dot{f}^2 + \dot{g}^2 = 1),\\
&\therefore \quad \mathbf{N} = \frac{\boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v}{\parallel \boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v \parallel} = (-\dot{g}\cos v, -\dot{g}\sin v, \dot{f}),\\
&\qquad\qquad \boldsymbol{\sigma}_{uu} = (\ddot{f}\cos v, \ddot{f}\sin v, \ddot{g}),\\
&\qquad\qquad \boldsymbol{\sigma}_{uv} = (-\dot{f}\sin v, \dot{f}\cos v, 0),\\
&\qquad\qquad \boldsymbol{\sigma}_{vv} = (-f\cos v, -f\sin v, 0),\\
&\therefore \quad L = \boldsymbol{\sigma}_{uu}\cdot\mathbf{N} = \dot{f}\ddot{g} - \ddot{f}\dot{g}, \ M = \boldsymbol{\sigma}_{uv}\cdot\mathbf{N} = 0, \ N = \boldsymbol{\sigma}_{vv}\cdot\mathbf{N} = f\dot{g},
\end{aligned}$$

so the second fundamental form is

$$(\dot{f}\ddot{g} - \ddot{f}\dot{g})du^2 + f\dot{g}dv^2.$$

For the unit sphere S^2 in latitude–longitude coordinates (Example 4.1.4), $u = \theta$, $v = \varphi$, $f(\theta) = \cos\theta$, $g(\theta) = \sin\theta$, giving the second fundamental form of S^2 as

$$d\theta^2 + \cos^2\theta\, d\varphi^2.$$

Note that this is the same as the *first* fundamental form of S^2 (see Example 6.1.3; the reason for this will appear in Section 8.2).

If the surface is the unit cylinder, we can take $f(u) = 1$, $g(u) = u$ (again, the conditions $f > 0$ and $\dot{f}^2 + \dot{g}^2 = 1$ are satisfied). This gives $L = M = 0, \ N = 1$, so the second fundamental form of the cylinder is dv^2.

EXERCISES

7.1.1 Compute the second fundamental form of the elliptic paraboloid

$$\boldsymbol{\sigma}(u,v) = (u, v, u^2 + v^2).$$

7.1.2 Suppose that the second fundamental form of a surface patch $\boldsymbol{\sigma}$ is zero everywhere. Prove that $\boldsymbol{\sigma}$ is an open subset of a plane. This is the analogue for surfaces of the theorem that a curve with zero curvature everywhere is part of a straight line.

7.1.3 Let a surface patch $\tilde{\boldsymbol{\sigma}}(\tilde{u}, \tilde{v})$ be a reparametrization of a surface patch $\boldsymbol{\sigma}(u,v)$ with reparametrization map $(u,v) = \Phi(\tilde{u}, \tilde{v})$. Prove that

$$\begin{pmatrix} \tilde{L} & \tilde{M} \\ \tilde{M} & \tilde{N} \end{pmatrix} = \pm J^t \begin{pmatrix} L & M \\ M & N \end{pmatrix} J,$$

where J is the Jacobian matrix of Φ and we take the plus sign if $\det(J) > 0$ and the minus sign if $\det(J) < 0$. Deduce from Exercise 6.1.4 that the second fundamental form of a surface patch is unchanged by a reparametrization of the patch which preserves its orientation.

7.1.4 What is the effect on the second fundamental form of a surface of applying an isometry of $\mathbb{R}^3$? Or a dilation?

7.2 The Gauss and Weingarten maps

Our second approach to defining the curvature of an oriented surface $\mathcal{S}$ is to consider its unit normal $\mathbf{N}$. The way that $\mathbf{N}$ varies clearly reflects the way in which $\mathcal{S}$ curves: $\mathbf{N}$ varies rapidly near a point at which the surface is highly curved and slowly where the surface is only slightly curved. If $\mathcal{S}$ is a plane, $\mathbf{N}$ is the same at all points of $\mathcal{S}$, i.e., $\mathbf{N}$ is a constant, and the curvature should be zero.

The values of $\mathbf{N}$ at the points of $\mathcal{S}$ are recorded by its *Gauss map* $\mathcal{G}_{\mathcal{S}}$ (or just $\mathcal{G}$ if there is no doubt as to which surface is intended). This is the map from $\mathcal{S}$ to the unit sphere S^2 that assigns to any point $\mathbf{p} \in \mathcal{S}$ the point $\mathbf{N}_{\mathbf{p}} \in S^2$, where $\mathbf{N}_{\mathbf{p}}$ is the unit normal of $\mathcal{S}$ at $\mathbf{p}$.

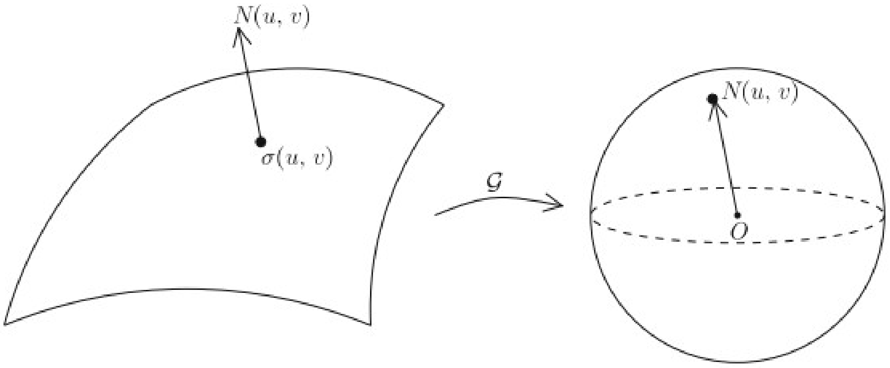

The rate at which $\mathbf{N}$ varies across $\mathcal{S}$ is measured by the derivative of $\mathcal{G}$:

$$D_{\mathbf{p}}\mathcal{G} : T_{\mathbf{p}}\mathcal{S} \to T_{\mathcal{G}}(\mathbf{p})S^2.$$

Now, the tangent plane at a point $\mathbf{q} \in S^2$ is the plane passing through the origin perpendicular to $\mathbf{q}$. Thus, $T_{\mathcal{G}}(\mathbf{p})S^2$ is the plane through the origin perpendicular to $\mathbf{N}_{\mathbf{p}}$, in other words, $T_{\mathbf{p}}\mathcal{S}$. Thus, the derivative of $\mathcal{G}$ is a linear map from the tangent plane of $\mathcal{S}$ to itself:

$$D_{\mathbf{p}}\mathcal{G} : T_{\mathbf{p}}\mathcal{S} \to T_{\mathbf{p}}\mathcal{S}.$$

Definition 7.2.1

Let $\mathbf{p}$ be a point of a surface $\mathcal{S}$. The *Weingarten map* $\mathcal{W}_{\mathbf{p},\mathcal{S}}$ of $\mathcal{S}$ at $\mathbf{p}$ is defined by

$$\mathcal{W}_{\mathbf{p},\mathcal{S}} = -D_{\mathbf{p}}\mathcal{G}. \tag{7.4}$$

The *second fundamental form* of $\mathcal{S}$ at $\mathbf{p} \in \mathcal{S}$ is the bilinear form on $T_{\mathbf{p}}\mathcal{S}$ given by

$$\langle\!\langle \mathbf{v}, \mathbf{w} \rangle\!\rangle_{\mathbf{p},\mathcal{S}} = \langle \mathcal{W}_{\mathbf{p},\mathcal{S}}(\mathbf{v}), \mathbf{w} \rangle_{\mathbf{p},\mathcal{S}}, \quad \mathbf{v}, \mathbf{w} \in T_{\mathbf{p}}\mathcal{S}.$$

The minus sign is introduced in Eq. 7.4 as a matter of convention (it will reduce the total number of minus signs later). We shall often omit the subscripts $\mathbf{p}$ and $\mathcal{S}$ from the Weingarten map if there is no danger of confusion.

The bilinearity asserted in this definition is easy to check. Let $\mathbf{v}_1, \mathbf{v}_2, \mathbf{w} \in T_{\mathbf{p}}\mathcal{S}$, $\lambda_1, \lambda_2 \in \mathbb{R}$. Then,

$$\begin{aligned}
\langle\!\langle \lambda_1\mathbf{v}_1 + \lambda_2\mathbf{v}_2, \mathbf{w} \rangle\!\rangle &= \langle \mathcal{W}(\lambda_1\mathbf{v}_1 + \lambda_2\mathbf{v}_2), \mathbf{w} \rangle \\
&= \langle \lambda_1\mathcal{W}(\mathbf{v}_1) + \lambda_2\mathcal{W}(\mathbf{v}_2), \mathbf{w} \rangle \\
&= \lambda_1\langle \mathcal{W}(\mathbf{v}_1), \mathbf{w} \rangle + \lambda_2\langle \mathcal{W}(\mathbf{v}_2), \mathbf{w} \rangle \\
&= \lambda_1\langle\!\langle \mathbf{v}_1, \mathbf{w} \rangle\!\rangle + \lambda_2\langle\!\langle \mathbf{v}_2, \mathbf{w} \rangle\!\rangle,
\end{aligned}$$

where we used the linearity of $\mathcal{W}$ in passing from the first line to the second and the bilinearity of $\langle\ ,\ \rangle$ in passing from the second line to the third. This proves that $\langle \mathbf{v}, \mathbf{w}\rangle$ is a linear function of $\mathbf{v}$ for each fixed $\mathbf{w}$. The proof that it is linear in $\mathbf{w}$ for each fixed $\mathbf{v}$ is similar but easier.

Proposition 7.2.2

Let $\mathbf{p}$ be a point of a surface $\mathcal{S}$, let $\boldsymbol{\sigma}(u,v)$ be a surface patch of $\mathcal{S}$ with $\mathbf{p}$ in its image, and let $Ldu^2 + 2Mdudv + Ndv^2$ be the second fundamental form of $\boldsymbol{\sigma}$ defined in Section 7.1. Then, for any $\mathbf{v}, \mathbf{w} \in T_{\mathbf{p}}\mathcal{S}$,

$$\langle \mathbf{v}, \mathbf{w}\rangle = Ldu(\mathbf{v})du(\mathbf{w}) + M(du(\mathbf{v})dv(\mathbf{w}) + du(\mathbf{w})dv(\mathbf{v})) + Ndv(\mathbf{v})dv(\mathbf{w}). \quad (7.5)$$

To prove this we shall need the following lemma, which will also be used elsewhere.

Lemma 7.2.3

Let $\boldsymbol{\sigma}(u,v)$ be a surface patch with standard unit normal $\mathbf{N}(u,v)$. Then,

$$\mathbf{N}_u \cdot \boldsymbol{\sigma}_u = -L, \quad \mathbf{N}_u \cdot \boldsymbol{\sigma}_v = \mathbf{N}_v \cdot \boldsymbol{\sigma}_u = -M, \quad \mathbf{N}_v \cdot \boldsymbol{\sigma}_v = -N,$$

where L, M and N are as defined in Eq. 7.2.

Proof

Since $\boldsymbol{\sigma}_u$ and $\boldsymbol{\sigma}_v$ are tangent vectors to the surface patch,

$$\mathbf{N} \cdot \boldsymbol{\sigma}_u = \mathbf{N} \cdot \boldsymbol{\sigma}_v = 0.$$

Differentiating these equations with respect to u and v gives

$$\begin{aligned} \mathbf{N}_u \cdot \boldsymbol{\sigma}_u &= -\mathbf{N} \cdot \boldsymbol{\sigma}_{uu} = -L, \quad & \mathbf{N}_v \cdot \boldsymbol{\sigma}_u &= -\mathbf{N} \cdot \boldsymbol{\sigma}_{uv} = -M, \\ \mathbf{N}_u \cdot \boldsymbol{\sigma}_v &= -\mathbf{N} \cdot \boldsymbol{\sigma}_{uv} = -M, \quad & \mathbf{N}_v \cdot \boldsymbol{\sigma}_v &= -\mathbf{N} \cdot \boldsymbol{\sigma}_{vv} = -N. \end{aligned} \qquad \square$$

Proof 7.2.2 Since both sides of the equation in the statement of the proposition define bilinear forms on $T_{\mathbf{p}}\mathcal{S}$, it suffices to verify that they agree when $\mathbf{v}$ and $\mathbf{w}$ are $\boldsymbol{\sigma}_u$ or $\boldsymbol{\sigma}_v$. Recalling that $du(\boldsymbol{\sigma}_u) = dv(\boldsymbol{\sigma}_v) = 1$, $du(\boldsymbol{\sigma}_v) = dv(\boldsymbol{\sigma}_u) = 0$, we have to prove that

$$\langle \boldsymbol{\sigma}_u, \boldsymbol{\sigma}_u\rangle = L, \quad \langle \boldsymbol{\sigma}_u, \boldsymbol{\sigma}_v\rangle = \langle \boldsymbol{\sigma}_v, \boldsymbol{\sigma}_u\rangle = M, \quad \langle \boldsymbol{\sigma}_v, \boldsymbol{\sigma}_v\rangle = N. \quad (7.6)$$

Let $\boldsymbol{\sigma}(u_0, v_0) = \mathbf{p}$. Then, with the derivatives evaluated at (u_0, v_0),

$$\mathcal{W}(\boldsymbol{\sigma}_u) = -\left.\frac{d}{du}\right|_{u=u_0} \mathcal{G}(\boldsymbol{\sigma}(u, v_0)) = -\left.\frac{d}{du}\right|_{u=u_0} \mathbf{N}(u, v_0) = -\mathbf{N}_u,$$

where $\mathbf{N}$ is the standard unit normal of $\boldsymbol{\sigma}$. Similarly, $\mathcal{W}(\boldsymbol{\sigma}_v) = -\mathbf{N}_v$. Hence,

$$\langle \boldsymbol{\sigma}_u, \boldsymbol{\sigma}_u \rangle = \langle \mathcal{W}(\boldsymbol{\sigma}_u), \boldsymbol{\sigma}_u \rangle = -\mathbf{N}_u \cdot \boldsymbol{\sigma}_u,$$

which is equal to L by Lemma 7.2.3. The other equations in (7.6) are proved similarly. ☐

Since the formula (7.5) for $\langle \mathbf{v}, \mathbf{w} \rangle$ is obviously unchanged when $\mathbf{v}$ and $\mathbf{w}$ are interchanged, we obtain

Corollary 7.2.4

The second fundamental form is a symmetric bilinear form. Equivalently, the Weingarten map is self-adjoint.

EXERCISES

7.2.1 Calculate the Gauss map $\mathcal{G}$ of the paraboloid $\mathcal{S}$ with equation $z = x^2 + y^2$. What is the image of $\mathcal{G}$?

7.2.2 Show that the Weingarten map changes sign when the orientation of the surface changes.

7.3 Normal and geodesic curvatures

It is obvious that the shape of a surface influences the curvature of curves on the surface. For example, a curve on a plane or a cylinder can have zero curvature everywhere, but this is not possible for curves on a sphere since no segment of a straight line can lie on a sphere. Thus, another natural way to investigate how much a surface curves is to look at the curvature of curves on the surface. We shall see that this leads, once again, to the second fundamental form of the surface.

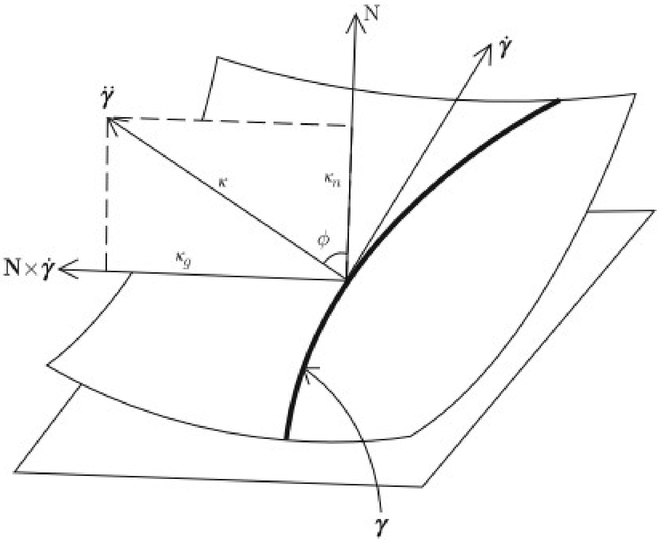
$\mathbf{N}$
$\dot{\boldsymbol{\gamma}}$
$\ddot{\boldsymbol{\gamma}}$
κ
κ_n
ϕ
κ_g
$\mathbf{N}\times\dot{\boldsymbol{\gamma}}$
$\boldsymbol{\gamma}$

If $\boldsymbol{\gamma}$ is a unit-speed curve on an oriented surface $\mathcal{S}$, then $\dot{\boldsymbol{\gamma}}$ is a unit vector and is, by definition, a tangent vector to $\mathcal{S}$. Hence, $\dot{\boldsymbol{\gamma}}$ is perpendicular to the unit normal $\mathbf{N}$ of $\mathcal{S}$, so $\dot{\boldsymbol{\gamma}}$, $\mathbf{N}$ and $\mathbf{N} \times \dot{\boldsymbol{\gamma}}$ are mutually perpendicular unit vectors. Again since $\boldsymbol{\gamma}$ is unit-speed, $\ddot{\boldsymbol{\gamma}}$ is perpendicular to $\dot{\boldsymbol{\gamma}}$, and hence is a linear combination of $\mathbf{N}$ and $\mathbf{N} \times \dot{\boldsymbol{\gamma}}$:

$$\ddot{\boldsymbol{\gamma}} = \kappa_n \mathbf{N} + \kappa_g \mathbf{N} \times \dot{\boldsymbol{\gamma}}. \tag{7.7}$$

Definition 7.3.1

The scalars κ_n and κ_g in Eq. 7.7 are called the *normal curvature* and the *geodesic curvature* of $\boldsymbol{\gamma}$, respectively.

Note that κ_n and κ_s both change sign when $\mathbf{N}$ is replaced by $-\mathbf{N}$, so on a general (not necessarily orientable) surface only the magnitudes of κ_n and κ_s are well defined.

Proposition 7.3.2

With the above notation, we have

$$\kappa_n = \ddot{\boldsymbol{\gamma}} \cdot \mathbf{N}, \quad \kappa_g = \ddot{\boldsymbol{\gamma}} \cdot (\mathbf{N} \times \dot{\boldsymbol{\gamma}}), \tag{7.8}$$

$$\kappa^2 = \kappa_n^2 + \kappa_g^2, \tag{7.9}$$

$$\kappa_n = \kappa \cos\psi, \kappa_g = \pm\kappa \sin\psi, \tag{7.10}$$

where κ is the curvature of $\boldsymbol{\gamma}$ and ψ is the angle between $\mathbf{N}$ and the principal normal $\mathbf{n}$ of $\boldsymbol{\gamma}$.

Proof

Equations 7.8 and 7.9 follow from Eq. 7.7 and the fact that $\mathbf{N}$ and $\mathbf{N} \times \dot{\boldsymbol{\gamma}}$ are perpendicular unit vectors. The first equation in (7.10) follows from $\ddot{\boldsymbol{\gamma}} = \kappa\mathbf{n}$, and the second then follows from Eq. 7.9. $\square$

If $\boldsymbol{\gamma}$ is regular, but not necessarily unit-speed, we define the geodesic and normal curvatures of $\boldsymbol{\gamma}$ to be those of a unit-speed reparametrization of $\boldsymbol{\gamma}$ (see Exercise 7.3.1). When a unit-speed parameter t is changed to another such parameter $\pm t + c$, where c is a constant, it is clear that $\kappa_n \mapsto \kappa_n$ and $\kappa_g \mapsto \pm\kappa_g$, so κ_n is well defined for any regular curve, while κ_g is well defined up to sign. Equations 7.9 and 7.10 continue to hold if $\boldsymbol{\gamma}$ is any regular curve.

The following proposition is the most important single fact about the normal curvature, and reveals its relation to the second fundamental form $\langle\!\langle\ ,\ \rangle\!\rangle$.

Proposition 7.3.3

If $\boldsymbol{\gamma}$ is a unit-speed curve on an oriented surface $\mathcal{S}$, its normal curvature is given by

$$\kappa_n = \langle\!\langle \dot{\boldsymbol{\gamma}}, \dot{\boldsymbol{\gamma}} \rangle\!\rangle.$$

If $\boldsymbol{\sigma}$ is a surface patch of $\mathcal{S}$ and $\boldsymbol{\gamma}(t) = \boldsymbol{\sigma}(u(t), v(t))$ is a curve in $\boldsymbol{\sigma}$,

$$\kappa_n = L\dot{u}^2 + 2M\dot{u}\dot{v} + N\dot{v}^2$$

in the notation of Section 7.1.

This result means that two curves which *touch* each other at a point $\mathbf{p}$ of a surface (i.e., which intersect at $\mathbf{p}$ and have parallel tangent vectors at $\mathbf{p}$) have the same normal curvature at $\mathbf{p}$.

Proof

Since $\dot{\boldsymbol{\gamma}}$ is a tangent vector to $\mathcal{S}$, $\mathbf{N}.\dot{\boldsymbol{\gamma}} = 0$. Hence, $\mathbf{N} \cdot \ddot{\boldsymbol{\gamma}} = -\dot{\mathbf{N}} \cdot \dot{\boldsymbol{\gamma}}$ so

$$\kappa_n = \mathbf{N} \cdot \ddot{\boldsymbol{\gamma}} = -\dot{\mathbf{N}} \cdot \dot{\boldsymbol{\gamma}} = \langle \mathcal{W}(\dot{\boldsymbol{\gamma}}), \dot{\boldsymbol{\gamma}} \rangle = \langle\!\langle \dot{\boldsymbol{\gamma}}, \dot{\boldsymbol{\gamma}} \rangle\!\rangle,$$

since

$$\dot{\mathbf{N}} = \frac{d}{dt}\mathcal{G}(\boldsymbol{\gamma}(t)) = -\mathcal{W}(\dot{\boldsymbol{\gamma}}).$$

The second part follows from the first and Proposition 7.2.2. $\square$

It turns out that, while the normal curvature depends on the second fundamental form of the surface, the geodesic curvature κ_g depends only on its *first* fundamental form (see Exercise 7.3.4). But we leave further discussion of κ_g to Chapter 9.

Here is a classical application of Proposition 7.3.3. It takes almost as long to state as to prove.

Proposition 7.3.4 (Meusnier's Theorem)

Let $\mathbf{p}$ be a point of a surface $\mathcal{S}$ and let $\mathbf{v}$ be a unit tangent vector to $\mathcal{S}$ at $\mathbf{p}$. Let Π_θ be the plane containing the line through $\mathbf{p}$ parallel to $\mathbf{v}$ and making an angle θ with the tangent plane $T_\mathbf{p}\mathcal{S}$, and assume that Π_θ is not parallel to $T_\mathbf{p}\mathcal{S}$. Suppose that Π_θ intersects $\mathcal{S}$ in a curve with curvature κ_θ. Then, $\kappa_\theta \sin\theta$ is independent of θ.

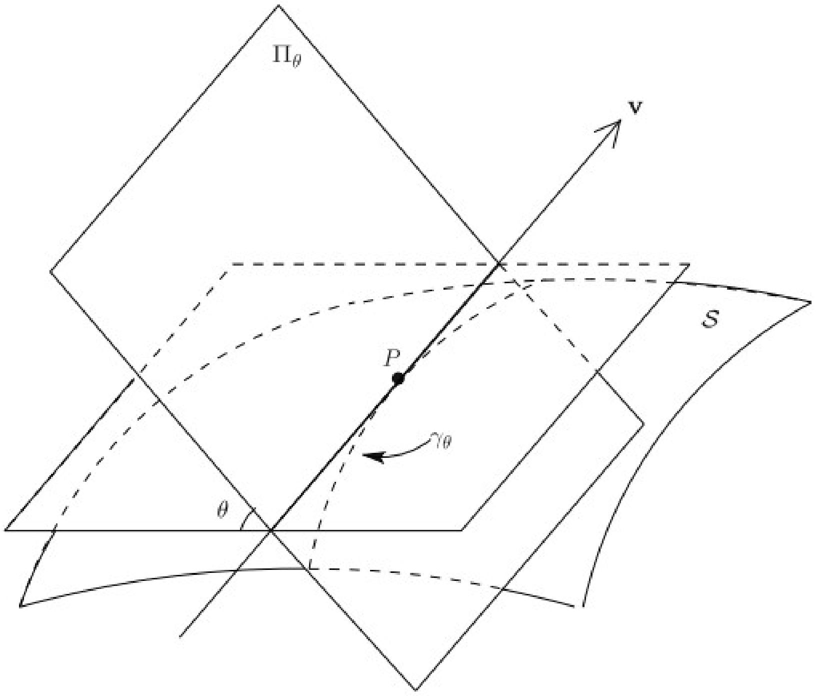

Proof

Assume that $\boldsymbol{\gamma}_\theta$ is a unit-speed parametrization of the curve of intersection of Π_θ and $\mathcal{S}$. Then, at $\mathbf{p}$, $\dot{\boldsymbol{\gamma}}_\theta = \pm\mathbf{v}$, so $\ddot{\boldsymbol{\gamma}}_\theta$ is perpendicular to $\mathbf{v}$ and is parallel to Π_θ. Thus, in the notation of Proposition 7.3.2, $\psi = \pi/2-\theta$ and so Eq. 7.10 gives

$$\kappa_\theta \sin\theta = \kappa_n.$$

But κ_n depends only on $\mathbf{p}$ and $\mathbf{v}$, and not on θ. ☐

An important special case is that in which $\boldsymbol{\gamma}$ is a *normal section* of the surface, i.e., $\boldsymbol{\gamma}$ is the intersection of the surface with a plane Π that is perpendicular to the tangent plane of the surface at every point of $\boldsymbol{\gamma}$.

Corollary 7.3.5

The curvature κ, normal curvature κ_n and geodesic curvature κ_g of a normal section of a surface are related by

$$\kappa_n = \pm\kappa, \quad \kappa_g = 0.$$

Proof

As in the proof of Proposition 7.3.3, $\kappa_n = \kappa \sin\theta$, where $\theta = \pm\pi/2$ for a normal section. This gives the first equation; the second follows from it and Eq. 7.9. ☐

EXERCISES

7.3.1 Let $\boldsymbol{\gamma}$ be a regular, but not necessarily unit-speed, curve on a surface. Prove that (with the usual notation) the normal and geodesic curvatures of $\boldsymbol{\gamma}$ are

$$\kappa_n = \frac{\langle \dot{\gamma}, \dot{\gamma} \rangle}{\langle \dot{\gamma}, \dot{\gamma} \rangle} \quad \text{and} \quad \kappa_g = \frac{\ddot{\gamma} \cdot (\mathbf{N} \times \dot{\gamma})}{\langle \dot{\gamma}, \dot{\gamma} \rangle^{3/2}}.$$

7.3.2 Show that the normal curvature of any curve on a sphere of radius r is $\pm 1/r$.

7.3.3 Compute the geodesic curvature of any circle on a sphere (not necessarily a great circle).

7.3.4 Show that, if $\boldsymbol{\gamma}(t) = \boldsymbol{\sigma}(u(t), v(t))$ is a unit-speed curve on a surface patch $\boldsymbol{\sigma}$ with first fundamental form $Edu^2 + 2Fdudv + Gdv^2$, the geodesic curvature of $\boldsymbol{\gamma}$ is

$$\kappa_g = (\ddot{v}\dot{u} - \dot{v}\ddot{u})\sqrt{EG - F^2} + A\dot{u}^3 + B\dot{u}^2\dot{v} + C\dot{u}\dot{v}^2 + D\dot{v}^3,$$

where A, B, C and D can be expressed in terms of E, F, G and their derivatives. Find A, B, C, D explicitly when $F = 0$.

7.3.5 Suppose that a unit-speed curve $\boldsymbol{\gamma}$ with curvature $\kappa > 0$ and principal normal $\mathbf{n}$ is a parametrization of the intersection of two oriented surfaces $\mathcal{S}_1$ and $\mathcal{S}_2$ with unit normals $\mathbf{N}_1$ and $\mathbf{N}_2$. Show that, if κ_1 and κ_2 are the normal curvatures of $\boldsymbol{\gamma}$ when viewed as a curve in $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively, then

$$\kappa_1 \mathbf{N}_2 - \kappa_2 \mathbf{N}_1 = \kappa(\mathbf{N}_1 \times \mathbf{N}_2) \times \mathbf{n}.$$

Deduce that, if α is the angle between the two surfaces,

$$\kappa^2 \sin^2 \alpha = \kappa_1^2 + \kappa_2^2 - 2\kappa_1\kappa_2 \cos\alpha.$$

7.3.6 A curve $\boldsymbol{\gamma}$ on a surface $\mathcal{S}$ is called *asymptotic* if its normal curvature is everywhere zero. Show that any straight line on a surface is an asymptotic curve. Show also that a curve $\boldsymbol{\gamma}$ with positive curvature is asymptotic if and only if its binormal $\mathbf{b}$ is parallel to the unit normal of $\mathcal{S}$ at all points of $\boldsymbol{\gamma}$.

7.3.7 Prove that the asymptotic curves on the surface

$$\boldsymbol{\sigma}(u, v) = (u\cos v, u \sin v, \ln u)$$

are given by

$$\ln u = \pm(v + c),$$

where c is an arbitrary constant.

7.4 Parallel transport and covariant derivative

Imagine that there is a road that runs along the Earth's equator and that you are driving along this road at constant speed. The road would appear perfectly straight – you would not have to turn to the right or left to continue along the road. Thus, you would perceive your *velocity* (and not just your speed) as being *constant*. On the other hand, an observer in space would see that your velocity is not constant as you are travelling in a circle rather than in a straight line. The resolution of this apparent paradox is that an observer restricted to the surface of the Earth perceives only the component of the acceleration tangential to the surface. From the point of view of the observer in space, the acceleration vector points towards the centre of the Earth, and so has zero tangential component.

In general, suppose that $\boldsymbol{\gamma}$ is a curve on a surface $\mathcal{S}$ and let $\mathbf{v}$ be a *tangent vector field* along $\boldsymbol{\gamma}$, i.e., a smooth map from an open interval (α, β) to $\mathbb{R}^3$ such that $\mathbf{v}(t) \in T_{\boldsymbol{\gamma}(t)}\mathcal{S}$ for all $t \in (\alpha, \beta)$. To an observer moving along the curve $\boldsymbol{\gamma}$, the perceived rate of change of $\mathbf{v}$ is the tangential component of $\dot{\mathbf{v}}$, i.e., the orthogonal projection of $\dot{\mathbf{v}} = d\mathbf{v}/dt$ onto $T_{\boldsymbol{\gamma}(t)}\mathcal{S}$. If $\mathbf{N}$ is a unit normal to $\boldsymbol{\sigma}$, the component of $\dot{\mathbf{v}}$ perpendicular to the surface is $(\dot{\mathbf{v}} \cdot \mathbf{N})\mathbf{N}$, so the tangential component is

$$\nabla_{\boldsymbol{\gamma}}\mathbf{v} = \dot{\mathbf{v}} - (\dot{\mathbf{v}} \cdot \mathbf{N})\mathbf{N}. \tag{7.11}$$

Note that this is unchanged if $\mathbf{N}$ is replaced by $-\mathbf{N}$, so $\nabla_{\boldsymbol{\gamma}}\mathbf{v}$ is well defined on any surface, orientable or not.

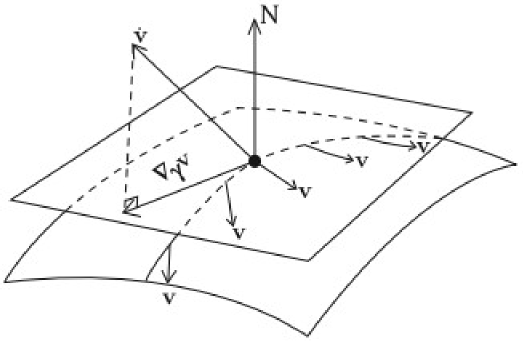

Definition 7.4.1

Let $\boldsymbol{\gamma}$ be a curve on a surface $\mathcal{S}$ and let $\mathbf{v}$ be a tangent vector field along $\boldsymbol{\gamma}$. The *covariant derivative* of $\mathbf{v}$ along $\boldsymbol{\gamma}$ is the orthogonal projection $\nabla_{\boldsymbol{\gamma}}\mathbf{v}$ of $d\mathbf{v}/dt$ onto the tangent plane $T_{\boldsymbol{\gamma}(t)}\mathcal{S}$ at a point $\boldsymbol{\gamma}(t)$.

In particular, an inhabitant of $\mathcal{S}$ would perceive $\mathbf{v}$ as being constant along $\boldsymbol{\gamma}$ if $\nabla_{\boldsymbol{\gamma}}\mathbf{v} = \mathbf{0}$. In this case, $\mathbf{v}$ is sometimes said to be *covariant constant*, but the usual terminology is contained in

Definition 7.4.2

With the notation in Definition 7.4.1, $\mathbf{v}$ is said to be *parallel along* $\boldsymbol{\gamma}$ if $\nabla_{\boldsymbol{\gamma}}\mathbf{v} = \mathbf{0}$ at every point of $\boldsymbol{\gamma}$.

Proposition 7.4.3

A tangent vector field $\mathbf{v}$ is parallel along a curve $\boldsymbol{\gamma}$ on a surface $\mathcal{S}$ if and only if $\dot{\mathbf{v}}$ is perpendicular to the tangent plane of $\mathcal{S}$ at all points of $\boldsymbol{\gamma}$.

8

Gaussian, mean and principal curvatures

In this chapter, we show how to extract geometric information from the second fundamental form of a surface or, equivalently, from its Weingarten map.

8.1 Gaussian and mean curvatures

We start by defining two new measures of the curvature of a surface.

Definition 8.1.1

Let $\mathcal{W}$ be the Weingarten map of an oriented surface $\mathcal{S}$ at a point $\mathbf{p} \in \mathcal{S}$. The Gaussian curvature K and mean curvature H of $\mathcal{S}$ at $\mathbf{p}$ are defined by

$$K = \det(\mathcal{W}), \quad H = \frac{1}{2}\text{trace}(\mathcal{W}).$$

Recall that the determinant and trace of a linear map (such as $\mathcal{W}$) can be computed as the determinant and the sum of the diagonal entries of the matrix of the linear map with respect to any basis (in this case of the tangent plane), and that they depend only on the linear map and not on the choice of basis.

When the sign of the unit normal of $\mathcal{S}$ is changed, the Weingarten map also changes sign (Exercise 7.2.2), thus leaving K unchanged. This implies that the Gaussian curvature is defined for *any* surface $\mathcal{S}$, orientable or not: to define K at a point $\mathbf{p} \in \mathcal{S}$, choose a surface patch $\boldsymbol{\sigma}$ with $\mathbf{p}$ in its image; this is an

Andrew Pressley, *Elementary Differential Geometry: Second Edition*,
Springer Undergraduate Mathematics Series, DOI 10.1007/978-1-84882-891-9_8,
© Springer-Verlag London Limited 2010

oriented surface, which may be used to define K, and the result is independent of the choice of $\boldsymbol{\sigma}$. On the other hand, on a surface that is not necessarily orientable, H is in general only well defined up to sign.

To get explicit formulas for H and K, we work in a surface patch of $\mathcal{S}$. Let $\boldsymbol{\sigma}(u, v)$ be a surface patch with first and second fundamental forms

$$Edu^2 + 2Fdudv + Gdv^2 \quad \text{and} \quad Ldu^2 + 2Mdudv + Ndv^2,$$

respectively. Define symmetric 2×2 matrices $\mathcal{F}_I$ and $\mathcal{F}_{II}$ by

$$\mathcal{F}_I = \begin{pmatrix} E & F \\ F & G \end{pmatrix}, \quad \mathcal{F}_{II} = \begin{pmatrix} L & M \\ M & N \end{pmatrix}.$$

Proposition 8.1.2

Let $\boldsymbol{\sigma}$ be a surface patch of an oriented surface $\mathcal{S}$. Then, with the above notation, the matrix of $\mathcal{W}_{\mathbf{p},\mathcal{S}}$ with respect to the basis $\{\boldsymbol{\sigma}_u, \boldsymbol{\sigma}_v\}$ of $T_{\mathbf{p}}\mathcal{S}$ is $\mathcal{F}_I^{-1}\mathcal{F}_{II}$.

Proof

By the proof of Proposition 7.2.2, $\mathcal{W}(\boldsymbol{\sigma}_u) = -\mathbf{N}_u$ and $\mathcal{W}(\boldsymbol{\sigma}_v) = -\mathbf{N}_v$, so the matrix of $\mathcal{W}$ is $\begin{pmatrix} a & c \\ b & d \end{pmatrix}$, where

$$-\mathbf{N}_u = a\boldsymbol{\sigma}_u + b\boldsymbol{\sigma}_v, \quad -\mathbf{N}_v = c\boldsymbol{\sigma}_u + d\boldsymbol{\sigma}_v.$$

Take the dot product of each of these equations with $\boldsymbol{\sigma}_u$ and $\boldsymbol{\sigma}_v$ and use Lemma 7.2.3; this gives

$$\begin{aligned} L &= aE + bF, \qquad & M &= cE + dF, \\ M &= aF + bG, \qquad & N &= cF + dG. \end{aligned}$$

These four scalar equations are equivalent to the single matrix equation

$$\begin{pmatrix} L & M \\ M & N \end{pmatrix} = \begin{pmatrix} E & F \\ F & G \end{pmatrix} \begin{pmatrix} a & c \\ b & d \end{pmatrix}$$
$$\text{i.e.,} \quad \mathcal{F}_{II} = \mathcal{F}_I \begin{pmatrix} a & c \\ b & d \end{pmatrix}.$$

Hence, the matrix of $\mathcal{W}$ with respect to the basis $\{\boldsymbol{\sigma}_u, \boldsymbol{\sigma}_v\}$ is

$$\begin{pmatrix} a & c \\ b & d \end{pmatrix} = \mathcal{F}_I^{-1}\mathcal{F}_{II}.$$

□

Corollary 8.1.3

We have

$$H = \frac{LG - 2MF + NE}{2(EG - F^2)}, \quad K = \frac{LN - M^2}{EG - F^2}.$$

Proof

By Definition 8.1.1,

$$K = \det\left(\mathcal{F}_I^{-1}\mathcal{F}_{II}\right) = \frac{\det(\mathcal{F}_{II})}{\det(\mathcal{F}_I)} = \frac{LN - M^2}{EG - F^2}.$$

To compute H, we need the trace of the matrix

$$\begin{aligned} \mathcal{F}_I^{-1}\mathcal{F}_{II} &= \frac{1}{EG - F^2}\begin{pmatrix} G & -F \\ -F & E \end{pmatrix}\begin{pmatrix} L & M \\ M & N \end{pmatrix} \\ &= \frac{1}{EG - F^2}\begin{pmatrix} LG - MF & MG - NF \\ ME - LF & NE - MF \end{pmatrix}. \end{aligned}$$

Thus,

$$2H = \operatorname{trace}\left(\mathcal{F}_I^{-1}\mathcal{F}_{II}\right) = \frac{LG - 2MF + NE}{EG - F^2}. \qquad \square$$

Example 8.1.4

In Examples 6.1.3 and 7.1.2 we considered the surface of revolution

$$\boldsymbol{\sigma}(u, v) = (f(u)\cos v, f(u)\sin v, g(u)),$$

where we can assume that $f > 0$ and $\dot{f}^2 + \dot{g}^2 = 1$ everywhere (a dot denoting d/du). We found that

$$E = 1,\ F = 0,\ G = f^2,\ \ L = \dot{f}\ddot{g} - \ddot{f}\dot{g},\ M = 0,\ N = f\dot{g}.$$

By Corollary 8.1.3, the Gaussian curvature is

$$K = \frac{LN - M^2}{EG - F^2} = \frac{(\dot{f}\ddot{g} - \ddot{f}\dot{g})f\dot{g}}{f^2}.$$

We can simplify this formula by noting that $\dot{f}^2 + \dot{g}^2 = 1$ implies (by differentiating with respect to u) that $\dot{f}\ddot{f} + \dot{g}\ddot{g} = 0$,

$$\begin{aligned} \therefore \quad (\dot{f}\ddot{g} - \ddot{f}\dot{g})\dot{g} &= -\dot{f}^2\ddot{f} - \ddot{f}\dot{g}^2 = -\ddot{f}(\dot{f}^2 + \dot{g}^2) = -\ddot{f}, \\ \therefore \quad K &= -\frac{\ddot{f}f}{f^2} = -\frac{\ddot{f}}{f}. \end{aligned}$$

We consider some special cases. If $\boldsymbol{\gamma}(u) = (u, 0, 0)$ is the x-axis, the corresponding surface of revolution is the xy-plane; since $f(u) = u$, we have $\dot{f} = 1$, $\ddot{f} = 0$, so $K = 0$. If $\boldsymbol{\gamma}(u) = (1, 0, u)$ is a straight line parallel to the z-axis, the corresponding surface is the unit cylinder; since $f(u) = 1$, $\ddot{f} = 0$, so $K = 0$. Finally, if $\boldsymbol{\gamma}(u) = (\cos u, 0, \sin u)$ is a circle of radius 1, the corresponding surface is the unit sphere; since $f(u) = \cos u$, $\dot{f} = -\sin u$, $\ddot{f} = -\cos u$ so $K = -\ddot{f}/f = -(-\cos u)/\cos u = 1$. Note that in each of these examples the curve $\boldsymbol{\gamma}$ is unit-speed.

Example 8.1.5

For a ruled surface, take a patch

$$\boldsymbol{\sigma}(u, v) = \boldsymbol{\gamma}(u) + v\boldsymbol{\delta}(u),$$

(see Example 5.3.1). Denoting d/du by a dot, we have $\boldsymbol{\sigma}_u = \dot{\boldsymbol{\gamma}} + v\dot{\boldsymbol{\delta}}$, $\boldsymbol{\sigma}_v = \boldsymbol{\delta}$, so

$$\boldsymbol{\sigma}_{uv} = \dot{\boldsymbol{\delta}}, \quad \boldsymbol{\sigma}_{vv} = \mathbf{0}.$$

Hence, if $\mathbf{N} = (\boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v)/ \parallel \boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v \parallel$ is the standard unit normal of $\boldsymbol{\sigma}$, then $M = \boldsymbol{\sigma}_{uv} \cdot \mathbf{N} = \dot{\boldsymbol{\delta}} \cdot \mathbf{N}$ and $N = 0$. So

$$K = \frac{LN - M^2}{EG - F^2} = \frac{-(\dot{\boldsymbol{\delta}} \cdot \mathbf{N})^2}{EG - F^2} \leq 0,$$

i.e., *the Gaussian curvature of a ruled surface is negative or zero.*

Gauss discovered a way to obtain the Gaussian curvature from the Gauss map itself, rather than from its derivative, the Weingarten map. His result is an analogue of Proposition 2.2.3, which shows that, if $\boldsymbol{\gamma}$ is a unit-speed plane curve, its signed curvature $\kappa_s = \dot{\varphi}$, where φ is the angle between its tangent vector $\dot{\boldsymbol{\gamma}}$ and a fixed direction, i.e., the (signed) *curvature is the rate of change of direction of the tangent vector of* $\boldsymbol{\gamma}$ *per unit length.* The 'direction' of the tangent plane to an oriented surface $\mathcal{S}$ is measured by its unit normal $\mathbf{N}$, so we might expect that a measure of the curvature of $\boldsymbol{\sigma}$ is the 'rate of change of $\mathbf{N}$ per unit area'. The values of $\mathbf{N}$ at points of $\mathcal{S}$ are recorded by the Gauss map $\mathcal{G}$, so if R is a small region on $\mathcal{S}$ containing a point $\mathbf{p}$, we should look at the ratio

$$\frac{\text{Area}(\mathcal{G}(R))}{\text{Area}(R)}$$

in the limit as the region R shrinks down to the point $\mathbf{p}$.

To make this idea precise, we work in a surface patch.

Theorem 8.1.6

Let $\boldsymbol{\sigma} : U \to \mathbb{R}^3$ be a surface patch, let $(u_0, v_0) \in U$, and let $\delta > 0$ be such that the closed disc

$$R_\delta = \{(u, v) \in \mathbb{R}^2 \mid (u - u_0)^2 + (v - v_0)^2 \leq \delta^2\}$$

with centre (u_0, v_0) and radius δ is contained in U. Then,

$$\lim_{\delta \to 0} \frac{\mathcal{A}_{\mathbf{N}}(R_\delta)}{\mathcal{A}_{\boldsymbol{\sigma}}(R_\delta)} = |K|,$$

where K is the Gaussian curvature of $\boldsymbol{\sigma}$ at $\boldsymbol{\sigma}(u_0, v_0)$.

Note that a δ with the properties in the statement of the theorem exists because U is open.

Proof

By Definition 6.4.1,

$$\frac{\mathcal{A}_{\mathbf{N}}(R_\delta)}{\mathcal{A}_{\boldsymbol{\sigma}}(R_\delta)} = \frac{\int_{R_\delta} \parallel \mathbf{N}_u \times \mathbf{N}_v \parallel \; dudv}{\int_{R_\delta} \parallel \boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v \parallel \; dudv}. \tag{8.1}$$

In the notation of the proof of Proposition 8.1.2,

$$\begin{aligned}
\mathbf{N}_u \times \mathbf{N}_v &= (a\boldsymbol{\sigma}_u + b\boldsymbol{\sigma}_v) \times (c\boldsymbol{\sigma}_u + d\boldsymbol{\sigma}_v) \\
&= (ad - bc)\boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v \\
&= \det(\mathcal{F}_I^{-1}\mathcal{F}_{II})\boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v \\
&= \frac{\det(\mathcal{F}_{II})}{\det(\mathcal{F}_I)}\boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v \\
&= \frac{LN - M^2}{EG - F^2}\boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v \\
&= K\boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v \quad \text{(by Corollary 8.1.3).}
\end{aligned} \tag{8.2}$$

Substituting in Eq. 8.1, we get

$$\frac{\mathcal{A}_{\mathbf{N}}(R_\delta)}{\mathcal{A}_{\boldsymbol{\sigma}}(R_\delta)} = \frac{\int_{R_\delta} |K| \parallel \boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v \parallel \; dudv}{\int_{R_\delta} \parallel \boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v \parallel \; dudv}.$$

Let ϵ be any positive number. Since $K(u, v)$ is a continuous function of (u, v) (see Exercise 8.1.3), we can choose $\delta > 0$ so small that

$$|K(u, v) - K(u_0, v_0)| < \epsilon$$

if $(u, v) \in R_\delta$. Since, for any real numbers a, b, $|a - b| \geq ||a| - |b||$, it follows that $||K(u,v)| - |K(u_0, v_0)|| < \epsilon$ if $(u, v) \in R_\delta$, i.e.,

$$|K(u_0, v_0)| - \epsilon < |K(u,v)| < |K(u_0, v_0)| + \epsilon$$

if $(u, v) \in R_\delta$. Multiplying through by $\| \boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v \|$ and integrating over R_δ, we get

$$\begin{aligned}(|K(u_0, v_0)| - \epsilon) \int \| \boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v \| \, du dv &< \int |K(u,v)| \, \| \boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v \| \, du dv \\ &< (|K(u_0, v_0)| + \epsilon) \int \| \boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v \| \, du dv, \\ \therefore \quad |K(u_0, v_0)| - \epsilon < \frac{\mathcal{A}_{\mathbf{N}}(R_\delta)}{\mathcal{A}_{\boldsymbol{\sigma}}(R_\delta)} &< |K(u_0, v_0)| + \epsilon \quad \text{(using Eq. 8.1)} \\ \therefore \quad \left| \frac{\mathcal{A}_{\mathbf{N}}(R_\delta)}{\mathcal{A}_{\boldsymbol{\sigma}}(R_\delta)} - |K(u_0, v_0)| \right| &< \epsilon.\end{aligned}$$

This proves the theorem. □

Although this proposition only gives the absolute value of the Gaussian curvature K, the sign can be recovered from the Gauss map if we define the *signed area* of $\mathcal{G}(R)$ to be $\pm \mathcal{A}_{\mathbf{N}}(R)$, where the sign is $+$ or $-$ according to whether $\mathbf{N}_u \times \mathbf{N}_v$ points in the same or the opposite direction as $\mathbf{N}$. By Eq. 8.3, this sign is that of K, so K is the limit of the ratio

$$\frac{\text{Signed area}(\mathcal{G}(R))}{\text{Area}(R)}$$

as the region R shrinks to the point $\mathbf{p}$.

As the following examples show, Theorem 8.1.6 sometimes allows one to find the Gaussian curvature of a surface with no calculation.

Example 8.1.7

For a plane, the unit normal is constant. Thus, for any R, $\mathcal{G}(R)$ is a single point, and thus has zero area. By the theorem, a plane has Gaussian curvature zero everywhere.

For a generalized cylinder, the unit normal is clearly always perpendicular to the rulings of the cylinder, so the image of the Gauss map is contained in the great circle on S^2 formed by intersecting S^2 with the plane passing through its centre perpendicular to the rulings of the cylinder. Any great circle obviously has zero area, so the cylinder has zero Gaussian curvature too.

Finally, for the unit sphere S^2 itself, the unit normal at a point $\mathbf{p}$ is clearly parallel to the radius vector from the centre of the sphere to $\mathbf{p}$. In other words, the Gauss map is the identity map or the antipodal map (depending on the choice of orientation). Both of these maps are obviously equiareal, so the absolute value of the Gaussian curvature of S^2 is 1. In fact, if $\boldsymbol{\sigma}$ is any surface patch of S^2, we have $\mathbf{N} = \pm\boldsymbol{\sigma}$ so with either choice of sign $\mathbf{N}_u \times \mathbf{N}_v = \boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v$ is a positive multiple of $\mathbf{N}$ and the Gaussian curvature is $+1$.

EXERCISES

8.1.1 Show that the Gaussian and mean curvatures of the surface $z = f(x, y)$, where f is a smooth function, are

$$K = \frac{f_{xx}f_{yy} - f_{xy}^2}{(1 + f_x^2 + f_y^2)^2}, \quad H = \frac{(1 + f_y^2)f_{xx} - 2f_x f_y f_{xy} + (1 + f_x^2)f_{yy}}{2(1 + f_x^2 + f_y^2)^{3/2}}.$$

8.1.2 Calculate the Gaussian curvature of the helicoid and catenoid (Exercises 4.2.6 and 5.3.1).

8.1.3 Show that the Gaussian and mean curvatures of a surface $\mathcal{S}$ are smooth functions on $\mathcal{S}$.

8.1.4 In the notation of Example 8.1.5, show that if $\boldsymbol{\delta}$ is the principal normal $\mathbf{n}$ of $\boldsymbol{\gamma}$ or its binormal $\mathbf{b}$, then $K = 0$ if and only if $\boldsymbol{\gamma}$ is planar.

8.1.5 What is the effect on the Gaussian and mean curvatures of a surface $\mathcal{S}$ if we apply a dilation of $\mathbb{R}^3$ to $\mathcal{S}$?

8.1.6 Show that the Weingarten map $\mathcal{W}$ of a surface satisfies the quadratic equation

$$\mathcal{W}^2 - 2H\mathcal{W} + K = 0,$$

in the usual notation.

8.1.7 Show that the image of the Gauss map of a generalized cone is a *curve* on S^2, and deduce that the cone has zero Gaussian curvature.

8.1.8 Let $\boldsymbol{\sigma} : U \to \mathbb{R}^3$ be a patch of a surface $\mathcal{S}$. Show that the image under the Gauss map of the part $\boldsymbol{\sigma}(R)$ of $\mathcal{S}$ corresponding to a region $R \subseteq U$ has area

$$\int_R |K| d\mathcal{A}_{\boldsymbol{\sigma}},$$

where K is the Gaussian curvature of $\mathcal{S}$.

5- <N2 <3

8.2.5 Suppose that two surfaces $\mathcal{S}_1$ and $\mathcal{S}_2$ intersect in a curve $\mathcal{C}$ that is a line of curvature of $\mathcal{S}_1$. Show that $\mathcal{C}$ is a line of curvature of $\mathcal{S}_2$ if and only if the angle between the tangent planes of $\mathcal{S}_1$ and $\mathcal{S}_2$ is constant along $\mathcal{C}$.

8.2.6 Let $\mathbf{\Sigma} : W \to \mathbb{R}^3$ be a smooth function defined on an open subset W of $\mathbb{R}^3$ such that, for each fixed value of u (resp. v, w), $\mathbf{\Sigma}(u, v, w)$ is a (regular) surface patch. Assume also that

$$\mathbf{\Sigma}_u \cdot \mathbf{\Sigma}_v = \mathbf{\Sigma}_v \cdot \mathbf{\Sigma}_w = \mathbf{\Sigma}_w \cdot \mathbf{\Sigma}_u = 0. \tag{8.5}$$

This means that the three families of surfaces formed by fixing the values of u, v or w constitute a triply orthogonal system (see Section 5.5).

(i) Show that $\mathbf{\Sigma}_u \cdot \mathbf{\Sigma}_{vw} = \mathbf{\Sigma}_v \cdot \mathbf{\Sigma}_{uw} = \mathbf{\Sigma}_w \cdot \mathbf{\Sigma}_{uv} = 0$.

(ii) Show that, for each of the surfaces in the triply orthogonal system, the matrices $\mathcal{F}_I$ and $\mathcal{F}_{II}$ are diagonal.

(iii) Deduce that the intersection of any surface from one family of the triply orthogonal system with any surface from another family is a line of curvature on both surfaces. (This is called *Dupin's Theorem.*)

8.2.7 Show that, if p, q and r are distinct positive numbers, there are exactly four umbilics on the ellipsoid

$$\frac{x^2}{p^2} + \frac{y^2}{q^2} + \frac{z^2}{r^2} = 1.$$

What happens if p, q and r are not distinct?

8.2.8 Show that the principal curvatures of a surface patch $\boldsymbol{\sigma} : U \to \mathbb{R}^3$ are smooth functions on U provided that $\boldsymbol{\sigma}$ has no umbilics. Show also that the principal curvatures either stay the same or both change sign when $\boldsymbol{\sigma}$ is reparametrized.

8.3 Surfaces of constant Gaussian curvature

We have seen in the examples in Section 8.1 some surfaces of zero and constant positive curvature. For an example of a surface with *constant negative* Gaussian curvature, however, we have to construct a new surface. To this end, we examine again the surface of revolution

$$\boldsymbol{\sigma}(u, v) = (f(u)\cos v, f(u)\sin v, g(u))$$

obtained by rotating the unit-speed curve $u \mapsto (f(u), 0, g(u))$ in the xz-plane around the z-axis. We found in Example 8.1.4 that its Gaussian curvature is

$$K = -\frac{\ddot{f}}{f}. \tag{8.6}$$

Suppose first that $K = 0$ everywhere. Then, Eq. 8.6 gives $\ddot{f} = 0$, so $f(u) = au + b$ for some constants a and b. Since $\dot{f}^2 + \dot{g}^2 = 1$, we get $\dot{g} = \pm\sqrt{1 - a^2}$ (so we must have $|a| \leq 1$) and hence $g(u) = \pm\sqrt{1 - a^2}u + c$, where c is another constant. By applying a translation along the z-axis we can assume that $c = 0$, and by applying a rotation by π about the x-axis, if necessary, we can assume that the sign is $+$. This gives the ruled surface

$$\boldsymbol{\sigma}(u, v) = (b\cos v, b\sin v, 0) + u(a\cos v, a\sin v, \sqrt{1 - a^2}).$$

If $a = 0$ this is a circular cylinder; if $|a| = 1$ it is the xy-plane; and if $0 < |a| < 1$ it is a circular cone (to see this, put $\tilde{u} = au + b$).

Now suppose that $K > 0$, say $K = 1/R^2$, where $R > 0$ is a constant. Then, Eq. 8.6 becomes

$$\ddot{f} + \frac{f}{R^2} = 0,$$

which has the general solution

$$f(u) = a\cos\left(\frac{u}{R} + b\right),$$

where a and b are constants. We can assume that $b = 0$ by performing a reparametrization $\tilde{u} = u + Rb, \tilde{v} = v$. Then, up to a change of sign and adding a constant,

$$g(u) = \int \sqrt{1 - \frac{a^2}{R^2}\sin^2\frac{u}{R}}\, du.$$

The integral in the formula for $g(u)$ can be evaluated in terms of 'elementary' functions only when $a = 0$ or $\pm R$. The case $a = 0$ does not give a surface, and if $a = R$ then $f(u) = R\cos\frac{u}{R}$, $g(u) = R\sin\frac{u}{R}$, and we have a sphere of radius R (the case $a = -R$ can be reduced to this by rotating the surface by π around the z-axis).

Suppose finally that $K < 0$. We can restrict ourselves to the case $K = -1$, as the general case can be obtained from this by applying a dilation of $\mathbb{R}^3$ (see Exercise 8.1.5). In view of the preceding case, we can think of a surface with $K = -1$ as a 'sphere of imaginary radius' $\sqrt{-1}$, or a 'pseudosphere'.

When $K = -1$ the general solution of Eq. 8.6 is

$$f(u) = ae^{u} + be^{-u},$$

where a and b are arbitrary constants. The function $g(u)$ can be expressed in terms of elementary functions only if one of a or b is zero. If $b = 0$ we can assume that $a = 1$ by a reparametrization $u \mapsto u + \text{constant}$, and the case in which $a = 0$ can be reduced to the case $b = 0$ by the reparametrization $u \mapsto -u$. Suppose then that $a = 1$ and $b = 0$; then, $f(u) = e^{u}$ and we can take

$$g(u) = \int \sqrt{1 - e^{2u}}\, du. \tag{8.7}$$

Note that we must have $u \leq 0$ for the integral in Eq. 8.7 to make sense, since otherwise $1 - e^{2u}$ would be negative. The integral can be evaluated by putting $\cos\theta = e^{u}$. Then,

$$\begin{aligned}\int \sqrt{1 - e^{2u}}\, du &= -\int \frac{\sin^2\theta}{\cos\theta}\, d\theta = \sin\theta - \ln(\sec\theta + \tan\theta)\\ &= \sqrt{1 - e^{2u}} - \ln(e^{-u} + \sqrt{e^{-2u} - 1}).\end{aligned}$$

We have omitted the arbitrary constant, but we can take it to be zero by a suitable translation of the surface parallel to the z-axis. Putting $x = f(u)$, $z = g(u)$, and noting that $\cosh^{-1}(v) = \ln(v + \sqrt{v^2 - 1})$, we see that the profile curve in the xz-plane has equation

$$z = \sqrt{1 - x^2} - \cosh^{-1}\left(\frac{1}{x}\right). \tag{8.8}$$

Rotating this curve around the z -axis thus gives a surface which has Gaussian curvature -1 everywhere. Note that, since $u \leq 0$, $x = e^{u}$ is restricted to the range $0 < x \leq 1$.

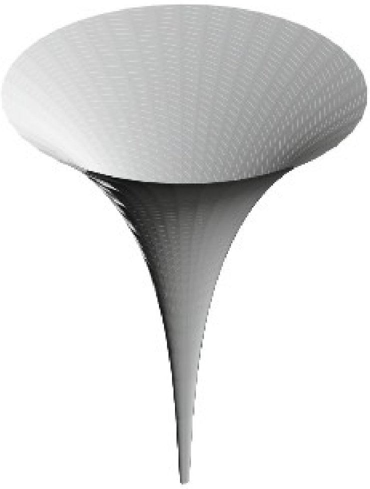

The curve defined by Eq. 8.8 is called the *tractrix*, and it has an interesting geometrical property. Consider the tangent line at a point P of its graph, and suppose that it intersects the z-axis at the point Q. Let us compute the distance from P to Q.

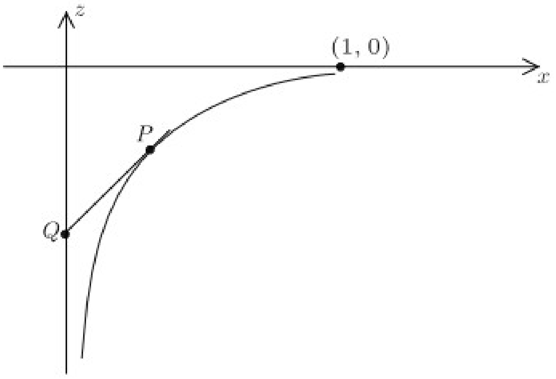

Suppose that P is the point (x_0, z_0). Either by a direct calculation or by inspecting the calculation of the integral (8.7), one finds that

$$\frac{dz}{dx} = \frac{\sqrt{1-x^2}}{x}.$$

Hence, the tangent line at P has equation

$$z - z_0 = \frac{\sqrt{1-x_0^2}}{x_0}(x - x_0).$$

This meets the z-axis at the point $(0, z_1)$, where

$$z_1 - z_0 = \frac{\sqrt{1-x_0^2}}{x_0}(0 - x_0) = -\sqrt{1-x_0^2}.$$

Hence, the square of the distance from P to Q is

$$x_0^2 + (z_1 - z_0)^2 = x_0^2 + 1 - x_0^2 = 1,$$

so the distance from P to Q is *constant* and equal to 1.

This means that the tractrix has the following description. Let a donkey pull a box of stones by a rope of length 1. Suppose that the donkey is initially at $(0, 0)$, the box is initially at $(1, 0)$, and let the donkey walk slowly along the negative z-axis. Then, the box of stones moves along the tractrix.

8.4 Flat surfaces

In Section 8.3, we gave some examples of surfaces of constant Gaussian curvature K, but this certainly falls well short of a complete classification of such surfaces. It is possible, however, to give a fairly complete description of *flat surfaces*, i.e., surfaces for which $K = 0$ everywhere. To do so, we shall make use of a special parametrization, valid for any surface, described in the following proposition.

Proposition 8.4.1

Let $\mathbf{p}$ be a point of a surface $\mathcal{S}$, and suppose that $\mathbf{p}$ is not an umbilic. Then, there is a surface patch $\boldsymbol{\sigma}(u, v)$ of $\mathcal{S}$ containing $\mathbf{p}$ whose first and second fundamental forms are

$$Edu^2 + Gdv^2 \quad \text{and} \quad Ldu^2 + Ndv^2,$$

respectively, for some smooth functions E,G,L and N.

We recall that a point $\mathbf{p}$ of a surface $\mathcal{S}$ is an umbilic if the two principal curvatures of $\mathcal{S}$ at $\mathbf{p}$ are equal. From Section 8.2, we see that for the patch $\boldsymbol{\sigma}$ in the statement of the proposition, $\boldsymbol{\sigma}_u$ and $\boldsymbol{\sigma}_v$ are principal vectors with corresponding principal curvatures L/E and N/G. We call $\boldsymbol{\sigma}$ a *principal patch.*

We assume Proposition 8.4.1 for the moment, and use it to give the proof of

Proposition 8.4.2

Let $\mathbf{p}$ be a point of a flat surface $\mathcal{S}$, and assume that $\mathbf{p}$ is not an umbilic. Then, there is a patch of $\mathcal{S}$ containing $\mathbf{p}$ that is a ruled surface.

Proof

We take a principal patch $\boldsymbol{\sigma} : U \to \mathbb{R}^3$ containing $\mathbf{p}$ as in Proposition 8.4.1, say $\mathbf{p} = \boldsymbol{\sigma}(u_0, v_0)$. By Corollary 8.1.3, the Gaussian curvature $K = LN/EG$. Since the Gaussian curvature is zero everywhere, either $L = 0$ or $N = 0$ at each point of U, and since $\mathbf{p}$ is not an umbilic L and N are not both zero. Suppose that $L(u_0, v_0) \neq 0$, say. Then, $L(u, v) \neq 0$ for (u, v) in some open subset of U containing (u_0, v_0). Hence, by shrinking U if necessary, we can assume that $L \neq 0$ at every point of U. Then, $N = 0$ everywhere, and the second fundamental form of $\boldsymbol{\sigma}$ is Ldu^2.

We shall prove that the parameter curves $u =$ constant are straight lines. Such a curve can be parametrized by $v \mapsto \boldsymbol{\sigma}(u_0, v)$, where u_0 is the constant value of u. A unit tangent vector to this curve is $\mathbf{t} = \boldsymbol{\sigma}_v/G^{1/2}$, so by Proposition 1.1.6 what we have to prove is that $\mathbf{t}_v = \mathbf{0}$.

By the proof of Proposition 8.1.2, the derivatives of the unit normal are

$$\mathbf{N}_u = -E^{-1}L\boldsymbol{\sigma}_u, \quad \mathbf{N}_v = \mathbf{0}. \tag{8.9}$$

Hence, $\mathbf{t}_v \cdot \boldsymbol{\sigma}_u = -EL^{-1}\mathbf{t}_v \cdot \mathbf{N}_u$. Now, $\mathbf{t} \cdot \mathbf{N}_u = 0$ and $\mathbf{N}_{uv} = \mathbf{0}$ by Eq. 8.9, so $\mathbf{t}_v \cdot \mathbf{N}_u = -\mathbf{t} \cdot \mathbf{N}_{uv} = 0$. Hence, $\mathbf{t}_v \cdot \boldsymbol{\sigma}_u = 0$. Next, $\mathbf{t}_v \cdot \mathbf{t} = 0$ since $\mathbf{t}$ is a unit vector by construction, so $\mathbf{t}_v \cdot \boldsymbol{\sigma}_v = 0$. Finally, $\mathbf{t}_v \cdot \mathbf{N} = -\mathbf{t} \cdot \mathbf{N}_v = 0$ by Eq. 8.9 again. Since the vectors $\boldsymbol{\sigma}_u, \boldsymbol{\sigma}_v$ and $\mathbf{N}$ form a basis of $\mathbb{R}^3$, we have proved that $\mathbf{t}_v = \mathbf{0}$. □

Our task, then, is to describe the structure of flat ruled surfaces. We parametrize the ruled surface as in Example 8.1.5:

$$\boldsymbol{\sigma}(u, v) = \boldsymbol{\gamma}(u) + v\boldsymbol{\delta}(u).$$

We found there that $\boldsymbol{\sigma}_u = \dot{\boldsymbol{\gamma}} + v\dot{\boldsymbol{\delta}}$, $\boldsymbol{\sigma}_v = \boldsymbol{\delta}$, the dot denoting d/du, and that the Gaussian curvature of $\boldsymbol{\sigma}$ is zero if and only if

$$\dot{\boldsymbol{\delta}} \cdot (\boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v) = 0.$$

Since

$$\boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v = \dot{\boldsymbol{\gamma}} \times \boldsymbol{\delta} + v\dot{\boldsymbol{\delta}} \times \boldsymbol{\delta},$$

and $\dot{\boldsymbol{\delta}} \cdot (\dot{\boldsymbol{\delta}} \times \boldsymbol{\delta}) = 0$,

$$K = 0 \quad \text{if and only if} \quad \dot{\boldsymbol{\delta}} \cdot (\dot{\boldsymbol{\gamma}} \times \boldsymbol{\delta}) = 0. \tag{8.10}$$

Thus, $K = 0$ *if and only if* $\dot{\boldsymbol{\gamma}}$, $\boldsymbol{\delta}$ *and* $\dot{\boldsymbol{\delta}}$ *are everywhere linearly dependent.*

To proceed further, let us assume, as we may, that $\boldsymbol{\delta}(u)$ is a unit vector for all values of u. Then, $\boldsymbol{\delta} \cdot \dot{\boldsymbol{\delta}} = 0$. Suppose first that $\dot{\boldsymbol{\delta}}(u) = \mathbf{0}$ for all values of u. Then, $\boldsymbol{\delta}$ is a constant vector and $\boldsymbol{\sigma}$ is a *generalized cylinder.*

Suppose now that $\dot{\boldsymbol{\delta}}$ is never zero. Then, $\boldsymbol{\delta}$ and $\dot{\boldsymbol{\delta}}$ are linearly independent as they are non-zero and perpendicular, so if $\dot{\boldsymbol{\gamma}}, \boldsymbol{\delta}$ and $\dot{\boldsymbol{\delta}}$ are linearly dependent, then

$$\dot{\boldsymbol{\gamma}}(u) = f(u)\boldsymbol{\delta}(u) + g(u)\dot{\boldsymbol{\delta}}(u)$$

for some smooth functions f and g. Assume first that $f = \dot{g}$ everywhere. Then, $\dot{\boldsymbol{\gamma}} = (g\boldsymbol{\delta})^{\cdot}$ and so $\boldsymbol{\gamma} = g\boldsymbol{\delta} + \mathbf{a}$, where $\mathbf{a}$ is a constant vector; hence,

$$\boldsymbol{\sigma}(u, v) = \mathbf{a} + (v + g(u))\boldsymbol{\delta}(u).$$

Putting $\tilde{u} = u$, $\tilde{v} = v + g(u)$, we see that this is a reparametrization of a *generalized cone.*

Suppose finally that $\dot{\boldsymbol{\delta}}$ and $f - \dot{g}$ are both nowhere zero. If we define

$$\tilde{\boldsymbol{\gamma}}(u) = \boldsymbol{\gamma}(u) - g(u)\boldsymbol{\delta}(u), \quad \tilde{v} = \frac{v + g(u)}{f(u) - \dot{g}(u)},$$

a short calculation gives

$$\boldsymbol{\sigma}(u,v) = \tilde{\boldsymbol{\gamma}}(u) + \tilde{v}\dot{\tilde{\boldsymbol{\gamma}}}(u),$$

so $\boldsymbol{\sigma}$ is a reparametrization of an open subset of the *tangent developable* of $\tilde{\boldsymbol{\gamma}}$.

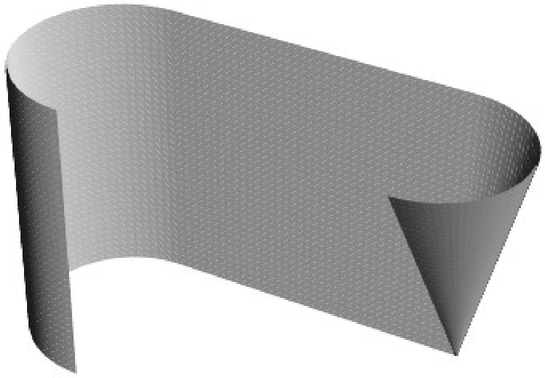

Of course, it could be that none of the conditions on $\boldsymbol{\delta}$, f and g considered above are satisfied. In fact, we have only shown that certain open subsets of the surface are parts of generalized cylinders, generalized cones or tangent developables. It is not true that the whole surface must be one of these three types, since flat surfaces of different types can be joined together to make a smooth surface, as shown in the diagram above. It can be shown that the most general flat surface is a patchwork consisting of pieces of generalized cylinders, generalized cones and tangent developables, joined together along segments of straight lines.

The remainder of this section is devoted to the proof of Proposition 8.4.1 and can safely be omitted by readers who are uncomfortable with the use of the inverse function theorem. In fact, we can prove a more general result with no additional effort:

Proposition 8.4.3

Let $\tilde{\boldsymbol{\sigma}} : \tilde{U} \to \mathbb{R}^3$ be a surface patch, and suppose that for all $(\tilde{u},\tilde{v}) \in \tilde{U}$ we are given tangent vectors

$$\boldsymbol{e}_1(\tilde{u},\tilde{v}) = a(\tilde{u},\tilde{v})\tilde{\boldsymbol{\sigma}}_{\tilde{u}} + b(\tilde{u},\tilde{v})\tilde{\boldsymbol{\sigma}}_{\tilde{v}}, \quad \boldsymbol{e}_2(\tilde{u},\tilde{v}) = c(\tilde{u},\tilde{v})\tilde{\boldsymbol{\sigma}}_{\tilde{u}} + d(\tilde{u},\tilde{v})\tilde{\boldsymbol{\sigma}}_{\tilde{v}},$$

whose components a, b, c, d are smooth functions of $(\tilde{u},\tilde{v})$. Assume that, at some point $(\tilde{u}_0,\tilde{v}_0) \in \tilde{U}$, the vectors $\boldsymbol{e}_1(\tilde{u}_0,\tilde{v}_0)$ and $\boldsymbol{e}_2(\tilde{u}_0,\tilde{v}_0)$ are linearly independent. Then, there is an open subset $\tilde{V}$ of $\tilde{U}$ containing $(\tilde{u}_0,\tilde{v}_0)$ and a reparametrization $\boldsymbol{\sigma}(u,v)$ of $\tilde{\boldsymbol{\sigma}}(\tilde{u},\tilde{v})$, for $(\tilde{u},\tilde{v}) \in \tilde{V}$, such that $\boldsymbol{\sigma}_u$ and $\boldsymbol{\sigma}_v$ are parallel to $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$, respectively.

Proposition 8.4.1 is a special case of Proposition 8.4.3. In fact, let $\tilde{\boldsymbol{\sigma}}$ be any surface patch of $\mathcal{S}$ containing $\mathbf{p}$, and let $\mathbf{p} = \tilde{\boldsymbol{\sigma}}(\tilde{u}_0, \tilde{v}_0)$. Since the principal curvatures κ_1 and κ_2 of $\tilde{\boldsymbol{\sigma}}$ are distinct at $\mathbf{p}$, and are continuous functions by Exercise 8.2.8, they remain distinct for $(\tilde{u}, \tilde{v})$ in some open set $\tilde{U}$ containing $(\tilde{u}_0, \tilde{v}_0)$ on which $\tilde{\boldsymbol{\sigma}}$ is defined. Let

$$\boldsymbol{e}_1 = \xi_1 \tilde{\boldsymbol{\sigma}}_{\tilde{u}} + \eta_1 \tilde{\boldsymbol{\sigma}}_{\tilde{v}}, \quad \boldsymbol{e}_2 = \xi_2 \tilde{\boldsymbol{\sigma}}_{\tilde{u}} + \eta_2 \tilde{\boldsymbol{\sigma}}_{\tilde{v}}$$

be unit principal vectors corresponding to κ_1 and κ_2; they are perpendicular by Proposition 8.2.1. Let $\boldsymbol{\sigma}(u, v)$ be a reparametrization of $\tilde{\boldsymbol{\sigma}}$ as in Proposition 8.4.3. Then, $\boldsymbol{\sigma}_u \cdot \boldsymbol{\sigma}_v = 0$ because $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are perpendicular, so the first fundamental form of $\boldsymbol{\sigma}$ is of the form $Edu^2 + Gdv^2$. Also, $\boldsymbol{\sigma}_u$ and $\boldsymbol{\sigma}_v$ are principal vectors corresponding to κ_1 and κ_2, so we have

$$(\mathcal{F}_{II} - \kappa_1 \mathcal{F}_I) \begin{pmatrix} 1 \\ 0 \end{pmatrix} = (\mathcal{F}_{II} - \kappa_2 \mathcal{F}_I) \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

where $\mathcal{F}_I$ and $\mathcal{F}_{II}$ are the matrices associated to the first and second fundamental forms of $\boldsymbol{\sigma}$. Since $\mathcal{F}_I = \begin{pmatrix} E & 0 \\ 0 & G \end{pmatrix}$, these equations imply that $\mathcal{F}_{II} = \begin{pmatrix} \kappa_1 E & 0 \\ 0 & \kappa_2 G \end{pmatrix}$, so the second fundamental form of $\boldsymbol{\sigma}$ is $Ldu^2 + Ndv^2$, where $L = \kappa_1 E$ and $N = \kappa_2 G$.

We are thus left with the proof of Proposition 8.4.3. To begin, we observe that, if

$$\boldsymbol{e} = A \tilde{\boldsymbol{\sigma}}_{\tilde{u}} + B \tilde{\boldsymbol{\sigma}}_{\tilde{v}},$$

where A and B are any given smooth functions of $(\tilde{u}, \tilde{v}) \in \tilde{U}$, we can find a curve $\boldsymbol{\gamma}$ in $\tilde{\boldsymbol{\sigma}}$ with $\dot{\boldsymbol{\gamma}} = \boldsymbol{e}$ and with any given point $\mathbf{q} = \tilde{\boldsymbol{\sigma}}(\alpha, \beta)$ as starting point $\boldsymbol{\gamma}(0)$. For, finding such a curve $\boldsymbol{\gamma}(t) = \tilde{\boldsymbol{\sigma}}(\tilde{u}(t), \tilde{v}(t))$ is equivalent to solving the pair of ordinary differential equations

$$\dot{\tilde{u}} = A(\tilde{u}, \tilde{v}), \quad \dot{\tilde{v}} = B(\tilde{u}, \tilde{v})$$

with initial conditions $\tilde{u}(0) = \alpha$, $\tilde{v}(0) = \beta$. It is proved in the theory of ordinary differential equations that this problem has a unique solution $\tilde{u}(t), \tilde{v}(t)$ defined on some open interval containing $t = 0$. Moreover, $\tilde{u}$ and $\tilde{v}$ are smooth functions of the three variables t, α and β.

Applying this observation to $\boldsymbol{e} = \boldsymbol{e}_1$, we can find a curve $\boldsymbol{\gamma}_1(s_1)$ in $\tilde{\boldsymbol{\sigma}}$ with $\boldsymbol{\gamma}_1(0) = \tilde{\boldsymbol{\sigma}}(\tilde{u}_0, \tilde{v}_0)$ and $d\boldsymbol{\gamma}_1/ds_1 = \boldsymbol{e}_1$. Now applying the same observation to $\boldsymbol{e} = \boldsymbol{e}_2$, we can find, for each value of s_1 close to 0, a curve $s_2 \mapsto \boldsymbol{\lambda}(s_1, s_2)$ in $\tilde{\boldsymbol{\sigma}}$ with $\partial \boldsymbol{\lambda}/\partial s_2 = \boldsymbol{e}_2$ and $\boldsymbol{\lambda}(s_1, 0) = \boldsymbol{\gamma}_1(s_1)$. Define $(\tilde{u}, \tilde{v})$ as functions of (s_1, s_2) by

$$\tilde{\boldsymbol{\sigma}}(\tilde{u}, \tilde{v}) = \boldsymbol{\lambda}(s_1, s_2). \tag{8.11}$$

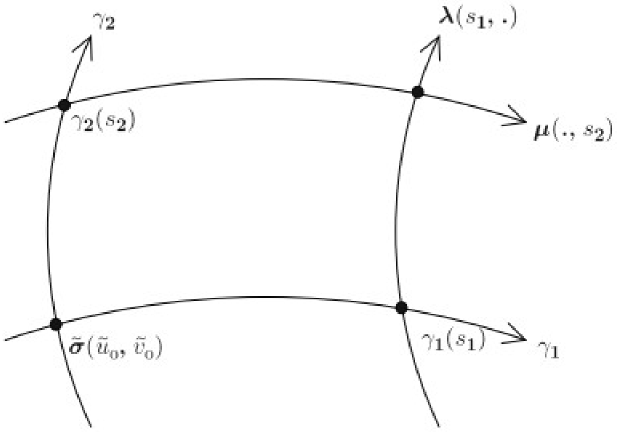
γ_2
$\boldsymbol{\lambda}(s_1, .)$
$\gamma_2(s_2)$
$\boldsymbol{\mu}(., s_2)$
$\tilde{\boldsymbol{\sigma}}(\tilde{u}_0, \tilde{v}_0)$
$\gamma_1(s_1)$
γ_1

Differentiating with respect to s_1 and s_2 gives

$$\tilde{\boldsymbol{\sigma}}_{\tilde{u}}\frac{\partial \tilde{u}}{\partial s_1} + \tilde{\boldsymbol{\sigma}}_{\tilde{v}}\frac{\partial \tilde{v}}{\partial s_1} = \boldsymbol{\lambda}_{s_1}, \quad \tilde{\boldsymbol{\sigma}}_{\tilde{u}}\frac{\partial \tilde{u}}{\partial s_2} + \tilde{\boldsymbol{\sigma}}_{\tilde{v}}\frac{\partial \tilde{v}}{\partial s_2} = \boldsymbol{\lambda}_{s_2}.$$

We have

$$\boldsymbol{\lambda}_{s_1}|_{s_2=0} = \frac{d}{ds_1}\boldsymbol{\lambda}(s_1, 0) = \frac{d\boldsymbol{\gamma}_1}{ds_1} = \boldsymbol{e}_1, \quad \boldsymbol{\lambda}_{s_2} = \frac{\partial \lambda}{\partial s_2} = \boldsymbol{e}_2. \tag{8.12}$$

Equating coefficients of $\tilde{\boldsymbol{\sigma}}_{\tilde{u}}$ and $\tilde{\boldsymbol{\sigma}}_{\tilde{v}}$, we see from the last two sets of equations that, at the point $\tilde{\boldsymbol{\sigma}}(\tilde{u}_0, \tilde{v}_0)$, where $s_1 = s_2 = 0$, the Jacobian matrix

$$\begin{pmatrix} \frac{\partial \tilde{u}}{\partial s_1} & \frac{\partial \tilde{u}}{\partial s_2} \\ \frac{\partial \tilde{v}}{\partial s_1} & \frac{\partial \tilde{v}}{\partial s_2} \end{pmatrix} = \begin{pmatrix} a & c \\ b & d \end{pmatrix}. \tag{8.13}$$

Since $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are linearly independent at $(\tilde{u}_0, \tilde{v}_0)$, this matrix is invertible. By the Inverse Function Theorem 5.6.1, Eq. 8.11 can be solved for (s_1, s_2) as smooth functions of $(\tilde{u}, \tilde{v})$ when $(\tilde{u}, \tilde{v})$ is in some open set $\tilde{W}$ of $\tilde{U}$ containing $(\tilde{u}_0, \tilde{v}_0)$. Thus, $\boldsymbol{\lambda}$ is an allowable surface patch; by Eq. 8.12, it has the property that $\boldsymbol{\lambda}_{s_1} = \boldsymbol{e}_1$ when $s_2 = 0$, and $\boldsymbol{\lambda}_{s_2} = \boldsymbol{e}_2$ everywhere.

We now repeat the procedure, this time starting with a curve $\boldsymbol{\gamma}_2(t_2)$ with $d\boldsymbol{\gamma}_2/dt_2 = \boldsymbol{e}_2$ and $\boldsymbol{\gamma}_2(0) = \tilde{\boldsymbol{\sigma}}(\tilde{u}_0, \tilde{v}_0)$, and then taking a curve $t_1 \mapsto \boldsymbol{\mu}(t_1, t_2)$ with $\partial\boldsymbol{\mu}/\partial t_1 = \boldsymbol{e}_1$ and $\boldsymbol{\mu}(0, t_2) = \boldsymbol{\gamma}_2(t_2)$. This gives an allowable patch $\boldsymbol{\mu}(t_1, t_2)$ such that

$$\boldsymbol{\mu}(t_1, t_2) = \tilde{\boldsymbol{\sigma}}(\tilde{u}, \tilde{v})$$

for $(\tilde{u}, \tilde{v})$ in some open subset $\tilde{Z}$ of $\tilde{U}$ containing $(\tilde{u}_0, \tilde{v}_0)$. This patch has the property that $\boldsymbol{\mu}_{t_1} = \boldsymbol{e}_1$ everywhere and $\boldsymbol{\mu}_{t_2} = \boldsymbol{e}_2$ when $t_1 = 0$.

The parametrization we want is $\boldsymbol{\sigma}(u,v)$, where $\boldsymbol{\sigma}(u,v)$ is the intersection of the curve $s_2 \mapsto \boldsymbol{\lambda}(u,s_2)$ with the curve $t_1 \mapsto \boldsymbol{\mu}(t_1,v)$. Thus, we consider the equations

$$\tilde{\boldsymbol{\sigma}}(\tilde{u},\tilde{v}) = \boldsymbol{\lambda}(u,s_2) = \boldsymbol{\mu}(t_1,v).$$

From Eq. 8.13,

$$\frac{\partial \tilde{u}}{\partial u} = a, \quad \frac{\partial \tilde{v}}{\partial u} = b,$$

and similarly

$$\frac{\partial \tilde{u}}{\partial v} = c, \quad \frac{\partial \tilde{v}}{\partial v} = d.$$

Hence, the Jacobian matrix

$$\begin{pmatrix} \frac{\partial \tilde{u}}{\partial u} & \frac{\partial \tilde{u}}{\partial v} \\ \frac{\partial \tilde{v}}{\partial u} & \frac{\partial \tilde{v}}{\partial v} \end{pmatrix} = \begin{pmatrix} a & c \\ b & d \end{pmatrix}.$$

As usual, the fact that this matrix is invertible means that (u,v) can be expressed as smooth functions of $(\tilde{u},\tilde{v})$, for $(\tilde{u},\tilde{v})$ in some open subset $\tilde{V}$ of $\tilde{W} \cap \tilde{Z}$ containing $(\tilde{u}_0,\tilde{v}_0)$, and we get a reparametrization $\boldsymbol{\sigma}(u,v)$ of $\tilde{\boldsymbol{\sigma}}(\tilde{u},\tilde{v})$. Finally, the equation $\boldsymbol{\sigma}(u,v) = \boldsymbol{\mu}(t_1,v)$ implies that

$$\boldsymbol{\sigma}_u = \frac{\partial t_1}{\partial u}\boldsymbol{\mu}_{t_1} = \frac{\partial t_1}{\partial u}\boldsymbol{e}_1,$$

and similarly

$$\boldsymbol{\sigma}_v = \frac{\partial s_2}{\partial v}\boldsymbol{e}_2,$$

so $\boldsymbol{\sigma}_u$ and $\boldsymbol{\sigma}_v$ are parallel to $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ everywhere. □

EXERCISES

8.4.1 Let **p** be a hyperbolic point of a surface $\mathcal{S}$ (see Section 8.2). Show that there is a patch of $\mathcal{S}$ containing **p** whose parameter curves are asymptotic curves (see Exercise 7.3.6). Show that the second fundamental form of such a patch is of the form $2M du dv$.

8.5 Surfaces of constant mean curvature

We now consider surfaces whose mean curvature H is constant. Such surfaces have an interesting physical interpretation: we shall show in Section 12.1 that soap bubbles always adopt the form of a surface of constant mean curvature. In this section we give two simple constructions of surfaces of constant non-zero mean curvature; the case in which $H = 0$ is treated in much more detail in Chapter 12.

The first of these gives a correspondence between surfaces of constant non-zero mean curvature and surfaces of constant positive Gaussian curvature.

Definition 8.5.1

Let $\mathcal{S}$ be an oriented surface and let $\lambda \in \mathbb{R}$. The *parallel surface* $\mathcal{S}^\lambda$ of $\mathcal{S}$ is

$$\mathcal{S}^\lambda = \{\mathbf{p} + \lambda \mathbf{N}_\mathbf{p} \mid \mathbf{p} \in \mathcal{S}\},$$

where $\mathbf{N}_\mathbf{p}$ is the unit normal of $\mathcal{S}$ at the point $\mathbf{p}$.

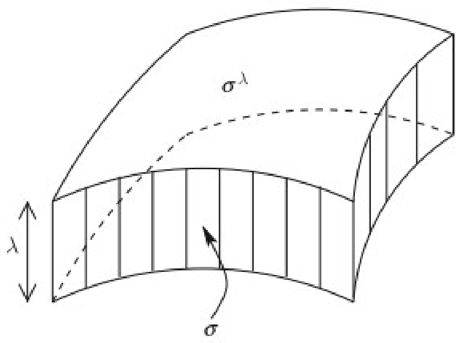

Roughly speaking, $\mathcal{S}^\lambda$ is obtained by translating the surface $\mathcal{S}$ at a distance λ perpendicular to itself (but this will not be a genuine translation since $\mathbf{N}_\mathbf{p}$ will in general depend on $\mathbf{p}$).

Proposition 8.5.2

Let κ_1 and κ_2 be the principal curvatures of an oriented surface $\mathcal{S}$, let $\lambda \in \mathbb{R}$ and let $\mathcal{S}^\lambda$ be the corresponding parallel surface of $\mathcal{S}$. Assume that neither κ_1 nor κ_2 is equal to $1/\lambda$ at any point of $\mathcal{S}$. Then,

(i) $\mathcal{S}^\lambda$ is a (smooth) oriented surface, the unit normal of $\mathcal{S}^\lambda$ at $\mathbf{p} + \lambda \mathbf{N}_\mathbf{p}$ being equal to $\epsilon \mathbf{N}_\mathbf{p}$, where ϵ is the sign of $(1 - \lambda\kappa_1)(1 - \lambda\kappa_2)$.

(ii) The principal curvatures of $\mathcal{S}^\lambda$ are $\epsilon\kappa_1/(1 - \lambda\kappa_1)$ and $\epsilon\kappa_2/(1 - \lambda\kappa_2)$, and the corresponding principal vectors are the same as those of $\mathcal{S}$ for the principal curvatures κ_1 and κ_2, respectively.

(iii) The Gaussian and mean curvatures of $\mathcal{S}^\lambda$ are

$$\frac{K}{1 - 2\lambda H + \lambda^2 K} \quad \text{and} \quad \frac{\epsilon(H - \lambda K)}{1 - 2\lambda H + \lambda^2 K},$$

respectively, where K and H are the Gaussian and mean curvatures of $\mathcal{S}$.

Proof

Let $\boldsymbol{\sigma}(u,v)$ be a surface patch of $\mathcal{S}$ with standard unit normal $\mathbf{N}(u,v)$. Define

$$\boldsymbol{\sigma}^{\lambda}(u,v) = \boldsymbol{\sigma}(u,v) + \lambda\mathbf{N}(u,v).$$

By Proposition 8.1.2,

$$\begin{aligned} \boldsymbol{\sigma}_u^{\lambda} &= \boldsymbol{\sigma}_u + \lambda\mathbf{N}_u = (1-\lambda a)\,\boldsymbol{\sigma}_u - \lambda b\,\boldsymbol{\sigma}_v, \\ \boldsymbol{\sigma}_v^{\lambda} &= \boldsymbol{\sigma}_v + \lambda\mathbf{N}_v = -\lambda c\,\boldsymbol{\sigma}_u + (1-\lambda d)\,\boldsymbol{\sigma}_v, \end{aligned} \tag{8.14}$$

where

$$\mathcal{W}_{\boldsymbol{\sigma}} = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$$

is the matrix of the Weingarten map of $\mathcal{S}$ with respect to the basis $\{\boldsymbol{\sigma}_u, \boldsymbol{\sigma}_v\}$ of the tangent plane. Hence,

$$\boldsymbol{\sigma}_u^{\lambda} \times \boldsymbol{\sigma}_v^{\lambda} = (1 - \lambda(a+d) + \lambda^2(ad-bc))\,\boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v.$$

Since κ_1 and κ_2 are the eigenvalues of $\mathcal{W}_{\boldsymbol{\sigma}}$ (see Section 8.2), and since the sum and product of the eigenvalues of a matrix are equal to the trace and the determinant of the matrix, respectively,

$$\kappa_1 + \kappa_2 = a + d, \quad \kappa_1\kappa_2 = ad - bc.$$

Hence,

$$\boldsymbol{\sigma}_u^{\lambda} \times \boldsymbol{\sigma}_v^{\lambda} = (1-\lambda\kappa_1)(1-\lambda\kappa_2)\,\boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v. \tag{8.15}$$

The assertions in part (i) follow from this equation.

The principal curvatures of $\mathcal{S}^{\lambda}$ are the eigenvalues of the matrix $\mathcal{W}_{\boldsymbol{\sigma}^{\lambda}}$ of the Weingarten map of $\mathcal{S}^{\lambda}$ with respect to the basis $\{\boldsymbol{\sigma}_u^{\lambda}, \boldsymbol{\sigma}_v^{\lambda}\}$. By the proof of Proposition 8.1.2, this is the negative of the matrix expressing $\mathbf{N}_u^{\lambda}$ and $\mathbf{N}_v^{\lambda}$ in terms of $\boldsymbol{\sigma}_u^{\lambda}$ and $\boldsymbol{\sigma}_v^{\lambda}$, where $\mathbf{N}^{\lambda}$ is the standard unit normal of $\boldsymbol{\sigma}^{\lambda}$. Equation 14 says that the matrix expressing $\boldsymbol{\sigma}_u^{\lambda}$ and $\boldsymbol{\sigma}_v^{\lambda}$ in terms of $\boldsymbol{\sigma}_u$ and $\boldsymbol{\sigma}_v$ is $I - \lambda\mathcal{W}_{\boldsymbol{\sigma}}$, and the fact that $\mathbf{N}^{\lambda} = \epsilon\mathbf{N}$ implies that $-\epsilon\mathcal{W}_{\boldsymbol{\sigma}}$ is the matrix expressing $\mathbf{N}_u^{\lambda}$ and $\mathbf{N}_v^{\lambda}$ in terms of $\boldsymbol{\sigma}_u$ and $\boldsymbol{\sigma}_v$. Combining these two observations we get

$$\mathcal{W}_{\boldsymbol{\sigma}^{\lambda}} = \epsilon(I - \lambda\mathcal{W}_{\boldsymbol{\sigma}})^{-1}\mathcal{W}_{\boldsymbol{\sigma}}.$$

If T is an eigenvector of $\mathcal{W}_{\boldsymbol{\sigma}}$ with eigenvalue κ, then T is also an eigenvector of $\mathcal{W}_{\boldsymbol{\sigma}^{\lambda}}$ with eigenvalue $\epsilon\kappa/(1-\lambda\kappa)$. The assertions in part (ii) follows from this.

Part (iii) follows from part (ii) by straightforward algebra. □

Corollary 8.5.3

If $\mathcal{S}$ has constant Gaussian curvature $1/R^2$, the parallel surfaces $\mathcal{S}^{\pm R}$ have constant mean curvature $1/2R$. Conversely, if $\mathcal{S}$ has constant mean curvature $1/2R$, the parallel surface $\mathcal{S}^R$ has constant Gaussian curvature $1/R^2$.

Proof

This follows from part (iii) of the proposition by straightforward algebra. For example, if $H = 1/2R$ the Gaussian curvature of $\mathcal{S}^R$ is

$$\frac{K}{1 - 2RH + R^2K} = \frac{K}{R^2K} = \frac{1}{R^2}. \qquad \square$$

The next construction gives a beautiful geometric description of the surfaces of revolution which have constant non-zero mean curvature in terms of the curve traced out by the focus of an ellipse that rolls without slipping along a straight line (cf. Exercise 2.2.10). Take the ellipse to be

$$\frac{x^2}{p^2} + \frac{(y-q)^2}{q^2} = 1,$$

where $p > q > 0$ are constants. Thus, the ellipse is tangent to the x-axis at the origin. The foci of the ellipse are the points $\mathbf{f}_1 = (-\epsilon p, q)$ and $\mathbf{f}_2 = (\epsilon p, q)$, where the eccentricity $\epsilon = \sqrt{1 - \frac{q^2}{p^2}}$.

Proposition 8.5.4

With the above notation, let $\mathcal{C}$ be the curve traced out by one of the foci of the ellipse as it rolls without slipping along the x-axis. Let $\mathcal{S}$ be the surface obtained by rotating $\mathcal{C}$ around the x-axis. Then, $\mathcal{S}$ has constant non-zero mean curvature.

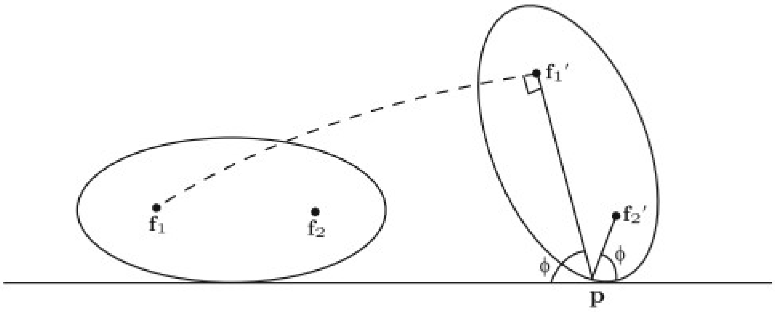

Proof

We consider a situation in which the ellipse has rolled along the x-axis so that its point of contact with the x-axis is at a point $\mathbf{p}$, the focus $\mathbf{f}_1$ has moved to a point ${\mathbf{f}_1}' = (x, y)$ on $\mathcal{C}$, and the focus $\mathbf{f}_2$ has moved to ${\mathbf{f}_2}' = (X, Y)$, say. Let φ be the angle between $\mathbf{p} - {\mathbf{f}_1}'$ and the x-axis; then φ is also the angle between $\mathbf{p} - {\mathbf{f}_2}'$ and the x-axis by Exercise 1.1.6(iii). Hence,

$$y = \parallel \mathbf{p} - {\mathbf{f}_1}' \parallel \sin\varphi, \quad Y = \parallel \mathbf{p} - {\mathbf{f}_2}' \parallel \sin\varphi$$

and so

$$y + Y = 2p\sin\varphi$$

by Exercise 1.1.6(i). But Exercise 1.1.6(ii) gives $yY = q^2$ so

$$y + \frac{q^2}{y} = 2p\sin\varphi. \tag{8.16}$$

Now, since the ellipse rolls without slipping, the point of contact of the ellipse with the x-axis is stationary. This implies that the point ${\mathbf{f}_1}'$ moves as if rotating instantaneously about $\mathbf{p}$, so that the tangent vector to $\mathcal{C}$ at ${\mathbf{f}_1}'$ is perpendicular to $\mathbf{p} - {\mathbf{f}_1}'$. (If this heuristic argument is unconvincing, an analytical proof can be found in Exercise 2.2.10.) It follows that

$$\frac{dy}{dx} = \cot\varphi. \tag{8.17}$$

Eliminating φ between Eqs. 8.16 and 8.17 gives

$$y^2 + q^2 = \frac{2py}{\sqrt{1 + \left(\frac{dy}{dx}\right)^2}}. \tag{8.18}$$

The surface $\mathcal{S}$ obtained by rotating $\mathcal{C}$ around the x-axis can be parametrized by

$$\boldsymbol{\sigma}(x, \theta) = (x, y\cos\theta, y\sin\theta)$$

where θ is the angle of rotation. The first and second fundamental forms of $\boldsymbol{\sigma}$ are

$$\left(1 + \left(\frac{dy}{dx}\right)^2\right) dx^2 + y^2 d\theta^2 \quad \text{and} \quad -\frac{d^2y}{dx^2}dx^2 + \frac{y}{\sqrt{1 + \left(\frac{dy}{dx}\right)^2}}d\theta^2,$$

respectively. Using the formula in Corollary 8.1.3, the mean curvature is found to be

$$H = \frac{1}{2y\sqrt{1 + \left(\frac{dy}{dx}\right)^2}} - \frac{\frac{d^2y}{dx^2}}{2\left(1 + \left(\frac{dy}{dx}\right)^2\right)^{3/2}}. \tag{8.19}$$

Differentiating both sides of Eq. 8.18 we get

$$2y\frac{dy}{dx} = \frac{2p\frac{dy}{dx}}{\sqrt{1+\left(\frac{dy}{dx}\right)^2}} - \frac{2py\frac{dy}{dx}\frac{d^2y}{dx^2}}{\left(1+\left(\frac{dy}{dx}\right)^2\right)^{3/2}}.$$

Dividing both sides by $4py\frac{dy}{dx}$ and comparing with Eq. 8.19 shows that the surface $\mathcal{S}$ has mean curvature $1/2p$. ☐

EXERCISES

8.5.1 Suppose that the first fundamental form of a surface patch $\boldsymbol{\sigma}(u,v)$ is of the form $E(du^2+dv^2)$. Prove that $\boldsymbol{\sigma}_{uu}+\boldsymbol{\sigma}_{vv}$ is perpendicular to $\boldsymbol{\sigma}_u$ and $\boldsymbol{\sigma}_v$. Deduce that the mean curvature $H=0$ everywhere if and only if the Laplacian

$$\boldsymbol{\sigma}_{uu}+\boldsymbol{\sigma}_{vv}=\mathbf{0}.$$

Show that the surface patch

$$\boldsymbol{\sigma}(u,v)=\left(u-\frac{u^3}{3}+uv^2, v-\frac{v^3}{3}+u^2v, u^2-v^2\right)$$

has $H=0$ everywhere. (A picture of this surface can be found in Section 12.2.)

8.5.2 Prove that $H=0$ for the surface

$$z=\ln\left(\frac{\cos y}{\cos x}\right).$$

(A picture of this surface can also be found in Section 12.2.)

8.5.3 Let $\boldsymbol{\sigma}(u,v)$ be a surface with first and second fundamental forms Edu^2+Gdv^2 and Ldu^2+Ndv^2, respectively (cf. Proposition 8.4.1). Define

$$\boldsymbol{\Sigma}(u,v,w)=\boldsymbol{\sigma}(u,v)+w\mathbf{N}(u,v),$$

where $\mathbf{N}$ is the standard unit normal of $\boldsymbol{\sigma}$. Show that the three families of surfaces obtained by fixing the values of u, v or w in $\boldsymbol{\Sigma}$ form a triply orthogonal system (see Section 5.5). The surfaces $w=$ constant are parallel surfaces of $\boldsymbol{\sigma}$. Show that the surfaces $u=$ constant and $v=$ constant are flat ruled surfaces.

5- 672 N 3

10

Gauss' Theorema Egregium

One of Gauss' most important discoveries about surfaces is that the Gaussian curvature is unchanged when the surface is bent without stretching. Gauss called this result 'egregium', and the Latin word for 'remarkable' has remained attached to his theorem ever since. We shall deduce the Theorema Egregium from two results which relate the first and second fundamental forms of a surface, and which have other important consequences.

10.1 The Gauss and Codazzi–Mainardi equations

It is natural to ask if there are any relations between the first and second fundamental forms of a surface. Note that, by Examples 6.1.2, 6.1.4, 7.1.1 and 7.1.2, the plane and the unit cylinder, when suitably parametrized, have the same first fundamental form but different second fundamental forms, and so the second fundamental form certainly cannot be 'deduced' from the first. Nevertheless, there are some nontrivial relations between the two forms.

Proposition 10.1.1 (Codazzi–Mainardi Equations)

Let

$$Edu^2 + 2Fdudv + Gdv^2 \quad \text{and} \quad Ldu^2 + 2Mdudv + Ndv^2$$

Andrew Pressley, *Elementary Differential Geometry: Second Edition*,
Springer Undergraduate Mathematics Series, DOI 10.1007/978-1-84882-891-9_10,
© Springer-Verlag London Limited 2010

5- /N2 N<3

(ii) Show that $u^2(\tilde{E}-1) = v^2(\tilde{G}-1)$, and by considering the Taylor expansions of $\tilde{E}$ and $\tilde{G}$ about $u = v = 0$, deduce that

$$G(r, \theta) = r^2 + kr^4 + \text{remainder}$$

for some constant k, where remainder$/r^4$ tends to zero as r tends to zero.

(iii) Show that $k = -K(P)/3$, where $K(P)$ is the value of the Gaussian curvature of $\mathcal{S}$ at P.

10.2.3 With the notation of Exercises 9.5.1 and 10.2.2, show that:

(i) The circumference of the geodesic circle with centre P and radius R is

$$C_R = 2\pi R\left(1 - \frac{K(P)}{6}R^2 + \text{remainder}\right),$$

where remainder$/R^2$ tends to zero as R tends to zero.

(ii) The area inside the geodesic circle in (i) is

$$A_R = \pi R^2\left(1 - \frac{K(P)}{12}R^2 + \text{remainder}\right),$$

where the remainder satisfies the same condition as in (i).

Verify that these formulas are consistent with those in spherical geometry obtained in Exercise 6.5.3.

10.2.4 Let A, B and C be the vertices of a triangle $\mathcal{T}$ on a surface $\mathcal{S}$ whose sides are arcs of geodesics, and let α, β and γ be its internal angles (so that α is the angle at A, etc.). Assume that the triangle is contained in a geodesic patch $\boldsymbol{\sigma}$ as in Exercise 9.5.1 with $P = A$. Thus, with the notation in that exercise, if we take $\mathbf{v}$ to be tangent at A to the side passing through A and B, then the sides meeting at A are the parameter curves $\theta = 0$ and $\theta = \alpha$, and the remaining side can be parametrized by $\boldsymbol{\gamma}(\theta) = \boldsymbol{\sigma}(f(\theta), \theta)$ for some smooth function f and $0 \leq \theta \leq \alpha$.

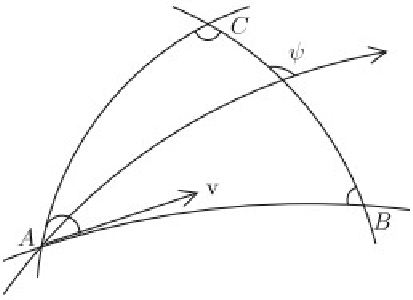

(i) Use the geodesic equations (9.2) to show that

$$f'' - \frac{f'\lambda'}{\lambda^2} = \frac{1}{2}\frac{\partial G}{\partial r},$$

where a dash denotes $d/d\theta$ and $\lambda = \parallel \boldsymbol{\gamma}' \parallel$.

(ii) Show that, if $\psi(\theta)$ is the angle between $\boldsymbol{\sigma}_r$ and the tangent vector to the side opposite A at $\boldsymbol{\gamma}(\theta)$, then

$$\psi'(\theta) = -\frac{\partial\sqrt{G}}{\partial r}(f(\theta), \theta).$$

(iii) Show that, if K is the Gaussian curvature of $\mathcal{S}$,

$$\int_{\mathcal{T}} K \, d\mathcal{A}_{\boldsymbol{\sigma}} = \alpha + \beta + \gamma - \pi.$$

This result will be generalized in Corollary 13.2.3.

10.2.5 Show that the Gaussian curvature of the Möbius band in Example 4.5.3 is equal to $-1/4$ everywhere along its median circle. Deduce that this Möbius band *cannot* be constructed by taking a strip of paper and joining the ends together with a half-twist. (The analytic description of the 'cut and paste' Möbius band is more complicated than the version in Example 4.5.3.)

10.2.6 Show that the only isometries from the catenoid to itself are products of rotations around its axis, reflections in planes containing the axis, and reflection in the plane containing the waist of the catenoid.

10.3 Surfaces of constant Gaussian curvature

The results of the preceding two sections enable us to gain a good understanding of surfaces of constant Gaussian curvature. We begin with the following local description.

Theorem 10.3.1

Any point of a surface of constant Gaussian curvature is contained in a patch that is isometric to an open subset of a plane, a sphere or a pseudosphere.

5- 7 2 T 3

11.1 Upper half-plane model

We saw in Example 9.3.3 that if the pseudosphere is parametrized as

$$\tilde{\boldsymbol{\sigma}}(v, w) = \left(\frac{1}{w} \cos v, \frac{1}{w} \sin v, \sqrt{1 - \frac{1}{w^2}} - \cosh^{-1} w \right),$$

where we must have $w > 1$ for $\tilde{\boldsymbol{\sigma}}$ to be well defined and smooth, its geodesics correspond to arcs of circles and straight lines in the vw-plane that intersect the v-axis perpendicularly. The line $w = 1$ appears to be a rather artificial boundary in the vw-plane, since the geodesics are well defined in the entire region $w > 0$. On the other hand, the line $w = 0$ is a 'real' boundary since the first fundamental form

$$\frac{dv^2 + dw^2}{w^2} \tag{11.1}$$

of the pseudosphere is undefined when $w = 0$. It is therefore natural to ask if there is a surface that corresponds to the whole of the half-plane $w > 0$ with this first fundamental form. In fact, there is no such surface for a celebrated theorem of Hilbert shows that there is no surface with constant negative Gaussian curvature that is 'geodesically complete', i.e., a surface for which all geodesics can be extended indefinitely in both directions (see Exercise 10.1.3).

One possible response to Hilbert's theorem is essentially to ignore it: all those properties of surfaces that depend only on the first fundamental form (lengths, angles, areas, geodesics, local isometries, ...) can still be studied for the half-plane

$$\mathcal{H} = \{(v, w) \in \mathbb{R}^2 \mid w > 0\}$$

equipped with the first fundamental form (11.1). They will then be called hyperbolic lengths, hyperbolic angles, etc. (We shall see the reason for the adjective 'hyperbolic' later.)

It will often be convenient to identify $\mathbb{R}^2$ with the set of complex numbers $\mathbb{C}$ via $(v, w) \leftrightarrow v + iw$, so that

$$\mathcal{H} = \{z \in \mathbb{C} \mid \mathfrak{Im}(z) > 0\}$$

is the set of complex numbers with positive imaginary part.

The first thing to observe is that $\mathcal{H}$ is a 'conformal model':

Proposition 11.1.1

Hyperbolic angles in $\mathcal{H}$ are the same as Euclidean angles.

Proof

This is just because the first fundamental form (11.1) of $\mathcal{H}$ is a multiple of $dv^2 + dw^2$ (see Corollary 6.3.4). $\square$

The 'hyperbolic lines' are the geodesics in $\mathcal{H}$, which were determined in Example 9.3.3.

Proposition 11.1.2

The geodesics in $\mathcal{H}$ are the half-lines parallel to the imaginary axis and the semicircles with centres on the real axis.

Here are some simple properties of hyperbolic lines.

Proposition 11.1.3

(i) There is a unique hyperbolic line passing through any two distinct points of $\mathcal{H}$.

(ii) The parallel axiom does not hold in $\mathcal{H}$.

In the following proof, and later in this chapter, 'lines' and 'circles' will mean Euclidean lines and circles ('hyperbolic line' means 'geodesic'). On the other hand, 'lengths' and 'angles' will always mean hyperbolic lengths and angles, unless explicitly stated otherwise.

Proof

(i) Let $a, b \in \mathcal{H}$, $a \neq b$. If the line passing through a and b is parallel to the imaginary axis, the unique hyperbolic line passing through the points a and b is the half-line containing them. If the line through a and b is not parallel to the imaginary axis, its perpendicular bisector intersects the real axis at some point c, say, and the unique hyperbolic line passing through a and b is the semicircle with centre c and radius $|a - c| = |b - c|$.
(ii) Take l to be the imaginary axis and let $a \in \mathcal{H}$ be any point not on l. For definiteness, assume that the real part $\mathfrak{Re}(a) > 0$. Then, the perpendicular bisector of the line joining a to the origin intersects the real axis at some point $b > 0$. Let c be a real number greater than b; then the semicircle with centre c passing through a is a hyperbolic line in $\mathcal{H}$ that does not intersect l. Of course,

the half-line through a parallel to the imaginary axis is another hyperbolic line with the same property (it can be regarded as the limiting case when $c \to \infty$). □

Since there is a unique hyperbolic line passing through any two points $a, b \in \mathcal{H}$, it makes sense to define the *hyperbolic distance* $d_{\mathcal{H}}(a, b)$ between a and b to be the length of the hyperbolic line segment joining them. It is shown in Exercise 11.2.1 that this is actually the hyperbolic length of the shortest curve joining a and b.

Proposition 11.1.4

The hyperbolic distance between two points $a, b \in \mathcal{H}$ is

$$d_{\mathcal{H}}(a, b) = 2 \tanh^{-1} \frac{|b - a|}{|b - \bar{a}|}.$$

In this formula, $\bar{a}$ denotes the complex conjugate of the complex number a. The appearance of the hyperbolic tangent gives an indication of the reason why the geometry of $\mathcal{H}$ is called 'hyperbolic geometry'.

Proof

There are two cases, depending on whether the hyperbolic line joining a and b is a semicircle or a half-line. We shall deal with the semicircle case, leaving the simpler case of the half-line to Exercise 11.1.2.

Suppose then that a and b lie on the semicircle with centre c on the real axis and radius r. The semicircle can be parametrized by

$$v = c + r\cos\theta, \quad w = r\sin\theta.$$

Writing d for $d_{\mathcal{H}}(a, b)$ and denoting $d/d\theta$ by a dot, we have

$$d = \int_{\varphi}^{\psi} \sqrt{\frac{\dot{v}^2 + \dot{w}^2}{w^2}}\, d\theta = \int_{\varphi}^{\psi} \sqrt{\frac{r^2 \sin^2\theta + r^2\cos^2\theta}{r^2\sin^2\theta}}\, d\theta = \int_{\varphi}^{\psi} \frac{d\theta}{\sin\theta},$$

where $\varphi = \arg(a - c)$, $\psi = \arg(b - c)$ (note that d is independent of the radius of the semicircle). Hence,

$$d = \ln \frac{\tan\frac{\psi}{2}}{\tan\frac{\varphi}{2}}.$$

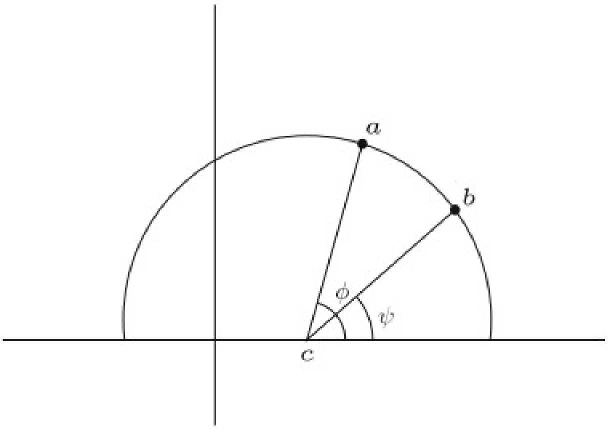

Now,

$$\tanh\frac{d}{2} = \frac{e^d - 1}{e^d + 1} = \frac{\tan\frac{\psi}{2} - \tan\frac{\varphi}{2}}{\tan\frac{\psi}{2} + \tan\frac{\varphi}{2}} = \frac{\sin\frac{\psi}{2}\cos\frac{\varphi}{2} - \cos\frac{\psi}{2}\sin\frac{\varphi}{2}}{\sin\frac{\psi}{2}\cos\frac{\varphi}{2} + \cos\frac{\psi}{2}\sin\varphi 2} = \frac{\sin\frac{\psi-\varphi}{2}}{\sin\frac{\psi+\varphi}{2}}. \tag{11.2}$$

On the other hand,

$$\begin{aligned} |b-a|^2 &= r^2((\cos\psi - \cos\varphi)^2 + (\sin\psi - \sin\varphi)^2) \\ &= 2r^2(1 - \cos(\psi - \varphi)) = 4r^2\sin^2\frac{\psi - \varphi}{2}, \end{aligned}$$

and similarly

$$|b - \bar{a}|^2 = 4r^2\sin^2\frac{\psi + \varphi}{2}.$$

Combining the last two equations with Eq. 11.2 gives

$$\tanh\frac{d}{2} = \frac{|b-a|}{|b-\bar{a}|}. \qquad \square$$

We conclude this section with another beautiful formula, this time for the area of a hyperbolic polygon, i.e., a polygon whose sides are hyperbolic lines.

Theorem 11.1.5

Let $\mathcal{P}$ be a n-sided hyperbolic polygon in $\mathcal{H}$ with internal angles $\alpha_1, \alpha_2, \ldots, \alpha_n$. Then, the hyperbolic area of the polygon is given by

$$\mathcal{A}(\mathcal{P}) = (n-2)\pi - \alpha_1 - \alpha_2 - \cdots - \alpha_n.$$

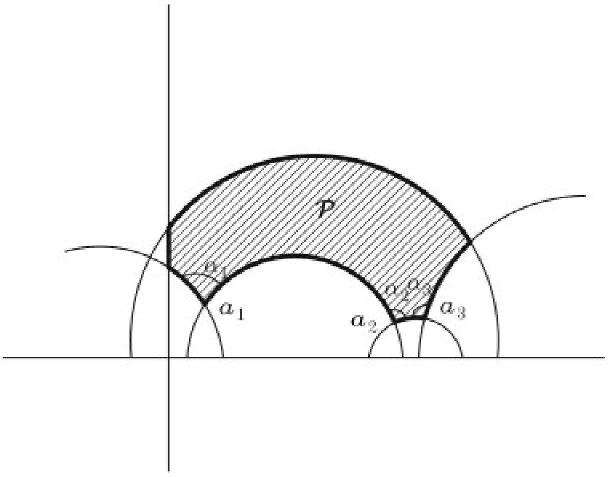

In particular, for a triangle with angles α, β, γ, the area is

$$\pi - \alpha - \beta - \gamma.$$

This should be compared with the well-known formula

$$\alpha + \beta + \gamma = \pi$$

for the sum of the angles of a Euclidean triangle with straight line sides, and the formula

$$\alpha + \beta + \gamma - \pi$$

for the area of a triangle on the unit sphere with geodesic (i.e., great circle) sides (Theorem 6.4.7).

Proof 11.1.5 Let $a_1, \ldots, a_n$ be the vertices of $\mathcal{P}$ and $\mathcal{C}$ its boundary, consisting of n hyperbolic line segments $a_1a_2, a_2a_3, \ldots, a_na_1$ (we assume that α_1 is the internal angle of $\mathcal{P}$ at the vertex a_1, etc.). Since the first fundamental form is $(dv^2 + dw^2)/w^2$, the area of $\mathcal{P}$ is

$$\int_{\mathcal{P}} \frac{dvdw}{w^2}.$$

We evaluate this integral by using Green's theorem (Section 3.2):

$$\int_{\mathcal{C}} pdv + qdw = \int_{\mathcal{P}} \left(\frac{\partial q}{\partial v} - \frac{\partial p}{\partial w} \right) dvdw,$$

where p and q are smooth functions of (v, w). Taking $p = 1/w$ and $q = 0$ gives

$$\int_{\mathcal{P}} \frac{dv dw}{w^2} = \int_{\mathcal{C}} \frac{dv}{w}. \tag{11.3}$$

To evaluate this integral we first prove the following lemma.

Lemma 11.1.6

Let a and b be the endpoints of a segment l of a hyperbolic line in $\mathcal{H}$ that forms part of a semicircle with centre p on the real axis, and suppose that the radius vectors joining p to a and p to b make angles φ and ψ, respectively, with the positive real axis (see the diagram in the proof of Proposition 11.1.4). Then,

$$\int_l \frac{dv}{w} = \varphi - \psi.$$

Note that the integral is independent of the radius of the semicircle, and that the formula is correct even if the hyperbolic line is part of a half-line, for in that case the integral vanishes since v is constant along the hyperbolic line.

Proof 11.1.6 We parametrize the hyperbolic line by $v = r\cos\theta$, $w = r\sin\theta$, where r is the radius of the semicircle. Then, the integral is

$$\int_{\varphi}^{\psi} \frac{-r\sin\theta\, d\theta}{r\sin\theta} = -\int_{\varphi}^{\psi} d\theta = \varphi - \psi. \qquad \square$$

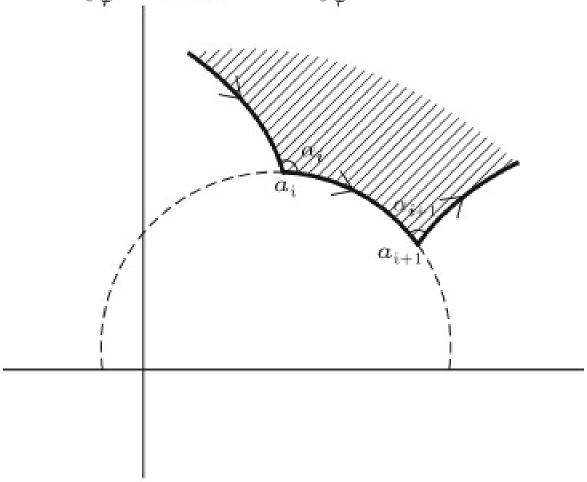

Returning to the proof of Theorem 11.1.5, let φ_i and ψ_i be the angles defined in the lemma corresponding to the side with endpoints a_i and a_{i+1}, for $i = 1, \ldots, n$ (it is understood that a_{n+1} means a_1). By Eq. 11.3 and the lemma,

$$\int_{\mathcal{P}} \frac{dvdw}{w^2} = \sum_{i=1}^{n} (\varphi_i - \psi_i). \tag{11.4}$$

We can simplify this sum by considering the change in direction of the outward-pointing normal of $\mathcal{P}$ as we traverse its boundary in an anticlockwise direction. As we traverse the side with endpoints a_i and a_{i+1}, the outward normal rotates anticlockwise through an angle $\psi_i - \varphi_i$, while at the vertex a_i it rotates by $\pi - \alpha_i$. Hence, as we traverse the boundary of $\mathcal{P}$, the outward normal rotates through an angle

$$n\pi + \sum_{i=1}^{n} (\psi_i - \varphi_i - \alpha_i).$$

But this angle of rotation is 2π (cf. Theorem 3.1.4), so we have the equation

$$2\pi = n\pi + \sum_{i=1}^{n} (\psi_i - \varphi_i - \alpha_i).$$

Rearranging, we get

$$\sum_{i=1}^{n} (\varphi_i - \psi_i) = (n-2)\pi - \sum_{i=1}^{n} \alpha_i.$$

By Eq. 11.4, this is the desired area. □

Note that the area of a hyperbolic triangle, i.e., a triangle whose sides are hyperbolic lines, depends only on its angles. We found in Proposition 6.5.8 that the same result holds in spherical geometry, but as we noted there this is completely different to the Euclidean situation, where we can change the size of a triangle (and hence its area) without changing its angles. In fact, we shall show in the next section that, as in spherical geometry (Exercise 6.5.2), two hyperbolic triangles with the same angles are *congruent*. But first we must discuss what congruence means in hyperbolic geometry.

EXERCISES

11.1.1 Show that, if l is a half-line geodesic in $\mathcal{H}$ and a is a point not on l, there are infinitely many hyperbolic lines passing through a that do not intersect l.

5-   /2 6T3

13

The Gauss–Bonnet theorem

The Gauss–Bonnet theorem is the most beautiful and profound result in the theory of surfaces. Its most important version relates the average of the Gaussian curvature to a property of the surface called its 'Euler number' which is 'topological', i.e., it is unchanged by any diffeomorphism of the surface. Such diffeomorphisms will in general change the value of the Gaussian curvature, but the theorem says that its average over the surface does *not* change. The real importance of the Gauss–Bonnet theorem is as a prototype of analogous results which apply in higher dimensional situations, and which relate *geometrical* properties to *topological* ones. The study of such relations was one of the most important themes of twentieth century mathematics, and continues to be actively studied today.

13.1 Gauss–Bonnet for simple closed curves

The simplest version of the Gauss–Bonnet Theorem involves *simple closed curves* on a surface. In the special case when the surface is a plane, these curves have been discussed in Section 3.1. For a general surface, we make the following definition.

Andrew Pressley, *Elementary Differential Geometry: Second Edition*,
Springer Undergraduate Mathematics Series, DOI 10.1007/978-1-84882-891-9_13,
© Springer-Verlag London Limited 2010

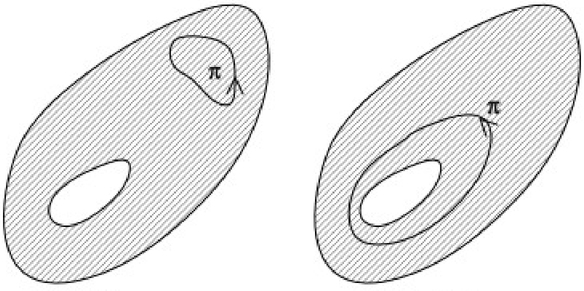

Definition 13.1.1

A curve $\boldsymbol{\gamma}(t) = \boldsymbol{\sigma}(u(t), v(t))$ on a surface patch $\boldsymbol{\sigma} : U \to \mathbb{R}^3$ is called a *simple closed curve* with *period* T if $\boldsymbol{\pi}(t) = (u(t), v(t))$ is a simple closed curve in $\mathbb{R}^2$ with period T such that the region $\text{int}(\boldsymbol{\pi})$ of $\mathbb{R}^2$ enclosed by $\boldsymbol{\pi}$ is entirely contained in U (see the diagrams above). The curve $\boldsymbol{\gamma}$ is said to be *positively-oriented* if $\boldsymbol{\pi}$ is positively-oriented. Finally, the image of $\text{int}(\boldsymbol{\pi})$ under the map $\boldsymbol{\sigma}$ is defined to be the *interior* $\text{int}(\boldsymbol{\gamma})$ of $\boldsymbol{\gamma}$.

We can now state the first version of the Gauss–Bonnet Theorem.

Theorem 13.1.2

Let $\boldsymbol{\gamma}(s)$ be a unit-speed simple closed curve on a surface patch $\boldsymbol{\sigma}$ of length $\ell(\boldsymbol{\gamma})$, and assume that $\boldsymbol{\gamma}$ is positively-oriented. Then,

$$\int_0^{\ell(\boldsymbol{\gamma})} \kappa_g ds = 2\pi - \int_{\text{int}(\boldsymbol{\gamma})} K d\mathcal{A}_{\boldsymbol{\sigma}},$$

where κ_g is the geodesic curvature of $\boldsymbol{\gamma}$, K is the Gaussian curvature of $\boldsymbol{\sigma}$ and $d\mathcal{A}_{\boldsymbol{\sigma}}$ is the area element of $\boldsymbol{\sigma}$ (see Section 6.4).

We use s to denote the parameter of $\boldsymbol{\gamma}$ to emphasize that $\boldsymbol{\gamma}$ is unit-speed. The double integral on the right-hand side of the equation in Theorem 13.1.2 is called the *total curvature* of the region $\text{int}(\boldsymbol{\gamma})$.

Proof

We start by computing the geodesic curvature of $\boldsymbol{\gamma}$. For this, we shall make use of a smooth *orthonormal* basis $\{\boldsymbol{e}', \boldsymbol{e}''\}$ of the tangent plane at each point of the surface patch, where 'smooth' means that $\boldsymbol{e}'$ and $\boldsymbol{e}''$ are smooth functions of the surface parameters (u, v). Then, $\{\boldsymbol{e}', \boldsymbol{e}'', \mathbf{N}\}$ is an orthonormal basis of $\mathbb{R}^3$ ($\mathbf{N}$ being the standard unit normal of $\boldsymbol{\sigma}$), and we shall assume that it is *right-handed*, i.e., that $\mathbf{N} = \boldsymbol{e}' \times \boldsymbol{e}''$. This can always be achieved by interchanging $\boldsymbol{e}'$ and $\boldsymbol{e}''$ if necessary. Note that the dashes on $\boldsymbol{e}'$ and $\boldsymbol{e}''$ have nothing to do with derivatives.

Let $\theta(s)$ be the oriented angle $\widehat{\boldsymbol{e}'\dot{\boldsymbol{\gamma}}}$ between the unit tangent vector $\dot{\boldsymbol{\gamma}}(s)$ of $\boldsymbol{\gamma}$ at $\boldsymbol{\gamma}(s)$ and the unit vector $\boldsymbol{e}'$ at the same point. Thus,

$$\dot{\boldsymbol{\gamma}} = \cos\theta \boldsymbol{e}' + \sin\theta \boldsymbol{e}''. \tag{13.1}$$

Then,

$$\mathbf{N} \times \dot{\boldsymbol{\gamma}} = -\sin\theta \boldsymbol{e}' + \cos\theta \boldsymbol{e}''. \tag{13.2}$$

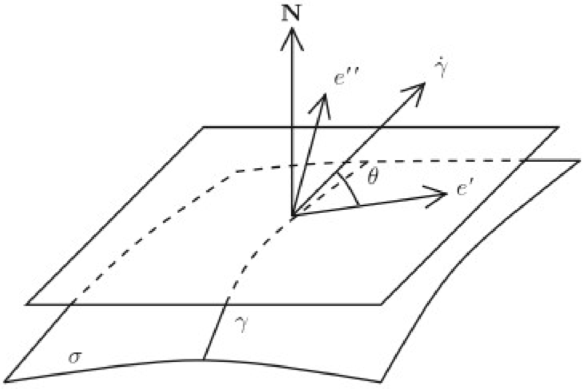

Now, by Eq. 13.1,

$$\ddot{\boldsymbol{\gamma}} = \cos\theta \dot{\mathbf{e}}' + \sin\theta \dot{\mathbf{e}}'' + \dot{\theta}(-\sin\theta \boldsymbol{e}' + \cos\theta \boldsymbol{e}''), \tag{13.3}$$

so by Eqs. 13.2 and 13.3 the geodesic curvature of $\boldsymbol{\gamma}$ is

$$\begin{aligned}
\kappa_g &= (\mathbf{N} \times \dot{\boldsymbol{\gamma}}) \cdot \ddot{\boldsymbol{\gamma}} \quad \text{(see Section 7.3)} \\
&= \dot{\theta}(-\sin\theta \boldsymbol{e}' + \cos\theta \boldsymbol{e}'') \cdot (-\sin\theta \boldsymbol{e}' + \cos\theta \boldsymbol{e}'') \\
&\qquad +(-\sin\theta \boldsymbol{e}' + \cos\theta \boldsymbol{e}'') \cdot (\cos\theta \dot{\mathbf{e}}' + \sin\theta \dot{\mathbf{e}}'') \\
&= \dot{\theta} + \cos^2\theta(\dot{\mathbf{e}}' \cdot \boldsymbol{e}'') - \sin^2\theta(\dot{\mathbf{e}}'' \cdot \boldsymbol{e}') \\
&\qquad + \sin\theta\cos\theta(\dot{\mathbf{e}}'' \cdot \boldsymbol{e}'' - \dot{\mathbf{e}}' \cdot \boldsymbol{e}') \quad \text{(by Eqs. 13.1 and 13.2).}
\end{aligned}$$

Since $\boldsymbol{e}'$ and $\boldsymbol{e}''$ are perpendicular unit vectors,

$$\boldsymbol{e}' \cdot \dot{\mathbf{e}}' = \boldsymbol{e}'' \cdot \dot{\mathbf{e}}'' = 0, \quad \dot{\mathbf{e}}' \cdot \boldsymbol{e}'' = -\boldsymbol{e}' \cdot \dot{\mathbf{e}}''.$$

Hence,

$$\kappa_g = \dot{\theta} - \boldsymbol{e}' \cdot \dot{\mathbf{e}}''. \tag{13.4}$$

Thus, to compute the left-hand side of the equation in Theorem 13.1.2, we must compute the integrals of $\dot{\theta}$ and of $\boldsymbol{e}' \cdot \dot{\boldsymbol{e}}''$ around the curve $\boldsymbol{\gamma}$. We begin with the latter, for which we shall need the following lemma.

Lemma 13.1.3

Let

$$Edu^2 + 2Fdudv + Gdv^2 \quad \text{and} \quad Ldu^2 + 2Mdudv + Ndv^2$$

be the first and second fundamental forms of $\boldsymbol{\sigma}$, respectively. Then, with the above notation, we have

$$\mathbf{e}'_u \cdot \mathbf{e}''_v - \mathbf{e}''_u \cdot \mathbf{e}'_v = \frac{LN - M^2}{(EG - F^2)^{1/2}}. \tag{13.5}$$

Assuming this for a moment, we compute

$$\int_0^{\ell(\boldsymbol{\gamma})} \boldsymbol{e}' \cdot \dot{\boldsymbol{e}}'' ds = \int_0^{\ell(\boldsymbol{\gamma})} \boldsymbol{e}' \cdot (\dot{u}\boldsymbol{e}''_u + \dot{v}\boldsymbol{e}''_v) ds = \int_{\boldsymbol{\pi}} (\boldsymbol{e}' \cdot \boldsymbol{e}''_u) du + (\boldsymbol{e}' \cdot \boldsymbol{e}''_v) dv.$$

By Green's theorem (see Section 3.2), this can be rewritten as a double integral:

$$\begin{aligned}
\int_0^{\ell(\boldsymbol{\gamma})} \boldsymbol{e}' \cdot \dot{\boldsymbol{e}}'' ds &= \int_{\mathrm{int}(\boldsymbol{\pi})} \{(\boldsymbol{e}' \cdot \boldsymbol{e}''_v)_u - (\boldsymbol{e}' \cdot \boldsymbol{e}''_u)_v\} dudv \\
&= \int_{\mathrm{int}(\boldsymbol{\pi})} \{(\boldsymbol{e}'_u \cdot \boldsymbol{e}''_v) - (\boldsymbol{e}'_v \cdot \boldsymbol{e}''_u)\} dudv \\
&= \int_{\mathrm{int}(\boldsymbol{\pi})} \frac{LN - M^2}{(EG - F^2)^{1/2}} dudv \quad \text{(by Lemma 13.1.3)} \\
&= \int_{\mathrm{int}(\boldsymbol{\pi})} \frac{LN - M^2}{EG - F^2} (EG - F^2)^{1/2} dudv \\
&= \int_{\mathrm{int}(\boldsymbol{\pi})} K d\mathcal{A}_{\boldsymbol{\sigma}} \quad \text{(see Section 6.4).}
\end{aligned} \tag{13.6}$$

If we combine Eqs. 13.4 and 13.6 and compare with the statement of the theorem, we see that what remains is to prove that

$$\int_0^{\ell(\boldsymbol{\gamma})} \dot{\theta}\, ds = 2\pi.$$

This is a version of Hopf's Umlaufsatz (cf. Theorem 3.1.4). We cannot give a fully satisfactory proof of it here but we offer the following heuristic argument.

The main observation is that, if $\tilde{\boldsymbol{\gamma}}$ is any other simple closed curve contained in the interior of $\boldsymbol{\gamma}$, there is a smooth family of simple closed curves $\boldsymbol{\gamma}^\tau$, defined for $0 \leq \tau \leq 1$, say, with $\boldsymbol{\gamma}^0 = \boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^1 = \tilde{\boldsymbol{\gamma}}$ (see Section 9.4 for the notion of a smooth family of curves). The existence of such a family is supposed to be 'intuitively obvious'.

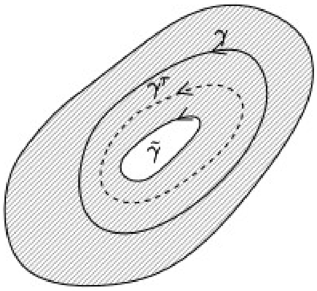

Note, however, that it is crucial that the interior of $\boldsymbol{\pi}$ is entirely contained in U, otherwise such a family will not exist, in general (see the diagram below).

Observe next that the integral $\int_0^{\ell(\gamma^\tau)} \dot{\theta}\, ds$ should depend continuously on τ. Further, since $\boldsymbol{\gamma}^\tau$ and $\boldsymbol{e}'$ return to their original values as one goes once round $\boldsymbol{\gamma}^\tau$, the integral is always an integer multiple of 2π. These two facts imply that the integral must be independent of τ – for by the intermediate value theorem a continuous variable cannot change from one integer to a different integer without passing through some non-integer value. To compute $\int_0^{\ell(\gamma)} \dot{\theta}\, ds$, we can therefore replace $\boldsymbol{\gamma}$ by any other simple closed curve $\tilde{\boldsymbol{\gamma}}$ in the interior of $\boldsymbol{\gamma}$, since this will not change the value of the integral. We take $\tilde{\boldsymbol{\gamma}}$ to be the image under $\boldsymbol{\sigma}$ of a small circle in the interior of $\boldsymbol{\pi}$. It is 'intuitively clear' that

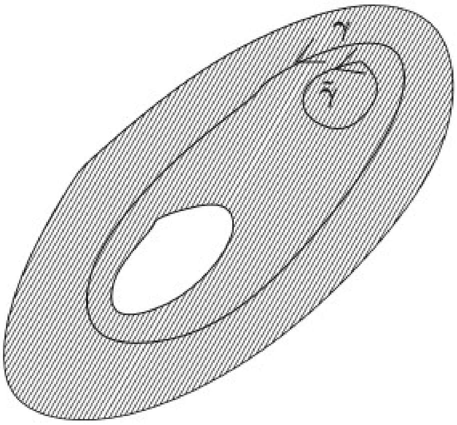

$$\int_0^{\ell(\tilde{\gamma})} \dot{\theta}\, ds = 2\pi,$$

because

(i) $\boldsymbol{e}'$ is essentially constant at all points of $\tilde{\boldsymbol{\gamma}}$ (because the circle is very small), and

(ii) the tangent vector to $\tilde{\boldsymbol{\gamma}}$ rotates by 2π on going once round $\tilde{\boldsymbol{\gamma}}$ because the interior of $\tilde{\boldsymbol{\gamma}}$ can be considered to be essentially part of a plane, and it is 'intuitively clear' that the tangent vector of a simple closed curve in the plane rotates by 2π on going once round the curve (indeed, this is the content of Theorem 3.1.4).

This completes the 'proof' of Hopf's Umlaufsatz. To complete the proof of Theorem 13.1.2, all that remains is to prove Lemma 13.1.3.

Proof 13.1.3 We can express the partial derivatives of $\boldsymbol{e}'$ and $\boldsymbol{e}''$ with respect to u and v in terms of the orthonormal basis $\{\boldsymbol{e}', \boldsymbol{e}'', \mathbf{N}\}$. Since both partial derivatives of $\boldsymbol{e}'$ are perpendicular to $\boldsymbol{e}'$, the $\boldsymbol{e}'$ components of $\boldsymbol{e}'_u$ and $\boldsymbol{e}'_v$ are zero (and similarly for $\boldsymbol{e}''$). Thus,

$$\begin{aligned}
\boldsymbol{e}'_u &= \qquad\quad \alpha\boldsymbol{e}'' + \lambda'\mathbf{N},\\
\boldsymbol{e}'_v &= \qquad\quad \beta\boldsymbol{e}'' + \mu'\mathbf{N},\\
\boldsymbol{e}''_u &= -\alpha'\boldsymbol{e}' \qquad\quad + \lambda''\mathbf{N},\\
\boldsymbol{e}''_v &= -\beta'\boldsymbol{e}' \qquad\quad + \mu''\mathbf{N},
\end{aligned}$$

for some scalars $\alpha, \beta, \alpha', \beta', \lambda', \mu', \lambda'', \mu''$ (which may depend on u and v). Moreover, by differentiating the equation $\boldsymbol{e}' \cdot \boldsymbol{e}'' = 0$ with respect to u, we see that $\boldsymbol{e}'_u \cdot \boldsymbol{e}'' = -\boldsymbol{e}' \cdot \boldsymbol{e}''_u$. i.e., $\alpha' = \alpha$ (and similarly $\beta' = \beta$). Thus,

$$\begin{aligned}
\boldsymbol{e}'_u &= \qquad\quad \alpha\boldsymbol{e}'' + \lambda'\mathbf{N},\\
\boldsymbol{e}'_v &= \qquad\quad \beta\boldsymbol{e}'' + \mu'\mathbf{N},\\
\boldsymbol{e}''_u &= -\alpha\boldsymbol{e}' \qquad\quad + \lambda''\mathbf{N},\\
\boldsymbol{e}''_v &= -\beta\boldsymbol{e}' \qquad\quad + \mu''\mathbf{N}.
\end{aligned} \tag{13.7}$$

It follows that

$$\boldsymbol{e}'_u \cdot \boldsymbol{e}''_v - \boldsymbol{e}''_u \cdot \boldsymbol{e}'_v = \lambda'\mu'' - \lambda''\mu'. \tag{13.8}$$

On the other hand, combining the formula

$$\mathbf{N}_u \times \mathbf{N}_v = K\boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v$$

(see Eq. 8.2) with the formulas

$$\mathbf{N} = \frac{\boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v}{\parallel \boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v \parallel}, \quad \parallel \boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v \parallel = (EG - F^2)^{1/2}$$

(see Proposition 6.4.2), we get

$$\mathbf{N}_u \times \mathbf{N}_v = \frac{LN - M^2}{(EG - F^2)^{1/2}}\mathbf{N},$$

and hence

$$(\mathbf{N}_u \times \mathbf{N}_v) \cdot \mathbf{N} = \frac{LN - M^2}{(EG - F^2)^{1/2}}. \tag{13.9}$$

Since $\mathbf{N} = \boldsymbol{e}' \times \boldsymbol{e}''$,

$$\begin{aligned}(\mathbf{N}_u \times \mathbf{N}_v) \cdot \mathbf{N} &= (\mathbf{N}_u \times \mathbf{N}_v) \cdot (\boldsymbol{e}' \times \boldsymbol{e}''),\\ &= (\mathbf{N}_u \cdot \boldsymbol{e}')(\mathbf{N}_v \cdot \boldsymbol{e}'') - (\mathbf{N}_u \cdot \boldsymbol{e}'')(\mathbf{N}_v \cdot \boldsymbol{e}')\\ &= (\mathbf{N} \cdot \boldsymbol{e}'_u)(\mathbf{N} \cdot \boldsymbol{e}''_v) - (\mathbf{N} \cdot \boldsymbol{e}''_u)(\mathbf{N} \cdot \boldsymbol{e}'_v) \qquad (13.10)\\ &= \lambda'\mu'' - \lambda''\mu' \quad \text{by Eq. 13.7,}\end{aligned}$$

where in passing from the second line to the third we used the equations

$$\begin{aligned}\mathbf{N}_u \cdot \boldsymbol{e}' &= -\mathbf{N} \cdot \boldsymbol{e}'_u, \quad \mathbf{N}_u \cdot \boldsymbol{e}'' = -\mathbf{N} \cdot \boldsymbol{e}''_u,\\ \mathbf{N}_v \cdot \boldsymbol{e}' &= -\mathbf{N} \cdot \boldsymbol{e}'_v, \quad \mathbf{N}_v \cdot \boldsymbol{e}'' = -\mathbf{N} \cdot \boldsymbol{e}''_v,\end{aligned}$$

which follow by differentiating $\mathbf{N} \cdot \boldsymbol{e}' = 0 = \mathbf{N} \cdot \boldsymbol{e}''$ with respect to u and v. Putting Eqs. 13.9 and 13.10 together shows that the right-hand sides of Eqs. 13.5 and 13.8 are equal. Since Eq. 13.8 has already been established, this proves Eq.13.5. □

EXERCISES

13.1.1 Suppose that a surface patch $\boldsymbol{\sigma}$ has Gaussian curvature ≤ 0 everywhere. Prove that there are no simple closed geodesics on $\boldsymbol{\sigma}$. How do you reconcile this with the fact that the parallels of a circular cylinder are geodesics?

13.1.2 Let $\boldsymbol{\gamma}$ be a unit-speed curve in $\mathbb{R}^3$ with nowhere vanishing curvature. Let $\mathbf{n}$ be the principal normal of $\boldsymbol{\gamma}$, viewed as a curve on S^2, and let s be the arc-length of $\mathbf{n}$. Show that the geodesic curvature of $\mathbf{n}$ is, up to a sign,

$$\frac{d}{ds}\left(\tan^{-1}\frac{\tau}{\kappa}\right),$$

where κ and τ are the curvature and torsion of $\boldsymbol{\gamma}$. Show also that, if $\mathbf{n}$ is a *simple closed* curve on S^2, the interior and exterior of $\mathbf{n}$ are regions of equal area (*Jacobi's Theorem*).

5- 6 2 //3

13.2 Gauss–Bonnet for curvilinear polygons

For the next version of Gauss–Bonnet, we shall have to generalize our notion of a curve by allowing the possibility of ' corners'. More precisely, we make the following definition.

Definition 13.2.1

A *curvilinear polygon* in $\mathbb{R}^2$ is a continuous map $\boldsymbol{\pi} : \mathbb{R} \to \mathbb{R}^2$ such that, for some real number T and some points $0 = t_0 < t_1 < \cdots < t_n = T$:

(i) $\boldsymbol{\pi}(t) = \boldsymbol{\pi}(t')$ if and only if $t' - t$ is an integer multiple of T.

(ii) $\boldsymbol{\pi}$ is smooth on each of the open intervals $(t_0, t_1), (t_1, t_2), \ldots, (t_{n-1}, t_n)$.

(iii) The one-sided derivatives

$$\dot{\boldsymbol{\pi}}^-(t_i) = \lim_{t \uparrow t_i} \frac{\boldsymbol{\pi}(t) - \boldsymbol{\pi}(t_i)}{t - t_i}, \quad \dot{\boldsymbol{\pi}}^+(t_i) = \lim_{t \downarrow t_i} \frac{\boldsymbol{\pi}(t) - \boldsymbol{\pi}(t_i)}{t - t_i} \tag{13.11}$$

exist for $i = 1, \ldots, n$ and are non-zero and not parallel.

The points $\boldsymbol{\gamma}(t_i)$ for $i = 1, \ldots, n$ are called the *vertices* of the curvilinear polygon $\boldsymbol{\pi}$, and the segments of it corresponding to the open intervals (t_{i-1}, t_i) are called its *edges*.

It makes sense to say that a curvilinear polygon $\boldsymbol{\pi}$ is positively-oriented: for all t such that $\boldsymbol{\pi}(t)$ is not a vertex, the vector $\mathbf{n}_s$ obtained by rotating $\dot{\boldsymbol{\pi}}$ anticlockwise by $\pi/2$ should point into $\operatorname{int}(\boldsymbol{\pi})$. (The region $\operatorname{int}(\boldsymbol{\pi})$ enclosed by $\boldsymbol{\pi}$ makes sense because the Jordan Curve Theorem applies to curvilinear polygons in the plane.)

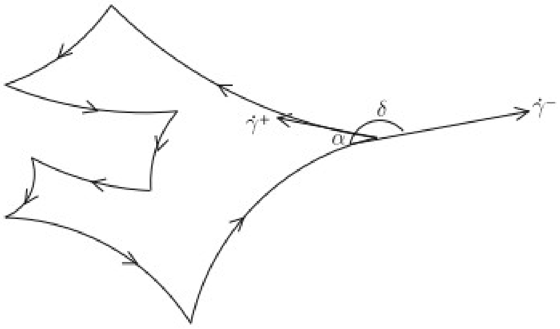

Now let $\boldsymbol{\sigma} : U \to \mathbb{R}^3$ be a surface patch and let $\boldsymbol{\pi} : \mathbb{R} \to U$ be a curvilinear polygon in U, as in Definition 13.2.1. Then, $\boldsymbol{\gamma} = \boldsymbol{\sigma} \circ \boldsymbol{\pi}$ is called a curvilinear polygon on the surface patch $\boldsymbol{\sigma}$, $\operatorname{int}(\boldsymbol{\gamma})$ is the image under $\boldsymbol{\sigma}$ of $\operatorname{int}(\boldsymbol{\pi})$, the vertices of $\boldsymbol{\gamma}$ are the points $\boldsymbol{\gamma}(t_i)$ for $i = 1, \ldots, n$, and the edges of $\boldsymbol{\gamma}$ are the segments of it corresponding to the open intervals (t_{i-1}, t_i). Since $\boldsymbol{\sigma}$ is allowable, the one-sided derivatives

$$\dot{\boldsymbol{\gamma}}^-(t_i) = \lim_{t\uparrow t_i} \frac{\boldsymbol{\gamma}(t)-\boldsymbol{\gamma}(t_i)}{t-t_i}, \quad \dot{\boldsymbol{\gamma}}^+(t_i) = \lim_{t\downarrow t_i} \frac{\boldsymbol{\gamma}(t)-\boldsymbol{\gamma}(t_i)}{t-t_i}$$

exist and are not parallel.

Let $\theta_i^{\pm}$ be the angles between $\dot{\boldsymbol{\gamma}}^{\pm}(t_i)$ and $\boldsymbol{e}'$, defined as in Eq. 13.1, let $\delta_i = \theta_i^+ - \theta_i^-$ be the external angle at the vertex $\boldsymbol{\gamma}(t_i)$, and let $\alpha_i = \pi - \delta_i$ be the internal angle. Since the tangent vectors $\dot{\boldsymbol{\gamma}}^+(t_i)$ and $\dot{\boldsymbol{\gamma}}^-(t_i)$ are not parallel, the angle δ_i is not a multiple of π. Note that all of these angles are well defined only up to multiples of 2π. We assume from now on that $0 < \alpha_i < 2\pi$ for $i = 1, \ldots, n$.

A curvilinear polygon $\boldsymbol{\gamma}$ is said to be unit-speed if $\parallel \dot{\boldsymbol{\gamma}} \parallel = 1$ whenever $\dot{\boldsymbol{\gamma}}$ is defined, i.e., for all t such that $\boldsymbol{\gamma}(t)$ is not a vertex of $\boldsymbol{\gamma}$. We denote the parameter of $\boldsymbol{\gamma}$ by s if $\boldsymbol{\gamma}$ is unit-speed. The period of $\boldsymbol{\gamma}$ is then equal to its length $\ell(\boldsymbol{\gamma})$, which is the sum of the lengths of the edges of $\boldsymbol{\gamma}$.

Theorem 13.2.2

Let $\boldsymbol{\gamma}$ be a positively-oriented unit-speed curvilinear polygon with n edges on a surface $\boldsymbol{\sigma}$, and let $\alpha_1, \alpha_2, \ldots, \alpha_n$ be the interior angles at its vertices. Then,

$$\int_0^{\ell(\boldsymbol{\gamma})} \kappa_g ds = \sum_{i=1}^n \alpha_i - (n-2)\pi - \int_{\mathrm{int}(\boldsymbol{\gamma})} K d\mathcal{A}_{\boldsymbol{\sigma}}.$$

Proof

Exactly the same argument as in the proof of Theorem 13.1.2 shows that

$$\int_0^{\ell(\boldsymbol{\gamma})} \kappa_g ds = \int_0^{\ell(\boldsymbol{\gamma})} \dot{\theta} ds - \int_{\mathrm{int}(\boldsymbol{\gamma})} K d\mathcal{A}_{\boldsymbol{\sigma}}.$$

We shall prove that

$$\int_0^{\ell(\boldsymbol{\gamma})} \dot{\theta} ds = 2\pi - \sum_{i=1}^n \delta_i. \tag{13.12}$$

Assuming this, we get

$$\begin{aligned}\int_0^{\ell(\boldsymbol{\gamma})} \kappa_g ds &= 2\pi - \sum_{i=1}^n \delta_i - \int_{\mathrm{int}(\boldsymbol{\gamma})} K d\mathcal{A}_{\boldsymbol{\sigma}} \\ &= 2\pi - \sum_{i=1}^n (\pi - \alpha_i) - \int_{\mathrm{int}(\boldsymbol{\gamma})} K d\mathcal{A}_{\boldsymbol{\sigma}} \\ &= \sum_{i=1}^n \alpha_i - (n-2)\pi - \int_{\mathrm{int}(\boldsymbol{\gamma})} K d\mathcal{A}_{\boldsymbol{\sigma}}.\end{aligned}$$

To establish Eq. 13.12, we imagine 'smoothing' each vertex of $\boldsymbol{\gamma}$ as shown in the following diagram.

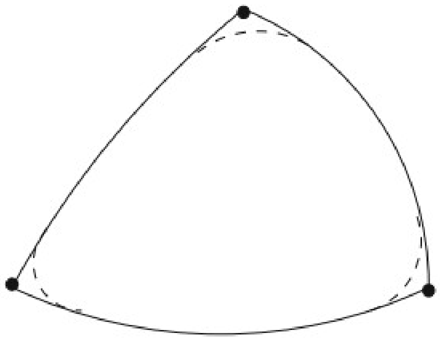

If the 'smoothed' curve $\tilde{\boldsymbol{\gamma}}$ is smooth (!), then, in an obvious notation,

$$\int_0^{\ell(\tilde{\boldsymbol{\gamma}})} \tilde{\dot{\theta}} ds = 2\pi. \tag{13.13}$$

Since $\boldsymbol{\gamma}$ and $\tilde{\boldsymbol{\gamma}}$ are the same except near the vertices of $\boldsymbol{\gamma}$, the difference

$$\int_0^{\ell(\tilde{\boldsymbol{\gamma}})} \tilde{\dot{\theta}} ds - \int_0^{\ell(\boldsymbol{\gamma})} \dot{\theta} ds \tag{13.14}$$

is a sum of n contributions, one from near each vertex. Near $\boldsymbol{\gamma}(s_i)$, the picture is

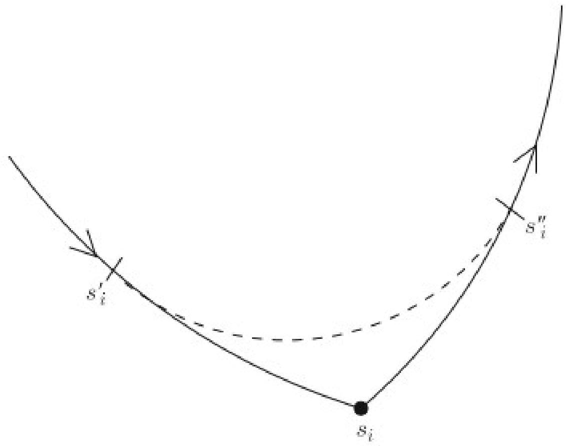

i.e., $\boldsymbol{\gamma}$ and $\tilde{\boldsymbol{\gamma}}$ agree except when s belongs to a small interval (s_i', s_i''), say, containing s_i, so the contribution from the ith vertex is

$$\int_{s_i'}^{s_i''} \tilde{\dot{\theta}} ds - \int_{s_i'}^{s_i} \dot{\theta} ds - \int_{s_i}^{s_i''} \dot{\theta} ds.$$

The first integral is the angle between $\dot{\tilde{\boldsymbol{\gamma}}}(s_i'')$ and $\dot{\tilde{\boldsymbol{\gamma}}}(s_i')$, which as s_i' and s_i'' tend to s_i becomes the angle between $\dot{\boldsymbol{\gamma}}^+(s_i)$ and $\dot{\boldsymbol{\gamma}}^-(s_i)$, i.e., δ_i. On the other hand, since $\boldsymbol{\gamma}(s)$ is smooth on each of the intervals (s_i', s_i) and (s_i, s_i''), the last two integrals go to zero as s_i' and s_i'' tend to s_i. Thus, the contribution to the expression (13.14) from the ith vertex tends to δ_i as s_i' and s_i'' tend to s_i . Summing over all the vertices, we get

$$\int_0^{\ell(\tilde{\boldsymbol{\gamma}})} \dot{\tilde{\theta}} ds - \int_0^{\ell(\boldsymbol{\gamma})} \dot{\theta} ds = \sum_{i=1}^{n} \delta_i.$$

Equation 13.12 now follows from this and Eq. 13.13. □

Corollary 13.2.3

If $\boldsymbol{\gamma}$ is a curvilinear polygon with n edges each of which is an arc of a geodesic, then the internal angles $\alpha_1, \alpha_2, \ldots, \alpha_n$ of the polygon satisfy the equation

$$\sum_{i=1}^{n} \alpha_i = (n-2)\pi + \int_{\operatorname{int}(\boldsymbol{\gamma})} K d\mathcal{A}_{\boldsymbol{\sigma}}.$$

Proof

This is immediate from Theorem 13.2.2, since $\kappa_g = 0$ along a geodesic. □

As a special case of Corollary 13.2.3, consider an n-gon in the plane with straight edges. Since $K = 0$ for the plane, Corollary 13.2.3 gives

$$\sum_{i=1}^{n} \alpha_i = (n-2)\pi,$$

a well-known result of elementary geometry.

For a curvilinear n-gon on S^2 whose sides are arcs of great circles, we have $K = 1$ so $\sum \alpha_i$ exceeds the plane value $(n-2)\pi$ by the area $\int d\mathcal{A}_{\boldsymbol{\sigma}}$ of the polygon. Taking $n = 3$, we get for the area of a spherical triangle with angles α, β, γ whose edges are arcs of great circles,

$$\mathcal{A} = \alpha + \beta + \gamma - \pi.$$

This is just Theorem 6.4.7, which is therefore a special case of Gauss–Bonnet.

Finally, for a geodesic n-gon on the pseudosphere (see Section 8.3), for which $K = -1$, we see that $\sum \alpha_i$ is *less* than $(n-2)\pi$ by the area of the polygon:

$$\mathcal{A} = (n-2)\pi - \alpha_1 - \alpha_2 - \cdots - \alpha_n,$$

which is Theorem 11.1.5.

5- 6 726/ 3

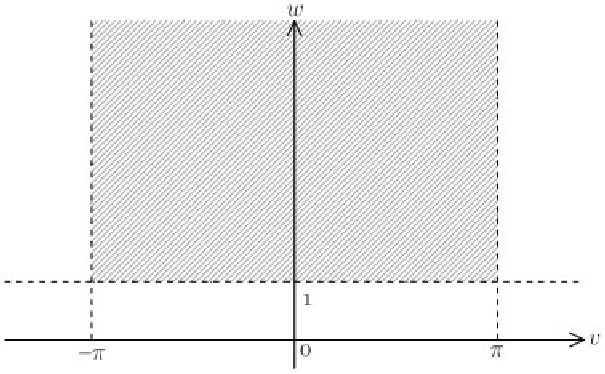

8.3.1 (i) Setting $\tilde{u} = v, \tilde{v} = w = e^{-u}$, we have $u = -\ln\tilde{v}, v = \tilde{u}$ so, in the notation of Exercise 6.1.4, $J = \begin{pmatrix} 0 & -\frac{1}{\tilde{v}} \\ 1 & 0 \end{pmatrix}$. Since J is invertible, $(u, v) \mapsto (v, w)$ is a reparametrization map. The first fundamental form in terms of v, w is given by $\begin{pmatrix} \tilde{E} & \tilde{F} \\ \tilde{F} & \tilde{G} \end{pmatrix} = J^t \begin{pmatrix} E & F \\ F & G \end{pmatrix} J = \begin{pmatrix} 0 & 1 \\ -\frac{1}{\tilde{v}} & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & f(u)^2 \end{pmatrix} \begin{pmatrix} 0 & -\frac{1}{\tilde{v}} \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{w^2} & 0 \\ 0 & \frac{1}{w^2} \end{pmatrix}$, so the first fundamental form is $(dv^2 + dw^2)/w^2$.

(ii) We find that the matrix $\tilde{J} = \begin{pmatrix} \frac{\partial v}{\partial V} & \frac{\partial v}{\partial W} \\ \frac{\partial w}{\partial V} & \frac{\partial w}{\partial W} \end{pmatrix} = \begin{pmatrix} v(w+1) & \frac{1}{2}(v^2 - (w+1)^2) \\ -\frac{1}{2}(v^2 - (w+1)^2) & v(w+1) \end{pmatrix}$, so the first fundamental form matrix in terms of V and W is $\tilde{J}^t \begin{pmatrix} \frac{1}{w^2} & 0 \\ 0 & \frac{1}{w^2} \end{pmatrix} \tilde{J} = \frac{(v^2+(w+1)^2)^2}{4w^2} I = \frac{4}{(1-V^2-W^2)^2} I$, after some tedious algebra.

In (i), $u < 0$ and $-\pi < v < \pi$ corresponds to $-\pi < v < \pi$ and $w > 1$, a semi-infinite rectangle in the upper half of the vw-plane.

To find the corresponding region in (ii), it is convenient to introduce the complex numbers $z = v + iw$, $Z = V + iW$. Then, the equations in (ii) are equivalent to $Z = \frac{z-i}{z+i}$, $z = \frac{Z+1}{i(Z-1)}$. The line $v = \pi$ in the vw-plane corresponds to $z + \bar{z} = 2\pi$ (the bar denoting complex conjugate), i.e., $\frac{Z+1}{i(Z-1)} - \frac{\bar{Z}+1}{i(\bar{Z}-1)} = 2\pi$, which simplifies to $|Z - (1 - \frac{i}{\pi})|^2 = \frac{1}{\pi^2}$; so $v = \pi$ corresponds to the circle in the VW-plane with centre $1 - \frac{i}{\pi}$ and radius $\frac{1}{\pi}$. Similarly, $v = -\pi$ corresponds to the circle with centre $1 + \frac{i}{\pi}$ and radius $\frac{1}{\pi}$. Finally, $w = 1$ corresponds to $z - \bar{z} = 2i$, i.e., $\frac{Z+1}{i(Z-1)} + \frac{\bar{Z}+1}{i(\bar{Z}-1)} = 2i$. This simplifies to $|Z - \frac{1}{2}|^2 = \frac{1}{4}$; so $w = 1$ corresponds to the circle with centre $1/2$ and radius $1/2$ in the VW-plane. The required region in the VW-plane is that bounded by these three circles:

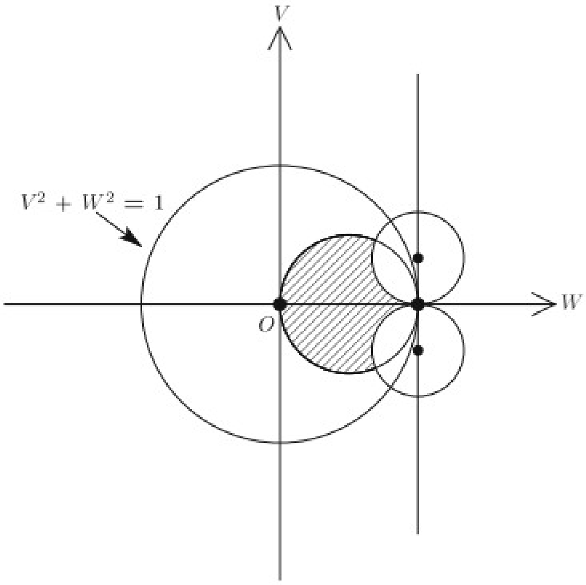

For (iii) we follow the hint and make use of polar coordinates on the disc, $V = r\cos\theta$, $W = r\sin\theta$, $\bar{V} = \bar{r}\cos\bar{\theta}$, $\bar{W} = \bar{r}\sin\bar{\theta}$. We find that $\bar{r} = \frac{2r}{r^2+1}$, $\bar{\theta} = \theta$. Suppose that the first fundamental form in terms of these parameters is $Ed\bar{r}^2 + 2Fd\bar{r}d\bar{\theta} + Gd\bar{\theta}^2$. Since $\frac{d\bar{r}}{dr} = \frac{2(1-r^2)}{(1+r^2)^2}$, the first fundamental form is $\frac{4(1-r^2)^2}{(1+r^2)^4}Edr^2 + \frac{4(1-r^2)}{(1+r^2)^2}Fdrd\theta + Gd\theta^2$. Equating this to $\frac{4(dV^2+dW^2)}{(1-V^2-W^2)^2} = \frac{4(dr^2+r^2d\theta^2)}{(1-r^2)^2}$, we get $E = \frac{(1+r^2)^4}{(1-r^2)^4} = \frac{1}{(1-\bar{r}^2)^2}$, $F = 0$, $G = \frac{4r^2}{(1-r^2)^2} = \frac{\bar{r}^2}{1-\bar{r}^2}$. Converting back to the parameters $(\bar{V}, \bar{W})$, we have $\bar{r}d\bar{r} = \bar{V}d\bar{V} + \bar{W}d\bar{W}, \bar{r}^2 d\bar{\theta} = \bar{V}d\bar{W} - \bar{W}d\bar{V}$, so the first fundamental form becomes $\frac{(\bar{V}d\bar{V}+\bar{W}d\bar{W})^2+(1-\bar{r}^2)(\bar{V}d\bar{W}-\bar{W}d\bar{V})^2}{\bar{r}^2(1-\bar{r}^2)^2} =$ $\frac{(\bar{V}^2+(1-\bar{r}^2)\bar{W}^2)d\bar{V}^2+2\bar{r}^2\bar{V}\bar{W}d\bar{V}d\bar{W}+(\bar{W}^2+(1-\bar{r}^2)\bar{V}^2)d\bar{W}^2}{\bar{r}^2(1-\bar{r}^2)^2}$
$= \frac{(1-\bar{W}^2)d\bar{V}^2+2\bar{V}\bar{W}d\bar{V}d\bar{W}+(1-\bar{V}^2)d\bar{W}^2}{(1-\bar{V}^2-\bar{W}^2)^2}$.

8.4.1 Let $\tilde{\boldsymbol{\sigma}}(\tilde{u}, \tilde{v})$ be a patch of $\mathcal{S}$ containing $\mathbf{p} = \tilde{\boldsymbol{\sigma}}(\tilde{u}_0, \tilde{v}_0)$. The Gaussian curvature K of $\mathcal{S}$ is < 0 at $\mathbf{p}$; since K is a smooth function of $(\tilde{u}, \tilde{v})$ (Exercise 8.1.3), $K(\tilde{u}, \tilde{v}) < 0$ for $(\tilde{u}, \tilde{v})$ in some open set $\tilde{U}$ containing $(\tilde{u}_0, \tilde{v}_0)$; then every point of $\tilde{\boldsymbol{\sigma}}(\tilde{U})$ is hyperbolic. Let κ_1, κ_2 be the principal curvatures of $\tilde{\boldsymbol{\sigma}}$, let $0 < \theta < \pi/2$ be such that $\tan\theta = \sqrt{-\kappa_1/\kappa_2}$, and let $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ be the unit tangent vectors of $\tilde{\boldsymbol{\sigma}}$ making angles θ and $-\theta$, respectively, with the principal vector corresponding to κ_1 (see Theorem 8.2.4). Applying Proposition 8.4.3 gives the result. For the last part, put $\dot{v} = 0$ in the formula for κ_n in Proposition 7.3.5: this shows that $L = 0$ if the parameter curves $v =$ constant are asymptotic. Similarly $N = 0$ if the parameter curves $u =$ constant are asymptotic.

!"#" $ % % & ' ( )) * +) , -)
.) @)) . @ *i 2 !.! % 3 4

,9 ,
*) TT
( ) 7
1 9 6
% T/

5- 6N2N 3 / f)) E -

5- 672N63 D) E * - @ A I

and write down those, if any, that satisfy the equation. However, it may be hard in practice, because there are 2^m sequences of values for m variables (two for the first variable; for each of these, two for the second variable; and so on). So the number of possibilities quickly becomes astronomical as m increases. And no known method for solving m-variable equations, mod 2, is substantially faster than trying all possibilities.

In fact, no fast method is known to decide even *existence* of solutions for m-variable equations, mod 2. This is a fundamental unsolved problem in computation and logic, as we will explain further in chapters 3 and 9.

2.5 Quadratic Integers

Let us begin with a curious fact about the numbers 25 and 27. 25 is a square and 27 is a cube, so these numbers give a solution of the equation $y^3 = x^2 + 2$; namely, $x = 5$ and $y = 3$. Nearly 2000 years ago, Diophantus specifically mentioned this equation and this solution in his *Arithmetica*, Book VI, Problem 17. After reading this passage in Diophantus, Fermat (1657) claimed that it is the *only* solution in the positive integers. Why the equation caught their attention we do not really know. But it was a turning point in the development of number theory when Euler (1770), p. 401, proved Fermat's claim by a new and audacious method.

Observing that $x^2 + 2 = (x + \sqrt{-2})(x - \sqrt{-2})$, Euler's attention was drawn to numbers of the form $a + b\sqrt{-2}$, where a and b are integers. Leaving aside the meaning of the "imaginary" $\sqrt{-2}$, these new numbers are also "integers" in some sense. This is because the sum, difference, and product of any two of them is another number of the same kind and they satisfy the algebraic rules of a ring (mentioned in section 1.3). The audacious part of Euler's solution was to assume that they also have more doubtful properties, such as unique prime factorization. To see why this assumption comes up, we follow Euler's train of thought.

Assume that x and y are ordinary integers such that

$$y^3 = x^2 + 2 = (x + \sqrt{-2})(x - \sqrt{-2}).$$

Then one might think, if the factors on the right behave like ordinary integers, that $\gcd(x+\sqrt{-2}, x-\sqrt{-2})=1$ and (imagining $x+\sqrt{-2}$ and $x-\sqrt{-2}$ factorized into primes, with no prime common to both) that each of them is a cube, since their product is the cube y^3. This is not quite right, but it is true when x is a solution of $y^3=x^2+2$, as we will see in the next subsection.

Then, writing $x+\sqrt{-2}$ as a cube, we get

$$\begin{aligned} x+\sqrt{-2} &= (a+b\sqrt{-2})^3 \\ &= a^3+3a^2\cdot b\sqrt{-2}+3a\cdot(b\sqrt{-2})^2+(b\sqrt{-2})^3 \\ &= a^3-6ab^2+(3a^2b-2b^3)\sqrt{-2}. \end{aligned}$$

It follows, by equating "real and imaginary parts," that

$$x=a^3-6ab^2 \quad \text{and} \quad 1=3a^2b-2b^3=b(3a^2-2b^2).$$

Now if $1=b(3a^2-2b^2)$ then b divides 1 and therefore $b=\pm1$. If $b=-1$ then $1=-(3a^2-2)$, which is impossible because $3a^2-2$ is either -2 (for $a=0$) or positive. We must therefore have $b=1$, in which case $1=3a^2-2$. So $a^2=1$ and hence $a=\pm1$.

Substituting $a=1$, $b=1$ in $x=a^3-6ab^2$ gives the negative value $x=-5$, so the only remaining option is $a=-1$, $b=1$. This gives the value $x=5$ claimed by Fermat, and with it the value $y=3$.

What Is Going on in Euler's Proof?

Euler came up with his proof before the "imaginary" number $\sqrt{-2}$ was well understood, not to mention the concept of the "integers" $a+b\sqrt{-2}$ and their greatest common divisors and primes. Today we know that it is valid to picture $a+b\sqrt{-2}$ as a point in the plane with real coordinate a and imaginary coordinate $b\sqrt{2}$, so that the distance of this point from the origin is $\sqrt{a^2+2b^2}$, by the Pythagorean theorem (figure 2.3). We also call this distance the *absolute value* of $a+b\sqrt{-2}$, and denote it by $|a+b\sqrt{-2}|$.

The square $|a+b\sqrt{-2}|^2$ of this distance is the ordinary integer a^2+2b^2, which is called the *norm* of $a+b\sqrt{-2}$. With its help we can

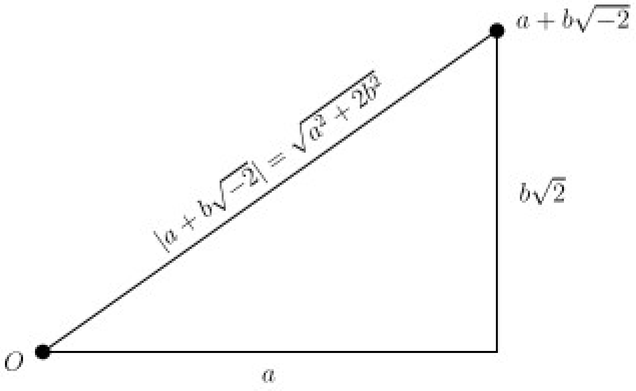

Figure 2.3: Distance to the point $a + b\sqrt{-2}$.

reduce questions about divisibility of these *quadratic integers* to questions about divisibility in the ordinary integers. The magical property of the norm that makes this possible is

$$|uv|^2 = |u|^2|v|^2, \tag{*}$$

or "norm of a product equals the product of the norms." We say more about this *multiplicative property* in the next section, but we first explain how it enables us to prove that $\gcd(x + \sqrt{-2}, x - \sqrt{-2}) = 1$ when x and y are integers such that $y^3 = x^2 + 2$.

The magical property (*) means that if v divides w (which means in turn that $w = uv$ for some u) then $|v|^2$ divides $|w|^2$. Also, the only numbers $a + b\sqrt{-2}$ with norm 1 are ±1, because $a^2 + 2b^2 = 1$ only if $a = \pm 1$, $b = 0$. Thus it suffices to prove that *any* common divisor of $x + \sqrt{-2}$, $x - \sqrt{-2}$ has norm 1 when $y^3 = x^2 + 2$.

First notice that

$$y^3 \equiv 0,\ 1,\ \text{or } 3 \pmod 4.$$

This can be seen by cubing the four possible values of y mod 4, namely 0, 1, 2, 3. On the other hand, for the even values of x mod 4 (0 and 2) we have

$$x^2 + 2 \equiv 2 \pmod 4.$$

So $y^3 = x^2 + 2$ only for x odd, and then the norm $x^2 + 2$ of $x \pm \sqrt{-2}$ is odd.

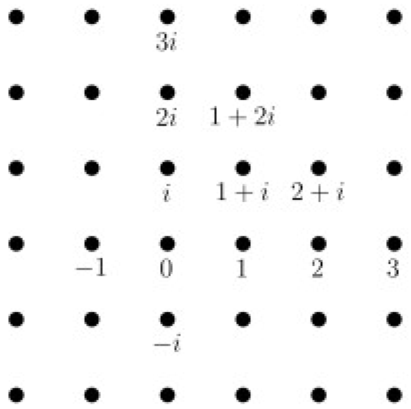

Figure 2.4: Gaussian integers near 0.

Now a divisor of $x + \sqrt{-2}$ and $x - \sqrt{-2}$ also divides their difference $2\sqrt{-2}$, which has norm 8. The gcd of the norms 8 and the odd number $x^2 + 2$ is 1, so any divisor of $x + \sqrt{-2}$, $x - \sqrt{-2}$ is a number whose norm divides 1; that is, 1 itself.

2.6 The Gaussian Integers

In the previous section we saw how a problem about ordinary integers draws our attention to weird "integers" involving the quantity $\sqrt{-2}$. Moreover, if we play along with these weird integers—granting them attributes such as gcd and primes—they enable us to solve the original problem very simply. It seems as though the factorization $x^2 + 2 = (x + \sqrt{-2})(x - \sqrt{-2})$ explains the behavior of $x^2 + 2$ better than we can in the language of ordinary integers.

However, we have yet to make sense of the concept of "prime" and "unique prime factorization" in the world of the quadratic integers $a + b\sqrt{-2}$. To pave the way for this, we look first at the simplest quadratic integers: those of the form $a + b\sqrt{-1}$, or $a + bi$, where a and b are ordinary integers. These are called the *Gaussian integers* because they were first studied by Gauss (1832). The Gaussian integers form a square grid in the plane of complex numbers, part of which is shown in figure 2.4.

Like the integers $a+b\sqrt{-2}$, the Gaussian integers form a ring because the sum, difference, and product of Gaussian integers are again Gaussian integers, and the various algebraic laws are easily checked. The distance from the origin to $a+bi$ (also known as the *absolute value*, $|a+bi|$) is $\sqrt{a^2+b^2}$ by the Pythagorean theorem, and the square of this distance, a^2+b^2, is an ordinary integer called the *norm* of $a+bi$. Again it is true that "the norm of a product equals the product of the norms," but this time we will check.

Consider Gaussian integers $a+bi$ and $c+di$, whose norms $|a+bi|^2$ and $|c+di|^2$ are a^2+b^2 and c^2+d^2, respectively. Their product is

$$\begin{aligned}(a+bi)(c+di) &= ac+adi+bci+bdi^2\\ &= (ac-bd)+(ad+bc)i \quad \text{because} \quad i^2=-1.\end{aligned}$$

It follows that the norm of the product is $(ac-bd)^2+(ad+bc)^2$, and miraculously

$$(ac-bd)^2+(ad+bc)^2=(a^2+b^2)(c^2+d^2)$$

—the product of the norms. This identity can be checked by multiplying out both sides, and finding that they both equal $a^2c^2+a^2d^2+b^2c^2+b^2d^2$.

Notice that this calculation does not assume that a, b, c, d are integers, so in fact it is true for any complex numbers u, v that

$$|uv|^2=|u|^2|v|^2.$$

This is why "norm of a product equals product of the norms" is also true for the integers $a+b\sqrt{-2}$.[1] As for the latter integers, it follows for the Gaussian integers that if v divides w then $|v|^2$ divides $|w|^2$, so divisibility of Gaussian integers depends on divisibility of ordinary integers.

It now makes sense for us to define a *Gaussian prime* to be a Gaussian integer, of norm greater than 1, which is not the product of

---

[1] It also follows that $|uv|=|u||v|$, a fact which has some geometric implications for the complex numbers, as we will see in section 10.3.

Gaussian integers of smaller norm. Since norms are ordinary integers, the process of factorizing a Gaussian integer into Gaussian integers of smaller norm must eventually halt, necessarily with Gaussian primes. So *every Gaussian integer has a Gaussian prime factorization.*

Examples of Gaussian Prime Factorizations

The smallest example is $2 = (1+i)(1-i)$. This is a factorization into numbers of smaller norm, because $|1+i|^2 = |1-i|^2 = 2$ and $|2|^2 = 4$. Also, neither $1+i$ nor $1-i$ can be split into factors of smaller norm, because their norm 2 is not the product of smaller integers.

It is the same story for any ordinary prime that is the sum of two squares, such as $37 = 6^2 + 1^2$. The ordinary prime splits into Gaussian factors of smaller norm, in this case

$$6^2 + 1^2 = (6+i)(6-i), \quad \text{where} \quad |6+i|^2 = |6-i|^2 = 37,$$
$$\text{whereas} \quad |37|^2 = 37^2.$$

But neither factor splits into factors of smaller norm, because 37 is not a product of smaller ordinary integers. Thus any ordinary prime that is a sum of two squares is a product of two Gaussian primes.

If we are given a random Gaussian integer, such as $3+i$, we can find its Gaussian prime factors from the ordinary prime factors of its norm. In this case

$$|3+i|^2 = 3^2 + 1^2 = 10 = 2 \cdot 5.$$

Thus we are looking for Gaussian factors of norms 2 and 5, and they will necessarily be Gaussian primes because their norms are ordinary primes. We already know some Gaussian primes of norm 2; namely, $1+i$ and $1-i$. Apart from changing the sign of 1, these are the only ones. And the only Gaussian primes of norm 5 are $2+i$ and similar numbers obtained by changing signs (since the only squares with sum 5 are 4 and 1). Among this small number of possibilities, we soon find that

$$3+i = (1-i)(2+i).$$

Division with Remainder

We now consider a rather different problem: *division with remainder*. For ordinary integers we have the visually evident *division property* that $a = qb + r$ with $|r| < |b|$, as we saw in section 2.1. Given a Gaussian integer, say $5 + 3i$, and a Gaussian integer of smaller norm, say $3 + i$, we would like to be able to divide $5 + 3i$ by $3 + i$ and obtain a remainder r of smaller norm than $3 + i$. This means finding Gaussian integers q and r ("quotient" and "remainder") with

$$5 + 3i = (3 + i)q + r, \quad \text{such that} \quad |r| < |3 + i|.$$

This can be done by looking at the multiples $3 + i$ in the neighborhood of $5 + 3i$, and picking the one, $(3 + i)q$, nearest to $5 + 3i$. Then the difference

$$r = 5 + 3i - (3 + i)q$$

is as small as possible, and hopefully it is smaller than $3 + i$.

Indeed, this always works, and for a rather striking reason: *the multiples of $3 + i$ form a grid of squares of side $|3 + i|$, and the distance from any point in a square to the nearest corner is less than the length of the side.*

Why squares? Well, any Gaussian integer multiple of $3 + i$, by $a + bi$ say, is the sum of a copies of $3 + i$ and b copies of $i(3 + i) = -1 + 3i$. The points $3 + i$ and $-1 + 3i$ lie at distance $|3 + 1| = \sqrt{10} = |-1 + 3i|$ from 0, and *in perpendicular directions*. Thus the points 0, $3 + i$, and $i(3 + i) = -1 + 3i$ form two sides of a square. Adding further copies of $3 + i$ and $-1 + 3i$ creates further squares, all of side length $|3 + i|$, as shown in figure 2.5.

We see from figure 2.5 that the multiple of $3 + i$ nearest to $5 + 3i$ is $2(3 + i)$, and that their difference is

$$r = 2(3 + i) - (5 + 3i) = 1 - i, \quad \text{with absolute value } \sqrt{2} < \sqrt{10}.$$

More generally, it is easy to prove (say, using the Pythagorean theorem) that the distance from any point in a square to the nearest corner is less than the side of the square. Also, the multiples uq of any Gaussian integer u form a grid of squares of side length $|u|$ by

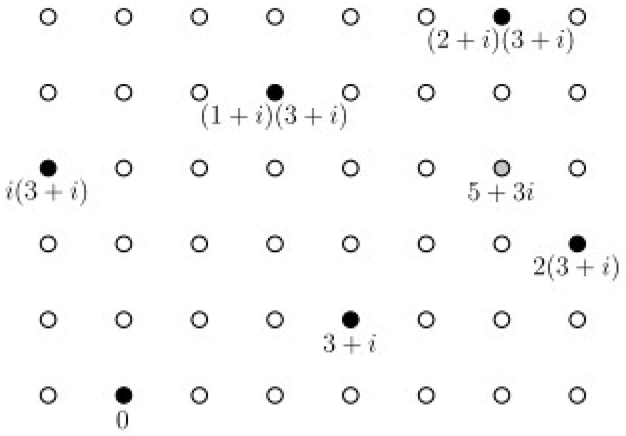

Figure 2.5: Multiples of $3+i$ near $5+3i$.

the argument just used for $u = 3+i$, hence any Gaussian integer v has difference

$$r = v - uq \quad \text{of absolute value } < |u|$$

from some uq. In other words, we have the:

Division property. For any Gaussian integers u, v there are Gaussian integers q, r such that

$$v = uq + r \quad \text{with} \quad |r| < |u|.$$

Since the remainder is smaller than the divisor u, the Euclidean algorithm by repeated division with remainder will terminate for any pair s, t of Gaussian integers, giving us their gcd in the form

$$\gcd(s, t) = ms + nt \quad \text{for some Gaussian integers } m \text{ and } n.$$

Then, by the same trick used for ordinary integers in section 2.3 we can prove:

Prime divisor property. If a Gaussian prime p divides a product uv of Gaussian integers, then p divides u or p divides v.

As we know from section 2.3, *unique prime factorization* now follows. The only difference is that it is slightly “less unique.” This is because, when one Gaussian prime divides another, the quotient is not necessarily 1: it could also be -1 or $\pm i$. For this reason we say that *Gaussian prime factorization is unique up to factors of* ± 1 *and* $\pm i$.

2.7 Euler's Proof Revisited

Now we know how to talk about "primes" among the integers $a + b\sqrt{-2}$ brought to light in Euler's solution of $y^3 = x^2 + 2$.

We measure the size of $a + b\sqrt{-2}$ by its norm $|a + b\sqrt{-2}|^2 = a^2 + 2b^2$, and say that $a + b\sqrt{-2}$ is *prime* if its norm is greater than 1 and $a + b\sqrt{-2}$ is not the product of integers of smaller norm. (Incidentally, the only integers $a + b\sqrt{-2}$ of norm 1 are ± 1, because $a^2 + 2b^2 = 1$ only for $a = \pm 1$ and $b = 0$.)

We also know that we can prove unique prime factorization for the integers $a + b\sqrt{-2}$ if we can prove the *division property*: for any integers u, v of the form $a + b\sqrt{-2}$ there are integers q, r (also of this form) such that

$$v = uq + r \quad \text{with} \quad |r| < |u|.$$

And we know how to find q and r: look at the multiples uq of u, and let

$$r = v - (\text{multiple } uq \text{ nearest to } v).$$

Thus it remains only to get a view of the multiples of u—clear enough that we can see that r is smaller than u.

As with the Gaussian integers, we will illustrate the idea with a particular example: in this case $u = 1 + \sqrt{-2}$, $v = 5 + \sqrt{-2}$. Figure 2.6 shows some of the relevant points in the plane of complex numbers.

The integers $a + b\sqrt{-2}$ form a grid of rectangles of width 1 and height $\sqrt{2}$. The multiples of $1 + \sqrt{-2}$ form a grid of rectangles of the *same shape*, but magnified by $|1 + \sqrt{-2}| = \sqrt{5}$ and rotated so that their short sides lie in the direction of $1 + \sqrt{-2}$.

The multiple of $1 + \sqrt{-2}$ nearest to $5 + \sqrt{-2}$ is

$$(2 - \sqrt{-2})(1 + \sqrt{-2}) = 4 + \sqrt{-2},$$

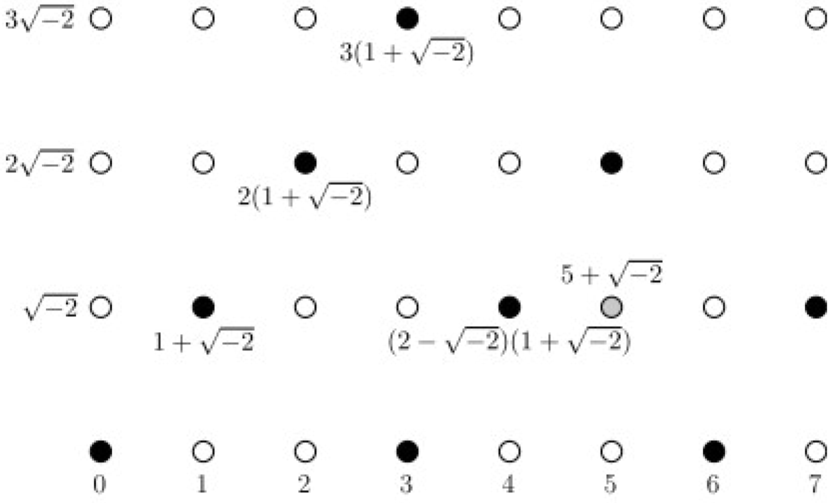

Figure 2.6: Multiples of $1+\sqrt{-2}$ near $5+\sqrt{-2}$.

so we get $q = 2 - \sqrt{-2}$ and $r = 1$. This r is certainly of smaller absolute value than the divisor $1+\sqrt{-2}$.

In general, the multiples of an integer $u = a + b\sqrt{-2}$ form a grid of rectangles of the same shape as the original grid, but magnified by $|u|$ and rotated so that their short sides are in the direction of u. Since the size $|r|$ of the remainder is the distance from v to the nearest multiple of u, it remains to show that *in a rectangle with short side* $|u|$ *and long side* $\sqrt{2}|u|$, *the distance from any point to the nearest corner is less than* $|u|$.

To see why this is true, consider the worst-case scenario shown in figure 2.7, where v is at the center of the rectangle.

By the Pythagorean theorem,

$$|r| = \sqrt{\frac{|u|^2}{2^2} + \frac{|u|^2}{2}} = \sqrt{\frac{3}{4}|u|^2} = \frac{\sqrt{3}}{2}|u| < |u|.$$

This finally proves the division property for the integers $a + b\sqrt{-2}$. As with ordinary and Gaussian integers, this gives the Euclidean algorithm, prime divisor property, and unique prime factorization. In fact, prime factorization is unique up to factors ±1 here, since the only integers $a + b\sqrt{-2}$ of norm 1 are ± 1.

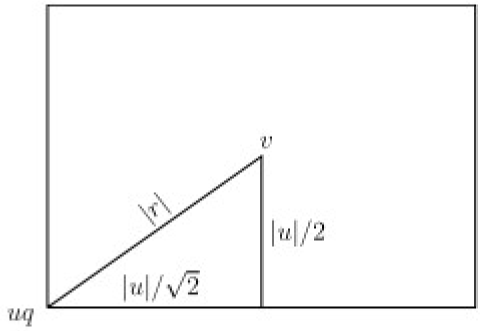

Figure 2.7: The point most distant from the corners.

Now we should recall *why* we want unique prime factorization for the integers $a + b\sqrt{-2}$. In section 2.5 we had

$$y^3 = x^2 + 2 = (x + \sqrt{-2})(x - \sqrt{-2}),$$

and we established that $x + \sqrt{-2}$ and $x - \sqrt{-2}$ have no common prime factor. But each prime factor p in y occurs as p^3 in y^3, so each prime factor p in $x + \sqrt{-2}$ occurs as $\pm p^3 = (\pm p)^3$, and each prime factor q in $x - \sqrt{-2}$ occurs as $\pm q^3 = (\pm q)^3$. In other words, $x + \sqrt{-2}$ *is a product of cubes and hence is a cube itself;* $x - \sqrt{-2}$ *is a product of cubes and hence is a cube itself.*

Euler's calculation and proof, starting with $x + \sqrt{-2} = (a + b\sqrt{-2})^3$, is now justified. The justification is probably far from any Euler had in mind, but it is based on ideas that now belong to elementary mathematics:

- The theory of divisibility and primes, based on the division property and the Euclidean algorithm.
- The multiplicative property of absolute value, $|uv| = |u||v|$, for complex numbers.
- Geometric representation of complex numbers and the Pythagorean theorem.

2.8 $\sqrt{2}$ and the Pell Equation

Another famous equation for which integer solutions are sought is

$$x^2 - 2y^2 = 1. \tag{*}$$

This equation was of interest to the ancient Greeks because, if x and y are large integers satisfying the equation, then x^2/y^2 is very close to 2. This implies that the fraction x/y is very close to $\sqrt{2}$.

In this section we will show that indeed there are arbitrarily large solutions of (*) and we will find a simple algorithm for producing them (and hence for finding arbitrarily close rational approximations to $\sqrt{2}$). We begin with the obvious solution $x = 3$, $y = 2$, which is the smallest positive integer solution.

Now since

$$1 = 3^2 - 2 \cdot 2^2$$

then

$$1 = (3^2 - 2 \cdot 2^2)^n \quad \text{for any positive integer } n.$$

Also,

$$3^2 - 2 \cdot 2^2 = (3 + 2\sqrt{2})(3 - 2\sqrt{2}),$$

so we have

$$1 = (3 + 2\sqrt{2})^n (3 - 2\sqrt{2})^n.$$

This prompts us to look at the positive integers x_n and y_n such that

$$(3 + 2\sqrt{2})^n = x_n + y_n\sqrt{2}. \tag{**}$$

For example,

$$(3 + 2\sqrt{2})^2 = 3^2 + 2 \cdot 3 \cdot 2\sqrt{2} + (2\sqrt{2})^2 = 17 + 12\sqrt{2},$$

so $x_2 = 17$ and $y_2 = 12$. We notice that $x = 17$, $y = 12$ is also a solution of (*).

In fact $x = x_n$, $y = y_n$ is a solution of (*) for each positive integer n. We prove this by induction.

Base step. The statement is true for $n = 1$ because $x_1^2 - 2y_1^2 = 3^2 - 2 \cdot 2^2 = 1$.

Induction step. We prove that if the statement is true for $n = k$ (that is, $x_k^2 - 2y_k^2 = 1$), then it is true for $n = k + 1$. By the definition (**) of x_{k+1} and y_{k+1},

$$\begin{aligned} x_{k+1} + y_{k+1}\sqrt{2} &= (3 + 2\sqrt{2})^{k+1} \\ &= (3 + 2\sqrt{2})^k (3 + 2\sqrt{2}) \\ &= (x_k + y_k\sqrt{2})(3 + 2\sqrt{2}) \\ &\qquad\qquad \text{(by definition of } x_k,\ y_k\text{)} \\ &= 3x_k + 4y_k + (2x_k + 3y_k)\sqrt{2}. \end{aligned}$$

Equating “rational and irrational parts” we get

$$x_{k+1} = 3x_k + 4y_k, \quad y_{k+1} = 2x_k + 3y_k. \tag{***}$$

It follows that

$$\begin{aligned} x_{k+1}^2 - 2y_{k+1}^2 &= (3x_k + 4y_k)^2 - 2(2x_k + 3y_k)^2 \\ &= 9x_k^2 + 24x_k y_k + 16y_k^2 - 2(4x_k^2 + 12x_k y_k + 9y_k^2) \\ &= x_k^2 - 2y_k^2 \\ &= 1, \qquad \text{(by induction hypothesis)} \end{aligned}$$

which completes the induction.

Thus $x_n^2 - 2y_n^2 = 1$ for all positive integers n.

It follows from the equations (***) that x_n and y_n increase with n (in fact, very rapidly). So x_n / y_n rapidly approaches $\sqrt{2}$ as n increases.

The first few approximations are

$$x_1/y_1 = 3/2 = 1.5$$

$$x_2/y_2 = 17/12 = 1.416\cdots$$

$$x_3/y_3 = 99/70 = 1.41428\cdots$$

$$x_4/y_4 = 577/408 = 1.4142156\cdots\ ,$$

which are correct in the first 1, 3, 5, 7 decimal places, respectively.

The Pell Equation

The equation $x^2 - 2y^2 = 1$ is a special case of the equation $x^2 - my^2 = 1$, for a nonsquare positive integer m, called the *Pell equation*. A proof similar to the one above shows that, if $x = x_1$, $y = y_1 \neq 0$ is one solution to the Pell equation, then we can find arbitrarily large solutions $x = x_n$, $y = y_n$ by the formula

$$x_n + y_n\sqrt{m} = (x_1 + y_1\sqrt{m})^n.$$

For example, $x = 2$, $y = 1$ is a solution of $x^2 - 3y^2 = 1$, and we get further solutions by the formula

$$x_n + y_n\sqrt{3} = (2 + \sqrt{3})^n.$$

In particular, using $(2 + \sqrt{3})^2 = 7 + 4\sqrt{3}$, we find the second solution $x = 7$, $y = 4$.

This is easy and fun to do, but two questions present themselves.

- How do we know that a solution of $x^2 - my^2 = 1$ *exists*?
- Given the smallest positive solution $x = x_1$ and $y = y_1$, does

  $$x_n + y_n\sqrt{m} = (x_1 + y_1\sqrt{m})^n$$

  give *all* positive integer solutions?

These questions are most easily answered by importing some ideas from outside number theory (the "pigeonhole principle" and some deeper algebra), so we will defer the answers until section 10.1.

5

Geometry

PREVIEW

Geometry was the first mathematical discipline to be developed in detail, and a large swath of elementary geometry is already in Euclid's *Elements*. The early sections of this chapter sample some of the highlights of Euclid's approach to geometry, which is an attractive blend of visualization and logic. In fact, for many people it is still the most compelling example of mathematical reasoning.

Euclid's geometry is literally "hands on," because it uses hand tools: the *straightedge* and the *compass*. These tools determine the subject matter of Euclid's geometry—straight lines and circles—but this subject matter embraces the measurement of length, area, and angles (and the sometimes unexpected connections between them, such as the Pythagorean theorem).

More surprisingly, Euclid's geometry has a rich *algebraic* content. This was unknown to Euclid, because the Greeks avoided numbers in geometry, so algebra was not even contemplated. It came to light only around 1630, when Fermat and Descartes introduced numbers and equations. The middle of this chapter describes the relationship, discovered by Descartes, between straightedge and compass constructions and *constructible numbers*—those obtainable from rational numbers by the operations of $+$, $-$, $\cdot$, $\div$, and $\sqrt{\ }$.

Finally, we discuss Euclid's modern incarnation: the *Euclidean geometry* of vector spaces with an *inner product*. Vector spaces capture the *linearity* of Euclid's geometry, traceable back to the straightedge, while the inner product captures the concept of length compatible with the Pythagorean theorem. Perhaps Euclid's geometry is so durable because it fits into both the ancient and modern worlds. Its ancient basis in hand tools—straightedge and compass—is exactly equivalent to its modern basis in vector spaces and inner products.

5.1 Numbers and Geometry

In modern mathematics, elementary geometry is *arithmetized* by reducing it to linear algebra over the real numbers. We will see how this is done in sections 5.7 and 5.8. Advanced geometry is arithmetized, too, by means of algebraic or differentiable functions of real or complex numbers. Yet somehow it is still easier for most humans to approach geometry visually, so visual objects re-emerge from the primordial soup of numbers and we find ways to simulate visual operations by operations on numbers. We represent points as ordered pairs (x, y) of real numbers, lines by linear equations $ax + by + c = 0$, circles by quadratic equations $(x - a)^2 + (y - b)^2 = r^2$; we find their intersection points by solving pairs of equations; we move objects by linear transformations of the set $\mathbb{R}^2$ of pairs (x, y); and so on. In this way we can model Euclid's geometry by higher-order concepts in the world of numbers.

This process has been recapitulated, in the last few decades, in the evolution of computers. Numbers are of course the native language of computers and, in their early decades of existence, computers were used mainly for processing numerical data. Then, in the 1970s, we saw the first crude printouts of images, typically made by assembling large arrays of symbols in a pattern printed on a long strip of paper. In the 1980s we started to program images, pixel by pixel, to be shown on computer screens. At the resolution then available (320×200) the results were usually hideous, as can be confirmed by looking at certain mathematics books of the period. When graphics commands were added to programming languages, it was a serious problem to program the best-looking "straight line" between two points. Any sloping line looked more or less like a staircase.

But resolution gradually improved and, by the 1990s, lines looked straight, curves looked smooth, and most pictures could be reproduced quite faithfully. This opened the floodgates of visual computing, leading

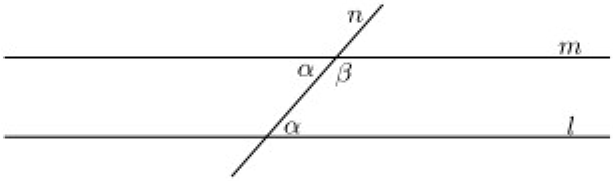

Figure 5.1: Angles involved in the parallel axiom.

to the world today in which users (particularly of phones and tablets) interact mainly with pictures on their screens: tapping them, translating and rotating them, magnifying them, and flipping them over by direct touch. It seems as though we *want* computers to be geometric, even if they have to be built on numbers, so programmers have to develop high-level programming concepts to model geometric operations.

In the present chapter we will study the interplay between numbers and geometry. There are reasons for doing geometry *without* a general concept of number, as we will see, and indeed Euclid did remarkable things without it. In the next two sections we review some of the big successes of his geometry: the theory of angles and the theory of area. But on balance geometry is easier with a suitable concept of number, and in the rest of the chapter we develop geometry on that basis.

5.2 Euclid's Theory of Angles

We are used to measuring angles by numbers: initially by degrees and at a higher level by radians, which involve the specific number π. But Euclid got by with the *right angle* as the unit of angle measure, and the concept of angle *equality*. His main conceptual tools for determining equality of angles were *congruence of triangles* and the *parallel axiom*. Indeed, the parallel axiom is the characteristic feature of Euclidean geometry, so we begin there. For convenience we use a slight variant of the parallel axiom, but it is equivalent to Euclid's.

Parallel Axiom. If lines l and m are parallel (that is, do not meet), and line n crosses them, then the angles α and β made by n with l and m, respectively, on the same side of n, have sum $\alpha + \beta$ equal to two right angles (figure 5.1).

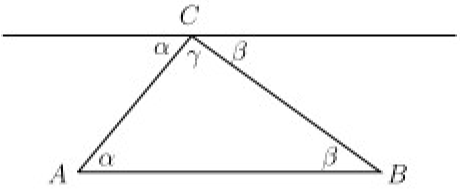

Figure 5.2: Angle sum of a triangle.

It follows that the angle adjacent to β is also α, since its sum with β is a straight angle, which equals two right angles. This immediately gives:

Angle sum of a triangle. *The angles α, β, γ of any triangle have sum $\alpha + \beta + \gamma$ equal to two right angles.*

Proof. Given a triangle ABC, with angles α, β, γ at the vertices A, B, C, respectively, consider the parallel to AB through the vertex C (figure 5.2).

Then it follows from the parallel axiom that angles α and β are adjacent to angle γ at C. Together, the angles at C below the parallel make a straight angle, so

$$\alpha + \beta + \gamma = \text{two right angles}. \qquad \square$$

Now we come to the use of congruence in establishing equality of angles. The relevant congruence axiom is the one today called SSS (for "side side side"): *if triangles ABC and $A'B'C'$ have corresponding sides equal then their corresponding angles are equal.* A famous consequence of SSS is a theorem about *isosceles* triangles (triangles with two equal sides).

Isosceles triangle theorem. *If ABC is a triangle with $AB = AC$, then the angle at B equals the angle at C.*

Proof. A stunning proof of this theorem (different from Euclid's) was given by Pappus, a Greek mathematician who lived some centuries later than Euclid.

Pappus noticed that triangle ABC is congruent to triangle ACB (yes, the same triangle, but "flipped," as it were) by SSS, because the

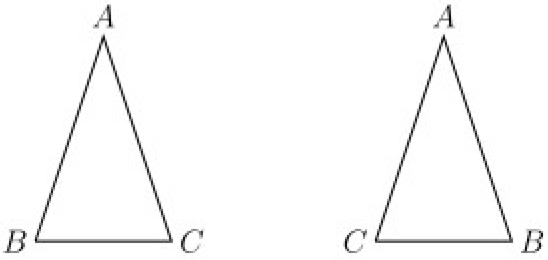

Figure 5.3: An isosceles triangle is congruent to itself.

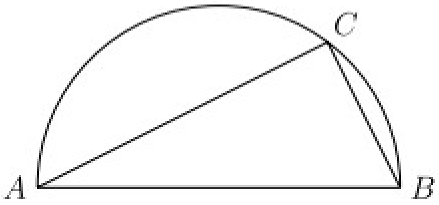

Figure 5.4: Angle in a semicircle.

corresponding sides

$$AB \text{ and } AC, \quad AC \text{ and } AB, \quad BC \text{ and } CB$$

are equal (see figure 5.3).

So the corresponding angles, at B and C, are also equal. □

Now we combine these two theorems about angles to obtain:

Angle in a semicircle. *If AB is the diameter of a circle, and C is another point on the circle, then the angle at C in triangle ABC is a right angle (figure 5.4).*

Proof. Let O be the center of the circle, and draw the line segment OC. This creates two isosceles triangles, OAC and OBC, because $OA = OB = OC$ (as radii of the circle). Then it follows from the isosceles triangle theorem that we have equal angles as shown in figure 5.5.

Also, since the angle sum of the triangle ABC is two right angles, we have

$$\alpha + (\alpha + \beta) + \beta = \text{two right angles.}$$

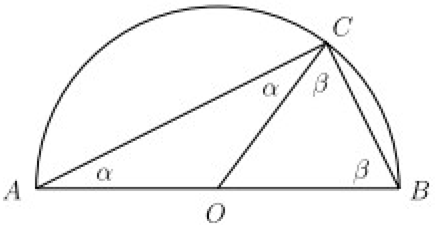

Figure 5.5: Isosceles triangles in a semicircle.

Therefore,

$$\alpha + \beta = \text{angle at C} = \text{right angle}. \qquad \square$$

5.3 Euclid's Theory of Area

According to legend, around 500 BCE the Pythagoreans discovered the irrationality of $\sqrt{2}$, or in their terms the *incommensurability* of the side and diagonal of the square. That is, there is no common unit of measure, u, such that the side and diagonal of the square are both integer multiples of u. Thus, the concept of positive integer—the *only* number concept the Pythagoreans had—is *inadequate* for describing all the lengths occurring in geometry. This momentous discovery led to a distinctive feature of Greek mathematics: the concept of *length* takes the place of number, but the arithmetic of lengths is quite limited. Lengths can be added and subtracted, but not multiplied. This means that area and volume have to be treated as separate kinds of magnitude, with even less capacity for addition and subtraction. In particular, the Greek concept of *equality* for areas and volumes is quite complicated.

Nevertheless, the Greeks were able to develop all the theory of area and volume needed for elementary mathematics. Here is how the theory of area gets off the ground.

Given lengths a and b, we modern mathematicians would like to say that the "product of a and b" is the area of the rectangle with adjacent sides a and b, which we may call the *rectangle of* a and b.

But what is area? It is certainly not a length, so from the Greek point of view there is no such thing as the area of the rectangle, other than the rectangle *itself* (figure 5.6). However there is a notion of *equality* for

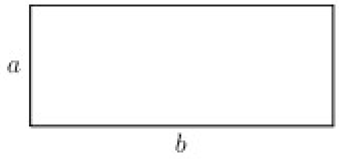

Figure 5.6: The rectangle of a and b.

rectangles which turns out to agree with the modern notion of equality of area, not only for rectangles but for all polygons. This notion of equality is based on the following principles, which are called "common notions" in Book I of Euclid's *Elements*.

1. Things which are equal to the same thing are also equal to each other.
2. If equals be added to equals, the wholes are equal.
3. If equals be subtracted from equals, the remainders are equal.
4. Things which coincide with one another are equal to one another.
5. The whole is greater than the part.

Notions 1, 2, 3, and 4 say essentially that objects are "equal" if one can be converted to the other by finitely often adding or subtracting identical ("coinciding") pieces. We already used these notions to prove the Pythagorean theorem in section 1.4. Notion 5 is not needed at this stage, because it turns out that Notions 1, 2, 3, and 4 suffice to prove that any polygons of equal area (in the modern sense) are "equal" in Euclid's sense.

The fact that "equal in area" is the same as "equal in Euclid's sense" (for polygons) is an elementary theorem. It was discovered in the nineteenth century by F. Bolyai and Gerwien, and is rather tedious to prove. Here we are content to show how the idea applies to the most important areas in elementary geometry—those of parallelograms and triangles. From now on we will say "equal in area" rather than "equal in Euclid's sense."

First, we see that the rectangle R of a and b is equal in area to any parallelogram P of height a and base b, because we can convert R to P by adding a triangle T to one side of R and then subtracting an identical triangle from the other side of the resulting shape (figure 5.7).

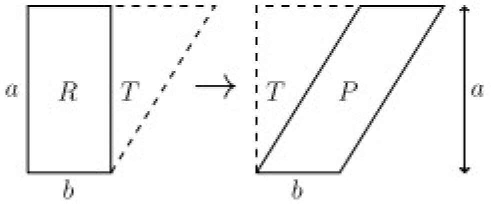

Figure 5.7: Converting a rectangle to a parallelogram.

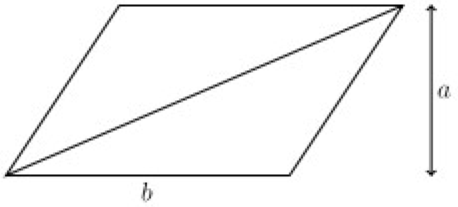

Figure 5.8: Making a parallelogram from two copies of a triangle.

Thus any parallelogram is equal in area to a rectangle with the same base and height. It follows that a triangle with height a and base b is equal in area to the rectangle with base b and half the height, $a/2$, because two copies of the triangle make a parallelogram of base b and height a (figure 5.8).

Since rectangles of equal area are actually "equal" in Euclid's sense we can *define* the product ab of lengths a and b to be the rectangle of a and b, and call products "equal" if the corresponding rectangles are "equal" in Euclid's sense. Then if $ab = cd$, we can *prove* this fact by converting the rectangle of c and d to the rectangle of a and b by adding and subtracting identical pieces. Thus, $ab = cd$ is actually *equivalent* to "equality" (in Euclid's sense) of the rectangle of a and b with the rectangle of c and d.

Even without knowing this general fact (which depends on the theorem of Bolyai and Gerwein) the Greeks could have proved "equality" of some interesting irrational rectangles, though it is not known whether they ever did so.

Geometric Proof that $\sqrt{2} \cdot \sqrt{3} = \sqrt{6}$

One of the supposed advantages of the theory of real numbers, introduced by Dedekind (1872), is that it permits rigorous proofs of

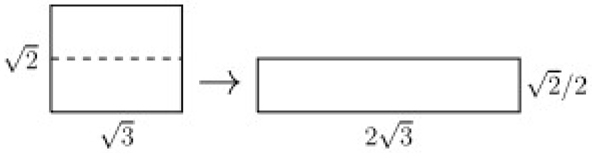

Figure 5.9: Converting the $\sqrt{2}$ by $\sqrt{3}$ rectangle to half height.

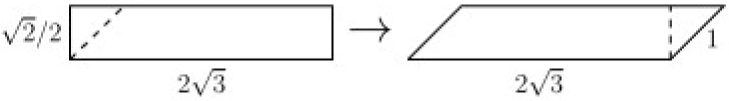

Figure 5.10: Converting rectangle to parallelogram.

results such as $\sqrt{2} \cdot \sqrt{3} = \sqrt{6}$. Indeed, Dedekind speculated that, with his definition of real number,

> we arrive at real proofs of theorems (as, e.g., $\sqrt{2} \cdot \sqrt{3} = \sqrt{6}$) that have never been established before.
>
> Dedekind (1901), p. 22

It may be true that $\sqrt{2} \cdot \sqrt{3} = \sqrt{6}$ had never really been proved before, but there was nothing to stop the Greeks from proving an equivalent proposition. Gardiner (2002), pp. 181–183 has pointed out that we can interpret and prove this equation quite easily in terms the Greeks would have accepted. Here is one such proof (a variation on Gardiner's), which is surely elementary.

By the Pythagorean theorem, $\sqrt{2}$ is the diagonal of the unit square and $\sqrt{3}$ is a side of the right-angled triangle with hypotenuse 2 and other side 1. Thus we can interpret $\sqrt{2}$ and $\sqrt{3}$ as lengths, and it is meaningful to speak of the rectangle of $\sqrt{2}$ and $\sqrt{3}$. We now transform this rectangle, by adding and subtracting equal pieces, into a rectangle of base 1.

First, cut the rectangle in two equal parts by a line parallel to the base $\sqrt{3}$, and then combine these two parts into a rectangle of height $\sqrt{2}/2$ and base $2\sqrt{3}$ (figure 5.9).

These rectangles are "equal" in Euclid's sense.

Next, we convert the new rectangle into a parallelogram by cutting off a right-angled triangle with equal sides $\sqrt{2}/2$ (and hence hypotenuse 1 by the Pythagorean theorem) from the left end and attaching it to the right (figure 5.10).

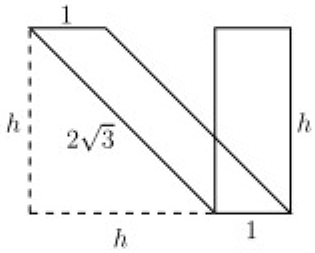

Figure 5.11: Converting parallelogram to rectangle with base 1.

The result is a parallelogram of the same width but sloping sides (at angle 45°) of length 1, and of course the same area again.

Finally, viewing the side 1 as the base of the parallelogram, we convert it to a rectangle with the same base and height, and hence the same area, by adding and subtracting a triangle (figure 5.11). It remains to find the height h of the rectangle.

We notice from figure 5.11 that h is the side of the isosceles right-angled triangle with hypotenuse $2\sqrt{3}$. Hence, by the Pythagorean theorem again,

$$h^2 + h^2 = (2\sqrt{3})^2 = 4 \times 3 = 12, \quad \text{so} \quad h^2 = 6.$$

This completes the proof that the rectangle with sides $\sqrt{2}$ and $\sqrt{3}$ is equal in area to the rectangle with sides $\sqrt{6}$ and 1. Hence (in modern terms) $\sqrt{2} \cdot \sqrt{3} = \sqrt{6}$.

The Concept of Volume

Euclid had a theory of volume that begins in the same way as his theory of area. The basic solid object is the *box* of a, b, and c; that is, with height a and base the rectangle of b and c, and all faces rectangular. "Equality" is defined by adding and subtracting identical pieces, and one finds that a *parallelepiped*[1] (a "squashed box" whose faces are parallelograms) is "equal" to the box with the same height, width, and depth (figure 5.12).

Next, by cutting a parallelepiped in half one finds that a *prism* is "equal" to the box with the same base but half the height. But that is

---

[1] This word is often misunderstood, misspelled, and mispronounced. It helps to read its parts as parallel-epi-ped, literally "parallel upon the foot," meaning that the top is parallel to the bottom, no matter which way is up.

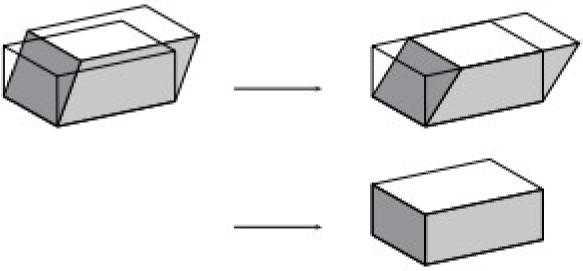

Figure 5.12: Transforming a squashed box to a box by adding and subtracting.

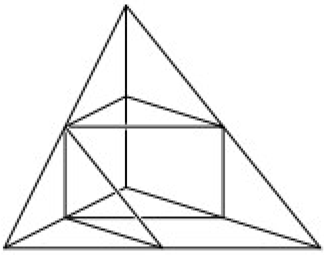

Figure 5.13: Two prisms in the tetrahedron.

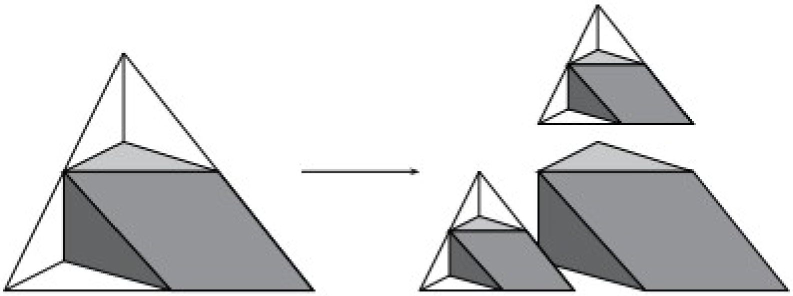

Figure 5.14: Continuing the dissection of the tetrahedron.

as far as one can go using finite numbers of pieces. In the case of the *tetrahedron*, Euclid was able to find an "equal" box only by cutting the tetrahedron into an *infinite* number of prisms. Figure 5.13 shows the first two prisms inside the tetrahedron, and figure 5.14 shows how to obtain further prisms, inside the tetrahedra that remain after the first two prisms are removed.

Much later, Dehn (1900) showed that the infinity is unavoidable: it is not possible to convert a regular tetrahedron to a cube by adding and subtracting finitely many pieces. This surprising discovery shows that volume is a deeper concept than area. If we suppose that volume belongs to elementary mathematics (as we presumably do), then we must accept that infinite processes have a place in elementary mathematics too. We pursue this train of thought further in the next chapter.

5.4 Straightedge and Compass Constructions

Many of the propositions in Euclid's *Elements* are *constructions*; that is, claims that certain figures in the plane can be constructed by drawing straight lines and circles. These constructions rest upon the ability to do two things that Euclid includes in his "postulates" (which we call axioms):

1. To draw a straight line from a given point to any other given point.
2. To draw a circle with given center and radius.

(Euclid actually splits the first axiom into two—drawing the line *segment* between given points and extending the segment arbitrarily far—because he does not admit infinite lines.) We now call the instruments for drawing lines and circles the "straightedge" and "compass," respectively,[2] so Euclid's constructions are called *straightedge and compass constructions*.

Euclid's very first proposition is a straightedge and compass construction: *drawing an equilateral triangle whose base is a given line segment* AB. It nicely illustrates how the compass is not only an instrument for drawing circles, but also for "storing" and "copying" lengths, because it allows a length to be carried from one place to another.

The third vertex C of the triangle is found at the intersection of the circles with centers A and B, respectively, and radius AB (figure 5.15). The triangle is then completed by drawing the line segments from A to C and from B to C.

---

[2] The straightedge is often called a ruler, but this wrongly suggests that there are marks on its edge that could be used for measurement. The compass used to be called a "pair of compasses." This term is falling into disuse, but I mention it because you may still see it in older books.

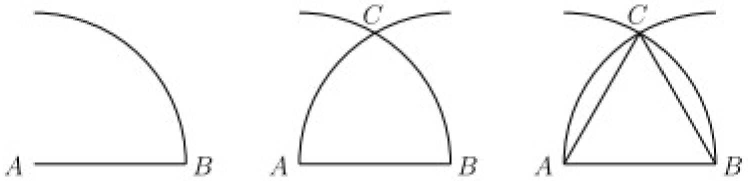

Figure 5.15: Constructing the equilateral triangle.

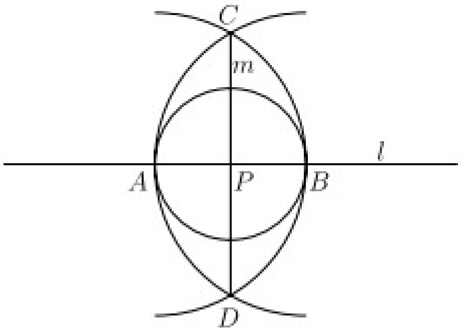

Figure 5.16: Constructing a perpendicular.

To prove that the construction is correct we observe that

$$\begin{aligned} & AC = AB \quad \text{because both are radii of the first circle,} \\ & BC = AB \quad \text{because both are radii of the second circle,} \\ \text{so} \quad & AB = AC = BC, \end{aligned}$$

because "things equal to the same thing are equal to one another."

Many other constructions in elementary geometry are variations on this one.

For example, the other intersection D of the two circles above gives a line CD *perpendicular* to AB and passing through its midpoint P. Conversely, by drawing a circle centered on an arbitrary point P on a line l we can make P the midpoint of a segment AB, and hence construct a line m perpendicular to l through P (figure 5.16).

Then, by constructing a perpendicular to m, we get a *parallel* n to the original line l. Parallels are a fundamental concept of Euclidean geometry, and they are the key to many constructions, because of the following theorem, which refers to figure 5.17.

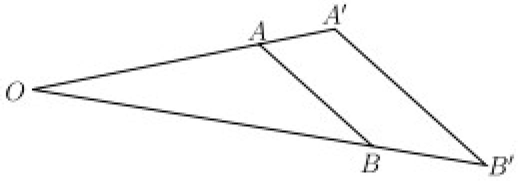

Figure 5.17: Setting of Thales' theorem.

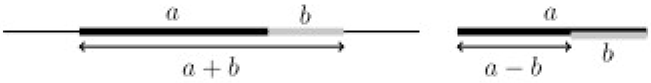

Figure 5.18: Forming the sum and difference of lengths.

Theorem of Thales. *If OAB and $OA'B'$ are triangles such that OAA' and OBB' are straight lines and AB is parallel to $A'B'$, then*

$$OA/OA' = OB/OB'.$$

Euclid's proof of this theorem is quite subtle, because his avoidance of numbers makes it hard to interpret the ratios OA/OA' and OB/OB'. When we accept that lengths are numbers we can argue much more simply (see section 5.7). This paves the way for straightedge and compass constructions that simulate all the arithmetic operations $+, -, \cdot, \div$ on lengths, as well as the $\sqrt{}$ operation. We carry out the most interesting parts of this program in the next section. But first we dispose of the $+$ and $-$ operations, which are simple even from Euclid's point of view and do not involve the theorem of Thales.

Given line segments a and b, we form $a + b$ by first extending the line segment a sufficiently far, then copying the line segment b with the compass and placing it alongside a in the same line. Similarly, if b is no greater than a, we form $a - b$ by carrying b to the interior of a with one end of b at the end of a (figure 5.18).

5.5 Geometric Realization of Algebraic Operations

As we saw at the end of the previous section, it is easy to obtain the sum and difference of lengths by constructions within the line. Product and quotient are also easy to obtain, but the construction makes use of the

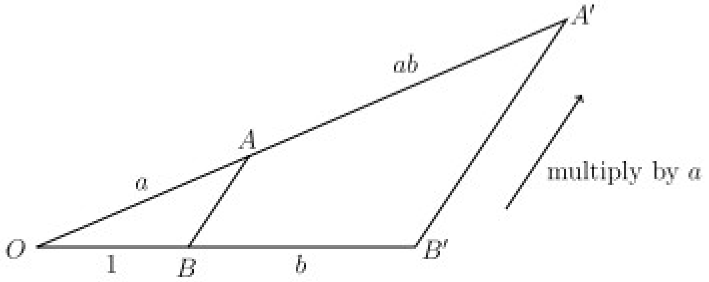

Figure 5.19: Constructing the product of lengths.

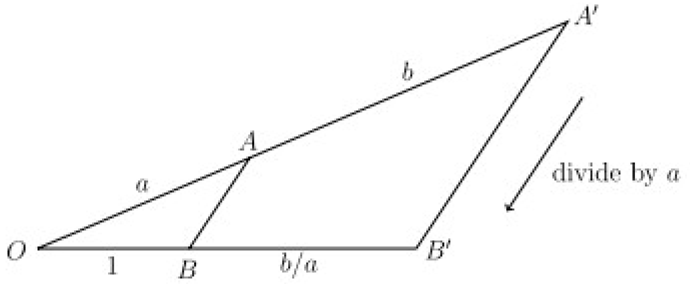

Figure 5.20: Constructing the quotient of lengths.

plane, and parallel lines. We also need to fix a *unit length*. But then, if we are given a length a, we can multiply any length b by a using the construction shown in figure 5.19.

The construction uses two lines through O. On one of them we mark the unit length $1 = OB$; on the other, the length $a = OA$. Then on the first line we add the length $b = BB'$. Finally, we draw AB and its parallel $A'B'$ through B'. By the theorem of Thales it follows that $AA' = ab$.

In effect, the parallels achieve "magnification by a," mapping the segments of length 1, b onto segments of length a, ab. The reverse direction gives "division by a," and hence we can obtain b/a for any lengths a, b as shown in figure 5.20.

Thus we now have geometric realizations of the operations $+, -, \cdot, \div$. To realize the operation $\sqrt{}$ is also surprisingly easy, but first we need the following consequence of the theorem of Thales.

Proportionality of Similar Triangles. *If ABC and $A'B'C'$ are triangles whose corresponding angles are equal (that is, angle at A equals angle*

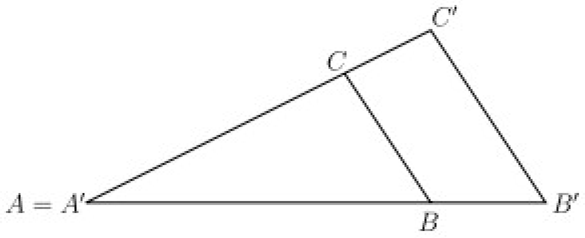

Figure 5.21: Similar triangles.

at A' and so on) then their corresponding sides are proportional. That is

$$\frac{AB}{A'B'} = \frac{BC}{B'C'} = \frac{CA}{C'A'}.$$

Proof. Move the triangle $A'B'C'$ to the position where $A = A'$, A, B, B' are in the same line, and A, C, C' are in the same line. We assume A, C, C' are in that order, in which case we have the situation shown in figure 5.21. (The argument is similar if the order is A, C', C.)

Since the angles at C and C' are equal, the lines BC and $B'C'$ are parallel. Then it follows from the theorem of Thales that the upper and lower sides are divided proportionally; that is, $AB/A'B' = AC/A'C'$. Similarly, by moving the triangles so that B coincides with B', we find that $BC/B'C' = BA/B'A' = AB/A'B'$. So in fact all pairs of corresponding sides are proportional. □

Now to construct $\sqrt{l}$ for any length l we make the construction shown in figure 5.22. This is Euclid's construction in his Book VI, Proposition 13.

As we saw in section 5.2, the angles α and β at A and B, respectively, have sum $\alpha + \beta$ equal to a right angle. It follows, since the angle sum of each triangle is two right angles, that the angles at C are also α and β as shown. Thus triangles ADC and CDB are similar. Comparing corresponding sides we obtain the equal ratios

$$1/h = h/l,$$

from which it follows that

$$h^2 = l, \quad \text{and therefore} \quad h = \sqrt{l}.$$

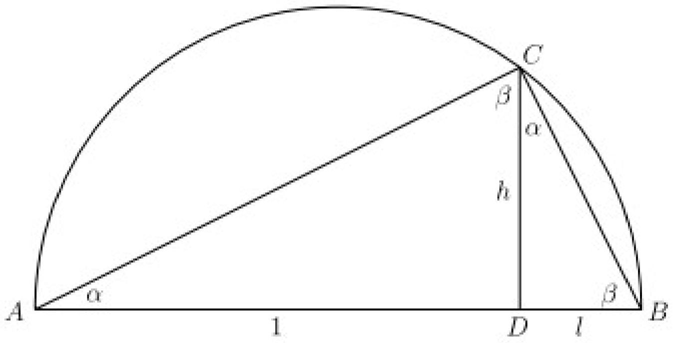

Figure 5.22: Constructing the square root of a length.

Moreover, we can *construct* $\sqrt{l}$ in the following steps.

1. Add l to 1 in the same line to form the line segment AB.
2. Bisect AB to find the center of the circle with diameter AB.
3. Draw this circle.
4. Construct the perpendicular to AB at the point D where the segments 1 and l meet.
5. Find the intersection C of this perpendicular with the circle of diameter AB and draw CD.

5.6 Algebraic Realization of Geometric Constructions

In the previous section we saw that straightedge and compass constructions include the algebraic operations of $+$, $-$, $\cdot$, $\div$, and $\sqrt{}$. In this section we will show that, when straightedge and compass operations are interpreted in the plane $\mathbb{R}^2$, *all* constructible lengths are obtainable from the unit length by the above operations. Thus we have an equivalence between geometric and algebraic concepts, and with it the opportunity to prove existence or nonexistence of geometric constructions by algebra. We will give an example of existence at the end of this section, and an example of nonexistence in section 5.9.

We begin by using coordinates to interpret the two basic constructions: drawing the line through two given points, and drawing the circle with given center and radius.

E 4)

( )))

E)-

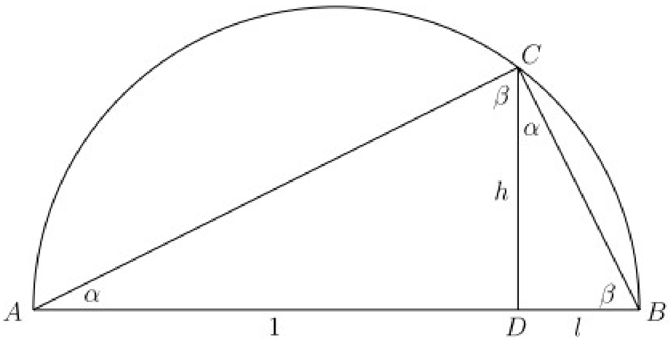
C
β
α
h
A
α
1
D
l
β
B

. - / - =)) -

% 8)
>
- 4
") - "&
& "& P) "&
!)
6 4 4) "&) 4 !
-)
/ . 4 4)
) "&)) !

5- N72 T63

y coordinates of the intersection come from the solutions for x and the first equation, again using $+, -, \cdot, \div$.

- Given two circles $(x-a_1)^2+(y-b_1)^2=r_1^2$ and $(x-a_2)^2+(y-b_2)^2=r_2^2$, we expand these two equations as

  $$x^2-2a_1x+a_1^2+y^2-2b_1y+b_1^2=r_1^2,$$
  $$x^2-2a_2x+a_2^2+y^2-2b_2y+b_2^2=r_2^2.$$

  Subtraction cancels the x^2 and y^2 terms, leaving a linear equation whose common solutions with either of the circle equations can be found as in the previous case.

To sum up: the coordinates of all new points, and hence the coefficients in the equations of all new lines and circles, are obtainable from the coordinates of the initially given points by the operations $+, -, \cdot, \div$, and $\sqrt{\ }$. Therefore, the coordinates of all points constructible by straightedge and compass are obtainable from 1 by these operations. Conversely, all points whose coordinates are numbers obtainable from 1 by the operations $+, -, \cdot, \div$, and $\sqrt{\ }$ (which we may call *constructible numbers*) are constructible by straightedge and compass.

Now here is the promised example: using algebra to prove constructibility.

The Regular Pentagon

Figure 5.23 shows the regular pentagon with sides of length 1 and a diagonal whose length x is sought. The other two diagonals are drawn to create some similar triangles. The symmetry of the regular pentagon ensures that each diagonal is parallel to the opposite side, and the various parallels imply that the triangles ABC, $A'BC$, and $A'B'C'$ have the equal angles α shown, and hence also equal angles β.

Thus $A'BC$ is similar to ABC and, since they have the common side BC, they are actually *congruent*. This gives the equal sides of length 1 shown in figure 5.24, hence also the sides of length $x-1$, since all diagonals have length x.

Then similarity of triangles ABC and $A'B'C'$ gives

$$\frac{1}{x}=\frac{x-1}{1}$$

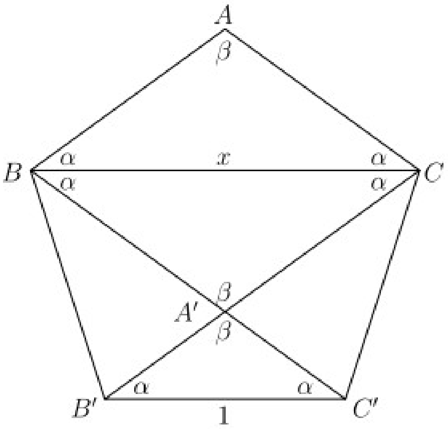

Figure 5.23: Angles in the regular pentagon.

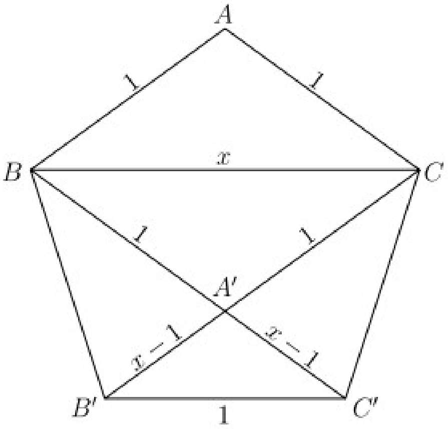

Figure 5.24: Lengths in the regular pentagon.

from the proportionality of corresponding sides. This is the quadratic equation $x^2 - x - 1 = 0$, which has the solutions $x = \frac{1 \pm \sqrt{5}}{2}$. The positive member of this pair is the length of the pentagon diagonal. It is a constructible number, so we can construct the diagonal by straightedge and compass. It is then easy to construct the triangle ABC, and to join a series of such triangles together to form the regular pentagon.

5- T 2 <73

among the next simplest, with constant (but negative) curvature at all points. In the modern geometry of curved spaces, Euclidean spaces have a special place as *tangent spaces*. A curved manifold has a tangent space at each point, and one often works in the tangent space to take advantage of its simpler structure (particularly, its nature as a vector space).

It is all very well to base geometry on the theory of real numbers, but how well do we understand the real numbers? What is *their* foundation? Hilbert (1899) raised this question, and he had an interesting answer: the real numbers can be based on geometry! More precisely, they can be based on a "completed" version of Euclid's geometry. Hilbert embarked on the project of completing Euclid's axioms in the early 1890s, first with the aim of filling in some missing steps in Euclid's proofs. As the project developed, he noticed that addition and multiplication arise from his axioms in a quite unexpected way, so that the field concept can be given a completely geometric foundation. See section 5.11. Then, by adding an axiom guaranteeing that the line has no gaps, he was able to recover a complete "number line" with the usual properties of $\mathbb{R}$.

*Non-Euclidean Geometry

In the 1820s, Janos Bolyai[5] and Nikolai Lobachevsky independently developed a rival geometry to Euclid's: a *non*-Euclidean geometry that satisfied all of Euclid's axioms except the parallel axiom. The parallel axiom has a different character from the other axioms, which describe the outcome of finite constructions or "experiments" one can imagine carrying out:

1. Given two points, draw the line segment between them.
2. Extend a line segment for any given distance.
3. Draw a circle with given center and radius.
4. Any two right angles are equal (that is, one can be moved to coincide with the other).

The parallel axiom, on the other hand, requires an experiment that involves an indefinite wait:

---

[5] Son of the Bolyai in the Bolyai-Gerwien theorem.

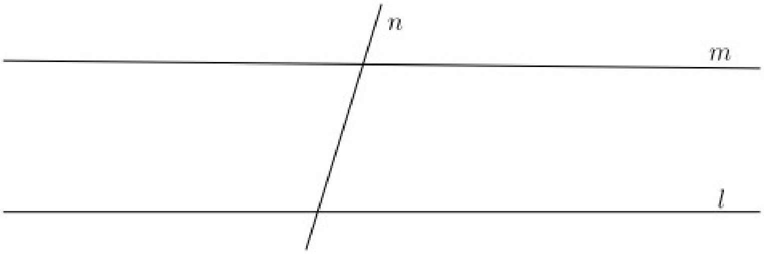

Figure 5.31: Two lines that are supposed to meet.

5. Given two lines l and m, crossed by another line n making interior angles with l and m together less than two right angles, the lines l and m will meet, *if produced indefinitely* (figure 5.31).

Since the time of Euclid, mathematicians have been unhappy with the parallel axiom, and have tried to prove it from the other, more constructive, axioms. The most determined attempt was made by Saccheri (1733), who got as far as showing that, if non-diverging lines l and m did *not* meet, they would have a common perpendicular at infinity. This, Saccheri thought, was "repugnant to the nature of straight lines." But it was not a contradiction, and in fact there is a geometry in which lines behave in precisely this fashion.

Bolyai and Lobachevsky worked out a large and coherent body of theorems that follow from Euclid's axioms 1 to 4 together with the axiom:

5′. There exist two lines l and m, which do not meet, although they are crossed by another line n making interior angles with l and m together less than two right angles.

Their results were eventually published in Lobachevsky (1829) and Bolyai (1832) (the latter an appendix to a book by Bolyai's father). They found no contradiction arising from this set of axioms, and Beltrami (1868) showed that no contradiction exists, because axioms 1, 2, 3, 4, and 5′ are satisfied under a suitable interpretation of the words "point," "line," and "angle." (We sketch one such interpretation below.) Thus Euclid's geometry had a rival, and deciding how to interpret the terms "point," "line," "distance," and "angle" became an issue.

As we have seen, Euclid's axioms had a ready-made interpretation in the coordinate geometry of Descartes: a "point" is an ordered pair (x, y) of real numbers, a "line" consists of the points satisfying an equation $ax + by + c = 0$, and the "distance" between (x_1, y_1) and (x_2, y_2) equals $\sqrt{(x_2 - x_1)^2 + (y_2 - y_2)^2}$.

For Bolyai's and Lobachevsky's axioms, Beltrami found several elegant interpretations, admittedly with a somewhat complicated concept of "distance." The simplest is probably the *half-plane model*, in which:

- "points" are points of the upper half-plane; that is, pairs (x, y) with $y > 0$,
- "lines" are the open semicircles in the upper half-plane with their centers on the x-axis, and the open half-lines $\{(x, y) : x = a, y > 0\}$,
- the "distance" between "points" P and Q is the integral of $\sqrt{dx^2 + dy^2}/y$ over the "line" connecting P and Q.

It turns out that "angle" in this model is just the ordinary angle between curves; that is, the angle between their tangents. This leads to some beautiful pictures of non-Euclidean geometric configurations, such as the one in figure 5.32. It shows a tiling of the half-plane by triangles which are "congruent" in the sense of non-Euclidean distance. In particular, they each have angles $\pi/2$, $\pi/3$, $\pi/7$. Knowing that they are congruent one can get a sense of non-Euclidean distance. One can see that the x-axis is infinitely far away—which explains why it is not included in the model—and perhaps also see that a "line" is the shortest path joining its endpoints, if one estimates distance by counting the number of triangles along a path.

It is also clear that the parallel axiom fails in the model. Take for example the "line" that goes straight up the center of the picture and the point on the far left with seven "lines" passing through it. Several of the latter "lines" do not meet the center line.

*Vector Space Geometry

Grassmann (1861), the *Lehrbuch der Arithmetik* mentioned in section 1.9, was not Grassmann's only great contribution to the

Numbers and Geometry

Now let us return to the axioms of Hilbert (1899), and what they tell us about the relationship between numbers and geometry. Hilbert's axioms probably capture Euclid's concept of the line (with the finer structure explored in Book V of the *Elements*). So, given the Descartes "model" $\mathbb{R}^2$ of Euclid's axioms, Hilbert has shown Euclidean geometry to be essentially *equivalent* to the algebra of $\mathbb{R}$. However, algebraists and logicians now prefer not to use the full set $\mathbb{R}$ in geometry. They point out that the set of constructible numbers suffices, because Euclid's geometry "sees" only the points arising from straightedge and compass constructions. Thus one can get by with an algebraically defined set of points, which is only a "potential" infinity, in contrast to the "actual" infinity $\mathbb{R}$. Logicians also prefer the theory of constructible numbers because its "consistency strength" is less than that of the theory of $\mathbb{R}$.

That is, it is easier to prove the consistency of the theory of constructible numbers (and hence the consistency of Euclid's axioms) than it is to prove the consistency of the theory of $\mathbb{R}$ (and hence the consistency of Hilbert's axioms).

Geometry and "Reverse Mathematics"

In recent decades, mathematical logicians have developed a field called *reverse mathematics*, whose motivation was stated by Friedman (1975) as follows:

> When the theorem is proved from the right axioms, the axioms can be proved from the theorem.

As logicians understand it, reverse mathematics is a technical field, concerned mainly with theorems about the real numbers (see section 9.9). However, if we understand reverse mathematics more broadly as the search for the "right axioms," then reverse mathematics began with Euclid.

He saw that the parallel axiom is the right axiom to prove the Pythagorean theorem, and perhaps the reverse—that the Pythagorean theorem proves the parallel axiom (given his other axioms). The same is true of many other theorems of Euclidean geometry, such as the theorem of Thales and the theorem that the angle sum of a triangle is π. All of these theorems are equivalent to the parallel axiom, so it is the "right axiom" to prove them.

To formalize this and other investigations in reverse mathematics we choose a *base theory* containing the most basic and obvious assumptions about some area of mathematics. It is to be expected that the base theory will fail to prove certain interesting but less obvious theorems. We then seek the "right" axiom or axioms to prove these theorems, judging an axiom to be "right" if it implies the theorem, and conversely, using only assumptions from the base theory.

Euclid began with a base theory now known as *neutral geometry*. It contains basic assumptions about points, lines, and congruence of triangles but *not* the parallel axiom. He proved as many theorems as he could before introducing the parallel axiom—only when it was needed to prove theorems about the area of parallelograms and ultimately the Pythagorean theorem. He also needed the parallel axiom to prove the theorem of Thales and that the angle sum of a triangle is π. We now know, conversely, that all of these theorems imply the parallel axiom in neutral geometry, so the latter is the "right" axiom to prove them.

Neutral geometry is also a base theory for *non*-Euclidean geometry, because the latter is obtained by adding to neutral geometry the "non-Euclidean parallel axiom" stating that there is more than one parallel to a given line through a given point.

Grassmann's theory of real vector spaces, as we have seen, can also be taken as a base theory for Euclidean geometry. It is quite different from the base theory of neutral geometry because the Euclidean parallel axiom *holds* in real vector spaces, and so does the theorem of Thales. Nevertheless, this new base theory is not strong enough to prove the Pythagorean theorem, or indeed to say anything about angles. Relative to the theory of real vector spaces, the "right" axiom to prove the Pythagorean theorem is existence of the inner product, because we can reverse the implication by using the Pythagorean theorem to define distance, hence angle and cosine, and then define the inner product by

$$\boldsymbol{u} \cdot \boldsymbol{v} = |\boldsymbol{u}| \cdot |\boldsymbol{v}| \cos \theta.$$

This raises the possibility of adding a different axiom to the theory of real vector spaces and obtaining a different kind of geometry, just as we obtain non-Euclidean geometry from neutral geometry by adding a different parallel axiom. Indeed we can, and simply by asserting the existence of a different kind of inner product. The inner product introduced by Grassmann is what we now call a *positive-definite* inner product, characterized by the property that $\boldsymbol{u} \cdot \boldsymbol{u} = 0$ only if $\boldsymbol{u}$ is the zero vector.

Non-positive-definite inner products also arise quite naturally. Probably the most famous is the one on the vector space $\mathbb{R}^4$ that defines the *Minkowski space* of Minkowski (1908). If we write the typical vector in $\mathbb{R}^4$ as $\boldsymbol{u} = (w, x, y, z)$ then the Minkowski inner product is defined by

$$\boldsymbol{u}_1 \cdot \boldsymbol{u}_2 = -w_1 w_2 + x_1 x_2 + y_1 y_2 + z_1 z_2.$$

In particular, the length $|\boldsymbol{u}|$ of a vector $\boldsymbol{u}$ in Minkowski space is given by

$$|\boldsymbol{u}|^2 = \boldsymbol{u} \cdot \boldsymbol{u} = -w^2 + x^2 + y^2 + z^2,$$

so $|\boldsymbol{u}|$ can certainly be zero when $\boldsymbol{u}$ is not the zero vector.

Minkowski space is famous as the geometric model of Einstein's *special relativity theory*. In this model, known as *flat spacetime*, x, y, and z are the coordinates of ordinary three-dimensional space and $w = ct$, where t is the time coordinate and c is the speed of light. As Minkowski (1908) said:

> The views of space and time which I wish to lay before you have sprung from the soil of experimental physics, and therein lies their strength. They are radical. Henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality.

Undoubtedly, relativity theory put non-positive-definite inner products on the map, making them as real and important as the ancient concept of distance. But in fact such inner products had already been considered by mathematicians, and one of them is involved in a model of non-Euclidean geometry discovered by Poincaré (1881)—the so-called *hyperboloid model*.

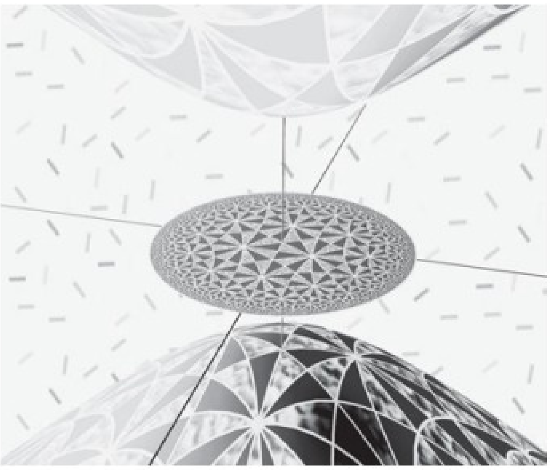

Figure 5.34: Triangle tessellation of the hyperboloid model.

To see where the hyperboloid comes from, consider the three-dimensional Minkowski space of vectors $\boldsymbol{u} = (w, x, y)$ with one time coordinate w and two space coordinates x, y. In this space, where $|\boldsymbol{u}|^2 = -w^2 + x^2 + y^2$, we consider the "sphere of imaginary radius"

$$\{\boldsymbol{u} : |\boldsymbol{u}| = \sqrt{-1}\} = \{(w, x, y) : -w^2 + x^2 + y^2 = -1\}.$$

This "sphere"[8] consists of the points (w, x, y) in $\mathbb{R}^3$ such that

$$w^2 - x^2 - y^2 = 1,$$

so it is actually a hyperboloid; namely, the surface obtained by rotating the hyperbola $w^2 - y^2 = 1$ in the (w, y)-plane about the w axis. If we take "distance" on either sheet of the hyperboloid to be the Minkowski distance, it turns out to be a model of the non-Euclidean plane. As with the other models of the non-Euclidean plane, calculating distance is a little troublesome so I omit the details. Instead, I offer figure 5.34—a black-and-white version of a picture due to Konrad Polthier of the Freie Universität, Berlin—which shows what the triangle tessellation of

---

[8] Long before the development of Minkowski space, or even the development of non-Euclidean geometry, Lambert (1766) speculated about the geometry of a "sphere of imaginary radius." In particular, he guessed that the angle sum of a triangle would be less than π on such a sphere—which is essentially the non-Euclidean parallel axiom.

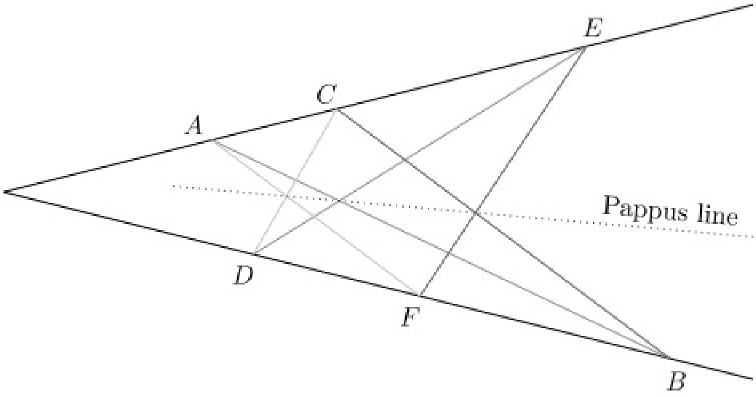

Figure 5.35: The Pappus configuration.

figure 5.32 looks like in the hyperboloid model. (It also relates the hyperboloid model to another model—the *conformal disk* model—which is a kind of intermediary between the hyperboloid model and the half-plane model shown in figure 5.32.)

This elegant relationship between Minkowski space and the non-Euclidean plane has been used for some textbook treatments of non-Euclidean geometry, such as Ryan (1986). Just as the positive-definite inner product is the "right axiom" to develop Euclidean geometry over the base theory of real vector spaces, the Minkowski inner product is the "right axiom" to develop non-Euclidean geometry.

*Projective Geometry

Another wonderfully "right" axiom is the theorem discovered by Pappus a few hundred years after Euclid. Pappus viewed his theorem as part of Euclidean geometry, but it does not really belong there. It is unlike typical Euclidean theorems in making no mention of length or angle, so its home should be a geometry that does not involve these concepts. The statement of the theorem is the following, which refers to the configuration shown in figure 5.35.

Theorem of Pappus. *If A, B, C, D, E, F are points of the plane lying alternately on two lines, then the intersections of the pairs of lines AB and DE, BC and EF, CD and FA, lie on a line.*

The Pappus theorem has a Euclidean *proof* using the concept of length, and also a coordinate proof using linear equations to define lines. But it seems to have no proof using only concepts "appropriate" to its statement: points, lines, and the membership of points in lines. The appropriate setting for the Pappus theorem is the *projective geometry* of the plane, a geometry which tries to capture the behavior of points and lines in a plane without regard to length or angle. In projective geometry, configurations of points and lines are considered the same if one can be projected onto the other. Projection can of course change lengths and angles, but the straightness of lines remains, as does the membership of points in lines.

If one seeks axioms for projective plane geometry, the following come easily to mind:

1. Any two points belong to a unique line.
2. Any two lines have a unique point in common.
3. There are four points, no three of which are in the same line.

The first axiom is also one of Euclid's. The second disagrees with Euclid in the case of parallel lines, but projective geometry demands it, because even parallel lines can be projected so that they meet—on the "horizon." The third axiom is there to ensure that we really have a "plane," and not merely a line. However, these simple axioms are very weak, and it can be shown that they do not suffice to prove the Pappus theorem. They do, however, form a natural base theory to which other axioms about points and lines can be added.

What are the "right axioms" to prove the Pappus theorem? The answer is no less than the Pappus theorem itself, thanks to what the Pappus theorem implies; namely, that the abstract plane of "points" and "lines" can be given *coordinates* which form a *field* as defined in section 4.3. Thus geometry springs fully armed from the Pappus theorem! The Pappus axiom (as we should now call it) is the right axiom to prove coordinatization by a field, because such a coordinatization allows us to prove the Pappus axiom. As remarked above, this follows by using linear equations to define lines. The field properties then enable us to find intersections of lines by solving linear equations, and to verify that points of intersection lie on a line.

The idea of reversing the coordinate approach to geometry began with von Staudt (1847), who used the Pappus axiom to define addition and multiplication of points on a line. Hilbert (1899) extended this idea to prove that the coordinates form a field, but he had to assume another projective axiom, the so-called *theorem of Desargues*, which was discovered around 1640. (Roughly speaking, the Pappus axiom easily implies that addition and multiplication are commutative, while the Desargues axiom easily implies that they are associative.) Quite remarkably, considering how long the Pappus and Desargues theorems had been around, Hessenberg (1905) discovered that the Pappus theorem implies the Desargues theorem. So the single Pappus axiom is in fact *equivalent* to coordinatization of the plane by a field.

This reversal of the Pappus theorem also tells us something remarkable about algebra: the nine field axioms follow from four geometric axioms—the three projective plane axioms plus Pappus!

( ) d)-

)-

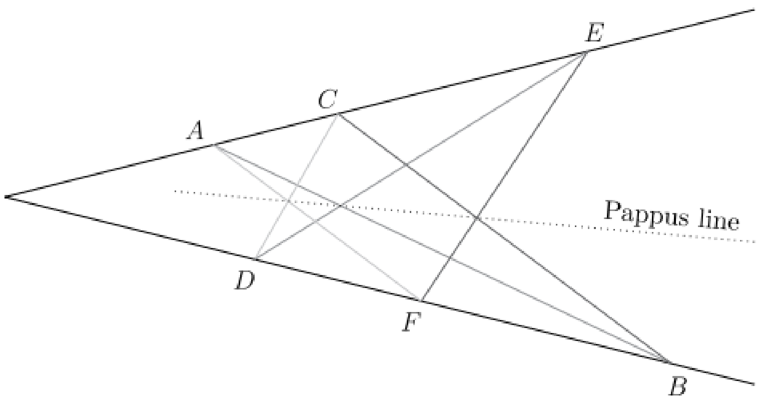

. - / / #)44 P-)

)44 E "8 &8 8 !8 @8 .) 4 4) + -
)) + 8 4) "&
) !@8 &) @. 8 !) . "8)

)44)) @) 4 - 4
- 8)))) 4 -) =) P
&)% 4 - + 4 G)44 4) H
) 4 8 8) 9 4 4 #
)44 4) -)44 4 J % - +
4) 8) - +)4 9)% 4
)) 4) -) -) - E 4 J %
- +8 P-) 4)))
) 9 4 J J)) -
-)) - 8 9) -) 8)
9 4 4
E '); 4 J % 4) - +8 -
) +
" + 4 9 -) =
" +)%) = 4
) 4 8))
P);) @ A #)- @
) 4)) 8 9 4 J % - +) 8 9)
% 4))) 9 4 J) + U G F H
#);)) +)%) G4) 8H)
+) 1 % 8 4);) % +)'8))
9) + P 4 %)44 # +
8 % 8))) 9) +);)9
4)) 9)
D)) G -); H 4 %)44 I #
)))44 8) ')
)44 4 \ ) +8))9) 4) G4 H)
G H) 9 - %)) P) P
6 # - + 4 - +))44 R #
)44); 2)) 3 -); 4 %
) F) 9+) P 8 9))) F))
4 %)44); ") ')9 % 8 9+ -

) =) P # P 4 4)9
P 9+ % -) =) 8) % +)
4)

) % -))44) - + 9 -)
%) 2 T6738)44); P)
) 4) 4) 1 9 2 T<<3 ;
) 4 %))) P 8 9))
) 4 J %); 8) !) - 8)
%) N6 2(- + 4)' -8)44);) +
4))) 4))) % 8
!) -);) + 4) +))) % 3 f) '
)9 +8 - -)44) !) -)
9) 8 1 9 - 2 < /3 %))44
4 !) - -)44);)
= %)) F) 4) 9+) P
%))44) -) '
)9)9) - 9) P); -
); U 4 J % 4)); 4)44 R

in the geometric series $\frac{1}{2}+\frac{1}{2^2}+\frac{1}{2^3}+\cdots$ (see the example in the next subsection). But it is harder to prove it is valid to differentiate term by term and hence to prove that $\frac{d}{dx}\exp(x)=\exp(x)$. This involves the advanced version of the continuity concept called *uniform* continuity, which is normally reserved for advanced calculus courses.

The Irrationality of e

The series $e=1+\frac{1}{1!}+\frac{1}{2!}+\frac{1}{3!}+\frac{1}{4!}+\cdots$ is good for computation, because the terms decrease rapidly in size. The rapid decrease is also of theoretical value, because it gives a proof that e is irrational (due to Joseph Fourier around 1815).

Irrationality of ***e*****.** *The number e is not equal to m/n for any positive integers m and n.*

Proof. Suppose on the contrary that

$$\frac{m}{n}=1+\frac{1}{1!}+\frac{1}{2!}+\frac{1}{3!}+\cdots+\frac{1}{n!}+\cdots .$$

If we multiply both sides of this equation by $n!$ we get

$$\begin{aligned} m\cdot(n-1)! &= n!+\frac{n!}{1!}+\frac{n!}{2!}+\frac{n!}{3!}+\cdots+\frac{n!}{n!} \\ &\quad +\frac{1}{n+1}+\frac{1}{(n+1)(n+2)}+\frac{1}{(n+1)(n+2)(n+3)}+\cdots \\ &= \text{integer}+\frac{1}{n+1}+\frac{1}{(n+1)(n+2)} \\ &\quad +\frac{1}{(n+1)(n+2)(n+3)}+\cdots . \end{aligned}$$

The left-hand side of the equation is also an integer, so we conclude that

$$\frac{1}{n+1}+\frac{1}{(n+1)(n+2)}+\frac{1}{(n+1)(n+2)(n+3)}+\cdots \quad \text{is an integer.}$$

But clearly $n \geq 1$ (as the denominator of a fraction), so

$$\frac{1}{n+1}+\frac{1}{(n+1)(n+2)}+\frac{1}{(n+1)(n+2)(n+3)}+\cdots$$

$$<\frac{1}{2}+\frac{1}{2^2}+\frac{1}{2^3}+\cdots=1.$$

So $\frac{1}{n+1}+\frac{1}{(n+1)(n+2)}+\frac{1}{(n+1)(n+2)(n+3)}+\cdots$ is *not* an integer, and we have a contradiction. Thus it is false to assume that $e = m/n$. □

6.8 *The Inverse Tangent Function and π

Calculus is needed for geometry at quite an elementary level, to understand the circular functions sine, cosine, and tangent, and to evaluate the number π. To illustrate how calculus contributes to geometry, we are going to derive the simplest possible expression for π, namely,

$$\frac{\pi}{4}=1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\cdots\ .$$

This expression is a consequence of the power series for the inverse tangent function,

$$\arctan x = x-\frac{x^3}{3}+\frac{x^5}{5}-\frac{x^7}{7}+\cdots\ ,$$

which we derive first.

We view $y = \tan\theta$ as the vertical side of a right-angled triangle with horizontal side 1 and angle θ as shown in figure 6.14. Thus $\tan\theta = y$ is measured along the vertical tangent to the unit circle. Its inverse function $\theta = \arctan y$ is so called because θ is also the arc length on the unit circle spanned by the angle θ.

We now calculate $\frac{d}{dy}\arctan y = \frac{d\theta}{dy}$ by estimating the ratio $\frac{\Delta\theta}{\Delta y}$ when $\Delta\theta$ is a small addition to the angle θ, causing a small addition Δy to the height y on the vertical tangent (figure 6.15).

By the Pythagorean theorem, $OB = \sqrt{1+y^2}$, so the arc length $A'B$ subtended by the angle $\Delta\theta$ on the circle of radius $\sqrt{1+y^2}$ is

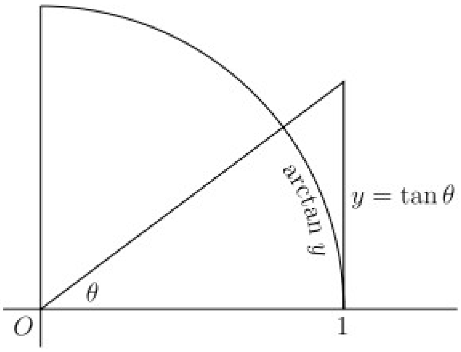
arctan y
$y = \tan\theta$
θ
O
1

Figure 6.14: Geometry of tan and arctan.

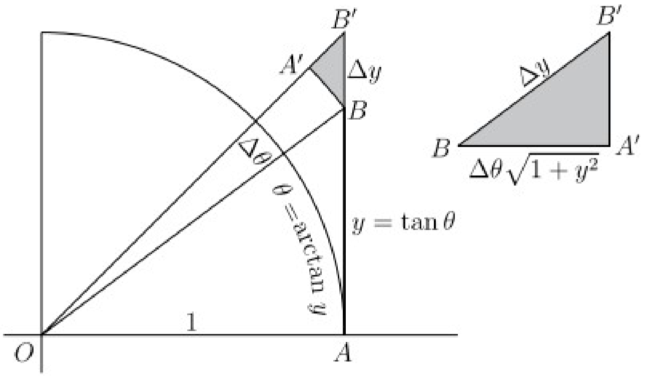
B'
A'
Δy
B
$\Delta\theta$
$\theta = $arctan y
$y = \tan\theta$
1
O
A
B'
Δy
B
A'
$\Delta\theta\sqrt{1+y^2}$

Figure 6.15: Estimating the derivative of arctan y.

$\Delta\theta\sqrt{1+y^2}$. As $\Delta\theta \to 0$ this arc approaches a straight line segment and $BA'B'$ approaches a right-angled triangle similar to OAB. Consequently, the ratio of the sides Δy and $\Delta\theta\sqrt{1+y^2}$ of $BA'B'$ (corresponding to hypotenuse and base) approaches $\sqrt{1+y^2}$, so

$$\frac{\Delta\theta}{\Delta y} \to \frac{1}{1+y^2} \quad \text{as} \quad \Delta\theta \to 0.$$

In other words,

$$\frac{d}{dy}\arctan y = \frac{d\theta}{dy} = \frac{1}{1+y^2}.$$

)-

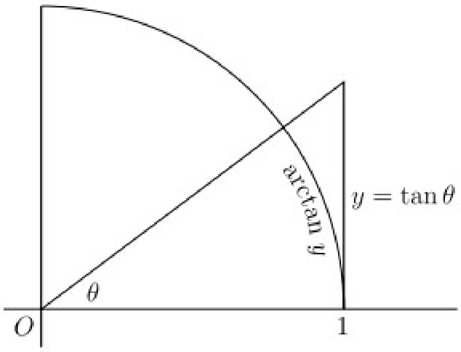

arctan y
$y = \tan\theta$
θ
O
1

. - N 6 * +))))

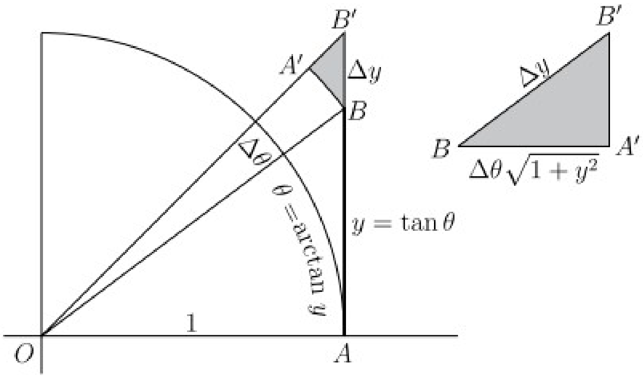

B′
A′
Δy
B
Δθ
θ = arctan y
y = tan θ
1
O
A
B′
Δy
B
A′
$\Delta\theta\sqrt{1+y^2}$

. - N / @) - %) %)) +

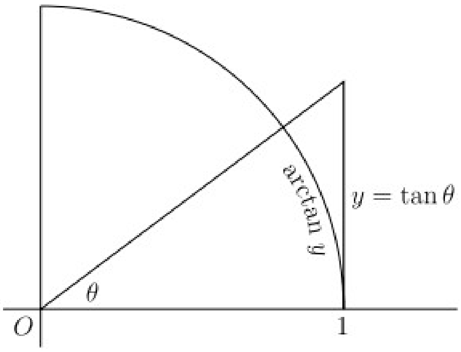
arctan y
$y = \tan\theta$
θ
O
1

Figure 6.14: Geometry of tan and arctan.

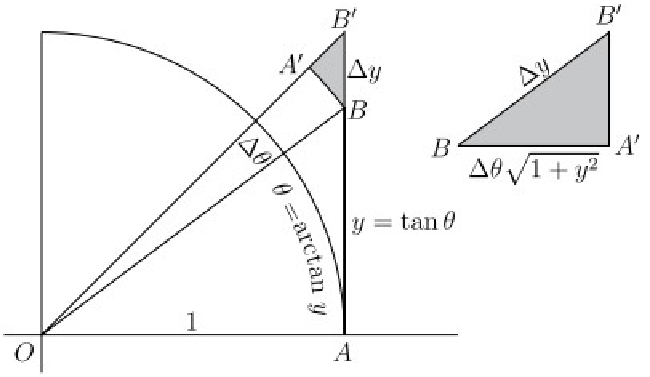
B'
A'
Δy
B
$\Delta\theta$
θ =arctan y
$y = \tan\theta$
1
O
A
B'
Δy
B
A'
$\Delta\theta\sqrt{1+y^2}$

Figure 6.15: Estimating the derivative of arctan y.

$\Delta\theta\sqrt{1+y^2}$. As $\Delta\theta \to 0$ this arc approaches a straight line segment and $BA'B'$ approaches a right-angled triangle similar to OAB. Consequently, the ratio of the sides Δy and $\Delta\theta\sqrt{1+y^2}$ of $BA'B'$ (corresponding to hypotenuse and base) approaches $\sqrt{1+y^2}$, so

$$\frac{\Delta\theta}{\Delta y} \to \frac{1}{1+y^2} \quad \text{as} \quad \Delta\theta \to 0.$$

In other words,

$$\frac{d}{dy}\arctan y = \frac{d\theta}{dy} = \frac{1}{1+y^2}.$$

It follows, integrating both sides, using $\arctan 0 = 0$ and the zero derivative theorem from section 6.3, that

$$\arctan x = \int_0^x \frac{dy}{1+y^2}.$$

Now we know from the theory of geometric series in section 6.1 that

$$\frac{1}{1+y^2} = 1 - y^2 + y^4 - y^6 + \cdots \quad \text{for} \quad |y| < 1.$$

Using this formula in the integral we get

$$\begin{aligned} \arctan x &= \int_0^x \left(1 - y^2 + y^4 - y^6 + \cdots\right)\, dy \quad \text{for} \quad |x| < 1 \\ &= \int_0^x 1\, dy - \int_0^x y^2\, dy + \int_0^x y^4\, dy - \int_0^x y^6\, dy + \cdots \\ &= x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots\ . \end{aligned}$$

To justify the first step—from the integral of an infinite sum to the infinite sum of integrals—we proceed as we did in section 6.7, using the *finite* series

$$\frac{1}{1+y^2} = 1 - y^2 + y^4 - \cdots \pm y^{2n} \mp \frac{y^{2n+2}}{1+y^2}$$

and showing that $\int_0^x \frac{y^{2n+2}}{1+y^2}\, dy \to 0$ as n increases.

This more rigorous argument has another advantage: *the finite series is also valid for* $y = 1$. So we can conclude that

$$\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots$$

is also valid for $x = 1$, which gives the wonderful formula

$$\frac{\pi}{4} = \arctan 1 = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots\ .$$

6.9 Elementary Functions

We began this chapter by looking at functions we can differentiate: starting with constant and identity functions and quickly reaching all rational functions, by rules for differentiating $+$, $-$, $\cdot$, $\div$ of functions whose derivatives are known. The next two rules, for inverse functions and composite functions, allow us to differentiate many algebraic functions, such as $\sqrt{x}$, $\sqrt{1+x^2}$, $\sqrt[3]{1+x^4}$, $\ldots$, and so on. Finally, the fundamental theorem of calculus,

$$\frac{d}{dx}\int_a^x f(t)\,dt = f(x),$$

allows us to differentiate functions defined by integrals, such as

$$\ln x = \int_1^x \frac{dt}{t} \quad \text{and} \quad \arctan x = \int_0^x \frac{dt}{1+t^2}.$$

The functions obtainable from the latter two functions and rational functions by inversion, composition, and the rational operations $+$, $-$, $\cdot$, $\div$ are called the *elementary functions*. It may seem rather arbitrary to stop at the ln and arctan integrals—since many other functions are definable as integrals of algebraic functions—but there are good reasons to consider this the right place for elementary calculus to stop.

The main reason is that, assuming the fundamental theorem of algebra, the *integrals of all rational functions reduce to integrals of the* ln *or* arctan *type*. To see why, recall that any rational function of t can be written as a quotient of polynomials, $p(t)/q(t)$. It follows from the fundamental theorem of algebra that $q(t)$ is a product of real linear or quadratic factors

$$at + b \quad \text{or} \quad ct^2 + dt + e.$$

Then we can apply the algebraic technique of *partial fractions* (for an example, see section 7.3) to write $p(x)/q(x)$ as a sum of terms

$$\frac{A}{at+b} \quad \text{and} \quad \frac{Bt+C}{ct^2+dt+e}.$$

Some further easy transformations then reduce the integral of $p(t)/q(t)$ to integrals of $\frac{1}{t}$ and $\frac{1}{1+t^2}$; that is, to integrals of the ln or arctan type. Thus, *the integral of any rational function is an elementary function.*

This result shows that the elementary functions are not as arbitrary as they look. To reinforce this claim, we look at the elementary functions obtainable from the arctan function. Obviously, they include its inverse function, $\tan\theta$, but they also include the other circular functions, such as $\cos\theta$ and $\sin\theta$. We devote the next subsection to deriving the connection between the sine, cosine, and tangent functions, since it also has a beautiful connection with number theory and geometry.

Rational Points on the Circle

Ever since the discovery of the Pythagorean theorem there has been interest in *Pythagorean triples*: positive integer triples (a, b, c) such that $a^2 + b^2 = c^2$. The simplest Pythagorean triple is (3,4,5), and others are (5,12,13), (7,24,25), and (8,15,17). Specific Pythagorean triples have been known since ancient times in Europe, the Middle East, India, and China. Around 300 BCE, Euclid (*Elements*, Book X, lemma following Proposition 28) found a formula that gives them all:

$$a = (p^2 - q^2)r, \quad b = 2pqr, \quad c = (p^2 + q^2)r,$$

as p, q, r run through all positive integers.

A few centuries later, Diophantus transformed the problem of finding these triples into one about rational numbers, which can be viewed as finding *rational points on the unit circle*. This version of the problem arises because if $a^2 + b^2 = c^2$ then

$$\left(\frac{a}{c}\right)^2 + \left(\frac{b}{c}\right)^2 = 1,$$

so (a, b, c) corresponds to a rational point $\left(\frac{a}{c}, \frac{b}{c}\right)$ on the circle $x^2 + y^2 = 1$. This view of Pythagorean triples is revealing, because the search for them can now be guided by geometry and algebra. See figure 6.16.

If P has rational coordinates, then the slope t from Q to P is a rational number (because the "rise" and "run" are rational). Conversely, if the line PQ has rational slope t we can find P as the common solution

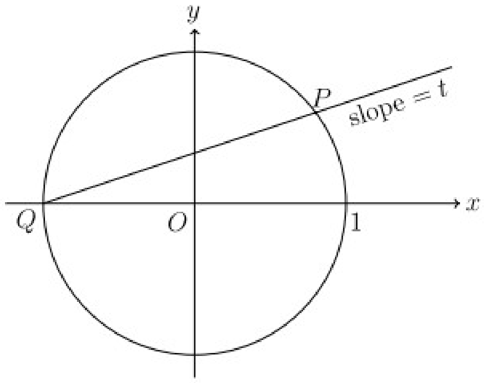

Figure 6.16: A rational point P on the unit circle.

of the linear equation

$$y = t(x+1) \qquad \text{(the equation of } PQ\text{)}$$

and the quadratic equation

$$x^2 + y^2 = 1 \qquad \text{(the equation of the unit circle).}$$

Since the coefficients of these two equations are rational, their common solutions are given by a quadratic equation in x with rational coefficients. We know that one of the solutions is $x = -1$ (for Q), so the second solution (giving the x-coordinate of P) must also be rational. We now work out exactly what the second solution is.

Substituting $y = t(x+1)$ in $x^2 + y^2 = 1$ gives the quadratic equation

$$x^2 + t^2(x+1)^2 = 1, \quad \text{that is,} \quad (1+t^2)x^2 + 2t^2x + t^2 - 1 = 0.$$

The quadratic formula gives the solutions

$$\begin{aligned} x &= \frac{-2t^2 \pm \sqrt{(2t^2)^2 - 4(1+t^2)(t^2-1)}}{2(1+t^2)} \\ &= \frac{-2t^2 \pm \sqrt{4t^4 - 4(t^4-1)}}{2(1+t^2)} \end{aligned}$$

$$= \frac{-2t^2 \pm 2}{2(1+t^2)}$$

$$= -1,\ \frac{1-t^2}{1+t^2}.$$

This confirms the solution $x = -1$ for Q, and gives $x = \frac{1-t^2}{1+t^2}$ as the x-coordinate of P. It follows from the equation of PQ that the y-coordinate of P is

$$y = t(x+1) = t\left(\frac{1-t^2+1+t^2}{1+t^2}\right) = \frac{2t}{1+t^2}.$$

Thus, *the rational points on* $x^2 + y^2 = 1$ *are* $(-1, 0)$ *and the points* $\left(\frac{1-t^2}{1+t^2}, \frac{2t}{1+t^2}\right)$ *as t runs through the rational numbers.*

For example, $t = \frac{1}{2}$ gives the point $\left(\frac{3}{5}, \frac{4}{5}\right)$ corresponding to the Pythagorean triple (3,4,5).

If we drop the restriction that t be rational, this makes no difference to the calculation of the coordinates of P,

$$x = \frac{1-t^2}{1+t^2}, \quad y = \frac{2t}{1+t^2}.$$

But now these formulas tell something interesting about the sine, cosine, and tangent functions, due to their association with the circle.

Sine and cosine formulas. *If* $t = \tan\frac{\theta}{2}$, *then*

$$\cos\theta = \frac{1-t^2}{1+t^2}, \quad \sin\theta = \frac{2t}{1+t^2}.$$

Proof. These formulas can be read off figure 6.17, once it is established that the angles are as shown. If we let θ be the angle between the line OP and the x-axis, then it is certainly true that $P = (\cos\theta, \sin\theta)$, by definition of the sine and cosine functions.

Also, it can be checked that $\theta/2$ is the angle of the line QP, using the facts that OPQ is an isosceles triangle, hence with equal angles at P and Q, and that the angle sum of a triangle is π.

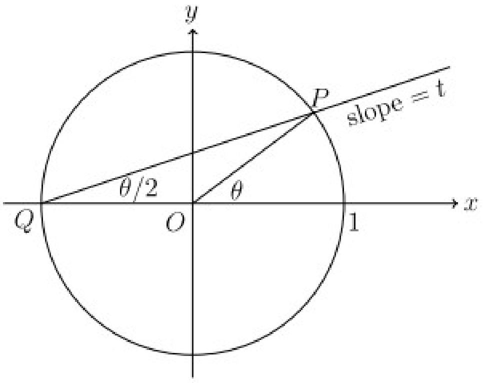

Figure 6.17: Angles in the circle.

Thus the slope t of PQ is $\tan \frac{\theta}{2}$, and hence

$$\left(\frac{1-t^2}{1+t^2}, \frac{2t}{1+t^2}\right) = (\cos\theta, \sin\theta),$$

which gives $\cos\theta = \frac{1-t^2}{1+t^2}$ and $\sin\theta = \frac{2t}{1+t^2}$. □

6.10 Historical Remarks

> Often I have considered the fact that most of the difficulties which block the progress of students trying to learn analysis stem from this: that although they understand little of ordinary algebra, still they attempt this more difficult art. From this it follows that they remain on the fringes, but in addition they entertain strange ideas about the concept of the infinite, which they must try to use.
>
> Euler (1748a), preface

Foundations of Calculus

Most of calculus as we know it, at the undergraduate level, was produced in a frenzy of activity in the late seventeenth and early eighteenth century. Both before and after that period there was concern about the

E)-

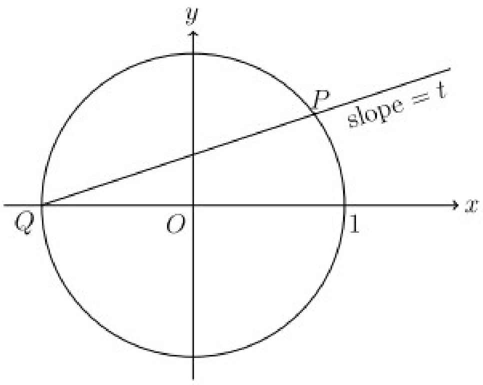
y
P
slope = t
Q
O
1
x

. -   N N ")   ) 4

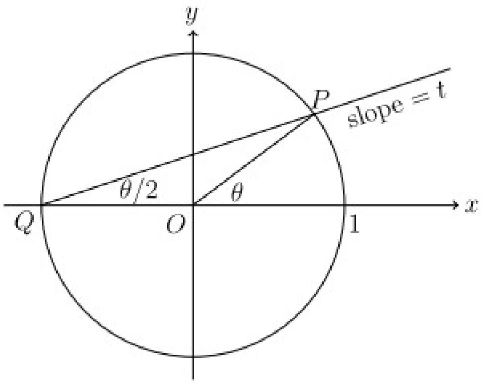
y
P
slope = t
θ/2
θ
Q
O
1
x

. -   N 7 " -

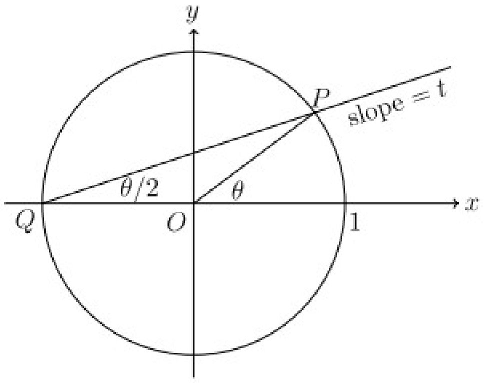

Figure 6.17: Angles in the circle.

Thus the slope t of PQ is $\tan\frac{\theta}{2}$, and hence

$$\left(\frac{1-t^2}{1+t^2}, \frac{2t}{1+t^2}\right) = (\cos\theta, \sin\theta),$$

which gives $\cos\theta = \frac{1-t^2}{1+t^2}$ and $\sin\theta = \frac{2t}{1+t^2}$. □

6.10 Historical Remarks

> Often I have considered the fact that most of the difficulties which block the progress of students trying to learn analysis stem from this: that although they understand little of ordinary algebra, still they attempt this more difficult art. From this it follows that they remain on the fringes, but in addition they entertain strange ideas about the concept of the infinite, which they must try to use.
>
> Euler (1748a), preface

Foundations of Calculus

Most of calculus as we know it, at the undergraduate level, was produced in a frenzy of activity in the late seventeenth and early eighteenth century. Both before and after that period there was concern about the infinite processes that underlie the calculus—the foundations of calculus that we now call *analysis*. Infinite processes were already a concern in ancient Greece, and the Greeks devised the method of exhaustion to avoid infinity as far as possible. (Certainly, it avoids *actual* infinity.) As we saw in section 6.1, Archimedes used exhaustion to determine the area of the parabolic segment. It allowed him to conclude, using an "arbitrarily large" but not infinite number of triangles, that the area of the parabolic segment is 4/3 the area of the largest triangle.

The seventeenth-century mathematicians were well-schooled in the method of exhaustion, from Euclid and Archimedes, so they assumed they could fall back on it if necessary. But as Huygens (1659), p. 337, remarked:

> Mathematicians will never have enough time to read all the discoveries in Geometry (a quantity which is increasing from day to day and seems likely in this scientific age to develop to enormous proportions) if they continue to be presented in a rigorous form according to the manner of the ancients.

The idea that calculus could *in principle* be presented "in the manner of the ancients" might have been true in 1659, when mathematicians were still investigating curves individually. But after Newton and Leibniz made calculus truly a method of *calculation*, applicable to seemingly arbitrary formulas, the question of rigor became more acute.

When calculus became this general it was natural to ask: What is a function? What is a continuous function? Is it the same as a differentiable function? Newton, who thought of functions in terms of continuous motion, seemed to assume that continuous functions are both integrable and differentiable. As we now define continuity, continuous functions are integrable (on closed intervals) but some of them are *nowhere* differentiable. These results could not be known until the nineteenth century, when the concept of continuous function was first defined. The foundations of the definition were laid by Bolzano (1817) and Cauchy (1821), but the properties of continuous functions were not rigorously provable until Dedekind gave a definition of the continuum $\mathbb{R}$ of real numbers in 1858 (published in 1872). This started the push to arithmetize geometry and analysis, mentioned in section 5.10, after the discovery of non-Euclidean geometry.

With a definition of $\mathbb{R}$ finally available (Cantor, Meray, and Weierstrass published equivalent definitions of $\mathbb{R}$ about the same time as Dedekind), the calculus was put on a sound foundation by Weierstrass in the 1870s. It can still be asked (and perhaps can never be answered) whether the nineteenth-century foundations of calculus are the best possible, but most of the alternatives proposed so far give the same results, and are not noticeably simpler.

Preparing Students for Calculus

As we have seen, the foundations of calculus involve advanced ideas, particularly about infinity. This was realized long before the full extent of the problem became known, and mathematicians tried to devise appropriate and interesting material to *prepare* students for calculus, without exposing them to all the challenges of infinity. The most brilliant and fascinating book along these lines was the *Introductio in analysin infinitorum* (Introduction to the analysis of the infinite) of Euler (1748a).

In a nutshell, Euler believed that the best preparation for calculus was the study of infinite series. He had good reasons for believing this. As we have seen, infinite geometric series were implicitly used by Euclid and Archimedes, so infinite series were in use long before calculus. Another pointer from history was the contribution of Oresme (1350). He made the important discovery that the so-called *harmonic series*,

$$\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\cdots,$$

does not have a finite sum, even though its terms tend to zero. His proof was by collecting the terms into the following groups,

$$\begin{aligned}&\frac{1}{2}+\left(\frac{1}{3}+\frac{1}{4}\right)+\left(\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}\right)\\ &+\left(\frac{1}{9}+\frac{1}{10}+\frac{1}{11}+\frac{1}{12}+\frac{1}{13}+\frac{1}{14}+\frac{1}{15}+\frac{1}{16}\right)+\cdots,\end{aligned}$$

each of which has twice as many terms as the group before. It is easy to see that

$$\frac{1}{3}+\frac{1}{4}>\frac{2}{4}=\frac{1}{2},$$

$$\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}>\frac{4}{8}=\frac{1}{2},$$

$$\frac{1}{9}+\frac{1}{10}+\frac{1}{11}+\frac{1}{12}+\frac{1}{13}+\frac{1}{14}+\frac{1}{15}+\frac{1}{16}>\frac{8}{16}=\frac{1}{2},$$

and so on. Hence each of the infinitely many groups has sum $\geq \frac{1}{2}$, and so the sum of the series grows beyond all bounds. This discovery of Oresme was the first sign that sums of infinite series have some subtleties of their own.

Unknown to Euler, and to anyone else in Europe until the nineteenth century, was a collection of results giving a really spectacular demonstration that infinite series can be studied before calculus. These results, from the Indian mathematician Madhava, who lived around 1350–1425, include

$$\sin x = x-\frac{x^3}{3!}+\frac{x^5}{5!}-\frac{x^7}{7!}+\cdots\ ,$$

$$\cos x = 1-\frac{x^2}{2!}+\frac{x^4}{4!}-\frac{x^6}{6!}+\cdots\ ,$$

$$\tan^{-1} x = x-\frac{x^3}{3}+\frac{x^5}{5}-\frac{x^7}{7}+\cdots\ ,$$

the last of which gives a famous formula for π by substituting $x=1$:

$$\frac{\pi}{4}=1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\cdots\ .$$

For more on these results, whose discovery by Indian mathematicians was little known until recently, see Plofker (2009).

Finally, just before the dawn of calculus, Wallis (1655) discovered that

$$\frac{\pi}{2}=\frac{2\cdot 2}{1\cdot 3}\cdot\frac{4\cdot 4}{3\cdot 5}\cdot\frac{6\cdot 6}{5\cdot 7}\cdot\frac{8\cdot 8}{7\cdot 9}\cdot\cdots\ ,$$

and his colleague Lord Brouncker discovered the mysteriously similar result

$$\frac{\pi}{4} = \cfrac{1}{1 + \cfrac{1^2}{2 + \cfrac{3^2}{2 + \cfrac{5^2}{2 + \cfrac{7^2}{2 + \cfrac{9^2}{\ddots}}}}}}$$

One of Euler's achievements was to explain how the three formulas for π are related. The relation between the series and the continued fraction is explained in section 10.9.

Thus Euler had good reason to think that infinite series, and a few similar infinite processes, are a rich field that can be explored before embarking upon calculus. All of the above results actually appear in his book, together with many of Euler's own. Indeed, Euler is now considered to be the most brilliant and creative virtuoso of infinite series who ever lived, so it is no surprise that his *Introductio* goes well beyond the needs of a prospective calculus student. Nevertheless, we can all marvel at what he did.

For example, by comparing the power series for sine, cosine, and the exponential function, he discovered the formula

$$e^{i\theta} = \cos\theta + i\sin\theta,$$

with its miraculous special case

$$e^{i\pi} = -1.$$

And by multiplying geometric series together he discovered that, for any $s > 1$,

$$\left(\frac{1}{1-2^{-s}}\right)\left(\frac{1}{1-3^{-s}}\right)\left(\frac{1}{1-5^{-s}}\right)\cdots\left(\frac{1}{1-p^{-s}}\right)\cdots$$
$$= \frac{1}{1^s} + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \frac{1}{5^s} + \cdots,$$

where p runs through all the prime numbers. The exponent $s > 1$ is there to ensure that the series on the right-hand side (the so-called *zeta* function of s) has a finite sum. For $s = 1$ the sum is infinite by Oresme's result, but this fact too is something we can exploit. If there are only finitely many primes, then the left side is *finite*, so we have a contradiction. Therefore, we have a new proof that there are infinitely many primes!

This result is just the first of a cornucopia of results about primes that fall out of Euler's formula.

The Geometric Series in Hardy's Pure Mathematics

> I have indeed in an examination asked a dozen candidates, including several future Senior Wranglers, to sum the series $1 + x + x^2 + \cdots$, and have not received a single answer that was not practically worthless—and this from men quite capable of solving difficult problems connected with the curvature and torsion of twisted curves.
>
> Hardy (1908), p. vi

The humble geometric series, which we have used as a foundation for much of this chapter, itself depends on a fundamental fact about limits: that $x^n \to 0$ as $n \to \infty$ when $|x| < 0$. For beginners, this fact seems obvious enough, and it was assumed in landmark works on the foundations of calculus, such as Cauchy (1821) and Jordan (1887). However, Hardy (1908), in his famous *Course of Pure Mathematics*, thought it worthwhile to probe more deeply, because his hope for his students was that

> accurate thought in connexion with these matters will become an integral part of their ordinary mathematical habit of mind. It is this conviction that has led me to devote so much space to the most elementary ideas of all connected with limits, to be purposely diffuse about fundamental points, to illustrate them by so elaborate a system of examples, and to write a chapter of fifty pages without advancing beyond the ordinary geometrical series.
>
> Hardy (1908), p. vii

!"#' $ % % & ' ( )) * +) , -)
.) @)))) 4 + 2 !.! % 3 4
*) /68 ( ) 7T8 1 9 / 8 % NT
5- /N2T 3

EXAMPLE 1.8.2

Electromagnetic Wave Equations One important application of this vector relation [Eq. (1.96)] is in the derivation of the electromagnetic wave equation. In vacuum Maxwell's equations become

$$\nabla \cdot \mathbf{B} = 0, \tag{1.97a}$$

$$\nabla \cdot \mathbf{E} = 0, \tag{1.97b}$$

$$\nabla \times \mathbf{B} = \varepsilon_0 \mu_0 \frac{\partial \mathbf{E}}{\partial t} = \frac{1}{c^2}\frac{\partial \mathbf{E}}{\partial t}, \tag{1.97c}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \tag{1.97d}$$

where **E** is the electric field, **B** the magnetic induction, ε_0 the electric permittivity, and μ_0 the magnetic permeability (SI units), so that $\varepsilon_0\mu_0 = 1/c^2$, where c is the velocity of light. This relation has important consequences. Because ε_0, μ_0 can be measured in any frame, the velocity of light is the same in any frame.

Suppose we eliminate **B** from Eqs. (1.97c) and (1.97d). We may do this by taking the curl of both sides of Eq. (1.97d) and the time derivative of both sides of Eq. (1.97c). Since the space and time derivatives commute,

$$\frac{\partial}{\partial t}\nabla \times \mathbf{B} = \nabla \times \frac{\partial \mathbf{B}}{\partial t},$$

and we obtain

$$\nabla \times (\nabla \times \mathbf{E}) = -\frac{1}{c^2}\frac{\partial^2 \mathbf{E}}{\partial t^2}.$$

Application of Eqs. (1.96) and (1.97b) yields

$$(\nabla \cdot \nabla)\mathbf{E} = \frac{1}{c^2}\frac{\partial^2 \mathbf{E}}{\partial t^2}, \tag{1.98}$$

the electromagnetic vector wave equation. Again, if **E** is expressed in Cartesian coordinates, Eq. (1.98) separates into three scalar wave equations, each involving a scalar Laplacian.

When external electric charge and current densities are kept as driving terms in Maxwell's equations, similar wave equations are valid for the electric potential and the vector potential. To show this, we solve Eq. (1.97a) by writing $\mathbf{B} = \nabla \times \mathbf{A}$ as a curl of the vector potential. This expression is substituted into Faraday's induction law in differential form [Eq. (1.97d)] to yield $\nabla \times (\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t}) = 0$. The vanishing curl implies that $\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t}$ is a gradient and therefore can be written as $-\nabla\varphi$, where $\varphi(\mathbf{r}, t)$ is defined as the (nonstatic) electric potential. These results

$$\mathbf{B} = \nabla \times \mathbf{A}, \qquad \mathbf{E} = -\nabla\varphi - \frac{\partial \mathbf{A}}{\partial t} \tag{1.99}$$

for the **B** and **E** fields solve the homogeneous Maxwell's equations.

We now show that the inhomogeneous Maxwell's equations,

$$\text{Gauss's law: } \boldsymbol{\nabla}\cdot\mathbf{E} = \rho/\varepsilon_0;$$
$$\text{Oersted's law: } \boldsymbol{\nabla}\times\mathbf{B} - \frac{1}{c^2}\frac{\partial \mathbf{E}}{\partial t} = \mu_0\mathbf{J} \qquad (1.100)$$

in differential form lead to wave equations for the potentials φ and $\mathbf{A}$, provided that $\boldsymbol{\nabla}\cdot\mathbf{A}$ is determined by the constraint $\frac{1}{c^2}\frac{\partial\varphi}{\partial t} + \boldsymbol{\nabla}\cdot\mathbf{A} = 0$. This choice of fixing the divergence of the vector potential is called the **Lorentz gauge** and serves to uncouple the partial differential equations of both potentials. This gauge constraint is not a restriction; it has no physical effect.

Substituting our electric field solution into Gauss's law yields

$$\frac{\rho}{\varepsilon_0} = \boldsymbol{\nabla}\cdot\mathbf{E} = -\boldsymbol{\nabla}^2\varphi - \frac{\partial}{\partial t}\boldsymbol{\nabla}\cdot\mathbf{A} = -\boldsymbol{\nabla}^2\varphi + \frac{1}{c^2}\frac{\partial^2\varphi}{\partial t^2},$$

the wave equation for the electric potential. In the last step, we used the Lorentz gauge to replace the divergence of the vector potential by the time derivative of the electric potential and thus decouple φ from $\mathbf{A}$.

Finally, we substitute $\mathbf{B} = \boldsymbol{\nabla}\times\mathbf{A}$ into Oersted's law and use Eq. (1.96), which expands $\boldsymbol{\nabla}^2$ in terms of a longitudinal (the gradient term) and a transverse component (the curl term). This yields

$$\mu_0\mathbf{J} + \frac{1}{c^2}\frac{\partial\mathbf{E}}{\partial t} = \boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\mathbf{A}) = \boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\mathbf{A}) - \boldsymbol{\nabla}^2\mathbf{A} = \mu_0\mathbf{J} - \frac{1}{c^2}\left(\boldsymbol{\nabla}\frac{\partial\varphi}{\partial t} + \frac{\partial^2\mathbf{A}}{\partial t^2}\right),$$

where we have used the electric field solution [Eq. (1.99)] in the last step. Now we see that the Lorentz gauge condition eliminates the gradient terms so that the wave equation

$$\frac{1}{c^2}\frac{\partial^2\mathbf{A}}{\partial t^2} - \boldsymbol{\nabla}^2\mathbf{A} = \mu_0\mathbf{J}$$

for the vector potential remains.

Finally, looking back at Oersted's law, taking the divergence of Eq. (1.100), dropping $\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\times\mathbf{B}) = 0$ and substituting Gauss's law for $\boldsymbol{\nabla}\cdot\mathbf{E} = \rho/\epsilon_0$, we find $\mu_0\boldsymbol{\nabla}\cdot\mathbf{J} = -\frac{1}{\epsilon_0 c^2}\frac{\partial\rho}{\partial t}$, where $\epsilon_0\mu_0 = 1/c^2$, that is, the continuity equation for the current density. This step justifies the inclusion of Maxwell's displacement current in the generalization of Oersted's law to nonstationary situations. ■

EXERCISES

1.8.1 Verify Eq. (1.96)

$$\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\mathbf{V}) = \boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\mathbf{V}) - (\boldsymbol{\nabla}\cdot\boldsymbol{\nabla})\mathbf{V}$$

by direct expansion in Cartesian coordinates. If symbolic software is available, check the identity for typical fields, such as $\mathbf{V} = \mathbf{r}, \mathbf{r}/r^3, \mathbf{a}\cdot\mathbf{r}\mathbf{b}, \mathbf{a}\times\mathbf{r}$.

1.8.2 Show that the identity

$$\nabla \times (\nabla \times \mathbf{V}) = \nabla(\nabla \cdot \mathbf{V}) - (\nabla \cdot \nabla)\mathbf{V}$$

follows from the *BAC–CAB* rule for a triple vector product. Justify any alteration of the order of factors in the *BAC* and *CAB* terms.

1.8.3 Prove that $\nabla \times (\varphi \nabla \varphi) = 0$.

1.8.4 Prove that $(\nabla u) \times (\nabla v)$ is solenoidal, where u and v are differentiable scalar functions. Start by formulating the problem as a mathematical equation.

1.8.5 φ is a scalar satisfying Laplace's equation, $\nabla^2 \varphi = 0$. Show that $\nabla \varphi$ is **both** solenoidal and irrotational.

1.8.6 With ψ a scalar function, show that

$$(\mathbf{r} \times \nabla) \cdot (\mathbf{r} \times \nabla)\psi = r^2 \nabla^2 \psi - r^2 \frac{\partial^2 \psi}{\partial r^2} - 2r \frac{\partial \psi}{\partial r}.$$

(This can actually be shown more easily in spherical polar coordinates; see Section 2.5.)

1.8.7 In the Pauli theory of the electron one encounters the expression

$$(\mathbf{p} - e\mathbf{A}) \times (\mathbf{p} - e\mathbf{A})\psi,$$

where ψ is a scalar function. $\mathbf{A}$ is the magnetic vector potential related to the magnetic induction $\mathbf{B}$ by $\mathbf{B} = \nabla \times \mathbf{A}$. Given that $\mathbf{p} = -i\nabla$, show that this expression reduces to $ie\mathbf{B}\psi$. Show that this leads to the orbital g-factor $g_L = 1$ upon writing the magnetic moment as $\boldsymbol{\mu} = g_L \mathbf{L}$ in units of Bohr magnetons. See also Example 1.7.1.

1.8.8 Show that any solution of the equation

$$\nabla \times \nabla \times \mathbf{A} - k^2 \mathbf{A} = 0$$

automatically satisfies the vector Helmholtz equation

$$\nabla^2 \mathbf{A} + k^2 \mathbf{A} = 0$$

and the solenoidal condition

$$\nabla \cdot \mathbf{A} = 0.$$

Hint. Let $\nabla \cdot$ operate on the first equation.

1.9 Vector Integration

The next step after differentiating vectors is to integrate them. Let us start with line integrals and then proceed to surface and volume integrals. In each case, the method of attack will be to reduce the vector integral to one-dimensional integrals over a coordinate interval.

Line Integrals

Using an increment of length $dr = \hat{\mathbf{x}}\,dx + \hat{\mathbf{y}}\,dy + \hat{\mathbf{z}}\,dz$, we often encounter the line integral

$$\int_C \mathbf{V}\cdot d\mathbf{r}, \tag{1.101}$$

in which the integral is over some contour C that may be open (with starting point and ending point separated) or closed (forming a loop) instead of an interval of the x-axis. The Riemann integral is defined by subdividing the curve into ever smaller segments whose number grows indefinitely. The form [Eq. (1.101)] is exactly the same as that encountered when we calculate the work done by a force that varies along the path

$$W = \int \mathbf{F}\cdot d\mathbf{r} = \int F_x(x, y, z)\,dx + \int F_y(x, y, z)\,dy + \int F_z(x, y, z)\,dz, \tag{1.102}$$

that is, a sum of conventional integrals over intervals of one variable each. In this expression, $\mathbf{F}$ is the force exerted on a particle. In general, such integrals depend on the path except for conservative forces, whose treatment we postpone to Section 1.12.

EXAMPLE 1.9.1

Path-Dependent Work The force exerted on a body is $\mathbf{F} = -\hat{\mathbf{x}}y + \hat{\mathbf{y}}x$. The problem is to calculate the work done going from the origin to the point (1, 1),

$$W = \int_{0,0}^{1,1} \mathbf{F}\cdot d\mathbf{r} = \int_{0,0}^{1,1} (-y\,dx + x\,dy). \tag{1.103}$$

Separating the two integrals, we obtain

$$W = -\int_0^1 y\,dx + \int_0^1 x\,dy. \tag{1.104}$$

The first integral cannot be evaluated until we specify the values of y as x ranges from 0 to 1. Likewise, the second integral requires x as a function of y. Consider first the path shown in Fig. 1.27. Then

$$W = -\int_0^1 0\,dx + \int_0^1 1\,dy = 1 \tag{1.105}$$

because $y = 0$ along the first segment of the path and $x = 1$ along the second. If we select the path $[x = 0, 0 \le y \le 1]$ and $[0 \le x \le 1, y = 1]$, then Eq. (1.103) gives $W = -1$. For this force, the work done depends on the choice of path. ■

EXAMPLE 1.9.2

Line Integral for Work Find the work done going around a unit circle clockwise from 0 to $-\pi$ shown in Fig. 1.28 in the xy-plane doing work against a force field given by

$$\mathbf{F} = \frac{-\hat{\mathbf{x}}y}{x^2+y^2} + \frac{\hat{\mathbf{y}}x}{x^2+y^2}.$$

Figure 1.27

A Path of Integration

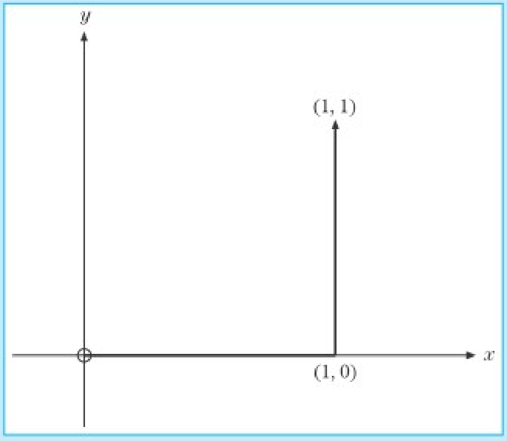

Figure 1.28

Circular and Square Integration Paths

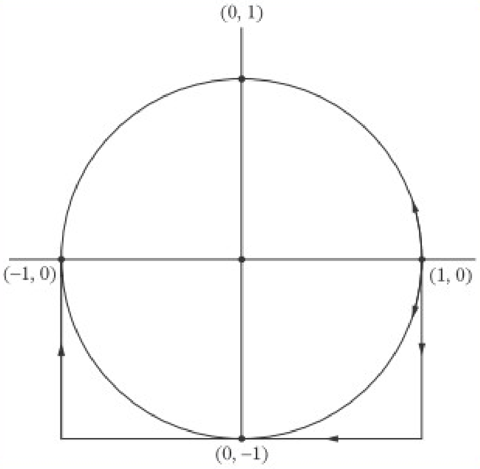

Let us parameterize the circle C as $x = \cos\varphi$, $y = \sin\varphi$ with the polar angle φ so that $dx = -\sin\varphi d\varphi$, $dy = \cos\varphi d\varphi$. Then the force can be written as $\mathbf{F} = -\hat{\mathbf{x}}\sin\varphi + \hat{\mathbf{y}}\cos\varphi$. The work becomes

$$-\int_C \frac{xdy - ydx}{x^2 + y^2} = \int_0^{-\pi} (-\sin^2\varphi - \cos^2\varphi)\, d\varphi = \pi.$$

Here we spend energy. If we integrate anticlockwise from $\varphi = 0$ to π we find the value $-\pi$ because we are riding with the force. The work is path dependent, which is consistent with the physical interpretation that $\mathbf{F} \cdot d\mathbf{r} \sim xdy - ydx = L_z$ is proportional to the z-component of orbital angular momentum (involving circulation, as discussed in Section 1.7).

If we integrate along the square through the points $(\pm 1, 0)$, $(0, -1)$ surrounding the circle, we find for the clockwise lower half square path of Fig. 1.28

$$\begin{aligned}
-\int \mathbf{F} \cdot d\mathbf{r} &= -\int_0^{-1} F_y \, dy|_{x=1} - \int_1^{-1} F_x \, dx|_{y=-1} - \int_{-1}^{0} F_y \, dy|_{x=-1} \\
&= \int_0^1 \frac{dy}{1+y^2} + \int_{-1}^{1} \frac{dx}{x^2+(-1)^2} + \int_{-1}^{0} \frac{dy}{(-1)^2+y^2} \\
&= \arctan(1) + \arctan(1) - \arctan(-1) - \arctan(-1) \\
&= 4 \cdot \frac{\pi}{4} = \pi,
\end{aligned}$$

which is consistent with the circular path.

For the circular paths we used the $x = \cos\varphi$, $y = \sin\varphi$ parameterization, whereas for the square shape we used the standard definitions $y = f(x)$ or $x = g(y)$ of a curve, that is, $y = -1 = \text{const.}$ and $x = \pm 1 = \text{const.}$ We could have used the implicit definition $F(x, y) \equiv x^2 + y^2 - 1 = 0$ of the circle. Then the total variation

$$dF = \frac{\partial F}{\partial x} dx + \frac{\partial F}{\partial y} dy = 2x \, dx + 2y \, dy \equiv 0$$

so that

$$dy = -x \, dx/y \text{ with } y = -\sqrt{1-x^2}$$

on our half circle. The work becomes

$$\begin{aligned}
-\int_C \frac{x \, dy - y \, dx}{x^2+y^2} = \int \left(\frac{x^2}{y} + y \right) dx = \int \frac{dx}{y} = \int_1^{-1} \frac{dx}{-\sqrt{1-x^2}} \\
= \arcsin 1 - \arcsin(-1) = 2 \cdot \frac{\pi}{2} = \pi,
\end{aligned}$$

in agreement with our previous results. ■

EXAMPLE 1.9.3

Gravitational Potential If a force can be described by a scalar function V_G as $\mathbf{F} = -\nabla V_G(\mathbf{r})$ [Eq. (1.65)], everywhere we call V_G its potential in mechanics and engineering. Because the total variation $dV_G = \nabla V_G \cdot d\mathbf{r} = -\mathbf{F}_G \cdot d\mathbf{r}$ is the work done against the force along the path $d\mathbf{r}$, the integrated work along any path from the initial point $\mathbf{r}_0$ to the final point $\mathbf{r}$ is given by a line integral $\int_{\mathbf{r}_0}^{\mathbf{r}} dV_G = V_G(\mathbf{r}) - V_G(\mathbf{r}_0)$, the potential difference between the end points of the path. Thus, to find the scalar potential for the gravitational force on a unit mass m_1,

$$\mathbf{F}_G = -\frac{Gm_1m_2\hat{\mathbf{r}}}{r^2} = -\frac{k\hat{\mathbf{r}}}{r^2}, \quad \text{radially } \textbf{inward} \tag{1.106}$$

we integrate from infinity, where V_G is zero into position $\mathbf{r}$. We obtain

$$V_G(r) - V_G(\infty) = -\int_{\infty}^{\mathbf{r}} \mathbf{F}_G \cdot d\mathbf{r} = +\int_{\mathbf{r}}^{\infty} \mathbf{F}_G \cdot d\mathbf{r}. \tag{1.107}$$

By use of $\mathbf{F}_G = -\mathbf{F}_{\text{applied}}$, the potential is the work done in bringing the unit mass in from infinity. (We can define only the potential difference. Here, we arbitrarily assign infinity to be a zero of potential.) Since $\mathbf{F}_G$ is radial, we obtain a contribution to V_G only when $d\mathbf{r}$ is radial or

$$V_G(r) = -\int_{r}^{\infty} \frac{k\,dr}{r^2} = -\frac{k}{r} = -\frac{Gm_1m_2}{r}. \tag{1.108}$$

The negative sign reflects the attractive nature of gravity. ■

Surface Integrals

Surface integrals appear in the same forms as line integrals, the element of area also being a vector, $d\boldsymbol{\sigma}$.[13] Often this area element is written $\mathbf{n}\,dA$, where $\mathbf{n}$ is a unit (normal) vector to indicate the positive direction.[14] There are two conventions for choosing the positive direction. First, if the surface is a closed surface, we agree to take the outward normal as positive. Second, if the surface is an open surface, the positive normal depends on the direction in which the perimeter of the open surface is traversed. If the right-hand fingers are curled in the direction of travel around the perimeter, the positive normal is indicated by the thumb of the right hand. As an illustration, a circle in the xy-plane (Fig. 1.29) mapped out from x to y to $-x$ to $-y$ and back to x will have its positive normal parallel to the positive z-axis (for the right-handed coordinate system).

Analogous to the line integrals, Eq. (1.101), surface integrals may appear in the form

$$\int \mathbf{V} \cdot d\boldsymbol{\sigma}. \tag{1.109}$$

This surface integral $\int \mathbf{V} \cdot d\boldsymbol{\sigma}$ may be interpreted as a flow or flux through the given surface. This is really what we did in Section 1.6 to understand the significance of the concept of divergence. Note that both physically and from the dot product the tangential components of the velocity contribute nothing to the flow through the surface.

---

[13] Recall that in Section 1.3 the area (of a parallelogram) is represented by a cross product **vector**.

[14] Although $\mathbf{n}$ always has unit length, its direction may well be a function of position.

Figure 1.29

Right-Hand Rule for the Positive Normal

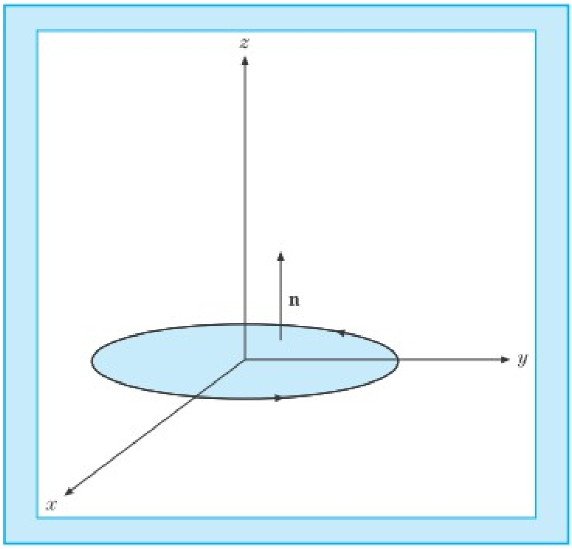

Figure 1.30

The Parabola $y = x^2$ for $0 \le y \le 1$ Rotated About the y-Axis

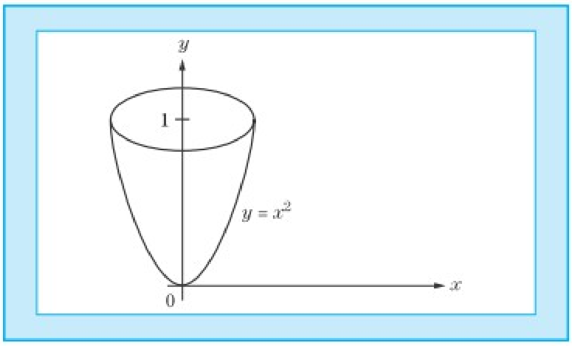

EXAMPLE 1.9.4

Moment of Inertia Let us determine the moment of inertia I_y of a segment of the parabola $y = x^2$ cut off by the line $y = 1$ and rotated about the y-axis (Fig. 1.30). We find

$$I_y = 2\mu \int_{x=0}^{1} \int_{y=x^2}^{1} x^2 \, dx \, dy = 2\mu \int_0^1 (1 - x^2) x^2 dx = 2\mu \left(\frac{x^3}{3} - \frac{x^5}{5} \right) \Big|_0^1 = \frac{4\mu}{15}.$$

The factor of 2 originates in the reflection symmetry of $x \to -x$, and μ is the constant mass density. ■

A surface in three-dimensional space may be explicitly given as $z = f(x, y)$ or by the coordinate functions of its points

$$x = x(u, v), \quad y = y(u, v), \quad z = z(u, v)$$

in terms of two parameters u, v or in implicit form $F(x, y, z) = 0$. The explicit form is a special case

$$F(x, y, z) \equiv z - f(x, y)$$

of the general implicit definition of a surface. We find the area $dA = dx\,dy/n_z$ over the projection $dx\,dy$ of the surface onto the xy-plane for the latter case. Here, $n_z = \cos\gamma$ is the z-component of the normal unit vector $\mathbf{n}$ at $\mathbf{r}$ on the surface so that γ is the angle of $d\mathbf{A}$ with the xy-plane. Thus, when we project $d\mathbf{A}$ to the xy-plane, we get $dA\cos\gamma = dx\,dy$, which proves this useful formula for measuring the **area of a curved surface**. From the gradient properties we also know that $\mathbf{n} = \nabla f/\sqrt{\nabla f^2}$.

EXAMPLE 1.9.5

A Surface Integral Here we apply the general formula for surface integrals to find the area on $z = xy = f(x, y)$ cut out by the unit circle in the xy-plane shown in Fig. 1.31. We start from

$$\frac{\partial f}{\partial x} = \frac{\partial z}{\partial x} = y, \quad \frac{\partial f}{\partial y} = \frac{\partial z}{\partial y} = x, \quad \frac{\partial f}{\partial z} = \frac{\partial z}{\partial z} = 1,$$

which we substitute into

$$n_z = 1\Big/\sqrt{1 + \left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2}$$

for the normal to yield the area

$$A = \int_{x=-1}^{x=1} \int_{y=-\sqrt{1-x^2}}^{y=\sqrt{1-x^2}} \sqrt{1 + x^2 + y^2}\, dx\, dy.$$

Figure 1.31

The Surface $z = xy$ Above and Below the Unit Circle $x^2 + y^2 = 1$

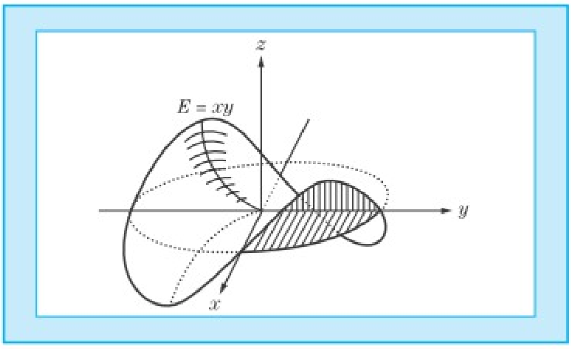

For the circular geometry plane polar coordinates r, φ are more appropriate, where the radial integral is evaluated by substituting $u = 1 + r^2$ in

$$A = \int_0^1 \sqrt{1+r^2}\,r\,dr \int_0^{2\pi} d\varphi = \pi \int \sqrt{u}\,du = \frac{2\pi}{3}\left|(1+r^2)^{3/2}\right|_0^1 = \frac{2\pi}{3}(2\sqrt{2}-1).$$

■

More examples of line and surface integrals are provided in Chapter 2.

Volume Integrals

Volume integrals are simpler because the volume element $d\tau$ is a scalar quantity.[15] We have

$$\int_v \mathbf{V}\,d\tau = \hat{\mathbf{x}}\int_v V_x\,d\tau + \hat{\mathbf{y}}\int_v V_y\,d\tau + \hat{\mathbf{z}}\int_v V_z\,d\tau, \tag{1.110}$$

again reducing the vector integral to a vector sum of scalar integrals.

If the vector

$$\mathbf{V} = V_\rho(\rho,\varphi,z)\hat{\boldsymbol{\rho}} + V_\varphi(\rho,\varphi,z)\hat{\boldsymbol{\varphi}} + V_z(\rho,\varphi,z)\hat{\mathbf{z}}$$

and its components are given in cylindrical coordinates $x = \rho\cos\varphi$, $y = \rho\sin\varphi$ with volume element $d\tau = \rho\,d\rho\,d\varphi\,dz$, the volume integral

$$\int_v \mathbf{V}\,d\tau = \hat{\mathbf{z}}\int_v V_z\,d\tau + \int\!\!\int\!\!\int (V_\rho\hat{\boldsymbol{\rho}} + V_\varphi\hat{\boldsymbol{\varphi}})\rho\,d\rho\,d\varphi\,dz$$

involves integrals over the varying unit vectors of the polar coordinates. To reduce them to scalar integrals, we need to expand the polar coordinate unit vectors in Cartesian unit vectors as follows. Dividing the plane coordinates by ρ, we find

$$\hat{\boldsymbol{\rho}} = \frac{1}{\rho}(x, y) = (\cos\varphi, \sin\varphi) = \hat{\mathbf{x}}\cos\varphi + \hat{\mathbf{y}}\sin\varphi.$$

Differentiating $\hat{\boldsymbol{\rho}}^2 = 1$, we see from $0 = \frac{d\hat{\boldsymbol{\rho}}^2}{d\varphi} = 2\hat{\boldsymbol{\rho}}\cdot\frac{d\hat{\boldsymbol{\rho}}}{d\varphi}$ that

$$\frac{d\hat{\boldsymbol{\rho}}}{d\varphi} = -\hat{\mathbf{x}}\sin\varphi + \hat{\mathbf{y}}\cos\varphi = \hat{\boldsymbol{\varphi}}$$

is perpendicular to $\hat{\boldsymbol{\rho}}$ and a unit vector; therefore, it is equal to $\hat{\boldsymbol{\varphi}}$. Substituting these expressions into the second integral yields the final result

$$\begin{aligned}\int_v \mathbf{V}\,d\tau &= \hat{\mathbf{z}}\int_v V_z\,d\tau + \hat{\mathbf{x}}\int\!\!\int\!\!\int [V_\rho\cos\varphi - V_\varphi\sin\varphi]\rho\,d\rho\,d\varphi\,dz \\ &\quad + \hat{\mathbf{y}}\int\!\!\int\!\!\int [V_\rho\sin\varphi + V_\varphi\cos\varphi]\rho\,d\rho\,d\varphi\,dz.\end{aligned} \tag{1.111}$$

The terms in brackets are the Cartesian components V_x, V_y expressed in plane polar coordinates.

---

[15] Frequently, the symbols d^3r and d^3x are used to denote a volume element in coordinate (xyz or $x_1x_2x_3$) space.

In spherical polar coordinates, all of the unit vectors depend on the coordinates, none can be pulled out of the integrals, and all have to be expanded in Cartesian unit vectors. This task of rewriting Eq. (1.110) is left as an exercise.

EXAMPLE 1.9.6

Volume of Rotated Gaussian Rotate the Gaussian $y = \exp(-x^2)$ about the z-axis leading to $z = \exp(-x^2 - y^2)$. Then the volume in the polar (cylindrical) coordinates appropriate for the geometry is given by

$$V = \int_{r=0}^{\infty} \int_{\varphi=0}^{2\pi} \int_{z=0}^{e^{-r^2}} r\,dr\,d\varphi\,dz = 2\pi \int_0^{\infty} re^{-r^2} dr = \pi \int_0^{\infty} e^{-u} du = \pi,$$

upon substituting $\exp(-x^2 - y^2) = \exp(-r^2)$, $dx\,dy = r\,dr\,d\varphi$, $u = r^2$, and $du = 2rdr$. ■

Integral Definitions of Gradient, Divergence, and Curl

One interesting and significant application of our surface and volume integrals is their use in developing alternate definitions of our differential relations. We find

$$\nabla\varphi = \lim_{\int d\tau \to 0} \frac{\int \varphi\, d\boldsymbol{\sigma}}{\int d\tau}, \tag{1.112}$$

$$\nabla \cdot \mathbf{V} = \lim_{\int d\tau \to 0} \frac{\int \mathbf{V} \cdot d\boldsymbol{\sigma}}{\int d\tau}, \tag{1.113}$$

$$\nabla \times \mathbf{V} = \lim_{\int d\tau \to 0} \frac{\int d\boldsymbol{\sigma} \times \mathbf{V}}{\int d\tau}. \tag{1.114}$$

In these three equations, $\int d\tau$ is the volume of a small region of space and $d\boldsymbol{\sigma}$ is the vector area element of this volume. The identification of Eq. (1.113) as the divergence of **V** was carried out in Section 1.6. Here, we show that Eq. (1.112) is consistent with our earlier definition of $\nabla\varphi$ [Eq. (1.64)]. For simplicity, we choose $d\tau$ to be the differential volume $dx\,dy\,dz$ (Fig. 1.32). This

Figure 1.32

Differential Rectangular Parallelepiped (Origin at Center)

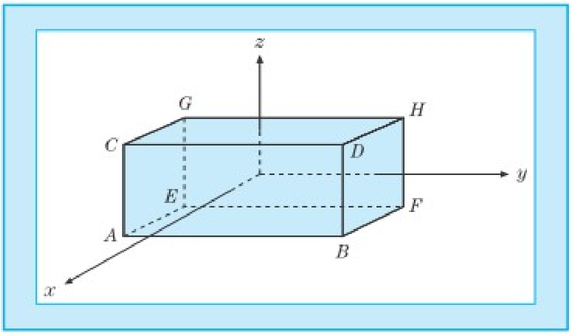

time, we place the origin at the geometric center of our volume element. The area integral leads to six integrals, one for each of the six faces. Remembering that $d\boldsymbol{\sigma}$ is outward, $d\boldsymbol{\sigma} \cdot \hat{\mathbf{x}} = -|d\boldsymbol{\sigma}|$ for surface *EFHG*, and $+|d\boldsymbol{\sigma}|$ for surface *ABDC*, we have

$$\begin{aligned}\int \varphi\, d\boldsymbol{\sigma} = {}& -\hat{\mathbf{x}} \int_{EFHG} \left(\varphi - \frac{\partial \varphi}{\partial x}\frac{dx}{2}\right) dy\,dz + \hat{\mathbf{x}} \int_{ABDC} \left(\varphi + \frac{\partial \varphi}{\partial x}\frac{dx}{2}\right) dy\,dz \\ & -\hat{\mathbf{y}} \int_{AEGC} \left(\varphi - \frac{\partial \varphi}{\partial y}\frac{dy}{2}\right) dx\,dz + \hat{\mathbf{y}} \int_{BFHD} \left(\varphi + \frac{\partial \varphi}{\partial y}\frac{dy}{2}\right) dx\,dz \\ & -\hat{\mathbf{z}} \int_{ABFE} \left(\varphi - \frac{\partial \varphi}{\partial z}\frac{dz}{2}\right) dx\,dy + \hat{\mathbf{z}} \int_{CDHG} \left(\varphi + \frac{\partial \varphi}{\partial z}\frac{dz}{2}\right) dx\,dy.\end{aligned}$$

Using the first two terms of a Maclaurin expansion, we evaluate each integrand at the origin with a correction included to correct for the displacement ($\pm dx/2$, etc.) of the center of the face from the origin. Having chosen the total volume to be of differential size ($\int d\tau = dx\,dy\,dz$), we drop the integral signs on the right and obtain

$$\int \varphi\, d\boldsymbol{\sigma} = \left(\hat{\mathbf{x}}\frac{\partial \varphi}{\partial x} + \hat{\mathbf{y}}\frac{\partial \varphi}{\partial y} + \hat{\mathbf{z}}\frac{\partial \varphi}{\partial z}\right) dx\,dy\,dz. \tag{1.115}$$

Dividing by

$$\int d\tau = dx\,dy\,dz,$$

we verify Eq. (1.112).

This verification has been oversimplified in ignoring other correction terms beyond the first derivatives. These additional terms, which are introduced in Section 5.6 when the Taylor expansion is developed, vanish in the limit

$$\int d\tau \to 0\,(dx \to 0,\, dy \to 0,\, dz \to 0).$$

This, of course, is the reason for specifying in Eqs. (1.112)–(1.114) that this limit be taken. Verification of Eq. (1.114) follows these same lines, using a differential volume $d\tau = dx\,dy\,dz$.

EXERCISES

1.9.1 Find the potential for the electric field generated by a charge q at the origin. Normalize the potential to zero at spatial infinity.

1.9.2 Determine the gravitational field of the earth taken to be spherical and of uniform mass density. Punch out a concentric spherical cavity and show that the field is zero inside it. Show that the field is constant if the cavity is not concentric.

5- NT2< 3

1.9.3 Evaluate

$$\frac{1}{3}\int_s \mathbf{r}\cdot d\boldsymbol{\sigma}$$

over the unit cube defined by the point (0, 0, 0) and the unit intercepts on the positive x-, y-, and z-axes. Note that (a) $\mathbf{r}\cdot d\boldsymbol{\sigma}$ is zero for three of the surfaces, and (b) each of the three remaining surfaces contributes the same amount to the integral.

1.9.4 Show by expansion of the surface integral that

$$\lim_{\int d\tau\to 0}\frac{\int_s d\boldsymbol{\sigma}\times\mathbf{V}}{\int d\tau}=\boldsymbol{\nabla}\times\mathbf{V}.$$

Hint. Choose the volume to be a differential volume, $dx\,dy\,dz$.

1.10 Gauss's Theorem

Here, we derive a useful relation between a surface integral of a vector and the volume integral of the divergence of that vector. Let us assume that the vector **V** and its first derivatives are continuous over the simply connected region (without holes) of interest. Then, **Gauss's theorem** states that

$$\int_S \mathbf{V}\cdot d\boldsymbol{\sigma}=\int_V \boldsymbol{\nabla}\cdot\mathbf{V}\,d\tau. \qquad (1.116\text{a})$$

In words, the surface integral of a vector over a closed surface equals the volume integral of the divergence of that vector integrated over the volume enclosed by the surface.

Imagine that volume V is subdivided into an arbitrarily large number of tiny (differential) parallelepipeds. For each parallelepiped,

$$\sum_{\text{six surfaces}}\mathbf{V}\cdot d\boldsymbol{\sigma}=\boldsymbol{\nabla}\cdot\mathbf{V}\,d\tau \qquad (1.116\text{b})$$

from the analysis of Section 1.6, Eq. (1.75), with $\rho\mathbf{v}$ replaced by **V**. The summation is over the six faces of the parallelepiped. Summing over all parallelepipeds, we find that the $\mathbf{V}\cdot d\boldsymbol{\sigma}$ terms cancel (pairwise) for all **interior** faces; only the contributions of the **exterior** surfaces survive (Fig. 1.33). Analogous to the definition of a Riemann integral as the limit of a sum, we take the limit as the number of parallelepipeds approaches infinity (→ ∞) and the dimensions of each approach zero (→ 0):

$$\sum_{\text{exterior surfaces}}\mathbf{V}\cdot_{\downarrow} d\boldsymbol{\sigma}=\sum_{\text{volumes}}\boldsymbol{\nabla}\cdot_{\downarrow}\mathbf{V}\,d\tau$$

$$\int_S \mathbf{V}\cdot d\boldsymbol{\sigma}=\int_V \boldsymbol{\nabla}\cdot\mathbf{V}\,d\tau.$$

The result is Eq. (1.116a), Gauss's theorem.

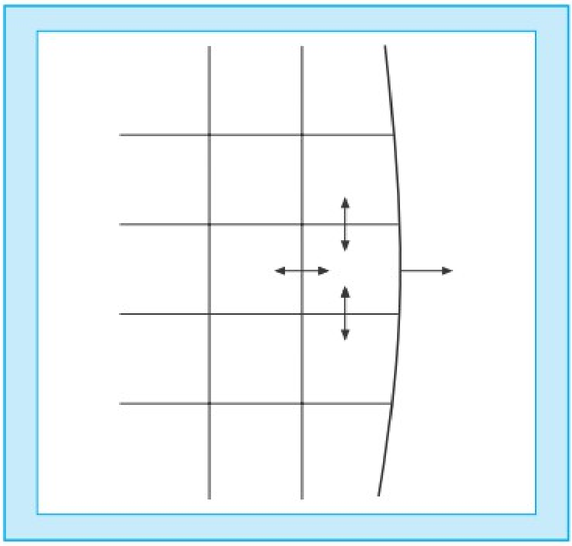

Figure 1.33

Exact Cancellation of $\mathbf{V} \cdot d\boldsymbol{\sigma}$'s on Interior Surfaces. No Cancellation on the Exterior Surface

From a physical standpoint, Eq. (1.75) has established $\nabla \cdot \mathbf{V}$ as the net outflow of field per unit volume. The volume integral then gives the total net outflow. However, the surface integral $\int \mathbf{V} \cdot d\boldsymbol{\sigma}$ is just another way of expressing this same quantity, which is the equality, Gauss's theorem.

Biographical Data

Gauss, Carl Friedrich. Gauss, a German mathematician, physicist, and astronomer, was born in Brunswick in 1777 and died in Göttingen in 1855. He was an infant prodigy in mathematics whose education was directed and financed by the Duke of Brunswick. As a teenager, he proved that regular n-polygons can be constructed in Euclidean geometry provided n is a Fermat prime number such as 3, 5, 17, and 257, a major advance in geometry since antiquity. This feat convinced him to stay in mathematics and give up the study of foreign languages. For his Ph.D., he proved the fundamental theorem of algebra, avoiding the then controversial complex numbers he had used to discover it. In his famous treatise *Disquisitiones Arithmetica* on number theory, he first proved the quadratic reciprocity theorem and originated the terse style and rigor of mathematical proofs as a series of logical steps, discarding any trace of the original heuristic ideas used in the discovery and checks of examples. Not surprisingly, he hated teaching. He is considered by many as the greatest mathematician of all times and was the last to provide major contributions to all then existing branches of mathematics. As the founder of differential geometry, he developed the intrinsic properties of surfaces, such as curvature, which later motivated B. Riemann to develop the geometry of metric spaces, the mathematical foundation of Einstein's General Relativity. In astronomy (for the orbit of the asteroid Ceres), he developed the method of least squares for fitting curves to data. In physics, he developed potential theory, and the unit of the magnetic induction is named after him in honor of his measurements and development of units in physics.

Green's Theorem

A frequently useful corollary of Gauss's theorem is a relation known as Green's theorem. If u and v are two scalar functions, we have the identities

$$\nabla \cdot (u\,\nabla v) = u\nabla \cdot \nabla v + (\nabla u) \cdot (\nabla v), \tag{1.117}$$

$$\nabla \cdot (v\,\nabla u) = v\nabla \cdot \nabla u + (\nabla v) \cdot (\nabla u), \tag{1.118}$$

which follow from the product rule of differentiation. Subtracting Eq. (1.118) from Eq. (1.117), integrating over a volume (u, v, and their derivatives, assumed continuous), and applying Eq. (1.116a) (Gauss's theorem), we obtain

$$\int_V (u\nabla \cdot \nabla v - v\nabla \cdot \nabla u)\,d\tau = \int_S (u\nabla v - v\nabla u) \cdot d\boldsymbol{\sigma}. \tag{1.119}$$

This is **Green's theorem**, which states that the antisymmetric Laplacian of a pair of functions integrated over a simply connected volume (no holes) is equivalent to the antisymmetric gradient of the pair integrated over the bounding surface. An alternate form of Green's theorem derived from Eq. (1.117) alone is

$$\int_S u\nabla v \cdot d\boldsymbol{\sigma} = \int_V u\nabla \cdot \nabla v\,d\tau + \int_V \nabla u \cdot \nabla v\,d\tau. \tag{1.120}$$

Finally, Gauss's theorem may also be extended to tensors (see Section 2.11).

Biographical Data

Green, George. Green, an English mathematician, was born in Nottingham in 1793 and died near Nottingham in 1841. He studied Laplace's papers in Cambridge and developed potential theory in electrodynamics.

EXERCISES

1.10.1 If $\mathbf{B} = \nabla \times \mathbf{A}$, show that

$$\int_S \mathbf{B} \cdot d\boldsymbol{\sigma} = 0$$

for any closed surface S. State this in words. If symbolic software is available, check this for a typical vector potential and specific surfaces, such as a sphere or cube.

1.10.2 Over some volume V, let ψ be a solution of Laplace's equation (with the derivatives appearing there continuous). Prove that the integral over any closed surface in V of the normal derivative of ψ ($\partial\psi/\partial n$, or $\nabla\psi\cdot\mathbf{n}$) will be zero.

1.10.3 In analogy to the integral definition of gradient, divergence, and curl of Section 1.10, show that

$$\nabla^2\varphi = \lim_{\int d\tau\to 0}\frac{\int \nabla\varphi\cdot d\boldsymbol{\sigma}}{\int d\tau}.$$

1.10.4 The electric displacement vector $\mathbf{D}$ satisfies the Maxwell equation $\nabla\cdot\mathbf{D} = \rho$, where ρ is the charge density (per unit volume). At the boundary between two media there is a surface charge density σ (per unit area). Show that a boundary condition for $\mathbf{D}$ is

$$(\mathbf{D}_2 - \mathbf{D}_1)\cdot\mathbf{n} = \sigma,$$

where $\mathbf{n}$ is a unit vector normal to the surface and out of medium 1. *Hint.* Consider a thin pillbox as shown in Fig. 1.34.

Figure 1.34

Pillbox

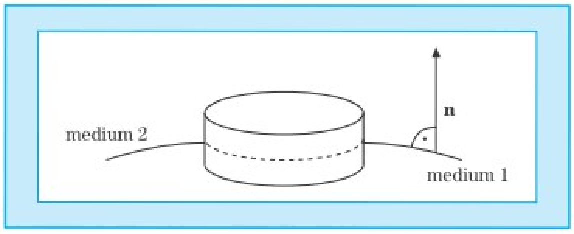

1.10.5 From Eq. (1.77) and Example 1.6.1, with $\mathbf{V}$ the electric field $\mathbf{E}$ and f the electrostatic potential φ, show that

$$\int \rho\varphi\, d\tau = \varepsilon_0 \int E^2\, d\tau.$$

This corresponds to a three-dimensional integration by parts.
Hint. $\mathbf{E} = -\nabla\varphi$, $\nabla\cdot\mathbf{E} = \rho/\varepsilon_0$. You may assume that φ vanishes at large r at least as fast as r^{-1}.

1.10.6 The creation of a **localized** system of steady electric currents (current density $\mathbf{J}$) and magnetic fields may be shown to require an amount of work

$$W = \frac{1}{2}\int \mathbf{H}\cdot\mathbf{B}\, d\tau.$$

Transform this into

$$W = \frac{1}{2}\int \mathbf{J}\cdot\mathbf{A}\, d\tau,$$

where $\mathbf{A}$ is the magnetic vector potential, $\nabla\times\mathbf{A} = \mathbf{B}$.

Hint. In Maxwell's equations, take the displacement current term $\partial\mathbf{D}/\partial t = 0$ and explain why using Ohm's law. If the fields and currents are localized, a bounding surface may be taken far enough out so that the integrals of the fields and currents over the surface yield zero.

1.11 Stokes's Theorem

Gauss's theorem relates the volume integral of a derivative of a function to an integral of the function over the closed surface bounding the volume. Here, we consider an analogous relation between the surface integral of a derivative of a function and the line integral of the function, the path of integration being the perimeter bounding the surface.

Let us take the surface and subdivide it into a network of arbitrarily small rectangles. In Section 1.7, we showed that the circulation about such a differential rectangle (in the xy-plane) is $\nabla \times \mathbf{V}|_z \; dx\,dy$. From Eq. (1.85) applied to **one** differential rectangle,

$$\sum_{\text{four sides}} \mathbf{V} \cdot d\boldsymbol{\lambda} = \nabla \times \mathbf{V} \cdot d\boldsymbol{\sigma}. \tag{1.121}$$

We sum over all the little rectangles as in the definition of a Riemann integral. The surface contributions [right-hand side of Eq. (1.121)] are added together. The line integrals [left-hand side of Eq. (1.121)] of all **interior** line segments cancel identically. Only the line integral around the perimeter survives (Fig. 1.35). Taking the usual limit as the number of rectangles approaches

Figure 1.35

Exact Cancellation on Interior Paths; No Cancellation on the Exterior Path

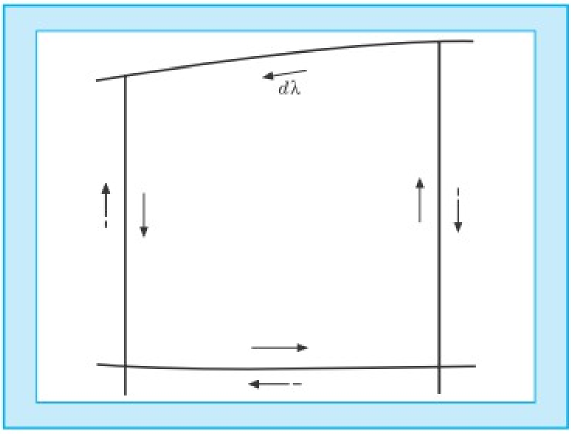

infinity while $dx \to 0$, $dy \to 0$, we have

$$\begin{array}{ccc} \displaystyle\sum_{\substack{\text{exterior line}\\ \text{segments}}} \mathbf{V}\cdot d\boldsymbol{\lambda} & = & \displaystyle\sum_{\text{rectangles}} \boldsymbol{\nabla}\times\mathbf{V}\cdot d\boldsymbol{\sigma} \\ \downarrow & & \downarrow \\ \displaystyle\oint \mathbf{V}\cdot d\boldsymbol{\lambda} & = & \displaystyle\int_S \boldsymbol{\nabla}\times\mathbf{V}\cdot d\boldsymbol{\sigma}. \end{array} \tag{1.122}$$

This is Stokes's theorem. The surface integral on the right is over the surface bounded by the perimeter or contour for the line integral on the left. The direction of the vector representing the area is out of the paper plane toward the reader if the direction of traversal around the contour for the line integral is in the positive mathematical sense as shown in Fig. 1.35.

This demonstration of Stokes's theorem is limited by the fact that we used a Maclaurin expansion of $\mathbf{V}(x, y, z)$ in establishing Eq. (1.85) in Section 1.7. Actually, we need only demand that the curl of $\mathbf{V}(x, y, z)$ exists and that it be integrable over the surface. Stokes's theorem obviously applies to an open, simply connected surface. It is possible to consider a closed surface as a limiting case of an open surface with the opening (and therefore the perimeter) shrinking to zero. This is the point of Exercise 1.11.4.

As a special case of Stokes's theorem, consider the curl of a two-dimensional vector field $\mathbf{V} = (V_1(x, y), V_2(x, y), 0)$. The curl $\boldsymbol{\nabla}\times\mathbf{V} = (0, 0, \frac{\partial V_2}{\partial x} - \frac{\partial V_1}{\partial y})$ so

$$\int_S \boldsymbol{\nabla}\times\mathbf{V}\cdot\hat{\mathbf{z}}\,dx\,dy = \int_S \left(\frac{\partial V_2}{\partial x} - \frac{\partial V_1}{\partial y}\right) dx\,dy = \int_C \mathbf{V}\cdot d\mathbf{r} = \int_C (V_1 dx + V_2 dy),$$

where the curve C is the boundary of the simply connected surface S that is integrated in the positive mathematical sense (anticlockwise). This relation is sometimes also called Green's theorem. In Chapter 6, we shall use it to prove Cauchy's theorem for analytic functions.

EXAMPLE 1.11.1

Area as a Line Integral For the two-dimensional Stokes's theorem, we first choose $\mathbf{V} = x\hat{\mathbf{y}}$, which gives the area $S = \int_S dx\,dy = \int_C x\,dy$, and for $\mathbf{V} = (y\hat{\mathbf{x}})$ we get similarly $S = \int_S dx\,dy = -\int_C y\,dx$. Adding both results gives the area

$$S = \frac{1}{2}\int_C (x\,dy - y\,dx). \quad ■$$

We can use Stokes's theorem to derive Oersted's and Faraday's laws from two of Maxwell's equations and vice versa, thus recognizing that the former are an integrated form of the latter.

EXAMPLE 1.11.2

Oersted's and Faraday's Laws Consider the magnetic field generated by a long wire that carries a stationary current I (Fig. 1.36). Starting from Maxwell's differential law $\boldsymbol{\nabla}\times\mathbf{H} = \mathbf{J}$ [Eq. (1.97c); with Maxwell's displacement current $\partial\mathbf{D}/\partial t = 0$ for a stationary current case by Ohm's law], we integrate over a closed area S perpendicular to and surrounding the wire and apply Stokes's

Figure 1.36

Oersted's Law for a Long Wire Carrying a Current

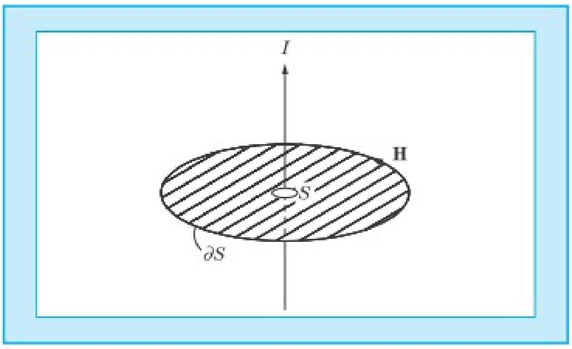

Figure 1.37

Faraday's Induction Law Across a Magnetic Induction Field

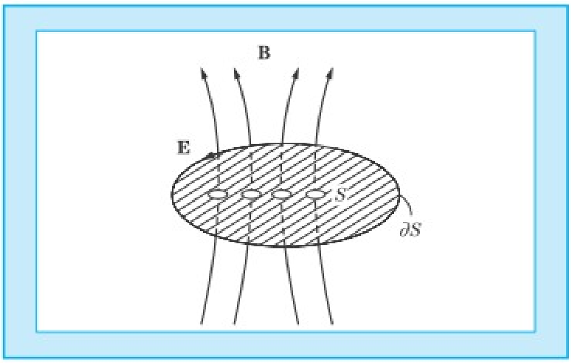

theorem to get

$$I = \int_S \mathbf{J} \cdot d\boldsymbol{\sigma} = \int_S (\boldsymbol{\nabla} \times \mathbf{H}) \cdot d\boldsymbol{\sigma} = \oint_{\partial S} \mathbf{H} \cdot d\mathbf{r},$$

which is Oersted's law. Here, the line integral is along ∂S, the closed curve surrounding the cross section area S.

Similarly, we can integrate Maxwell's equation for $\boldsymbol{\nabla} \times \mathbf{E}$ [Eq. (1.97d)] to yield Faraday's induction law. Imagine moving a closed loop (∂S) of wire (of area S) across a magnetic induction field $\mathbf{B}$ (Fig. 1.37). At a fixed moment of time we integrate Maxwell's equation and use Stokes's theorem, yielding

$$\int_{\partial S} \mathbf{E} \cdot d\mathbf{r} = \int_S (\boldsymbol{\nabla} \times \mathbf{E}) \cdot d\boldsymbol{\sigma} = -\frac{d}{dt} \int_S \mathbf{B} \cdot d\boldsymbol{\sigma} = -\frac{d\Phi}{dt},$$

which is Faraday's law. The line integral on the left-hand side represents the voltage induced in the wire loop, whereas the right-hand side is the change with time of the magnetic flux Φ through the moving surface S of the wire. ■

SUMMARY

Both Stokes's and Gauss's theorems are of tremendous importance in a wide variety of problems involving vector calculus in electrodynamics, where they allow us to derive the local form of Maxwell's differential equations from the global (integral) form of the experimental laws. An indication of their power and versatility may be obtained from the exercises in Sections 1.10 and 1.11 and the development of potential theory in Section 1.12.

Biographical Data

Stokes, Sir George Gabriel. Stokes, a British mathematician and physicist, was born in Skreen, Ireland, in 1819 and died in Cambridge in 1903. Son of a clergyman, his talent for mathematics was already evident in school. In 1849, he became Lucasian professor at Cambridge, the chair Isaac Newton once held and currently held by S. Hawking. In 1885, he became president of the Royal Society. He is known for the theory of viscous fluids, with practical applications to the motion of ships in water. He demonstrated his vision by hailing Joule's work early on and recognizing X-rays as electromagnetic radiation. He received the Rumford and Copley medals of the Royal Society and served as a member of Parliament for Cambridge University in 1887–1892.

EXERCISES

1.11.1 The calculation of the magnetic moment of a current loop leads to the line integral

$$\oint \mathbf{r} \times d\mathbf{r}.$$

(a) Integrate around the perimeter of a current loop (in the xy-plane) and show that the scalar magnitude of this line integral is twice the area of the enclosed surface.

(b) The perimeter of an ellipse is described by $\mathbf{r} = \hat{\mathbf{x}}a\cos\theta + \hat{\mathbf{y}}b\sin\theta$. From part (a), show that the area of the ellipse is πab.

1.11.2 In steady state, the magnetic field $\mathbf{H}$ satisfies the Maxwell equation $\nabla \times \mathbf{H} = \mathbf{J}$, where $\mathbf{J}$ is the current density (per square meter). At the boundary between two media there is a surface current density $\mathbf{K}$ (perimeter). Show that a boundary condition on $\mathbf{H}$ is

$$\mathbf{n} \times (\mathbf{H}_2 - \mathbf{H}_1) = \mathbf{K},$$

where $\mathbf{n}$ is a unit vector normal to the surface and out of medium 1. *Hint.* Consider a narrow loop perpendicular to the interface as shown in Fig. 1.38.

Figure 1.38

Loop Contour

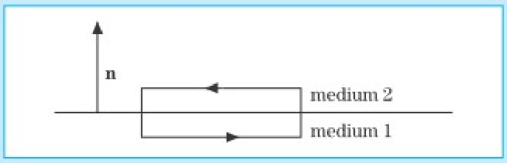

5- T 2 73

Figure 1.42

Gauss's Law

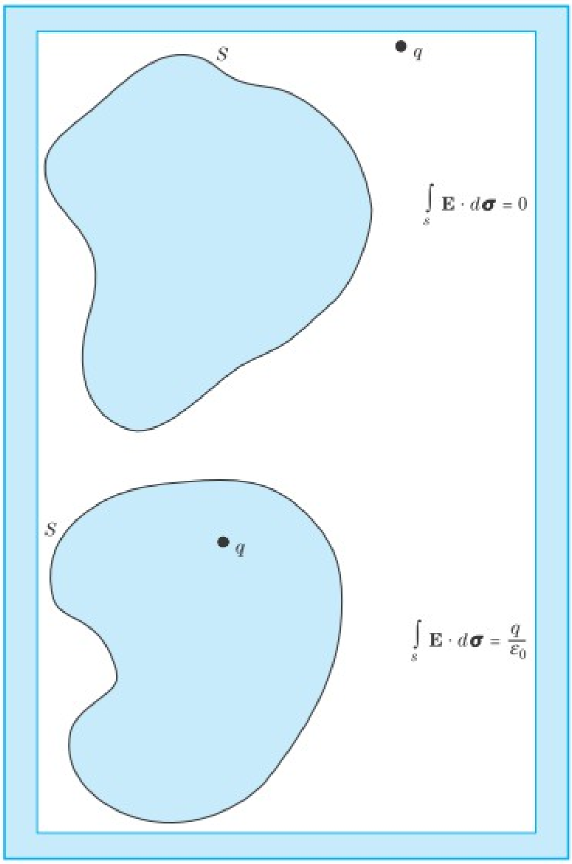
S
q
$\int_S \mathbf{E} \cdot d\boldsymbol{\sigma} = 0$
S
q
$\int_S \mathbf{E} \cdot d\boldsymbol{\sigma} = \frac{q}{\varepsilon_0}$

1.13 Gauss's Law and Poisson's Equation

Gauss's Law

Consider a point electric charge q at the origin of our coordinate system. This produces an electric field $\mathbf{E}$[16] given by

$$\mathbf{E} = \frac{q\hat{\mathbf{r}}}{4\pi\varepsilon_0 r^2}. \tag{1.136}$$

We now derive Gauss's law, which states that the surface integral in Fig. 1.42 is q/ε_0 if the closed surface S includes the origin (where q is located) and zero

---

[16] The electric field $\mathbf{E}$ is defined as the force per unit charge on a small stationary test charge q_t: $\mathbf{E} = \mathbf{F}/q_t$. From Coulomb's law, the force on q_t due to q is $\mathbf{F} = (qq_t/4\pi\varepsilon_0)(\hat{\mathbf{r}}/r^2)$. When we divide by q_t, Eq. (1.136) follows.

Figure 1.43

Exclusion of the Origin

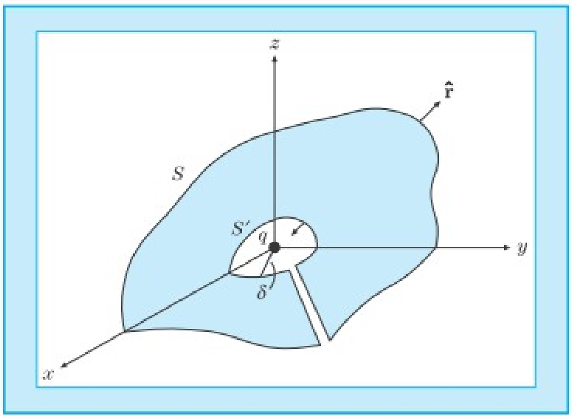

if the surface does not include the origin. The surface S is any closed surface; it need not be spherical.

Using Gauss's theorem [Eq. (1.116a)] (and neglecting the scale factor $q/4\pi\varepsilon_0$), we obtain

$$\int_S \frac{\hat{\mathbf{r}}\cdot d\boldsymbol{\sigma}}{r^2} = \int_V \nabla\cdot\left(\frac{\hat{\mathbf{r}}}{r^2}\right) d\tau = 0 \tag{1.137}$$

by Example 1.6.1, provided the surface S does not include the origin, where the integrands are not defined. This proves the second part of Gauss's law.

The first part, in which the surface S must include the origin, may be handled by surrounding the origin with a small sphere S' of radius δ (Fig. 1.43). So that there will be no question as to what is inside and what is outside, imagine the volume outside the outer surface S and the volume inside surface $S'(r < \delta)$ connected by a small hole. This joins surfaces S and S', combining them into one **single, simply connected closed surface**. Because the radius of the imaginary hole may be made vanishingly small, there is no additional contribution to the surface integral. The inner surface is deliberately chosen to be spherical so that we will be able to integrate over it. Gauss's theorem now applies to the volume between S and S' without any difficulty. We have

$$\int_S \frac{\hat{\mathbf{r}}\cdot d\boldsymbol{\sigma}}{r^2} + \int_{S'} \frac{\hat{\mathbf{r}}\cdot d\boldsymbol{\sigma}'}{\delta^2} = 0. \tag{1.138}$$

We may evaluate the second integral for $d\boldsymbol{\sigma}' = -\hat{\mathbf{r}}\delta^2\, d\Omega$, in which $d\Omega$ is an element of solid angle. The minus sign appears because we agreed in Section 1.9 to have the positive normal $\hat{\mathbf{r}}'$ **outward** from the volume. In this case, the outward $\hat{\mathbf{r}}'$ is in the negative radial direction, $\hat{\mathbf{r}}' = -\hat{\mathbf{r}}$. By integrating over all angles, we have

$$\int_{S'} \frac{\hat{\mathbf{r}}\cdot d\boldsymbol{\sigma}'}{\delta^2} = -\int_{S'} \frac{\hat{\mathbf{r}}\cdot\hat{\mathbf{r}}\delta^2 d\Omega}{\delta^2} = -4\pi, \tag{1.139}$$

independent of the radius δ. With the constants from Eq. (1.136), this results in

$$\int_S \mathbf{E}\cdot d\boldsymbol{\sigma} = \frac{q}{4\pi\varepsilon_0}4\pi = \frac{q}{\varepsilon_0}, \tag{1.140}$$

completing the proof of Gauss's law. Notice that although the surface S may be spherical, it **need not** be spherical.

Going a bit further, we consider a distributed charge so that

$$q = \int_V \rho\, d\tau. \tag{1.141}$$

Equation (1.140) still applies, with q now interpreted as the total distributed charge enclosed by surface S:

$$\int_S \mathbf{E}\cdot d\boldsymbol{\sigma} = \int_V \frac{\rho}{\varepsilon_0}\, d\tau. \tag{1.142}$$

Using Gauss's theorem, we have

$$\int_V \boldsymbol{\nabla}\cdot\mathbf{E}\, d\tau = \int_V \frac{\rho}{\varepsilon_0}\, d\tau. \tag{1.143}$$

Since our volume is completely arbitrary, the integrands must be equal or

$$\boldsymbol{\nabla}\cdot\mathbf{E} = \frac{\rho}{\varepsilon_0}, \tag{1.144}$$

one of Maxwell's equations. If we reverse the argument, Gauss's law follows immediately from Maxwell's equation by integration.

Poisson's Equation

Replacing $\mathbf{E}$ by $-\boldsymbol{\nabla}\varphi$, Eq. (1.144) becomes

$$\boldsymbol{\nabla}\cdot\boldsymbol{\nabla}\varphi = -\frac{\rho}{\varepsilon_0}, \tag{1.145}$$

which is Poisson's equation. We know a solution,

$$\varphi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(\mathbf{r}')d\tau'}{|\mathbf{r}-\mathbf{r}'|},$$

from generalizing a sum of Coulomb potentials for discrete charges in electrostatics to a continuous charge distribution.

For the condition $\rho = 0$ this reduces to an even more famous equation, the **Laplace equation**.

$$\boldsymbol{\nabla}\cdot\boldsymbol{\nabla}\varphi = 0. \tag{1.146}$$

We encounter Laplace's equation frequently in discussing various curved coordinate systems (Chapter 2) and the special functions of mathematical physics that appear as its solutions in Chapter 11.

From direct comparison of the Coulomb electrostatic force law and Newton's law of universal gravitation,

$$\mathbf{F}_E = \frac{1}{4\pi\varepsilon_0}\frac{q_1q_2}{r^2}\hat{\mathbf{r}}, \qquad \mathbf{F}_G = -G\frac{m_1m_2}{r^2}\hat{\mathbf{r}}.$$

All of the potential theory of this section therefore applies equally well to gravitational potentials. For example, the gravitational Poisson equation is

$$\nabla \cdot \nabla \varphi = +4\pi G\rho, \tag{1.147}$$

with ρ now a mass density.

Biographical Data

Poisson, Siméon Denis. Poisson, a French mathematician, was born in Pithiviers, France in 1781 and died in Paris in 1840. He studied mathematics at the Ecole Polytechnique under Laplace and Lagrange, whom he so impressed with his talent that he became professor there in 1802. He contributed to their celestial mechanics, Fourier's heat theory, and probability theory, among others.

EXERCISES

1.13.1 Develop Gauss's law for the two-dimensional case in which

$$\varphi = -q\frac{\ln \rho}{2\pi \varepsilon_0}, \quad \mathbf{E} = -\nabla\varphi = q\frac{\hat{\boldsymbol{\rho}}}{2\pi\varepsilon_0\rho},$$

where q is the charge at the origin or the line charge per unit length if the two-dimensional system is a unit thickness slice of a three-dimensional (circular cylindrical) system. The variable ρ is measured radially outward from the line charge. $\hat{\boldsymbol{\rho}}$ is the corresponding unit vector (see Section 2.2). If graphical software is available, draw the potential and field for the $q/2\pi\varepsilon_0 = 1$ case.

1.13.2 (a) Show that Gauss's law follows from Maxwell's equation

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0}$$

by integrating over a closed surface. Here, ρ is the charge density.

(b) Assuming that the electric field of a point charge q is spherically symmetric, show that Gauss's law implies the Coulomb inverse square expression

$$\mathbf{E} = \frac{q\hat{\mathbf{r}}}{4\pi\varepsilon_0 r^2}.$$

1.13.3 Show that the value of the electrostatic potential φ at any point P is equal to the average of the potential over any spherical surface centered on P. There are no electric charges on or within the sphere.

Hint. Use Green's theorem [Eq. (1.119)], with $u^{-1} = r$, the distance from P, and $v = \varphi$.

1.13.4 Using Maxwell's equations, show that for a system (steady current) the magnetic vector potential $\mathbf{A}$ satisfies a vector Poisson equation

$$\nabla^2\mathbf{A} = -\mu\mathbf{J},$$

provided we require $\nabla \cdot \mathbf{A} = 0$ in Coulomb gauge.

1.14 Dirac Delta Function

From Example 1.6.1 and the development of Gauss's law in Section 1.13,

$$\int \nabla \cdot \nabla \left(\frac{1}{r}\right) d\tau = -\int \nabla \cdot \left(\frac{\hat{\mathbf{r}}}{r^2}\right) d\tau = \begin{cases} -4\pi \\ 0 \end{cases}, \tag{1.148}$$

depending on whether the integration includes the origin $\mathbf{r} = 0$ or not. This result may be conveniently expressed by introducing the Dirac delta function,

$$\nabla^2 \left(\frac{1}{r}\right) = -4\pi\,\delta\,(\mathbf{r}) = -4\pi\,\delta(x)\delta(y)\delta(z). \tag{1.149}$$

This Dirac delta function is **defined** by its assigned properties

$$\delta\,(x) = 0, \qquad x \neq 0 \tag{1.150}$$

$$f\,(0) = \int_{-\infty}^{\infty} f\,(x)\,\delta(x)\,dx, \tag{1.151}$$

where $f(x)$ is any well-behaved function and the integration includes the origin. As a special case of Eq. (1.151),

$$\int_{-\infty}^{\infty} \delta(x)\,dx = 1. \tag{1.152}$$

From Eq. (1.151), $\delta(x)$ must be an infinitely high, infinitely thin spike at $x = 0$, as in the description of an impulsive force or the charge density for a point charge.[17] The problem is that **no such function exists** in the usual sense of function. However, the crucial property in Eq. (1.151) can be developed rigorously as the limit of a **sequence** of functions, a distribution. For example, the delta function may be approximated by the sequences of functions in n for $n \to \infty$ [Eqs. (1.153)–(1.156) and Figs. 1.44–1.47]:

$$\delta_n(x) = \begin{cases} 0, & x < -\frac{1}{2n} \\ n, & -\frac{1}{2n} < x < \frac{1}{2n} \\ 0, & x > \frac{1}{2n} \end{cases} \tag{1.153}$$

$$\delta_n(x) = \frac{n}{\sqrt{\pi}} \exp(-n^2x^2) \tag{1.154}$$

$$\delta_n(x) = \frac{n}{\pi} \cdot \frac{1}{1 + n^2x^2} \tag{1.155}$$

$$\delta_n(x) = \frac{\sin nx}{\pi x} = \frac{1}{2\pi} \int_{-n}^{n} e^{ixt}\,dt. \tag{1.156}$$

---

[17] The delta function is frequently invoked to describe very short-range forces such as nuclear forces. It also appears in the normalization of continuum wave functions of quantum mechanics.

Figure 1.44

δ Sequence Function Eq. (1.153)

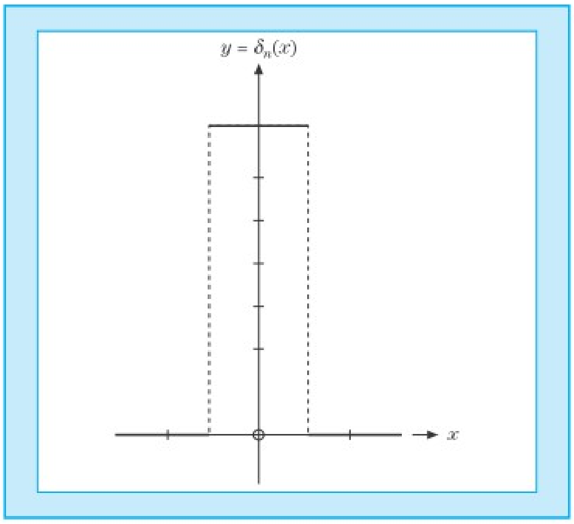

Figure 1.45

δ Sequence Function Eq. (1.154)

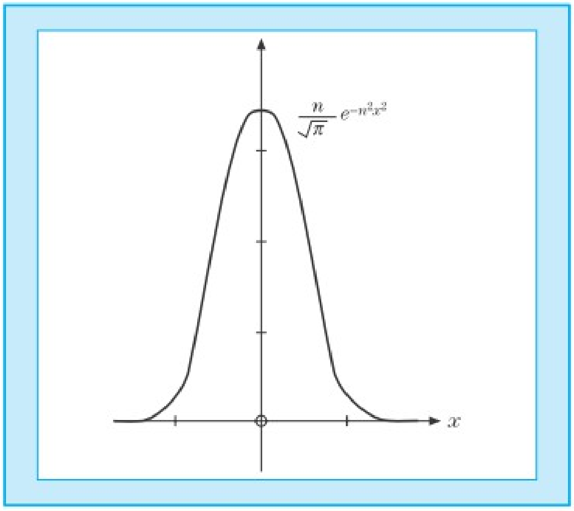

EXAMPLE 1.14.1 Let us evaluate $\int_{-\pi}^{\pi} \cos x\delta(x)dx = \cos 0 = 1$ using the sequence of Eq. (1.153). We find

$$\int_{-1/2n}^{1/2n} n\cos x\, dx = n\sin x|_{-1/2n}^{1/2n} = n\left(\sin\left(\frac{1}{2n}\right) - \sin\left(-\frac{1}{2n}\right)\right)$$

$$= 2n\sin\frac{1}{2n} = 2n\left(\frac{1}{2n} + O(1/n^3)\right) \to 1 \text{ for } n \to \infty.$$

Figure 1.46

δ Sequence Function Eq. (1.155)

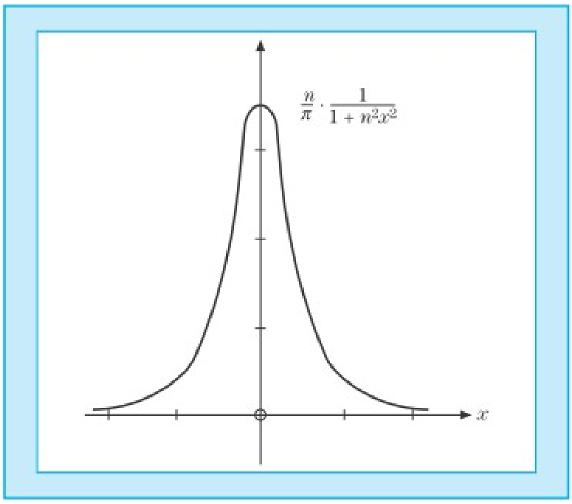

Figure 1.47

δ Sequence Function Eq. (1.156)

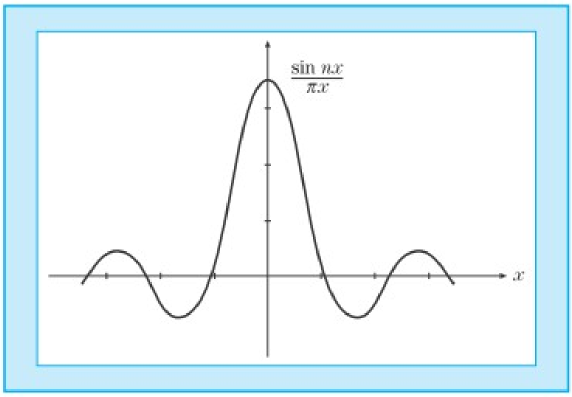

Notice how the integration limits change in the first step. Similarly, $\int_{-\pi}^{\pi} \sin x\delta(x) \cdot dx = \sin 0 = 0$. We could have used Eq. (1.155) instead,

$$\begin{aligned}
\frac{n}{\pi}\int_{-\pi}^{\pi}\frac{\cos x\,dx}{1+n^2x^2} &= \frac{n}{\pi}\int_{-\pi}^{\pi}\frac{1-x^2/2+\cdots}{1+n^2x^2}dx = \frac{n}{\pi}\int_{-\pi}^{\pi}\frac{dx}{1+n^2x^2}\\
&= \frac{1}{\pi}\int_{-n\pi}^{n\pi}\frac{dy}{1+y^2} = \frac{1}{\pi}[\arctan(n\pi) - \arctan(-n\pi)]\\
&= \frac{2}{\pi}\arctan(n\pi) \to \frac{2}{\pi}\frac{\pi}{2} = 1, \quad \text{for } n \to \infty,
\end{aligned}$$

by keeping just the first term of the power expansion of $\cos x$. Again, we could have changed the integration limits to $\pm\pi/n$ in the first step for all terms with positive powers of x because the denominator is so large, except close to $x = 0$ for large n. This explains why the higher order terms of the $\cos x$ power series do not contribute. ■

These approximations have varying degrees of usefulness. Equation (1.153) is useful in providing a simple derivation of the integral property [Eq. (1.151)]. Equation (1.154) is convenient to differentiate. Its derivatives lead to the Hermite polynomials. Equation (1.156) is particularly useful in Fourier analysis and in its applications to quantum mechanics. In the theory of Fourier series, Eq. (1.156) often appears (modified) as the Dirichlet kernel:

$$\delta_n(x) = \frac{1}{2\pi}\frac{\sin[(n+\frac{1}{2})x]}{\sin(\frac{1}{2}x)}. \tag{1.157}$$

In using these approximations in Eq. (1.151) and later, we assume that $f(x)$ is integrable—it offers no problems at large x.

For most physical purposes such approximations are quite adequate. From a mathematical standpoint, the situation is still unsatisfactory: The limits

$$\lim_{n\to\infty} \delta_n(x)$$

do not exist.

A way out of this difficulty is provided by the theory of distributions. Recognizing that Eq. (1.151) is the fundamental property, we focus our attention on it rather than on $\delta(x)$. Equations (1.153)–(1.156), with $n = 1, 2, 3, \ldots$, may be interpreted as **sequences** of normalized functions:

$$\int_{-\infty}^{\infty} \delta_n(x)\,dx = 1. \tag{1.158}$$

The sequence of integrals has the limit

$$\lim_{n\to\infty}\int_{-\infty}^{\infty} \delta_n(x) f(x)\,dx = f(0). \tag{1.159}$$

Note that Eq. (1.158) is the limit of a sequence of integrals. Again, the limit of $\delta_n(x)$, $n\to\infty$, does not exist. [The limits for all four forms of $\delta_n(x)$ diverge at $x = 0$.]

We may treat $\delta(x)$ consistently in the form

$$\int_{-\infty}^{\infty} \delta(x) f(x)\,dx = \lim_{n\to\infty}\int_{-\infty}^{\infty} \delta_n(x) f(x)\,dx. \tag{1.160}$$

$\delta(x)$ is labeled a distribution (not a function) defined by the sequences $\delta_n(x)$ as indicated in Eq. (1.158). We might emphasize that the integral on the left-hand side of Eq. (1.160) is not a Riemann integral.[18] It is a limit.

---

[18] It can be treated as a Stieltjes integral if desired. $\delta(x)\,dx$ is replaced by $du(x)$, where $u(x)$ is the Heaviside step function.

This distribution $\delta(x)$ is only one of an infinity of possible distributions, but it is the one we are interested in because of Eq. (1.151).

From these sequences of functions, we see that Dirac's delta function must be even in x, $\delta(-x) = \delta(x)$.

Let us now consider a detailed application of the Dirac delta function to a single charge and illustrate the singularity of the electric field at the origin.

EXAMPLE 1.14.2

Total Charge inside a Sphere Consider the total electric flux $\oint \mathbf{E} \cdot d\boldsymbol{\sigma}$ out of a sphere of radius R around the origin surrounding n charges e_j located at the points $\mathbf{r}_j$ with $r_j < R$ (i.e., inside the sphere). The electric field strength $\mathbf{E} = -\boldsymbol{\nabla}\varphi(\mathbf{r})$, where the potential

$$\varphi = \sum_{j=1}^{n} \frac{e_j}{|\mathbf{r} - \mathbf{r}_j|} = \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d^3r'$$

is the sum of the Coulomb potentials generated by each charge and the total charge density is $\rho(\mathbf{r}) = \sum_j e_j \delta(\mathbf{r} - \mathbf{r}_j)$. The delta function is used here as an abbreviation of a pointlike density. Now we use Gauss's theorem for

$$\oint \mathbf{E} \cdot d\boldsymbol{\sigma} = -\oint \boldsymbol{\nabla}\varphi \cdot d\boldsymbol{\sigma} = -\int \boldsymbol{\nabla}^2 \varphi d\tau = \int \frac{\rho(\mathbf{r})}{\varepsilon_0} d\tau = \frac{\sum_j e_j}{\varepsilon_0}$$

in conjunction with the differential form of Gauss's law $\boldsymbol{\nabla} \cdot \mathbf{E} = -\rho/\varepsilon_0$ and $\sum_j e_j \int \delta(\mathbf{r} - \mathbf{r}_j) d\tau = \sum_j e_j$. ■

The integral property [Eq. (1.151)] is useful in cases in which the argument of the delta function is a function $g(x)$ with **simple zeros** on the real axis, which leads to the rules

$$\delta(ax) = \frac{1}{a}\delta(x), \quad a > 0, \tag{1.161}$$

$$\delta(g(x)) = \sum_{\substack{a, \\ g(a)=0, \\ g'(a)\neq 0}} \frac{\delta(x-a)}{|g'(a)|}. \tag{1.162}$$

To obtain Eq. (1.161) we change the integration variable in

$$\int_{-\infty}^{\infty} f(x)\delta(ax)dx = \frac{1}{a}\int_{-\infty}^{\infty} f\left(\frac{y}{a}\right)\delta(y)dy = \frac{1}{a}f(0)$$

and apply Eq. (1.151). To prove Eq. (1.162) , we decompose the integral

$$\int_{-\infty}^{\infty} f(x)\delta(g(x))dx = \sum_a \int_{a-\varepsilon}^{a+\varepsilon} f(x)\delta((x-a)g'(a))dx \tag{1.163}$$

into a sum of integrals over small intervals containing the first-order zeros of $g(x)$. In these intervals, $g(x) \approx g(a) + (x-a)g'(a) = (x-a)g'(a)$. Using Eq. (1.161) on the right-hand side of Eq. (1.163), we obtain the integral of Eq. (1.162).

EXAMPLE 1.14.3

Evaluate $I \equiv \int_{-\infty}^{\infty} f(x)\delta(x^2-2)dx$ Because the zeros of the argument of the delta function, $x^2 = 2$, are $x = \pm\sqrt{2}$, we can write the integral as a sum of two contributions:

$$\begin{aligned} I &= \int_{\sqrt{2}-\epsilon}^{\sqrt{2}+\epsilon} \delta(x-\sqrt{2}) \frac{f(x)dx}{\frac{d(x^2-2)}{dx}|_{x=\sqrt{2}}} dx + \int_{-\sqrt{2}-\epsilon}^{-\sqrt{2}+\epsilon} \delta(x+\sqrt{2}) \frac{f(x)dx}{\frac{d(x^2-2)}{dx}|_{x=-\sqrt{2}}} \\ &= \int_{\sqrt{2}-\epsilon}^{\sqrt{2}+\epsilon} \delta(x-\sqrt{2}) \frac{f(x)dx}{2\sqrt{2}} + \int_{-\sqrt{2}-\epsilon}^{-\sqrt{2}+\epsilon} \delta(x+\sqrt{2}) \frac{f(x)dx}{2\sqrt{2}} \\ &= \frac{f(\sqrt{2}) + f(-\sqrt{2})}{2\sqrt{2}}. \end{aligned}$$ ■

This example is good training for the following one.

EXAMPLE 1.14.4

Phase Space In the scattering theory of relativistic particles using Feynman diagrams, we encounter the following integral over energy of the scattered particle (we set the velocity of light $c = 1$):

$$\begin{aligned} \int d^4p\,\delta(p^2 - m^2) f(p) &\equiv \int d^3p \int dp_0 \delta\left(p_0^2 - \mathbf{p}^2 - m^2\right) f(p) \\ &= \int_{E>0} \frac{d^3p f(E, \mathbf{p})}{2\sqrt{m^2 + \mathbf{p}^2}} - \int_{E<0} \frac{d^3p f(E, \mathbf{p})}{2\sqrt{m^2 + \mathbf{p}^2}}, \end{aligned}$$

where we have used Eq. (1.162) at the zeros $E = \pm\sqrt{m^2 + \mathbf{p}^2}$ of the argument of the delta function. The physical meaning of $\delta(p^2 - m^2)$ is that the particle of mass m and four-momentum $p^\mu = (p_0, \mathbf{p})$ is on its mass shell because $p^2 = m^2$ is equivalent to $E = \pm\sqrt{m^2 + \mathbf{p}^2}$. Thus, the on-mass-shell volume element in momentum space is the Lorentz invariant $\frac{d^3p}{2E}$, in contrast to the nonrelativistic d^3p of momentum space. The fact that a negative energy occurs is a peculiarity of relativistic kinematics that is related to the antiparticle. ■

Using integration by parts we can also **define the derivative** $\delta'(x)$ of the Dirac delta function by the relation

$$\int_{-\infty}^{\infty} f(x)\delta'(x - x')\,dx = -\int_{-\infty}^{\infty} f'(x)\delta(x - x')\,dx = -f'(x'). \qquad (1.164)$$

It should be understood that our Dirac delta function has significance only as part of an integrand. Thus, the Dirac delta function is often regarded as a linear operator: $\delta(x - x_0)$ operates on $f(x)$ and yields $f(x_0)$:

$$\mathcal{L}(x_0) f(x) \equiv \int_{-\infty}^{\infty} \delta(x - x_0) f(x)\,dx = f(x_0). \qquad (1.165)$$

It may also be classified as a linear mapping or simply as a generalized function. Shifting our singularity to the point $x = x'$, we write the Dirac delta function as $\delta(x - x')$. Equation (1.151) becomes

$$\int_{-\infty}^{\infty} f(x)\delta(x - x')\,dx = f(x'). \tag{1.166}$$

As a description of a singularity at $x = x'$, the Dirac delta function may be written as $\delta(x - x')$ or as $\delta(x' - x)$. Expanding to three dimensions and using spherical polar coordinates, we obtain

$$\begin{aligned} f(0) &= \int_0^{2\pi}\int_0^{\pi}\int_0^{\infty} f(\mathbf{r})\delta(\mathbf{r})\,r^2\,dr\sin\theta\,d\theta\,d\varphi \\ &= \iiint_{-\infty}^{\infty} f(x, y, z)\delta(x)\,\delta(y)\,\delta(z)\,dx\,dy\,dz, \\ &\int_0^{\infty}\frac{\delta(r)}{r^2}r^2 dr\int_{-1}^{1}\delta(\cos\theta)\,d\cos\theta\int_0^{2\pi}\delta(\varphi)\,d\varphi = 1, \end{aligned} \tag{1.167}$$

where each one-dimensional integral is equal to unity. This corresponds to a singularity (or source) at the origin. Again, if our source is at $\mathbf{r} = \mathbf{r}_1$, Eq. (1.167) generalizes to

$$\iiint f(\mathbf{r}_2)\delta(\mathbf{r}_2 - \mathbf{r}_1)\,r_2^2 dr_2\sin\theta_2\,d\theta_2\,d\varphi_2 = f(\mathbf{r}_1), \tag{1.168}$$

where

$$\int_0^{\infty}\frac{\delta(r_2 - r_1)}{r_2^2}r_2^2 dr_2\int_{-1}^{1}\delta(\cos\theta_2 - \cos\theta_1)d\cos\theta_2\int_0^{2\pi}\delta(\varphi_2 - \varphi_1)d\varphi_2 = 1.$$

SUMMARY

We use $\delta(x)$ frequently and call it the Dirac delta function—for historical reasons.[19] Remember that it is not really a function. It is essentially a shorthand notation, defined implicitly as the limit of integrals in a sequence, $\delta_n(x)$, according to Eq. (1.160).

Biographical Data

Dirac, Paul Adrien Maurice. Dirac, an English physicist, was born in Bristol in 1902 and died in Bristol in 1984. He obtained a degree in electrical engineering at Bristol and obtained his Ph.D. in mathematical physics in 1926 at Cambridge. By 1932, he was Lucasian professor, like Stokes, the chair Newton once held. In the 1920s, he advanced quantum mechanics, became one of the founders of quantum field theory, and, in 1928, discovered his relativistic equation for the electron that predicted antiparticles for which he was awarded the Nobel prize in 1933.

---

[19] Dirac introduced the delta function to quantum mechanics. Actually, the delta function can be traced back to Kirchhoff, 1882. For further details, see M. Jammer (1966). *The Conceptual Development of Quantum Mechanics*, p. 301. McGraw-Hill, New York.

E 4)
Y) % RRR , % j 4 Q 4))) 5) Q)

!"#' $ % % & ' ( )) * +) , -)
.)))) - 2 !.! % 3 4
*) 68 ( ) N<8 1 9 <<8 % 7T
5- < 2 <3 "("OO@O #1'# "! E# " , ,$ @(@$!E YO"(

A startling consequence of Theorem 6.2 is that in hyperbolic geometry a segment can be determined with the aid of an angle; for example, an angle of an equilateral triangle determines the length of a side uniquely. This is sometimes stated more dramatically by saying that hyperbolic geometry has an *absolute unit of length* (see Gauss' letter to Taurinus, quoted earlier in this chapter). If the geometry of the physical universe were hyperbolic, it would no longer be necessary to keep a unit of length carefully guarded in the Bureau of Standards (the same is true for elliptic geometry).

PARALLELS THAT ADMIT A COMMON PERPENDICULAR

In Chapter 5, commenting on the flaw in Proclus' attempted proof of the parallel postulate, I remarked that it was presumptuous to assume that parallel lines looked like railroad tracks, that is, that they were everywhere equidistant from each other. Let us now make this remark more precise. Given lines l and l' and points A, B, C, . . . on l. Drop perpendiculars AA′, BB′, CC′, . . . from these points to l'. We will say that points A, B, C, . . . are *equidistant* from l' if all these perpendicular segments are congruent to one another (Figure 6.5).

THEOREM 6.3. In hyperbolic geometry if l and l' are any distinct parallel lines, then any set of points on l equidistant from l' has at most two points in it.

Proof:
Assume on the contrary there is a set of three points A, B, and C on l equidistant from l'. Then quadrilaterals □A′B′BA, □A′C′CA,

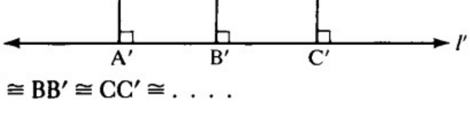

FIGURE 6.5 AA′ ≅ BB′ ≅ CC′ ≅ . . . .

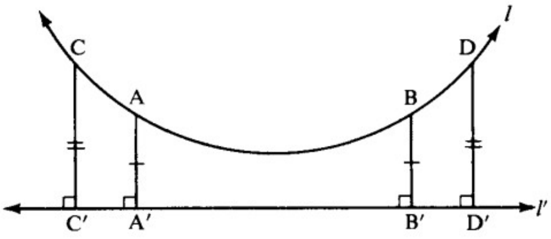

FIGURE 6.6

and □B′C′CB are Saccheri quadrilaterals (the base angles are right angles and the sides are congruent). In Exercise 1, Chapter 5, you showed that the summit angles of a Saccheri quadrilateral are congruent. Thus, ∢A′AB ≅ ∢B′BA, ∢A′AC ≅ ∢C′CA, and ∢B′BC ≅ ∢C′CB. By transitivity (Axiom C-5), it follows that the supplementary angles ∢B′BA and ∢B′BC are congruent to each other; hence, by definition, they are right angles. Therefore, these Saccheri quadrilaterals are all rectangles. But rectangles do not exist in hyperbolic geometry (Lemma 6.1). This contradiction shows that A, B, and C cannot be equidistant from *l*′. ■

The theorem states that *at most* two points at a time on *l* can be equidistant from *l*′. It allows the possibility that there are pairs of points (A, B), (C, D), . . . on *l* such that each pair is equidistant from *l*′—thus, dropping perpendiculars, AA′ ≅ BB′ and CC′ ≅ DD′, but AA′ is not congruent to CC′. A diagram for this might be Figure 6.6, which suggests that the point "in the middle" of *l* is closest to *l*′, with *l* moving away from *l*′ symmetrically on either side of this middle point. We will prove that this is indeed the case (Theorems 6.4 and 6.5 and Exercises 4 and 10).

Note, however, that Theorem 6.3 allows another possibility, that there is no pair of points on *l* equidistant from *l*′! An diagram for this might be Figure 6.7, where the points on *l* are at varying distances from *l*′; *l* moves away from *l*′ in one direction and approaches *l*′ in the other direction without ever meeting it. Thus, different pairs of parallel lines need not resemble each other—some may look like the first diagram, some like the second.

THEOREM 6.4. In hyperbolic geometry if *l* and *l*′ are parallel lines for which there exists a pair of points A and B on *l* equidistant from *l*′, then

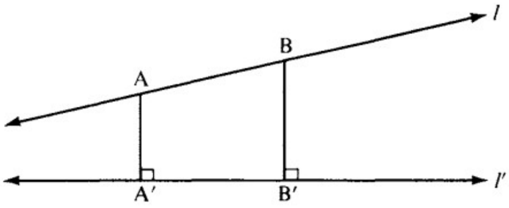

FIGURE 6.7 BB′ > AA′.

l and *l*′ have a common perpendicular segment that is also the shortest segment between *l* and *l*′.

Proof:
Suppose A and B on *l* are equidistant from *l*′. Then □A′B′BA is a Saccheri quadrilateral, where A′ and B′ are the feet on *l*′ of the perpendiculars from A and B. Let M be the midpoint of AB and M′ the midpoint of A′B′ (Proposition 4.3; see Figure 6.8). The theorem will follow from the next lemma. ■

LEMMA 6.2. The segment joining the midpoints of the base and summit of a Saccheri quadrilateral is perpendicular to both the base and the summit, and this segment is shorter than the sides (see Figure 6.8).

Proof:
We know that ∢A ≅ ∢B (Exercise 1, Chapter 5). Hence, ΔA′AM ≅ ΔB′BM (SAS). Therefore, the corresponding sides A′M and B′M are congruent. This implies ΔA′M′M ≅ ΔB′M′M (SSS). Therefore, the corresponding angles ∢A′M′M and ∢B′M′M are congruent. Since these are supplementary angles, they must be right angles, proving MM′ perpendicular to the base

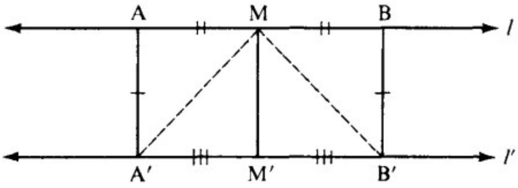

FIGURE 6.8

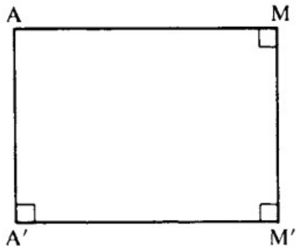

FIGURE 6.9

A′B′. From the two pairs of congruent triangles, we also have ∢A′MM′ ≅ ∢B′MM′ and ∢A′MA ≅ ∢B′MB. Adding the degrees of these angles, we have $(\sphericalangle AMM')^\circ = (\sphericalangle BMM')^\circ$ (Theorem 4.3(3)), i.e., the supplementary angles ∢AMM′ and ∢BMM′ have the same number of degrees. Hence, they are right angles and MM′ is also perpendicular to the summit AB.

Consider next quadrilateral □A′M′MA (Figure 6.9). It has three right angles, so it is what we call a Lambert quadrilateral (Exercise 4, Chapter 5). In hyperbolic geometry the fourth angle must be acute, since rectangles do not exist (Lemma 6.1). You showed in Exercise 4(c), Chapter 5, that AA′ > MM′, i.e., that MM′ is shorter than AA′. The remainder of the proof that MM′ is shorter than any other segment between *l* and *l′* is left for Exercise 3. ■

THEOREM 6.5. In hyperbolic geometry if lines *l* and *l′* have a common perpendicular segment MM′, then they are parallel and MM′ is unique. Moreover, if A and B are any points on *l* such that M is the midpoint of segment AB, then A and B are equidistant from *l′*.

Proof:
The fact that *l* and *l′* are parallel follows from the first corollary to the alternate interior angle theorem (Theorem 4.1). If *l* and *l′* had another common perpendicular segment NN′, then □M′N′NM would be a rectangle, which cannot exist (Lemma 6.1). Suppose now that M is the midpoint of AB. Drop perpendiculars AA′ and BB′ to *l*. We must prove that AA′ ≅ BB′. (See Figure 6.10.) First, ΔAM′M ≅ ΔBM′M (SAS), AM′ ≅ BM′, and ∢AM′M ≅ ∢BM′M. Therefore, $(\sphericalangle A'M'A)^\circ = 90^\circ - (\sphericalangle AM'M)^\circ = 90^\circ -$

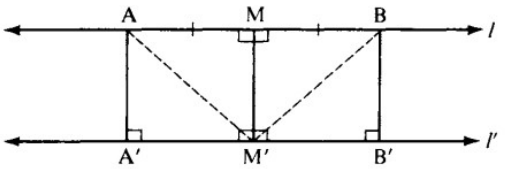

FIGURE 6.10

$(\sphericalangle BM'M)^\circ = (\sphericalangle B'M'B)^\circ$ (by Theorem 4.3) so that $\sphericalangle A'M'A \cong \sphericalangle B'M'B$. Hence, $\triangle AA'M' \cong \triangle BB'M'$ (AAS), so that the corresponding sides AA′ and BB′ are congruent. ■

LIMITING PARALLEL RAYS

Theorems 6.4 and 6.5 and Exercises 4 and 10 give us a good understanding of the first type of parallel lines. We know that such lines actually exist from the usual construction: start with any line l and any point P not on it (Figure 6.11). Drop perpendicular $\overleftrightarrow{PQ}$ to l and let m be the perpendicular through P to $\overleftrightarrow{PQ}$. Then m and l have the common perpendicular segment PQ. Pairs of points on m situated symmetrically about $\overleftrightarrow{PQ}$ are equidistant from l. By the universal hyperbolic theorem, there exist other lines n through P parallel to l. We cannot yet say that any such n is the second type of parallel, for n and l might have a common perpendicular going through a point other than P.

How then do we know that parallels of the second type exist? Here the axioms of continuity come in. The following is the intuitive idea (see Figure 6.12). Consider one ray $\overrightarrow{PS}$ of m, and consider various rays between $\overrightarrow{PS}$ and $\overrightarrow{PQ}$. Some of these rays, such as $\overrightarrow{PR}$, will intersect l; others, such as $\overrightarrow{PY}$, will not. A continuity argument shows that as R recedes endlessly on l from Q, $\overrightarrow{PR}$ will approach a certain limiting ray

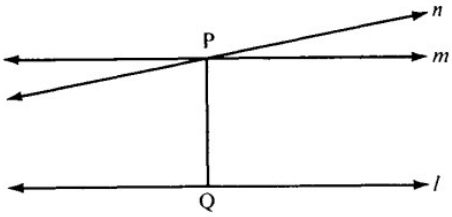

FIGURE 6.11

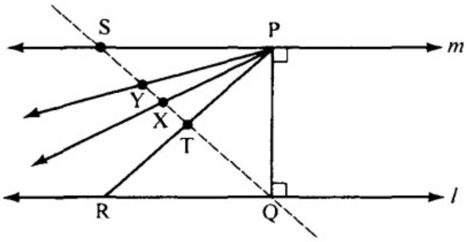

FIGURE 6.12

$\overrightarrow{PX}$ that does *not* meet l. The ray $\overrightarrow{PX}$ is "limiting" in the following precise sense: any ray between $\overrightarrow{PX}$ and $\overrightarrow{PQ}$ intersects l, whereas any other ray $\overrightarrow{PY}$ such that $\overrightarrow{PX}$ is between $\overrightarrow{PY}$ and $\overrightarrow{PQ}$ does not intersect l. The ray $\overrightarrow{PX}$ may be called the *left limiting parallel ray* to l through P. Similarly, there is a *right limiting parallel ray* on the opposite side of $\overrightarrow{PQ}$.

THEOREM 6.6. For every line l and every point P not on l, let Q be the foot of the perpendicular from P to l. Then there are two unique nonopposite rays $\overrightarrow{PX}$ and $\overrightarrow{PX'}$ on opposite sides of $\overleftrightarrow{PQ}$ that do not meet l and have the property that a ray emanating from P meets l if and only if it is between $\overrightarrow{PX}$ and $\overrightarrow{PX'}$. Moreover, these limiting rays are situated symmetrically about $\overleftrightarrow{PQ}$ in the sense that $\sphericalangle XPQ \cong \sphericalangle X'PQ$.

Proof:
To prove rigorously that $\overrightarrow{PX}$ exists, consider the line $\overleftrightarrow{SQ}$ (Figure 6.12). Let Σ_1 be the set of all points T on segment SQ such that $\overrightarrow{PT}$ meets l, together with all points on the ray opposite to $\overrightarrow{QS}$; let Σ_2 be the complement of Σ_1 (so $Q \in \Sigma_1$ and $S \in \Sigma_2$). By the crossbar theorem (Chapter 3), if point T on segment SQ belongs to Σ_1, then the entire segment TQ (in fact, $\overrightarrow{TQ}$) is contained in Σ_1. Hence, (Σ_1, Σ_2) is a Dedekind cut. By Dedekind's axiom (Chapter 3), there is a unique point X on $\overleftrightarrow{SQ}$ such that for P_1 and P_2 on $\overleftrightarrow{SQ}$, $P_1 * X * P_2$ if and only if $X \neq P_1$ or P_2, $P_1 \in \Sigma_1$, and $P_2 \in \Sigma_2$.

By definition of Σ_1 and Σ_2, rays below $\overrightarrow{PX}$ all meet l and rays above $\overrightarrow{PX}$ do not. We claim that $\overrightarrow{PX}$ does not meet l either. Assume on the contrary that $\overrightarrow{PX}$ meets l in a point U (Figure 6.13). Choose any point V on l to the left of U, i.e., $V * U * Q$ (Axiom B-2). Since V and U are on the same side of $\overleftrightarrow{SQ}$ (Exercise 9, Chapter 3), V and P are on opposite sides, so VP meets SQ in a point Y. We have

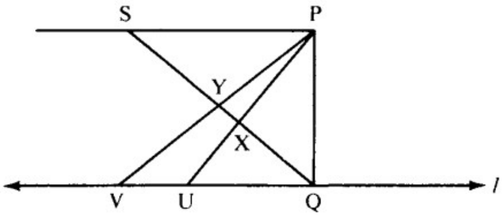

FIGURE 6.13

$Y * X * Q$ (Proposition 3.7), so $Y \in \Sigma_2$, contradicting the fact that $\overrightarrow{PY}$ meets l. It follows that $\overrightarrow{PX}$ is the left limiting parallel ray (we obtain the right limiting parallel ray in a similar manner).

To prove symmetry, assume on the contrary that angles $\sphericalangle XPQ$ and $\sphericalangle X'PQ$ are not congruent, e.g., $(\sphericalangle XPQ)^\circ < (\sphericalangle X'PQ)^\circ$. By Axiom C-4, there is a ray between $\overrightarrow{PX'}$ and $\overrightarrow{PQ}$ that intersects l (by definition of limiting ray) in a point R′ such that $\sphericalangle R'PQ \cong \sphericalangle XPQ$ (Figure 6.14). Let R be the point on the opposite side of $\overleftrightarrow{PQ}$ from R′ such that $R * Q * R'$ and $RQ \cong R'Q$ (Axiom C-1). Then $\Delta RPQ \cong \Delta R'PQ$ (SAS). Hence, $\sphericalangle RPQ \cong \sphericalangle R'PQ$, and by transitivity (Axiom C-5), $\sphericalangle RPQ \cong \sphericalangle XPQ$. But this is impossible, because $\overrightarrow{PR}$ is between $\overrightarrow{PX}$ and $\overrightarrow{PQ}$ (Axiom C-4). ■

Either of the congruent angles $\sphericalangle XPQ$ and $\sphericalangle X'PQ$ is called (by an abuse of language) the *angle of parallelism* at point P with respect to l. Its degree measure is usually denoted $\Pi(PQ)^\circ$. Note that $\Pi(PQ)^\circ < 90^\circ$, for $\Pi(PQ)^\circ = 90^\circ$ would contradict the universal hyperbolic theorem (see Exercise 7(a)). It can be shown that as P varies, $\Pi(PQ)^\circ$ takes on all possible values between 0° and 90° (see Major Exercise 9). One of the greatest discoveries by J. Bolyai and Lobachevsky is their formula for this number of degrees (see

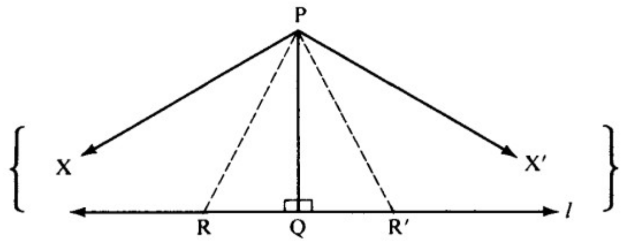

FIGURE 6.14

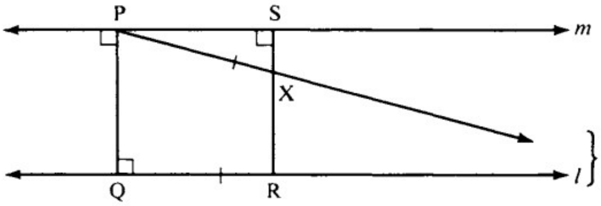

FIGURE 6.15

Theorem 10.2 in Chapter 10). A *natural unit segment* OI in hyperbolic geometry is any segment OI such that $\Pi(OI)^\circ = 45^\circ$. Major Exercise 5 shows that all such segments are congruent to each other.

We have proved the existence of limiting parallel rays by a continuity argument conforming to today's standards of rigor. Gauss, J. Bolyai, and Lobachevsky took this existence for granted, but J. Bolyai discovered a simple *straightedge-and-compass construction* for the limiting rays. Let Q be the foot of the perpendicular from P to *l*, *m* the line through P perpendicular to $\overleftrightarrow{PQ}$, R any point on *l* different from Q, and S the foot of the perpendicular from R to *m* (Figure 6.15). Then PR > QR (Exercise 3) and PS < QR (Exercise 4(c), Chapter 5, on Lambert quadrilaterals). By the elementary continuity principle, a compass with center P and radius congruent to QR will intersect segment SR in a unique point X between S and R. It can be proved that $\overrightarrow{PX}$ is the right limiting parallel ray to *l* through P! (The proof is complicated. See p. 269 in Chapter 7, Project 4 in this chapter, or Theorem 10.9, Chapter 10.)

CLASSIFICATION OF PARALLELS

We have discussed two types of parallels to a given *l*. The first type consists of parallels *m* such that *l* and *m* have a common perpendicular; *m* diverges from *l* on both sides of the common perpendicular. The second type consists of parallels that approach *l* asymptotically in one direction (i.e., they contain a limiting parallel ray in that direction) and which diverge from *l* in the other direction. If *m* is the second type of parallel, Exercises 7 and 8 show that *l* and *m* do not have a common perpendicular. We have implied that these two are the only types of parallels, and this is the content of the next theorem.

THEOREM 6.7. Given m parallel to l such that m does not contain a limiting parallel ray to l in either direction. Then there exists a common perpendicular to m and l (which is unique by Theorem 6.5).

This theorem is proved by Borsuk and Szmielew (1960, p. 291) by a continuity argument, but their proof gives you no idea of how to actually find the common perpendicular. There is an easy way to find it in the Klein and Poincaré models, discussed in the next chapter. Hilbert gave a direct construction, which we will sketch. (Project 1 gives another.)

Proof:
Hilbert's idea is to find two points H and K on l that are equidistant from m, for once these are found, the perpendicular bisector of segment HK is also perpendicular to m (see Lemma 6.2). Choose any two points A and B on l and suppose that the perpendicular segment AA′ from A to m is longer than the perpendicular segment BB′ from B to m. (See Figure 6.16.) Let E be the point between A′ and A such that A′E ≅ B′B. On the same side of $\overleftrightarrow{AA'}$ as B, let $\overrightarrow{EF}$ be the unique ray such that ∢A′EF ≅ ∢B′BG, where A * B * G. The key point that will be proved in Major Exercises 2–6 is that $\overrightarrow{EF}$ intersects $\overrightarrow{AG}$ in a point H. Let K be the unique point on $\overrightarrow{BG}$ such that EH ≅ BK. Drop perpendiculars $\overleftrightarrow{HH'}$ and $\overleftrightarrow{KK'}$ to m. The upshot of these constructions is that □EHH′A′ is congruent to □BKK′B′ (just divide them into triangles). Hence, the corresponding sides HH′ and KK′ are congruent, so that the points H and K on l are equidistant from m, as required. ■

To sum up, given a point P not on l, there exist exactly two limiting parallel rays to l through P, one in each direction. There are infinitely many lines through P that do not enter the region between the limiting

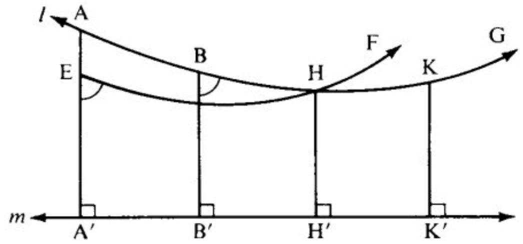

FIGURE 6.16

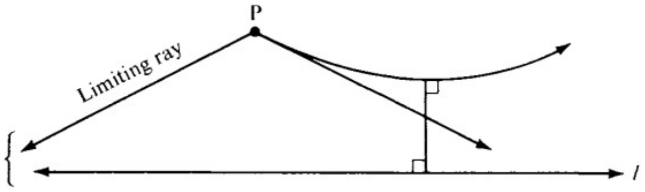

FIGURE 6.17

rays and l. Each such line is divergently parallel to l and admits a unique common perpendicular with l (for one of these lines the common perpendicular will go through P, but for all the rest the common perpendicular will pass through other points).

A Note on Terminology. In most books on hyperbolic geometry the word "parallel" is used only for lines that contain limiting parallel rays in our sense. The other lines, which admit a common perpendicular, have various names in the literature: "nonintersecting," "ultraparallel," "hyperparallel," and "superparallel." We will continue to use the word "parallel" to mean "nonintersecting." A parallel to l that contains a limiting parallel ray to l will be called an *asymptotic parallel,* and a parallel to l that admits a common perpendicular to l will be called a *divergently parallel line.* Rays that are limiting parallel will be denoted by a brace in diagrams (see Figure 6.17).

STRANGE NEW UNIVERSE?

In this chapter we have only begun to investigate the "strange new universe" of hyperbolic geometry. You can develop more of this geometry by doing the exercises, reading Chapter 10, and examining works in the bibliography at the end of the book. You will encounter new entities such as asymptotic triangles, ideal and ultra-ideal points, equidistant curves, horocycles, and pseudospheres.

If you consider this geometry too "far out" to pursue, you are in for a surprise. We will see in the next chapter that if the undefined terms of hyperbolic geometry are suitably interpreted, hyperbolic geometry can be considered a part of Euclidean geometry!

Meanwhile, notice how we have deepened our understanding of the role of the parallel postulate *P* in Euclidean geometry. Any statement *S* in the language of our geometry that is a theorem in Euclidean geometry ($P \Rightarrow S$) and whose negation is a theorem in hyperbolic geometry ($\sim P \Rightarrow \sim S$) is equivalent (in neutral geometry) to the parallel postulate (by Logic Rule 9(c)). For example, by Exercise 14, Chapter 5, "The angle sum of every triangle is 180°" is such a statement. By Exercise 12 in this chapter, "Every triangle has a circumscribed circle" is another such statement. By Theorem 5.1, "Every point interior to an acute angle lies on a line intersecting both sides of the angle in two distinct points" is a third such statement. I leave the enjoyment of providing a long list of such statements to you (Exercise 15). I urge your instructor to give a prize to the student(s) with the longest list.

REVIEW EXERCISE

Which of the following statements are correct?

(1) The negation of Hilbert's parallel postulate states that for every line *l* and every point P not on *l* there exist at least two lines through P parallel to *l*.

(2) It is a theorem in neutral geometry that if lines *l* and *m* meet on a given side of a transversal *t*, then the sum of the degrees of the interior angles on that given side of *t* is less than 180°.

(3) Gauss began working on non-Euclidean geometry when he was 15 years old.

(4) The philosopher Kant taught that our minds could not conceive of any geometry other than Euclidean geometry.

(5) The first mathematician to publish an account of hyperbolic geometry was the Russian Lobachevsky.

(6) The crossbar theorem asserts that a ray emanating from a vertex A of $\triangle ABC$ and interior to $\sphericalangle A$ must intersect the opposite side BC of the triangle.

(7) It is a theorem in hyperbolic geometry that for any segment AB there exists a square having AB as one of its sides.

(8) Every Saccheri quadrilateral is a convex quadrilateral.

(9) In hyperbolic geometry if $\triangle ABC$ and $\triangle DEF$ are equilateral triangles and $\sphericalangle A \cong \sphericalangle D$, then the triangles are congruent.

E)-

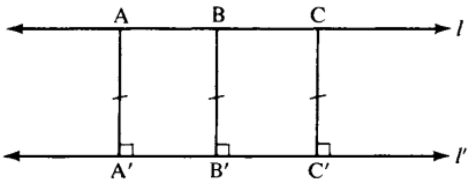

. - N /

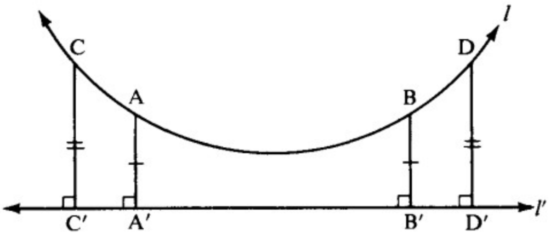
l
C
D
A
B
C′
A′
B′
D′
l′

. - N N

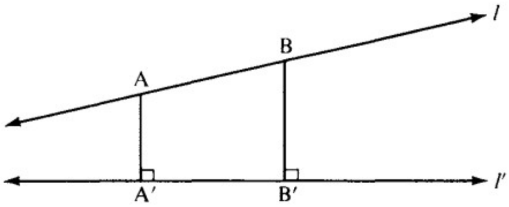

. - N 7

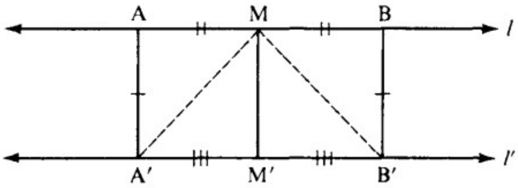

. - N T

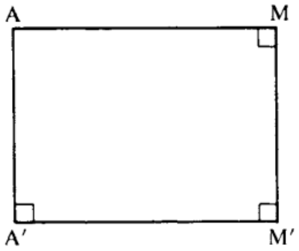

. - N <

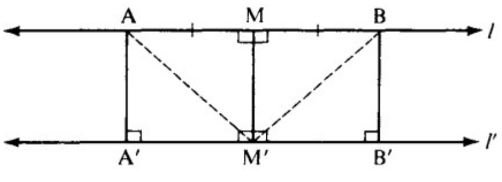

. - N

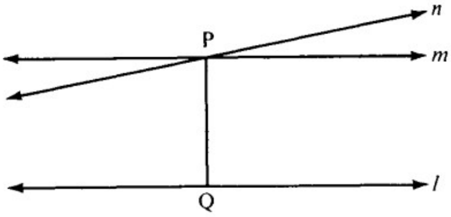

. - N

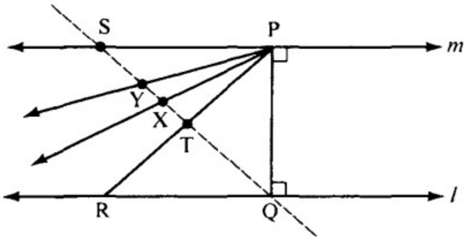

. - N

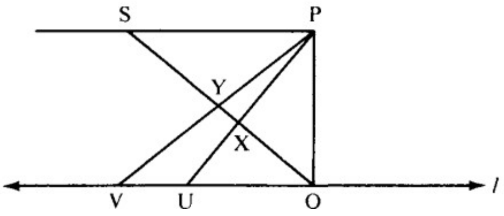

. - N

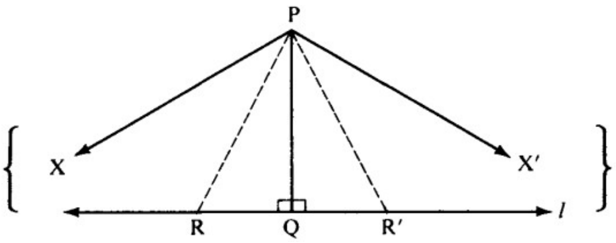

. - N 6

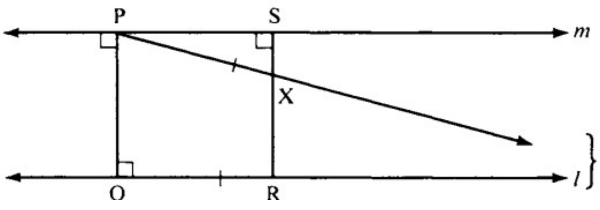

. - N /

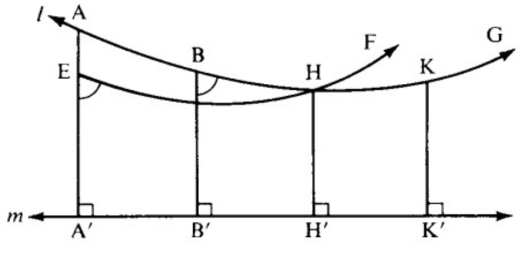
l
A
E
B
F
H
K
G
m
A′
B′
H′
K′

. - N N

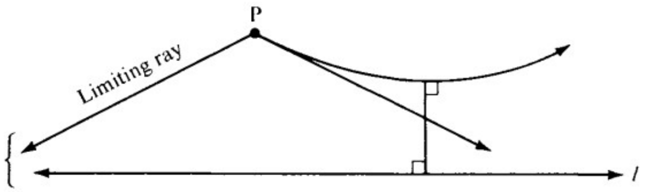
P
Limiting ray
l

. - N 7

CHAPTER 7

INDEPENDENCE OF THE PARALLEL POSTULATE

> All my efforts to discover a contradiction, an inconsistency, in this non-Euclidean geometry have been without success. . . .
>
> C. F. GAUSS

CONSISTENCY OF HYPERBOLIC GEOMETRY

In the previous chapter you were introduced to hyperbolic geometry and presented with some theorems that must seem very strange to someone accustomed to Euclidean geometry. Even though you may admit that the proofs of these theorems are correct, given our assumptions, you may feel that the basic assumption of hyperbolic geometry — the hyperbolic axiom — is a false assumption. Let's examine what might be meant by saying it's false.

Suppose I assume that when I drop some object, say, a stone, it will "fall" upward. I can go out and drop rocks and, unless I have rocks in my head, I will discover that my assumption was false.

Now what sort of experiment could I perform to show that the hyperbolic assumption is false, or, equivalently, to show that its negation, the Euclidean parallel postulate, is true? First of all, I would have to understand what this statement means. In the above example I understood very well the meaning of "stone" and what it means to "drop" one, so I could act upon this understanding. But what does it mean that *l* is a "line," that P is a "point" not "on" *l*, or that there is a "unique parallel" to *l* through P? I might represent "points" and

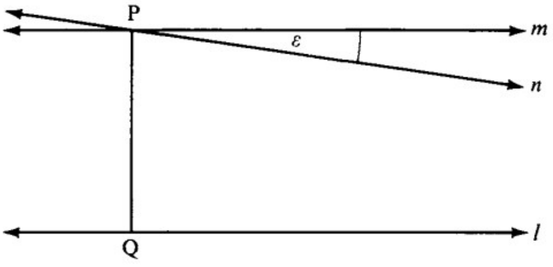

FIGURE 7.1

"lines" with paper, pencil, and straightedge. Suppose I draw $\overleftrightarrow{PQ}$ perpendicular to *l* and *m* through P perpendicular to $\overleftrightarrow{PQ}$, and then draw a line *n* through P, making a very small angle of $\epsilon°$ with *m*. Using Euclidean trigonometry, I can calculate exactly how far out on *n* I would have to go to get to the point where *n* is supposed to intersect *l*, but if ϵ is small enough, that point might be millions of miles away. Thus, I could not physically perform the experiment to prove that the hyperbolic axiom is false.

But is geometry about lines that we can draw? Applied geometry (engineering) is; but pure geometry is about ideal lines, which are concepts, not objects. The only experiments we can perform on these ideal lines are thought-experiments. So the question should be: Can we *conceive* of a non-Euclidean geometry? Kant said no, that any geometry other than Euclidean is inconceivable. At the time, of course, no one had yet conceived of a different geometry. It is in this sense that Gauss, J. Bolyai, and Lobachevsky created a "new universe."

Other questions can be raised. Mathematicians reject many of their own ideas because they either lead to contradictions or do not lead anywhere, i.e., do not prove fruitful, useful, or interesting. Does the hyperbolic axiom lead to a contradiction? Saccheri thought it would, and tried to prove the parallel postulate that way. Is hyperbolic geometry fruitful, useful, or interesting?

Let us postpone this question and take up the former: Is hyperbolic geometry *consistent?* Nowadays we refer to this as a question in *metamathematics,* i.e., a question outside of a mathematical system about the system itself. The question is not about lines or points or other geometric entities; it is a question about the *whole system* of hyperbolic geometry.

If hyperbolic geometry were inconsistent, an ordinary mathematical argument could derive a contradiction. Saccheri tried to do this and failed. Could it be that he wasn't clever enough, that someday some genius will find a contradiction?

On the other hand, can it be proved that hyperbolic geometry is *consistent*—can it be proved that there is no possible way to derive a contradiction?

We might ask the same question about Euclidean geometry—how do we *know* it is consistent? Of course, this was never a burning question before the discovery of non-Euclidean geometry simply because everyone *believed* Euclidean geometry to be consistent. Remarkably enough, if we make this belief an explicit assumption (albeit a metamathematical assumption), it is possible to give a proof that hyperbolic geometry is consistent. Let us state this possibility as a theorem:

METAMATHEMATICAL THEOREM 1. If Euclidean geometry is consistent, then so is hyperbolic geometry.

Granting this result for the moment, we get the following important corollary.

COROLLARY. If Euclidean geometry is consistent, then no proof or disproof of the parallel postulate from the rest of Hilbert's postulates will ever be found, i.e., the parallel postulate is independent of the other postulates.

> To prove the corollary, assume on the contrary that a proof of the parallel postulate exists. Then hyperbolic geometry would be inconsistent, since the hyperbolic axiom contradicts a proved result. But Metamathematical Theorem 1 asserts that hyperbolic geometry is consistent relative to Euclidean geometry. This contradiction proves that no proof of the parallel postulate exists (RAA). The hypothesis that Euclidean geometry is consistent ensures that no *dis*proof exists either. ■

Thus, 2000 years of efforts to prove Euclid V were in vain. There is no more hope of proving it than there is of finding a method for trisecting every angle using straightedge and compass alone.

Of course, when we say this, we are *assuming* the consistency of the venerable Euclidean geometry. Had Saccheri, Legendre, F. Bolyai, or any of the dozens of other scholars succeeded in proving Euclid V from the other axioms, with the noble intention of making Euclidean geometry more secure and elegant, they would have instead completely destroyed Euclidean geometry as a consistent body of thought! (I urge you, dear reader, to go over the preceding statements very carefully to make sure you have understood them. If you have not understood, you have missed the main point of this book.)

In the form given here, Metamathematical Theorem 1 is due to Eugenio Beltrami (1835–1900); a different proof was later given by Felix Klein (1849–1925).[1] Beltrami proved the relative consistency of hyperbolic geometry in 1868 using differential geometry (see The Pseudosphere, Chapter 10). Klein recognized that projective geometry could be used to give another proof. In 1871 he applied the method Arthur Cayley used (in 1859) to express distance and angle measure by projective coordinates.

To prove Metamathematical Theorem 1, we have to again ask ourselves, What is a "line" in hyperbolic geometry—in fact, what is the hyperbolic plane? The honest answer is that we don't know; it is just an abstraction. A hyperbolic "line" is an undefined term describing an abstract concept that resembles the concept of a Euclidean line except for its parallelism properties. Then how shall we visualize hyperbolic geometry? In mathematics, as in any other field of research, posing the right question is just as important as finding answers.

The question of "visualizing" means finding Euclidean objects that represent hyperbolic objects. This means finding a Euclidean *model* for hyperbolic geometry. In Chapter 2 we discussed the idea of models for an axiom system; there we showed that the Euclidean parallel postulate is independent of the axioms for incidence geometry by exhibiting three-point and five-point models of incidence geometry that are not Euclidean. Here we want to know whether the parallel postulate is independent of a much *larger* system of axioms, namely,

---

[1] Beltrami made important contributions to differential geometry. Klein was a master of many branches of mathematics and an influential teacher. His book on the history of nineteenth-century mathematics shows how familiar he was with all aspects of the subject. Klein's famous inaugural address in 1872, his *Erlanger Programme*, made the study of groups of transformations and their invariants the key to geometry (see Chapter 9).

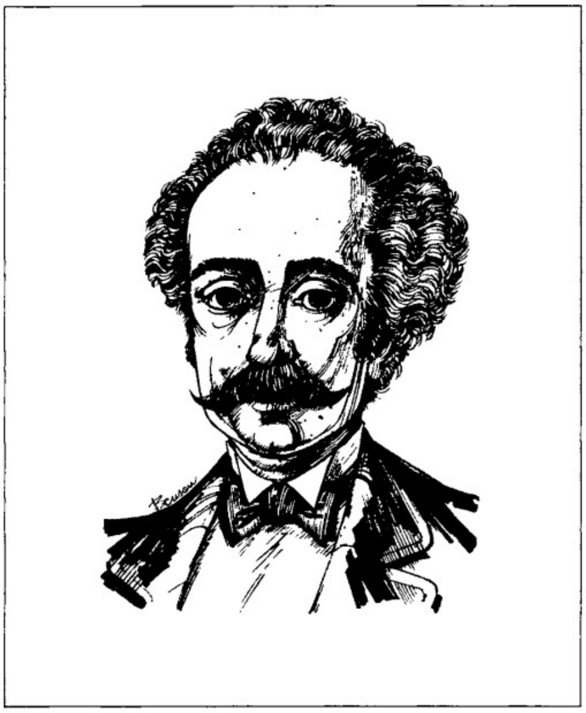

Eugenio Beltrami

neutral geometry. We can show that it is, and by the same method—by exhibiting models for hyperbolic geometry.[2]

THE BELTRAMI-KLEIN MODEL

For brevity, we will refer to this first model (the Beltrami-Klein model) as the "Klein model." We fix once and for all a circle γ in the Euclidean plane (which Cayley referred to as "the absolute"). If O is

---

[2] Unlike the situation for incidence geometry, we cannot construct a model for neutral geometry in which the elliptic parallel property holds, because it is a theorem in neutral geometry that parallel lines exist (see Corollary 2 to Theorem 4.1). If you worked through Major Exercise 13, Chapter 6, you will easily understand the motivation for the Beltrami-Klein model. It is the projective completion of the hyperbolic plane!

Felix Klein

the center of γ and OR is a radius, the *interior* of γ by definition consists of all points X such that $OX < OR$ (see Figure 7.2). In Klein's model the points in the interior of γ represent the points of the hyperbolic plane.

Recall that a chord of γ is a segment AB joining two points A and B on γ. We wish to consider the segment without its endpoints, which we will call an *open chord* and denote by A)(B. In Klein's model the open chords of γ represent the lines of the hyperbolic plane. The relation "lies on" is represented in the usual sense: P lies on A)(B means that P lies on the Euclidean line $\overleftrightarrow{AB}$ and P is between A and B. The hyperbolic relation "between" is represented by the usual Euclidean relation "between." This much is easy. The representation of "congruence" is much more complicated, and we will discuss it later in this chapter (The Projective Nature of the Beltrami-Klein Model).

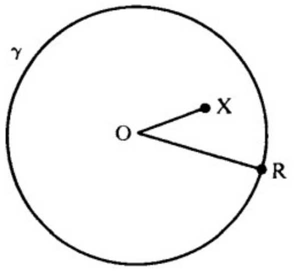

FIGURE 7.2

It is immediately clear from Figure 7.3 that the hyperbolic axiom holds in this representation.

Here the two open chords *m* and *n* through P are both parallel to the open chord *l*—for what does "parallel" mean in this representation? The definition of "parallel" states that two lines are parallel if they have no point in common. In Klein's representation this becomes: two open chords are parallel if they have no point in common (in the definition of "parallel," replace the word "line" by "open chord"). The fact that the three chords, when extended, may meet outside the circle γ is irrelevant—points outside of γ do not represent points of the hyperbolic plane. So let us summarize the Beltrami-Klein proof of the relative consistency of hyperbolic geometry as follows:

First, a glossary is set up to "translate" the five undefined terms ("point," "line," "lies on," "between," and "congruent") into their interpretations in the Euclidean model (we have done this for the first four terms). All the defined terms are then interpreted by "translating" all occurrences of undefined terms. For instance, the defined term "parallel" was interpreted by replacing every occurrence of the

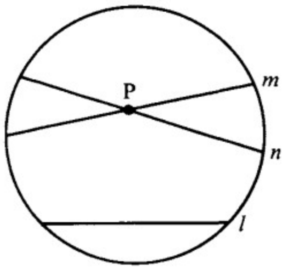

FIGURE 7.3

word "line" in the definition by "open chord." Once all the defined terms have been interpreted, we have to interpret the axioms of the system. Incidence Axiom 1, for example, has the following interpretation in the Klein model:

INCIDENCE AXIOM 1 (Klein). Given any two distinct points A and B in the interior of circle γ. There exists a unique open chord l of γ such that A and B both lie on l.

We must prove that this is a theorem in Euclidean geometry (and similarly, prove the interpretations of all the other axioms). Once all the interpreted axioms have been proved to be theorems in Euclidean geometry, any proof of a contradiction within hyperbolic geometry could be translated by our glossary into a proof of a contradiction in Euclidean geometry. From our assumption that Euclidean geometry is consistent, it follows that no such proof exists. Thus, if Euclidean geometry is consistent, so is hyperbolic geometry.

We must now backtrack and prove that the interpretations of the axioms of hyperbolic geometry in the Klein model are theorems in Euclidean geometry. Let us prove Axiom I-1 (Klein) stated above:

Proof
Given A and B interior to γ. Let $\overleftrightarrow{AB}$ be the Euclidean line through them (see Figure 7.4). This line intersects γ in two distinct points C and D. Then A and B lie on the open chord C)(D, and, by Axiom I-1 for Euclidean geometry, this is the only open chord on which they both lie. ■

In the second step of the proof we used a theorem from Euclidean geometry that states that a line passing through the interior of a circle

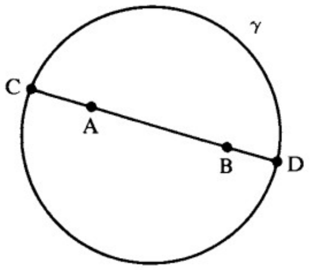

FIGURE 7.4

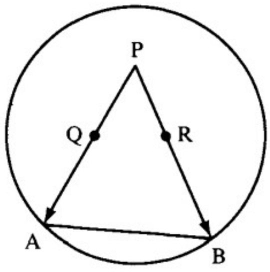

FIGURE 7.5 Limiting parallel rays.

intersects the circle in two distinct points. This can be proved from the circular continuity principle (see Major Exercise 1, Chapter 4). Verifications of the interpretations of the other incidence axioms, the betweenness axioms, and Dedekind's axiom are left as exercises; the congruence axioms are verified later in this chapter.

One nice aspect of the Klein model is that it is easy to visualize the limiting parallel rays (see Figure 7.5). Let P be a point interior to γ and not on the open chord A)(B. A and B are points on the circle and therefore do not represent points in the hyperbolic plane; they are said to represent *ideal points* and are called the *ends* of the hyperbolic line represented by A)(B (see Note 2 following Major Exercise 3, Chapter 6). Then the limiting parallel rays to A)(B from P are represented by the segments PA and PB with the endpoints A and B omitted. It is clear that any ray between these limiting parallel rays intersects the open chord A)(B, whereas all other rays emanating from P do not. The symmetry and transitivity of limiting parallelism (which, as you saw in Major Exercises 2 and 3, Chapter 6, were tricky to prove) are utterly obvious in the Klein model, as is the fact that every angle has a *line of enclosure* (given $\sphericalangle$QPR, if A is the end of $\overrightarrow{PQ}$ and B is the end of $\overrightarrow{PR}$, then A)(B is the line of enclosure of $\sphericalangle$QPR—see Major Exercise 8, Chapter 6).

Let us conclude this section by considering the interpretation in the Klein model of "congruence," the subtlest part of the model. One method of interpretation is to use a system of numerical measurement of angle degrees and segment lengths. Two angles would then be interpreted as congruent if they had the same number of degrees, and two segments would be interpreted as congruent if they had the same length (compare Theorem 4.3). The catch is that Euclidean methods of measuring degrees and lengths cannot be used. If we use Euclidean length, for example, then every line (i.e., open chord) would have a finite length less than or equal to the length of a diameter of γ. This would invalidate the interpretations of Axioms B-2 and C-1, which ensure that lines are infinitely long.

We will further discuss the matter in this chapter (in the sections Perpendicularity in the Beltrami-Klein Model and The Projective Nature of the Beltrami-Klein Model), but first let's consider the Poincaré models, in which congruence of angles is easier to describe.

THE POINCARÉ MODELS

A disk model due to Henri Poincaré (1854–1912)[3] also represents points of the hyperbolic plane by the points *interior* to a Euclidean circle γ, but lines are represented differently. First, all open chords that pass through the center O of γ (i.e., all *open diameters* l of γ) represent lines. The other lines are represented by *open arcs of circles orthogonal to* γ. More precisely, let δ be a circle orthogonal to γ (at each point of intersection of γ and δ the radii of γ and δ through that point are perpendicular). Then intersecting δ with the interior of γ gives an open arc m, which by definition represents a hyperbolic line in the Poincaré model. So we will call *Poincaré line,* or "P-line," either an open diameter l of γ or an open circular arc m orthogonal to γ (see Figure 7.6).

A point interior to γ "lies on" a Poincaré line if it lies on it in the Euclidean sense. Similarly, "between" has its usual Euclidean interpretation (for A, B, and C on an open arc coming from an orthogonal circle δ with center P, B is *between* A and C if $\overrightarrow{PB}$ is between $\overrightarrow{PA}$ and $\overrightarrow{PC}$).

The interpretation of *congruence for segments* in the Poincaré model is complicated, being based on a way of measuring length that is different from the usual Euclidean way, just as in the Klein model (see

---

[3] Poincaré was the cousin of the president of France. Like Gauss, Poincaré made profound discoveries in many branches of mathematics and physics; he even started a new branch of mathematics, algebraic topology. He used his models of hyperbolic geometry to discover new theorems about automorphic functions of a complex variable. Poincaré is also important as a philosopher of science (see Chapter 8).

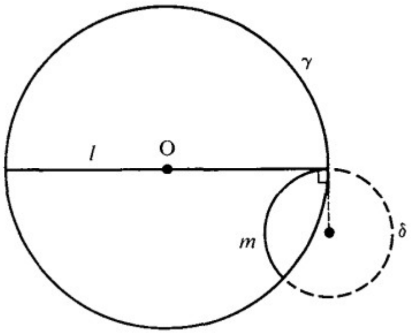

FIGURE 7.6

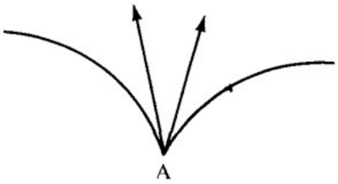

FIGURE 7.7

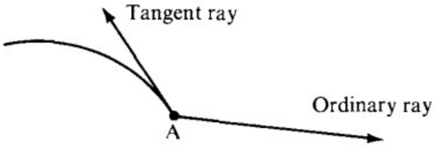

FIGURE 7.8

p. 248). *Congruence for angles* has the usual Euclidean meaning, however, and this is the main advantage of the Poincaré model over the Klein model.[4] Specifically, if two directed circular arcs intersect at a point A, the number of degrees in the *angle* they make is by definition the number of degrees in the angle between their tangent rays at A (see Figure 7.7). Or, if one directed circular arc intersects an ordinary ray at A, the number of degrees in the *angle* they make is by definition the number of degrees in the angle between the tangent ray and the ordinary ray at A (see Figure 7.8).

Having interpreted all the undefined terms of hyperbolic geometry in the Poincaré model, we get (by substitution) interpretations of all

---

[4] Technically, we say that the Poincaré model is *conformal*— it represents angles accurately —while the Klein model is not. Another example of a conformal model is Mercator's map of the surface of the earth.

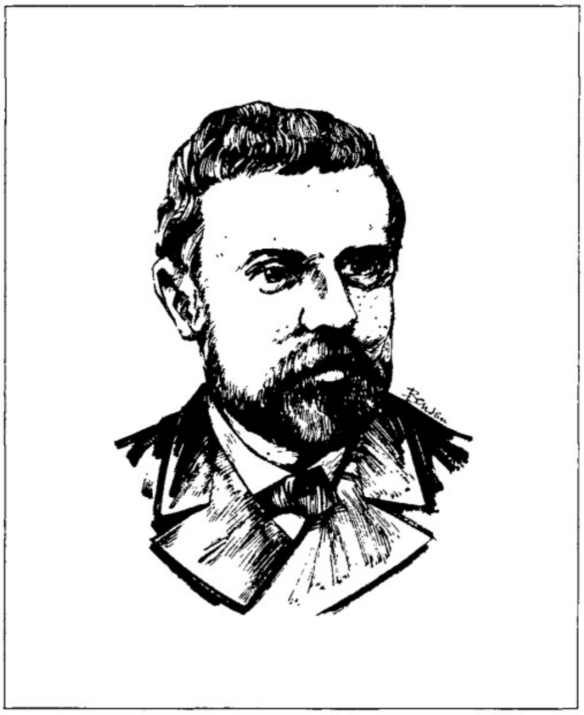

Henri Poincaré

the defined terms. For example, two Poincaré lines are *parallel* if and only if they have no point in common. Then all the axioms of hyperbolic geometry get translated into statements in Euclidean geometry, and it will be shown in the section Inversion in Circles later in this chapter that these interpretations are theorems in Euclidean geometry. Hence, the Poincaré model furnishes another proof that if Euclidean geometry is consistent, so is hyperbolic geometry.

The limiting parallel rays in the Poincaré model are illustrated in Figure 7.9. Here we have chosen l to be an open diameter A)(B; the rays are circular arcs that meet $\overleftrightarrow{AB}$ at A and B and are tangent to this line at those points. You can see how these rays approach l asymptotically as you move out toward the ideal points represented by A and B.

Figure 7.10 illustrates two parallel Poincaré lines with a common perpendicular. The diagram shows how m diverges from l on either side of the common perpendicular PO.

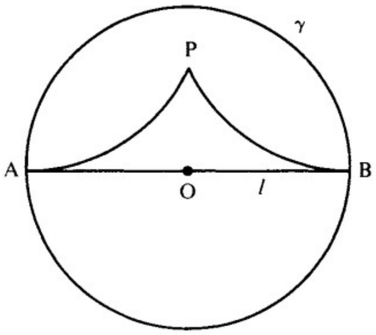

FIGURE 7.9

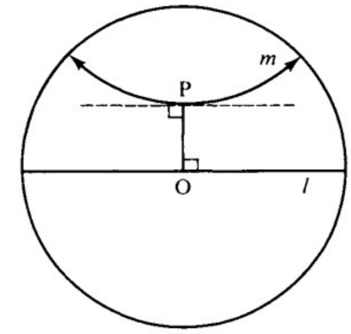

FIGURE 7.10

Figure 7.11 illustrates a Lambert quadrilateral. You can see that the fourth angle is acute. By adding the mirror image of this Lambert quadrilateral we get a diagram illustrating a Saccheri quadrilateral (Figure 7.12).

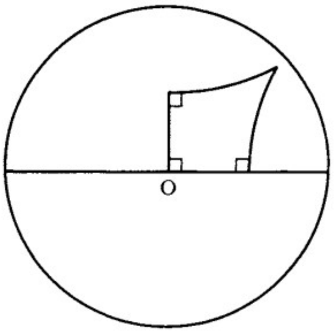

FIGURE 7.11

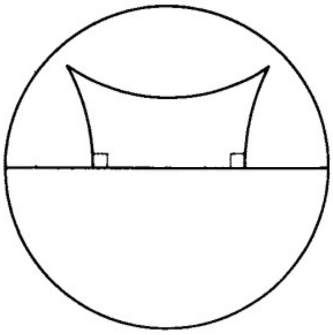

FIGURE 7.12

You may be surprised that we have two different models of hyperbolic geometry, one due to Klein and the other to Poincaré. (There is a third model, also due to Poincaré, soon to be described.) Yet you may have the feeling that these models are not "essentially different." In fact, these models are *isomorphic* in the technical sense that one-to-one correspondences can be set up between the "points" and "lines" in one model and the "points" and "lines" in the other so as to preserve the relations of incidence, betweenness, and congruence. Such isomorphism is illustrated in Figure 7.13. We start with the Klein model and consider, in Euclidean three-space, a sphere of the same radius sitting on the plane of the Klein model and tangent to it at the origin. We project upward orthogonally the entire Klein model onto the lower hemisphere of this sphere; by this projection, the chords in the Klein model become arcs of circles orthogonal to the equator. We then project stereographically from the north pole of the sphere onto the

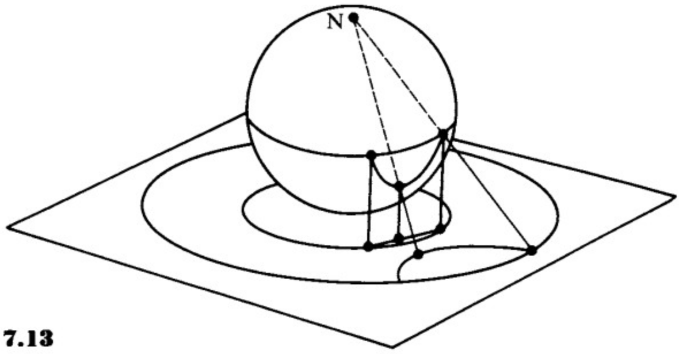

FIGURE 7.13

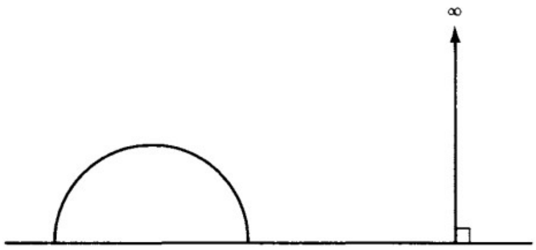

FIGURE 7.14

original plane. The equator of the sphere will project onto a circle larger than the one used in the Klein model, and the lower hemisphere will project stereographically onto the inside of this circle. Under these successive transformations, the chords of the Klein model will be mapped one-to-one onto the diameters and orthogonal arcs of the Poincaré model. In this way the isomorphism of the models may be established.

One can actually prove that *all possible models of hyperbolic geometry are isomorphic to one another,* i.e., that the axioms for hyperbolic geometry are *categorical.* The same is true for Euclidean geometry. The categorical nature of Euclidean geometry is established by introducing Cartesian coordinates into the Euclidean plane. Analogously, the categorical nature of hyperbolic geometry is established by introducing Beltrami coordinates into the hyperbolic plane (for which hyperbolic trigonometry must first be developed).[5]

In the other Poincaré model mentioned here, the points of the hyperbolic plane are represented by the points of one of the Euclidean half-planes determined by a fixed Euclidean line. If we use the Cartesian model for the Euclidean plane, it is customary to make the x axis the fixed line and then to use for our model the upper half-plane consisting of all points (x, y) with $y > 0$. Hyperbolic lines are represented in two ways:

1. As rays emanating from points on the x axis and perpendicular to the x axis;
2. As semicircles in the upper half-plane whose center lies on the x axis (see Figure 7.14).

Incidence and betweenness have the usual Euclidean interpretation.

---

[5] See Chapter 10 as well as Borsuk and Szmielew (1960), Chapter 6.

This model is conformal also (degrees of angles are measured in the Euclidean way). Measurement of lengths will be discussed later.

To establish isomorphism with the previous models, choose a point E on the equator of the sphere in Figure 7.13, and let Π be the plane tangent to the sphere at the point diametrically opposite to E. Stereographic projection from E to Π maps the equator onto a line in Π and the lower hemisphere onto the lower half-plane determined by this line. Notice that the points on this line represent ideal points. However, one ideal point is missing: the point E got lost in the stereographic projection. It is customary to imagine an ideal "point at infinity" ∞ that corresponds to E; it is the common end of all the vertical rays.

PERPENDICULARITY IN THE BELTRAMI-KLEIN MODEL

The Klein model is not conformal. Congruence of angles is interpreted differently from the usual Euclidean way, and will be explained later in this chapter (p. 260). Here we will describe only those angles that are congruent to their supplements, namely, right angles.

Let l and m be open chords of γ. To describe when $l \perp m$ in the Klein model, there are two cases to consider:

Case 1. One of l and m is a diameter. Then $l \perp m$ in the Klein sense if and only if $l \perp m$ in the Euclidean sense. (See Figure 7.15.)

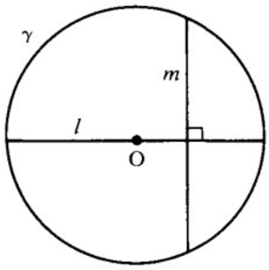

FIGURE 7.15

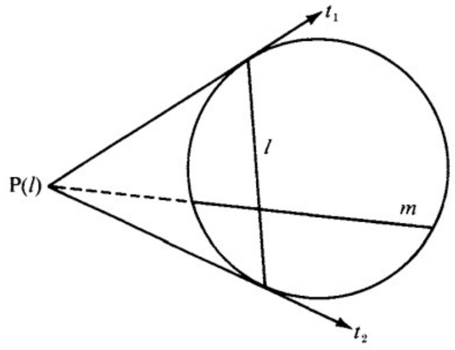

FIGURE 7.16

Case 2. Neither l nor m is a diameter. In this case we associate to l a certain point P(l) outside of γ called the *pole* of l, defined as follows. Let t_1 and t_2 be the tangents to γ at the endpoints of l. Then by definition P(l) is the unique point common to t_1 and t_2 (t_1 and t_2 are not parallel, because l is not a diameter); see Figure 7.16.

It turns out that *l is perpendicular to m in the sense of the Klein model if and only if the Euclidean line extending m passes through the pole of l.*

This description of perpendicularity will be justified later (pp. 260–261). We can use it to see more easily why divergently parallel lines have a common perpendicular — Theorem 6.7. In the hypothesis of Theorem 6.7 we are given two parallel lines that do not contain limiting parallel rays. In the Klein model this means that we are given open chords l and m that do not have a common end. The conclusion of Theorem 6.7 is that l and m have a common perpendicular k. How do we find k? Let's discuss case 2, leaving case 1 as an exercise. By the above description of perpendicularity if k were perpendicular to both l and m, the extension of k would have to pass through the pole of l and the pole of m. Hence, to construct k, we need only join these poles by a Euclidean line and take k to be the open chord of γ cut out by this line (Figure 7.17).[6]

There is a nice language that describes the behavior of pairs of lines in the Klein model. Let us call the points inside circle γ (which

---

[6] If l and m did have a common end Ω, the Euclidean line joining P(l) to P(m) would be tangent to γ at Ω. That is why Saccheri claimed that asymptotically parallel lines have "a common perpendicular at infinity," and this he found repugnant.

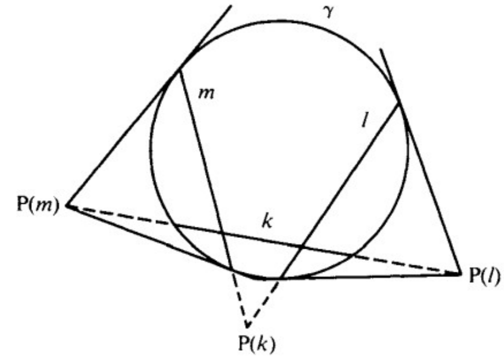

FIGURE 7.17

represent all the points in the hyperbolic plane) *ordinary points.* We already called the points on the circle γ *ideal points.* Let us call the points outside γ *ultra-ideal points.* Finally, for every diameter of γ, let us imagine another point "at infinity" such that all the Euclidean lines parallel in the Euclidean sense to this diameter meet in this point at infinity, just as railroad tracks appear to meet at the horizon. These points at infinity will also be called *ultra-ideal.* We can then say that two Klein lines "meet" at an ordinary point, an ideal point, or an ultra-ideal point, depending on whether they are intersecting, asymptotically parallel, or divergently parallel, respectively. The ultra-ideal point at which divergently parallel Klein lines l and m "meet" is the pole P(k) of their common perpendicular k (see Figure 7.17).

This language is suggestive of further theorems in hyperbolic geometry. For example, we know that two ordinary points determine a unique line, and we have seen that two ideal points also determine a unique line, the line of enclosure of Major Exercise 8, Chapter 6. We can ask the same question about two points that are ultra-ideal or about two points of different species. For example, an ordinary point and an ideal or ultra-ideal point always determine a unique line, but two ultra-ideal points may or may not (see Figure 7.18). Let us translate back from this language, say, in the case of an ordinary point O and an ultra-ideal point P(l) that is the pole of a Klein line l. What is the Klein line "joining" O to P(l)? It is the unique Klein line m through O that is perpendicular in the sense of the Klein model to the line l (see Figure 7.16). We leave the other cases for exercises.

If you did most of the exercises in hyperbolic geometry in Chapter 6, deriving results without having reliable diagrams to guide you, the

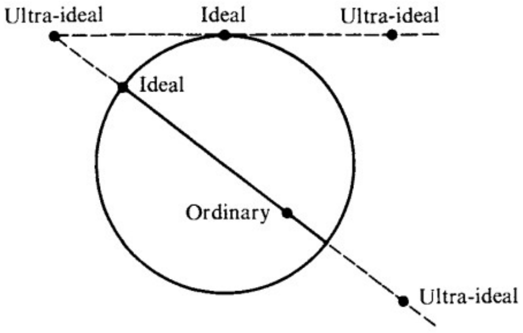

FIGURE 7.18

Klein and Poincaré models must come as a great relief. It is a useful exercise to take an absurd diagram like Figure 6.26 and draw those divergently parallel perpendicular bisectors of the triangle more accurately in one of the models. It is amazing that J. Bolyai and Lobachevsky were able to visualize hyperbolic geometry without such models, especially since they worked in three dimensions. They must have had non-Euclidean eyesight.

A MODEL OF THE HYPERBOLIC PLANE FROM PHYSICS

This model comes from the theory of special relativity. In Cartesian three-space $\mathbb{R}^3$, with coordinates denoted x, y and t (for *time*), distance will be measured by the Minkowski metric

$$ds^2 = dx^2 + dy^2 - dt^2.$$

Then with respect to the Minkowski metric, the surface of equation

$$x^2 + y^2 - t^2 = -1$$

is a "sphere" centered at the origin $O = (0, 0, 0)$ of imaginary radius $i = \sqrt{-1}$. (As was mentioned in Chapter 5, Lambert was the first to wonder if such a model existed.) In Euclidean terms, it is a two-sheeted *hyperboloid* (surface of revolution obtained by rotating the hyperbola $t^2 - x^2 = 1$ around the x axis). We choose the sheet Σ: $t \geq 1$ as our model. It looks like an infinite bowl (see Figure 7.19).

5- < 2 T3

CHAPTER 8

PHILOSOPHICAL IMPLICATIONS

> I have had my solutions for a long time, but I do not yet know how I am to arrive at them.
>
> C. F. GAUSS

WHAT IS THE GEOMETRY OF PHYSICAL SPACE?

We have shown that if Euclidean geometry is consistent, so is hyperbolic geometry, since we can construct models for it within Euclidean geometry. Conversely, it can be proved that if hyperbolic geometry is consistent, so is Euclidean geometry, for the "horocycles" on the "horosphere" in hyperbolic space form a model of the lines on the Euclidean plane (see Kulczycki, 1961, §17). *Thus, the two geometries are equally consistent.*

You may grant now that, logically speaking, hyperbolic geometry deserves to be put on an equal footing with Euclidean geometry. But you may also feel that hyperbolic geometry is just an amusing intellectual pastime, whereas Euclidean geometry accurately represents the physical world we live in and is therefore far more important. Let's examine this idea a little more closely.

Certainly, engineering and architecture are evidence that Euclidean geometry is extremely useful for ordinary measurement of distances that are not too large. However, the representational accuracy of Euclidean geometry is less certain when we deal with larger distances. For example, let us interpret a "line" physically as the path traveled by a light ray. We could then consider three widely separated light sources forming a physical triangle. We would want to measure the angles of this physical triangle in order to verify whether the sum is 180° or not (such an experiment would presumably settle the question of whether space is Euclidean or hyperbolic).

F. W. Bessel, a friend of Gauss, performed such a measurement, using the angle of parallax of a distant star. The results were inconclusive. Why? Because any physical experiment involves experimental error. Our instruments are never completely accurate. Suppose the sum did turn out to be 180°. If the error in our measurement were at most 1/100 of a degree, we could conclude only that the sum was between 179.99° and 180.01°. We could never be sure that it actually was 180°.

Suppose, on the other hand, that measurement gave us a sum of 179°. Although we could conclude only that the sum was between 178.99° and 179.01°, we would be certain that the sum was less than 180°. In other words, the only conclusive result of such an experiment would be that space is hyperbolic![1] The inconclusiveness of Bessel's experiment shows only that if space is hyperbolic, the defects of terrestrial triangles are extremely small.

To repeat the point: Because of experimental error, a physical experiment can never prove conclusively that space is Euclidean — it can prove only that space is non-Euclidean.

The discussion can be made more subtle. We must question the nature of our instruments — aren't they designed on the basis of Euclidean assumptions? We must question our interpretation of "lines" — couldn't light rays travel on curved paths? We must question whether space, especially space of cosmic dimensions, cannot be described by geometries other than these two.

The latter is in fact our present scientific attitude. According to Einstein, space and time are inseparable and the geometry of space-time is affected by matter, so that light rays are indeed curved by the gravitational attraction of masses. Space is no longer conceived of as an empty Newtonian box whose contours are unaffected by the rocks put into it. The problem is much more complicated than Euclid or Loba-

---

[1] If the measurement gave us a sum of 181° with error at most .01°, we would conclude that space is elliptic.

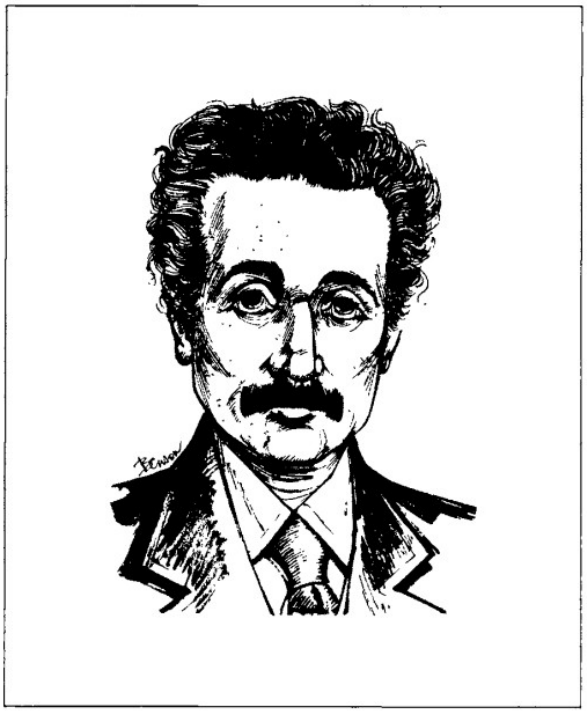
Albert Einstein

chevsky ever imagined — neither of their geometries is adequate for our present conception of space. This does not diminish the historical importance of non-Euclidean geometry. Einstein said, "To this interpretation of geometry I attach great importance, for should I not have been acquainted with it, I never would have been able to develop the theory of relativity."[2]

Here is the famous response of Poincaré to the question of which geometry is true:

> If geometry were an experimental science, it would not be an exact science. It would be subjected to continual revision. . . . *The geometrical axioms are therefore neither synthetic a priori intuitions nor experimental*

---

[2] See George Gamow (1956), which tells how Einstein developed a geometry appropriate to general relativity from the ideas of Georg Friedrich Bernhard Riemann (1826–1866).

> *facts. They are conventions.* Our choice among all possible conventions is guided by experimental facts; but it remains free, and is only limited by the necessity of avoiding every contradiction, and thus it is that postulates may remain rigorously true even when the experimental laws which have determined their adoption are only approximate. In other words, *the axioms of geometry* (I do not speak of those of arithmetic) *are only definitions in disguise.* What then are we to think of the question: Is Euclidean Geometry true? It has no meaning. We might as well ask if the metric system is true and if the old weights and measures are false; if Cartesian coordinates are true and polar coordinates false. *One geometry cannot be more true than another: it can only be more convenient.* [italics added] [3]

You may think that Euclidean geometry is the most convenient — it is for ordinary engineering, but not for the theory of relativity. Moreover, R. K. Luneburg contends that visual space, the space mapped on our brains through our eyes, is most conveniently described by hyperbolic geometry.[4]

Philosophers are still arguing about Poincaré's philosophy of conventionalism. One school, which includes Newton, Helmholtz, Russell, and Whitehead, contends that space has an intrinsic metric or standard of measurement. The other school, which includes Riemann, Poincaré, Clifford, and Einstein, contends that a metric is stipulated by convention. The discussion can become very subtle (see Torretti, 1978, Chapter 4).

WHAT IS MATHEMATICS ABOUT?

The preceding discussion sheds new light on what geometry, and in general, mathematics, is about. Geometry is not about light rays, but the path of a light ray is one possible *physical interpretation* of the undefined geometric term "line." Bertrand Russell once said that "mathematics is the subject in which we do not know what we are

---

[3] H. Poincaré (1952), p. 50. See also Essay Topic 18 at the end of this chapter.

[4] R. K. Luneburg (1947), and his article in the *Optical Society of America Journal,* October, 1950, p. 629. See also the articles by O. Blank in that same journal, December, 1958, p. 911, and March, 1961, p. 335, and the explanation in Trudeau (1987), pp. 251 – 254.

5- TN26 63

CHAPTER 10

FURTHER RESULTS IN HYPERBOLIC GEOMETRY

> The theorems of this geometry appear to be paradoxical and, to the uninitiated, absurd; but calm, steady reflection reveals that they contain nothing at all impossible.
>
> C. F. GAUSS

In this chapter we will penetrate more deeply into the "strange new universe" of hyperbolic geometry. To facilitate this adventure, we will sometimes glance quickly at certain unusual forms, not pausing for rigorous proofs of existence; at other times we will be more rigorous but allow ourselves to operate within the Klein or Poincaré models where our familiar Euclidean tools are available, instead of laboriously reasoning directly from the hyperbolic axioms. Hopefully, by the end of this excursion, the terrain will become sufficiently familiar to you that you will be comfortable enough to make further explorations with other guides.

AREA AND DEFECT

In 1799, in answer to a letter from Farkas Bolyai in which Bolyai claimed to have proved Euclid's fifth postulate, Gauss wrote:

> . . . the way in which I have proceeded does not lead to the desired goal, the goal that you declare you have reached, but instead to a doubt of the validity of [Euclidean] geometry. I have certainly achieved

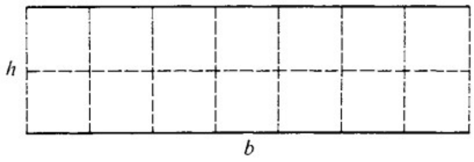

FIGURE 10.1 Area = bh.

> results which most people would look upon as proof, but which in my eyes prove almost nothing; if, for example, one can prove that there exists a right triangle whose area is greater than any given number, then I am able to establish the entire system of [Euclidean] geometry with complete rigor. Most people would certainly set forth this theorem as an axiom; I do not do so, though certainly it may be possible that, no matter how far apart one chooses the vertices of a triangle, the triangle's area still stays within a finite bound. I am in possession of several theorems of this sort, but none of them satisfy me.[1]

One of the most surprising facts in hyperbolic geometry is that there is an upper limit to the possible area a triangle can have, even though there is *not* an upper limit to the lengths of the sides of the triangle.

To see how this can be, we have to review the way in which area is calculated in Euclidean geometry. The simplest figure is a rectangle, whose area we calculate as the length of the base times the length of the side (Figure 10.1). This formula is arrived at by noticing that exactly bh unit squares fill up the interior of the rectangle, where a unit square is a square whose side has length one. Keep in mind that the unit of length is arbitrary, so that if we measure area in square inches, we get a different number than if we measure in square feet (but the latter number is always proportional to the former, the proportionality factor being $144 = 12^2$).

From the area of a rectangle we can calculate the area of a right triangle. A diagonal of a rectangle divides it into two congruent right triangles, and since we want congruent triangles to have the same area, the area of the right triangle must be half the area of the rectangle. (See Figure 10.2.)

We can decompose the interior of an arbitrary triangle into the union of the interiors of two right triangles by dropping an appropriate

---

[1] R. Bonola (1955).

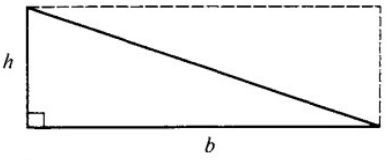

FIGURE 10.2 Area $= \frac{1}{2}bh$.

altitude. Since we want the area of the whole to be the sum of the area of its parts, we find the area of the triangle to be $\frac{1}{2}b_1h + \frac{1}{2}b_2h$, and since $b = b_1 + b_2$, we again get half the base times the height for the area of a general triangle (Figure 10.3).

You can see that by taking b and h to be sufficiently large, the area $\frac{1}{2}bh$ can be made as large as you like.

So why doesn't this work equally well in hyperbolic geometry? Because the whole system of measuring area is based on square units, and as we have seen (Lemma 6.1) rectangles (in particular, squares) do not exist in hyperbolic geometry.

What then does "area" mean in hyperbolic geometry? We can certainly say intuitively that it is a way of assigning to every triangle a certain positive number called its *area,* and we want this area function to have the following properties:

1. *Invariance under congruence.* Congruent triangles have the same area.
2. *Additivity.* If a triangle T is split into two triangles T_1 and T_2 by a segment joining a vertex to a point of the opposite side, then the area of T is the sum of the areas of T_1 and T_2 (Figure 10.4).

Having defined area, we then ask how it is calculated. It can be proved rigorously that in hyperbolic geometry the area of a triangle cannot be calculated as half the base times the height (see Moise, 1990, p. 411). So how do you calculate it? Here we find one of the

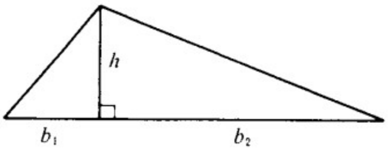

FIGURE 10.3 Area $= \frac{1}{2}bh$ where $b = b_1 + b_2$.

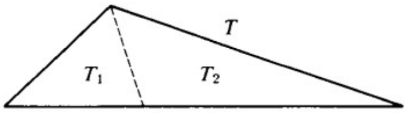

FIGURE 10.4 Area $T =$ area $T_1 +$ area T_2.

most beautiful aspects of mathematics, a direct relationship between two concepts that at first seem totally unrelated. You may have recognized this relationship in reading area properties 1 and 2, for in Theorem 4.7 and Exercise 1, Chapter 4, we proved that the *defect* also has these properties. Recall that in hyperbolic geometry the angle sum of any triangle is always less than 180° (Theorem 6.1), so that if we define the defect to be 180° minus the angle sum, we get a positive number. When a mathematician sees two functions with the same properties, he suspects they are related. Gauss discovered this relationship as early as 1794 (he was only 17 years old), and called it the first theorem in the subject.[2]

THEOREM 10.1. In hyperbolic geometry there is a positive constant k such that for any $\triangle ABC$

$$\text{area}(\triangle ABC) = \frac{\pi}{180} k^2 \times \text{defect}(\triangle ABC).$$

For the proof, which is not difficult although it is somewhat lengthy, see Moise (1990, p. 413). The theorem says that the area of any triangle is proportional to its defect, with proportionality constant $(\pi/180)k^2$. This constant depends on the unit of measurement, i.e., on whichever triangle is taken to have area equal to 1.

We can now see why there is an upper limit to the area of all triangles. Namely, the defect measures how much the angle sum is less than 180. Since the angle sum can never get below 0, the defect can never get above 180. Thus, we have the following corollary.

---

[2] This theorem seems to imply that Euclidean and hyperbolic area theories have little in common. Yet J. Bolyai discovered a wonderful theorem on area that is valid in *both* geometries. Define the area of a polygon to be the sum of the areas of the triangles used to triangulate the polygon. (It is not difficult to show that this definition does not depend on the choice of triangulation.) Bolyai's theorem states that two polygons S and S' have the same area if and only if for some n, polygon S (respectively, S') has a triangulation $\{T_1, \ldots, T_n\}$ (respectively, $\{T'_1, \ldots, T'_n\}$) such that $T_j \cong T'_j$ for all $j = 1, \ldots, n$. (For a proof, see E. E. Moise (1990, pp. 394–410). To appreciate the depth of this theorem, try to prove it for two rectangles in the Euclidean plane.)

COROLLARY. In hyperbolic geometry the area of any triangle is at most πk^2.

Of course, there is no finite triangle whose area equals the maximal value πk^2, although we can approach this area as closely as we wish (and achieve it with an infinite trebly asymptotic triangle). However, J. Bolyai showed how to construct a *circle* of area πk^2 and a regular 4-sided polygon with a 45° angle that also has this area (see Exercise 29 or Bonola, 1955, pp. 106–110).

If we interpret hyperbolic geometry in our physical world, it is clear that since defects of terrestrial triangles are immeasurably small, while their areas are measurable, the proportionality constant k^2 must be extremely large. According to Kulczycki (1961, pp. 153–155) the measurements of the parallaxes of fixed stars "elicit that the constant k is not less than about six hundred trillion miles." Thus, $k^2 > 36 \times 10^{28}$. These measurements are based on the fact that the right triangle whose one side is half the major axis of the earth's orbit around the sun and whose opposite vertex is at a fixed star has defect less than the parallax of the star (Figure 10.5). Of course, since Einstein, we do not use hyperbolic geometry to model the geometry of the universe.

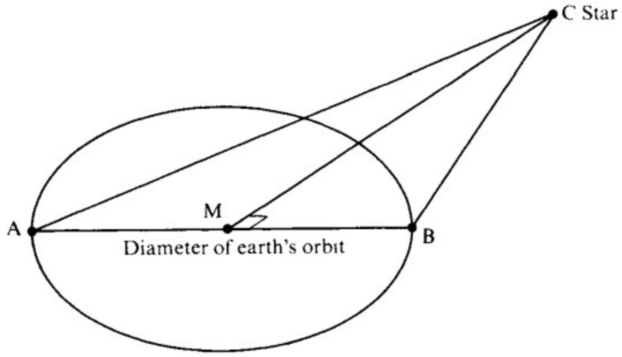

FIGURE 10.5 Defect $(\triangle ACM)^\circ = 90^\circ - (\sphericalangle CAM)^\circ - (\sphericalangle ACM)^\circ < 90^\circ - (\sphericalangle CAM)^\circ =$ parallax.

THE ANGLE OF PARALLELISM

Recall that given any line l and any point P not on l, there exist limiting parallel rays $\overrightarrow{PX}$ and $\overrightarrow{PY}$ to l that are situated symmetrically about the perpendicular PQ from P to l (Figure 10.6; see Chapter 6). We proved that $\sphericalangle XPQ \cong \sphericalangle YPQ$ (Theorem 6.6), so either of these angles can be called the *angle of parallelism* for P with respect to l.

It is not difficult to show that the number α of radians in the angle of parallelism depends only on the distance d from P to Q, not on the particular line l or the particular point P (see Major Exercise 5, Chapter 6). In the notation used after the proof of Theorem 6.6, $\alpha = (\pi/180)\Pi(PQ)^\circ$.

The following formula relating α and d was discovered by J. Bolyai and Lobachevsky.

THEOREM 10.2. Formula of Bolyai-Lobachevsky:

$$\tan\frac{\alpha}{2} = e^{-d/k}$$

where k is the constant whose square occurs in the proportionality factor of area to defect in Theorem 10.1.

This is certainly one of the most remarkable formulas in all of mathematics, and it is astonishing how few mathematicians know it. In this formula the number e is the base for the natural logarithms (e is approximately 2.718 . . .), and $\tan \alpha/2$ is the trigonometric tangent of half of α. Proofs of this formula will be found in Kulczycki (1961, §20) and in Borsuk and Szmielew (1960, Chapter 6, §26). We have checked this formula for the Poincaré disk model (where $k = 1$) —see Theorem 7.2, p. 256.

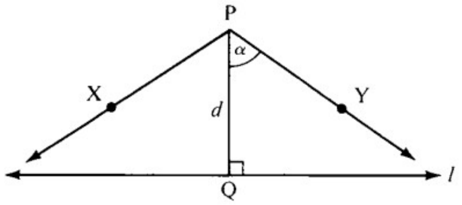

FIGURE 10.6

CYCLES

The various proofs of the Bolyai-Lobachevsky formula (Theorem 10.2) all make use of a curve that is peculiar to hyperbolic geometry, called either a *limiting curve* or a *horocycle* in the literature. It is obtained as follows (see also Exercises 41–44, Chapter 9).

Start with a line *l*, point Q on *l*, and erect perpendicular $\overleftrightarrow{PQ}$ to *l* through Q. Then consider the circle δ with center P and radius PQ, which is tangent to *l* at Q (see Figure 10.7). Now let P recede from *l* along the perpendicular. The circle δ will increase in size, remaining tangent to *l*, and will approach a limiting position as P recedes arbitrarily far from Q. In Euclidean geometry the limiting position of δ would just be the line *l*, but in hyperbolic geometry the limiting position of δ is a new curve *h* called a *limiting curve* or *horocycle.*

We can visualize this in the Poincaré model as follows. Let *l* be a diameter of the Euclidean circle γ whose interior represents the hyperbolic plane, and let Q be the center of γ. It can be proved that the hyperbolic circle with hyperbolic center P is represented by a Euclidean circle whose Euclidean center R lies between P and Q (Exercise P-5, Chapter 7); see Figure 10.8.

As P recedes from Q toward the ideal point represented by S, R is pulled up to the Euclidean midpoint of SQ, so that the horocycle *h* is a Euclidean circle tangent to γ at S and tangent to *l* at Q. It can be shown in general that all horocycles are represented in the Poincaré model by Euclidean circles inside γ and tangent to γ. Moreover, all the Poincaré lines passing through the ideal point S are orthogonal to *h*; a hyperbolic ray from a point of *h* out to the ideal point S is called a *diameter* of *h* (see Chapter 9, Exercises 41, 44, and 46).

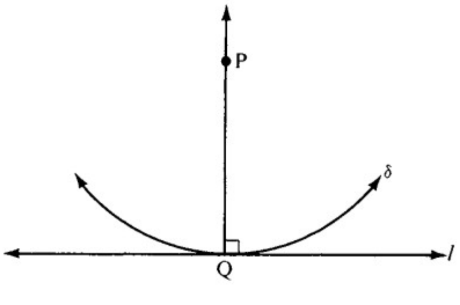

FIGURE 10.7

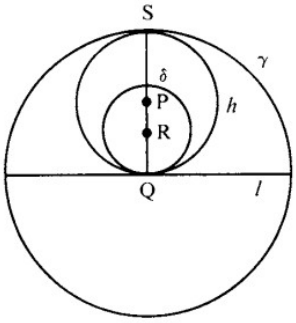

FIGURE 10.8

In the Poincaré model two horocycles tangent to γ at S are said to be *concentric* (see Figure 10.9). It is by studying the ratio of corresponding arcs on concentric horocycles that the Bolyai-Lobachevsky formula can be derived (see Kulczycki, 1961, §18).

There is an analogous construction in hyperbolic space called the *horosphere.* Instead of taking the limit of circles to get the horocycle, one takes the limit of spheres to get the horosphere or *limiting surface.*

Another strange curve in hyperbolic geometry that has no Euclidean counterpart is the *equidistant curve* (Figure 10.10). Start with a line l and a point P not on l. Consider the locus of all points on the same side of l as P and at the same perpendicular distance from l as P. In Euclidean geometry this locus would just be the unique line through P parallel to l, but in hyperbolic geometry it is not a line, it is the *hypercycle,* or *equidistant curve,* through P.

In the Poincaré model let A and B be the ideal endpoints of l. It turns out that the equidistant curve to l through P is represented by

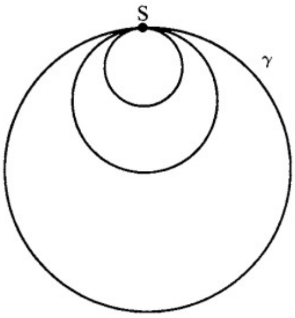

FIGURE 10.9 Concentric horocycles.

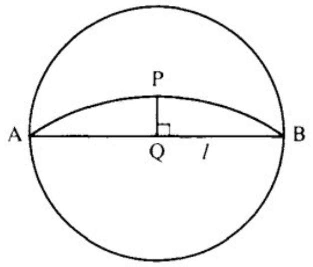

FIGURE 10.10 Equidistant curve.

the arc of the Euclidean circle passing through A, B, and P (Chapter 9, Exercises 40, 45, and 47). This curve is orthogonal to all Poincaré lines perpendicular to the line *l*.

In the Euclidean plane three points either lie on a uniquely determined line or on a uniquely determined circle. Not so in the hyperbolic plane—look at the Poincaré model. A Euclidean circle represents:

1. A *hyperbolic circle* if it is entirely inside γ;
2. A *horocycle* if it is inside γ except for one point where it is tangent to γ;
3. An *equidistant curve* if it cuts γ nonorthogonally in two points;
4. A *hyperbolic line* if it cuts γ orthogonally.

It follows that in the hyperbolic plane three noncollinear points lie either on a circle, a horocycle, or a hypercycle accordingly as the perpendicular bisectors of the triangle are "concurrent" in an ordinary, ideal, or ultra-ideal point—see the last section of this chapter.

THE PSEUDOSPHERE

One of the difficulties with the Poincaré model is that, although it faithfully represents angles of the hyperbolic plane (i.e., it is a *conformal model*), it distorts distances. So it is natural to ask whether another model exists that also represents hyperbolic lengths faithfully by Euclidean lengths. If there is such a model, it would be called *isometric*. An equally natural idea is to seek as a model some surface in Euclidean three-dimensional space. The lines of the hyperbolic plane would

5- 6 <26673

(8) The representation by Weierstrass coordinates helps make sense of Lambert's description of the hyperbolic plane as a "sphere of imaginary radius *i*."

(9) The curvature of the hyperbolic plane is $1/k^2$, where k^2 times the defect in radians equals the area of a triangle.

(10) The analogue of the Pythagorean theorem is the formula $\cosh c = \cosh a \cosh b$ (for a right triangle with right angle at C, standard notation, $k = 1$).

EXERCISES

1. Verify all the identities for hyperbolic functions listed in the table on p. 401.
2. Verify formulas (5), (6), (7) in which Π provides a link between hyperbolic and circular functions. Graph the function $\Pi(x)$.
3. The proof of Theorem 10.3 required a complicated argument using the Poincaré model. Give a different proof using the Klein model. (Hint: According to the note at the end of Chapter 7, you can assume A = O, the center of the absolute. Show that $\cos A = \overline{AC}/\overline{AB}$ and $\sin A = \overline{BC}/\overline{AB}$ (Euclidean lengths), $\overline{AB} = \tanh c$ (where $c = d'(AB)$ is the Klein length), $\overline{AC} = \tanh b$, and $\overline{BC} = (1 - \overline{AC}^2)^{1/2} \tanh a = \tanh a/\cosh b$ (use Theorem 7.4). Conclude by deducing the formula $\cosh c = \cosh a \cosh b$ from the Pythagorean theorem. See Figure 10.29.)
4. Prove Theorem 10.4.

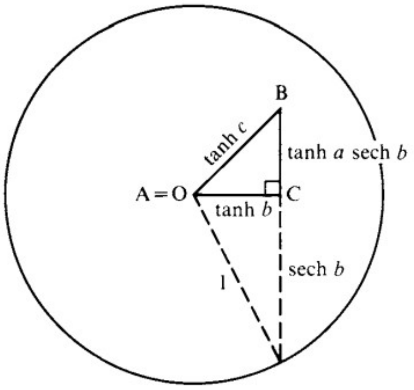

FIGURE 10.29

5. Verify step 2 in the proof of Theorem 10.6.
6. Verify formulas (18) through (21) for complementary lengths. Graph the function
$$f(x) = x^* = \ln \frac{e^x + 1}{e^x - 1}.$$
7. Prove the assertions about Weierstrass coordinates in Theorem 10.10. (Hint: Derive the equation of a line in Weierstrass coordinates from the equation of a line in Beltrami coordinates.)
8. Verify formulas (28) and (29) in the proof of Theorem 10.10 and use them to reconcile with the distance formula in Exercise K-14, Chapter 7.
9. Generalize Lemmas 10.1 and 10.2 by showing that if (x_1, y_1) and (x_2, y_2) are distinct points in the Klein model, then the midpoint and perpendicular bisector of the segment they determine are given respectively by
$$\left(\frac{x_1s_2 + x_2s_1}{s_1 + s_2}, \frac{y_1s_2 + y_2s_1}{s_1 + s_2}\right)$$
$$(x_1s_2 - x_2s_1)x + (y_1s_2 - y_2s_1)y + (s_1 - s_2) = 0$$
where $s_i = \sqrt{1 - x_i^2 - y_i^2}$, $i = 1, 2$. (Hint: Use Lemma 10.2 to find the point Q in the Cartesian plane where the perpendicular bisectors of OP_1 and OP_2 meet. Then joining Q to the pole of $\overleftrightarrow{P_1P_2}$ gives the perpendicular bisector of P_1P_2, and intersecting it with P_1P_2 gives the midpoint.)
10. Prove the corollary to Theorem 10.11.
11. In a right triangle with right angle at C, prove that the circumscribed cycle is a
$$\left.\begin{matrix}\text{Circle}\\ \text{Horocycle}\\ \text{Equidistant curve}\end{matrix}\right\} \Leftrightarrow \frac{a}{2} \left\{\begin{matrix} < \\ = \\ > \end{matrix}\right\} \left(\frac{b}{2}\right)^*.$$
(Hint: Apply the second criterion of Theorem 10.11 with $\sphericalangle$C the largest angle, using the fact that Π is a decreasing function. Or else argue directly from the definition of complementary lengths.)
12. Verify A. P. Kotelnikov's rule for remembering the relations among the parts of a right triangle with the right angle at C, standard notation (x^* denoting the complementary length to x): in Figure 10.30, *the sine of each angle is equal to the product of the tangents of the two adjacent angles and is equal to the product of the cosines of the two opposite angles.* For example,
$$\sin A = \tan \Pi(a^*) \tan \Pi(c) = \cos \Pi(b^*) \cos B.$$
(This rule is the hyperbolic analogue of the rule John Napier published in 1614 for the trigonometry of a right triangle on a unit sphere in

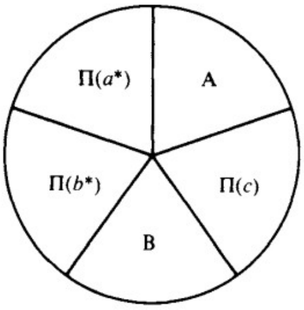

FIGURE 10.30

Euclidean space. For Napier's rule, use a, b, A′, c', B′ in cyclic order, where A′ denotes the complementary angle to ∢A and $c' = \pi/2 - c$. J. Bolyai and Lobachevsky discovered that spherical trigonometry in hyperbolic space is the same as in Euclidean space.)

13. In Euclidean geometry, every circle can be inscribed in a triangle (the tangents at three appropriate points on the circle meet to form a triangle). Show that in hyperbolic geometry (with $k = 1$), an inscribed circle of a triangle must have diameter less than ln 3. (Hint: In Figure 10.32, show that $(\sphericalangle \text{AIB}')^r + (\sphericalangle \text{BIC}')^r + (\sphericalangle \text{CIA}')^r = \pi$, and that each of these three angles is less than $\Pi(r)$; then apply the Bolyai-Lobachevsky formula to find x such that $\Pi(x) = \pi/3$.)
14. Show that (with $k = 1$) a trebly asymptotic triangle has an inscribed circle of diameter ln 3 (see Exercise K-13, Chapter 7). Show that the area of this circle is
$$2\pi\left(\frac{2}{\sqrt{3}} - 1\right)$$
and its circumference is $2\pi/\sqrt{3}$ (use Theorems 10.5 and 10.7).
15. Show that for any three positive numbers α, β, γ such that $\pi < \alpha + \beta + \gamma$, there exists a triangle having these numbers as the radian measures of its angles. (Hint: Use Theorem 10.4.)
16. In a singly asymptotic triangle ABΩ, if $c = \overline{\text{AB}}$, then
$$\cosh c = \frac{\cos \text{A} \cos \text{B} + 1}{\sin \text{A} \sin \text{B}}.$$
(Hint: Let C approach Ω in formula (15), Theorem 10.4. For a proof without using continuity, note that when ∢A and ∢B are acute,
$$c = \Pi^{-1}(\alpha) + \Pi^{-1}(\beta),$$
where $\alpha = (\sphericalangle\text{A})^r$, $\beta = (\sphericalangle\text{B})^r$—see Figure 10.31.) Show that the gen-

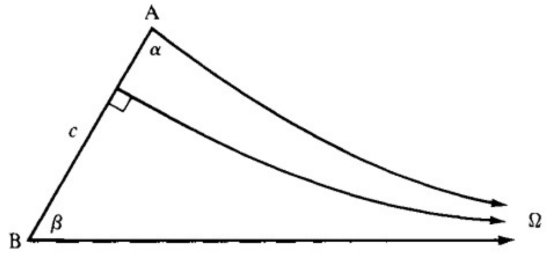

FIGURE 10.31

eralization of the Bolyai-Lobachevsky formula to the case when $\beta < \pi/2$ is

$$\tan\frac{\alpha}{2} = e^{-c}\cot\frac{\beta}{2}.$$

17. Write down equations which show how the side and the angle of an equilateral triangle determine each other.
18. (a) In a right triangle with standard notation and right angle at C, show that $\tan A = \tanh a/\sinh b$.
    (b) Deduce that in an isosceles triangle with base b and side a, summit at B, and one base angle at A,

$$\tanh a \cos\frac{B}{2} = \tan A \sinh\frac{b}{2}$$

$$\sin A \cosh\frac{b}{2} = \cos\frac{B}{2}.$$

    (Hint: Drop the altitude to the base.)
19. In a right triangle ΔABC with right angle at C (and standard notation), show that

$$\sin K = \frac{\sinh a \sinh b}{1 + \cosh a \cosh b},$$

    where K = the area = the defect of ΔABC. (Hint: Use Theorem 10.3 and trigonometric identities.)
20. Given ΔABC, if h is the length of the altitude from vertex B, show that (in standard notation) the product $\sinh b \sinh h$ is independent of the choice of which vertex is labeled B; this is the hyperbolic analogue of the Euclidean theorem that bh is constant. (Hint: Show that $\sinh b \sinh h = S \sinh a \sinh b \sinh c$, where S is the constant ratio occurring in the law of sines.) The next exercises will shed light on the

geometric significance of the constant $\frac{1}{2}\sinh b \sinh h$, which we will denote by H (for Heron); by the hint, $2H = \sin C \sinh a \sinh b = \sin B \sinh c \sinh a = \sin A \sinh b \sinh c$.

In Exercises 21–28, s will denote the *semiperimeter* $\frac{1}{2}(a+b+c)$ of $\triangle ABC$.

21. Show that (in standard notation)

$$\sin\frac{A}{2} = \sqrt{\frac{\sinh(s-b)\sinh(s-c)}{\sinh b \sinh c}}$$
$$\cos\frac{A}{2} = \sqrt{\frac{\sinh s \sinh(s-a)}{\sinh b \sinh c}}.$$

(Hint: Square both sides; use identities and Theorem 10.4.)

22. Show that $H = \sqrt{\sinh s \sinh(s-a)\sinh(s-b)\sinh(s-c)}$. (Hint: Use the identity $\sin A = 2\sin(A/2)\cos(A/2)$.)
23. "Infinitesimally," the Heron is equal to $\sqrt{s(s-a)(s-b)(s-c)}$. Show that in Euclidean geometry, this quantity is equal to the area of $\triangle ABC$. (Hint: See Coxeter, 1969, p. 12.)
24. Suppose the inscribed circle of $\triangle ABC$ has radius r and touches BC at A′, CA at B′, and AB at C′. Show that in neutral geometry, $\overline{AB'} = s - a = \overline{AC'}$, $\overline{BC'} = s - b = \overline{BA'}$, $\overline{CA'} = s - c = \overline{CB'}$ (see Figure 10.32). (Hint: Review the construction of the inscribed circle in Exercise 16, Chapter 6.)
25. Deduce from Exercise 24 that in hyperbolic geometry

$$\tanh r \sinh s = H$$

whereas in Euclidean geometry $rs =$ the area of $\triangle ABC$. (Hint: In hyperbolic geometry, use Exercises 18, 21, 22, and 24 to compute $\tan(A/2)\sinh(s-a)$; in Euclidean geometry, add up the areas of triangles IAB, IBC, and ICA in Figure 10.32.)

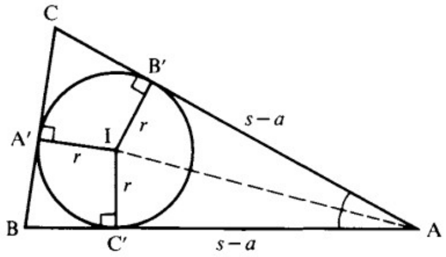

FIGURE 10.32

26. Prove Gauss' equations:

$$\begin{aligned}\cosh \tfrac{1}{2} c \sin \tfrac{1}{2} (\mathrm{A}+\mathrm{B}) &= \cosh \tfrac{1}{2} (a-b) \cos \tfrac{1}{2} \mathrm{C}\\ \cosh \tfrac{1}{2} c \cos \tfrac{1}{2} (\mathrm{A}+\mathrm{B}) &= \cosh \tfrac{1}{2} (a+b) \sin \tfrac{1}{2} \mathrm{C}\\ \sinh \tfrac{1}{2} c \sin \tfrac{1}{2} (\mathrm{A}-\mathrm{B}) &= \sinh \tfrac{1}{2} (a-b) \cos \tfrac{1}{2} \mathrm{C}\\ \sinh \tfrac{1}{2} c \cos \tfrac{1}{2} (\mathrm{A}-\mathrm{B}) &= \sinh \tfrac{1}{2} (a+b) \sin \tfrac{1}{2} \mathrm{C}.\end{aligned}$$

(Hint: Use identities such as

$$\sinh x + \sinh y = 2 \sinh \tfrac{1}{2} (x+y) \cosh \tfrac{1}{2} (x-y)$$

and analogous identities for the circular functions, then apply the half-angle formulas of Exercise 21.)

27. Show that a hyperbolic analogue of Heron's Euclidean area formula in Exercise 23 is the formula

$$\sin \frac{K}{2} = \frac{H}{2 \cosh \frac{a}{2} \cosh \frac{b}{2} \cosh \frac{c}{2}},$$

where K = area = defect of $\triangle$ABC. (Hint: Use Gauss' equations, the identity $\sin K/2 = \cos \tfrac{1}{2}(\mathrm{A}+\mathrm{B}+\mathrm{C})$, other trigonometric identities, and the formula $H = \tfrac{1}{2} \sin \mathrm{C} \sinh a \sinh b$.)

28. If $\triangle$ABC has a circumscribed circle of radius R, show that

$$\tanh R = \frac{2 \sinh \frac{a}{2} \sinh \frac{b}{2} \sinh \frac{c}{2}}{H},$$

which by Exercise 27 is equivalent to formula (30) of Theorem 10.12. (Hint: In Figure 10.33 $\sin \mathrm{A} = \sin (\beta' + \gamma')$, $H = \tfrac{1}{2} \sin \mathrm{A} \sinh b \sinh c$;

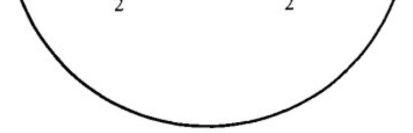

FIGURE 10.33

use Theorem 10.3 to determine $\cos \gamma'$ and $\cos \beta'$, and use Exercise 18 to determine $\sin \gamma'$ and $\sin \beta'$, obtaining with the help of identities the formula

$$H \tanh R = \left[\cos \frac{\gamma}{2} \sinh \frac{b}{2} + \cos \frac{\beta}{2} \sinh \frac{c}{2}\right] 2 \sinh \frac{b}{2} \sinh \frac{c}{2}.$$

Show finally that the term in brackets equals $\sinh a/2$ by using Theorem 10.3 to derive expressions for $\sinh a/2$, $\sinh b/2$, and $\sinh c/2$ and plugging in $\sin (\gamma + \beta)/2 = \sin \alpha/2$.)

29. Let $\Omega\Sigma\Lambda$ be a trebly asymptotic triangle and $\triangle$ABC its *pedal triangle* formed by the feet of the perpendiculars from each ideal vertex to the opposite side. Since all trebly asymptotic triangles are congruent to one another by Proposition 9.16(c), $\triangle$ABC is equilateral. Show that in the Poincaré upper half-plane model, the radian measure θ of an angle of $\triangle$ABC is given by $\tan \frac{1}{2}\theta = \frac{1}{2}$ or $\tan \theta = \frac{4}{3}$, $\sin \theta = \frac{4}{5}$, $\cos \theta = \frac{3}{5}$, and the length c of a side is given by $\cosh c = \frac{3}{2}$. Deduce that *a circle whose radius is a side of* $\triangle$ABC *has area equal to* π. Show further that the Heron H, the circumradius R, and the inradius r of $\triangle$ABC are given by

$$\begin{aligned} H &= \tfrac{1}{2} \\ \tanh R &= \tfrac{1}{2} \\ \tanh r &= \tfrac{1}{4}. \end{aligned}$$

(Hint: There are many ways to obtain these results using the previous exercises and the models. In the Poincaré upper half-plane model, taking $\Omega = -1$, $\Sigma = \infty$, $\Lambda = 1$ gives $A = i$, $B = 1 + 2i$, $C = -1 + 2i$. Show that $\overleftrightarrow{BC}$ is the upper semicircle of $x^2 + y^2 = 5$, $\overleftrightarrow{AB}$ the upper semicircle of $(x - 2)^2 + y^2 = 5$, and that the tangents to these circles at B have slopes $-\frac{1}{2}$, $\frac{1}{2}$, respectively.

This and the double angle formulas yield the assertions about θ. Exercises 18(a), 20, 25, and 28 can then be applied. Or use the Klein model, choosing $\Omega\Sigma\Lambda$ so that the origin is the incenter and circumcenter of $\triangle$ABC.)

Exercises 30–33 develop some more trigonometric formulas that William Thurston uses in his application of hyperbolic geometry to topology; see Casson and Bleiler (1988).

30. Given a quadrilateral with two adjacent right angles, labeled as in Figure 10.34. Prove that

$$\cosh d = \cosh a \cosh b \cosh c - \sinh a \sinh c.$$

(Hint: Imitate the proof of Theorem 10.8.)

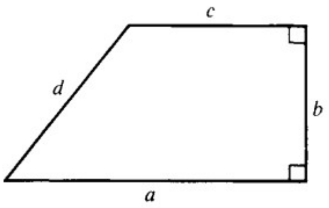

FIGURE 10.34

31. Given a pentagon with at least four right angles, labeled as in Figure 10.35. Prove that

$$\sinh e \cosh a = \sinh c \cosh b.$$

(Hint: Drop altitude h from the fifth vertex to the opposite side and apply the corollary to Theorem 10.8 to the two Lambert quadrilaterals obtained.)

32. Given a pentagon with five right angles (in Exercise K-5, Chapter 7, you showed that such pentagons exist), labeled as in Figure 10.36. Prove that

$$\cosh d = \sinh a \sinh b = \coth c \coth e.$$

(Hint: Here is one method; perhaps you can find one simpler: first apply Exercise 31 five times to all the combinations of four consecutive sides; then by considering the broken lines in Figure 10.36 and Exercise 30 and Theorem 10.3, deduce the equations

$$\cosh a \cosh b \left(\frac{\cosh c}{\cosh e}\right) - \sinh a \left(\frac{\sinh c}{\cosh e}\right)$$
$$= \cosh d = \cosh a \cosh b \left(\frac{\cosh e}{\cosh c}\right) - \sinh b \left(\frac{\sinh e}{\cosh c}\right).$$

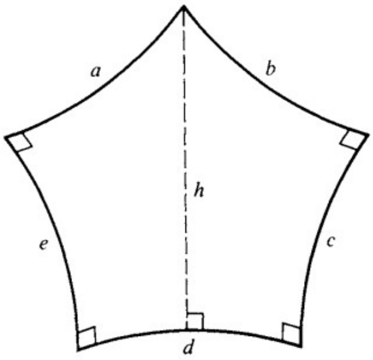

FIGURE 10.35

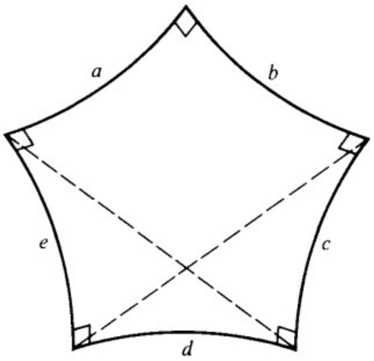

FIGURE 10.36

Substitute for the terms in parentheses from four of the earlier equations and do some algebra using the identity $\cosh^2 x = 1 + \sinh^2 x$ to get the result.)

33. In an all right-angled hexagon (again review Exercise K-5, Chapter 7) labeled as in Figure 10.37, prove the following remarkable analogues of the laws of sines and cosines given in Theorem 10.4 for a triangle:

$$\frac{\sinh a}{\sinh \alpha} = \frac{\sinh b}{\sinh \beta} = \frac{\sinh c}{\sinh \gamma}$$

$$\cosh a = \frac{\cosh \beta \cosh \gamma + \cosh \alpha}{\sinh \beta \sinh \gamma}.$$

(Hint: First show that the common perpendicular segment of length h falls inside the hexagon as shown; then apply Exercise 32 to the two all right-angled pentagons to get two expressions for $\cosh h$ and obtain the "law of sines"; for the "law of cosines," apply Exercise 32 to get expressions for $\cosh a$, $\cosh \beta$, $\cosh \alpha$, and $\cosh h$, use Exercise 31 to eliminate $\sinh c_1$, and use the identity for $\sinh \gamma_2 = \sinh(\gamma - \gamma_1)$.)

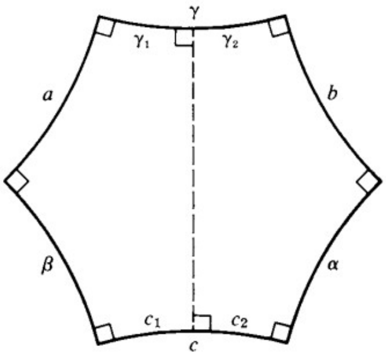

FIGURE 10.37

)-

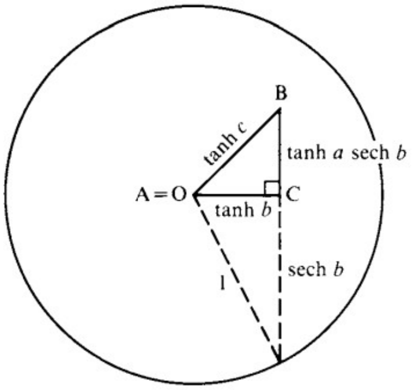
B
tanh c
tanh a sech b
A = O
C
tanh b
sech b
1

. - <

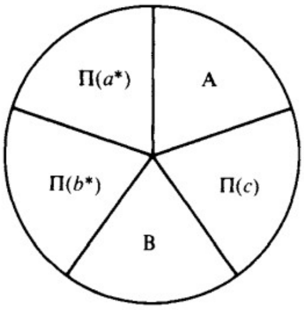
Π(a*)
A
Π(b*)
Π(c)
B

. -

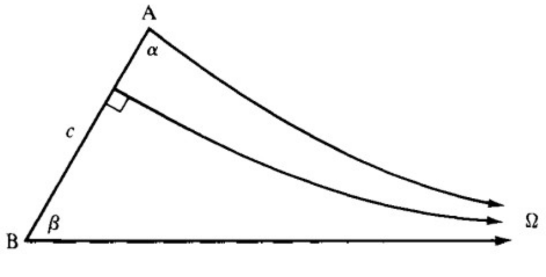
A
α
c
β
B
Ω

. -

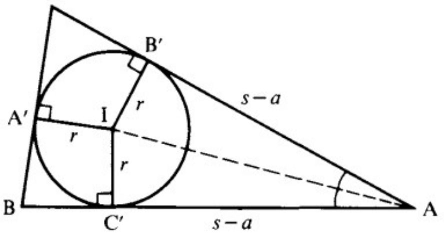
C
B′
s − a
A′
I
r
r
r
B
C′
s − a
A

. -

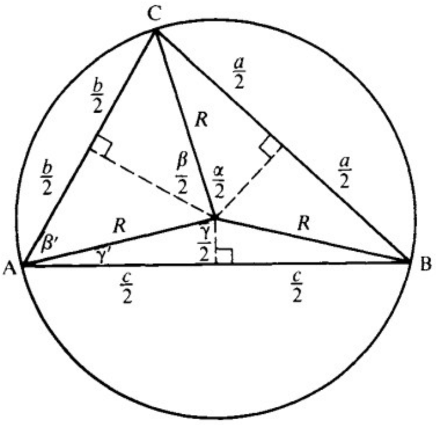
C
$\frac{b}{2}$
$\frac{a}{2}$
R
$\frac{b}{2}$
$\frac{\beta}{2}$
$\frac{\alpha}{2}$
$\frac{a}{2}$
R
R
β'
γ'
$\frac{\gamma}{2}$
A
B
$\frac{c}{2}$
$\frac{c}{2}$

. -

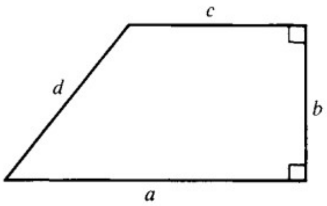
c
d
b
a

. - 6

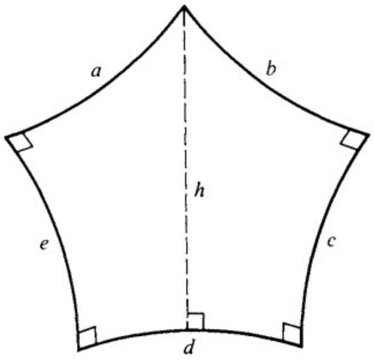
a
b
h
e
c
d

. - /

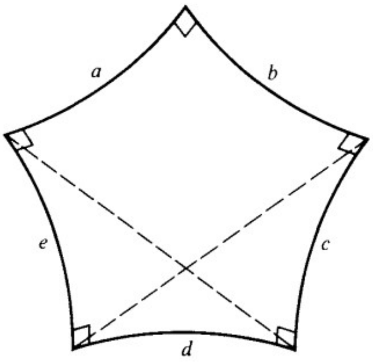
a
b
e
c
d

. - N

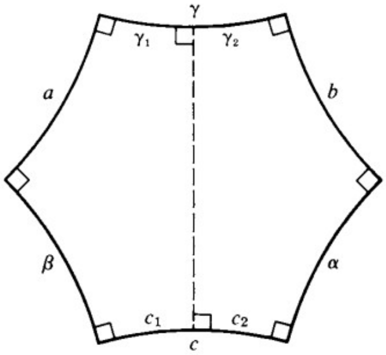
γ
γ₁
γ₂
a
b
β
α
c₁
c₂
c

. -   7

5- 66 26N 3 ( )) #   +

Clifford-parallel if they are equidistant from each other; the lines are joined to each other by a continuous family of common perpendicular segments of the same length. Such lines cannot lie in a plane (in an elliptic plane two lines must intersect), so they are skew lines. Moreover, in general, in elliptic space for any point P not on a line *l* there exist exactly two lines through P that are Clifford-parallel to *l*, called the *right and left Clifford parallels* to *l* through P. We say "in general" because there is a special line l^*, called the *absolute polar* of *l*: if P lies on l^*, there is only one Clifford parallel to *l* through P, which is l^*. (Naturally, this is difficult to visualize! See Coxeter, 1968, Chapter 7.)

Elliptic space is finite but unbounded—finite because all lines have finite length and look like circles, and unbounded because there is no boundary, just as on the surface of a sphere there is no boundary. In a universe having this geometry, with light rays traveling along elliptic lines, you could conceivably look through a very powerful telescope and see the back of your own head! (Although you might have to wait a few billion years for the light to travel all the way around.)

RIEMANNIAN GEOMETRY

It is impossible to rigorously explain the ideas of Riemannian geometry without using the language of the differential and integral calculus, so in this appendix we can only attempt to understand very roughly the intuitive idea. The basic notion we shall attempt to grasp is *curvature.*

The two simplest smooth one-dimensional figures in the Euclidean plane are a line and a circle. We think of a line as being "straight" i.e., not curved, so that if we were to assign any numerical curvature k to a line, we would assign the value $k = 0$. A circle γ, on the other hand, is certainly "curved," and how much it is curved depends on its radius r. The larger the radius of γ, the more γ approaches a line, i.e., the less it curves; it is therefore natural to define the curvature k of a circle of radius r by $k = 1/r$ (the curvature is inversely proportional to the radius); see Figure A.7.

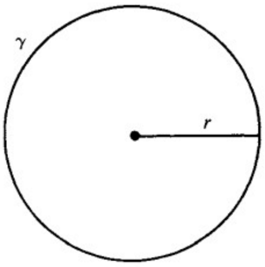

FIGURE A.7 Curvature $k = 1/r$.

Consider next an arbitrary smooth curve γ in the Euclidean plane. By "smooth curve" we mean one that has a continuously turning tangent line at each point. The tangent to a point on a circle is the perpendicular to the radius at that point. We can define the tangent to a point P on general curve γ as follows. P and another point Q on γ determine a line l. We fix point P and let Q approach P. The limiting position achieved by line l as Q approaches P is by definition the *tangent line t* to γ at P (see Figure A.8).

Besides the fixed point P, we can also consider two other points P_1 and P_2 on γ. These three points determine a circle δ. Fix P and let P_1 and P_2 both approach P along γ. The limiting position of circle δ as P_1 and P_2 approach P is the circle that "best fits" the curve γ at P, and is called the *osculating circle* to γ at P (from the Latin *osculari,* "to kiss"); see Figure A.9. It is reasonable to define the *curvature* of γ at P as the curvature of its osculating circle at P, i.e., the reciprocal $k = 1/r$ of the radius r of the osculating circle (r is also called the *radius of curvature* of γ at P).

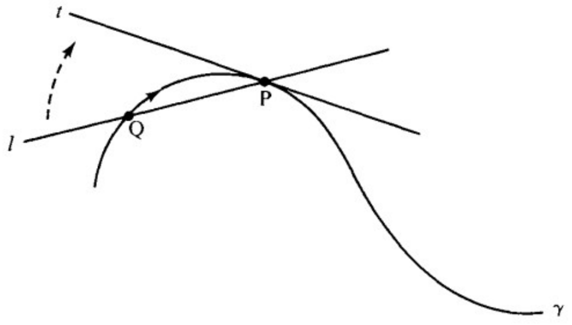

FIGURE A.8

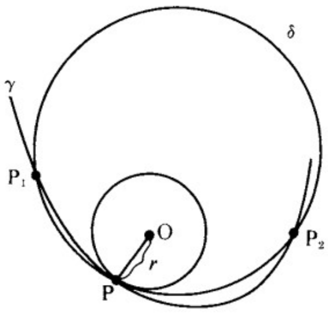

FIGURE A.9

It is clear from Figure A.10 that the osculating circle will vary in size as we move along the curve γ, so that the curvature k varies from point to point along γ. Notice also that the tangent to γ at a point P is also the tangent to the osculating circle at P.

Notice also in Figure A.10 that the osculating circle may be on different sides of the curve γ. It is convenient to *redefine curvature* so that it is positive on one side of γ and negative on the other. Once this is done, it becomes clear that in Figure A.10 there must be a point I between A and B on γ at which the curvature is zero, since we assume γ is smooth enough for the curvature to vary continuously. This point I is called a *point of inflection,* and at such a point the osculating "circle" degenerates into a line, the tangent line at I (see Figure A.11).

What we have said about plane curves applies equally well to curves in Euclidean space, with the following modifications. The osculating circle lies in a unique plane through P called the *osculating plane* Π of γ at P (except in the degenerate case of a point P at which γ has curvature zero). Since Π varies with P if γ is not a plane curve, we no longer have a smooth way to assign positive and negative values to the

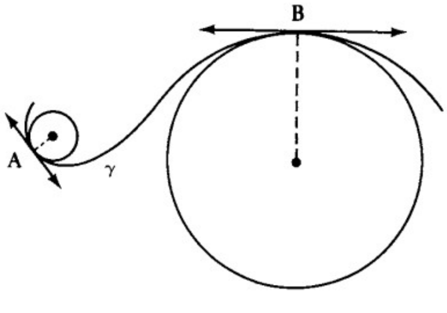

FIGURE A.10

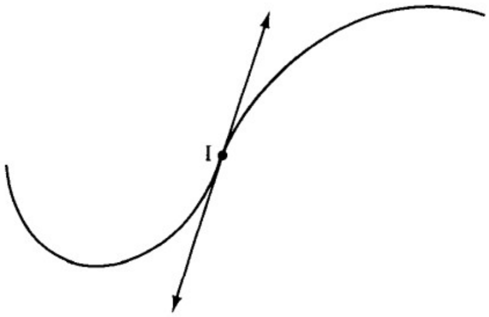

FIGURE A.11 Point of inflection $k = 0$.

curvature. We assign a *curvature vector* **k** lying in the osculating plane, emanating from P and perpendicular to the tangent line to γ at P (which also lies in the osculating plane), of length $1/r$ and pointing toward the center of the osculating circle (called the *center of curvature*). If γ has curvature zero at P, we take **k** to be the zero vector.

We next pass to smooth surfaces embedded in Euclidean three-dimensional space. "Smooth" now means that there is a continuously turning *tangent plane* for each point P of the surface. The tangent plane T at P separates Euclidean three-space into two half-spaces, one of which we arbitrarily consider positive and the other negative. Then we can define a *signed curvature* k to any curve γ through P lying on the given surface Σ which is positive or negative according as the center of curvature of γ at P lies in the positive or negative half-space. (For these signs to vary smoothly with P, we have to *orient* the surface, which can always be done locally—i.e., in a neighborhood of P.)

Consider the line through P that is perpendicular to T, called the *normal line* at P. A plane that contains the normal line intersects the surface in a plane curve. We can imagine this plane rotating around the normal line, and as it does so, it will cut out different curves on the surface passing through P. We have already explained how to define the curvature at P for each of these normal sections. In general, these curvatures will vary as we rotate around the normal line. (In the special case of a sphere these curvatures are constant and equal to the reciprocal of the radius of the sphere, since the curves cut out are all great circles on the sphere.) It can be proved by methods of differential geometry that these curvatures achieve a maximum value k_1 and a minimum value k_2 as we rotate, and the corresponding normal sections (called *principal curves*) are perpendicular to each other. The product

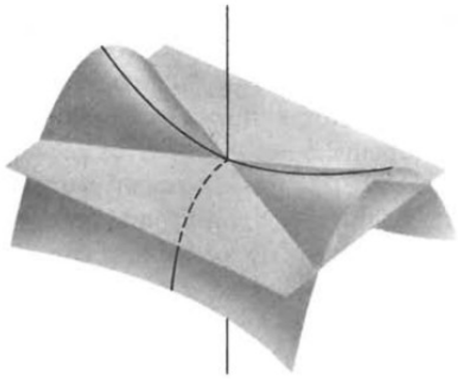

FIGURE A.12

$K = k_1k_2$ of these maximum and minimum curvatures is now called the *Gaussian curvature* (after Gauss, who first studied it), or simply "the curvature," of the surface at the point P. Once again, K will in general change as P varies over the surface; if K happens to stay constant, we obtain the three geometries discussed in Chapter 10, according as K is negative (pseudosphere), zero (plane), or positive (sphere).

In Figure A.12, the tangent plane, normal line, and principal curves for a saddle-shaped surface are shown. For point P on this surface the Gaussian curvature will be a negative number, according to our convention, since the osculating circles for the two principal curves lie on different sides of the tangent plane. On the other hand, for the egg-shaped surface in Figure A.13, the two principal curves lie on the same side of the tangent plane, so the Gaussian curvature is positive.

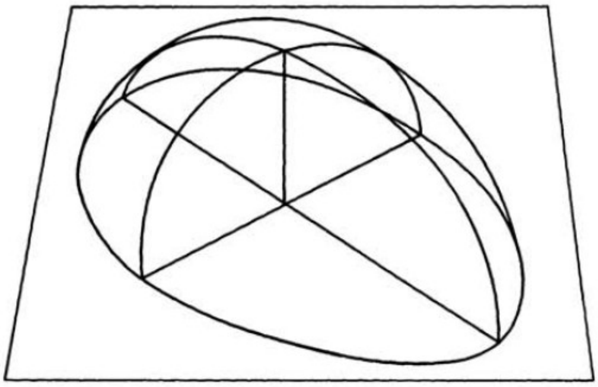

FIGURE A.13

Refer back to the picture of a pseudosphere in Figure 10.12, which is obtained by revolving a tractrix around its asymptote. At any point on this surface, it can be shown that the two principal curves are the (horizontal) circle of revolution through that point and the (vertical) tractrix through that point. Since these curves lie on opposite sides of the tangent plane, we see that the surface curvature is negative. As we move up the surface, the circle shrinks and its curvature increases indefinitely, while the tractrix flattens and its curvature decreases to zero (the curvature of its asymptote); this makes it plausible that the product K of the principal curvatures could stay constant (but of course the proof of that requires a calculation).

Similarly, in the case of a cylinder obtained, say, by rotating a vertical line around a parallel vertical line, the principal curves at any point are the horizontal circle and the vertical line, which has curvature $k_2 = 0$; hence, the Gaussian curvature $K = k_1k_2$ at any point on a cylinder will also be zero (see Figure A.14). We can better grasp this surprising result if we think of a cylinder as a "rolled-up plane." Surely, in any sensible definition of surface curvature a flat plane should be assigned zero curvature. In the process of "rolling-up" a rectangular plane strip the arc lengths and angles between curves on the strip are not changed, and in this sense the "intrinsic geometry" is not changed. Gauss was looking for a definition of surface curvature that depended only on the intrinsic geometry of the surface, not on the particular way the surface was embedded in Euclidean three-space. He was able to prove that his curvature K did not change if the surface was subjected to a "bending" in which arc lengths and angles of all curves on the surface are left invariant. Thus, K describes the intrinsic curvature of the surface independent of the way it is bent to fit into Euclidean three-space. This is all the more remarkable because the principal curvatures k_1 and k_2 may change under such a "bending"; nevertheless, their product $K = k_1k_2$ stays the same. Gauss was so excited about this result that he named it the *theorema egregrium,* "the extraordinary theorem." In a letter to the astronomer Hansen he wrote: "These investigations deeply affect many other things; I would go so far as to say they are involved in the metaphysics of the geometry of space."

Gauss also solved the problem of determining this intrinsic curvature K without reference to the ambient three-space. Imagine a two-dimensional creature living on a surface and having no conception of a

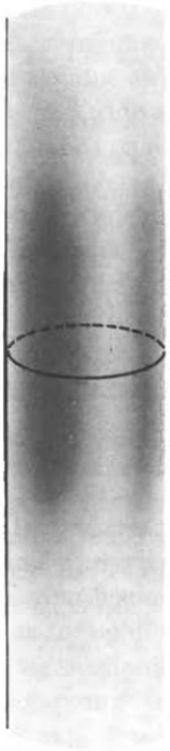

FIGURE A.14 Gaussian curvature of a cylinder is zero.

third dimension, being unable to conceive of the normal lines we used to define the curvature K. How could this creature calculate K? We will have to use the language of the differential calculus to give Gauss' answer.

In the Euclidean plane a point is determined by its x and y coordinates. If these coordinates are subjected to infinitesimal changes denoted dx and dy, then the point moves an infinitesimal distance ds whose square is given by the Pythagorean formula $ds^2 = dx^2 + dy^2$. Now on a smooth surface a point will also be determined locally by two coordinates x and y. If these coordinates are subjected to infinitesimal changes dx and dy, then the point moves a distance ds on the surface whose square is given by the more complicated expression

$$ds^2 = E\,dx^2 + 2F\,dx\,dy + \mathrm{G}\,\mathrm{dy}^2,$$

where E, F, and G may vary as the point varies. The functions E, F and G could in principle be determined by the two-dimensional creature making measurements on his surface. Gauss then showed that his curvature K is given in terms of E, F, and G by a not-too-complicated formula (see Lanczos, 1970, p. 183). Thus, the creature could also calculate K from this formula and discover that his world was curved, although he would have difficulty visualizing what that might mean.

Although this talk about a two-dimensional creature may seem bizarre, it is not, as Riemann demonstrated. Riemann reasoned that we are in an entirely analogous situation, living in a three-dimensional universe in which an infinitesimal change in distance ds is given by an analogous formula involving the three infinitesimals dx, dy, and dz:

$$ds^2 = g_{11}\, dx^2 + g_{22} dy^2 + g_{33} dz^2 + 2g_{23} dy\, dz + 2g_{31} dz\, dx + 2g_{12} dx\, dy.$$

From this formula Riemann was able to define a "curvature tensor" analogous to the Gaussian curvature for a surface, only more complicated: Gauss' curvature involved only a single number K, whereas Riemann's depended on six different numbers. Riemann discovered this curvature almost accidentally in his research on heat transfer. In fact, he developed such a curvature tensor for abstract geometries of any dimension n, and Einstein was able to apply Riemann's ideas to his four-dimensional space-time continuum.

So we are in the same position as that poor two-dimensional creature. We can make measurements to determine the Riemannian curvature of our universe. Astronomers have been performing such measurements. If we find that the Riemannian curvature is not zero, we know that the geometry is not Euclidean. However, this does not mean that our space is embedded in some higher-dimensional physical space in which it is somehow "curved." When we say, loosely, that "space is curved," we mean only that its geometric properties differ from the properties of Euclidean space in a very specific way given by Riemann's formulas.

It was in his 1854 inaugural lecture *Ueber die Hypothesen welche der Geometrie zugrunde liegen* ("On the hypotheses that form the foundation for geometry") that Riemann introduced the idea of an n-dimensional space whose intrinsic geometry is determined by a quadratic formula for the infinitesimal change in distance ds. Such a structure is now called a *Riemannian manifold* (see Spivak, 1970, or Do Carmo, 1976, for the definition). Different manifolds yield different geometries, so that Riemann discovered an infinite number of new geometries.

Let us return to the two-dimensional case of a surface Σ. For each point P on Σ and each curve γ lying on Σ and passing through P, the curvature vector $\mathbf{k}$ decomposes naturally into a vector sum

$$\mathbf{k} = \mathbf{k}_n + \mathbf{k}_g$$

of its projection $\mathbf{k}_n$ on the normal line and its projection $\mathbf{k}_g$ on the tangent plane T to Σ at P; these projections are called, respectively, the *normal* and *tangential curvature vectors.* The length of $\mathbf{k}_g$ is called the *geodesic curvature* k_g of the curve γ at P relative to Σ. We call γ a *geodesic* if $k_g = 0$. In that case, γ has *zero curvature relative to the surface* Σ; it may have nonzero curvature relative to Euclidean three-space, but then its curvature vector points along the normal line to the surface at P. Another way to describe a geodesic γ that has nonzero $\mathbf{k} = \mathbf{k}_n$ is that its osculating plane Π at P is perpendicular to the tangent plane T (since Π contains the normal line at P). From this description we see immediately that the geodesics on a sphere are its great circles ($\Pi \perp T$ iff Π passes through the center of the sphere); Figure A.15.

This description of geodesic curvature refers to the ambient Euclidean three-space. But F. Minding in 1830 showed that it too is an intrinsic quantity for Σ: it depends only on the functions E, F, and G and the curve γ. Hence the notion of "geodesic" can be defined on a Riemannian manifold. And it gives us the correct interpretation of the heretofore confusing term *straight line* on such a manifold. Riemann recognized that the geodesics would be of finite length if the curvature of the surface was greater than some positive constant.

A proposed alternative interpretation of the term "straight line

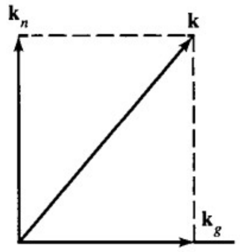

FIGURE A.15

segment" is "the shortest path on the surface joining two points on the surface." It can be proved that *if such a shortest path exists, it must be an arc of a geodesic.* But a shortest path may not exist: let Σ be a punctured plane or a punctured sphere; two points on opposite sides of the puncture cannot be joined by a shortest path on Σ. On *complete* manifolds (meaning that every Cauchy sequence of points converges), it can be proved that shortest paths always exist (theorem of Hopf-Rinow). However, an arc of a geodesic need not be the shortest path: consider the longer great-circular arc joining two nonantipodal points on a sphere, or consider the helical arc joining two points on a vertical line on the cylinder in Figure A.16. So the definition of "geodesic" (found in several books by nonmathematicians) as "the shortest path" is inadequate because it excludes such arcs.

It is important to note that, although we have emphasized Riemann's generalization of Gauss' ideas from two dimensions to higher dimensions, Riemann's formulation gives new information about *surfaces* that cannot be embedded in Euclidean three-space. For example, the hyperbolic plane can be described as a complete two-dimensional Riemannian manifold of constant negative curvature, and the elliptic plane can be described as a complete two-dimensional Riemannian manifold of constant positive curvature on which any two points lie on a unique geodesic. Neither of these manifolds can be analytically embedded in Euclidean three-space.

Some idea of Riemann's influence on modern mathematics can be gleaned from the following list of concepts, methods, and theorems

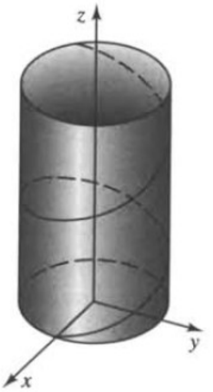

FIGURE A.16 A helix is a geodesic on a cylinder.

that have been named after him: Riemannian curvature of Riemannian manifolds, Riemann integral, Riemann-Lebesgue lemma, Riemann surfaces, Riemann-Roch theorem, Riemann matrices, Riemann hypothesis about the Riemann zeta function, Riemann's method in the theory of trigonometrical series, Riemann's method for hyperbolic partial differential equations, Riemann mapping theorem, and Cauchy-Riemann equations.[2]

---

[2] For the story of Riemann's difficult life, see Bell (1961); for a mathematically precise treatment of Riemannian geometry, including an explanation of what Riemann said in his famous 1854 inaugural dissertation lecture, see Spivak (1970), Vol. 2; for applications to the theory of relativity, see Lanczos (1970); for an intuitive discussion of how gravity can be explained in terms of the curvature of space, see the last part of Taylor and Wheeler (1992).

)-

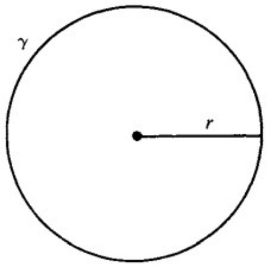

. - " 7

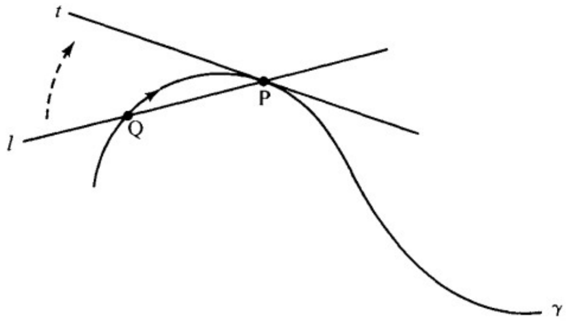

. - " T

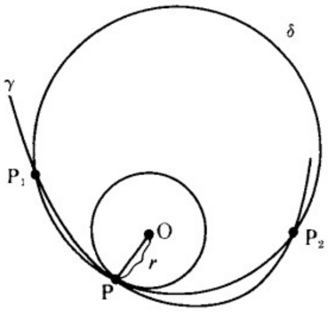

. - " <

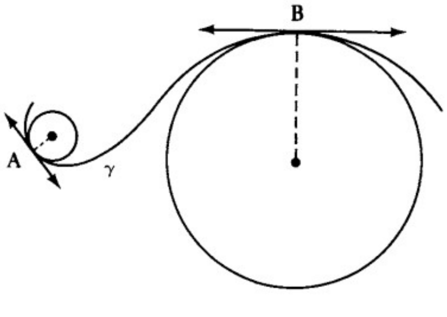

. - "

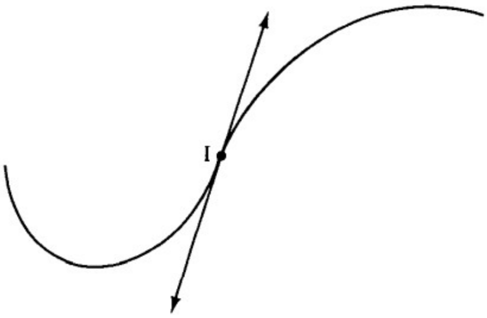
I

. - "

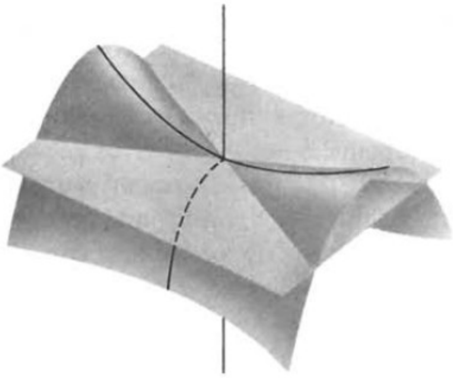

. - "

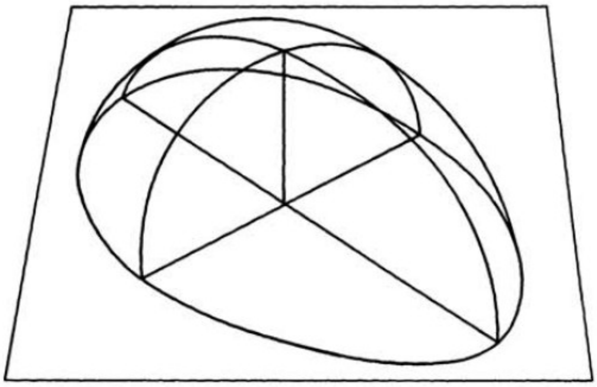

. - "

. - " 6

*)) %)) + F

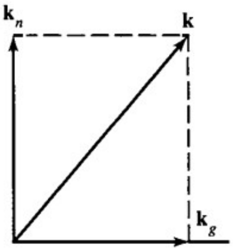
$\mathbf{k}_n$
$\mathbf{k}$
$\mathbf{k}_g$

. - " /

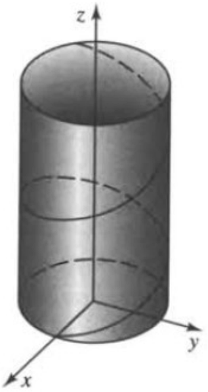

. - " N

" ;) -) +

5- 6/6267 3

APPENDIX B
GEOMETRY WITHOUT CONTINUITY

> Any serious student should become familiar with the great discovery, made at the end of the last century, that large parts of geometry do not depend on continuity.
>
> H. BUSEMANN AND P. J. KELLY

If Euclid could have returned to life at the beginning of the twentieth century and heard the criticisms made of the gaps in his work, he undoubtedly would have accepted the incidence, betweenness, and congruence axioms that replaced his first four postulates. A model of those axioms is called an H-*plane* ("H" for Hilbert, Hessenberg, and Hjelmslev). However, it is easy to imagine Euclid bristling at Dedekind's continuity axiom, although he could be expected to admit Archimedes' axiom and the elementary continuity principle.

Much work has been done in the last century to develop geometries without continuity assumptions. The most comprehensive such treatise is F. Bachmann's *Aufbau der Geometrie aus dem Spiegelungsbegriff* (Construction of Geometry based on the Concept of Reflection, 1973), in which plane geometries based only on axioms of incidence, perpendicularity, and reflections are studied.[1] Bachmann is justified in calling this study *plane absolute geometry,* since it includes elliptic, hyperbolic, and Euclidean geometries as special cases (as well as other unusual geometries). Bachmann has succeeded in dispensing with betweenness axioms as well as continuity axioms. It is unfortunate that he calls models of his axioms "metric" planes—since "metric spaces" have a quite different meaning in which a "metric" measures distance by real numbers; I propose calling them *absolute planes.*

---

[1] For a presentation of Bachmann's axioms in English, see Ewald (1971). I. M. Yaglom, in the foreword to the Russian translation of Bachmann's book, calls it "indisputably the most significant development in the foundations of geometry in decades." You were introduced to Bachmann's methods on pp. 340–342 and in Exercise 50 of Chapter 9.

The key ideas for this study are:

1. Embed the absolute plane in a projective plane.
2. Show that the points and lines in this projective plane have homogeneous coordinates from some field K (as in Major Exercise 10, Chapter 2).
3. Show that two lines are perpendicular if and only if their coordinates satisfy a particular homogeneous quadratic equation.
4. Use algebra to get further information about the absolute plane.

This program has been completely successful in classifying H-planes (the problem of determining all absolute planes remains unsolved). In particular, it is shown that (up to isomorphism) the only H-planes in which Archimedes' axiom and the elementary continuity principle hold are either the Cartesian model or the Klein model over an Archimedean Euclidean field K (Archimedean ordered field in which every positive element has a square root). This result is particularly remarkable because it gives the existence of limiting parallel rays in the non-Euclidean case without appealing to Dedekind's continuity axiom (compare the proof of Theorem 6.6, p. 196).

Let us describe the embedding briefly. The intuitive idea comes from our discussion of the Klein model, in which new "ideal" and "ultra-ideal" points are introduced so that lines that were previously parallel now meet when extended through these new points (see Major Exercise 13, Chapter 6, and p. 240). Abstractly, the new points are simply pencils of lines. The old points are in one-to-one correspondence with the pencils of the first kind, i.e., A corresponds to the pencil p(A) of all lines through A.

An ultra-ideal point is a pencil of the second kind; it is a pole P(t), which for a fixed line t consists of all the perpendiculars to t. But the pencils of the third kind must be described carefully to avoid circular reasoning. We described such a pencil in Chapter 9 (p. 340) as consisting of all lines through a fixed ideal point, the ideal point being defined as an equivalence class of limiting parallel rays. Since we don't know limiting parallel rays exist now, we instead use the theorem on three reflections (Proposition 9.19) as our definition: the pencil $p(lm)$ determined by parallel lines l and m which do not have a common perpendicular consists of all lines n such that the product

$$R_l R_m R_n$$

is a reflection. Certain properties that were previously obvious now require a considerable amount of ingenuity to prove — for example, if $h, k \in p(lm)$, then $p(hk) = p(lm)$. Hjelmslev discovered how to prove that a pencil of the first kind $p(\text{D})$ and a pencil of the third kind $p(a'c')$ have a unique line b in common (see Figure B.1). He dropped perpendiculars a, c from D to a', c' at points A, C, then dropped perpendicular d from D to $\overleftrightarrow{AC}$. Then the line b is uniquely determined by the equation

$$R_a R_b R_c = R_d.$$

We now know the new points of our projective plane. Our old lines a can be extended to new lines $l(a)$ so that any pencil containing a is by definition incident with $l(a)$. But we need more new lines to fill out our projective plane. For example, for any old point A, the polar $l(\text{A})$, consisting of all poles of lines through A, should be a new line; in the Klein model, this line lies entirely outside the absolute circle (in the Cartesian model, it's the line at infinity); see Figure B.2.

But how are we to describe the lines in the Klein model that are tangent to the absolute circle, and how are we to verify the axioms for a projective plane? To accomplish this, Hjelmslev discovered a remarkable device: he fixed a point O and fixed a pair u, v of nonperpendicu-

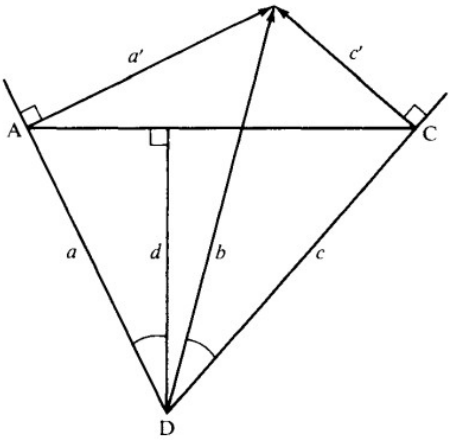

FIGURE B.1

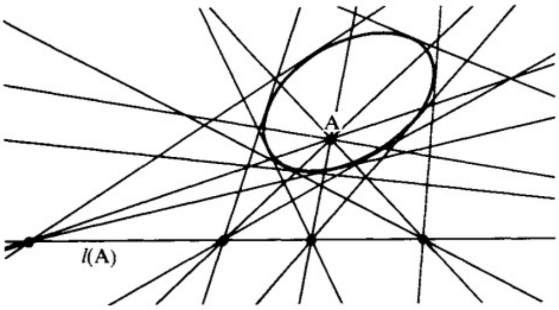

FIGURE B.2 The polar of a point.

lar lines through O — think of these lines as determining an acute angle θ. He then defined a transformation that fixed O and sent any $A \neq O$ to the midpoint A^* of the segment joining A to its image under rotation $R_u R_v$ about O through angle 2θ (Figure B.3); this transformation is called the *half-rotation* (or *snail map*) about O corresponding to u, v.

Hjelmslev observed that in the Klein model, half-rotations extend to collineations of the projective plane, and that any projective line except $l(O)$ could be mapped onto an extended Klein line $l(a)$ by a suitable half-rotation about O. So he proposed to call a set of pencils a "line" if it is mapped onto some $l(a)$ by some half-rotation about O, or if it is $l(O)$. With this definition, he was then able to verify the axioms for a projective plane.

The execution of idea 2 — construction of the field K of coordinates — requires even more technique. The key tool is a complicated theorem of Hessenberg, which generalizes the Euclidean theorem that tells when a quadrilateral can be circumscribed by a

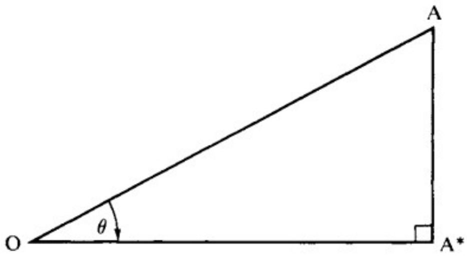

FIGURE B.3

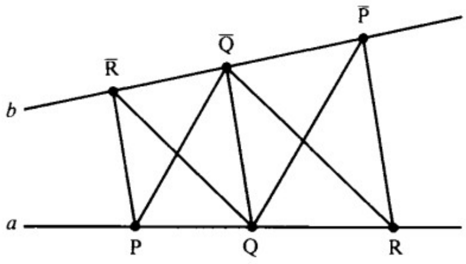

FIGURE B.4 Pappus' theorem.

circle. The method of constructing K is the standard method for any affine plane in which Pappus' theorem holds (Ewald, 1971, Chapter 3); here the affine plane is obtained by removing the polar $l(\mathrm{O})$ of O, and Pappus' theorem states: Let P, Q, R lie on a, $\overline{\mathrm{P}}$, $\overline{\mathrm{Q}}$, $\overline{\mathrm{R}}$ lie on b, such that a does not meet b in any of these six points (Figure B.4). If $\mathrm{P}\overline{\mathrm{Q}} \parallel \overline{\mathrm{P}}\mathrm{Q}$ and $\mathrm{Q}\overline{\mathrm{R}} \parallel \overline{\mathrm{Q}}\mathrm{R}$, then $\mathrm{P}\overline{\mathrm{R}} \parallel \overline{\mathrm{P}}\mathrm{R}$. Bachmann's approach is to verify Pappus' theorem by brute force; there is another approach due to Lingenberg that I believe gives more insight (see Lenz, 1967, p. 206 ff., on the Euclidean "pseudoplane"). Pappus' theorem implies that the definition given for "lines" in our projective plane does not depend on the choice of O.

If we started with an H-plane, it is then possible to define an order on K in terms of the betweenness relation in the H-plane. We need only specify the set P of positive scalars (since $x < y \Leftrightarrow y - x > 0$).

Choose any two points A, B from the H-plane. For any third point X collinear with A, B, the ratio

$$\mathrm{AX}:\mathrm{BX}$$

is defined to be that unique scalar in K which, multiplying the vector from B to X, gives the vector from A to X. We then define P to consist of 1 and all ratios AX : BX as X runs over all third points on the affine line $\overleftrightarrow{\mathrm{AB}}$ that do *not* lie between A and B (this includes all ideal and ultra-ideal points on $\overleftrightarrow{\mathrm{AB}}$). It can be shown from invariance of the ratio under parallel projection that this definition of P does not depend on the choice of A, B.

Moreover, a theorem of Menelaus (compare Exercise H-5, Chapter 7) can be proved for these ratios, and can be used to demonstrate that P has all the properties required for a set of positive numbers that makes K an *ordered field* (see Lenz, 1967, p. 223 ff.). In turn, this enables us to extend the betweenness relation to all triples of collinear points in the affine plane, and to show that the H-plane is embedded as a convex open subset of the affine plane, containing the origin O. The embedding is locally Euclidean at O in that perpendicularity for lines through O has the familiar meaning from Cartesian analytic geometry.

It can be proved (from free mobility—Lemma 9.3, p. 358) that the field K is Pythagorean (the sum of two squares is a square), but K is not Euclidean unless the elementary continuity principle holds (a neat algebraic criterion for this geometric property).

As for idea 3 on our list, further argument shows that there is a constant $k \in K$ such that lines having homogeneous coordinates $[a_1, a_2, a_3]$, $[b_1, b_2, b_3]$ are perpendicular if and only if

$$a_1b_1 + a_2b_2 + ka_3b_3 = 0.$$

Here "perpendicularity" has been extended to all pairs of lines in the projective plane, and there may exist certain lines (called *isotropic*) that are perpendicular to themselves (e.g., when $k = -1$, the line $[0, 1, -1]$, i.e., $y = 1$). To each line $[a_1, a_2, a_3]$ is associated its *pole* $[a_1, a_2, ka_3]$ (except for line at infinity $[0, 0, 1]$ when $k = 0$), and we see that the perpendiculars to this line are precisely the lines passing through its pole; isotropic lines pass through their own poles, and for $k \neq 0$, the locus of poles of isotropic lines is given by the affine equation

$$x^2 + y^2 = -k^{-1}$$

and may be called *the absolute*. The points (x, y) in the H-plane satisfy the inequality $|k|\,(x^2 + y^2) < 1$.

The fourth angle of a Lambert quadrilateral is acute, right, or obtuse accordingly as $k < 0$, $k = 0$, or $k > 0$. If $k = 0$ and the H-plane fills up the affine plane, we obtain the *generalized Euclidean plane* over K. However, if $k = 0$, if K is non-Archimedean, and if we restrict our H-plane points to those (x, y) that are infinitesimal, we obtain Dehn's *semi-Euclidean plane*, in which the Euclidean parallel postulate fails but rectangles exist.

If $k = -1$, the absolute is the unit circle, and if every point in its interior belongs to the H-plane, we obtain the *generalized hyperbolic plane* over K, provided K is Euclidean. In case K is non-Archimedean, we can again restrict (x, y) to infinitesimal values, obtaining Schur's *semihyperbolic plane,* in which limiting parallel rays do not exist, even though the elementary continuity principle holds (the ideal points fill up an entire region inside and on the absolute).

If neither k nor $-k$ is a square (so K is not Euclidean), yet every sum $1 + kx^2$ is either a square or k times a square, the points (x, y) satisfying $x^2 + y^2 < |k^{-1}|$ form a *semielliptic* H-plane, in which parallel lines have a unique common perpendicular (the absolute is empty).

Finally, if K is non-Archimedean, $k = 1$, and (x, y) are again restricted to infinitesimal values, we obtain Dehn's *non-Legendrian plane,* in which the fourth angle of a Lambert quadrilateral is obtuse.

You can see that there are some unusual H-planes, most of them living on non-Archimedean fields; on Archimedean fields, only Euclidean, hyperbolic, and semielliptic planes survive. For the complete algebraic classification of H-planes (discovered by W. Pejas, 1960), see Hessenberg and Diller (1967). For an introduction to absolute plane geometry in English, see Chapter 5 of Bachmann, Behnke, and Fladt (1974). Finally, for an expanded version of this appendix, see my article "Euclidean and Non-Euclidean Geometries without Continuity," *American Mathematical Monthly,* November 1979.

E)-

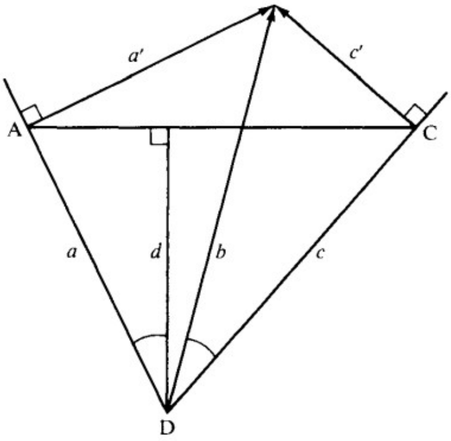

. -   &

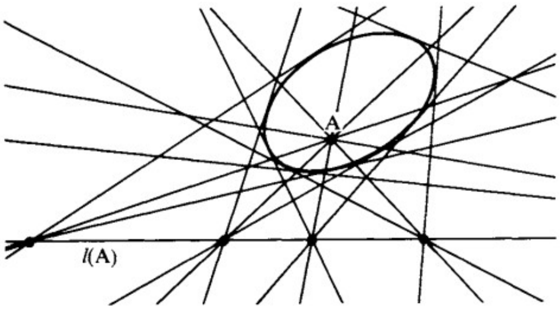

. -   &
4)   ) 4

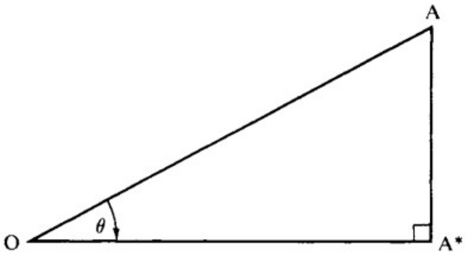

. -   &

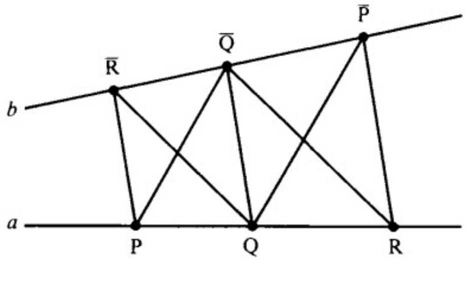

. -   & 6   )44 q

!"#" $ % % & ' ( )) * +) , -)
.) @)) $ @) * + " ") + "44) 2 !.! % 3 4

,9 Y) % RRR)- 2" (+) 3

!"#" $ % % & ' ( )) * +) , -)
.) .) @)) $ @) * + 2 !.! % 3 4

5- T 2</3 / 6 #);)

by Axiom 5 and hypothesis $B*Y*P$ and $B*P*A$, then by Theorem 7 $B*Y*P*A$ and, hence, $B*Y*A$. By Axiom 5 $A*Y*B$. ●

We shall certainly want a point on a segment to "divide" the segment into two parts. We can now prove a theorem analagous to Theorem 12.

Theorem 21. Any point P on a segment $A\text{–}B$ separates $A\text{–}B$ into two nonempty subsets $A\text{–}P$, $P\text{–}B$ which together with P form a partition of $A\text{–}B$.

Proof: By Theorems 16 and 20 $A\text{–}P$, $P\text{–}B$ are nonempty and subsets.

(a) Partition: Let P be any point on $A\text{–}B$ and $X \neq P$ any point on $A\text{–}B$. We want to show that X is an element of one and only one of the subsets $A\text{–}P$, $P\text{–}B$. By Theorem 18 we are permitted to apply Axiom 8. By Axiom 8, if $A*P*B$ then $A*P*B*X$, $A*P*X*B$, $A*X*P*B$, or $X*A*P*B$. But since $A*X*B$ we have $A*P*X*B$ or $A*X*P*B$ but not both. Thus, for any point X either $P*X*B$ or $A*X*P$ but not both; hence, any point X of $A\text{–}B$ is on one and only one of the sets $A\text{–}P$, $P\text{–}B$. Therefore, $A\text{–}P$, $P\text{–}B$ together with P (Theorem 15) form a partition of $A\text{–}B$.

(b) Separation: Now let X_1, Y_1 be two distinct elements of set $A\text{–}P$. Then $A*X_1*P$ and $A*Y_1*P$. By Axiom 5 we may write $P*X_1*A$ and $P*Y_1*A$, from which it follows by Theorem 10a that $P*X_1*Y_1$ or $P*Y_1*X_1$; therefore it follows from Axiom 7 that X_1*P*Y_1 cannot hold. Similarly for X_2, Y_2 as elements of $P\text{–}B$.

Suppose now that X_1, X_2 are any two elements, one from each set; it follows from part (a) that they are distinct. We want to show that X_1*P*X_2. $A*X_1*P$ and $A*P*B$ imply $A*X_1*P*B$ and, hence X_1*P*B. X_1*P*B and $P*X_2*B$ imply X_1*P*X_2*B; hence X_1*P*X_2. ●

EXERCISES 5.3

1–5. Prove Theorems 15–19. (*Theorem 19)

*6. Prove that a segment has an infinite number of points. (Hint: by contradiction; if not, there must exist an empty segment.)

7. Prove that a line has an infinite number of points.

5.4 The Axiom of Pasch

In many standard proofs of high-school geometry, lines through a vertex of a triangle are considered—angle bisectors, for example, as illustrated in Figure 5.7. In all such cases it is *assumed* that these lines intersect the opposite side of the triangle, but on what basis is such an assumption

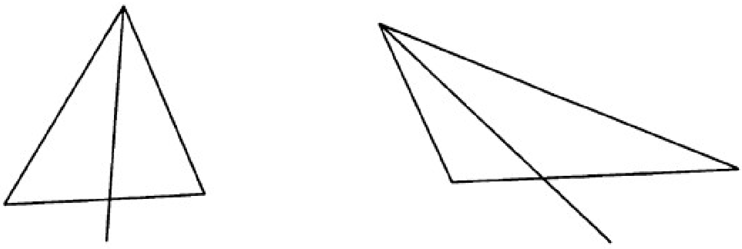

FIGURE 5.7

justified? Pasch, in the late nineteenth century, is credited with first maintaining not only that such a statement needs justification but that it is impossible to prove from the standard Euclidean axioms. So he introduced the following axiom, which is illustrated in Figure 5.8.

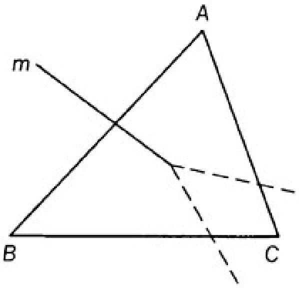

FIGURE 5.8

Axiom 10. ***(Pasch)*** If A, B, C are three distinct and noncollinear points, m is a line not passing through any of them, and m contains a point of segment $A\text{–}B$, then m also contains a point of segment $A\text{–}C$ or $B\text{–}C$.

This axiom does not immediately solve the problem about angle bisectors; before we can solve such a problem we must first be able to state it precisely. Note, as in Figure 5.9, that not every line through a

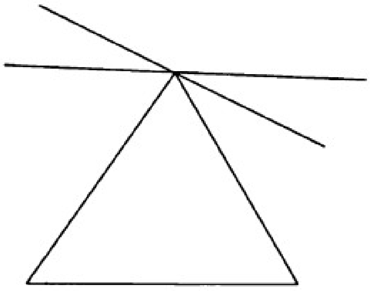

FIGURE 5.9

vertex of a triangle passes through the opposite side; it must also pass through a point *inside* the triangle.

In order to express and prove such a theorem we shall want definitions of "triangle" and "inside," as well as some related theorems. Thus we shall postpone some of this until we have established the groundwork. Let us start with a theorem that shows that the alternative expressed in Pasch's axiom is an exclusive one. See Figure 5.10.

Theorem 22. If A, B, C are noncollinear points, any line m containing a point of segment A–B, and a point of segment A–C, cannot contain a point of segment B–C.

Proof: Suppose that there exists a line m with three such points; that is, a line with points D, E, and F such that D belongs to A–B, E belongs to B–C, and F belongs to A–C. By Axiom 7 one point, say F on A–C must lie between the others, so $D{*}F{*}E$. Since E is contained in B–C, it is contained in line BC—this follows from Theorem 18. Similarly, D is contained in line AB. It follows, therefore, that D, E, and B are noncollinear, for otherwise Theorem 4 is contradicted. (Why)? Thus by Axiom 10 the line AC which intersects D–E at F must intersect either B–E or B–D. But by Theorem 4 and Axiom 7 this is impossible. Hence, the assumption is false. The argument proceeds similarly if $D{*}E{*}F$ or $F{*}D{*}E$. ●

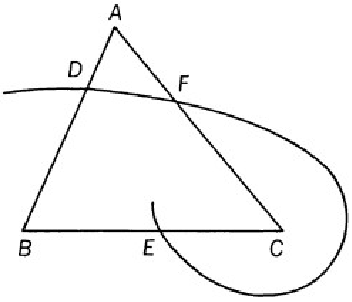

FIGURE 5.10

We might have introduced the following definition at the beginning of the section. By postponing it until now, we have shown merely that this definition is unnecessary; we now hope to show that it helps us to make statements more succinctly, clearly, and in keeping with our intuition.

Definition 8. Given three distinct noncollinear points A, B, C, the new set which is the union of A, B, C together with A–B, A–C, and B–C, is called a *triangle*, and:

A, B, C are called the *vertices;* each one a *vertex.*
A–B, A–C, B–C are called the *sides.*
The lines containing the sides are called the *sidelines.*

Because of Theorems 17 and 19 and the above definition we may state the following rules:

Language Rule 9: *A triangle is* DETERMINED *by its vertices.*

Notation Rule 8: *The triangle determined by A, B, C will be denoted by:*

$$\triangle ABC$$

We shall read this and often denote it "triangle ABC."

It should be noted that because endpoints are not parts of segments, vertices are not parts of sides of triangles.

We may now state Axiom 10 in the simpler form:

If a line not through a vertex intersects one side of a triangle, it intersects at least one of the other sides.

Theorem 22 may be restated in either of the following more intuitive forms:

If a line intersects one side of a triangle, it intersects at most one of the other sides.

A line cannot intersect all three sides of a triangle.

The machinery has now been built to enable us to prove a very important theorem. But first we shall want a definition and two preliminary theorems:

Definition 9. Let m be any line and P a point not on it. Let S_1 be the set of all points consisting of P and all points X not on m such that P–X contains no point of m. Let S_2 be the set of all points Y, such that P–Y contains a point of m. Then S_1, S_2 are called *half planes of line m.*

Theorem 23. Half planes of a line m are nonempty.

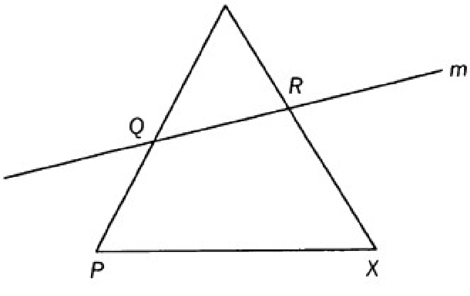

FIGURE 5.11

Proof: Given a line m, there exists a point P not on m and a point Q on m. There exists a point Y such that $P*Q*Y$ and, hence, P–Y contains a point of m. There exists another point R on m and a point X such that $Y*R*X$. Then by Theorem 22 m cannot contain a point of P–X. ●

Theorem 24. There exist exactly two half planes of a line m (relative to a point P) which together with m form a partition of the set of all points.

Proof: Every point is either on m or not. No point is both. Consider the points not on m. By the argument of Theorem 23, if M is a point of m, then there exist points P, Q not on m and such that $P*M*Q$. Let X be any point not on m. If X is collinear with P and Q, then exactly one of the following holds: $X*P*Q$, $P*X*Q$, $P*Q*X$, $X = P$ or $X = Q$. Then exactly one of the segments Q–X, P–X meets m. (Proof in the first three cases is nontrivial and is left as an exercise.) If X is not collinear with P and Q, then once again exactly one of the segments Q–X, P–X meets m.

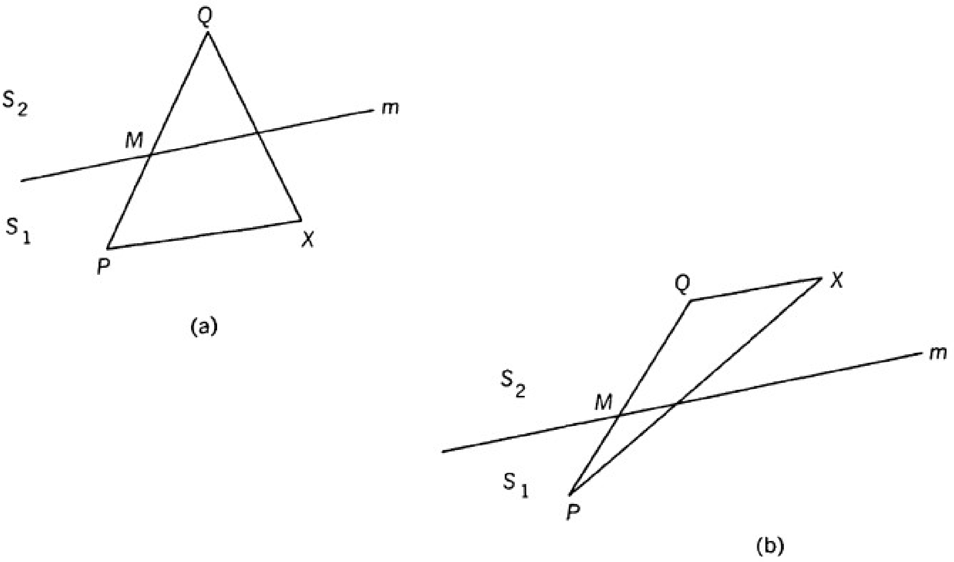

(a)

(b)

FIGURE 5.12

For if neither meets m, then since $P*M*Q$, Axiom 10 is contradicted. If both P–X, Q–X meet m, then Theorem 22 is contradicted. Thus, if P–X does not meet m, X is in S_1 as given in Definition 9; and if Q–X does not meet m, then P–X does and X is in S_2 as given in Definition 9. Since these are the only possibilities and each point is in exactly one of the sets m, S_1, S_2 the theorem is proved. ●

It is now possible to prove one of the basic theorems of the foundations of Euclidean geometry. It is often assumed as an axiom (in place of

Pasch's) and called *The Plane Separation Postulate:* any line separates the set of all points into two half planes. To understand the proof, however, it may be necessary to review the meaning of "separates."

Definition 5 is stated in terms of any set S separating two others, S_1 and S_2. S_1 and S_2 must be distinct nonempty disjoint sets which are disjoint from S and are such that two conditions hold: if A is an element of S_1 and B is an element of S_2 there exists a point of S between them; conversely, if there exists a point of S between A and B, they must belong to distinct sets. The converse, if stated in contrapositive form, says: if A and B belong to the same set, there does not exist a point of S between them.

Theorems 23 and 24 show that half planes of a line m, together with m, satisfy the properties of S_1, S_2, and S of the definition. It must now be shown that they satisfy case (a) and case (b) of Definition 5.

Theorem 25. Any line m *separates* the set of all points into two half planes.

Proof: By Theorems 23 and 24 we know that the half planes are nonempty sets disjoint from each other and m. Let P be a point of one half plane, and let A, A' be any two points not on m such that A–P, A'–P contain no point of m. Let B, B' be any two points of the other half plane; that is, not on m but such that P–B, P–B' do contain a point of m.

Case a. If any 3, 4, or 5 of the points P, A, A', B', B are collinear, then it follows that m intersects A–B, A'–B', A'–B, A–B'. (Proof left as an exercise.) If no three of the points are collinear, then it follows directly from Axiom 10 that m intersects A–B, A'–B', A'–B, A–B'.

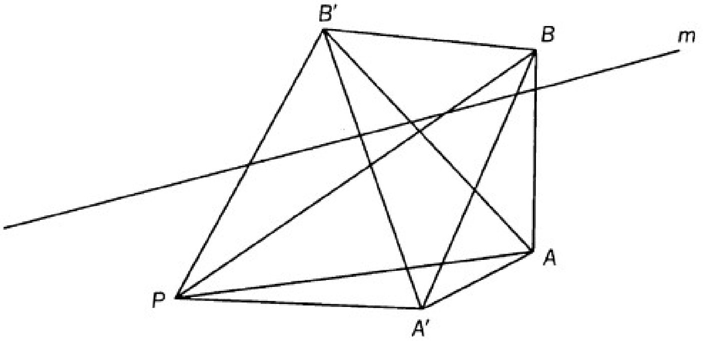

FIGURE 5.13

Case b. The converse says that if A–B intersects m, then A, B belong to different half planes of m. In its contrapositive form, this says that if A, A' belong to the same set, then A–A' does not intersect m. If either A, P, A' are collinear, or B, P, B' are collinear, then it follows that neither

A–A' nor B–B' intersect m. (Proof left as an exercise.) If A, P, A' are not collinear, then since by hypothesis neither P–A nor P–A' intersect m, it follows that Axiom 10 would be contradicted if m intersected A–A'. If B, P, B' are not collinear, then since by hypothesis P–B and P–B' each intersect m, by Theorem 22 m does not intersect B–B'. ●

Language Rule 10: *The sets S_1, S_2 in Definition 9 are said to be* DETERMINED BY m AND P.

The following theorem establishes the uniqueness of half planes determined by a line m and any point not on it.

Theorem 26. Let m be a line, P, P' two points not on m. The half planes determined by m, P are the same as those determined by m and P'.

Proof: Left as an exercise.

Theorem 27. Let m, m' be distinct lines. The half planes of line m are distinct from the half planes of line m'.

Proof: Left as an exercise.

EXERCISES 5.4

1. Complete Theorem 24.
2. Complete Theorem 25.
*3. Prove Theorem 26.
4. Prove Theorem 27.
5. Prove that if $A*F*B$, and $B*C*D$, then there exists a point E on F–D such that $A*E*C$. A, B, C are noncollinear.
6. Prove that if A, B, C are three distinct noncollinear points and D, E points such that $B*C*D$, and $C*E*A$, then there exists a point F on line DE such that $A*F*B$.
7. In exercise 6, show that $D*E*F$.
*8. If in a $\triangle ABC$, $A*D*B$, $A*E*C$, and $B*F*C$, then A–F meets D–E.
9. Prove that in $\triangle ABC$, if $A*D*B$, $B*E*C$, then A–E and C–D intersect.

5.5 Convex Sets

Definition 10. A set S is called a *convex set* if and only if for any two points P, Q, belonging to S, the entire segment P–Q is contained in S.

Theorem 28.

(a) Every line is a convex set.
(b) Each half line of a point O is a convex set.

)-

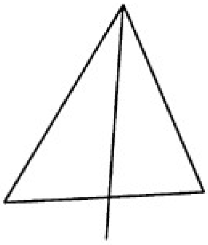

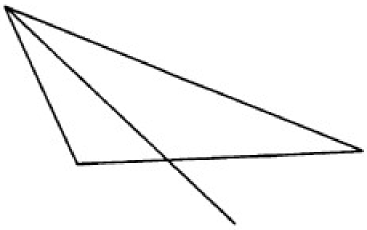

. - / 7

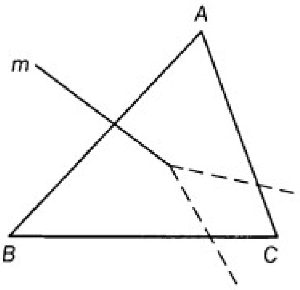

. - / T

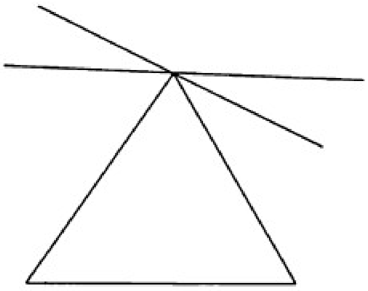

. - / <

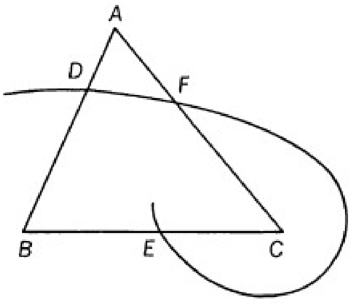

. - /

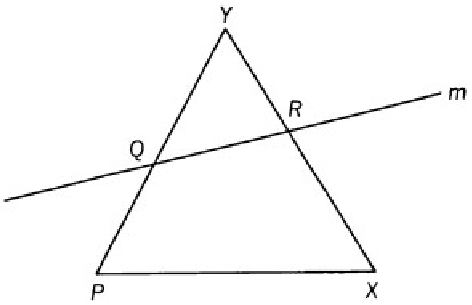

. - /

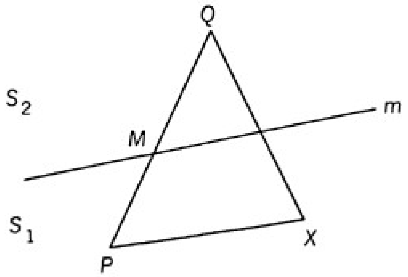

2)3

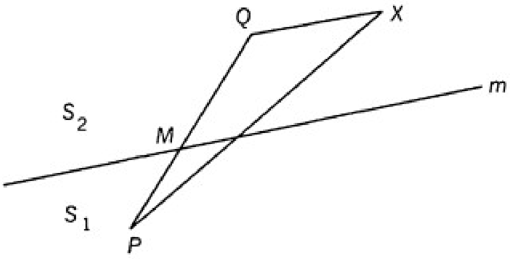

293

. - /

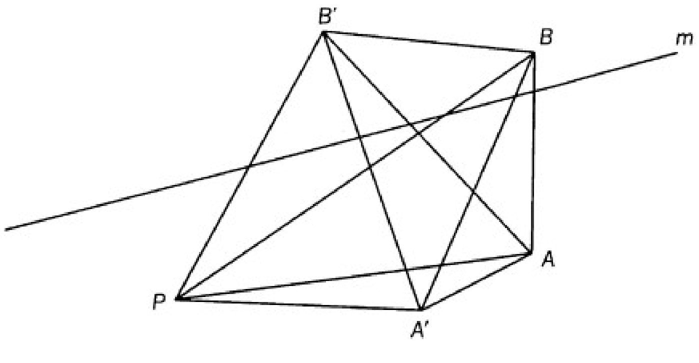

. - /

5-   N 2 7N3

part 3

Non-Euclidean Geometry

introduction

> That, if a straight line falling on two straight lines makes the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which are the angles less than the two right angles.[1]

Euclid presented this statement as his fifth postulate (axiom) and thereby profoundly influenced the course of mathematical history. Although historians disagree about the authenticity of many parts of the *Elements*, they usually agree that this statement originated with Euclid. That he placed it among his assumptions rather than introducing it as a theorem must be considered a monument to his genius.

It is possible that outside of this text the reader has never before seen this statement and is, therefore, somewhat confused as to how or why it is such an important one. It may help if we say that this statement is known as Euclid's Parallel Postulate; on the other hand, this may cause more bewilderment, for most readers undoubtedly conceive of a parallel axiom along such lines as: through a point not on a given line exactly one line exists parallel to the given line. Indeed, it turns out that this statement is equivalent to the one quoted above. But had Euclid made his statement in the more familiar form it might never have played such an important historical role.

---

[1] Whenever we wish to refer to this particular statement we shall call it the *Parallel Postulate*, or the *Fifth*.

It should be remembered that at the time of Euclid and for centuries thereafter a postulate was regarded not merely as a truth but as a self-evident truth, a truth beyond doubt. Even today this notion is widely held, as indicated by the definitions found in any standard dictionary and by the fact that many people still regard mathematics as the absolute science. This opinion was held all the more firmly in ancient times, and thus incited attacks upon the Fifth.

A look at the other postulates[2] of the *Elements* helps one to understand why the Fifth might cause such furor. Not only is it much more complicated, not only is it not self-evident, it is actually open to doubt.

It was argued that the truth of the Fifth is, at best, merely plausible. For, the arguments continued, is it not true that there exist "lines" which are asymptotic—as in the hyperbola and conchoid? And does not the fact that some lines exist which, though converging, do not meet no matter how far extended make it open to question whether the same might not be true of straight lines? If so, if open to question, then the statement cannot be a postulate.

Centuries of trying to prove the Fifth from the other postulates followed. These attempts led to the development of new geometries and to an entirely new and vastly richer approach to mathematics in general. Exactly how this came about, and a discussion of some of the results, will be the subject of the rest of the text.

---

[2] See Appendix, pp. 218–219.

9 The Concept of Parallelism

In discussing some of the attempts to prove Euclid's fifth postulate we shall learn that there are many statements equivalent to the Fifth. A survey of a list of such statements should convince the reader that the Parallel Postulate is rich in implications. He may wonder, therefore, what would occur if such a statement were discarded or negated. What would happen if we introduced a new parallel postulate? This will be seen in this chapter.

9.1 History

Although we shall not review the entire history of attempts to prove the Fifth,[1] it is necessary to discuss some of these attempts in order to learn how they have led to new discoveries. It is hoped that the following presentation will provide some answers and induce the reader to seek out others.

Proclus ***(410–485*** A.D.)

Proclus is the source of much of the knowledge we have about Euclid and his earliest critics and commentators. According to Proclus, the first attempts to prove the Parallel Postulate began almost immediately and were coordinated with revised statements of Euclid's definition of *parallel*. One should not confuse the postulate with his twenty-third definition:

---

[1] For a more complete history, see R. Bonola, *Non-Euclidean Geometry*, translated by H. S. Carslaw (Open Court Publishing Company, 1912).

Parallel straight lines are straight lines which, being in the same plane and being produced indefinitely in both directions, do not meet one another in either direction.

Of the many attempts listed by Proclus, the most significant involves changing this definition of parallel lines to: parallel lines are lines which are everywhere equidistant. How this is then used to attempt to prove the Fifth need not be discussed. For Proclus was the first to point out that there *might* exist parallel lines in Euclid's sense that are not everywhere equidistant. Thus, to use his definition, it would be necessary to prove first that if two straight lines do not meet they are equidistant. And to do this requires the Fifth Postulate or a statement equivalent to it.

As is often the case, the good critic is not necessarily an innovator. The barbs Proclus throws at the attempts of others are sharp and well-directed; but, when he tries a proof of his own, he falls equally short of success.

First, he argues that the Fifth cannot be an axiom. His reasoning, if accurately recorded, must be taken as evidence of how distorted man's thinking can become when he is obsessed with proving a point. Proclus argues, correctly, that Proposition 17 of the *Elements* is the converse of the Fifth; but he then says that it is impossible that a statement whose converse can be proved is not capable itself of being proved!(!)(!)

With reasoning such as this to justify his attempt, he then proves the Fifth by assuming a statement equivalent to it, namely, that parallel lines remain a finite distance apart.

It is for this type of argument that Proclus criticizes all others who had attempted a proof. In each case he points out that, to prove the postulate, either an assumption equivalent to it or some paradoxical assumption about infinite areas and lengths behaving like finite ones has been made. This criticism applies equally well to most later attempts.

Through the ages, in the many attempts to prove the Parallel Postulate, a long list of statements equivalent to it has been gathered. Most of the important ones were known to Proclus. Among them are the following:

1. There exist straight lines everywhere equidistant from one another.
2. If a straight line intersects one of two parallels, it will intersect the other.
3. Straight lines parallel to the same straight line are parallel to one another.
4. The sum of the angles of a triangle is equal to two right angles.
5. Two straight lines which intersect one another cannot both be parallel to the same line.

> Parallel straight lines are straight lines which, being in the same plane and being produced indefinitely in both directions, do not meet one another in either direction.

Of the many attempts listed by Proclus, the most significant involves changing this definition of parallel lines to: parallel lines are lines which are everywhere equidistant. How this is then used to attempt to prove the Fifth need not be discussed. For Proclus was the first to point out that there *might* exist parallel lines in Euclid's sense that are not everywhere equidistant. Thus, to use his definition, it would be necessary to prove first that if two straight lines do not meet they are equidistant. And to do this requires the Fifth Postulate or a statement equivalent to it.

As is often the case, the good critic is not necessarily an innovator. The barbs Proclus throws at the attempts of others are sharp and well-directed; but, when he tries a proof of his own, he falls equally short of success.

First, he argues that the Fifth cannot be an axiom. His reasoning, if accurately recorded, must be taken as evidence of how distorted man's thinking can become when he is obsessed with proving a point. Proclus argues, correctly, that Proposition 17 of the *Elements* is the converse of the Fifth; but he then says that it is impossible that a statement whose converse can be proved is not capable itself of being proved!(!)(!)

With reasoning such as this to justify his attempt, he then proves the Fifth by assuming a statement equivalent to it, namely, that parallel lines remain a finite distance apart.

It is for this type of argument that Proclus criticizes all others who had attempted a proof. In each case he points out that, to prove the postulate, either an assumption equivalent to it or some paradoxical assumption about infinite areas and lengths behaving like finite ones has been made. This criticism applies equally well to most later attempts.

Through the ages, in the many attempts to prove the Parallel Postulate, a long list of statements equivalent to it has been gathered. Most of the important ones were known to Proclus. Among them are the following:

1. There exist straight lines everywhere equidistant from one another.
2. If a straight line intersects one of two parallels, it will intersect the other.
3. Straight lines parallel to the same straight line are parallel to one another.
4. The sum of the angles of a triangle is equal to two right angles.
5. Two straight lines which intersect one another cannot both be parallel to the same line.

6. Through a given point not on a given line only one line can exist parallel to the given line.

Statement (6) is probably the form in which most readers have encountered the parallel postulate. Although known to Proclus, it is called "Playfair's Axiom" after a British geometer of the eighteenth century.

When we say that these statements are each equivalent to the Fifth we mean that they (together with other assumptions) mutually imply each other. That the Parallel Postulate implies each of these statements is evident to anyone who has studied the *Elements*, for these statements are either proved there as theorems or follow quite directly from theorems that are proved. That each of these statements in turn imply the Fifth has been shown at various times by many mathematicians; some of the proofs are quite complicated, but this should come as no surprise since, after all, it requires thirty theorems before Euclid proves any of them from the Fifth.

If we were merely attempting to prove the Parallel Postulate as a theorem in some system, there would be nothing wrong with using any of these equivalent assumptions in proving the Fifth. Because we are attempting to prove the Parallel Postulate from the other four postulates of Euclid, care must be taken. Suppose, for example, that we use the fact that the sum of the angles of a triangle is equal to two right angles to prove the Fifth, and suppose the proof is correct. The problem with such a course lies in determining whether or not the assumed statement about the sum of the angles is a true statement. It is either an axiom or a theorem. To take it as an axiom defeats the purpose of the undertaking. To take it as a theorem involves one in a circularity problem, for the only way to prove it is to use the Fifth or a statement equivalent to it.

Gerolamo Saccheri (1667–1733)

During the centuries that followed Proclus, countless other attempts were made to prove the Parallel Postulate and numerous new flaws and equivalent statements were uncovered. We shall pass over these attempts until we come to Gerolamo Saccheri, an Italian mathematician of the seventeenth and eighteenth centuries, who came up with a new approach.

Saccheri assumes the first four postulates, all the theorems proved from them (the first 28), and assumes that the Fifth is false. In other words, he attempts to prove the truth of the postulate by the indirect method.

As a basic figure in his work, he takes a quadrilateral $ABCD$, with right angles at A and B, and $A\text{–}D \cong B\text{–}C$. He then proves that the angles at C and D must be congruent to each other and goes on to consider three hypotheses:

1. $\angle C$ and $\angle D$ are right angles.
2. $\angle C$ and $\angle D$ are obtuse angles.
3. $\angle C$ and $\angle D$ are acute angles.

He proves that if any one of these hypotheses is true for a single quadrilateral it is true for every quadrilateral; he also proves that depending upon whether hypothesis 1, 2, or 3 is assumed the sum of the angles of a triangle will be, respectively, equal to, greater than, or less than two right angles. He then goes on to show that on assuming the second hypothesis a contradiction is reached, and therefore the assumption must be discarded. To complete his argument, he attempts to eliminate the third hypothesis by assuming it is true.

To state all the theorems he derives under this assumption would needlessly repeat many of the theorems of the remaining sections of this book. Rather, let it be pointed out that instead of trusting to logic, Saccheri, predisposed to find a contradiction, finds one after a circuitous search located at a point at infinity—a point, needless to say, that causes a great deal of trouble.

Saccheri himself was not satisfied with the "contradiction" he reached and tried other methods, most of which had been tried many times over the centuries. If he had placed more faith in his reasoning and less faith in his faith, he would have been credited with the discovery of non-Euclidean geometry a hundred years before it was finally discovered. As it was, his work was soon forgotten and more or less remained so until it was resurrected late in the nineteenth century.

Johann H. Lambert (1728–1777)

Johann Lambert, a Swiss mathematician of the eighteenth century, was probably aware of Saccheri's work. His method of approach is quite similar except that, instead of a figure with two right angles, he starts out with a figure with three right angles. He, too, considers three hypotheses; all are about the nature of the fourth angle:

1. The fourth angle is a right angle.
2. The fourth angle is obtuse.
3. The fourth angle is acute.

On assuming the second hypothesis he easily obtains a contradiction. But once again in assuming the hypothesis of the acute angle a contradiction does not easily follow; however, he does reach several new conclusions. He proves, as did Saccheri, that the sum of the angles of any triangle is less than two right angles (the amount of the difference they call the *defect*); but, he goes a step further than Saccheri and proves that the defect of any polygon is *proportional to the area* of the polygon.

Furthermore, he comes up with still more astounding information, namely, that under this hypothesis there would be an *absolute unit of measure*. We shall discuss this concept later.

Once again, however, intuition takes precedence over logic and Lambert concludes that, since there can be no absolute unit of measure, this hypothesis must be false.

Gauss, Bolyai, Lobachevsky

Attacks on the Parallel Postulate continued and, indeed, grew more intense until the early 1820s. At that time a subtle change in outlook and objective became discernable among some of the mathematicians concerned with the problem.

In Germany, Carl Friederich Gauss (1777–1855) had been bringing his great gifts to bear upon the problem for some time. It was unfortunate for his contemporaries that he did not publish his findings. In letters to his friends, made public after his death, he expressly asked them not to reveal his findings while he lived. In these letters one discovers that he had gone a long way toward discovering and developing some of the properties of a new geometry that he called *non-Euclidean*. It is probable that he was the first to use this name.

Although it may never be known exactly what motivated him to keep his findings secret, it is likely that fear for his reputation may have been a factor. For, on the one hand, Gauss was recognized in his own lifetime as one of the greatest of mathematicians. On the other hand, he was aware that his discoveries would be shocking to layman and mathematician alike; indeed, they ran so counter to one's intuition that devulging them might have opened him to ridicule. At the same time, keeping his work secret meant that he never received the honor that might have been his due; the credit must go instead to the son of a friend of his and a young contemporary.

In Hungary, Johan Bolyai (1802–1860), and in Russia, Nicolai Ivanovich Lobachevsky (1793–1856) had each been working on the problem from 1820–1830. At the end of the decade they finally published their findings; ironically, it did not create any great stir and neither man achieved any fame for his accomplishment during his lifetime. Gauss read Bolyai's work within a year of publication and was pleased with it; he became aware of Lobachevsky's work somewhat later and was equally impressed. However, his praise was limited to correspondence with his friends and thus his immense prestige did nothing toward making either work known or accepted.

Instead of being pleased by Gauss's praise, Bolyai was disappointed and suspicious. He knew that his father and Gauss had corresponded for years; he had written his father of his discoveries in great detail years before they were finally published. Therefore, when Gauss praised his work but added that it was of the same kind that he himself had been working on, Bolyai felt that Gauss was attempting to gain credit for his own discovery.

According to the best historical detective work, each of the three men was ignorant of what the other was doing. It is not at all uncommon in the history of mathematics and the sciences for discoveries to be made simultaneously by different men, in scattered places, each ignorant of the others' work. It would seem that when an idea is ripe it will be forthcoming.

The work of Gauss, Bolyai, and Lobachevsky will be analyzed in detail as we continue.

B. Riemann (1826–1866)

The next great breakthrough occurred in 1854. Riemann, a student of Gauss and later one of the great mathematicians of his time, presented a paper in which he made a distinction between the infinitude of a line and its unboundedness. We have had occasion to refer to this before in regard to Euclid's Proposition 16, the exterior angle theorem. Let us recapitulate.

Suppose we consider a line drawn around a sphere, that is, a circle. It is not infinite in length; it has a definite circumference. Yet it is unbounded in the sense that we can continue around it without stop. Suppose now that we not only contradict the Parallel Postulate but also the assumption that lines are of infinite length. What happens then?

Because no one had considered this before the time of Riemann, no one had observed that in Saccheri's and Lambert's work the reason that the hypothesis of the obtuse angle leads to a contradiction is that they both assume that lines are of infinite length. If this assumption had not been made they would have reached no contradiction in pursuing the hypothesis of the obtuse angle.

Riemann suggested such a geometry, a geometry in which there are no parallel lines, lines may enclose a space, and the sum of the angles of a triangle is greater than two right angles. Unfortunately, the techniques needed to deal adequately with a discussion of Riemann's suggested geometry are beyond the limited scope of this text.

Felix Klein (1849–1925)

Work related to still another type of geometry developed in the nineteenth century, projective geometry, led Klein to classify the subject into three types: parabolic, elliptic, and hyperbolic geometry. They correspond respectively to the geometry of Euclid, Riemann, and Bolyai-Lobachevsky. The classifications are commonly used today.

In 1871 Klein pointed out that there are really two types of Riemann geometry. There is the spherical type in which two straight lines have two points in common and the elliptic type where two straight lines have but one point in common; because of the intersection properties of lines, these two geometries are often classified according to Klein's scheme as double-elliptic and single-elliptic respectively.

This discovery of Klein's helps us to solve the problem raised in Chapter 7 concerning the proof of the exterior angle theorem. It will be recalled that to prove that theorem we first had to prove Theorem 75, that through a point outside a given line there exists at least one parallel. This statement is consistent with both Euclidean and Bolyai-Lobachevskian geometry but it cannot hold in Riemann's. Euclid assumes that there is exactly one parallel; Bolyai and Lobachevsky assume that there is more than one; but Riemann assumes that there is none.

However, Theorem 75 was proved without assuming any parallel postulate; therefore, some axiom or axioms of our system must be eliminated if we are to get a geometry without parallels. (See Figure 7.8.)

An analysis of the proof of Theorem 75 indicates that the contradiction comes from a line intersecting another in two distinct points. For this to be a contradiction two facts must hold: there must be two distinct points of intersection and it must be impossible for two lines to intersect in more than one point. The second condition is Theorem 4, which follows from the Axiom that two points determine a unique line; the first condition is assured by D and G being on opposite sides of line AB, which is assured by the plane separation theorem, which in turn is assured by Pasch's Axiom. One begins to see that the geometry of Riemann is a strange one indeed. It turns out that in the spherical geometry two lines intersect in two points and that it is true that every line separates the plane; in single elliptic, on the other hand, two lines intersect in exactly one point but no line separates the plane.

The Age of Hilbert

In the late nineteenth century a new approach, the beginning of critical examinations into the foundations of geometry came into being. This period also brought an interesting epilogue to the history of attempts to prove the Parallel Postulate.

It was discovered that a considerable amount of Euclidean geometry could be developed without the Postulate of Archimedes; that is, without an assumption of continuity. Saccheri and Lambert and all the others had made generous use of continuity assumptions in their arguments. Now, with the foundations of geometry undergoing intense investigation, it was discovered that if the Postulate of Archimedes is rejected the Fifth does not follow from the assumption that the sum of the angles of a triangle is equal to two right angles. In other words, there can exist a non-Archimedean geometry in which the sum of the angles of every triangle is equal to two right angles but where the Fifth does not hold.

This is a fitting climax to the chain of events started by the Parallel Postulate. Now, in every way, Euclid has been vindicated. Not only has the parallel postulate been shown to be independent of the others (we shall discuss its independence in the last chapter) but his form of it has been proved superior to the other form; those statements that were once regarded as equivalent to the Fifth have proved to be so only when propped up by the Postulate of Archimedes.

Hopefully, the reader is fully prepared to investigate a new and unfamiliar geometry. We might, however, make a final clarification. Strictly speaking, any geometry that is not "Euclidean" may be called "non-Euclidean." Thus the finite geometries introduced in the beginning of the text are "non-Euclidean" geometries. Any geometry denying *any* statement of Euclidean geometry may be called "non-Euclidean." It is customary, however, to restrict this title to those geometries that deny the parallel postulate.

EXERCISES 9.1

1. Check the references listed at the end of Part III to find examples of fallacious proofs of the Fifth that involve direction, infinite areas, or some assumption equivalent to the Fifth.

2. What is wrong with the following: Given a line m and a perpendicular to m from a point A not on it. Let it intersect m at a point B. Through A there exists exactly one line perpendicular to line AB. This line is parallel (non-intersecting) to m. Therefore, we have proved that through a point not on a given line there exists at most one line parallel to the given line. (Suggestion: compare this with exercise 1, Exercises 7.2.)

9.2 The Geometry of Bolyai-Lobachevsky

To study the geometry of Bolyai and Lobachevsky, often called *hyperbolic* geometry, all we will need is a new axiom, namely, any denial of some form of the Euclidean parallel postulate. The particular form we choose makes a difference in the development of the system.

Since the time of Proclus most geometers have taken "Playfair's Axiom" as their choice for a parallel postulate. If this is the statement chosen, the contradiction takes the following form:

> Through a point not on a given line, more than one line can be drawn not intersecting the given line.

Indeed, this is the form used by Gauss, Bolyai, and Lobachevsky in beginning their studies of the new geometry.

Unfortunately, in order to begin in this manner, we are forced to introduce considerations of continuity. For it is known that there are two types of lines through a point not on a given line: those that intersect the given line and those that do not. It can be shown that there are an infinite number of each type. In the usual development it is then stated that the *first* of the nonintersecting lines are to be called the *parallels* and that there are two such first, or limiting, lines. But to assume or to prove that such lines exist one must implicitly or explicitly introduce continuity considerations.

To avoid this we shall adopt an axiom first suggested by Hilbert. We have modified it to make it much stronger than need be and thereby to get to the heart of the subject matter more quickly.

***Axiom 17* (*Hyperbolic Parallel Postulate, or HPP*).** If P is any point not on a given line m, then there exist exactly two rays, say $\overrightarrow{PR}$ and $\overrightarrow{PS}$, which:

(a) *are not opposite rays*

(b) *do not intersect* m

(c) *are such that any ray, say* $\overrightarrow{PQ}$, *intersects* m *iff* $\overrightarrow{PQ}$ *is between* $\overrightarrow{PR}$ *and* $\overrightarrow{PS}$.

***Definition 50*.** Each of the two rays whose existence is given in HPP is called a *parallel to* m *through* P.

Having introduced Axiom 17 it is now possible to prove some of the characteristic theorems of this new geometry. It should be kept in mind that every theorem proved in Chapters 5, 6, 7, and 8 will hold in the new geometry.

Axiom 17 tells us that any ray between the parallel rays to a line m must intersect m. A modification of this is presented as the first theorem in the new geometry.

***Theorem 108*.** With any point P (not on a line m) as endpoint, if a ray lies between a ray parallel to m, and a ray intersecting m, then it intersects m.

Restatement: *If* $\overrightarrow{PR}$ *and* $\overrightarrow{PS}$ *are the parallels through* P *to a line* m, *and if* B *is any point on* m, *then any ray* $\overrightarrow{PQ}$ *falling between* $\overrightarrow{PR}$ *and* $\overrightarrow{PB}$ *intersects* m.

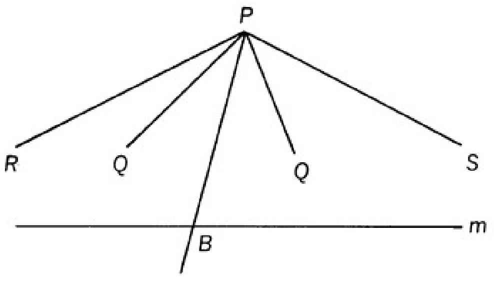

FIGURE 9.1

Proof: For if $\overrightarrow{PB}$ intersects m, then it follows from HPP that $\overrightarrow{PB}$ is between $\overrightarrow{PR}$ and $\overrightarrow{PS}$. If now $\overrightarrow{PQ}$ is a ray falling between $\overrightarrow{PR}$ and $\overrightarrow{PB}$, then it follows from Theorem 51a that $\overrightarrow{PQ}$ is between $\overrightarrow{PR}$ and $\overrightarrow{PS}$. Hence, from HPP it must intersect m.

Similarly, if ray $\overrightarrow{PQ}$ is between $\overrightarrow{PB}$ and $\overrightarrow{PS}$. ●

It will soon become evident that there are many nonintersecting lines that are not parallels. The next theorem says, in effect, that a parallel is the "first" of the nonintersecting lines.

Theorem 109. If $\overrightarrow{PQ}$ does not intersect a line m, but every ray between $\overrightarrow{PQ}$ and $\overrightarrow{PB}$, where B is a point on m, does intersect m, then $\overrightarrow{PQ}$ is parallel to line m.

Proof: Let m be a line, P a point not on it, $\overrightarrow{PQ}$ a ray that does not intersect m. By HPP there exist two rays $\overrightarrow{PR}$ and $\overrightarrow{PS}$ through P and parallel to m. Let B be any point on m such that it is false that B, P, and Q are collinear. Then Q falls on one side or the other of line PB. It follows from HPP that $\overrightarrow{PR}$ and $\overrightarrow{PS}$ must be on opposite sides of line PB; suppose Q falls on the R side of $\overrightarrow{PB}$.

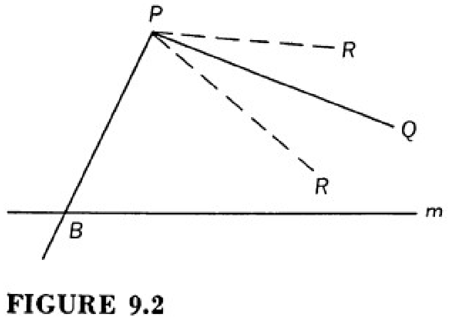

FIGURE 9.2

Now either $\overrightarrow{PR}$ is between $\overrightarrow{PB}$ and $\overrightarrow{PQ}$, $\overrightarrow{PQ}$ is between $\overrightarrow{PB}$ and $\overrightarrow{PR}$, or $\overrightarrow{PR} = \overrightarrow{PQ}$. (Why?)

By hypothesis, $\overrightarrow{PQ}$ does not intersect m; hence, it follows from Theorem 108 that it cannot lie between $\overrightarrow{PB}$ and $\overrightarrow{PR}$.

Also by hypothesis, every ray between $\overrightarrow{PQ}$ and $\overrightarrow{PB}$ intersects m; hence, by HPP, $\overrightarrow{PR}$ cannot lie between $\overrightarrow{PB}$ and $\overrightarrow{PQ}$.

It follows, therefore, that $\overrightarrow{PQ} = \overrightarrow{PR}$.

A similar argument holds if Q is on the S side of PB. ●

It follows as an immediate corollary of HPP that any perpendicular from a point P to a line m must lie between the parallels through P to line m. We can now show:

Theorem 110. If Q is the foot of the perpendicular from a given point P to a given line m, and $\overrightarrow{PR}$ and $\overrightarrow{PS}$ are the parallels through P to m, then $\angle RPQ \cong \angle SPQ$.

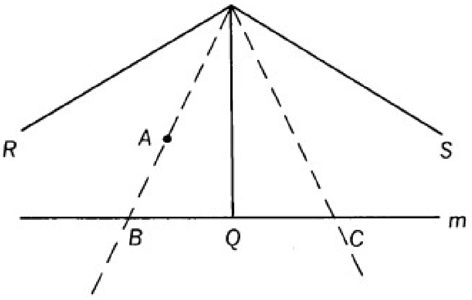

FIGURE 9.3

Proof: Suppose not. Then either $\angle SPQ < \angle RPQ$ or $\angle RPQ < \angle SPQ$. If $\angle SPQ < \angle RPQ$, then there exists a point A interior to $\angle RPQ$ such that $\angle APQ \cong \angle SPQ$ and, by Theorem 108, $\overrightarrow{PA}$ intersects m at a point B.

Now on m take C such that $B{*}Q{*}C$ and $B\text{–}Q \cong Q\text{–}C$ so that $\triangle BQP \cong \triangle CQP$ by SAS and therefore $\angle CPQ \cong \angle BPQ$. But $\angle APQ = \angle BPQ \cong \angle SPQ$. So by transitivity $\angle CPQ \cong \angle SPQ$, which implies that $\overrightarrow{PC}$ and $\overrightarrow{PS}$ are the same. (Why?)

But this is impossible because $\overrightarrow{PC}$ intersects m and $\overrightarrow{PS}$ does not. Hence, our assumption is false.

A similar argument eliminates $\angle RPQ < \angle SPQ$.

Hence, it follows that $\angle RPQ \cong \angle SPQ$. ●

The converse of Theorem 110 is also true.

Theorem 111. If $\overrightarrow{PR}$ and $\overrightarrow{PS}$ are the parallels through P to a given line m, then the bisector of $\angle RPS$ is perpendicular to line m.

Proof: Given that $\overrightarrow{PA}$ is the bisector of $\angle RPS$, it follows that $\overrightarrow{PA}$ must intersect m (why?) at some point Q, and that $\angle RPQ \cong \angle SPQ$. Suppose that $\overrightarrow{PQ}$ is not perpendicular to m.

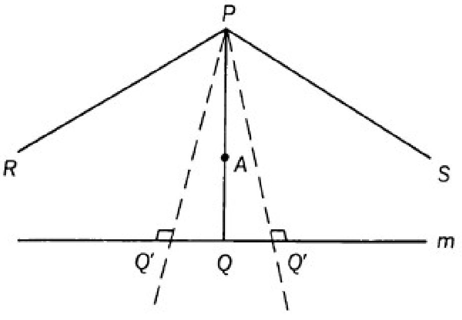

FIGURE 9.4

Then by Theorem 93 and Theorem 110 there exists a line PQ' through P and perpendicular to m, and such that $\angle RPQ' \cong \angle SPQ'$. Thus PQ' is also a bisector of $\angle RPS$, which contradicts Theorem 86.

It follows, therefore, that if $\overrightarrow{PA}$ is an angle bisector, it is also the perpendicular to m. ●

Although the diagrams have been drawn to indicate that the interior of the angle formed by parallel rays $\overrightarrow{PR}$ and $\overrightarrow{PS}$ contain m, this has not yet been established. Of course it would be quite embarrassing if false, for it would be almost painful to visualize condition (c) of Axiom 17 if the rays pointed upward. The next theorem assures us that the diagrams have not misled us. (See Figure 9.5.)

Theorem 112. If Q is the foot of the perpendicular from a point P to a given line m, and $\overrightarrow{PR}$ and $\overrightarrow{PS}$ are the parallels through P to m, then $\angle RPQ$ and $\angle SPQ$ are acute.

Proof: Suppose not. By Theorem 110 they must be congruent to each other. If they are both right angles, then $\overrightarrow{PR}$ and $\overrightarrow{PS}$ are opposite, contradicting HPP. If they are both obtuse, then $\overrightarrow{PE}$, the ray making a right angle with $\overrightarrow{PQ}$, would lie within $\angle SPQ$ or $\angle RPQ$ and hence by Theorem 108 would have to intersect m. By Theorem 78 this is impossible. The angles must, therefore, both be acute. ●

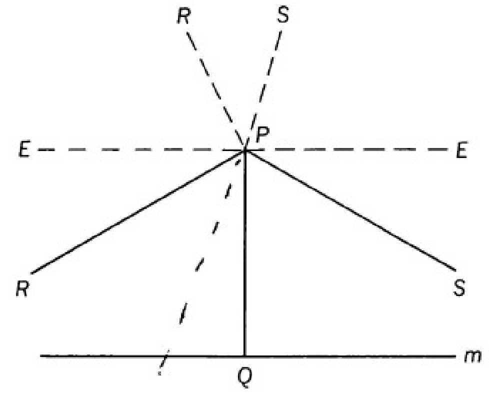

FIGURE 9.5

This theorem helps us to clarify a problem raised in Chapter 5. It was stated there that no angle analogue for Theorem 40 exists; that it cannot be proved that any line through an interior point of an angle, but not containing the vertex, cuts one of the sides. It should now be evident that to prove such a statement is, in effect, to prove the contradictory of Axiom 17. Therefore, it should be equally evident that to prove an analogue for Theorem 40 a Euclidean parallel postulate would be needed.

Using Theorem 112, it can now be proved that:

Theorem 113. If $\overrightarrow{PS}$ is a parallel through P to a line m, and N is a point of $\overrightarrow{PS}$ and X is any point such that $N*P*X$, then:

(a) It is false that $\overrightarrow{PX}$ is parallel to m
(b) It is false that $\overrightarrow{PX}$ intersects m
(c) Line PX ($= PS$) is nonintersecting with m.

Proof: Left as an exercise.

This theorem formally establishes some concepts that may have been intuitively evident; the theorem is not completely trivial, however. In this geometry that which is intuitively evident may be trivially true; on the other hand, it may be actually false. One must be careful of putting too much trust in one's intuition.

In keeping with this admonition, we shall carefully define some concepts whose meaning under other circumstances might very well have been left as part of the universal language.

Definition 51. Two noncollinear rays $\overrightarrow{PR}$ and $\overrightarrow{QS}$, where P and Q are distinct points, are said to be *in the same direction* iff they are on the same side of line PQ.

Definition 52. If $\overrightarrow{PS}$ is a parallel to m through P, and if $\overrightarrow{QX}$ is any ray on m, in the same direction as $\overrightarrow{PS}$, then $\overrightarrow{PS}$ is said to be *parallel to* $\overrightarrow{QX}$.

Let us carefully note the significance of Definition 52. Whereas HPP and Definition 50 tell us when a ray is parallel to a line, Definition 52 tells us when a ray is parallel to a ray. It tells us, for example, that while a ray may be parallel to a line, it is not parallel to *any* ray on that line. As illustrated in Figure 9.6, $\overrightarrow{PS}$ is parallel to m, and to $\overrightarrow{QX}$; but, $\overrightarrow{PS}$ is not parallel to $\overrightarrow{QY}$.

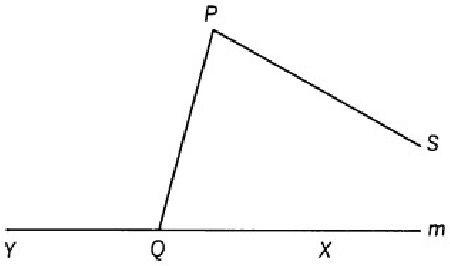

FIGURE 9.6

It is still not possible for us to speak of two *lines* being parallel to each other. In fact, because of the way in which parallels have been introduced, we have yet to answer the following question: if a ray is parallel through a point P to a line m, must it follow that at any other point on the line containing the ray there exists another ray on the line and parallel to m? The answer to this question is supplied by the next theorem.

Theorem 114. Let m be a line, P a point not on m, $\overrightarrow{PR}$ a parallel to m through P; if A is any point on the line containing $\overrightarrow{PR}$, that is, if *either* $P*A*R$, *or* $A*P*R$, then $\overrightarrow{AR}$ is parallel to m.

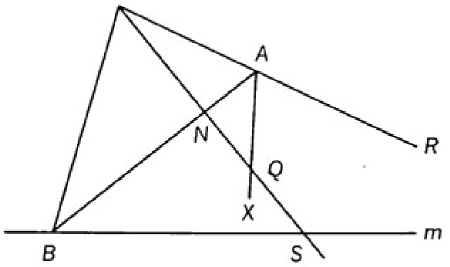

FIGURE 9.7

Proof:

Case 1. Let A be any point such that $P*A*R$; let B be any point on m; let X be any point interior to $\angle BAR$; and, let Q be any point on $\overrightarrow{AX}$ on the R side of m.

Now Q is interior to $\angle BAR$ (why?); it is also interior to $\angle BPR$. (Why?) It then follows, from Theorem 108 and Definition 18, that $\overrightarrow{PQ}$ intersects m in some point, say S. It is also the case that $\overrightarrow{PQ}$ intersects B–A in a point N. (Why?) So from Pasch's axiom applied to $\triangle BNS$ it follows that $\overrightarrow{AQ} = \overrightarrow{AX}$ intersects B–S. (Why?)

It was given that $\overrightarrow{AR}$ (as a subset of $\overrightarrow{PR}$) does not intersect line m. It has been shown that any ray $\overrightarrow{AX}$ interior to $\angle BAR$ intersects line m; therefore, by Theorem 109, $\overrightarrow{AR}$ is a parallel through P to line m.

(Because R is an arbitrary point, and because we wished to keep the statement of the theorem from becoming any more complicated, we have ignored the case $P*R*A$. But to include this case would lead to no difficulty because there always exists a point R' on $\overrightarrow{PR}$ such that $P*A*R'$ and the proof proceeds as in Case 1—except that the conclusion would then read that $\overrightarrow{AR'}$ is parallel to m.)

Case 2. Suppose now that A is a point such that $A*P*R$, B is any point on m, and X is any interior point of $\angle BAR$. Then there exists a point Q on $\overrightarrow{AX}$ such that Q is on the A side of m, and the A side of line BP. (Why?)

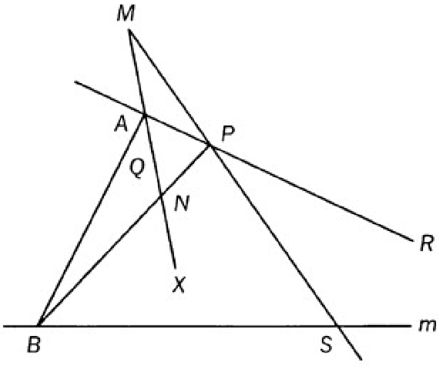

FIGURE 9.8

Let M be a point such that $Q*A*M$. Then $\overrightarrow{MP}$ intersects line m in some point, say S. (Why?)

In $\triangle BAP$, $\overrightarrow{AQ}$ intersects B–P in some point N; in $\triangle BPS$, $\overrightarrow{AQ} = \overrightarrow{AX}$ intersects B–S. (Why?)

From Theorem 113 it follows that $\overrightarrow{AR}$ (as a subset of line AR) does not intersect line m. It has been shown that any ray $\overrightarrow{AX}$ interior to $\angle BAR$ must intersect line m; therefore, by Theorem 109, $\overrightarrow{AR}$ is a parallel through A to line m. ●

Once Theorems 113 and 114 have been proved it is possible to offer the following definition with the assurance that it is meaningful.

Definition 53. Given $\overrightarrow{PS}$ is a parallel to m and given X, Y are points on line PS, then $\overrightarrow{XY}$ is said to be *parallel to m in the same sense* as $\overrightarrow{PS}$ iff $\overrightarrow{XY}$ is not opposite $\overrightarrow{XS}$.

Language Rule 17. *If $\overrightarrow{PS}$ is parallel to m, then it is possible to speak of* LINE *PS being* PARALLEL TO *m* IN THE SAME SENSE; *if this language is used, however, one must then distinguish between line sense PS and line sense SP.*

We conclude this section by proving two theorems that show how parallels of this geometry are quite like those of Euclidean geometry and by proving one theorem that shows how different they are.

Although Definition 52 tells us under what circumstances $\overrightarrow{PS}$ can be said to be parallel to $\overrightarrow{QX}$, it does not guarantee that $\overrightarrow{QX}$ will be parallel to $\overrightarrow{PS}$. Theorem 115 fills this gap.

Theorem 115. Two rays are always mutually parallel.

Restatement: *If $\overrightarrow{AR}$ is a parallel to $\overrightarrow{BS}$, it is to be shown that $\overrightarrow{BS}$ is a parallel to $\overrightarrow{AR}$.*

Proof: From A drop a perpendicular to $\overrightarrow{BS}$ intersecting at a point C. We know that $\overrightarrow{CS}$ would have to be nonintersecting with $\overrightarrow{AR}$; to show that it is parallel we shall use Theorem 109. Thus, it is to be shown that any ray interior to $\angle ACS$ intersects $\overrightarrow{AR}$.

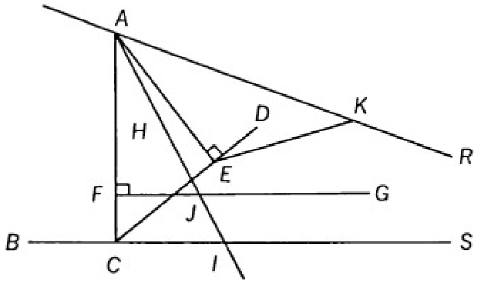

FIGURE 9.9

Let $\overrightarrow{CD}$ be a ray interior to $\angle ACS$, with D on the B side of $\overrightarrow{AR}$. Now from A let $A-E$ be perpendicular to line CD; E must fall on $\overrightarrow{CD}$ because $\angle DCS$, $\angle ACD$ are acute angles and, therefore, E cannot coincide with C or be on the opposite side of AC without contradicting the exterior angle theorem. Hence, E lies interior to $\angle ACS$.

Therefore, $\angle ACE$ is an acute angle, and $\angle AEC$ is a right angle of $\triangle AEC$, so by Theorem 82 $A-E < A-C$.

Now on $A-C$ there exists a point F such that $A-F \cong A-E$; let FG be perpendicular to $A-C$ and on the same side of $A-C$ as R and S.

E also lies interior to $\angle CAR$ (or else, $\overrightarrow{CE}$ meets $\overrightarrow{AR}$), so $\angle EAR < \angle FAR$. Let $\angle FAH \cong \angle EAR$. Then by Theorem 108 $\overrightarrow{AH}$ must meet $\overrightarrow{BS}$ in some point I. So in $\triangle ACI$, line FG must meet $A-I$ in some point J.

On $\overrightarrow{AR}$ let K be a point such that $A-K \cong A-J$. Then $\triangle AFJ \cong \triangle AEK$ by SAS. So corresponding angles $\angle AFJ$, $\angle AEK$ are congruent. But $\angle AFJ$ is a right angle; therefore, $\angle AEK$ is a right angle. By construction, $\angle AED$ is a right angle, therefore, $\overrightarrow{ED} = \overrightarrow{EK}$ and $\overrightarrow{CE}$ intersects $\overrightarrow{AR}$ at K.

It follows from Theorem 109 that $\overrightarrow{CS}$ is parallel to $\overrightarrow{AR}$, and hence by Theorem 114 that $\overrightarrow{BS}$ is parallel to $\overrightarrow{AR}$. ●

This theorem assures us that the parallel property for rays is a symmetric property. What about transitivity? Before one jumps to a conclusion on this question a look at HPP might be of value. Observe that while $\overrightarrow{PR}$ and $\overrightarrow{PS}$ are each parallel to m they are not parallel to each other. Thus if there is to be a transitivity property it must be restricted as in Theorem 116.

Theorem 116. If two rays are each parallel (in the same direction) to a third, then they are parallel to one another.

Proof: There are two cases to be considered: the case in which the two rays are on opposite sides of the third and the case in which they are on the same side of the third.

Case 1. Let $\overrightarrow{AR}$, $\overrightarrow{CT}$ both be parallel to $\overrightarrow{BS}$ on opposite sides of $\overrightarrow{BS}$. Then $A-C$ must meet $\overrightarrow{BS}$ in a point D. (See Figure 9.10.)

Let $\overrightarrow{AE}$ be any ray through A between $\overrightarrow{AC}$ and $\overrightarrow{AR}$. Then by Theorem 108 this line will meet $\overrightarrow{BS}$ in a point, say F. Join $C-F$. Since $\overrightarrow{CT}$ is parallel to $\overrightarrow{BS}$, then by Theorem 115, $\overrightarrow{BS}$ is parallel to $\overrightarrow{CT}$, and by Theorem 114, $\overrightarrow{FS}$ is parallel to $\overrightarrow{CT}$. $\overrightarrow{FE}$ is interior to $\angle CFS$ (why?), so $\overrightarrow{FE} = \overrightarrow{AE}$ will meet $\overrightarrow{CT}$.

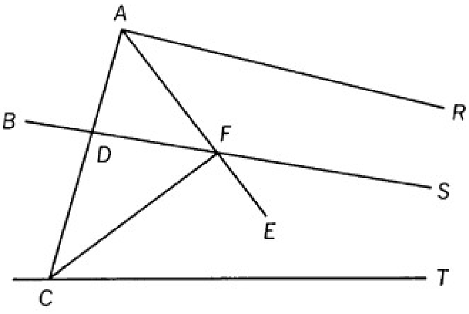

FIGURE 9.10

Since $\overrightarrow{AR}$ does not intersect $\overrightarrow{CT}$ (why?), but any line through A interior to $\angle CAR$ does, it follows from Theorem 109 that $\overrightarrow{AR}$ and $\overrightarrow{CT}$ are parallel.

Case 2. Suppose now that $\overrightarrow{AR}$ and $\overrightarrow{CT}$ are on the same side of $\overrightarrow{BS}$ and each is parallel to it. Then $\overrightarrow{AR}$ and $\overrightarrow{CT}$ cannot intersect. For if they did intersect, say in a point X, then through X there would exist two rays parallel (in the same direction) to a third. This is impossible. (Why?) Thus, they do not intersect, and either $\overrightarrow{AR}$ and $\overrightarrow{BS}$ are on opposite sides of $\overrightarrow{CT}$ or $\overrightarrow{CT}$ and $\overrightarrow{BS}$ are on opposite sides of $\overrightarrow{AR}$.

The proof is similar in either case; let us suppose the former, that is, that $\overrightarrow{AR}$ and $\overrightarrow{BS}$ are on opposite sides of $\overrightarrow{CT}$.

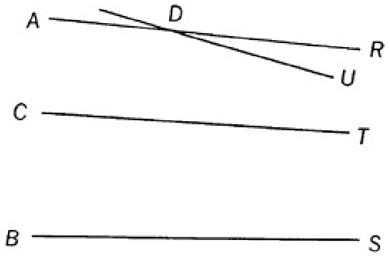

FIGURE 9.11

Suppose $\overrightarrow{AR}$ is not parallel to $\overrightarrow{CT}$. Then through any point D of $\overrightarrow{AR}$ there exists a ray $\overrightarrow{DU}$ parallel to $\overrightarrow{CT}$. It follows from Case 1 that $\overrightarrow{DU}$ is parallel to $\overrightarrow{BS}$. We already have that $\overrightarrow{AR}$ is parallel to $\overrightarrow{BS}$, so from Theorem 114 it follows that $\overrightarrow{DR}$ is parallel to $\overrightarrow{BS}$.

But from HPP, in the same direction, only one parallel to line BS can exist through a point D not on it. Therefore, $\overrightarrow{DU}$ must coincide with $\overrightarrow{DR}$ and thus $\overrightarrow{AR}$ is parallel to $\overrightarrow{CT}$. ●

We have now shown that if certain restrictions are made the parallels in this geometry have some of the properties of parallels in Euclidean geometry. It has also been shown that in this geometry there exists a distinction between "nonintersecting" lines and "parallel" lines; for, while all parallels are nonintersecting, it is not the case that all nonintersecting lines are parallels. Theorem 109, in effect, showed that parallels are the first of the nonintersecting lines. The next theorem (together with Theorem 77) establishes a more basic difference in their characteristics.

Theorem 117. Corresponding angles formed by two parallels and any transversal are not congruent.

Restatement: *Let $\overrightarrow{XR}$ and $\overrightarrow{YS}$ be parallel rays and line AB a transversal with A on $\overrightarrow{XR}$ and B on $\overrightarrow{YS}$; let C be a point such that $B*A*C$. It is to be proved that $\angle CAR$ is not congruent to $\angle CBS$.*

Proof: By Theorem 114 and Definition 52, $\overrightarrow{AR}$ and $\overrightarrow{BS}$ are parallel rays. Suppose now that $\angle CAR \cong \angle CBS$.

They cannot be right angles, for then $\angle BAR \cong \angle ABS$, contradicting Theorem 112.

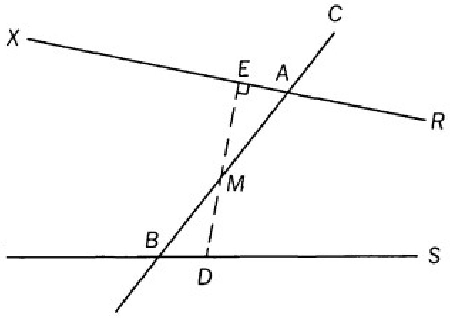

FIGURE 9.12

Let M be the midpoint of A–B and let the perpendicular from M to line AR intersect at a point E, which by the preceding statement is distinct from A. Whichever side of line AB point E falls on, take D on the opposite side of line AB such that B–$D \cong A$–E. It can then be shown that $\triangle MEA \cong \triangle MDB$ by SAS, and thus concluded that $\angle EMA \cong \angle DMB$.

Now since $\angle DMB$, $\angle DMC$ form a linear pair and $\angle EMA \cong \angle DMB$, then $\angle EMA$, and $\angle DMC$ are supplementary. (Why?) So by Theorem 62, Corollary 3, $\angle DMC$ and $\angle EMA$ form a linear pair; and, by Definitions 13 and 25, EMD is a (straight) line.

From the congruent triangles it follows that $\angle AEM \cong \angle MDB$; so, from Theorem 90 it follows that $\angle MDB$ is a right angle and by Definition 38 that $\angle MDS$ is a right angle.

It has now been shown that there exists a transversal ED of parallels $\overrightarrow{ER}$ and $\overrightarrow{DS}$ (using Theorem 114), such that $\angle RED$ and $\angle SDE$ are right angles. This contradicts Theorem 112.

Thus the assumption must be false and $\angle CAR$ and $\angle CBS$ are not congruent. ●

Corollary: *Alternate interior angles formed by two parallel lines and any transversal are not congruent.*

EXERCISES 9.2

1. Complete the proof of Theorems 110 and 111.
2. Prove Theorem 113.
3. Complete the proof of Theorem 114.
4. Complete the proof of Theorem 116.

9.3 A New Kind of Triangle

When two nonintersecting lines are met by a transversal the resulting figure is given no special name. When two parallel lines are met by a transversal, however, the resulting figure has some interesting properties and is sometimes called a *limit triangle* or an *asymptotic triangle.* Each of these names is descriptive in its own way but is also rather suggestive. So we shall merely call it a T-V-T (two-vertices-triangle). The following definition makes this more precise.

Definition 54. Given: any two parallel rays $\overrightarrow{XR}$ and $\overrightarrow{YS}$ met by a transversal m, at points A and B with $X{*}A{*}R$ and $Y{*}B{*}S$:

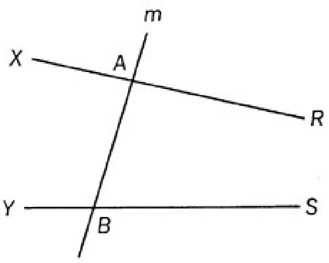

FIGURE 9.13

(a) The union of $\overrightarrow{AR}$, $\overrightarrow{BS}$, A, B and $A\text{–}B$ is called a *T-V-T*.

(b) The segment $A\text{–}B$ is called the *side* of the T-V-T.

(c) The angles $\angle BAR$ and $\angle ABS$ are called the *angles* of the T-V-T; also the *interior angles*.

(d) A, B are called the *vertices*.

(e) The angles $\angle XAB$, $\angle YBA$ and their respective vertical angles are called *exterior* angles of the T-V-T.

(f) An interior angle and an exterior angle of a T-V-T that are not adjacent are called *opposite interior* and *opposite exterior* angles relative to one another.

(g) The *interior* of a T-V-T is the intersection of:

- (i) The R side of line AB
- (ii) The B side of line AR
- (iii) The A side of line BS

(h) The *exterior* of a T-V-T is the set of all points which are neither on the T-V-T nor in its interior.

Notation Rule 19. *A* T-V-T *with rays* $\overrightarrow{AR}$ *and* $\overrightarrow{BS}$ *joined by segment* $A\text{–}B$ *will be denoted by:*

$$RABS$$

The figure has some of the properties of any ordinary triangle as shown by the following theorems.

Theorem 118. If a line passes through a point P interior to a T-V-T $RABS$, then:

(a) a line through P and either vertex intersects the other ray;

(b) a line through P and parallel to either ray intersects the side.

Proof:

Case a. Suppose that P lies interior to $RABS$. Then $\overrightarrow{AP}$ must intersect $\overrightarrow{BS}$ in some point C and, similarly, $\overrightarrow{BP}$ must intersect $\overrightarrow{AR}$; in each case this follows from Theorem 108.

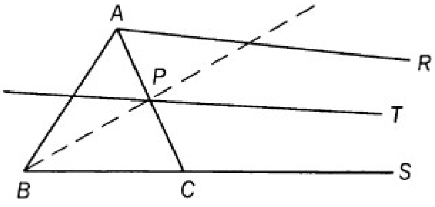

FIGURE 9.14

Case b. Suppose P lies interior to $RABS$ and $\overrightarrow{PT}$ is a ray parallel to either $\overrightarrow{AR}$ or $\overrightarrow{BS}$. Since $\overrightarrow{PT}$ intersects $A\text{–}C$ at P and is parallel to $\overrightarrow{BS}$, then by Pasch's axiom it must intersect $A\text{–}B$. ●

Theorem 119. If a line not through a vertex of a T-V-T:

(a) intersects either ray, then it intersects the other ray, or the side;

(b) intersects the side, and is not parallel to either ray, then it intersects one of the two rays.

Proof:

Case a. If line m intersects a ray, say $\overrightarrow{AR}$ of $RABS$ at C, and does not go through a vertex, then m is not the same line as line BC. Therefore, there exists a point X which is interior to $RABS$ and either on the R side of BC or on the A side of BC. (Why does such a point exist?)

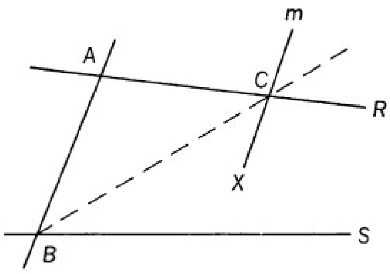

FIGURE 9.15

If it is on the R side, then it lies in $\angle BCR$ and so, by Theorems 108 and 114, it must intersect $\overrightarrow{BS}$. If it is on the A side of line BC, then in $\triangle ACB$, m intersects $A\text{–}B$, by Theorem 39.

Case b. If m intersects $A\text{–}B$ at some point, say C, and is not parallel to either ray, then let $\overrightarrow{CT}$ be such a parallel. Then $RACT$ and $TCBS$ are T-V-T triangles. Line m has a point X either in $\angle ACT$ or $\angle BCT$ (but not

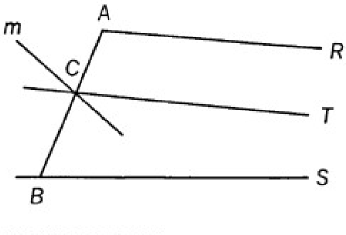

FIGURE 9.16

Case b. Suppose P lies interior to $RABS$ and $\overrightarrow{PT}$ is a ray parallel to either $\overrightarrow{AR}$ or $\overrightarrow{BS}$. Since $\overrightarrow{PT}$ intersects $A–C$ at P and is parallel to $\overrightarrow{BS}$, then by Pasch's axiom it must intersect $A–B$. ●

Theorem 119. If a line not through a vertex of a T-V-T:

(a) intersects either ray, then it intersects the other ray, or the side;

(b) intersects the side, and is not parallel to either ray, then it intersects one of the two rays.

Proof:

Case a. If line m intersects a ray, say $\overrightarrow{AR}$ of $RABS$ at C, and does not go through a vertex, then m is not the same line as line BC. Therefore, there exists a point X which is interior to $RABS$ and either on the R side of BC or on the A side of BC. (Why does such a point exist?)

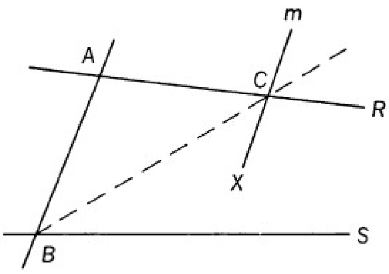

FIGURE 9.15

If it is on the R side, then it lies in $\angle BCR$ and so, by Theorems 108 and 114, it must intersect $\overrightarrow{BS}$. If it is on the A side of line BC, then in $\triangle ACB$, m intersects $A–B$, by Theorem 39.

Case b. If m intersects $A–B$ at some point, say C, and is not parallel to either ray, then let $\overrightarrow{CT}$ be such a parallel. Then $RACT$ and $TCBS$ are T-V-T triangles. Line m has a point X either in $\angle ACT$ or $\angle BCT$ (but not

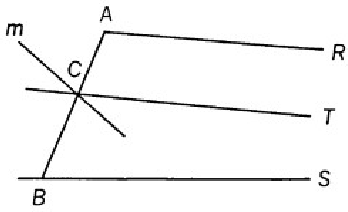

FIGURE 9.16

Theorem 121. If in two T-V-T triangles, an angle of one is congruent to an angle of the other, and the sides are congruent, then the remaining angles are congruent.

Restatement: *Given two* T-V-T *triangles* $RABS$ *and* $R'A'B'S'$, *in which* $\angle B \cong \angle B'$, *and* $A\text{–}B \cong A'\text{–}B'$, *prove that* $\angle A \cong \angle A'$.

Proof: Suppose that $\angle A' < \angle A$. Then there exists a point C such that $\angle BAC \cong \angle B'A'R'$, and C is interior to $\angle BAR$; hence, by Theorem 108, $\overrightarrow{AC}$ intersects $\overrightarrow{BS}$ at some point D.

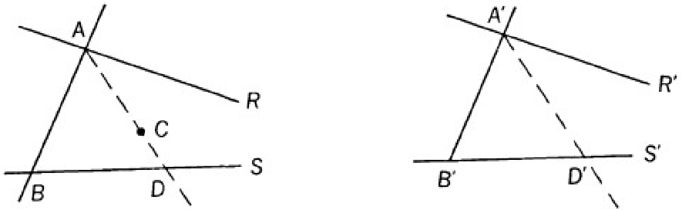

FIGURE 9.18

On $\overrightarrow{B'S'}$ there exists a point D' such that $B\text{–}D \cong B'\text{–}D'$. Then by SAS, $\triangle ABD \cong A'B'D'$, and $\angle BAD \cong \angle B'A'D'$. But $\angle BAD \cong \angle B'A'R'$; hence, by transitivity, $\angle B'A'D' \cong \angle B'A'R'$, which contradicts Axiom 14 (why?).

Similarly, if $\angle A < \angle A'$.

Therefore, by the Trichotomy theorem it follows that $\angle A \cong \angle A'$. ●

The next theorem is astonishingly different from anything in Euclidean geometry.

Theorem 122. If in two T-V-T triangles the angles of one are congruent to the angles of the other, then the sides are congruent.

Proof: Given that $\angle A \cong \angle A'$, $\angle B \cong \angle B'$, it is to be shown that $A\text{–}B \cong A'\text{–}B'$ in the two T-V-T triangles $RABS$ and $R'A'B'S'$.

Suppose $A'\text{–}B' < A\text{–}B$. Then there exists a point C on $A\text{–}B$ such that $B\text{–}C \cong B'\text{–}A'$, and through C there exists a parallel, say $\overrightarrow{CT}$, to $\overrightarrow{AR}$ and hence, by Theorem 116, to $\overrightarrow{BS}$.

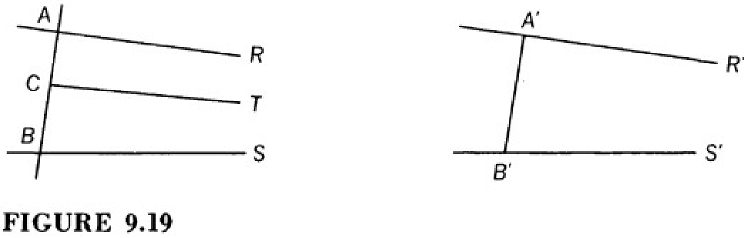

FIGURE 9.19

By Theorem 121 $\angle BCT \cong \angle A'$; but, since $\angle A' \cong \angle A$, it follows that $\angle A \cong \angle BCT$, which contradicts Theorem 117 (as well as Theorem 120). Hence, our assumption is false.

Similarly, if $A\text{–}B < A'\text{–}B'$.

So, by the Trichotomy for Segments theorem, $A\text{–}B \cong A'\text{–}B'$. ●

EXERCISES 9.3

1. Prove that in a T-V-T $RABS$, the bisector of each of the two angles intersects in a point interior to $RABS$.

2. A T-V-T is called *isosceles* if its two *angles* are congruent. Prove that if the sides of two isosceles *T-V-T*s are congruent, then the angles of each are congruent to the angles of the other.

3. Prove that the perpendicular bisector of the side $A\text{–}B$ of an isosceles T-V-T $RABS$ is parallel to the rays $\overrightarrow{AR}$ and $\overrightarrow{BS}$.

4. Prove that if $\overrightarrow{CT}$ is the perpendicular bisector of $A\text{–}B$ in isosceles T-V-T $RABS$, then the perpendiculars from any point on $\overrightarrow{CT}$ to $\overrightarrow{AR}$ and $\overrightarrow{BS}$ are congruent.

5. Prove that if the perpendicular bisector $\overrightarrow{CT}$ of $A\text{–}B$ in $RABS$ is parallel to $\overrightarrow{AR}$ and $\overrightarrow{BS}$, then $RABS$ is isosceles.

)-

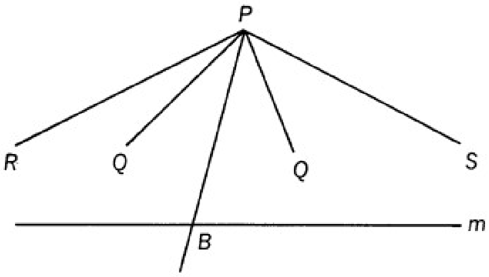

. -   <

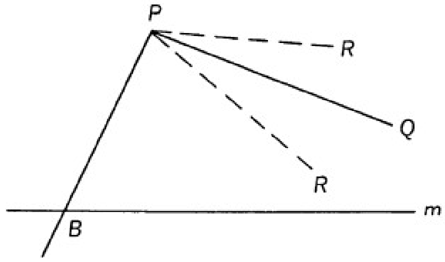

. -   <

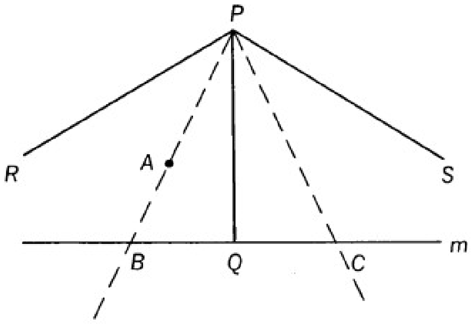

. -   <

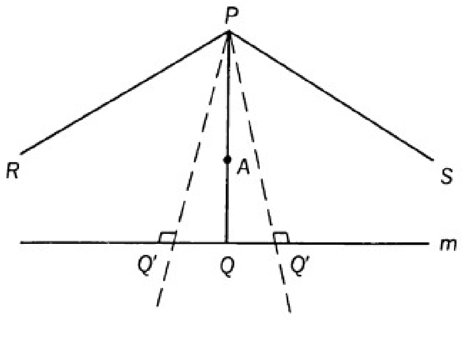

. -   < 6

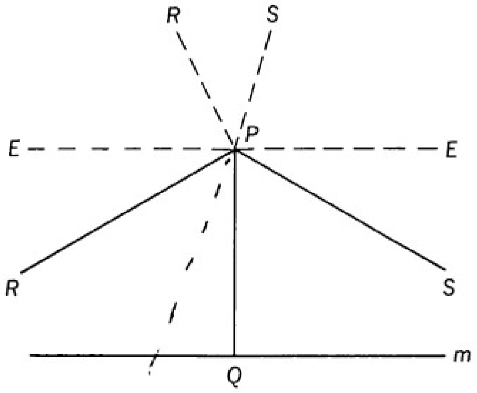

. -   < /

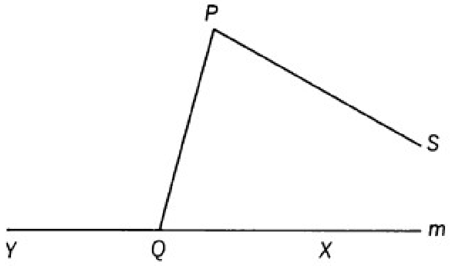

. -   < N

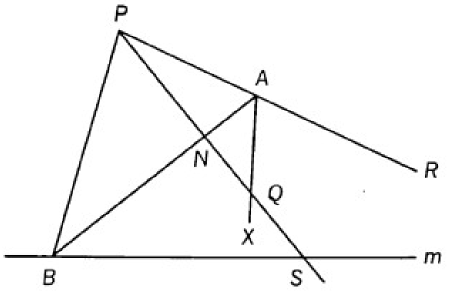

. -   < 7

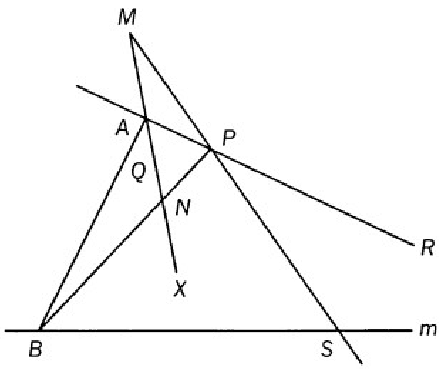
M
A
P
Q
N
R
X
m
B
S

. -   < T

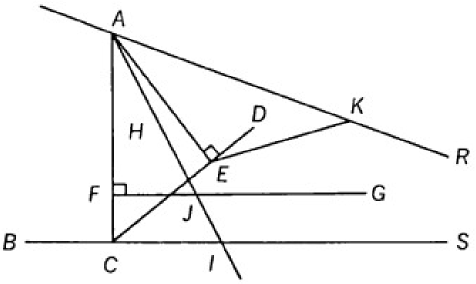
A
K
D
H
R
E
F
G
J
B
S
C
I

. -   < <

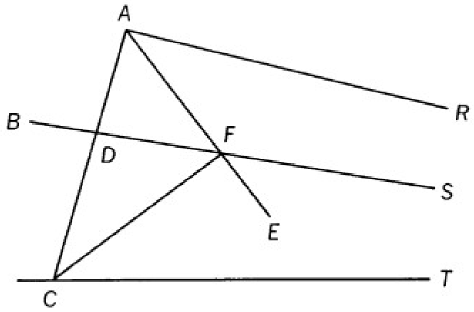
A
R
B
F
D
S
E
T
C

. -   <

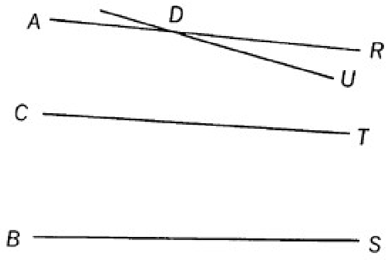
A
D
R
U
C
T
B
S

. -   <

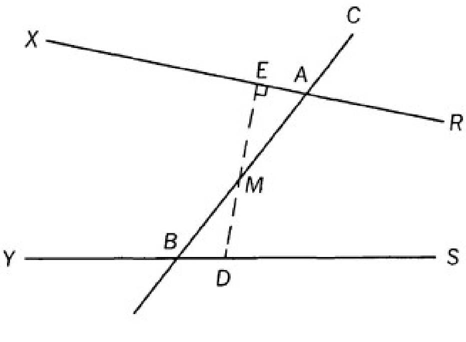
C
X
E
A
R
M
B
Y
S
D

. -   <

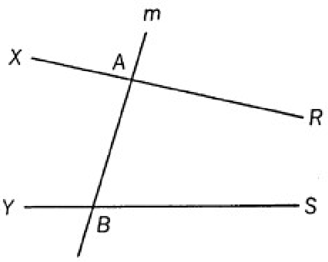

. -   <

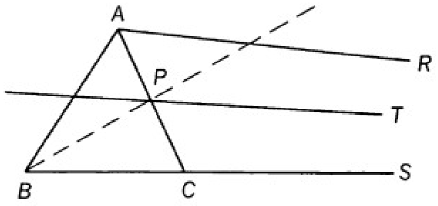

. -   < 6

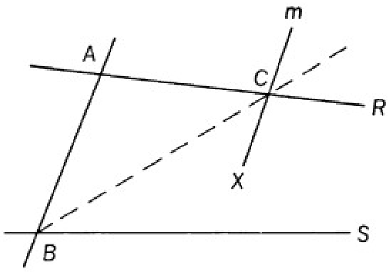

. -   < /

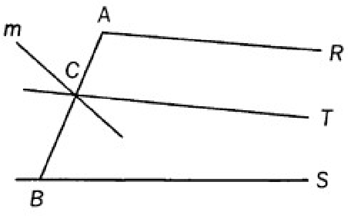

. -   < N

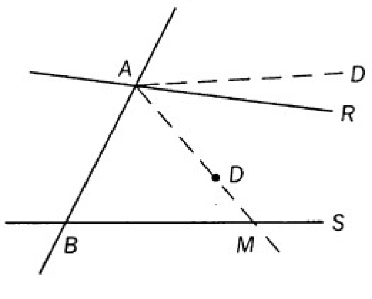

. -   < 7

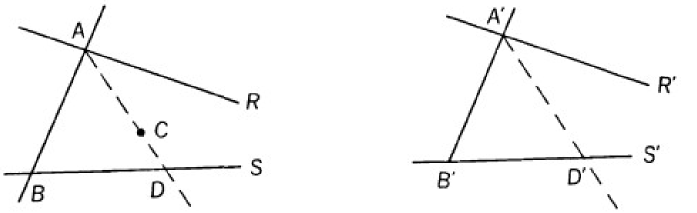
A
R
C
S
B
D
A′
R′
S′
B′
D′

. - < T

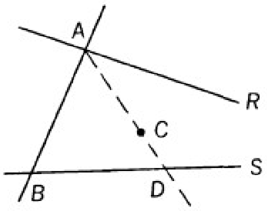

. - < T 2)3

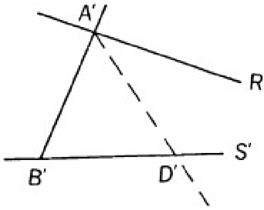

. - < T 293

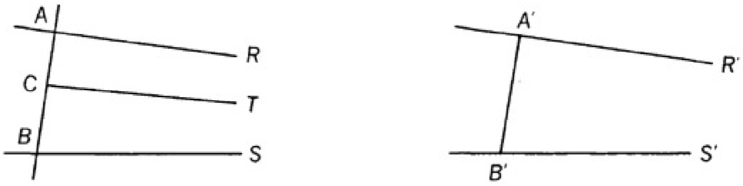
A
R
C
T
B
S
A′
R′
S′
B′

. - < <

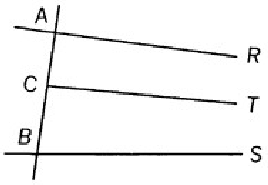

. - < < 2)3

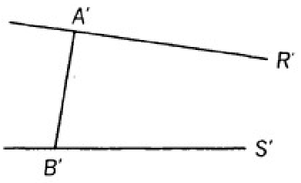

. - < < 293

10 New Shapes

Even when an author treats his subject as comprehensively as possible, he must use a certain amount of selectivity to keep the book from becoming unwieldy. Since it is not our aim to be comprehensive, a great deal of material must be left out or slighted. One should not assume that the deleted material is less important than that which is included.

If the subject matter is unadaptable or extraneous to the development of the text, there is no problem. However, many of the topics that might have been discussed in this chapter fall into the category of borderline cases both as to adaptability and relevance. In such a case the choice of materials to be included or deleted is arbitrary. It would have been possible, for example, to discuss the concept of *corresponding points,* which have as a locus the curve called a *horocycle* or the *equidistant* curve. We also might have examined *equivalent polygons* and their relevance to the *measurement of area.*

We have chosen to consider topics that are more obviously relevant to the development of the rest of the text and that easily generate interesting and instructive exercises. But the reader is encouraged to pursue additional topics on his own; many stimulating ones will be found in the references listed at the end of Part III.

10.1 Absolute Geometry

In Chapter 2 of this text we presented and discussed three axiom systems: 1, 2, and 3. It will be remembered that Axiom Sets 1 and 2 differed in only one axiom. In Axiom Set 1 there was a Euclidean parallel axiom; in Axiom Set 2 there was an axiom which said that no nonintersecting lines exist. It should be recognized that this situation is comparable to the relationship between Euclidean and hyperbolic geometry. For here, too, both systems have the same set of axioms except for one; but in this case the denial of the Euclidean parallel axiom takes the form of asserting the existence of more than one parallel rather than saying that there are none. The comparison carries further.

Axiom Set 3 was created by taking the set of all axioms common to the other two. Thus it was evident that all of the theorems derived from this set must of necessity hold in both of the other systems. In fact, Axiom System 3 was a *part* of each of the other systems; the same theorems would have resulted if we had worked in either of the other systems and merely refrained from ever using Axiom 5 in a proof. A completely analogous situation holds in the relationship between Euclidean geometry and hyperbolic geometry.

The founders of hyperbolic geometry were aware of this common core of subject matter. Bolyai, in particular, was very much interested in the theorems common to the new geometry and Euclidean geometry, and gave them a name. He called this system of axioms and theorems *absolute* geometry.

The reader should now be able to appreciate why, in developing the material of Chapters 5, 6, 7, and 8, no use was made of a parallel postulate. It was excluded to make the material available in Chapters 9 and 10.

As stated in Section 8.5, there was no need to stop at Theorem 107. There is literally no end to the number of theorems that could have been proved. Those theorems that would be needed were proved and those for which there was no immediate need were not proved. On the same basis we must now introduce several more theorems of absolute geometry.

Theorem 123. A point is interior to a convex quadrilateral iff it lies between: two points on opposite sides; two points on adjacent sides; a vertex and a point on a side not having that vertex as endpoint; opposite vertices.

Proof: Left as an exercise.

Theorem 124. The segment between points on opposite sides of a convex quadrilateral must intersect the segment between points on the other two opposite sides.

Proof: Left as an exercise.

Theorem 125. In a convex quadrilateral the (rays containing the) diagonals lie interior to the angles with which they have a common vertex.

Proof: Left as an exercise.

Theorem 126. (a) The line joining points on opposite sides of a convex quadrilateral "divides" it into two convex quadrilaterals.

(b) A diagonal of a convex quadrilateral "divides" it into two triangles.

Proof: Left as an exercise.

Theorem 127. If $ABCD$ is any convex quadrilateral with right angles at $\angle A$ and $\angle B$, and E is a point such that $A{*}E{*}B$, then any line perpendicular to line AB at E must intersect segment $C–D$.

Proof: Left as an exercise.

Theorem 128. If $ABCD$ is any convex quadrilateral with right angles at $\angle A$ and $\angle B$, and E is any point such that $C{*}E{*}D$, then the line which passes through E and which is perpendicular to line AB intersects segment $A–B$.

Proof: Left as an exercise.

EXERCISES 10.1

1. Prove Theorem 123.
2. Prove Theorem 124.
3. Prove Theorem 125.

*4. Prove Theorem 126a.

*5. Prove Theorem 126b.

6. Prove Theorem 127.
7. Prove Theorem 128.

10.2 New Quadrilaterals

In this section we shall introduce two figures briefly mentioned in the section on history. Both Saccheri and Lambert introduced quadrilaterals from which they attempted to prove the Parallel Postulate. It will be remembered that before claiming a tenuous victory, each had derived a great many theorems from his hypothesis of the acute angle.

Since we have already introduced the HPP we shall be approaching the material from a different perspective. This in no way detracts from the fascination of these figures. The Saccheri quadrilateral, in particular, we find to be one of the most interesting figures in geometry. We trust that before long the reader will agree.

Definition 55. A convex quadrilateral $ABCD$ in which $\angle A$ and $\angle B$ are right angles and $A\text{–}D \cong B\text{–}C$ is called a *Saccheri Quadrilateral.* $A\text{–}B$ is called the *base.* $D\text{–}C$ is called the *summit.* Angles $\angle D$ and $\angle C$ are called the *summit angles.*

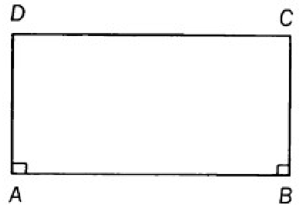

FIGURE 10.1

Notation Rule 20. *The quadrilateral defined in Definition 55 will be denoted by:*

S.Q. $ABCD$

where the first two letters of the quadrilateral named always represent the right angles.

According to the definition, any quadrilateral with two right angles and congruent *sides* qualifies as a Saccheri Quadrilateral. Therefore, as far as the definition is concerned, it is possible to refer to a square or a rectangle as a Saccheri Quadrilateral. This is true if we remain in absolute geometry, as we do in the next four theorems.

Theorem 129. If a line passes through the midpoints of the base and summit of a S.Q., then it is perpendicular to both of them.

Proof: Given a S.Q. $ABCD$ with points E, F, such that $A*E*B$, $D*F*C$, and $A\text{–}E \cong E\text{–}B$, $D\text{–}F \cong F\text{–}C$, it is to be shown that $EF \perp AB$ and $EF \perp DC$.

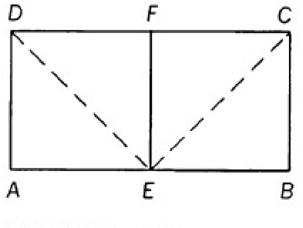

FIGURE 10.2

It follows immediately from the definition of S.Q. and from what is given that $\triangle DAE \cong \triangle CBE$. Hence, $\angle AED \cong \angle BEC$ and $D\text{–}E \cong C\text{–}E$.

$\triangle DFE \cong \triangle CFE$ by *SSS*. Hence, $\angle DFE \cong \angle CFE$ and, therefore, by definition, are right angles.

It must be the case that $\overrightarrow{ED}$ is between $\overrightarrow{EA}$ and $\overrightarrow{EF}$, and that $\overrightarrow{EC}$ is between $\overrightarrow{EB}$ and $\overrightarrow{EF}$. (Why?)

By corresponding parts of congruent triangles, $\angle DEF \cong \angle CEF$, and since we have already established that $\angle AED \cong \angle BEC$, it follows from Theorem 64 on addition of adjacent angles, that $\angle AEF \cong \angle BEF$. Since by definition they form a linear pair, they must be right angles.

By definition it now follows that $EF \perp AB$ and $EF \perp DC$. ●

Corollary: *The lines containing the base and summit of a S.Q. are nonintersecting and nonparallel lines.*
Proof: Follows from Theorems 129, 78, and 117.

Theorem 130. If a line is the perpendicular bisector of the base of a S. Q. it is also the perpendicular bisector of the summit.

Proof: Given a S.Q. *ABCD* with a point *E* such that $A*E*B$ and $A\text{–}E \cong E\text{–}B$, it is to be shown that there exists a point *F* such that $D*F*C$, such that the perpendicular bisector of $A\text{–}B$ passes through *F* and such that $EF \perp DC$ and $D\text{–}F \cong F\text{–}C$.

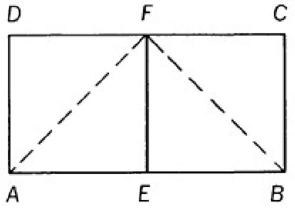

FIGURE 10.3

It follows from Theorem 127 that there exists a point *F* such that $D*F*C$ and such that the perpendicular bisector of $A\text{–}B$ passes through *F*.

$\triangle AEF \cong \triangle BEF$ by *SAS*. Hence $A\text{–}F \cong B\text{–}F$, $\angle EAF \cong \angle EBF$.

It must be the case that $\overrightarrow{AF}$ is between $\overrightarrow{AD}$ and $\overrightarrow{AE}$ and that $\overrightarrow{BF}$ is between $\overrightarrow{BE}$ and $\overrightarrow{BC}$. (Why?)

By definition, $\angle A$, $\angle B$ are right angles and thus $\angle A \cong \angle B$. It then follows (from what has just been established) from Theorem 65 on the subtraction of adjacent angles, that $\angle DAF \cong \angle CBF$. Since $D\text{–}A \cong C\text{–}B$, it follows by *SAS* that $\triangle DAF \cong \triangle CBF$ and thus that $D\text{–}F \cong F\text{–}C$.

But now, by Theorem 64 (Addition), it follows that $\angle DFE \cong \angle CFE$ and, since they form a linear pair, that they are right angles. So by definition $EF \perp DC$.

Therefore, by definition, EF is the perpendicular bisector of D–C. ●

Theorem 131. The perpendicular bisector of the summit of a S.Q. is also the perpendicular bisector of the base.

Proof: Left as an exercise. (Hint: prove by contradiction.)

Theorem 132. The summit angles of a S.Q. are congruent.

Proof: Left as an exercise.

If we ignore the Corollary to Theorem 129, which requires a hyperbolic theorem and definition in its proof, it will be seen that the proofs of these theorems use only the theorems from absolute geometry; each of them is true for squares and rectangles.

It is not often, however, that squares and rectangles are called Saccheri Quadrilaterals. This is because of the characteristic properties the latter have in hyperbolic geometry. This is brought out sharply in the next theorem.

In studying the proof of this fundamental theorem it should not pass unnoticed how much of the preceding development of this text is required for its solution.

Theorem 133. The summit angles of a S.Q. are acute.

Proof: In S.Q. $ABCD$, let E be a point such that $D*C*E$, and let F be a point such that $A*B*F$.

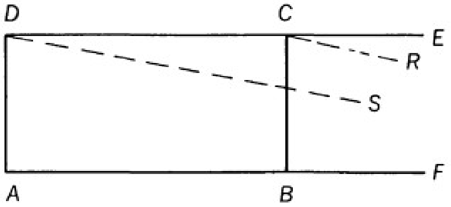

FIGURE 10.4

Let $\overrightarrow{DS}$ be parallel to $\overrightarrow{AB}$ and let $\overrightarrow{CR}$ be parallel to $\overrightarrow{BF}$, then $\overrightarrow{DS}$ and $\overrightarrow{CR}$ are both parallel to $\overrightarrow{AB}$ and to each other. (Why?)

By the Corollary to Theorem 129, the base and summit of a S.Q. are nonparallel and nonintersecting; hence, it follows that $\overrightarrow{DS}$ is between $\overrightarrow{DA}$ and $\overrightarrow{DC}$ and that $\overrightarrow{CR}$ is between $\overrightarrow{CB}$ and $\overrightarrow{CE}$. (Why?)

Now from Theorem 121 it follows that in the two T-V-T triangles *SDAF* and *RCBF*, $\angle SDA \cong \angle RCB$. Furthermore, from Theorem 120 it follows that in T-V-T *RCDS*, $\angle CDS < \angle ECR$. Hence, using Theorem 104 and Definitions 46 through 49, it follows that $\angle CDA < \angle ECB$.

According to Theorem 132 $\angle CDA \cong \angle DCB$; therefore, from Theorem 67 (Presubstitution) it follows that $\angle DCB < \angle ECB$.

Because $\angle DCB$, $\angle ECB$ form a linear pair they are supplementary. Because $\angle DCB < \angle ECB$ they are not right angles. Therefore, using Definition 38, Theorems 89b, 89a, and 88, it follows that $\angle DCB$ is acute.

Thus, it follows that $\angle CDA$ is acute. (Why?) ●

This theorem has a traumatic effect on one's intuition. It is one thing to agree that there can exist two parallels to a given line through a given point; one can almost do this abstractly. When this is accepted, one still has to cope with the idea that the parallel rays "point downward"; if they are to be nonintersecting, intuition would have them "pointing upward." This is strange, but one can adjust to it. Now, however, something new enters the discussion.

We have never examined the properties of the lines in this system. We know that "line" is an undefined term; by the axioms assumed, especially the *betweenness* axioms, it is plausible to regard them as "straight" lines. This has been done in all of the figures presented up to this point, and if we had not entered the realm of hyperbolic geometry no question would have been raised. But what now?

Observe that the sides of a Saccheri Quadrilateral are congruent; it has now been shown that the summit angles are acute. How then does one draw the summit? Notice (in Figure 10.5) that if one draws the summit straight, the angles do not look acute; if one draws the angles acute, the summit does not look straight.

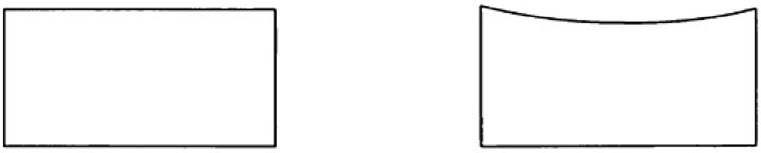

FIGURE 10.5

The next figure introduced will hardly improve matters. It is fortunate, indeed, that diagrams are not required for proofs. They might be quite misleading in hyperbolic geometry.

Definition 56. A quadrilateral containing three right angles is called a *Lambert Quadrilateral.*

Notation Rule 21. *The quadrilateral defined in Definition 56 will be denoted by:*

L.Q. $ABCD$

where the first three letters of the quadrilateral named will always represent the right angles.

As far as the definition is concerned it would be possible to refer to a square or a rectangle as a Lambert Quadrilateral. Once again it should be pointed out that in Euclidean geometry this is seldom done and, as the next theorem shows, in Hyperbolic geometry it cannot be done.

Theorem 134. In a L.Q. the fourth angle is acute.

Proof: Let $ABCD$ be a L.Q. with right angles at A, B, and C. Let E be a point such that $A*B*E$, and let F be a point such that $D*C*F$. Finally, let $B\text{–}A \cong B\text{–}E$ and $C\text{–}D \cong C\text{–}F$.

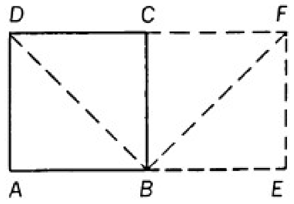

FIGURE 10.6

$\triangle DCB \cong \triangle FCB$ by SAS and thus their corresponding angles are congruent, as well as sides $D\text{–}B$ and $B\text{–}F$.

From Theorem 125 it follows that $\overrightarrow{BD}$ is between $\overrightarrow{BA}$ and $\overrightarrow{BC}$; it is also the case that $\overrightarrow{BF}$ is between $\overrightarrow{BC}$ and $\overrightarrow{BE}$.

It now follows that $\angle DBA \cong \angle FBE$ (why?); so $\triangle DBA \cong \triangle FBE$ by SAS, $\angle A \cong \angle E$, and $D\text{–}A \cong F\text{–}E$.

It was given that $\angle A$ is a right angle; it has been shown that $\angle A \cong \angle E$. Therefore, $\angle A$ and $\angle E$ are right angles.

By definition, it is now the case that quadrilateral $AEFD$ is a S.Q. and so by Theorem 133 $\angle ADC$ is acute. ●

It is now possible to prove the converse of Theorem 129; contrary to one's first impression, this is not a theorem from absolute geometry.

Theorem 135. If a line is perpendicular to both the base and summit of a S.Q., it joins the midpoints.

Proof: Left as an exercise (Hint: This statement is equivalent to "the line joining the midpoint of the base and summit of a S.Q. is the only line perpendicular to both of them" when used in conjunction with Theorems 129 and 134.)

The following theorem and its converse are very useful. It should be noted that they are about *any* convex quadrilaterals that have two right angles (and satisfy the other conditions).

Theorem 136. If, in a convex quadrilateral $ABCD$, the angles at A and B are right angles, and if $A\text{–}D < B\text{–}C$, then $\angle C < \angle D$.

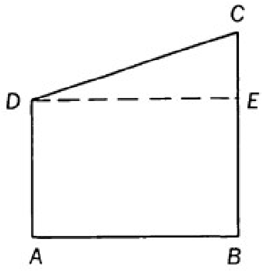

FIGURE 10.7

Proof: Since $A\text{–}D < B\text{–}C$, there exists a point E such that $B{*}E{*}C$ and $A\text{–}D \cong B\text{–}E$; but then quadrilateral $ABED$ is a S.Q., and from Theorem 132

$$\text{(a)}\quad \angle ADE \cong \angle BED$$

From the exterior angle theorem it follows that

$$\text{(b)}\quad \angle BCD < \angle BED$$

and from the definition of *less than*, that

$$\text{(c)}\quad \angle ADE < \angle ADC$$

Now, from Theorem 67 (Pre-Substitution) and lines (a) and (c), we have

$$\text{(d)}\quad \angle BED < \angle ADC$$

and from (b) and (d) together with Theorem 69 (Transitivity), it can be concluded that

$$\angle BCD < \angle ADC, \text{ that is, } \angle C < \angle D. \quad \bullet$$

Theorem 137. If, in a convex quadrilateral $ABCD$, the angles at A and at B are right angles, and if $\angle C < \angle D$, then $A\text{–}D < B\text{–}C$.

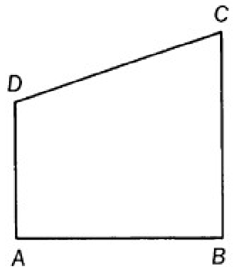

FIGURE 10.8

Proof: Suppose that $A\text{–}D \cong B\text{–}C$. Then, by definition, $ABCD$ is a *S.Q.*, and the assumption that $\angle C < \angle D$ contradicts Theorem 132.

Suppose $B\text{–}C < A\text{–}D$. Then, by Theorem 136, $\angle D < \angle C$, and this contradicts our hypothesis (by Theorem 66, the Trichotomy Theorem for angles).

Thus, by Theorem 57, the Trichotomy Theorem for segments, it is the case that $A\text{–}D < B\text{–}C$. ●

Theorem 138. In any convex quadrilateral $ABCD$, if $\angle A$ and $\angle B$ are right angles, and $\angle C \cong \angle D$, then the figure is a S.Q.

Proof: Left as an exercise.

EXERCISES 10.2

1. If a perpendicular is erected *at* any point *on* the side of a S.Q. it must intersect the opposite side. Answer true or false. If true, prove it; if false, give a counterexample.
2. If a perpendicular is erected *at* any point *on* the summit of a S.Q. it must intersect the base. Answer true or false. If true, prove it; if false, give a counterexample.
3. If a perpendicular is *dropped from* a point on a side of a S.Q. *to* the opposite side line, it must intersect the side. Answer true or false. If true, prove it; if false give, a counterexample.
4. Prove Theorem 131.
5. Prove Theorem 132.
6. Complete the proof of Theorem 133.
7. Prove Theorem 135.
8. Prove Theorem 138.
9. Prove that in a L. Q. the sides adjacent to the acute angle are greater than their respective opposite sides.
10. Prove that in a S.Q. the summit is greater than the base.

11. Prove that the diagonals of a S.Q. are congruent.

12. Prove that the diagonals of a S.Q. do not bisect each other.

13. Prove that the perpendicular bisector of the base of a S.Q. passes through the intersection of the diagonals.

14. Prove that if from the extremities of one side of a triangle, lines are drawn perpendicular to the line passing through the midpoints of the other two sides, a S.Q. is formed.

15. Prove that the perpendicular bisector of any side of a triangle is perpendicular to the line joining the midpoints of the other two sides.

*16. Prove that a line through the midpoint of one side of a triangle and perpendicular to the line that bisects a second side at right angles bisects the third side.

17. Prove that the line joining the midpoints of the equal sides of a S.Q. is perpendicular to the line joining the midpoints of the base and summit; prove that it bisects the diagonals.

18. Prove that the line joining the midpoints of the equal sides of a S.Q. does not pass through the intersection of the diagonals.

10.3 Congruent Quadrilaterals

Notation Rule 22: *By $ABCD \leftrightarrow EFGH$ we shall mean:*

$$\begin{array}{lll} A \leftrightarrow E & \angle A \leftrightarrow \angle E & A\text{–}B \leftrightarrow E\text{–}F \\ B \leftrightarrow F & \angle B \leftrightarrow \angle F & B\text{–}C \leftrightarrow F\text{–}G \\ C \leftrightarrow G & \angle C \leftrightarrow \angle G & C\text{–}D \leftrightarrow G\text{–}H \\ D \leftrightarrow H & \angle D \leftrightarrow \angle H & A\text{–}D \leftrightarrow E\text{–}H \end{array}$$

Definition 57. Given any two quadrilaterals $ABCD$, $EFGH$, if there exists at least one correspondence $ABCD \leftrightarrow WXYZ$, where W, X, Y, Z are E, F, G, H in some order such that

$$\begin{array}{ll} \angle A \cong \angle W & A\text{–}B \cong W\text{–}X \\ \angle B \cong \angle X & B\text{–}C \cong X\text{–}Y \\ \angle C \cong \angle Y & C\text{–}D \cong Y\text{–}Z \\ \angle D \cong \angle Z & A\text{–}D \cong W\text{–}Z \end{array}$$

then this correspondence is called a *congruent correspondence* or a *congruence.* The two quadrilaterals are called *congruent.*

Notation Rule 23: *If $ABCD \leftrightarrow EFGH$ is a congruent correspondence for quadrilaterals $ABCD$ and $EFGH$, this will be denoted by:*

$$ABCD \cong EFGH$$

Some of the techniques used to prove the following theorems are quite different from those used to prove most elementary congruence theorems in Euclidean geometry.

Theorem 139. If in L.Q.'s $ABCD$ and $EFGH$, $ABCD \leftrightarrow EFGH$, $\angle D \cong \angle H$ and, if any *one* of the following holds:

$$\text{(a)}\quad A{-}D \cong E{-}H \quad \text{or} \quad \text{(b)}\quad D{-}C \cong H{-}G$$

or

$$\text{(c)}\quad A{-}B \cong E{-}F \quad \text{or} \quad \text{(d)}\quad B{-}C \cong F{-}G$$

then, $ABCD \cong EFGH$.

Proof: The proofs of (a) and (b) are similar, as are the proofs of (c) and (d) but, as we shall see, the proofs of (a) and (c) are quite different.

Case (a). By hypothesis $A{-}D \cong E{-}H$. Suppose $D{-}C < H{-}G$. Then there exists a point I such that $H{*}I{*}G$ and $D{-}C \cong H{-}I$; $\triangle ADC \cong \triangle EHI$ by SAS.

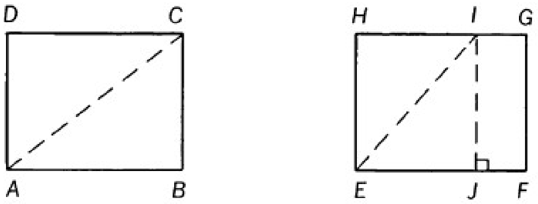

FIGURE 10.9

At I drop a perpendicular to line EF intersecting at a point J such that $E{*}J{*}F$ (follows from Theorem 128).

Now, $\overrightarrow{AC}$ is between $\overrightarrow{AD}$ and $\overrightarrow{AB}$; also $\overrightarrow{EI}$ is between $\overrightarrow{EH}$ and $\overrightarrow{EJ}$. (Why?) Therefore, $\triangle ABC \cong \triangle EJI$. (Why?)

Now $\angle DCA \cong \angle HIE$, $\angle ACB \cong \angle EIJ$; therefore, $\angle DCB \cong \angle HIJ$. Since $\angle C$ is a right angle, so is $\angle HIJ$, and hence so is $\angle JIG$. By construction, so is $\angle IJF$. Therefore, quadrilateral $IJFG$ has four right angles, contradicting Theorem 134.

A similar argument holds if $H{-}G < D{-}C$. Therefore, from the Trichotomy Theorem it follows that $D{-}C \cong H{-}G$.

It now follows that $\triangle ADC \cong \triangle EHG$; then, since $\overrightarrow{EG}$ is between $\overrightarrow{EH}$ and $\overrightarrow{EF}$, it follows by ASA (or SAA) that $\triangle ABC \cong \triangle EFG$.

Thus, $ABCD \cong EFGH$.

Case (c). By hypothesis, $A{-}B \cong E{-}F$. Let H' be a point such that $D{*}A{*}H'$ and $A{-}H' \cong E{-}H$. Then $\triangle BAH' \cong \triangle FEH$ (why?), and $H'{-}B \cong H{-}F$.

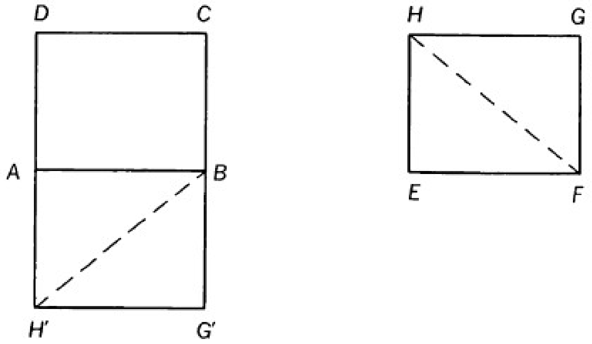

FIGURE 10.10

Let G' be a point such that $C*B*G'$ and $B\text{–}G' \cong F\text{–}G$. Join G', H'. Now $\overrightarrow{BH'}$ is between $\overrightarrow{BA}$ and $\overrightarrow{BG'}$ (why?) and $\overrightarrow{FH}$ is between $\overrightarrow{FE}$ and $\overrightarrow{FG}$. Thus, $\triangle BG'H' \cong \triangle FGH$. (Why?)

It now follows that $\angle AH'G' \cong \angle EHG$ and $ABG'H' \cong EFGH$. Therefore, by transitivity $\angle H' \cong \angle D$. Also $\angle G'$ and $\angle C$ are right angles. So by Theorem 138, $G'CDH'$ is a S.Q.

By definition of S.Q., $D\text{–}C \cong H'\text{–}G'$. By Theorem 135, $B\text{–}C \cong B\text{–}G'$, and $D\text{–}A \cong A\text{–}H'$. Hence, by transitivity and definition, $ABCD \cong EFGH$. ●

There are many ways to attempt the above proof, and because a Lambert Quadrilateral is actually being constructed, one must be wary not to assume what is being proved.

This theorem, once proved, provides an easy test for congruence of Lambert Quadrilaterals that can be used in the following proof about Saccheri Quadrilaterals:

Theorem 140. If in S.Q.'s $ABCD$ and $EFGH$, $ABCD \leftrightarrow EFGH$, $\angle D \cong \angle H$ and, if any *one* of the following holds:

(a) $C\text{–}D \cong G\text{–}H$ *or* (b) $A\text{–}B \cong E\text{–}F$ *or* (c) $B\text{–}C \cong F\text{–}G$

then $ABCD \cong EFGH$.

Proof:

Case (*a*). By hypothesis, $C\text{–}D \cong G\text{–}H$, $\angle D \cong \angle H$.

Suppose $H\text{–}E < D\text{–}A$. Then there exists a point I such that $D*I*A$ and $D\text{–}I \cong H\text{–}E$. Then $\triangle DIC \cong \triangle HEG$.

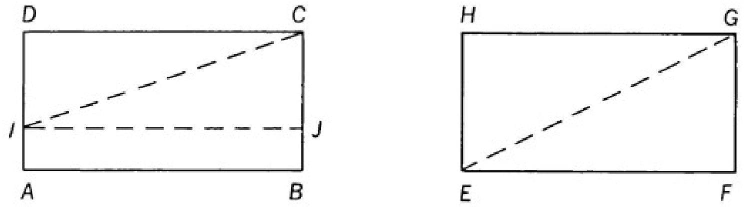

FIGURE 10.11

On $\overrightarrow{CB}$, let $C\text{–}J \cong G\text{–}F$. Then $\overrightarrow{CI}$ is between $\overrightarrow{CJ}$ and $\overrightarrow{CD}$. (Why?) Hence, $\triangle ICJ \cong \triangle EGF$. (Why?)

It now follows that $C{*}J{*}B$. (Why?)

Since it can be shown that $\angle DIJ$ and $\angle IJC$ are right angles, it follows that $\angle AIJ$ and $\angle BJI$ are right angles. This contradicts Theorem 134.

Similarly, if $D\text{–}A < H\text{–}E$. Therefore, $H\text{–}E \cong D\text{–}A$.

From this, it can be shown that $\triangle ADC \cong \triangle EHG$ and $\triangle ACB \cong \triangle EGF$, and thus finally that $ABCD \cong EFGH$.

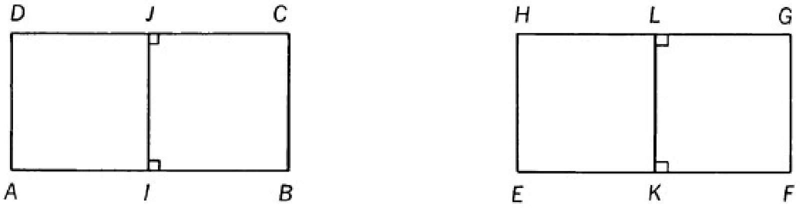

FIGURE 10.12

Case (*b*). This follows from constructing the perpendiculars at the midpoints of $A\text{–}B$ and $E\text{–}F$ and using Theorem 139, case (c).

Case (*c*). Follows as in Case (b) from Theorem 139, Case (b). ●

EXERCISES 10.3

1. Fill in the proofs of Theorems 139 and 140, Case (a).
2. Complete the proof of Theorem 140, Case (b).
3. Complete the proof of Theorem 140, Case (c).
4. If $ABCD$ and $EFGH$ are two L.Q.'s with acute angles at D and H, prove that:

(a) If $A\text{–}B \cong E\text{–}F$ and $B\text{–}C \cong F\text{–}G$, then $ABCD \cong EFGH$.

(b) If $D\text{–}A \cong H\text{–}E$ and $A\text{–}B \cong E\text{–}F$, then $ABCD \cong EFGH$.

(c) If $D\text{–}A \cong H\text{–}E$ and $B\text{–}C \cong F\text{–}G$, then $ABCD \cong EFGH$.

*(d) If $D\text{–}A \cong H\text{–}E$ and $D\text{–}C \cong H\text{–}G$, then $ABCD \cong EFGH$.

5. If $ABCD$ and $EFGH$ are S.Q.'s with acute angles at C, D, G, and H, prove that:

(a) If $A\text{–}B \cong E\text{–}F$ and if $B\text{–}C \cong F\text{–}G$, then $ABCD \cong EFGH$.
(b) If $A\text{–}B \cong E\text{–}F$ and if $C\text{–}D \cong G\text{–}H$, then $ABCD \cong EFGH$.
(c) If $B\text{–}C \cong F\text{–}G$ and if $C\text{–}D \cong G\text{–}H$, then $ABCD \cong EFGH$.

10.4 A Comparison of Hyperbolic and Euclidean Properties

In this section we will briefly survey some of the properties of hyperbolic geometry that contrast most strikingly with Euclidean properties. There will be no attempt to present this material formally. Some properties will be introduced by means of theorems, others will merely be mentioned or briefly discussed.

Parallels

In Definitions 51 and 52 we stated that in order for two rays to be parallel they must be *in the same direction*, and it was stated what that meant. Let us now refer to that direction as the *direction of parallelism*.

By now the reader must have anticipated the next theorem. It says, in effect, that parallel lines (rays) converge in the direction of parallelism.

Theorem 141. If $\overrightarrow{PR}$ and $\overrightarrow{QS}$ are any two parallel rays, and A, B, are points such that $P{*}A{*}B{*}R$, let C, D be the points where the perpendiculars from A and B to line QS intersect QS, then $B\text{–}D < A\text{–}C$.

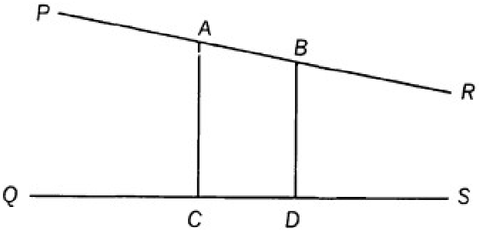

FIGURE 10.13

Proof: Because $\overrightarrow{PR}$ is parallel to $\overrightarrow{QS}$, so is $\overrightarrow{AR}$ and $\overrightarrow{BR}$. Then $\angle CAR$ and $\angle DBR$ are each acute angles.

But since $A{*}B{*}R$, it follows that $\angle DBA$, $\angle DBR$ form a linear pair and, hence, that $\angle DBA$ is obtuse.

Then in (convex?) quadrilateral $CDBA$ $\angle C$, $\angle D$ are right angles and $\angle CAB < \angle DBA$. So, by Theorem 137, $B\text{–}D < A\text{–}C$. ●

It should be immediately apparent from the proof of this theorem that if X is any point such that $A*B*X$, and Y is the foot of the perpendicular from X to line QS, then $X\text{–}Y < B\text{–}D$. This says that any segment "farther out" in the direction of parallelism is less than the preceding ones. But this does not imply that parallels become less than any length no matter how small; that is, that the distance between parallels converges to zero; or, that the lines are asymptotic. The next theorem, in effect, says just that.

Theorem 142. Let $X\text{–}Y$ be any segment whatsoever; then there exists a segment $W\text{–}Z$ such that W is on $\overrightarrow{PR}$, a parallel to QS, and $W\text{–}Z$ is a perpendicular to QS, and $W\text{–}Z < X\text{–}Y$.

Proof: Let $\overrightarrow{PR}$ and $\overrightarrow{QS}$ be parallel and PA perpendicular to QS. Let $X\text{–}Y$ be any segment whatsoever; then, if $P\text{–}A < X\text{–}Y$ we are through. If $P\text{–}A \cong X\text{–}Y$ then this theorem follows from Theorem 141. Suppose then, that $X\text{–}Y < P\text{–}A$.

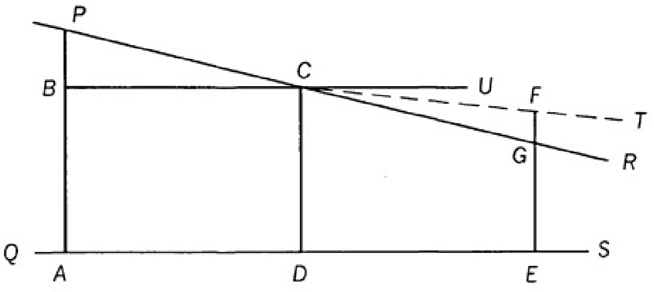

FIGURE 10.14

Let B be a point such that $A*B*P$ and $A\text{–}B \cong X\text{–}Y$. Through B let $\overrightarrow{BU}$ be perpendicular to PA and in the direction $\overrightarrow{QS}$. Then, by Theorem 119, $\overrightarrow{BU}$ intersects $\overrightarrow{PR}$ at some point, say C. Let CD be the perpendicular from C to QS.

Then $\angle RCD < \angle BCD$. (Why?) So let $\overrightarrow{CT}$ be such that $\angle BCD \cong \angle TCD$; then $\angle RCD < \angle TCD$.

Let E be a point such that $A*D*E$ and $A\text{–}D \cong D\text{–}E$, and let F be a point on $\overrightarrow{CT}$ such that $B\text{–}C \cong C\text{–}F$.

Now it can be shown that $F\text{–}E \cong A\text{–}B$, and FE is perpendicular to QS

(left as an exercise). Furthermore, $\overrightarrow{PC}$ must intersect $F\text{–}E$ at some point, say G. (Why?)

Thus $E\text{–}G < E\text{–}F$; $E\text{–}F \cong A\text{–}B$; $A\text{–}B \cong X\text{–}Y$. So $E\text{–}G < X\text{–}Y$, where G is a point on $\overrightarrow{PR}$ and EG is perpendicular to $\overrightarrow{QS}$. Hence $E\text{–}G$ is the segment $W\text{–}Z$. This completes the proof. ●

It should be apparent also that parallel rays diverge in the opposite direction and become larger than any given segment. This could be proved now, but we leave this as an exercise and turn to other interesting implications of the theory of parallels.

While discussing the historical development of non-Euclidean geometries, we stated that in the new geometry there exists an *absolute unit* of length. A rigorous discussion of this concept lies beyond the synthetic approach, but the following intuitive explanation may help to clarify it.

Given points A, Y, P, and X, such that $A*Y*P*X$; and given parallels $\overrightarrow{AR}$, $\overrightarrow{YS}$, $\overrightarrow{PT}$, and $\overrightarrow{XU}$, it follows immediately from Theorem 120 that $\angle AXU < \angle APT < \angle AYS$. (See Figure 10.15.)

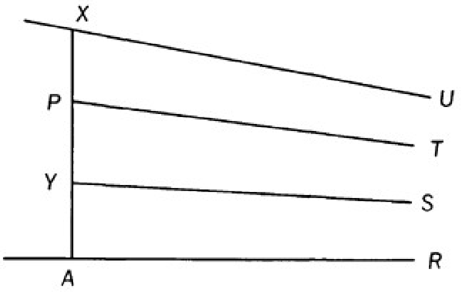

FIGURE 10.15

Furthermore, given two *T-V-T*'s $RABS$ and $R'CDS'$, such that $\angle B$ and $\angle D$ are right angles, then it follows immediately from Theorem 121 that if $A\text{–}B \cong C\text{–}D$, then $\angle A \cong \angle C$; and from Theorem 122 it follows that if $\angle A \cong \angle C$, then $A\text{–}B \cong C\text{–}D$. (See Figure 10.16.)

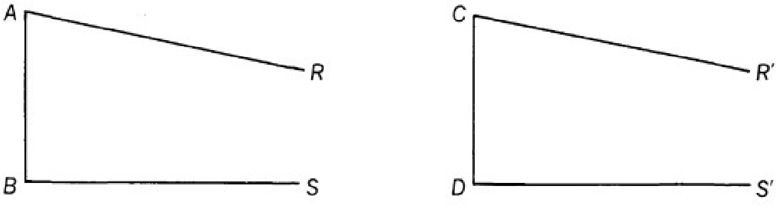

FIGURE 10.16

The following definition will make it easier to discuss this:

Definition 58. Given $\overrightarrow{PR}$ a parallel to m through P, and Q the intersection point of the perpendicular from P to m, then $\angle QPR$ is called *the angle of parallelism for P–Q.*

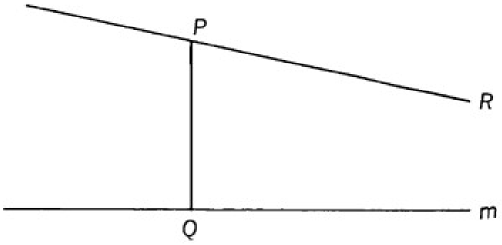

FIGURE 10.17

We can now say that the angles of parallelism corresponding to equal distances are equal, and conversely. Furthermore, as the distances become greater, the angle of parallelism becomes smaller, and as the distances become smaller, the angles of parallelism become larger. There is a functional relationship between the measure of the angle of parallelism $\angle QPR$ and the length of the segment P–Q; furthermore, to every acute angle there corresponds a distance for which the angle is the angle of parallelism.

In both Euclidean and hyperbolic geometries there is an absolute standard to which all angles can be compared, the right angle. No one needs to keep a right angle hidden in a vault at the Bureau of Standards to preserve it, for it can always be reconstructed. Thus, there exists a natural *unit* of measure (some given fractional part of a right angle) that could be reconstructed if it were ever lost. This is what is meant by saying that angles have an *absolute* unit of measure in both geometries.

In Euclidean geometry there is no absolute standard for units of length; they are quite arbitrary. However, once it has been determined in hyperbolic geometry that to every angle there corresponds a definite segment, namely the segment associated with that angle when it is regarded as an angle of parallelism, an interesting new property comes into being. From the unit of angular measure one can determine a corresponding unit of linear measure, a unit having a structural connection with the geometry and thus one that is an absolute unit of length.

Nonintersecting Nonparallel Lines

It has already been shown that if two lines have a common perpendicular they are nonintersecting and nonparallel. A proof, originated by

Hilbert, establishes the converse and occurs in most texts on hyperbolic geometry. Because the proof is long and complicated and does not introduce any new techniques that make its reproduction here worthwhile, we shall omit it and merely state the theorem.

Theorem 143. If two straight lines neither intersect nor are parallel, they must have a common perpendicular.

This is an *existence* theorem (in fact, Hilbert's proof actually shows how to construct such a line if one knows how to construct parallels). The *uniqueness* of the common perpendicular was established in effect in the proof of Theorem 134.

A property more interesting than the existence of a common perpendicular to two nonintersecting lines is:

Theorem 144. Let the common perpendicular to two nonintersecting lines m, n, meet at P, Q. Let A, B be points such that $P{*}A{*}B$ and let perpendiculars through A, B to n meet at C, D respectively, then

$$P\text{–}Q < A\text{–}C,\ A\text{–}C < B\text{–}D$$

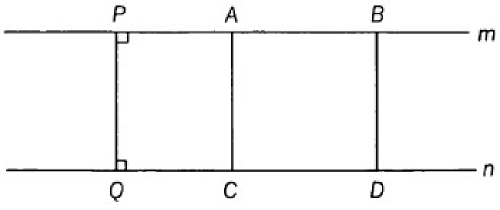

FIGURE 10.18

Proof: $PQCA$ is a L.Q. Hence, $\angle CAP$ is acute and, therefore, $P\text{–}Q < A\text{–}C$ by Theorem 137.

Because $\angle CAP$ is acute, $\angle CAB$ is obtuse. Since quadrilateral $PQDB$ is a L.Q., $\angle DBP$ is acute. But $\angle DBP = \angle DBA$. Hence, in (convex?) quadrilateral $CDBA$, $\angle DBA < \angle CAB$.

Therefore, $A\text{–}C < B\text{–}D$ by Theorem 137.

A similar argument holds if the points are on the opposite sides of line PQ. ●

This theorem, in effect, shows that nonintersecting nonparallel lines diverge as the distance from the common parallel increases. It can also be shown that the distance apart becomes greater than any arbitrarily chosen segment. Thus, it would seem that nonintersecting lines are as shown in Figure 10.19.

Hilbert, establishes the converse and occurs in most texts on hyperbolic geometry. Because the proof is long and complicated and does not introduce any new techniques that make its reproduction here worthwhile, we shall omit it and merely state the theorem.

Theorem 143. If two straight lines neither intersect nor are parallel, they must have a common perpendicular.

This is an *existence* theorem (in fact, Hilbert's proof actually shows how to construct such a line if one knows how to construct parallels). The *uniqueness* of the common perpendicular was established in effect in the proof of Theorem 134.

A property more interesting than the existence of a common perpendicular to two nonintersecting lines is:

Theorem 144. Let the common perpendicular to two nonintersecting lines m, n, meet at P, Q. Let A, B be points such that $P{*}A{*}B$ and let perpendiculars through A, B to n meet at C, D respectively, then

$$P\text{–}Q < A\text{–}C,\ A\text{–}C < B\text{–}D$$

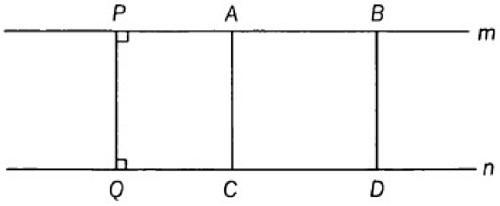

FIGURE 10.18

Proof: $PQCA$ is a L.Q. Hence, $\angle CAP$ is acute and, therefore, $P\text{–}Q < A\text{–}C$ by Theorem 137.

Because $\angle CAP$ is acute, $\angle CAB$ is obtuse. Since quadrilateral $PQDB$ is a L.Q., $\angle DBP$ is acute. But $\angle DBP = \angle DBA$. Hence, in (convex?) quadrilateral $CDBA$, $\angle DBA < \angle CAB$.

Therefore, $A\text{–}C < B\text{–}D$ by Theorem 137.

A similar argument holds if the points are on the opposite sides of line PQ. ●

This theorem, in effect, shows that nonintersecting nonparallel lines diverge as the distance from the common parallel increases. It can also be shown that the distance apart becomes greater than any arbitrarily chosen segment. Thus, it would seem that nonintersecting lines are as shown in Figure 10.19.

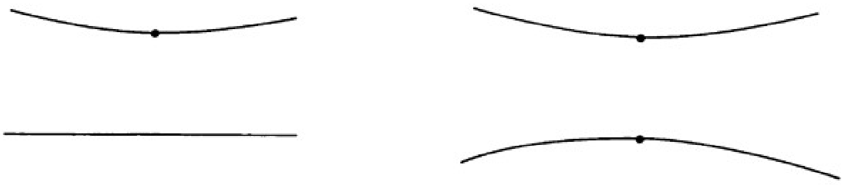

FIGURE 10.19

Triangles

As previously stated, there are many interesting concepts in hyperbolic geometry that one might pursue in discussing triangles. We have chosen to omit discussions of area and defects of triangles, but several properties occur in this geometry which, although not well suited to the synthetic approach, should not be ignored. Accordingly, the remainder of the theorems will illustrate these properties.

Definition 59. Any *three* angles $\angle A$, $\angle B$, and $\angle C$ are said to be *taken together less than two right angles*, iff:

(a) $m(\angle X) + m(\angle Y) = m(\angle Z)$ exists, where $\angle X$, $\angle Y$ are any two of the three angles; and

(b) $\angle V$ and $\angle W$ are two angles taken together less than two right angles, where $\angle V$ is some angle of the set $m(\angle Z)$ and $\angle W$ is the remaining one of the three angles.

Theorem 145. The three angles of every right triangle are taken together less than two right angles.

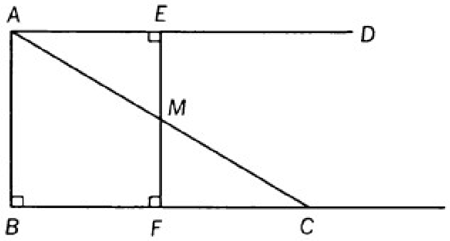

FIGURE 10.20

Proof: Let ABC be any triangle with a right angle at B. At A, on the opposite side of line AC from B, let $\angle CAD \cong \angle BCA$. Let M be the midpoint of A–C, and lines ME and MF the perpendiculars from M to AD and BC, respectively.

Then $\triangle FCM \cong \triangle EAM$ by SAA and, therefore, $\angle AME \cong \angle CMF$. But $\angle CMF$, $\angle FMA$ form a linear pair; therefore, $\angle AME$, $\angle FMA$ are supplementary. Since they are adjacent, by Theorem 62, Corollary 3, they form a linear pair.

Thus, $BFEA$ is a quadrilateral with right angles at B, F, and E. Hence, its fourth angle is acute. But $m(\angle BAE) = m(\angle BAC) + m(\angle BCA)$.

Therefore, $\angle B$, $\angle A'$, and $\angle C$ of triangle ABC are taken together less than two right angles. ●

Theorem 146. The three angles of *any* triangle are taken together less than two right angles.

Proof: Let ABC be any triangle. By Theorem 91, at least two of its angles are acute, say A and B.

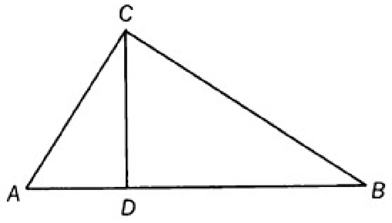

FIGURE 10.21

Then a line through C perpendicular to AB must intersect A–B at some point D. (Why?)

Therefore, C–D divides triangle ABC into two right triangles ADC and BDC.

From the preceding theorem, $m(\angle DAC) + m(\angle ACD) = m(\angle M)$ and $m(\angle M) < m(\angle X)$ where X is any right angle.

And $m(\angle DBC) + m(\angle DCB) = m(\angle N)$, and $m(\angle N) < m(\angle X)$.

Therefore, $\angle M$ and $\angle N$ are taken together less than two right angles. ●

These results, strange as they are, could have been anticipated from the discussion of the history of attempts to prove the Fifth. For if the statement:

The sum of the angles of any triangle equals two right angles.

is equivalent to the Parallel Postulate, then when one statement is denied so is the other. Of course, whether the sum was to be less than or greater than two right angles still had to be determined.

The last of the properties to be discussed in this section may also have been anticipated—a clue was given by Theorem 122, where the congruence of the angles of two T-V-T triangles implied the congruence of the sides. The property is introduced below as statement (S). Unfortunately, it cannot be introduced as a theorem because the proof does not lend itself to the synthetic method. In such a proof we would be required to say that:

The sum of four angles (of a quadrilateral) is less than 360 degrees (four right angles).

There is no easy way, in the synthetic approach, to express a statement equivalent to this. The problem is not only to find a definition but to find a definition that is usable.

If such a definition were available, then the proof of the following statement would be relatively easy.

(S) If two triangles ABC and DEF are such that $ABC \leftrightarrow DEF$, $\angle A \cong \angle D$, $\angle B \cong \angle E$, $\angle C \cong \angle F$, then $\triangle ABC \cong \triangle DEF$.

This statement introduces one of the most astounding properties of hyperbolic geometry. In effect, it says that there is no such thing as two figures being similar but not congruent. Thus, for a photograph to resemble its subject it would have to be a lifesize photograph: moreover, all the fields that depend upon similarity and proportion would be drastically altered. Consider what would happen to painting, designing, drafting, and architecture. Consider what would happen to the world around us. It would be a very strange new world.

Yet such a world is a possible one. The last chapter will explain how it might even be that this geometry most closely describes the space world around us.

EXERCISES 10.4

1. In the proof of Theorem 142:

   (a) prove that $\overrightarrow{CT}$ is between $\overrightarrow{CU}$ and $\overrightarrow{CR}$;
   (b) prove that $E\text{-}F \cong B\text{-}A$, and $\angle DEF$ is a right angle.

2. Complete the proof of Theorem 146.

3. Prove that parallel rays diverge in the direction opposite to the direction of parallelism.

4. Try to define what it means to say that the sum of four angles is less than (or equal to) four right angles.

5. Assuming one has the definition of exercise 4, prove statement (S) of the text.

Which of the following statements are true in *both* Euclidean and hyperbolic geometry. Explain.

6. There exist equilateral quadrilaterals.

7. There exist parallelograms.

8. Two lines are parallel if they have no points in common.

9. If two lines are perpendicular to the same line, then they are parallel.

10. The perpendicular to a given line from a point not on it is unique.

11. The exterior angle of a triangle is greater than the opposite interior angle.

12. Given a line and a point not on the line, there exists at least one line which passes through the given point and is parallel to the given line.

13. If two sides of a triangle are not congruent, then the angles opposite them are not congruent and the greater angle is opposite the longer side.

14. The opposite sides of a parallelogram are equal.

15. All straight angles are equal to two right angles.

16. Given two lines and a transversal, if a pair of corresponding angles are congruent, then a pair of alternate interior angles are congruent.

17. If a S.Q. $ABCD$ has right angles at A and B, then the angles at C and D are congruent to each other.

18. All squares have four equal angles.

19. The Pythagorean Theorem holds.

5-   2 N3

11 A Model World

From the list of theorems in the last two chapters, one may choose numerous denials of the Euclidean parallel axiom and its equivalent forms:

1. Through a point outside a given line, more than one parallel to the given line exists.
2. If two lines are cut by a transversal such that the interior angles on the same side of the transversal are, taken together, less than two right angles, the lines do not necessarily meet on that side.
3. Parallel lines are not everywhere equidistant.
4. The sum of the angles of any triangle is less than two right angles.

These are only a few. Each denies some statement of Euclidean geometry; but the statements on the list are consistent with each other. Or so we are told.

How do we know? How do we know that neither Saccheri nor Lambert actually reached a contradiction? How do we know that the system of Bolyai-Lobachevsky is free from contradictions, that no theorem derived contradicts another, and that no such theorems will ever be found?

It is often said that Bolyai and Lobachevsky proved the independence of the parallel postulate but this is not quite accurate. It was not until some years later that a model was found for their system. Perhaps the first to do so was E. Beltrami (1835–1900). Another model, the one we shall discuss, was presented by H. Poincaré (1854–1912) in his work *Science and Hypothesis*. Both of these models are often credited to Klein, who discussed them in relation to projective geometry.

11.1 A Model

We shall now return to the Euclidean plane and assume the full extent of its properties. The purpose is to construct a model for hyperbolic geometry from a part of that plane.

Consider a circle, which will be called the *fundamental circle*. A Euclidean point will be considered a point in this model if and only if it is interior to the fundamental circle. Points on the circle are to be "points at infinity."

Consider now a circle that cuts the fundamental circle orthogonally; that is, in such a manner that the tangents to each circle at the points of intersection are perpendicular to each other. By a "straight line" we shall mean that *part* of such a circle that is interior to the fundamental circle.

By various techniques of modern Euclidean geometry, many interesting properties of such "lines" can be derived. For example, it can be shown that through two points interior to the fundamental circle, exactly one circle exists that is orthogonal to the fundamental circle; and, that two circles orthogonal to the fundamental circle can intersect in at most one point interior to the fundamental circle.

Furthermore, a distance function can be defined in such a way that the distance between any interior point and points nearer and nearer to the fundamental circle grows larger without bound. Thus this model, in effect, may be regarded as infinite in extent.

In such a model it is actually easy to "see" some of the properties that seemed so strange and were impossible to diagram in a strictly Euclidean framework.

Figure 11.1 shows two parallels PR and PS through a point P not on a line m. It indicates the angle θ that these two lines make with each other.

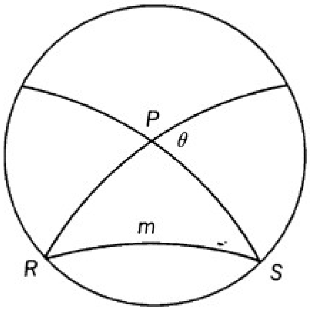

FIGURE 11.1

Figure 11.2 shows a T-V-T $RABS$, in which $R = S$ (a point at infinity). It also shows an ordinary triangle ABD in which the sum of the angles are apparently less than two right angles.

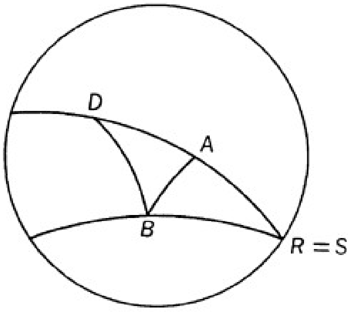

FIGURE 11.2

Figure 11.3 shows a S.Q. $ABCD$ with right angles at A and at B; and a L.Q. $BADE$ with right angles at B, A, and D.

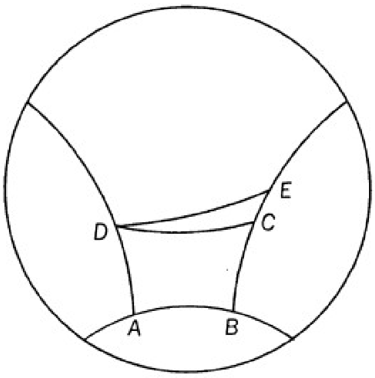

FIGURE 11.3

What makes this model so valuable is not so much that it is easy to diagram the results but that all the properties of hyperbolic geometry can be rigorously established in it; that it can easily be generalized to three dimensions by taking a fundamental sphere and spheres orthogonal to it; and, finally, that all of these results are actually being obtained in Euclidean geometry. The model exists in the Euclidean plane (or space) although it is a model for hyperbolic geometry.

The model, therefore, establishes a relative test for the consistency of hyperbolic geometry. That is, hyperbolic geometry is as consistent as Euclidean geometry; a contradiction could arise only if it already existed in the Euclidean model. It is a matter of fact that over the centuries no one has found an inconsistency in Euclidean geometry; therefore, it is very likely that there is none. This is not a very satisfactory test for consistency but it is the best one we have. But whether or not there exists an absolute test for the consistency of Euclidean geometry is irrelevant to our purposes. The main result of the model is that it shows that hyperbolic geometry is on as firm a logical base as Euclidean geometry; furthermore, it shows once and for all that the Parallel Postulate of Euclidean geometry is independent of the others.

However, the existence of the model also raises new questions. For, if they are both shown to be consistent, how does one decide which is the "true" geometry? Which geometry better describes the space world around us?

11.2 A Model World

Imagine a world which is a large sphere with the property that the temperature changes from point to point within the sphere. As one approaches the center of this sphere the temperature becomes hot and as one approaches the boundary of the sphere the temperature becomes very cold, approaching absolute zero. Furthermore, suppose that all material objects (including the inhabitants of this world) are extremely sensitive to changes in temperature and change instantaneously in size with each change in temperature in such a manner that they grow larger or indefinitely smaller as the temperature increases or decreases, respectively.

To an inhabitant of such a world it would seem to be of infinite extent, for as he approached the boundary of his world he would become smaller and smaller. To reach the boundary would require an infinite number of steps, and he would be as sure that his world was infinite in extent as we seem to be in regard to ours. There would be no way for him to become aware of his situation because any yardstick he carried with him would change in the same proportion.

It can be shown rigorously (using the calculus of variations) that for an inhabitant of such a world the shortest distance between two points would be along the arc of a circle that cuts the bounding sphere orthogonally. This is plausible also, for as he walked from one point to another and if he followed a path which swerved toward the center, his steps would be longer there; thus this path would give a smaller number of steps.

Imagine yet another fact about this world, namely, that light travels along the "shortest paths" along the arcs of circles orthogonal to the bounding sphere. Then, if a man walked toward an object following the shortest path, these shortest paths would actually look straight to him.

What would some of the other properties of this world be? The angle θ shown in Figure 11.1 might very well be visible to the inhabitants. But this angle can be shown to decrease as the point P and line m approach the center of the bounding sphere. Or, to put it another way, if the point P and line m are at a given distance from the center of the sphere, the angle θ decreases indefinitely as the radius of the bounding sphere increases. Thus, the angle θ may be made very small indeed.

Suppose now that Earth is at the center of such a sphere and that the radius of the sphere is so large that the angle θ is smaller than any instrument of man can detect. *In such a situation, the physical world around us would differ in no respect from that to which we are accustomed. But the geometry of Lobachevsky would apply to this world as legitimately as the geometry of Euclid.*

This, then, creates a paradox. If θ is non-existent, we are living in a Euclidean world; if θ exists, no matter how small, we are living in a Lobachevskian world. But how can we ever know that the angle θ is not so small that it remains undetectable even though present? How can we know which geometry is the one that actually describes the world around us?

11.3 Is Euclidean Geometry True?

Modern physics suggests that our world is more likely to be a non-Euclidean one than a Euclidean one, although the particular geometry that physicists have adopted is a modification of Riemann's geometry rather than that of Lobachevsky. But this is not what we want to stress here.

As far as mathematics is concerned, the question as to which geometry is the true one is not merely totally irrelevant but meaningless. As Poincaré says, to ask which is true is like asking which form of measurement is more correct, centimeters or inches? Is the polar coordinate system the "true" one or is it the rectangular system? These are meaningless questions.

This is not to suggest that mathematics need not keep in touch with reality, for almost everyone would agree that a mathematics divorced from reality would soon become sterile. But this is quite different from being concerned with "truth." Rather let us ask which geometry is the most convenient. To this there is a simple answer: Euclidean geometry is the simplest and most convenient to use and shall probably always be used in elementary applications.

This does not mean, however, that no practical value can be found in studying other geometries. All of the geometries, finite and infinite, Euclidean and non-Euclidean, are worthy of investigation. This is so even if our prime objective is to obtain the "truest" geometry, the geometry with the widest range of applications. For the history of mathematics teaches us that some of the most useful results in mathematics have often come from pursuits having no apparent "useful" purpose. We have seen an illustration of this in the unexpectedly rich results that poured forth when mathematicians attempted to answer a seemingly innocent question: can Euclid's Fifth postulate be proved from the others?

references

Bonola, Roberto. *Non-Euclidean Geometry*. Translated by H. S. Carslaw. Chicago: The Open Court Publishing Company, 1912. Reprinted by Dover Publications, Inc., 1955.

Carslaw, H. S. *Non-Euclidean Plane Geometry and Trigonometry*. New York: Longmans, Green and Company, 1916. Reprinted by Chelsea Publishing Company, 1960.

Kulczycki, Stepfan. *Non-Euclidean Geometry*. Translated by S. Knapowski. New York: Pergamon Press, 1961.

Lobachevsky, Nicholai. *Geometrical Researches on the Theory of Parallels*. Translated by G. B. Halsted. Chicago: The Open Court Publishing Company, 1914.

Moise, E. E. *Elementary Geometry from an Advanced Standpoint*. Reading, Massachusetts: Addison-Wesley Publishing Company, Inc., 1963.

Poincaré, Henri. *The Foundations of Science*. Translated by G. B. Halsted. New York and Garrison, New York: The Science Press, 1921.

Sommerville, D. M. Y. *The Elements of Non-Euclidean Geometry*. London: G. Bell and Sons, Ltd., 1914. Reprinted by Dover Publications, Inc., 1958.

Wolfe, Harold E. *Introduction to Non-Euclidean Geometry*. New York: Holt, Rinehart and Winston, Inc., 1945.

Young, J. W. *Lectures on Fundamental Concepts of Algebra and Geometry*. New York: The MacMillan Company, 1911.

)-

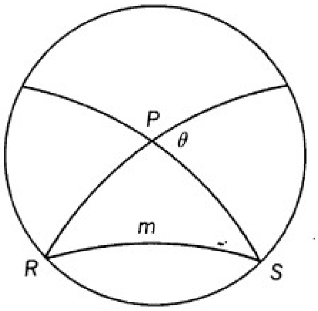

. -

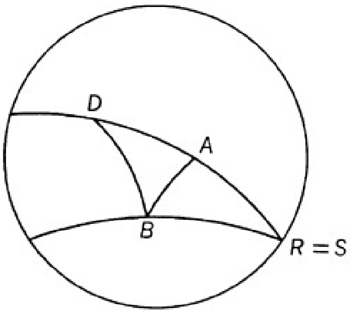

. -

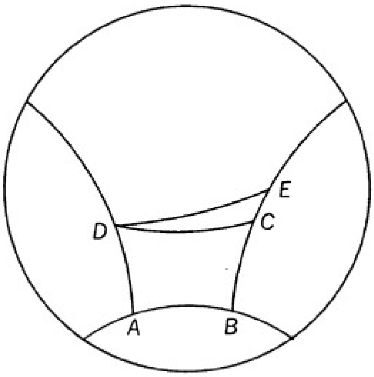

. -

.
: : 7 /<

Chapter 2 - Part 2

Volume-02---20211103

!"#$# % & & !' (! !) * * + , * - . *
/ * !/ * * . ,0 1 2 3 * 3 4 * . , * 3 1 * * *
. 5 "/" & 6 13

-7
-
+* 80
) * 8 0
9 7 0
& :

;. : 5 6 # * .,) * * + ,

) * ($ < , &* . < , = 7 *
>< , ? 7 7 * 3 - ,
3 @ * * * 3 +* '
* 5 . 3 * 1* 0 3 6

8.2 Analogy with Riemannian Geometry

There is an interesting analogy between formula (8.4) (or (8.15)) and H. Weyl's formula [Wey3] for the volume of the tubular ε-neighborhood of a compact, n-dimensional Riemannian submanifold of Euclidean space $\mathbb{R}^d$. Indeed, by [BergGo, Theorem 6.9.9, p. 235], this volume is given by the formula

$$V_M(\varepsilon) = a_0\varepsilon^{d-n} + a_1\varepsilon^{d-n+1} + \cdots + a_n\varepsilon^d. \tag{8.23}$$

When $V_M(\varepsilon)$ is the volume of the two-sided ε-neighborhood, the odd-numbered coefficients $a_1, a_3, \ldots$ all vanish, and for $j = 0, 1, 2, \ldots$, the even numbered coefficient a_{2j} is given by

$$a_{2j} = \frac{1}{d-n+2j}\int_M K_{2j}\delta, \tag{8.24}$$

where K_{2j} is a universal j-th degree polynomial in the curvature tensor of M, and δ is the canonical density on M (see [BergGo, Proposition 6.6.1, p. 214]). Thus, formula (8.23) is a polynomial in ε whose coefficients (8.24) are expressed in terms of the Weyl curvatures in different (integer) dimensions.

If $V_M(\varepsilon)$ is the volume of the one-sided ε-neighborhood (which corresponds more closely to our definition of $V(\varepsilon)$), then the counterpart of (8.23) still holds. In that case, the odd coefficients do not necessarily vanish. We refer to [BergGo], Sections 6.6–6.9, and especially Section 6.9.8 and Theorem 6.9.9, and also to [Gra], for a much more detailed discussion and various helpful examples. For additional references, see also the endnotes in Section 8.5.

In the explicit formula (8.4) (or (8.15)), the volume $V(\varepsilon)$ is expressed as a sum of terms $\varepsilon^{1-\omega}$, where ω ranges over the visible complex dimensions. This suggests that the complex dimensions of fractal strings and the associated residues[7] could have a direct geometric interpretation. The work [Pe2] and [LapPe2–4, LapPeWi1–3] indicates a close relationship to some aspects of geometric measure theory [Fed1, 2] and a possible new notion of fractal curvature measures indexed by the underlying complex dimensions, both in the present one-dimensional context of fractal strings and in higher dimensions. See Section 13.1 for a discussion of the above work.

Remark 8.12. The Weyl curvatures occurring in Weyl's tube formula were shown in [Wey3] to be intrinsic to the submanifold. One of them is closely tied to the notion of Euler characteristic and hence to the Gauss–Bonnet formula (or its higher-dimensional counterpart, due to Chern).

---

[7] Or, more generally, the Laurent expansions at these poles, when they have higher multiplicities.

!"#$# % & & !' (! !) * * + , * - . *
/ * !/) * " * + , * 3) * & , 5 "/" & 6 13

+*
) * A8
9 7 8
&

;. 85: 6

+ 37* * 3 B @ 0 +* 0 0 * . * .
7 * * *) * * * 0 * * ,
B * A < * &* 1 *1
* B * * * C 3 1 2 * * , 0 1
. * 0 1 .,0 3 * . ,D * * @ = B
) * D* 3 * * 3 * 1* $ * 1
* * * 3 1 , * ( ) * * 3
3 3 * * 3 +* " 1 1 0 *. 0 E , *
*33 3 & * *1 1 7 (
$ C * 7 +* * 3) * (1 * 1 7*7 , :80 7
+* 7 * C * ,) * = F * , * 0 * * 3* 3
% & 7 * 3 * * 3 > ?0) * * *7
1 * 0 * 3 * . * 7, +* G , *
H * I " . , * *, ' 5 : :A60) * 3
3, . , +* = 1*1 F * * 0 3* 3 *, 0 :A0
5 * * > ?6
" * * * . 3 *& * 7 * , < * * 0 F *
3 *& 7 7 * F * * , *3 . 7, .0 *3
1*1 7, +* * * *&* *7 * , F *& (3 , * .
* * . ,* 7 * , * (+* 0 1 *.
1 4 0 * 3 C * , J . . " "* -7 &* , F * 3, .
" . * *, ' 0) * 3 7, " 1
* 3 +* = (9 * 5 3* 0 H 3 > ?6
, 1 * 4* " E * * * *
* +* 1 * F *& *11 3 , & , ,
7J 0 0 F 1
) .* 3 . 1 3 0) * 3 3 +* = 1 1 1* * 3
* * 5 7 1 6

After Goldbach and his influence on Euler, we turn to Gauss, who, among the large number of mathematicians with whom Riemann was in contact, was certainly the most influential on him.[9] We shall see in the various sections of the present chapter that this influence was crucial for what concerns the fields of complex analysis, elliptic integrals, topology, differential geometry—the same list as for Euler's influence on Riemann—and also for what concerns his ideas on space. There are other topics in mathematics and physics which were central in the work of Riemann and where he used ideas he learned from Gauss: the Dirichlet principle, magnetism, etc.; they are addressed in several other chapters of the present book.

The first contact between Gauss and Riemann took place probably in 1846, before Gauss became officially Riemann's mentor. In that year, in a letter to his father dated November 5 and translated in [233], Riemann informs the latter about the courses he plans to follow, and among them he mentions a course by Gauss on "the theory of least squares."[10] During his two year stay in Berlin (1847–1849), Riemann continued to study thoroughly Gauss's papers. In another letter to his father, dated May 30, 1849, he writes (translation in [233]):

> Dirichlet has arranged to me to have access to the library. Without his assistance, I fear there would have been obstacles. I am usually in the reading room by nine in the morning, to read two papers by Gauss that are not available anywhere else. I have looked fruitlessly for a long time in the catalog of the royal library for another work of Gauss, which won the Copenhagen prize, and finally just got it through Dr. Dale of the Observatory. I am still studying it.

During the same stay in Berlin, Riemann followed lectures by Dirichlet on topics related to Gauss's works. He writes to his father (letter without date, quoted in [233]):

> My own course of specialization is the one with Dirichlet; he lectures on an area of mathematics to which Gauss owes his entire reputation. I have applied myself very seriously to this subject, not without success, I hope.

Regarding his written production, Riemann endorsed Gauss' principle: *pauca sed matura* (few but ripe).

Riemann, as a child, liked history. In a letter to his father, dated May 3, 1840 (he was 14), he complains about the fact that at his Gymnasium there were fewer

---

[9] Some historians of mathematics claimed that when Riemann enrolled the University of Göttingen, as a doctoral student of Gauss, the latter was old and in poor health, and that furthermore, he disliked teaching. From this, they deduced that Gauss's influence on Riemann was limited. This is in contradiction with the scope and the variety of the mathematical ideas of Riemann for which he stated, in one way or another, but often explicitly, that he got them under the direct influence of Gauss or by reading his works. The influence of a mathematician is not measured by the time spent talking with him or reading his works. Gauss died the year after Riemann obtained his habilitation, but his imprint on him was permanent.

[10] The other courses are on the Cultural History of Greece and Rome, Theology, Recent Church history, General Physiology and Definite Integrals. Riemann had also the possibility to choose courses among Probability, Mineralogy and General Natural History. He adds: "The most useful to me will be mineralogy. Unfortunately it conflicts with Gauss's lecture, since it is scheduled at 10 o'clock, and so I'd be able to attend only if Gauss moved his lecture forward, otherwise it looks like it won't be possible. General Natural History would be very interesting, and I would certainly attend, if along with everything else I had enough money.".

;. 5: 6 3 * 3 +* ,1 . $
0 &* 0 3 3 7, @ F * . *
5/ 3* . * * 60 * * & >A ? 5 8 60 * 3
& * 1*1 7, 0 * 3 7, +* $,1 . * * ,

$$1+\frac{\alpha\beta}{1.\gamma}x+\frac{\alpha(\alpha+1)\beta(\beta+1)}{1.2\gamma(\gamma+1)}x^2+\frac{\alpha(\alpha+1)(\alpha+2)\beta(\beta+1)(\beta+2)}{1.2.3\gamma(\gamma+1)(\gamma+2)}x^3+\dots$$

K0 L0 M * 1* * * 3 &* *7 2

+* 3 3 * * * , * 3 * 7 * 3 * ,
1 . 7, * . . 1 * &* 1* * ', 1 & 3 .
* 7 *3 * 0 3 ,1 . C 3
* * , 1 3 1 1* ' 3 @ * 3 +* 0 1 3) *
C 3 3 * * 3 N
F 80 3 * 4 * $,
* 3 7* ((. 5 8 8 86 1 7 C 3 .
&* C & H * . F 3 30 *
*7 * * H *7 * 1* 3 7, H
*7 & 7 * 3 1

integrals of algebraic functions, is analyzed in Chap. 4 of the present volume, written by Houzel [141].

Section 5 is concerned with the so-called Gauss hypergeometric series. These series, in various forms, were studied by Euler in his *Institutiones calculi integralis* (Foundations of integral calculus), a treatise in three volumes [92] (1768–1770), and in several other papers by him, and by Gauss. The hypergeometric series is a family of functions of the form

$$1+\frac{\alpha\beta}{1.\gamma}x+\frac{\alpha(\alpha+1)\beta(\beta+1)}{1.2\gamma(\gamma+1)}x^2+\frac{\alpha(\alpha+1)(\alpha+2)\beta(\beta+1)(\beta+2)}{1.2.3\gamma(\gamma+1)(\gamma+2)}x^3+\dots$$

where α, β, γ are parameters and where the variable is x.

Gauss considered that almost any transcendental function is obtained from a hypergeometric series by assigning special values to the parameters. By providing such a broad class of functions, the introduction of the hypergeometric series in the field of analysis opened up new paths. Besides Euler and Gauss, the predecessors of Riemann in this field include Pfaff and Kummer.

In Sect. 6, we deal with the zeta function. The history of this function is sometimes traced back to the work of Pietro Mengoli (1625–1686) on the problem of finding the value of the infinite series of inverses of squares of integers. Indeed, it is reasonable to assume that questions about this series were accompanied by questions about the series of inverses of cubes and other powers. But it was Euler again who studied $\sum_{1}^{\infty}\frac{1}{n^s}$ as a function of s (for s real), establishing the functional equation that it satisfies, and the relation with prime numbers. This was the starting point of Riemann's investigations on what became later known as the Riemann zeta function.

In Sect. 7, we make a quick review of some works done by Riemann's predecessors on the notion of space. This is essentially a philosophical debate, but it has a direct impact on mathematics and in particular on Riemann's work on geometry, more especially on his habilitation dissertation. It is in his reflections on space that Riemann introduced in mathematics the notion of Mannigfaltigkeit, which he borrowed from the philosophical literature. This notion reflects Riemann's multi-faced view on space, and it is an ancestor of the modern notion of manifold. Our review of space is necessarily very sketchy, since this notion is one of the most fundamental notions of philosophy, and talking seriously about it would require a whole essay. In particular, there is a lot to say on the philosophy of space in the works of Newton, Euler and Riemann and the comparison between them, but it is not possible to do it in the scope of the present chapter. Our intent here is just to indicate some aspects of the notion of space as it appears in the works of these authors and those of some other philosophers, including Aristotle, Descartes and Kant, and, as much as possible in a short survey, to give some hints on the context in which they emerge.

It is also important to say that the effect of this discussion on space goes far beyond the limits of philosophy. Euler's theories of physics are strongly permeated with his philosophical ideas on space. Gauss's differential geometry was motivated

;. 5 6

De repraesentatione superficiei sphaericae super plano (On the representation of spherical surfaces on a plane) [104], *De proiectione geographica superficiei sphaericae* (On the geographical projections of spherical surfaces) [105], and *De proiectione geographica Deslisliana in mappa generali imperii russici usitata* (On Delisle's geographic projection used in the general map of the Russian empire) [106]. The three memoirs were published in 1777. In the development of the theory, he used complex numbers to represent angle-preserving maps. Lagrange also studied angle-preserving maps, in his memoir *Sur la construction des cartes géographiques* (On the construction of geographical maps) [165], published in 1779.

In fact, the notion of angle-preserving map can be traced back to Greek antiquity, see the survey [194]. We already recalled that Euler, in his didactical treatise *Introductio in analysin infinitorum*, refers explicitly to functions in which the variable is a complex number. De Moivre, already in 1730, considered polynomials defined on the complex plane, and it is conceivable that other mathematicians before him did the same [182]. Remmert, who, besides being a specialist of complex analysis, is a highly respected historian in this field, writes in his *Theory of complex variables* [213] that the theory was born at the moment when Gauss sent a letter to Bessel, dated December 18, 1811, in which he writes[24]:

> At the beginning I would ask anyone who wants to introduce a new function in analysis to clarify whether he intends to confine it to real magnitudes (real values of the argument) and regard the imaginary values as just vestigial—or whether he subscribes to my fundamental proposition that in the realm of magnitudes the imaginary ones $a + b\sqrt{-1} = a + bi$ have to be regarded as enjoying equal rights with the real ones. We are not talking about practical utility here; rather analysis is, to my mind, a self-sufficient science. It would lose immeasurably in beauty and symmetry from the rejection of any fictive magnitudes. At each stage truths, which otherwise are quite generally valid, would have to be encumbered with all sorts of qualifications.

In fact, the letter also shows that at that time Gauss was already aware of the concept of complex integration, including Cauchy's integral theorem; cf. [126] Vol. 8, p. 90–92.

Cauchy, in his *Cours d'analyse* [34] (1821), starts by defining functions of real variables (p. 19), and then passes to complex variables. There are two distinct definitions in the real case, for functions of one or several variables:

> When variable quantities are so tied to each other that, given the value of one of them, we can deduce the values of all the others, we usually conceive these various quantities expressed in terms of one of them, which then bears the name *independent variable*; and the other quantities expressed in terms of the independent variable are what we call functions of that variable.
>
> When variable quantities are so tied to each other that, given the values of some of them, we can deduce the values of all the others, we usually conceive these various quantities expressed in terms of several of them, which then bear the name *independent variables*; and the remaining quantities expressed in terms of the independent variables, are what we call functions of these same variables.[25]

---

[24] The translation is Remmert's; cf. [213] p. 1.

[25] Lorsque des quantités variables sont tellement liées entre elles que, la valeur de l'une d'elles étant donnée, on puisse en conclure les valeurs de toutes les autres, on conçoit d'ordinaire ces

It was only at the publication of Gauss's *Collected Works*,[42] containing in particular his famous notebook, that it became clear that Gauss's assertion concerning the fact that he had discovered before Abel most of the properties of elliptic functions, including their double periodicity, was correct. One of the first results of Abel concerns integrals of arcs of lemniscate, a curve which he showed to be divisible by ruler and compass into n equal parts, for the same values of n for which the circle is divisible into n equal parts. The same result was stated without proof in Gauss's *Disquisitiones arithmeticae* [121].

Abel's first major results on elliptic functions are contained in his 1827 paper *Recherches sur les fonctions elliptiques* (Researches on elliptic functions) [1]. He explains there the double periodicity of these functions, as well as their multiplication and division properties. The analogy with circular functions is again highlighted. At the beginning of his paper, Abel talks about his famous predecessors, Euler, Lagrange and Legendre. He writes (p. 101):

> The first idea of these [elliptic] functions were given by the immortal Euler, who showed that the separable equation
>
> $$\frac{\partial x}{\sqrt{\alpha+\beta x+\gamma x^2+\delta x^3+\epsilon x^4}}+\frac{\partial y}{\sqrt{\alpha+\beta y+\gamma y^2+\delta y^3+\epsilon y^4}}=0$$
>
> is algebraically integrable. After Euler, Lagrange added something, when he gave his elegant theory of the transformation of the integral
>
> $$\int\frac{R\,dx}{\sqrt{(1-p^2x^2)(1-q^2x^2)}},$$
>
> where R is a rational function of x. But the first, if I am not mistaken, who went thoroughly into the nature of these functions, is Mr. Legendre, who, first in a memoir on elliptic functions, and then in his excellent *Exercices de mathématiques*, developed numerous elegant properties of these functions, and showed their usefulness.[43]

Riemann was already interested in elliptic functions while he was a student in Berlin. Klein, in his *Development of mathematics in the 19th century* [162] (Chap-

---

[42] Gauss's collected works, *Carl Friedrich Gauss' Werke*, in twelve volumes, were published between 1863 and 1929.

[43] La première idée de ces fonctions a été donnée par l'immortel Euler, en démontrant que l'équation séparée

$$\frac{\partial x}{\sqrt{\alpha+\beta x+\gamma x^2+\delta x^3+\epsilon x^4}}+\frac{\partial y}{\sqrt{\alpha+\beta y+\gamma y^2+\delta y^3+\epsilon y^4}}=0$$

est intégrable algébriquement. Après Euler, Lagrange y a ajouté quelque chose, en donnant son élégante théorie de la transformation de l'intégrale

$$\int\frac{R\,dx}{\sqrt{(1-p^2x^2)(1-q^2x^2)}},$$

où R est une fonction rationnelle de x. Mais le premier, et si je ne me trompe, le seul, qui ait approfondi la nature de ces fonctions, est M. Legendre, qui d'abord dans un mémoire sur les fonctions elliptiques, et ensuite dans ses excellents exercices de mathématiques, a développé nombre de propriétés élégantes de ces fonctions, et a montré leur application.

;. 5886 9,1 .

5 Hypergeometric Series

The theory of the hypergeometric series is another topic which Riemann tackled and whose roots involve in an essential way the works of Euler and Gauss. Riemann's main paper on the subject is *Beiträge zur Theorie der durch die Gauss'sche Reihe* $F(\alpha, \beta, \gamma, x)$ *darstellbaren Functionen* (Contribution to the theory of functions representable by Gauss's series $F(\alpha, \beta, \gamma, x)$) [222], published in 1857. The work in this paper was used by Riemann later in his development of the theory of analytic differential equations. There are also fragments on the same subject published in Riemann's *Collected works.*

The hypergeometric series is a function of the form

$$F(\alpha, \beta, \gamma, x) = 1 + \frac{\alpha\beta}{1.\gamma}x + \frac{\alpha(\alpha+1)\beta(\beta+1)}{1.2\gamma(\gamma+1)}x^2 + \frac{\alpha(\alpha+1)(\alpha+2)\beta(\beta+1)(\beta+2)}{1.2.3\gamma(\gamma+1)(\gamma+2)}x^3 + \dots$$

where x is the variable.

The term "hypergeometric series" appears in Euler's *Institutiones calculi integralis* [92] (1769), Chapter XI. The series is a solution of the so-called Euler hypergeometric differential equation which appears in Chapters VIII and XI of the same treatise. As a matter of fact, this name was given to several different but closely related objects. Euler, in one of his earliest memoir *De progressionibus transcendentibus seu quarum termini generales algebraice dari nequeunt* (On transcendental progressions, that is, those whose general terms cannot be given algebraically) [65], published in 1738, starts by mentioning Wallis's "hypergeometric series" $1! + 2! + 3! + 4! + \cdots$ (without the factorial notation). The terminology here refers to the fact that in analogy with the case of geometric progressions, where each term is obtained from the preceding one by multiplying it by a constant, one defined a hypergeometric progression as a progression in which each term is obtained from the preceding one by multiplying it by a factor which increases by a unit at each step. Wallis's papers on this subject include [251] (1, Scholium to Proposition 190) and [252] (p. 315).

Gauss mentions a hypergeometric series in his doctoral dissertation *Demonstratio nova theorematis omnem functionem algebraicam rationalem integram unius variabilis in factores reales primi vel secundi gradus resolvi posse* [119] (New proof of the theorem that every rational integral algebraic function of one variable can be resolved into real factors of the first or second degree) (1799).

We refer the reader to the paper [59] for a comprehensive history of the hypergeometric series.

At the beginning of his announcement of his memoir [222], Riemann states: "This memoir treats a class of functions which are useful to solve various problems in mathematical physics." As a matter of fact, these functions are still commonly used today in mathematical physics. Riemann notes that the name *hypergeometric series* was first proposed by Pfaff, for a more general series, whereas Euler, after Wallis, used such a name for a series which is slightly different. Pfaff was Gauss's friend, and had been his teacher. He studied this function in his book *Disquisitiones analyticae maxime ad calculum integralem et doctrinam serierum pertinentes* (Analytic investigations most relevant for integral calculus and the doctrine of series) [195] (1797). Gauss has a series of unpublished results on the hypergeometric series, which he communicated to the astronomer Bessel, who was also his friend, in a letter dated September 3, 1805. The results were used by Gauss in his later works. In his writings on the subject, Gauss used continued fractions in his study of the quotient of two hypergeometric series. He developed these ideas in his paper *Disquisitiones generales circa serium* $1 + \frac{\alpha\beta}{1.\gamma}x + \frac{\alpha(\alpha+1)\beta(\beta+1)}{1.2\gamma(\gamma+1)}xx + \frac{\alpha(\alpha+1)(\alpha+2)\beta(\beta+1)(\beta+2)}{1.2.3\gamma(\gamma+1)(\gamma+2)}x^3 + \ldots$ *etc.* (General investigations on the series $1 + \frac{\alpha\beta}{1.\gamma}x + \frac{\alpha(\alpha+1)\beta(\beta+1)}{1.2\gamma(\gamma+1)}xx + \frac{\alpha(\alpha+1)(\alpha+2)\beta(\beta+1)(\beta+2)}{1.2.3\gamma(\gamma+1)(\gamma+2)}x^3 + \ldots$ etc.) [122]. The same year, he wrote another paper on the same subject which he never published but which is contained in his *Collected Works* edition [123]. Riemann, in his paper [222], proved that these fractions converge in the complex plane cut along the subset $[2, +\infty]$ of the x-axis. In the same memoir, he introduced in the study of the hypergeometic functions a new method, which applies to all functions that satisfy linear differential equations with algebraic coefficients. He recalls in the announcement of that memoir, published in the *Göttinger Nachrichten*, No. 1, 1857, that Euler and Gauss made a thorough study of these functions from the theoretical point of view.

In the introduction to his paper [122] Gauss declares that practically any transcendental function that appears in analysis may be obtained as a special case of the hypergeometric series. In fact, it is known that functions like $\log(1 + z)$, $\arcsin z$ and several orthogonal polynomials, including Legendre polynomials and Chebyshev polynomials, can be expressed using hypergeometric functions. The so-called confluent hypergeometric function (or Kummer's function) is a limit of the hypergeometric function.

The introduction of the hypergeometric series brought a whole new class of new functions to the field of analysis which, at least in the times of Euler, consisted in the study of functions.

6 The Zeta Function

This section is concerned with Riemann's article *Über die Anzahl der Primzahlen unter einer gegebenen Grösse* (On the number of primes less than a given magnitude) [219]. This memoir, which is only 8 pages long, changed the course of mathematics. Riemann wrote it at the occasion of his election to the Berlin Academy of Sciences, on August 11, 1859. Every newly elected member at that academy was asked to report on his most recent research, and Riemann chose this topic. A short history of the subject will show that the list of predecessors of Riemann in this field includes names which are familiar to us now: Euler, as always, then Legendre, Dirichlet and Gauss.

Riemann starts his memoir by recalling that Gauss and Dirichlet had been interested in this subject several years before him. He displays the following formula, which he recalls was noted by Euler, and which was his own departure point:

O * /

> and even engines and motion in character, in the same way as algebra represents numbers in magnitude.[65]

In the same letter ([176] p. 570), Leibniz adds:

> I found the elements of a new characteristic, completely different from Algebra and which will have great advantages for the exact and natural mental representation, although without figures, of everything that depends on the imagination. Algebra is nothing but the characteristic of undetermined numbers or magnitudes. But it does not directly express the place, angles and motions, from which it follows that it is often difficult to reduce, in a computation, what is in a figure, and that it is even more difficult to find geometrical proofs and constructions which are enough practical even when the Algebraic calculus is all done.[66]

Together with his letter to Huygens, Leibniz included the manuscript of an essay he wrote on the new subject. He writes, in the same letter ([176] p. 571):

> But since I don't see that anybody else has ever had the same thought, which makes me fear that it might be lost if I do not get enough time to complete it, I will add here an essay which seems to me important, and which will suffice at least to rendre my aim more credible and easier to conceive, so that if something prevents its realization now, it will serve as a monument for posterity and give the possibility to somebody else to finish it.[67]

He then explains in more detail his vision of this new domain of mathematics, and where it stands with respect to algebra and geometry, giving several examples of a formalism to denote loci, showing how this formalism expresses statements such that the intersection of two spherical surfaces is a circle, and the intersection of two planes is a line.

Leibniz' letter ends with the words ([174] p. 25):

---

[65] Après tous les progrès que j'ai faits en ces matières, je ne suis pas encore content de l'Algèbre, en ce qu'elle ne donne ni les plus courtes voies, ni les plus belles constructions de Géométrie. C'est pourquoi lorsqu'il s'agit de cela, je crois qu'il nous faut encore une autre analyse proprement géométrique ou linéaire, qui nous exprime directement *situm*, comme l'algèbre exprime *magnitudinem*. Et je crois d'en avoir le moyen, et qu'on pourrait représenter des figures et même des machines et mouvements en caractères, comme l'algèbre représente les nombres en grandeurs. [We have modernized the French.].

[66] J'ai trouvé quelques éléments d'une nouvelle caractéristique, tout à fait différente de l'Algèbre, et qui aura de grands avantages pour représenter à l'esprit exactement et au naturel, quoique sans figures, tout ce qui dépend de l'imagination. L'Algèbre n'est autre chose que la caractéristique des nombres indéterminés ou des grandeurs. Mais elle n'exprime pas directement la situation, les angles et les mouvements, d'où vient qu'il est souvent difficile de réduire dans un calcul ce qui est dans la figure, et qu'il est encore plus difficile de trouver des démonstrations et des constructions géométriques assez commodes lors même que le calcul d'Algèbre est tout fait.

[67] Mais comme je ne remarque pas que quelqu'autre ait jamais eu la même pensée, ce qui me fait craindre qu'elle ne se perde, si je n'y ai pas le temps de l'achever, j'ajouterai ici un essai qui me paraît considérable, et qui suffira au moins à rendre mon dessein plus croyable et plus aisé à concevoir, afin que si quelque hasard en empêche la perfection à présent, ceci serve de monument à la postérité, et donne lieu à quelque autre d'en venir à bout.

> I have only one remark to add, namely, that I see that it is possible to extend the characteristic to things which are not subject to imagination. But this is too important and it would lead us too far for me to be able to explain myself on that in a few words.[68]

When Leibniz started his correspondence with Huygens, the latter was already a well established scientist whose achievements were behind him, and it was not easy to convince him of the usefulness of a new theory. Huygens thought that the theory was too abstract and he remained skeptical about it. He was above all a geometer working on concrete geometrical problems.

One may recall that when Leibniz sent him the above letter, Huygens was considered as a world authority in geometry and physics. He was settled in Paris since 15 years, and he was a leading member of the *Académie Royale des Sciences*. Leibniz had studied mathematics with Huygens, who was seventeen years older than him, and he considered him as his mentor. Huygens responded to Leibniz in a letter dated November 22, 1679 ([176] p. 577):

> I have examined carefully what you are asking me regarding your new characteristic, and to be frank with you, I cannot not conceive the fact that you have so much expectations from what you spread on me. Because your example of places concerns only realities which were already perfectly known, and the proposition saying that the intersection of a plane and a spherical surface makes the circumference of a circle does not follow clearly. Finally, I cannot see in what way you can apply your characteristic to which you seem you want to reduce all these different matters, like the quadratures, the invention of curves by the properties of tangents, the irrational roots of equations, Diophantus' problems, the shortest and the most beautiful constructions of the geometric problems. And what still appears to me stranger than anything else, the invention and the explanation of machines. I say it to you unsuspiciously, in my opinion this is only wishful thinking, and I need other proofs in order to believe that there could be some reality in what you present. I would nevertheless restrain myself from saying that you are mistaken, knowing the subtlety and the deepness of your mind. I only beg you that the magnificence of the things you are searching won't let you postpone from giving us those which you already found, like this Arithmetic Quadrature you discovered, concerning the roots of the equations beyond the cubical, if you are still satisfied with it.[69]

---

[68] Je n'ai qu'une remarque à ajouter, c'est que je vois qu'il est possible d'étendre la caractéristique jusqu'aux choses, qui ne sont pas sujettes à l'imagination ; mais cela est trop important et va trop loin pour que je me puisse expliquer là-dessus en peu de paroles.

[69] J'ai examiné attentivement ce que vous me demandez touchant votre nouvelle caractéristique, mais pour vous l'avouer franchement, je ne conçois pas parce que vous m'en étalez, que vous y puissiez fonder de si grandes espérances. Car votre exemple des Lieux ne regarde que des vérités qui nous étaient déjà fort connues, et la proposition de ce que l'intersection d'un plan et d'une surface sphérique fait la circonférence d'un cercle, s'y conclut assez obscurément. Enfin, je ne vois point de quel biais vous pourriez appliquer votre caractéristique à toutes ces choses différentes qu'il semble que vous y vouliez réduire, comme les quadratures, l'invention des courbes par la propriétés des tangentes, les racines irrationnelles des Équations, les problèmes de Diophante, les plus courtes et plus belles constructions des problèmes géométriques. Et ce qui me paraît encore le plus étrange, l'invention et l'explication des machines. Je vous le dis ingénument, ce ne sont là à mon avis que de beaux souhaits, et il me faudrait d'autres preuves pour croire qu'il y eût de la réalité dans ce que vous avancez. Je n'ai pourtant garde de dire que vous vous abusiez, connaissant d'ailleurs la subtilité et profondeur de votre esprit. Je vous prie seulement que la grandeur des choses que vous cherchez ne vous fasse point différer de nous donner celles que vous avez déjà trouvées, comme est

In another letter dated January 11, 1680 ([176] p. 584) Huygens writes:

> For what concerns the effects of your characteristic, I see that you insist on being persuaded of them, but as you say yourself, the examples will be more important than reasonings. This is why I am asking you much simpler examples, but capable of overcoming my incredulity, because that of the places, I confess, does not seem to me of that sort.[70]

The essay that Leibniz sent did not obtain Huygens' backing and it remained hidden among other manuscripts in Huygens' estate. It was published for the first time in 1833, and drew the attention of several nineteenth-century mathematicians, including Grassmann (1809–1877), the founder of the theory of vector spaces, who realized its importance for the new field of topology. There are two recent editions of this text, both included in doctoral dissertations, by J. Acheverría [175] (1995), in France, and by de Risi, [234] (2007), in Germany. The two dissertations contain other texts by Leibniz on the same subject.

Leibniz used several names for the new field, including *analysis situs, geometria situs, characteristica situs, characteristica geometrica, analysis geometrica, speciosa situs*, etc.

The first mathematician who worked consciously on topological questions is Euler. These questions include the definition and the invariance of the Euler characteristic of a convex polyheron, the problem known as that of the Königsberg seven bridges, another question related to the Knight's tour on the chessboard, and a musical question concerning a graph known as the *speculum musicum*. This graph was introduced in Euler's *Tentamen novae theoriae musicae ex certissimis harmoniae principiis dilucide expositae* (A attempt at a new theory of music, exposed in all clearness according to the most well-founded principles of harmony) [71]. Its vertices are the twelve notes of the chromatic scale, and the edges connect two elements which differ by a fifth or a major third with the property that one may traverse all the edges of the graph passing exactly once by each note. The article [189] is a detailed survey of the work of Euler on these questions. In the present section, we start by reviewing the work of Euler on the question of the seven bridges of Königsberg. This work shows that Euler considered himself as the direct heir of Leibniz for what concerns the field of topology. We shall then describe in detail the works of Euler and Descartes on the Euler characteristic, a question which is directly related to the topological classification of surfaces, which was one of Riemann's major achievements in topology. We recall that Euler formulated this result for a surface which is the boundary of a convex polyhedron having F faces, A edges and S vertices; the formula is then:

$$F - A + S = 2.$$

(Footnote 69 continued)
cette Quadrature Arithmétique et que vous avez découvert pour les racines des équations au-delà du cube, si vous en êtes content vous-même.

[70] Pour ce qui est des effets de votre caractéristique, je vois que vous persistez à en être persuadé, mais, comme vous dites vous-même, les exemples toucheront plus que les raisonnements. C'est pourquoi je vous en demande des plus simples, mais propres à convaincre mon incrédulité, car celui des lieux, je l'avoue, ne me paraît pas de cette sorte.

We start with the problem of the Königsberg bridges.

In the eighteenth century, the city of Königsberg[71] consisted of four quarters separated by branches of the river Pregel and related by seven bridges. The famous "problem of the seven bridges of Königsberg" asks for a path in that city that starts at a given point and returns to the same point after crossing once and only once each of the seven bridges. At the time of Euler, this was a popular question among the inhabitants of Königsberg.

Euler showed that such a path does not exist. He presented his solution to the Saint Petersburg Academy of Sciences on August 26, 1735, and in the same year he wrote a memoir on the solution of a more general problem entitled *Solutio problematis ad geometriam situs pertinentis* (Solution of a problem relative to the geometry of position) [74]. Euler learned about the problem from a letter, dated March 7, 1736, sent to him by Carl Leonhard Gottlieb Ehler, one of his friends who was the mayor of Danzig[72] and who had worked as an astronomer in Berlin. Euler solved the problem just after he received the letter. In a letter dated March 13, 1736, written to Giovanni Marioni, an Italian astronomer working at the court of Vienna, Euler declares that he became interested in this question because he realized that the problem could not be solved using geometry, algebra or combinatorics, and that therefore he wondered whether "it belonged to the 'geometry of position,' (*geometria situs*) which Leibniz has so much sought for." In the same letter, Euler announced to Marioni that after some thought, he found a proof which applies not only to that case, but to all similar problems.

In the introduction of his paper, Euler writes (translation from [29]):

> In addition to that branch of geometry which is concerned with magnitudes, and which has always received the greatest attention, there is another branch, previously almost unknown, which Leibniz first mentioned, calling it the geometry of position. This branch is concerned only with the determination of position and its properties; it does not involve measurements, nor calculations made with them. It has not yet been satisfactorily determined what kind of problems are relevant to this geometry of position, or what methods should be used in solving them. Hence, when a problem was recently mentioned, which seemed geometrical but was so constructed that it did not require the measurement of distances, nor did calculation help at all, I had no doubt that it was concerned with the geometry of position, especially as its solution involved only position, and no calculation was of any use. I have therefore decided to give here the method which I have found for solving this kind of problem, as an example of the geometry of position.

Euler's work on this problem is commented in several articles and books.

We now come to Euler's polyhedron formula, and we start with Descartes.

Long before Euler came out with his formula $F - A + S = 2$ relating the faces (F), edges (A) and vertices (S) of a convex polyhedron, Descartes obtained an equivalent result, with a geometric proof, involving the solid angles and the dihedral angles between the faces. Described in modern terms, Descartes' proof consists in computing in two different manners the total curvature of the boundary of the polyhedron.

---

[71] Today, the city of Königsberg, called Kaliningrad, is part of a Russian exclave between Poland and Lithuania on the Baltic Sea.

[72] Today, Danzig is the city of Gdansk, in Poland.

-7
< * 1 7 NP . 7 . 7 3.

;. 885 6

> the surface dependent on that of the various sections that one can perform on it by cutting it with planes.[89]

Meusnier's work is surveyed by Darboux in the paper [49].

Dupin is mostly known for the so-called Dupin indicatrix, a geometric device which characterizes the shape of a surface at a given point and which turned out to be related to the Gaussian curvature at that point. His famous work on geometry bears the title *Développements de géométrie, avec des applications à la stabilité des vaisseaux, aux déblais et remblais, au défilement, à l'optique, etc. pour faire suite à la géométrie descriptive et à la géométrie analytique de M. Monge : Théorie.* (Developments in geometry, with applications to the stability of vessels, cuts and fills, scrolling, optics, etc. as a sequel to the descriptive and analytic geometry of Mr. Monge: Theory) [58], 1813.

Rodrigues introduced, before Gauss, the Gauss map, and showed that at a given point the ratio of the area of the image to the area on the surface is equal (at the infinitesimal level) to the product of the two principal curvatures (those defined by Euler), cf. [235]. This result pre-dates that of Gauss, but the fact that curvature is an isometry invariant is however absent from Rodrigues' work.

Finally, it is fair to mention that Riemann's work in high dimensions was prepared by works of other mathematicians done in higher dimensions, and we mention in particular Jacobi [145] on the reduction of pairs of quadratic forms, Grassmann on higher-dimensional linear algebra [131], and Cayley [41] on higher-dimensional analytic geometry.

10 Trigonometric Series

Riemann's habilitation dissertation (*Habilitationsschrift*) is concerned with trigonometric series. Its title is *Über die Darstellbarkeit einer Function durch eine trigonometrische Reihe* (On the representability of a function by a trigonometric series) [215]. The dissertation is divided into two parts. The first part is a survey of the history of the problem of representing by a trigonometric series a function which is arbitrary, but which—according to Riemann's words at the beginning of his memoir—is "given graphically." It is important to note the last fact, since, for instance, a function which is discontinuous at a dense set of points cannot be given graphically. Dealing with the most general functions is part of the subject of the second part of Riemann's memoir.

We shall report on the historical part of Riemann's memoir. It involves several mathematicians, in particular Euler, although not directly his work on trigonometric series. We therefore note right away that Euler uses trigonometric series in his two

---

[89] M. Euler a traité la même matière dans un fort beau mémoire, imprimé en 1760 parmi ceux de l'Académie de Berlin. Cet illustre géomètre envisage la question d'une manière différente de celle que nous venons d'exposer ; il fait dépendre la courbure d'un élément de surface de celle des différentes sections qu'on peut y faire en le coupant par des plans.

" 1 , (* 3 " 1 3 * 2 0 * . 3 &
* * 4 *1 * * * * . & 1 * 3 3
7 * 3 +* * &* * * 1

We conclude with Grothendieck's reference to Riemann. In his *Sketch of a program*, [134], he writes (p. 240 of the English translation)[111]:

> Whereas in my research before 1970, my attention was systematically directed towards objects of maximal generality, in order to uncover a general language adequate for the world of algebraic geometry [...] here I was brought back, via objects so simple that a child learns them while playing, to the beginnings and origins of algebraic geometry, familiar to Riemann and his followers!

Topic	**Euler**	**Riemann**
Functions of a complex variable	• Introductio in analysin infinitorum (1748) • De repraesentatione superficiei sphaericae super plano (1777)	• Grundlagen für eine allgemeine Theorie der Functionen einer veränderlichen complexen Grösse (1851) • Theorie der Abel'schen Functionen (1857)
Elliptic and Abelian integrals	• Specimen de constructione aequationum differentialium sine indeterminatarum separatione (1738) • Observationes de comparatione arcuum curvarum irrectificibilium (1761) • De integratione aequationis differentialis $\frac{mdx}{\sqrt{1-x^4}} = \frac{ndy}{\sqrt{1-y^4}}$ (1761)	• Grundlagen für eine allgemeine Theorie der Functionen einer veränderlichen complexen Grösse (1851) • Theorie der Abel'schen Functionen (1857) • Über das Verschwinden der ϑ-Functionen (1857)
Hypergeometric series	• De summatione innumerabilium progressionum (1738) • Institutionum calculi integralis volumen secundum (1769) • Specimen transformationis singularis serierum (1778)	• Beiträge zur Theorie der durch die Gauss'sche Reihe $F(\alpha, \beta, \gamma, x)$ darstellbaren Functionen (1857)
The zeta function	• Variae observationes circa series infinitas (1744) • Remarques sur un beau rapport entre les series des puissances tant directes que réciproques (1749)	• Über die Anzahl der Primzahlen unter einer gegebenen Grösse (1859)
Integration	• Institutionum calculi integralis (1768–1770)	• Über die Darstellbarkeit einer Function durch eine trigonometrische Reihe (1854)

---

[111] The English translation is by Lochak and Schneps.

Topic	**Euler**	**Riemann**
Space and philosophy of nature	• Anleitung zur Naturlehre (1745) • Reflexions sur l'espace et le temps (1748)	• Grundlagen für eine allgemeine Theorie der Functionen einer veränderlichen complexen Grösse (1851)
Topology	• Solutio problematis ad geometriam situs pertinentis (1741) • Elementa doctrinae solidorum (1758) • Demonstratio nonnullarum insignium proprietatum, quibus solida hedris planis inclusasunt praedita (1758)	• Grundlagen für eine allgemeine Theorie der Functionen einer veränderlichen complexen Grösse (1851) • Theorie der Abel'schen Functionen (1857) • Über die Hypothesen, welche der Geometrie zu Grunde liegen (1854)
Differential geometry	• Introductio in analysin infinitorum (1748) • Recherches sur la courbure des surfaces (1767)	• Über die Hypothesen, welche der Geometrie zu Grunde liegen (1854) • Commentatio mathematica, qua respondere tentatur quaestioni ab Ill$^{\mathrm{ma}}$ Academia Parisiensi propositae (1861) • Ein beitrag zu den Untersuchungen über die flüssigen Bewegung eines gleichartigen Ellipsoides (1861) • Über die Fläche vom kleinsten Inhalt bei gegebener Begrenzung (1867)
Trigonometric series	• Recherches sur la question des inegalités du mouvement de Saturne et de Jupiter (1748)	• Über die Darstellbarkeit einer Function durch eine trigonometrische Reihe (1854)
Acoustics	• Dissertatio physica de sono (1727) • Sur la vibration des cordes (1748)	• Über die Fortpflanzung ebener Luftwellen von endlicher Schwingungsweite (1860)

Acknowledgements I would like to thank Vincent Alberge, Jeremy Gray and Marie Pascale Hautefeuille who read a preliminary version of this chapter and suggested corrections.

References

1. N.H. Abel, Recherches sur les fonctions elliptiques. J. für die Reine und Angew. Math. **2**, 101–181 (1827)
2. L.V. Ahlfors, Quasiconformal mappings and their applications. Lect. Modern Math. **2**, 151–164 (1964)
3. d'Alembert, Jean le Rond, Opuscules mathématiques, Paris, Claude-Antoine Jombert, 1780
4. d'Alembert, Jean le Rond, Recherches sur les vibrations des cordes sonores & Supplément, Opuscules mathématiques, Paris, 1761, tome 1, pp. 1–73

;. 5 6

$$ds = \sqrt{\sum (dx)^2}.$$

Riemann noted that there are other possibilities. For example, ds could be the fourth root of a fourth power expression, but he did not see many possibilities for geometry there and he set this aside.

The quadratic form, however, did interest him. It contains $n(n+1)/2$ coefficients, of which n can be altered by a change of variables, so it depends essentially on $n(n-1)/2$ coefficients that are determined by the manifold. The example of $\sqrt{\sum (dx)^2}$ is therefore special, and Riemann proposed to call such manifolds flat.

To proceed further, Riemann considered the infinitesimal triangle with one vertex at the origin, one on a geodesic out of the origin to the point $(x_1, x_2, \ldots, x_n)$, and one on a geodesic out of the origin to the point $(dx_1, dx_2, \ldots, dx_n)$. The quotient of $\sqrt{\sum (dx)^2}$ by the area of this triangle measures the departure of this infinitesimal region from flatness, and divided by $-3/4$ is in fact the Gaussian curvature of the surface. So the curvature of an n-manifold can be understood by knowing $n(n-1)/2$ surface curvatures (the sectional curvatures, as we would say).

This led Riemann to explain the difference between intrinsic and extrinsic properties of a surface. He explained that for a sphere, which has an intrinsic geometry different from a plane, the Gaussian curvature multiplied by the area of an infinitesimal geodesic triangle is half the excess of the sum of its angles over π. This allowed him to express the belief that the geometry of an n-dimensional manifold could be understood by understanding its sectional curvatures.

Flat manifolds, he observed, have every sectional curvature zero. They are therefore a special case of the manifolds of constant curvature (that is, having the same sectional curvatures everywhere) and in these manifolds geometric figures can be freely moved around without stretching. To give examples of such manifolds he wrote down the metric

$$ds = \frac{1}{1 + \frac{\alpha}{4} \sum x^2} \sqrt{\sum (dx)^2},$$

where α is the curvature.

It was evident that the surfaces of constant positive curvature are spheres; the sphere of radius r has curvature r^{-2}. Riemann gave a complicated description of how to fit all the surfaces of constant curvature into one family. On this description, the cylinder is the example of a surface of zero curvature, and surfaces of constant negative curvature are locally like the saddle-shaped part of a torus.

In the third and final part of his paper [24] Riemann discussed how his ideas might apply to space. If all the sectional curvatures are zero, the space is Euclidean. But if we assume only that there is free mobility of bodies then space is described as a three-dimensional manifold of constant curvature, which can be determined from the knowledge of the sum of the angles in any triangle. Or, one could assume that length is independent of position but not direction.

$ H 5 32 6Q 7, * * * . * 3 1* C *
. B* 0 * 3 3 & 3 3 7, R : * +* * &* *

Riemann's Work on Minimal Surfaces

Sumio Yamada

Abstract Three months before his death in 1866, Riemann left a set of notes to K. Hattendorff, a disciple of his, on minimal surfaces with boundary. Afterwards, Hattendorff supplied the text to the notes mostly consisting of computations, which became the two papers on the subject: "On the surface of least area with a given boundary" and "Examples of surfaces of least area with a given boundary." We will go over the expositions and provide an overview from the modern viewpoint, make some comments on Riemann-Hattendorff's text, and compare the work with that of Weierstrass on the same subject.

Keywords Riemann · Minimal surface · Calculus of variations

2000 Mathematics Subject Classication: 53A10 · 30F45

1 Introduction

We first recall the statement of the Riemann mapping theorem,

Theorem 1.1 *If* Ω *is a non-empty simply connected open proper subset of the complex plane* $\mathbb{C}$*, then there exists a bi-holomorphic mapping* f *from* Ω *onto the open unit disk* $D = \{z \in \mathbb{C} \mid |z| < 1\}$.

We remark that the resulting plane region Ω is a minimal surface spanned by the boundary $\partial\Omega$ when the complex plane is identified with the xy-plane in the three dimensional ambient space $\mathbb{R}^3$.

It was in Riemann's thesis [8] in 1851 where the cerebrated Riemann Mapping Theorem was first presented. There he utilized the Dirichlet principle in order to obtain a map which is harmonic and conformal from a simply connected region of the complex plane to the unit disc. Riemann's proof was incomplete as the existence of such harmonic functions minimizing the Dirichlet energy functional is not always

---

S. Yamada (✉)
Department of Mathematics, Gakushuin University, Tokyo 171-8588, Japan
e-mail: yamada@math.gakushuin.ac.jp

© Springer International Publishing AG 2017

L. Ji et al. (eds.), *From Riemann to Differential Geometry and Relativity*,
DOI 10.1007/978-3-319-60039-0_5

guaranteed, a defect first pointed out by Weierstrass in 1859. The subsequent important historical development in analysis and geometry initiated by Riemann's idea of using the Dirichlet principle, which led to the wealth of mathematics around the so-called Plateau problem, is comprehensively described in Courant's book [2].

Hans Lewy, in his introduction [5] to an edition of collected works of Riemann (1953), speculates that the work of Riemann on the subject of minimal surfaces in $\mathbb{R}^3$ [9, 10] may well be an attempt to rectify his own proof of the Riemann Mapping Theorem, as the inverse map $f^{-1} : D \to \Omega \subset \mathbb{R}^3$ of the Riemann mapping provides a harmonic and conformal parameterization of the *minimal* surface $\Omega \subset xy$-plane, where the xy-plane is identified as $\mathbb{C}$. The set of minimal surfaces Riemann succeeded in constructing, however, have boundary sets only of special types; lines, line segments, circles, with which one would not be able to approximate an arbitrary boundary curve $\partial\Omega$ as required by the statement of the Riemann Mapping Theorem.

Having stated this observation about a failed attempt, however, it is important to recognize the true value of what Riemann created in the field of minimal surfaces within the two posthumously published papers [9, 10]. Historically, the study of minimal surfaces and later minimal submanifolds has led to many interesting applications, not only in differential geometry, but also in general relativity, material sciences, industrial design, among others. Riemann was correct to foresee the scientific potential the subject offered. As for the minimal surfaces in $\mathbb{R}^3$, much of the subsequent development on the subject up to the present day is based on the so-called Weierstrass-Enneper representation formula, which is comprehensively presented in [11].

The goal of this chapter is to illustrate that inside the notes Riemann left to Hattendorff, based on the computations Riemann had made over 1860–61, much of the well-known classical results on minimal surfaces including the Weierstrass-Enneper representation, Schwarz's explicit construction of minimal surfaces, as well as the Schwarz-Christoffel transformation, are contained in essence. We will demonstrate this by reading the text of [9] through §1–13 with comments added as appropriate. In particular in the second to the last section, we provide a direct comparison between the Weierstrass-Enneper representation and the much less known Riemann representation of minimal surface.

2 On the Surface of Least Area with a Given Boundary

Regarding the treatment of surfaces in $\mathbb{R}^3$ in this article, it is safe to assume that Riemann was fully informed of the surface theory of Gauss [4] where every surface is locally parameterized by two independent real variables u, v.

In §1 and §2, a disc-type surface Ω is parameterized by two parameters p, q which are effectively the polar coordinates in the two-dimensional disc D_r; $p \in [0, r]$ on the radial set $\{q = \text{const.}\}$ and $q \in [0, 2\pi\rho)$ on the circle $\{p = \rho\}$. Then a

;. 8 5 AA6

in his paper [91], also formulated problems concerning the projection of subsets of the sphere onto the Euclidean plane in terms of partial differential equations. Likewise, Lagrange used the notion of conformal map in his papers on cartography [88], and the same notion is inherent in Gauss's work. The terminology "isothermal coordinates" which the latter introduced, referring to a locally conformal map between a subdomain of the plane and a subdomain of the surface, indicates the relation with physics. Riemann, in his dissertation, refers to an 1822 paper by Gauss, published in the *Astronomische Abhandlungen* in 1825 (but written several years before).[24] The title of that paper is *Allgemeine Auflösung der Aufgabe: die Theile einer gegebnen Fläche auf einer andern gegebnen Fläche so abzubilden, daß die Abbildung dem Abgebildeten in den kleinsten Theilen ähnlich wird* (General solution of the problem: to represent the parts of a given surface on another so that the smallest parts of the representation shall be similar to the corresponding parts of the surface represented), a paper presented to a prize question proposed by the Royal Society of Sciences at Copenhagen, [61]. In this paper, Gauss shows that every sufficiently small neighborhood of a point in an arbitrary real-analytic surface can be mapped conformally onto a subset of the plane.[25]

After recalling the definition of a conformal map, Riemann passes to the equivalent condition expressed in terms of partial differential equations. Here, we are given a function f of a complex variable which is composed of two functions u and v of two real variables x and y:

$$f(x + iy) = u + iv.$$

The functions u and v are differentiable and satisfy the Cauchy–Riemann equations. They appear as *potentials* in the space of the two variables x and y. Klein, in his article on *Riemann and his significance for the development of modern mathematics* writes ([86] p. 168):

> Riemann's method can be briefly characterized by saying that *he applies to these parts u and v the principles of the theory of the potential*. In other words, his starting point lies in the domain of mathematical physics.

In the same article (p. 170), after explaining some of Riemann's tools, Klein adds:

> All these new tools and methods, created by Riemann for the purpose of pure mathematics out of the physical intuition, have again proved of the greatest value for mathematical physics. Thus, for instance, we now always make use of Riemann's methods in treating the *stationary* flow of a fluid within a two-dimensional region. A whole series of most interesting problems, formerly regarded as insolvable, had thus been solved completely. One of the best known problems of this kind is Helmholtz's determination of the shape of a free liquid jet.

---

[24] The paper won a prize for a question proposed by the Copenhagen Royal Society of Sciences in 1822. The subject of the competition was: "To represent the parts of a given surface onto another surface in such a way that the representation is similar to the original in its infinitesimal parts." A letter from Gauss to Schumacher dated July 5, 1816 shows that the solution was already known to Gauss at that time; cf. Gauss's *Werke* Vol. 8, p. 371.

[25] Gauss did not solve the problem of mapping conformally an arbitrary finite portion of the surface; this was one of the questions considered by Riemann. An English translation of Gauss's paper is published in [153], Volume 3.

- 7 (1 7 (3 9 4= 3 * *1 *
H 3 J

the Habilitationsvortrag in 1919, [164], together with a commentary, making the link with relativity theory. One of the main features of the local geometry conceived by Riemann is that it is well suited to the study of gravity and more general fields in physics. Relativity theory, which encompasses the largest part of modern physics, relies in a crucial way on the notions introduced by Riemann.

From the purely mathematical point of view, the most important contribution of the Habilitationsvortrag is that it sets the bases of what we call today Riemannian geometry, with the introduction of the curvature tensor and its consequences, including several results such as the fact that the homogeneity (with the inherent notion of transformation group) corresponds to constant curvature. This new geometry can be considered as a far-reaching generalization of Gauss's work on the intrinsic geometry of surfaces, and at the same time it is a generalization of the non-Euclidean geometry (of constant curvature) discovered by Lobachevsky, Bolyai and Gauss, in the few decades that preceded Riemann. Furthermore, Riemann sets in this memoir the bases of several developments made in several directions by Clifford, Christoffel, Bianchi, Ricci, Beltrami, Levi-Civita, Élie Cartan, Einstein and many others.

It is known that the full importance of the Habilitationsvortrag was not recognized in the first years after it was delivered. A report by Dedekind on the mathematical content of the memoir was published only in 1868.[44] But it is also known that Gauss, who, as Riemann's mentor, was present at the lecture, expressed his complete satisfaction with it. This reaction to Riemann's Habilitationsvortrag is described in Dedekind's biography of Riemann published in the Collected Works [141]. Gauss's praise was certainly a cause for Riemann's own contentedness, because Gauss was known to be sparing with compliments.

In the introduction to his dissertation, Riemann declares that he is unexperienced in philosophical questions, and that in preparing the lecture he could rely only on some remarks that Gauss made in his second paper on biquadratic residues and in his Jubilee-book, and some philosophical researches of Herbart. Riemann discusses the question of space and that of manifoldness and its specialization. This specialization may be either continuous or discrete. In the chapter [119] contained in the present volume, Plotnitsly emphasizes that Riemann speaks of discrete manifolds, and then says that, rather than space itself, it is "the reality underlying space" that may be discrete.[45] "Manifolds" as we intend them today are particular cases of manifoldnesses. Riemann writes:

> Manifoldnesses in which, as in the plane and in space, the line-element may be reduced to the form $\sqrt{\sum dx^2}$, are therefore only a particular case of the manifoldnesses to be here investigated; they require a special name, and therefore these manifoldnesses in which the square of the line-element may be expressed as the sum of the squares of complete differentials I will call *flat*.

---

[44] *Abhandlungen der Königlichen Gesellschaft der Wissenschaften zu Göttingen, 13.*

[45] It may be useful to note that in modern physics, spacetime is studied in its both characters, discrete (e.g., in lattice gauge theories, which are often considered as mathematical discrete approximations) or as continuous (e.g., in general relativity).

;. 5 :6 $. , * 7 3 3 * * * * . . * 4* +* =
(. , * 0 * 3 * * * . * 4*
@ 3 * . , 5 * &* 6 3 & 3 7, S 7* & (,0 ' ,* * 3 +* 0
3 *3 * 1 3 3) *

Riemann declares that notions whose specializations form a continuous manifoldness are the positions of perceived objects (die Orte der Sinnengegenstände) and colors. It is conceivable that Riemann, in his mention of colors, refers to the fact that one can continuously move from a color to another one, a color being characterized by the proportions of red, green and violet it contains. This makes color a three-dimensional manifoldness. Weyl writes, in his *Space, time, matter*, that Riemann's reference to color "is confirmed by Maxwell's familiar construction of the color triangle" ([165], p. 84 of the English translation). There are also writings of Helmholtz and Thomas Young on this matter. There are particular portions of a manifoldness called *quanta*,[46] whose nature is different from that which is characterized by the discrete and the continuous. "Their comparison with regard to quantity is accomplished in the case of discrete magnitudes by counting, in the case of continuous by measuring." The notions of measurement and of dimension are discussed in the Aristotelean style. The relation between measurement and the axioms of geometry is said to be fundamental, and in some sense this question concerns the relation between the axioms of geometry and the reality of space, that is, between mathematical and empirical truth. Riemann writes:

> Either therefore the reality which underlies space must form a discrete manifoldness, or we must seek the ground of its metric relations outside it, in binding forces which act upon it.[47]

It is important to recall Riemann's words. He declares that geometry depends at the same time on axioms and on observational and experimental physics. He considers that classical geometry, with the first principles and axioms that it assumes and the connections between them, does not lead anywhere, because "we do not perceive the necessity of these connections." What is missing is a notion of "multiply-extended magnitude" (mehrfach ausgedehnte Grösse), a notion which makes space a particular "triply-extended magnitude." He proposes that the properties that distinguish space from other conceivable triply-extended magnitudes be deduced from experience. In particular, the space that Riemann talks about, although built from undefined notions and axioms connecting them, is not the space of traditional geometry. He suggests that this space should reflect the material world around us. He formulates the problem of finding the "simplest matters of fact from which the metric relations (Massverhältnisse) of space may be determined." He declares that these matters of fact are "not of necessity, but only of empirical certainty." They are the "hypotheses" that are referred to in the title of the dissertation. He says that he will investigate the "probability" of these matters of fact, "within the limits of observation," and see whether they may be extended "beyond the limits of observation, both on the side of the infinitely great and of the infinitely small." The most important among these matters of fact is related to the work of Euclid. The geometry that Riemann will

---

[46] In Aristotle's language, the latin word *quantum* is used as the translation of the word *ποσόν* (quantity, or "quantified thing"), which is one of Aristotle's categories. In his *Metaphysics*, Aristotle mentions four types of change: of substance, *quale*, *quantum*, or place. (Metaphysics, 1069b9–13).

[47] Es muss also entweder das dem Raume zu Grunde liegende Wirkliche eine discrete Mannigfaltigkeit bilden, oder der Grund der Massverhältnisse ausserhalb, in darauf wirkenden bindenden Kräften, gesucht werden.

F &*7 *) * 0 0 * *
* , & * * 0 * 7 . * * 4 3
7, 1 1 3 0 . * 3 & * $ *(*
3 * * 3 < , 0 1* 0 0 * 0 *) * =

G C 3 7, *2 = * *
* . I 5> 8 ?0 1 : @ . * * 6

construct will be Euclidean at the infinitesimal level.[48] We already noted by the way that the notion of "infinitely small" is treated in several works of the Ancient Greeks. The same notion is thoroughly discussed by Galileo Galilei in his *Discorsi; First day*, to whom Riemann refers in his habilitation, though in a different context.

The rest of the Habilitationsvortrag is a development of the ideas expressed in the introduction. The first part concerns the notion of n-dimensional Mannigfaltigkeit. This term is sometimes translated into English by "manifoldness." Riemann also talks about a "multiply extended magnitude." This is an ancestor to the mathematical notion of manifold. But the meaning of Mannigfaltigkeit, in Riemann's terminology, and that of the mathematical notion of manifold, as it is used today, do not coincide, even though in German the word Mannigfaltigkeit is used for "manifold."[49] There are discrete and continuous manifoldnesses, and there are manifoldnesses which are not mathematical. Riemann says that notions with specializations to discrete manifoldness are very common, but that, by contrast, there are very few notions whose specialization form a continuous manifoldness. We note incidentally that Poincaré pondered this terminology. In a letter to the mathematician and historian of mathematics Gustav Eneström, dated November 19, 1883 (cf. [120] p. 143), he writes:

> I prefer the translation of *Mannigfaltigkeit* by multiplicity, because the two words have the same etymological meaning. The word set is more adapted to the *Mannigfaltigkeiten* considered by Mr. Cantor and which are discrete. It would be less adapted to those which I consider and which are discontinuous. What is the opinion of Mr. Mittag on this matter?[50]

Eneström responds, on November 23:

> Mr. Mittag-Leffler thinks that you may be right, and, consequently, one should prefer the word multiplicity.[51]

In fact, Poincaré used the word multiplicity in its French form ("multiplicité") to denote Riemann's moduli space.

In the treatise *Théorie des fonctions algébriques de deux variables indépendantes* (Theory of algebraic functions of two independent variables) by Picard and Simart

---

[48] It may be worth recalling that Lobachevsky, in his various works on non-Euclidean geometry that he started in the late 1820s, systematically checked that the formulae that obtained in his new geometry give, at the infinitesimal level, the Euclidean formulae. See e.g. [93] p. 31.

[49] Jost, in [147], tries to sort out this complex terminology. He writes on p. 29: "The English of Clifford may appear somewhat old-fashioned for a modern reader. For instance, he writes 'manifoldness' instead of the simpler modern translation 'manifold' of Riemann's term 'Mannigfaltigkeit.' But Riemann's German sounds likewise somewhat old-fashioned, and for that matter, 'manifoldness' is the more accurate translation of Riemann's term. In any case, for historical reasons, I have selected that translation here."

[50] Je préfère la traduction de *Mannigfaltigkeit* par multiplicité, car les deux mots ont même sens étymologique. Le mot ensemble convient bien aux *Mannigfaltigkeiten* envisagés par M. Cantor et qui sont discrètes; il conviendrait moins à celles que je considère et qui sont discontinues. Qu'en pense M. Mittag à ce sujet?

[51] M. Mittag-Leffler pense que vous pouvez avoir raison, et que, par conséquent, il faut préférer le mot multiplicité.

;. 8 5 A 6

Appel-Goursat-Fatou

Riemann surfaces are also thoroughly studied in the first volume of the treatise *Théorie des fonctions algébriques et de leurs intégrales et des transcendantes qui s'y rattachent* (Theory of algebraic functions and their integrals, and their related transcendentals) [6] by Appell, Goursat and Fatou, which appeared in 1929. In reality, the treatise is a revised edition, by Fatou, of the treatise [4] by Appell and Goursat. Fatou was at the same time a mathematician and an astronomer. In 1906, he defended a thesis entitled *Séries trigonométriques et séries de Taylor* (Trigonometric series and Taylor series), [29, 30], whose subject is Lebesgue's integration theory, which in some sense is a refinement of Riemann's integration theory (see Sect. 5 below). It is in this thesis that we find the famous Fatou Lemma (also called the Fatou-Lebesgue Lemma) on the comparison between the integral of a lower limit of positive measurable functions and the lower limit of their integrals. The lemma is a key element in the proof of the Dominated Convergence Theorem. In the same year, Fatou started his work on the iteration of rational maps of the plane, a work that was revived in the last two decades of the twentieth century by Sullivan, Thurston and others. Fatou also worked on the dynamics of transcendental functions.

The title of the first volume of the treatise by Appell, Goursat and Fatou is *Étude des fonctions analytiques sur une surface de Riemann* (Study of analytic functions on a Riemann surface) [6]. In that treatise, Riemann surfaces are still represented, like in the 19th-century treatises, in an anthropomorphic fashion, (using Weyl's expression; see Sect. 2 of the present article). Figure 4 is extracted from that volume, and is already contained in the first edition by Appell and Goursat (Fig. 1 in Sect. 2 above). The authors declare, concerning the surface considered: "This surface is analogous to that represented in Fig. 10, with the difference that, in reality, the two leaves are infinitely close and the apertures are infinitely narrow.

Chapter III of this volume is entitled *Connexion des surfaces à deux feuillets. Périodicité des intégrales hyperelliptiques* (Connectivity of two-sheeted surfaces and periodicity of hyperelliptic integrals). The authors start by saying (p. 99):

> In what follows, we consider surfaces as leaves without thickness, in such a way that a point or a line drawn on that surface will be visible for a observer situated on one side or the other. These surfaces will be considered as *perfectly elastic and rip-stop*.[28]

Halphen

Among the other treatises that are related to Riemann surfaces, we mention Halphen's *Traité des fonctions elliptiques et de leurs applications* (A treatise on elliptic fonctions and their applications) in 3 volumes, published in 1886, 1888 and 1891 [41].

---

[28] Dans ce qui suit, nous considérons des surfaces comme des feuillets sans épaisseur, de sorte qu'un point ou une ligne tracée sur la surface seront visibles pour un observateur placé d'un côté ou de l'autre. Les surfaces seront en outre regardées comme *parfaitement élastiques et indéchirables*.

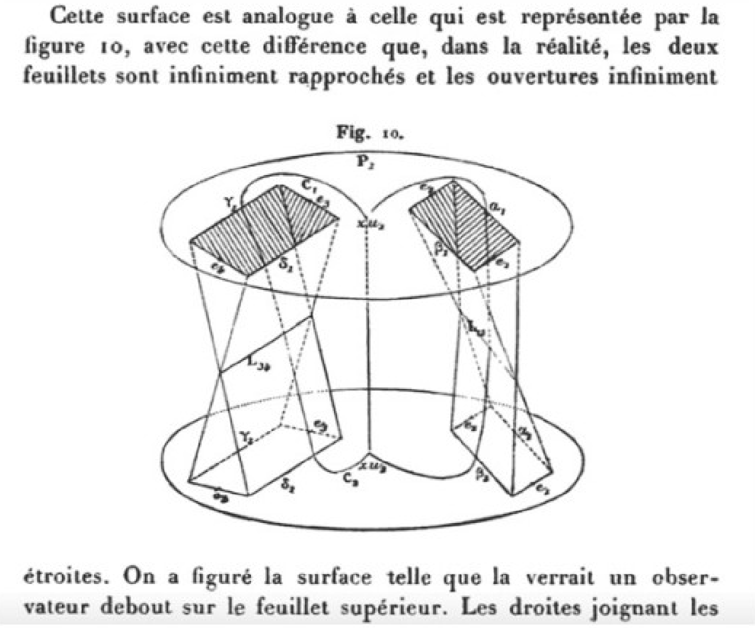

Cette surface est analogue à celle qui est représentée par la figure 10, avec cette différence que, dans la réalité, les deux feuillets sont infiniment rapprochés et les ouvertures infiniment

Fig. 10.

P₂
C₁
γ₁
δ₁
α₁
β₁
xu₂
γ₂
δ₂
C₂
β₂

étroites. On a figuré la surface telle que la verrait un observateur debout sur le feuillet supérieur. Les droites joignant les

Fig. 4 A picture from the treatise by Appell, Goursat and Fatou [6]. One can read in the text above the figure: "This surface is analogous to the one presented in Fig. 10, with the difference that, in reality, the two sheets are infinitely close, etc.," and, below the figure: "We have represented the surface in the way an observer standing on the upper sheet would see it."

This treatise had a certain impact on students in algebra and analysis. Georges-Henri Halphen, was a graduate of the École Polytechnique,[29] and he started with a career in the army. He submitted a doctoral dissertation on 1878, titled *Sur les invariants différentiels* (On differential invariants) [40], in which he determined the invariants of planar or skew curves under projective transformations. His thesis committee consisted of Hermite, Bouquet and Darboux. Haplhen participated to the 1870 French-German war. In 1872, he was appointed *répétiteur*[30] at the École Polytechnique. He was a specialist, among other things, of differential invariants, elimination theory, and singularities of algebraic curves. Picard, in biography of Halphen [82], writes the following (p. x of the Introduction):

> Riemann, in his theory of Abelian functions, had introduced the major notion of genus of elliptic curves, and he classified them into different classes, two curves being in the same class whenever there is a uniform correspondence between them. The famous geometer, who liked the great horizons, passed quickly on more than one difficult point, in particular, for what concerns higher singularities. Halphen gave a general formula, which applies to all cases, for the determination of the genus of an algebraic curve. Then, passing to the study of curves belonging to the same class, he went deeper into a remarkable proposition of

---

[29] We remind the reader that the École Polytechnique is a military school.

[30] A kind of a teaching assistant.

> Mr. Nœther according to which one may find in every class curves that have only ordinary singularities [...][31]

The first part of Halphen's treatise concerns the general theory of elliptic functions. The second part makes this treatise special compared to the other treatises on the same subject: it concerns the applications of elliptic functions to various branches of mathematics and physics. The subtitle of that volume is *Applications à la mécanique, à la physique, à la géodésie, à la géométrie et au calcul intégral* (Applications to mechanics, physics, geodesy, geometry and integral calculus). It was known since the eighteenth century, that is, since the birth of the theory of elliptic functions, that these functions have many applications in physics. It suffices to recall in this respect that these functions are in some sense generalizations of the familiar trigonometric functions, and that they can be used to represent a large class of periodic phenomena. For instance, whereas the small oscillations of a pendulum are represented by the sine functions (which is the inverse function of the elliptic integral $\int_0^x \frac{dt}{\sqrt{1-t^2}}$), for large oscillations, one needs (inverses of) more general elliptic integrals. By the time of Riemann, elliptic integrals were used in problems of gravitation and electromagnetism. We recall in this respect that the famous treatise of Legendre, *Exercices de calcul intégral* (Exercises of integral calculus) [63] contains a substantial part on elliptic integrals and their applications to problems in geometry and mechanics. We also note that the subtitle of the first volume of Legendre's *Traité des fonctions elliptiques et des intégrales eulériennes* (Treatise on elliptic functions and Eulerian integrals) [64] is: *Contenant la théorie des fonctions elliptiques et son application à différents problèmes de géométrie et de mécanique* (Containing the theory of elliptic functions and its application to various problems of geometry and mechanics). One may also mention in this respect that expressions of the lengths of arcs of an ellipse (which are precisely given by elliptic integrals) are obviously useful in celestial mechanics, since Kepler's first law says that orbits of planets in the solar system are ellipses with the Sun at one of their two foci. His second law says that a segment joining a planet and the Sun sweeps out equal areas during equal intervals of time. We also recall that Gauss was also an astronomer, and his interest in elliptic functions was motivated by his work on the trajectories of planets. Finally, Abel's 1827 famous paper on elliptic functions that we already mentioned, starts by mentioning the "beautiful properties" of Abelian functions "and their applications." He writes ([1] p. 101):

> Since a long time, the logarithmic functions, and the exponential and circular functions were the only transcendental functions that attracted the attention of the geometers. It is only in

---

[31] Riemann, dans sa théorie des fonctions abéliennes, avait introduit la notion capitale du genre des courbes algébriques, et partagé celles-ci en différentes classes, deux courbes étant de la même classe quand elles se correspondent uniformément. L'illustre géomètre, qui aimait les grands horizons, avait peu insisté sur plus d'un point difficile, en particulier sur ce qui concerne les singularités élevées. Halphen donne une formule générale, applicable à tous les cas, pour la détermination du genre d'une courbe algébrique; puis, passant à l'étude des courbes d'une même classe, il approfondit une proposition remarquable donnée par M. Noether, d'après laquelle on peut trouver dans toute classe des courbes n'ayant que des singularités ordinaires [...].

> recent times that some other functions started to be considered. Among them one has to distinguish the so-called elliptic functions, at the same time because of their beauty and of their use in the various branches of mathematics."[32]

The applications to geodesy mentioned by Halphen concern the geodesics on an ellipsoid of revolution whose ratio of major to minor axis is close to 1. Such a body is a representation of the shape of the Earth. It is also well known that Gauss was highly interested in geodesy. The applications of elliptic functions to geodesy were also considered by Jacobi in his paper [53]. In that paper, Jacobi solves a problem in geodesy which was addressed by Gauss. More details on elliptic functions are given in Chap. 1 of the present volume [76].

The third volume of Halphen's treatise contains fragments on elliptic functions which were collected after Halphen's death and published by Stieltjes.[33] The volume also contains Picard's biography of Halphen [82] which we already quoted. Picard declares there that Halphen was "one of the most eminent geometers in Europe."

Tannery and Molk

We now review the 4-volume treatise *Éléments de la théorie des fonctions elliptiques* (Elements of the theory of elliptic functions) [107] by Tannery and Molk. A few words on the authors are in order.

Jules Tannery (1848–1910) was a geometer, philosopher and writer. He edited the correspondence between Lagrange and d'Alembert.

In 1874, Tannery defended a doctoral dissertation whose title is *Propriétés des intégrales des équations différentielles linéaires à coefficients variables* (Properties of the integrals of linear differential equations with variable coefficients) [105] and [106]. The thesis committee consisted of Hermite, Briot and Bouquet. The dissertation starts with the following:

> The study of functions of an imaginary variable defined by an equation, a study which was substituted to the research, often unworkable, of the explicit form of these functions, profoundly renewed analysis in this century. It is well known that the glory of having shown this new way goes to Cauchy. The works of Mr. Puiseux on the solutions of algebraic equations, those of Messrs. Briot and Bouquet on doubly periodic functions and on differential equations, have largely proved the fertility of the idea of Cauchy in France. In Germany, the

---

[32] Depuis longtemps les fonctions logarithmiques, et les fonctions exponentielles et circulaires ont été les seules fonctions transcendantes qui ont attiré l'attention des géomètres. Ce n'est que dans les derniers temps qu'on a commencé à en considérer quelques autres. Parmi celles-ci il faut distinguer les fonctions, nommées elliptiques, tant pour leurs belles propriétés analytiques que pour leur application dans les diverses branches des mathématiques.

[33] Thomas Johannes Stieltjes (1856–1894) was Dutch but he decided to live in France. He acquired the French citizenship and in 1886 he became professor at the Faculté des Sciences de Toulouse. Stieltjes is known for several works on analysis and number theory, in particular on the so-called Stieltjes integral, elliptic functions, Dirichlet series, and is considered as the founder of the analytic theory of continued fractions. Stieltjes is also remembered for a failed attempt to prove the Riemann hypothesis, which he announced in his paper [104].

$ *11 * . 3 , 3 7, 9* 1 . 3 *
1 3 & * *J *2 * 7 3,
* 1 * *1 @* F * (* +* *
. , 3 . 3 ,

;. :586

Planes of reference give rise to these formations, which define the disciplinary nature of mathematics and science, and essentially shape mathematical and scientific thinking and practice—essentially, but, I would argue, not completely or even most centrally, at least in creative mathematics and science. One might say that creative or, to adopt Thomas Kuhn's language, revolutionary, thought in philosophy and mathematics or science alike is defined by planes of immanence and creation of concepts in Deleuze and Guattari's sense, which always carry individual signatures underneath them [8, p. 50]. Just as there are Plato's ideas, Descartes's cogito, and Leibniz's monads, there are Gauss's curvature, Riemann's manifolds, Dedekind's ideals, and Grothendieck's topoi in mathematics, or Einstein's spaces curved by gravity in general relativity and Heisenberg's matrix variables or Dirac's spinors in quantum physics. This is true even though the functioning of mathematical and scientific concepts does require planes of reference, functions (or other formal mathematical entities), propositions, and so forth.

The difference between philosophy and mathematics or science may instead be, to stay with Kuhn's idiom, in the nature of *normal*, rather than *revolutionary*, practice in each domain. In philosophy, the normal practice consists primarily in understanding, interpreting, and commenting on concepts, while in mathematics and science, the normal practice consists primarily in creating, by means of planes of references, frames of reference, functions and other mathematical or scientific formations, propositions, and so forth. It is true that for Deleuze and Guattari creative, revolutionary philosophical practice is the only true philosophy. However, leaving aside an arguably too restrictive character of this view of philosophy, this is not in conflict with the view of creative mathematics or science advocated here.

Deleuze and Guattari do allow that creative mathematical and scientific thought, such as that of Riemann (one of their primary examples), could have philosophical or artistic, compositional, aspects. But they appear to associate these aspects with *philosophical* or, in Deleuze's language, *inexact* (but philosophically rigorous) thought within mathematical or scientific thought. This philosophical thought is either operative alongside mathematical and scientific thought or enters by way of interference (in the positive sense of interfering wave fronts rather than in the more negative sense of inhibition) between mathematics and philosophy [8, pp. 217–218]. Both types of association are pertinent and important, certainly in Riemann's case. My argument is different, however. I argue that creative *technical*, *exact* mathematical and scientific thought is defined by planes of immanence and multi-component mathematical or scientific concepts, the architecture of which is analogous to that of philosophical concepts in Deleuze and Guattari's sense. That is, even apart from their philosophical strata, *mathematical* and *scientific* planes of immanence and the nature of mathematical or scientific concepts are *analogous* to philosophical thought. It is not only a matter of mathematical and scientific thought becoming philosophical at certain junctures, but a matter of the mathematical and philosophical thought creating parallel homomorphic (partially corresponding to each other), although not isomorphic (fully corresponding to each other), architectures of mathematical and philosophical concepts. They are not isomorphic because of technical, exact, aspects of

T * * * = 3 * 0 " * = . 0 * 3 S 7 4= *3 0 * +* =
&* 0) * = * 3 0 " 3 (3= 3 * 0* 3 + 3 (= 1 * * 0
@ = 1* & 3 7, . *& , . * * & , * 3 9 7 .= * 2 &* *7

mathematical and scientific concepts, demanded by the disciplinary nature of mathematics and science, aspects, generally, not found in philosophical concepts.

To bring this point home, I need to say more about Deleuze and Guattari's conceptions of planes of immanence and reference. The plane of immanence, as the plane of the movement of thought, is not "a concept that is or can be thought," but is "the image of thought, the image that thought gives itself of what it means to think" [8, p. 37]. As they say: "Concepts are like multiple waves, rising and falling, but the plane of immanence is the single wave that roles them up and unrolls them" [8, p. 36]. The present argument aims to extend, rather than to juxtapose, the plane of immanence (and the relationship between it and concepts) to mathematical and scientific thought, and to join this plane with the plane of reference. For Deleuze and Guattari, mathematics or science "relinquishes the infinite, infinite speed [of thought], in order to a gain *a reference able to actualize the virtual* [of chaos]. [It] gives reference to the virtual, a reference that actualizes the virtual through functions [or other mathematical objects]" [8, p. 118; translation modified]. Thought's enactment of this process constitutes a plane of reference. Planes of reference do play a major, indeed irreducible, role in mathematical and scientific thinking, especially in the disciplinary functioning of mathematics and science. This, however, is not inconsistent with the view that planes of immanence and the creation of concepts are found in mathematics and science. Mathematical and scientific thought combines both planes (sometimes, as does philosophy, also adding planes of composition) and creates its concepts from this fusion. The processes of thought defined the plane of immanence and specifically the creation of concepts (in Deleuze and Guattari's sense) are equally found in mathematics and science, and, again, define creative thinking there most essentially, analogously to the way it happens in philosophy. Such mathematical planes of immanence and the concepts they give rise to may coexist and interact with philosophical planes of immanence and concepts, but they are not reducible to philosophical planes and concepts. This is because of the equally irreducible interaction between these mathematical planes and concepts with technical, exact aspects of mathematical and scientific thinking.

Thus, in his rethinking of spatiality and geometry, Riemann, not only laid out a philosophical plane of immanence that gives rise to philosophical concepts and architectures, as Deleuze and Guattari rightly argue [7, pp. 483–486] but also introduced *a new mathematical plane of immanence*, which gives rise to multicomponent mathematical concepts, alongside *a mathematical plane of reference*. Riemann's thought shaped the plane of immanence of modern mathematical thought arguably more than that of any other mathematician (although Newton, Gauss, and Galois before Riemann, and Poincaré and Hilbert after him offer some competition). This plane extends well beyond geometry. Riemann made major transformative contributions, especially of a conceptual nature, to many areas of modern mathematics: geometry, topology, analysis, algebra, and number theory–not the least by bringing these fields to bear on each other, and his contributions to physics or philosophy were part of this interactive thinking and practice.

The interactive heterogeneity of Riemann's practice is a crucial aspect of Riemann's mathematics, and it extends beyond mathematics and shapes its mathematical operation from this exterior. Thus, the plane of immanence of Riemann's thought also has a more strictly philosophical dimension and thus creates more strictly philosophical concepts, emphasized by Deleuze and Guattari, rather than only mathematical concepts analogous to philosophical concepts by virtue of their multi-component architecture, stressed here. Riemann's philosophical thought was uncommonly and even nearly uniquely significant for his mathematical or physical thought, although one could think of a few competing cases, such as Weyl, who might have been inspired by Riemann in this respect as well, as he was by many other aspects of Riemann's thought. So were the physical dimensions of Riemann's thought, and this, too, is shared by Weyl, or, earlier Gauss, although this is more common. This role of physics in their work is also essentially connected to the role of philosophy there, which is, again, uncommon, if one speaks of such essential connections, making one's philosophical thinking a constitutive part of one's mathematical thinking, rather than of general philosophical reflections concerning mathematics and science, on the part of mathematicians and scientists.[8] The creation of new mathematics in Riemann was enabled by all three dimensions—mathematical, physical, and philosophical—of Riemann's thought and concepts. Riemann lays out a new plane of immanence of mathematical thought by changing both how to think about geometry or space and how to pursue thinking differently mathematically, via bringing together different mathematical fields and combining them with physical and philosophical concepts. A similar claim could also be made about his thinking concerning physics.

Riemann's Habilitation lecture is a magnificent unfolding of this plane (a geometrical metaphor of "plane" is fitting here). It is a major contribution not only to mathematics but also to physics and philosophy, especially to the philosophy of mathematics and physics, but far from exclusively so, as, for example, Deleuze and Guattari's use of Riemann's concepts shows. It is difficult to overestimate the significance and impact of Riemann's thinking concerning spatiality and geometry in mathematics and physics, in shaping the planes of immanence of thought of both. In mathematics the list of even major areas of impact is nearly inexhaustible, and I shall only mention a few of, arguably, the most important ones. First of all, Riemann's rethinking of geometry in terms of manifoldness crucially expanded the idea of the multiplicity of spaces and geometries themselves. Riemann's view of geometry as, in Deleuze and Guattari's language, topology and typology of manifolds led from the late-nineteenth century on to the extraordinary (a still ongoing) progress of geometry, beginning with the work of Sophus Lie and Felix Klein, and a bit later Élie Cartan [7, p. 483]. This work also connected differential geometry of manifolds and the theory of groups, specifically Lie groups. These connections eventually proved to have a major significance for quantum theory, especially in the theory of elementary particles, which are classified by using Lie groups as symmetry groups, although in connection with the infinite-dimensional spaces. The concept of manifold was also crucial for the development of topology. Initially, it was Riemann's earlier work on Riemann surfaces that had a greater impact. Eventually, topology came to be defined by understanding, which, as I shall explain, follows Riemann, of topological spaces as composed of local neighborhoods of points and (open) subspaces of a given space. It is true that Riemann did not have a concept of topological space. My main concern, however, is the development, transformation of Riemann's concepts (such as manifold) leading to new concepts, such as topological space, which makes Riemann's concepts alive, makes them "live on" in new concepts.

---

[8] Poincaré's extensive (much more extensive than Riemann's) philosophical works (e.g. [28, 29]), while influenced, as were his mathematical works in geometry, by Riemann's Habilitation lecture, may be seen along these lines. I realize that this claim may be challenged, and make it with caution. I would, nevertheless, argue that Riemann's philosophical thinking plays a greater constitutive role in his mathematical thinking than Poincaré's philosophical thinking in his mathematical thinking. The situation is of course different when it comes to physics, which is a major part of Riemann's and Poincaré's mathematical thinking alike, and both made major contributions to physics.

3 Mathematics: Space, Geometry, and the Concept of Manifold

The significance of Riemann's lecture for mathematics, physics, and philosophy, and its impact in all three fields were immense. This impact was delayed until its publication, in 1868, two years after Riemann's death, and fourteen years after it was presented in 1854, although some of Riemann's key ideas contained there and in Riemann's related works became known and had their impact earlier. One can only surmise (a tantalizing surmise!) the consequences for the history of mathematics and physics if the lecture was published more immediately. Riemann opens as follows:

> As is well known, geometry presupposes the concept of space, as well as assuming the basic principles for construction in space. It gives only nominal definitions of these things, while their essential specification appears in the form of axioms. The relationship between these presuppositions is left in the dark; we do not see whether, or to what extent, any connection between them is necessary, or a priori whether any connection between them is even possible.
>
> From Euclid to Lagrange this darkness has been dispelled neither by the mathematicians nor the philosophers who have concerned themselves with it. The reason [ground] [*Grund*] for this is undoubtedly because the general concept of multiply extended magnitudes [*Grösse*], which includes spatial magnitudes, remains completely unexplored. I have therefore first set myself the task of constructing the concept of a multiply extended magnitude from general notions of magnitude. It will be shown that a multiply extended magnitude is susceptible of various metric relations, so that space constitutes only a special case of a triply extended magnitude. From this, however, it is a necessary consequence that the theorems of geometry cannot be deduced from general notions of magnitude, but that those properties that distinguish space from other conceivable triply extended magnitudes can only be deduced from experience. Thus arises the problem of seeking out the simplest data from which the metric relations of space can be determined, a problem that by its very nature is not completely determined, for there may be several systems of simple data that suffice to determine the metric relations of space; for the present purposes, the most important system is that laid down as a foundation of geometry by Euclid. These data are—like all data—not logically necessary, but only of empirical certainty, they are hypotheses [*Hypothesen*]; one can therefore investigate their likelihood, which is certainly very great within the bounds of observation, and afterwards decide on the legitimacy of extending them beyond the bounds of observation, both in the direction of the immeasurably large [*Unmessbargrosse*] and in the direction of the immeasurably small [*Unmessbarkleinen*]. [31, p. 23; translation modified]

These introductory reflections are already profound and far-reaching, and Riemann develops the ideas suggested here quite a bit further in the lecture. First of all, Riemann does not appear to be interested in the axiomatic approach to geometry and is even suspicious of axioms of geometry. In contrast to most previous works on non-Euclidean geometry, the parallel postulate is not the starting point of his investigation. The reason is clearly that the concepts of space and of measuring distances in space are not adequately defined. This is what leaves "the relationships between [axioms] in the dark. We do not see whether, or to what extent, any connection between them is necessary, or a priori whether any connection between them is even possible." This questioning of the axioms of geometry was unusual at Riemann's time, and it was much deeper than customary doubts concerning the parallel postulate. The axiomatic approach to non-Euclidean geometry is, thus, abandoned and even implicitly questioned by Riemann, even though it was this approach that led to the discovery of non-Euclidean geometry by Lobachevsky and Bolyai. This may have been one of the reasons why only Gauss, rather than either Lobachevsky (highly regarded by Gauss) or Bolyai, was expressly invoked in the lecture.[9] It is true that Gauss was ultimately unable to establish a possible existence (in the sense of logical consistency) of non-Euclidean geometry, because, unlike Lobachevsky and Bolyai, he did not do this for the three-dimensional case. In spirit, however, Gauss's work on the geometry of two-dimensional surfaces and his concept of curvature as intrinsic to a given surface was much closer to Riemann. For Riemann, Gauss was pursuing a trajectory of thought better suited for the foundations of geometry and more fruitful for its conceptual development. In his famous extraordinary theorem (*theorema egrerium*, as he called it), Gauss proved that the curvature of a surface, which he defined as well, was intrinsic to the surface. More precisely, the theorem states that the Gaussian curvature of a surface does not change if one only bends the surface but does not stretch it. This means that curvature can be fully determined by measuring angles and distances on the surface itself, without considering the way it is embedded in the ambient (three-dimensional) Euclidean space or, in Riemann's terms, manifold, which makes the Gaussian curvature an intrinsic invariant of a given surface.

One nearly has here a two-dimensional conceptual architecture that suggests and even approaches that of Riemann, although major additional thinking is necessary to give this architecture the type of generality Riemann is able to do. Gauss's theorem suggests that one could see the surface as an independent curved manifold and then to generalize this concept to higher dimensions via "the concept of a multiply extended magnitude," mentioned in this passage and the concept of manifold, which is what Riemann did. These concepts also helped him to generalize to higher dimensions Gauss's concept of curvature. To do so required yet another new concept, another

---

[9] There were other, more extrinsic, reasons, beginning with the fact that Gauss was Riemann's mentor and the chair of his Habilitation committee. Indeed, Gauss selected this topic among three proposed by Riemann (following the rules). The philosopher R. H. Lotze, a fervent opponent of non-Euclidean geometry, was a member of the philosophy faculty, to which Riemann's Habilitation was presented. Later on, Lotze criticized Riemann's approach anyway, as part of his general critique of non-Euclidean geometry (see [16, pp. 222–226] and [33, pp. 97–112]).

F 1 0 & 0 +* = (. , 3 * * * 3 1 &* * * . &
* *) *
/) * 0 +* * 1 . * *J , . 7 3 3* . , * 3
1 * 3 & 1 F * 2 * 3 * , 5 * . 0 * * 3 60 +* 1 & 3 * &*
* * 0 3 C 3 * 0 * * /) * 0 +* * 1 . * *J , . 7

3 3* . ,*3 1*3& 1 F * 2*3*,
5 *. 0* *3 60+* 1&3 * &* * *0 3C3* 0*
*

great invention of Riemann, the tensor of curvature, and a new form of differential calculus, tensor calculus on manifolds, a generalization of differential calculus. This calculus was fully developed later on, and it played a major role in Einstein's general relativity. The independence of the curvature or of geometry of a given manifold of any dimension from its embedding also allows one to determine, at least in principle, intrinsically whether this space is flat or curved, which is crucial for determining the nature of the physical space we inhabit. For example, as indicated earlier, space is curved in the immediate vicinity of the solar system, because of the gravity of the Sun, planets, and other material entities in the solar system (or locally around other stars), but appears to be on the overage flat on the scale of the observable Universe.[10] In any event, by virtue of inventing his concepts and by using them, Riemann provided a general rigorous grounding to all geometry, Euclidean or non-Euclidean, rather than merely establishing, as Lobachevsky and Bolyai did, the logical possibility of the non-Euclidean geometry of (constant) negative curvature in three dimensions. The absence of a concept, such as that of manifold, leaves us in "the dark," because "we do not see whether, or to what extent, any connection between [axioms] is necessary, or a priori whether any connection between them is even possible."

Riemann's thinking is conceptual-problematic rather than axiomatic-theorematic: he grounds mathematics, as well as physics, in hypotheses and concepts, concepts-problems, such as the concept of manifold and its subconcepts (distance, curvature, tensor, and so forth). The problem of space and geometry is now posed in terms of finding concepts that are necessary to define space and to give it geometry. A general concept of manifold was assumed to be applicable to any possible space, while specific manifolds, flat or curved, define different spaces or subspaces. In this respect, as foundational thinking, that of Riemann is different from that conforming to Hilbert's concept of "foundations" [*Grundlagen*] in the sense of axiomatic foundations, an idea that Hilbert also tried to apply to physics. One of the problems of his famous 1900 list (Problem 6) was in fact that of the development of (mathematized) axiomatic foundations, system(s) of axioms, for physics, on the model of his own *Foundations of Geometry* [13].[11] That such a system is possible is inevitably a hypothesis. It became clear subsequently that the existence of such a system is a hypothesis even in mathematics, and in view of Gödel's incompleteness theorems in 1931, ultimately a wrong hypothesis, insofar as such a system cannot be proven to be free of contradiction, once it is large enough to include arithmetic, as geometry is.

---

[10] As I qualified earlier, at least this is a workable and widely accepted view, widely but not universally. It has never been established definitively or, in any event, agreed upon whether such a determination is ever rigorously possible, as opposed to having a practically effective and possibly, within its proper limits, the best available theory or, as Poincaré would have it, "convention," without making a real claim concerning "the reality underlying space" [31, p. 33]. Einstein had his doubts too, although he was ultimately inclined to accept the possibility of such a determination, at least in principle, as, it appears, was Riemann, but not Poincaré, with whose position Einstein, nevertheless, had to contend and which he tried to accommodate within his own (e.g. [11, pp. 324–328]).

[11] For the development of Hilbert's ideas, as reflected in different editions of the *Grundlagen*, see [4]. In the first version of the book, Hilbert was closer to Riemann, and he later returned to a more Riemannian view of geometry in the wake of general relativity to which he made important contributions.

;. 8 5: 6

3.1 Discrete Conformality of Polyhedral Surfaces

A closed surface S together with a non-empty finite subset of points $V \subset S$ will be called a *marked surface*. A triangulation $\mathcal{T}$ of a marked surface (S, V) is a topological triangulation of S such that the vertex set of $\mathcal{T}$ is V. We use $E = E(\mathcal{T})$, $V = V(\mathcal{T})$ to denote the sets of all edges and vertices in $\mathcal{T}$ respectively. A (Euclidean) *polyhedral metric* on (S, V), to be called a *PL metric* on (S, V) for simplicity, is a flat cone metric on (S, V) with cone points contained in V. We call the triple (S, V, d) a polyhedral surface. All PL metrics are obtained by isometric gluing of Euclidean triangles along pairs of edges. For instance boundaries of convex polytopes are PL metrics on $(\mathbb{S}^2, V)$. The *discrete curvature* of a PL metric d is the function $K_d : V \to (-\infty, 2\pi)$ sending a vertex v to 2π minus the cone angle at v. For a closed surface S, it is well known that the Gauss-Bonnet theorem $\sum_{v \in V} K_d(v) = 2\pi\chi(S)$ holds. If $\mathcal{T}$ is a triangulation of (S, V) with a PL metric d for which all edges in $\mathcal{T}$ are geodesic, we say $\mathcal{T}$ is *geometric* in d and d is a PL metric on $(S, V, \mathcal{T})$. In this case, we can represent the PL metric d by the length function $l_d : E(\mathcal{T}) \to \mathbb{R}_{>0}$ sending an edge to its length. Thus the polyhedral surface (S, V, d) can be represented by $(S, \mathcal{T}, l)$. This is a way of coding a polyhedral surface by a finite-dimensional vector $l_d \in \mathbb{R}^E$.

In general, a polyhedral surface (S, V, d) admits infinitely many different geometric triangulations. However, each polyhedral surface (S, V, d) has a natural *Delaunay triangulation* $\mathcal{T}_d$ which is a geometric triangulation with vertices V such that for each edge, the sum of two angles facing e is at most π. Delaunay triangulations are the most commonly used triangulations in scientific computing. It can be constructed from the Voronoi decomposition $\{R(v)|v \in V\}$ of (S, V, d) as follows. Here a Voronoi 2-cell $R(v)$ for $v \in V$ is defined to be $\{x \in S | d(x, v) \le d(x, v'), \forall v' \in V\}$. The Delaunay tessellation of (S, V, d) is the dual cell decomposition of $\{R(v)|v \in V\}$ whose vertices are V and each 1-dimensional connected component of $R(v) \cap R(v')$ corresponds to a (geodesic) edge from v to v'. A Delaunay triangulation is a subdivision of the Delaunay tessellation into triangles without introducing extra vertices. Any two Delaunay triangulations of (S, V, d) are related by a sequence of Delaunay triangulations such that adjacent ones differ by a diagonal switch along an edge. See for instance [4].

Suppose d is a PL metric on a triangulated surface $(S, \mathcal{T})$ whose edge length function is $l_d : E(\mathcal{T}) \to \mathbb{R}_{>0}$. For a positive function $u : V(\mathcal{T}) \to \mathbb{R}_{>0}$, the vertex scaling of l_d by u is the new function $u * l_d : E(\mathcal{T}) \to \mathbb{R}_{>0}$ such that $u * l_d(vv') = u(v)u(v')l_d(vv')$ where vv' is an edge with end points v, v'. If d, d' are two PL metrics on $(S, \mathcal{T})$, then they differ by a *vertex scaling* if $l_d = u * l_{d'}$ for some $u : V \to \mathbb{R}_{>0}$. The notation of vertex scaling change of PL metrics was introduced in [28] and in [16].

The definition of discrete conformality involves Delaunay triangulations and vertex scaling.

Definition 3.1 ([11]) Two PL metrics d and d' on a marked closed surface (S, V) are *discrete conformal* if there is a sequence of PL metrics $d_1 = d, d_2, ..., d_n = d'$ and a sequence of triangulations $\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_n$ of (S, V) such that

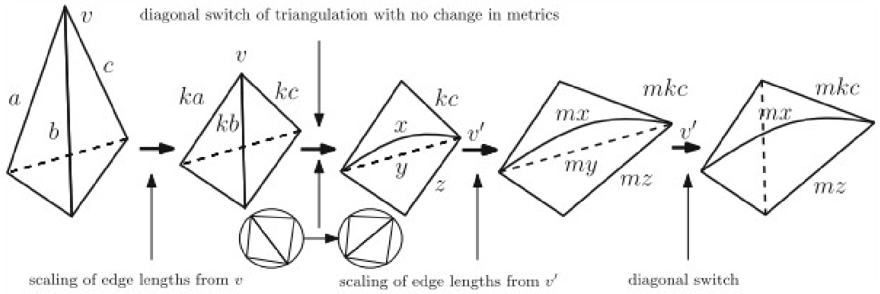
v
a
b
c
diagonal switch of triangulation with no change in metrics
v
ka
kb
kc
kc
x
y
z
v′
mx
my
mz
mkc
v′
mx
mkc
mz
scaling of edge lengths from v
scaling of edge lengths from v′
diagonal switch

Fig. 3 Discrete conformal change of PL metrics from an arbitrary tetrahedron to one with constant curvature π. All triangulations involved are Delaunay

(a) each $\mathcal{T}_i$ is Delaunay in d_i,

(b) if $\mathcal{T}_i \neq \mathcal{T}_{i+1}$, then there is an isometry h_i from (S, V, d_i) to (S, V, d_{i+1}) such that h_i is homotopic to the identity map on (S, V), and

(c) if $\mathcal{T}_i = \mathcal{T}_{i+1}$, there is a function $u_i : V \to \mathbb{R}_{>0}$ such that for each edge $e = vv'$ in $\mathcal{T}_i$, the lengths $l_{d_i}(vv')$ and $l_{d_{i+1}}(vv')$ of e in d_i and d_{i+1} are related by

$$l_{d_{i+1}}(vv') = u_i(v)u_i(v')l_{d_i}(vv'), \tag{3.1}$$

i.e., $l_{d_{i+1}} = u_i * l_{d_i}$.

The original motivation in [16] for introducing vertex scaling $u * l_d$ as an approximation to conformal change is the following. Since a PL polyhedral metric l_d on $(S, \mathcal{T})$ is a discretization of a Riemannian metric g and a function $u : V(\mathcal{T}) \to \mathbb{R}_{>0}$ is a discretization of a positive function λ on S, the conformal Riemannian metric λg should be approximated by the PL metric defined by $u * l_d$. The deeper reason for $u * l_d$ to be a discrete conformal change is due to the following observation in Riemannian geometry ([18]). Given a Riemannian metric g on a compact connected manifold M and $\lambda : M \to \mathbb{R}_{>0}$, there exists a constant $C = C(M, g, \lambda)$ such that for any $p, q \in M$, we have

$$|d_{\lambda^4 g}(p, q) - \lambda(p)\lambda(q)d_g(p, q)| \leq Cd_g(p, q)^3$$

where $d_g(p, q)$ is the distance between p, q in the metric g.

The relationship between discrete conformal geometry and hyperbolic geometry is the following [5, 11]. Given a Delaunay triangulated polyhedral surface $(S, \mathcal{T}, d)$ with $V = V(\mathcal{T})$, one can naturally associate to d a cusped hyperbolic metric d^* on $S - V$. Here is the construction. Take a Euclidean triangle τ in $(\mathcal{T}, d)$ considered as the Euclidean convex hull of vertices $v_1, v_2, v_3 \in \mathbb{C}$. Let τ^* be the hyperbolic convex hull $C_{\mathbb{H}}(v_1, v_2, v_3)$ of v_1, v_2, v_3 in the hyperbolic 3-space $\mathbb{H}^3$. Here we consider $\mathbb{C}$ to be in the sphere at infinity of the upper half-space model $\mathbb{C} \times \mathbb{R}_{>0}$ of $\mathbb{H}^3$. Now if σ and τ are two Euclidean triangles in $\mathcal{T}$ glued by a Euclidean isometry f along an edge, since each isometry f of $\mathbb{C}$ extends naturally to an isometry f^* of $\mathbb{H}^3$, we glue

τ^* and σ^* along the corresponding edge by the isometry f^*. In this way, we obtain a complete finite area hyperbolic metric d^* on $S - V$. It follows from the construction that d^* is independent of the choices of the Delaunay triangulations $\mathcal{T}$ used in the construction. It is proved in [11, Theorem 43] that two PL metrics d_1 and d_2 on a closed marked surface (S, V) are discrete conformal in the sense of Definition 3.1 if and only if their associated hyperbolic metrics d_1^* and d_2^* are isometric by an isometry homotopic to the identity (respecting V). Conversely, if S is a closed surface and $\hat{d}$ is a complete finite area hyperbolic metric on $S - V$, then there exists a polyhedral metric d on (S, V) such that $d^* = \hat{d}$. Thus for closed surfaces, there exists a bijection between the space of all discrete conformal classes of polyhedral metrics on (S, V) and the Teichmüller space of cusped metrics on $S - V$.

By this construction, if $\mathcal{T}$ is a Delaunay triangulation of the plane $(\mathbb{C}, d_{st})$ with $V = V(\mathcal{T})$ and d_{st} being the standard flat metric on $\mathbb{C}$, then the associated hyperbolic metric d_{st}^* is the boundary of the convex hull $C_{\mathbb{H}}(V)$ of V in $\mathbb{H}^3$. To see this, we note that codimension-1 faces of $C_{\mathbb{H}}(V)$ correspond to the circum-disks of triangles $\tau \in \mathcal{T}$ due to the Delaunay condition. This shows the relationship between discrete conformal geometry and convex hull construction in the hyperbolic 3-space $\mathbb{H}^3$ and the essential role of Delaunay condition in discrete conformality.

The main theorems proved in [11, 12] are:

Theorem 3.2 ([11]) *Given any PL metric d on a closed marked surface (S, V) and any $K^* : V \to (-\infty, 2\pi)$ such that $\sum_{v \in V} K^*(v) = 2\pi\chi(S)$, there exists a PL metric d^* on (S, V), unique up to scaling and isometries homotopic to the identity map on (S, V), such that*

(a) d^* is discrete conformal to d, and

(b) the discrete curvature K_{d^*} is equal to K^*.

Furthermore, the PL metric d^* can be found by a finite-dimensional variational principle.

For the constant function $K^* = 2\pi\chi(S)/|V|$ in Theorem 3.2, we obtain a constant curvature PL metric d^*, unique up to scaling and isometries homotopic to the identity, discrete conformal to d. We call d^* *the discrete uniformization metric* associated to d. The existence and uniqueness of d^* is a discrete version of the uniformization theorem for closed surfaces.

Theorem 3.2 for the torus $S = \mathbb{S}^1 \times \mathbb{S}^1$ with $K^* = 0$ is equivalent to a theorem of Fillastre [9]. Theorem 3.2 shows that every polyhedral torus $(\mathbb{S}^1 \times \mathbb{S}^1, V, d)$ is discrete conformal to a flat torus $(\mathbb{S}^1 \times \mathbb{S}^1, V, d_{flat})$. Translating it into the language of hyperbolic metrics, we can replace d by any cusped hyperbolic metric $\hat{d}$ on the punctured torus $\mathbb{S}^1 \times \mathbb{S}^1 - V$. The hyperbolic metric associated to $(\mathbb{S}^1 \times \mathbb{S}^1, V, d_{flat})$ is constructed as follows. Take a lattice $L = \mathbb{Z} + \tau\mathbb{Z}$ in $\mathbb{C}$ and consider the boundary $\partial C_{\mathbb{H}}(V^*)$ of the convex hull of V^* in $\mathbb{H}^3$ where V^* is a discrete set in $\mathbb{C}$ invariant under the action of L. Then by the discussion above, d_{flat}^* is isometric to the cusped hyperbolic metric $\partial C_{\mathbb{H}}(V^*)/L$. Furthermore, the lattice L is unique up to complex linear transformations. This is the result proved in [9]. To be more precise, Fillastre proved the following version of convex embedding theorem. For any cusped hyperbolic metric $\hat{d}$ on $\mathbb{S}^1 \times \mathbb{S}^1 - V$, there exist a lattice $L \subset \mathbb{C}$ and a finite set V' in the conformal infinite of the hyperbolic manifold $\mathbb{H}^3/L$ such that $\hat{d}$ is isometric to the boundary of the convex hull of V' in $\mathbb{H}^3/L$.

This shows a close connection between discrete conformal geometry and the convex isometric embedding program of Weyl, Alexandrov, Nirenberg, Pogorelov and others.

Theorem 3.3 ([12]) *Given two PL metrics on a closed marked surface (S, V) such that the lengths of edges are algebraic numbers, there exists an algorithm to decide if they are discrete conformal.*

Theorem 3.3 is proved in our joint work with Ren Guo in [12]. The counterpart of Theorem 3.2 for hyperbolic polyhedral surfaces is also proved in [12].

An important question is whether discrete conformality defined above approximates smooth conformality. To this end, let us recall discrete conformal maps between polyhedral surfaces [5, 18]. Given a closed polyhedral surface (S, V, d). Let d^* be the hyperbolic metric on $S - V$ associated d constructed using ideal hyperbolic triangles associated to Euclidean triangles. The vertical projection of the ideal hyperbolic triangle $\tau^* = C_{\mathbb{H}}(v_1, v_2, v_3)$ to the Euclidean triangle $\tau = C_{\mathbb{E}}(v_1, v_2, v_3)$ produces a piecewise projective homeomorphism Φ_d from $(S - V, d^*)$ to $(S - V, d|_{S-V})$. If d_1, d_2 are two discrete conformal PL metrics on (S, V), then the *discrete conformal map* from (S, V, d_1) to (S, V, d_2) is defined to be (the extension to S) of the composition $\Phi_{d_2} \circ \Psi \circ \Phi_{d_1}^{-1}$ where $\Psi : (S - V, d_1^*) \to (S - V, d_2^*)$ is the isometry homotopic to the identity. Discrete conformal maps are piecewise projective.

Our recent work with Sun and Wu [18] shows that discrete conformality does converge to the smooth conformality. Given a simply connected marked polygonal domain with a PL metric (D, V, d) and three boundary vertices $p, q, r \in V$, let the metric double of (D, V, d) along the boundary be the marked polyhedral 2-sphere $(\mathbb{S}^2, V', d')$. Using Theorem 3.2, one produces a new polyhedral surface $(\mathbb{S}^2, V', d^*)$ such that (1) $(\mathbb{S}^2, V', d^*)$ is discrete conformal to $(\mathbb{S}^2, V', d')$, (2) the discrete curvatures of d^* at p, q, r are $4\pi/3$, (3) the discrete curvatures of d^* at all other vertices are zero and (4) its area is $\sqrt{3}/2$. Thus $(\mathbb{S}^2, V', d^*)$ is the metric double of an equilateral triangle ΔABC of edge length 1. Here A, B, C correspond to p, q, r. Let $F : (\mathbb{S}^2, V', d') \to (\mathbb{S}^2, V', d^*)$ be the associated discrete conformal map sending $\{p, q, r\}$ to $\{A, B, C\}$ respectively. Due to the uniqueness part of Theorem 3.2, we see that $f = F|_D : D \to \Delta ABC$ sending p, q, r to the vertices A, B, C respectively. We call f the *discrete uniformization map* associated to $(D, \mathcal{T}, l, \{p, q, r\})$

Theorem 3.4 ([18]) *Suppose Ω is a Jordan domain in the complex plane with three distinct points p, q, r in the boundary of Ω. There exists a sequence of simply connected polygonal domains $(\Omega_n, \mathcal{T}_n, \{p_n, q_n, r_n\})$ with triangulations $\mathcal{T}_n$ by equilateral triangles of edge lengths converging to 0 where p_n, q_n, r_n are three boundary vertices such that the following hold*

(i) $\Omega_n \subset \Omega_{n+1}$ and $\Omega = \cup_{n=1}^{\infty} \Omega_n$,

(ii) the discrete uniformization maps f_n associated to $(\Omega_n, \mathcal{T}_n, \{p_n, q_n, r_n\})$ converge uniformly on compact sets to the Riemann mapping $f : (\Omega, \{p, q, r\}) \to (\Delta ABC, \{A, B, C\})$.

On the Positive Mass Theorem for Closed Riemannian Manifolds

Andreas Hermann and Emmanuel Humbert

Abstract The Positive Mass Conjecture for asymptotically flat Riemannian manifolds is a famous open problem in geometric analysis. In this article we consider a variant of this conjecture, namely the Positive Mass Conjecture for closed Riemannian manifolds. We explain why the two positive mass conjectures are equivalent. After that we explain our proof of the following result: If one can prove the Positive Mass Conjecture for one closed simply-connected non-spin manifold of dimension $n \geq 5$ then the Positive Mass Conjecture is true for all closed manifolds of dimension n.

1 Introduction

General Relativity is a geometrical theory: the background is a 4-Lorentzian manifold $(\mathcal{M}, G)$ whose metric satisfies:

$$\mathrm{Ric}_G - \frac{1}{2} s_G G = 8\pi T. \tag{1}$$

where Ric_G and s_G denote the Ricci curvature and the scalar curvature of the metric G respectively. The tensor T is the *energy momentum tensor* and contains all the information on the matter and Formula (1) explains how the physical objects (the tensor T) are related to the geometrical background (the curvature). After Gauss' work on the curvature of surfaces which resulted in his famous theorema egregium, it was Riemann who generalized the notion of curvature to higher dimensions by introducing a quantity which today is called the Riemann curvature tensor. In this way he founded the language of modern differential geometry which was later used

---

A. Hermann (✉)
Institut für Mathematik, Universität Potsdam, Karl-Liebknecht-Str. 24-25, 14476
Potsdam, Germany
e-mail: hermanna@uni-potsdam.de

E. Humbert
LMPT, Université de Tours, Parc de Grandmont, 37200 Tours, France
e-mail: emmanuel.humbert@lmpt.univ-tours.fr

© Springer International Publishing AG 2017

L. Ji et al. (eds.), *From Riemann to Differential Geometry and Relativity*,
DOI 10.1007/978-3-319-60039-0_17

by Einstein to formulate the field Eq. (1). Many physical laws can be read through this formula: in particular, it describes gravitation in terms of curvature which at the beginning of the 20th century was the real breakthrough of the theory. This formulation completely denies the existence of a global time which nevertheless has a strong meaning for each human being. One way to overcome this difficulty is to consider an appropriate space-type hypersurface M, which is a Riemannian manifold when equipped with the restriction of the metric G. This manifold M is called a *Cauchy hypersurface* and can be interpreted as a picture of the universe at some fixed time. General Relativity is a deterministic theory: specifying enough data on the chosen Cauchy hypersurface allows to reconstruct the whole spacetime: this is a result by Choquet-Bruhat and Geroch [7]. This is one reason for which many problems can be formulated in terms of Riemannian geometry.

In this survey, we present the Positive Mass Theorem which illustrates what is explained above: the physical question of whether a local positive energy density of an isolated system gives rise to a positive total energy can be formulated as a problem of pure Riemannian geometry. So it is not surprising that in this context, geometrical tools were crucial to (partially) answer this question. On the other hand it is really surprising that the Positive Mass Theorem was the key point in the final argument of R. Schoen for the solution of the Yamabe problem, which was a challenging question in Riemannian geometry and partial differential equations for more than 20 years and which has nothing to do with General Relativity.

In the paper, we will start by introducing the ADM mass of an asymptotically flat manifold and we will explain the reason why it can be interpreted as the total energy of an isolated system. The approach we choose in this text is inspired by the book of Vaugon and the second author [14] where the reader will find many more details. The ADM mass measures how an isolated system affects the trajectories of massless particles evolving far from the system. Note that in the chapter written by Jacques Franchi [8], it is explained how such a system also perturbs Brownian motions (see e.g. the study on Schwarzschild solution). A famous open problem is to know if, under natural physical assumptions, the ADM mass is non-negative together with a rigidity result (for many other interesting rigidity results, the reader may refer to the chapter written by Marc Mars [12]): it should vanish if and only if the spacetime does not contain any matter nor gravity fields and hence is flat.

We will formulate the Positive Mass Conjecture for asymptotically flat manifolds which has been solved in the context of General Relativity (that is, in dimension 3), but which is still considered as open in dimension greater or equal to 8 if the manifold is not spin. Then we will study the mass of closed manifolds, more precisely of manifolds which are compactifications of some asymptotically flat manifolds. As in the chapter written by Jean-Philippe Nicolas [13], the situation appears much clearer in the compact setting giving rise to new applications. We will explain how Schoen defined the mass of a closed Riemannian manifold which can be regarded as the ADM mass of a blown-up manifold. This will give rise to a Positive Mass Conjecture for closed manifolds which was the main point in the last argument for solving the Yamabe problem. It is well known that the Positive Mass Conjectures for closed and for asymptotically flat manifolds are actually equivalent: the proof of this equivalence is essentially contained in the work of Schoen, Lohkamp or Lee-Parker [16, 17, 19]. But, even if all the ingredients of the proof are in the literature, it is quite difficult to find a reference where this equivalence is explicitly stated and proved. We do it in this paper.

In the last paragraph we present our recent results which reduce the Positive Mass Conjecture to proving that it holds for every metric on one fixed closed non spin and simply connected manifold in each dimension.

2 ADM Mass in General Relativity

The goal of this paragraph is to explain how we can define the energy of an isolated system in General Relativity. We will not be very precise in this section: our goal is mainly to explain the physical considerations which lead to the definition of the ADM Mass. For more information on this subject, the reader may refer to [14].

2.1 Modeling Isolated Systems in General Relativity

An *isolated system* is a physical system contained in a "bounded region". Obviously, the notion of "bounded region" should be understood in the sense "bounded in space" and not in time. The general idea is that the gravitational field induced by the system decreases as the distance to the system increases and vanishes at infinity.

It is generally considered as natural that the spacetime $(\mathcal{M}, G)$ (a 4-Lorentzian manifold satisfying Eq. 1) is *globally hyperbolic*, i.e. possesses a *Cauchy hypersurface* M, which means that any inextensible causal curve of $\mathcal{M}$ meets M exactly once. The manifold $\mathcal{M}$ is then homeomorphic to $M \times \mathbb{R}$ and the hypersurfaces $M \times \{t\}$ can be interpreted as "pictures of the universe" at some time t as was proved by Geroch [9]. Bernal-Sánchez [6] were able to strengthen this result by showing that $\mathcal{M}$ is diffeomorphic to $M \times \mathbb{R}$. With this assumption, it is natural to impose that:

- the energy-momentum tensor T is compactly supported in each hypersurface $M \times \{t\}$ which means that the system is contained in a bounded region of space;
- the metric g becomes flat at infinity, in a sense to be made precise, which means that the gravitational field decreases far from the system.

The first assumption is in general too strong. For example it is not satisfied for the Reissner-Nordström metric. Actually, in all physical applications, we just need to consider the second assumption: it does not imply the first one but implies that the energy-momentum tensor T vanishes at infinity. A good way to take into account these considerations is to assume that the spacetime $(\mathcal{M}, G)$ is *asymptotically flat*: there exists a coordinate system (t, x, y, z) such that the hypersurfaces $t =$ constant correspond to Cauchy hypersurfaces and such that in these coordinates we have

On Local Characterization Results in Geometry and Gravitation

Marc Mars

Abstract An important problem in differential geometry and in gravitation is to identify metrics in a fully coordinate independent manner. In fact, the very foundation of Riemannian geometry is based on the existence of a tensor, the Riemann or curvature tensor, which vanishes if and only if the metric is locally flat. Many other such local characterizations of metrics are known. The aim of this article is to present a brief selection of them as an example of the fruitful interplay between differential geometry and gravity.

1 Introduction

Albert Einstein's great insight was that the gravitational interaction was a manifestation of the curvature of spacetime. With this radically new point of view, Einstein gave physical reality to non-euclidean geometries, introduced in Bernhard Riemann's inaugural lecture "Über die Hypothesen, welche der Geometrie zu Grunde liegen" in 1854 [47], about sixty years before Einstein's theory of general relativity was founded. It is most remarkable that Riemann himself, in the last paragraphs of his dissertation, had already realized that the grounds for the metric relations in physical space had to come from binding forces that act upon it, in a view that was already radically different from Newton's conception of space and time as a fixed and invariable entity completely unaffected by physical processes occurring within it. Building on Riemann's mathematical ideas, Einstein could transform his outstanding physical intuition into one of the most elegant physical theories of all times. One of the fundamental consequences of this theory is that differential geometry is not only a fundamental branch of mathematics but also a basic tool for doing physics, and the

---

Work partially supported by the projects FIS2012-30926, FIS2015-65140-P (Spanish MINECO/FEDER) and P09-FQM-4496 (Junta de Andalucía/FEDER).

M. Mars (✉)
Instituto de Física Fundamental y Matemáticas, IUFFyM, Universidad de Salamanca,
Plaza de la Merced s/n, 37008 Salamanca, Spain
e-mail: marc@usal.es

© Springer International Publishing AG 2017

L. Ji et al. (eds.), *From Riemann to Differential Geometry and Relativity*,
DOI 10.1007/978-3-319-60039-0_18

interplay between these two aspects has given a tremendous boost to this scientific discipline.

One of the basic questions that differential geometry has to address is: when are two Riemannian manifolds[1] isometric? Obviously this is also a fundamental question in general relativity, as it addresses the question: when two spacetimes are physically equivalent? This question can be stated, and it is relevant, also at the local level: When are two Riemannian manifolds locally isometric? In fact, answering this question in the case of Euclidean space led Riemann to the introduction of his curvature tensor and to the statement that the vanishing of the curvature at every point is the necessary and sufficient condition for the metric being (locally) flat. The vanishing of the Riemann curvature tensor as a local characterization of Euclidean space is one of the most fundamental results in Riemannian geometry. A natural question is whether similar local characterizations for other geometries also exist. A satisfactory local characterization of a geometry is a very useful tool both in geometry and in general relativity (or any other geometric theory of gravitation). Many results along those lines are known, some coming from pure geometry and some from gravitational theory. The purpose of this chapter is to review a selection of such characterizations as an example of the fruitful interplay between differential geometry and gravity. It must be emphasized that I have no intention of presenting a comprehensive review of known results in this area of research. This would require much more space and a more knowledgeable author. The selection is based solely on my own taste, so many interesting local characterizations will be left out, for which I apologize in advance.

2 Classical Characterizations

Local characterizations involve the notion of local isometry, in the following standard sense. Two Riemannian manifolds $(\mathcal{M}_1, g_1)$ and $(\mathcal{M}_2, g_2)$ are locally isometric if given any point $p \in \mathcal{M}_1$, there exist a connected, open neighbourhood $\mathcal{U}_p$ of p and a smooth map $\Phi_p : (\mathcal{U}_p, g_1|_{\mathcal{U}_p}) \longrightarrow (\mathcal{M}_2, g_2)$ which is an isometry onto its image. Often it is necessary to extend this definition to a situation when $(\mathcal{M}_2, g_2)$ is not a single space but belongs to a class $\{(\mathcal{M}_2^{\mathfrak{a}}, g_2^{\mathfrak{a}})\}$. In this case, the target of the local isometry Φ_p is allowed to depend on p. All manifolds considered here are finite dimensional, smooth, Hausdorff and connected.

As already mentioned, the most important local characterization of a geometry is that of Euclidean space or, more generally, of flat space of signature (p, q), $\mathbb{M}^{p,q} := (\mathbb{R}^{p+q}, \eta^{(p,q)} = -\sum_{i=1}^{p}(dx^i)^2 + \sum_{i=p+1}^{p+q}(dx^i)^2)$. Depending on the point of view, the Riemann tensor can be defined in several ways. The modern approach is to define the curvature operator as

---

[1] My convention here is that a Riemannian metric is not necessarily positive definite, but can have any non-degenerate signature. The term *strictly Riemannian* is reserved for the positive definite case.

$$\mathrm{Curv}^g(X,Y)Z = \left(\nabla_X\nabla_Y - \nabla_Y\nabla_X - \nabla_{[X,Y]}\right)Z, \qquad X,Y,Z \in \mathfrak{X}(\mathcal{M})$$

where ∇ is the metric-compatible, torsion-free connection associated to g

$$\mathrm{Riem}^g(W,Z,X,Y) = \langle W, \mathrm{Curv}^g(X,Y)Z\rangle, \qquad W,X,Y,Z \in \mathfrak{X}(\mathcal{M})$$

where $\langle\, ,\, \rangle$ denotes scalar product with g. In terms of this tensor Riemann's characterization of flat space is

Theorem 2.1 (Local characterization of flat space, Riemann [47]) *A Riemannian manifold $(\mathcal{M}, g)$ is locally isometric to the flat space with the same dimension and signature if and only if $Riem^g = 0$.*

Being intended to all faculty members, Riemann's lecture contains hardly any formula (see Chap. 4 in Spivak's comprehensive treatise [73] for a detailed exposition of Riemann's lecture) so no detailed proof of this claim is given there. Not even an explicit form of the Riemann tensor in a general coordinate system appears there. Following Spivak [73], such an explicit expression, and a proof of the fact that its non-vanishing is an obstruction for a metric being locally flat, appeared first in an unpublished paper, known as the *Commentatio*, submitted in 1861 to the Paris Academy and which appeared in the second edition of the collected works of Riemann [62]. However, as emphasized in [26] (see also [34]), this unpublished paper was primarily concerned with a problem of heat conduction proposed by the Academy in 1858, so its primary concern was not in Geometry, and it is debatable whether our modern understanding of the field might be biasing the interpretation of its results (see the interesting discussion on this *Commentatio* in [34] and references therein). In any case, an explicit expression of what we now know as the Riemann tensor appeared for the first time in this unpublished paper by Riemann. The first explicit proof that the vanishing of this tensor is sufficient for a metric being (locally) flat was given by Elwin Bruno Christoffel [17]. The proof of Theorem 2.1 is so central in differential geometry that many different proofs have been given. Spivak's treatise includes seven explicit proofs, which very nicely illustrate how differential geometry has evolved over the years from a discipline where the use of coordinates was central, to a mature theory where geometric objects fully independent of coordinates are the basic entities.

Another fundamental local characterization of Riemannian manifolds involves manifolds of constant sectional curvature. Recall that the sectional curvature of a two-plane $\Pi \subset T_p\mathcal{M}$, $p \in (\mathcal{M}, g)$ with non-degenerate induced first fundamental form h is defined as

$$K(p;\Pi) = \frac{\mathrm{Riem}^g(X,Y,X,Y)}{h(X,X)h(Y,Y) - h(X,Y)^2}, \qquad \{X,Y\} \text{ any basis of } \Pi.$$

The sectional curvature $K(p;\Pi)$ agrees with the Gauss curvature at p of the surface ruled by geodesics starting at p with tangent vector $X \in \Pi$. A Riemannian manifold has constant curvature k if $K(p;\Pi) = k$ for all $p \in \mathcal{M}$ and all non-degenerate

$ * &* N51E16*. +* &* * 1 *
37,.3 * .*1 *. & U 1

plane $\Pi \in T_p\mathcal{M}$. A well-known lemma by Friedrich Schur [70] asserts that any Riemannian manifold with sectional curvature $K(p, \Pi) = k(p)$ independent of Π is in fact of constant curvature.

For any $k \in \mathbb{R}$, define the manifold $\mathbb{M}_k^{(p,q)}$ as the subset $B(K) \subset \mathbb{R}^{p+q}$

$$B(K) := \{x \in \mathbb{R}^{p+q}; 1 + \frac{k}{4}(\eta_{ij}^{(p,q)} x^i x^j) > 0\},$$

endowed with the metric g_k given by

$$g_k := \frac{1}{\left(1 + \frac{k}{4}(\eta_{ij} x^i x^j)\right)^2} \eta_{ij}^{(p,q)} dx^i dx^j,$$
$$\eta^{(p,q)} := \mathrm{diag}(\underbrace{-1, \ldots, -1}_{p}, \underbrace{1, \ldots, 1}_{q})$$

(it is easily seen that $B(K)$ is connected for all values of k and all signatures (p, q)). Spaces of constant curvature are all isometric to this space. This is already mentioned without proof in Riemann's inaugural lecture [47] and it was also known to Eugenio Beltrami [4].

Theorem 2.2 *Let $(\mathcal{M}, g)$ be a Riemannian manifold of constant sectional curvature k and signature (p, q). Then $(\mathcal{M}, g)$ is locally isometric to $\mathbb{M}_k^{(p,q)}$.*

Riemannian manifolds of constant curvature are also locally characterized by the condition that the Killing algebra of $(\mathcal{M}, g)$ has maximal dimension. Recall that a Killing vector is a vector field $\xi \in \mathfrak{X}(\mathcal{M})$ satisfying $\mathcal{L}_\xi g = 0$ where $\mathcal{L}$ denotes Lie derivative. The set of Killing vectors $\mathcal{K}$ in a manifold defines a finite-dimensional algebra (with respect to the Lie bracket) called Killing algebra. If the manifold has dimension n, the Killing algebra has dimension $\dim(\mathcal{K}) \leq \frac{n(n+1)}{2}$ and whenever equality is attained, the manifold is called maximally symmetric. A classical result by Luigi Bianchi [6] asserts that maximally symmetric spaces are precisely those of constant sectional curvature.

Theorem 2.3 (Bianchi [6]) *A Riemannian manifold $(\mathcal{M}, g)$ is maximally symmetric if and only if it has constant sectional curvature.*

Another important class of Riemannian manifolds singled out by the properties of its isometry group is the class of Riemannian symmetric spaces. These are defined as Riemannian manifolds $(\mathcal{M}, g)$ such that, at any point $p \in \mathcal{M}$ there exists an isometry $\Psi_p : \mathcal{M} \to \mathcal{M}$, $\Psi_p^\star(g) = g$ satisfying $\Psi_p(p) = p$ and $d\Psi_p|_p = -\mathrm{Id}|_p$, i.e. this isometry leaves p invariant and reverses any vector at p. Riemannian symmetric spaces can be locally characterized in terms of the Riemann tensor, as first proved by Élie Cartan.

Theorem 2.4 (Locally symmetric spaces, Cartan [12]) *A Riemannian manifold $(\mathcal{M}, g)$ is locally isometric to a Riemannian symmetric space if and only if the curvature tensor satisfies*

$$\nabla Riem^g = 0.$$

The Riemann tensor can be decomposed in its trace and its trace-free parts. The trace part defines the Ricci tensor $\mathrm{Ric}^g(X, Y) = \mathrm{tr}_g \mathrm{Curv}^g(X, \cdot)Y$ from which the scalar curvature is defined as $\mathrm{Scal}^g = \mathrm{tr}_g \mathrm{Ric}^g$. Useful tensors defined in terms of the Ricci tensor are the Schouten tensor Sch^g (in dimension $n > 2$) and the Einstein tensor Ein^g:

$$\mathrm{Sch}^g := \frac{1}{n-2}\left(\mathrm{Ric}^g - \frac{1}{2(n-1)}\mathrm{Scal}^g g\right), \qquad \mathrm{Ein}^g = \mathrm{Ric}^g - \frac{1}{2}\mathrm{Scal}^g.$$

The Weyl tensor is the trace-free part of the Riemann tensor, defined as

$$\mathrm{Weyl}_g = \mathrm{Riem}^g - \mathrm{Sch}^g \odot g$$

where $\odot$ is the Kulkarni-Nomizu product on symmetric tensors

$$\begin{aligned}(A \odot B)(X_1, X_2, X_3, X_4) =& A(X_1, X_3)B(X_2, X_4) - A(X_1, X_4)B(X_2, X_3) \\ &+ A(X_2, X_4)B(X_1, X_3) - A(X_2, X_3)B(X_1.X_4).\end{aligned}$$

Hermann Weyl introduced [79] the tensor named after him in an attempt to find a theory of gravitation invariant under scale (conformally invariant). The Weyl tensor has the fundamental property that $\mathrm{Weyl}_{\Omega^2 g} = \mathrm{Weyl}_g$ where $\Omega \in C^\infty(\mathcal{M}, \mathbb{R}^+)$. Thus, it vanishes automatically in any locally conformally flat space. Schouten [66] proved that the vanishing of this tensor is also sufficient in dimension $n \geq 4$ (in dimension $n = 2$ the Weyl tensor is not defined and in $n = 3$ it vanishes identically). As is well-known, in dimension $n = 2$ any space is locally conformally flat. In dimension three, the necessary and sufficient condition for local conformal flatness was obtained by Émile Cotton [16] in terms of the vanishing of the so-called Cotton tensor, which in dimension $n > 2$ is defined as

$$\mathrm{Cott}(X, Y, Z) = (n-2)\left((\nabla_Z \mathrm{Sch})(X, Y) - (\nabla_Y \mathrm{Sch})(X, Z)\right),$$

where $X, Y, Z \in \mathfrak{X}(\mathcal{M})$. This tensor is conformally invariant in any space with vanishing Weyl tensor (hence always in dimension three). Summarizing the discussion above, we have

Theorem 2.5 (Locally conformally flat spaces, Cotton, Weyl and Schouten) *A Riemannian manifold* $(\mathcal{M}^n, g)$ $n \geq 3$ *of signature* (p, q) *is locally isometric to a conformally flat space* $(\mathbb{R}^n, \Omega^2 \eta^{(p,q)})$, $\Omega \in C^\infty(\mathbb{R}^n, \mathbb{R}^+)$ *if and only if*

(i) If $n = 3$*, the Cotton tensor vanishes.*
(ii) If $n \geq 4$*, the Weyl tensor vanishes.*

3 Local Characterizations of the Schwarzschild and Kruskal Spacetimes

Among indefinite Riemannian manifolds, the Lorentzian ones (i.e. with signature $(-, +, \dots, +)$) play a fundamental role in physics, as they describe spacetimes. For the purposes of this article by **spacetime** we simply mean a Lorentzian manifold[2] $(\mathcal{M}, g)$ of dimension $n \geq 2$. In Einstein's theory of General Relativity the spacetime geometry is linked to the energy-momentum contents of the non-gravitational fields via the Einstein field equations

$$\mathrm{Ein}^{g} = \chi T$$

where χ is a coupling constant depending on Newton's gravitational constant G and T is the energy-momentum tensor of the fields. In particular, a spacetime is *vacuum* whenever its Einstein tensor (or equivalently its Ricci tensor) vanishes.

With the exception of the flat Minkowski spacetime, perhaps the most important spacetime in general relativity is the Schwarzschild spacetime of mass $M \in \mathbb{R}$ found by Karl Schwarzschild [67] shortly after the Einstein field equations were proposed (see [77] for the corresponding solution in higher dimensions). To be specific, by this solution we mean the manifold $\mathcal{M}_{\rm Sch} := \mathbb{R} \times (r_0, \infty) \times \mathbb{S}^{m-1}$ $m \geq 3$, where $r_0 = 0$ if $M \leq 0$ and $r_0 = (2M)^{\frac{1}{m-2}}$ if $M > 0$, endowed with the metric (in global coordinates $(t, r) \in \mathbb{R} \times (r_0, \infty)$)

$$g_{\rm Sch} = -\left(1 - \frac{2M}{r^{m-2}}\right) dt^2 + \left(1 - \frac{2M}{r^{m-2}}\right)^{-1} dr^2 + r^2 \gamma^{\mathbb{S}^{m-1}} \tag{1}$$

in units where $8\pi G = (m-1)\omega_{m-1}$ and $\gamma^{\mathbb{S}^k}$ is the standard metric of the k-sphere $\mathbb{S}^k$ and ω_k the corresponding total volume. The Schwarzschild spacetime is spherically symmetric and vacuum. Spherically symmetric means that the spacetime admits an action of $SO(m)$ as a group of isometries with orbits being either spacelike codimension-two surfaces or points. The fact that the metric (1) is defined also in another manifold, namely $\mathbb{R} \times (0, r_0) \times \mathbb{S}^2$ was found (for $m = 3$) independently by Droste [20] and Hilbert [40].

The Schwarzschild spacetime of mass $M \neq 0$ admits a maximal extension that keeps the properties of being spherically symmetric and vacuum. This is the Kruskal spacetime, which is defined as follows. Let $\mathcal{M}_{\rm Kr} = \mathcal{U} \times \mathbb{S}^{m-1}$ where $\mathcal{U} \subset \mathbb{R}^2$ is an open set defined by the inequality $\mathcal{U} := \{(u, v) \in \mathbb{R}^2; r(u \cdot v) > 0\}$ where $r := (a, b) \subset \mathbb{R} \to \mathbb{R}$ is the maximal solution of the differential equation

$$\frac{dr}{dx} = \frac{1}{(m-2)\,x}(2|M|)^{\frac{1}{m-2}}\left(1 - \frac{2M}{r^{m-2}}\right). \tag{2}$$

---

[2] Usually, the notion of spacetime requires also time-orientability, but this is irrelevant for the local characterization theorems we are describing here.

For $M < 0$, the maximal domain of definition does not include $x = 0$. However, for $M > 0$, the solution to this equation extends smoothly through $x = 0$ and has a non-zero gradient at $x = 0$. The Kruskal metric g_{Kr} is defined on $M_{\text{Kr}} = \mathcal{U} \times \mathbb{S}^{m-1}$ and reads

$$g_{\text{Kr}} = \frac{4}{(m-2)^2} |2M|^{\frac{2}{m-2}} \frac{1 - 2Mr(u \cdot v)^{2-m}}{uv} du dv + r(u \cdot v)^2 \gamma^{\mathbb{S}^{m-1}}.$$

The case $m = 3$ is best known, as it corresponds to the original extension by Martin Kruskal [46] (also found independently by Georges Szekeres [75]) of the Schwarzschild spacetime of dimension $n = 4$. In this case the ODE (2) for $M > 0$ can be integrated in simple terms as

$$uv = \left(1 - \frac{r}{2M}\right) e^{\frac{r}{2M}} \quad \Longrightarrow \quad g_{\text{Kr}} = -\frac{32M^3}{r(uv)} e^{-\frac{r}{2M}} du dv + r^2(uv) \gamma^{\mathbb{S}^{m-1}}.$$

The original Schwarzschild spacetime corresponds to the domain $\{v > 0\} \cap \{u < 0\}$ and describes the exterior of the black hole spacetime. The conformal diagram of this exterior domain can be found in [56] along with interesting results on scattering and peeling of test fields on this and other backgrounds. The process of relativistic diffusion on the Schwarzschild spacetime is discussed in [30].

The most important local characterization theorem in gravitational theory is the Birkhoff theorem [7]. This theorem appears in the literature in two different guises, either as a local characterization theorem of the Kruskal spacetime, or as a statement that under suitable circumstances, a spherically symmetric spacetime must admit an additional Killing vector.[3] The two guises appear because the Schwarzschild spacetime was obtained under the assumptions of vacuum, spherically symmetry, time-independence and the implicit assumption that ∇r is spacelike. Birkhoff's proof that any vacuum, spherically symmetric spacetime admits an additional Killing vector and the fact that this Killing is timelike whenever ∇r is spacelike, provides therefore a local characterization theorem for the Schwarzschild metric among vacuum spherically symmetric spacetimes. In fact, the first result concerning the existence of an additional static symmetry is due to Jebsen [42] in a work virtually forgotten for a long time despite the fact that due credit was given e.g. in [33, 69]. This work has been brought back to light in more recent times, first in [64] and later in [19], see [43] and references therein for further details on the discovery of the Schwarzschild spacetime and Birkhoff's theorem.

Historical remarks aside, the fact of the matter is that the Schwarzschild and Kruskal spacetimes are unique among vacuum spherically symmetric n-dimensional Lorentzian spacetimes. This result has enormous physical significance as it means that, like in Newtonian theory, a spherically symmetric gravitational field outside its sources depends on a single constant, namely the total mass of the configuration. It also states that spherical sources cannot emit gravitational radiation, even if they are themselves evolving. The precise local characterization result is as follows.

---

[3] The second guise is sometimes stated that spherically symmetric, say vacuum, spacetimes are static, but this is clearly not true in this generality as the Kruskal extension admits no global timelike Killing vector, see [22] where this fact is discussed in some more detail.

Theorem 3.1 (Local characterization of the Kruskal spacetime) *Let* $(\mathcal{M}^n, g)$ *be a Lorentzian* n*-dimensional space such that the group* $SO(n-1)$ *acts as a group of isometries with orbits which are either codimension-two spacelike submanifolds or points. If* $(\mathcal{M}, g)$ *is vacuum, then there exists* $M \in \mathbb{R}$ *such that* $(\mathcal{M}, g)$ *is locally isometric to the Kruskal spacetime* $(\mathcal{M}_{Kr}, g_{Kr})$ *of mass* M*. Let* $r : \mathcal{M} \to \mathbb{R}$ *be defined as*

$$r(p) = \left(\frac{|\mathcal{O}_p|}{\omega_{n-2}} \right)^{\frac{1}{n-2}}$$

where $\mathcal{O}_p$ *is the orbit of* $SO(n-1)$ *containing* p*,* $|\mathcal{O}_p|$ *its* $(n-2)$*-volume and* ω_{n-2} *the* $(n-2)$*-volume of the round unit sphere* $\mathbb{S}^{n-2}$*. If* $\langle \nabla r, \nabla r \rangle_p \neq 0$*, then* $(\mathcal{M}, g)$ *is locally isometric near* p *to the Schwarzschild spacetime* $(\mathcal{M}_{Sch}, g_{Sch})$ *of mass* M*.*

Analogous local characterization theorems of spherically symmetry spacetimes exist for many other matter models, including Λ-vacuum (i.e. $\mathrm{Ric}^g = (n-1)\Lambda g$, $\Lambda \in \mathbb{R}$), electrovacuum, Einstein-Maxwell-dilaton and many others. A detailed description of all such result is beyond the scope of this note, see [68] and references therein for results in this direction.

3.1 Characterization of Spherically Symmetric Spacetimes with an Additional Killing Vector

As already mentioned, the Birkhoff theorem for the vacuum spherically symmetric spacetime can be viewed as a local characterization result of the Kruskal spacetime or as a statement that spherically symmetric vacuum spacetimes admit an additional local isometry with generator orthogonal to the spherically symmetric orbits and which is timelike on regions where ∇r is spacelike. Viewed from this perspective, the Birkhoff theorem is a particular case of a result proved by Eiesland in [23] (and announced at the Eastern Meeting of the American Mathematical Society at Chicago in 1921 [24] shortly before the result by Birkhoff appeared). The theorem by Eiesland, although correct, uses a rather strange convention of the Einstein tensor. In terms of the usual definition $\mathrm{Ein}^g = \mathrm{Ric}^g - \frac{1}{2}\mathrm{Scal}^g g$ this theorem has been recently quoted in [76]. We follow the notation in this reference (with a couple of typo corrections).

Theorem 3.2 (Eiesland [23]) *The necessary and sufficient conditions that a locally spherically symmetric Lorentzian manifold with line-element*

$$ds^2 = -A^2 dt^2 + B^2 dr^2 + C^2 \left(d\theta^2 + \sin^2 \theta d\phi^2 \right)$$

A, B and C being arbitrary, non-zero functions of t and r, and C not a constant, shall admit a one-parameter group of isometries generated by a vector field $K = k_0(t,r)\partial_t + k_1(t,r)\partial_r$ are

$$A^2 Ein^t_r = \frac{2\Psi'}{C}\partial_r C \partial_t C,$$
$$A^2B^2\left(Ein^t_t - Ein^r_r\right) = \frac{2\Psi'}{C}\left[A^2(\partial_r C)^2 + B^2(\partial_t C)^2\right],$$

where $\Psi = \Psi(C)$ is some differentiable function of C, or a constant, and Ein^r_t, Ein^t_t etc. denote the components of the Einstein tensor (with one index raised) in the coordinate basis $\{\partial_t, \partial_r, \partial_\theta, \partial_\phi\}$. Under these conditions the Killing field K is, up to a constant multiple, given by

$$k_0(t,r) = -\frac{1}{e^\Psi AB}\partial_r C, \qquad k_1(t,r) = \frac{1}{e^\Psi AB}\partial_t C.$$

This theorem has been extended in [65] to Einstein spaces which are locally warped products $g = h + r^2\gamma$, where h is a Lorentzian metric and r a positive non-constant function on a two-dimensional space $\mathbb{Q}$, and γ is a metric on an m-dimensional manifold Σ^m, $m \geq 2$. Let us state and prove a theorem that generalizes Eiesland's result to arbitrary warped space with two-dimensional base.

Theorem 3.3 *Let $(\mathcal{M}, g)$ be locally isometric to a warped product space $(\mathbb{Q} \times \Sigma^m, g)$ $(m \geq 2)$, i.e. $g = h + r^2\gamma$, where r is a smooth positive function and h a metric on a two-dimensional manifold $\mathbb{Q}$ and (Σ, γ) is a (pseudo-)Riemannian manifold. Let $\epsilon = 1$ if h is Lorentzian and $\epsilon = -1$ if h is positive or negative definite.*

Assume that there is no open set in $\mathbb{Q}$ where $|\nabla r|^2_h$ vanishes identically. Then $(\mathcal{M}, g)$ admits a Killing vector $K \in \mathfrak{X}(\mathbb{Q})$ if and only if, away from the points where $|\nabla r|^2_h = 0$, there exists a smooth, nowhere-zero, function $H(r)$ such that

$$Ric^g|_{\mathbb{Q}} - \frac{1}{2}tr_h\left(Ric^g|_{\mathbb{Q}}\right)h = \frac{(m-1)H'}{2Hr}\left(dr \otimes dr + \epsilon dr^\star \otimes dr^\star\right)$$

where $Ric^g|_{\mathbb{Q}}$ is the restriction of the Ricci tensor of g to the tangent planes of $\mathbb{Q}$ and $\star$ is the Hodge dual in $(\mathbb{Q}, h)$. Moreover, the one-form $\boldsymbol{K} := h(K,\cdot)$ is $\boldsymbol{K} = H(r)dr^\star$

Proof It is immediate to check that a vector field K tangent to the $\mathbb{Q}$ factor is Killing if and only if K is Killing for $(\mathbb{Q}, h)$ and $K(r) = 0$. On the open set where dr is not null for the metric h (which is dense on $\mathbb{Q}$ by assumption), dr and $dr^\star$ are a basis of the cotangent space of any point in $\mathbb{Q}$. We work on this open (possibly disconnected) set from now on. Note that $\langle dr, dr^\star\rangle = 0$ and $\langle dr, dr\rangle = -\epsilon\langle dr^\star, dr^\star\rangle$ where $\langle\cdot,\cdot\rangle$ denotes scalar product with h. The condition $K(r) = 0$ is equivalent to the existence of a function $H \in C^\infty(\mathbb{Q}, \mathbb{R})$ such that $\boldsymbol{K} = H dr^\star$ or, in index notation $K_i = H\boldsymbol{\eta}_{ij}\nabla^j r$ where $\boldsymbol{\eta}$ is the volume form of h and ∇ denotes the Levi-Civita derivative of h. It is known (and follows e.g. by explicit computation) that

$$\mathrm{Ric}^g|_{\mathbb{Q}} = K^h h - \frac{m-1}{r}\mathrm{Hess}_h r$$

where the Hessian of a functions is, as usual, defined by $(\mathrm{Hess} f)(X, Y) = \langle \nabla_X \mathrm{grad}\, f, Y \rangle$ and where K^h is the Gauss curvature of h. Let us denote by $\mathring{\mathrm{Ric}}^g_{\mathbb{Q}}$ the trace-free part of $\mathrm{Ric}^g|_{\mathbb{Q}}$ so that we have

$$\mathrm{Hess}_h r = -\frac{r}{m-1}\mathring{\mathrm{Ric}}^g_{\mathbb{Q}} + f h \tag{3}$$

where f is a scalar function whose form does not concern us. The Killing equations for K read

$$\begin{aligned} 0 = \nabla_i K_j + \nabla_j K_i &= \eta_{jl} \nabla_i H \nabla^l r + H \eta_j{}^l \nabla_i \nabla_l r + \eta_{il} \nabla_j H \nabla^l r + H \eta_i{}^l \nabla_j \nabla_l r \\ &= \eta_{jl} \left(\nabla_i H \nabla^l r - \frac{Hr}{m-1} (\mathring{\mathrm{Ric}}^g_{\mathbb{Q}})^l{}_i \right) + \eta_{il} \left(\nabla_j H \nabla^l r - \frac{Hr}{m-1} (\mathring{\mathrm{Ric}}^g_{\mathbb{Q}})^l{}_i \right) \end{aligned}$$

after inserting (3) and noticing that the term proportional to f drops. An endomorphism A^i_j satisfying $\eta_{jl} A^l_i + \eta_{il} A^l_j = 0$ must be proportional to the identity. Thus, there exists a function F on $\mathbb{Q}$ such that

$$\nabla_i H \nabla_j r - \frac{Hr}{m-1} (\mathring{\mathrm{Ric}}^g_{\mathbb{Q}})_{ij} = F h_{ij} \qquad \Longrightarrow \qquad A = \frac{1}{2} \langle \nabla H, \nabla r \rangle \tag{4}$$

where the implication follows simply by taking trace. The antisymmetric part of (4) implies that there exists a function $\hat{H} : I \to \mathbb{R}$ such that $H = \hat{H} \circ r$. We make the usual abuse of notation of writing simply $H(r)$. Note that H cannot vanish anywhere (on the open set $|\nabla r|^2_h \neq 0$) because if $H = 0$ at one point p, then (4) implies that H' also vanishes at that point, and the Killing K vanishes at p together with its covariant derivative, which can only happen if K is identically zero.

A basis for trace-free symmetric tensors is given by $dr \otimes dr + \epsilon dr^\star \otimes dr^\star$ and $dr \otimes dr^\star + dr^\star \otimes dr$. Expanding $\mathring{\mathrm{Ric}}^g_{\mathbb{Q}}$ in this basis and using $h = \frac{1}{|\nabla r|^2_h}(dr \otimes dr - \epsilon dr^\star \otimes dr^\star)$ it follows that

$$\mathring{\mathrm{Ric}}^g_{\mathbb{Q}} = \frac{(m-1)H'}{2Hr} \left(dr \otimes dr + \epsilon dr^\star \otimes dr^\star \right),$$

as stated in the theorem. □

Eiesland's Theorem 3.2 follows from this result because on open connected sets where $|\nabla r|^2_h \neq 0$, we can choose $H > 0$ (possible after reversing the Killing K). Writing $H = e^{-\Psi}$, Theorem 3.2 is recovered after a straightforward calculation.

Bernhard Riemann and His Work

Lizhen Ji

Abstract This chapter attempts to give a brief summary of the life and work of Riemann. It tries to address the following issues: the influence of his education and early life on his work, and a summary of his major works and an overview of his work, and the general impact of his work through the concepts and terminology named after him.

1 Introduction

Bernhard Riemann (1826–1866) was one of the most original mathematicians in the history of mathematics. Though he had written only a small number of papers, he changed the way we view and do mathematics. For example, to get a taste of his impact, it suffices to think of several concepts named after him: Riemann surfaces, Riemann-Roch theorem, Riemann zeta function, and Riemannian geometry.

Specifically, before Riemann, complex functions were studied on domains of the complex plane $\mathbb{C}$, and the issue of analytic continuation and the ensuing problem of multi-valuedness were complicated. Now after the work of Riemann, the proper way to understand a complex analytic function and its analytic continuation, in particular to understand properly an algebraic function defined by an algebraic equation, is to consider its associated Riemann surface. One can even turn things around so that Riemann surfaces become the more basic spaces, and holomorphic functions on them can be viewed as maps between Riemann surfaces. (In the language of categories, Riemann surfaces become objects and holomorphic maps become morphisms between the spaces.) This thinking led to the Riemann mapping theorem, which in turn led to the uniformization theorem for Riemann surfaces, which is one of the most important theorems in mathematics. Consequently, geometry and analysis are now intimately connected. Riemann proved the crucial part of Riemann's inequality in the Riemann-Roch Theorem. In doing this, he initiated the theory of algebraic topology and made topology an essential part of geometric function theory and geometric

---

L. Ji (✉)
Department of Mathematics, University of Michigan, Ann Arbor, MI 48109, USA
e-mail: lji@umich.edu

© Springer International Publishing AG 2017

L. Ji et al. (eds.), *From Riemann to Differential Geometry and Relativity*,
DOI 10.1007/978-3-319-60039-0_20

analysis. The Riemann-Roch theorem also motivated the important index theories for higher dimensional manifolds and algebraic varieties.

Similarly, the Riemann zeta function brought the methods of complex analysis into analytic number theory, and it is now hard to imagine what remains if we remove ideas of complex analysis, for example, the Riemann zeta function and generalized Riemann zeta functions, from analytic number theory. Generalized Riemann zeta functions include L-functions which are also crucial in algebraic number theory and arithmetic geometry. For example, without them, it is almost impossible to formulate the celebrated Langlands program.

In differential geometry, the idea of Riemannian metric, Riemannian connections and their non-positive definite generalizations in the theory of relativity are indispensable to most people working in the subject, and higher dimensional generalizations of the Riemannian mapping theorem can also be expressed in terms of Riemannian geometry.

Besides his geometric and global view, Riemann's work has several other features. Riemann's intuitive reasoning played a crucial role in his work. It might fall somewhat short of a rigorous proof, but the brilliant ideas in his works are so much clearer, because his work is not overly filled with lengthy computations (though heavy computations were important for him to gain his understanding and intuition). One example of Riemann's intuitive reasoning is his application of Dirichlet's principle to various problems such as the Riemann mapping theorem and the Riemann-Roch theorem.

Integration between mathematics and physics runs through a large part of Riemann's work. According to Klein [8, p. 167], this integration is the source of inspiration for Riemann:

> he endeavored again and again to find a general mathematical formulation for the laws underlying all natural phenomena.... *these physical views are the mainspring of Riemann's purely mathematical investigations.*

Riemann had a bigger perspective and was a better philosopher than most mathematicians. For example, his paper on the hypothesis which lies at the basis of geometry contains a substantial philosophic part. Not only the intrinsic geometry is important to him, its relation to reality, or to the real space we live in, is also important to him. According to Freudenthal [4], Riemann was

> one of the most profound and imaginative mathematicians of all time, he had a strong inclination to philosophy, indeed, was a great philosopher. Had he lived and worked longer, philosophers would acknowledge him as one of them.

Riemann was also capable of doing very extensive and heavy calculations. This can be confirmed by some existing calculating sheets or scratching papers used by him at the Nachlass of Riemann in the Historical Library of the University of Göttingen. This is consistent with Riemann's ability and joy in doing computation when he was a young boy. Riemann resembled Gauss in the aspect of powerful computation. Computation was crucial to Riemann, and his intuition, deep insights and understanding might result from, but is certainly supported by, his extensive computations.

Besides his originality, Riemann was also broad. His work covered almost all major subjects in mathematics. It is important and interesting to get an overview of the work of Riemann. This task is probably beyond the ability of any single person, though several people including Klein [8], the editor of the Russian edition of Riemann's Collected works, and the author of the book [9], had tried. Narasimhan edited the 1990 edition of Riemann's collected works [16]. In the preface, he also summarized and commented on some of Riemann's major works after writing "No one person is capable of a full analysis of Riemann's work, its history, its development and its influence on mathematics."

In this book, there are several chapters devoted to summarizing and analyzing some aspects of Riemann's works, for example, [5, 12–14, 18]. On the other hand, in this chapter, we want to give a brief overview of all major aspects of Riemann's work and life. Given the long lasting impact of Riemann on the development of mathematics, it is worthwhile to view Riemann and his work from different perspectives. This chapter should complement other chapters of this book. We also hope to understand better, or give some explanation for, his work from the perspective of his education and interaction with others.

2 Riemann's Work I: His Best Known Works

As we mentioned earlier, Riemann made deep contributions to many different subjects. In the opinion of most mathematicians, he is probably best known for his work in the following four subjects:

1. Complex analysis.
2. Real analysis.
3. Riemannian geometry.
4. Number theory.

We will start with the most famous result of his work during his life time, and then summarize some of his major results in the above four subjects.

1.a. *Abelian functions and the Jacobian inversion problem.*
Given what most people know about Riemann, it might be surprising to point out that during Riemann's life time, he first became famous and probably was best known for his solution of the Jacobi inversion problem for Riemann surfaces (or algebraic functions) of higher genus.[1] Though this particular result of Riemann is

---

[1] As it is known, the theory of elliptic integrals motivated the theory of Abelian integrals. One basic insight in the theory of elliptic integrals is to consider the inverse of an elliptic integral. The general framework for Abelian integrals on a fixed compact Riemann surface is the map via integration of holomorphic 1-forms from the group of divisors of degree 0 to the Jacobian variety of the Riemann surface. To invert this map, we need to identify its image. One version of the Jacobi inversion problem, as presented in most modern textbooks on Riemann surfaces, asks to prove that this map is surjective. The full version asks, in addition, for a precise description of the inverse images of this map.

!"#$# % & & !' (! !) * * + , * - . *
/ * !+ 3, * +* . / 3 - + , V* . * 3 + *& * *
+* . $ 5 "/" & 6 13
;. A 5 6 $ 1* *

Chapter 6
Teleparallelism

According to FEYNMAN (1962/63), “gravity is that field which corresponds to a gauge invariance with respect to displacement transformations.” Taking this literally would favor Einstein’s teleparallelism equivalent of GR, which has been recast (MIELKE 1992; MIELKE et al. 1996) into a Yang–Mills-type *gauge theory of translations*. On the other hand, Ashtekar’s reformulation ASHTEKAR (1986, 1991) of general relativity (GR) is a complex reformulation that essentially projects out the right and left helicity modes of the graviton.

6.1 Chiral Teleparallelism

Originally, these complex variables were developed in the Hamiltonian approach. In the equivalent Lagrangian formulation, this change of variables is more simply induced by a *generating function* (DOLAN 1989; MIELKE 1990) that involves a boundary term dC_{TT} constructed from a *translational Chern–Simons term* (CS) multiplied by the imaginary unit. This is facilitated in the framework of Riemann–Cartan (RC) geometry, where the coframe ϑ^{α} surfaces as a soldered *translational* gauge potential (MIELKE et al. 1993; TRESGUERRES & MIELKE 2000) having $T^{\alpha} := D\vartheta^{\alpha}$ as its corresponding translational field strength, i.e., torsion.

Via a sort of duality rotation, CS-type boundary terms dC transform into a viable gravitational Lagrangian: the standard Hilbert action of general relativity (GR), or, as was already suggested by EINSTEIN (1928), a theory with *teleparallelism* ($\mathrm{GR}_{\|}$). Recently, it was shown (MIELKE 1992) how a complex Yang–Mills-type version of $\mathrm{GR}_{\|}$ arises by a canonical transformation induced via dC_{TT}. Moreover, the naturalness of Sparling’s energy complex in this approach has been pointed out; cf. DUBOIS-VIOLETTE & MADORE (1987).

© Springer International Publishing Switzerland 2017

E.W. Mielke, *Geometrodynamics of Gauge Fields*,
Mathematical Physics Studies, DOI 10.1007/978-3-319-29734-7_6

Since the resulting chiral version of gravity is surprisingly close to Yang–Mills, we can adopt a related Hamiltonian formulation in which the Poisson brackets for the constraints are on a par with the local Poincaré algebra. Interestingly enough, the corresponding Gauss constraint of spatial diffeomorphisms is annihilated, on the operator level, by a state vector that is proportional to the exponentiated translational CS term, thus improving on MIELKE (1998, 1999). Classical configurations dominating this action are *torsion instantons*, thus hinting at another close formal parallelism of internal Yang–Mills with chiral $\mathrm{GR}_{\|}$. Moreover, this similarity with quantum gauge theories paves the way to a Becchi–Rouet–Stora–Tyutin (BRST) quantization of gravity based on *translational* ghost operators. The correlation function for Wilson loops is known (GUADAGNINI et al. 1990) to yield, to first order, the Gauss self-linking number. For the full expectation value, the generalized Jones polynomial related to the Kauffman bracket of knots (BRÜGMANN et al. 1992) arises once a *framing* has been chosen. In the case of *chiral teleparallelism*, the loops automatically carry an orthonormal frame along their path, a concept that was realized much earlier by CARTAN (1924). Thus such *Cartan circuits* winding around torsion instantons may provide another clue to a field quantization of gravity based on chiral $\mathrm{GR}_{\|}$.

One may speculate that teleparallel ("flat") spacetime may reveal, on the Planck scale $\ell := \sqrt{8\pi\hbar G_{\mathrm{N}}/c^3} \simeq 10^{-33}$ cm, a topologically rich spectrum of dislocations and knotlike loops.

6.2 Parity-Violating Topological Invariants in Gravity

In the one-dimensional harmonic oscillator model (MIELKE 1998) with q as generalized coordinate, a *canonical transformation* can be induced by a boundary term derived from the *Chern–Simons* (CS)-type term $\mathscr{C} = q^2/2$ as *generating function*. After quantization, the corresponding operator $\mathscr{S} = \exp(-\mathscr{C}) = \exp\left(-q^2/2\right)$ induces a well-known renormalization of the Schrödinger wave function. On the other hand, for diffeomorphism-invariant topological field theories, HOROWITZ (1989) has shown that $\Psi = N \exp\left(i\theta \int \underline{\mathscr{C}}\right)$ is, up to an overall factor, the unique solution of the Hamiltonian constraints. It is worth investigating whether this carries over to gravity:

Since the Poincaré group $P := \mathbb{R}^4 \rtimes SO(1,3)$ is the semidirect product of translations and Lorentz rotations, its gauging leads to the two associated Chern–Simons (CS) three-forms:

$$C_{\mathrm{TT}} := \frac{1}{2\ell^2}\left(\vartheta^\alpha \wedge T_\alpha\right) = -\frac{(-1)^{\mathrm{sig}}}{2\ell^2}\, {}^*\mathscr{A}, \qquad (6.2.1)$$

F . , . 0 1 3 . +* * 1* * 3 1 * * 3 0
1 * & 0 7, * * & * 1 1 * 21 * 3 * * *
0 1 & . (5 AA 0 AAA6

$$
\begin{aligned}
C_{\mathrm{RR}} &:= -\mathrm{Tr}\left(\Gamma \wedge \Omega - \frac{1}{3}\Gamma \wedge \Gamma \wedge \Gamma\right) \\
&= -\frac{1}{2}\left(\Gamma_\alpha{}^\beta \wedge R_\beta{}^\alpha + \frac{1}{3}\Gamma_\alpha{}^\beta \wedge \Gamma_\beta{}^\gamma \wedge \Gamma_\gamma{}^\alpha\right) .
\end{aligned}
\tag{6.2.2}
$$

They are *translational* and Lorentz-rotational CS terms, where $\mathscr{A} := {}^*(\vartheta_\alpha \wedge T^\alpha) = \mathscr{A}_i dx^i$ is the *axial torsion* one-form. In the first *translational* CS three-form, there necessarily occurs a fundamental length ℓ for dimensional reasons.

In a gauge theory with the linear group $SL(5, \mathbb{R})$ as structure group, containing the de Sitter group $\mathrm{SO}(1, 4)$ or $\mathrm{SO}(2, 3)$ as a subgroup, both CS three-forms are intimately interrelated: via a Wigner–Inönü-type contraction, the CS decomposition

$$
\hat{C} = C_{\mathrm{RR}} - 2C_{\mathrm{TT}} \tag{6.2.3}
$$

into linear and translational terms was realized already in HEHL et al. (1995) (see footnote 31), and later by NIEH (2007), MERCURI (2009). In contrast to the metric-free Pontryagin form, in the NY term (6.2.4), a metric $g_{\alpha\beta}$ is needed to raise and lower the indices, for instance in $T_\alpha = g_{\alpha\beta}T^\beta$.

The corresponding boundary or Nieh–Yan term (NIEH & YAN 1982; GUO et al. 1999) can be obtained by exterior differentiation

$$
dC_{\mathrm{TT}} \equiv \frac{1}{2\ell^2}\left(T^\alpha \wedge T_\alpha + R_{\alpha\beta} \wedge \vartheta^\alpha \wedge \vartheta^\beta\right), \tag{6.2.4}
$$

whereas the corresponding Pontryagin term

$$
dC_{\mathrm{RR}} = -\mathrm{Tr}\ (\Omega \wedge \Omega) = \frac{1}{2}\, R^{\alpha\beta} \wedge R_{\alpha\beta} \tag{6.2.5}
$$

is more familiar. Its integration is proportional to the Pontryagin number. Both are *parity violating* (MIELKE 1999).

Whereas the Pontryagin term (6.2.5) is a topological Lagrangian whose variation returns the second Bianchi identity

$$
DR_{\alpha\beta} \equiv 0, \tag{6.2.6}
$$

the less well known torsion identity (6.2.4) is based, after exterior multiplication by ϑ_α, on the first Bianchi identity

$$
DT^\alpha \equiv R_\beta{}^\alpha \wedge \vartheta^\beta \tag{6.2.7}
$$

in Riemann–Cartan (RC) geometry.

!"#$# % & & !' (! !) * * + , * - . *
/ * !+ ,# 1 & 5 "/" & 6 13
* : :: : : : #
+* 8 0) * 0 9 7 0 9 4 0 & :

II

CIRCLES

The circle, in various guises, occurs in many branches of mathematics. In this chapter we shall consider those properties which have the habit of appearing in different contexts. Our methods will be mixed. We shall use Euclidean geometry or analytical geometry, according to our needs at the time. We begin with the consideration of certain one-to-one mappings of the plane onto itself, and the product of these mappings. (See 0.11.) Later on we study inversion, which is only one-to-one onto if we cut out a point of the plane, or better still, add a point to the plane.

12.1 The nine-point circle

We have already encountered this circle (§7.3). It is a circle which passes through nine points intimately connected with any given triangle *ABC*. (We shall not be using vector methods, and we shall use the customary notation for the points connected with a triangle). The nine-point circle is the first really exciting one to appear in any course on Euclidean geometry which goes far enough.

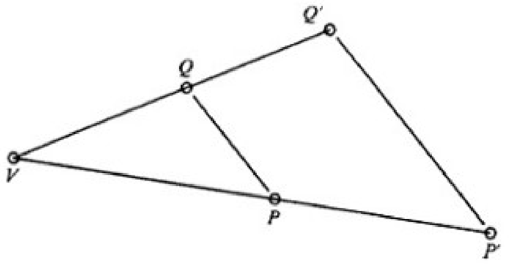

Fig. 12.1

To prove the existence of the nine-point circle of a triangle, we make use of *a one-to-one mapping* of the Euclidean plane onto itself. This mapping is called *a central dilatation*. We choose a point *V*, and call it *the center of the dilatation*

(Fig 12.1).

If P is any point in the plane its image, or map P' is found by choosing the unique point P' which lies on VP, is on the same side of V as is P, and is such that $\overline{VP'} = k\overline{VP}$ where k is a fixed real positive number. If $P = V$, we define V' to be V.

This mapping is evidently one-to-one. It is also called *a central similarity*. It will occur again later (§41.5) as one of the fundamental mappings of the Euclidean plane onto itself. Here we are only interested in the case when k is positive, and when the point P moves on a circle, center O (Fig. 12.2). If O' is the map of O, then the triangles VOP and $VO'P'$ are similar, so that $\overline{VP'} = k\overline{VP}$, Since O is a fixed point, and $| OP | = r$, the radius of the circle on which P moves, we deduce that P' moves on a circle, center O' of radius kr.

Let us now recall some properties of a triangle ABC which we wish to use (Fig 12.3). We know that the perpendiculars AD', BE' and CF' from the vertices A, B and C onto the opposite sides all pass through a point H, the orthocenter (§7.2). If AD' is produced to cut the circumcircle of triangle ABC again at D, it is easy to see that $\overline{HD'} = \overline{D'D}$. since $\measuredangle BCD = \measuredangle BAD$(angles on the same arc), and $\measuredangle BCH$ is complementary to $\measuredangle B$, and is therefore equal to $\measuredangle BAD$. Hence the triangles HCD', DCD' are congruent, and $\overline{HD'} = \overline{D'D}$.

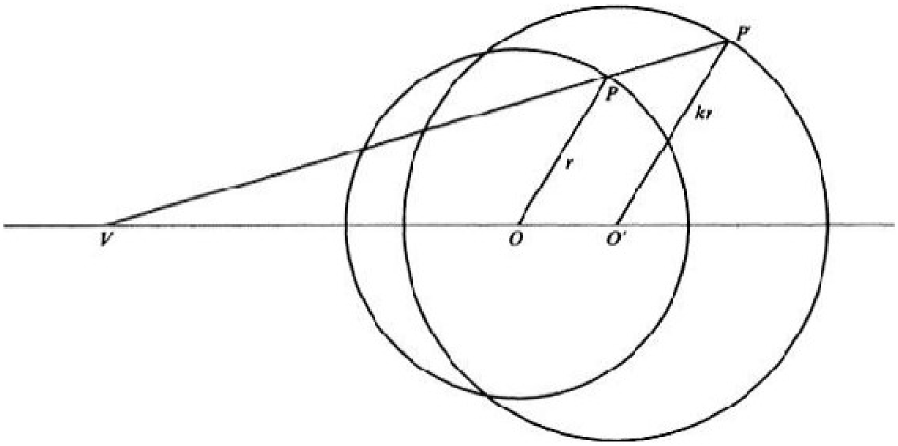

Fig. 12.2

Now consider the central dilatation which has vertex H, and for which $k = 1/2$. This maps every point P on P', where $\overline{HP'} = \overline{HP}/2$. Let P move around the circumcircle of triangle ABC. We know that P' will move on a circle of half the radius, with center bisecting HO, where O is the center of the circumcircle of triangle ABC. This circle will pass through the midpoints of HA, HB and HC, since *A*, *B* and *C* are on the circumcircle, and also through the points *D*′, *E*′ and *F*′, since we have shown that *D*′ is the midpoint of *HD*. We have therefore found *six* points connected with the triangle *ABC* which lie on a circle.

Now, a circle is uniquely determined if we are given *three* non-collinear points through which it must pass. Consider the triangle *BHC*. Its orthocenter is *A*. By what we have just proved, there is a circle which passes through *D*′, *E*′ and *F*′, the feet of the perpendiculars from *H*, *B* and *C* onto the opposite sides *BC*, *HC* and *HB*, and through the midpoints of *AH*, *AB* and *AC*, *A* being the orthocenter of triangle *HBC*. But there is only one circle through the points *D*′, *E*′ and *F*′. Hence the circle connected with triangle *ABC* which passes through *D*′, *E*′ and *F*′ and the midpoints of *HA*, *HB* and *HC* also passes through the midpoints of *AB* and *AC*. By considering the triangle *HCA* we see that this same circle also passes through the midpoint of *BC*. We have therefore established the existence of a circle which passes through the midpoints of the sides of triangle *ABC*, the midpoints of *HA*, *HB* and *HC*, and through the feet of the perpendiculars from the vertices of *ABC* onto the opposite sides. We also know that its center bisects *OH*, and that its radius is half that of the circumcircle.

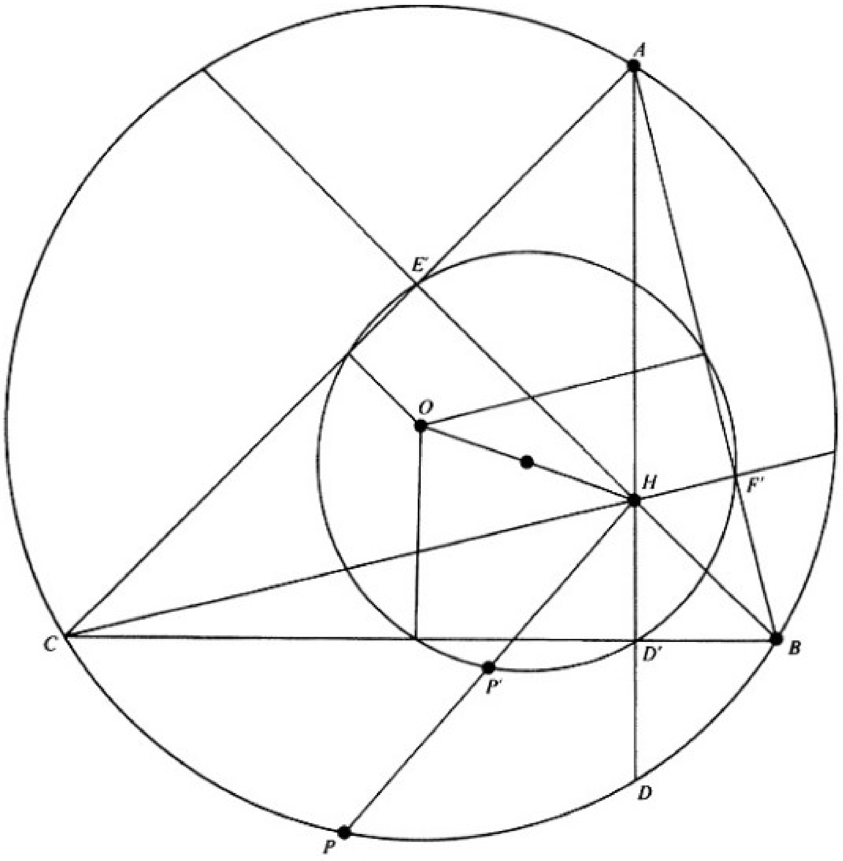

Fig. 12.3

This is a first example of the use of a mapping to establish a theorem in geometry. We shall come across other examples fairly soon.

Exercises

12.1 Let AO (where O is the circumcenter of triangle ABC) intersect the circumcircle again in Y. Prove that $BHCY$ is a parallelogram, and deduce that the circle through six points found above also passes through the midpoints of BC, CA and AB.

12.2 Let A', B', C' be the midpoints of BC, CA and AB, and let L', M' and N' be the midpoints of HA, HB and HC. Using the theorem that the join of the midpoints of two sides of a triangle is parallel to the third side prove that both $B'C'M'N'$ and $C'A'N'L'$ are rectangles, and deduce that $A'L'$, $B'M'$ and $C'N'$ are three diameters of one circle. Prove that this circle passes through the feet of the perpendiculars, D', E' and F' from A, B and C onto the sides BC, CA and AB, and so complete another proof of the existence of the nine-point circle.

12.3 We know that *N*, the center of the nine-point circle, bisects *OH* (in the notation of §12.1). Deduce from Exercise 12.1 that if *A′* is the midpoint of *BC*, then since *A′* bisects *HY* and *O* bisects *AY* $\overline{AH} = 2\overline{OA'}$. Now join *O* to *H* and show that the centroid of *ABC*, which is at the point *G* on *AA′*, where $\overline{AG} = 2\overline{GA'}$, must lie on *OH*, and $\overline{HG} = 2\overline{GO}$. (The line which contains the circumcenter, centroid, nine-point center and orthocenter of triangle *ABC* is called the *Euler line* of triangle *ABC*.)

12.4 If *L′* is the midpoint of *AH*, show that *AL′A′O* is a parallelogram, so that | *A′L′* | = | *OA* | = radius of circumcircle of *ABC*, and therefore, since *A′* and *L′* both lie on the nine-point circle, *A′L′* must be a diameter of the nine-point circle. (Compare Exercise 12.2.)

12.5 On the arc *A′D′* of the nine-point circle, take the point *X* one-third the way from *A′* to *D′*. Take similar points *Y* and *Z* on the arcs *B′E′* and *C′F′*. Show that the triangle *XYZ* is equilateral. (Hint: the triangle *X′Y′Z′* which corresponds to *XYZ* under the mapping, center *H*, which maps the circumcircle onto the nine-point circle, must be equilateral.)

12.6 If *A′*, *B′*, *C′* be the respective midpoints of the sides *BC*, *CA* and *AB* of a triangle *ABC*, show that *O*, the circumcenter of *ABC*, is the orthocenter of *A′B′C′*. From the parallel case of the Desargues Theorem (§5.1) show that *AA′*, *BB′*, *CC′* pass through a point *G*. (This, is of course, the centroid of *ABC*). There is a dilatation, center *G*, which maps *ABC* onto *A′B′C′*. Show that *H* maps onto *O*, and deduce that *H*, *G* and *O* are collinear, and $\overline{HG} = 2\overline{GO}$. (See Exercise 12.3.) Deduce also that $\overline{AH} = 2\overline{OA'}$.

12.7 Show that *O* is the incenter of the triangle *PQR* produced by drawing tangents to the circumcircle of *ABC* at the vertices. Show also that *H* is the incenter of the triangle *D′E′F′*, where *D′* is the foot of the perpendicular from *A* onto *BC*, *etc.* Prove that the sides of *PQR* and *D′E′F′* are respectively parallel. Deduce that the joins *D′P*, *E′Q*, *F′R* pass through a point *V*, the center of the dilatation which maps *D′E′F′* on *PQR*, and that *V* lies on *OH*, the Euler line of *ABC*. Deduce further that since the circumcenter of *D′E′F′* is the nine-point center of *ABC*, the circumcenter of *PQR* lies on the join of *V* to the nine-point center of *ABC*, that is, on the Euler line of *ABC*.

13.1 The centers of similitude of two circles

If *V* is a center for a dilatation which lies outside a circle $\mathscr{C}$ of center *O* (see Fig 13.1), we know already that the image $\mathscr{C}'$ of $\mathscr{C}$ is a circle center *O′*, where $\overline{VO'} = k\overline{VO}$, and that the radius of $\mathscr{C}'$ $=kr_r$,where *r* is the radius of $\mathscr{C}$:. The points *P* and *P′* lie on the same side of *V*, the radii *OP* and *O′P′* are parallel, and $\overline{VO}:\overline{VO'}$ =1:k= of radius of $\mathscr{C}$ If *P*, *Q* are distinct points on $\mathscr{C}$ and *P′*, *Q′* the image points, the chord *PQ* is parallel to the chord *P′Q′*. If *Q* moves towards *P* on $\mathscr{C}$ the point *Q′* moves towards *P′* on $\mathscr{C}$. Hence each of the two tangents from *V* to $\mathscr{C}$ will be a tangent to $\mathscr{C}$ and *V* is the intersection of two common tangents to $\mathscr{C}$ and to $\mathscr{C}$ each tangent having $\mathscr{C}$ and $\mathscr{C}'$ on the same side of it. We call such tangents *external* common tangents.

If *V* does not lie outside $\mathscr{C}$ all the statements made above continue to be true, but there are no tangents from *V* to $\mathscr{C}$ and therefore none from *V* to $\mathscr{C}'$. (see Fig 13.2).

These situations can be reproduced with any two given circles. Let us suppose that we are given two circles $\mathscr{C}$ and $\mathscr{C}'$ with distinct centers O and O'. Draw parallel radii OP and $O'P'$ *in the same sense* through O and O' respectively. Then the join of the points P and P' will intersect the line of centers OO' of the two circles in a fixed point V (unless the circles are of equal radii), and $\overline{VO}:\overline{VO'}$ of $\mathscr{C}$ radius of $\mathscr{C}'$. This point V is called *the external center of similitude* of $\mathscr{C}$ and $\mathscr{C}'$. It divides OO' *externally* in the ratio of the radii of $\mathscr{C}$ and $\mathscr{C}'$. We see that $\mathscr{C}$ can always be mapped onto $\mathscr{C}'$ by a central dilatation, vertex V.

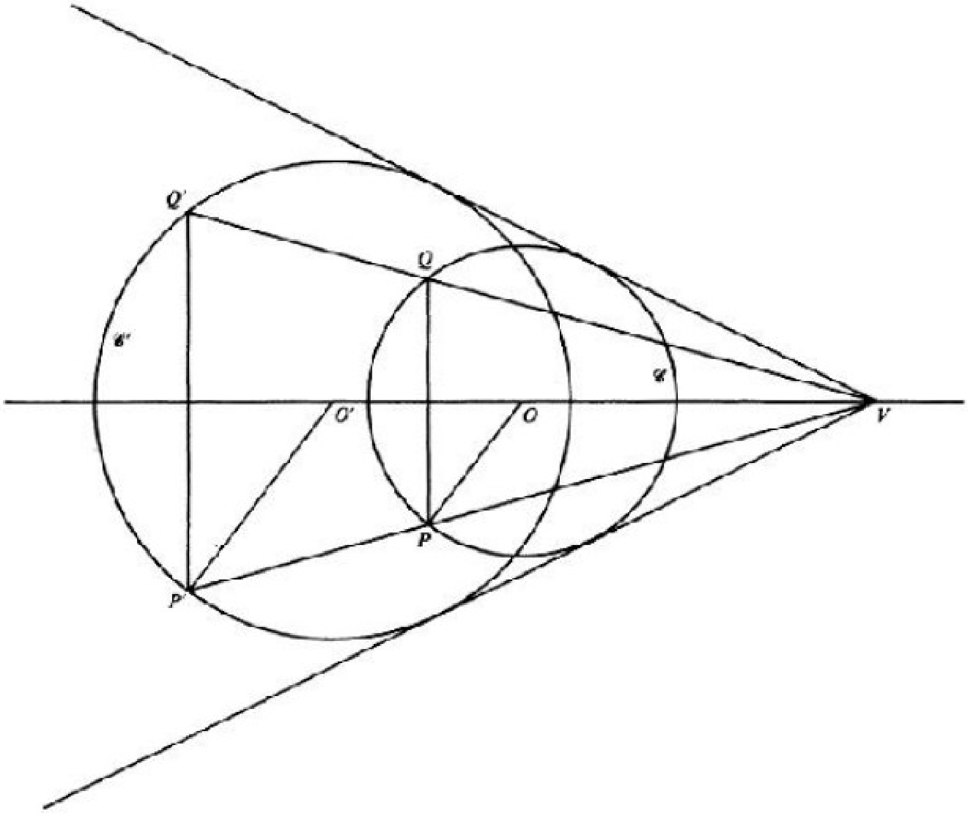

Fig. 13.1

If, in our description of central dilatations, we choose the image point P' so that P and P' are on *opposite* sides of V (Fig 13.3), the map of a circle is still a circle, the center of the circle $\mathscr{C}$ is the map O' of the center O of $\mathscr{C}$ and this time V divides OO' internally in the ratio of the radii of $\mathscr{C}$ and $\mathscr{C}'$ and OP and $O'P'$ are parallel but in opposite directions. If V lies outside $\mathscr{C}$ the tangents from V to $\mathscr{C}$ are also tangents to $\mathscr{C}'$ but $\mathscr{C}$ and $\mathscr{C}'$ lie on opposite sides of each common tangent.

The tangents are called *internal common tangents*, or *transverse* common tangents.

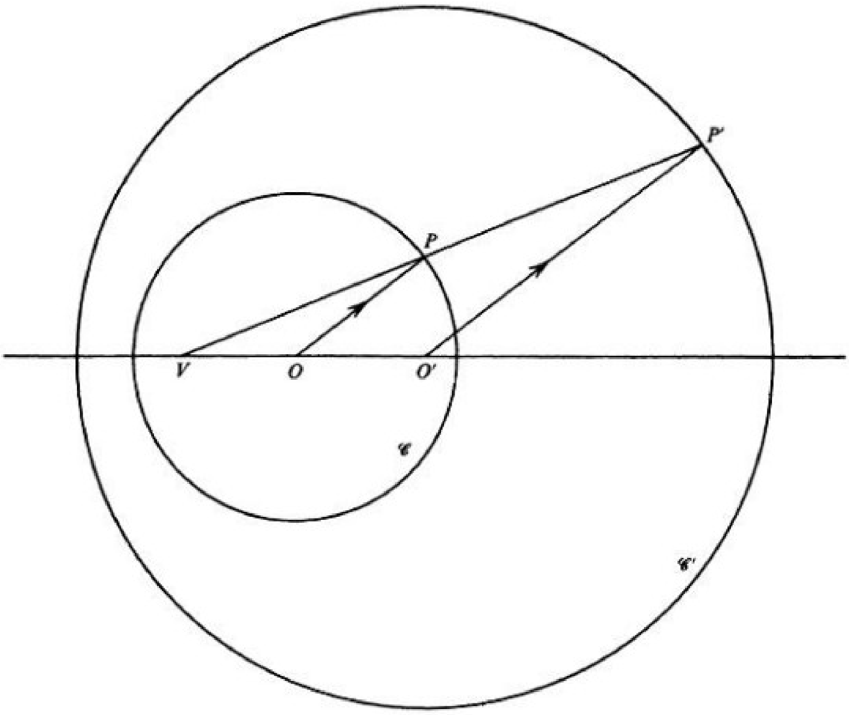

Fig. 13.2

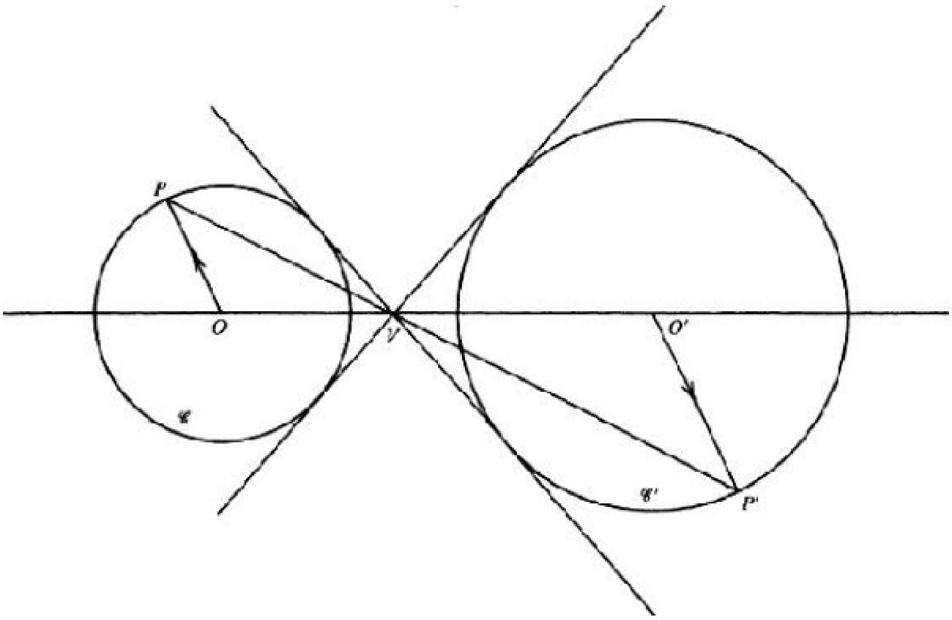

Fig. 13.3

Again, if we are given any two circles $\mathscr{C}$ and $\mathscr{C}'$ with centers O and O' the above situation can be reproduced. Draw parallel radii OP and $O'P'$ in *opposite senses* through O and O' respectively. Then the join of the points P and P' will intersect the line of centers OO' in a fixed point V, and $|VO| : |VO'|$ = radius of $\mathscr{C}$:radius of $\mathscr{C}'$. This point V is called *the internal center of similitude of* $\mathscr{C}$ *and* $\mathscr{C}'$. It divides OO' *internally* in the ratio of the radii of $\mathscr{C}$ and $\mathscr{C}'$.

Exercises

13.1 Show that if the circles $\mathscr{C}$ and $\mathscr{C}'$ are concentric, the common center must be considered to be both the internal and external center of similitude. Where are the internal and external centers when: (*a*) $\mathscr{C}$ and $\mathscr{C}'$ touch *externally*, lying on opposite sides of the tangent at the point of contact and: (*b*) $\mathscr{C}$ and $\mathscr{C}'$ touch *internally*, lying on the same side of the tangent at the point of contact?

13.2 If we have three circles in the plane, no two with equal radius, and the centers forming a triangle, prove that the six centers of similitude (two from each pair of circles) lie in sets of three on four lines. (Two proofs are possible, one using the Desargues Theorem (§5.2), and the other the Menelaus Theorem (§4.1.)) The four lines are called *axes of similitude of the three circles*.

13.3 In the preceding Exercise show that the join of two internal centers of similitude always passes through an external center of similitude. What deduction can you make about the join of the points of contact of a circle which touches two other circles externally?

*.

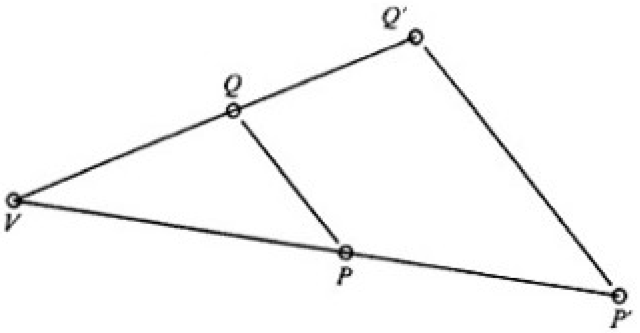
Q'
Q
V
P
P'

/ .

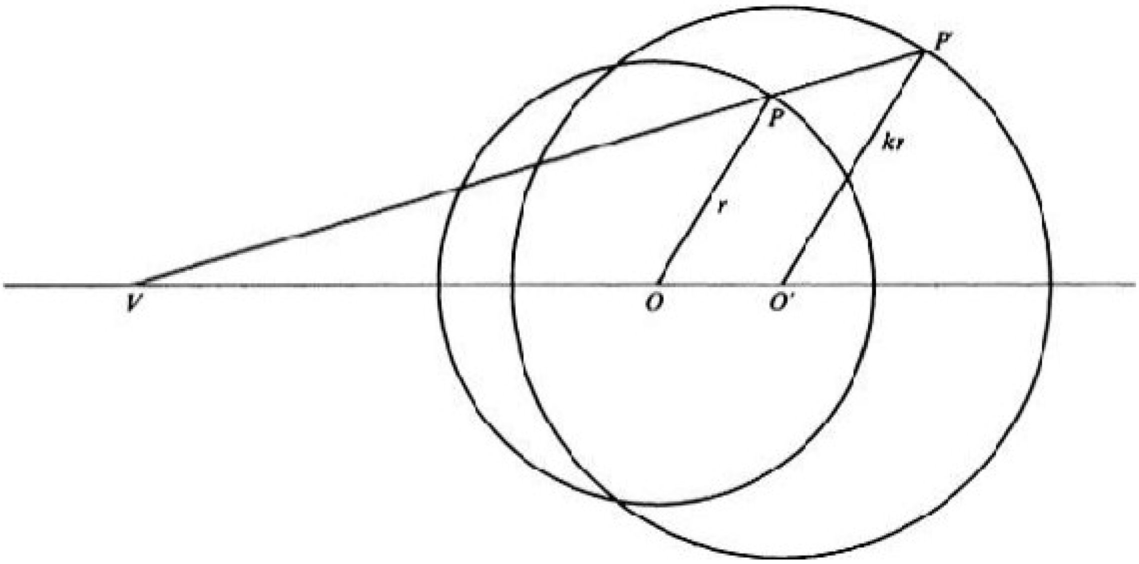
P'
P
kr
r
V
O
O'

/ .

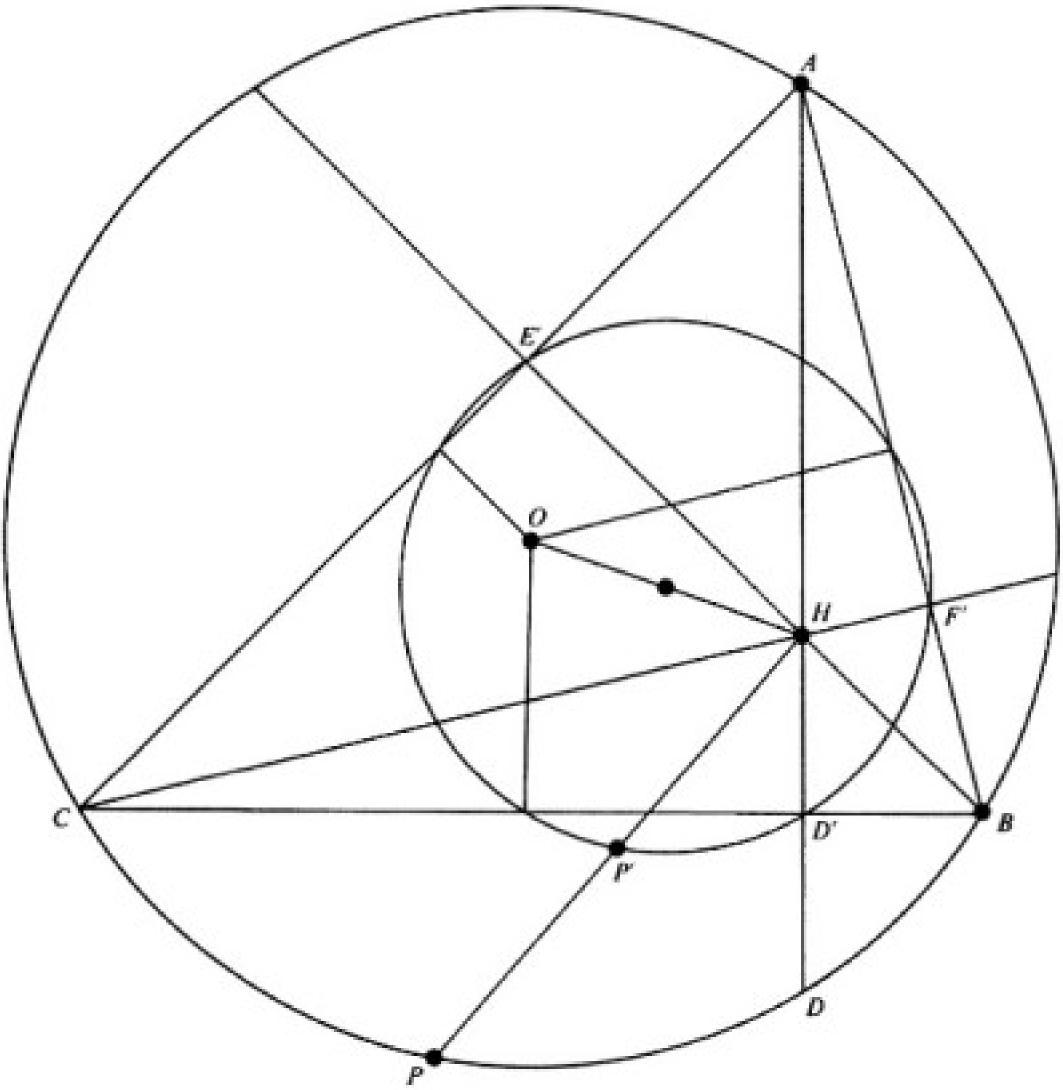
A
E'
O
H
F'
C
B
D'
P'
D
P

/ .

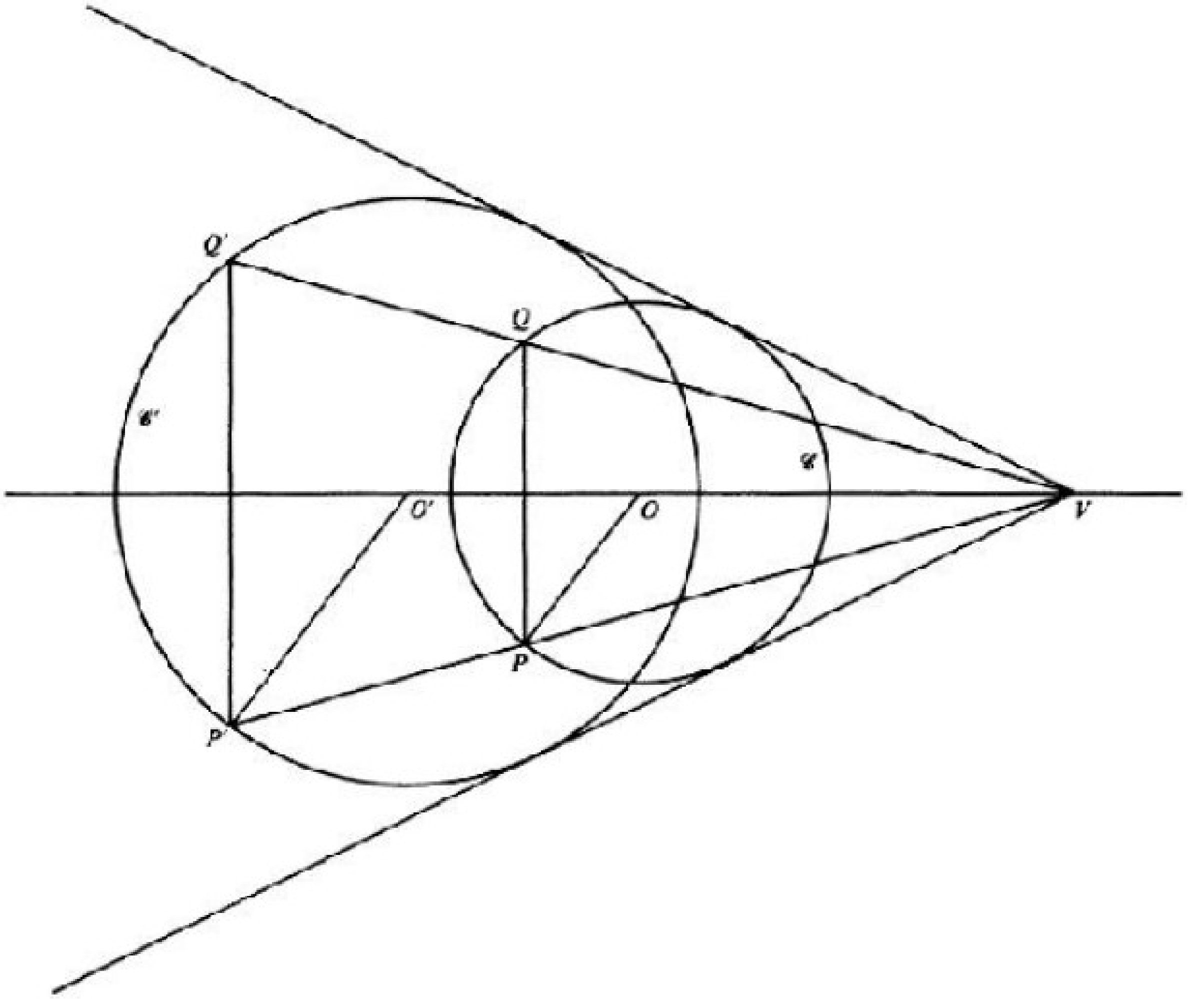
Q′
Q
𝒞′
𝒞
O′
O
V
P
P′

/ .

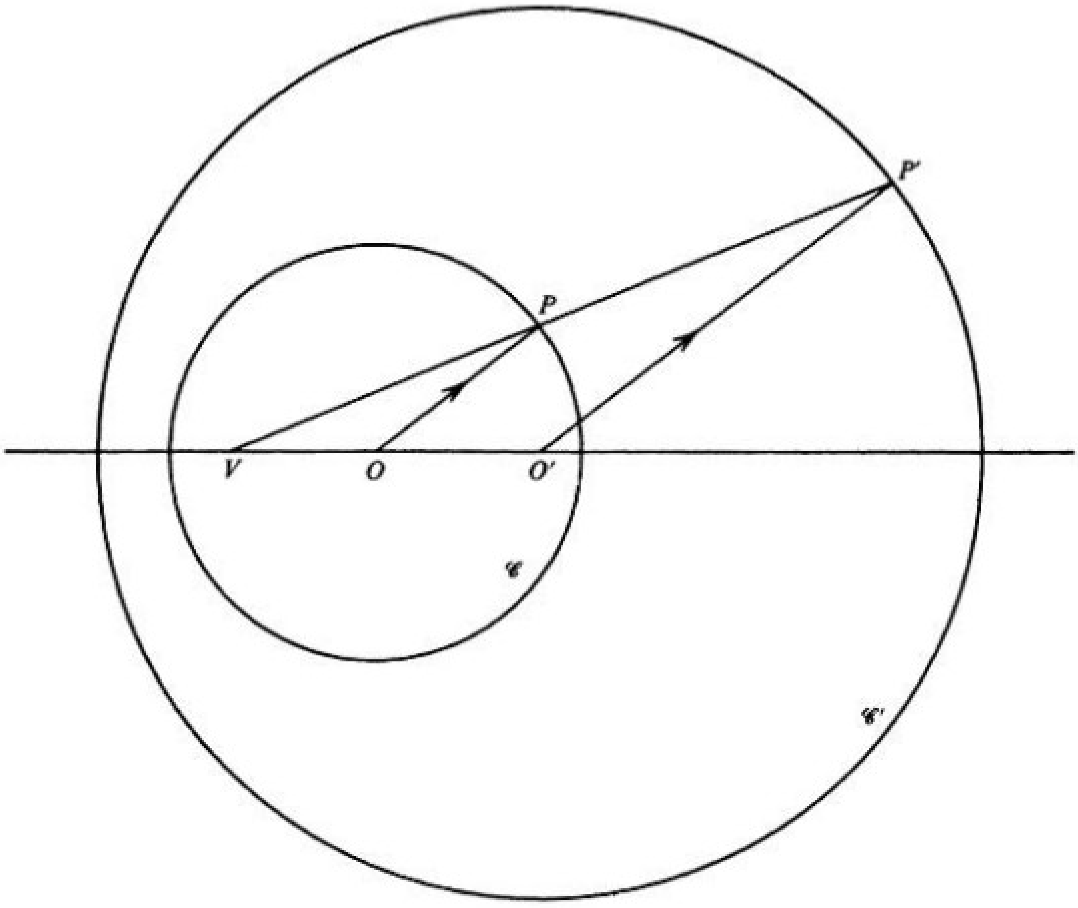
P'
P
V
O
O'
𝒞
𝒞'

/ .

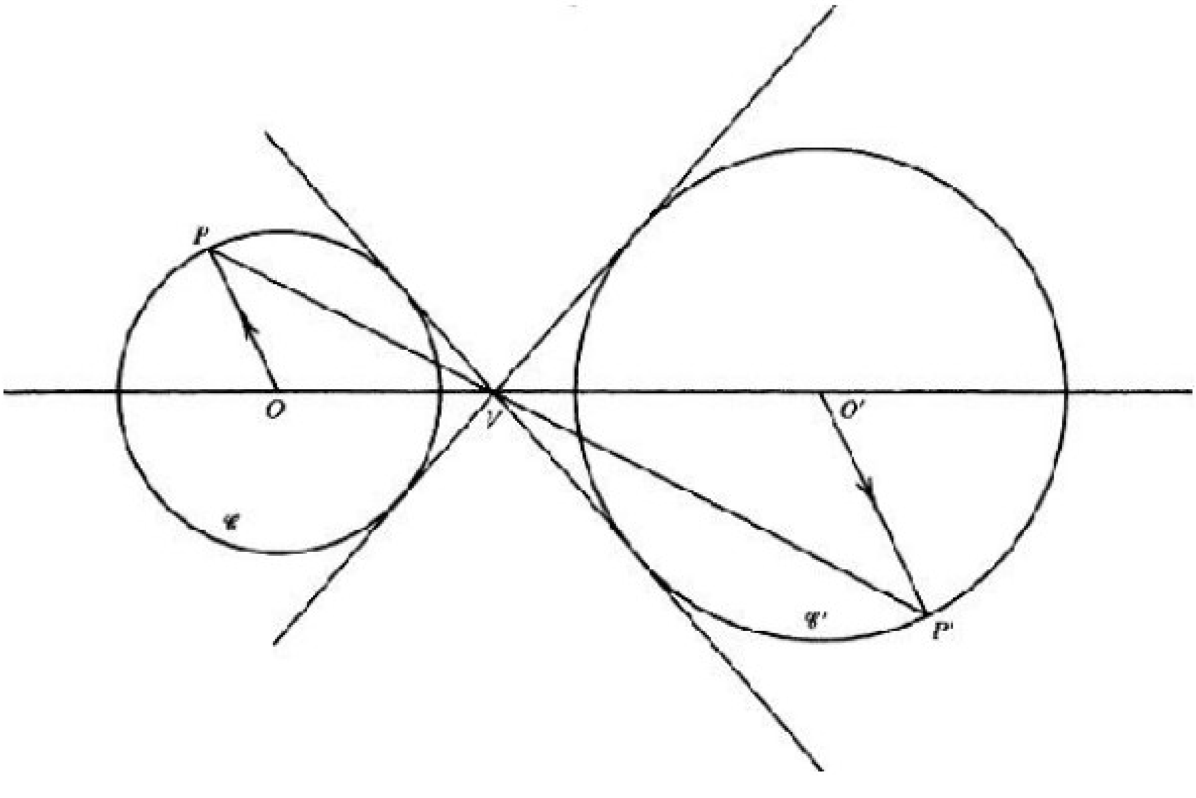
P
O
𝒞
V
O'
𝒞'
P'

/ .

Since C and C' are fixed points, where is P when the left-hand sum is a minimum? Show that when this is the case the angles subtended at P by BC and AB and AC are each 120°.

25.1 A fixed point theorem

Theorem *If* $\mathscr{C}$ *and* $\mathscr{D}$ *are two non-intersecting circles, any circle* $\mathscr{E}$ *which cuts them both orthogonally passes through two fixed points on their line of centers.* (See Fig 25.1.)

Proof. We anticipate the result of §29.4 which indicates that we may choose our coordinate axes so that the equations of the given circles $\mathscr{C}$ and $\mathscr{D}$ are respectively:

$$X^2 + Y^2 + 2gX + p^2 = 0, X^2 + Y^2 + 2gX + p^2 = 0,$$

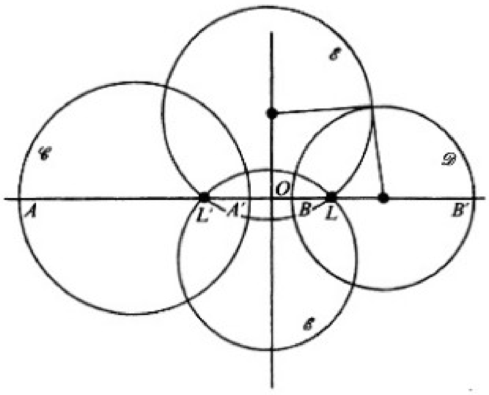

Fig. 25.1

The X–axis is the line of centers, and the Y–axis is a special line related to both circles called *the radical axis*. This line does not intersect either circle, as we see on putting $X = 0$ in both equations.

If $\mathscr{E}$ is the circle with equation

$$X^2 + Y^2 + 2kX+2/Y+c=0,$$

the orthogonality condition (Eqn (2), §22.1) applied to $\mathscr{E}$ and $\mathscr{C}$ and $\mathscr{E}$ and $\mathscr{D}$ is

$$2gk=p^2+c, 2g'k= p^2+c$$

Since $g \neq g'$, the circles $\mathscr{C}$ and $\mathscr{D}$ being distinct circles, we have $k = 0$, and $c = -p^2$. Hence the equation of $\mathscr{E}$ is

$$X^2 + Y^2 + 2lY - P^2 = 0$$

an equation in which l is the only variable parameter. This shows that a circle which cuts both $\mathscr{C}$ and $\mathscr{D}$ orthogonally must have its center on the radical axis (the Y–axis in this case) of the two circles, and such a circle cuts the line of centers, $Y = 0$, in the points given by $X^2 - p^2 = 0$, that is in the two fixed points $(p, 0)$ and $(-p, 0)$.

If we call these fixed points L and L', and the points where the line of centers cuts $\mathscr{C}$ and $\mathscr{D}$ respectively A, A' and B, B', we have

$$| OL | = | OL' |^2 = p^2 = \overline{OA} \,.\, \overline{OA'} = \overline{OB} \,.\, \overline{OB'},$$

since if we put $Y = 0$ in the equation of each circle we obtain the equations $X^2+2gX+p^2=0, X^2+2g^1X+p^2=0$,which show that $\overline{OA} \,.\, \overline{OA'} = p^2 = \overline{OB} \,.\, \overline{OB}$. From §19.1 it follows that L, L' are harmonic conjugates with respect to A, A' and to B, B'. Hence we have:

Corollary. Two non-overlapping pairs of points on a line have a pair of common harmonic conjugates.

This pair of common harmonic conjugates is unique. For if M, M' be another pair of harmonic conjugates to A, A' and B, B', and O' be the midpoint of MM', we should have $| O'M |^2 = | O'M' |^2 \ \overline{O'A} \,.\, \overline{O'A'} = \overline{O'B} \,.\, \overline{O'B'}$. and so the point O' would have the same *power* (§18.1) with respect to the circles $\mathscr{C}$ and $\mathscr{D}$. Such a point must lie on the radical axis of $\mathscr{C}$ and $\mathscr{D}$ (§27.3) and is therefore uniquely defined. Hence $O' = O$, and thence M, M' coincide with L, L'.

25.2 The inverse of a circle and a pair of inverse points

Theorem *If P and Q are points inverse in a circle $\mathscr{C}$, inversion in any circle center O maps P', Q' on points P', Q' inverse in the inverse $\mathscr{C}'$ of $\mathscr{C}$.*

Proof. All circles through P and Q are orthogonal to $\mathscr{C}$ (Thm 20.2). Hence after inversion all circles through P' and Q' are orthogonal to $\mathscr{C}'$. One such circle is the line $P'Q'$, the inverse of circle POQ. Hence $P'Q'$ is a diameter of $\mathscr{C}'$, and since circles through P' and Q' are orthogonal to $\mathscr{C}'$, the points P' and Q' are inverse points in $\mathscr{C}'$.

Remark. The proof holds with only minor modifications if $\mathscr{C}'$ is a line. Then P' and Q' are mirror images in the line.

Exercises

25.1 $\mathscr{C}$ is a circle which cuts a line l. From a point O on l a tangent OT is drawn to $\mathscr{C}$, and with center O and $|OT|$ as radius a circle is drawn which is orthogonal to $\mathscr{C}$. Show that this circle does *not* intersect the diameter of $\mathscr{C}$ which is perpendicular to l.

25.2 Let $\mathscr{C}$ and $\mathscr{D}$ be two circles which intersect in the points A and B. By inversion with respect to A as center show that the operations of inversion in distinct intersecting circles $\mathscr{C}$ and $\mathscr{D}$ are commutative if and only if $\mathscr{C}$ and $\mathscr{D}$ are orthogonal. (If inversion in $\mathscr{C}$ is represented by $\mathrm{T}^{\mathscr{C}}$, and in $\mathscr{D}$ by $T_{\mathscr{D}}$, we wish to show that if P is any point, $PT^{\mathscr{C}}T^{\mathscr{D}}=PT^{\mathscr{D}}T^{\mathscr{C}}$ if and only if $\mathscr{C}$ and $\mathscr{D}$ are orthogonal.)

25.3 Show that four mutually orthogonal circles can be inverted into a circle $\mathscr{C}$, two perpendicular diameters l and m of $\mathscr{C}$, and a concentric circle of radius r', where $r'^2 + r^2 = 0$, r being the radius of $\mathscr{C}$, so that this last circle is a virtual circle (§0.9). If the process of inversion in four mutually orthogonal circles be denoted by T_1, T_2, T_3, T_4, show that

$$P(T_1 T_2 T_3 T_4)=P,$$

for any point P. (This is the first encounter with inversion in a virtual circle, defined in §20.1.)

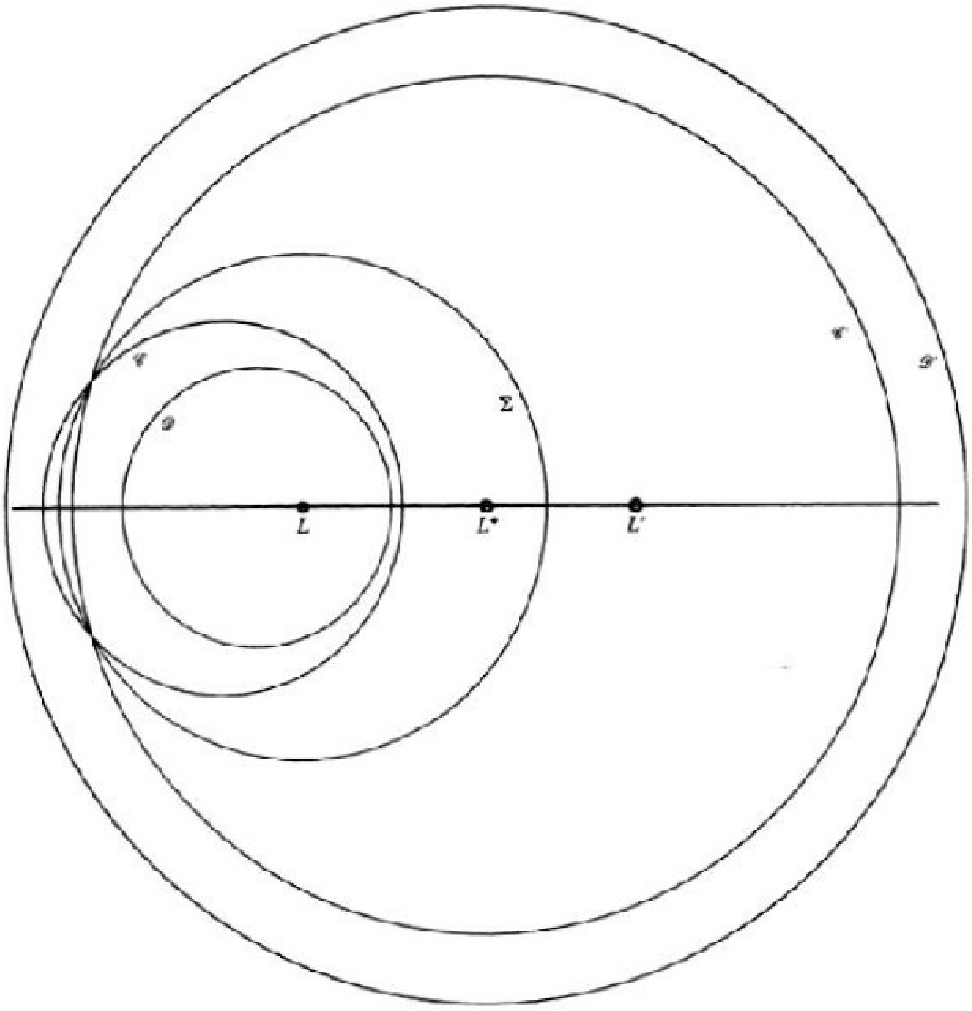
$\mathscr{C}$
$\mathscr{D}$
Σ
$\mathscr{C}'$
$\mathscr{D}'$
L
L^*
L'

Fig. 25.2

25.3 Circles into concentric circles

Theorem *Any two non-intersecting circles* $\mathscr{C}$ *and* $\mathscr{D}$ *may be inverted simultaneously into a pair of concentric circles*. (See Fig 25.2.)

Proof. By Thm 25.1 there exists a pair of points L, L' on the line of centers of $\mathscr{C}$ and $\mathscr{D}$ such that all circles orthogonal to $\mathscr{C}$ and to $\mathscr{D}$ pass through L and L', and, of course, such that all circles through L and L' are orthogonal to $\mathscr{C}$ and to $\mathscr{D}$. We invert with respect to one of these points, say L. Let L^* be the inverse of L'. Then since all circles through L and L' are orthogonal to $\mathscr{C}$ and to $\mathscr{D}$, all lines through L^* are orthogonal to $\mathscr{C}'$ and $\mathscr{D}'$, the inverses of $\mathscr{C}$ and of $\mathscr{D}$. Hence L^* is the common center of $\mathscr{C}'$ and $\mathscr{D}'$.

25.4 The Steiner porism

We are given two circles $\mathscr{C}$ and $\mathscr{D}$, one lying inside the other (Fig 25.3), and we begin to draw circles which touch both $\mathscr{C}$ and $\mathscr{D}$. Beginning with one such circle $\mathscr{C}_1$, the next circle we draw, $\mathscr{C}_2$, touches $\mathscr{C}$ and $\mathscr{D}$ and also $\mathscr{C}_1$. Then we draw $\mathscr{C}_3$ to touch $\mathscr{C}$ and $\mathscr{D}$ and $\mathscr{C}_2$, and so on, $\mathscr{C}\iota$ touching $\mathscr{C}$ and $\mathscr{D}$ and $\mathscr{C}_{i-1}$. We ask whether there exists a finite chain of circles $\mathscr{C}_1\mathscr{C}_2,\ldots\mathscr{C}_{n-1}\,\mathscr{C}_n$where each circle of the chain touches $\mathscr{C}$ and $\mathscr{D}$ and the two numerically adjacent circles, $\mathscr{C}\iota$ touching $\mathscr{C}_{i-1}$ and $\mathscr{C}_n$, and $\mathscr{C}_{n-1}$ itself touching $\mathscr{C}_{n-1}$ and $\mathscr{C}_1$?

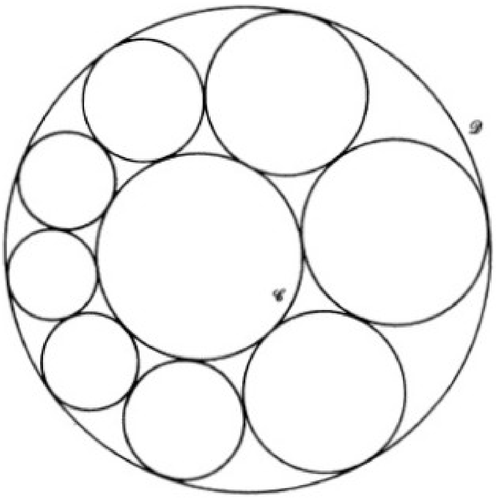

Fig. 25.3

If there is such a chain does it depend on where we start with $\mathscr{C}_1$, or may we draw $\mathscr{C}_1$ in any position, as long as it touches $\mathscr{C}$ and $\mathscr{D}$?

The answer to this problem is simple if we invert $\mathscr{C}$ and $\mathscr{D}$ into two concentric circles, which we may do by Thm 25.3, since they do not intersect. Our problem is reduced to the case when $\mathscr{C}'$ and $\mathscr{D}'$ are concentric. It is clear now that a chain only exists when there is a special relationship between *n* and the radii of $\mathscr{C}'$ and $\mathscr{D}'$, but that when this relationship *does* exist, since the whole configuration can be rotated about the common center of $\mathscr{C}'$ and $\mathscr{D}'$, it is independent of the position of the initial circle of the chain. Hence the problem of inscribing a chain of circles as described above has *no* solution or it has *an infinity* of solutions. This is why it is described as a porism. It was first formulated by Steiner.

25.5 The Steiner formula

It is of interest to obtain the formula given by Steiner which relates the radii of $\mathscr{C}$ and $\mathscr{D}$, the distance between their centers, and *n*, in the case when a chain of *n* circles can be inscribed to touch each other and $\mathscr{C}$ and $\mathscr{D}$.

Let us suppose that a common diameter of $\mathscr{C}$ and $\mathscr{D}$ intersects them in the points *A*, *B* and *C*, *D* respectively (Fig 25.4), and that after inversion with respect to *L*, which is a point on the common diameter, the points invert into A', B', and C', D', which will be the points of intersection of the same diameter with the concentric circles $\mathscr{C}'$ and $\mathscr{D}'$. Suppose that the radii of $\mathscr{C}$ and $\mathscr{D}$ are *R* and *r*, and the distance between their centers is *d*. Let the radii of the concentric circles $\mathscr{C}'$ and $\mathscr{D}'$ be R_1 and R_2. We use the result of §23.2, that $\{AB, CD\} = \{A'B', C'D'\}$, that is, that

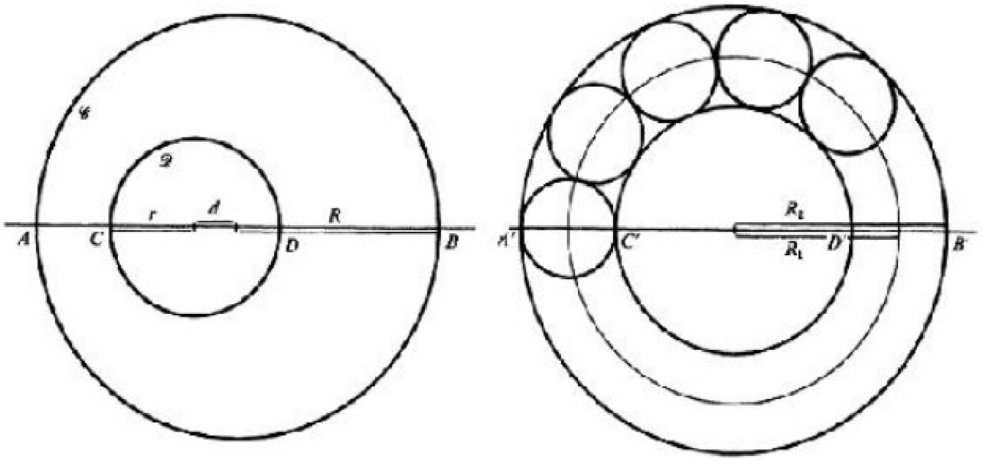

Fig. 25.4

$$\overline{AC}\,.\,\overline{BD}/\overline{AD}\,.\,\overline{BC} = \overline{A'C'}\,.\,\overline{B'D'}/\overline{A'D'}\,.\,\overline{B'C'}.$$

The left-hand side is readily seen to equal

$$(R-d-r)(R+d-r)/(R-d+r)=[(R-r)^2-d^2]/[(R+r)^2-d^2],$$

and the right-hand side is

$$(R_1-R_2)^2/R_1+R_2)^2$$

Equating these two expressions, a little reduction leads to

$$(R-r)^2-d^2=R_r(R_1/R_2/R_1-2).$$

Now if a chain of n circles can be inscribed to touch $\mathscr{C}'$ and $\mathscr{D}'$ (Fig 25.5), we have the simple relation

$$(R_2-R_2)/(R_2+R_1)=\sin(\pi/n),$$

so that

$$R_2/R_1=(1+\sin(\pi/n))/(1-\sin(\pi/n)),$$

and substitution gives us

$$(R-r)^2-d^2=4Rr\tan^2(\pi/n),$$

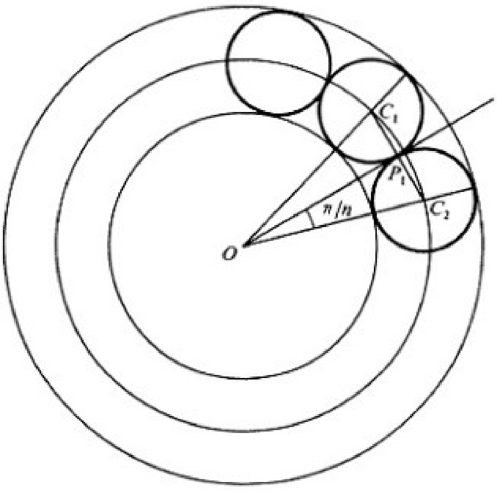

Fig. 25.5

for the relation between the radii of circles $\mathscr{C}$ and $\mathscr{D}$ and the distance between their centers which has to be satisfied for a chain of n circles to be inscribed, touching both circles.

Exercises

25.4 Show that two circles $\mathscr{C}$ and $\mathscr{D}$ intersecting at B and C are orthogonal if and only if there exist two circles touching $\mathscr{D}$ at B and C respectively which also touch each other at A, where A is any point on $\mathscr{C}$ distinct from B and C.

25.5 $\mathscr{C}$ and $\mathscr{D}$ are two distinct circles. Find the locus of a point P which is such that two circles, tangent both to $\mathscr{C}$ and $\mathscr{D}$ are also tangent to each other at P. (There are three cases to examine: $\mathscr{C}$ and $\mathscr{D}$ intersect: $\mathscr{C}$ and $\mathscr{D}$ touch, and $\mathscr{C}$ and $\mathscr{D}$ do not intersect.)

25.6 If the conditions for a chain of circles touching $\mathscr{C}$ and $\mathscr{D}$ are satisfied (The Steiner Porism, §25.4), show that the mutual points of contact of circles of the chain lie on a circle, and that this circle serves for inverting $\mathscr{C}$ into $\mathscr{D}$.

25.7 Continuing with Exercise 25.2, show that the operations of inversion in $\mathscr{C}$ and $\mathscr{D}$ cannot be commutative if $\mathscr{C}$ and $\mathscr{D}$ touch, or if $\mathscr{C}$ and $\mathscr{D}$ do not intersect.

26.1 The problem of Apollonius

Given three circles $\mathscr{C}_1\mathscr{C}_2$ and $\mathscr{C}_3$ how many circles can be drawn to touch all of them?

This is an example of a problem in which the variation or specialization of the given conditions (the circles in this case) may produce a change in the *number* of distinct solutions. A case encountered in the beginnings of Euclidean geometry is when the three circles are lines, and we consider how many circles can be drawn to touch the sides of a given triangle. There are *four* such circles, the incircle and the three excircles. Later we shall consider the matter algebraically (§38.1) and we shall see that the maximum *finite* number of solutions to the Apollonius problem is *eight*. Specialization of the circles to lines has halved the number of solutions.

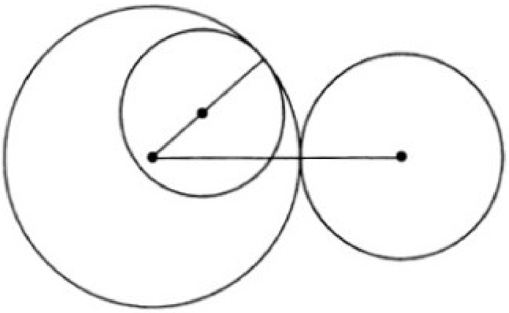

Fig. 26.1

In this section we shall examine the different possible cases of the Apollonius problem. If the circles $\mathscr{C}_1$, $\mathscr{C}_2$ and $\mathscr{C}_3$ are concentric, there are no real circles touching all three of them. We should stress, perhaps, that it is always *real* solutions we are seeking. (See §38.1.)

The condition for the contact of two circles can always be expressed thus:

(Distance between centers)= Sum or difference of radii,

according to the kind of contact, external or internal.

If the two circles be $\mathscr{C}$, $\mathscr{C}'$ (Fig 26.1), given in the usual form (§22.1), this condition becomes

$$(2gg'+2ff'-c-c')^2=4(g^2+f^2-c)(g'^2+f'^2-c'),$$

expressing the condition that if θ be the angle of intersection, $\cos\theta = \alpha 1$. Suppose now that the circle $\mathscr{C}'$ is a *point-circle*, that is a circle of zero radius, so that $g'^2 + f'^2 - c' = 0$. Then the condition for contact between $\mathscr{C}$ and a point-circle is

$$2gg'+2ff'-c-c'=0,$$

which is the condition that the point-circle itself, which is merely the point $(-g', -f')$, should lie on the circle $\mathscr{C}$, which is

$$g'^2+f'^2-2gg'-2ff'+c=0.$$

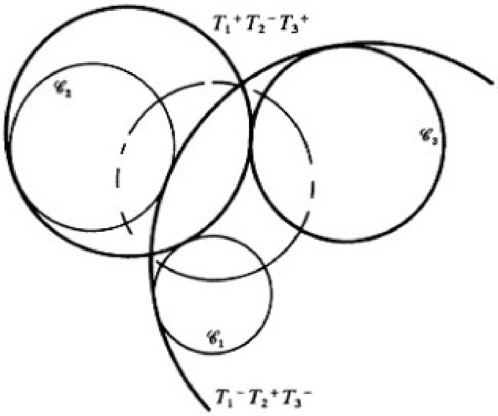

Fig. 26.2

This is a condition we could have reached intuitively, but it is satisfactory to have an algebraic verification. (In §37.1, we shall see that a similar result is true when a point-circle cuts another circle *orthogonally*.)

If the three given circles are each specialized to a point, the condition for a circle $\mathscr{C}$ to touch them is the condition that $\mathscr{C}$ should pass through the three given points. There is only one circle through three points, so we have a case when there is only *one* solution to the problem of Apollonius.

We shall set some special cases, which yield two, three, four, five and six solutions, as Exercises. In the meantime let us see why the number *eight* appears to be the correct number of solutions when the given circles are not specialized. We assume that there is at least one circle which touches the three given circles. Now suppose that we denote a circle (Fig 26.2) which touches $\mathscr{C}_1$ externally by the symbol T_1^+, and one which touches $\mathscr{C}_1$ internally by the symbol T_1^-. Then a combination of the three symbols such as

$$T_1^+T_2^-T_3^+$$

describes a circle which touches $\mathscr{C}_1$ externally, $\mathscr{C}_2$ internally, and $\mathscr{C}_3$ externally. There are *eight* combinations of a triad of plus or minus symbols, namely + + +,

+ + –, + – –, +– +, –++, – +–, – – –, – – +, and therefore if there is a unique circle which corresponds to a triad of the symbols (for instance, a unique circle which touches all of the given circles externally), we expect a total of eight contact circles in all, since there seems no reason, if the circles are not specialized, for one kind of contact to be favoured more than another.

This sounds a rather unmathematical argument, but it can be made to depend on mathematical reasoning. We observe that the distinction between external and internal contact of a circle $\mathscr{C}$ with $\mathscr{C}_1$ vanishes if $\mathscr{C}_1$ becomes a point circle, and we expect, if our reasoning has anything in it, that there are only four circles which pass through a point and touch two given circles $\mathscr{C}_3$ and $\mathscr{C}_3$. In fact, inversion with respect to the point $\mathscr{C}_1$ shows this at once, since the circles $\mathscr{C}_{2'}$ and $\mathscr{C}_{3'}$ have four common tangents, and these arise from the circles through $\mathscr{C}_1$ which touch $\mathscr{C}_2$ and $\mathscr{C}_3$.

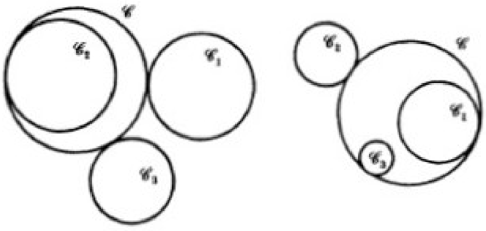

Fig. 26.3

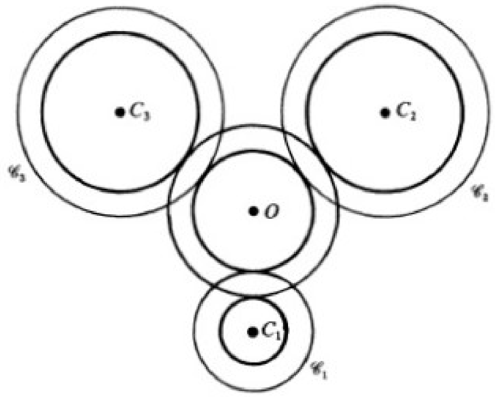

Fig. 26.4

We may think that the case in which $\mathscr{C}_1$ $\mathscr{C}_2$ and $\mathscr{C}_3$ all lie outside each other is more general than the case when, say, $\mathscr{C}_2$ and $\mathscr{C}_3$ lie inside $\mathscr{C}_1$. But if we invert with respect to $\mathscr{C}_1$ this goes into itself, and if $\mathscr{C}_2$ and $\mathscr{C}_3$ lie outside $\mathscr{C}_1$ they invert into circles inside $\mathscr{C}_1$. Again, if we have a circle $\mathscr{C}$ (Fig 26.3) which touches $\mathscr{C}_1$ externally, $\mathscr{C}_2$ internally, and $\mathscr{C}_3$ externally, and we invert with respect to $\mathscr{C}$, the case $T_1^+T_2^-T_3^+$ does turn into $T_1^-T_2^+T_3^-$, since $\mathscr{C}$ inverts into itself, and $\mathscr{C}_1$ which lies outside $\mathscr{C}$, inverts into a circle inside $\mathscr{C}$ the circle $\mathscr{C}_2$ which lies inside $\mathscr{C}$ inverts into a circle outside $\mathscr{C}$ and $\mathscr{C}_3$ which lies outside $\mathscr{C}$ inverts into a circle inside $\mathscr{C}$. Hence if we can show that there is one circle which satisfies the condition $T_1^+T_2^+T_3^+$, and one which satisfies, say, the condition $T_1^-T_2^+T_3^+$, inversions will prove that circles satisfying the other six conditions exist, since we may evidently permute the numbers 1, 2 and 3 in these considerations. (As soon as we have established the existence of a circle Σ orthogonal to $\mathscr{C}_1$,$\mathscr{C}_2$ and $\mathscr{C}_3$ (§28.2), inversion in Σ maps each $\mathscr{C}_1$ onto itself, and can interchange the signs for any given circle condition, but see the remark below.)

We choose our circles $\mathscr{C}_1$,$\mathscr{C}_2$ and $\mathscr{C}_3$ (Fig 26.4) so that the radii of $\mathscr{C}_2$ and $\mathscr{C}_3$ are equal, and so that the center of $\mathscr{C}_1$ is equidistant from the centers of $\mathscr{C}_2$ and $\mathscr{C}_3$. We also choose the radius of $\mathscr{C}_1$ to be less than that of $\mathscr{C}_2$ or $\mathscr{C}_3$. If there is a circle $\mathscr{C}$ of center O which touches the three given circles externally, decreasing the radii of the three circles by the same amount enables us to construct a circle center O which still touches the diminished circles externally. All we have to do is to increase the radius of $\mathscr{C}$ by the amount the other three radii are decreased.

Eventually we arrive at a point circle center C_1, and two equal circles centers C_2 and C_3. If we can draw a circle through C_1 to touch the equal circles centers C_2 and C_3 externally, then we can move back to the circle touching the three original circles externally. If there is only one such circle, we may argue that in a more general case there is only one. We naturally invert with respect to C_1 (Fig 26.5), but we choose as circle of inversion the circle center C_1 which is orthogonal to both circles $\mathscr{C}_2$ and $\mathscr{C}_3$. Since their centers are at equal distances from C_1, and their radii are equal, this is possible. After inversion therefore, the circles centers C_2 and C_3 are unchanged, but the circle through C_1 which touches them has inverted into a common tangent. It is clear that it is the external common tangent shown, the other common tangents inverting into circles through C_1 which do not have external contact with both of the circles $\mathscr{C}_2$ and $\mathscr{C}_3$.

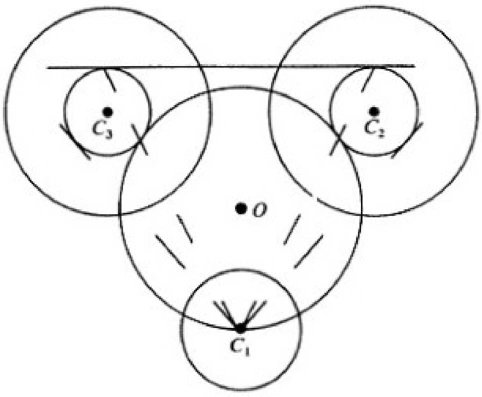

Fig. 26.5

To show that there is one circle satisfying the condition $T_1^-T_2^+T_3^+$ is left as an Exercise.

Remark. Nothing in the above discussion indicates that the eight circles which touch three given circles *always* conform to the pattern we have described. For example, if the centers of $\mathscr{C}_1$, $\mathscr{C}_2$ and $\mathscr{C}_3$ are collinear, the circle orthogonal to the three circles is the line of centers, and inversion in this circle is equivalent to reflexion in the line. It is clear that if there is a $T_1^+T_2^+T_3^+$ contact circle, reflexion in the line of centers produces another contact circle *of the same type*, and similarly all existing types of contact circles occur in pairs. What we have shown above is that there is a case (or there are cases) where our heuristic argument is valid.

Exercises

26.1 Show that a variation of the method given above enables us to assert that there is a unique circle satisfying the condition $T_1^-T_2^+T_3^+$.

26.2 Discuss the types of contact circle given by the other common tangents to the reduced circles $\mathscr{C}_2$ and $\mathscr{C}_3$ in the method for finding a circle with three external contacts described above. In combination with the method of the first Exercise above, are all the eight contact circles produced by the two methods?

26.3 $\mathscr{C}_1$ and $\mathscr{C}_2$ are point-circles, and $\mathscr{C}_3$ is a proper circle. Show that there are two contact circles, touching all three circles.

26.4 $\mathscr{C}_1$ is a point-circle, and $\mathscr{C}_2$ and $\mathscr{C}_3$ are proper circles which touch each other. Show there are three solutions to the problem of Apollonius.

26.5 $\mathscr{C}_1$ and $\mathscr{C}_2$ are proper circles, and $\mathscr{C}_3$ touches both $\mathscr{C}_2$ and $\mathscr{C}_3$ but these latter circles do not touch. Show there are four solutions to the Apollonian problem, or five if we count $\mathscr{C}_1$ in the solutions.

26.6 $\mathscr{C}_1$ and $\mathscr{C}_2$ all touch each other. Show that there are two circles distinct from the given ones which touch the three circles, so that, since each of the given circles is a solution of the problem, there are five solutions in all.

26.7 $\mathscr{C}_1$ and $\mathscr{C}_2$ touch, and $\mathscr{C}_3$ is a proper circle which does not touch $\mathscr{C}_1$ or $\mathscr{C}_2$. Show that there are six distinct circles which touch the three given circles.

†**26.8** Can you find a case when there are seven distinct solutions? (None of the many papers on the problem of Apollonius mentions this case specifically. Can you prove that such a case is impossible?)

26.9 If $\mathscr{C}_1$ contains both $\mathscr{C}_2$ and $\mathscr{C}_3$ discuss the possible types for the eight contact circles.

26.10 l and m are two parallel tangents to a circle $\mathscr{C}$ and a further tangent n intersects l in P and m in Q. Prove that the circle on PQ as diameter contains the center of $\mathscr{C}$.

Deduce by inversion that if three equal circles $\mathscr{C}_1, \mathscr{C}_2$ and $\mathscr{C}_3$ pass through the same point O, $\mathscr{C}_1$ and $\mathscr{C}_2$ touching at O, then if $\mathscr{C}_1$ intersects $\mathscr{C}_3$ again at P, and $\mathscr{C}_2$ intersects $\mathscr{C}_3$ again at Q, PQ is a diameter of $\mathscr{C}_3$. (There is also a more direct and simple proof.)

---

† See Pedoe, D., '*The Missing Seventh Circle', Elemente der Mathematik*, **25**, Jan. 1970, p. 14–15.

Since

$$L_{12} + L_{23} \equiv C_1 - C_3 = 0 \equiv L_{13}$$

the radical axis L_{13} of $\mathscr{C}_1$ and $\mathscr{C}_3$ passes through the intersection of L_{12} and L_{23}, if they have an intersection, or L_{13} is parallel to both lines, if they are parallel.

28.2 The circle orthogonal to three given circles

The theorem of the preceding Section shows us that there is at least one point, finite or at infinity, which has the same power with respect to three given circles $\mathscr{C}_1$ $\mathscr{C}_2$ $\mathscr{C}_3$(Fig. 28.2). If the three circles belong to the same coaxal system, every point on the common radical axis has this property. If the circles are not in the same coaxal system there is a unique point with this property. It is called *the radical center* of the three circles. If it lies outside one, and therefore each of the three circles, and the point be designated as V, and the common power as $|VT|^2$, a circle radius $|VT|$ and center V will cut each of the three circles $\mathscr{C}_i$ orthogonally. If the radical center lies inside one of the circles, and therefore inside each of them, and the common power is $-|VT|^2$, the common orthogonal circle is a *virtual* circle, with equation

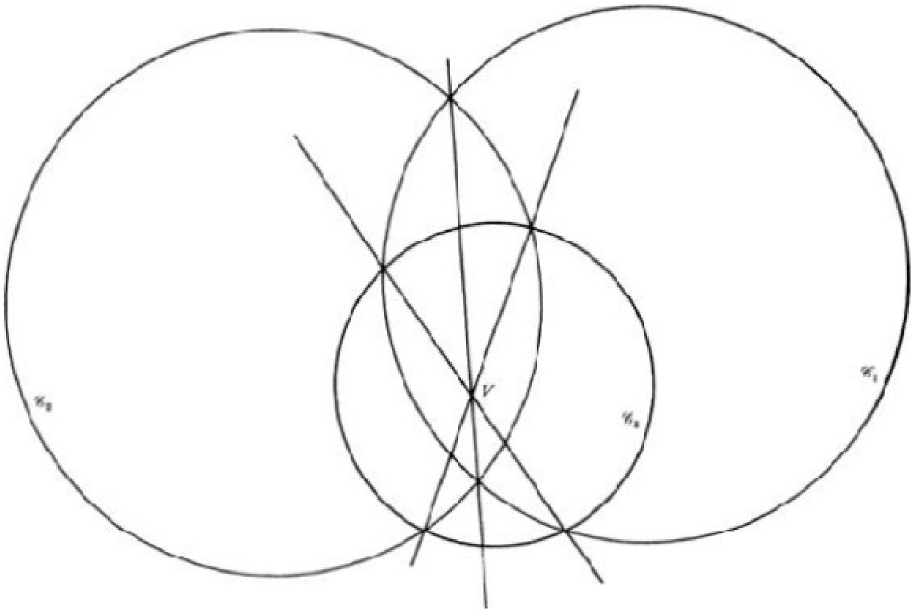

Fig. 28.2

$$(X-a)^2+(Y-b)^2+r^2=0,$$

where $r^2=|VT|^2$, and the coordinates of V are (a, b). Algebraically, this virtual circle satisfies the condition for orthogonality with each of the circles $\mathscr{C}_i$.

The common orthogonal circle, if taken as circle of inversion, inverts each circle $\mathscr{C}_i$ into itself, and, in particular, inverts any circle $\mathscr{C}$ which touches the three circles into a distinct circle $\mathscr{C}'$ which touches the three circles. We assume that $\mathscr{C}$ is not also orthogonal to the common orthogonal circle. Hence we expect the circles which touch the three given circles to fall into pairs, mutually inverse in the common orthogonal circle. (See §26.1 and §38.1.) The consideration of inversion in a virtual circle arises quite naturally in this context.

When the centers of the three circles $\mathscr{C}_i$ are collinear, the common orthogonal circle is a line, the line of centers, and the radical center is at infinity. Inversion in the common orthogonal circle is reflexion in the line of centers in this case.

Exercises

28.1 If $\mathscr{C}_1$, $\mathscr{C}_2$ are two non-intersecting circles, and $\mathscr{C}_3$ intersects $\mathscr{C}_1$ in the points P, Q and $\mathscr{C}_2$ in the points R, S, prove that the point $PQ \cap RS$ lies on the radical axis of $\mathscr{C}_1$ and $\mathscr{C}_2$.

28.2 A circle drawn through two given points P, P' intersects a given circle $\mathscr{C}$ in the points Q and R. Prove that the lines QR form a pencil, that is, they either all pass through a point, or they are all parallel to each other. Hence prove that two circles through P and P' touch $\mathscr{C}$, and show how to construct them.

28.3 Prove the following construction, due to Gergonne, for a circle which touches three given circles $\mathscr{C}_1$, $\mathscr{C}_2$ and $\mathscr{C}_3$. Draw the circle which is orthogonal to the three given circles. (See §28.2.) Draw the three chords of intersection of this circle with $\mathscr{C}_1$, $\mathscr{C}_2$ and $\mathscr{C}_3$ respectively. From the points where these chords meet an axis of similitude of the three circles $\mathscr{C}_1$, $\mathscr{C}_2$, $\mathscr{C}_3$ (See Exercise 13.2) draw pairs of tangents to $\mathscr{C}_1$, $\mathscr{C}_2$ and $\mathscr{C}_3$. Then the two circles drawn through these six points of contact will touch $\mathscr{C}_1$, $\mathscr{C}_2$ and $\mathscr{C}_3$. (We shall see later, in §38.1, an algebraic verification of the theorem that the circles touching three given circles $\mathscr{C}_1$, $\mathscr{C}_2$ and $\mathscr{C}_3$ are inverse in pairs in the common orthogonal circle of the three given circles. But since circles are invariant under inversion in a circle orthogonal to them, the circles $\mathscr{C}_1$, $\mathscr{C}_2$ and $\mathscr{C}_3$ invert into themselves under inversion in their common orthogonal circle, and a circle which touches them at A, B and C say inverts into a circle which touches them at A', B' and C', where the primes denote points inverse to A, B and C in the common orthogonal circle. Hence circle ABC, the orthogonal circle, and circle $A'B'C'$ are in a coaxal system (Exercise 21.6). Therefore, by Thm 28.1, the radical axes of $\mathscr{C}_1$ and each of the three coaxal circles we are discussing meet in a point. Two of the radical axes are tangents at A and A' to $\mathscr{C}_1$, and the third is the common chord of $\mathscr{C}_1$ and the common orthogonal circle. Finally, if P is the point of intersection of these three lines, P must be on an axis of similitude (See Exercise 13.2) of the three given circles, because $|PA| = |PA'|$ shows that P lies on the radical axis of circles ABC, $A'B'C'$, and by Exercise 18.3 we know that if two circles touch two others tdical axis of either pair passes through a center of similitude of the other pair, and since circles ABC, $A'B'C'$ touch $\mathscr{C}_1$, $\mathscr{C}_2$and $\mathscr{C}_3$, the radical axis of the two touching circles passes through three centers of similitude: it is an axis of similitude of $\mathscr{C}_1$, $\mathscr{C}_2$ and $\mathscr{C}_3$

28.4 Since three circles have four axes of similitude, the above construction yields eight Apollonian circles (compare §26.1). Draw them.

28.3 Another proof of the nine-point circle theorem

In §12.1 we saw that if A, B and C are the vertices of a triangle, D', E' and F' the feet of the perpendiculars from the vertices onto the opposite sides, and α, β, γ the midpoints of BC, CA and AB respectively, then there is a circle which passes through these six points and also the three midpoints of HA, HB and HC, where H is the orthocenter, the intersection of AD', BE' and CF' (Fig. 28.3). We remark that $|CD'|.|CB| = |CE'| . |CA|$, since a circle on AB as diameter passes through D' and E'. We deduce that $|C\alpha| . |CD'| = |C\beta| . |CE'|$, by halving each side of the equation. Hence the points α, D', β and E' are concyclic. Let them lie on a circle $\mathscr{C}_3$. Similarly the points α, D', γ, F' are concyclic. Let them lie on a circle $\mathscr{C}_2$. If the circles $\mathscr{C}_3$ and $\mathscr{C}_2$ do not coincide, they have the chord of intersection $\alpha D'$, which is the line BC. The points γ, F', β, E' are concyclic. Let them lie on a circle $\mathscr{C}_1$. If $\mathscr{C}_1 \neq \mathscr{C}_2$, their chord of intersection is AB, and if $\mathscr{C}_1 \neq \mathscr{C}_3$, their chord of intersection is AC. Hence, if the circles $\mathscr{C}_1$, $\mathscr{C}_2$and $\mathscr{C}_3$ are distinct, their radical axes, taking the circles in pairs, form the sides of a triangle. This contradicts Thm 28.1. Hence the points α, β, γ, D', E' and F' all lie on one circle. As in §12.1, by applying this result to the triangle BHC, we show that this circle also contains the midpoints of HA, HB and HC.

touching circles passes through three centers of similitude: it is an axis of similitude of $\mathscr{C}_1$, $\mathscr{C}_2$ and $\mathscr{C}_3$

28.4 Since three circles have four axes of similitude, the above construction yields eight Apollonian circles (compare §26.1). Draw them.

28.3 Another proof of the nine-point circle theorem

In §12.1 we saw that if A, B and C are the vertices of a triangle, D', E' and F' the feet of the perpendiculars from the vertices onto the opposite sides, and α, β, γ the midpoints of BC, CA and AB respectively, then there is a circle which passes through these six points and also the three midpoints of HA, HB and HC, where H is the orthocenter, the intersection of AD', BE' and CF' (Fig. 28.3). We remark that $|CD'|.|CB| = |CE'| \,.\, |CA|$, since a circle on AB as diameter passes through D' and E'. We deduce that $|C\alpha| \,.\, |CD'| = |C\beta| \,.\, |CE'|$, by halving each side of the equation. Hence the points α, D', β and E' are concyclic. Let them lie on a circle $\mathscr{C}_3$. Similarly the points α, D', γ, F' are concyclic. Let them lie on a circle $\mathscr{C}_2$. If the circles $\mathscr{C}_3$ and $\mathscr{C}_2$ do not coincide, they have the chord of intersection $\alpha D'$, which is the line BC. The points γ, F', β, E' are concyclic. Let them lie on a circle $\mathscr{C}_1$. If $\mathscr{C}_1 \neq \mathscr{C}_2$, their chord of intersection is AB, and if $\mathscr{C}_1 \neq \mathscr{C}_3$, their chord of intersection is AC. Hence, if the circles $\mathscr{C}_1$, $\mathscr{C}_2$and $\mathscr{C}_3$ are distinct, their radical axes, taking the circles in pairs, form the sides of a triangle. This contradicts Thm 28.1. Hence the points α, β, γ, D', E' and F' all lie on one circle. As in §12.1, by applying this result to the triangle BHC, we show that this circle also contains the midpoints of HA, HB and HC.

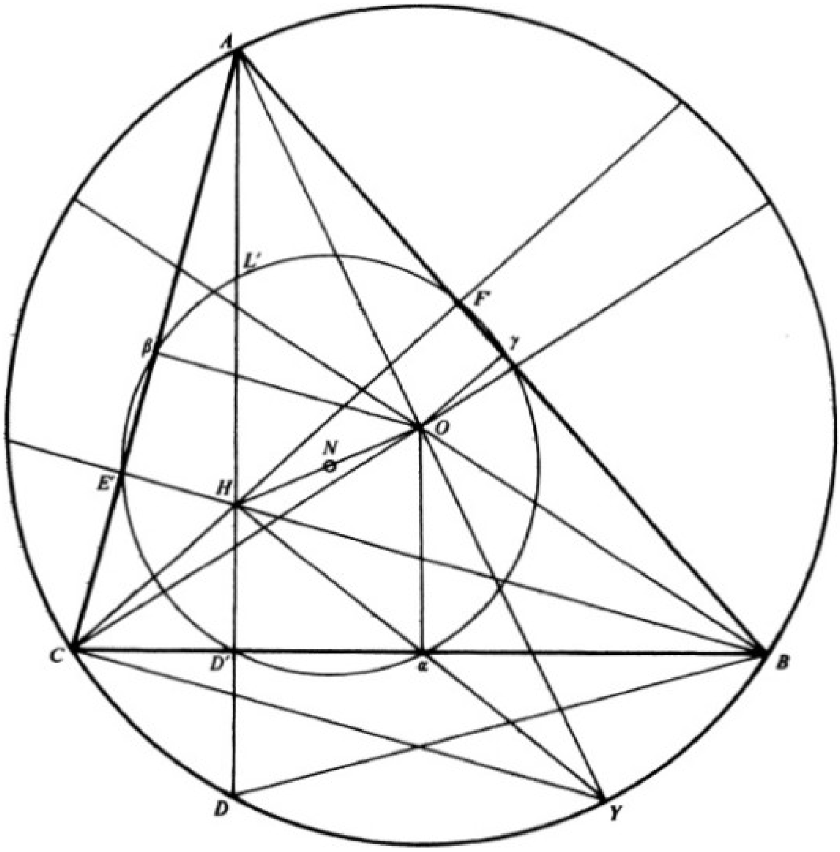

Fig. 28.3

28.4 Clerk Maxwell's derivation of reciprocal figures

We use the notation of §6.1. Let ABC be a triangle, and D a point not on the sides. The triangle PQR is such that QR is parallel to DA, RP is parallel to DB and PQ is parallel to DC. We wish to show the existence of a point F such that BC is parallel to FP, CA is parallel to FQ and AB is parallel to FR.

Maxwell, who was a geometer as well as a great physicist, drew circles with centers at the points A, B, C and D respectively (Fig. 28.4), and remarked that the chord of intersection of each pair of circles is perpendicular to the line joining the centers, and that the chords of intersection of the pairs which can be taken from three circles pass through one point (Thm 28.1). Hence the four

*.

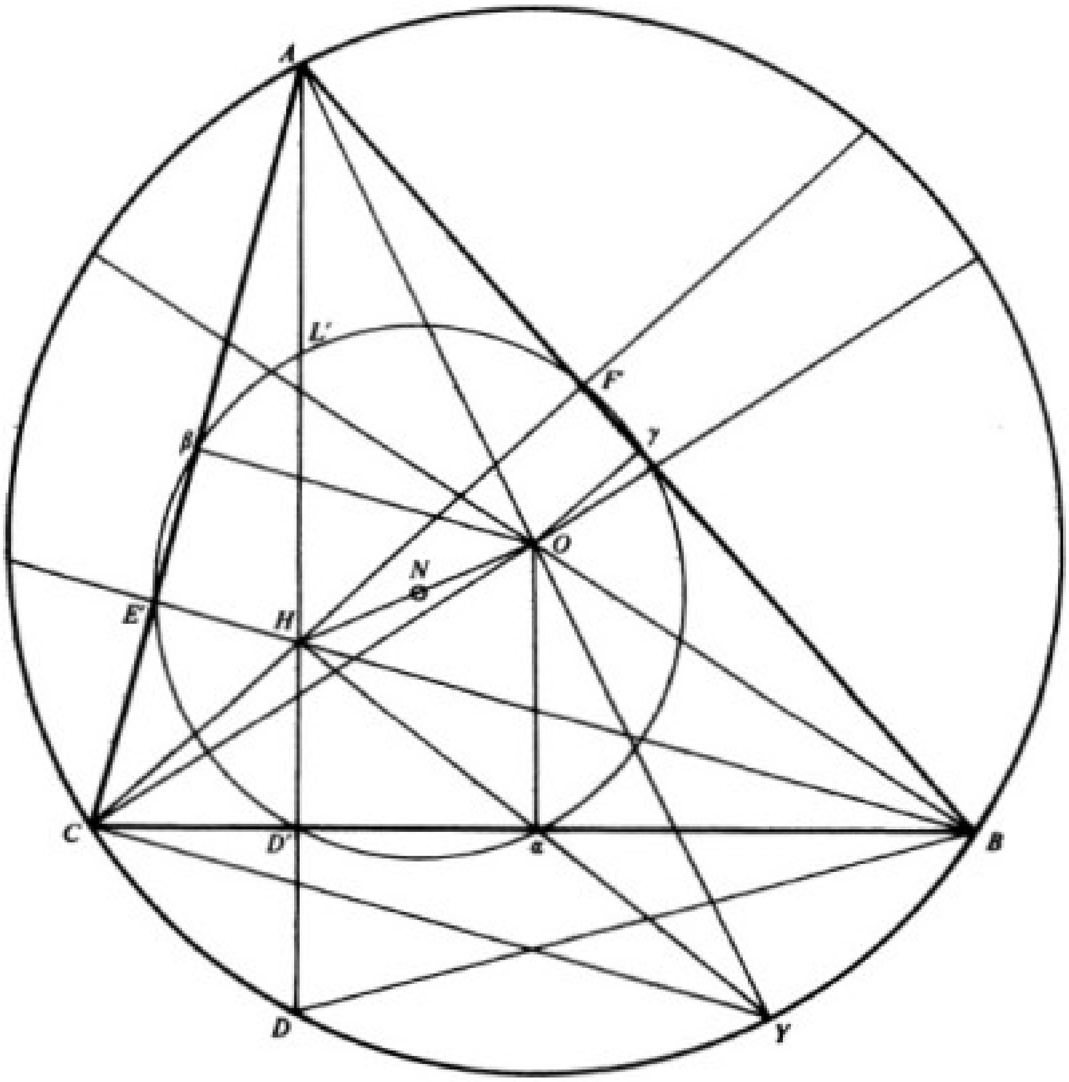
A
L'
F'
β
γ
O
N
H
E'
C
D'
α
B
D
Y

/ .

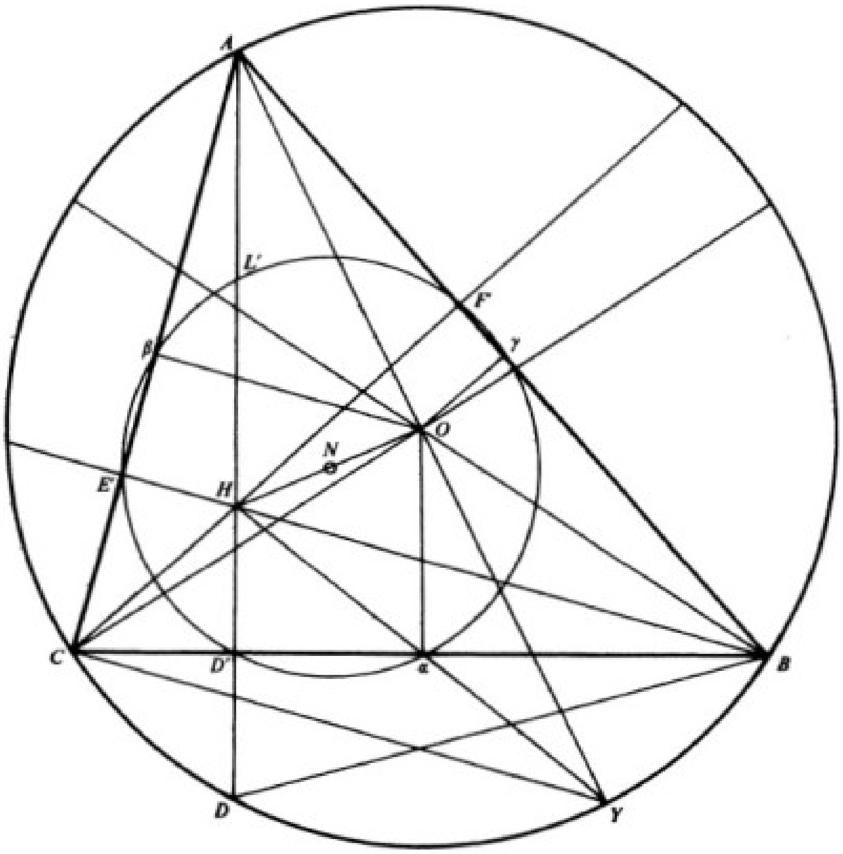

Fig. 28.3

28.4 Clerk Maxwell's derivation of reciprocal figures

We use the notation of §6.1. Let ABC be a triangle, and D a point not on the sides. The triangle PQR is such that QR is parallel to DA, RP is parallel to DB and PQ is parallel to DC. We wish to show the existence of a point F such that BC is parallel to FP, CA is parallel to FQ and AB is parallel to FR.

Maxwell, who was a geometer as well as a great physicist, drew circles with centers at the points A, B, C and D respectively (Fig. 28.4), and remarked that the chord of intersection of each pair of circles is perpendicular to the line joining the centers, and that the chords of intersection of the pairs which can be taken from three circles pass through one point (Thm 28.1). Hence the four circles constructed have common chords which pass in threes through four points *A'*, *B'*, *C'* and *D'*, where *B'C'* is perpendicular to *DA*, *C'A'* is perpendicular to *DB*, *A'B'* is perpendicular to *DC*, and *D'A'* is perpendicular to *BC*, *D'B'* is perpendicular to *CA* and *D'C'* is perpendicular to *AB*.

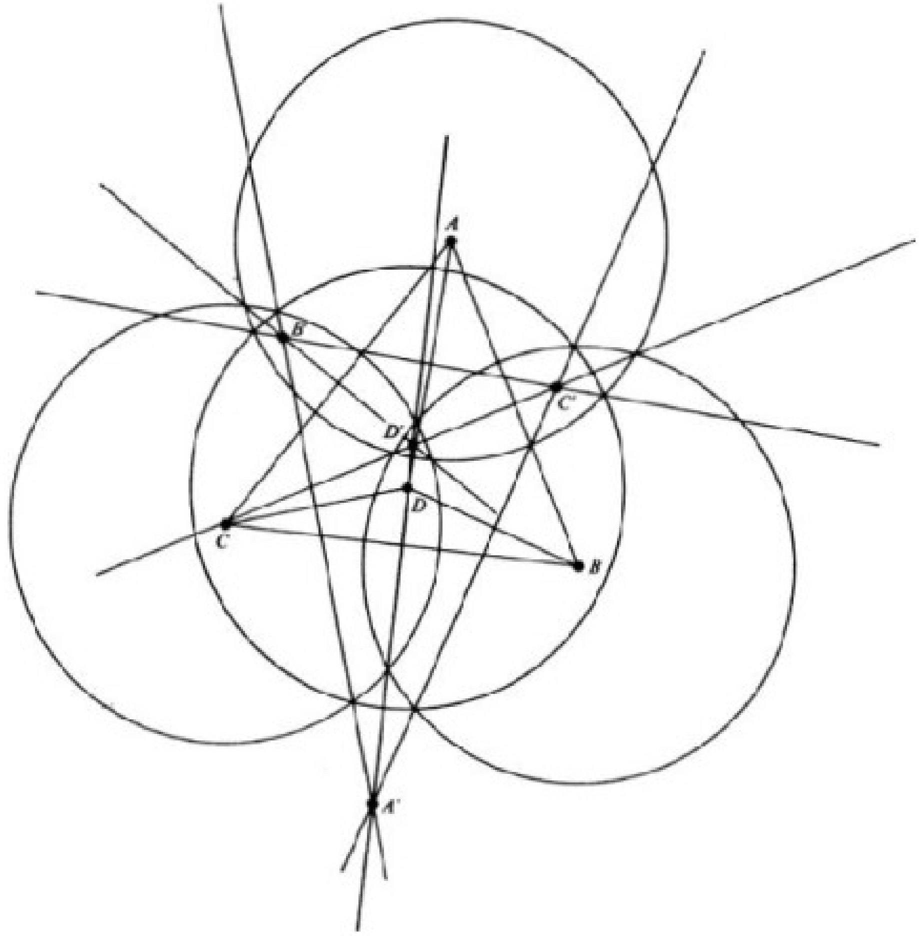

Fig. 28.4

If we now rotate the figure *A'B'C'D'* through a right angle, the figure reciprocal to *ABCD* is in existence! The triangle *A'B'C'* is similar to the given triangle *PQR*, and the point *F* is immediately obtained from the point *D'* which corresponds to it in triangle *A'B'C'*. Through *P* we draw a line parallel to *BC*, and through *Q* a line parallel to *CA*. The intersection *F* of these two lines will be such that *FR* is parallel to *AB*.

*.

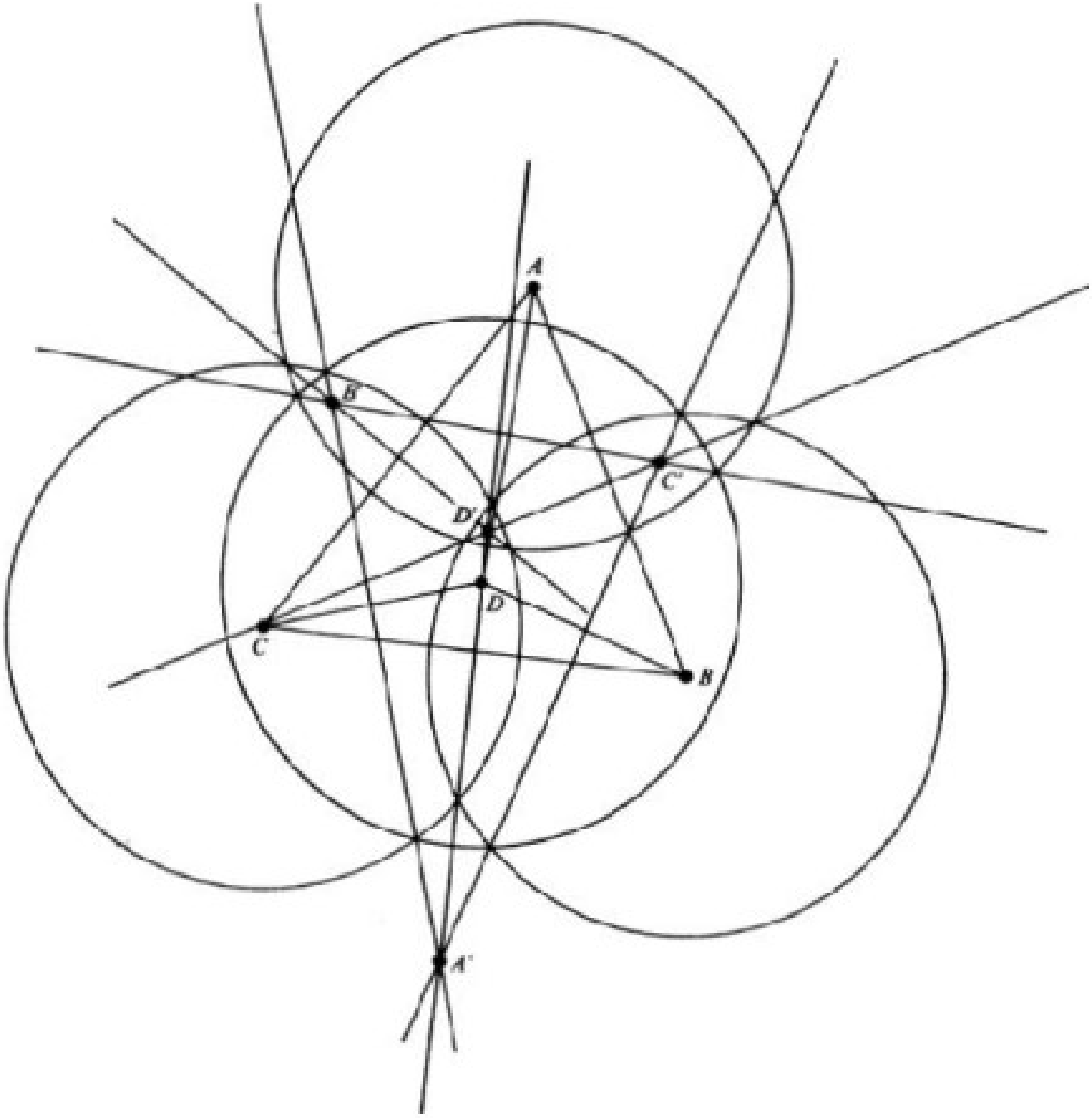
A
B'
C'
D'
D
C
B
A'

/ .   :

V

MAPPINGS OF THE EUCLIDEAN PLANE

In this chapter we propose to investigate the fundamental bijective (one-to-one onto) mappings of the Euclidean plane, using algebraic methods, and, in particular, complex numbers. Some of these mappings have already appeared in purely geometric guise (§12.1, §14.1). Some of the proofs given previously may seem to depend on the acceptance of the correctness of diagrams, and it is important to show that algebraic proofs are also possible. The algebra of complex numbers is very useful in this connection, but we shall try not to lose sight of the geometrical meaning of theorems and, in fact, it will be evident in many cases that the algebra is motivated by the geometry.

In §0.4 we gave the properties of complex numbers which we shall be using. We saw that we may regard the Euclidean plane as a representative of the field of complex numbers, the point (a, b) being associated with the complex number $a + ib$. When the Euclidean plane is regarded in this way we shall refer to it as *the Gauss plane*. Complex numbers will be denoted by lower-case letters, such as z, a, b, etc., and although we have just mapped the point (a, b), where a, b are real numbers, onto the complex number $z = a + ib$, there will be no misunderstanding if we talk of the mapping $z' = az$, where $|a| = 1$, it being understood that a is a complex number in this case. We shall naturally talk of ‘the point z' instead of ‘the point of the Euclidean plane associated with the complex number z'.

41.1 Translations

Let b be a fixed complex number, and consider the mapping

$$z' = z + b.$$

It is clear that every point z is moved through a fixed distance and in a fixed direction, given by the complex number b regarded as a vector. Such a mapping is called a *translation* (see Fig. 41.1). If $b = 0$, the mapping is the identity mapping.

41.2 Rotations

The mapping $z' = (\cos\varphi + i\sin\varphi)z$, leaves the modulus of z unchanged, but increases its amplitude by φ, and therefore moves the point representing z around the circle, center at the origin, which passes through the point z, and through a fixed angle φ. This is *a rotation about the origin* (see Fig. 41.2).

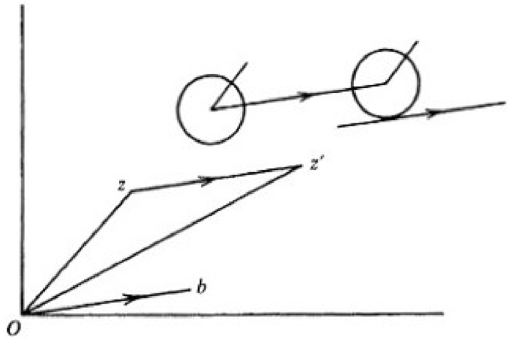

Fig. 41.1

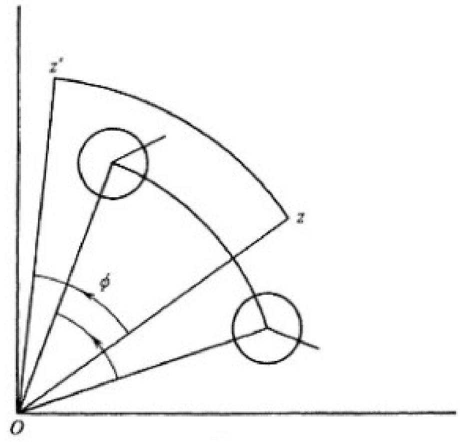

Fig. 41.2

This mapping may also be written $z' = az$, where $|a| = 1$, so that $a = \cos\varphi + i\sin\varphi$. If $a = 1$, we obtain the identity mapping.

41.3 Reflexions

The mapping $z' = \bar{z}$ is a reflexion in the X_1-axis (Fig. 41.3). If we require a reflexion in the line which passes through the origin, and makes an angle α with the X_1-axis, it is given by $z' = (\cos 2\alpha + i \sin 2\alpha)\bar{z}$, since we have at once $|z| = |z'|$, and $am\, z' + am\, z = 2\alpha$, so that the given line always bisects the angle between the joins of the origin to the points z and z'. The difference between the mappings when the map is a rotation and when it is a reflexion should be noted.

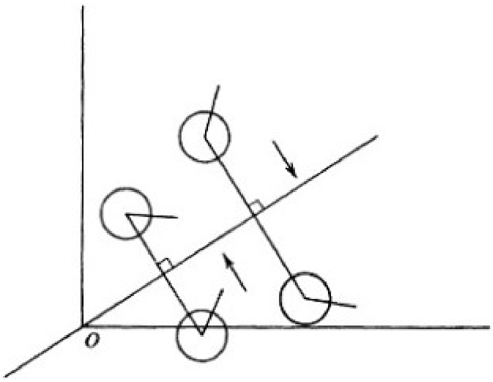

Fig. 41.3

41.4 Inversion

This is not a mapping of the Gauss plane, but of the plane with the center of inversion removed (Fig. 41.4). We shall overcome this difficulty in the next chapter, but in the meantime we remark that $z' = 1/z$ is equivalent to inversion in the unit circle, center at the origin, followed by reflexion in the X_1-axis. For if $z = r(\cos \theta + i \sin \theta), z^{-1} = r^{-1}(\cos (-\theta) + i \sin (-\theta))$, and the inverse of z is the point $r^{-1}(\cos (\theta) + i \sin (\theta))$. If we wish to obtain the inverse of z directly, we write $z' = 1/\bar{z}$.

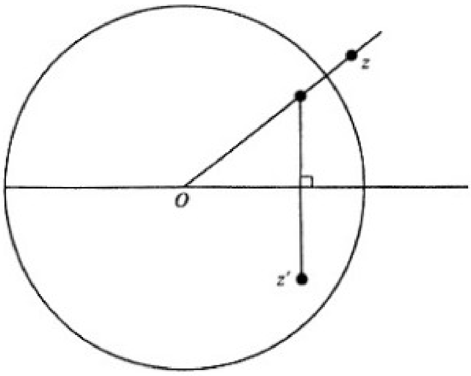

Fig. 41.4

41.5 Central dilatations

We can combine mappings (Fig. 41.5). Let us consider the mapping $z' = az$, where

$$a = |a|(\cos \alpha + i \sin \alpha).$$

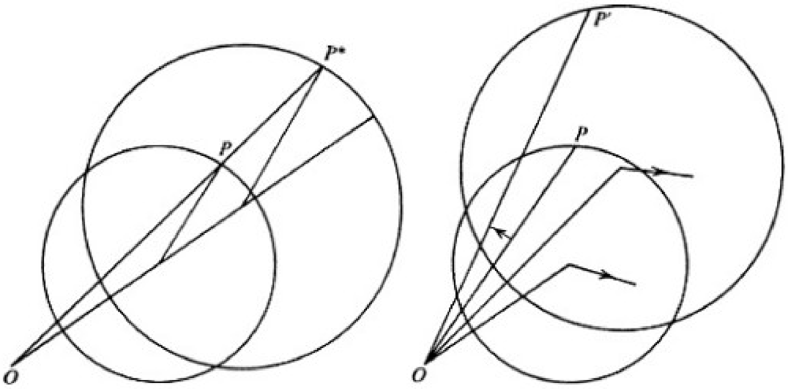

Fig. 41.5

Then $z = r(\cos \theta + i \sin \theta)$ is mapped on

$$z' = |a|r(\cos(\theta + \alpha) + i\sin(\theta + \alpha)).$$

The point P representing z is stretched to the point P^*, where $\overline{OP^*} = |a|\overline{OP}$, (that is, stretched in a fixed ratio), and OP^* is then rotated about O through the fixed angle α to OP', where P' represents z'.

We note that the order in which the two operations, stretching and rotation, are carried out does not affect the final result. That is, they are *commutative* operations. If a is real, so that $\alpha = 0$, the mapping is called either a *central similarity*, or a *central dilatation*, or merely a *magnification* or *enlargement*, but if rotation is also involved, it is called a *dilative rotation* about O, or a *spiral similarity*. Each of these terms is fairly easy to remember. We encountered central dilatations in §12.1, reflexions in a line in §14.1, rotations in the same section, and inversions in §20.1. Translations, of course, are implicit in the equivalence relation defined for vectors in §1.1. A free vector is obtained from a bound vector by means of a translation.

All of the mappings mentioned above can be used to solve certain types of geometrical problem. We have already seen the power of inversion in the § 20.1–onwards, and we have used reflexions to solve certain problems and to prove special theorems. The Exercises below can be solved by the use of the mappings we have discussed. We omit Exercises which need inversion.

Exercises

41.1 Two circles $\mathscr{C}_1$ and $\mathscr{C}_2$ are given, and a line l. Find a point A on $\mathscr{C}_1$ and a point B on $\mathscr{C}_2$ such that AB is parallel to l, and the distance $AB = d$, where d is given. (What is the map of a given circle under a translation?)

41.2 Two circles $\mathscr{C}_1$ and $\mathscr{C}_2$ are given, and a line l. Construct squares $ABCD$ with A and C on l, with B on $\mathscr{C}_1$, and D on $\mathscr{C}_2$. (What is the reflexion of a given circle in a given line?)

41.3 Show how to construct an equilateral triangle which has one vertex at a given point A, and one vertex on each of two given lines l and m. (What locus do we obtain if we rotate a given line through a given angle A about a given point?)

41.4 A point A and lines l and m are given. Show how to construct a square $ABCD$, with vertex A at A, vertex B on l, and vertex C on m. (What is the locus of l under the spiral similarity, center A, given by the equation $z' = az$, where $|a| = \sqrt{2}\,1$ and $am\ a = + \pi/4$, or $- \pi/4$?)

41.5 Show that a central dilatation, $z' = pz$, where p is real, maps a line onto a parallel line. Show also that if $z'' = qz'$ be another central dilatation with the same center, this maps the point z' onto the point $z'' = qpz$. Show that if the two mappings be combined, with the second first, we obtain $z'' = pqz$ as a result of the composition, so that the final result is independent of the order in which the mappings are combined. Use this result to obtain the parallel case of the Pappus Theorem of §9.5. (The points A, B, C lie on l, and the points A', B', C' lie on m. Given that AC' is parallel to CA', and CB' is parallel to BC', we wish to prove that AB' is parallel to BA'. Use central dilatations with center at $l \cap m$. Let $C = AT_1$, so that $A' = C'T_1$, and let $B = CT_2$, so that $C' = B'T_2$. Show that $B = AT_1T_2$, and $A' = B'T_2T_1$. From $T_1T_2 = T_2T_1$, deduce the theorem.)

42.1 Isometries

We note that the mappings: *translation*, *rotation* and *reflexion*, are all such that the distance between two points is unchanged by one of these mappings. If the two points are z and w, a translation gives

$$|z' - w'| = z + b - (w + b)| = |z - w|,$$

a rotation gives and reflexion in a line through the origin, which we can write $z' = a^{\bar{z}}$, where $|a| = 1$, gives so that our assertion is proved in all three cases.

$$\begin{aligned} |z' - w'| &= |(\cos\varphi + i\sin\varphi)z - (\cos\varphi + i\sin\varphi)w| \\ &= |(\cos\varphi + i\sin\varphi)||(z - w)| = |z - w|, \end{aligned}$$

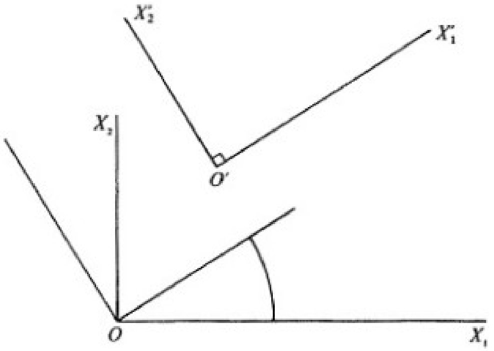

Fig. 42.1

$$|z' - w'| = |a\overline{(z - w)}| = |z - w|,$$

We know that inversion and central dilatations do change the distance between a pair of points.

Definition. An *isometry* of the Gauss plane is a one-to-one onto mapping of the plane such that if P is mapped on P' and Q is mapped on Q', the distance PQ = distance $P'Q'$, whatever the points P and Q.

Such mappings are also called *congruent transformations*, or *rigid motions*. The mappings: *translation*, *rotation* and *reflexion* are all isometries, and compositions of them are naturally also isometries. Two important ones must be familiar to the reader.

42.2 Coordinate transformations

We assume that our coordinate axes are orthogonal cartesian, and we *rotate* them through an angle φ about the origin, and in a positive direction, from the

X_1-axis towards the X_2-axis (Fig. 42.1).

We then *translate* the new position of the axes so that the original origin O falls on the point $O' = b$, where b is a complex number referred to the original axes. Calling the original axes OX_1, OX_2, call the new position of the axes $O'X_1'$, $O'X_2'$.

After the rotation the point which was originally z with respect to the original axes has become

$$z(\cos(-\varphi) + i\sin(-\varphi))$$

with respect to the axes rotated through φ about O, and the point b has become $b(\cos(-\varphi) + i\sin(-\varphi))$ with respect to these axes. After the translation, the point z becomes, with respect to the axes $O'X_1'$, $O'X_2'$:

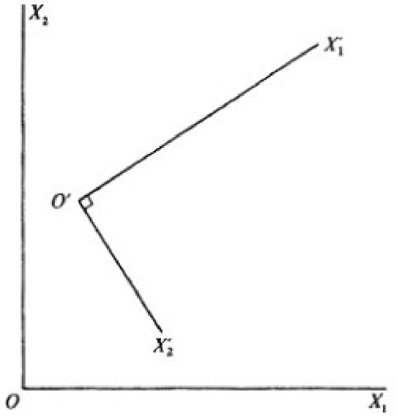

Fig. 42.2

$$Z' = z(\cos(-\varphi) + i\sin(-\varphi)) - b(\cos(-\varphi) + i\sin(-\varphi)),$$

which we may write

$$Z' = az + c, \text{ where } |a| = 1.$$

The same point now has two different names, or aliases: z with respect to the original axes, and $Z' = az + c$ with respect to the new axes.

If we suppose that the coordinate axes remain unchanged, and consider the mapping $Z' = az + c$ ($|a| = 1$), we are not surprised to find that this is also an isometry. In fact, if z and w are mapped on Z' and W',

$$|Z' - W'| = |az + c - aw - c| = |a||z - w| = |z - w|.$$

In our transformation of coordinate axes above, we *preserved* the *orientation* of the axes: that is, the sense of the rotation from $O'X_1'$ towards $O'X_2'$ is the same as that of OX_1 towards OX_2. If, however, we *reverse* the direction of *one* of the new axes (Fig. 42.2), say $O'X_2$, the point z with respect to OX_1, OX_2 now has the form

$$Z' = a\bar{z} + c \qquad (|a| = 1)$$

with respect to the new axes $O'X_1'$, $O'X_2'$.

Once again, if we suppose the coordinate axes to remain unchanged, and consider the mapping $Z' = a^{\bar{z}} + c$ ($|a| = 1$), we find that it is also an isometry, since

$$|Z' - W'| = |a\bar{z} + c - a\bar{w} - c| = |a||\bar{z} - \bar{w}| = |\overline{z - w}| = |z - w|$$

We shall find that these two isometries of the Gauss plane, which arise in this natural way by coordinate transformations, are the fundamental isometries of the Gauss plane, and there are no others.

Exercises

42.1 The X_1 and X_2-axes are interchanged, so that if $z = a + ib$, $Z = b + ia$ after the transformation. Show that the following composition of coordinate transformations produces the above result: $Z_1 = iz = -b + ia$, $Z_2 = \bar{z}_1 = -b - ia$, $Z = -Z_2 = b + ia$. Show also that the composition of these transformations is equivalent to $Z = i\bar{z}$.

42.2 Where are the new coordinate axes which induce the coordinate transformations

$$\text{(i)}Z = \bar{z} \ \text{(ii)}Z = i\bar{z} \ \text{(iii)}Z = -i\bar{z} + 1?$$

42.3 Find the coordinate transformation produced if the new system has its origin at the point $1 + i$, the new negative X_1-axis passes through O, and the new positive X_2-axis intersects the original positive X_2-axis. In the same circumstances, find the transformation if the new positive X_2-axis intersects the old positive X_1-axis.

43.1 The main theorem on isometries of the Gauss plane

Theorem *The set* $\mathscr{I}$ *of all isometries of the Gauss plane is composed of two classes* $\mathscr{I}_+$ *and* $\mathscr{I}_-$. *The class* $\mathscr{I}_+$ *consists of all isometries of the form* $z' = az + b$ $(|a| = 1)$, *and the class* $\mathscr{I}_-$ *of all isometries of the form* $z' = c\bar{z} + d$ $(|c| = 1)$.

We recall that an isometry is a one-to-one onto mapping of the Gauss plane which preserves the distance between a pair of points, and that such a mapping is also called a *congruent* or *rigid* transformation, or *motion*, if the plane is thought of as being moved over itself.

We have already seen that the mappings $\mathscr{I}_+$ and $\mathscr{I}_-$ are such that distance is preserved, but we have not checked that both are one-to-one onto mappings. In each case this is immediate, since if $z' = az + b$, z' is uniquely defined by z, and $z = a^{-1}(z' - b)$, so that, since $a \neq 0$, z is uniquely defined by z'. If $z' = c\bar{z} + d$, z' is uniquely defined by z, and $\bar{z} = c^{-1}(z' - d)$, so that $a \neq 0$, z is uniquely defined by z'.

We have some auxiliary theorems to prove before we establish the main theorem, Thm 43.1.

43.2 An auxiliary theorem

Theorem *Given two pairs of points* z_0, z_1 *and* w_0, w_1, *where* $|z_0 - z_1| = |w_0 - w_1| \neq 0$, *there is just one mapping of type* $\mathscr{I}_+$ *and just one of type* $\mathscr{I}_-$ *which maps* z_0 *onto* w_0 *and* z_1 *onto* w_1.

Proof. All that we have to do is to determine the constants a, b for a mapping of the first kind, and the constants c, d for a mapping of the second kind. Since we must have $az_0 + b = w_0$, and $az_1 + b = w_1$, where $|a| = 1$, subtraction gives us $a(z_0 - z_1) = w_0 - w_1$, which defines a, since $z_0 - z_1 \neq 0$, and also gives us a value of a for which $|a| = 1$. Knowing the value of a, we substitute, and find that $b = w_0 - az_0$. Hence a and b are both uniquely determined.

The proof that c and d are uniquely determined is very similar.

One of the things which interests us for any given mapping of the Gauss plane is whether points which are collinear are mapped into points which are collinear.

Definition. A *collineation* of the Gauss plane is a mapping which preserves collinearity, so that straight lines are mapped into straight lines.

A collineation need not be an isometry, as we see from the example of a central dilatation. But if we consider isometries, we have:

43.3 Isometries are collineations

Theorem *Every isometry of the Gauss plane is a collineation.*

Proof. In the real Euclidean plane we have a very powerful notion, derived from that of distance, namely the notion of a point lying on a line *between* two distinct points. If u, v and w are the complex numbers attached to three distinct collinear points, we know that v lies on the line joining the two other points, and between them, if and only if

$$|v - u| + |w - v| = |w - u|.$$

If under an isometry the points become u', v' and w' respectively, we have

$$|v' - u'| + |w' - v'| = |w' - u'|.$$

and since these points are also in the real Euclidean plane, the points represented by the complex numbers u', v' and w' must necessarily lie, in this order, on a line.

We deduce from this theorem that although Thm 43.2 tells us that there are at least two isometries which map z_0 on w_0 and z_1 on w_1, where $|z_0 - w_0| = |z_1 - w_1|$, each mapping maps a point of the line joining z_0 and z_1 onto a point of the line joining w_0 to w_1, and the effects of the mappings on points of the object line are identical.

43.4 Isometries and parallel lines

Theorem *An isometry of the Gauss plane maps parallel lines onto parallel lines.* (See Fig. 43.1.)

Proof. If P is a point in the Gauss plane, and l a line which does not contain P, the minimum distance of P from the line is well defined, and given by PR, where R is the foot of the perpendicular from P onto the line. Under an isometry, if P is mapped onto P', the line l onto a line l', and R onto R', the point R' lies on l', and $P'R'$ is the minimum distance of P' from l'. Hence $P'R'$ is perpendicular to l'.

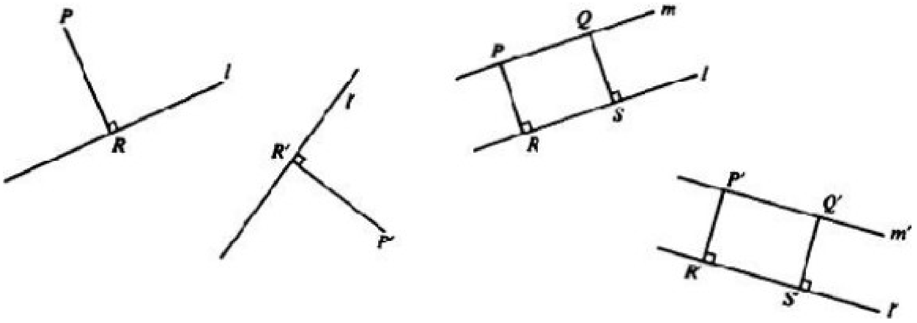

Fig. 43.1

Given two parallel lines *l* and *m*, take two distinct points *P* and *Q* on *m*, and let *R* and *S* be the respective feet of perpendiculars from *P* and *Q* onto the line *l*. Then $PR = QS$, and the isometry maps *l* and *m* into lines *l*′ and *m*′ which are likewise equidistant from each other, and therefore parallel.

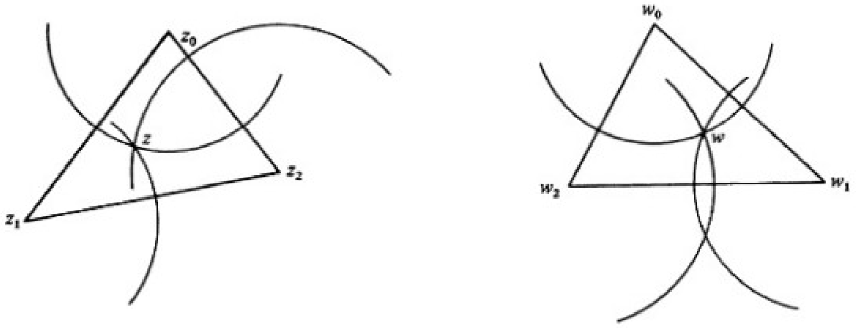

Fig. 43.2

Alternatively, and more rapidly, we may argue that if the maps *l*′ and *m*′ of *l* and *m* intersected at a point *P*′, the object of *P*′ would have to be a point lying on *l* and also on *m*, and by assumption these are parallel lines.

43.5 Determination of an isometry

Theorem *An isometry of the Gauss plane is uniquely determined by the assignment of the congruent map of a given triangle*. (See Fig. 43.2.)

Proof. We are given three distinct non-collinear points z_0, z_1 and z_2, and their respective images w_0, w_1 and w_2, where

$$|z_i - z_i| = |w_i - w_j| \ (i, j = 0, 1, 2).$$

We wish to show that the image of a point z distinct from the three given points z_i is uniquely defined under an isometry which satisfies the given conditions.

We may argue that a point is uniquely determined if its distances from three given non-collinear points are given. For the assignment of its distances from two given points determines the point as the intersection of two distinct circles, which intersect in two points, and the point also lies on a third circle, center at the third given point.

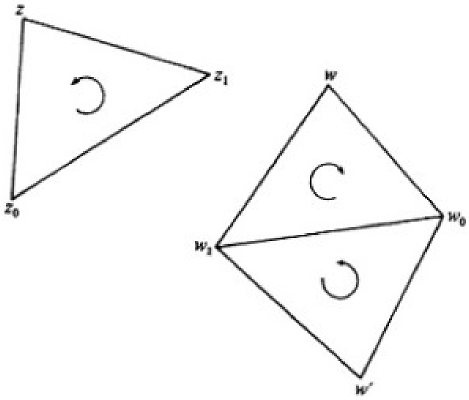

Fig. 43.3

This circle cannot pass through both points of intersection of the first two circles constructed, since its center would have to be collinear with the first two centers. Hence it passes through only one of the two points, and so the point is uniquely determined by its distances from three given non-collinear points.

Hence the image of z under an isometry is uniquely determined, since we are given the distances of z from the points z_i, and we are therefore also given the distances of the image w from the points w_i.

Another proof is indicated in Exercise 43.3.

43.6 An auxiliary theorem

Theorem *There are precisely two isometries of the Gauss plane which map two given points* z_0 *and* z_1 *onto two given points* w_0 *and* w_1, *where* $|z_0 - z_1| = |w_0 - w_1| \neq 0$. (See Fig. 43.3.)

Proof. We have already found two isometries which do this, in Thm 43.2, one of type $z' = az + b$ ($|a| = 1$), and the other of type $z' = c^{\bar{z}} + d$ ($|c| = 1$). We show that there are not more than two.

By an argument similar to the one used in Thm 43.5, we know that if a point z is not on the line joining the points z_0 and z_1, its image point w is determined as the intersection of two circles, one center w_0 and of radius $|z - z_0|$, and the other center w_1 and radius $|z - z_1|$. Hence there are two possible images of z. Call these points w and w'. By the preceding theorem, the assignment of w_0, w_1 and w as image points of z_0, z_1 and z determines the isometry uniquely, and similarly the assignment of w_0, w_1 and w' as the image points of z_0, z_1 and z determines the isometry uniquely. Hence there are at most two isometries which map z_0 on w_0 and z_1 on w_1, and since we know two isometries which do this, there are exactly two, and these are of the form $z' = az + b$ ($|a| = 1$) and $z' = c^{\bar{z}} + d$ ($|c| = 1$).

We have now proved our Main Thm 43.1, which we repeat here:

Main theorem on isometries of the Gauss plane. The set of all isometries $\mathscr{I}$ of the Gauss plane is composed of two classes $\mathscr{I}_+$ and $\mathscr{I}_-$. The class $\mathscr{I}_+$ consists of all isometries of the form $z' = az + b$ ($|a| = 1$), and the class $\mathscr{I}_-$ of all isometries of the form $z' = c^{\bar{z}} + d$ ($|c| = 1$).

Remark. We point out once again that the isometry which maps the points of the line joining z_0 to z_1 onto the points of the line joining w_0 to w_1 is unique. The fact that two distinct isometries can effect the mapping is only visible when we consider the mappings of points not on the line joining z_0 to z_1. We notice that the sense of rotation (w_0, w_1, w) is opposite to the sense of rotation (w_0, w_1, w') in the proof of the theorem above. One isometry, $z' = az + b$ ($|a| = 1$) *preserves* orientation, and the other, $z' = c^{\bar{z}} + d$ ($|c| = 1$) *reverses* orientation. Of course, we have not proved these statements as yet. The isometry $z' = az + b$ is called a *direct* isometry, and the isometry $z' = c^{\bar{z}} + d$ is called an *indirect* (or *opposite*) isometry.

Exercises

43.1 Find the isometries of the Gauss plane which map the point i onto the point O, and the point O onto the point 1. Which isometry also maps the point 1 onto the point $1 + i$? What is the image of the point 1 under the other isometry? Compare the orientations of the two image triangles.

43.2 l and m are two parallel lines, the points A, B lie on l, the points C, D on m, $AB = CD$, and AD intersects BC in E. Show that $AE = ED$, and $BE = EC$. Under an isometry, the points map into A', B', C', D' and E', and $A'B' = C'D'$, and $A'E' = E'D'$, while $B'E' = E'C'$. Prove that $A'B'$ must be parallel to $C'D'$, and deduce Thm 43.4.

43.3 Give another proof of Thm 43.5, using the theorem that under an isometry parallel lines map onto parallel lines. (If z is not on a side of triangle $z_0z_1z_2$, draw parallels to z_0z_1 and z_0z_2 respectively through z, intersecting z_0z_2 and z_0z_1 respectively in z'' and z'. Prove that w is uniquely determined by a reversal of this construction.)

43.4 Show that in cartesian coordinates a direct isometry is given by equations of the form

$$x' = x \cos \alpha - y \sin \alpha + p,$$

$$y' = x \sin \alpha + y \cos \alpha + q,$$

whereas an indirect isometry is given by equations of the form

$$x' = x \cos \alpha + y \sin \alpha + p,$$

$$y' = x \sin \alpha - y \cos \alpha + q,$$

Show that the inverse mappings which give (x, y) in terms of (x', y') have similar forms.

44.1 Subgroups

The axioms which determine a group have already appeared in §0.10. Here we give some more of the theory, for the reader who is not too familiar with the subject.

A *subgroup* of a given group G is a group whose elements lie in G, and whose operations are the same as those of G. For example, the rational numbers form a group under the operation of addition, and so do the integers, and the latter is a subgroup of the former. The set of even integers is a subgroup of the set of integers, and therefore also a subgroup of the set of rationals.

If we consider the set of all non-zero complex numbers, the operation now being *multiplication*, they form a group, with identity element the complex number 1, since any non-zero complex number has an inverse which is a complex number, the product of two non-zero complex numbers is also a non-zero complex number, and multiplication is associative. Within this group if we consider those complex numbers whose modulus is equal to 1, this set forms a subgroup, since $|z_1z_2| = |z_1|\ |z_2|$, and$|1/z| = |z|^{-1}$, shows that the product of two unimodular numbers is unimodular, and the inverse of a unimodular number is unimodular.

44.2 Conditions for a subgroup

Theorem *A non-empty subset H of a group G is a subgroup of G if and only if for every a, $b \in H$* (i) *b^{-1} and $ab \in H$, or* (ii) *$ab^{-1} \in H$.*

F *.

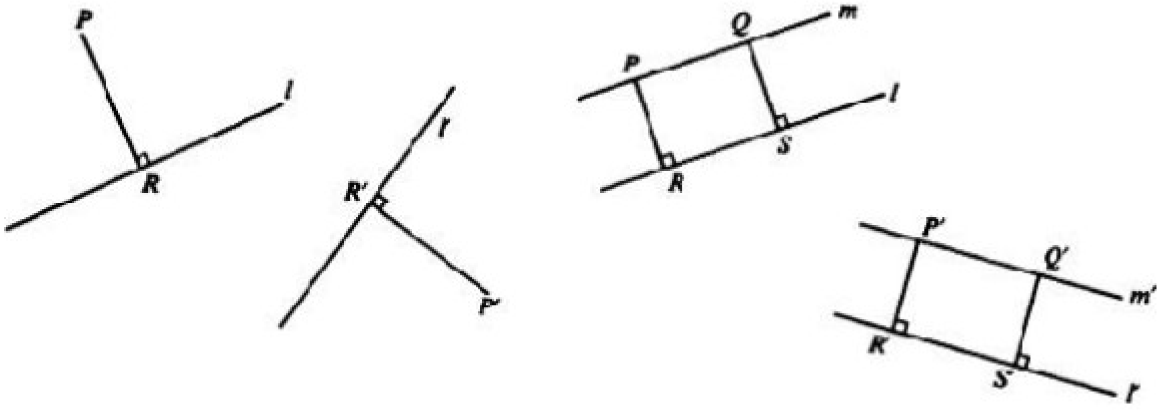
P
l
R
l'
R'
P'
m
Q
P
l
S
R
P'
Q'
m'
R'
S'
l'

/ . :

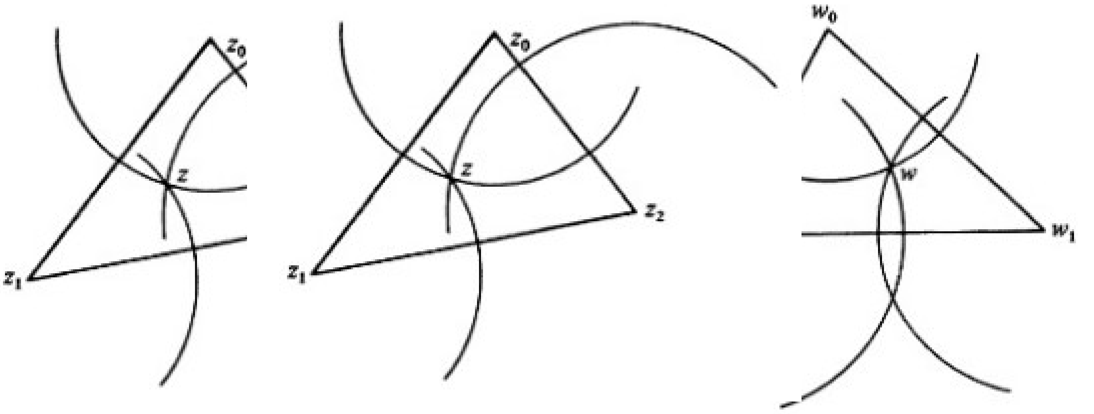
z_0
z
z_1
z_0
z
z_2
z_1
w_0
w
w_1

/ . :

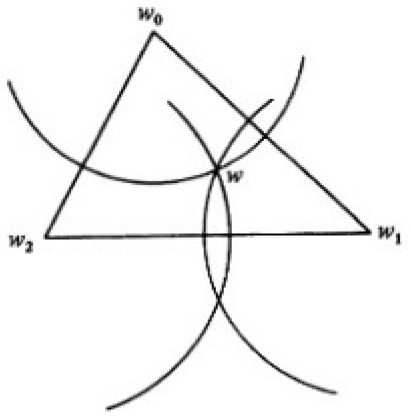

5*6

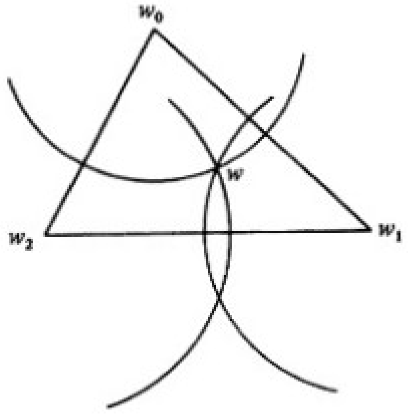

576

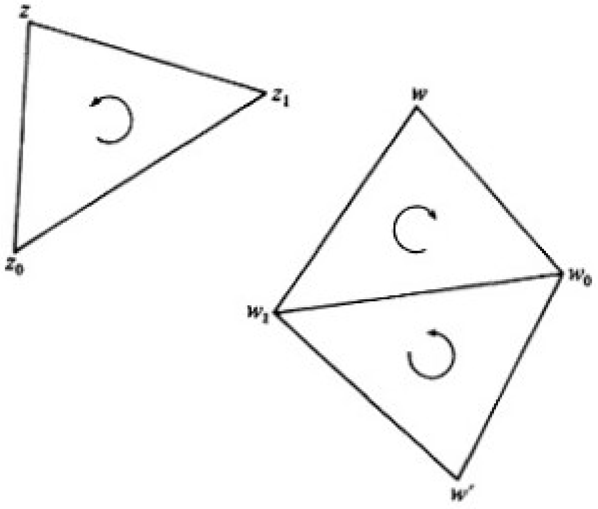
z
z_1
z_0
w
w_0
w_1
w'

/ . :

VI

MAPPINGS OF THE INVERSIVE PLANE

We now consider a larger group of mappings of the extended Gauss plane, named after Moebius, and to do so we extend the plane of the complex numbers by adjoining one additional point, which we call *the point at infinity*, and denote by the symbol ∞.

We stipulate that $1/z = \infty$, when $z = 0$, and calculations with the symbol ∞ will be performed according to the following rules:

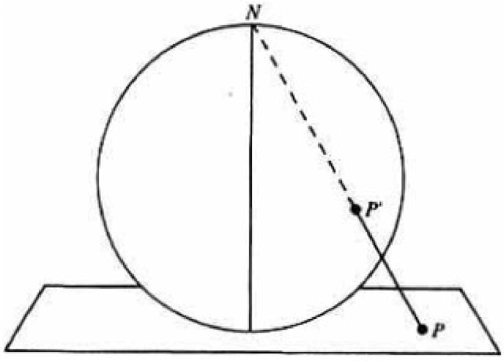

Fig. 52.1

For any finite complex number c,

$$c \pm \infty = \infty,\ c/\infty = 0,$$

and for non-zero c,

$$c \,.\, \infty = \infty,\ c/0 = \infty.$$

The expressions $\infty - \infty$, ∞/∞, $0 \,.\, \infty$ and $0/0$ are considered to be undefined.

Definition. The Gauss plane, with the point ∞ adjoined to it, is called *the inversive plane*.

Remark. If we recall that the complex number plane, the Gauss plane, is in one-to-one correspondence, by means of stereographic projection, with the surface of a punctured sphere, a sphere from which the point N of projection has been removed, all that we are doing now is to take the whole surface of the sphere as a map of the plane of complex numbers, adding the point N, to which we attach the symbol ∞ (Fig.52.1).

52.1 Moebius transformations

These are transformations of the inversive plane given by

$$z' = \frac{az + b}{cz + d},$$

where a, b, c and d are complex numbers with the restriction that

$$\Delta = ad - bc \neq 0.$$

The number Δ is called the *determinant* of the mapping.

We do not intend to give a detailed account of these mappings. Our purpose in this chapter is to develop the theory sufficiently to discuss the Poincare model of a hyperbolic non-Euclidean geometry. To do this we shall have to use many of the ideas we have introduced in preceding chapters.

52.2 The determinant of a Moebius mapping

Let us assume that in the expression $(az + b)/(cz + d)$, the number $c \neq 0$. Then we may write

$$\frac{az + b}{cz + d} = \frac{a}{c} - \frac{(ad - bc)}{c(cz + d)}.$$

If $\Delta = ad - bc = 0$, our mapping would be $z' = a/c$, a mapping of the whole inversive plane onto one point. If $c = 0$, the mapping is $z' = (az + b)/d$, which is a dilative rotation if $d \neq 0$. But if $\Delta = 0$ also, $ad = 0$. If $d = 0$, all points of the inversive plane are mapped onto the one point $z' = \infty$, and if $a = 0$, $z' = b/d$. If both $a = 0$ and $d = 0$, $z' = \infty$. Hence we see that there is a very sound reason for insisting that $\Delta \neq 0$, since we wish to consider non-degenerate mappings. Degenerate mappings, in which all points of the inversive plane are mapped onto one point, are of little interest.

We now see that the condition $\Delta \neq 0$ is sufficient for a Moebius transformation to be a one-to-one onto mapping of the inversive plane, a bijective mapping. In the first place $(az + b)/(cz + d)$ defines a unique z' unless $cz + d = 0$. But if $cz + d = 0$, so that $z = -(d/c)$, we cannot also have $az + b = 0$ for this value of z, since this would imply that $a(-d/c) + b = 0$, or $\Delta = ad - bc = 0$. Hence $cz + d = 0$ gives $z' =$ (non-zero constant)/0, and we have agreed that this gives $z' = \infty$.

The point $z = \infty$ has a unique image-point, since

$$\frac{az + b}{cz + d} = \frac{a + b/z}{c + d/z} = \frac{a}{c}$$

when $z = \infty$, by our rules. Hence every point of the inversive plane maps onto a uniquely defined point of the inversive plane.

The inverse mapping is given by $z = (-dz' + b)/(cz' - a)$, and this is again a Moebius transformation, since the determinant of this mapping is

$$(-d)(-a)-bc = ad-bc = \Delta\neq 0.$$

Hence the object-point z of an image-point z' is uniquely determined, and since every point z of the inversive plane has an image z', and every point z' of the inversive plane arises from a uniquely defined point z of the inversive plane, the mapping given by our Moebius transformation $z' = (az + b)/(cz + d)$ is one-to-one and onto, in other words, *a bijective mapping*.

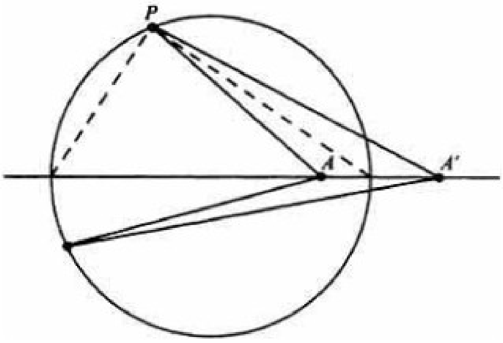

Fig. 52.2

By choosing special values of the constants a, b, c and d, we obtain a number of mappings which we have already encountered. But we can also *dissect* the given mapping into a product of familiar mappings. Before we do this, we examine the effect of a Moebius mapping on lines and circles.

52.3 The map of a circle or line under a Moebius transformation

Theorem *A Moebius transformation is a circular transformation, that is, it maps the set of circles and lines into the set of circles and lines.*

Proof. We can prove this theorem and an additional one by taking the equation of a circle or line in the z-plane in the form $|z - p| = k|z - q|$, where k is a real number (Fig.52.2). This is an expression of the theorem that any circle is the locus of a point which moves so that the ratio of its distances from two fixed points is a constant, these fixed points being inverse points for the circle. When the locus is a line, $k = 1$, and the points are mirror images in the line. The circle is called an Apollonius circle with respect to the two points (see §18.3). If we now transform this equation by the Moebius mapping $Z = (az + b)/(cz + d)$, we must write $z = (-dZ + b)/(cZ - a)$, and we obtain the equation

$$\left|\frac{-dZ+b}{cZ-a} - p\right| = k\left|\frac{-dZ+b}{cZ-a} - q\right|$$

and, after a little reduction this leads to

$$|Z(-d-pc)+b+pa| = k|Z(-d-qc)+b+qa|$$

or

$$\left|Z - \frac{ap+b}{cp+d}\right| = k\left|\frac{d+qc}{d+pc}\right|\left|Z - \frac{aq+b}{cq+d}\right|,$$

which is of the form

$$|Z-p'| = k'|Z-q'|,$$

where p', q' are the maps of the points p and q under the Moebius transformation, and $k' = k|d + qc|/|d + pc|$.

The form of the equation for Z shows that the transform is either a line or a circle, and we also have:

Corollary. The map of a circle and a pair of inverse points under a Moebius transformation is a circle and a pair of inverse points, where if the circle is a line, the points are mirror images in the line.

Other proofs of this theorem and Corollary are indicated in Exercises 52.1, 52.3.

52.4 Dissection of a Moebius transformation

We may write

$$z' = \frac{az + b}{cz + d} = \frac{a}{c} + \frac{(bc - ad)}{c(cz + d)} = \frac{a}{c} + \frac{(bc - ad)}{c^2}\left(\frac{1}{z + d/c}\right).$$

We now write

$$\text{(i)}\ T_{a/c}: z_1 = z + d/c,$$

$$\text{(ii)}\ N: z_2 = 1/z_1,$$

$$\text{(iii)}\ D: z_3 = \frac{(bc - ad)}{c^2} z_2,$$

$$\text{(iv)}\ T_{a/c}: z' = a/c + z_3.$$

The mapping (i) is a translation, (ii) is an inversion in the unit circle, center at the origin, followed by a reflexion in the real axis, (iii) is a dilative rotation, and (iv) is also a translation. If we give names to these familiar mappings as indicated, then

$$z' = \frac{az + b}{cz + d} = zT_{d/c}NDT_{a/c}.$$

In other words, a general Moebius transformation is the composition of a translation, an inversion in the unit circle followed by a reflexion in the *X*-axis, a dilative rotation and a translation. Many important consequences follow from this dissection, as the exercises following indicate.

Of course, a Moebius transformation is not, in general, a direct isometry, nor a direct similitude, although these mappings are included amongst Moebius transformations as special cases.

52.5 The group of Moebius transformations

Theorem *Moebius transformations form a group* $\mathscr{B}$ *under composition of mappings. If B and C are two Moebius mappings,* Δ_B *and* Δ_C *their determinants, then* $\Delta_{BC} = \Delta_B\Delta_C$ *is the determinant of the Moebius mapping BC.*

Proof. We have seen already that if B is a Moebius transformation, so is B^{-1}. The identity mapping $z' = z$ is a Moebius transformation. If $B: z' = (az + b)/(cz + d)$, $\Delta_B = ad - bc$, $C: z' = (a'z + b')/(c'z + d')$, $\Delta_C = a'd' - b'c'$ are two Moebius transformations, then the composition is

$$BC: z' = \frac{a'(az+b) + b'(cz+d)}{c'(az+b) + d'(cz+d)} = \frac{(aa' + b'c)z + a'b + b'd}{(ac' + cd')z + bc' + dd'},$$

and $\Delta_{BC} = (aa' + b'c)(bc' + dd') - (a'b + b'd)(ac' + cd') = (ad - bc)(a'd' - b'c')$, after some cancellation, so that $\Delta_{BC} = \Delta_B \Delta_C$, as we wished to prove, and since $\Delta_B \neq 0$, $\Delta_C \neq 0$, $\Delta_{BC} \neq 0$, and so the composition is also a Moebius transformation. The associative law holds for composition of bijective mappings, and therefore all the group axioms are satisfied, and the theorem is proved.

Exercises

52.1 By looking at the dissection of a Moebius transformation show that a Moebius transformation maps lines onto lines or circles, and circles onto lines or circles.

52.2 Show that the mapping $z' = 1/z$ maps the points 2, $1 + i$, $1 - i$ of the circle $|z - 1| = 1$ onto the collinear points $1/2$, $1/2 - i/2$, $1/2 + i/2$. Can you explain why this is so?

52.3 Give another proof of the Corollary to Thm 52.3, by examining the effect of the component parts of a Moebius transformation.

52.4 In the proof of Thm 52.3, the condition that the map of our circle or line be a line is $k' = k|d + qc|/|d + pc| = 1$. Show that an equivalent condition is that the point $z = -d/c$ which maps onto the point $Z = \infty$ lie on the circle $|z - p| = k|z - q|$. Why is this equivalence to be expected?

52.5 Show that the condition that $Z = (az + b)/(cz + d)$ should map the circle $|Z| = 1$ onto a line in the z-plane is $|a| = |c|$.

52.6 If a given Moebius transformation maps a circle $\mathscr{C}$ onto a circle $\mathscr{C}'$, show that a point P moving continuously on $\mathscr{C}$ is mapped onto a point P' which moves continuously on $\mathscr{C}'$, but that clockwise motion on $\mathscr{C}$ may be changed into anti-clockwise motion on $\mathscr{C}'$.

52.7 By Thm 52.5, the determinant of the mapping inverse to $z' = (az + b)/(cz + d)$ should be $(ad - bc)^{-1}$, whereas the mapping $z = (-dz' + b)/(cz' - a)$, which is the inverse mapping, has determinant $ad - bc$. Can you explain this?

53.1 Fixed points of a Moebius transformation

Theorem *If a Moebius transformation has more than two distinct fixed points, it is the identity mapping, $z' = z$.*

Proof. If w is a fixed point, then $w = (aw + b)/(cw + d)$. This leads to the quadratic equation $cw^2 + (d - a)w - b = 0$. If this equation has more than two distinct roots, it must be the identical equation $0 . w^2 + 0 . w + 0 = 0$, so that $c = d - a = b = 0$. Then the mapping is $z' = z$, since $\Delta = ad \neq 0$, and so neither a nor d is zero, and $z' = az/d = z$.

It is sometimes possible to say that a given mapping is uniquely determined when a certain number of distinct points and their images under the mapping are given. We came across theorems of this nature with isometries (Thm 43.5). A Moebius transformation depends on the ratios of four constants, a, b, c, d, subject to the conditions $ad - bc \neq 0$, and we expect such a transformation to be given if three distinct points and their images are assigned. But a stronger theorem is true.

53.2 The determination of a Moebius transformation

Theorem *A Moebius transformation is uniquely determined by the assignment of three distinct points z_j and their three distinct image points z_j'* ($j = 1, 2, 3$).

Proof. Consider the mapping $B{:}z' = (z - z_2)(z_1 - z_3)/(z - z_1)(z_2 - z_3)$. This is a Moebius transformation, since its determinant is

$$\Delta c = (z_2' - z_3')(z_1' - z_3')(-z_1' - z_2') \neq 0,$$

Under the mapping B, the image of $z = z_1$ is $z' = \infty$, the image of $z = z_2$ is $z' = 0$, and the image of $z = z_3$ is $z' = 1$.

Now define a mapping $C{:}z' = (z - z_2')(z_1' - z_3')/(z - z_1')(z_2' - z_3')$. Here the determinant is

$$\Delta c = (z_1' - z_3')(z_2' - z_3')(-z_1' - z_2') \neq 0,$$

so that this is also a Moebius transformation, and under this mapping the points z_1', z_2' and z_3' are mapped, respectively, onto the points ∞, 0 and 1. The composition of transformations BC^{-1} is again a Moebius mapping, and maps z_1 onto ∞ and then onto z_1', z_2 onto 0 and then onto z_2', and finally z_3 onto 1 and then onto z_3'.

Hence BC^{-1} is a Moebius transformation which maps z_1, z_2, z_3 onto the points z_1', z_2', z_3'. Is it unique? If A were a distinct Moebius mapping with the same properties, the Moebius mapping $(BC^{-1})A^{-1}$ would have three distinct fixed points, z_1, z_2 and z_3, and by Thm 53.1, the transformation $(BC^{-1})A^{-1} = \text{Identity}$, which means that $BC^{-1} = A$.

We are on ground familiar to those who know something about projective geometry (§69.1), and without saying so, we have been using the concept of *cross-ratio*.

53.3 The cross-ratio of four complex numbers

Definition. The cross-ratio of the four points (or complex numbers) z_j ($j = 1, 2, 3, 4$) is the complex number

$$(z_1, z_2; z_3, z_4) = \frac{(z_1 - z_3)}{(z_2 - z_3)} \Big/ \frac{(z_1 - z_4)}{(z_2 - z_4)}$$
$$= \frac{(z_1 - z_3)(z_2 - z_4)}{(z_2 - z_3)(z_1 - z_4)}.$$

It will be noticed that the order of the points is significant, and in fact there are, on the face of it, 4! = 24 different values of the cross-ratio of four given points, if the order of the points is not assigned. But these fall into six sets, since there are relations between the values which arise after a permutation of the points. The full details will be given later (§69.1). Here we need one property of the permutations:

Theorem $(z_1, z_2; z_3, z_4) = (z_2, z_1; z_4, z_3)$.

Proof. This is simple verification:

$$(z_2, z_1; z_4, z_3) = \frac{(z_2 - z_4)}{(z_2 - z_3)} \Big/ \frac{(z_1 - z_4)}{(z_1 - z_3)} = \frac{(z_1 - z_3)(z_2 - z_4)}{(z_2 - z_3)(z_1 - z_4)}$$
$$= (z_1, z_2; z_3, z_4).$$

Remark. The above is a special case of the theorem that *interchanging one pair* of the four numbers *and also the other pair* leaves the cross-ratio unchanged (compare Exercise 69.1).

One reason for the importance of the cross-ratio of four numbers is its *invariance* under any Moebius transformation. We prove this in the next section, but remark here that we proved in §23.2 that the cross-ratio of four collinear points is unchanged if we *invert* with respect to a center of inversion on the line. Now, inversion by itself is not a Moebius mapping, but *inversion followed by reflexion* in the line on which the points lie is a Moebius mapping, given by $z' = 1/z$, and the points are unchanged by the reflexion in the line. Hence we can regard the theorem of §23.2 as a very special case of the one we shall now prove.

53.4 An invariant under Moebius transformations

Theorem *The cross-ratio of four points is an invariant under Moebius* *transformations.*

Proof. This can be shown by direct verification. If the four points are $z_j (j = 1, 2, 3, 4)$, and these are mapped onto the points z_j', where $z_j' = (az_j + b)/(cz_j + d)$, then straightforward calculation shows that

$$\frac{(z_1 - z_3)(z_2 - z_4)}{(z_2 - z_3)(z_1 - z_4)} = \frac{(z_1' - z_3')(z_2' - z_4')}{(z_2' - z_3')(z_1' - z_4')}.$$

Alternatively, we may use the method of Thm 53.2. With the notations of that theorem, we have $z_j BC^{-1} = z_j'$, which means that $z_j B = z_j' C$. If we write out what this means in algebraic terms, we have

$$\frac{(z_j - z_2)(z_1 - z_3)}{(z_j - z_1)(z_2 - z_3)} = \frac{(z_j' - z_2')(z_1' - z_3')}{(z_j' - z_1')(z_2' - z_3')},$$

and writing $z_j = z_4$, and $z_j' = z_4'$, we have the desired result. Of course, the proof of Thm 53.2 was devised with a previous knowledge of the cross-ratio function in mind.

We now prove a very important geometric property of the cross-ratio.

53.5 The reality of the cross-ratio

Theorem *The cross-ratio of four distinct points is real if and only if the four points lie on a straight line or a circle.*

Remark. This theorem can be proved geometrically, more or less, by a method indicated in Exercise 53.3. The proof we give here is purely algebraic.

Proof. Let the four distinct points be $z_j (j = 1, 2, 3, 4)$, and suppose in the first instance that the cross-ratio $(z_1, z_2; z_3, z_4) = k$, where k is real. Let T be the unique Moebius transformation which maps z_1, z_2 and z_3 onto the three distinct points z_1', z_2' and z_3' of the real axis (Thm 53.2). Let $z_4 T = z_4'$. By the invariance of the cross-ratio (Thm 53.4), we have

$$\frac{(z_4 - z_2)(z_1 - z_3)}{(z_4 - z_1)(z_2 - z_3)} = \frac{(z_4' - z_2')(z_1' - z_3')}{(z_4' - z_1')(z_2' - z_3')} = k.$$

Hence

$$(z_4' - z_2')/(z_4' - z_1') = \mathrm{k}(z_2' - z_3')/(z_1' - z_3')$$

= a real number

Since z_1', z_2', z_3' are real, if we solve this equation for z_4', we find that z_4' = a real number, since the real numbers form a field. Hence $z_4T = z_4'$ also lies on the real axis. The four points z_1, z_2, z_3 and z_4 therefore lie on the image of (real line)T^{-1}. But we know that a Moebius transformation maps a line onto a circle or a line. Since T^{-1} is a Moebius transformation, and the real line is a line, it follows that the four points z_j ($j = 1, 2, 3, 4$) lie on a circle or a line when the cross-ratio of the four points is real.

Now let us assume that the four given points lie on a line or a circle. There is a unique Moebius transformation T which maps z_1, z_2 and z_3 into three given distinct points z_1', z_2', z_3' of the real axis, and since T maps circles and lines onto circles or lines the map of the line or circle which contains the four points z_j must be a line or circle through the three points z_1', z_2', z_3'. Since these points lie on a line, the real axis, the map must be this line. Hence the map of z_4 is also a point z_4' on the real axis. The cross-ratio of the four points z_j' on the real axis is a real number, and since cross-ratios are invariant under a Moebius transformation, it follows that the cross-ratio $(z_1, z_2; z_3, z_4)$ of the four points on a line or a circle is the same real number.

53.6 Lines in the inversive plane and the point ∞

The point ∞ was added to the Euclidean plane, considered as a representative of the complex line, as soon as we introduced inversion in §20.1. We then found that inversion is a one-to-one onto transformation of the completed plane, which we call the inversive plane. In §21.1 it was seen to be reasonable to consider all lines of the inversive plane to pass through this point ∞. The theorem that the inverse of a line which does not pass through the center of inversion is a curve (a circle, in fact) which does pass through the center of inversion is then an immediate deduction. When we consider Moebius transformations of the Gauss plane, inversion is the one component of the dissected Moebius transformation which is not one-to-one onto, and by adding the point ∞ to the Gauss plane, obtaining the inversive plane, we ensure that a Moebius transformation is one-to-one onto for the inversive plane.

We have already mentioned stereographic projection, which gives a one-to-one correspondence between the points of the Gauss plane and the points of a sphere from which one point has been removed. Adding the point ∞ to the Gauss plane is equivalent to patching the hole in the punctured sphere. We shall see later (Exercise 88.1) that circles in the Gauss plane project into circles on the sphere, if we project from the point *N* (Fig.52.1), the Gauss plane being regarded as the tangent plane at the point on the sphere diametrically opposite to *N*. If we join *N* to a line in the Gauss plane, the intersection of the resulting plane with the sphere is a circle which passes through *N*. The inversive plane can be regarded as the surface of a sphere, the point *N* being the point ∞. This was already remarked at the beginning of the Chapter. The further point made here is that lines in the inversive plane all pass through the point ∞.

Exercises

53.1 Show that any Moebius transformation with the two distinct fixed points p and q may be written in the form

$$\frac{z'-p}{z'-q}=k\,\frac{(z-p)}{(z-q)},$$

where k is a complex number.

53.2 Deduce from Thm 53.5 that a Moebius transformation maps the set of lines and circles into the set of lines and circles.

53.3 If the complex numbers z_1, z_2, z_3, z_4 are attached to the points A, B, C and D respectively of the real Euclidean plane, prove that the cross-ratio

$$(z_1,z_2;z_3,z_4) = \mathrm{k}(\cos\theta + \mathrm{i}\sin\theta),$$

where $k = |AC| \,.\, |BD|/|AD| \,.\, |BC|$, and $\theta = \alpha - \beta$, where α and β are the measures of the angles CAD and CBD respectively. Deduce Thm 53.5.

53.4 Show that the points of any given circle in the inversive plane may be represented thus:

$$z=(at+b)/(ct+d)\ (ad-bc\neq 0,)$$

where a, b, c, and d are complex numbers, and t is real.

53.5 Prove that if z_1, z_2, p, q and r are distinct complex numbers, the following relation between cross-ratios is true:

$$(z_1,z_2;p,r) = (z_1,z_1;P,q).(z_1,z_2;q,r).$$

53.6 Prove that

$$(z_1,z_2;z_3,z_4) = (z_2,z_1;z_3,z_4) = 1/(z_1,z_2;z_3,z_4)$$

53.7 Show that if the points z_1, z_2, z_3 and z_4 which lie on a circle are mapped by a Moebius transformation onto the points A, B, C and ∞ of a line, then

$$(z_1,z_2;z_3,z_4)=(A, B ;C, \infty) = \overline{AC}/\overline{BC}.$$

53.8 Prove the following theorem for cross-ratios:

$$\frac{(z_1, w_2; z_2, w_1)}{(z_2, w_3; z_3, w_2)} \cdot \frac{(z_3, w_4; z_4, w_3)}{(z_4, w_1; z_1, w_4)}.$$

53.9 Prove that if $\mathscr{C}_1,\mathscr{C}_2,\mathscr{C}_3$ and $\mathscr{C}_4$, are four circles (or lines), and $\mathscr{C}_1,\mathscr{C}_2$ intersect in P_1, Q_1, the circles $\mathscr{C}_1$ intersect in P_2, Q_2, the circles $\mathscr{C}_2,\mathscr{C}_3$ intersect in P_3, Q_3, and finally the circles $\mathscr{C}_4,\mathscr{C}_1$ intersect in P_4, Q_4, then if the four points P_i lie on a line or circle, so do the four points Q_i. (Use the preceding Exercise. This result appears again as Thm III of §94.1, with applications. Note that one point of the intersection of two circles, when the circles are lines, is always the point ∞.)

53.10 In the stereographic projection of the surface of a sphere onto the inversive plane (Fig.52.1), the lines in the plane arise from circles through the point N. What can you say about the two circles through N which arise from two parallel lines in the plane?

53.11 If $z_i = a + r(\cos \alpha_i + i \sin \alpha_i)$ ($i = 1, 2, 3, 4$), so that the four points z_i are points on a circle, center a, radius r, angular coordinates α_i, show that

$$(z_1, z_2; z_3, z_4) = \frac{\sin \frac{1}{2}(\alpha_1 - \alpha_3) \sin \frac{1}{2}(\alpha_2 - \alpha_4)}{\sin \frac{1}{2}(\alpha_1 - \alpha_4) \sin \frac{1}{2}(\alpha_2 - \alpha_3)}.$$

53.12 If the four points A, B, C, D represented by the complex numbers z_1, z_2, z_3, z_4 form a set of *orthocyclic points* (see Exercise 21.9), show that $(z_1, z_2; z_3, z_4)$ has modulus 1.

53.13 Prove that the triangles ABC and $A'B'C'$ are directly similar if and only if $(a, b; c, \infty) = (a', b'; c', \infty)$, where a, b, c are the complex numbers representing the points A, B, C, and a', b', c' the complex numbers representing the points A', B', C'.

54.1 The map of the inside of a circle

Theorem *If a circle $\mathscr{C}$ be mapped onto a circle $\mathscr{C}'$ by a Moebius mapping, then the interior of $\mathscr{C}$ is mapped onto the interior of $\mathscr{C}'$, or onto the exterior of $\mathscr{C}'$.*

Proof. We saw that a given Moebius transformation may be written as the composition of a translation, an inversion in the unit circle followed by a reflexion in the real axis, a dilative rotation and a translation. All these mappings, with the exception of the inversion, map the inside of a circle onto the inside of the mapped (image) circle. Inversion may map the inside of the object circle onto the inside of the image circle, or it may map the inside of the object circle onto the outside of the image circle. A simple example is the effect of inversion on the circle of inversion itself. Points on the circle of inversion map onto themselves, but points *inside* the circle are mapped onto points *outside* the circle. On the other hand, if the circle of inversion is *orthogonal* to the circle $\mathscr{C}$, points on $\mathscr{C}$ are merely interchanged by the inversion, and it is clear that the interior of $\mathscr{C}$ maps into the interior of $\mathscr{C}' = \mathscr{C}$.

Algebraic proofs of the theorem in inversion: The interior of a circle $\mathscr{C}$ inverts into the interior or the exterior of the image circle $\mathscr{C}'$, and this depends on whether the center of inversion is outside or inside $\mathscr{C}$, are indicated in the following Exercise (Exercise 54.3).

54.2 Map of a circle into itself

We are especially interested in those Moebius transformations which map a given circle $\mathscr{C}$ onto itself, and the inside of $\mathscr{C}$ onto the inside of $\mathscr{C}$. We take the unit circle, center at the origin, $|z| = 1$, as the circle $\mathscr{C}$.

There are, of course, Moebius transformations which map $|z| = 1$ onto $|z| = 1$, and the inside onto the inside. The identity mapping $Z = z$ is one such mapping, so the set of transformations is not empty. Another Moebius mapping which will serve is the rotation $Z = az$, where $|a| = 1$. This merely rotates the unit circle about the origin. If we take the composition of two such mappings, we have a mapping of the unit circle onto itself, inside mapping onto inside, and more generally we see that *under composition the set of Moebius transformations which map a circle and its interior onto themselves form a group*. This is merely a special case of the theorem discussed in §0.12: 'One-to-one onto mappings of a set form a group under composition.'

It is interesting to determine what form a Moebius transformation takes when it maps $|z| = 1$ onto $|Z| = 1$ and the interior onto the interior. If the transformation is $Z = (az + b)/(cz + d)$, we want $z^{\bar{z}} = 1$ to map onto $Z\overline{Z} = 1$, in the first instance. The map of $Z\overline{Z} = 1$ arises from

$$\frac{(az+b)}{(cz+d)}\frac{\overline{(az+b)}}{\overline{(cz+d)}} - 1 = 0,$$

which simplifies to

$$z\bar{z}(a\bar{a} - c\bar{c}) + z(a\bar{b} - c\bar{d}) + \bar{z}(\bar{a}b - \bar{c}d) + b\bar{b} - d\bar{d} = 0.$$

This is the same as the equation $z^{\bar{z}} - 1 = 0$ if

$$a\bar{b} - c\bar{d} = \bar{a}b - \bar{c}d = 0,$$

and

$$a\bar{a} - c\bar{c} + b\bar{b} - d\bar{d} = 0.$$

We therefore have only two conditions:

$$a\bar{b} - c\bar{d} = 0,$$

and

$$|a|^2 + |b|^2 = |c|^2 + |d|^2.$$

These are sufficient conditions. Write $c = k\bar{b}$. Then $\bar{c}d - \bar{a}b = 0$ leads to $\bar{k}bd - \bar{a}b = 0$, so that $d = \bar{a}/\bar{k}$. If we now substitute in the condition $|c|^2 + |d|^2 = |a|^2 + |b|^2$, this leads to $|k|^2|b|^2 + |a|^2/|k|^2 = |a|^2 + |b|^2$, and so we can take $|k| = 1$. Then $1/\bar{k} = k$, and so the form of the Moebius transformation is

$$Z = \frac{az + b}{k(\bar{b}z + \bar{a})}, \text{ where } |k| = 1,$$

and, of course, we may also write this as $Z = k(az + b)/(\bar{b}z + \bar{a})$, $|k| = 1$.

This mapping transforms $|z| = 1$ into $|Z| = 1$. The determinant of the mapping is $\Delta = k^2(a\bar{a} - b\bar{b}) = k^2(|a|^2 - |b|^2)$. If we choose $k = 1$, and $\Delta = 1$, then $|a|^2 - |b|^2 = 1$, and so $|b| < |a|$. The point 0 is then mapped onto the point $b/\bar{a}$, and since $|b| < |a|$, this point is *inside* the circle $|Z| = 1$. This is sufficient for us to be able to assert that the mapping $Z = (az + b)/(\bar{b}z + \bar{a})$, where $|a|^2 - |b|^2 = 1$, is a Moebius transformation which maps the circle $|z| = 1$ onto the circle $|Z| = 1$, and the inside of $|z| = 1$ onto the inside of $|Z| = 1$.

54.3 Conformal property of Moebius transformations

Theorem *The circles $\mathscr{C}$ and $\mathscr{D}$ intersect at the point z_0, and u is a unit tangent vector to $\mathscr{C}$ at z_0, v is a unit tangent vector to $\mathscr{D}$ at z_0. If $\mathscr{C}$ and $\mathscr{D}$ are mapped by a Moebius transformation onto the circles $\mathscr{C}'$ and $\mathscr{D}'$ intersecting at Z_0, and u is mapped on u', v is mapped on v', then u' and v' are tangent circles at Z_0 to $\mathscr{C}'$ and $\mathscr{D}'$ respectively, and the angle between u' and v' is equal to that between u and v both in measure and in sense*. (See Fig. 54.1.)

Proof. We assume that $cz_0 + d \neq 0$ in the Moebius mapping $Z = (az + b)/(cz + d)$. Let z_1 be a point on $\mathscr{C}$ distinct from z_0, and z_2 a point on $\mathscr{D}$ distinct from z_0. Then

$$Z_0 - Z_1 = \frac{az_0 + b}{cz_0 + d} - \frac{az_1 + b}{cz_1 + d} = \frac{(ad - bc)(z_0 - z_1)}{(cz_0 + d)(cz_1 + d)},$$

$$Z_0 - Z_2 = \frac{(ad - bc)(z_0 - z_2)}{(cz_0 + d)(cz_2 + d)}.$$

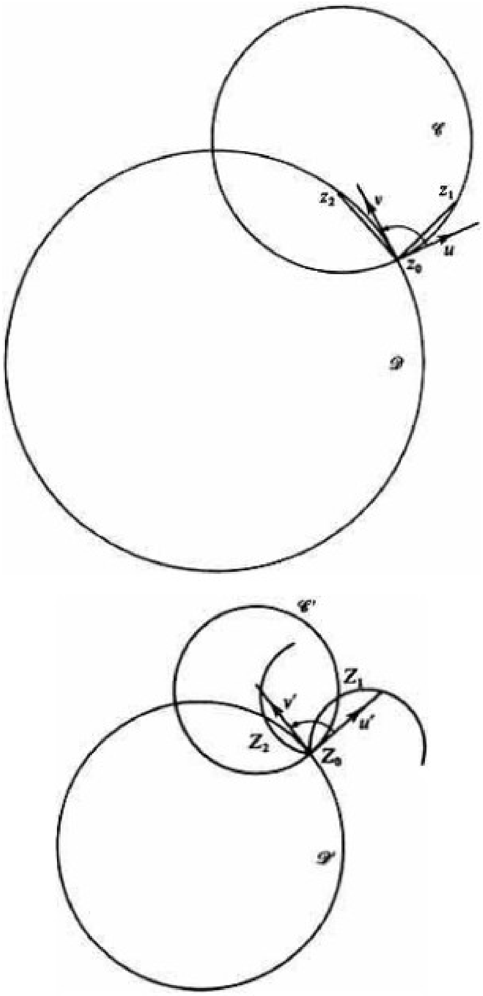

$\mathscr{C}$
z_2
v
z_1
z_0
u
$\mathscr{D}$
$\mathscr{C}'$
Z_1
v'
u'
Z_2
Z_0
$\mathscr{D}'$

Fig. 54.1

Hence, since $ad - bc \neq 0$,

$$\frac{Z_0 - Z_1}{Z_0 - Z_2} = \frac{(z_0 - z_1)}{(z_0 - z_2)} \cdot \frac{(cz_2 + d)}{(cz_1 + d)}.$$

As z_1 approaches z_0 on $\mathscr{C}$, Z_1 approaches Z_0 on $\mathscr{C}'$, and since the *line* joining z_0 to z_1 is mapped onto a circle through Z_0 and Z_1 (which may be a line), the unit tangent vector u to $\mathscr{C}$ at z_0 is mapped onto a circle tangent to $\mathscr{C}'$ at Z_0. Similarly, the unit tangent vector v to $\mathscr{D}$ at z_0 is mapped onto a circle tangent to $\mathscr{D}'$ at Z_0. As z_2 and z_1 both approach z_0 on the circles $\mathscr{C}$ and $\mathscr{D}$ respectively, we have, in the limit

$$\frac{Z_0 - Z_1}{Z_0 - Z_2} = \frac{z_0 - z_1}{z_0 - z_2},$$

and comparison with Thm 47.9 shows us that the triangles $Z_0Z_1Z_2$ and $z_0z_1z_2$ are *directly similar*, and therefore the angle between u' and v' is equal to that between u and v both in measure and sense.

Another proof of this theorem is suggested in Exercise 54.1.

Exercises

54.1 From the dissection of a Moebius transformation (§52.4) prove Thm 54.3, by examining the effect of the component transformations on the measure and the sense of the angle between two circles.

54.2 Prove directly that the mappings $Z = (az + b)/(\bar{b}z + \bar{a})$, where $a\bar{a} - b\bar{b} = 1$, map the circle $|z| = 1$ onto the circle $|Z| = 1$, and the interior onto the interior, and that the set of mappings form a group under composition.

54.3 The *inside* of the circle $\mathscr{C}$, given by the equation

$$C(X, Y) = X^2 + Y^2 - 2pX - 2qY + r = 0$$

may be characterized as the set of points (x, y) for which the *power* $C(x, y)$ of (x, y) with respect to the circle $\mathscr{C}$ is not positive. Using the formula of §21.1, show that, after inversion in the circle $X^2 + Y^2 = k^2$,

$$C(x, y) = r\left\{X^2 + Y^2 - \frac{2k^2}{r}(pX + qY) + \frac{k^4}{r}\right\}\Big/(X^2 + Y^2),$$

where the point (x, y) is mapped onto the point (X, Y) by the inversion, so that if the circle $\mathscr{C}$ is mapped onto the circle $\mathscr{C}'$, we have

Power of (x,y) with respect to $\mathscr{C}$=r(Power of(X,Y) with respect to $\mathscr{C}'$)/$(X^2 + Y^2)$.

Deduce that if the center of inversion $(0, 0)$ is *inside* $\mathscr{C}$, the inside of $\mathscr{C}$ inverts into the *outside* of $\mathscr{C}'$, and that if the center of inversion is *outside* $\mathscr{C}$, the inside of $\mathscr{C}$ inverts into the *inside* of $\mathscr{C}'$.

55.1 Special Moebius transformations

We now concentrate on the subgroup of Moebius transformations which map the unit circle onto itself, and the inside onto the inside. We know that these transformations do form a subgroup of the group of all Moebius transformations (§54.2), and we shall call these special Moebius mappings *M*-transformations.

We also give the boundary of the unit circle $|z| = 1$ a special name, ω, and we call the inside of the unit circle Ω. The points of ω are not included among the points of Ω, which is regarded as an open disk.

Our aim is to set up a geometry in Ω, defining points and lines and incidence properties in the first instance. Our points are ordinary Euclidean points in Ω, but our lines are circles orthogonal to ω, and since we are only considering the space Ω, our lines are the arcs of circles orthogonal to ω which lie in Ω. We shall show that we can obtain a geometry in Ω which marches step by step alongside Euclidean geometry, until there is a dramatic parting of ways when we arrive at the concept of parallel lines in the two geometries. In Euclidean geometry there is a unique parallel to a given line l through a given point P. We shall find in our space Ω that there are *two* parallels. This geometry is called *hyperbolic geometry*, and is an example of a *non-Euclidean* geometry.

We shall show that the group of M-transformations corresponds to the group of direct isometries of the Euclidean plane, in the sense that similar theorems are true for each group in the respective geometries. We begin our study of Ω by proving some theorems on M-transformations.

55.2 Theorems on *M*-transformations

Theorem *Let X and A be two points of* Ω, ξ *and* α *directions through X and A respectively. Then there is a unique M-transformation which maps X on A, and the direction* ξ *onto the direction* α. (See Fig. 55.1.)

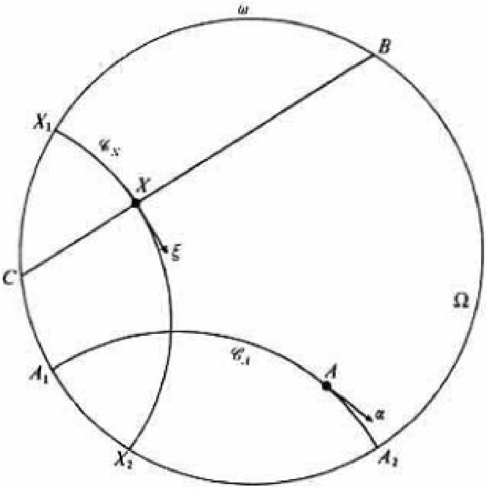

Fig. 55.1

transformations reverse orientation. We list those properties of M_--transformations which differ from M_+-transformations as Exercises (Exercises 57.7, 57.9).

We must now redefine p-congruence. If we call the elements of our extended group M^*-*transformations*, we say that $AB \overset{p}{=} A'B'$ if there is an M^*-transformation which maps A on A' and B on B'. Angles, if congruent in the Euclidean sense are congruent under M^*-transformations, and we now have *the general SAS theorem*, which is independent of the orientation of the triangles ABC and $A'B'C'$: the formulation is as in Thm 57.1, but the phrase 'with equal orientation' is omitted.

Theorem II *Given two p-triangles ABC and A'B'C', the congruences* $AB \overset{p}{=} A'B'$ $AC \overset{p}{=} A'C'$ *and* $\sphericalangle CAB \overset{p}{=} \sphericalangle C'A'B'$ *imply the p-congruence of the triangles, so that* $\sphericalangle ABC \overset{p}{=} \sphericalangle A'B'C'$, $\sphericalangle ACB \overset{p}{=} \sphericalangle A'C'B'$ *and* $BC \overset{p}{=} B'C'$.

57.2 The parallel axiom

As far as we have gone, there seems to be a complete correspondence between our p-geometry and that part of Euclidean geometry which is developed before the introduction of the parallel axiom. This axiom, in one form, (shown by Playfair to be equivalent to Euclid's formulation), says:

Through a given point P which does not lie on a given line l there passes a unique line which is parallel to l.

In Euclidean geometry, of course, two coplanar lines are said to be *parallel* if they do not intersect.

We could say that two p-lines in our p-geometry are p-parallel if they do not intersect.

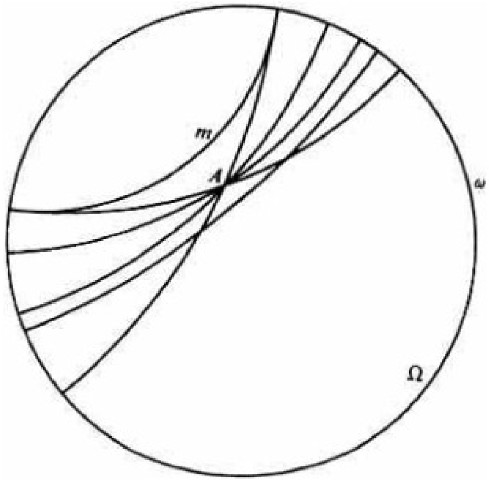

Fig. 57.3

Let A be a point in Ω (Fig.57.3), and let m be a given p-line which does not contain A. Then we see that through A there is an infinity of p-lines which do not intersect m in Ω. This property is independent of whether m is convex or concave to A. We could say, therefore, that there is *an infinity* of p-lines through A which are p-parallel to the given p-line m.

But we notice (see Exercise 57.1) that there are exactly *two* p-lines through A which touch the given p-line m, the points of contact lying on ω. These two p-lines through A separate the p-lines through A into two classes: the p-lines which cut the p-line m are in one class, and those which do not intersect m are in the second class.

We reserve the term *p-parallel* for *the two p-lines through A which touch the given p-line m at points of* ω. We now have, in our geometry, instead of the Euclidean parallel axiom, the theorem:

Theorem *Through any given p-point A which does not lie on a given p-line m there pass exactly two p-lines p-parallel to the given p-line.*

Hence our p-geometry is not Euclidean geometry, and it is naturally called *non-Euclidean geometry*. Our p-plane is a *model* for such a geometry. If Euclidean geometry is an internally consistent system, without contradictions, then so is the model, and its existence shows that the parallel axiom in Euclidean geometry is *independent* of the other axioms, that is *it cannot be deduced from them*.

Exercises

57.1 Prove that through any given point A in Ω which does not lie on a given p-line m there pass just two p-lines which touch m, and these do so on ω.

57.2 Show that through any given p-point A there exists a unique p-line l' perpendicular to a given p-line l, and that if m, n be the p-lines through A which are p-parallel to l (it is assumed that A does not lie on l), then m and n make equal angles with l'.

57.3 Prove that two distinct p-lines which are perpendicular to the same p-line have no intersection.

57.4 If we allow the vertices A, B and C of a p-triangle to lie on ω, what can we say about the angles of this triangle?

57.5 Prove that two distinct p-lines which have no intersection have a unique p-line perpendicular to both of them.

57.6 Prove that the totality of M_+ and M_- transformations, with composition as group multiplication, form a group, which is a subgroup of the group of all Moebius and conjugate Moebius transformations.

57.7 If z_1, z_2, z_3 and z_4 are mapped by a conjugate Moebius transformation into z_1', z_2', z_3' and z_4', prove that the cross-ratio of the object set is equal to the conjugate of the cross-ratio of the image set:

$$(z_1, z_2; z_3, z_4) = \overline{(z_1', z_2'; z_3', z_4')}$$

57.8 Show that the composition of the two mappings

$$Z' = (\bar{a}\bar{z} + \bar{b})/(\bar{c}\bar{z} + \bar{d}), \quad Z = \bar{Z}' \, (\bar{a}\bar{d} - \bar{b}\bar{c} \neq 0)$$

produces the mapping

$$Z = (az + b)/(cz + d), \quad (ad - bc \neq 0).$$

Deduce that Thm 52.3 and its Corollary are true for conjugate Moebius transformations.

57.9 Show that Thm 54.3 continues to hold for conjugate Moebius transformations, except that the sense of an angle is changed by the transformation.

57.10 Show that if z_1, z_2, z_3, z_4 are four points on a p-line, and z_1', z_2', z_3', z_4' are their images under a Moebius *or* a conjugate Moebius transformation, then

$$(z_1, z_2; z_3, z_4) = (z_1', z_2'; z_3', z_4').$$

57.11 Show that inversion in the circle $cz\bar{\bar{z}} - z\bar{\bar{p}} - \bar{\bar{z}}p - b = 0$, where b and c are real, the center is the complex number p/c, the square of the radius is given by $r^2 = (p\bar{\bar{p}} + bc)/c^2$, can be obtained by writing

$$(Z - p/c)\overline{(z - p/c)} = r^2 = (p\bar{p} + bc)/c^2,$$

which leads to the conjugate Moebius transformation

$$Z = \frac{p\bar{z} + b}{c\bar{z} - \bar{p}} \quad (p\bar{p} + bc > 0).$$

57.12 Show that the circles $cz\bar{\bar{z}} - z\bar{\bar{p}} - \bar{\bar{z}}p - b = 0$, $ez\bar{\bar{z}} - z\bar{q} - \bar{\bar{z}}q - d = 0$ are orthogonal if and only if $be + cd + q\bar{\bar{p}} + \bar{q}p = 0$.

57.13 Show that the composition of inversion in two circles is equivalent to a Moebius transformation.

57.14 Show that if the Moebius transformation of the preceding Exercise is of the form $Z = (az + b)/(cz + d)$, then if the circles are orthogonal we find that $a + d = 0$.

57.15 Verify that a Moebius transformation $Z = (az + b)/(cz + d)$, where $a + d = 0$ is an involutory transformation, that is, if the map of any point z is Z, then the map of Z is z. (Compare Exercise 25.2. Since $(T_{\mathscr{C}})^{-1} = T_{\mathscr{C}}$, and $(T_{\mathscr{D}})^{-1} = T_{\mathscr{D}}$, $T_{\mathscr{C}}T_{\mathscr{D}} = T_{\mathscr{D}}T_{\mathscr{C}}$ leads to $(T_{\mathscr{C}}T_{\mathscr{D}})^2 =$ Identity. See also Thm I, §76.2, and Appendix III, on *the complex line*.)

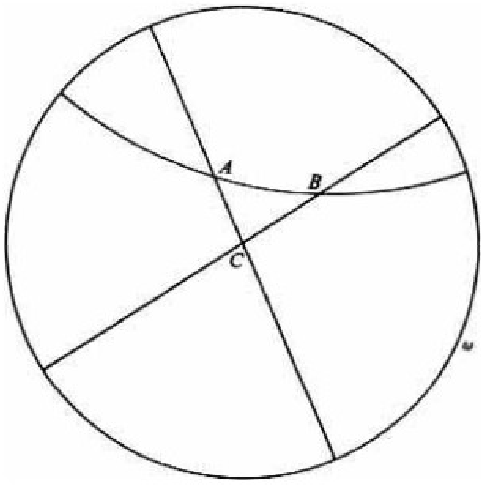

Fig. 58.1

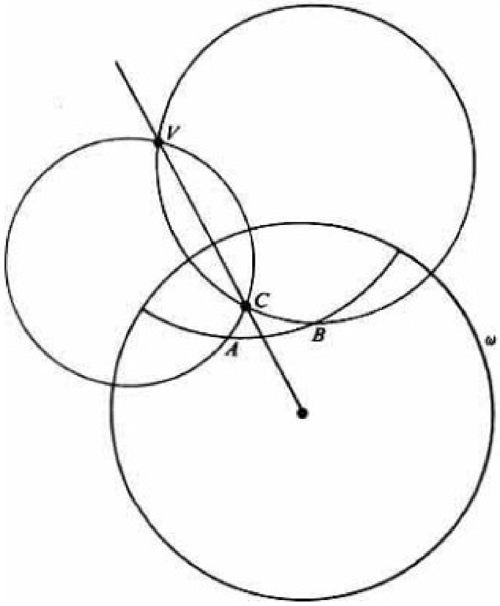

Fig. 58.2

*.

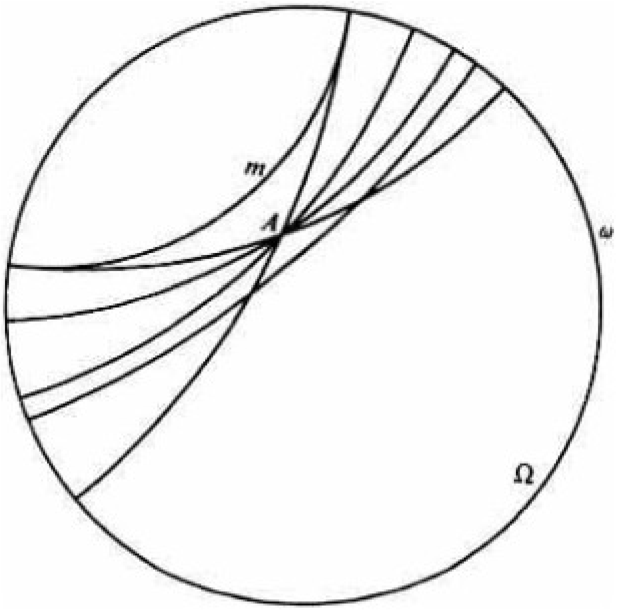
m
A
ω
Ω

/ .

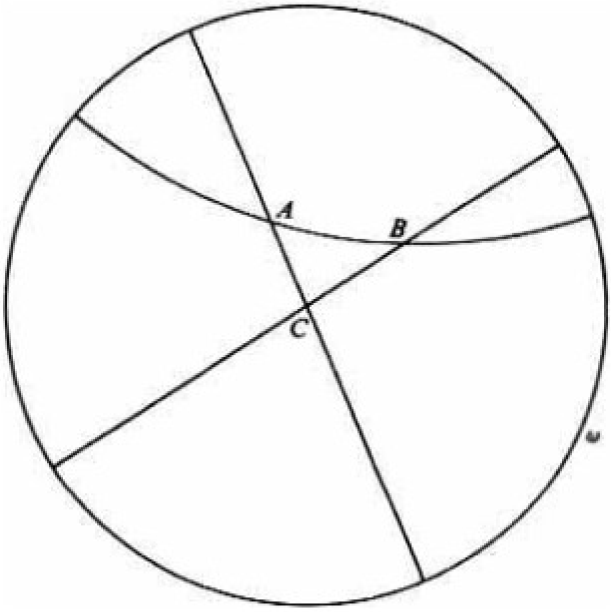
A
B
C
ω

/ .

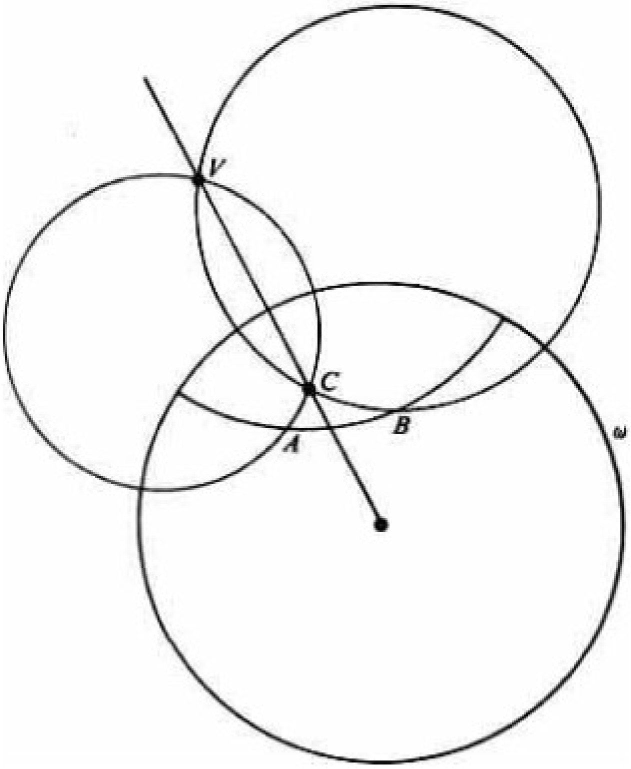
V
C
A
B
ω

/ .

58.1 The angle-sum of a *p*-triangle

Theorem *The sum of the angles of a p-triangle ABC is less than π.*

Proof. If C is the center of ω (Fig.58.1), then the p-lines CA and CB are diameters of ω, and the p-line AB is an arc of a circle which is orthogonal to ω. If two circles are orthogonal, their centers do not both lie inside one of the circles. The center of the p-line AB therefore lies outside ω, and the center C of ω lies outside the circle which determines the p-line AB. Hence the p-line AB is *convex* to C, as shown in Fig. 58.1, and the angles of the curvilinear triangle ABC at A and at B are less than the angles of the Euclidean triangle ABC at A and at B. Since the angle-sum of the Euclidean triangle ABC is equal to π, the angle-sum of the p-triangle ABC is less than π.

If no vertex of the p-triangle ABC is the center of ω (Fig.58.2), let V be the other intersection of the circles which determine the p-lines AC and BC. Then V and C are inverse points in ω, being the intersection of two circles orthogonal to ω. Choose as circle of inversion the circle $\mathscr{C}$ which has V as center and which is orthogonal to ω.

If we invert in this circle, ω inverts into itself, circles through V invert into lines, and since C is the intersection of two circles orthogonal to ω, C inverts into the center of ω, and the circles through V invert into diameters of ω. Hence the p-triangle ABC inverts into a p-triangle $A'B'C'$, where C' is the center of ω, and so we are back in the case we examined first. Since angles are unchanged by inversion, the angle-sum of p-triangle ABC = the angle-sum of triangle $A'B'C'$, and this is less than π.

58.2 The exterior angle is greater than either of the interior and opposite angles

Theorem *In any p-triangle, if one of the p-sides be produced, the exterior angle is greater than either of the interior and opposite angles.*

Proof. This follows directly from the proof of Thm 58.1, since angles of p-triangle ABC are unchanged by inversion, and if we are considering p-line CB produced, say, the exterior angle at B of p-triangle ABC is equal to the exterior angle at B' of p-triangle $A'B'C'$. Since the p-line $B'A'$ is convex to C', the exterior angle at B' of p-triangle $A'B'C'$ is greater than the same angle for the Euclidean triangle $A'B'C'$, and the angle at A' for the p-triangle $A'B'C'$ is less than the same angle for the Euclidean triangle $A'B'C'$. Since, for the Euclidean triangle $A'B'C'$

the exterior angle at B' exceeds the angle A', this is even more the case for the *p*-triangle $A'B'C'$.

58.3 Non-Euclidean distance

We have shown how to compare two *p*-segments for *p*-congruence. We now obtain a *distance-function* for any two *p*-points A and B. This function will satisfy the usual conditions:

$$d(A,B) = d(B,A) \geqq 0:$$

$$d(A,B) = 0 \text{ if and only if } \mathrm{B} = \mathrm{A}:$$

$$d(A,B) + d(B,C) \geqq d(A,C),$$

with equality if and only if B lies in the *p*-segment connecting A and C. This last inequality is called *the triangle inequality*.

Our M^*-transformations are equivalent to the isometries of the Euclidean plane, and these, of course, are distance-preserving, and we hope that our distance-function will be invariant under M^*-transformations. We shall see that this is the case.

Our distance-function is to be defined in terms of the cross-ratio $(\alpha, \beta; A, B)$, where the *p*-line cuts ω in the points α and β (Fig.58.3). We therefore begin by obtaining some further properties of this cross-ratio.

We know that there is a unique Moebius transformation which maps the points α, A and β onto three given distinct points α', A' and β' of the real line, where we choose A' to lie between α' and β'. A point moving continuously on the *p*-segment $\alpha\beta$ is mapped by this Moebius transformation onto a point which moves continuously on the segment of the real line defined by α' and β'. (Compare Exercise 52.6.) Since we are considering a pair of points A, B on the *p*-line whose terminal points on ω are α and β, all points A', B' will lie on the segment of the real line $\alpha'\beta'$. We also know that cross-ratios are invariant under a Moebius transformation, and so

$$(\alpha,\beta A,B) = (\alpha',\beta';A',B'),$$

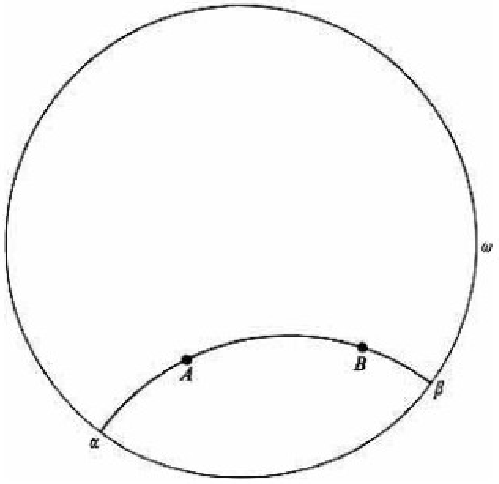

Fig. 58.3

which is a real number. We found this result in Thm 53.5. But we now obtain a stronger result. The order α, A, B, β on the p-line is the same as the order α', A', B', β' on the real line, and the cross-ratio of the real numbers is equal to $\frac{\alpha'A'}{A'\beta'}\Big/\frac{\alpha'B'}{B'\beta'}$. As A' moves on the real line from α' towards β', the ratio $\alpha'A'/A'\beta'$ increases from 0 to $+\infty$, in a continuous manner. It therefore follows, that if B' is *between* A' and β', the ratio $\alpha'B'/B'\beta'$ is *greater* than the ratio $\alpha'A'/A'\beta'$. We have therefore proved that *if the points* α, A, B, β *are in this order on a p-line*, *the cross-ratio* $(\alpha, \beta\, A, B)$ *is a positive real number less than* 1.

If α, β be interchanged, the cross-ratio is the reciprocal of this (Ex. 53.6), and similarly if A, B be interchanged: so that interchanging either α, β or A, B produces a positive real number greater than 1.

Definition. Let the p-line ab cut ω in the points α, β. Then

$$d(A,B) = |\log(\alpha,\beta A,B)|.$$

We have just seen that the cross-ratio in the definition is a positive real number, so that the logarithm exists, and the logarithm is positive if the cross-ratio exceeds 1, and negative if the cross-ratio is less than 1. Since we define $d(A, B)$ to be the modulus of the cross-ratio, we have

$$d(A,B) \geqq 0.$$

Again,

$$d(B,A) = |\log(\alpha,\beta B,A)| = |\log 1/(\alpha,\beta A,B)|$$

$$= |-\log(\alpha,\beta A,B)| = d(A,B).$$

Furthermore, if $d(A, B) = 0$, then we must have $(\alpha, \beta; A, B) = 1$, and since α, β are distinct points, this leads to $A = B$. Hence $d(A, B) = 0$ if and only if $B = A$.

We note also that interchanging α and β leaves $d(A, B)$ unchanged, the proof being similar to the proof that $d(A, B) = d(B, A)$.

Theorem I *If A, B and C are p-points, in this order, on a p-line then*

$$d(A,C)= d(A,B) + d(B,C)$$

Proof. Suppose that α, A, B, C, β lie, in this order, on a p-line, α, β being the intersections of the p-line with ω. Then we know that the cross-ratios $(\alpha, \beta; A, B)$, $(\alpha, \beta; B, C)$ and $(\alpha, \beta; A, C)$ are all positive real numbers less than 1. Their logarithms are therefore all negative real numbers. From Exercise 53.5 we may write

$$(\alpha,\beta A,C) = (\alpha,\beta A,B).(\alpha,\beta B,C)$$

(this result is independent of the order of A, B and C). Taking logarithms,

$$\log(\alpha,\beta A,C) = \log(\alpha,\beta A,B) + \log(\alpha,\beta B,C),$$

and since these three logarithms are all negative numbers, we now have

$$|\log(\alpha,\beta A,C)| = |\log(\alpha,\beta A,B)| + |\log(\alpha,\beta B,C)|$$

and so

$$d(A,C) = d(A,B) + d(B,C)$$

Before proving the general triangle inequality, we remark that our distance-function $d(A, B)$ is, in fact, invariant under an M^*-transformation. We know that an M_+-transformation leaves cross-ratios unchanged (Thm 53.4), and an M_--transformation maps a cross-ratio onto its conjugate complex value (Exercise 57.7). But when the four points under consideration are on a p-line, the cross-ratio is a real number (Thm 53.5), and so both M_+ and M_--transformations leave the cross-ratio invariant. Hence, if the points α, β, A and B on a p-line are mapped onto the points α', β', A' and B' by an M^*-transformation, these points are again on a p-line, α' and β' lie on ω, and since

$$(\alpha,\beta A,B) = (\alpha',\beta';A',B'),$$

we have

$$\mathrm{d}(A,B) = d(A',B').$$

Conversely, if we are given that $d(A, B) = d(A', B')$, we can prove that there is an M_+-transformation which maps A on A' and B on B'. It will be recalled that our original definition of *p*-congruence between *p*-segments (§56.2) involved the existence of an M_+-transformation which mapped A on A' and B on B'. The result we prove now ties the concepts of *p*-congruence and *p*-distance together.

We assume that the *p*-line AB cuts ω in the points α and β, and that the *p*-line $A'B'$ cuts ω in the points α' and β'. We name the points of intersection of the *p*-lines with ω so that the order α, A, B, β on the one *p*-line corresponds to the order α', A', B', β' on the other *p*-line. We know there is a unique M_+-transformation which maps α on α', A on A' and β on β'; the proof is exactly as in that of Thm 55.2. We assert that this M_+-transformation also maps B on B'.

We are given that $|\log(\alpha, \beta; A, B)| = |\log(\alpha', \beta'; A', B')|$, and because of our method of choosing α' and β' we know that each cross-ratio is a positive real number less than 1. Hence we can say that the two logarithms are equal, dropping the modulus signs, and that therefore the cross-ratios are equal. Since α is mapped on α', A on A' and β on β', this means that B is mapped on B'.

The proof that the cross-ratio $(\alpha, \beta; A, B)$ is a positive real number less than 1 also shows that if C is a *p*-point in the *p*-segment $B\beta$, the cross-ratio $(\alpha, \beta; A, C)$ is also positive, real and less than 1 but *less than the value of the cross-ratio* $(\alpha, \beta; A, B)$. Since the logarithms of a sequence of steadily decreasing positive real numbers less than 1 tend steadily towards $-\infty$, and the distance-function is the modulus of the logarithm of this cross-ratio, we see that:

The distance-function $d(A, B)$ *for fixed* A *is a monotone increasing function of the position of the p-point* B *on the p-line* $A\beta$. *As* B *moves from* A *towards* β, $d(A, B)$ *increases steadily from* 0 *to* $+\infty$.

Of course, a similar result is true if A moves steadily from B towards α.

It follows that if we have four *p*-points A, B, C and D (Fig.58.4), and we know that

$$d(A,B)>d(C,D)$$

and we find the unique M-transformation which maps A on C, and the direction along the *p*-line AB at A onto the direction along the *p*-line CD at C (Thm 55.2), the point B will be mapped onto a *p*-point on the *p*-segment $D\beta'$; that is, it will fall on a point *outside* the *p*-segment CD.

Hence, there is an exact equivalent in our *p*-geometry to the Euclidean construction: ‘Mark off from the segment CD, or CD produced, a length equal to a given length AB.’

Suppose now that A, B and C are *p*-points (Fig.58.5), and we consider the *p*-lines through *B* which intersect the *p*-line *AC* in *p*-points which lie between *A* and *C*. The *p*-line *BA* intersects the *p*-line *AC* at *A*, and if *B′* be the *p*-point of intersection of a *p*-line through *B* and the *p*-line *AC*, we see that as *B′* moves continuously towards *C* from *A*, the *p*-angle *ABB′* increases continuously. We also note that if *B′* be any *p*-point in the *p*-segment *AC*, the *p*-line *BB′* lies inside the *p*-angle *ABC* and the *p*-angle *ABB′* is less than the *p*-angle *ABC*.

With these preparatory remarks *we may now prove theorems in our p-geometry with the terminology and the diagrams of Euclidean geometry*.

We shall not, of course, use the Euclidean parallel axiom!

Theorem II *In a p-triangle ABC, if d*(*A*, *B*) = *d*(*A*, *C*), *then the p-angles ABC and ACB are congruent.*

Proof. By the General SAS Theorem of §57.1, if we consider the two triangles *ABC* and *ACB*, we have $AB \overset{p}{=} AC$, $AC \overset{p}{=} AB$,, and $\sphericalangle BAC \overset{p}{=} \sphericalangle CAB$. The two triangles are therefore *p*-congruent, and so $\sphericalangle ABC \overset{p}{=} \sphericalangle ACB$.

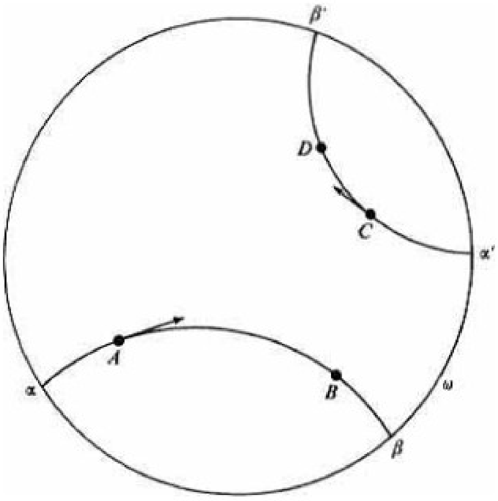

Fig. 58.4

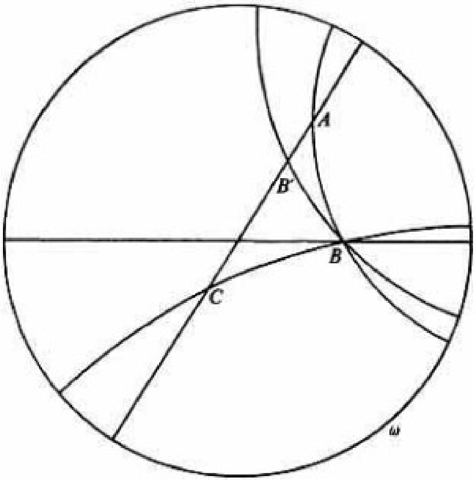

Fig. 58.5

Note that triangles *ABC* and *ACB* have opposite orientation, so that we must use the General SAS Theorem. A direct proof is indicated in Exercise 58.4.

Theorem III *In any p-triangle ABC, the greatest side is opposite the greatest angle*

Proof. Suppose that $d(A, C) > d(A, B)$ (see Fig. 58.6), and mark off a *p*-segment AB' on *p*-line *AC* equal to *AB*, so that $d(A, B') = d(A, B)$. Then B' lies in the segment *AC*, and the *p*-line BB' lies inside the angle *ABC*. Since $d(A, B') = d(A, B)$, the angles ABB' and $AB'B$ are equal. Therefore angle *ABC* is greater than angle $AB'B$, since angle *ABC* is greater than angle ABB'. But angle $AB'B$ is the exterior angle to angle $BB'C$ in *p*-triangle $BB'C$, and therefore exceeds the interior angle $B'CB$ which is the angle *ACB* (Thm 58.2). Hence angle *ABC* is greater than angle *ACB*.

Corollary. In a right-angled *p*-triangle, the longest side is that opposite the right angle.

Proof. Since the angle-sum of a *p*-triangle is less than two right angles, (Thm 58.1), the right angle is the largest angle.

We can now prove the triangle inequality for *p*-triangles and *p*-distances.

Theorem IV *In any p-triangle ABC*, $d(A, B) + d(B, C) > d(A, C)$.

Proof. If $d(A, C) \geq d(A, B)$ and $d(A, C) \geq d(B, C)$, then

$$d(A,C) + d(C,B) > d(A,B) \text{ and } d(A,C) + d(A,B) > d(B,C),$$

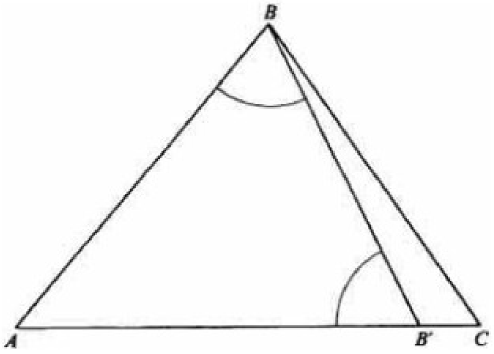

Fig. 58.6

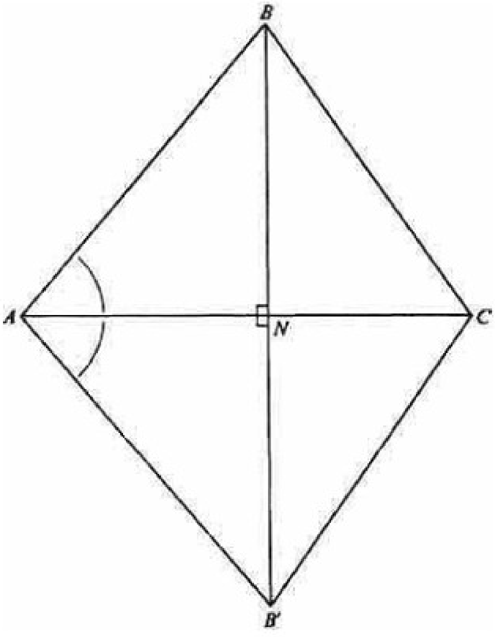

Fig. 58.7

so that if *AC* is the longest side of the *p*-triangle *ABC* (Fig.58.7), the proof of the theorem is only necessary with this assumption for the general inequality to be true. On the opposite side of the *p*-line *AC* to *B* mark off angle *B′AC* = angle *BAC*, where $d(A, B') = d(A, B)$. Let the intersection of the *p*-lines *BB′* and *AC* be *N*. The *p*-triangles *ANB* and *ANB′* are *p*-congruent, by the S A S theorem (Thm

57.1, II), and since the sum of the angles BNA, $B'NA$ is two right angles, both BNA and $B'NA$ are right angles. Therefore $d(A, B) > d(A, N)$, and $d(B, C) > d(N, C)$. Hence, *if N lies in the segment AC*,

$$d(A,B) + d(B,C) > d(A,N) + d(N,C) = d(A,C)$$

(by Thm I, above).

If N lay outside the p-segment AC, we still have $d(A, B) > d(A, N)$, and since in this case $d(A, N) > d(A, C)$, we should have $d(A, B) > d(A, C)$, which is a contradiction to our initial hypothesis.

Exercises

58.1 Show that the stronger form of the triangle inequality is: If A, B and C are three p-points, then $d(A, B) + d(B, C) \geq d(A, C)$, with equality if and only if A, B and C lie, in this order, on a p-line.

58.2 Does Thm III of §58.3 imply Thm II of the same Section? (Examine the proofs of both Theorems carefully before answering.)

58.3 If A is at the center of ω, and B, C are p-points such that $d(A, B) = d(A, C)$, show that the line-segments AB and AC are equal in length from the Euclidean point of view, and deduce that $\sphericalangle ABC = \sphericalangle ACB$. .

58.4 Using the preceding Exercise, prove that if in a p-triangle ABC we have $d(A, B) = d(A, C)$, then the p-angles ABC and ACB are equal.

58.5 Prove the ASA Theorem: (If for two triangles ABC, $A'B'C'$ (p-triangles, of course) $AB \overset{p}{=} A'B'$, $\sphericalangle BAC \overset{p}{=} \sphericalangle B'A'C'$ and $\sphericalangle ABC \overset{p}{=} \sphericalangle A'B'C'$, then the triangles are p-congruent).

58.6 Prove the SSS Theorem: (If for two p-triangles ABC, $A'B'C'$ we have $AB \overset{p}{=} A'B'$, $BC \overset{p}{=} B'C'$ and $CA \overset{p}{=} C'A'$ then the corresponding angles of the two triangles are also p-congruent). (Hint: Copy triangle $A'B'C'$ onto triangle ACQ, where $AQ \overset{p}{=} A'B'$, $\sphericalangle CAQ \overset{p}{=} \sphericalangle C'A'B'$, and Q is on the opposite side of AC to B. Prove triangles ABC, AQC are p-congruent, and deduce the theorem.)

58.7 Prove the SAA Theorem: (If for two p-triangles ABC, $A'B'C'$ we have $AB \overset{p}{=} A'B'$, $\sphericalangle CAB \overset{p}{=} \sphericalangle C'A'B'$ and $\sphericalangle ACB \overset{p}{=} \sphericalangle A'C'B'$, then the two triangles are p-congruent.)

59.1 Similar *p*-triangles are *p*-congruent

We have already seen (Thm 58.1) that in our p-geometry the sum of the angles of a p-triangle is always less than π, whereas in Euclidean geometry the angle-sum is always equal to π. Perhaps an even more surprising difference is the fact that a p-triangle is completely determined by its angles.

Theorem *Two p-triangles are p-congruent if the three angles of the one are respectively equal to the three angles of the other.*

Remark. We have established above that we may prove theorems in our p-geometry with the terminology and diagrams of Euclidean geometry, and have

*.

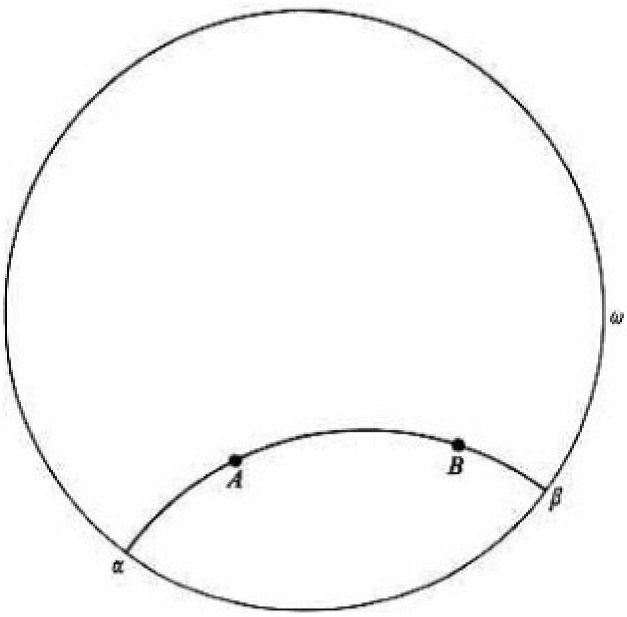
ω
A
B
β
α

/ .

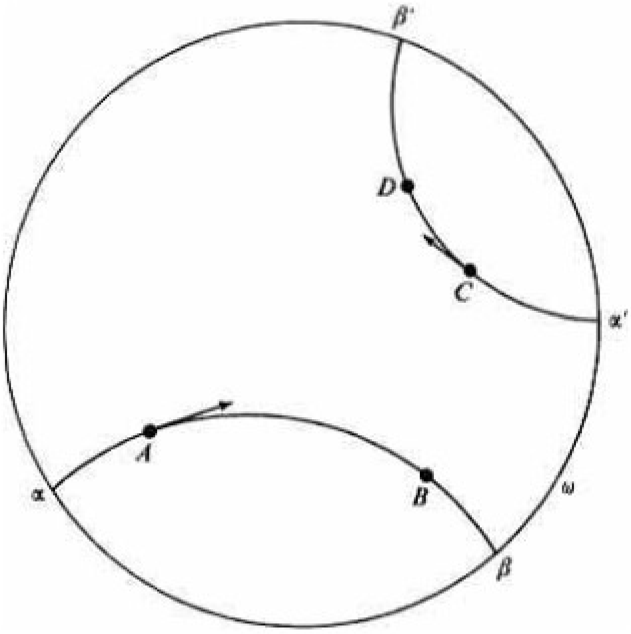
β′
D
C
α′
A
α
B
ω
β

/ . :

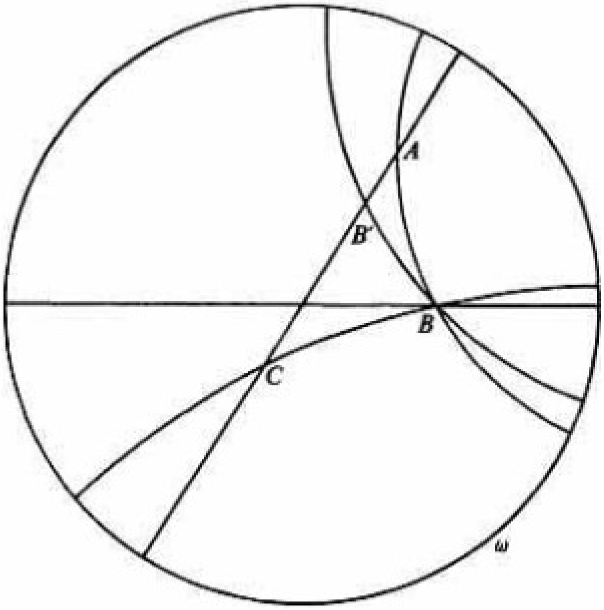

/ .

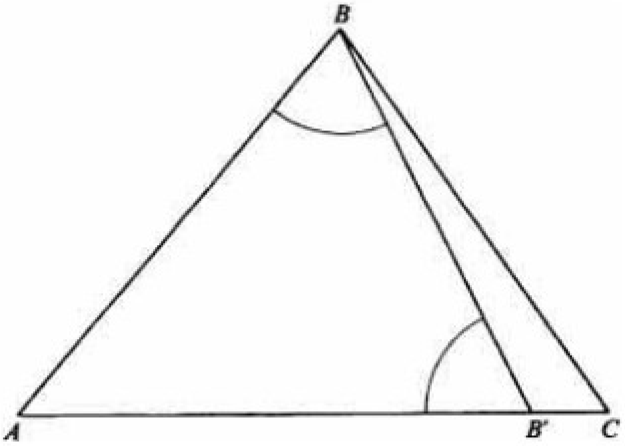

/ . 8

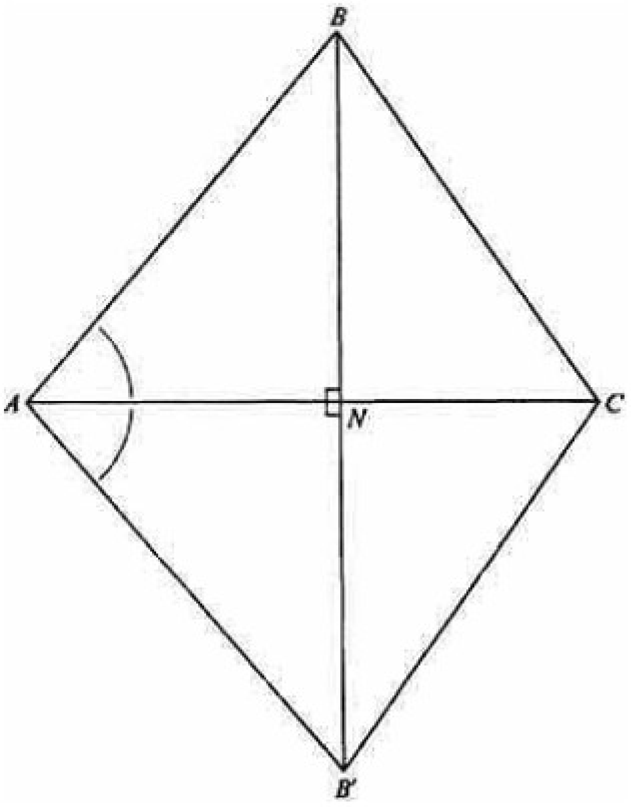
B
A
N
C
B′

/ .

57.1, II), and since the sum of the angles BNA, $B'NA$ is two right angles, both BNA and $B'NA$ are right angles. Therefore $d(A, B) > d(A, N)$, and $d(B, C) > d(N, C)$. Hence, *if N lies in the segment AC*,

$$d(A,B) + d(B,C) > d(A,N) + d(N,C) = d(A,C)$$

(by Thm I, above).

If N lay outside the p-segment AC, we still have $d(A, B) > d(A, N)$, and since in this case $d(A, N) > d(A, C)$, we should have $d(A, B) > d(A, C)$, which is a contradiction to our initial hypothesis.

Exercises

58.1 Show that the stronger form of the triangle inequality is: If A, B and C are three p-points, then $d(A, B) + d(B, C) \geq d(A, C)$, with equality if and only if A, B and C lie, in this order, on a p-line.

58.2 Does Thm III of §58.3 imply Thm II of the same Section? (Examine the proofs of both Theorems carefully before answering.)

58.3 If A is at the center of ω, and B, C are p-points such that $d(A, B) = d(A, C)$, show that the line-segments AB and AC are equal in length from the Euclidean point of view, and deduce that $\sphericalangle ABC = \sphericalangle ACB$. .

58.4 Using the preceding Exercise, prove that if in a p-triangle ABC we have $d(A, B) = d(A, C)$, then the p-angles ABC and ACB are equal.

58.5 Prove the ASA Theorem: (If for two triangles ABC, $A'B'C'$ (p-triangles, of course) $AB \overset{p}{=} A'B'$, $\sphericalangle BAC \overset{p}{=} \sphericalangle B'A'C'$ and $\sphericalangle ABC \overset{p}{=} \sphericalangle A'B'C'$, then the triangles are p-congruent).

58.6 Prove the SSS Theorem: (If for two p-triangles ABC, $A'B'C'$ we have $AB \overset{p}{=} A'B'$, $BC \overset{p}{=} B'C'$ and $CA \overset{p}{=} C'A'$ then the corresponding angles of the two triangles are also p-congruent). (Hint: Copy triangle $A'B'C'$ onto triangle ACQ, where $AQ \overset{p}{=} A'B'$, $\sphericalangle CAQ \overset{p}{=} \sphericalangle C'A'B'$, and Q is on the opposite side of AC to B. Prove triangles ABC, AQC are p-congruent, and deduce the theorem.)

58.7 Prove the SAA Theorem: (If for two p-triangles ABC, $A'B'C'$ we have $AB \overset{p}{=} A'B'$, $\sphericalangle CAB \overset{p}{=} \sphericalangle C'A'B'$ and $\sphericalangle ACB \overset{p}{=} \sphericalangle A'C'B'$, then the two triangles are p-congruent.)

59.1 Similar *p*-triangles are *p*-congruent

We have already seen (Thm 58.1) that in our p-geometry the sum of the angles of a p-triangle is always less than π, whereas in Euclidean geometry the angle-sum is always equal to π. Perhaps an even more surprising difference is the fact that a p-triangle is completely determined by its angles.

Theorem *Two p-triangles are p-congruent if the three angles of the one are respectively equal to the three angles of the other.*

Remark. We have established above that we may prove theorems in our p-geometry with the terminology and diagrams of Euclidean geometry, and have taken advantage of this freedom in the proofs of Theorems II and III and IV in §58.3. We continue in the same vein in the following proof.

Proof. In Fig. 59.1 suppose that $\sphericalangle A = \sphericalangle A'$, $\sphericalangle B = \sphericalangle B'$, $\sphericalangle C = \sphericalangle C'$ but that the p-triangles ABC and $A'B'C'$ are not p-congruent. Then we may assume that $d(A, B) \neq d(A', B')$, and that in fact $d(A, B) > d(A', B')$. On AB and AC mark off lengths equal to $A'B'$ and $A'C'$ respectively, calling the endpoints on AB and AC the points D and E. Then D certainly lies in the segment AB, but E may lie on C, or on AC produced, or in the segment AC. If E falls on C, then by the SAS Theorem (§57.1) the triangles ADC and $A'B'C'$ are congruent, so that the angle ADC = angle $A'B'C'$ = angle ABC. But by Thm 58.2, angle ADC > angle ABC, Hence E cannot fall on C. If E falls on AC produced then again triangles ADE and $A'B'C'$ are congruent, and angle AED = angle $A'C'B'$ = angle ACB, and by Thm 58.2 we should have angle ACB > angle AED. Hence E can only fall between A and C. If we now consider the quadrilateral $BCED$, the sum of the angles is 2π. But this contradicts Thm 58.1, since a quadrilateral can be divided into two triangles, and the angle-sum of one of the triangles would be $\geq \pi$. Hence the original triangles ABC, $A'B'C'$ are p-congruent.

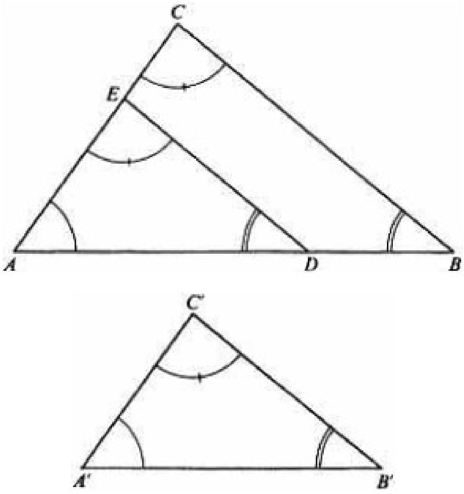

Fig. 59.1

59.2 *p*-circles in the *p*-plane

There are some obvious constructs in our p-geometry which merit investigation.

The definition of a *p*-circle is evidently as follows:

Definition. The locus of a *p*-point which moves so that its *p*-distance from a given *p*-point A is a constant r is a *p*-circle, center (*p*-center) A and *p*-radius r.

Theorem I *A p-circle, center A is a Euclidean circle orthogonal to the family of p-lines which pass through A.*

Proof. The *p*-lines through A all pass through the fixed point A' (Fig.59.2), where A' is the inverse of A in ω. Let P be the point whose locus we are considering, and $\mathscr{C}$ the *p*-line through P and A. This passes through A'. We wish to find the locus of P, subject to the condition $|\log(\alpha, \beta: A, P)| = r$, for a constant r, where α, β are the points in which $\mathscr{C}$ intersects ω. If we invert with respect to a circle, center A', orthogonal to ω, ω inverts into itself, the inside of ω inverts into the inside of ω, and A inverts into the center O of ω. The circle $\mathscr{C}$ inverts into the line OP', where P' is the inverse of P. Let the diameter OP' meet ω in the points α', β', which are the inverses of α, β. We know that although the inversion is a conjugate Moebius transformation, the cross-ratios $(\alpha, \beta; A, P)$ and $(\alpha', \beta'; O, P')$ are equal (Exercise 57.10). Hence

$$|\log(\alpha',\beta';o,P')| = |\log(\alpha,\beta A,P)| = r$$

This is readily seen to imply that P' is at a given distance from O, on either side of O, and therefore the locus of P' is a circle in the Euclidean sense, center O. This circle is orthogonal to the pencil of lines through O, and therefore arises from the Euclidean circle which is orthogonal to the family of circles through A and A' in the original figure.

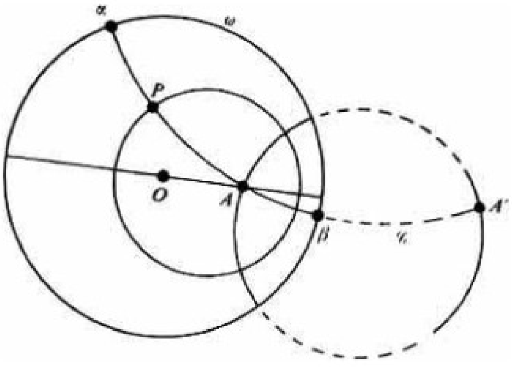

Fig. 59.2

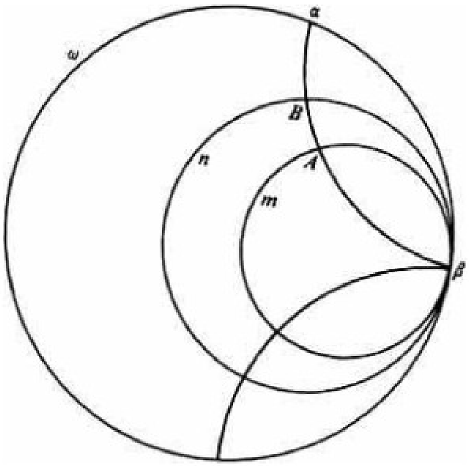

Fig. 59.3

We note that the Euclidean center of this circle only coincides with the p-center A in one case. We leave the elucidation of this happening as an Exercise. (Exercise 59.3.)

Again, as the center A moves towards ω, and we keep P fixed, the p-distance $d(AP)$ tends towards infinity, and the p-circle center A tends towards a Euclidean circle which touches ω internally. Such a circle is called a *horocycle*.

Theorem II *Two horocycles tangent to* ω *at the same point* β *cut off equal p-distances on the p-lines through* β.

Proof. Let m and n be two horocycles touching ω at the point β (Fig.59.3), and suppose that a p-line through β meets ω again in the point α, and intersects m in the p-point A and n in the p-point B. We know (Exercise 57.11) that inversion is a conjugate Moebius transformation. We also know (Exercise 57.10) that p-lengths are unaffected by a conjugate Moebius transformation. Let us therefore invert the given configuration in a circle center β (Fig.59.4). The circles ω, m and n invert into parallel lines ω', m' and n', the p-line AB inverts into a line $\alpha'B'A'$, where α' is the point where it meets ω', B' is the point where it meets n', and A' is the point where it meets m'. The point β itself has inverted into the point ∞. What has happened to the cross-ratio $(\alpha, \beta; A, B)$? It has become the cross-ratio $(\alpha', \infty; A', B')$, and this is simply the ratio $\overline{\alpha'A'}/\overline{\alpha'B'}$. But for fixed parallel lines we know that this is constant. Hence $d(A, B)$ is independent of the particular p-line through β which cuts A and B on the given horocycles.

An alternative proof is indicated in Exercise 59.1.

It has been made clear that to the possible two-dimensional inhabitants of our space Ω, the boundary circle ω is unattainable, and corresponds to the line at infinity in the real Euclidean plane. Poincaré suggested physical conditions which might obtain in Ω which would make ω unattainable. He imagined a variation of temperature in Ω which would make the dimensions of the two-dimensional inhabitants shrink to zero as they approach ω, the Great Divide. As they shrink, their steps also shrink, and the law of shrinkage can be so chosen that the geodesic, or shortest path, in the Euclidean sense, between two points in Ω is precisely the *p*-line through the two points. The Poincaré model is an illustration of Hamlet's: 'I could be bounded in a nutshell and count myself a king of infinite space, were it not that I have bad dreams'.

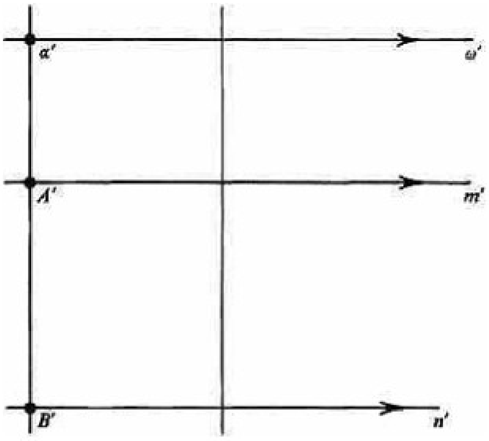

Fig. 59.4

We have seen that in our hyperbolic non-Euclidean geometry there are two *p*-parallels through a given *p*-point to a given *p*-line. It is possible to have a non-Euclidean geometry in which there are *no* parallels through a given point to a line.

We shall not investigate other types of non-Euclidean geometry. The Poincaré model, which we now leave, is a model of *hyperbolic* non-Euclidean geometry, and a non-Euclidean geometry with no parallels through a point to a given line is called an *elliptic* non-Euclidean geometry. In Exercise 59.2 a model for such a geometry is introduced. The invention of non-Euclidean geometries in the nineteenth century was one of the great leaps forward of the human mind. We can well understand the excitement of Janos Bolyai, one of the inventors of non-Euclidean geometry, when he wrote to his father, who first taught him mathematics: 'Out of nothing I have created a new, different, world!' For further explorations of this different world we refer the interested reader to other books (**4, 19, 20, 23**).

We have listed the axioms of Euclid in an Appendix (p. 432) so that the reader may be able to refer to them, and compare the Euclidean theorems deduced from them with those we have obtained. It is hoped that the impact of a non-Euclidean geometry may serve to interest the reader in the foundations of geometry.

In our next chapter we discuss *projective geometry*. In this geometry two lines in a plane always intersect, and so this geometry also ranks as a non-Euclidean geometry.

Exercises

59.1 Let m and n be two horocycles touching ω at the point β, and suppose that one p-line through β meets m again in A, and n again in B, and that another p-line through β meets m again in A', and n again in B'. Show that, by an inversion in a suitable circle of the coaxal system defined by the two p-lines, the p-line βAB can be inverted into the p-line $\beta A'B'$, and that this inversion maps ω and the horocycles onto themselves, ω onto ω, m onto m and n onto n, so that A is mapped onto A' and B onto B'. Deduce, that since an inversion is a conjugate Moebius transformation, $d(A, B) = d(A', B')$.

59.2 S is the surface of a sphere in real three-dimensional Euclidean space, and O is the center. Show that we have a model of an elliptic non-Euclidean geometry if we take S as our Plane, and identify a *pair* of points on S which are diametrically opposed (that is, their join passes through the center) as a Point in the Plane. The Lines in our Plane are to be given by the intersections with S of Euclidean planes through O. Show that two distinct Points determine a unique Line, and that two distinct Lines always intersect in a unique Point, so that there are no parallels to a given Line through a given Point which does not lie on it.

59.3 The points A and A' are inverse in a circle ω, A lying inside ω, and $\mathscr{C}$ is a circle of radius >0 which lies inside ω and is a member of the coaxal system polar to the coaxal system defined by circles which pass through A and A'. Show that if A' is a finite point, the center of $\mathscr{C}$ is never at A. When is the center of $\mathscr{C}$ at A?

59.4 Consider a Euclidean half-plane Ω bounded by the line ω, and all (semi-) circles in Ω which have their centers on ω, together with the set of all lines in Ω perpendicular to ω. Show how to set up a model of a hyperbolic geometry with this arrangement. What is the connexion with the model discussed in § 56.1?

VII

THE PROJECTIVE PLANE AND PROJECTIVE SPACE

We shall now investigate a geometry which is more fundamental than Euclidean geometry, in the sense that there will be no mention of metrical properties, distances or angle-measure, and the only concept we shall make use of at first is that of *incidence*. This geometry is called *projective geometry*, since its properties are invariant under *projection*, which is a geometrical transformation we shall investigate later. We shall enquire into the properties of projective space of n dimensions, after a special study of the cases $n = 2$ and $n = 3$, and since our treatment will be algebraic (non-algebraic treatments are also possible), we must say what kind of algebraic construct our numbers are to be taken from.

We shall work in a *commutative field*, although some of our theorems are also valid in a *skew* (non-commutative) field. For most of our purposes it will suffice to assume that our field is that of *the complex numbers*.

The outstanding property of this field which we may want to use is sometimes called *the fundamental theorem of algebra*: any polynomial with coefficients in the field of complex numbers splits into linear factors with coefficients which are complex numbers.

60.1 The complex projective plane

We must define the points of our geometry. A *point* is defined as an ordered triad of complex numbers (y_0, y_1, y_2) *not all zero*, with the understanding that the ordered triad (y_0k, y_1k, y_2k), where k is *any non-zero complex number*, represents the same point.

We must show that this relation is *an equivalence relation*, which is easily done, but this becomes evident if we remark that the ordered triad (y_0, y_1, y_2) represents a point distinct from the origin in three-dimensional complex affine space, and the set of points (y_0k, y_1k, y_2k) as k varies is the *ray*, or line through the origin joining the origin to the point. We are therefore defining a *line* or *ray* through the origin to be a *point* in our geometry, which is that of the complex projective plane.

The coordinates (y_0, y_1, y_2) we are using are usually called *homogeneous coordinates*, since only the ratios $y_0{:}y_1{:}y_2$ of the coordinates are significant, these being the same for the point (y_0k, y_1k, y_2k).

A *line* in our complex projective plane is defined to be the subset of points in the plane whose coordinates satisfy a linear homogeneous equation

$$u^0X_0 + u^1X_1 + u^2X_2 = 0,$$

where not all the u^i $(i = 0, 1, 2)$ are zero. Since, multiplying on the right by $k \neq 0$ we also have

$$u^0(X_0k) + u^1(X_1k) + u^2(X_2k) = 0,$$

it follows that if one member of the equivalence class representing a point of our projective plane satisfies the equation, so does every member.

We note that in three-dimensional affine space the equation above is that of *a plane through the origin*. Hence rays through the origin are to be points of the projective plane, and *planes* through the origin are to be *lines* of the projective plane. A *point* of the projective plane is *incident* with, or *lies on* a *line* of the projective plane if and only if the *ray* corresponding to the point lies in the *plane* corresponding to the line.

We now prove two incidence theorems which explain why our algebraic formulation is chosen in what may appear, at first sight, to be a curious manner.

Theorem I *In the projective plane, two distinct points determine a unique line with which they are incident.*

Theorem II *In the projective plane two distinct lines intersect in a unique point.*

Remark. It is the second theorem which distinguishes this geometry from ordinary Euclidean geometry, where lines which are distinct may not intersect. In this geometry distinct lines *always* intersect.

Proof. We prove the second theorem first. Let

$$u^0X_0 + u^1X_1 + u^2X_2 = 0 = \upsilon^0X^0 + \upsilon^1X_1 + \upsilon^2X_2.$$

be the equations of the two lines. Since they are distinct, we do *not* have

$$u^0:u^1:u^2 = \upsilon^0:\upsilon^1:\upsilon^2.$$

Solving the two homogeneous equations for the point of intersection of the two lines gives us as point of intersection:

$$(u^1\upsilon^2 - u^2\upsilon^1, u^2\upsilon^0 - u^0\upsilon^2, u^0\upsilon^1 - u^1\upsilon^0).$$

This triad could only fail to be a point if *each* of the elements of the triad displayed above were zero. This could only be the case if

$$[u^0:u^1:u^2] = k[\upsilon^0,\upsilon^1,\upsilon^2]\ (k\neq 0),$$

which would mean that the two lines are identical, and this possibility is contrary to hypothesis.

To prove Thm I we remark that if the two distinct points are

$$(x_0,x_1,x_2) \text{ and } (y_0,y_1,y_2),$$

then we do *not* have

$$(x_0,x_1,x_2) = (y_0k,y_1k,y_2k)\ (k\neq 0),$$

and the line

$$u^0X_0 + u^1X_1 + u^2X_2 = 0$$

contains both points, where

$$u^0 = x_1y_2 - x_2y_1,\ u^1 = x_2y_0 - x_0y_2,\ u^2 = x_0y_1 - x_1y_0,$$

as can be immediately verified. This line exists, since we do not have $u^0 = u^1 = u^2 = 0$, the two points (x_0, x_1, x_2) and (y_0, y_1, y_2) being distinct points. Hence there is *a* line through the two points. There cannot be more than one, since we should then have two distinct lines intersecting in more than one point, contrary to Thm II.

To Theorems I and II we add a third, designed to ensure that the geometry we are considering is not trivial.

Theorem III The projective plane contains at least four distinct points, no three of which are collinear.

Proof. The points (1, 0, 0), (0, 1, 0), (0, 0, 1) and (1, 1, 1) are easily seen to satisfy this condition over the field of the complex numbers, since the first three points lie, in pairs, on the lines $X_2 = 0$, $X_0 = 0$ and $X_1 = 0$ respectively, and the fourth point does not lie on any of these lines.

It will be noticed that we use capital letters for the variables in an equation, lowercase letters for actual coordinates of points, and also for what we may evidently call *the coordinates of a line*: $[u^0, u^1, u^2]$, the line being

$$u^0X_0 + u^1X_1 + u^2X_2 = 0.$$

We note, that just as with points, the ordered triad $[ku^0, ku^1, ku^2]$ $(k \neq 0)$ represents the same line as does the ordered triad $[u^0, u^1, u^2]$.

60.2 The principle of duality in the projective plane

The symmetrical nature of the incidence relation

$$u^0y_0 + u^1y_1 + u^2y_2 = 0.$$

which expresses the fact that the line $[u^0, u^1, u^2]$ *contains* the point (y_0, y_1, y_2), or equivalently, that the point (y_0, y_1, y_2) *lies on* the line $[u^0, u^1, u^2]$ gives rise to the *Principle of Duality* in the projective plane. This asserts that by an automatic interchange of the terms *point* and *line*, *lying on* and *passing through*, *join* and *intersection*, *collinear* and *concurrent*, and so on, any theorem in the projective plane which involves only incidence properties of points and lines becomes, on transliteration with the help of the dictionary of interchanges we have just listed, a theorem involving lines and points.

To give an immediate example, the theorem of Desargues, which we have already encountered (§5.2) and which is essentially a theorem of the projective plane, says that if ABC, $A'B'C'$ are two triangles in the projective plane which are such that the lines AA', BB' and CC' are concurrent (that is, pass through the same point), then the three points of intersection of corresponding sides, $BC \cap B'C'$,$CA \cap C'A'$ and $AB \cap A'B'$are collinear (lie on a line).

The theorem obtained by applying the Principle of Duality says that if abc, $a'b'c'$ are triangles (we should change the term *triangle*, which refers to a configuration formed by three *points*, to *triline*, to be precise, since we now consider two configurations formed by triads of lines, a, b and c and a', b' and c', these not being concurrent: but we shall not do this) in the projective plane which are such that the points aa', bb' and cc' are collinear (lie on a line) then the lines joining corresponding vertices, $bc \cup b'c'$,$ca \cup c'a$and $ab \cup a'b'$ are concurrent (pass through a point).

We shall see, in fact, that exactly the same algebra which we shall use to establish the Desargues Theorem will establish the dual theorem, which is the converse of the Desargues Theorem (§62.2).

It will be noted that we are using the symbol AB to denote the *line* joining two distinct points A and B, and dually the symbol ab to denote the *point* of intersection of the two distinct lines a and b. Since we use the symbol $\cap$ for the intersection of two point-sets, the point of intersection of two lines AB and $A'B'$ is written as $AB \cap A'B'$, and therefore dually we write $ab \cup a'b'$ for the line joining the points ab and $a'b'$, although the $\cup$ symbol here does *not* stand for the join of two point-sets. Another symbol, if used, would also create difficulties. Some authors use the symbol $A + B$ for the set of points on the line joining A to B. We shall not do this, and we do not think that there will be any misunderstanding in interpreting our symbolism.

The reader may wonder why we talk of the Principle of Duality, when our statement describing it amounts to a proof. Historically the statement was put in the form of a Principle. For us it is almost a self-evident theorem.

Exercises

60.1 Show that the equation of the line joining the two points (x_0, x_1, x_2) and (y_0, y_1, y_2) is the result of eliminating U^0, U^1, U^2 from the three equations $U^0X_0 + U^1X_1 + U^2X_2 = 0$, $U^0x_0 + U^1x_1 + U^2x_2 = 0$, $U^0y_0 + U^1y_1 + U^2y_2 = 0$, and is therefore the determinantal equation

$$\begin{vmatrix} X_0 & X_1 & X_2 \\ x_0 & x_1 & x_2 \\ y_0 & y_1 & y_2 \end{vmatrix} = 0.$$

60.2 Write out the dual of the Exercise above, changing the phrase ‘equation of the line’ to ‘equation of the point’. (Note that the equation of a *line* is the condition that a point (X_0, X_1, X_2) lie on a line $[u^0, u^1, u^2]$. The equation of a *point* is the condition that a line $[U^0, U^1, U^2]$ pass through a point (x_0, x_1, x_2).)

60.3 Four distinct points A, B, C and D, no three of which are collinear, are said to form a complete quadrangle (or four-point), with *sides* formed by joining the points in pairs, and with *diagonal triangle* formed by the *diagonal points*, which are the intersections of the sides, other than A, B, C and D. Write down the dual of this definition, drawing a diagram in each case.

60.4 The Pappus Theorem says that if A, B, C are three distinct points on a line l, and A', B', C' three distinct points on a distinct line m, then the three points $BC' \cap B'C, CA' \cap C'A$ and $AB' \cap A'B$ are collinear. Write down the dual of this theorem, drawing a diagram in each case.

60.5 A set of points $(P, Q, R, \ldots,)$ on a line l is said to be *in perspective* with a set of points $(P', Q', R', \ldots)$ on a line l', if the joins PP', QQ', RR', . . . all pass through a point V. Formulate a dual concept for lines $(p, q, r, \ldots)$ on a point L, and lines $(p', q', r', \ldots)$ on a point L'.

60.6 The points on the surface of a given sphere which are diametrically opposed are identified as Points, and great circles cut on the surface of the sphere by planes through the center are identified as Lines. Show that these Points and Lines satisfy Theorems I, II and III of §60.1, the ground field being that of the real numbers.

60.7 Show that the points $A = (1, 0, 0)$, $B = (0, 1, 0)$, $C = (0, 0, 1)$, $D = (0, 1, 1)$, $E = (1, 0, 1)$, $F = (1, 1, 0)$ and $G = (1, 1, 1)$, where the field is that of the integers, mod 2 (that is $1 + 1 = 0$) form a finite projective geometry, with three points on every line and three lines through every point (See Pedoe, **12,** p. 85).

60.8 If all the elements involved, with the exception of the symbol ∞, are real numbers, show that the set of points and lines described satisfy Thms. I, II and III of §60.1. The points are: (*i*) ordinary points with coordinates (a, b); (*ii*) infinite points (m), together with the point ∞. The lines are given by the equations $X = c$, and $Y = mX + b$, and there is also the line at infinity ω. The line at infinity contains the point ∞ and all points (m), the line $X = c$ contains the point ∞ and all points (c, d), and the line $Y = mX + b$ contains the point (m) and all the points $(a, ma + b)$.

61.1 A model for the projective plane

We have already associated a *ray* through the origin of a three-dimensional space with a *point* of our projective plane. We pursue this further. Let V_3 be the complex vector space of vectors $y = (y_0, y_1, y_2)$, where the y_i are complex numbers. If $y \neq (0, 0, 0)$, the zero vector, then we may define a *ray* of V_3 as the set of vectors $yk = (y_0k, y_1k, y_2k)$ $(k \neq 0)$. We say that a *point* of our complex projective plane is a *ray* of V_3.

Two vectors $y = (y_0, y_1, y_2)$ and $z = (z_0, z_1, z_2)$ of V_3 are said to be *linearly dependent* if there exist complex numbers λ and μ, not both zero, so that

$$y\lambda + z\mu = 0$$

where 0 is the zero vector. If such a relation necessarily implies that $\lambda = \mu = 0$, we say that y and z are *linearly independent*. Since a point of the complex projective plane is defined to be a ray of V_3, we see that if two non-zero vectors y and z are linearly dependent, so that they define the same ray of V_3, then they also define the same point of our complex projective plane.

If y and z are *independent*, and therefore neither is the zero vector, each defines a ray of V_3, and the set of vectors linearly dependent on y and z form a subspace of V_3. In fact, if $x = y\lambda + z\mu$, and we write down the three equations for the coordinates of the vectors on each side of this equation, we have

$$\begin{aligned} y_0\lambda + z_0\mu - x_0 &= 0, \\ y_1\lambda + z_1\mu - x_1 &= 0, \\ y_2\lambda + z_2\mu - x_2 &= 0, \end{aligned}$$

and on eliminating λ and μ from these equations we have the determinantal equation

$$\begin{vmatrix} X_0 & X_1 & X_2 \\ y_0 & y_1 & y_2 \\ z_0 & z_1 & z_2 \end{vmatrix} = 0.$$

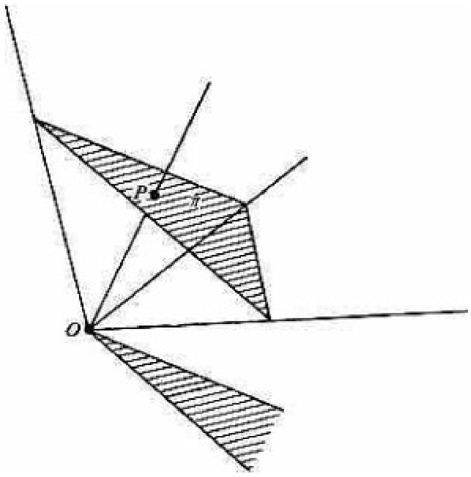

Fig. 61.1

This is a linear homogeneous equation in (X_0, X_1, X_2), and represents both the *plane* in V_3 defined by two distinct rays $y\lambda$ and $z\mu$, and the *line* in our projective plane which joins the two distinct points y and z.

It is clear that we may use a single symbol such as y or z for a point in our projective plane, with the understanding that $y\lambda$ represents the same point as y. We have seen that the set of points linearly dependent on the points y and z is the set $y\lambda + z\mu$. This gives the set of points on the *line* joining y and z.

If we assume that we are perfectly familiar with vector spaces V_3 over the complex number field, we have set up a *model* for the complex projective plane, since all the mathematical properties of the projective plane can be described or verified for V_3, and we assume that the mathematical properties of V_3 are free from contradiction. If we wish to *see* something, and this is what a model is for, let us consider how to set up homogeneous coordinates in a plane in real affine space. Call the plane π, and choose a system of three-dimensional coordinates for which the origin $O = (0, 0, 0)$ does not lie in π (Fig.61.1). Then a ray through O which is not parallel to π intersects π in a unique point P, and if the ray is the set of points (y_0k, y_1k, y_2k), where k varies over the real number field, the point P can be assigned the coordinates (y_0, y_1, y_2), where (y_0k, y_1k, y_2k) are also the coordinates of the same point P. This is precisely the development already given.

Here there is a difference, since rays through O may be parallel to π. If such a ray is $(z_0, z_1, z_2)k$, we say that the point (z_0, z_1, z_2) is *an ideal point* in π, or *a point at infinity*. If the equation of the plane π is $u^0X_0 + u^1X_1 + u^2X_2 = \text{constant}$, in our system of three-dimensional coordinates, the equation of the plane through O which is *parallel* to π is simply the equation $u^0X_0 + u^1X_1 + u^2X_2 = 0$, and we may take this same equation as *the equation of the line at infinity in the plane* π.

By introducing ideal points into the plane π, we have made it into a projective plane, in which all points have equal status, and two distinct lines always intersect.

Returning to our complex projective plane, which we shall also label π, we know that the vector space V_3 contains sets of 3 linearly independent vectors. One such set is $E_0 = (1, 0, 0)$, $E_1 = (0, 1, 0)$, $E_2 = (0, 0, 1)$. Vectors are, of course, linearly independent if there is no non-trivial linear homogeneous relation between them; that is, if the vectors be x, y and z, they are linearly independent when any linear relation $x\lambda + y\mu + z\nu = 0$ necessarily implies $\lambda = \mu = \nu = 0$. The vectors E_0, E_1 and E_2 are clearly independent. Since we may write any vector $y = (y_0, y_1, y_2)$ in V_3 as

$$y = E_0y_0 + E_1y_1 + E_2y_2,$$

and this same expression holds for points in π, we have shown that:

The system of homogeneous coordinates in π arises from the expression of a point P being linearly dependent on three fixed points which form a proper triangle in π.

We know that any three independent vectors in V_3 form *a basis* for the vectors of V_3: that is, any vector of V_3 can be expressed linearly in terms of three independent vectors. If the three independent vectors be named E_0^*, E_1^*, E_2^* then any vector y of V_3 has a unique expression in the form

$$y = E_0^*\lambda + E_1^*\mu + E_2^*\nu.$$

Independent vectors in V_3 correspond, if there are two vectors, to two distinct points of π, and if there are three vectors, to three non-collinear points of π, forming a proper triangle. The theory of the basis for vectors in V_3 shows us that if any three non-collinear points be chosen in π, with coordinates E_0^*, E_1^* and

E_2^*, then any point in π may be expressed in the form

$$P = E_0^*\lambda + E_1^*\mu + E_2^*\nu,$$

where the ordered triad (λ, μ, ν) is any one of an equivalence class $(\lambda k, \mu k, \nu k)$ $(k \neq 0)$ (Fig.61.2.)

If we call the triangle we have chosen *the triangle of reference*, we may call (λ, μ, ν) *the coordinates of the point P with respect to the given triangle of reference*.

We shall investigate, in §64.1 the effect of a change of triangle of reference on the coordinates of a given point. But for a given triangle of reference we now have enough apparatus to prove theorems on points, lines, their joins and intersections in the projective plane π.

Since we shall be investigating the projective geometry of n dimensions in §66.1, we have introduced a suffix notation (y_0, y_1, y_2) for our coordinates, but we may now revert to a simpler notation, calling the triangle of reference ABC, and the coordinates of a point in π: (x, y, z).

We note that the coordinates of A, B and C are respectively $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$ referred to ABC as triangle of reference. Any point on BC is $(0, 1, 0)y + (0, 0, 1)z = (0, y, z)$, and in fact the equation of the line BC is $X = 0$.

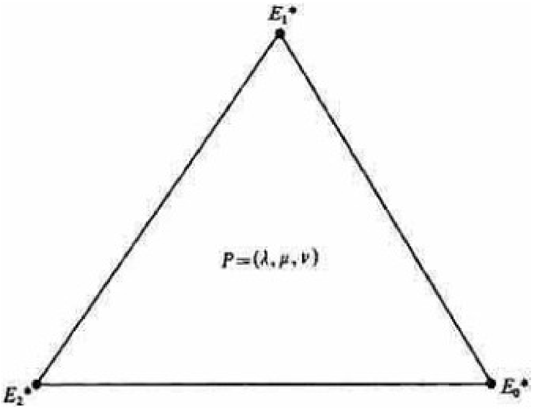

Fig. 61.2

If A' is the point $(0, q, r)$ on BC, any point on the line AA' is given by $(1, 0, 0)p + (0, q, r) = (p, q, r)$. Conversely, if the point (p, q, r) be joined to the vertex A, the intersection of this line with the side BC is the point $(0, q, r)$, and the equation of the line joining A to the point (p, q, r) is simply $rY - qZ = 0$.

The reader will find it interesting to compare the treatment of points and lines given here with that of points and lines in the real Euclidean plane given in Chapter I. Some of the results appear to be similar, but we do not deal with metrical results in the projective plane. We have a greater freedom, since our coordinates are not normalized (for instance, if A, B and C are three non-collinear points, any point P may be expressed as $P = Ax + By + Cz$, without having $x + y + z = 1$, as in Thm 4.2), but we do not have a metrical interpretation.

On the other hand, some theorems, such as the Theorem of Ceva, do have a corresponding theorem in the projective plane, since the metrical interpretation is not fundamental to the theorem.

Example. The lines AP, BP and CP cut the opposite sides of triangle ABC in the points L, M and N respectively. If $L = xB + x'C$, $M = yC + y'A$, $N = zA + z'B$, then

$$xyz = x'y'z'.$$

If $P = (p, q, r)$ with respect to ABC as triangle of reference, then $L = (0, q, r)$, $M = (p, 0, r)$ and $N = (p, q, 0)$, Since we are given that $L = (0, x, x')$, $M = (y', 0, y)$ and $N = (z, z', 0)$, we have

$$x{:}x' = q{:}r,\ y{:}y' = r{:}p,\ z{:}Z' = p{:}q,$$

and therefore

$$(x{:}x')(y{:}y')(z{:}z') = (q{:}r)(r{:}p)(p{:}q) = 1.$$

Example. Theorem of Menelaus (compare Thm 4.1). A line cuts the sides BC, CA and AB of a triangle ABC in the points L, M and N respectively. If $L = Bx + Cx'$, $M = Cy + Ay'$ and $N = Az + Bz'$, then $xyz = -x'y'z'$.

Let the equation of the line be $eX + fY + gZ = 0$. Then L is given by the equation $fY + gZ = 0$, so that $L = (0, -g, f) = B(-g) + C(f)$. Hence $x{:}x' = -g{:}f$. Similarly $y{:}y' = -e{:}g$, and $z{:}z' = -f{:}e$, and the theorem follows immediately.

Exercises

61.1 Prove that if $A' = B + Ck$ is a point on BC, and if M be taken on AA' distinct from A or A', then if BM intersects CA in Q and CM intersects AB in R, the line QR will intersect BC in the point $B - Ck$.

61.2 The sides BC, CA and AB of the triangle of reference have coordinates $[1, 0, 0]$, $[0, 1, 0]$ and $[0, 0, 1]$ respectively. If we write

$$a = [1,0,0],\ b = [0,1,0],\ c = [0,0,1],$$

show that $a' = b + kc$ is a line through the point bc. Write down the theorem dual to that given in Exercise 61.1, and prove it, using essentially the same algebra.

61.3 If ABC be the triangle of reference, show that the equation $lX + nZ = 0$ represents a line through the point B, and the equation $pX + qY = 0$ represents a line through the point C. If the first line intersects the side AC in the point E, and the second line intersects the side AB in the point F, show that the equation of the line EF is $lpX + lqY + npZ = 0$.

61.4 If the lines $uX + \upsilon Y + wZ = 0$, $u'X + \upsilon'Y + w'Z = 0$ intersect at the point P, show that the equation

$$(uX + \upsilon Y + wZ)\gamma + (u'X + \upsilon'Y + w'Z)\mu = 0$$

represents a line which always passes through P, for all values of λ and μ except $\lambda = \mu = 0$. Find the equation of the line which joins P to the vertex A of the triangle of reference.

62.1 The normalization theorem for points on a line

Theorem *If P, Q and R are three distinct collinear points in π, then a coordinate system may be set up in which the homogeneous coordinates of P, Q and R, written as vectors y, z and t, satisfy the equation*

$$t = y + z.$$

Remark. We use a single symbol to represent a vector. When we use a single symbol to represent a point in π, this means that we have chosen a specific representative from the equivalence class of ordered triads which represent the point.

Proof. In any system of homogeneous coordinates where P, Q and R are represented by the vector symbols y', z' and t' respectively, we have a relation

$$y'\lambda + z'\mu + t'\nu = 0,$$

since the three points are linearly dependent. The λ, μ, υ are not all zero. In fact, none of the multipliers is zero, since this would involve two of the points under consideration being identical, and this is excluded. Hence $\nu \neq 0$, and we may write

$$t' = y'\rho + z'\sigma,$$

where neither ρ nor σ is zero. Now write

$$y'\rho = y,\ z'\sigma = z \text{ and } t' = t,$$

so that y is a specific representative of the point P, z a specific representative of the point Q and t a specific representative of the point R, and we have

$$t = y + z.$$

Remark. If we set up a coordinate system on the line for which P and Q are base points, any point on the line being $X = Px_0 + Qx_1$, our procedure assigns the unit point, with coordinates (1, 1), to R.

This tidying-up process enables us to prove the Desargues Theorem very simply.

62.2 The Desargues theorem

Theorem *If two triangles ABC, $A'B'C'$ in π are such that the lines AA', BB' and CC' pass through a point V, then the three points $BC \cap B'C'$, $CA \cap C'A'$ and $AB \cap A'B'$ all lie on a line.*

Proof. We assume that the seven points V, A, B, C, A', B', C' are distinct (Fig.62.1). In a system of homogeneous coordinates let $^{\varepsilon}$ denote the point V, and let $A = \alpha$, $B = \beta$, $C = \gamma$, $A' = \alpha'$, $B' = \beta'$, and $C' = \gamma'$.

By the Normalization Theorem of §62.1, we may write

$$\varepsilon = \alpha + \alpha' = \beta + \beta' = \gamma + \gamma'$$

simultaneously for the three pairs of points with which V is collinear. Let us write:

$$\rho = \beta - \gamma = \gamma' - \beta', \sigma = \gamma - \alpha = \alpha' - \gamma', \tau = \alpha - \beta = \beta' - \alpha'.$$

Then the point represented by ρ is linearly dependent on B and C, and therefore lies on the line BC, and it is also linearly dependent on B' and C', and therefore also lies on the line $B'C'$. If these lines are distinct, ρ is therefore the symbol for the point of intersection $BC \cap B'C'$. Similarly,

$$\sigma = CA \cap C'A' \text{ and } \tau = AB \cap A'B'$$

But

$$\rho + \sigma + \tau = \beta - \gamma + \gamma - \alpha + \alpha - \alpha = 0.$$

Therefore ρ, σ and τ are linearly dependent, and represent three collinear points. Hence the points

$$BC \cap B'C', CA \cap C'A' \text{ and } AB \cap A'B'$$

are collinear.

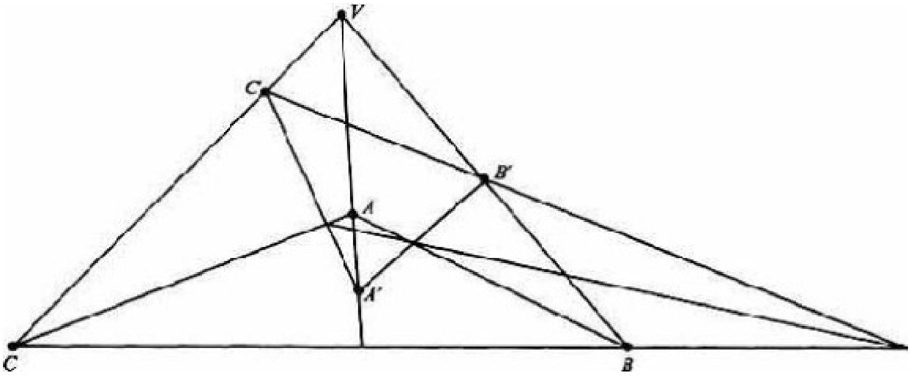

Fig. 62.1

Remark. With a similar definition of homogeneous coordinates to that we have given, and with lines defined to be sets of points linearly dependent on pairs of given distinct points, the projective plane can be defined over a skew field. The Normalization Theorem holds for such a field, and so does the Theorem of Desargues. It will be observed that commutativity under multiplication has not been used in the two proofs above.

When projective geometry is built up from incidence axioms, the Desargues Theorem for the plane is often invoked as an axiom. We can prove the Theorem here because we have assumed a system of homogeneous coordinates for our projective plane. It can be shown that the assumption of the Desargues Theorem leads to the possibility of the construction of a system of homogeneous coordinates over a skew field.

Finally, we note that when two triangles ABC, $A'B'C'$ are such that the lines AA', BB' and CC' pass through a point V, they are said to be *in perspective* from the point V, which is called *the center of perspective*. The line which contains the points $BC \cap B'C'$ $CA \cap C'A'$ and $AB \cap A'B'$ is called *the axis of perspective*. We may also say that the two triangles are *in perspective* from the axis of perspective, and the Desargues Theorem says that two triangles in perspective from a point are also in perspective from a line. The dual, or converse, reverses the order of these statements.

Exercises

62.1 State and prove the dual to the Normalization Theorem (§62.1).

62.2 State and prove the dual of the Desargues Theorem.

62.3 Show that the dual of the Desargues Theorem is implied by the Desargues Theorem, without the intervention of the Principle of Duality. (Hint: we wish to show that three certain lines pass through a point.

This can be shown to be equivalent to three certain points lying on a line. Choosing two triangles which satisfy the Desargues Theorem, this follows.)

62.4 If the sides of a variable triangle pass through three fixed collinear points, while two vertices move along fixed lines, prove that the third vertex must move along a third fixed line concurrent with the other two.

62.5 Three triangles $A_iB_iC_i$ ($i = 1, 2, 3$) are such that any two of them are in perspective, the center of perspective of $A_iB_iC_i$ and $A_jB_jC_j$ being V_{ij}. If V_{12}, V_{23} and V_{31} are collinear, prove that the three triangles have a common axis of perspective l, and that any two of the triangles $A_1A_2A_3$, $B_1B_2B_3$ and $C_1C_2C_3$ are in perspective, the three centers of perspective lying on l.

62.6 A, B, C, D is a quadrangle. P is a point on BD, Q, R are points on AB, AD, and S, T are points on CB, CD. Prove that if P, Q, R are collinear, and also P, S, T, then the lines QS, RT and AC are concurrent.

62.7 D is a fixed point on the side AB of a given triangle ABC, H and K are fixed points which do not lie on any side of the triangle. P is a variable point on CD, and AP intersects DH in Q, and BP intersects DK in R. Prove that QR meets AB in a fixed point. (If P' be another position of P, and Q', R' the corresponding constructed points, we wish to show that QR, AB, and $Q'R'$ are concurrent. Application of the converse of Desargues' Theorem to two suitable triangles produces the result.)

62.8 D is a point on the side BC of triangle ABC, and P, Q are points on AB. The line PD meets AC in H, the line QD meets AC in K, and CP meets AD in M, while CQ meets AD in N. Prove that KM and HN meet on AB. (For a geometrical proof, not using coordinates, see Exercise 63.6.)

63.1 The normalization theorem for points in a projective plane

Theorem I *If A, B, C and P are four distinct points in a projective plane π, and no three are collinear, then a system of homogeneous coordinates can be set up so that the symbols x, y, z* and *t for the respective points satisfy the relation $t = x + y + z$.*

Proof. Since four points in π are linearly dependent, if x', y', z' and t' are the vector symbols for A, B, C and P in our system of homogeneous coordinates, we have a relation $x'\lambda + y'\mu + z'\nu + t'\tau = 0$, where not all of the multipliers λ, μ, ν, τ are zero. In fact, none of them can be zero since if, say, $\lambda = 0$, this would imply the linear dependence, and therefore collinearity of the points B, C and P. Hence we may write

$$t' = x'\lambda' + y'\mu' + z'\nu',$$

where none of the multipliers λ', μ', ν' is zero. We may now *absorb* the multipliers λ', μ' and ν', and write

$$x = x'\lambda', y = y'\mu', z = z'\nu'$$

where x is a definite symbol for A, y for B and z for C, and writing $t = t'$ we have

$$t = x + y + z$$

Remark. If A, B and C are chosen as vertices of the triangle of reference in π, our theorem asserts that any point P *not on the sides* of ABC can be taken to have the coordinates (1, 1, 1). This is called *fixing the unit point* of the coordinate system.

Care has to be taken in applying this theorem, so that we do not assign the same coordinates to two distinct points in the plane. Students tend to do this at first, and are therefore able to prove all true and untrue theorems they set their minds to!

The dual of the above theorem is also important, and we state it without proof.

Theorem II *If a, b, c and p are four distinct lines in a projective plane π, and no three are concurrent, then a system of homogeneous coordinates may be set up so that the symbols l, m, n and q for the respective lines satisfy the relation*

$$q = l + m + n$$

Remark. If a, b and c are chosen as sides of the triangle of reference, the fourth line p, which does not pass through any vertex, can be written as

$$X + Y + Z = 0.$$

Example. ABC is a triangle, and O is a point in the plane of the triangle which does not lie on any side. The line AO meets BC in A', the line BO meets CA in B', and the line CO meets AB in C'. We prove that the points

$$L = B'C' \cap BC, M = C'A' \cap CA, N = A'B' \cap \mathrm{AB}$$

are collinear, and that a coordinate system may be chosen so that the equation of the line LMN is $X + Y + Z = 0$.

Remark. The line LMN is called *the polar line* of the point O with respect to the triangle ABC.

We choose ABC as triangle of reference (Fig.63.1), and O as unit point, as we may do, by Thm I above. Then $A' = (0, 1, 1)$, $B' = (1, 0, 1)$ and $C' = (1, 1, 0)$. To find L, we consider the set of points $(1, 0, 1)\lambda + (1, 1, 0)\mu$, which is the set of points on $B'C'$, and choose λ, μ so that the resulting point lies on BC, which means that its first coordinate is zero. This is evidently done by taking $\lambda = 1$ and $\mu = -1$, so that $L = (1, 0, 1) - (1, 1, 0) = (0, -1, 1)$.

Similarly $M = (1, 0, -1)$ and $N = (-1, 1, 0)$, and these three points satisfy the equation $X + Y + Z = 0$.

Theorem III *The Pappus Theorem. If D, A, E are distinct points on a line l, and C, F, B are distinct points on another line m, coplanar with l, then the three points $DF \cap AC = v$,$DB \cap CE = U$ and $EF \cap AB = W$ are collinear*. (See Fig.

;. *11 $

63.2.)

Remark. The Pappus Theorem appeared in §9.5, when the line of collinearity *UVW* is the line at infinity. Our naming of the points here is not the standard one, since we wish to name the triangle of reference we intend to use to prove the theorem as *ABC*, and the unit point as *U*.

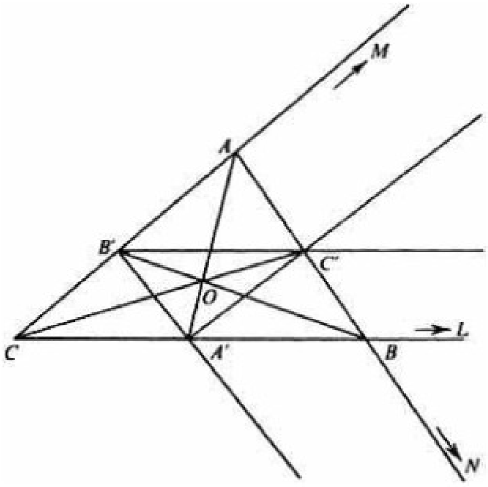

Fig. 63.1

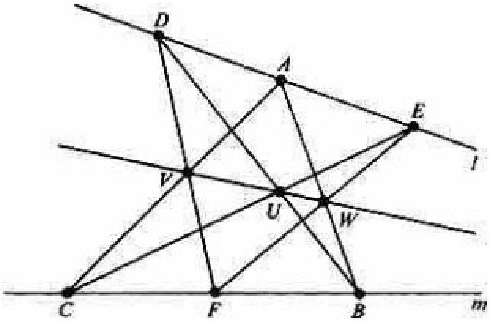

Fig. 63.2

Proof. We choose *ABC* as triangle of reference, and *U* as the unit point, so that $U = A + B + C$. We then take *E*, which is on *UC* as $E = U + Ck$. We now find $D = AE \cap BU$. Since $E = A + B + C(1 + k)$, we see that

$$\begin{aligned} E + AK &= A(1+\mathrm{k}) + B(1+\mathrm{k}) + C(1+\mathrm{k}) - Bk \\ &= U(1+\mathrm{k}) - Bk, \end{aligned}$$

and therefore the point represented, lying on AE and BU, must be D. Hence

$$D = E + Ak = A(1+k) + B + C(1+k)\, Bk$$

The point F is not yet defined, and we take

$$F = B + Ck'$$

V lies on AC and FD, and we find that

$$v = -F + D = A(1+\mathrm{k}) + C(1+\mathrm{k}-\mathrm{k}').$$

Finally, W lies on AB and EF, and to find W we have to eliminate C from

$$E = A + B + C(1+\mathrm{k}). \text{ and } F = B + Ck'$$

and

$$Ek' - F(1+\mathrm{k}) = Ak' + B(k' - 1 - k) = W$$

We note that in obtaining this result we assumed that $k'(1 + k) = (1 + k)k'$, that is, we assumed the commutativity of multiplication for the symbols of our ground field. Since we began by assuming that it is the field of complex numbers, we may do this.

We now have

$$\begin{gathered} U = A + B + C \\ V = A(1+k) + B.0 + C(1+k-k') \\ W = Ak' + B(k' - 1 - k) + C.0, \end{gathered}$$

and evidently

$$U(1+k-k') - V + W = 0.$$

so that the points U, V and W are collinear.

The proof is not as elegant as that of the Desargues Theorem, and algebraically we expect this, since it can be shown that the Pappus Theorem is equivalent to the ground field being commutative, and so at some stage of the proof this assumption enters. No such assumption, as we saw, entered into the proof of the Desargues Theorem. Both theorems will appear again later, when we shall have the means to prove them without using algebra (Thm V, §71.2, Exercise 71.5).

Exercises

63.1 In the proof of the Pappus Theorem (Thm III above) we tacitly assumed, in choosing U as unit point with respect to triangle ABC, that U does not lie on any side of ABC. Show that if U does lie on a side of

ABC the initial assumptions of the theorem are invalidated, or the theorem becomes trivial.

63.2 The line *LMN* is the polar line of the point *O* with respect to the triangle *ABC*, and $A'AO \cap BC, B' = BO \cap CA, C' = CO \cap AB$. The point *P* does not lie on a side of the triangle *ABC*, and *AP*, *BP* and *CP* intersect the polar line in the points *F*, *G* and *H* respectively. Prove that the lines *FA'*, *GB'* and *HC'* are concurrent. (With the notation of the preceding Example, p. 254, take *P* as (p, q, r), and show that the equation of the line *FA'* is $(q - r)(X + Y + Z) + 2rq(Y/q - Z/r) = 0$. If the lines *FA'*, *GB'* and *HC'* are $u = 0$, $v = 0$ and $w = 0$ respectively, we have

$$pu + qv + rw = 0.$$

which proves that the lines are linearly dependent, that is, concurrent.)

63.3 *ABC* is a given triangle, and *D*, *E* and *F* are fixed non-collinear points on *BC*, *CA* and *AB* respectively. A variable line through *A* meets *DE* and *DF* in *P* and *Q* respectively. Prove that the intersection *R* of *BP* and *CQ* lies on a fixed line.

63.4 *L*, *M* and *N* are collinear points on the sides *BC*, *CA* and *AB* of a triangle *ABC*, and *A'*, *B'* and *C'* are chosen on *AL*, *BM* and *CN* respectively. *B'C'*, *C'A'* and *A'B'* meet *BC*, *CA* and *AB* respectively in the points *P*, *Q* and *R*. Prove that *A'P*, *B'Q* and *C'R* are concurrent.

63.5 If *ABC* and *DEF* are two coplanar triangles such that *AD*, *BE*, *CF* are concurrent, and also *AE*, *BF*, *CD* are concurrent, prove that *AF*, *BD* and *CE* are concurrent. (With *ABC* as triangle of reference, and the points *D*, *E*, *F* as $(1, p, p')$, $(q', 1, q)$ and $(r, r', 1)$, the first condition gives $pqr = p'q'r'$. Proceed from this result. The problem can also be solved by applying the Pappus Theorem.)

63.6 *D* is a point on the side *BC* of a triangle *ABC*, and *P*, *Q* are points on *AB*. The line *PD* meets *AC* in *H*, the line *QD* meets *AC* in *K*, and *CP* meets *AD* in *M*, while *CQ* meets *AD* in *N*. Prove that *KM* and *HN* meet on *AB*. (This can be proved by using the Pappus Theorem, noting that the line *AB* is also the line *PQ*. For a proof involving only the Desargues Theorem, see Exercise 69.8.)

64.1 Change of triangle of reference

Since we wish to introduce matrix notation, we now switch back to the suffix notation for coordinates. Let *ABC* and *A'B'C'* be two triangles in the projective plane π, and suppose that a point *P* has coordinates (y_0, y_1, y_2) with respect to *ABC* as triangle of reference, and coordinates (y_0', y_1', y_2') with respect to *A'B'C'* as triangle of reference. This means that we may write

$$P = Ay_0 + By_1 + Cy_2 \text{ and } P = A'y'_0 + B'y'_0 + C'y'_2.$$

Now suppose that *A* has coordinates (a_{00}, a_{10}, a_{20}) with respect to triangle *A'B'C'*, so that

$$A = A'a_{00} + B'a_{10} + C'a_{20},$$

and similarly suppose that

$$B = A'a_{01} + B'a_{11} + C'a_{21},$$
$$C = A'a_{02} + B'a_{12} + C'a_{22}.$$

Then

$$\begin{aligned} P &= (A'a_{00} + B'a_{10} + C'a_{20})y_0 + (A'a_{01} + B'a_{11} + C'a_{21})y_1 \\ &\qquad\qquad\qquad\qquad\qquad\qquad + (A'a_{02} + B'a_{12} + C'a_{22})y_2 \\ &= A'(a_{00}y_0 + a_{01}y_1 + a_{02}y_2) + B'(a_{10}y_0 + a_{11}y_1 + a_{12}y_2) \\ &\qquad\qquad\qquad\qquad\qquad\qquad + C'(a_{20}y_0 + a_{21}y_1 + a_{22}y_2) \\ &= A'y_0' + B'y_1' + C'y_2'. \end{aligned}$$

Since A', B' and C' are linearly independent, any linear relation between them must be trivial, with zero multipliers. Hence we may equate the multipliers of A', B' and C' on both sides of our last equation, and we have

$$\begin{aligned} y_0' &= a_{00}y_0 + a_{01}y_1 + a_{02}y_2, \\ y_1' &= a_{10}y_0 + a_{11}y_1 + a_{12}y_2, \\ y_2' &= a_{20}y_0 + a_{21}y_1 + a_{22}y_2. \end{aligned}$$

so that, in matrix form

$$y' = (y_0' + y_1' + y_2')^{\mathrm{T}} = Ay,$$

where $y = (y_0, y_1, y_2)^T$, and $\mathbf{A} = [a_{ij}]$ $(i, j = 0, 1, 2)$.

Of course, an equivalent equation is

$$y'k = Ay \ (k \neq 0),$$

since we are dealing with homogeneous coordinates.

We can make an immediate statement about the square matrix **A.** Since we may also express y in terms of y', *the matrix* **A** *has an inverse*, and the expression of y in terms of y' must be

$$y = A^{-1}y'$$

We have therefore proved

Theorem I *A change of triangle of reference in the projective plane is equivalent to a non-singular linear transformation of coordinates.*

We consider the effect of a change of triangle of reference on the equation of a line. If the equation of a line is $u^0\ Y_0 + u^1\ Y_1 + u^2\ Y_2 = 0$, we saw that we may talk of *line coordinates* $[u^0, u^1, u^2]$, and designate line coordinates by a single symbol $u = [u^0, u^1, u^2]$. Now suppose that the equation of a line is $v^0\ Y_0' + v^1\ Y_1' + v^2\ Y_2' = 0$ in the primed system of coordinates. Then its original equation in the unprimed system was

$$v^0(a_{00}Y_0 + a_{01}\ Y_1 + a_{02}Y_2) + v^1(a_{10}Y_0 + a_{11}Y_1 + a_{12}Y_2) + v_2(a_{20}Y_0 + a_{21}Y_1 + a_{22}Y_2) = 0,$$

that is

$$(v^0a_{00} + v^2\ a_{10} + v^2\mathrm{a}_{20})Y_0 + (v^0a_{01} + v^1a_{11} + v^2a_{21})Y_1 + (v^0a_{02} + v^1a_{12} + v^2a_{22}Y_2) = 0.$$

Hence, if $y' = \mathbf{A}y$ is the matrix transformation of coordinates,

$$[u^0, u^1, u^2] = [v^0, v^1, v^2]\begin{bmatrix} a_{00} & a_{01} & a_{02} \\ a_{10} & a_{11} & a_{12} \\ a_{20} & a_{21} & a_{22} \end{bmatrix}$$

is the matrix transformation between the original and the new line coordinates. Once more, if $y' = \mathbf{A}y$ gives the new point-coordinates y' in terms of the original point-coordinates y, then $u' = u\mathbf{A}^{-1}$ gives the new line-coordinates u' in terms of the original line-coordinates u.

We note that a point is written as a *column*-vector, and a line as a *row*-vector in these computations. Of course, the calculations may be considerably simplified by using matrix notation. The equation of a line is $uy = 0$, and in the new coordinates it is $vy' = 0$. If we substitute $y' = \mathbf{A}y$, the equation becomes $v\mathbf{A}y = 0$, so that $u = v\mathbf{A}$, or if we call the new line coordinates u', then $u = u'\mathbf{A}$, which is the result obtained above, with the working given in detail.

We therefore have

Theorem II *If we change to a new triangle of reference by the non-singular transformation of point coordinates* $y' = \mathbf{A}y$, *then the corresponding transformation in line-coordinates is given by* $u' = u\mathbf{A}^{-1}$.

Example. Let O be a point not on the sides of a triangle ABC. Let $AO \cap BC = A'BO \cap CA = B'$ and $CO'AB = C'$. We have proved in the Example of §63.1 that we may choose the coordinates so that the equation of the line on which the points $L = B'C' \cap BC$, $M = C'C' \cap CA$and $N = A'B' \cap AB$ lie is given by the equation $X_0 + X_1 + X_2 = 0$. We now show that with respect to $A'B'C'$ as triangle of reference, the equation of this line can also be taken as $X_0' + X_1' + X_2' = 0$, and we shall deduce *that O has the same polar line with respect to the triangles ABC* and $A'B'C'$.

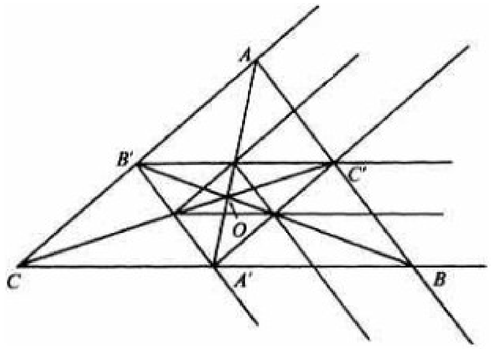

Fig. 64.1

Once again we take O as unit point with respect to triangle ABC (Fig.64.1), so that $A' = (0, 1, 1)$, $B' = (1, 0, 1)$ and $C' = (1, 1, 0)$.

Since A', B' and C' form a proper triangle, we may take $A'B'C'$ as a new triangle of reference, and express the points A, B and C linearly in terms of A', B' and C'. Clearly

$$\begin{aligned}-2A &= (-2,0,0) = (0,1,1) - (1,0,1) - (1,1,0),\\ -2B &= (0,-2,0) = -(0,1,1) + (1,0,1) - (1,1,0),\\ -2C &= (0,0,-2) = -(0,1,1) - (1,0,1) + (1,1,0),\end{aligned}$$

and the above theory tells us that it is the transpose of this matrix, which gives A, B and C in terms of A', B' and C' which is the matrix $\mathbf{A}$ we are looking for. The matrix is symmetric, and it is

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix}.$$

The point $O = (1, 1, 1) = (A' + B' + C')/2$, and so its coordinates are still $(1, 1, 1)$ with respect to the triangle $A'B'C'$. Now let us see what happens to the equation $X_0 + X_1 + X_2 = 0$, which has line coordinates $[1, 1, 1]$. The new line coordinates u' are connected with the original ones by the transformation $u' = [1, 1, 1]\mathbf{A}^{-1}$.

Since so that the new line-coordinates of the polar line may also be taken as $[1, 1, 1]$. But these are the line coordinates of the polar line of O with respect to the triangle $A'B'C'$, since O has coordinates $(1, 1, 1)$ with respect to this triangle. Hence *O has the same polar line with respect to triangles ABC and $A'B'C'$.*

$$\begin{bmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \end{bmatrix}\begin{bmatrix} 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix} = 2\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$u' = [1, 1, 1]\begin{bmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \end{bmatrix} = [-2, -2, -2]$$

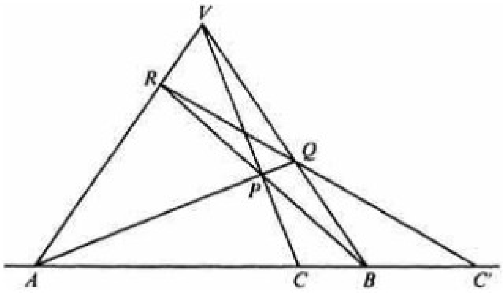

Fig. 64.2

The geometric meaning of this result is left as an exercise for the reader, and the explanation is simple in terms of harmonic ranges and pencils, with which we end this discussion of the projective geometry of the plane, leaving many other important matters to be dealt with later.

64.2 Harmonic ranges

A, B and C are three distinct collinear points, and V is any point not on the line AB. We take any point P on VC distinct from V and from C, and call the intersection of BP and VA the point R and let Q be the intersection of AP and VB. Then if the line QR intersects AB in C', we show that C' depends only on A, B and C, and not on the choice of V and P (Fig.64.2).

In fact, we show that if $C = A + B\lambda$, then $C' = A - B\lambda$.

Since P is on VC, we may write

$$P = C + V\mu = A + B\lambda + V\mu.$$

Therefore

$$P - A = B\mu + V\mu = Q,$$

since the point represented by each term of this equation lies on AP and also on BV. Again

$$P - B\lambda = A + V\mu = \mathrm{R}.$$

Also, C' is on RQ and also on AB, and since

$$R - Q = P - B\lambda - (P - A) = A - B\lambda$$

this must be the point C', so that we have obtained

$$C' = A - B\lambda.$$

We call C' *the harmonic conjugate* of C with respect to A and B. Our algebraic result shows that if we start with C', A and B, and perform the above construction, we arrive at C, since $-(-\lambda) = \lambda$. Hence C and C' are paired, just as A and B are, to start with. But the pairing is in the same sense, since

$$2B\lambda = B = (A+B\lambda) - (A - B\lambda) = C - C'$$

shows that if we begin with C and C' instead of A and B, and B instead of C, the harmonic construction leads to

$$C + C' = (A + B\lambda) + (A - B\lambda) = 2A = A.$$

Hence, *if we say that C and C′ are harmonic conjugates with respect to A and B, we have also proved that A and B are harmonic conjugates with respect to C and C′*. The points A, B and C, C' are said to form *a harmonic range*.

Metrically, we encountered harmonic ranges in §19.1. Points which divide a segment AB internally and externally in the same numerical ratio are harmonic conjugates with respect to A and B. This is indicated by the symbols $A + B\lambda$ and $A - B\lambda$ for C and C'.

There is a dual theory for harmonic pencils, and an important connecting link which asserts that *the range of points given by the intersections of a line with the four rays of a harmonic pencil is a harmonic range*, but we leave the proof of this until §69.1.

Here we remark that the point of intersection, in the construction above, of VC and RQ can be written as $R + Q = 2V\mu + (A + B\lambda)$, and since $R - Q = C'$, this shows that *the harmonic conjugate of* C' with respect to R and Q *is the point* $VC \cap RQ$. Hence, if $BR \cap QC = R'$ and $AQ \cap CR = Q'$, the line $Q'R'$ passes through C'.

We now see why the polar line of O, in the Example of §64.1, with respect to both triangles, ABC and $A'B'C'$, is the same. Another method of proof is indicated in Exercise 64.1.

64.3 Abstract projective planes

As we have already indicated in Exercises 60.7 and 60.8, the concept of *projective plane* can be generalized. Suppose that we have a finite or infinite set S of elements $P_1, P_2, \ldots$ together with a finite or infinite collection of *special subsets* $l_1, l_2, \ldots$ of S, and the concept of *inclusion*. Thus, if P is an element of S which is in a special subset l of S, we shall write $P \in I$. Then an *abstract projective plane* is a set S, which with its special subsets l satisfies the following three properties:

I. If P and Q are any two distinct elements of S, there is a unique special subset l of S such that $P \in I$ and $Q \in I$.

II. If l_1 and l_2 are any two distinct special subsets of S, there is a unique element P of S such that $P \in I_1$ and $P \in I_2$.

III. There are at least four distinct elements of S such that no three of them are elements of a special subset l of S.

In §60.1, these properties were written as Theorems I, II and III, our complex projective plane being a special case of an abstract projective plane. In Exercise 60.7 a finite projective plane with seven points was introduced, the ground field being that of the integers modulo 2, and in Exercise 60.8 it was indicated that once coordinates have been introduced into the real affine plane, further 'ideal' points and lines can be added so that we obtain a projective plane.

The subject of finite projective planes is a most important one. It will not be discussed here. Further discussion can be found in Pedoe (**12**), Chapter V. But we give some Exercises which indicate some consequences of the three axioms which are to be satisfied by the elements (points) and special subsets (lines) of an abstract projective plane.

Exercises

64.1 Prove the result of the Example in §64.1 by considering the triangles ABC and $A'B'C'$ as perspective triangles, and then the triangles $A'B'C'$ and $A''B''C''$ as perspective triangles, where $A'' = B'C' \cap AA'$,$B'' = C'A' \cap BB'$and $C'' = A'B' \cap CC'$.

64.2 Show that if we have a harmonic range A, B, $A + B\lambda$, $A - B\lambda$ referred to one system of coordinates, and with respect to another triangle of reference A becomes A' and B becomes B', then $A + B\lambda$ becomes $A' + B'\lambda$ and $A - B\lambda$ becomes $A' - B'\lambda$.

64.3 If A, B, C, D are four points in a plane, no three of which are collinear, prove that their coordinates a, b, c, d may be so chosen that $a + b + c + d = 0$. If $P = AB \cap CD$,$Q = AC \cap BD$ and $R = AD \cap BC$, so that P, Q and R are the diagonal points of the quadrangle formed by A, B, C and D, show that the coordinates p, q and r of P, Q and R satisfy the equations

$$\begin{aligned} p &= a + b = -c - d, \\ q &= a + c = -b - d, \\ r &= a + d = -b - c. \end{aligned}$$

64.4 Deduce from the preceding Exercise that any two diagonal points of a complete quadrangle are harmonic conjugates with respect to the points in which their join meets the two sides of the quadrangle which do not contain the two diagonal points.

64.5 State and prove the dual theorems corresponding to the two preceding Exercises.

64.6 OA, OB, OC are given lines. Variable lines MPQ, $MP'Q'$ meet OA in M, OB in P, P' and OC in Q, Q'. Finally PQ' meets $P'Q$ in R. Prove that R lies on a fixed line.

64.7 Let the diagonal triangle of a complete quadrangle be LMN, and let P be a given point in the plane. Prove that the harmonic conjugates of LP, MP and NP relative to the pairs of opposite sides of the quadrangle which meet in L, M and N respectively are concurrent.

64.8 ABC is a triangle, and P, Q are given points. A', B' and C' are the harmonic conjugates relative to P, Q of the points in which PQ meets BC, CA and AB respectively. Prove that AA', BB' and CC' are concurrent.

64.9 Consider the non-singular linear mapping given by the equations

$$(X_0', X_1', X_2') = (3X_0 - X_1 + 2X_2, -2X_0 - 2X_1 - 2X_2, 2X_0 - X_1 + 3X_2),$$

and let the transforms of P and Q be P' and Q'. Prove that PP' passes through the point $(1, -1, 1)$, that the line PP' is mapped onto itself, and that the lines PQ and $P'Q'$ intersect on the line $2X_0 - X_1 + 2X_2 = 0$.

64.10 Prove that in an abstract projective plane at least three points lie on every line, and at least three lines pass through every point.

64.11 A tennis match is played between two teams, each player playing one or more members of the other team. We are told that:

(i) any two members of the same team have exactly one opponent in common,

(ii) no two members of the same team play all the members of the other team between them. Prove that two players who do not play each other have the same number of opponents. Deduce that any two players, whether in the same or different teams, have the same number of opponents.

64.12 Let $n \geq 2$ be any integer, and suppose that we are considering an abstract projective plane with a finite number of points. Show that the following properties are equivalent:

(1) One line contains exactly $n + 1$ points.

(2) One point is on exactly $n + 1$ lines.

(3) Every line contains exactly $n + 1$ points.

(4) Every point is on exactly $n + 1$ lines.

(5) There are exactly $n^2 + n + 1$ points.

(6) There are exactly $n^2 + n + 1$ lines.

65.1 Projective space of three dimensions

We now introduce homogeneous coordinates into what we call a *three-dimensional projective space*. We consider a four-dimensional vector space V_4 over the complex field, and define *a point of three-dimensional projective space*, which we shall call S_3, to be *a ray*, which arises from *a vector* of V_4. That is, a point of S_3 is to be the equivalence class (y_0k, y_1k, y_2k, y_3k) $(k \neq 0)$, where $y = (y_0, y_1, y_2, y_3)$ is *a non-zero vector* of V_4.

Once again, linear dependence of points in S_3 is defined in the obvious way, and the points in S_3 linearly dependent on two distinct points of S_3 determine a line. If y and z are any representatives of the two points, the set of points $y\lambda + z\mu$ determine a line which contains the two points. But now that we obtain S_3 from a V_4, we may consider the set of points linearly dependent on three points, not already linearly dependent, which means that they are not collinear. If the points be denoted by the symbols y, z and t, the set is $y\lambda + z\mu + t\nu$, and we say this set of points determines *a plane* in S_3, the plane (yzt).

In V_4 there are sets of four independent vectors, but any five vectors are dependent. Hence in S_4 there are sets of four independent points, but any set of five points is a dependent set. Suppose that we now let A, B, C, D and E be the symbols for five points in S_3. Then we have a relation

$$Aa + Bb + Cc + Dd + Ee = 0,$$

where not all of the multipliers are zero. If we now write

$$Aa + Bb + Cc= -Dd - Ee,$$

the left-hand side represents a point in the plane ABC, and the right-hand side a point on the line DE. Hence we have one of the Propositions of Incidence in S_3:

A plane either contains a given line, or intersects it in one point.

Suppose now that we have two distinct planes, the one defined by the points A, B, C and the other by the points D, E, and F. Then the line DE either lies in the plane ABC or intersects it in a point, and the line EF either lies in the plane ABC or intersects it in a point. Since the planes are distinct, we do not have both lines, DE and EF in the plane ABC. If one of these lines lies in the plane ABC, the two planes ABC and DEF already intersect in a line. If neither lies in the plane ABC, we have two points in the plane DEF which lie in the plane ABC, and therefore the line joining the two points lies in ABC. Hence

Two planes which are distinct always intersect in the points of a line.

We note that there are no exceptions to the Propositions of Incidence in S_3, since planes are never parallel.

We have not written all the Propositions of Incidence in a form which would indicate that the Principle of Duality (§60.2) also applies to S_3. It does, of course, but in S_3 the fundamental interchange is between point and plane. Algebraically this is seen as soon as we show that a plane can also be given by homogeneous coordinates. If

$$(X_0,X_1,X_2,X_2,X_3) = (a_0,a_1,a_2,a_3)\lambda + (b_0,b_1,b_2,b_3)\mu + (c_0,c_1,c_2,c_3,)\nu,$$

the elimination of the parameters λ, μ, and ν leads to the determinantal equation

$$\begin{vmatrix} X_0 & X_1 & X_2 & X_2 \\ a_0 & a_1 & a_2 & a_3 \\ b_0 & b_1 & b_2 & b_3 \\ c_0 & c_1 & c_2 & c_3 \end{vmatrix} = 0,$$

which is of the form

$$u^0X_0 + u^1X_1 + u^2X_2 + u^3X_3 = 0,$$

a homogeneous linear equation. Hence a plane can be assigned homogeneous coordinates $[u^0, u^1, u^2, u^3]$, and belongs to an equivalence class $k[u^0, u^1, u^2, u^3]$ $(k \neq 0)$. Two distinct planes intersect in a unique line, and the equation of a line is often given by the assignment of two distinct planes which contain it. We may use matrix notation, and write the equation of a plane as

$$uX = 0,$$

where $u = [u^0, u^1, u^2, u^3]$ and $X = (X_0, X_1, X_2, X_3)^T$, so that once again we write *plane-coordinates* as *row-vectors* and *point-coordinates* as *column-vectors*. Then the equation of a line may be written

$$uX = 0 = vX$$

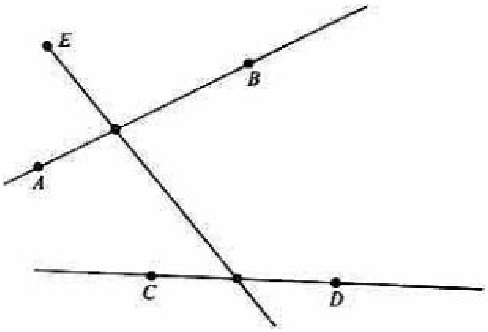

Fig. 65.1

where u and v are two distinct planes through the line.

The condition that a given plane u should contain a given point x is

$$uX = 0,$$

and this is a symmetrical relation between the plane-coordinates u and the point-coordinates x. Hence we have a Principle of Duality for S_3, in which *point* and *plane* and *plane* and *point* are interchanged. Since a line is determined by two points, and also by two planes, *lines remain unchanged* in the transliteration which produces a dual theorem from a given theorem.

We may prove easily, or regard the Proposition as an application of the Principle of Duality, that since three distinct points determine a unique plane, provided the points are linearly independent, so *three distinct planes determine a unique point*, provided they are linearly independent, *which lies in each of them*.

We consider some applications of the Propositions of Incidence.

(*a*) From a point which does not lie on either of two skew (non-intersecting) lines in S_3 one and only one line can be drawn to intersect each line. (Such a line is called a *transversal* of the two lines).

Take points A, B on one line, and C, D on the other (Fig.65.1), and let E be the symbol for the given point. Since five points are always linearly dependent, we have a relation

$$Aa + Bb + Cc + Dd + Ee = 0$$

between the symbols for the five points, where not all of a, b, c, d and e are zero. Writing this as

$$-Ee = (Aa + Bb) + (Cc + Dd),$$

we see that the point E is collinear with the point $Aa + Bb$ on the one line and the point $Cc + Dd$ on the other.

If there were two distinct transversals to the two lines, through E, the two lines would lie in the plane determined by the two transversals, which intersect at E, and therefore the given lines, being coplanar, would intersect, whereas we are told that they are skew.

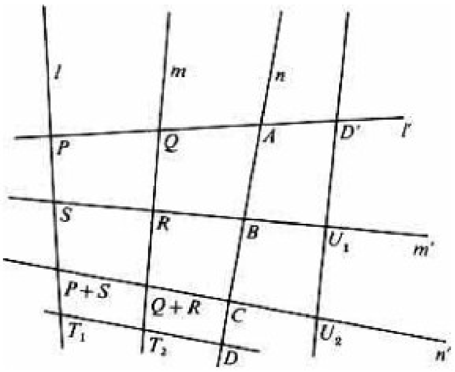

Fig. 65.2

Another way of obtaining this result is suggested in Exercise 65.2.

(*b*) We investigate the set of lines which are transversal to *three* given mutually skew lines in S_3 (Fig.65.2). Let l, m and n be mutually skew lines, and draw the transversals l', m' and n' from three distinct points A, B and C lying on n to l and m. From the point D on n, distinct from A, B and C, we draw a fourth transversal to l and m, and from D', any point on l' which is not on l, m or n we draw the unique transversal to m' and n'. We prove that, subject to certain conditions, the transversal from D to l and m intersects the transversal from D' to m' and n'.

We take the four points shown in the diagram as base-points, P, Q, R and S. They are independent, since dependence would involve l and m being coplanar, and we have assumed that they are skew. By the Normalization Theorem, §62.1, we may choose $A = P + Q$, where $A = n \cap l'$, and $B = R + S$, where $B = n \cap m'$. With A and B determined except for a multiplying constant, we may also choose $C = A + B$, where $C = n \cap n'$. (This possibility should be examined carefully.) We then take $D = A + B\lambda$, for some λ, and $D' = P + Q\mu$, for some μ, where D is the point chosen on n distinct from A, B and C, and D' is the point chosen on l' distinct from A, P and Q.

Since $C = A + B = P + Q + R + S = (P + S) + (Q + R)$, the transversal from C to l and m intersects these lines at the points $P + S$ and $Q + R$ respectively. Similarly $D = A + B\lambda = P + Q + (R + S)\lambda = (P + S\lambda) + (Q + R\lambda)$, and therefore the transversal from D to l and m intersects these lines in the points $T_1 = P + S\lambda$ and $T_2 = Q + R\lambda$.

Finally, we note that $D' = P + Q\mu = -(S + R\mu) + (P + S) + (Q + R)\mu$, and therefore the transversal from D' to m' and n' intersects these lines in the points

$$U_1 = -(S + R\mu) \text{ and } U_2 = (P + S) + (Q + R)\mu.$$

We have now obtained the coordinates of the points we are considering. We wish to show that there is a point on the line U_1U_2 which is the same point as one on the line T_1T_2: that is, we require that for some values of v, v' and k,

$$\begin{aligned}(T_1 + T_2\nu)k = [(P + S\lambda) + (Q + R\lambda)\nu]k &= U_2 - U_1\nu' \\ &= (P + S) + (Q + R)\mu + (S + R\mu)\nu'.\end{aligned}$$

Since P, Q, R and S are independent, the multipliers of P, Q, R and S on each side of this equation are proportional, and therefore

$$1 : v : \lambda\ v\lambda = 1 : \mu : \mu + \mu v' : 1 + v'$$

are equations to be solved, if possible, for v and v'. Taking $v = \mu$, then we want both

$$\lambda\mu = \mu + \mu v'$$

and

$$\lambda = 1 + v'$$

If we assume commutativity for our multipliers, we have

$$\lambda\mu = \mu\lambda = \mu + \mu v'\ \mu[\lambda - (1 + v')] = 0,$$

and since $\mu \neq 0$, we may take

$$v' = \lambda - 1,$$

and the final equation is also satisfied.

In this case the point of intersection of the transversals through D and D' is

$$\begin{aligned}T_1 + T_2\nu &= (P + S\lambda) + (Q + R\lambda)\nu \\ &= P + Q\nu + R\nu\lambda + S\lambda \\ &= P + Q\mu + R\mu\lambda + S\lambda,\end{aligned}$$

so that the coordinates of the point may be taken as

$$(y_0, y_1, y_2, y_3) = (1, \mu, \mu\lambda, \lambda)$$

and this always lies on the surface

$$X_1X_3 - X_0X_2 = 0.$$

Such a surface is called *a quadric surface*, and it is one of the fundamental constructs in S_3. We shall investigate quadric surfaces in more detail in §81.1.

We remark that the surface contains two systems of *generators*, or lines which lie in the surface, one system containing the lines l, m and n, and the other containing the lines l', m' and n'. The lines in a system are mutually skew, but any line in one system intersects any line in the other system. The reader is asked to prove this result in Exercise 65.5.

We have already mentioned that the assumption of commutativity of our coordinates is equivalent to the assumption of the Pappus Theorem, and in some texts the intersection of the lines $U_1\ U_2$ and T_1T_2 is derived directly from that theorem [Baker, (2), Vol. 1]. Another proof is suggested in Exercises 71.16, 17, 18, using the Fundamental Theorem of Projective Geometry, which is also equivalent to the Pappus Theorem. This theorem is proved in §71.1.

65.2 Moebius tetrahedra

Theorem *If two tetrahedra $ABCD$, $A'B'C'D'$ are such that $A' \in (BCD)$,$B' \in (CDA)$,$C' \in (DAB)$,$D' \in (ABC)$ so that the first tetrahedron is* escribed *to the second), and also $A \in (B'C'D')$,$B \in (C'D'A')$ and $C \in (D'A'B')$, then necessarily $D \in (A'B'C')$.*

Such mutually inscribed and escribed tetrahedra are called *Moebius* tetrahedra, after their discoverer. They are a surprising phenomenon, since it is not possible in S_2 to have two triangles which are mutually inscribed and escribed (Exercise 65.7). We leave the proof of the theorem as stated for an Exercise (65.6), and demonstrate the existence of Moebius tetrads by remarking that if we take $ABCD$ as tetrad of reference, and

$$\begin{gathered}A' = (0, q_1, r_1, s_1)\\ B' = (-q_1, 0, r_2, s_2)\\ C' = (-r_1, -r_2, 0, s_3)\\ D' = (-s_1, -s_2, s_3, 0)\end{gathered}$$

then we have a tetrad which is inscribed in $ABCD$. To show that it is also escribed, we have to show that the plane $A'B'C'$ contains the point $D = (0, 0, 0, 1)$, the plane $B'C'D'$ contains the point $A = (1, 0, 0, 0)$, the plane $C'D'A'$ contains the point $B' = (0, 1, 0, 0)$, and the plane $D'A'B'$ the point $C' = (0, 0, 1, 0)$. If we prove that the plane $A'B'C'$ contains D, this will indicate the method. Since $(0, q_1, r_1, s_1)r_2 - (-q_1, 0, r_2, s_2)r_1 + (-r_1, -r_2, 0, s_3)q_1 = (0, 0, 0, k)$, where $k = s_1r_2 - s_2r_1 + s_3q_1$, this is so. We also remark that the matrix of entries for A', B', C', D' is skew-symmetric. Another proof of our last result arises from the theorem that the determinant of a skew-symmetric matrix of odd order is zero. We shall see in §85.4 that Moebius tetrads and skew-symmetric matrices are linked.

This section on S_3 has been kept short since we shall obtain results in S_n, projective space of n dimensions, which naturally apply to the case $n = 3$, and the methods are not essentially different. The Exercises which follow indicate what a fuller treatment of the case $n = 3$ would imply. We must stress a point which is often not clear to students when the properties of S_3 are first studied: a line in S_3 is not different from a line in S_2, and a plane in S_3 is a projective plane, so that properties in S_2 continue to hold in any plane in S_3. For example, if we have any three distinct collinear points in S_3, which we denote by A, B and C, and if $C = A + B\lambda$, then the harmonic conjugate of C with respect to A and B can be constructed, as in §64.2, in any plane which contains the line ABC, and we find that the harmonic conjugate $C' = A - B\lambda$, and therefore it is independent of the particular plane chosen through the line.

Exercises

65.1 Show that in S_3 a plane is determined by the assignment of a point and a line which does not contain the point, if the plane is to contain both the point and the line. Show also that two intersecting lines determine a plane, which is to contain them.

65.2 Give another proof that from a point which does not lie on either of two skew lines in S_3 there is a unique transversal to the two lines, by considering the intersection of the one line with the plane defined by the point and the other line.

65.3 State and prove a Normalization Theorem for five points A, B, C, D and E in S_3, where E does not lie in any face of the tetrahedron formed by A, B, C and D.

65.4 Show that in Fig. 65.2 the point C does not lie in any face of the tetrahedron formed by P, Q, R and S, so that, by the Normalization Theorem, we could have taken $C = P + Q + R + S$ to start with, and then deduced that $A = P + Q$ and $B = R + S$, since there is a unique transversal through C to the lines l' and m'.

65.5 Show that the quadric surface $X_1X_3 - X_0X_2 = 0$ contains the line of intersection of the two planes $\lambda X_1 - X_2 = 0 = \lambda X_0 - X_3$ for any value of λ, and the line of intersection of the two planes $\mu X_0 - X_1 = 0 = X_2 - \mu X_3$ for any value of μ. If we call the first set of lines the λ-system, and the second set the μ-system, show that the λ-system consists of mutually skew lines, as does the μ-system, but that every line of the λ-system intersects every line of the μ-system. Finally, show that the point of intersection of a line of the λ-system and a line of the μ-system is the point $(1, \mu, \mu\lambda, \lambda)$ of the surface. Do any more lines in the surface pass through this point?

65.6 Prove the theorem on Moebius tetrads of §65.2. (Begin with A' as given, and show that if we take $B' = (-q_1, 0, r_2, s_2)$, $C' = (-r_1, q_3, 0, s_3)$, and $D' = (-s_1, q_4, r_4, 0)$, as we may, then we obtain $q_3 = -r_2$, $r_4 = -s_3$ and $q_4 = -s_2$, so that $(A'B'C')$ contains D.)

65.7 Show that it is impossible to find two proper triangles in S_2 which are mutually inscribed and escribed.

65.8 AB, CD are given skew lines in S_3. The transversal of AB, CD through a variable point P meets AB, CD in R, S. The point P' is the harmonic conjugate of P with respect to R, S. Find the locus of P' if P moves on (i) a given plane, (ii) a given line.

65.9 If ABC, $A'B'C'$ are two triangles in S_3, which are not coplanar, and the lines AA', BB' and CC' are concurrent at a point V (so that the triangles are perspective from a point), show that the points $BC \cap B'C'$,$CA \cap C'A'$ and $AB \cap A'A'B'$ exist and are collinear, using only the Propositions of Incidence.

65.10 Prove the converse of the Desargues Theorem in S_3, that if two non-coplanar triangles are perspective from a line, then they are also perspective from a point.

65.11 Given five mutually skew lines p, a, b, c, d in S_3, a transversal is drawn to a and c from a point P on p, intersecting a in A and c in C, and another transversal is drawn to b and d, intersecting b in B and d in D. Prove that the intersection Q=$AB \cap CD$ of AB and CD describes a line as P describes p.

65.12 Show that a change of tetrad of reference in S_3 is equivalent to a non-singular transformation $y' = \mathbf{A}y$ between coordinates. Show that plane-coordinates change in the manner indicated by the matrix equation $u = u'\mathbf{A}$, where $y' = \mathbf{A}y$ is the relation between point-coordinates.

65.13 Restate the Normalization Theorem (Exercise 65.3) as a theorem on non-singular linear transformations in S_3.

65.14 Let A_i (i = 1, 2, 3, 4) be a tetrad of points in S_3 which are not coplanar. If P be any given point, and PA_1 meet the plane $A_2A_3A_4$ in the point P_1, let P_1' be the harmonic conjugate of P with respect to the points A_1 and P_1. Let us write $PT_1 = P_1'$, where the transformation T_1 is called *harmonic homology* (see §39.1). Show that T_1 is represented by a non-singular matrix. Prove further that if T_2, T_3 and T_4 be defined in a similar way,

$$(P)T_1T_2T_3T_4 = P.$$

(Compare Exercise 25.3, and the link to this Exercise provided by §39.1.)

65.15 If A, B, C and D be four non-coplanar points in S_3, and

$$P = Ax_0 + Bx_1 - Cx_2 - Dx_3$$

show that the harmonic conjugate of P with respect to the pair of points in which the transversal from P to AB and CD intersects them is

$$P' = Ax_0 + Bx_1 - Cx_2 - Dx_3$$

and that the relation between P and P' is represented by a non-singular matrix. What can we say about the square of this matrix?

65.16 $ABCD$ is a tetrahedron, A' an arbitrary point not lying in a face. The transversal is drawn from A' to AB, CD, and B' is the harmonic conjugate of A' with respect to the intersections of this transversal with AB and CD. Points C' and D' are defined similarly, with regard to AC and BD and AD and BC. Prove that the two tetrads $ABCD$, $A'B'C'D'$ are such that each edge of either meets a pair of opposite edges of the other, and is divided harmonically by them.

Prove also that AA', BB', CC' and DD' meet in a point A'', that AB', BA', CD' and DC' meet in a point B'', that AC', BD', CA', DB' meet in a point C'', and that AD', BC', CB', DA' meet in a point D''. Finally, show that the three tetrads $ABCD$, $A'B'C'D'$ and $A''B''C''D''$ form a completely symmetrical system, in the sense that the edge of any one of the tetrads meets a pair of opposite edges of each of the others and is divided harmonically by them. (Such a set of tetrahedra is called a *desmic* system. Begin with $A' = A + B + C + D$, and then $B' = (A + B) - (C + D)$, *etc.* The solution is shorter than the enunciation of the problem! For a real, and simple model of a set of desmic tetrads, see Exercise 88.6.)

65.17 π and π' are distinct planes in S_3, and V is a point not in either plane, W a distinct point not in either plane. The points P in π' are projected into points Q in π from V, and into points Q' in π from W,($VP \cap$

$\pi = Q, WP \cap \pi = Q'$) Show that the mapping $Q \rightarrow Q'$ in π is a collineation, that the joins of corresponding points, QQ', pass through a fixed point, and that the intersection of corresponding lines, $m \cap m'$, where m' is the map of m, always lies on a fixed line. After doing this geometrically (without the use of coordinates) set up the equations for the collineation $Q \rightarrow Q'$ in π, and verify your results.

!"#$# % & & !' (! !) * * + , * - . *
/ * ! + , F * 3 # F * 3 F 3 @ 3 * * 3 9,1 7 *
+ , 5 "/" & 6 13

11

The Axiom on Parallels

Euclidean geometry is distinguished from neutral geometry by one axiom, one that describes the nature of parallel lines. There are many formulations of the parallel axiom for Euclidean geometry, but the one that gets right to the heart of the matter is *Playfair's axiom*, named after the Scottish mathematician John Playfair.

> *(Playfair's axiom) Let ℓ be a line, and let P be a point that is not on ℓ. Then there is exactly one line through P that is parallel to ℓ.*

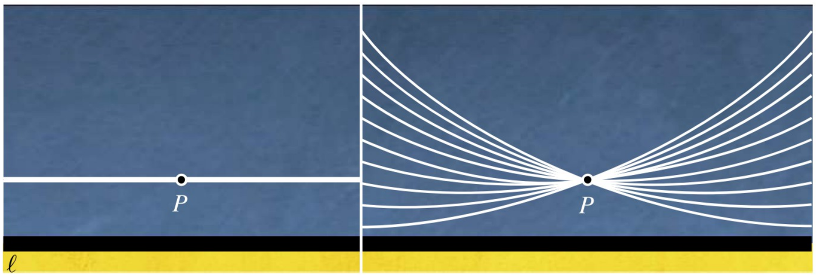

Figure 1. On the left, Euclidean geometry: a single line through P parallel to ℓ. On the right, more than one line through P parallel to ℓ.

In this chapter, we look at a small collection of theorems that are almost immediate consequences of Playfair's axiom, and as such are at the very core of Euclidean geometry. The first of these is Euclid's fifth postulate. This is the controversial postulate in the *Elements*, but also the one that guarantees the same parallel behavior that Playfair's axiom provides. Euclid's fifth is a little unwieldy, particularly when compared to Playfair's axiom, but it is the historical impetus for much of what followed. So let's use Playfair's axiom to prove Euclid's fifth postulate.

Theorem 11.1 *(Euclid's fifth postulate) If lines ℓ_1 and ℓ_2 are crossed by a transversal t, and the sum of adjacent interior angles on one side of t measure less than 180°, then ℓ_1 and ℓ_2 intersect on that side of t.*

Proof First some labels.

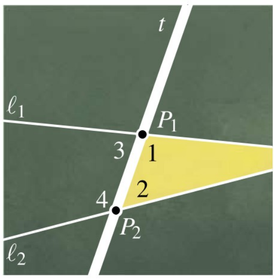

Figure 2. The labels.

Start with lines ℓ_1 and ℓ_2 crossed by transversal t. Label P_1 and P_2, the points of intersection of t with ℓ_1 and ℓ_2 respectively. On one side of t, the two adjacent interior angles should add up to less than 180°. Label the one at P_1 as $\angle 1$ and the one at P_2 at $\angle 2$. Label the supplement of $\angle 1$ as $\angle 3$ and label the supplement of $\angle 2$ as $\angle 4$.

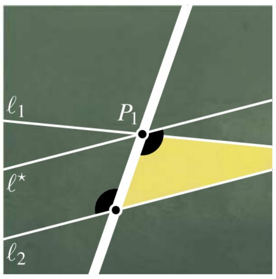

Figure 3. Constructing the unique parallel.

This theorem asserts that the lines intersect, and says where the intersection occurs. We will tackle those two statements in that order. Note that $\angle 1$ and $\angle 4$ are alternate interior angles, but they are not congruent: if they were, their supplements $\angle 2$ and $\angle 3$ would be too, and then

$$(\angle 1) + (\angle 2) = (\angle 1) + (\angle 3) = 180^\circ.$$

There is, however, another line $\ell^\star$ through P_1 that forms an angle congruent to $\angle 4$ (because of the angle construction axiom), and by the alternate interior angle theorem, $\ell^\star$ must be parallel to ℓ_2. Because of Playfair's axiom, $\ell^\star$ is the only parallel to ℓ_2 through P_1. Therefore ℓ_1 intersects ℓ_2.

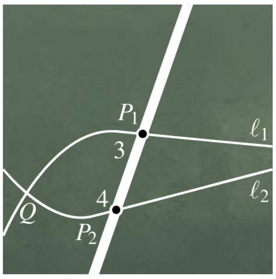

Figure 4. An impossible triangle on the wrong side of t.

The second part of the proof is to figure out on which side of t that ℓ_1 and ℓ_2 cross. Let's see what would happen if they intersected at a point Q on the wrong side of t, the side with $\angle 3$ and $\angle 4$. Then $\triangle P_1P_2Q$ would have two interior angles, $\angle 3$ and $\angle 4$, whose measures add up to more than 180°. This would violate the Saccheri-Legendre theorem. So ℓ_1 and ℓ_2 cannot intersect on the side of t with $\angle 3$ and $\angle 4$ and so they must intersect on the side with $\angle 1$ and $\angle 2$.

□

One of the truly useful theorems of neutral geometry is the alternate interior angle theorem. In fact, we just used it in the last proof. You may recall from high school geometry that the converse of the theorem is often even more useful. The problem is that the converse of the alternate interior angle theorem can't be proved using just the axioms of neutral geometry. It depends on the Euclidean behavior of parallel lines.

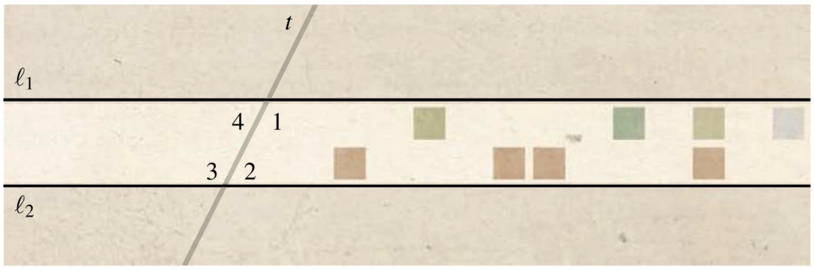

Figure 5. The converse of the alternate interior angle theorem: if $\ell_1 \parallel \ell_2$, then $\angle 1 \simeq \angle 3$ and $\angle 2 \simeq \angle 4$.

Theorem 11.2 *(Converse of the alternate interior angle theorem) If ℓ_1 and ℓ_2 are parallel, then the pairs of alternate interior angles formed by a transversal t are congruent.*

Proof Consider two parallel lines crossed by a transversal. Label adjacent interior angles: $\angle 1$ and $\angle 2$, and $\angle 3$ and $\angle 4$, so that $\angle 1$ and $\angle 4$ are supplementary and $\angle 2$ and $\angle 3$ are supplementary. That means that the pairs of alternate interior angles are $\angle 1$ and $\angle 3$, and $\angle 2$ and $\angle 4$. Now, we just have to do a little arithmetic. From the two pairs of supplementary angles:

(i): $(\angle 1) + (\angle 4) = 180^\circ$
(ii): $(\angle 2) + (\angle 3) = 180^\circ$.

Adding all four angles together,

$$(\angle 1) + (\angle 2) + (\angle 3) + (\angle 4) = 360^\circ.$$

Now we turn to Euclid's fifth postulate, or rather its contrapositive. The lines ℓ_1 and ℓ_2 are parallel, which means that they do not intersect on either side of t. Therefore Euclid's fifth says that on neither side of t may the sum of adjacent interior angles be less than 180°:

- $(\angle 1) + (\angle 2) \geq 180^\circ$
- $(\angle 3) + (\angle 4) \geq 180^\circ$.

If either sum was greater than 180°, though, the sum of all four angles would have to be more than 360°. We saw above that is not the case, so the inequalities are actually equalities:

(iii): $(\angle 1) + (\angle 2) = 180^\circ$
(iv): $(\angle 3) + (\angle 4) = 180^\circ$.

Now we have two systems of equations with four unknowns, and it is basic algebra from here. Subtract *(iv)* from *(i)* to get

$$(\angle 1) - (\angle 3) = 0 \implies (\angle 1) = (\angle 3).$$

Subtract *(i)* from *(iii)* to get

$$(\angle 2) - (\angle 4) = 0 \implies (\angle 2) = (\angle 4).$$

Therefore the alternate interior angles are congruent. □

One of the key theorems we proved in the neutral geometry section was the Saccheri-Legendre theorem: that the angle sum of a triangle is at most 180°. That is all you can say

with the axioms of neutral geometry, but in a world with Playfair's axiom and the converse of the alternate interior angle theorem, we can make a more bold statement about angle sums of triangles.

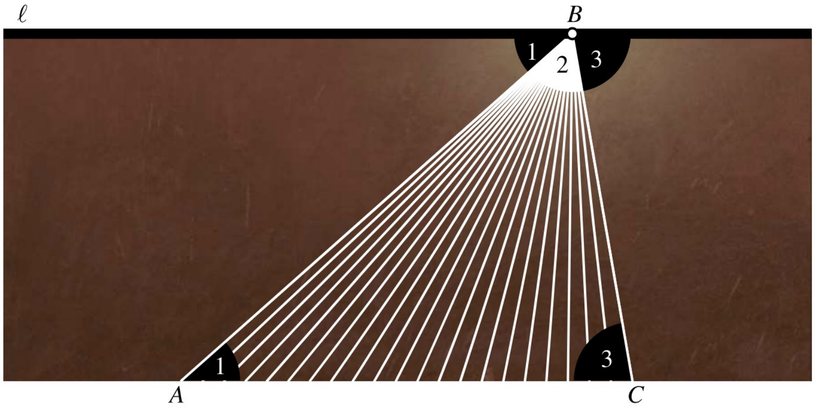

Figure 6. Use the converse of the alternate interior angle theorem to find the angle sum of a triangle.

Theorem 11.3 *The angle sum of a triangle is* 180°.

Proof Given $\triangle ABC$, by Playfair's axiom, there is a unique line ℓ through B that is parallel to $\overline{AC}$. That line and the rays $\overrightarrow{BA}$ and $\overrightarrow{BC}$ form three angles, $\angle 1$, $\angle 2$, and $\angle 3$ as shown in figure 6. By the converse of the alternate interior angle theorem, two pairs of alternate interior angles are congruent: $\angle 1 \simeq \angle A$ and $\angle 3 \simeq \angle C$. Therefore, the angle sum of $\triangle ABC$ is

$$s(\triangle ABC) = (\angle A) + (\angle B) + (\angle C) = (\angle 1) + (\angle 2) + (\angle 3) = 180^\circ. \qquad \square$$

So, the three angles in an equilateral triangle all measure 60° and the measures of the two acute angles in a right triangle must add up to 90°.

In the last chapter on quadrilaterals we talked a little bit about the uncertain status of rectangles in neutral geometry, saying that it is easy to make a convex quadrilateral with three right angles, but that once you have done that, there is no guarantee that the fourth angle will be a right angle. Here it is now in the Euclidean context:

Theorem 11.4 *Let* $\angle ABC$ *be a right angle. Let* r_A *and* r_C *be rays so that* r_A *has endpoint* A*, is on the same side of* $\overline{AB}$ *as* C*, and is perpendicular to* $\overline{AB}$*;* r_C *has endpoint* C*, is on the same side of* $\overline{BC}$ *as* A*, and is perpendicular to* $\overline{BC}$*. Then* r_A *and* r_C *intersect at a point* D*, and the angle formed at this intersection,* $\angle ADC$*, is a right angle. Therefore* $\square ABCD$ *is a rectangle.*

Proof First we need to make sure that r_A and r_C intersect. Let ℓ_A and ℓ_C be the lines containing r_A and r_C respectively. By the alternate interior angle theorem, the right angles at A and B imply that ℓ_A and $\overline{BC}$ are parallel. So $\overline{BC}$ is the one line parallel to ℓ_A through C, and that implies that ℓ_C cannot be parallel to ℓ_A: it has to intersect ℓ_A. Let's call that point of intersection D. This

theorem claims that it is the rays, not the lines, that intersect. Therefore we need to rule out the possibility that the intersection of ℓ_A and ℓ_C might happen on one (or both) of the opposite rays. Since ℓ_A is parallel to $\overline{BC}$, all the points of ℓ_A are on the same side of $\overline{BC}$ as A. None of the points of r_C^{op} is on that side of BC, so D cannot be on r_C^{op}. Likewise, all the points of ℓ_C are on the same side of $\overline{AB}$ as C. None of the points of r_A^{op} is on that side of AB, so D cannot be on r_A^{op}.

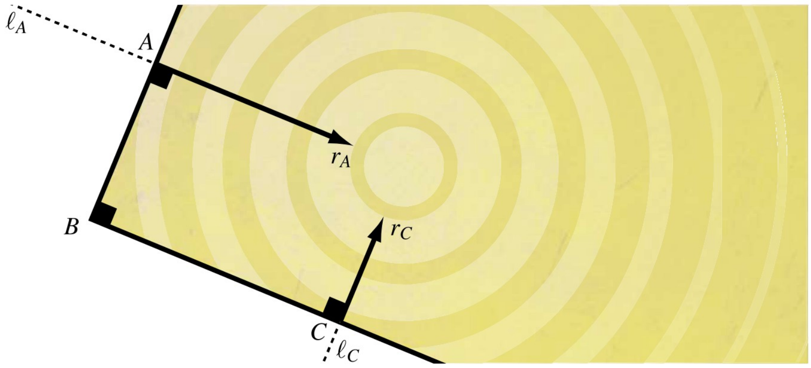

Figure 7. Constructing a rectangle.

So now $\square ABCD$ has three right angles, $\angle A$, $\angle B$, and $\angle C$. It is actually a convex quadrilateral too (verifying this is left as an exercise), so the diagonal AC divides $\square ABCD$ into two triangles $\triangle ABC$ and $\triangle ADC$ (see figure 8). Then, since the angle sum of a triangle is 180°,

$$\begin{gathered}
s(\triangle ABC) + s(\triangle ADC) = 180^\circ + 180^\circ \\
(\angle CAB) + (\angle B) + (\angle ACB) + (\angle CAD) + (\angle D) + (\angle ACD) = 360^\circ \\
(\angle A) + (\angle B) + (\angle C) + (\angle D) = 360^\circ \\
90^\circ + 90^\circ + 90^\circ + (\angle D) = 360^\circ \\
(\angle D) = 90^\circ.
\end{gathered}$$

Therefore $\angle D$ is a right angle. □

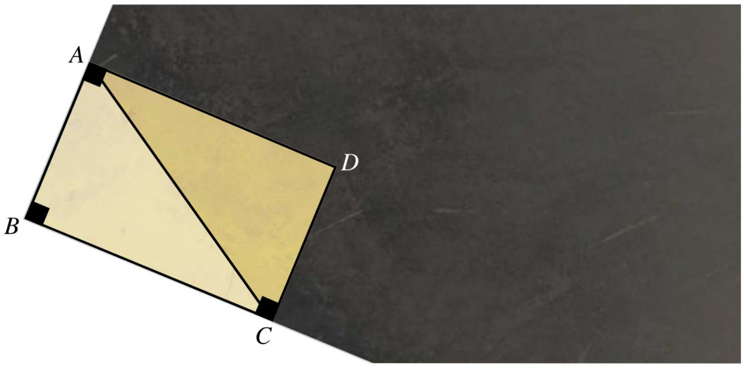

Figure 8. Verifying that the fourth angle is right.

Exercise 11.1 Fill in the gap in the proof by explaining why $\square ABCD$ must be convex.

We see then that, yes, rectangles exist in Euclidean geometry. Two essential properties of rectangles are provided in the following (easy) exercises.

Exercise 11.2 Let $\square ABCD$ be a rectangle. Prove that $\overline{AB}$ is parallel to $\overline{CD}$ and $\overline{AD}$ is parallel to $\overline{BC}$.

Exercise 11.3 Let $\square ABCD$ be a rectangle. Prove that $AB \simeq CD$ and $AD \simeq BC$ and $AC \simeq BD$.

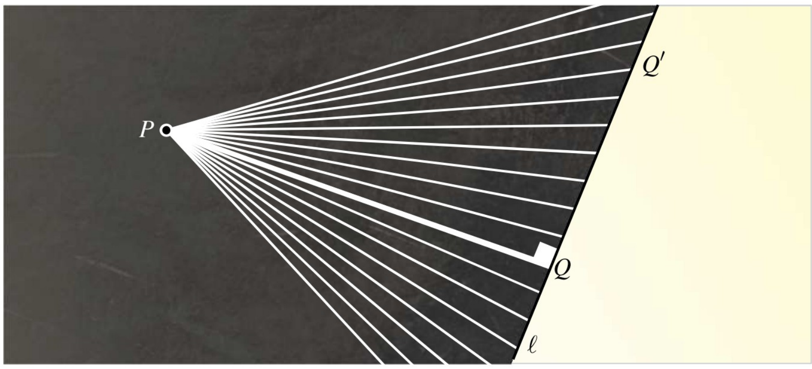

Figure 9. The distance from P to ℓ is measured along the segment from P to ℓ that is perpendicular to ℓ.

For the last result of this section, we return to parallel lines. One of the things that we will see when we study non-Euclidean geometry is that parallel lines tend to diverge from each other. That doesn't happen in Euclidean geometry. This is one of the key differences between the two geometries. Let's make this more precise. Suppose that P is a point that is not on a line ℓ. Define the distance from P to ℓ to be the minimum distance from P to a point on ℓ:

$$d(P,\ell) = \min \left\{ |PQ| \;\middle|\; Q \text{ is on } \ell \right\}.$$

Exercise 11.4 Prove that the minimum actually occurs when Q is the foot of the perpendicular to ℓ through P (this can be proved using solely the axioms of neutral geometry).

For a pair of parallel lines, that distance as measured along perpendiculars does not change.

Theorem 11.5 *If ℓ and ℓ' are parallel lines, then the distance from a point on ℓ to ℓ' is constant. In other words, if P and Q are points on ℓ, then*

$$d(P,\ell') = d(Q,\ell').$$

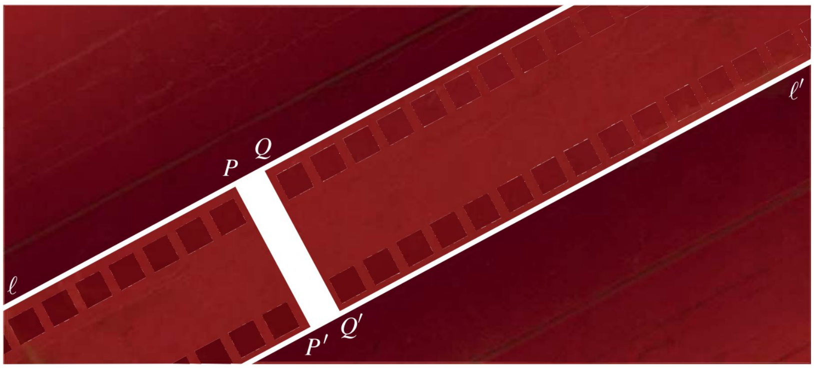

Figure 10. Perpendicular segments PP' and QQ' between parallel lines have the same length. That is, parallel lines are equidistant.

Proof Let P' and Q' be the feet of the perpendiculars on ℓ' from P and Q respectively. That way, by the previous exercise,

$$d(P,\ell') = |PP'| \text{ and } d(Q,\ell') = |QQ'|.$$

Then $\angle PP'Q'$ and $\angle QQ'P'$ are right angles. By the converse of the alternate interior angle theorem, $\angle P$ and $\angle Q$ are right angles too, so $\square PQQ'P'$ is a rectangle. According to exercise 11.2, PP' and QQ', which are opposite sides of a rectangle, are congruent. □

11.1 Exercises

11.5. Let ℓ be a line and P a point not on ℓ. Using only the results of neutral geometry, prove that there is at least one line through P that is parallel to ℓ.

11.6. Suppose that ℓ_1, ℓ_2, and ℓ_3 are three distinct lines such that ℓ_1 and ℓ_2 are parallel, and ℓ_2 and ℓ_3 are parallel. Prove that ℓ_1 and ℓ_3 are parallel.

11.7. Suppose that ℓ_1 and ℓ_2 are parallel lines, and that t is a line that intersects (but is distinct from) ℓ_1. Prove that t intersects ℓ_2.

11.8. Given three distinct parallel lines, prove that there is a labeling of them as ℓ_1, ℓ_2, and ℓ_3 so that

- no point of ℓ_1 lies between a point of ℓ_2 and a point of ℓ_3
- every point of ℓ_2 lies between a point of ℓ_1 and of ℓ_3
- no point of ℓ_3 lies between a point of ℓ_1 and a point of ℓ_2.

In this case, we can say that ℓ_2 is *between* ℓ_1 and ℓ_3.

11.9. This exercise is an extension of the previous one. Given n distinct parallel lines, prove that there is an ordering of them as $\ell_1, \ell_2, \ldots, \ell_n$ so that if $i < j < k$ then ℓ_j is between ℓ_i and ℓ_k.

11.10. Find the angle sum of a convex n-gon as a function of n.

11.11. What is the measure of an interior angle of a regular n-gon (as a function of n)?

11.12. Consider a convex quadrilateral $\square ABCD$. Prove that the two diagonals of $\square ABCD$ bisect each other if and only if $\square ABCD$ is a parallelogram.

11.13. Show that a parallelogram $\square ABCD$ is a rectangle if and only if $AC \simeq BD$.

11.14. Suppose that the diagonals of a convex quadrilateral $\square ABCD$ intersect one another at a point P and that

$$AP \simeq BP \simeq CP \simeq DP.$$

Prove that $\square ABCD$ is a rectangle.

11.15. Suppose that the diagonals of a convex quadrilateral bisect one another at right angles. Prove that the quadrilateral must be a rhombus.

11.16. Let $\triangle ABC$ be isosceles with $AB \simeq AC$, and let a be the midpoint of BC. Let b and c be points on the line through A that is parallel to BC so that $b * A * c$ and so that $bA \simeq cA \simeq Ba$. Prove that $\triangle ABC \simeq \triangle abc$.

11.17. Verify that the Cartesian model (as developed through the exercises in chapters 1–3) satisfies Playfair's axiom.

12

Parallel Projection

In the course of this chapter, we will need to use a few properties of parallelograms. It is rather common to confuse the properties of a parallelogram with its definition, so first, recall the definition of a parallelogram.

Definition 12.1 A *parallelogram* is a simple quadrilateral whose opposite sides are parallel.

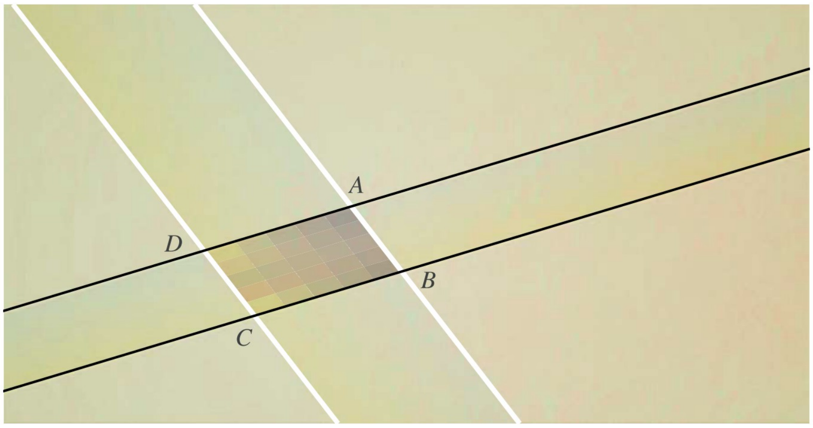

Figure 1. A parallelogram.

Now on to the properties of parallelograms that we will need for this chapter.

Exercise 12.1 Prove that in a parallelogram, the two pairs of opposite sides are congruent and the two pairs of opposite angles are congruent.

Exercise 12.2 Prove that if a convex quadrilateral has one pair of opposite sides that are both parallel and congruent, then it is a parallelogram.

Exercise 12.3 Let $\square ABB'A'$ be a simple quadrilateral. Verify that if AA' and BB' are parallel, but AB and $A'B'$ are not, then AA' and BB' cannot be congruent.

12.1 Parallel projection

The purpose of this chapter is to introduce a mechanism called *parallel projection*, a particular kind of mapping from the points on one line to the points on another. Parallel projection is the piece of machinery that you have to have in place to understand similarity, which is in turn essential for much of what we will be doing in the next chapters. The primary goal at this point is to understand how distances between points may be distorted by the parallel projection mapping. Once that is figured out, we will be able to turn our attention to the geometry of similarity.

Definition 12.2 A *parallel projection* from one line ℓ to another ℓ' is a map $\Phi : \ell \to \ell'$ that assigns to each point P on ℓ a point $\Phi(P)$ on ℓ' so that all the lines connecting a point and its image are parallel to one another.

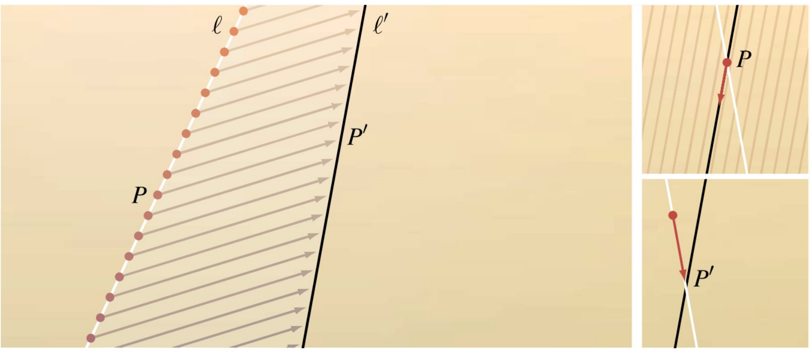

Figure 2. The path from a point P on ℓ to a point P' on ℓ' defines a parallel projection as long as neither P nor P' is the intersection of ℓ and ℓ' (as shown at right).

It is easy to construct parallel projections. Any point P on ℓ and its image $\Phi(P)$ on ℓ' completely determines the projection: for any other point Q on ℓ there is a unique line that passes through Q and is parallel to $\overline{P\,\Phi(P)}$. Wherever the line intersects ℓ' will have to be $\Phi(Q)$. There are only two cases where this construction will not work: if P is the intersection of ℓ and ℓ', then the lines of projection run parallel to ℓ' and so fail to provide a point of intersection; and if $\Phi(P)$ is the intersection of ℓ and ℓ', then the lines of projection actually coincide rather than being parallel.

Theorem 12.3 *A parallel projection is both one-to-one and onto.*

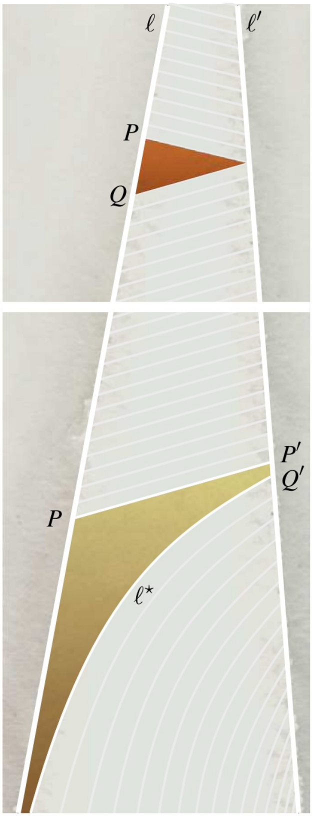

Figure 3. Using proofs by contradiction, we can show that parallel projection is both one-to-one (top) and onto (bottom).

Proof Let $\Phi : \ell \to \ell'$ be a parallel projection. First let's see why Φ is one-to-one. Suppose that it is not. That is, suppose that P and Q are two distinct points on ℓ but that $\Phi(P) = \Phi(Q)$. Then the two projecting lines $\overline{P\,\Phi(P)}$ and $\overline{Q\,\Phi(Q)}$, which ought to be parallel, actually share a point. This can't happen.

Now let's see why Φ is onto. Suppose that Q' is a point on ℓ' that is not in the image of Φ. To get a sense of how Φ is casting points from ℓ to ℓ', consider a point P on ℓ and its image $P' = \Phi(P)$ on ℓ'. By Playfair's axiom, there is a unique line $\ell^\star$ that passes through Q' and is parallel to $\overline{PP'}$. If $\ell^\star$ intersects ℓ at a point Q, then $\Phi(Q) = Q'$, contradicting the initial supposition that Q' is not in the image of Φ. If $\ell^\star$ does not intersect ℓ, then both ℓ and $\overline{PP'}$ are lines that pass through P and are parallel to $\ell^\star$, in violation of Playfair's axiom. In either case, we arrive at a contradiction. Therefore, an arbitrary point Q' of ℓ must be in the image of Φ, and so Φ is onto. □

Now that we have established that parallel projection is a bijection, it helps to use a naming convention to makes things more readable. I will use a prime mark $'$ to indicate the parallel projection of a point, so $\Phi(P) = P'$, $\Phi(Q) = Q'$, and so on.

12.2 Parallel projection, order, and congruence

So far we have seen that parallel projection establishes a correspondence between the points of one line and the points of another. What about the order of the points? Can points get shuffled up in the process of a parallel projection? No.

Theorem 12.4 *Let $\Phi : \ell \to \ell'$ be a parallel projection. If A, B, and C are points on ℓ and $A * B * C$, then $A' * B' * C'$.*

Proof Because B is between A and C, A and C must be on opposite sides of $\overline{BB'}$. But, $\overline{AA'}$ does not intersect $\overline{BB'}$, so A' has to be on the same side of $\overline{BB'}$ as A; and $\overline{CC'}$ does not

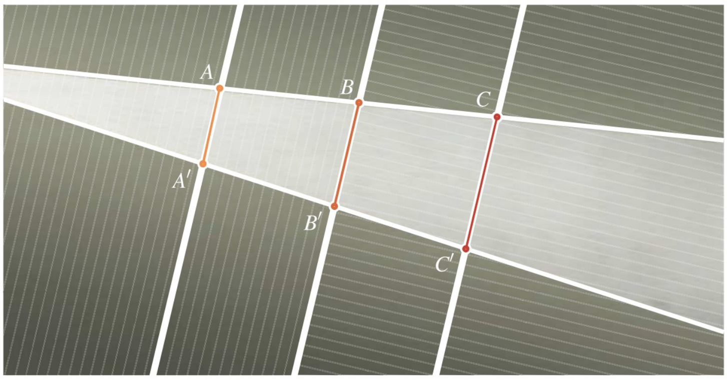

Figure 4. Parallel projection keeps points in order.

intersect $\overline{BB'}$, so C' has to be on the same side of $\overline{BB'}$ as C. Therefore A' and C' have to be on opposite sides of $\overline{BB'}$, and so the intersection of $\overline{BB'}$ and $A'C'$, which is B', must be between A' and C'. □

Now what about congruence?

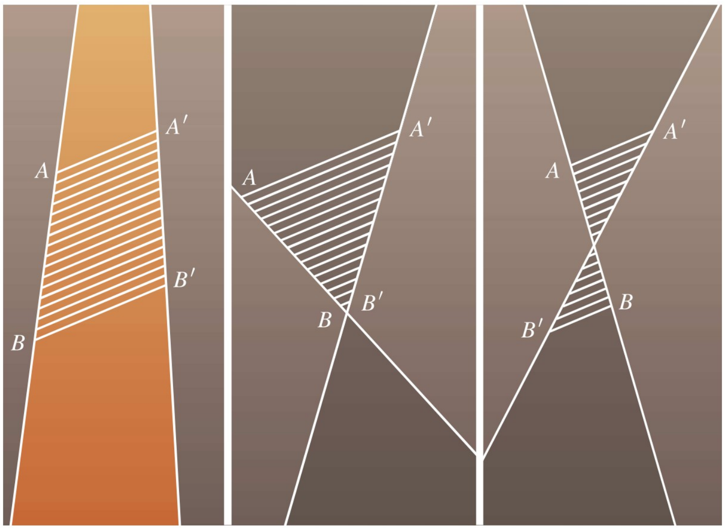

Figure 5. There are three positions for A and B relative to the image line: both on the same side, one on the image line, or one on each side. Likewise, there are three positions for a and b. There are nine cases in all.

Theorem 12.5 *Let $\Phi : \ell \to \ell'$ be a parallel projection. If a, b, A, and B are points on ℓ and if $ab \simeq AB$, then $a'b' \simeq A'B'$.*

Proof There are several cases, depending on the positions of ab and AB relative to ℓ'. They could lie on the same side of ℓ', they could lie on opposite sides of ℓ', one or both could straddle

ℓ', or one or both could have an endpoint on ℓ'. Each case must be handled differently. We will look at the case where both segments are on the same side of ℓ'. The others are left as exercises. There are still two cases to consider, depending on whether or not ℓ and ℓ' are parallel.

Case 1: when ℓ and ℓ' are parallel (see figure 6). This is the simpler case, but it helps illuminate the more general case. Because aa' is parallel to bb' and ab is parallel to $a'b'$, $\square aa'b'b$ is a parallelogram; AA' is parallel to BB' and AB is parallel to $A'B'$ so $\square AA'B'B$ is also a parallelogram. Because the opposite sides of a parallelogram are congruent (exercise 12.1), $a'b' \simeq ab$ and $AB \simeq A'B'$. Since $ab \simeq AB$, $a'b' \simeq A'B'$.

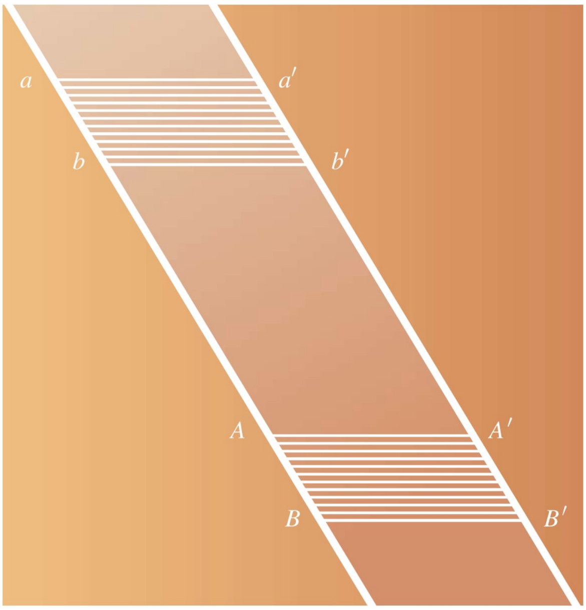

Figure 6. Case 1, when the two lines are parallel.

Case 2: when ℓ and ℓ' are not parallel (see figure 7). This is the more likely case, in which $\square aa'b'b$ and $\square AA'B'B$ will not be parallelograms. We can takes a similar approach to case 1, though, by building some parallelograms into the problem. Because ℓ and ℓ' are not parallel, the segments aa' and bb' cannot have the same length (exercise 12.3), and the segments AA' and BB' cannot have the same length. Let's assume that aa' is shorter than bb' and that AA' is shorter than BB'. If this is not the case, then it is just a matter of switching some labels to make it so. Then

- there is a point c between b and b' so that $bc \simeq aa'$, and
- there is a point C between B and B' so that $BC \simeq AA'$.

This creates four shapes of interest: the quadrilaterals $\square a'abc$ and $\square A'ABC$ that are parallelograms (exercise 12.2), and the triangles $\triangle a'b'c$ and $\triangle A'B'C$. The key here is to prove that $\triangle a'b'c \simeq \triangle A'B'C$.

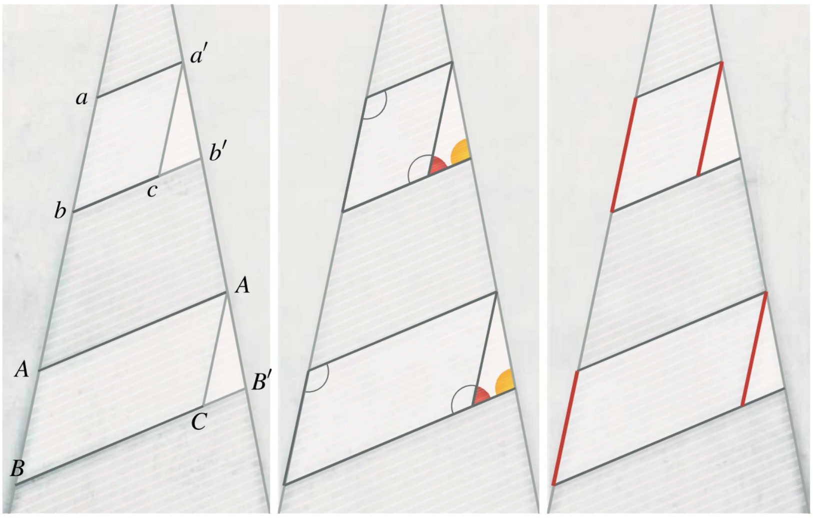

Figure 7. The non-parallel case. (l) First construct parallelograms. (c) The two angles and (r) the sides needed for AAS to show that $\triangle a'b'c' \simeq \triangle A'B'C'$.

- **A:** $\angle b' \simeq \angle B'$. The lines $\overline{cb'}$ and $\overline{CB'}$ are parallel (they are two of the projecting lines) and they are crossed by the transversal ℓ'. By the converse of the alternate interior angle theorem, $\angle a'b'c$ and $\angle A'B'C$ are congruent.
- **A:** $\angle c \simeq \angle C$. The opposite angles of the two parallelograms are congruent. Therefore $\angle a'cb \simeq \angle a'ab$ and $\angle A'AB \simeq \angle A'CB$. But $\overline{aa'}$ and $\overline{AA'}$ are parallel lines cut by the transversal ℓ, so $\angle a'ab \simeq \angle A'AB$, so $\angle a'cb \simeq \angle A'CB$, and so their supplements $\angle a'cb'$ and $\angle A'CB'$ are also congruent.
- **S:** $a'c \simeq A'C$. The opposite sides of the two parallelograms are congruent too. Therefore $a'c \simeq ab$ and $AB \simeq A'C$, and since $ab \simeq AB$, we get $a'c \simeq A'C$.

By AAS, then, $\triangle a'b'c \simeq \triangle A'B'C$. The corresponding sides $a'b'$ and $A'B'$ have to be congruent.

□

Exercise 12.4 Fill in some gaps:

(1) Prove the theorem in the case when both a and A lie on ℓ' (at the intersection of ℓ and ℓ').
(2) Prove the theorem in the case when a and A lie on opposite sides of ℓ' from b and B.

12.3 Parallel projection and distance

That brings us to the question at the very heart of parallel projection. If Φ is a parallel projection and A and B are two points on ℓ, how do the lengths $|AB|$ and $|A'B'|$ compare? In case 1 of the last proof, AB and $A'B'$ ended up being congruent, but that was because ℓ and ℓ' were parallel. In general, AB and $A'B'$ do not have to be congruent. But it turns out that in the process of parallel projecting from one line to another, all distances are scaled by a constant multiple.

Theorem 12.6 *If* $\Phi : \ell \to \ell'$ *is a parallel projection, then there is a constant* k *such that*

$$|A'B'| = k|AB|$$

for all points A *and* B *on* ℓ.

It is helpful to get a little perspective on this before the proof. First note that the previous theorem gives us a way to narrow the scope of the problem. Fix a point O on ℓ and let r be one of the two rays along ℓ with O as its endpoint. The segment construction axiom says that every segment AB on ℓ is congruent to a segment OP where P is some point on r. We have just seen that parallel projection maps congruent segments to congruent segments. So if Φ scales all segments of the form OP by a factor of k, then it must scale all the segments of ℓ by the same factor.

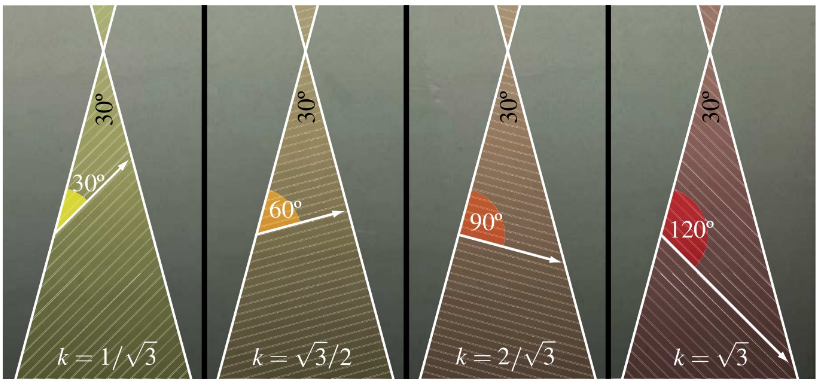

Figure 8. Some parallel projections and their scaling constants.

Second, we can extend that by considering what happens when we parallel project end-to-end congruent copies of a segment. For this, let me introduce another convenient notation convention: for the rest of this argument, let P_d denote the point on r that is a distance d from O. Now, pick a positive real value x, and let

$$k = |O'P'_x|/|OP_x|,$$

so that Φ scales the segment OP_x by a factor of k. Eventually we will have to show that Φ scales all segments by that factor, but for now let's restrict our attention to the segments OP_{nx}, where n is a positive integer. Between O and P_{nx} are P_x, P_{2x}, … $P_{(n-1)x}$ in order:

$$O * P_x * P_{2x} * \cdots * P_{(n-1)x} * P_{nx}.$$

We have seen that parallel projection preserves the order of points, so

$$O' * P'_x * P'_{2x} * \cdots * P'_{(n-1)x} * P'_{nx}.$$

Each segment $P_{ix}P_{(i+1)x}$ is congruent to OP_x and consequently each parallel projection $P'_{ix}P'_{(i+1)x}$ is congruent to $O'P'_x$. Add their lengths:

$$\begin{aligned}|O'P'_{nx}| &= |O'P'_x| + |P'_xP'_{2x}| + |P'_{2x}P'_{3x}| + \cdots + |P'_{(n-1)x}P'_{nx}|\\ &= kx + kx + kx + \cdots + kx \quad (n \text{ times})\\ &= k \cdot nx\end{aligned}$$

and so Φ scales OP_{nx} by a factor of k.

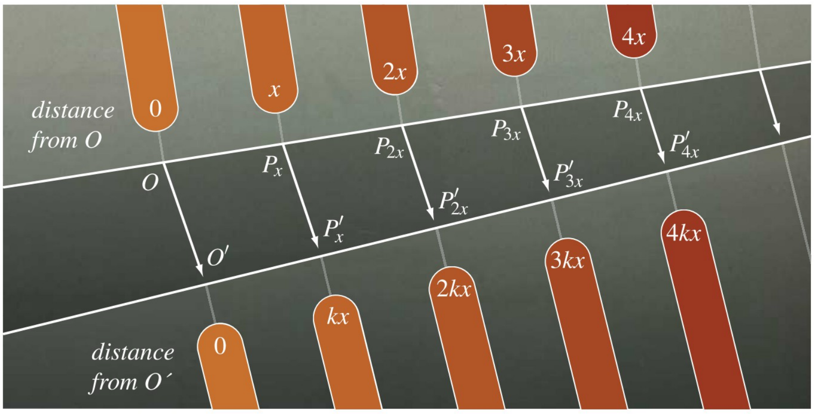

Figure 9. Scaling of segments of length kx, where k is an integer.

Sadly, no matter what x is, the points P_{nx} account for an inconsequential portion of the set of all points of r. However, if OP_x and OP_y were to have two different scaling factors we could use this end-to-end copying to magnify the difference between them. Third, then, consider an example that demonstrates how this works, and which shows how this ultimately prevents there being two different scaling factors. In this example, let's suppose that $|O'P'_1| = 2$, so all integer length segments on ℓ are scaled by a factor of 2. Let's take a look at what this means for $P_{3.45}$. Let k be the scaling factor for $OP_{3.45}$. Figure 10 shows how the first few end-to-end copies of $OP_{3.45}$ begin to limit the possible values of k.

Fourth, note that the floor function, $f(x) = \lfloor x \rfloor$, assigns to each real number x the largest integer that is less than or equal to it. The ceiling function, $f(x) = \lceil x \rceil$, assigns to each real number x the smallest integer that is greater than or equal to it.

It is finally time to prove that parallel projection scales distance.

Proof Let $k = |O'P'_1|$ so that k is the scaling factor for the segment of length 1 (and consequently for all integer length segments). Now take a point P_x on ℓ and let k' be the scaling factor for the segment OP_x. We want to show that $k' = k$ and we can do that by following the same strategy as in the example above, by capturing k' in an increasingly narrow band around k by looking at the parallel projection of P_{nx} as n increases. Start by working back and forth between

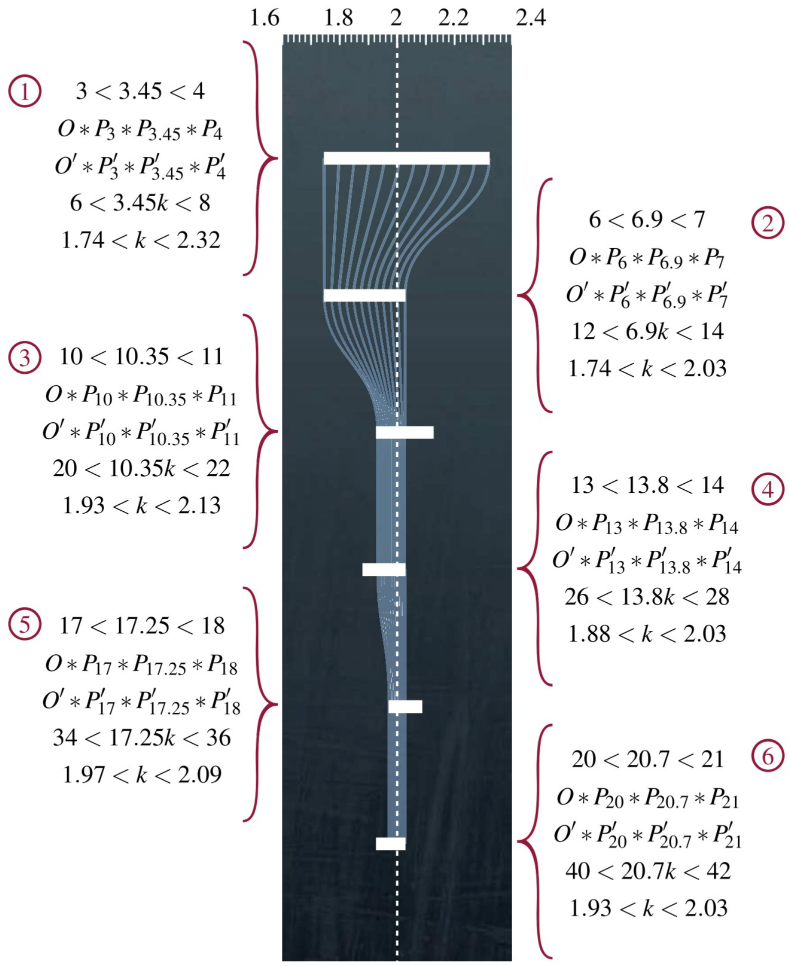
1.6
1.8
2
2.2
2.4
①
$3 < 3.45 < 4$
$O * P_3 * P_{3.45} * P_4$
$O' * P'_3 * P'_{3.45} * P'_4$
$6 < 3.45k < 8$
$1.74 < k < 2.32$
②
$6 < 6.9 < 7$
$O * P_6 * P_{6.9} * P_7$
$O' * P'_6 * P'_{6.9} * P'_7$
$12 < 6.9k < 14$
$1.74 < k < 2.03$
③
$10 < 10.35 < 11$
$O * P_{10} * P_{10.35} * P_{11}$
$O' * P'_{10} * P'_{10.35} * P'_{11}$
$20 < 10.35k < 22$
$1.93 < k < 2.13$
④
$13 < 13.8 < 14$
$O * P_{13} * P_{13.8} * P_{14}$
$O' * P'_{13} * P'_{13.8} * P'_{14}$
$26 < 13.8k < 28$
$1.88 < k < 2.03$
⑤
$17 < 17.25 < 18$
$O * P_{17} * P_{17.25} * P_{18}$
$O' * P'_{17} * P'_{17.25} * P'_{18}$
$34 < 17.25k < 36$
$1.97 < k < 2.09$
⑥
$20 < 20.7 < 21$
$O * P_{20} * P_{20.7} * P_{21}$
$O' * P'_{20} * P'_{20.7} * P'_{21}$
$40 < 20.7k < 42$
$1.93 < k < 2.03$

Figure 10. An example. Projecting end-to-end congruent copies of a segment of length 3.45 by a parallel projection that scales all integer length segments by a factor of two.

inequalities of real numbers and orders of points:

$$\lfloor nx \rfloor \le nx \le \lceil nx \rceil$$
$$O * P_{\lfloor nx \rfloor} * P_{nx} * P_{\lceil nx \rceil}$$
$$O' * P'_{\lfloor nx \rfloor} * P'_{nx} * P'_{\lceil nx \rceil}$$
$$k\lfloor nx \rfloor \le k'nx \le k\lceil nx \rceil.$$

(The notation here is a bit casual—if nx is an integer, then $\lfloor nx \rfloor = nx$, so $P_{\lfloor nx \rfloor}$ is not actually between O and P_{nx}. This does not cause any problems for the proof.) Because $(nx - 1) \le \lfloor nx \rfloor$

and $\lceil nx \rceil \leq (nx + 1)$ we can move to a less precise but more manageable set of inequalities:

$$k(nx - 1) \leq k'nx \leq k(nx + 1).$$

Divide by nx to get

$$k\left(\frac{nx - 1}{nx}\right) \leq k' \leq k\left(\frac{nx + 1}{nx}\right).$$

As n increases, the ratios $(nx - 1)/(nx)$ and $(nx + 1)/(nx)$ both approach 1. In the limit as n goes to infinity, they are 1. Since the inequalities have to be true for all n, the only possible value for k' , then, is k. □

12.4 Exercises

12.5. Suppose that Φ is a parallel projection from ℓ to ℓ'. If ℓ and ℓ' intersect at P, prove that $\Phi(P) = P$.

12.6. Prove that if ℓ and ℓ' are parallel, then the scaling factor of a parallel projection between them must be one, but that if ℓ and ℓ' are not parallel, then there is a parallel projection with every possible scaling factor k where $0 < k < \infty$.

12.7. Let $\Phi : \ell \to \ell'$ be a parallel projection. Show that the inverse map $\Phi^{-1} : \ell' \to \ell$ is also a parallel projection.

12.8. Let $\Phi_1 : \ell \to \ell'$ and $\Phi_2 : \ell' \to \ell''$ be parallel projections. Is the composition $\Phi_2 \circ \Phi_1 : \ell \to \ell''$ necessarily a parallel projection?

12.9. In chapter 6, we constructed a distance function, and one of the keys to the construction was locating the points on a ray that were a distance of $m/2^n$ from its endpoint. In Euclidean geometry, there is a construction that locates all the points on a ray that are any rational distance m/n from its endpoint. Take two non-opposite rays r and r' with a common endpoint O. Along r, lay out m congruent copies of a segment of length 1, ending at the point P_m. Along r', lay out n congruent copies of a segment of length 1, ending at the point Q_n. Mark the point Q_1 on r' that is a distance 1 from O. Verify that the line that passes through Q_1 and is parallel to $P_m Q_n$ intersects r a distance of m/n from O.

;. 5 : 6 * ,

13

Similarity

In the chapters on neutral geometry, we spent a lot of effort to gain an understanding of polygon congruence, particularly congruence of triangles and quadrilaterals. Remember that polygon congruence is an equivalence relation (it is reflexive, symmetric, and transitive). It turns out that congruence is not the only important equivalence relation between polygons, though, and the purpose of this chapter is to investigate another: similarity. Similarity is a less demanding relation than congruence. Intuitively, we tend to think of congruent polygons as exactly the same, just positioned differently. Similar polygons are not just positioned differently, but are scaled versions of one another: the same shape, but possibly different sizes.

Definition 13.1 Two n-sided polygons $P_1P_2\dots P_n$ and $Q_1Q_2\dots Q_n$ are *similar* to one another if they meet two sets of conditions:

(1) corresponding interior angles are congruent:

$$\angle P_i \simeq \angle Q_i, \quad 1 \le i \le n.$$

(2) corresponding side lengths differ by the same constant multiple k:

$$|P_iP_{i+1}| = k \cdot |Q_iQ_{i+1}|, \quad 1 \le i \le n.$$

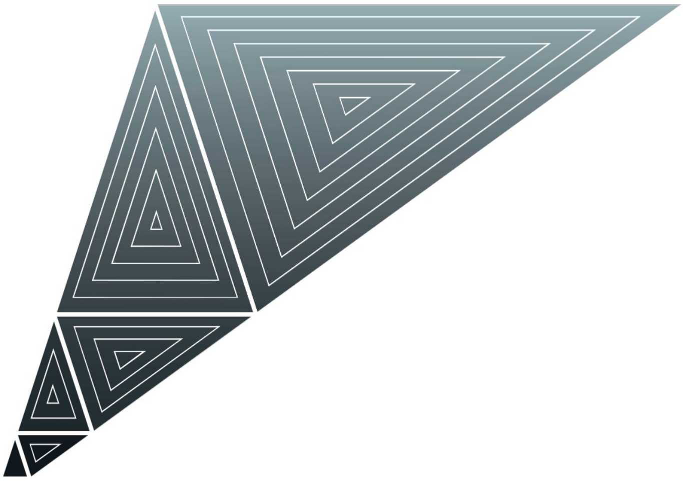

Figure 1. An arrangement of similar triangles.

To indicate that two polygons $P_1P_2\dots P_n$ and $Q_1Q_2\dots Q_n$ are similar, write $P_1P_2\dots P_n \sim Q_1Q_2\dots Q_n$. There are a few things worth noting here. First, if polygons are congruent, they will be similar as well with the scaling constant 1. Second, it is easy to see that similarity is reflexive, symmetric and transitive and so is an equivalence relation. Third, when you jump from one polygon to another similar polygon, the corresponding segment lengths are scaled by the same amount. That behavior echoes the work we did in the last chapter, and for good reason: parallel projection underlies everything that we are going to do in this chapter.

Much of the time, when working with either parallel projection or similarity, the scaling constant is not important. The only thing that matters is that there is one. Fortunately, the existence of a scaling constant can be indicated without ever mentioning what it is. The key to doing this is ratios. Consider a parallel projection from line ℓ to line ℓ'. Let A, B, a, and b be points on ℓ and let A', B', a', and b' be their respective images on ℓ'. The main result of the last chapter was that there is a scaling constant k so that

$$|A'B'| = k \cdot |AB| \quad\text{and}\quad |a'b'| = k \cdot |ab|.$$

The ratios are only a step away from this pair of equations.

Ratio 1: Solve for k in both equations and set them equal to each other:

$$\frac{|A'B'|}{|AB|} = \frac{|a'b'|}{|ab|}.$$

Ratio 2: Starting from the first ratio, multiply by $|AB|$ and divide by $|a'b'|$:

$$\frac{|A'B'|}{|a'b'|} = \frac{|AB|}{|ab|}.$$

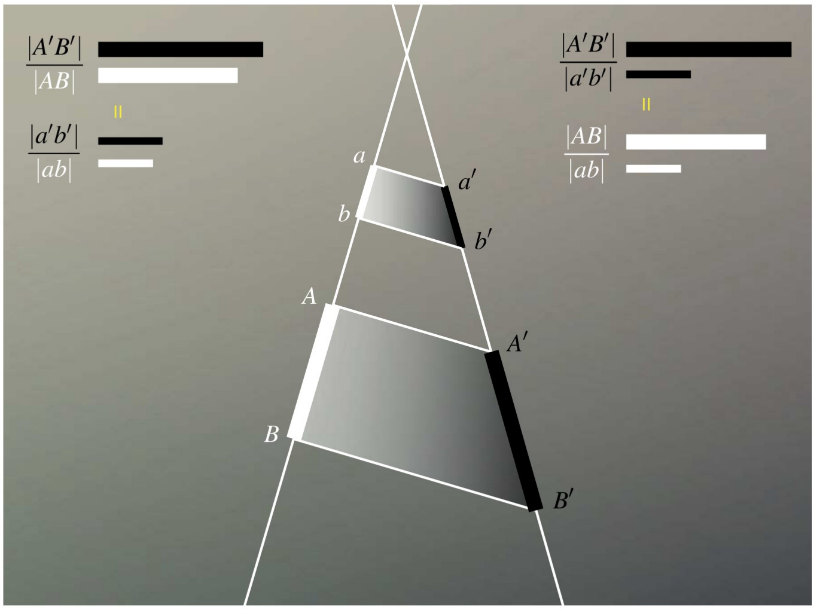

Figure 2. Two invariant ratios of a parallel projection.

13.1 Triangle similarity theorems

Let's begin with a few theorems that deal with similarity of triangles. I like to think of them as degenerations of the triangle congruence theorems, where the strict condition of side congruence, $A'B' \simeq AB$, is replaced with the more flexible condition of constant scaling, $|A'B'| = k|AB|$. The first is the SAS similarity theorem.

Theorem 13.2 *(SAS similarity) In $\triangle ABC$ and $\triangle A'B'C'$, if $\angle A \simeq \angle A'$ and if there is a constant k so that*

$$|A'B'| = k \cdot |AB| \text{ and } |A'C'| = k \cdot |AC|,$$

then $\triangle ABC \sim \triangle A'B'C'$.

Proof First note that as with parallel projection, the second condition in the SAS similarity theorem can be recast in terms of ratios:

$$\frac{|A'B'|}{|AB|} = \frac{|A'C'|}{|AC|} \text{ or } \frac{|A'B'|}{|A'C'|} = \frac{|AB|}{|AC|}.$$

With that said, what we need to do in the proof is to establish two more angle congruences, that $\angle B \simeq \angle B'$ and $\angle C \simeq \angle C'$, and one more ratio of sides, that $|B'C'| = k|BC|$. Two parallel projections form the backbone of this proof.

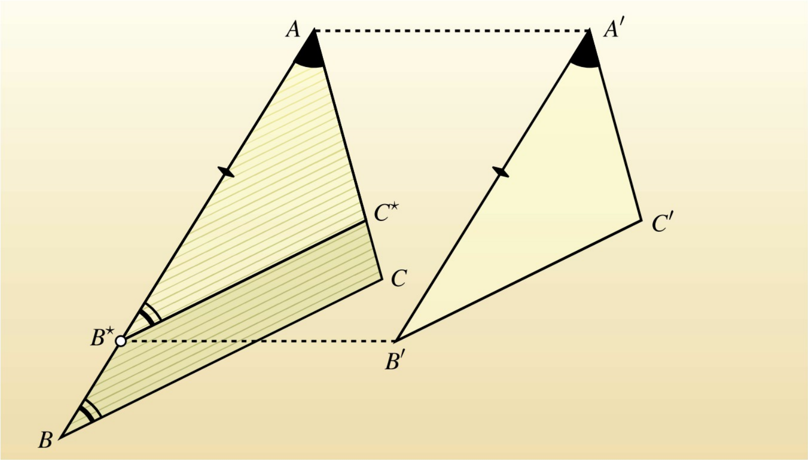

Figure 3. Proving SAS similarity. The first parallel projection.

The first parallel projection. The primary purpose of the first projection is to build a transitional triangle that is congruent to $\triangle A'B'C'$ but positioned on top of $\triangle ABC$. Begin by locating the point $B^\star$ on $\overrightarrow{AB}$ so that $AB^\star \simeq A'B'$. We cannot know the exact location of $B^\star$ relative to B on this ray. It depends on whether $A'B'$ is shorter or longer than AB. For this argument, assume that $A'B'$ is shorter than AB, which will place $B^\star$ between A and B (the other case is not substantially different). Consider the parallel projection

$$\Phi_1 : (\overline{AB}) \longrightarrow (\overline{AC})$$

that maps B to C. Because A is the intersection of these two lines, $\Phi_1(A) = A$. Label $C^\star = \Phi_1(B^\star)$. Let's see how the newly formed $\triangle AB^\star C^\star$ compares with $\triangle A'B'C'$. By parallel projection,

$$\frac{|AC^\star|}{|AC|} = \frac{|AB^\star|}{|AB|},$$

but we chose $B^\star$ so that $|AB^\star| = |A'B'|$. Therefore

$$\frac{|AC^\star|}{|AC|} = \frac{|A'B'|}{|AB|} = \frac{|A'C'|}{|AC|},$$

and so $|AC^\star| = |A'C'|$. Put that together with what we already know, that $AB^\star \simeq A'B'$ and $\angle A \simeq \angle A'$, and by SAS, $\triangle A'B'C'$ and $\triangle AB^\star C^\star$ are congruent. In particular, that means $\angle B' \simeq \angle B^\star$ and $\angle C' \simeq \angle C^\star$. Now let's turn back to see how $\triangle AB^\star C^\star$ relates to $\triangle ABC$. To locate $C^\star$, we used a projection that was parallel to $\overline{BC}$. Hence $\overline{B^\star C^\star}$ and $\overline{BC}$ are parallel, and so, by the converse of the alternate interior angle theorem, $\angle B^\star \simeq \angle B$ and $\angle C^\star \simeq \angle C$. Since angle congruence is transitive, we now have the required angle congruences, $\angle B \simeq \angle B'$ and $\angle C \simeq \angle C'$.

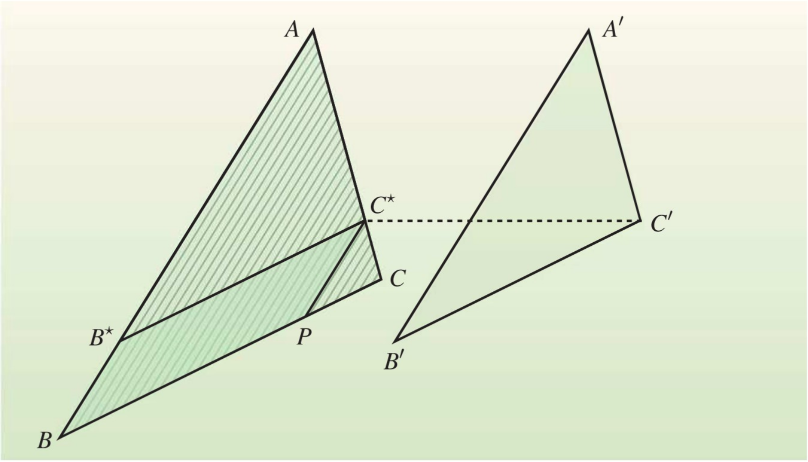

Figure 4. Proving SAS similarity. The second parallel projection.

The second parallel projection. Now consider the parallel projection

$$\Phi_2 : (\overline{AC}) \longrightarrow (\overline{BC})$$

that maps A to B. Since C is the intersection of $\overline{AC}$ and $\overline{BC}$, $\Phi_2(C) = C$. The other point of interest this time is $C^\star$. Let $P = \Phi_2(C^\star)$. In doing so, we have effectively carved out a parallelogram $\square BB^\star C^\star P$. Recall (from the last chapter) that the opposite sides of a parallelogram are congruent, so $B^\star C^\star \simeq BP$, and therefore $|B'C'| = |BP|$. The parallel projection Φ_2 establishes equal ratios $|BP|/|BC| = |AC^\star|/|AC|$, and so

$$\frac{|B'C'|}{|BC|} = \frac{|BP|}{|BC|} = \frac{|AC^\star|}{|AC|} = \frac{|A'C'|}{|AC|} = k.$$

Thus, $|B'C'| = k|BC|$, as needed. □

Back in the neutral geometry chapters, after SAS we next encountered ASA and AAS. Unlike SAS, both theorems reference only one pair of sides in the triangles. Let's look at what happens when we try to modify those congruence conditions into similarity conditions.

ASA congruence	ASA similarity (?)
$\angle A \simeq \angle A'$	$\angle A \simeq \angle A'$
$AB \simeq A'B'$	$\|A'B'\| = k \cdot \|AB\|$
$\angle B \simeq \angle B'$	$\angle B \simeq \angle B'$

AAS congruence	AAS similarity (?)
$\angle A \simeq \angle A'$	$\angle A \simeq \angle A'$
$\angle B \simeq \angle B'$	$\angle B \simeq \angle B'$
$BC \simeq B'C'$	$\|B'C'\| = k \cdot \|BC\|$

In each of these conversions, the condition on the one side is automatically satisfied: there will always be a real value of k that makes the equation true. That suggests that it may take only two angle congruences to guarantee similarity.

Theorem 13.3 *(AA similarity) In $\triangle ABC$ and $\triangle A'B'C'$, if $\angle A \simeq \angle A'$ and $\angle B \simeq \angle B'$, then $\triangle ABC \sim \triangle A'B'C'$.*

We have plenty of information about the angles, so what we need is some information about ratios of sides. If we can show that

$$\frac{|A'B'|}{|AB|} = \frac{|A'C'|}{|AC|},$$

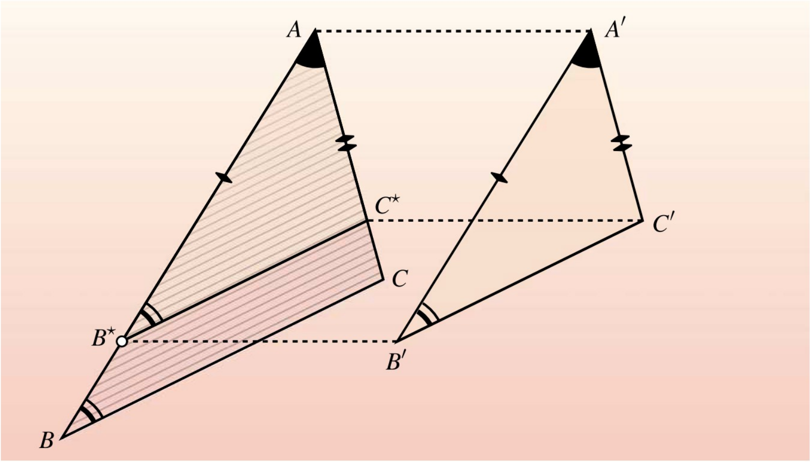

Figure 5. Proof of AA similarity, by parallel projection.

then that, along with the congruence $\angle A \simeq \angle A'$, will be enough to use the SAS similarity theorem.

Proof Start by constructing a transition triangle: one that is positioned on top of $\triangle ABC$ but that is congruent to $\triangle A'B'C'$. To do that, locate $B^\star$ on $\overrightarrow{AB}$ so that $AB^\star \simeq A'B'$, and $C^\star$ on $\overrightarrow{AC}$ so that $AC^\star \simeq A'C'$. By SAS, $\triangle AB^\star C^\star$ and $\triangle A'B'C'$ are congruent. Consider the congruent angles

$$\angle B^\star \simeq \angle B' \simeq \angle B.$$

According to the alternate interior angle theorem, $\overline{B^\star C^\star}$ and $\overline{BC}$ must be parallel. Therefore the parallel projection from $\overline{AB}$ to $\overline{AC}$ that maps B to C and A to itself will also map $B^\star$ to $C^\star$ and so

$$\frac{|A'B'|}{|AB|} = \frac{|AB^\star|}{|AB|} = \frac{|AC^\star|}{|AC|} = \frac{|A'C'|}{|AC|}.$$

The first and last terms give the ratio we need. That, together with the known congruence $\angle A \simeq \angle A'$, is enough for SAS similarity, so $\triangle ABC \sim \triangle A'B'C'$. □

While AAA was not enough to guarantee congruence, thanks to the result above, we now know that it is (more than) enough to guarantee similarity. The last of the triangle similarity theorems is SSS similarity (SSA, which just misses as a congruence theorem, is done in again by the same counterexample).

Theorem 13.4 *(SSS similarity) In $\triangle ABC$ and $\triangle A'B'C'$, if there is a constant k so that*

$$|A'B'| = k \cdot |AB|, \quad |B'C'| = k \cdot |BC|, \quad \textit{and} \quad |C'A'| = k \cdot |CA|,$$

then $\triangle ABC \sim \triangle A'B'C'$.

Exercise 13.1 Prove the SSS triangle similarity theorem.

13.2 The Pythagorean theorem

Before we close this chapter, let's meet one of the real celebrities of the subject, the Pythagorean theorem. There are many, many proofs of this theorem. This one involves dividing the triangle into two smaller triangles, each similar to the initial triangle, and then working with ratios.

Theorem 13.5 *(The Pythagorean theorem) Let $\triangle ABC$ be a right triangle whose right angle is at the vertex C. Let*

$$a = |BC|, \quad b = |AC|, \quad \textit{and} \quad c = |AB|.$$

Then $c^2 = a^2 + b^2$.

Proof Let D be the foot of the perpendicular to AB through C. The segment CD divides $\triangle ABC$ into two smaller triangles: $\triangle ACD$ and $\triangle BCD$. Label the lengths of their sides as $c_1 = |AD|$, $c_2 = |BD|$, and $d = |CD|$, and note that $c = c_1 + c_2$. Now $\triangle ADC$ shares $\angle A$ with $\triangle ACB$, and they both have a right angle, so by the AA similarity theorem, $\triangle ADC \sim \triangle ACB$. Similarly, $\triangle BDC$ shares $\angle B$ with $\triangle ACB$, and they both have a right angle as well, so again by AA similarity, $\triangle BDC \sim \triangle BCA$. From these similarities, there are many ratios, but the two that we need are

$$\frac{a}{c} = \frac{c_2}{a} \implies a^2 = c \cdot c_2 \text{ and } \frac{b}{c} = \frac{c_1}{b} \implies b^2 = c \cdot c_1.$$

All that is left is to add the equations and simplify to get the Pythagorean theorem:

$$a^2 + b^2 = c \cdot c_2 + c \cdot c_1 = c(c_2 + c_1) = c^2. \qquad \square$$

Here is a first application of the Pythagorean theorem. The triangle inequality says that if a, b, and c are the lengths of the sides of a triangle, then

$$|a - b| < c < a + b.$$

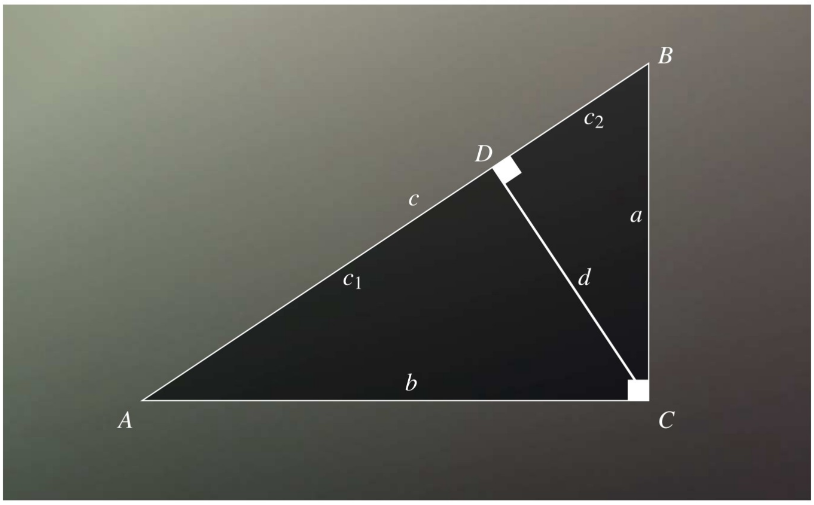

Figure 6. A proof of the Pythagorean theorem via similarity.

What about the converse? If a, b, and c are positive reals satisfying the triangle inequality conditions, can we put together a triangle with sides of those lengths?

Theorem 13.6 *Let a, b, and c be positive real numbers. Suppose that c is the largest of them and that $c < a + b$. Then there is a triangle with sides of length a, b, and c.*

Proof Start with a segment AB whose length is c. We need to place a third point C so that it is a distance a from B and b from A (according to SSS, there is only one such triangle up to congruence, so this may not be too easy). The strategy is to build the triangle out of two right triangles. Mark D on $\overrightarrow{AB}$ and label $d = |AD|$. Mark C on one of the rays with endpoint D

that is perpendicular to AB and label $e = |CD|$. Then $\triangle ACD$ and $\triangle BCD$ are right triangles. Furthermore, by sliding D and C along their respective rays, we can make d and e any positive numbers.

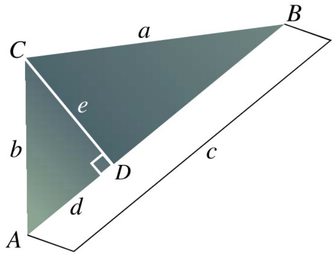

Figure 7. Building a triangle out of two right triangles.

We need to see if it is possible to position the two so that $|AC| = b$ and $|BC| = a$. To get $|AC| = b$, we will need

$$d^2 + e^2 = b^2.$$

To get $|BC| = a$, we will need

$$(c - d)^2 + e^2 = a^2.$$

With a little algebra, we can find d and e. Solve for e^2 in the equations and set them equal:

$$\begin{gathered} b^2 - d^2 = e^2 = a^2 - (c-d)^2 \\ b^2 - d^2 = a^2 - c^2 + 2cd - d^2 \\ b^2 = a^2 - c^2 + 2cd \\ (b^2 - a^2 + c^2)/2c = d. \end{gathered}$$

Since we required $c > a$, d will be positive. Substitute to find e:

$$e^2 = b^2 - d^2 = b^2 - \left(\frac{b^2 - a^2 + c^2}{2c}\right)^2.$$

Here is the essential part. Because we will have to take a square root to find e, the right hand side of this equation has to be positive; if not the equation has no solution and the triangle cannot be built. Let's go back to see if the triangle inequality condition on the three sides will help:

$$\begin{gathered} c < a + b \\ c - b < a \\ (c-b)^2 < a^2 \\ c^2 - 2bc + b^2 < a^2 \\ c^2 - a^2 + b^2 < 2bc \\ (c^2 - a^2 + b^2)/2c < b \\ ((c^2 - a^2 + b^2)/2c)^2 < b^2 \\ 0 < b^2 - ((c^2 - a^2 + b^2)/2c)^2 \end{gathered}$$

which is exactly what we want. As long as $c < a + b$, then, a value for e can be found, and the triangle can be built. $\square$

13.3 Exercises

13.2. Prove that similarity of polygons is an equivalence relation.

13.3. State and prove the SASAS similarity theorem for convex quadrilaterals.

13.4. State and prove the ASASA similarity theorem for convex quadrilaterals.

13.5. State and prove the SSSSA similarity theorem for convex quadrilaterals.

13.6. Is AAAA a valid similarity theorem for quadrilaterals?

13.7. Suppose that $P_1P_2\cdots P_n$ and $Q_1Q_2\cdots Q_n$ are similar polygons, and P_1P_j is a diagonal that lies entirely in the interior of $P_1P_2\cdots P_n$ (other than its endpoints). Prove that Q_1Q_j lies entirely in the interior of $Q_1Q_2\cdots Q_n$ (other than its endpoints).

13.8. Suppose that $P_1P_2\cdots P_n$ and $Q_1Q_2\cdots Q_n$ are similar polygons, and P_1P_j is a diagonal that lies entirely in the interior of $P_1P_2\cdots P_n$ as in the last exercise. Prove that $P_1P_2\cdots P_j \sim Q_1Q_2\cdots Q_j$.

13.9. For $\triangle ABC$, the lines

- through A, parallel to BC;
- through B, parallel to AC; and
- through C, parallel to AB

intersect to form a triangle. Prove that the triangle is similar to $\triangle ABC$.

13.10. Let $\square ABCD$ be a trapezoid, with $AB \parallel CD$. Let M_1 be the midpoint of AB, M_2 the midpoint of CD, N_1 the midpoint of BC, and N_2 the midpoint of AD. Then the segments M_1M_2 and N_1N_2 cut $\square ABCD$ into four pieces. Under what conditions (on the original $\square ABCD$) are the four pieces each similar to the original $\square ABCD$?

13.11. Let $\triangle ABC$ be an isosceles triangle with $AB \simeq AC$. Label D so that $B * C * D$ and so that $AC \simeq CD$ (so that $\triangle ACD$ is isosceles). What angle measure or measures must $\angle B$ have in order for the large triangle $\triangle DAB$ to be similar to $\triangle ABC$?

13.12. Consider a rectangle $\square ABCD$ with $|AB| < |BC|$, and suppose that it has the property that if a square $\square ABEF$ is constructed inside $\square ABCD$, then the remaining rectangle $\square ECDF$ is similar to the original $\square ABCD$. A rectangle with this property is called a golden rectangle. Find the value of $|BC|/|AB|$, a value known as the golden ratio.

13.13. Let $\triangle ABC$ be a right triangle with right angle at C, and suppose that $|AB| = a$ and $|BC| = a/2$. Let M be the midpoint of AB. The line through M that is perpendicular to AB intersects AC at a point N. What is $|MN|$ in terms of a?

13.14. Let $\triangle ABC$ be an equilateral triangle with sides of length a. Prove that the three segments that connect each vertex to the midpoint of the opposite side intersect at a point P. What is the distance from a vertex to P in terms of a?

The six trigonometric functions can be defined, for values of θ between 0 and 90°, as ratios of pairs of sides of a right triangle with an interior angle θ. If the length of the hypotenuse is h, the length of the leg adjacent to θ is a, and the length of the leg opposite θ is o, then these functions are defined as

$$\begin{aligned}
&\text{sine: } \sin(\theta) = o/h\\
&\text{cosine: } \cos(\theta) = a/h\\
&\text{tangent: } \tan(\theta) = o/a\\
&\text{cotangent: } \cot(\theta) = a/o\\
&\text{secant: } \sec(\theta) = h/a\\
&\text{cosecant: } \csc(\theta) = h/o.
\end{aligned}$$

In chapter 21, we will extend these functions to accept all real values. In the meantime, we will have need of these functions, and some of the basic identities relating them as listed in the exercises below.

13.15. Verify that the six trigonometric functions are well-defined. That is, show that it does not matter which right triangle with interior angle θ you choose—these six ratios will not change.

13.16. Verify the Pythagorean identities (for values of θ between 0 and 90°).

$$\begin{aligned}
\sin^2\theta + \cos^2\theta &= 1\\
1 + \tan^2\theta &= \sec^2\theta\\
1 + \cot^2\theta &= \csc^2\theta
\end{aligned}$$

13.17. Verify the cofunction identities (for values of θ between 0 and 90°).

$$\begin{aligned}
\sin(90^\circ - \theta) &= \cos\theta\\
\cos(90^\circ - \theta) &= \sin\theta\\
\tan(90^\circ - \theta) &= \cot\theta\\
\cot(90^\circ - \theta) &= \tan\theta\\
\sec(90^\circ - \theta) &= \csc\theta\\
\csc(90^\circ - \theta) &= \sec\theta
\end{aligned}$$

;. 5 6

14

Circles

This is the first of two chapters that deal with circles. This chapter gives some basic definitions and some elementary theorems, the most important of which is the inscribed angle theorem. In the next, we will tackle the important issue of circumference and see how that leads to the radian angle measurement system.

14.1 Definitions

Circles are basic geometric objects, but are making a late entrance in this book. In fact, circles can be defined very early in neutral geometry, the only prerequisites being points, segments, and congruence. However, most of the interesting theorems involving circles are specific to Euclidean geometry (or, in the case of the circumference formula, are different in Euclidean and non-Euclidean geometry). In any case, now is the time for a proper definition.

Definition 14.1 For a point O and positive real number r, the *circle* with *center* O and *radius* r is the set of points that are a distance r from O.

A few observations.

(1) A circle is a set, but we say that a point in the set is *on* the circle, rather than *in* the circle. The term *in* is reserved for points in the interior of the circle.
(2) In the definition, the radius is a number. We often talk about the radius as a geometric entity though, as one of the segments from the center to a point on the circle.
(3) We tend to think of the center of a circle as a fundamental part of it, but it is not a point on the circle.
(4) It is not common to talk about circles as congruent or not congruent. If you did, you would say that two circles are congruent if and only if they have the same radius.

Before we get into anything complicated, let's make some related definitions.

Definition 14.2 A line segment with both endpoints on a circle is called a *chord* of the circle. A chord that passes through the center of the circle is called a *diameter* of the circle.

Just like the term radius, the term diameter plays two roles, a numerical one and a geometric one. The diameter in the numerical sense is just the length of the diameter in the geometric sense (and is twice the radius).

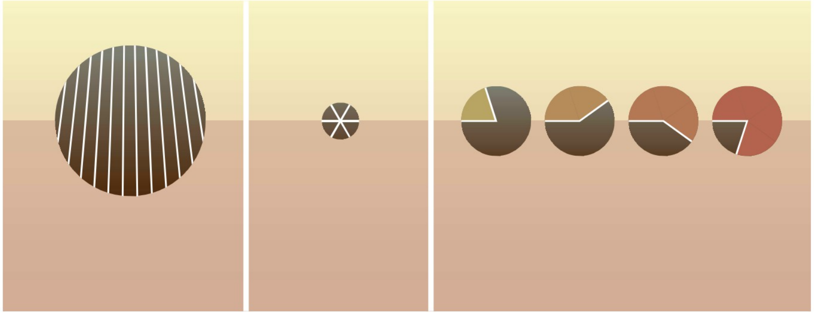

Figure 1. Twelve chords, three diameters, and four central angles.

Definition 14.3 An angle with its vertex at the center of a circle is called a *central angle* of the circle.

We will see in the next section that a line intersects a circle at most twice. Therefore, if AB is a chord of a circle, then all the points of the circle other than A and B are on one side or the other of $\overline{AB}$. Thus $\overline{AB}$ separates the points into two sets, called *arcs* of the circle. There are three types of arcs—semicircles, major arcs, and minor arcs—depending on where the chord crosses the circle.

Definition 14.4 Let AB be a diameter of a circle $\mathcal{C}$. All the points of $\mathcal{C}$ that are on one side of $\overline{AB}$, together with the endpoints A and B, form a *semicircle*.

Each diameter divides the circle into two semicircles, overlapping at the endpoints A and B.

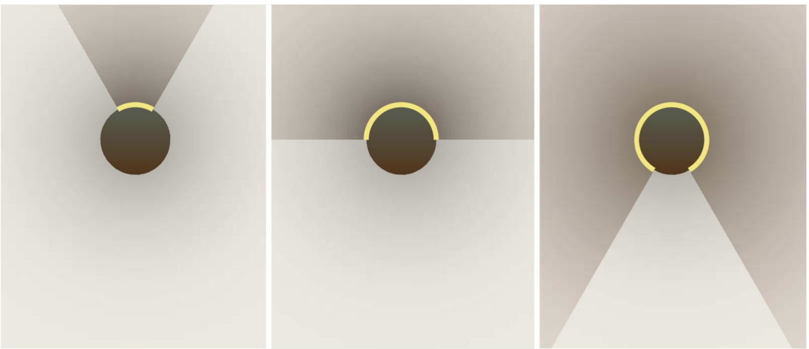

Figure 2. (l) A minor arc, (c) a semicircle, and (r) a major arc.

Definition 14.5 Let AB be a chord of a circle $\mathcal{C}$ that is not a diameter, and let O be the center of the circle. The points of $\mathcal{C}$ that are on the same side of $\overline{AB}$ as O, together with the endpoints A and B, form a *major arc*. The points of $\mathcal{C}$ that are on the opposite side of $\overline{AB}$ from O, together with the endpoints A and B, form a *minor arc*.

Like the two semicircles defined by a diameter, the major arc and minor arc defined by a chord overlap only at the endpoints A and B. We will use the notation $\overset{\frown}{AB}$ for arcs, including diameters. The notation is ambiguous, as it does not specify whether we are referring to the minor arc or to the major arc. In practice, this ambiguity rarely causes problems. Furthermore, we are usually interested in the minor arc.

There is a simple, direct, and important correspondence between arcs and central angles. Recall from chapter 9, that two rays with a common endpoint define not one, but two angles, a proper angle and a reflex angle, that are related to the minor and major arcs as described in the next exercise.

Exercise 14.1 Let AB be a chord of a circle $\mathcal{C}$ with center O that is not a diameter. Prove that the points of the minor arc $\widehat{AB}$ are A, B, and the points of $\mathcal{C}$ in the interior of the proper angle $\angle AOB$. Prove that the points of the major arc $\widehat{AB}$ are A, B, and the points of $\mathcal{C}$ in the interior of the reflex angle $\angle AOB$ (that is, the points exterior to the proper angle).

14.2 Intersections

Circles are different from the shapes we have studied up this point because they are not built out of lines or line segments. Circles share at least one characteristic with simple polygons though: they have an interior and an exterior.

Definition 14.6 For a circle $\mathcal{C}$ with center O and radius r, and for a point P that is not on $\mathcal{C}$,

- if $|OP| < r$, then P is inside $\mathcal{C}$,
- if $|OP| > r$, then P is outside $\mathcal{C}$.

The set of points inside the circle is the *interior* of the circle. The set of points outside the circle is its *exterior*.

Just like simple polygons, circles separate their interior and exterior. To get a better sense of that, we need to look at how circles intersect other geometric objects.

Theorem 14.7 *A line will intersect a circle in 0, 1, or 2 points.*

Proof Let O be the center of a circle $\mathcal{C}$ of radius r, and let ℓ be a line. It is easy to find points on ℓ that are far from $\mathcal{C}$, but are there points on ℓ that are close to $\mathcal{C}$? The easiest way to figure out how close ℓ gets to $\mathcal{C}$ is to look at the closest point on ℓ to the center O. We have seen

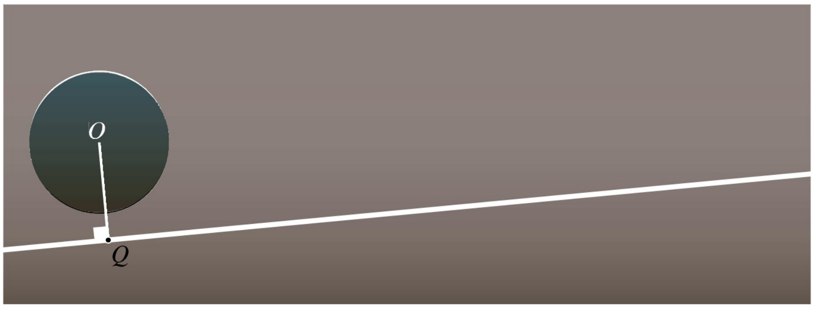

Figure 3. A circle and line that do not intersect.

(lemma 10.3) that the closest point to O on ℓ is the foot of the perpendicular, call it Q. There are three possibilities.

Zero intersections: If $|OQ| > r$, then the other points of ℓ are even farther from O, so none of the points on ℓ can be on $\mathcal{C}$.

One intersection: If $|OQ| = r$, then Q is an intersection, but it is the only intersection because the other points on ℓ are farther away from O.

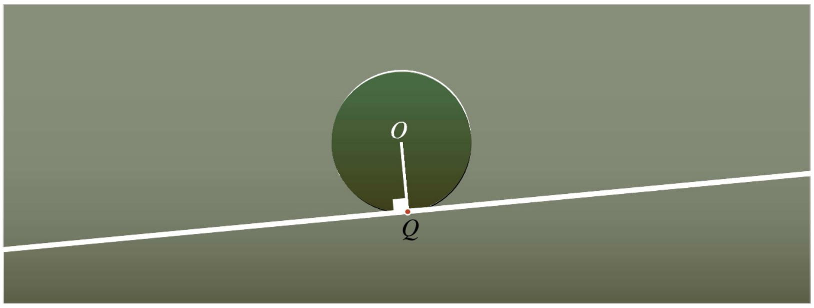

Figure 4. A circle and line that intersect at a single point.

Two intersections: If $|OQ| < r$, then the line lies both inside and outside the circle. We just need to find where the line crosses into, and then back out of, the circle. The idea is to relate a point's distance from O to its distance from Q, and we can do that with the Pythagorean theorem. If P is a point on ℓ other than Q, then $\triangle OQP$ will be a right triangle with side lengths that are related by the Pythagorean theorem:

$$|OQ|^2 + |QP|^2 = |OP|^2.$$

If P is on the circle, $|OP| = r$. Therefore $|PQ| = \sqrt{r^2 - |OQ|^2}$. Since $|OQ| < r$, this expression is a positive real number. There are then exactly two points on ℓ, one on each side of Q, that are this distance from Q.

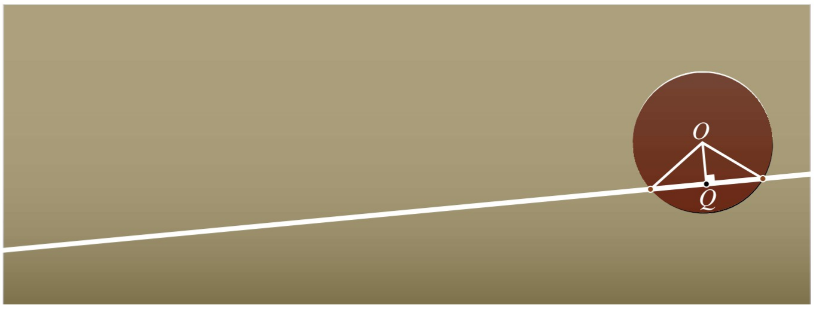

Figure 5. A circle and line that intersect at two points. □

A line that intersects a circle once (at the foot of the perpendicular) is called a *tangent* line to the circle. A line that intersects a circle twice is called a *secant* line of the circle.

There is a important corollary that turns this last theorem about lines into a statement about segments.

Corollary 14.8 *If a point P is inside a circle, and a point Q is outside it, then the segment PQ intersects the circle (exactly once).*

Proof Label the center of the circle O. From the last theorem, we know that $\overline{PQ}$ intersects the circle twice, and that the two intersections are separated by F, the foot of the perpendicular to PQ through O. The important intersection here is the one that is on the same side of the foot of the perpendicular as Q, call it R. According to the Pythagorean theorem (with $\triangle OFQ$ and $\triangle OFR$),

$$|FQ| = \sqrt{|OQ|^2 - |OF|^2} \text{ and } |FR| = \sqrt{|OR|^2 - |OF|^2}.$$

Since $|OQ| > |OR|$, $|FQ| > |FR|$, which places R between F and Q. We cannot say whether P and Q are on the same side of F, though. If they are on opposite sides of F, then $P * F * R * Q$, so R is between P and Q as needed. If P and Q are on the same side of F, then we need to look at the right triangles $\triangle OFP$ and $\triangle OFR$. They tell us that

$$|FP| = \sqrt{|OP|^2 - |OF|^2} \text{ and } |FR| = \sqrt{|OR|^2 - |OF|^2}.$$

Since $|OP| < |OR|$, $|FP| < |FR|$, which places P between F and R. Finally, if $F * P * R$ and $F * R * Q$, then R has to be between P and Q. □

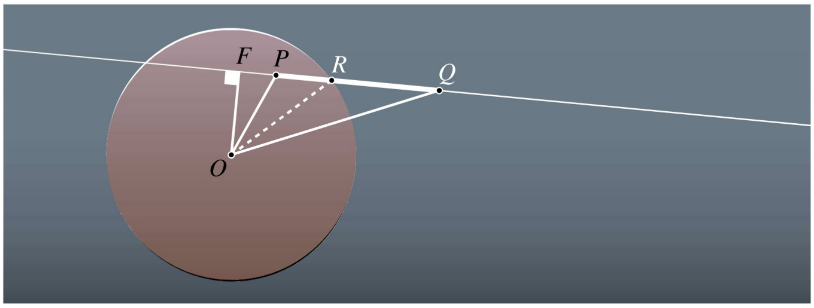

Figure 6. Intersection of a circle and a segment.

There is another important question of intersections: how can two circles intersect?

Theorem 14.9 *Two circles intersect at 0, 1, or 2 points.*

Proof Three factors come in to play here: the radii of the circles and the distance between their centers. Label

O_1, O_2 : the centers of the two circles,
r_1, r_2 : the radii of the two circles, and
c : the distance between the centers.

An intersection point X will have to be a distance r_1 from O_1 and r_2 from O_2. We know that $|O_1O_2| = c$, so the question is: can we build a triangle with sides of lengths r_1, r_2, and c? We studied this question at the end of the last chapter. There are three possibilities.

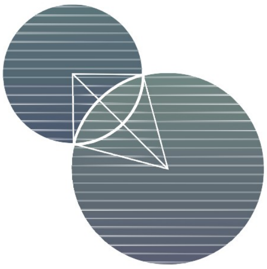

Figure 7. Two intersections.

Two intersections: when $|r_1 - r_2| < c < r_1 + r_2$.
In this case, the proposed side lengths satisfy the triangle inequality conditions. Therefore, there are exactly two triangles, $\triangle O_1XO_2$ and $\triangle O_1YO_2$, one on each side of O_1O_2, with sides of the required lengths. There are then exactly two intersections of the circles.

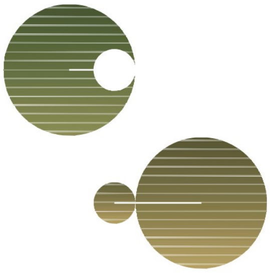

Figure 8. One intersection.

One intersection: when $c = |r_1 - r_2|$ *or* $c = r_1 + r_2$.
In these limiting cases, the triangle devolves into a line segment and the two intersections merge. In the first, either $O_1 * O_2 * X$ or $X * O_1 * O_2$, depending on which radius is larger. In the second, $O_1 * X * O_2$.

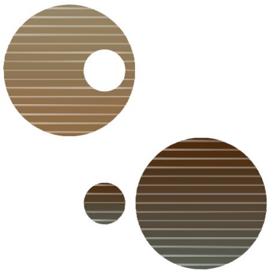

Figure 9. No intersections.

Zero intersections: when $c < |r_1 - r_2|$ *or* $c > r_1 + r_2$.
In this case, the proposed side lengths do not satisfy the triangle inequality conditions. No triangle can be formed with sides of these lengths, so the two circles cannot intersect. In the first case, one circles lies entirely inside the other. In the second, they are separated from one another.

□

As you established in exercise 14.1, there is a correspondence between central angles and arcs. In the next chapter we will establish a metrical connection between the measure of the central angle and the length of the corresponding arc (which is the basis for radian measure). In the meantime, the correspondence is a helpful way to simplify illustrations. Using an arc

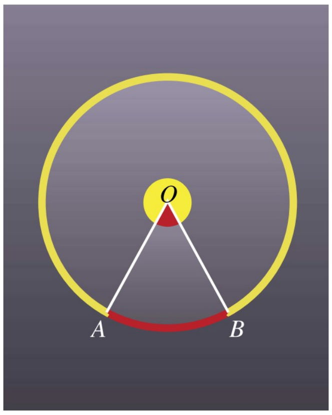

Figure 10. The major arc corresponds to the reflex angle $\angle AOB$. The minor arc corresponds to the proper angle $\angle AOB$.

to indicate a central angle keeps a picture from getting too crowded around the center of the circle.

14.3 The inscribed angle theorem

In this section we will prove the inscribed angle theorem, a result that is indispensable when working with circles. I suspect that this theorem is the most elementary result of Euclidean geometry that is generally not known to the average calculus student. Before stating the theorem, we must define an inscribed angle, the subject of the theorem.

Definition 14.10 If A, B, and C are points on a circle, then $\angle ABC$ is an *inscribed angle* on the circle.

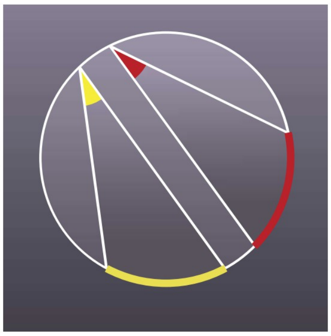

Figure 11. Two angles inscribed in a circle.

Given an inscribed angle $\angle ABC$, either proper or reflex, the points A and C are the endpoints of two arcs, either a minor and a major arc or two semicircles. Excluding the endpoints, one of the arcs will be contained in the interior of $\angle ABC$. We say, then, that $\angle ABC$ is inscribed on the arc $\overset{\frown}{AC}$. The inscribed angle theorem describes the relationship between an inscribed angle and the central angle on the same arc.

Theorem 14.11 *(The inscribed angle theorem) If $\angle ABC$ is an inscribed angle on a circle with center O, then*

$$(\angle ABC) = \frac{1}{2}(\angle AOC).$$

The proof of the inscribed angle theorem is a good demonstration of the benefits of starting with an easy case. There are three cases in the proof, depending on the location of the vertex B relative to the lines $\overline{OA}$ and $\overline{OC}$. The first case establishes the theorem for two particular locations of B, where things work out easily. But the special cases are the key that unlocks everything else.

Proof *Case 1. Where B is the intersection of $\overrightarrow{OC}^{op}$ with the circle, or where B is the intersection of $\overrightarrow{OA}^{op}$ with the circle.* We will prove the result where B is on $\overrightarrow{OC}^{op}$. To deal with the other possibility, where B is on $\overrightarrow{OA}^{op}$, just switch the letters A and C. Label $\angle AOB$ as $\angle 1$ and $\angle AOC$ as $\angle 2$. These angles are supplementary, so

$$(i) \quad (\angle 1) + (\angle 2) = 180^\circ.$$

The angle sum of $\triangle AOB$ is 180°, but in that triangle $\angle A$ and $\angle B$ are opposite congruent segments (radii), so by the isosceles triangle theorem they are congruent. Therefore

$$(ii) \quad 2(\angle B) + (\angle 1) = 180^\circ,$$

and if we subtract (ii) from (i), we get $(\angle 2) - 2(\angle B) = 0$, so $(\angle AOC) = 2(\angle ABC)$.

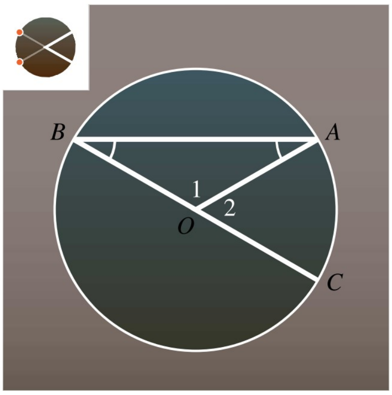

Figure 12. The inscribed angle theorem, case 1: an inscribed angle with one ray along a diameter.

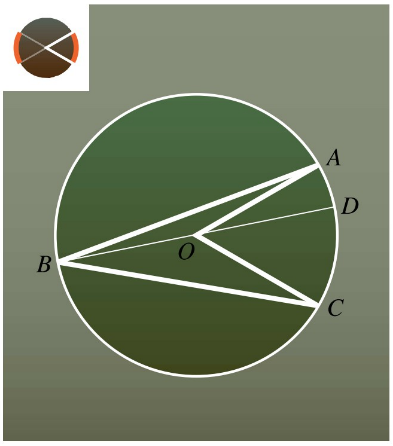

Figure 13. The inscribed angle theorem, case 2.

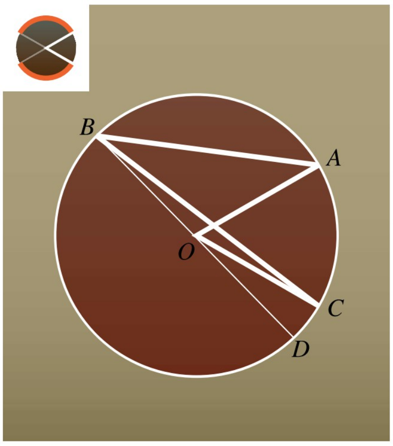

Figure 14. The inscribed angle theorem, case 3.

Case 2. Where B is in the interior of $\angle AOC$*, or where B is in the interior of the angle formed by* $\overrightarrow{OA}^{op}$ *and* $\overrightarrow{OC}^{op}$*, or when* $A * O * C$. There are three cases here—in the first the central angle is reflex, in the second it is proper, and in the third it is a straight angle—but the proof is the same for all of them. In each case, $\overline{OB}$ splits both the inscribed and the central angles. To identify the four angles, label one more point: D is the second intersection of $\overline{OB}$ with the circle (so BD is a diameter of the circle). Using angle addition in conjunction with the previous results,

$$\begin{aligned}(\angle AOC) &= (\angle AOD) + (\angle DOC)\\ &= 2(\angle ABD) + 2(\angle DBC)\\ &= 2((\angle ABD) + (\angle DBC))\\ &= 2(\angle ABC).\end{aligned}$$

Case 3. Where B is in the interior of the angle formed by $\overrightarrow{OA}$ *and* $\overrightarrow{OC}^{op}$*, or where B is in the interior of the angle formed by* $\overrightarrow{OC}$ *and* $\overrightarrow{OA}^{op}$. As in the last case, label D so that BD is a diameter. The difference this time is that we need to use angle subtraction instead of angle addition. Since subtraction is less symmetric than addition, the two cases will differ slightly (in terms of lettering). In the first,

$$\begin{aligned}(\angle AOC) &= (\angle AOD) - (\angle DOC)\\ &= 2(\angle ABD) - 2(\angle DBC)\\ &= 2((\angle ABD) - (\angle DBC))\\ &= 2(\angle ABC).\end{aligned}$$

To get the second, just switch A and C. □

There are two important and immediate corollaries to this theorem. First, because all inscribed angles on a given arc share the same central angle,

Corollary 14.12 *All inscribed angles on a given arc are congruent.*

Second, the special case where the central angle $\angle AOC$ is a straight angle, so that the inscribed angle $\angle ABC$ is a right angle, is important enough to earn its own name.

Corollary 14.13 *(Thales' theorem) If C is a point on a circle with diameter AB (and C is neither A nor B), then $\triangle ABC$ is a right triangle.*

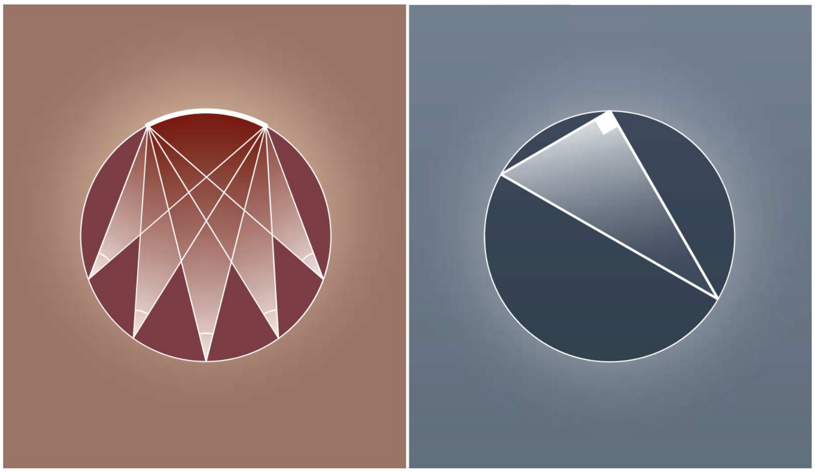

Figure 15. (l) Five congruent angles inscribed on the same arc. (r) A right angle inscribed on a semicircle.

Thales' theorem is often presented as an if-and-only-if statement, and both directions are quite useful.

Exercise 14.2 Prove that if $\triangle ABC$ is a right triangle with right angle at C, then C is a point on the circle with diameter AB.

14.4 Applications of the inscribed angle theorem

Using the inscribed angle theorem, we can establish several nice relationships between chords, secants, and tangents associated with a circle. We will look at two of these results to end this chapter.

Theorem 14.14 *(The chord-chord formula) Let $\mathscr{C}$ be a circle with center O. Suppose that AC and BD are chords of $\mathscr{C}$, and suppose that they intersect at a point P. Label the angle of intersection, $\theta = \angle APD \simeq \angle BPC$. Then*

$$(\theta) = \frac{(\angle AOD) + (\angle BOC)}{2}.$$

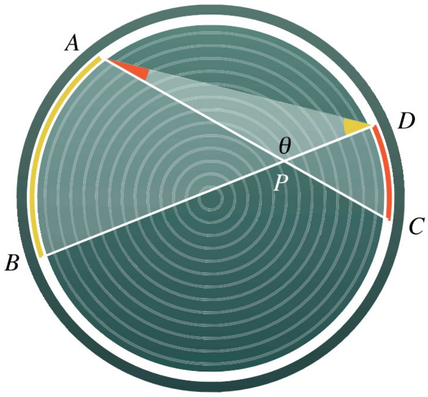

Figure 16. Two intersecting chords in a circle.

Proof The angle θ is an interior angle of $\triangle APD$, so

$$(\theta) = 180^\circ - (\angle A) - (\angle D).$$

$\angle A$ and $\angle D$ are inscribed angles: $\angle A$ is inscribed on the arc $\overset{\frown}{CD}$ and $\angle D$ is inscribed on the arc $\overset{\frown}{AB}$. According to the inscribed angle theorem, they are half the size of the corresponding central angles, so

$$\begin{aligned}(\theta) &= 180^\circ - \tfrac{1}{2}(\angle COD) - \tfrac{1}{2}(\angle AOB)\\ &= \tfrac{1}{2}(360^\circ - (\angle COD) - (\angle AOB)).\end{aligned}$$

This relates θ to central angles, but alas, not to the central angles in the formula. If we add all four central angles around O, though,

$$(\angle AOB) + (\angle BOC) + (\angle COD) + (\angle DOA) = 360^\circ$$

and so

$$(\angle BOC) + (\angle DOA) = 360^\circ - (\angle COD) - (\angle AOB).$$

Now substitute to get the desired formula:

$$(\theta) = \tfrac{1}{2}((\angle BOC) + (\angle DOA)). \qquad \square$$

According to the chord-chord formula, as long as the intersection point P is inside the circle, θ can be computed as the average of two central angles. What would happen if P moved outside the circle? Then we would not be talking about chords, since chords stop at the circle boundary, but about the secant lines containing them.

Theorem 14.15 *(The secant-secant formula) Suppose that A, B, C, and D are points on a circle arranged so that $\square ABCD$ is a simple quadrilateral, and that the secant lines AB and CD intersect at a point P that is outside the circle. Label the angle of intersection, $\angle APD$, as θ. Let $\angle AOD$ and $\angle BOC$ be taken so that P is in their interiors (that is, the angles may be reflex). If P occurs on the same side of AD as B and C, then*

$$(\theta) = \frac{(\angle AOD) - (\angle BOC)}{2}.$$

If P occurs on the same side of BC as A and D, then

$$(\theta) = \frac{(\angle BOC) - (\angle AOD)}{2}.$$

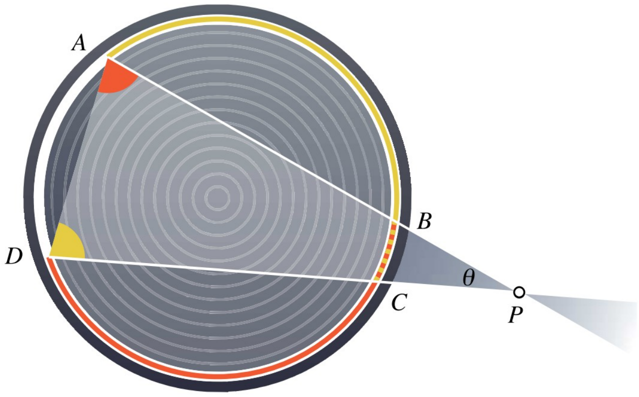

Figure 17. Two lines that intersect outside of the circle.

Proof There is obviously a great deal of symmetry between the two cases, so let me just address the first. The same principles apply here as in the last proof. Angle θ is an interior angle of $\triangle APD$, so

$$(\theta) = 180^\circ - (\angle A) - (\angle D).$$

$\angle A$ and $\angle D$ are inscribed angles— $\angle A$ is inscribed on arc $\widehat{BD}$ and $\angle D$ is inscribed on arc $\widehat{AC}$. We need to use the inscribed angle theorem to relate the angles to central angles, and in this case, the central angles overlap, so we will need to break them down further, but the rest is straightforward.

$$\begin{aligned}(\theta) &= 180^\circ - \tfrac{1}{2}(\angle BOD) - \tfrac{1}{2}(\angle AOC)\\ &= \tfrac{1}{2}[360^\circ - (\angle BOD) - (\angle AOC)]\\ &= \tfrac{1}{2}[360^\circ - (\angle BOC) - (\angle COD) - (\angle AOB) - (\angle BOC)]\\ &= \tfrac{1}{2}[[360^\circ - (\angle AOB) - (\angle BOC) - (\angle COD)] - (\angle BOC)]\\ &= \tfrac{1}{2}[(\angle AOD) - (\angle BOC)].\end{aligned}$$

□

14.5 Exercises

14.3. Prove that no line is entirely contained in any circle.

14.4. Prove that a circle is convex. That is, prove that if points P and Q are inside a circle, then all the points on the segment PQ are inside the circle.

14.5. Prove that for any circle there is a triangle entirely contained in it, so all the points of the triangle are inside the circle.

14.6. Prove that for any circle there is a triangle which entirely contains it, so all the points of the circle are in the interior of the triangle.

14.7. Let P be a point on a line ℓ, and let Q be a point that is not on ℓ. Prove that there is a unique circle $\mathcal{C}$ that passes through Q and is tangent to ℓ at P.

14.8. Suppose that $\square ABCD$ is cyclic, so there is a circle through all four of its vertices. Prove that $(\angle A) + (\angle C) = 180^\circ$ and $(\angle B) + (\angle D) = 180^\circ$.

14.9. Given a circle $\mathcal{C}$ with center O and points A and C on $\mathcal{C}$, prove that:

- if a point B is inside $\mathcal{C}$, then $(\angle ABC) > \frac{1}{2}(\angle AOC)$, and
- if a point B is outside $\mathcal{C}$, then $(\angle ABC) < \frac{1}{2}(\angle AOC)$.

Note that you will have to think about when $\angle ABC$ and $\angle AOC$ should be taken as the reflex angle.

14.10. Let $\angle ABC$ be an inscribed angle on a circle. Prove that, excluding the endpoints, exactly one of the two arcs $\overset{\frown}{AC}$ lies in the interior of $\angle ABC$.

14.11. Let $\mathcal{C}$ be a circle and P be a point outside of it. Prove that there are exactly two lines that pass through P and are tangent to $\mathcal{C}$. Let Q and R be the points of tangency for the two lines. Prove that PQ and PR are congruent.

14.12. Let A be a point outside of a circle $\mathcal{C}$ with center O. Let B be the point on $\mathcal{C}$ so that $A * O * B$. There are two lines through A that are tangent to $\mathcal{C}$. Label their intersections with $\mathcal{C}$ as P and Q. What is the relationship between the angle measures $(\angle PAQ)$ and $(\angle PBQ)$?

14.13. Let $\mathcal{P}$ be a regular polygon. Prove that there is a unique circle $\mathcal{C}$ contained in $\mathcal{P}$ so that $\mathcal{C}$ is tangent to each of the sides of $\mathcal{P}$. This is called the circle *inscribed* in $\mathcal{P}$.

14.14. Suppose that a parallelogram is inscribed in a circle. Prove that the parallelogram must be a rectangle.

14.15. Let $\square ABCD$ be a trapezoid with exactly one set of parallel sides that is inscribed in a circle. Prove that the non-parallel sides of $\square ABCD$ must be congruent.

14.16. Suppose that circles $\mathcal{C}_1$ with center O_1 and $\mathcal{C}_2$ with center O_2 intersect at two points P and Q. Prove that $\angle O_1PO_2 \simeq \angle O_1QO_2$.

14.17. Let $\triangle ABC$ be equilateral, with sides of length a. Let $\mathcal{C}_A$, $\mathcal{C}_B$, and $\mathcal{C}_C$ be three circles, each with radius $a/2$, centered at A, B, and C. The three circles are tangent to one another and determine a curved but approximately triangular region in the center of $\triangle ABC$. Show that a circle can be inscribed in the region, tangent to $\mathcal{C}_A$, $\mathcal{C}_B$, and $\mathcal{C}_C$. What is the radius of this circle? Hint: see exercise 13.14.

14.18. The "tangent-tangent" formula. Let P be a point that is outside of a circle $\mathscr{C}$. Consider the two tangent lines to $\mathscr{C}$ that pass through P and let A and B be the points of tangency between those lines and the circle. Prove that

$$(\angle APB) = \frac{(\angle 1) - (\angle 2)}{2}$$

where $\angle 1$ is the reflex central angle corresponding to the major arc $\widehat{AB}$ and $\angle 2$ is the proper central angle corresponding to the minor arc $\widehat{AB}$.

14.19. Let AC and BD be two chords of a circle that intersect at a point P inside the circle. Prove that

$$|AP| \cdot |CP| = |BP| \cdot |DP|.$$

14.6 References

I learned of the chord-chord, secant-secant, and tangent-tangent formulas in the Wallace and West book *Roads to Geometry* [WW04]. They use the names two-chord angle theorem, two-secant angle theorem, and two-tangent angle theorem.

;. :A5 886

15

Circumference

15.1 A theorem on perimeters

In chapter 9 on polygons, we defined the perimeter of a polygon $\mathcal{P} = P_1 \cdots P_n$ as

$$|\mathcal{P}| = \sum_{i=1}^{n} |P_i P_{i+1}|,$$

but then left it at that. In this lesson we are going to use perimeters of cyclic polygons to find the circumference of the circle. Along the way, we will need to use the following result that compares the perimeters of two convex polygons when one is contained in the other. We will prove the theorem by a process of "shaving down" the sides of the larger polygon until we reach the smaller one.

Theorem 15.1 *If $\mathcal{P}$ and $\mathcal{Q}$ are convex polygons and all the points of $\mathcal{P}$ are on or inside $\mathcal{Q}$, then $|\mathcal{P}| \leq |\mathcal{Q}|$.*

Proof Some of the edges of $\mathcal{P}$ may run along the edges of $\mathcal{Q}$, but unless $\mathcal{P} = \mathcal{Q}$, at least one edge of $\mathcal{P}$ must pass through the interior of $\mathcal{Q}$. Let s be one of those interior edges. The line containing s intersects $\mathcal{Q}$ twice—call the intersections a and b—dividing $\mathcal{Q}$ into two smaller polygons that share the side ab, one on the same side of s as $\mathcal{P}$, the other on the opposite side. Define a new polygon $\mathcal{Q}_1$ that consists of

- points of $\mathcal{Q}$ on the same side of s as $\mathcal{P}$, and
- points on the segment ab.

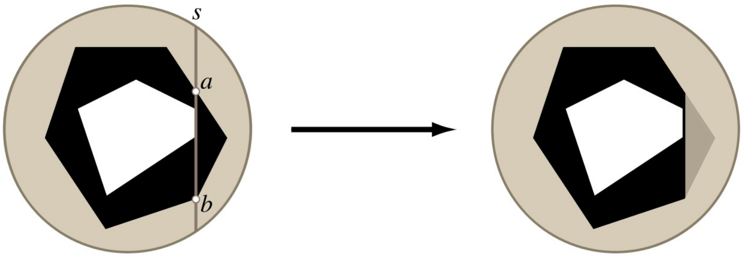

Figure 1. Shaving a corner off of a polygon.

There are two things to notice about $\mathcal{Q}_1$. First, $\mathcal{Q}_1$ and $\mathcal{P}$ have one more coincident side (the side s) than $\mathcal{Q}$ and $\mathcal{P}$ had. Second, the portions of $\mathcal{Q}$ and $\mathcal{Q}_1$ on the side of s with $\mathcal{P}$ are identical, so the segments making up that part contribute the same amount to their perimeters. On the other side, though, the path that $\mathcal{Q}$ takes from a to b is longer than the direct route along the segment ab of $\mathcal{Q}_1$ (because of the triangle inequality). Combining the two parts, we get $|\mathcal{Q}_1| \leq |\mathcal{Q}|$.

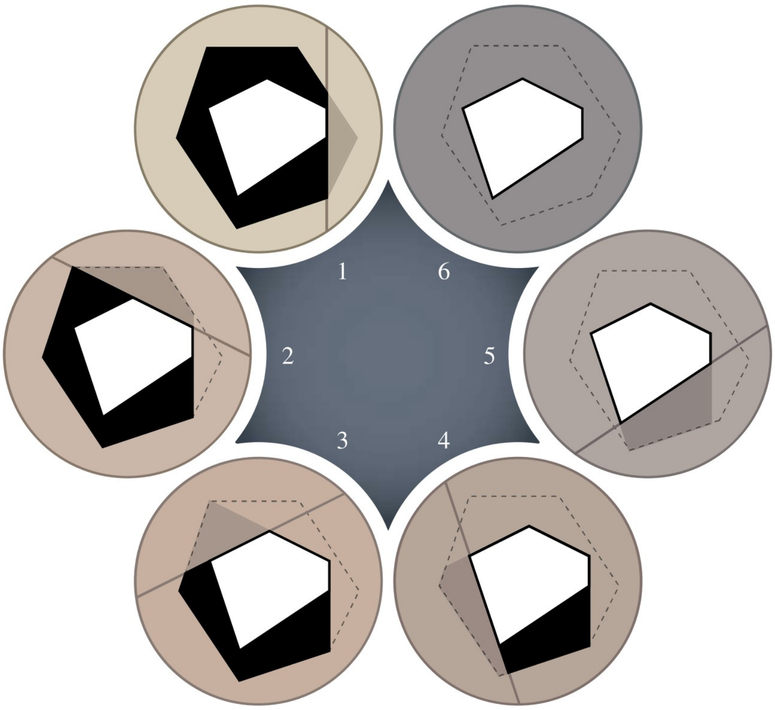

Figure 2. One at a time, shave the edges of the outer polygon down to the inner one.

Now we can repeat this process with $\mathcal{P}$ and $\mathcal{Q}_1$, generating $\mathcal{Q}_2$ with even smaller perimeter than $\mathcal{Q}_1$ and another coincident side with $\mathcal{P}$. And again, to get $\mathcal{Q}_3$. Eventually, though, after say m steps, we run out of sides that pass through the interior, at which point $\mathcal{P} = \mathcal{Q}_m$. Then

$$|\mathcal{P}| = |\mathcal{Q}_m| \leq |\mathcal{Q}_{m-1}| \leq \cdots \leq |\mathcal{Q}_2| \leq |\mathcal{Q}_1| \leq |\mathcal{Q}|. \qquad \square$$

15.2 Circumference

The purpose of this chapter is to understand the connection between the radius of a circle, and the distance around it, its circumference. The final result, the formula $C = 2\pi r$, ranks with the Pythagorean theorem in celebrity, although I think it is an often misunderstood celebrity. Let's be clear about what this equation is not. It is not an equation comparing two known quantities C and $2\pi r$. Instead, it is saying that the ratio $C/(2r)$ is a constant. The constant is what we have come to call π.

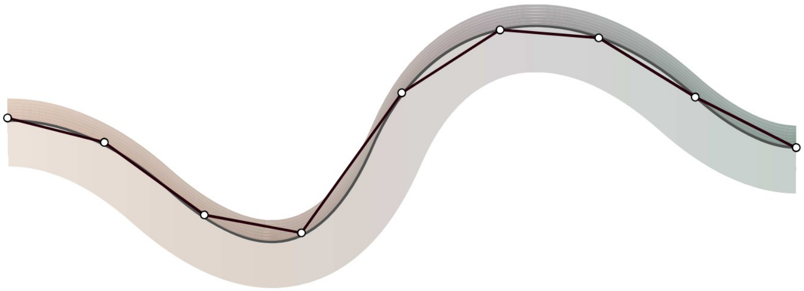

Figure 3. An approximation of an arc by segments.

To define the circumference of a circle, let's borrow an idea from calculus, the idea of approximating a curve by straight line segments, and then refining the approximation by increasing the number of segments. For a circle $\mathcal{C}$, the approximating line segments will be the edges of a simple cyclic polygon $\mathcal{P}$ inscribed in the circle. We will want the circumference of $\mathcal{C}$ to be bigger than the perimeter of $\mathcal{P}$. We should also expect that by adding in additional vertices to $\mathcal{P}$, we should be able to get the perimeter of $\mathcal{P}$ as close as we want to the circumference of $\mathcal{C}$. All this suggests that to get the circumference of $\mathcal{C}$, we need to find out how large the perimeters of inscribed polygons can be.

Definition 15.2 The circumference of a circle $\mathcal{C}$, written $|\mathcal{C}|$, is

$$|\mathcal{C}| = \sup \left\{ |\mathcal{P}| \middle| \mathcal{P} \text{ is a simple cyclic polygon inscribed in } \mathcal{C} \right\}.$$

There is nothing in the definition to guarantee that the supremum exists. It is conceivable that the lengths of the approximating perimeters might grow with bound. An example is the *Koch snowflake*. It is constructed as follows. Take an equilateral triangle with sides of length 1. The perimeter of this triangle is 3. Divide each side into thirds. On each middle third, build an equilateral triangle by adding two more sides then remove the original side. You have made a shape with $3 \cdot 4$ sides, each with length 1/3, for a perimeter of 4. Now iterate: divide the sides into thirds; build equilateral triangles on each middle third, and remove the base. That will make $3 \cdot 16$ sides of length 1/9, for a perimeter of 16/3. After the next step, there are $3 \cdot 64$ sides of length 1/27 for a perimeter of 64/9. Generally, after n iterations, there are $3 \cdot 4^n$ sides of length $1/3^n$ for a total perimeter of $4^n/3^{n-1}$, and

$$\lim_{n\to\infty} \frac{4^n}{3^{n-1}} = \lim_{n\to\infty} 3 \left(\frac{4}{3}\right)^n = \infty.$$

The Koch snowflake, which is the limiting shape in this process, has infinite perimeter! The first thing we need to do, then, is to make sure that circles are better behaved than this.

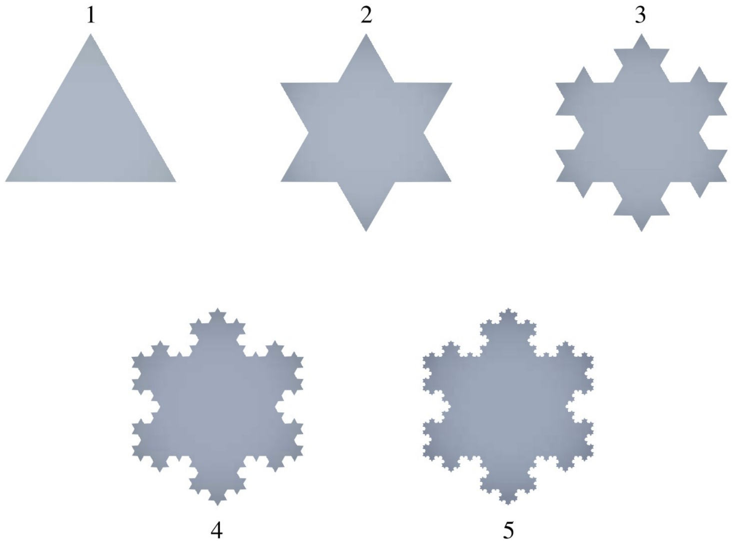

Figure 4. The first few steps in the construction of the Koch snowflake.

Theorem 15.3 *If* $\mathcal{C}$ *is a circle of radius* r*, then* $|\mathcal{C}| \leq 8r$.

Proof The first step is to build a circumscribing square around $\mathcal{C}$, the smallest square that still contains $\mathcal{C}$. Begin by choosing two perpendicular diameters d_1 and d_2. Each will intersect $\mathcal{C}$ twice, for a total of four intersections, P_1, P_2, P_3, and P_4. For each i between 1 and 4, let t_i be the tangent line to $\mathcal{C}$ at P_i. The tangents intersect to form the circumscribing square. The length of each side of the square is equal to the diameter of $\mathcal{C}$, so the perimeter of the square is $4 \cdot 2r = 8r$.

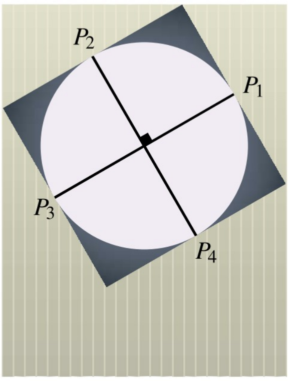

Figure 5. The circumscribing square.

Now we turn to the theorem we proved to start this chapter. Each simple cyclic polygon inscribed in $\mathcal{C}$ is a convex polygon contained in the circumscribing square. Therefore the perimeter of any such approximating polygon is bounded above by $8r$. Remember that we have defined $|\mathcal{C}|$ to be the supremum of all the approximating perimeters, so it cannot exceed $8r$ either. □

Now that we know that a circle has a circumference, the next step is to find a way to calculate it. The key to that is the next theorem.

Theorem 15.4 *The ratio of the circumference of a circle to its radius is a constant.*

Proof Let's suppose that the ratio is not a constant, so that there are two circles $\mathcal{C}_1$ and $\mathcal{C}_2$ with centers O_1 and O_2 and radii r_1 and r_2, but with unequal ratios

$$|\mathcal{C}_1|/r_1 > |\mathcal{C}_2|/r_2.$$

As we have defined circumference, there are approximating cyclic polygons to $\mathcal{C}_1$ whose perimeters are arbitrarily close to its circumference. In particular, there is an approximating cyclic polygon $\mathcal{P} = P_1P_2\ldots P_n$ for $\mathcal{C}_1$ so that

$$|\mathcal{P}|/r_1 > |\mathcal{C}_2|/r_2.$$

We can construct a cyclic polygon $\mathcal{Q}$ on $\mathcal{C}_2$ that is similar to $\mathcal{P}$. Intuitively, we just need to scale $\mathcal{P}$ so that it fits in the circle $\mathcal{C}_2$. The formal construction is:

- Begin by placing a point Q_1 on circle $\mathcal{C}_2$.
- Locate Q_2 on $\mathcal{C}_2$ so that $\angle P_1O_1P_2$ is congruent to $\angle Q_1O_2Q_2$ (there are two choices for Q_2).
- Locate Q_3 on $\mathcal{C}_2$ and on the opposite side of O_2Q_2 from Q_1 so that $\angle P_2O_1P_3$ is congruent to $\angle Q_2O_2Q_3$.
- Continue placing points on $\mathcal{C}_2$ in this fashion until Q_n has been placed to form the polygon $\mathcal{Q} = Q_1Q_2\ldots Q_n$. At the end of the process $\angle P_nO_1P_1$ is congruent to $\angle Q_nO_2Q_1$ by angle subtraction.

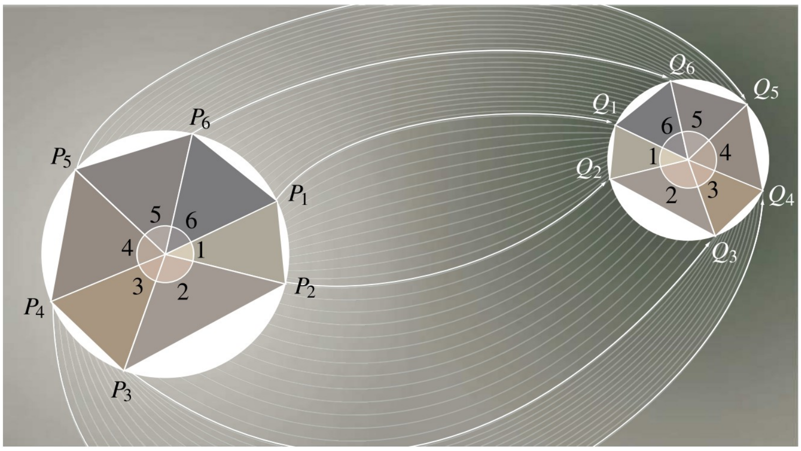

Figure 6. The ratio $|Q_iQ_{i+1}|/|P_iP_{i+1}|$ is a constant.

Then

$$\frac{|O_2Q_i|}{|O_1P_i|} = \frac{r_2}{r_1} = \frac{|O_2Q_{i+1}|}{|O_1P_{i+1}|} \text{ and } \angle Q_iO_2Q_{i+1} \simeq \angle P_iO_1P_{i+1},$$

so by SAS similarity, $\triangle Q_iO_2Q_{i+1}$ is similar to $\triangle P_iO_1P_{i+1}$. That gives us the ratio of the third sides of the triangle as

$$\frac{|Q_iQ_{i+1}|}{|P_iP_{i+1}|} = \frac{r_2}{r_1}.$$

Thus we can describe the perimeter of $\mathcal{Q}$ as

$$|\mathcal{Q}| = \sum_{i=1}^{n} |Q_iQ_{i+1}| = \sum_{i=1}^{n} \frac{r_2}{r_1}|P_iP_{i+1}| = \frac{r_2}{r_1}\sum_{i=1}^{n} |P_iP_{i+1}| = \frac{r_2}{r_1}|\mathcal{P}|.$$

That would mean that

$$\frac{|\mathcal{Q}|}{r_2} = \frac{|\mathcal{P}|}{r_1} > \frac{|\mathcal{C}_2|}{r_2}.$$

That is, $|\mathcal{Q}| > |\mathcal{C}_2|$, when in fact the circumference of $\mathcal{C}_2$ is supposed to be greater than the perimeter of any of its approximating cyclic polygons. □

Definition 15.5 The constant π is the ratio of the circumference of a circle to its diameter: $\pi = |\mathcal{C}|/(2r)$.

It is more convenient to express this definition in terms of a limit rather than a supremum. Fortunately, the supremum can be reached as a limit of perimeters of a sequence of regular polygons as follows. Arrange n angles each measuring $360^\circ/n$ around the center of a circle $\mathcal{C}$. The rays of the angles intersect $\mathcal{C}$ n times, and the points P_i are the vertices of a regular n-gon, $\mathcal{P}_n = P_1P_2 \ldots P_n$. The tangent lines to $\mathcal{C}$ at the neighboring points P_i and P_{i+1} intersect at a point Q_i. Taken together, the n points are the vertices of another regular n-gon $\mathcal{Q}_n = Q_1Q_2 \ldots Q_n$. The polygon $\mathcal{P}_n$ is just one of the many cyclic polygons inscribed in $\mathcal{C}$ so $|\mathcal{P}_n| \leq |\mathcal{C}|$. The polygon $\mathcal{Q}_n$ circumscribes $\mathcal{C}$, and every cyclic polygon inscribed on $\mathcal{C}$ lies inside $\mathcal{Q}_n$, so $|\mathcal{Q}_n| \geq |\mathcal{C}|$.

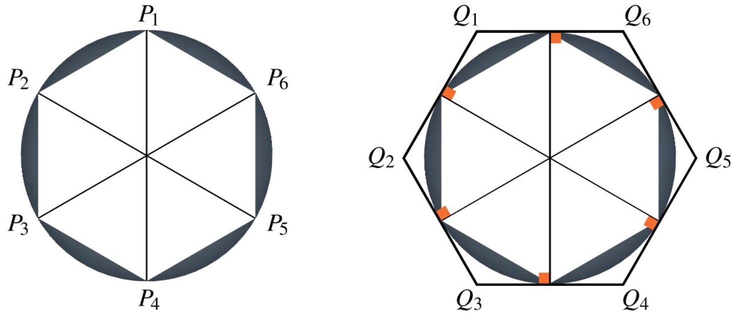

Figure 7. Regular inscribed (l) and circumscribed (r) hexagons.

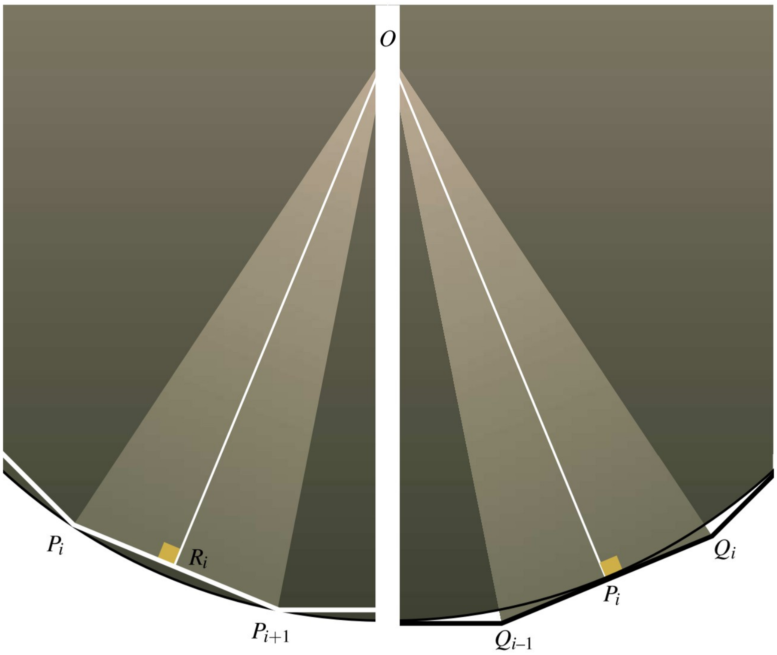

Figure 8. Approximation by inscribed and circumscribed polgons.

The lower bound prescribed by $\mathcal{P}_n$. $\overrightarrow{OQ_i}$ is a perpendicular bisector of P_iP_{i+1}, intersecting it at a point R_i and dividing $\triangle OP_iP_{i+1}$ in two. By the HL congruence theorem for right triangles, the two parts, $\triangle OR_iP_i$ and $\triangle OR_iP_{i+1}$, are congruent. Then $\mathcal{P}_n$ is built from $2n$ segments of length $|P_iR_i|$. Now for a little trigonometry:

$$\sin(360^\circ/2n) = \frac{|P_iR_i|}{r} \implies |P_iR_i| = r\sin(360^\circ/2n)$$

so $|\mathcal{P}_n| = 2nr\sin(360^\circ/2n)$.

The upper bound prescribed by $\mathcal{Q}_n$. $\overrightarrow{OP_i}$ is a perpendicular bisector of $Q_{i-1}Q_i$, intersecting it at P_i and dividing $\triangle OQ_{i-1}Q_i$ in two. By SAS, the two parts, $\triangle OP_iQ_{i-1}$ and $\triangle OP_iQ_i$, are congruent. Then $\mathcal{Q}_n$ is built from $2n$ segments of length $|P_iQ_i|$. Because

$$\tan(360^\circ/2n) = |P_iQ_i|/r \implies |P_iQ_i| = r\tan(360^\circ/2n),$$

$|\mathcal{Q}_n| = 2nr\tan(360^\circ/2n)$.

Let's compare $|\mathcal{P}_n|$ and $|\mathcal{Q}_n|$ as n increases (the key to the calculation is that as x approaches 0, $\cos(x)$ approaches 1):

$$\begin{aligned}
\lim_{n\to\infty}\frac{|\mathcal{P}_n|}{|\mathcal{Q}_n|} &= \lim_{n\to\infty}\frac{2nr\tan(360^\circ/2n)}{2nr\sin(360^\circ/2n)}\\
&= \lim_{n\to\infty}\frac{\sin(360^\circ/2n)}{\cos(360^\circ/2n)}\cdot\frac{1}{\sin(360^\circ/2n)}\\
&= \lim_{n\to\infty}\frac{1}{\cos(360^\circ/2n)}\\
&= 1.
\end{aligned}$$

Therefore $\lim_{n\to\infty} |\mathcal{P}_n| = \lim_{n\to\infty} |\mathcal{Q}_n|$. Since $|\mathcal{C}|$ is trapped between $|\mathcal{P}_n|$ and $|\mathcal{Q}_n|$ for all n, and since they approach the same number as n goes to infinity, $|\mathcal{C}|$ also approaches this number. That gives a more comfortable equation for circumference as

$$|\mathcal{C}| = \lim_{n\to\infty} 2nr \sin(360^\circ/2n),$$

and since $|\mathcal{C}| = 2\pi r$, we can make a definition of π as

$$\pi = \lim_{n\to\infty} n \sin(360^\circ/2n).$$

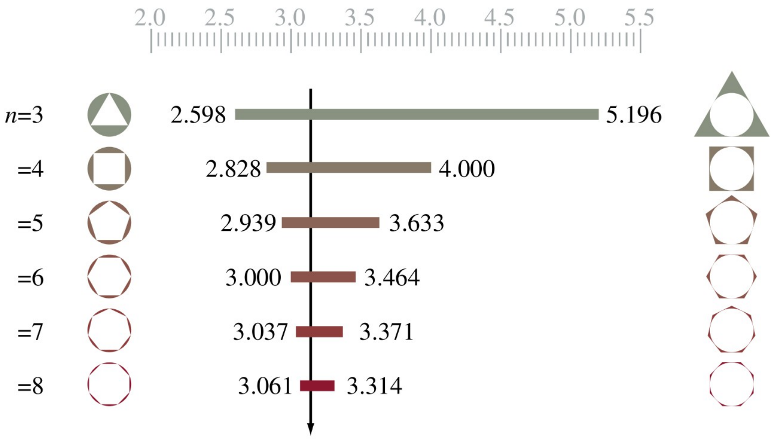

Figure 9. Using perimeters of regular inscribed and circumscribed polygons to determine upper and lower bounds for π.

15.3 Lengths of arcs and radians

It doesn't take much modification to get a formula for the length of an arc. The 360° in the formula for $|\mathcal{C}|$ is the measure of the central angle corresponding to an arc that goes completely around the circle. To get the measure of any other arc, we can replace the 360° with the measure of the corresponding central angle.

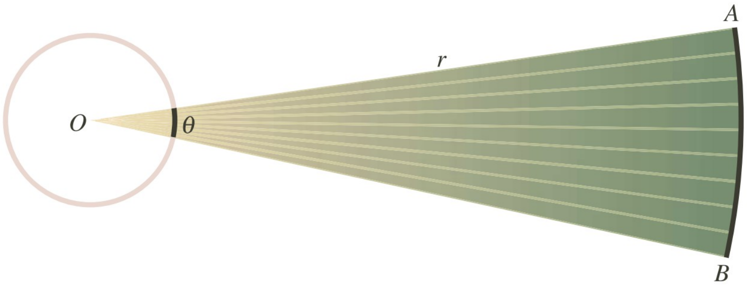

Figure 10. An arc.

Theorem 15.6 *If $\widehat{AB}$ is the arc of a circle with radius r, and if θ is the measure of the central angle $\angle AOB$, then*

$$|\widehat{AB}| = \frac{\pi}{180^\circ}\theta \cdot r.$$

Proof To start, replace the 360° in the circumference formula with θ:

$$|\widehat{AB}| = \lim_{n\to\infty} 2nr \sin(\theta/2n) = 2r \cdot \lim_{n\to\infty} n \sin(\theta/2n).$$

This limit is clearly related to the one that defines π. We can absorb the difference between the two into the variable with the substitution $n = m \cdot \theta/360^\circ$. As n approaches infinity, m will as well, so

$$\begin{aligned} |\widehat{AB}| &= 2r \cdot \lim_{m\to\infty} \frac{m\cdot\theta}{360^\circ} \sin\left(\frac{\theta}{2m\cdot\theta/360^\circ}\right) \\ &= \frac{2r\theta}{360^\circ} \cdot \lim_{m\to\infty} m \sin(360^\circ/2m) \\ &= \frac{\theta}{180^\circ} r\pi. \end{aligned}$$

□

There is one more thing to notice before the end of this chapter. The arc length formula provides a direct connection between angle measure (of the central angle) and distance (along the arc). And yet, the $\frac{\pi}{180^\circ}$ factor in that formula suggests that distance and the degree measurement system are a little out of sync with one another. This can be fixed by modernizing our method of angle measurement. The preferred angle measurement system, and the one that we will use, is *radian* measurement.

Definition 15.7 One radian is $180^\circ/\pi$.

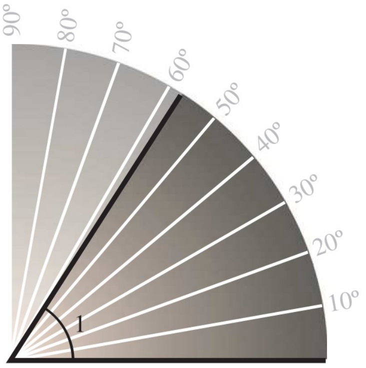

Figure 11. One radian is approximately 57.296°.

Then the measure of a straight angle is π radians. The measure of a right angle is $\pi/2$ radians. One complete turn of the circle is 2π radians. If $\theta = (\angle AOB)$ is measured in radians, then $|\widehat{AB}| = r \cdot \theta$.

15.4 Exercises

15.1. Let A and B be points on a circle $\mathcal{C}$ with radius r. Let θ be the measure of the central angle corresponding to the minor arc (or semicircle) $\widehat{AB}$. What is the relationship (in the form of an equation) between θ, r, and $|AB|$?

15.2. Let A and B be points on a circle $\mathcal{C}$ with center O and radius r. If $(\angle BAO) = \theta$, what is $|\widehat{AB}|$ (in terms of r and θ)?

15.3. Let AB be a diameter of a circle $\mathcal{C}$, and let P be a point on AB. Let $\mathcal{C}_1$ be the circle with diameter AP and let $\mathcal{C}_2$ be the circle with diameter BP. Show that the sum of the circumferences of $\mathcal{C}_1$ and $\mathcal{C}_2$ is equal to the circumference of $\mathcal{C}$ (the shape formed by the three semicircles on one side of AB is called an *arbelos*).

15.4. What is the radius of the circle that passes through the three vertices of a equilateral triangle with sides of length 1?

15.5. Let $\mathcal{C}_1$ be the circle inscribed in an equilateral triangle $\triangle ABC$ with sides of length 1. There is a line that is parallel to AB and tangent to $\mathcal{C}_1$ that divides $\triangle ABC$ into a trapezoid and a smaller equilateral triangle. Let $\mathcal{C}_2$ be the circle inscribed in the smaller equilateral triangle. Continuing in this way, we can form a stack of smaller and smaller circles. Is the sum of their circumferences finite, and if so, what is it?

15.6. In the construction of the Koch snowflake, the middle third of each segment is replaced with two-thirds of an equilateral triangle. Suppose, instead, that middle third was replaced with three of the four sides of a square. What is the perimeter of the n-th stage of this operation? Would the limiting perimeter still be infinite?

15.7. This problem deals with the possibility of angle measurement systems other than degrees or radians. Let $\mathscr{A}$ be the set of angles in the plane. Consider a function

$$\star : \mathscr{A} \to (0, \infty) : \angle A \to (\angle A)^\star$$

that satisfies

- if $\angle A \simeq \angle B$, then $(\angle A)^\star = (\angle B)^\star$
- if D is in the interior of $\angle ABC$, then

$$(\angle ABC)^\star = (\angle ABD)^\star + (\angle DBC)^\star.$$

Prove that the $\star$ measurement system is a constant multiple of the degree measurement system (or, for that matter, the radian measurement system). That is, prove that there is a $k > 0$ such that for all $\angle A \in \mathscr{A}$,

$$(\angle A)^\star = k \cdot (\angle A).$$

15.5 References

The Koch snowflake is an example of a fractal. Gerald Edgar's book *Measure, Topology, and Fractal Geometry* [Edg90], deals with these objects and their measures.

;.   :A5 886   )

23

Reflections

This chapter introduces the first type of isometry, reflection across a line. As it turns out, reflections are the building blocks for all isometries. In this chapter we will see why, in the three reflections theorem. This theorem provides the strategy that we will use over the next few chapters to classify all isometries. First, we must define a reflection.

Definition 23.1 Given a line ℓ, the *reflection* across ℓ is the mapping $s : \mathbb{E} \to \mathbb{E}$ so that

- if P is on ℓ, then $s(P) = P$, and
- if P is not on ℓ, then $s(P)$ is the unique point on the line through P perpendicular to ℓ that is the same distance from ℓ as P, but that is on the opposite side of ℓ from P.

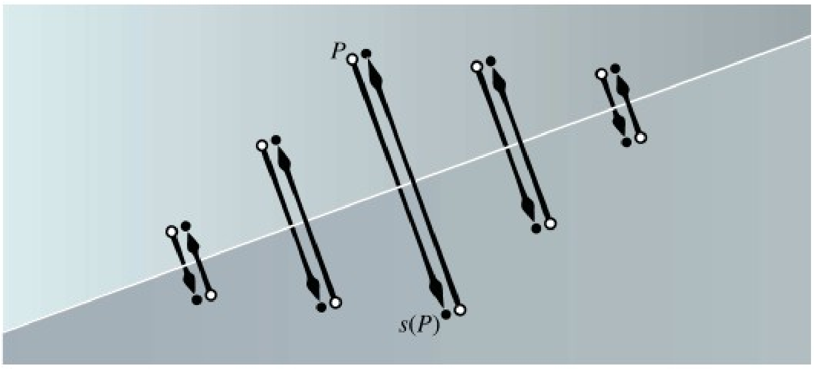

Figure 1. Reflecting points across a line.

We first verify that a reflection is an isometry.

Theorem 23.2 *A reflection is an isometry.*

Proof It is easy to see that a reflection s is a bijection. Look at the composition $s \circ s$: the swap of points done by the first application of s is immediately undone by the second application of s, so that $s^2 = id$. Therefore s is its own inverse, and for a mapping to have an inverse, it must be a bijection.

-7 X* H*3 & Y 4 *1 JZ[51 J & . ,6 X* *7\

The other step is to show that s preserves distances, that $|s(PQ)| = |PQ|$ for any points P and Q. The only thing that makes this part difficult is that there are many possible positions of P and Q relative to each other and to ℓ, the line of reflection:

- **I:** P and Q are both on ℓ.
- **II:** One of P and Q is on ℓ, while the other is not.
  - [1] $\overline{PQ}$ is perpendicular to ℓ
  - [2] $\overline{PQ}$ is not perpendicular to ℓ
- **III:** Neither P nor Q is on ℓ.
  - [1] $\overline{PQ}$ is perpendicular to ℓ
    - [i] P and Q are on the same side of ℓ
    - [ii] P and Q are on opposite sides of ℓ
  - [2] $\overline{PQ}$ is not perpendicular to ℓ
    - [i] P and Q are on the same side of ℓ
    - [ii] P and Q are on opposite sides of ℓ

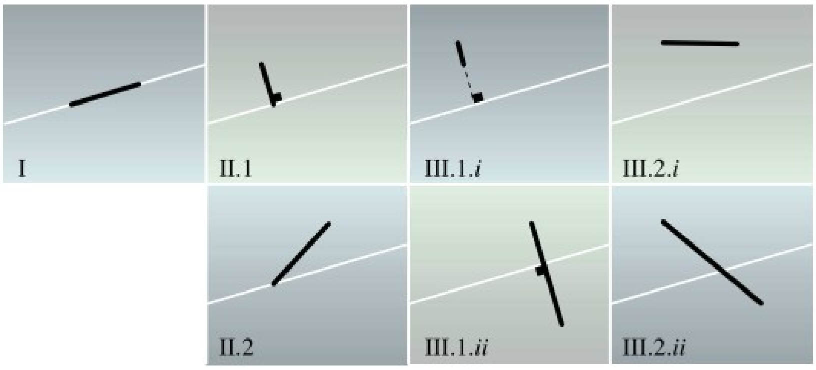

Figure 2. Verifying that a reflection is an isometry. There are many cases, depending on the position of the segment PQ relative to the reflecting line ℓ.

At this point, none of these cases should cause any trouble. Consider case III.2.*i*, which is the archetypal case in this proof. To verify the claim in this case, first label two more points (both of which are fixed by s):

F_P**:** the foot of the perpendicular to ℓ through P, and
F_Q**:** the foot of the perpendicular to ℓ through Q.

From the definition of a reflection,

$$PF_P \simeq s(PF_P) \quad \text{and} \quad QF_Q \simeq s(QF_Q)$$

and the angles at F_P and F_Q are right angles. Because $F_P F_Q$ is congruent to itself, by SASAS, $\square PF_PF_QQ$ and $s(\square PF_PF_QQ)$ are congruent, and therefore PQ and $s(PQ)$ are the same length. □

We saw in the last chapter that if an isometry fixes two points, it must fix all the points on the line through them. Of course, every reflection fixes all the points of a line. A good question

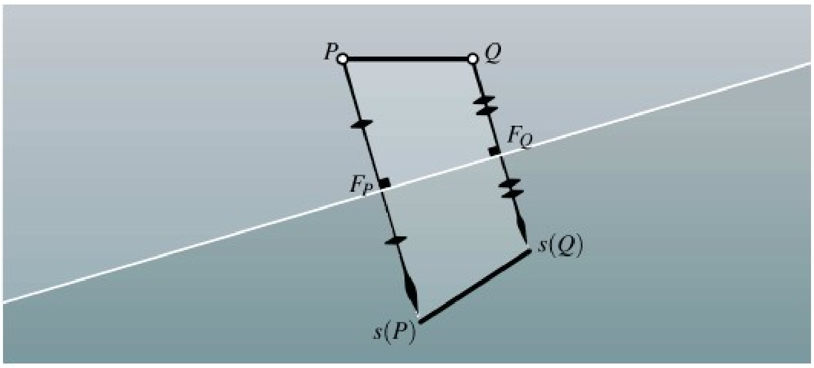

Figure 3. Use SASAS to verify that reflections preserve isometries in case III.2.*i*.

to ask, then, is how common is this "line-fixing" behavior? Not that common, as it turns out, and so this is a useful characterization of a reflection.

Theorem 23.3 *If an isometry fixes all the points of a line but is not the identity, then it must be a reflection.*

Proof Let f be an isometry that fixes all the points on a line ℓ but is not the identity. Let s be the reflection across ℓ. We already know that f and s agree on ℓ, so we just need to show that they agree on one point that isn't on ℓ. Take two points A and B on ℓ and a third point C that is not on ℓ. Since an isometry preserves distance, and since A and B are fixed,

$$AC \simeq f(AC) \simeq Af(C) \implies f(C) \text{ is on the circle with center } A \text{ and radius } |AC|\text{, and}$$
$$BC \simeq f(BC) \simeq Bf(C) \implies f(C) \text{ is on the circle with center } B \text{ and radius } |BC|.$$

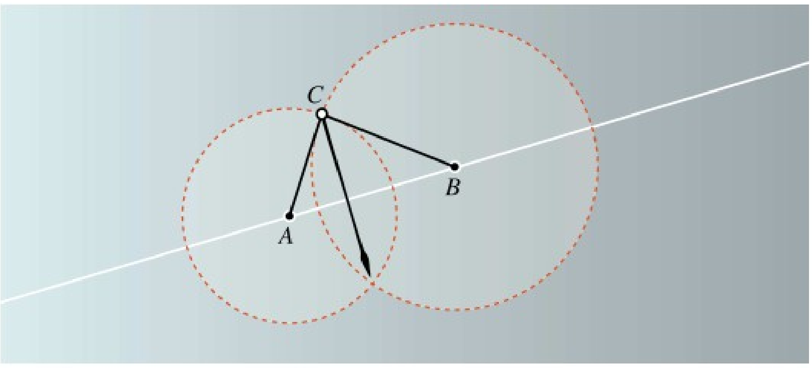

Figure 4. If a non-identity isometry fixes a line, then it is a reflection.

Therefore $f(C)$ is at an intersection of these two circles, and there are only two (distinct circles intersect at most twice). Furthermore, one of them is C, and if $f(C) = C$, then f would fix three non-collinear points and would be the identity. We excluded that possibility, so $f(C)$ has to be the other point of intersection. For the same reasons, $s(C)$ must be the second intersection. Therefore $f(C) = s(C)$; the two isometries agree on three non-collinear points, A, B, and C and so they must be equal. ☐

Theorem 23.4 *(Three reflections theorem) Any isometry can be written as a reflection, a composition of two reflections, or a composition of three reflections.*

We saw in the last chapter that when isometries agree on three non-collinear points they have to be the same. That is how we will proceed. We just need to find a composition of up to three reflections $s_3 \circ s_2 \circ s_1$ that agrees with T on three non-collinear points. There are three steps to this, as we build up each of the three reflections. At each step we will make the composition of reflections match T on another point without moving any of the previously set points.

Proof Let A, B, and C be three non-collinear points and let T be an isometry.

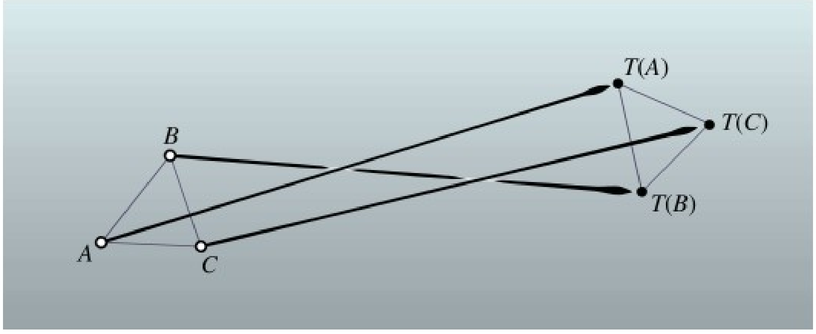

Figure 5. The image of a triangle under an isometry T.

Step One. With the first isometry, s_1, we put A into position. If $A = T(A)$, let s_1 be the identity isometry. If $A \neq T(A)$, let s_1 be the reflection across the perpendicular bisector of $AT(A)$. Either way, $s_1(A) = T(A)$.

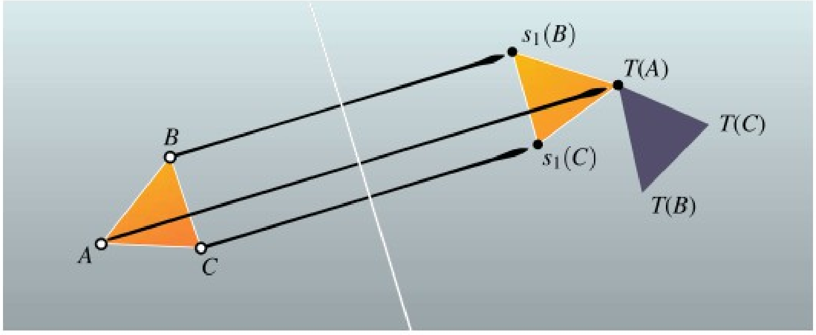

Figure 6. The first reflection.

Step Two. With the second isometry, s_2, we put B into position. To do this, we look at where $s_1(B)$ ended up after step one. It is possible (but unlikely) that $s_1(B)$ ended up on the line $T(\overline{AB})$. If that is the case, then because

$$|s_1(AB)| = |AB| = |T(AB)|,$$

there are only two possible spots for $s_1(B)$, one on either side of $T(A)$. If $s_1(B)$ is on the same side of $T(A)$ as $T(B)$, then $s_1(B) = T(B)$ already, so we can let s_2 be the identity isometry. If $s_1(B)$ is on the opposite side of $T(A)$ from $T(B)$, then let s_2 be the reflection across the line that passes through $T(A)$ and is perpendicular to $s_1(B)T(B)$. That reflection fixes $T(A)$ and maps $s_1(B)$ to $T(B)$.

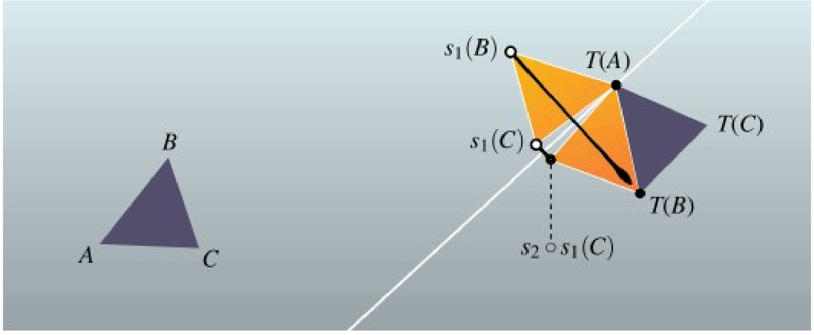

Figure 7. The second reflection.

The more likely possibility is that $s_1(B)$ is not on $T(\overline{AB})$. In that case, let s_2 be the reflection across the bisector of $\angle s_1(B)T(A)T(B)$. Then $T(A)$ is on the line of reflection, so it will be fixed by s_2. Furthermore, the reflecting line cuts $\triangle s_1(B)T(A)T(B)$ in two pieces, that, by SAS, are congruent. Therefore the reflecting line is the perpendicular bisector to $s_1(B)T(B)$, and that means s_2 will map $s_1(B)$ to $T(B)$. Here is where we stand after step two:

$$s_2 \circ s_1(A) = s_2 \circ T(A) = T(A),$$
$$s_2 \circ s_1(B) = T(B).$$

Step Three. That leaves point C. As in the previous step, what we do next depends upon where $s_2 \circ s_1(C)$ is. There aren't many possibilities. We know that $s_2 \circ s_1(AB) = T(AB)$, and we know that $s_2 \circ s_1(\triangle ABC)$ is congruent to $\triangle ABC$, which is congruent to $T(\triangle ABC)$. There are only two ways to build that triangle on a given side of $T(AB)$, one on either side of it. If $s_2 \circ s_1(C)$ is on the same side of $T(AB)$ as $T(C)$, then $s_2 \circ s_1(C) = T(C)$ already, so let s_3 be the identity map. If $s_2 \circ s_1(C)$ is on the opposite side of $T(AB)$ from $T(C)$, then let s_3 be the reflection across the line $T(AB)$. That fixes $T(A)$ and $T(B)$, but maps $s_2 \circ s_1(C)$ to $T(C)$. Putting it all together,

$$s_3 \circ s_2 \circ s_1(A) = T(A)$$
$$s_3 \circ s_2 \circ s_1(B) = T(B)$$
$$s_3 \circ s_2 \circ s_1(C) = T(C).$$

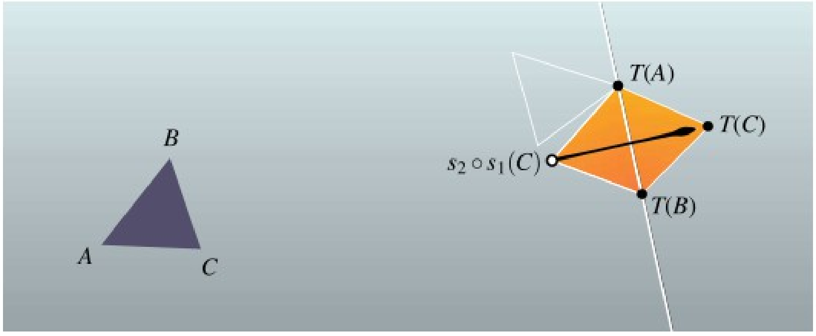

Figure 8. The third reflection.

Since the two isometries agree on three non-collinear points, they must be the same. As long as at least one of s_1, s_2, and s_3 is a reflection, we have met the requirements of the theorem. What if all of them are the identity map? In that case, T is the identity map, and the identity can be written as the composition of any reflection s with itself: $T = s \circ s$. □

Over the next few chapters, we will use this result to classify all isometries. In the next chapter, we will look at what happens when we compose two reflections. Then, after a brief discussion of orientation, we will look at what happens when we tack on a third reflection.

23.1 The analytic viewpoint

It is a little messy to try to work out an equation for an arbitrary reflection at this point. We can, however, find an equation for a reflection across a line that passes through the origin. Let's close out this chapter by doing so.

Theorem 23.5 *Let ℓ be a line through the origin, and let (a, b) be the coordinates of an intersection of ℓ with the unit circle. Then the reflection s across ℓ is given by*

$$s\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a^2 - b^2 & 2ab \\ 2ab & b^2 - a^2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}.$$

Proof Since (a, b) is on the unit circle, it can be written as $(\cos\theta, \sin\theta)$ for some real number θ. Let D be the distance from the point (x, y) to the origin and let ϕ be its angle measure as measured from the x-axis, in the counterclockwise direction, so that

$$\begin{matrix} \cos\phi = x/D \\ \sin\phi = y/D \end{matrix} \implies \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} D\cos\phi \\ D\sin\phi \end{pmatrix}.$$

If α is the angle between ϕ and θ, $\alpha = \phi - \theta$, then $s(x, y)$ will still be at a distance D from the origin, but at an angle

$$\phi - 2\alpha = \phi - 2(\phi - \theta) = 2\theta - \phi.$$

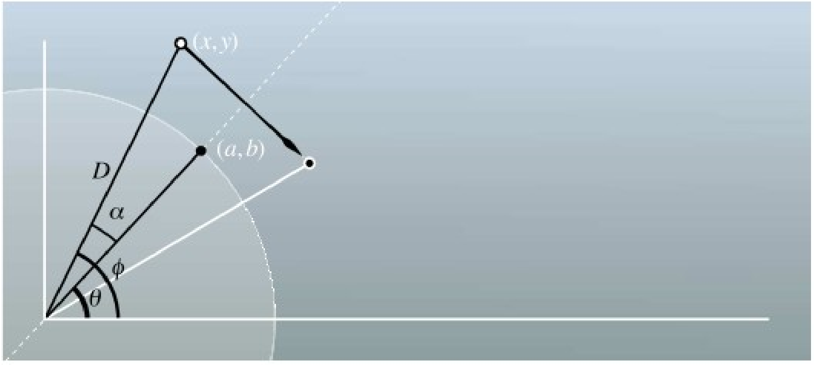

Figure 9. A reflection across a line that passes through the origin. In this case $\phi > \theta$ so that $\alpha > 0$, but the calculations are the same if $\alpha < 0$.

Therefore

$$s\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} D\cos(2\theta - \phi) \\ D\sin(2\theta - \phi) \end{pmatrix}$$

and we can use the addition rules for sine and cosine:

$$\begin{aligned} s\begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} D\cos(2\theta)\cos(-\phi) - D\sin(2\theta)\sin(-\phi) \\ D\sin(2\theta)\cos(-\phi) + D\cos(2\theta)\sin(-\phi) \end{pmatrix} \\ &= \begin{pmatrix} D\cos(2\theta)\cos\phi + D\sin(2\theta)\sin\phi \\ D\sin(2\theta)\cos\phi - D\cos(2\theta)\sin\phi \end{pmatrix}. \end{aligned}$$

This can factored into a matrix form, and from there, the double angle formulas will take us the rest of the way.

$$\begin{aligned} s\begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} \cos(2\theta) & \sin(2\theta) \\ \sin(2\theta) & -\cos(\theta) \end{pmatrix} \begin{pmatrix} D\cos\phi \\ D\sin\phi \end{pmatrix} \\ &= \begin{pmatrix} \cos^2\theta - \sin^2\theta & 2\sin\theta\cos\theta \\ 2\sin\theta\cos\theta & \sin^2\theta - \cos^2\theta \end{pmatrix} \begin{pmatrix} D\cos\phi \\ D\sin\phi \end{pmatrix} \\ &= \begin{pmatrix} a^2 - b^2 & 2ab \\ 2ab & b^2 - a^2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}. \end{aligned}$$

□

There are two special cases worth noting. The equation for reflecting across the x-axis is

$$s\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ -y \end{pmatrix},$$

and the equation for reflecting across the y-axis is

$$s\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -x \\ y \end{pmatrix}.$$

23.2 Exercises

23.1. Given $\triangle ABC$, let s be the reflection across the perpendicular bisector of BC. Prove that if $\triangle ABC$ is invariant under s (that is, $s(\triangle ABC) = \triangle ABC$), then $\triangle ABC$ is isosceles.

23.2. Given $\square ABCD$, let E and F be midpoints of AB and CD respectively. Let s be the reflection across the line EF. Prove that if $s(A) = B$ and $s(C) = D$, then $\square AEFD \simeq \square BEFC$.

23.3. Given a convex quadrilateral $\square ABCD$, let s_1 be the reflection across the diagonal AC, and let s_2 be the reflection across the diagonal BD. If $s_1(B) = D$ and $s_2(A) = C$, what can you say about $\square ABCD$?

23.4. What is the matrix equation for a reflection across the line $y = x$?

23.5. What is the matrix equation for a reflection across the horizontal line $y = k$?

23.6. Suppose that s_1 and s_2 are reflections, and that for some point P, $s_1(P) = s_2(P) \neq P$. Is it necessarily true that $s_1 = s_2$?

23.7. Suppose that s_1 and s_2 are reflections, and that for two distinct points P and Q, $s_1(P) = s_2(P)$ and $s_1(Q) = s_2(Q)$. Is it necessarily true that $s_1 = s_2$?

23.8. Given lines ℓ_1 and ℓ_2 is there necessarily a reflection s so that $s(\ell_1) = \ell_2$? If so, is it unique?

23.9. Let s_1 and s_2 be reflections across perpendicular lines ℓ_1 and ℓ_2 that intersect at a point P. Show that if Q is any other point, then P is the midpoint of the segment connecting Q to $s_2 \circ s_1(Q)$.

23.10. Let s_1 and s_2 be reflections across perpendicular lines ℓ_1 and ℓ_2. For any point P, prove that

$$P = s_2 \circ s_1 \circ s_2 \circ s_1(P),$$

and that there is a circle which passes through the points P, $s_1(P)$, $s_2 \circ s_1(P)$, and $s_1 \circ s_2 \circ s_1(P)$.

24

Translations and Rotations

The big result of the last chapter was that every isometry can be written as a reflection or as a composition of two or three reflections. In this chapter we will look at the types of isometries that we can get by composing two reflections. Any reflection composed with itself results in the identity, so we are interested in compositions of two distinct reflections. Then there are two cases: either the reflecting lines are parallel or the reflecting lines are intersecting.

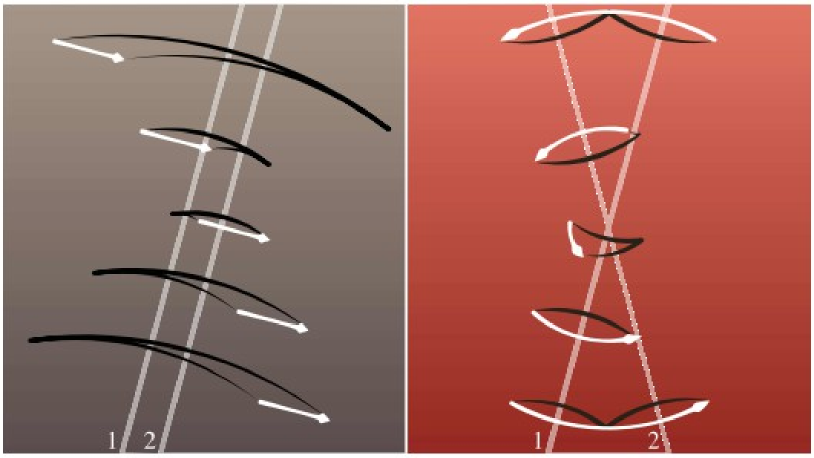

Figure 1. Pairs of reflections: (l) across parallel lines and (r) across intersecting lines.

The two cases describe fundamentally different types of isometries. In the second, the intersection point of the two lines is fixed by the composition of isometries. This doesn't happen in the first, since there is no intersection point, and in fact, this composition does not have any fixed points.

Definition 24.1 A *translation* is a composition of reflections across parallel lines. A *rotation* is a composition of reflections across intersecting lines.

By defining translations and rotations as compositions of isometries, it is automatically true that they will be isometries as well. But the definitions do not do a good job of revealing what a translation or rotation actually looks like. Doing that is the purpose of this chapter.

24.1 Translation

First, let's tackle the case of the translation. To do that, it is helpful to take a more measured look at the behavior of a single reflection. Consider a reflection s across a line ℓ. Let P be a point that is not on ℓ and let $\ell_\perp$ be the line through P that is perpendicular to ℓ. Now let's set up $\ell_\perp$ as a number line. That is, choose an arbitrary point O to be the origin, and a ray from O that points in the positive direction; then every point on $\ell_\perp$ has a coordinate, its signed distance from O. Suppose that P is at coordinate x and that ℓ and $\ell_\perp$ intersect at the point Q with coordinate y. Given the definition of a reflection, $s(P)$ has to be somewhere on $\ell_\perp$, and so it too must correspond to some coordinate. What coordinate? The distance from P to Q is $|y-x|$. Since s is an isometry and Q is a fixed point, the distance from $s(P)$ to Q is $|y-x|$ too. That limits the coordinate for $s(P)$ to one of two possibilities:

$$y+|y-x| = \begin{cases} y+(y-x) = 2y-x & \text{if } y-x \geq 0 \\ y-(y-x) = x & \text{if } y-x < 0 \end{cases}$$

or

$$y-|y-x| = \begin{cases} y-(y-x) = x & \text{if } y-x \geq 0 \\ y-(-(y-x)) = 2y-x & \text{if } y-x < 0. \end{cases}$$

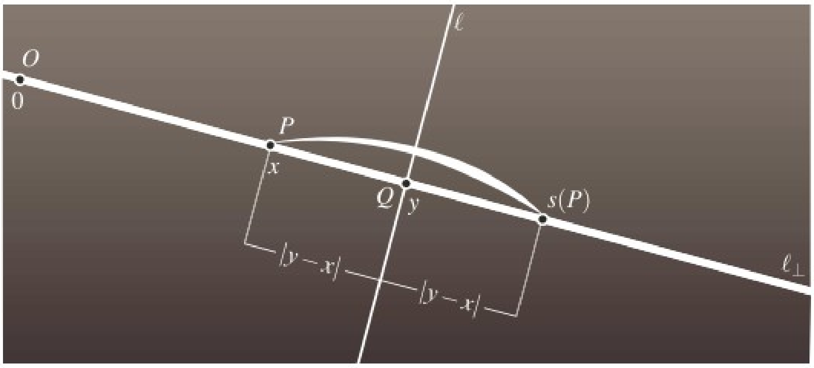

Figure 2. Measuring reflections along a line.

Since P is not on ℓ it is not a fixed point, so $s(P)$ is not at the coordinate x. The only other possibility, then, is that $s(P)$ is at the coordinate $2y-x$. The formula still works even if P is on ℓ. For in that case P is fixed, so $s(P)$ is also at coordinate x. That is what the formula reveals: if P is on ℓ, $y=x$, and so $2y-x=x$. Having this formula in hand will make it a little easier to compose parallel reflections.

Theorem 24.2 *Suppose that t is the translation $s_2 \circ s_1$ where s_1 and s_2 are reflections across parallel lines ℓ_1 and ℓ_2 that are separated by a distance d. Then for any point P, $t(P)$ is located*

- *on the line through P that is perpendicular to both ℓ_1 and ℓ_2,*
- *in the direction of the ray that points from ℓ_1 to ℓ_2,*
- *at a distance $2d$ from P.*

Proof Take a point P, and let $\ell_\perp$ be the line through P that is perpendicular to ℓ_1 and ℓ_2. By definition, $s_1(P)$ will still be on $\ell_\perp$, and then so will $s_2(s_1(P))$. Let's just look along this line, and, as in the preceding discussion, lay out a number line along it. It does not matter where we put the origin on the line, but it does help the discussion to choose the positive direction so that going from ℓ_1 to ℓ_2 moves in the positive direction. Then mark the coordinates

x : coordinate of P,
y_1 : coordinate for the intersection of $\ell_\perp$ and ℓ_1,
y_2 : coordinate for the intersection of $\ell_\perp$ and ℓ_2.

According to our previous calculations, $s_1(x)$ will be at the coordinate $2y_1 - x$ and $s_2 \circ s_1(x)$ will be at the coordinate

$$2y_2 - (2y_1 - x) = x + 2(y_2 - y_1) = x + 2d.$$

Therefore $s_2 \circ s_1(P)$ will be $2d$ farther along the line $\ell_\perp$ than P in the direction pointing from ℓ_1 to ℓ_2.

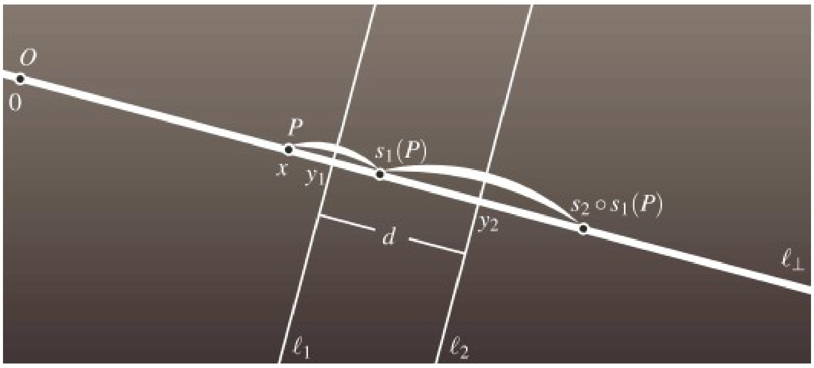

Figure 3. Measuring a composition of parallel reflections along a line. ☐

To recap, a translation moves all points along lines that are perpendicular to ℓ_1 and ℓ_2. They move in parallel, in the same direction, over the same distance. All of that—the parallel lines, the direction, and the distance—can be determined by looking at the effect of the translation on a single point. Therefore a translation is completely determined by its behavior on a single point. Because of that, we can get a precise idea of how many translations there are.

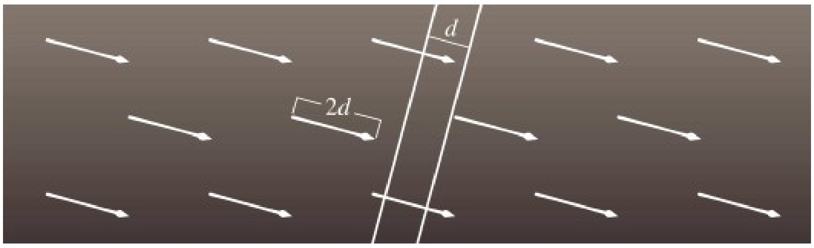

Figure 4. Transations shift all points the same amount in the same direction.

Theorem 24.3 *Given two distinct points P and Q, there is exactly one translation t so that $t(P) = Q$.*

Proof To prove existence, we will describe (as a composition of reflections) a translation that maps P to Q. The two reflections, s_1 and s_2, will be across lines that are perpendicular to PQ (and hence are parallel to one another). Let s_1 be the reflection across the line through P. Let s_2 be the reflection across the line through the midpoint of PQ. Then $s_2 \circ s_1$ is a translation and

$$s_2 \circ s_1(P) = s_2(P) = Q.$$

Proving uniqueness is easy: since a translation is completely determined by its behavior on one point, there can be only one translation taking P to Q. ☐

It is cumbersome to try to think of a translation as a composition of reflections. The properties we derived give a much better sense of the effects of a translation, and they can be formalized as follows. A *directed segment* is a line segment that distinguishes between its ends: one is called the initial endpoint, the other the terminal endpoint. We can define an equivalence relation $\sim$ on the set of all directed segments as follows: two directed segments σ_1 and σ_2 are equivalent if there is a translation t mapping σ_1 to σ_2 so that initial point is mapped to initial point and terminal point is mapped to terminal point.

Exercise 24.1 Verify that the relation $\sim$ is an equivalence relation on the set of all directed segments.

Definition 24.4 A *vector* is an equivalence class of directed segments under the $\sim$ equivalence relation.

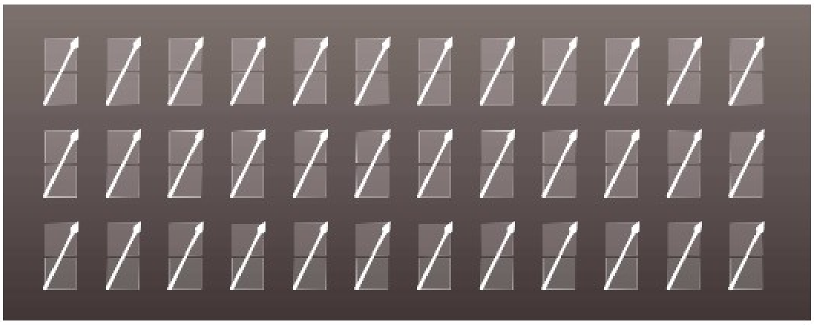

Figure 5. Some equivalence class representatives of a vector (one over, two up).

Associated to any transformation t is the vector that is represented by directed segments of the form $Pt(P)$ with initial point P and terminal point $t(P)$. The vector defines and is defined by t. It is called the *translation vector* of t. It is almost always more convenient and natural to think about a translation in terms of its translation vector than as a composition of reflections. For instance, if you think of a translation t as a composition of reflections, it might not be that clear that t has no fixed points. If you think of that translation in terms of its translation vector, it is clear that no point P can be fixed by t, since $Pt(P)$ is always a directed segment with two distinct endpoints.

24.2 Rotations

The illustrations at the start of this chapter suggest that when ℓ_1 and ℓ_2 intersect, the rotation $r = s_2 \circ s_1$ acts by turning points around the intersection point O. To measure the effect of the turning, we establish an angular coordinate system around O (just as we established a linear coordinate system on $\ell_\perp$ when ℓ_1 and ℓ_2 were parallel). Choose a ray $\overrightarrow{OR}$ with endpoint O. This marks the zero angle and a direction around that point (this amounts to a choice of clockwise or counterclockwise, which is discussed in the next chapter). After choosing, every ray from O will form an angle with $\overrightarrow{OR}$ and we can then associate each point on the ray with the angle measure. Before attempting two reflections, let's try to understand how the angular coordinate of a point behaves when we apply one reflection s across a line ℓ to it. Pick a point O on ℓ and set up an angular coordinate system as described. Let P be an arbitrary point that is not on ℓ. Then label

θ : the angular coordinate at P
ϕ : the angular coordinate of one of the rays from O that make up ℓ.

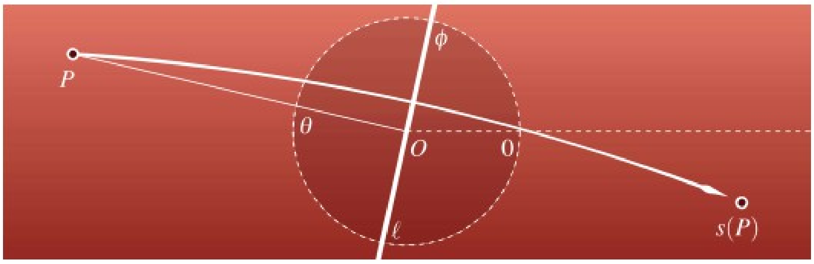

Figure 6. Measuring reflections around a fixed point.

There are two choices for ϕ that differ by π, but for the calculation, it makes no difference which one we pick. The angle between ℓ and OP has a measure of $|\phi - \theta|$. Since isometries preserve angle measure and ℓ is fixed by s, the angle between ℓ and $Os(P)$ also has a measure of $|\phi - \theta|$. That limits the possibilities for the angular coordinates of $s(P)$ to

$$\phi + |\phi - \theta| = \begin{cases} 2\phi - \theta & \text{if } \phi - \theta \geq 0 \\ \theta & \text{if } \phi - \theta < 0 \end{cases}$$

or

$$\phi - |\phi - \theta| = \begin{cases} \theta & \text{if } \phi - \theta \geq 0 \\ 2\phi - \theta & \text{if } \phi - \theta < 0. \end{cases}$$

Since P is not on ℓ, it is not fixed, and therefore $s(P)$ will not be at angle θ. The only other possibility is that $s(P)$ is at angle $2\phi - \theta$. Furthermore, the formula still holds when P is on ℓ. In that case, P is fixed, so $s(P)$ is also at angle θ. That is what the formula indicates: if P is on ℓ, then $\phi = \theta$ and so $2\phi - \theta = \theta$. Now let's take the formula and use it to see what happens when we compose two intersecting reflections.

Theorem 24.5 *Suppose that r is the rotation $s_2 \circ s_1$ where s_1 and s_2 are reflections across lines ℓ_1 and ℓ_2 intersecting at a point O and at an angle of θ to one another. For any point P, $r(P)$ is located*

- *on the circle centered at O that passes through P*
- *so that OP and $Os(P)$ form an angle with measure 2θ,*
- *in the direction indicated by the arc from ℓ_1 to ℓ_2.*

Proof Since $s_2 \circ s_1$ preserves distances and O is a fixed point, the distance from O to $s(P)$ is the same as the distance from O to P. That places $s(P)$ on the circle centered at O passing through P. Where precisely? As in the discussion above, set up an angular coordinate system centered at O. Mark the coordinates

α **:** the angular coordinate for P,
ϕ_1 **:** the angular coordinate for ℓ_1,
ϕ_2 **:** the angular coordinate for ℓ_2.

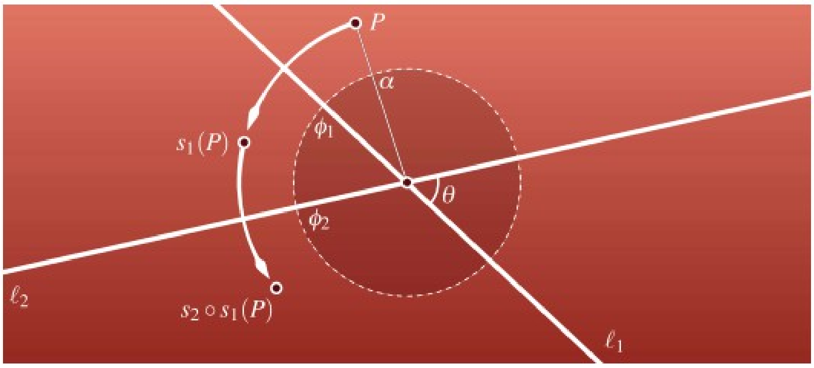

Figure 7. Measuring a composition of reflections around the fixed point.

The intersection of ℓ_1 and ℓ_2 forms two vertical angle pairs. It is helpful to measure angles in the direction so that the directed angle from ℓ_1 to ℓ_2 is the smaller of them (if ℓ_1 and ℓ_2 intersect at right angles, it doesn't matter which orientation you choose). According to the previous discussion, $s_1(P)$ will have the coordinate $2\phi_1 - \alpha$. Then $s_2(s_1(P))$ will have the angular coordinate

$$2\phi_2 - (2\phi_1 - \alpha) = \alpha + 2(\phi_2 - \phi_1) = \alpha + 2\theta.$$

Therefore OP and $Os(P)$ form an angle of 2θ, measured in the direction from ℓ_1 to ℓ_2. ☐

It is generally more convenient to think of a rotation in terms of the angle 2θ, the *rotation angle*, and the fixed point, the *center of rotation*, rather than as a composition of reflections. For instance, by thinking of a rotation in terms of its rotation angle and center, it is clear that a rotation only has one fixed point, the center of rotation. This viewpoint also gives a good perspective on just how common rotations are. The proof of the following result is left to the reader.

Theorem 24.6 *For a point O and angle measure $0 < \theta < 2\pi$, there is exactly one rotation in each of the two directions around O that has rotation center O and rotation angle θ. When $\theta = \pi$, the rotations in those two directions coincide (a half-turn).*

Exercise 24.2 Prove theorem 24.6.

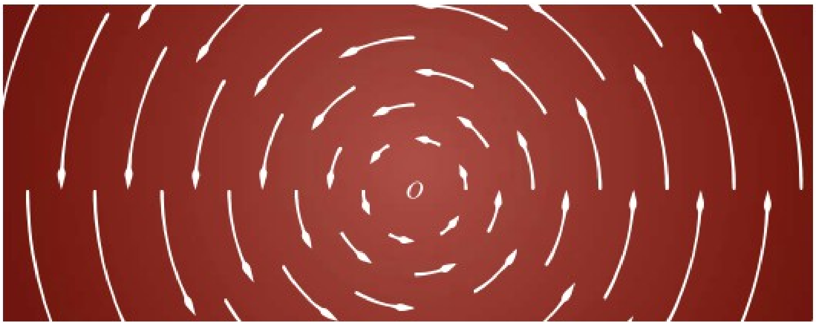

Figure 8. A counterclockwise rotation by $\pi/6$ centered at O.

24.3 The analytic viewpoint

Analytically, translations are the simplest of the isometries. If we break the translation vector of a translation T down into a horizontal component h and a vertical component k, then

$$T\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x+h \\ y+k \end{pmatrix}.$$

The equations for rotations are more challenging. For now, let's restrict our attention to rotations that are centered at the origin.

Theorem 24.7 *The analytic equation for a rotation r around the origin by an angle θ in the counterclockwise direction is*

$$r\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}.$$

Proof We can realize this rotation as a composition of two reflections across lines through the origin. For convenience, let's choose the reflections s_1 across ℓ_1 and s_2 across ℓ_2, where ℓ_1 is the x-axis and ℓ_2 forms an angle of $\theta/2$ (counterclockwise) with the x-axis.

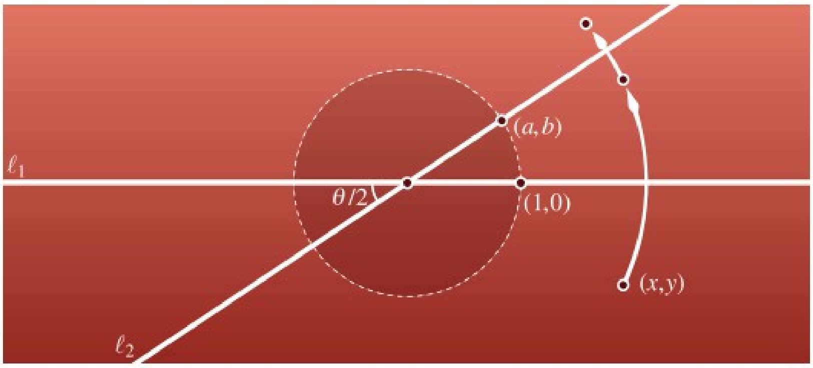

Figure 9. A rotation by θ around the origin can be written as a composition of reflections across two lines which form a $\theta/2$ angle at the origin.

Then $s_2 \circ s_1$ will be a rotation by an angle of $2 \cdot \theta/2 = \theta$. In the last chapter, we found out that equations for these reflections take the form

$$s\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a^2 - b^2 & 2ab \\ 2ab & b^2 - a^2 \end{pmatrix}$$

where (a, b) marks the intersection of the line and the unit circle. We can use that equation now. The first line intersects the unit circle at $(1, 0)$, so

$$s_1\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}.$$

The second line intersects the unit circle at $(\cos\theta/2, \sin\theta/2)$, so

$$s_2\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos^2(\theta/2) - \sin^2(\theta/2) & 2\cos(\theta/2)\sin(\theta/2) \\ 2\cos(\theta/2)\sin(\theta/2) & \sin^2(\theta/2) - \cos^2(\theta/2) \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}.$$

We can use the double angle formulas to write

$$\begin{aligned} \cos^2(\theta/2) - \sin^2(\theta/2) &= \cos(\theta), \\ 2\cos(\theta/2)\sin(\theta/2) &= \sin(\theta), \end{aligned}$$

which simplifies the matrix considerably to

$$s_2\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}.$$

To compute the composition of the transformations, multiply the matrices:

$$r\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}. \qquad \square$$

24.4 Exercises

24.3. Prove that every isometry T can be written as a composition $t_1 \circ t_2$ where t_1 is a translation (or the identity) and t_2 is an isometry that fixes the origin.

24.4. Find the analytic equations for reflections across the lines $x = a$ and $x = b$. Then verify that their composition is a translation.

24.5. Suppose that r_1 and r_2 are counterclockwise rotations about the origin by angles of θ_1 and θ_2 respectively. From the matrix equations for r_2 and r_1, show that the matrix equations for $r_2 \circ r_1$ have the form of a rotation or the identity.

24.6. Suppose that ℓ is an invariant line of a rotation r. That is, if P is a point on ℓ, then $r(P)$ is also on ℓ. Show that ℓ passes through the center of rotation and the angle of rotation is π.

24.7. Let r be a rotation that maps the point $(0, 0)$ to the point $(1, 0)$ and the point $(1, 0)$ to the point (a, b). What conditions are required of a and b for this to be possible?

24.8. In exercise 21.17 of chapter 21, we found the equation of a parabola with focus $(0, a)$ and directrix line $y = -a$, where $a \neq 0$. Find the equation of a parabola with focus $(a, 0)$ and directrix line $x = -a$ in two ways. First, do a direct calculation, as in chapter 21. Second, write a point on the parabola from exercise 21.17 in the parametrized form $(t, t^2/(4a))$ and apply a rotation.

24.9. Working from the previous exercise, find the equation for a parabola with focus $(3, 5)$ and directrix line $x = -1$.

24.10. Show that the parabola with focus at $(1, 1)$ and directrix line $y = -2 - x$ has the equation

$$(x - y)^2 - 8(x + y) = 0.$$

24.11. Find the equation of the ellipse with foci $(2, 0)$ and $(6, 0)$ that passes through the point $(0, 0)$.

24.12. Find the equation of the ellipse with foci $Q_1 = (1, \sqrt{3})$ and $Q_2 = (-1, -\sqrt{3})$ so that for a point $P = (x, y)$ on it

$$d(P, Q_1) + d(P, Q_2) = 8.$$

25

Orientation

The remaining step in the classification of isometries is to tackle compositions of three reflections. We would like to know whether compositions of three reflections can devolve into any of the previously identified isometries (reflection, translation, rotation). It turns out that composing three reflections can yield another reflection, but it will never result in a translation or rotation. This is helpful. It comes down to the issue of orientation, which we will investigate in this chapter.

An *orthonormal frame* is an ordered pair of perpendicular unit length segments that share a common endpoint. One such frame, $\mathscr{F}_+ = \{OP_x, OP_y\}$, centered at the origin with

$$O = (0, 0), \quad P_x = (1, 0), \quad \text{and} \quad P_y = (0, 1),$$

is at the heart of the coordinate system. There is another such frame, $\mathscr{F}_- = \{OP_x, OP'_y\}$, that shares the same first segment as $\mathscr{F}_+$, but that has $P'_y = (0, -1)$. In general, a frame can be viewed as a way to represent information about orientation, whether it is clockwise or counterclockwise. Until now, we have only made that choice at the origin: in $\mathscr{F}_+$, the directed minor arc from P_x to P_y points in the counterclockwise direction; in $\mathscr{F}_-$, the directed minor arc from P_x to P_y points in the clockwise direction. By using translations and rotations, we can now propagate that choice across the rest of the plane. Consider a frame $\mathscr{F} = \{QQ_x, QQ_y\}$. Let r be the rotation centered at O so that the vectors $r(OP_x)$ and QQ_x are parallel and pointed in the same direction. Let t be the translation which maps the point O to Q. Then

$$t \circ r(O) = Q \quad \text{and} \quad t \circ r(P_x) = Q_x.$$

What about P_y and P'_y? Both $t \circ r(OP_y)$ and $t \circ r(OP'_y)$ are perpendicular to QQ_x, and both $t \circ r(P_y)$ and $t \circ r(P'_y)$ are located a distance of 1 from Q along their segments. Because of the segment and angle construction axioms, there are only two points that meet both requirements. One of them is Q_y. Therefore, exactly one of two things may happen: either $t \circ r(P_y) = Q_y$ or $t \circ r(P'_y) = Q_y$. In the first case, $t \circ r$ maps $\mathscr{F}_+$ to $\mathscr{F}$; then we call $\mathscr{F}$ a counterclockwise frame, and the directed minor arc from Q_x to Q_y is in the counterclockwise direction. In the second case, $t \circ r$ maps $\mathscr{F}_-$ to $\mathscr{F}$; then we call $\mathscr{F}$ a clockwise frame, and the directed minor arc from Q_x to Q_y is in the clockwise direction.

This is a two-step process, in the order $t \circ r$, where r is a rotation around the origin, followed by a translation t. As long as we stick to that way of doing it, there is one and only one way to map one of the two original frames to any other. Hence there is no issue about making a consistent choice because there is no choice to make. This is not a trivial issue because not all

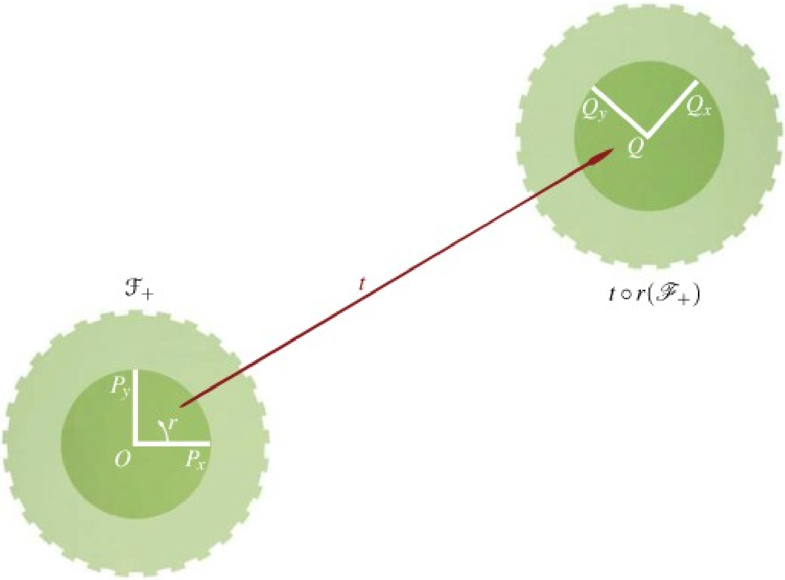

Figure 1. A rotation followed by a translation maps the frame $\mathcal{F}_+$ to the frame $\mathcal{F} = \{QQ_x, QQ_y\}$. Therefore $\mathcal{F}$ is a counterclockwise frame.

surfaces can be oriented consistently like this. The most famous non-orientable surface is the Möbius strip. It is formed by taking a strip, giving it a half-twist, and joining the ends. A frame $\mathcal{F}$ on the Möbius strip can be translated from one point to another in two different ways, t_1 and t_2, and the resulting frames $t_1(\mathcal{F})$ and $t_2(\mathcal{F})$ are not oriented the same way. The unfortunate side effect of the two-step process is that it doesn't determine whether other combinations of rotations and translations might actually reverse orientation, even though we instinctively know that to not be true. We will show that soon as an easy consequence of the fact that a reflection reverses the orientation of a frame, which we will tackle next. That argument will require the following lemma.

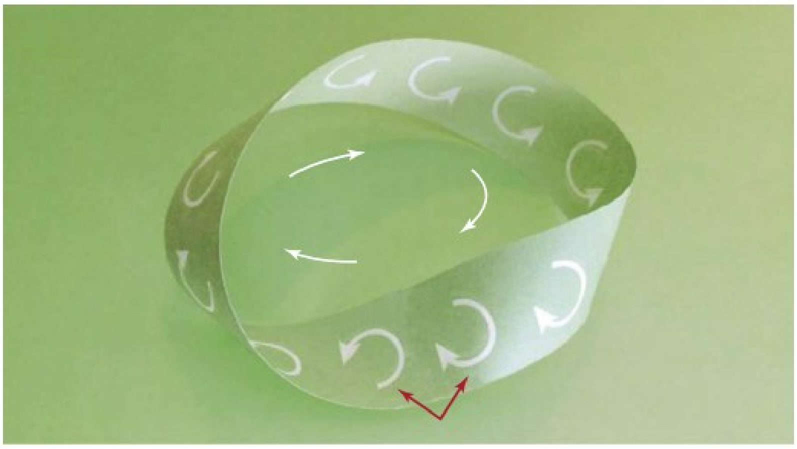

Figure 2. One lap around the Möbius strip flips orientation.

Lemma 25.1 *Let r be a rotation by an angle θ (not a half-turn), and let ℓ be a line. Then the angle between ℓ and $r(\ell)$ is θ.*

Exercise 25.1 Prove the previous lemma.

Theorem 25.2 *Let s be a reflection and $\mathscr{F}$ be a frame. Then $\mathscr{F}$ and $s(\mathscr{F})$ have the opposite orientation.*

Proof Let $\mathscr{F}$ be a counterclockwise frame (the clockwise case is essentially the same). Now let's suppose that $s(\mathscr{F})$ is also counterclockwise and try to arrive at a contradiction. Because both are counterclockwise, there are rotations r_1 and r_2 around the origin and translations t_1 and t_2 so that

$$t_1 \circ r_1(\mathscr{F}_+) = \mathscr{F} \quad \text{and} \quad t_2 \circ r_2(\mathscr{F}_+) = s(\mathscr{F})$$

so

$$s \circ t_1 \circ r_1(\mathscr{F}_+) = s(\mathscr{F}) = t_2 \circ r_2(\mathscr{F}_+).$$

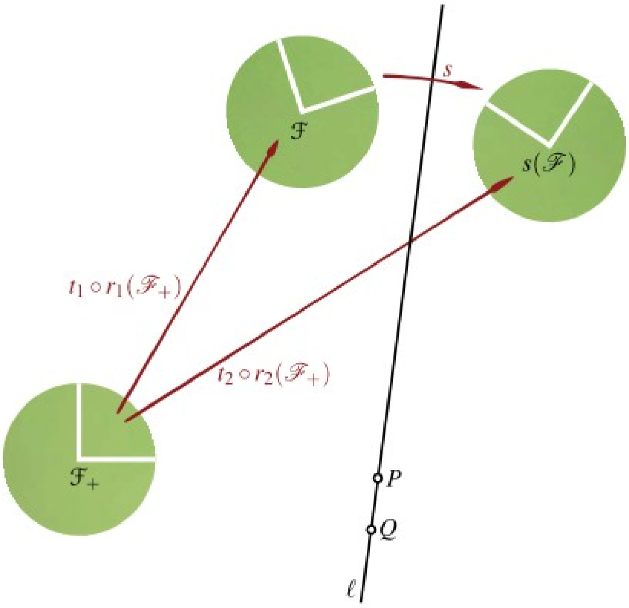
s
$\mathscr{F}$
$s(\mathscr{F})$
$t_1 \circ r_1(\mathscr{F}_+)$
$t_2 \circ r_2(\mathscr{F}_+)$
$\mathscr{F}_+$
P
Q
ℓ

Figure 3. A reflection will always reverse the orientation of a frame because the composition of a rotation, translation, and reflection cannot have the same effect as the composition of just a rotation and a translation.

We have seen that if isometries agree on three non-collinear points, then they must be the same. The frame $\mathscr{F}_+$ contains three non-collinear points, and thus, $s \circ t_1 \circ r_1$ and $t_2 \circ r_2$ are the same. Therefore

$$\begin{aligned} s \circ t_1 \circ r_1 &= t_2 \circ r_2 \\ s \circ t_1 &= t_2 \circ r_2 \circ r_1^{-1} \\ s &= t_2 \circ r_2 \circ r_1^{-1} \circ t_1^{-1}. \end{aligned}$$

On the right hand side, $r_2 \circ r_1^{-1}$ is a composition of two rotations around the origin. Either the two rotations cancel, leaving the identity map, or they compose to form another rotation around the origin. Let ρ denote this composition. On the left, the map s is a reflection, so it fixes a line of points. Choose two distinct points on it, P and Q. Then

$$\begin{aligned} t_2 \circ \rho \circ t_1^{-1}(P) = s(P) = P &\implies \rho \circ t_1^{-1}(P) = t_2^{-1}(P) \\ t_2 \circ \rho \circ t_1^{-1}(Q) = s(Q) = Q &\implies \rho \circ t_1^{-1}(Q) = t_2^{-1}(Q). \end{aligned}$$

According to the last lemma, the angle between $\rho \circ t_1^{-1}(PQ)$ and $t_1^{-1}(PQ)$ is the angle of rotation of ρ, say θ. Both t_1^{-1} and t_2^{-1} are translations, however, so both $t_1^{-1}(PQ)$ and $t_2^{-1}(PQ)$ are parallel to PQ, and hence to each other. Therefore $\rho \circ t_1^{-1}(PQ)$ also forms an angle of θ with $t_2^{-1}(PQ)$. The only way the two segments can coincide, then, is if ρ is either a half-turn or the identity. We may rule out the half-turn since a half-turn would reverse the direction of the vector PQ. Hence ρ must be the identity. That means that $s = t_2 \circ t_1^{-1}$ and we arrive at a contradiction: the composition $t_2 \circ t_1^{-1}$ of two translations is another translation, not a reflection. □

This shows that a reflection is a wholesale swap of orientations. All the clockwise frames reflect to counterclockwise ones and all the counterclockwise frames reflect to clockwise ones.

Definition 25.3 An isometry is *orientation-preserving* if it maps clockwise frames to clockwise frames and counterclockwise frames to counterclockwise frames. A isometry is *orientation-reversing* if it swaps clockwise and counterclockwise frames.

We have just proved that reflections are orientation-reversing. Because every isometry can be written as a composition of reflections, determining whether an isometry is orientation preserving or reversing is a matter of counting flips.

Corollary 25.4 *A composition of two orientation-preserving maps is orientation-preserving; a composition of two orientation-reversing maps is orientation-preserving; a composition of one orientation-preserving map and one orientation-reversing map is orientation-reversing.*

Exercise 25.2 Prove the corollary.

Corollary 25.5 *Translations, rotations, and the identity map are orientation-preserving. They are the only orientation-preserving isometries.*

Exercise 25.3 Prove the corollary.

Let's now recap our progress in the classification of isometries.

# of reflections	isometry	orientation	fixed points
1	reflection	reversing	line
2	identity	preserving	all
	translation	"	none
	rotation	"	point
3	?	reversing	?

In the next chapter we find what goes in place of the question mark.

25.1 Exercises

25.4. Show that if τ is an orientation-preserving isometry that fixes two points, then it is the identity. Show that if τ is an orientation-preserving isometry that has at least one fixed point and at least one non-fixed point, then it is a rotation.

25.5. Let τ_1 be a counterclockwise rotation by $\pi/2$ about the origin. Let τ_2 be a counterclockwise rotation by $\pi/2$ about the point $(1, 0)$. Show that $\tau_1 \circ \tau_2$ is a rotation.

25.6. Another way to think about orientation is with normal vectors formed by cross products. Label the three points $O = (0, 0)$, $P = (1, 0)$, and $Q = (0, 1)$ and let T be an isometry. Form the vectors

v_x**:** the vector from O to P,
v_y**:** the vector from O to Q,
w_x**:** the vector from $T(O)$ to $T(P)$, and
w_y**:** the vector from $T(O)$ to $T(Q)$.

Compare the cross products $v_1 \times v_2$ and $T(v_1) \times T(v_2)$, where T is

- a reflection across a line through O,
- a rotation around O,
- a translation.

25.7. Let T be an orientation-preserving isometry. Prove that T is a rotation if and only if there exist points P and Q so that $|PT(P)| \neq |QT(Q)|$.

25.8. Let t be a translation and r be a rotation. Prove that $t \circ r$ and $r \circ t$ are rotations.

25.9. Consider the matrix equation

$$T\begin{pmatrix} x \\ y \end{pmatrix} = M\begin{pmatrix} x \\ y \end{pmatrix}$$

(where M is a 2×2 matrix) for (a) a reflection across a line through the origin and (b) a rotation around the origin. Show that for a reflection, $\det(M) = -1$ and that for a rotation, $\det(M) = 1$. In this way, the determinant of M captures information about orientation.

The remaining exercises look at surfaces that can be formed by bending and warping a flat sheet and gluing its sides together. More complete investigations can be found in many places. For instance, Massey has a nice write-up in the first chapter of *A Basic Course in Algebraic Topology* [Mas91].

25.10. Let S be the square with vertices $(0, 0)$, $(1, 0)$, $(0, 1)$, and $(1, 1)$. Orient each of those edges with vectors

$\vec{a}$: pointing from $(0, 0)$ to $(1, 0)$
$\vec{b}$: pointing from $(0, 0)$ to $(0, 1)$
$\vec{A}$: pointing from $(0, 1)$ to $(1, 1)$
$\vec{B}$: pointing from $(1, 0)$ to $(1, 1)$.

Imagine that a square S is made from a flexible material. Show that by bending and warping S, and gluing $\vec{a}$ to $\vec{A}$ and $\vec{b}$ to $\vec{B}$ so that they are oriented consistently, the resulting shape, a *torus*, can be made to look like an like an inner tube.

25.11. Let S be the square in the last exercise, with $\vec{a}$, $\vec{b}$, and $\vec{A}$ oriented as before. Orient $\vec{B}$ so that it points from $(1, 1)$ to $(1, 0)$. The resulting shape, a *Klein bottle* cannot be embedded in three-dimensional space without a self-intersection. Show that it is a non-orientable surface. Hint: show that it contains a Möbius strip.

25.12. Let S be the octagon $P_1 P_2 P_3 \cdots P_8$. Glue the following vectors along its edges, matching the orientation of those vectors:

- $\overrightarrow{P_1 P_2}$ with $\overrightarrow{P_4 P_3}$,
- $\overrightarrow{P_2 P_3}$ with $\overrightarrow{P_5 P_4}$,
- $\overrightarrow{P_5 P_6}$ with $\overrightarrow{P_8 P_7}$, and
- $\overrightarrow{P_6 P_7}$ with $\overrightarrow{P_1 P_8}$.

Show that the result can be deformed to form a inner-tube shape with two holes (a "two-holed torus").

37

Non-Euclidean Parallels

From here on out, let's assume that Playfair's axiom is false. That means that our geometry is not Euclidean. It means our triangle angle sums are strictly less than π. It means that Saccheri quadrilaterals are not rectangles, because rectangles do not exist, and so the summit angles are acute (not right). And other absurdities.

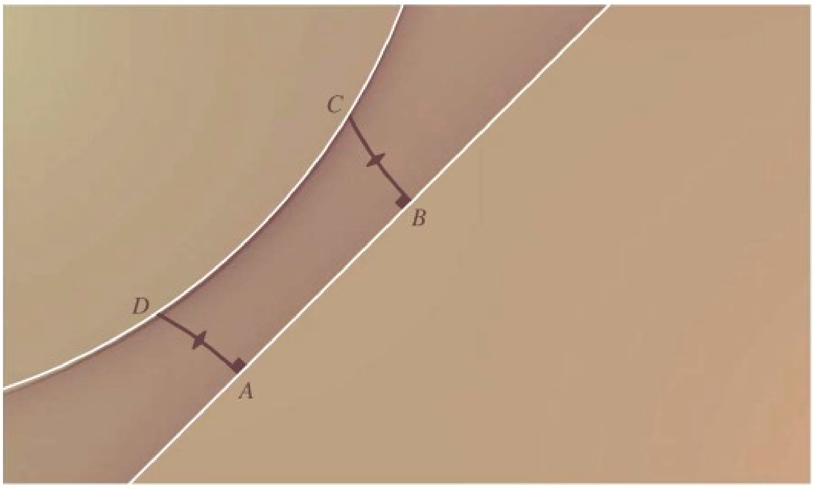

Figure 1. The base and summit of a Saccheri quadrilateral are on parallel lines.

It may have been implied, but in the last chapter we never proved that the base and summit of a Saccheri quadrilateral are parallel.

Exercise 37.1 Let $\square ABCD$ be a Saccheri quadrilateral with base AB and summit CD. Prove that $\overleftrightarrow{AB}$ and $\overleftrightarrow{CD}$ are parallel.

In this section, we will explore parallels more systematically. By the end of it, we should have a better sense of what to look for in a non-Euclidean geometry. To start, let's see that non-Euclidean behavior is pervasive, and not just an anomaly that occurs at a few isolated points.

Theorem 37.1 *Let ℓ be a line and let P be a point that is not on ℓ. Then there are infinitely many lines through P that are parallel to ℓ.*

Proof There are two pieces to this proof. First that there are always at least two parallels. Then that there are infinitely many.

Part 1, that there are (at least) two parallels. Drop the perpendicular from P to ℓ, and mark its intersection with ℓ as Q. Let ℓ' be the line through P that is perpendicular to PQ. According to the alternate interior angle theorem, ℓ' is parallel to ℓ. Now suppose that it is the only parallel. To get a contradiction, we are going to build a Saccheri quadrilateral and see that it would have to be a rectangle. Pick a point on ℓ other than Q and label it R. On the line through R that is perpendicular to ℓ, mark the point S so that $RS \simeq PQ$ (choose S on the same side of ℓ as P). Then $\square PQRS$ is a Saccheri quadrilateral with base QR and summit PS. According to exercise 37.1, $\overline{PS}$ and $\overline{QR} = \ell$ are parallel. But there is only one line through P that is parallel to ℓ, namely ℓ'. Therefore $\overline{PS} = \ell'$. In $\square PQRS$, the summit angle at P is a right angle; and the other summit angle must be congruent to it. Therefore all four angles of $\square PQRS$ are right angles and it is a rectangle. There are no rectangles in non-Euclidean geometry, so there must be more than one parallel to ℓ that passes through P.

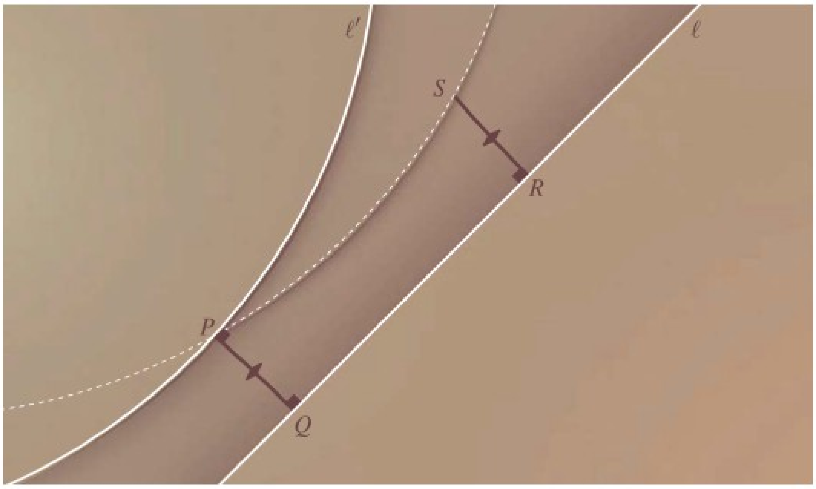

Figure 2. If there are no rectangles, then there must be at least two lines through P that are parallel to $\overline{QR}$.

Part 2, that there are infinitely many parallels. Let ℓ_1 and ℓ_2 be two lines through P that are parallel to ℓ. We can box-in P by surrounding it with more labeled points:

A, A' **:** on ℓ_1 so that $A * P * A'$,
B **:** on ℓ_2 and on the same side of PQ as A,
B' **:** on ℓ_2 and on the same side of PQ as A' (so $B * P * B'$).

Let C be a point on the segment AB. We will argue that $\overline{PC}$ is also parallel to ℓ. Suppose $\overline{PC}$ intersects ℓ at a point D. Then one of $\overrightarrow{PA}$, $\overrightarrow{PA'}$, $\overrightarrow{PB}$, or $\overrightarrow{PB'}$ would have to lie in the interior of

$\angle QPD$. According to the crossbar theorem, the ray would have to intersect QD. This cannot happen since ℓ_1 and ℓ_2 (which contain all four rays) are parallel to ℓ.

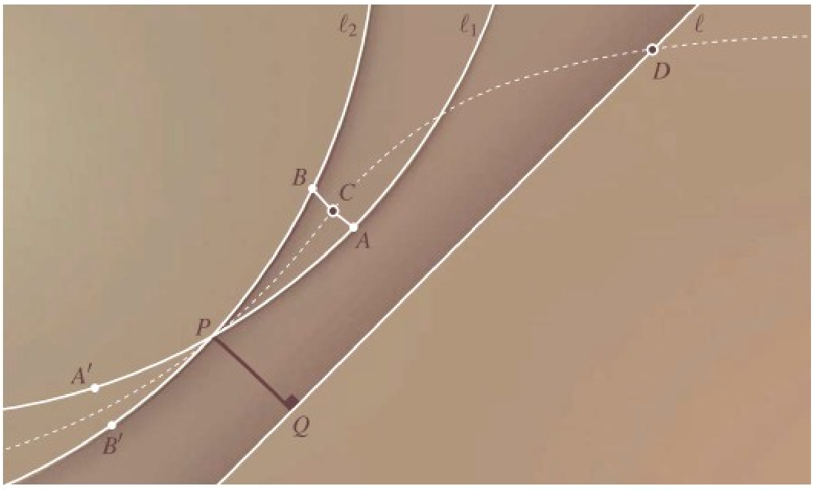

Figure 3. Given two parallels through P, we can find infinitely more between them. □

In Euclidean geometry, parallel lines are everywhere equidistant, so we can measure the distance between them with a single number. Non-Euclidean geometry is not so simple. Given parallel lines ℓ_1 and ℓ_2 in non-Euclidean geometry, the distance from a point P on ℓ_2 to the line ℓ_1 is measured as usual: drop a perpendicular from P to ℓ_1 and mark its foot (the intersection with ℓ_1) as F_P. Then the distance from P to ℓ_1 is $|PF_P|$.

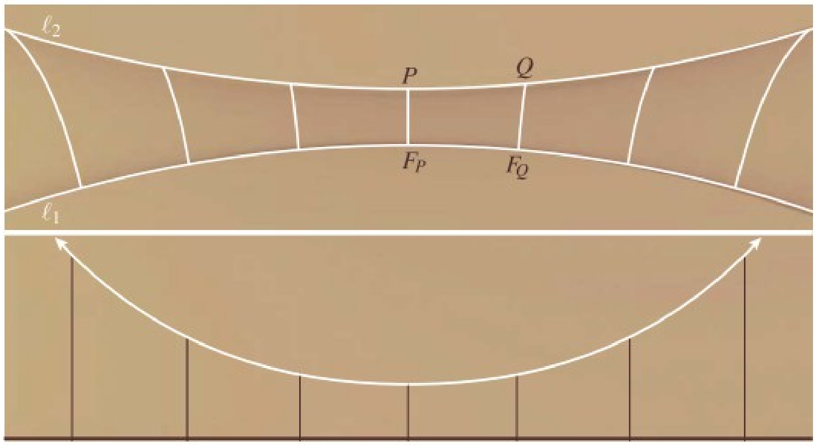

Figure 4. We can view the distance between parallel lines ℓ_1 and ℓ_2 as a function of the points on ℓ_2.

We can define a function $f : \ell_2 \to \mathbb{R}$ by setting $f(P)$ to be the distance from P to ℓ_1. The rest of this chapter investigates it. Before we go any further, we need the following lemma (note its similarity to the theorem in the last chapter that compared the length of the summit to the length of the base).

Lemma 37.2 *Let ℓ_1 and ℓ_2 be parallel lines, let P and Q be points on ℓ_2, and let F_P and F_Q be the respective feet of the perpendiculars on ℓ_1. Then $|F_PF_Q| < |PQ|$.*

Exercise 37.2 Prove the lemma.

Theorem 37.3 *The function f measuring the distance from points of ℓ_2 to the line ℓ_1 is continuous.*

Proof It is always possible to drop a perpendicular from a point P on ℓ_2 to the line ℓ_1, and it is always possible to measure the distance from P to its foot F_P, so the function f is always defined. For f to be continuous at the point P, we need a bit more: as a point Q on ℓ_2 approaches P, the value of $f(Q)$ must approach $f(P)$. To be more precise,

> For every $\epsilon > 0$, there must be a $\delta > 0$ so that if $|PQ| < \delta$ then $|f(P) - f(Q)| < \epsilon$.

The verification of this hinges on a few simple geometric observations that follow from the triangle inequality.

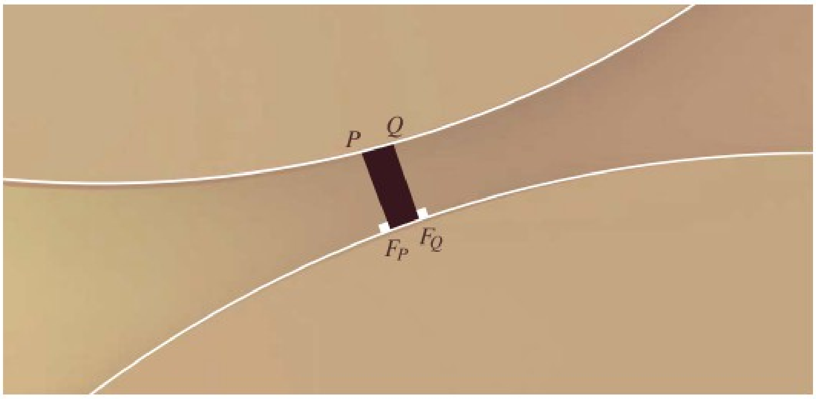

Figure 5. To show that f is continuous, consider two points P and Q that are close to one another.

- $|PF_P| \le |PQ| + |QF_Q| + |F_PF_Q|$.
- $|QF_Q| \le |PQ| + |PF_P| + |F_PF_Q| \implies |PF_P| \ge |QF_Q| - |PQ| - |F_PF_Q|$.
- Therefore

$$|QF_Q| - |PQ| - |F_PF_Q| \le |PF_P| \le |PQ| + |QF_Q| + |F_PF_Q|.$$

- Subtract $|QF_Q|$:

$$-|PQ| - |F_PF_Q| \le |PF_P| - |QF_Q| \le |PQ| + |F_PF_Q|.$$

- We don't know whether $|PF_P| > |QF_Q|$ or if it is the other way around. Either way, we can say that

$$\big||PF_P| - |QF_Q|\big| \leq |PQ| + |F_PF_Q|.$$

With that information, verifying continuity is straightforward. Fix $\epsilon > 0$ and put $\delta = \epsilon/2$. If $|PQ| < \delta = \epsilon/2$, then $|F_PF_Q| < \epsilon/2$ as well (according to lemma 37.2), and so

$$|f(P) - f(Q)| = \big||PF_P| - |QF_Q|\big| \leq |PQ| + |F_PF_Q| < \epsilon/2 + \epsilon/2 = \epsilon.$$

Therefore f is continuous at P. Since P was an arbitrary point on ℓ_2, f is continuous on all of ℓ_2. □

Theorem 37.4 *For a real number r, $|\{f^{-1}(r)\}| \leq 2$. That is, there are at most two points on ℓ_2 that are at a distance r from ℓ_1.*

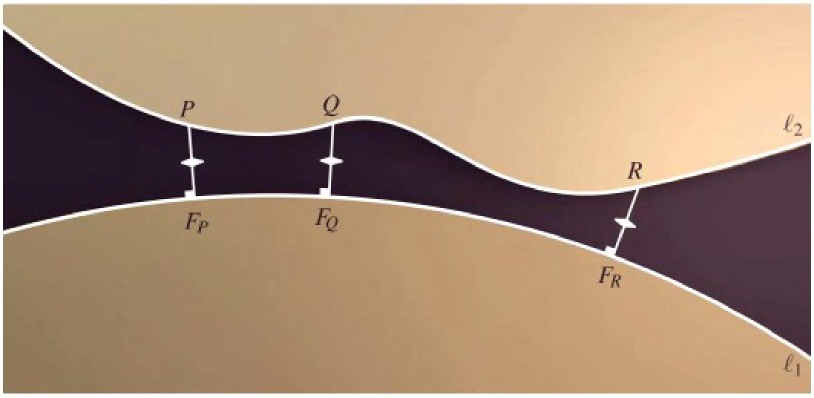

Figure 6. It is not possible for all of P, Q, and R to be the same distance from ℓ_1.

Proof Suppose that there is a real number r so that $|\{f^{-1}(r)\}| > 2$. Then there are (at least) three distinct points P, Q, and R on ℓ_2 with $P * Q * R$ so that $f(P) = f(Q) = f(R) = r$. If we label the feet of their respective perpendiculars as F_P, F_Q, and F_R, then

$$|PF_P| = |QF_Q| = |RF_R| = r.$$

Thus $\square F_PF_QQP$ and $\square F_QF_RRQ$ are Saccheri quadrilaterals. They cannot be rectangles, though, so their summit angles $\angle PQF_Q$ and $\angle RQF_Q$ must be acute. But the angles are supplementary so they cannot both be acute. □

Theorem 37.5 *The function f has no local maxima and at most one local minimum.*

Proof Suppose that f has a local maximum at a point B. Then, since f is continuous, for a sufficiently small value positive value ϵ, there are points A and C on either side of B so that

$$f(A) = f(C) = f(B) - \epsilon.$$

Let F_A, F_B, and F_C be the feet of the perpendiculars to ℓ_1 from A, B, and C respectively. Then

$$|AF_A| = f(A) = f(C) = |CF_C|$$

so $\square F_A F_C C A$ is a Saccheri quadrilateral with base $F_A F_C$. The segment BF_B connects the summit to the base and is perpendicular to the base. We saw at the end of the last chapter that such segments cannot be longer than the legs. But the legs of this Saccheri quadrilateral are AF_A and CF_C, and they are shorter (by ϵ) than BF_B. This is a contradiction, and so f cannot have a local maximum.

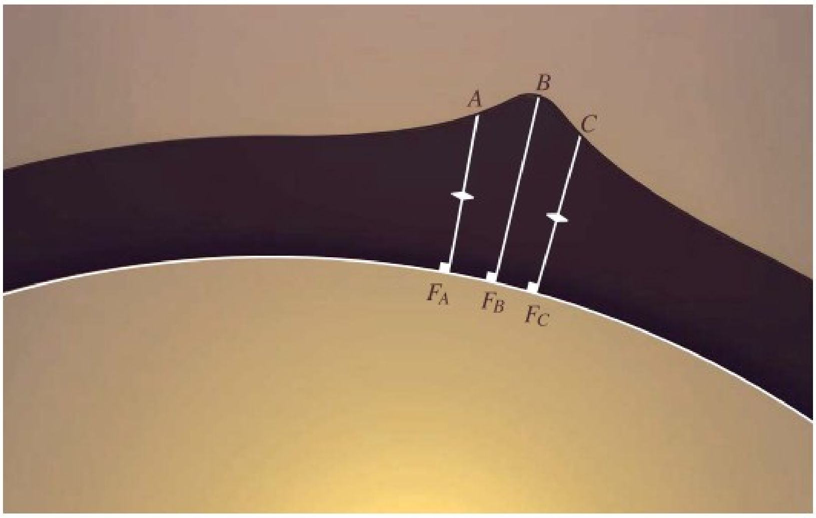

Figure 7. It is not possible for f to have a local maximum (as is suggested here at the point B).

The second part of the theorem follows immediately: if f had more than one minimum, then, because it is continuous, it would have a local maximum in between them. ☐

Theorem 37.6 *Let P and Q be points on ℓ_2 and suppose that $f(Q) > f(P)$. Then f approaches infinity along $\overrightarrow{PQ}$.*

It should be clear that the value of f continues to increase past Q. If at some point it began to decrease, then we would have found a local maximum, and we have just seen that f has none. But could it be that f continues to grow, but is bounded above? Is there an upper bound that f approaches in the limit but never surpasses? This is a more delicate question. The strategy of the proof is to build a wall of congruent Saccheri quadrilaterals that appears to contain $\overrightarrow{PQ}$, and then see why that cannot happen.

Proof Suppose that f does have an upper bound so that it approaches but never exceeds a number k. Starting from the point F_Q, the foot of the perpendicular from Q, mark off congruent segments along ℓ_1 on the same side of PF_P as $\overrightarrow{PQ}$. That is, label points F_i so that

$$F_Q = F_0 * F_1 * F_2 * F_3 * \cdots$$

and so that the segments F_iF_{i+1} are congruent. Above each point F_i (on the same side of ℓ_1 as ℓ_2), erect a perpendicular segment F_iP_i of length k. Then each $\square F_iF_{i+1}P_{i+1}P_i$ is a Saccheri quadrilateral, and by SASAS they are all congruent.

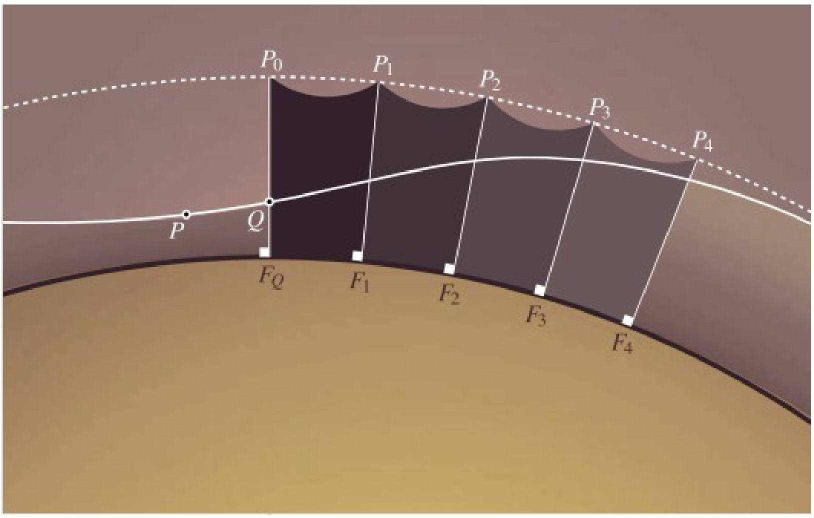

Figure 8. It is not possible for f to level off at some asymptotic value, as suggested here.

Their altitudes are shorter than their legs, by an amount that we can call ϵ (since the quadrilaterals are congruent, the discrepancy will be the same for all of them). Now let's get back to $\overrightarrow{PQ}$. For a while, it may run below the summits of the quadrilaterals we have assembled. But f has a limiting value of k along it, so eventually there is a point beyond which f is greater than $k - \epsilon$. Look beyond that point, at the interaction of $\overrightarrow{PQ}$ with one of the summits P_nP_{n+1}. Points P_n and P_{n+1} are on the opposite side of ℓ_2 from ℓ_1. But the altitude between them is on the same side of ℓ_2 as ℓ_1. This would imply that ℓ_2 intersects P_nP_{n+1} twice. Two distinct lines can intersect only once, so this is a contradiction. ☐

Looking in the other direction, if $f(Q) > f(P)$, then $f(P) < f(Q)$, and it is natural to wonder what happens as we progress along $\overrightarrow{QP}$. Does the distance to ℓ_2 continue to decrease, or does it at some point begin to increase again?

Definition 37.7 If f achieves a minimum, then ℓ_1 and ℓ_2 are called *divergent* parallels. If instead f approaches zero, then ℓ_1 and ℓ_2 are called *asymptotic* parallels.

If ℓ_1 and ℓ_2 are divergent parallels, then once $\overrightarrow{QP}$ has crossed over the minimum value, the value of f will begin to go up. Once that happens, by the previous result, f will go to infinity. Both types of parallels exist in non-Euclidean geometry. The base and summit of a Saccheri quadrilateral are divergent parallels. Asymptotic parallels are more elusive, but one of the exercises outlines a proof of their existence. Beyond that, there is one more obvious question: are there any parallels that are neither divergent nor asymptotic? Suppose ℓ_1 and ℓ_2 are not divergent. Then f does not achieve its minimum. It is easy then to pick points P and Q

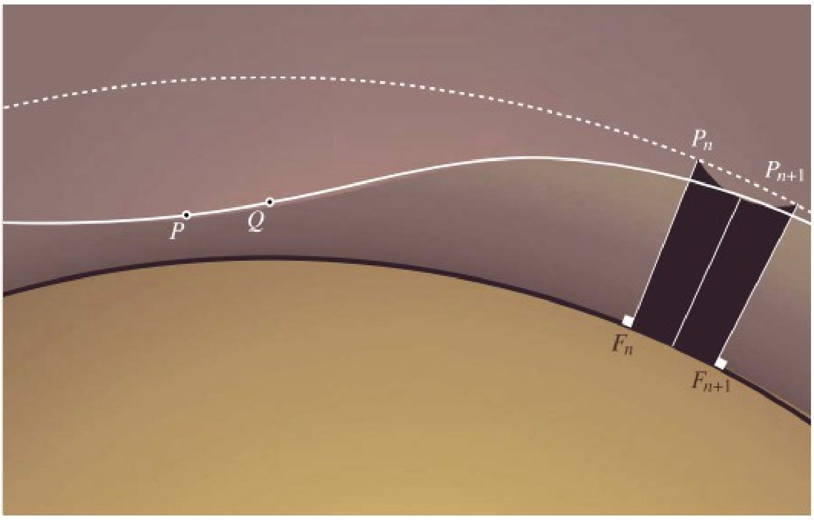

Figure 9. If $\overrightarrow{PQ}$ approached an asymptote, it would eventually intersect the summit of a constructed Saccheri quadrilateral more than once.

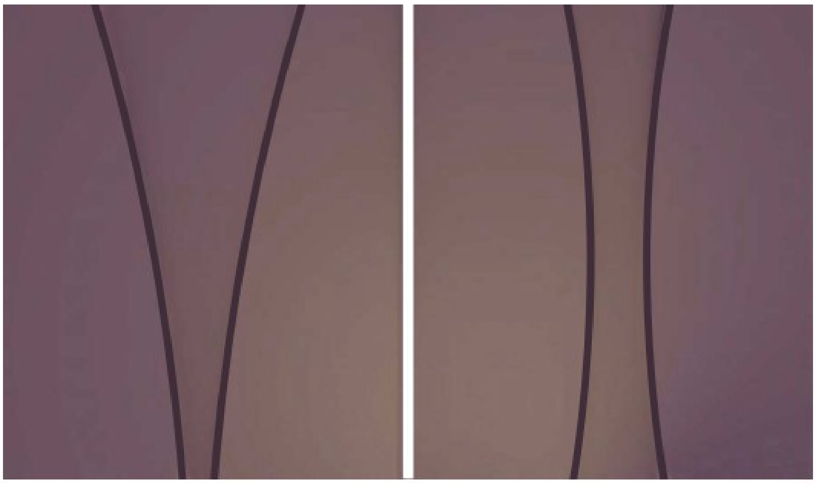

Figure 10. (l) asymptotic parallels; (r) divergent parallels.

so that $f(P) < f(Q)$, so f must continue to decrease past P as we move out along $\overrightarrow{QP}$. By definition, the lines are asymptotic if the value of f decreases to zero. Is it possible that the value of f is always decreasing, but never drops below some positive number k? No. The argument against a nonzero lower bound is essentially the same as the argument against an upper bound in the proof above.

Theorem 37.8 *If ℓ_1 and ℓ_2 are parallel, then they are either divergent or asymptotic.*

Exercise 37.3 Prove theorem 37.8.

The results of this chapter should give you some sense of the look of non-Euclidean geometry. But can such a geometry exist? If we want to answer YES, we need to find a model that meets all the requirements of neutral geometry, and has the non-Euclidean characteristics. In the next chapters, we will look at such models.

37.1 Exercises

37.4. Let r_1 and r_2 be two non-opposite rays with endpoint A. Label point B and C on r_1 and B' and C' on r_2 so that $A * B * C$ and $A * B' * C'$. Prove that the angle sum of $\triangle ACC'$ is less than the angle sum of $\triangle ABB'$.

37.5. Suppose that ℓ_1 and ℓ_2 are divergent parallels. Let P be the point on ℓ_2 that is the minimum distance from ℓ_1, and let Q be the foot of the perpendicular from P. Prove that QP intersects ℓ_2 at a right angle.

37.6. Let ℓ be a line and let P be a point that is not on ℓ. Prove that there is a line through P which is asymptotic to ℓ. Hint: let r_1 and r_2 be rays from P, one that intersects ℓ and one that is along a divergent parallel to ℓ. Let A be a point on r_1 and B be a point on r_2. A point Q on AB falls into one of two sets: either PQ is parallel to ℓ or it isn't. By Dedekind's axiom, there is a point Q_0 that separates those two sets. Show that PQ_0 is an asymptotic parallel.

37.7. Suppose that lines ℓ_1 and ℓ_2 intersect at a point O. Suppose that P and Q are on ℓ_2 so that $O * P * Q$, and that F_P and F_Q are the feet of the perpendiculars, from P and Q respectively, to ℓ_1. Prove that $|PF_P| < |QF_Q|$.

37.8. Let r_1 and r_2 be two rays with a common endpoint O. Prove that for a positive real number x, there is a point P on r_1 so that $d(P, r_2) > x$.

;. : 5: 6 $ 3 1

38

The Pseudosphere

In time, some geometers began to doubt that Playfair's axiom could be proved using only the axioms of neutral geometry. So they began to search for a model of a neutral geometry that would exhibit non-Euclidean behavior. In such a geometry parallel lines would have to bend away from one another. Viewed from a Euclidean perspective, the non-Euclidean plane would look warped. One idea was to use a curved surface as the model. In it, points would be modeled by the points on the surface, while lines will be modeled by geodesics on the surface. The distance between two points would be measured as the arc length of the geodesic connecting them, and the angle between two geodesics would be the angle between their tangent vectors at the point of intersection. The undefined terms *on*, *between*, and *congruent* would then be defined in the obvious way.

A geodesic is (roughly speaking) the shortest path between two points. For example, the geodesics on a plane are lines. Geodesics on a sphere are *great circles*, circles on the sphere whose centers coincide with the center of the sphere. Neither the plane nor the sphere works as a model for non-Euclidean geometry. The plane is flat and the sphere, while warped, seems to be warped in the wrong way because the great circles bend toward one another rather than away.

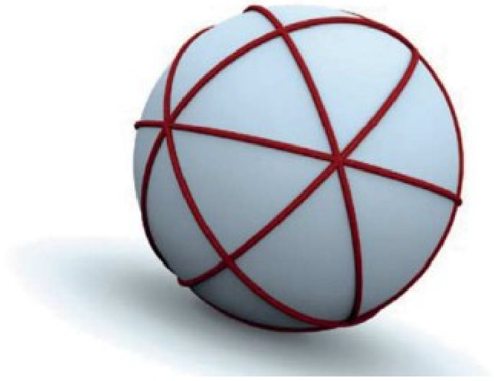

Figure 1. Some geodesics (great circles) on a sphere.

There are, however, some surfaces that exhibit the right kind of warping, those with constant negative Gaussian curvature. One is called the pseudosphere. The purpose of this chapter is to find equations for it. We will have to use differential geometry in order to do so, so this chapter is advanced. My primary reference for this is Spivak's five volume Differential Geometry series, the calculations are in volume 3 [Spi99b].

We will get into the details as this chapter unfolds, but there are two parts to a surface of constant negative curvature: the constant part and the negative part. The curvature has to be constant so that the model will be "isotropic", that is it looks the same at every point. Without this requirement, the surface won't satisfy all of the neutral axioms. The curvature has to be negative so that parallels will bend away from one another as we would expect in a non-Euclidean geometry.

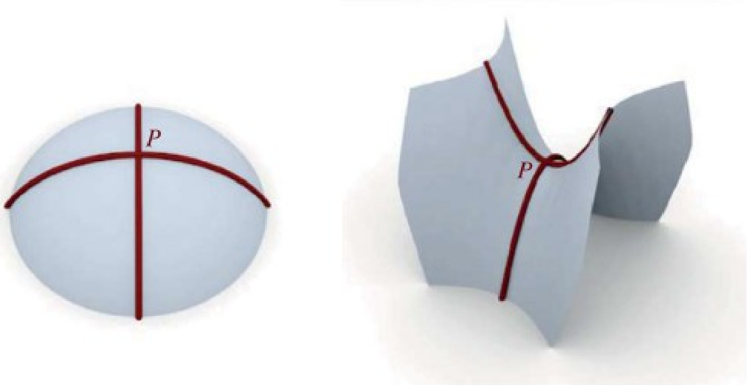

Figure 2. The curves are geodesics on the surfaces. On the left, P is a point of positive curvature. On the right, P is a point of negative curvature. For negatively curved surfaces, the geodesics bend away from each other.

38.1 Surfaces

There are several ways to describe a surface in $\mathbb{R}^3$. Perhaps the most elementary is when one variable can be written as a function of the other two such as $z = f(x, y)$. Because that form can be limiting, a more flexible approach is to parametrize the surface, writing coordinates as a function of two variables:

$$X(s, t) = (f(s, t), g(s, t), h(s, t)).$$

Given a surface S parametrized by $X(s, t)$, assuming that the surface is relatively smooth and that the parametrization function is relatively well-behaved, we can compute partial derivatives

$$\begin{aligned} X_s &= \left.\frac{\partial X}{\partial s}\right|_{(s_0,t_0)} = \langle f_s(s_0, t_0),\ g_s(s_0, t_0),\ h_s(s_0, t_0)\rangle\,, \\ X_t &= \left.\frac{\partial X}{\partial t}\right|_{(s_0,t_0)} = \langle f_t(s_0, t_0),\ g_t(s_0, t_0),\ h_t(s_0, t_0)\rangle \end{aligned}$$

at a point $P = X(s_0, t_0)$ on S. The results are vectors that point in the direction of an infinitesimal change in s and an infinitesimal change in t. They form a basis for T_P, the *tangent plane* to S at P.

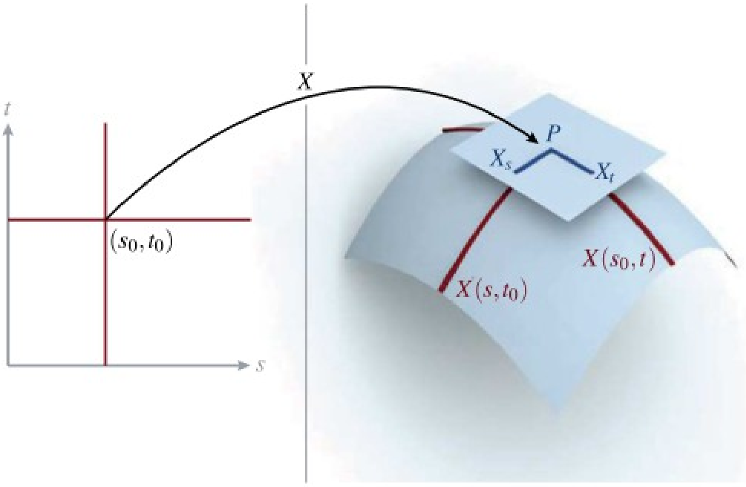

Figure 3. The parametrization X maps the st-domain space to the surface. The partial derivatives $X_s(s_0, t_0)$ and $X_t(s_0, t_0)$ form a basis for the tangent plane to the surface at P.

In the domain space, the line through (s_0, t_0) in the direction of the vector $\langle a, b\rangle$ can be parametrized as

$$\rho(w) = (s_0 + aw, t_0 + bw).$$

Then $X(\rho(w))$ is a curve on S,

$$X(\rho(w)) = \Big(f(s_0 + aw, t_0 + bw),\ g(s_0 + aw, t_0 + bw),\ h(s_0 + aw, t_0 + bw)\Big),$$

and the tangent vector at P can be computed by the chain rule:

$$\begin{aligned}[X(\rho(w))]' &= \langle f_s \cdot a + f_t \cdot b,\ g_s \cdot a + g_t \cdot b,\ h_s \cdot a + h_t \cdot b\rangle \\ &= a\langle f_s, g_s, h_s\rangle + b\langle f_t, g_t, h_t\rangle \\ &= aX_s + bX_t.\end{aligned}$$

Therefore, the vectors of the tangent plane T_P are all tangent vectors to parametrized curves on S that pass through P. So T_P is the best linear approximation to S in an infinitesimally small neighborhood around P.

Let's turn now to specifics. There are several surfaces that meet the constant negative curvature requirement. The one we are after, the pseudosphere, is relatively easy to work with because it is a surface of revolution. Start with a curve $\mathcal{C}$ in the yz-plane, parametrized as

$$r(t) = \big(0,\ y(t),\ z(t)\big).$$

It will simplify calculations if we place another restriction on the parametrization, and require that $\mathcal{C}$ be parametrized by arc length. That is, choose a parametrization that moves along $\mathcal{C}$ at a constant speed of 1, so that

$$\sqrt{[y'(t)]^2 + [z'(t)]^2} = 1.$$

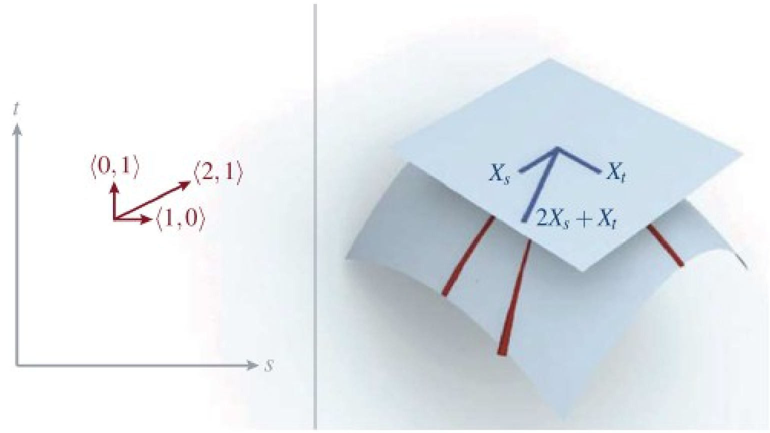

Figure 4. Any tangent vector in the tangent plane can be written as a linear combination of X_s and X_t.

Revolving $\mathcal{C}$ around the y-axis will generate a surface of revolution that can be parametrized as

$$X(s,t) = \big(z(t)\sin s,\ y(t),\ z(t)\cos s\big), \quad 0 \le s \le 2\pi.$$

At a point on this surface, the partial derivatives are

$$\begin{aligned} X_s &= \Big\langle z(t)\cos s,\ 0,\ -z(t)\sin s\Big\rangle \\ X_t &= \Big\langle z'(t)\sin s,\ y'(t),\ z'(t)\cos s\Big\rangle. \end{aligned}$$

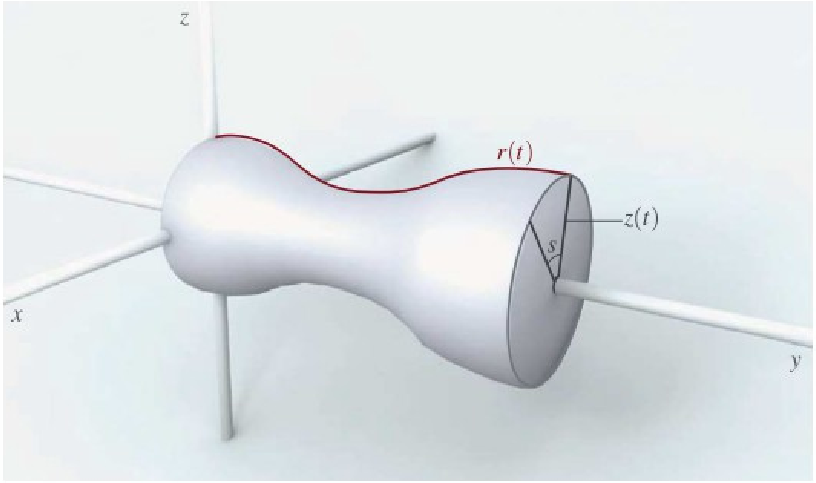

Figure 5. Form a surface by revolving the curve $r(t) = \langle 0, y(t), z(t)\rangle$ around the y-axis.

38.2 Maps between surfaces. The Gauss map

Because surfaces are now the objects of study, we consider maps between them. Let $f : S_1 \to S_2$ be a map from the points of a surface, S_1, to the points of another surface, S_2. Let P be a point on the first surface. The tangent plane T_P is the best linear approximation to S_1 around P. Likewise, on S_2, the tangent plane $T_{f(P)}$ is the best linear approximation to S_2 around $f(P)$. We want to define a linear map between the tangent planes $df : T_P \to T_{f(P)}$ that best approximates f near the points.

Let $X : D \to S_1$ be a parametrization of S_1 with $X(s_0, t_0) = P$. Every vector in T_P can be written as a linear combination $aX_s + bX_t$ and is therefore the tangent vector to the curve $X(\rho(w))$ where $\rho(w) = (s_0 + aw, t_0 + bw)$. The composition $f \circ X : D \to S_2$ is not likely to be a parametrization of S_2 since it will not typically be a bijection. Nevertheless, we can calculate the partial derivatives $(f \circ X)_s$ and $(f \circ X)_t$. A parametrized line $\rho(w)$ in D is mapped to a curve $f \circ X(\rho(w))$ on S_2, whose tangent vector at $f(P)$ is (by the chain rule)

$$\begin{aligned}[f \circ X(\rho(w))]' &= (f \circ X)_s \cdot \frac{ds}{dw} + (f \circ X)_t \cdot \frac{dt}{dw} \\ &= (f \circ X)_s \cdot a + (f \circ X)_t \cdot b.\end{aligned}$$

We define the differential of f at P, $df : T_P \to T_{f(P)}$, by

$$df(aX_s + bX_t) = a(f \circ X)_s + b(f \circ X)_t.$$

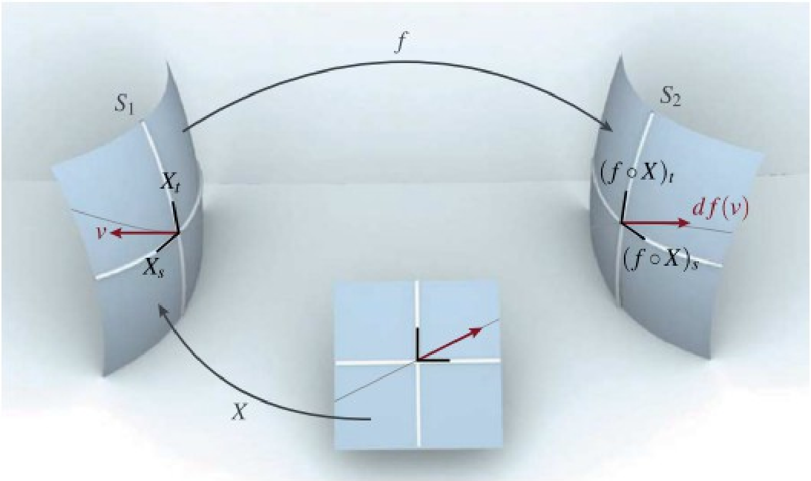

Figure 6. A map $f : S_1 \to S_2$ between two surfaces S_1 and S_2. The differential df is the map between tangent planes.

Gauss's idea was to use a particular map between surfaces, now called the Gauss map, to study curvature. The basic idea is simple and elegant. Imagine moving in a small neighborhood around a point P on a surface S, with an eye on the tangent plane as you do so. The tangent plane balances on the surface like a surfboard, and so as the surface bends, the tangent plane must turn. The more the tangent plane has to turn to keep up with the surface, the more highly curved the surface is. We can measure how fast the tangent plane is turning by looking at the rate of change of its unit normal vector (it is tempting to try to extend the surfboard analogy, and to call the unit normal vector the surfer on the surfboard, but that assumes the surfer is always perpendicular to the board).

Definition 38.1 Let $\mathbb{S}$ denote the sphere of radius one centered at the origin. For a point P on a smooth surface S, let $n(P)$ be the unit normal vector to S at P. If the tail of $n(P)$ is translated to the origin, then its head will be at a point $\nu(P)$ on $\mathbb{S}$. The *Gauss map* $\nu : S \to \mathbb{S}$ is defined by $P \mapsto \nu(P)$.

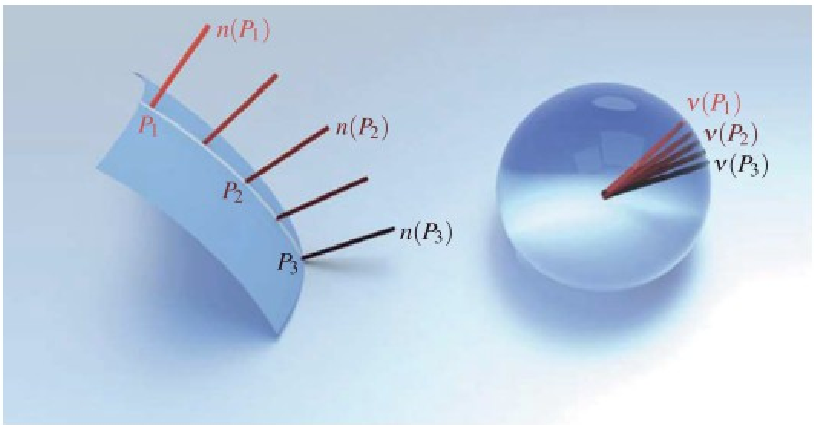

Figure 7. The Gauss map $\nu : S \to \mathbb{S}$.

The Gauss map may be complicated. However, the curvature at a point P only depends on what is happening near P, and there ν can be replaced by its linear approximation $d\nu$. The differential of the Gauss map has the advantage over the differentials of other maps that the tangent planes T_P and $T_{\nu(P)}$ are parallel. Because they have the same vectors in them, $\{X_s, X_t\}$ is a basis of T_P and of $T_{\nu(P)}$. In it, $d\nu$ can be written as a 2×2 matrix.

Let's now compute the Gauss map and its differential for our surface of revolution. To do that, we need to know the unit normal vector. The cross product of X_s and X_t is perpendicular to both of them, so it is a normal vector to the tangent plane:

$$\begin{aligned} N &= X_s \times X_t \\ &= \begin{vmatrix} i & j & k \\ z(t)\cos s & 0 & -z(t)\sin s \\ z'(t)\sin s & y'(t) & z'(t)\cos s \end{vmatrix} \\ &= \left\langle y'(t)z(t)\sin s,\ -z(t)z'(t)\cos^2 s - z(t)z'(t)\sin^2 s,\ y'(t)z(t)\cos s \right\rangle \\ &= \left\langle y'(t)z(t)\sin s,\ -z(t)z'(t),\ y'(t)z(t)\cos s \right\rangle. \end{aligned}$$

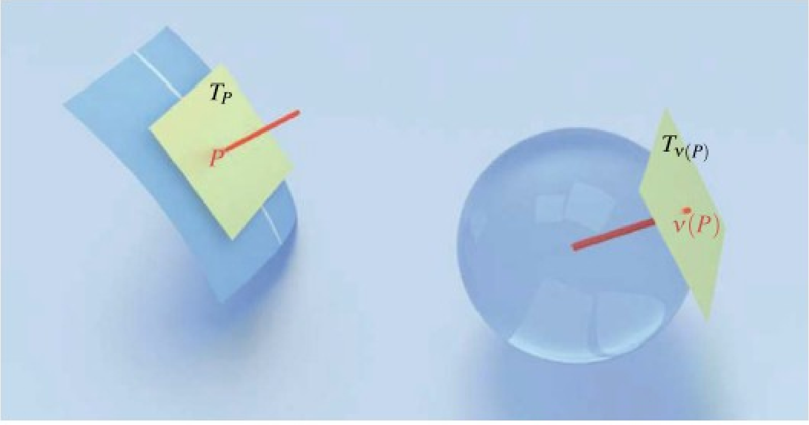

Figure 8. The differential of the Gauss map.

It may not be a unit vector, so we need to divide by its length

$$\begin{aligned}|N| &= \sqrt{\big(y'(t)z(t)\sin s\big)^2 + \big(-z(t)z'(t)\big)^2 + \big(y'(t)z(t)\cos s\big)^2}\\ &= \sqrt{\big(y'(t)z(t)\big)^2 + \big(-z(t)z'(t)\big)^2}\\ &= z(t)\sqrt{\big(y'(t)\big)^2 + \big(z'(t)\big)^2}\\ &= z(t).\end{aligned}$$

The last step takes advantage of the fact that $\mathcal{C}$ is parametrized by arc length. The unit normal is then

$$n = \frac{N}{|N|} = \Big\langle y'(t)\sin s,\ -z'(t),\ y'(t)\cos s\Big\rangle.$$

This calculation gives us the Gauss map as well. If $P = X(s,t)$, then

$$\nu(P) = \big(y'(t)\sin s,\ -z'(t),\ y'(t)\cos s\big).$$

Now let's find the $d\nu$ matrix. Compute the partial derivatives,

$$\begin{aligned}(\nu\circ X)_s &= \Big\langle y'(t)\cos s,\ 0,\ -y'(t)\sin s\Big\rangle\\ (\nu\circ X)_t &= \Big\langle y''(t)\sin s,\ -z''(t),\ y''(t)\cos s\Big\rangle.\end{aligned}$$

We rewrite them as linear combinations of the basis elements X_s and X_t, where

$$\begin{aligned}X_s &= \Big\langle z(t)\cos s,\ 0,\ -z(t)\sin s\Big\rangle\\ X_t &= \Big\langle z'(t)\sin s,\ y'(t),\ z'(t)\cos s\Big\rangle.\end{aligned}$$

The first is easy:

$$\begin{aligned}(\nu \circ X)_s &= \langle y'(t)\cos s,\ 0,\ -y'(t)\sin s\rangle \\ &= \frac{y'(t)}{z(t)} \cdot \Big\langle z(t)\cos s, 0, -z(t)\sin s\Big\rangle \\ &= \frac{y'(t)}{z(t)} \cdot X_s.\end{aligned}$$

The second requires a trick. Because $\mathcal{C}$ is parametrized by arc length,

$$[y'(t)]^2 + [z'(t)]^2 = 1.$$

If we differentiate with respect to t and solve for y'', we get

$$2y'(t)y''(t) + 2z'(t)z''(t) = 0 \implies y''(t) = -\frac{z'(t)z''(t)}{y'(t)}.$$

Substitute to get

$$\begin{aligned}(\nu \circ X)_t &= \Big\langle y''(t)\sin s,\ -z''(t),\ y''(t)\cos s\Big\rangle \\ &= \Big\langle -\frac{z'(t)z''(t)}{y'(t)}\sin s,\ -z''(t),\ -\frac{z'(t)z''(t)}{y'(t)}\cos s\Big\rangle \\ &= -\frac{z''(t)}{y'(t)} \cdot \Big\langle z'(t)\sin s,\ y'(t),\ z'(t)\cos s\Big\rangle \\ &= -\frac{z''(t)}{y'(t)} \cdot X_t.\end{aligned}$$

Therefore

$$\begin{aligned}d\nu(aX_s + bX_t) &= a(\nu \circ X)_s + b(\nu \circ X)_t \\ &= a(y'/z)X_s + b(-z''/y')X_t.\end{aligned}$$

In the $\{X_s, X_t\}$ basis, $d\nu$ has the matrix form

$$d\nu \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} y'/z & 0 \\ 0 & -z''/y' \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}.$$

38.3 Gaussian curvature

The *Gaussian curvature* $K(P)$ of S at the point P is defined to be the determinant of the $d\nu$ matrix at P. In our case

$$K(P) = -\frac{y'(t)}{z(t)} \cdot \frac{-z''(t)}{y'(t)} = -\frac{z''(t)}{z(t)}.$$

What is the geometric significance of that number?

Why we want K to be negative. The $d\nu$ matrix is a symmetric matrix, which is obvious in our example, and true in general. Therefore it has real eigenvalues, and K, its determinant, is their product. If we apply the $d\nu$ map to the circle of unit vectors $\langle\cos\theta, \sin\theta\rangle$, we get a set of vectors that trace out an ellipse. Because the dv matrix is symmetric, the eigenvectors are along the major and minor axes of the ellipse, and the corresponding eigenvalues are the signed distances from the origin to the ellipse along the axes. This is true in general, and easy to see in our example. The eigenvectors point in the directions of maximum and minimum curvature, then, and the corresponding eigenvalues are the curvatures in those directions (they are called the principal curvatures). If K is negative, then one of the eigenvalues is positive and the other is negative. So in the direction of one eigenvector, the surface will bend towards the normal vector $n(P)$; in the direction of the other, the surface will bend away from $n(P)$. Hence, negative K will make the surface bend so that lines on it diverge.

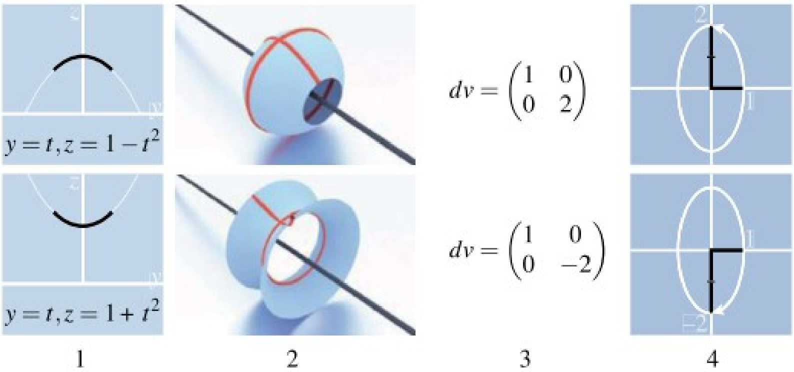

$$dv = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$$

$$dv = \begin{pmatrix} 1 & 0 \\ 0 & -2 \end{pmatrix}$$

Figure 9. (1) A curve is (2) revolved around the y-axis. We calculate (3) the dv matrix, and can then (4) see its effects on the unit circle.

Why we want $|K|$ to be constant. The magnitude of $\det(dv)$ is the factor by which dv scales area. The dv map will transform a 1×1 square into a parallelogram with an area of $|\det(dv)|$, and since dv is linear, all other areas will follow suit. In our example, the parallelogram is a rectangle, so this is clear. In a small region (say of size ϵ) near P, the dv map is a good approximation of v, and so v also approximately scales area by $|\det(dv)|$. To get a sense of the scaling near P, consider the circle of points that are a distance of ϵ from P on the surface—a "circle" on the surface—and look at its circumference. In the plane that circumference is $2\pi\epsilon$, but on a curved surface it is not. The ratio depends on the curvature at P. Intuitively that is clear because if $|\det(dv)|$ is large, then v will wrap the circle farther around $\mathbb{S}$ than if it is small. For instance, consider the points P and Q shown in figure 10.

Both are points of positive curvature, but the curvature at P is larger than the curvature at Q. Let C_P and C_Q be circles of radius ϵ, centered at P and Q respectively. While both circles have the same radius, the circumference of C_P is smaller than the circumference of C_Q. As a model for neutral geometry, this would create the following problem. The circumference of a circle is the limit of perimeters of regular inscribed polygons. Since C_P and C_Q have the same radius, a regular n-gon inscribed in C_P is congruent to a regular n-gon inscribed in C_Q. Hence their perimeters would be the same, and so the circumferences of C_P and C_Q, as limits of them, should be the same (the Bertrand-Diquet-Puiseux theorem gives a precise relationship between the curvature and the circumference/radius ratio and that is a place to look if you want a more rigorous explanation [Spi99a]).

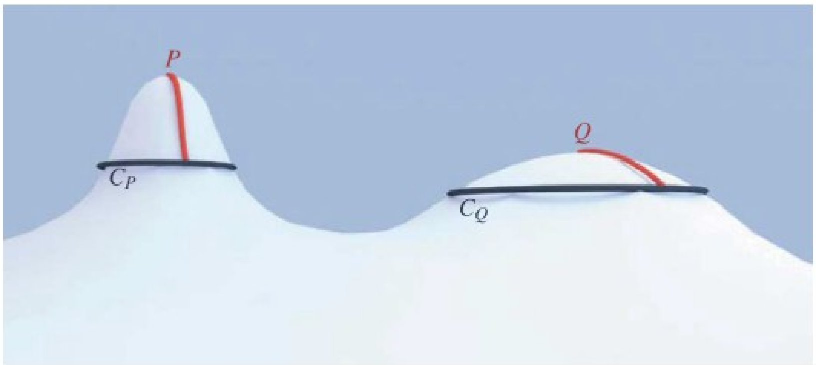

Figure 10. The curvature is greater at P than it is at Q. A circle centered at P of radius ϵ would have a smaller circumference than a circle centered at Q with the same radius.

38.4 The tractrix and pseudosphere

So it comes down to this: we want the curvature K to be a negative constant, say $K = -1$. That is,

$$-\frac{z''(t)}{z(t)} = -1 \implies z''(t) = z(t).$$

One solution to this differential equation (not the only one) is $z(t) = e^{t}$. What then, is $y(t)$? We get

$$\begin{aligned} (y'(t))^2 + (z'(t))^2 &= 1 \\ (y'(t))^2 + e^{2t} &= 1 \\ y'(t) &= -\sqrt{1 - e^{2t}} \\ y(t) &= \int -\sqrt{1 - e^{2t}}\, dt \end{aligned}$$

(the solution curve comes in two pieces corresponding to the positive and negative roots and the negative root is the one that gives the standard piece). We leave evaluating the integral as an exercise.

Exercise 38.1 Show that

$$\int -\sqrt{1 - e^{2t}}\, dt = \ln\left(\frac{1 + \sqrt{1 - e^{2t}}}{e^t}\right) - \sqrt{1 - e^{2t}} + C.$$

Any value for C will work, so put $C = 0$. For the term inside the square root to be positive, t must be between $-\infty$ and zero. The result is a parametrization of $\mathcal{C}$. Let's go ahead and do one more step and write y directly in terms of z. It is easy since $z = e^t$:

$$y = \ln\left(\frac{1 + \sqrt{1 - z^2}}{z}\right) - \sqrt{1 - z^2}, \quad 0 < z \le 1.$$

The curve is called the *tractrix*. Its graph is shown in figure 11. Rotating the tractrix around the y-axis generates the pseudosphere, shown in figure 12.

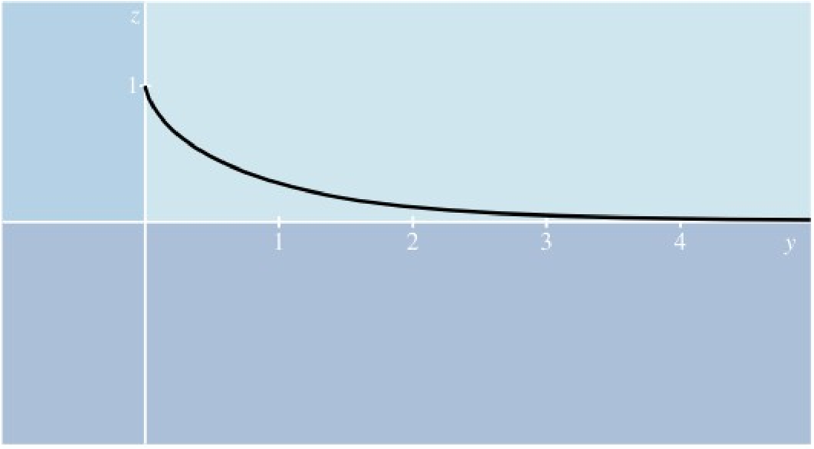

Figure 11. The graph of the tractrix (for positive y).

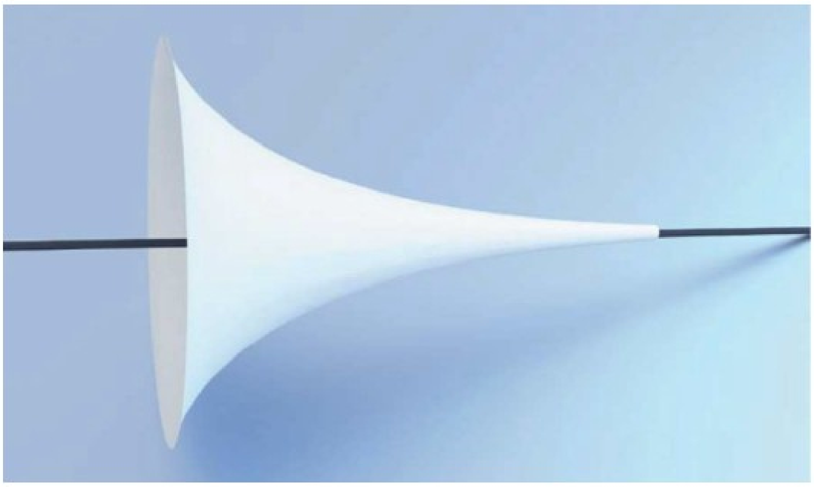

Figure 12. The pseudosphere.

The pseudosphere models a non-Euclidean neutral geometry locally, but globally there are some problems. One is apparent already: the pseudosphere ends at $y = 0$ where the tractrix has a vertical tangent. A neutral geometry plane cannot just stop like that. It is probably reasonable to hope that some other $K = -1$ surface (not a surface of revolution) would avoid this problem, but unfortunately this is not the case. Hilbert proved that no complete regular surface with constant negative curvature can be smoothly embedded in $\mathbb{R}^3$ [Spi99b]. In spite of its limitations, we will stick with the pseudosphere for a few more chapters. It will show the way to another model without the limitations.

38.5 Exercises

38.2. A torus can be realized as a surface of revolution formed by revolving a circle with radius a and center $(b, 0)$ around the y-axis (where $b > a$). Parametrize the surface.

38.3. Verify that the Gaussian curvature of the plane $z = 0$ is zero.

38.4. Show that the cylinder $x^2 + y^2 = 1$ has Gaussian curvature of zero.

38.5. Show that the Gaussian curvature of a sphere of radius r is $1/r^2$ (use spherical coordinates).

38.6. Compute the Gaussian curvature of the saddle surface $z = x^2 - y^2$ at the point $(0, 0, 0)$.

38.7. Let $P = (y_0, z_0)$ be a point on the tractrix. Let ℓ be the tangent line to the tractrix at P, and let $Q = (y_1, 0)$ be its intersection with the y-axis. Show that the distance from P to Q is 1 (so the distance from P to Q is independent of the coordinates of P).

39

Geodesics on the Pseudosphere

In the last chapter, we found a parametrization of the pseudosphere, a surface with constant negative curvature. The plan was to use that surface as a model for a non-Euclidean neutral geometry in which lines are represented by geodesics along the surface. The goal of this chapter is to calculate the geodesics.

39.1 Geodesics, the theory

Roughly, a geodesic between two points on a surface is the shortest path on the surface that connects them. To be more precise, we will talk about the calculus of variations. Let S be a surface parametrized by $X : D \to \mathbb{R}^3$ where

$$X(s,t) = (f(s,t), g(s,t), h(s,t))$$

and consider a parametrized curve $\rho(w)$ in the domain D (it can then be projected onto the surface S by X). We want to perturb $\rho(w)$ to see if the arc length of the curve on S increases or decreases. If the arc length always increases, no matter the variation, then $\rho(w)$ is a geodesic. What are the mechanics of perturbing?

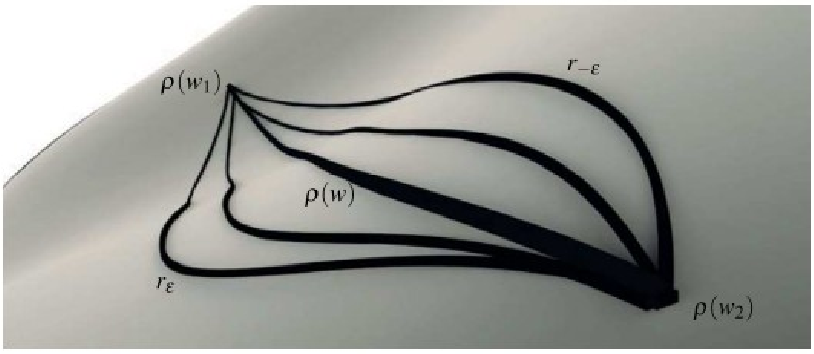

Figure 1. Some variations of the curve $\rho(w)$.

Definition 39.1 Let $\rho : [w_1, w_2] \to D$ be a parametrized curve in a domain D. A family of curves in D

$$\{r_v : [w_1, w_2] \longrightarrow D\}, \quad -\epsilon < v < \epsilon,$$

is a (smooth) *variation of* ρ if $R(w, v) = r_v(w)$ is a smooth map that satisfies:

- $R(w_1, v) = \rho(w_1)$ for all v; i.e., they all start at the same point,
- $R(w_2, v) = \rho(w_2)$ for all v; i.e., they all end at the same point,
- $R(w, 0) = \rho(w)$ (at $v = 0$, the original curve, unperturbed).

In general the length of a parametric curve $r(t)$ is calculated by integrating $|r'(t)|$. In our situation, the X map projects r_v of a variation onto the surface S, where we can calculate its arc length L as a function of v:

$$L(v) = \int_{w_1}^{w_2} \left|X(r_v(w))\right| dw = \int_{w_1}^{w_2} \sqrt{\left[X(r_v(w))\right]' \cdot \left[X(r_v(w))\right]'}\, dw.$$

For $\rho(w) = R(w, 0)$ to be a geodesic, $v = 0$ must be a local minimum of $L(v)$. This has to happen for every such variation.

Definition 39.2 Suppose that S is a surface parametrized by $X : D \longrightarrow \mathbb{R}^3$ and ρ is a parametrized curve in D. Then $X(\rho)$ is a *geodesic* of S if for every variation r_v of $\rho(w)$, $L(v)$, the length of $X(r_v)$, has a local minimum at $v = 0$.

Finding the local minimum across a perturbation requires locating the critical points, which requires computing the derivative with respect to v, of $L(v)$. We need to be careful with the variables: the integral is with respect to w, but the derivative is with respect to v. There is a trick that simplifies future calculations. The arc length integral involves a square root, which tends to make things difficult. There is, however, a related energy integral

$$\mathcal{E}(v) = \frac{1}{2} \int_{w_1}^{w_2} \left[X(r_v(w))\right]' \cdot \left[X(r_v(w))\right]' dw$$

that has the same critical points as L but not the square root (this is a commonly used trick in optimization problems; the calculations are not fundamentally different, but they are cleaned up). So to find the geodesics, we can solve the simple-looking equation

$$\frac{d}{dv}\mathcal{E}(v)\bigg|_{v=0} = 0.$$

Alas, it is not as simple as it looks.

39.2 Geodesics, the calculations

In the last chapter, we parametrized the pseudosphere as a surface of revolution

$$X(s, t) = \big(z(t)\sin s,\ y(t),\ z(t)\cos s\big), \quad 0 \le s \le 2\pi,$$

and found the equation for y in terms of z that would give the surface constant negative curvature. The result was the tractrix:

$$y = \ln\left(\frac{1+\sqrt{1-z^2}}{z}\right) - \sqrt{1-z^2}, \quad 0 < z \leq 1.$$

The tractrix came to us in a parametrized form where $z = e^t$, for $-\infty < t \leq 0$. We again need a parametrization of the tractrix, but this time a different one. Put $z(t) = 1/t$ with $1 \leq t < \infty$. Then

$$\begin{aligned} y(t) &= \ln\left(\frac{1+\sqrt{1-(1/t)^2}}{1/t}\right) - \sqrt{1-(1/t)^2} \\ &= \ln\left(t + t\sqrt{1-(1/t)^2}\right) - \sqrt{1-(1/t)^2} \\ &= \ln\left(t + \sqrt{t^2-1}\right) - \sqrt{1-(1/t)^2}. \end{aligned}$$

In this parametrization, X is the map from $D = [0, 2\pi) \times [1, \infty)$ to $\mathbb{R}^3$ given by

$$X(s,t) = \left(\frac{1}{t}\sin s,\ \ln\left(t+\sqrt{t^2-1}\right) - \sqrt{1-(1/t)^2},\ \frac{1}{t}\cos s\right).$$

Our goal is to find curves $\rho(w)$ in D so that $X(\rho(w))$ is a geodesic, i.e., curves that minimize energy. That calculation generates a pair of differential equations that $\rho(w)$ must satisfy. Solving the system will give the equations of the geodesics of the pseudosphere.

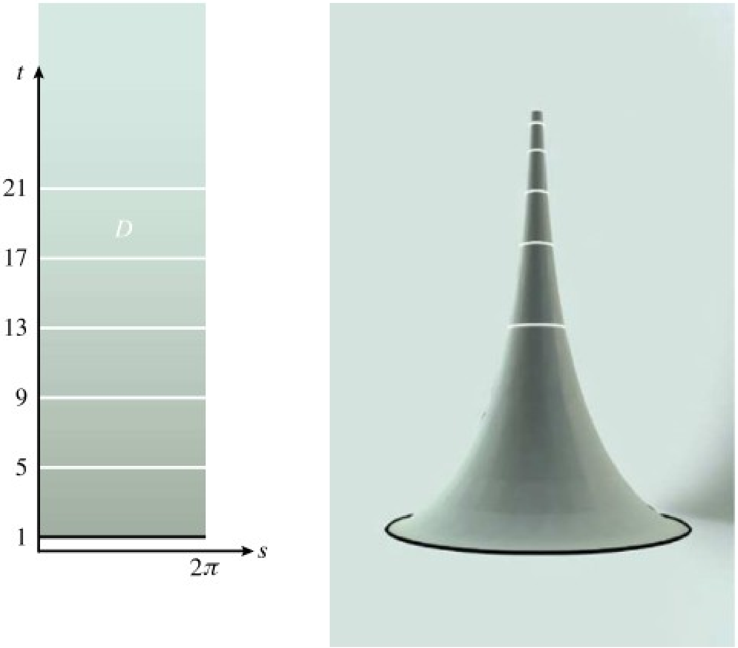

Figure 2. The domain D that is projected onto the pseudosphere by the parametrization X.

Given a variation $r_v(w)$, the energy integral is

$$\mathcal{E}(v) = \frac{1}{2}\int_{w_1}^{w_2} \frac{\partial X}{\partial w}\cdot\frac{\partial X}{\partial w}\,dw.$$

We need to find $\partial X/\partial w$. It may be helpful to review our variables. Because

$$X = \big(z(t)\sin s,\ y(t),\ z(t)\cos s\big),$$

X depends on s and t, which depend upon v and w (for the first calculation, we are setting up an integral to calculate the length of one curve in the variation, so the v variable is held constant). We have found y and z in terms of t, and it might be tempting to substitute now, but it is better to wait. We can use the chain rule to calculate

$$\begin{aligned}\frac{\partial X}{\partial w} &= \Big\langle \frac{\partial}{\partial s}(z\sin s)\cdot\frac{\partial s}{\partial w} + \frac{\partial}{\partial t}(z\sin s)\cdot\frac{\partial t}{\partial w},\ \frac{dy}{dt}\frac{\partial t}{\partial w},\\ &\qquad\qquad \frac{\partial}{\partial s}(z\cos s)\cdot\frac{\partial s}{\partial w} + \frac{\partial}{\partial t}(z\cos s)\cdot\frac{\partial t}{\partial w}\Big\rangle\\ &= \Big\langle z\cos s\frac{\partial s}{\partial w} + \sin s\frac{dz}{dt}\frac{\partial t}{\partial w},\ \frac{dy}{dt}\frac{\partial t}{\partial w},\ -z\sin s\frac{\partial s}{\partial w} + \cos s\frac{dz}{dt}\frac{\partial t}{\partial w}\Big\rangle,\end{aligned}$$

and then

$$\begin{aligned}\frac{\partial X}{\partial w}\cdot\frac{\partial X}{\partial w} &= \left[z\cos s\frac{\partial s}{\partial w} + \sin s\frac{dz}{dt}\frac{\partial t}{\partial w}\right]^2 + \left[\frac{dy}{dt}\frac{\partial t}{\partial w}\right]^2 + \left[-z\sin s\frac{\partial s}{\partial w} + \cos s\frac{dz}{dt}\frac{\partial t}{\partial w}\right]^2.\\ &= \left[z^2\cos^2 s\left(\frac{\partial s}{\partial w}\right)^2 + 2z\sin s\cos s\frac{\partial s}{\partial w}\frac{dz}{dt}\frac{\partial t}{\partial w} + \sin^2 s\left(\frac{dz}{dt}\frac{\partial t}{\partial w}\right)^2\right] + \left[\frac{dy}{dt}\frac{\partial t}{\partial w}\right]^2\\ &\quad + \left[z^2\sin^2 s\left(\frac{\partial s}{\partial w}\right)^2 - 2z\sin s\cos s\frac{\partial s}{\partial w}\frac{dz}{dt}\frac{\partial t}{\partial w} + \cos^2 s\left(\frac{dz}{dt}\frac{\partial t}{\partial w}\right)^2\right].\end{aligned}$$

Thanks to the Pythagorean identity $\sin^2 x + \cos^2 x = 1$, this simplifies to

$$\frac{\partial X}{\partial w}\cdot\frac{\partial X}{\partial w} = z^2\left(\frac{\partial s}{\partial w}\right)^2 + \left(\frac{dz}{dt}\frac{\partial t}{\partial w}\right)^2 + \left(\frac{dy}{dt}\frac{\partial t}{\partial w}\right)^2.$$

Now we can substitute y and z. We know $z = 1/t$, so $dz/dt = -1/t^2$. The dy/dt term is a little more difficult.

Lemma 39.3 *The derivative of*

$$y(t) = \ln(t + \sqrt{t^2 - 1}) - \sqrt{1 - (1/t)^2}, \quad t \geq 1,$$

is $y'(t) = \sqrt{t^2 - 1}/t^2$, *and thus* $[y'(t)]^2 = t^{-2} - t^{-4}$.

Exercise 39.1 Verify lemma 39.3.

Substituting,

$$\begin{aligned}\frac{\partial X}{\partial w}\cdot\frac{\partial X}{\partial w} &= \frac{1}{t^2}\left(\frac{\partial s}{\partial w}\right)^2+\frac{1}{t^4}\left(\frac{\partial t}{\partial w}\right)^2+\left(\frac{1}{t^2}-\frac{1}{t^4}\right)\left(\frac{\partial t}{\partial w}\right)^2\\ &= \frac{1}{t^2}\left[\left(\frac{\partial s}{\partial w}\right)^2+\left(\frac{\partial t}{\partial w}\right)^2\right].\end{aligned}$$

We have our energy function. For a variation $\{r_v(w)\}$ of ρ,

$$\mathcal{E}(v)=\frac{1}{2}\int_{w_1}^{w_2}\frac{1}{t^2}\left[\left(\frac{\partial s}{\partial w}\right)^2+\left(\frac{\partial t}{\partial w}\right)^2\right]dw.$$

The geodesics will be critical points of the function. To find them, let's differentiate with respect to v (use the product rule):

$$\begin{aligned}\frac{d\mathcal{E}}{dv} &= \frac{d}{dv}\left(\frac{1}{2}\int_{w_1}^{w_2}\frac{1}{t^2}\left[\left(\frac{\partial s}{\partial w}\right)^2+\left(\frac{\partial t}{\partial w}\right)^2\right]dw\right)\\ &= \frac{1}{2}\int_{w_1}^{w_2}\frac{\partial}{\partial v}\left(\frac{1}{t^2}\cdot\left[\left(\frac{\partial s}{\partial w}\right)^2+\left(\frac{\partial t}{\partial w}\right)^2\right]\right)dw\\ &= \frac{1}{2}\int_{w_1}^{w_2}\frac{1}{t^2}\left[2\frac{\partial s}{\partial w}\frac{\partial^2 s}{\partial w\partial v}+2\frac{\partial t}{\partial w}\frac{\partial^2 t}{\partial w\partial v}\right]-\frac{2}{t^3}\frac{\partial t}{\partial v}\left[\left(\frac{\partial s}{\partial w}\right)^2+\left(\frac{\partial t}{\partial w}\right)^2\right]dw\\ &= \int_{w_1}^{w_2}\frac{1}{t^2}\frac{\partial s}{\partial w}\frac{\partial^2 s}{\partial w\partial v}+\frac{1}{t^2}\frac{\partial t}{\partial w}\frac{\partial^2 t}{\partial w\partial v}-\frac{1}{t^3}\left[\left(\frac{\partial s}{\partial w}\right)^2+\left(\frac{\partial t}{\partial w}\right)^2\right]\frac{\partial t}{\partial v}\,dw.\end{aligned}$$

The variation variable is v so we have very little control over it. Therefore, it makes sense to try to isolate the v terms as much as possible. The mixed partials are a challenge, but fortunately there is a trick.

Lemma 39.4

$$\int_{w_1}^{w_2}\frac{1}{t^2}\frac{\partial s}{\partial w}\frac{\partial^2 s}{\partial w\partial v}\,dw=\int_{w_1}^{w_2}\frac{1}{t^3}\left[2\frac{\partial s}{\partial w}\frac{\partial t}{\partial w}-t\frac{\partial^2 s}{\partial w^2}\right]\frac{\partial s}{\partial v}\,dw$$

$$\int_{w_1}^{w_2}\frac{1}{t^2}\frac{\partial t}{\partial w}\frac{\partial^2 t}{\partial w\partial v}\,dw=\int_{w_1}^{w_2}\frac{1}{t^3}\left[2\left(\frac{\partial t}{\partial w}\right)^2-t\frac{\partial^2 t}{\partial w^2}\right]\frac{\partial t}{\partial v}\,dw.$$

Proof Both formulas are an application of integration by parts. We will do the first calculation, and leave the similar second calculation as an exercise. Put

$$U=\frac{1}{t^2}\cdot\frac{\partial s}{\partial w},\qquad dU=\left(\frac{1}{t^2}\frac{\partial^2 s}{\partial w^2}-\frac{2}{t^3}\frac{\partial t}{\partial w}\frac{\partial s}{\partial w}\right)dw,$$

$$dV=\frac{\partial^2 s}{\partial w\partial v}\,dw,\qquad V=\frac{\partial s}{\partial v},$$

so that

$$\int_{w_1}^{w_2} \frac{1}{t^2}\frac{\partial s}{\partial w}\frac{\partial^2 s}{\partial w \partial v}\,dw = \left(\frac{1}{t^2}\frac{\partial s}{\partial w}\frac{\partial s}{\partial v}\right)\Bigg|_{w_1}^{w_2} - \int_{w_1}^{w_2} \frac{\partial s}{\partial v}\left[\frac{1}{t^2}\frac{\partial^2 s}{\partial w^2} - \frac{2}{t^3}\frac{\partial t}{\partial w}\frac{\partial s}{\partial w}\right] dw.$$

The variation is fixed at both endpoints, so $\partial s/\partial v(w_1) = \partial s/\partial v(w_2) = 0$. The first term contributes nothing, so

$$\begin{aligned}\int_{w_1}^{w_2} \frac{1}{t^2}\frac{\partial s}{\partial w}\frac{\partial^2 s}{\partial w \partial v}\,dw &= 0 - \int_{w_1}^{w_2} \frac{\partial s}{\partial v}\left[\frac{1}{t^2}\frac{\partial^2 s}{\partial w^2} - \frac{2}{t^3}\frac{\partial t}{\partial w}\frac{\partial s}{\partial w}\right] dw \\ &= \int_{w_1}^{w_2} \frac{1}{t^3}\left[2\frac{\partial t}{\partial w}\frac{\partial s}{\partial w} - t\frac{\partial^2 s}{\partial w^2}\right]\frac{\partial s}{\partial v}\,dw.\end{aligned}$$

□

Exercise 39.2 Verify the second equation in the last lemma.

Substitute into $d\mathcal{E}/dv$ to get

$$\begin{aligned}\frac{d\mathcal{E}}{dv} &= \int_{w_1}^{w_2} \frac{1}{t^3}\left[2\frac{\partial s}{\partial w}\frac{\partial t}{\partial w} - t\frac{\partial^2 s}{\partial w^2}\right]\frac{\partial s}{\partial v} + \frac{1}{t^3}\left[2\left(\frac{\partial t}{\partial w}\right)^2 - t\frac{\partial^2 t}{\partial w^2}\right]\frac{\partial t}{\partial v} \\ &\qquad - \frac{1}{t^3}\left[\left(\frac{\partial s}{\partial w}\right)^2 + \left(\frac{\partial t}{\partial w}\right)^2\right]\frac{\partial t}{\partial v}\,dw \\ &= \int_{w_1}^{w_2} \frac{1}{t^3}\left[2\frac{\partial s}{\partial w}\frac{\partial t}{\partial w} - t\frac{\partial^2 s}{\partial w^2}\right]\frac{\partial s}{\partial v} + \frac{1}{t^3}\left[\left(\frac{\partial t}{\partial w}\right)^2 - t\frac{\partial^2 t}{\partial w^2} - \left(\frac{\partial s}{\partial w}\right)^2\right]\frac{\partial t}{\partial v}\,dw.\end{aligned}$$

The critical points of $\mathcal{E}(v)$ occur when the integral is zero. Still, we know little about $\partial s/\partial v$ and $\partial t/\partial v$. There may be some variations where the $\partial s/\partial v$ and $\partial t/\partial v$ components cancel each other out. But geodesics occur only when energy is minimized for all variations. The only way to guarantee this regardless of variation is if when $v = 0$, at $\rho(v)$,

- $2\dfrac{ds}{dw}\dfrac{dt}{dw} - t\dfrac{d^2 s}{dw^2} = 0$, and
- $\left(\dfrac{dt}{dw}\right)^2 - t\dfrac{d^2 t}{dw^2} - \left(\dfrac{ds}{dw}\right)^2 = 0$,

the differential equations that describe a geodesic ρ. The next step is to put them into a more useful form (one that does not involve second derivatives) with the following lemmas.

Lemma 39.5 *If*

$$2\frac{ds}{dw}\frac{dt}{dw} - t\frac{d^2 s}{dw^2} = 0,$$

then s'/t^2 is a constant.

Proof Assuming as given that

$$-ts'' + 2t's' = 0,$$

use the quotient rule to compute the derivative of s'/t^2:

$$\left(\frac{s'}{t^2}\right)' = \frac{t^2 s'' - s' \cdot 2tt'}{(t^2)^2} = \frac{ts'' - 2s't'}{t^3} = 0.$$

Since the derivative of s'/t^2 is zero, s'/t^2 is a constant. ☐

Lemma 39.6 *If*

$$2\frac{ds}{dw}\frac{dt}{dw} - t\frac{d^2 s}{dw^2} = 0 \quad \textit{and} \quad \left(\frac{dt}{dw}\right)^2 - t\frac{d^2 t}{dw^2} - \left(\frac{ds}{dw}\right)^2 = 0,$$

then $(s')^2 + (t')^2 = Ct^2$ *for some constant* C.

Proof Multiply both sides of the second equation by -1 to get $(s')^2 - (t')^2 + tt'' = 0$. Then add $(s')^2 + (t')^2$ to both sides and multiply by t':

$$\begin{aligned} (s')^2 - (t')^2 + tt'' &= 0 \\ 2(s')^2 + tt'' &= (s')^2 + (t')^2 \\ 2(s')^2 t' + tt't'' &= [(s')^2 + (t')^2]t'. \end{aligned}$$

According to the first equation, $2s't' = ts''$ so

$$\begin{aligned} s'(ts'') + tt't'' &= [(s')^2 + (t')^2]t' \\ t[s's'' + t't''] &= [(s')^2 + (t')^2]t' \\ [s's'' + t't'']/[(s')^2 + (t')^2] &= t'/t. \end{aligned}$$

Integrate with respect to w. On the right

$$\int \frac{t'}{t}\, dw = \ln(t) + C_1.$$

The left hand side is similar: put $u = (s')^2 + (t')^2$ so that $du = (2s's'' + 2t't'')\, dw$. Then

$$\int \frac{s's'' + t't''}{(s')^2 + (t')^2}\, dw = \int \frac{1/2}{u}\, du = \frac{1}{2}\ln|u| + C_2 = \frac{1}{2}\ln\left((s')^2 + (t')^2\right) + C_2.$$

Combining, we get

$$\begin{aligned} \tfrac{1}{2}\ln\left((s')^2 + (t')^2\right) + C_2 &= \ln(t) + C_1 \\ \ln\left((s')^2 + (t')^2\right) &= 2\ln(t) + 2C_1 - 2C_2 \\ (s')^2 + (t')^2 &= e^{\ln(t^2) + 2C_1 - 2C_2} \\ (s')^2 + (t')^2 &= Ct^2 \end{aligned}$$

where $C = e^{2C_1 - 2C_2}$. ☐

The constant C eventually gets wrapped into other constants, so there is no harm in assuming that $C = 1$ now. Doing that means

$$X(\rho)' \cdot X(\rho)' = \frac{1}{t^2}\left[(s')^2 + (t')^2\right] = \frac{1}{t^2} \cdot t^2 = 1.$$

The parametrized curve $X(\rho)$ on the pseudosphere moves at a constant speed of 1, so the curve is parametrized by arc length.

Now let's assume that a curve ρ satisfies the differential equations, and try to see what ρ will look like. Assuming $s'/t^2 = k$ for some constant k, write $s' = kt^2$. There is one case where this is easy: if $k = 0$, then $s' = 0$, so s is a constant, so ρ is a vertical line in D.

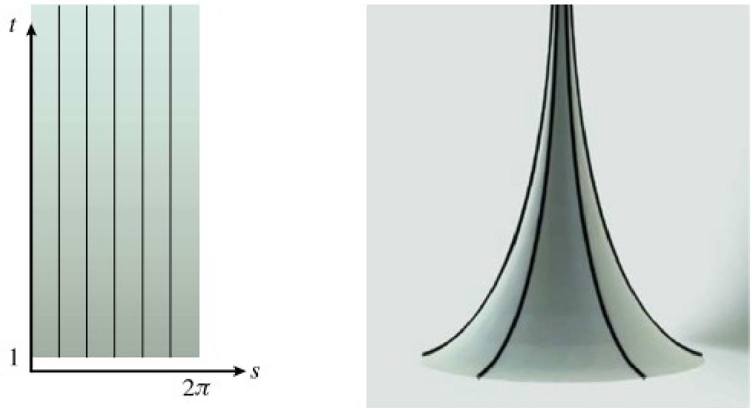

Figure 3. Geodesics on the pseudosphere that are projections of vertical lines.

If $k \neq 0$, the calculation is more difficult. Start by substituting $s' = kt^2$ into the second equation:

$$\begin{aligned}(s')^2 + (t')^2 = t^2 &\implies (kt^2)^2 + (t')^2 = t^2\\ &\implies (t')^2 = t^2 - k^2t^4\\ &\implies t' = \sqrt{t^2 - k^2t^4} = t\sqrt{1 - k^2t^2}.\end{aligned}$$

Since s and t are functions of w, we can eliminate w to get a single differential equation relating s and t:

$$\frac{ds}{dt} = \frac{ds/dw}{dt/dw} = \frac{kt^2}{t\sqrt{1 - k^2t^2}} = \frac{kt}{\sqrt{1 - k^2t^2}}.$$

Integrating gives s in terms of t, the equations of the geodesics.

Lemma 39.7 *If $k \neq 0$, then*

$$\int \frac{kt}{\sqrt{1 - k^2t^2}}\,dt = -\frac{1}{k}\sqrt{1 - k^2t^2} + C.$$

Proof Let $u = 1 - k^2t^2$ so $du = -2k^2t\,dt$ and

$$\begin{aligned}\int \frac{kt}{\sqrt{1-k^2t^2}}\,dt &= \int \frac{-1/(2k)}{\sqrt{u}}\,du \\ &= -\frac{1}{2k}\int u^{-1/2}\,du \\ &= -\frac{1}{2k}2u^{1/2} + C \\ &= -\frac{1}{k}\sqrt{1-k^2t^2} + C.\end{aligned}$$ ☐

Therefore, if ρ is a geodesic, it satisfies

$$s = -\frac{1}{k}\sqrt{1-k^2t^2} + C \implies k(s-C) = -\sqrt{1-k^2t^2}.$$

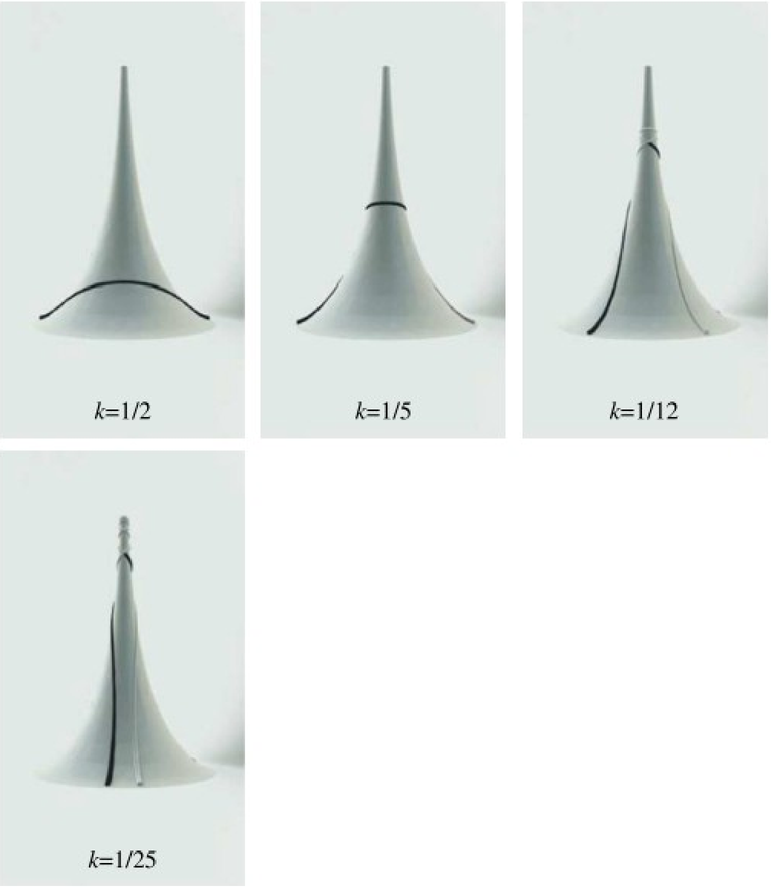

k=1/2

k=1/5

k=1/12

k=1/25

Figure 4. Geodesics on the pseudosphere that are the projections of half-circles. The projection of the entire half-circle is shown. The portion that lies in the domain D is shaded black. Extending the curves beyond D creates self-intersections that would be undesirable in a model for a non-Euclidean neutral geometry.

Square both sides:

$$\begin{aligned} k^2(s-C)^2 &= 1-k^2t^2 \\ k^2(s-C)^2+k^2t^2 &= 1 \\ (s-C)^2+t^2 &= (1/k)^2. \end{aligned}$$

The curves in D that project to geodesics on the pseudosphere are circles that are centered on the s-axis (or at least the portions of them that lie in D).

The geodesics reveal another problem with the pseudosphere model. When the circles ρ reach the edge of D, the corresponding geodesics on the pseudosphere just stop. It is easy to get around that problem by extending D horizontally so that X wraps around the pseudosphere more than once. But then the geodesics cross over themselves, creating undesirable self-intersections.

39.3 References

Spivak discusses geodesics in the third volume of his *Comprehensive Introduction to Differential Geometry* [Spi99b], but my approach here more closely follows Thorpe's *Elementary Topics in Differential Geometry* [Tho79].

39.4 Exercises

39.3. The standard parameterization (in terms of sine and cosine) of the circle $(s-C)^2+t^2=(1/k)^2$ does not satisfy the differential equation for the arc length criterion or either of the differential equations for the geodesic criterion. The parametrization that works is

$$s(w)=\frac{1}{k}\tanh(w), \quad t(w)=\frac{1}{k}\operatorname{sech}(w).$$

Verify that the parametrization satisfies the differential equations (the $\operatorname{sech}(w)$ and $\tanh(w)$ functions are covered in chapter 44 if you are not familiar with them).

39.4. Show that lines are geodesics in the plane $z=0$.

39.5. Show that the circle $r(t)=\langle\cos t, \sin t, k\rangle$ (k constant) is a geodesic on the cylinder $x^2+y^2=1$. Show that the vertical line $r(t)=\langle 1, 0, t\rangle$ is also a geodesic on the cylinder.

39.6. Show that the equatorial circle $r(t)=\langle\cos t, \sin t, 0\rangle$ is a geodesic on the sphere $x^2+y^2+z^2=1$.

46

Hyperbolic Area

When we developed Euclidean area in chapter 30, we described it as a positive real number associated with each polygon so that

- congruent polygons would have the same area,
- if a polygon could be broken into smaller pieces, the area of that polygon would be the sum of the areas of the pieces, and
- the area of a rectangle would be given by the formula $A = bh$.

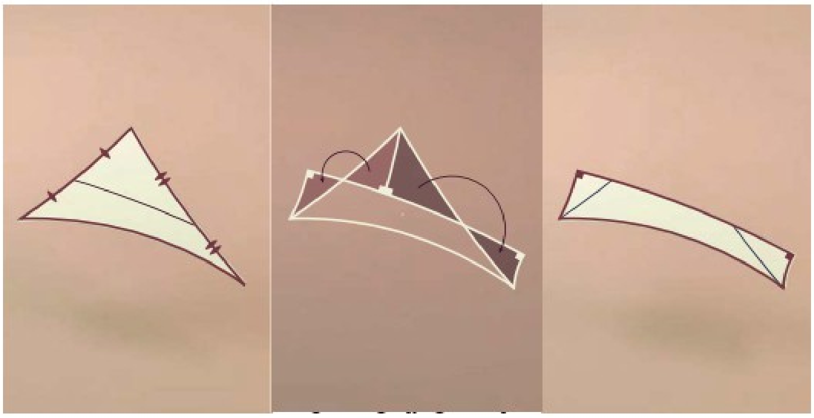

Figure 1. A triangle can be cut and the pieces rearranged to form a Saccheri quadrilateral. The triangle and the Saccheri quadrilateral should have the same area.

Hyperbolic area should retain the first two properties, but the last one isn't going to work because there are no rectangles. Hyperbolic area will have to be built from some other shape. The obvious choice is a triangle (and you could argue that it would have been better to build up Euclidean areas from triangles as well). The area of a Euclidean triangle is $A = bh/2$ but the formula doesn't work in the hyperbolic realm (a specific example is investigated in the exercises). So where can we look to find the formula for the area of a hyperbolic triangle? We will go back to the pseudosphere.

46.1 Area on the pseudosphere

In multivariable calculus the area $\mathscr{A}$ of a parametrized surface $\Omega = X(D)$ can be calculated as the double integral

$$\mathscr{A} = \iint_{\Omega} 1\, d\sigma = \iint_{D} |N| dt ds$$

where $|N|$ is the length of the normal vector. We have parametrized the pseudosphere by

$$X(s,t) = \left(\frac{\sin s}{t},\ \ln(t + \sqrt{t^2 - 1}) - \sqrt{1 - \frac{1}{t^2}},\ \frac{\cos s}{t} \right),$$

where $0 \le s \le 2\pi$ and $1 \le t < \infty$. The restriction on t is essential, but the $[0, 2\pi]$ restriction on s is there just to ensure that the domain wraps around the pseudosphere only once. That horizontal constraint is fairly restricting but we can ignore it, so long as we understand that this will cause X to wrap around the pseudosphere repeatedly.

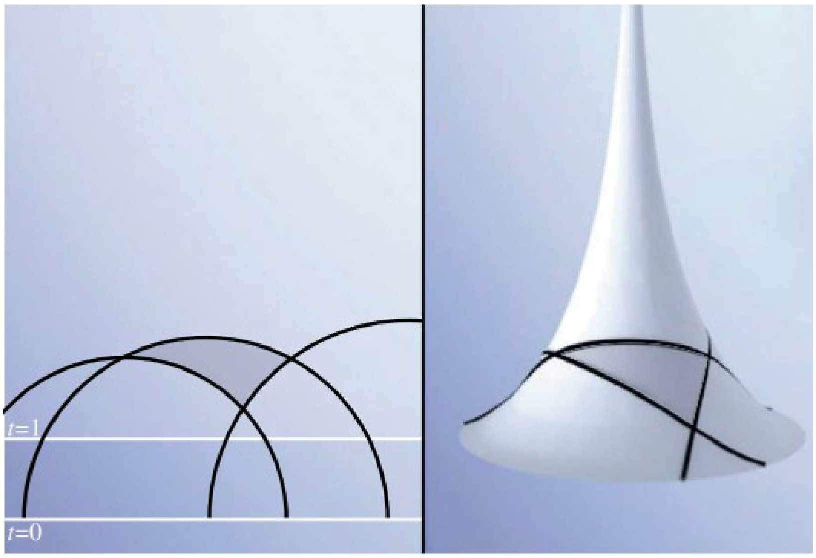

Figure 2. The projection of a triangle from the UHP to the pseudosphere.

The N vector is the cross product of the partial derivatives,

$$X_s = \left\langle \frac{\cos s}{t},\ 0,\ -\frac{\sin s}{t} \right\rangle$$
$$X_t = \left\langle -\frac{\sin s}{t^2},\ \frac{\sqrt{t^2 - 1}}{t^2},\ -\frac{\cos s}{t^2} \right\rangle$$

(the calculation of $\partial y/\partial t$ in X_t was an exercise in chapter 39). Then

$$\begin{aligned}
N &= \begin{vmatrix} i & j & k \\ \dfrac{\cos s}{t} & 0 & -\dfrac{\sin s}{t} \\ -\dfrac{\sin s}{t^2} & \dfrac{\sqrt{t^2-1}}{t^2} & -\dfrac{\cos t}{t^2} \end{vmatrix} \\
&= \left\langle \frac{\sin s}{t} \cdot \frac{\sqrt{t^2-1}}{t^2},\ \frac{\sin s}{t} \cdot \frac{\sin s}{t^2} + \frac{\cos s}{t} \frac{\cos s}{t^2},\ \frac{\cos s}{t} \cdot \frac{\sqrt{t^2-1}}{t^2} \right\rangle \\
&= \left\langle \frac{\sin s\sqrt{t^2-1}}{t^3},\ \frac{1}{t^3},\ \frac{\cos s\sqrt{t^2-1}}{t^3} \right\rangle,
\end{aligned}$$

and so

$$\begin{aligned}
|N| &= (N \cdot N)^{1/2} \\
&= \left[\frac{\sin^2 s(t^2-1) + 1 + \cos^2 s(t^2-1)}{t^6} \right]^{1/2} \\
&= \frac{[(t^2-1)+1]^{1/2}}{t^3} \\
&= 1/t^2.
\end{aligned}$$

Substitution yields a pleasant integral:

$$\mathscr{A} = \iint_D \frac{1}{t^2}\, ds dt.$$

We are not interested in the surface area of the pseudosphere as a whole. We want to know the areas of hyperbolic triangles on it. For that, the tricky part is to set up the proper domain D to cover them. We proceed in several steps.

Lemma 46.1 *Let*

$$D = \{(s,t) \big| (s^2+t^2) < r^2,\ 0 \le a \le s \le b,\ t \ge 1\},$$

an approximately quadrilateral region in the UHP. Then the area of $X(D)$ *is*

$$\mathscr{A} = (b-a) + (\sin^{-1}(a/r) - \sin^{-1}(b/r)).$$

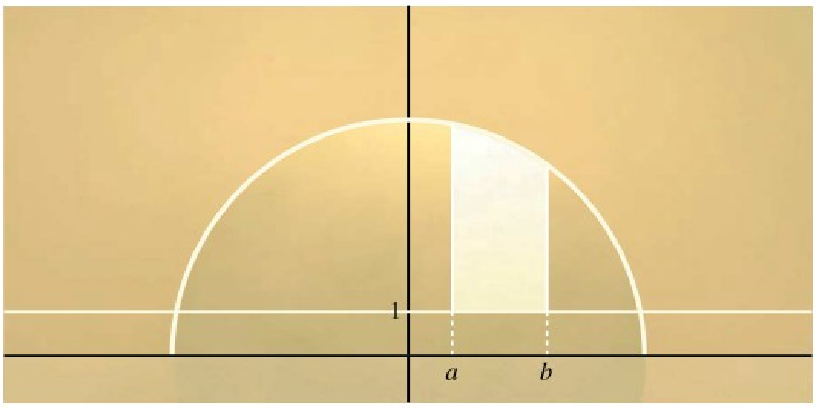

Figure 3. The first domain, with $0 \leq a \leq b$.

Proof This is a straightforward calculation.

$$
\begin{aligned}
\mathscr{A} &= \iint_D \frac{1}{t^2}\, ds dt \\
&= \int_a^b \int_1^{\sqrt{r^2-s^2}} \frac{1}{t^2}\, dt ds \\
&= \int_a^b -\frac{1}{t}\Big|_1^{\sqrt{r^2-s^2}} ds \\
&= \int_a^b 1 - \frac{1}{\sqrt{r^2-s^2}}\, ds \\
&= \int_a^b 1\, ds - \int_a^b \frac{1}{\sqrt{r^2-s^2}}\, ds \\
&= (b-a) - \int_a^b \frac{1}{\sqrt{r^2-s^2}}\, ds.
\end{aligned}
$$

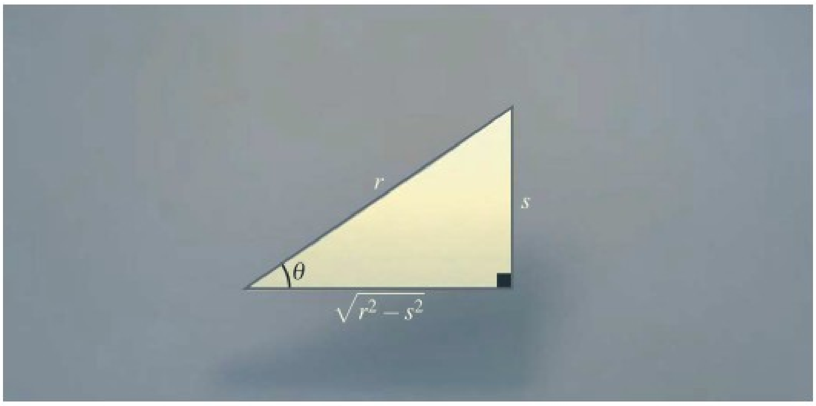

Figure 4. A trigonometric substitution.

Use trigonometric substitution: put $s = r\sin\theta$, so that $\sqrt{r^2 - s^2} = r\cos\theta$ and $ds = r\cos\theta\, d\theta$. Then

$$\begin{aligned}
\mathscr{A} &= (b-a) - \int_{s=a}^{s=b} \frac{r\cos\theta}{r\cos\theta}\, d\theta \\
&= (b-a) - \left(\theta\Big|_{s=a}^{s=b}\right) \\
&= (b-a) - \left(\sin^{-1}(s/r)\Big|_a^b\right) \\
&= (b-a) - \left[\sin^{-1}(b/r) - \sin^{-1}(a/r)\right] \\
&= (b-a) + \sin^{-1}(a/r) - \sin^{-1}(b/r).
\end{aligned}$$

□

The previous lemma ended with two inverse sine functions. They can be rewritten in a more useful form.

Lemma 46.2 *Let D be the same region as in the previous lemma, but now label its top-most vertices*

$$A = (a, \sqrt{r^2 - a^2}) \text{ and } B = (b, \sqrt{r^2 - b^2}).$$

Then

$$\mathscr{A} = (b-a) + (\pi - (\angle A) - (\angle B)),$$

where $(\angle A)$ and $(\angle B)$ are computed by measuring the angles between the vertical lines and the tangent lines to D at A and B.

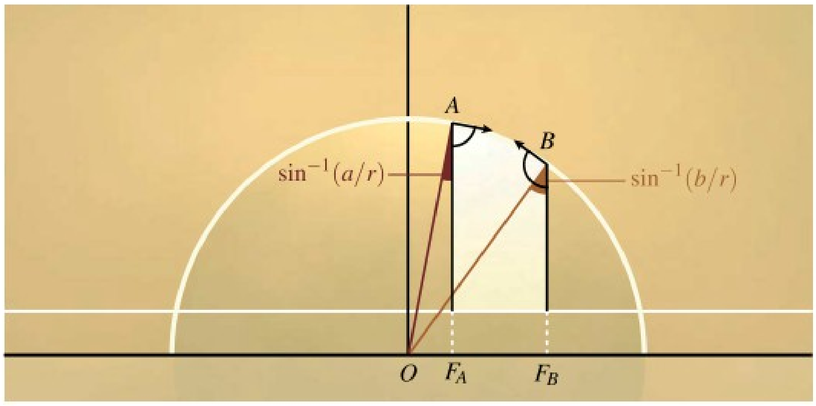

Figure 5. Relating the area to the measures of $\angle A$ and $\angle B$.

Proof We start with labels: let O be the origin, $F_A = (a, 0)$, and $F_B = (b, 0)$. The right triangle $\triangle OF_AA$ has hypotenuse r and a side of length a, so

$$\sin(\angle OAF_A) = \frac{a}{r} \implies (\angle OAF_A) = \sin^{-1}(a/r).$$

This angle is complementary to $\angle A$, so $\sin^{-1}(a/r) = \pi/2 - (\angle A)$. The behavior at B is different. Again

$$\sin(\angle OBF_B) = \frac{b}{r} \implies (\angle OBF_B) = \sin^{-1}(b/r),$$

but this time $\sin^{-1}(b/r) = (\angle B) - \pi/2$. Substituting into the formula for $\mathscr{A}$,

$$\begin{aligned}\mathscr{A} &= (b-a) + (\pi/2 - (\angle A)) - ((\angle B) - \pi/2)\\ &= (b-a) + (\pi - (\angle A) - (\angle B)).\end{aligned}$$

□

In the first two lemmas, we put restrictions on D that are not really necessary: that a and b have to be positive, and that the circle must be centered at the origin. Now let's remove them in the next two lemmas.

Lemma 46.3 *If*

$$D = \left\{(s,t)\middle|(s^2+t^2) < r^2,\ a \le s \le b,\ t \ge 1\right\},$$

then $\mathscr{A} = (b-a) + (\pi - (\angle A) - (\angle B))$.

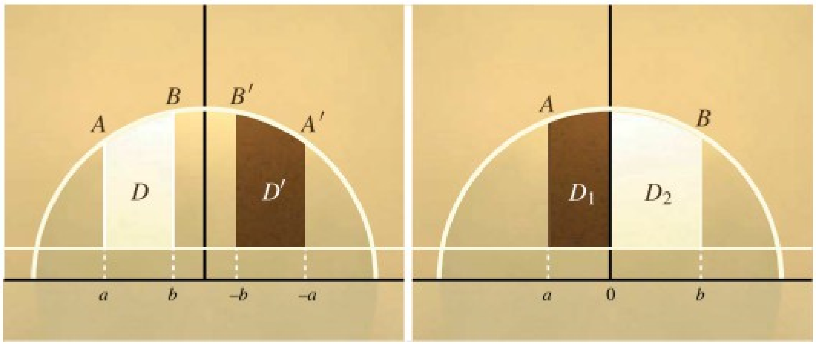

Figure 6. Extending to other types of domains: (l) $a \le b \le 0$ and (r) $a \le 0 \le b$.

Proof Step 1. Suppose that $a < b \le 0$. The reflection across the vertical axis will transform D into a congruent shape D', mapping A to A' and B to B'. The area of $X(D')$ is the same as the area of $X(D)$, though, and $\angle A \simeq \angle A'$ and $\angle B \simeq \angle B'$, so

$$\begin{aligned}\mathscr{A}(D) &= \mathscr{A}(D')\\ &= ((-a) - (-b)) + (\pi - (\angle B') - (\angle A'))\\ &= (b-a) + (\pi - (\angle A) - (\angle B)).\end{aligned}$$

Step 2. Suppose that $a < 0 < b$. Then D can be broken up into two smaller regions, D_1 from $x = a$ to $x = 0$, and D_2 from $x = 0$ to $x = b$. We have

$$\begin{aligned}\mathscr{A}(D_1) &= (0-a) + (\pi - (\angle A) - \pi/2) = -a + \pi/2 - (\angle A)\\ \mathscr{A}(D_2) &= (b-0) + (\pi - \pi/2 - (\angle B)) = b + \pi/2 - (\angle B),\end{aligned}$$

and so

$$\begin{aligned}\mathscr{A}(D) &= \mathscr{A}(D_1) + \mathscr{A}(D_2)\\ &= (-a + \pi/2 - (\angle A)) + (b + \pi/2 - (\angle B))\\ &= (b - a) + (\pi - (\angle A) - (\angle B)).\end{aligned}$$ □

Lemma 46.4 *If*

$$D = \left\{(s,t)\big|(s-h)^2 + t^2 < r^2,\ a \le s \le b,\ t \ge 1\right\},$$

then $\mathscr{A} = (b - a) + (\pi - (\angle A) - (\angle B))$.

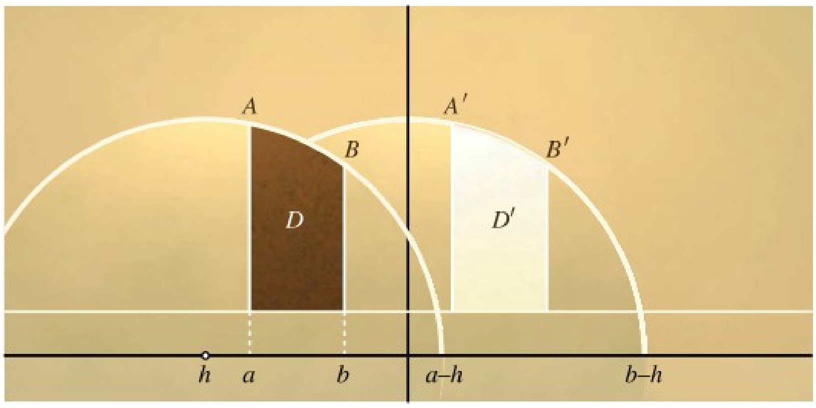

Figure 7. Extending to other types of domains: circles that are not centered at the origin.

Proof The horizontal translation by $-h$ shifts D to a congruent domain D' whose bounding circle is centered at the origin, mapping A to A' and B to B'. It does not alter area: the area of $X(D)$ is the same as the area of $X(D')$. Nor does it alter angle measures, so $(\angle A) = (\angle A')$ and $(\angle B) = (\angle B')$. Therefore

$$\begin{aligned}\mathscr{A}(D) &= \mathscr{A}(D')\\ &= ((b-h) - (a-h)) + (\pi - (\angle A') - (\angle B'))\\ &= (b - a) + (\pi - (\angle A) - (\angle B)).\end{aligned}$$ □

Now we can cut apart and piece together domains to find the area of a triangle.

Theorem 46.5 *For* $\triangle ABC$ *in the UHP, the area of* $X(\triangle ABC)$ *is*

$$\mathscr{A} = \pi - (\angle A) - (\angle B) - (\angle C).$$

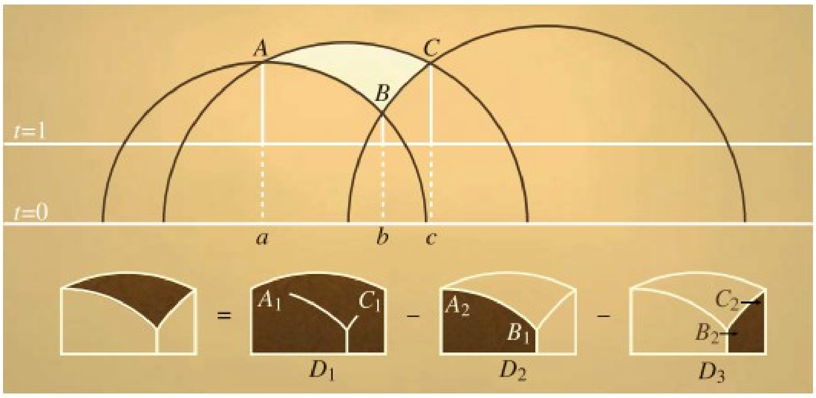

Figure 8. Calculating the area of a hyperbolic triangle.

Proof There are a few configurations. We will look at just one, as illustrated and labeled in figure 8. In this case, the area of $X(\triangle ABC)$ is

$$\mathscr{A}(\triangle ABC) = \mathscr{A}(D_1) - \mathscr{A}(D_2) - \mathscr{A}(D_3)$$

where

$$\begin{aligned}
\mathscr{A}(D_1) &= (c-a) + [\pi - (\angle A_1) - (\angle C_1)] \\
\mathscr{A}(D_2) &= (b-a) + [\pi - (\angle A_2) - (\angle B_1)] \\
\mathscr{A}(D_3) &= (c-b) + [\pi - (\angle B_2) - (\angle C_2)],
\end{aligned}$$

and so

$$\begin{aligned}
\mathscr{A}(\triangle ABC) &= (c-a) + [\pi - (\angle A_1) - (\angle C_1)] \\
&\quad - (b-a) - [\pi - (\angle A_2) - (\angle B_1)] \\
&\quad - (c-b) - [\pi - (\angle B_2) - (\angle C_2)] \\
&= [(\angle A_2) - (\angle A_1)] + [(\angle B_2) + (\angle B_1)] + [(\angle C_2) - (\angle C_1)] - \pi \\
&= -(\angle A) + [2\pi - (\angle B)] - (\angle C) - \pi \\
&= \pi - (\angle A) - (\angle B) - (\angle C).
\end{aligned}$$

□

46.2 Areas of polygons in the Poincaré disk

The UHP model inherits its metrical properties from the pseudosphere, so $\triangle ABC$ in the UHP should also have area $\pi - (\angle A) - (\angle B) - (\angle C)$. The map $\sigma \circ \rho$ from the UHP to D is conformal, so the image of $\triangle ABC$ in D will have the same angles. Therefore, the pseudosphere area calculation carries over directly to suggest an area formula in the Poincaré disk.

Definition 46.6 The area of $\triangle ABC$ is $\pi - (\angle A) - (\angle B) - (\angle C)$.

In spite of all the calculations that have led to this point, this is still (I think) a little surprising. We are saying that triangle areas are calculated entirely in terms of angle sums! The

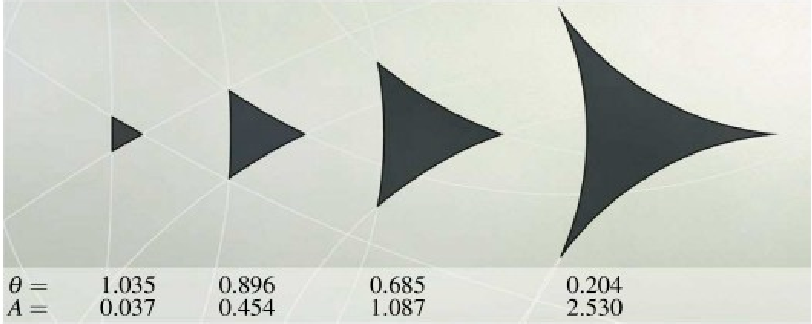

Figure 9. Hyperbolic areas of equilateral triangles, given by the formula $\mathscr{A} = \pi - 3\theta$.

term $\pi - (\angle A) - (\angle B) - (\angle C)$ is called the *defect* of the triangle. In Euclidean geometry, it would always be zero, so the defect measures how far the triangle deviates from a Euclidean triangle. This formula is saying that small triangles (in the sense of area) are the ones with small defects. That jibes with the general principal of the last lesson, that Euclidean and hyperbolic geometry look approximately the same for small objects, but diverge from one another for bigger objects. It also means that no hyperbolic triangle can have an area greater than π! Surprising or not, it turns out that the formula meets the requirements of an area function.

In chapter 31, we found a formula for the area of a general simple polygon, using an inductive argument. It was based on the idea that a simple polygon can be split into smaller pieces, and then the area of the whole is the sum of the areas of the pieces (every simple polygon can be triangulated and its area is the sum of the areas of the triangles). The key to the argument was that every simple polygon with more than three sides has a diagonal (theorem 31.1) that breaks the polygon into two smaller pieces. The argument for the existence of a diagonal works just fine in hyperbolic geometry.

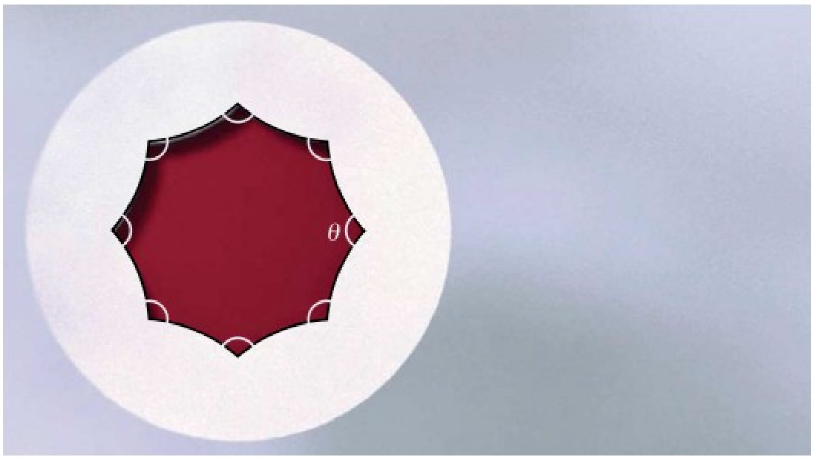

Figure 10. The area of a regular octagon with interior angle θ is $\mathscr{A} = 6\pi - 8\theta$.

Exercise 46.1 Check the proof of theorem 31.1 to make sure that it is still valid in hyperbolic geometry.

Theorem 46.7 *The area of a simple polygon* $\mathcal{P} = P_1P_2\cdots P_n$ *is*

$$(n-2)\pi - \sum_{i=1}^{n}(\angle P_i).$$

Proof As in the Euclidean case, we proceed by induction on the number of sides of $\mathcal{P}$. The base case consists of polygons of three sides: triangles. For them, the formula matches the defect formula. Now assume that polygons with n or fewer sides have areas that are given by the formula, and consider an $(n+1)$-sided polygon $\mathcal{P} = P_1P_2\cdots P_{n+1}$. It has a diagonal cutting $\mathcal{P}$ into two smaller pieces $\mathcal{P}_A$ and $\mathcal{P}_B$. To simplify the notation, let's relabel the vertices of $\mathcal{P}$ so that

$$\begin{aligned}\mathcal{P}_A &= P_1P_2\cdots P_j\\ \mathcal{P}_B &= P_1P_jP_{j+1}\cdots P_{n+1}.\end{aligned}$$

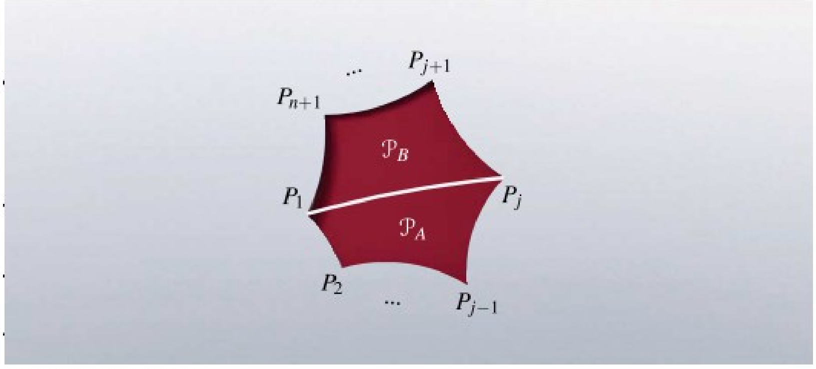

Figure 11. To prove the area formula, we use an inductive proof, cutting the polygon into two smaller pieces with a diagonal.

Between $\mathcal{P}_A$ and $\mathcal{P}_B$, there are $n+3$ sides (the $n+1$ original sides and the diagonal that is counted twice because it occurs in both pieces). The first piece, $\mathcal{P}_A$, has j of the sides. The other piece, $\mathcal{P}_B$, must have the remaining $n+3-j$ sides. According to the inductive hypothesis,

$$\begin{aligned}\mathscr{A}(\mathcal{P}_A) &= (j-2)\pi - \left[(\angle P_jP_1P_2) + \sum_{i=2}^{j-1}(\angle P_i) + (\angle P_{j-1}P_jP_1)\right]\\ \mathscr{A}(\mathcal{P}_B) &= (n+1-j)\pi - \left[(\angle P_1P_jP_{j+1}) + \sum_{i=j+1}^{n+1}(\angle P_i) + (\angle P_{n+1}P_1P_j)\right].\end{aligned}$$

Then

$$\begin{aligned}\mathscr{A}(\mathcal{P}) &= \mathscr{A}(\mathcal{P}_A) + \mathscr{A}(\mathcal{P}_B)\\ &= (n-1)\pi - \left[(\angle P_1) + \sum_{i=2}^{j-1}(\angle P_i) + (\angle P_j) + \sum_{i=j+1}^{n+1}(\angle P_i)\right]\\ &= ((n+1)-2)\pi - \sum_{i=1}^{n+1}(\angle P_i),\end{aligned}$$

as needed. By induction, this is a valid formula for any simple polygon. □

Now that we have this formula for the area of a general polygon, we can check that it satisfies the requirements of an area function.

Exercise 46.2 Verify that the formula assigns to every simple polygon a positive real number. Verify that the formula assigns the same number to two congruent polygons.

The last requirement, that the area of the whole is the sum of the areas of the pieces, is more complicated to verify. The next exercise confirms it in a few specific cases.

Exercise 46.3 Let D be a point in the interior of $\triangle ABC$. Prove that

$$\mathscr{A}(\triangle ABC) = \mathscr{A}(\triangle ABD) + \mathscr{A}(\triangle BCD) + \mathscr{A}(\triangle ACD).$$

Given a convex quadrilateral $\square ABCD$, let $A * E * B$ and $C * F * D$. Prove that

$$\mathscr{A}(\square ABCD) = \mathscr{A}(\square AEFD) + \mathscr{A}(\square BEFC).$$

46.3 Area of a circle

To end the chapter, let's use the formula for the area of a polygon to find the area of a circle. The formula depends upon two things. One is lemma 45.6, that as n approaches infinity, $n\sin(x/n)$ approaches x. The other is a kind of hyperbolic cofactor identity:

Lemma 46.8 *Let α and β be the two acute angles in a right triangle and let a be the length of the side opposite α. Then*

$$\cos\alpha = \sin\beta \cdot \cosh a.$$

Exercise 46.4 Prove the lemma. Hint: to get started, rewrite the cosine term in terms of hyperbolic tangents.

The derivation of the area formula for circles is like the derivation of the circumference formula: we approximate the circle with regular inscribed polygons and then take a limit as the number of sides approaches infinity. The area formula would seem to make this easy, since it only depends on the interior angles and all the interior angles of a regular polygon are congruent. It turns out that there are some computational complications with this approach, and so we will instead do something a little less direct.

Theorem 46.9 *The area of a circle with radius r is*

$$\mathscr{A} = 2\pi(\cosh r - 1).$$

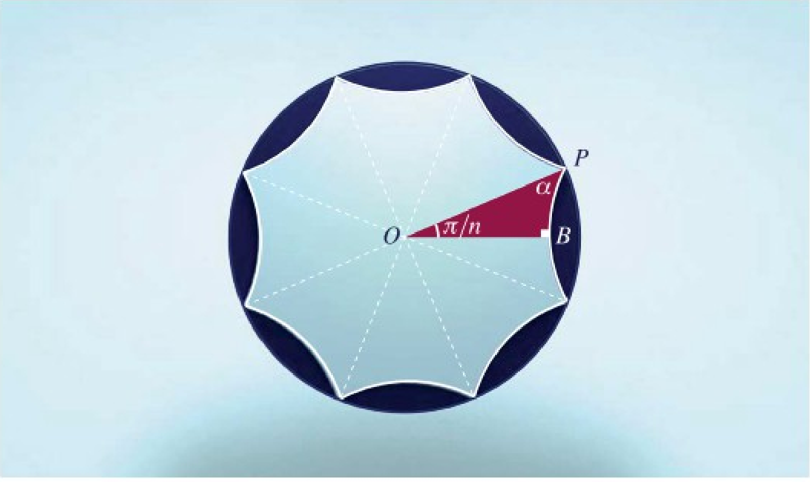

Figure 12. Use inscribed regular polygons to approximate the area inside a circle.

Proof Write $\mathscr{A}$ for the area of a circle with radius r. Let $\mathscr{A}_n$ denote the area of a regular n-gon $\mathcal{P}_n$ inscribed in the circle. It can be broken up into $2n$ right triangles each of which is congruent to $\triangle OPB$, where O is the center of the circle, P is one of the vertices of the n-gon, and B is the right angle. The angle at P is the key. Let's say it has a measure of α. Then each interior angle of $\mathcal{P}_n$ has a measure of 2α and therefore

$$\mathscr{A}_n = (n-2)\pi - \sum_{i=1}^{n} 2\alpha = (n-2)\pi - 2n\alpha.$$

Rather than trying to find the value of α, we solve the equation for α:

$$2n\alpha = (n-2)\pi - \mathscr{A}_n \implies \alpha = \pi/2 - \pi/n - \mathscr{A}_n/2n.$$

The trick is to look not at α itself, but instead at $\cos\alpha$. There are two ways we can work with that expression. One way is to use the cofactor identity in lemma 46.8. In $\triangle OPB$, the angle at O measures π/n. Therefore

$$\cos\alpha = \sin(\pi/n)\cosh(|OB|).$$

The other way is via the traditional cofactor identity:

$$\begin{aligned} \cos\alpha &= \cos\left(\frac{\pi}{2} - \frac{\pi}{n} - \frac{\mathscr{A}_n}{2n}\right) \\ &= \cos\left(\frac{\pi}{2} - \left(\frac{\pi}{n} + \frac{\mathscr{A}_n}{2n}\right)\right) \\ &= \sin\left(\frac{\pi}{n} + \frac{\mathscr{A}_n}{2n}\right). \end{aligned}$$

Setting them equal we get

$$\sin(\pi/n)\cosh(|OB|) = \sin\left(\frac{\pi}{n} + \frac{\mathscr{A}_n}{2n}\right).$$

As n approaches infinity, both sides of the equation approach zero, but if we multiply by n, then

$$n\sin(\pi/n)\cosh(|OB|) = n\sin\left(\frac{\pi + \mathscr{A}_n/2}{n}\right).$$

As n approaches infinity,

- $|OB|$ approaches r ,
- $\mathscr{A}_n$ approaches $\mathscr{A}$,
- the small angle formula describes what happens to the sine terms.

The result is

$$\pi \cdot \cosh(r) = \pi + \mathscr{A}/2,$$

and we can solve for $\mathscr{A}$:

$$\mathscr{A}/2 = \pi\cosh r - \pi \implies \mathscr{A} = 2\pi(\cosh r - 1). \qquad \square$$

The formula looks different from the Euclidean one. Look, though, at what happens when it is rewritten using the Taylor series for hyperbolic cosine:

$$\begin{aligned}\mathscr{A} &= 2\pi\left(\left(\sum_{n=0}^{\infty}\frac{r^{2n}}{(2n)!}\right) - 1\right)\\ &= 2\pi\left(\left(1 + \frac{r^2}{2!} + \frac{r^4}{4!} + \cdots\right) - 1\right)\\ &= 2\pi\left(\frac{r^2}{2} + \frac{r^4}{24} + \cdots\right)\\ &= \pi r^2 + \frac{\pi r^4}{12} + \cdots.\end{aligned}$$

For small values of r, the later terms in the series contribute very little to the value of $\mathscr{A}$. That is, when $r \approx 0$, $\mathscr{A} \approx \pi r^2$. Once again, the hyperbolic world appears to be similar to the Euclidean world when it is viewed at a small enough scale.

46.4 Exercises

46.5. A torus is formed by revolving the circle

$$(x-b)^2 + y^2 = a^2, \quad 0 < a < b,$$

around the y-axis (see exercise 38.2 in lesson 38). Compute the surface area of the torus.

46.6. Let $\triangle ABC$ be a right triangle with vertices $A = 0$, $B = 0.5$, and $C = 0.5i$. First consider AB to be its base, and calculate the lengths of its base and height. Then consider BC as the base, and calculate the lengths of the base and height. Observe that the values of $bh/2$ are not the same.

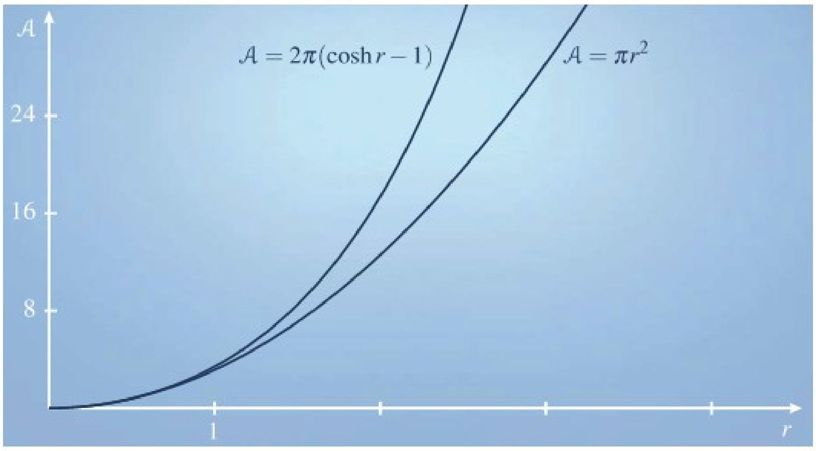

Figure 13. Comparisons of the areas of hyperbolic and Euclidean circles as a function of their radii.

46.7. Compute the hyperbolic area of the triangle with vertices $A = 0$, $B = 0.5$, and $C = 0.5i$.

46.8. In the proof of the formula for the area of a triangle, we considered the case where the vertex B is below AC. Show that the formula is valid if B is above AC.

46.9. Let R be a regular quadrilateral inscribed in a circle of radius r. Find a formula for the area of R as a function of r.

46.10. Prove that there are circles that cannot be contained in the interior of any triangle.

46.11. What is the area of the convex quadrilateral $\square ABCD$ with $A = 0.25 + 0i$, $B = 0 + 0i$, $C = 0 + 0.25i$ and right angles at A, B, and C (see exercise 45.6)?

46.12. What is the area of the Saccheri quadrilateral $\square EFGH$ that has summit vertices $E = (-1/3) + (1/3)i$ and $F = (1/3) + (1/3)i$ and its base along the real axis (see exercise 45.10)?

46.13. Suppose that $\triangle ABC$ is decomposed into a collection of smaller triangles $\{T_i\}$. Explain why the area of $\triangle ABC$ is the sum of their areas. You will need to use Euler's formula for planar polygons, that $v - e + f = 1$, where v is the total number of vertices, e is the total number of edges, and f is the total number of faces in the decomposition.

-7 X 4* & 3]]]
!"#$# % & & !' (! !) * * + , * - . *

/ * ! + , $. 9 , @ 3 * 0 9,1 7 0 * 3 J & +
5 "/" & 6 13
+* 0) * 0 9 7 :0 9 4 0 & :
;. 8 5 6 *1 8 3 9,1 7 *

Chapter 6

Models for the Hyperbolic Plane

6.1 Introduction

The role of models in geometry is tied to the subject's historical connection to axiomatics. While *The Elements* of Euclid is the archetype of an axiomatic system, the real plane, $\mathbb{R}^2$, is the archetype of a model for an axiomatic system, in this case, that of the Euclidean plane. The existence of a model for an axiomatic system is proof that the system is consistent.

We have seen that Euclidean and hyperbolic geometries share so much that we might think of hyperbolic geometry as nearly Euclidean. Indeed, Theorem 2.28 implies that when we look at a region of the hyperbolic plane that is sufficiently small, the angular defect of a triangle is so small as to be indetectable. In practical terms, if space were hyperbolic, our perception of space as Euclidean may be correct, if human scale is "sufficiently small."

Carl Friedrich Gauss (1777–1855) claimed to have discovered, and secreted, Bolyai's and Lobachevsky's results years before the latter two made their work public. He may well have. Gauss was the leading mathematician of his time and he worked on a broad array of problems, among them the problem of the parallels. His influence was such that he could make or break a career and he probably broke Bolyai's career—and spirit—by brushing off the younger man's work when he saw it by saying, in effect, "Yes this is great work. I know because I did it years ago." The story is that he kept his work on non-Euclidean geometries secret because he feared the outcry that would attend its public release.

As we work through our discussion, we will see that the connections between Euclidean and hyperbolic geometries go beyond formalities. Models for the hyperbolic plane and for the Euclidean plane can be constructed from one another. Establishing this was the critical step in the resolution of the problem of the parallels. The relative consistency of hyperbolic geometry meant that the axioms for hyperbolic geometry are consistent if and only if the axioms

© Springer International Publishing AG, part of Springer Nature 2018

M. I. Dillon, *Geometry Through History*,
https://doi.org/10.1007/978-3-319-74135-2_6

for Euclidean geometry are consistent. Understanding this is what enabled the mathematics community to move on, once and for all, from the problem of the parallels.

6.2 Beltrami

Eugenio Beltrami (1835–1900) in 1868 published two articles about hyperbolic geometry and surfaces of constant negative curvature. In the first, he described how to realize the geometry of Lobachevsky on a surface of negative curvature. The date is significant: 1868 is also the year in which Riemann's work on geometry was published.

Georg Bernhard Riemann (1826–1866) studied with Gauss in Göttingen and received his doctoral degree there in 1851. To teach in the university, Riemann had to complete a second project called the *habilitation*. The habilitation culminated in a lecture to the faculty, its topic chosen by the examiners from among different subjects supplied by the candidate. Riemann's topic, chosen by Gauss, was foundations of geometry. The lecture he gave included ideas about generalized surfaces called manifolds, higher-dimensional spaces, and the notion that a surface has intrinsic properties, that is, properties that are independent of any surrounding space. Riemann's work was to have an enormous impact on the mathematics of the nineteenth and twentieth centuries. Indeed, its influence is still felt today. Before the force of his blow could be fully effected, though, Riemann's ideas had to circulate through the community. By 1868, they were known well enough abroad to give Beltrami the fortitude to go ahead with his account about realizing Lobachevsky's geometry. (See [10] and [17].)

The discovery and dissemination of ideas associated to the resolution of the parallel postulate is a story rich in irony and this part of the tale is no exception. Lambert, the eighteenth century mathematician whom we met in the context of ibn al-Haytham-Lambert quadrilaterals, is associated with several ideas in mathematics, among them the problem of the parallels. Quite apart from that, he also studied and developed the theory of hyperbolic trigonometric functions. Though Lambert viewed his assault on the problem of the parallel postulate as a failure, his hyperbolic trigonometry turned out to be a key element in work leading up to Beltrami's first model of a non-Euclidean plane. While Lambert had both things in his hands—the problem of the parallels and hyperbolic trigonometry—the connection between the two escaped him entirely. It would be another 100 years before the problem of the parallels was resolved. (See [17] for details.)

Beltrami's paper starts with an analytic description of a line element on a surface of *constant negative curvature*. There is not a great deal that we will say about this; we want simply to establish some context. Curvature, by which we mean *Gaussian curvature*, is a measure of how a surface bends. As expected, a plane has zero curvature but so does the surface of a cylinder: A cylinder can be "uwrapped" and laid flat on a plane. A sphere has constant positive curvature: constant, because the curvature does not change from point

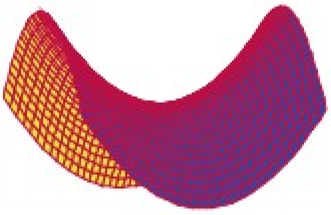

Figure 6.1: A saddle has negative curvature

to point, positive, because if you make a cut in a sphere in an attempt to lay it flat on the plane, pieces of the sphere on either side of the cut have to be pulled apart. The surface of a sphere has too little "material" to form something flat. A hyperbolic paraboloid—a saddle—has negative curvature. (See Fig. 6.1.) If you make a cut in a saddle in an attempt to lay it flat on the plane, pieces of the saddle on either side of the cut have to be pulled together: The surface has too much "material" to form something flat.

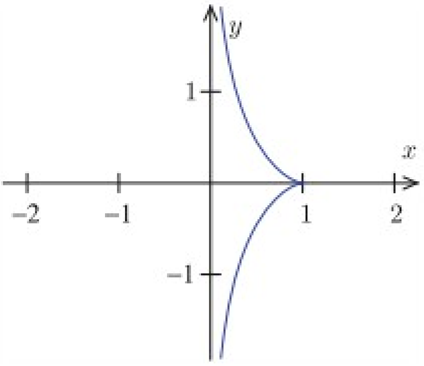

Figure 6.2: The tractrix defined by $x = 1/\cosh t, y = t - \tanh t$

The *pseudosphere*, or *tractoid*, is a surface of constant negative curvature that can be generated by rotating a curve called a *tractrix* about its vertical axis. (See Fig. 6.2.) Beltrami identified a particular region on this surface, and used differential geometry to establish that Lobachevsky's geometry is realized there. An important device in his discussion is the fact that the region of interest on the surface can be mapped onto an open disk in the plane. The circle defining the disk in the plane is the *limit circle*. (See Fig. 6.3.) Its points correspond to points "at infinity" on the surface, indicated in pink in Fig. 6.3. Points outside

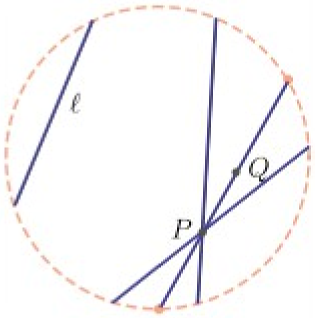

Figure 6.3: The Klein-Beltrami model for the hyperbolic plane

the disk are not part of the model. Lines on the surface are represented by chords of the limit circle. In Fig. 6.3, the line determined by points P and Q is shown, as are several parallels to the line ℓ through the point P. This is the model often referred to as the *Klein-Beltrami model* for the hyperbolic plane.

If Beltrami's work established the relative consistency of Euclidean and non-Euclidean geometry—that is, hyperbolic geometry—Klein's work established the relative consistency of projective geometry, spherical geometry, Euclidean geometry, and non-Euclidean geometry. [28] At this point in our tale, though, the most accessible model for a hyperbolic plane is the Poincaré disk.

Henri Poincaré (1854–1912) started working in hyperbolic geometry when he was thinking about quadratic forms, solutions to certain differential equations, and complex analysis, areas of mathematics that at that time, were not seen as part of geometry. Poincaré's random insight was that the transformations he had been working on were also transformations of the hyperbolic plane. Once his eyes were opened, Poincaré saw the relevance of non-Euclidean geometry all over mathematics. He quickly set about bringing hyperbolic geometry to bear in several different areas of mathematics, including number theory. Non-Euclidean geometry, which until Poincaré's time had seemed a bizarre curiosity, was suddenly everywhere. Poincaré's work might properly signal the true arrival of hyperbolic geometry into mainstream mathematics. (See [28].)

6.3 The Poincaré Disk

There are two models attributed to Poincaré: the upper half-plane model and the disk model. We study the disk model. Like the upper half-plane model, the disk model is constructed within the Euclidean plane, which we designate $\mathcal{E}$. Since $\mathcal{E}$ is modeled by $\mathbb{R}^2$, and we will be concerned with certain mappings from $\mathcal{E}$ to $\mathbb{R}$, calculus is available to us and we employ it to study the Poincaré disk.

Let $\mathcal{C}$ be the circle with center O and radius 1. The choice of 1 for the radius of $\mathcal{C}$ is arbitrary. Let $\mathcal{D}$ be the open disk in $\mathcal{E}$ associated to $\mathcal{C}$. $\mathcal{D}$ is the foundation for our model of a hyperbolic plane. When we think of $\mathcal{D}$ as modeling the hyperbolic plane, we designate it $\mathcal{P}$.

The points of $\mathcal{P}$ are the points of $\mathcal{D}$. Lines in the Poincaré disk are of two types: **type-1 lines** are intersections of diameters of $\mathcal{C}$ with $\mathcal{D}$, and **type-2 lines** are intersections of circles orthogonal to $\mathcal{C}$ with $\mathcal{D}$.

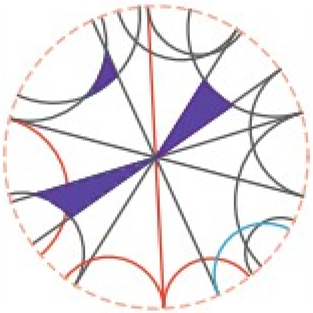

Figure 6.4: The Poincaré disk model

Proceeding through our discussion, we will be dealing with $\mathcal{D}$ sometimes as a subset of $\mathcal{E}$—which, when convenient, we can model using $\mathbb{R}^2$, with which we have some familiarity—and sometimes as modeling $\mathcal{P}$, which we are endeavoring to understand. Going forward, every subset of $\mathcal{D}$ wears two hats: one Euclidean, and one hyperbolic. We must be careful to distinguish which hat it wears when we pick up an object to discuss its nature.

We continue to refer to Euclidean lines simply as lines. Following [23], we refer to type-1 and type-2 lines together as **L-lines** ("L" for Lobachevsky).

A goal for us right now is understanding what it means for one point to be between two other points in our model, $\mathcal{P}$. Once we establish that we have a well-behaved notion of betweenness in $\mathcal{P}$, we will have license to talk about L-segments. In other words, if A and B are points on an L-line ℓ, then the associated L-segment is the collection of points $X \in \ell$ so that $X = A$, or $A * X * B$, or $X = B$.

Angles in $\mathcal{P}$ are identical to angles in $\mathcal{E}$. The Euclidean tangents at the point of intersection of two L-lines determine the angles formed by the L-lines.

Figure 6.4 shows some L-lines and regions in $\mathcal{P}$. The red L-lines are all parallel. The light blue arc, a type-2 line, is parallel to three of the red L-lines, but not to the fourth. This is significant: In a neutral plane, transitivity of parallelism is one of the conditions equivalent to the parallel postulate. That parallelism is not transitive is thus a defining feature of a hyperbolic plane. Three points in $\mathcal{P}$ that do not lie on a single L-line determine an **L-triangle**. The filled areas in the figure are regions enclosed by L-triangles.

Our first task is to show that $\mathcal{P}$ satisfies the first postulate of Euclid, that is, Axioms I.1–2 of Hilbert. These are the only incidence axioms we need. The

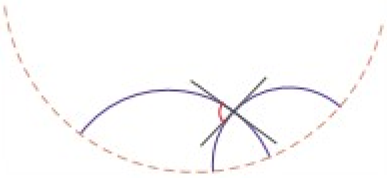

Figure 6.5: An angle between L-lines is an angle between their tangents

others are either automatic, since our model employs Euclidean points, or irrelevant, since we are discussing a plane.

Lemma 6.1. Two points in $\mathcal{P}$ determine a unique L-line.

Proof. Let P and Q be points in $\mathcal{P}$. If P belongs to a Euclidean diameter of $\mathcal{C}$ that runs through Q, then that diameter determines a unique type-1 line containing P and Q. Suppose in this case that there were also a type-2 line containing P and Q. That L-line would be determined by a circle $\mathcal{C}'$ in $\mathcal{E}$ orthogonal to $\mathcal{C}$. By Lemma 5.21, both P and Q would have to be different from O. Under inversion through $\mathcal{C}$, $\overline{PQ} \setminus O$ is mapped to itself. As an orthogonal to $\mathcal{C}$, $\mathcal{C}'$ is also mapped to itself under inversion through $\mathcal{C}$. If the $\mathcal{C}$ inverses of P and Q are P' and Q' respectively, this means P, Q, P', Q' would belong both to a line in $\mathcal{E}$ and to a circle in $\mathcal{E}$, a violation of Prop. III.10. We conclude that if P and Q lie on a diameter of $\mathcal{C}$, there is a unique L-line that contains the two points.

Next suppose P and Q do not lie on a diameter of $\mathcal{C}$. In this case neither P nor Q can be O. We must show that there is a single type-2 line containing them. Let P' be the $\mathcal{C}$ inverse of P so that the Euclidean line determined by P and P' is $\overline{OP}$. Then P, Q, and P' are noncollinear, so by Prop. III.10, they determine a unique circle, $\mathcal{C}'$. By Theorem 5.24, $\mathcal{C}$ and $\mathcal{C}'$ are orthogonal, thus, in this case, P and Q determine a type-2 line. Also by Theorem 5.24, any circle orthogonal to $\mathcal{C}$ and containing P and Q must contain their inverses, so the L-line determined by P and Q is unique, again by Prop. III.10. □

Let ℓ be an L-line in $\mathcal{P}$. In $\mathcal{E}$, the Euclidean arc or line segment that determines ℓ intersects $\mathcal{C}$ in two points. We call these the **points at infinity** of ℓ. Note that these are Euclidean points that are not in $\mathcal{P}$.

We can establish that both the axioms of order and the continuous nature of a line hold in $\mathcal{P}$ by coordinatizing L-lines using $\mathbb{R}$. Once we coordinatize lines, several of Hilbert's axioms will be verified immediately.

First, we consider a notion of non-Euclidean distance in $\mathcal{P}$.

Distance in the Poincaré Disk

A measure of distance on a set that conforms to our understanding of how distance ought to work is usually defined by a *metric*.

Definition 6.2. A **metric** on a set $\mathcal{S}$ is a mapping $d : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ that satisfies the following for all x, y, z in $\mathcal{S}$:

(1) $d(x, y) \geq 0$ with equality only if $x = y$;

(2) $d(x, y) = d(y, x)$;

(3) $d(x, z) \leq d(x, y) + d(y, z)$.

When $\mathcal{S}$ has a metric d, $(\mathcal{S}, d)$ is a **metric space**.

The last condition of Definition 6.2 is a version of Euclid's Triangle Inequality, Prop. I.20. In this case, the inequality is not strict because x, y, z are not necessarily vertices of a triangle. Going forward, we use "triangle inequality" to refer to condition (3) in Definition 6.2, or to

$$|a + b| \leq |a| + |b|,$$

the triangle inequality as it arises in $\mathbb{R}$, or to the inequality in the next lemma.

Lemma 6.3. Let x, y, z be points in a metric space $(\mathcal{S}, d)$. Then

$$d(x, y) \geq |d(x, z) - d(y, z)|.$$

Proof. Since $d(x, z) \leq d(x, y) + d(y, z)$, we have $d(x, z) - d(y, z) \leq d(x, y)$. On the other hand, $d(y, z) \leq d(x, y) + d(x, z)$ implies $d(y, z) - d(x, z) \leq d(x, y)$. Putting the two together, we get the result. □

The archetype for a metric space is $\mathbb{R}^n$ with

$$d(P, Q) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{6.1}$$

where $P = (x_1, x_2, \ldots, x_n)$ and $Q = (y_1, y_2, \ldots, y_n)$. Notice that in $\mathbb{R}$, (6.1) becomes

$$d(x, y) = |x - y|.$$

Proofs that (6.1) satisfies the triangle inequality abound in the literature. We leave it as an exercise to prove the case for $n = 1$. Equation (6.1) defines the **Euclidean metric** on $\mathbb{R}^n$.

Lemma 6.4. Let $(\mathcal{S}, d)$ be a metric space. Let $A \subseteq \mathcal{S}$. Then (A, d) is also a metric space.

Proof. Let $x, y \in A$. Since $x, y \in \mathcal{S}$, properties (1) and (2) of Definition 6.2 hold. Now let z be an element in A as well. Since x, y, z all belong to $\mathcal{S}$, property (3) of Definition 6.2 must hold as well. □

Since $\mathcal{D}$ is a subset of a metric space, $\mathbb{R}^2$, it is also a metric space under the Euclidean metric. The problem with this is that Postulate II will not hold in this metric space. We need a metric on $\mathcal{P}$ that will allow us to stretch line segments indefinitely.

Definition 6.5. Let P and Q be points in $\mathcal{P}$, and let ℓ be the L-line they determine. Let S and T be the points at infinity of ℓ. The **L-distance between P and Q** is

$$L(P,Q) = \left| \ln \frac{|PS|/|PT|}{|QS|/|QT|} \right| \tag{6.2}$$

where $|PS|$ is the Euclidean distance between points P and S, (or the Euclidean length of PS), and ln is the natural logarithm.

Theorem 6.6. *L is a metric on $\mathcal{P}$.*

Proof. Notice first that when $P = Q$, $\frac{|PS|/|PT|}{|QS|/|QT|} = 1$, and $\ln 1 = 0$. Note further that by the definition, the L-distance between points is always nonnegative.

The triangle inequality and condition (2) of Definition 6.2 are easy to show using properties of logarithms and absolute values. Likewise, if P, Q determine a type-1 line, it is easy to show that $L(P,Q) = 0$ implies $P = Q$. We leave those verifications to the exercises. We must show that condition (1) holds if P, Q determine a type-2 line.

Suppose then that $P \neq Q$ lie on a type-2 line in $\mathcal{P}$. Notice that $\overrightarrow{SP}$ and $\overrightarrow{SQ}$ must be different, otherwise S, P, Q would lie on both a line and a circle, in violation of Prop. III.10. Since $\overrightarrow{SQ}$ and $\overrightarrow{SP}$ lie on the same side of ST, $\angle TSP$ and $\angle TSQ$ must be different.

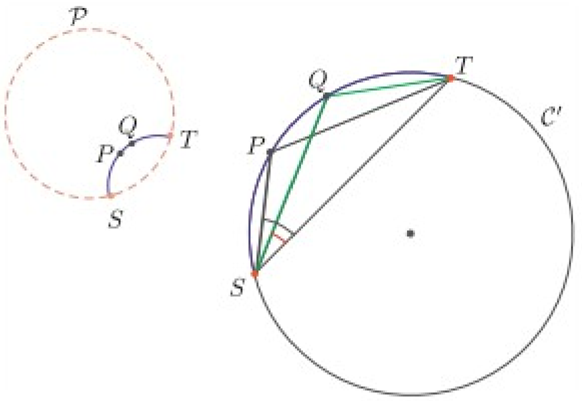

Figure 6.6: A type-2 line in $\mathcal{P}$ and the underlying set in $\mathcal{E}$

Assume now that $L(P,Q) = 0$. We then have

$$\frac{|PS|/|PT|}{|QS|/|QT|} = 1, \text{ that is, } \frac{|PS|}{|QS|} = \frac{|PT|}{|QT|}.$$

By Prop. III.21, $\angle SPT \cong \angle SQT$. We thus have SAS for similar Euclidean triangles (Prop. VI.6) $\Delta STP \sim \Delta STQ$. That implies $\angle TSP \cong \angle TSQ$, which is a contradiction. We conclude that if $L(P, Q) = 0$, then $P = Q$ as desired. $\square$

As we proceed through our discussion, $|PQ|$ will always designate the Euclidean length of the line segment PQ, or equivalently, the Euclidean distance between P and Q.

The notion of a metric allows us to consider what it means for mappings to be continuous on spaces that are more general than $\mathbb{R}$ with the Euclidean metric.

Definition 6.7. Let f be a mapping from a metric space $(\mathcal{S}_1, d_1)$ to a metric space $(\mathcal{S}_2, d_2)$. Then f is **continuous at a point** $x \in \mathcal{S}_1$ provided for any real number $\epsilon > 0$ there is a real number $\delta > 0$ so that if for some $y \in \mathcal{S}_1$, $d_1(x, y) < \delta$, then $d_2(f(x), f(y)) < \epsilon$.

Continuous mappings on metric spaces have many of the properties we are used to from calculus. We have to be careful, though, to bear in mind that an arbitrary metric space may not come equipped with operations such as addition or multiplication, for example. In general, a sum of mappings between metric spaces may not make sense. Composition of mappings, though, makes sense as long as the domain of the second mapping contains the range of the first.

Lemma 6.8. Let $(\mathcal{S}_1, d_1)$, $(\mathcal{S}_2, d_2)$, and $(\mathcal{S}_3, d_3)$ be metric spaces. Let $f : \mathcal{S}_1 \to \mathcal{S}_2$ be a continuous mapping and let $g : \mathcal{R} \to \mathcal{S}_3$ be a continuous mapping, where $\mathcal{R} \subseteq f(\mathcal{S}_1)$. Then $g \circ f$ is a continuous mapping from $\mathcal{S}_1$ to $\mathcal{S}_3$.

Proof. Let x be in $\mathcal{S}_1$. Given $\epsilon > 0$, we must produce $\delta > 0$ so that if $d_1(x, y) < \delta$ for $y \in \mathcal{S}_1$, then $d_3(g(f(x)), g(f(y))) < \epsilon$.

Since g is continuous, we know there is $\xi > 0$ so that if $d_2(f(x), f(y)) < \xi$ for $f(y) \in f(\mathcal{S}_1)$, then $d_3(g(f(x)), g(f(y))) < \epsilon$. With that ξ fixed, we can in turn choose $\delta > 0$ so that if $d_1(x, y) < \delta$ for some $y \in \mathcal{S}_1$, then $d_2(f(x), f(y)) < \xi$, from which it follows that $d_3(g(f(x)), g(f(y))) < \epsilon$, as was to be shown. $\square$

Now that we have a metric on $\mathcal{P}$ and a notion of continuity, we can approach the problem of coordinatizing L-lines using $\mathbb{R}$. We start by considering the same problem in the more familiar setting of $\mathbb{R}^2$.

Let ℓ be the line in $\mathbb{R}^2$ that passes through points (x_0, y_0) and (a, b). When we think about identifying the points of ℓ with the points of $\mathbb{R}$, we are usually thinking of something more than just a point-by-point correspondence. An "identification" should show some respect to distances between points of ℓ and distances between corresponding points of $\mathbb{R}$. Even so, there are many different ways to make this sort of identification using a parametrization. We can parametrize ℓ with a variable t and then think of a point X on ℓ as identified with the associated $t \in \mathbb{R}$. For example,

$$x = x_0 + t(a - x_0),\ y = y_0 + t(b - y_0),\ t \in \mathbb{R}. \tag{6.3}$$

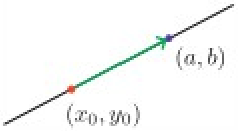

Figure 6.7: A line in $\mathbb{R}^2$ with a "starting point" and a direction

Here the point (x_0, y_0) is identified with 0, and the point (a, b) is identified with 1. The vector $\mathbf{v} =< a - x_0, b - y_0 >$ (shown in Fig. 6.7 in green) determines the positive direction on ℓ: As t increases, the associated points on ℓ can be found farther to the right of (x_0, y_0). Note, moreover, that points on ℓ that are close, are associated to values of t that are also close. Equation (6.3) thus gives us a way of assigning coordinates to the points on ℓ, so that distances between coordinates of points bear a certain fidelity to distances between points on ℓ.

Let us make this more precise.

Definition 6.9. A **coordinatization** of a curve $\mathcal{S}$ in a space with a metric d is a continuous bijective mapping $f : \mathcal{S} \to \mathbb{R}$.

The bijectivity condition guarantees a point-by-point correspondence between $\mathcal{S}$ and $\mathbb{R}$. The continuity condition guarantees that a coordinatization respects the notions of nearness on the curve $\mathcal{S}$ and in $\mathbb{R}$.

We get to our coordinatization through a sequence of lemmas.

Lemma 6.10. *Let ℓ be an L-line in $\mathcal{P}$. Let T, S be the points at infinity of ℓ. Let $|ST| = \tau$. Define $R : \ell \to \mathbb{R}$ by $R(X) = |XT|$. If we view ℓ as a subset of $\mathbb{R}^2$ with the Euclidean metric, then R is a continuous bijective mapping of ℓ onto the interval $(0, \tau) \subset \mathbb{R}$.*

Proof. Let everything be as hypothesized and suppose X is fixed on ℓ. Let $\epsilon > 0$ be given. If we take $\delta = \epsilon$, then if $|XY| < \delta$ for some $Y \in \ell$, Lemma 6.3 gives us

$$d(R(X), R(Y)) = ||XT| - |YT|| \leq |XY| < \epsilon,$$

which shows that R is continuous.

Next we show that R is bijective onto $(0, \tau)$. The case in which ℓ is a type-1 line is routine and we leave it to the reader. Suppose then that ℓ is a type-2 line.

To show that R is one-to-one we suppose it is not and seek a contradiction. Suppose then that for two different points X, Y on ℓ, $R(X) = R(Y)$. Since $|XT| = |YT|$, ΔXYT is isosceles. By Exercise 5.4.7, the median of ΔXYT from T (shown in red in Fig. 6.8) is the perpendicular bisector of XY. But XY is a chord of $\mathcal{C}'$, the Euclidean circle that determines ℓ. By Corollary 5.2, the perpendicular bisector of XY passes through the center of $\mathcal{C}'$. Since it also passes through T and a point on $\widehat{ST}$, that means the center of $\mathcal{C}'$ is interior to

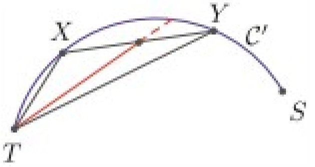

Figure 6.8: $|XT| = |YT|$ implies ΔXYT is isosceles

$\mathcal{C}$, the circle that determines $\mathcal{P}$. By Lemma 5.21 that is impossible because $\mathcal{C}'$ and $\mathcal{C}$ are orthogonal. We conclude that R is indeed one-to-one.

Finally, we show that R is onto the interval $(0, \tau)$. Notice that for $X \in \ell$, $|XT| \leq |XS| + |ST|$ so $|XT| < |ST|$. This shows that $R(X) \in (0, \tau)$. Let $\rho \in (0, \tau)$. Consider the circle in $\mathcal{E}$ centered at T with radius ρ. That circle must intersect $\widehat{ST}$ at some point X. It follows that $R(X) = \rho$, which establishes that R is onto $(0, \tau)$ and completes the proof. ☐

The proof of our next major result benefits from a theorem we learn in middle school or high school. This is close to Prop. II.11, 13, although Euclid does not mention cosines.

> **The Law of Cosines** Let a, b, c be the lengths of sides of a triangle and suppose the side with length a subtends the angle α. Then
>
> $$a^2 = b^2 + c^2 - 2bc \cos \alpha.$$

Lemma 6.11. Let ℓ be an L-line in $\mathcal{P}$ with points at infinity S and T. Let $\tau = |ST|$. If ℓ is a type-2 line, let $\varphi = \angle TXS$, for any X on ℓ. If ℓ is a type-1 line, let $\varphi = \pi$. If $x = |XT|$ and $y = |XS|$, then

$$y = x \cos \varphi + \sqrt{\tau^2 - x^2 \sin^2 \varphi}.$$

Proof. When ℓ is a type-1 line, $\cos \varphi = -1$ and $\sin^2 \varphi = 0$. The result in this case follows nearly immediately.

Suppose ℓ is a type-2 line determined by a circle $\mathcal{C}'$. Since the base on the circumference of $\mathcal{C}'$ for $\widehat{ST}$ is the major arc associated to ST, φ is obtuse, so $\cos \varphi < 0$. (The verification that φ is obtuse is left as an exercise.)

The Law of Cosines gives us $\tau^2 = x^2 + y^2 - 2xy \cos \varphi$ so that

$$y^2 - (2x \cos \varphi) y + (x^2 - \tau^2) = 0.$$

By the quadratic formula,

$$y = \frac{2x \cos \varphi \pm \sqrt{4x^2 \cos^2 \varphi - 4(x^2 - \tau^2)}}{2} = x \cos \varphi \pm \sqrt{x^2(\cos^2 \varphi - 1) + \tau^2}.$$

Since $\cos\varphi < 0$, the positive square root must be what gives us y. Note too that $\cos^2\varphi - 1 = -\sin^2\varphi$. This implies the result. $\square$

Notice that by Prop. III.21, the angle φ in Lemma 6.11 is constant.

Now we dust off a few more tools from single variable calculus.

Recall that an *algebraic function* is a mapping from a subset of $\mathbb{R}$ to $\mathbb{R}$ that is defined using arithmetic (addition, subtraction, multiplication, and division), powers, and roots. Examples of algebraic functions are polynomials, rational functions, $f(x) = \sqrt{x}$, and $g(x) = \sqrt[3]{\frac{x^2+1}{3x^2+5}}$. Trigonometric functions, hyperbolic functions, exponentials, and logarithms are examples of non-algebraic functions. Algebraic functions are continuous on their domains.

Recall also the following important theorem from Calculus I.

Theorem 6.12 (Intermediate Value Theorem). Let f be a continuous function from an interval $I \subseteq \mathbb{R}$ to $\mathbb{R}$. If $f(a) < f(b)$ for some a, b in I, then for any $y \in (f(a), f(b))$ there is x between a and b so that $f(x) = y$.

Lemma 6.13. Let ℓ be an L-line in $\mathcal{P}$ with points at infinity S and T. Let $\tau = |ST|$. If ℓ is a type-2 line, let $\varphi = \angle TXS$, for any X on ℓ. If ℓ is a type-1 line, let $\varphi = \pi$. For $x \in (0, \tau)$, define

$$g(x) = \frac{x}{x\cos\varphi + \sqrt{\tau^2 - x^2\sin^2\varphi}}.$$

Then g is a continuous bijection from $(0, \tau)$ onto $\mathbb{R}_+$.

Proof. That g is continuous on its domain follows from the fact that g is algebraic in x. (Note that τ, and φ are constants so $\cos\varphi$ and $\sin\varphi$ are also constants.) In Lemma 6.11 we established that the denominator of g is $|XS|$, thus for $x \in (0, \tau)$, it also lies strictly between 0 and τ. This confirms that g is defined and continuous on all of $(0, \tau)$.

Note that g is differentiable on its domain with

$$g'(x) = \frac{\tau^2}{(x\cos\varphi + \sqrt{\tau^2 - x^2\sin^2\varphi})^2\sqrt{\tau^2 - x^2\sin^2\varphi}}.$$

As $g'(x) > 0$ for all $x \in (0, \tau)$, g is strictly increasing. This shows that g is one-to-one.

Next notice that

$$\lim_{x\to 0^+} g(x) = 0,$$

and that

$$\lim_{x\to\tau^-} g(x) = \lim_{x\to\tau^-} \frac{1}{\cos\varphi + \sqrt{\frac{\tau^2}{x^2} - \sin^2\varphi}}.$$

Since $\cos\varphi < 0$, and $\sqrt{1 - \sin^2\varphi} = |\cos\varphi|$, the denominator $\cos\varphi + \sqrt{\frac{\tau^2}{x^2} - \sin^2\varphi}$ approaches 0 from above as x approaches τ from below, thus $\lim_{x\to\tau^-} g(x) = \infty$. By the Intermediate Value Theorem, g must assume all values in $\mathbb{R}_+$, thus, g is onto $\mathbb{R}_+$. $\square$

Theorem 6.14. Let ℓ be an L-line in $\mathcal{P}$. Let P be a fixed point on ℓ and suppose S and T are the points at infinity of ℓ. Let

$$f(X) = \ln \frac{|PS|/|PT|}{|XS|/|XT|}. \tag{6.4}$$

Then f is a coordinatization of ℓ by $\mathbb{R}$.

Proof. Let everything be as hypothesized.

We may view f as a composition, $f = f_1 \circ f_2$, where $f_2 : \ell \to \mathbb{R}_+$ by $f_2(X) = \frac{|PS|/|PT|}{|XS|/|XT|}$, and $f_1 : \mathbb{R}_+ \to \mathbb{R}$ is the natural logarithm function. Bijectivity and continuity of $f : \ell \to \mathbb{R}$ will follow if we show that f_1 and f_2 are both bijective and continuous.

By our calculus courses, we know that the natural logarithm is a bijective continuous mapping from $\mathbb{R}_+$ onto $\mathbb{R}$, so we need only address the qualifications of f_2. Notice that

$$f_2 = \frac{|PS|}{|PT|} g \circ R,$$

where $|PS|/|PT|$ is a constant, $R : \ell \to (0, \tau)$, $\tau = |ST|$, is the mapping we discussed in Lemma 6.10, and $g : (0, \tau) \to \mathbb{R}_+$ is the mapping from Lemma 6.13. As f_2 is the composition of continuous bijections, it must also be a continuous bijection. We leave the details underlying that statement to the reader. □

Suppose now that ℓ is an L-line with the line coordinatization f given (6.4). We can order the points on ℓ according to their coordinates: For points A, B, C on ℓ, say B is between A and C provided $f(A) < f(B) < f(C)$ or $f(C) < f(B) < f(A)$. That Hilbert's axioms of line order and continuity apply in $\mathcal{P}$ then follows from properties of $\mathbb{R}$, properties one studies, for instance, in a first course in real analysis. It is not too difficult, and we leave it as an exercise, to show that Pasch's Axiom holds in $\mathcal{P}$.

The L-metric on $\mathcal{P}$ allows us to define congruence of segments: L-segments AB and CD are congruent if and only if $L(A, B) = L(C, D)$. Since we are transferring the problem of comparing segments to the problem of comparing numbers, difficulties that Hilbert had to manage disappear. For instance, congruence of segments is an equivalence relation because equality of numbers is an equivalence relation. Angles in $\mathcal{P}$ are identical to angles in $\mathcal{E}$, so the axioms that apply to angle congruence in $\mathcal{E}$ carry over to $\mathcal{P}$. Note that once Hilbert's axioms are established, the associated theorems follow. We do not have to prove any of them.

Determining that SAS holds in $(\mathcal{P}, L)$ is more interesting. This is where congruence of segments and congruence of angles meet, as it were. Our strategy for proving SAS involves the use of L-line reflections.

L-Reflections on $\mathcal{P}$

Definition 6.15. The **type-1 reflections** on $\mathcal{P}$ are Euclidean reflections determined by type-1 lines. The **type-2 reflections** on $\mathcal{P}$ are circular inversions determined by type-2 lines. The type-1 and type-2 reflections together are called **L-reflections** on $\mathcal{P}$.

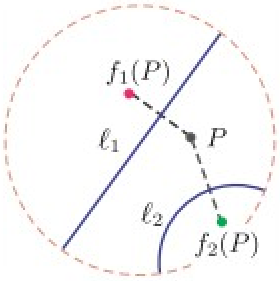

Figure 6.9: Two types of L-reflections

Just as Euclidean reflection leaves Euclidean distance unchanged, a type-2 reflection on $\mathcal{P}$ leaves L-distance unchanged. To understand this, we take a moment to consider the ratio of Euclidean segment lengths in (6.2). That ratio is the **planar cross ratio** of Euclidean points P, Q, S, T.

Lemma 6.16. The planar cross ratio of points in $\mathcal{E} \setminus O$ is invariant under circular inversion of $\mathcal{E} \setminus O$.

Proof. Consider points P, Q, S, T in $\mathcal{E} \setminus O$ and let P', Q', S', T' be their respective images under an inversion with radius a and center O. Assuming noncollinearity, we have the following four pairs of similar triangles, by Observation 5.16: $\Delta POS \sim \Delta S'OP'$, $\Delta POT \sim \Delta T'OP'$, $\Delta QOS \sim \Delta S'OQ'$, and $\Delta QOT \sim \Delta T'OQ'$. It follows that

$$\frac{|PS|}{|OP|} = \frac{|P'S'|}{|OS'|}, \text{ and } \frac{|PT|}{|OP|} = \frac{|P'T'|}{|OT'|}. \tag{6.5}$$

Dividing the second equation into the first we get

$$\frac{|PS|}{|PT|} = \frac{|P'S'|}{|P'T'|}\frac{|OT'|}{|OS'|}.$$

Similarly,

$$\frac{|QS|}{|QT|} = \frac{|Q'S'|}{|Q'T'|}\frac{|OT'|}{|OS'|}.$$

Dividing the latter two equations, we get

$$\frac{|PS|}{|PT|}\frac{|QT|}{|QS|} = \frac{|P'S'|}{|P'T'|}\frac{|Q'T'|}{|Q'S'|},$$

as desired.

Next, consider what happens if O, P, S are collinear. We may assume the points are labeled so that $|PS| = |OS| - |OP|$ and $|P'S'| = |OP'| - |OS'|$. Then $|PS||OS'| = a^2 - |OP||OS'|$, and $|P'S'||OP| = a^2 - |OS'||OP|$ so $|PS||OS'| = |P'S'||OP|$ implying that the first statement in (6.5) is still true. Thus, if any of the four "triangles" is not actually a triangle, we get the same equations and we can still conclude that the planar cross ratio is unchanged by inversion. □

Theorem 6.17. The L-metric is invariant under circular inversion.

Proof. The proof is immediate by Lemma 6.16. □

We study Euclidean reflections and rotations in some detail in later chapters after developing the requisite algebraic machinery. For now, we simply observe that under reflection across a diagonal, as well as rotation about its center, $\mathcal{P}$ is mapped to itself, L-lines are mapped to L-lines, and Euclidean length and angle are preserved. Our intuition tells us this must be so, and for now, we leave it at that.

Since Euclidean length is preserved, L-length is also preserved under type-1 reflections. (We leave the verification as an exercise.) Theorem 5.23 implies that a type-2 reflection on $\mathcal{P}$ leaves $\mathcal{P}$ fixed as a set. By Lemma 5.25, angles formed by L-lines are unchanged by inversion. L-reflections, then, map $\mathcal{P}$ to $\mathcal{P}$, and respect L-length and angle. The next result is less obvious.

Lemma 6.18. A type-2 reflection of $\mathcal{P}$ maps L-lines to L-lines.

Proof. We continue to use $\mathcal{C}$, with center O and radius 1, as the circle in $\mathcal{E}$ defining $\mathcal{P}$. Let ℓ be a type-2 line in $\mathcal{P}$ with underlying circle $\mathcal{C}'$ with center A and radius a. From Exercise 5.3.9, we know that $\mathcal{C}$ and $\mathcal{C}'$ are orthogonal if and only if

$$a^2 + 1 = b^2, \tag{6.6}$$

where $b = |OA|$.

Fix a coordinate system on $\mathcal{E}$ based on $\mathcal{C}'$, so that A has rectangular coordinates $(0, 0)$, O has coordinates $(b, 0)$, and $\mathcal{C}'$ is given by $x^2 + y^2 = a^2$. We use rectangular and polar coordinates to show that a type-1 line in $\mathcal{P}$ is mapped to a type-2 line under the type-2 reflection determined by $\mathcal{C}'$. Because circular inversion is an involution, this is enough to show that type-2 reflections map L-lines to L-lines in $\mathcal{P}$.

Note that a diameter of $\mathcal{C}$ belongs to a line that can be described in rectangular coordinates by

$$x - b = cy,$$

for some $c \in \mathbb{R}$. Switching to polar coordinates, the line is given by $r\cos\theta - b = cr\sin\theta$. Under inversion through radius a, the line is mapped to the circle in $\mathcal{E}$ given by

$$a^2 r\cos\theta - br^2 = a^2 cr\sin\theta + br^2.$$

Switching back to rectangular we have

$$x^2 - \frac{a^2}{b}x + y^2 + \frac{a^2 c}{b}y = 0,$$

which can be written

$$\left(x-\frac{a^2}{2b}\right)^2+\left(y+\frac{a^2c}{2b}\right)^2=\frac{a^4(1+c^2)}{4b^2}.$$

The line through O is thus mapped to the circle $\mathcal{C}''$ in $\mathcal{E}$ with radius $a^2\sqrt{1+c^2}/(2b)$ and center Q (in rectangular coordinates) given by

$$Q=\left(\frac{a^2}{2b},-\frac{a^2c}{2b}\right).$$

Again invoking Exercise 5.3.9, we know that $\mathcal{C}''$ is orthogonal to $\mathcal{C}$ if and only if $|OQ|^2$ is the sum of the squares of the radii of $\mathcal{C}$ and $\mathcal{C}''$. We see then that ℓ, the L-line underlying the diameter of $\mathcal{C}$, is mapped to a type-2 line if and only if

$$|OQ|^2=1+\frac{a^4(1+c^2)}{4b^2}.$$

With an eye toward Equation (6.6), we see that

$$|OQ|^2=\left(\frac{a^2}{2b}-b\right)^2+\left(-\frac{a^2c}{2b}\right)^2=\frac{a^4}{4b^2}-a^2+b^2+\frac{a^4c^2}{4b^2}$$

$$=\frac{a^4(1+c^2)}{4b^2}+(b^2-a^2)=\frac{a^4(1+c^2)}{4b^2}+1,$$

as was to be shown. □

We can proceed now under the assumption that any L-reflection maps an L-line to an L-line.

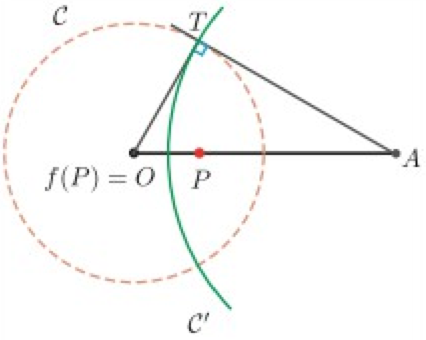

Figure 6.10: Every point can be L-reflected to O

Lemma 6.19. For every point P in $\mathcal{P}$ there is an L-reflection f with $f(P)=O$.

Proof. Given a point $P \in \mathcal{P}$ different from O, let $k = |OP|$. Since the radius of $\mathcal{C}$ is 1, $k < 1$. Let $A \in \overrightarrow{OP}$ satisfy $|OA| = 1/k$. (See Fig. 6.10.) Notice that since $0 < k < 1$, $1/k > 1$ so A is outside $\mathcal{C}$.

Form a tangent to $\mathcal{C}$ from A and say the point of tangency is T. Let $a = |AT|$. Notice that the circle $\mathcal{C}'$ with center A and radius a is indeed orthogonal to $\mathcal{C}$ so it does determine a type-2 line in $\mathcal{P}$, thus an L-reflection, f. We claim that $f(P) = O$. It suffices to show that $|Af(P)| = 1/k$.

We have $|AP||Af(P)| = a^2$. Note

$$|AP| = \frac{1}{k} - k = \frac{(1-k^2)}{k}.$$

By the Pythagorean Theorem, $|OA|^2 = 1 + a^2$ so

$$a^2 = \frac{1}{k^2} - 1 = \frac{1-k^2}{k^2} = \frac{(1-k^2)}{k}\frac{1}{k} = |AP|\frac{1}{k} = |AP||Af(P)|$$

in other words, $|Af(P)| = 1/k$, as was to be shown. □

Hilbert's Axiom III.5 says that if ΔABC and $\Delta A'B'C'$ satisfy $|AB| \cong |A'B'|$, $\angle BAC \cong \angle B'A'C'$, and $|AC| \cong |A'C'|$, then $\angle ABC \cong \angle A'B'C'$. In Hilbert's narrative, it is a short leap from Axiom III.5 to triangle congruence. With the tools in our hands, the route to the result that SAS implies L-triangle congruence is even more direct. This approach to the proof is from [18] and [23].

Theorem 6.20. SAS implies L-triangle congruence in $\mathcal{P}$.

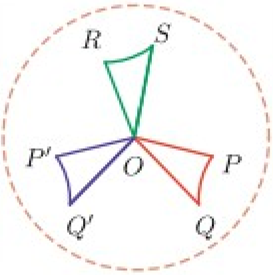

Figure 6.11: Euclidean reflection maps $\Delta OP'Q'$ to ΔOPQ. Rotation maps ΔORS to ΔOPQ

Sketch of proof. Let ΔABC and $\Delta A'B'C'$ be L-triangles in $\mathcal{P}$ with

$$L(A, B) = L(A', B'),\ L(A, C) = L(A', C'),\ \text{and } \angle BAC \cong \angle B'A'C'.$$

Let f be an L-reflection that maps A to O, the center of the Euclidean circle defining $\mathcal{P}$. Since L-reflections respect L-length, L-lines, and angle, $f(\Delta ABC) = \Delta OPQ$ is congruent to ΔABC.

Letting g be the L-reflection that maps A' to O, we get a fourth L-triangle

$$g(\Delta A'B'C') = \Delta OP'Q' \cong \Delta A'B'C'.$$

Note now that the sides OP, OQ, OP', OQ' all lie on diagonals of $\mathcal{C}$. Since $\angle POQ \cong \angle P'OQ'$, we can use either a Euclidean reflection or rotation to map $\Delta OP'Q'$ to a new L-triangle, $\Delta OP''Q''$, where P'' is on OP and Q'' is on OQ. Because

$$L(O,Q) = L(O,Q') = L(O,Q''), \text{ and } L(O,P) = L(O,P') = L(O,P''),$$

points P and P'' have to coincide, as do points Q and Q''. Since two points in $\mathcal{P}$ determine a unique L-line, the L-segments PQ and $P''Q''$ must coincide as well. Since the legs of $\angle QPO$ and $\angle Q''P''O$ coincide, the angles must be congruent as well. The same argument establishes that $\angle PQO \cong \angle P''Q''O$ so that

$$\Delta OPQ \cong \Delta OP''Q'' \cong \Delta OP'Q'.$$

Since congruence is an equivalence relation, we have $\Delta ABC \cong \Delta A'B'C'$, which proves the theorem. □

Authors who use this argument to prove SAS in $\mathcal{P}$ (see [11], besides [18] and [23], for example) usually note that it employs superposition, the method that Euclid used to argue Prop. I. 4. Remember that Euclid has been chastised through the ages for using superposition! Superposition is a valid technique if it is developed as part of the discussion of allowable transformations or *isometries* on the geometry in question. We have handled circular inversion fairly thoroughly, and appealed to intuition in our use of Euclidean reflections and rotations. In the next chapter, we meet transformations formally and in earnest.

Exercises 6.1. 1. Is the following statement true or false? Justify your answer with a proof, if it is true. If it is false, adjust the statement to make it true and supply details: "Any set of three noncollinear points in $\mathcal{P}$ is noncollinear in both the Euclidean and the hyperbolic sense, thus, by identifying a triangle with its vertices, we find that the collection of Euclidean triangles and the collection of L-triangles in $\mathcal{P}$ is the same."

2. There is a statement in the proof of Lemma 6.1 that says "P, Q and P' are noncollinear." Why is this true?

3. (a) Show that $d(x,y) = |x-y|$ is the Euclidean metric on $\mathbb{R}$.

   (b) Show that for x, y in $\mathbb{R}$, $|x+y| \le |x| + |y|$.

4. Finish the verifications that were left out of the proof of Theorem 6.6.

5. Consider points P, Q in $\mathcal{P}$ with S, T the points at infinity of the associated L-line. What happens to $L(P, Q)$ if we keep Q fixed but move P closer (in the Euclidean sense) to S? What happens if we move P closer to T? Now think of orthogonal circles that are very small at the edge of $\mathcal{D}$ and their associated L-lines in $\mathcal{P}$. What does the L-distance tell us about the type-2 lines that those circles determine? In other words, would those type-2 lines somehow seem shrunken in a hyperbolic world, as the underlying Euclidean circles are so small?

6. Show that by parametrizing the line ℓ as we did in (6.3), we get a line coordinatization that satisfies Definition 6.9.

7. Parametrize the line in $\mathbb{R}^2$ that passes through the points $(-1, 1)$ and $(2, 3)$, in three different ways, using the formula as given in (6.3), as follows. In each case, identify the point on ℓ associated with the real number 2.

    (a) Identify $(-1, 1)$ with 0 and the positive direction with $\mathbf{v} =< 3, 2 >$;

    (b) Identify $(2, 3)$ with 0 and the positive direction with $\mathbf{u} =< -3, -2 >$;

    (c) Identify $(-1, 1)$ with 0 and the positive direction with $\mathbf{u}$ as given in part (b).

8. Show that R as defined in Lemma 6.10 is a bijection in the case in which ℓ is a type-1 line.

9. Finish the proof of Lemma 6.11 for the case of a type-1 line.

10. Verify this statement, as it appears in the proof of Lemma 6.11: "Since the base on the circumference of $\mathcal{C}'$ for $\widehat{ST}$ is the major arc associated to ST, φ is obtuse."

11. Show that g' is as given in the proof of Lemma 6.13.

12. Finish the proof of Theorem 6.14 by showing that $f_2 : \ell \to \mathbb{R}_+$ is bijective and continuous. (Hint: It may be easiest to argue that (1) multiplication by a constant is a continuous bijective mapping of $\mathbb{R}$ to $\mathbb{R}$, and (2) if $h_1 : A \to B$ and $h_2 : B \to C$ are bijections, then $h_2 \circ h_1 : A \to C$ is a bijection.)

13. Show that Pasch's Axiom holds in $\mathcal{P}$.

14. Show that any mapping that leaves Euclidean distance invariant also leaves L-distance invariant. Give an example of a mapping that leaves L-distance invariant but does not leave Euclidean distance invariant.

15. Figure 6.11 depicts two different ways congruent L-triangles with a vertex at O could be oriented relative to one another. (See the proof of Theorem 6.20.) What are the other possibilities? For each possibility, identify the reflection or rotation you would use to map one of the L-triangles to ΔOPQ. For instance, reflection across the angle bisector of $\angle Q'OQ$ would

map $\Delta OP'Q'$ to ΔOPQ. What mapping would take ΔORS to $\Delta OP'Q'$? What about other configurations, for example if the L-triangles share a side?

16. Fix a type-1 line and a point P not on it in $\mathcal{P}$. Sketch the limit parallels through P. Repeat the exercise for a type-2 line in $\mathcal{P}$.

17. Describe an L-circle in $\mathcal{P}$. Are L-circles Euclidean circles as well?

;. : 5 :A6 *1 # F 3 J & + ,

Chapter 8

An Introduction to Projective Geometry

8.1 Introduction

Projective geometry has a long history going back to the late classical era with the work of Pappus. Early in the nineteenth century, Poncelet wrote the first treatise on the subject but it did not develop into one of the building blocks of modern mathematics until late in the nineteenth century. Nowadays, projective geometry is part of the everyday working tool set for mathematicians in geometry, topology, and algebra. It is also of interest in its own right.

Like affine geometry, projective geometry is a type of incidence geometry closely related to Euclidean geometry. It takes a mere three axioms to define a projective plane. Despite the seemingly spartan nature of such a configuration, a certain class of projective planes provides more than enough material to keep us occupied here.

The projective planes we study here are those that can be seen as sitting inside higher-dimensional projective spaces. Whether or not a given projective plane can be embedded naturally into projective 3-space turns on an algebraic property of the underlying coordinate system, which may be more general than a field. If the underlying coordinate system is a field, then we can model our projective plane using a vector space. Since finite-dimensional vector spaces always embed naturally into higher-dimensional vector spaces, in this setting, our projective plane embeds into projective 3-space.

This is not the first time we have seen the connection between the nature of the coordinatizing set for a line and features of the geometry itself. With a study of projective geometry, we come to a veritable precipice. Swaying above the turquoise depths of modern algebra roiling several stories below, we resist the impulse to jump, and content ourselves with a study of projective planes that arise from vector spaces over fields. (Jumpers may consult [8].)

© Springer International Publishing AG, part of Springer Nature 2018

M. I. Dillon, *Geometry Through History*,
https://doi.org/10.1007/978-3-319-74135-2_8

We start by defining projective space in terms of affine space, with the addition of so-called ideal points or points at infinity. We then show that we can shed the labels and treat all points in a projective space as one class of objects, all lines as another class of objects, all planes as another, and so on.

8.2 Projective Space

Because we need it for context, we start with a study of projective 3-space, rather than a projective plane. Before adding points to $\mathbb{FA}^3$, we reprise a definition we saw in Chapter 7.

Definition 7.67. The **direction of an affine line** ℓ in $\mathbb{FA}^n$, is any vector $\mathbf{v}$ in the ambient vector space V, where, for some $\mathbf{x} \in V$, $\ell = \{\mathbf{x} + t\mathbf{v} | t \in \mathbb{F}\}$. A **direction in an affine** k**-plane** in $\mathbb{FA}^n$, is a direction for an affine line in that k-plane. For $k \in \{1, \ldots, n-1\}$ fixed, k**-planes in** $\mathbb{FA}^n$ **are parallel** provided they have the same directions and do not intersect. Two k-planes in $\mathbb{FA}^n$ are **skew** if they have different directions and do not intersect.

A pair of distinct parallel lines in $\mathbb{FA}^3$ is contained in a unique plane. (This was part of Theorem 7.68.) There is no affine plane containing a pair of skew lines.

Partition the lines in $\mathbb{FA}^3$ by direction. For instance, lines in $\mathbb{R}^3$ with direction vector $(0, 0, 1)$ are the z-axis and all lines parallel to the z-axis. (See Fig. 8.1.) These comprise one equivalence class of lines in $\mathbb{R}^3$.

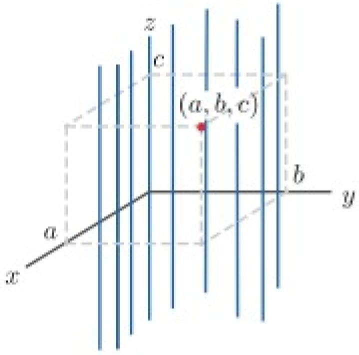

Figure 8.1: Lines parallel to the z-axis in $\mathbb{R}^3$

Add a single point, called an *ideal point* or *point at infinity*, to each line in a given equivalence class of parallel lines in $\mathbb{FA}^3$ and to each plane containing that line. The lines in each equivalence class now intersect in the unique ideal point associated to that class. Note that we add a single ideal point to $\mathbb{FA}^3$ for each different direction in $\mathbb{FA}^3$ and that this point is appended only to the lines and planes with that direction. We distinguish the points of the original affine space by calling them *ordinary points*.

Definition 8.1. The points of **projective** 3**-space over** $\mathbb{F}$, $\mathbb{FP}^3$, are the ordinary and ideal points associated to $\mathbb{FA}^3$. The **ideal plane** or **plane at infinity** is the collection of all ideal points in $\mathbb{FP}^3$. An **ordinary plane** in $\mathbb{FP}^3$ is the collection of all ordinary points associated to a plane in $\mathbb{FA}^3$ together with the ideal points associated to the directions in that affine plane. The collection of ideal points in an ordinary plane is the **ideal line** or **line at infinity** in that plane. A **projective plane** is an ordinary or ideal plane in $\mathbb{FP}^3$. An **ordinary line** is a line with ordinary points and a single ideal point. A **projective line** is an ordinary or ideal line in $\mathbb{FP}^3$. The projective points, lines, and planes in $\mathbb{FP}^3$ are its **projective objects**. Two projective objects in $\mathbb{FP}^3$ are **incident** provided the underlying sets have nonempty intersection.

An ordinary plane contains exactly one point at infinity for each direction in the underlying affine plane. It also contains exactly one line at infinity, the collection of all of its ideal points. The line at infinity for an ordinary plane α is the intersection of α with the plane at infinity.

Our first task is to establish the projective analogs to our main theorems on incidence in $\mathbb{FA}^3$, Theorem 7.68, and Theorem 7.93.

There is no concept of order in an arbitrary projective space so we will not have occasion to discuss segments in this chapter. As such, we use the notation AB for the line through the points A and B.

Theorem 8.2. (1) Two distinct points determine a unique line in $\mathbb{FP}^3$.

(2) If two lines intersect in $\mathbb{FP}^3$, their intersection is a single point.

(3) Two lines in a given plane in $\mathbb{FP}^3$ intersect.

(4) A point and a line not through the point lie in a unique plane in $\mathbb{FP}^3$.

(5) If two points are in a plane in $\mathbb{FP}^3$, the line they determine is in the plane.

(6) A line that does not lie in a plane in $\mathbb{FP}^3$ intersects the plane in one point.

(7) Two intersecting lines in $\mathbb{FP}^3$ lie in a unique plane.

(8) Three noncollinear points in $\mathbb{FP}^3$ lie in a unique plane.

(9) Two planes intersect in a unique line in $\mathbb{FP}^3$.

Proof. Let $V \cong \mathbb{F}^3$ be the ambient vector space for $\mathbb{FA}^3$.

(1) Two ordinary points in $\mathbb{FP}^3$ determine a unique affine line in $\mathbb{FA}^3$ by Theorem 7.68. The points of that affine line comprise the ordinary points on a unique ordinary line in $\mathbb{FP}^3$.

Suppose P is an ordinary point and Q_∞ an ideal point in $\mathbb{FP}^3$. P corresponds to a unique vector $\mathbf{v} \in V$, that is, a point in $\mathbb{FA}^3$, whereas Q_∞ corresponds to a direction in $\mathbb{FA}^3$ determined by a vector $\mathbf{w} \in V$. Then $\mathbf{v} + \operatorname{span}\{\mathbf{w}\}$ is a line in $\mathbb{FA}^3$, comprising the ordinary points of an ordinary line ℓ in $\mathbb{FP}^3$ that includes P and Q_∞. If there is any other line containing P and Q_∞ in $\mathbb{FP}^3$, it must be ordinary, as it contains P, and its ordinary points must coincide with $\mathbf{v} + \operatorname{span}\{\mathbf{w}\}$ because this is the unique line in $\mathbb{FA}^3$ through $\mathbf{v}$ and $\mathbf{w} - \mathbf{v}$, which means it must be ℓ.

Suppose we are given P_∞ and Q_∞, two ideal points in $\mathbb{FP}^3$. There is no ordinary line containing them both as an ordinary line contains exactly one ideal point. P_∞ and Q_∞ correspond to directions in $\mathbb{FA}^3$, say given by vectors $\mathbf{v}$, $\mathbf{w}$ in V, where $\mathbf{w} \neq c\mathbf{v}$ for any $c \in \mathbb{F}$. Then $\{\mathbf{v}, \mathbf{w}\}$ is a basis for W, a two-dimensional subspace of V. The directions in W, and all planes in $\mathbb{FA}^3$ parallel to W, are identical, so every plane in the parallel class determined by W contains P_∞, Q_∞, and the unique ideal line determined by the set of directions common to all the planes in that parallel class. That unique line at infinity is the line determined by P_∞ and Q_∞.

(2) Suppose we have two intersecting lines in $\mathbb{FP}^3$. If both are ordinary and their ordinary points form parallel lines in $\mathbb{FA}^3$, then in $\mathbb{FP}^3$ the lines intersect in a unique ideal point that represents their common direction. If both are ordinary and they intersect in a unique ordinary point, then there can be no other ordinary point in the intersection, by Theorem 7.68. Since their ordinary points form lines with different directions, they cannot have an ideal point in common in $\mathbb{FP}^3$, so in this case, the intersection is a single ordinary point. If one line is ordinary and the other is ideal, then they can share at most one ideal point representing the direction of the ordinary line. If both lines are ideal, then one arises as the directions in one parallel class of affine planes, and the other as the directions in a second parallel class of affine planes. Nonparallel planes in $\mathbb{FA}^3$ intersect in a line and the direction of that line is represented as an ideal point in $\mathbb{FP}^3$, which must be the single point of intersection of our two ideal lines.

(3) Consider lines ℓ, m in a plane α in $\mathbb{FP}^3$. Suppose α is ideal so that ℓ and m are both ideal. In this case, ℓ and m represent parallel classes of affine planes in $\mathbb{FA}^3$. Let β', γ' be affine planes, one from each of these two parallel classes. By Theorem 7.93, $\beta' \cap \gamma' = \ell'$, a single line the direction of which is represented by a unique point in $\ell \cap m$.

If α is ordinary, then at least one of the lines must be ordinary. Say it is ℓ. Let ℓ' be the affine line comprised of the ordinary points of ℓ. If m is also ordinary, let m' be the affine line comprised of the ordinary points of m. Either $\ell' \cap m'$ is a single affine point or ℓ' and m' are parallel. If $\ell' \cap m'$ is a single affine point, then the two affine lines have different directions, thus, ℓ and m have no ideal points in common so $\ell \cap m$ is a single ordinary point. If ℓ' and m' are parallel, there is a single ideal point P in α representing the direction of the two lines. In this case, $\ell \cap m = P$. If m is the line at infinity in α, then it must contain a point P associated to the direction of ℓ'. In this case $\ell \cap m = P$.

(4) Suppose we are given a point and a line not through the point. If the point and line are both ideal, they lie in the unique ideal plane by construction.

If either the line, the point, or both are ordinary, let $\mathcal{S}$ be the union of the point and the points on the line. Since affine lines contain at least two points, and since there are at least three directions in $\mathbb{FA}^3$, $\mathcal{S}$ contains at least three points. Since at least one of our objects is ordinary, $\mathcal{S}$ must contain an ordinary point, P. From the remaining points in $\mathcal{S}$, pick two more, Q, R, so that P, Q, R are noncollinear. By part (1) above, P and Q determine a unique line ℓ in $\mathbb{FP}^3$. P and R determine a second unique line m in $\mathbb{FP}^3$. By assumption, ℓ and m are distinct lines, intersecting at P. Since they contain an ordinary point, P, ℓ and m are ordinary lines so their ordinary points determine a unique affine plane, α'. That affine plane underlies a projective plane with ideal points determined by the directions in α'. Since α' is unique in $\mathbb{FA}^3$, the associated projective plane is unique in $\mathbb{FP}^3$, which proves the result.

(5) Let P and Q belong to a plane α in $\mathbb{FP}^3$. Suppose both P and Q are ideal. If α is also ideal, it contains all ideal points, so contains the unique ideal line determined by P and Q. If α is ordinary, its ordinary points comprise an affine plane, α'. Let $\mathbf{v}$ and $\mathbf{w}$ be vectors in V with the directions represented by P and Q. Then $\alpha' = \mathbf{x} + \operatorname{span}\{\mathbf{v}, \mathbf{w}\}$, for some $\mathbf{x}$ in V. The directions in $\operatorname{span}\{\mathbf{v}, \mathbf{w}\}$, which determine the ideal line determined by P and Q, are the directions in α'. It follows that α contains the ideal line determined by P and Q.

If one point is ordinary and the other ideal, then α is ordinary and the ideal point represents a direction in α', the plane in $\mathbb{FA}^3$ comprised of the ordinary points of α. Say the ordinary point is associated to some $\mathbf{v}$ in V and the ideal point is associated to $\mathbf{w}$ in V. The affine line $\ell = \mathbf{v} + \operatorname{span}\{\mathbf{w}\}$ is in α' so the associated projective line, which includes P and Q, must then be in α.

If both points are ordinary, they determine a unique affine line, ℓ', in $\mathbb{FA}^3$. The ordinary points of α comprise an affine plane that contains ℓ'. Then ℓ' is the set of ordinary points of a unique projective line determined by P and Q, and lying in α.

(6) Let α be a plane in $\mathbb{FP}^3$ and ℓ a line in $\mathbb{FP}^3$ not contained in α. The ideal plane contains all ideal lines so α and ℓ cannot both be ideal. Suppose both are ordinary. Let ℓ' be the affine line comprised of the ordinary points of ℓ. Let α' be the affine plane comprised of the ordinary points of α. We know that $\alpha' \cap \ell'$ is either empty or a single point, by Theorem 7.68. If empty, then ℓ' is parallel to lines in α' so $\alpha \cap \ell$ is the single ideal point associated to that class of parallel lines in $\mathbb{FA}^3$. If $\alpha' \cap \ell'$ contains a single point, then the direction of ℓ' is not in α'. It follows that $\alpha \cap \ell$ cannot contain an ideal point. Thus, in this case, there can be no more than the single ordinary point in $\alpha \cap \ell$.

If ℓ is ordinary and α ideal, then $\alpha \cap \ell$ is the unique ideal point associated to the parallel class of the affine line underlying ℓ. If ℓ is ideal and α ordinary, then ℓ is associated to a parallel class of planes in $\mathbb{FA}^3$ and α', the affine plane comprised of the ordinary points of α, is not in that parallel class: If it were, ℓ would be contained in α. Let β' be an affine plane in the parallel class of affine planes determined by ℓ. Since α' is not parallel to β', $\alpha' \cap \beta' = \ell''$, where ℓ'' is an affine line. Let P be the ideal point associated to the direction of ℓ''. Notice that $P \in \alpha \cap \ell$. Moreover, if β'' is another affine plane in the parallel class determined by ℓ, then $\alpha' \cap \beta''$ is a line parallel to ℓ''. It follows that $P = \alpha \cap \ell$.

(7) Let ℓ and m be intersecting lines in $\mathbb{FP}^3$, with Q the point of intersection. Let P be a point on m that is not on ℓ. By (4), P and ℓ lie in a unique plane, $\alpha \subset \mathbb{FP}^3$. Since Q is on $\ell \subset \alpha$, (5) implies that $PQ = m$ is in α as well. Now any point on m, apart from Q, together with ℓ determine α and by (4), there is no other plane containing the point and ℓ. Since this is true for all points of m, except the one that also belongs to ℓ, it follows that α is the only plane containing both ℓ and m.

(8) This also follows by (1) and (7).

(9) Suppose α and β are ordinary planes in $\mathbb{FP}^3$. Either their ordinary points comprise planes in $\mathbb{FA}^3$ that are parallel or their ordinary points comprise planes in $\mathbb{FA}^3$ that intersect in a unique affine line, ℓ'. In the former case, every direction in α is a direction in β and vice versa so the two planes intersect in their common line at infinity. In the latter case, Theorem 7.93 guarantees that ℓ' comprises all the ordinary points in the intersection $\alpha \cap \beta$. By the same theorem, the direction of ℓ' is the only direction the ordinary points of the two planes have in common so the only ideal point the two can have in common is the one associated to ℓ'. In this case, then, $\alpha \cap \beta = \ell$, the projective line associated to ℓ'.

Now suppose α is ordinary and β is ideal. By definition, the collection of ideal points in α, the intersection of α and β, is the line at infinity for α.

□

Corollary 8.3. Three distinct planes in $\mathbb{FP}^3$ that do not have a common line intersect in a unique point.

Proof. The proof is left as an exercise. □

Next is a basic theorem about counting points in $\mathbb{FP}^3$. This is the projective version of Theorem 7.70.

Theorem 8.4. (1) A line in $\mathbb{FP}^3$ has at least three points.

(2) A plane in $\mathbb{FP}^3$ has four points, no three of which are collinear.

(3) There are four noncoplanar points.

Proof. Let α_∞ be the plane at infinity in $\mathbb{FP}^3$.

(1) An ordinary line contains two ordinary points by Theorem 7.70. In $\mathbb{FP}^3$, it has an additional ideal point so the total number of points on any ordinary line is at least three.

If ℓ_∞ is an ideal line in $\mathbb{FP}^3$, then it lies in some ordinary plane, α. Let $\alpha' \subset \mathbb{FA}^3$ be the underlying affine plane and let $V \cong \mathbb{F}^3$ be the ambient vector space for $\mathbb{FA}^3$. We can use an element from $\mathrm{Aff}(V)$, if necessary, to map α' to an isomorphic copy of $\mathrm{span}\{(1,0,0),(0,1,0)\}$, which we will call the xy-plane. A direction in the xy-plane can be identified with the nonzero scalar multiples of a nonzero vector in the plane. We identify three different directions in the xy-plane, then, corresponding to the vectors $(1,0,0),(0,1,0),(1,1,0)$. Since the correspondence between α' and the xy-plane is via an isomorphism of the affine geometry, there must be lines with three different directions in α'. Those three different directions in α' are represented by three different points on ℓ_∞.

(2) In Theorem 7.70.2. we established that in any affine plane, there are four points, no three of which are collinear. It follows that in any ordinary plane in $\mathbb{FP}^3$, there are four ordinary points, no three of which are collinear.

Next we consider α_∞. Let ℓ be a line in α_∞ and say that A, B, C are three points on ℓ. Note that ℓ represents directions in some parallel class of affine planes. Let α' be an affine plane in that parallel class. By Theorem 7.70, there are four noncoplanar points in $\mathbb{FA}^3$ so we can take an affine point not in α'. Take also any point in α'. The affine line determined by these two affine points has a direction that is not in α'. That direction is realized as a point, $Q \in \alpha_\infty$, where $Q \notin \ell$.

Let $m = AQ$. Notice that $m \subset \alpha_\infty$. Take D to be any third point on m. Now we have $\{A, Q, D\} \subset m$, $\{A, B, C\} \subset \ell$, $\ell \cap m = A$, and $\ell \cup m \subset \alpha_\infty$. Consider the set $\mathcal{S} = \{B, C, Q, D\}$. Any choice of three of these points gives us two points on ℓ and a point on m not on ℓ, or two points on m, and a point on ℓ not on m. This shows that $\mathcal{S}$ is a set of four points in α_∞, no three of which are collinear.

(3) Theorem 7.70 guarantees that $\mathbb{FP}^3$ has four ordinary points which do not lie in an ordinary plane.

□

Theorems 8.2 and 8.4 establish that there is no need to distinguish between ordinary and ideal objects in $\mathbb{FP}^3$. In terms of incidences, all points are alike, all lines are alike, and all planes are alike in $\mathbb{FP}^3$.

Our construction and the attendant theorems also show that any field can be the foundation for a projective space.

Appending ideal points to affine space is one way to construct projective space. Another approach is axiomatic. Here is a set of axioms we could use to define $\mathbb{FP}^3$.

Axioms for Projective Space

Axiom PS1. Two points determine a unique line.

Axiom PS2. Three noncollinear points lie in a unique plane.

Axiom PS3. If two points lie in a plane, the line they determine lies in that plane.

Axiom PS4. Two coplanar lines intersect in a unique point.

Axiom PS5. Two planes intersect in at least two points.

Axiom PS6. A plane and a line that does not lie in the plane intersect in exactly one point.

Axiom PS7. A line contains at least three points.

Axiom PS8. A plane contains four points, no three of which are collinear.

Axiom PS9. There exist four noncoplanar points.

The primary focus of our interest is $\mathbb{FP}^2$. We could have defined a projective plane using the following set of axioms. These apply to all projective planes, not just those constructed from affine space. Among projective planes defined with this set of axioms are those that are still the subject of active research.

Axioms for Projective Plane

Axiom PP1. Two points determine a unique line.

Axiom PP2. Two lines intersect in a unique point.

Axiom PP3. There exist four points, no three of which are collinear.

An advantage to using axioms is that they highlight the *principle of duality*, an intriguing feature of projective geometry.

Definition 8.5. The **dual** of a logical statement about a projective plane is the statement we get by switching the words "point" and "line" and the phrases "lie on" and "intersect in."

Note that in Axiom PP1, "determine" means "lie on." Axiom PP2 could be rephrased, *two lines determine a unique point.* Using the word "determine" in the two different senses, we highlight the duality of the two statements.

The following are dual to the axioms for a projective plane.

Dual Axioms for a Projective Plane

Axiom PP1*. Two distinct lines intersect in a unique point.

Axiom PP2*. Two distinct points determine a unique line.

Axiom PP3*. There are four lines, no three of which are concurrent.

The dual of *collinear,* a word that refers to the relative positions of points in a set, is *concurrent*, which refers to the relative positions of lines in a set. As we progress through our study of projective geometry, we will see that various locutions regarding points and lines have natural duals that we will not necessarily discuss. The dual of a statement captures a certain idea. The exact choice of words to express that idea is variable.

Principle of Duality. Any statement about a projective plane is true if and only if its dual statement is true.

Once we show that Axioms PP1–3 are equivalent to Axioms PP1*–3* in $\mathbb{FP}^2$, we will have established the principle of duality for projective planes constructed from affine planes.

Theorem 8.6. Axioms PP1–3 are equivalent to Axioms PP1*–3* in $\mathbb{FP}^2$.

Proof. It is clear that the pair Axioms PP1, 2 is equivalent to the pair Axioms PP1*, 2*. Suppose then that Axiom PP3 is true. Let $\{A, B, C, D\}$ be four points in $\mathbb{FP}^2$, no three of which are collinear. Consider the set of lines $\mathcal{S} = \{AB, BC, CD, AD\}$. (See Fig. 8.2.) By Axiom PP3, the four lines are distinct: for instance, because A, B, C are not collinear, AB and BC are distinct lines. To see that no three of the lines are concurrent, consider BC, CD, AD. We have $BC \cap CD = C$, but by assumption, C cannot belong to AD, so BC, CD, AD cannot be concurrent. We leave it as an exercise to show that lines in the other three three-line subsets of $\mathcal{S}$ are not concurrent. This shows that Axiom PP3 implies Axiom PP3*.

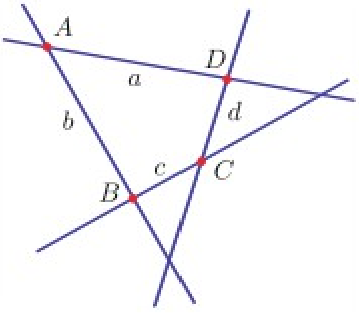

Figure 8.2: Four points, no three of which are collinear and four lines, no three of which are concurrent

Now assume Axiom PP3*. Let $\mathcal{T} = \{a, b, c, d\}$ be four lines in $\mathbb{FP}^2$, no three of which are concurrent. Since any two lines have a common point, we have four points $A = a \cap b$, $B = b \cap c$, $C = c \cap d$, $D = d \cap a$. Since the point of intersection of two lines is unique, and since no three of these lines are concurrent, the four points must be distinct. We claim that no three of the points in the set $\mathcal{U} = \{A, B, C, D\}$ are collinear. Consider $\{B, C, D\}$. If these are collinear, then $D \in c$, but that would make a, c, d concurrent as $D = a \cap d$. The other three cases follow similarly. We leave them as an exercise. □

Duality manifests in higher-dimensional projective spaces as well. Given a statement about $\mathbb{FP}^3$, we obtain its dual by switching the words "point" and "plane" and the phrases "lie on" and "intersect in." Lines are self-dual in projective 3-space. We leave it as an exercise to dualize the axioms we stated above for projective 3-space.

Relatively simple assemblages of points and lines get a good amount of attention in projective geometry and turn out to be surprisingly subtle. The following definition applies in any kind of plane or space.

Definition 8.7. The **pencil of lines through a point** P is the collection of all lines through P.

A line ℓ in $\mathbb{FP}^2$ that does not go through P intersects each line in the pencil through P exactly once. We say ℓ is a *section* of the pencil. The dual of a pencil of lines through a point in a projective plane is the collection of points on a line. It is called the *range* of points on the line.

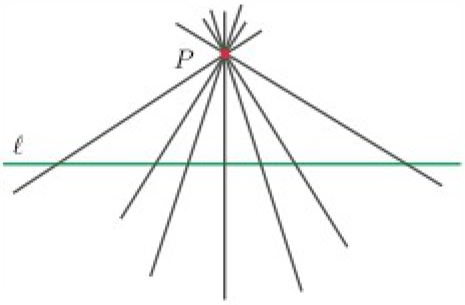

Figure 8.3: The pencil of lines through P is sectioned by ℓ

We noted in Chapter 7 that triangles form an important organizational unit in an affine plane. Triangles play a critical role in projective planes as well.

The definition of triangle does not vary much in different geometries but we repeat it here for the projective setting so that we can address the nature of its dual.

Definition 8.8. A **triangle** is a set of three noncollinear points (vertices) and the lines they determine (sides). A **trilateral** is a set of three nonconcurrent lines and the points they determine.

Triangles are **self-dual**, that is, the dual of a triangle—a trilateral—is the same collection of points and lines as the triangle itself.

Four-sided figures are not self-dual in projective geometry. We start with the four-sided analog of a triangle.

Definition 8.9. A **complete quadrangle** is a configuration in the projective plane composed of a set of four points (vertices), no three of which are collinear, together with the six lines (sides) they determine. **Opposite sides** of a complete quadrangle are two sides without a common vertex. The **diagonal points** of a complete quadrangle are the points where opposite sides intersect.

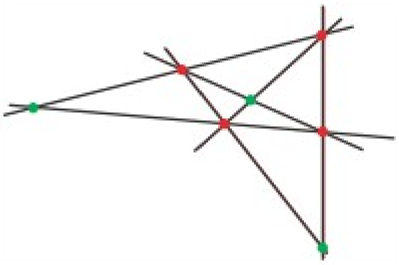

Figure 8.4: A complete quadrangle with diagonal points shown in green

Note that a complete quadrangle is an assemblage of four points and six lines. If we say "quadrangle" in projective geometry we always mean "complete quadrangle," likewise for the dual of a complete quadrangle, a *complete quadrilateral*.

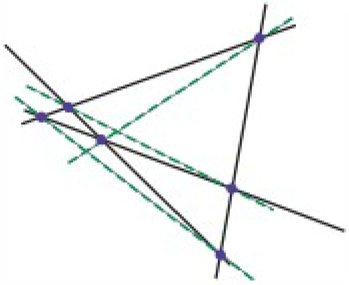

Figure 8.5: A complete quadrilateral with diagonals shown as broken green lines

Definition 8.10. A **complete quadrilateral** is a configuration in the projective plane composed of a set of four lines (sides), no three of which are concurrent, together with the six points (vertices) they determine. **Opposite vertices** of a complete quadrilateral are two vertices without a common side. The **diagonal lines** of a complete quadrilateral are lines that pass through opposite vertices.

Exercises 8.1. 1. In the proof of part (4) of Theorem 8.2, it is suggested that after choosing an ordinary point in $\mathcal{S}$, P, we could choose two more points, Q, and R, so that P, Q, R are noncollinear. Why is this true?

2. Prove Corollary 8.3.

3. In the proof of Theorem 8.4, we refer to three different directions in $\mathbb{FA}^3$ given by vectors $(1,0,0),(0,1,0),(1,1,0)$. Write down parametric and vector versions of the line through $\mathbf{0}$ and each of the three given points in $\mathbb{R}^3$. Do the same, but also write a list of the points on each line in $\mathbb{F}_2^3$, and again in $\mathbb{F}_3^3$.

4. Consider the collection of four lines referred to in the proof of Theorem 8.6, $\mathcal{S} = \{AB, BC, CD, AD\}$. The proof addresses the fact that the lines in the subset $\{BC, CD, AD\}$ are not concurrent.

   (a) Write down every other subset of $\mathcal{S}$ that contains exactly three lines.

   (b) For each subset you found in part (a), argue that the lines are not concurrent.

5. Consider the collection of four points referred to in the proof of Theorem 8.6, $\mathcal{U} = \{A, B, C, D\}$. The proof addresses the fact that the points in the subset $\{B, C, D\}$ are not collinear.

   (a) Write down every other subset of $\mathcal{U}$ that contains exactly three points.

   (b) For each subset you found in part (a), argue that the points are not collinear.

6. Dualize each of Axioms PS1-9. Remember that in projective space, points and planes are dual to one another, and lines are self-dual.

7. Write the dual of each of the following objects and statements in a projective plane.

   (a) Four points, no three of which are collinear

   (b) Two lines, and a point on neither line

   (c) The line determined by a given point and the point of intersection of two given lines

   (d) All lines in the projective plane

   (e) Distinct concurrent lines have only one common point.

   (f) Given a line and a point not on the line, distinct lines through the given point meet the given line in distinct points.

8. Write the dual for each of the following, as statements or objects in projective 3-space.
   (a) the set of points on a line
   (b) the set of points on a plane
   (c) the set of planes containing a given line
   (d) the set of lines passing through a common point
   (e) the set of lines lying in a given plane
   (f) the set of planes in space
   (g) the set of lines in space
   (h) the set of all points in space
9. Is a complete quadrangle self-dual? Explain.
10. How many sides go through a vertex of a complete quadrangle? How many sides go through a vertex of a complete quadrilateral?
11. Consider Fig. 8.6. This is a depiction of a planar geometry in which the points are represented as dots. The lines are the collections of points connected by the sides of the triangle, the cevians as shown, and what appears to be the incircle, as shown. The incidence geometry associated to the picture is called the *Fano plane*. Note that the triangle, segments, and circle are just schematic devices for capturing the incidences in the Fano plane.

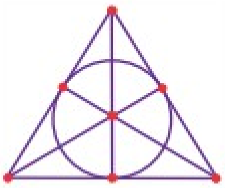

Figure 8.6: The Fano plane

   (a) Verify that the points and lines in the figure satisfy Axioms PP1–3. How many points and how many lines do we have in this plane?
   (b) Identify three different triangles in the Fano plane.
   (c) Identify a complete quadrangle in the Fano plane.
   (d) Identify a complete quadrilateral in the Fano plane.
   (e) Find the diagonal points on the complete quadrangle that you found above. Repeat the exercise with a different quadrangle. What do you notice about the diagonal points?

8.3 The Theorem of Desargues

Pappus's Theorem is often cited as the first theorem in projective geometry. Given that Pappus lived during the fourth century, and that projective geometry was not properly realized until the middle of the nineteenth century, it is clear that Pappus would not have been thinking about projective geometry per se. Pappus's Theorem has a distinctly projective flavor, though. The cleanest way to present it is in the projective plane, where intersection points of lines are guaranteed.

Desargues's Theorem is likewise often cited as the second theorem in projective geometry. Girard Desargues lived during the seventeenth century, so he is a bit closer to the development of modern projective geometry, which arose partly through the observations of Renaissance artists. Desargues was an architect, and belonged to a circle of thinkers that included the Pascals, Etienne (father) and Blaise (son), Fermat, Descartes, and Mersenne. Because his work was unorthodox—he invented his own terminology, for example—it was largely ignored except by the younger Pascal. Pascal's name is given to a result that is closely related to Pappus's Theorem. Desargues's Theorem, viewed in a certain light, is a generalization of Pappus's Theorem.

Pappus's proof of Pappus's Theorem involved heavy use of Euclidean geometry. Of course, in Pappus's time, there was no other geometry. Our approach is to use projective methods to prove Desargues's Theorem first. Deploying Desargues's Theorem after we develop projective transformations, we will find an easy proof of Pappus's Theorem.

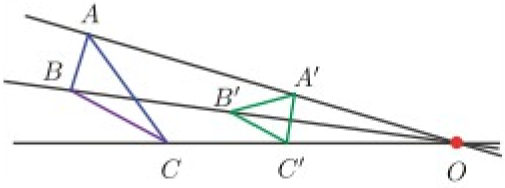

Figure 8.7: Triangles perspective from a point

Definition 8.11. Two triangles ΔABC, $\Delta A'B'C'$ are **perspective from a point** O provided AA', BB', CC' are concurrent at O. In this case, O is a **center of perspectivity**.

When triangles—or any other objects—are perspective from a point, they appear to overlap perfectly if one's eye is at the point of perspectivity. This is an essential idea in projective geometry and in art, especially Renaissance art, in which the effective illusion of space on a flat surface was an important goal.

Definition 8.12. Triangles ΔABC, $\Delta A'B'C'$ are **perspective from a line** ℓ

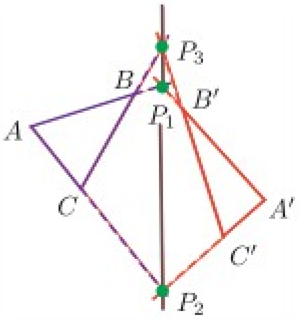

Figure 8.8: Triangles perspective from a line

provided the points $AB \cap A'B' = P_1$, $AC \cap A'C' = P_2$, and $BC \cap B'C' = P_3$ are collinear. In this case, ℓ is an **axis of perspectivity**.

Theorem 8.13 (Desargues). Two triangles are perspective from a point in $\mathbb{FP}^3$ if and only if they are perspective from a line.

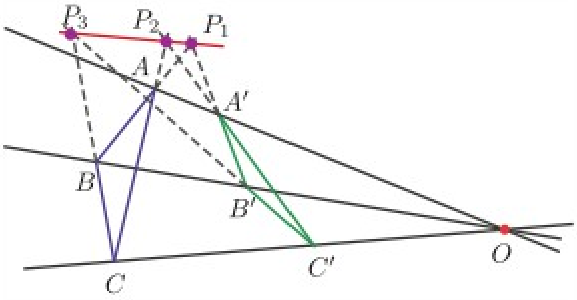

Figure 8.9: The theorem of Desargues

Proof. For the first part of the proof, we assume that the triangles lie in two different planes. Say ΔABC lies in the plane α and $\Delta A'B'C'$ lies in the plane α'.

If ΔABC and $\Delta A'B'C'$ are perspective from a point O, then since the lines AA' and BB' intersect at O, Theorem 8.2 guarantees that they determine a plane. That plane contains A, A', B, B' and O. Now the lines AB and $A'B'$ are in this plane so they intersect at a point P_1. Note that $P_1 \in \alpha \cap \alpha'$. The identical argument produces points $P_2 = AC \cap A'C'$ and $P_3 = BC \cap B'C'$, also in $\alpha \cap \alpha'$. This proves that $\alpha \cap \alpha'$ is the axis of perspectivity for the triangles.

Next suppose ΔABC and $\Delta A'B'C'$ are perspective from a line. As intersecting lines, AB and $A'B'$ lie in a plane β, again by Theorem 8.2. We then have A, B, A', B' all in β so AA' and BB' intersect in β. Similarly, AA' and CC' intersect in a plane β', and BB' and CC' intersect in a third plane β''. We must argue that the planes β, β', β'' intersect in a point, O, that is, that they do not share a line.

If the planes did share a line, that line would be $\beta \cap \beta'$, which contains AB, and at the same time, $\beta \cap \beta''$, which contains BC. As vertices of a triangle, A, B, C are noncollinear, so the line of intersection of β and β' must be different from the line of intersection of β and β''. It follows again by Theorem 8.2 that the intersection of the three planes is a single point, O. Now $AA' = \beta \cap \beta'$, $BB' = \beta \cap \beta''$, and $CC' = \beta' \cap \beta''$. The three lines AA', BB', CC' are thus concurrent at O.

Next, assume that ΔABC and $\Delta A'B'C'$ lie in a single plane, γ. The strategy for the proof in this case is to build another triangle in a different plane and then invoke the first part of the theorem.

Suppose the two triangles are perspective from a point $O \in \gamma$. Let ℓ be a line through O, not lying in γ. Choose P, P' distinct points on ℓ, both different from O. Let PAB be the plane determined by the point P and the line AB. Likewise define PAC, and PBC, each the plane containing P and one side of the triangle ΔABC.

Notice that PAB, PAC, and PBC must be distinct planes: If any two were equal, then P, A, B, and C would be coplanar, contradicting our choice of P. Moreover, $PAB \cap PAC = PA$, $PAB \cap PBC = PB$ and $PAC \cap PBC = PC$. Since the lines PA and $P'A'$ lie in the plane determined by intersecting lines AA', and PP', they must intersect in a point A''. We define points B'', C'' in a similar manner.

Next we argue that A'', B'', C'' cannot be collinear: If they were, then since $A''B''$ is in PAB, and $A''C''$ is in PAC, and $B''C''$ is in PBC, the three planes would have to intersect in a line containing P and lying in γ. This contradicts P lying outside γ. Thus, we have a new triangle, $\Delta A''B''C''$.

By construction, ΔABC and $\Delta A''B''C''$ are perspective from P, while $\Delta A'B'C'$ and $\Delta A''B''C''$ are perspective from P'. The line of perspectivity in both cases is the intersection of γ with the plane of $\Delta A''B''C$". This gives us that $A''B'' \cap AB$ and $A''B'' \cap A'B'$ fall on this line of intersection. Since $A''B''$ meets this line just once, AB and $A'B'$ intersect there as desired. The other sides of the two triangles meet on this line as well, for the same reasons. We conclude that ΔABC and $\Delta A'B'C'$ are perspective from the line $\gamma \cap \xi$, where ξ is the plane of $\Delta A''B''C''$.

We leave the last part of the proof to the reader. □

Notice that the proof of Desargues's Theorem is much easier when the triangles occupy different planes. Though it holds in Euclidean space, as long as all the intersection points exist, Desargues's Theorem is fundamentally about projective space. This business of embedding the plane in 3-space is not just a trick in the proof. The theorem is actually false in projective planes that cannot be viewed as occupying a part of projective 3-space. Projective planes for which Desargues's Theorem holds are *Desarguesian* and those for which it does not hold are *non-Desarguesian*. Non-Desarguesian planes are not defined over fields. They are not based on vector spaces or, equivalently, affine spaces.

Figure 8.10 shows the arrangement in Fig. 8.9 without specifying a particular role for any point or line. It turns out that any point in the configuration can play the role of O, the point of perspectivity. Likewise, any line can play the role of the axis of perspectivity. In the exercises, you will have a chance to explore these ideas a little bit.

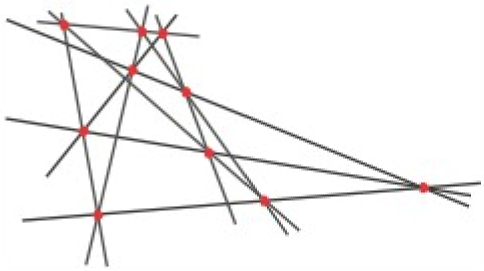

Figure 8.10: Any point can be the point of perspectivity in a Desarguesian configuration

Exercises 8.2. 1. In the proof of the second half of Desargues's Theorem, we have triangles in a single plane γ, perspective from a point O in γ. Then we let ℓ be a line not in γ, but through the point O. Why must such a line ℓ exist?

2. Make a sketch to illustrate the part of the proof of Desargues's Theorem that applies to triangles that lie in one plane.

3. Finish the proof of Desargues's Theorem.

4. Below we will learn that the Fano plane is Desarguesian. Identify two triangles perspective from a point in the Fano plane. Find the axis of perspectivity.

5. An assemblage of ten points and ten lines so that each line is incident to exactly three of the points and each point is incident to exactly three of the lines is a *Desarguesian configuration*. Verify that Fig. 8.9 depicts a Desarguesian configuration. Relabel the points in that figure so that the point of perspectivity is a point other than O. (See Fig. 8.10.) Similarly, any line can play the role of the axis of perspectivity. Relabel the points several different ways to illustrate this principle.

6. Referring to the previous exercise, does a given point of perspectivity uniquely determine the two triangles in perspective, if the choice is made in a Desarguesian configuration?

8.4 Harmonic Sequences

Desargues's Theorem is about triangles. In this section, we consider quadrangles, quadrilaterals, and associated arrangements of points and lines.

Let A, B, C, D be the vertices of a complete quadrangle, as prescribed by Definition 8.9. We designate this quadrangle simply as $ABCD$. The diagonal points of $ABCD$ are $AB \cap CD$, $AC \cap BD$, and $AD \cap BC$. If our projective plane is defined over a field of characteristic different from two, these diagonal points are noncollinear. (This is not obvious and we omit the proof. See [9], for example, where it is proved using an underlying vector space.) The triangle they determine is called the *diagonal triangle* of the quadrangle. A complete quadrilateral also has an associated diagonal triangle, which you can define using duality. (See the exercises.)

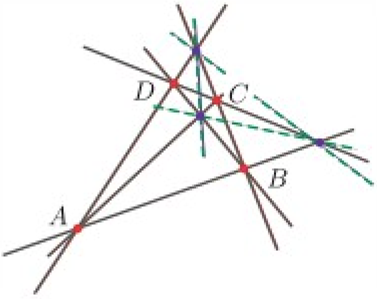

Figure 8.11: The quadrangle $ABCD$ with its diagonal points in blue and the sides of its diagonal triangle in green

Definition 8.14. A **harmonic sequence of points**, $H(P, Q; R, S)$, in a projective plane, is a set of four collinear points P, Q, R, S where P, Q are diagonal points of a quadrangle and R, S are on the sides of the quadrangle that pass through the third diagonal point. In this case, we say that S is the **harmonic conjugate** of R with respect to P and Q.

Dual to a harmonic sequence of points is a harmonic sequence of lines. When we dualize, we must remember that the dual of a quadrangle is a quadrilateral.

Definition 8.15. A **harmonic sequence of lines**, $H(p, q; r, s)$, in a projective plane, is a set of four concurrent lines p, q, r, s, where p, q are diagonal lines of a complete quadrilateral and r, s pass through the vertices of the third diagonal

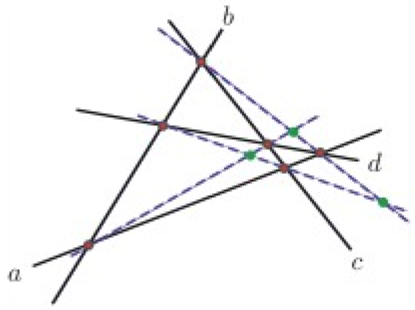

Figure 8.12: The quadrilateral *abcd* with its diagonal sides in blue, and vertices of the diagonal triangle in green

side of the quadrilateral. In this case, we say that s is the **harmonic conjugate** of r with respect to p and q.

The ordering of the points or lines in a harmonic sequence matters. While $H(P,Q;R,S)$ means $H(Q,P;S,R)$, it does not mean $H(P,R;Q,S)$, for instance.

The line determined by two of the diagonal points of a quadrangle must intersect the two sides of the third diagonal point in our projective planes. Consequently there is a harmonic sequence associated to any quadrangle. Indeed, a harmonic sequence of points is a *section* of the quadrangle: Every side of the figure intersects the line in one point.

Constructing a Harmonic Conjugate

We start our study by constructing a harmonic conjugate of a point. Our approach here closely follows [9].

Suppose P, Q, R are distinct points on a line ℓ in a projective plane. Choose C not on ℓ and $B \in QC$, B different from Q and C, as in Fig. 8.13. Form RB

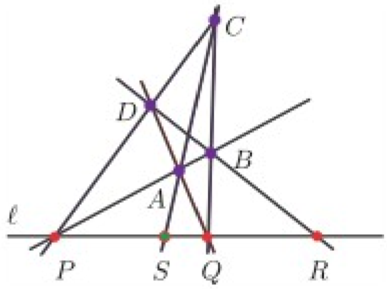

Figure 8.13: Construct the harmonic conjugate of R with respect to P and Q

and PC. Let D be their point of intersection. Next form PB and QD. Let A be their point of intersection. Form AC and let $S = AC \cap \ell$. We leave it as an exercise to verify that $ABCD$ forms a quadrangle and that S is a harmonic conjugate to R.

Our first theorem on the topic establishes that the harmonic conjugate of a point is unique.

Theorem 8.16. Given distinct points P, Q, R on a line ℓ in a projective plane, there is a unique point S on ℓ so that $H(P, Q; R, S)$.

Proof. Let P, Q, R be on ℓ in a projective plane. We have established that there is S so that $H(P, Q; R, S)$. Referring to the construction above, suppose $A'B'C'D'$ is a second quadrangle that produces a point S' so that $H(P, Q; R, S')$. We have $AC \cap \ell = S$ and $A'C' \cap \ell = S'$. Notice that $S' = S$ if and only if $AC \cap A'C'$ is on ℓ. (See Fig. 8.14.) We prove that $AC \cap A'C' \in \ell$ using Desargues's Theorem three times. First notice that ΔBCD and $\Delta B'C'D'$ are perspective from ℓ, thus, BB', CC', and DD' are concurrent. (In Fig. 8.14 they appear to be parallel, so we can think of them as projectively concurrent.) Similarly, ΔABD and $\Delta A'B'D'$ are perspective from ℓ, so AA', BB', DD' are concurrent. From there we have AA', BB', CC' concurrent so that ΔABC and $\Delta A'B'C'$ are perspective from a point, thus, $AB \cap A'B' = P$, $BC \cap B'C' = Q$, and $AC \cap A'C' = S = S'$ are collinear, as desired. □

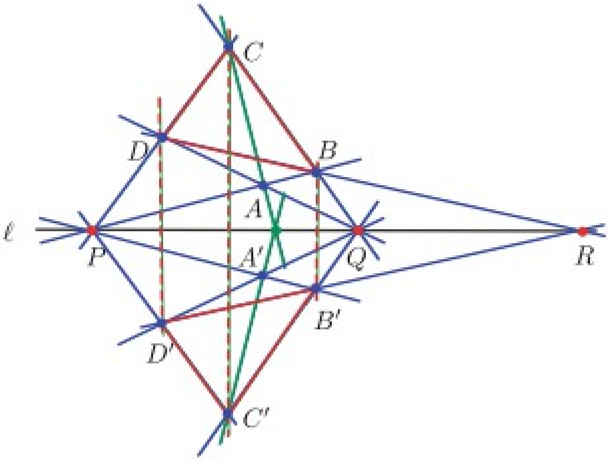

Figure 8.14: Since ΔBCD and $\Delta B'C'D'$ are perspective from ℓ, the lines BB', CC', DD' must be concurrent

The principle of duality in a projective plane allows us to make the following statement without proof.

Theorem 8.17. If p, q, r are concurrent lines in a projective plane, then among the lines in the pencil at the point of concurrency, there is a unique line s so that $H(p, q; r, s)$.

Theorem 8.18. If a line ℓ meets concurrent lines p, q, r, s in the points P, Q, R, S in a projective plane, then $H(P, Q; R, S)$ if and only if $H(p, q; r, s)$.

Proof. Suppose p, q, r, s are lines concurrent at C. Let ℓ meet lines p, q, r, s in points P, Q, R, S, respectively so that $H(P, Q; R, S)$. Using C, we can form S from P, Q, R. In the process, we construct a quadrangle $ABCD$ so that pairs of opposite sides—AB, CD; and AD, BC—intersect in P and Q, and diagonals—AC and BD—intersect PQ at R and S, respectively. We have $P = AB \cap CD$, $Q = AD \cap BC$, $R = BD \cap PQ$, and $S = AC \cap PQ$.

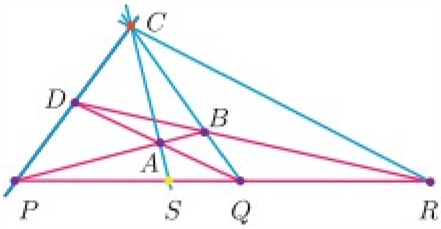

Figure 8.15: C is the point of concurrency of diagonal lines associated to the quadrilateral with pink sides

If we are to view p, q, r, s themselves so that $H(p, q; r, s)$, we need to see p, q as diagonals in a complete quadrilateral, and r, s as passing through the distinct vertices on the third diagonal side of the quadrilateral.

Consider the quadrilateral $\mathcal{Q}$ with sides $PB = PA$, $QD = QA$, $PQ = PR$, $BD = BR$, and vertices B, D, P, Q, A, R. Here $\{P, D\}$, $\{B, Q\}$, and $\{A, R\}$ are opposite vertex pairs. It follows that $PD = PC = p$ and $BQ = QC = q$ are diagonals for $\mathcal{Q}$ and that $r = RC$ and $s = AC$ satisfy $H(p, q; r, s)$.

The converse follows by duality. ☐

Exercises 8.3. 1. How many sides go through each vertex of a complete quadrangle in a projective plane? How many sides go through each diagonal point?

2. Use duality to define the diagonal triangle of a complete quadrilateral in a projective plane.

3. In Fig. 8.16 you can see Fig. 8.11 purely as a configuration. Likewise, Fig. 8.17 shows Fig. 8.12 purely as a configuration. Describe these configurations by specifying the number of points, the number of lines, and the incidences enjoyed by each point and each line. Are the two the same type of configuration? Are they dual? (See Exercise 8.2.5.)

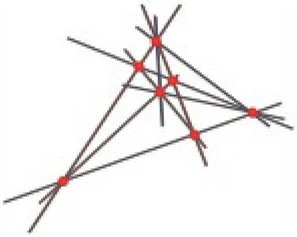

Figure 8.16: The configuration associated to a complete quadrangle and its diagonal triangle

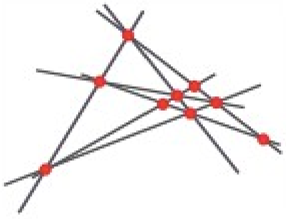

Figure 8.17: The configuration associated to a complete quadrilateral and its diagonal triangle

4. Verify that the line determined by the diagonal points of a quadrangle in a projective plane is a section of the quadrangle.

5. Verify that the construction we described as in Fig. 8.13 does in fact yield a complete quadrangle, $ABCD$, for which $H(P,Q;R,S)$.

8.5 Transformations and Pappus's Theorem

There are several different ways to approach projective geometry. We could have approached it strictly from the transformational point of view in which we identify the objects left invariant under certain transformations, declare those to be the objects of interest, and study the features of those objects that the transformations respect. This is Erlangen geometry. Our point of view is less algebraic, but from any point of view, the notions of perspectivity and projectivity are fundamental. As promised, these will lead us to Pappus's Theorem.

Our context here is a projective plane, $\mathbb{FP}^2$.

Definition 8.19. Let ℓ and ℓ' be lines in $\mathbb{FP}^2$. A **perspectivity** is a mapping from ℓ to ℓ' so that every line joining a point on ℓ to its image on ℓ' passes through a fixed point P, called the **center** of the perspectivity.

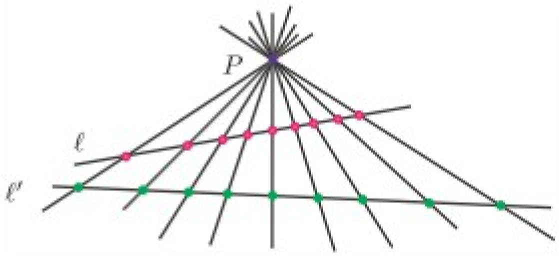

Figure 8.18: A perspectivity with center P

A perspectivity is its own inverse: In Fig. 8.18, we can view the mapping as going from ℓ to ℓ' or from ℓ' to ℓ. As long as ℓ and ℓ' are different lines, the center of a perspectivity is unique. We leave the verification of that statement as an easy exercise.

Definition 8.20. A **projectivity** or **projective transformation** is the composition of a finite number of perspectivities.

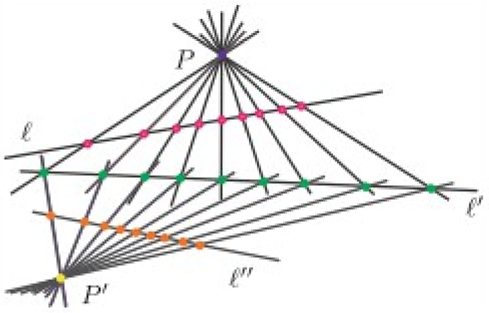

Figure 8.19: A projectivity from ℓ to ℓ'' (or from ℓ'' to ℓ) composed of two perspectivities, one through P, the second through P'

Since a perspectivity is its own inverse, a projectivity is also invertible: If $T = T_1 \cdots T_n$ is a product of perspectivities, then $T^{-1} = T_n \cdots T_1$.

If $A, B, C, \ldots$ are collinear points that are mapped to $A', B', C', \ldots$ respectively under a projectivity T, we write $T(A, B, C, \ldots) = (A', B', C', \ldots)$.

Theorem 8.21. Projectivities map harmonic sequences of points to harmonic sequences of points in a projective plane.

Proof. It is enough to argue that a single perspectivity sends a harmonic sequence of points in a projective plane to another harmonic sequence of points.

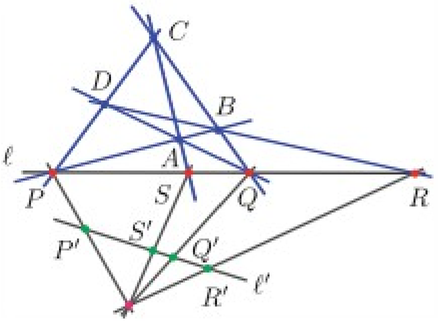

Figure 8.20: A perspectivity maps a harmonic sequence of points to a harmonic sequence of points

Suppose T is a perspectivity and $H(P,Q;R,S)$ where $\ell = PQ$. Referring to Fig. 8.20, we see this is the scenario addressed by Theorem 8.18: The line ℓ meets the concurrent lines PP', QQ', RR', SS' in the points P,Q,R,S with $H(P,Q;R,S)$ so $H(PP',QQ';RR',SS')$. By the same theorem, we also have $H(P',Q';R',S')$. □

As noted in [9], the content of the theorem suggests that there is no guarantee of the existence of a perspectivity that will map four arbitrary collinear points to another specified set of four collinear points. We can certainly map a given pair of points to any pair on another line via a perspectivity so pairs of points can be made to correspond using projective transformations, but not arbitrary quadruples. The next theorem answers the question about triplets.

Theorem 8.22 (Fundamental Theorem of Projective Geometry). Any set of three collinear points in a projective plane can be mapped to any other set of three collinear points in that plane by means of a unique projective transformation.

Sketch of proof. As in [9], we sketch the proof. First we argue for existence, then we address uniqueness. Let A,B,C and A',B',C' be arbitrary triples of collinear points in $\mathbb{FP}^2$.

Define $B'' = AB' \cap A'B$, $C'' = AC' \cap A'C$, and $A'' = AA' \cap B''C''$. If T_1 is the perspectivity through A', then $T_1(A,B,C) = (A'',B'',C'')$. Taking T_2 to be the perspectivity through A, we then have $T_2(A'',B'',C'') = (A',B',C')$. (See Fig. 8.21.) This gives us $T(A,B,C) = (A',B',C')$, where $T = T_2 \circ T_1$. The difficult part of the proof is arguing that T is unique, in other words, that knowing $T(A,B,C)$ is enough to know $X' = T(X)$ for any $X \in AB$. The approach is to argue that we can describe a given point on ℓ, or get arbitrarily close to a given point on ℓ, by means of harmonic sequences.

Theorem 8.16 implies there is unique D^* on ℓ so that $H(A,B;C,D^*)$. By Theorem 8.21, $H(A',B';C',D^{*\prime})$ where $T(D^*) = D^{*\prime}$. If $X = D^*$, we are done.

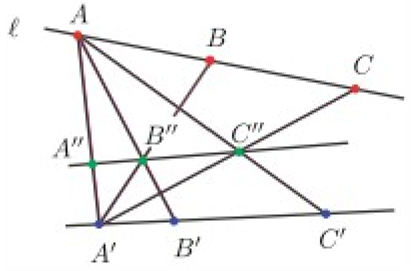

Figure 8.21: There is a projectivity relating any two sets of three collinear points

If not, repeat, noting there is unique E^* so that $H(B, C; D^*, E^*)$, and if $X \neq E^*$, there is unique F^* so that $H(C, D^*; E^*, F^*)$, etc. Depending on the nature of the underlying field—for example, whether it is $\mathbb{R}$ or $\mathbb{C}$ or finite—we either use up all the points on ℓ this way, eventually accounting for X as a member of a harmonic sequence, or we get arbitrarily close to X using harmonic sequences. Since each harmonic sequence of points is mapped to a well-defined harmonic sequence of points under T, this means $T(X)$ is determined by $T(A, B, C)$ as desired. $\square$

Corollary 8.23. Let ℓ and ℓ' be distinct lines in $\mathbb{FP}^2$. A projectivity from ℓ to ℓ' that leaves a point on ℓ fixed is a perspectivity.

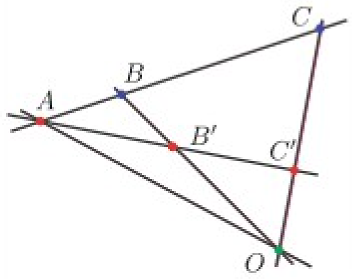

Figure 8.22: A nonidentity projectivity leaving one point fixed is a perspectivity

Proof. Let T be a projectivity with $T(A, B, C) = (A, B', C')$ for A, B, C collinear in some projective plane. Say $\ell = AB$ and $\ell' = AB'$ where $\ell \neq \ell'$. Let $O = BB' \cap CC'$. Then the perspectivity with center O is identical to T at A, B, C so by the theorem, that perspectivity must be T. $\square$

Ironically, we conclude this section with the first theorem of projective geometry, Pappus's Theorem.

Theorem 8.24 (Pappus's Theorem). Suppose A, B, C, A', B', C' are points in a projective plane so that A, B, C are collinear, and A', B', C' are collinear. Then the points $AB' \cap A'B$, $AC' \cap A'C$, $BC' \cap B'C$ are collinear.

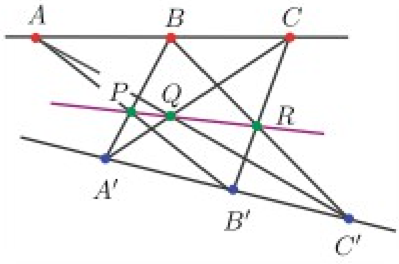

Figure 8.23: Pappus's Theorem says the green points are collinear

Proof. Let $P = AB' \cap A'B$, $Q = AC' \cap A'C$, $R = BC' \cap B'C$. We identify some additional points for the proof. Let $D = PQ \cap AC$, $E = AB' \cap A'C$, and $F = AC' \cap A'B$. Let T be the perspectivity from AB' to AC' with center A'. Then $T(A, P, E, B') = (A, F, Q, C')$. Let S be the perspectivity from AC' to PQ with center B, so that $S(A, F, Q, C') = (D, P, Q, R')$, where $R' = PQ \cap BC'$. We want to show that $R' = R$.

Composing the two perspectivities, we get the projectivity

$$(S \circ T)(A, P, E, B') = (D, P, Q, R').$$

By Corollary 8.23, $S \circ T$ is a perspectivity. Since AD and EQ meet at C, C is the center of the perspectivity and $R' = PQ \cap B'C$, implying $R' = R$, as desired. □

Exercises 8.4. 1. Show that the center of a perspectivity in a projective plane is unique.

2. Describe how to map a given pair of points in a projective plane to any pair of points on another line via a perspectivity.

3. If $T = T_1 T_2 \cdots T_n$ is a projectivity of a projective plane, with T_i a perspectivity, show that $T^{-1} = T_n \cdots T_2 T_1$.

4. Describe how to map a given pair of points in a projective plane to any other specified pair on the same line using a projectivity.

5. What is the dual of Pappus's Theorem?

6. Can you rearrange the labels A, B, C on one line and/or A', B', C' on the other line in Fig. 8.23 so that the line PQ changes? Keep the designation $P = AB' \cap A'B$, etc.

8.6 Homogeneous Coordinates

We return now to the idea that the projective planes we are studying here are those that arise from vector spaces. This time we take a slightly different point of view from the one we assumed when we first defined projective space as affine space with ideal points appended.

Let V be a three-dimensional vector space over $\mathbb{F}$ so $V \cong \mathbb{F}^3$. We use V itself to realize a projective plane over $\mathbb{F}$. In the process, we get an introduction to homogeneous coordinates. Again, remember that this cannot be done for every projective plane, just the ones based on fields. Note further that to realize a projective plane—a two-dimensional setting—we use a three-dimensional vector space.

A one-dimensional subspace of V is the span of any nonzero vector in V. In $\mathbb{R}^3$, a one-dimensional subspace looks like a line through the origin. In $\mathbb{F}_2^3$, it looks like a set of discrete points but we still refer to it as a line through the origin. **Projective points** in this model are lines through the origin in V. **Projective lines** in this model are two-dimensional subspaces of V, that is, the planes through the origin. **Incidence** is nonempty intersection. Before anything else, we establish that this model satisfies the axioms for a projective plane.

Theorem 8.25. The projective points and lines associated to $V \cong \mathbb{R}^3$ satisfy the axioms for a projective plane.

Proof. Axiom PP1 says two projective points determine a unique projective line. In our model, let $P = \operatorname{span}\{\mathbf{v}\}$ and $Q = \operatorname{span}\{\mathbf{w}\}$ be our two projective points. Here $\{\mathbf{v}, \mathbf{w}\}$ is a linearly independent set in V. P and Q determine the projective line $\ell = \operatorname{span}\{\mathbf{v}, \mathbf{w}\}$. Since the plane in V that contains $\mathbf{v}$ and $\mathbf{w}$ is unique, so is ℓ. This proves Axiom PP1.

Next, consider two projective lines $\ell = \operatorname{span}\{\mathbf{v}, \mathbf{w}\}$, and $m = \operatorname{span}\{\mathbf{u}, \mathbf{x}\}$, each modeled on a two-dimensional subspace of V. Since $\dim V = 3$, $\{\mathbf{v}, \mathbf{w}, \mathbf{u}, \mathbf{x}\}$ is linearly dependent. We assume ℓ and m are distinct, implying that

$$\dim \operatorname{span}\{\mathbf{v}, \mathbf{w}, \mathbf{u}, \mathbf{x}\} = 3.$$

There is thus a one-dimensional subspace common to $\operatorname{span}\{\mathbf{v}, \mathbf{w}\}$ and $\operatorname{span}\{\mathbf{u}, \mathbf{x}\}$. This one-dimensional subspace of V, a line through the origin, models a unique projective point, P, the point of intersection of ℓ and m. This shows that Axiom PP2 holds.

Next, let $P_1 = \operatorname{span}\{(1,0,0)\}$, $P_2 = \operatorname{span}\{(0,1,0)\}$, $P_3 = \operatorname{span}\{(0,0,1)\}$, and $P_4 = \operatorname{span}\{(1,1,1)\}$. Take any pair of these, for example P_1, P_2. These determine a unique projective line $\ell = \operatorname{span}\{(1,0,0),(0,1,0)\}$. Notice that neither P_3 nor P_4 lies on $\ell = \{(a, b, 0) | \, a, b \in \mathbb{R}\}$. The other two cases are easy to check, allowing us to conclude that our model contains four points, no three of which of are collinear. This verifies Axiom PP3. $\square$

Next we coordinatize our model.

Define $\sim$ on nonzero $\mathbf{v}, \mathbf{w} \in \mathbb{F}^3$ by $\mathbf{v} \sim \mathbf{w}$ provided $\mathbf{v} = c\mathbf{w}$ for some $c \in \mathbb{F}$. We leave it as an exercise to check that $\sim$ is an equivalence relation on $\mathbb{F}^3 \setminus \mathbf{0}$.

Designate the equivalence class containing $\mathbf{v} = (a, b, c) \neq \mathbf{0}$ by $[\mathbf{v}] = [a, b, c]$. Each equivalence class corresponds to a unique projective point in our model. Each point in the model now has coordinates $[a, b, c]$, where not all a, b, c are zero.

Definition 8.26. The **homogeneous coordinates** for a point $P \in \mathbb{FP}^2$ determined by the span of a nonzero vector $\mathbf{v} = (a, b, c) \in \mathbb{F}^3$ is the equivalence class $[a, b, c] = \{t(a, b, c) \,|\, t \neq 0\}$.

If P is a projective point with $P = [a, b, c]$, at least one of the numbers a, b, c must be nonzero. There is thus always a representative of P that has one coordinate equal to 1.

It is instructive to reconcile this model of $\mathbb{FP}^2$ with the first model we had, where we appended ideal points to $\mathbb{FA}^2$. Because our experience is mostly with $\mathbb{R}^3$, we can use that to flesh out the connection between the two. This is purely for the convenience of familiarity. Nothing we say here depends on properties of $\mathbb{R}$ as the field of scalars.

Consider the pencil of lines through $\mathbf{0}$ in $\mathbb{R}^3$. This pencil corresponds to the set of projective points in our model. An arbitrary line in the pencil has the form $\ell = \operatorname{span}\{(a, b, c)\}$, where $(a, b, c) \neq (0, 0, 0)$. If $c \neq 0$, $1/c$ is defined in $\mathbb{R}$ and we have the point $(a/c, b/c, 1)$ on ℓ. In other words, ℓ intersects the plane $z = 1$ at $(a/c, b/c, 1)$. In the usual coordinate system for $\mathbb{R}^3$, the plane $z = 1$ is one unit above, and parallel to, the xy-plane.

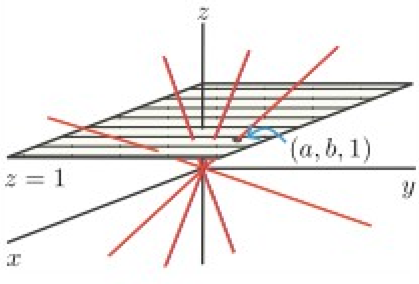

Figure 8.24: In the lines-through-the-origin model of $\mathbb{RP}^2$, the lines with homogeneous coordinates $[a, b, 1]$ correspond to affine points, and those with homogeneous coordinates $[a, b, 0]$ correspond to points at infinity

Any line ℓ in the pencil that does not intersect the plane $z = 1$ must be the span of a vector with the form $(a, b, 0)$. In other words, ℓ lies in the xy-plane. We thus have two kinds of lines in the pencil: those that intersect the plane $z = 1$, and those that lie in the xy-plane.

The projective points associated to the lines that intersect the plane $z = 1$ have homogeneous coordinates $[a, b, 1]$. The projective points associated to the lines in the xy-plane have homogeneous coordinates $[a, b, 0]$. Since $z = 1$ is a copy of $\mathbb{RA}^2$, we can identify the projective points having homogeneous coordinates $[a, b, 1]$ with the points of $\mathbb{RA}^2$, that is, the ordinary points of $\mathbb{RP}^2$. We then identify the projective points having homogeneous coordinates $[a, b, 0]$ with the ideal points appended to $\mathbb{RA}^2$.

There is yet another related model for $\mathbb{RP}^2$. Think about the pencil of lines through $\mathbf{0}$ in $\mathbb{R}^3$, this time in its relationship to the unit sphere in $\mathbb{R}^3$, also called the **unit 2-sphere**:

$$S^2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}.$$

Each line intersects the surface of S^2 in exactly two Euclidean points. Those two points are called **antipodal points** on S^2. A pair of antipodal points on S^2 determines the single projective point that the line represents. In this S^2-based model of $\mathbb{RP}^2$, projective lines are based on the great circles on the sphere—that is, circles centered at $\mathbf{0}$. A projective line, then, is the set of antipodal point pairs on a great circle on S^2. Since any plane through $\mathbf{0}$ in $\mathbb{R}^3$ intersects S^2 in a great circle, and any great circle on S^2 determines a plane through $\mathbf{0}$, we have a neat correspondence between the elements in one model and the elements in the other model.

Antipodal point pairs on S^2 give us a good model for thinking about the physical nature of $\mathbb{RP}^2$, that is, what it would be like to live in $\mathbb{RP}^2$. Consider

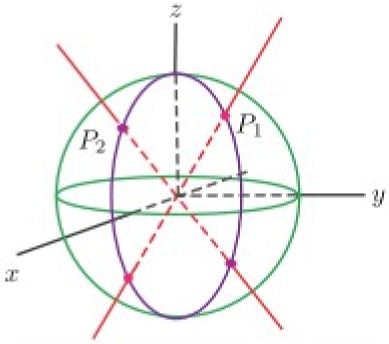

Figure 8.25: In the S^2-based model for $\mathbb{RP}^2$, a projective point is realized as an antipodal point pair on the sphere. A projective line is realized as the antipodal point pairs on a great circle.

a point P on S^2 and say P' is its antipodal point. If we move along any great circle through P, we must eventually arrive at P'. In $\mathbb{RP}^2$, then, if move along a line through a point, we meet back up again with that point. This suggests that if light travels in a line, then in a projective plane, observers who could see far enough would actually be looking at the backs of their own heads.

Finally, we employ homogeneous coordinates to do a calculation in $\mathbb{RP}^2$.

Example 8.27. Let $P = [1, 0, 0]$, $Q = [0, 1, 0]$, and $R = [1, 1, 0]$ in $\mathbb{RP}^2$. Notice that P, Q, R are collinear: This follows because the two-dimensional vector subspace of $\mathbb{R}^2$, $\alpha = \text{span}\{(1, 0, 0), (0, 1, 0)\}$ contains the one-dimensional subspace of $\mathbb{R}^2$, $\ell = \text{span}\{(1, 1, 0)\}$. We would like to find the harmonic conjugate of R with respect to P and Q.

Referring to the construction we carried out in Section 8.4, we choose a projective line through Q, different from PQ, and take two points on that line, different from Q. Let the projective line be modeled by the Euclidean plane

$$\beta = \text{span}\{(0, 1, 0), (0, 0, 1)\} = \{(0, a, b) |\ a, b, \in \mathbb{R}\}.$$

The projective points $B = [0, 1, 1]$ and $C = [0, 2, 1]$ are on that line and different from Q.

Let $D = PC \cap RB$. The projective line PC is determined by the Euclidean plane $\{(a, 2b, b) |\ a, b \in \mathbb{R}\}$. RB is determined by $\{(a, a + b, b) |\ a, b \in \mathbb{R}\}$. It follows that $D = [1, 2, 1]$. The projective line PB is represented by the Euclidean plane $\{(a, b, b) |\ a, b \in \mathbb{R}\}$, while QD is represented by $\{(a, b, a) |\ a, b \in \mathbb{R}\}$. The projective point $A = PB \cap QD$ is then $[1, 1, 1]$. We have AC represented by $\{(a, a + 2b, a + b) |\ a, b \in \mathbb{R}\}$ and PQ represented by $\{(a, b, 0) |\ a, b \in \mathbb{R}\}$. It follows that $S = AC \cap PQ$, the harmonic conjugate of R with respect to P and Q, is $[1, -1, 0]$.

Exercises 8.5.

1. Show that $\sim$ defined on $\mathbb{F}^3 \setminus \mathbf{0}$ by $\mathbf{v} \sim \mathbf{w}$ provided $\mathbf{v} = c\mathbf{w}$ for some $c \in \mathbb{F}$ is an equivalence relation.

2. If $V = \mathbb{R}^3$, the lines $\ell_1 = \text{span}\{(1, 0, 0)\}$ and $\ell_2 = \text{span}\{(0, 1, 0)\}$ determine what plane? What is the plane determined by lines ℓ_1 and $\ell_3 = \text{span}\{(0, 0, 1)\}$? What about the plane determined by lines ℓ_2 and ℓ_3? What is the plane determined by ℓ_1 and $\ell_4 = \text{span}\{(1, 1, 1)\}$? Describe it. How is it oriented in V?

3. Let $\alpha = \text{span}\{\mathbf{v}, \mathbf{w}\}$, and $\beta = \text{span}\{\mathbf{u}, \mathbf{x}\}$ be distinct planes in V where $\dim V = 3$. Detail the argument that $\{\mathbf{v}, \mathbf{w}, \mathbf{u}, \mathbf{x}\}$ is linearly dependent, thus that there is one-dimensional subspace in common to α and β. Use a dependence relation to describe the line of intersection of α and β.

4. Finish the proof of Theorem 8.25.

5. In Example 8.27, we note that if $C = [0, 2, 1]$ and $P = [1, 0, 0]$, then the projective line PC is determined by the Euclidean plane $\{(a, 2b, b) |\ a, b \in \mathbb{R}\}$. Verify that this is correct. Do the same for the rest of the Euclidean

planes in that example and verify that the intersection points, given in homogeneous coordinates, are correct. Note in particular the Euclidean plane underlying the projective line QD. Why can we say that plane contains any Euclidean point of the form (a, b, a)?

6. Rework Example 8.27 using different points B and C.
7. Rework Example 8.27 over $\mathbb{F}_2$ and again over $\mathbb{F}_3$.
8. Show that the Fano plane is a model for $\mathbb{F}_2\mathbb{P}^2$.
9. Use homogeneous coordinates to count the number of points in $\mathbb{F}_3\mathbb{P}^2$.

!"#$# % & & !' (! !) * * + , * - . *
/ * !9 . * * , * 3 @ . . * * * 3 1 * ,
, 5 "/" & 6 13
+* 0) * 0 9 7 A 0 9 4 0 & A

!"#$# % & & !' (! !) * * + , * - . *
/ * !F 3 + , * 3 $ 1 ., 5 "/" & 6 13

+* 0) * A0 9 7 0 9 4 0 &

;. 5 6 $ + , 7 * 3

4

The Geometry of Submanifolds

Werner Ballmann

© Springer Basel 2018
W. Ballmann, *Introduction to Geometry and Topology*, Compact Textbooks in Mathematics,
https://doi.org/10.1007/978-3-0348-0983-2_4

In this chapter we will discuss the geometry of submanifolds of Euclidean spaces. We assume that the reader is familiar with the fundamentals of Euclidean geometry, that is, the geometry of $\mathbb{R}^m$ equipped with the Euclidean scalar product and the associated metric $d(p,q) = \|p-q\|$. A motion of $\mathbb{R}^m$ is a map $B\colon \mathbb{R}^m \longrightarrow \mathbb{R}^m$ of the form $B(x) = Sx + t$ for $S \in \mathrm{O}(m)$ and $t \in \mathbb{R}^m$. As a warm-up exercise, show that a map $\mathbb{R}^m \longrightarrow \mathbb{R}^m$ preserves distances if and only if it is a motion.

The geometry of submanifolds consists of two parts: interior and exterior geometry. Interior geometry relates to measurements within the submanifold, exterior geometry to the form of the submanifold relative to the surrounding Euclidean space. We will first briefly discuss these two aspects for curves in Euclidean spaces, for which interior geometry naturally consists only of measuring the lengths of curve segments. Afterwards, we will consider the interior geometry of submanifolds, then their exterior geometry, and conclude by proving the Theorema Egregium of Gauß.[1]

We are interested in *geometric properties and invariants* of submanifolds $M \subseteq \mathbb{R}^n$. These should not depend on the coordinates of M, and should not change if M is replaced by $B(M)$, where B is a motion of $\mathbb{R}^n$. The lengths of curves are examples of such invariants, but there are less obvious, deeper, properties and invariants.

Throughout, we identify the tangent space $T_x\mathbb{R}^n$ with $\mathbb{R}^n$ as in Example 2.2.3.1, and the tangent spaces of submanifolds $M \subseteq \mathbb{R}^n$ with linear subspaces of $\mathbb{R}^n$ as in Proposition 2.3.2.2. Compare with Example 2.2.3.2. More generally, we consider immersions $f\colon M \longrightarrow \mathbb{R}^n$ instead of submanifolds.

With the above identification of the tangent spaces, a *vector field along* f is, in this chapter, a map $X\colon M \longrightarrow \mathbb{R}^n$. We then imagine $X(p)$ as a vector with basepoint $f(p)$. We call such an f *tangential*, if $X(p) \in \operatorname{im} f_{*p}$ for all $p \in M$, that is, when there is a vector field Y on M with $X = df \circ Y$. For a smooth map φ from M to a vector space V, we also commonly write $X\varphi$ instead of $d\varphi \circ X$. Compare with (2.19).

---

[1] Johann Carl Friedrich Gauß (1777–1855).

4.1 Curves

We begin our investigations with curves in Euclidean spaces. If $c\colon I \longrightarrow \mathbb{R}^n$ is a smooth curve, we typically interpret the parameter $t \in I$ as time, and therefore call $\|\dot{c}\|$ the *speed* of c. We call c *regular*, if $\dot{c}(t) \neq 0$ for all $t \in I$. If c is regular, then the *field of directions* of c,

$$e\colon I \longrightarrow \mathbb{R}^3, \quad e(t) := \dot{c}(t)/\|\dot{c}(t)\|, \tag{4.1}$$

is a smooth vector field along c with constant norm 1. We will sometimes also refer to e as the *direction field* of c.

Example 4.1.1

As curves $c\colon \mathbb{R} \longrightarrow \mathbb{R}^2$,

1. $c(t) = (t, t^3)$ is regular,
2. $c(t) = (t^2, t^3)$ is not regular at $t = 0$ and its image has a crease at $(0, 0)$,
3. $c(t) = (t^3, t^3)$ is not regular at $t = 0$, although its image, a line, looks regular.

■

If $c\colon I \longrightarrow \mathbb{R}^n$ is a smooth curve and $\varphi\colon J \longrightarrow I$ is a diffeomorphism, then $\bar{c} = c \circ \varphi$ is regular if and only if c is regular. We refer to such a φ as a *change of parameters* and $\bar{c}$ as a *reparameterization* of c. More generally, we call $c \circ \varphi$ a *monotone reparameterization* of c, if $\varphi\colon [\alpha, \beta] \longrightarrow [a, b]$ is monotone, surjective, and smooth.

4.1.1 Length and Energy

Geometry in Euclidean spaces is based on length measurements. Let $c\colon [a, b] \longrightarrow \mathbb{R}^n$ be a smooth curve. Then

$$L(c) := \int_a^b \|\dot{c}(t)\|\, dt \quad \text{and} \quad E(c) = \frac{1}{2} \int_a^b \|\dot{c}(t)\|^2\, dt \tag{4.2}$$

are respectively called the *length* and *energy* of c. It follows from the Cauchy-Schwarz inequality[2] that

$$L^2(c) = \Big(\int_a^b 1 \cdot \|\dot{c}\| \Big)^2 \leq \Big(\int_a^b 1^2 \Big) \Big(\int_a^b \|\dot{c}\|^2 \Big) = 2(b - a) E(c), \tag{4.3}$$

and equality holds if and only if c has constant speed, that is, when the function $\|\dot{c}\|$ is constant. Clearly, the length and energy of curves are invariant under motions: If B is a motion of $\mathbb{R}^n$, then $L(B \circ c) = L(c)$ and $E(B \circ c) = E(c)$. The lengths of curves are, moreover, independent of the parameterization of the curve, see Exercise 4.6.1.1. Length is therefore a geometric invariant of curves in the sense described above.

---

[2] Hermann Amandus Schwarz (1843–1921).

We say that a smooth curve $c\colon I \longrightarrow \mathbb{R}^n$ is *parameterized by arc length*, if c has constant speed $\|\dot{c}(t)\| = 1$. Then c is regular, and the lengths of segments of c correspond to the lengths of the parameter intervals.

Proposition 4.1.2 *If $c\colon I \longrightarrow \mathbb{R}$ is a regular curve and $t_0 \in I$, there is a change of parameters $\varphi\colon J \longrightarrow I$ with $0 \in J$ such that $\varphi(0) = t_0$ and $c \circ \varphi$ is parameterized by arc length.*

Proof

Define $\psi\colon I \longrightarrow \mathbb{R}$ via $\psi(t) := \int_{t_0}^{t} \|\dot{c}\|$. Since c is regular, ψ is smooth with $\dot{\psi}(t) = \|\dot{c}(t)\| \neq 0$. Therefore ψ is a diffeomorphism on an interval $J \subseteq \mathbb{R}$ with $\psi(t_0) = 0$. The inverse map $\varphi\colon J \longrightarrow I$ is then the desired change of parameters. □

We now discuss a few elementary facts about length and energy. The arguments are chosen so that they may be used again in later extensions of these claims.

Proposition 4.1.3 *For $x, y \in \mathbb{R}^n$ and all smooth curves $c\colon [a, b] \longrightarrow \mathbb{R}^n$ from x to y, $L(c) \geq d(x, y)$. Equality holds if and only if c is a monotone reparameterization of the line $ty + (1 - t)x$, $0 \leq t \leq 1$, from x to y.*

Proof

We assume $x \neq y$ and set $v := (y - x)/\|y - x\|$, the unit vector pointing from x to y. Let $h : \mathbb{R}^n \longrightarrow \mathbb{R}$, $h(z) := \langle v, z\rangle$, be the height function associated to v. Then $\operatorname{grad} h = v$, and therefore

$$\begin{aligned} L(c) = \int_a^b \|\dot{c}\| \geq \int_a^b \langle v, \dot{c}\rangle = \int_a^b \langle \operatorname{grad} h, \dot{c}\rangle \\ = h(c(b)) - h(c(a)) = h(y) - h(x) = \|y - x\| = d(x, y). \end{aligned}$$

Equality implies that $\langle v, \dot{c}\rangle = \|\dot{c}\|$. □

Corollary 4.1.4 *For $x, y \in \mathbb{R}^n$,*

$$d(x, y) = \min\{L(c) \mid c \text{ is a smooth curve from } x \text{ to } y\},$$

and the minimum is realized precisely by the monotone reparameterizations of the line segment $ty + (1 - t)x$, $0 \leq t \leq 1$, from x to y. □

Energy possesses better analytic properties than length, it is *not* invariant under change of parameters, but rather prefers curves of constant speed:

Proposition 4.1.5 *Let $x, y \in \mathbb{R}^n$ and*

$$c_0\colon [a, b] \longrightarrow \mathbb{R}^n, \quad c_0(t) = \frac{1}{b - a}\big((t - a)y + (b - t)x\big),$$

be the line segment from x to y. Then for all smooth curves $c\colon [a,b] \longrightarrow \mathbb{R}^n$ from x to y, $E(c) \geq E(c_0)$, and equality implies $c = c_0$.

Proof

We assume $x \neq y$. With terminology as in Proposition 4.1.3 we then obtain

$$\begin{aligned}
2E(c) &= \int_a^b \|\dot{c}\|^2 \geq \int_a^b \langle v, \dot{c}\rangle^2 \\
&\geq \frac{1}{b-a}\Big(\int_a^b \langle v, \dot{c}\rangle\Big)^2 \\
&= \frac{1}{b-a}\big(h(y) - h(x)\big)^2 = 2E(c_0).
\end{aligned}$$

Equality implies that $\langle v, \dot{c}\rangle$ is constant and $\dot{c} = \langle v, \dot{c}\rangle v$. □

4.1.2 Curvature

While the derivative of a curve at a point is the best approximation of the curve by a line, we are now interested in approximations of second order. To this end, let $c\colon I \longrightarrow \mathbb{R}^n$ be a smooth curve and t_0 be a point in the interior of I such that $\dot{c}(t_0)$ and $\ddot{c}(t_0)$ are linearly independent.

Proposition 4.1.6 *For $t_1 < t_2 < t_3$ in I close enough to t_0, $c(t_1)$, $c(t_2)$ and $c(t_3)$ are not collinear. Moreover, for $t_1, t_2, t_3 \longrightarrow t_0$, the unique circle $K(t_1, t_2, t_3)$ through $c(t_1)$, $c(t_2)$, and $c(t_3)$ converges to a limit circle $K(t_0)$; see Fig. 4.1. This limit circle passes through $c(t_0)$, is tangent to c at $c(t_0)$, and lies in the affine plane $E(t_0)$ through $c(t_0)$ spanned by $\dot{c}(t_0)$ and $\ddot{c}(t_0)$. Its center $M = M(t_0) \in E(t_0)$ is determined by the following system of linear equations:*

$$\begin{aligned}
\langle M - c(t_0), \dot{c}(t_0)\rangle &= 0, \\
\langle M - c(t_0), \ddot{c}(t_0)\rangle &= \|\dot{c}(t_0)\|^2.
\end{aligned}$$

Fig. 4.1 Circles approximating a curve

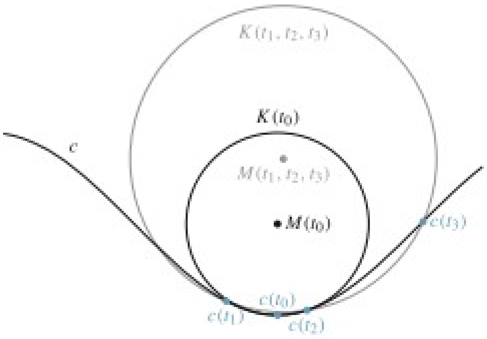

Proof

The first claim follows easily from the second-order Taylor approximation[3] to c at t_0. We now denote by $M(t_1, t_2, t_3)$ the center of the circle $K(t_1, t_2, t_3)$. Then the function

$$f = f(t) := \|c(t) - M(t_1, t_2, t_3)\|^2$$

takes the same value at the points t_1, t_2, and t_3, namely, the square of the radius of $K(t_1, t_2, t_3)$. Its first derivative

$$2\langle c(t) - M(t_1, t_2, t_3), \dot{c}(t)\rangle$$

therefore has zeros at points $s_1 \in (t_1, t_2)$ and $s_2 \in (t_2, t_3)$, and so its second derivative

$$2\langle c(t) - M(t_1, t_2, t_3), \ddot{c}(t)\rangle + 2\|\dot{c}(t)\|^2$$

has a zero at a point $r \in (s_1, s_2)$. □

Definition 4.1.7

For c and t_0 as above, we call $E(t_0)$ the *osculating plane* and $K(t_0)$ the *osculating circle* of c at t_0. We call $M(t_0)$ the *center of curvature*, $R(t_0) = \|M - c(t_0)\|$ the *radius of curvature*, and $\kappa(t_0) = 1/R(t_0)$ the *curvature* of c at t_0.

In light of the first equation in Proposition 4.1.6, the second equation says that $R(t_0)$ times the norm of the component of $\ddot{c}(t_0)$ perpendicular to $\dot{c}(t_0)$ is equal to $\|\dot{c}(t_0)\|^2$. In other words,

$$\kappa(t_0) = \frac{\|\ddot{c}(t_0) - \langle e(t_0), \ddot{c}(t_0)\rangle e(t_0)\|}{\|\dot{c}(t_0)\|^2}, \tag{4.4}$$

where $e = \dot{c}/\|\dot{c}\|$ denotes the field of directions of c as in (4.1). We thereby obtain a second definition of the curvature $\kappa(t_0)$ of c at t_0, consistent with the first, that also applies when $\ddot{c}(t_0)$ is linearly dependent on $\dot{c}(t_0)$.

If c is parameterized by arc length, then $\langle \dot{c}, \dot{c}\rangle/2$ is by definition constant, and therefore has derivative $\langle \dot{c}, \ddot{c}\rangle = 0$. Then $\ddot{c}(t)$ is perpendicular to $\dot{c}(t)$ for all $t \in I$ and consequently

$$\kappa(t_0) = \|\ddot{c}(t_0)\|. \tag{4.5}$$

In Exercises 4.6.2 and 4.6.3 we will check that the curvature of curves is a geometric invariant, which adequately replicates our conception of curvature.

---

[3] Brook Taylor (1685–1731).

4.1.3 Plane Curves

We call curves $c\colon I \longrightarrow \mathbb{R}^2$ *plane curves*. If $c\colon I \longrightarrow \mathbb{R}^2$ is a regular plane curve, then there is exactly one vector field n along c perpendicular to the field of directions $e = \dot{c}/\|\dot{c}\|$ of c such that $(e(t), n(t))$ is a positively oriented basis of $\mathbb{R}^2$ for all $t \in I$, namely

$$n = (-e^2, e^1)$$

with $e = (e^1, e^2)$; see Fig. 4.2. We call $n = n(t)$ the *principal normal (vector) field* of c. This allows us to give a sign to the curvature of regular plane curves,

$$\kappa_o(t) := \frac{\langle n(t), \ddot{c}(t)\rangle}{\|\dot{c}(t)\|^2}. \tag{4.6}$$

We call κ_o the *oriented curvature* of c. We can rewrite the formula (4.6) as

$$\kappa_o(t) = \frac{\det(\dot{c}(t), \ddot{c}(t))}{\|\dot{c}(t)\|^3}. \tag{4.7}$$

If c is parameterized by arc length, we obtain

$$\kappa_o(t) = \langle n(t), \ddot{c}(t)\rangle = \det(\dot{c}(t), \ddot{c}(t)). \tag{4.8}$$

It is clear that, for the curvature defined in the preceding section, $\kappa = |\kappa_o|$.

Fig. 4.2 The vector fields e and n

Proposition 4.1.8 *For the field of directions e and the principal normal field n of a regular plane curve $c\colon I \longrightarrow \mathbb{R}^2$,*

$$\dot{e} = \|\dot{c}\|\kappa_o n \quad \textit{and} \quad \dot{n} = -\|\dot{c}\|\kappa_o e.$$

Proof
Since the functions $\langle e, e\rangle$ and $\langle n, n\rangle$ are constant, their derivatives vanish, so $\langle e, \dot{e}\rangle = \langle n, \dot{n}\rangle = 0$. Therefore $\dot{e}$ is a multiple of n and $\dot{n}$ is a multiple of e. Now $\|\dot{c}\|e = \dot{c}$, and therefore $\|\dot{c}\|\langle n, \dot{e}\rangle = \langle n, \ddot{c}\rangle = \|\dot{c}\|^2\kappa_o$. This proves the first equation. The function $\langle n, e\rangle$ is also constant, so $\langle \dot{n}, e\rangle = -\langle n, \dot{e}\rangle$. The second equation follows. □

With the aid of the differential equations from Proposition 4.1.8, we obtain further interpretations of the curvature: Let $c\colon I \longrightarrow \mathbb{R}^2$ be a regular plane curve, and $v \in \mathbb{R}^2$

be a fixed unit vector. Let $\alpha = \alpha(t) \in \mathbb{R}/2\pi\mathbb{Z}$ be the oriented angle from v to $e(t)$, that is $\cos\alpha = \langle v, e\rangle$ and $\sin\alpha = -\langle v, n\rangle$. With Proposition 4.1.8, it follows that

$$\dot\alpha \sin\alpha = -\langle v, \dot e\rangle = \|\dot c\|\kappa_o \sin\alpha,$$
$$\dot\alpha \cos\alpha = -\langle v, \dot n\rangle = \|\dot c\|\kappa_o \cos\alpha.$$

Therefore, we have

$$\dot\alpha = \|\dot c\|\,\kappa_o. \tag{4.9}$$

If $I = \mathbb{R}$ and c is periodic with period $\omega > 0$, then $e(\omega) = e(0)$, so $\alpha(\omega) = \alpha(0)$ modulo $2\pi\mathbb{Z}$ and therefore

$$\frac{1}{2\pi}\int_0^\omega \kappa_0(t)\|\dot c(t)\|\,dt = k \in \mathbb{Z}. \tag{4.10}$$

The number k is called the *winding number* of c (with respect to the given period ω). The winding number counts the number of times e circles the origin during one period. The *Winding Number Theorem* states that the winding number is ±1 if c is a Jordan curve, that is, the curve does not repeat points except when forced to by the period. See, for example, section 2.2 in [Kl].

A further interpretation of the curvature refers to the change in the length of c when we pass to parallel curves. To this end, let $I = [a, b]$ and

$$c_s \colon [a, b] \longrightarrow \mathbb{R}^2, \quad c_s(t) := c(t) + sn(t), \quad s \in \mathbb{R}, \tag{4.11}$$

be the family of *curves parallel to* $c = c_0$. Then by Proposition 4.1.8,

$$\dot c_s = \dot c - s\|\dot c\|\kappa_o e = (1 - s\kappa_o)\dot c.$$

Therefore the derivative δL of the length $L(c_s)$ as a function of s at $s = 0$ is given by

$$\delta L = -\int_a^b \kappa_o(t)\|\dot c(t)\|\,dt. \tag{4.12}$$

We can therefore interpret the oriented curvature as a measure of the change in length of a plane curve in the normal direction.

For a given speed and a given curvature, the equations from Proposition 4.1.8 are a linear system of ordinary differential equations for the fields e and n. In preparation for the next proposition, we write these equations in matrix form,

$$\dot F = FS \tag{4.13}$$

with

$$F := \begin{pmatrix} e^1 & n^1 \\ e^2 & n^2 \end{pmatrix} \quad \text{and} \quad S := \|\dot c\| \cdot \begin{pmatrix} 0 & -\kappa_o \\ \kappa_o & 0 \end{pmatrix}. \tag{4.14}$$

Here, the coefficients of F and S are functions of $t \in I$, and differentiating F means differentiating the coefficients of F. The important point in (4.13) is the skew-symmetry of S.

Proposition 4.1.9 *Let $T: I \longrightarrow (0, \infty)$ and $\kappa_o: I \longrightarrow \mathbb{R}$ be smooth functions. Then:*

1. *(Existence) For $t_0 \in I$ and $x_0, e_0 \in \mathbb{R}^2$ with $\|e_0\| = 1$, there is a regular plane curve $c: I \longrightarrow \mathbb{R}^2$ with $c(t_0) = x_0$ and $e(t_0) = e_0$ such that T is the speed of c and κ_o is the oriented curvature of c.*
2. *(Uniqueness) For every two regular plane curves $c_1, c_2: I \longrightarrow \mathbb{R}^2$ with speed T and oriented curvature κ_o, there is precisely one orientation-preserving motion B of $\mathbb{R}^2$ with $c_2 = B \circ c_1$.*

Proof

Let $F: I \longrightarrow \mathbb{R}^{2\times 2}$ be a solution of the linear ordinary differential equation (4.13) with $\|\dot{c}\|$ replaced by T and with $F(t_0) \in \mathrm{SO}(2)$. By the skew-symmetry of S,

$$\frac{d}{dt}(FF^*) = \dot{F}F^* + F\dot{F}^* = FSF^* + FS^*F^* = FSF^* - FSF^* = 0.$$

It therefore follows that $F(t)$ is in $\mathrm{O}(2)$ for all $t \in I$. However, F is a smooth function of t, and therefore so is $\det F$. Since $\det F(t) = \pm 1$ for all $t \in I$ and $\det F(t_0) = 1$, we must have $\det F(t) = 1$ for all $t \in I$. From this it follows that $F(t)$ is in $\mathrm{SO}(2)$ for all $t \in I$. We now choose the particular solution $F: I \longrightarrow \mathbb{R}^{2\times 2}$ with

$$F(t_0) = \begin{pmatrix} e_0^1 & -e_0^2 \\ e_0^2 & e_0^1 \end{pmatrix} \in \mathrm{SO}(2).$$

By what we have just proved, $F(t)$ is in $\mathrm{SO}(2)$ for all $t \in I$. Moreover, the first column e and the second column n of F satisfy the differential equations from Proposition 4.1.8 (with $\|\dot{c}\|$ replaced by T). Since $F(t) \in \mathrm{SO}(2)$, we then have that $(e(t), n(t))$ is a positively oriented basis of $\mathbb{R}^2$ for all $t \in I$. Therefore,

$$c = c(t) := x_0 + \int_{t_0}^{t} T \cdot e$$

is a smooth plane curve with field of directions e, principal normal field n, speed T, and oriented curvature κ_o. Claim 1 of the proposition follows from this.

Now let $c_1, c_2: I \longrightarrow \mathbb{R}^2$ be curves with speed T and oriented curvature κ_o. Let $B = Ax + a$ be the unique, orientation-preserving motion of $\mathbb{R}^2$ with

$$A(c_1(t_0)) + a = c_2(t_0), \quad A(e_1(t_0)) = e_2(t_0), \quad A(n_1(t_0)) = n_2(t_0),$$

where e_1, e_2, n_1 and n_2 denote the fields of directions and principal normal fields associated to c_1 and c_2. Then $B \circ c_1$ is a curve parameterized by arc length with direction field $A \circ e_1$, principal normal field $A \circ n_1$, and oriented curvature κ_o. Therefore, $A \circ e_1$ and $A \circ n_1$ are solutions to the equations of Proposition 4.1.8 with the same initial conditions as e_2 and n_2. It therefore follows that $A \circ e_1 = e_2$ and $A \circ n_1 = n_2$. Finally, since $A(c_1(t_0)) + a = c_2(t_0)$, it follows that $c_2 = B \circ c_1$. □

Remark 4.1.10 The Lie algebra of SO(2) (the tangent space at the identity matrix) consists of the skew-symmetric (2×2)-matrices, compare with Example 2.3.8.3. The matrix $S = S(t)$ in (4.13) is skew-symmetric, so $\mathrm{SO}(2) \ni B \mapsto BS(t)$ is a left-invariant vector field on SO(2) for all $t \in I$; see also Exercise 2.7.14.1. The solutions of the corresponding differential equations on SO(2) therefore lie in SO(2). The first part of the above proof shows this with elementary methods.

Corollary 4.1.11 *A regular plane curve moves counter-clockwise around a circle of radius $R > 0$ if and only if its oriented curvature is constant with value $1/R$.*

4.1.4 Space Curves

We now move to the next case, $n = 3$, that of *space curves*, i.e. curves in $\mathbb{R}^3$. Let $c\colon I \longrightarrow \mathbb{R}^3$ be a smooth space curve, such that $\dot{c}(t)$ and $\ddot{c}(t)$ are linearly independent for all $t \in I$. Then c is regular, and, in addition to the field of directions $e = \dot{c}/\|\dot{c}\|$, we obtain a second vector field along c,

$$n := \frac{\ddot{c} - \langle e, \ddot{c}\rangle e}{\|\ddot{c} - \langle e, \ddot{c}\rangle e\|}, \tag{4.15}$$

the *principal normal field* along c. Together with the *binormal field*

$$b = e \times n, \tag{4.16}$$

where × denotes the cross product, we obtain the *Frenet frame* $e, n, b\colon I \longrightarrow \mathbb{R}^3$ of c. The triple $(e(t), n(t), b(t))$ is a positively oriented orthonormal basis of $\mathbb{R}^3$ for all $t \in I$. The corresponding differential equations are called the *Frenet-Serret formulas*[4]:

$$\dot{e} = \|\dot{c}\|\kappa\, n, \quad \dot{n} = -\|\dot{c}\|\kappa\, e + \|\dot{c}\|\tau\, b, \quad \dot{b} = -\|\dot{c}\|\tau\, n. \tag{4.17}$$

Here, $\tau = \tau(t) := \langle \dot{n}(t), b(t)\rangle = -\langle n(t), \dot{b}(t)\rangle$ is called the *torsion* of c.

As in the case of plane curves, it is instructive and helpful to write the Frenet-Serret formulas in matrix form,

$$\dot{F} = FS \tag{4.18}$$

with

$$F := \begin{pmatrix} e^1 & n^1 & b^1 \\ e^2 & n^2 & b^2 \\ e^3 & n^3 & b^3 \end{pmatrix} \quad \text{and} \quad S := \|\dot{c}\| \cdot \begin{pmatrix} 0 & -\kappa & 0 \\ \kappa & 0 & -\tau \\ 0 & \tau & 0 \end{pmatrix}. \tag{4.19}$$

The skew-symmetry of S is again important. Using it, we obtain a theorem about space curves analogous to Proposition 4.1.9:

---

[4] Jean Frédéric Frenet (1816–1900), Joseph Alfred Serret (1819–1885).

Proposition 4.1.12 *Let* $T, \kappa : I \longrightarrow (0, \infty)$ *and* $\tau : I \longrightarrow \mathbb{R}$ *be smooth functions.*

1. *(Existence) For* $t_0 \in I$ *and* $x_0, e_0, n_0 \in \mathbb{R}^3$ *with* (e_0, n_0) *orthonormal, there is a regular space curve* $c : I \longrightarrow \mathbb{R}^3$ *with* $c(t_0) = x_0$, $e(t_0) = e_0$ *and* $n(t_0) = n_0$, *such that* T *is the speed of* c, κ *the curvature, and* τ *the torsion.*
2. *(Uniqueness) For every two regular space curves* $c_1, c_2 : I \longrightarrow \mathbb{R}^3$ *with speed* T, *curvature* κ, *and torsion* τ, *there is exactly one orientation-preserving motion* B *of* $\mathbb{R}^3$ *with* $c_2 = B \circ c_1$.

Proof

The proof is, *mutatis mutandis*, identical to the proof of Theorem 4.1.9: In some places, the numeral 2 is replaced by 3; the matrices F and S are now (3×3)-matrices as in (4.19), and the initial conditions read

$$F(t_0) := (e_0, n_0, e_0 \times n_0) \in \mathrm{SO}(3).$$

We leave reworking the details as an exercise. □

4.1.5 Parallelism from a New Perspective

We can only define the Frenet frame along a space curve c if $\dot{c}$ and $\ddot{c}$ are pointwise linearly independent. Moreover, the differential equations (4.17) of Frenet and Serret involve third derivatives of c. An analogous theory in arbitrary dimensions, in particular for curves in $\mathbb{R}^n$ with $n \geq 4$, involves the first n derivatives of the curve and requires that the first $n - 1$ thereof are pointwise linearly independent. Compare, for example, with sections 1.2 and 1.3 in [Kl]. This is rather a lot to ask when n is large.

To arrive at a new canonical class of vector fields along c, we bring another idea into play, that of curve-dependent parallelism. This is a central notion in modern differential geometry, that we first encounter here in a simple situation. To this end, let $c : I \longrightarrow \mathbb{R}^n$ be a regular curve.

Definition 4.1.13

We call a vector field $X : I \longrightarrow \mathbb{R}^n$ along c a *normal field*, if X is perpendicular to c, i.e., if $\langle \dot{c}(t), X(t) \rangle = 0$ for all $t \in I$. For a normal field X along c, we call the component of $\dot{X}$ normal to c the *covariant derivative* of X along c, written as $\nabla X/dt$ or X':

$$\frac{\nabla X}{dt} = X' := \dot{X} - \langle e, \dot{X} \rangle e.$$

We call a normal field X along c *parallel* if $X' = 0$.

If X is a normal field along c, then the function $\langle e, X\rangle$ vanishes identically, and therefore so does its derivative. We thereby obtain

$$\langle e, \dot{X}\rangle = -\langle \dot{e}, X\rangle. \tag{4.20}$$

The component $\langle e(t), \dot{X}(t)\rangle e(t)$ of $\dot{X}(t)$ tangent to c therefore only depends on the value of X at t. It moreover follows that, for given $t \in I$ and $v \in \mathbb{R}^n$ perpendicular to $e(t)$, the derivative of a normal field X along c with $X(t) = v$ has norm at least $|\langle \dot{e}(t), v\rangle|$, with equality holding if and only if $\dot{X}(t)$ is tangent to c. Parallel normal fields are, in this sense, the most parsimonious normal fields.

Proposition 4.1.14 *Let X and Y be normal fields along c. Then the following hold:*

1. *(Linearity) For $\alpha, \beta \in \mathbb{R}$, $(\alpha X + \beta Y)' = \alpha X' + \beta Y'$.*
2. *(Product rule) The function $\langle X, Y\rangle$ has derivative $\langle X', Y\rangle + \langle X, Y'\rangle$.*

Proof

Equation (1) is clear. To prove (2), we compute

$$\frac{d}{dt}\langle X, Y\rangle = \langle \dot{X}, Y\rangle + \langle X, \dot{Y}\rangle = \langle X', Y\rangle + \langle X, Y'\rangle,$$

where we use $\langle e, X\rangle = \langle e, Y\rangle = 0$ for the equality on the right. □

Corollary 4.1.15 *For parallel normal fields X and Y along c, the following hold:*

1. *Linear combinations $\alpha X + \beta Y$ are themselves parallel along c.*
2. *The function $\langle X, Y\rangle$ is constant.* □

Proposition 4.1.16 *For $t_0 \in I$ and $X_0 \in \mathbb{R}^n$ perpendicular to $\dot{c}(t_0)$, there is precisely one parallel normal field X along c with $X(t_0) = X_0$.*

Proof

By Definition 4.1.13 and (4.20), parallel normal fields X along c satisfy the linear ordinary differential equation

$$\dot{X} = -\frac{\langle \ddot{c}, X\rangle}{\|\dot{c}\|^2}\dot{c}. \tag{4.21}$$

Conversely, let $X : I \longrightarrow \mathbb{R}^n$ be the solution of (4.21) with $X(t_0) = X_0$. Scalar multiplication of the left- and right-hand sides of (4.21) by $\dot{c}$ shows that $\langle \dot{c}, X\rangle$ has vanishing derivative, and therefore is constant. Since $\langle \dot{c}(t_0), X(t_0)\rangle = 0$, we conclude from this that X is a normal field along c. In summation, then, it follows that a vector field X along c is a parallel normal field along c if and only if it is a solution of (4.21) and $X(t)$ is perpendicular to $\dot{c}(t)$ for some $t \in I$. □

Corollary 4.1.17 *Let $c \colon I \longrightarrow \mathbb{R}^n$ be a regular curve. For $t_0 \in I$ and orthonormal vectors $x_2, \ldots, x_n \in \mathbb{R}^n$ perpendicular to $\dot{c}(t_0)$, let $X_2, \ldots, X_n$ be the parallel normal fields along*

c with $X_i(t_0) = x_i$ for all $2 \le i \le n$. Then $(e(t), X_2(t), \dots, X_n(t))$ is an orthonormal basis of $\mathbb{R}^n$ for all $t \in I$. □

A parallel normal field along c does not generally consist of vectors that are parallel in the usual sense—if that were the case, the normal field would be constant. To determine the parallel normal fields along c, one must solve the differential equation (4.21).

Example 4.1.18

Let c be a space curve such that $\dot{c}$ and $\ddot{c}$ are pointwise linearly independent. The Frenet frame $(e, n, b = e \times n)$ is then easy to determine. Determining the parallel normal fields is more difficult. We can express the normal fields as linear combinations $X = \alpha n + \beta b$ with real functions α and β. With the aid of (4.17), the condition that X be parallel then translates to the differential equations

$$\dot{\alpha} = \|\dot{c}\|\tau\beta \quad \text{and} \quad \dot{\beta} = -\|\dot{c}\|\tau\alpha$$

for the coefficients. If the speed and torsion of c are constant, then this system of equations is easy to solve: If α solves the differential equation $\ddot{\alpha} + \|\dot{c}\|\tau\alpha = 0$, then the pair α and $\beta := \dot{\alpha}/\|\dot{c}\|\tau$ solves the system of differential equations above. ■

Proposition 4.1.19 *For smooth functions $T\colon I \longrightarrow (0, \infty)$ and $\kappa_2, \dots, \kappa_n\colon I \longrightarrow \mathbb{R}$, the following hold:*

1. *(Existence) For $t_0 \in I$ and $x_0, \dots, x_n \in \mathbb{R}^n$ with $(x_1, \dots, x_n)$ orthonormal, there exist a regular curve $c\colon I \longrightarrow \mathbb{R}^n$ with speed T and parallel normal fields $X_2, \dots, X_n$ along c with $c(t_0) = x_0$, $e(t_0) = x_1$, $X_i(t_0) = x_i$, and $\langle \dot{e}, X_i \rangle = \kappa_i T$ for all $2 \le i \le n$.*
2. *(Uniqueness) For every two regular curves $c_1, c_2\colon I \longrightarrow \mathbb{R}^n$ with speed T and parallel and pointwise orthonormal normal fields $X_2, \dots, X_n$ along c_1 and $Y_2, \dots, Y_n$ along c_2 with $\langle \dot{e}_1, X_i \rangle = \kappa_i T$ and $\langle \dot{e}_2, Y_i \rangle = \kappa_i T$, there is exactly one motion $B = Ax + a$ of $\mathbb{R}^n$ with $c_2 = B \circ c_1$ and $Y_i = A \circ X_i$ for all $2 \le i \le n$.*

Proof

We once again write the differential equations for e and $X_2, \dots, X_n$ in matrix form,

$$\dot{F} = FS \tag{4.22}$$

with

$$F := \begin{pmatrix} e^1 & X_2^1 & \dots & X_n^1 \\ \vdots & \vdots & & \vdots \\ e^n & X_2^n & \dots & X_n^n \end{pmatrix} \quad \text{and} \quad S := T \cdot \begin{pmatrix} 0 & -\kappa_2 & \dots & -\kappa_n \\ \kappa_2 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ \kappa_n & 0 & \dots & 0 \end{pmatrix}. \tag{4.23}$$

The skew-symmetry of S is again important. The rest proceeds as before. □

Compare the matrices S from (4.19) and (4.23) and the corresponding systems of differential equations (4.18) and (4.22). This "small change" to the matrix S has a large effect, namely that we now have canonical normal fields along regular curves in Euclidean spaces regardless of the dimension of the space. Moreover, the differential equation (4.23) only involves second derivatives of c (and first derivatives of the X_i). We will end with this insight, and not delve any deeper, as we only introduced the concepts of covariant derivatives and parallelism along curves to give motivation for analogous concepts in our later discussions.

4.2 Interior Geometry

After our introductory remarks about curves, we now consider the general case of submanifolds in the Euclidean space $\mathbb{R}^n$. There are two ways these may be specified: firstly via equations, that is as level sets of mappings, or via embeddings or immersions $M \longrightarrow \mathbb{R}^n$. The second case includes the first, since we can think of a submanifold $M \subseteq \mathbb{R}^n$ with inclusion $i\colon M \longrightarrow \mathbb{R}^n$ as an embedding. Nevertheless, it is often worthwhile to give results separately for the first case, as their formulations are usually simpler and easier to grasp.

The interior geometry of submanifolds deals with measurements of geometric objects which are contained in the submanifold. The measuring stick is provided by the ambient space $\mathbb{R}^n$, namely, the Euclidean scalar product. In every other respect, however, the surrounding space is ignored, in some sense because the inhabitants of the submanifold are incapable of perceiving the space outside the submanifold.[5] In the preceding discussion of curves, their lengths are measurements belonging to their interior geometry, but their curvature is not.

In the following, we denote by M a manifold of dimension m, and by $f\colon M \longrightarrow \mathbb{R}^n$ an immersion. For $p \in M$, we call the linear subspaces

$$T_p f := \operatorname{im} df(p) \quad \text{and} \quad N_p f := [\operatorname{im} df(p)]^{\perp} \tag{4.24}$$

of $\mathbb{R}^n$ the *tangent space* and *normal space* to f at p. We denote by

$$\pi_p^T : \mathbb{R}^n \longrightarrow \mathbb{R}^n \quad \text{and} \quad \pi_p^N : \mathbb{R}^n \longrightarrow \mathbb{R}^n \tag{4.25}$$

the orthogonal projections from $\mathbb{R}^n$ to $T_p f$ and $N_p f$. If $M \subseteq \mathbb{R}^n$ is a submanifold with inclusion f, then $T_p f = T_p M$ under the usual identification of $T_p M$ as in Proposition 2.3.2. Therefore, we then write $T_p M$ and $N_p M$ instead of $T_p f$ and $N_p f$.

Example 4.2.1

1) For a regular curve $c\colon I \longrightarrow \mathbb{R}^n$, $T_t c = \mathbb{R} \cdot \dot{c}(t) = \mathbb{R} \cdot e(t)$. For regular plane curves c, $N_t c = \mathbb{R} \cdot n(t)$, for regular space curves c, $N_t c$ is the span of the principal normal $n(t)$ and the binormal $b(t)$.

---

[5] This perceptual difficulty is one of the topics of the novella *Flatland*, written by Edwin Abbott Abbott (1838–1926) and published in 1884.

;. 5 6 F + ,

$. , 7 * 3 3 * * .
7J * * 3 7 * 3 $ * . (1 & 3 3 7,
* 7 1*) 0 * ,0 @ 3 * * * 1 3 F & , 1 0
& 0 3 . 1* . 30 7 * *7 *
7 * 3 * *1*7 1 & . 1* 3 7 * 3 F
1 3 . 3 & 0 . * * 7 . .
. ,0 7 &*

Compare the matrices S from (4.19) and (4.23) and the corresponding systems of differential equations (4.18) and (4.22). This "small change" to the matrix S has a large effect, namely that we now have canonical normal fields along regular curves in Euclidean spaces regardless of the dimension of the space. Moreover, the differential equation (4.23) only involves second derivatives of c (and first derivatives of the X_i). We will end with this insight, and not delve any deeper, as we only introduced the concepts of covariant derivatives and parallelism along curves to give motivation for analogous concepts in our later discussions.

4.2 Interior Geometry

After our introductory remarks about curves, we now consider the general case of submanifolds in the Euclidean space $\mathbb{R}^n$. There are two ways these may be specified: firstly via equations, that is as level sets of mappings, or via embeddings or immersions $M \longrightarrow \mathbb{R}^n$. The second case includes the first, since we can think of a submanifold $M \subseteq \mathbb{R}^n$ with inclusion $i\colon M \longrightarrow \mathbb{R}^n$ as an embedding. Nevertheless, it is often worthwhile to give results separately for the first case, as their formulations are usually simpler and easier to grasp.

The interior geometry of submanifolds deals with measurements of geometric objects which are contained in the submanifold. The measuring stick is provided by the ambient space $\mathbb{R}^n$, namely, the Euclidean scalar product. In every other respect, however, the surrounding space is ignored, in some sense because the inhabitants of the submanifold are incapable of perceiving the space outside the submanifold.[5] In the preceding discussion of curves, their lengths are measurements belonging to their interior geometry, but their curvature is not.

In the following, we denote by M a manifold of dimension m, and by $f\colon M \longrightarrow \mathbb{R}^n$ an immersion. For $p \in M$, we call the linear subspaces

$$T_p f := \operatorname{im} df(p) \quad \text{and} \quad N_p f := [\operatorname{im} df(p)]^{\perp} \tag{4.24}$$

of $\mathbb{R}^n$ the *tangent space* and *normal space* to f at p. We denote by

$$\pi_p^T : \mathbb{R}^n \longrightarrow \mathbb{R}^n \quad \text{and} \quad \pi_p^N : \mathbb{R}^n \longrightarrow \mathbb{R}^n \tag{4.25}$$

the orthogonal projections from $\mathbb{R}^n$ to $T_p f$ and $N_p f$. If $M \subseteq \mathbb{R}^n$ is a submanifold with inclusion f, then $T_p f = T_p M$ under the usual identification of $T_p M$ as in Proposition 2.3.2. Therefore, we then write $T_p M$ and $N_p M$ instead of $T_p f$ and $N_p f$.

Example 4.2.1

1) For a regular curve $c\colon I \longrightarrow \mathbb{R}^n$, $T_t c = \mathbb{R} \cdot \dot{c}(t) = \mathbb{R} \cdot e(t)$. For regular plane curves c, $N_t c = \mathbb{R} \cdot n(t)$, for regular space curves c, $N_t c$ is the span of the principal normal $n(t)$ and the binormal $b(t)$.

---

[5] This perceptual difficulty is one of the topics of the novella *Flatland*, written by Edwin Abbott Abbott (1838–1926) and published in 1884.

2) The sphere $S_r^m = \{x \in \mathbb{R}^{m+1} \mid \|x\|^2 = r^2\}$ of radius $r > 0$ is a submanifold of $\mathbb{R}^{m+1}$ with $T_x S^m = \{y \in \mathbb{R}^{m+1} \mid \langle x, y\rangle = 0\}$ and $N_x M = \mathbb{R} \cdot x$ for all $x \in S_r^m$.

■

Let $X\colon M \longrightarrow \mathbb{R}^n$ be a vector field along f. Then for $p \in M$, we call

$$X^T(p) := \pi_p^T(X(p)) \quad \text{and} \quad X^N(p) = \pi_p^N(X(p)) \tag{4.26}$$

the *tangential* and *normal components* of X at p. We also write

$$X^T = \pi^T \circ X \quad \text{und} \quad X^N = \pi^N \circ X. \tag{4.27}$$

4.2.1 The First Fundamental Form

Measurements in M only depend on the restriction of the Euclidean scalar product of the ambient space $\mathbb{R}^n$ to the tangent spaces $T_p f$. The first fundamental form gets to the heart of this fact.

Definition 4.2.2

For p in M, we call the scalar product g_p on T_pM,

$$g_p(v, w) := \langle df(p)(v), df(p)(w)\rangle, \quad v, w \in T_pM,$$

the *first fundamental form* of f at p. We call the family of scalar products $g = (g_p)_{p\in M}$ the *first fundamental form* of f.

Instead of $g_p(v, w)$, we sometimes write $\langle v, w\rangle_p$ or $\langle v, w\rangle$, and similarly for the corresponding norms. These notations are particularly useful for submanifolds $M \subseteq \mathbb{R}^n$, as f is then the inclusion and $df(p)$ is the usual identification of T_pM with a linear subspace of $\mathbb{R}^n$.

We next clarify the regularity of the first fundamental form's dependence on the point p. To this end, let (U, x) be a chart on M. For all $p \in U$, the coordinate fields $\partial/\partial x^1(p), \ldots, \partial/\partial x^m(p)$ then form a basis T_pM. Since $df(p)(\partial/\partial x^i(p)) = (\partial f/\partial x^i)(p)$, the coefficients of the fundamental matrix of the first fundamental form with respect to this basis are given by

$$g_{ij}(p) := \Big\langle \frac{\partial}{\partial x^i}(p), \frac{\partial}{\partial x^j}(p)\Big\rangle_p = \Big\langle \frac{\partial f}{\partial x^i}(p), \frac{\partial f}{\partial x^j}(p)\Big\rangle. \tag{4.28}$$

Now the partial derivatives $\partial f/\partial x^i\colon U \longrightarrow \mathbb{R}^n$ are smooth, and therefore so are the functions $g_{ij}\colon U \longrightarrow \mathbb{R}$. Since the g_p are scalar products, the matrix of the g_{ij} is pointwise symmetric and positive definite.

Example 4.2.3

1) If $c : I \longrightarrow \mathbb{R}^n$ is a regular curve, then $g_{tt} = \|\dot{c}\|^2$.[6]
2) *Surfaces of revolution:* Let $c = (r, h) = (r(t), h(t))$, $t \in I$, be a regular curve in the (x, z)-plane with $r > 0$. We call c the *profile curve* of the *surface of revolution*

$$f : I \times \mathbb{R} \longrightarrow \mathbb{R}^3, \quad f(t, \varphi) := (r(t)\cos(\varphi), r(t)\sin(\varphi), h(t)).$$

A concrete example from this class is the torus considered in Exercise 2.7.8.1.
Based on the corresponding terms from geography, we call the curves $\varphi = \text{const}$ *meridians* or *longitudes*, and the curves $t = \text{const}$ *latitudes* of f. The partial derivatives of f are

$$\frac{\partial f}{\partial t} = (\dot{r}\cos(\varphi), \dot{r}\sin(\varphi), \dot{h}), \quad \frac{\partial f}{\partial \varphi} = (-r\sin(\varphi), r\cos(\varphi), 0).$$

Now $\partial f/\partial t$ and $\partial f/\partial \varphi$ are pointwise $\neq 0$ and perpendicular to one another, and are consequently pointwise linearly independent. Therefore, f is an immersion. For all $p \in M$, the vectors $(\partial f/\partial t)(p)$ and $(\partial f/\partial \varphi)(p)$ form a basis of $T_p f$, and their cross product is the generator of the line $N_p f$. The coefficients of the fundamental matrix of the first fundamental form are

$$g_{tt} = \langle \frac{\partial f}{\partial t}, \frac{\partial f}{\partial t} \rangle = \|\dot{c}\|^2, \quad g_{t\varphi} = g_{\varphi t} = \langle \frac{\partial f}{\partial t}, \frac{\partial f}{\partial \varphi} \rangle = 0, \quad g_{\varphi\varphi} = \langle \frac{\partial f}{\partial \varphi}, \frac{\partial f}{\partial \varphi} \rangle = r^2.$$

Therefore, the fundamental matrix

$$\begin{pmatrix} \|\dot{c}\|^2 & 0 \\ 0 & r^2 \end{pmatrix}$$

of the first fundamental form is diagonal. If c is parameterized by arc length, that is, has speed 1, then $g_{tt} = \|\dot{c}\|^2 = 1$.
3) *Generalized helicoids:* Let $a \neq 0$ be a constant and $c = c(t) = (x(t), y(t))$, $t \in I$, be a regular curve in the (x, y)-plane. We call c the *profile curve* of the *generalized helicoid*

$$f(t, \varphi) = (x(t)\cos\varphi - y(t)\sin\varphi, x(t)\sin\varphi + y(t)\cos\varphi, a\varphi).$$

A concrete example in this class is the *helicoid*, where $f(t, \varphi) = (t\cos\varphi, t\sin\varphi, a\varphi)$.
The partial derivatives of f are

$$\frac{\partial f}{\partial t} = (\dot{x}\cos\varphi - \dot{y}\sin\varphi, \dot{x}\sin\varphi + \dot{y}\cos\varphi, 0),$$
$$\frac{\partial f}{\partial \varphi} = (-x\sin\varphi - y\cos\varphi, x\cos\varphi - y\sin\varphi, a).$$

---

[6] Here and in other examples we use the variable names as indices.

These are pointwise linearly independent, since $\partial f/\partial t \neq 0$ and $a \neq 0$. The coefficients of the fundamental matrix of the first fundamental form are therefore

$$g_{tt} = \|\dot{c}\|^2, \quad g_{t\varphi} = g_{\varphi t} = x\dot{y} - \dot{x}y, \quad g_{\varphi\varphi} = \|c\|^2 + a^2.$$

These do not depend on the parameter φ.

4) *Ruled surfaces:* Let $c : I \longrightarrow \mathbb{R}^3$ be a regular space curve and $X : I \longrightarrow \mathbb{R}^3$ be a smooth vector field along c, such that $\dot{c}(t)$ and $X(t)$ are linearly independent for all $t \in I$. The associated *ruled surface* is then

$$f(s, t) = c(t) + sX(t).$$

A concrete example is the helicoid as above with $f(s, t) = (s \cos t, s \sin t, at)$.
Each line $t =$ const is called a *generator*, and each curve $s =$ const a *directrix* of f. The partial derivatives of f are

$$\frac{\partial f}{\partial s} = X \quad \text{and} \quad \frac{\partial f}{\partial t} = \dot{c} + s\dot{X}.$$

Therefore, there is a neighborhood U of $\{s = 0\}$ in $\mathbb{R} \times I$, such that f is an immersion on U. The coefficients of the first fundamental form are

$$g_{ss} = \|X\|^2, \quad g_{st} = g_{ts} = \langle X, \dot{c} + s\dot{X}\rangle, \quad g_{tt} = \langle \dot{c} + s\dot{X}, \dot{c} + s\dot{X}\rangle.$$

5) *Graphs:* Let $W \subset \mathbb{R}^m$ be open and $h : W \longrightarrow \mathbb{R}$ be smooth. Then $f : W \longrightarrow \mathbb{R}^{m+1}$, $f(x) := (x, h(x))$, is an embedding with $\partial f/\partial x^j = (e_j, \partial h/\partial x^j)$, where e_j denotes the j-th standard basis vector of $\mathbb{R}^m$. For each $p \in M$, the tangent space $T_p f$ is the span of the $(\partial f/\partial x^j)(p)$. The coefficients of the first fundamental form with respect to this basis are

$$g_{ij} = \left\langle \frac{\partial f}{\partial x^i}, \frac{\partial f}{\partial x^j} \right\rangle = \delta_{ij} + \frac{\partial h}{\partial x^i}\frac{\partial h}{\partial x^j}.$$

In the case $m = 2$ the normal spaces to f are generated by the $(\partial f/\partial x^1) \times (\partial f/\partial x^2)$.

■

Now let (U, x) again be a chart on M, and let X and Y be vector fields on U. We write X and Y as linear combinations

$$X = \xi^i \frac{\partial}{\partial x^i} \quad \text{and} \quad Y = \eta^i \frac{\partial}{\partial x^i}$$

of the coordinate vector fields. At every $p \in U$, then

$$\langle X(p), Y(p)\rangle = \langle \xi^i(p) \frac{\partial}{\partial x^i}(p), \eta^j(p) \frac{\partial}{\partial x^j}(p)\rangle = g_{ij}(p)\xi^i(p)\eta^j(p). \tag{4.29}$$

For ease of notation, we omit the p, and write (4.29) as an equality of functions,

$$\langle X, Y\rangle = g_{ij}\xi^i\eta^j. \tag{4.30}$$

Since the g_{ij} are smooth, we conclude that the function $\langle X, Y\rangle$ is smooth if X and Y are smooth. We also write

$$g = g_{ij}dx^i dx^j, \tag{4.31}$$

since $dx^i(X) = \xi^i$ and

$$dx^i dx^j(X, Y) := dx^i(X)dx^j(Y) = \xi^i\eta^j. \tag{4.32}$$

We will often need the inverse of the fundamental matrix (g_{ij}). In light of the Einstein summation convention, it is convenient to denote its coefficients by g^{ij}. Then, by definition,

$$g^{ij}g_{jk} = g_{kj}g^{ji} = \delta^i_k. \tag{4.33}$$

The formulas for the inverse matrix from linear algebra show that the g^{ij} are also smooth.

Lemma 4.2.4 *Let $p \in M$ and (U, x) be a chart on M around p. Let $w \in \mathbb{R}^n$ be a vector and $w = w^T + w^N$ be the decomposition of w into tangential and normal components with respect to the decomposition $\mathbb{R}^n = T_p f \oplus N_p f$ as in (4.24). Then there is exactly one tangent vector $v \in T_pM$ with $w^T = df(p)(v)$:*

$$v = \xi^j \frac{\partial}{\partial x^j}(p) \quad \textit{and} \quad w^T = \xi^j \frac{\partial f}{\partial x^j}(p) \quad \textit{with} \quad \xi^j = \langle w, \frac{\partial f}{\partial x^i}(p)\rangle g^{ij}(p).$$

In particular, the tangential and normal components of a smooth vector field along f are smooth.

Proof

The first claim is clear, since $df(p)$ is injective. Now write $w^T = \xi^j(\partial f/\partial x^j)(p)$. Then

$$\langle w, \frac{\partial f}{\partial x^i}(p)\rangle = \langle w^T, \frac{\partial f}{\partial x^i}(p)\rangle = \xi^k\langle\frac{\partial f}{\partial x^k}(p), \frac{\partial f}{\partial x^i}(p)\rangle = \xi^k g_{ki}(p).$$

From this it follows that

$$\xi^j = \xi^k g_{ki}(p)g^{ij}(p) = \langle w, \frac{\partial f}{\partial x^i}(p)\rangle g^{ij}(p). \qquad \square$$

Remark 4.2.5 In the case of a submanifold $M \subseteq \mathbb{R}^n$, the correspondence between the tangential component w^T of w and tangent vector v of M can be seen as the usual identification of the tangent spaces of M with linear subspaces of $\mathbb{R}^n$, and is, in this sense, trivial.

;. 5 86 $ * 3 *

4.2.2 The Internal Distance

The first fundamental form is defined so that, for all $p \in M$, $df(p)\colon T_pM \longrightarrow \mathbb{R}^n$ preserves norms. We define the length and energy of a piecewise smooth curve $c\colon [a,b] \longrightarrow M$ as before,

$$L(c) = \int_a^b \|\dot{c}(t)\|\,dt \quad \text{and} \quad E(c) = \frac{1}{2}\int_a^b \|\dot{c}(t)\|^2 dt. \tag{4.34}$$

Since $d(f \circ c)/dt = df \circ \dot{c}$, we then have $L(c) = L(f \circ c)$ and $E(c) = E(f \circ c)$.

In Corollary 4.1.4 we saw that the distance between points in Euclidean spaces is realized by the lengths of curves. By analogy, we arrive at the *internal metric* or *internal distance* d on M,

$$d(p,q) = \inf L(c), \qquad p, q \in M, \tag{4.35}$$

where the infimum is taken over all piecewise smooth curves c in M running from p to q. The external metric $\|f(p) - f(q)\|$, $p, q \in M$, therefore satisfies $\|f(p) - f(q)\| \leq d(p,q)$ (and is only a true metric if f is injective). The external metric usually requires the use of measurements outside of M (resp. the image of f), as the shortest curves between points in $\mathbb{R}^n$ are lines, which are not typically contained in M (resp. the image of f). See ◘ Fig. 4.3.

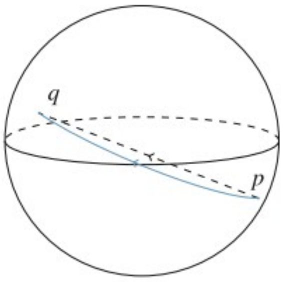

◘ **Fig. 4.3** The internal distance and the Euclidean distance between points on the unit sphere

Remark 4.2.6 We allow d to take the value ∞: $d(p,q) = \infty$ if and only if p and q lie in different connected components of M.

Proposition 4.2.7 *The internal metric on M is a metric on M, and it induces the topology of M.*

Proof

Let $p \in M$ and $x\colon U \longrightarrow U'$ be a chart on M around p with $x(p) = 0$. Choose $\varepsilon > 0$ such that the open Euclidean ball B' of radius ε around 0 is contained in U' and such that, for all $q \in x^{-1}(B')$ and $\xi \in \mathbb{R}^m$,

$$\varepsilon^2 \delta_{ij}\xi^i\xi^j \leq g_{ij}(q)\xi^i\xi^j \leq \varepsilon^{-2}\delta_{ij}\xi^i\xi^j.$$

Now, M is a Hausdorff space. Therefore, every continuous path from p to a point q outside of $B = x^{-1}(B')$ must first run through B to $x^{-1}(\partial B')$. Such a path segment has length $\geq \varepsilon^2$. It therefore follows $d(p, q) > 0$.

For the length of a piecewise smooth curve c in B,

$$\varepsilon L(c) \leq L_E(x \circ c) \leq L(c)/\varepsilon,$$

where $L_E(x \circ c)$ denotes the Euclidean distance of $x \circ c$. This estimate carries over to the distance from p to $q \in B$,

$$\varepsilon d(p, q) \leq \|x(p) - x(q)\| \leq d(p, q)/\varepsilon.$$

It therefore follows that $d(p, q) > 0$ whenever $p \neq q$. Since d is symmetric and satisfies the triangle inequality, it is a metric on M. The claim about the topology follows directly from our estimate of distances in B. □

Remark 4.2.8 Since, in general, f need not be injective, we advise caution in the above proof of the positivity of $d(p, q) > 0$.

Example 4.2.9

We use the argument from the proof of Proposition 4.1.3 to show that the internal distance $d(x, y)$ on the sphere S^m of radius 1 in $\mathbb{R}^{m+1}$ is given by the angle $\angle(x, y)$.

For a given $x \in S^m$, we choose $\varphi = \varphi(y) = \angle(x, y)$ as a replacement for the height function h in the proof of Proposition 4.1.3. Now we can write points $y \in S^m \setminus \{\pm x\}$ uniquely as

$$y = \cos(\varphi)x + \sin(\varphi)z$$

with $z = z(y)$ in the equator of x, that is $\langle x, z\rangle = 0$, $\langle z, z\rangle = 1$, and $\varphi = \varphi(y) \in (0, \pi)$. The function $\varphi \colon S^m \setminus \{\pm x\} \longrightarrow (0, \pi)$ is smooth, with gradient (see Exercise 4.6.11)

$$(\operatorname{grad} \varphi)(\cos(\varphi)x + \sin(\varphi)z) = -\sin(\varphi)x + \cos(\varphi)z.$$

We now show that $d(x, y) \geq \varphi(y)$. It suffices to show that every piecewise smooth curve $c : [a, b] \longrightarrow S^m$ with $c(a) = x$ has length $L(c) \geq \varphi(c(b))$. To this end, we can assume that $a = \sup\{t \in [a, b] \mid c(t) = x\}$ and $b = \inf\{t \in [a, b] \mid \varphi(c(t)) = \varphi(c(b))\}$. For all $t \in (a, b)$, then, $c(t)$ lies in $S^m \setminus \{\pm x\}$. Since $\|\operatorname{grad} \varphi\| = 1$, we obtain

$$L(c) = \int_a^b \|\dot c(t)\|\, dt \geq \int_a^b \langle \operatorname{grad} \varphi(c(t)), \dot c(t)\rangle\, dt = \varphi(y).$$

On the other hand, the great circle

$$c\colon [0, 1] \longrightarrow S^m, \quad c(t) = \cos(t\varphi(y))x + \sin(t\varphi(y))z(y)$$

is smooth, with length $\varphi(y) = \angle(x, y)$. The claim that $d(x, y) = \angle(x, y)$ follows. It also follows that the shortest curves connecting two points are monotone reparameterizations of arcs along great circles. ■

4.2.3 Variations and Geodesics

We now address the question of which conditions a piecewise smooth curve $c\colon [a, b] \longrightarrow M$ must satisfy if it is the shortest curve between its endpoints $p = c(a)$ and $q = c(b)$. We consider the length L as a functional on the space of piecewise smooth curves from p to q.[7] The shortest curves are then those for which L attains a minimum, and so will be critical points of L. To make this notion of critical points of L precise, one considers families of piecewise smooth curves, or so-called *variations*, such that L is differentiable along these families. We have already encountered variations in a special case. Compare (4.11) and (4.12).

Definition 4.2.10

A *variation* of c is a map

$$h\colon (-\varepsilon, \varepsilon) \times [a, b] \longrightarrow M \quad \text{with} \quad h(0, t) = c(t),$$

such that there is a subdivision $a = t_0 < t_1 < \cdots < t_k = b$ of $[a, b]$ such that the restriction of h to $(-\varepsilon, \varepsilon) \times [t_{i-1}, t_i]$ is smooth for all $1 \leq i \leq k$.
The variation consists of the piecewise smooth curves $c_s := h(s, .)$. We call the piecewise smooth vector field $V = V(t) = (\partial_s h)(0, t)$ along c the *variation field* of h. Instead of h, we sometimes write (c_s). We call a variation $h = (c_s)$ of $c = c_0$ *proper*, if $c_s(a) = c(a)$ and $c_s(b) = c(b)$ for all $s \in (-\varepsilon, \varepsilon)$. See ◘ Fig. 4.4.

◘ **Fig. 4.4** A proper variation

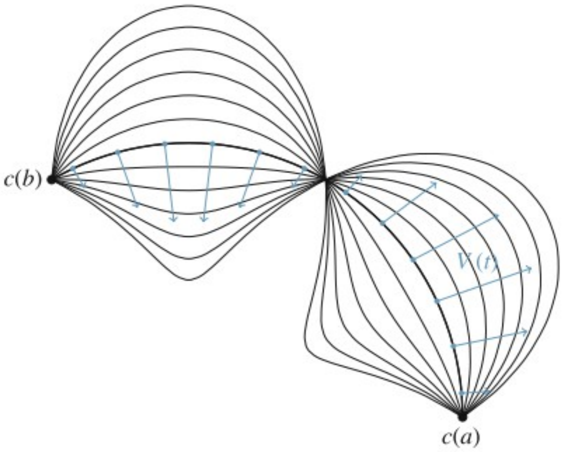

---

[7] Since the domain of L is a space of functions, one does not simply call L a function, but rather uses the hifalutin term *functional*, which does not, however, have a deeper meaning.

Let $c\colon [a,b] \longrightarrow M$ be a piecewise smooth curve. As in the smooth case, we call c *regular* if $\dot{c}(t) \neq 0$ for all $t \in [a,b]$. Now let c be regular and let $h = (c_s)$ be a variation of $c = c_0$. Then the c_s are also regular if we only consider sufficiently small s. By definition of the first fundamental form,

$$\|\dot{c}_s(t)\| = \|\partial_t(f \circ h)(s,t)\|.$$

With a subdivision as in Definition 4.2.10 it therefore follows that $\|\dot{c}_s(t)\|$ is smooth on the rectangles $(-\varepsilon,\varepsilon) \times [t_{i-1},t_i]$, $1 \leq i \leq k$. From this it follows that the lengths $L(c_s)$ depend smoothly on s.

Definition 4.2.11

Let $c\colon [a,b] \longrightarrow M$ be regular and let $h = (c_s)$ be a variation of $c = c_0$. Then we call the derivative δL of $L(c_s)$ by s at $s = 0$ the *the first variation of arc length* of (c_s).

Proposition 4.2.12 *Let $c\colon [a,b] \longrightarrow M$ be piecewise smooth with constant speed $T \neq 0$. Then for a variation $h = (c_s)$ of $c = c_0$ with variation field V,*

$$\begin{aligned}
\delta L &= \frac{1}{T}\left[\sum_{i=1}^{k} \big\langle V, \dot{c}\big\rangle\Big|_{t_{i-1}}^{t_i} - \int_a^b \Big\langle df \circ V, \frac{d^2(f \circ c)}{dt^2}\Big\rangle\, dt\right] \\
&= \frac{1}{T}\left[\big\langle V, \dot{c}\big\rangle\Big|_a^b + \sum_{i=1}^{k-1} \big\langle V(t_i), \Delta_i\big\rangle - \int_a^b \Big\langle df \circ V, \frac{d^2(f \circ c)}{dt^2}\Big\rangle\, dt\right]
\end{aligned}$$

with $\Delta_i := \dot{c}(t_i-) - \dot{c}(t_i+)$, $1 \leq i \leq k-1$.

Proof

We may assume that $\dot{c}_s(t) \neq 0$ for all s and t. Then $\|\dot{c}_s(t)\|$ is smooth on the rectangles $(-\varepsilon,\varepsilon) \times [t_{i-1},t_i]$ as in Definition 4.2.10. Therefore,

$$\begin{aligned}
\frac{d(L(c_s))}{ds} &= \frac{d}{ds}\Big(\int_a^b \|\dot{c}_s(t)\|\, dt\Big) \\
&= \int_a^b \frac{d}{ds}\sqrt{\langle \partial_t(f \circ h), \partial_t(f \circ h)\rangle}\, dt.
\end{aligned}$$

With the requirement that $\|\dot{c}(t)\| = T = \text{const}$, we thereby obtain

$$\begin{aligned}
\delta L &= \frac{1}{T}\int_a^b \big\langle \partial_s\partial_t(f \circ h), \partial_t(f \circ h)\big\rangle(0,t)\, dt \\
&= \frac{1}{T}\int_a^b \big\langle \partial_t\partial_s(f \circ h), \partial_t(f \circ h)\big\rangle(0,t)\, dt
\end{aligned}$$

$$= \frac{1}{T}\int_a^b \partial_t\big\langle \partial_s(f\circ h), \partial_t(f\circ h)\big\rangle(0,t)\,dt - \frac{1}{T}\int_a^b \big\langle \partial_s(f\circ h), \partial_t^2(f\circ h)\big\rangle(0,t)\,dt$$

$$= \frac{1}{T}\sum_{i=1}^{k}\big\langle \partial_s(f\circ h), \partial_t(f\circ h)\big\rangle(0,t)\Big|_{t_{i-1}}^{t_i} - \frac{1}{T}\int_a^b \big\langle \partial_s(f\circ h), \partial_t^2(f\circ h)\big\rangle(0,t)\,dt.$$

However, at $s = 0$, by definition,

$$\partial_s(f\circ h) = df\circ V, \quad \partial_t(f\circ h) = df\circ\dot{c} \quad \text{and} \quad \partial_t^2(f\circ h) = \frac{d^2(f\circ c)}{dt^2}. \qquad \square$$

We now come to a result of Johann Bernoulli[8] from 1698 (unpublished), one of the first results ever in differential geometry; compare with [HT, pp. 117].

Proposition 4.2.13 *A piecewise smooth curve $c\colon [a,b] \longrightarrow M$ has constant speed T and first variation of arc length $\delta L = 0$ for every proper variation of c if and only if c is smooth and the tangential component of the second derivative of the second derivative of $f\circ c$ vanishes.*

Proof

We first assume that c is smooth and the tangential component of the second derivative of $f\circ c$ vanishes. By the definition of the first fundamental form

$$\frac{d}{dt}\langle\dot{c},\dot{c}\rangle = \frac{d}{dt}\Big\langle\frac{d(f\circ c)}{dt}, \frac{d(f\circ c)}{dt}\Big\rangle = 2\Big\langle\frac{d(f\circ c)}{dt}, \frac{d^2(f\circ c)}{dt^2}\Big\rangle = 0,$$

since $d(f\circ c)/dt$ is tangent to f. Therefore c has constant speed T.

We may now further assume that c is not constant, that is that $T > 0$. Let $h = (c_s)$ be a proper variation of $c = c_0$ with variation field V.

Since h is proper and c is smooth, it follows that $V(a) = 0$, $V(b) = 0$, and $\Delta_i = 0$, $1 \le i \le k-1$. Therefore, the first terms in the formula for δL vanish. The integrand of the integral now vanishes pointwise, since $df\circ V$ is tangential and $d^2(f\circ c)/dt^2$ is normal to f. Therefore $\delta L = 0$.

In the opposite direction, we can assume that c is not constant. Then let $a = t_0 < t_1 < \cdots < t_k = b$ be a subdivision, such that c is smooth on the intervals $[t_{i-1}, t_i]$. We first show, that the tangential component of the second derivative of $f\circ c$ vanishes on the intervals $[t_{i-1}, t_i]$. Since, by Lemma 4.2.4, the tangential component is smooth on these intervals, it suffices to show that it vanishes on the open intervals (t_{i-1}, t_i). Assume that this is not the case for some $t' \in (t_{i-1}, t_i)$. Then there is a vector $v \in T_{c(t')}M$ with

$$\Big\langle df(c(t'))(v), \frac{d^2(f\circ c)}{dt^2}(t')\Big\rangle > 0.$$

---

[8] Johann Bernoulli (1667–1748). He, of course, only discussed curves on surfaces in $\mathbb{R}^3$, but this amounts to much the same thing in proofs.

Now let (U, x) be a chart on M around $c(t')$, and let $\xi \in \mathbb{R}^m$ be the principal part of v with respect to x. Then

$$v = \xi^i \frac{\partial}{\partial x^i}\Big|_{c(t')} \quad \text{and} \quad df(c(t')) \cdot v = \xi^i \frac{\partial f}{\partial x^i}(c(t')).$$

Since the derivatives of f and c are continuous, there is an $\varepsilon > 0$ with $(t' - \varepsilon, t' + \varepsilon) \subseteq (t_{i-1}, t_i)$, such that $c((t' - \varepsilon, t' + \varepsilon)) \subseteq U$ and

$$\Big\langle \xi^i \frac{\partial f}{\partial x^i}(c(t)), \frac{d^2(f \circ c)}{dt^2}(t) \Big\rangle > 0$$

for all $t \in (t' - \varepsilon, t' + \varepsilon)$. Now let $\varphi : \mathbb{R} \longrightarrow \mathbb{R}$ be a bump function with

$$0 \le \varphi \le 1, \quad \operatorname{supp} \varphi \subseteq (t' - \varepsilon, t' + \varepsilon), \quad \text{and} \quad \varphi(t') = 1.$$

Set

$$h(s, t) = \begin{cases} x^{-1}\big(x(c(t)) + s\,\varphi(t)\,\xi\big) & \text{for } t \in (t' - \varepsilon, t' + \varepsilon), \\ c(t) & \text{otherwise}. \end{cases}$$

Then h is a proper variation of c. The variation field V of h satisfies

$$df(c(t))(V(t)) = \varphi(t)\,\xi^i \frac{\partial f}{\partial x^i}(c(t))$$

for all $t \in (t' - \varepsilon, t' + \varepsilon)$, and $V(t) = 0$ otherwise. However, the choice of v or ξ respectively implies $\delta L \neq 0$, which is a contradiction. Therefore, the tangential component of the second derivative of $f \circ c$ vanishes on the intervals $[t_{i-1}, t_i]$.

We next show that c is continuously differentiable, that is, that $\dot{c}(t_i-) = \dot{c}(t_i+)$ for all $0 < i < k$. To find a contradiction, we assume that $\Delta_i = \dot{c}(t_i-) - \dot{c}(t_i+) \neq 0$ for some i. We again choose a chart (U, x) on M, this time about $c(t_i)$, and a bump function φ as above, but now with support in $(t_i - \varepsilon, t_i + \varepsilon) \subseteq (t_{i-1}, t_{i+1})$ and with $\varphi(t_i) = 1$, such that $c((t_i - \varepsilon, t_i + \varepsilon)) \subseteq U$, and define a proper variation h of c as above, where ξ^i are the coefficients of Δ_i with respect to x. Since the tangential derivatives of $f \circ c$ are normal to f, it follows that $\delta L = \langle V(t_i), \Delta_i \rangle = \|\Delta_i\|^2 \neq 0$, a contradiction. So c is continuously differentiable.

It now only remains to show that c is smooth. For this, we refer to the discussion below: By the choice of t_i, c is already smooth on the intervals $[t_{i-1}, t_i]$. Now let $0 < i < k$ and (U, x) be a chart around $c(t_i)$. Let $\varepsilon > 0$ be chosen so that $c((t_i - \varepsilon, t_i + \varepsilon))$ is contained in U. Since the tangential component of the second derivative of $f \circ c$ vanishes on the intervals $(t_i - \varepsilon, t_i]$ and $[t_i, t_i + \varepsilon)$, the $c^j := x^j \circ c$ satisfy the second-order ordinary differential equations (4.40) on these intervals. Now, the values and the derivatives of the $c^j|_{(t_i - \varepsilon, t_i]}$ and $c^j|_{[t_i, t_i + \varepsilon)}$ agree at $t = t_i$, since c is continuously differentiable. Thus the c^j solve the equations in (4.40) on $(t_i - \varepsilon, t_i + \varepsilon)$. Since the coefficients Γ_{ij}^k of the equation are smooth, the c^j are smooth on these intervals. Therefore, c is smooth on $(t_i - \varepsilon, t_i + \varepsilon)$ and therefore on $[a, b]$. □

Definition 4.2.14

We call a smooth curve $c\colon I \longrightarrow M$ a *geodesic*, if the tangential component of the second derivative of $f \circ c$ vanishes,

$$\left[\frac{d^2(f \circ c)}{dt^2}\right]^T = 0.$$

Example 4.2.15

Let $p, q \in M$. Suppose there is a curve $c_0\colon [a, b] \longrightarrow M$ such that $f \circ c_0$ is the constant-speed line segment in $\mathbb{R}^n$ from $f(p)$ to $f(q)$. Then c_0 is smooth and has constant speed $\|f(q) - f(p)\|/|b - a|$. The second derivative $f \circ c_0$ vanishes, so c_0 is a geodesic.

$f \circ c_0$ is the shortest among all the piecewise smooth curves in $\mathbb{R}^n$ from $f(p)$ to $f(q)$, and therefore, *a fortiori*, is the shortest among all curves of the form $f \circ c$. Thus, c_0 is a shortest curve connecting p to q in M. The first variation of arc length of every proper variation of c_0 must therefore vanish, so we again see that c_0 is a geodesic.

For the ruled surfaces of Example 4.2.3.4, the generators $t = \text{const}$ are lines in $\mathbb{R}^3$ and therefore are geodesics in the surface. ■

4.2.4 Covariant Derivative and Geodesics

The definitions of the length of piecewise smooth curves in M and of the geodesics as critical points of the length functional only involve the first fundamental form on M (resp. f). We therefore seek a formula for geodesics that only requires the first fundamental form. It is even advisable to aim higher. Let $X\colon I \longrightarrow TM$ be a smooth *vector field along* c, , i.e., $X(t) \in T_{c(t)}M$ for all $t \in I$. Then $Xf = df \circ X$ is a smooth tangent vector field along $f \circ c$. In general, $d(Xf)/dt$ is no longer tangent to f. The tangential part of $d(Xf)/dt$ is our focus. See also Lemma 4.2.4.

Definition 4.2.16

We call the unique vector field $X'\colon I \longrightarrow TM$ along c with

$$df \circ X' = [d(Xf)/dt]^T$$

the *covariant derivative* of X, and we sometimes write $\nabla X/dt$ instead of X'. We call X *parallel* (along c) if $X' = 0$.

Geodesics are therefore, by Definition 4.2.14, characterized by the condition that the covariant derivative $\nabla\dot{c}/dt$ vanishes, or, in other words, that $\dot{c}$ is parallel along c.

Remark 4.2.17

1) To the best of my knowledge, we have Levi-Civita[9] to thank for the insight that the orthogonal projection onto the tangential component as in Definition 4.2.16 leads to a sensible kind of derivative: the covariant derivative.
2) For a vector field X along a curve c in M, X' is once again a vector field along c, while the usual derivative $\dot{X}$ of X takes values in TTM. The second derivative $\ddot{X}$ then has values in $TTTM$, and so forth. In contrast, higher covariant derivatives of X always remain vector fields along c.
3) We have already encountered the strategy of considering orthogonal projections of usual derivatives in our discussion of the normal fields of curves; see Definition 4.1.13 and the discussion which follows it.

We now discuss the computation of covariant derivatives in terms of local coordinates. Let (U, x) be a chart on M and $c\colon I \longrightarrow U$ a smooth curve. Let $X\colon I \longrightarrow M$ be a smooth vector field along c. Then

$$X = \xi^i \frac{\partial}{\partial x^i} \quad \text{and} \quad Xf = df \circ X = \xi^i \frac{\partial f}{\partial x^i} \quad \text{with} \quad \xi^j = \left\langle X, \frac{\partial}{\partial x^i}\right\rangle g^{ij},$$

where these equations are to be read as equations along c. With $c^i := x^i \circ c$ and since $f \circ c = (f \circ x^{-1}) \circ (x \circ c)$, we therefore obtain

$$\frac{d(Xf)}{dt} = \frac{d}{dt}\big(\xi^i \frac{\partial f}{\partial x^i}\big) = \dot{\xi}^i \frac{\partial f}{\partial x^i} + \dot{c}^j \xi^i \frac{\partial^2 f}{\partial x^j \partial x^i}. \tag{4.36}$$

Since the partial derivatives $\partial f/\partial x^i$ are tangent to f, the first term is already tangent to f. The second term is a linear combination of the second partial derivatives of f with respect to x, and is not, in general, tangent to f. The tangential components of the second partial derivatives are, however, pointwise linear combinations of the first partial derivatives, which are bases of their respective tangent spaces $T_p f$,

$$\left[\frac{\partial^2 f}{\partial x^i \partial x^j}\right]^T =: \Gamma_{ij}^k \frac{\partial f}{\partial x^k}. \tag{4.37}$$

The coefficients $\Gamma_{ij}^k : U \longrightarrow \mathbb{R}$ are called *Christoffel symbols*.[10] Using Definition 4.2.16 and (4.36), we can now compute the covariant derivative of X to be

$$\frac{\nabla X}{dt} = X' = \big(\dot{\xi}^k + \Gamma_{ij}^k \dot{c}^i \xi^j\big) \frac{\partial}{\partial x^k}, \tag{4.38}$$

---

[9] Tullio Levi-Civita (1873–1941).

[10] Elwin Bruno Christoffel (1829–1900).

where Γ_{ij}^k, strictly speaking, should be read as $\Gamma_{ij}^k \circ c$. By Lemma 4.2.4, or by Proposition 4.2.23 below, the Christoffel symbols are smooth. Therefore, the covariant derivative of X is also smooth. Since the second derivatives $\partial^2 f/\partial x^i \partial x^j$ are symmetric in i and j, the Christoffel symbols are symmetric in their two lower indices, $\Gamma_{ij}^k = \Gamma_{ji}^k$.

Proposition 4.2.18 *Let X and Y be smooth vector fields along c. Then the following hold:*

1. *(Linearity) For $\alpha, \beta \in \mathbb{R}$, $(\alpha X + \beta Y)' = \alpha X' + \beta Y'$.*
2. *(Product rule 1) For smooth $\varphi \colon I \longrightarrow \mathbb{R}$, $(\varphi X)' = \dot{\varphi} X + \varphi X'$.*
3. *(Product rule 2) The function $\langle X, Y \rangle$ has derivative $\langle X', Y \rangle + \langle X, Y' \rangle$.*

Proof

Claims 1 and 2 follow immediately from Definition 4.2.16 or from (4.38). By the definition of the first fundamental form $\langle X, Y \rangle = \langle Xf, Yf \rangle$, and therefore

$$\begin{aligned} d\langle X, Y \rangle/dt &= d\langle Xf, Yf \rangle/dt = \langle d(Xf)/dt, Yf \rangle + \langle Xf, d(Yf)/dt \rangle \\ &= \langle df \circ X', df \circ Y \rangle + \langle df \circ X, df \circ Y' \rangle = \langle X', Y \rangle + \langle X, Y' \rangle. \end{aligned}$$

In passing from the first line to the second, we have used that, in the first term, the normal part of $d(Xf)/dt$ is by definition perpendicular to f and therefore is perpendicular to $Yf = df \circ Y$. We reason analogously for the second term. □

Corollary 4.2.19 *For parallel vector fields X and Y along c the following hold:*

1. *Linear combinations $\alpha X + \beta Y$ are themselves parallel along c.*
2. *The function $\langle X, Y \rangle$ is constant.* □

Let $W \subseteq \mathbb{R}^2 = \{(s, t) \mid s, t \in \mathbb{R}\}$ be an open subset, and let $\varphi : W \longrightarrow M$ be smooth. Then the partial derivatives $\partial\varphi/\partial s$ and $\partial\varphi/\partial t$ are vector fields along the s- and t-coordinate lines. We can therefore consider their covariant derivatives along these curves.

Proposition 4.2.20 *Let W be an open subset of the (s, t)-plane and let $\varphi : W \longrightarrow M$ be smooth. Then*

$$\frac{\nabla}{\partial s}\frac{\partial\varphi}{\partial t} = \frac{\nabla}{\partial t}\frac{\partial\varphi}{\partial s}.$$

Proof

By definition,

$$\frac{\nabla}{\partial s}\frac{\partial\varphi}{\partial t} = \left[\frac{\partial^2\varphi}{\partial s \partial t}\right]^T = \left[\frac{\partial^2\varphi}{\partial t \partial s}\right]^T = \frac{\nabla}{\partial t}\frac{\partial\varphi}{\partial s}.$$

□

With respect to local coordinates x and using (4.38), a vector field X along a smooth curve c is parallel, in the sense of Definition 4.2.16, if the coefficients ξ^i of X with respect to x solve the first-order ordinary differential equation

$$\dot{\xi}^k + \Gamma_{ij}^k \dot{c}^i \xi^j = 0. \tag{4.39}$$

The equation is linear in the ξ^i, so maximal solutions are defined on the entire domain of definition of the differential equation.

Corollary 4.2.21 *Let $c : I \longrightarrow M$ be a smooth curve and $t_0 \in I$. Then the following hold:*

1. *For $v \in T_{c(t_0)}M$ there is precisely one parallel vector field X along c with $X(t_0) = v$.*
2. *For a basis $(v_1, \ldots, v_m)$ of $T_{c(t_0)}M$ and parallel vector fields X_i along c with $X_i(t_0) = v_i$, $(X_1(t), \ldots, X_m(t))$ is a basis of $T_{c(t)}M$ for all $t \in I$. Furthermore, $\langle X_i(t), X_j(t)\rangle = \langle v_i, v_j\rangle$.* □

By definition, geodesics are smooth curves c with $\nabla \dot{c}/dt = 0$. Therefore, by (4.39), they are characterized as solutions to the second-order differential equation

$$\ddot{c}^k + \Gamma_{ij}^k \dot{c}^i \dot{c}^j = 0. \tag{4.40}$$

In contrast to (4.39), however, this equation is not linear in its solutions. Nonetheless, we can extract some important consequences from the fact that geodesics are solutions to an ordinary differential equation:

Corollary 4.2.22

1. *If $c_1 \colon I_1 \longrightarrow M$ and $c_2 \colon I_2 \longrightarrow M$ are geodesics with $I_1 \cap I_2 \neq \emptyset$ and $c_1|_{I_1 \cap I_2} = c_2|_{I_1 \cap I_2}$, then the concatenation $c \colon I_1 \cup I_2 \longrightarrow M$ of c_1 and c_2 is also a geodesic.*
2. *If c is a geodesic, so is $\tilde{c} = \tilde{c}(t) = c(at + b)$ for all $a, b \in \mathbb{R}$.*
3. *For $t \in \mathbb{R}$, $p \in M$, and $v \in T_pM$ there is precisely one maximal geodesic $c \colon I \longrightarrow M$ with $c(t) = p$ and $\dot{c}(t) = v$.* □

By *maximal*, we mean that the domain of every other geodesic with the given initial conditions is contained in I. In particular, I is open, since in local coordinates geodesics are solutions of the differential equation (4.40).

Proposition 4.2.23 *The Christoffel symbols associated to a chart x on M can be calculated from the coefficients of the first fundamental form in x. More explicitly,*

$$\Gamma_{ij}^l = \frac{1}{2} g^{kl} \Big(\frac{\partial g_{jk}}{\partial x^i} + \frac{\partial g_{ik}}{\partial x^j} - \frac{\partial g_{ij}}{\partial x^k} \Big).$$

Proof

We have

$$\frac{\partial g_{ij}}{\partial x^k} = \frac{\partial}{\partial x^k} \Big\langle \frac{\partial f}{\partial x^i}, \frac{\partial f}{\partial x^j} \Big\rangle = \Big\langle \frac{\partial^2 f}{\partial x^k \partial x^i}, \frac{\partial f}{\partial x^j} \Big\rangle + \Big\langle \frac{\partial f}{\partial x^i}, \frac{\partial^2 f}{\partial x^k \partial x^j} \Big\rangle.$$

Now $\partial f/\partial x^j$ and $\partial f/\partial x^i$ are tangent to f, so, on the right-hand side, only the tangential components of the second partial derivatives of f contribute. With (4.37), we then obtain

$$\frac{\partial g_{ij}}{\partial x^k} = \Big\langle \Gamma^l_{ki}\frac{\partial f}{\partial x^l}, \frac{\partial f}{\partial x^j}\Big\rangle + \Big\langle \frac{\partial f}{\partial x^i}, \Gamma^l_{kj}\frac{\partial f}{\partial x^l}\Big\rangle = \Gamma^l_{ki}g_{lj} + \Gamma^l_{kj}g_{il},$$

and similarly for $\partial g_{jk}/\partial x^i$ and $\partial g_{ik}/\partial x^j$. With the aid of the symmetry of the lower indices of the Christoffel symbols, we therefore obtain

$$\frac{\partial g_{jk}}{\partial x^i} + \frac{\partial g_{ik}}{\partial x^j} - \frac{\partial g_{ij}}{\partial x^k} = 2\,\Gamma^l_{ij}g_{lk}.$$

Finally, therefore

$$\begin{aligned} 2\,\Gamma^l_{ij} &= 2\,\Gamma^\mu_{ij}\,\delta^l_\mu = 2\,\Gamma^\mu_{ij}(g_{\mu k}g^{kl}) \\ &= 2\left(\Gamma^\mu_{ij}g_{\mu k}\right)g^{kl} = g^{kl}\Big(\frac{\partial g_{jk}}{\partial x^i} + \frac{\partial g_{ik}}{\partial x^j} - \frac{\partial g_{ij}}{\partial x^k}\Big). \qquad \square \end{aligned}$$

We see that the Christoffel symbols can be explicitly computed from the coefficients of the first fundamental form and their derivatives. Therefore, the Christoffel symbols are quantities belonging to interior geometry! With (4.39) and (4.40), geodesics and parallel vector fields are therefore also explicitly characterized as objects from interior geometry.

4.3 Exterior Geometry

The tangent space $T_p f$ approximates the image of f around p to first order, but does not describe, how that image bends in the ambient space. As in the case of curves, we therefore consider approximations of second order.

For $p \in M$, we consider the orthogonal decomposition $\mathbb{R}^n = T_p f \oplus N_p f$ with the corresponding projections π^T_p and π^N_p. Compare with (4.24) and (4.26). Then for the smooth map

$$x\colon M \longrightarrow T_p f, \quad x(q) = \pi^T_p(f(q) - f(p)), \tag{4.41}$$

we have $x(p) = 0$, and, since π^T_p is linear, $dx(p) = \pi^T_p \circ df(p) = df(p)$. Therefore $dx(p) : T_pM \longrightarrow T_p f$ is an isomorphism. The Inverse Function Theorem thus implies that there are open neighborhoods U of p in M and U' of 0 in $T_p f$, such that $x\colon U \longrightarrow U'$ is a diffeomorphism with $x(p) = 0$ and $dx(p) = df(p)$. Up to the choice of an isomorphism $T_p f \simeq \mathbb{R}^m$, (U, x) is thus the chart on M about p which best fits the region of M around p in the ambient space $\mathbb{R}^n$.

Proposition 4.3.1 (Local Normal Form) *Let* $h = \pi^N_p \circ (f - f(p)) \circ x^{-1} : U' \longrightarrow N_p f$ *be the* $N_p f$ *component of* $(f - f(p)) \circ x^{-1}$. *Then* $h(0) = 0$, $dh(0) = 0$, *and*

$$(f \circ x^{-1})(u) = f(p) + u + h(u) \quad \textit{for all } u \in U'.$$

Now $\partial f/\partial x^j$ and $\partial f/\partial x^i$ are tangent to f, so, on the right-hand side, only the tangential components of the second partial derivatives of f contribute. With (4.37), we then obtain

$$\frac{\partial g_{ij}}{\partial x^k} = \left\langle \Gamma^l_{ki}\frac{\partial f}{\partial x^l}, \frac{\partial f}{\partial x^j}\right\rangle + \left\langle \frac{\partial f}{\partial x^i}, \Gamma^l_{kj}\frac{\partial f}{\partial x^l}\right\rangle = \Gamma^l_{ki}g_{lj} + \Gamma^l_{kj}g_{il},$$

and similarly for $\partial g_{jk}/\partial x^i$ and $\partial g_{ik}/\partial x^j$. With the aid of the symmetry of the lower indices of the Christoffel symbols, we therefore obtain

$$\frac{\partial g_{jk}}{\partial x^i} + \frac{\partial g_{ik}}{\partial x^j} - \frac{\partial g_{ij}}{\partial x^k} = 2\,\Gamma^l_{ij}g_{lk}.$$

Finally, therefore

$$\begin{aligned} 2\,\Gamma^l_{ij} &= 2\,\Gamma^\mu_{ij}\,\delta^l_\mu = 2\,\Gamma^\mu_{ij}\,(g_{\mu k}g^{kl}) \\ &= 2\left(\Gamma^\mu_{ij}g_{\mu k}\right)g^{kl} = g^{kl}\left(\frac{\partial g_{jk}}{\partial x^i} + \frac{\partial g_{ik}}{\partial x^j} - \frac{\partial g_{ij}}{\partial x^k}\right). \end{aligned} \qquad \square$$

We see that the Christoffel symbols can be explicitly computed from the coefficients of the first fundamental form and their derivatives. Therefore, the Christoffel symbols are quantities belonging to interior geometry! With (4.39) and (4.40), geodesics and parallel vector fields are therefore also explicitly characterized as objects from interior geometry.

4.3 Exterior Geometry

The tangent space $T_p f$ approximates the image of f around p to first order, but does not describe, how that image bends in the ambient space. As in the case of curves, we therefore consider approximations of second order.

For $p \in M$, we consider the orthogonal decomposition $\mathbb{R}^n = T_p f \oplus N_p f$ with the corresponding projections π^T_p and π^N_p. Compare with (4.24) and (4.26). Then for the smooth map

$$x\colon M \longrightarrow T_p f, \quad x(q) = \pi^T_p(f(q) - f(p)), \tag{4.41}$$

we have $x(p) = 0$, and, since π^T_p is linear, $dx(p) = \pi^T_p \circ df(p) = df(p)$. Therefore $dx(p) : T_p M \longrightarrow T_p f$ is an isomorphism. The Inverse Function Theorem thus implies that there are open neighborhoods U of p in M and U' of 0 in $T_p f$, such that $x\colon U \longrightarrow U'$ is a diffeomorphism with $x(p) = 0$ and $dx(p) = df(p)$. Up to the choice of an isomorphism $T_p f \simeq \mathbb{R}^m$, (U, x) is thus the chart on M about p which best fits the region of M around p in the ambient space $\mathbb{R}^n$.

Proposition 4.3.1 (Local Normal Form) *Let $h = \pi^N_p \circ (f - f(p)) \circ x^{-1} : U' \longrightarrow N_p f$ be the $N_p f$ component of $(f - f(p)) \circ x^{-1}$. Then $h(0) = 0$, $dh(0) = 0$, and*

$$(f \circ x^{-1})(u) = f(p) + u + h(u) \quad \textit{for all } u \in U'.$$

Proof

Since $x^{-1}(0) = p$, $h(0) = 0$. Furthermore,

$$dh(0) = \pi_p^N \circ df(p) \circ dx^{-1}(0).$$

Now $\operatorname{im} df(p) = T_p f$, so $\pi_p^N \circ df(p) = 0$. The rest is clear. □

By Proposition 4.3.1, the Taylor expansion of h around 0 is given by

$$h(u) = \frac{1}{2} H_p(u, u) + \text{terms of third or higher order in } u, \tag{4.42}$$

where $H_p := D^2 h|_0$. Up to translation by $f(p)$ and terms of third or higher order in u,

$$Q_p = \{u + v \in \mathbb{R}^n \mid u \in T_p f,\ v = H_p(u, u)/2 \in N_p f\} \tag{4.43}$$

therefore describes the image of f around p. We call Q_p the *osculating paraboloid* to f at p. The osculating paraboloid to f at p is an m-dimensional submanifold of $\mathbb{R}^n$ and describes $f - f(p)$ about p to second order.

With the chart (U, x) on M around p as above, we now let $c\colon I \longrightarrow U$ be a smooth curve with $c(t_0) = p$ for some $t_0 \in I$. Then, by Proposition 4.3.1,

$$f \circ c = (f \circ x^{-1}) \circ (x \circ c) = f(p) + x \circ c + h \circ (x \circ c).$$

It therefore follows from (4.42) that

$$\begin{aligned}
\frac{d^2(f \circ c)}{dt^2}(t_0) &= \frac{d^2(x \circ c + h \circ (x \circ c))}{dt^2}(t_0) \\
&= \frac{d^2(x \circ c)}{dt^2}(t_0) + H_p\big(dx(p)(\dot{c}(t_0)), dx(p)(\dot{c}(t_0))\big) \\
&= \frac{d^2(x \circ c)}{dt^2}(t_0) + H_p\big(df(p)(\dot{c}(t_0)), df(p)(\dot{c}(t_0))\big).
\end{aligned}$$

The first term on the right is tangent to f at p, the second is normal. Therefore, the right-hand side is an orthogonal decomposition of the second derivative of $f \circ c$ at $t = t_0$. In this section, the second, normal, term is our focus.

Definition 4.3.2

We call the symmetric bilinear form

$$S_p : T_p M \times T_p M \longrightarrow N_p f, \quad S_p(v, w) := H_p(df(p)(v), df(p)(w)),$$

the *second fundamental form* of f at p.

We can summarize the computations above in the following equation:

$$\frac{d^2(f \circ c)}{dt^2} = df \circ \frac{\nabla \dot c}{dt} + S(\dot c, \dot c), \tag{4.44}$$

where $\nabla \dot c/dt$ denotes the covariant derivative of $\dot c$, see Definition 4.2.16. The reference points have been omitted in (4.44) for ease of reading. The first term is tangent to f, the second is normal. By (4.44), we can determine the second fundamental form S of f by polarization, without computing the special charts x or the local normal form of f as in Proposition 4.3.1.

Remark 4.3.3 In Definition 4.3.2 and (4.44) we again encounter the natural identification $df(p)\colon T_pM \longrightarrow T_p f$. In the case of a submanifold, this is the usual identification of T_pM with a subspace of $\mathbb{R}^n$ and will therefore be omitted from our notation. For submanifolds, the above formulas simplify to

$$S_p = H_p \quad \text{and} \quad \ddot c = \frac{\nabla \dot c}{dt} + S(\dot c, \dot c), \tag{4.45}$$

where $\nabla \dot c/dt$ is the tangential part of $\ddot c$, and $S(\dot c, \dot c)$ is the normal part.

Example 4.3.4

1) Let $M = I$ be an open interval and $c\colon I \longrightarrow \mathbb{R}^n$ be a regular curve, that is, an immersion. Then for $t \in I$, $T_t c = \mathbb{R}\cdot e(t)$ with $e(t) = \dot c(t)/\|\dot c(t)\|$ as in (4.1). The second fundamental form associates to $\partial/\partial t$ the normal component of $\ddot c$ (at the corresponding point), that is, $S(\partial/\partial t, \partial/\partial t) = \ddot c - \langle \ddot c, e\rangle e$. The length of these vectors is $\kappa \cdot \|\dot c\|^2$. Compare with (4.4).
2) Let $M = S^m_r \subseteq \mathbb{R}^{m+1}$ be the sphere with center 0 and radius r (and f be the inclusion). For $x \in S^m_r$, let $U = \{y \in S^m_r \mid \langle x, y\rangle > 0\}$. Then the orthogonal projection π^T_x : $U \longrightarrow T_x S^m_r = \{u \in \mathbb{R}^{m+1} \mid \langle x, u\rangle = 0\}$ is a diffeomorphism onto its image U' and is therefore a chart on S^m_r around x. Up to conflicts in the notation, it corresponds to the charts in (4.41) and Proposition 4.3.1, there denoted by x. As in Proposition 4.3.1 (for the case $f =$ inclusion), we can now write U as the graph of the map

$$h\colon U' \longrightarrow \mathbb{R}\,x = N_x S^m_r, \quad h(u) = \left(\sqrt{r^2 - \|u\|^2} - r\right)\frac{x}{r}.$$

Therefore, the second fundamental form of $M = S^m_r$ at x is given by

$$S_x(v, w) = -\frac{\langle v, w\rangle}{r^2} \cdot x.$$

With (4.44), we can determine $S_x(v, v)$ for a unit vector $v \in T_x S^n_r$ by determining the normal portion of the second derivative of a smooth curve c through x with derivative $\dot c(0) = v$. Here, the curve $c = c(t) = \cos(t)x + r\sin(t)v$ recommends itself, since $\ddot c = -c$ is normal to the sphere at all points, and therefore c is a geodesic. It thereby follows that $S_x(rv, rv) = -x$. Polarization then leads to the formula for S_x which we obtained in a different way above.

■

Proposition 4.3.5 *For smooth vector fields X, Y on M,*

$$S(X, Y) = [XYf]^N.$$

Proof

Let $p \in M$ with the chart x around p as in (4.41) and the map h as in Proposition 4.3.1. Since $dh(0) = 0$, it then follows that

$$\begin{aligned}[XYf]^N(p) &= [X(df \circ Y)]^N(p) = X(d(h \circ x) \circ Y)(p) \\ &= D^2h(0)\big(df(p)(X(p)), df(p)(Y(p))\big) = S_p(X(p), Y(p)).\end{aligned}$$ □

By (4.44), we already knew that we need not determine the special chart x as in (4.41) and h as in Proposition 4.3.1 to compute the second fundamental form. Proposition 4.3.5 provides another argument for this fact.

Now let $x : U \longrightarrow U'$ be an arbitrary chart on M. It then follows from Proposition 4.3.5 that the (vector-valued) entries in the fundamental matrix of the second fundamental form with respect to x are given by

$$h_{ij} := S\Big(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\Big) = \left[\frac{\partial^2 f}{\partial x^i \partial x^j}\right]^N. \tag{4.46}$$

In conclusion, by (4.37) and (4.46), the tangential component of the second derivative of f determines the Christoffel symbols, and the normal component determines the coefficients of the second fundamental form.

4.3.1 Hypersurfaces

The classical case of submanifolds of Euclidean spaces are surfaces in $\mathbb{R}^3$. Since the discussion is essentially the same, we consider hypersurfaces $M \subseteq \mathbb{R}^{m+1}$ or immersions $f : M \longrightarrow \mathbb{R}^{m+1}$. Then, for all $p \in M$, $\dim N_p f = 1$. In particular, $N_p f$ contains precisely two unit vectors. For a given $p \in M$, let $n = n_p$ be one of these. Then we can write the second fundamental form S_p as

$$S_p(v, w) = S_p^n(v, w)\, n_p \quad \text{with} \quad S_p^n(v, w) := \langle S_p(v, w), n_p \rangle. \tag{4.47}$$

We call S_p^n the *second fundamental form of f at p with respect to n_p*. If x is a chart on M about p, then we obtain, by (4.46),

$$h_{ij}^n(p) := S_p^n\Big(\frac{\partial}{\partial x^i}(p), \frac{\partial}{\partial x^j}(p)\Big) = \Big\langle \frac{\partial^2 f}{\partial x^i \partial x^j}(p), n_p \Big\rangle \tag{4.48}$$

for the coefficients of the fundamental matrix of the second fundamental form. The sign of $h_{ij}^n(p)$ depends on the choice of the vector $n = n_p$.

The *Weingarten map*[11] is the self-adjoint endomorphism $L_p\colon T_pM \longrightarrow T_pM$ corresponding to the second fundamental form, which is characterized by

$$\langle L_p v, w\rangle = S_p^n(v, w) \quad \text{for all } v, w \in T_pM. \tag{4.49}$$

We know from linear algebra that the characteristic values and directions of a symmetric bilinear form on a Euclidean vector space correspond precisely to the eigenvalues and eigenvectors of the associated self-adjoint endomorphism of the vector space. Correspondingly, the characteristic values and directions of the second fundamental form are the eigenvalues and eigenvectors of the Weingarten map.

Definition 4.3.6

The characteristic values of the second fundamental form S_p^n are called the *principal curvatures* and the corresponding direction in T_pM *principal directions* of M or f at p. Regular curves $c\colon I \longrightarrow M$, such that $\dot{c}(t)$ is a principal direction for all $t \in I$ are called *lines of curvature*.

The principal directions do not depend on the choice of n_p, and only the signs of the principal curvatures do.

We now describe three classical results for surfaces in $\mathbb{R}^3$ connected with the correspondence between symmetric bilinear forms and self-adjoint endomorphisms:

1. The *Theorem of Rodrigues*[12] (1816) states that the minimum and the maximum of the function $S_p(v, v)$, where v runs through the unit vectors in T_pM, are attained at eigenvectors of L_p.
2. Let $E \subseteq \mathbb{R}^{m+1}$ be an affine plane through $f(p)$ tangent to n_p and a unit vector $u \in T_pf$. Furthermore, let $x : U \longrightarrow U'$ be the chart in (4.41). For a smooth curve $c\colon I \longrightarrow U$ through p, the image of $\sigma := f \circ c$ is contained in E if and only if $x \circ c$ lies in $\mathbb{R}\cdot u$. If this holds and $\dot{\sigma}(0)$ is a positive multiple of u, then we call σ a *normal section* of M (resp. f) through p in the direction u. See ◘ Fig. 4.5. With h as in Proposition 4.3.1, for example,

$$\sigma_u = \sigma_u(t) = (f \circ x^{-1})(tu) = f(p) + tu + h(tu) \tag{4.50}$$

is one such normal section. Since normal sections in the direction u only differ in their parameterizations locally around $t = 0$, their oriented curvatures (at least at $t = 0$), in the plane E oriented by (u, n_p), are independent of the choice of normal section and therefore define a function of the unit vector u, written as $\kappa_o(u)$. *Euler's Theorem* states that κ_o (as a function of u) is either constant or attains its maximum and minimum in directions perpendicular to one another. To see this, write $u =$

---

[11] Julius Weingarten (1836–1910).

[12] Benjamin Olinde Rodrigues (1794–1851).

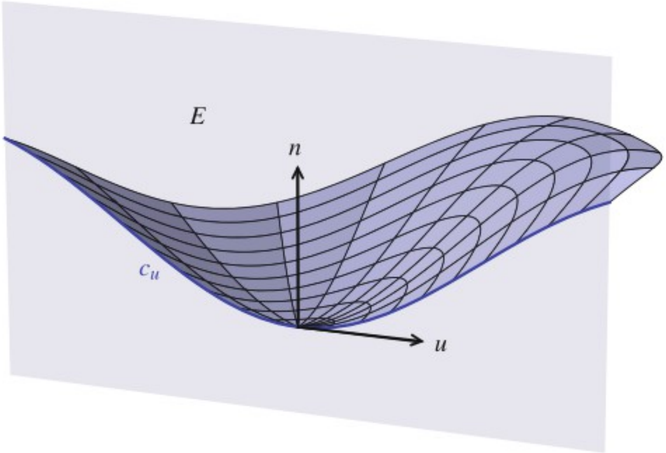

Fig. 4.5 A normal section

$df(p)(v)$ with $v \in T_pM$. With σ_u as above, we then have $\langle \dot{\sigma}_u(0), \ddot{\sigma}_u(0) \rangle = 0$ and therefore

$$\kappa_o(u) = \langle \ddot{\sigma}_u(0), n_p \rangle = S^n_p(v, v). \tag{4.51}$$

Therefore the maximum and minimum of κ_o correspond precisely to the extreme values of $S^n_p(v, v)$, where v runs through the unit vectors in T_pM. These will be attained in directions perpendicular to one another if the function $S^n_p = S^n_p(v, v)$ is not constant.

1) *Meusnier's Theorem*[13] (1776) complements Euler's Theorem: Let $v \in T_pM$ be a unit vector with $S^n_p(v, v) \neq 0$ and $c\colon I \longrightarrow M$ a smooth curve with $c(0) = p$ and $\dot{c}(0) = v$. Then for $\sigma := f \circ c$, $u := \dot{\sigma}(0) = df(p)(v)$ and $\ddot{\sigma}(0)$ are linearly independent. Let σ_u be a normal section of f through p in the direction u. Meusnier's Theorem now says that the orthogonal projection of the center of curvature M_u of σ_u at $t = 0$ onto the affine plane E through $f(p)$, spanned by $u = \dot{\sigma}(0)$ and $\ddot{\sigma}(0)$, is the center of curvature of σ at $t = 0$. Since $\dot{\sigma}(0)$ has norm 1, the curvature of σ at $t = 0$ is, by (4.4), given by $\|\ddot{\sigma}(0) - \langle \dot{\sigma}(0), \ddot{\sigma}(0) \rangle \dot{\sigma}(0)\|$. Therefore, Meusnier's Theorem follows from (4.44) and (4.51), since

$$\|\ddot{\sigma}(0) - \langle \dot{\sigma}(0), \ddot{\sigma}(0) \rangle \dot{\sigma}(0)\| \cos\theta = |\langle \ddot{\sigma}(0), n_p \rangle| = |S^n_p(v, v)|,$$

where $\theta \in [0, \pi/2)$ denotes the angle between E and N_pf. In particular, it follows from Meusnier's Theorem, that the osculating circles of all such $\sigma = f \circ c$ with

---

[13] Jean Baptiste Marie Charles Meusnier de la Place (1754–1793).

$\dot{c}(0) = v$ form a sphere with center M_u and radius $1/|S^n_p(v, v)|$, the so-called *Meusnier sphere*.

At every $p \in M$, we have a choice of two unit vectors $\pm n_p$. Naturally, it is disadvantageous to choose n_p separately for each point.

Definition 4.3.7

For an open subset $W \subseteq M$ a smooth map $n : W \longrightarrow S^m \subseteq \mathbb{R}^{m+1}$ is called a *Gauß map* on M (resp. f), if $n_p = n(p) \in N_p f$ for all $p \in W$.

Example 4.3.8

If $m = 2$ and (U, x) is a chart on M, then

$$\tilde{n} := \frac{\partial f}{\partial x^1} \times \frac{\partial f}{\partial x^2} \tag{4.52}$$

is perpendicular to f, and therefore $n := \tilde{n}/\|\tilde{n}\|$ is a Gauß map on U. For the case $m > 2$ there is an analogous formula. (Exercise: check the latter.) ■

Remark 4.3.9 We call a connected hypersurface $M \subseteq \mathbb{R}^{m+1}$ *two-sided* if there is a global Gauß map $n : M \longrightarrow S^m$, otherwise we call M one-sided. Two-sidedness and one-sidedness are equivalent to the orientability and non-orientability of M respectively. Compare with ► Sect. 3.6. The Möbius strip is one-sided.

In the following, let $n\colon M \longrightarrow S^m$ be a Gauß map on f. For vector fields X and Y on M, we then have $\langle Yf, n\rangle = 0$, and therefore

$$0 = X\langle n, Yf\rangle = \langle n, XYf\rangle + \langle Xn, Yf\rangle.$$

By Proposition 4.3.5, $\langle XYf, n\rangle = S^n(X, Y)$. It therefore follows that

$$S^n(X, Y) = \langle n, XYf\rangle = -\langle Xn, Yf\rangle = -\langle df \circ Y, dn \circ X\rangle. \tag{4.53}$$

The second fundamental form thus agrees with the bilinear form $-\langle df, dn\rangle$, also written as $-df \cdot dn$. Now, for every $p \in M$, $T_p f$ is the orthogonal complement to $n(p) \in S^m$, so $T_p f = T_{n(p)} S^m$. Therefore the ranges of $df(p)$ and $dn(p)$ agree. Since $df(p)\colon T_p M \longrightarrow T_p f$ is by definition of the first fundamental form an orthogonal transformation, we can reformulate (4.53) as follows:

$$dn = -df \circ L, \tag{4.54}$$

where L denotes the Weingarten map. If, in particular, $M \subseteq \mathbb{R}^{m+1}$ and f is the inclusion, then $L = -dn$.

For a chart (U, x) on M, we write $L(\partial/\partial x^i) = a_i^l\,\partial/\partial x^l$ with as yet undetermined coefficients a_i^l. Since

$$h_{ij}^n = S^n\Big(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\Big) = \Big\langle L\Big(\frac{\partial}{\partial x^i}\Big), \frac{\partial}{\partial x^j}\Big\rangle = a_i^l g_{lj}$$

it follows that $a_i^k = a_i^l g_{lj} g^{jk} = h_{ij}^n g^{jk}$. Therefore, we finally obtain

$$L\Big(\frac{\partial}{\partial x^i}\Big) = h_{ij}^n g^{jk} \frac{\partial}{\partial x^k}. \tag{4.55}$$

In other words, with respect to the basis $\partial/\partial x^i$, L is represented by the matrix with entries $h_{ij}^n g^{jk}$.

Up to sign, the elementary symmetric functions in the characteristic values of the second fundamental form correspond to the coefficients of the characteristic polynomial of the Weingarten map, and therefore, by (4.55), to the coefficients of the characteristic polynomial of the matrix $(h_{ij}^n g^{jk})$.

Definition 4.3.10

The determinant $K := \det L$ is called the *Gauß-Kronecker curvature*[14] of M (resp. f). The arithmetic mean $H := \operatorname{tr} L/m$ is called the *mean curvature* of M (resp. f).

If $\kappa_1, \ldots, \kappa_m$ denote the characteristic values of the second fundamental form S_p^n, that is, the eigenvalues of the Weingarten map L_p, then

$$K(p) = \kappa_1 \cdots \kappa_m \quad \text{and} \quad H(p) = (\kappa_1 + \cdots + \kappa_m)/m. \tag{4.56}$$

The sign of H always depends on the choice of n_p, the sign of K does so only when m is odd.

We ought not neglect the important classical case:

Definition 4.3.11

If M is a surface, that is, $m = 2$, then

$$K := \det L = \det(h_{ij}^n)/\det(g_{ij})$$

is called the *Gaußian curvature* of M (resp. f). See ▪ Fig. 4.6.

---

[14] Leopold Kronecker (1823–1891).

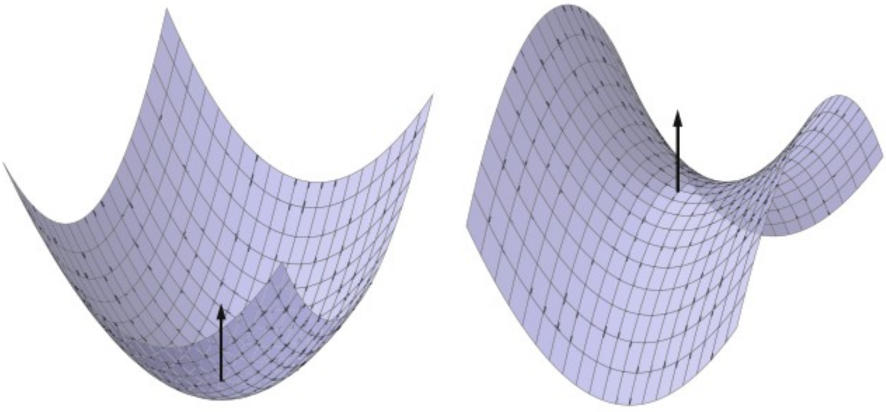

$K > 0$ $K < 0$

Fig. 4.6 Positive and negative Gaußian curvature

Example 4.3.12

1) For the *surface of revolution* $f = f(t, \varphi)$ as in Example 4.2.3.2, we have

$$\tilde{n} = (-\dot{h}\cos(\varphi), -\dot{h}\sin(\varphi), \dot{r})$$

in the sense of Example 4.3.8, and $n = \tilde{n}/\|\tilde{n}\|$ is the associated Gauß map of f. The coefficients of the second fundamental form of f, as in (4.48), are

$$h^n_{tt} = \frac{\dot{r}\ddot{h} - \ddot{r}\dot{h}}{\sqrt{\dot{r}^2 + \dot{h}^2}}, \quad h^n_{t\varphi} = h^n_{\varphi t} = 0, \quad h^n_{\varphi\varphi} = \frac{r\dot{h}}{\sqrt{\dot{r}^2 + \dot{h}^2}}.$$

Since the first and second fundamental forms are both in diagonal form, meridians and lines of latitude are lines of curvature. The corresponding principal curvatures are

$$\kappa_t = \frac{\dot{r}\ddot{h} - \ddot{r}\dot{h}}{(\dot{r}^2 + \dot{h}^2)^{3/2}} \quad \text{and} \quad \kappa_\varphi = \frac{\dot{h}}{r\sqrt{\dot{r}^2 + \dot{h}^2}}.$$

If the profile curve is parameterized by arc length, then the Gaußian curvature is $K = -\ddot{r}/r$.

2) For the *generalized helicoid* $f = f(t, \varphi)$ as in Example 4.2.3.3,

$$\tilde{n} = (a\dot{x}\sin\varphi + a\dot{y}\cos\varphi, -a\dot{x}\cos\varphi + a\dot{y}\sin\varphi, x\dot{x} + y\dot{y})$$

in the sense of Example 4.3.8, and $n = \tilde{n}/\|\tilde{n}\|$ is the associated Gauß map of f. Therefore,

$$h^n_{tt} = a\frac{\dot{y}\ddot{x} - \dot{x}\ddot{y}}{\|\tilde{n}\|}, \quad h^n_{t\varphi} = h^n_{\varphi t} = -a\frac{\dot{x}^2 + \dot{y}^2}{\|\tilde{n}\|}, \quad h^n_{\varphi\varphi} = a\frac{\dot{x}y - x\dot{y}}{\|\tilde{n}\|}$$

with $\|\tilde{n}\|^2 = a^2(\dot{x}^2 + \dot{y}^2) + (x\dot{x} + y\dot{y})^2$. The formula for the Gaußian curvature is somewhat long, and so will not be written out here.

3) For the *ruled surface* $f = f(s, t)$ as in Example 4.2.3.4, $\tilde{n} = X \times (\dot{c} + s\dot{X})$ in the sense of Example 4.3.8, and $n = \tilde{n}/\|\tilde{n}\|$ is the associated Gauß map of f. Therefore,

$$h^n_{ss} = 0, \quad h^n_{st} = h^n_{ts} = \frac{\langle \dot{X}, X \times \dot{c}\rangle}{\|\tilde{n}\|}, \quad \text{and} \quad h^n_{tt} = \frac{\langle \ddot{c} + s\ddot{X}, X \times (\dot{c} + s\dot{X})\rangle}{\|\tilde{n}\|}.$$

Since $h^n_{ss} = 0$, the Gaußian curvature K is non-positive.

■

Definition 4.3.13

A point $p \in M$ is called an *umbilic point*, if the function $S^n_p(v, v)$ is constant, where v runs through the unit vectors in T_pM.

Clearly, $p \in M$ is an umbilic point if and only if the Weingarten map L_p is a multiple of the identity or, equivalently by (4.54), if $dn(p)$ is a multiple of $df(p)$.

Proposition 4.3.14 *If M is connected, $m \geq 2$, and all points in M are umbilic, then the image of f is contained in an affine hyperplane or a sphere.*

Proof

Since all points in M are umbilic, there is, by (4.54), a smooth function λ on M with $dn = \lambda\, df$. Now let (U, x) be a chart on M with U connected. Then $\partial n/\partial x^i = \lambda\, \partial f/\partial x^i$, and therefore

$$\frac{\partial\lambda}{\partial x^i}\frac{\partial f}{\partial x^j} + \lambda\frac{\partial^2 f}{\partial x^i \partial x^j} = \frac{\partial^2 n}{\partial x^i \partial x^j} = \frac{\partial^2 n}{\partial x^j \partial x^i} = \frac{\partial\lambda}{\partial x^j}\frac{\partial f}{\partial x^i} + \lambda\frac{\partial^2 f}{\partial x^j \partial x^i}.$$

Since the second terms on the left and right are equal, the first terms are as well. For $i \neq j$, however, $\partial f/\partial x^i$ and $\partial f/\partial x^j$ are linearly independent. Since $m \geq 2$, it therefore follows that the partial derivatives of λ vanish, so λ is constant on U.

For $\lambda = 0$ it follows that n is constant on U. Then the image of U under f is contained in an affine hyperplane of $\mathbb{R}^{m+1}$. For $\lambda \neq 0$ it follows that $f - n/\lambda = x_0 =$ constant on U. Since $\|n\| = 1$, we therefore obtain that $\|f - x_0\| = 1/|\lambda|$ on U. This shows the claim in the case $M = U$. But M is connected, so the claim follows. □

Affine planes in $\mathbb{R}^3$ and cylinders over regular plane curves are surfaces in space with vanishing Gaußian curvature. With an appropriate parameterization, one can (clearly) even ensure that their first fundamental forms have constant coefficients δ_{ij}. Consequently, the first fundamental form by no means determines the external appearance of a surface in space. Generalized helicoids provide an interesting example of this.

Example 4.3.15

Let $f = f(t,\varphi)$ be a generalized helicoid as in Example 4.2.3.3. We will show that—up to a t-dependent shift in the φ-parameter (that is, a diffeomorphism of the parameter domain)—there is a surface of revolution with the same first fundamental form. We first consider surfaces of the form

$$\tilde{f}(t,\varphi) = f(t, \varphi + \alpha(t)).$$

The partial derivatives of $\tilde{f}$ are

$$\begin{aligned}\frac{\partial \tilde{f}}{\partial t}(t,\varphi) &= \frac{\partial f}{\partial t}(t, \varphi + \alpha) + \frac{\partial f}{\partial \varphi}(t, \varphi + \alpha)\,\dot{\alpha},\\ \frac{\partial \tilde{f}}{\partial \varphi}(t,\varphi) &= \frac{\partial f}{\partial \varphi}(t, \varphi + \alpha).\end{aligned}$$

With x, y, and a as in Example 4.2.3.3, the coefficients of the first fundamental form of $\tilde{f}$ can be computed to be

$$\begin{aligned}\tilde{g}_{tt} &= \dot{x}^2 + \dot{y}^2 + 2(x\dot{y} - \dot{x}y)\,\dot{\alpha} + (x^2 + y^2 + a^2)\,\dot{\alpha}^2,\\ \tilde{g}_{t\varphi} = \tilde{g}_{\varphi t} &= x\dot{y} - \dot{x}y + (x^2 + y^2 + a^2)\,\dot{\alpha},\\ \tilde{g}_{\varphi\varphi} &= x^2 + y^2 + a^2.\end{aligned}$$

Now the first fundamental form of $\tilde{f}$ ought to also be that of a surface of revolution. Therefore, we must have $\tilde{g}_{t\varphi} = \tilde{g}_{\varphi t} = 0$ and $\tilde{g}_{\varphi\varphi} = r^2$, where the profile curve of the surface of revolution as in Example 4.2.3.2 is given by radius r and height h. So we must have

$$\dot{\alpha} = \frac{\dot{x}y - x\dot{y}}{r^2} \quad \text{with} \quad r = \sqrt{x^2 + y^2 + a^2}$$

and

$$\frac{(x\dot{x} + y\dot{y})^2}{r^2} + \dot{h}^2 = \dot{r}^2 + \dot{h}^2 = \tilde{g}_{tt} = \dot{x}^2 + \dot{y}^2 - \frac{(x\dot{y} - \dot{x}y)^2}{r^2}.$$

It therefore follows that

$$\dot{h} = \pm\frac{a}{r}\sqrt{\dot{x}^2 + \dot{y}^2}, \quad \text{so} \quad h = \pm\int \frac{a}{r}\sqrt{\dot{x}^2 + \dot{y}^2}.$$

We are free to choose the sign and constant of integration of h. Geometrically, these choices correspond to a reflection or a translation in the z-direction respectively. The constant of integration of α can also be chosen freely.

We can also assume that the profile curve of f is parameterized by arc length. Then

$$h = \pm\int \frac{a}{r}\,dt.$$

;. :5:6 +* *3 $ @. .

For the helicoid $f = (t\cos\varphi, t\sin\varphi, 1)$, we obtain (with appropriate choices)

$$r(t) = \sqrt{t^2+1} \quad \text{and} \quad h(t) = \operatorname{arcsinh} t.$$

So the corresponding surface of revolution, $t = \sinh h$ and $r = \cosh h$, is a catenoid. ■

4.4 Gauß Equations and the Theorema Egregium

We now return to the covariant derivative.

Definition 4.4.1

For $Y \in \mathcal{V}(M)$, the *covariant derivative* of Y in the direction $v \in T_pM$ is the unique tangent vector $\nabla_v Y \in T_pM$ with

$$df \circ \nabla_v Y = [v(Yf)]^T = [d(Yf)(v)]^T.$$

For $X \in \mathcal{V}(M)$, the covariant derivative of Y in the direction of X is correspondingly given by $(\nabla_X Y)(p) := \nabla_{X(p)} Y = [(XYf)(p)]^T$.

Let $p \in M$ and (U, x) be a chart on M about p. Let $v = \xi^i\,\partial/\partial x^i(p)$ and $Y = \eta^i\,\partial/\partial x^i$ be a smooth vector field on U. Then $Yf = \eta^i\,\partial f/\partial x^i$ and so

$$v(Yf) = v(\eta^i)\frac{\partial f}{\partial x^i}(p) + \xi^i\eta^j(p)\frac{\partial^2 f}{\partial x^i \partial x^j}(p).$$

By the definition of the Christoffel symbols, therefore,

$$\nabla_v Y = \left(v(\eta^k) + \Gamma^k_{ij}(p)\xi^i\eta^j(p)\right)\frac{\partial}{\partial x^k}(p), \tag{4.57}$$

see (4.37). If X is a smooth vector field on U with $X = \xi^i\,\partial/\partial x^i$, then correspondingly

$$\nabla_X Y = \left(X\eta^k + \Gamma^k_{ij}\xi^i\eta^j\right)\frac{\partial}{\partial x^k}. \tag{4.58}$$

We see that $\nabla_X Y$ is again a smooth vector field.

Corollary 4.4.2

1. *If Y is a smooth vector field on M and $c\colon I \longrightarrow M$ is a smooth curve with $\dot c(t_0) =: v \in T_pM$, then $Y \circ c$ is a smooth vector field along c with $(Y \circ c)'(t_0) = \nabla_v Y$.*
2. *If X and Y are smooth vector fields on M, then*

$$XYf = [XYf]^T + [XYf]^N = df \circ \nabla_X Y + S(X, Y). \tag{4.59}$$

Proof

The first claim follows directly from (4.38) and (4.57), the second from Proposition 4.3.5 and Definition 4.4.1. □

With (4.58), we see that the covariant derivative $\nabla_X Y$ with respect to a chart corresponds to the usual derivative of the coefficients of Y in the direction of X up to a correction term of order 0. Furthermore, Proposition 4.2.23 shows us that the covariant derivative belongs to the interior geometry of M.

Proposition 4.4.3 *For the map* $\nabla\colon \mathcal{V}(M)\times\mathcal{V}(M)\longrightarrow\mathcal{V}(M)$, $(X, Y)\mapsto \nabla_X Y$, *the following hold:*

1. ∇ *is bilinear, that is,* $\mathbb{R}$*-linear in both* X *and* Y.
2. *For all* $\varphi \in \mathcal{F}(M)$, $\nabla_{\varphi X} Y = \varphi\nabla_X Y$ *and* $\nabla_X(\varphi Y) = X(\varphi)Y + \varphi\nabla_X Y$.
3. ∇ *is* symmetric*:* $\nabla_X Y - \nabla_Y X = [X, Y]$.
4. ∇ *is* metric *(product rule):* $X\langle Y, Z\rangle = \langle\nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle$.

Proof

The first two assertions follow from (4.58). By the symmetry of the second fundamental form and (4.59),

$$\begin{aligned} df\circ[X, Y] &= [X, Y]f = XYf - YXf \\ &= df\circ\nabla_X Y - df\circ\nabla_Y X = df\circ(\nabla_X Y - \nabla_Y X). \end{aligned}$$

Assertion 3 follows. The product rule follows from Proposition 4.2.18.3 together with Corollary 4.4.2.1. □

A map $\nabla\colon \mathcal{V}(M)\times\mathcal{V}(M)\longrightarrow\mathcal{V}(M)$ that satisfies the first two conditions of Proposition 4.4.3 is called a *connection* or sometimes a *covariant derivative* on M. The special connection from Definition 4.4.1 is called the *Levi-Civita connection* of M (resp. f). It is distinguished among all connections on M by the last two conditions in Proposition 4.4.3.

We next discuss the *Riemann curvature tensor*,[15]

$$R(X, Y)Z := \nabla_X\nabla_Y Z - \nabla_Y\nabla_X Z - \nabla_{[X,Y]}Z, \tag{4.60}$$

a further object of interior geometry. Patient computation yields the following formula for vector fields X, Y, Z on the domain U of a chart x on M:

$$R(X, Y)Z = \left\{\partial\Gamma^l_{jk}/\partial x^i - \partial\Gamma^l_{ik}/\partial x^j + \Gamma^l_{i\nu}\Gamma^\nu_{jk} - \Gamma^l_{j\nu}\Gamma^\nu_{ik}\right\}\xi^i\eta^j\zeta^k\frac{\partial}{\partial x^l}. \tag{4.61}$$

We see that R is pointwise linear in each of the coefficients ξ^i, η^j, and ζ^k of X, Y, and Z, and no derivatives of these coefficients appear. This justifies the designation of R as a tensor.

---

[15] Georg Friedrich Bernhard Riemann (1826–1866).

Proof

The first claim follows directly from (4.38) and (4.57), the second from Proposition 4.3.5 and Definition 4.4.1. □

With (4.58), we see that the covariant derivative $\nabla_X Y$ with respect to a chart corresponds to the usual derivative of the coefficients of Y in the direction of X up to a correction term of order 0. Furthermore, Proposition 4.2.23 shows us that the covariant derivative belongs to the interior geometry of M.

Proposition 4.4.3 *For the map $\nabla\colon \mathcal{V}(M)\times\mathcal{V}(M)\longrightarrow\mathcal{V}(M)$, $(X, Y)\mapsto \nabla_X Y$, the following hold:*

1. *∇ is bilinear, that is, $\mathbb{R}$-linear in both X and Y.*
2. *For all $\varphi \in \mathcal{F}(M)$, $\nabla_{\varphi X} Y = \varphi\nabla_X Y$ and $\nabla_X(\varphi Y) = X(\varphi)Y + \varphi\nabla_X Y$.*
3. *∇ is* symmetric*: $\nabla_X Y - \nabla_Y X = [X, Y]$.*
4. *∇ is* metric *(product rule): $X\langle Y, Z\rangle = \langle\nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle$.*

Proof

The first two assertions follow from (4.58). By the symmetry of the second fundamental form and (4.59),

$$\begin{aligned} df \circ [X, Y] &= [X, Y]f = XYf - YXf \\ &= df \circ \nabla_X Y - df \circ \nabla_Y X = df \circ (\nabla_X Y - \nabla_Y X). \end{aligned}$$

Assertion 3 follows. The product rule follows from Proposition 4.2.18.3 together with Corollary 4.4.2.1. □

A map $\nabla\colon \mathcal{V}(M)\times\mathcal{V}(M)\longrightarrow\mathcal{V}(M)$ that satisfies the first two conditions of Proposition 4.4.3 is called a *connection* or sometimes a *covariant derivative* on M. The special connection from Definition 4.4.1 is called the *Levi-Civita connection* of M (resp. f). It is distinguished among all connections on M by the last two conditions in Proposition 4.4.3.

We next discuss the *Riemann curvature tensor*,[15]

$$R(X, Y)Z := \nabla_X\nabla_Y Z - \nabla_Y\nabla_X Z - \nabla_{[X,Y]}Z, \tag{4.60}$$

a further object of interior geometry. Patient computation yields the following formula for vector fields X, Y, Z on the domain U of a chart x on M:

$$R(X, Y)Z = \left\{\partial\Gamma^l_{jk}/\partial x^i - \partial\Gamma^l_{ik}/\partial x^j + \Gamma^l_{i\nu}\Gamma^\nu_{jk} - \Gamma^l_{j\nu}\Gamma^\nu_{ik}\right\}\xi^i\eta^j\zeta^k\frac{\partial}{\partial x^l}. \tag{4.61}$$

We see that R is pointwise linear in each of the coefficients ξ^i, η^j, and ζ^k of X, Y, and Z, and no derivatives of these coefficients appear. This justifies the designation of R as a tensor.

---

[15] Georg Friedrich Bernhard Riemann (1826–1866).

Proposition 4.4.4 (Gauß Equations) *For vector fields X, Y, V, W on M,*

$$\langle R(X,Y)V, W\rangle = \langle S(X,W), S(Y,V)\rangle - \langle S(X,V), S(Y,W)\rangle.$$

Proof

Since Vf is a map to the vector space $\mathbb{R}^n$,

$$XYVf - YXVf = [X,Y]Vf.$$

With (4.59) and Proposition 4.4.3.4, we obtain

$$\begin{aligned}\langle XYVf, Wf\rangle &= X\langle YVf, Wf\rangle - \langle YVf, XWf\rangle \\ &= X\langle [YVf]^T, Wf\rangle - \langle [YVf]^T, [XWf]^T\rangle - \langle [YVf]^N, [XWf]^N\rangle \\ &= X\langle \nabla_X Y, W\rangle - \langle \nabla_Y V, \nabla_X W\rangle - \langle S(Y,V), S(X,W)\rangle \\ &= \langle \nabla_X \nabla_Y V, W\rangle - \langle S(Y,V), S(X,W)\rangle,\end{aligned}$$

and analogously, $\langle YXVf, Wf\rangle = \langle \nabla_Y \nabla_X V, W\rangle - \langle S(X,V), S(Y,W)\rangle$. Finally, by (4.59), it further holds that

$$\langle [X,Y]Vf, Wf\rangle = \langle df \circ \nabla_{[X,Y]}V + S([X,Y],V), Wf\rangle = \langle \nabla_{[X,Y]}V, W\rangle.$$

Altogether, we obtain

$$\begin{aligned}0 &= \langle XYVf - YXVf - [X,Y]Vf, Wf\rangle \\ &= \langle \nabla_X\nabla_Y V - \nabla_Y\nabla_X V - \nabla_{[X,Y]}V, W\rangle \\ &\qquad - \langle S(Y,V), S(X,W)\rangle + \langle S(X,V), S(Y,W)\rangle \\ &= \langle R(X,Y)V, W\rangle - \langle S(Y,V), S(X,W)\rangle + \langle S(X,V), S(Y,W)\rangle. \qquad \square\end{aligned}$$

Proposition 4.4.4 contains, in a more general form, one of the fundamental geometric insights of Gauß, namely, that the Gaußian curvature, defined with the aid of the second fundamental form, is a quantity belonging to interior geometry. In our formulation, Gauß's insight[16] reads as follows:

Theorema Egregium 4.4.5 (Gauß (1827)) *Let M be a surface and $f\colon M \longrightarrow \mathbb{R}^3$ be an immersion. Let $p \in M$, and let $v, w \in T_pM$ be linearly independent. Then*

$$K(p) = \frac{\langle R(v,w)w, v\rangle}{\|v^2\|\|w\|^2 - \langle v,w\rangle^2}.$$

---

[16] Gauß called this insight the *Theorema Egregium*, meaning "remarkable theorem". The word "egregium" is related etymologically to the word "egregious".

Proof

By Proposition 4.4.4, it follows that $\langle R(v,w)w,v\rangle = S^n_p(v,v)S^n_p(w,w) - S^n_p(v,w)^2$, independently of the choice of a normal vector n_p at p. Therefore, the right-hand side of the asserted equality becomes

$$\frac{S^n_p(v,v)S^n_p(w,w) - S^n_p(v,w)^2}{\|v\|^2\|w\|^2 - \langle v,w\rangle^2}.$$

The number does not depend on the choice of v and w, as long as v and w are linearly independent. We can, for example, choose a chart x around p, and set $v = (\partial/\partial x^1)(p)$ and $w = (\partial/\partial x^2)(p)$. Then

$$\begin{aligned}\frac{S^n_p(v,v)S^n_p(w,w) - S^n_p(v,w)^2}{\|v^2\|\|w\|^2 - \langle v,w\rangle^2} &= \frac{h^n_{11}h^n_{22} - (h^n_{12})^2}{g_{11}g_{22} - g^2_{12}}(p)\\ &= \det(h^n_{ik}g^{kj})(p) = K(p).\end{aligned}$$ □

More generally, the terms on the right-hand side of the equation in Proposition 4.4.4 coming from the second fundamental form are quantities belonging to interior geometry, since the Riemann curvature tensor is. Among these quantities is the *sectional curvature*, introduced by Riemann, that associates a curvature to tangent planes to M: if $p \in M$, P is a linear 2-dimensional subspace of T_pM, and (v, w) is a basis of P, then

$$K(P) := \frac{\langle R(v,w)w,v\rangle}{\|v^2\|\|w\|^2 - \langle v,w\rangle^2} \tag{4.62}$$

is the sectional curvature of P. This generalizes the Gaußian curvature and was introduced by Riemann in his famous Habilitation[17] lecture (1854) before he had defined his curvature tensor.

In the following examples we replace $\mathbb{R}^n$ by a general finite-dimensional Euclidean vector space W. By means of an orthonormal transformation, we can then identify W with $\mathbb{R}^{\dim W}$ and, accordingly, use the previous concepts and results with W as our ambient space.

Example 4.4.6

Let $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}, \mathbb{H}\}$ and V be an n-dimensional vector space over $\mathbb{K}$ together with a positive definite sesquilinear form, denoted by (v, w). After a choice of an orthonormal basis, $V \cong \mathbb{K}^n$ with sesquilinear form $(x, y) = \sum \bar{x}_i y_i$.

1) Let $G = \{A \in \mathrm{End}(V) \mid A^*A = \mathrm{id}\}$ be the Lie group of all endomorphisms which preserve the sesquilinear form on V. In Example 2.3.8.3 we saw, modulo the

---

[17] In the German-speaking world, a Habilitation is a post-doctoral qualification necessary to teach at the university level.

identification of V with $\mathbb{K}^n$ seen as a Euclidean space, that G is a submanifold of $W = \mathrm{End}(V)$ with

$$T_E G = \{C \in \mathrm{End}(V) \mid C^* = -C\}. \tag{4.63}$$

The real part of the sesquilinear form $(A, B) := \mathrm{tr}(A^* B)$ is a scalar product on $\mathrm{End}(V) \cong \mathbb{R}^{dn \times n}$, $d = \dim_{\mathbb{R}} \mathbb{K}$. The corresponding orthogonal complement of $T_E G$ in $\mathrm{End}(V)$ is

$$N_E G = H(V) = \{C \in \mathrm{End}(V) \mid C^* = C\}. \tag{4.64}$$

For all $A \in G$, $A\, T_E G\, A^{-1} = T_E G$ and $A\, H(V)\, A^{-1} = H(V)$, and left and right translation by A preserve the sesquilinear form on $\mathrm{End}(V)$:

$$(AC, AD) = (CA, DA) = (C, D) = \mathrm{tr}(C^* D)$$

for all $C, D \in \mathrm{End}(V)$. In particular,

$$T_A G = A T_E G = T_E G A \quad \text{and} \quad N_A G = A H(V) = H(V) A. \tag{4.65}$$

For all $A \in G$ and $C \in \mathrm{End}(V)$, it therefore follows that

$$\pi_A^T(AC) = \frac{1}{2} A(C - C^*) \quad \text{and} \quad \pi_A^N(AC) = \frac{1}{2} A(C + C^*). \tag{4.66}$$

The (tangential) left-invariant vector fields on G are of the form $X_C(A) = AC$ with $C \in T_E G$, compare with Exercise 2.7.14.1. Since X_C is the restriction of right multiplication by C to G, and the latter is linear on $\mathrm{End}(V)$,

$$dX_C(A)(AD) = ADC = X_{DC}(A)$$

for all $A \in G$ and $D \in T_E G$. For left invariant vector fields X_C and X_D (with $C, D \in T_E G$), we can use (4.59) to show that

$$dX_D(X_C) = \nabla_{X_C} X_D + S(X_C, X_D) = X_{(CD-DC)/2} + X_{(CD+DC)/2} \tag{4.67}$$

for the covariant derivative and second fundamental form of G. For $B, C, D \in T_E G$, the curvature tensor R of G then becomes

$$R(X_B, X_C) X_D = X_F \quad \text{with} \quad F = -\frac{1}{4}[[B, C], D] \in T_E G. \tag{4.68}$$

Since $B^* = -B$ and $[B, C]^* = -[B, C]$, we therefore obtain

$$\langle R(X_B, X_C) X_C, X_B \rangle = \frac{1}{4} \mathrm{Re}\, \mathrm{tr}([[B, C], C] B) = \frac{1}{4} \|[B, C]\|^2 \tag{4.69}$$

for the sectional curvature of G. In particular, the sectional curvature is non-negative.

Now let $A = A(t) = Be^{tC}$, with $B \in G$, $C \in T_E G$, and let the exponential series be as in Exercise 2.7.6. Then A is a smooth curve in G with

$$\dot{A} = AC \quad \text{and} \quad \ddot{A} = AC^2. \tag{4.70}$$

Now $C^2 \in N_E G$, so $\ddot{A} \in N_A G$. Therefore, A is a geodesic, and so we have characterized all the geodesics of G.

2) Let $G_k(V)$ be the Grassmannian manifold of k-dimensional $\mathbb{K}$-linear subspaces in V as in Example 2.1.15.5. For $P \in G_k(V)$, let $F(P) \colon V \longrightarrow V$ be the orthogonal projection of V onto P with image P and kernel $P^\perp$. If $(e_1, \dots, e_k)$ is an orthonormal basis of P, then

$$F(P)(v) = \sum_{1 \le i \le k} e_i\,(e_i, v). \tag{4.71}$$

We now extend $(e_1, \dots, e_k)$ to an orthonormal basis $E = (e_1, \dots, e_n)$ of V and consider the chart κ_E on $G_k(V)$ around P associated to E as in Example 2.1.15.5. Then for $A = (a_i^\mu) \in \mathbb{K}^{n-k,k}$, $\kappa_E^{-1}(A)$ is the graph of the linear map $P \longrightarrow P^\perp$, $e_i u^i \mapsto e_{k+j} a_i^j u^i$. Therefore,

$$F(\kappa_E^{-1}(A)) = \begin{pmatrix} 1 & 0 \\ A & 0 \end{pmatrix} \begin{pmatrix} 1 & -A^* \\ A & 1 \end{pmatrix}^{-1} \tag{4.72}$$

is the matrix of $F(\kappa_E^{-1}(A))$ with respect to E. In particular, it therefore follows that F is smooth with differential

$$dF(P)(A) = \begin{pmatrix} 0 & A^* \\ A & 0 \end{pmatrix},$$

where $T_P G_k(V)$ is identified with $\mathrm{Hom}(V, V^\perp) \cong \mathbb{K}^{n-k,k}$ as in Example 2.2.3.3. In particular, $F \colon G_k(V) \longrightarrow \mathrm{End}(V)$ is an embedding.

We equip $\mathrm{End}\, V$ with the real part of the sesquilinear form $\mathrm{tr}(X^*Y)$ as a scalar product. Then the first fundamental form of F in P becomes

$$\begin{aligned} \langle A, B \rangle_P &= \langle dF(P)(A), dF(P)(B) \rangle \\ &= \mathrm{Re}(\mathrm{tr}(A^*B) + \mathrm{tr}(AB^*)) = 2\,\mathrm{Re}\,\mathrm{tr}(A^*B). \end{aligned}$$

To determine the geodesics and the second fundamental form, we now write P_0 in place of P. Again let $A \in \mathbb{K}^{n-k,k}$ and

$$P = P(t) = e^{tB} P_0 \quad \text{with} \quad B = \begin{pmatrix} 0 & -A^* \\ A & 0 \end{pmatrix}. \tag{4.73}$$

Then P is a smooth curve in $G_k(V)$ with $P(0) = P_0$ and $\dot{P}(0) = A$. Now $(e^{tB})^* = e^{-tB}$, so the e^{tB} preserve the sesquilinear form $(.,.)$ on V. It therefore follows that

$$F(P(t)) = e^{tB} F(P_0) e^{-tB} = e^{tB} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} e^{-tB} \tag{4.74}$$

and $T_P F = e^{tB}\, T_{P_0} F\, e^{-tB}$. The second derivative of $F \circ P$ is computed to be

$$\frac{d^2(F \circ P)}{dt^2} = e^{tB} \begin{pmatrix} -2A^*A & 0 \\ 0 & 2AA^* \end{pmatrix} e^{-tB}. \tag{4.75}$$

Therefore, the second derivative of $F \circ P$ is normal to F, so $P = P(t)$ is a geodesic, and so we have characterized all the geodesics of $G_k(V)$. By polarization, we furthermore obtain the second fundamental form F at P_0,

$$S_{P_0}(A, B) = \begin{pmatrix} -A^*B - B^*A & 0 \\ 0 & AB^* + BA^* \end{pmatrix}.$$

Since the $F(P)$ are orthogonal projections onto k-dimensional subspaces of V, $F(P)^* = F(P)$ and $\operatorname{tr}(F(P)) = k$. The image of F therefore lies in a sphere of radius $\sqrt{k}$ in the space of self-adjoint endomorphisms $W = H(V) \subseteq \operatorname{End}(V)$. Thus, the $T_P F$ and the images of the second fundamental forms S_P of F are contained in $H(V)$, and our formulas above show this explicitly.

■

The curvature tensor and sectional curvature are quantities belonging to interior geometry. This is the beginning of Riemannian geometry based on Riemannian metrics, that is, smooth families g_p of scalar products on the T_pM, which are no longer required to be the first fundamental form of a given immersion $M \longrightarrow \mathbb{R}^n$. As in the case of the first fundamental form, one obtains a Levi-Civita connection, a curvature tensor, and the associated sectional curvature, and, armed with these, enters the realm of Riemannian geometry.

4.5 Supplementary Literature

Good sources to round off the discussion in this chapter are [Kl, dC], [Sp2, chapters 1&2], [Sp3, Ho] and [ST]. The more classical [MP] and [St1, St2, St3] are also interesting. In [Ch], introductory articles on many aspects of global differential geometry can be found. Beside these, there are also a number of newer introductions to differential geometry, for example [Bä, EJ, Le], and [Kü].

4.6 Exercises

Exercise 4.6.1

1. The length of a smooth curve $c\colon [a,b] \longrightarrow \mathbb{R}^n$ is invariant under monotone reparameterization: If $\varphi\colon [\alpha,\beta] \longrightarrow [a,b]$ is monotone, surjective, and smooth, then $L(c \circ \varphi) = L(c)$.
2. Show that analogues of the statements about length and energy in the corresponding part of ▶ Sect. 4.1 hold for piecewise smooth curves (as in (3.4)).
3. Let $c\colon [a,b] \longrightarrow \mathbb{R}^n$ be a piecewise smooth curve. For subdivisions

$$U : a = t_0 < \cdots < t_k = b$$

of $[a,b]$, let $\delta(U) = \max(t_i - t_{i-1})$ and $L(U)$ be the length of the piecewise linear curve with corners at $c(t_0), c(t_1), \dots, c(t_k)$. Show: if (U_n) is a sequence of subdivisions of $[a,b]$ with $\delta(U_n) \longrightarrow 0$, then $L(U_n) \longrightarrow L(c)$.

4. Let $c\colon [0,\infty) \longrightarrow \mathbb{R}^2$ be a smooth curve of the form $c(t) = (t, y(t))$. If $\dot{y}(t)$ converges for $t \to \infty$, then $\lim_{t\to\infty} L(c|_{[0,t]})/\|(t,y(t)) - (0,y(0))\| = 1$. ■

Exercise 4.6.2

Compute the speed and curvature of the curves $c\colon \mathbb{R} \longrightarrow \mathbb{R}^2$ given by

1. $c(t) = (t, t^k)$ with $k \geq 0$ and, more generally, $c(t) = (t, f(t))$;
2. $c(t) = \exp(t)(\cos t, \sin t)$ and, more generally, $c(t) = f(t)(\cos t, \sin t)$ with $f > 0$;
3. $c(t) = (a\cos t, b\sin t)$.

Sketch each curve or curve type. ■

Exercise 4.6.3

Let $c\colon I \longrightarrow \mathbb{R}^n$ be a regular curve with curvature κ.

1. If the image of c lies in a circle K of radius R, then, for all $t \in I$, K is the osculating circle of c and $\kappa \equiv 1/R$.
2. The image of c lies in a line if and only if κ vanishes.
3. For $x \in \mathbb{R}^n$, suppose the function $r = r(t) = \|c(t) - x\|$ has a relative maximum at $t_0 \in I$. Then $\kappa(t_0) = 1/R(t_0) \geq 1/r(t_0)$.
4. If $\varphi\colon J \longrightarrow I$ is a change of parameters, then $\tilde{\kappa} = \kappa \circ \varphi$, where $\tilde{\kappa}$ denotes the curvature of the curve $\tilde{c} := c \circ \varphi$.
5. If $B : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is a motion, then the curve $B \circ c$ has the same speed and curvature as c.

■

Exercise 4.6.4

Let $c\colon I \longrightarrow \mathbb{R}^2$ be a regular plane curve.

1. If $\varphi\colon J \longrightarrow I$ is a change of parameters with $\dot{\varphi} > 0$, then $\tilde{\kappa}_o = \kappa_o \circ \varphi$, where $\tilde{\kappa}_o$ denotes the oriented curvature of $\tilde{c} := c \circ \varphi$.
2. If $B : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ is an orientation-preserving motion, then the regular plane curve $B \circ c$ has the same oriented curvature as c.

■

Exercise 4.6.5

In the following, let $c\colon I \longrightarrow \mathbb{R}^2$ be a regular plane curve with field of directions e, principal normal field n and oriented curvature κ_o.

1. If $c+n$ is constant, then c moves along a circle of radius 1.
2. If $\kappa_o(t) \neq 0$ for all $t \in I$, then the curve $a = a(t) = c(t) + n(t)/\kappa_o(t)$ sending t to the center of curvature of c at t is called the *evolute* of c. Compute e, n, κ_o, and a for the parabola (t, t^2) and the catenary $(t, \cosh t)$.
3. Let $\kappa_o(t) \neq 0$ for all $t \in I$ and a be the evolute of c. Show that $\dot{a}(t)$ is a multiple of $n(t)$ for all $t \in I$ and conclude that, for all points where it is defined, the tangent to the evolute intersects the curve c perpendicularly at $c(t)$. For $s < t$ in I, compute the length of the arc $a|_{[s,t]}$ of the evolute, and compare it with the radius of curvature of c at s and t.
4. Again let $\kappa_o(t) \neq 0$ for all $t \in I$ and c be parameterized by arc length for simplicity. For $\beta \in \mathbb{R} \setminus I$, we then call $b = b(t) = c(t) + (\beta - t)\dot{c}(t)$, $t \in I$, the *evolvent* of c. One can image b as the endpoint of a thread wound tautly around c. Compute the field of directions, the principal normal field, and the curvature of b, and conclude that the curve c is the evolute of its evolvent.

■

Exercise 4.6.6

Compute the field of directions, principal normal field, and binormal field of the *helix* $c\colon \mathbb{R} \longrightarrow \mathbb{R}^3$, $c(t) = (r\cos t, r\sin t, ht)$ with $r, h \in \mathbb{R}$, $r > 0$. Show that the speed, curvature, and torsion of the helix are given by

$$\|\dot{c}(t)\| = \sqrt{r^2+h^2}, \quad \kappa(t) = \frac{r}{r^2+h^2}, \quad \text{and} \quad \tau(t) = \frac{h}{r^2+h^2},$$

and are therefore, in particular, constant. Up to reparameterization and motions we therefore obtain all space curves with constant curvature $\kappa > 0$ and torsion τ. Also verify that the helix is a *generalized helix*, that is, that $\dot{c}$ and the z-axis enclose a constant angle. ■

Exercise 4.6.7

Let $c\colon I \longrightarrow \mathbb{R}^3$ be a smooth space curve such that $\dot{c}(t)$ and $\ddot{c}(t)$ are linearly independent for all $t \in I$.

1. The curvature and torsion of c are given by

$$\kappa(t) = \frac{\|\dot{c}(t) \times \ddot{c}(t)\|}{\|\dot{c}(t)\|^3} \quad \text{and} \quad \tau(t) = \frac{\det(\dot{c}(t), \ddot{c}(t), \dddot{c}(t))}{\|\dot{c}(t) \times \ddot{c}(t)\|^2}.$$

2. The image of c lies in an affine plane in $\mathbb{R}^3$ if and only if the torsion of c vanishes.
3. If c is parameterized by arc length and $t_0 = 0 \in I$, then

$$c(t) = c(0) + \big(t - \frac{t^3}{6}\kappa_0^2\big)e_0 + \big(\frac{t^2}{2}\kappa_0 + \frac{t^3}{6}\dot{\kappa}_0\big)n_0 + \frac{t^3}{6}\kappa_0\tau_0 b_0 + o(t^3),$$

   where in each case the index 0 denotes the value at 0. Sketch the projections of c onto the *osculating*, *normal*, and *rectifying planes* of c at $t=0$, that is, the planes spanned by the pairs of vectors (e_0, n_0), (n_0, b_0), and (e_0, b_0) respectively.

■

Exercise 4.6.8 (Concerning Definition 4.1.13)

1. Determine the parallel normal fields along a regular space curve with vanishing torsion.
2. Determine the parallel normal fields along the helix; see Example 4.1.18 and Exercise 4.6.6.

■

Exercise 4.6.9

1. Write $S^2 \setminus \{(0, 0, \pm 1)\}$ as a surface of revolution.
2. Write the one-sheeted *hyperboloid* $x^2 + y^2 - z^2 = 1$ as a surface of revolution and—in two different ways—as a ruled surface.
3. Represent the Möbius strip as a ruled surface.

■

Exercise 4.6.10

Let $c = (r, h) : I \longrightarrow \mathbb{R}^2$ be the profile curve of a surface of revolution $f = f(t, \varphi)$ as in Example 4.2.3.2. Show that one can reparameterize c so that

(a) $g_{tt} \equiv 1$, or

(b) $g_{tt} = g_{\varphi\varphi}$, or

(c) $g_{tt} g_{\varphi\varphi} = 1$.

In each case, we have $g_{t\varphi} = g_{\varphi t} = 0$. Therefore f (or more precisely $df(t, \varphi)$ for all $(t, \varphi) \in I \times \mathbb{R}$) is *angle-preserving* or *conformal* in case b), and is *area-preserving* in case c).

■

Exercise 4.6.11

The *gradient* of a smooth function φ on M is the vector field grad φ, such that $\langle \operatorname{grad} \varphi(p), v\rangle = d\varphi(p)(v)$ for all p in M and $v \in T_pM$. Show: With respect to a chart (U, x) on M,

$$\operatorname{grad} \varphi = g^{ij} \frac{\partial \varphi}{\partial x^i} \frac{\partial}{\partial x^j} \quad \text{on } U. \tag{4.76}$$

■

Exercise 4.6.12

1. For a geodesic c, $\|\dot{c}\|$ is constant.
2. Up to reparameterization, the meridians of surfaces of revolution are geodesics. Under what conditions on the profile curve are they geodesics? Which latitudes are geodesics?
3. For $x, y \in \mathbb{R}^{m+1}$ such that $\|x\| = \|y\| = 1$ and $\langle x, y\rangle = 0$, the curve $c(t) = r\cos(t)x + r\sin(t)y$, $t \in \mathbb{R}$, is a geodesic on the sphere S^m_r as in Example 4.2.1.2. Up to reparameterization, every geodesic on S^m_r is of this form.
4. Let $c\colon I \longrightarrow \mathbb{R}^3$ be a space curve such that $\dot{c}$ and $\ddot{c}$ are pointwise linearly independent, and let $f = f(s, t) = c(t) + sb(t)$ be the ruled surface spanned by c and the binormal field b of c. Show: f is an immersion of $M = \mathbb{R} \times I$, and the curve $(t, 0)$, $t \in I$, is a geodesic in M.

■

Exercise 4.6.13

1. For a regular curve $c\colon I \longrightarrow M$ with field of directions $e := \dot{c}/\|\dot{c}\|$ in M, we call $\|\nabla\dot{c}/dt - \langle e, \nabla\dot{c}/dt\rangle e\|/\|\dot{c}\|^2$ the *geodesic curvature* of c; compare with (4.4). Show: Up to reparameterization c is a geodesic if and only if the geodesic curvature of c vanishes.
2. Let $f\colon M \longrightarrow \mathbb{R}^n$ be an immersion of the surface M and $c\colon I \longrightarrow M$ be a regular curve. Let n be one of the two vector fields along c with $\langle \dot{c}, n\rangle = 0$ and constant norm 1. The corresponding oriented geodesic curvature of c is then $\kappa_o := \langle \nabla\dot{c}/dt, n\rangle/\|\dot{c}\|^2$ with $e = \dot{c}/\|\dot{c}\|$. Prove that

$$\nabla e/dt = \|\dot{c}\|\kappa_o n \quad \text{and} \quad \nabla n/dt = -\|\dot{c}\|\kappa_o e.$$

3. Determine the oriented geodesic curvature of the latitudes of a surface of revolution as in Example 4.2.3.2. Discuss the special case of spheres.

■

Exercise 4.6.14

1. Let $M \subseteq \mathbb{R}^n$ be a linear subspace and let $c\colon I \longrightarrow M$ be a smooth curve. Show: A vector field $X\colon I \longrightarrow M$ along c is parallel in the sense of Definition 4.2.16 if and only if it is parallel in the usual sense.
2. Verify with the use of (4.37) that, with respect to the coordinates (t, φ), the Christoffel symbols of the surface of revolution with profile curve $c = (r, h)$ are given by

$$\Gamma^t_{tt} = \frac{\dot{r}\ddot{r} + \dot{h}\ddot{h}}{\|\dot{c}\|^2}, \quad \Gamma^t_{\varphi\varphi} = -\frac{r\dot{r}}{\|\dot{c}\|^2}, \quad \Gamma^\varphi_{\varphi t} = \Gamma^\varphi_{t\varphi} = \frac{\dot{r}}{r}$$

and $\Gamma^\varphi_{tt} = \Gamma^\varphi_{\varphi\varphi} = \Gamma^t_{\varphi t} = \Gamma^t_{t\varphi} = 0$. List the covariant derivatives of the basis fields $\partial/\partial t$ and $\partial/\partial\varphi$ as vector fields along the meridians and latitudes. Determine the space of parallel vector fields along these curves. Discuss the special case of spheres.
3. If the first fundamental form of an immersion $f\colon M \longrightarrow \mathbb{R}^n$ is in diagonal form with respect to a chart (U, x) on M, that is, if $g_{ij} \equiv 0$ for $i \neq j$, then $g^{ii} = g_{ii}^{-1}$, $g^{ij} = 0$ for $i \neq j$ and therefore

$$\Gamma^k_{ik} = \frac{1}{2g_{kk}}\frac{\partial g_{kk}}{\partial x^i} \quad \text{and, for } i \neq k, \quad \Gamma^k_{ii} = \frac{-1}{2g_{kk}}\frac{\partial g_{ii}}{\partial x^k}.$$

■

Exercise 4.6.15

Let $f\colon M \longrightarrow \mathbb{R}^3$ be an immersed surface together with a Gauß map $n\colon M \longrightarrow S^2$.

1. A smooth curve $c\colon I \longrightarrow M$ is a line of curvature if and only if there is a smooth function $\lambda\colon I \longrightarrow \mathbb{R}$ with $dn \circ \dot{c} = \lambda \cdot df \circ \dot{c}$.
2. (Joachimsthal's Theorem[18]) For $f_1\colon M_1 \longrightarrow \mathbb{R}^3$ and $f_2\colon M_2 \longrightarrow \mathbb{R}^3$, let c_1 in M_1 and c_2 in M_2 be curves with $f_1 \circ c_1 = f_2 \circ c_2 =: c$. Further let the intersection of f_1 and

---

[18] Ferdinand Joachimsthal (1818–1861).

f_2 along c be transversal, that is, $T_{c_1(t)}f_1 \neq T_{c_2(t)}f_2$ for all t. Show each two of the following statements implies the third:

(a) c_1 is a line of curvature;
(b) c_2 is a line of curvature;
(c) f_1 and f_2 intersect at a fixed angle along c.

3. We call $v \in T_pM$, $v \neq 0$, an *asymptotic direction* if $S^n_p(v, v) = 0$. Regular curves $c: I \longrightarrow M$, such that $\dot{c}(t)$ is an asymptotic direction for all $t \in I$, are called *asymptotes*. Check that the s-parameter lines on the ruled surface as in Example 4.3.12.3 are asymptotes. Note that $K(p) \leq 0$ if T_pM contains an asymptotic direction and that, up to collinearity, there are two asymptotic directions at p if $K(p) < 0$. ■

Exercise 4.6.16

1. The *catenoid* is the surface of revolution generated by the catenary,

$$f = f(t, \varphi) = (\cosh t \cos \varphi, \cosh t \sin \varphi, t).$$

Compare the first fundamental form of the catenoid with the (somewhat differently parameterized than above) helicoid

$$f = f(t, \varphi) = (\sinh t \cos \varphi, \sinh t \sin \varphi, \varphi).$$

Determine the second fundamental forms of the catenoid and the helicoid and show that they are *minimal surfaces*, that is, that their mean curvatures H vanish. Compute their Gaußian curvatures. Sketch both surfaces.

2. Classify surfaces of revolution with constant Gaußian curvature and surfaces of revolution that are minimal surfaces, that is, have mean curvature $H \equiv 0$.
3. Let $f: M \longrightarrow \mathbb{R}^3$ be an immersed surface. For $x \in \mathbb{R}^3$, suppose the function $r: M \longrightarrow \mathbb{R}$, $r(p) = \|f(p) - x\|$ has a maximum at $p_0 \in M$. Show that $K(p_0) \geq 1/r(p_0)^2$ and conclude that M has a point with positive Gaußian curvature, if M is compact. (Compare with Exercise 4.6.3.3.) ■

Exercise 4.6.17

Determine the curvature tensor of $G_k(V)$ and show that the sectional curvature of $G_k(V)$ is non-negative. ■

!"#$# % & & !' (! !) * * + , * - . *
/ * !F 3 * * * # $ * #3&* 3 * *
5 "/" & 6 13
/ : 8 :

!"#$# % & & !' (! !) * * + , * - . *
/ * !S + ,0 * + , 5 "/" & 6 13

+* :0 & :8

* : 3 * 5 86 1* * * 0 7* (10 1 *3 * & 0
* *Z[3 1 0 * *Z[3 * 1 * * 0 * *Z[3 1 . * * * *Z[3
* H & 1 *

F * & * (0 3 1 * S 7 - H ; * * 1* 3 1 . * * 3
< 3 N' AAA 80 H *J 3* ; * 3 4 3 1 * 3 ^ 1* * ^
- 1 *3 1
) 3* 3 Z* 1 3 1 *3

8 : - 1 *3 * * 4*3 * \ *. *

F * 3 * [3 * H & 3 1 H *

Book I

On the Straight Line

CHAPTER I

On Angles

10. The figure formed by two rays issuing from the same endpoint is called an angle. The point is called the vertex of the angle, and the two rays are its sides.

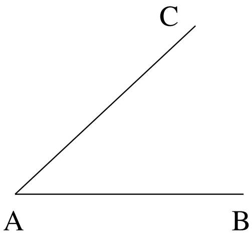

Figure 7

We denote an angle by the letter of its vertex, placed between two other letters which indicate the sides, often surmounted by a special symbol. If, however, the figure contains only one angle with the vertex considered, that letter will suffice to designate the angle. Thus, the angle formed by the two rays AB, AC (Fig. 7) will be denoted by BAC or, more simply, by A.

Two angles are said to be congruent, in agreement with the definition of congruent figures (3) if they can be superimposed by a rigid motion.[1]

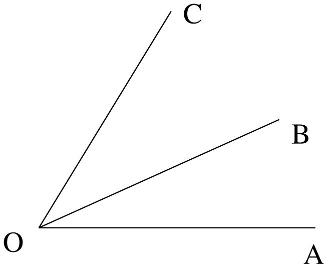

Figure 8

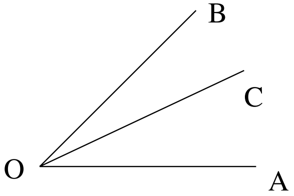

Figure 9

Two congruent angles, BAC, B′A′C′, can be superimposed in two different ways: namely, either AB takes the direction of A′B′ and AC the direction of A′C′, or the other way around. We pass from one to the other by turning the angle around onto itself, for instance, by moving the angle BAC in such a way that AC occupies the position previously occupied by AB, and vice-versa.

---

[1] But see the footnote to 3. transl.

11. We say that two angles are adjacent if they have the same vertex, a common side, and so that they are located on opposite sides of this common side.

When angles AOB, BOC are adjacent (Fig. 8), angle AOC is called the sum of the two angles.

The sum of two or more angles is independent of the order of the angles added.

To compare angles, we move them so they have a common vertex, and a common side, and lie on the same side of this common side. Assume that AOB, AOC are placed in this manner. If, rotating around point O, we encounter the sides in the order OA, OB, OC (Fig. 8), angle AOC is equal to the sum of AOB and another angle BOC; in this case angle AOC is said to be larger than AOB, and the latter is smaller than AOC; if, on the contrary (Fig. 9) the order is OA, OC, OB, then angle AOC is smaller than AOB. The angle BOC which, added to the smaller angle, yields the larger angle, is the difference of the two angles.

Finally, in the intermediate case in which OB coincides with OC, the two angles are congruent (see 10).

In the interior of every angle BAC there is a ray AM which divides this angle into two congruent parts; it is called the bisector of the angle. The rays contained in the angle BAM make a smaller angle with AB than with AC; the opposite is true for rays contained in the angle MAC.

An angle is said to be the double, triple, etc., of another if it is the sum of two, three, etc., angles equal to this other angle. The smaller angle will then be called a half, third, etc., of the larger one.

Remark. The size of an angle does not depend on the size of its sides, which are rays (5), and which we must imagine as extending indefinitely.

12. As we have said, an angle is determined by two rays, such as OA, OB (Fig. 10). If we extend OA past point O to form OA′, and in the same way extend OB to form OB′, we obtain a new angle A′OB′.

Two angles AOB, A′OB′ such that the sides of one are extensions of the sides of the other are called vertical angles.

Theorem. Two vertical angles are congruent.

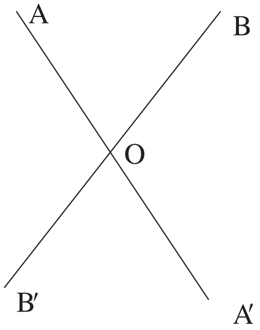

Figure 10

Indeed, let us turn (10) the angle BOA onto itself (Fig. 10). The side OB will take the place of OA , and, since 0A will take the former position of OB, this means that ray OA, which is the extension of OA , will coincide with OB , the extension of OB. Thus angle AOB is superimposed on angle A OB , and therefore these two angles are congruent.

13. Arcs and angles. Every ray issuing from the center of a circle meets this circle in one and only one point.

Every angle AOB (Fig. 4) with its vertex at the center O of a circle determines an arc $\overset{\frown}{AB}$ having as endpoints the intersections of the sides of the angle with the circle. In general, (see however 20b), we consider the arc which is less than a semicircle.

Conversely, every arc less than a semicircle can be thought of as determined by an angle with its vertex at the center of the circle, formed by the rays passing through the endpoints of the arc.

Theorem. On the same circle, or on equal circles:

1 . To equal arcs (smaller than a semicircle) there correspond equal central angles;[2]

2 . To unequal arcs (smaller than a semicircle) there correspond unequal central angles, and the greater angle corresponds to the greater arc;

3 . If an arc (smaller than a semicircle) is the sum of two others, the corresponding central angle is the sum of the angles associated with the smaller arcs.

1 , 2 . Let $\overset{\frown}{AB}$, $\overset{\frown}{AC}$ (Fig. 4) be the two arcs, drawn on the same circle, starting from the same point A, in the same direction (8). The two angles AOB, AOC are therefore placed as in 11. But then rays OA, OB, OC are placed in the same order as points A, B, C on the circle. Moreover, if rays OB, OC coincide, then so do the points B, C, and conversely.

3 . Let us recall that the sum of two arcs (8b) is formed by placing them as arcs $\overset{\frown}{AB}$, $\overset{\frown}{BC}$ are placed (Fig. 4). Then the angle AOC corresponding to the sum of these arcs will be the sum of AOB and BOC, because these angles are adjacent.

According to this result, in order to compare angles, one can draw circles with the same radius, centered at the vertices of the angles, and compare the arcs intercepted on these circles.

The division of angles into two or more equal parts corresponds to the division into equal parts of the corresponding arc of a circle centered at the vertex of the angle.

14. Perpendiculars. Right angles. We say that two lines are perpendicular to each other if, among the four angles they form, two adjacent angles are equal to each other. For example, line AOA (Fig. 11) is perpendicular to BOB if the angles numbered 1 and 2 in the figure are equal. In such a case, the four angles at O are equal to each other, since angles 3 and 4 (Fig. 11) are equal respectively to 1 and 2, because they are vertical angles.

An angle whose sides are perpendicular is called a right angle.

---

[2] This is the standard term in American texts. Hadamard uses the locution angles with their vertex at the center or angles at the center . transl.

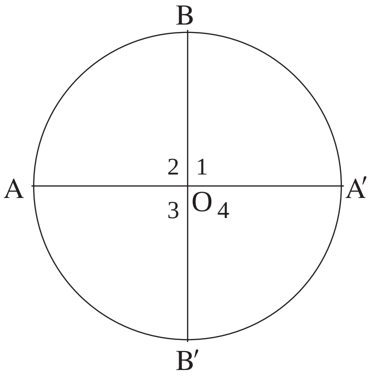

Figure 11

Theorem. In a given plane, through a given point on a line, one can draw one, and only one, perpendicular to this line.

Suppose we want to draw a perpendicular to line AA′ through point O (Fig. 11). It suffices to draw a circle with center O which cuts the line at A and A′, then find point B which divides semicircle AA′ into two equal parts. Then OB will be the required perpendicular; and conversely, a perpendicular to AA′ drawn through point O must divide arc $\overset{\frown}{AA'}$ in half.

Corollary. We see that a right angle intercepts an arc equal to one fourth of a circle centered at the vertex of the angle.

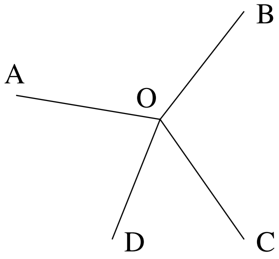

Figure 12

All right angles are equal, since two such angles will intercept equal arcs on equal circles centered at their vertices.

15. If several rays are drawn issuing from the same point, the sum of the successive angles thus formed (AOB, BOC, COD, DOA, Fig. 12) is equal to four right angles.

Indeed, the sum of the arcs intercepted by these angles on a circle centered at their vertex is the entire circle.

If several rays are drawn issuing from a point on a line, in such a way that all the rays are on one side of the line (Fig. 13), the sum of the successive angles thus formed is equal to two right angles.

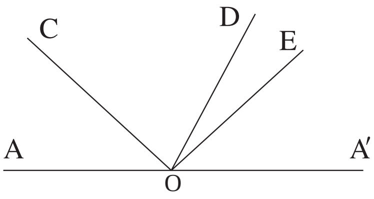

Figure 13

Indeed, the sum of the intercepted arcs is a semicircle.

Conversely, if two or more angles with the same vertex, and each adjacent to the next, (AOC, COD, DOE, EOA′, Fig. 13) have a sum equal to two right angles, then their outer sides form a straight line.

Indeed, these outer sides cut a circle centered at the common vertex of the angles at two diametrically opposite points, since the arc they intercept is a semicircle.

Remark. The angle AOA′ (Fig. 13) whose sides form a straight line is called a straight angle.

15b. Theorem. The bisectors of the four angles formed by two concurrent lines form two straight lines, perpendicular to each other.

Let AA′, BB′ (Fig. 14) be two lines which intersect at O and form the angles AOB, BOA′, A′OB′, B′OA, whose bisectors are Om, On, Om′, On′. We claim:

1. That Om, Om′ are collinear, as are On, On′.

2. That the two lines thus formed are perpendicular.

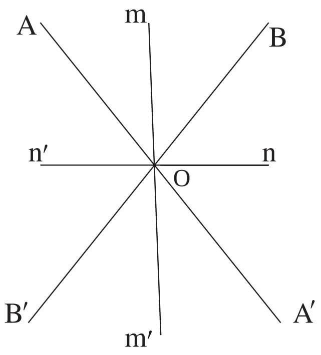

Figure 14

First, Om is perpendicular to On because, since the sum of AOB and BOA′ is two right angles, half of each, mOB and BOn, must add up to a right angle.

Applying the same reasoning to angles A′OB′ and B′OA, we see that Om′ and On are perpendicular. Therefore Om′ is the extension of Om, and likewise On′ is the extension of On.

16. An angle which is less than a right angle is called an acute angle; an angle greater than a right angle is called an obtuse angle.

Two angles are said to be complementary if their sum is a right angle; supplementary if their sum is equal to two right angles.

17. Angle Measure. The ratio of two quantities of the same kind is[3] the number which expresses how many times one of the quantities is contained in the other. For instance, if, in dividing the segment AB into five equal parts, one of these parts is contained exactly three times in the segment BC, then the ratio of BC to AB is said to be equal to 3/5. If, on the other hand, a fifth of AB is not contained an exact number of times in BC, for example if it is contained in BC more than three times, but less than four times, then 3/5 would be an approximate value of the ratio $\frac{BC}{AB}$: it would be within one fifth less than the value of the ratio (the value 4/5 would be within 1/5 more).

The ratio of two quantities a, b of the same kind is equal to the ratio of two other quantities a', b' also of the same kind (but not necessarily of the same kind as the first two) if, for every n, the approximation to within $1/n$ of the first ratio is equal to the approximation within $1/n$ of the second ratio.

The measure of a quantity, relative to a quantity of the same kind chosen as unit, is the ratio of the given quantity to the unit.

In addition we have the following properties (Leçons d'Arithmetique by J. Tannery):

1°. Two quantities of the same kind, which have the same measure relative to the same unit, are equal;

2°. The ratio of two quantities of the same kind is equal to the ratio of the numbers which serve as their measures relative to the same unit;

3°. The ratio of two numbers is the same as the quotient of the two numbers; etc.

Theorem. In the same circle, or in equal circles, the ratio of two central angles is equal to the ratio of their subtended arcs.

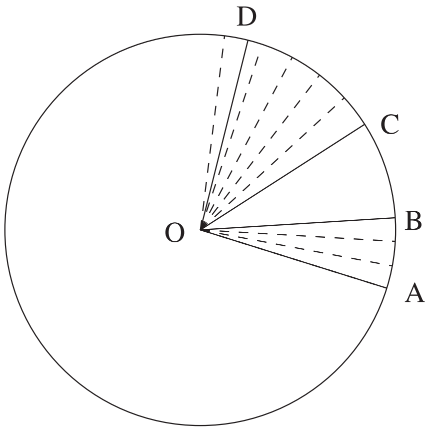

Figure 15

---

[3] See Leçons d'Arithmetique by J. Tannery, chapters X and XIII.

Consider[4] (Fig. 15) two arcs $\overset{\frown}{AB}$, $\overset{\frown}{CD}$ on circle O. Divide central angle COD into three equal parts (for example), and suppose that one of these thirds can fit four, but not five times, into the angle AOB; the value (smaller by less than 1/3) of the ratio $\frac{AOB}{COD}$ is $\frac{4}{3}$.

But dividing the angle COD into three equal parts, we have at the same time divided arc $\overset{\frown}{CD}$ into three equal parts (13). If a third of COD fits four, but not five times, into angle AOB, it follows the same way that a third of arc $\overset{\frown}{CD}$ fits four, but not five times, into arc AB. The values to within 1/3 of the two ratios are therefore equal, and similarly the values to within $1/n$ will be equal for any integer n. The theorem is therefore proved.

Corollary. If we take as the unit of angle measure the central angle which intercepts a unit arc, then every central angle will have the same measure as the arc contained between its sides.

This statement is exactly the same as the preceding one, since the measure of a quantity is its ratio to the unit.

Supposing, as we will from now on, that on each circle we choose as unit arc an arc intercepted by a unit central angle, the preceding corollary can be stated in the abridged form: A central angle is measued by the arc contained between its sides.

18. The definitions reviewed above allow us to establish an important convention.

From now on, we will be able to suppose that all the quantities about which we reason have been measured with an appropriate unit chosen for each kind of quantity. Then, in all the equations we write, the quantities which appear on the two sides of an equality will represent not the quantities themselves, but rather their measures.

This will allow us to write a number of equations which otherwise would have no meaning. For instance, we might equate quantities of different kinds, since we will be dealing only with the numbers which measure them, the meaning of which is perfectly clear. We will also be able to consider the product of any two quantities, since we can talk about the product of two numbers, etc.

In fact, whenever we write the equality of two quantities of the same kind, this equality will have the same meaning as before, since the equality of these quantities is the same as the equality of their measures.

Following this convention we can write:

$$AOB = \text{arc } \overset{\frown}{AB},$$

where AB is an arc of a circle and O is the center. It is important however to insist on the fact that the preceding equality assumes that the unit of angle measure and the unit of arc measure have been chosen in the manner specified above.

---

[4] This theorem becomes obvious if we consider the following proposition from arithmetic (Tannery, Leçons d'Arithmetique: Two quantities are proportional if: 1 To any value of the first, there corresponds an equal value of the second, and 2 the sum of two values of the first corresponds to the sum of the corresponding values of the second. These two conditions hold here (13).

The argument in the text does no more than reproduce, for this particular case, the proof of this general theorem of arithmetic.

18b. Traditionally, the circle has been divided into 360 equal parts called degrees, each of which contains 60 minutes, which are themselves divided into 60 seconds. One can then measure arcs in degrees and, correspondingly, angles will also be measured in degrees, and the number of degrees, minutes, and seconds of the angle will be the same as the number of degrees, minutes, and seconds of the arc intercepted by this angle on a circle centered at its vertex. A right angle corresponds to one fourth of a circle; that is, to 90 degrees. It follows that the measure of an angle at the center of a circle does not depend on the radius of the circle on which one measure the arcs, since the chosen unit of angle measure (the degree) has a value independent of this radius, namely one ninetieth of a right angle.

In writing angles (or arcs) in degrees, minutes, and seconds, we use an abridged notation: an angle of 87 degrees, 34 minutes, and 25 seconds is written: 87 34 25 .

The introduction of decimal notation in all other kinds of measurements has led to the use of another mode of division, in which the circle is divided into 400 equal parts called gradients (or grads). The grad, a little smaller than the degree, is, as we see, one hundredth of a right angle. The grad is divided decimally, so there is actually no need for special names for its parts, which are written using the rules of decimal notation. Thus we can speak of the angle $3^G.5417$ (that is, 3 grads and 5417 ten-thousandths).

Nonetheless, a hundredth of a grad is often called a centesimal minute, and is indicated by the symbol ‘ (to distinguish it from the hexagesimal minute, which is a sixtieth of a degree). Likewise, one hundredth of a centesimal minute is called a centesimal second, denoted by the symbol ‘‘. The angle just considered could then be written as $3^G.54$‘17‘‘.

A grad is equal to $\frac{360}{400}$, that is $\frac{9}{10}$ of a degree, or 54 . A degree equals $\frac{400}{360}^G$ = $1^G.11$‘11‘‘1... (in other words, $\frac{10}{9}$ of a grad).

19. Theorem. Given a line, and a point not on the line, there is one, and only one, perpendicular to the line passing through the given point.

1 . There is one perpendicular. Consider point O and line xy (Fig. 16). Using xy as a hinge, let us turn the half-plane containing O until it overlaps the other half-plane. Supposing point O falls on O , we join OO . This line intersects xy because it joins two points in different half-planes. If I is the point of intersection, the angles O Ix and OIx are equal because one of them can be superimposed on the other by rotation around xy. Thus the lines xy and OO are perpendicular.

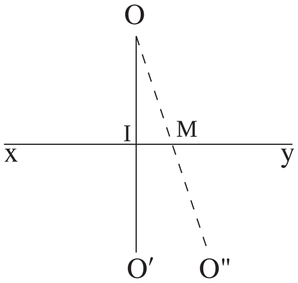

Figure 16

2 . There is only one perpendicular. Indeed, assume that OM is a perpendicular to xy passing through O. Extend the segment OM by its own length to MO .

Again turning the half-plane containing O around xy, the segment MO will fall on MO because the angles OMx and O Mx are right angles, and hence are equal. Since MO = MO, the point O falls on O, and therefore O and O coincide, as do the lines OO and OO. QED

19b. The reflection of a point O in a line xy is the endpoint of the perpendicular from O to the line, extended by the length of a segment equal to itself. It follows from the preceding considerations that this reflection is none other than the new position occupied by O after a rotation around xy.

Given an arbitrary figure, we can take the reflection of each of its points. The set of these reflections constitutes a new figure, called the reflection of the first. We see that in order to obtain the reflection of a given figure in line xy, we can turn the plane of the figure around xy, so that each half-plane determined by the line falls on the other, then note the position taken by the original figure. It follows that:

Theorem. A plane figure is congruent to its reflection.

Corollary. The reflection of a line is a line.

When a figure coincides with its reflection in line xy, we say that it is symmetric in this line, or that this line is an axis of symmetry of the figure.

20. To make a figure F coincide with its reflection F, we had to use a motion which took the figure out of its plane. We must note that this superposition is not possible without such a movement; this holds because the sense of rotation is reversed in the two figures. We will now explain what this means.

First, let us remark that the plane of the figure divides space into two regions. For brevity, let us call one of these the region situated above the plane; the other, the region below the plane.

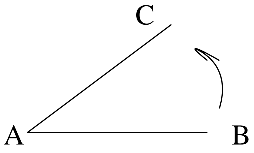

Figure 17

Consider now an angle BAC in the figure F, which can be viewed as being described by a ray moving inside this angle from the position AB to the position AC (Fig. 17). Viewed from above this angle BAC will be said to have an inverse sense of rotation or a direct sense, according as the moving ray turns clockwise or counterclockwise.[5] To be definite, let us consider the second situation. In this case, an observer lying along AB, with his feet at A, his head in the direction of B, and looking down, will see the side AC on his left; and therefore, if still lying on AB, he faces AC, the region below the plane will be on his right.

It is clear that to describe the sense of rotation of an angle viewed from below, this discussion can be repeated, with the word above replaced by the word below, and vice-versa.

---

[5] We note that in order to define the sense of rotation, we must take into account the order of the sides. Thus angle BAC has the opposite sense from CAB.

Since an observer lying on AB and facing AC will necessarily have the region above the plane on his left, if the region below is on his right, and vice-versa, we see that the sense of rotation changes depending on whether we view the angle from one side of the plane or from the other.[6]

Let us now suppose that an angle is moved in any way at all, without ever leaving the plane. An observer moving along with the angle will not change his position relative to the regions of space above and below the plane, and therefore the sense of rotation is not altered by a motion which does not leave the plane.

To prove that such a motion cannot make a figure F coincide with its reflection F′, it suffices therefore to see that the senses of rotation of the two figures are opposite. But we have seen that we can make F coincide with F′ by turning the plane around xy (19). As a result of this rotation, the points above the plane are moved below it, and vice-versa. The sense of rotation of an angle in F, viewed from above, is therefore the same as the sense of F′ viewed from below, so that the two angles, viewed from the same side, have opposite senses of rotation. QED

20b. **Remarks.** I. We say that a plane is oriented when a sense of rotation for angles has been chosen as the direct sense. According to the preceding paragraph, orienting a plane amounts to deciding which region of space will be said to be above the plane.

II. It is convenient to regard an angle with vertex O as being described by a ray starting from O which at first coincides with the first side, and then turns about O in the plane (in the direct or inverse sense) until it coincides with the second side. If the ray has made a quarter of a complete turn, the angle is a right angle. If it has made one half of a complete turn, the angle is a straight angle, such as AOA′ (Figures 11 or 13).

Nothing prevents us, by the way, from considering angles greater than two right angles, since our ray can make more than half a complete turn.

III. Clearly, an arc of a circle, like an angle, can have a direct or an inverse sense, which depends, of course, on the order in which one gives its endpoints.

Two points A, B on a circle divide it into two different arcs of which (unless the two points are diametrically opposite) one is a minor arc (less than a semicircle), and the other a major arc. It must be noted that, since the endpoints A, B are given in a certain order (for example if A is first and B is second), these two arcs have opposite senses.

Exercises

Exercise 1. Given a segment AB and its midpoint M, show that the distance CM is one half the difference between CA and CB if C is a point on the segment. If C is on line AB, but not between A and B, then CM is one half the sum of CA and CB.

Exercise 2. Given an angle AOB, and its bisector OM, show that angle COM is one half the difference of COA and COB if ray OC is inside angle AOB; it is the supplement of half the difference if ray OC is inside angle A′OB′ which is vertical to AOB; it is one half the sum of COA and COB if OM is inside one of the other angles AOA′ and BOB′ formed by these lines.

---

[6] For this reason, writing viewed through a transparent sheet of paper appears reversed.

Exercise 3. Four rays OA, OB, OC, OD issue from O (in the order listed) such that AOB = COD and BOC = DOA. Show that OA and OC are collinear, as are OB and OD.

Exercise 4. If four consecutive rays OA, OB, OC, OD are such that the bisectors of angles AOB, COD are collinear, as are the bisectors of BOC, AOD, then these rays are collinear in pairs.

CHAPTER II

On Triangles

21. A plane region bounded by line segments is called a polygon (Fig 18). The segments are called the sides of the polygon, and their endpoints are called its vertices.

In fact, we will generally use the term polygon only for a portion of the plane bounded by a single contour which can be drawn with a single continuous stroke. The plane region which is shaded in Figure 19 will not be a polygon for us.

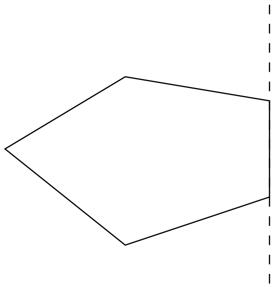

Figure 18

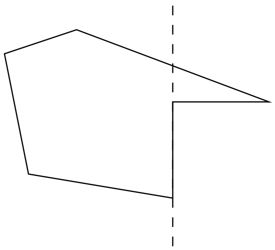

Figure 18b

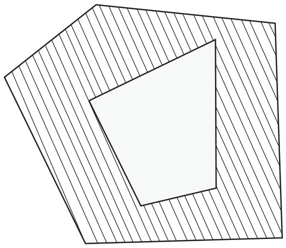

Figure 19

A polygon is said to be convex (Fig. 18) if, when we extend any side indefinitely, none of the lines thus formed crosses the polygon. In the opposite case (Fig. 18b) the polygon is said to be concave.

Polygons are classified according to the number of sides they have. Thus the simplest polygons are: the polygon with three sides or triangle, the polygon with 4 sides or quadrilateral, the polygon with 5 sides, or pentagon, the polygon with six sides or hexagon. We will also consider polygons with 8, 10, 12, and 15 sides, called octagons, decagons, dodecagons, and pentadecagons.

Any segment joining two non-consecutive vertices of a polygon is called a diagonal.

Remark. More generally, an arbitrary broken line, even one whose sides cross (as in Fig. 19b), will sometimes be called a polygon. In this case, when the broken line does not bound a unique region, the polygon is said to be improper. However, when we want to specify that the polygon is of the first type (Figures 18 and 18b), and not like the one in Figure 19b, we will speak of a proper polygon.

22. Among triangles we distinguish in particular:

The isosceles triangle. This name is used for a triangle with two equal sides. The common vertex of the equal sides is called the vertex of the isosceles triangle, and the side opposite the vertex is called the base;

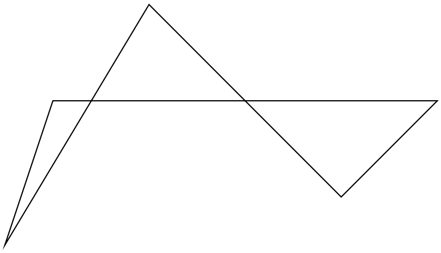

Figure 19b

The equilateral triangle, which has all three sides equal;

The right triangle, which has a right angle. The side opposite the right angle is called the hypotenuse.[1]

A perpendicular dropped from a vertex of a triangle onto the opposite side is called an altitude of the triangle; a median is a line which joins a vertex with the midpoint of the opposite side.

23. Theorem. In an isosceles triangle, the angles opposite the equal sides are equal.

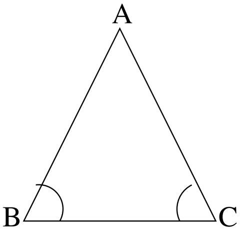

Figure 20

Suppose ABC is an isosceles triangle (Fig. 20). Let us turn angle BAC onto itself (10) so that AB lies along the line of AC and vice-versa. Since AB and AC are equal, point B will fall on C, and C on B. Angle ABC will therefore fall on ACB, so that these angles are equal. QED

Converse. If two angles of a triangle are equal, the triangle is isosceles.

In triangle ABC, suppose B = C. Let us move this triangle so that side BC is turned around, switching points B and C. Since ABC = ACB, side BA will assume the direction of of CA, and conversely. Point A, which is the intersection of BA and CA, will therefore retain its original position, so that segment AB falls on segment AC. QED

Corollary. An equilateral triangle is also equiangular (that is, its three angles are equal), and conversely.

---

[1] The original text does not use term leg for a side of a right triangle that is not the hypotenuse. However, we will use this standard English term freely in the translation. transl.

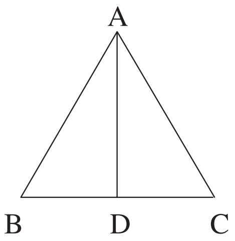

Figure 21

Theorem. In an isosceles triangle, the bisector of the angle at the vertex is perpendicular to the base, and divides it into two equal parts.

In isosceles triangle ABC (Fig. 21), let AD be the bisector of A. In turning angle BAC around on itself, this bisector does not move, and therefore neither does the point D where this bisector cuts the base. Segment DB falls on DC and angle ADB on ADC. Therefore DB = DC and ADB = ADC. QED

Remark. In triangle ABC we can consider:

1. The bisector of A;
2. The altitude from A;
3. The median from the same point;
4. The perpendicular to the midpoint of BC.

In general, these four lines are distinct from each other (see Exercise 17). The preceding theorem shows that, in an isosceles triangle, all of these are the same line, which is a line of symmetry of the triangle (19b).

This theorem can thus be restated: the altitude of an isosceles triangle is also an angle bisector and a median; or the median of an isosceles triangle is at the same time an angle bisector and an altitude; the perpendicular bisector of the base passes through the vertex and bisects the angle at the vertex.

Corollary. In an isosceles triangle, the altitudes dropped from the endpoints of the base are equal; the same is true about the medians from the endpoints of the base, about the bisectors of the angles at these points, etc., because these segments are symmetric to each other.

24. The following propositions, known under the name of cases of congruence of triangles, give necessary and sufficient conditions for two triangles to be congruent.

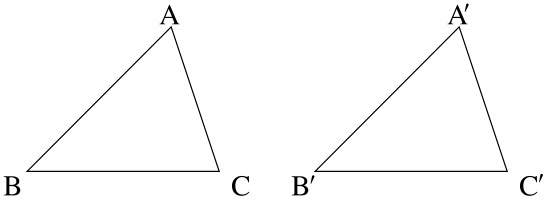

Figure 22

1st case: ASA. Two triangles are congruent if they have an equal side contained between corresponding equal angles.

Suppose the triangles are ABC, A B C (fig. 22) in which BC = B C , B = B , and C = C . Let us move angle B onto angle B so that side B A lies along the line of BA and B C along the line of BC. Since BC = B C , the point C falls on C. Now, since C = C, side C A assumes the direction of CA, and therefore point A falls on the intersection of BA and CA; that is, on A. This establishes the congruence of the two figures.

2nd case: SAS. Two triangles are congruent if they have an equal angle contained between corresponding equal sides.

Suppose the two triangles are ABC, A B C (Fig. 22) in which A = A , AB = AC, and AC = A C .

Let us move angle A over angle A so that A B assumes the direction of AB and A C assumes the direction of AC. Since A B = AB, point B will fall on B, and likewise C on C. Therefore side B C coincides with BC, and the superposition of the two figures is complete.

3rd case: SSS. Two triangles are congruent if they have three equal corresponding sides.

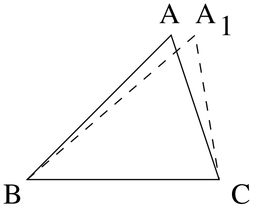

Figure 23

Consider triangles ABC and A B C such that AB = A B , AC = A C , and BC = B C , and let us move the second triangle so that side B C coincides with BC, and the two triangles are on the same side of line BC. Denote the new position of A by A_1. We claim that point A_1 coincides with A. This is obvious if BA_1 has the same direction as BA or if CA_1 has the same direction as CA. If this were not the case, we would have formed isosceles triangles BAA_1 and CAA_1 (Fig. 23) and the perpendicular bisector of AA_1 would have to pass through B and C (23, Corollary); in other words, this bisector would have to be line BC. This however is not possible, because points A and A_1 are on the same side of BC, so that BC cannot pass through the midpoint of AA_1. Thus the only possibility is that points A and A_1 must be the same point; in other words, the given triangles coincide.

Remarks. I. In order to establish that $A = A_1$, we have investigated what would happen if these two points were distinct. Arriving, in this case, at a conclusion which is clearly false, we concluded that this possibility does not arise. This is a very useful method of reasoning, called proof by contradiction.

II. In a triangle there are six main elements to consider; namely, the three angles and the three sides. We have seen that if we establish the equality of three of these elements (properly chosen) in two triangles, we can conclude that the two triangles are congruent and, in particular, that the remaining elements are equal as well.

III. Two congruent triangles (or, more generally, polygons) may differ in their sense of rotation (20). In this case, they can only be superimposed by a motion outside the plane. On the other hand, if the sense of rotation is the same, the two polygons can be superimposed by moving them within the plane, as we will have occasion to see later on.

25. The angle formed by a side of a convex polygon and the extension of the next side is called an exterior angle of the polygon.

Theorem. An exterior angle of a triangle is greater than either of the non-adjacent interior angles.

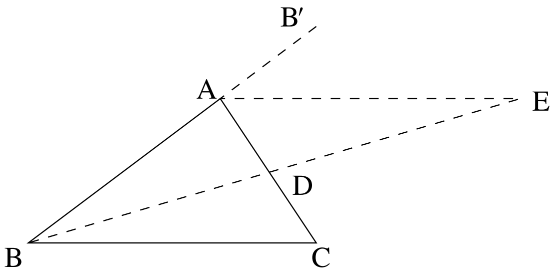

Figure 24

Let the triangle be ABC, and construct the exterior angle B AC (Fig. 24). We claim that this angle is greater, for example, than the interior angle C. To show this, we construct median BD, which we extend by its own length to a point E. The point E being inside[2] angle B AC, this last must be greater than angle EAC.

But this last angle is precisely equal to C; indeed, triangles DAE, DBC are congruent, having an equal angle between equal sides: the angles at D are equal because they are vertical angles, and $AD = DC$, $BD = DE$ by construction. Thus exterior angle B AC is greater than interior angle C. QED

Exterior angle B AC is the supplement of interior angle A. Since angle C is smaller than the supplement of A, the sum $A + C$ is less than two right angles. Our theorem can thus be restated: The sum of any two angles of a triangle is less than two right angles. In particular, a triangle cannot have more than one right or obtuse angle.

Theorem. In any triangle, the greater angle lies opposite the greater side.

In triangle ABC, suppose $AB > AC$. we will show that $C > B$. To show this, we take, on AB, a length $AD = AC$ (Fig. 25). It follows from the hypothesis that point D is between A and B, and therefore angle ACD is less than C. But in

---

[2] Point E is on the same side of line BAB as D (otherwise line DE, which has the common point B with BAB , would have to cut it again between D and E, and this is impossible), and thus on the same side as C. But B and E are on different sides of AC, which BE crosses at D.

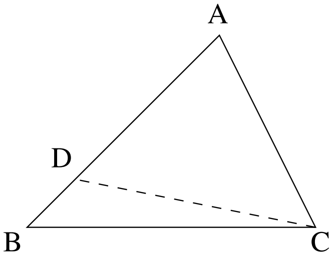

Figure 25

isosceles triangle ACD, $ACD = ADC$, which is greater than B by the preceding theorem applied to triangle DCB. The theorem is thus proved.

Conversely, the greater side corresponds[3] to the greater angle.

This statement is obviously equivalent to the preceding one.

26. *Theorem.* Any side of a triangle is less than the sum of the other two.

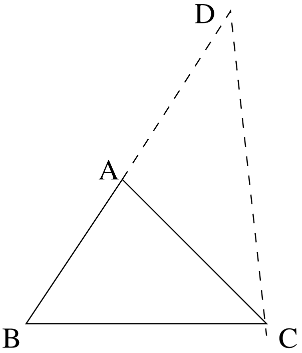

Figure 26

In triangle ABC, we extend side AB to a point D such that $AD = AC$ (Fig. 26). We must prove that[4] $BC < BD$. Drawing CD, we see that angle D, which is equal to angle ACD (23), is therefore less than BCD.

The desired inequality thus follows from the preceding theorem applied to the triangle BCD.

Corollaries. I. Any side of a triangle is greater than the difference between the other two.

Indeed, the inequality $BC < AB + AC$ gives, after subtracting AC from both sides:

$$BC - AC < AB.$$

---

[3] The side corresponding to an angle is the side opposite it.

[4] The theorem is obvious if BC is not the largest side of the triangle.

II. For any three points A, B, C, each of the distances AB, BC, AB is at most equal to the sum of the other two, and at least equal to the difference of the other two, equality holding if the three points are collinear.

Theorem. A line segment is shorter than any broken line with the same endpoints.

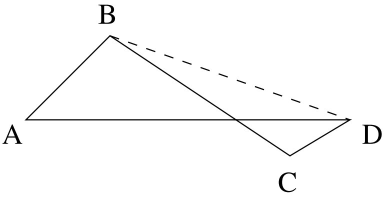

Figure 27

If the broken line has only two sides, the theorem reduces to the preceding one. Consider next a broken line with three sides ABCD (Fig. 27). Drawing BD, we will have $AD < AB + BD$ and, since $BD < BC + CD$, we have

$$AD < AB + BD < AB + BC + CD.$$

The theorem is thus proven for a broken line with three sides. The same reasoning can be used successively for broken lines with 4 sides, 5 sides, etc. Therefore the theorem is true no matter how large the number of sides.

27. The sum of the sides of a polygon, or of a broken line, is called its perimeter.

Theorem. The perimeter of a convex broken line is less than that of any broken line with the same endpoints which surrounds it.

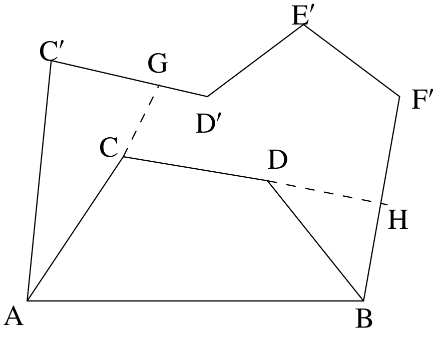

Figure 28

Consider the convex line ACDB and the surrounding line AC D E F B (Fig. 28). We extend sides AC, CD in the sense indicated by ACDB; that is, the side AC past C and the side CD past D. These extensions do not intersect the interior of polygon ACDB because of its convexity. Suppose they cut the surrounding line in G and H respectively.

The path ACDB is shorter than ACHB because they have a common portion ACD, and the remainder BD of the first is less than the remainder DHB of the second. In turn, the path ACHB is shorter than AGD E F B because, after removing the common parts AC, HB, the segment CH which is left is smaller than the broken line CGD E F H. In the same way, AGD E F B is less than AC D E F B, because AG is less than AC G. Thus

ACDB < ACHB < ACGD E F B < AC D E F B.

QED

Corollary. The perimeter of a convex polygon is less than the perimeter of any closed polygonal line surrounding it completely.

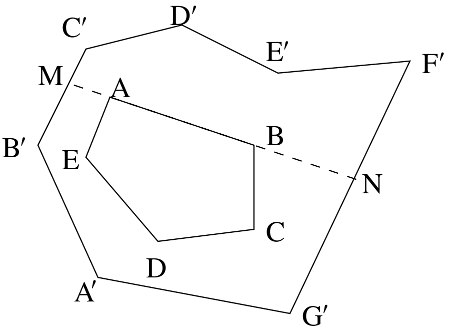

Figure 29

Consider (Fig. 29) convex polygon ABCDE, and polygonal line A B C D E F G A which surrounds it completely. We extend side AB in both directions until it intersects the surrounding polygon at M, N. By the preceding result, the length of the path AEDCB is less than AMB A G NB, and therefore the perimeter of AEDCBA is less than the perimeter of the polygon NMB A G N. This perimeter, in turn, is less than the surrounding line because the part MB A G N is common to both, and MN < MC D E F N.

28. Theorem. If two triangles have a pair of unequal angles contained by two sides equal in pairs, and the third sides of the triangles are unequal, then the greater side is opposite the greater angle.

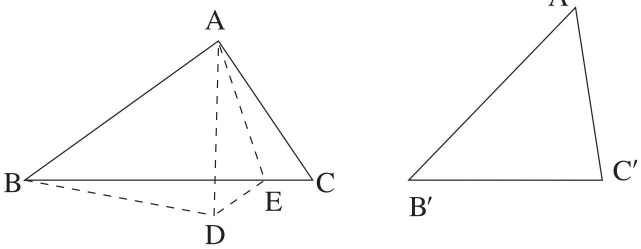

Figure 30

Consider triangles ABC, $A'B'C'$ such that $AB = A'B'$, $AC = A'C'$, and $A > A'$ (Fig. 30). We want to prove that $BC > B'C'$. Let us move the second triangle so that $A'B'$ coincides with AB. Since A' is less than A, side $A'C'$ will move to a position AD in the interior of angle BAC. We construct the bisector AE of DAC. This segment is also inside BAC, and, as B and C lie on different sides of this line, it must intersect side BC in some point E located between B and C. If we draw DE, we see that triangles ACE and ADE are congruent, because they have an equal angle (with vertex A) between two corresponding equal sides (AE in common, $AC = A'C' = AD$). Therefore $DE = EC$. The inequality $BD < BE + ED$ provided by the triangle BDE then gives us

$$BD < BE + EC,$$

or

$$BD < BC.$$

QED

Conversely. If, in two triangles, two pairs of sides are equal, but the third sides are unequal, then the angles opposite the unequal sides are unequal, and the greater angle is opposite the greater side.

This statement is equivalent to the preceding one.

Remark. The preceding theorem does not require the third case of congruence for triangles (SSS). It therefore provides another proof of that case.

Indeed, if we have $AB = A'B'$, $AC = A'C'$ and, in addition $BC = B'C'$, the angles A and A' would have to be equal, or else BC and $B'C'$ could not be equal. Knowing now that $A = A'$ the two triangles are congruent by the second case (SAS).

Exercises

Exercise 5. Prove that a triangle is isosceles:

1. if an angle bisector is also an altitude;
2. if a median is also an altitude;
3. if an angle bisector is also a median.

Exercise 6. On side Ox of some angle, we take two lengths OA, OB, and on side Ox' we take two lengths OA', OB', equal respectively to the first two lengths. We draw AB', BA', which cross each other. Show that point I, where these two segments intersect, lies on the bisector of the given angle.

Exercise 7. If two sides of a triangle are unequal, then the median between these two sides makes the greater angle with the smaller side. (Imitate the construction in 25.)

Exercise 8. If a point in the plane of a triangle is joined to the three vertices of a triangle, then the sum of these segments is greater than the semi-perimeter of the triangle; if the point is inside the triangle, the sum is less than the whole perimeter.

Exercise 8b. If a point in the plane of a polygon is joined to the vertices of the polygon, then the sum of these segments is greater than the semi-perimeter of the polygon.

Exercise 9. The sum of the diagonals of a [convex[5]] quadrilateral is between the semi-perimeter and the whole perimeter.

Exercise 10. The intersection point of the diagonals of a [convex[6]] quadrilateral is the point in the plane such that the sum of its distances to the four vertices is as small as possible.

Exercise 11. A median of a triangle is smaller than half the sum of the sides surrounding it, and greater than the difference between this sum and half of the third side.

Exercise 12. The sum of the medians of a triangle is greater than its semi-perimeter and less than its whole perimeter.

Exercise 13. On a given line, find a point such that the sum of its distances to two given points is as small as possible. Distinguish two cases, according to whether the points are on the same side of the line or not. The second case can be reduced to the first (by reflecting part of the figure in the given line).

Exercise 14. (Billiard problem.) Given a line xy and two points A, B on the same side of the line, find a point M on this line such that $AMx = BMy$.

We obtain the same point as in the preceding problem.

Exercise 15. On a given line, find a point with the property that the difference of its distances to two given points is as large as possible. Distinguish two cases, according to whether the points are on the same side of the line or not.

---

[5] This condition is not in the original text. See solution. transl.

[6] See the previous footnote. transl.

CHAPTER III

Perpendiculars and Oblique Line Segments

29. Theorem. If, from a given point outside a line, we draw a perpendicular and several oblique line segments:

1. The perpendicular is shorter than any oblique segment;

2. Two oblique segments whose feet are equally distant from the foot of the perpendicular are equal;

3. Of two oblique segments, the longer is the one whose foot is further from the foot of the perpendicular.

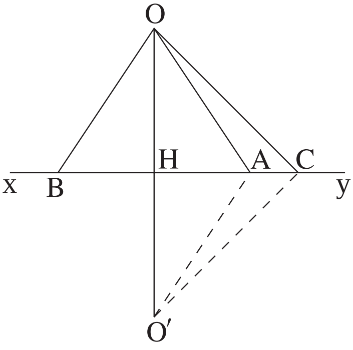

Figure 31

1. Consider perpendicular OH and oblique segment OA from point O to line xy (Fig. 31). If we extend OH by a length HO' equal to itself, then O' is symmetric to O with respect to line xy. Therefore $O'A$ is the symmetric image of OA, and the two are equal. Now in triangle $OO'A$ we have

$$OO' < OA + O'A,$$

and we can replace OO' with $2OH$ and $OA + O'A$ with $2OA$. Thus we find $2OH < 2OA$, or $OH < OA$.

2. Consider next the oblique segments OA, OB such that $HA = HB$. These two oblique segments will be equal by symmetry with respect to line OH.

3. Finally, let OA and OC be oblique segments such that $HC > HA$ (Fig. 31). Suppose first that points A and C are on the same side of H. Then point A is inside triangle $OO'C$. By (27) we have

$$OA + O'A < OC + O'C$$

and, as we saw before, $OA = O'A$ and $OC = O'C$. Dividing by two, we have, as before,

$$OA < OC.$$

If we had started with an oblique segment OB less distant than OC, but on the other side of H, it would suffice to construct a length HA = HB in the direction of HC. The oblique segment OB then would be equal to OA (2°), and hence smaller than OC as we have seen above.

30. Conversely. If two oblique segments are equal, their feet are equally distant from the foot of the perpendicular, otherwise they are unequal; if they are unequal, the longer is the more distant from the perpendicular.

Corollary. There are no more than two oblique segments of the same length from the same point O to a line xy.

This is true because the feet of these oblique segments are equally distant from H, and there are only two points on xy at a given distance from H.

31. The length of the perpendicular dropped from a point to a line is called the distance from the point to the line. The preceding theorem shows that this perpendicular is in fact the shortest path from the point to the line.

32. Theorem.

1°. Every point on the perpendicular bisector of a segment is equally distant from the endpoints of the segment;

2°. A point not on the perpendicular bisector is not equally distant from the endpoints of the segment.

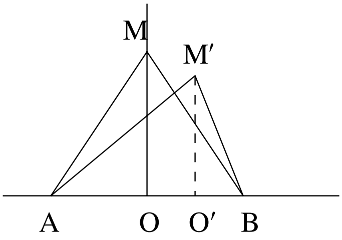

Figure 32

1°. If M is on the perpendicular bisector of AB (Fig. 32), the segments MA, MB are equal, because they are oblique lines which are equally distant from the foot of the perpendicular MO.

2°. Let M′ be a point not on the perpendicular bisector, and suppose it is on the same side of the bisector as point B. The foot O′ of the perpendicular dropped from M′ onto line AB will be on the same side of the bisector (otherwise this perpendicular would meet the bisector and, from their intersection, there would be two lines perpendicular to AB).

We will then have O′A > O′B and therefore (29)

$$M'A > M'B.$$

QED

Remark I. We could have proved the second part in a different way by establishing the following equivalent proposition: Any point which is equidistant from A and B is on the perpendicular bisector of AB. This follows from the converse of 30 (the feet of two equal oblique segments are equidistant from the foot of the perpendicular). One could also use the properties of isosceles triangles (23, Remark). However, one must observe that our way of proceeding has the advantage of showing which of the two distances is the greater when they are unequal.

Remark II. The proposition which we just stated: Any point which is equidistant from A and B is on the perpendicular bisector of AB, is the converse of the first part of the preceding theorem. We have here two different ways of proving a converse. The first consists in following the original reasoning in reverse. This is what we did in the preceding remark. The original reasoning (1° in the previous theorem) started from the hypothesis that point M is on the perpendicular bisector; in other words, that the feet of MA, MB were equally distant from the perpendicular from M, and deduced from this that MA and MB are equal. This time, we started from the hypothesis that MA and MB are equal, and we concluded that their feet are equally distant.

The second method of proving the converse consists in proving what we call the inverse statement. This name is given to the statement whose hypothesis is the negation of the original hypothesis, and whose conclusion is the negation of the original conclusion. Thus the second part of the preceding theorem is the inverse of the first, and is equivalent to its converse.

We will find later on (see, for example, 41) a third method of proving a converse.

33. Let us now make use of the definition of 1b.

Using this definition, the preceding theorem can be restated as follows:

Theorem. The locus of points equidistant from two given points is the perpendicular bisector of the segment joining these two points.

This is true because the figure formed by the points equidistant from A and B is in fact the perpendicular bisector of AB.

We remark that, to establish this fact, one must prove, as we have, that:

1°. every point on the perpendicular bisector satisfies the given condition;

2°. every point satisfying this condition is on the perpendicular bisector; or, equivalently, that no point outside the bisector satisfies the condition. This kind of double argument is necessary in all problems about geometric loci.

Exercises

Exercise 16. If the legs of a first right triangle are respectively smaller than the legs of a second, then the hypotenuse of the first is smaller than the hypotenuse of the second.

Exercise 17. If the angles B and C of a triangle ABC are acute, and the sides AB, AC unequal, then the lines starting from A are, in decreasing order of length, as follows: larger side, median (see Exercise 7 in Chapter II), angle bisector, smaller side, altitude.

Exercise 18. The median of a non-isosceles triangle is greater than the bisector from the same vertex, bounded by the third side.

CHAPTER IV

Cases of Congruence for Right Triangles. A Property of the Bisector of an Angle

34. Cases of congruence for right triangles. The general cases of congruence for triangles apply, of course, to right triangles as well. For instance, two right triangles are congruent if their legs are respectively equal (2^{nd} case of congruence for any triangles).

Aside from these general cases, right triangles present two special cases of congruence.

First case of congruence. Two right triangles are congruent if they have equal hypotenuses, and an equal acute angle.

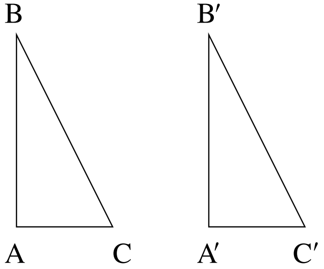

Figure 33

Consider two right triangles $ABC, A'B'C'$ (Fig. 33) in which $BC = B'C'$, $B = B'$. We move the second triangle over the first so that the angles B and B' coincide. Then $B'C'$ will assume the direction of BC, and, since these two segments are equal, C' will fall on C. $B'A'$ will assume the direction of BA, and therefore $C'A'$ will coincide with the perpendicular dropped from C on BA; that is, with CA.

Second case of congruence. Two right triangles are congruent if they have equal hypotenuses and one pair of corresponding legs equal.

Suppose the two triangles are ABC, $A'B'C'$, and that $BC = B'C'$, $AB = A'B'$. We move the second triangle onto the first so that $A'B'$ coincides with AB. The side AC will assume the direction of $A'C'$. We then have two oblique lines from point B to line AC; namely, BC and the new position of $B'C'$. These oblique lines are equal by hypothesis and therefore (30), equidistant from the foot of the perpendicular. This gives $A'C' = AC$, from which the congruence of the triangles follows.

35. Theorem. If two right triangles have equal hypotenuses and an unequal acute angle, then the sides opposite the unequal angles are unequal, and the larger side is opposite the larger angle.

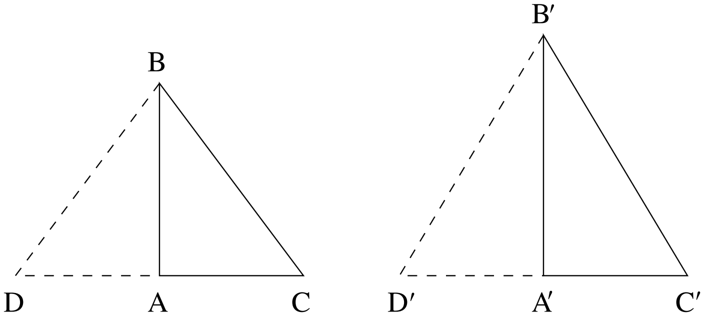

Figure 34

Let the two triangles be ABC, $A'B'C'$ (Fig. 34), for which $BC = B'C'$, $B > B'$. We claim that $AC > A'C'$.

To see this, we extend AC by its own length to get AD, and similarly extend $A'C'$ to get $A'D'$. We immediately have (29) $BD = BC = B'C' = B'D'$. Moreover, in isosceles triangle BDC median BA is also an angle bisector, so that angle DBC is twice the original angle B. Likewise, angle $D'B'C'$ is twice the original angle B', from which we have $DBC > D'B'C'$.

The two triangles DBC, $D'B'C'$ therefore have an unequal angle between equal sides, from which it follows that $DC > D'C'$, and therefore $AC > A'C'$.

36. Theorem. The bisector of an angle is the locus of points in the interior of the angle which are equidistant from the two sides.

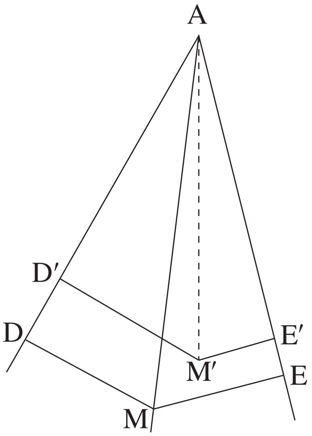

Figure 35

As explained earlier (33), the proof consists of two parts:

1°. Every point on the bisector is equidistant from the two sides.

Suppose the angle is ABC, and M is a point on its bisector. If we drop perpendiculars MD, ME to the sides of the angle, then right triangles AMD and AME are congruent, because they have a common hypotenuse, and an equal acute angle (at A) by hypothesis. Therefore the perpendiculars MD, ME are equal.

2. The distances from any point inside the angle, but not on its bisector, to the two sides are unequal.

Suppose point M lies, for instance, between the bisector and side AC. Then angle BAM will be greater than M AC. If we drop perpendiculars M D, M E onto sides AB, AC, then right triangles AM D, AM E will have a common hypotenuse and an unequal angle at A. Therefore M D will (35) be greater than M E.

As in the case of the theorem of 32, we could have proved, in place of the inverse proposition in 2 above, the converse proposition: Any point inside an angle, and equidistant from its two sides, is on its bisector. To do this, following the original reasoning in reverse, we would have considered a point M equidistant from the two sides, and would have applied the second case of congruence (34) to the two right triangles AMD, AME, which have a common hypotenuse and in which MD = ME. We would have concluded that the angles at A are equal, so that AM is the angle bisector. However, we would not have established which is the larger distance when they are unequal.

Corollaries. I. This theorem allows us to give a second definition of the bisector of an angle, namely: The bisector of an angle is the locus of the points inside the angle which are equidistant from the sides.

We observe that this second definition is exactly equivalent to the one given in 11.

II. The locus of all points equidistant from two intersecting lines is composed of the two bisectors (16) of the angles formed by these two lines.

Exercises

Exercise 19. Show that a triangle is isosceles if it has two equal altitudes.

Exercise 20. More generally, in any triangle, the smaller altitude corresponds to the larger side.

CHAPTER V

Parallel Lines

37. When two lines (Fig. 36) are intersected by a third (called a transversal), this last line forms eight angles with the first two, which are numbered in the figure. The relative positions of these angles are described as follows.

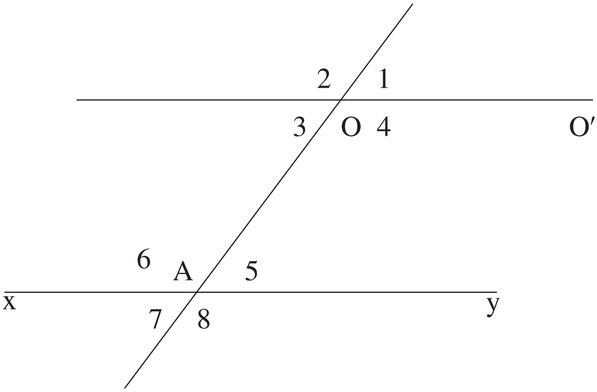

Figure 36

Two angles such as 3 and 5, situated between the two lines, and on different sides of the transversal, are called alternate interior angles.

Two angles such as 3 and 6 situated between the two lines, but on the same side of the transversal, are said to be interior on the same side.

Two angles such as 6 and 2 on the same side of the transversal, one between the two lines, one outside, are said to be corresponding.

38. Definition. Two lines in the same plane are said to be parallel if they do not intersect, no matter how far extended in either direction.

Theorem. Two lines intersected by the same transversal are parallel:

1. If the interior angles on the same side are supplementary;[1]
2. If alternate interior angles are equal;
3. If corresponding angles are equal.

1. If the two lines were to intersect, on either side of the transversal, they would form a triangle in which (25) the sum of two interior angles on the same side would have to be less than two right angles.

---

[1] If the angles 3 and 6 are supplementary, then so are 4 and 5, since the sum of these four angles is four right angles.

The other two cases can be reduced to the first:

2 . If 3 = 5, this is equivalent to saying that 3 is the supplement of 6, or that the interior angles on the same side are supplementary.

3 . If 6 = 2, then again 3 and 6 are supplementary, because 3 is the supplement of 2.

This theorem can be used to prove that two lines are parallel.

Corollary. In particular, two lines perpendicular to a third line are parallel.

39. Theorem. Through any point not on a given line, a parallel to the given line can be drawn.

Consider point O and line xy (Fig. 36). If we join point O to any point A of xy, then line OO , which makes an angle with OA such that AOO + OAy is equal to two right angles, will be parallel to xy.

40. Because the preceding construction can be made in infinitely many ways (since point A can be chosen anywhere on line xy), it would seem that there are infinitely many different parallels.

This is not true, however, if we adopt the following axiom:

Axiom. Through any point not on a given line, only one parallel to the given line can be drawn.[2]

Corollaries. I. Two distinct lines parallel to a third line are themselves parallel; since if the two lines had a point in common, two lines would pass through it, each parallel to the third line.

II. If two lines are parallel, any third line which intersects one of them must intersect the other, otherwise two parallels to the second line would intersect each other.[3]

41. The most important proposition in the theory of parallels is the following converse of the theorem in 38.

Converse. When two parallel lines are intersected by the same transversal:

1 . The interior angles on the same side are supplementary;

2 . The alternate interior angles are equal;

3 . The corresponding angles are equal.

The proof is the same in all three cases. Consider the parallel lines AB, CD intersected by the transversal EFx (Fig. 37). We claim, for example, that the corresponding angles xEB and xFD are equal. Indeed, we can construct, at point E, an angle xEB equal to angle xFD. Line EB will then be parallel to CD, and therefore it will coincide with EB.

Corollaries. I. If two lines determine, with a common transversal, two interior angles on the same side which are not supplementary, then the two lines are not parallel, and they intersect on the side of the transversal where the sum of the interior angles is less than two right angles.

II. When two lines are parallel, any line perpendicular to one of them is perpendicular to the other.

---

[2] This axiom is known as Euclid's Postulate. In fact it should be viewed as part of the definition in fundamental notions. (See Note B at the end of this volume.)

[3] We have here another example of a proof by contradiction (see 24, Remark I.)

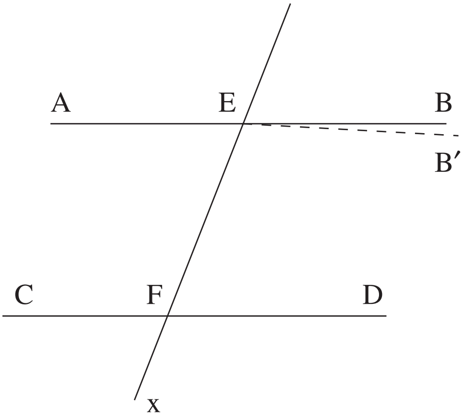

Figure 37

This is true because the perpendicular must intersect the other line (40, Corollary II) and the angle it forms will be a right angle, because of the theorem we have just proved.

Remarks. I. The corresponding equal angles xEB and xFD have the same direction of rotation.

The two parallel directions EB, FD, both situated on the same side of a common transversal, are called parallel and in the same direction.

II. We have used a new method of proof of a converse, different from the ones used in 32, and which consists in establishing the converse with the help of the original theorem. It should be noted here that the argument basically depends on the fact that the parallel through E to CD is unique.

42. According to the theorem of 38 and its converse, the definition of parallels amounts to the following: Two lines are parallel if they form, with an arbitrary transversal, equal corresponding angles (or equal alternate interior angles, or supplementary interior angles on the same side of the transversal).

This definition, equivalent to the first, is usually more convenient to use.

In place of the phrase parallel lines we often substitutes lines with the same direction, whose meaning is clear from the preceding propositions.

Remark. Because of what we have just noted, two lines which coincide must be viewed as a particular case of two parallel lines.

43. Theorem. Two angles with two pairs of parallel sides are either equal or supplementary. They are equal if the sides both lie in the same direction or both in opposite directions; they are supplementary if a pair of sides lie in the same direction, and the other pair opposite.

First, two angles with a common side, and whose second sides are parallel in the same direction (Fig. 38) are equal because they are corresponding angles. Two angles whose sides are parallel and in the same direction are then equal because one side of the first angle and one side of the second angle will form a third angle equal to the first two. If one of the sides is in the same direction, with the other

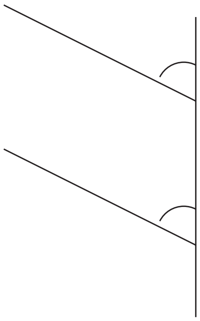

Figure 38

opposite, then extending the side which is in the opposite direction we obtain an angle supplementary to the first angle, and equal to the second.

If both sides are in opposite directions, we extend both sides of the first angle. Thus we form an angle equal to the first because they are vertical angles, and also equal to the second.

Remark. Two corresponding angles and, therefore, two angles with sides parallel and in the same direction, have the same sense of rotation. We can therefore say: Two angles with parallel sides are equal or supplementary, according as whether they have the same sense of rotation or not.

Theorem. Two angles whose sides are perpendicular in pairs are equal or supplementary, according as whether they have the same sense of rotation or not.

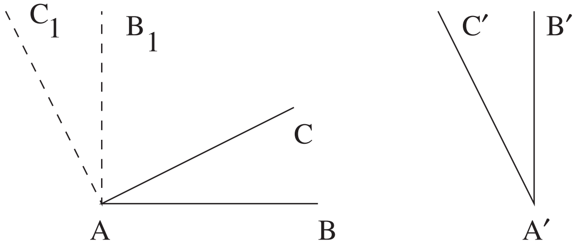

Figure 39

Consider angles BAC, $B'A'C'$ (Fig. 39) such that $A'B'$ and $A'C'$ are perpendicular respectively to AB and AC. We draw AB_1 perpendicular to AB, and turn angle B_1AC around on itself: side AB_1 falls on AC, and hence line AB, perpendicular to AB_1, will occupy a position AC_1 perpendicular to AC. We now have an angle B_1AC_1 which is equal to BAC and has the same sense (because it was constructed by reflection of angle CAB, whose sense is opposite to that of BAC) and whose sides, perpendicular respectively to those of the first angle, are therefore parallel to those of $B'A'C'$. Since the angles BAC, $B'A'C'$ are equal or supplementary, according as their sides have the same or opposite senses of rotation, the same is true of the given angles.

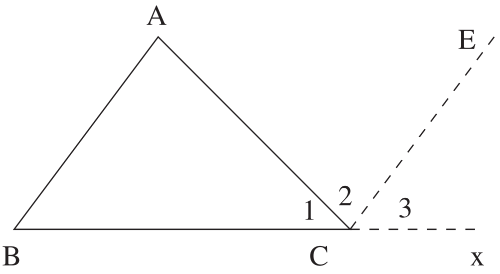

Figure 40

44. Theorem. The sum of the angles of a triangle is equal to two right angles.

In triangle ABC (Fig. 40), we extend BC in the direction Cx, and draw CE parallel to AB. We form three angles at C (numbered 1 to 3 in the figure), whose sum is equal to two right angles. These angles are equal to the three angles of the triangle; namely: 1 is the angle C of the triangle; 2 = A (these are alternate interior angle formed by transversal AC with the parallel lines AB, CE); 3 = B (these are corresponding angles formed by transversal BC with the same parallels).

Corollaries. I. An exterior angle of a triangle is equal to the sum of the non-adjacent interior angles.

II. The acute angles of a right triangles are complementary.

III. If two triangles have two pairs of equal angles, then the third pair of angles is equal as well.

44b. Theorem. The sum of the interior angles of a convex polygon[4] is equal to two less than the number of sides times two right angles.

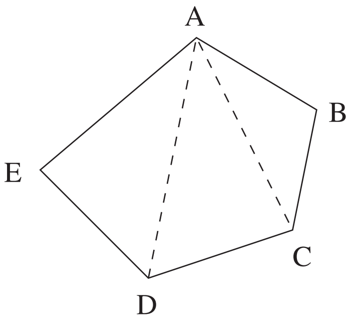

Figure 41

Let the polygon be ABCDE (Fig. 41). Joining A to the other vertices, we decompose the polygon into triangles. The number of triangles is equal to the number of sides minus two because if A is taken as the common vertex of these triangles, all the other sides of the polygon are the sides opposite A, except the two sides which end at A. The sum of the angles of these triangles gives us the sum of the angles of the polygon. The theorem is proved.

---

[4] The theorem remains true if one takes, as the interior angle at any vertex pointing inwards (Fig. 19b), the one which is directed towards the interior of the polygon, an angle which in this case is greater than two right angles. In the case of Fig. 19b the proof would proceed as in the text, drawing the diagonals starting from the vertex pointing inwards. If there are several inward-pointing vertices, the theorem would still be true, but the proof would be more difficult.

If n is the number of sides of the polygon, the sum of the angles is $2(n - 2)$ or $2n - 4$ right angles.

Corollary. The sum of the exterior angles of a convex polygon, formed by extending the sides in the same sense, is equal to four right angles.

Indeed, an interior angle plus the adjacent exterior angle gives us two right angles. Adding the results for all the n vertices we obtain 2n right angles, of which $2n - 4$ are given by the sum of the interior angle. The sum of the exterior angles is equal to the four missing right angles.

Exercises

Exercise 21. In a triangle ABC, we draw a parallel to BC through the intersection point of the bisectors of B and C. This parallel intersects AB in M and BC in N. Show that $MN = BM + CN$.

What happens to this statement if the parallel is drawn through the intersection point of the bisectors of the exterior angles at B and C? Or through the intersection of the bisector of B with the bisector of the exterior angle at C?

Sum of the angles of a polygon.

Exercise 22. Prove the theorem in 44b by decomposing the polygon into triangles using segments starting from an interior point of the polygon.

Exercise 23. In triangle ABC we draw lines AD, AE from point A to side BC, such that the first makes an angle equal to C with AB, while the second makes an angle equal to B with AC. Show that triangle ADE is isosceles.

Exercise 24. In any triangle ABC:

1. The bisector of A and the altitude from A make an angle equal to half the difference between B and C.

2. The bisectors of B and C form an angle equal to $\frac{1}{2}A$ + one right angle.

3. The bisectors of the exterior angles of B and C form an angle equal to one right angle $- \frac{1}{2}A$.

25. In a convex quadrilateral:

1. The bisectors of two consecutive angles form an angle equal to one half the sum of the other two angles.

2. The bisectors of two opposite angles form a supplementary angle to the half the difference of the other two angles.

CHAPTER VI

On Parallelograms. On Translations

45. Among quadrilaterals, we consider in particular trapezoids and parallelograms. A quadrilateral is called a trapezoid (Fig. 42) if it has two parallel sides. These parallel sides are called the bases of the trapezoid. A quadrilateral with two pairs of parallel sides is called a parallelogram. (Fig. 43).

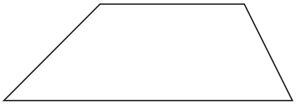
Figure 42

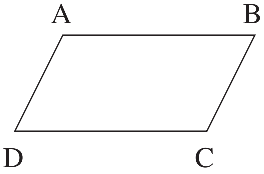

Figure 43

Theorem. In a parallelogram, opposite angles are equal, and angles adjacent to the same side are supplementary.

Indeed, in parallelogram ABCD (Fig. 43), angles A and C, adjacent to side AB, are interior angles on the same side formed by parallels AD, BC cut by transversal AB; therefore they are supplementary. The opposite angles A, C are equal because they have parallel sides in opposite directions.

Remark. We see that it suffices to know one angle of a parallelogram in order to know all of them.

Converse. If, in a quadrilateral, the opposite angles are equal, then the quadrilateral is a parallelogram.

Indeed, the sum of the four angles of a quadrilateral equals four right angles (44b). Since A = C, B = D, the sum of the four angles A + B + C + D can be writen as 2A + 2B. We then have A + B = two right angles, so the lines AD, BC are parallel, since they form two supplementary interior angles on the same side of transversal AB. Similarly, we can show that AB is parallel to CD.

46. Theorem. In any parallelogram, the opposite sides are equal.

In parallelogram ABCD (Fig. 44), we draw diagonal AC. This diagonal divides the parallelogram into two triangles ABC, CDA, which are congruent because they have the common side AC between equal angles: $A_1 = C_1$ are alternate interior angles formed by parallels AB, CD, and $A_2 = C_2$ are alternate interior angles formed by parallels AD, BC.

These congruent triangles yield AB = CD, AD = BC.

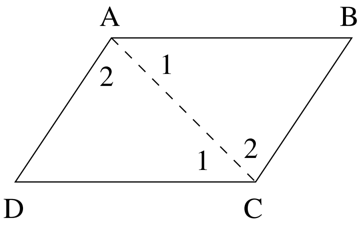

Figure 44

In this theorem, the hypothesis consists of two parts:
1 . two sides are parallel;
2 . the other two sides are also parallel.

The conclusion also has two parts:
1 . two opposite sides are equal;
2 . the other two sides are also equal.

Since we can form a converse by using either the whole or a part of the original conclusion, and vice versa, this theorem has two converses.

Converses. A quadrilateral is a parallelogram:
1 . If the opposite sides are equal;
2 . If two opposite sides are parallel and equal.

1 . Suppose that AB = CD and AD = BC in quadrilateral ABCD (Fig. 44). We again draw diagonal AC. Triangles ABC, CDA will be congruent, since they have three sides equal in pairs. Thus angles A_1 and C_1 are equal, and since these are alternate interior angles with respect to transversal AC, lines AB, CD must be parallel. Likewise, the equality of A_2, C_2 proves that AD and BC are parallel.

2 . Assume now that AB = CD and AB is parallel to CD. Triangles ABC, CDA are again congruent, because they have an equal angle ($A_1 = C_1$ are alternate interior angles) between two pairs of equal sides: AC is a common side and AB = CD. From this congruence we again find that $A_2 = C_2$, and that sides AD, BC are parallel.

46b. Remark I. A quadrilateral can have two sides AB, CD equal, and the other two BC, AD parallel, without being a parallelogram (it is then called an isosceles trapezoid).

Choosing an arbitrary side AB, it suffices to take the line symmetric to AB (D_1C, Fig. 45) with respect to any line xy in the plane, provided that the line xy is not parallel to AB, and that xy intersects the line AB in a point I on the extension of AB (and not on segment AB itself). Then quadrilateral $ABCD_1$ will have two parallel sides (both perpendicular to xy) and the other two equal (because they are symmetric to each other). These last two sides are not parallel because they intersect at point I.

Conversely, every quadrilateral with two sides BC, AD parallel, and the other two equal is either a parallelogram, or (in the case of an isosceles trapezoid) a figure with a line of symmetry. Indeed, let xy be the perpendicular bisector of BC. We have already found two oblique lines from C to the line AD, both equal to AB, namely: CD_1 which is symmetric to AB with respect to xy, and CD_2 which,

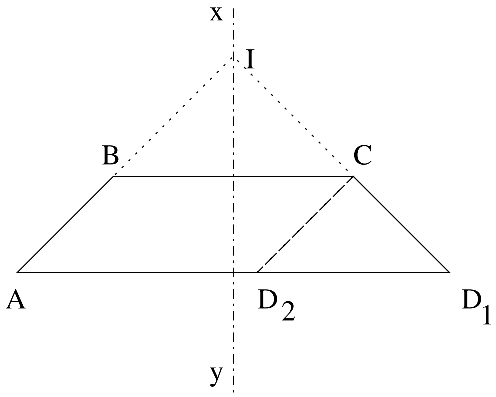

Figure 45

together with A, B, and C forms a parallelogram. These oblique lines certainly have endpoints on AD, since that line is parallel to BC. It follows that D must be either D_1 or D_2, since there are only two oblique lines equal to AB from C to AD.

The reasoning above may seem to be incorrect if D_2 coincides with D_1; that is, when CD_1 is parallel to AB. This requires, as we have just seen, that AB also be parallel to xy, and therefore perpendicular to AD, and indeed in this case there is only one oblique line from C with length equal to AB (the perpendicular).

Remark II. In the first part of the preceding converse it is essential that the quadrilateral be proper (21, Remark): it is only if the triangles ABC, ADC of Fig. 44 are on different sides of the common side AC that the angles A_1, C_1 are alternate interior.

It is easy to construct an improper quadrilateral (called a anti-parallelogram) whose opposite sides are equal. It suffices, in parallelogram ABCD (Fig. 45b), to replace the point D by point E symmetric to it with respect to the diagonal AC.

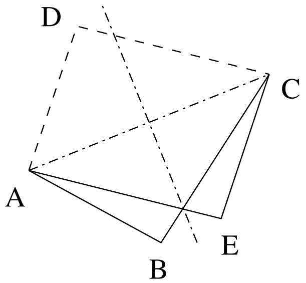

Figure 45b

We can also obtain an anti-parallelogram $ABCE_1$ by taking E_1 to be symmetric to B with respect to the perpendicular bisector of AC, so that ABE_1C is an isosceles trapezoid. But this point E_1 is none other than E, because (24, case 3) there is, on any one side of AC, only one point E which is both at a distance AE = BC from A and also at a distance CE = AB from C. Every improper quadrilateral with equal opposite sides is thus formed by the non-parallel sides and the diagonals of an isosceles trapezoid.

Remark III. Two quadrilaterals with four pairs of equal sides need not be congruent. In other words, we can deform a quadrilateral ABCD (proper or not) without changing the lengths of its sides.

Indeed, let AB = a, BC = b, CD = c, DA = d be the constant lengths of the four sides. With sides a, b and an arbitrary included angle B, we can construct a triangle ABC, whose construction will determine the length of AC (a diagonal of the quadrilateral). To each value of this diagonal (under the conditions for existence given in 86, Book II) there corresponds a triangle ACD with this base AC and with the other two sides CD, DA having length c, d. The angle B can therefore be arbitrary (at least within certain limits).

A quadrilateral which can be deformed under these conditions is called an articulated quadrilateral. This notion is important in practical applications of geometry.

According to the preceding results, a parallelogram remains a parallelogram when it is articulated and, likewise, an anti-parallelogram remains an anti-parallelogram under these conditions.[1]

47. Theorem. In a parallelogram, the two diagonals divide each other into equal parts.

In parallelogram ABCD (Fig. 46), we draw diagonals AC, BD, which intersect at O. Triangles ABO, CDO are congruent because they have equal angles, and an equal side AB = CD (by the preceding theorem). Therefore AO = CO, BO = DO. QED

Converse. A quadrilateral is a parallelogram if its diagonals divide each other into equal parts.

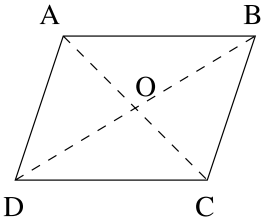

Figure 46

Indeed, assume (Fig. 46) that AO = CO, BO = DO. Triangles ABO, CDO are again congruent because they have an equal angle (the angles at O are vertical) included between pairs of equal sides. Therefore their angles at A and C will be equal, so that AB is parallel to CD. The congruence of triangles ADO, BCO also shows that AD is parallel to BC.

---

[1] This reasoning fails only if a parallelogram is transformed into an anti-parallelogram, or inversely (since a parallelogram and an anti-parallelogram are the only quadrilaterals with opposite sides equal). If the deformation occurs continuously, this requires that the quadrilateral first flatten into a line, so that one pair of adjacent sides ends up as extensions of each other, as does the other pair.

Remark. We have proved the converses in 46 and 47 by retracing the original reasoning in the opposite direction, as explained in 32, Remark II.

48. A quadrilateral whose angles are all equal, and are therefore are all right angles, is called a rectangle. A rectangle is a parallelogram, since its opposite angles are equal.

A quadrilateral whose sides are all equal is called a rhombus. A rhombus is a parallelogram since it has equal opposite sides.

Thus, in a rectangle, as in a rhombus, the diagonals intersect at their common midpoint.

Theorem. The diagonals of a rectangle are equal.

In rectangle ABCD (Fig. 47), the diagonals are equal because triangles ACD, BCD are congruent: they have side DC in common, ADC = DCB since they are both right angles, and AD = BC since they are opposite sides of a parallelogram.

Corollary. In a right triangle, the median from the vertex of the right angle equals half the hypotenuse.

This is true because if we draw parallels to the sides of the right angle through the endpoints of the hypotenuse, we form a rectangle, in which the median in question is half the diagonal.

Converse. A parallelogram with equal diagonals is a rectangle.

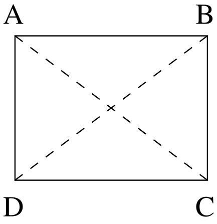

Figure 47

Suppose (Fig. 47) that AD = BC in parallelogram ABCD. We know that AD = BC: consequently, triangles ADC, BCD are congruent, since their three sides are equal in pairs. Angles ADC, BCD are therefore equal and, since they are supplementary, they must be right angles, which shows that the parallelogram is a rectangle.

Corollary. A triangle in which a median is one half the corresponding side is a right triangle.

Corollary. In a rhombus, the diagonals are perpendicular, and bisect the vertex angles.

If ABCD (Fig. 48) is a rhombus, then triangle ABD is isosceles. Diagonal AC, being a median of this triangle, is also an altitude and an angle bisector.

Converse. A parallelogram with perpendicular diagonals is a rhombus.

Indeed, each vertex is equidistant from the adjacent vertices, since it lies on the perpendicular bisector of the diagonal joining these vertices.

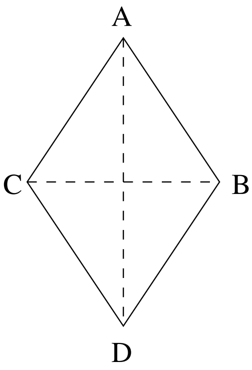

Figure 48

49. A square is a quadrilateral in which all sides are equal, and all the angles right angles.

Thus a square is both a rhombus and a rectangle, so that the diagonals are equal, perpendicular, and intersect each other at their common midpoint.

Conversely, any quadrilateral whose diagonals are equal, perpendicular, and bisect each other is a square.

Two squares with the same side are congruent.

50. Translations.

Lemma. Two figures F, F' are congruent, with the same sense of rotation, if their points correspond in such a way that, if we take points A, B, C from one figure, and the corresponding points A', B', C' from the other, then the triangles thus formed are always congruent, with the same sense of rotation, no matter which point we take for C.

Indeed, take two points A, B in figure F, and the homologous[2] points A', B'. Segment AB is then clearly equal to $A'B'$. Let us move the second figure onto the first in such a way that these two equal segments coincide. We claim that then the two figures coincide completely. Indeed, let C be a third point of the first figure, and let C' be its homologous point. Since the two triangles ABC, $A'B'C'$ are congruent, angle $B'A'C'$ is equal to BAC, and has the same sense of rotation. Therefore, when $A'B'$ is made to coincide with AB, the line $A'C'$ will assume the direction of AC. Since $A'C' = AC$ as well, we conclude that C' coincides with C. This argument applies to all the points of the figure, so the two figures must coincide completely.

Remarks. I. We have just provided a sufficient condition for two figures to be congruent; this conditions is also clearly necessary.

From the preceding reasoning it also follows that:

II. In order to superimpose two equal figures with the same orientation, it suffices to superimpose two points of one of the figures onto their homologous points.

---

[2] This is the name given to pairs of corresponding points in the two figures.

51. Theorem. If, starting at each point of a figure, we draw equal parallel segments in the same direction, then their endpoints will form a figure congruent to the first.

First consider two points A, B of the first figure, and the corresponding points A', B' of the second. Since the segments AA' and BB' are parallel and equal, $ABA'B'$ is a parallelogram. Therefore $A'B'$ is equal and parallel to AB, with the same orientation. Thus segments joining homologous pairs of points are equal, parallel, and have the same direction.

It follows that any three points of the first figure correspond to three points forming a congruent triangle, and since the angles of these triangles have their sides parallel and in the same sense, their sense of rotation is the same. The figures are therefore congruent.

The operation by which we pass from the first figure to the second is called a translation. We note that a translation is determined if we are given the length, direction, and sense of a segment, such as AA', joining a point to its homologue. We can therefore designate a translation by the letters of such a segment: for example, we speak of the translation AA'.

Corollaries. I. If, through each point on a line, we draw equal parallel segments in the same sense, the locus of their endpoints is a line parallel to the first.

In particular, the locus of points on the same side of a line, and a at a given distance from the line, is a parallel line.

II. Two parallel lines are everywhere equidistant.

We can therefore speak of the distance between two parallel lines.

III. The locus of points equidistant from two parallel lines is a third line, parallel to the first two.

Exercises

Parallelograms.

Exercise 26. The angle bisectors of a parallelogram form a rectangle. The bisectors of the exterior angles also form a rectangle.

Exercise 27. Any line passing through the intersection of the diagonals of a parallelogram is divided by this point, and by two opposite sides, into two equal segments.

For this reason, the point of intersection of the diagonals of a parallelogram is called the center of this polygon.

Exercise 28. Two parallelograms, one of which is inscribed in the other (that is, the vertices of the second are on the sides of the first) must have the same center.

Exercise 29. An angle of a triangle is acute, right, or obtuse, according as whether its opposite side is less than, equal to, or greater than double the corresponding median.

Exercise 30. If, in a right triangle, one of the acute angles is double the other, then one of the sides of the right angle is half the hypotenuse.

Translations.

Exercise 31. Find the locus of the points such that the sum or difference of its distances to two given lines is equal to a given length.

Exercise 32. Given two parallel lines, and two points A, B outside these two parallels, and on different sides, what is the shortest broken line joining the two points, so that the portion contained between the two parallels has a given direction?

CHAPTER VII

Congruent Lines in a Triangle

52. Theorem. In any triangle, the perpendicular bisectors of the three sides are concurrent.

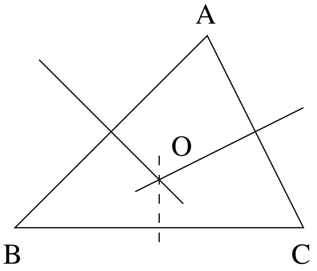

Figure 49

Suppose the triangle is ABC (Fig. 49). The perpendicular bisectors of sides AB, AC are not parallel (otherwise lines AB, AC would coincide), so they intersect at some point O. We must show that point O is also on the perpendicular bisector of BC.

Because point O, is on the perpendicular bisector of AB, it is equidistant from A and B; likewise, because it is on the perpendicular bisector of AC, it is equidistant from A and C. It is therefore equidistant from B and C and thus it is on the perpendicular bisector of BC.

53. Theorem. In any triangle, the three altitudes are concurrent.

Suppose the triangle is ABC (Fig. 50). We draw a parallel to BC through A, a parallel to AC through B, and a parallel to AB through C. This forms a new triangle A B C . We will show that the altitudes of ABC are the perpendicular bisectors of the sides of the new triangle, from which it follows that they are concurrent.

Parallelogram ABCB gives us BC = AB , and parallelogram ACBC gives BC = AC , so that A is indeed the midpoint of B C . Altitude AD of ABC thus passes through the midpoint of B C , and is perpendicular to it because it is perpendicular to the parallel line BC.

Since this reasoning can be repeated for the other altitudes, the proof is complete.

54. Theorem. In any triangle:

1 . the three angle bisectors are concurrent;

2 . the bisector of an angle, and the bisectors of the two non-adjacent exterior angles, are concurrent.

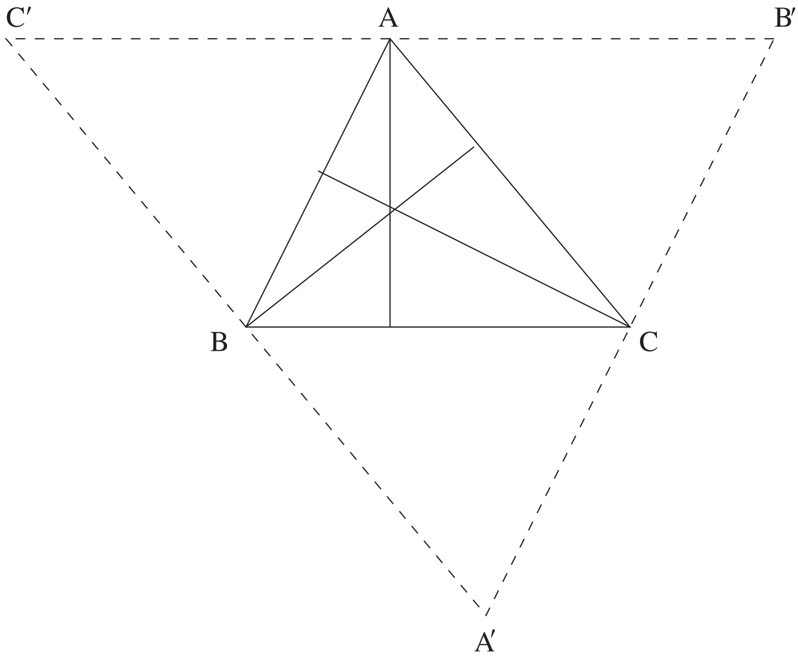

Figure 50

1 . In triangle ABC (Fig.51), we draw the bisectors of angles B and C; they intersect at a point O inside the triangle. This point, being on the bisector of angle B, is equidistant from the sides AB and BC. Likewise, being on the bisector of angle C, the point O is equidistant from AC and BC. It is therefore equidistant from AB and AC and, being interior to angle A, is on the bisector of this angle.

2 . Since the exterior angles CBx, BCy have a sum less than four right angles, halves of each have a sum less than two right angles. The bisectors of these angles will therefore meet (41, Corollary) at a point O inside angle A. This point O , like point O, will be equidistant from the three sides of the triangle. Therefore it will belong to the bisector of A.

55. Theorem. The segment joining the midpoints of two sides of a triangle is parallel to the third side, and equal to half of it.

In triangle ABC (Fig. 52), let D be the midpoint of AB and let E be the midpoint of AC. We extend line DE past E by its own length, to a point F. Quadrilateral ADCF will be a parallelogram (47), and therefore CF will be equal and parallel to DA, or, equivalently, to BD. Thus quadrilateral BDCF is also a parallelogram. Therefore:

1 . DE is parallel to BC;
2 . DE, which is equal to half of DF, is also half of BC.

56. Theorem. The three medians of a triangle are concurrent at a point situated on each of them one third of its length from the corresponding side.

First, let BE, CF be two medians of triangle ABC (Fig. 53). We claim that their intersection G lies at one third the length of each of them. To see this, let M

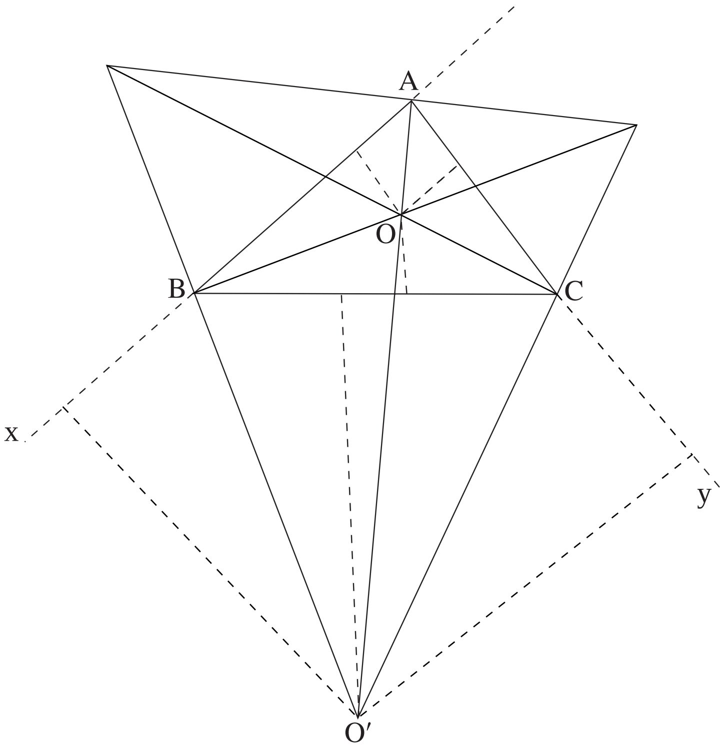

Figure 51

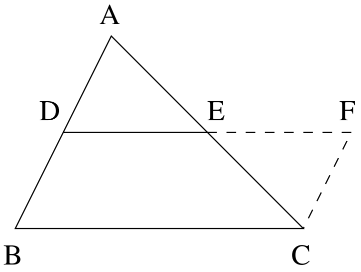

Figure 52

and N be the midpoints of BG and CG. Segment MN, which joins the midpoints of two sides of triangle BCG, is parallel to BC and equal to half of it. But EF is parallel to BC and equal to half of it. This means that EFMN is a parallelogram, whose diagonals divide each other in half. Therefore EG = GM = MB and FG = GN = NC.

Thus median BE passes through the point situated at one third the length of CF. But the same reasoning can also be applied to show that the median AD passes through the same point. The thereom is proved.

Remark. The point where the medians meet is also called the center of mass of the triangle. The reason for this name is given in the study of mechanics.

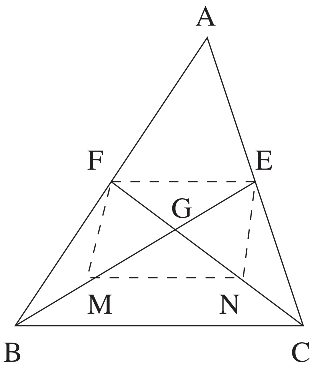

Figure 53

Exercises

Exercise 33. Join a given point to the intersection of two lines, which intersect outside the limits of the diagram (53).

Exercise 34. In a trapezoid, the midpoints of the non-parallel sides and the midpoints of the two diagonals are on the same line, parallel to the bases. The distance between the midpoints of the non-parallel sides equals half the sum of the bases; the distance between the midpoints of the diagonals is equal to half their difference.

Exercise 35. If, from two points A, B and the midpoint C of AB, we drop perpendiculars onto an arbitrary line, the perpendicular from C is equal to half the sum of the other two perpendiculars, or to half their difference, according as whether these two perpendiculars have the same or opposite sense.

Exercise 36. The midpoints of the sides of any quadrilateral are the vertices of a parallelogram. The sides of this parallelogram are parallel to the diagonals of the given quadrilateral, and equal to halves of these diagonals. The center of the parallelogram is also the midpoint of the segment joining the midpoints of the diagonals of the given quadrilateral.

Exercise 37. Prove that the medians of a triangle ABC are concurrent by extending the median CF (Fig. 53) beyond F by a length equal to FG.

Exercise 38. Given three lines passing through the same point O (all three distinct), and a point A on one of them, show that there exists:

1 . A triangle with a vertex at A and having the three lines as its altitudes (one exception);

2 . A triangle with a vertex at A and having the three lines as its medians;

3 . A triangle with a vertex at A and having the three lines as bisectors of its interior or exterior angles (one exception);

4 . A triangle with a midpoint of one of its sides at point A and having the three lines as perpendicular bisectors of the sides (reduce this to 1).

Problems for Book 1

Exercise 39. In a triangle, the larger side corresponds to the smaller median[1]. (Consider the angle made by the third median with the third side.) A triangle with two equal median is isosceles.

Exercise 40. Let us assume that a billiard ball which strikes a flat wall will bounce off in such a way that the two lines of the path followed by the ball (before and after the collision) make equal angles with the wall. Consider n lines $D_1, D_2, \ldots, D_n$ in the plane, and points A, B on the same side of all of these lines. In what direction should a billiard ball be shot from A in order that it arrive at B after having bounced off each of the given lines successively? Show that the path followed by the ball in this case is the shortest broken line going from A to B and having successive vertices on the given lines.[2]

Special Case. The given lines are the four sides of rectangle, taken in their natural order; the point B coincides with A and is inside the rectangle. Show that, in this case, the path traveled by the ball is equal to the sum of the diagonals of the rectangle.

Exercise 41. The diagonals of the two rectangles of exercise 26 are situated on the same two lines, parallel to the sides of the given parallelogram (analogous to 54). One of these diagonals is half the difference, and the other half the sum, of the sides of the parallelogram.

Exercise 42. In an isosceles triangle, the sum of the distances from a point on the base to the other sides is constant. What happens if the point is taken on the extension of the base?

In an equilateral triangle, the sum of the distances from a point inside the triangle to the three sides is constant. What happens when the point is outside the triangle?

Exercise 43. In triangle ABC, we draw a perpendicular through the midpoint D of BC to the bisector of angle A. This line cuts off segments on the sides AB, AC equal to, respectively, $\frac{1}{2}(AB + AC)$ and $\frac{1}{2}(AB - AC)$.

Exercise 44. Let $ABCD$, $DEFG$ be two squares placed side by side, so that sides DC, DE have the same direction, and sides AD, DG are extensions of each other. On AD and on the extension of DC, we take two segments AH, CK equal to DG. Show that quadrilateral $HBKF$ is also a square.

Exercise 45. On the sides AB, AC of a triangle, and outside the triangle, we construct squares $ABDE$, $ACGF$, with D and F being the vertices opposite A. Show that:

1°. EG is perpendicular to the median from A, and equal to twice this median;

2°. The fourth vertex I of the parallelogram with vertices EAG (with E and G opposite vertices) lies on the altitude from A in the original triangle;

---

[1] For similar statements concerning the altitudes of a triangle, see Exercises 19, 20, and for angle bisectors see exercises 362, 362b at the end of this volume.

[2] In the case where there is only one line, the problem reduces to the subject of Exercises 13–14. Once these exercises are solved, one tries to find a way to use the solution for the case of one line to treat the case for two lines; then to extend it to three lines, and so on.

3. CD, BF are equal to and perpendicular to BI, CI respectively, and their intersection point is also on the altitude from A.

Exercise 46. We are given a right angle AOB, and two perpendicular lines through a point P, the first intersecting the sides of the angle in A, B, and the second intersecting the same sides in C, D. Show that the perpendiculars from the points D, O, C to the line OP intercept on AB segments equal to AP, PB, respectively, but having the opposite sense.

Book II
On the Circle

CHAPTER I

Intersection of a Line with a Circle

57. Theorem. Through three non-collinear points we can always draw one and only one circle.

In other words, a circle is determined by three points not on a straight line.

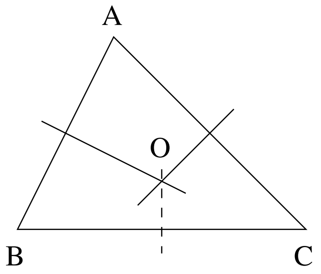

Figure 54

Indeed, assume that points A, B, C (Fig. 54) are not collinear. We proved earlier (52) that the perpendicular bisectors of segments BC, CA, AB pass through the same point O, equidistant from A, B, C. The circle with center O and radius OA passes through the three given points. It is the only circle with this property, because the center of a circle passing through A, B, C must necessarily belong to the three perpendicular bisectors just mentioned.

Corollary. We see that a circle cannot have two different centers nor, consequently, two different radii.

58. Theorem. A line cannot intersect a circle in more than two points.

If the distance from the center to the line is greater than the radius, the line does not intersect the circle.

If this distance is less than the radius, the line intersects the circle in two points.

Finally, if the distance to the line is equal to the radius, the line and the circle have only one common point.

In the last case we say that the line is tangent to the circle.

Assume that we are given a circle with center O and a line D. We drop (Fig. 55) perpendicular OH to the line.

1 . The circle cannot have more than two common points with line D. This amounts to saying that there are at most two oblique line segments equal in length to the radius R from O to D (31, Corollary).

2 . If OH is greater than the radius, the same is true a forteriori of all the oblique segments to D from O (29); thus all the points of D are outside the circle.

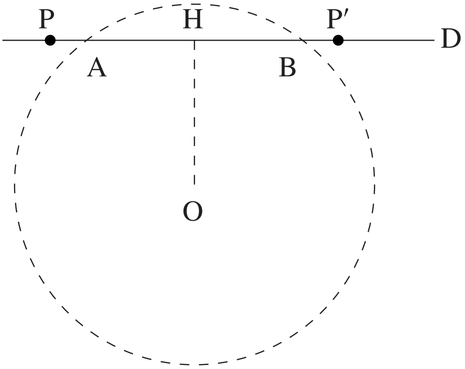

Figure 55

3 . If, on the contrary, OH (Fig. 55) is less than the radius, then point H is inside the circle; but on the two rays determined by H, there exist points outside the circle. To obtain such points, it is enough to construct points P, P on these rays such that HP, HP are equal to the radius: the distances OP, OP must then be greater than the radius. Therefore there will be an intersection point between H and P, and another between H and P ; further, these are the only points of intersection (1).

4 . Finally, if OH is equal to the radius (Fig. 56), the point H is on the circle but, as in 2 , we see that all the other points of the line are outside the circle.

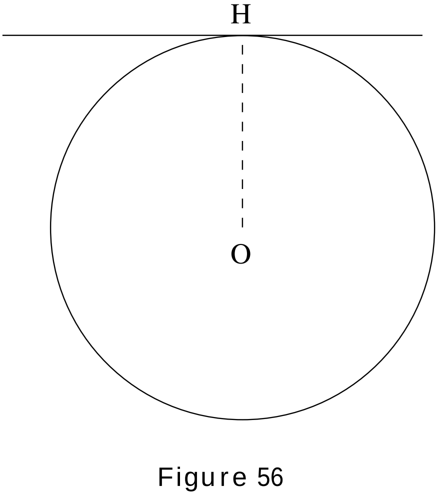

Figure 56

Corollary. Through a point on a circle we can draw one, and only one, tangent to the circle; namely, the perpendicular to the radius which ends at that point.

59. The definition just given for the tangent to a circle is not appropriate for an arbitrary curve.

Definition. A tangent to an arbitrary curve (Fig. 57) at a point M of this curve is the limiting position of line MM' as the point M' on the curve approaches the point M.

In other words,[1] line MT will be said to be tangent at M if, for any angle ε, we can find, on the two sides of M, arcs $\overset{\frown}{MM_1}$, $\overset{\frown}{MM_2}$ such that for any point M' on these arcs, line MM' makes an angle less than ε with line[2] MT.

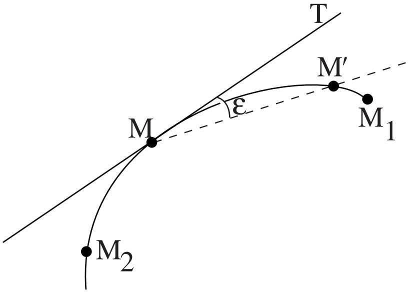

Figure 57

In the case of a circle, we will see that this definition is equivalent to the one we gave earlier.

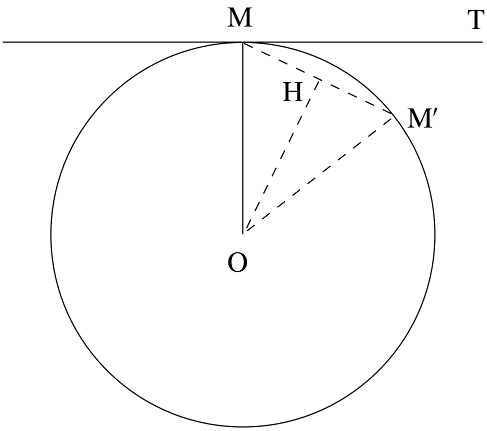

Figure 56b

From a point M on the circle we draw perpendicular MT to radius OM (Fig. 56b), and from O we drop a perpendicular OH to chord MM'. This perpendicular is both an altitude of the isosceles triangle OMM', and also the bisector of the angle at O. Angle TMM' is equal to MOH (they have pairs of perpendicular sides), and thus is half of MOM. But this angle can be made arbitrarily small by taking the point M sufficiently close to M.

---

[1] See the definition of the word limit in Leçons d'Arithmetique by J. Tannery (Paris, Librarie Armand Colin, 1894), chap. VII, n° 237, and chap. XII.

[2] One can show, as it is shown in arithmetic (Leçons de J. Tannery, n° 239), that if the line MT exists, then it is unique.

60. The line perpendicular to the tangent at its point of contact with a curve is called the normal to the curve at that point. The normal to a circle at any point is simply the radius ending at that point.

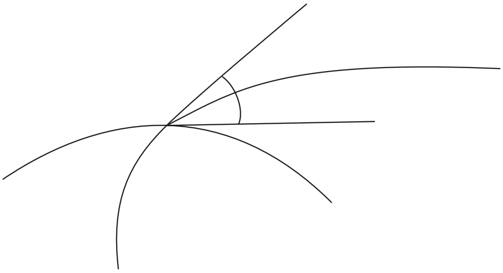

Figure 57b

On an arbitrary circle, there exist two points (and only two) such that the normals at these points pass through a given point P on the plane (different from the center): these are the endpoints of the diameter passing through P.

60b. The angle between two curves at one of their intersection points is the angle formed by the tangents to the curves at this point (Fig. 57b). The angle at which two circles intersect is therefore equal to the angle between the two radii ending at a common point, or to its supplement.

Exercises

Exercise 47. Through every point on a circle, we draw segments equal and parallel to a given segment. What is the locus of the endpoints of these segments?

Exercise 48. Find the locus of the midpoints of the segments joining a fixed point with each point on a circle.

Exercise 49. Consider a diameter AB of a circle O, a point C on the extension of the diameter, and a secant CDE passing through C which cuts the circle in D, E. If the exterior segment CD of the secant is equal to the radius, show that angle EOB is triple DOA.

CHAPTER II

Diameters and Chords

61. Using the definition of 19b, the remarks in 9 can be stated as follows:

Theorem. Every diameter is a line of symmetry for the interior of its circle, and also for its circumference.

We see that a circle has infinitely many lines of symmetry.

62. A line segments whose endpoints are the endpoints of an arc of a circle is called a chord. The arc is said to be subtended by the chord. We note that every chord subtends two different arcs, one larger than, the other smaller than, a semicircle (or both equal to a semicircle if the chord is a diameter).

63. Theorem. The diameter perpendicular to a chord cuts this chord, and its two subtended arcs, into equal parts.

Indeed, the diameter of a circle O perpendicular to a chord AB is a line of symmetry for the circle and at the same time a line of symmetry for the isosceles triangle OAB. Therefore this diameter is a line of symmetry for the whole figure formed by the circle and the chord.

Corollary. The same line is determined by any two of the following conditions:

1 . It is perpendicular to a chord;

2 . It passes through the center of the circle;

3 . It passes through the midpoint of the chord;

4 , 5 . It passes through the midpoint of one of the subtended arcs.

Corollary. The locus of the midpoints of a series of parallel chords is the diameter perpendicular to these chords.

Corollary. A tangent is parallel to any chord bisected by the diameter through the point of contact.

Theorem. Two parallel lines intercept equal arcs between them.

Indeed (Fig. 58), these arcs are symmetric with respect to the diameter which is perpendicular to these lines.

64. Theorem. On any circle, the closest and furthest point from a given point P of the plane are the feet of the normals (60) passing through P.

If A is the one of these points on ray OP, and B the point on the opposite ray (Figures 59 and 60), the distance PA is the difference between OP and the radius, and PB is the sum of the same segments. An arbitrary point M of the circle is

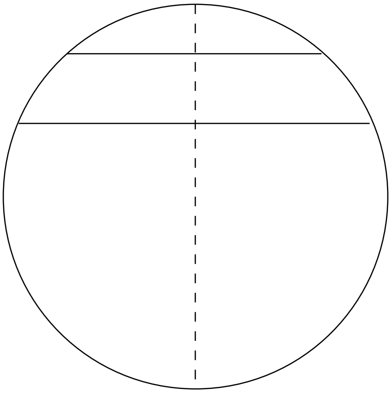

Figure 58

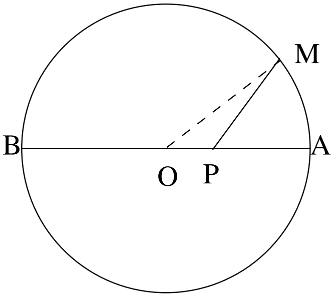

Figure 59

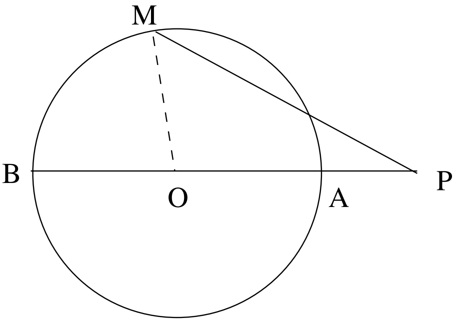

Figure 60

therefore at a distance from P greater than PA and smaller than PB, since PM is the third side of triangle OPM.

The distance PM increases as the point M moves along the circle from A to B, because, in triangle OPM, it is opposite to an angle which increases, but is contained between two constant sides.

Corollary. The diameter is the longest chord of a circle.

Indeed, if we take point P to coincide with A, we see that chord PM is smaller than the diameter PB.

65. Theorem. In the same circle, or in equal circles we have:

1 . Equal chords correspond to equal arcs, and conversely;

2 . For two unequal arcs, both less than a semicircle, the greater arc corresponds to the greater chord.

1 . If we superimpose the two equal arcs, their endpoints will coincide, and so will the corresponding chords. Conversely, if the chords are equal, the angles at the center corresponding to these chords are also equal, by the third case of congruence for triangles (SSS). Therefore the corresponding arcs are equal as well.

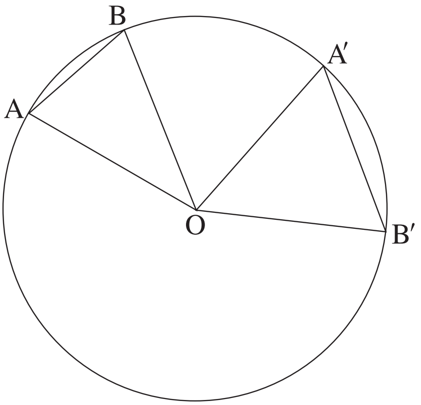

Figure 61

2°. If arc $\widehat{AB}$ is less than $\widehat{A'B'}$ (Fig. 61), and therefore the angle AOB is less than $A'OB'$, then chord AB will be less than $A'B'$ by the theorem of 28 applied to triangles OAB, $OA'B'$.

66. Theorem. In the same circle, or in equal circles:

1°. Two equal chords are equidistant from the center, and conversely;

2°. Of two unequal chords, the longer is closer to the center.

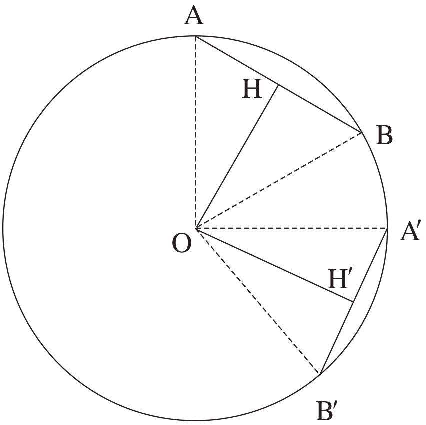

Figure 62

1°. Two equal chords of the same circle correspond to equal arcs. It suffices to superimpose the two arcs to see that the midpoints of the two chords are equidistant from the center.

Conversely, if two chords AB, $A'B'$ (Fig. 62) of circle O are equidistant from the center, then right triangles OHA, $OH'A'$ have equal hypotenuses and an equal pair of sides $OH = OH'$; thus $HA = H'A'$ and $AB = A'B'$.

2°. Suppose that chord AB is greater than $A'B'$ (Fig. 63). It follows that angle AOB is greater than $A'OB'$ and, dropping perpendiculars OH, OH', we see that

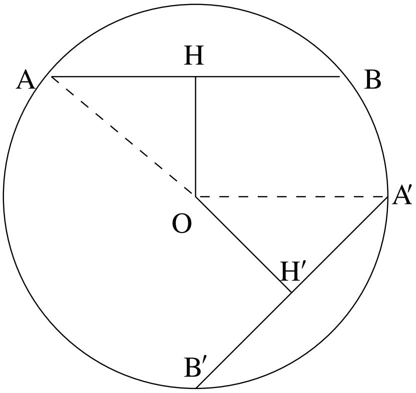

Figure 63

AOH is greater than $A'OH'$. Angle OAH, the complement of the first angle, is less than $OA'H'$, the complement of the second. Thus right triangles OHA, $OH'A'$ have equal hypotenuses, and a pair of unequal angles, from which we deduce (35) that $OH < OH'$.

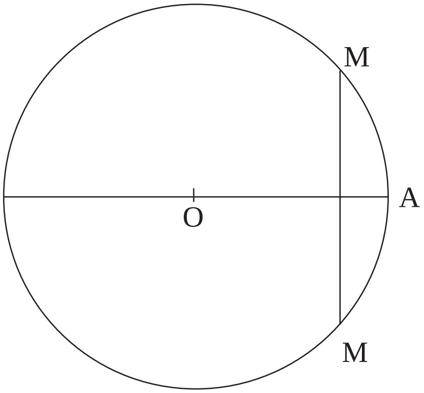

Figure 64

67. Consider a chord MM' (Fig. 64) whose distance to the center, at first less than the radius, increases until it equals the radius. Assume, to be definite, that this chord moves while remaining perpendicular to a fixed diameter OA. According to the preceding theorem, the length MM' decreases as this chord approaches the tangent at A, and we can take point M so close to point A (that is, the chord so close to the tangent), as to make this length as small as we wish, since $MM' < 2MA$. We see thus that the points M, M' approach each other and tend to coincide with the point A; this is expressed informally by saying that the tangent has a double point in common with the circle. We will see that this manner of speaking allows us to simplify the statements of certain theorems.

Exercises

Exercise 50. A circle passes through two fixed points A, B. Let C be one of the points where this circle meets a fixed line perpendicular to AB. Find the locus of the point diametrically opposite to C as the circle varies while passing through A and B.

Exercise 51. By dividing a chord into three equal parts and joining the division points to the center, we do not obtain three equal angles. Which is the greatest of the three angles? (Exercise 7.) Generalize to more than three equal parts.

Exercise 52. If two chords of a circle are equal, and we extend them (if necessary) to their point of intersection, the segments joining this intersection point with the endpoints of the two chords form two equal pairs.

Exercise 53. What is the locus of the midpoints of the chords of a given length of a circle?

Exercise 54. What is the shortest chord which passes through a given point inside a circle?

CHAPTER III

The Intersection of Two Circles

68. According to the theorem in 57, two circles cannot have more than two points in common.

Theorem. When two circles intersect, the line joining their centers is perpendicular to their common chord and divides it into equal parts. When there is only one common point, this point is on the line joining the centers. Conversely, if there is a common point on the line of their centers, the circles have only one point in common.

Indeed, the line of the centers is a common line of symmetry of the two circles. If the two circles meet at a point not on this line, they must meet again at the symmetric point. If there is only one intersection point, then it must belong to the line of the centers.

Conversely, if there is a common point on the line of the centers, then this is the only common point. Otherwise, the second common point is either on the line of the centers, in which case the circles would have a common diameter, or outside this line, in which case there would be a third common point. Under either of these circumstances, the two circles must coincide.

69. **Definition.** We say that two curves are tangent if they have the same tangent line at a common point.

According to the preceding theorem, in the case of two circles this definition becomes: Two circles are tangent when they have exactly one common point, because the tangents at this common point must be the same (58. Corollary).

70. Consider two circles with centers O, O' and radii R, R' and suppose, to be definite, that $R \geq R'$. Five cases are possible:

1°. $OO' > R + R'$ (Fig. 65). Let M be a point inside or on circle O', so that $O'M \leq R'$. We will have (26, Corollary)

$$OM \geq OO' - O'M > R + R' - O'M \geq R.$$

Therefore, every point in the second circle is exterior to the first, and every point on the first is exterior to the second. The two circles are said to be exterior to each other.

2°. $OO' = R + R'$. In this case the segment OO' can be viewed as the sum of two segments OA, $O'A$ (Fig. 66) equal to R, R', respectively. The point A is common to the two circles, and for all other points the preceding reasoning applies. The two circles are then tangent externally.

3°. $R + R' > OO' > R - R'$. In this case, R' is contained between the sum and the difference of the two lengths OO' and R.

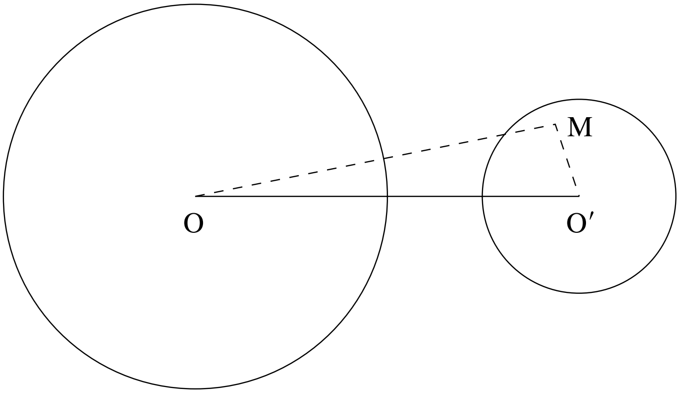

Figure 65

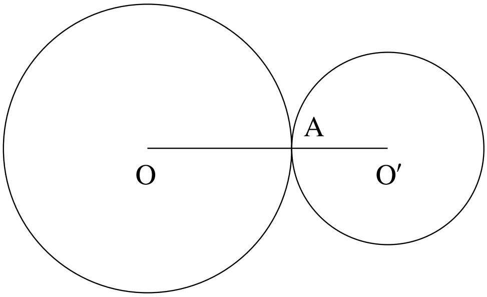

Figure 66

Therefore, of the two points A, B where circle O intersects the line of the centers, one is interior and the other exterior to circle O′ (Fig. 67). It follows that circle O, which is a curve joining A and B, meets circle O′ in a point different from either A or B, and therefore not on the line of the centers.

The circles have two common points, and are said to be intersecting.

4°. $OO' = R - R'$. In this case, OO' can be viewed as the difference of two segments OA, O′A (Fig. 68) equal to R, R′, respectively. Point A on the line of the centers is common to both circles, so the two circles are tangent.

Let M′ be any point on or inside the circle O′. We have

$$OM' \le OO' + O'M' \le OO' + R';$$

in other words, $OM' \le R$.

The entire circle O′, except for point A, is therefore interior to O. The two circles are said to be tangent internally.

5°. $OO' < R - R'$ (Fig. 69). Consider again a point M′ on or inside circle O′. We obtain

$$OM' \le OO' + O'M' \le OO' + R' < R.$$

Thus circle O′ is entirely inside O. The circle O′ is said to be interior to O.

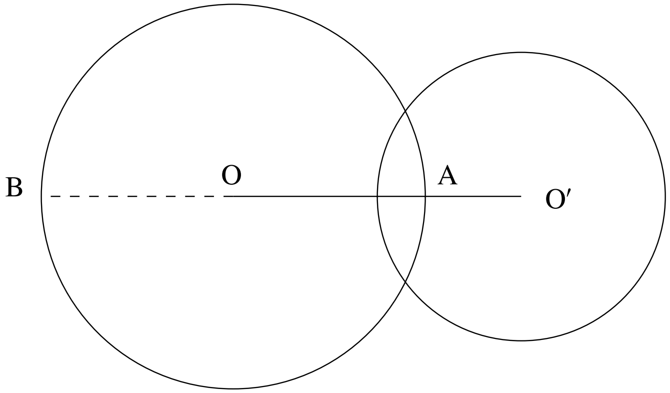

Figure 67

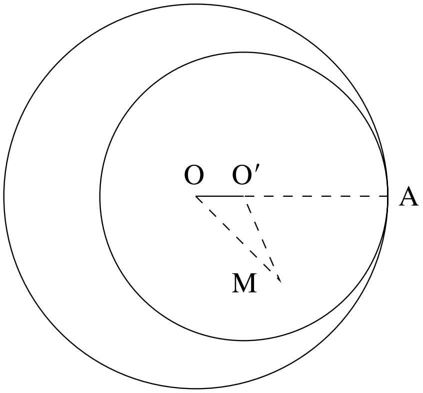

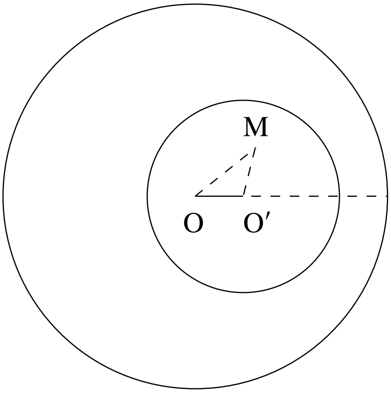

Figure 68

Figure 69

71. The preceding list exhausts the possibilities. It follows that the converses of the conclusions are also true. For example, if two circles intersect, the distance between the centers is contained between the sum and the difference of their radii: if this distance were greater than the sum of the radii, the circles would be exterior; if this distance were equal to the sum, the circles would be tangent externally, etc.

In any case, all of these converses are easily proved directly. Thus, in the case of intersecting circles, the point of intersection forms a triangle with the two centers, and the theorems of 26 yields the preceding conclusion.

We can therefore state the following theorem.

Theorem. 1 . If two circles are exterior, the distance between their centers is greater than the sum of the radii, and conversely.

2 . If two circles are tangent externally, the distance between their centers is equal to the sum of the radii, and conversely.

3 . If two circles intersect, the distance between their centers is less than the sum of the radii and greater than their difference, and conversely.

4 . If the circles are tangent internally, the distance between their centers is equal to the difference of the radii, and conversely.

5°. If one circle is interior to another, the distance between their centers is less than the difference of the radii, and conversely.

One can also say: Two circles are exterior, interior, intersecting, or tangent according as whether the segments they determine on the line of the centers are exterior to each other, are interior to each other, are extensions of each other, or overlap with a common endpoint.

72. If two circles, at first intersecting, vary so as to become tangent at A, their common points approach the point A indefinitely (see Exercise 55). One can informally say, as before (67), that two tangent circles have two common points which coincide, or a double common point.

Exercises

Exercise 55. Consider a circle O of radius R, and a circle O' of radius R' intersecting the first. Let A be one of the points where the line OO' intersects the circle O, and let B be another point on the same circle chosen arbitrarily, with particular interest when it is close to A.

Show that the two circles have one of their common points on the little arc $\overset{\frown}{AB}$:

When the point O' is on the extension of OA past A, if the difference $R + R' - OO'$ is less than $OB + O'B - OO'$;

When O' is between O and A, if the difference $OO' - (R - R')$ is less than $OO' - (OB - O'B)$;

When O is between O' and A, if the difference $OO' - (R' - R)$ is less than $OO' - (O'B - OB)$.

Exercise 56. What are the shortest and longest segments one can draw joining two circles?

Exercise 57. What is the locus of the centers of circles of fixed radius which are tangent to a fixed circle?

Exercise 58. If a line passing through the common point of two tangent circles intersects these circles again in two other points, then the radii ending at these points are parallel.

Exercise 59. Two circles which are tangent internally remain tangent if their radii are increased or decreased by the same quantity, without changing their centers. Two circles which are tangent externally remain tangent if, without changing their centers, one radius is increased by some quantity and the other is decreased by the same quantity.

CHAPTER IV

Property of the Inscribed Angle

73. An angle is said to be inscribed in a circle if it is formed by two chords AB, AC with a common endpoint. Thus an inscribed angle has its vertex on the circle.

Theorem. The measure of an angle inscribed in a circle equals one half the arc contained between its sides. (The endpoints B, C determine two arcs on the circle. The arc $\overset{\frown}{BC}$ which concerns us is the one which does not contain A.)

An inscribed angle equals half the central angle which subtends the same arc.[1]

N.B. The first part of the above statement assumes the convention of 17 (corollary).

We distinguish three cases.

First Case. One of the sides of the inscribed angle passes through the center.

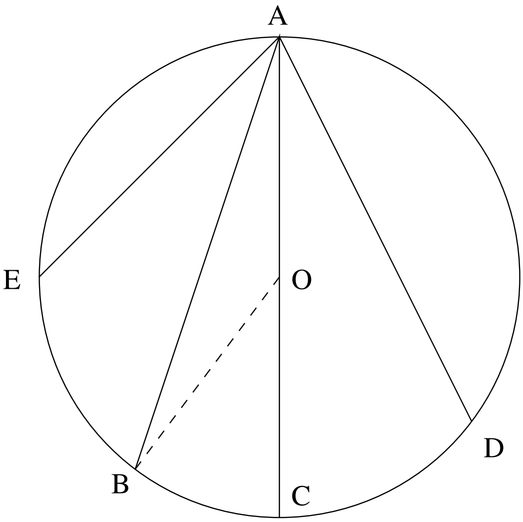

Figure 70

Consider inscribed angle BAC (Fig. 70) whose side AC passes through the center O. We join BO. Thus we form an isosceles triangle OAB whose angles A and B are equal. The exterior angle BOC, equal to the sum A + B, is therefore twice A. Since the measure of BOC is precisely that of arc $\widehat{BC}$, the measure of BAC is one half of this arc.

---

[1] The intercepted arc thus can be either minor arc $\overset{\frown}{BC}$ (20b) or major arc $\overset{\frown}{BC}$. In the last case, the corresponding central angle will be greater than two right angles, and the inscribed angle will be obtuse.

Second Case. The center is inside the inscribed angle.

Consider inscribed angle BAD (Fig. 70). Joining AO, which intersects the circle in C, we decompose the inscribed angle into two parts BAC, CAD for which the theorem has already been proved (first case). We have therefore (with the convention of 18)

$$BAC = \frac{1}{2}\text{arc } \widehat{BC}, \quad CAD = \frac{1}{2}\text{arc } \widehat{CD},$$

and adding,

$$BAD = \frac{1}{2}\text{arc } \widehat{BD}\ .$$

Third Case. The center is outside the inscribed angle.

Consider inscribed angle BAE (Fig. 70). We again construct diameter AC, and we can write

$$BAC = \frac{1}{2}\text{arc } \widehat{BC}, \quad EAC = \frac{1}{2}\text{arc } \widehat{EC},$$

and subtracting,

$$BAE = \frac{1}{2}\text{arc } \widehat{BE}\ .$$

Corollaries. I. All the angles inscribed in the same arc of a circle are equal, because they have the same measure.

II. An angle inscribed in a semicircle is a right angle, because its measure is a quarter of a circle.

74. Theorem. The angle formed by a tangent and a chord ending at the point of contact is equal to one half the arc it intercepts on the circle.

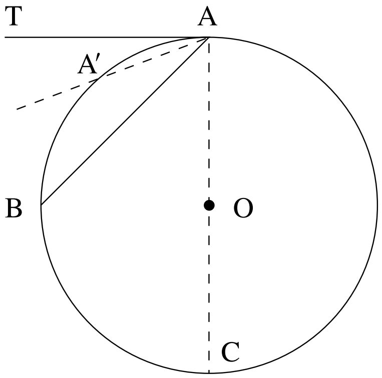

Figure 71

Angle BAT, formed by tangent AT and chord AB (Fig. 71) can in effect be considered as the limit of the angle formed by AB with chord AA when point A becomes infinitely close to A.

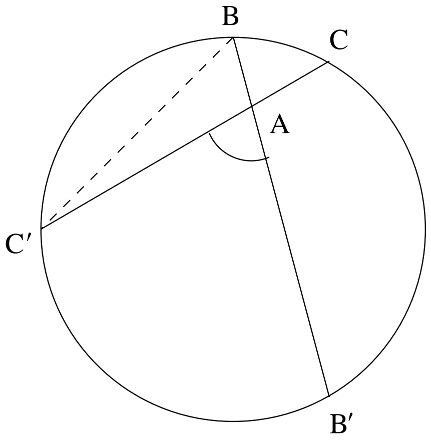

Figure 72

Alternatively, the proof given for the third case above can also be applied here; the equality $EAC = \frac{1}{2}$arc $\overset{\frown}{EC}$ must simply be replaced by

$$TAC = \frac{1}{2}\text{arc } \overset{\frown}{AC},$$

which follows from the fact that angle TAC is a right angle and arc $\overset{\frown}{AC}$ is a semicircle.

75. **Theorem.** The angle formed by two secants which intersect inside a circle is equal to one half the sum of two intercepted arcs: one which is between its sides and the other which is between the extensions of its sides.

Consider angle BAC (Fig. 72) whose sides intersect the circle again in B and C, and join BC. Angles C and B are one half the arcs $\overset{\frown}{BC}$, $\overset{\frown}{B\ C}$. But these two angles have a sum equal to A, an exterior angle for triangle ABC.

76. **Theorem.** The angle formed by two secants which meet outside a circle is equal to the half the difference of the two arcs contained between its sides.

Consider angle BAC formed by secants $AB\ B$, $AC\ C$. Angle $BC\ C$, exterior to triangle ABC, is equal to the sum $A + B$. Thus we can measure angle A as the difference $BC\ C \quad B$. This proves the theorem, since inscribed angle $BC\ C$ intercepts arc $\overset{\frown}{BC}$, and B intercepts $\overset{\frown}{B\ C}$.

Remark. If one of the secants, say $AB\ B$, is replaced by the tangent AT (Fig. 74), the theorem and its proof hold without change, if we let T play the roles of both B and B; in other words, by using the fact that the tangent meets the circle in two points coinciding at T, following the convention of 67.

77. **Theorem.** The locus of points situated on one side of a line, from which a given segment of the line subtends a given angle, is a circular arc ending at the endpoints of the given segment.

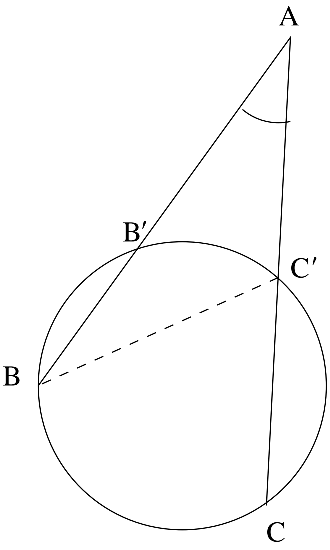

Figure 73

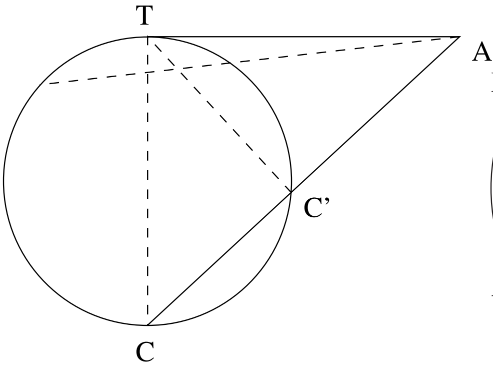

Figure 74

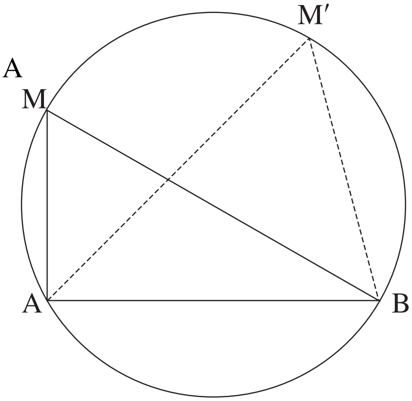

Figure 75

Indeed, let the given segment be AB, and let M be a point on the locus;[2] in other words, such that the angle AMB is equal to the given angle (Fig. 75). If we draw an arc of a circle through A, M, and B, with endpoints at A and B, any point on this arc will belong to the locus, because AM′B = AMB (by 73). On the other hand, any point N of the plane, situated on the same side of AB but not on the arc, is either outside or inside this circle. In the first case, angle ANB will be smaller than the given angle (76); in the second it will be larger (75), and therefore N will not belong to the locus. QED

---

[2] Such points certainly exist. One of them is obtained by constructing, through A and B, parallels to any angle equal to the given one and choosing, among the vertices of the parallelogram thus formed, the one which is on the required side of the line AB.

The circular arc described above is called the arc capable of the given angle relative to AB.[3]

If we do not require the points on the locus to be on one side of the line AB, then the locus will consist of two circular arcs, symmetric witih respect to AB.

78. When the given angle is a right angle, the preceding result combined with Corollary II in 73 shows that the locus of points where a given segment subtends a right angle is the circle having the segment as diameter.

This also follows from the two corollaries in 48.

79. Four randomly chosen points will not generally belong to the same circle, since any three of them determine a circle which, in general, will not pass through the fourth.

A necessary condition for four points to belong to a circle is given in the following result.

Theorem. In any convex quadrilateral inscribed in a circle, the opposite angles are supplementary.

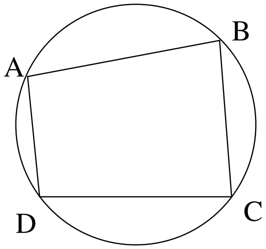

Figure 76

In quadrilateral ABCD inscribed in a circle (Fig. 76):

Angle A is equal to half arc $\overset{\frown}{DCB}$;

Angle C is equal to half arc $\overset{\frown}{DAB}$.

The sum of these two arcs is the whole circle, so that the sum A + C is equal to a central angle corresponding to a semicircle; that is, to two right angles.[4]

80. Converse. If two opposite angles of a convex quadrilateral are supplementary, then the quadrilateral is cyclic.

Suppose the quadrilateral is ABCD (Fig. 76) in which angles A and C are supplementary. The arc of the circle joining D, A, B, extended beyond BD, is the locus of the points from which BD subtends a constant angle (77), which is (79) supplementary to angle A. This arc therefore passes through point C.

81. Remark. In any quadrilateral ABCD one can consider the four angles A, B, C, D, and also, by drawing the diagonals AC, BD, the eight angles numbered 1 to 8 in Figure 77. If the quadrilateral is cyclic, then:

---

[3] The term in French is simply the word capable, which has the same ordinary meaning as in English. But it is not used in English in this mathematical sense. We will not use it in this edition. Transl.

[4] Hadamard uses the term inscribable (inscriptible), but without defining it, for a quadrilateral whose vertices lie on a circle. From this point on, we will use the standard term cyclic. Transl.

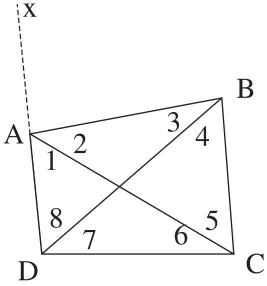

Figure 77

1 . A and C are supplementary;

2 . B and D are supplementary;

3 . 1 and 4 are equal;

4 . 5 and 8 are equal;

5 . 2 and 7 are equal;

6 . 3 and 6 are equal.

Conversely, if any one of these conditions holds, the quadrilateral is cyclic (77, 80).

Thus, any one of the conditions 1 , 2 , 3 , 4 , 5 , 6 implies the others.

82. Conditions 1 and 2 in the preceding remark can be rephrased in a form closer to 3 6 .

We note first that since points A, B (Fig. 77) are on the same side of CD, triangles CDA, CDB have the same sense of rotation, so that the equal angles DAC, DBC have the same sense. The same is true of angles ACB, ADB, etc. On the other hand, angles DAB, DCB have the opposite sense, since A and C are on different sides of BD. So we extend DA beyond point A along Ax, to form an angle xAB with the same sense as DCB: we see that these angles are equal if the angles DAB, DCB are supplementary, and we arrive at the following conclusion:

If four points A, B, C, D are on the same circle, the angle of the same sense formed by AC, AD on the one hand, and BC, BD on the other, are equal; and any statement that can be derived from this one by permuting the letters A, B, C, D is also true.

Conversely, if any one of the above six conditions holds, then the quadrilateral is cyclic. Therefore, each of the conditions 1 , 2 , 3 , 4 , 5 , 6 implies all the others.

82b. The statement of 77 can also be replaced by the following: The locus of vertices of all angles of a given size, whose sides (extended, if necessary) pass through two given points, is a circle passing through these two points.

Since every circle can be obtained in this manner, this statement can be considered as a new definition of a circle, equivalent to the first, and which can replace it if necessary.

Exercises

Exercise 60. Find the locus of the midpoints of the chords of a circle passing through a fixed point.

Exercise 61. On each radius of a circle we lay off, starting from the center, a length equal to the distance from the endpoint of the radius to a fixed diameter. Find the locus of the endpoints of the segments thus obtained.

Exercise 62. Given a fixed circle, and a fixed chord AB in this circle. Let CD be a variable chord, whose length stays fixed.

1°. Find the locus of the intersection I of the lines AC, BD.

2°. Find the locus of the intersection K of the lines AD, BC.

3°. Find the locus of the centers of the circles circumscribing the two triangles ICD, KCD. Show that these are respectively congruent to the two loci found above.

Exercise 63. Points A and B are fixed on a circle, and M is a variable point on the same circle. We extend segment MA by a length MN = MB. Find the locus of point N.

Exercise 64. Points A, B, C are three points on a circle. We join the midpoints of the arcs $\overset{\frown}{AB}$, $\overset{\frown}{AC}$. Show that this line intercepts equal segments (starting from A) on chords AB, AC.

Exercise 65. If we draw two arbitrary secants through the common points A, B of two circles, the chords joining the new intersections of these lines with the two circles are parallel.

Exercise 66. The bisectors of the angles of an arbitrary quadrilateral form a cyclic quadrilateral. The same is true for the bisectors of the exterior angles.

Exercise 67. Through the midpoint C of arc $\overset{\frown}{AB}$ of a circle, we draw two arbitrary lines which meet the circle in D, E, and chord AB in F, G. Show that DEFG is cyclic.

Exercise 68. We are given a circle, a fixed point P on this circle, and also a line, and a fixed point Q on this line. Through points P, Q we draw a variable circle which cuts the given circle again at R, and the given line at S. Show that line RS intersects the given circle in a fixed point.

Exercise 69. Two circles intersect at A and B. Through A we pass a variable line which cuts the circles again in C, C'. Show that the segment CC' subtends a constant angle at B, equal to the angle of the two radii ending in C, C'.

Through point A we draw a second secant, cutting the circles in D and D'. Show that the angle between the chords CD, $C'D'$ is equal to the angle found earlier, or to its supplement.

What becomes of the last statement when the two secants coincide?

Exercise 70. In any triangle, the points which are symmetric to the intersection of the altitudes with respect to the three sides lie on the circumscribed circle.

Exercise 71. Show that the altitudes of a triangle are the bisectors of the triangle formed by their feet.

We can apply the method of 82 to the quadrilaterals formed by two altitudes and two sides, show that these quadrilaterals are cyclic, and use the resulting angle properties.

The same method applies in a number of situations, such as the one in the following problem.[5]

Exercise 72. The feet of the perpendiculars, dropped from a point of the circumscribed circle onto the three sides of a triangle, are on the same line (the Simson Line).

Conversely, if the feet of the perpendiculars, dropped from a point in the plane onto the three sides of a triangle, are on the same line, this point must be on the circumscribed circle.

---

[5] This kind of reasoning is also useful in many of the problems proposed at the end of Book II.

CHAPTER V

Constructions

83. Since ancient times, it has been the practice to reserve the name geometric constructions for those constructions done with straightedge and compass. These constructions, theoretically, have absolute precision. In practice, they are effectively very precise, but nonetheless, as with any such process, offer some opportunity for error (such as the width of the marks left by a pencil).

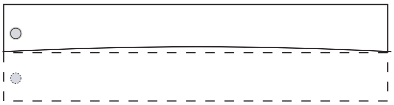

Figure 78

The straightedge is an instrument designed to draw straight lines. The accuracy of a straightedge, that is, the fact that its edge is really straight, depends on the care taken in its manufacture. To verify that this condition is satisfied, we use the actual definition of a straight line. We trace a first line along the edge of the straightedge (Fig. 78); then we turn the instrument around, placing it on the other side of the drawing. We should be able, in this new position, to trace along the edge of the straightedge a second line coinciding with the first.

The compass is composed of two pointed, articulated rods. The opening of the compass is the distance which separates the points of the two rods, and this instrument allows one to transfer distances, or to draw a circle with radius equal to this opening, with any center. This instrument is accurate if the opening does not change while the compass is moved.

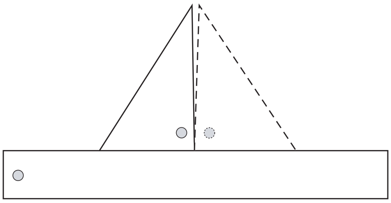

Figure 79

84. Besides the instruments just discussed, there are others which are used in practice, such as the set-square and the protractor. The set-square is a flat piece of wood (or other material) in the shape of a right triangle. A good set-square must satisfy two conditions:

1 . the sides must be straight;

2 . the angle must be a right angle.

The first of these is verified as in the case of the straightedege; it is usually satisfied in common set-squares. To verify the second one, we place the set-square on the edge of a straightedege (Fig. 79) and we trace a straight line along the side which should be perpendicular. Then, placing the set-square in the opposite sense, we should be able to draw a straight line coinciding with the first. Often, this condition is found to be only approximately satisfied, and one cannot rely on it for exact geometric constructions.

Finally, the protractor is an instrument for measuring angles. It is usually in the shape of a semicircle made of horn or copper, and divided into 180 parts. Since this division can only be done approximately, the protractor, while useful in practice, is not a geometric instrument.

85. Construction 1. (First fundamental construction). Construct the perpendicular bisector of a given segment.

This perpendicular is the locus of points equidistant from the endpoints A, B of a segment (33). So if we draw arcs, with centers A and B, which are sufficiently large that they intersect, the intersection points will belong to the perpendicular bisector; it remains simply to join these points.

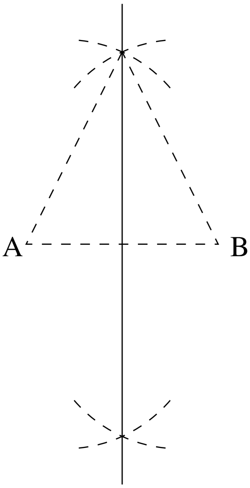

Figure 80

The following two constructions reduce to this one.

Construction 2. Drop a perpendicular from a point to a line.

With the given point O (Fig. 81 or 81b) as center, we draw a circle which intersects the line in points A, B. The perpendicular bisector of AB passes through O because OA = OB; it is therefore the desired perpendicular.

If point O is not on the line, we can take the distance OA as the common radius of the circles from construction 1 (used to find the perpendicular bisector of AB): thus we need not change the opening of the compass. These two circles will intersect at O and again at a point O′, symmetric to the first with respect to

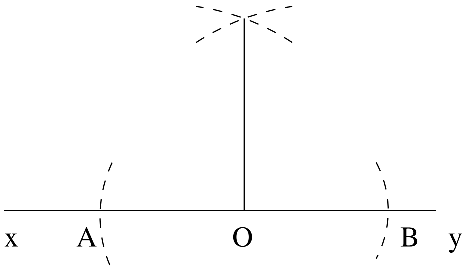

Figure 81

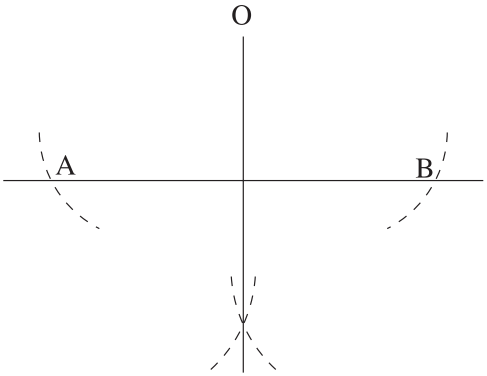

Figure 81b

AB. This simplification is not possible when O is on the line, because in that case O = O. Even when O is very close to the line it is preferable to use a different radius, as the proximity of O and O will not allow for a good determination of the line between them.

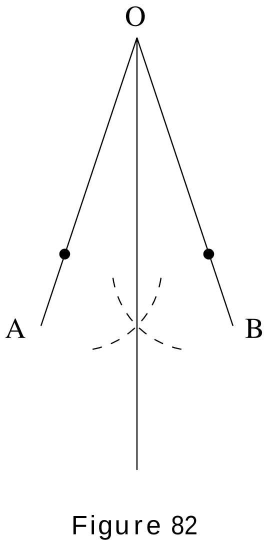

Figure 82

Construction 3. Construct the bisector of a given angle AOB (Fig. 82).

We lay off two equal segments OA, OB on the sides of the angle, so as to form an isosceles triangle. The perpendicular bisector of AB is the desired angle bisector (23).

The same simplification as before applies to the radii of the circles.

86. Construction 4. (Second fundamental construction.) Construct a triangle, knowing its three sides.

Taking segment BC equal to the first side (Fig. 83), we draw circles centered at B, C with radii equal to the other two sides; their intersection A gives the third vertex of the triangle. Thus there are two possible choices for this third vertex, since two circles have two points in common, but the two resulting triangles are symmetric with respect to BC, and so are congruent.

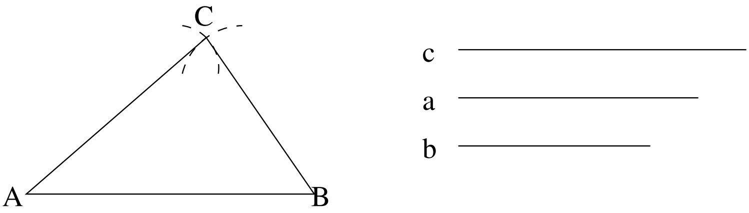

Figure 83

Condition for this construction to be possible. In order for the two circles to intersect, it is necessary and sufficient (70, 71) that the first side be less than then sum of the others, and greater than their difference or (26, Corollary), that each side be less than the sum of the other two.

If the order of the lengths of the three sides is known, it is enough to write that the longest side is smaller than the sum of the other two.

Construction 5. Through a point on a line, construct another line, forming with the first an angle equal to a given angle.

On the sides of the given angle O, we take two arbitrary points A, B. We then construct a triangle congruent to OAB, starting with a side equal to OA on the given line, and with an endpoint at the given point.

In practice, the triangle OAB is taken to be isosceles, so that we only need two different settings for the opening of the compass.

Construction 6. Knowing two angles of a triangle, find the third.

We construct two adjacent angles equal respectively to the two known angles. By extending one of the exterior sides we forms an angle supplementary to the sum of the first two. This is the required angle.

Condition for this construction to be possible. The sum of the two given angles must be less than two right angles.

Construction 7. Construct a triangle knowing two sides and the angle they contain.

On the sides of an angle equal to the given one, we lay off segments, starting from the vertex, equal to the given sides.

Construction 8. Construct a triangle knowing one side and two angles.

First, knowing two angles, we can find the third. In particular, we know the two angles adjacent to the given side; we place these angles at the endpoints of this side (Construction 5).

The conditions for this construction to be possible are the same as those for Construction 6.

87. Construction 9. Construct a triangle knowing two sides, and the angle opposite one of them.

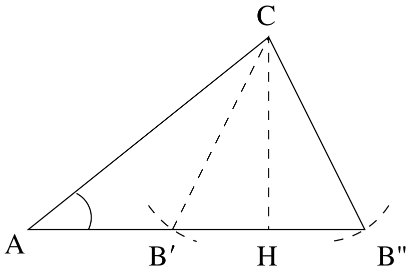

Figure 84

Let the triangle be ABC, in which we are given angle A and sides a, b, opposite A, B, respectively. After constructing an angle equal to A (Fig. 84), we lay off a segment AC equal to b on one of its sides. This determines point C, and we can then find B by drawing a circle with center C and radius a. The intersection of this circle with the second side of angle A gives us vertex B, which determines the triangle.

Discussion. We have a solution whenever the circle constructed above intersects ray Ax (Fig. 84), the second side of angle A.

We drop a perpendicular CH onto Ax.

1 . If CH > a, the circle does not intersect Ax and the construction is impossible.

2 . If CH < a, the circle intersects line Ax in two points B , B . It remains to see whether these points are on ray Ax or on its extension.

We distinguish two cases:

(a) Angle A is acute. In this case H, the midpoint of B B , is on ray Ax. At least one of the points B , B will also be on this ray, so that there always is at least one solution. There is a second solution if the second point is between A and H, which happens if and only if a < b.

(b) Angle A is obtuse. Then point H will not be on ray Ax; therefore at least one of the points B , B will not provide a solution, and thus there is at most one solution. There is one if the second point is farther from H than A is, which happens if and only if a > b.

87b. Finally, if angle A is a right angle, the problem can be restated thus: Construct a right triangle, knowing the hypotenuse and a side of the right angle. It has a solution if the hypotenuse is greater than the given side.

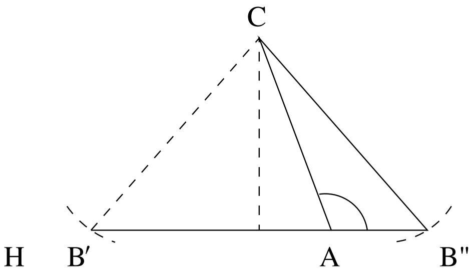

Figure 85

Remark. The following theorem is a result of the previous considerations:

Theorem. When two triangles have two sides equal in pairs, as well as the angle opposite one of them, the angles opposite the second side are either equal or supplementary. The first case (which implies the congruence of the triangles) always holds if the given equal angles are right or obtuse.

88. Construction 10. Through a point, construct a parallel to a given line.

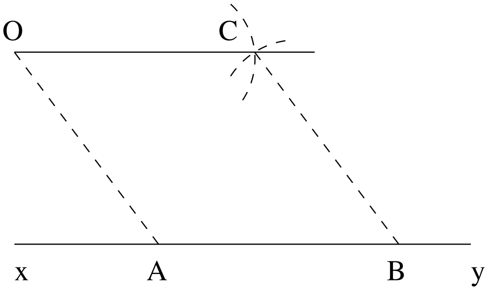

Figure 86

On the given line, we take two arbitrary points A, B (Fig. 86), then draw a circle with center O and radius AB, and a circle with center B and radius OA. These two circles will intersect (on the side of OB opposite to A) in a point C which belongs to the desired parallel. Indeed, OABC is a parallelogram because it has equal opposite sides.

It is convenient to take AB = OA so that all the circles drawn have the same radius.

89. Constructions 2 and 10 can also be performed with the set-square. To perform Construction 2 (To construct a perpendicular to a given line through a given point), we first apply the straightedge to the given line, and then slide one of the sides of the set-square along the straightedge, until its second side, perpendicular to the first, passes through the given point (Fig. 87) and thus gives the desired perpendicular.

This construction assumes the condition, usually not fully satisfied, that the angle of the set-square is a right angle, so it is not used in very precise sketches.

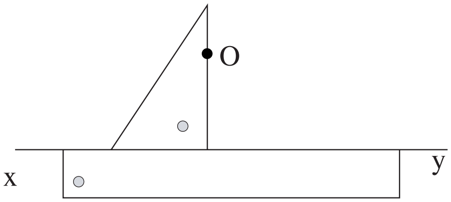

Figure 87

To perform Construction 10: Through a point, construct a parallel to a given line, we first place one side of the set-square along the given line (Fig. 88). Next, applying the straightedge to another side of the set-square and holding it in this position, we slide the set-square along the straightedge until the edge originally placed on the given line passes through the given point. The line thus obtained is parallel to the original one because it makes equal corresponding angles (A, A′ in the figure) with the same transversal (the edge of the straightedge).

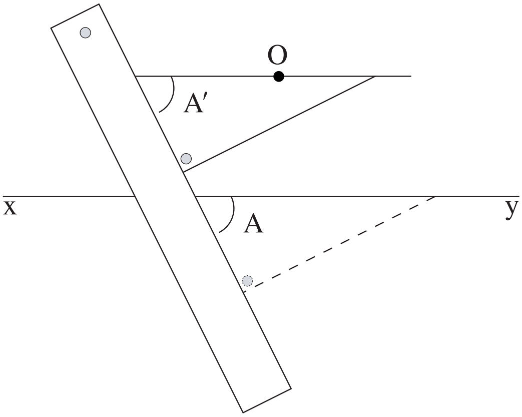

Figure 88

This construction supposes only that the edges of the straightedge and of the set-square are straight. Since it is as accurate as, and simpler than, the compass construction, it is the one that is used in practice.

90. Construction 11. Divide a circular arc into two equal parts.

We draw the perpendicular bisector of the corresponding chord (Construction 1).

Construction 12. Construct the circle passing through three given points A, B, C, not on the same line.

We construct the perpendicular bisectors of two of the segments joining the points: their intersection O is the center of the circle, and OA is its radius.

The circle thus obtained is said to be circumscribed about the triangle formed by the three points.

In general a polygon (or a broken line) is said to be inscribed in a curve, which is then said to be circumscribed about the polygon, when all the vertices of the polygon are on the curve.

Remark. If the three points A, B, C were collinear, there would not be any circle passing through them; we can consider line ABC as replacing it. A straight line can be viewed as a limiting case of a circle, whose center becomes infinitely distant. In fact, line AB is the locus of points at which segment AB subtends a constant angle; namely, an angle of zero degrees, or two right angles. One can also show directly that a circle passing through two fixed points AB, and whose center recedes indefinitely, has the line AB as its limit (one must, of course, give a precise meaning to this statement.)

Construction 13. Construct a circle knowing two of its points A, B, and the tangent AT at one of them.

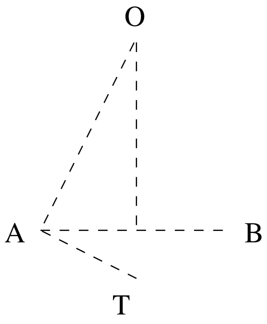

Figure 89

The center O will be the intersection of the perpendicular bisector of AB with the perpendicular to AT at the point A. The circle with center O and radius OA will indeed be tangent to AT, and will pass through B (Fig. 89).

Remark. This construction is a limiting case of the preceding one, following the general definition of tangents (59).

Construction 14. On a given segment, construct the arc of a circle from which the segment subtends a given angle.

Let AB be the given segment. Construct one point M of the locus (see footnote 2 to 77) and construct the circle AMB; or, note that the tangent at A to this circle forms with AB an angle equal to the given one (73, 74), which allows us to construct it (Construction 5), and we have reduced the problem to the preceding construction.

91. Construction 15. Construct the tangent to a given circle at a given point.

We draw a perpendicular to the radius at the point of contact.

Construction 16. Construct a tangent to a circle, parallel to a given line.

The diameter passing through the point of contact is precisely the diameter perpendicular to the given line; this diameter has two endpoints, and therefore there are two solutions.

Construction 17. Construct a tangent to a circle through a point in the plane.

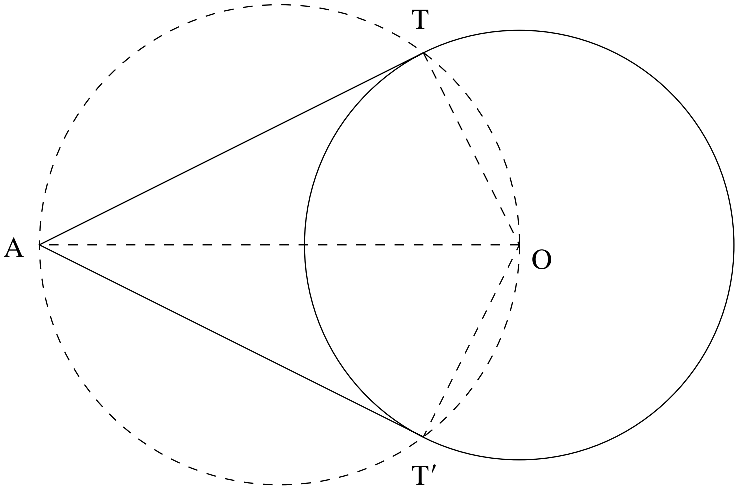

Figure 90

Let the given point be A, and the given circle have center O (Fig. 90). Assume that the problem is solved, and let T be the point of contact of a tangent passing through A. Since ATO is a right angle, point T belongs to the circle with diameter OA. Conversely, if a point T is on the given circle, and on the circle with diameter OA, then the tangent through T passes through A, because ATO is a right angle.

Therefore the points T which satisfy the conditions of the problem are the common points of the given circle and the circle with diameter OA.

Another method. Let the desired point of contact be T. Extend the segment OT beyond T by a length TO′ = OT (Fig. 91). Since TA must be perpendicular to OT, point O′ is symmetric to O with respect to AT, and therefore AO = AO′. Thus point O′ belongs to the circle with center A and radius AO. As it also belongs to a circle centered at O and with a radius twice the radius of the given circle, it is determined by the intersection of these two curves. Conversely, if point O′ is such that AO′ = AO, the midpoint T of OO′ is the point of contact of a tangent from A to the circle with center O and radius OT, because of the properties of an isosceles triangle.

Discussion. Let us start, for example, with the second construction. In order for a solution to exist, it is necessary that the two circles which determine O′ intersect; that is, (66) that the distance between their centers be less than the sum of the radii, and more than the difference. If R is the radius of the given circle,

(1) $$AO \quad 2R + AO$$

and

(2) $$AO \quad AO \quad 2R$$

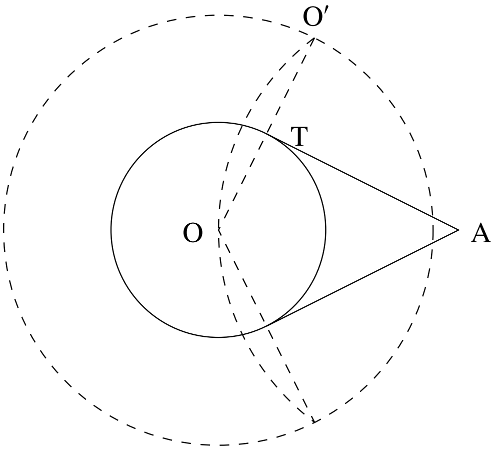

Figure 91

or

$$(3) \qquad AO \quad 2R \quad AO,$$

depending on whether AO is greater than or less than 2R. The inequalities (1) and (2) are obviously satisfied. Inequality (3) can be written 2AO 2R, or AO R. Thus the problem is impossible if A is inside the circle. On the other hand, if A is outside, there are two solutions, since the circles intersect in two points. Finally, if point A is on the circle, the two circles which determine point O are tangent, and there is only one solution, the tangent at A. Or rather, there are two identical solutions because, as we remarked (72), if the point A approaches infinitely close to the circle, the two points of contact, and therefore the two tangents, tend to coincide.

92. Theorem. The two tangents to a circle passing through a given point are equal, and they form equal angles with the diameter passing through the point. This diameter is perpendicular to the chord joining the two points of contact.

Indeed, the two points of contact are symmetric with respect to OA (Fig. 90) because they are the intersection points of two circles with centers on this line (see 68).

93. Construction 18. Draw the common tangents to two circles.

Consider two circles with centers O, O and radii R, R . The common tangents can be of two kinds: they can be external (Fig. 92), which means that both circles are on one side of the line, or internal (Fig. 93) when the two circles are on different sides of the tangent.

1 . Common external tangents. Suppose that a solution exists, and let AA be the required tangent (Fig. 92), with A, A being the points of contact, so that OA, O A are perpendicular to AA . We daw a parallel through O to AA , extended until it intersects OA in C. We will determine this point C.

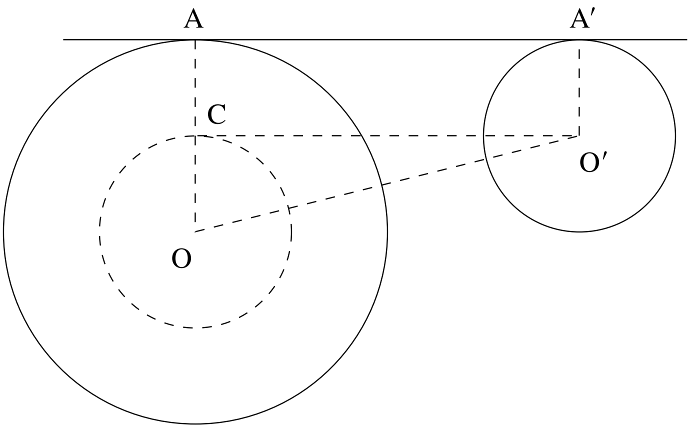

Figure 92

To this end, observe that CAA O is a rectangle, so that AC = R and therefore[1] OC = R R .

The point C therefore lies on a circle we can draw, since we have its center, O, and its radius, which is the difference between the given radii. Also, line CO is tangent to this circle, since angle OCO is a right angle. Thus we are led to Construction 17.

Once point C is determined, extending radius OC gives us point A. The tangent at A to the circle O is then the common tangent, because if we drop a perpendicular O A onto AA , the equality OC = R R gives O A = AC = OA OC = R .

The construction is possible (91) only if OO R R , in other words (70), if the two circles are external to each other, tangent externally, intersecting, or tangent internally. In the first three cases there are two solutions, since we can draw two tangents from O to circle OC; in the fourth case, there is only one solution, which can be considered two coinciding solutions, since the two tangents from O to circle OC coincide.

2 . Common internal tangents. Let BB be a common internal tangent, with points of contact at B and B (Fig. 93). As in the first case, we draw a parallel O D to BB , where D is the intersection of the parallel with line OB. Quadrilateral DBB O is a rectangle, and therefore BD = R , and D belongs to the circle centered at O with radius R + R . We can draw this circle, and draw a tangent to it from O . Once point D is determined, the intersection of OD with the first circle provides the point B. Following this reason backwards, we see that the tangent at B is a common tangent.

The construction is possible only if OO R + R , that is, if the two circles are exterior to each other, or externally tangent. In the first case there are two solutions, in the second only one (or two identical solutions).

In summary:

Two circles external to each other have two common external tangents and two common internal tangents;

---

[1] To be definite, we assume that the circle O has the greater radius. This assumption, which we have the right to make, is not essential here. In the opposite case, we will have OC = R R.

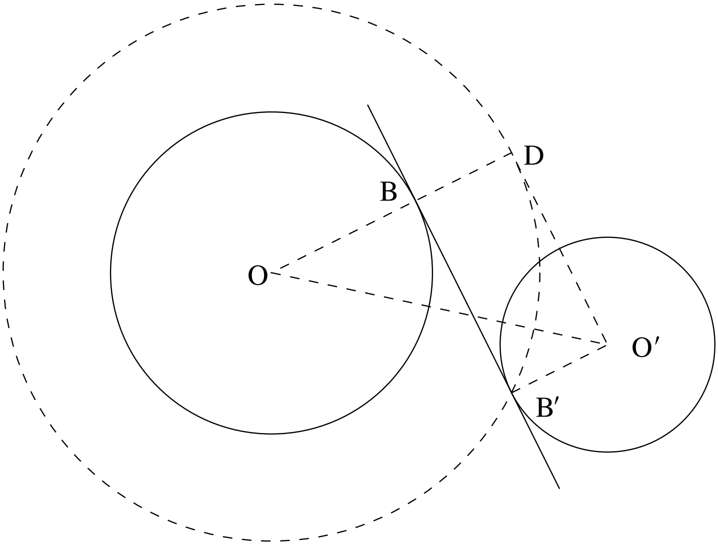

Figure 93

Two circles which are tangent externally have two common external tangents, and one common internal tangent (at their point of contact);

Two intersecting circles have two common external tangents;

Two circles which are tangent internally have one common external tangent (at their point of contact);

Two circles, one of which is interior to the other, have no common tangents.

94. Construction 19. Construct the circles tangent to three lines.

The problem amounts to this: find the points equidistant from three lines. A circle with its center at such a point, and the common distance as radius, will answer the question.

Now, the locus of points equidistant from two intersecting lines consists of the two bisectors of the angles formed by these lines.

Let us assume first that each pair of the three lines intersect, but not all three pass through the same point, so that they form a triangle (Fig. 94). In this triangle we construct the bisectors of the interior and exterior angles. We already know that: 1° the bisectors of the interior angles are concurrent; 2° the bisector of any interior angle has a point in common with the bisectors of the non-adjacent exterior angles. The four points thus defined (the intersection of the interior bisectors, and the three vertices of the triangle formed by the exterior bisectors) are therefore the desired points, and there are no others.

The first of the circles with centers at these points is inside the triangle. It is called the inscribed circle: in general, a curve is inscribed in a polygon when it is tangent to all of its sides. The three other circles, also tangent to the three sides, but outside the triangle, are said to be escribed: each one of them is situated inside one of the angles of the triangle, but separated from its vertex by the opposite side, as indicated in the figure.

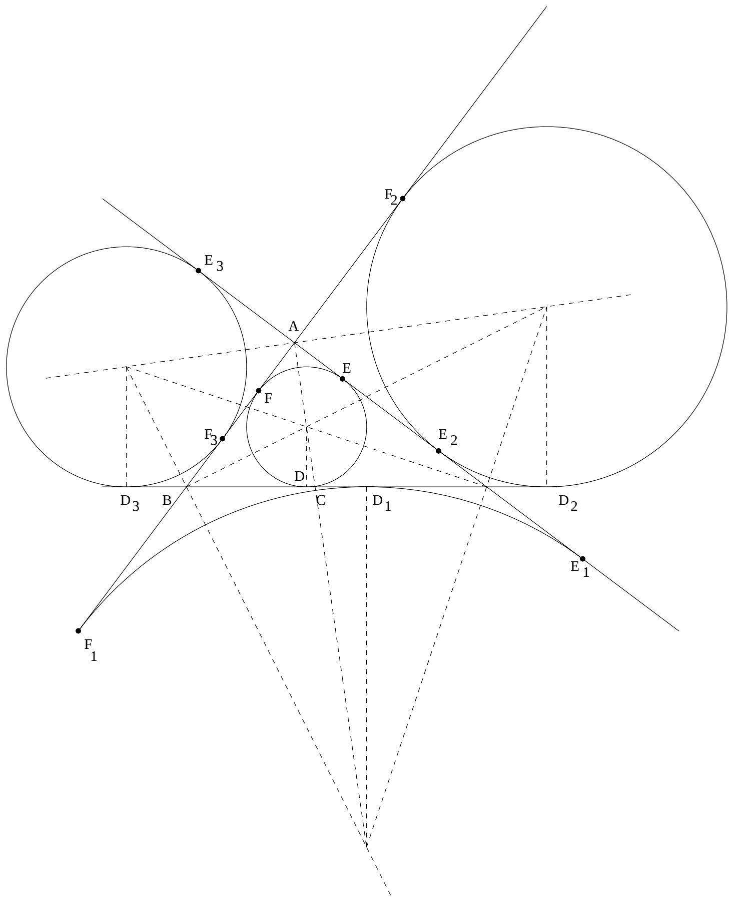

Figure 94

If the three given lines pass through the same point, there is no solution properly speaking, because one can draw only two tangents from a point to a circle. The preceding construction gives a circle reduced to one point; namely the, common point of the three lines.

If two of the lines are parallel, and the third is a transversal (Fig. 95), the bisectors of the angles formed by the transversal with the other two lines form two parallel pairs. Therefore we obtain only two circles, on either side of the transversal.

Finally, if the three lines are parallel, there is no solution, because a circle has only two tangents with a given direction.

Exercises

Exercise 73. Construct a circle of given radius:

(i) which passes through two given points;

(ii) which passes through a given point and is tangent to a given line;

(iii) which is tangent to two given lines;

(iv) which is tangent to a given line and a given circle.

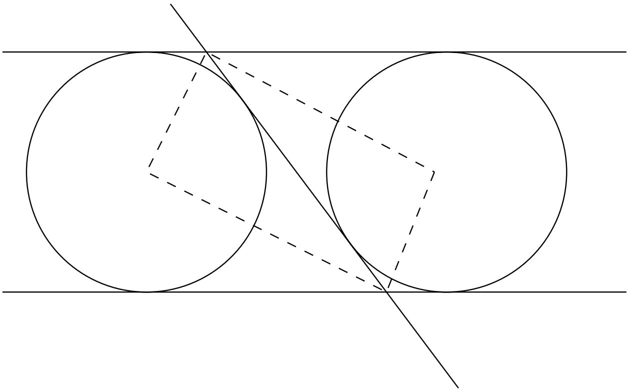

Figure 95

Exercise 74. Construct a circle tangent to a given line at a given point and also tangent to a given circle.

Exercise 75. Between two given lines (or circles), construct a segment of given length and direction.

Exercise 76. Construct a tangent to a given circle, on which a given line determines a segment of a given length.

Exercise 77. Construct a triangle knowing a side, the opposite angle, and an altitude (distinguish two cases, according as the given altitude corresponds to the given side or not).

Exercise 78. Divide a segment into three equal parts using the theorem of no. 56 on the medians of a triangle.

Exercise 79. Construct a triangle knowing:

(i) two sides and a median (distinguish two cases, according as the given median falls on one of the given sides or is contained between them);

(ii) one side and two medians (two cases);

(iii) the three medians.

Exercise 80. Construct a triangle knowing a side, an altitude, and a median (five cases).

Exercise 81. Construct a triangle knowing an angle, an altitude, and a median (five cases).

Exercise 82. Construct a triangle knowing one side, the opposite angle, and the sum or the difference of the other sides.

Exercise 83. Same problem, but the angle is adjacent to the given side.

Exercise 84. Construct a triangle knowing an angle, an altitude, and the perimeter (two cases).

Exercise 85. Construct a triangle knowing one side, the difference of the adjacent angles, and the sum or difference of the other sides.

Exercise 86. Construct the bisector of the angle formed by two lines which cannot be extended to their intersection.

Theorems on Tangents (see 92).

Exercise 87. In a quadrilateral circumscribed about a circle (and containing the circle in its interior), the sum of two opposite sides is equal to the sum of the other two opposite sides. Conversely, a convex quadrilateral in which the sums of the opposite sides are equal can be circumscribed about a circle.

While we cannot be sure that a quadrilateral circumscribed about a circle contains the circle in its interior, we can nonetheless always say that the sum of two side of the quadrilateral, appropriately chosen, is equal to the sum of the two others, and conversely.

Exercise 88. A circle tangent to three sides of a convex quadrilateral is on the same side as the interior of the quadrilateral, relative to these three sides. Show that this circle intersects or does not intersect the fourth side according as the sum of this side and the opposite side is greater than or less than the sum of the other sides.

Exercise 88b. If two points are outside a circle, the line joining them is exterior to or secant to this circle according as this segment between the two points is or is not contained between the sum and the difference of the tangents from these points to the circle.

Exercise 89. The segment intercepted by two fixed tangents of a circle on a variable tangent subtends a constant angle at the center. Special case: the fixed tangents are parallel.

Exercise 90. If a variable tangent to a circle intersects two fixed tangents, and the point of contact of the variable tangent is on the smaller arc determined by the point of contact of the fixed tangents, the triangle formed by the three tangents has a constant perimeter.

What can we say about this situation if the point of contact of the variable tangent lies on the larger arc determined by the fixed tangents?

Exercise 90b. Denote by a, b, c the three sides of a triangle ABC, and let p be the semiperimeter of the triangle (so that $2p = a + b + c$). Let D, E, F be the points of contact of the inscribed circle, and D_1, E_1, F_1, D_2, E_2, F_2, D_3, E_3, F_3 be the points of contact of the escribed circles in angles A, B, C, respectively. Then the segments intercepted by these points have the following values:

$$AE = AF = CD_2 = CE_2 = BD_3 = BF_3 = p - a;$$

$$BF = BD = AE_3 = AF_3 = CE_1 = CD_1 = p - b;$$

$$CD = CE = BF_1 = BD_1 = AF_2 = AE_2 = p - c;$$

$$AE_1 = AF_1 = BF_2 = BD_2 = CD_3 = CE_3 = p.$$

Exercise 91. Construct three pairwise tangent circles with centers at the three vertices of a given triangle.

CHAPTER VI

On the Motion of Figures

95. We have defined (20) the notion of congruent figures with the same orientation. We know (50) that if two points of a figure coincide with the corresponding points in a congrent figure, then the two figures coincide completely. In other words, there is only one way to place a figure F congruent to a given figure F with the same orientation[1] in such a way that two points of this figure (the points corresponding to two given points A, B of F) have a given position.

In fact, with this information (knowledge of the homologues A , B of A, B), we can determine the point C corresponding to another point C of F.

Indeed, it suffices to construct the triangle A B C congruent to A, B, C (Construction 4) then choose, from the two possibilities for the third vertex, the one which corresponds to the same orientation.

Since this can be done for every point of F , we see that we can solve (at least for figures composed of finitely many points and lines) the following:

Problem. Construct a plane figure F congruent to a given figure F (and with same orientation), knowing the points corresponding to two given points.

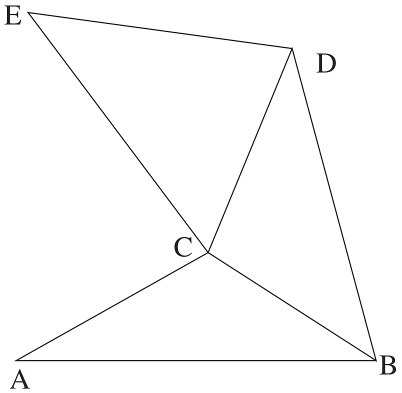

Figure 96

We should note that in constructing each point F , we are not obliged to start directly from the given points A , B .

---

[1] It is clear that without this restriction, nothing prevents us from replacing the figure F with the figure symmetric to it, with respect to the line joining the two points.

On the contrary, we can obtain the point corresponding to any point using the triangle formed, not with the original points, but with two other points already constructed.[2]

In this way we can construct a figure ABCDE (Fig. 96) by constructing successively three triangles congruent to ABC, BCD, CDE.

96. The operation called translation transforms every figure into a congruent figure with the same orientation. This motion can be thought of as a continuous sliding of the figure in its plane. It suffices to make the translation vary continuously from the value zero (for which the transformed figure coincides with the original one) to the final value of the translation.

In this parallel motion, the points move along parallel lines, in the same direction.

97. A rotation is the name given to a motion in which each point of the figure turns around a fixed point, called the center of the rotation, in a determined sense, through an angle which defines the magnitude of the rotation. In other words, a point A of the original figure F being given, we find the corresponding point A' of F' by joining A to the center O of rotation, forming an angle AOA' equal to the angle of rotation (and with the same orientation) then taking, on the second side of this angle, a length OA' equal to OA (Fig. 97).

We see that a rotation is determined by its center, its magnitude, and its sense.

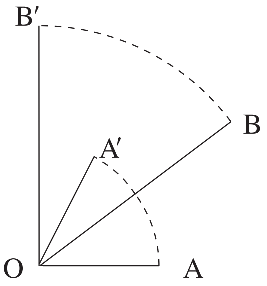

Figure 97

The figure F' obtained from F by an arbitrary rotation is congruent to F. To see this, take any two points A, B of F (Fig. 97) and consider the homologues A', B'. Angle $A'OB'$ is equal to AOB (and has the same orientation). Suppose, to be definite, that AOB has the same orientation as the rotation.[3] Replacing OB by OB', we increase this angle by a quantity equal to the angle of rotation, but then replacing OA by OA' diminishes the angle by the same quantity.

We conclude that the triangles AOB, $A'OB'$ are congruent, with same orientation, because they have an equal angle contained between equal sides. ($OA = OA'$, $OB = OB'$). Since this holds for all points B, the congruence of the figures is established.

---

[2] We will see how we can proceed in just the same manner in drawing maps (see Solid Geometry, Book X, 565).

[3] Recall (20) that angle AOB is taken to be drawn in the sense going from AO towards OB. We can make the assumption of the text by taking, as side AO, the first side we encounter, as we rotate in the given sense.

A rotation can also be obtained as a sliding of the figure in the plane, making the angle of rotation vary continously from zero to its final value. In this motion, each point moves along a circular arc with its center at the center of rotation. These arcs correspond to equal central angles.

98. Theorem. In two equal figures, derived from each other by a rotation around a point O (Fig. 98):

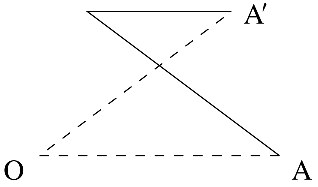

Figure 98

1°. Two corresponding lines, taken in the corresponding sense, form an angle equal to the angle of rotation, and in the same sense;

2°. The center of rotation, two arbitrary corresponding points A, A', and the intersection I of two corresponding lines passing through A and A', respectively, belong to the same circle.

1°. The first part of the theorem is true, by definition, for two corresponding lines passing through the center of rotation. It is therefore true for arbitrary corresponding lines, because each one can be replaced by a parallel line passing through the center of rotation.

2°. The four points A, A', O, I are on the same circle; namely, the locus of the vertices of angles, equal to the angle of rotation, whose sides pass through points A, A'.

99. Rotations through an angle equal to two right angles, or 180°, are of particular interest. The point A' corresponding to a point A can be constructed by joining OA, and extending this segment by a length $OA' = OA$. The point A' thus obtained is called the point symmetric to A with respect to the point O. Thus, in plane geometry, symmetry with respect to a point, and rotation about the same point by 180°, are two identical operations.

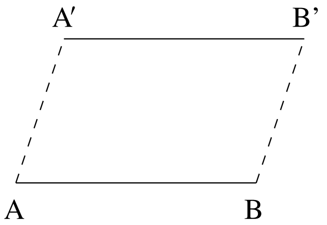

Figure 99

100. If, in two equal figures with the same orientation, two corresponding segments AB, $A'B'$ (Fig. 99) are parallel and have the same sense, the two figures

can be superimposed by a translation. Indeed, the quadrilateral $ABA'B'$ is a parallelogram, and therefore AA' is also parallel to BB', equal to it, and in the same sense: the translation AA' moves the points A, B to their respective homologues, which implies that this translation will make the two figures coincide completely.

If, in two equal figures with the same orientation, a point O coincides with its homologue, the two figures can be superimposed by a rotation around this point.

Indeed, if A, A' are two other corresponding points, the rotation with angle AOA' around O brings the point A to A', while the point O remains its own homologue.

101. We have seen (20) that two congruent figures with different orientations cannot be superimposed without one of them leaving the plane. On the other hand, two congruent figures with the same orientation can always be superimposed by a motion in the plane; namely, by a translation followed by a rotation. Indeed, if A, A' are two corresponding points, the translation AA' applied to the first figure will make these two points coincide, after which a rotation about the point A' will suffice to make the figures coincide (by the preceding section).

Observing that a translation does not change the directions of lines, we deduce that two arbitrary corresponding lines of two congruent figures with same orientation form a constant angle: thus we can talk of the angles between the two figures.

102. However, of the two preceding operation (translation and rotation), we can show that one only is necessary. Indeed, we have the following theorem:

Theorem. Two congruent figures with the same orientation can be superimposed either by a translation, or by a rotation about a properly chosen point.

This point is called the center of rotation.

Proof. Consider two points A, B of the first figure, and their homologues A', B' in the second figure. We will have superimposed the two figures if we make the segements AB and $A'B'$ coincide.

1°. If the segments AB, $A'B'$ are parallel in the same sense (Fig. 99), the figures can be superimposed by a translation;

2°. Assume that this is not the case. Let us see if these figures can be superimposed by a rotation. We know the angle of this rotation, which is equal to the angle between these figures, for instance the angle between AB and $A'B'$ (Fig. 100). We must therefore find a point O such that the rotation around O with this angle takes A to A'. This amounts to finding the center of a circular arc ending at A, A' and subtended by an angle equal to half the angle of rotation, and in the same sense (since the angle at the center equals twice the inscribed angle intercepting the same arc).

The point O will therefore be on the perpendicular bisector of the segment AA'; on the other hand, it will also be on the perpendicular to the tangent to the circle, constructed as indicated in 90 (Construction 14).

Conversely, point O being thus determined, the rotation with this center and with an angle equal to the angle of the two figures, will take A onto A'; it will also take the segment AB to a new position, parallel to $A'B'$ and with same direction. Since the two segments have the endpoint A' in common, this new position must be $A'B'$ itself. Thus the figures are superimposed. QED

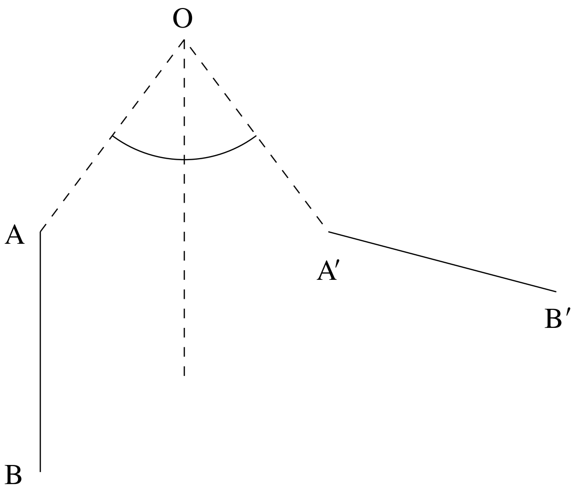

Figure 100

Remark. We can also prove this theorem by finding point O as the intersection of the perpendicular bisectors of AA' and BB' or (98) as the intersection of the circumscribed circles of triangles IAA', IBB', where point I is the intersection of AB and $A'B'$. With either of these constructions, one would establish that the triangles OAB and $OA'B'$ are congruent.

102b. The preceding theorem can also be proved by decomposing rotations and translations into two successive symmetries.

Lemma. Successive reflection in two lines D_1, D_2 is equivalent to:

1°. a translation perpendicular to the common direction of the lines, and twice the translation which would bring the first line onto the second, when the lines are parallel;

2°. a rotation around their intersection point and double the one which would bring the first line onto the second, when the lines intersect.

Proof. Let F be a figure, let F' be its reflection in D_1, and let F'' be the reflection of F' in D_2.

1°. Suppose that D_1 and D_2 are parallel (Fig. 101). If M is a point on F, we can find its reflection M' in D_1, by extending the perpendicular MH_1 from M to D_1 by its own length. We can then find the reflection M'' of M' in D_2 by extending the perpendicular $M'H_2$ from M' to D_2 by its own length. We see immediately that the three points M, M', M'' are on the same common perpendicular to D_1 and D_2. To be definite, we assume that M' is between D_1 and D_2. In this case the segment MM'', equal to the sum of MM' and $M'M''$ (Fig. 101), is twice the segment H_1H_2 which is the sum of $M'H_1$ and $M'H_2$.

This conclusion remains true if the point M' is exterior to the two parallel lines: in that case segment MM'' is the difference of MM' and $M'M''$, and H_1H_2 is the difference of $M'H_1$ and $M'H_2$. Thus the translation represented by a segment perpendicular to the two lines and double in length to H_1H_2 in fact does take any point of F onto the corresponding point of F''.

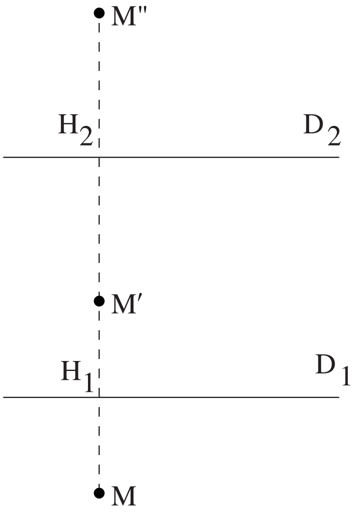

Figure 101

2°. Assume that the two lines of symmetry meet at O (Fig. 102), and let M, M', M'' be corresponding points of F, F', F'', so that M, M' are symmetric with respect to D_1 and M', M'' are symmetric with respect to D_2. Points M, M', M'' are clearly on the same circle with center O. Moreover, angle MOM'' is equal to the sum or difference of MOM' and $M'OM''$, and therefore is twice the angle between D_1 and D_2, and can be thought of as the sum or the difference of the angles made by OM' with D_1 and D_2. Therefore the rotation defined in the statement of the lemma moves each point M onto its corresponding point M''. QED

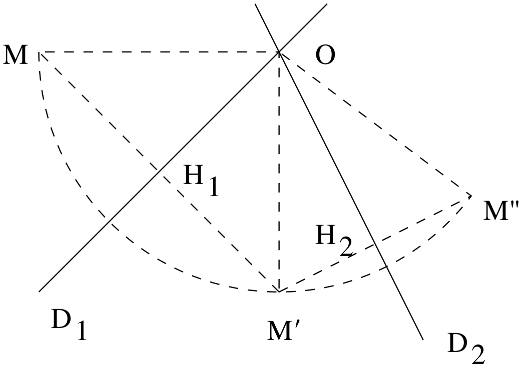

Figure 102

Remark. Two successive reflections in the same line cancel each other, since each point returns to its original position.

Corollary. Conversely, every translation can be replaced by two successive line reflections, in lines perpendicular to the direction of the translation, one of which can be chosen arbitrarily;

Every rotation can be replaced by two successive reflections in lines passing through the center of the rotation, and one of these can be chosen arbitrarily.

Indeed, in the case of a translation, it is enough to take the line that is not arbitrarily chosen to be at a distance from the arbitrary line equal to half that of the translation; and, in the case of a rotation, to take it so as to make, with the arbitrarily chosen line, an angle half that of the rotation.

103. Theorem. Any number of successive translations or rotations can be replaced with a single translation or a single rotation.

This process called composition of displacements: the final motion is called the result of the given displacements.

It suffices to know how to compose two displacements; if we have three, we begin by composing the first two, and then compose the result with the third; one would continue in an analogous fashion if there were more than three displacements.

Two successive translations replace a line with a parallel one, in the same sense, and therefore the result is a translation (100).

We need therefore consider only a translation followed or preceded by a rotation, or two rotations.

In the first case, we assume, to be definite, that the translation is performed first. This translation can be decomposed into two reflections in the lines D_1, D_2, and the rotation can also be decomposed into reflections in the lines D_1, D_2, so that we must perform four reflections altogether, in lines D_1, D_2, D_1, D_2. Now the line D_2 can be chosen arbitrarily from among those perpendicular to the direction of the translation, and the line D_1 can be chosen arbitrarily from among those passing through the center of the rotation. So we can choose both to coincide with the line through the center of rotation and perpendicular to the direction of translation.

With this choice, the reflections in these two lines will cancel, and we are left with two reflections, in lines D_1, D_2 which will yield (by the preceding numbered paragraph) a translation or a rotation.

If we had to compose two rotations, we would have taken $D_2 = D_1$ to be the line joining the two centers of rotation. This proves the theorem in all cases.

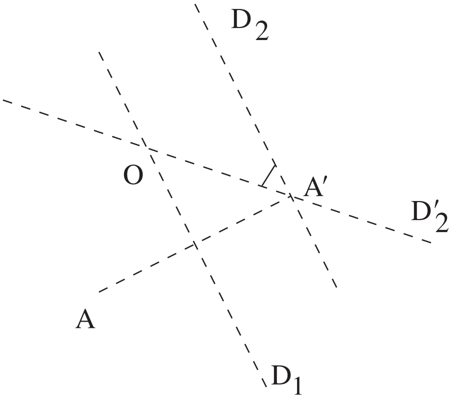

Figure 103

The theorem in 102 can now be regarded as a corollary of the one we just proved. An arbitrary motion of a figure in the plane is a translation followed by a rotation, namely (if A, A' are two corresponding points of the original figure and of the displaced figure) the translation AA' followed by a rotation with center A' through an angle equal to the angle of the two figures (Fig. 103). Lines D_2, D_1 will coincide with the perpendicular from A' to AA'; line D_1 will then be the perpendicular bisector of AA', and D_2 will be obtained by rotating D_2 about A' through half the angle of rotation. The intersection point of D_1, D_2 gives the center of rotation. We see that the construction is entirely in agreement with the one given earlier, because in constructing (Construction 14) the tangent to the circular arc on AA' which subtends half the angle of rotation, we are actually constructing a perpendicular to D_2.

104. **Theorem.** When a figure moves in the plane, without change shape, at each instant the normals (60) to the trajectories of the various points (if they are not all parallel) will be concurrent at a point which we call the instantaneous center of rotation.

Leaving aside the case where all normals to the moving points $M, N, P, \ldots$ are parallel, we assume that two of these normals, the normals at M, N for instance, intersect in a point O (Fig. 104).

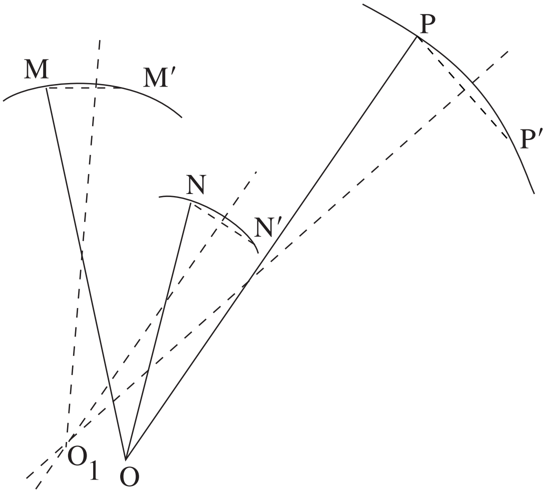

Figure 104

Consider a second position of the moving figure, close to the first, where the new positions of M, N, $P, \ldots$ are M', N', P' … The perpendicular bisectors of MM', NN', PP', … pass through the same point O_1, the center of the rotation which brings the first position of the figure onto the second. If we now imagine the second position coming arbitrarily close to the first, the midpoint of MM' will approach M, the line MM' will approach the tangent at M to the trajectory of this point, and the perpendicular bisector of MM' will approach the normal to this trajectory. In the same way, the perpendicular bisector to NN' approaches the normal to the trajectory of N. The point O_1 will then approach the point O, so that the normals to the trajectories of P ..., which are the limiting positions of the perpendicular bisectors of PP ,..., will also pass through O.[4]

Corollary. If we know the tangents to the trajectories of two points of a moving figure which does not change shape, we can find the tangent to the trajectory of any other point of the figure.

Indeed, the perpendicular to the known tangents determine the instantaneous center of rotation and, therefore, allow us to draw the normal to the trajectory at any point.

Exercises

Exercise 92. If a figure moves, without changing shape, in such a way that two of its lines each pass through a fixed point, then there exist infinitely many lines in the figure which, during the motion, also pass through a fixed point. Every other line remains tangent to a fixed circle. The instantaneous center of rotation for any two positions of the figure remains on a certain fixed circle.

Exercise 93. Given three figures F_1, F_2, F_3, all congruent and with the same orientation, we determine the three centers of the rotations which would superimpose one of these figures on another. Show that the angles of the triangle formed by these centers of rotation are equal or supplementary to half the angles between pairs of these figures.

Exercise 94. Compose two rotations with the same magnitude but opposite sense, and with different centers of rotation.

Exercise 95. Two congruent figures with opposite orientation can be superimposed:

1 . In infinitely many ways, by three successive line reflections;

2 . In infinitely many ways, by a translation followed by a line reflection;

3 . In only one way, by a translation preceded or followed by a reflection in a line parallel to the translation.

Exercise 96. Construct an equilateral triangle with vertices on three given parallel lines, or on three given concentric circles.

Exercise 97. On a given circle, find an arc, equal to a given arc, such that the lines joining the endpoints to two given points are parallel.

---

[4] We assume here, as axioms, propositions such as these: If a variable point M_1 has a limit position M, and a variable direction D_1 has a limit position D, the parallel through M_1 to D_1 has as limiting position the parallel through M to D. If two lines D_1, D_1 have limiting positions D, D , then the intersection point of D_1, D_1 has limiting position the intersection of D, D . In reality, these propositions are theorems which should be proved; we omit the proofs because they could not be performed with the desired simplicity, without assuming a knowledge of limits relying on notions of differential calculus.

Problems for Book II

Exercise 98. We join a point M on a circle with the points of contact A, B of the tangents passing through a point P in the plane. We then draw a parallel through P to the tangent at M. Show that lines MA, MB intercept on this parallel a segment whose length is independent of M, and which is divided by P into two equal parts.

Exercise 99. Let ABC be an equilateral triangle inscribed in a circle, and let M be a point on arc $\overset{\frown}{BC}$. Show that $MA = MB + MC$.

Exercise 100. Through the three vertices of an equilateral triangle, draw lines such that the bisectors of the angles they form with the altitudes from the same vertex are parallel. Show that these three lines intersect on the circumscribed circle.

Exercise 101. The midpoints of the sides of a triangle, the feet of the altitudes, and the midpoints of the segments, on each altitude, between the vertex and their intersection, are nine points on the same circle (the nine point circle). The center of this circle is the midpoint of the segment which joins the intersection of the altitudes to the center of the circumscribed circle; its radius is half that of the circumscribed circle. Deduce from this the result of Problem 70.

Exercise 101b. In any triangle, the circle passing through the centers of the three escribed circles has its center at the intersection of the radii of these three circles to the points of contact with the corresponding sides; also, its center is symmetric to the center of the inscribed circle with respect to the center of the circumscribed circle. Its radius is twice that of the circumscribed circle.

Exercise 102. In any triangle, the six feet of the perpendiculars dropped from the foot of each altitude onto the other sides lie on the same circle.

Exercise 103. In triangle ABC, the perpendicular bisector of BC and the angle bisector of A intersect on the circumscribed circle. Their point of intersection is equidistant from B, C, from the center of the inscribed circle, and from the center of the escribed circle contained in the angle A.

Exercise 103b. Construct a triangle knowing the lengths of the altitude, the median, and the bisector from the same vertex.

Exercise 104. Two tangents to a circle and the chord joining their points of contact divide into equal parts the perpendicular dropped from an arbitrary point on this chord to the line which joins it with the center.

Exercise 105. On the three sides of the triangle ABC, and outside the triangle, we construct equilateral triangles BCA , CAB , ABC , and join AA , BB , and CC . Show that:

1 . These three line segments are equal;

2 . They pass through the same point, at which the three sides of the original triangle subtend equal angles;

3 . If this point is inside the triangle, the sum of its distances to the three vertices is equal to the common length of AA , BB , CC (Problem 99).

Exercise 106. Lines ABE, CDE,BCF, ADF intersect in pairs, and form four triangles. Show that the circles circumscribing these triangles all pass through the same point I.

The centers of these four circles lie on another circle G, which also passes through point I (use Problem 69).[5]

A necessary and sufficient condition for quadrilateral $ABCD$ to be cyclic is that line EF, which connects the intersections of opposite sides of the quadrilateral, contain point I. This line is then the bisector of the angles AIC, BID subtended by the diagonals of the quadrilateral at point I, or of their supplements (the angle formed by IA and the extension of IC, or by IB and the extension of ID).[6]

The circumcenter O of quadrilateral $ABCD$ lies on circle G. This circumcenter also lies on the circles AIC and BID. Conclude from this that I is the foot of the perpendicular from O to EF.

Exercise 107. The bisectors of the angles formed by the (extensions of) the opposite sides of a cyclic quadrilateral are perpendicular. They are parallel to the bisectors of the angles formed by the diagonals of the quadrilateral, and conversely.

Exercise 107b. If any two lines intersect two given circles, the chords joining any pairs of points of intersection of the lines with the first circle intersect the chords joining any pairs of points of the intersection of these lines with the second circle in four concyclic points.

Exercise 108. Given triangle ABC, find the locus of points M such that the three perpendiculars erected from A, B, C to AM, BM, CM respectively are concurrent.

Exercise 109. Find the locus of the vertices of the rectangles considered in Exercise 41, if the given parallelogram is articulated (46b) while one of its sides is held fixed.

Exercise 110. On two consecutive segments AB, BC of the same line, we draw two variable circles with the same radius. These circles intersect at B, and at one additional point, whose locus is required.

Exercise 111. Let OA, OB be two variable perpendicular radii of a circle O. Through A, B construct two lines, each parallel to one of two fixed perpendicular directions. Find the locus of the intersection of these lines, as the right angle AOB turns around the center.

Exercise 112. Two circles, each of which touches a fixed line at a fixed point, vary while remaining tangent to each other. Find the locus of their point of contact.

Exercise 113. Find the locus of the vertex of the right angle of a right triangle which does not change shape, as the two other vertices slide along two perpendicular lines.

Exercise 114. Draw a line along which two given circles intercept given lengths.

---

[5] The reference to Problem 69 is probably an error in the original text. We give a proof using Problem 72. transl.

[6] The second case is the one that always holds. [This footnote, in the original, is the author s correction of an error in an earlier edition. transl.]

Exercise 115. In a given circle, inscribe a triangle whose sides are parallel to three given lines; or two of whose sides are parallel to given lines, and whose third side passes through a given point.

Exercise 116. Given a line xy and two points A, B, find a point M on the line such that the angle AMx is twice BMy.

Exercise 117. Construct a pentagon (or, more generally, a polygon with an odd number of sides) knowing the midpoints of the sides. What happens when the polygon has an even number of sides?

Exercise 118. Construct a trapezoid knowing its sides. More generally, contruct a quadrilateral knowing its four sides and the angle between (the extensions of) two opposite sides.[7]

Exercise 119. Cut the sides AB, AC of a triangle with a transversal having a given direction, and intercepting equal segments on these sides, starting at B and C.

Exercise 120. Construct a segment perpendicular to two given parallel lines, which subtends a given angle at a given point.

Exercise 121. We are given a circle, points P, Q on the circle, and a line. Find a point M on the circle such that the lines MP, MQ intercept on the line a segment IK of given length.

Exercise 121b. Same problem as above, except that we are given not the length, but the midpoint, of IK.

Exercise 122. Construct a square whose sides pass through four given points. Can the problem have infinitely many solutions? In that case, what will be the locus of the centers of the squares described in the problem?

Exercise 123. In an articulated quadrilateral $ABCD$ (46b, Remark III), what conditions must the lengths of the sides satisfy if a vertex, say B, is to pivot through a complete revolution; in other words, such that the size of the angle at B can be arbitrary? Show that the vertex D is not, in general, the center of a complete revolution. What is the exceptional case?

---

[7] With relation to this problem and the following ones, see 284, Note A.

CHAPTER II

Similarity of Triangles

117. **Definition.** Two triangles are similar if they have equal angles and proportional corresponding sides.

Remarks. Two congruent triangles are similar. Two triangles similar to a third are also similar to each other. In particular, if two triangles are similar, any triangle congruent to the first is similar to the second.

Theorem. Every parallel to one side of a triangle forms a triangle with the other two sides which is similar to the first.

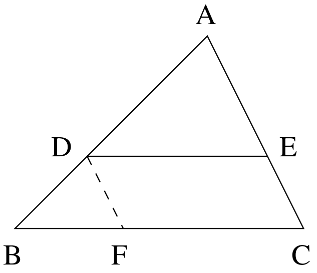

Figure 115

Consider parallel DE to side BC of triangle ABC (Fig. 115). We claim that the new triangle ADE is similar to ABC.

First, the angles of the two triangles are equal, A being common to both, and angles D, E being corresponding angles with B, C, respectively.

Second, we have (114) $\frac{AD}{AB} = \frac{AE}{AC}$, and it remains only to prove that the common value of these ratios is also equal to $\frac{DE}{BC}$.

To show this, we draw a parallel DF to AC, forming parallelogram DEFC. Line DF cuts sides BA, BC into proportional parts, so that $\frac{DE}{BC} = \frac{FC}{BC} = \frac{AD}{AB}$.

QED

Remark. Line DE can be inside (Fig. 115) or outside (Figs. 116, 117) triangle ABC without affecting the validity of the proof. However, in the case represented in figure 117, the corresponding angles of the two triangles are equal because they are alternate interior angles, rather than corresponding angles.

118. The following theorems (cases of similarity for triangles) give necessary and sufficient conditions for two triangles to be similar. They correspond to the three cases of congruence, to which they are reduced by use of the preceding theorem.

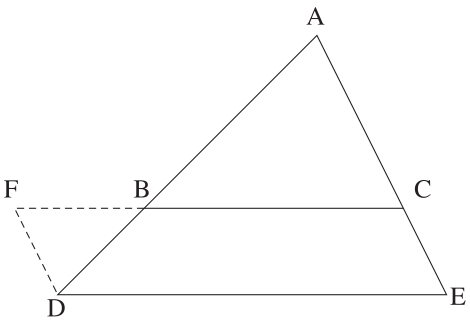

Figure 116

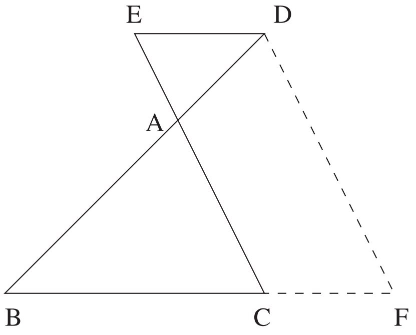

Figure 117

First case of similarity. Two triangles are similar if they have two equal corresponding angles.

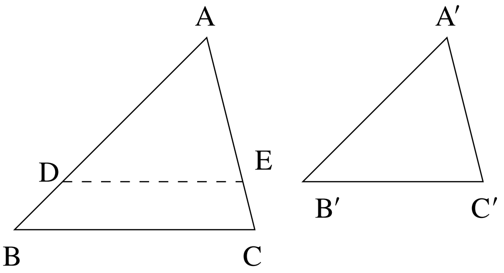

Figure 118

Consider triangles ABC, A B C in which we have A = A , B = B (Fig. 118). On AB we construct a length AD = A B and through point D we draw a parallel DE to BC. Triangle ADE is similar to ABC (by the preceding theorem) and congruent to A B C , having an equal side between two equal angles.

Second case. Two triangles are similar if they have an equal angle contained between proportional sides.

Consider triangles ABC, A B C in which we have A = A and $\frac{AB}{AC} = \frac{A\ B}{A\ C}$. On side AB we take a length AD = A B and we draw parallel DE to BC. Triangle ADE is similar to ABC and congruent to A B C . Indeed, they have an equal angle (A = A) contained between equal sides: AD = A B by construction, and the two proportions $\frac{AB}{AC} = \frac{A\ B}{A\ C}$ (given) and $\frac{AB}{AC} = \frac{AD}{AE}$, which have three common terms, give us A C = AE.

Third case. Two triangles are similar if they have three proportional sides.

Suppose the two triangles are ABC, A B C in which $\frac{AB}{A\ B} = \frac{AC}{A\ C} = \frac{BC}{B\ C}$. On side AB we take a length AD = A B and draw parallel DE to BC. Triangle ADE is similar to ABC and congruent to A B C , having three pairs of corresponding sides equal. Indeed, AD = A B by construction, and the proportions $\frac{AB}{AD} = \frac{AC}{AE}$ and $\frac{AB}{AD} = \frac{BC}{DE}$ (by the previous result) have three terms in common with the proportions $\frac{AB}{A'B'} = \frac{AC}{A'C'}$ and $\frac{AB}{A'B'} = \frac{BC}{B'C'}$ (given by hypothesis), respectively. We conclude that in fact $AE = A'C'$, $DE = B'C'$.

Remark. The definition of similar triangles consists of five conditions: three given by the equality of the angles, two by the proportionality of the sides. Corollary III of 44 (Book I) shows that one of the first three conditions can be omitted, which reduces the total to four. The cases of similarity show that, to establish similarity, we only need to prove two of these conditions, if chosen properly.

119. Theorem. Two triangles whose corresponding sides are parallel or perpendicular are similar.

Indeed, in this situation, the corresponding angles are either equal in pairs or supplementary (Book I, 43), so that if the triangles are ABC and $A'B'C'$, we have

$$A = A' \quad \text{or} \quad A + A' = 2 \text{ right angles},$$

$$B = B' \quad \text{or} \quad B + B' = 2 \text{ right angles},$$

$$C = C' \quad \text{or} \quad C + C' = 2 \text{ right angles}.$$

The three equations in the second column cannot all be satisfied simultaneously, because they would imply that the total sum of the angles in the two triangles is equal to six right angles, rather than four. Nor can even two of these equalities hold simultaneously; for instance, $A + A' = 2$ right angles and $B + B' = 2$ right angles would give a total sum for the angles of the two triangles equal to 4 right angles $+C + C'$. Therefore at least two equalities in the first column must hold, and then the triangles have angles which are equal in pairs (first case of similarity).

120. Theorem. Two right triangles are similar if the ratio of one of the sides of the right angle to the hypotenuse is the same in the two triangles.

Indeed, we can proceed as in 118 by forming a third triangle similar to the first, and having the same hypotenuse as the second. The second and third triangles will then be congruent, by the second case of congruence for right triangles.

Remark. Aside from this case of similarity, we can apply to right triangles the cases for similarity for arbitrary triangles. For instance, two right triangles will be similar if they have an equal acute angle (first case of similarity for arbitrary triangles); or if the sides of the right angle are proportional (second case of similarity).

Because of this, all right triangles ABC which have the same acute angle C have proportional sides. If we are given the angle C, we can then determine:

1°. The ratio of the opposite side AB to the hypotenuse BC (this ratio is called the sine of C and is denoted $\sin C$);

2°. The ratio of AC to the hypotenuse (called the cosine of C and denoted $\cos C$;

3°. The ratio of AB to AC (called the tangent of C and denoted $\tan C$).

These facts form the basis of trigonometry.

121. Theorem. A family of concurrent lines cuts off proportional segments on two parallel lines.

Consider, for instance, the three concurrent lines SAA', SBB', SCC' (Fig. 119) which meet two parallels ABC, $A'B'C'$. Similar triangles (117) SAB, $SA'B'$

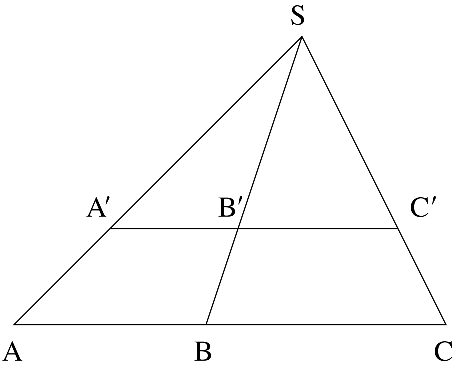

Figure 119

yield

$$\frac{A'B'}{AB} = \frac{SB'}{SB},$$

and similar triangles SBC, SB′C′ give

$$\frac{B'C'}{BC} = \frac{SB'}{SB},$$

from which it follows that A′B′, B′C′ are proportional to AB, BC.

Remark. The corresponding segments all have the same sense, or each pair has the opposite sense, according as point S is exterior or interior to the two parallels.

Converse. If three lines intercept proportional segments on two parallel lines (all in the same sense, or each pair in the opposite sense), these three lines are either concurrent or parallel.

If the corresponding segments are equal and in the same sense, the three lines are parallel (46, converse). Assume then that lines AA′, BB′, CC′ intercept proportional segments on parallel lines ABC, A′B′C′, so that

$$\frac{A'B'}{AB} = \frac{B'C'}{BC} = \frac{A'C'}{AC};$$

the value of this ratio being different from 1 if the segments are in the same sense. This last condition makes it impossible (46) for lines AA′, BB′ to be parallel; they therefore intersect at some point S.

Joining SC, we see that this line cuts the parallel A′B′C′ in a point C_1 which must coincide with the point C′ because $\frac{A'C_1}{B'C_1} = \frac{AC}{BC}$ (by the preceding theorem), which equals $\frac{A'C'}{B'C'}$ (by hypothesis).

Exercises

Exercise 129. Through the intersection of the diagonals of a trapezoid we draw a parallel to the bases. Show that this line is divided into equal segments by the non-parallel sides.

Exercise 130. Let a = AB and b = CD be the two bases of a trapezoid. We divide one of the non-parallel sides into the ratio $\frac{EA}{EC} = \frac{m}{n}$, and through point E we draw a parallel to the bases. Show that the segment of this parallel contained inside the trapezoid is equal to $\frac{m \cdot CD + n \cdot AB}{m+n}$. Special case: E is the midpoint of AC.

Exercise 131. From the vertices of a triangle, and from the intersection of its medians, we drop perpendicular segments onto a line exterior to the triangle. Show that the last of these segments is equal to the arithmetic mean of the first three. (Preceding exercise.)

Exercise 132. Through vertex A of parallelogram ABCD we draw a line which intersects diagonal BD in E, and sides BC, CD in F, G. Show that AE is the geometric mean of EF and EG.

Exercise 133. Find the locus of the points which divide the segment intercepted by the sides of an angle, on the lines parallel to a given direction, in a given ratio.

Exercise 134. Find the locus of points at which two given circles subtend the same angle.[1]

---

[1] The angle subtended by a curve at a given point is the angle formed by the tangents to the curve through that point.

CHAPTER III

Metric Relations in a Triangle

122. **Definition.** The foot of the perpendicular dropped from a point onto a line is called the orthogonal projection (or, briefly, projection) of the point onto that line. The segment whose endpoints are the projections of the endpoints of a given segment is called the projection of the segment onto the line.

123. Let ABC (Fig. 120) be a triangle with a right angle at A. From this vertex A, drop altitude AD onto the hypotenuse. We will discover certain relations among the elements of this figure.

Theorem. In a right triangle, each side of the right angle is the geometric mean between the whole hypotenuse and its projection onto the hypotenuse.

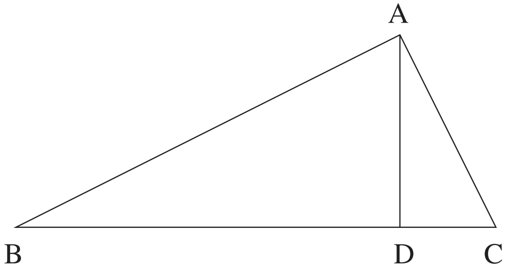

Figure 120

For instance, AB is the geometric mean between BD and BC. Indeed, right triangles ABD, ABC, having angle B in common, are similar (120, Remark). The sides corresponding to AB, BD in the first are sides BC, AB in the second, so that

$$\frac{BD}{AB} = \frac{AB}{BC}, \quad \text{or} \quad AB^2 = BD \cdot BC.$$

QED

Corollary. Any chord in a circle is the geometric mean between a diameter and its projection on the diameter passing through one of its endpoints (Fig. 121).

This is true because the diameter and the chord are the hypotenuse and a side of a right triangle (73, Corollary II).

Remark. Similar triangles ABD, ABC (Figures 120, 121) also give

$$\frac{AB}{BC} = \frac{AD}{AC} \quad \text{or} \quad AB \cdot AC = BC \cdot AD.$$

Thus, the product of the sides of the right angle in a right triangle, is equal to the product of the hypotenuse and the corresponding altitude.

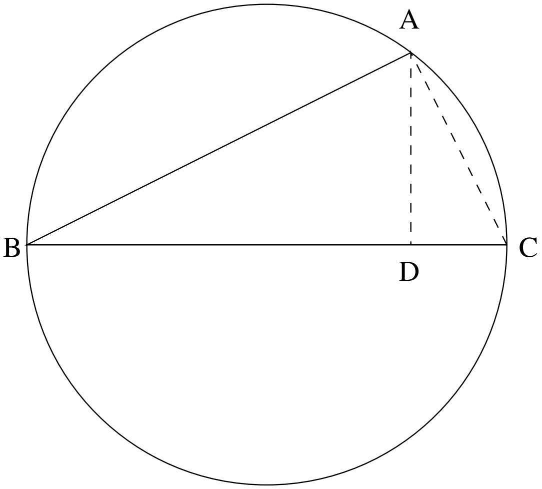

Figure 121

124. Theorem of Pythagoras. In a right triangle, the square of the hypotenuse is equal to the sum of the squares of the other two sides.

We have just seen that

$$AB^2 = BC \cdot BD.$$

The same theorem applied to the side AC gives

$$AC^2 = BC \cdot CD.$$

Adding these equalities we obtain

$$AB^2 + AC^2 = BC(BD + CD) = BC^2.$$

QED

This theorem can be used to calculate any side of a right triangle when we know the other two sides. Thus, let us calculate the hypotenuse of a right triangle in which the legs are 3 and 4 meters. If the unit is the meter, the square of the number representing the measure of the hypotenuse will be the sum of the squares of the numbers 3 and 4. This sum is $3^2 + 4^2 = 25$, whose square root is 5; thus the hypotenuse will equal 5 meters.

Let us now calculate one leg of a right triangle, knowing that the hypotenuse is equal to 10 meters and the other leg is 7 meters. The square of the required side, added to $49 (= 7^2)$ must give $100 (= 10^2)$; this square is therefore equal to $100 - 49 = 51$. The square root of 51, (which is 7.14, correct to the nearest centimeter) will give the measure, in meters, of the desired side.

125. Theorem. In a right triangle, the altitude from the right angle is the geometric mean between the two segments which it determines on the hypotenuse.

Indeed, triangles ABD, ACD (Fig. 120) are similar because they have perpendicular sides. Segments BD, AD are therefore proportional to AD, DC.

Corollary. The perpendicular from any point on a circle to a diameter is the geometric mean between the two segments it determines on the diameter.

126. *Theorem.* The difference of the squares of two sides of a triangle is equal to the difference of the squares of their projections on the third side.

In triangle ABC (Fig. 123), we project vertex B to point H on AB. The equations

$$AB^2 = AH^2 + BH^2,$$

$$BC^2 = CH^2 + AH^2$$

yield, by subtraction, $BC^2 - AB^2 = CH^2 - AH^2$.

Theorem. In any triangle:

1. The square of the side opposite an acute angle is equal to the sum of the squares of the other two sides, minus twice the product of one of these sides and the projection of the other onto it;

2. The square of the side opposite an obtuse angle is equal to the sum of the squares of the other two sides, plus twice the product of one of these sides and the projection of the other onto it.

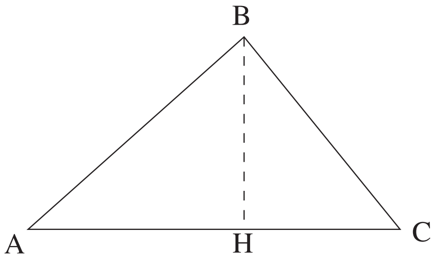

Figure 122

1. In triangle ABC, consider side BC opposite acute angle A. Drop altitude BH from B to AC. By the preceding theorem we have

$$BC^2 = AB^2 + CH^2 - AH^2.$$

Now CH is the difference between AC and AH, so that CH^2 can be replaced[1] by

$$AC^2 - 2AC \cdot AH + AH^2,$$

yielding the desired result

$$BC^2 = AB^2 + AC^2 - 2AC \cdot AH.$$

2. In triangle ABC, consider side BC opposite obtuse angle A. Drop altitude BH from B to AC. Again we have

$$BC^2 = AB^2 + CH^2 - AH^2,$$

---

[1] We assume that the formulas for the squares of the sum and the difference of two numbers are known:

$$(a + b)^2 = a^2 + 2ab + b^2,$$

$$(a - b)^2 = a^2 - 2ab + b^2.$$

See Tannery, Leçons d'Arithmetique, chap. II, n^{os} 83, 85.

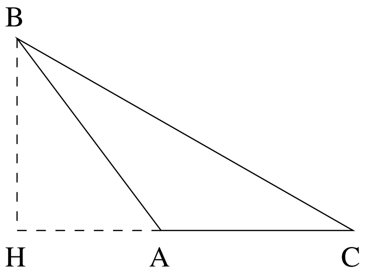

Figure 123

but CH is the sum of AC and AH, so that CH^2 can be replaced by

$$AC^2 + 2AC \cdot AH + AH^2,$$

yielding

$$BC^2 = AB^2 + AC^2 + 2AC \cdot AH.$$

Corollary. An angle of a triangle is acute, right, or obtuse, according as the square of the opposite side is less than, equal to, or greater than the sum of the squares of the two other sides.

127. Stewart's Theorem. Given triangle ABC and point D on the base between B and C, we have

$$AB^2 \cdot DC + AC^2 \cdot BD - AD^2 \cdot BC = BC \cdot DC \cdot BD.$$

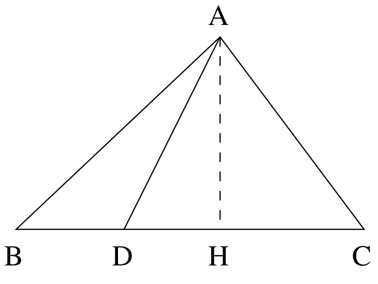

Figure 124

We drop a perpendicular AH from A to BC (Fig. 124) and suppose, to be definite, that H and C are on the same side of D. The two theorems of the preceding section apply, one to triangle ACD and the other to triangle ABD, to yield

$$AC^2 = AD^2 + DC^2 - 2CD \cdot DH,$$

$$AB^2 = AD^2 + BD^2 + 2BD \cdot DH.$$

We multiply the first equation by BD, and the second by CD, then add the resulting equations. The term $2BD \cdot CD \cdot DH$, which appears once with a + sign and once with a − sign, disappears and we have

$$AC^2 \cdot BD + AB^2 \cdot CD = AD^2(BD + CD) + DC^2 \cdot BD + BD^2 \cdot CD$$
$$= AD^2 \cdot BC + BC \cdot DC \cdot BD.$$

128. Application to the calculation of various notable lines in a triangle. Consider an arbitrary triangle ABC whose sides BC, CA, AB have known measures a, b, c, respectively. Let us calculate the lengths of the medians, bisectors, and altitudes of this triangle.

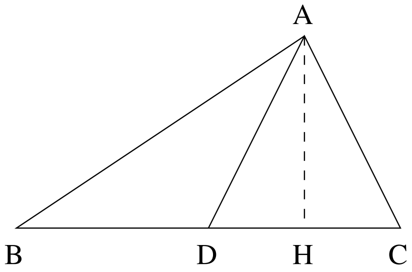

Figure 125

1 . Medians. Let AD be the median from vertex A (Fig. 125). In the equation given by the previous theorem, we replace BC, CA, AB by a, b, c, respectively; DC and BD by a/2. Dividing by a we obtain

$$\frac{b^2 + c^2}{2} = AD^2 + \left(\frac{a}{2}\right)^2,$$

and therefore

$$AD^2 = \frac{b^2 + c^2}{2} - \left(\frac{a}{2}\right)^2.$$

Thus, the sum of the squares of two sides of a triangle equals twice the square of half the third side, plus twice the square of the corresponding median.

128b. On the other hand, if, in the equalities

$$AC^2 = AD^2 + DC^2 - 2CD \cdot DH,$$
$$AB^2 = AD^2 + BD^2 + 2BD \cdot DH$$

we replace BC, CA, AB by a, b, c and DC and BD by a/2, then substracting them term by term, we get

$$c^2 - b^2 = 2a \cdot DH.$$

Thus the difference between two sides of a triangle is equal to twice the product of the third side and the projection of the median onto this side.

Corollary. The locus of points C, such that the difference of the distances from C to two fixed points A and B is constant, is a line perpendicular to AB.

The reason is that the difference $AB^2 - AC^2$ is constant, so the projection H of C onto AB is a fixed point.

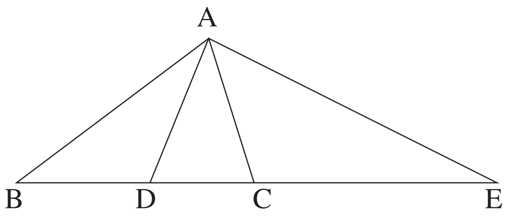

Figure 126

129. 2 . Bisectors. We now denote by AD the bisector of A (Fig. 126). Point D divides side BC into segments proportional to AB and AC, so that

$$\frac{BD}{c} = \frac{CD}{b} = \frac{BC}{b+c} = \frac{a}{b+c},$$

or

$$BD = \frac{ac}{b+c}, \quad CD = \frac{ab}{c+b}.$$

In the result of the theorem in 127, we replace BD and CD with these values, and BC, CA, AB with their values. Dividing by a, we obtain

$$\frac{c^2b}{b+c} + \frac{b^2c}{b+c} - AD^2 = \frac{ab}{b+c} \cdot \frac{ac}{b+c},$$

so that

$$AD^2 = \frac{c^2b + b^2c}{b+c} - \frac{a^2bc}{(b+c)^2} = bc\frac{(b+c)^2 - a^2}{(b+c)^2}.$$

Now we consider bisector AE of the exterior angle at A (Fig. 126) and, to be definite, we assume that AB is greater than AC, so that point E, which divides segment BC externally into parts proportional to AB and AC, is on the extension of BC past C. We then have

$$\frac{BE}{c} = \frac{CE}{b} = \frac{BC}{c-b} = \frac{a}{c-b},$$

and therefore

$$BE = \frac{ac}{c-b}, \quad CE = \frac{ab}{c-b}.$$

We apply the theorem of 127 to triangle ABE with point C on base BE. Replacing BC, CA, AB, BE, CE by their values, and dividing by a, we have

$$\frac{c^2b}{c-b} + EA^2 - \frac{b^2c}{c-b} = \frac{ac}{c-b} \cdot \frac{ab}{c-b},$$

so that

$$AE^2 = \frac{a^2bc}{(c-b)^2} - \frac{c^2b - b^2c}{c-b} = bc\frac{a^2 - (c-b)^2}{(c-b)^2}.$$

130. 3 . Altitudes. We drop altitude AH from A (Fig. 127). Of the two angles B, C, at least one is acute. We assume, to be definite, that B is acute, and apply the theorem of 126 to the side AC = b opposite this angle to obtain

$$b^2 = a^2 + c^2 - 2a \cdot BH,$$

or

$$BH = \frac{a^2 + c^2 - b^2}{2a}.$$

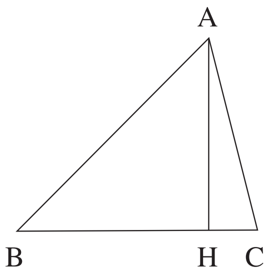

Figure 127

Now in right triangle AHB we have

$$AH^2 = c^2 - BH^2 = c^2 - \frac{(a^2 + c^2 - b^2)^2}{4a^2}.$$

This equation, whose right-hand member is a difference of two squares, can be written as[2]

$$\begin{aligned} AH^2 &= \left(c - \frac{a^2 + c^2 - b^2}{2a}\right)\left(c + \frac{a^2 + c^2 - b^2}{2a}\right) \\ &= \frac{(2ac - a^2 - c^2 + b^2)(2ac + a^2 + c^2 - b^2)}{4a^2} = \frac{[b^2 - (a - c)^2][(a + c)^2 - b^2]}{4a^2}. \end{aligned}$$

But each of the factors appearing in the numerator of the right-hand member of this equation is also a difference of two squares. Thus this last formula can be written as

$$AH^2 = \frac{(b - a + c)(b + a - c)(a + c - b)(a + b + c)}{4a^2}.$$

If we denote by p the semi-perimeter of the triangle, so that $a + b + c = 2p$, the quantities $b + c - a$, $c + a - b$, $a + b - c$ are equal to $2p - 2a$, $2p - 2b$, $2p - 2c$, respectively, so that

$$AH^2 = \frac{4p(p - a)(p - b)(p - c)}{a^2}.$$

On the other hand, if we want to use the product $[(a + c)^2 - b^2][b^2 - (a - c)^2]$, we can write the the following formula:

$$AH^2 = \frac{1}{4a^2}[4b^2c^2 - (b^2 + c^2 - a^2)^2] = \frac{1}{4a^2}(2b^2c^2 + 2c^2a^2 + 2a^2b^2 - a^4 - b^4 - c^4).$$

130b. 4°. **Radius of the circumscribed circle.** The product of two sides of a triangle is equal to the altitude drawn to the third side, multiplied by the diameter of the circumscribed circle.

In triangle ABC (Fig. 128), let AH be the altitude from A, and let AA' be the diameter of the circumscribed circle. Triangles AHC, ABA' have right angles at H and B, respectively. They have an equal acute angle ($A' = C$ since they intercept the same arc $\overset{\frown}{AB}$). They are therefore similar, so that $\frac{AH}{AC} = \frac{AB}{AA'}$, or $AB \cdot AC = AH \cdot AA'$. QED

---

[2] See Leçons d'Arithmétique, by J. Tannery, chap. II, n° 87.

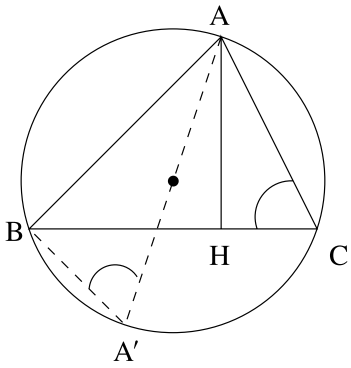

Figure 128

Using the value just found for AH, this theorem gives the following formula for the radius R of the circumscribed triangle:

$$R = \frac{AA'}{2} = \frac{bc}{2AH} = \frac{abc}{4\sqrt{p(p-a)(p-b)(p-c)}}.$$

Exercises

Exercise 135. The product of the segments intercepted by a variable tangent of a circle on two parallel tangents is constant.

Exercise 136. The inverse of the square of the altitude of a right triangle is equal to the sum of the inverses of the squares of the sides of the right angle.

Exercise 137. What is the ratio of the sum of the squares of the medians to the sum of the squares of the sides of a triangle?

Exercise 138. The sum of the squares of the distances from an arbitrary point in the plane to two opposite vertices of a parallelogram differs by a constant quantity from the sum of the squares of the distances from the same point to the other two vertices. Special case: the rectangle.

Exercise 139. The sum of the squares of the four sides of a quadrilateral is equal to the sum of the squares of its diagonals plus four times the square of the segment joining their midpoints.

Exercise 140. If ABC is a triangle, G the intersection of its medians, and M an arbitrary point in the plane, we have $MA^2 + MB^2 + MC^2 = GA^2 + GB^2 + GC^2 + 3MG^2$.

Exercise 141. Find the locus of points such that the squares of their distances to two fixed points, multiplied by two given numbers, have a given sum or difference. Use this to find a new proof of the theorem in 116.

Exercise 142. Find the locus of points such that the squares of their distances to three points in the plane, multiplied by three given numbers, has a given constant sum. Same problem with an arbitrary number of points.

Exercise 143. The square of the bisector of an angle of a triangle is equal to the product of the sides which contain it, diminished by the product of the segments it intercepts on the opposite side. State and prove an analogous statement for the bisector of the exterior angle.

Exercise 144. Deduce the inequalities indicated in Exercises 11 and 18 from the formulas found for the medians and bisectors.

Exercise 145. If the median of a triangle is the geometric mean between the sides which contain it, then the square with side $b \quad c$ has a diagonal equal to the third side of the triangle.

Exercise 146. The altitudes of a triangle are inversely proportional to the sides to which they are drawn.

Exercise 147. The product of the distances from a point on a circle to two opposite sides of an inscribed quadrilateral is equal to the product of the distances of the same point to the other sides, or to the diagonals. What happens to this statement when the two opposite sides become tangents?

CHAPTER IV

Proportional Segments in a Circle. Radical Axis

131. Theorem. If a variable secant to a circle is drawn from a point A in the plane of the circle, the product of the distances from A to the two intersection points with the circle is constant.

We distinguish two cases.

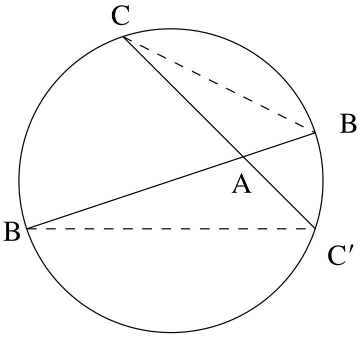

Figure 129

1 . Point A is inside the circle (Fig. 129). Consider two secants ABB', ACC', and join BC', CB'. Triangles ABC', ACB' are similar, since they have equal angles. Indeed, the angles at A are equal because they are vertical angles, and B = C because both have measure equal to one half arc $\overset{\frown}{B'C'}$. The proportionality of the sides then gives $\frac{AB}{AC} = \frac{AC'}{AB'}$, and the equality of the products of the means and the extremes gives $AB \cdot AB' = AC \cdot AC'$. QED

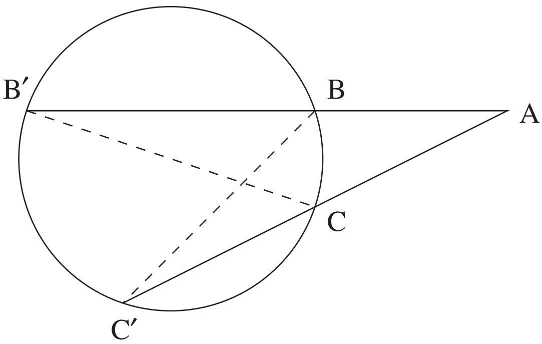

Figure 129b

2 . Point A is outside the circle (Fig. 129b). Consider two secants ABB , ACC , and join BC , CB . Triangles ABC , ACB are again similar, having equal angles: the angle at A in common, and B = C intercept the same arc. The conclusion follows as before.

131b. Converse. If, on two lines ABB , ACC through a point A (Fig. 130) we take four points A, A , B, B (the point A being either exterior to both segments BB , CC , or interior to both) such that AB • AB = AC • AC , then these four points are on the same circle.

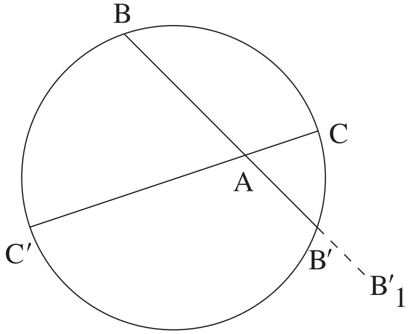

Figure 130

Indeed, the three points B, C, C determine a circle, and this circle intersects ABB in a point B_1 such that $AB \cdot AB_1$ = AC • AC . This equality, combined with the one in the hypothesis, shows that AB_1 = AB , and therefore the point B_1 coincides with B .

132. Theorem. If, through a point exterior to a circle, we draw a tangent and a secant to the circle, the tangent is geometric mean between the entire secant and its external segment.

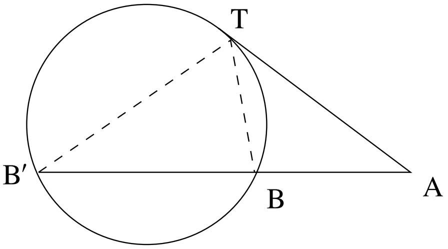

Figure 131

Indeed, consider secant ABB and tangent AT (Fig. 130). We repeat the proof of 131 (2) with no modification except replacing both letters C, C by the letter T.

The preceding argument is another example of the fact (Book II, 67) that the tangent should be thought of as having a double intersection with the circle at the point of contact.

Conversely, If we take points B, B' on line ABB', both on the same side of A, and on another line AT we take point T such that AT is geometric mean between AB and AB', then the three points B, B', T are on a circle tangent to AT at T.

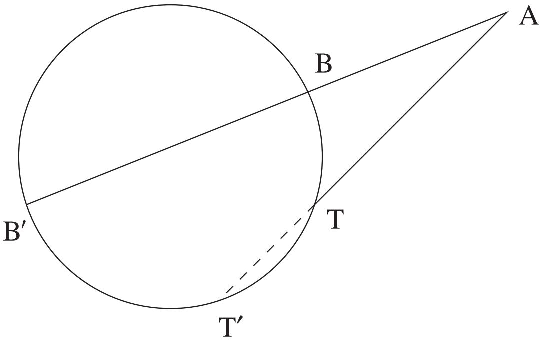

Figure 132

Indeed, circle $BB'T$ has the common point T with line AT (Fig. 132); if it had another common point T', then (131) $AB \cdot AB' = AT \cdot AT'$. This equation, compared with the hypothesis $AB \cdot AB' = AT^2$, shows that point T' must coincide with T.

133. Definition. For any point A, the product of the segments from A to the points of intersection of a secant with a given circle passing through the point is called the power of point A with respect to the given circle. According to 131, this product is independent of the secant. The product is preceded by a + sign if the point A is outside the circle, and by a − sign if A is inside. When the point is exterior to the circle, the power is equal to the square of the tangent from this point.

134. The power of the point A with respect to a circle of center O is equal to the difference of the square of OA and the square of the radius.

Indeed, let us take line OA as our secant (in place of ABB'). The segments AB, AB' represent the sum and the difference between OA and the radius of the circle; their product is therefore the difference of the squares of these quantities.

If one takes into account the sign of the power, this is always equal to $d^2 - R^2$ (where d is the distance OA and R is the radius).

135. When two circles meet at a right angle, the square of the radius of each one is equal to the power of its center with respect to the other, and conversely.

If circles O, O' (Fig. 133) intersect in A at a right angle, the tangent to circle O' at this point is precisely OA, and the power of point O with respect to circle O' is indeed OA^2.

Conversely, if the power of point O with respect to circle O' is OA^2, then OA is tangent to this circle, and therefore the two circles are orthogonal.

136. Theorem. The locus of points which have the same power with respect to two given circles is a line perpendicular to the line of the centers.

This line is called the radical axis of the two circles.

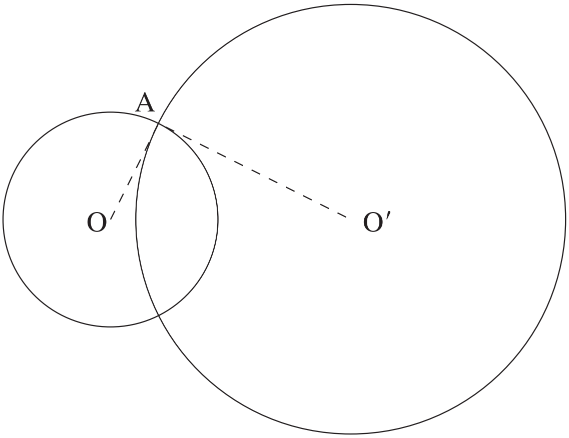

Figure 133

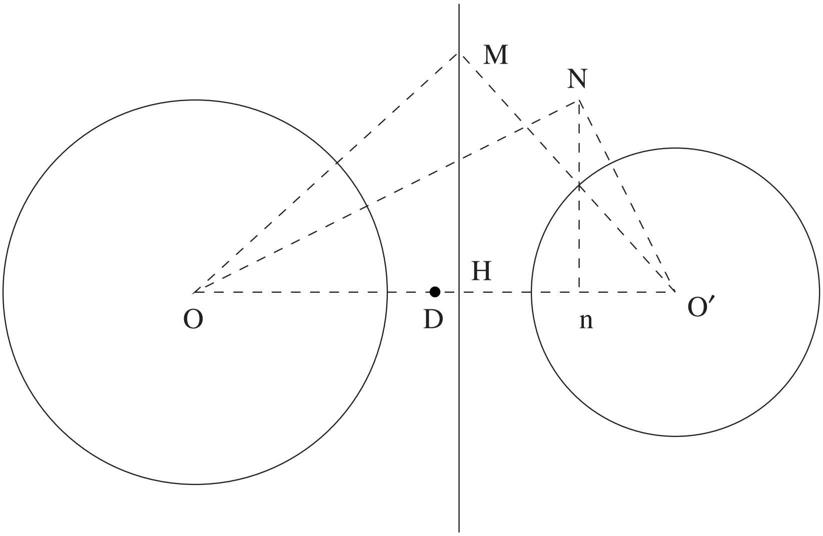

Figure 134

Consider two circles with centers O, O' (Fig. 134) and radii r, r'. If the point M has the same power with respect to the two circles, we have $OM^2 - r^2 = O'M^2 - r'^2$, which can be rewritten as $OM^2 - O'M^2 = r^2 - r'^2$ and gives us the desired conclusion (128b).

According to the results of 128b, the distance between the intersection H of the radical axis with the line of the centers, and the midpoint D of this segment, is given by the formula $2DH \cdot OO' = r^2 - r'^2$.

Remarks. I. The preceding proof remains valid if one of the circles has a radius equal to zero, and so reduces to its center. Thus, the locus of points such that their power with respect to a given circle is equal to the square of their distance to a given point is a line perpendicular to the line joining the given point to the center of the circle. We can call this line the radical axis of the circle and the point.

II. Two concentric circles do not have a radical axis.

III. The difference between the powers of an arbitrary point with respect to two circles is equal to twice the product of the distance of this point to the radical axis and the distance between the centers (Fig. 134).

Indeed, let N be the point in question, and suppose n is its projection onto OO'. The difference of the powers of this point with respect to the two circles will be

$$ON^2 - r^2 - (O'N^2 - r'^2) = ON^2 - O'N^2 - (r^2 - r'^2).$$

Since $ON^2 - O'N^2 = 2Dn \cdot OO'$ and $r^2 - r'^2 = 2DH \cdot OO'$, we obtain

$$ON^2 - O'N^2 - (r^2 - r'^2) = 2Hn \cdot OO'.$$

137. If two circles intersect, their radical axis is their common chord.

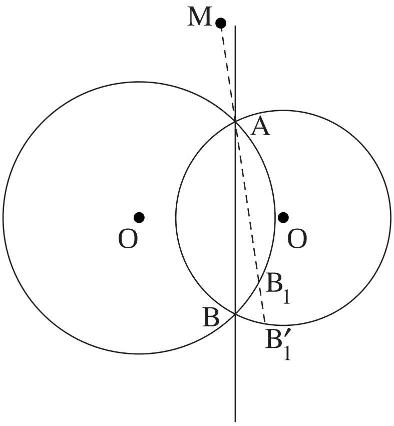

Figure 135

Indeed, it is clear that every point of the common chord AB (Fig. 135) belongs to the locus. It follows that, conversely, every point of the radical axis belongs to this chord. This can be seen directly by drawing MA: if this line intersects the two circles in two distinct new points B_1, B_1', the two powers $MA \cdot MB_1$ and $MA \cdot MB_1'$ will be different.

In the same way, if two circles are tangent, their radical axis is their common tangent.

138. The radical axis of two circles (or at least the part of the axis which is exterior to the two circles) is the locus of the centers of circles cutting the first two at right angles.

This is true because the center of such a circle has the same power with respect to the given circles; namely, the square of the radius.

The radical axis divides a common tangent into two equal parts.

139. Theorem. Given three circles, the three radical axes determined by pairs of them are either parallel or concurrent.

This is true because the point of intersection of two of the radical axes will have same power with respect to the three axes, and will therefore belong to the third radical axis as well. This point is called the radical center of the three circles; if it is exterior, it is the center of a circle cutting all three of them at right angles.

Remark. If two of the radical axes coincide, the preceding argument shows that the third must coincide with the first two. The three circles then have the same radical axis. Every circle orthogonal to the first two will be orthogonal to the third as well.

Exercises

Exercise 148. We are given a circle and two points A, B in its plane. Through A we draw a variable secant AMN which cuts the circle in M, N. Show that the circle passing through M, N, B also passes through another fixed point.

Exercise 149. The locus of points such that the ratio of their powers with respect to two given circles is equal to a given number is a circle which has the same radical axis with the first two.

Exercise 150. The preceding exercise is equivalent to exercise 128 if the two given circles intersect in two points.

Exercise 151. Consider a triangle ABC, points D, D' on BC, E, E' on CA, and F, F' on AB. If it is known that there is a circle passing through D, D', E, E', a circle passing through E, E', F, F', and a circle passing through F, F', D, D', it follows that all six points D, D', E, E', F, F' are on the same circle.

Exercise 152. When do all the circles orthogonal to two given circles O, O' intersect line OO'? Show that if they all do, then the intersection points are always the same (the limit points of Poncelet); namely, the points that have the same radical axis with the given circles.

Exercise 153. Four points A, B, C, D are given on the same line. A variable circle passes through A, B, and another variable circle passes through C, D. Show that the common chords of these circles pass through a fixed point.

Exercise 154. If the radical center of three circles is interior to these circles, then it is the center of a circle divided into two equal parts by each of the first three circles.

CHAPTER VI

Constructions

151. Construction 1. Divide a segment into equal parts, or into parts proportional to given segments.

First method (Fig. 148). Suppose segment AB must be divided, say, into three equal parts. On an arbitrary line passing through A, and starting from this point, we construct three equal segments AC, CD, DE.

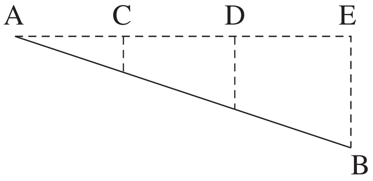

Figure 148

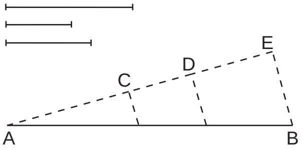

Figure 149

It suffices to draw BE, then draw parallels to this line through C and D.

These parallels will divide AB into three equal parts (113). If we need to divide AB into parts proportional to three given segments (Fig. 149), we take AC, CD, DE equal to the three segments in question.

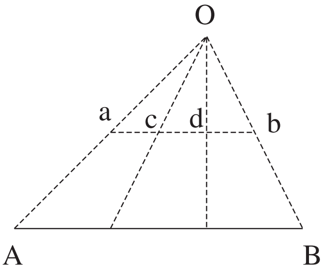

Figure 150

Second method (Fig. 150). Again, suppose segment AB must be divided into three equal parts. We draw an arbitrary parallel ab to AB, and on it take equal segments ac, cd, db. We then draw Aa, Bb, which will intersect in O. Lines Oc, Od then divide AB into three equal parts (121).

If we need to divide AB into parts proportional to three given segments (Fig. 149), we take ac, cd, db equal to the three segments in question.

Construction 2. Find the fourth proportional to three given segments.

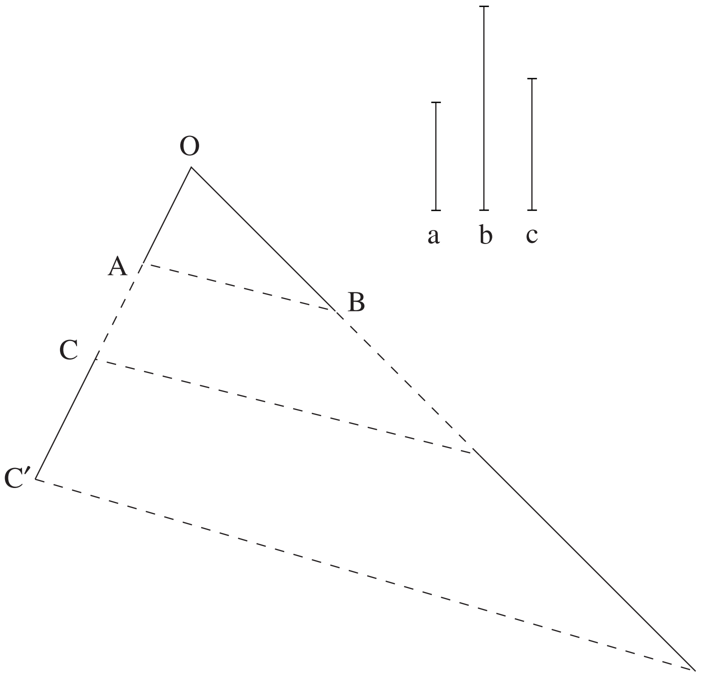

Figure 151

The constructions are similar to the preceding.

First method (Fig. 151). Consider the given segments a, b, c. We must find x such that $\frac{a}{b} = \frac{c}{x}$. On the sides of an arbitrary angle AOB, we construct segments $OA = a$ and $OB = b$. Then segment c can be placed anywhere on the side OA (in the figure, it is CC'). The lines parallel to AB through the endpoints of this segment determine on OB a segment equal to the required fourth proportional.

If we take one of the endpoints of segment c to coincide with point A or point O, we need, in general, to draw only a single parallel.

We leave it to the reader to apply the second method used in the preceding construction to this problem.

152. Construction 3. On a given segment, construct a triangle similar to a given triangle.

Suppose the given the triangle is ABC, and the given segment is A_1B_1. We find a point D on AB such that $AD = A_1B_1$, and draw a parallel DE to BC (Fig. 152), extending it to intersect AC at E. Then it remains only to construct a triangle on A_1B_1 congruent to ADE (Book II, 86, Construction 4).

This construction obviously gives a solution to the following problem:

Problem. Given two points A, B of a figure F and their corresponding points A', B' in a similar figure F', find the point corresponding to an arbitrary third point of F.

In particular, repeated application of this construction will allow us to perform

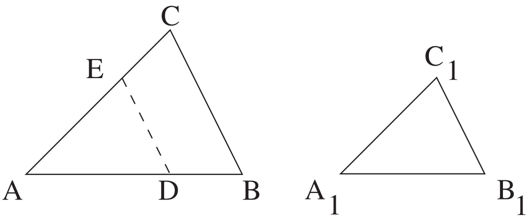

Figure 152

Construction 3b. On a given segment, construct a polygon similar to one given.

153. Construction 4. Construct the geometric mean of two segments.

We have stated three theorems leading to geometric means (123, 125, 132). Each of these theorems provides a method for solving the proposed problem.

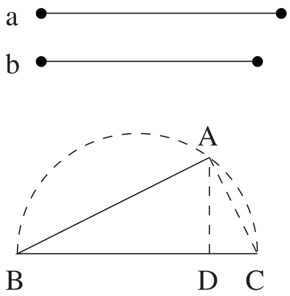

Figure 153

First method (Fig. 153). Starting from a point B on a line, we mark off, in the same sense, segments BD, BC equal to the given segments a, b (labeling the smaller of the two BD). We will construct a right triangle with hypotenuse BC, in which BD will be the projection of another side onto the hypotenuse. To do this, it suffices to place the vertex A of the right angle at the intersection of the perpendicular erected from D to BC with the circle on diameter BC. Then AB will be the required geometric mean (123).

Second method (Fig. 154). From an arbitrary point D on an arbitrary line, and in opposite senses, we construct segments DB, DC equal to the given segments a, b. We will consider them to be the projections of the legs of a right triangle on its hypotenuse. To do this it will suffice to place the vertex A of the right angle at the intersection of the perpendicular from D to BC with the circle with diameter BC. Then AD will be the geometric mean (125).

Third method (Fig. 155). Starting from a point O on a line, and in the same sense, we construct segments OA, OB equal to the given segments a, b. We draw an arbitrary circle through A and B, and draw a tangent OT to this circle. This tangent is the geometric mean (132).

154. Construction 5. Construct a segment whose square is the sum of the squares of two given segments.

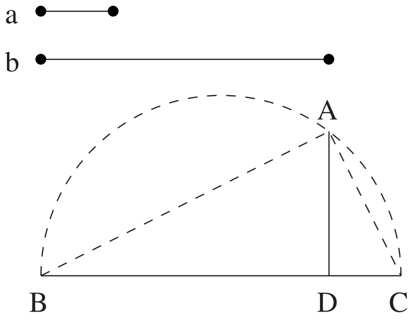

Figure 154

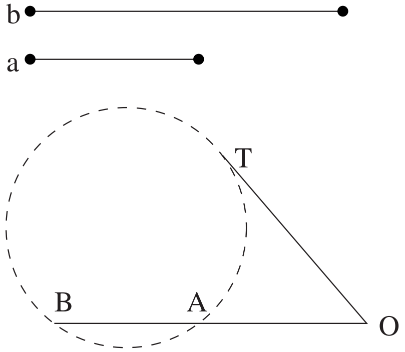

Figure 155

We construct segments equal to the given ones on the two sides of a right angle, starting from the vertex. The required segment will be the hypotenuse of the right triangle thus formed.

Construction 6. Construct a segment whose square is the difference of the squares of two given segments.

We construct (Book II, 87b, Construction 9) a right triangle having the larger of the given segments as hypotenuse, and the other as a side of the right angle. The other side of the right angle gives the required segment.

155. Construction 7. Construct two segments, knowing their sum and their product.

Supppose we must find two segments whose sum is equal to a given segment $a = BC$ and whose product is equal to the product of two given segments b, c. We may assume that these last two lengths are laid off (as BB , CC along the perpendiculars to the endpoints of BC) in the same sense (Fig. 156). Since the sum of the required segments is BC, they can be represented as BM and CM, for some point M on BC. Now, the product $BM \cdot CM$ must equal $BB \cdot CC$, and therefore

$$\frac{BB}{BM} = \frac{CM}{CC},$$

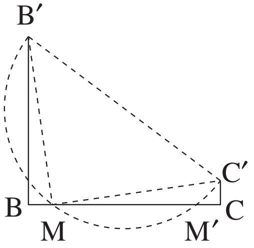

Figure 156

which shows that the right triangles $BB'M$, CMC' are similar. Thus angle BMB' is equal to the corresponding angle $MC'C$, and therefore complementary to CMC', so that angle $B'MC'$ is a right angle. Point M will be on the circle with diameter $B'C'$.

Conversely, if the circle with diameter $B'C'$ intersects BC in M, triangles $BB'M$, $CC'M$ will be similar, since they have perpendicular sides, and the equality $BM \cdot CM = BB' \cdot CC'$ follows.

Conditions for existence. We have taken lines b and c arbitrarily, but we may assume that they are equal: we may replace them, without changing their product, by their geometric mean. Assume thus that $BB' = CC'$, so that $BB'CC'$ is a rectangle. The midpoint of $B'C'$ is then at a distance from BC equal to BB', and the radius of the circle is one half $B'C'$, which is the same as one half BC. Therefore the circle will intersect BC if BB' is at least equal to half of BC; in other words, if the given product is at most equal to the square of half of BC.

If M is a point satisfying the conditions of the problem, then, turning the segment BC around on itself or taking the symmetric figure with respect to the midpoint of BC (which is the same thing), we obtain a point M' which will have the same property as the first, since $BM' = CM$, $BM = CM'$. We see thus that the intersections M, M' of BC with the circle with diameter $B'C'$ are symmetric with respect to the midpoint of BC, which can also be verified directly with the help of 62.

If the given product is equal to the square of half of BC (which, as we have just shown, is the greatest value it can take), the circle with diameter $B'C'$ is tangent to BC; the points M, M' coincide with each other and with the midpoint of BC, so that $BM = CM$. We thus see that the product of two segments whose sum is constant is maximal when these two segments are equal.

Remark. Let x denote segment BM. We have $CM = a - x$, and the equality $BM \cdot CM = BB' \cdot CC'$ can be written $x(a - x) = bc$, or

$$x^2 - ax + bc = 0.$$

The preceding construction allows us to find a segment x satisfying

$$x^2 - ax + q = 0,$$

where a is a given segment, and q is the product of two given segments.

Construction 8. Construct two segments, knowing their difference and their product.

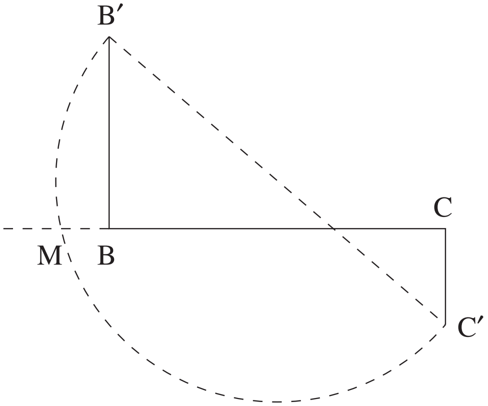

Figure 157

Suppose we wish to find two segments whose difference is equal to a given segment $a = BC$, and whose product is equal to the product of two given segments b, c. We may assume that these last two lengths are laid off as BB', CC', perpendicular to the endpoints of BC in opposite senses (Fig. 157). Since the difference of the required segments is BC, they can can be represented as BM, CM, where M is a point on an extension of BC (in one of two directions) satisfying $BM \cdot CM = BB' \cdot CC'$. As in the preceding construction, we conclude that point M must belong to the circle with diameter $B'C'$ and, conversely, a point of intersection of this circle with an extension of BC satisfies the conditions of the problem.

Since the points B', C' are on different sides of BC, the circle always intersects the line, and therefore a solution is always possible.

Remark. Let x be the smaller of the segments found above; the other can be represented as $y = a + x$, and the equality $xy = bc$ becomes either

$$x(a + x) = bc, \quad x^2 + ax - bc = 0,$$

or

$$(y - a)y = bc, \quad y^2 - ay - bc = 0.$$

We can therefore find segments satisfying either of the following two conditions

$$x^2 + ax - q = 0,$$

$$y^2 - ay - q = 0,$$

where a is a given segment, and q is the product of two given segments.

156. We say that a line is divided into extreme and mean ratio if the greater segment is geometric mean between the whole segment and the smaller part.

Construction 9. Divide a segment into extreme and mean ratio.

Let the given line segment be BC (Fig. 158), on which we need to find a point D such that $\frac{BC}{BD} = \frac{BD}{CD}$. In this proportion, let us add the numerators and denominators. We obtain

$$\frac{BC}{BD} = \frac{BC + BD}{BC}.$$

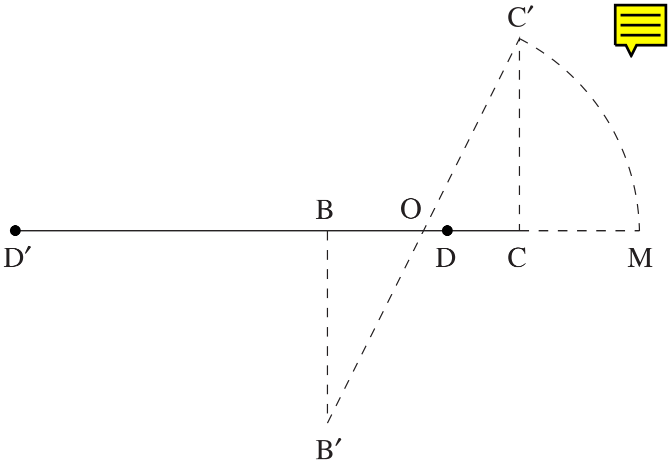

Figure 158

Thus the two segments BD and BC + BD, whose difference is equal to BC, have a product equal to BC^2. This observation leads us to the preceding construction. We must construct perpendicular segments BB′, CC′ on the two endpoints of BC, and in opposite senses, both equal to BC. The circle with diameter B′C′ will intersect the extension past C of BC in a point M. Then if we construct BD = CM, point D is the required point of division.

Starting from B, and in the sense opposite to BC, we construct a segment BD′ = BM. The point D′ thus constructed has a property analogous to that of D: its distance to B is the geometric mean between its distance to C and the segment BC. Indeed, we have BD′ = BC + BD, so that the proportion $\frac{BC}{BD} = \frac{BC+BD}{BC}$ can be rewritten as $\frac{BC}{BD} = \frac{BD'}{BC} = \frac{CD'}{BD'}$.

The point D′ is said to divide the segment BC into mean and extreme ratio externally.[1]

Note that in order to have this property, the point D′ must be chosen as described, because the proportion $\frac{BD'}{BC} = \frac{CD'}{BD'}$ gives, conversely, $\frac{BD'}{BC} = \frac{BC}{BD'-BC}$ (by taking the difference of the numerators and denominators), so that BD′ is one of the segments whose difference is BC and whose product is BC^2.

Let BD = a: we propose to calculate BD and BD′. We remark first that since BB′ and CC′ are equal, the circle with diameter B′C′ has as its center the midpoint O of BC, so that $OC = \frac{a}{2}$. The theorem about the hypotenuse of right triangle, applied to triangle OCC′, tells us that:

$$OC' = OM = \sqrt{\left(\frac{a}{2}\right)^2 + a^2} = \sqrt{\frac{5}{4}a^2} = \frac{a\sqrt{5}}{2}.$$

We therefore have

$$CM = BD = a\frac{\sqrt{5}-1}{2}, \quad BM = BD' = a\frac{\sqrt{5}+1}{2}.$$

---

[1] The points D and D′ are not conjugate with respect to AB (see Exercise 173).

157. Problem. Find the locus of points with the property that the ratio of their distances to two given lines is equal to a given ratio.

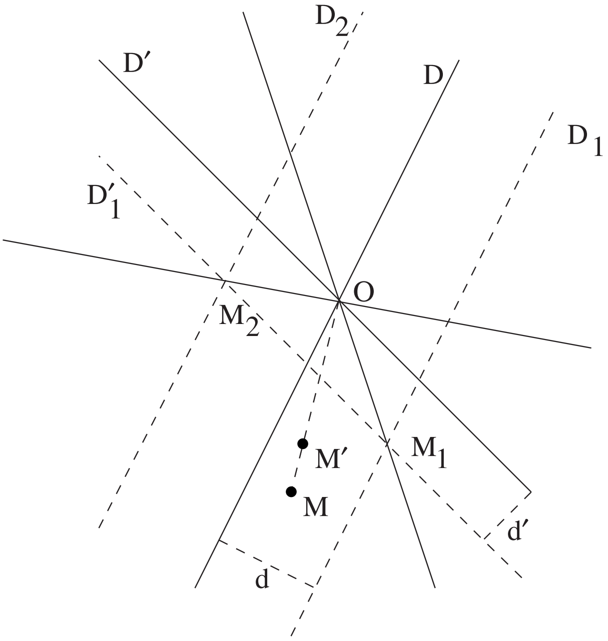

Figure 159

Let D, D' be the given lines, and assume that they meet at point O (Fig. 159). We are looking for points whose distances to these lines are in the ratio $d : d'$ of two given segments. If M is a point on this locus, every point M' on line OM will also be on the locus, because the perpendicular segments from M to D, D', and the corresponding perpendicular segments from M', obviously form two homothetic figures with center of homothecy O. It follows that the locus consists of lines passing through O. We can obtain some points on this locus by finding points whose distances to D, D' are equal to d, d', respectively. We know that the locus of points at a distance d from D consists of two lines D_1, D_2 parallel to D. Likewise, the locus of points at a distance d' from D' consists of two lines D'_1, D'_2 parallel to D'. Let M_1, M_2 be the intersections of D_1, D_2 with D'_1. The lines OM_1, OM_2 are part of our locus. They are in fact the entire locus. Indeed, if M is a point on the locus, then line OM intersects D'_1 in a point M' which, being at distance d' from D', must be at distance d from D, and must therefore coincide with M_1 or with M_2.

If the two lines D, D' are parallel, there would be two points of the locus on any arbitrary common perpendicular:[2] these would be the two conjugate points which divide this perpendicular in the given ratio. The locus consists of parallels to the original lines through these two points.

---

[2] Except when the given ratio is one. In this case there is only one point, the midpoint of the common perpendicular, the other point being thrown to infinity.

Construction 10. Find all points whose distances to three given lines have given ratios.

We first construct the two lines which are the loci of points whose distances to the first two lines are in the given ratio; we then do the same for the second and third given line. In this way we obtain two new lines which will intersect the first two parallels in four points which satisfy the conditions of the problem.

158. Construction 11. Construct the common tangents to two circles.

We have solved this problem in Book II, 93. The preceding theorems allow us to give an entirely different solution. It suffices to draw two parallel radii, so as to determine the centers of similarity (143), and to draw from either of these centers a tangent to one of the circles: this line will be tangent to the second as well because of the homothecy of the two figures.

Construction 12. Find the radical axis of two circles.

If the two circles are secant or tangent, we need only draw the common chord or the common tangent.

If not, we draw an arbitrary third circle which intersects the two circles. The intersection of its two common chords with the given circles will belong (139) to the required radical axis. It suffices then to drop a perpendicular from this point to the line of the centers. Alternatively, we can repeat the construction with another circle in order to determine a second point.

Remark. This construction can be applied to a circle and a line, which leads us to state that the radical axis of a circle and a line is the line itself.

The construction can also be applied in the limiting case when one of the circles is reduced to one point. The auxilliary circle must then pass through this point.

The preceding construction also allows us to find the radical center of three given circles, and therefore (139) to solve the following problem:

Construction 13. Construct the circle orthogonal to three given circles. In particular, construct:

a circle passing thorough a given point and orthogonal to two given circles; or

a circle passing through two points and orthogonal to a given circle.

159. Construction 14. Construct a circle passing through two given points and tangent to a given line.

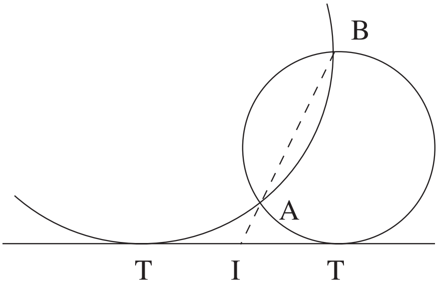

Figure 160

Consider points A, B (Fig. 160) and line xy; let T be the point of contact of the required circle with xy. Extend AB to its intersection I with xy. The segment IT is known: it is the geometric mean of IA and IB.

Conversely, if point T verifies this condition, there exists (132) a circle passing through A, B and tangent to xy at T: it is determined by construction 13 of Book II, 90.

Since the geometric mean of IA, IB can be placed on either side of point I on xy, there are two solutions to the problem.

The solution is clearly possible only when A, B are on the same side of line xy.

Remarks. I. The preceding method does not apply when AB is parallel to xy. In this case the problem has only one solution, the point of contact being the intersection of xy with the perpendicular bisector of AB.

II. On the other hand, this method does apply when A and B coincide on the line IA intersecting xy at I; that is, when we are required to find a circle tangent to IA at A and also tangent to xy. The point of contact T can be found by laying off, in either sense, the segment IT = IA.

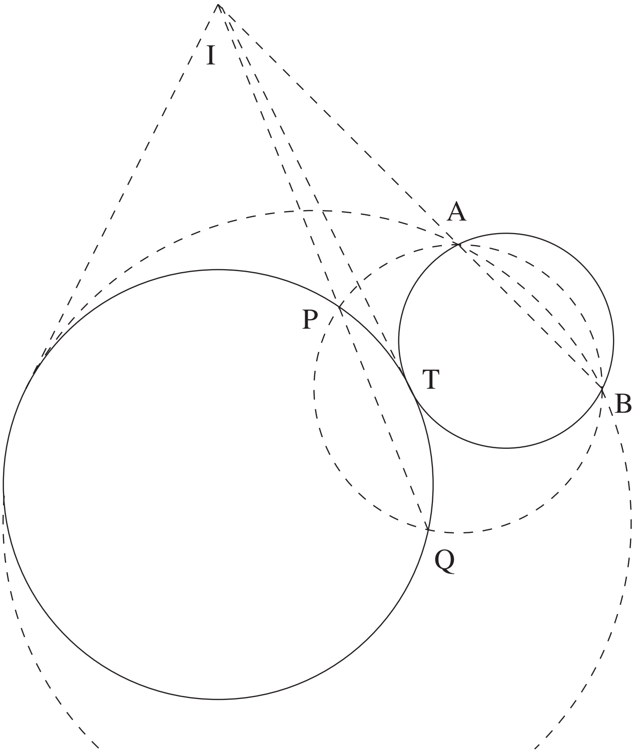

Figure 161

Construction 15. Construct a circle passing through two given points and tangent to a given circle.

Consider two points A, B (Fig. 161) and circle C. We are looking for the point of contact T with the required circle.

To find this point, let I be the point of intersection of AB with the tangent at T. Take an arbitrary circle passing through A, B, and intersecting C in P, Q. Point I will be on line PQ (139).

Since the intersection of AB with PQ gives us point I, it suffices to construct tangents to C through this point. The problem thus has two solutions.

Condition for existence. The point I must be exterior to circle C. For this to happen, it is necessary and sufficient that this point lie on the extention in one direction or the other of segment PQ; that is, outside the auxiliary circle. This will happen when points A, B are both on the same arc determined on the auxilliary circle by P, Q; that is, when A and B are both inside or both outside the given circle.

The remarks of the preceding construction apply here, with respect to the case when A and B are two identical points on a given line.

Exercises

Exercise 163. Find a segment whose ratio to a given segment is equal to the ratio of the squares of two given segments.

Exercise 164. Find a segment such that the ratio of its square to the square of a given segment is equal to the ratio of two given segments.

Exercise 165. Draw a segment through a given point which is divided into a given ratio by two given lines.[3]

Exercise 166. Through a point exterior to a circle, draw a secant which is divided into mean and extreme ratio by the circle.

Exercise 167. Given three consecutive segments AB, BC, CD on a line, find a point at which they subtend the same angle.

Exercise 168. Through two points on the same diameter of a circle, draw two equal chords which intersect on the circle.

Exercise 169. Construct a triangle knowing two sides and the bisector of the angle they include.

Exercise 170. Construct a triangle knowing a side, the corresponding altitude, and the product of the other two sides.

Exercise 171. Construct a triangle knowing the angles and the perimeter; or the angles and the sum of the medians; or the angles and the sum of the altitudes, etc.

Exercise 172. Construct a triangle knowing its three altitudes (Exercise 146).

Exercise 173. The harmonic conjugate of point B with respect to segment DD (Fig. 158, 156) is obtained by extending segment BC by its own length. The harmonic conjugate of D with respect to BC is the point symmetric to D with respect to the midpoint of BC. Also show that the circle with diameter DD passes through the vertices (other than B, C) of the square with diagonal BC.

---

[3] See solution for a note on this problem statement. transl.

Exercise 174. Given a line and two points A, B, find a point on the line at which AB subtends the greatest possible angle.

This problem can be treated indirectly, by first seeking points on the line at which AB subtends a given angle, then finding the greatest angle for which the construction is possible.

The next two problems can be approached in the same way.

Exercise 175. Draw a common perpendicular to two given parallel lines which subtends the greatest possible angle at a given point.

Exercise 176. Find a circle passing through two given points, and intercepting a chord of given length on a given line. Find the minimum possible length when the two points are on different sides of the line.

Exercise 177. Construct a circle passing through a given point and having the same radical axis as two given circles.

CHAPTER VII

Regular Polygons

160. A regular polygon is a convex polygon all of whose sides are equal, and all of whose angles are equal. A regular broken line is a broken line all of whose sides are equal, and all of whose angles are equal and have the same orientation.

161. Theorem. If a circle is divided into an arbitrary number n of equal parts, then:

1 . The division points are the vertices of a regular polygon;

2 . The tangents at these points are the sides of a second regular polygon.

1 . Two consecutive sides of the polygon with vertices at the points of division are obviously symmetric with respect to the radius ending at their common endpoint; two consecutive angles of this polygon are symmetric with respect to the radius perpendicular to their common side.

2 . Two consecutive angles of the polygon formed by the tangents at the points of division are obviously symmetric with respect to the radius perpendicular to their common side; two consecutive sides of this polygon are symmetric with respect to the radius perpendicular to the chord joining their points of contact.

162. Theorem. Conversely, every regular polygon and, more generally, every regular broken line, can be inscribed in a circle and circumscribed about another circle.

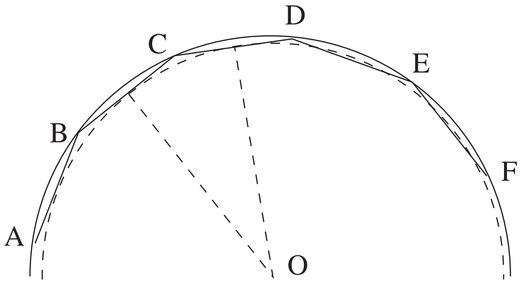

Figure 162

Consider, for instance, the regular broken line ABCDEF (Fig. 162). We draw the circle cirumscribing triangle ABC, whose center O is on the perpendicular bisector of BC. We claim that this circle passes through the point D. To see this, it suffices to observe that the sides BA, CD are symmetric with respect to the perpendicular bisector of BC; indeed, the segment symmetric to BA coincides with CD in direction (because of the equality of the angles at B and C) and length (because $BA = CD$). Thus we have $OA = OD$. Likewise, we see that this same circle, circumscribed about BCD passes through E, and so on.

Now, chords AB, BC, etc., are equal chords of the circumscribed circle, and they are therefore equally distant from the center O. Therefore the sides of the broken line are tangent to a second circle with center O.

The radius of this second circle, or the distance from an arbitrary side to the center, is called the apothem of the regular broken line (or of the regular polygon).

Remark. Every regular polygon with n sides can be superimposed on itself by certain rotations (namely the rotations whose angle is a number of times the n^{th} part of the circle) and by certain line reflections (reflections in the perpendicular bisectors of the sides or the bisectors of the angles).

The number of axes of symmetry is always n. If $n = 2p$ is even, the vertices form diametrically opposite pairs, and this gives p axes of symmetry, each of which is the common bisector of a pair of opposite angles. Likewise, the sides give p axes of symmetry, each perpendicular to two opposite sides.

If $n = 2p + 1$ is odd, each axis of symmetry bisects an angle, and is at the same time the perpendicular bisector of the opposite side.

163. Thus there are infinitely many regular polygons with the same number n of sides, but all of them are similar.

Theorem. Two regular polygons with the same number of sides are similar.

Indeed, two regular polygons with the same number of sides inscribed in congruent circles are obviously congruent: they will coincide if we superimpose their circumscribed circles and one vertex.

If we now are given two regular polygons P, P' inscribed in circles C, C', we can draw a circle concentric to C and congruent to C'. If we look at the intersections of this circle with the radii of C drawn to the vertices of P, we will find a regular polygon which is homothetic to P with respect to the center of C, and congruent to P'.

Remark. In particular, all regular polygons with the same number n of sides have the same angle. It is easy to calculate this angle. Since the sum of the angles of one of these polygons is $2n - 4$ right angles, each one will be the nth part of this measure; that is, $2 - \frac{4}{n}$ right angles.

164. Having divided a circle into n equal parts, let us join the points of division separated by p equal arcs, starting from one division point, until we visit some vertex for the second time. If P is the first vertex which is visited twice, this vertex can only be the one where we started, because if it were the endpoint of a side NP, already traversed in the sense NP, it is clear that we would have visited N first before returning to P.

The figure thus obtained is (21, Remark) an improper polygon, and is called a regular star polygon. For instance, Fig. 163 represents the star pentagon obtained by dividing the circle into five equal parts at the points A, B, C, D, E and joining them in the order $ACEBDA$ ($n = 5$, $p = 2$).

We may assume that n and p are relatively prime. Indeed, if they had a common divisor $d > 1$, we would obtain the same construction replacing n, p by $\frac{n}{d}$, $\frac{p}{d}$.

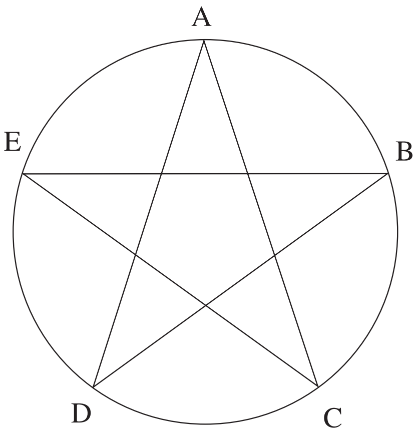

Figure 163

If n and p are relatively prime, the star polygon will have precisely n sides. Indeed, since each side spans p parts of the circle, tracing k sides we will have spanned kp parts. We will have returned to the starting point after traveling a whole number of circles, that is if kp is a multiple of n. It is shown in arithmetic[1] that this happens for some value $k < n$ if p and n are not relatively prime; but if, on the other hand, p is relatively prime to n, this happens for the first time when $k = n$.

We can thus form star polygons with n sides by dividing the circle into n equal parts, and joining every p^{th} division point, where p is any number less than, and relatively prime to, n. Every star polygon is obtained in two ways by this construction, because if its side is the chord corresponding to an arc equal to p divisions, it is also the chord corresponding to $n - p$ divisions. For instance, the star polygon in Figure 163 can be obtained either by joining every second point, or by joining every third point. Thus one half of the values of p must be discarded if we want to find each polygon once. The value $p = 1$ (and therefore $p = n - 1$) corresponds to the usual regular polygon.

Example. Consider $n = 15$. Besides 1 and 14, the numbers less than 15 and relatively prime to it are 2, 4, 7, 8, 11, 13. We need not consider the values $p = 8$, 11, 13, which give the same polygons as 7, 4, 2. Thus there is one proper regular pentadecagon, and three star pentadecagons.

165. Construction of inscribed regular polygons. If we know how to inscribe a regular polygon, we also know how to circumscribe the regular polygon with the same number of sides.

This last is formed by the tangents to the circle through the vertices of the first.

If we know how to inscribe a regular polygon, we also know how to inscribe a regular polygon with double the number of sides.

It suffices to divide the arc intercepted by the first polygon into two equal parts.

---

[1] See Tannery, Leçons d'Arithmetique, chap. IV.

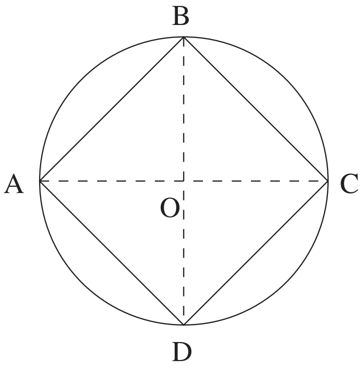

Figure 164

166. The Square. To inscribe a square in a circle O (Fig. 164), it suffices to draw two perpendicular diameters AC, BD: they will divide the circle into four equal parts.

The preceding remark allows us to construct the inscribed regular octagon, then the regular polygons with 16, 32, . . . and, generally, 2^n sides.

The side of the square inscribed in a circle of radius R is $c_4 = R\sqrt{2}$.

This is true because in right triangle AOB (Fig. 164), we have

$$AB^2 = AO^2 + BO^2 = 2R^2.$$

The apothem of an inscribed regular polygon can be calculated by the following rule: the square of the apothem equals the square of the radius of the circumscribed circle minus the square of half a side. This follows from the fact that the apothem, radius, and half a side form a right triangle.

Thus, the apothem a of a polygon with side c inscribed in a circle of radius R is $a = \sqrt{R^2 - \frac{c^2}{4}}$.

For the square, this apothem is

$$a_4 = \sqrt{R^2 - \frac{c_4^2}{4}} = \sqrt{R^2 - \frac{2R^2}{4}} = \frac{R\sqrt{2}}{2}.$$

It is equal to half the side, which was evident a priori.

167. The Hexagon.

Theorem. The side of a regular hexagon is equal to the radius of its circumscribed circle.

Let AB (Fig. 165) be the side of a regular hexagon inscribed in a circle with center O. We claim that triangle OAB is equilateral. We see right away that it is isosceles. Also, the angle at O intercepts a sixth of the circle, so it is equal to $\frac{4}{6} = \frac{2}{3}$ of a right angle. There remain $2 - \frac{2}{3} = \frac{4}{3}$ of a right angle for the sum of the angles at A and B. Since these two angles are equal, both are $\frac{2}{3}$ of a right angle. The triangle is equiangular, and so equilateral.

The apothem of the regular hexagon is

$$a_6 = \sqrt{R^2 - \frac{R^2}{4}} = \frac{R\sqrt{3}}{2}.$$

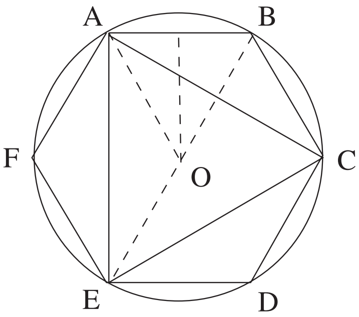

Figure 165

The Equilateral triangle. Joining every other vertex of our hexagon, we form inscribed equilateral triangle ACE (Fig. 165).

Side AC of this triangle, being perpendicular to radius OB, is equal to twice the altitude of triangle AOB; that is (since the three altitudes of an equilateral triangle are equal), twice the apothem of the hexagon: $c_3 = R\sqrt{3}$.

This can also be seen by observing that B and E are diametrically opposite (since the arc between them is $\frac{3}{6}$ of the circle). Triangle ABE thus has a right angle at A, and side AE is twice the length of the perpendicular from the midpoint O of BE to AB.

The apothem of the inscribed equilateral triangle is $a_3 = \frac{R}{2}$.

Knowing the construction for the inscribed hexagon, we can derive the construction of inscribed polygons with $12, 24, \ldots, 3 \times 2^n$ sides.

168. The Decagon.

Theorem. The side of the inscribed regular decagon is equal to the greater segment of the two that divide the radius into mean and extreme ratio.

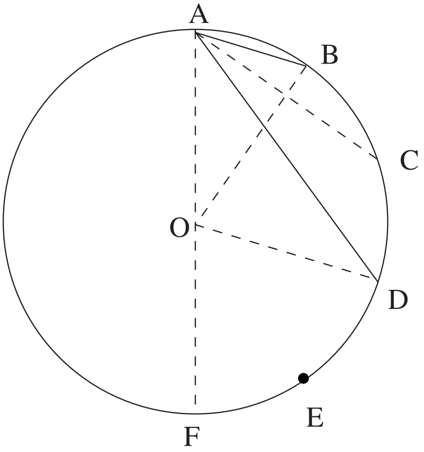

Figure 166

Let AB (fig. 166) be the side of the regular decagon inscribed in a circle with center O. The angle at O is $\frac{4}{10} = \frac{2}{5}$ of a right angle. The sum of the angles OAB, OBA is therefore $2 - \frac{2}{5} = \frac{8}{5}$ of a right angle, so each of them is $\frac{4}{5}$ of a right angle. We construct the bisector AI of OAB (with I on OB), which forms, with OA and AB, angles equal to $\frac{2}{5}$ of a right angle. Angle AIB, exterior to triangle AIO is equal to the sum of the two remote interior angles, which is here $\frac{4}{5}$ of a right angle. It follows that triangles AIB, AIO are isosceles and that AB = AI = IO. The property of the angle bisector (115), $\frac{OI}{IB} = \frac{OA}{AB}$, can thus be rewritten as

$$\frac{OI}{IB} = \frac{OB}{OI}.$$

QED

The side of the regular inscribed decagon can then be determined by Construction 9 in 156. This side is:

$$c_{10} = R\frac{\sqrt{5} - 1}{2},$$

and the apothem is:

$$a_{10} = \sqrt{R^2 - R^2\left(\frac{\sqrt{5} - 1}{4}\right)^2} = \frac{R}{4}\sqrt{10 + 2\sqrt{5}}.$$

169. Aside from the ordinary regular decagon, there is also a star regular decagon, obtained by joining every third vertex of the vertices of the ordinary one.

We can find the side of the ordinary decagon, and the side of the star decagon, from the following result:

Theorem. The difference of the sides of the two regular inscribed decagons is equal to the radius, and their product is the square of the radius.

Once again, we let AB (Fig. 166) be the side of the ordinary regular decagon, so that semicircle $\overset{\frown}{AF}$ is divided into five equal parts by the points B, C, D, E. Then AD is the side of the star decagon, and line AD coincides with the bisector AI of angle OAB, since the inscribed angles FAD, DAB intercept equal arcs on the circle. Radius OD is parallel to AB because alternate interior angles ODA, DAB are both equal to OAD. Thus triangles AIB, DOI are similar; they are also isosceles, as we have seen in the preceding section. We then have AD − AB = AD − AI = ID = OD.

As for the relation $AB \cdot AD = AI \cdot AD = OA^2$, this follows from the similarity of triangles AIO, AOD which have equal angles.

We observe that the sides of the two regular decagons correspond to the two divisions (interior and exterior) of the radius into mean and extreme ratio, obtained in Construction 9.

The side of the star decagon is $c_{10} = R\frac{\sqrt{5}+1}{2}$, and its apothem is:

$$a_{10} = \sqrt{R^2 - R^2\left(\frac{\sqrt{5} + 1}{4}\right)^2} = \frac{R}{4}\sqrt{10 - 2\sqrt{5}}.$$

170. The Pentagon. Joining every other vertex of a regular decagon we obtain the ordinary regular pentagon. Joining every other vertex of the pentagon,

which amounts to joining every fourth vertex of the decagon, we obtain the regular star pentagon.

Side AC of the ordinary regular pentagon is twice the altitude from A in triangle AOB (Fig. 166). Thus we can compute it, knowing the side of the decagon, since we can compute the altitudes of a triangle, given its sides. But, more simply, the side of the ordinary regular pentagon is twice the apothem of the star decagon. This can be seen either from triangle AIO whose altitudes (namely, half the side of the pentagon, and the apothem of the star decagon) are equal, or in right triangle ADF, where side DF is equal to the side of the pentagon, and the parallel through O to this side is equal to the apothem of the star decagon. The side and apothem of the ordinary regular pentagon are therefore

$$c_5 = 2a_{10} = \frac{R}{2}\sqrt{10 - 2\sqrt{5}}, \quad a_5 = \frac{R}{4}(\sqrt{5} + 1).$$

In the same way, by considering right triangle ABF, we see that the side of the star pentagon is twice the apothem of the ordinary decagon, and therefore

$$c_5 = \frac{R}{2}\sqrt{10 + 2\sqrt{5}}, \quad a_5 = \frac{R}{4}(\sqrt{5} - 1).$$

Remark. In general, the same reasoning shows that the calculation of the side and apothem of a polygon (ordinary or star) obtained by dividing the circle into 2n + 1 equal parts, and joining every p^th^ point of division, amounts to the calculation of the polygon obtained by dividing the circle into 4n + 2 equal parts and joining every q^th^ point of division, where q is derived from p by the relation[2]

$$\frac{p}{2n+1} + \frac{q}{4n+2} = \frac{1}{2}, \quad \text{or} \quad 2p + q = 2n + 1.$$

171. The Pentadecagon. Next we consider inscribing a proper regular pentadecagon in a circle. We will know how to do this, once we have constructed the hexagon and decagon, from the following result.

Theorem. The arc intercepted by the side of the regular inscribed pentadecagon is the difference of the arcs intercepted by the side of the hexagon and the side of the decagon.

Indeed, this arc is a 15th of the circle, and the ratio $\frac{1}{15}$ is equal to the difference $\frac{1}{6} - \frac{1}{10}$.

Thus we can start from point A (Fig. 167), and lay off two openings of the compass AB, AC, on the circle, one equal to the side of the regular hexagon, the other to the side of the decagon. Segment BC will be the required side.

Aside from the ordinary regular pentadecagon, there exist three star pentadecagons (164) whose sides intercept $\frac{2}{15}$, $\frac{4}{15}$, and $\frac{7}{15}$ of the circle. These arcs can be easily deduced from the first, and can also be obtained by constructions analogous to the one used for the ordinary pentedecagon.

---

[2] Conversely, this remark allows to reduce the calculations for a polygon with 4n + 2 sides to calculations for a polygon with 2n + 1 sides. This is true because a polygon with 4n + 2 sides corresponds to an odd number q (see 164); 2n + 1 − q will then be even and the relation 2p = 2n + 1 − q will yield an integer p, relatively prime to 2n + 1 and corresponding to a polygon with 2n + 1 sides, for which it will suffice to calculate the side and the apothem. In this way, the calculations done for the pentadecagon (174,175) can be used to find the corresponding results for polygons with 30 sides.

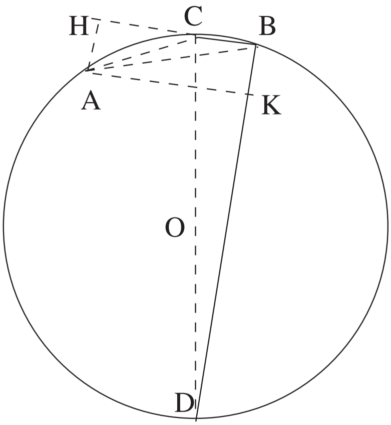

Figure 167

The arc intercepted by the side of the first star pentadecagon is the difference of the arcs intercepted by the sides of the star decagon and that of the hexagon.

The arc intercepted by the side of the second star pentadecagon is the sum of the arcs intercepted by the sides of the proper decagon and that of the hexagon.

The arc intercepted by the side of the third star pentadecagon is the sum of the arcs intercepted by the sides of the star decagon and that of the hexagon.

Indeed, we have:

$$\frac{2}{15} = \frac{3}{10} - \frac{1}{6},$$
$$\frac{4}{15} = \frac{1}{10} + \frac{1}{6},$$
$$\frac{7}{15} = \frac{3}{10} + \frac{1}{6}.$$

172. More generally, if we know how to inscribe regular polygons with m sides and n sides, with m, n relatively prime, we also know how to inscribe a regular polygon with mn sides.

Indeed, the sides of the two known polygons intercept an m^{th} and an n^{th} of the circle. By adding p times the first arc, and subtracting q times the second, we obtain the arc intercepted by the new polygon if

$$\frac{p}{m} - \frac{q}{n} = \frac{1}{mn}, \quad \text{or} \quad np - mq = 1.$$

It is shown in arithmetic[3] that this relation is not possible if m and n have a common factor (different from one), but that it is possible if m and n are relatively prime.

For instance, the regular dodecagons can be inscribed using the constructions for the square and the triangle, because $\frac{1}{12} = \frac{1}{3} - \frac{1}{4}$, $\frac{5}{12} = \frac{2}{3} - \frac{1}{4}$.

The remark of 170 is a special case of this one, corresponding to $m = 2$ and n odd.

---

[3] Tannery, Leçons d Arithmetique, chap. XIV, n 511

173. The geometer Gauss proved that we can inscribe, with straightedge and compass, any regular polygon whose number of sides is of the form $2^n + 1$. Thus we were able to inscribe the regular triangle $(2 + 1 = 3)$ and pentagon $(2^2 + 1 = 5)$. Next come the polygons with $17(= 2^4 + 1)$, $257(= 2^8 + 1)$, etc., sides.[4]

Combining this proposition with those already obtained, we see that we can inscribe, with straightedge and compass, a regular polygon with N sides if the prime factorization of N contains only: 1°. prime factors of the form $2^n + 1$, all distinct; 2° the factor 2 to an arbitrary power.

Conversely, we can prove that the construction cannot be done by straightedge and compass if the number N is not of the type just described.

Thus, the regular polygon with $170(= 2 \times 5 \times 17)$ can be inscribed, but not the polygon with 9 sides, because the number 9 is a power of 2 plus one, but is not prime; alternatively, its factors $(9 = 3 \times 3)$ are certainly of the form $2^n + 1$, but are equal.

174. To calculate the side of the ordinary regular pentadecagon inscribed in a circle of radius R, we again consider sides AB, AC (Fig. 167) of the inscribed decagon and hexagon, so that BC is the required side. If we drop a perpendicular AH onto BC, we will have $BC = BH - CH$.

Now angle ABH, as an inscribed angle, intercepts arc $\overset{\frown}{AC}$, and is therefore equal to half the central angle corresponding to a side of the decagon, so that BH is the apothem of the regular decagon inscribed in a circle with radius AB. Thus

$$BH = \frac{AB}{4}\sqrt{10 + 2\sqrt{5}} = \frac{R}{4}\sqrt{10 + 2\sqrt{5}},$$

since $AB = R$.

Similarly, angle ACH is equal to the inscribed angle intercepting arc $\overset{\frown}{ACB}$ (since both these angles are supplementary to ACB), so it is equal to half the central angle corresponding to the side of the regular hexagon. Thus CH is the apothem of the regular hexagon inscribed in the circle with radius AC, or:

$$CH = \frac{AC\sqrt{3}}{2} = R\frac{\sqrt{5} - 1}{2} \cdot \frac{\sqrt{3}}{2},$$

and consequently

$$BC = \frac{R}{4}\left(\sqrt{10 + 2\sqrt{5}} - \sqrt{3}(\sqrt{5} - 1)\right).$$

It is clear that the preceding method applies whenever we must solve problems of the following type: knowing the chords of two arcs in a circle of given radius, find the chord of the difference of the two arcs.

In particular, to calculate the side of the first star pentadecagon, we use the same reasoning, replacing the side and apothem of the ordinary decagon with the side and apothem of the star decagon. The required side will be

$$R\frac{\sqrt{5} + 1}{2}\frac{\sqrt{3}}{2} - \frac{R}{4}\sqrt{10 - 2\sqrt{5}} = \frac{R}{4}\left((\sqrt{5} + 1)\sqrt{3} - \sqrt{10 - 2\sqrt{5}}\right).$$

---

[4] It is easily proved (it is enough to apply n° 48 and n° 49, application III, of Bourbet, Leçons d'Algébre) that for $2^n + 1$ to be prime it is necessary that n itself be a power of 2; this condition is not sufficient.

Next let us calculate the side of the second star pentadecagon. We again consider the sides AB, AC′ of the hexagon and ordinary decagon, but we draw them in opposite senses along the circle, so that BC′ is the required segment (Fig. 167b).

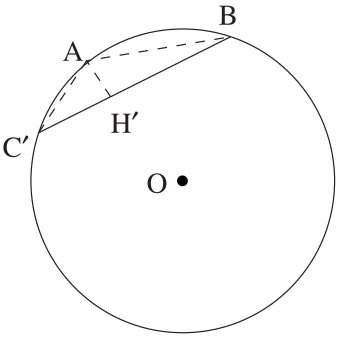

Figure 167b

We drop perpendicular AH′ onto BC′. The two segments BH′, C′H′ will again be equal, the first to the apothem of the ordinary decagon inscribed in the circle with radius AB, the second to the apothem of the hexagon inscribed in the circle with radius AC′. Their sum BC′ will then be

$$BC' = \frac{R}{4}\left(\sqrt{10+2\sqrt{5}}+\sqrt{3}(\sqrt{5}-1)\right).$$

Clearly we can operate the same way whenever we find ourselves in the position of knowing the chords of two arcs inscribed in a circle of given radius, and seeking the chord of the sum of the two arcs.

In particular, for the side of the third star pentadecagon we find the value

$$\frac{R}{4}\left((\sqrt{5}+1)\sqrt{3}+\sqrt{10-2\sqrt{5}}\right).$$

175. We could deduce the apothem of the pentadecagon from the value found for its side, but this would require the extraction of an additional square root. We can instead compute the apothem (without requiring an additional square root), using a method similar to the one used for the side.

In Fig. 167, we construct diameter CD from point C. Triangle BCD shows that BD is again twice the required apothem. We now drop perpendicular AK from A to BD. Angle ADB (as an inscribed angle) is equal to half the central angle corresponding to a side of the hexagon; the segment DK is equal to the apothem of the hexagon inscribed in the circle of radius AD; and thus

$$DK = AD\frac{\sqrt{3}}{2} = \frac{R}{2}\sqrt{10+2\sqrt{5}}\,\frac{\sqrt{3}}{2},$$

because AD is twice the apothem of the regular decagon inscribed in the given circle.

On the other hand, angle ABD is equal to ACD; that is, to the complement of half the central angle corresponding to the regular decagon; angle KAB is equal to half this central angle; and BK is half the side of the decagon inscribed in the circle with radius AB, so:

$$BK = AB\frac{\sqrt{5}-1}{4} = R\frac{\sqrt{5}-1}{4}.$$

The required apothem is thus

$$\frac{BD}{2} = \frac{BK+DK}{2} = \frac{R}{8}\left(\sqrt{3}\sqrt{10+2\sqrt{5}}+\sqrt{5}-1\right).$$

Analogous reasoning gives the following value for the apothems of the first star pentadecagon:

$$\frac{R}{8}\left(\sqrt{3}\sqrt{10-2\sqrt{5}}+\sqrt{5}+1\right).$$

For the second star pentadecagon, we get:

$$\frac{R}{8}\left(\sqrt{3}\sqrt{10+2\sqrt{5}}-(\sqrt{5}-1)\right).$$

For the third star pentadecagon, we get:

$$\frac{R}{8}\left(\sqrt{3}\sqrt{10-2\sqrt{5}}-(\sqrt{5}+1)\right).$$

176. Measurement of the circumference of a circle.

Theorem. Let P be the perimeter of a regular circumscribed polygon, and let p be the perimeter of the inscribed polygon with the same number of sides. If we keep doubling the number of sides indefinitely, then P and p tend to a common limit L.

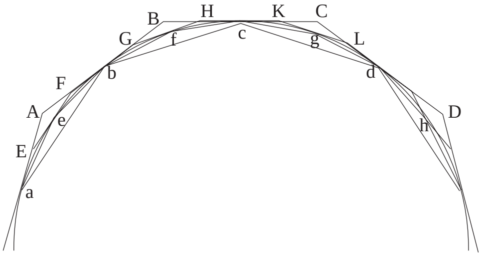

Figure 168

Consider inscribed polygon $abcd\cdots$ (see Fig. 168) and circumscribed polygon $ABCD\cdots$ whose sides have, as points of contact, the vertices of the first. Let us double the number of sides of these polygons, dividing each of the arcs $\overset{\frown}{ab}$, $\overset{\frown}{bc}$, $\overset{\frown}{cd}, \ldots$ into two equal parts, so as to form the new regular inscribed polygon $aebfcg\cdots$ and the corresponding new circumscribed polygon $EFGHKL\cdots$ (Fig. 168). We operate in the same way on the new polygons, and so on, indefinitely. We claim that the perimeters p of the inscribed polygons and the perimeters P of the circumscribed polygons tend to a common value.

We will prove this with the help of the following remarks.

1 . The perimeters p increase as we double the number of sides; for instance, polygon $aebfcg\cdots$ has a greater perimeter than polygon $abcd\ldots$, since this last polygon is inside it;

2 . The perimeters P decrease as we double the number of sides; for example polygon $EFGHKL\cdots$ has a smaller perimeter than polygon $ABC\cdots$, which contains it in its interior;

3 . Any perimeter p is smaller than any perimeter P, because each inscribed polygon is inside each circumscribed polygon.

The quantity p, which increases while always remaining less than a fixed quantity (namely, any value of P) tends to a limit (Leçons d Arithmetique de M. Tannery, chap. XII, n 470).

Likewise, the quantity P, which decreases while always remaining greater than a fixed quantity (namely, any value of p), also tends to a limit.

These limits are equal. Indeed, the two polygons are always similar, and their perimeters are in the same ratio as their apothems. Now the apothem of the circumscribed polygon is equal to the radius R of the given circle, so that

$$\frac{P}{p} = \frac{R}{a},$$

where a denotes the apothem of the inscribed polygon. But as we keep doubling the number of sides indefinitely, the apothem a tends to R. Thus the limit of the ratio $\frac{P}{p}$, which is the ratio of the two limits referred to above (Leçons d Arithmetique chap. XII, n 466) is equal to one. QED

177. We now claim that the perimeter of any convex polygon, inscribed or circumscribed,[5] all of whose sides decrease indefinitely, tends to the limit of the preceding theorem.

In particular, this limit is independent of the initial regular polygon.

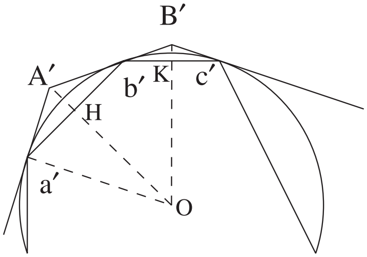

Figure 169

Indeed, consider an arbitrary inscribed polygon $a\,b\,c\cdots$ (Fig. 169), and also the convex circumscribed polygon $A\,B\cdots$ formed by the tangents at $a\,,b\,,c\,\ldots$. Let p and P be the perimeters of the two polygons. The limit L of the preceding theorem is contained between p and P because p, for instance, is smaller than any of the perimeters P considered above, which have limit L, so that p L. In the same way, we find that P L.

---

[5] It is understood that a circumscribed polygon includes the circle in its interior.

On the other hand, line OA intersects ab at its midpoint H, and

$$\frac{aA + Ab}{ab} = \frac{aA}{aH} = \frac{R}{OH},$$

since right triangles AaH, oaA are similar, having a common acute angle. Likewise,

$$\frac{bB + Bc}{bc} = \frac{R}{OK},$$

where K denotes the midpoint of bc, and so on.

Let us add the numerators and denominators of the left hand sides of these equalities. We obtain a ratio contained between the smallest and the greatest[6] of the original ratios; thus

$$\frac{P}{p} = f,$$

where f is between the smallest and largest of the quantities $\frac{R}{OH}, \frac{R}{OK}, \ldots$. Now if the polygon varies in such a way that all the sides decrease indefinitely, all the distances OH tend to R, and all the ratios $\frac{OH}{R}, \frac{OK}{R}$ tend to one. Therefore the same happens for $\frac{P}{p}$, and therefore for $\frac{L}{p}$ and $\frac{P}{L}$ which are both between 1 and $\frac{P}{p}$. Thus P and p tend to L. QED

Definition. The length L, the common limit of the inscribed and circumscribed convex polygons as the sides decrease indefinitely (a limit whose existence has just been proved), is called the circumference or length of the circle.

178. Theorem.

The ratios of the lengths of two circles is the ratio of their radii.

Consider circles C, C_1 with radii R, R_1. In these circles, we inscribe regular polygons with the same number of sides. These polygons are similar, and therefore (146, Corollary) the ratio of their perimeters P, P_1 is the ratio of their radii. As the number of sides increases indefinitely, the ratio $\frac{P}{P_1}$ tends to the ratio $\frac{C}{C_1}$ of the circumferences of the circles. This proves the theorem.

Corollary. The ratio of the length of a circle to the diameter is constant.

Indeed, the proportion $\frac{C}{C_1} = \frac{R}{R_1}$ can be rewritten $\frac{C}{R} = \frac{C_1}{R_1}$ or $\frac{C}{2R} = \frac{C_1}{2R_1}$. Thus the ratio of the circumference to the diameter is the same for any two circles.

This constant number, the ratio of the circumference to the diameter, is denoted by the Greek letter π.

Corollary. The circumference of a circle of radius R is $2\pi R$.

179. Length of an arc of a circle.

Definition. The length of an arc of a circle is the limit of the perimeters of convex broken lines inscribed in the arc, ending at the endpoints of the arc, as all the sides decrease indefinitely.

---

[6] Leçons d Arithmetique, chap. VI, n 212.

The existence of this limit is proved by the same reasoning used for the whole circle. We first consider a regular inscribed or circumscribed broken line, and we double the number of sides indefinitely. Then we pass to an arbitrary broken line.[7]

Every arc of a circle is longer than its chord, since it is the limit of broken lines longer than this chord, whose perimeters increase.

For analogous reasons, it is shorter than any broken line surrounding it, with the same endpoints.[8]

The ratio of an arc of a circle to its chord tends to one as the arc tends to zero (if the circle remains fixed). This is true because the arc $\overset{\frown}{ab}$ (Fig. 169) is contained between its chord and the broken line aAb formed by the tangents at its endpoints: as we have seen (177) the ratio of these quantities goes to one when the arc $\overset{\frown}{ab}$ decreases indefinitely.

179b. Two congruent arcs have equal lengths, since the broken lines used in the definition of their lengths can be taken to be equal in pairs. The sum of two arcs has a length equal to the sum of the lengths of the two arcs, since two broken lines inscribed in the two arcs, taken together, constitute a broken line inscribed in the total arc.

Thus, according to a principle invoked several times before, the ratio of the lengths of two arcs of the same circle is equal to the ratio of the arcs (17). In particular, the ratio of these lengths is equal to the ratio of the measures of the arcs in degrees or gradients.

Since the circumference of a circle of radius R is $2\pi R$, and this circle is divided into 360 degrees, it follows that the length of each degree will be $\frac{2\pi R}{360} = \frac{\pi R}{180}$. The length of a minute will be equal to one sixtieth of this quantity, or $\frac{\pi R}{180 \times 60}$, and a second $\frac{\pi R}{180 \times 60 \times 60}$. Thus, on a circle with radius R, an arc of m degrees, n minutes,

---

[7] This method of proof can be extended to any convex curve A; that is, to a curve such that none of its arcs intersects its chord. It suffices to show that to every number $m < 1$ (but as close to 1 as we want) we can associate a length ε such that for any arc $\overset{\frown}{ab}$ contained in A whose endpoints are closer to each other than ε, the ratio of the chord ab to the sum $ac + bc$ (see figure) of the tangents at a, b (bounded by their points of intersection) is contained between m and 1. This condition is satisfied (as is proved in differential calculus) under certain assumptions which are always satisfied in the case of the curves we consider. We are then assured (as in 177), that the ratio of the perimeter of an inscribed broken line to the perimeter of the corresponding circumscribed line tends to one, as the number of sides increases and each of them tends to zero.

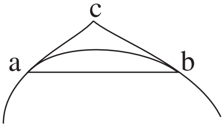

This can be used to prove the existence of a limit so long as the vertices of each broken line are included among the vertices of the next one (by the argument in 176). Then we pass to the general case as in 177.

Finally, a non-convex arc can often be broken into convex pieces, allowing us to define its length.

[8] These conclusions extend to any arc of a curve defined in accordance with the considerations of the preceding note. In particular, the line segment is the shortest path from one point to another. A convex arc (see previous note) is shorter than any line surrounding it (a broken line or a curve) with the same endpoints.

p seconds, has length

$$\frac{R}{180}\quad m+\frac{n}{60}+\frac{p}{60\times 60}\quad ;$$

an arc of n gradients will have length $\frac{Rn}{200}$.

180. Thus we see that we can calculate the length of any arc of a given circle, if we know the number .

Calculation of . Method of perimeters. To calculate this number, or equivalently, to calculate the length of a circle of a given radius R, we have to calculate the perimeters of regular inscribed polygons whose number of sides doubles indefinitely. Each of these perimeters will furnish an estimate from below of the required length, and this approximation gets close as the number of sides increases. If, at the same time that we calculate the perimeter of the inscribed polygon, we can calculate the perimeter of the corresponding circumscribed polygon, we will obtain an estimate from above, and the difference of these two estimates will give an upper bound to the error committed by adopting one of them.

We have learned to calculate the sides of a number of regular inscribed polygons, for instance, that of a square. To deduce from this the sides of the polygons with $4\times 2, 4\times 2^2, \ldots, 4\times 2^n$ sides, it will suffice to solve the following problem:

Problem. Knowing the side c of a regular polygon inscribed in a circle with given radius R, calculate the side of the polygon inscribed in the same circle, with twice the number of sides.

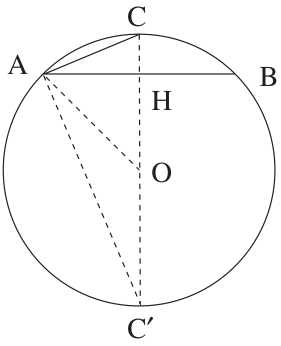

Figure 170

Let AB = c (Fig. 170) be the side of the first polygon inscribed in a circle with radius OA = R. Dividing arc $\widehat{AB}$ into two equal parts with radius OC, perpendicular to the midpoint H of AB, we obtain side AC of the inscribed polygon with twice the number of sides. Now in triangle OAC we have:

$$AC^2 = OA^2 + OC^2 \quad 2OC\bullet OH = 2R^2 \quad 2R\bullet OH = 2R(R \quad OH),$$

where OH equals $\overline{R^2 \quad \frac{c^2}{4}}$. Thus the required side c_1 will be

$$c_1 = \overline{2R \quad R \quad \overline{R^2 \quad \frac{c^2}{4}}}\ .$$

Thus, since the side of the square is $R\sqrt{2}$ and its perimeter $R \times 4\sqrt{2}$, the side of the octagon and its perimeter will be

$$R\sqrt{2-\sqrt{2}}, \quad R \times 8\sqrt{2-\sqrt{2}}.$$

The side and perimeter of the regular inscribed polygon with sixteen sides will be

$$R\sqrt{2-\sqrt{2+\sqrt{2}}}, \quad R \times 16\sqrt{2-\sqrt{2+\sqrt{2}}},$$

and so on.

180b. In order to get an upper estimate for the length of the circle, we need to calculate the perimeter of a circumscribed regular polygon, and we must therefore solve the following problem:

Problem. Knowing the side c of a regular polygon inscribed in a circle of radius R, calculate the side of the regular polygon with the same number of sides circumscribed about the same circle.

The required side c' is found immediately by noting that the two polygons are similar, and the ratio of their sides is the ratio of their apothems. Since the apothem of the circumscribed polygon is R, and that of the inscribed polygon $\sqrt{R^2 - \frac{c^2}{4}}$, the required side is given by the proportion

$$\frac{c'}{c} = \frac{R}{\sqrt{R^2 - \frac{c^2}{4}}}.$$

We have just learned how to calculate the sides of the inscribed polygons with large numbers of sides: the preceding proportion gives us the sides of the corresponding circumscribed polygons.

181. Knowing the side c of a regular polygon inscribed in a circle of radius R, we can also calculate the side c_1 and the apothem a_1 of the regular polygon with twice the number of sides inscribed in the same circle (which gives us the side of the circumscribed polygon without an additional square root) by the following method.

In Fig. 170, let C' be the point where the extension of radius OC intersects the circle. Triangle ACC' is a right triangle and gives us (123, Remark)

$$AC \cdot AC' = CC' \cdot AH = 2R \times \frac{c}{2} = Rc.$$

On the other hand,

$$AC^2 + AC'^2 = CC'^2 = 4R^2.$$

Multiplying the first equation by 2 and adding it to the second, we get:

$$AC^2 + AC'^2 + 2AC \cdot AC' = (AC + AC')^2 = 4R^2 + 2Rc.$$

Thus

$$AC + AC' = \sqrt{4R^2 + 2Rc} = 2\sqrt{R\left(R + \frac{c}{2}\right)}.$$

If we multiply the first equation by 2, then subtract it from the second, we have:

$$AC^2 + AC'^2 - 2AC \cdot AC' = (AC' - AC)^2 = 4R^2 - 2Rc.$$

Thus

$$AC' - AC = \sqrt{4R^2 - 2Rc} = 2\sqrt{R\left(R - \frac{c}{2}\right)}.$$

Adding and subtracting these equalities member by member, these two equations give us

$$AC = \sqrt{R\left(R + \frac{c}{2}\right)} - \sqrt{R\left(R - \frac{c}{2}\right)}; \quad AC' = \sqrt{R\left(R + \frac{c}{2}\right)} + \sqrt{R\left(R - \frac{c}{2}\right)},$$

which solves the problem, since AC is equal to the required chord c_1 and AC' is twice the apothem a_1.

We have thus learned how to calculate the chords of half of the two arcs intercepted by a given chord on a circle of given radius.

182. The isoperimetric method. We can present the preceding method in a slightly different form.

In effect, the problem consist in calculating the ratio of the perimeter p of an inscribed regular polygon to the radius r of the circumscribed circle, or to the apothem a, which is the radius of the inscribed circle, and to do this for polygons whose number of sides increases indefinitely. But it makes no difference whether or not these polygons are inscribed in the same circle, since the ratios $\frac{p}{r}$ and $\frac{p}{a}$ depend only on the number of sides of the polygon (163).

In the isoperimetric method we consider regular polygons whose number of sides doubles indefinitely, and which have the same perimeter. We must therefore solve the following problem:

Problem. Knowing the radius r and apothem a of a regular polygon, calculate the radius r' and apothem a' of a regular polygon with twice as many sides, isoperimetric to the first.

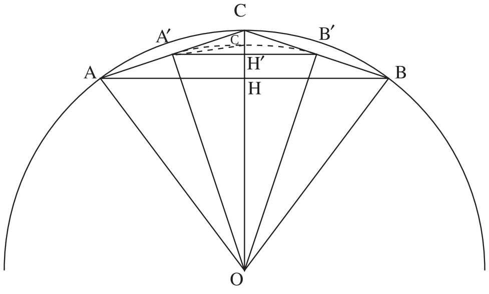

Figure 171

Let AB (Fig. 171) be a side of the regular polygon with n sides inscribed in a circle with radius $OA = r$. We drop perpendicular $OH = a$ to AB, and extend it to its intersection C with the circumscribed circle, which is also the midpoint of arc $\widehat{AB}$. We draw AC and BC, and let A', B' be the midpoints of these segments. Let $A'B'$ intersects OC in a point H'.

Segment $A'B'$ is the side of a regular polygon with $2n$ sides inscribed in the circle with radius OA'.

Indeed, since angle AOB intercepts an arc equal to the n^{th} part of the circle with radius OA, angle $A'OB'$ which is half of it (because $A'OC$ is half of AOC and

B OC half of BOC) will intercept an arc equal to the 2nth part of the circle with radius OA .

This polygon is isoperimetric to the first.

This is true because side A B joins the midpoints of AC, BC, and is therefore half of the original side AB, and because the new polygon has twice as many sides.

The required lengths will thus be r = OA , a = OH .

Now H is the midpoint of CH, so we have

$$OH \quad OH = OC \quad OH .$$

This can be written as

$$2OH = OC + OH,$$

or

$$OH = \frac{OC + OH}{2}.$$

In other words, OH is the arithmetic mean of OC and OH.

On the other hand, right triangle OA C gives

$$OA = \overline{OC \times OH} .$$

That is, OA is the geometric mean of OC and OH .

Thus we can calculate a by the formula

$$a = \frac{r + a}{2},$$

and r by the formula

$$r = \overline{ra} .$$

Repeating this operation of doubling, we obtain values of r and a for polygons with larger and larger numbers of sides. The ratios of these quantities to the common semi-perimeter of these polygons will give approximate values for $\frac{1}{}$, one less than the limit, the other greater, and both growing closer.

Suppose, to simplify the situation, that the common perimeter of the polygons is equal to two units of length, and take a square as the initial polygon. The apothem is half the side, or $\frac{1}{4}$, and the radius is equal to the side divided by $\overline{2}$, or $\frac{1}{2\ 2}$. Taking a = $\frac{1}{4}$, r = $\frac{1}{2\ 2}$ in the preceding formulas will give the values of a , r for the octagon, and so on.

But if we take[9] a = 0, r = $\frac{1}{2}$, our formulas yield precisely a = $\frac{1}{4}$, r = $\frac{1}{2\ 2}$. We then obtain the following proposition (a theorem of Schwab).

Theorem. If we form a sequence of numbers starting with 0 and $\frac{1}{2}$, and each successive term is alternately the arithmetic mean or the geometric mean of the preceding two terms, then the numbers in this sequence tend to $\frac{1}{}$.

183. Since the numbers a and r, consecutive terms of the preceding sequence, approach the limit $\frac{1}{}$ from below and from above respectively, the error committed in taking either for an approximate value of $\frac{1}{}$ is less than r a.

---

[9] These numbers correspond to the division of the circle into two equal parts. If the division points are A and B, diameter AB can be thought of as the side of a regular polygon with two sides, whose perimeter is 2AB. If this perimeter is 2, the radius must be $\frac{1}{2}$ and the apothem (the distance from the side to the center) is zero.

Also we have

$$r \quad a < \frac{r \quad a}{4}.$$

To prove this, we revisit Figure 171 and lay off on OC a length OC = OA = r . Segment C H can represent r a . The two angles C A H , C A C are equal, since on the circle OA they are measured by one half the equal arcs $\overset{\frown}{C\ B}$, $\overset{\frown}{A\ C}$ (one is an inscribed angle, the other is formed by a tangent and a secant). The angle bisector theorem (see 115) then shows that the ratio of H C to C C is equal to the ratio of H A to A C. Thus H C is less than half of H C, which is equal to a quarter of CH; that is, to $\frac{r \quad a}{4}$.

The first two terms in the Schwab sequence give r a = $\frac{1}{2}$. Thus, replacing $\frac{1}{}$ by the 2nth term in this sequence, we commit an error less than

$$\frac{1}{2 \times 4^n} = \frac{1}{2^{2n+1}}.$$

184. The Greek geometer Archimedes, who was the first to define and calculate the length of the circle, used the method of perimeters to find the value for

$$\frac{22}{7}$$

which is greater than by less than one hundredth.

Note that one can obtain only approximate values for , and never the exact value, because it can be proved that this number is irrational; that is, it is not an integer or a ratio of two integers.

The first decimals of and $\frac{1}{}$ are[10]

$$= 3.141592653\ldots, \qquad \frac{1}{} = 0.3183098\ldots.$$

We often use the value 3.1416, which is slightly too large, for .

The famous problem of the quadrature of the circle amounts (as we will see later) to constructing a segment equal to the length of a circle with given radius. It is impossible to solve this problem using only straightedge and compass. This impossibility, which does not necessarily follow[11] from the irrationality of , has also been proved.[12]

---

[10] These values have not been obtained by the methods just explained, but by incomparably faster ones borrowed from the differential calculus. One can easily remember these decimals, using the following verse: Que j aime á faire apprendre un nombre utile aux sages in which the number of letters of the words give the consecutive digits of . (This French mnemonic given above has various analogues in English; for example, May I have a large container of coffee? transl.)

The value $\frac{355}{113}$, greater than by less than one millionth, was given by Adrien Metius.

[11] For instance, the ratio of the diagonal of a square to its side is irrational, but one can easily construct the diagonal knowing the side.

[12] The irrationality of was proved by Lambert (1770). The impossibility of the quadrature of the circle is due to Lindemann, who proved it in 1882, by generalizing results of the great French mathematician Hermite.

CHAPTER II

Transversal

192. Theorem. If sides BC, CA, AB of a triangle are intersected by a line at a, b, c (Fig. 174) we have the following relation among the segments determined on the sides:

$$\frac{aB}{aC} \bullet \frac{bC}{bA} \bullet \frac{cA}{cB} = 1. \qquad (2)$$

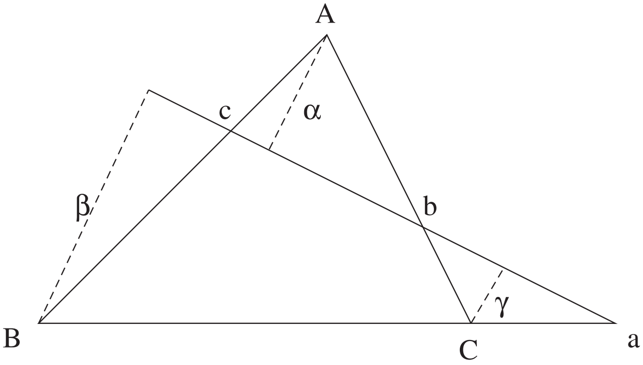

Figure 174

To prove this, we draw three segments parallel to an arbitrary direction through the vertices of the triangle , on which we choose the same positive sense, and extend them to intersect the transversal. Let , , be the distances from the vertices to the transversal, measured along these parallel lines. We have, in magnitude and sign,

$$\frac{aB}{aC} = \frac{}{},$$
$$\frac{bC}{bA} = \frac{}{},$$
$$\frac{cA}{cB} = \frac{}{}.$$

Multiplying, we obtain $\frac{aB}{aC} \bullet \frac{bC}{bA} \bullet \frac{cA}{cB} = \frac{}{} = 1.$[1] QED

---

[1] If the transversal is parallel to BC, the point a should be considered as thrown to infinity, and the ratio $\frac{aB}{aC}$ taken equal to one. The relation in the statement then reduces to $\frac{bA}{bC} = \frac{cA}{cB}$; that is, to the theorem of 114. If sides AB, AC of the triangle are parallel, and point A is at

193. Converse. Theorem of Menelaus.[2] If we choose three points a, b, c on sides BA, CA, AB of triangle ABC which satisfy the the relation

$$\frac{aB}{aC} \cdot \frac{bC}{bA} \cdot \frac{cA}{cB} = 1, \tag{2}$$

then these three points lie on a straight line.

This is true because line ab intersects side AB in a point c such that

$$\frac{aB}{aC} \cdot \frac{bC}{bA} \cdot \frac{c\,A}{c\,B} = 1.$$

Comparing this relation with relation (2), we see that

$$\frac{cA}{cB} = \frac{c\,A}{c\,B},$$

and therefore the points c and c coincide. QED

Remark. This theorem is actually equivalent to the one proved in 144. Indeed, we can find three lengths , , such that we have (in magnitude and sign):

$$\frac{aB}{aC} = \frac{}{}, \quad \frac{bC}{bA} = \frac{}{},$$

which, together with (2), imply

$$\frac{cA}{cB} = \frac{}{}.$$

Therefore, if we draw three homothetic figures in which A, B, C are three corresponding points, and , , three corresponding segments, then their centers of similarity are a, b, c.

194. This theorem is useful in showing that three points are collinear.

Example I. The midpoints of the three diagonals of a complete quadrilateral (see 145) lie on a straight line.

Consider the complete quadrilateral ABCDEF (Fig. 175), whose diagonals AB, CD, EF have midpoints L, M, N. Consider the triangle ACE formed by three sides of the quadrilateral, and let a, c, e be the midpoints of CE, EA, AC. Line ce is parallel to CE and passes through point L; line ea is parallel to EA and passes through M; line ac is parallel to AC and passes through N. To show that L, M, N are on a straight line it suffices to prove the relation

$$\frac{Lc}{Le} \cdot \frac{Me}{Ma} \cdot \frac{Na}{Nc} = 1.$$

Parallel lines Lec, BCE give

$$\frac{Lc}{Le} = \frac{BE}{BC},$$

and likewise we obtain

$$\frac{Me}{Ma} = \frac{DA}{DE}, \quad \frac{Na}{Nc} = \frac{FC}{FA}.$$

---

infinity, the expression $\frac{aB}{aC} \cdot \frac{bC}{bA} \cdot \frac{cA}{cB}$ can be rewritten in the form

$$\frac{aB}{aC} \cdot \frac{bC}{cB} \cdot \frac{cA}{bA},$$

and we can replace $\frac{cA}{bA}$ by 1. We then obtain the statement in 117.

[2] The name Theorem of Menelaus in more recent sources refers variously to this theorem, to its converse in 192, or to both statements taken together. transl.

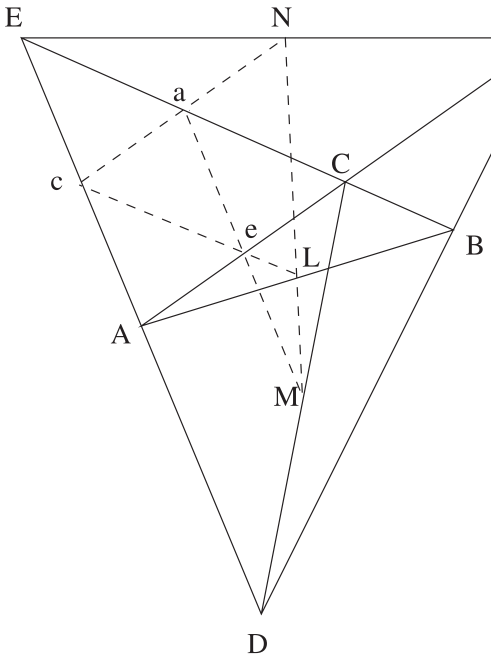

Figure 175

The product of the right hand sides

$$\frac{BE}{BC} \cdot \frac{DA}{DE} \cdot \frac{FC}{FA}$$

is indeed 1 since points B, D, F, taken on the sides of triangle ACE, lie on a straight line.

195. Example II. If the vertices of triangles abc, a b c are situated so that the lines aa , bb , cc , joining corresponding vertices, intersect in a point o, then the intersection points of the corresponding sides are on a straight line.

Consider the intersections of bc, b c , m of ca, c a , and n of ab, a b (Fig. 176). To show that these points are on a straight line, we must verify that

$$\frac{b}{c} \cdot \frac{mc}{ma} \cdot \frac{na}{nb} = 1.$$

Now triangle abc is cut by transversal b c , giving us

$$\frac{b}{c} \cdot \frac{c\,c}{c\,o} \cdot \frac{b\,o}{b\,b} = 1.$$

Likewise, triangles oca, oab cut, respectively, by mc a , na b give

$$\frac{ma}{mc} \cdot \frac{a\,a}{a\,o} \cdot \frac{c\,o}{c\,c} = 1, \quad \frac{na}{nb} \cdot \frac{b\,b}{b\,o} \cdot \frac{a\,o}{a\,a} = 1.$$

Multiplying these equalities, the factors a a, b b, c c, a o, b o, c o cancel out, and we obtain the required relation.

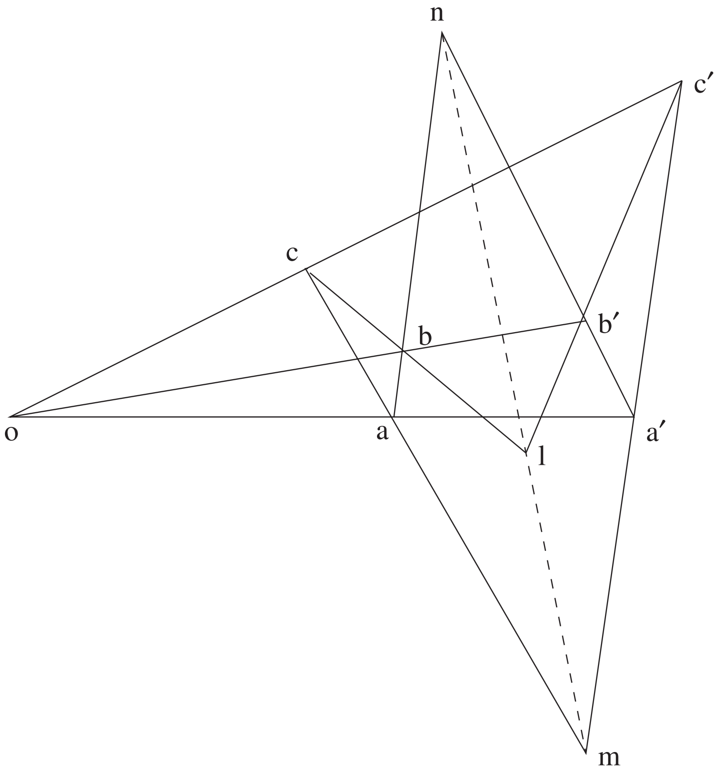

Figure 176

Two triangles, situated so that the lines joining the corresponding vertices are concurrent, are said to be homologic.[3]

196. Example III (Pascal s Theorem). In a hexagon inscribed in a circle, the intersection points of opposite sides are on a straight line.

Suppose the inscribed hexagon is ABCDEF (Fig. 177), and that opposite sides AB, DE intersect in L, BC, EF meet in M, and CD, FA meet in N. Points L, M, N are on sides JK, KI, IJ of the triangle IJK formed by the lines AB, CD, EF. They will be on a straight line if

$$\frac{LJ}{LK}\cdot\frac{MK}{MI}\cdot\frac{NI}{NJ}=1. \tag{3}$$

Cutting the triangle IJK by the remaining sides DE, BC, FA of the hexagon, we obtain

$$\frac{LJ}{LK}\cdot\frac{EK}{EI}\cdot\frac{DI}{DJ}=1,\quad \frac{MK}{MI}\cdot\frac{CI}{CJ}\cdot\frac{BJ}{BK}=1,\quad \frac{NI}{NJ}\cdot\frac{AJ}{AK}\cdot\frac{FK}{FI}=1.$$

We now multiply these relations, grouping the factors in the numerator and denominator appropriately, to obtain

$$\frac{LJ}{LK}\cdot\frac{MK}{MI}\cdot\frac{NI}{NJ}\cdot\frac{CI\cdot DI}{EI\cdot FI}\cdot\frac{AJ\cdot BJ}{CJ\cdot DJ}\cdot\frac{EK\cdot FK}{AK\cdot BK}=1.$$

---

[3] The more modern terminology for such triangles is perspective. transl.

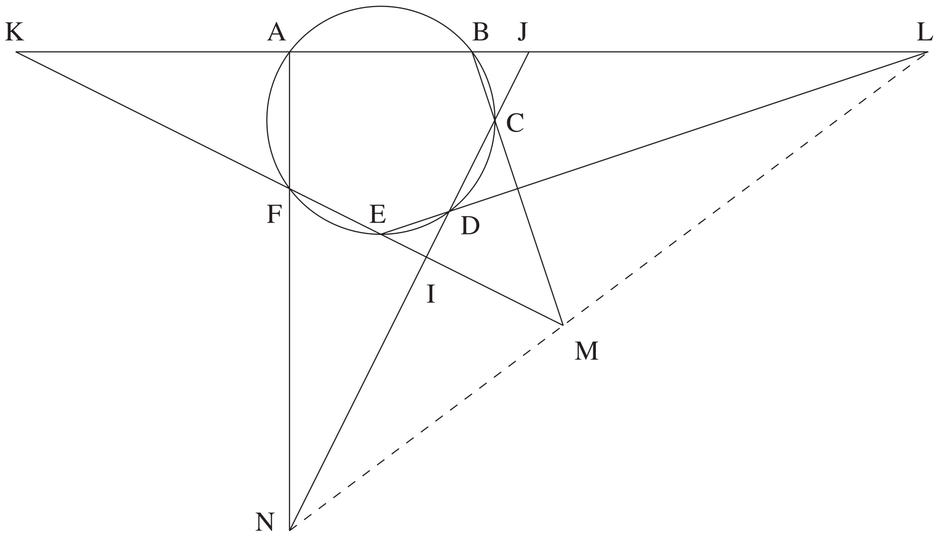

Figure 177

Each of the last three fractions is equal to one; for instance, the products $CI \cdot DI$ and $EI \cdot FI$ are both equal to the power of I with respect to the circle.

We conclude that (3) is true, and the theorem is proved.

Remark. The preceding proof remains valid if the pairs of points A, B; C, D; E, F are made to coincide, so that the sides of the triangle IJK become tangents to the circle. The theorem becomes the following statement: The tangents to the circumscribed circle at the vertices of a triangle intersect the corresponding sides in three collinear points.

197. Theorem. Through the vertices A, B, C of a triangle, we draw segments Aa, Bb, Cc concurrent at a point O, where a, b, c are on the corresponding sides BC, CA, AB. Then:

$$\frac{aB}{aC} \cdot \frac{bC}{bA} \cdot \frac{cA}{cB} = 1. \tag{4}$$

Indeed, triangle AaC (Fig. 178), cut by transversal BOb gives us[4]

$$\frac{Ba}{BC} \cdot \frac{bC}{bA} \cdot \frac{OA}{Oa} = 1;$$

triangle AaB, cut by transversal Cc gives

$$\frac{Ca}{CB} \cdot \frac{cB}{cA} \cdot \frac{OA}{Oa} = 1.$$

We divide these two equalities; the ratio $\frac{OA}{Oa}$ cancels out, and we obtain

$$\frac{aB}{aC} \cdot \frac{bC}{bA} \cdot \frac{cA}{cB} \cdot \frac{CB}{BC} = 1,$$

---

[4] This reasoning remains correct when point O is thrown to infinity; that is, when lines Aa, Bb, Cc are parallel. Indeed, we have seen that the theorem in 192 is true when a transversal is parallel to one of the sides of the triangle.

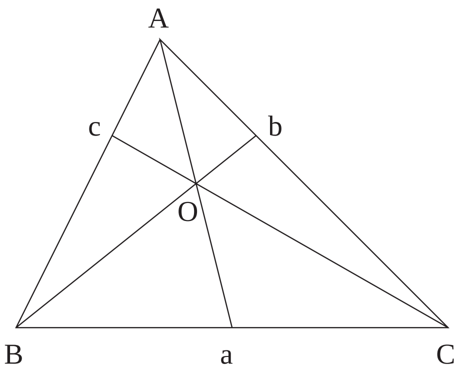

Figure 178

which is equivalent to relation (4), because $CB = BC$.

198. Converse (Theorem of Ceva[5]). If we take points a, b, c on the sides BC, CA, AB of a triangle, such that

$$\frac{aB}{aC} \cdot \frac{bC}{bA} \cdot \frac{cA}{cB} = 1,$$

then the lines Aa, Bb, Cc are concurrent.

Indeed, consider the intersection O of Aa and Bb. The line CO intersects AB in a point c such that

$$\frac{aB}{aC} \cdot \frac{bC}{bA} \cdot \frac{c\ A}{c\ B} = 1.$$

This, combined with the hypothesis, gives $\frac{cA}{cB} = \frac{c\ A}{c\ B}$, so that the points c, c coincide. QED

If the lines Aa, Bb were parallel, the line Cc would be parallel to them, and the three lines must be considered as concurrent at infinity.

This result is useful in proving that three lines are concurrent.

Example. The medians of a triangle are concurrent.

This is true because, if a, b, c denote the midpoints of the three sides of the triangle, each of the ratios $\frac{aB}{aC}, \frac{bC}{bA}, \frac{cA}{cB}$ is equal to 1.

One can prove analogously that the angle bisectors of a triangle are concurrent, etc.

Exercises

Exercise 223. Deduce the theorem of Menelaus, in the case for which the division points are external, from the fact that the three circles which determine the required point in Exercise 127 have the same intersection (see 189).

---

[5] The term Ceva s Theorem in recent sources, can refer to this theorem, to its converse, or to both statements taken together. See the footnote to 193. transl.

Exercise 224. A line intersects the sides of a triangle ABC in three points a, b, c. Take the point symmetric to each of these points with respect to the midpoint of the side which contains it. Show that the points a′, b′, c′ obtained in this way are also collinear.

If the points a, b, c are the projections of a point on the circumscribed circle onto the three sides (Exercise 72), the same is true for a′, b′, c′.

Exercise 225. One side OAB of an angle is fixed, as well as the two points A, B on it. The other side, OA′B′, turns around O, carrying with it two points A′, B′ at fixed distances from O. What is the locus of the intersection of AA′ and BB′?

Exercise 226. What does Ceva's theorem (198) give us when applied to the altitudes of a triangle?

Exercise 227. Three concurrent lines through the vertices of triangle ABC intersect the opposite sides in a, b, c. We take points a′, b′, c′ symmetric to these points with respect to the midpoints of the sides which contain them. Show that Aa′, Bb′, Cc′ are also concurrent.

Exercise 228. The lines joining the vertices of a triangle to the points of contact of the opposite sides with the inscribed circle are concurrent.

Exercise 229. Consider triangle A′B′C′ obtained (as in 53) from triangle ABC by drawing, through each vertex, a parallel to the opposite side. Let a, b, c be points on BC, CA, AB, respectively.

If Aa, Bb, Cc are concurrent, so are A′a, B′b, C′c.

Exercise 230. Suppose lines Aa, Bb, Cc are concurrent. The harmonic conjugates of a, b, c with respect to the sides BC, CA, AB, on which they are situated, are on a straight line.

Application: Aa, Bb, Cc are the angle bisectors.

Exercise 231. If lines Aa, Bb, Cc, bounded by the sides BC, CA, AB of triangle ABC are concurrent, and if circle abc intersects the sides again at a′, b′, c′, then lines Aa′, Bb′, Cc′ are also concurrent.

CHAPTER V

Inverse Figures

214. The inverse of a point M with respect to some point O (called the pole of inversion, Figures 186, 187) is a point M' on line OM such that:

$$OM \cdot OM' = k,$$

for some constant k, called the power of inversion. Point M' is also called the transform of M with respect to reciprical radius vectors.

As usual, the segment[1] OM' is taken in the direction of OM (Fig. 186) or in the opposite direction (Fig. 187) according as k is positive or negative.

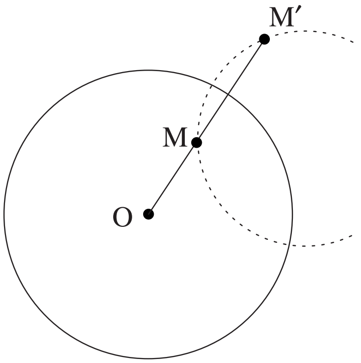

Figure 186

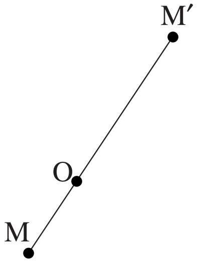

Figure 187

We see immediately that:

1°. Every point has an inverse except for the pole, whose inverse is thrown to infinity.

2°. If M' is the inverse of M then, reciprocally, M is the inverse of M'.

The figure F' formed by the inverses of all the points in a figure F is called the inverse figure of F.

215. Theorem. Two figures, both inverse to a third figure F with respect to the same pole O, are homothetic to each other.

Indeed, let M be a point of F; let M', M'_1 be the points which correspond to it under the two inversions with common pole O and with powers k, k_1. Division of the equation

$$OM \cdot OM' = k$$

---

[1] The segments OM, OM' are called radius vectors of the points M, M'.

by the equation

$$OM \cdot OM_1 = k$$

gives the equation $\frac{OM_1}{OM} = \frac{k_1}{k}$. QED

Thus the power of the inversion does not influence the shape of the figure obtained. This shape is changed only by changing the pole of inversion.

216. When the power of inversion k is positive, we can construct a circle centered at the pole of inversion and with radius $\sqrt{k}$ (Fig. 186). This circle, which clearly suffices to define the inversion, is called the circle of inversion. It is the locus of points which coincide with their inverses.

Two inverse points are conjugate with respect to the circle of inversion, and belong to the same diameter.

Every circle passing through two inverse points intersects the circle of inversion at a right angle (135). Conversely, if two points M, M′ have the property that every circle passing through them is orthogonal to a given circle, the two points are inverses with respect to this circle (135–137).

When two points are symmetric with respect to a line, any circle passing through these points cuts the line at right angles, since the line is a diameter of the circle. We are thus led to consider symmetry with respect to a line as a limiting case of inversion, the circle of inversion being replaced by the line and, consequently, the pole thrown to infinity.

217. Any two points A, B and their inverses A′, B′ are on the same circle (131b).

It follows (82) that the angle made by AB with the line OAA′ is equal to the angle made by A′B′ with OBB′, but in the opposite sense of rotation. Thus if we know the direction of AB and of the radius vectors, we also know the direction of A′B′.

This direction A′B′ is said to be antiparallel to AB with respect to angle AOB. It is parallel to the reflection of AB in the bisector of this angle since, in this reflection, the line OA becomes OB and angle BAO becomes an angle which is equal, but in the opposite sense.

218. **Problem.** Knowing the distance between two points A, B, and their radius vectors, find the distance between their inverses A′, B′.

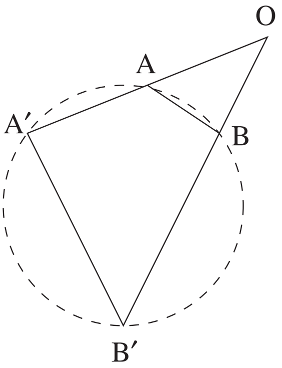

Figure 188

Similar triangles OAB, OB A (Fig. 188) give us $\frac{A\ B}{OA} = \frac{BA}{OB}$.

Solving this equation for A B , then replacing OA by $\frac{k}{OA}$, we obtain A B = BA $\cdot\, \frac{k}{OA \cdot OB}$.

Remark. This argument does not work when A, B, O are collinear, but the result is still true, even in magnitude and sign (if a positive direction has been chosen on the common radius vector). This is easily seen if we replace A B and AB by OB OA = $\frac{k}{OB}$ $\frac{k}{OA}$ and OA OB.

219. **Theorem.** The tangents at corresponding points of two inverse curves make equal angles, but in opposite senses, with the common radius vector of their points of contact.

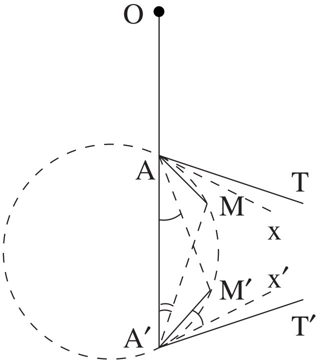

Figure 189

Let A (Fig. 189) be a point on curve C, and A the corresponding point on the inverse C . Consider a point M on C, close to A, and let M be its inverse. We draw tangents Ax, A x to the circle passing through A, A , M, M (217). When M approaches A as close as we like (and, consequently, M approaches A) angles AA M, A AM will tend to zero, and the same will be true for angles MAx, M A x , which are equal to the first two. Therefore, if the line AM tends to a limiting position AT, then Ax will tend to the same limiting position. Therefore line A x will have a limiting position A T which makes the same angle as AT with line OAA , but in the opposite sense. Since angle M A x tends to zero, A M will also tend to A T . QED

Corollary. The tangents at corresponding points of two inverse curves are symmetric with respect to the perpendicular bisector of their points of contact.

Remark. The preceding results are also obviously true if the inversion becomes a line reflection.

Theorem. Two curves intersect at the same angle as their inverses (or their symmetric curves), except for the sense of rotation.

Indeed, the angle formed by the tangents to the curves at their common point A, and the angle formed by the tangents to the inverse curves at the corresponding point A , are symmetric with respect to the perpendicular bisector of AA .

Corollary. If two curves are tangent, the same is true for their inverses.

220. Inverse of a line.

Theorem. The inverse figure of a line is a circle passing through the pole of inversion.

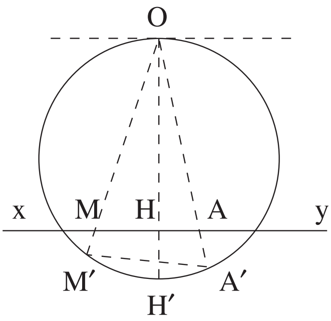

Figure 190

Let xy (Fig. 190) be a line whose inverse we seek, and let O be the pole of the inversion. Take any point A,[2] and let A' be its inverse. Next take a variable point M on xy and its inverse M'. The angle between $M'A'$ and OM' is equal to the angle between AM and OA, and in the opposite sense. Therefore this angle is constant as point M varies along the line, and the locus of M' is a circle passing through O and A. QED

In particular, we can choose, for point A, the projection H of O onto xy. Let H' be its inverse. Angle $OM'H'$ is therefore a right angle, and we see that the tangent at O of the inverse circle is parallel to the given line, and its diameter is $\frac{k}{\ }$, where k is the power of the inversion, and is the distance OH from the pole to the line.

Remark. The preceding statement is valid for lines which do not pass through the pole. A line passing through the pole is its own inverse.

Corollary. The inverse figure of a circle passing through the pole is a line; namely, the line perpendicular to the diameter of the pole at the inverse point of the point diametrically opposite to the pole.

221. The inverse of an arbitrary circle.

Theorem. The inverse of a circle not passing through the pole is a circle.

1 . First, if the power of inversion k is equal to the power p of the pole with respect to the given circle, this circle is its own inverse, as the intersection points of any secant from the pole are inverses of each other.

2 . If the power k is arbitrary (Fig. 191), the inverse figure will be a circle homothetic to the one considered above (215) with the pole as center of similarity and ratio of similarity $\frac{k}{p}$.

---

[2] i.e. on line xy. transl.

222. Conversely, two arbitrary circles C, C′ (Fig. 191) can be considered as inverses of each other in two different ways, since they are homothetic in two different ways. The pole of inversion is then a center of similarity, and the power of inversion is equal to the power of this pole with respect to the first circle, multiplied by the ratio of similarity.

These similarities are the only two which transform one of these circles into the other. Indeed, if they are inverse with respect to some pole S, the two circles must be homothetic with respect to S, since one of them is its own inverse with respect to this pole.

223. By the preceding argument, if a secant from a center of similarity S intersects C in M, N, and C′ in M′, N′, so that M′ corresponds to M and N′ corresponds to N under a homothecy beween the two circles, then M and N′ on the one hand, and M′ and N on the other, are pairs of inverse points.

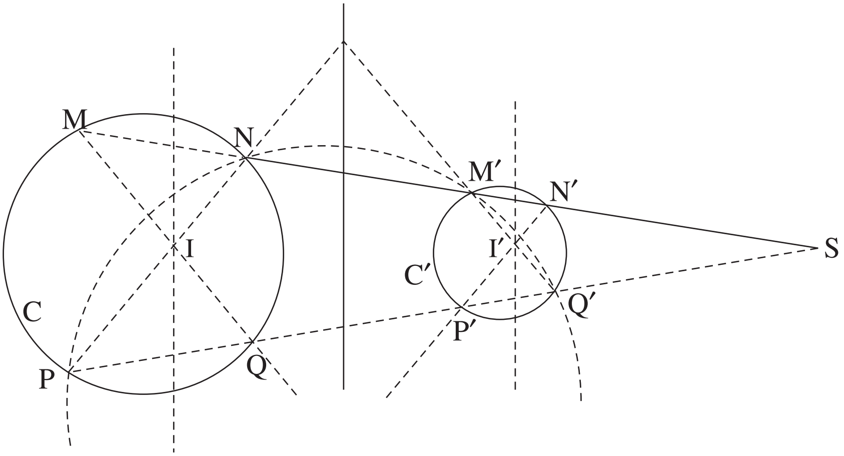

Figure 191

The points M and N′ are sometimes called antihomologous points.

The only pairs of points which are both homologous and antihomologous are the points of contact of the common tangents from point S.

The common points of two circles, if any exist, are their own anti-homologues.

The antihomologous chord of a chord given by two points in the first circle, is the chord in the second circle determined by the anti-homologues of the first two points.

224. Two pairs of antihomologous points lie on the same circle (217) and consequently two antihomologous chords intersect on the radical axis (139).

Moreover, two antihomologous chords intersect the polars of the center of similarity, with respect to their respective circles, in two corresponding points. This is true because the chord N′P′, antihomologous to MQ (Fig. 191), corresponds to NP, which intersects MQ in a point I situated on the polar of S with respect to circle C (211) and whose corresponding point I′ is the intersection of N′P′ with the polar of S with respect to C′.

This double property allows us to determine the anti-homologue of a given chord without involving the endpoints of the chord.

This applies, of course, to the tangents at antihomologous points, which are the limiting cases of the preceding chords.

Remark. If the two circles are identical (case 1 of 221), the two chords MQ, NP, joining two points to their inverses, intersect on the polar of the pole S of the inversion. From this point of view, this polar plays the role of the radical axis of the two identical circles, if they are considered as inverses of each other with respect to the point S.

225. If the two circles are congruent, the exterior center of similarity is thrown to infinity, and the corresponding homothecy reduces to a translation, while the corresponding inversion reduces to a line reflection; the properties of antihomologous points are the same as in the general case.

226. A line and a circle can also be regarded as inverse figures in two different ways. The poles of inversion will be (220) the endpoints of the diameter perpendicular to the line. The theory of antihomologous points leads us to consider these as centers of similarity of the line and the circle.

Of two antihomologous chords, one is on the given line, and therefore we can still say that these two chords intersect on the radical axis, in the sense of 158 (Construction 12, remark).

Finally, two lines are symmetric to each other in two different ways, the axes of symmetry being the bisectors of the angles formed by these lines. The antihomologous chords are the lines themselves, and therefore intersect in a fixed point.

227. When two circles are inverse to each other, every circle which passes through two antihomologous points is its own inverse, since the power of the pole with respect to is equal to the power of inversion. Such a circle must therefore intersect the two circles at the same angle and, if it is tangent to one, it must be tangent to the other.

Conversely, any circle which intersects two circles C, C at the same angle (Fig. 192) intersects them in two pairs of antihomologous points.

Indeed, let A, B; A , B be the four points of intersection, labeled so that the angles formed by circles C, at A, and of the circles C and at A (which are equal by hypothesis) are in opposite senses, and so that the same is true for B, B . If S is the intersection of AA and BB , and k is the power of S with respect to circle , then the inversion with pole S and power k does not change : it takes A onto A and B onto B . Thus this inversion transforms C into a circle which passes through A , B , and is tangent to C at A (by hypothesis, and 219); that is, this circle coincides with C (90).

This conclusion remains true if is tangent to C and C . To see this, it suffices to repeat the preceding argument while taking S to be the intersection of AA and BB , with A and A being the points of contact of , and B, B the second intersections of C, C with any circle passing through A, A (Fig. 192).

In this case, the proposition also follows from the theorem of 144, because the points of contact are centers of similarity, one for and C, the other for and C . Therefore they are collinear with a center S of similarity of C and C . We also see that the center of similarity is external if the points of contact are of the same kind[3], and internal otherwise.

---

[3] That is, both external or both internal.

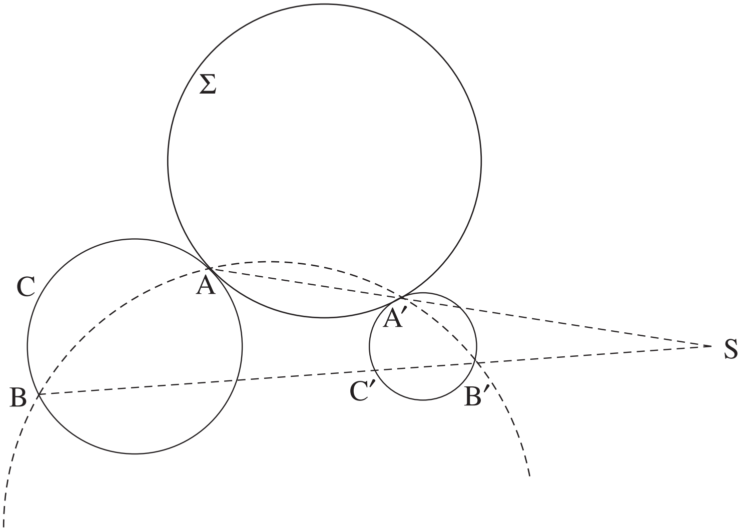

Figure 192

Of course, the arguments just presented remain valid when one or both of the circles C, C , or circle , are replaced by lines.

227b. If circle cuts circles C, C at right angles, we can consider any pair of the points of intersection, chosen arbitrarily, as antihomologous. Indeed, two right angles can always be considered equal and of opposite senses (by replacing, if necessary, one side by its extension). Thus, a circle which cuts C, C at right angles corresponds to itself in either of the two inversions which transform C into C .

228. If the inversion with pole S interchanges circles C and C and it has a circle of inversion , this circle cuts at right angles, since the power of point S with respect to is equal to the power of the inversion.

This property corresponds to the property of the bisectors of the angles formed by two lines. Indeed, these bisectors are the locus of the centers of circles tangent to the two lines, which amounts to saying that any one of these circles intersects one of these bisectors at right angles. Here we see that any circle , tangent to C and C , or, more generally, any circle which cuts C and C at equal angles, is orthogonal to circle , or to the analogous circle centered at the other center of similarity (if these circles exist).

The circle always exists if the circles C, C intersect: in this case it passes through their common points.

In general, when it exists, it has the same radical axis as C and C since these three circles have the same set of orthogonal circles (by the preceding section).

Exercises

Exercise 242. When two points are inverses with respect to a circle, the ratio of their distances to a variable point on the circle is constant.

Exercise 243. Show that Exercise 68, in the case where the given circle intersects the line in a point I, can be reduced to Exercise 65 by means of an inversion with pole I.

Exercise 244. Through a common point A of two circles C, C' we draw secants AMM', ANN', which intersect C in M, N and C' in M', N'. Circles AMN', ANM' intersect at A and at another point. Find the locus of this point as the two secants turn, independent of one another, about point A.

Exercise 245. The inverses of circles with a common radical axis are also circles with a common radical axis.

Exercise 246. Transform the definition of a circle by inversion (7) to obtain a new proof of the theorem of 116.

Exercise 247. When we take the inverse of a circle, what is the point whose image is the center of the transformed circle?

Exercise 248. Two given circles can always be transformed, by the same inversion, either into two parallel lines, or into two concentric circles.

Exercise 249. Given three points on a line, find a fourth point on the same line such that if we invert with respect to this point, the three given points divide the line into two equal parts.

Exercise 250. Two figures are inverse to each other with respect to an inversion S. We apply some inversion T to both figures. Show that the new figures obtained are again inverse to each other, and find the new pole of inversion. In particular, examine the case when the power of the inversion S is positive. (The new circle of inversion is obtained by applying the inversion T to the circle of inversion of S.)

Exercise 251. An inversion S is applied to a figure A to obtained a figure B, which is then transformed into a figure A' by an inversion S'. Assume that the powers of inversion are positive.

1°. Show that by applying an appropriate inversion T to A, A', we can transform these figures into congruent or homothetic figures, these two possibilities being mutually exclusive;[4]

2°. Show that we can find infinitely many pairs of inversions S_1, S_1', equivalent to S, S'; that is, such that consecutive application of S_1 and S_1' to the figure A yields figure A'. In particular, we can always replace the two given inversions by an inversion preceded or followed by a line symmetry, except in one case (the case where figures A, A' are similar);

3°. When are the figures obtained by applying T (of 1°) to A, A' images of each other under translation?

4°. What happens if we apply the operations S, S' several times in succession (that is, we apply S to transform A' into B', then apply S' to transform B' into A'', then apply S to transform A'' into B'', then apply S' to transform B'' into A''', etc.)? Can it happen that we eventually obtain the original figure A?

---

[4] Except in the case mentioned in 3°, which can be considered a common limiting case, since a translation can be viewed as a limiting case of a homothecy.

Exercise 252. We successively apply inversions S, S', S'', etc. (as in the preceding exercise). Show (limiting the discussion to the case of positive powers of inversion) that this sequence of operations can be replaced by a single inversion preceded or followed by one, two, or three line symmetries, unless the final figure and the original figure are similar (preceding exercise, 2°, and Exercise 94).

Exercise 253. In the preceding exercise, assume that an odd number of inversions S, S', S'', . . . has been applied. Find a point which returns to its original position.

Exercise 253b. In a given circle, inscribe a polygon whose sides pass through given points, or are parallel to given directions (preceding exercise).

Exercise 254. On a fixed tangent to a circle, with point of contact T, we draw variable segments TM, TN whose product is constant. Let T' be the point on the give circle which is diametrically opposite T.

1°. Show that the line joining the second intersections of $T'M$, $T'N$ with the circle passes through a fixed point;

2°. Same problem for the line joining the points of contact of the second tangents from M, N (reduce this to the preceding case).

3°. Find the locus of the intersection of these two tangents.

Exercise 255. Given a point O on the plane of a circle, we draw a variable secant which intersects the circle in M, N. The circles with diameters OM, ON intersect the given circle again in M', N'. Find the locus of the intersection of MN', $M'N$ as the secant turns about point O.

Exercise 256. When a variable circle intersects two fixed circles at a constant angle, it also cuts any fixed circle having the same radical axis as the first two at a constant angle.

Exercise 257. On two segments AB, CD on the same line, we draw arcs of circles with the same central angle V. As the angle V varies, find:

1°. the locus of the midpoint of the common chord of the two circles of which these arcs are a part;

2°. the locus of the intersection of these circles; in other words, the locus of points at which the given segments subtend equal or supplementary angles. (Use Exercise 219b.)

CHAPTER VI

The Problem of Tangent Circles

229. **Problem.** Draw a circle passing through two given points, and tangent to a given line or a given circle.

We have already solved this problem (159). Inversion allows us to give another solution. Let us transform the figure by an inversion with its pole at one of the given points. The required circle becomes a line passing through a known point and tangent to a known circle (the inverses of the second point and of the given circle).

230. **Problem.** Draw a circle passing through a given point and tangent to two given lines or circles.

First method. The required circle corresponds to itself in one of two inversions (or line symmetries) which transform one of the lines into the other. Thus, in addition to the given point, we know a second point; namely the point which is inverse or symmetric to the first. We are thus led to the preceding problem. There can be four solutions, since the preceding problem has two.

Second method. Invert the given figure, with the given point as pole. We are led to the construction of the common tangents to two circles.

231. **Problem.** Draw a circle tangent to three given circles.

This can be reduced to the preceding problem.

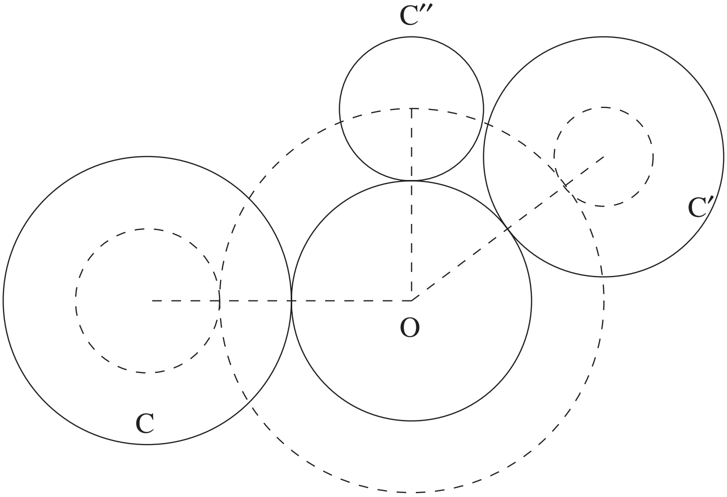

Figure 193

Indeed, consider, the given circles C, C , C (Fig. 193) with radii r, r , r , and a circle with center O and radius R, which, let us say, is externally tangent to the three circles. Then a circle with center O and radius R + r passes through the center of C , and is tangent to the circles concentric to C, C with radii equal to the difference between r and the radii r, r , respectively.

The cases where the required circle is tangent internally to one or more of the given circles are treated analogously, except that certain differences of radii are replaced by sums.

232. We can also solve the same problem using an entirely different method (due to Gergonne).

Let A, B, C be the given circles. We first look for a circle (Fig. 194) which is tangent in the same way to the three given circles, at points a, b, c.

If such a circle exists, there must necessarily be a second[1] circle . Indeed, the radical center I of the three circles has same power with respect to these circles. The inversion with I as a pole and this common power as the power of inversion will not change the given circles, and will transform into a circle tangent to A, B, C at points a , b , c which are the inverses of a, b, c. The contacts will be of the same type; that is, all of the same type as for or the opposite type, according as whether I is an external or internal center of similarity for the circles and .

We will prove that the radical axis xy of the circles , is precisely the axis of direct similarity (145) of the given circles. To see this, consider the intersection S of the lines ab, a b . This point is on the radical axis of , , because the points a, a and b, b are inverse pairs with respect to point I. This point S is the center of direct similarity for the circles A, B (227). In the same way, we can show that the radical axis xy passes through the other two centers of direct similarity of the given circles.

Having established this, we draw the common tangents of A and , which pass through a, a . The point where these tangents intersect will be on xy, because points a, a are inverse with respect to I. This point will be the pole of aa with respect to circle A.

Thus the chord aa passes through the pole p of the direct axis of similarity with respect to the circle A.

We are thus led to the following construction:

Determine the radical center I and the axis of direct similarity xy of the given circles. Join I to the poles p, q, r of xy with respect to these circles. The lines obtained in this way intersect the corresponding circles in the required points a, a ; b, b ; c, c .

233. It remains to prove that this construction will indeed provide circles satisfying the conditions of the problem.

To do this, we will show that the points a, b, c, a , b , c obtained above are antihomologous in pairs in the given circles.

Let us look for the chord in circle B which is antihomolgous to chord aa in circle A. This chord is determined (224) by the following two conditions:

1 . the two chords intersect on the radical axis of the two circles;

---

[1] Circle is different from . Indeed, unless the line Ia is tangent to A, point a is different from a. Likewise, Ib would be tangent to B, and Ic to C. But if a, b, c were chosen this way, the circle abc would be orthogonal, rather than tangent, to A, B, C.

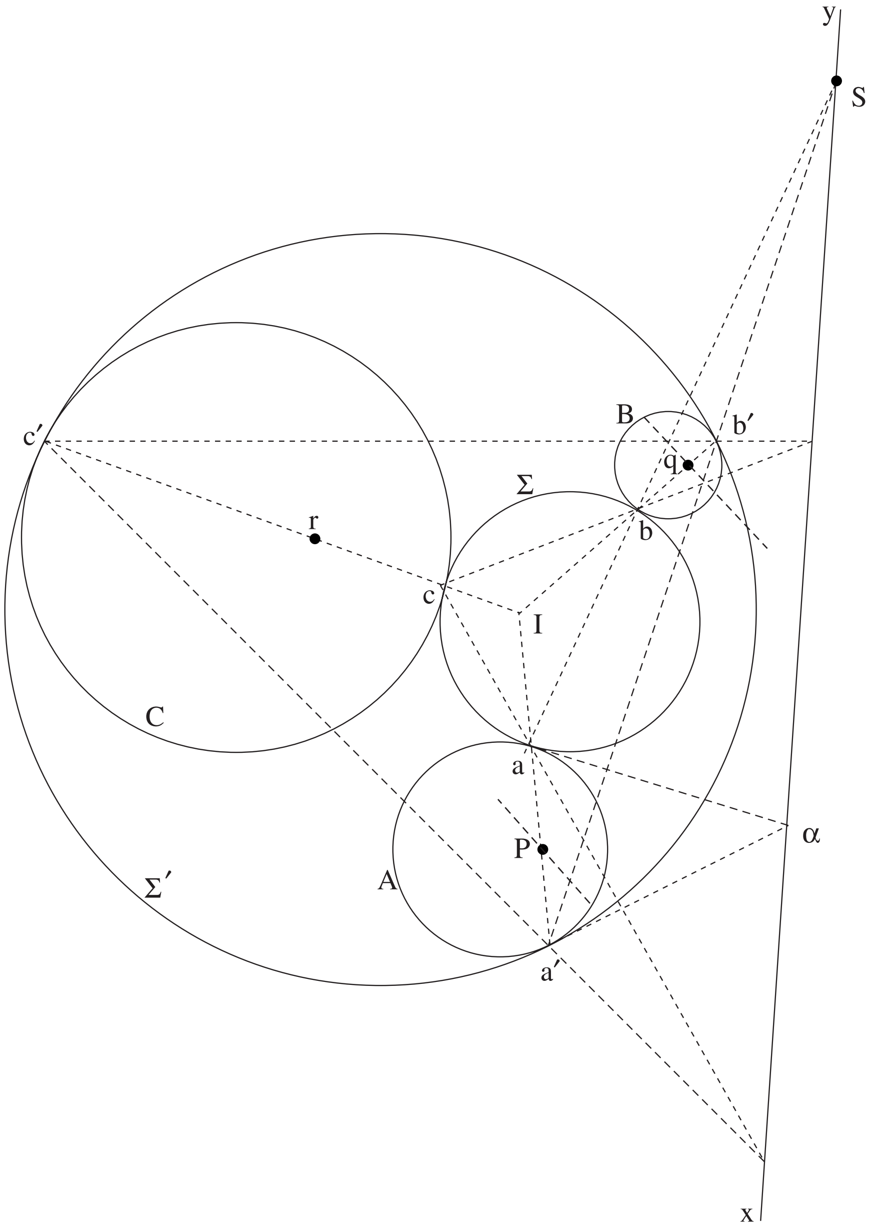

Figure 194

2 . the points where they meet the polars of S with respect to the two circles correspond to each other in the homothecy with center S taking one circle onto the other.

Now condition 1 is satisfied by aa and bb since their intersection I in fact belongs to the radical axis of A and B. Condition 2 is also satisfied. Indeed, points p, q are the poles of xy with respect to circles A, B, so they lie on the polars of S with respect to these circles. And these points correspond to each other in the homothecy with center S which brings circle A onto circle B, since line xy corresponds to itself, and so its poles in the two circles correspond to each other. The chord antihomologous to aa is therefore bb. In the same way, we can show that chords cc, aa are antihomolgous in circles C, A, and chords cc, bb are antihomoogous in circles C, B.

We will denote by b, c the two antihomologous points for a (without having yet proved that these two points are antihomologous); then b, c will be the antihomologues points for a. Let us draw a circle through a, b, c and a circle through a, b, c.

These two circles are inverse to each other with respect to point I, since the the pairs a, a; b, b; c, c are inverse with respect to this point. The circle inverse to that of abc is therefore a b c.

The radical axis of these circles is xy because this radical axis must pass through the intersection of ab, a b, and through the intersection of ac, a c.

Now let $_1$ be the intersection of the tangents to circle A at a, a. This point is on xy, since aa passes through point p; it is also on the perpendicular bisector of aa.

But the point , where the tangents at a, a to , intersect, is also on xy, because this line is the radical axis of , ; it is also on the perpendicular bisector of aa because the tangents a, a, to circles and respectively, are equal.[2]

Therefore points and $_1$ coincide, and the circles , are tangent to A at a, a. They are likewise tangent to B and C at b, b; c, c (227).

234. We have sought a circle tangent to A, B, C with contacts of the same kind. We would find circles which have contacts of different kinds with the circles A, B, C by replacing the direct axis of similarity with an inverse axis of similarity in the preceding reasoning.

As there are four axes of similarity, the problem of tangent circles can have eight solutions. However, some or all of these solutions may be missing: this will happen, for instance, if line Ip does not intersect circle A. We can see that if the power of the point I with respect to the three circles is negative, and therefore the point I is interior to the circles, all eight solutions exist. These eight solutions also exist if the three circles are exterior to each other, as can easily be seen solving the problem by the method of 231.

235. Gergonne's solution can also be applied when two of the given circles are replaced by points or lines. This can be seen by repeating the reasoning of 232 using these new conditions. However, the construction only yields the points of contact with the circles. In particular, it does not yield any result if all circles are replaced by points or lines.

236. Gergonne's solution fails to apply in certain special cases. Indeed, if the three centers are collinear, the radical center and the three poles of each axis of similarity are thrown to infinity in the direction perpendicular to this axis. This position can be avoided by first applying an inversion to the figure.

Conversely, three circles can, in general, be transformed by an inversion into three circles with collinear centers. For this to be possible, it is sufficient that the common power of the radical center be positive and consequently that there exist a circle cutting the three given circles at right angles. If we take any point at all on this circle as the pole of inversion, the given circles are transformed into three circles which are intersected at right angles by the same line, and therefore their centers will be on that line.

---

[2] The original text here refers to the tangents $_1$a, $_1$a . Since the points and $_1$ in fact coincide, this statement is correct, but the passage seems clearer if we refer to the point here as . transl.

The inconvenience which we just noted would therefore disappear if we could formulate a solution in terms of properties which do not change under an inversion, such as the tangency of two circles, the angle between them, etc. One can in fact modify Gergonne's solution in this way.[3]

Exercises

Exercise 258. Construct a circle passing through two points and cutting a given circle at a given angle.

Exercise 259. Construct a circle orthogonal to two given circles and tangent to a third given circle. More generally, construct a circle orthogonal to two given circles and cutting a third given circle at a given angle. (See Exercise 248.)

Exercise 260. Through two points A, B we draw two circles tangent to the same circle C, and a third one which is orthogonal to C. Show that this last circle divides the angle made by the other two into equal parts, and that its center is the intersection of their common tangents (see 229).

Exercise 261. Through a point A we draw two tangent circles (with contacts of the same kind) to two given circles, and a third one which is orthogonal to them. Show that this last circle passes through the second common point B of the first two, and possesses the properties indicated in the preceding exercise.

Exercise 262. Two tangent circles are drawn through a point A as in the preceding exercise. Let P, Q be their points of contact with the first given circle, and P', Q' the points of contact with the second.

1°. The circles APQ, $AP'Q'$ are tangent. They are orthogonal to the third circle of the preceding problem;

2°. The circles APQ, BPQ (where B is the second common point of the circles APP', AQQ', as in the preceding exercise) are inverse to each other with respect to the first of the given circles; the circles $AP'Q'$, $BP'Q'$ are inverse with respect to the second.

3°. What happens to these statements when the given circles are replaced by lines? Show that in this case the four circles APQ, $AP'Q'$, BPQ, $BP'Q'$ are equal.

Exercise 263. Draw a circle passing through a given point and tangent to two given lines by first drawing an arbitrary circle C tangent to the two lines, then noting that the required circle must be homothetic to C with respect to the intersection point O of the lines.

Exercise 264. Use an analogous method to construct a circle tangent to two given lines and to a given circle. (The line joining O to the point of contact cuts circle C and the given circle at equal angles.)

---

[3] See Note C at the end of this volume.

Exercise 265. The same problem, but the two lines are replaced by two concentric circles. (Use a circle concentric to the given circles, and passing the point of contact of the required circle with the third given circle.)

Exercise 266. Two variable tangent circles are each tangent to two fixed circles. Find the locus of their point of contact. (The first method of 230 shows when two circles tangent to two given circles can have two identical intersection points.)

Exercise 267. Pairs of the centers of the eight circles tangent to two given circles are on the perpendiculars dropped from the radical center to the four axes of similarity.

Exercise 268. Through a point inside an angle, draw a transversal forming a triangle of minimum perimeter with the sides of the angle (not extended past the vertex). (Use the indirect method (Exercise 174) and Exercise 90.)

CHAPTER VII

Properties of Cyclic Quadilaterals. Peaucellier's Inverter

237. **Ptolemy's Theorem.** The product of the diagonals of a cyclic quadrilateral is equal to the sum of the products of opposite sides.

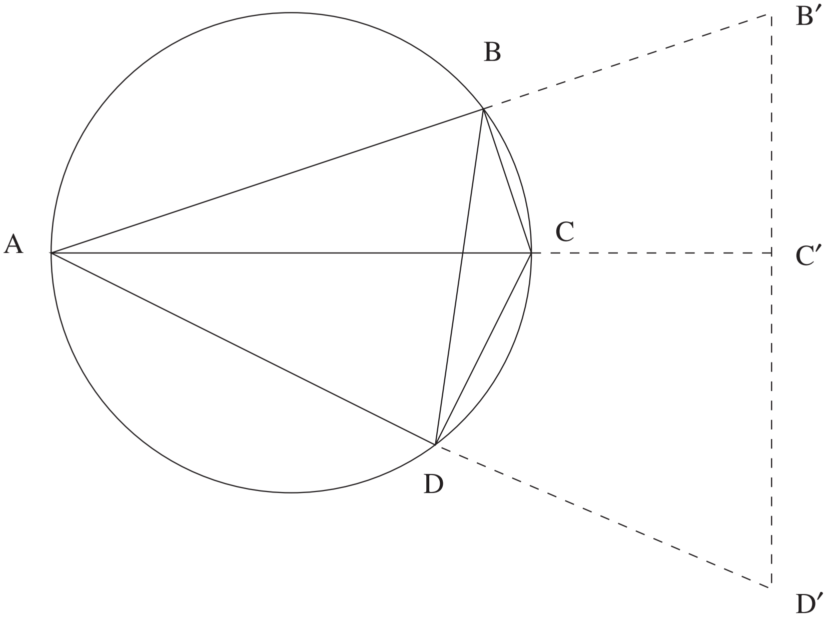

Figure 195

Suppose quadrilateral ABCD (Fig. 195) is inscribed in a circle. We transform the figure by an inversion with its center at point A. Circle ABCD is transformed into a line which passes through the inverses B′, C′, D′ of B, C, D. If C is the vertex of the given quadrilateral opposite to A, the points B, D will be on different sides of the line AC, and therefore so will points B′, D′. Thus point C′ is between B′ and D′, and we have (in absolute value)

$$B'D' = B'C' + C'D'.$$

But if k is the power of inversion, we have (218):

$$B'D' = DB \cdot \frac{k}{AB \cdot AD}, \quad B'C' = CB \cdot \frac{k}{AB \cdot AC}, \quad C'D' = DC \cdot \frac{k}{AC \cdot AD}.$$

Substituting these values into the preceding equality, then multiplying by $AB \cdot AC \cdot AD$, and dividing by k, we obtain the stated relation:

$$AC \cdot BD = AD \cdot BC + AB \cdot CD.$$

237b. The preceding property actually characterizes cyclic quadrilaterals. This is the result of the following theorem:

Theorem. If a quadrilateral is not cyclic, then the product of the diagonals is smaller than the sum of the products of opposite sides (and greater than their difference).

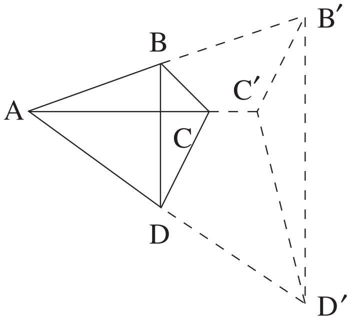

Figure 195b

Suppose $ABCD$ (Fig. 195b) is a quadrilateral which is not cyclic. If we repeat the preceding construction, points B', C', D' are no longer collinear, and they will form a triangle.

The values found in the previous section for $C'D'$, $D'B'$, $B'C'$ are proportional to the products $DC \cdot AB$, $BD \cdot AC$, $CB \cdot AD$. Therefore any of these products is less than the sum of the other two.[1]

238. The relation in Ptolemy's theorem is still true, and is proved in the same way, when the four points A, B, C, D are situated in this order on a straight line rather than a circle (Fig. 196). Thus, if four points are collinear, the product of the overlapping segments they form is equal to the sum of the products of the non-overlapping segments.

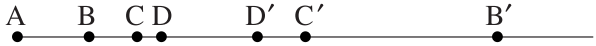

Figure 196

Taking signs into account, this relation can be written as

$$AB \cdot CD + AC \cdot DB + AD \cdot BC = 0.$$

---

[1] Hadamard's text does not mention this directly, but this observation also proves the assertion made about the difference between the products of opposite sides. transl.

This is the result of repeating the argument of 237: the relationships we used are, in fact, true in magnitude and in sign (218, Remark).

239. Problem. Knowing the chords AB, AC (Fig. 197) of two arcs in a circle of radius R, find the chord of the sum or difference of the two arcs.

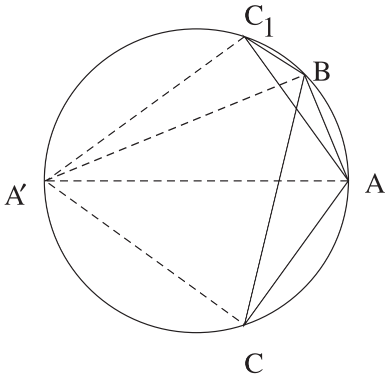

Figure 197

This is the problem we had to solve in 174 in order to calculate the sides of the various regular pentadecagons.

1°. Assume that arcs $\overset{\frown}{AB}$, $\overset{\frown}{AC}$ are described in contrary senses, so that arc $\overset{\frown}{BC}$ is their sum. If we draw diameter AA', we have:

$$AA' = 2R, \quad A'B = \sqrt{4R^2 - AB^2}, \quad A'C = \sqrt{4R^2 - AC^2}.$$

The cyclic quadrilateral $ABA'C$ gives

$$BC \cdot AA' = AB \cdot A'C + AC \cdot A'B,$$

a relation in which every quantity is known, except BC, and which gives us:

$$BC = \frac{AB\sqrt{4R^2 - AC^2} + AC\sqrt{4R^2 - AB^2}}{2R}.$$

2°. Assume that the arcs $\overset{\frown}{AB}$, $\overset{\frown}{AC_1}$ are described in the same sense, so that arc $\overset{\frown}{BC_1}$ is their difference. The cyclic quadrilateral $AA'BC_1$ now gives us:

$$A'B \cdot AC_1 = AB \cdot A'C_1 + AA' \cdot BC_1$$

so that

$$BC_1 = \frac{A'B \cdot AC_1 - AB \cdot A'C_1}{AA'} = \frac{AC_1\sqrt{4R^2 - AB^2} - AB\sqrt{4R^2 - AC_1^2}}{2R}.$$

240. Theorem. The ratio of the diagonals of a cyclic quadrilateral is equal to the ratio of the sums of the products of the sides which meet at their endpoints.

Suppose the diagonals of cyclic quadrilateral $ABCD$ (Fig. 128) intersect at O. Triangles OAD, OBC are similar (131), and give us:

$$\frac{OA}{OB} = \frac{AD}{BC} = \frac{OD}{OC},$$

which can be rewritten as

$$\frac{OA}{AB \cdot AD} = \frac{OB}{AB \cdot BC}, \qquad \frac{OC}{BC \cdot CD} = \frac{OD}{AD \cdot CD}.$$

Likewise, the similar triangles OAB, OCD yield

$$\frac{OA}{OD} = \frac{AB}{CD}, \quad \text{or} \quad \frac{OA}{AB \cdot AD} = \frac{OD}{AD \cdot CD}.$$

Thus the four ratios

$$\frac{OA}{AB \cdot AD}, \quad \frac{OB}{AB \cdot BC}, \quad \frac{OC}{BC \cdot CD}, \quad \frac{OD}{CD \cdot DA}$$

are equal. Adding the numerators and denominators of the first and third, and of the second and fourth, we obtain the desired relation

$$\frac{AC}{AB \cdot AD + BC \cdot CD} = \frac{BD}{AB \cdot BC + CD \cdot AD}.$$

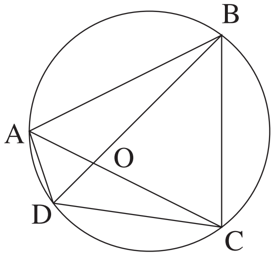

Figure 198

Another proof. Let a = AB, b = BC, c = CD, d = DA be the four sides of the quadrilateral. Permuting these sides in all possible ways, we obtain other quadrilaterals inscribed in the same circle, because the sum of the arcs corresponding to these four chords will always be the whole circle. Clearly, we can always take a to be the first side, so we have the six possible permutations

abcd, adcb,
acdb, abdc,
adbc, acbd.

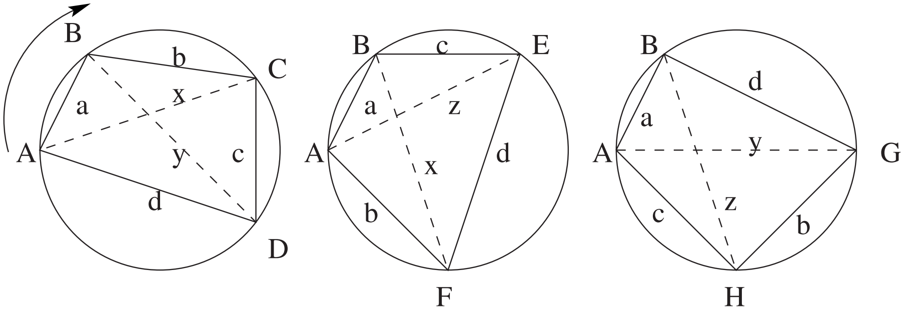

Figure 198b

However, the permutations in the same row give the same quadrilateral: for example, in quadrilateral $ABCD$ (Fig. 198b) the order of the sides is a, b, c, d if we read them in the order indicated by the arrow, and a, d, c, b if we read in the opposite direction. The remaining three permutations are $abcd$, $acdb$, $adbc$ which correspond to quadrilaterals $ABCD$, $ABEF$, $ABGH$ (Fig. 198b). Arc $\overset{\frown}{BF}$ is equal to $\overset{\frown}{AC}$, since it is the sum of equal arcs. Likewise, arc $\overset{\frown}{BH}$ is equal to $\overset{\frown}{AE}$ and $\overset{\frown}{BD}$ to $\overset{\frown}{AG}$. The three quadrilaterals have only three distinct diagonals: $AC = BF = x$, $BD = AG = y$, $AE = BH = z$, and applying Ptolemy's theorem to quadrilaterals $ABEF$, $ABGH$ gives

$$xz = ad + bc, \quad yz = ab + cd,$$

which gives us, by division, the required ratios.

240b. **Problem.** Calculate the diagonals x, y of a cyclic quadrilateral with sides a, b, c, d.

We know the product and the ratio of the diagonals:

$$xy = ac + bd, \quad \frac{x}{y} = \frac{ad + bc}{ab + cd}.$$

Multiplication yields

$$x^2 = \frac{(ac + bd)(ad + bc)}{ab + cd},$$

and we can find y^2 similarly.[2]

Problem. Calculate the radius of the circumscribed circle for a cyclic quadrilateral with sides a, b, c, d.

In triangle ABD (Figures 195 or 198), the radius R of the circumscribed circle is given by the formula (130, 130b):

$$R^2 = \frac{a^2 d^2 \cdot BD^2}{[(a + d)^2 - BD^2][BD^2 - (a - d)^2]}.$$

Replacing $BD = y$ by the value found above we obtain

$$(a + d)^2 - BD^2 = \frac{(a + d)^2(ad + bc) - (ac + bd)(ab + cd)}{ad + bc} = \frac{ad[(a + d)^2 - (b - c)^2]}{ad + bc}$$
$$= \frac{ad(a + d + b - c)(a + d + c - b)}{ad + bc};$$
$$BD^2 - (a - d)^2 = \frac{ad(b + c + a - d)(b + c + d - a)}{ad + bc}.$$

Thus we find

$$R^2 = \frac{(ac + bd)(ab + cd)(ad + bc)}{(b + c + d - a)(c + d + a - b)(d + b + a - c)(b + a + c - d)},$$

an expression independent of the order of the sides, which is consistent with our earlier remarks (240). Denoting by p the semi-perimeter (so that $2p = a + b + c + d$), we can rewrite the formula as follows:

$$R^2 = \frac{(ac + bd)(ab + cd)(ad + bc)}{16(p - a)(p - b)(p - c)(p - d)}.$$

---

[2] That is, by dividing. — transl.

241. Peaucellier's inversion cell.

Theorem. Consider a rhombus $MPM'Q$ and a point O such that $OP = OQ$ (Fig. 199). Assume that the rhombus is articulated (46b), the (equal) lengths OP, OQ held constant, and the point O remains fixed. Then the points M and M' move along two figures inverse to each other.

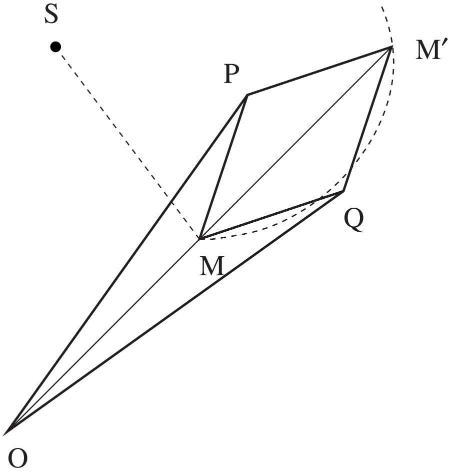

Figure 199

Indeed, points M, M', O are equidistant from P, Q, and they are therefore on the same line; namely, the perpendicular bisector of PQ. Let us draw a circle with center P and passing through M and M'. The power $OM \cdot OM'$ of O with respect to this circle is then equal to $OP^2 - PM^2$, and it therefore remains constant. QED

This theorem allows one to solve a problem of theoretical, if not practical, importance.

The apparatus used to draw circles, the compass, is accurate by definition: the two points will remain at a constant distance provided there is sufficient friction at the hinge of the compass.

On the other hand, the correctness of a ruler depends on its edge having the form of a straight line. This can be realized only by approximation, which is more or less exact depending on the care with which the instrument is manufactured.

Can one draw a straight line without having such a line in advance, using only properties of figures we can be sure of?

This is precisely what the apparatus described in the preceding result, due to Peaucellier, allows us to do. Indeed, assume that OP, OQ, PM, PM', QM, QM' are rigid rods[3] attached to each other, and the point O is fixed. Then M, M' must be inverses of each other with respect to point O. Let us make point M describe a circle, by attaching it to a fixed point S using a seventh rod (Fig. 199). If this circle passes through point O (that is, if $SM = SO$), point M' will describe a straight line.

---

[3] In Figure 199, we have represented OP, OQ, ... by straight lines, but it is not necessary that they should have this form. It is only necessary that the points O and P, for instance, should be kept at a constant distance, no matter how this is done.

241b. **Hart's inversion cell.** Let ABCD be a anti-parallelogram (46b, Remark II; Fig. 199b) formed by the diagonals and the non-parallel sides of an isosceles trapezoid. Let O, M, M′ be the three points where sides AB, AD, BC intersect a line parallel to the bases of the trapezoid. If the anti-parallelogram is articulated while the common length a of sides AB, CD remains constant, as well as the common length b of BC and AD, and while points O, M, M′ remain attached to the corresponding sides (that is, the lengths AO, AM, BM′ remain constant in magnitude and sense) and finally, if the point O remains fixed, then the points M and M′ move along two figures inverse to each other with respect to the pole O.

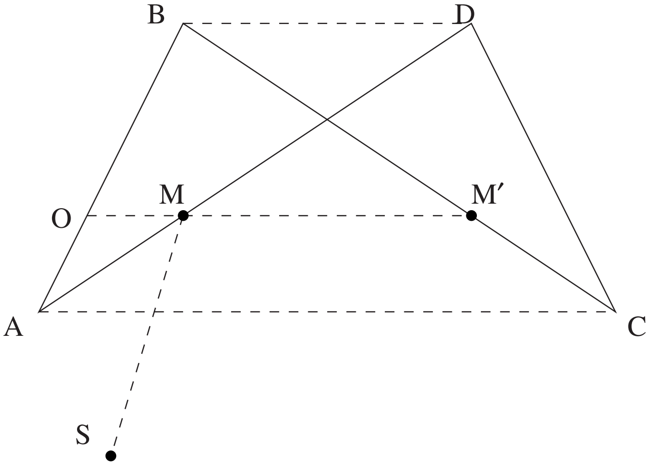

Figure 199b

Proof. We know (46b) that the quadrilateral will remain an anti-parallelogram.

On the other hand, the lines OM, OM′, which we assumed were parallel to the bases of the trapezoid in the initial position of the figure, will remain so, because this parallelism is characterized by the proportions

$$\frac{AM}{AD} = \frac{AO}{AB}, \quad \frac{BM'}{BC} = \frac{BO}{BA},$$

which, by hypothesis, do not change.

In particular, we see that points O, M, M′ remain collinear.

Having seen this, assume for the moment that not only O, but the entire side AB remains fixed. Then points M, M′ will move along circles with centers A, B, and radii AM, BM′ = M′D, respectively.

Since the ratio of these radii is equal to the ratio of the distances OA, OB, point O is one of the centers of similarity for the two circles. Since the radii with endpoints M, M′ are not parallel, these points will be (223) antihomologous, and therefore inverse. The product OM • OM′ therefore remains constant as the anti-parallelogram is deformed, and while points A, B remain fixed.

Since this product is not changed by a rigid rotation around O, the theorem is proved. QED

Corollary. If point M is attached to a fixed point S by a rigid rod of length equal to SO, point M will move along a straight line.

Remark. The product of the bases AC, BD remains constant as the figure is deformed. Indeed, the ratio of BD to OM is constant and equal to $\frac{1}{h} = \frac{AB}{AO}$, and the ratio of AC to OM is equal to $\frac{1}{h} = \frac{BA}{BO}$.

Exercises

Exercise 269. The converse of Exercise 99: If a point M in the plane of an equilateral triangle ABC is such that $MA = MB + MC$, this point is on the circumscribed circle. Otherwise, $MA < MB + MC$.

Exercise 270. If we had started the proof in 237 and 238 with a vertex other than A, we would have obtained a triangle analogous to $B\ C\ D$.

1 . Show that all of these triangles are similar;

2 . Calculate the angles of any one of these triangles, knowing the angles made by the sides and diagonals of the given quadrilateral;

3 . Show that we would obtain triangles similar to the preceding ones by dropping perpendiculars from one of the four vertices A, B, C, D to the sides of the triangle formed by the other three; or by joining B and D to the third vertex of a triangle similar to ABC, with the same sense of rotation, in which the side corresponding to AC is AD;

4 . Show that the shape of the similar triangles considered above is preserved, except for the sense of rotation, if points A, B, C, D are all subject to the same inversion; in other words, show that if we operate on these transformed points as we have operated in 1 , 2 or 3 on the points A, B, C, D, we will obtain triangles similar to the first ones;

5 . Conversely, if two quadrilaterals $ABCD$, $A_1B_1C_1D_1$ are such that the triangles resulting from the preceding constructions applied to each of them are similar, show that there exists[4] an inversion which transforms A, B, C, D into four points forming a quadrilateral congruent to $A_1B_1C_1D_1$. Determine this inversion.

Exercise 270b. Find an inversion which transforms three given points into three other points forming a triangle congruent to a given triangle.

Exercise 271. Peaucellier s inversion cell would still function if the articulated quadrilateral $MPM\ Q$ were not a rhombus, but if $MP = M\ P$, $MQ = M\ Q$, and $OP^2 \quad OQ^2 = MP^2 \quad MQ^2$.

Exercise 271b. In Hart s inversion cell, calculate the power of the inversion knowing the sides a, b of the anti-parallelogram, as well as the ratios h and $h \ = 1 \quad h$. Calculate the product of the bases of the trapezoid knowing a and b.

Problems for the Complements to Book III

Exercise 272. The radical axis and the center of similarity divide the segment joining two antihomologous points in a constant cross ratio.

Exercise 273. The cross ratio of four points on a circle is equal to the cross ratio of their inverses on a circle (or line) inverse to the first.

---

[4] The case where $ABCD$, $A\ B\ C\ D$ are similar is an exception transl.

Exercise 274. The cross ratio of four points on a circle is obtained by dividing the products of the opposite sides of the quadrilateral they form, or dividing one of these products by the product of the diagonals.

Exercise 275. Which inversions transform two given circles into equal circles? Find an inversion which transforms three given circles into three equal circles.

Exercise 276. Given three circles, we draw the three circles (228) with respect to each pair of these circles, and with centers at the three centers of similarity situated on the same axis. Show that these three circles have the same radical axis.

Exercise 277. What is the locus of the centers of the inversions which transform two given circles into two others such that the first divides the second into equal parts; or, more generally, such that the first intercepts on the second an arc corresponding to a given angle at the center of the second?

Exercise 278. The segments intercepted by two circles on a line subtend, at the limit points (Exercise 152), angles having the same bisector, or perpendicular bisectors. Special case: the line is tangent to one of the circles.

Exercise 279. Prove Pascal's theorem (196) by forming three circles, each pair of which has a center of similarity at the intersections of two opposite sides of the hexagon. (Each of these circles passes through two opposite vertices.)

Exercise 280. Use Exercise 253 to construct the circles tangent to three given circles.

Exercise 281. A right angle pivots around a fixed point in the plane of a circle. We draw tangents to the circle at the intersection points with the sides of the angle. Find the locus of the vertices of the quadrilateral obtained in this way. (This locus is the inverse of that obtained in Exercise 201.)

Exercise 282. A quadrilateral ABCD is inscribed in a circle O and circumscribed around a circle o. Let a, b, c, d be its points of contact with o.

1 . The point P where the diagonals of ABCD and abcd intersect (exercise 239) is a limit point of the circles O, o (Exercise 241, part 2).

2 . The diagonals of abcd bisect the angles formed by the diagonals of ABCD (Exercise 278).

3 . If we apply the preceding exercise to circle o and point P, we obtain circle O.

4 . Conclude that, given two circles, we cannot generally find a quadrilateral inscribed in one and circumscribed about the other. For this to be possible, a certain relation must be satisfied by the radii and the distance between the centers. If this relation is satisfied, there are infinitely many quadrilaterals satisfying the conditions.

Exercise 283. If a triangle has an obtuse angle, there exists a circle (and only one) with respect to which this triangle is conjugate (211, Remark). The center of this circle is the intersection of the altitudes of the triangle.

Exercise 283b. For the existence of a triangle inscribed in a given circle, and conjugate to another given circle, it is necessary that the square of the radius of the second circle be equal to one half the power of its center with respect to the other circle. If this condition is satisfied, there are infinitely many triangles with both required properties. (Use Exercise 70.)

Exercise 284. Given three circles, each pair of which intersect, we construct a curvilinear triangle whose three vertices are intersection points of the circles, and whose sides are arcs of these circles which do not contain any other intersection points. Show that the sum of the interior angles of the three circles is less than, or greater than, two right angles, according as whether there exists or not a circle orthogonal to the three given circles.[5]

Exercise 285. Since the necessary and sufficient condition for a quadrilateral $ABCD$ to be cyclic is that the points B', C', D', inverse to B, C, D with respect to A, are collinear, all properties of cyclic quadrilaterals can be deduced from this condition.

In particular, deduce the theorem on the ratio of the diagonals. (Application of Stewart's theorem, 127.)

Exercise 286. Through two points which divide a diameter of a circle harmonically, we draw perpendiculars to this diameter. These lines are intersected by an arbitrary tangent at two points such that the ratio of their distances to the center is constant.

---

[5] See Note A, 290.

Book IV

On Areas

CHAPTER I

The Measure of Areas

242. Two polygons are adjacent (Figs. 200, 200b) if they have one or more common sides or portions of sides (Fig. 200b), without having any interior point in common.

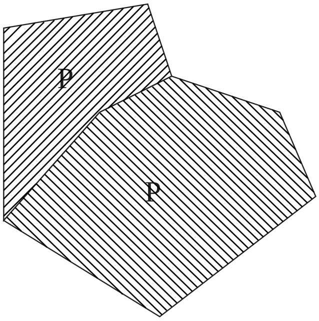

Figure 200

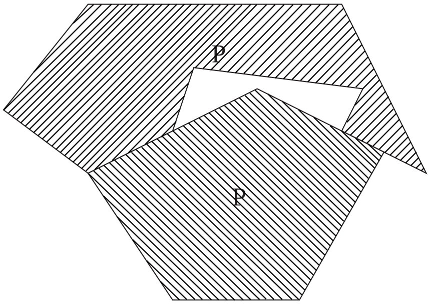

Figure 200b

If we remove the common sides of two adjacent polygons P, P, we form[1] a third polygon P called the sum of the first two. The interior of this polygon consists of the interior points of the original polygons, and only these points.

243. We define the areas of plane polygons by associating with each polygon a quantity (called the area or surface area of the polygon) satisfying the following properties:

I. Two congruent polygons have the same area, regardless of their position in space;

II. The sum P of two adjacent polygons P, P has an area equal to the sum of the areas of P and P.

We assume that such a correspondence can be made.[2]

244. Areas can be defined in infinitely many ways, because if we associate a quantity with each planar polygon so as to satisfy conditions I and II, a quantity proportional to the one we have used will also satisfy these conditions. In order to measure areas, as we are about to do, we must start by designating a polygon whose area is taken as the unit area. The measure of an arbitrary area will be the ratio of this area to the unit.

---

[1] In Book IV we suppress the restriction on the meaning of the word polygon formulated in 20 of Book I. The shaded portion of the plane in Figure 19 will therefore be treated in this book as a polygon. This polygon can be formed by taking the sum of two ordinary adjacent polygons, as shown in Figure 200. On the other hand, we consider in this book only proper polygons.

[2] This assumption is not really necessary, as it can be proved. See note D.

We agree from now on to take, as the unit of area, the area of a square whose side is the unit of length. The theorems we will state assume this convention, and it will not be necessary to repeat it in each individual case.

245. Two polygons with the same area are said to be equivalent. Two congruent polygons are therefore equivalent. The converse, naturally, is not true: two equivalent polygons need not be congruent. For instance, we will show that one can construct a square equivalent to any given polygon.

246. A base of a rectangle is one of its sides. The length of the side perpendicular to the base is then called the altitude of the rectangle. More generally, we can call one of the sides of a parallelogram its base; the altitude is then the distance from the base to the opposite side (measured, of course, along a common perpendicular).

Finally, a base of a triangle is just an arbitrary side, and an altitude is the perpendicular dropped on this side from the opposite vertex.

247. Theorem. The ratio of the areas of two rectangles with the same base is equal to the ratio of their altitudes.

Indeed:

1. Two rectangles with the same base and same altitude are congruent, and therefore equivalent by property I.

2. If three rectangles $ABCD$, $A'B'C'D'$, $A''B''C''D''$ (Fig. 201) have the same base, and the altitude of the third is the sum of the altitudes of the first two, the area of the third will be the sum of the areas of the first two (property II). Indeed, the third can be considered as the sum of two rectangles $A''B''EF$, $C''D''EF$ which are congruent, respectively, to the first two.

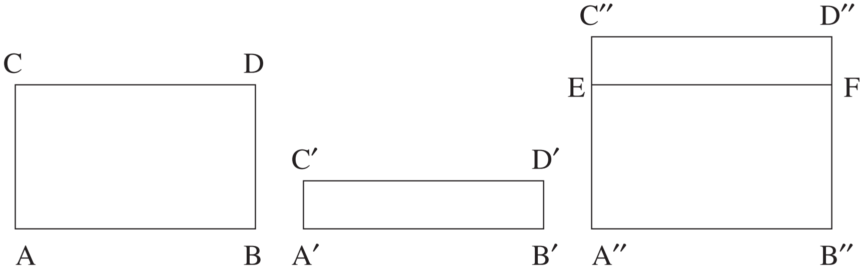

Figure 201

From these two facts we can prove, as in 113 (Book III) or 17 (Book I), and as is explained in the study of arithmetic[3] that an approximation to within $\frac{1}{n}$ of the ratio of the two areas is equal to an approximation within $\frac{1}{n}$ of the two altitudes, which proves the theorem.

Corollary. Since either side of a rectangle can be taken as a base, or as an altitude, any result concerning altitudes is also true for bases. In other words, the ratio of the areas of two rectangles with one equal side is equal to the ratio of the unequal sides.

---

[3] Leçons d'Arithmetique, chap. XIII, n° 493.

Theorem. The ratio of the areas of two rectangles is equal to the product of the ratios of the corresponding dimensions.

Indeed, we have just seen that the area of a rectangle is proportional to its base, and it is also proportional to its altitude, when these dimensions vary separately. The area is therefore proportional to the product of the base and the altitude.

We could also repeat the reasoning usually followed in arithmetic in such cases. Let A be the area of a rectangle with sides a, b, and A' the area of a rectangle with sides a', b'. Consider a rectangle with sides a' and b. The area A'' of this rectangle satisfies

$$\frac{A}{A''} = \frac{a}{a'}, \quad \frac{A''}{A'} = \frac{b}{b'}.$$

In this equation we can think of A, A', A'' as representing the numbers measuring the corresponding areas (relative to a given unit), rather than the areas themselves. For example, we might choose (as we have here), the unit area to be the area of a square constructed on a line segment of unit length. Under these conditions, we can multiply the two equations, and write the product $\frac{A}{A''} \bullet \frac{A''}{A'}$ as $\frac{AA''}{A''A'} = \frac{A}{A'}$. Thus the ratio $\frac{A}{A'}$ is indeed equal to the product of the ratios $\frac{a}{a'}$ and $\frac{b}{b'}$.

Theorem. The area of a rectangle is equal to the product of its two dimensions.

This theorem is the same as the preceding one, applied to the given rectangle and to the square with unit side. The area A' of the square is the unit area, and its dimensions a', b' are both equal to one, so that the ratio $\frac{A}{A'}$ is just the measure of the area A, and the ratios $\frac{a}{a'}$, $\frac{b}{b'}$ are the measures of the lengths a, b.

Remark. 1°. The statement above only makes sense if read according to the conventions of 18 (which we also recalled in 106, Book III). The meaning of the statement is: the *number* which measures the area of a rectangle is equal to the product of the *numbers* which measure the base and the altitude.

2°. This statement is precise only if read according to the convention of 244, which is therefore assumed here implicitly. It is clear *a priori* that the stated equality does not hold if the unit of length and the unit of area are chosen arbitrarily, and independently of each other.

We see from this that we can choose the unit of length arbitrarily, but once this choice is made, the choice for the unit of area is determined. We sometimes say that the unit of area is a derived unit.

248. *Theorem.* The area of a parallelogram is the product of its base and its altitude.

Suppose the parallelogram is ABCD (Fig. 202). We draw perpendiculars from A, B to side AB, and extend them to their intersections c, d with line CD, to form rectangle ABcd. This rectangle is equivalent to the parallelogram, because if we add to these quadrilaterals the right triangles BdD, AcC (which are congruent, having a pair of equal angles and a pair of equal sides) we obtain the same sum; namely, trapezoid AcBD. The rectangle has dimensions equal to the base and altitude of the parallelogram, so the theorem is proved.

249. *Theorem.* The area of a triangle is equal to half the product of the base and the altitude.

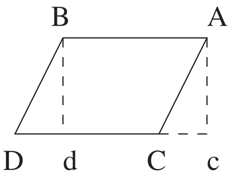

Figure 202

Suppose the triangle is ABC (Fig. 203). We draw parallel AD to BC and parallel CD to AB. The parallelogram ABCD thus formed has the same base and altitude as the triangle. This parallelogram is divided into two congruent triangles ABC, ADC by its diagonal AC (46). Thus triangle ABC is half the parallelogram.

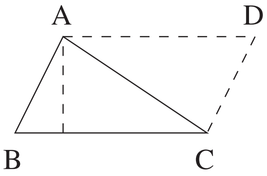

Figure 203

250. Corollary. The locus of the vertices of a triangle with base AB and constant area is composed of two lines parallel to AB.

This is because the vertices in question (Fig. 204) are just those points which are at a fixed distance from AB.

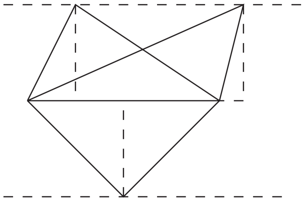

Figure 204

251. Problem. Calculate the area of a triangle knowing the three sides.

We denote the sides of the triangle by a, b, c, and the semi-perimeter by p (so that $a + b + c = 2p$). We have seen (130) that the altitude AH corresponding to side a is

$$AH = \sqrt{\frac{4p(p-a)(p-b)(p-c)}{a^2}} = \frac{2}{a}\sqrt{p(p-a)(p-b)(p-c)}.$$

The area is then given by

$$\begin{aligned} S = \frac{a \cdot AH}{2} &= \sqrt{p(p-a)(p-b)(p-c)} \\ &= \frac{1}{4}\sqrt{2b^2c^2 + 2c^2a^2 + 2a^2b^2 - a^4 - b^4 - c^4}. \end{aligned}$$

Remark. The product of the three sides of a triangle is equal to four times the product of its area and the radius of the circumscribed circle.

In triangle ABC, where AH is the altitude from A, and R the radius of the circumscribed circle, we have (130b) $AB \bullet AC = 2R \bullet AH$. Multiplying both sides by BC, we obtain

$$AB \bullet AC \bullet BC = 2R \bullet AH \bullet BC = 4RS.$$

252. The area of a polygon. The area of an arbitrary polygon can be obtain by decomposing it into triangles, then adding their areas.

The following two results are just examples of this method.

Theorem. The area of a trapezoid is equal to half the sum of the bases times the altitude.

Suppose the trapezoid is ABCD (Fig. 205). Drawing diagonal AD, we divide the base into triangles ABD, ACD. We take their bases to be AB, CD. Then the corresponding altitudes DD , AA are equal to each other and to altitude h of the trapezoid. The area of ABCD is therefore equal to

$$h \bullet \frac{AB}{2} + h \bullet \frac{CD}{2} = \frac{h(AB + CD)}{2}.$$

252b. Corollary. The area of a trapezoid is the product of its altitude and the segment joining the midpoints of the non-parallel sides.

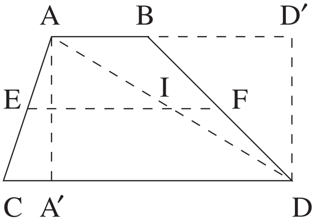

Figure 205

Indeed, the midpoints E, F of sides AC, BD (Fig. 205) and midpoint I of AD are on the same parallel to the bases, and segment EF is equal to half the sum of the bases, since it consists of the pieces EI, IF which are equal to $\frac{CD}{2}$, $\frac{AB}{2}$, respectively.

253. Theorem. The area of a regular polygon is equal to the perimeter times half the apothem.

Indeed, we decompose regular polygon ABCDEF (Fig. 206) into triangles OAB, OBC, . . . , which are all congruent and therefore have the same altitude, equal to the apothem OH. Thus

$$\text{area } OAB = OH \cdot \frac{AB}{2}$$
$$\text{area } OBC = OH \cdot \frac{BC}{2}$$
$$\cdots\cdots\cdots$$
$$\text{area } OFA = OH \cdot \frac{FA}{2},$$

and addition gives us:

$$\text{area } ABCDEF = \frac{OH(AB + BC + \cdots + FA)}{2}.$$

QED

Remark. If the polygon has an even number of sides, its area is equal to half the radius times the perimeter of the polygon obtained by joining every other vertex.

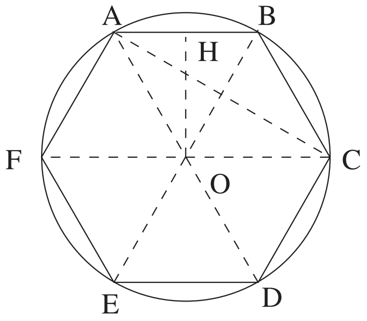

Figure 206

This is true because the area of triangle OAB (Fig. 206) is also equal to half the product of OB and the altitude AI from A, which is equal to half of AC.

253b. A polygon bounded by a broken line inscribed in a circle (Fig. 207), and by the two radii of the circumscribed circle corresponding to the endpoints of the line, is called a polygonal sector.

The polygonal sector is said to be regular if the broken line which serves as its base is regular.

Theorem. The area of a regular polygonal sector is the product of the perimeter of the base and half the apothem.

The proof of this result is identical to the preceding one.

254. Theorem. The area of a convex polygon circumscribed about a circle (and containing the circle in its interior) is equal to the product of its perimeter and half the radius of the inscribed circle.

The polygon (Fig. 208) can be decomposed into triangles with a common vertex at the center of the circle. These triangles, which have as bases the various sides, also have a common altitude equal to the radius.

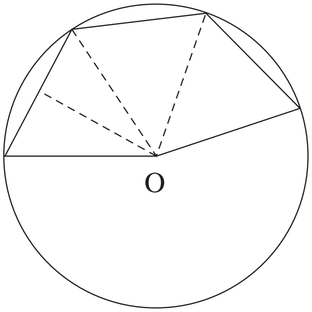

Figure 207

255. Problem. Find the area of a cyclic quadrilateral knowing its four sides.

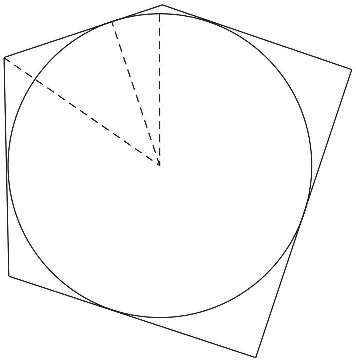

Figure 208

Consider cyclic quadrilateral ABCD in which AB = a, BC = b, CD = c, and DA = d. If R is the radius of the circumscribed circle, we have (251, Remark)

$$a \cdot b \cdot AC = 4R \times \text{area } ABC, \quad c \cdot d \cdot AC = 4R \times \text{area } ACD.$$

The required area S is the sum of the areas of ABC and ACD, so that

$$4RS = AC(ab + cd) = \sqrt{(ac + bd)(ad + bc)(ab + cd)}.$$

Replacing R by the value found in 240b, we have:

$$S = \sqrt{(p - a)(p - b)(p - c)(p - d)}.$$

Remark. The result is independent of the order of the sides, which is obvious a priori if we note that quadrilaterals ABCD, ABEF, ABGH (see 240) are equivalent, being composed of pairs of congruent triangles.

Exercises

Exercise 287. Find the area of an equilateral triangle with side a.

Exercise 288. What is the side of an equilateral triangle with an area of 1 square meter?

Exercise 289. We travel around the perimeter of a square in a determined direction, and join each vertex to the midpoint of the side which precedes the opposite vertex. Show that the lines obtained in this way form a new square, whose area is one-fifth the area of the original square.

Exercise 290. Through a point on diagonal AC of parallelogram ABCD we draw lines parallel to the sides. The parallelogram is divided into four smaller parallelograms, two of which have a diagonal on AC. Show that the other two are equivalent.

Exercise 291. Among all triangles with the same base and same angle at the opposite vertex, which one has the greatest area?

Exercise 292. In a trapezoid, the triangles formed by each diagonal with one of the non-parallel sides are equivalent, and conversely.

Exercise 293. Through the midpoint of each diagonal of a quadrilateral, we draw a parallel to the other diagonal, then join the intersection point of these parallels to the midpoints of the four sides. Show that the quadrilateral is divided by these segments into four parts with equal areas.

Exercise 294. Through each vertex of a quadrilateral, we draw a parallel to the diagonal which does not pass through that vertex. Show that the parallelogram formed this way has twice the area of the quadrilateral.

Two quadrilaterals in which corresponding diagonals are equal and intersect at the same angle are equal in area.

Exercise 295. Find a point inside a triangle such that when this point is joined to the three vertices, three equivalent triangles are formed. More generally, find a point such that the three triangles formed are proportional to three given lengths or numbers.

Exercise 295b. Using area considerations (see the preceding exercise), show that if we join a point to the three vertices of a triangle, the lines formed divide the opposite sides into ratios whose product is one. (This theorem was proved in 197.)

Exercise 296. We join a point O in the plane to the vertices of a parallelogram ABCD (in which AC, BD are diagonals).

1 . If the point O is inside the parallelogram, the sum of the triangles OAB, OCD is equivalent to the sum of OBC, ODA.

2 . For any point O, the triangle OAC is equivalent to the sum or difference of OAB, OAD.

Exercise 297. The area of a trapezoid is equal to the product of a non-parallel side and the perpendicular dropped to this side from the midpoint of the opposite side. Prove this in two different ways:

1 by rewriting the expression given in 252b;

2 by showing directly that the trapezoid is equivalent to a parallelogram with the base and altitude indicated above.

If we join the midpoints of the non-parallel sides of a trapezoid to the endpoints of the opposite side, the triangle obtained has half the area of the trapezoid.

Exercise 298. The sum of the distances of a point inside a regular polygon to its sides is constant.

Exercise 299. The area of a triangle is equal to the product of the semi-perimeter and the radius of the inscribed circle.

The area of a triangle is also equal to the product of the radius of an escribed circle and the difference between the semi-perimeter and the corresponding side (or the radius of an escribed circle and half the difference between this side and the sum of the other two).

Exercise 300. The inverse of the radius of the inscribed circle of a triangle is equal to the sum of the inverses of the radii of the three escribed circles.

Exercise 301. If x, y, z denote the distances from a point inside a triangle to the three sides, and h, k, ℓ are the corresponding altitudes, then

$$\frac{x}{h} + \frac{y}{k} + \frac{z}{\ell} = 1.$$

What happens if the point is outside the triangle?

CHAPTER II

Comparison of Areas

256. Theorem. The ratio of the areas of two triangles with a pair of equal (or supplementary) angles is equal to the ratio of the products of the sides containing these angles.

We can superimpose the angles if they are equal, or make them adjacent if they are supplementary. The two triangles will then be ABC, AB C , with sides AC, AC in the same direction, while the sides AB, AB are either in the same direction (Fig. 209), or are extensions of each other (Fig. 210).

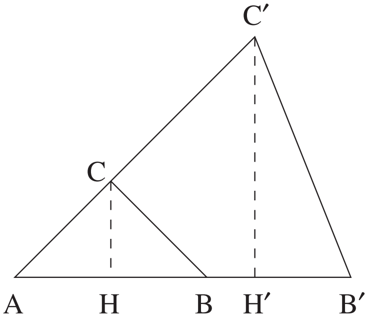

Figure 209

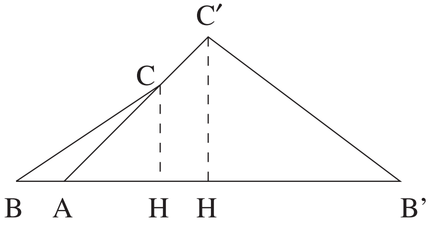

Figure 210

The ratio of the two triangles will be equal to the ratio of their bases AB, AB , multiplied by the ratio of their altitudes CH, CH , and this last ratio is obviously equal to $\frac{AC}{AC}$. QED

257. Theorem. The ratio of the areas of two similar polygons is equal to the square of their ratio of similarity.

We will distinguish two cases.

1 . If we are dealing with two similar triangles ABC, A B C , it suffices to observe that these triangles will have angle A = A . Therefore the ratio of the areas is equal to the product of the ratios $\frac{AB}{A\ B}$ and $\frac{AC}{A\ C}$, that is, to the square of one of them, since these two ratios are equal.

2 Now consider two arbitrary similar polygons, ABCDE, A B C D E (Fig. 211), with a ratio of similarity equal to k.

These polygons can be decomposed into similar triangles, arranged similarly: ABC, ACD, ADE; A B C , A C D , A D E , and as seen in case 1,

$$\frac{ABC}{A\ B\ C} = \frac{ACD}{A\ C\ D} = \frac{ADE}{A\ D\ E} = k^2.$$

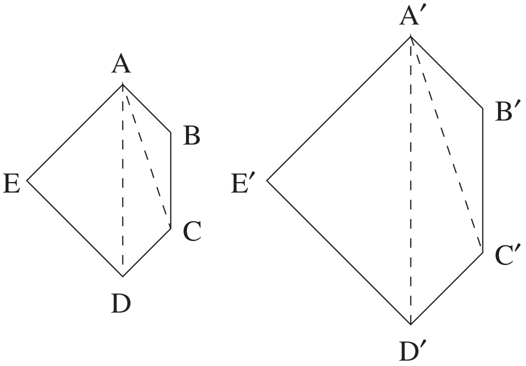

Figure 211

Taking the sum of the numerators and of the denominators we obtain

$$\frac{ABCDE}{ABCDE} = k^2.$$

QED

258. Theorem. The square constructed on the hypotenuse of a right triangle is equivalent to the sum of the squares constructed on the two legs.

Let the right triangle be ABC, and construct squares ABEF, ACGH, and BCIJ, outside this triangle (Fig. 212), on legs AB, AC and on hypotenuse BC.

We drop perpendicular AD from A to BC, and extend it to its intersection K with IJ. We claim that rectangle BDIK is equivalent to square ABEF.

To prove this, we join AI and CE. Then triangle ABI has the same base BI and the same altitude AL = BD as rectangle BDIK; it is therefore equivalent to half this rectangle. Also, triangle EBC has the same base BE and the same altitude CM = AB as square ABEF, so it is equivalent to half this square. Now triangles ABI, EBC are congruent, since they have an equal angle at A (angle B of the triangle plus a right angle) contained between equal corresponding sides (AB = BE and BI = BC). Thus we see that rectangle BDIK is indeed equivalent to square ABEF.

In the same way, we can show that rectangle CDKJ is equivalent to square ACGH. Thus square BCIJ, the sum of the two rectangles, is indeed equivalent to the sum of the two squares.

Remark. This theorem is the same as the one we proved in 124, since the three squares ABEF, ACBH, BCIJ have areas equal to the squares of the numbers which measure the sides. The proof also follows the same line, because the equivalence of rectangle BDIK to square ABEF expresses the same idea as the equation $AB^2 = BC \cdot BD$, which we used at that point.

Exercises

Exercise 302. The three sides of a triangle are divided into given ratios, and a new triangle is formed whose vertices are the points of division. Find the ratio of the area of the new triangle to that of the original triangle. Examine the case

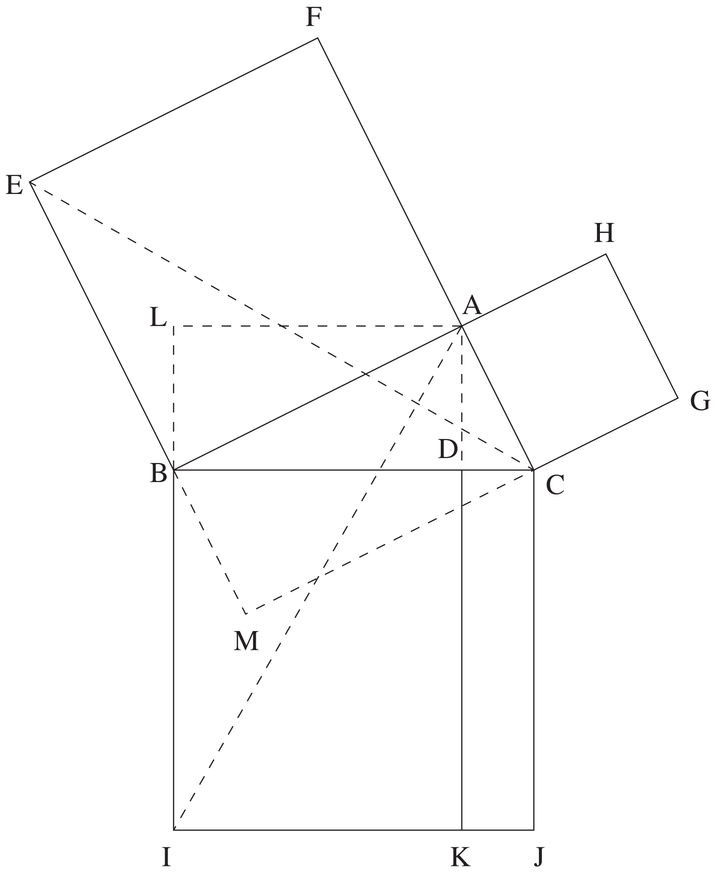

Figure 212

in which one or several of the division points is outside the triangle. Deduce from this proofs for the theorems in 192 193.

Exercise 303. Divide a triangle into two equivalent parts by a line parallel to the base.

Exercise 304. Divide a trapezoid into two equivalent parts by a line parallel to the bases.

Exercise 305. An equilateral triangle is inscribed in a given circle. The radii passing through its vertices are extended to intersect a new circle, the one circumscribing a square which is circumscribed about the original circle. Prove that these points form a triangle equal in area to the regular hexagon inscribed in the original circle.

Exercise 306. Through each vertex of a quadrilateral, we draw a line parallel to a fixed direction, and find the intersection of each of these parallels with the diagonal which does not pass through that vertex. Prove:

1 . that the quadrilateral formed by these intersections is equivalent to the original quadrilateral;

2 . that this new quadrilateral is a trapezoid or a parallelogram, if the original is. More generally, the ratios in which the opposite sides (extended) divide each other are the same in the two quadrilaterals.

Exercise 307. Given some arbitrary polygons in the plane, we draw parallels through each of their vertices in a fixed direction, then measure off on each parallel a distance proportional to the distance from the vertex considered to a fixed line (in the sense corresponding to the sense of the segment from the vertex to the fixed line).

Show that the endpoints of these segments form polygons whose areas are proportional to those of the original polygons (Exercise 297).

When are the new polygon equal in area to the original polygons?

Exercise 308. Show that the preceding exercise contains Exercise 306 as a special case (use Exercise 129).

Alternative Proofs of the Theorem on the Square of the Hypotenuse

Exercise 309. In Exercise 43 (Book I), show that the third square $HBKF$ is equivalent to the sum of the two given squares. Deduce from this the theorem of 258.

Exercise 310. Prove the theorem on the square of the hypotenuse using the theorem of 257, and observing that the right triangle is the sum of the triangles into which it is divided by its altitude.

Exercise 311. Given an arbitrary triangle ABC, we construct any parallelograms $ABEF$, $ACGH$, on bases AB, AC, and outside the triangle. Then we extend EF, GH to their intersection point M. Show that the parallelogram having BC for one side, and the other parallel and equal to AM, is equivalent to the sum of the first two parallelograms.

Show that the previous result is a special case of the theorem on the square of the hypotenuse.

CHAPTER III

Area of the Circle

259. We define the area of a circle as the limit approached by the area of an inscribed or circumscribed polygon, all of whose sides tend to zero.

To show that this limit exists and is independent of the way in which the sides tend to zero, we follow reasoning similar to that used to study the length of the circle. We start by considering regular inscribed polygons with a number of sides which is doubled indefinitely, and the corresponding circumscribed polygons (Fig. 168). Under these conditions:

The areas of the inscribed polygons increase, because each one contains the preceding one. Any of these areas is smaller than the area of an arbitrary circumscribed polygon. Therefore these areas tend to a limit.

In the same way, the areas of the circumscribed polygons decrease, because each one is contained in the preceding one. Each of these areas is greater than the area of any inscribed polygon. Therefore the areas of the circumscribed polygons also tend to a limit.

These two limits are equal, since the ratio of the area of an inscribed polygon and the corresponding circumscribed polygon is equal to the square of the ratio of similarity, which tends to one (176).

260. Let S be the common value of these limits, obtained by starting from the square, for instance, and considering the polygons with $4, 8, 16 \ldots, 2^n \ldots$ sides. Now consider arbitrary inscribed polygons a b c ... (Fig. 169) and the corresponding circumscribed polygons A B ..., imposing only the condition that the number of sides of these polygons increases indefinitely, in such a way that all the sides tend to zero.

The area of the inscribed polygon a b c ... is smaller than S because S is the limit of areas of circumscribed polygons, which are all greater than a b c ....

Since S is contained between the inscribed and circumscribed polygons, it differs from each of them by less than they differ from each other.

Now this difference tends to zero. Indeed, it is composed of triangles a b A , b c B ,..., and is thus equal in area to

$$\frac{1}{2}(ab \cdot AH + bc \cdot BK + \cdots) < \frac{1}{2}(ab + bc + \cdot) ,$$

where A H, B K,... (Fig. 169) are the altitudes of these triangles, and denotes the greatest of these altitudes.

The factor a b + b c + ••• approaches the length of the circle, while the altitudes A H, B K,... all tend to zero. Indeed, the lengths OH, OA , for instance, both tend to the radius of the circle, and therefore their difference A H tends to zero.

Thus, the areas of inscribed and circumscribed polygons tend to a unique limit S which is the area of the circle.[1]

261. Theorem. The area of a circle is numerically equal to the length of the circumference, multiplied by half the radius.

Indeed, the area of a regular convex polygon is the product of its perimeter and half the apothem. As the number of sides increases indefinitely, the perimeter approaches the length of the circle, and the apothem approaches the radius.

Remark. We obtain the same result using the theorem of 254 about the circumscribed polygon. Indeed, the perimeter of such a polygon will approach the length of the circle as the number of sides increases indefinitely.

Corollary. The area of the circle with radius R is R^2.

Indeed, $2\ R \times \frac{R}{2} = \ R^2$.

262. A circular sector is a portion of the plane bounded by a circular arc and the radii corresponding to its endpoints.

The area of such a sector is the limit of the areas of inscribed polygonal sectors as all the sides decrease indefinitely. The existence of this limit is established as in the case of the circle.

Theorem. The area of a circular sector is equal to the arc which bounds it, multiplied by half the radius.

Indeed, the perimeter of the regular inscribed broken line approaches the length of the arc, and the apothem approaches the radius.

The length of an arc of n gradients in a circle of radius R is $\frac{nR}{200}$, and therefore the area of the corresponding sector is $\frac{R^2 n}{400}$.

Analogously, the area of a sector of a circle with radius R, corresponding to an angle equal to $m\ n\ p$ will be

$$\frac{R^2}{360}\ \ m + \frac{n}{60} + \frac{p}{3600}\ \ .$$

263. Areas bounded by circular arcs. A segment of a circle (Figures 213 and 214) is the area contained between an arc and its chord.

Clearly, the area of a segment is obtained from the area of the corresponding sector by subtracting the triangle with the chord as base and the center as its vertex, provided that the arc is less than a semicircle (Fig. 213), and by adding the same triangle in the opposite case (Fig. 214).

More generally, the calculation of the area of a region bounded by arcs of circles can be reduced to that of a polygonal area by adding or subtracting sectors or segments of circles.

---

[1] The distance ab approaches zero, independent of the position of arc $\overset{\frown}{ab}$ on the given curve. This condition is satisfied if the analogous condition in the note to 179 is satisfied. We are then assured that the difference between the areas of the inscribed and circumscribed polygons approaches zero as the sides of these polygons approach zero.

As has been already indicated in the note mentioned above, the argument can be divided into two parts, corresponding to sections 259 and 260 in the text.

As for non-convex areas, they may be evaluated as sums or differences of convex ones.

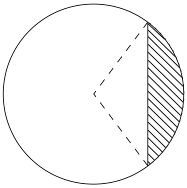
Figure 213

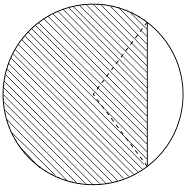
Figure 214

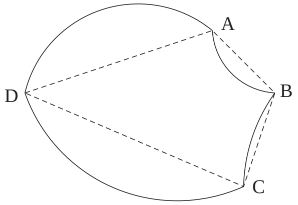

Figure 215

For instance, the curvilinear area ABCD (Fig. 215) is equal to the area of the ordinary quadrilateral ABCD, diminished by the segments AB, BC, and increased by the segments CD, DA.

In conclusion, we know how to find the area of any portion of the plane bounded by lines and circular arcs.

Exercises

Exercise 312. What is the radius of a circle with an area of one square meter?

Exercise 313. What is the radius of the circle in which a sector of $15^{G}25$ has an area of one square meter?

Exercise 314. What is the radius of a circle in which the segment contained between an arc of 60 and its chord has an area of one square meter?

Exercise 315. Prove that the area of the annulus bounded by two concentric circles is equal to the area of a circle whose diameter is the chord of the larger circle which is tangent to the smaller one.

Exercise 316. On a quarter circle AC we take points B, D at equal distances from the endpoints. We then drop perpendiculars BE, DF to radius OC. Show that the curvilinear trapezoid BEFD is equivalent to sector OBD.

Exercise 317. On the legs of a right triangle as diameters, we draw semicircles outside the triangle. We draw a third semicircle on the hypotenuse as diameter, but this time on the same side of the hypotenuse as the triangle. Show that the sum of the two lunes, or crescents, contained between the smaller semicircles and the large one is equal to the area of the triangle.

Exercise 318. On side AB of a square inscribed in circle O, we draw a new semicircle, outside the square. Radius OMN intersects this semicircle at N and the original circle at M. Show that the area bounded by segments MN and by arcs $\widehat{MA}$, $\widehat{NA}$ can be squared; that is, we can construct, with straightedge and compass, a square equivalent to it.

Note A:
On the Methods of Geometry

268. We would like to collect here some advice which we believe to be very useful for the understanding of mathematics in general and, in particular, for solving problems.

We must convince students that they will not be able to benefit from their study of mathematics, will not be able to pursue this study without excessive effort, and will not be able to form a proper idea of what geometry consists in, unless they not only understand the proofs they read, but also construct for themselves, to a greater or lesser extent, proofs of theorems or solutions to problems.

Contrary to common prejudice, this understanding can be achieved by anyone, or at least by anyone willing to think and to direct their thoughts methodically. The precepts which we will indicate derive from the most ordinary common sense. There is not one among them which may not seem a pure banality to the reader. However, experience shows that forgetting one or another of these obvious rules is the cause of almost every difficulty encountered in the solution of elementary problems; this is also true, more often than one might think, in research in more or less advanced parts of the mathematical sciences.

(a) Theorems to Prove

269. To prove a theorem means to pass, by way of reasoning, from the hypothesis to the conclusion.

Consider the theorem (Book I, 36): Every point on the bisector of an angle is equidistant from the two sides of this angle. The hypothesis and conclusion are: Hypothesis: If the point M is on the bisector of the angle BAC (Fig. 35) Conclusion: Then it will be equidistant from AB and AC.

We have to deduce the second from the first, to transform the properties stated in the hypothesis, so as to derive the truth of those which constitute the conclusion.

Obviously, it is necessary to know very precisely what the hypothesis is and what the conclusion is, of the theorem we want to prove. The student must therefore practice stating them without any hesitation.

270. Before we continue, we can already make an important observation. In every proof, one proposes to show that the conclusion holds when the hypothesis is supposed true. If the hypothesis is not considered certain, nothing can assure us of the truth of the conclusion. Thus, in the preceding example, if the point M were not on the bisector, we in fact know (36) that it would not be equidistant from the two sides.

It is clear that it would not serve any purpose to suppose that something is true, if the reasoning carried no trace of this assumption, if this assumption were not used at some point. We see that the hypothesis must be used in the proof and even, generally, the entire hypothesis.

271. We will return a little later to the preceding rule. But first, we have an immediate need to formulate another, analogous rule. We must call special attention to it, since, while indispensible, it is often misunderstood. It refers to the definitions of the terms used.

On the one hand, it is obvious that we could not know how to reason about notions which have not been defined; on the other, as before, it is clear that it is the same thing to ignore a definition, or not to use it in our reasoning.

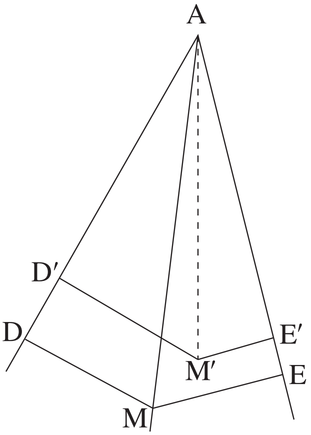

Figure 35

Therefore it is of paramount importance that we refer to the definitions of all the notions involved in the situation at hand.

Example. Consider the same theorem:

Every point on the bisector of an angle is equidistant from the two sides of this angle.

We must start by asking the following questions:

What does it mean that the line MA is the angle bisector?

ANSWER: That it divides the angle A into two equal parts.

What is the distance from M to line AB?

ANSWER: It is the length of the perpendicular dropped from this point to this line.

So our statement becomes:

Hypothesis: $MAD = MAE$; MD is perpendicular to AD; ME is perpendicular to AE.

Conclusion: $MD = ME$

We hope that this example will clarify sufficently the meaning of this rule, which Pascal places at the basis of all logic: we must substitute the definition in the place of the term which it defines.[1]

---

[1] It is even necessary, in general, to use all the parts of the definition when there are several. We may note in this respect the remarks made, a little later, about the hypothesis (275).

272. The definition of the same term can sometimes be given in several forms, among which we must choose the one which is the most convenient for the purpose we have in mind. Thus, we can define the bisector AM of the angle as the line which forms an angle equal to one half of the original angle with one of its sides (in the appropriate sense of rotation).

This formulation of the definition would not have been advantageous for the preceding theorem. On the other hand, it is the form which served best, for instance, in the proof of the theorem in 16 (Book I).

Certain theorems allow us, in the same way, to replace a definition with an equivalent one. Thus, the original definition of parallels (38) is no longer used in the original form of 39, where we learned to replace it by the following form, which is exactly equivalent to the first: Two lines are parallel if they form equal alternate interior angles with the same transversal (or equal corresponding angles, or supplementary interior angles on the same side).

273. The rule just stated is perhaps the most important one that we are going to discuss. Its importance is apparent when we think of the great number of auxiliary constructions, which sometimes seem arbitrary, but which are a direct consequence of this rule.

To give just one example, at the beginning of Book II, when we reason about an arbitrary point on a circle, we always start by joining it to the center. The reader who has reflected on the preceding remarks will understand that there is nothing artificial about this construction: its necessity is immediately apparent. In fact, it comes from the definition of a circle, according to which, if we say that a point M belongs to a circle with center O, we are saying that the distance OM is equal to the radius of the circle.

Starting with Chapter IV (73 and later) we see things change: it is not always the case that a point on the circle must be joined to the center in order to reason about it. We have now learned to replace the original definition of the circle with another, given in 82. Using this definition to express the fact that a point M belongs to a circle, we can join M to three points A, B, C on the curve, and show that quadrilateral $ABCM$ is cyclic, using one of the conditions listed in 81. From this point on, in every proof involving a circle, we can choose between these two definitions. The one which we use depends on the circumstances.[2]

274. Having been cautioned by what we have just mentioned, we must, as we have said, transform the given hypothesis so as to find in it evidence for the conclusion.

In the simplest cases, we can find a theorem which allows us to make such a transformation.

Example. The hypothesis we discuss in 271 immediately furnishes the corresponding conclusion, through one of the cases of congruence for right triangles.

In other cases, however, we must pass through one or several intermediary transformations. For example, we can seek to state the hypothesis in a form as close a possible to the conclusion.

Example. Suppose we want to prove the theorem (25): In any triangle, the greater angle is opposite the greater side.

---

[2] Later, we encounter other definitions, equivalent to the original; for instance, in Book III, 131.

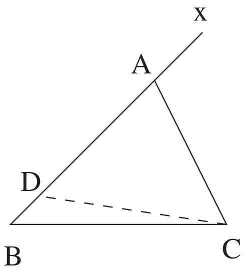

Figure 217

We need to express the fact that AB is greater than AC, so we take point D on AB, between A and B, such that $AD = AC$. The hypothesis and conclusion are therefore

(I) Hypothesis: DA is the extension of DB, $DA = AC$.

Conclusion: $ACB > ABC$.

Using isosceles triangle ADC, in which the base angles must be equal, allows us to reformulate the hypothesis in a new way.

(II) Hypothesis: DA is the extension of DB, $ADC = ACD$.

Conclusion: $ACB = ADC + DCB > ABC$.

Or, more simply,

(III) Hypothesis: Dx is the extension of DB (Fig. 217).

Conclusion: $xDC + DCB > xBC$,

which is obvious from the theorem on the exterior angle of a triangle (same section).

We have arrived at the result by a series of successive transformations.

275. In performing these transformations, it is especially important that we not neglect our first rule (270) and that we examine whether any part of the hypothesis remains unused or has been abandoned. In this way we can make sure that the new hypothesis is the same as the old, or that the two are entirely equivalent.

Example. In the preceding example, form (II) of the hypothesis is completely equivalent to form (I). In other words, if hypothesis (I) is true, then so is hypothesis (II), and conversely. Indeed, the only difference is that the equality $AC = AD$ has been replaced by $ADC = ACD$. But we know that either of these conditions implies the other. Thus hypothesis (II) can be entirely substituted for hypothesis (I): to assume one or the other amounts to the same thing.

While it may happen that an element of the hypothesis can be abandoned without consequence,[3] this is not generally the case.[4] When we find ourselves blocked in the course of a proof, we must ask ourselves whether the difficulty in reaching the goal is not due to the fact that we have lost part of the given hypothesis on the way.

---

[3] This is what happens in the preceding example with hypothesis (III) and hypothesis (II). Hypothesis (II) is exactly equivalent to (II′): Dx is the extension of DB, and there is a point on Dx such that the triangle with this vertex, and with base DC, has equal angles at the base. Indeed, such a point could obviously be labeled A. But this point might not exist, even if hypothesis (III) (which states that Dx is the extension of DB), is satisfied. For such a point to exist, the angle xDC must be acute, according to the exterior angle theorem.

Thus hypothesis (III) can hold, without hypothesis (II) being true.

[4] One must take care to state results in a form in which the hypothesis does not contain any useless element. In the most advanced research, and particularly in applications of mathematics,

Example. Suppose we must prove the following theorem:

Given point M in the plane of triangle ABC, we construct angle BAP = MAC and length AP = AM. Likewise, we construct CBQ = MBA, BQ = BM and ACR = MCB, CR = CM. Then points M, P, Q, R lie on the same circle (Fig. 218).

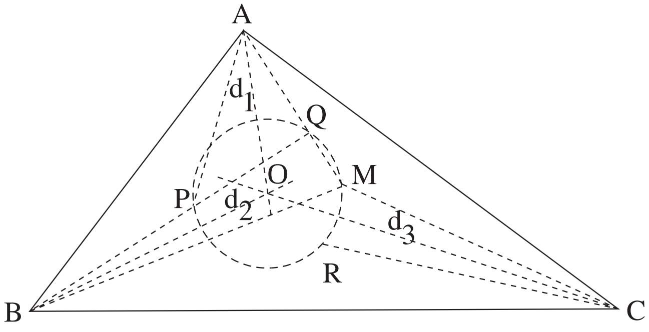

Figure 218

In other words,

(I) Hypothesis:

$$BAP = MAC, \quad AP = AM$$
$$CBQ = MBA, \quad BQ = BM$$
$$ACR = MCB, \quad CR = CM.$$

Conclusion: M, P, Q, R lie on the same circle.

Let d_1 be the bisector of angle A. Then sides AB, AC are symmetric with respect to this line, and because BAP = MAC, we see that AM and AP are also symmetric with respect to d_1. Thus point P is symmetric to M with respect to d_1. Likewise, Q is symmetric to M with respect to the bisector d_2 of B, and R is symmetric to M with respect to the bisector d_3 of C. Thus we are tempted to transform the hypothesis as follows:

(II) Hypothesis:
P is symmetric to M with respect to d_1,
Q is symmetric to M with respect to d_2,
R is symmetric to M with respect to d_3.

Conclusion: M, P, Q, R lie on the same circle.

However, if we started from this form of the statement, it would be impossible to find a proof; indeed, the proposition thus formulated is false. It is not true that a point M and its symmetric points P, Q, R with respect to three arbitrary lines

---

the greatest difficulty often consists in recognizing, among the given elements of the problem, those that will be useful.

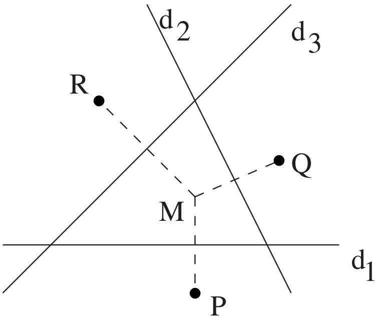

Figure 219

are on the same circle. We can see this simply by inspecting figure 219, or by noting that any three points P, Q, R can be viewed as symmetric to a point M with respect to three lines d_1, d_2, d_3; namely, the perpendicular bisectors of MP, MQ, MR.

Thus we must have made an error in substituting form (II) of the hypothesis in place of (I). The error consists in the fact that lines d_1, d_2, d_3 are not arbitrary: they are the angle bisectors of the given triangle and, as such, are concurrent. The correct form of (II) is then[5]

(II) Hypothesis:

P is the point symmetric to M with respect to d_1,
Q is the point symmetric to M with respect to d_2,
R is the point symmetric to M with respect to d_3;
lines d_1, d_2, d_3 are concurrent at point O,

and this leads immediately to a proof, because we see that the four points M, P, Q, R are on a circle with center O.

276. Instead of transforming the hypothesis in order to make it more similar to the conclusion, it is often advantageous to operate on the conclusion, and to replace it by another which implies the original, and which can be deduced more easily from the hypothesis.

Example. Suppose we want to demonstrate the following theorem (Exercise 72):

From a point M on the circumscribed circle of triangle ABC, we drop perpendiculars MP, MQ, MR to the sides of the triangle. Then the feet of these perpendiculars are on a straight line.

We will have proved that P, Q, R are collinear (Fig. 220) if (joining PQ, PR) we show that the angles BPR, CPQ, which are opposite and have the same vertex, are equal. Thus, if the statement is:

Hypothesis: Points A, B, C, M are on the same circle;
MP, MQ, MR are perpendicular to BC, CA, AB, respectively,

we can give the conclusion the form

Conclusion: BPR = CPQ.

---

[5] This hypothesis can indeed be substituted for the original one: three concurrent lines can generally be viewed as the bisectors of the angles of a triangle (Exercise 38).

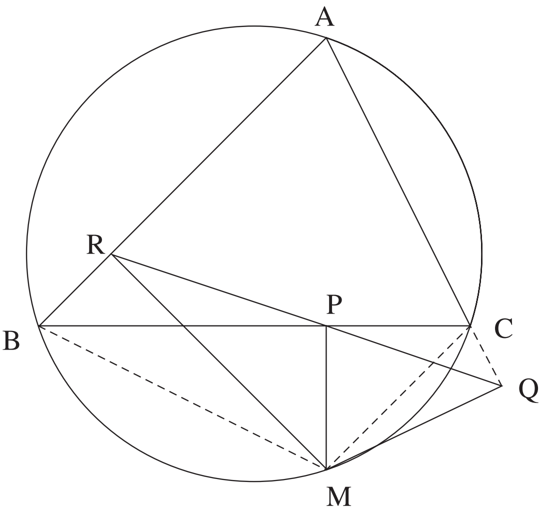

Figure 220

But, because of the right angles BRM, BPM, quadrilateral BRPM is cyclic (81), so that BPR = BMR; likewise, quadrilateral CQMP is cyclic, so that CPQ = CMQ. It will suffice to establish (with the same hypothesis):

Conclusion: BMR = CMQ.

The original conclusion is therefore replaced by another, whose proof is simpler and can easily be found by the reader.[6]

277. We now return to an important remark which we were constrained to neglect, but which must be applied to the reading of any statement of a problem.

This is the observation that many theorems can be stated in several different forms. We have already noted some examples in the text.

Example I. We have observed in 32 that the proposition:

Every point equidistant from two points A, B is on the perpendicular bisector of AB.

can also be stated as follows: A point which is not on the perpendicular bisector of AB is not equidistant from A and B.

We have seen that this is an instance of a general principle: the contrapositive of an arbitrary proposition is equivalent to the original proposition.

This idea is related to the mode of proof by contradiction which consists in showing that if we assume that the hypothesis true, and at the same time the conclusion false, we are led to a contradiction.

Example II. Consider the statement (23):

The bisector of the angle at the vertex of an isosceles triangle is perpendicular to the base, and divides it into equal segments.

---

[6] We have given an argument based on the diagram as drawn in fig. 220. We can modify this proof to make it independent of this particular position. This is easily done using the remarks in 83. However, in solving the problem itself, we can always arrange the argument to use the case we have described; it suffices to change, if necessary, the order of the letters A, B, C.

This statement amounts, as we have seen,[7] to either of the following statements:

The altitude from the vertex falls on the midpoint of the base, and divides the angle at the vertex into two equal parts.

The perpendicular bisector of the base passes through the vertex, and is the bisector of the angle at the vertex.

Clearly this example, like the preceding one, is an instance of a general principle. We have encountered this principle, for instance, in 63, Book II. Indeed, it is encountered at the very beginning of geometry. The proof of 41 (the converse of the theorem of 38) is obviously just a remark of this kind.

These are two general categories of situations in which a statement can be replaced by an equivalent statement, but they are not the only such situations. Some reflection should, in any particular case, help find the different forms which a proposition can take. It is essential to review these forms, in order to choose the one which lends itself the most to proof; in other words, to ask the question in the form in which the solution is easiest to find.[8]

278. This last observation concludes the exposition of the fundamental rules which we wanted to indicate. It will be useful to study the proofs given in the book from the point of view of the application of these principles, and to ask questions such as the following:

To prove the theorem on the medians of a triangle (56), we have taken the midpoints of BG and CG (Fig. 53). Was it logically necessary to think of this construction? Could it be replaced by others?[9]

In Exercise 8, do both parts of the conclusion assume that the given point is inside the triangle? Does the answer to this question indicate which theorem must be used for the proofs of the two parts?

In the proof of 27, at which point do we use the fact that the enveloping polygon is convex?

Etc., etc.

(b) Geometric Loci Construction Problems

279. What we have just said about the proof of theorems allows us to shorten our remarks on other possible types of questions, whose solutions must be sought according to the same principles. We are about to see this in problems concerning geometric loci and construction.

Geometric loci. In the text, we have learned how to find a number of geometric loci: the locus of points equidistant from two given points, or two given lines, the locus of points at a given distance from a given line, etc.

Other geometric loci are obvious on their own. For instance, the locus of points with the property that the lines joining them to a fixed point A are parallel to a fixed line xy, is simply the parallel to xy passing through A.

Having seen this, if we are given a property of a point M or, more generally, properties of a variable figure containing this point, the locus of M can be found by transforming the given properties into others which correspond, for this point, to a known locus.

---

[7] We recall (see 41) that this equivalence depends on the fact that there is only one bisector of the angle at the vertex, only one altitude, only one perpendicular bisector of the base, etc.

[8] One should give the problem a form in which it can always be solved (Abel).

[9] One of these is given in Exercise 37.

We are thus faced with a question analogous to the one encountered in proving a theorem. In that case, we had to deduce, from a collection of given properties (the hypothesis), another collection, also provided (the conclusion). Here the same sort of transformation must be performed. The only difference is that now, while the starting point is given, the end point, or conclusion, is not. We only know that it must be of the sort that will furnish the required locus. It is clear that we must look for common properties of the various positions of the variable figure and, particularly, the elements of this figure which do not change as the figure varies.

The path we must follow is the same as in the preceding case, and we would only need to repeat here the precepts just formulated.

One of these applies even more strictly here than in the case of the proof of a theorem. For proofs, we have seen that that part of the hypothesis may be abandoned in our successive transformations. This can never be allowed in searching for a geometric locus, because the locus must contain the points possessing the given properties, and only these. We must therefore make sure that the conclusion is exactly equivalent to the hypothesis.

This is in fact what we have done for most of the loci obtained in the text (33, 36, 77, etc.); we have omitted this second part of the reasoning only in those cases where it was so easy to supply that it was not necessary to do so explicitly.

280. Construction problems. Suppose now that we must construct a figure with given properties. The nature of the answer to such a question can vary significantly, depending on whether or not the given conditions are enough to determine exactly the unknown figure.

Examples. (1) Construct a line tangent to a given circle.

The tangent at an arbitrary point on the circle will solve this problem. There are infinitely many solutions: the given conditions do not suffice to determine the figure. The problem is said to be underdetermined.

(2) Construct a line tangent to two given circles.

Here we have an additional condition. This time the problem is determined: there are (at most) four solutions (93).

(3) Construct a line tangent to three given circles.

In general there is no line solving this problem. The first two circles have (at most) four common tangents, so the required line can only be one of these four; but an arbitrary third circle, will in general be tangent to none of these lines. Thus the problem is, except in special cases, impossible. Too many conditions are required of the figure described.

The problems proposed for students will, generally, be determined.

281. The problem often reduces to the construction of a single point.

Example I. Construct a circle passing through three given points (90).

It is enough to determine the center.

Example II. Construct a triangle knowing a side, the opposite angle, and the corresponding altitude.

If we place the given side arbitrarily, we need only find the opposite vertex.

In this situation, the method generally used is the construction by intersecting geometric loci, which consists in deducing from the given conditions two curves on which the unknown point must be found: the intersection of the curves will locate this point.

Example I. In order to find the center of the circle passing through three points A, B, C, it was enough to observe that the condition that this center be equidistant from A and B provides a first geometric locus, and the condition that it be equidistant from B and C a second locus.

Example II. In order to construct a triangle ABC knowing the side BC, the angle A, and the corresponding altitude, having chosen a location for BC, we have two loci for A: 1 The circular arc along which segment BC subtends the given angle; 2 a line parallel to BC, at a distance equal to the given altitude.

Let us say that a condition on a point is a simple condition if the set of points satisfying the condition is a curve. A point is then determined by two simple conditions, and if the two loci corresponding to the simple conditions are known[10], the required point will be their intersection.

More generally, a figure is determined by a certain number of simple conditions (we will return to this point in the second volume, Chapter X, and Note E). A polygon with n sides is determined in size and position by $2n$ simple conditions, since there are n points to be determined. To determine the polygon in size and shape only, $2n - 3$ simple conditions will suffice, since we can choose one of the vertices, and the direction of one of the sides ending in that vertex, both arbitrarily. This number $2n - 3$ is equal to n for a triangle, but larger when n is greater than 3: thus the remarks of 46b (Remark III) and 147.

282. According to the observation above, when we are given a construction problem, we try to change it so that it is reduced to questions whose solution we know; for example, by deducing two geometric loci for one of the points of the required figure.

Thus, one can consider a figure which is assumed to satisfy the given conditions,[11] and these conditions will serve as our hypothesis. We will then have to find the conclusion and, for this purpose, we will have to ask what the known elements in the figure are.

In this instance again, the general rules are the same as in the proof of a theorem. As in the case of a geometric locus, the conclusions we obtain must be fully equivalent to the hypothesis. This is true because, on the one hand, the conditions of the problem must imply the construction we find, and, on the other hand, every figure formed by this construction must satisfy all the conditions in question.

We will not dwell any longer on the particular features presented by construction problems, but simply refer the reader to the excellent work Methods and Theories for the Solution of Construction Problems in Geometry by Petersen,[12] which we have found very useful in the preparation of this volume.

---

[10] In elementary geometry, we limit ourselves to the solution of construction problems with straightedge and compass, so that the loci in question must always be straight lines or circles.

[11] This is what is meant, for instance in 93, by saying that we assume the problem solved.

[12] Translated by O. Chemin, Gauthier-Villars, 1880

(c) The Method of Transformations

283. The student who is familiar with the advice in the preceding sections, who can substitute, more or less mechanically, the defined terms for the definitions, who knows how to find quickly the various forms in which a problem he is trying to solve can be presented, will soon be able to solve a great number of problems posed in elementary geometry. Other problems, however, will seem to him inaccessible or very difficult, although they admit in reality of very simple solutions. It is just that these solutions do not depend simply on the direct form of reasoning which we have discussed, but require the aid of simplifications which we have yet to discuss: the method of transformations.

Properly speaking, after the preceding discussion, all geometric methods can be legitimately called “transformation methods”. However, we reserve this name for methods which consist in passing from the properties of a figure to the properties of another figure.

To define a transformation is to associate to each figure another figure, following a certain rule, such that the second figure is completely determined by the first, and conversely. From each property of one of the figures one can derive a property of the other, which is, in some sense, a restatement of that property.

Examples. We have defined a figure F' homothetic to a figure F, with respect to a given center of homothecy and ratio of similitude. Given figure F, we can construct figure F'. We have seen that to each line of F there corresponds a parallel line in F', to each triangle in F there corresponds a similar triangle in F', to each circle in F there corresponds a circle in F', and so on.

In the same way, given a figure F, we can construct a figure F' derived from the first by a rotation, or a translation, or a given line reflection; or more generally, a figure F' similar to F, and such that the points corresponding to two given points A, B are two other given points A', B'. The properties of F immediately imply corresponding properties of F'.

284. In some cases, a transformation need not be applied to the entire figure. It is often useful to transform only part of a figure.

This happens, in particular, with the simple transformations mentioned above: translation, line reflection, homothecy, or a general similarity. There is usually no advantage in applying such a transformation to the entire figure, since the properties of the transformed figure are neither simpler nor more complicated than those of the original figure: they are exactly the same.[13] On the other hand, it is often necessary to apply one of these simple transformations to a certain part of the figure.

Example I. Consider the problem (Exercise 32): Given two parallel lines, and two points A and B outside these parallels, and situated on different sides, what is the shortest broken line joining these two points in such a way that the portion contained between the parallel lines has a given direction?

Suppose the required line is $AMNB$ (Fig. 221). Point N is obtained from M by a translation which is obviously known, since the length of the segment intercepted by the given parallels on any line with a given direction is the same. We can apply this translation to segment AM; point A is changed into a known point C, and

---

[13] Geometry studies exactly those properties which are not changed by rigid motions.

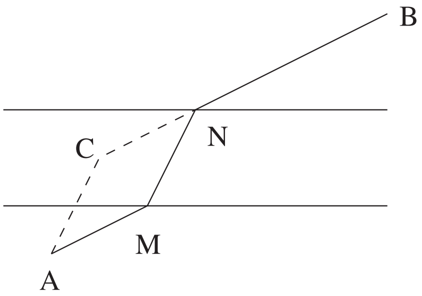

Figure 221

segment AM into an equal segment CN. We deduce easily that points B, N, C are on a straight line.

Analogous considerations apply to Exercises 118 122 (Book II).

Example II. In Exercise 14, segment AM, after reflection in the given line xy, will be collinear with the segment BM.

285. In solving certain construction problems, however, it is useful to transform an entire figure, by similarity or rigid motion. The advantage obtained by such a transformation is that it can transform unknown parts of the figure into known parts of the transformed figure.

Example. In quadrilateral ABCD, inscribe another quadrilateral similar to a given quadrilateral mnpq.

There exists a figure similar to the one required, in which the required quadrilateral MNPQ corresponds to quadrilateral mnpq. Quadrilateral ABCD will correspond, in this similar figure, to a quadrilateral abcd, circumscribed about mnpq, and which can be constructed[14] by the solution of Exercise 213.

286. Other transformations that we have encountered change the figures to which they are applied in a more or less remarkable way.

An example is transformation by inversion, which changes a line of the original figure into a circle of the transformed figure. Another example is transformation by reciprocal polars. These transformations have been used in the complements to Book III to simplify the proofs of a number of theorems.

Another transformation, which we have only alluded to, but must be mentioned, is perspective. As we have defined it in plane geometry, it is only applied to figures composed of points on a straight line. We obtain the transformed figure by joining every point on the original figure to a fixed point, outside the line, then cutting the rays obtained with another line, then subjecting the figure obtained on this transversal to an arbitrary displacement.

287. It is appropriate to note an essential difference which distinguishes the transformation by polars from others considered above, such as similarity, inversion,

---

[14] The author has in mind here the following construction: using the method of Exercise 213, construct a quadrilateral similar to ABCD with sides passing through points m, n, p, q. Then the required points M, N, P, Q are those that correspond to m, n, p, q under the similarity in question. transl.

or perspective. These last are point transformations, in the sense that each point in the original figure corresponds to a definite point in the transformed figure. This is not true of a transformation by polars, in which a line in one figure corresponds to a point in the other.

Another transformation used in the text, and which shares this character of not being a point transformation, is the one which may be called a dilation, and which we have applied to figures formed by circles and lines. In this operation one increases or decreases (as the case requires) the radius of each circle by a certain quantity a. It may happen, of course, that the radius of a circle becomes zero, and the circle becomes a point. Conversely, under this transformation a point can be considered as a circle of zero radius, and may be transformed into a circle of radius a. Each line in the figure will be transported in a perpendicular direction, in one sense or the other, by the same quantity.[15]

The most important property of a dilation is that two tangent curves remain tangent, after the dilation has been applied appropriately (Exercise 59).[16] It is this property that we have used in 93 and 231.

288. Reduced forms. Since the purpose of a transformation is to simplify the figure we are working with, we should try to simplify the situation as much as possible.

To this end, we remark that the various kinds of transformations just mentioned involve some arbitrary elements. For example, for a homothecy we must specify the center and the ratio of similarity; for an inversion, we must specify the pole and the power;[17] for a dilatation, we must specify the quantity a discussed in the previous paragraph; etc.

Each type of transformation therefore includes infinitely many particular transformations.

Among these we must choose a particular transformation, such that the transformed figure satisfies one or several required conditions, and in general we want it to satisfy as many of these conditions as we can, taking into account what we noted in 280. The transformed figure will then be called a reduced form of the original figure.

Example I. We can always subject a circle to a dilation which reduces it to a point; it suffice to dilate by a quantity equal to the radius of the circle.

A reduced form, reduced by dilation, of a figure containing a circle, is one in which this circle is reduced to a point. We have used such a reduced form in 93 and 231.

Example II. Two circles with a common point can be transformed by a single inversion into two lines (by taking the common point as the pole). Two circles with no common point can be transformed into concentric circles (Exercise 248). In other words, the reduced form of a figure consisting of two circles is made up of two lines or two concentric circles, according as whether the original circles had or did not have a common point.

---

[15] A dilation of a circle or a line can be performed in two different ways, between which one can choose according to each particular case

[16] Dilation can be extended (using methods which go beyond the framework of this book) to figures which contain arbitrary curves, and it can be proved that this extended transformation still preserves the property of two curves being tangent. This property also holds for the transformation by polars, generalized in an analogous fashion. (See Volume II (Solid Geometry), 771.)

[17] We have seen (215) that the choice of the pole is the more important.

Example III. In the example of 285, we have used a reduced form of the required figure. The figure was reduced using a similarity, and the reduced form was determined by the requirement that quadrilateral MNPQ be transformed into the given quadrilateral mnpq.

289. Invariants. The advantage of a transformation can prove illusory, if, by simplifying certain properties of the figure, we have made others more complicated. We should apply a transformation only in those cases where the various properties in the statement are sufficiently simplified.

As it happens, in almost all the categories of transformations reviewed above, there are certain properties which are not changed at all, which remain, as we say, invariant.

Example I. We have seen that in a homothecy, the angles of the transformed figure are the same as those in the original figure. Likewise, the ratios of lengths do not change, and so on.

Example II. We have seen that a dilation changes tangent circles into tangent circles. Thus tangency is an invariant property for dilation.

Example III. The angle of two curves is an invariant for inversion (219).

Etc.

290. There can be no inconvenience in applying a transformation if only invariant properties are mentioned in the statement of the problem. In these cases, we can always assume the figure is in its reduced form.

Examples. Since dilations preserve tangency, we can assume, in constructing the common tangents to two circles, that one of the circles is reduced to a point. We followed this path in 93.

The same observation applies to the construction of a circle tangent to three given circles (231).

Tangency is also invariant for inversion. Thus, when searching for circles tangent to three given circles, we may assume that two of them are reduced to two lines or two concentric circles (Exercises 264–265).

Similar observations apply to the proof of theorems such as the following: All the circles which intersect two given circles at constant angles, are tangent to two fixed circles; or to the result of Exercise 285, etc.

291. Groups.[18] The fact that the various transformations we discussed possess invariants is a consequence of another property about which we will say a few words.

The transformation equivalent to the successive[19] application of two transformations is called the product of these two transformations. In other words, if the transformation S changes figure F into figure F', and the transformation T, applied to figure F , changes it into figure F , the product of these transformations will be the transformation that changes F to F .

---

[18] The final part of this note is addressed especially to those readers who have studied the complements to Book III.

[19] In general, the product of two transformations depends on the order of the factors. Thus, the product of two reflections in two different lines is a rotation or a translation which is double the rotation or translation which would make the first line coincide with the second.

Example. The result of 102 can be stated as follows: The product of two line reflections is a rotation or a translation.

Having said this, a set of transformations is said to form a group if the product of any two of them is also a member of this set.[20]

For example, the set of all homothecies is a group: this amounts to saying that two figure homothetic to a third are homothetic to each other. The set of all homothecies whose centers are on a given line also form a group because, as we have seen, the center of the product of two homothecies is on the line passing through their centers.

The set of all rigid motions form a group, because two plane figures, congruent to a third and with the same orientation, are also equal.

292. The collection of inversions does not form a group. The product of two inversions is not an inversion, but we may consider the group of transformations obtained by successive application of several inversions.

For the sake of brevity, let us call the product of any number of inversions, or line reflections, an S-transformation. This category includes, in particular, all rigid motions, because each rigid motion is the product of two line reflections; all homothecies, because a homothecy results from the application of two inversions with the same pole and different powers. Therefore all similarities are also S-transformations since a similarity is a homothecy, possibly followed by a rigid motion and a line reflection.

It would seem, from their definition, that to obtain all S-transformations we should consider all products of n inversions, n denoting an arbitrary positive integer, and that stopping at some fixed value of n would not suffice.

This however is not the case: every S-transformation amounts to an inversion followed or preceded by one, two, or three line reflections, except for the case when the transformation is a similarity (Exercise 252). Therefore, four simple operations (inversions or line reflections) suffice to produce any S-transformation.

Since the product of two S-transformations is again an S-transformations, the S-transformations form a group which one may call, for the sake of brevity, the group of inversions.

293. Two figures F, F which are changed into one another by a transformation in a group are said to be homologous with respect to this group. For instance, two congruent figures with the same orientation are homologous with respect to the group of rigid motions.

According to the definition of a group, two figures F, F , homologous to a third F , are also homologous to each other, since the transformation that passes from F to F is the product of the transformations which pass from F to F and from F to F .

The reduced form of a figure F with respect to a given group is a figure homologous to F with respect to this group, which satisfies certain additional conditions. If these conditions are sufficient in number, and well chosen, they will suffice to determine completely the reduced form F_0 of F.

---

[20] The original text makes no mention of the existence of an inverse transformation for each element of a group of transformations. transl.

Suppose this has been done. If we then apply one of the transformations of the group to F, the new figure F' obtained will also have the reduced form F_0, since the figures homologous to F' are exactly those which are homologous to F.

The reduced form of F is thus also the reduced form of any figure homologous to F, and therefore every property of the reduced form is an invariant property of F.

294. We will clarify this remark by an example.

Consider the figure formed by four arbitrary points A, B, C, D. We propose to find the properties of this figure which are invariant for the group of inversions. For this purpose, take the inverse of the figure, using an inversion I with point A as its pole. If the inverse points for B, C, D are b, c, d respectively, they form a figure homologous to the original (a reduced form) with respect to the group, in which one of the four points (the one that corresponds to A) is thrown to infinity. We will see, by retracing our general argument as applied to this special case, that the angles of triangle bcd are the invariants of the figure $ABCD$. (A direct proof of this fact is the object of Exercise 270.)

Let A', B', C', D' be the points corresponding to A, B, C, D under an arbitrary transformation T of the group. We operate on these points just as we operated on the original points; that is, we apply an inversion I' with pole A' to obtain the homologous points b', c', d'. We can pass from the figure formed by triangle bcd and a point at infinity to the figure formed by triangle $b'c'd'$ and a point at infinity by a transformation in our group; namely, the product of the three inversions I, T, I'. As we know, this transformation is either a similarity, or an inversion followed by line reflections. The second alternative occurs only if T itself is an inversion, because then it would send a point at infinity into a finite point (namely, the pole). We conclude that the triangles bcd, $b'c'd'$ are in fact similar. QED

Thus, the angles and the ratios of the sides of the triangle bcd are indeed the invariants of the figure $ABCD$ relative to the group of inversions. There are only two independent invariants, that is, knowledge of two of them is sufficient to calculate the others: two angles of triangle bcd are indeed sufficient to construct a triangle similar to it.

295. We have just seen that the existence of invariants follows from the existence of of a group; conversely, the set of all transformations which have the same invariants forms a group. Indeed, if two transformations do not change a certain property, the same is true for their product.

Example I. A homothecy changes every segment into a parallel, proportional segment. Conversely, every point transformation with this property is a homothecy (142). Thus, the collection of homothecies forms a group. The reasoning of 144 is simply a different form of the argument just made.

Example II. A perspective preserves the cross-ratio of four points and, conversely, every transformation applied to a system of points on a straight line, and which conserves the cross-ratio, is a perspective.[21]

Thus the set of perspectives forms a group.

---

[21] This can be proved by moving the transformed figure so that one point coincides with its corresponding point, while the two lines remain distinct. We are then led to Exercise 233.

We will limit ourselves to these remarks on the general properties of transformations. As for their use, we refer the reader to the work of Petersen already mentioned.

Note B:
On Euclid's Postulate

I

296. We have assumed, as an axiom (40), that through a point there cannot pass more than one parallel to a given line.

This statement appears in the Elements of the Greek geometer Euclid,[22] where the principles of geometry are described completely for the first time, and with remarkable perfection, as well as in all subsequent expositions (which are all based on this first exposition). More precisely, Euclid gives the statement of 41 (Corollary I), which is equivalent to it:

If two lines make, with the same transversal, two interior angles on the same side whose sum is different from two right angles, then the lines are not parallel, and they intersect on the side of the transversal where the sum of the interior angles is less than two right angles.

This statement is taken to be among those considered obviously true.

The followers of Euclid, from antiquity to the middle ages and into modern times, have not stopped marveling at the place assigned to this proposition, which indeed does not exhibit the same obvious character shared by other propositions accepted without proof. Indeed, it does not seem any more evident than many of the propositions in whose proof it is used. In particular (and we will soon return to this point), they found it strange that Euclid accepted as obvious for straight lines a fact which is not at all true for curves in general.[23]

297. Many attempts were therefore made to prove this postulate: all of them failed. In particular, trying to argue by contradiction, and deducing all the possible consequences if the postulate is assumed false, one obtains a series of conclusions which are very different from the conclusions of the usual theory of parallel lines. But however far we push these conclusions, they never yield (at least when they are deduced correctly) a contradiction among themselves, or with earlier propositions, a contradiction which would prove the impossibility of the initial hypothesis.

298. The great mathematician Gauss[24] asked himself whether such a contradiction did in fact exist, if perhaps the opposite of Euclid's postulate is in fact compatible with the other axioms of geometry and with the consequences which would follow. In other words, perhaps it is impossible to prove the proposition in question.

---

[22] Around 300 B.C.

[23] Wallis, 1663. Cf. Stäckel et Engel, Die Theorie des Parallelinien, von Euklid bis Gauss, Leipzig, Teubner, 1895.

[24] 1777–1855.

Around the same time Lobachevsky[25] and Bolyai[26] looked at this same hypothesis independently, and constructed a geometry having in common with ordinary geometry all the propositions preceding, or independent of, Euclid s postulate, but in which all the others statements are modified. The results arrived at in such a geometry (called non-Euclidean geometry) often take an intriguing, paradoxal form, and contradict our usual way of visualizing things. But no matter how amazing these results may seem at first, there is none whose absurdity can be proved. Here are some of the simplest examples.

In non-Euclidean geometry:

Through a point A outside a line D, but in the plane which contains them, there are infinitely many lines which do not intersect[27] D. All of these non-secant lines are situated within a certain angle with vertex A (and in the corresponding vertical angle). This angle, called angle of parallelism, increases with the distance from the point to the line;

Every line D interior to the angle of parallelism has a unique common perpendicular with line D, which gives the shortest distance between points on the two lines. If a point M moves along D and away from the foot of the common perpendicular, its distance to D increases constantly and indefinitely. The same is not true of lines D_1, D_2 which are the sides of the angle of parallelism. These lines do not intersect D, but they approach D more and more closely as we move along them away from the point A, in the direction which makes an acute angle with the perpendicular from A to D. If a point M moves along one of these lines in this direction, its distance to D tends to zero. The locus of the points in the plane which are equidistant from a fixed line is a curve.

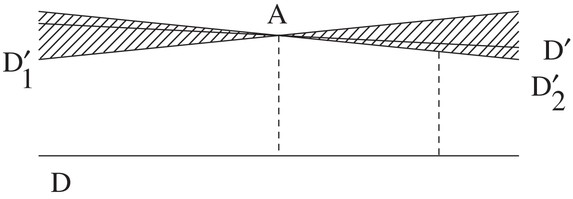

Figure 222

The sum of the angles of a triangle is less than two right angles, and the difference is proportional to the area of the triangle. It follows that this area is always less than a certain quantity, no matter how long the sides of the triangle may be (when the sides are very long, the angles become very small, and the triangle becomes very thin, as indicated in the figure). In the same way, the sum of the angles of a polygon with n sides is less than 2n 4 right angles, and the difference is proportional to the area of the polygon. Consequently, there are no rectangles: if a quadrilateral has three right angles, the fourth must be acute.

---

[25] 1793 1856.

[26] 1802 1860. Non-Euclidean geometry was also discovered, more or less completely, in a variety of places (Cf. the reference above to Stäckel and Engel).

[27] We can show that the Euclidean postulate is always true, or always false: we cannot show the existence of a single point outside a single line, for which the parallel through the point to the line is unique, without also being able to prove this uniqueness for all possible points and lines.

There are no similar figures[28] except, of course, congruent figures. Two triangles are congruent if they have equal corresponding angles. The assumption that there exist two non-congruent similar triangles is equivalent to the assumption of Euclid's postulate.
Etc.

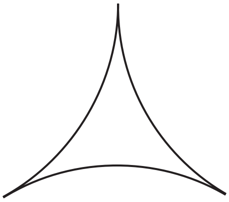

There are in fact infinitely many non-Euclidean geometries. The relations between the various parts of a figure, under Lobachevsky's hypothesis, include in fact a certain number k^2, determined once and for all, but which can have an arbitrary value: for instance the ratio (which is constant, as we just mentioned) between the area of a triangle and the supplement of the sum of its angles. Euclidian geometry can be viewed as a limiting case of the non-Euclidean geometries: it corresponds to $k =$.

299. Is non-Euclidean geometry, which seems paradoxical, but also seems to be logically possible, really possible, as believed by its advocates? It may be that this is only an impression, due to the fact that the consequences have not been pushed far enough. Would it be possible, through deeper or better investigation, to encounter a contradiction which presented itself neither to Gauss, nor to Bolyai, nor to Lobachvsky?

We can state with confidence that this is not the case. Without developing in detail the reasoning which validates this statement,[29] we will borrow a simple and striking manner of presenting them from H. Poincare.[30]

Imagine a sphere S inside of which there is a medium whose temperature and index of refraction are variable. Suppose there are objects moving in this medium, and that their their movements are sufficiently slow and their specific heat sufficiently small that they immediately achieve temperature equilibrium with the medium; moreover, all of these objects will have the same coefficient of dilation, so that we can define the surrounding temperature by the length of any one of them. Let R be the radius of the sphere, and the distance from a point in this medium to the center of the sphere: we will assume that the absolute temperature[31] is $R^2 \quad ^2$ and that the index of refraction is $\frac{1}{R^2 \quad ^2}$.

---

[28] Contrary to our usage in the rest of this book, here we are speaking of similar figures as figures with equal corresponding angles, and proportional corresponding sides.

[29] We may note an analogy between the results mentioned above and certain results of spherical geometry, to which they are opposed in a certain sense. Thus, the sum of the angles in a spherical triangle is greater than two right angles, and the difference is proportional to the area of the triangle; two spherical triangles with equal corresponding angles are either congruent or symmetric, etc. In fact, the geometry of certain surfaces (called pseudospherical surfaces) is identical to Lobachevski's geometry. It is in this way that the logical possibility of non-Euclidean geometries was established, but this proof is not complete: 1 because pseudospherical surfaces cannot be considered to be limitless in all directions, as we would expect from a plane; 2 because they only pertain to plane non-Euclidean geometry, and would not prevent in principle a proof of the postulate by considerations of geometry in space.

[30] Revue generale des sciences pures et appliques, Volume III, 1892, p. 75.

What would an intelligent being think, who had never left such a world?

(1) Since the dimensions of two small objects transported from one point to another vary in the same ratio, because they have the same coefficient of dilation, these beings would believe that these dimensions have not changed. They would have no idea of what we call differences in temperature. No thermometer could reveal it, since the dilation of the container would be the same as that of the thermometric liquid.

(2) They would believe that this sphere S is infinite: they could never reach its surface because, as they approach it, they enter colder and colder regions; they would become smaller and smaller and, without realizing it, would make smaller and smaller steps.[32]

(3) What they would call straight lines would be circles orthogonal to the sphere S, for three reasons:

1 : They would be the trajectories of light rays;

2 : Measuring various curves with a ruler, our imaginary beings would find that these circles are the shortest paths from a point to another: indeed, their ruler would contract or dilate when passing from one region to another, and they would not doubt this phenomenon;

3 : If a solid body rotated so that one of its lines remained fixed, this line could only be one of these circles: this is because, if a cylinder turns slowly and is heated on one side only, the locus of the points which do not move will be a convex curve on the heated side, and not a straight line.

It follows that these beings would adopt Lobachevski's geometry.[33]

We see now that it is in fact impossible to prove Euclid's postulate with the help of earlier propositions: if such a proof existed, it would have to be admitted by the fictitious beings just described (since all earlier propositions are still true, from their point of view); but it would lead to an incorrect result since, for these beings, the postulate is false.

II

300. What role then must we assign to this proposition, which is not as obvious as the axioms, and which cannot be proved as a theorem?

This role is that of a definition. To explain what we mean by this, we refer to the preceding note (Note A, 271).

As we have seen there, we must have a definition for each term figuring in our statements, a definition whose role in the reasoning can be made clear by substituting it for the defined term each time it occurs.

---

[31] In physics, this is the temperature measured from a zero point, so that it is proportional to the volume of a thermometric body. (In physics, it is specifically assumed that this body is a gas sufficiently far from its liquefying point; here it can be any object, since all of them have the same coefficient of dilation, and the absolute temperature is proportional not to the volume, but to an arbitrary linear dimension.)

[32] We could add that, by the properties we have assumed for the index of refraction, they could not see what happens outside the sphere.

[33] The arbitrary constant introduced in non-Euclidean geometry is represented here by the radius R of the sphere.

Of course, there are terms which have not been, and cannot be, defined, because a notion can only be defined by use of earlier notions,[34] which is impossible for the first notions to be introduced.

But, since these notions are clear in themselves and have a number of obvious properties, the role of the definition (which is still needed in each case, as we have just recalled) is replaced by these properties, which we assume without proof. We proceeded this way in the case of the straight line, which we did not actually define in the proper sense, but rather defined indirectly, by assuming its fundamental properties.

301. It is important that we assume enough of these propreties to be able to characterize the notion they define, that is, to distinguish it from every different notion. For instance, we would have given a bad definition of the straight line if, instead of assuming, as we have done:

(1) that every figure congruent to a straight line is a straight line; and conversely, two lines can be superimposed in infinitely many ways;
(2) that one and only one line passes through two points,

we would have assumed the first property, but not the second. This first property is not, in fact, peculiar to straight lines: it is also satisfied, for instances, by circles with a radius of 1 meter. It would have been impossible to prove, starting from this definition, any property of straight lines which is not also shared by these circles. For instance, we could not have proved that the sum of two angles of a triangle is less than two right angles because this is not always true for curvilinear triangles whose sides are arcs of congruent circles.

302. All geometry rests on a fundamental notion, that of a rigid motion, which we have introduced in 2. At that point, we considered as self-evident the idea of a figure which can be moved without changing its shape or size; in other words, what we have called an invariant figure. Let us try to study how this notion is defined by its properties.

Given a figure which is subject to a rigid motion, and an arbitrary point M, we can imagine that this point is permanently attached to this figure and follows it in its motion, so that it comes to occupy some position M'. We can thus say that the rigid motion associates to each point M in space another point M', which is the position to which the point M is transported. A rigid motion is therefore a point transformation[35] of space.

It is also an obvious property of the notion we consider that two rigid motions, applied successively, are equivalent to another rigid motion. In other words, these transformations form a group.[36]

We can thus say:

An invariant figure is a figure to which we apply only transformations from a certain group (called the group of rigid motions), having the following properties:

(I)

There are infinitely many transformations in the group which transport a point A onto an arbitrary position A'.

---

[34] For instance, we cannot define a circle as the locus of points in a plane situated at a given distance from a given point of this plane , without having defined earlier the notions of distance, plane, and geometric locus.

[35] See Note A, 287, 291, and the passages that follow.

[36] Ibid.

> In general, there are no transformations of the group which, at the same time, bring two given points A, B to two given positions A′, B′. In order for such a transformation to exist, a certain quantity, depending on A and B, must be equal to the corresponding quantity for A′, B′.
>
> If there is a transformation in the group which takes A, B onto A′, B′, then there are infinitely many. In particular, there exist infinitely many transformations which leave two points A, B fixed but, for all of these transformations, there are infinitely many other points which are fixed. These points form a unique line, which extends indefinitely, called a straight line. One and only one straight line passes through two points. There exist surfaces (called planes) such that every line which has two points in a plane is entirely contained in it; one such surface passes through three arbitrary points in space,

etc.[37]

303. It is nonetheless essential to observe that the preceding definition, like all other indirect definitions, implies the following axiom:

AXIOM (A): There is a group possessing property (I).

We will also remark that the notions of straight line and plane derive from the notion of rigid motion, without which they cannot be defined.

304. Euclideans and non-Euclideans agree that the group of rigid motions satisfies property (I), and they therefore assume axiom (A). They are divided by the question as to whether the following statement is true or not:

(II)

> Through a point outside a line there is only one line parallel to a given line.

This statement must then express a property of the group of rigid motions, because straight lines and planes are defined by this group. The question then takes the following form:

Does the group of rigid motions defined by property (I), also satisfy property (II)?

Now, if it is not possible to answer this question, the reason can only be that it is ill posed: that the question does not have a precise meaning.

Indeed, the group of rigid motions is not well defined by property (I). If (in agreement with axiom (A), which we assume) there exists a group with these properties, then there are infinitely many such groups. The inhabitants of Poincaré’s world would have an entirely different notion from ours of an “invariant figure”, since the objects which they move would dilate or contract in the process. However, their group of rigid motions would satisfy property (I) in the interior of the sphere (that is, for all of space about which they could reason, which is infinite from their point of view).

The resolution of the question is now clear. Statement (II) will be precise if, among all the groups satisfying property (I), we reserve the name group of rigid motions for a group satisfying the condition of this statement; in other words, if this group is defined not just by property (I), but by properties (I) and (II).

---

[37] We do not pretend to list here all the properties needed to define the group of rigid motions. By “Property (I)” we mean those properties which we have assumed in Book I, prior to Euclid’s postulate.

305. But an objection can still be raised. Just as the original definition implied axiom (A), the one we now propose is not possible unless we resolve the following question:

Does there exist a group which satisfies both conditions (I) and (II)?

The answer is in the affirmative. One proves (assuming, of course, axiom (A)) that among the infinitely many groups satisfying condition (I) there are some which are non-Euclidean (for which Euclid's postulate is false) and some which are Euclidean (for which it is true).

Now every difficulty is removed. Euclid's postulate is part of the definition of the fundamental notions of geometry.

306. Are we then saying that one should not ask whether this postulate is true or false, that such a question is totally devoid of meaning?

We would have the right to do this, if we were free to define the notions of geometry in a completely arbitrary manner. But this is not the case: these notions are provided us by experience. The idea of an invariant figure is suggested by those invariant figures (namely, the solid bodies) which nature offers. We must define invariant figures and rigid motions with these solids and their displacements in mind if we want geometry to be applicable to real objects.

307. We see therefore that there is indeed a question about Euclid's postulate: the question is whether the definition given above accords with experience, whether or not the properties of natural motions which we observe are analogous to the properties of the Euclidean group.

But this is no longer a mathematical question. The solution of such a problem does not depend on reasoning, but on observation.

If we have been led to develop Euclid's conception rather that Lobachevsky's, it is because our senses, within their limitations, show us that the postulate is reasonably exact. We see that two lines parallel to a third are parallel, that there exist similar figures with arbitrary ratios of similarity, that there exist rectangles, etc.

308. This is, of course, a crude verification. To submit it to deeper examination, we must measure, with the greatest possible precision, the angles of a triangle, to discover whether their sum is equal to two right angles. We should use as large a triangle as possible, since it is for large triangles (298) that the discrepancy between the two hypotheses is greatest. Experiments have shown (to the extent allowed by experimental difficulties) that this equality is verified (or at least that the discrepancy is smaller than observational errors).

We would then be justified in asserting that the geometry which represents reality most accurately is Euclidean or, at least, it differs very little from Euclidean (in other words, the constant k is very large, or, if the geometry is comparable to Poincaré's imaginary world, the radius R is enormous relative to all usual lengths). In one word, Euclidean geometry is physically true.

308b. Things have changed following the recent[38] developments in our ideas about physics (the Theory of Relativity).

This new theory has profoundly modified the science of motion or kinematics (which is a direct application of geometry). The models used for velocities in the old kinematics were based on principles of Euclidean geometry (mainly on properties of parallelograms). This is still assumed to be practically correct for "very small" speeds — that is, measured by very small numbers in comparison to the speed V of light (which is approximately 300000 kilometers per second) — and so for all ordinary speeds,[39] but not for speeds comparable to V. For such speeds, it is Lobachevski's geometry which corresponds to physical reality. The role of the radius R is played by V.

---

[38] That is, at the time the author was writing. — transl.

Non-Euclidean geometry (and even other, more general, geometries) enter into still other parts of the theories of physics (General Relativity). However, one can see from the preceding discussion that none of these changes has any influence on everyday life (for instance, on the work of the engineer). Euclid's geometry retains its validity for all figures which we can view at a glance, or measure directly.

---

[39] Since V is so large, the speed of a military projectile, for instance, is "very small".

Note C:
On the Problem of Tangent Circles

309. As we discussed in 236, Gergonne s method for finding the circles tangent to three given circles does not apply in all possible cases: it gives no result if the three centers are collinear. We added that this inconvenience disappears if the solution is presented in a form in which the properties utilized do not change when an arbitrary inversion is applied. This is what we propose to do in this note.

Let the given circles be A, B, C, and take point b on B, antihomologuous[40] to a point a of A, with respect to a center of similarity S_{12} of the circles A and B. Let c be the point antihomologuous in C to b, with respect to a center of similarity S_{23} of B, C. The circle passing through a, b, c cuts the three circles again in a , b , c ; it cuts them at equal angles (227), so that a , b are antihomologous with respect to S_{12}, b , c with respect to S_{23} and, moreover, the pairs a, c ; c, a are antihomologous[41] with respect to a center of similarity S_{13} of circles A, C.

Replacing a by another point a_1 of circle A, we obtain a new circle $_1$, analogous to . Point S_{12} has the same power with respect to and $_1$ (equal to the power of the inversion which transforms A into B); the same is true for point S_{23}. Thus the radical axis xy of circles , $_1$ is an axis of similarity for the given circles, and it is clear therefore that the center of similarity S_{13} which we are discussing, belongs, along with the first two, to this axis of similarity.[42]

310. We see now that there are four series of circles (corresponding to the four axes of similarity), and that the circles in each series have the same radical axis. Conversely, every circle which has a common radical axis with two circles of one of these series must belong to that series (since it corresponds to itself in the two inversions with poles S_{12} and S_{13}). Therefore each of these series is determined by two of its circles, or by one circle and the radical axis xy. We can, generally, take this circle to be the circle $_0$, belonging to the series, which is orthogonal to the three given circles (227b).

The locus of the centers of the circles in one of the series is the perpendicular from the radical center of A, B, C to one of their axes of similarity.

---

[40] See 227b. transl.

[41] The fact that a is antihomologous to c rather than c is established by noting (cf. 227) the angles made by with A, C at the points a, c have the same orientation (rather than opposite orientation, as would be the case for antihomologous points).

[42] The reasoning does assume that the point S_{13} is the same for and for $_1$. But if this were not true, at least one of the two points S_{13} corresponding to and $_1$, for instance, the one that corresponds to $_1$, would be collinear with S_{12} and S_{23}. This point would then have the same power with respect to $_1$ as to ; namely, the power of the inversion which transforms A into C. Therefore the circle $_1$ would intersect A and C in points which correspond to each other in this inversion.

311. The circles which are tangent to the given circles obviously belong to the series just mentioned and, conversely, every circle in one of these series which is tangent to one of the given circles must be tangent to the others as well.

The problem of tangent circles is thus reduced to the following:

Find a circle with the same radical axis as two given circles , $_1$, and tangent to a third given circle A (Fig. 223).

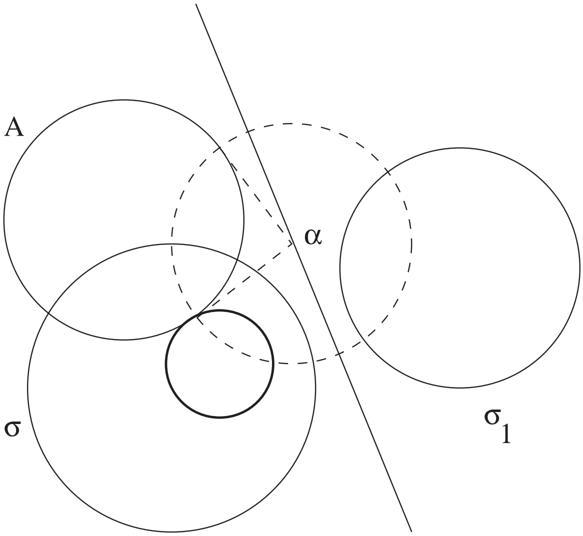

Figure 223

This last question is easily solved: if a is the point of contact of the required circle with the circle A, the circle passing through a and orthogonal to and $_1$ will also be orthogonal to the required circle, and therefore to A. These conditions allow us to find a (158, Construction 13). In other words, we have to draw a tangent to A passing through the radical center of , $_1$ and A, to find the required point of contact. Conversely, it is easy to see that the points obtained in this manner correspond to solutions of the problem.

As we remarked above, one of the circles , $_1$ can be replaced by their radical axis xy, so that the solution to the problem of tangent circles can also be reformulated as follows:

Through a point on circle A and its two antihomologous points, draw a circle . The common chord of A and intersects the axis of similarity xy at a point . The tangents from to circle A give the points of contact of the required circles.

311b. When circle is taken to be the circle $_0$ whose center is the radical center I of the three given circles, and which intersects them at right angles, we recover Gergonne s original solution. The common points of A and $_0$ are indeed the points of contact of the tangents from I to A, so that the common chord of the two circles is the polar of the point I with respect to A. The polar of the intersection point of this chord and xy is thus in fact the line joining I to the pole of xy.

We see now why this solution does not work when the centers of the given circles are collinear: both the circle $_0$ and the line xy coincide with the line of the centers. In order to remove the difficulty, it suffices, as we have said, to use a circle different from $_0$.

The method we offer also has the advantage (over Gergonne's method) of being applicable when one or several of the given circles are replaced by points or lines, yielding directly the points of contact with one of the lines: these are situated on a circle whose center is the intersection of the given line with the axis of similarity xy, and which intersects the circle $_1$ at right angles. This construction obviously generalizes Construction 14 (159). This construction even applies when there is no circle left, which is not the case for Gergonne's solution. It fails only when all three circles are reduced to points.

312. It follows from 309 that the common chord of circles A and is precisely the line we have called aa , where a is the point antihomologus to c.

Thus the construction indicated above consists in:

determining the point b antihomologous to a (which is taken arbitrarily on circle A) with respect to the center of similarity S_{12}, the point c antihomologous to b with respect to S_{23}, and the point a antihomologus to a with respect to S_{31};

joining aa ;

repeating this construction starting with another point a_1 on A.

The intersection of the two chords aa , a_1a_1 thus obtained is the point , from which we must draw tangents to circle A in order to obtain the points of contact of the required circles.

Discussion. Unlike Gergonne's form of the solution, ours allows a simple discussion of the number of actual circles which solve the problem.

We observe first that the above construction is or is not possible according as whether the point is exterior or interior to A.

Take an arbitrary point a_1 close to a. Clearly, the point will be inside A if the small arcs aa_1, a a_1 have the same sense, and outside if they have the opposite sense.

We must therefore determine whether, as point a moves along the circle A, point a moves in the same sense, or in the opposite sense.

To answer this question, we start by studying the sense in which the point b will move.

Since the point homothetic to a with respect to S_{12} clearly varies on B in the same sense as a on A, we see immediately (cf. 223) that b turns in the same sense as a if S_{12} is interior to B, and in the opposite sense if S_{12} is exterior.

Let us agree to say that a center of similarity S_{12} of two circles A, B is positive if it is outside these two circles, and negative if it is inside the circles. This amounts to saying that a center of similarity is considered positive or negative depending on whether, through this point, there are or are not real common tangents to the two circles.

If two circles are exterior, their centers of similarity are positive. If they intersect at two points, the exterior center of similarity is positive, and the other negative. If one circle is inside the other, both centers of similarity are negative.

According to this convention, whether the sense of the motion of b is the same as that of a depends on whether the center of similarity S_{12} is negative or positive.[43]

Similarly, the point c moves on C in the same sense as b on B, or in the opposite sense, according as S_{23} is negative or positive; and a moves in the same sense as c, or in the opposite sense, according as S_{31} is negative or positive.

Combining these remarks with the earlier ones, we see that is outside A in other words, there are two real points of contact if, among the three centers of similarity on the axis xy, one or three are positive.

To calculate the total number of real circles which answer the problem, it will suffice to apply this remark to each of the four groups of collinear centers of similarity.

In what follows, we will use the digit 0 to denote the group of three direct centers of homothecy; the digit 1 for the group for which B, C are directly homothetic, and for which both are inversely homothetic to A; and the digits 2 and 3 for the analogous groups obtained by replacing A by B, and then by C.

312b. Having said this, it is easy to see that (excluding the cases when two of the circles are tangent[44]) there are eleven possible positions for the three circles.[45]

In each case, the preceding considerations would immediately yield the result. For instance, in case I, all the centers of similarity are positive, so that there are eight solutions, while in case V, there are only four, because the centers of similarity of A with B or C are all negative. We would have to use the positive center of B and C, (that is, the exterior center of similarity) in order to obtain real solutions.

We thus arrive at the following table which indicates, in each of the eleven cases, the positive and negative centers (the letters S indicate the exterior centers of similarity, and the letters S the interior ones), the number of solutions, and the homothecy groups (numbered 0, 1, 2, 3, as above) which produce them.

---

[43] That is, the sense of motion is the same if the center of similarity is negative, and opposite if the center of similarity is positive. transl.

[44] These must be considered as limiting cases of those where the circles have 0 or 2 intersections, between which they are intermediary. If two of the given circles are tangent, then two of the solutions will coincide.

[45] In each of the positions shown in Figure 223b, the letters A, B, C can be permuted. The different figures which then correspond to these permutations will allow the same number of solutions, which is why we have kept only one of them in our enumeration.

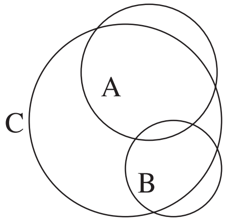

Figure XI

Figure XI is meant to include other possible positions in which the intersections of A and B, for instance, are both interior to C (instead of one inside and one outside, as in Fig. 223b). This would not change the result of the current discussion. The same remark could be applied to other problems, such as finding the circle orthogonal to A, B, C. These are numbered I to XI in Figure 223b.

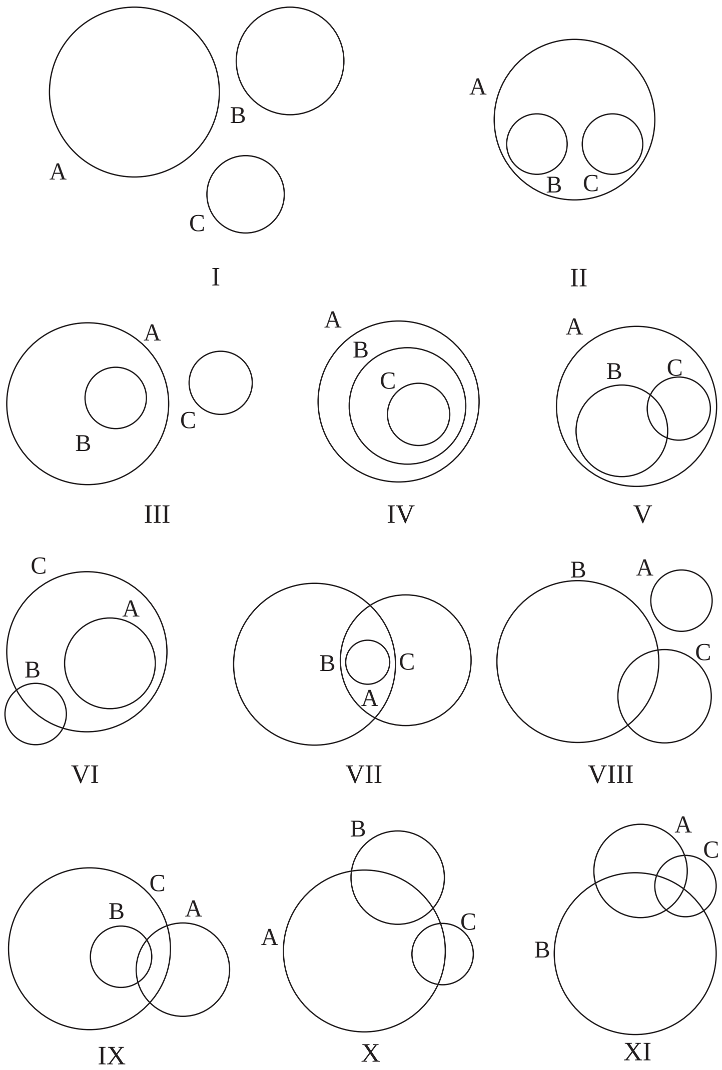
A
B
C
I
A
B
C
II
A
B
C
III
A
B
C
IV
A
B
C
V
C
A
B
VI
B
C
A
VII
B
A
C
VIII
C
B
A
IX
B
A
C
X
A
C
B
XI

Figure 223b

	positive centers of similitude	negative centers of similitude	number of solutions	corresponding groups
I	all	none	8	
II	S_{23}, S_{23}	S_{12}, S_{12}, S_{13}, S_{13}	8	
III	S_{13}, S_{13}, S_{23}, S_{23}	S_{12}, S_{12}	none	
IV	none	all	none	
V	S_{23}	S_{23}, S_{12}, S_{12}, S_{13}, S_{13}	4	(0,1)
VI	S_{12}, S_{12}, S_{23}	S_{23}, S_{13}, S_{13}	4	(2,3)
VII	S_{23}	S_{23}, S_{12}, S_{12}, S_{13}, S_{13}	4	(0,1)
VIII	S_{12}, S_{12}, S_{13}, S_{13}, S_{23}	S_{23}	4	(0,1)
IX	S_{12}, S_{13}	S_{12}, S_{13}, S_{23}, S_{23}	4	(2,3)
X	S_{12}, S_{13}, S_{23}, S_{23}	S_{12}, S_{13}	4	(0,1)
XI	S_{12}, S_{13}, S_{23}	S_{12}, S_{13}, S_{23}	8	

We may observe that the number of cases to examine could have been reduced considerably by taking into account that cases I and II can be reduced to each other by an inversion; the same is true of cases III, IV, cases V, VI, VII, VIII, and cases IX, X.

Note D:
On the Notion of Area

313. In Book IV of this work, we have taken the usual path of assuming a priori (243) that it is possible to define the area of a polygon; that is, that we can associate a number to each polygon (called its area) possessing the following properties:

I. Two congruent polygons have the same area, regardless of their position in space;

II. The polygon P , the sum of two adjacent polygons P, P , has area equal to the sum of the areas of P and P .

The possibility of such a correspondence amounts, in this theory, to a postulate. But this postulate is useless. The fact in question need not be assumed, but can be proved by the following method, which must be regarded, for this reason, as superior to the previous one.

314. Theorem. In any triangle, the product of a side and the corresponding altitude is the same, regardless of the chosen side.

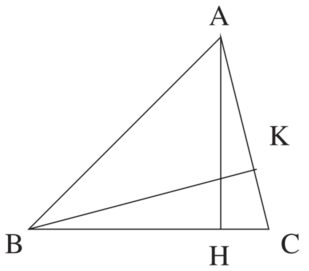

Figure 224

Let the given triangle be ABC (Fig. 224), in which altitudes AH, BK correspond to the sides BC, CA respectively. Right triangles ACH, BCK have angle C in common; therefore they are similar and give us

$$\frac{AH}{AC} = \frac{BK}{BC},$$

so that $BC \cdot AH = AC \cdot BK$. QED

The area of the triangle is what we call the preceding product, multiplied by a constant k, chosen once and for all, and whose choice will be discussed soon. This area is zero when the triangle ceases to exist properly speaking (when the three vertices are collinear) and only in that case.

The areas of triangles with the same altitude are proportional to their bases. It is clear that two congruent triangles have the same area.

315. Consider an arbitrary point O in the plane of a triangle. By joining it to the three vertices, we form three new triangles with the three sides of the original triangle as bases, and vertex O. One of these triangles will be said to be additive (for instance, triangle OBC in figure 228) if it is on the same side of the original triangle, relative to the common base; it will be called subtractive (for instance, triangle OBC in Figure 229) otherwise.[46]

Theorem. If an arbitrary point O is given in the plane of a triangle, and joined to its vertices, the sum of the areas of the additive triangles, minus the sum of the areas of the subtractive triangles, equals the area of the original triangle.

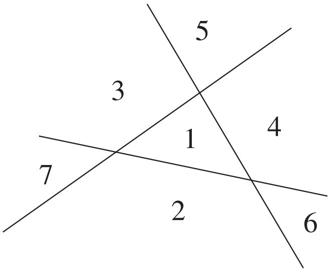

Figure 225

The sides of the given triangle divide the plane into seven regions (figure 225): one interior region; three regions (2–4 in the figure) separated from the first by one of the three sides; and three last regions (5–7) in the vertical angles corresponding to the angles of the triangle.

We now distinguish five cases.

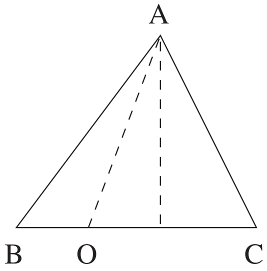

Figure 226

1°. Point O is on one of the sides. If point O is taken on side BC of triangle ABC (Fig. 226), there is no triangle OBC. Then the sum of the areas of the two triangles OAB, OAC is equal to the area of ABC because these triangles have the same altitude (the perpendicular from A to BC) and the third base BC is the sum of the first two bases OB, OC.

---

[46] If the vertices of the triangle are given in a fixed (but arbitrary) order, and those of one of the triangles with vertex O are given in such a way that the common vertices are in the same order, we can also say that this last triangle is additive or subtractive, depending on whether it has the same sense of rotation as the original triangle or not.

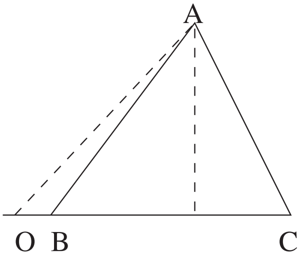

Figure 227

2°. Point O is on the extension of a side. If point O is on the extension of BC (Fig. 227), there is no triangle OBC. As for the triangles OAC, OAB, their difference is indeed ABC: these triangles have the same altitude, and the third base BC is the difference of the first two OC, OB.

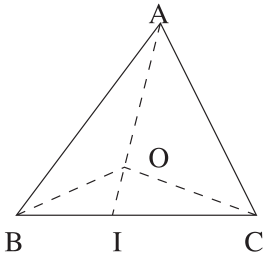

Figure 228

3°. Point O is inside the triangle (Fig. 228). We extend OA to its intersection I with BC. Triangle ABC is equal (1°) to the sum of the triangles ABI, ACI, which can in turn be decomposed into AOB + BOI, AOC + COI. Now the triangles AOB, AOC are additive triangles, and the triangles BOI, COI together give a third additive triangle BOC.

4°. Point O is outside triangle ABC, but inside one of its angles, for example A. If I is the intersection of OA and BC, the sum of the two additive triangles AOB, AOC can be replaced (1°) by the sum of the four triangles AIB, AIC, OIB, OIC. The sum of the first two of these is ABC, and the sum of the other two is OBC. Thus we have

$$AOB + AOC = ABC + OBC,$$

or

$$OAB + OAC - OBC = ABC,$$

which is the required relation, since triangle OBC is subtractive.

5°. Point O is in a vertical angle corresponding to one of the angles of the triangle, for instance A (Fig. 230). In this case A is inside triangle OBC, so that (3°)

$$OBC = ABC + OAB + OAC,$$

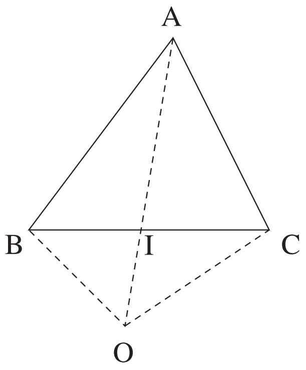

Figure 229

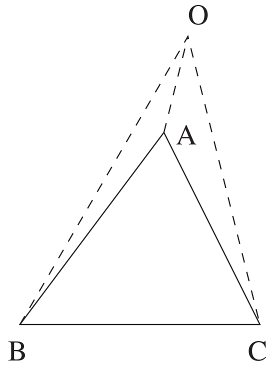

Figure 230

or

$$OBC - OAB - OAC = ABC,$$

which is the required relation because triangle OBC is additive, and the other two subtractive.

316. Now consider an arbitrary polygon ABCDE (Fig. 231) and an arbitrary point O in its plane. Joining O to the vertices, we form triangles with vertex O and the various sides as bases. Each of these triangles will be considered additive or subtractive depending on whether it is on the same side as the polygon,[47] relative to the common side.

Theorem. Consider a polygon, decomposed in an arbitrary manner into a number n of triangles, and a point O of the plane, joined to all its vertices. The difference S between the sum of the areas of the additive triangles with vertex O and the sum of the areas of the subtractive triangles (or just the first sum, if the second one does not exist) is equal to the sum Σ of the areas of the n triangles into which the polygon was decomposed.

---

[47] When we talk of the position of the polygon, relative to one of its sides, we refer to the region which is immediately next to this side. For instance, triangle OAB (Fig. 231) is additive because it is on the same side of AB as the area shaded in the figure.

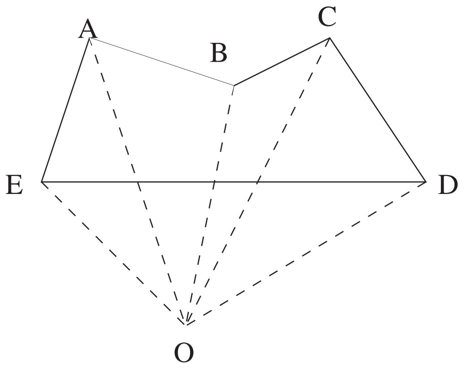

Figure 231

If the theorem is true for two adjacent polygons P, P decomposed into triangles, it is also true for their sum P . Indeed, we can distinguish two parts in the perimeters of P and P : the sides (or parts of sides) which are not common to both polygons, but which form part of the perimeter of P , together with their additive and subtractive triangles,[48] and the common sides, or parts of sides. To these common sides correspond triangles which are additive for P or P , and subtractive for the other (since P, P are situated on different sides of the common side) and which will therefore disappear when calculating the sum of the quantities S relative to these two polygons. Thus this sum is indeed the quantity S formed with polygon P . On the other hand, the quantity for P is obviously the sum of the analogous quantities for P and P , so that the equality of the quantities S and is true for this third polygon P if it is true for the first two.

We can now consider the proof finished, because the theorem has been proved when n = 1 (in which case it reduces to the preceding theorem) and, if it is true for some value of n, it will be true for the following value n + 1 (since a polygon composed of n + 1 triangles is the sum of a triangle and a polygon composed of n triangles).

Corollaries. The quantity S is independent of the choice of the point O, since the quantity does not depend on this choice.

Likewise, the quantity is independent of the decomposition of the polygon into triangles.

317. The common value of the numbers S and will be called the area of the polygon.

Two congruent polygons have the same area since they can be decomposed into pairs of congruent triangles; also, the preceding proof shows that, when two polygons are adjacent, the area of their sum is equal to the sum of the partial areas.

In other words, the areas thus defined possess properties I and II.

---

[48] We may have to divide a side of the polygon P into several segments, some of which will be shared with P , and some not, and to replace the triangle with vertex O and this segment as base with partial triangles whose bases are the smaller segments. This will not change the quantity S, since these partial triangles add up to the total triangle (by the preceding theorem). We can do the same with P and P .

318. It follows that it is impossible to decompose a polygon into parts which, assembled differently (but so that they are still adjacent to each other), form a polygon that lies inside the original one.

This is true because the new polygon would have the same area as the first.

This proposition is not proved by the theory given in the text, since the existence of areas is taken as a postulate.

319. We have not yet discussed the choice of the constant k. It is clear that changing this number amounts to changing all areas into proportional areas. We have already remarked (244) that such a change would not affect its two fundamental properties.

We can then determine k in such a way that the square with a side equal to the unit of length has unit area. Now this square is composed of two triangles, each of which has both a base and an altitude equal to the unit length. Its area is 2k, so that we choose $k = \frac{1}{2}$. The area of any triangle will then be half the product of its base and altitude. The areas determined in this way are then just those which we learned how to calculate in the text.[49]

---

[49] The arguments given in the text show that the method we have used earlier to calculate areas is the only one which has the preceding properties, while at the same time satisfying the condition that the square with unit side has unit area.

Appendix:
Malfatti's Problem

The principles established in the text (227 236) and in Note C, concerning tangent circles and circles isogonal to two given circles (that is, intersecting them at equal angles) allow us to present a solution to a famous problem, due to Malfatti, and stated as follows:

Given three lines (a_1), (a_2), (a_3) in a plane,[58] find three circles (x_1), (x_2), (x_3) each pair of which are tangent and, in addition, (x_1) is tangent to (a_2, a_3), (x_2) is tangent to (a_3, a_1), and (x_3) is tangent to (a_1, a_2).[59]

The solution was first given by Steiner, and we will follow an exposition by Schröter. It involves the property established in 228.

We have seen there that the circles which are tangent to two given circles C, C or, more generally, which are isogonal to C, C, are divided into two series which can be characterized (in the case where the inversions which change C into C have positive power) by saying that the circles of the first series are orthogonal to a certain circle, and those in the second series are orthogonal to an analogous circle $_1$.

The circles , $_1$ are, as we said, analogous to the bisectors of the angles formed by two lines, and we will call them the bisecting circles of C and C.

One of these bisectors reduces to a line when C and C are equal (225), and only in this case.

When the circles C, C are tangent, one series of isogonal circles is obviously formed by the circles passing through their point of contact. The corresponding bisecting circle is reduced to that point. In this case, one cannot define an inversion which transforms C into C, since the power of this inversion would have to be zero. In what follows, when we talk about circles isogonal to two tangent circles C, C, we do not refer to those just mentioned, but to the proper series, whose circles do not pass through the same point, and are preserved by an inversion (whose power is not zero).

If an inversion transforms two circles C, C into C_1, C_1, it also transforms the bisecting circles of C, C into those of C_1, C_1. Indeed, the definition of the bisecting circles is based solely on the notion of the angle of two circles, and this is not changed by inversion.

---

[58] We state the problem as if the three lines form a triangle with vertices A_1, A_2, A_3, which is the general case. The solution does not change if, for example, two of the lines are parallel.

[59] We suppose that the nine points of contact (those of the required circles with the given lines, and those of each pair of circles) are distinct. In other words, we do not consider the situation when, for instance, (a_1) is tangent at the point of contact of (x_2) and (x_3) (a situation which could, however, be treated with the same methods).

Let us now consider circles ($_1$), ($_2$), ($_3$) with centers $_1$, $_2$, $_3$ and pairwise tangent (externally, to be definite) at points $_1$, $_2$, $_3$ (see fig. 232). Join $_2$ $_3$, and let μ_1, $_1$ be the new points where this line intersects ($_2$), ($_3$), respectively.

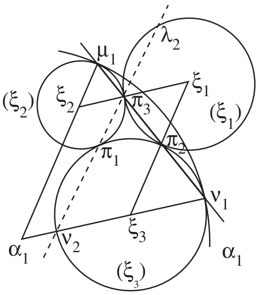

Figure 232

Through the points μ_1, $_1$ we can draw a new circle ($_1$), which is tangent to ($_2$), ($_3$) at these points, and whose center $_1$ is at the intersections of the radii μ_1 $_2$, $_1$ $_3$. The radius of this circle is the sum of the radii of the given circles.

To see this, observe that μ_1 $_2$ is parallel to $_1$ $_3$ (because $_3$ is a center of similarity for ($_1$), ($_2$)) and $_1$ $_3$ is parallel to $_1$ $_2$ (for an analogous reason). In parallelogram $_1$ $_2$ $_1$ $_3$, the equality of the opposite sides yields the indicated value for both $_1\mu_1$ and $_1$ $_1$.

Thus, if we also join points $_1$, $_3$ with a line which intersects ($_1$) again at $_2$ and ($_3$) at $_2$, the circle ($_2$) which is tangent at these points to ($_1$), ($_2$), respectively, is equal to ($_1$), since the preceding argument can be applied to it as well.

It is in fact be easy to see that the equality of the circles ($_1$), ($_2$) would still hold if the contacts of ($_1$), ($_2$), ($_3$) were not all exterior.[60]

We deduce immediately the following lemma, on which our subsequent reasoning will be based.

Lemma. If three circles (x_1), (x_2), (x_3) are pairwise tangent[61] at P_1, P_2, P_3, let (a_1) be a circle tangent to (x_2), (x_3) at m_1, n_1, and let (a_2) be a circle tangent to (x_1), (x_3) at $_2$, n_2, so that points m_1, n_1, P_2, P_3 are (227, 224)[62] on a circle (k_1), and the points $_2$, n_2, P_1, P_3 are on a circle (k_2).

Then the common point N_3 (other than P_3) of (k_1), (k_2) is on the bisecting[63] circle (g_3) of (a_1), (a_2).

---

[60] In this case, two of the circles ( ) would have to be tangent externally to each other, and internally to the third. Each circle ( ) would have a radius equal to the difference between the radius of the large circle and the sum of the radii of the small ones.

[61] Each point P is the point of contact of the circles (x) with indices different from its own.

[62] In order to apply the proposition of 224, we should generally make sure that (a_1), (x_1) belong to the same series of tangent circles to (x_2), (x_3), but in our situation there is only one proper series.

[63] Circle (x_3) belongs to one of two series of circles tangent to (a_1), (a_2); we are referring here to the bisecting circle corresponding to this series of circles.

Indeed, transform the figure by inversion, taking N_3 as the pole. Circles (x_1), (x_2), (x_3) will give us new circles ($_1$), ($_2$), ($_3$); (a_1), (a_2) will give circles ($_1$), ($_2$), while (k_1), (k_2) become two lines. We are led to the figure considered above, and we have just established that ($_1$), ($_2$) are equal. Their bisecting circle will thus be a line, so that the bisector of (a_1), (a_2) will pass through N_3. QED

Moreover, the circle through n_1, N_3, n_2 is orthogonal to (x_3), (a_1), (a_2), and (g_3).

Circle (k_1) intersects (a_1), (g_3) at equal angles, and likewise, (k_2) intersects (a_2), (g_3) at equal angles.

The angle made by circle (k_1) with (g_3) at N_3 is of the opposite sense from its angle with (a_1) at (n_1).

All of this is easily seen on a figure transformed by inversion with pole N_3, on which figure the points $_1$, $_2$ will be diametrically opposite on the circle ($_3$).

Having established these points, we can now attack the proposed problem.

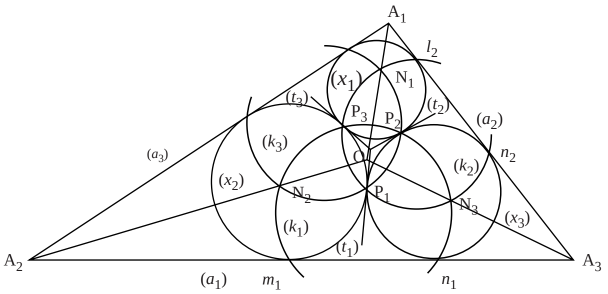

Figure 233

I. As above, denote by P_1, P_2, P_3 the points of contact of the required circles (x_1), (x_2), (x_3) (Fig. 233). We can apply the lemma just proved to these circles, with circles (a_1), (a_2) replaced by the given lines. We then know

1 . that there is a circle (k_1) passing through points P_2, P_3 and the points of contact m_1, n_1 of (x_2), (x_3) with (a_1); a circle (k_2) through P_3, P_1 and the points of contact of (x_3), (x_1) with (a_2); and a circle (k_3) through P_1, P_2 and the points of contact of (x_1), (x_2) with (a_3).

2 . that, for instance, (k_1), (k_2) intersect at a point N_3 (other than P_3) on the bisector (g_3) of the angle between (a_1) and (a_2), which here replaces the bisecting circle. (If, as we are now assuming, the circles (x) are all inside the triangle, we are dealing with the bisector of the interior angle.) Similarly, circles (k_2), (k_3) intersect at a point N_1 on the bisector (g_1) of the angle between (a_2) and (a_3), and circles (k_3), (k_1) intersect at a point N_2 on the bisector (g_2) of the angle between (a_3) and (a_1).

3 . that (k_1) intersects (a_1), (g_3) at equal angles.

We can observe immediately that, for an analogous reason, the angle made by circle (k_1) with (g_2) is equal to the angles made by the same circle with (a_1) or with (g_3).

It is also equal to the angle made by (k_1) with circle (x_2) at point m_1 (where this circle is tangent to (a_1)) or at the point P_3 (which is the same, but with the opposite orientation).[64]

II. Denote by (t_1), (t_2), (t_3) the common tangents to the required circles at the points of contact P_1, P_2, P_3, respectively. We see that line (t_3) intersects (k_1) at the same angle as (g_3) since this angle, measured at P_3, has the opposite sense from the analogous angle of (k_1) and (g_3) at N_3, as can easily be seen from what we have said so far.

Point P_3 is obviously symmetric to N_3 with respect to the line of centers of (k_1), (k_2). Thus (t_3) is symmetric to (g_3) with respect to the same line (since it forms the same angle with (k_1), but in the opposite sense).

For the same reason, (t_2) intersects (k_1) at the same angle, as it is symmetric to (g_2) with respect to the line of centers of (k_3), (k_1).

It is clear, however, that all the lines intersecting a given circle at a given angle are tangent to a second circle, concentric with the first. Therefore, the five lines (a_1), (g_2), (g_3), (t_2), (t_3) are tangent to a circle[65] (k_1) concentric with (k_1).

This last result is equivalent to a solution of the problem. Circle (k_1) can be considered to be known, since we know three of the five tangents (namely, (a_1), (g_2), (g_3)). Similar considerations apply to the analogous circles (k_2), (k_3) which are concentric to (k_2), (k_3), and we arrive at the following construction:

Let O be the center of a circle tangent to the three given lines, and let (g_1), (g_2), (g_3) be the lines joining the vertices of the triangle formed by these lines to O. We draw a circle (k_1) tangent to the lines (a_1), (g_2), (g_3), a circle[66] (k_2) tangent to the lines (a_2), (g_3), (g_1), and a circle (k_3) tangent to the lines (a_3), (g_1), (g_2).

Let (t_1) be the second common tangent of (k_2), (k_3), symmetric to (g_1) with respect to the line of the centers of these circles. Let (t_2) be the common tangent to (k_3), (k_1) symmetric to (g_2) with respect to the line of their centers. Let (t_3) be the common tangent to (k_1), (k_2) symmetric to (g_3) with respect to the line of their centers.

The first circle (x_1) is tangent[67] to lines (a_2), (a_3), (t_2), (t_3); likewise, (x_2) is tangent to (a_3), (a_1), (t_3), (t_1), and (x_3) is tangent to (a_1), (a_2), (t_1), (t_2).

---

[64] To determine the orientation of the angles under consideration, we rely on the fact that two circles (or a circle and a line) which have two common points form, at these two points, equal angles of opposite orientation, which is obvious, as these angles are symmetric relative to a line.

[65] In order not to make the figure too complicated, we have not drawn in circle (k_1) or the analogous circles (k_2), (k_3).

[66] There exist four circles tangent to (a_1), (g_2), (g_3). If one wants the circles (x) to be inside the given triangle, one must obviously choose the circle inscribed in the triangle formed by these lines. If not, any of these four circles can be chosen, but the other two analogous circles are then determined: indeed, we see (by performing another inversion relative to the pole N_3) that the chords N_3n_1, N_3n_2 must be symmetric with respect to (g_3), and the same must be true of the bisectors of the angle between (a_1) and (g_3) and the angle between (a_2) and (g_3), which pass through the centers of (k_1) and (k_2) respectively.

[67] Its center is on line (g_1) (and not on the bisector perpendicular to this line); one can show, by assigning directions to the lines (g), (t), that it is also on the bisector of the angle formed by (t_2), (t_3) which passes through the center of (k_1).

Lines (g_1), (g_2), (g_3) can be the bisectors of any of the angles formed by the given lines, provided that these bisectors intersect.[68]

It remains to prove that the construction indicated above does effectively produce circles solving the problem. We will not establish this fact, which was treated by Petersen.[69]

It is remarkable that the preceding solution extends to the case in which (a_1), (a_2), (a_3) are arbitrary circles, instead of being lines.[70]

In this case (g_1), (g_2), (g_3) will be the bisecting circles of $(a_2), (a_3)$; $(a_3), (a_1)$; $(a_1), (a_2)$, respectively.[71] Using this fact, all the reasoning involving only the circles (a_1), (a_2), (a_3), (g_1), (g_2), (g_3), (k_1), (k_2), (k_3) (whose definitions are not changed), that is, all the reasoning done until Part II, remain true without modification.[72]

In order to apply the results of this reasoning to actual situations, we must state what is meant by (t_1), (t_2), (t_3).

To do this, let (C) be the circle orthogonal to (a_1), (a_2), (a_3) (assuming, to be definite, that it exists).[73] We denote by (t_1) the circle orthogonal to (C), and tangent to (x_2) and (x_3) at their point of contact P_3.

Similarly, let (t_2), (t_3) be two circles orthogonal to (C), and each tangent to two of the given circles at their point of tangency.

---

[68] This follows from the fact that the points of contact of the required circles must be external, and the lines (a) must be common external tangents. (See an analogous remark later in this note for the case when (a_1), (a_2), (a_3) are replaced by circles.)

[69] Crelle's Journal, vol. 89, pp. 130–135. Clearly one must establish first that the lines (a_2), (a_3), (t_2), (t_3) are tangent to the same circle. The reader who has solved Exercise 422 will convince himself that this is indeed the case; that this is so even when O is any point, not necessarily the center of the inscribed circle; and that the lines (t_1), (t_2), (t_3) are concurrent.

On the other hand, the fact that these circles are tangent depends on the particular position of the point O (see Petersen's work). We must remark (limiting ourselves to the case in which the circles (x) are inside the triangle) that the existence of a solution is obvious by continuity. Indeed, draw an arbitrary circle (x_1) tangent to (a_2) , (a_3) and inside the triangle: there exists a circle (x_2) inside the triangle and tangent to (a_1), (a_3), (x_1), and a circle (x_3) tangent to (a_1), (a_2), (x_1).

These circles (x_2), (x_3) clearly intersect in two points when (x_1) is very small, and do not intersect at all when (x_1) is the inscribed circle in the triangle. Thus, for some value of the radius of (x_1), the circles (x_2), (x_3) will be tangent.

[70] The new problem would immediately reduce to the first if circles (a_1), (a_2), (a_3) had a common point S (an inversion with pole S would show this). But the solution must be considered differently if this is not the case.

[71] Among the six possible circles bisecting (a_1), (a_2), (a_3) in pairs, we must choose three for (g_1), (g_2), (g_3) which have (Exercise 276) the same radical axis. Assume, to be definite, that the required circles are tangent externally. Then circle (a_1) will have contacts of the same kind with (x_2), (x_3) as (x_3) itself , since there is only one proper series of common tangent circles to these two. Likewise, (a_2) has contacts of the same kind with (x_3), (x_1), and (a_3) with (x_1), (x_2). By considering the various kinds of contacts possible, we will be able to verify that the circles (g) satisfy the relation indicated.

The case where two of the circles are tangent internally can be reduced to the preceding one by inversion (and this discussion can thus be simplified).

[72] It follows that the bisecting circles in question exist whenever there is a solution to the problem. For instance, (g_1) must have its center at one of the centers of similarity of (a_2), (a_3), and must pass through N_3.

[73] This circle is reduced to one point in the case of footnote IV [i.e. when the three given circles have a common point. — transl.].

The reasoning developed above shows that circle (k_1) forms equal angles[74] with (x_2), (x_3), and with (a_1), (g_2), (g_3).

The last three are orthogonal to (C).

It follows (see Note C, especially 310) that any circle (k_1), having the same radical axis[75] as (k_1), (C), is also isogonal to (a_1), (g_2), (g_3), and that if it is tangent to one of them (311, construction) it will also be tangent to the others. Circle (t_3) is in fact the image[76] of (g_3) under the inversion which does not change (C), (k_1), or (k_2). The same is true if we replace (k_1), (k_2) by (k_1), (k_2).

Thus the construction is as follows:

Let (a_1), (a_2), (a_3) be given circles, and let (g_1), (g_2), (g_3) be bisecting circles of (a_2), (a_3); (a_3), (a_1); (a_1), (a_2) respectively, these circles being chosen so as to have the same radical axis. Let (C) be the circle orthogonal to the given circles

Determine circle (k_1) tangent to (a_1), (g_2), (g_3), circle (k_2) tangent to (a_2), (g_3), (g_1), and circle (k_3) tangent to (a_3), (g_1), (g_2)[77]; or, more generally[78], determine circle (k_1) isogonal to (a_1), (g_2), (g_3), circle (k_2) isogonal to (a_2), (g_3), (g_1), and circle (k_3) isogonal to (a_3), (g_1), (g_2).

Let (t_1) be the image of (g_1) under the inversion which does not change (C), (k_2), or (k_3); let (t_2) be the image of (g_2) under the inversion which does not change (C), (k_3), or (k_1); let (t_3) be the image of (g_3) under the inversion which does not change (C), (k_1), or (k_2).

Of the required circles, the first is tangent to (a_2), (a_3), (t_2), (t_3); the second is tangent to (a_3), (a_1), (t_3), (t_1); and the third is tangent to (a_1), (a_2), (t_1), (t_2).

It is easy to verify that what we have said remains true when the circle (C) does not exist, when the radical center I of (a_1), (a_2), (a_3) is interior to these circles. We just replace the words circle orthogonal to (C) by circle relative to which the power of I is the same as its power relative to (a_1), (a_2), (a_3) , and we say that a circle (k_1) has the same radical axis as (k_1), (C) if every circle orthogonal to (k_1), such that the power of I is the same as that with respect to (a_1), is also orthogonal to (k_1).

---

[74] With the same remark as before concerning the orientation of these equal angles.

[75] It is clear that, unlike the first case [where the (a) are lines transl.], the circle (k_1) is not generally concentric with (k_1).

[76] This follows as before from the fact that this image satisfies all the conditions defining (t_3), including its intersection with (g_1) and the size of the angle at that point, as well as concerning its orthogonality to C. The inversion involved here has as its pole the radical center of the circles under consideration.

[77] As before, we can take (k_1) to be any of the circles tangent to (a_1), (g_2), (g_3). Once this choice is made, it determines the series of circles isogonal to (a_2), (g_3), (g_1) to which (k_2), (k_2) belong.

[78] It is important to be aware of this generalization. It can indeed happen that the circle (k_1) does not exist (when the construction of 311 is impossible), even though our problem does have solutions.

Chapter 3

Chapter 3 - Part 1

Volume-03---20211108

!"#$# % & & !' (! !) * * + , * - . *
/ * !0 + ,1 * + , 2 "/" & 3 45

-6 7 & *6 5*5

* 5 * 4 8 * 6 9* 8 :*5* * 51 ;* * 8 & 5 . 5 <
"* * % & & 9* 8 :*5* * 5 2 * 8 & 3 * 8 & 4 *5
4* * * 4* * 5

* 1 < * 5 < 2* 4* 5 * 8 & *6* = 3

!"#$# % & & !' (! !) * * + , * - . *
/ * ! * * * 5 , 2 7532. &> 3 2 "/" & 3 45

!#)?! ! !9* 8 :*5* * 5!9* 8 :*5* * 5 0 @ 5
.A A A A * B 2.A A 4 * 3 # * 5 2 C 3 45

-6 7 & D * E 1 5 . * * &
!"#$# % & & !' (! !) * * + , * - . *
/ * !0 + ,1 * + , 2 "/" & 3 45

!"#$# % & & !' (! !9* 8 :*5* * 5!9* 8 :*5* * 5
0 @ 5 .A A A A * BB 2.A A 5* F 4* 3 # * 5 2 C 3 45

-6 * * * 2 *@G 3

!"#$# % & & !' (! !-4 * 5 5 , 977
2#5&* 531 5 5 !' * * * B 977 2#5&* 53 7%+#+7 07#)%B%+
2 C3 45

D. 6

!"#$# % & & !' (! !-4 * 5 5 , 977
2#5&* 531 5 5 !' * * -4 * 5 5 , 977 2#5&* 53
7%+#+7 B%"B# 2 C3 45

!"#$# % & & !' (! ! & . 6 B - 4* *
H 5! & . 6 B - 4* * H 5 45

-6 0 & * 0 5 III

!"#$# % & & !' (! !# ' 5. 0 *
. 6 *!#J' 5. J J0 * J# . 6 * 45

-6 0 & * 0 5 III

!"#$# % & & !' (! ! 5 , 1 5 5 ! 5
, 1 5 5 45

-6 0 & * 0 5 III

!"#$# % & & !' (! !) * 1 #6 * 1 # 5
#44 5 # B 5 !) * J J#6 * J# 5J#44 5J J# JB 5 45

!"#$# % & & !' (! ! 5 , 1 5 5 !K
K * 5 , 9 H , L B 2 C3 45

!"#$# % & & !' (! !0 F *, + ,!0 F *,
+ , B) 2 C M3 45

!"#$# % & & !' (! ! # * * * 75
!+ , $ 0 * 5 ## 2 3 45

+* 1) * 1 : 6 1 : ; & 2 3

!"#$# % & & !' (! !$ + , % 6 !$
+ , % 6 45

M 2 *@G 3

!"#$# % & & !' (! !9* 8 :*5* * 5!9* 8 :*5* * 5
0 @ 5 .A A A A * BB 2.A A 5* F 4* 3 # * 5 2 C 3 45

-6 - = N5 5 & G *4 * *. 1 * D * & G 8 *
D 5 <

!#)?! ! !9* 8 :*5* * 5!9* 8 :*5* * 5 0 @ 5
.A A A A * BB 2.A A 5* F 4* 3 # * 5 2 C 3 45

5 4D. *

ERRATA DE LA GÉOMÉTRIE PLANE

	Au lieu de :	*Lire :*
Table, p. XII, n°s des paragraphes du Livre II, ch. VII,	**100** *bis* **101** **102**	**101** **102** **102** *bis*
Page 51, n° **58**, 2°,	Si, la distance, OH	Si la distance OH
— 52, *ligne* 3e avant dernière du n° **59**,	$\widehat{TMM'}$ égal à $\widehat{MOH}$	$\widehat{TMM'}$, égal à $\widehat{MOH}$
Page 89, exercice 88 *bis*,	coupe ou ne coupe pas ce cercle, suivant qu'elle est plus grande ou plus petite que la somme des tangentes...	est extérieure ou sécante à ce cercle, suivant qu'elle est ou non comprise entre la somme et la différence des tangentes...
— 96, dernier alinéa (*ligne* 5e par en bas),	Le théorème du n° **101**	Le théorème du n° **102**
Page 107, *ligne* 4,	*la direction de* Ax'	*la direction* Ax'
— 109, *ligne* 10,	CB	CD
— 165, *ligne* 3e avant-dernière de la note,	le côté de l'apothème	le côté et l'apothème
Page 186, *ligne* 8e par en bas,	Pour le démontrer, observer	Pour le démontrer, observons
— 241, *ligne* 7,	n° **224**	n° **244**
— 250, exercice 305,	les rayons qui vont au sommet	les rayons qui vont aux sommets
— 277, *ligne* 30,	par une inversion de pôle A'	par une inversion I' de pôle A'
— 287, *ligne* 15 du n° **309**,	de plus, les points b', c'; c', b'	de plus, les points b, c'; c, b'
— — fin du n° **309**,	sur ce même axe similitude.	sur ce même axe de similitude.
— 303, *ligne* 5 (fin du premier alinéa de l'exercice 391),	coupe AB en un points fixe	coupe AB en un point fixe
Page 304, dernière *ligne* (exercice 401),	trois cercles de centre A, B, C	trois cercles de centres A, B, C
Page 305, *ligne* 1,	concentriques aux premières	concentriques aux premiers
— 307, 5e *ligne* de l'exercice 418,	côté BC	côté BC_1

LIVRE VII

DÉPLACEMENTS — SYMÉTRIES SIMILITUDE

CHAPITRE PREMIER

DÉPLACEMENTS

410. *Deux figures* F, F' *sont égales, si l'on peut trouver trois points non en ligne droite* A, B, C *de l'une, auxquels correspondent les points* A', B', C' *de l'autre, de manière que, si* M *est un point quelconque de* F *et* M' *son homologue dans* F', *la figure* $ABCM$ *soit égale à la figure* $A'B'C'M'$.

Supposons, en effet, qu'il en soit ainsi et transportons la figure F' de manière que le triangle $A'B'C'$ vienne sur le triangle ABC (qui lui est manifestement égal dans ces conditions).

Soient alors M et M' deux points correspondants des deux figures. En général, le point M' n'est pas dans le plan $A'B'C'$ et les droites $A'B'$, $A'C'$, $A'M'$ forment un trièdre égal, par hypothèse, à $A.BCM$. Lorsque $A'B'C'$ est superposé à ABC, ces deux trièdres coïncident (**380**, Rem.) et $A'M'$ prend la direction AM. Comme $A'M' = AM$, le point M' vient sur M.

La même conclusion subsiste si M' est dans le plan $A'B'C'$, en vertu de ce que nous savons (Pl., **50**, Rem. II) sur les figures planes égales.

Les figures F, F' ont donc été ainsi complètement superposées.

Le raisonnement montre en outre que *deux figures égales qui ont trois points homologues communs, non en ligne droite, coïncident.*

Il en résulte que *deux déplacements qui produisent le même effet sur trois points* A, B, C *non en ligne droite, sont identiques entre eux.*

C. Q. F. D.

Car les deux figures F′, F qui résultent d'une même figure quelconque F par ces deux déplacements, coïncident, comme ayant trois points homologues communs.

411. Rotations.

Considérons deux figures égales ayant *deux* points homologues communs A, B.

D'après la définition de la ligne droite (Pl., **5**), la droite AB a la même position dans les deux figures, et il en est de même d'un point quelconque *m* de cette droite, puisque le segment A*m* doit avoir la même grandeur et le même sens de part et d'autre.

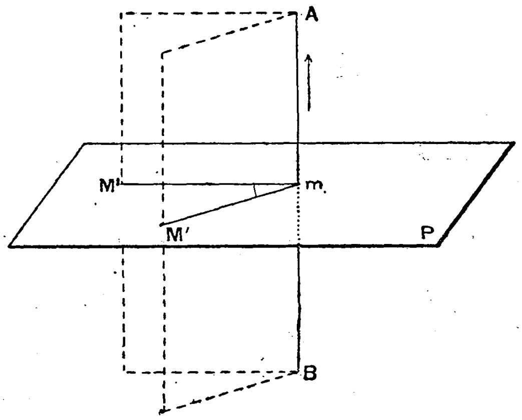

Fig. 316.

Ainsi *les deux figures ont en commun tous les points d'une droite.*

Considérons la portion de la figure située dans un plan quelconque perpendiculaire à AB, par exemple dans le plan P qui est perpendiculaire à AB au point *m* (*fig.* 316), et composée, par conséquent, de points M tels que M*m* soit perpendiculaire à AB. A ces points correspondront, dans la seconde figure donnée, des points jouissant de la même propriété (le point *m* restant le même) et, par conséquent, situés dans le même plan P.

Nous avons donc, dans ce plan, deux figures égales et de même sens (1), qui ont un point commun *m* et, par suite, se déduisent l'une de l'autre par une rotation autour de ce point (Pl., **100**). L'angle de cette rotation est l'angle plan du dièdre que forme un plan quelconque passant par AB dans la première figure avec le plan homologue de la seconde. *Cet angle de rotation est donc le même dans tous les plans* P.

Définition. — On donne, en géométrie de l'espace, le nom de *rotation* à l'opération par laquelle on déduit l'une de l'autre les deux figures dont il vient d'être parlé ; autrement dit, à l'opération qui consiste, étant donnée une droite AB (dite *axe de rotation*), et un angle (dit *angle de rotation*), à faire tourner chaque point de la figure primitive, dans le plan mené par ce point perpendiculairement à l'axe et autour du point de rencontre du plan en question avec l'axe, d'un angle égal à l'angle de rotation (*fig.* 316).

---

(1) Parce que deux droites issues de *m* forment, avec l'axe, un trièdre dont la disposition doit être la même dans les deux figures données.

On doit avoir choisi un sens sur l'axe et indiqué, une fois pour toutes, si les rotations effectuées sont directes ou rétrogrades par rapport au sens ainsi choisi (**349**).

Si l'on est libre de choisir à sa volonté le sens de l'axe, on peut évidemment toujours le prendre tel que la rotation soit directe.

Réciproquement, *deux figures déduites l'une de l'autre par une rotation, sont égales.*

Car, si M, M' en sont deux points correspondants, m leur projection (commune, en vertu de l'hypothèse) sur l'axe AB, on peut déplacer la première figure de manière à amener l'angle droit $\widehat{MmA}$ sur son égal $\widehat{M'mA}$. Ce déplacement, n'ayant pas changé les positions des points de AB, est, d'après ce qui précède, une rotation dont l'axe est AB et l'angle évidemment égal à $\widehat{MmM'}$. Cette rotation qui, par hypothèse, change la première figure en une figure égale, coïncide donc avec la rotation donnée, puisqu'elle a même axe et même angle.

412. Symétrie par rapport à une droite.

L'angle de rotation peut être de 180° : alors le point M' correspondant à un point quelconque M de la première figure sera le symétrique (Pl., **99**) de M par rapport à sa projection m sur l'axe (*fig.* 317).

On donne à la rotation de 180° autour d'une droite le nom de *symétrie* ou encore de *transposition* par rapport à cette droite.

Dans ce cas, il est inutile d'indiquer le sens de la rotation. La symétrie est *réciproque*, c'est-à-dire qu'en appliquant la même construction au point trouvé M', on retomberait sur le point M.

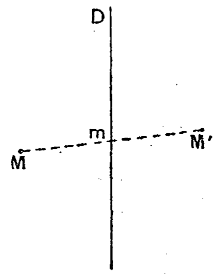

FIG. 317.

Il résulte de ce qui précède que *deux figures symétriques l'une de l'autre par rapport à une droite, sont égales entre elles.*

413. Translations.

Définition. — Si, par les différents points A, B, C, ... d'une figure F, on mène des droites AA′, BB′, CC′, ... toutes parallèles entre elles, de même sens et égales, les points A′, B′, C′, ... forment une figure F′ qui est dite (voir Pl., **51**) déduite de la première par la *translation* AA′ (*fig.* 318).

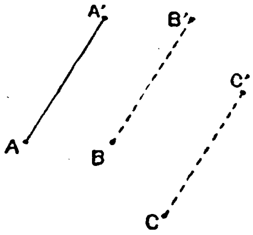

Fig. 318.

Cette nouvelle figure est égale à F.

En effet, si A et B sont deux points de F, il leur correspondra, dans F′, deux points A′ et B′, tels que A′B′ soit égal et parallèle à AB, avec le même sens.

A trois points A, B, C de F, correspondront trois points A′, B′, C′, tels que le triangle A′B′C′ soit égal à ABC. En particulier, l'angle $\widehat{B'A'C'}$, par exemple, est égal à $\widehat{BAC}$. De plus, le plan A′B′C′ est parallèle à ABC.

Soient alors A, B, C, M quatre points de F. Si ces points sont dans un même plan, la figure ABCM est (Pl., **50**) égale à celle que forment les points homologues A′, B′, C′, M′.

Si, au contraire, les quatre points A, B, C, M ne sont pas dans un même plan, le trièdre ABCM est égal au trièdre homologue A′B′C′M′ (**382** *bis*). Si l'on fait coïncider ces deux trièdres, comme A′B′, A′C′, A′M′ sont respectivement égaux à AB, AC, AM, les deux figures ABCM, A′B′C′M′ coïncident : d'où (**410**) l'égalité des figures F, F′.

Par exemple, *les projections d'une même figure sur deux plans parallèles sont deux figures égales.* Car elles se déduisent l'une de l'autre par une translation égale à la distance des deux plans et perpendiculaire à leur direction.

414. Déplacements hélicoïdaux.

Définition. — On nomme *déplacement hélicoïdal* (1) le déplacement résultant (2) d'une rotation R et d'une translation T, cette dernière étant parallèle à l'axe de la première.

L'ordre des deux opérations est indifférent (ce qui n'arriverait pas si la translation n'était pas parallèle à l'axe de rotation (3)). Autrement dit, si M′ est la nouvelle position que vient occuper le point M, après une rotation R d'un angle donné autour de l'axe D (*fig.* 319); M'_1, la nouvelle position que vient occuper le point M' par une translation déterminée T parallèle à D, on retomberait sur le même point M'_1 en faisant subir au point M la translation T, de manière à obtenir un point M_1, et faisant ensuite tourner ce point M_1 de l'angle donné autour de D.

---

(1) La raison de cette dénomination sera donnée plus loin, (**547**).

(2) Voir, pour le sens de ce mot, Pl., **103**, liv. II.

(3) Voir exercice 622.

Si, en effet, m est la projection commune des points M,M' sur l'axe; m_1, la projection du point M_1, la figure $M_1m_1M'_1$ se déduit évidemment de la figure MmM' par la translation MM_1. Elle est donc, comme celle-ci, un triangle isoscèle dont le plan est perpendiculaire à D et dont l'angle est celui de la rotation R : ce qui démontre l'exactitude de notre affirmation.

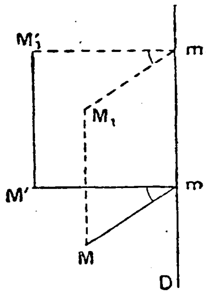

Fig. 319.

Le déplacement hélicoïdal comprend évidemment, comme cas particuliers, la rotation (lorsque la translation T est nulle) et la translation (lorsque la rotation R est nulle.)

Remarque. — Lorsqu'on se donne un déplacement hélicoïdal, un sens déterminé est par là même donné sur l'axe, à savoir le sens de la translation. On dira donc qu'un déplacement hélicoïdal est direct (ou encore *sinistrorsum*), ou bien au contraire qu'il est rétrograde (ou *dextrorsum*) suivant que la rotation est directe ou rétrograde par rapport au sens de la translation.

415. Nous avons vu que deux figures égales qui ont deux points communs (1) peuvent être amenées à coïncider par une rotation.

Deux figures égales qui ont un point commun peuvent être amenées à coïncider par deux rotations successives.

Soient, en effet, A le point commun; B,B' deux points homologues. On a $AB = AB'$ et, par conséquent, le point B' viendra coïncider avec B si l'on fait tourner la seconde figure autour d'une perpendiculaire au plan BAB', d'un angle égal à $\widehat{B'AB}$. Les deux figures ayant alors deux points communs, une seconde rotation amènera la coïncidence complète.

Deux figures égales quelconques peuvent être amenées à coïncider par une translation suivie de deux rotations. Car, si A,A' sont deux points homologues quelconques, la translation A'A imprimée à la seconde figure amène le point A' sur le point A, après quoi on est ramené au cas précédent.

Mais ces combinaisons de translations et de rotations peuvent être simplifiées et remplacées par un seul déplacement hélicoïdal.

A cet effet, nous procéderons comme en géométrie plane (Pl., **102-103**), en étudiant d'abord la composition des transpositions.

416. Deux symétries par rapport à la même droite se détruisent : nous savons (**412**) que, si M' est le symétrique du point M par rapport à la droite D, M'' le symétrique de M' par rapport à cette même droite, le point M'' coïncide avec le point M.

---

(1) C'est-à-dire telles que deux points coïncident avec leurs homologues.

Théorème. — *Deux transpositions successives, par rapport à deux axes différents, équivalent :*

1° *Si les deux axes sont parallèles, à une translation parallèle à la perpendiculaire commune aux deux axes et double de celle qui amènerait le premier axe sur le second;*

2° *Si les deux axes sont concourants, à une rotation autour de leur perpendiculaire commune, double de celle qui amènerait le premier axe sur le second;*

3° *Si les deux axes ne sont pas dans un même plan, à un déplacement hélicoïdal, ayant pour axe leur perpendiculaire commune et double de celui qui amènerait le premier axe sur le second.*

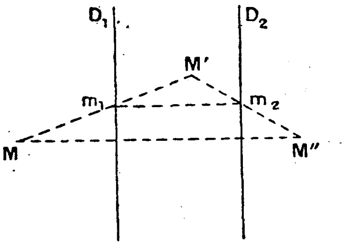

Fig. 320.

1° Soient les axes parallèles D_1, D_2 (*fig.* 320), M un point quelconque de la figure primitive, projeté en m_1 sur D_1 ; M′, son symétrique par rapport à D_1 ($m_1M' = Mm_1$), projeté en m_2 sur D_2 ; M″, le symétrique de M′ par rapport à D_2 ($m_2M'' = M'm_2$). Le plan mené par M perpendiculairement aux droites D_1, D_2 contient MM′ (qui est perpendiculaire à D_1) et M′M″ (qui est perpendiculaire à D_2). Donc m_1m_2 est une perpendiculaire commune aux deux axes et il est clair (Pl., **55**) que M″ se déduit de M par une translation parallèle à cette perpendiculaire commune et double de cette perpendiculaire commune.

2° Soient les deux axes D_1, D_2 (*fig.* 321) concourants en O ; P, leur plan ; Ox, la perpendiculaire à ce plan menée par le point O; et, comme tout à l'heure, M, M′, M″ trois points dont les deux premiers sont symétriques l'un de l'autre par rapport à D_1 (la droite qui les joint coupant D_1 en m_1), les deux derniers par rapport à D_2 (la droite qui les joint coupant D_2 en m_2).

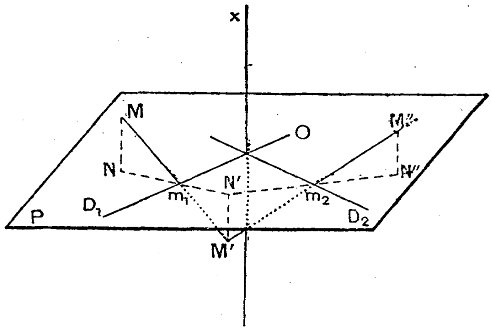

Fig. 321.

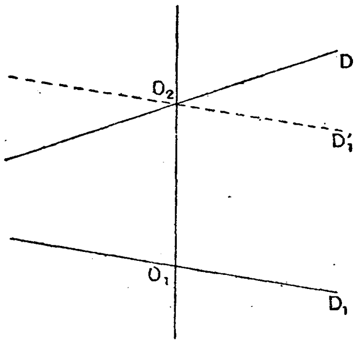

Fig. 322.

Le plan P étant son propre symétrique tant par rapport à D_1 que par rapport à D_2, si nous projetons les points M, M′, M″ en N, N′, N″ sur ce plan, les points N et N″ seront symétriques de N′ par rapport à D_1 et D_2 respectivement. N″ se déduira donc de N (Pl., **102**) par une rotation autour de Ox, d'angle égal au double de celui des deux axes.

Mais les droites NM, N″M″ sont égales, parallèles toutes deux à Ox et de même sens (contraire à celui de $N'M'$) : il est, dès lors, clair que la même rotation d'axe Ox qui amène N sur N'' amène aussi M sur M''.

3° Soient les deux axes D_1, D_2 (*fig.* 322), non situés dans un même plan; O_1O_2, leur perpendiculaire commune. Par le point O_2, menons la parallèle D'_1 à D_1. Entre les deux rotations de 180° que nous avons à composer, intercalons deux rotations successives de 180° autour de D_1' : ce qui est indifférent puisque ces deux rotations se détruisent.

Les deux symétries successives par rapport à D_1, D'_1, se composent en une translation parallèle à O_1O_2 et double de O_1O_2; les deux symétries relatives à D'_1, D_2, en une rotation autour de O_1O_2, double de celle qui amènerait D'_1 sur D_2 : le théorème est donc démontré.

Réciproque. — *Tout déplacement peut être décomposé en deux symétries par rapport à deux droites différentes.*

Ces deux droites doivent :

S'il s'agit d'une translation, être perpendiculaires à cette translation;

S'il s'agit d'une rotation ou d'un déplacement hélicoïdal, être perpendiculaires à l'axe et le rencontrer.

A ces conditions près, on peut choisir arbitrairement l'une des droites, l'autre étant alors déterminée.

Si, par exemple, c'est l'axe de la première symétrie qui est pris arbitrairement sous les conditions indiquées, le second axe se déduira du premier par une translation égale à la moitié de la translation et une rotation égale à la moitié de la rotation données. Ce second axe étant ainsi déterminé, les deux symétries produisent bien, en vertu du théorème précédent, un déplacement résultant identique au déplacement donné.

417. Théorème. — *Un nombre quelconque de déplacements hélicoïdaux ont pour déplacement résultant un déplacement hélicoïdal unique.*

S'il s'agit de rotations concourantes, le déplacement résultant est une rotation qui concourt avec les premières.

S'il s'agit de translations, le déplacement résultant est une translation.

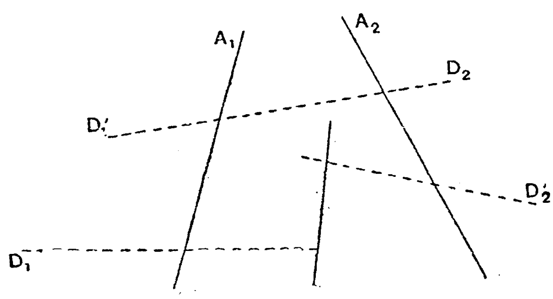

FIG. 323.

Il suffit évidemment[1] de démontrer le théorème pour deux déplacements : car pour en composer trois, on pourra composer les deux premiers, puis le déplacement résultant avec le troisième ; et ainsi de suite.

Soient donc deux déplacements hélicoïdaux[2] d'axes A_1, A_2 (*fig.* 323).

---

[1] Comparer Pl., **103**.

[2] Si l'un des déplacements donnés, le premier par exemple, est une translation, on prendra pour l'axe A, dans le raisonnement qui va suivre, une droite quelconque parallèle à cette translation.

Nous pourrons remplacer le premier par deux symétries d'axes D_1, D'_1, et cela, en prenant arbitrairement la seconde droite D'_1 parmi celles qui coupent à angle droit l'axe A_1. Nous pourrons de même remplacer le second déplacement par deux symétries d'axes D_2, D'_2, et cela, en prenant arbitrairement le premier axe D_2 parmi ceux qui coupent à angle droit A_2.

Nous ferons en sorte que D'_1 et D_2 coïncident : il suffira, pour cela de les faire coïncider toutes deux avec la perpendiculaire commune à A_1, A_2. Alors les symétries par rapport à ces deux axes se détruisent, et il reste les symétries par rapport à D_1, D'_2, lesquelles donnent bien un déplacement hélicoïdal unique.

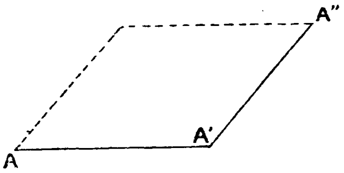

Fig. 323 *bis*.

Si les déplacements donnés sont des rotations et que les axes A_1, A_2 concourent en un point O, la droite D'_1 passera par ce même point et il en sera de même de D_1 et de D'_2. Le déplacement résultant sera donc une rotation d'axe passant par O.

Si les déplacements donnés sont des translations, les droites D'_1, D_1, D'_2 sont parallèles entre elles et le déplacement résultant est une translation.

Ce dernier fait est, d'ailleurs, évident *à priori*, et il est clair que la translation résultante est la diagonale d'un parallélogramme qui a pour côtés les deux composantes (*fig.* 323 *bis*).

C. Q. F. D.

418. Il suffit évidemment de combiner le théorème précédent avec les remarques du n° **415** pour arriver à l'énoncé que nous avions en vue :

Théorème. — *Deux figures égales peuvent toujours être amenées à coïncider :*

S'il y a un point commun, par une rotation ; dans le cas général, par un déplacement hélicoïdal.

EXERCICES

577. On fait tourner une figure invariable S autour d'un axe fixe D, l'angle de rotation étant quelconque; après quoi, on fait tourner une partie S' de la figure S autour d'un axe D' qui fait partie de S (le reste de la figure S restant fixe), l'angle de rotation étant également quelconque.

Montrer qu'on peut, en général, disposer des deux angles de rotation de manière à rendre une droite donnée de S' parallèle à une droite donnée arbitraire de l'espace (ou, ce qui revient au même, un plan arbitraire de S' parallèle à un plan donné arbitraire de l'espace). Quels sont les cas d'exception ?

578. On fait tourner une figure invariable S autour d'un axe fixe D, l'angle de rotation étant quelconque; après quoi, on fait tourner une partie S' de la figure S autour d'un axe D' qui fait partie de S, puis une partie S'' de S' autour d'un axe D'' qui fait partie de S', les deux angles de rotation étant également quelconques.

Montrer qu'on peut disposer des angles des trois rotations de manière que toutes les droites de la figure S'' deviennent parallèles aux droites correspondantes d'une figure donnée S''_0 égale à S''.

579. Trouver une rotation transformant respectivement deux points donnés en deux autres points donnés (la distance des deux premiers points étant, bien entendu, supposée égale à celle des deux autres).

Dans quels cas le problème est-il indéterminé ?

580. Construire l'axe de la rotation qui transforme deux demi-droites concourantes données OA, OB en deux demi-droites OA', OB' concourantes avec les premières (l'angle $\widehat{A'OB'}$ étant égal à $\widehat{AOB}$).

581. Étant admis que deux figures égales F, F' qui ont un point commun, dérivent l'une de l'autre par une rotation, en déduire que deux figures égales quelconques dérivent l'une de l'autre par un déplacement hélicoïdal.

(On montrera qu'il y a des droites de F qui sont parallèles à leurs homologues de F'. Soient alors P un plan de F perpendiculaire aux droites en question; P', son homologue; f, la partie de F située dans P; f', la partie correspondante de F'. On appliquera à f et à la projection de f' sur P le théorème du n° **102.** (Pl., liv. II.))

582. Quels sont les différents déplacements qui laissent inaltérée une droite donnée ?

583. Il y a une infinité de rotations qui transforment l'une dans l'autre deux droites données D, D'. Les axes de ces rotations ne sont autres que les droites appelées G_1, G_2 dans l'exercice 454.

Parmi ces rotations, il y en a deux dont l'angle est de 180°.

584. Tout déplacement hélicoïdal qui transforme l'une dans l'autre les deux droites données D, D' a son axe A parallèle à celui d'une des rotations considérées dans l'exercice précédent.

Quel est le lieu de l'axe A, lorsqu'en donne la direction de cet axe ?

Trouver ce lieu : 1° directement; 2° en utilisant la composition des déplacements et les deux exercices précédents. On prouvera que l'axe A rencontre toujours, à angle droit, l'une ou l'autre de deux droites fixes (suivant qu'il est parallèle à une droite G_1 ou à une droite G_2).

585. Trouver le lieu de l'axe d'un déplacement hélicoïdal, connaissant la direction de cet axe et deux points homologues.

586. Construire l'axe d'un déplacement hélicoïdal, connaissant un point de cet axe, la grandeur de la translation et deux points homologues.

587. Un déplacement donné quelconque peut, en général, être décomposé en deux rotations, dont une suivant une droite donnée quelconque. Quels sont les cas d'exception ?

588. On compose un déplacement hélicoïdal donné avec une quelconque des translations T parallèles à une direction donnée. Lieu de l'axe du nouveau déplacement ainsi obtenu, lorsqu'on fait varier la grandeur de la translation T. Montrer qu'en général il y a un et un seul de ces déplacements qui se réduit à une rotation.

589. On compose une rotation donnée avec une rotation donc l'axe est fixe et rencontre le premier, mais dont l'angle est variable. Lieu de l'axe de la rotation résultante.

590. Composer une rotation donnée avec une rotation d'axe donné rencontrant le premier mais d'angle inconnu, de manière que le déplacement résultant soit une transposition (ou plus généralement, une rotation d'angle donné).

591. Composer un déplacement hélicoïdal donné avec un déplacement inconnu, mais d'axe donné, de manière que le déplacement résultant soit une transposition.

592. Deux rotations (d'angles non nuls) dont les axes ne sont pas dans un même plan, ne se composent jamais suivant une rotation.

593. Lorsqu'un déplacement hélicoïdal ne se réduit ni à une translation, ni à une rotation, il n'existe aucun plan qui reste inaltéré par ce déplacement.

594. Mener, par un point donné, deux droites qu'un déplacement donné transforme l'une dans l'autre.

595. Étant donné un déplacement, trouver deux droites homologues entre elles, qui soient situées dans un même plan donné.

596. Construire un angle polyèdre, connaissant les bissectrices des faces.

Le problème peut-il être indéterminé?

Connaissant toutes les bissectrices, moins deux, trouver le lieu de ces dernières, de manière que l'indétermination ait lieu.

597. Déduire de la composition des rotations le lieu des symétriques d'un point donné, par rapport à toutes les droites qui passent par un point fixe et sont situées dans un plan fixe.

598. Décomposer une rotation R en deux rotations S, S′ dont les angles soient entre eux, l'axe de la rotation S étant, d'autre part donné et concourant avec celui de R. Quel est le lieu décrit par l'axe de la rotation S′ lorsque l'axe de la rotation S décrit un plan?

598 *bis*. Décomposer un déplacement hélicoïdal en deux déplacements hélicoïdaux S, S′ composés de translations égales entre elles et de rotations égales entre elles, tous deux dextrorsum ou tous deux sinistrorsum, l'axe du déplacement S étant donné.

599. Trouver deux déplacements hélicoïdaux d'axes donnés qui se composent suivant un déplacement d'axe également donné.

CHAPITRE II

SYMÉTRIES

419. Définitions. — Deux points M, M′ sont dits (comme en géométrie plane) *symétriques par rapport à un point* O (ou encore par rapport au *centre* O) (*fig.* 324) lorsque le point O est le milieu de la droite MM′.

Deux points M, M′ sont dits *symétriques par rapport à un plan* P (*fig.* 325), lorsque la droite qui les joint est perpendiculaire à ce plan et divisée par lui en deux parties égales.

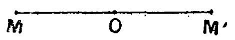

Fig. 324.

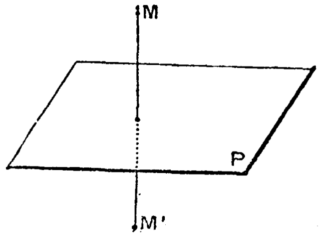

Fig. 325.

La *symétrique d'une figure* F, par rapport à un point ou à un plan, est la figure formée par les symétriques des différents points de F.

419 *bis*. Coïncident avec leurs symétriques par rapport à un point :

— le centre de symétrie ;

— les droites qui passent par ce point, et non d'autres droites : car toute droite qui coïncide avec sa symétrique doit contenir deux points symétriques l'une de l'autre, ce qui exige qu'elle passe par le centre ;

— les plans qui passent par le centre (et non d'autres plans, pour une raison tout analogue).

Coïncident avec leurs symétriques par rapport à un plan :

— les points du plan de symétrie, et non d'autres points ;

— les droites situées dans le plan de symétrie et celles qui lui sont perpendiculaires ;

— le plan de symétrie et les plans qui lui sont perpendiculaires.

420. Théorème. — *Deux figures symétriques d'une même troisième, par rapport à deux centres différents, sont égales entre elles.*

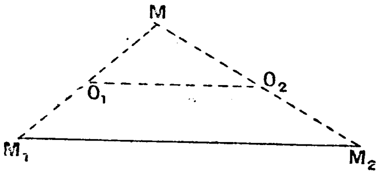

Fig. 326.

Si, en effet, M_1 et M_2 sont les symétriques de M par rapport aux deux centres O_1, O_2 (*fig.* 326), la droite M_1M_2 est (Pl., **55**) parallèle à O_1O_2 et double de O_1O_2.

Donc, les figures symétriques d'une même troisième, par rapport à O_1 et O_2 respectivement, se déduisent l'une de l'autre par une translation parallèle à O_1O_2 et double de O_1O_2.

Théorème. — *Deux figures symétriques d'une même troisième, l'une par rapport à un point, l'autre par rapport à un plan, sont égales entre elles.*

Il suffit, d'après le théorème précédent, de faire la démonstration pour une situation déterminée du centre de symétrie : elle sera, par cela même, faite pour toutes les autres positions de ce point.

Nous pouvons, par conséquent, supposer que le centre de symétrie O est situé dans le plan de symétrie P. Dans ces conditions, nous allons faire voir que *les symétriques d'une même figure* F, *l'une par rapport au point* O, *l'autre par rapport au plan* P, *se déduisent l'une de l'autre par une rotation de* 180° *autour d'une droite :* cette droite est la perpendiculaire Ox au plan P, menée par le point O.

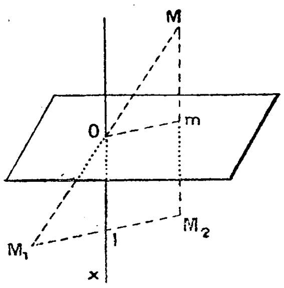

Fig. 327.

Soient, en effet, M un point quelconque de F (*fig.* 327) ; M_1, son symétrique par rapport à O ; M_2, son symétrique par rapport à P, de sorte que la droite MM_2, perpendiculaire au plan P, a son milieu m dans ce plan. La droite Ox, parallèle à MM_2 et menée par le milieu de MM_1, rencontre en son milieu le troisième côté M_1M_2 du triangle MM_1M_2.

D'ailleurs, M_1M_2 est parallèle à Om (qui joint les milieux de MM_1 et de MM_2) et, par suite, perpendiculaire à Ox. Donc, les points M_1, M_2 sont bien symétriques l'un de l'autre par rapport à Ox.

C. Q. F. D.

Corollaire. — *Deux figures symétriques d'une même troisième* F, *par rapport à deux plans différents, sont égales entre elles*, puisqu'elles sont égales à la symétrique de la figure F par rapport à un point quelconque de l'espace.

On peut d'ailleurs (exercice 600) constater directement que ces figures se déduisent l'une de l'autre par une rotation ou une translation convenable.

421. Théorème. — *Une figure plane est égale à sa symétrique par rapport à un point ou à un plan quelconque.*

Le théorème est évident lorsque le plan de symétrie coïncide avec le plan de la figure, puisque celle-ci coïncide alors avec sa symétrique. Il est dès lors vrai dans tous les cas, d'après le numéro précédent.

En particulier, *la figure symétrique d'un plan est un plan ;*

La figure symétrique d'une droite est une droite ; la figure symétrique d'un segment de droite est un segment de droite égal au premier ;

Deux angles symétriques l'un de l'autre sont égaux ;

La figure symétrique d'une circonférence est une circonférence ; etc.

421 *bis*. **Corollaires.** — I. *Une droite et un plan, perpendiculaires entre eux, ont pour symétriques une droite et un plan perpendiculaires entre eux :* cela résulte de la définition du plan perpendiculaire à une droite, puisqu'un angle droit a pour symétrique un angle droit.

II. *L'angle d'une droite et d'un plan est égal à l'angle de leurs symétriques :* cela résulte de la définition et du corollaire précédent. De même :

III. *Deux dièdres symétriques sont égaux.*

422. Théorème. — *Deux dièdres symétriques l'un de l'autre, par rapport à un point ou à un plan, sont de sens contraires* (*lorsque les sens choisis sur les arêtes se correspondent*).

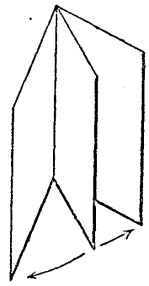

Fig. 328.

Si, en effet, le plan de symétrie est le plan de l'une des faces (*fig.* 328), les arêtes coïncident (et cela en direction *et sens*) pendant que les faces non communes sont de part et d'autre de la face commune.

Corollaire. — *Dans deux figures symétriques, les trièdres correspondants ont leurs dispositions inverses.*

REMARQUE. — Cette dernière proposition résulte d'ailleurs immédiatement du n° **372** : car si on prend pour centre de symétrie le sommet d'un des trièdres considérés, on retombe évidemment sur la construction des trièdres symétriques indiquée au n° **370**.

Les *trièdres symétriques* considérés en cet endroit satisfont donc bien à la définition générale des figures symétriques, telles que nous les considérons actuellement.

D'après ce qui précède — au lieu que deux figures symétriques par rapport à une droite sont superposables — deux figures symétriques par rapport à un point ou à un plan, tout en ayant tous leurs côtés, tous leurs angles, tous leurs dièdres égaux à chacun, *ne sont pas en général superposables*, parce que leur disposition est inverse.

423. Théorème. — *Deux polyèdres symétriques sont équivalents.*

Nous distinguerons deux cas :

1° *Les polyèdres considérés sont des pyramides.*

Ces pyramides ont leurs bases égales (**421**) ; leur hauteurs sont d'ailleurs symétriques l'une de l'autre (**421** *bis*, Coroll. I) et, par conséquent, égales : elles sont donc équivalentes ;

2° *Cas général.* Décomposons l'un des polyèdres donnés en pyramides. L'autre sera décomposable en pyramides, symétriques respectivement des premières. Ces pyramides étant équivalentes chacune à chacune, les polyèdres totaux le sont aussi.

C. Q. F. D.

424. On dit qu'une figure *admet pour axe de symétrie* la droite D, lorsqu'elle coïncide avec sa symétrique par rapport à D.

De même, une figure est dite *admettre un plan* ou *un centre de symétrie*, lorsqu'elle coïncide avec sa symétrique par rapport à ce plan ou à ce centre.

Une figure F qui a un plan P ou un centre C de symétrie (¹) est

---

(1) Il est clair que la même conclusion ne s'appliquerait pas au cas où, au lieu d'un centre ou d'un plan de symétrie, la figure F admettrait un *axe* de symétrie.

superposable à sa symétrique F′ par rapport à un point C′ ou à un plan P′ quelconque, puisque F et F′ peuvent être considérées comme symétriques d'une même figure (à savoir F elle-même) l'une par rapport à P ou à C, l'autre par rapport à P′ ou à C′.

Toutefois, il faut bien observer que, dans la superposition de F et de F′, ce ne sont pas, en général, les points homologues (c'est-à-dire symétriques l'un de l'autre par rapport à P′ ou à C′) que l'on arrive à faire coïncider. C'est, par exemple, ce qui se produit dans le cas du trièdre isoscèle (**383**) : un trièdre isoscèle est effectivement (voir exercice 480) une figure qui a un plan de symétrie.

EXERCICES

600. Deux figures sont symétriques d'une même troisième, par rapport à deux plans différents. Quel est le déplacement qui amène la première à coïncider avec la seconde ?

(Distinguer deux cas, suivant que les deux plans sont concourants ou parallèles).

601. Même question pour deux figures symétriques d'une même troisième, l'une par rapport à un plan, l'autre par rapport à un point non situé dans ce plan.

602. Une figure limitée en tous sens peut-elle avoir deux centres de symétrie ?

603. Tout plan mené par le point d'intersection des diagonales d'un parallépipède divise celui-ci en deux parties équivalentes.

604. *Symétrie oblique.* — Un point quelconque M étant considéré, soit M′ le point tel que la droite MM′ soit parallèle à une droite fixe D et divisée en deux parties égales par un plan P (sécant à D) : le point M′ est dit dériver de M par *symétrie oblique*.

Prouver que :

La figure obliquement symétrique d'une droite est une droite ;

La figure obliquement symétrique d'un plan est un plan.

Deux polyèdres obliquement symétriques l'un de l'autre sont équivalents. (On peut utiliser pour cela l'ex. 565.)

605. Même question lorsque le point M′ est défini, à l'aide du point M, par la condition que la droite MM′ soit parallèle à un plan fixe P et divisée en deux parties égales par une droite fixe D (sécante à P).

On démontrera que cette transformation peut se ramener à deux symétries obliques, telles qu'elles ont été définies à l'exercice précédent, ou encore à une symétrie oblique et une symétrie par rapport à un point.

606. Appliquer les conclusions de l'exercice précédent à l'exercice 568.

607. Si une figure F est égale à la symétrique, par rapport à un plan, d'une figure F′, les deux figures peuvent être amenées à coïncider par une rotation précédée ou suivie d'une symétrie par rapport à un plan perpendiculaire à l'axe de rotation ou par rapport à un point situé sur cet axe.

608. Inscrire, dans un angle polyèdre convexe donné, un angle polyèdre tel que la somme de ses faces soit minima. Cas du trièdre.

CHAPITRE III

HOMOTHÉTIE ET SIMILITUDE

425. Définition. — De même qu'en géométrie plane, nous nommerons *homothétique* d'un point quelconque M, par rapport à un point O, appelé *centre d'homothétie*, et à un nombre k, appelé *rapport d'homothétie* ou *de similitude*, le point M′ obtenu en joignant OM et portant sur cette droite, à partir du point O′, un segment donné par la relation

$$\frac{\mathrm{OM}'}{\mathrm{OM}} = k.$$

Pour achever de définir l'homothétie, on doit indiquer s'il faut prendre le segment OM′ dans le sens OM (homothétie *directe*) ou dans le sens opposé (homothétie *inverse*).

La symétrie par rapport à un point est, ainsi que nous l'avons noté (Pl., **140**, Rem.), un cas particulier d'homothétie inverse, le cas où le rapport de similitude est égal à 1.

426. Le théorème déjà énoncé en géométrie plane (n° **141**) :

Théorème. — *Dans deux figures homothétiques, la droite qui joint deux points quelconques de l'une des figures et celle qui joint les points homologues de l'autre sont toujours parallèles et dans le rapport de similitude; elle sont de même sens ou de sens contraires, suivant que l'homothétie est directe ou inverse*,

subsiste, avec sa démonstration, en géométrie de l'espace. Il en est de même des

Corollaires. — I. *La figure homothétique d'une droite est une droite parallèle à la première;*

d'où résulte, moyennant le n° **330** :

II. *La figure homothétique d'un plan est un plan parallèle au premier ;*

Il en est de même également des corollaires.

III. *Deux angles homothétiques l'un de l'autre sont égaux.*

La figure homothétique d'un triangle est un triangle semblable au premier

et (moyennant le corollaire II et le théorème du n° **393**)

IV. *La figure homothétique d'un polygone plan est un polygone semblable au premier;*

Plus généralement, *la figure homothétique d'une figure plane est une figure plane semblable à la première* (1).

V. *La figure homothétique d'une circonférence est une circonférence, les centres étant homothétiques l'un de l'autre et le rapport des rayons étant égal au rapport de similitude.*

Enfin les corollaires I et II donnent :

VI. *Une droite et un plan perpendiculaires ont pour homothétiques une droite et un plan perpendiculaires*

et, moyennant les n° **358** *bis*, **384** *bis* :

VII. *Dans deux figures homothétiques,*

Les dièdres homologues sont égaux;

Les angles polyèdres homologues sont égaux si l'homothétie est directe; symétriques, si l'homothétie est inverse.

427. La réciproque du théorème précédent :

Réciproque. — *Deux figures étant données, s'il existe deux points* O, O' *tels que la droite qui joint le point* O *à un point* M *quelconque de la première figure et celle qui joint le point* O' *au point* M' *homologue de la seconde soient constamment parallèles et dans un rapport donné* k (*toujours de même sens ou toujours de sens contraires*), *les deux figures sont homothétiques*,

subsiste également et se démontre comme en géométrie plane, moyennant la convention qui consiste à considérer, comme cas limite de deux figures homothétiques, deux figures égales qui se déduisent l'une de l'autre par une translation.

---

(1) Il peut sembler que cette proposition ne soit autre que la définition des figures semblables (Pl., **146**). Il n'en est rien : deux figures semblables sont en effet telles que l'une soit égale à une homothétique de l'autre *par rapport à un point de son plan* (seule espèce d'homothétie considérée en géométrie plane); au lieu que cette dernière restriction n'est point supposée dans l'énoncé actuel.

Il en résulte, comme en géométrie plane (**144**) :

Théorème. — *Deux figures homothétiques d'une même troisième sont homothétiques entre elles; le rapport d'homothétie est le quotient des rapports primitifs, et cette homothétie est directe ou inverse, suivant que les homothéties primitives sont de même nom ou de noms contraires.*

Les trois centres d'homothétie sont sur une même ligne droite (dite *axe d'homothétie*).

La démonstration de cette dernière partie de l'énoncé : *les trois centres sont en ligne droite*, peut toutefois être simplifiée lorsqu'on fait intervenir l'homothétie dans l'espace.

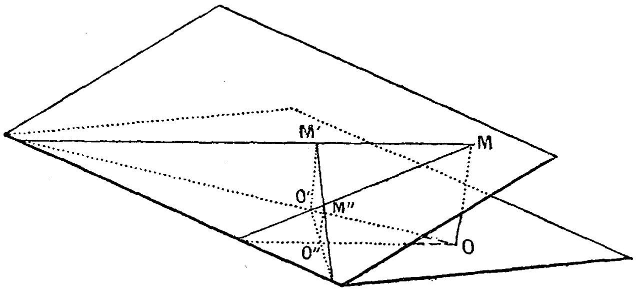

Fig. 329.

Soient, en effet, O un point de la première figure (*fig.* 329), O' et O'' ses homologues dans la seconde et dans la troisième; M, un point de la première figure, situé en dehors du plan $OO'O''$; M', M'' ses homologues. Le centre d'homothétie des deux premières figures est à l'intersection de OO', MM'; celui de la première et de la troisième, à l'intersection de OO'', MM''; celui des deux dernières, à l'intersection de $O'O''$, $M'M''$. Ces trois centres sont donc évidemment sur une même droite, à savoir l'intersection des plans $OO'O''$, $MM'M''$.

Remarque. — *Si les deux homothéties primitives sont de même nom, avec le même rapport de similitude, les deux figures homothétiques d'une même troisième se déduisent l'une de l'autre par translation.* C'est le cas limite auquel il a été fait allusion tout à l'heure.

428. Théorème. — *Lorsque quatre figures sont homothétiques entre elles, les six centres d'homothétie sont dans un même plan* (dit *plan d'homothétie*) *et forment un quadrilatère complet dont les côtés sont les quatre axes d'homothétie des figures prises trois à trois.*

Désignons par F_1, F_2, F_3, F_4 les quatre figures; par S_{12}, S_{13}, etc., les centres d'homothétie des figures F_1, F_2; F_1, F_3; etc. Les points S_{12}, S_{13}, S_{23} sont sur une même droite, l'axe d'homothétie des figures F_1, F_2, F_3. Le point S_{12} est d'ailleurs, avec les points S_{14} et S_{24}, sur une seconde droite, l'axe d'homothétie des figures F_1, F_2, F_4. Un plan passant par ces deux droites (et ce plan existe toujours (¹), puisque les deux droites ont un point commun) contiendra les deux autres axes d'homothétie : celui de F_1, F_3, F_4 (dont il contiendra les points S_{13}, S_{14}) et celui de F_2, F_3, F_4 (dont il contiendra les points S_{23}, S_{24}) : il contiendra donc les six centres, lesquels auront bien la disposition indiquée par l'énoncé.

429. Définition. — Deux figures sont dites *semblables* si l'une d'elles est égale à l'une des homothétiques *directes* de l'autre.

Cette définition concorde d'ailleurs (**426**, Coroll. IV), dans le cas des figures planes, avec celle qui a été donnée en géométrie plane.

Des différents corollaires du n° **426** résulte d'ailleurs :

Théorème. — *Deux polyèdres semblables ont leurs faces homologues semblables, avec le même rapport de similitude, et leurs angles polyèdres homologues égaux chacun à chacun.*

Nous allons démontrer la réciproque :

Réciproque. — *Si deux polyèdres ont leurs faces semblables chacune à chacune, avec le même rapport de similitude, leurs angles polyèdres égaux chacun à chacun, ces éléments étant pareillement assemblés* (²) *: ils sont semblables.*

Nous allons d'abord démontrer que *si deux polyèdres ont leurs faces égales et leurs angles polyèdres égaux chacun à chacun et pareillement assemblés, ils sont égaux.*

Pour cela, nous transporterons l'un des polyèdres sur l'autre, de manière qu'une de ses faces — soit, par exemple, la face F, dont

---

(1) Il peut ne pas être unique, si les deux axes d'homothétie qui le déterminent coïncident : les centres S_{12}, S_{13}, ... sont alors tous en ligne droite.

(2) C'est-à-dire que ces éléments se correspondent dans les deux polyèdres de manière qu'à deux faces du premier, contiguës suivant une arête, correspondent, dans le second, deux faces respectivement semblables aux premières et aussi contiguës entre elles; qu'à des faces formant un angle polyèdre, correspondent des faces (semblables aux premières) formant l'angle polyèdre égal au premier, etc.

les sommets sont A, B, C, D, E (*fig.* 330) — vienne coïncider avec la face homologue (et par conséquent égale) F′ (aux sommets A′, B′, C′, D′, E′) de l'autre.

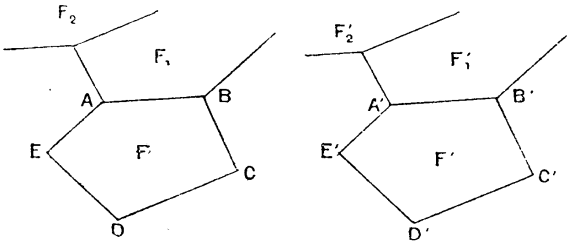

Fig. 330.

Alors l'angle polyèdre en A et l'angle polyèdre en A′, figures égales entre elles qui ont trois points communs non en ligne droite (puisque les faces F et F′ sont entièrement superposées), coïncident; et comme on peut raisonner de même pour chaque sommet de F, on voit que toute face F_1 du premier polyèdre, contiguë à F, coïncide avec son homologue.

Mais on peut recommencer le raisonnement précédent, en partant de F_1, et démontrer la même conclusion pour toute face F_2 contiguë à F_1, ; continuant ainsi de proche en proche, on peut manifestement montrer que la coïncidence est établie pour toutes les faces. Les deux polyèdres sont donc bien égaux.

Prenant alors, conformément à l'énoncé, deux polyèdres P, P′ qui ont toutes leurs faces semblables, avec le même rapport de similitude k, et leurs angles polyèdres égaux, ces éléments étant pareillement assemblés, nous considérerons le polyèdre P_1, homothétique direct de P avec k pour rapport d'homothétie. Ce polyèdre aura toutes ses faces égales et tous ses angles polyèdres égaux aux faces et aux angles polyèdres homologues de P′; il sera donc égal à P′, et les polyèdres P, P′ sont semblables.

C. Q. F. D.

430. Théorème. — *Deux polyèdres semblables peuvent être décomposés en pyramides semblables et pareillement assemblées.*

Il suffira, à cet effet, après avoir amené les deux solides dans la position où ils sont homothétiques, de décomposer l'un d'eux en pyramides (**395**) et l'autre, en pyramides homothétiques des premières.

On démontrerait d'ailleurs, par une marche analogue à celle qui a été suivie dans le plan (Pl., **149**) et à celle du n° précédent, que, *réciproquement, deux polyèdres, composés de pyramides semblables et pareillement assemblées sont semblables.*

431. Théorème. — *Le rapport des volumes de deux polyèdres semblables est égal au cube du rapport de similitude.*

Comme précédemment (**423**), nous distinguerons deux cas :

1° *Les polyèdres sont des pyramides.*

Soient B et H la base et la hauteur de la première pyramide ; B′ et H′, la base et la hauteur de la seconde ; k le rapport de similitude : on a (Pl., **257**)

$$B' = k^2B, \; H' = kH$$

et le rapport des volumes sera

$$\frac{\frac{1}{3}B'.H'}{\frac{1}{3}.B.H} = \frac{B'}{B}. \; \frac{H'}{H} = k^3.$$

2° *Cas général.* — On décomposera les polyèdres donnés en pyramides semblables et pareillement assemblées. Deux pyramides correspondantes quelconques étant dans le rapport k^3, leurs sommes sont, en vertu d'un théorème connu sur les proportions (¹), dans le même rapport.

C. Q. F. D.

EXERCICES

609. Si deux figures se correspondent point par point, de manière que la ligne joignant deux points quelconques de l'une soit parallèle à celle qui joint les points homologues de l'autre, elles sont homothétiques.

610. Si deux figures se correspondent point par point, de manière que A, B, C, étant trois points quelconques de l'une, A′, B′, C′ leurs homologues dans l'autre, l'angle $\widehat{B'A'C'}$ soit toujours égal à $\widehat{BAC}$, elles sont semblables, ou l'une est semblable à la symétrique de l'autre.

---

(1) Tannery, *Leçons d'Arithmétique*, n° **218**. Comparer Pl., **257**.

611. Deux figures semblables (mais non égales) peuvent être amenées à coïncider à l'aide d'une rotation, suivie d'une homothétie directe par rapport à un point situé sur l'axe de rotation.

Énoncé analogue pour deux figures dont l'une est semblable à la symétrique de l'autre.

612. Deux figures limitées en tout sens ne peuvent pas être homothétiques entre elles de plus de deux façons différentes.

613. Lieu du centre d'une homothétie dans laquelle les homologues de trois points donnés sont respectivement dans trois plans donnés.

614. Lieu du centre d'une homothétie dans laquelle le rapport de similitude est donné, sachant que l'homologue d'une droite donnée D rencontre une droite donnée D′.

615. Lieu des droites menées par un point donné et divisées dans un rapport donné par ce point et deux plans donnés ; par trois plans concourants donnés.

616. Prouver que l'on peut trouver, d'une infinité de façons, deux polyèdres, P, Q, tels que leurs volumes soient dans le même rapport que leurs surfaces. Déterminer Q, connaissant P et un polyèdre semblable à Q.

PROBLÈMES

PROPOSÉS SUR LE SEPTIÈME LIVRE

617. Deux tétraèdres sont égaux ou symétriques :

1° S'ils ont un dièdre égal compris entre faces égales chacune à chacune ;

2° S'ils ont un angle trièdre égal ou symétrique, compris entre arêtes égales chacune à chacune ;

3° S'ils ont une face égale adjacente à trois dièdres égaux chacun à chacun ;

4° S'ils ont une arête égale adjacente à deux trièdres égaux ou symétriques chacun à chacun ;

5° S'ils ont leurs six arêtes égales chacune à chacune.

Il est toutefois sous-entendu, dans tous ces énoncés, que les éléments égaux ou symétriques sont assemblés de la même façon.

618. Déduire des cas d'égalité énoncés à l'exercice précédent, des cas de similitude des tétraèdres.

619. Étant donnés une droite D et deux points A, B, trouver sur D le point M tel que $MA + MB$ soit minimum, et le point N tel que $NA - NB$ soit maximum.

620. Trouver, sur une droite donnée, le point tel que la somme de ses distances à deux droites parallèles données soit minima.

621. Que deviennent les énoncés des exercices 604 et 605 lorsque la droite MM′ est divisée par le plan P (ex. 604) ou la droite D (ex. 605) dans un rapport donné quelconque, et non en deux parties égales ?

622. Une rotation et une translation obliques entre elles se composent suivant un déplacement hélicoïdal dont l'axe est parallèle à celui de la rotation. Si l'on change l'ordre des deux opérations, on a deux mouvements dont les axes sont symétriques l'un de l'autre par rapport au plan qui passe par l'axe de la rotation et par la perpendiculaire commune menée à ce dernier et à la translation.

623. Plus généralement, en composant deux déplacements donnés dans un ordre déterminé, puis dans l'ordre inverse, on obtient deux déplacements dont les axes sont symétriques l'un de l'autre par rapport à la perpendiculaire commune aux deux axes primitifs (la grandeur de la rotation et celle de la translation résultantes étant d'ailleurs les mêmes dans les deux cas).

624. L'ordre dans lequel on compose deux déplacements n'est indifférent que si ces déplacements sont :

1° Ou deux déplacements de même axe;

2° Ou deux translations;

3° Ou deux rotations de 180° autour d'axes qui se coupent à angle droit.

625. Étant donnés trois plans P, Q, R passant par une même droite D et, dans l'un d'eux, une droite SA sécante à D en S, il existe, en général :

1° Un trièdre ayant une arête suivant SA et tels que P, Q, R soient les plans bissecteurs de ses dièdres ou de leurs suppléments (ex. 488). Discuter, suivant les cas, si c'est l'une ou l'autre de ces deux circonstances qui se présente. Lorsque, P, Q, R et S restant fixes, SA varie, le plan de la face opposée à SA passe par une droite fixe ;

2° Un trièdre tel que P, Q, R soient chacun perpendiculaire à une des deux faces, en passant par la bissectrice de cette face ou celle de son supplément, l'une de ces bissectrices étant SA. Discussion analogue à celle de 1°. Lieu des arêtes du trièdre lorsque SA varie, P, Q, R et S restant fixes;

3° Un trièdre tel que P, Q, R passent chacun par une de ses arêtes et la bissectrice de la face opposée ou de son supplément, l'une de ces bissectrices étant SA.

626. Étant données trois droites concourantes Sa, Sb, Sc d'un même plan, il existe, en général :

1° Un trièdre tel que Sa, Sb, Sc soient les droites dont il est question dans l'exercice 489, l'une des faces étant dans un plan donné P passant par Sa. Lieux des arêtes de ce trièdre lorsque P tourne autour de Sa;

2° Un trièdre tel que Sa, Sb, Sc soient les droites dont il est question dans l'exercice 491, l'un des plans bissecteurs mentionnés en cet endroit étant un plan donné P mené par Sa;

3° Un trièdre tel que Sa, Sb, Sc soient les droites dont il est question dans l'exercice 493, l'un des plans bissecteurs mentionnés en cet endroit étant un plan donné P passant par Sa.

627. (Généralisation de la 1^{re} partie de l'ex. précédent). Construire un angle polyèdre, connaissant les bissectrices des suppléments de ses faces. — Le problème est, en général, possible lorsque le nombre des faces est pair. Lorsque ce nombre est impair, le problème est, au contraire, impossible ou indéterminé. Quel est, dans ce dernier cas, le lieu d'une arête de l'angle polyèdre cherché ?

Lorsqu'on donne toutes les bissectrices moins une, quel est le lieu sur lequel doit se trouver cette dernière pour que le problème soit possible ?

628. Étant donnés trois plans P, Q, R passant par une même droite D et, dans l'un d'eux, une droite SA sécante à D en S, trouver un trièdre tel que P, Q et R passent chacun par une de ses arêtes et la bissectrice de la face opposée, l'une des arêtes étant SA. (Utiliser ex. 597.)

LIVRE VIII

LES CORPS RONDS

CHAPITRE PREMIER

DÉFINITIONS GÉNÉRALES — CYLINDRE

432. Parmi les surfaces qui ne sont pas planes, les plus simples sont les surfaces *cylindriques*, les surfaces *coniques* et les surfaces de *révolution*.

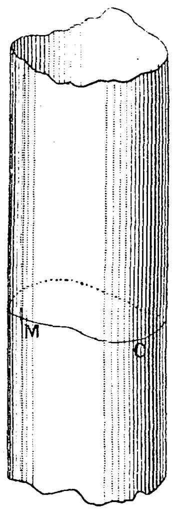

FIG. 331.

On nomme surface *cylindrique* (*fig*. 331), ou simplement *cylindre*, la surface engendrée par une droite, dite *génératrice*, qui se déplace en restant parallèle à une droite fixe.

Il est clair qu'on déterminera une surface cylindrique en se donnant : 1° la direction commune des génératrices ; 2° un point de chacune d'elles. On prend, en général, ces points de manière qu'ils varient continûment lorsque la génératrice varie continûment et forment une ligne (C, *fig*. 331), qui est dite la *directrice*.

Il est clair qu'on peut considérer comme directrice toute ligne tracée sur la surface et rencontrant toutes les génératrices. Il est souvent avantageux de prendre pour directrice une courbe plane.

Le plan est une surface cylindrique, à savoir celle qu'on obtient (**332**) en prenant pour directrice une droite.

433. Définition. — Une droite est dite *tangente à une surface*, en un point A de cette surface, lorsqu'elle est tangente en A à une ligne tracée sur la surface.

Théorème. — *Toutes les tangentes que l'on peut mener à une surface cylindrique, en un point de cette surface, sont dans un même plan.*

Ce plan est dit le *plan tangent au cylindre*, mené au point considéré.

Le plan tangent au cylindre en un point contient la génératrice qui passe par ce point et est le même tout le long de cette génératrice.

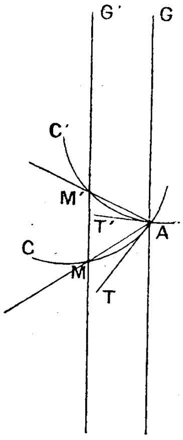

Fig. 332.

Soient A le point donné (*fig.* 332); G, la génératrice qui passe par ce point; C, une courbe tracée sur la surface par le point A et admettant une tangente AT, distincte (¹) de G; C', une autre courbe tracée par A sur la surface et admettant une tangente AT'. Soient maintenant M' un point de C', voisin de A; M, le point où la génératrice G' du point M' vient rencontrer la courbe C (²) : les deux droites AM, AM' sont, avec G, dans un même plan (le plan des deux génératrices G, G'). Donc leurs positions limites seront aussi (³) dans un même plan avec G : autrement dit, toutes les droites telles que AT' seront dans le même plan GAT.

Ce plan contient bien la droite G et est bien le même en tous les points de G (comme position limite (³) du plan GG').

434. *Une surface cylindrique peut être considérée comme le lieu des positions que vient occuper la directrice, lorsqu'on lui fait subir successivement les différentes translations parallèles aux génératrices.* Car, dans ces différentes translations, chaque point de la directrice se déplace évidemment sur la génératrice correspondante et décrit toute cette génératrice.

Les sections d'une surface cylindrique par des plans parallèles entre

---

(1) Nous ne considérerons pas le cas où il ne passerait pas, par le point A, de courbe ayant une tangente distincte de G. Ce cas, qui n'est pas théoriquement impossible, ne se présente pas dans les cylindres que l'on a à considérer habituellement.

(2) On démontre que ce point existe nécessairement lorsque la tangente AT est distincte de G.

(3) Nous admettons ici (voir Pl., **104**, note) les propositions suivantes : *Quand trois droites variables issues d'un même point sont dans un même plan, il en est de même de leurs positions limites (si celles-ci existent). — Quand une droite variable et un plan variable qui la contient tendent chacun vers une position limite, la position limite du plan contient la position limite de la droite.* Ces propositions se démontrent d'ailleurs aisément (V. Exercices 486 et 487).

eux (mais non parallèles aux génératrices) (*fig.* 333) *sont égales.* Car elles se déduisent l'une de l'autre par translation (**335**).

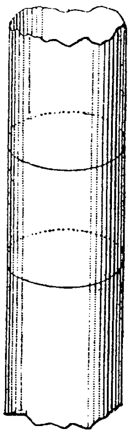

Fig. 333.

En particulier, on désigne sous le nom de *sections droites* du cylindre, les sections par des plans perpendiculaires aux génératrices. On voit que toutes les sections droites d'un même cylindre sont égales entre elles.

Dans le cas où le plan sécant est parallèle à la direction commune des génératrices, la section se compose évidemment d'une ou plusieurs de celles-ci.

435. On nomme plus spécialement *cylindre* le volume obtenu en coupant une surface cylindrique par deux plans parallèles (*fig.* 334) et limité, par conséquent, par une portion de surface cylindrique et deux aires planes égales entre elles (dites *bases*).

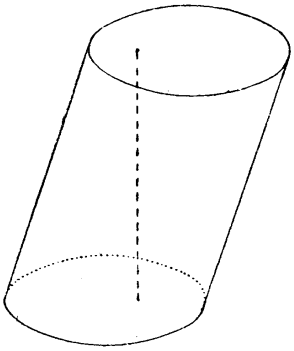

Fig. 334.

La *hauteur* du cylindre est la distance des plans des bases.

Un cylindre est dit *droit*, lorsque ses génératrices sont perpendiculaires aux plans des bases : autrement dit, lorsque celles-ci sont des sections droites.

436. On nomme surface *conique* (*fig.* 335) ou plus simplement *cône*, la surface engendrée par une droite (dite *génératrice*) qui se déplace en passant par un point fixe (dit *sommet* du cône).

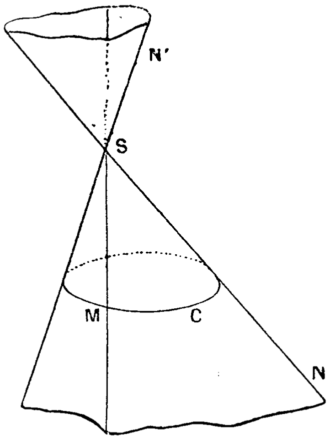

Fig. 335.

Il est clair qu'on déterminera une surface conique en se donnant : 1° le sommet; 2° un point de chaque génératrice. On prend, en général, ces points de manière qu'ils varient continûment et forment une ligne (C, *fig.* 335), qui est dite *directrice* ou *base* du cône. On peut considérer comme directrice toute ligne tracée sur la surface et rencontrant toutes les génératrices.

Le plan est une surface conique, à savoir celle qu'on obtient en prenant pour directrice une droite (**323**).

Théorème. — *Les tangentes que l'on peut mener à une surface conique, en un point autre que le sommet, sont toutes dans un même plan.*

Ce plan est dit le *plan tangent* au cône au point considéré.

Le plan tangent au cône contient la génératrice du point de contact et est le même en tous les points d'une même génératrice.

Même démonstration que pour le cylindre (**433**).

Une surface conique peut être considérée comme le lieu des positions que vient occuper la directrice, lorsqu'on lui fait subir successivement toutes les homothéties qui ont pour centre le sommet.

Lorsque la directrice est fermée, ses homothétiques directes forment une première portion ou *nappe* du cône, pendant que ses homothétiques inverses forment la seconde nappe.

Ces deux nappes sont alors séparées l'une de l'autre par le sommet (*fig.* 335).

Les sections d'un même cône par des plans parallèles sont semblables (**393**, Rem.).

La section d'un cône par un plan passant par le sommet se compose évidemment d'une ou plusieurs génératrices.

437. On appelle, plus spécialement, *cône*, le volume obtenu en coupant une nappe de surface conique par un plan, autrement dit, limité par une aire plane (dite *base*) et une portion de surface conique (*fig.* 336).

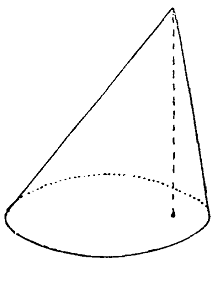

Fig. 336.

La *hauteur* du cône est la distance du sommet au plan de la base.

On nomme *tronc de cône* (*fig.* 337) le volume obtenu en coupant un cône par un plan parallèle à la base; autrement dit, limité par une portion de surface conique et deux aires planes semblables (dites *bases*). La *hauteur* du tronc du cône est la distance des plans des bases.

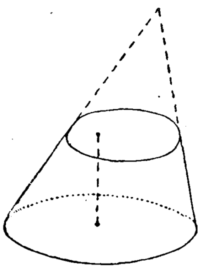

Fig. 337.

438. On nomme *surface de révolution*, la surface engendrée par une ligne C qui se déplace en tournant autour d'un axe fixe, l'angle de rotation prenant successivement toutes les valeurs possibles.

Dans ces conditions, un point quelconque M de C se déplacera, d'après ce que nous savons, dans le plan mené par ce point perpendiculairement à l'axe : il décrira dans ce plan, un cercle ayant son centre sur l'axe.

Ce cercle est dit un *parallèle* de la surface.

Comme il passe un parallèle par tout point de la surface, on voit qu'*une surface de révolution peut être considérée comme le lieu décrit par un cercle qui varie* (en changeant, en général, de rayon) *de manière que son centre décrive une droite, son plan restant perpendiculaire à cette droite.*

On achèvera de définir la surface en imposant au cercle variable la condition de rencontrer constamment la ligne donnée C. Comme, une fois l'axe donné, il suffit de se donner un point quelconque d'un parallèle pour le déterminer entièrement, on voit que *la ligne* C *peut être remplacée par n'importe quelle autre courbe tracée sur la surface et rencontrant tous les parallèles.*

On choisit, le plus souvent, pour la ligne C, la section de la surface par un plan quelconque passant par l'axe. Cette section MM′ (*fig.* 338), dite *méridienne* de la surface, est évidemment symétrique par rapport à l'axe, puisqu'elle contient les points diamétralement opposés de chaque parallèle. On pourra, par conséquent, se contenter d'en considérer une moitié, par exemple la partie située d'un seul et même côté de l'axe.

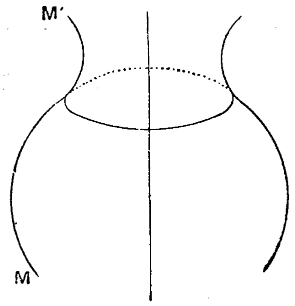

FIG. 338.

On nomme *méridiens* les diverses positions occupées par la méridienne, lorsque son plan tourne autour de l'axe. Par chaque point de la surface, il passe évidemment un méridien.

Une surface de révolution admet une infinité de plans de symétrie : elle est symétrique par rapport à tout plan passant par l'axe, car un parallèle quelconque jouit de cette propriété, en vertu du n° **419** *bis* et du n° **62** (Pl., liv. II).

On démontre qu'en un point quelconque d'une surface de révolution existe, en général, un *plan tangent*, c'est-à-dire que les tangentes que l'on peut mener aux différentes courbes tracées sur la surface par ce point, sont dans un même plan. Ce plan doit, d'après ce qui précède, coïncider avec son symétrique par rapport au plan méridien qui passe au point considéré, et, par suite, être perpendiculaire à ce dernier plan.

439. Après le plan, les surfaces cylindriques les plus simples sont celles qui ont pour directrices des cercles.

Parmi celles-là, on considère, en particulier, celle qui a pour section droite un cercle et qui est la *surface cylindrique de révolution :* cette surface est, en effet, celle qu'on obtient en faisant tourner une droite D autour d'un axe A qui lui est parallèle, car, dans ce mouvement, la droite reste constamment parallèle à A, pendant qu'un quelconque de ces points décrit un cercle dont le plan est perpendiculaire à A; et inversement, tout cylindre dont la section droite est un cercle peut être obtenu de cette manière, l'axe A étant la parallèle menée par le centre du cercle à la direction des génératrices.

En coupant la surface cylindrique de révolution par deux plans de sections droites, on obtient le *cylindre droit à base circulaire* ou *cylindre de révolution* (*fig.* 339), que l'on peut encore considérer comme la figure engendrée par un rectangle AA'DD' (*fig.* 339) tournant autour d'un de ses côtés.

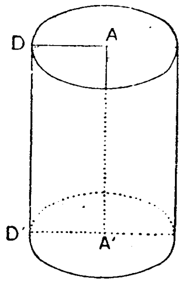

Fig. 339.

Un cylindre de révolution est manifestement défini quand on se donne le cercle de base et la hauteur (ainsi que le sens dans lequel elle doit être portée). Deux cylindres de révolution qui ont même rayon de base et même hauteur, sont égaux.

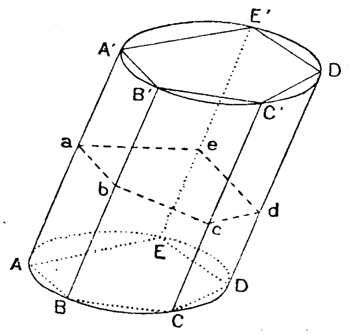

Fig. 340.

440. Surface latérale du cylindre.

Un cylindre étant donné (*fig.* 340), inscrivons dans la première courbe de base un polygone quelconque : ce polygone sera la base d'un prisme ABCDEA'B'C'D'E' (*fig.* 340) ayant pour arêtes des génératrices du cylindre et pour seconde base un polygone inscrit dans la seconde base de celui-ci : prisme qui sera dit *inscrit* au cylindre.

L'*aire latérale* du cylindre est, par définition, la limite vers laquelle tend l'aire latérale d'un prisme inscrit, lorsque le nombre des côtés du polygone de base augmente indéfiniment, de manière que chacun d'eux tende vers zéro.

Dans le cas du cylindre droit à base circulaire, nous allons prouver l'existence de cette limite et en trouver la valeur.

Théorème. — *L'aire latérale d'un cylindre droit à base circulaire est égale au produit du périmètre de la base par la hauteur.*

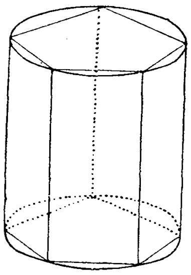

Fig. 341.

En effet, l'aire latérale du prisme inscrit (*fig.* 341) est égale (**388**), au périmètre de la section droite (qui se confond ici avec la base) multiplié par l'arête latérale.

Si l'on augmente indéfiniment le nombre des côtés du polygone de base, de manière que chacun d'eux tende vers zéro, le périmètre de ce polygone tend, ainsi que nous l'avons démontré (Pl., **176-177**), vers une limite, qui est la longueur de la circonférence de base du cylindre; l'arête du prisme étant d'ailleurs constamment égale à la hauteur du cylindre, le théorème est démontré (1).

Corollaire. — Soient R le rayon de base du cylindre, h sa hauteur. *L'aire latérale est* $2\pi Rh$, puisque la circonférence de base a pour longueur $2\pi R$.

Pour avoir l'*aire totale* du cylindre, il faudrait ajouter, bien entendu, les aires des deux cercles de bases. L'aire totale est donc $2\pi Rh + 2\pi R^2 = 2\pi R\,(h + R)$.

441. Volume du cylindre.

Le volume d'un cylindre est, par définition, la limite vers laquelle tend le volume d'un prisme inscrit, lorsque le nombre des côtés du polygone de base augmente indéfiniment, de manière que chacun d'eux tende vers zéro.

Dans le cas du cylindre à base circulaire, droit ou non, nous allons prouver l'existence de cette limite et en trouver la valeur.

---

(1) Le raisonnement s'applique à un cylindre quelconque, du moment que l'on a pu définir (Pl., **179**, note) la longueur de la courbe de section droite de ce cylindre : c'est-à-dire, du moment que le périmètre d'un polygone inscrit dans cette courbe et dont le nombre de côtés augmente indéfiniment tend vers une limite (voir fig. 340). La conclusion est évidemment la suivante : *L'aire latérale d'un cylindre quelconque est égale au périmètre de la section droite, multiplié par l'arête.*

Théorème. — *Le volume du cylindre à base circulaire est égal à la surface de base, multipliée par la hauteur.*

Le volume du prisme inscrit est, en effet, égal au produit de la surface de sa base par sa hauteur : cette dernière est commune au prisme et au cylindre, pendant que la surface de base du prisme a pour limite la surface de base du cylindre (1).

Corollaire. — Soient, comme précédemment, R le rayon du cylindre ; h sa hauteur. *Le volume est* $\pi R^2 h$.

Remarque. — Nous avons défini l'aire latérale et le volume d'un cylindre comme limites de l'aire latérale et du volume d'un prisme inscrit. Nous serions évidemment arrivés à la même limite en considérant l'aire latérale et le volume d'un prisme *circonscrit* (on désigne sous ce nom un prisme ayant pour bases des polygones circonscrits aux courbes de bases et dont les arêtes sont parallèles et égales aux génératrices du cylindre).

EXERCICES

629. Si une droite a plus de deux points communs avec une surface cylindrique à base circulaire, elle est une génératrice de la surface.

630. Trouver le lieu des milieux des cordes interceptées par une surface cylindrique à base circulaire sur des droites issues d'un point fixe.

631. Mener, par un point donné de l'espace, un plan tangent à un cylindre de révolution donné.

632. Lieu des centres des homothéties de rapport donné et telles que l'homologue d'une droite donnée rencontre un cercle donné.

633. Trouver le lieu des points de l'espace tels que leurs projections sur les côtés d'un triangle donné soient trois points en ligne droite.

634. Démontrer que les seules surfaces de révolution, qui soient en même temps des cylindres, sont les cylindres de révolution définis au n° **439**.

635. Calculer les dimensions du litre qui sert à mesurer les grains et du litre qui sert à mesurer les liquides, sachant que ces deux instruments ont tous deux la forme cylindrique, la hauteur étant, dans le premier, égale au diamètre de la base et, dans le second, double de ce diamètre.

636. Quel est le rapport des volumes engendrés par un rectangle tournant successivement autour de deux côtés consécutifs ?

637. Étant donnés un cercle et deux diamètres rectangulaires, trouver un rectangle ayant deux côtés suivant ces diamètres, un sommet sur la circonférence et et qui, en tournant autour d'un de ses côtés, engendre un cylindre de surface totale donnée. Maximum de cette surface totale.

---

(1) Le raisonnement s'applique évidemment à un cylindre de base quelconque, en admettant toutefois que la définition (Pl., **260**, note) de l'aire d'une courbe fermée s'applique à la courbe de base de ce cylindre. Moyennant cette condition, on a donc le théorème suivant : *Le volume d'un cylindre quelconque est égal au produit de sa base par sa hauteur.*

638. Étant donné un cylindre droit à base circulaire, évaluer la portion de l'aire latérale comprise entre la base, un plan P passant par un diamètre de la base et deux génératrices quelconques (voir la figure ci-contre). Montrer que l'aire ainsi définie est *quarrable*, c'est-à-dire qu'on peut construire un rectangle équivalent (on suppose qu'on a tracé le cercle de base, marqué sur sa circonférence les pieds des deux génératrices, donné la longueur interceptée par le plan P sur l'une d'elles et la direction du diamètre commun aux deux plans).

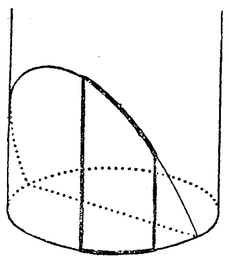

(On imitera la méthode du n° **440**, en remplaçant d'abord la surface cylindrique par une surface prismatique inscrite).

639. En coupant une surface cylindrique à base circulaire par deux plans non parallèles entre eux (mais ne se coupant pas à l'intérieur de la surface), on obtient un solide (qu'on peut appeler *tronc de cylindre*). Montrer que l'aire latérale et le volume de ce solide sont les mêmes que ceux d'un cylindre limité par la même surface cylindrique et deux plans parallèles à la base du cylindre menés aux points où la parallèle aux génératrices menée par le centre de cette base rencontre les plans donnés.

CHAPITRE II

CONE — TRONC DE CONE

442. Parmi les cônes, on distingue ceux qui ont pour bases des cercles.

En particulier, un cône dont la base est un cercle et dont le sommet est sur la perpendiculaire élevé au plan du cercle par son centre, a reçu le nom de *cône droit à base circulaire* ou *cône de révolution*.

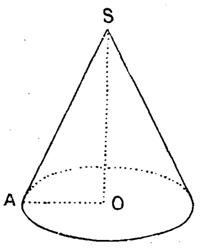

Fig. 342.

Ce cône (*fig.* 342), considéré comme surface conique illimitée, est, en effet la surface de révolution qui a pour méridienne une droite sécante à l'axe; considéré comme solide limité, c'est la figure engendrée par un triangle rectangle qui tourne autour d'un de ses côtés.

Un cône droit est évidemment déterminé lorsqu'on donne son cercle de base et sa hauteur, avec le sens dans lequel celle-ci doit être portée. Deux cônes droits de même rayon de base et de même hauteur soit égaux.

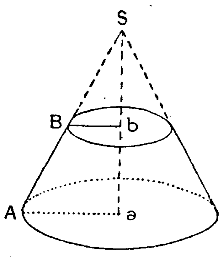

Fig. 343.

La longueur commune des génératrices d'un cône droit a reçu le nom d'*arête latérale* ou d'*apothème* du cône. On donne le nom d'*angle au sommet* du cône à l'angle que forment entre elles deux génératrices situées dans un même plan méridien, angle évidemment double de l'angle compris entre l'une d'elles et l'axe.

Il est clair que le *tronc de cône de révolution* peut s'obtenir en faisant tourner un trapèze rectangle AB*ab* (*fig.* 343) autour du côté perpendiculaire aux bases.

Remarque. — Si la génératrice est perpendiculaire à l'axe, la surface dégénère en un plan.

443. Aire latérale du cône de révolution.

Si, dans la base d'un cône quelconque, on inscrit un polygone quelconque, la pyramide qui a pour base ce polygone et dont le sommet est le même que celui du cône, est dite *inscrite* à celui-ci.

On nomme *aire latérale* du cône, la limite vers laquelle tend l'aire latérale d'une pyramide inscrite, lorsque le nombre des côtés du polygone de base augmente indéfiniment, de manière que chacun d'eux tende vers zéro.

Nous allons, dans le cas du cône de révolution, prouver l'existence de cette limite et en trouver la valeur.

Théorème. — *L'aire latérale du cône droit est égale au demi-produit de la circonférence de base par l'arête latérale.*

Dans le cône donné, de sommet S (*fig.* 344), inscrivons la pyramide SABCDE. L'aire latérale de cette pyramide est la somme des triangles SAB, SBC, SCD, ..., autrement dit, la somme des demi-produits obtenus en multipliant chacune des bases AB, BC, CD .... de ces triangles par la hauteur correspondante. Elle est donc égale (¹) à la demi-somme $\frac{AB + BC + \dots}{2}$, multipliée par une quantité a intermédiaire entre la plus grande et la plus petite des hauteurs.

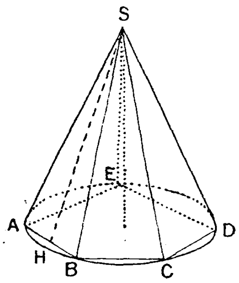

Fig. 344.

Lorsque le nombre des côtés du polygone de base augmente indéfiniment de manière que chacun d'eux tende vers zéro, la somme AB + BC + ... tend vers la longueur de la circonférence de base. Quant aux hauteurs, si nous considérons, par exemple, la hauteur SH (*fig.* 344) du triangle SAB, laquelle tombe au milieu H de AB, nous voyons, dans le triangle SAH, qu'elle est comprise entre SA et SA — AH, lequel est égal à $SA - \frac{AB}{2}$. Elle tend donc vers SA lorsque AB tend vers zéro et, lorsque le plus grand des côtés AB, BC, ... tend vers zéro, la plus grande et la plus petite des hauteurs considérées précédemment tendent vers l'arête latérale du cône. Il en est donc de même de a et le théorème est démontré (²).

---

(1) Tannery, *Leçons d'Arithmétique*, n° **212**, page 180.

(2) Contrairement à ce qui a été vu pour le cylindre, et à ce qu'on va voir plus loin pour le volume du cône, la démonstration relative à l'aire du cône de révolution n'est point valable

Corollaire. — Appelons maintenant a l'arête latérale du cône, et soit R le rayon de base. *L'aire latérale sera* $\frac{1}{2}\,2\pi R \times a = \pi R a$.

L'aire totale du cône s'obtiendra en ajoutant, à l'aire latérale, celle de la base, ce qui donnera $\pi R a + \pi R^2 = \pi R\,(a + R)$.

Remarque. — Si, au lieu de pyramides inscrites au cône, nous avions considéré des pyramides *circonscrites*, c'est-à-dire ayant pour bases des polygones circonscrits à la base du cône (le sommet restant le même), nous aurions constaté que les surfaces latérales de ces pyramides tendent vers la même limite que les surfaces latérales des pyramides inscrites. Le raisonnement précédent peut en effet se répéter dans ces nouvelles conditions, avec cette seule modification que, dans une pyramide circonscrite SA′B′C′D′,... (*fig.* 345), les hauteurs des triangles SA′B′, SB′C′,... sont constamment égales à l'apothème du cône, car elles tombent (**364**) aux points de contact des côtés A′B′, B′C′,... avec le cercle de base.

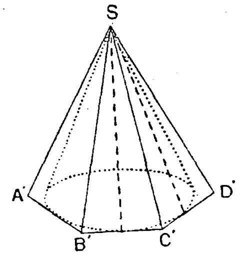

Fig. 345.

444. Volume du cône.

On nomme *volume* d'un cône quelconque, la limite vers laquelle tend le volume d'une pyramide inscrite, lorsque le nombre des côtés du polygone de base augmente indéfiniment de manière que chacun d'eux tende vers zéro.

Dans le cas du cône à base circulaire, nous allons prouver l'existence de cette limite et en trouver la valeur.

Théorème. — *Le volume d'un cône à base circulaire est égal au tiers du produit de la surface de base par la hauteur.*

Ce théorème se déduit immédiatement du théorème relatif au volume de la pyramide, en remarquant que l'aire du polygone qui sert de base à la pyramide tend, par définition, vers l'aire du cercle qui sert de base au cône, lorsqu'on augmente le nombre des côtés dans les conditions indiquées (¹).

Corollaire. — Soient R le rayon de base; h la hauteur. *Le volume est* $\frac{1}{3}\pi R^2 h$.

en dehors de ce cas spécial. Pour les autres espèces de cône (et en particulier pour le cône *oblique* à base circulaire), on peut démontrer l'existence de la limite, mais non en donner d'évaluation élémentaire, du moins en général.

(1) Le raisonnement et sa conclusion s'étendent évidemment à un cône quelconque, pourvu que l'on puisse définir l'aire de la base.

Remarque. — Il est clair qu'on serait arrivé au même résultat en substituant, aux pyramides inscrites, des pyramides circonscrites.

445. Aire latérale du tronc de cône de révolution.

Si, dans le cône dont fait partie un tronc de cône donné on inscrit une pyramide, le plan de base supérieure du tronc détache de cette pyramide un tronc de pyramide, qui est dit *inscrit* au tronc de cône.

On nomme *aire latérale* du tronc de cône, la limite vers laquelle tend l'aire latérale d'un tronc de pyramide inscrit, lorsque le nombre des côtés des polygones augmente indéfiniment, de manière que chacun de ces côtés tende vers zéro.

Cette définition est manifestement équivalente à la suivante : L'aire latérale d'un tronc de cône est la différence des aires des deux cônes qui font partie de la même surface conique que le tronc et qui ont bases respectives les deux bases de ce tronc.

Théorème. — *L'aire latérale d'un tronc de cône de révolution est égale à la demi-somme des deux circonférences de bases, multipliée par l'arête latérale.*

Soit le tronc de cône $ABA'B'$ (*fig.* 346), dont l'arête latérale est AA' et qui fait partie d'un cône SAB de sommet S. L'aire latérale de ce cône est la différence des aires latérales du cône SAB et du cône $SA'B'$ qui a pour base la petite base du tronc.

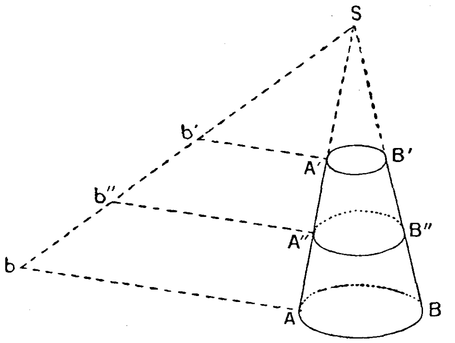

Fig. 346.

Par le point A, menons à SA une perpendiculaire Ab égale à la longueur de la circonférence de base AB, et joignons Sb. Si, par le point A', nous menons une parallèle $A'b'$ à Ab, jusqu'à rencontre en b' avec Sb, cette droite $A'b'$ sera égale à la longueur de la seconde circonférence de base $A'B'$. Car le rapport des deux circonférences de bases $A'B'$, AB, c'est-à-dire le rapport de leurs rayons, est égal à $\frac{SA'}{SA}$ ou à $\frac{A'b'}{Ab}$, et l'on a, d'autre part, circ. $AB = Ab$.

Dès lors, l'aire latérale du cône SAB est égale (**443**) à celle du triangle SAb et l'aire latérale du cône $SA'B'$, à celle du triangle $SA'b$. L'aire du tronc de cône donné équivaut donc à celle trapèze rectangle $AA'bb'$, laquelle a bien l'expression indiquée par l'énoncé.

Corollaires. — I. *Si* R, R′ *sont les rayons des bases; a, l'arête, l'aire latérale est* $\pi(R + R')a$.

II. *L'aire latérale d'un tronc de cône est égale au produit de l'arête par la circonférence obtenue en coupant le solide par un plan parallèle aux bases et équidistant de ces bases.*

Car un tel plan passe par le milieu A'' (*fig.* 346) de AA' : la longueur de la circonférence de section est (d'après un raisonnement tout semblable à celui qui a été fait tout à l'heure) égale à la droite $A''b''$ parallèle à Ab et limitée à Sb. Le produit de cette circonférence par l'arête AA' équivaut donc (Pl., **252** *bis*) à l'aire du trapèze $AA'bb'$.

C. Q. F. D.

Remarque. — On peut considérer le cylindre et le cône comme des cas limites du tronc de cône : le premier correspondant à l'hypothèse où le trapèze qui engendre le tronc de cône devient un rectangle (ses bases étant égales) ; le second, à l'hypothèse où ce trapèze se réduit à un triangle (une des bases étant nulle).

Les corollaires précédents continuent à s'appliquer et donnent bien : pour $R' = R$, l'aire du cylindre (**440**) ; pour $R' = 0$, l'aire du cône (**443**).

446. Volume du tronc de cône.

Le *volume du tronc de cône* est la limite vers laquelle tend le volume d'un tronc de pyramide inscrit, lorsque les côtés des polygones de bases tendent tous vers zéro. Il est évidemment égal à la différence des volumes de deux cônes, les mêmes dont il a été question dans le numéro précédent, à propos de l'aire latérale.

Théorème. — *Un tronc de cône à base circulaire est équivalent à la somme de trois cônes ayant pour hauteur commune la hauteur du tronc et, pour bases respectives, les deux bases de ce tronc et une moyenne proportionnelle entre ces deux bases.*

Car le tronc de pyramide inscrit est, d'après le théorème du n° **408**, équivalent à la somme de trois pyramides, qui tendent respectivement vers les trois cônes dont parle l'énoncé.

Corollaire. — *Si* R *et* R′ *sont les rayons des bases, h la hauteur, le volume est* $\frac{1}{3}\pi h\,(R^2 + R'^2 + RR')$.

Car les bases du tronc sont respectivement mesurées par πR^2 et $\pi R'^2$, dont la moyenne proportionnelle est $\sqrt{\pi R^2 \times \pi R'^2} = \pi RR'$.

EXERCICES

640. Si une droite a plus de deux points communs avec une surface conique à base circulaire, elle est une génératrice de la surface.

641. Montrer que les seules surfaces de révolution qui soient en même temps des cônes sont les cônes de révolution définis au n° **442**.

642. Mener, par un point donné, un plan tangent à un cône à base circulaire donné.

643. Si un cône de révolution est tangent aux deux faces d'un dièdre, les génératrices de contact font des angles égaux avec l'arête. Le plan mené par l'axe du cône et l'arête du dièdre fait des angles égaux avec les faces, et fait aussi des angles égaux avec les plans méridiens des génératrices de contact.

644. Quel est le lieu des axes des cônes de révolution tangents à deux plans donnés ?

645. Quel est le lieu des axes des cônes de révolution qui passent par deux droites concourantes données?

646. Lieu des points tels que le rapport de leurs distances à un point et à un plan passant par ce point soit constant.

647. Démontrer qu'un cône à base circulaire, mais quelconque d'ailleurs (c'est-à-dire, en général, non de révolution), admet toujours un plan de symétrie et que la surface conique dont il fait partie admet un axe de symétrie.

648. Faire passer un cône de révolution par trois droites données concourantes, mais non situées dans un même plan. Quel est le nombre des solutions?

Montrer que deux quelconques de cônes de révolution qui passent par les trois droites données ont une quatrième droite commune (laquelle peut être confondue avec l'une des trois premières). Construire la position de cette quatrième droite.

649. La condition pour qu'un angle polyèdre convexe à quatre faces soit inscriptible à un cône de révolution est que la somme de deux dièdres opposés soit égale à la somme des deux autres, ou sinon, qu'il en soit lorsqu'on remplace une ou plusieurs arêtes par leurs prolongements.

650. Quelle est la condition pour qu'un angle polyèdre à quatre faces soit inscriptible à deux cônes de révolution? — pour qu'il soit inscriptible à trois cônes de révolution? Dans ce dernier cas, les axes des trois cônes forment un trièdre trirectangle.

651. Trouver un cône (ou un cylindre) de révolution tangent à trois plans donnés. Quel est le nombre des solutions?

Deux cônes quelconques tangents aux trois plans donnés ont un quatrième plan tangent commun. Construire ce plan.

652. Quelles conditions doivent remplir les faces d'un angle polyèdre à quatre faces pour que cet angle soit circonscriptible à un cône de révolution? Dans quels cas l'angle polyèdre est-il conscriptible à plus d'un cône de révolution?

653. Parmi les génératrices d'un cône de révolution, quelle est celle qui fait le plus petit ou le plus grand angle avec une droite donnée?

654. L'angle au sommet d'un cône de révolution est plus grand que l'angle de deux génératrices non situées dans le même plan méridien.

655. On coupe un cône à base circulaire par un plan parallèle à la base, mais situé au delà du sommet et formant, par conséquent, avec la seconde nappe de la surface conique, un second cône homothétique, du premier. Évaluer le volume du solide (*tronc de cône de seconde espèce*) formé par l'ensemble des deux cônes, connaissant les rayons des deux cercles et la distance de leurs plans.

656. Inscrire, dans un cône de révolution donné, un cylindre de surface latérale donnée. Maximum de celle-ci.

657. On donne deux cônes de révolution égaux SAB, S'A'B', placés de façon que les plans des cercles de bases AB, A'B' sont parallèles et que le sommet de chacun des cônes est dans le plan du cercle de base de l'autre. On coupe ces deux cônes par un plan P parallèle aux plans des deux bases et situé entre ces plans; ce plan P coupe le premier cône suivant un cercle CD et le second cône suivant un cercle C'D'. On désigne par r, l, h, le rayon de la base, l'arête, la hauteur de chacun des cônes, par x la distance du sommet S au point de rencontre du plan P et de l'arête SA et par y la distance du sommet S au plan P.

1° Déterminer x de façon que le rapport de la somme des surfaces latérales des deux troncs de cône ABCD, A'B'C'D' et de la surface latérale du cône SAB soit égal à un nombre donné λ. Discuter.

2° Déterminer y de façon que le rapport de la somme des volumes des troncs de cônes ABCD, A'B'C'D' au volume du cône SAB soit égal à un nombre donné μ. Discuter.

CHAPITRE III

GÉNÉRALITÉS SUR LA SPHÈRE

447. Définition. — On nomme *sphère* (*fig.* 347), le lieu géométrique des points de l'espace situés à une distance donnée d'un point donné, appelé *centre* de la sphère.

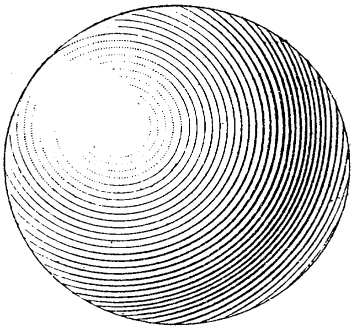

FIG. 347.

La droite qui joint le centre à un point de la surface est dite un *rayon*. Une droite qui joint deux points de la surface et qui, de plus, passe par le centre, est dite un *diamètre*. Il résulte de la définition que tous les rayons sont égaux entre eux et qu'un diamètre est double d'un rayon.

Un plan qui passe par le centre de la sphère est dit *plan diamétral*. La sphère sépare l'une de l'autre deux régions de l'espace : l'une, comprenant les points dont la distance au centre est moindre que le rayon, est dite *intérieure* à la sphère (¹) ; au contraire, les points dont la distance au centre est plus grande que le rayon sont dits *extérieurs* à la surface. On ne peut passer de l'une des deux régions à l'autre par un chemin continu sans traverser la surface.

La définition de la sphère, jointe à celle de la circonférence (Pl., **56**) montre évidemment que la section d'une sphère par un plan diamétral est une circonférence de même centre et de même rayon que la sphère : une telle circonférence est dite *grand cercle* de la sphère.

Théorème. — *La surface engendrée par la révolution d'une demi-circonférence autour de son diamètre est une sphère.*

---

(1) On donne également, dans le langage courant, le nom de *sphère* à la région située à l'intérieur de la surface : cette manière de parler n'introduit aucune confusion dans les raisonnements.

Réciproquement, *toute sphère peut être considérée comme engendrée de cette façon, l'axe étant un diamètre quelconque de la surface.*

1° Soit une demi-circonférence du centre O (*fig.* 348) tournant autour de son diamètre AB. La distance OM d'un point quelconque de la demi-circonférence au. centre O ne changera pas dans ce déplacement, de sorte que le point M sera constamment sur une sphère S ayant même centre et même rayon que la circonférence donnée.

Inversement, tout point M′ de la sphère ainsi obtenue est tel que la distance OM′ soit égale à OA. Il appartient donc à une circonférence, section de la sphère par le plan AM′B, et qui est une des positions que vient prendre la circonférence donnée dans son mouvement de révolution autour de l'axe;

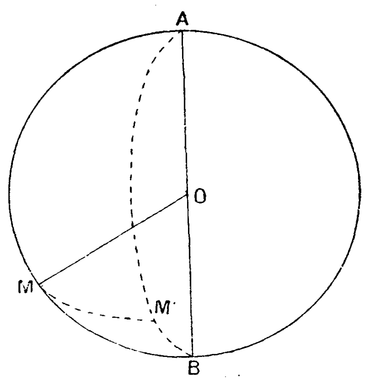

FIG. 348.

2° Soit AB un diamètre quelconque d'une sphère donnée O. Par AB, faisons passer un plan qui coupe la sphère suivant une circonférence de centre O. La sphère engendrée par la révolution d'une des moitiés de cette circonférence autour de l'axe AB coïncide manifestement avec la sphère donnée.

REMARQUE. — On voit que la *sphère est de révolution autour d'un quelconque de ses diamètres.*

Par conséquent (**438**), *la sphère admet tous ses plans diamétraux comme plans de symétrie.*

448. Intersection d'une droite et d'une sphère.

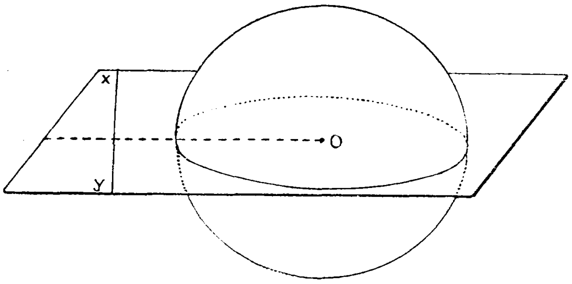

FIG. 349.

Pour étudier l'intersection d'une droite xy et d'une sphère S (*fig.* 349), faisons passer un plan par la droite xy et le centre O de la sphère. Tout point commun à la droite et à la sphère appartenant forcément au grand cercle, section de la sphère par ce plan, les conclusions du n° **58** (Pl., liv. II), nous permettent d'énoncer les propositions suivantes :

1° *Une droite ne rencontre pas une sphère, si la distance de cette droite au centre est plus grande que le rayon ;*

2° *Si cette distance est plus petite que le rayon, la droite et la sphère se coupent en deux points. La corde interceptée sur la droite est d'autant plus petite que la distance de la droite au centre est plus grande; le centre se projette au milieu de cette corde ;*

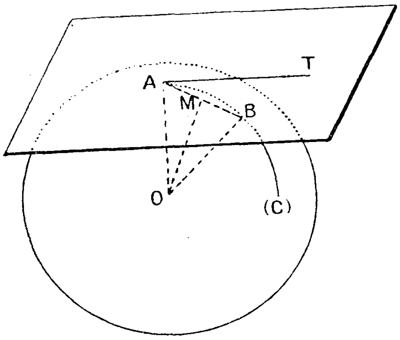

Fig. 350.

3° *Enfin, dans le cas intermédiaire où la distance de la droite au centre est égale au rayon, la droite n'a, avec la sphère, qu'un point commun.*

On dit alors qu'elle est *tangente* à la sphère (AT, *fig.* 350).

On voit qu'*une tangente à la sphère est perpendiculaire au rayon du point de contact*, et qu'inversement, *toute droite perpendiculaire à l'extrémité d'un rayon est tangente.*

449. La définition que nous venons de donner de la tangente à la sphère n'est pas celle que nous avons indiquée, d'une façon générale, pour la tangente à une surface quelconque (**433**). Nous allons faire voir que ces deux définitions sont entièrement équivalentes l'une à l'autre.

Par un point A de la sphère (*fig.* 350), soit tracée une courbe quelconque (C), que nous supposerons avoir une tangente AT. Nous avons à démontrer que AT est tangente à la sphère, au sens du numéro précédent.

Or la tangente AT est la limite d'une sécante AB, lorsque le point B tend vers le point A. Le point M, projection du centre AB, est le milieu de AB : il tend donc, dans les conditions dont nous venons de parler, vers le point A ; d'où résulte nécessairement la conséquence demandée.

Inversement, toute droite AT perpendiculaire à l'extrémité d'un rayon de la sphère est tangente au moins à une courbe tracée sur la surface, à savoir le grand cercle situé dans le plan OAT.

449 *bis*. De ce qui précède, résulte le théorème suivant :

Théorème. — *Le lieu des tangentes à la sphère en un de ses points est le plan perpendiculaire au rayon qui aboutit en ce point* (*fig.* 350).

Ce plan est dit *plan tangent à la sphère*, au point considéré.

Nous avons donc démontré, pour la sphère, un théorème analogue à celui qui avait été précédemment établi pour les cylindres (**433**) et pour les cônes (**436**).

Toutefois, nous avons à noter une différence importante : un plan tangent à la sphère n'a manifestement qu'*un seul* point de contact. Le cylindre et le cône présentaient, au contraire, cette particularité que le plan tangent était le même en *une infinité* de points (à savoir, tous les points d'une même génératrice).

450. Intersection d'une sphère et d'un plan.

Théorème. — *Si la distance d'un plan au centre d'une sphère est supérieure au rayon de celle-ci, les deux surfaces n'ont aucun point commun.*

Si la distance du plan au centre de la sphère est moindre que le rayon, les deux surfaces se coupent suivant un cercle, dont le centre est la projection du centre de la sphère sur le plan.

Enfin, dans le cas intermédiaire où il y a égalité, le plan n'a qu'un point commun avec la sphère; il est (**449** *bis*) *tangent à cette surface.*

Si, en effet, par la projection C du centre O de la sphère sur le plan considéré P, on mène, dans celui-ci, une série de droites (*fig.* 351, 352), la distance du centre à l'une quelconque d'entre elles est la même que celle du même point au plan.

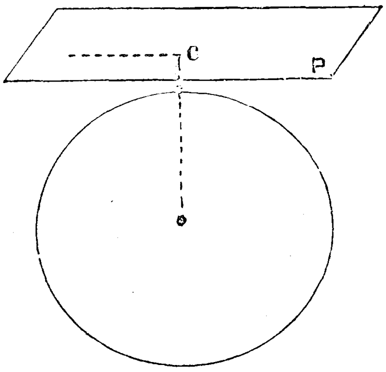

FIG. 351.

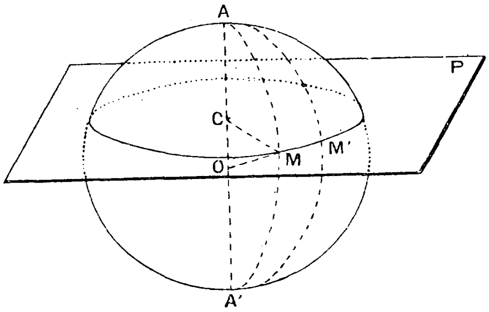

FIG. 352.

Si donc cette distance est plus grande que le rayon, aucune des droites ainsi menées ne coupe la sphère (**448**).

Si cette distance est égale au rayon, les droites en question sont tangentes.

Si elle est plus petite que le rayon, toutes ces droites coupent la sphère. Le lieu des points communs est, d'après la définition de la sphère et le théorème du n° **344**, une circonférence de centre C.

Ce dernier point résulte d'ailleurs de ce que le diamètre OC peut être considéré (**447**, Rem.), comme un axe de révolution de la sphère. Si, en effet, M est un point commun à celle-ci et au plan, ce point, en tournant autour de OC, engendre un cercle de centre C qui appartient à la fois à la sphère (engendrée par la révolution du demi-grand cercle AMA′ situé dans le plan OCM) et au plan P (engendré par la demi-droite CM ; **442**, Rem.).

Inversement, si M′ est un autre point quelconque commun à la sphère et au plan, le demi-grand cercle AMA′ peut être amené, par une rotation autour de OC, à coïncider avec le demi-grand cercle AM′A′, situé dans le plan OCM′. Le point M qui, dans ce mouvement, n'a pas cessé d'appartenir au plan P, se trouve alors amené à l'intersection de ce plan et du demi-grand cercle AM′A′, c'est-à-dire au point M′. Ce dernier est donc bien une des positions que vient prendre le point M en tournant autour de OC.

Corollaires. — I. *Une sphère de rayon* R *et un plan situé à une distance d du centre se coupent suivant un cercle dont le rayon r est donné par la formule*

$$r = \sqrt{R^2 - d^2}.$$

C'est ce qui se voit dans le triangle rectangle OCM (*fig.* 352), lequel a pour hypoténuse R et pour côtés de l'angle droit r, d.

La formule précédente justifie la dénomination de *grand cercle*, donnée plus haut à la section de la sphère par un plan passant par le centre, section dont le rayon est R. On voit en effet que toutes les autres sections de la sphère (lesquelles sont appelées *petits cercles*) ont un rayon inférieur à R.

II. *Toute sphère qui contient trois points d'un cercle, contient ce cercle tout entier.* Car le plan de ce cercle coupe la sphère suivant un cercle, forcément identique au premier (Pl., **57**).

III. *Le lieu des centres des sphères qui passent par un cercle donné est la droite* (dite *axe* du cercle) *perpendiculaire au plan du cercle en son centre.*

En effet, il résulte de ce que nous venons de voir que les centres des sphères qui contiennent le cercle donné sont tous sur la droite en question, et que, réciproquement, la sphère qui a pour centre un point quelconque de cette droite et qui passe par un point du cercle donné, contient ce cercle tout entier.

451. Théorème. — *Deux cercles qui ont deux points communs sans être situés dans un même plan, déterminent une sphère.*

Soient les deux cercles C, C' (*fig.* 353) dont les plans sont P, P' et qui ont les deux points communs A, B.

Toute sphère passant par le cercle C aura son centre sur la perpendiculaire Cx, élevée par le centre de ce cercle au plan P, et réciproquement, toute sphère ayant son centre sur cette droite et passant par A contiendra tout le cercle C. En particulier, il résulte tout d'abord de là et du n° **340** que la droite Cx est dans le plan perpendiculaire au milieu de AB.

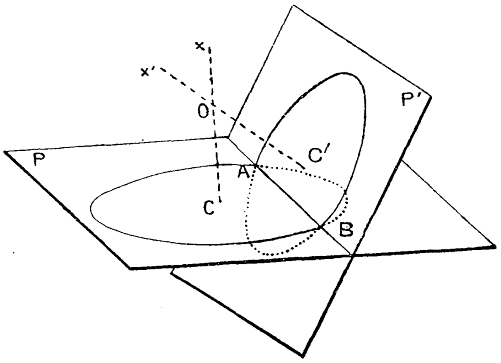

FIG. 353.

De même, toute sphère passant par le cercle C aura son centre sur la perpendiculaire $C'x'$, élevée par le centre de ce cercle au plan P' ; et, réciproquement, toute sphère ayant son centre sur cette droite et passant par le point A, contiendra tout le cercle C'. En particulier, cette droite est dans le plan perpendiculaire au milieu de AB.

Les deux droites Cx, $C'x'$ sont dans un même plan ; elles ne sont ni parallèles ni confondues (sans quoi les plans P, P', qui leur sont respectivement perpendiculaires, seraient eux-mêmes parallèles ou confondus, ce qui n'est pas). Elles se coupent donc en un point unique O.

La sphère de centre O et de rayon OA répond, et répond seule, à la question.

REMARQUES. — I. On verrait de même que *deux cercles, qui sont tangents entre eux sans être dans un même plan* (*fig.* 354), *déterminent une sphère.*

Il faudrait seulement, dans la démonstration précédente, remplacer le plan perpendiculaire au milieu de AB par le plan mené au point de contact, perpendiculairement à la tangente commune.

II. Si les deux cercles donnés étaient dans un même plan, ce serait ce plan qui remplacerait la sphère. Le plan est un cas-limite de la sphère, comme la droite est un cas-limite du cercle (Pl., **90**, Rem.).

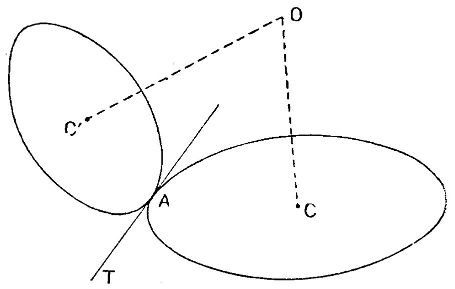

Fig. 354.

Théorème. — *Un cercle et un point extérieur à son plan déterminent une sphère.*

Quatre points, non situés dans un même plan, déterminent une sphère.

1° La condition, imposée à la sphère cherchée de passer par un cercle et un point extérieur à son plan, peut être remplacée (**450**, Coroll. II) par celle de contenir deux cercles qui ont deux points communs : à savoir, le cercle donné et celui qu'on peut mener par deux points de ce cercle et le point donné ;

2° De même, la condition, imposée à une sphère, de passer par quatre points donnés A,B,C,D, peut être remplacée par celle de contenir deux cercles qui se coupent en deux points : le cercle ABC et le cercle ABD.

Corollaire. — *Une sphère ne peut avoir deux centres, ni, par suite, deux rayons différents.*

452. Cône et cylindre circonscrits à la sphère.

Théorème. — *Le lieu des tangentes que l'on peut mener à une sphère par un point extérieur est un cône de révolution.*

Ces tangentes sont toutes égales entre elles.

Le lieu de leurs points de contact est un petit cercle de la sphère.

Les plans tangents au cône, lieu des tangentes, sont tangents à la sphère. Ce sont les seuls plans jouissant de cette propriété qu'on puisse mener par le point donné.

Soient O le centre de la sphère donnée (*fig.* 355), S le point donné. Soit T le point de contact d'une tangente menée par S à l'un quelconque des grands cercles dont les plans passent par OS. Si nous faisons tourner la droite ST autour de l'axe OS, elle ne cessera pas d'être tangente à la sphère, puisque celle-ci est de révolution autour de OS.

Les tangentes ainsi obtenues (à savoir les positions successives de ST) et qui satisfont évidemment aux conditions de l'énoncé, sont d'ailleurs les seules que l'on puisse mener du point S à la sphère : car une telle tangente est, d'après la démonstration du n° **449**, tangente au grand cercle situé dans le plan qui passe par son point de contact et les points O, S.

En chacune des positions du point T, le cône et la sphère ont même plan tangent, à savoir celui qui est déterminé par la droite ST et la tangente T*t* au cercle lieu du point T.

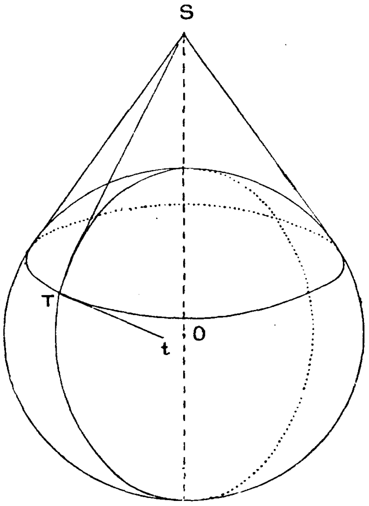

FIG. 355.

Nous obtenons ainsi une série de plans tangents menés du point S à la sphère : ce sont d'ailleurs les seuls, car, si le plan tangent en un point T passe par le point S, la droite ST est tangente.

REMARQUE. — On dit que le cône dont nous venons de parler est *circonscrit* à la sphère, et celle-ci *inscrite* au cône.

Inversement, *le long d'un petit cercle quelconque, on peut inscrire un cône à la sphère :* le sommet de ce cône sera le point où le plan tangent en un point quelconque du cercle coupe le diamètre perpendiculaire au plan de ce cercle.

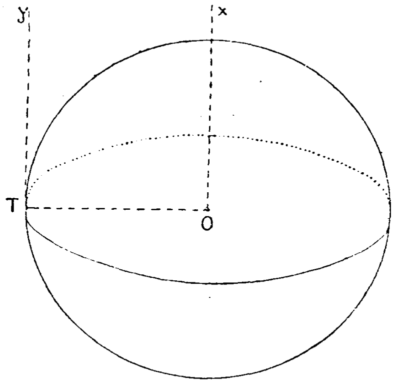

FIG. 356.

Théorème. — *Le lieu des tangentes menées à une sphère, parallèlement à une droite donnée, est un cylindre de révolution. Les plans tangents à ce cylindre sont tangents à la sphère. Ce sont les seuls plans qui jouissent de cette propriété, tout en étant parallèles à la droite donnée. Le lieu de leurs points de contact est le grand cercle dont le plan est perpendiculaire à la droite donnée.*

Soit Ox (*fig.* 356) la parallèle à la droite donnée, menée par le centre de la sphère. Soit Ty une tangente menée, parallèlement à cette droite, à un grand cercle dont le plan passe par Ox. On pourra, comme tout à l'heure, faire tourner cette droite Ty autour de Ox, sans qu'elle cesse d'être tangente à la sphère. Elle décrira, dans ces conditions, un cylindre de révolution et le point T décrira un grand cercle, car le rayon OT, perpendiculaire à la tangente, décrira le plan mené par O perpendiculairement à la direction donnée.

On verrait, comme plus haut, que les seules tangentes à la sphère parallèles à Ox, sont les génératrices du cylindre dont nous venons de parler et que ce cylindre est *circonscrit* à la sphère, c'est-à-dire tangent (¹) à cette surface en tous les points du cercle que décrit le point T.

Inversement, *le long d'un grand cercle quelconque, on peut circonscrire à la sphère un cylindre*, celui qui a pour section droite le grand cercle.

452 *bis*. *Par une droite entièrement extérieure à une sphère, on peut mener deux plans tangents à cette sphère.*

Soient D la droite donnée (*fig.* 357) ; C, le cercle de contact du cône circonscrit à la sphère, ayant pour sommet un point P de cette droite. Le plan du cercle C coupe D en un point I, extérieur à C (puisqu'il est extérieur à la sphère). Du point I, on peut mener deux tangentes IT, IT′ au cercle C. Le plan tangent à la sphère au point T n'est autre (n° précédent) que le plan TIP.

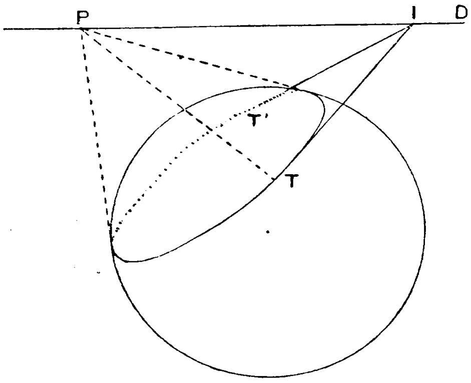

Fig. 357.

Inversement, tout plan tangent à la sphère passant par P doit (n° précédent) être tangent au cône dont il vient d'être question. S'il passe également par I, il coïncide avec l'un des deux que nous venons d'obtenir.

On peut mener à une sphère deux plans tangents parallèles à un plan donné. Leurs points de contact sont les extrémités du diamètre perpendiculaire au plan donné.

453. Intersection de deux sphères.

Théorème. — *Si deux sphères distinctes ont un point commun en dehors de la ligne des centres, leur intersection est un cercle, dont*

---

(1) Par analogie avec la définition donnée en géométrie plane, on dit que deux surfaces sont *tangentes* en un de leurs points communs, lorsqu'elles ont même plan tangent en ce point.

le plan est perpendiculaire à cette ligne et le centre, situé sur cette ligne.

Soient, en effet, O, O′ deux sphères qui ont en commun un point A, situé en dehors de la ligne OO′. Le point A, en tournant autour de OO′, engendrera un cercle qui appartiendra à la fois aux deux surfaces. Celles-ci n'ont d'ailleurs aucun point commun en dehors de ce cercle, en vertu du n° **451**.

453 *bis*. La situation respective de deux sphères dépend de l'ordre de grandeur dans lequel sont rangées la distance OO′ des centres, la somme R + R′ des rayons et la différence R — R′ de ces mêmes rayons. Cette dépendance est indiquée par le théorème suivant :

Théorème. — *Deux sphères sont :*

1° *Extérieures* (*fig.* 358), *si la distance des centres est supérieure à la somme des rayons ;*

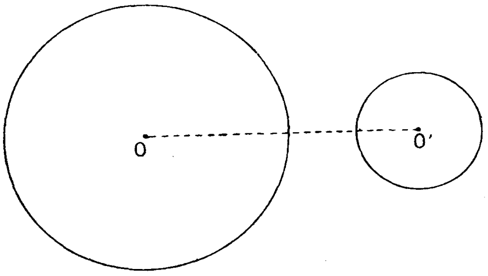

FIG. 358.

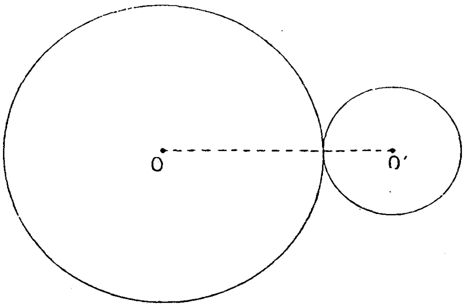

FIG. 359.

2° *Tangentes extérieurement* (*fig.* 359), *si la distance des centres est égale à la somme des rayons ;*

3° *Sécantes suivant un cercle* (*fig.* 360), *si la distance des centres est comprise entre la somme des rayons et leur différence ;*

4° *Tangentes intérieurement* (*fig.* 361), *si la distance des centres est égale à la différence des rayons.*

5° *Intérieures* (*fig.* 362), *si la distance des centres est plus petite que la différence des rayons.*

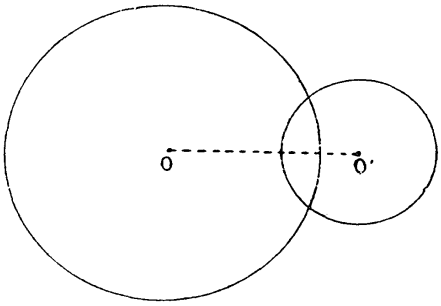

FIG. 360.

La démonstration est identiquement la même qu'en géométrie plane (Pl., **66**) sauf en 3°, où il suffit de remarquer qu'en coupant la figure par un plan quelconque passant par la ligne des centres, on obtient, dans les deux sphères, deux grands cercles sécants entre eux, de sorte que ces surfaces ont un point commun en dehors de la ligne des centres : après quoi, on est ramené au numéro précédent.

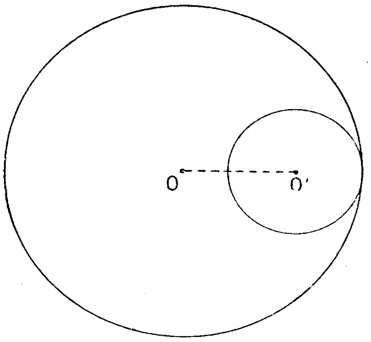

Fig. 361.

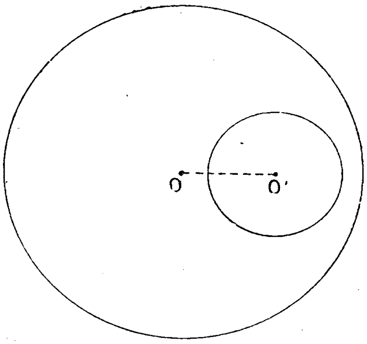

Fig. 362.

454. Intersection de trois sphères.

Si trois sphères ont un point commun en dehors du plan qui contient les trois centres, elles en ont un second, symétrique du premier par rapport à ce plan, puisque celui-ci est un plan de symétrie commun des trois surfaces.

D'après cela trois sphères peuvent :

Soit n'avoir aucun point commun ;

Soit avoir un seul point commun situé dans le plan des centres (les trois plans tangents se coupant suivant une ligne droite, perpendiculaire au plan des centres) ;

Soit avoir deux points communs ;

Soit avoir en commun trois points *et, par suite, un cercle entier.*

455. Puissance d'un point par rapport à une sphère.

Théorème. — *Si, par un point donné de l'espace, on mène différentes sécantes à une sphère, le produit des segments issus du point donné et aboutissant respectivement aux deux points d'intersection de chacune d'elles avec la surface est le même pour toutes les sécantes.*

Soient A le point donné (*fig.* 363) ; ABB′, ACC′, deux quelconques des sécantes. Le plan de ces deux droites coupe la sphère suivant un cercle, lequel donne bien

$$AB \cdot AB' = AC \cdot AC'.$$

Le produit dont il est question dans l'énoncé ne dépend donc que de la sphère et de la position du point donné A.

Ce produit, précédé du signe + si le point A est extérieur, du signe — si ce point est intérieur, est dit la *puissance* du point par rapport à la sphère.

La démonstration précédente montre que *si, par un point donné, on mène différents plans coupant une sphère donnée suivant des cercles,*

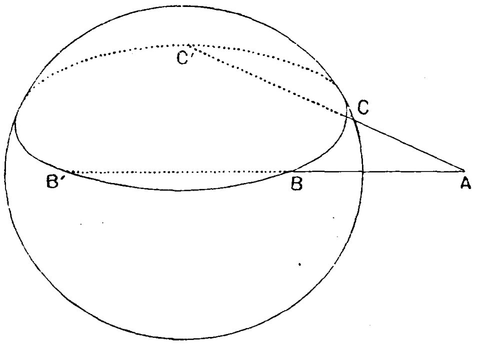

Fig. 363.

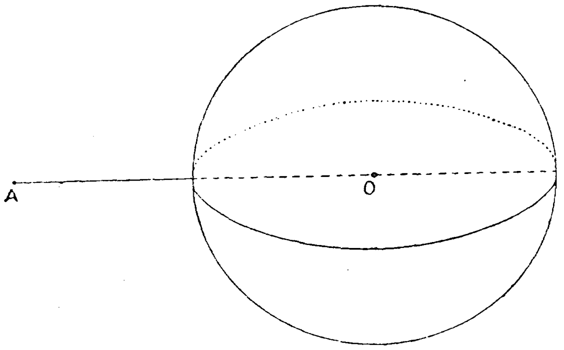

Fig. 364.

le point donné a, par rapport à tous ces cercles, la même puissance, à savoir la puissance de ce point par rapport à la sphère.

Considérons, en particulier, un plan passant par le point donné et coupant la sphère suivant un grand cercle : nous voyons immédiatement (*fig.* 364) que, comme pour le cercle en géométrie plane, *la puissance d'un point par rapport à une sphère est représentée par l'expression* $d^2 - R^2$, *où* R *est le rayon de la sphère, d la distance du point au centre* (1).

Dans le cas du point extérieur, la puissance est égale au carré de la tangente.

456. On appelle *angle de deux surfaces*, en un de leurs points communs, l'angle dièdre formé par les deux plans tangents.

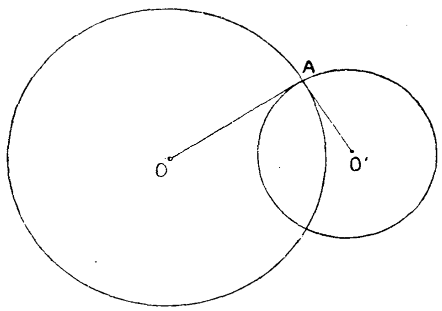

Fig. 365.

D'après cela, l'angle de deux sphères est (**378**) égal à l'angle des rayons qui aboutissent au point commun, ou à son supplément. Si les deux sphères se coupent orthogonalement, ces deux rayons sont perpendiculaires entre eux.

---

(1) Comparer Pl., **134**.

Lorsque deux sphères sont orthogonales, le carré du rayon de chacune d'elles est égal à la puissance de son centre par rapport à l'autre (¹). Car le rayon de la première sphère qui aboutit en un point commun est tangent à la seconde sphère (*fig.* 365). Réciproquement, *si le carré du rayon d'une sphère est égal à la puissance de son centre par rapport à une sphère, ces deux sphères se coupent à angle droit.*

457. Plan radical de deux sphères.

Théorème. — *Le lieu des points qui ont même point par rapport à deux sphères, est un plan perpendiculaire à la ligne des centres.*

Ce plan est dit *plan radical* des deux sphères considérées.

Démonstration. — Un plan quelconque P, mené par la ligne des centres des deux sphères données, les coupe suivant deux grands cercles C,C′ (*fig.* 366). La portion du lieu cherché, située dans le plan P, se compose de l'axe radical de ces deux cercles.

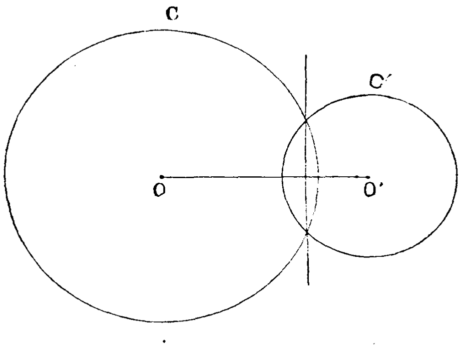

Fig. 366.

Lorsque le plan P tourne autour de la ligne des centres, l'axe radical qui vient d'être considéré engendre un plan, lequel constitue le lieu cherché.

Lorsque les deux sphères se coupent, le plan radical n'est autre que le plan du cercle *c* d'intersection, car tout point situé dans ce plan a pour puissance, tant par rapport à l'une que par rapport à l'autre sphère, sa puissance par rapport au cercle *c*. Ce qui précède montre que les points de ce plan sont les seuls à avoir même puissance par rapport aux deux sphères ; le même fait peut d'ailleurs se reconnaître directement (Comparer Pl., **137**).

Le plan radical ou, du moins, la portion de ce plan extérieure aux deux sphères (dans le cas où celles-ci se coupent) est (²) le lieu des centres des sphères orthogonales à la fois aux deux premières.

---

(1) Comparer Pl., **135**.
(2) Comparer Pl., **138**.

Si les sphères données sont concentriques, le plan radical est rejeté à l'infini(1).

458. Théorème. — *Les plans radicaux de trois sphères prises deux à deux se coupent suivant une même droite (sauf dans le cas où les trois centres sont en ligne droite et où les trois plans radicaux sont parallèles entre eux ou tous confondus).*

En effet, si S, S′, S″ sont les trois sphères considérées dont les centres ne sont pas en ligne droite, le plan radical de S et de S′ coupe le plan radical de S et de S″ suivant une droite, laquelle est le lieu des points qui ont même puissance par rapport aux trois sphères et se trouve, par conséquent, située dans le plan radical de S′ et de S″.

La droite dont nous venons de démontrer l'existence se nomme *axe radical* des trois sphères. Elle n'est d'ailleurs autre que la perpendiculaire au plan des trois centres, menée par le centre radical des grands cercles suivant lequel ce plan coupe les sphères données.

L'axe radical (ou du moins sa partie extérieure aux trois sphères), est le lieu des centres des sphères qui coupent les premières orthogonalement.

Si les trois centres sont en ligne droite, les trois plans sont tous perpendiculaires à cette droite.

Si le plan radical de S, S′ coïncide avec le plan radical de S, S″, il sera aussi le plan radical de S′, S″, puisque tous les points de ce plan auront même puissance par rapport aux trois sphères.

Théorème. — *Les six plans radicaux de quatre sphères prises deux à deux (ou les quatre axes radicaux de ces quatre sphères prises trois à trois) se coupent en un même point* (dit *centre radical* des quatre sphères), *sauf le cas où les quatre centres sont dans un même plan, les plans radicaux étant alors parallèles à une même droite.*

Car si les plans radicaux de S avec S′, de S avec S″, de S avec S‴ se coupent en un point I, ce point aura même puissance par rapport aux quatre sphères S, S′, S″, S‴ et appartiendra, par suite, aux autres plans radicaux.

Il sera (s'il est extérieur aux quatre sphères) le centre d'une sphère qui les coupera toutes quatre orthogonalement.

---

(1) Comparer Pl., **136**, Rem. II.

459. Sphères homothétiques.

La figure homothétique d'une sphère est une sphère, le rapport des rayons étant égal au rapport d'homothétie et les deux centres se correspondant.

Si, en effet, O est le centre de la sphère donnée, O′ son homothétique ; M un point quelconque de ladite sphère, M′ son homologue, le segment O′M′ est, avec OM, dans un rapport égal au rapport d'homothétie (**426**) : ce segment est donc bien constant.

460. Réciproquement, *deux sphères quelconques sont deux figures homothétiques, et cela de deux façons différentes* (¹) : *l'homothétie étant directe dans un cas, inverse dans l'autre.*

En effet, si O, O′ sont les centres des deux sphères; OM, O′M′ deux rayons parallèles et de même sens, mais quelconques d'ailleurs (le point M prenant successivement toutes les positions possibles sur la première sphère), les segments OM, O′M′ satisfont aux conditions énumérées dans l'hypothèse du nº **427**.

La même conclusion subsisterait si OM, O′M étaient des rayons parallèles et de sens contraires : le théorème est donc démontré.

Le théorème résulte encore, d'ailleurs, de ce que les grands cercles obtenus en coupant les sphères par un plan passant par la ligne des centres sont homothétiques entre eux, le centre et le rapport d'homothétie étant indépendants du choix du plan sécant.

Remarque. — *Deux sphères ne peuvent être homothétiques de plus de deux façons différentes.* Car, d'après le numéro précédent, dans toute homothétie qui transforme la première sphère en la seconde, les centres se correspondent et le rapport d'homothétie est égal au rapport des rayons. Or il n'y a que deux points qui divisent la ligne des centres dans un rapport égal à celui des rayons (²).

On nomme, comme en géométrie plane, *centre de similitude externe* et *centre de similitude interne*, les centres des deux homothéties dont l'existence vient d'être constatée.

461. La propriété qui appartient aux sphères, d'être homothétiques de deux manières différentes, leur est commune avec toutes les figures douées de centres de symétrie.

---

(1) Toutefois, il importe de remarquer que deux sphères égales ne peuvent être considérées comme homothétiques que moyennant l'extension donnée à ce mot au nº **427**.

(2) Voir aussi exercice 612.

Soient, en effet, F une figure quelconque; F_1, sa symétrique par rapport à un point O. Toute homothétique F' de F sera (**427**) homothétique de F_1, le centre d'homothétie étant en général différent du premier et les homothéties étant de noms contraires.

Si maintenant la figure F admet le point O comme centre de symétrie, elle coïncide avec F_1 : elle est, par conséquent, homothétique à F' de deux manières différentes.

Les deux homologues M, M_1 (*fig.* 367) d'un même point M' de F' sont toujours symétriques l'un de l'autre par rapport au point O. On choisira entre les deux homothéties en indiquant lequel de ces deux points on fait correspondre au point M'.

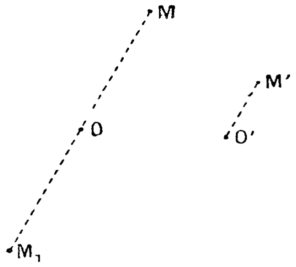

FIG. 367.

462. *Trois sphères ont quatre axes d'homothétie,* comme trois cercles en géométrie plane.

Ces quatre axes d'homothétie sont, en effet, les mêmes que ceux des grands cercles déterminés par le plan des trois centres.

Quatre sphères ont huit plans d'homothétie (**428**). Car, ayant pris un point M sur la première sphère O, et menant, dans les trois autres O', O'', O''', les diamètres $M'M'_1$, $M''M''_1$, $M'''M'''_1$ parallèles au rayon OM, on pourra choisir, parmi les extrémités de chacun de ces diamètres, le point qui sera l'homologue du point M dans l'homothétie correspondante. Une fois choisis ainsi les trois homologues du point M, on aura un plan d'homothétie bien déterminé. Or ce triple choix peut se faire de huit façons différentes : car on peut d'abord choisir entre les points M'_1, M'_1, ainsi qu'entre les points M'', M''_1, ce qui peut se faire de quatre manières, et à chacune des quatre combinaisons ainsi obtenues en correspondent deux, différentes entre elles par le choix fait entre M''' et M'''_1.

Chaque axe d'homothétie des trois premières sphères, correspondant au choix fait entre M' et M'_1 d'une part, M'' et M''_1 de l'autre, *est contenu dans deux plans d'homothétie; chaque centre d'homothétie, dans quatre.*

463. Plans tangents communs à deux sphères.

Tout plan tangent commun à deux sphères passe par un centre de similitude : le centre de similitude externe, si le plan tangent est *extérieur* (*c'est-à-dire laisse les deux sphères d'un même côté*); *le centre de similitude interne, si le plan tangent est intérieur* (*c'est-à-dire laisse les deux sphères de part et d'autre*).

Cela résulte de ce que les rayons OA, O'A' (*fig.* 368) qui aboutissent aux deux points de contact, sont parallèles entre eux, comme étant tous deux perpendiculaires au plan tangent commun.

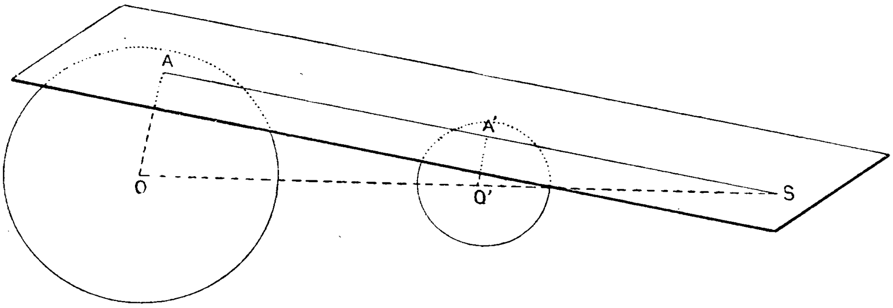

FIG. 368.

Réciproquement, tout plan tangent à l'une des sphères, mené par un des centres de similitude, est tangent à l'autre sphère, en raison de l'homothétie des deux figures par rapport à ce point.

Les plans tangents communs à deux sphères sont donc les plans tangents à l'un ou à l'autre des deux cônes circonscrits à l'une d'elles et ayant pour sommets les centres de similitude, si ces cônes existent.

Si on se reporte aux raisonnements du n° **452**, on voit immédiatement que l'un quelconque de ces deux cônes a pour méridienne une tangente commune menée, par le centre de similitude correspondant, aux grands cercles C, C' déterminés, dans les deux sphères données, par un plan passant par la ligne des centres. Les deux cônes existeront donc si les deux sphères sont extérieures; un seul d'entre eux, si les sphères se coupent; aucun, si elles sont intérieures l'une à l'autre.

464. Plans tangents communs à trois sphères.

Tout plan tangent commun à trois sphères passe par un de leurs axes de similitude : il doit, en effet (n° précéd.), passer par un centre de similitude de la première et de la seconde sphère donnée et par un centre de similitude de la première et de la troisième.

Inversement, tout plan, tangent à l'une des sphères et passant par un de leurs axes de similitude, est tangent aux deux autres sphères.

D'après cela, à chaque axe de similitude des trois sphères, pourvu qu'il soit sans point commun avec ces sphères, correspondent deux plans tangents communs passant par cet axe. Si aucun des quatre axes de similitude ne coupe les sphères, il y aura huit plans tangents communs. Si cette condition n'est pas remplie, le nombre des plans tangents s'abaisse de deux ou plusieurs unités : il peut n'y avoir aucun plan tangent commun (par exemple si deux des sphères données sont intérieures l'une à l'autre).

EXERCICES

658. Une surface telle que le cercle qui passe par trois quelconques de ses points y soit contenu tout entier, est une sphère (ou un plan).

659. Si des cercles, en nombre quelconque, sont tels que deux quelconques d'entre eux se coupent en deux points :

Ou bien tous ces cercles ont deux points communs ;

Ou bien ils appartiennent tous à une même sphère.

660. Quelle est la surface engendrée par une circonférence, lorsqu'elle tourne autour d'un axe qui se projette sur son plan suivant un de ses diamètres ?

661. Lieu des projections d'un point donné sur les plans qui passent par un point fixe.

661 *bis*. Lieu des centres des sections faites dans une sphère par des plans qui passent par un point fixe ou par une droite fixe.

662. Le centre d'une sphère est fixe, pendant que son rayon varie. Lieu du cercle de contact du cône circonscrit à cette sphère, ayant pour sommet un point donné.

663. Lieu des points d'où l'on peut mener à une sphère trois tangentes formant un trièdre trirectangle.

664. Lieu des points d'où l'on peut mener à une sphère trois plans tangents formant un trièdre trirectangle.

665. Le lieu des points tels que le rapport de leurs distances à deux points donnés soit constant, est une sphère.

666. Le lieu des points tels que leurs distances à trois points donnés soient proportionnelles à trois nombres donnés, est un cercle orthogonal à toutes les sphères qui passent par les trois points donnés.

667. Lieu des points tels que les cônes circonscrits à deux sphères données et ayant ces points pour sommets soient égaux.

Même problème, lorsqu'il y a trois sphères données.

668. Lieu des points tels que la somme des carrés de leurs distances à deux points donnés, multipliés respectivement par des coefficients donnés, soit constante.

669. Lieu des points dont les puissances par rapport à deux sphères données soient entre elles comme deux nombres donnés.

Même problème, lorsqu'il y a trois sphères données et trois nombres donnés.

670. Lieu des extrémités d'un segment égal et parallèle à un segment fixe, sachant que ces extrémités sont respectivement sur deux sphères données.

672. Trouver un triangle se déduisant par une translation d'un triangle donné et ayant ses sommets sur trois sphères données.

673. Un plan fixe quelconque est coupé par des sphères qui passent par un même cercle suivant des cercles ayant même axe radical.

674. Trouver une sphère, connaissant un cercle et le plan tangent en un point de ce cercle.

675. Plus généralement, trouver une sphère qui passe par un cercle donné et soit tangente à un plan ou à une sphère donnée.

676. Faire passer, par un cercle donné, une sphère orthogonale à une sphère donnée.

677. Trouver une sphère ayant son centre sur une droite donnée, tangente à une droite donnée et passant par un point donné.

678. Trouver une sphère ayant son centre sur une droite donnée, tangente à une droite donnée et à un plan donné.

679. Trouver une sphère passant par un cercle donné et tangente à un autre cercle donné. Discussion.

680. Trouver une sphère coupant à angle droit deux cercles donnés (1). Cas d'impossibilité.

681. Une sphère varie en passant par un cercle fixe C. Lieu du cercle de contact du cône circonscrit à la sphère, ayant pour sommet un point donné du plan de C. — Lieu des points de contact des plans tangents menés à la sphère par une droite donnée du plan de C. — Lieu de la droite d'intersection des plans tangents en deux points du cercle.

Qu'arrive-t-il lorsqu'une sphère varie en passant par deux points fixes A, B et restant tangente à une droite fixe qui rencontre AB prolongée?

682. Une sphère variable passe par deux points fixes, et est tangente à un plan fixe ou à une sphère fixe. Lieu du point de contact.

682 *bis*. Trouver une sphère passant par deux points donnés et tangente à deux plans donnés.

683. Une sphère variable est tangente à deux droites fixes et a son centre dans le plan mené parallèlement à ces deux droites, à égale distance de l'une et de l'autre. Trouver le lieu de ce centre.

684. Il existe, en général, une sphère et une seule tangente aux côtés d'un quadrilatère gauche. Mais si la somme de deux côtés est égale à celle des deux autres, le nombre des sphères tangentes aux quatre côtés est infini. Trouver, dans ces conditions, le lieu des centres des sphères. Quelle est celle dont le rayon est le plus petit ?

Déduire de là la solution de l'exercice 558.

---

(1) Une sphère et un cercle se coupent à angle droit lorsque la tangente au cercle, menée par un point commun, est perpendiculaire au plan tangent à la sphère au même point.

684 *bis*. Quelles sont les conditions pour qu'il existe une sphère tangente aux six arêtes d'un tétraèdre?

685. Inscrire ou ex-inscrire une sphère à un tétraèdre.

686. Une sphère varie en restant tangente à un plan fixe en un point fixe. Lieu des points de contact des plans tangents parallèles à un plan donné.

687. Si un cercle variable rencontre deux cercles fixes chacun en deux points, son plan passe par un point fixe.

688. Trouver un cercle qui divise deux cercles donnés en deux parties égales.

688 *bis*. Trouver un cercle qui soit divisé par deux cercles donnés en deux parties égales.

689. Trouver un cercle tangent à deux cercles donnés.

690. Si deux sphères n'ont aucun point commun, il existe deux points (*points limites*) tels que les sphères réduites à ces points, aient, avec les sphères données, même plan radical. Toute sphère orthogonale aux deux premières passe par les points limites.

691. Si trois sphères qui n'ont pas même plan radical n'ont aucun point commun, les sphères qui leur sont orthogonales passent par un cercle fixe. Ce cercle est le lieu des points tels que les sphères réduites à ces points aient, avec les sphères données, même axe radical.

La sphère orthogonale à quatre sphères données (si elle existe) est le lieu des points tels que les sphères réduites à ces points aient, avec les sphères données, même centre radical.

Quel est le lieu des points limites de deux sphères dont chacune varie en passant par un cercle fixe?

692. Autour d'un point fixe comme sommet, on fait tourner un trièdre trirectangle dont les arêtes rencontrent une sphère donnée. Montrer :

1° Que la somme des carrés des cordes interceptées par la sphère sur les trois arêtes reste constante;

2° Qu'il en est de même de la somme des carrés des six segments qui vont du sommet aux points d'intersection des arêtes avec la sphère;

3° Que la somme des aires des cercles d'intersection de la sphère avec les trois faces est également constante.

693. Dans les mêmes conditions, si l'on fait passer un plan par trois des points d'intersection considérés à l'exercice précédent, le lieu de la projection du sommet du trièdre sur ce plan est une sphère (ex. 513, 1° et Pl., ex. 70).

Si le sommet du trièdre est sur la sphère, le plan en question passe par un point fixe, lequel est le centre de gravité du triangle formé par les trois points d'intersection des arêtes avec la sphère.

694. On peut choisir trois sphères de manière que leurs huit plans tangents communs (**464**) existent. (Il suffira, les centres étant choisis, de prendre les rayons suffisamment petits.)

CHAPITRE IV

GÉOMÉTRIE SPHÉRIQUE — TRIANGLES SPHÉRIQUES

465. *Par deux points de la sphère, il passe un grand cercle ; et un seul, sauf si ces deux points sont diamétralement opposés.*

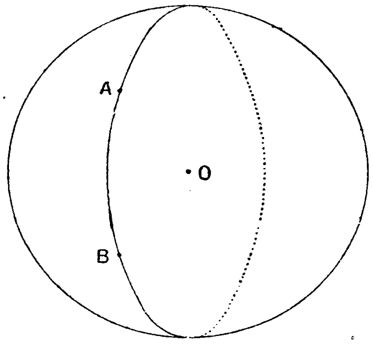

Fig. 369.

Car par deux points A, B (*fig.* 369) de la surface et le centre O passe un plan, lequel est unique si A et B ne sont pas en ligne droite avec le centre.

Remarques I. — La propriété que nous venons de constater établit une analogie entre les grands cercles de la sphère et les lignes droites en géométrie plane. Mais cette analogie n'est pas complète, puisqu'il y a exception pour les points diamétralement opposés.

II. Les points A et B divisent le grand cercle qui les contient en deux arcs. Lorsqu'on parle de *l'arc de grand cercle* qui joint A et B, on a plus spécialement en vue celui des deux qui est plus petit qu'une demi-circonférence.

466. Le diamètre perpendiculaire au plan d'un cercle de la sphère coupe la surface en deux points, qu'on nomme les *pôles* de ce cercle (*fig.* 370).

Chacun d'eux est également distant de tous les points du cercle. Si celui-ci est un grand cercle, la distance d'un quelconque de ses points au pôle est égale à la corde d'un quadrant, ainsi qu'on le reconnaît immédiatement à l'inspection de la figure 371.

Inversement, le lieu des points d'une sphère situés à une distance constante (moindre que le diamètre) d'un point de cette surface, est un cercle ayant ce point pour pôle.

D'après cela, on voit qu'on peut décrire un cercle sur une sphère comme sur un plan, à l'aide d'un compas dont l'une des pointes est placée en l'un des deux pôles de ce cercle. Toutefois, il est nécessaire (surtout si l'ouverture est un peu grande) d'employer un compas (dit compas *sphérique*) à branches courbes et non droites.

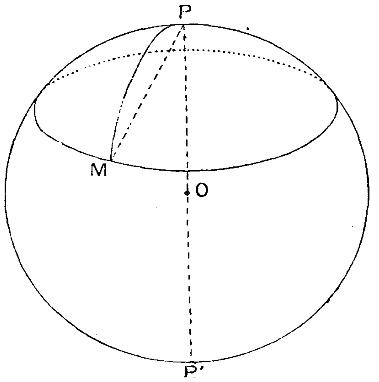

FIG. 370.

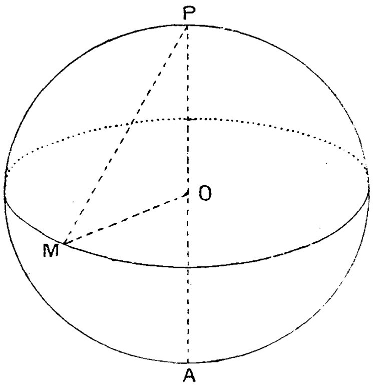

FIG. 371.

L'arc de grand cercle qui joint le pôle d'un cercle de la sphère à un point quelconque de ce cercle a également la même grandeur, quel que soit ce point. On lui donne le nom de *rayon sphérique* du cercle. Le rayon sphérique d'un grand cercle est égal à un quadrant.

Un cercle quelconque divise la sphère en deux régions, dites *calottes sphériques*, situées de part et d'autre du plan de ce cercle. Chacune de ces calottes comprend un des pôles et est formée par les points dont la distance à ce pôle est plus petite que la distance de ce même pôle aux points du cercle.

S'il s'agit d'un petit cercle, on donne souvent le nom de région *intérieure* à ce petit cercle, à la plus petite des deux calottes sphériques, celle dont le rayon sphérique est inférieur à un quadrant. Cette région intérieure est évidemment celle qui est située du côté du plan du cercle où n'est pas le centre de la sphère.

467. Angle de deux grands cercles.

Théorème. — *L'angle* (Pl., **60** *bis*) *de deux demi-grands cercles, terminés à leur diamètre commun, est égal à l'angle des demi-plans qui les contiennent. Il a pour mesure l'arc intercepté, entre eux, sur le grand cercle qui a pour pôles leurs points communs.*

Soient les deux demi-grands cercles AMA′, ANA′ (*fig.* 372), qui se coupent en A, A′.

Soient Ax, Ay les tangentes à ces courbes au point A. Ces tangentes, étant toutes deux perpendiculaires à AA′, déterminent l'angle plan du dièdre M-AA′-N.

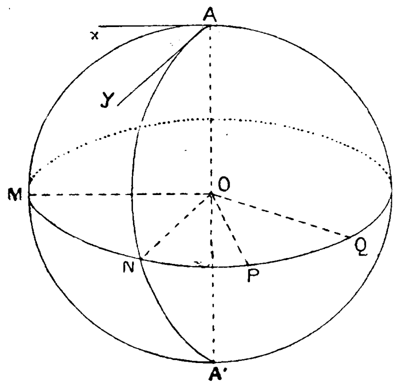

Fig. 372.

D'autre part, le grand cercle qui a pour pôles A, A′, a son plan perpendiculaire à AA′. Ce plan coupe les demi-plans AA′M, AA′N suivant deux droites OM, ON qui forment également entre elles l'angle plan du dièdre M-AA′-N. Mais cet angle est précisément l'angle au centre correspondant à l'arc de grand cercle MN : ce qui démontre la seconde partie du théorème.

468. Une autre expression de l'angle de deux grands cercles est encore donnée par le théorème suivant :

Théorème. — *L'angle de deux demi-grands cercles est mesuré par l'arc de grand cercle qui joint leurs pôles, ou par son supplément.*

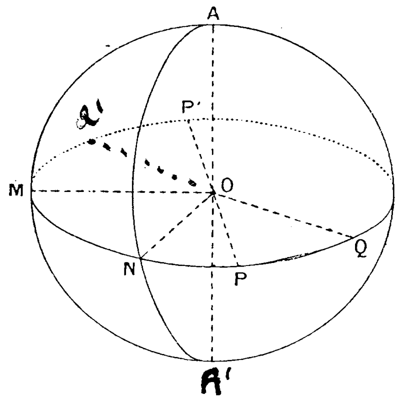

Fig. 373.

Soient, comme tout à l'heure, AMA′, ANA′ (*fig.* 373) deux demi-grands cercles qui coupent en M, N le grand cercle dont le plan est perpendiculaire à AA′. Ce dernier grand cercle contient les pôles P, P′, Q, Q′ des deux premiers : car les diamètres PP′, QQ′, respectivement perpendiculaires à AMA′, ANA′, sont tous deux perpendiculaires à AA′. De plus PP′, QQ′ sont respectivement perpendiculaires aux rayons OM, ON de la sphère : leur angle est donc égal à $\widehat{MON}$ ou à son supplément.

Remarque. — Chacun des deux grands cercles considérés ayant deux pôles, l'arc de grand cercle qui joint un pôle de l'un à un pôle de l'autre peut être choisi de quatre façons différentes.

Supposons que les pôles P, Q (*fig.* 373) soient tels que les trièdres O·AMP, O·ANQ aient même disposition. Alors ces deux trièdres, étant trirectangles, sont égaux et, comme ils ont l'arête OA commune, ils se déduisent l'un de l'autre par rotation autour de cette droite. Ceci revient à dire que les angles $\widehat{MON}$ (angle des deux demi-grands cercles) et $\widehat{POQ}$ sont égaux.

D'ailleurs, si l'on remplace le pôle P par son opposé P′ (*fig.* 373), la disposition du trièdre O·AMP change évidemment. Comme l'angle $\widehat{MOP'}$ est supplémentaire de $\widehat{MOP}$, on voit que l'*arc de grand cercle qui joint les deux pôles a même mesure que l'angle des deux demi-grands cercles ou que son supplément, suivant que les trièdres* O·AMP, O·ANQ *ont ou non même disposition.*

La disposition du trièdre O·AMP dépend d'ailleurs du sens de rotation de l'angle $\widehat{AOM}$, vu du point P. Donc *l'angle de deux demi-grands cercles, en un de leurs points communs, est égal à l'arc de grand cercle qui joint leurs pôles, si l'on choisit ceux-ci de telle manière que chacun des deux demi-grands cercles, vu de son pôle, paraisse tourner dans le sens direct, quand on le suit à partir du point commun considéré.*

469. Théorème. — *Le lieu des points d'une sphère également distants de deux points donnés sur cette surface, est le grand cercle perpendiculaire au milieu de celui qui joint les deux points donnés.*

Soient A, B, les deux points donnés (*fig.* 374). Le plan perpendiculaire au milieu de AB passe par le centre O de la sphère (puisque OA = OB) et coupe, par conséquent, la surface suivant un grand cercle, qui est le lieu cherché (**340**). Celui-ci passe d'ailleurs par le milieu de l'arc de grand cercle AB (point équidistant de A et de B); de plus, en vertu du n° **467**, il est perpendiculaire à cet arc; car leurs deux plans sont évidemment perpendiculaires entre eux.

Remarque. — Le lieu précédent divise la surface en deux hémisphères qui contiennent (**340**, Remarque) l'un les points plus rapprochés de A que de B, l'autre les points plus rapprochés de B que de A.

470. Problème. — *Trouver le rayon d'une sphère solide, à l'aide de constructions effectuées sur la surface et de constructions planes.*

Première solution. — Prenant deux points quelconques A, B sur la surface (*fig.* 374), traçons, avec une même ouverture de compas, deux cercles ayant ces deux points comme pôles respectifs. Un point M, commun à ces deux cercles, étant également distant de A et de B, appartiendra au grand cercle obtenu au n° précédent. En répétant la même construction avec d'autres ouvertures de compas, on aura deux autres points N, P du même grand cercle. Relevant alors au compas les distances MN, NP, PM, on pourra construire, sur un plan (*fig.* 375), un triangle égal à MNP, et le cercle circonscrit à ce triangle aura, pour rayon, le rayon cherché.

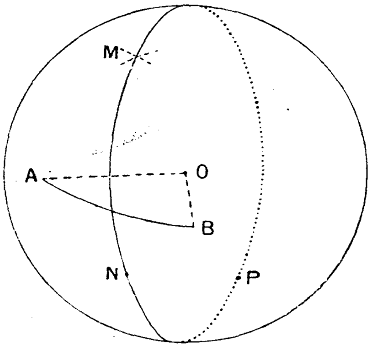

FIG. 374.

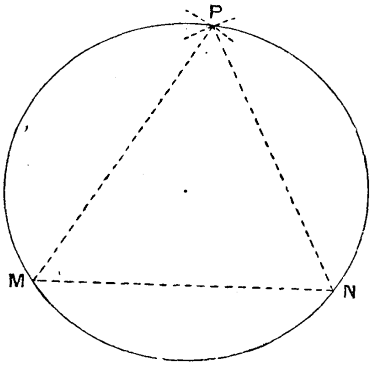

FIG. 375.

Deuxième solution. — D'un point quelconque C de la sphère comme pôle, avec une ouverture de compas déterminée, décrivons un cercle, sur lequel nous prendrons trois points M, N, P. Nous pourrons, comme il vient d'être expliqué, déterminer le rayon de ce cercle en construisant un triangle égal à MNP.

Or la connaissance de ce rayon et de celle de la distance CM (égale à l'ouverture de compas initiale) permettent de trouver le rayon de

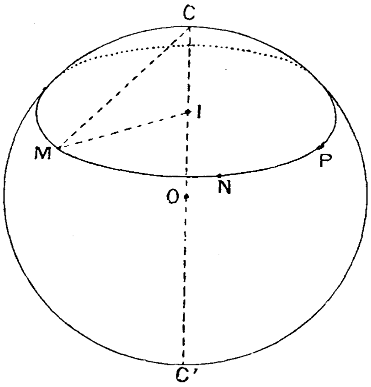

FIG. 376.

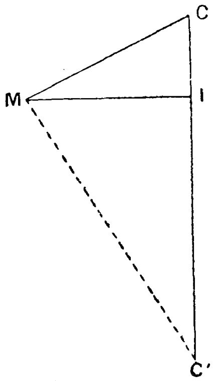

FIG. 376 *bis*.

la sphère; car, si I est le centre du cercle MNP (*fig.* 376), le triangle CIM est rectangle et peut être construit sur un plan (puisqu'on en connaît l'hypoténuse CM et un côté MI) : prolongeant alors (*fig.* 376 *bis*) le côté CI et menant, en M, la perpendiculaire à CM jusqu'à rencontre en C′ avec CI prolongé, la longueur CC′ sera le diamètre de la sphère.

470 *bis.* **Problème.** — *Joindre deux points donnés d'une sphère solide par un arc de grand cercle.*

De chacun des points donnés comme pôle, décrivons un grand cercle (ce que nous savons faire après la résolution du problème précédent). Ces deux grands cercles se coupent en deux points, qui sont manifestement les pôles du grand cercle cherché.

Remarque. — Il est clair que la solution de ce problème, combinée avec la deuxième solution du problème précédent, permet de *tracer le grand cercle perpendiculaire au milieu de l'arc de grand cercle qui joint deux points donnés* (**462**).

471. Polygones sphériques.

On donne le nom de *polygone sphérique* à toute portion de sphère limitée par des arcs de grands cercles (appelés *côtés* du polygone), plus petits qu'une demi-conférence, limités à leurs intersections successives.

Un polygone sphérique est dit *convexe* (*fig.* 377) lorsqu'il est situé d'un seul côté par rapport à chacun des grands cercles dont font partie ses côtés; *concave* (*fig.* 378), dans le cas contraire.

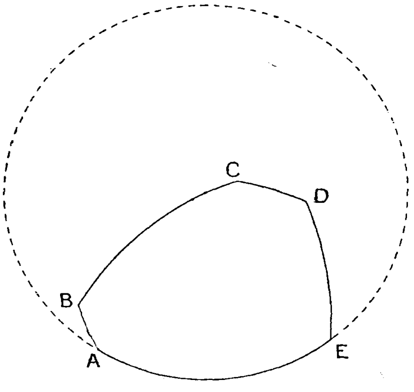

Fig. 377.

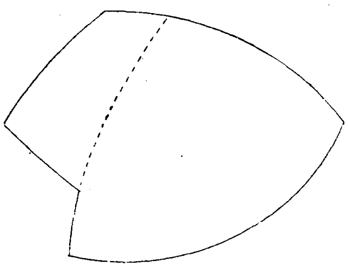

Fig. 378.

Les polygones sphériques se classent, comme les polygones plans, d'après le nombre de leurs côtés, le plus simple (1) étant le *triangle sphérique* (*fig.* 379).

Un triangle sphérique est toujours un polygone convexe. Car, dans le triangle ABC (*fig.* 379), si l'arc de grand cercle AC n'était pas tout entier du même côté par rapport au grand cercle AB, il couperait ce dernier en un point (autre que A) situé entre A et C et serait, par conséquent, plus grand qu'une demi-circonférence, contrairement à la définition.

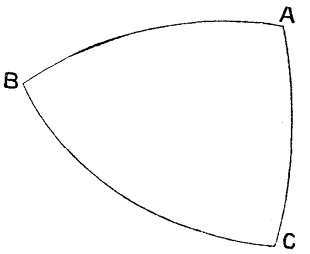

Fig. 379.

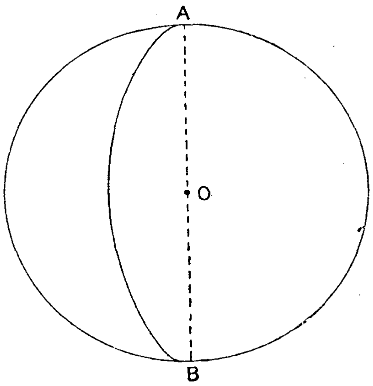

Fig. 380.

Relations entre les polygones sphériques et les angles polyèdres.

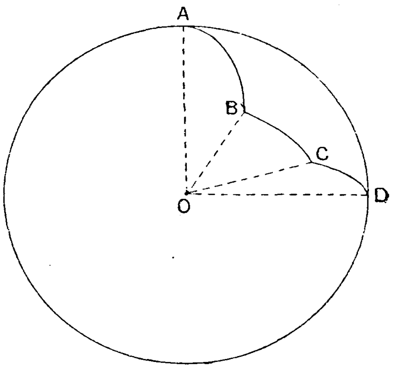

Fig. 381.

A tout polygone sphérique correspond un angle polyèdre, ayant pour sommet le centre de la sphère et pour arêtes les droites qui joignent ce centre aux différents sommets (*fig.* 381).

Les faces $\widehat{AOB}$, $\widehat{BOC}$, etc. (*fig.* 381) *de cet angle polyèdre sont les angles au centre correspondant aux côtés* AB, BC, etc., *du polygone.*

Les dièdres de l'angle polyèdre ont, d'après le théorème du n° **469**, *leurs rectilignes égaux aux angles du polygone.*

Inversement, tout angle polyèdre ayant pour sommet le centre d'une sphère coupe celle-ci suivant un polygone sphérique ayant, avec l'angle polyèdre, les relations que nous venons d'indiquer.

---

(1) Deux demi-grands cercles terminés au même diamètre comprennent entre eux une figure (fig. 380), dite *fuseau sphérique* (Compléments, n° **693**), qui peut évidemment être considérée comme un *biangle.* Seulement les côtés de cette sorte de polygone sont *égaux* et non *inférieurs* à une demi-circonférence.

471 *bis*. D'après cela, on voit que, de toute propriété relative aux faces et aux dièdres d'un angle polyèdre, on peut déduire une propriété relative aux côtés et aux angles du polygone sphérique correspondant, et inversement.

Par exemple, le théorème du n° **373** nous donnera le suivant : *Un côté quelconque d'un polygone sphérique est plus petit que la somme des autres* (d'où résulte : *un côté quelconque d'un triangle sphérique est plus grand que la différence des deux autres*).

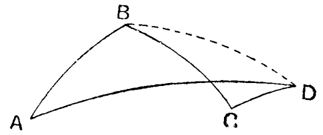

Fig. 382.

On voit dès lors, comme en géométrie plane (Pl., **26**), que *l'arc de grand cercle* (*moindre qu'une demi-circonférence*) *qui joint deux points est plus court que toute ligne polygonale sphérique terminée aux mêmes extrémités* (*fig.* 382) (1).

Cet arc de grand cercle est dit mesurer la *distance sphérique* des deux points donnés.

En nous bornant au cas d'un triangle sphérique, le théorème du n° **375** nous montre qu'inversement, trois arcs de grands cercles, moindres qu'une demi-circonférence et dont chacun est moindre que la somme des deux autres, tandis que la somme des trois est inférieure à une circonférence entière, sont égaux aux côtés d'un même triangle sphérique.

472. *Quand un polygone sphérique convexe est intérieur à un polygone sphérique quelconque* (*un ou plusieurs sommets ou côtés pouvant être communs*) (*fig.* 383, 383 *bis*), *le périmètre du*

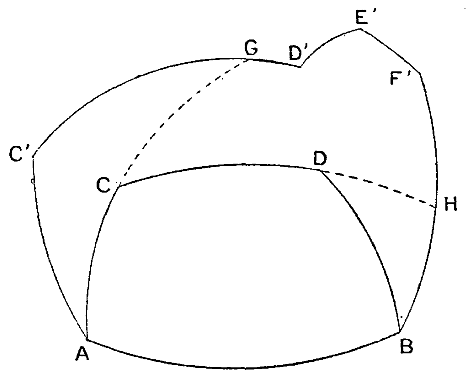

Fig. 383.

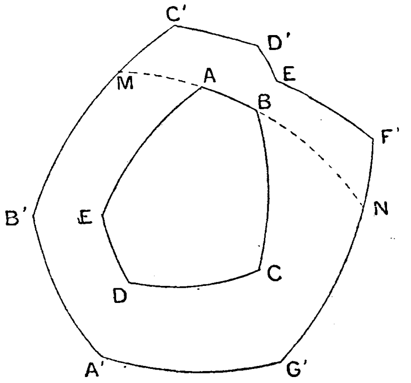

Fig. 383 *bis*.

---

(1). La figure montre la construction à effectuer.

polygone enveloppé est inférieur à celui du polygone enveloppant.

Cet énoncé est, d'après ce que nous venons de dire, identique à celui du n° **373** *bis*, un polygone sphérique étant manifestement convexe ou concave en même temps que l'angle polyèdre correspondant. Mais on peut refaire le raisonnement sur la figure sphérique : c'est ce que montrent les figures 383, 383 *bis*, où, pour plus de facilité, les notations ont été prises conformes à celles de la Géométrie plane (Pl., **27**).

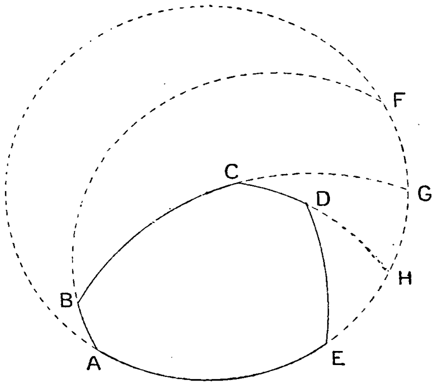

Fig. 384.

La démonstration continue à s'appliquer quand le polygone enveloppant est remplacé par une circonférence de grand cercle (voir *fig.* 384). Donc :

Le périmètre d'un polygone sphérique convexe est inférieur à la circonférence d'un grand cercle, énoncé qui équivaut à celui du n° **374**.

473. La *disposition* d'un triangle sphérique ABC est le sens de l'angle $\widehat{BAC}$ de ce triangle, vu de l'extérieur de la sphère. Il est clair que cette disposition est la même que celle du trièdre correspondant.

Deux triangles sphériques sont dits *symétriques*, lorsqu'ils ont tous leurs éléments égaux chacun à chacun, mais diffèrent par la disposition.

C'est ce qui arrive, en particulier, pour deux triangles symétriques l'un de l'autre par rapport au centre de la sphère, c'est-à-dire tels que les sommets de l'un soient diamétralement opposés aux sommets de l'autre.

474. Triangles sphériques polaires.

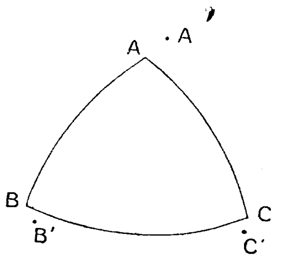

Fig. 385.

Soit ABC un triangle sphérique (*fig.* 385). Soient A′, celui des deux pôles du grand cercle BC qui est situé (par rapport à ce grand cercle) dans le même hémisphère que A ; B′, celui des deux pôles du grand cercle CA qui est (par rapport au grand cercle CA), dans le même hémisphère que B ; C′, celui des pôles du grand cercle AB qui est dans le même hémisphère que C. Le triangle sphérique A′B′C′ est dit le triangle *polaire* de ABC.

Si l'on se reporte à la définition des trièdres supplémentaires (**377**), on voit que *le trièdre qui correspond au triangle* A′B′C′ *est le supplémentaire du trièdre qui correspond au triangle* ABC.

Dès lors, les théorèmes des n^{os} **377**, **378** donnent immédiatement les conclusions suivantes :

Si un triangle sphérique est polaire d'un autre : réciproquement, celui-ci est polaire du premier.

Si deux triangles sphériques sont polaires l'un de l'autre, chaque côté de l'un est supplémentaire de l'angle correspondant de l'autre.

474 *bis.* Nous venons de déduire les propriétés des triangles polaires de celles des trièdres supplémentaires. Mais on peut également démontrer directement les mêmes propositions, en suivant, relativement aux triangles sphériques, la marche correspondante à celle qui a été suivie pour les trièdres.

On démontrera d'abord les lemmes suivants :

Lemmes. — I. *L'arc de grand cercle qui joint un point quelconque* B *de la sphère à l'un des pôles* A *d'un grand cercle donné, est plus petit ou plus grand qu'un quadrant, suivant que le pôle* A *est ou non, par rapport au grand cercle donné, dans le même hémisphère que le point* B.

Car le grand cercle AB et le grand cercle donné se coupent en un point I (*fig.* 386, 386 *bis*), tel que l'arc AI soit égal à un quadrant. Cet arc AI est d'ailleurs évidemment supérieur ou inférieur à AB,

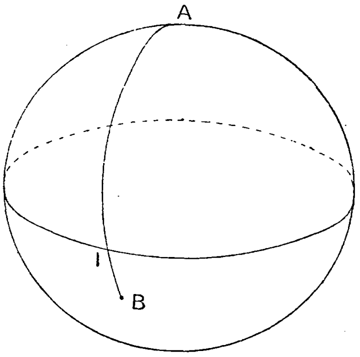

Fig. 386.

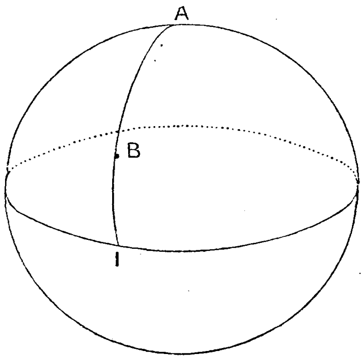

Fig. 386 *bis*.

suivant que les points A et B sont du même côté (*fig.* 386) ou de côtés différents (*fig.* 386 *bis*) du point I.

II. *Étant donnés deux demi-grands cercles* AMA′, ANA′ (*fig.* 387) *terminés à un diamètre commun, si* P *est le pôle du premier situé, par rapport à celui-ci, dans le même hémisphère que le second, et* Q, *le pôle du second situé, par rapport à lui, dans le même hémisphère que le premier, l'arc de grand cercle* PQ [1] *est supplémentaire de l'angle des deux demi-grands cercles donnés.*

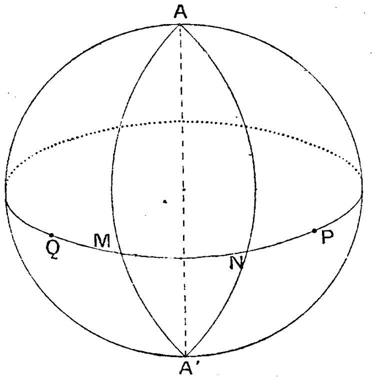

Fig. 387.

D'après le n° **468**, il suffit, pour arriver à cette conclusion, de démontrer que, si M et N sont les points où nos deux demi-grands cercles rencontrent le grand cercle de pôle A, les arcs MP et NQ sont de sens contraires. Or, cela est manifeste, puisque, d'après l'hypothèse, l'arc MP a le sens MN et l'arc NQ, le sens NM.

Cela posé, il est aisé de démontrer la première propriété des triangles polaires, à savoir que *si un triangle est polaire d'un autre, inversement celui-ci est polaire du premier.*

Soit, en effet, ABC un triangle sphérique (*fig.* 385) dont A′B′C′ est le polaire. C′ étant un pôle de AB, l'arc de grand cercle AC′ est égal à un quadrant, et il en est de même de l'arc AB′, pour une raison analogue : d'où résulte que A est un pôle de l'arc de grand cercle B′C′ (**470** *bis*).

Ce pôle est d'ailleurs bien, par rapport au grand cercle B′C′, dans le même hémisphère que le point A. Car cette conclusion et la partie de l'hypothèse d'après laquelle le point A′ est, par rapport au grand cercle BC, du même côté que A, expriment, en vertu du lemme I, un seul et même fait : à savoir, que l'arc de grand cercle AA′ est inférieur à un quadrant.

Quant à la propriété d'après laquelle *chaque côté d'un triangle sphérique est le supplément de l'angle correspondant du triangle polaire du premier*, elle résulte évidemment du lemme II : car les sommets du premier triangle sont bien déduits des côtés du second comme le veut l'hypothèse de ce lemme.

---

[1] Voir **465**, Remarque II.

Des propriétés fondamentales des triangles polaires, on déduit, comme au n° **379**, les théorèmes suivants :

Chaque angle d'un triangle sphérique, augmenté de deux droits, est plus grand que la somme des deux autres ;

La somme des angles d'un triangle sphérique est comprise entre deux droits et six droits.

475. La notion de *polygones sphériques polaires* s'étend au cas où le nombre des côtés est quelconque. C'est ce que nous allons voir en nous bornant toutefois, pour simplifier, aux polygones convexes.

Soit, par exemple, le polygone ABCDE (*fig.* 388), lequel est supposé convexe. Désignons par P le pôle du grand cercle AB, pris dans l'hémisphère qui contient le polygone par Q, le pôle de BC, pris dans l'hémisphère qui contient le polygone ; par R, le pôle de CD, par S le pôle de DE, par T celui de EA, chacun de ces pôles étant pris, par rapport au grand cercle correspondant, du même côté que le polygone donné.

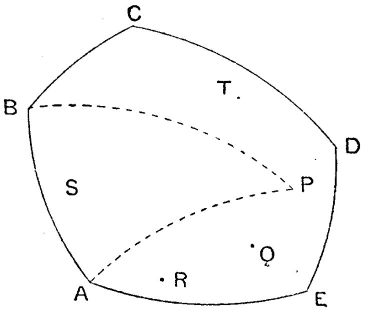

Fig. 388.

Les points P,Q,R,S,T sont les sommets d'un nouveau polygone sphérique, qui est dit le *polaire* du premier.

Nous allons voir que, réciproquement, le polaire du polygone PQRST est le polygone donné ABCDE.

Pour cela, nous remarquerons que les distances sphériques PA, PB sont égales à un quadrant, mais que les distances PC, PD, PE sont inférieures à un quadrant, puisque les points C,D,E sont, par rapport au grand cercle AB, dans l'hémisphère qui contient le point P. De même les distances QB, QC sont égales à un quadrant, les distances QD, QE, QA inférieures à un quadrant; et ainsi de suite, chaque pôle étant distant d'un quadrant des sommets situés sur le côté qui lui correspond, et situé à une distance moindre qu'un quadrant des autres sommets.

Les distances PB, QB étant égales à un quadrant, le point B est un pôle du grand cercle PQ ; de même, C est un pôle du grand cercle QR ; et ainsi de suite.

D'ailleurs, les distances BR, BS, BT sont, d'après ce qui vient d'être dit, inférieures à un quadrant. Donc (n° précédent, lemme I) les points R,S,T sont d'un même côté du grand cercle PQ, celui où est situé le point B.

Le même raisonnement pouvant se répéter pour les autres côtés du polygone PQRST, nous voyons : 1° que ce polygone est convexe comme le premier ; 2° que les points B,C,D... sont bien déduits de ses côtés, comme les points P,Q,R... l'ont été des côtés du polygone ABCDE.

Quant à cette proposition que les côtés d'un polygone sont les suppléments des angles du polygone polaire, elle se démontre comme pour les triangles.

476. De même que les propositions des numéros précédents, celles qui vont suivre correspondent, pour la plupart, à des théorèmes démontrés au livre V (nos **380-384**) et peuvent soit se déduire de ces théorèmes, soit se démontrer directement par une marche exactement calquée sur celle qui a été suivie au livre V.

Cas d'égalité des triangles sphériques.

Sur une même sphère ou sur des sphères égales, deux triangles sphériques sont égaux ou symétriques :

1° *Lorsqu'ils ont un côté égal adjacent à deux angles égaux chacun à chacun ;*

2° *Lorsqu'ils ont un angle égal compris entre deux côtés égaux chacun à chacun ;*

3° *Lorsqu'ils ont les trois côtés égaux chacun à chacun ;*

4° *Lorsqu'ils ont les trois angles égaux chacun à chacun.*

Propriétés du triangle sphérique isoscèle.

Tout triangle sphérique isoscèle est superposable à son symétrique ; inversement, tout triangle sphérique égal à son symétrique est isoscèle.

Dans un triangle sphérique isoscèle, les angles opposés aux côtés égaux sont égaux. Inversement, tout triangle sphérique qui a deux angles égaux est isoscèle.

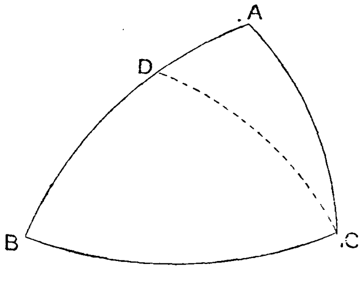

Fig. 389.

Théorème. — *Dans un triangle sphérique, à des angles inégaux correspondent des côtés inégaux, et inversement ; au plus grand angle correspond le plus grand côté* (*fig.* 389) (1).

476 *bis*. **Théorème.** — *Si deux triangles sphériques d'une même sphère ont un angle inégal compris entre côtés égaux chacun à chacun, les troisièmes côtés sont inégaux et au plus grand angle est opposé le plus grand côté.*

Soient les deux triangles sphériques ABC, A'B'C', tels que l'on ait

---

(1) Pour suivre la marche analogue à celle qui a été suivie au Ve livre (n° **384**), il faut, étant donné le triangle ABC dans lequel $\widehat{B} < \widehat{C}$, tracer, jusqu'à rencontre avec AB, l'arc de grand cercle CD tel que $\widehat{DCB} = \widehat{B}$. Cette construction est représentée sur la figure 389.

$\widehat{A} > \widehat{A'}$, $AB = A'B'$, $AC = A'C'$. Supposons que ces deux triangles aient même disposition et transportons le second sur le premier, de manière à faire coïncider les côtés égaux A'B', AB. Le côté A'C' viendra alors suivant un arc de grand cercle AC_1 égal à AC, mais intérieur à l'angle $\widehat{A}$ (*fig* 390).

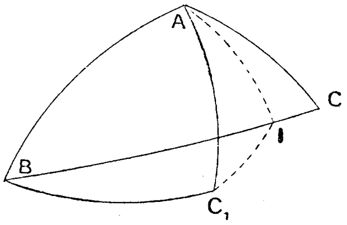

Fig. 390.

Menons l'arc de grand cercle qui divise en deux parties égales l'angle $\widehat{C_1AC}$, Cet arc de grand cercle passe, lui aussi, à l'intérieur de l'angle $\widehat{BAC}$ et coupe, par conséquent, l'arc BC en un point I situé entre B et C. Menons encore l'arc de cercle C_1I.

Les deux triangles sphériques ACI, AC_1I sont symétriques (1), comme ayant un angle égal (en A, par construction) compris entre côtés égaux chacun à chacun (AI commun; $AC = AC_1$). Donc on a $C_1I = CI$.

Or le triangle BC_1I donne

$$BC_1 < BI + IC_1.$$

Donc aussi

$$BC_1 < BI + IC,$$

ou

$$BC_1 < BC.$$

C. Q. F. D.

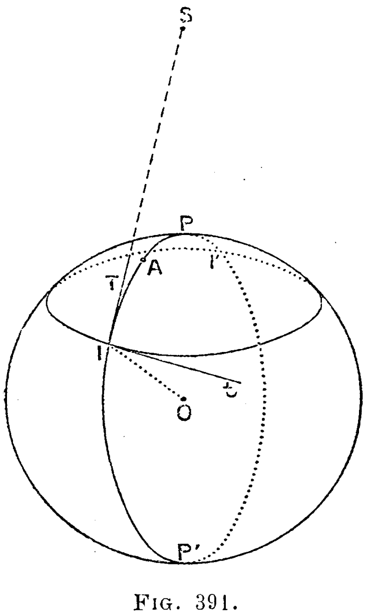

Fig. 391.

477. Théorème. — *La condition nécessaire et suffisante pour qu'un grand cercle et un cercle quelconque se coupent à angle droit est que le premier contienne les pôles du second.*

Soient I un point commun aux deux cercles; IT, I*t*, les tangentes au grand et au petit cercle en ce point ; P, P' les pôles du petit cercle; O, le centre de la sphère (*fig.* 391).

La condition indiquée dans l'énoncé est suffisante : car, si le grand cercle passe par P, P', son plan contient deux droites non parallèles entre elles et toutes deux perpendiculaires à I*t* : à savoir, le diamètre PP′ et le rayon OI. Ce plan, et par suite la tangente IT, sont donc perpendiculaires à I*t*.

---

(1) Les dispositions sont inverses, puisque les deux triangles sont de part et d'autre du côté commun AI.

La condition est d'ailleurs nécessaire : car, si les deux cercles se coupent à angle droit, le plan du grand cercle contient les deux droites IT, OI, perpendiculaires à I*t*. Il est donc lui-même perpendiculaire à cette droite, par suite au plan du petit cercle et contient, dès lors, le diamètre PP′, perpendiculaire à ce dernier plan menée par le point O.

Corollaire. — *Par un point pris sur une sphère, on peut mener un grand cercle perpendiculaire à un cercle donné de cette sphère: et on n'en peut mener qu'un seul, si le point donné n'est pas un pôle du cercle donné.*

Le grand cercle répondant à la question, sera déterminé par le point donné A et les pôles P, P′ du cercle cherché.

On remarquera qu'il existe deux *arcs* de grand cercle issus du point A et perpendiculaires au cercle donné; à savoir, ceux qui aboutissent aux deux points I, I′ où ce cercle est coupé par le grand cercle dont l'existence vient d'être établie (1).

477 *bis*. Le théorème qui précède est un cas particulier du suivant:

Théorème. — *La condition nécessaire et suffisante pour que deux cercles de la sphère se coupent à angle droit, est que le plan de l'un d'eux passe par le sommet du cône (ou du cylindre) circonscrit suivant l'autre.*

Tout d'abord, dans le cas où le premier cercle est un grand cercle, cette proposition résulte de la démonstration précédente: car, dans celle-ci, la droite IT, tangente au grand cercle IPP′, passe (**452**) par le sommet S du cône circonscrit suivant le petit cercle considéré.

Mais tout cercle qui coupe celui-ci à angle droit en I doit être tangent à IT; car IT est la seule tangente à la sphère en I qui soit perpendiculaire à I*t*. Donc le plan d'un tel cercle devra passer par le point S.

Inversement, tout cercle passant en I et dont le plan passe par S est tangent à IT et, par conséquent, la condition est aussi suffisante.

478. Théorème. — *Si, par un point d'une sphère, on mène les deux arcs de grand cercle perpendiculaires à un cercle donné et divers arcs de grands cercles obliques au même cercle, les arcs perpendiculaires sont, l'un plus court, l'autre plus long que tous les arcs obliques.*

---

(1) Nous considérons exclusivement, ici, les arcs issus du point A et terminés à leur *premier* point d'intersection avec le cercle donné. Si l'on ne tenait pas compte de cette restriction, le nombre des arcs perpendiculaires serait supérieur à deux : par exemple, l'arc AP′I′ (fig. 391) répondrait encore à la question.

Un arc oblique est d'autant plus long que son extrémité est plus éloignée de celle de l'arc perpendiculaire le plus court.

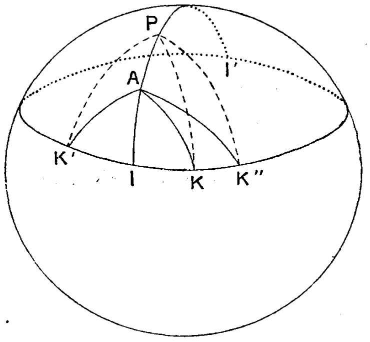

FIG. 392.

Soient A le point donné; P, celui des pôles du cercle donné qui est, par rapport à ce cercle, du même côté que le point A; AI, AI', les deux arcs de grand cercle perpendiculaires, dont le second est celui qui contient le point P à son intérieur; AK, AK', AK'' les divers arcs obliques (*fig.* 392).

1° L'arc AK est supérieur à AI, mais inférieur à AI'. Car si nous menons l'arc de grand cercle PK, le triangle sphérique APK donne

$$AK > PK - PA, \quad AK < PK + PA,$$

alors que l'on a

$$PK - PA = PI - PA = AI$$
$$PK + PA = PI' + PA = AI';$$

2° Supposons que les points K, K' du cercle donné soient tels que les arcs IK, IK' soient égaux. Il en est alors de même des cordes de ces arcs et le point I est également distant des deux points K, K'. Comme le point P jouit de la même propriété, le lieu des points de la sphère également distants de K et de K' est le grand cercle PI. Celui-ci contenant le point P, les cordes AK, AK' sont bien égales, et aussi, par conséquent, les arcs de grands cercles correspondants.

3° Soit maintenant, sur le cercle donné, un point K'' tel que $IK'' > IK$. Nous pouvons d'ailleurs supposer (¹), moyennant ce qui vient d'être établi (2°), que les deux points K, K'' sont du même côté du point I. Menons les arcs de grands cercles PK, PK''. Le point K étant intérieur à l'angle $\widehat{K''PI}$, on a $\widehat{KPI} < \widehat{K''PI}$.

Les triangles sphériques APK, APK'' ont donc un angle inégal (en A) compris entre côtés égaux chacun à chacun, et l'on a bien $AK < AK''$.

C. Q. F. D.

---

(1) Comparer Pl. **29**.

L'arc de grand cercle AI est dit la *distance sphérique* du point A au cercle KIK'I'. Cette dénomination est justifiée par le théorème qui précède.

479. Outre les constructions que nous avons précédemment appris à réaliser sur la sphère (n° **470**-**470** *bis*), on peut effectuer, sur cette surface, la plupart des constructions analogues à celles que nous avons indiquées en géométrie plane (liv. II, chap. VI). Nous donnons, à cet égard, un certain nombre d'exemples aux exercices (exercice 702).

EXERCICES

695. Si on joint un point pris dans l'intérieur d'un triangle sphérique aux trois sommets par des arcs de grands cercles, la somme de ces arcs est plus petite que le périmètre du triangle et plus grande que la moitié de ce périmètre.

696. Si deux triangles sphériques rectangles sont tels que les côtés de l'angle droit du premier soient respectivement plus petits que les côtés de l'angle droit du second, l'hypoténuse du premier est plus petite que celle du second.

697. Si la médiane d'un triangle sphérique (arc de grand cercle qui joint un sommet au milieu du côté opposé) est égale à un quadrant, elle est en même temps bissectrice de l'angle qui la comprend.

Si elle est plus petite qu'un quadrant, elle est plus grande que la bissectrice de l'angle en question, limitée au troisième côté. Elle est, au contraire, plus petite que cette bissectrice si elle est plus grande qu'un quadrant.

698. Si, dans un quadrilatère sphérique convexe, les côtés opposés sont égaux, les diagonales se coupent mutuellement en parties égales. Leur point d'intersection est le pôle du grand cercle qui passe par les points d'intersection des côtés opposés (Utiliser liv. VII). Deux sommets consécutifs et les points diamétralement opposés aux deux autres sommets sont sur un même cercle.

Dans un quadrilatère (*losange sphérique*) dont les quatre côtés sont égaux, les diagonales sont, en outre, perpendiculaires entre elles.

Si, les côtés opposés étant égaux, les diagonales sont égales, leur point d'intersection et les points d'intersection des côtés opposés sont les sommets d'un triangle trirectangle. Les quatre angles du quadrilatère sont alors égaux entre eux.

699. Si un grand cercle varie en restant constamment tangent à un petit cercle fixe C, et qu'on prenne, parmi les deux pôles du grand cercle, celui qui est situé dans le même hémisphère que C, ce pôle décrit un petit cercle C', qui est dit *polaire* du premier.

La relation entre C et C' est réciproque, c'est-à-dire que le polaire de C' n'est autre que C.

700. Les grands cercles bissecteurs des angles d'un triangle sphérique se coupent aux deux mêmes points diamétralement opposés. Il en est de même des trois médianes et des trois hauteurs (grands cercles menés par chaque sommet perpendiculairement au côté opposé).

701. Connaissant la distance des pôles de deux cercles et leurs rayons sphériques, trouver les conditions pour que ces cercles se coupent. Examiner successivement le cas où les rayons sont plus petits qu'un quadrant, de sorte que les pôles donnés sont intérieurs, et le cas où cette restriction n'est pas donnée.

702. Résoudre *effectivement* (voir page 59, note), à l'aide de constructions sphériques et, au besoin, de constructions planes, les problèmes suivants :

a) Trouver le point diamétralement opposé à un point donné ;

b) Faire passer un grand cercle par deux points donnés ;

c) Faire passer un cercle par trois points donnés de la sphère. Trouver les pôles d'un cercle donné ;

d) Mener, d'un point donné, un grand cercle perpendiculaire sur un cercle donné ;

e) Diviser en deux parties égales l'angle de deux grands cercles. Diviser en deux parties égales un arc de cercle quelconque.

f) Construire un triangle sphérique connaissant les trois côtés; discussion (ex. 701) ;

g) Par un point donné de la sphère, mener un grand cercle faisant avec un grand cercle donné un angle égal à un angle donné. Minimum de cet angle ;

Construire un triangle sphérique :

h) connaissant deux côtés et l'angle compris ;

i) connaissant deux côtés et l'angle opposé à l'un d'eux ;

j) connaissant un côté et deux angles (1) (discussion);

k) connaissant les trois angles;

l) Construire un triangle sphérique rectangle, connaissant l'hypoténuse et un angle adjacent ou un côté ;

m) Tracer un cercle tangent à un cercle donné en un point donné et passant par un autre point donné.

n) Tracer le grand cercle tangent à un cercle donné en un point donné de ce cercle ;

o) Mener, d'un point quelconque de la sphère, un grand cercle tangent à un petit cercle donné (pour cette construction et la suivante, utiliser ex. 699). Discussion;

p) Mener un grand cercle tangent à deux petits cercles donnés. Discussion ;

q) Mener, d'un point donné comme pôle, un cercle qui coupe un cercle donné à angle droit ;

r) Construire un petit cercle tangent à trois grands cercles donnés ;

s) Construire un cercle tangent à un cercle donné en un point donné et tangent à un autre cercle donné.

703. Tracer un grand cercle sur lequel deux petits cercles donnés interceptent des arcs égaux à des arcs donnés.

704. Étant donnés, sur une sphère, un grand cercle C et, sur ce grand cercle un point A, placer (à l'aide de constructions planes et sphériques) le point qui, lorsqu'on considère la sphère donnée comme sphère céleste, le grand cercle donné comme équateur et le point A comme origine des ascensions droites, a une ascension droite et une déclinaison (2) données.

705. Étant donnés trois cercles sur la sphère, il existe un cercle et un seul qui partage chacun d'eux en deux parties égales. Quel est le plan de ce cercle ?

---

(1) On remarquera que, si les angles donnés ne sont pas tous deux adjacents au côté donné, on ne peut pas procéder comme en Géométrie plane (**85**, constr. 8) : il faut recourir à la considération des triangles polaires (**474**), comme pour la construction suivante.

(2) TISSERAND et ANDOYER, *Leçons de Cosmographie*, n^{os} **9-12**.

706. Étant donnés deux grands cercles tangents à un même petit cercle C, un troisième grand cercle quelconque tangent à C forme avec les deux premiers un triangle sphérique de périmètre constant.

707. Construire (effectivement) un triangle sphérique connaissant un côté, un angle adjacent et la somme des deux autres côtés (donnée par un arc de grand cercle égal à cette somme).

708. Construire (effectivement) un triangle sphérique, connaissant un angle, une hauteur (ex. 700) et le périmètre. (Utiliser ex. 706 ; deux cas à distinguer).

709. Trouver les conditions pour qu'un quadrilatère sphérique soit circonscriptible à un cercle. (Comparer ex. 652.)

710. D'un point de la sphère, extérieur à un petit cercle donné, on lui mène les deux grands cercles tangents (ex. 702, *o*); pour quelle position du point ces deux grands cercles font-ils le plus petit angle ?

711. Le grand cercle qui passe par les milieux M, N des côtés AB, AC d'un triangle sphérique coupe le grand cercle BC aux milieux des deux arcs qui ont pour extrémités le point B et le point diamétralement opposé à C. Le pôle du grand cercle MN est également distant de B et de C et les arcs de grands cercles qui le joignent à B et à C font entre eux un angle double de l'angle au centre correspondant à l'arc de grand cercle MN.

CHAPITRE V

SURFACE ET VOLUME DE LA SPHÈRE

480. Théorème. — *La surface engendrée par un segment de droite tournant autour d'un axe situé avec lui dans un même plan et ne le traversant pas, a pour mesure la projection du segment sur l'axe, multipliée par la circonférence qui a son centre sur l'axe et touche le segment en son milieu* (1).

La surface engendrée par un segment de droite AB (*fig.* 394) tournant autour d'un axe xy situé avec lui dans un même plan et ne le traversant pas, est en général, celle d'un tronc de cône dont les circonférences de bases ont pour rayons les perpendiculaires Aa, Bb abaissées des points A, B sur l'axe.

Exceptionnellement, le tronc de cône se réduit à un cône, lorsque le segment a une extrémité sur l'axe (*fig.* 393), ou à un cylindre, lorsque le segment est parallèle à l'axe (*fig.* 393 *bis*) (2).

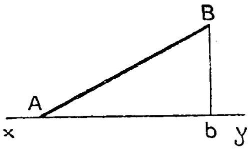

Fig. 393.

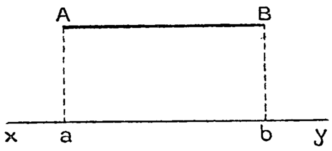

Fig. 393 *bis*.

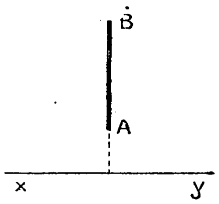

Fig. 393 *ter*.

Soient M le milieu de AB ; Mm, la perpendiculaire abaissée de ce point sur l'axe. La surface latérale du tronc de cône a pour mesure (**445**) le produit 2π AB. M*m*; et cette expression est encore valable lorsque le tronc de cône est remplacé, soit par un cône, soit par un cylindre (**445**, Rem.).

---

(1) L'expression ainsi obtenue cesse d'avoir un sens lorsque le segment est perpendiculaire à l'axe (fig. 393 *ter*). Cette circonstance ne se présentera pas dans les raisonnements qui suivront.

(2) Dans le cas, exclu par la note précédente, où le segment donné serait perpendiculaire à l'axe, on aurait une couronne circulaire ou un cercle.

Par le point M, menons la perpendiculaire à AB, jusqu'à rencontre en O avec l'axe. D'autre part, par le point A, menons la parallèle AH à l'axe, jusqu'à rencontre en H avec B*b*. Le segment AH est égal à *ab*, c'est-à-dire à la projection de AB sur l'axe. D'autre part, les triangles ABH, M*m*O sont semblables, comme ayant leurs côtés perpendiculaires, et donnent

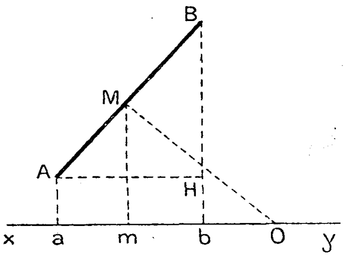

Fig. 394.

$$\frac{AB}{OM} = \frac{AH}{Mm},$$

ce qui s'écrit

$$AB \cdot Mm = OM \cdot AH.$$

Donc

$$\text{Surf. } AB = 2\pi\ AB.\ Mm = 2\pi\ OM \cdot AH = 2\pi\ OM.\ ab.$$

C. Q. F. D.

481. Aire de la zone.

Définition. — On nomme *zone*, la portion de surface sphérique comprise entre deux plans parallèles (*fig.* 395). Les circonférences situées dans ces deux plans et qui limitent, par conséquent, la zone, en sont dites les *bases*. La distance des deux plans de bases est la *hauteur* de la zone.

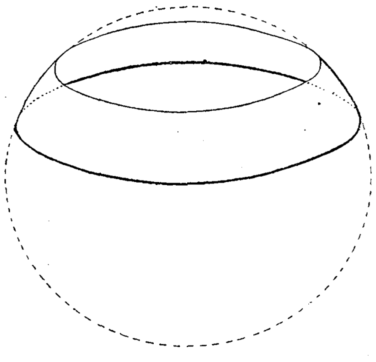

Fig. 395.

L'une quelconque des deux portions que l'on détermine dans une surface sphérique en la coupant par un plan, est dite une *calotte sphérique*. On peut évidemment considérer la calotte sphérique comme une zone dans laquelle un des plans de bases est tangent à la sphère.

La zone (ou la calotte) peut être encore définie comme la surface engendrée par la révolution d'un arc de cercle AB (*fig.* 396, 396 *bis*) autour d'un diamètre sans point commun avec lui (cas de la zone) ou passant par une de ses extrémités (cas de la calotte). La *hauteur* est alors la projection de l'arc AB sur l'axe.

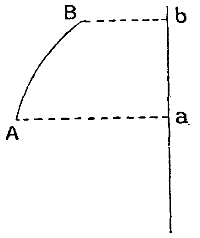

Fig. 396.

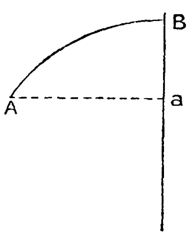

Fig. 396 *bis*.

482. Pour définir l'*aire* de la zone, nous remplacerons d'abord l'arc AB par une ligne brisée (ACDEB, *fig*. 397) inscrite dans cet arc. L'aire de la zone sera, par définition, la limite vers laquelle tend l'aire engendrée par cette ligne en tournant autour de l'axe, lorsque le nombre des côtés de la ligne augmente indéfiniment de manière que chacun d'eux tende vers zéro.

L'existence de cette limite et son expression résultent des deux théorèmes suivants :

Théorème. — *L'aire engendrée par une ligne brisée inscrite à un cercle, en tournant autour d'un diamètre qui ne la coupe pas, est égale à la projection de la ligne brisée sur le diamètre, multipliée par la longueur d'une circonférence dont le rayon est compris entre la plus petite et la plus grande des distances du centre aux côtés de la ligne brisée.*

Soit la ligne brisée ACDEB (*fig*. 397) inscrite dans un cercle de centre O et tournant autour du diamètre xy, lequel ne la traverse pas (une des extrémités de la ligne brisée, ou les deux, pouvant être sur xy). Soient a, c, d, e, b, les projections des sommets sur xy.

Les perpendiculaires aux milieux H, K, L, M des côtés de la ligne brisée concourant toutes en O, le théorème du n° **480** donne ici (¹) :

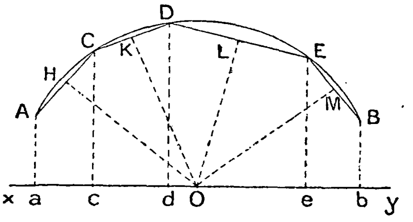

Fig. 397.

$$\text{Surf. AC} = 2\pi\,\overline{ac}\cdot\overline{OH}$$
$$\text{Surf. CD} = 2\pi\,\overline{cd}\cdot\overline{OK}$$
$$\text{Surf. DE} = 2\pi\,\overline{de}\cdot\overline{OL}$$
$$\text{Surf. EB} = 2\pi\,\overline{eb}\cdot\overline{OM}$$

---

(1) Le cas d'exception signalé à la page 169, note 2, ne peut se présenter dans le raisonnement actuel, une corde d'un cercle ne pouvant être perpendiculaire à un diamètre sans être traversée par lui.

La surface engendrée par la ligne ACDEB est donc mesurée par

$$2\pi\,(\overline{ac}\cdot\overline{\mathrm{OH}}+\overline{cd}\cdot\overline{\mathrm{OK}}+\overline{de}\cdot\overline{\mathrm{OL}}+\overline{eb}\cdot\overline{\mathrm{OM}}).$$

Or, d'après un théorème connu d'arithmétique (¹), l'expression entre parenthèses est égale au produit de la somme $ac + cd + de + eb$, c'est-à-dire de ab, par une quantité comprise entre la plus grande et la plus petite des quantités OH, OK, OL, OM : ce qui démontre le théorème.

Corollaire. — Si la ligne brisée est régulière, les distances OH, OK, OL, OM, sont toutes égales à l'apothème de cette ligne brisée.

Donc l'*aire engendrée par une ligne brisée régulière tournant autour d'un axe situé dans son plan*, *passant par son centre et ne la traversant pas*, *est égale à la projection de cette ligne sur l'axe*, *multipliée par la longueur de la circonférence inscrite.*

483. Théorème. — *L'aire engendrée par un arc de cercle tournant autour d'un diamètre qui ne le traverse pas*, *est égale à la projection de l'arc sur le diamètre*, *multipliée par la longueur de la circonférence entière.*

Autrement dit, l'*aire de la zone est égale au produit de sa hauteur par la circonférence d'un grand cercle.*

Soient, en effet, l'arc AB, de centre O, tournant autour du diamètre xy ; ab, la projection de AB sur xy. Si, dans l'arc AB, nous inscrivons une ligne brisée ACDEB, celle-ci, en tournant autour de xy, engendrera une aire égale au produit de $2\pi\cdot\overline{ab}$ par une quantité intermédiaire entre la plus grande et la plus petite des distances du centre aux côtés de la ligne brisée.

Si maintenant on augmente indéfiniment le nombre des côtés, de manière que chacun d'eux tende vers zéro, toutes les distances dont il vient d'être question tendent vers le rayon $\mathrm{OA} = \mathrm{R}$ de la circonférence.

Donc l'aire engendrée a bien une limite, indépendante de la loi suivant laquelle on fait croître le nombre des côtés de la ligne brisée (pourvu que tous ces côtés tendent vers zéro) et cette limite est

$$2\pi\,\mathrm{R}\cdot\overline{ab}.$$

C. Q. F. D.

---

(1) TANNERY, *Leçons d'Arithmétique*, nº **212**, page 180.

Remarque. — Le raisonnement précédent et sa conclusion sont applicables au cas d'une calotte sphérique.

Corollaire. — On voit, par le théorème précédent, que *deux zones d'une même sphère sont proportionnelles à leurs hauteurs.*

484. Aire de la sphère.

Théorème. — *L'aire de la sphère de rayon* R *est* $4\pi R^2$.

En effet, le raisonnement du numéro précédent est encore applicable lorsque l'arc AB devient la demi-circonférence. La projection *ab* est alors le diamètre 2R ; la zone engendrée n'est autre que la sphère entière. La surface de cette sphère est donc mesurée par le produit

$$2\pi R \times 2R = 4\pi R^2.$$

C. Q. F. D.

Corollaire. — *La surface de la sphère est égale à quatre fois la surface d'un grand cercle.*

Remarque. — On voit que *les surfaces de deux sphères sont proportionnelles aux carrés de leurs rayons.*

485. Théorème. — *Le volume engendré par un triangle tournant autour d'un axe situé dans son plan, passant par un de ses sommets et ne le traversant pas, a pour mesure le produit de la surface que décrit le côté opposé au sommet situé sur l'axe, par le tiers de la hauteur correspondante.*

Nous distinguerons trois cas :

1° *Un des côtés du triangle est situé sur l'axe.*

Soit le triangle ABC (*fig.* 398, 398 *bis*) dont le côté AB est situé sur l'axe *xy*. Par le sommet C, menons, sur cet axe, la perpendiculaire C*c*. Les deux triangles rectangles AC*c*, BC*c* engendreront,

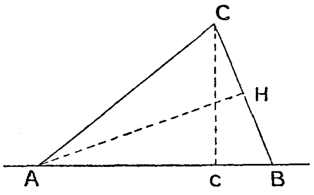

Fig. 398.

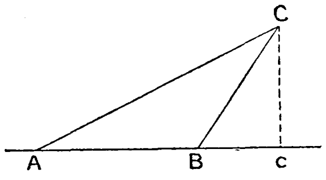

Fig. 398 *bis*.

en tournant autour de xy, deux cônes ayant pour base commune le cercle de rayon cC, pour hauteurs respectives Ac, Bc, et dont les volumes seront respectivement $\frac{1}{3}\pi\overline{Cc}^2\cdot\overline{Ac}$, $\frac{1}{3}\pi\overline{Cc}^2\cdot\overline{Bc}$. Le volume engendré par ABC sera la somme ou la différence des deux cônes ainsi obtenus, suivant que le point c sera sur le segment AB (*fig.* 398) ou sur un de ses prolongements, par exemple au delà du point B (*fig.* 398 *bis*) : on aura donc

$$\text{Vol. ABC} = \frac{1}{3}\pi\cdot\overline{Cc}^2\cdot\overline{Ac} + \frac{1}{3}\pi\overline{Cc}^2\cdot\overline{BC} = \frac{1}{3}\overline{Cc}^2(\overline{Ac}+\overline{Bc})$$

dans le premier cas, et

$$\text{Vol. ABC} = \frac{1}{3}\pi\overline{Cc}^2\cdot\overline{Ac} - \frac{1}{3}\pi\overline{Cc}^2\cdot\overline{Bc} = \frac{1}{3}\pi\overline{Cc}^2(\overline{Ac}-\overline{Bc})$$

dans le second.

Mais comme on a, dans le premier cas,

$$\text{AB} = \text{A}c + \text{B}c$$

et, dans le second,

$$\text{AB} = \text{A}c - \text{B}c,$$

il vient, en toute hypothèse,

$$\text{Vol. ABC} = \frac{1}{3}\pi\overline{Cc}^2\cdot\text{AB}.$$

D'autre part, si AH est la hauteur du triangle issue du point A, on a

$$\text{C}c\cdot\text{AB} = \text{BC}\cdot\text{AH} :$$

car ces deux produits représentent tous deux le double de l'aire du triangle ABC. On peut donc écrire

$$\text{Vol. ABC} = \frac{1}{3}\pi\overline{Cc}\cdot\text{BC}\cdot\text{AH},$$

ce qui donne

$$\text{Vol. ABC} = \frac{1}{3}\text{AH}\cdot\text{Surf. BC},$$

car la surface décrite par le côté BC est celle d'un cône et a pour mesure (**443**) $\pi\ \overline{BC} \cdot \overline{Cc}$.

2° *Le côté opposé au sommet situé sur l'axe est parallèle à l'axe.*

Soit le triangle ABC (*fig.* 399, 399 *bis*) dans lequel le sommet A est situé sur l'axe xy, tandis que le côté BC est parallèle à cet axe. Menons encore la hauteur AH et projetons les points B, C en b, c sur l'axe. Lorsqu'on fait tourner la figure autour de cet axe, le rectangle BbAH engendre un cylindre et le triangle ABb engendre un cône ; comme ces deux solides ont même base (le cercle de rayon bB) et même hauteur (Ab), le cône est le tiers du cylindre : leur différence, c'est-à-dire le volume engendré par le triangle ABH, est les $\frac{2}{3}$ du cylindre engendré par le rectangle AbBH.

De même, le volume engendré par le triangle ACH est les $\frac{2}{3}$ du cylindre engendré par AcCH.

Donc le volume engendré par le triangle ABC, lequel est la somme (*fig.* 399) ou la différence (*fig.* 399 *bis*) des volumes engendrés par les deux triangles précédents, est les $\frac{2}{3}$ du cylindre engendré par le rectangle BbCc, ce cylindre étant la somme ou la différence des deux cylindres que nous venons de considérer. Autrement dit, on a

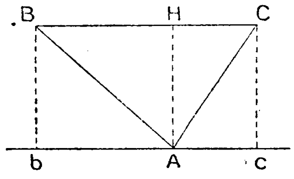

FIG. 399.

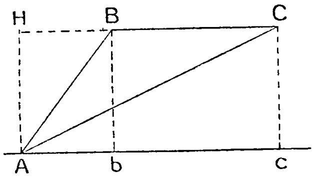

FIG. 399 *bis*.

$$\text{Vol. ABC} = \frac{2}{3}\,\text{Vol. B}b\text{C}c = \frac{2}{3}\,\pi\,\overline{Bb}^2 \cdot \overline{BC}$$

$$= \frac{1}{3}\,\text{AH} \times 2\pi\,\text{B}b \cdot \text{BC} = \frac{1}{3}\,\text{AH} \times \text{surf. BC}.$$

3° *Cas général.*

Soit le triangle ABC (*fig.* 400) dont le sommet A est sur l'axe xy, le côté BC n'étant pas parallèle à cet axe. Soit D le point de rencontre de xy avec BC prolongé. On a (1°), en appelant toujours AH la hauteur issue de A,

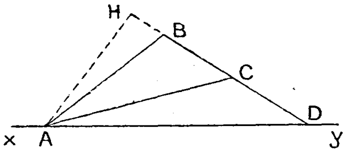

Fig. 400.

$$\text{Vol. ABD} = \text{surf. BD} \times \frac{1}{3}\,\text{AH},$$

$$\text{Vol. ACD} = \text{surf. CD} \times \frac{1}{3}\,\text{AH},$$

d'où, par différence,

$$\text{Vol. ABC} = \text{surf. BC} \times \frac{1}{3}\,\text{AH}.$$

C. Q. F. D.

486. Volume du secteur sphérique.

Définition. — On nomme *secteur sphérique*, la figure engendrée par un secteur circulaire tournant autour d'un diamètre qui ne le traverse pas. Dans ce mouvement, l'arc qui sert de base au secteur circulaire engendre une zone, qui est dite *base* du secteur sphérique.

Pour définir le volume du secteur sphérique, nous remplacerons le secteur circulaire par un secteur polygonal inscrit. Le volume du secteur sphérique sera, par définition, la limite vers laquelle tend le volume engendré par ce secteur polygonal, lorsque le nombre des côtés de la ligne brisée qui lui sert de base augmente indéfiniment, de manière que chacun d'eux tende vers zéro.

Les théorèmes suivants montrent l'existence de cette limite et en fournissent l'expression, laquelle est indépendante de la loi suivant laquelle on construit les lignes brisées inscrites.

Théorème. — *Le volume engendré par un secteur polygonal* (Pl., **253** *bis*) *en tournant autour d'un axe situé dans son plan, passant par son centre et ne le traversant pas, est égal au tiers de la surface engendrée par la ligne brisée qui sert de base au secteur, multiplié par une quantité intermédiaire entre la plus grande et la plus petite des distances du centre aux côtés de cette ligne.*

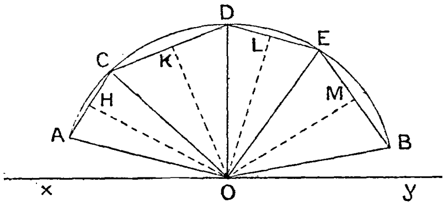

Fig. 401.

Soit le secteur polygonal OACDEB (*fig.* 401), tournant autour d'un axe xy passant par son centre O, situé dans son plan et ne le traversant pas. Les triangles OAC, OCD, ODE, OEB engendrent respectivement des volumes dont l'expression est fournie par le théorème précédent. On a ainsi, en désignant par OH, OK, OL, OM, les distances du centre aux côtés AC, DC, DE, EB :

$$\text{Vol. OAC} = \frac{1}{3}\ \text{surf. AC} \times \text{OH}$$

$$\text{Vol. OCD} = \frac{1}{3}\ \text{surf. CD} \times \text{OK}$$

$$\text{Vol. ODE} = \frac{1}{3}\ \text{surf. DE} \times \text{OL}$$

$$\text{Vol. OEB} = \frac{1}{3}\ \text{surf. EB} \times \text{OM}.$$

En faisant la somme et recourant encore au théorème d'arithmétique déjà appliqué n° **482**, on voit bien que le volume engendré par le secteur est égal au tiers de la somme surf. AC + surf. CD + surf. DE + surf. EB, — c'est-à-dire au tiers de la surface qu'engendre la ligne ACDEB, — multiplié par une quantité intermédiaire entre la plus grande et la plus petite des quantités OH, OK, OL, OM.

C. Q. F. D.

Corollaire. — Si le secteur polygonal est régulier, les lignes OH, OK, OL, OM sont toutes égales à l'apothème de la ligne brisée régulière ACDEB.

Donc *le volume engendré par un secteur polygonal régulier tournant autour d'un axe situé dans son plan, passant par son centre et ne le traversant pas, est égal à la surface engendrée par la ligne brisée qui sert de base au secteur, multipliée par le tiers de l'apothème.*

Théorème. — *Le volume du secteur sphérique est égal à la zone qui lui sert de base, multipliée par le tiers du rayon.*

En effet, si l'on augmente indéfiniment le nombre des côtés d'une ligne brisée inscrite dans l'arc AB (*fig.* 401) qui engendre la zone de base, — et cela de manière que chacun d'eux cités tende vers zéro, — la surface engendrée par cette ligne brisée inscrite tend vers la surface de zone, pendant que les distances OH, OK, etc. (*fig.* 401), des différents côtés au centre tendent toutes vers le rayon. Donc le volume engendré par le secteur polygonal correspondant tend bien vers la limite indiquée dans l'énoncé.

Corollaire. — *Si* R *est le rayon de la sphère, h la hauteur de la zone qui sert de base au secteur, le volume de celui-ci est* $\frac{1}{3}\pi R^2 h$.

Cette expression est bien, en effet, le produit de $\frac{1}{3}$ R par l'aire de la zone, trouvée au nº **483**.

487. Volume de la sphère.

Théorème. — *Le volume de la sphère de rayon* R *est* $\frac{4}{3}\pi R^3$.

En effet, les raisonnements qui précèdent sont applicables à la figure engendrée par un demi-cercle tournant autour de son diamètre, c'est-à-dire au volume de la sphère. On doit alors, dans l'expression indiquée au corollaire du numéro précédent, remplacer la hauteur h par 2R; ce qui conduit bien au résultat annoncé.

Corollaire. — *Le volume de la sphère de diamètre* D *est* $\frac{1}{6}\pi D^3$.

Il suffit, pour le voir, de remplacer R, par $\frac{D}{2}$, dans l'expression $\frac{4}{3}\pi R^3$.

Remarque. — On voit que *les volumes des deux sphères sont proportionnels aux cubes de leurs rayons* (*ou de leurs diamètres*).

488. Volume de l'anneau sphérique.

On nomme *anneau sphérique* le solide engendré par la révolution d'un segment de cercle (Pl., **263**) autour d'un diamètre qui ne le traverse pas.

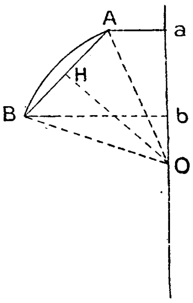

Fig. 402.

Le volume engendré par un segment de cercle tournant autour d'un diamètre qui ne le traverse pas (volume de l'*anneau sphérique*) *est le sixième du volume du cylindre qui a pour base la corde du segment et pour hauteur la projection de cette corde sur l'axe.*

Soit le segment de cercle compris entre l'arc AB d'une circonférence O (*fig.* 402) et sa corde, tournant autour de l'axe xy qui passe par O. Soit encore ab la projection AB sur xy.

Lorsque la figure tourne autour de xy, le secteur circulaire OAB engendre un volume qui est mesuré (**486**) par le produit $\frac{2}{3}\pi\overline{OA}^2\,\overline{ab}$.

D'autre part, le triangle rectiligne OAB engendre, dans les mêmes conditions, un volume égal (**485**) à surf. AB $\times \frac{1}{3}$ OH (en désignant par OH la distance du centre à la corde AB), c'est-à-dire (**480**) à

$$2\pi\,\overline{ab}.\,\overline{OH} \times \frac{1}{3}\,\overline{OH} = \frac{2}{3}\pi\,\overline{OH}^2.\,\overline{ab}.$$

La différence des deux volumes précédents est évidemment constituée par le volume de l'anneau sphérique : celui-ci a donc pour mesure

$$\frac{2}{3}\pi\,\overline{OA}^2.\overline{ab} - \frac{2}{3}\pi\,\overline{OH}^2.\overline{ab} = \frac{2}{3}\pi\,\overline{ab}\,(\overline{OA}^2 - \overline{OH}^2).$$

Mais, dans le triangle OAH, on a

$$\overline{OA}^2 - \overline{OH}^2 = \overline{AH}^2 = \frac{\overline{AB}^2}{4}.$$

Donc

$$\text{Vol. anneau AB} = \frac{2}{3}\pi\,\overline{ab}\,\frac{\overline{AB}^2}{4} = \frac{1}{6}\pi\,\overline{AB}^2.\,\overline{ab}.$$

C. Q. F. D.

489. Volume du segment sphérique.

On nomme *segment sphérique* la portion de sphère comprise entre deux plans parallèles : volume limité, en général, par une zone et deux cercles dont les plans sont parallèles entre eux (*fig.* 403), ces cercles étant dits les *bases* du segment. Le segment sphérique peut aussi être à une seule base, c'est-à-dire limité par un cercle et une calotte sphérique (*fig.* 403 *bis*); le plan de l'autre base doit être

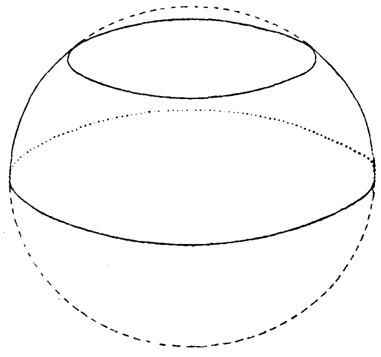
FIG. 403.

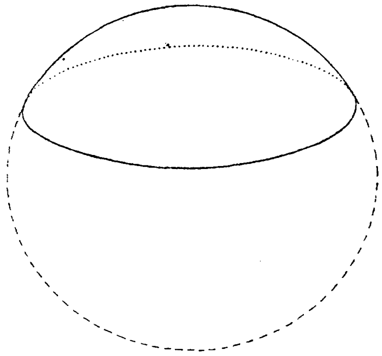
FIG. 403 *bis*.

alors, pour l'application des raisonnements, réputé tangent à la sphère.

La *hauteur* du segment est la distance des plans des deux bases.

Le segment sphérique est équivalent à la demi-somme des deux cylindres qui ont pour hauteur commune, la hauteur du segment et pour bases respectives les deux bases de ce segment, augmentée de la sphère qui a la hauteur pour diamètre.

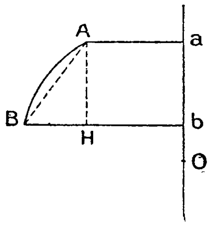

Fig. 404.

Soit le segment sphérique engendré par la révolution du trapèze mixtiligne $aABb$ (*fig.* 404) autour de l'axe ab et limité, par conséquent, d'une part par deux cercles de rayons respectifs aA, bB, d'autre part par la zone qu'engendre l'arc AB. Ce solide est manifestement la somme du tronc de cône engendré par le trapèze *rectiligne* $AaBb$ et de l'anneau sphérique AB.

Le tronc de cône a pour volume

$$\frac{1}{3}\pi\overline{ab}\left(\overline{Aa}^2+\overline{Bb}^2+\overline{Aa}\cdot\overline{Bb}\right);$$

l'anneau sphérique a pour volume

$$\frac{1}{6}\pi\overline{ab}\cdot\overline{AB}^2.$$

Nous trouvons donc comme somme

$$\frac{1}{6}\pi\overline{ab}\left(\overline{AB}^2+2\overline{Aa}^2+2\overline{Bb}^2+2\overline{Aa}\cdot\overline{Bb}\right).$$

Mais $\overline{AB}^2$ peut s'exprimer à l'aide de Aa, Bb et ab ; car, si l'on mène par le point A la parallèle AH à l'axe, jusqu'à rencontre en H avec Bb, le triangle rectangle ABH donne

$$\overline{AB}^2=\overline{AH}^2+\overline{BH}^2,$$

et comme, d'autre part, on a (¹)

$$AH=ab;\quad BH=Bb-Hb=Bb-Aa,$$

(¹) Nous supposons, pour fixer les idées, que Bb soit plus grand que Aa : autrement dit, que l'on a désigné par B celle des extrémités de l'arc donné qui est le plus éloignée de l'axe.

il vient

$$\overline{AB}^2 = \overline{ab}^2 + (Bb - Aa)^2.$$

Le volume du segment est donc

$$\frac{1}{6}\pi\overline{ab}\left[\overline{ab}^2 + (Bb - Aa)^2 + 2\overline{Aa}^2 + 2\overline{Bb}^2 + 2\overline{Aa}\cdot\overline{Bb}\right]$$

$$= \frac{1}{6}\pi\overline{ab}\left(\overline{ab}^2 + \overline{Aa}^2 - 2\overline{Aa}\cdot\overline{Bb} + \overline{Bb}^2 + 2\overline{Aa}^2 + 2\overline{Aa}\cdot\overline{Bb} + 2\overline{Bb}^2\right),$$

ou, toutes réductions faites,

$$\frac{1}{6}\pi\overline{ab}^3 + \frac{1}{2}\pi\overline{ab}\left(\overline{Aa}^2 + \overline{Bb}^2\right),$$

ce qui est bien la somme de la sphère et des deux demi-cylindres indiqués dans l'énoncé.

EXERCICES

712. Partager la surface d'une sphère en parties équivalentes par des plans passant par une droite donnée, extérieure à la sphère.

713. Une sphère varie en passant constamment par le centre d'une sphère fixe. Montrer que celle-ci intercepte sur la sphère variable une calotte d'aire constante.

714. Un cylindre est circonscrit à une sphère. Montrer que la zone de la sphère comprise entre deux plans perpendiculaires à l'axe du cylindre est équivalente à l'aire cylindrique comprise entre les deux mêmes plans.

715. Un cylindre étant circonscrit à une sphère, on mène à celle-ci un plan tangent P perpendiculaire à l'axe du cylindre et l'on prend la section droite ainsi déterminée dans ce dernier comme base d'un cône ayant pour sommet le centre de la sphère. Montrer :

1° Que si l'on coupe les trois solides par un même plan parallèle à P, le cercle, section du cylindre, a une aire équivalente à la somme des cercles suivant lesquels sont coupés respectivement le cône et la sphère.

2° Que si l'on coupe les trois solides par deux plans parallèles à P, le volume intercepté, entre ces deux plans, dans le cylindre, est la somme des volumes interceptés dans le cône et dans la sphère.

716. Montrer que les volumes du cylindre, du cône, du tronc de cône et du segment sphérique satisfont tous à la relation :

$$V = \frac{h}{6}\left(B + B' + 4B''\right),$$

h, B, B′, B″ désignant, comme à l'exercice 575, la hauteur du solide, les aires des deux bases (la base supérieure du cône étant nulle) et l'aire de la section menée, parallèlement aux bases, à égale distance de ces deux bases.

717. Deux petits cercles polaires l'un de l'autre (ex 699) sont tels que si l'on prenait pour unité de longueur la longueur d'un grand cercle et pour unité de surface l'aire de la demi-sphère, le nombre qui mesure la longueur de l'un serait égal à l'unité, moins le nombre qui mesure la calotte sphérique intérieure à l'autre.

718. Pour tous les polyèdres circonscrits à une même sphère, le rapport du volume à la surface est le même.

719. L'énoncé précédent subsiste encore pour un cylindre, un cône ou un tronc de cône circonscrits à la sphère (on dit qu'un cylindre, un cône ou un tronc de cône sont circonscrits à la sphère, si : 1° la surface cylindrique ou conique est circonscrite à cette sphère ; 2° les plans de bases sont tangents à la même sphère) ; ou plus généralement, pour tout solide limité par des portions de surfaces cylindriques ou coniques circonscrites et par des plans tangents.

720. Quel est le rapport des volumes engendrés par un parallélogramme tournant successivement autour de deux côtés adjacents !

721. Construire (effectivement) un triangle, connaissant les rayons des sphères équivalentes aux volumes qu'engendre ce triangle en tournant successivement autour de chacun de ses côtés.

722. Calculer le rayon d'un cercle tracé sur une sphère de rayon R, sachant que l'aire de ce cercle est égale à la différence des calottes qu'il détermine.

Quelle est la hauteur du cône circonscrit suivant ce cercle ?

723. Trouver, sur une sphère donnée, une calotte qui soit dans un rapport donné avec l'aire du cercle qui lui sert de base.

724. Couper une sphère par un plan de manière que l'un des segments sphériques (à une base) obtenus soit dans un rapport donné avec le secteur sphérique limité par la même calotte.

725. On prend un cercle d'une sphère comme base d'un cône qui a pour sommet l'un des pôles du cercle. Trouver le cercle de manière que le volume du cône soit dans un rapport donné avec le segment sphérique qui a pour base unique le même cercle et qui contient le cône.

Vers quelle limite tend le rapport du cône au segment sphérique, lorsque leur hauteur commune tend vers zéro?

726. Construire un segment sphérique appartenant à une sphère donnée, connaissant son volume (égal à celui d'une sphère de rayon donné) et l'aire de la zone qui le limite (égale à celle d'un cercle donné).

Quel est le maximum du volume du segment, lorsque l'aire de la zone est donnée?

727. Tracer, sur une sphère, trois cercles tangents entre eux deux à deux, ayant leurs plans parallèles à une même droite, dont les deux premiers soient égaux entre eux et divisent en quatre parties équivalentes l'une des calottes déterminées par le troisième.

728. Soient A le point de contact de deux circonférences O et O′ tangentes extérieurement ; T, T′ les points de contact d'une tangente commune extérieure à ces circonférences. Montrer que le volume compris entre le tronc de cône engendré par TT′ (dans sa révolution autour de la ligne des centres) et la sphère qui passe par les cercles de bases de ce tronc de cône est double de la portion de ce même tronc de cône

extérieure aux sphères S, S′, engendrées par les circonférences données, et que cette portion est équivalente à la somme des anneaux sphériques situés respectivement dans les sphères S, S′, à l'extérieur des cônes engendrés par les cordes AT, AT′. Exprimer ces volumes à l'aide des rayons des circonférences données.

PROBLÈMES

PROPOSÉS SUR LE LIVRE VIII

729. Lieu des centres des sphères qui coupent deux sphères données suivant des grands cercles. — Même problème lorsqu'il y a trois sphères données.

730. Lieu des centres des sphères qui sont coupées par deux (ou trois) sphères données suivant des grands cercles.

731. Si le centre radical de quatre sphères leur est intérieur, il est le centre d'une sphère qui est coupée par chacune des premières suivant un grand cercle.

732. Lieu des génératrices de contact des plans tangents menés par un point donné de l'espace aux cônes de révolution qui passent par les côtés d'un angle donné (se ramène à l'exerc. 681 par la considération des sphères inscrites aux cônes considérés).

Par deux points donnés A, B d'une sphère, on fait passer un petit cercle variable et, par un point C situé le grand cercle AB, on mène un grand cercle tangent à ce petit cercle. Quel est le lieu du point T de contact ? Montrer qu'on a $\overline{CT}^2 = \frac{CA.\ CB.\ R}{d}$, en désignant par R le rayon de la sphère et par d la distance du centre à la corde AB.

733. On considère deux droites rectangulaires entre elles, non situées dans un même plan, et on les coupe par un plan variable, mais parallèle à un plan fixe. Montrer que la sphère qui a pour points diamétralement opposés les deux points d'intersection passe par un cercle fixe.

734. Sur deux droites données D, D′ on prend, à partir de deux points fixes A, A′ deux segments AM, A′M′ dont le rapport $\frac{A'M'}{AM}$ soit égal à un nombre donné k, et on considère la sphère (ex. 665) lieu des points P tels que $\frac{PM'}{PM} = k$.

Montrer que si les points M, M′ varient sur leurs droites respectives (de manière que le rapport $\frac{A'M'}{AM}$ soit constamment égal à k), les sphères obtenues, comme il vient d'être dit, appartiennent à l'une ou à l'autre de deux séries (suivant les sens dans lesquels sont portés les segments AM, A′M′), les sphères de la première série passant toutes (¹) par un même cercle C_1 ; les sphères de la seconde série passant toutes par un même cercle C_2.

La direction du plan du cercle C_1, comme celle du plan du cercle C_2, est indépendante du choix des points A, A′ lorsque les droites D, D′ et le rapport k sont donnés. Lorsque k varie, ces plans restent parallèles à une droite fixe.

---

(1) Il est d'abord nécessaire de prouver que deux sphères de la même série se coupent. C'est ce que l'on fera aisément à l'aide de l'exercice 733.

Tout point appartenant à un cercle C_1 ou C_2 est tel que ses distances aux deux droites données soient entre elles dans le rapport k. Inversement, par tout point tel que le rapport de ses distances à D et à D′ soit égal à k, il passe un cercle C_1 et un cercle C_2.

735. Étant données deux sphères et un point P, le lieu d'une droite passant par P, joignant entre eux deux points pris respectivement sur les deux sphères et qui soit divisée par ces points et le point P dans un rapport donné, est un cône à base circulaire.

Ce cône peut se réduire à un plan. Trouver le lieu du point P pour qu'il en soit ainsi.

736. Le lieu des arêtes des dièdres droits dont les faces passent respectivement par deux droites concourantes données est un cône à base circulaire (ex. 462).

737. Le lieu des points tels que le rapport de leurs distances à deux droites concourantes données soit constant est un cône circulaire oblique (ex. 734).

738. Le lieu des bissectrices des angles dont un côté est fixe et dont l'autre côté décrit un plan fixe (en passant constamment par le point d'intersection de ce plan avec le premier côté) est un cône (généralement oblique) à base circulaire.

739. Lieu des sommets d'un angle polyèdre dont les quatre faces passent respectivement par les côtés d'un quadrilatère plan donné, sachant que cet angle polyèdre peut être coupé (ex. 500) suivant un rectangle.

Même problème, lorsqu'au lieu d'être un rectangle, on sait que la section peut être un losange.

Même problème, lorsqu'on sait que l'angle polyèdre peut être coupé suivant un carré. Lieu du centre du carré (cône oblique à base circulaire).

740. On considère un tronc de cône dans lequel la hauteur est moyenne proportionnelle entre les diamètres des deux bases : Démontrer qu'on peut inscrire une sphère (ex. 719) à ce solide (Pl., ex. 135).

Montrer que la surface latérale de ce tronc de cône est équivalente à celle d'un cercle ayant pour rayon l'arête latérale.

Construire les rayons des bases, connaissant la hauteur et l'arête latérale.

741. Construire trois sphères passant respectivement par les sommets d'un triangle donné, tangentes en ces sommets au plan du triangle et tangentes entre elles deux à deux. Calculer leur rayons, connaissant les côtés du triangle.

742. Sur la sphère céleste, on considère les points M tels que leur ascension droite soit égale à leur déclinaison (¹) :

1° Trouver le lieu des projections des points M sur le plan de l'équateur :

2. Trouver le lieu des droites AM, A étant, sur l'équateur, le point origine des ascensions droites.

743. Étant donnés trois points A, B, C, en ligne droite (B entre A et C), on décrit, sur BC comme diamètre, un demi-cercle auquel on mène la tangente AD; soit M un point de l'arc BD. On propose de déterminer le point M de sorte qu'en joignant MC, MA et en faisant tourner la figure autour de AC, le volume engendré par le triangle MAC soit partagé dans le rapport donné k par la zone qu'engendre l'arc MB. Trouver les limites de k en fonction du rapport h de OA à OB, O étant le milieu de BC.

744. On donne un carré ABCD, dont le côté est égal à $2a$, et on considère un point M dans le plan de ce carré.

---

(1) Tisserand et Andoyer, *Leçons de Cosmographie*, nᵒˢ 9-12.

1° Calculer la somme S des volumes engendrés par les triangles MAB, MBC, MCD, MDA tournant respectivement, le premier autour de AB, le second autour de BC, le troisième autour de CD, le quatrième autour de DA.

2° Démontrer que la somme S ne dépend que des quantités a et d, en désignant par d la distance du point M au centre du carré, et trouver le lieu géométrique du point M lorsque ce point se déplace de manière que la somme S reste équivalente au volume du cône de hauteur a et de rayon donné r.

3° Déterminer, sur le lieu précédent, les positions de M pour lesquelles le rapport des volumes engendrés par le triangle MAB, tournant successivement autour de MA et de MB, est égal à un nombre donné m.

745. Soit un trapèze OAMB dans lequel le côté OB est perpendiculaire aux bases OA, BM. On suppose que les sommets O et A sont fixes et que le sommet B se déplace sur la droite fixe Oy perpendiculaire à OA. On fait tourner la figure autour de Oy.

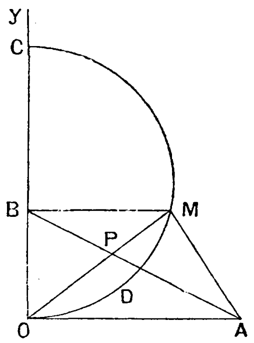

1° On suppose que les volumes engendrés par les triangles OAB et ABM sont équivalents et on demande de trouver, dans le plan AOy, le lieu géométrique du point M ainsi que le lieu géométrique du point de rencontre P des diagonales OM et AB.

2° Soit ODMC la demi-circonférence tangente en O à OA et passant par M; on fait tourner la figure autour de Oy et on suppose que les volumes engendrés par le triangle OAB et par la surface ODMA (limitée par les droites OA, AM et par l'arc de cercle ODM) sont équivalents. Trouver, dans cette deuxième hypothèse, le lieu décrit par M dans le plan AOy.

3° Déterminer le point M, sachant que les volumes décrits par les triangles OAB et ABM et par la surface ODMA sont équivalents. (On désignera par a la longueur donnée OA.)

LIVRE X

NOTIONS SUR LA TOPOGRAPHIE

CHAPITRE PREMIER

GÉNÉRALITÉS. PLANIMÉTRIE

551. La *topographie* est l'art de déterminer la forme d'un terrain. Nous nous proposons uniquement, ici, d'exposer les principes géométriques sur lesquels repose la topographie, en renvoyant, pour les indications pratiques et les détails de toute espèce, aux traités spéciaux. Il est, d'ailleurs, à peine utile d'ajouter que la lecture de ceux-ci ne saurait être poursuivie utilement si on ne la complète en effectuant soi-même, sur le terrain, les opérations étudiées.

Nous nous bornons aux méthodes approximatives qui suffisent dans la pratique courante. Les opérations de grande précision nécessitent une série de précautions auxquelles nous ferons allusion, chemin faisant, mais dont nous ne saurions donner même une idée sans dépasser le cadre de cet ouvrage.

552. On appelle *verticale* en un point la direction de la pesanteur en ce point.

On détermine pratiquement cette direction comme position d'équilibre du *fil à plomb*, c'est-à-dire d'un fil auquel est suspendu un corps pesant. Lorsqu'on voudra, par exemple, s'assurer qu'une tige est verticale, il faudra constater qu'elle est parallèle à la direction du fil à plomb.

Cette direction n'est pas la même en des points différents de la surface du globe, puisque la terre a sensiblement la forme d'une sphère (¹) et que les verticales vont concourir au centre de cette sphère. Mais elle change très lentement (²) et l'on peut admettre, dans la topographie usuelle (mais non, bien entendu, dans les opérations de précision), que les verticales menées par les différents points du terrain que l'on veut étudier sont parallèles entre elles.

Un plan est dit vertical lorsqu'il est parallèle à la verticale.

553. Un plan *horizontal* est celui qui est perpendiculaire à la verticale.

Une droite est également dite horizontale quand elle est perpendiculaire à la verticale.

Lorsqu'un liquide est en équilibre, sa surface libre (considérée sur une étendue telle qu'on puisse y admettre le parallélisme des verticales) a la forme d'un plan horizontal (³).

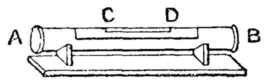

Fig. 466.
Niveau à bulle d'air.

On utilise cette propriété pour s'assurer qu'une direction est horizontale. L'instrument employé à cet effet est connu sous le nom de *niveau à bulle d'air*. C'est un tube de verre légèrement convexe, enchâssé dans une monture métallique AB (*fig.* 466) et dans lequel on a versé un liquide qui le remplit incomplètement de manière à y laisser une bulle d'air. On reconnaît que l'axe du tube est horizontal lorsque la surface libre du liquide affleure entre deux traits C, D (*fig.* 466), marqués à l'avance sur le verre.

On voit que le niveau à bulle d'air sert à reconnaître l'horizontalité d'une droite.

Pour constater l'horizontalité d'un plan, on constate l'horizontalité de deux droites de ce plan.

Il faut que ces deux droites ne soient pas parallèles, et

---

(1) Tisserand et Andoyer, *Leçons de Cosmographie*, livre II.

(2) Il résulte de la manière même dont a été obtenu le mètre que les verticales en deux points du globe font entre elles un angle de 90° lorsque ces points sont distants d'environ 10 000 kilomètres. Dès lors, pour une altération de 1° dans la direction de la verticale, il faut un déplacement de $\frac{10\,000}{90} = 111$ km. Une altération d'une minute correspond encore, comme on le voit, à un déplacement d'environ 2 kilomètres.

(3) Le fait cesse d'être exact aux bords de la surface, la direction de celle-ci étant alors altérée par l'effet d'une action due aux parois et nommée *capillarité* (voir ci-après, n° 574).

on doit même les choisir à peu près perpendiculaires l'une à l'autre (¹).

554. Pour rendre un plan matériel (par exemple une planche ou un plateau métallique) horizontal, on opère par tâtonnements, en modifiant progressivement la direction de ce plan, jusqu'à ce que, par l'emploi du niveau à bulle d'air, on constate l'horizontalité parfaite.

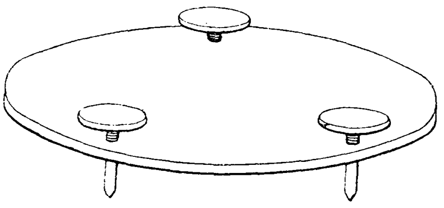
FIG. 468.

Pour pouvoir déplacer le plan de manière à modifier arbitrairement sa direction, on emploie plusieurs dispositifs.

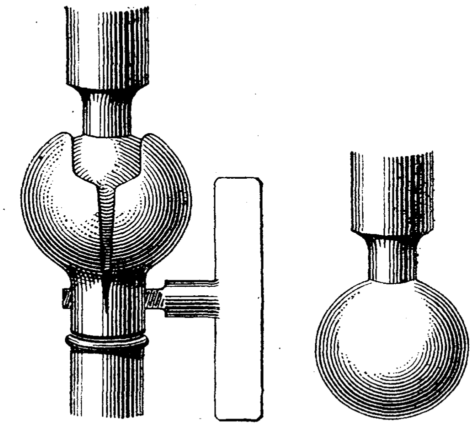
FIG. 469.

1° On peut faire supporter le plan en question par trois vis (*fig.* 468) dont l'action permet d'élever ou d'abaisser à volonté trois points du plan. Cette disposition, permettant d'opérer les déplacements par degrés insensibles, est adoptée, à l'exclusion de toute autre, toutes les fois qu'une certaine précision est nécessaire.

2° On obtient le même résultat, mais beaucoup plus grossièrement, à l'aide du *genou à coquilles* (*fig.* 469). Dans ce mode de suspension, le plan mobile porte une sphère pleine qui vient remplir exactement une sphère creuse fixée au support (²), la première sphère pouvant tourner arbitrairement à l'intérieur de la seconde.

3° Dans un autre système de genou, le plan mobile n'est pas lié à son support directement, mais par l'intermédiaire d'un cylindre intermédiaire C (*fig.* 470), de manière que le plan peut tourner autour de l'axe AA′ du cylindre, ce dernier pouvant,

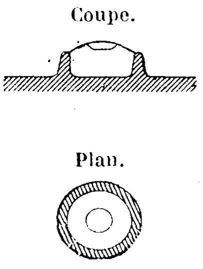
Coupe.
Plan.
FIG. 467.

---

(1) Dans certains appareils, on emploie des niveaux à bulle d'air de forme arrondie (*fig.* 467) : lorsqu'il y a horizontalité, la bulle d'air vient se placer dans un cercle tracé sur la face supérieure de l'instrument ; on n'a alors besoin que d'une seule opération pour reconnaître l'horizontalité d'un plan.

(2) La sphère creuse est formée de deux parties ou *coquilles*, d'où le nom de l'instrument.

d'autre part, tourner autour d'un axe B (*fig.* 470) perpendiculaire au premier et fixé au support. Cette disposition permet bien (ex. 577, 578) de donner au plan une direction arbitraire.

Dans l'un comme dans l'autre des deux appareils précédents, il existe des vis permettant de supprimer toute mobilité, une fois l'horizontalité obtenue.

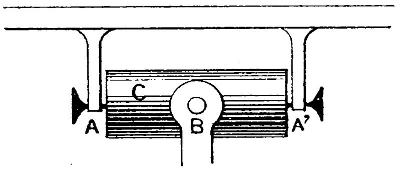

Fig. 470.

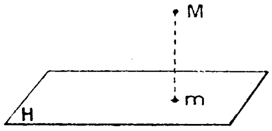

Fig. 471.

555. Soit choisi une fois pour toutes un plan horizontal déterminé H (*fig.* 471), dit *plan de comparaison*. Un point quelconque M sera déterminé si l'on se donne sa *projection horizontale m* (*fig.* 471) (projection sur le plan H) et sa *cote* (¹) *m*M. Il serait, toutefois, d'une manière générale, nécessaire d'indiquer le sens dans lequel cette cote doit être portée ; mais cette nécessité disparaît si l'on a pris, ainsi qu'on le fait généralement en topographie, le plan de comparaison au-dessous de tous les points de la figure étudiée (²).

D'après cela, on voit qu'on connaîtra entièrement la forme d'une figure, en particulier d'un terrain, si l'on a déterminé :

1° Sa projection horizontale, qui en est encore dite le *plan;*

2° Les cotes des différents points,

d'où deux sortes d'opérations en topographie : la *planimétrie* et le *nivellement*.

556. Levé du plan.

Lever le plan d'un terrain, c'est noter tous les éléments qui déterminent la forme et les dimensions de ce plan.

Lorsqu'on a levé le plan d'un terrain, on est à même de construire, sur le papier, une figure semblable à la projection horizontale de ce terrain, avec un rapport de similitude donné. C'est ce qu'on nomme *rapporter* le plan sur le papier, et le rapport de similitude en question se nomme l'*échelle* du plan ainsi tracé.

---

(1) Voir le cours de Géométrie descriptive.

(2) Il est d'ailleurs clair qu'au besoin on pourrait indiquer le sens de la longueur *m*M par le signe donné à la cote.

Les opérations fondamentales de la planimétrie sont : 1° la détermination d'une droite; 2° la mesure d'une longueur; 3° la mesure d'un angle.

557. Détermination d'un alignement.

Au point de vue de la planimétrie, un point du terrain peut être évidemment considéré comme déterminé si l'on connaît sa projetante, c'est-à-dire la verticale qui passe par ce point. On représente sur le terrain cette projetante par un *jalon*, c'est-à-dire par une tige, munie d'une marque destinée à la faire reconnaître de loin, et que l'on plante dans le sol bien verticalement (à l'aide du fil à plomb).

De même la direction d'une droite n'est intéressante à connaître en planimétrie que par le plan vertical ou *alignement* qui la contient.

L'alignement est déterminé par les deux extrémités de la droite. Si ces extrémités sont assez éloignées l'une de l'autre, les jalons qui les représentent peuvent, dans les opérations usuelles, être assimilés à des droites géométriques, sans que l'incertitude due à leur épaisseur dépasse la grandeur des erreurs admises.

Pour des raisons diverses, on peut avoir besoin de *jalonner* un alignement, c'est-à-dire de placer des jalons intermédiaires sur cet alignement, ou encore de le prolonger au delà de l'extrémité primitive.

On reconnaît que trois jalons font partie du même alignement si l'on peut placer l'œil de manière à ce que le premier jalon masque en même temps les deux autres.

Le jalonnement d'un alignement nécessite — comme d'ailleurs les opérations de topographie en général — la présence de deux opérateurs. L'un d'eux vérifie l'alignement, comme il vient d'être dit; l'autre se déplace sur les indications du premier, en transportant le jalon à poser, jusqu'à ce que l'alignement soit réalisé.

558. On a besoin, ainsi qu'on le verra plus loin, de pouvoir déterminer un alignement par deux verticales dont la distance ne dépasse pas les dimensions des instruments que nous décrirons aux n^{os} **560-561**.

On arrive à ce résultat par l'emploi de l'*alidade à pinnules*.

Une *pinnule* est une plaque allongée (*fig.* 472) percée, sur une moitié de sa longueur, d'une fente rectiligne étroite *ab*, et sur l'autre d'une fenêtre rectangulaire que traverse un fil fin *cd*. Le fil étant dans le prolongement de la fente, une droite est ainsi déterminée.

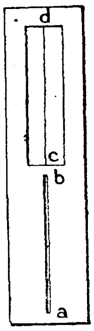

Fig. 472.

Une *alidade* est une règle munie à ses deux extrémités, de pinnules perpendiculaires à son plan (voir *fig.* 476). Ces pinnules sont disposées de manière que le fil de l'une soit en face de la fente de l'autre et inversement. En plaçant l'œil de manière que le premier fil masque la seconde fente et que la première fente laisse apercevoir le second fil, on vise un plan déterminé avec l'exactitude voulue.

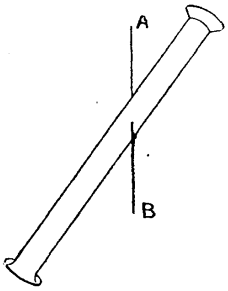

Fig. 473.

Une précision plus grande est obtenue par l'emploi des lunettes. Si une lunette est mobile autour d'une droite AB (*fig.* 473) perpendiculaire à son axe, ce dernier se meut dans un plan, lequel est vertical si la droite AB est horizontale.

559. Mesure directe d'une longueur.

On mesure les longueurs sur le terrain à l'aide de la *chaîne d'arpenteur*, qui est un décamètre divisé en *chaînons* de 20 centimètres chacun.

Si la droite à mesurer est parfaitement horizontale, on portera la chaîne sur cette droite autant de fois qu'on pourra le faire ; il restera, en général, un dernier segment moindre qu'un décamètre, et qu'on évaluera en comptant les chaînons et mesurant la fraction de chaînon qu'il comprend. On devra prendre soin :

1° Que, dans chacune de ses positions successives, la chaîne soit bien tendue ;

2° Qu'elle soit bien dans l'alignement donné (n° précéd.) ;

3° Que dans chaque position, l'extrémité postérieure de la chaîne soit à la place exacte où était l'extrémité antérieure dans la position précédente.

On marque, à cet effet, à l'aide d'une petite tige de fer, appelée *fiche*, que l'on plante dans le sol, l'endroit où arrive l'extrémité antérieure de la chaîne avant de relever celle-ci.

Mais dans la pratique, la droite à lever n'est jamais horizontale. C'est alors (d'après ce qui précède) sa projection horizontale qu'il s'agit de mesurer.

On y arrive en tendant la chaîne horizontalement, dans chacune de ses positions successives, et plaçant chaque fiche (à l'aide du fil à plomb) dans la verticale qui passe par l'extrémité antérieure de la chaîne (¹). Par exemple, la figure 474 représente un alignement AB sur lequel la chaîne a été reportée trois fois en AB_1, A_2B_2, A_3B_3, un dernier segment A_4B_4 étant plus petit que la longueur de la chaîne. Si tous ces segments sont dans un même plan et horizontaux, que les points A_2, A_3, A_4, B_4 soient bien sur les verticales menées respectivement par B_1, B_2, B_3, B, la somme des quatre segments A_1B_1, A_2B_2, A_3B_3, A_4B_4 (soit trois fois la longueur de la chaîne, plus le segment A_4B_4) représentera bien la projection horizontale de AB.

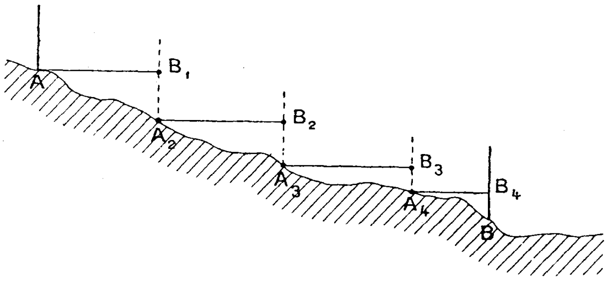

Fig. 474.

L'horizontalité de la chaîne, sa parfaite tension, etc., ne peuvent être obtenues qu'assez approximativement dans la pratique : néanmoins, avec des opérateurs exercés, on arrive à ne pas dépasser une erreur de $\frac{1}{1000}$. Dans les opérations très précises, on est obligé, pour éviter les causes d'erreur dont nous venons de parler et d'autres sur lesquelles nous n'insisterons pas, à une série de précautions délicates et compliquées. Aussi, dans ces derniers levés et même dans les levés usuels, réduit-on les mesures *directes* de longueur au moindre nombre possible, toutes les autres longueurs étant évaluées indirectement, comme il est expliqué ci-après (nº **563**) (²).

---

(1) Nous nous plaçons dans l'hypothèse où la droite à lever est descendante, ce qui revient à dire que l'on part de l'extrémité la plus élevée de cette droite. C'est d'ailleurs ainsi que l'on procède généralement.

(2) Il ressort avec évidence de ce qui est dit au nº **563**, que l'on ne peut se dispenser de mesurer directement au moins une longueur.

Dans les opérations de précision, on mesure *deux* droites ou *bases*, situées aussi loin que possible l'une de l'autre. Si les opérations ont été bien conduites, la valeur de la seconde base, calculée à l'aide de la première, doit coïncider avec la valeur mesurée directement.

560. Mesure directe d'un angle.

La projection horizontale de l'angle de deux droites AB, AC n'est évidemment autre que l'angle plan du dièdre formé par les alignements qui les contiennent : c'est ce que montre la simple inspection de la *fig.* 475.

L'instrument le plus généralement usité pour la mesure des angles sur le terrain est le *graphomètre* (*fig.* 476).

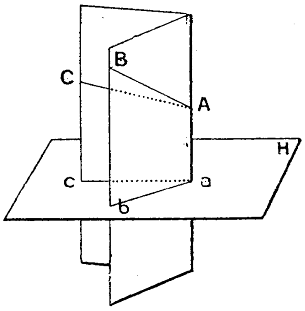

FIG. 475.

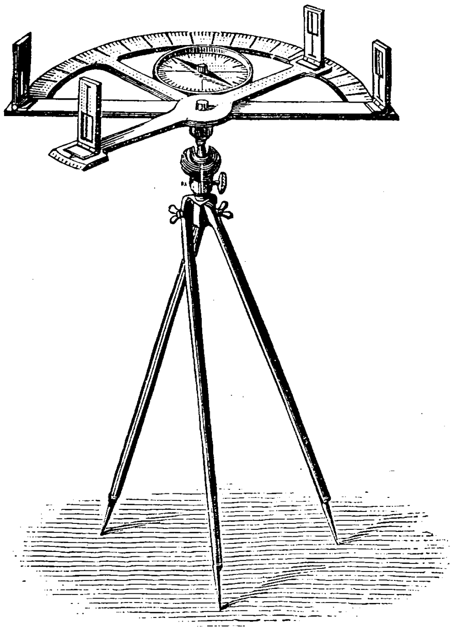

FIG. 476. — Graphomètre.

Il se compose d'un demi-cercle métallique ou *limbe*, monté sur genou (**554**) et, de plus, mobile dans son plan. Ce limbe, sur lequel est tracée une division en degrés et minutes (1), porte en outre deux alidades : l'une est fixe et a l'un de ses bords le long de la ligne 0° — 180° de la graduation ; l'autre alidade est mobile autour du centre du limbe.

Pour mesurer l'angle de deux alignements, on installe le graphomètre de manière que son centre soit dans la verticale du sommet de l'angle. On rend le limbe horizontal en faisant mouvoir le genou, puis on fait tourner ce limbe dans son plan de manière à amener l'alidade fixe dans le premier des deux alignements donnés. Plaçant ensuite la seconde alidade dans le second de ces deux alignements, il ne reste plus qu'à lire l'angle cherché.

Dans des instruments plus exacts, on substitue aux alidades une lunette mobile. C'est ce qui est réalisé dans le *théodolite* (2), instrument qui remplace le graphomètre dans les mesures de précision.

---

(1) En réalité, le limbe n'est gradué qu'en degrés et demi-degrés. Un dispositif spécial appelé *vernier*, et sur lequel nous n'insisterons pas, permet de lire les minutes.

(2) TISSERAND et ANDOYER, *Leçons de Cosmographie*, n° 5, page 6.

On sait (¹) qu'on appelle *azimuth* d'une direction, l'angle que fait l'alignement vertical qui contient cette direction avec un plan vertical fixe pris comme origine. L'angle de deux alignements peut évidemment être considéré comme la différence de leurs azimuths.

On prend souvent, comme direction origine, la direction sud-nord (plan méridien du lieu). Pour permettre d'évaluer les azimuths ainsi définis, le graphomètre est, en général, muni d'une boussole.

561. La *planchette* est un instrument qui permet, en même temps, de lever un plan et de le rapporter sur le papier.

Elle se compose, comme son nom l'indique, d'une planche bien plane, montée sur genou, pouvant également pivoter et glisser sur elle-même dans son plan et sur laquelle on tend une feuille de papier à dessin (*fig.* 477).

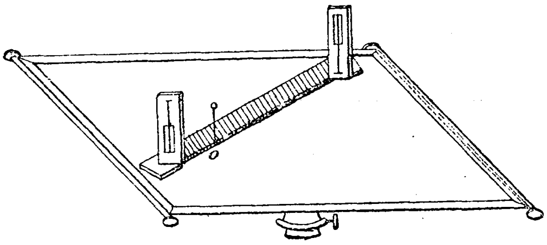

Fig. 477. — Planchette.

Une alidade mobile peut être posée à volonté sur la planchette. Le bord de cette alidade porte une petite échancrure qui permet de la faire pivoter autour d'une aiguille *o* (*fig.* 477) piquée à volonté dans le papier.

Soit à lever un angle à la planchette.

En général, le sommet *o* de l'angle et un côté *oa* sont marqués d'avance sur le papier. Il faut alors :

1° A l'aide du genou, rendre l'instrument horizontal ;

2° Par les mouvements de l'appareil dans son plan, s'arranger pour que le sommet *o* marqué sur le papier soit dans la verticale du sommet donné sur le terrain (²) et pour que, plaçant l'alidade suivant *oa*, cette alidade soit dans le premier alignement donné.

Toute cette première partie du travail se nomme la *mise en station* de l'instrument.

On fait alors pivoter l'alidade autour de l'aiguille piquée en *o*

---

(1) Tisserand et Andoyer, *Leçons de Cosmographie*, n° 4, page 5.

(2) Le point *o* peut être à 3 ou 4 centimètres de la verticale du sommet donné sans qu'il en résulte d'erreur dépassant les limites admises.

jusqu'à ce qu'elle vise le second alignement donné, et on trace un trait de crayon ob le long du bord, dans cette nouvelle position. L'angle $\widehat{boa}$ représente évidemment l'angle cherché.

Cette manière de procéder peut sembler, au premier abord, plus exacte que celle dans laquelle on fait usage du graphomètre, puisque, *théoriquement*, l'angle $\widehat{boa}$ obtenu est rigoureusement égal à l'angle qu'il s'agit d'évaluer, tandis que la lecture au graphomètre ne donne cet angle qu'à une ou deux minutes près.

En réalité, c'est le contraire qui a lieu : en raison des petites dimensions du dessin tracé sur la planchette, les erreurs dues à l'épaisseur du trait de crayon, aux petites déviations de l'alidade, etc., sont notablement supérieures à celles qu'introduit l'emploi du graphomètre. Mais la planchette, instrument inférieur au point de vue de l'exactitude, a souvent pour elle l'avantage de la rapidité.

562. Levé d'un triangle.

Un levé de plan quelconque peut se ramener (voir ci-après) à une série de levés de triangles. Or, cette dernière opération peut toujours être effectuée à l'aide de celles que nous avons décrites aux trois numéros précédents. Pour lever un triangle, il suffit d'avoir mesuré :

1° Un côté et deux angles ;

2° Ou deux côtés et l'angle compris ;

3° Ou les trois côtés.

En particulier, on peut alors rapporter le triangle sur le papier, à une échelle donnée quelconque. Si, par exemple, on a mesuré un côté AB et les deux angles adjacents $\widehat{A}$, $\widehat{B}$, on tracera sur la feuille un segment de droite ab qui soit avec la longueur mesurée dans un rapport égal à l'échelle donnée (1). Puis, si l'on a opéré avec le graphomètre, il restera à construire, à l'aide du rapporteur (2), deux droites ac, bc faisant avec ab des angles respectivement égaux à ceux qui ont été levés sur place.

---

(1) Comme précédemment pour les lectures angulaires, des artifices spéciaux permettent de déterminer le segment ab au 1/10 de millimètre près, quoique les règles servant aux mesures ne soient divisées qu'en millimètres.

(2) Le rapporteur, en raison de ses petites dimensions, ne permet d'obtenir que des approximations assez grossières dans la mesure des angles. Aussi, lorsqu'on veut une certaine exactitude, doit-on construire le triangle à l'aide de ses trois côtés, calculés s'ils ne sont pas connus directement (voir numéro suivant).

Si, au contraire, on a employé la planchette, le tracé des droites *ac*, *bc* aura été effectué (**561**) au moment même du levé sur le terrain.

563. Mesure indirecte des longueurs et des angles. De ce que nous venons de dire résulte qu'on peut connaître une longueur ou un angle sans les avoir mesurés directement. Il suffit de lever un triangle ayant pour un de ses éléments la longueur ou l'angle en question ; cet élément peut alors être considéré comme déterminé.

Si l'on veut obtenir effectivement la mesure — en mètres, s'il s'agit d'une longueur; en degrés et minutes, s'il s'agit d'un angle — on pourra rapporter le triangle sur le papier, à une échelle prise arbitrairement. La grandeur cherchée sera :

Si c'est un angle, égale à l'angle homologue de la figure ainsi tracée ;

Si c'est une longueur, égale à la longueur homologue, multipliée par un rapport inverse de l'échelle choisie.

Si celle-ci n'est pas trop petite, l'erreur commise dans la mesure sur le papier, multipliée par le rapport inverse dont nous venons de parler, pourra ne pas donner un produit supérieur à l'erreur admise, et en particulier à l'erreur provenant des opérations sur le terrain qui ont donné les éléments directement connus.

S'il en est autrement, il faudra, soit construire sur le terrain un triangle égal ou semblable à celui qu'on a levé — moyen évidemment très incommode et que nous ne mentionnons que pour mémoire — soit recourir à la trigonométrie, laquelle fait connaître le résultat cherché sans introduire aucune erreur nouvelle (¹).

564. La remarque qui précède est d'une application constante en topographie. Citons-en immédiatement deux conséquences :

1° *On peut lever un angle sans employer d'autre instrument que la chaîne.*

Il suffira de lever les trois côtés d'un triangle ayant l'angle donné parmi ses éléments.

Par conséquent, *on peut lever un plan quelconque avec la chaîne*

(¹) Bourlet, *Leçons de Trigonométrie*, livre III. — Les quantités fournies par les calculs trigonométriques ne le sont qu'avec une certaine approximation ; mais l'erreur provenant de ce fait est d'ordre de beaucoup inférieur à celui des erreurs qui résultent des mesures sur le terrain.

seule, puisque nous ramènerons tout levé de plan à des mesures de longueurs et d'angles.

2° *On peut mesurer la distance* (*horizontale*) *de deux points* A, B (*fig.* 478) *dont le second est inaccessible, quoique visible.* Il suffit, évidemment, de mesurer une distance AC et de viser les deux angles $\widehat{BAC}$, $\widehat{BCA}$.

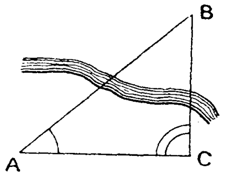

Fig. 478.

On voit seulement qu'il est nécessaire, pour déterminer exactement la position du point B, d'observer ce point, non seulement du point A, mais d'un autre point C. La visée faite du point A ne peut servir qu'à faire connaître la direction de la droite AB et ne peut donner à elle seule la longueur de cette droite.

Remarquons encore que le procédé est en défaut lorsqu'on prend le point C sur la droite AB. Il est même inapplicable si l'angle en B est trop aigu, car alors l'intersection des droites AB, CB détermine mal le point B.

565. Triangulation.

Il est aisé de comprendre comment un levé quelconque peut être ramené à des levés de triangles.

Supposons qu'on ait déjà levé un plan comprenant un certain nombre de points A, B, C, ..., K, L (*fig.* 479) et qu'on veuille *rattacher* à ce levé un nouveau point M, c'est-à-dire lever la figure formée par les points A, B, C, ..., K, L, M.

Il suffira, pour cela, de lever le triangle formé par le point M et deux des points appartenant à la figure primitive, les points K et L, par exemple (triangle dans lequel le côté KL doit déjà être regardé comme connu, puisqu'il appartient à la figure déjà levée). Si l'on a noté en outre le sens de ce triangle, autrement dit si l'on a observé de quel côté se trouve M par rapport à la droite KL, le problème est bien résolu : car si, connaissant par avance les points A, B, C, ..., K, L, on se donnait les éléments et le sens du triangle KLM, cela suffirait pour que la position du point M fût déterminée.

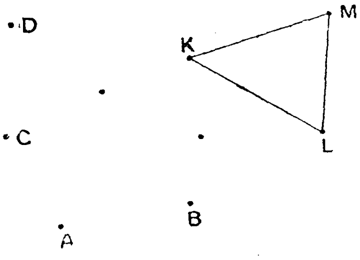

Fig. 479.

Donc, pour lever le plan de la figure composée des points A, B, C, ..., on mesurera d'abord la distance de deux d'entre eux, A, B, puis on rattachera un troisième point C aux deux premiers; un quatrième point D au levé ABC, et ainsi de suite.

La méthode comporte évidemment un certain degré d'arbitraire. En particulier, on peut évaluer une quelconque des distances CD, CE, ... de plusieurs manières différentes. On a ainsi un moyen de vérification des opérations.

La série de levés de triangles par laquelle on obtient le plan d'un terrain quelconque se nomme une *triangulation*.

566. Dans les levés simples que l'on rencontre usuellement, il est commode de réduire au moindre nombre possible les mises en station du graphomètre ou de la planchette : d'où les deux méthodes dites par *intersections* et par *rayonnement*.

Dans la méthode par *intersections*, on chaîne un alignement AB, auquel on rattache tous les autres points de la figure. A cet effet, M étant un point quelconque, on visera les angles en A et B du triangle ABM. Cette méthode n'exige, comme on le voit, que deux mises en station, en A et en B. Elle suppose qu'on ait choisi la *base* AB de manière à ce qu'aucun des angles tels que $\widehat{AMB}$ ne soit trop aigu (**564**).

Dans la méthode par *rayonnement*, on donne à tous les triangles un même sommet A et l'on mesure : 1° les angles en A ; 2° les côtés issus de A. On n'a ainsi qu'une seule mise en station, en A.

567. Levé d'un polygone par cheminement.

On peut lever un polygone par une méthode toute semblable à celle qui sert pour un triangle, à savoir, en en mesurant les côtés et les angles.

Il n'est pas nécessaire, théoriquement, de mesurer tous les côtés et tous les angles du polygone; il suffit de connaître tous ces éléments, moins trois. Par exemple, on aura levé le pentagone ABCDE (*fig.* 480) si on a déterminé les angles $\widehat{B}$, $\widehat{C}$, $\widehat{D}$ et les côtés AB, BC, CD, DE : car on aura ainsi rattaché les différents sommets les uns aux autres par l'intermédiaire des triangles ABC, BCD, CDE (dans chacun desquels on connaît deux côtés et l'angle compris), et il ne restera plus qu'à joindre EA.

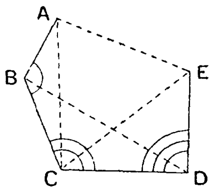

Fig. 480.

On voit, en même temps, que le cheminement n'est qu'un mode particulier de triangulation.

Dans la pratique, on relève tous les côtés et tous les angles, afin d'obtenir des vérifications.

La première de ces vérifications peut être faite dès la levée des angles : elle consiste à s'assurer que la somme de ceux-ci est bien égale à autant de fois deux droits qu'il y a de côtés moins deux.

La seconde ne se fait (du moins si l'on n'emploie pas la trigonométrie) qu'au moment où l'on rapporte le plan sur papier. Le polygone doit se fermer exactement si le levé a été bien fait (1).

568. Emploi de l'équerre d'arpenteur.

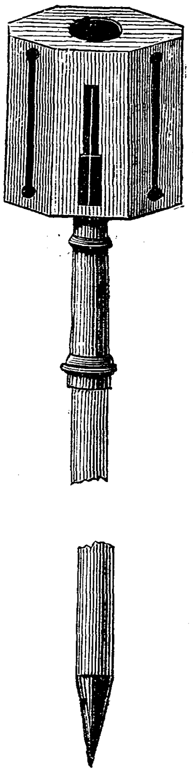

Fig. 481.
Équerre d'arpenteur.

L'équerre d'arpenteur est un instrument qui sert à mener des perpendiculaires sur le terrain. Elle consiste en un prisme octogone (*fig.* 481) dont quatre faces (2) sont munies de pinnules. Ce prisme est porté sur une tige, parallèle aux arêtes du prisme, et que l'on enfonce *verticalement* dans le sol. Si p, p' (*fig.* 482) sont les pinnules portées par deux faces opposées ; q, q', celles qui appartiennent à deux autres faces également opposées, on a ainsi déterminé deux alignements pp', qq'. L'équerre est *juste* lorsque ces deux alignements sont perpendiculaires entre eux.

Les deux problèmes fondamentaux que l'on peut résoudre avec l'équerre supposée juste sont les suivants :

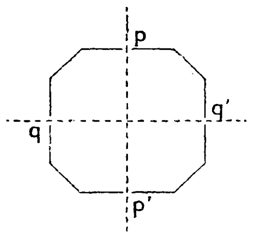

Fig. 482.

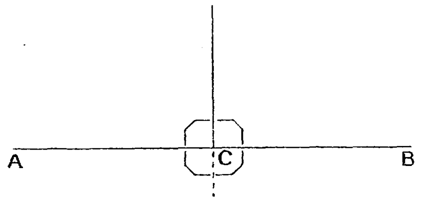

Fig. 483.

1° *Élever une perpendiculaire à un alignement* AB (*fig.* 483) *par un point* C *pris sur cet alignement.*

---

(1) Dans la pratique on n'arrive pas, en général, à une fermeture rigoureusement exacte. Mais l'erreur de fermeture ne doit pas dépasser une limite assez faible.

(2) En réalité, les huit faces portent des pinnules. Mais nous ne décrivons ici, pour simplifier, que ce qui est essentiel pour le fonctionnement de l'appareil.

On place l'équerre au point C de manière qu'un de ses plans de visée soit dans l'alignement donné ; le plan de visée perpendiculaire donnera l'alignement cherché.

2° *Abaisser d'un point* C, *extérieur à un alignement* AB, *une perpendiculaire sur cet alignement.*

Cette opération est la plus délicate de la topographie usuelle. Il faut, en effet, procéder par tâtonnement et déplacer l'équerre jusqu'à ce qu'on l'ait amenée à une position O située dans l'alignement AB et telle que la perpendiculaire élevée en O à AB (problème précédent) passe par C.

569. Soient un alignement jalonné sur le terrain, projeté horizontalement suivant Ox (*fig.* 484), et un autre alignement, perpendiculaire au premier, projeté horizontalement suivant Oy, et qui ne joue un rôle que dans le raisonnement. On peut déterminer, à l'aide de la chaîne et de l'équerre, les coordonnées de la projection horizontale d'un point quelconque M par rapport aux axes Ox, Oy. Il suffira d'abaisser, du point M, la perpendiculaire MM' sur Ox (n° précédent) et de chaîner OM', M'M.

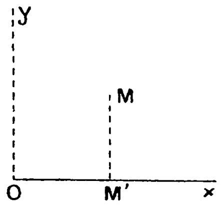

Fig. 484.

La connaissance de ces coordonnées détermine entièrement, comme nous le savons, la position de M par rapport aux axes Ox, Oy. Il suffira donc de répéter l'opération précédente avec un nombre quelconque de points donnés pour *lever*, *avec la chaîne et l'équerre*, *le plan de la figure formée par ces points.*

Rien n'empêche, d'ailleurs, de lever séparément différentes parties de la figure avec des axes différents pour chacune de ces parties.

570. En général, lorsqu'on veut lever le plan d'un terrain quelconque, on divise l'opération en deux parties.

On fait d'abord choix d'un certain nombre de points importants suffisamment éloignés les uns des autres et formant un polygone (*polygone topographique* ou *canevas*). On lève ce polygone avec le plus grand soin possible : on peut employer à cet effet l'une quelconque des méthodes que nous venons d'indiquer et opérer :

1° *Par cheminement*, à la chaîne et au graphomètre — ou même à la chaîne seule (en employant la remarque du n° **564**). Toutefois, cette dernière méthode, très longue et même peu exacte, n'est employée que dans les levés très peu étendus ;

2° *Par intersections*, si l'on a pu trouver une base qui ne soit vue d'aucun des sommets du polygone sous un angle trop aigu ;

3° *Par rayonnement ;*

4° Plus généralement, par une *triangulation* quelconque ;

5° *A l'équerre* (numéro précédent).

Toutefois, il faut observer que les deuxième et troisième méthodes (et souvent la dernière) deviennent inapplicables lorsque le terrain offre trop d'obstacles à la vue.

On n'emploie pas, en général, la planchette dans le levé du canevas, à cause de sa moindre exactitude.

571. Le canevas une fois levé, on y rattache les autres points intéressants en appliquant toujours les mêmes principes, mais visant un peu plus à la rapidité. La planchette est employée avec avantage dans cette partie du travail.

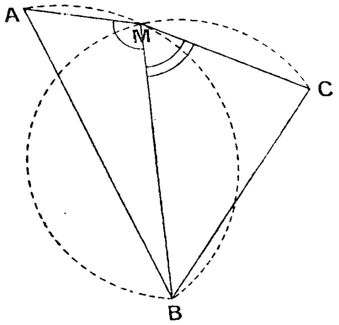

FIG. 485.

En dehors des méthodes précédemment décrites, un moyen simple de rattacher un point M au canevas consiste à viser de ce point trois points déjà relevés, A, B, C, et à noter les angles $\widehat{AMB}$, $\widehat{BMC}$. La connaissance de ces deux angles donne deux lieux du point M, à avoir deux segments de cercles (*fig.* 485) décrits l'un sur la corde AB, l'autre sur la corde BC. On n'a plus qu'à prendre, sur le papier, le point, autre que B, où se coupent les deux cercles. La trigonométrie permet d'ailleurs (¹) de déterminer tous les éléments des triangles ABM, BCM.

---

(1) *Leçons de trigonométrie rectiligne* de M. BOURLET, n° 179.

572. La remarque du n° **563** se généralise évidemment en ce sens que, pour connaître une longueur ou un angle, on peut rattacher les deux extrémités de la longueur, ou le sommet de l'angle et deux points pris sur ses côtés, à un levé que l'on effectue.

C'est ainsi que l'on procède, par exemple : pour *trouver la distance de deux points* A, B, *tous deux inaccessibles*, *mais visibles* (*fig.* 486). On mesure une base CD à laquelle on rattache les deux points donnés, par intersections (¹).

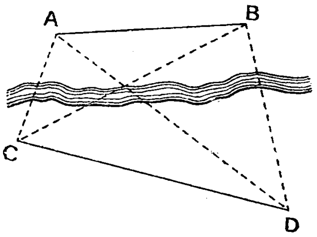

Fig. 486.

573. Lorsque le plan à lever contient une ligne (par exemple, une route ou le bord d'une rivière (*fig.* 487)), on lèvera des points M, M′, M″, ..., pris sur cette ligne et assez rapprochés les uns des autres pour faire connaître suffisamment sa forme.

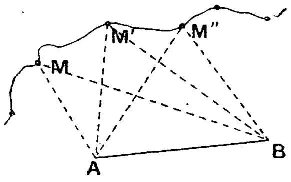

Fig. 487.

---

(1) *Leçons de trigonométrie rectiligne* de M. Bourlet, n° 177.

CHAPITRE II

NIVELLEMENT

574. Nivellement direct.

On mesure la différence des cotes de deux points à l'aide du *niveau d'eau*.

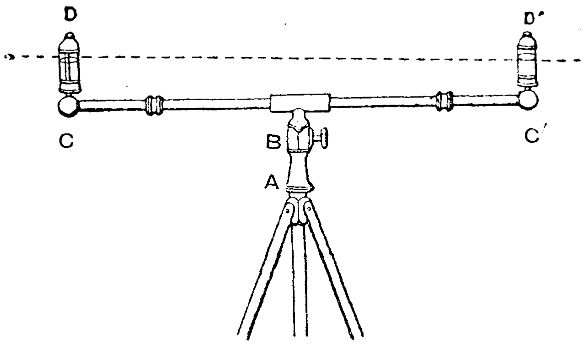

Fig. 488. — Niveau d'eau.

Cet appareil se compose (*fig.* 488) de deux fioles en verre F, F', communiquant entre elles par un tube TT' perpendiculaire à leur direction, qui aboutit au fond de chacune d'elles, et que l'on remplit incomplètement d'eau colorée. Le tout est porté sur un axe A que l'on installe verticalement. En vertu du principe des *vases communiquants*, établi en hydrostatique, les surfaces libres du liquide dans les deux fioles font partie d'un seul et même plan horizontal.

Le principe des vases communiquants n'est toutefois pas applicable sans modification, à cause de l'action de la *capillarité*, action dont l'effet est : 1° de relever toute la surface libre de l'eau dans chaque fiole, d'une petite quantité dépendant du diamètre de cette fiole; 2° de relever les bords de cette surface, de manière que celle-ci cesse d'être horizontale dans le voisinage de la paroi et prend la forme indiquée (*fig.* 489).

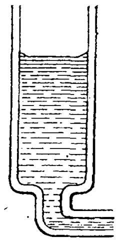

Fig. 489.

Mais, si les fioles sont exactement de même calibre, les actions capillaires sont identiques de part et d'autre. On a soin, en construisant l'appareil, que cette condition soit remplie. Lorsqu'il en est ainsi, les bords *ab*, *a'b'* des deux surfaces libres font partie d'un seul et même plan horizontal.

En plaçant l'œil suivant une tangente commune intérieure aux deux circonférences *ab*, *a'b'* et s'assurant qu'on voit bien ces deux lignes sous l'aspect de deux petites droites en prolongement l'une de l'autre (*fig.* 490), on vise une droite horizontale bien déterminée.

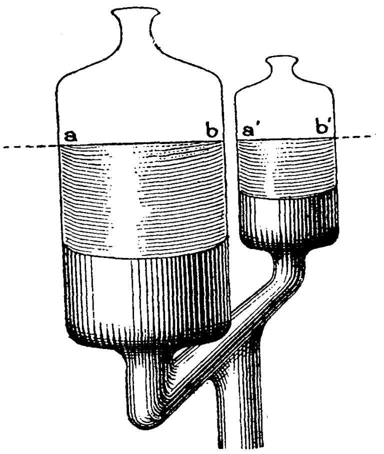

Fig. 490.

575. Pour pouvoir viser successivement toutes les directions d'un même plan horizontal, il suffit évidemment de faire tourner l'appareil autour de l'axe A, si celui-ci est bien vertical. La hauteur de l'eau dans les fioles reste alors invariable et la droite considérée tout à l'heure engendre un plan horizontal, dit *plan de visée* de l'appareil.

Lorsque l'axe A autour duquel tourne le niveau n'est pas exactement vertical, la conclusion précédente n'est pas applicable en général. Toutefois, on peut encore admettre, même dans ces conditions, que le niveau du liquide reste invariable dans la rotation, si l'on a constaté : 1° que les deux fioles sont de même diamètre intérieur (c'est ce que nous avons déjà supposé) ; 2° que l'axe aboutit au milieu du tube.

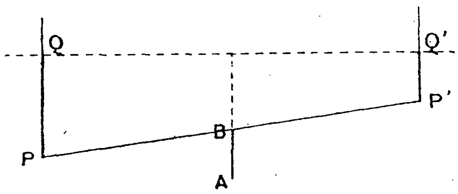

Fig. 491.

En effet, en toute hypothèse, le volume du liquide compris dans l'appareil demeure constant. Or, ce volume comprend : 1° le volume du tube horizontal, lequel est toujours le même ; 2° les volumes compris dans les fioles. Ceux-ci, puisque les fioles sont cylindriques et de même calibre, sont proportionnels aux hauteurs (¹) PQ, P'Q' (*fig.* 491) occupées par le liquide dans chacune d'elles, de sorte que leur somme est proportionnelle à la somme PQ + P'Q'.

Le volume total étant constant, nous voyons que, lors d'une

---

(1) Nous raisonnons comme si les fioles restaient verticales, ce qui n'est évidemment pas exact; mais, lorsque les inclinaisons sont petites, l'erreur provenant de ce fait est tout à fait négligeable.

rotation autour d'un axe non vertical, la hauteur occupée par le liquide dans l'une des fioles augmentera pendant qu'elle diminuera dans l'autre, mais que *la somme de ces hauteurs ne changera pas.*

Or, le trapèze (¹) PQ P′Q′ montre que cette somme est double de la distance verticale qui existe entre le milieu du tube et le niveau supérieur du liquide. Ce dernier devra donc rester invariable si le milieu du tube est sur l'axe. Il est même aisé de voir (exercice 855) que si la distance du milieu du tube à l'axe est petite, en même temps que l'angle de cet axe avec la verticale, la variation du plan de visée sera négligeable. On peut donc toujours, dans la pratique, considérer ce plan de visée comme fixe.

576. L'opération fondamentale exécutable avec le niveau d'eau est la suivante :

Étant donné un point M (*fig.* 492) *situé plus bas que le plan de visée de l'appareil*, *déterminer la cote de ce point au-dessous de ce plan.*

Il suffira à cet effet, de trouver, sur la verticale du point M, le point M′ qui est situé dans le plan de visée de l'appareil, et de noter la distance MM′.

C'est ce qu'on réalise à l'aide de la *mire*, c'est-à-dire d'une règle graduée portant une plaque carrée mobile, appelée *voyant*, laquelle est partagée en quatre parties peintes alternativement en blanc et en couleur (*fig.* 493) de manière que le centre en soit très nettement visible de loin.

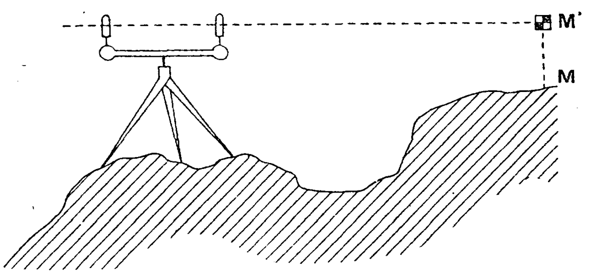

FIG. 492.

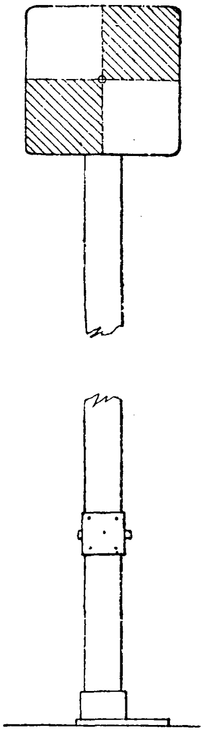

FIG. 493.

Un aide place cette règle bien verticalement au point M, puis, sur les indications de l'opérateur, fait monter ou descendre le

---

(¹) Nous réduisons, pour simplifier, le tube et les fioles à des droites.

voyant le long de la règle jusqu'à ce que son centre soit en M', dans le plan de visée du niveau. Il ne reste plus qu'à lire sur la graduation de la règle, la longueur MM'.

Il faut supposer toutefois : 1° que le point M, ou tout au moins le point M' peut être aperçu de l'endroit où est placé le niveau ; 2° que la distance horizontale des deux appareils ne soit pas trop grande, car la petite erreur commise par l'opérateur sur la direction de la visée a évidemment une influence croissante avec la distance en question ; 3° que la différence de cote cherchée ne soit pas supérieure à 2 mètres si la mire est *simple*, et à 4 mètres si la mire est à *coulisse*, c'est-à-dire se compose de deux règles pouvant glisser l'une sur l'autre, de manière à pouvoir s'allonger jusqu'au double de la longueur primitive. Cette dernière ne peut d'ailleurs dépasser 2 mètres, afin que l'extrémité de la première règle puisse être atteinte sans difficulté.

577. Nivellement simple.

Soit à *déterminer la différence de cote de deux points* M, N.

On installe le niveau d'eau de manière que son plan de visée soit supérieur aux deux points donnés et l'on mesure (n° précéd.) la cote de chacun de ces points au-dessous de ce plan de visée. Il ne reste plus qu'à soustraire l'une de l'autre les cotes ainsi mesurées.

On pourra faire les deux lectures sans aucun dérangement du niveau, si celui-ci est dans l'alignement MN. Sinon, on fera tourner l'appareil autour de son axe. On pourra, en employant ce dernier procédé, déterminer avec une seule mise en station du niveau, les différences existant entre les cotes d'une série de points visibles d'un même poste d'observation.

Pour que la méthode précédente soit applicable, il faut (n° précéd.) : 1° que l'on ait pu placer le niveau de manière à apercevoir en même temps les points donnés ; 2° que la distance horizontale de ces points au niveau d'eau ne soit pas trop grande ; 3° que les différences de cote cherchées soient assez petites.

578. Nivellement composé.

Si l'une ou l'autre des conditions que nous venons d'énumérer n'est pas remplie, il faut recourir au *nivellement composé*, c'est-à-dire faire choix d'un certain nombre de points auxiliaires P, Q,... S, tels que l'on puisse mesurer, par nivellement simple, les différences de cote qui existent entre M et P, entre P et Q, ..., entre S et N. La somme algébrique des quantités trouvées donne évidemment le résultat cherché.

579. Nivellement indirect.

De même que la mesure des angles et des longueurs horizontales, le nivellement peut s'effectuer indirectement.

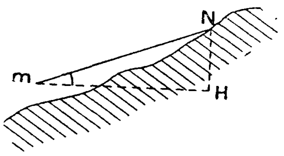

FIG. 494.

Soit, en effet, à mesurer la différence de cote entre un point m où l'on peut placer l'œil et un point N quelconque.

Soit H (*fig.* 494) la projection du point m sur la verticale du point N. Le triangle rectangle mNH sera complètement déterminé et, par conséquent, la différence de cote cherchée NH connue, si l'on a pu mesurer : 1° Le côté mH; 2° l'angle en m, lequel n'est autre que l'angle de la direction mN avec le plan horizontal, ou la *hauteur* (¹) de cette direction.

La première mesure est du ressort de la planimétrie : elle s'effectue, — directement ou indirectement suivant les cas, — à l'aide des méthodes exposées au chapitre précédent.

Pour mesurer la hauteur de la direction mN, on peut se servir du graphomètre, dont on installe le limbe verticalement de manière que le plan de visée de l'alidade fixe soit horizontal, en faisant ensuite passer le plan de visée de l'alidade mobile par la droite mN.

Mais, ainsi que nous venons de le voir, le nivellement indirect comporte une opération de planimétrie : de sorte que, la plupart du temps, lorsqu'on emploie cette méthode, on effectue concurremment le levé et le nivellement. Il y a, dans ces conditions, tout avantage à employer le théodolite, qui permet de mesurer, *à la fois et pour une même position de l'instrument*, l'azimuth et la hauteur d'une direction.

580. Pour obtenir effectivement la mesure de NH, on pourra, ainsi que nous l'avons dit plus haut :

1° Soit employer la trigonométrie (²) : c'est ce que l'on fait toujours lorsqu'on vise à une grande exactitude ;

---

(1) TISSERAND et ANDOYER, *Leçons de Cosmographie*, n° 4, page 5.

(2) *Leçons de trigonométrie rectiligne* de M. BOURLET, liv. III, ch. I et aussi n° 173-174.

2° Soit construire sur le papier un triangle semblable à mNH ;

3° Soit construire un tel triangle sur le terrain. Ceci conduit à un procédé de mesure très simple, en ce qu'il n'exige l'emploi d'aucun instrument angulaire. On plante, bien verticalement, dans l'alignement mN, une tige H'N' (*fig.* 495), sur laquelle on marque le point H' qui est situé dans le même plan horizontal que m et le point N qui est situé sur le rayon visuel mN. On a ainsi réalisé le triangle mH'N' semblable à mHN et la mesure de la distance horizontale mH, ainsi que des longueurs mH', H'N' fait connaître la cote cherchée par une simple proportion.

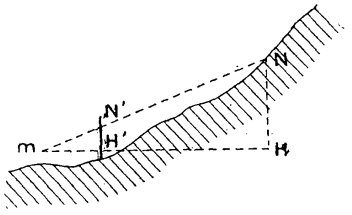

Fig. 495.

581. Si maintenant on veut la différence de cote existant entre deux points M, N du terrain, on pourra, suivant les cas :

Soit placer l'œil en m, au-dessus du point M, et ajouter (ou retrancher) à la différence de cote NH la distance verticale des points m,M, évaluée directement. C'est, par exemple, ainsi qu'on procède lorsqu'on emploie la méthode indiquée en 3° au numéro précédent ;

Soit opérer par différence, absolument comme avec le niveau d'eau.

Le nivellement indirect peut ainsi être considéré comme une opération tout à fait analogue au nivellement direct, le plan horizontal mené par le point m considéré aux deux numéros précédents remplaçant le plan de visée du niveau d'eau.

Seulement, au lieu que ce dernier devait toujours être *au-dessus* des points visés, la différence de cote ne dépassant d'ailleurs pas 4 mètres, on peut mesurer, par nivellement indirect, des dénivellations de sens et de grandeurs quelconques. Le nivellement indirect pourra donc être *simple* dans des circonstances où le nivellement direct serait forcément *composé.*

Mais il est, de plus, une série de circonstances (mesure de la hauteur d'un clocher, d'une montagne inaccessible, etc., etc.) où le nivellement direct est impraticable, au lieu que le nivellement indirect s'applique à tout point visible.

582. Ayant ainsi déterminé les différences mutuelles des cotes, il suffit, pour connaître ces cotes elles-mêmes, de se donner l'une d'entre elles. Cette dernière inconnue dépend d'ailleurs évidemment du choix du plan horizontal de comparaison, de sorte qu'on peut la considérer comme arbitraire. Par exemple, on pourra faire passer le plan de comparaison par le point le plus bas du terrain donné et, par conséquent, prendre pour ce point la cote zéro.

Pour changer de plan de comparaison, il suffit d'ailleurs d'augmenter ou de diminuer toutes les cotes d'une même quantité, égale à la distance de l'ancien plan au nouveau. C'est ce qu'on devra faire, par exemple, lorsqu'on aura nivelé, indépendamment l'un de l'autre et avec des plans de comparaison différents, deux terrains voisins et qu'on voudra réunir en un seul les deux nivellements.

Si enfin on veut comparer entre eux des nivellements portant sur des régions très différentes, il importe d'avoir un niveau (1) de comparaison commun pour toutes ces opérations. On compte alors les cotes au-dessus du *niveau de la mer :* les cotes ainsi estimées prennent le nom d'*altitudes* des points correspondants.

On a effectué, par des opérations de précision, un *nivellement général* du territoire français, de manière à faire connaître les altitudes d'un très grand nombre de points. Pour rapporter un nivellement quelconque au niveau de la mer, il suffira donc d'y rattacher un point qui ait fait partie du nivellement général.

583. **Représentation des cotes. Courbes de niveau.**

Un plan étant rapporté sur le papier, il faut encore arriver à donner une idée nette du relief du terrain ainsi figuré. On arrive à ce résultat à l'aide des *courbes de niveau.*

On appelle *courbe de niveau* (*fig.* 496, 501) le lieu des points qui ont une même cote donnée, autrement dit la section de la surface du terrain par un plan horizontal.

On détermine une telle courbe en plantant (à l'aide du niveau d'eau) des piquets en des points suffisamment rapprochés les uns des autres et ayant la cote donnée, puis relevant, par intersections ou par toute autre méthode de planimétrie, les positions de ces piquets.

On représente le relief d'un terrain en figurant les courbes de niveau dont les cotes diffèrent les unes des autres d'une même quantité ou *équidistance* (dépendant, bien entendu, de la nature du terrain et de l'échelle du plan).

---

(1) Nous disons *niveau* de comparaison et non *plan* de comparaison, puisque, pour des points très éloignés les uns des autres, il n'est plus possible d'admettre le parallélisme des plans horizontaux.

Si l'équidistance a été convenablement choisie, la forme et la disposition des courbes de niveau permettront de se rendre compte de toutes les particularités du relief. En particulier, le terrain sera d'autant plus incliné que les courbes de niveau, pour une même équidistance, seront plus rapprochées les unes des autres.

Sur beaucoup de cartes topographiques, le relief est figuré par des hachures tracées entre les courbes de niveau successives, normalement à ces courbes, par conséquent d'autant plus courtes que ces dernières sont plus voisines les unes des autres ou que la pente est plus forte, les hachures étant d'ailleurs d'autant plus rapprochées et plus grosses qu'elles sont plus courtes. Ce mode de figuration a toutefois l'inconvénient de surcharger la carte sans donner aucun renseignement que ne fournisse le tracé des courbes de niveau.

Indépendamment du figuré général du terrain, la carte porte indication des cotes d'un certain nombre de points importants.

584. Lorsqu'on veut représenter les ondulations du terrain le long d'une ligne donnée dont la projection est figurée en L (*fig.* 496) sur le plan, on a recours aux *profils*.

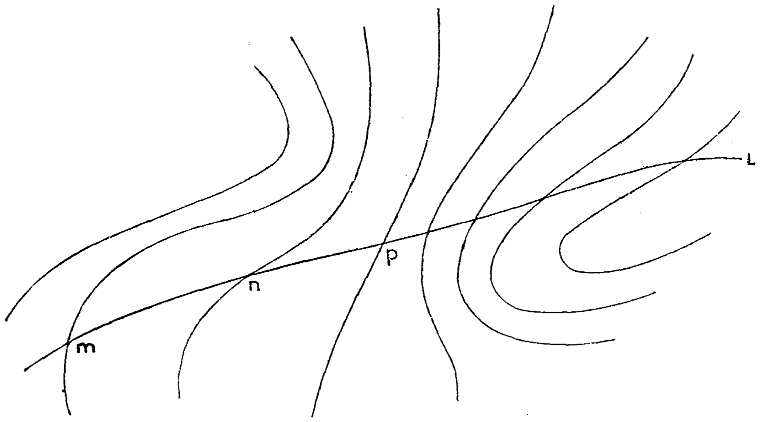

Fig. 496.

On nomme *profil* la figure obtenue en développant (**538**) le cylindre qui projette horizontalement la ligne considérée. Il est clair que si L est une ligne droite, le profil n'est autre que la section même du terrain par le plan vertical qui projette cette droite.

Puisque le tracé des courbes de niveau donne, à lui seul, la forme du terrain, il doit permettre de construire le profil suivant une ligne quelconque L.

On remarquera, à cet effet, qu'on connaît les ordonnées ou cotes d'un certain nombre de points m, n, p, ... (*fig.* 496) de la ligne L, à savoir, des points où cette ligne coupe les courbes de niveau. On relèvera, sur le plan, les abscisses curvilignes des mêmes points, que l'on reportera en O_1m_1, O_1n_1, O_1p_1, ... (*fig.* 496 *bis*), et l'on aura ainsi, aux extrémités des ordonnées m_1M_1, n_1N_1, ... une série de points du profil cherché.

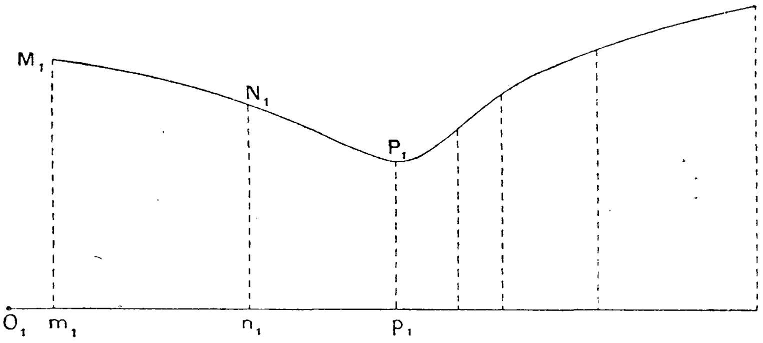

Fig. 496 *bis*.

585. Pour terminer ce qui concerne la planimétrie et le nivellement, nous dirons un mot de quelques simplifications qui peuvent être apportées dans ces deux opérations.

Les *méthodes photographiques*, qui présentent l'avantage important de réduire à la plus courte durée possible les opérations à effectuer sur place, sont fréquemment appliquées à la topographie. La place occupée par l'image d'un objet sur l'épreuve photographique dépend, en effet, de la direction du rayon lumineux joignant ce point au centre optique de la lentille objective. On peut dès lors en déduire, moyennant certaines précautions dans le détail desquelles nous n'entrerons pas, l'azimuth et la hauteur de cette direction. Il suffira donc d'avoir deux photographies, prises de deux points séparés l'un de l'autre par une distance connue, pour en déduire, par *intersections*, le levé et le nivellement du terrain ainsi doublement reproduit.

586. Nous avons, d'autre part, remarqué, à propos de la méthode par intersections, que nous ne pouvions pas obtenir la position relative de deux points A, B en visant simplement le point B du point A et qu'il était nécessaire de compléter cette visée, soit en chaînant l'alignement AB, soit en rattachant le point B à un *troisième* point C, non en ligne droite avec les deux premiers.

Il existe cependant un moyen d'échapper à cette nécessité. C'est de se servir d'un principe emprunté au *Cours de Cosmographie* (¹) :

Les diamètres apparents d'un même objet, vu à différentes distances, sont inversement proportionnels à ces distances.

D'après cela, si un aide se transporte au point B avec un *voyant* de dimensions connues, l'observation du diamètre apparent sous lequel ce voyant apparaît du point A fera connaître la distance AB.

Il faut observer que c'est la distance AB elle-même qui est ainsi mesurée et non pas (comme lorsqu'on opère à la chaîne) sa projection horizontale AH (*fig.* 497). Mais on pourra évaluer cette dernière si on lit la hauteur du rayon visuel AB, puisque, dans le triangle rectangle ABH, on connaîtra l'hypoténuse et un angle aigu.

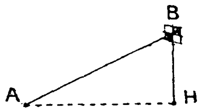

Fig. 497.

On peut, en se fondant sur ce principe, lire et noter sur le papier, en même temps, l'azimuth, la hauteur et la longueur de la droite qui joint un point quelconque M à un point O choisi une fois pour toutes, et, par conséquent, obtenir par *rayonnement* le levé et le nivellement d'un terrain, avec une seule visée pour chaque point.

---

(1) Tisserand et Andoyer, *Leçons de Cosmographie*, page 88.

CHAPITRE III

ARPENTAGE

587. L'*arpentage* a pour objet de trouver l'aire d'un terrain, ou, plus exactement, l'aire de sa projection horizontale. C'est, en général, cette dernière qui est utile à connaître. Les terrains que l'on arpente sont, en effet, le plus ordinairement destinés à la culture ou à la bâtisse : or, comme c'est verticalement que les végétaux croissent et que les constructions sont élevées, les dimensions des édifices, ou le nombre de plantes d'une espèce déterminée, qui peuvent trouver place sur un terrain donné, ne dépendent que de la projection horizontale de ce terrain.

588. Toute méthode par laquelle on a pu lever le plan d'un terrain permet, par cela même, de l'arpenter. Pour mesurer, en effet, l'aire d'un terrain polygonal, on le considérera comme une somme de triangles et, du moment que l'on aura les éléments nécessaires pour déterminer chaque triangle (lesquels seront fournis par le levé), les formules de la trigonométrie permettront d'en calculer la surface.

Si l'on a mesuré les trois côtés de chaque triangle, on en aura l'aire par la formule (Pl., **251**)

$$S = \sqrt{p(p-a)(p-b)(p-c)}.$$

589. L'arpentage est le plus ordinairement effectué par des opérateurs qui ne sont pas en état de se livrer à des calculs trigonométriques : l'instrument employé est alors l'équerre d'arpenteur.

Pour arpenter un polygone, on pourrait le décomposer en triangles dont on mesurerait la base et (avec l'équerre) la hauteur.

Mais on préfère opérer différemment et faire choix tout d'abord d'une droite *xy* (*fig.* 498) ou *base*, par exemple la plus grande diagonale AE du polygone, sur laquelle on abaisse des perpendiculaires B*b*, C*c*, D*d*, F*f*, G*g* (*fig.* 498) de tous les sommets. On a ainsi décomposé le polygone en triangles (AB*b*, AG*g*, D*d*E, F*f*E, *fig.* 498) et en trapèzes rectangles (B*b*C*c*, C*c*D*d*, G*g*F*f*, *fig.* 498), dont on sait évaluer les aires par les formules connues.

Sur la figure 498, tous les triangles et trapèzes sont *additifs :* il peut en être autrement : sur la figure 498 *bis*, par exemple, l'aire du polygone s'obtient en retranchant, de la somme AB*b* + B*b*C*c* + C*c*D*d* + D*d*E*e* + FG*g* + G*g*H*h* + K*k*A, la somme FE*e* + H*h*K*k*.

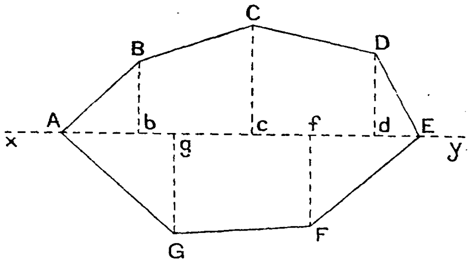

FIG. 498.

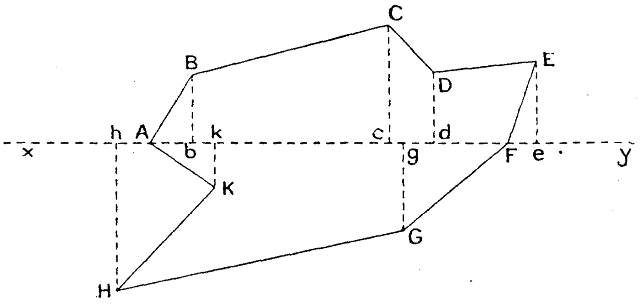

FIG. 498 *bis*.

Il est clair que, dans cette manière d'arpenter, on opère comme s'il s'agissait d'un levé à l'équerre, la base jouant le rôle d'axe des abscisses.

590. Quant au cas où le terrain est limité par un contour curviligne, il se ramène au précédent en remplaçant la courbe par un polygone d'un nombre suffisamment grand de côtés.

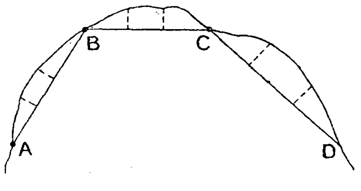

FIG. 499.

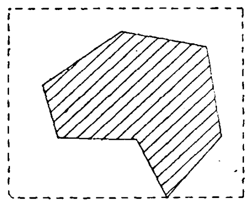

FIG. 500.

On commence par inscrire un premier polygone ABCD... (*fig.* 499) d'un nombre de côtés peu élevé, que l'on arpente comme nous l'avons dit au numéro précédent, puis on évalue la partie comprise entre la courbe et chaque côté en prenant ce côté pour base (1).

---

(1) Il existe des formules d'approximation permettant d'arriver, pour ces aires mixtilignes, à des résultats un peu plus exacts que ceux que l'on obtiendrait par la sommation pure et simple des trapèzes.

591. Il peut être nécessaire d'arpenter un terrain dans lequel on ne peut pénétrer (*fig.* 500). On entoure alors ce terrain d'un polygone facile à mesurer, par exemple d'un rectangle (*fig.* 500) et l'on arpente la partie de ce dernier comprise à l'extérieur du terrain proposé : il ne reste plus qu'à faire la différence des deux résultats.

592. Cubage.

On peut se proposer d'évaluer le volume d'un terrain compris à l'intérieur du cylindre qui projette sur un plan horizontal déterminé une aire S (*fig.* 501) prise sur ce terrain.

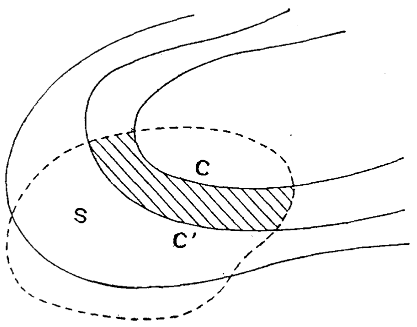

Fig. 501.

Supposons d'abord que la surface S soit comprise entre deux courbes de niveau consécutives C, C'. Alors on pourra, avec une approximation suffisante, considérer tous ses points comme ayant au-dessus du plan horizontal donné, une même cote, par exemple une moyenne entre la cote de C et la cote de C'. Le volume à évaluer sera alors celui d'un cylindre ayant pour hauteur cette cote et pour base, la projection horizontale de S.

Si maintenant l'aire S est quelconque, on pourra toujours (*fig.* 501) la décomposer en parties comprises entre courbes de niveau consécutives et que l'on traitera comme il vient d'être dit.

592 *bis*. Le plus ordinairement, toutefois, l'aire S a la forme d'une bande étroite tracée dans le voisinage d'une certaine ligne L (*fig.* 502). C'est ce qui arrive lorsque le cubage proposé est destiné à l'établissement d'une route.

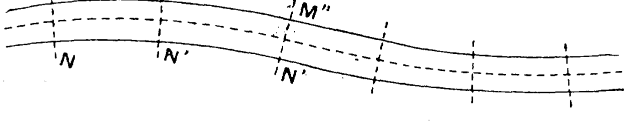

Fig. 502.

Dans ce cas, la méthode que nous venons d'indiquer n'est pas la plus commode : il convient plutôt de déterminer une série de *profils* *en travers*, c'est-à-dire de profils (**584**) suivant des alignements MN, M'N'... normaux à L et de considérer le volume compris entre chacun de ces profils et le suivant comme un cylindre ayant pour hauteur la distance des deux profils et pour base l'un d'eux.

EXERCICES

852. Vérifier si une équerre d'arpenteur est juste.

853. Mener, à un alignement donné, une perpendiculaire par un point extérieur sans employer d'autre instrument que la chaîne et sans tâtonnement (appliquer Pl., **126**).

854. Diviser un alignement en deux parties égales, à l'aide du graphomètre seul.

854 *bis*. Porter sur un alignement donné, à partir d'un point donné, une longueur égale à celle d'un autre alignement donné à l'aide du graphomètre seul.

855. Un niveau d'eau a un tube horizontal de 1 mètre de long, porté par un axe qui aboutit à 1 centimètre du milieu du tube. On a installé l'appareil de manière que l'axe passe avec la verticale un angle dont la tangente est $\frac{1}{10}$. Calculer approximativement la variation maximum du plan de visée, lorsque l'appareil tourne autour de son axe. Le résultat obtenu dépend-il de la longueur du tube?

856. Prolonger un alignement au delà d'un obstacle qui arrête la vue.

857. Montrer qu'on peut remplacer le niveau d'eau par un miroir plan dont on peut faire varier la direction en la maintenant verticale.

858. Montrer que le point M (fig. 485, n° **571**) d'un plan, tel que les droites qui le joignent à trois points donnés A, B, C fassent entre elles des angles donnés, peut s'obtenir de la manière suivante : on construit sur BC, CA, AB comme bases, des triangles BC*a*, CA*b*, AB*c* tous semblables au triangle T dont les angles sont égaux aux angles donnés ou à leurs suppléments (les sommets homologues étant pris convenablement chaque fois). Les droites A*a*, B*b*, C*c* concourent en un même point, qui répond à la question. Les points *a*, *b*, *c* ne sont autres que ceux où les droites MA, MB, MC rencontrent respectivement les segments de cercles MBC, MCA, MAB.

COMPLÉMENTS DE GÉOMÉTRIE DANS L'ESPACE

CHAPITRE PREMIER

CENTRE DES DISTANCES PROPORTIONNELLES

593. Problème. — *Étant données, en grandeur et signe, les abscisses* IA *et* IB *de deux points* A, B *d'une droite, rapportées à un point* I *de la même droite, trouver l'abscisse du point* M *qui divise le segment* AB *dans un rapport donné (en grandeur et en signe)* $\frac{\mathrm{MA}}{\mathrm{MB}} = -\frac{q}{p}$.

On a

$$\mathrm{MA} = \mathrm{IA} - \mathrm{IM}$$
$$\mathrm{MB} = \mathrm{IB} - \mathrm{IM}$$

et, par conséquent,

$$\frac{\mathrm{IA} - \mathrm{IM}}{\mathrm{IB} - \mathrm{IM}} = -\frac{q}{p}.$$

Cette équation, où tout est connu, sauf OM, est du premier degré par rapport à cette quantité. Elle fait connaître l'abscisse cherchée

$$\mathrm{IM} = \frac{p.\overline{\mathrm{IA}} + q.\overline{\mathrm{IB}}}{p + q}.$$

Comme nous devions nous y attendre (Pl., **110**), cette expression est infinie pour $q = -p$, et dans ce cas seulement.

Corollaire. — Par le point I, faisons passer une droite ou un plan quelconque (*fig.* 503) et soient $\mathrm{AA}_1 = a$, $\mathrm{BB}_1 = b$, $\mathrm{MM}_1 = m$ les distances des points A, B, M à cette droite ou à ce plan. Nous supposons d'ailleurs ces distances comptées en grandeur et signe, un sens positif ayant été choisi sur leur direction commune. On aura alors

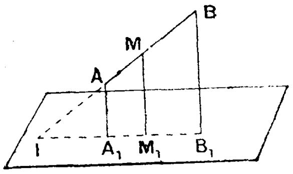

Fig. 503.

$$m = \frac{pa + qb}{p + q}.$$

En effet, les grandeurs a, b, m sont, en grandeur et signe (Pl., **190**), proportionnelles à IA, IB, IM. L'équation précédente étant homogène par rapport à ces dernières quantités, on peut donc les y remplacer par a, b, m : ce qui donne le résultat annoncé.

Remarque I. — Ce résultat ne serait d'ailleurs aucunement mis en défaut si les distances a, b, m, au lieu d'être comptées sur des perpendiculaires au plan considéré, étaient comptées suivant des parallèles à une même direction quelconque (non parallèle à ce plan).

Remarque II. — Nous savons (Pl., **113**) que le rapport $\frac{M_1A_1}{M_1B_1}$ a également la même valeur $\frac{-q}{p}$.

594. Soient donnés un certain nombre de points A, B, C, D, ... (fig. 504) et les nombres p, q, r, s, ..., positifs ou négatifs, mais non tous nuls, correspondant respectivement à ces points. Soient M le point qui divise AB dans le rapport $-\frac{q}{p}\left(\frac{MA}{MB} = -\frac{q}{p}\right)$; M′, le point qui divise MC dans le rapport $-\frac{r}{p+q}\left(\frac{M'M}{M'C} = -\frac{r}{p+q}\right)$; M″, le point qui divise M′D dans le rapport $\frac{-s}{p+q+r}\left(\frac{M''M'}{M''D} = -\frac{s}{p+q+r}\right)$; etc.

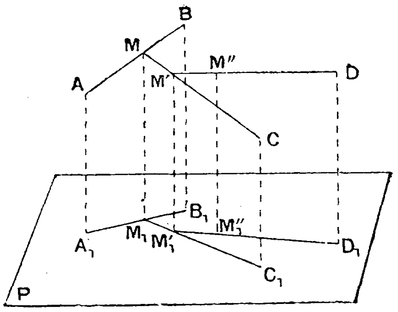

Fig. 504.

Si a, b, c, d, ... sont les distances des points A, B, C, D, ... à un plan quelconque P, la distance m du point M au même plan sera (n° précédent)

$$m = \frac{pa + qb}{p + q}.$$

La distance m' du point M' à ce plan sera

$$m' = \frac{(p+q)m + rc}{(p+q)+r} = \frac{pa + qb + rc}{p+q+r};$$

de même la distance du point M'' sera

$$m'' = \frac{(p+q+r)m' + sd}{(p+q+r)+s} = \frac{pa + qb + rc + sd}{p+q+r+s}.$$

Continuons ainsi jusqu'à ce que nous ayons utilisé tous les points donnés et chacun d'eux une seule fois; nous obtiendrons un point O qu'on appelle le *centre des distances proportionnelles* des points A, B, C, D, ... affectés des coefficients p, q, r, s, ...

La distance de ce point au plan P est

$$o = \frac{pa + qb + rc + sd + \ldots}{p+q+r+s+\ldots}.$$

Nous avons dit que l'on devait utiliser chacun des points donnés une seule fois; mais nous n'avons pas spécifié l'*ordre* dans lequel on les faisait intervenir. Il semblerait, au premier abord, que la position du point final O dût dépendre de l'ordre en question. Le résultat qui précède montre qu'il n'en est rien. La valeur trouvée pour la distance du point O au plan P est, en effet, indépendante de cet ordre (pourvu, bien entendu, qu'à chacun des points donnés corresponde toujours le même coefficient) et, par conséquent, si, en changeant l'ordre en question, on trouvait un point O' différent du premier, la droite OO' devrait être parallèle au plan P, et cela quel que soit le plan P, ce qui est absurde.

Remarque. — Soient A_1, B_1, C_1, M_1, M'_1,.., O_1 les projections sur le plan P, des points A, B, C... M, M'... O (*fig.* 504). Le point M_1 divisera A_1B_1 dans le rapport $-\frac{q}{p}$; le point M'_1 divisera C_1M_1 dans le rapport $-\frac{r}{p+q}$; etc. En un mot, le *point* O_1 *sera le centre des distances proportionnelles des points* A_1, B_1, C_1,... *affectés des mêmes coefficients que les points donnés* A, B, C... et la même chose a lieu si les droites AA_1, BB_1,... au lieu d'être perpendiculaires au plan P, sont parallèles à une même direction quelconque.

595. Le *centre des distances proportionnelles* n'est évidemment autre chose que ce que l'on nomme en statique le *centre des forces parallèles*, les points donnés étant les points d'application des forces, et les intensités de celles-ci étant représentées par les coefficients p, q, r,... lesquels sont positifs ou négatifs suivant le sens des forces correspondantes.

596. Il y a une circonstance où la suite d'opérations que nous venons d'indiquer tomberait en défaut, c'est celle où l'on aurait à diviser une droite dans un rapport égal à $+1$, par exemple, où les deux premiers coefficients seraient égaux et de signes contraires. Il est aisé de voir qu'on peut toujours tourner la difficulté en intervertissant l'ordre des points donnés, sauf dans un cas, celui où la somme de *tous* les coefficients donnés est nulle. Dans le langage de la statique, ce cas correspond à celui où les forces données équivalent à un couple[1].

597. Lorsque tous les coefficients sont égaux à $+1$, le centre des distances proportionnelles prend le nom de *centre des moyennes distances*. Le centre des moyennes distances d'un système de points peut donc être considéré comme défini par cette propriété que sa distance à un plan quelconque est la moyenne arithmétique des distances des points donnés.

598. Nous allons considérer en particulier les cas où les points donnés sont au nombre de deux, trois ou quatre.

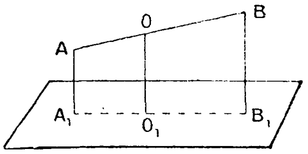

Fig. 505.

Supposons d'abord qu'il n'y ait que deux points A, B (*fig.* 505) affectés des coefficients p, q. Le centre des distances proportionnelles est alors un point de la droite AB. D'autre part, A et B étant donnés, *ce centre peut coïncider avec un point quelconque* O *de la droite* AB, *pour un choix convenable de p et de q :* il suffit évidemment de prendre

$$\frac{q}{p} = -\frac{\mathrm{OA}}{\mathrm{OB}}.$$

---

[1] Dans le cas où les forces données se font équilibre, le centre des distances proportionnelles n'est plus rejeté à l'infini, mais indéterminé. Dans la suite d'opérations indiquées dans le texte, on arriverait encore finalement à deux points affectés de coefficients égaux et de signes contraires, mais ces deux points coïncideraient entre eux.

Introduisons le nombre t donné par la relation

$$p + q + t = 0;$$

la relation précédemment obtenue

$$o = \frac{pa + qb}{p + q}$$

entre les distances des points A, B, O, devient

$$pa + qb + to = 0.$$

Ainsi, *étant donnés trois points quelconques en ligne droite, on peut trouver trois nombres non tous nuls, mais dont la somme est nulle et qui donnent lieu à la relation précédente entre les distances des trois points donnés à un plan quelconque.*

La condition précédente détermine d'ailleurs complètement, sinon les nombres p, q, t eux-mêmes, du moins leurs rapports mutuels : on a (puisque $\frac{p}{q} = -\frac{\text{IA}}{\text{IB}}$ et que $p + q + t = 0$)

$$\frac{p}{\text{BO}} = \frac{q}{\text{OA}} = \frac{p + q}{\text{BO} + \text{OA}} = -\frac{t}{\text{BA}} = \frac{t}{\text{AB}}.$$

599. Soient à présent trois points A, B, C (*fig.* 506), que nous supposerons former un triangle ; soient encore p, q, r les coefficients correspondants.

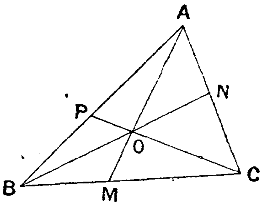

Fig. 506.

Si nous divisons BC dans un rapport $\frac{\text{MB}}{\text{MC}} = -\frac{r}{q}$, le centre des distances proportionnelles cherché O est situé sur la droite AM. On voit que *ce point est nécessairement dans le plan* ABC.

En divisant de même CA dans un rapport $\frac{\text{NC}}{\text{NA}} = -\frac{p}{r}$ et AB dans un rapport $\frac{\text{PA}}{\text{PB}} = -\frac{q}{p}$, les droites BN, CP donnent deux nouveaux lieux du point O.

Il résulte de là une nouvelle démonstration du théorème du n° **197** (Pl., Compl. du livre III).

Si, en effet, les côtés d'un triangle ABC sont divisés par les points M, N, P (*fig.* 506), de manière que $\frac{MB}{MC} \cdot \frac{NC}{NA} \cdot \frac{PA}{PB} = -1$, on peut évidemment poser

$$\frac{MB}{MC} = -\frac{r}{q}$$

$$\frac{NC}{NA} = -\frac{p}{r}$$

et, en vertu de la relation $\frac{MB}{MC} \cdot \frac{NC}{NA} \cdot \frac{PA}{PB} = -1$, l'on aura nécessairement

$$\frac{PA}{PB} = -\frac{q}{p}.$$

Les remarques précédentes montrent dès lors que les trois droites AM, BN, CP concourent en un même point, à savoir le centre des distances proportionnelles des points A, B, C.

599 *bis.* S'il s'agit du *centre des moyennes distances* des trois points A, B, C, les droites AM, BN, CP sont les médianes du triangle formé par ces points. Le centre des moyennes distances n'est donc autre chose que le point nommé précédemment *centre de gravité* du triangle ABC.

600. Coordonnées barycentriques planes.

Inversement, soient donnés le triangle ABC et le point O de son plan. Joignons OA, et soit M le point où cette droite coupe BC. Si nous prenons deux nombres q, r, satisfaisant à la relation

$$\frac{MB}{MC} = -\frac{r}{q},$$

puis un nombre p déterminé par la proportion

$$\frac{OM}{OA} = -\frac{p}{q+r},$$

le point O sera le centre des distances proportionnelles des points A, B, C, affectés des coefficients p, q, r.

Ainsi, *étant donnés quatre points dont les trois premiers ne sont pas* *en ligne droite, on peut affecter ceux-ci de coefficients tels que leur centre des distances proportionnelles soit le quatrième point donné.*

Le raisonnement semble en défaut lorsque la droite OA est parallèle à BC; mais on pourra alors le recommencer, soit en considérant la droite OB et le côté CA, soit en considérant la droite OC et le côté AB (1).

Les rapports mutuels des quantités p, q, r sont d'ailleurs déterminés quand on donne le point O (puisque ce sont les rapports dans lesquels les droites OA, OB, OC divisent respectivement les côtés BC, CA, AB).

On donne le nom de *coordonnées barycentriques* du point O, par rapport au *triangle de référence* ABC, aux coefficients p, q, r ainsi choisis. Les coordonnées barycentriques d'un point déterminé, par rapport à un triangle déterminé, sont déterminées à un facteur commun près.

Inversement, à tout système de coordonnées p, q, r, tel que la somme $p+q+r$ soit différente de zéro, correspond un point déterminé situé dans le plan du triangle.

Remarque. — Si l'on mène, par les points A, B, C, O, des parallèles à une même direction quelconque, jusqu'à rencontre en A_1, B_1, C_1, O_1 avec un plan P, le point O_1 aura, en vertu de la définition précédente et du n° **594** (Remarque), les mêmes coordonnées barycentriques, par rapport au triangle $A_1 B_1 C_1$, que le point O par rapport au triangle ABC.

601. La considération des aires permet de donner une interprétation simple des coordonnées barycentriques planes.

Considérons, en effet, les deux triangles ABM, ACM. Ces deux triangles, qui ont évidemment même hauteur, sont entre eux comme leurs bases MB, MC. Pour la même raison, ce même rapport $\frac{\mathrm{MB}}{\mathrm{MC}}$ est égal au rapport des triangles OMB, OMC. Il vient donc, en vertu d'un théorème connu sur les proportions

$$\frac{\mathrm{MB}}{\mathrm{MC}}=\frac{\mathrm{AMB}}{\mathrm{AMC}}=\frac{\mathrm{OMB}}{\mathrm{OMC}}=\frac{\mathrm{AOB}}{\mathrm{AOC}}$$

---

(1) Il est évidemment impossible que les droites qui joignent le point O aux sommets du triangle ABC soient toutes trois parallèles aux côtés du même triangle, puisque les parallèles menées aux côtés d'un triangle par les sommets opposés ne passent pas par un même point.

puisque le triangle AOB équivaut à la somme ou à la différence des triangles AMB et OMB, le triangle AOC étant en même temps équivalent à la somme ou à la différence des triangles AMC et OMC.

Le rapport $\frac{\text{AOB}}{\text{AOC}}$ est donc égal, en valeur absolue, au rapport $\frac{r}{q}$. Mais, de plus, nous pouvons remarquer que si la droite AO coupe le côté BC lui-même (c'est-à-dire si $\frac{r}{q}$ est positif) les deux triangles AOB, AOC sont tous deux additifs (voir Pl.; Note D, **315**, *fig*. 228-229) ou tous deux soustractifs (*ibid.*, *fig*. 230), au lieu que si la droite AO coupe l'un des prolongements de BC, ces deux triangles sont l'un additif, l'autre soustractif, en même temps que le rapport $\frac{r}{q}$ deviendra négatif. Si donc on fait précéder chacun des angles BOC, COA, AOB du signe $+$ ou du signe $-$ suivant qu'il est additif ou soustractif, *le rapport* $\frac{r}{q}$ *sera, en grandeur et signe, égal au rapport des triangles* AOB, AOC.

En faisant des raisonnements analogues relativement aux autres côtés du triangle ABC, on arrive évidemment à la conclusion suivante :

Les coordonnées barycentriques d'un point sont, en grandeur et signe (moyennant la convention qui vient d'être établie), *proportionnelles aux aires des triangles qui ont pour sommet commun ce point et, pour bases respectives, les côtés du triangle de référence.*

602. Introduisons le nombre t défini par la relation

$$p + q + r + t = 0$$

Alors la relation, précédemment trouvée, entre les distances a, b, c, o des points A, B, C, O à une droite quelconque du plan ABC ou à un plan quelconque

$$o = \frac{pa + qb + rc}{p + q + r}$$

peut s'écrire

$$pa + qb + rc + to = 0.$$

Ainsi, *quatre points quelconques étant donnés dans un plan, on peut trouver quatre nombres p, q, r, t dont la somme est nulle et tels* *que les distances des quatre points donnés à un plan quelconque, multipliées respectivement par ces quatre nombres, donnent une somme nulle.*

Le raisonnement précédent cesse d'être valable lorsque les trois points A, B, C sont en ligne droite. Mais alors nous savons (**598**) qu'on peut trouver les nombres p, q, r tels que l'on ait

$$pa + qb + rc = 0,$$

ce qui est la relation précédente, t étant pris égal à zéro.

603. Coordonnées barycentriques dans l'espace. — Considérons maintenant le cas où l'on part de quatre points A, B, C, D (*fig.* 507) non situés dans un même plan. p, q, r, s étant les coefficients affectés à ces points, on pourra commencer par prendre le centre des distances proportionnelles des points A, B, C affectés des coefficients p, q, r, — autrement dit, le point M du plan ABC qui a pour coordonnées barycentriques, par rapport à ce triangle, les nombres p, q, r, — puis prendre sur DM le point O tel que

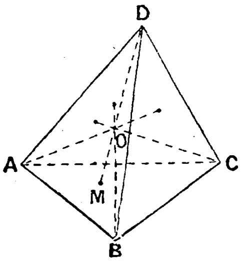

Fig. 507.

$$\frac{\text{OM}}{\text{OD}} = -\frac{s}{p+q+r}.$$

Par exemple, si les coefficients p, q, r, s sont tous égaux à $+1$, le point O appartiendra à la droite qui joint le point D au centre de gravité du triangle ABC, de sorte que cette droite et les trois autres droites joignant le point A au centre de gravité du triangle BCD, le point B au centre de gravité de CDA, le point C au centre de gravité de DAB, concourent en un même point.

Inversement, O (*fig.* 507) étant un point quelconque de l'espace, on pourra toujours déterminer les coefficients p, q, r, s de manière que le centre des distances proportionnelles obtenu soit le point O. Il suffira de prendre pour p, q, r les coordonnées barycentriques, par rapport au triangle ABC, du point M où la droite DO coupe le plan ABC, et pour s le nombre défini par la relation écrite tout à l'heure.

Les coefficients p, q, r, s, qui sont ainsi déterminés à un facteur commun près, sont dits les *coordonnées barycentriques* du point O, par rapport au *tétraèdre de référence* ABCD.

Nous laissons au lecteur le soin de démontrer que ces coordonnées barycentriques sont proportionnelles aux quatre tétraèdres ayant pour sommet commun le point considéré et pour bases respectives les faces du tétraèdre de référence, et que cette proportionnalité a lieu en grandeur et signe, si l'on a soin d'affecter les tétraèdres en question du signe + ou du signe — suivant qu'ils sont additifs ou soustractifs (¹).

604. Introduisons le nombre t défini par la relation

$$p+q+r+s+t=0.$$

La relation

$$o=\frac{pa+qb+rc+sd}{p+q+r+s},$$

qui a lieu entre les distances des cinq points A, B, C, D, O à un plan quelconque, pourra s'écrire

$$pa+qb+rc+sd+to=0.$$

Donc, étant donnés cinq points de l'espace, on peut trouver les cinq coefficients p, q, r, s, t donnant lieu à la relation précédente. On prouvera d'ailleurs, comme tout à l'heure (**602**), que la conclusion n'est pas en défaut si les points A, B, C, D cessent de former un tétraèdre.

605. On définit en mécanique le *centre de gravité* d'une figure quelconque. Cette définition, dans le cas d'un polygone ou d'un polyèdre *homogène* (cette condition, absolument essentielle, est implicitement supposée dans ce qui va suivre) conduit aux conclusions suivantes :

Le centre de gravité d'un triangle est le point de concours de ses médianes, conformément à la dénomination que nous avons adoptée depuis le n° **55** *bis*.

Le centre de gravité d'un polygone plan quelconque s'obtient en le décomposant en triangles, dont on prend les centres de gravité, et prenant le centre des distances proportionnelles des points ainsi obtenus,

---

(1) Voir note F à la fin du volume.

affectés de coefficients respectivement égaux aux aires des triangles correspondants (¹).

Le centre de gravité d'un tétraèdre est le point où se coupent (**603**) *les droites qui joignent chaque sommet au centre de gravité de la face opposée.*

Le centre de gravité d'un polyèdre quelconque s'obtient en le décomposant en tétraèdres et opérant comme il vient d'être expliqué pour les polygones dans le plan (le mot *aire* étant remplacé par le mot *volume*).

606. Théorème. — *Si on coupe un prisme par différents plans, les centres de gravité des sections ainsi obtenues sont sur une même parallèle aux arêtes.*

Pour démontrer ce théorème, nous nous appuierons sur le lemme suivant :

Lemme. — *La surface d'un polygone situé dans un plan fixe* P *est, avec sa projection sur un autre plan fixe* P′, *dans un rapport constant, quel que soit le polygone.*

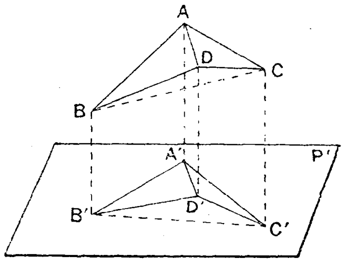

FIG. 508.

Dans la démonstration, nous distinguerons plusieurs cas :

1° Soient d'abord, dans le plan P, deux triangles ABD, ACD (*fig.* 508) qui ont un côté commun AD et dont les projections sur le plan P′ sont A′B′D′, A′C′D′. Si p, q, r sont les coordonnées barycentriques du point D, rapportées au triangle ABC, autrement dit si le point D est le centre des distances proportionnelles des points A, B, C affectés des coefficients p, q, r, le rapport $\frac{q}{r}$ sera celui des triangles ACD, ABD. Ce rapport sera donc égal à celui des triangles A′C′D′, A′B′D′, puisque le point D′ a, relativement au triangle A′B′C′, les mêmes coordonnées barycentriques que le point D, relativement au triangle ABC.

2° Ce qui précède montre que le rapport d'un triangle du plan P à sa projection sur P′ ne change pas si l'on remplace le triangle consi-

---

(1) Il résulte de la définition mécanique du centre de gravité, que ce point est indépendant du mode adopté pour la décomposition du polygone en triangles : c'est d'ailleurs ce qui peut se voir directement (exercice 869).

déré par un autre (situé également dans le plan P) ayant avec le premier un côté commun. Ce rapport est, par conséquent, le même pour deux triangles ABC, ADE (*fig.* 509) du plan P ayant un sommet commun, car on peut trouver un triangle (le triangle ABD) ayant avec ABC un côté commun et avec ADE un côté commun. Il est donc aussi le même pour deux triangles quelconques du plan P, puisqu'on peut toujours trouver un triangle ayant un sommet commun avec le premier et un sommet commun avec le second.

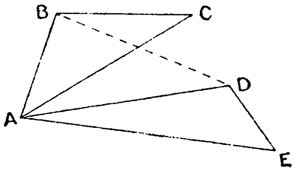

FIG. 509.

3° Si enfin il s'agit d'un polygone quelconque, on peut décomposer ce polygone en triangles. Le rapport en question est le même pour tous ces triangles et a, par conséquent, encore la même valeur pour le polygone entier, d'après un théorème connu sur les proportions (comparer Pl., **257**).

Corollaires. — I. *Le rapport de l'aire d'un polygone plan à celle de sa projection est égal au cosinus de l'angle des deux plans.*

Nous savons, en effet, que ce rapport est indépendant du polygone choisi. Nous prendrons, pour ce polygone, un triangle ABC ayant sa base BC sur l'intersection des deux plans P et P′ (*fig.* 510). Si BCA′ est la projection de ABC sur le plan P′ et que AH soit la hauteur du triangle ABC, A′H est (théorème des trois perpendiculaires), la hauteur du triangle BCA′. De plus, l'angle en H du triangle rectangle AA′H est égal à l'angle des deux plans P, P′ et, par conséquent, le rapport $\frac{\text{A}'\text{H}}{\text{AH}}$ est égal au cosinus de cet angle. Or ce rapport est égal au rapport des deux triangles A′BC, ABC, puisque ceux-ci ont même base.

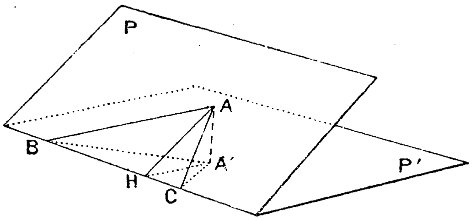

FIG. 510.

II. Le lemme précédent s'applique *à une aire curviligne quelconque* S *située dans le plan* P *et à sa projection.*

En effet, par définition (Pl., **260**, note), l'aire S (*fig.* 511) est la limite d'une aire polygonale inscrite *s*. Si S′, *s′* sont les projections de S, *s* sur le plan P′, lorsque le nombre des côtés du polygone *s* augmentera indéfiniment de manière que chacun d'eux tende vers zéro, les mêmes circonstances se présenteront pour le polygone s'.

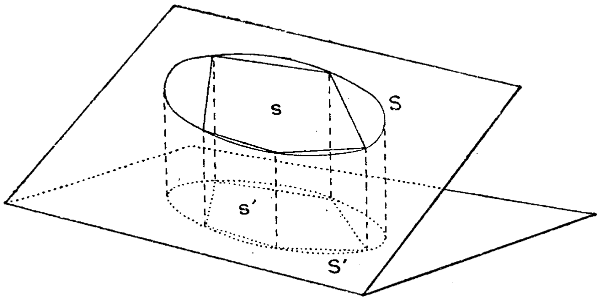

Fig. 511.

Or nous venons de voir qu'on a

$$s' = s \cos \alpha,$$

α étant l'angle des deux plans.

Donc, lorsque s tendra vers S, s' tendra vers une limite S' égale à S cos α.

C. Q. F. D.

607. A l'aide du lemme qui vient d'être démontré, il va nous être facile de démontrer le théorème précédemment énoncé.

Nous allons démontrer que *si l'on mène, dans une surface prismatique, une section droite et une section oblique, le centre de gravité de la section oblique se projette, sur le plan de la section droite, au centre de gravité de cette dernière.*

Si d'abord le prisme est triangulaire, le fait résulte (*fig.* 512) de la remarque du n° **594**, en vertu du n° **599** *bis*.

Si maintenant le prisme est polygonal, on peut le décomposer en prismes triangulaires (*fig.* 513) qui donneront, dans le plan de la section oblique, les triangles T_1, T_2, T_3,... et, dans le plan de la sec-

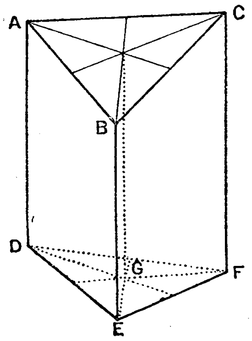

Fig. 512.

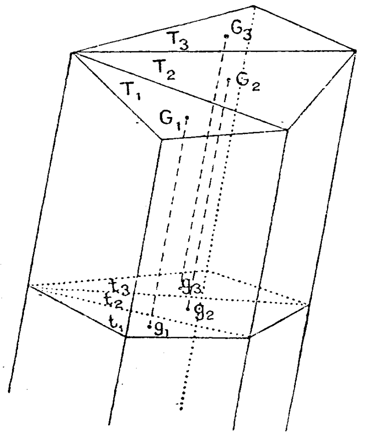

Fig. 513.

tion droite, les triangles t_1, t_2, t_3,... Soient G_1, G_2, G_3, les centres de gravité de T_1, T_2, T_3,...; g_1, g_2, g_3, les centres de gravité de t_1, t_2, t_3,... lesquels sont (comme nous venons de le voir) les projections de G_1, G_2, G_3, sur le plan de la section droite. Le centre de gravité G de la section oblique est le centre des distances proportionnelles des points G_1, G_2, G_3, affectés de coefficients proportionnels aux aires T_1, T_2, T_3,...; et, d'autre part, le centre de gravité g de la section droite est le centre des distances proportionnelles des points g_1, g_2, g_3,..., affectés des coefficients t_1, t_2, t_3,..., lesquels sont (lemme précédent) proportionnels aux premiers. Donc le point g est (**594**, Rem.) la projection de G.

Donc aussi, lorsque le plan de la section oblique varie, le point G décrit la parallèle aux arêtes menée par g.

C. Q. F. D.

608. Volume du tronc de prisme.

Théorème. — *Le volume d'un tronc de prisme quelconque est égal au produit de l'une des bases par la distance du centre de gravité de l'autre base au plan de la première.*

Nous distinguerons deux cas :

1° *Le tronc de prisme est triangulaire.* Soient ABC, DEF (*fig.* 512) les deux bases, G le centre de gravité du triangle DEF ; d, e, f, g, les distances des points D, E, F, G au plan ABC.

Le volume du tronc de prisme est (**409**)

$$\mathrm{ABC} \times \frac{d+e+f}{3}.$$

Mais (**599** *bis*) $\frac{d+e+f}{3}$ est égale à g. Le volume est donc bien mesuré par le produit ABC $\times\, g$.

2° *Le tronc de prisme est polygonal.* On le décomposera en troncs de prismes triangulaires dont les bases sont T_1, T_2, T_3,..., dans le plan P de la base inférieure ; T'_1, T'_2, T'_3,... dans le plan P′ de la base supérieure. Si g_1, g_2, g_3,... sont les distances des points G'_1, G'_2, G'_3,... au plan P, on a (1°), pour le volume du tronc :

$$V = g_1 T_1 + g_2 T_2 + g_3 T_3 + \cdots$$

Mais, si g est la distance, au plan P, du centre de gravité de la base supérieure, on a

$$g = \frac{g_1 T_1' + g_2 T_2' + g_3 T_3' + \ldots}{T_1' + T_2' + T_3' + \ldots} = \frac{g_1 T_1 + g_2 T_2 + g_3 T_3 + \ldots}{T_1 + T_2 + T_3 + \ldots}$$

puisque les triangles T_1', T_2', T_3'... sont proportionnels à T_1, T_2, T_3... (comme ayant les mêmes projections sur un plan de section droite). Tirant $g_1 T_1 + g_2 T_2 + g_3 T_3 + \ldots$ de cette égalité, il vient bien, conformément à l'énoncé,

$$V = g\,(T_1 + T_2 + T_3 + \ldots.).$$

Corollaires. — I. *Un tronc de prisme est équivalent au prisme que l'on obtiendrait en substituant à la base supérieure une section menée par le centre de gravité de cette base et parallèle à la base inférieure.*

Par conséquent, aussi :

II. *Le volume du tronc de prisme a pour mesure le produit de la section droite par la distance des centres de gravité des deux bases.*

III. Le théorème précédent montre bien que *la position trouvée, pour le centre de gravité d'un polygone*, par le procédé du n° **605**, *est indépendante du mode de décomposition du polygone en triangles.* En premier lieu, en effet, les raisonnements des n^os^ **607-608** ne supposent nullement que cette indépendance soit établie (ils sont valables en supposant que le prisme ait été décomposé *d'une manière déterminée* en prismes triangulaires). Dès lors, un polygone plan quelconque étant donné, considérons-le comme l'une des bases d'un tronc de prisme, l'autre base étant dans un plan quelconque P. Si S est l'aire de cette seconde base, g la distance du centre de gravité du polygone donné au plan P, V le volume du tronc de prisme, on aura (théor. préc.)

$$g = \frac{V}{S}.$$

Cette formule montre que la distance g du centre de gravité G au plan P est indépendante de la manière dont on a décomposé le polygone en triangles pour trouver le point G. Comme le plan P est de direction quelconque, il en résulte (comparer **594**) que la position du point G est indépendante de ce mode de décomposition.

609. Problème. — *Trouver le lieu des points tels que les carrés de leurs distances à des points donnés, multipliés respectivement par des coefficients donnés (positifs ou négatifs), aient une somme algébrique égale à une quantité donnée.*

Soient les points donnés A, B, C, D, ... et les coefficients correspondants (positifs ou négatifs) p, q, r, s, ... Nous avons à chercher le lieu des points tels que l'on ait

$$p\cdot\overline{MA}^2 + q\cdot\overline{MB}^2 + r\cdot\overline{MC}^2 + \ldots \qquad = k,$$

k étant un nombre donné.

Supposant la somme $p+q+r+\ldots$ différente de zéro, soit O le centre des distances proportionnelles des points donnés, affectés des coefficients p, q, r, s, ... Joignons OM et soit Oa (*fig.* 514) la projection de OA sur cette droite, le segment Oa étant compté en grandeur et signe avec OM comme sens positif.

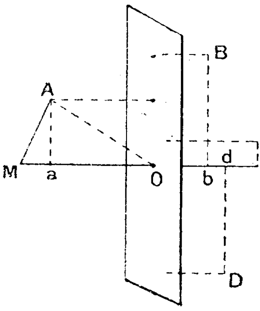

FIG. 514.

On aura, quelle que soit la position du point a,

$$\overline{MA}^2 = \overline{OM}^2 + \overline{OA}^2 - 2\overline{OM}\cdot\overline{Oa}.$$

Car si le point a est du même côté de O que le point M, le segment Oa est positif, l'angle $\widehat{AOM}$ aigu et l'on retrouve l'énoncé du n° **126**, 1° (Pl., livre III); si, au contraire, le point a est sur OM prolongé au delà de O, le segment Oa est négatif, l'angle $\widehat{AOM}$ obtus, et l'égalité précédente résulte de la seconde partie du même théorème.

Si, de même, Ob, Oc, Od ... sont les projections de OB, OC, OD, ... sur OM, comptées en grandeur et signe dans les mêmes conditions, on aura

$$\overline{MB}^2 = \overline{OM}^2 + \overline{OB}^2 - 2\overline{OM}\cdot\overline{Ob}$$
$$\overline{MC}^2 = \overline{OM}^2 + \overline{OC}^2 - 2\overline{OM}\cdot\overline{Oc}$$

. . . .

Multipliant l'égalité obtenue tout à l'heure par p, celles que nous venons d'écrire par q, r, s, ... respectivement, et ajoutant, il vient

$$\begin{aligned} &p\cdot\overline{\mathrm{MA}}^2+q\cdot\overline{\mathrm{MB}}^2+r\cdot\overline{\mathrm{MC}}^2+\dots \\ &=(p+q+r+\dots)\,\overline{\mathrm{OM}}^2+p\cdot\overline{\mathrm{OA}}^2+q\cdot\overline{\mathrm{OB}}^2+r\cdot\overline{\mathrm{OC}}^2+\dots \\ &\qquad -2\overline{\mathrm{OM}}\,(p\cdot\overline{\mathrm{O}a}+q\cdot\overline{\mathrm{O}b}+r\cdot\overline{\mathrm{O}c}+\dots). \end{aligned}$$

Mais, au second membre, la parenthèse du dernier terme

$$p\cdot\overline{\mathrm{O}a}+q\cdot\overline{\mathrm{O}b}+r\cdot\overline{\mathrm{O}c}+\dots$$

est nulle, en vertu du choix du point O; car les quantités Oa, Ob, Oc ... peuvent être considérées comme les distances (comptées en grandeur et signe) des points A, B, C, ... à un même plan mené par O, celui qui est perpendiculaire à OM (*fig.* 514).

L'égalité précédente se réduit donc à

$$p\overline{\mathrm{MA}}^2+q\cdot\overline{\mathrm{MB}}^2+r\cdot\overline{\mathrm{MC}}^2+\dots=(p+q+r+\dots)\,\overline{\mathrm{OM}}^2+l$$

en désignant par l la quantité $p\cdot\overline{\mathrm{OA}}^2+q\cdot\overline{\mathrm{OB}}^2+r\cdot\overline{\mathrm{OC}}^2+\dots$, *laquelle est indépendante de la position du point* M.

Ainsi écrite, cette égalité montre que la condition nécessaire et suffisante pour que la somme algébrique $p\cdot\overline{\mathrm{MA}}^2+q\cdot\overline{\mathrm{MB}}^2+r\cdot\overline{\mathrm{MC}}^2+\dots$ soit constante est que OM soit constant.

Donc le lieu demandé est une sphère de centre O.

610. Le raisonnement précédent serait en défaut si la somme $p+q+r+\dots$ était nulle. Dans ce cas, en effet, le lieu ne serait pas une sphère, mais un plan.

Pour le voir, divisons les points donnés en deux groupes, de manière que, dans chacun de ces deux groupes, la somme des coefficients soit différente de zéro (ce qui est évidemment possible dès que les coefficients ne sont pas tous nuls (¹)), le centre des distances proportionnelles du premier groupe étant O, celui du second O′. Supposons, par exemple, pour fixer les idées, qu'il y ait quatre points dont deux, A, B, appartiennent au premier groupe, les deux autres, C, D, au second. Les coefficients p, q, r, s sont alors tels que les deux sommes $p+q$, $r+s$, différentes de zéro, sont égales et de signes contraires.

---

(1) Il est clair que le problème cesserait de se poser si tous les coefficients étaient nuls.

Nous aurons, comme précédemment,

$$d \cdot \overline{MA}^2 + q \cdot \overline{MB}^2 = (p + q)\, \overline{OM}^2 + l$$
$$r \cdot \overline{MC}^2 + s \cdot \overline{MD}^2 = (r + s)\, \overline{O'M}^2 + l' = -(p + q)\, \overline{O'M}^2 + l'$$

en désignant par l, l' respectivement, les sommes $p \cdot \overline{OA}^2 + q \cdot \overline{OB}^2$, $r \cdot \overline{O'C}^2 + s \cdot \overline{O'D}^2$. En ajoutant ces deux égalités, il vient :

$$p \cdot \overline{MA}^2 + q \cdot \overline{MB}^2 + r \cdot \overline{MC}^2 + s \cdot \overline{MD}^2 = (p + q)(\overline{OM}^2 - \overline{O'M}^2) + l + l'.$$

Donc la condition $p \cdot \overline{MA}^2 + q \cdot \overline{MB}^2 + r \cdot \overline{MC}^2 + s \cdot \overline{MD}^2 =$ const. équivaut à $\overline{OM}^2 - \overline{O'M}^2 =$ const., c'est-à-dire donne pour lieu du point M un plan (¹).

611. Dans l'équation finale

$$p \cdot \overline{MA}^2 + q \cdot \overline{MB}^2 + r \cdot \overline{MC}^2 \ldots = (p + q + r + \ldots)\, \overline{OM}^2 + l$$

du n° **609**, introduisons (comme au n° **604**) la quantité t définie par la relation

$$p + q + r + \ldots + t = 0 :$$

il vient

$$p \cdot \overline{MA}^2 + q \cdot \overline{MB}^2 + r \cdot \overline{MC}^2 + \ldots + t \cdot \overline{MO}^2 = l.$$

Supposons qu'il y ait trois points A, B, C et soit O un point situé dans leur plan. Nous savons qu'on peut déterminer les coefficients p, q, r de manière que O soit le centre des distances proportionnelles des points A, B, C.

Ainsi, *étant donnés quatre points* A, B, C, O *d'un plan* (²) *il existe toujours quatre nombres* p, q, r, t non tous nuls, mais dont la somme est nulle, *tels que la somme* $p \cdot \overline{MA}^2 + q \cdot \overline{MB}^2 + r \cdot \overline{MC}^2 + t\, \overline{MO}^2$ *soit indépendante de la position du point* M.

Par un raisonnement tout semblable, on voit que, *étant donnés cinq points quelconques* A, B, C, D, O *de l'espace*, *il existe cinq nombres* p, q, r, s, t, *non tous nuls*, *mais dont la somme est nulle tels que la quantité* $p \cdot \overline{MA}^2 + q \cdot \overline{MB}^2 + r \cdot \overline{MC}^2 + s \cdot \overline{MD}^2 + t \cdot \overline{MO}^2$ *soit indépendante de la position du point* M.

---

(1) Dans le cas (n° **596**, note) où le point O′ coïnciderait avec le point O, il est clair que la *quantité* $p.\ MA^2 + q.\ MB^2 + \ldots$ *serait constante, quel que soit le point* M.

(2) Le raisonnement serait en défaut si les points A, B, C étaient en ligne droite ; on verrait alors, comme au n° **602**, que le résultat subsiste, t étant nul. La formule ainsi obtenue ne serait autre que celle de Stewart (Pl., **127**).

612. Relations entre les distances mutuelles de quatre points d'un plan.

Soient A, B, C, O quatre points d'un plan. Désignons par x, y, z, a, b, c, respectivement, les distances OA, OB, OC, BC, CA, AB.

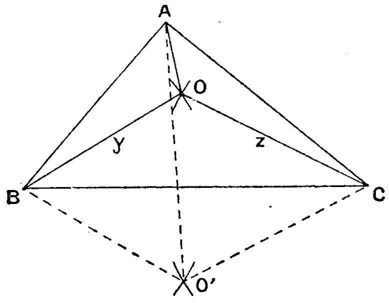

Fig. 515.

Il existe une relation entre les six quantités x, y, z, a, b, c. En effet, connaissant a, b, c, on peut construire le triangle ABC (*fig.* 515), après quoi, les circonférences de centres respectifs B, C et de rayons y, z donnent deux lieux du point O. Ces deux circonférences se coupent en deux points O, O′ (*fig.* 515) et, par conséquent, lorsqu'on donne a, b, c, y, z, la quantité x ne peut avoir que deux valeurs.

Pour trouver la relation dont nous venons de constater l'existence, nous emploierons le résultat obtenu au numéro précédent. Soient donc p, q, r, t, quatre nombres de somme nulle et tels que l'on ait

$$p.\,\overline{\mathrm{MA}}^2 + q.\,\overline{\mathrm{MB}}^2 + r.\,\overline{\mathrm{MC}}^2 + t.\,\overline{\mathrm{MO}}^2 = l$$

l étant indépendant de la position du point M.

Nous ferons coïncider successivement le point M avec les points A, B, C, O ; il viendra, en tenant compte des désignations précédemment adoptées pour les distances mutuelles des quatre points en question,

$$(1)\qquad \begin{array}{rrrrl} & qc^2 & + rb^2 & + tx^2 & = l \\ pc^2 & & + ra^2 & + ty^2 & = l \\ pb^2 & + qa^2 & & + tz^2 & = l \\ px^2 & + qy^2 & + rz^2 & & = l \end{array}$$

Si nous tenons compte de la condition

$$(2)\qquad p + q + r + t = 0,$$

nous aurons cinq équations homogènes du premier degré entre lesquelles nous pourrons éliminer les cinq inconnues, non toutes nulles, p, q, r, t, l.

Pour faire cette élimination (¹), écrivons, par exemple, les trois premières équations sous la forme

$$\frac{q}{b^2}+\frac{r}{c^2}=\frac{a^2\,(l-tx^2)}{a^2\,b^2\,c^2}$$

$$\frac{r}{c^2}+\frac{p}{a^2}=\frac{b^2\,(l-ty^2)}{a^2\,b^2\,c^2}$$

$$\frac{p}{a^2}+\frac{q}{b^2}=\frac{c^2\,(l-tz^2)}{a^2\,b^2\,c^2};$$

d'où résulte

$$\begin{aligned} p &= a^2\left[\frac{b^2\,(l-ty^2)+c^2(l-tz^2)-a^2\,(l-tx^2)}{2\,a^2\,b^2\,c^2}\right] \\ &= a^2\,\frac{l\,(b^2+c^2-a^2)-t\,(b^2\,y^2+c^2\,z^2-a^2\,x^2)}{2\,a^2\,b^2\,c^2} \\ q &= b^2\,\frac{l\,(c^2+a^2-b^2)-t\,(c^2\,z^2+a^2\,x^2-b^2\,y^2)}{2\,a^2\,b^2\,c^2} \\ r &= c^2\,\frac{l\,(a^2+b^2-c^2)-t\,(a^2\,x^2+b^2\,y^2-c^2\,z^2)}{2\,a^2\,b^2\,c^2}. \end{aligned}$$

Si nous substituons ces valeurs dans la dernière équation (1) et dans l'équation (2), celles-ci deviennent, l'une

$$(3)\qquad l\Sigma a^2(b^2+c^2-a^2)x^2-t\Sigma a^2x^2(b^2y^2+c^2z^2-a^2x^2)=2ta^2b^2c^2$$

(en désignant par $\Sigma a^2(b^2+c^2-a^2)x^2$ la somme $a^2(b^2+c^2-a^2)x^2+b^2(c^2+a^2-b^2)y^2+c^2(a^2+b^2-c^2)z^2$ et, de même, par $\Sigma a^2x^2\,(by^2+c^2z^2-a^2x^2)$ la somme de la quantité $a^2x^2(b^2y^2+c^2z^2-a^2x^2)$ et des deux autres analogues); l'autre

$$(4)\quad l\Sigma a^2(b^2+c^2-a^2)-t\Sigma a^2(b^2y^2+c^2z^2-a^2x^2)+2ta^2b^2c^2=0.$$

La quantité

$$\begin{aligned}\Sigma a^2(b^2+c^2-a^2) &= a^2(b^2+c^2-a^2)+b^2(c^2+a^2-b^2)+ \\ &\quad c^2(a^2+b^2-c^2)=2b^2c^2+2c^2a^2+2a^2b^2-a^4-b^4-c^4\end{aligned}$$

---

(1) Le lecteur familiarisé avec la théorie des déterminants écrira immédiatement le résultat de cette élimination sous la forme

$$\begin{vmatrix} 0 & c^2 & b^2 & x^2 & 1 \\ c^2 & 0 & a^2 & y^2 & 1 \\ b^2 & a^2 & 0 & z^2 & 1 \\ x^2 & y^2 & z^2 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{vmatrix} = 0.$$

représente (Pl., **251**) 16 fois le carré de la surface S du triangle ABC. De même, la quantité $\Sigma a^2x^2(b^2y^2 + c^2z^2 - a^2x^2)$ représente 16 fois le carré de la surface du triangle dont les côtés sont mesurés par ax, by, cz.

Enfin, il faut remarquer que la somme

$$\begin{aligned}\Sigma a^2(b^2y^2 + c^2z^2 - a^2x^2) = a^2(by^2 + c^2z^2 + a^2x^2) + b^2(c^2z^2 + a^2x^2 - b^2y^2)\\ + c^2(a^2x^2 + b^2y^2 - c^2z^2)\end{aligned}$$

est identiquement égale à $\Sigma a^2(b^2 + c^2 - a^2)x^2$. Les deux équations (3) et (4) s'écrivent donc

$$(3') \quad l[\Sigma a^2(b^2 + c^2 - a^2)x^2 - 2a^2b^2c^2] = t\Sigma a^2x^2(b^2y^2 + c^2z^2 - a^2x^2)$$

et

$$(4') \qquad 16S^2l = t[\Sigma a^2(b^2 + c^2 - a^2)x^2 - 2a^2b^2c^2].$$

L'élimination de l et de t est immédiate et donne

$$(5) \quad [\Sigma a^2(b^2 + c^2 - a^2)x^2 - 2a^2b^2c^2]^2 = 16S^2.\Sigma a^2x^2(b^2y^2 + c^2z^2 - a^2x^2).$$

On peut ordonner cette équation par rapport à x, y, z. On obtient ainsi aisément (en divisant par $4a^2b^2c^2$) :

$$(5') \quad \Sigma a^2(x^2 - y^2)(x^2 - z^2) - \Sigma a^2(b^2 + c^2 - a^2)x^2 + a^2b^2c^2 = 0.$$

613. Nous avons à nous demander si, réciproquement, la relation précédente entre six quantités x, y, z est suffisante pour que ces quantités mesurent les distances mutuelles de quatre points d'un plan. C'est ce que nous allons démontrer *en supposant que l'on puisse construire un triangle ayant pour côtés* a, b, c.

A cet effet, nous remarquerons que, si les quantités a, b, c, x, y, z satisfont à l'équation (5), on peut (ainsi qu'on s'en assure aisément en refaisant en sens inverse les calculs précédents) trouver les quantités p, q, r, t, l satisfaisant aux équations (1) et (2).

Construisons alors le triangle ABC qui a pour côtés a, b, c, et soit O le point qui a pour coordonnées barycentriques par rapport à ce triangle, les quantités p, q, r. Si x', y', z' désignent les distances

OA, OB, OC, il existe, ainsi que nous l'avons vu, un nombre l' tel que l'on ait

$$(1')\quad \left\{\begin{array}{lllll} qc^2 & +\, rb^2 & +\, tx'^2 & = l' \\ pc^2 & +\, ra^2 & +\, ty'^2 & = l' \\ pb^2 \,+\, qa^2 & & +\, tz'^2 & = l' \\ px'^2 \,+\, qy'^2 & +\, rz'^2 & & = l' \end{array}\right.$$

La comparaison de ces équations avec les équations (1), (2) donne aisément $x' = x$, $y' = y$, $z' = z$, $l' = l$: d'où la conclusion demandée.

Notre raisonnement suppose, nous l'avons dit, que l'on puisse construire un triangle ayant pour côtés a, b, c. Toutefois, il serait encore valable (1) si l'on pouvait construire l'un des triangles qui ont respectivement pour côtés a, y, z; b, z, x; c, x, y : il suffirait, en effet, de le recommencer avec un changement de notation, le point O étant permuté avec l'un des points A, B, C.

Mais il est absolument nécessaire d'introduire l'une ou l'autre des suppositions précédentes. On peut trouver six quantités a, b, c, x, y, z satisfaisant à la relation (5) et qui, cependant, ne représentent pas les distances mutuelles de quatre points du plan, parce qu'aucun des triangles ayant pour côtés a, b, c; a, y, z; b, z, x; c, x, y, ne peut être construit (voir exercices 876, 877).

614. Volume du tétraèdre en fonction des arêtes. Soient $a = BC$, $b = CA$, $c = AB$, $\alpha = DA$, $\beta = DB$, $\gamma = DC$ les six arêtes d'un tétraèdre ABCD (*fig.* 516). La connaissance de ces six longueurs détermine le polyèdre (ex. 617) et, par conséquent, son volume. Nous allons montrer comment on peut, dans ces conditions, calculer celui-ci.

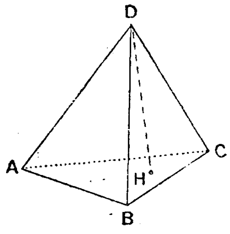

FIG. 516.

Du point D, soit abaissée la hauteur DH $= h$, et désignons par x, y, z, les distances HA, HB, HC : on aura, dans les triangles rectangles DHA, DHB, DHC,

$$\begin{aligned} \overline{HA}^2 = x^2 &= \alpha^2 - h^2 \\ y^2 &= \beta^2 - h^2 \\ z^2 &= \gamma^2 - h^2 \end{aligned}$$

---

(1) Il résulte de là que, lorsque les quantités a, b, c, x, y, z satisfont à la relation (5), si l'un des triangles qui ont respectivement pour côtés a, b, c; a, y, z; b, z, x; c, x, y peut être construit, il en est de même des trois autres.

Substituons ces valeurs dans la relation (5′), laquelle a lieu entre les longueurs a, b, c, x, y, z, puisque les points H, A, B, C sont dans un même plan. Il vient

$$\Sigma\, a^2 (\alpha^2 - \beta^2)(\alpha^2 - \gamma^2) - \Sigma\, a^2 (b^2 + c^2 - a^2)(\alpha^2 - h^2) + a^2b^2c^2 = 0$$

ou, en faisant passer tous les termes en h^2 dans le second membre,

$$\begin{aligned} &\Sigma\, a^2 (\alpha^2 - \beta^2)(\alpha^2 - \gamma^2) - \Sigma\, a^2 (b^2 + c^2 - a^2)\, \alpha^2 + a^2\, b^2\, c^2 \\ &= - h^2\, \Sigma a^2 (b^2 + c^2 - a^2) = - 16\, S^2 h^2, \end{aligned}$$

S désignant toujours la surface du triangle ABC. Mais Sh représente trois fois le volume V cherché. Donc l'expression

$$\Sigma a^2 (\alpha^2 - \beta^2)(\alpha^2 - \gamma^2) - \Sigma\, a^2 (b^2 + c^2 - a^2)\, \alpha^2 + a^2\, b^2\, c^2,$$

qui est nulle lorsque a, b, c, α, β, γ, sont les distances mutuelles de quatre points d'un plan, est, d'une façon générale, égale à $- 16 \times 3^2 . V^2$, en désignant par V le volume du tétraèdre qui a pour arêtes a, b, c, α, β, γ.

615. Relation entre les distances mutuelles de cinq points de l'espace.

Soient A, B, C, D, O cinq points de l'espace ; $a =$ BC, $b =$ CA, $c =$ AB, $\alpha =$ DA, $\beta =$ DB, $\gamma =$ DC, $x =$ OA, $y =$ OB, $z =$ OC, $u =$ OD leurs distances mutuelles. Il existe entre ces dix quantités une relation, comme on le verra en constatant (par une marche analogue à celle qui a été suivie tout à l'heure) que la connaissance des longueurs a, b, c, α, β, γ détermine le tétraèdre ABCD; la connaissance des longueurs a, b, c, x, y, z, le tétraèdre ABCO : après quoi la quantité u ne peut plus avoir que deux valeurs.

Pour trouver la relation en question, on s'appuiera sur l'existence des nombres p, q, r, s, t, de somme nulle et tels que la quantité $p.\overline{\mathrm{MA}}^2 + q.\overline{\mathrm{MB}}^2 + r.\overline{\mathrm{MC}}^2 + s.\overline{\mathrm{MD}}^2 + t.\overline{\mathrm{MO}}^2$ soit une quantité l indépendante du choix du point M ; et, faisant coïncider successivement le point M avec A, B, C, D, O, on écrira les équations

$$\begin{array}{lllllll} & qc^2 & + rb^2 & + s\alpha^2 & + tx^2 & = l \\ pc^2 & & + ra^2 & + s\beta^2 & + ty^2 & = l \\ pb^2 & + qa^2 & & + s\gamma^2 & + tz^2 & = l \\ p\alpha^2 & + q\beta^2 & + r\gamma^2 & & + tu^2 & = l \\ px^2 & + qy^2 & + rz^2 & + su^2 & & = l \\ p & + q & + r & + s & + t & = 0 \end{array}$$

entre lesquelles il faudrait éliminer p, q, r, s, t, l pour avoir la relation cherchée (¹).

EXERCICES

859. Le centre des distances proportionnelles de plusieurs points affectés de coefficients dont la somme est nulle, s'il n'est pas rejeté à l'infini, est en général indéterminé. Si on le recherche par la méthode du n° **594**, le point obtenu n'est pas indépendant de l'ordre dans lequel on prend les points donnés.

860. Si l'on fait tendre les coordonnées barycentriques d'un point M vers des nombres fixes non tous nuls, mais dont la somme est nulle, le point M s'éloigne à l'infini dans une direction déterminée.

861. Lorsqu'un point du plan d'un triangle décrit une droite, ses coordonnées barycentriques p, q, r sont liées par une relation de la forme $ap + bq + cr = 0$, les nombres a, b, c étant constants.

(Il suffit de prendre pour a, b, c les distances des sommets du triangle à la droite considérée.)

Réciproquement, le lieu des points dont les coordonnées barycentriques satisfont à une relation de la forme précédente est une droite, sauf dans les cas où a, b, et c sont égaux entre eux, et où tous les points du lieu sont rejetés à l'infini (²).

Quelles sont les propriétés analogues dans l'espace ?

862. $ap + bq + cr = 0$, $a'p + b'q + c'r = 0$ étant respectivement les *équations* de deux droites, c'est-à-dire les relations entre les coordonnées barycentriques qui les caractérisent, quelles conditions doivent vérifier a, b, c, a' b' c' pour que ces droites soient parallèles ?

Quel est le problème analogue dans l'espace ?

863. Appliquer l'exercice 861 à la recherche du lieu des points dont la somme des distances à trois droites données (dans le plan) ou à quatre plans donnés (dans l'espace) est constante.

864. Sur chaque face d'un tétraèdre, on prend un point ayant, par rapport à cette face, des coordonnées barycentriques données. Dans quel rapport le volume du tétraèdre qui a pour sommets les quatre points ainsi déterminés, est-il avec le volume du tétraèdre primitif ?

865. A tout polygone plan, on peut faire correspondre un segment tel que sa projection sur une droite quelconque de l'espace soit mesurée par le même nombre que l'aire de la projection du polygone sur un plan perpendiculaire à cette droite.

(¹) A l'aide des déterminants, le résultat s'écrit : $\begin{vmatrix} 0 & c^2 & b^2 & \alpha^2 & x^2 & 1 \\ c^2 & 0 & a^2 & \beta^2 & y^2 & 1 \\ b^2 & a^2 & 0 & \gamma^2 & z^2 & 1 \\ \alpha^2 & \beta^2 & \gamma^2 & 0 & u^2 & 1 \\ x^2 & y^2 & z^2 & u^2 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{vmatrix} = o$

(²) Ce cas peut d'ailleurs être considéré comme rentrant dans le cas général (voir ci-après, n° **621**).

(Ce segment est perpendiculaire au plan du polygone et sa longueur est mesurée par le même nombre que l'aire du polygone).

866. A tout polygone gauche, on peut faire correspondre un segment tel que sa projection sur une droite quelconque, soit mesurée par le même nombre que l'aire limitée par la projection du polygone sur le plan perpendiculaire à cette droite (cette projection étant supposée ne se couper elle-même en aucun point) (¹).

Le théorème étant vrai pour un triangle (ex. précéd.), on remarquera que si on a formé les segments demandés pour deux polygones qui ont une partie commune, on peut former un segment analogue, — le résultant (²) des deux premiers — pour le contour formé en réunissant ces deux polygones et supprimant la partie commune.

(Le théorème s'étend, par voie de passage à la limite, à un contour curviligne).

867. Faire passer, par un point donné, un plan qui divise un prisme donné en deux troncs de prisme dont le rapport soit égal à un nombre donné.

868. Trouver toutes les sphères tangentes à quatre plans donnés, en remarquant qu'on connaît, au signe près, les coordonnées barycentriques du centre d'une telle sphère.

Montrer que le nombre de ces sphères est, en général, de huit. Quelles relations doit-il y avoir entre les aires des faces pour qu'une ou plusieurs d'entre elles cessent d'exister?

869. Démontrer directement que le centre de gravité d'un polygone (**605**) est indépendant de son mode de décomposition en triangles.

On montrera (comparer Pl., note D) que, quel que soit le mode de décomposition, ce centre de gravité est le même que celui qu'on obtiendrait en affectant les différents triangles qui ont pour bases respectives les côtés et pour sommet commun un point quelconque O du plan, de coefficients proportionnels aux aires de ces triangles précédées du signe + ou du signe — suivant qu'ils sont additifs ou soustractifs, et prenant, dans ces conditions, le centre des distances proportionnelles des centres de gravité de ces triangles.

Démontrer la même proposition pour le centre de gravité d'un polyèdre (comparer note F).

870. Le centre des distances proportionnelles d'un système de points d'un plan affectés de coefficients positifs, ou le centre de gravité d'un polygone plan, est à l'intérieur de tout polygone convexe comprenant ces points ou ce polygone.

Le centre des distances proportionnelles d'un système de points de l'espace affectés de coefficients positifs (ou le centre de gravité d'un polyèdre) est à l'intérieur de tout polyèdre convexe comprenant ces points ou ce polyèdre.

871. On sait qu'on appelle *centre de gravité d'une ligne brisée (supposée homogène)* le centre des distances proportionnelles des milieux des côtés de cette ligne, affectés de coefficients proportionnels aux longueurs de ces côtés.

Montrer que la surface latérale d'un tronc de prisme est égale au périmètre de la section droite, multipliée par la longueur interceptée par les bases, sur la parallèle aux arêtes menées par le centre de gravité de ce périmètre.

---

(1) Le théorème s'étend au cas où la projection du contour se croise elle-même; mais alors cette projection délimite plusieurs aires et non une seule; et il faut, pour que le théorème reste exact, l'appliquer à une somme algébrique de ces aires, multipliées par des coefficients positifs ou négatifs convenables.

(2) Voir le *Cours de Mécanique*, ou encore la *Composition des translations* (**417**).

872. Quelles conditions doivent remplir six longueurs pour qu'elles puissent être les arêtes d'un même tétraèdre ?

873. Prouver directement que si six longueurs a, b, c, x, y, z satisfont à la relation (5) du n° **612**, les deux triangles dont l'un a pour côtés a, b, c et l'autre a, y, z peuvent être construits tous deux ou ne peuvent être construits ni l'un ni l'autre.

On considérera la relation (5) comme équation en x^2. La quantité qui doit être positive pour que cette équation ait ses racines réelles est le produit de

$$(a+b+c)\ (b+c-a)\ (a+c-b)\ (a+b-c) \qquad (\text{Pl.}, \mathbf{130}),$$

par

$$(a+y+z)\ (y+z-a)\ (z+a-y)\ (a+y-z).$$

874. Quelle relation doit-il y avoir entre les arêtes a, $b, c, \alpha, \beta, \gamma$ (n° **614**) d'un tétraèdre ABCD, pour que le centre de la sphère circonscrite soit dans le plan de la face ABC ?

Réponse :

$$a^2\alpha^2(b^2+c^2-a^2)+b^2\beta^2(c^2+a^2-b^2)+c^2\gamma^2(a^2+b^2-c^2)=2a^2b^2c^2.$$

875. Si A, B, C, D sont quatre points d'un plan et M un point quelconque de l'espace, on a

$$\overline{\text{MA}}^2.(\text{DBC})+\overline{\text{MB}}^2.(\text{DCA})+\overline{\text{MC}}^2.(\text{DAB})+\overline{\text{MD}}^2.(\text{CBA})=\pm\Sigma.$$

où (DBC), (DCA), etc., désignent les aires des triangles DBC, DCA, précédées du signe + ou du signe —, suivant les sens de rotation de ces triangles, les sommets étant pris, bien entendu, dans l'ordre où ils sont écrits, et Σ est l'aire du triangle dont les côtés sont mesurés par les produits AB.CD, AC.BD, AD.BC.

Si les points A, B, C, D sont sur une même circonférence, on a

$$\overline{\text{MA}}^2.(\text{DBC})+\overline{\text{MB}}^2.(\text{DCA})+\overline{\text{MC}}^2.(\text{DAB})+\overline{\text{MD}}^2.(\text{CBA})=0.$$

(Appliquer le premier théorème du n° **611** en remarquant que les coordonnées barycentriques d'un des points A, B, C, D par rapport au triangle formé par les trois autres sont liées aux aires des triangles DBC..., comme il a été expliqué au n° **601**. On transformera ensuite les valeurs de la constante trouvée en considérant les inverses des points A, B, C par rapport à D pris comme pôle d'inversion).

876. Un angle étant donné dans un plan, appelons *distance hyperbolique* de deux points quelconques A et B du plan, le côté du carré équivalent au parallélogramme qui a ses côtés parallèles à ceux de l'angle et deux sommets opposés en A et B, cette distance étant dite de *première* ou de *seconde espèce*, suivant que la parallèle à AB menée par le sommet de l'angle donné est située dans cet angle ou dans son supplément.

Appelons également *perpendiculaire hyperbolique* à une droite, tout autre droite qui forme avec la première et les parallèles aux côtés de l'angle donné menées par leur point commun, un faisceau harmonique ; *distance hyperbolique* d'un point à une droite, la longueur hyperbolique de la perpendiculaire hyperbolique menée de ce point à cette droite.

Montrer :

1° Que si trois points A, B, C sont en ligne droite, l'une des longueurs hyperboliques BC, CA, AB est égale à la somme des deux autres ;

2° Que l'aire d'un triangle a pour mesure le demi-produit de la longueur hyperbolique d'un côté par la distance hyperbolique de ce côté au sommet opposé ;

3° Que si un triangle ABC a deux côtés AB, BC hyperboliquement perpendiculaires, et que la longueur hyperbolique AC soit de l'espèce de AB, la longueur hyperbolique perpendiculaire AB est *plus grande* que la longueur hyperbolique AC, et que le carré de cette dernière est égal à la *différence* des carrés des longueurs hyperboliques AB et BC;

4° Que si trois points A, B, C sont tels que les longueurs hyperboliques BC, CA, AB soient de même espèce, le carré de la première est égal à la somme des carrés des deux autres, moins le double produit des longueurs hyperboliques AB et AH, en désignant par H le point de rencontre de la droite AB avec sa perpendiculaire hyperbolique menée par C, et comptant le double produit dont il vient d'être question en grandeur et signe (suivant que les segments AC et AH sont de même sens ou de sens contraires).

(Imiter Pl., **126**).

5° Que, dans les mêmes conditions, l'une des distances hyperboliques BC, CA, AB est *plus grande* que la somme des deux autres (sauf dans le cas où les trois points A, B, C sont en ligne droite).

6° Que si les quatre points A, B, C, D du plan sont tels que leurs distances hyperboliques mutuelles soient de même espèce, ces distances satisfont à la relation (5′) du n° **612**.

Comment les énoncés 4° et 6° doivent-ils être modifiés si les distances hyperboliques des points considérés deux à deux ne sont pas toutes de même espèce?

Évaluer l'aire d'un triangle en fonction des longueurs hyperboliques de ses côtés (comparer Pl., **130**, **251**).

877. (Réciproque de l'exercice précédent, 6°). Si six longueurs a, b, c, x, y, z satisfont à la relation (5′) du n° **612** :

ou bien il existe quatre points d'un plan dont les distances mutuelles sont a, b, c, x, y, z ;

ou bien il existe quatre points d'un plan dont les distances hyperboliques mutuelles, toutes de même espèce, sont a, b, c, x, y, z.

Les deux hypothèses s'excluent l'une l'autre, sauf si les quatre points sont en ligne droite.

CHAPITRE VII

ELLIPSE CONSIDÉRÉE COMME PROJECTION D'UN CERCLE. HYPERBOLE RAPPORTÉE A SES ASYMPTOTES

731. Théorème. — *Un cercle étant rapporté à deux diamètres perpendiculaires entre eux, le lieu du point qu'on déduit d'un point quelconque de la circonférence en augmentant l'ordonnée dans un rapport constant, sans changer l'abscisse, est une ellipse.*

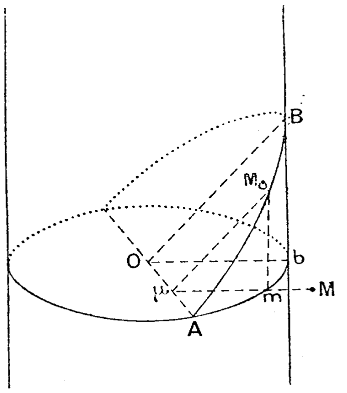

Fig. 603.

Soient O le cercle donné (*fig.* 603); OA, Ob, les deux diamètres rectangulaires; m, un point de la circonférence, ayant pour abscisse Oμ et pour ordonnée μm; M, un point pris sur μm prolongé de manière que $\frac{\mu M}{\mu m}$ soit égal à un nombre constant k plus grand que 1. Je dis que le lieu de M est une ellipse.

C'est ce qui résulte de ce fait que la section du cylindre droit qui a pour base le cercle donné par un plan passant par OA est une ellipse. Nous savons, en effet, (**360**, **364**), que si M_0 est un point de cette section; m, le point du cercle donné, projection de M_0 sur le plan de ce cercle, les points M_0, m ont même projection μ sur OA et que le rapport $\frac{\mu M_0}{\mu m}$ est constant (il est inverse du cosinus de l'angle $\widehat{M_0 \mu m}$, c'est-à-dire du cosinus de l'angle α que fait le plan de section avec le plan de base). μM_0 sera donc constamment égal à μM et, par conséquent, l'ellipse de section sera égale au lieu cherché (que l'on obtiendra en *rabattant* cette ellipse sur le plan de base, c'est-à-dire en la faisant tourner autour de OA, d'un angle égal à α), si ce rapport constant est égal à k. On arrivera évidemment à ce résultat en faisant passer le plan de section par un point B pris sur la génératrice issue du point b, de manière que OB soit égal à $k.\overline{Ob}$ (autrement dit, en déterminant l'angle des deux plans de manière que son cosinus soit égal à $\frac{1}{k}$) ce qui est possible, puisque k est supérieur à 1.

L'ellipse obtenue a d'ailleurs pour petit axe le diamètre OA commun aux deux plans.

732. Théorème. — *Un cercle étant rapporté à deux diamètres rectangulaires, le lieu du point qu'on déduit d'un point quelconque de la circonférence en diminuant l'ordonnée dans un rapport constant, sans changer l'abscisse, est une ellipse ayant le cercle donné pour cercle principal.*

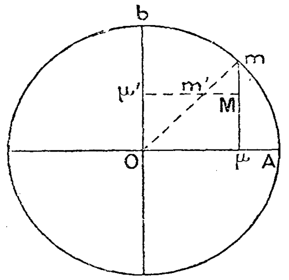

Fig. 604.

Soient encore OA, Ob les deux diamètres rectangulaires auxquels est rapporté le cercle donné ; m un point de ce cercle projeté en μ sur OA ; M un point pris sur μm de manière que $\frac{\mu \mathrm{M}}{\mu m}$ soit égal à une constante k plus petite que 1 (*fig.* 604).

Projetons le point M en μ' sur Ob et soit m' le point où Mμ' rencontre le rayon Om. Les parallèles Mm', μO et les parallèles mM, Oμ' montrent : 1° que $\frac{\mathrm{O}m'}{\mathrm{O}m} = \frac{\mu\mathrm{M}}{\mu m} = k$, de sorte que O$m'$ est constant ; 2° que $\frac{\mu'\mathrm{M}}{\mu' m'} = \frac{\mathrm{O}m}{\mathrm{O}m'} = \frac{1}{k}$. Nous voyons donc que le point m' décrit une circonférence de centre O et que, si Ob est pris pour axe des abscisses, le point M se déduit du point m' en augmentant l'ordonnée dans le rapport constant $\frac{1}{k}$ (lequel est plus grand que 1) sans changer l'abscisse. Donc le point M décrit une ellipse.

733. Projection orthogonale du cercle.

Théorème. — *La projection orthogonale d'un cercle sur un plan est une ellipse.*

On peut toujours (**413**) supposer que le plan de projection passe par le centre du cercle (*fig.* 605). Alors, si l'on prend le diamètre OA situé dans ce plan pour axe des abscisses tant du cercle que de sa projection, le rapport $\frac{\mu M}{\mu m_0}$ des ordonnées correspondant à une même abscisse $O\mu$ est une constante plus petite que 1. Donc, (théorème précédent) la projection est bien une ellipse ayant pour cercle principal un cercle égal au cercle donné. Le point m du cercle principal situé sur l'ordonnée μM prolongée est celui que l'on obtient en rabattant le point m_0 du cercle donné qui a pour projection M, c'est-à-dire en le faisant tourner, autour de OA, d'un angle égal à l'angle α des deux plans.

On peut d'ailleurs faire le rabattement de deux manières différentes, à savoir : de façon que les segments μM, μm soient toujours

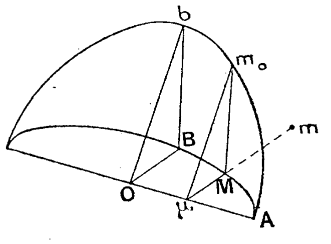

Fig. 605.

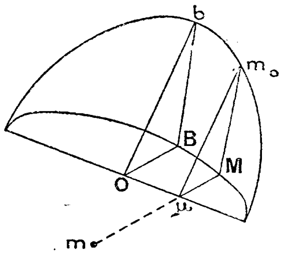

Fig. 605 *bis*.

de même sens (*fig.* 605), ou de façon que ces segments soient toujours de sens contraires (*fig.* 605 *bis*).

Remarque. — L'angle α a pour cosinus le nombre k, lequel est, d'ailleurs *égal au rapport* $\frac{b}{a}$ en désignant par $2a$ et $2b$ les axes de l'ellipse) : car lorsque le point μ est en O, on a manifestement $\mu m = Ob$ (*fig.* 605) $= a$; $\mu M = OB$ (*fig.* 605) $= b$.

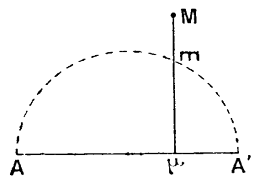

Fig. 606.

Corollaire I. — *Le lieu des points* M *d'un plan, projetés en* μ *à l'intérieur d'un segment* AA' (*fig.* 606) *et tels que* μM *soit dans un rapport constant avec la moyenne proportionnelle de* μA et de $\mu A'$, *est une ellipse*.

Car la moyenne proportionnelle de μA et de $\mu A'$ n'est autre (Pl., **125**, Coroll.) que l'ordonnée μm du cercle qui a AA' pour diamètre. L'énoncé précédent résulte donc des deux premiers théorèmes du numéro précédent.

II. *Équation de l'ellipse.* — Soient a, b les deux demi-axes; x, y, l'abscisse $O\mu$ et l'ordonnée μM du point M. L'ordonnée μm du cercle principal est évidemment donnée par

$$\overline{\mu m}^2 + x^2 = a^2.$$

Par conséquent, y qui est égal à $k.\,\overline{\mu m}$ ou (Rem. précéd. à $\frac{b}{a}.\,\overline{\mu m}$ sera donné par la relation

$$y^2 = \frac{b^2}{a^2}\,\overline{\mu m}^2 = \frac{b^2}{a^2}(a^2 - x^2).$$

ou

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1,$$

laquelle représente l'*équation de l'ellipse*, rapportée aux deux axes de symétrie de la courbe.

734. L'ellipse et le cercle principal sont (**630-633**) deux *figures homographiques*, dans lesquelles les points à l'infini se correspondent. Les points du grand axe sont également communs aux deux figures.

Un point P étant donné dans la figure F dont fait partie l'ellipse, on construira son homologue p dans la figure f dont fait partie le cercle en augmentant l'ordonnée de P dans le rapport $\frac{a}{b}$ sans changer l'abscisse. On trouvera d'ailleurs l'homologue d'une droite de la figure F en construisant les homologues de deux de ses points ou, plus simplement, en construisant l'homologue d'un de ses points que l'on joindra au point de rencontre de la droite avec l'axe (ce dernier point se correspondant à lui-même).

L'ellipse et le cercle du n° **731** sont également deux figures homographiques et on peut leur appliquer des considérations toutes semblables aux précédentes : ce sont, cette fois, les points du petit axe qui sont communs aux deux figures.

On peut tirer de ce que nous venons de dire des solutions nouvelles pour les problèmes déjà traités au livre IX, chap. I, tels que :

Intersection d'une droite et d'une ellipse. — On déterminera l'homologue de la droite donnée dans la figure f. Les points cherchés sont les homologues (dans la figure F) des points de rencontre de la droite ainsi déterminée avec le cercle principal (*fig.* 607).

Tangentes à une ellipse par un point de son plan. — On déterminera l'homologue p de ce point dans la figure f (*fig.* 608) : les tangentes cherchées correspondront aux tangentes menées par p au cercle principal.

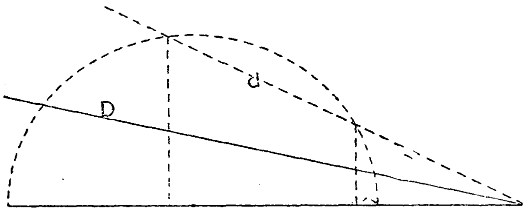

Fig. 607.

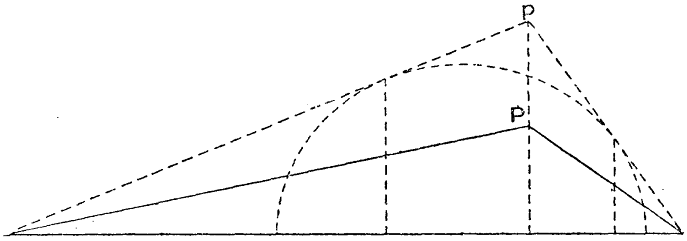

Fig. 608.

On mènera d'une manière analogue les *tangentes à l'ellipse parallèles à une direction donnée.*

735. Diamètres de l'ellipse.

Théorème. — *Le lieu des milieux des cordes parallèles à une direction donnée est une droite passant par le centre* (ou, du moins, la portion de cette droite intérieure à la courbe).

En effet, dans le cercle qui a pour projection l'ellipse (*fig.* 609), les cordes qui sont correspondantes à celles dont il est question dans l'énoncé, sont elles-mêmes parallèles à une direction fixe. Le lieu de leurs milieux est donc un diamètre (perpendiculaire à leur direction) dont la projection sera le lieu cherché.

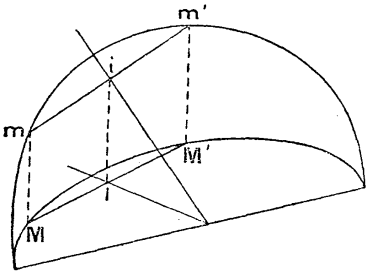

Fig. 609.

Ce lieu est nommé le *diamètre* de l'ellipse, *conjugué* de la direction considérée.

736. *Le diamètre conjugué d'une direction passe par les points de contact des tangentes parallèles à cette direction. Il passe également par le point de rencontre des tangentes aux extrémités d'une corde quelconque ayant la direction considérée.*

Car ces propriétés, qui appartiennent au cercle, se conservent dans la projection parallèle (1).

737. Diamètres conjugués.

Théorème. — *Si deux diamètres d'une ellipse sont tels que l'un est*

---

(1) Ces deux théorèmes peuvent d'ailleurs se démontrer directement, comme nous le ferons dans le cas de l'hyperbole (n^{os} **746-747**).

parallèle aux cordes conjuguées de l'autre, réciproquement celui-ci est parallèle aux cordes conjuguées du premier.

Car, la condition nécessaire et suffisante pour qu'un diamètre D d'une ellipse soit parallèle aux cordes conjuguées d'un autre diamètre D′ est (**735**) que D, D′ correspondent à deux diamètres rectangulaires du cercle qui a pour projection l'ellipse considérée : et il est clair que cette condition est réciproque, c'est-à-dire ne change pas lorsqu'on permute l'un avec l'autre les deux diamètres considérés.

Remarques. — I. *Les diamètres conjugués* (c'est-à-dire tels que chacun d'eux divise en deux parties égales les cordes parallèles à l'autre) *sont les rayons correspondants d'une involution*, projection de celle qui est formée par les rayons rectangulaires du cercle.

II. *Les seuls diamètres conjugués qui soient rectangulaires sont les axes* (**363**).

738. Théorème. — *Les cordes qui joignent un point d'une ellipse à deux points diamétralement opposés de cette courbe sont parallèles à deux diamètres conjugués.*

On donne souvent le nom de *cordes supplémentaires* à deux cordes ayant une extrémité commune et leurs secondes extrémités diamétralement opposées, de sorte que le théorème peut encore s'énoncer : *Deux cordes supplémentaires sont parallèles à deux diamètres conjugués.*

Car deux cordes supplémentaires d'un cercle sont rectangulaires, comme formant un angle inscrit dans une demi-circonférence [1].

739. Théorème *d'Apollonius.* — *Dans une ellipse :* 1° *La somme des carrés de deux demi-diamètres conjugués quelconques est constante, et égale à la somme des carrés des deux demi-axes;*

2° *L'aire du parallélogramme construit sur deux demi-diamètres conjugués est constante et égale à l'aire du rectangle construit sur les deux demi-axes.*

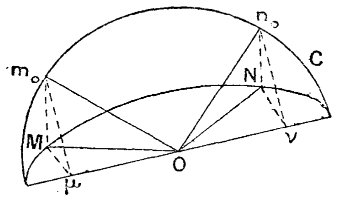

Fig. 610.

1° Soient une ellipse, projection d'un cercle C (*fig.* 610); deux diamètres conjugués OM, ON, projections de deux diamètres rectangulaires Om_0, On_0 du cercle C. Soient μ la projection commune de m_0

---

[1] Le théorème peut aussi se démontrer directement : voir le cas de l'hyperbole (n° **748**).

et de M sur OA; ν, la projection de n_0, N sur la même droite. Les deux triangles rectangles m_0MO, n_0NO montrent que la somme des carrés de OM et de ON est égale à la somme des carrés des deux hypoténuses (soit $2a^2$) diminuée de $\overline{m_0M}^2 + \overline{n_0N}^2$. Mais cette dernière somme est dans un rapport constant avec $\overline{m_0\mu}^2 + \overline{n_0\nu}^2$, car on a :

$$(21) \qquad \frac{a^2-b^2}{a^2} = \frac{\overline{m_0\mu}^2 - \overline{M\mu}^2}{\overline{m_0\mu}^2} = \frac{\overline{m_0M}^2}{\overline{m_0\mu}^2} = \frac{\overline{n_0N}^2}{\overline{n_0\nu}^2} = \frac{\overline{m_0M}^2 + \overline{n_0N}^2}{\overline{m_0\mu}^2 + \overline{n_0\nu}^2}$$

Enfin, on a $\overline{m_0\mu}^2 + \overline{n_0\nu}^2 = a^2$, car les triangles rectangles $Om_0\mu$, $On_0\nu$, ayant même hypoténuse et étant équiangles (puisqu'ils ont leurs côtés perpendiculaires), sont égaux, de sorte qu'on a $n_0\nu = O\mu$, d'où $\overline{m_0\mu}^2 + \overline{n_0\nu}^2 = \overline{m_0\mu}^2 + \overline{O\mu}^2 = \overline{Om_0}^2$.

Dès lors, $\overline{m_0M}^2 + \overline{n_0N}^2$, est égal, d'après l'équation (21), à $a^2 - b^2$. Donc, $\overline{OM}^2 + \overline{ON}^2$ est constant : il a d'ailleurs bien la valeur $2a^2 - (\overline{m_0M}^2 + \overline{n_0N}^2) = 2a^2 - (a^2 - b^2) = a^2 + b^2$: ce qui, au reste, est évident, car les axes forment évidemment un système de diamètres conjugués.

2° Le parallélogramme construit sur deux demi-diamètres conjugués est la projection du carré construit sur deux rayons rectangulaires du cercle C. Son aire est donc égale à a^2 (aire du carré) multiplié par le cosinus de l'angle des deux plans; soit $a^2 . \frac{b}{a} = ab$.

740. Théorème. — *Lorsqu'une droite de longueur constante se déplace de manière que ses deux extrémités décrivent deux droites fixes rectangulaires, un point quelconque de la droite mobile décrit une ellipse ayant ses axes suivant les droites fixes.*

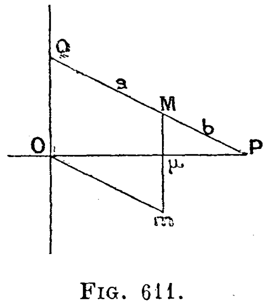

Fig. 611.

Soient PQ la droite mobile, les points P et Q décrivant respectivement les droites perpendiculaires Ox, Oy (*fig.* 611); M, un point de cette droite, tel que $MP = b$, $MQ = a$. Achevons le parallélogramme $OQMm$, dont le côté Mm, parallèle à OQ, coupe Ox en μ. Le point m décrit évidemment une circonférence de centre O et de rayon $Om = MQ = a$. Mais, d'autre part, le rapport $\frac{\mu M}{\mu m}$ est égal à $\frac{PM}{Om}$:

il a donc la valeur constante $\frac{b}{a}$. Le lieu de M est donc bien une ellipse, dont les axes ont pour directrices Ox, Oy, et pour longueurs a et $a\frac{b}{a} = b$.

Remarques. — I. On voit que *le rayon du cercle homographique correspondant à chaque position du point* M *est parallèle à la sécante mobile.*

II. Le raisonnement est le même, que le point M soit sur PQ lui-même (*fig.* 611) ou sur un de ses prolongements (*fig.* 611 *bis*). Une même ellipse admet donc deux modes de génération, les deux sécantes qui donnent un même point de la courbe étant également inclinées par rapport aux axes (PQ, P_1Q_1, *fig.* 611 *bis*).

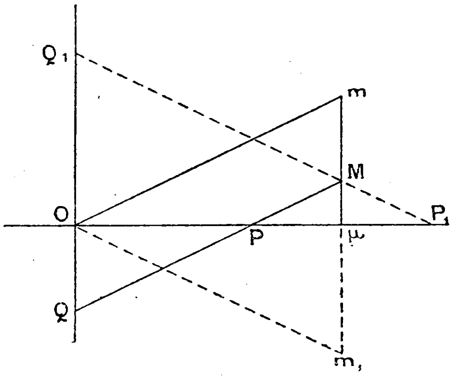

Fig. 611 *bis*.

Ce théorème donne un moyen très simple de construire par points une ellipse : il n'est pas, pour cela, besoin d'autre instrument qu'une *bande de papier* sur laquelle sont reportées, à partir d'un même point (dans le même sens ou en sens contraires) deux longueurs égales aux demi-axes.

On a fondé sur le même principe un *compas elliptique* permettant de tracer la courbe d'un mouvement continu.

741. Normale à l'ellipse.

On peut déduire du théorème qui précède une nouvelle construction de la tangente à l'ellipse :

Théorème. — *La normale à l'ellipse décrite par un point lié à une droite de longueur constante qui glisse entre deux droites rectangulaires fixes, passe par le quatrième sommet du rectangle qui a deux côtés suivant ces deux droites et une diagonale suivant la droite mobile.*

Ceci n'est autre chose que la proposition démontrée au n° **104** (Pl., liv. II). Car la normale au lieu du point P est la perpendiculaire à OP, la normale au lieu du point Q, la perpendiculaire à OQ, et par conséquent *le centre instantané de rotation* est le quatrième sommet R du rectangle qui a trois sommets en O, P, Q.

Le même théorème peut se démontrer directement. On remarquera, pour cela, que la tangente en M est (**736**) parallèle au diamètre conjugué de OM. Or, deux diamètres conjugués correspondent à deux rayons rectangulaires du cercle homographique, donc (n° précéd.) à deux positions rectangulaires de la droite mobile. On obtiendra, par conséquent, le diamètre $O'M'$ conjugué de OM en amenant le rectangle OPQR, par une rotation d'un angle droit autour du point O, à la position $OP'Q'R'$ (*fig.* 612) telle que $OP' = OQ$, $OQ' = OP$, puis prenant sur la nouvelle diagonale PQ', la longueur $P'M' = PM$.

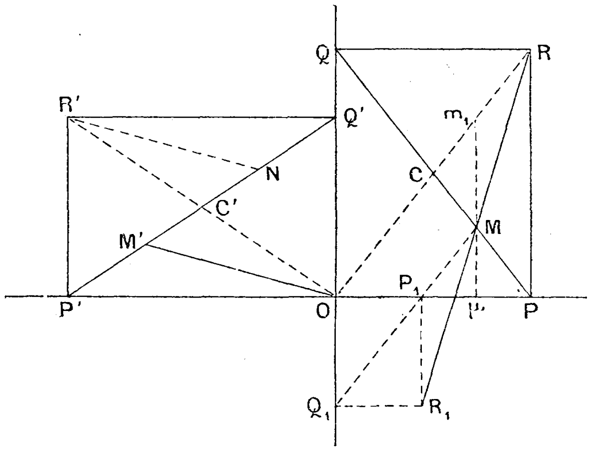

Fig. 612.

Mais si nous prenons, sur cette même diagonale, la longueur $Q'N$ égale à la précédente, la droite $R'N$ sera parallèle à OM' (car ces deux droites sont symétriques l'une de l'autre, par rapport au centre C' du rectangle $OP'Q'R'$). Or $R'N$ est la nouvelle position prise par RM dans la rotation d'un angle droit qui amène OPQR sur $OP'Q'R'$. Donc RM est bien perpendiculaire à OM', c'est-à-dire normal en M.

Remarques. — Dans la figure précédente, *la droite* RM *est égale au demi-diamètre conjugué de* OM : on a bien, en effet, $RM = R'N = OM'$.

Si l'on prolonge cette même droite RM d'une longueur MR_1 égale à elle-même, *le point* R_1 *est sur la droite symétrique de* OR *par rapport aux axes, à une distance du centre* O *égale à* $a - b$ *ou à* $a + b$, *suivant que le point* M *est intérieur ou extérieur au segment* PQ (c'est-à-dire suivant que $OR = PQ$ est lui-même égal à $a + b$ ou à $a - b$).

Car, C étant le centre du rectangle OPQR, la droite CM qui joint les milieux de OR et de RR_1, a une direction symétrique de celle de OR par rapport aux axes et une longueur égale à $\frac{a-b}{2}$ dans le cas où $CP = CQ = \frac{a+b}{2}$.

D'ailleurs on peut voir directement que *le point* R *est le quatrième sommet du rectangle* ($OP_1Q_1R_1$, *fig.* 612) *qu'on obtient en considérant le point* M *comme construit par le second mode de génération* mentionné au n° **740** (Rem. II), c'est-à-dire comme situé sur une droite P_1Q_1, de longueur constante $a - b$ mobile entre les axes, à la distance b du point P_1 et à la distance a du point Q_1 : il suffit de remarquer que la projection du point R_1 sur Ox est symétrique de P par rapport à la projection μ du point M sur la même droite, de sorte que cette projection coïncide avec le point P_1 (un raisonnement analogue s'appliquant à la projection de R_1 sur Oy).

Le point m_1 du cercle homographique que l'on obtient (**740**) en achevant le parallélogramme OQ_1Mm_1, n'est évidemment autre que l'intersection de $M\mu$ avec OR.

742. Les considérations précédentes donnent immédiatement la solution du problème suivant :

Problème. — *Construire les axes d'une ellipse, connaissant deux demi-diamètres conjugués en grandeur et position.*

OM, OM′ étant ces deux demi-diamètres (*fig.* 612) on mènera par M deux droites MR, MR_1 égales à OM′ et perpendiculaires à cette droite. Les axes seront dirigés suivant les bissectrices des angles formés par OR, OR_1 et leurs longueurs seront données par les relations $OR = a + b$, $OR_1 = a - b$.

743. Le théorème du n° **740** n'est qu'un cas particulier du suivant :

Théorème de *la Hire.* — *Lorsqu'une droite de longueur constante se déplace entre deux droites fixes quelconques, un point quelconque invariablement lié à la droite mobile décrit une ellipse.*

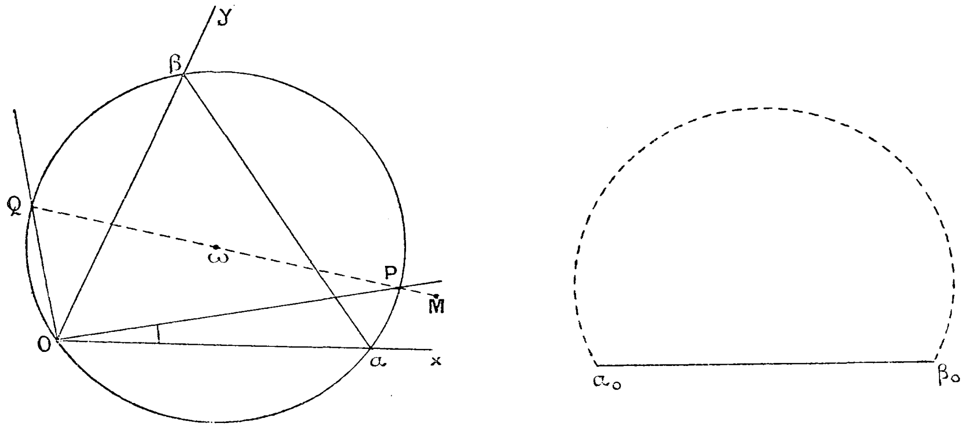

Fig. 613.

Soient αβ (*fig.* 613) une droite, constamment égale à une droite fixe

$\alpha_0\beta_0$, qui se déplace de manière que ses extrémités α, β décrivent respectivement deux droites fixes Ox, Oy ; M, un point qui se meut de manière que la figure $\alpha\beta M$ soit (Pl., **2**) invariable. Circonscrivons un cercle au triangle $O\alpha\beta$: ce cercle aura un rayon constant, celui du segment de cercle décrit sur $\alpha_0\beta_0$ et capable de l'angle $\widehat{xOy}$: il pourra dès lors être considéré comme invariablement lié à la figure mobile $\alpha\beta M$. Si donc nous menons, dans ce cercle, le diamètre PQ qui passe par le point M, les longueurs MP et MQ seront constantes.

Mais la droite OP est de direction fixe : car l'angle $\widehat{xoP}$ a pour mesure la moitié de l'arc constant αP ; et il en est de même de la droite OQ. Ces deux droites étant d'ailleurs rectangulaires, nous sommes revenus au lieu obtenu au n° **740**.

Remarque. — Toutefois, il existe une infinité de positions du point M pour lesquelles le lieu est une portion de droite (ellipse infiniment aplatie) ; ce sont les points de la circonférence $O\alpha\beta$.

744. Propriété des sécantes à l'hyperbole.

Théorème. — *Toute corde d'une hyperbole a même milieu que le segment intercepté sur elle par les asymptotes.*

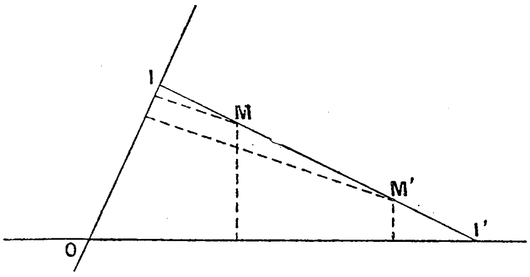

Fig. 614.

Soient MM' la corde considérée ; I, I' les points où elle coupe les asymptotes. Les distances des points M, M' à OI sont entre elles (*fig.* 614) comme les segments IM, IM' ; les distances de ces mêmes points à OI' sont entre elles comme I'M, I'M'. Or le rapport des deux premières distances est inverse du rapport des deux secondes, en vertu du théorème du n° **729**. Donc on a

$$\frac{IM}{IM'} = \frac{I'M'}{I'M}.$$

Cette égalité a d'ailleurs lieu *en grandeur et en signe*, car les points M, M' sont, par rapport aux asymptotes, dans un même angle (auquel cas les rapports $\frac{IM}{IM'}$ et $\frac{I'M'}{I'M}$ sont tous deux positifs ; *fig.* 614) ou dans deux angles opposés par le sommet (les rapports $\frac{IM}{IM'}$, $\frac{I'M'}{I'M}$ étant alors tous deux négatifs ; *fig.* 614 *bis*).

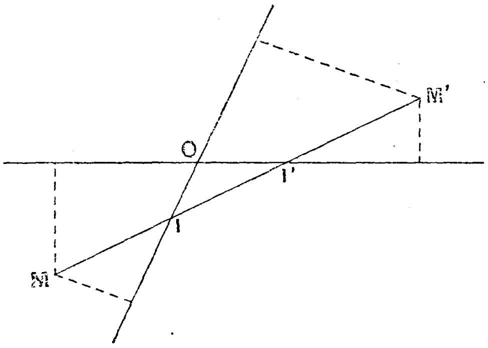

Fig. 614 *bis*.

Or il n'y a qu'un point qui divise MM' dans un rapport $\frac{IM}{IM'}$ égal en grandeur et en signe à $\frac{I'M'}{I'M}$, c'est le symétrique de I' par rapport au milieu de MM'.

Ce théorème donne un moyen simple de construire par points une hyperbole (dont on connaît les asymptotes et un point.) Il suffit de faire passer par ce point une série de sécantes (*fig.* 615) et de déterminer le second point M' où chacune d'elles coupe la courbe en portant, à partir du point d'intersection avec une des asymptotes, un segment égal à celui qui est intercepté entre la seconde asymptote et le point connu.

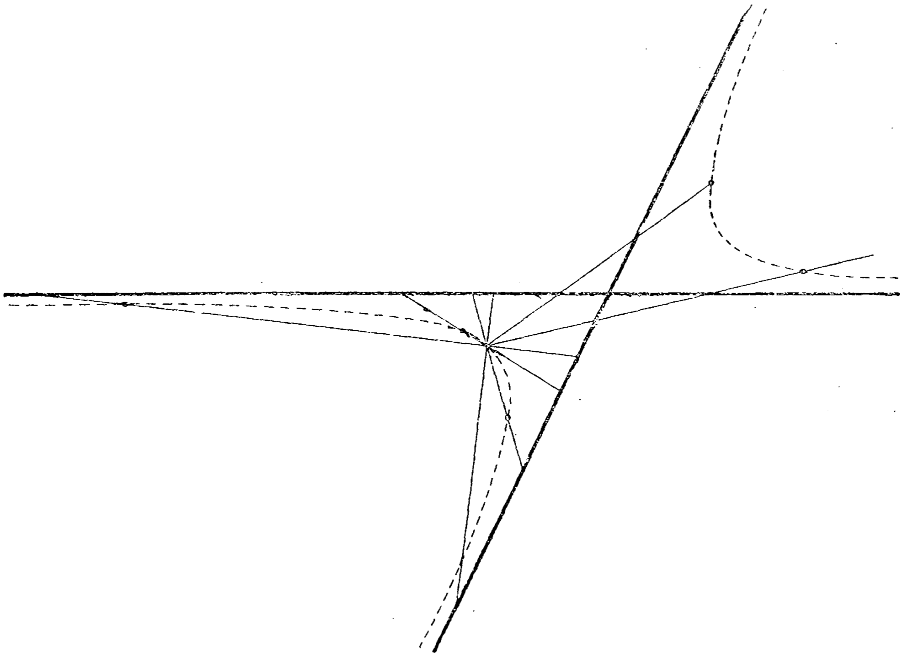

Fig. 615.

Tout point ainsi construit appartient d'ailleurs bien à l'hyperbole, comme on le voit en reprenant en sens inverse les raisonnements précédents.

745. Diamètres de l'hyperbole.

Théorème. — *Dans une hyperbole, le lieu des milieux des cordes parallèles à une direction donnée est une droite passant par le centre.*

Car le lieu des milieux des segments interceptés, sur des droites parallèles à une même direction, par les côtés de l'angle fixe, formé par les asymptotes, est évidemment une droite passant par le sommet de cet angle.

Le lieu ainsi défini se nomme le *diamètre* conjugué de la direction de cordes considérée.

746. On nomme diamètres *conjugués* l'un de l'autre deux diamètres tels que le premier soit parallèle à la direction conjuguée du second.

Théorème. — *Deux diamètres conjugués forment avec les asymptotes un faisceau harmonique et réciproquement.*

Car les parallèles à l'un des rayons d'un tel faisceau sont divisées en deux parties égales par les trois autres.

Corollaires. — I. *Si un diamètre est conjugué d'un autre, réciproquement celui-ci est conjugué du premier.*

II. *Les diamètres conjugués sont des rayons correspondants d'une même involution* (**655**).

III. *Dans l'hyperbole équilatère, les diamètres conjugués sont également inclinés sur les asymptotes* (Pl., **201**).

Un diamètre passe par les points de contact des tangentes parallèles à la direction conjuguée de ce diamètre (si ces tangentes existent).

Car, les deux points d'intersection d'une tangente étant tous deux confondus avec le point de contact, il en est de même du milieu de la corde interceptée.

On voit, de plus, qu'*une tangente est divisée en deux parties égales par son point de contact et les asymptotes.*

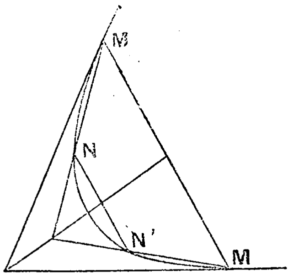

Fig. 616.

747. *Un diamètre passe par le point de rencontre des tangentes aux extrémités d'une corde conjuguée.*

Car soient MM′ (*fig.* 616) une telle corde, NN′ une corde parallèle à la première. La droite qui joint les milieux de MM′, NN′ n'est autre que le diamètre considéré : or cette droite passe (Pl., **121**) par le point d'intersection de MN, M′N′. Le diamètre passe donc aussi par l'intersection des tangentes en M, M′, car celles-ci sont les limites vers lesquelles tendent MN, M′ N′, lorsque la seconde corde se rapproche indéfiniment de la première.

748. *Deux cordes supplémentaires* (c'est-à-dire deux cordes qui ont une extrémité commune et leurs secondes extrémités diamétralement opposées) *sont parallèles à deux diamètres conjugués.* En effet (*fig.* 617), soient les deux cordes supplémentaires MN, MN′. Ces deux cordes sont respectivement parallèles aux droites qui joignent le centre O de l'hyperbole aux milieux m, m' de MN, MN′. Or Om est le diamètre conjugué de la direction MN.

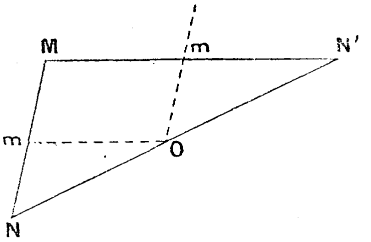

FIG. 617.

REMARQUE. — Les théorèmes que nous venons de démontrer depuis le nº **745** sont entièrement analogues à ceux que nous avons rencontrés dans la théorie de l'ellipse (**735-738**).

Des théorèmes tout semblables (sauf ceux qui regardent les diamètres conjugués) ont été également démontrés pour la parabole (**532**, **532** *bis*).

749. *De deux diamètres conjugués, il y en a toujours un qui coupe la courbe et l'autre qui ne la coupe pas.* Car deux diamètres conjugués, formant avec les asymptotes un faisceau harmonique, sont situés dans des angles différents par rapport à ces droites.

Si un diamètre d'une hyperbole ne rencontre pas la courbe, il rencontre toujours l'hyperbole conjuguée. On convient alors d'appeler *longueur* de ce diamètre la longueur b interceptée sur lui par l'hyperbole conjuguée. Les longueurs ainsi définies pour les diamètres non transverses donnent lieu à des énoncés analogues à ceux qui concernent les diamètres de l'ellipse, mais à condition de changer toujours b'^2 en $-b'^2$, comme nous l'avons déjà indiqué pour l'axe non transverse (**507**). C'est, par exemple, ce qui arrive pour les théorèmes d'Apollonius (ex. 1057) et pour l'*équation* de l'hyperbole, rapportée à deux diamètres conjugués (ex. 1058 *bis*).

EXERCICES

1045. Lieu des points qui divisent dans un rapport donné les cordes d'un cercle parallèles à une direction fixe.

1046. L'ellipse et son cercle principal sont deux figures homologiques avec centre à l'infini.

1047. Construire une ellipse, connaissant deux sommets opposés et un point ou une tangente.

1048. La corde qui joint les extrémités de deux diamètres conjugués de l'ellipse est tangente à une ellipse fixe. Lieu du point qui divise cette corde dans un rapport donné quelconque.

1049. Inscrire à l'ellipse un triangle d'aire maximum. Il y a une infinité de triangles inscrits qui ont l'aire maximum.

Même problème pour un polygone d'un nombre quelconque de côtés.

1050. Circonscrire à l'ellipse un triangle comprenant la courbe à son intérieur et d'aire maximum.

1051. Deux demi-diamètres quelconques d'une ellipse ou d'une hyperbole forment avec les tangentes menées à l'extrémité de chacun d'eux deux triangles équivalents (utiliser la symétrie oblique de la courbe par rapport à chacun de ses diamètres).

1051 *bis*. Les tangentes menées d'un point à une ellipse ou à une hyperbole sont entre elles comme les diamètres qui leur sont parallèles.

Il en est de même des cordes des arcs qu'interceptent entre elles deux sécantes parallèles.

1052. Si deux ellipses homothétiques se coupent et que, par un point de la corde commune prolongée, on leur mène des tangentes, celles-ci sont entre elles comme les diamètres de l'une des courbes qui leur sont parallèles.

1053. Quand un parallélogramme est circonscrit à une ellipse, les diagonales sont deux diamètres conjugués de la courbe. Démontrer le même théorème pour l'hyperbole.

1054. La projection parallèle d'une hyperbole sur un plan quelconque est une hyperbole.

1055. Deux hyperboles conjuguées sont obliquement symétriques l'une de l'autre par rapport aux asymptotes : c'est-à-dire (voir exercice 604) que chaque point M de l'un correspond à un point M' de l'autre, de manière que la droite MM' soit parallèle à une asymptote et coupée en son milieu par l'autre. Les points M et M' sont les extrémités (**749**) de deux diamètres conjugués.

1056. Construire les axes d'une hyperbole connaissant en grandeur et position deux demi-diamètres conjugués.

1057. Le parallélogramme construit sur deux demi-diamètres conjugués d'une hyperbole a une aire constante.

La différence des carrés de ces deux demi-diamètres est également constante.

1058. Si a', b' sont les longueurs de deux demi-diamètres conjugués de l'ellipse, l'équation de la courbe rapportée à ces deux demi-diamètres est $\frac{x^2}{a'^2} + \frac{y^2}{b'^2} = 1$.

1058 *bis*. Si a', b' sont les longueurs de deux demi-diamètres conjugués de l'hyperbole, l'équation de la courbe rapportée à ces deux demi-diamètres est

$$\frac{x^2}{a'^2} - \frac{y^2}{b'^2} = 1.$$

On prouvera que la base et la hauteur du parallélogramme dont l'aire est constante, en vertu du n° **729**, sont proportionnelles, l'une à $\frac{x}{a'} - \frac{y}{b'}$, l'autre à $\frac{x}{a'} + \frac{y}{b'}\Big)$.

1059. Si x, y sont les coordonnées de l'extrémité d'un diamètre d'une ellipse rapportée à ses axes, dont les longueurs sont a et b, les extrémités du diamètre conjugué ont leurs coordonnées x', y' données par les formules

$$\frac{x'}{a} = \pm \frac{y}{b}, \qquad \frac{y'}{b} = \mp \frac{x}{a}.$$

Trouver un énoncé analogue pour l'hyperbole.

1060. Lieu des milieux des cordes d'une conique qui passent par un point fixe. Cas de la parabole.

Lieu des points M tels qu'en les joignant à un point fixe A du plan, le segment compris entre M et chacun des points d'intersection de AM avec une conique donnée, soit égal au segment compris entre A et l'autre point d'intersection. Montrer que ce lieu ne change pas si on remplace le point A par un point déterminé quelconque du lieu.

1061. On donne un angle xOy et un point A. Lieu du point obtenu en prenant sur une sécante quelconque menée de A, à partir du point où cette sécante rencontre Ox, un segment qui soit dans un rapport donné avec le segment intercepté, sur la même sécante, entre le point A et la droite Oy.

1062. Le lieu des points qui divisent dans un rapport donné les segments interceptés par un angle fixe sur les droites issues d'un point fixe est une hyperbole passant par le point fixe et ayant ses asymptotes parallèles aux côtés de l'angle (se ramène au précédent).

1063. Trouver, sur une hyperbole donnée, deux points qui forment avec le centre de la courbe un triangle semblable à un triangle donné.

Même problème pour l'ellipse.

1064. Une tangente mobile d'une ellipse intercepte, sur deux tangentes fixes parallèles entre elles, à partir des points de contact, des segments dont le produit est constant.

Deux tangentes parallèles mobiles interceptent, sur une tangente fixe, des segments dont le produit est constant.

Étendre ces théorèmes à l'hyperbole (ex. 1052).

1065. Réciproquement, quand une droite varie de manière à intercepter, sur deux droites parallèles données, à partir de deux points donnés sur ces droites, des segments dont le produit est constant, elle est tangente à une ellipse ou à une hyperbole fixe.

1066. Sur deux parallèles données, respectivement, on prend deux points M et N tels que leurs distances à deux points fixes A et B du plan soient liées par la relation $\overline{AM}^2 - K^2 . \overline{BN}^2 =$ constante (K étant un nombre donné). Montrer que MN est tangente à une conique fixe (se ramène, en général, au précédent). Qu'arrive-t-il pour $K = 1$?

1067. Ramener le problème du n° **742** (recherche des axes d'une ellipse dont on donne en grandeur et direction deux demi-diamètres conjugués) à l'exercice 903.

La construction à laquelle on arrive est la suivante: On construit un carré ayant pour sommets opposés les extrémités A, B des deux demi-diamètres donnés. La circonférence qui passe par les deux autres sommets de ce carré et le centre donné de l'ellipse coupe la droite AB en deux points diamétralement opposés, par lesquels passent respectivement les deux axes cherchés.

1068. Les droites qui joignent deux points fixes d'une hyperbole à un point variable de la courbe interceptent sur une asymptote un segment de longueur constante et égale à celle qu'interceptent, sur la même asymptote, les parallèles à l'autre asymptote menées par les points fixes.

1069. Trouver les diamètres conjugués communs à deux coniques concentriques.

Ces diamètres existent toujours si les coniques ne sont pas toutes deux des hyperboles.

1070. Construire les directions des axes d'une conique, connaissant en direction deux couples de diamètres conjugués.

1071. Trouver, dans l'ellipse ou dans l'hyperbole, deux diamètres conjugués faisant entre eux un angle donné.

1072. Trouver, dans l'ellipse, deux diamètres conjugués égaux. Montrer que ces diamètres sont les diamètres conjugués qui font entre eux l'angle maximum.

1073. Deux diamètres conjugués d'une hyperbole équilatère sont symétriques l'un de l'autre par rapport aux asymptotes.

1074. Lieu d'un point M d'un plan tel que les droites qui le joignent à deux points donnés soient également inclinées sur une direction donnée.

1075. Par deux points fixes du plan, respectivement, on mène des parallèles à deux diamètres conjugués d'une conique donnée. Quel est le lieu du point d'intersection de ces deux droites?

1076. Deux hyperboles équilatères concentriques et qui ont leurs asymptotes à 45° les unes sur les autres, se coupent à angle droit en tous leurs points communs.

1077. On multiplie les coordonnées des points d'une ellipse E (ou d'une hyperbole) rapportée à ses axes, respectivement par des coefficients donnés.

1° Le lieu du point dont les coordonnées sont mesurées par les nombres ainsi obtenus est une nouvelle ellipse (ou hyperbole) E';

2° Si les coefficients choisis sont tels que E et E' soient homofocales et que M, N soient deux points quelconques de E; M′, N′, les points correspondants de E′, la distance MN′ est égale à la distance M′N.

(Cette proposition (ou plutôt son analogue en géométrie à trois dimensions) se présente dans une question de Mécanique céleste).

1078. La normale à une ellipse est divisée par les deux axes dans un rapport constant.

Trouver le lieu des points qui divisent l'un des segments ainsi obtenus dans un rapport donné quelconque.

1079. La perpendiculaire au milieu d'une corde quelconque de l'ellipse divise en deux parties égales le segment compris entre les sommets des rectangles OPQR (n° **741**) correspondant respectivement à ces points (Pl., **102**). Elle divise également en deux parties égales le segment intercepté sur un axe par les normales en ces points (Pl., ex. 356).

1080. Par deux points A, B d'une hyperbole, on mène des parallèles aux asymptotes. Prouver :

1° Que la seconde diagonale du parallélogramme ainsi formé passe par le centre de la courbe ;

2° Que la moitié de cette seconde diagonale est moyenne proportionnelle entre les distances du centre du parallélogramme au point où elle rencontre la tangente en A et au centre de la courbe ;

3° Que cette même moitié est moyenne proportionnelle entre les distances du centre du parallélogramme aux points où la seconde diagonale rencontre la courbe.

1081. Un demi-diamètre d'une ellipse est moyen proportionnel entre les segments déterminés, à partir du centre, par une corde conjuguée à ce diamètre et le point de concours des tangentes aux extrémités de cette corde.

1082. Même proposition pour l'hyperbole (imiter la méthode du n° **536** : on est ramené aux théorèmes des transversales). Qu'arrive-t-il si le diamètre considéré est non transverse ?

1083. Les définitions étant les mêmes qu'à l'exercice 876 :

1° Une hyperbole ayant ses asymptotes parallèles aux côtés de l'angle donné est le lieu des points tels que leurs distances hyperboliques, d'espèce déterminée, au centre de la courbe soient constantes ;

2° La tangente à cette hyperbole est hyperboliquement perpendiculaire au diamètre du point de contact ;

3° Les deux tangentes menées d'un point à cette hyperbole ont même longueur hyperbolique ;

4° Deux sécantes parallèles interceptent entre elles, sur cette hyperbole, deux arcs dont les cordes ont même longueur hyperbolique ;

5° Dans un quadrilatère circonscrit à cette hyperbole, la somme des longueurs hyperboliques de deux côtés est égale à la somme des longueurs hyperboliques des deux autres ;

6° Si on appelle *angle hyperbolique* ou *pseudo-angle hyperbolique* de deux directions, le logarithme du rapport anharmonique formé par ces directions et celles des côtés de l'angle donné, ou le logarithme de la valeur absolue de ce rapport anharmonique, suivant que ce rapport anharmonique (pris dans un ordre tel qu'il soit égal à 1 quand les deux directions coïncident), est positif ou négatif (c'est-à-dire suivant que ces directions sont celles de deux longueurs hyperboliques de même espèce ou de deux longueurs hyperboliques d'espèces différentes), et si un triangle a deux côtés de longueurs hyperboliques égales (et de même espèce), les angles ou pseudo-angles) hyperboliques opposés à ces côtés sont aussi égaux ;

7° Si les côtés d'un triangle sont de même espèce, mais que le côté AB ait une longueur hyperbolique plus grande que AC, l'angle hyperbolique ACB opposé à AC est aussi plus grand que l'angle hyperbolique ABC opposé à AC. Mais il n'en est pas de même si, les côtés AB et AC étant de même espèce, le côté BC est d'espèce différente ;

8° Une hyperbole H ayant ses asymptotes parallèles aux côtés de l'angle donné, est le lieu des points M tels que l'angle hyperbolique des droites qui joignent chacun d'eux à deux points A et B de la courbe soit constant.

Évaluer les angles hyperboliques d'un triangle, connaissant les longueurs hyperboliques de ses côtés (on commencera par le cas particulier où deux de ces côtés sont hyperboliquement perpendiculaires).

1084. Si un quadrilatère est circonscrit à une ellipse ou à une hyperbole, la somme de deux des triangles qui ont pour sommet commun le centre et pour base les côtés, est équivalente à la somme des deux autres.

Le lieu des centres des coniques inscrites à un quadrilatère est la droite qui passe par les milieux des diagonales (Pl., ex. 371).

1085. Deux sécantes parallèles à une hyperbole coupent cette courbe en quatre points tels que leurs distances à une des asymptotes soient en proportion.

1086. Le produit des segments interceptés, sur une droite, entre une hyperbole et ses deux asymptotes est égal et de signe contraire au carré du diamètre parallèle à la droite.

1087. Le produit des segments interceptés, à partir d'un point fixe, sur une sécante issue de ce point, par une hyperbole, est au produit des segments interceptés, à partir du même point, par les asymptotes, dans un rapport qui ne dépend pas de la direction de la sécante.

(Appliquer le théorème des transversales et l'ex. 1068).

1088. Théorème de *Newton*. Si, par deux points O, O′ du plan d'une ellipse, on mène deux sécantes parallèles, dont l'une coupe la courbe en m, n, l'autre en m', n', le rapport $\frac{Om \cdot On}{O'm' \cdot O'n'}$ est indépendant de la direction commune des sécantes.

Si, autour du point O, on fait pivoter une sécante, le produit des segments Om, On, interceptés par la courbe sur cette sécante, est dans un rapport constant avec le carré du demi-diamètre parallèle à la sécante.

Lieu de l'extrémité d'une longueur portée à partir de O sur la sécante et moyenne proportionnelle entre Om et On.

1089. Mêmes questions pour l'hyperbole (utiliser ex. 1087).

1090. Une corde focale d'une conique (c'est-à dire une corde passant par le foyer) est proportionnelle au carré du diamètre qui lui est parallèle.

La somme ou la différence de deux cordes focales parallèles à deux diamètres conjugués est constante.

1091. On circonscrit un cercle à un triangle inscrit dans une ellipse. Montrer que le rayon R du cercle est donné par la formule

$$R = \frac{\delta\delta'\delta''}{ab}$$

où $\delta, \delta' \delta''$ sont les demi-diamètres parallèles aux côtés du triangle; a, b, les demi-axes de l'ellipse.

(Appliquer Pl., **130** *bis* au triangle considéré et au triangle (inscrit au cercle de rayon a), dont le premier est la projection).

On suppose que les trois sommets du triangle tendent simultanément vers un même point de la courbe. Montrer que le cercle circonscrit tend vers une position limite déterminée (cercle *osculateur*) indépendante de la loi suivant laquelle les trois sommets se rapprochent de M. Construire cette position limite.

1092. Trouver le lieu du centre du cercle inscrit au triangle qui a pour sommets un point quelconque M d'une ellipse donnée et les deux foyers (exercice 1077).

Montrer que si M′ est un autre point de la courbe, C′ le cercle analogue à C construit à l'aide de M′, les droites qui joignent chacun des points M, M′ au centre du cercle correspondant ont même projection sur M M′ (ex. 1079).

En conclure que l'axe radical des cercles C, C′ passe par le milieu de M M′. La puissance de M par rapport au cercle C′ est égale à la puissance de M′ par rapport au cercle C.

CHAPITRE VIII

QUADRATURE DES CONIQUES

750. Aire de l'ellipse.

Théorème. — *L'ellipse d'axes* $2a$ *et* $2b$ *a pour aire* πab.

En effet, cette ellipse est la projection d'un cercle de rayon a et d'aire par conséquent égale à πa^2. Le cosinus de l'angle des deux plans est d'ailleurs égal (**733**, Rem.) à $\frac{b}{a}$; l'aire projetée est donc (1) (**606**. Coroll. II) $\pi a^2 \times \frac{b}{a} = \pi ab$.

Remarque. — Le même raisonnement s'applique *au secteur d'ellipse*, c'est-à-dire à la portion d'ellipse comprise entre deux demi-diamètres, car cette aire est manifestement la projection d'un secteur circulaire. On peut, dès lors (comparer Pl., **263**), évaluer toute aire limitée par des droites, des arcs de cercles et des arcs d'ellipses.

751. Aire du secteur d'hyperbole.

Nous nommons *secteur d'hyperbole* (*fig.* 618) l'aire comprise entre un arc d'hyperbole MN et les rayons qui joignent ses extrémités au centre.

On démontre (2) que la définition de l'aire (Pl., **260**) s'applique à la portion de plan ainsi définie, c'est-à-dire qu'en substituant à l'arc d'hyperbole une ligne brisée inscrite (*fig.* 619), on obtient une aire polygonale qui tend vers une limite déterminée lorsque le nombre des côtés de la ligne brisée va en augmentant indéfiniment et chacun d'eux en tendant vers zéro suivant une loi quelconque.

---

(1) Le raisonnement du n° **606** démontre l'existence de l'aire de la portion projetée, en même temps qu'il en établit l'expression.

(2) Il suffit (Pl. **260**, note 1, page 252) de montrer que la distance d'une corde au point de rencontre des tangentes en ses extrémités tend vers zéro avec la longueur de la corde : ce qu'on déduira aisément de l'exercice 1080.

Nous allons apprendre à évaluer l'aire ainsi définie. Nous pouvons d'ailleurs remarquer que cette aire est équivalente à celle qui est délimitée par l'arc MN, une des asymptotes, et les parallèles Mm,Nn (*fig.* 618) menés par les points MN à l'autre asymptote. En effet, ces deux aires s'obtiennent respectivement en retranchant de l'aire mixtiligne $OMNn$ les deux triangles rectilignes OMm,ONn, lesquels sont équivalents entre eux comme moitiés de parallélogrammes équivalents.

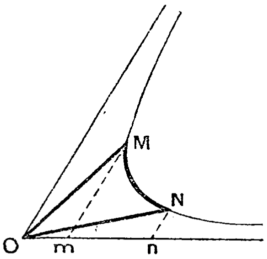

FIG. 618.

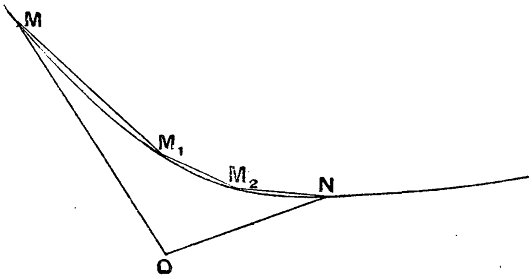

FIG. 619.

Les droites Mm,Nn sont les ordonnées des points M et N, lorsque l'hyperbole est rapportée à ses asymptotes.

Théorème. — *Dans une même hyperbole supposée rapportée à ses asymptotes, l'aire d'un secteur est proportionnelle au logarithme du rapport des abscisses des extrémités de l'arc correspondant.* Comme dans d'autres démonstrations analogues, nous prouverons tout d'abord les deux points suivants :

1° *A deux valeurs égales du rapport des abscisses correspondent des secteurs équivalents.* — Soient les deux arcs MN,M′N′ tels que le rapport des abscisses Om,On (*fig.* 620) des points M,N soit égal au rapport des abscisses Om',On' des points M′,N′. Entre M et N, pre-

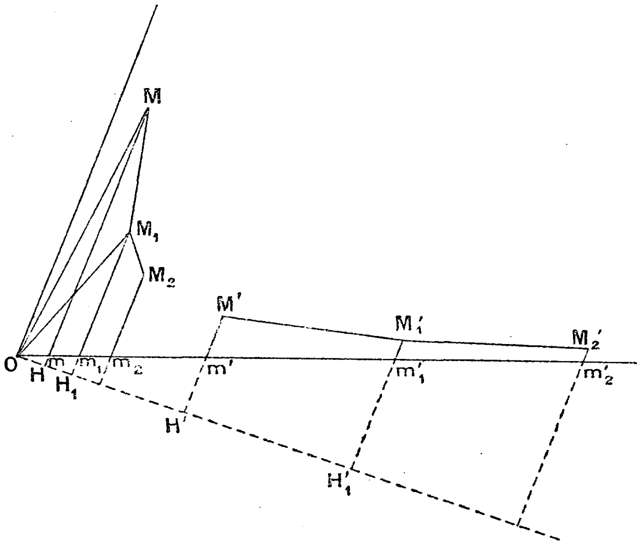

FIG. 620.

nons sur la courbe les points $M_1, M_2, \ldots$; entre M′ et N′ les points $M'_1, M'_2, \ldots$, tels que les abscisses $Om'_1, Om'_2, \ldots$ de $M'_1, M'_2, \ldots$ soient respectivement aux abscisses $Om_1, Om_2, \ldots$ de $M_1, M_2, \ldots$ comme Om' est à Om. Si nous démontrons que le secteur polygonal $OMM_1M_2\ldots N$ est équivalent à $OM'M'_1M'_2\ldots N'$, il sera établi que le secteur hyperbolique OMN est équivalent au secteur hyperbolique OM′N′, car le secteur hyperbolique OMN est, par définition, la limite du secteur polygonal $OMM_1M_2\ldots$ N lorsque les points $M, M_1M_2, \ldots$ vont en se rapprochant indéfiniment les uns des autres.

Comme tout à l'heure, le triangle OMM_1 est équivalent au trapèze MM_1mm_1 (parce que le triangle OMm est équivalent à OM_1m_1), et, de même, le triangle $OM'M'_1$ au trapèze $M'M'_1m'm'_1$. Le trapèze MM_1mm_1 a pour mesure le demi-produit de la somme $Mm + M_1m_1$ par la hauteur, que nous compterons en HH_1 (*fig.* 620) sur la perpendiculaire aux deux bases menées par le point O, et le trapèze $M'M'_1m'm'_1$ a pour mesure le demi-produit de la somme $M'm' + M'_1m'_1$ par la hauteur $H'H'_1$ comptée sur la même perpendiculaire.

Or les rapports $\frac{M_1m_1}{Mm}, \frac{M'_1m'_1}{M'm'}$ sont respectivement inverses des rapports $\frac{OH_1}{OH}, \frac{OH'_1}{OH'}$ (puisque le triangle OMm est équivalent à OM_1m et le triangle $OM'm'$ à $OM'_1m'_1$); ces rapports sont égaux entre eux, puisque l'on a $\frac{Om_1}{Om} = \frac{Om'_1}{Om'}$. On peut écrire :

$$\frac{M'm' + M'_1m'_1}{Mm + M_1m_1} = \frac{M'm'}{Mm}.$$

Quant aux hauteurs $H'H'_1$, HH_1, elles sont entre elles comme OH′ est à OH, à cause de l'égalité $\frac{OH'_1}{OH'} = \frac{OH_1}{OH}$.

Donc les trapèzes en question sont entre eux comme les produits $Mm \times OH$ et $M'm' \times OH'$, c'est-à-dire comme les triangles OMm, $OM'm'$; ces trapèzes, et aussi les triangles $OMM_1, OM'M'_1$, sont, par conséquent, équivalents entre eux.

De même, le triangle $OM'_1M'_2$ sera équivalent à OM_1M_2, et ainsi de suite. La conclusion est donc démontrée.

2° *Si l'on donne au rapport des abscisses trois valeurs telles que la troisième soit le produit des deux premières, on obtient trois secteurs* *tels que le troisième soit équivalent à la somme des deux premiers.* — C'est ce qui est évident si les deux premiers secteurs ont pour base deux arcs consécutifs MN, NP (*fig.* 621), le troisième secteur ayant pour base l'arc total MP. Mais, d'autre part, il est clair qu'on peut ramener tout autre cas à celui-là, en vertu de 1°.

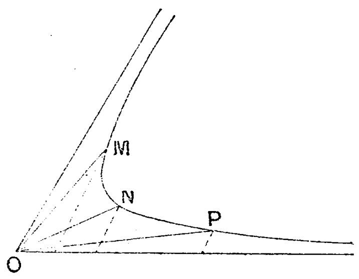

Fig. 621.

Par conséquent, *si le logarithme du rapport des abscisses prend trois valeurs telles que la troisième soit la somme des deux premières, le troisième secteur obtenu est équivalent à la somme des deux premiers :* cet énoncé revient au précédent, puisque la somme des logarithmes de deux nombres est égale au logarithme de leur produit (¹).

Des deux propositions qui viennent d'être démontrées résulte, ainsi qu'il a été expliqué à plusieurs reprises, le théorème énoncé.

752. Nous venons d'apprendre à comparer deux secteurs correspondant à une même hyperbole. Comparons maintenant deux secteurs empruntés à deux hyperboles différentes.

Nous rapporterons chacune des hyperboles à ses asymptotes. L'aire du parallélogramme ayant un sommet en un point quelconque d'une de ces hyperboles et deux côtés suivant les asymptotes de cette courbe, est une constante que nous appellerons, pour abréger, la *constante relative* à l'hyperbole considérée.

Cela posé, on aura le théorème suivant :

Théorème. — *Dans deux hyperboles dont chacune est rapportée à ses asymptotes, deux secteurs correspondant à la même valeur du rapport des abscisses sont entre eux comme les constantes relatives aux deux courbes.*

En effet, si O, O' sont les centres des deux courbes, M, M_1 deux points de la première; M', M'_1 deux points de la seconde, tels que leurs abscisses $O'm'$, $O'm'_1$ soient entre elles comme les abscisses Om, Om_1 des points M, M_1 (*fig.* 622), on démontrera, comme au numéro précédent (1°), que les triangles OMM_1, $O'M'M'_1$ sont entre eux

---

(1) Voir Bourlet, *Leçons d'Algèbre*, livre V, ch. III, n° **137**, page 426.

comme OMm, O$'$M$'m'$, c'est-à-dire comme les constantes relatives aux deux courbes. Dès lors, la même conclusion s'étendra (toujours comme au numéro précédent, 1°) au rapport de deux secteurs OMN, O$'$M$'$N$'$ tels que les abscisses des points M$'$,N$'$ soient entre elles comme celles des points M, N.

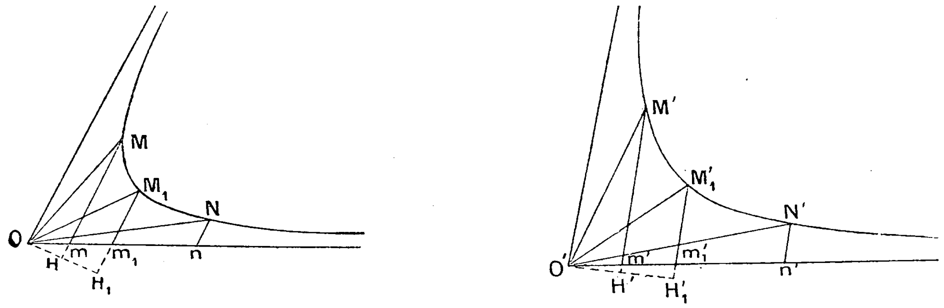

FIG. 622.

Des deux théorèmes précédents résulte :

Théorème. — *L'aire du secteur correspondant à l'arc compris entre deux points d'une hyperbole quelconque est proportionnelle au produit de la constante relative à la courbe par le logarithme du rapport des abscisses de ces deux points.*

En effet, nous venons de voir que cette aire est proportionnelle : 1° au logarithme du rapport des abscisses ; 2° à la constante relative à la courbe, lorsque ces deux quantités varient séparément : elle est donc proportionnelle à leur produit.

752 *bis*. **Corollaire.** — Tout ce que nous venons de dire subsiste quel que soit le système de logarithmes employés.

Prenons maintenant pour base du système de logarithmes le rapport des abscisses de deux points M, N, pris sur une même branche d'hyperbole et tels que l'aire du secteur compris entre ces deux points soit égale à la constante relative à la courbe. Alors l'aire d'un secteur d'hyperbole quelconque sera *égale* au produit de la constante relative à la courbe par le logarithme de rapport des abscisses correspondantes, car cette propriété a lieu pour le secteur particulier que nous venons de définir et s'étend, par suite, à tous les autres secteurs d'hyperboles, en vertu des théorèmes précédents.

Le système de logarithmes ainsi obtenu n'est autre que le système des logarithmes *népériens*, le premier qui ait été inventé (¹). Sa base est désignée ordinairement par la lettre e : les premières décimales de ce nombre sont $e = 2,7182818...$

Le nombre e n'est pas constructible à l'aide de la règle et du compas : c'est-à-dire qu'on ne peut construire à l'aide de la règle et du compas deux lignes dont le rapport soit égal à e. Plus généralement, il n'existe aucun secteur d'hyperbole tel qu'on puisse construire géométriquement le carré qui lui est équivalent, en même temps que les asymptotes de la courbe et les extrémités de l'arc qui sert de base au secteur (²).

Remarque. — *L'aire d'un secteur d'hyperbole augmente indéfiniment lorsqu'une des extrémités de l'arc correspondant s'éloigne à l'infini, l'autre restant fixe,* puisque le logarithme de l'abscisse variable augmente indéfiniment.

753. Aire du segment de parabole.

Le *segment de parabole* est l'aire comprise entre un arc de parabole et sa corde. Contrairement à ce qui a lieu pour l'ellipse et l'hyperbole, on peut construire avec la règle et le compas un carré équivalent à un segment de parabole donné quelconque.

Théorème. — *Un segment de parabole équivaut aux deux tiers du triangle formé par la corde du segment et les tangentes à ses extrémités* (³).

---

(1) Bourlet, *Leçons d'Algèbre*, page 428.

(2) Les Mathématiques supérieures permettent d'établir, entre les nombres e et π, une relation qui fait intervenir les qualités imaginaires et dont nous ne pouvons par conséquent donner une idée ici, mais dont l'existence ne doit pas surprendre, puisque l'un de ces nombres est lié à l'aire de l'ellipse (**750**) et l'autre à l'aire de l'hyperbole.

Le théorème relatif à l'impossibilité de construire le nombre e (ou plutôt une proposition plus étendue dont ce théorème découle (voir note E à la fin du volume)), est dû à M. Hermite (1874) : c'est le résultat auquel nous avons fait allusion au liv. III (Pl., page 181, note 2). Le théorème plus général que nous énonçons ensuite est celui qui a été déduit des travaux de M. Hermite par M. Lindemann. En vertu de la relation, mentionnée tout à l'heure, qui existe entre les nombres e et π, ce théorème entraîne l'impossibilité de la quadrature du cercle.

(3) Ce théorème et le mode de démonstration que nous en donnons sont dus à *Archimède*.

La démonstration suppose que les polygones inscrits se succèdent suivant une loi déterminée (chaque polygone ayant pour sommets les sommets du précédent et les points de la courbe situés sur les diamètres conjugués des côtés de ce précédent). Pour démontrer que la limite obtenue est la même, quelle que soit la loi d'inscription (pourvu que chaque côté tende vers zéro), il suffit (Pl., page 252, note 1) de faire voir que la distance d'une corde au point d'intersection des tangentes en ses extrémités tend vers zéro avec la longueur de la corde. Or, c'est ce qui résulte aisément du n° **536**.

Soit le segment de parabole déterminé par la corde aa' (*fig.* 623), les tangentes menées à la courbe en a, a' se coupant en A. Soit S l'aire du triangle Aaa'.

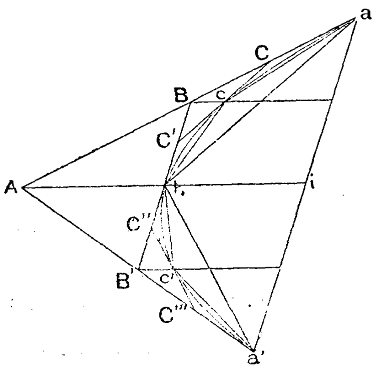

FIG. 623.

Le diamètre qui passe en A coupe la courbe en un point b où la tangente est parallèle à aa' ; soient B, B' les points où cette tangente coupe respectivement Aa et Aa'. Le diamètre Ab est divisé Ab, en deux parties égales par les points et le point i où il coupe aa' (**526**). Si donc nous joignons ab et $a'b$, les triangles iab et $ia'b$ sont respectivement les moitiés de iaA et de $ia'A$, donc, le triangle $aa'b$ a pour aire $\frac{S}{2}$. Le triangle ABB', qui est semblable à Aaa', avec le rapport de similitude $\frac{1}{2}$, a pour mesure $\frac{S}{4}$, de sorte que la somme des deux triangles bBa et $bB'a$ est égale ment $\frac{S}{4}$.

Menons maintenant les diamètres des points B et B' lesquels coupent respectivement la courbe en c, c' : les tangentes en ces points sont respectivement parallèles à ab et $a'b$; la première coupe aB en C et Bb en C' ; la seconde coupe bB' en C'' et $B'a'$ en C'''. Si nous joignons ac, cb, bc', $c'a'$, nous constaterons, comme tout à l'heure : 1° que le triangle acb est la moitié du aBb et le triangle $bc'a'$, la moitié du triangle $bB'a'$; 2° que les deux triangles BCC' et B'C''C''' sont respectivement les quarts de Bab et $B'ba'$. Dès lors, il est clair que la somme des deux triangles inscrits acb et $bc'a'$ est égale à $\frac{S}{8}$, de sorte que le polygone inscrit $acbc'a'$ a pour aire $\frac{S}{2}+\frac{S}{2\times 4}$.

En même temps on voit que la somme des deux triangles circonscrits BCC' et B'C''C''' est $\frac{S}{16}$. De plus, les deux polygones $acbc'a'$ et ACC'C''C''' sont séparés par quatre triangles intermédiaires aCc, $cC'b$, $bC''c'$, $c'C'''a'$, ayant pour somme $\frac{S}{4^2}$.

Nous continuerons de même en menant les diamètres des points C, C′, C″, C‴. Les points où ces diamètres couperont la courbe formeront, avec les points précédemment obtenus, a, c, b, c', a', les sommets d'un nouveau polygone inscrit dont l'aire sera $\frac{S}{2}+\frac{S}{2\times 4}+\frac{S}{2\times 4^2}$; l'espace compris entre ce polygone inscrit et le polygone circonscrit correspondant sera $\frac{S}{4^3}$.

Si l'on continue ainsi indéfiniment, on voit que l'aire d'un quelconque des polygones inscrits successifs est mesurée par une expression de la forme $\frac{S}{2}+\frac{S}{2\times 4}+\frac{S}{2\times 4^2}+\ldots\ldots+\frac{S}{2\times 4^m}$, c'est-à-dire par la somme d'un certain nombre de termes de la progression géométrique qui a pour premier terme $\frac{S}{2}$ et pour raison $\frac{1}{4}$. Lorsque m augmente indéfiniment, cette expression tend (¹) vers une limite égale à $\frac{S}{2\left(1-\frac{1}{4}\right)}=2\frac{S}{3}$. Le théorème est donc démontré.

L'aire compris entre un polygone inscrit et le polygone circonscrit correspondant est d'ailleurs égale à $\frac{S}{4^m}$ et tend bien vers zéro lorsque m augmente indéfiniment.

EXERCICES

1093. L'aire d'un triangle inscrit à une parabole est double de celle du triangle circonscrit formé par les tangentes en ses sommets.

Proposition analogue pour un polygone inscrit d'un nombre quelconque de côtés. En déduire l'aire du segment de parabole.

1094. Deux diamètres conjugués divisent une ellipse en quatre parties équivalentes.

1095. On mène, dans une ellipse, deux demi-diamètres tels que le secteur compris entre eux et la courbe ait une aire constante. Montrer 1° que la corde qui joint les

---

(1) Bourlet, *Leçons d'Algèbre*, livre V, ch. II, n° **133**, p. 414.

extrémités de ces deux demi-diamètres est tangente à une ellipse fixe, qui la touche en son milieu;

2° Que le segment compris entre la courbe et cette corde a également une aire constante;

3° Qu'il en est de même du triangle qui a pour côtés cette corde et les tangentes menées à ses extrémités. Lieu du point de concours de ces tangentes.

1095 *bis*. Étant donnée une ellipse, quel est le lieu d'un point pour lequel le rayon du cercle considéré à l'exercice 757 est constant?

1096. Deux sécantes parallèles interceptent sur l'hyperbole des arcs qui servent de bases à des secteurs équivalents. Démontrer cette proposition sans se servir des théorèmes sur l'aire du secteur (n^{os} **751-752** *bis*) et en utilisant seulement le théorème sur les diamètres de l'hyperbole (**745**). En déduire à nouveau l'aire du secteur d'hyperbole (ex. 1085).

1097. Résoudre, pour l'hyperbole, l'exercice 1095.

1098. En adoptant les définitions des exercices 876 et 1083 :

1° Si deux lignes brisées L et L′, terminées aux mêmes extrémités, ont tous leurs côtés de même espèce, et que L enveloppe L′, laquelle est convexe, la somme des longueurs hyperboliques des côtés de L est *plus petite* que la somme correspondante relative à L′;

2° Dans une hyperbole H ayant ses asymptotes parallèles aux côtés de l'angle donné, le rapport de la longueur hyperbolique d'une corde AB à la somme des longueurs hyperboliques des tangentes AB, BC menées en ses extrémités et terminées à leur point de rencontre C tend vers l'unité lorsque la corde tend vers zéro (mener le diamètre du point C et appliquer exercices 1080, 1083);

3° Si on inscrit à un arc AB de l'hyperbole H une ligne brisée dont le nombre des côtés augmente indéfiniment de manière que chacun d'eux tende vers zéro, la somme des longueurs hyperboliques de ces côtés tend vers une limite l, qu'on peut appeler la longueur hyperbolique de l'arc AB;

4° Deux arcs de l'hyperbole H, qui ont même longueur hyperbolique, sont vus du centre sous des angles hyperboliques égaux et correspondent à des secteurs d'hyperbole équivalents.

Plus généralement, l'aire d'un secteur de l'hyperbole H est proportionnelle à la longueur hyperbolique de l'arc qui lui sert de base, ou à l'angle hyperbolique des deux rayons qui le limitent.

1099. Si on joint un point variable d'une ellipse à deux points fixes de la courbe et qu'on mène, par le centre, des parallèles aux droites ainsi tracées, on détermine, dans l'ellipse, un secteur d'aire constante et égale à la moitié de l'aire du secteur formé par les demi-diamètres qui aboutissent aux deux points fixes.

Même problème dans l'hyperbole.

1100. Trouver le lieu d'un point tel que le segment de parabole compris entre la courbe et la corde de contact des tangentes issues de ce point ait une aire constante. Montrer que ce lieu est une parabole qui se déduit de la parabole donnée par translation. La projection de la corde de contact sur la directrice est constante (employer une remarque analogue à celle de l'exercice 1096).

1101. Mener, par un point intérieur à une conique, une corde qui détache le plus petit segment possible.

1102. Diviser un secteur d'ellipse en deux parties équivalentes par un diamètre. Même problème pour un secteur d'hyperbole (¹).

1103. Circonscrire à un triangle la plus petite ellipse possible.

(On étudiera (ex. 1049) le rapport de l'aire d'un triangle à celle d'une ellipse circonscrite et l'on en cherchera le maximum).

1103 *bis*. Inscrire, à un triangle donné, la plus grande ellipse possible (parmi celles qui sont comprises à l'intérieur du triangle).

(Même méthode).

1104. Lieu du centre d'une hyperbole qui passe par deux points fixes A, B, de manière que le secteur compris entre la courbe et les demi-diamètres aboutissant en ces points ait une aire donnée, la constante relative à la courbe étant également donnée. Montrer que les asymptotes passent chacune par un point fixe. Trouver le lieu de l'intersection des tangentes à l'hyperbole en A et B.

1105. Une ellipse varie de manière que ses axes restent dirigés suivant deux droites fixes et que son aire reste constante. Montrer que cette ellipse est tangente à une hyperbole équilatère fixe. Généraliser au cas où l'ellipse, gardant toujours une aire constante, a deux diamètres conjugués suivant deux droites données.

1106. Calculer l'aire du triangle qui a pour sommets le centre d'une hyperbole et deux points de la courbe, connaissant la constante relative à l'hyperbole et le rapport des abscisses des deux points (lorsqu'on rapporte la courbe à ses asymptotes).

1107. Montrer que le logarithme de $(1 + a)$, dans le système dont la base est le nombre e, est la limite, pour n infini, de l'expression

$$a\left(\frac{1}{n} + \frac{1}{n+a} + \frac{1}{n+2a} + \dots + \frac{1}{n+na}\right)$$

(On remarquera que cette expression représente une aire polygonale formée à l'aide d'un contour inscrit à l'hyperbole).

---

(1) Le problème de diviser un secteur d'ellipse ou d'hyperbole en un nombre de parties équivalentes qui n'est pas une puissance de 2 n'est pas résoluble par la règle et le compas.

L'introduction des imaginaires permet de réunir en un seul et même problème la question précédente relative à l'ellipse et la question relative à l'hyperbole, quoique ces deux questions soient d'apparence très différentes. (L'une conduit à la division d'un angle en un nombre quelconque de parties égales; l'autre, à trouver deux lignes dont le rapport soit $\sqrt[m]{\frac{a}{b}}$, où a et b sont deux lignes données.)

NOTE E

SUR LA RÉSOLUBILITÉ DES PROBLÈMES DE GÉOMÉTRIE

788. Ainsi que nous avons eu l'occasion de le faire remarquer à plusieurs reprises, les divers problèmes de construction que l'on peut rencontrer en géométrie sont loin de pouvoir être tous résolus comme nous l'avons indiqué au livre II de la Géométrie plane, c'est-à-dire uniquement par le moyen de la règle et du compas.

A quels caractères peut-on reconnaître si un problème donné quelconque admet une solution de cette nature? Nous ne pouvons donner à cet égard que quelques indications, car, ainsi que nous allons le voir, les méthodes qui permettent de répondre à cette question appartiennent au domaine de l'algèbre.

Nous avons vu, en effet, au livre X, comment une figure plane quelconque peut être considérée comme déterminée par des données numériques convenablement choisies : c'est ce que nous avons appelé faire le *levé* de la figure en question. Nous avons même vu qu'on peut, pour ces données, choisir exclusivement des mesures de longueur (levé à la chaîne seule, ou à la chaîne et à l'équerre). *Nous supposerons essentiellement, dans ce qui va suivre, que cette dernière précaution ait été prise:* si, par exemple, la figure considérée contient un angle, il est entendu que la détermination de cet angle sera ramenée à celle des trois côtés d'un triangle dont il fait partie.

789. Il est vrai que ces données numériques, nécessaires pour déterminer la figure, sont quelquefois en nombre infini : c'est ce qui arrive si cette figure contient des lignes de forme quelconque qu'il faut construire par points, ceux-ci étant en nombre infini. Mais ce cas ne se présentera pas dans les problèmes dont nous nous occuperons. Nous supposerons, en effet, que les figures considérées sont de celles que l'on peut reproduire à l'aide de la règle et du compas — et cela, par un nombre fini d'opérations : autrement dit, ces figures se composeront toujours de points, de droites et de cercles en nombre fini.

Par exemple, si, parmi les données de la question, figure une ellipse, il ne faudra pas supposer cette ellipse tracée — puisqu'on ne peut le faire d'un mouvement continu, en n'employant que la règle et le compas —

mais donnée par ses foyers et son grand axe, ou par cinq points (¹), ou par toutes autres données équivalentes à celles-là. De même, si une ellipse figure parmi les inconnues, il est clair qu'on ne pourra demander qu'elle soit tracée : elle devra être regardée comme obtenue lorsqu'on en connaîtra cinq points.

Toute figure satisfaisant à la condition précédente sera déterminée par un nombre fini de points, puisqu'une droite est déterminée par deux points et un cercle par trois.

790. Pour déterminer la situation d'un point, nous pourrons prendre, en particulier, les *coordonnées* de ce point (**533**), un système d'axes, que nous supposerons rectangulaires, ayant été choisi dans le plan de la figure : la connaissance de celle-ci sera alors fournie par la connaissance des coordonnées d'un certain nombre de points.

791. Soit maintenant un problème de construction, ayant pour objet la recherche d'une figure F′ ayant avec une figure donnée F des relations données.

Si nous voulions déterminer la figure F par des mesures de longueur, autrement dit en faire le levé (à la chaîne seule, ou à la chaîne et à l'équerre, bien entendu), nous obtiendrions une certaine série de nombres N. De même, le problème étant supposé résolu et la figure F′ construite, le levé de la figure formée par l'ensemble de F et de F′ conduirait à adjoindre aux nombres N d'autres nombres que nous désignerons par N′.

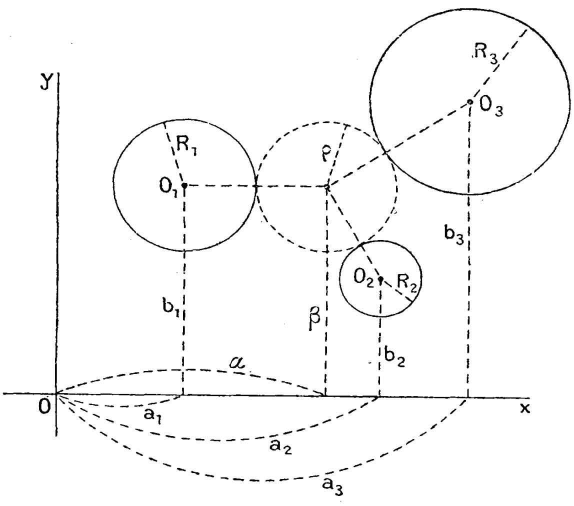

Fig. 643.

Exemple. — Dans le problème du *cercle tangent à trois cercles donnés*, les nombres N seront les coordonnées a_1, b_1; a_2, b_2; a_3, b_3 des centres de trois cercles et leurs rayons R_1, R_2, R_3 : ces nombres détermineront entièrement la figure formée par ces trois cercles (*fig.* 643). Les nombres N′ seront les coordonnées α, β du centre du cercle cherché et le rayon ρ de ce cercle.

On pourra encore considérer les distances mutuelles O_2O_3. O_3O_1, O_1O_2 (*fig.* 643) des centres des cercles donnés et les rayons R_1, R_2, R_3, comme constituant les quantités N, car ces longueurs suffisent à faire le levé de la figure formée par les cercles donnés. Les nombres N′ seront alors, par exemple, les distances x, y, z du centre du cercle cherché aux points O_1, O_2, O_3, et le rayon ρ du même cercle.

Cela posé, imaginons pour un instant qu'au lieu de *construire* la figure

---

(1) Ces deux manières de donner une conique sont équivalentes au point de vue qui nous occupe, car on peut (ex. 1175) construire, par la règle et le compas, les foyers d'une conique dont on connaît cinq points.

F′, on se propose de *calculer* les nombres N′ en supposant donnés les nombres N.

Dans ce but, il faudra former les conditions auxquelles doivent satisfaire les nombres N et N′ pour que les figures F et F′ aient entre elles les relations indiquées dans l'énoncé du problème : on aura ainsi un certain nombre d'équations qu'il faudra *résoudre* par rapport aux nombres N′. Cette résolution est un problème d'algèbre.

La condition nécessaire et suffisante pour que la figure cherchée F′ *puisse se construire à l'aide de la règle et du compas est que le calcul des nombres* N′ *puisse être effectué en ne résolvant que des équations du premier et du second degré, à une inconnue chacune* (ces équations successives étant d'ailleurs en nombre quelconque), *les coefficients de chacune d'elles étant supposés formés rationnellement avec :* 1° *des nombres* N ; 2° *des nombres déjà calculés à l'aide des équations précédentes ;* 3° *des nombres* **entiers** *connus* (¹).

Par exemple, le côté du pentagone régulier inscrit à un cercle de rayon donné R peut être construit à l'aide de la règle et du compas. Or ce côté est mesuré (Pl., **170**) par $\frac{R}{2}\sqrt{10 - 2\sqrt{5}}$ autrement dit sa valeur x est donnée par la résolution successive des équations du second degré :

$$(10\,R - z)^2 = 20\,R^2$$
$$\text{d'où } z = R\left(10 - 2\sqrt{5}\right)$$
$$4x^2 = Rz,$$

dont la première ne contient, dans ses coefficients, que le rayon donné R et les nombres entiers 10 et 20 ; la seconde, que le nombre z déjà calculé, le rayon R et l'entier 4.

792. Dans un grand nombre de cas, on trouve plus ou moins aisément (²) les conditions auxquelles doivent satisfaire les nombres N et N′ et on constate que ces conditions sont *algébriques*, c'est-à-dire s'expriment par un certain nombre d'équations de la forme

$$P = 0, \tag{23}$$

P désignant chaque fois un polynôme entier par rapport aux nombres N et N′, dont les coefficients sont des nombres entiers.

Exemple. — Dans le cas du problème des cercles tangents, les nombres cherchés α, β, ρ (voir au numéro précédent) seront déterminés par les équations

$$\begin{cases} (\alpha - a_1)^2 + (\beta - b_1)^2 = (R_1 \pm \rho)^2 \\ (\alpha - a_2)^2 + (\beta - b_2)^2 = (R_2 \pm \rho)^2 \\ (\alpha - a_3)^2 + (\beta - b_3)^2 = (R_3 \pm \rho)^2 \end{cases}$$

Si, au contraire, on s'est donné (comme nous l'avons supposé en second lieu) les

---

(1) Nous avons donné, en Géométrie plane (**155**), le moyen de construire une longueur donnée par une équation du second degré.

(2) Cette partie du problème est, en général, du ressort de la géométrie analytique.

distances mutuelles des points O_1, O_2, O_3, au lieu des coordonnées de ces mêmes points, on écrira les relations de contact

$$(24) \qquad x = \pm R_1 \pm \rho, \quad y = \pm R_2 \pm \rho, \quad z = \pm R_3 \pm \rho,$$

auxquelles on devra adjoindre l'équation obtenue en subtituant x, y, z, O_2O_3, $O_3 O_1$, $O_1 O_2$dans la relation (5′) (nº **612**).

On obtient bien ainsi quatre équations de la forme (23).

793. Si, après avoir formé ces équations, on réussit à les résoudre par le moyen d'équations du premier et du second degré satisfaisant aux conditions précédemment indiquées, on aura décidé dans le sens de l'affirmative la question posée : *Peut-on construire la figure* F′ *à l'aide de la règle et du compas?*

EXEMPLE. — Dans le problème des cercles tangents, si l'on remplace x, y, z (notations des deux numéros précédents) par leurs valeurs $\pm R_1 \pm \rho$, $\pm R_2 \pm \rho$, $\mp R_3 \pm \rho$ tirées de l'équation (24) et que l'on reporte ces valeurs dans l'équation (5′) du nº **612**, on aura pour ρ une équation du second degré (¹).

Le problème est donc bien résoluble.

Si, au contraire, les divers essais auxquels on peut se livrer pour effectuer cette solution échouent, cela peut provenir de ce que l'on a mal opéré ou, au contraire, de ce que cette résolution est impossible à obtenir. Rien, au premier abord, ne force à considérer l'une des deux explications comme la véritable plutôt que l'autre.

Il est clair que le problème qui consiste à décider entre ces deux explications est des plus difficiles. Il existe, en effet, une infinie variété de méthodes que l'on peut essayer d'appliquer à la résolution proposée, une infinité de manières dont on peut combiner des équations du premier et de second degré pour arriver à cette résolution, et il s'agit de savoir si *aucune* de toutes ces méthodes ou de ces combinaisons *imaginables* ne permet d'atteindre le but.

Aussi la réponse n'a-t-elle pu être fournie que dans ce siècle, par le moyen de la théorie des groupes (²) : cette théorie permet de décider en toute certitude la question précédente (toutes les fois que les conditions qui déterminent les N′ sont algébriques) : autrement dit, elle permet, — ou bien d'opérer la résolution demandée, — ou bien de démontrer en toute rigueur que cette résolution est impossible.

Parmi les problèmes dont l'impossibilité (au point de vue de la construction par la règle et le compas) a pu être ainsi démontrée, nous citerons :

(¹) C'est ce que l'on reconnaît en formant d'abord les binômes $y^2—z^2$, $z^2—x^2$, $x^2—y^2$ (qui figurent dans la relation du nº **612**) et constatant qu'ils sont du premier degré en ρ.

(²) La notion de *groupe* a été introduite, précisément en vue du problème dont nous parlons en ce moment et de problèmes analogues, par divers mathématiciens du commencement de ce siècle; mais les découvertes capitales sur ce sujet sont dues à GALOIS (savant français, mort à vingt ans (1826-1846)), qui a pu, grâce à elles, résoudre complètement les problèmes en question. En même temps, il reconnaissait l'importance générale de la notion de groupe, qui, depuis ses travaux, tend à jouer un rôle prépondérant dans toutes les parties des mathématiques.

Le problème de la *trisection de l'angle*, qui consiste à diviser un angle, donné arbitrairement, en trois parties égales (ou plus généralement, en un nombre quelconque de parties égales, ce nombre n'étant pas une puissance de 2);

Le problème de la *duplication du cube*, ou problème *délique* (¹), qui consiste à trouver un cube dont le volume soit double de celui d'un cube donné, — ou, d'une manière générale, à trouver deux lignes dont le rapport soit égal à $\sqrt[3]{n}$, n étant le rapport de deux lignes données (²) ;

Le problème de l'inscription du polygone régulier de N côtés, lorsque N n'appartient pas à l'une des catégories énumérées au n° **173** de la Géométrie plane;

La recherche des points communs à deux coniques données quelconques (³), ou (ce qui revient au même) celle de leurs sécantes communes ou celle de leurs pôles doubles;

La recherche des plans de symétrie et des sections circulaires d'un cône ayant pour base une conique (**757**) (⁴).

794. Problèmes transcendants.

L'impossibilité du problème est démontrée à plus forte raison, si l'on constate que les relations données ne peuvent pas s'exprimer par des équations algébriques à coefficients entiers (de la forme (23)) entre les données N et les inconnues N′. Le problème est alors dit *transcendant*.

Seulement, on peut se trouver en présence d'une difficulté tout analogue à celle que nous avons rencontrée tout à l'heure. Pour affirmer qu'un problème est transcendant, il faut *prouver* qu'il est impossible de déterminer les inconnues du problème par des équations algébriques de la forme indiquée.

On ne possède pas de méthode entièrement générale pour triompher de cette dernière difficulté.

Mais elle a pu être surmontée pour les deux principaux problèmes transcendants que nous ayons rencontrés en géométrie élémentaire, ceux qui sont relatifs aux nombres e et π. Les découvertes de MM. Hermite et Lindemann à cet égard ont précisément consisté à démontrer que *l'un comme l'autre de ces nombres n'est racine d'aucune équation algébrique à coefficients*

---

(1) La légende raconte qu'Apollon, pour faire cesser la peste qui ravageait l'île de Délos, aurait demandé que l'on doublât le volume de l'autel élevé en son honneur, et qui était de forme cubique : d'où le nom donné à ce problème.

(2) Ainsi que nous l'avons dit (ex. 1102, note), on est conduit, en algèbre, à réunir en une seule cette question et celle de la trisection de l'angle.

(3) On suppose, bien entendu, que les coniques sont données ainsi qu'il est expliqué au n° **789**. On voit ici qu'il n'est pas indifférent de savoir si l'on applique ou non la règle formulée en cet endroit : car si les coniques étaient entièrement tracées, le problème qui consiste à trouver leurs intersections serait tout résolu.

(4) Nous avons indiqué (particulièrement au livre II) un certain nombre de problèmes ayant pour objet la construction d'un triangle déterminé par trois éléments. Si l'on choisit ces trois éléments de toutes les façons possibles parmi les côtés, les hauteurs, les médianes, les bissectrices intérieures ou extérieures, les rayons des cercles circonscrits, inscrits et exinscrits, on obtient 244 questions (en ne comptant que celles qui sont véritablement distinctes les unes des autres). Sur ces 244, on constate que 69 au moins ne sont pas résolubles par la règle et le compas (*Korselt*, Archiv. der Math. und Phys., Leipzig, 1900.)

entiers. Par conséquent aussi il est impossible de construire avec la règle et le compas deux lignes dont le rapport soit égal à e ou deux lignes dont le rapport soit égal à π: c'est de là que ressort l'impossibilité de la quadrature du cercle.

La recherche de la longueur d'un arc de cercle donné quelconque est également un problème transcendant : il n'existe pas d'angle que l'on puisse construire avec la règle et le compas, en même temps que la longueur de l'arc qu'il intercepte sur un cercle de rayon donné.

795. Constructions effectuées par la règle seule.

Quels sont les problèmes de construction que l'on pourrait résoudre si, au lieu d'une règle et d'un compas, on ne pouvait employer que la règle ?

Avant de répondre à cette question, il est nécessaire de spécifier comment l'on entend l'emploi de la règle. Nous supposerons que celle-ci ait un bord rectiligne *et un seul*, l'instrument étant, dans tous les autres sens, terminé par des bords de forme inconnue, dont on s'interdit, d'ailleurs, de faire usage, de quelque façon que ce soit, au cours des constructions. Autrement dit, les seules opérations que nous pourrons effectuer avec notre règle seront les suivantes :

1° Tracer la droite qui passe par deux points donnés ;

2° Marquer le point commun à deux droites données.

Cela posé, nous considérerons encore la figure donnée F et la figure cherchée F′ comme déterminées par des nombres. Mais le choix entre les différents modes de levé possibles, choix qui était indifférent dans le cas où l'on pouvait se servir de la règle et du compas, ne l'est plus ici : nous supposerons essentiellement que les longueurs mesurées, pour déterminer les deux figures ne soient autres que les coordonnées de leurs différents points (voir n° **790**) par rapport à deux axes déterminés.

Dans ces conditions, *il faudra que les nombres inconnus* N′ *soient déterminés exclusivement par des équations du premier degré* formées rationnellement à l'aide des nombres connus N et de nombres entiers.

Cette condition est nécessaire ; mais elle n'est pas suffisante. Par exemple, le problème de *mener par un point donné* A, *une parallèle à une droite donnée* D, satisfait à cette condition (¹) : il est aisé, cependant, de voir que ce problème ne peut pas être résolu à l'aide de la règle seule.

Supposons, en effet, que l'on ait pu obtenir la solution dans ces conditions. Faisons la perspective de la figure considérée sur un autre plan P′ : le point A aura pour perspective un point A′ et la droite D, une droite D′.

Or les deux opérations précédemment indiquées comme exécutables avec la règle, sont toutes deux projectives. Par conséquent, la série des constructions effectuées dans le plan P, à partir du point A et de la droite

---

(1) C'est-à-dire qu'étant données les coordonnées du point A et celles de deux points de la droite D, on peut déterminer, par des équations du premier degré, les coordonnées d'un second point de la parallèle cherchée.

D aura pour image, sur le plan P′, une série de constructions toutes semblables effectuées à l'aide du point A′ et de la droite D′. Cette dernière série de constructions devrait dès lors fournir la parallèle menée par le point A′ à la droite D′ : ce qui est absurde, puisque nous savons que cette parallèle n'*est pas* la perspective de la parallèle menée par A à D.

Il est clair qu'on peut raisonner de même dans tout autre cas analogue, et que la conclusion est la suivante : *pour qu'un problème soit résoluble à l'aide de la règle seule, il faut que son énoncé ne contienne* (ou tout au moins *puisse être amené à ne contenir*) *que des propriétés projectives*, du moins *tant qu'on ne connaît, entre les éléments de la figure donnée, aucune relation non projective.*

796. Inversement, l'ensemble des deux conditions que nous venons d'indiquer : à savoir, *que le problème soit du premier degré* et *qu'il soit projectif*, est *suffisant* pour que la construction soit possible par la règle seule. C'est ainsi que l'on peut construire, avec la règle, l'homologue d'un point quelconque, dans une homographie déterminée par trois couples de points homologues (**645**) ; le conjugué harmonique d'un point d'une droite par rapport à un segment de cette droite (Pl., **203**) ; le second point d'intersection d'une conique avec une droite menée par un point connu de la courbe ; etc.

797. Ainsi que nous l'avons spécifié dans l'énoncé qui vient d'être donné, la conclusion peut se trouver modifiée si l'on connaît entre les éléments de la figure donnée une relation non projective.

Supposons, par exemple, que l'on donne quelque part dans le plan de la figure un parallélogramme. Alors on pourra mener, par un point quelconque, une parallèle à une droite quelconque (1) (voir ex. 895) ; par suite aussi, on peut (Pl., **151**) diviser une droite en segments proportionnels à des segments donnés *d'une même droite*, ou encore (en combinant cette construction avec l'exercice 129) diviser une droite donnée en un nombre quelconque donné de parties égales. D'une manière générale, on peut résoudre, dans ces conditions, *tout problème du premier degré dont l'énoncé se conserve par projection parallèle.*

798. Enfin si, dans le plan du dessin, on se donne un carré, il n'y a plus de condition de projectivité : il suffit que le problème soit du premier degré (exemple : mener d'un point une perpendiculaire sur une droite) (2).

---

(1) La possibilité de ce problème, dans les conditions indiquées, résulte de ce que le problème suivant : *mener, par un point donné, une droite coupant une droite donnée L en un point situé sur une droite* Δ *non directement donnée, mais qui passe par l'intersection de deux droites données* A, B *et par l'intersection de deux droites données* C, D, est possible par la règle seule (car il est du premier degré et projectif). Lorsque A et B sont parallèles entre elles ainsi que C, D, la droite Δ est la droite de l'infini, et on retombe sur le problème proposé. La seule difficulté nouvelle provient de ce qu'ici la droite Δ ne peut être tracée. On a vu, à l'exercice 895, que cette difficulté peut être tournée en construisant une figure homographique de la première et où l'homologue de Δ soit à distance finie.

(2) On peut concevoir que ce problème soit résoluble dans ces conditions par la règle seule, si l'on remarque que son énoncé peut être mis sous la forme suivante. *Étant donnés un*

799. On peut se demander également quelles constructions sont possibles à l'aide du compas seul. La réponse peut se donner en un mot : *on peut résoudre, par le compas seul, les mêmes problèmes qu'avec la règle et le compas* (1).

800. Admettons enfin qu'on n'ait à sa disposition qu'une règle, mais que cette règle ait *deux bords* que l'on sache *rectilignes et parallèles entre eux*. On démontre également que l'*usage de cet instrument remplace entièrement celui de la règle et du compas*. Il en est de même si, outre la règle (à un seul bord), on a à sa disposition une équerre, même si l'angle de cette équerre n'est pas droit, pourvu que ses bords soient bien rectilignes.

801. Enfin, rappelons que, dans ce que nous avons dit précédemment sur l'usage de la règle, nous avons supposé (d'une manière analogue à ce qui a été dit au n° **789** pour les problèmes constructibles par la règle et le compas) que les données comprenaient exclusivement des figures exécutables avec la règle seule, autrement dit, des droites et des points : les conclusions précédemment énoncées ne sont nécessairement vraies que moyennant cette restriction. C'est ainsi que nous avons appris (Pl., **211**) à tracer par la règle seule les tangentes menées d'un point à un cercle, problème qui n'est pas du premier degré : seulement le cercle était tracé d'avance.

On démontre qu'il suffit d'avoir tracé, une fois pour toutes, *un seul* cercle dont le centre soit connu, pour que tout problème constructible par la règle et le compas puisse être ensuite construit à l'aide de la règle seule.

NOTE F

SUR LA DÉFINITION DES VOLUMES

802. Nous avons admis dans le texte (liv. VI) qu'à chaque polyèdre on peut faire correspondre une grandeur appelée *volume*, possédant les deux propriétés suivantes :

I. *Deux polyèdres égaux ont le même volume, quelles que soient leurs situations dans l'espace.*

---

point P *et sept droites* A, A′, B, B′ (les côtés du carré), C, D (les diagonales), L (la droite donnée), *mener par le point* P *une droite* M, *telle que, sur la droite* Δ (ici, droite de l'infini) *qui joint le point d'intersection de* A, A′ *au point d'intersection de* B, B′, *les droites* A, B ; C, D ; L, M *interceptent trois segments en involution*. La difficulté provenant de ce que Δ ne peut être tracée se tourne comme précédemment (voir note précédente).

(1) Toutefois, bien entendu, si la figure cherchée comprend des droites, on doit considérer une de ces droites comme obtenue dès qu'on en connaît deux points.

II. *Le polyèdre* P'', *somme de deux polyèdres adjacents* P *et* P', *a pour volume la somme des volumes de* P *et de* P'.

Nous allons montrer qu'on peut effectivement réaliser une pareille correspondance. La marche suivie à cet effet sera toute semblable à celle que nous avons adoptée en Géométrie plane (note D), ce qui nous dispensera d'insister sur certains points que nous avons développés en cet endroit.

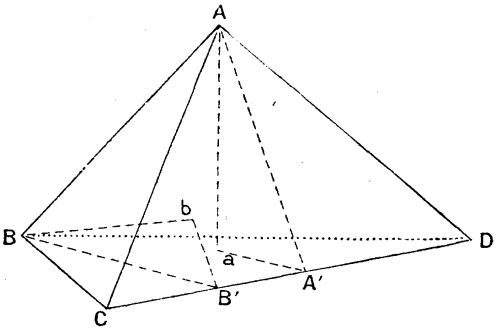

FIG. 644.

Théorème. — *Dans tout tétraèdre, chaque face, multipliée par la hauteur correspondante, donne le même produit.*

Considérons, dans le tétraèdre ABCD (*fig.* 644), les deux faces BCD, ACD, auxquelles correspondent respectivement les hauteurs Aa, Bb. Les deux triangles BCD, ACD, ayant la base commune CD, ont entre eux le même rapport que leurs hauteurs BB', AA'. Il suffit de montrer que ce rapport est inverse du rapport $\frac{Aa}{Bb}$, c'est-à-dire que l'on a

$$\frac{AA'}{Aa} = \frac{BB'}{Bb}.$$

Or c'est ce qui se voit dans les triangles AaA', BbB', rectangles, l'un en a, l'autre en b, et qui ont tous deux un angle aigu égal au dièdre A.CD.B, ou à son supplément (comparer **360**).

Donc on a aussi : $$\frac{\text{surf. BCD}}{\text{surf. ACD}} = \frac{Bb}{Aa}$$

ou $$Aa \times \text{surf. BCD} = Bb \times \text{surf. ACD}.$$

C. Q. F. D.

La valeur commune des produits précédents, multipliée par une constante numérique *K* dont nous choisirons la valeur un peu plus loin, sera dite le *volume* du tétraèdre. Ce volume est nul si les quatre points A, B, C, D sont dans un même plan, et dans ce cas seulement.

803. Nous allons maintenant démontrer la proposition qui correspond à celle que nous avons établie au n° **315** (Pl., note D). Mais, afin de ne pas avoir à distinguer un grand nombre de dispositions de figures, nous considérerons les volumes comme affectés de signes, d'après la convention suivante :

Nous désignerons par (ABCD) le volume du tétraèdre ABCD, précédé du signe + ou du signe —, suivant que ce tétraèdre est *direct* ou *rétrograde*, c'est-à-dire suivant que le trièdre A.BCD est direct ou rétrograde. Le signe de l'expression (ABCD) dépend évidemment de l'ordre dans lequel on nomme les quatre points A,B,C,D; il est aisé de voir qu'il change quand on permute deux d'entre eux (1).

804. Le signe de l'expression (ABCD) change lorsque A, B, C restant fixes, le point D change de côté par rapport au plan ABC, car alors le dièdre C.AB.D change le sens. Ce signe change dès lors aussi lorsque l'un quelconque des quatre points, A, par exemple, change de côté par rapport au plan des trois autres qui restent fixes, puisque (ABCD) est égal et de signe contraire à (DBCA).

Le point A restant fixe ainsi que le plan BCD, le signe de l'expression (ABCD) dépend (Pl., **20**) du sens de rotation du triangle BCD dans ce plan.

805. Théorème. — A, B, C, D, E *étant cinq points quelconques de l'espace, on a toujours*

$$(EBCD) + (AECD) + (ABED) + (ABCE) = (ABCD). \qquad (25)$$

1° Nous traiterons d'abord le cas particulier où le point E est dans le plan ABC (*fig.* 645). Le volume du tétraèdre EABC est alors nul et l'équation précédente pourra s'écrire (2)

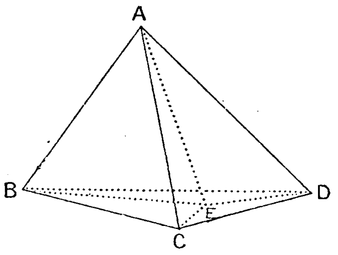

Fig. 645.

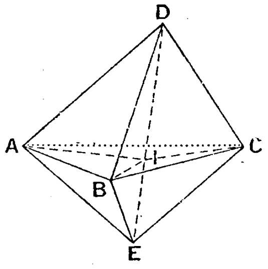

Fig. 646.

$$(EBCD) + (ECAD) + (EABD) = (ABCD).$$

Or les quatre tétraèdres EBCD, ECAD, EABD, ABCD ont la hauteur abaissée du point D commune. Leurs volumes seront donc proportionnels à leurs bases et les expressions (EBCD), (ECAD), (EABD), (EBCD) seront proportionnelles à ces mêmes bases affectées des signes + ou — suivant leurs sens de rotation (puisque ceux-ci correspondent (n° précédent) aux dispositions des tétraèdres). Or, dans ces conditions, nous savons que la base ABC est la somme algébrique des bases EBC, ECA, EAB (Pl., **315**).

---

(1) Cela est évident (**371**) si, parmi les sommets permutés ne figure pas A ; d'autre part, si A est permuté avec C, par exemple, le dièdre B.AC.D changera de sens (puisque, les faces étant les mêmes et nommées dans le même ordre, on aura changé le sens choisi sur l'arête) : or le sens de ce dièdre décide de la disposition du trièdre A.BCD.

(2) L'expression (ECAD), par exemple, est égale à (AECD), car toutes deux sont (**803**) égales et de signe contraire à (EACD).

2° Passons maintenant au cas général. Soit I le point de rencontre de la droite DE et du plan ABC (*fig.* 646); on aura (1°)

$$(26) \qquad (IBCD) + (AICD) + (ABID) = (ABCD).$$

Mais, E étant dans les plans ADE, BDI, CDI, ABC, on a aussi

$$\begin{aligned} (IBCD) &= (EBCD) + (IECD) + (IBED) + (IBCE) \\ (AICD) &= (EICD) + (AECD) + (AIED) + (AICE) \\ (ABID) &= (EBID) + (AEID) + (ABED) + (ABIE) \\ 0 = (ABCI) &= (EBCI) + (AECI) + (ABEI) + (ABCE) \end{aligned}$$

Portant ces valeurs dans l'équation (26), il vient bien

$$(ABCD) = (EBCD) + (AECD) + (ABED) + (ABCE)$$

[Tous les autres termes se détruisent deux à deux comme dérivant l'un de l'autre par permutation des lettres I et E (par exemple (IECD) et (EICD)].

806. Corollaire. — Si l'on convient d'appeler *additif* ou *soustractif* (comparer Pl., **315**) chacun des tétraèdres EBCD, AECD, ABED, ABCE, suivant qu'il est ou non du même côté que le tétraèdre ABCD, par rapport à la face commune, le volume du tétraèdre ABCD est égal à la somme des volumes des tétraèdres additifs, diminuée de celle des tétraèdres soustractifs.

En effet, dans l'équation (25), supposons que l'ordre des points ABCD ait été pris tel que le tétraèdre ABCD soit direct. Les expressions (EBCD),... seront positives ou négatives suivant que la disposition des tétraèdres correspondants sera ou non la même que celle du tétraèdre ABCD, c'est-à-dire suivant qu'ils seront additifs ou soustractifs.

Inversement, le théorème du n° **315** donne, entre quatre points d'un plan, la relation

$$(OBC) + (OCA) + (OAB) = (ABC)$$

où (ABC) est l'aire du triangle ABC précédée du signe + ou du signe —, suivant le sens de rotation de ce triangle.

807. On appellera *volume d'une pyramide quelconque* le produit de la base par la hauteur et par la constante K. Il est évident que ce volume est la somme des tétraèdres obtenus en décomposant la base en triangles d'une façon quelconque.

808. Théorème. — *Soient un polyèdre décomposé, d'une façon quelconque, en un nombre quelconque de tétraèdres, et un point quelconque O de l'espace, que l'on joint à tous les sommets. Une quelconque des pyramides qui ont pour sommet commun O et pour bases respectives les faces du polyèdre étant considérée comme additive ou comme soustractive, suivant qu'elle est ou non du même côté que le polyèdre, par rapport à la face commune, la différence S entre la somme des volumes des pyramides additives et la somme des volumes* *des pyramides soustractives* (s'il en existe) *est égale à la somme* Σ *des tétraèdres en lesquels le polyèdre a été décomposé.*

Corollaire. — *La quantité* S *est indépendante du choix du point* O *et la quantité* Σ, *du mode de décomposition du polyèdre en tétraèdres.*

Démonstrations calquées sur celles du n° **316**.

La quantité ainsi définie sera dite le *volume* du polyèdre : elle possède les deux propriétés fondamentales énoncées tout à l'heure (comparer **317**).

Les volumes ainsi définis coïncident avec ceux que nous avons mesurés dans le texte, lorsqu'on prend $K = \frac{1}{3}$: c'est la valeur que doit avoir cette constante pour que le cube d'arête 1 ait son volume égal à l'unité. Les raisonnements donnés dans le texte prouvent qu'il n'y a qu'une seule manière de déterminer les volumes de manière à satisfaire à cette condition et à celles que nous avions imposées en commençant.

Il est impossible de décomposer un polyèdre en polyèdres partiels qui, autrement assemblés, forment un polyèdre intérieur au premier. Cette proposition résulte des raisonnements que nous venons de présenter et non de ceux du livre VI (comparer **318**).

NOTE G

SUR LES NOTIONS DE LONGUEUR, D'AIRE ET DE VOLUME RELATIVES A DES LIGNES ET A DES SURFACES QUELCONQUES

809. Longueur d'un arc de courbe gauche.

Un arc AB d'une courbe gauche (C) étant donné, projetons-le orthogonalement sur un plan quelconque, suivant un arc *ab* (*fig*. 643) d'une courbe (*c*). Supposons que la définition de la longueur d'une courbe plane donnée dans la note de la page 173 (Pl., liv. III) s'applique à la courbe (*c*) : on pourra, en d'autres termes, mesurer la longueur d'un arc quelconque de (*c*). Si les hypothèses énumérées dans la note en question sont vérifiées, le rapport d'un arc de (*c*) à sa corde tendra vers l'unité quand l'arc tendra vers zéro ; et même, $1 + \alpha$ étant un nombre aussi peu supérieur qu'on le veut à 1, on pourra toujours (¹) trouver un nombre ε tel que, pour tous

---

(¹) En effet la longueur d'un arc *pq* faisant partie de *ab* est (la courbe *ab* étant supposée convexe le long de cet arc) plus petite que la somme des tangentes en *p* et en *q*, prolongées jusqu'à leur point d'intersection et les hypothèses faites à l'endroit cité de la Géométrie plane les arcs faisant partie de l'arc ab et dont la corde est moindre que ε, ce rapport soit plus petit que $1 + \alpha$. Nous pourrons développer sur un plan (**538**) le cylindre qui projette la courbe AB, celle-ci venant suivant une nouvelle courbe A_1B_1 et sa projection ab suivant un segment de droite a_1b_1 (*fig.* 647) égal à la longueur de l'arc ab. Supposons que la définition de la longueur d'un arc de courbe plane s'applique également au développement de (c) et soit l la longueur de l'arc A_1B_1.

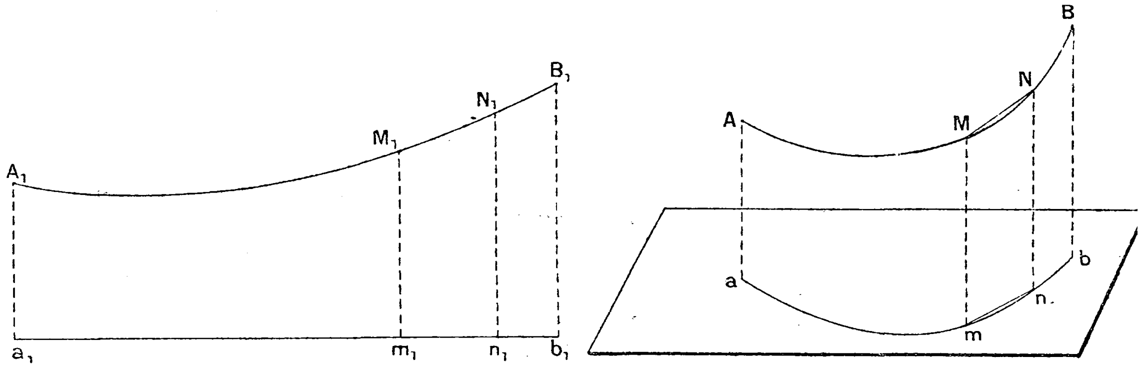

FIG. 647.

Dans ces conditions, je dis que le *périmètre d'une ligne polygonale inscrite à l'arc* AB *et ayant ses extrémités en* A *et* B *tend vers* l, *lorsque le nombre des côtés de cette ligne brisée augmente indéfiniment de manière que chacun d'eux tende vers zéro.*

Soient, en effet, M, N deux sommets consécutifs de la ligne polygonale ; m, n leurs projections sur le plan de (c) ; M_1, N_1, m_1, n_1, les points correspondants à M, N, m, n dans le développement du cylindre. On a $m_1M_1 = mM$; $n_1N_1 = nN$; mais mn, corde d'un arc de (c), est plus petit que m_1n_1 (lequel est égal à la longueur de cet arc).

Portons le trapèze $mMnN$ en $m_1M_1n'N'$ sur le trapèze $m_1M_1n_1N_1$ (*fig.*643 *bis*) de manière que mM vienne suivant m_1M_1 et mn suivant la direction m_1n_1. On aura évidemment $MN = M_1N' < M_1N_1$; mais *le rapport* $\frac{M_1N_1}{M_1N} = \frac{M_1N_1}{MN}$ *sera plus petit que le rapport* $\frac{m_1n_1}{m_1n'}$ (c'est-à-dire que $\frac{m_1n_1}{mn}$). Car $\frac{M_1N_1}{M_1N'} - 1$ ou $\frac{M_1N_1 - M_1N'}{M_1N'}$ est plus petit que $\frac{N\,N'}{M_1N'}$, lequel est à son tour plus petit que $\frac{nn'}{m_1n'}$, c'est-à-dire que $\frac{m_1n_1}{m_1n'} - 1$.

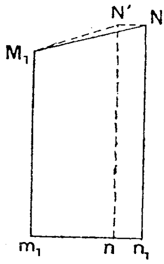

FIG. 647 *bis*.

Or nous avons vu que l'on pouvait rendre le rapport $\frac{m_1n_1}{mn}$ plus petit que $1 + \alpha$, quel que soit le nombre positif α, et qu'il suffisait, pour cela, de

consistent précisément à admettre que le rapport de cette somme à la corde pq devient plus petit que $1 + \alpha$, toutes les fois que l'arc pq, intérieur à ab, a une corde inférieure à une quantité convenablement choisie ε.

prendre mn inférieur à une certaine quantité ε. Si donc nous avons pris tous les côtés MN de la ligne polygonale (et par conséquent, *a fortiori*, toutes les cordes mn) inférieures à ε, tous les rapports tels que $\frac{M_1N_1}{MN}$ sont plus petits que $1 + \alpha$: d'après une proposition souvent rappelée d'arithmétique, il en est de même du rapport formé avec la longueur de la ligne brisée $A_1 \dots M_1N_1 \dots B_1$, somme de toutes les cordes M_1N_1, et la longueur de la ligne polygonale A... MN... B, dont nous sommes partis et qui est la somme des cordes telles que MN.

Autrement dit (puisque α peut être pris aussi petit qu'on veut) le rapport $\frac{A_1 \dots M_1N_1 \dots B_1}{A. \; M\,N \dots B}$ tend vers 1, lorsque les côtés de la ligne polygonale A,.. M N... B tendent tous vers zéro.

Or le numérateur tend, par hypothèse, vers l : il en est donc de même du dénominateur. C. Q. F. D.

Il résulte de là, en particulier, qu'en projetant la courbe (c) sur un plan différent du premier et répétant les mêmes opérations dans ces nouvelles conditions, la longueur l trouvée ne pourrait avoir une valeur autre que la première.

La longueur l, limite vers laquelle tend le périmètre d'un polygone inscrit à l'arc AB et dont tous les côtés tendent vers zéro, est dite la *longueur de l'arc de courbe* AB. Il est clair, d'après ce qui précède, que, *si on développe un cylindre sur un plan, toute ligne tracée sur la surface a même longueur que son développement.*

La ligne droite est, dans l'espace, le plus court chemin d'un point A *à un autre* B. Car tout autre chemin est, soit une ligne brisée, plus longue que la ligne droite, soit une ligne courbe dont la longueur est la limite d'une ligne brisée (¹).

Le rapport d'un arc de courbe gauche à sa corde tend vers l'unité lorsque l'arc tend vers zéro, car l'arc MN de la courbe AB (*fig.* 647) est égal à l'arc M_1N_1 de la courbe A_1B_1 ; le rapport $\frac{\text{arc } M_1N_1}{\text{corde } M_1N_1}$ tend vers l'unité et nous avons vu qu'il en est de même pour le rapport $\frac{\text{corde } M_1N_1}{\text{corde } MN}$.

810. Plus court chemin d'un point à un autre sur la sphère.

Appliquons ce qui précède à une ligne sphérique AB. Inscrivons dans cette ligne le polygone A... MN... B, puis remplaçons chaque côté par l'arc de grand cercle qui a mêmes extrémités : nous obtenons ainsi une ligne composée d'arc de grands cercles successifs, ou *ligne brisée sphérique*.

Lorsque les côtés du polygone tendent vers zéro, le rapport de chaque

---

(1) La longueur de la ligne courbe ne peut être *égale* à celle de la ligne droite, car, si C est un point extérieur à la droite, mais situé sur la courbe, celle-ci est au moins égale à AC + BC, lequel est plus grand que la droite AB.

arc de grand cercle à la corde correspondante tend vers 1, et il en est de même, par conséquent, du rapport de la ligne brisée sphérique à la ligne brisée rectiligne, de sorte que ces deux quantités ont même limite. Ainsi *la longueur d'une courbe sphérique est la limite vers laquelle tend la longueur d'une ligne brisée sphérique inscrite, dont tous les côtés tendent vers zéro.*

L'arc de grand cercle qui joint A et B est plus court que toute ligne brisée sphérique joignant les mêmes points. Il est donc aussi plus court que toute courbe sphérique allant de A en B : *l'arc de grand cercle est le plus court chemin entre deux points de la sphère.*

811. Considérons un cône quelconque, ou la portion d'un cône comprise entre deux génératrices OA, OB (*fig.* 648). Une sphère quelconque décrite du sommet O comme centre coupera ce cône suivant une ligne sphérique AB. Ayant décrit dans un plan, d'un point quelconque O_1 comme centre, un cercle de même rayon que la sphère, et pris sur la circonférence un point arbitraire A_1, nous ferons correspondre, à un point quelconque M de la ligne sphérique, le point M_1 pris sur la circonférence de manière que l'arc de cercle A_1M_1 ait même longueur que l'arc AM de la ligne sphérique; puis à chaque point P de la surface conique, le point P_1 pris à une distance de O_1 égale à OP, sur un rayon tel que son extrémité M_1 corresponde (au sens que nous venons d'indiquer) au point M où la génératrice OP coupe la sphère.

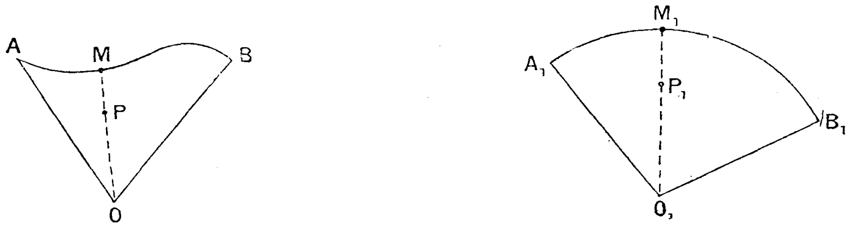

Fig. 648.

La figure formée par les points tels que P_1 est dite le *développement* de la surface conique. Le lecteur démontrera aisément, par des procédés analogues à ceux que nous avons employés pour le cylindre : 1° que *toute ligne tracée sur le cône a même longueur que son développement ;* 2° que *deux lignes se coupent sous le même angle que leurs développements.*

812. On dit qu'une surface est *développable* lorsqu'on peut faire correspondre ses points à ceux d'un plan de manière que les lignes qui se correspondent aient même longueur de part et d'autre. C'est ce qui a lieu pour le cylindre et le cône. Il existe une infinité d'autres surfaces développables. Mais il faut se garder de croire que cette propriété appartienne à une surface quelconque : pour une surface prise au hasard, au contraire, elle n'a, en général, pas lieu. La sphère, par exemple, ne la possède pas.

C'est ce dont on peut se rendre compte en considérant les angles que font entre elles les lignes tracées sur une surface développable.

Soient S une telle surface ; Ax, Ay (*fig.* 649) deux lignes concourantes de S, auxquelles correspondent, lorsque cette surface est développée sur un plan (ce que nous supposons possible), les deux lignes A_1x_1, A_1y_1. Coupons Ax, Ay par une troisième ligne BC, ayant pour développement B_1C_1 et supposons que les points B et C se rapprochent indéfiniment du point A. Dans ces conditions, les directions des lignes AB, AC ; A_1B_1, A_1C_1 tendent respectivement vers celles des tangentes aux lignes Ax, Ay, A_1x_1, A_1y_1 en A et en A_1. De plus, les rapports des arcs AB, AC, A_1B_1, A_1C_1 aux cordes correspondantes tendront vers l'unité.

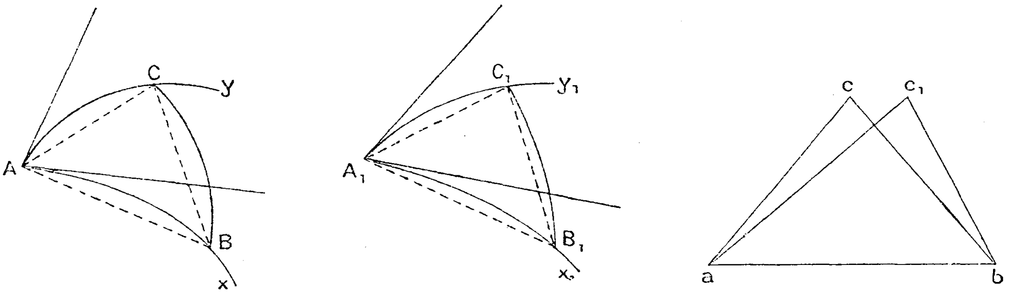

FIG. 649.

Nous ne pouvons pas étendre cette conclusion aux arcs BC, B_1C_1, puisque les courbes BC et B_1C_1 sont variables. Mais on démontre que, moyennant certaines restrictions simples imposées à S, on peut choisir les courbes variables BC, B_1C_1, de manière que les rapports des arcs BC, B_1C_1 à leurs cordes respectives tendent vers 1 et que, de plus, le rapport $\frac{AB}{AC}$ tende vers une limite.

En tout cas, si S était une sphère, un tel choix serait assurément possible : il suffirait de prendre pour BC un arc de grand cercle(1).

Mais d'après l'hypothèse, les arcs AB et A_1B_1 ont même longueur, de même que AC, A_1C_1 et BC, B_1C_1. Donc le rapport de la corde AB à la corde A_1B_1 tend vers l'unité, et il en est de même pour chacune des cordes AC, BC et sa correspondante.

Par conséquent, si, sur une droite fixe ab comme base, nous construisons un triangle abc semblable à ABC et un triangle abc_1 semblable à $A_1B_1C_1$, le point c tendra vers une position limite (puisque les rapports $\frac{AC}{AB}$, $\frac{BC}{AB}$ admettent chacun une valeur limite) et le point c_1 tendra vers la même limite (puisque $\frac{ac}{ac_1} = \frac{AC}{A_1C_1} : \frac{AB}{A_1B_1}$ tend vers l'unité, et de même $\frac{bc}{bc_1}$). Donc l'angle $\widehat{bac} = \widehat{BAC}$ tend vers la même limite que $\widehat{bac_1} = \widehat{B_1A_1C_1}$: autrement dit, l'angle des deux courbes A_1x_1, A_1y_1 est égal à celui des deux courbes Ax, Ay. Ainsi, il devra toujours arriver (comme nous l'avons trouvé dans les cas du cône et du cylindre) que l'angle de deux courbes quelconques de la surface soit égal à celui de leurs développements.

---

(1) On va voir, au numéro suivant que, dans ces conditions, B_1C_1 est un segment de droite.

813. Cela posé, admettons que la surface S soit une sphère. Alors la ligne la plus courte entre deux points de la sphère devra évidemment avoir pour image la ligne la plus courte entre deux points du plan : autrement dit, les grands cercles de la sphère auront pour images les droites du plan et un triangle sphérique aura pour image un triangle rectiligne du plan. Or cela est absurde, puisque la somme des angles d'un triangle rectiligne est égale à deux droits et celle des angles d'un triangle sphérique, toujours plus grande que deux droits.

Donc la sphère, ou même n'importe quelle aire sphérique, si petite qu'elle soit (puisqu'on pourra toujours y trouver un triangle sphérique), ne pourra jamais être développée sur le plan (1).

814. Volumes limités par des surfaces courbes.

Soit une portion d'espace S limitée par des surfaces quelconques.

Supposons que nous puissions trouver deux séries indéfinies de polyèdres $P_1, P_2, \ldots, P_n, \ldots$; $Q_1, Q_2, \ldots, Q_n, \ldots$ satisfaisant aux conditions suivantes :

1° Quel que soit n, le polyèdre (2) P_n n'a aucun point extérieur à S (tout ou partie des sommets, ou des arêtes ou même des faces de P_n pouvant faire partie des surfaces limites de S);

2° Quel que soit n, l'espace S n'a aucun point extérieur à Q_n (une partie quelconque de la surface limite de S pouvant être sur la surface de Q_n);

3° La différence $Q_n - P_n$ des volumes des deux polyèdres de même indice tend vers zéro lorsque cet indice n augmente indéfiniment; ou encore, le rapport $\frac{Q}{P_n}$ tend vers l'unité dans les mêmes conditions.

Alors je dis que *les volumes* P_n, Q_n *tendent vers une limite commune* V.

Pour le démontrer, je remarque tout d'abord que tout polyèdre Q_n de la seconde série renferme tout polyèdre P_n de la première à son intérieur et que, par conséquent, tout volume de la seconde série est plus grand que n'importe quel volume de la première.

Cela posé, supposons d'abord que, parmi les P_n, il y en ait un, P_α, plus grand que tous les suivants (ou au moins égal à l'un quelconque d'entre eux)

Alors, lorsque l'indice n est plus grand que α, les quantités Q_n, P_n comprennent entre elles P_α et comme leur différence tend vers zéro ou leur rapport vers 1, il est évident qu'elles tendent vers la limite commune P_α.

Supposons, en second lieu, que jamais une des quantités P_n ne soit plus grande que toutes les suivantes : autrement dit, si nous prenons une quelconque d'entre elles, soit P_α, il en existera une suivante, P_β, plus grande que P_α; puis une autre, P_γ, suivant P_β, et plus grande que P_β ; et ainsi de suite indéfiniment.

---

(1) De même, deux sphères de rayons différents ne sont pas *applicables* l'une sur l'autre, c'est-à-dire qu'il est impossible de faire correspondre point par point une portion de l'une à une portion de l'autre, de manière que les lignes qui se correspondent aient même longueur.

(2) Nous ne supposerons pas que le polyèdre P_n soit nécessairement d'un seul tenant : on pourra prendre, pour P_n, l'ensemble de plusieurs polyèdres non contigus, pourvu que ces polyèdres et leur volume total vérifient les conditions indiquées dans le texte. Nous aurons plus loin (**815** *bis*) à utiliser cette remarque.

Puisque les quantités P_α, P_β, P_γ, ... vont en croissant et qu'elles restent inférieures à une quantité déterminée (à savoir l'une quelconque des Q_n), elles tendent vers une limite V, et les quantités Q_α, Q_β, Q_γ, ... tendent également vers V (puisque les différences $Q_\alpha - P_\alpha$, $Q_\beta - P_\beta$, $Q_\gamma - P_\gamma$, ... ont pour limite zéro ou que les rapports $\frac{Q_\alpha}{P_\alpha}$, ... ont pour limite 1).

V est au moins égal à chacun des P_n (puisqu'il est la limite de la suite Q_α, Q_β, Q_γ, ..., dont tous les termes sont supérieurs à P_n) et au plus égal à chacun des Q_n (comme limite de la suite P_α, P_β, P_γ, ...). Donc il est la limite commune de P_n et de Q_n, puisqu'il est constamment compris entre ces deux quantités, dont la différence tend vers zéro ou le rapport vers 1.

C. Q. F. D.

De plus, *si on forme, d'une manière quelconque, deux autres séries de polyèdres,* P'_n *et* Q'_n, *satisfaisant* (*comme* P_n et Q_n) *aux conditions précédemment indiquées par rapport à la même portion d'espace* S, *la limite commune de* P'_n *et de* Q'_n *aura la même valeur* V *que la limite commune de* P_n *et de* Q_n.

Cette proposition n'est pas distincte de la précédente : car on peut former une suite comprenant alternativement des polyèdres P_n et des polyèdres P'_n, ainsi qu'une composée des polyèdres Q_n et Q'_n correspondants : ces deux suites mixtes satisferont aux mêmes conditions que les premières et auront, par conséquent, une limite commune, ce qui ne se peut que si P_n et P'_n, Q_n et Q'_n ont la même limite.

La quantité V, limite des volumes P_n et des volumes Q_n, et indépendante (d'après ce que nous venons de constater) de la manière dont on choisit ceux-ci (pourvu que les uns comprennent S, que les autres soient compris dans S et que la différence des uns et des autres tende vers zéro ou leur rapport vers 1) est dite le *volume* de l'espace S. Il est clair que la notion de volume, une fois définie, possédera les propriétés indiquées au livre VI (**397**) ; c'est-à-dire que deux espaces égaux ont le même volume et que l'espace formé par la réunion de deux espaces contigus aura un volume égal à la somme de leurs volumes (¹).

815. Il reste à savoir si, une portion d'espace S étant donnée, on peut trouver les polyèdres P_n et Q_n satisfaisant aux conditions indiquées.

Supposons que, à l'intérieur de S, on puisse trouver un point O tel que la droite qui le joint à un point quelconque de la surface limite de S (cette droite n'étant pas prolongée au delà de O) ne perce cette surface limite en aucun point. Soit S' l'homothétique de S par rapport au point O, avec un rapport d'homothétie $1 - \varepsilon$ plus petit que l'unité; S'', l'homothétique de S par rapport au centre O, avec le rapport de similitude $1 + \varepsilon$: il est clair

---

(1) Sur ce dernier point, voir plus loin, **815** *bis*.

que chacun des solides S′, S, S″ est entièrement intérieur au suivant. Soit alors R un polyèdre intérieur à S″, mais comprenant S′ à son intérieur. L'homothétique P de R par rapport à O avec le rapport d'homothétie $\frac{1}{1+\varepsilon}$ sera intérieur à S (puisque R est intérieur à S) ; et l'homothétique Q de R par rapport à O, avec le rapport d'homothétie $\frac{1}{1-\varepsilon}$ comprendra S à son intérieur (puisque R comprend S″). D'ailleurs le rapport des volumes de P et de Q est (**431**) égal à $\left(\frac{1-\varepsilon}{1+\varepsilon}\right)^3$ et tend vers 1 lorsque ε tend vers zéro.

Dès lors, si on donne à ε une série de valeurs qui tendent vers zéro, et si, pour chacune de ces valeurs, on construit le polyèdre R et, par conséquent, les polyèdres P et Q, on a une double série de polyèdres remplissant toutes les conditions demandées.

815 *bis*. Si la forme de la région S est telle que le raisonnement précédent ne s'y applique pas, on pourra, en général, la décomposer en deux ou plusieurs régions partielles auxquelles il soit applicable. On formera alors chaque polyèdre P avec l'ensemble des polyèdres P relatifs à ces régions partielles (¹) et le polyèdre Q avec l'ensemble des polyèdres Q relatifs à ces mêmes régions.

816. Soit *d* la distance minima de la surface limite de S à celle de S′ et à celle de S″, c'est-à-dire une quantité telle que tout point de la première soit au moins à la distance *d* de tout point de la seconde et au moins à la distance *d* de tout point de la troisième.

Inscrivons à la surface limite de S un polyèdre comprenant le point O à son intérieur et tel que la plus grande dimension de chaque face soit inférieure à *d*. Alors la surface de ce polyèdre ne pourra évidemment avoir aucun point commun avec la surface de S″, ni avec la surface de S′, ni même avec l'intérieur de S′. Donc ce polyèdre est compris dans S″ et comprend S′ : on peut le prendre pour le polyèdre R du numéro précédent, et si l'on forme de pareils polyèdres pour des valeurs de plus en plus petites de ε, leurs volumes tendent vers V.

Comme ε tend vers zéro avec *d*, on voit que, *si l'on inscrit à la surface limite de* S *un polyèdre* (*comprenant toujours à son intérieur un point déterminé intérieur à* S) *et tel que la plus grande dimension de chacune de ses faces tende vers zéro, le volume de ce polyèdre tendra vers le volume de* S. On voit d'ailleurs immédiatement que cette conclusion s'étend au cas où l'on ne peut appliquer directement le raisonnement du n° **815** et où on est obligé de décomposer S en deux ou plusieurs parties comme il est expliqué au n° **815** *bis*.

Lorsqu'il s'agit de la sphère, on peut prendre pour le point O le centre et les solides S′ S″ seront des sphères concentriques à la première. La distance *d* sera la différence des rayons de S et de S′ et on pourra prendre,

---

(1) Voir la note 2, page 535.

pour R, tout polyèdre inscrit dont les dimensions sont toutes plus petites que cette différence (1).

817. Dans le cas du cône ou du cylindre, la définition précédente concorde avec celles que nous avons données au livre VIII. Prenant, par exemple, le cas du cylindre, il est clair que les prismes inscrits et circonscrits sont précisément des polyèdres P_n et Q_n tels que nous les avons introduits au n° **815**.

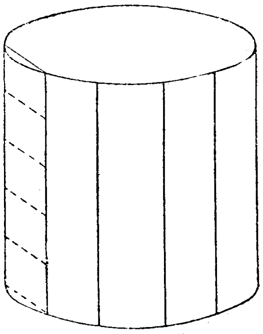

Fig. 651.

D'ailleurs, le prisme inscrit tel que les côtés de sa base tendent vers zéro, peut être considéré comme un polyèdre inscrit dans lequel la plus grande dimension de chaque face tend vers zéro. Il suffira, à cet effet, de diviser chaque arête latérale du prisme en un certain nombre de parties égales, lequel nombre ira en augmentant indéfiniment en même temps que le nombre des côtés de la base (*fig.* 651). Les points de division feront partie, comme les sommets extrêmes eux-mêmes, de la surface cylindrique, et, par conséquent, les petits rectangles qu'ils forment (*fig.* 651) pourront être considérés comme faces d'un polyèdre inscrit, — plusieurs de ces faces étant seulement en prolongement les unes des autres, ce qui ne change rien à la validité des raisonnements.

Des considérations toutes semblables s'appliquent évidemment au cône.

818. Considérons, de même, le volume V engendré par un polygone plan tournant autour d'un axe situé dans son plan et qui ne le traverse pas.

Considérons p positions du polygone, ABCD, A′B′C′D′ (*fig.* 652) étant,

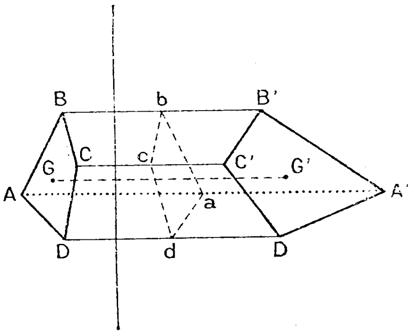

Fig. 652.

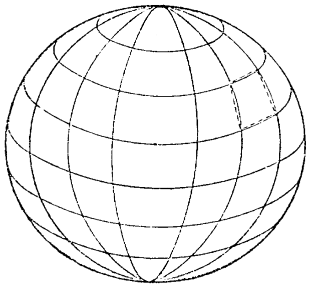

Fig. 650.

---

(1) Par exemple, on pourra diviser la sphère par des méridiens et des parallèles (*fig.* 650) dont les longitudes ou les latitudes diffèrent entre elles d'une quantité λ telle que l'arc de grand cercle d'angle au centre λ ait une longueur inférieure à d. Les points d'intersection de ces méridiens et de ces parallèles seront les sommets d'un polyèdre inscrit dont toutes les faces, qui seront des trapèzes ou des triangles, auront leurs dimensions plus petites que d.

par exemple, deux consécutives d'entre elles. Les polygones ABCD, A'B'C'D' sont symétriques l'un de l'autre par rapport à un plan, le plan bissecteur du dièdre dans les faces duquel ils sont respectivement contenus. Il existe donc un tronc de prisme qui a pour base ces deux polygones, les arêtes étant perpendiculaires à ce plan de symétrie. Considérons le polyèdre formé par l'ensemble des troncs de prismes qui ont ainsi pour bases les couples de positions consécutives du polygone. Si nous augmentons indéfiniment le nombre de ces positions de manière que les dièdres successifs que forment entre elles leurs plans tendent vers zéro, le volume du polyèdre dont nous venons de parler tendra vers V, ainsi qu'il résultera de considérations toutes pareilles à celles que nous venons de présenter au n° précédent.

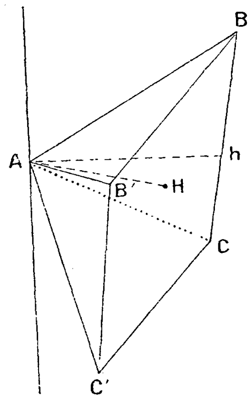

Fig. 653.

Cette remarque permet de démontrer très simplement certaines propositions importantes. Supposons, par exemple, que le polygone considéré soit un triangle ABC ayant son sommet A sur l'axe, alors les troncs de prismes se réduiront à des pyramides telles que ABCB'C' (*fig.* 653).

Chacune de ces pyramides a pour mesure le tiers du produit de sa base BCB'C' par sa hauteur AH, de sorte que leur somme est égale au tiers de la somme des bases multipliée par une quantité intermédiaire entre la plus petite et la plus grande des hauteurs. Or, lorsqu'on augmente indéfiniment le nombre des positions successives du triangle, la somme des trapèzes BCB'C' tend vers la surface engendrée par BC dans sa révolution autour de l'axe, et quant à la hauteur AH, elle tend vers la hauteur A*h* du triangle ABC, car la distance *h*H, qui est la distance du point H à la droite BC, tend vers zéro.

Nous avons donc immédiatement le théorème du n° **485** : *le volume engendré par un triangle tournant autour d'un axe situé dans son plan, passant par son centre et ne le traversant pas, est égal à la surface engendrée par le côté opposé au sommet situé sur l'axe, multipliée par le tiers de la hauteur correspondante.*

819. Revenons au cas où le polygone tournant est quelconque : la remarque précédente nous donne encore (du moins pour le cas d'un polygone) le *théorème de Guldin.*

Théorème. — *Le volume engendré par une aire plane qui tourne autour d'un axe situé dans son plan et ne le traversant pas, est égal au produit de cette aire par la circonférence que décrit son centre de gravité.*

Soient, en effet, G le centre de gravité du polygone ABCD, G' le centre de gravité de la position suivante A'B'C'D' (*fig.* 652).

Le tronc de prisme ABCDA'B'C'D' a pour mesure (**608**) le produit de sa section droite par la distance GG'.

Or la section droite, qui n'est autre que la section par le plan bissecteur du dièdre formé par les bases, tend vers le polygone ABCD lui-même et, d'autre part, la somme de toutes les longueurs telles que GG' tend vers la longueur de la circonférence décrite par le point G. Donc le théorème est démontré.

820. Nous avons défini le volume V d'un solide quelconque par cette propriété que deux volumes polyédraux P_n et Q_n, dont le premier est intérieur au solide et dont le second comprend ce solide, tendent vers V lorsque leur différence tend vers zéro. Mais il est clair, une fois V défini, que deux volumes quelconques, polyédraux ou non, tendront vers V s'ils satisfont aux autres conditions imposées à P_n et à Q_n, puisque l'un sera toujours plus petit que V et l'autre plus grand que V.

Soit, par exemple, une aire plane curviligne A tournant autour d'un axe qui est situé dans son plan et ne la traverse pas. Inscrivons et circonscrivons à cette aire des polygones qui auront A pour limite commune. Ces polygones, en tournant autour de l'axe, engendreront des volumes dont les uns seront inscrits au volume V engendré par A, les autres circonscrits au même volume, et la différence entre les volumes inscrits et les volumes circonscrits tendra vers zéro (1). Donc ces volumes auront V pour limite.

On voit par là que la définition des volumes du secteur sphérique et de la sphère, telle que nous l'avons donnée au livre VIII (chap. v), concorde avec la définition actuelle.

On voit aussi que le théorème de Guldin, démontré au numéro précédent pour les polygones, est également vrai pour les aires planes curvilignes, si l'on admet que le centre de gravité de l'aire plane considérée est la position limite vers laquelle tend le centre de gravité d'un polygone inscrit dont tous les côtés tendent vers zéro. Alors, en effet, dans l'égalité exprimée par ce théorème, les deux membres sont les limites de quantités analogues dans lesquelles l'aire curviligne est remplacée par un polygone inscrit.

821. Soit encore un volume quelconque que nous couperons par des plans parallèles à une même direction, la distance de deux plans successifs ne dépassant pas une certaine longueur h. Soient s la portion du solide comprise entre deux consécutifs de ces plans; C un cylindre ayant ses bases dans les plans en question et compris dans s; C' un cylindre ayant ses bases dans les mêmes plans et comprenant s à son intérieur (si les deux courbes de section du solide par les deux plans considérés sont telles que la projection de l'une d'elles sur le plan de l'autre comprenne celle-ci, on pourra prendre pour C et C′ les cylindres droits qui ont respectivement ces courbes pour bases).

---

(1) En effet, cette différence a pour mesure le produit de l'aire annulaire comprise entre un polygone inscrit et un polygone circonscrit, — laquelle tend évidemment vers zéro, — par 2π et par la distance de l'axe au centre de gravité de cette aire, laquelle n'augmente pas indéfiniment (ex. 870).

Dans tous les solides que l'on a habituellement à considérer, on peut prendre h assez petit pour que, quels que soient les deux plans, pourvu que leur distance ne dépasse pas h, le rapport $\frac{C}{C'}$ soit compris entre l'unité et un nombre $1 - \alpha$ aussi voisin qu'on le veut de l'unité (ou, s'il n'en est pas ainsi, le solide peut être décomposé en parties dont chacune satisfasse à cette condition).

Dès lors, le solide étant décomposé en plusieurs portions par une série de plans parallèles entre eux dont les distances successives ne dépassent pas h, formons, pour chacune de ces portions les cylindres C et C′ et appelons P le solide formé par l'ensemble des cylindres C, Q le solide formé par l'ensemble des cylindres C′. D'après ce que nous venons de dire, le rapport de chaque cylindre C au cylindre C′ correspondant et, par conséquent aussi, le rapport $\frac{P}{Q}$, seront compris entre 1 et $1 - \alpha$. Le nombre α pouvant être choisi aussi petit qu'on veut si h est suffisamment petit, on voit que lorsque h tend vers zéro, les volumes P et Q tendent simultanément vers le volume du solide, puisque le rapport $\frac{P}{Q}$ tend vers 1 et que P est intérieur à ce solide, lequel est intérieur à Q.

822. Aire d'une surface courbe.

La définition de l'aire d'une portion de surface courbe offre de beaucoup plus grandes difficultés que les précédentes. Si, en effet, on considère une surface polyédrale inscrite à la surface courbe et qu'on suppose seulement, comme nous l'avons fait, que la plus grande dimension de chaque face diminue indéfiniment, sans ajouter à cette supposition d'autres hypothèses convenablement choisies sur la loi de variation de la surface polyédrale, il n'arrive pas nécessairement que l'étendue de celle-ci tende vers une limite déterminée : il peut même arriver que cette étendue augmente indéfiniment.

Nous nous bornerons à définir l'aire d'une surface fermée convexe.

Dans ce but, nous établirons d'abord les deux lemmes suivants :

Lemmes. — I. *Une face d'un polyèdre quelconque est plus petite que la somme des autres.*

Projetons, en effet, toutes ces autres faces sur le plan de la première. L'ensemble de ces projections recouvrira évidemment toute cette première face (il pourra même arriver qu'il la dépasse, ou en recouvre tout ou partie plusieurs fois). Or, chaque face est (**606**) plus grande que sa projection.

II. *La surface d'un polyèdre convexe est plus petite que celle d'un polyèdre enveloppant quelconque.*

Soient S la surface du polyèdre convexe, S′ la surface du polyèdre enveloppant. Celle-ci pourra avoir une ou plusieurs faces communes (1) avec la première (pourvu qu'elle n'ait aucune partie intérieure à la première) : soit n le nombre des faces non communes sur S. Si $n = 1$, le théorème est démontré, car l'unique face non commune de S est plus petite que la somme des faces non communes de S′, lesquelles forment avec elle un polyèdre.

Si n est différent de 1, prolongeons le plan P d'une des faces non communes de S : nous partageons ainsi S′ en deux parties, l'une S'_1, située par rapport à P du même côté que S, l'autre S'_2 située du côté opposé. Nous diminuerons cette dernière (lemme précédent) en la remplaçant par la portion du plan P terminée au même contour. Or, nous obtenons ainsi un nouveau polyèdre enveloppant S, mais avec lequel ce dernier aura une face commune de plus.

Il est clair qu'en continuant cette réduction, nous serons finalement ramenés au cas de $n = 1$. La proposition est donc démontrée.

Remarque. — Une proposition toute semblable s'appliquerait (avec le même mode de démonstration) à une surface polyédrale convexe ouverte et à une surface polyédrale enveloppante terminée au même contour.

823. Cela posé, soit une surface fermée convexe S. Nous considérerons encore des polyèdres P_n et Q_n dont les uns seront intérieurs à S et les autres la comprendront à leur intérieur. Mais nous imposerons cette fois à ces polyèdres *la condition d'être convexes*. Dès lors, en vertu du lemme II, la surface de chaque polyèdre P_n sera inférieure à la surface de n'importe quel polyèdre Q_n. Si donc on suppose que la différence entre la surface de P_n et celle de Q_n tend vers zéro (ou que leur rapport tend vers 1), ces surfaces auront une limite commune, indépendante de la manière dont on aura choisi les P_n et les Q_n pourvu que ceux-ci remplissent les diverses conditions que nous venons d'énumérer. Il n'y a, à cet égard, rien à changer aux raisonnements du n° **814**.

On obtient d'ailleurs des polyèdres P_n et Q_n par le même procédé qui a été expliqué au n° **815**, en ayant soin, toutefois, de prendre pour R un polyèdre convexe. En particulier, on verra, exactement comme au n° **816**, *qu'une surface fermée convexe est la limite d'un polyèdre convexe inscrit, dans lequel la plus grande dimension de chaque face tend vers zéro.*

824. Lorsqu'il s'agit de la longueur d'un arc de courbe plane, on peut, en général, si l'arc n'est pas convexe, le décomposer en arcs convexes. Il importe de remarquer que ce mode de raisonnement n'est pas applicable ici. On rencontre, et cela parmi les surfaces usuelles, des surfaces dont aucune portion, si petite qu'elle soit, n'est convexe et dont, en un quelconque de leurs points, on ne peut pas dire qu'elles tournent leur convexité dans un sens ou dans l'autre. C'est, par exemple, le cas de la *surface gauche de révolution* (ex. 1031) dont la forme générale est représentée fig. 654 (¹).

---

(1) Une face de S sera considérée comme commune avec S′ si elle est située dans le plan d'une face de S′.

825. Dans le cas du cylindre, du cône ou du tronc de cône, la définition précédente coïncide avec celles que nous avons données au livre VIII : on peut, en effet, évaluer la surface *totale* de l'un de ces corps comme limite de la surface totale des prismes, pyramides ou troncs de pyramides inscrits (comparer **817**) et comme les surfaces de bases de ces polyèdres tendent vers les surfaces de bases des corps ronds correspondants, il en est de même des surfaces latérales.

On peut également raisonner sur la surface engendrée par une ligne brisée tournant autour d'un axe extérieur à elle et situé dans son plan, comme nous avons raisonné au nº **818** sur le volume engendré par un polygone : il est clair qu'en appliquant, non plus le théorème du nº **608**, mais l'exercice 871, nous obtiendrons (du moins pour une ligne brisée tournante) le second *théorème de Guldin :*

Théorème. — *La surface engendrée par une ligne qui tourne autour d'un axe situé dans son plan et ne le traversant pas, est égale au produit de la longueur de cette ligne par la circonférence que décrit son centre de gravité.*

Ce théorème s'étendra d'ailleurs à une ligne courbe tournante, par voie de passage à la limite : les considérations qui précèdent permettent de faire cette extension par une voie analogue à celle que nous avons employée au nº **820**, toutes les fois qu'il s'agit d'une ligne convexe tournant sa concavité vers l'axe.

825 *bis*. Nous pouvons encore définir l'aire de la zone, puisque nous pouvons définir l'aire totale du segment sphérique ; et cette définition concorde (comparer **820**) avec celle du nº **482**.

On remarquera que ces définitions permettent d'obtenir immédiatement les théorèmes relatifs aux volumes sphériques.

Si, en effet, nous considérons un polyèdre convexe inscrit à la sphère, ce polyèdre sera décomposable en pyramides ayant pour

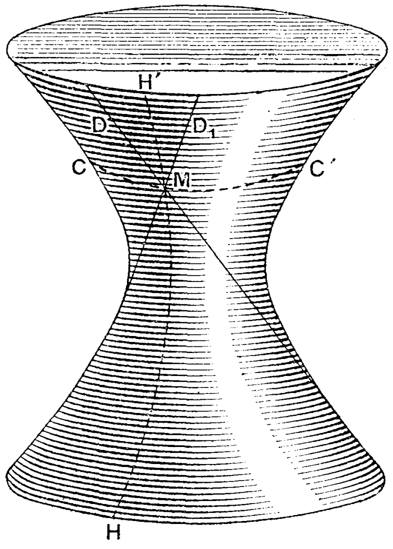

Fig. 654.

(¹) M étant un point quelconque de la surface, le plan tangent en M coupe la surface suivant deux droites (ex. 1273) MD, MD_1, lesquelles partagent celui-ci en quatre parties dont deux opposées l'une à l'autre (celles qui contiennent le méridien HH' (*fig.* 654)) sont, par rapport au plan tangent, du côté extérieur, et les deux autres (celles qui contiennent le parallèle CC') du côté intérieur.

bases les différentes faces et pour sommet commun le centre. Son volume sera égal au tiers de la somme des bases, — c'est-à-dire de la surface, — multipliée par une quantité intermédiaire entre la plus grande et la plus petite des hauteurs. Or ces dernières tendent toutes vers le rayon, puisque le polyèdre et, par conséquent (puisqu'il est convexe), le plan de chacune de ses faces sont extérieurs (**816**) à une sphère quelconque concentrique à la première et de rayon plus petit. Quant à la surface du polyèdre, elle tend vers celle de la sphère. On a donc bien ce théorème que *le volume de la sphère est égal au produit de sa surface par le tiers du rayon.* Le même raisonnement s'applique d'ailleurs évidemment au secteur sphérique.

826. Lorsqu'une surface S peut être développée sur un plan, on démontre que l'aire de toute figure tracée sur S est égale à celle de son développement (voir exerc. 822).

Bien entendu, la réciproque n'est pas vraie : nous voulons dire que, si l'on a fait correspondre les points d'une surface S aux points d'un plan de manière que les figures qui se correspondent aient même aire, il n'en résulte en aucune façon que S soit développable. Par exemple, dans l'exercice 1301, nous indiquons une correspondance de cette espèce entre une sphère (c'est-à-dire une surface non développable) et un cylindre (lequel peut être étalé sur un plan).

NOTE H

SUR LES POLYÈDRES RÉGULIERS ET LES GROUPES DE ROTATIONS

827. Nous avons démontré, au nº **708**, que *tout polyèdre régulier admet un certain nombre de déplacements,* c'est-à-dire que ces déplacements font coïncider le polyèdre avec lui-même, les sommets venant prendre la place les uns des autres.

Soient R et S deux déplacements qui laissent ainsi inaltéré le polyèdre considéré, ou, plus généralement, qui laissent inaltérée une figure donnée F quelconque. Il est clair que la figure F ne changera pas davantage, si l'on effectue successivement les deux déplacements R et S, autrement dit, si l'on effectue le déplacement RS, *produit* (Pl., **291**) de R et de S.

Par conséquent, en introduisant une définition donnée au nº **291**, si nous considérons l'ensemble des déplacements qui n'altèrent pas la figure F, *ces déplacements forment un groupe*, puisque R et S étant deux quelconques d'entre eux, les produits (¹) RS et SR appartiennent au même ensemble.

Il est bien entendu que S peut n'être autre que R : le produit RR est désigné par la notation R^2 et de même les produits de trois, quatre.... facteurs tous égaux à R sont respectivement désignés par les notations R^3, R^4..., et sont dits les *puissances* successives de R.

Si R est un déplacement hélicoïdal composé d'une rotation d'angle α et d'une translation t parallèle à l'axe de la rotation, on voit immédiatement que R^m (en désignant par m un entier quelconque) sera un déplacement de même axe, composé d'une rotation d'angle $m\alpha$ et d'une translation de grandeur mt.

Il faut observer, d'autre part, que nous comptons parmi les opérations appartenant au groupe celle dans laquelle on ne change rien aux positions des différents points : elle est dite *l'opération identique* et il nous arrivera de la représenter par le chiffre 1. Il est aisé de voir que la conclusion obtenue tout à l'heure (à savoir que si R et S laissent inaltérée la figure F, il en est de même de leur produit) suppose essentiellement la convention qui vient d'être faite, relativement à l'opération identique.

En effet, R étant un déplacement quelconque, qui a pour effet d'amener un point quelconque M de l'espace en une nouvelle position M', soit S le déplacement *inverse* de R, c'est-à-dire celui qui, appliqué au point M', donne pour résultat M (R étant un déplacement hélicoïdal composé d'une translation de grandeur t et d'une rotation d'angle α, le déplacement inverse a le même axe que R, sa translation et sa rotation ayant les mêmes grandeurs t et α, mais étant de sens opposés aux premiers). Si R laisse inaltérée la figure F, il en est de même de S et, par conséquent, si nous voulons pouvoir dire, comme nous l'avons fait tout à l'heure, que les déplacements qu'admet F forment un groupe, il faut compter au nombre de ces déplacements les produits RS et SR. Or chacun de ces deux produits donne, en vertu de la définition de S, l'opération identique (chacune des deux opérations R et S ayant pour effet de ramener à leurs positions primitives les figures transformées par l'autre).

L'opération inverse de R est représentée par le symbole R^{-1}; ses puissances successives, par R^{-2}, R^{-3}.... On voit immédiatement que R^{-k} est inverse de R^k et que l'inverse de RS (en désignant par S une seconde opération quelconque) est $S^{-1}R^{-1}$.

828. Nous venons de remarquer que, *si le groupe des déplacements qu'admet une figure quelconque* F *comprend un déplacement* R, *il comprend aussi le déplacement inverse* R^{-1}.

La même propriété appartient à la plupart des groupes que l'on a à considérer dans les applications de cette notion aux diverses parties des mathématiques. On démontre, en particulier, qu'elle a lieu pour tout

---

(1) Rappelons que les deux produits RS et SR ne sont pas, en général, identiques entre eux (voir Pl., page 276, note 2 et exercice 624).

groupe composé d'un nombre fini de transformations ; nous la démontrerons d'ailleurs, chemin faisant (**830** *bis*), pour les groupes dont nous allons avoir à nous occuper.

Elle entraîne, pour les groupes G qui la possèdent, cette conséquence que *si une figure* F′ *est homologue* (¹) *d'une figure* F (relativement à un groupe G), *réciproquement* F *est homologue de* F′ ; car si R est l'opération du groupe qui transforme F en F′, l'inverse R^{-1} transformera F′ en F.

829. Voici une nouvelle conséquence de la propriété précédente :

Soient R un déplacement (²), composé d'une translation de grandeur t, parallèle à un axe A et d'une rotation d'angle α autour du même axe (t ou α pouvant, bien entendu, être nul), et qui change une figure quelconque f en f' (*fig.* 655); S, un autre déplacement quelconque d'axe B. Appliquons à la figure formée par l'axe A, la figure f et la figure f', le déplacement S : f viendra en f_1, f' en f'_1, l'axe A en A_1. Nous voyons dès lors que le déplacement R_1 qui a pour axe A_1, la translation t et la rotation α étant les mêmes que celles de R, change la figure f_1 (transformée de f' par S) en f'_1 (c'est-à-dire en la transformée de f' par S).

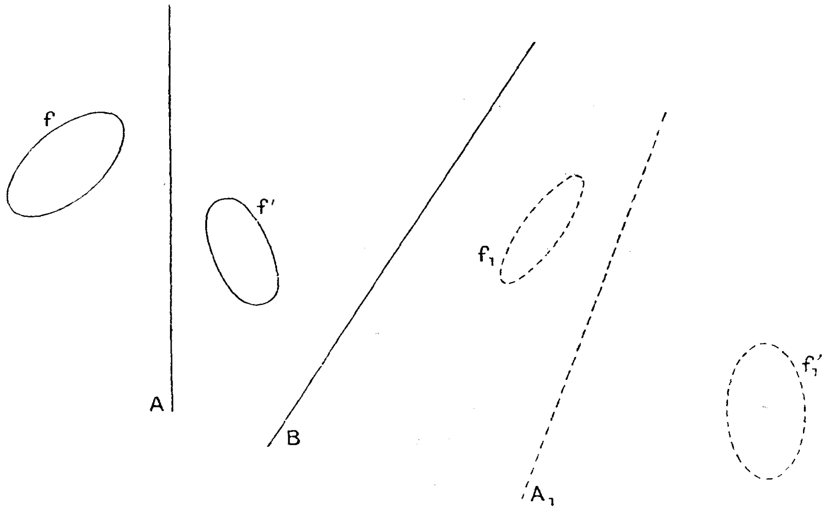

Fig. 655.

On donne à ce nouveau déplacement R_1 le nom de *transformé de* R *par* S. Deux déplacements R, R_1, dont l'un est le transformé de l'autre par un troisième déplacement quelconque, sont dits *semblables*. On voit que deux déplacements semblables ont (en grandeur) même translation et même rotation, et qu'ils sont tous deux *dextrorsum* ou tous deux *sinistrorsum* (**414**, Rem.); ces conditions sont, d'ailleurs, suffisantes, car si elles sont remplies par deux déplacements R, R_1, on les transformera l'un dans l'autre par tout déplacement qui transformera les axes l'un dans l'autre

---

(1) Pour le sens de ce mot, nous renvoyons à la géométrie plane (**293**).

(2) Nous raisonnons ici sur des déplacements, mais la notion d'opérations *semblables* et son application aux groupes subsistent pour des opérations quelconques.

(le sens de translation donné sur le premier axe venant suivant le sens de translation donné sur le second).

Cela posé, admettons que les déplacements R et S, considérés il y a un instant, fassent partie d'un même groupe G, qui comprenne également (d'après la propriété énoncée au n° précédent) R^{-1} et S^{-1}. Alors f et f_1 sont homologues entre eux par rapport à G, de même que f_1 et f'_1; et, comme f et f' sont aussi homologues l'une de l'autre, il en est de même Pl., **293**) pour f_1 et f'_1 ; donc le déplacement R_1, qui change f_1 en f'_1, appartient à G.

Ainsi, *si deux déplacements* R *et* S *appartiennent à un groupe* G, *il en est de même du transformé* R_1 *de* R *par* S.

Les deux déplacements R et R_1 sont dits *homologues* l'un de l'autre par rapport au groupe. Bien entendu, pour que deux déplacements soient homologues, il ne suffit pas qu'ils soient semblables, il faut encore que, parmi les opérations qui les transforment l'un dans l'autre, il y en ait au moins une qui fasse partie du groupe.

---

830. Dans le cas où la figure F est un polyèdre régulier, le nombre des déplacements qu'elle admet est fini.

Nous allons nous proposer de rechercher, d'une manière générale, tous les groupes composés de déplacements en nombre fini. Le nombre N de ces déplacements (le déplacement identique compris, bien entendu), sera dit l'*ordre* du groupe.

Après avoir formé les groupes en question, nous chercherons s'il leur correspond des polyèdres réguliers. Nous nous appuierons, à cet effet, sur le théorème du n° **710.** Par contre, *nos raisonnements seront entièrement indépendants des considérations présentées aux n*^{os} **712-717**; ils nous fourniront, dès lors, une nouvelle démonstration des résultats établis, relativement à l'existence et à la nature des polyèdres réguliers.

Mais l'utilité de la recherche que nous allons entreprendre n'est nullement limitée à l'étude des polyèdres réguliers. D'une part, en effet, on constate que les groupes cherchés jouent un rôle important dans le problème de la résolution des équations (voir note 2 de la page 522); d'autre part, certaines propriétés que nous démontrerons (telles que celle du *domaine fondamental*, n° **845**) se transportent d'elles-mêmes à d'autres groupes analogues, composés d'une infinité d'opérations et dont l'étude est nécessaire à la solution de certaines questions importantes de mathématiques.

Soit G l'un des groupes cherchés. S'il contient un déplacement R, il contiendra aussi ses puissances successives R^2, R^3... Il faudra donc, tout d'abord, que ces puissances ne soient pas en nombre infini.

Dès lors, R *ne peut pas être une translation :* car, si t était la grandeur de cette translation, ses puissances successives seraient des translations de grandeurs respectives $2t$, $3t$,... : de telles translations sont nécessairement toutes distinctes entre elles et en nombre infini.

De même R *ne peut être un déplacement hélicoïdal :* car, si t est la translation qui entre dans R, les déplacements R^2, R^3... comprendraient les translations $2t$, $3t$... et seraient, par conséquent, aussi tous différents les uns des autres.

Donc *le groupe* G *est exclusivement composé de rotations.*

830 *bis.* De plus, soit α l'angle d'une rotation R appartenant à G : les puissances successives de R auront pour angles respectifs 2α, 3α... Pour que ces puissances ne soient pas en nombre infini, il faut qu'une certaine puissance R^p ne soit pas distincte d'une puissance précédente $R^{p'}$. Il est clair que ceci aura lieu lorsque les angles de R^p et de $R^{p'}$ différeront d'un nombre entier de circonférences, et dans ce cas seulement. Donc, *si nous supposons que les angles soient mesurés en prenant pour unité la circonférence entière,* on aura :

$$p\alpha = p'\alpha + m,$$

m étant un entier. Cette équation donne :

$$\alpha = \frac{m}{n}, \tag{27}$$

en posant $p - p' = n$.

On voit alors que la puissance n^{me} de R donnera une rotation composée d'un nombre entier m de circonférences, c'est-à-dire se réduira au déplacement identique : la puissance R^{n+1} ne sera, dès lors, pas distincte de R, la puissance R^{n+2} de R^2....

Nous pouvons d'ailleurs toujours supposer que n ait été choisi le plus petit possible (c'est-à-dire que R^n soit la première puissance de R qui se réduise au déplacement identique). Cela revient évidemment à supposer que la fraction $\frac{m}{n}$ est réduite à sa plus simple expression. Lorsqu'il en est ainsi, on dit que n est l'*ordre* de la rotation R. Par exemple, une rotation d'ordre 2 n'est autre que ce que nous avons appelé précédemment *transposition.*

L'angle $n\alpha$ étant un multiple de la circonférence, la rotation R^{n-p}, laquelle est d'angle $(n - p)\alpha$, revient à une rotation d'angle $p\,\alpha$ effectuée en sens contraire, c'est-à-dire à R^{-p}. En particulier, l'*inverse* $R^{-1} = R^{n-1}$ *de* R *figure parmi les puissances de* R *et, par conséquent, dans le groupe* G.

Enfin, on peut remarquer que si n est l'ordre de R, la rotation ayant pour axe celui de R et pour angle $\frac{1}{n}$ figure également parmi les puissances de R. Car les nombres m et n qui figurent dans la relation (27) étant premiers entre eux, on peut (¹) trouver deux entiers h et k tels que l'on ait

$$km - hn = 1$$

---

(1) TANNERY, *Leçons d'Arithmétique*, page 466.

Cette relation donne

$$k\alpha = \frac{km}{n} = h + \frac{1}{n},$$

et montre, par conséquent, que la rotation R^k a son angle égal au $n^{\text{ème}}$ de la circonférence, à un nombre entier h de circférenee près, lequel peut être laissé de côté.

831. Considérons maintenant deux rotations R et S du groupe.

Leurs axes sont dans un même plan; l'hypothèse contraire serait, en effet, incompatible avec le résultat que nous venons de trouver au n° **829**, en vertu de la proposition que nous énonçons à l'exercice 592.

De plus, *ces axes (s'ils sont distincts) ne peuvent être parallèles*, car, dans ce cas, ou bien R et S auraient leurs angles égaux et de sens contraires, et le déplacement RS serait une translation non nulle (comparer Pl., ex. 94); ou bien les angles de R et de S seraient inégaux (ou égaux et de même sens) et les deux opérations RS et SR seraient deux rotations égales, mais (ex. 623-624) autour d'axes forcément distincts entre eux, de sorte que l'opération $RS(SR)^{-1}$ (ou $RSR^{-1}S^{-1}$) serait une translation non nulle.

Les deux axes (s'ils sont distincts) se coupent donc forcément en un certain point O.

832. Le groupe G renferme d'ailleurs nécessairement une rotation T dont l'axe passe par le point O et est situé en dehors du plan des deux premiers. Si, en effet, R et S sont toutes deux des transpositions, cette rotation T est donnée (**416**) par le produit RS; si, au contraire R, par exemple, n'est pas une transposition, la transformée de S par R a évidemment son axe en dehors du plan qui contient ceux de R et de S.

Dès lors, l'axe de toute rotation U du groupe doit passer par le point O, puisqu'il doit (n° préc.) rencontrer les axes de R, de S et de T.

833. Nous arrivons donc à cette première conclusion :

Tous les déplacements du groupe sont des rotations dont les axes passent par un même point.

Nous prendrons le point O où se coupent tous les axes comme centre d'une sphère Σ : il est clair qu'il suffira, pour étudier les rotations du groupe G, de les appliquer aux seuls points de la sphère Σ.

Chaque axe de rotation coupera Σ en deux points, les seuls que la rotation correspondante laisse inaltérés sur la sphère ; pour abréger, nous donnerons aux points ainsi obtenus le nom de *pôles de rotation.*

Chaque pôle de rotation est, en général, commun à plusieurs rotations du groupe. La plus petite (¹) R d'entre elles a pour angle une partie aliquote de la circonférence : car si cet angle était mesuré par $\frac{m}{n}$ (la fraction $\frac{m}{n}$

---

(1) C'est-à-dire celle dont l'angle est le plus petit, le déplacement identique étant, bien entendu, excepté.

étant irréductible), nous savons que le groupe renfermerait la rotation de même axe et d'angle $\frac{1}{n}$, qui serait plus petite que la première, si m était plus grand que 1 : donc $m = 1$.

D'ailleurs, toutes les autres rotations du groupe ayant les mêmes pôles que R seront des puissances de R : car si l'angle α de l'une d'elles était mesuré par un nombre compris entre $\frac{p}{n}$ et $\frac{p+1}{n}$, en la multipliant par R^{-p}, on obtiendrait évidemment une rotation de même axe et dont l'angle serait plus petit que $\frac{1}{n}$, par conséquent plus petit que celui de R (contrairement à l'hypothèse). Nous dirons que n est l'*ordre du pôle de rotation* considéré.

Un pôle d'ordre n est évidemment commun à $n - 1$ rotations, le déplacement identique non compris.

834. N étant l'ordre du groupe, un point quelconque P de la sphère a N homologues, qu'on obtient en lui faisant subir successivement les N déplacements du groupe (déplacement identique compris). Si P n'est pas un pôle de rotation, ces homologues sont tous distincts entre eux; car si deux ou plusieurs d'entre eux coïncidaient en P', le point P' serait pôle d'une rotation R', et le point P, pôle d'une rotation R homologue de R' (**829**).

Si, au contraire, P est un pôle de rotation d'ordre n, commun à la rotation R $\left(\text{d'angle } \frac{1}{n}\right)$ et à ses puissances $R^2 \ldots$, R^{n-1}, il coïncidera avec $n - 1$ de ses homologues. De plus, soit P' un autre homologue de P, déduit de P par un déplacement S du groupe et ne coïncidant pas avec P : il y aura n homologues de P qui coïncideront avec P'; à savoir, ceux qu'on déduit de P par les rotations S, RS, $R^2S \ldots$, $R^{n-1}S$. Il n'y en aura d'ailleurs pas d'autres : car, sans cela, le point P' serait pôle de rotation d'un ordre n' supérieur à n et le point P serait également, en vertu de ce que nous avons dit au n° **829**, pôle de rotation d'ordre n' et non d'ordre n.

Ainsi, le nombre des homologues de P réunis en P' est exactement n; et, puisque le point P' a été pris quelconque parmi les homologues de P, on peut dire que *les* N *homologues de* P *viennent se confondre entre eux n à n*. On voit, par conséquent, que N est divisible par n et qu'*il y a* $\frac{N}{n}$ *homologues* **distincts** *de* P (le point P lui-même étant compté dans le nombre).

835. Le résultat que nous venons d'obtenir va nous mener à une relation fondamentale entre l'ordre N du groupe et les ordres des différents pôles de rotation.

Soit en effet, P_1 un premier pôle de rotation d'ordre n_1 : nous venons de voir que ce point avait $\frac{N}{n_1}$ homologues distincts. Mais, d'autre part, le pôle

P_1 est commun à $n_1 - 1$ rotations (déplacement identique non compris) : ses homologues seront chacun (**834**) communs à $n_1 - 1$ rotations homologues des premières : soit en tout $\frac{N}{n_1}(n_1 - 1) = N\left(1 - \frac{1}{n_1}\right)$ rotations. Si maintenant, il existe un pôle P_2, non homologue de P_1 et d'ordre (1) n_2, ce point aura $\frac{N}{n_2}$ homologues distincts, fournissant $N\left(1 - \frac{1}{n_2}\right)$ rotations. Un troisième pôle P_3 (s'il en existe), distinct de P_1, de P_2 et de leurs homologues, aura $\frac{N}{n_3}$ homologues fournissant $N\left(1 - \frac{1}{n_3}\right)$ rotations, etc. Nous aurons donc finalement, pour le nombre total de ces rotations, une expression de la forme

$$N\left(1 - \frac{1}{n_1}\right) + N\left(1 - \frac{1}{n_2}\right) + \ldots + N\left(1 - \frac{1}{n_p}\right).$$

Dans ce nombre, le déplacement identique n'est pas compté; mais chacun des $N - 1$ autres déplacements du groupe figure deux fois (et deux fois seulement), une fois pour chacun des deux pôles qu'il possède: d'où la relation à laquelle nous voulions arriver

$$N\left(1 - \frac{1}{n_1}\right) + N\left(1 - \frac{1}{n_2}\right) + \ldots + N\left(1 - \frac{1}{n_p}\right) = 2(N - 1),$$

ou, en divisant par N,

$$(28) \qquad \left(1 - \frac{1}{n_1}\right) + \left(1 - \frac{1}{n_2}\right) + \ldots + \left(1 - \frac{1}{n_p}\right) = 2 - \frac{2}{N}.$$

C'est cette équation que nous allons résoudre par rapport aux entiers positifs (plus grands que l'unité) $n_1, n_2, \ldots, n_p$, N.

Puisque les nombres $n_1, n_2, \ldots, n_p$ sont au moins égaux à 2, chacune des parenthèses du premier membre est au moins égale à $\frac{1}{2}$: donc *le nombre p de ces parenthèses est inférieur* (*et non égal*) *à 4*, car le second membre est plus petit que 2 (l'égalité étant exclue).

D'autre part, le second membre étant au moins égal à 1 (l'égalité n'ayant lieu que pour $N = 2$) pendant que chacune des parenthèses du premier membre est plus petite que 1, on ne peut pas avoir $p = 1$.

836. Si p est égal à 2, l'équation (28) se réduit à

$$\frac{1}{n_1} + \frac{1}{n_2} = \frac{2}{N}$$

et n'a pas d'autres solutions acceptables pour nous que $n_1 = n_2 = N$. Car, il faudrait, sans cela, que l'un des nombres n_1, n_2 fût plus grand que N : ce qui est absurde, puisque N doit être divisible par n_1 et par n_2.

---

(1) Il est bien entendu (voir plus haut, **829**) que n_2 peut être égal à n_1 sans que P_1 et P soient homologues.

$n_1 = n_2 = N$ montre qu'il n'y a que deux pôles de rotation distincts $\left(\frac{N}{n_1} = \frac{N}{n_2} = 1\right)$ et, par conséquent, qu'un seul axe : nous savons qu'alors, les rotations autour de cet axe sont toutes des puissances de l'une d'entre elles. Inversement, il est clair qu'on a bien ainsi un groupe : tel est le groupe des rotations qu'admet (**707**) un angle polyèdre régulier quelconque.

837. Envisageons maintenant le cas de $p = 3$. Dans ce cas, il y a au moins un des nombres n_1, n_2, n_3 qui est égal à 2. Car si on avait $n_1 \geqslant 3$, $n_2 \geqslant 3$, $n_3 \geqslant 3$, chacune des parenthèses du premier membre de l'équation (28) serait au moins égale à $\frac{2}{3}$: leur somme serait donc au moins égale à 2, ce qui ne se peut. Nous prendrons donc $n_3 = 2$, et la relation (28) devient

$$(29) \qquad \frac{1}{n_1} + \frac{1}{n_2} = \frac{1}{2} + \frac{2}{N}.$$

Nous pouvons remarquer tout de suite que cette équation coïnciderait avec l'équation (15′) du n° **715** si l'on changeait n_1, n_2, N, respectivement en m, n, 2A.

Seulement, ici, l'un des nombres cherchés, n_1 par exemple, peut prendre la valeur 2, après quoi l'autre peut être choisi arbitrairement, le nombre N étant égal (comme le montre l'équation (29)) à $2n_2$. Il n'y a alors que deux pôles, et, par conséquent, un seul axe, d'ordre n_2, les rotations qui ne sont pas effectuées autour de cet axe étant toutes des transpositions ; on verra aisément que les axes de celles-ci doivent être perpendiculaires au premier et faire des angles égaux les uns avec les autres.

Les groupes ainsi obtenus ont reçu le nom de *groupes diédraux* (¹) : un tel groupe est celui qu'admet un polygone plan régulier quelconque ou la figure formée par un angle polyèdre régulier et son symétrique, formé en prolongeant les arêtes au delà du sommet.

On notera le cas où n_2 est lui-même égal à 2. Le groupe se compose alors de trois transpositions dont les axes forment un trièdre trirectangle.

Si, maintenant, nous écartons le cas de $n_1 = 2$, les raisonnements du n° **715** sont valables : n_1 est nécessairement à 3 ; n_3 a l'une des valeurs 3, 4, 5. Les valeurs correspondantes de N sont $N = 12, 24, 60$.

838. Démontrons qu'à chacune des combinaisons ainsi obtenues, correspond bien un groupe d'ordre fini (²).

---

(1) Si on divise un grand cercle de la sphère en n parties égales, les n arcs ainsi obtenus peuvent être considérés, à un certain point de vue, comme les côtés (en prolongement les uns des autres) d'un polygone sphérique régulier comprenant un hémisphère entier et la sphère est divisée en deux tels polygones. Le polyèdre correspondant (dièdre) aura deux faces confondues l'une avec l'autre, ainsi qu'il arrivait pour le polygone régulier de deux côtés. (Pl., page 179, note.)

(2) Cela est évident si nous admettons la théorie des polyèdres réguliers, telle qu'elle est exposée aux Compléments, ch. V, puisque nous savons que les cinq espèces de polyèdres réguliers donnent précisément trois groupes qui correspondent aux trois combinaisons en

Les raisonnements sont absolument pareils pour les trois cas : nous traiterons plus particulièrement le plus compliqué, celui de $n_2 = 5$.

Nous savons que le groupe doit comprendre : 1° des rotations d'ordre n_2; 2° des rotations d'ordre $n_1 = 3$; 3° des transpositions. Nous allons trouver une rotation d'ordre 3 et une transposition dont le produit soit une rotation d'ordre n_2.

A cet effet, nous construirons un triangle sphérique (*fig.* 656) ayant pour angles (en A, B, C respectivement) la $n_1^{\text{ème}}$ partie, la $n_2^{\text{ème}}$ partie et la $n_3^{\text{ème}}$ partie de la demi-circonférence (le dernier angle étant droit, puisque $n_3 = 2$). Ce triangle peut être construit, car ses angles satisfont aux conditions de possibilité énoncées au n° **379**. Nous considérerons :

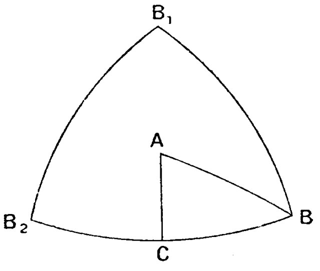

FIG. 656.

1° une rotation R, d'angle égal à la $n_1^{\text{ème}}$ partie de la circonférence, ayant un pôle en A et de même sens que la rotation d'angle égal à BAC $\left(\text{soit à } \frac{1}{2\,n_1}\right)$ qui amènerait le grand cercle AB dans la direction AC;

2° une rotation S, d'angle égal à la $n_2^{\text{ème}}$ partie de la circonférence, ayant un pôle en B et de même sens que la rotation d'angle ABC $= \frac{1}{2\,n_2}$ qui amènerait le grand cercle BA dans la direction BC;

3° une transposition T ayant un pôle en C.

Cette dernière opération transforme B en un point B_2, symétrique de B par rapport au diamètre du point C et par conséquent aussi (puisque l'angle en C est droit) par rapport au plan du grand cercle AC. L'angle sphérique BAB_2 est dès lors mesuré par $\frac{1}{n_1}$ et le point B_2 n'est autre que le transformé de B par R.

Si à ce point B_2, lui-même, on applique la rotation R, on aura un troisième point B_1. Une troisième application de la rotation R ramènerait au point B, car, puisque $n_1 = 3$, on a $R^3 = 1$.

Les triangles sphériques ABB_2, AB_2B_1, AB_1B étant isoscèles et égaux entre eux (1), l'angle sphérique B_1BB_2, double de l'angle à la base du

---

question. Mais, ainsi que nous l'avons dit, nous ne nous appuyons pas sur les résultats des n$^{\text{os}}$ **712-717**, que nous allons, au contraire, démontrer à nouveau dans ce qui va suivre.

(1) Les triangles sphériques ABC, AB_2C sont entièrement extérieurs l'un à l'autre et forment par leur ensemble le triangle sphérique ABB_2. De même les triangles sphériques ABB_2, AB_2B_1. AB_1B sont entièrement extérieurs les uns aux autres et forment par leur ensemble le triangle $BB_1\,B_2$. C'est ce que l'on verra immédiatement en remarquant : 1° que l'on peut construire un triangle sphérique BB_1B_2 ayant ses angles mesurés par $\frac{1}{n_2}$; 2° qu'en divisant ce triangle par les arcs de grands cercles qui joignent le pôle intérieur (**466**) du cercle circonscrit aux sommets et aux milieux des côtés, on obtient des triangles rectangles ayant les angles que nous avons assignés au triangle ABC. Au reste, il importe d'observer qu'aucune de ces remarques n'est nécessaire à la suite de nos raisonnements.

triangle ABB_2, est mesuré par $\frac{1}{n_2}$; donc le point B_2 est le transformé de B_1 par la rotation S.

Mais si nous effectuons successivement les déplacements R et T, le point B reste inaltéré (puisque R l'amène en B_2 et que T ramène B_2 en B) et le point B_1 vient en B_2. Le déplacement RT produit donc sur trois points non en ligne droite (B, B_1 et le centre de la sphère) le même effet que S et, par conséquent (**410**), on a

$$RT = S \tag{30}$$

ce qui peut encore s'écrire

$$S^{-1} = T^{-1}R^{-1} = TR^{-1}, \tag{31}$$

car, T étant une transposition, on a $T^{-1} = T$.

839. Lorsqu'on effectue plusieurs fois de suite la rotation S, le point B_1 a n_2 transformés distincts (le $n_2^{\text{ème}}$ transformé coïncidant avec B_1 lui-même) : soit B_1, B_2, B_3 (*fig.* 657) si $n_2 = 3$; B_1, B_2, B_3, B_4 (*fig.* 658), si $n_2 = 4$; B_1, B_2, B_3, B_4, B_5 (*fig.* 659), si $n_2 = 5$; et les n_2 triangles sphériques $B\,B_1B_2$, $BB_2\,B_3$,.... sont tous équilatéraux et égaux entre eux.

La considération de ces triangles va nous servir à étudier le déplacement

$$U = RTR^{-1}T.$$

Si l'on effectue successivement les déplacements R, T, R^{-1}, T, en partant du point B ou du point B_1, on arrive au point B_3 dans le premier cas, au point B_2 dans le second (¹). Toute rotation, autour d'un axe passant par le centre de la sphère, qui transformera B en B_3 et B_1 en B_2, sera donc (toujours en vertu du n° **410**) identique au déplacement U.

Or, si $n_2 = 3$, la transposition ayant un pôle au milieu C_1 (*fig.* 657) de l'arc de grand cercle BB_3 change bien B en B_3; de plus, le triangle équilatéral BB_3B_1 est changé en un triangle équilatéral égal ayant aussi un côté suivant BB_3, mais placé par rapport à ce côté, dans l'hémisphère où n'est pas B_1 : le troisième sommet de ce triangle, nouvelle position du point B_1, ne peut être autre que B_2.

Donc U est la transposition qui a un pôle en C_1.

Si $n_2 = 4$, le triangle sphérique équilatéral BB_3B_4 (*fig.* 658) (deuxième transformé de $BB_1\,B_2$ par S) admet une rotation d'angle égal à $\frac{1}{3}$ de circonférence, transformée de R par S^2 et dont nous représentons en A_2 (*fig.* 658) le pôle, transformé de A. Cette rotation change B en B_3; en même temps —

(¹) Par exemple, le point B est changé par le déplacement R en B_2, que T change en B, lequel est changé par $R-1$ en B_1, lequel est changé par T en B_3.

le point B_3 venant en B_4 et le point B_4 en B, — le point B_1, symétrique de B_3 par rapport au plan du grand cercle BB_4, vient en un point symétrique de B_4 par rapport au plan du grand cercle BB_3, c'est-à-dire au point B_2. Donc, cette rotation, d'ordre 3, est identique au déplacement U.

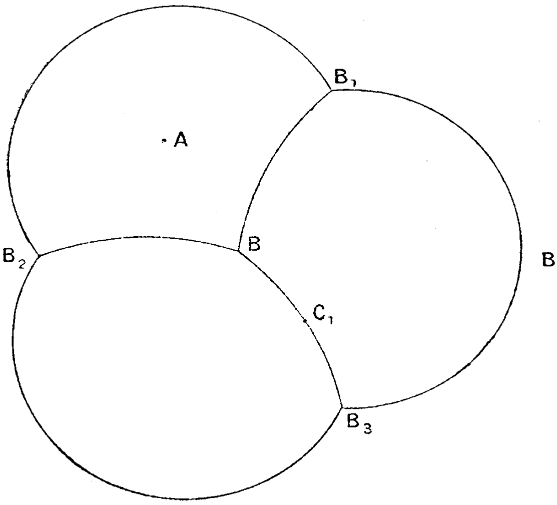

Fig. 657.

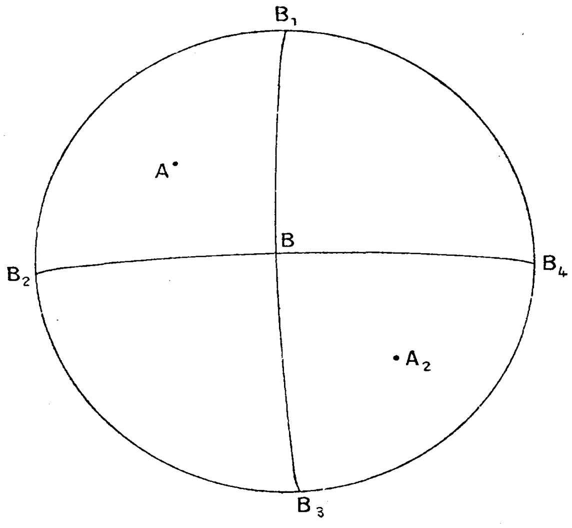

Fig. 658.

Enfin, si $n_2 = 5$, une rotation d'angle égal au $\frac{1}{5}$ de la circonférence, ayant un pôle en B_4 (*fig.* 659), amène B en B_3 (puisque le triangle sphérique BB_3B_4 est équilatéral et a son angle égal à $\frac{1}{5}$); elle amène de même le point B_5 en B et, par conséquent, le point B_1 vient au point B_2, puisque l'un est le symétrique de B_4 par rapport au plan du grand cercle BB_5, et l'autre le symétrique de B_4 par rapport au plan du grand cercle BB_3.

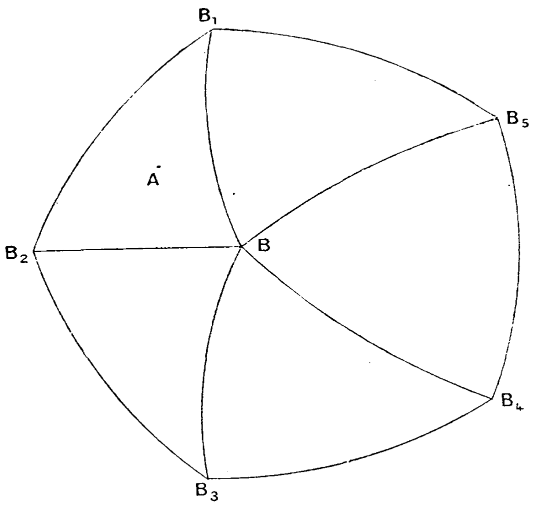

Fig. 659.

Donc, cette rotation d'ordre 5 coïncide avec U.

840. Cela posé, considérons les produits que l'on peut former avec les facteurs R, S, T pris en nombre quelconque et dans tous les ordres possibles. Dans chacun de ces produits, on peut évidemment remplacer plusieurs facteurs *consécutifs* par leur produit effectué, mais il est bien entendu, ainsi que nous l'avons déjà rappelé, que l'on n'a pas, en général, le droit d'intervertir l'ordre des facteurs. Dans ces conditions, il semble que les produits ainsi obtenus soient en nombre infini : nous allons démontrer qu'il n'en est rien et que l'on trouve ainsi des déplacements en nombre limité. Nous aurons ainsi fourni la démonstration demandée, car les produits dont nous parlons forment, d'après la manière même dont il sont définis, un groupe.

Pour arriver à cette démonstration, nous commencerons par remplacer chaque facteur S par sa valeur RT, en vertu de la relation (30). Les symboles que nous aurons à considérer seront donc composés de facteurs R et T. Chaque fois qu'il y aura deux lettres T consécutives, nous pourrons les supprimer, puisque l'on a $T^2 = 1$. De même R^3 étant égal à 1, non seulement nous ne conserverons jamais plus de deux lettres R consécutives, mais lorsqu'il s'en trouvera deux, nous remplacerons leur produit R^2 par R^{-1}. Dans ces conditions, chacun des déplacements en question sera représenté par un symbole composé de lettres T alternant avec des lettres R ou R^{-1}.

Mais un même déplacement peut être représenté par plusieurs symboles de la forme précédente : nous supposerons (ce que nous avons le droit de faire) que, parmi ces symboles équivalents entre eux, ont ait choisi le plus court. Nous considérerons donc un tableau formé de tous les symboles de la forme indiquée tels qu'aucun d'entre eux ne puisse être remplacé par un symbole équivalent, mais plus court; et nous allons montrer que le nombre des symboles du tableau est limité.

841. En premier lieu, la succession RT ne peut figurer plus de deux fois consécutives (pour $n_2 = 5$ ou $n_2 = 4$; plus d'une fois pour $n_2 = 3$) : car la substitution $RT = S$ est d'ordre 5 et, par conséquent, $S^3 = RTRTRT$ peut se remplacer par le symbole le plus court (¹) $S^{-2} = TR^{-1}TR^{-1}$.

De même, le signe R^{-1} (alternant avec T) ne peut figurer plus de deux fois consécutivement : la suite $TR^{-1}TR^{-1}TR^{-1} = S^{-3}$ se remplaçant par $S^{-2} = RTRT$.

Mais il y a plus : la succession RT, par exemple, ne peut figurer deux fois de suite qu'à l'une des extrémités du symbole : en effet, si elle se trouvait deux fois de suite à l'intérieur de ce symbole, elle serait précédée d'une lettre T et suivie du signe R^{-1} (ou exceptionnellement (²) R) ce qui donnerait la succession :

$$\ldots TR\,TR\,TR^{-1}\ldots.$$

(¹) Le symbole ainsi abrégé se réduirait encore, en général, puisque la première lettre T de la succession $TR^{-1}TR^{-1}$ serait en contact avec une lettre T précédente, ce qui permettrait de les supprimer toutes deux, et de même pour la dernière lettre R^{-1} (comparer ci-après).

(²) Cette dernière hypothèse est d'ores et déjà inadmissible si R n'est pas la dernière lettre du symbole, car elle serait suivie d'une lettre T, donnant une troisième fois la succession RT.

Mais si, dans celle-ci, on remplace

$$RTRT = S^2 \text{ par } TR^{-1}TR^{-1}TR^{-1} = S^{-3},$$

il vient

$$\ldots TTR^{-1}TR^{-1}TR^{-1}R^{-1}\ldots\ldots,$$

ce qui (puisque $T^2 = 1$ et $R^{-2} = R$), se réduit à

$$\ldots R^{-1}TR^{-1}TR\ldots,$$

symbole plus court d'une lettre que le premier.

De même, la successsion TR^{-1} ne peut figurer deux fois de suite qu'à l'une des extrémités du symbole.

Par conséquent, sauf aux extrémités, on voit que les lettres R et R^{-1} (entremêlées de lettres T) *alternent régulièrement :* autrement dit que l'ordre est

$$\ldots RTR^{-1}T\,RTR^{-1}T\,RTR^{-1}T\ldots\ldots$$

Mais le produit $RTR^{-1}T$ n'est autre que le déplacement que nous avons nommé U, et nous avons constaté que ce déplacement était d'ordre fini (d'ordre 5, pour $n_2 = 5$).

Dès lors, notre conclusion est établie : notre symbole comprend, au plus, le produit U, écrit une ou deux fois (¹), pouvant être précédé et pouvant être suivi de lettres R ou R^{-1} en nombre au plus égal à 2 (entremêlées de lettres T); et il est clair qu'il n'existe qu'un nombre fini de symboles ainsi formés (en tenant compte des différentes réductions que nous venons d'indiquer, on vérifiera d'ailleurs bien que ce nombre est de 60, et qu'il est de 12 pour $n_2 = 3$, de 24 pour $n_2 = 4$).

842. Nous allons montrer maintenant qu'à chacune des valeurs trouvées pour n_2 ne correspond qu'un seul groupe d'ordre fini (en ne considérant pas comme distincts deux groupes *semblables* entre eux, c'est-à-dire tels que les rotations de l'un soit respectivement les transformées de celles de l'autre par un même déplacement).

Autrement dit, nous montrerons que le groupe correspondant à l'une des combinaisons précédemment obtenues comprend bien trois rotations R, S, T disposées comme nous l'avons indiqué au n° **838** : il comprendra, dès lors, tous les produits qu'on peut former avec les facteurs R, S, T combinés de toutes les façons possibles et coïncidera bien, par conséquent, avec un des groupes qui viennent d'être formés.

Nous prendrons encore $n_3 = 2$, mais nous n'allons pas spécifier, au moins provisoirement, lequel des deux entiers n_1, n_2 est égal à 3 et lequel peut recevoir les valeurs 3, 4 ou 5.

---

(1) Si le produit U figurait trois fois, on le remplacerait (ainsi qu'on a fait pour S^3 par $U^{-2} = TRTR^{-1}\,TRTR^{-1}$, symbole plus court.

Soit alors A un pôle de rotation d'ordre n_1, et considérons, parmi les autres pôles homologues de A, ceux qui sont le plus rapprochés de A : soient A_0 l'un d'eux; A_1, ... , (*fig.* 660) ceux qu'on déduit successivement de A_0 par rotation d'angle $\frac{1}{n_1}$ autour de A.

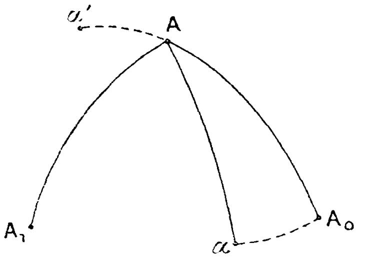

FIG. 660.

Outre ces points A_0, A_1.... il ne saurait exister de point α, homologue de A, aussi rapproché de A que A_1. Car l'arc $A\alpha = AA_0$ devrait tomber dans l'un des angles que font entre eux les arcs AA_0, AA_1..., par exemple, dans l'angle sphérique A_0AA_1 (*fig.* 660) : il devrait alors faire avec l'un des côtés de cet angle, AA_0 par exemple, un angle au plus égal à $\frac{A_0AA_1}{2} = \frac{1}{2\,n_1}\cdot$ Dès lors, dans le triangle isoscèle $AA_0\alpha$, la valeur commune des angles à la base devrait être supérieure à $\frac{1}{2\,n_1}$ (sans quoi la somme des angles serait au plus égale à $\frac{3}{2\,n_1}$, par conséquent — puisque $n_1 \geqslant 3$, — à la demi-circonférence, ce qui est impossible dans un triangle sphérique) et le côté $A_0\alpha$ serait plus petit que AA_0.

Mais il existe une rotation du groupe qui transforme A_0 en A. Cette même rotation transformerait α en un point α' — homologue, lui aussi, de A, — et tel que $A\alpha' < AA_0$, ce qui est contraire à l'hypothèse d'après la quelle A_0 est un homologue de A aussi rapproché que possible de A.

843. Ayant ainsi établi que tout homologue de A, aussi rapproché de A que A_0, est le transformé de A_0 par une puissance convenable de la rotation R qui a pour pôle A et pour angle $\frac{1}{n_1}$, considérons une rotation S du groupe qui transforme A_0 en A. Cette même rotation transformera A en un autre homologue A_1 de A tel que $AA_1 = AA_0$. D'après ce que nous venons de démontrer, A_1 se déduira de A_0 par un déplacement de la forme R^p et on peut évidemment supposer (en multipliant S par une puissance convenable de R) que $p = 1$, de sorte que l'angle sphérique A_0AA_1 sera mesuré par $\frac{1}{n_1}$ (*fig.* 661).

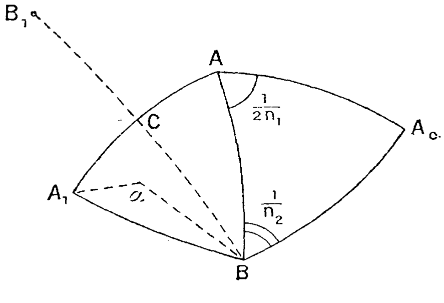

FIG. 661.

Le déplacement S ainsi déterminé aura ses deux pôles de rotation sur le grand cercle qui divise en deux parties égales l'angle A_0AA_1. Soit B celui de ces pôles qui est tel que l'arc de grand cercle AB inférieur à la demi-circonférence soit compris à l'intérieur de l'angle A_0AA_1, de sorte que les deux triangles sphériques isoscèles et égaux AA_0B, AA_1B (*fig.* 661) auront leurs angles à la base mesurés par $\frac{1}{2n_1}$. Ces mêmes triangles auront leur angle au sommet au moins égal à $\frac{1}{n_1}$ ou à $\frac{1}{n_2}$ (n_1 ou n_2 étant l'ordre de la rotation S). Or les nombres $\frac{1}{n_1}$ et $\frac{1}{n_2}$ sont tous deux supérieurs (pour $n_1, n_2 = 3, 4, 5$) à $\frac{1}{2n_1}$: donc on a $AB < AA_0$. Il en résulte (d'après l'hypothèse faite sur A_0) que B n'est pas un homologue de A, autrement dit que S n'est homologue ni de R ni d'une de ses puissances : B est un pôle de rotation d'ordre n_2 (que n_2 soit d'ailleurs égal à n_1 ou différent de n_1) et S a son angle mesuré par $\frac{1}{n_2}$ (¹).

Enfin la rotation $T = R^{-1}S$ transformera A_1 en A et A en A_1 : ce sera donc une transposition ayant un pôle au milieu C de AA_1 ; la relation $T = R^{-1}S$ donne d'ailleurs (en multipliant les deux membres à gauche par R) $S = RT$. En un mot, les rotations R, S, T ont bien entre elles les relations annoncées.

844. Remarquons, de plus, qu'il ne saurait y avoir d'homologue de A plus rapproché de B que A. Si, en effet, α était un tel homologue (par conséquent tel que l'arc de grand cercle $B\alpha$ (*fig.* 661), soit plus petit que BA), il y aurait au moins un des grands cercles BA, $BA_1, \ldots BA_0$, transformés de BA par les puissances successives de S, qui ferait avec $B\alpha$ un angle au plus égal à $\frac{1}{2n_2}$ (comparer **842**) : soit par exemple, BA_1 ce grand cercle. Le plus grand angle du triangle sphérique $BA_1\alpha$ serait alors autre que B_1, sans quoi la somme totale serait au plus égale à $\frac{B}{2n_2}$, par conséquent à la demi-circonférence, ce qui est impossible (comparer encore **842**), et, par conséquent, le côté $A_1\alpha$ serait inférieur au plus grand des deux autres côtés, c'est-à-dire à A_1B, lequel est lui-même (ainsi que nous l'avons vu) plus petit que AA_0. Or ceci ne peut avoir lieu, car une rotation du groupe qui transforme A_1 en A changerait α en un homologue de A plus voisin de ce dernier que A_0 (comparer **842**).

Il ne saurait non plus y avoir d'homologue β de B plus voisin de A que B, sans quoi la rotation qui change β en B changerait A en un homologue α de A tel que $\alpha B < AB$.

La distance AB *est inférieure à un quadrant :* car le point diamétralement

---

(¹) Cet angle ne sera pas un multiple de $\frac{1}{n_2}$, sans quoi la rotation de pôle B et d'angle égal à $\frac{1}{n_2}$ (laquelle appartient au groupe) transformerait A en un de ses homologues plus voisin de lui que A_0, contrairement à l'hypothèse).

opposé à B est un pôle de rotation homologue, soit de A, soit de B : dans les deux cas, nous savons que sa distance à A, est plus grande (¹) que AB.

845. Proposons-nous enfin de faire voir qu'à chacun des groupes obtenus correspondent deux polyèdres réguliers (sauf pour $n_1 = n_2 = 3$, qui ne donne pas deux polyèdres différents) : il nous suffira (**710**) de montrer que ces groupes donnent chacun deux divisions de la sphère en polygones sphériques réguliers et égaux.

A cet effet, considérons un point A de la sphère et appelons *domaine* de ce point la région formée par les points qui sont plus rapprochés de A que d'aucun de ses homologues $A_1 . A_2 \ldots$. Les points de ce domaine sont évidemment ceux qui sont dans le même hémisphère que A, par rapport à chacun des grands cercles perpendiculaires aux milieux de AA_1, de AA_2, etc : le domaine de A est donc un polygone sphérique convexe, limité par des arcs empruntés à tout ou partie des grands cercles en question.

Les points $A_1, A_2, \ldots$ homologues de A, auraient de même leurs domaines respectifs D_1, $D_2 \ldots$. *Tous ces domaines sont extérieurs les uns aux autres et recouvrent, dans leur ensemble, la sphère entière :* en effet, un point quelconque P de la surface appartient toujours à un de ces domaines et, en général, à un seul, à savoir celui qui correspond à l'homologue de A le plus rapproché (²) de P.

Supposons d'abord que le point A ne soit pas un pôle de rotation, de sorte que les points $A_1, A_2 \ldots$, tous distincts entre eux, sont en nombre égal *à* l'ordre N du groupe.

Tous les domaines D, D_1, $D_2 \ldots$, *sont égaux entre eux : chacun d'eux est le transformé de* D *par un déplacement du groupe et par un seul.* Le domaine D_1, par exemple, est le transformé de D par le déplacement qui change A en A_1 : car si P est un point de D, P_1 son transformé, la distance A_1P_1 (égale à AP) est plus petite que la distance du point P_1 à A_2, par exemple (cette distance étant égale à la distance P à l'homologue A' du point A qui est venu en A_2).

Ainsi chacun des domaines D, D_1, $D_2 \ldots$ correspond à un déplacement parfaitement déterminé du groupe.

Tout point P_0 *de la sphère a un homologue et, en général* (³), *un seul dans le domaine* D, puisque P_0 est contenu dans un des homologues de D.

---

(1) Il semblerait que les deux distances puissent être égales. Mais ceci ne pourrait se produire que si le point diamètralement opposé à B était un homologue de B, déduit de B par la rotation de pôle A : il faudrait, pour cela, que cette rotation soit d'ordre pair et l'on peut toujours appliquer le raisonnement du texte au pôle d'ordre 3. Au reste, si la distance AB était d'un quadrant, le triangle BAA_1 aurait ses angles à la base droits, ce qui n'est pas.

(2) S'il existe deux ou plusieurs homologues de A ayant avec P la distance minima, le point P est évidemment mitoyen à plusieurs domaines adjacents entre eux.

(3) Le cas exceptionnel est encore celui où le point P_0 est mitoyen à deux ou plusieurs domaines contigus.

On nomme *domaine fondamental* du groupe G tout domaine qui, comme D, possède cette propriété de renfermer un homologue et un seul de chaque point de la sphère.

846. Supposons maintenant que A soit un pôle de rotation d'ordre (¹) n_1 ou n_2. Alors les points A_1, A_2..., sont en nombre inférieur à N et les domaines D, D_1, D_2,... cessent d'être des domaines fondamentaux. Mais ils sont encore tous égaux entre eux, extérieurs les uns aux autres et recouvrent entièrement la sphère par leur ensemble (les raisonnements présentés tout à l'heure à cet égard continuant à être valables). Nous allons voir que chacun de ces domaines (D, par exemple) est un polygone sphérique régulier.

Pour cela, A étant supposé un pôle d'ordre n_1, reprenons les notations du n° **843** : le point B, pôle d'ordre n_2, sera (**844**) au moins aussi rapproché de A que tout autre homologue de B : chacun des grands cercles limites du domaine D laissera donc A et B dans le même hémisphère ou passera par B : il laissera donc dans le même hémisphère tout point de l'arc de grand cercle (²) AB. Cet arc de grand cercle appartiendra, par conséquent, au domaine D et B, à la frontière de ce domaine. Il en sera de même des points B_1. B_2,..., transformés successifs de B par la rotation R de pôle A et d'angle $\frac{1}{n}$.

Les arcs de grands cercles BB_1, B_1B_2,... déterminent un polygone sphérique régulier, inscrit à un cercle dont A est le pôle intérieur (puisque la distance AB est inférieure à un quadrant). Un grand cercle limite du domaine D ne saurait couper les côtés BB_1, ..., sans quoi il laisserait deux des points B, B_1,.., dans des hémisphères différents, ce que nous savons ne pas être : chacun de ces grands cercles laisse, dès lors, tout le polygone dans un seul et même hémisphère. Donc tout le polygone sphérique BB_1B_2... fera partie du domaine D.

Mais chaque côté de ce polygone est frontière de D : car le symétrique de A par rapport au plan du grand cercle BB_1 est un homologue de A, le point A_1 (*fig.* 661). Donc, notre conclusion est démontrée : le domaine du point A est un polygone sphérique régulier.

Nous aurons bien ainsi deux divisions de la sphère en polygones réguliers égaux : à savoir, les domaines des pôles homologues de A, d'une part; les domaines des pôles homologues de B, de l'autre.

---

(1) Le lecteur étudiera sans difficulté le cas où A est un pôle d'ordre 2, lequel n'offre d'ailleurs aucun intérêt particulier.

(2) Il est bien entendu (**465**) qu'il s'agit de l'arc de grand cercle moindre qu'une demi-circonférence.

!"#$# % & & !' (! !) * * + , * - . *
/ * !0 + ,1 * + , 2 "/" & 3 45

-6 7 & *6 5*5

* 5 * 4 8 * 6 9* 8 :*5* * 51 ;* * 8 & 5 . 5 <
"* * % & & 9* 8 :*5* * 5 2 * 8 & 3 * 8 & 4 *5
4* * * 4* * 5

* 1 < * 5 < 2* 4* 5 * 8 & *6* = 3

!"#$# % & & !' (! !) * * + , * - . *
/ * ! * * * 5 , 2 7532. &> 3 2 "/" & 3 45

!#)?! ! !9* 8 :*5* * 5!9* 8 :*5* * 5 0 @ 5
.A A A A * B 2.A A 4 * 3 # * 5 2 C 3 45

-6 7 & D * E 1 5 . * * &
!"#$# % & & !' (! !) * * + , * - . *
/ * !0 + ,1 * + , 2 "/" & 3 45

!"#$# % & & !' (! !9* 8 :*5* * 5!9* 8 :*5* * 5
0 @ 5 .A A A A * BB 2.A A 5* F 4* 3 # * 5 2 C 3 45

-6 * * * 2 *@G 3

!"#$# % & & !' (! !-4 * 5 5 , 977
2#5&* 531 5 5 !' * * * B 977 2#5&* 53 7%+#+7 07#)%B%+
2 C3 45

D. 6

!"#$# % & & !' (! !-4 * 5 5 , 977
2#5&* 531 5 5 !' * * -4 * 5 5 , 977 2#5&* 53
7%+#+7 B%"B# 2 C3 45

!"#$# % & & !' (! ! & . 6 B - 4* *
H 5! & . 6 B - 4* * H 5 45

-6 0 & * 0 5 III

!"#$# % & & !' (! !# ' 5. 0 *
. 6 *!#J' 5. J J0 * J# . 6 * 45

-6 0 & * 0 5 III

!"#$# % & & !' (! ! 5 , 1 5 5 ! 5
, 1 5 5 45

-6 0 & * 0 5 III

!"#$# % & & !' (! !) * 1 #6 * 1 # 5
#44 5 # B 5 !) * J J#6 * J# 5J#44 5J J# JB 5 45

!"#$# % & & !' (! ! 5 , 1 5 5 !K
K * 5 , 9 H , L B 2 C3 45

!"#$# % & & !' (! !0 F *, + ,!0 F *,
+ , B) 2 C M3 45

!"#$# % & & !' (! ! # * * * 75
!+ , $ 0 * 5 ## 2 3 45

+* 1) * 1 : 6 1 : ; & 2 3

!"#$# % & & !' (! !$ + , % 6 !$
+ , % 6 45

M 2 *@G 3

!"#$# % & & !' (! !9* 8 :*5* * 5!9* 8 :*5* * 5
0 @ 5 .A A A A * BB 2.A A 5* F 4* 3 # * 5 2 C 3 45

-6 - = N5 5 & G *4 * *. 1 * D * & G 8 *
D 5 <

!#)?! ! !9* 8 :*5* * 5!9* 8 :*5* * 5 0 @ 5
.A A A A * BB 2.A A 5* F 4* 3 # * 5 2 C 3 45

5 C 4D. *

!#)?! ! ! 4* 45

* $ 5 * 8 & 6 8 G 5 5 5 <

!"#$# % & & !' (! !) * * + , * - . *
/ * ! * * * 5 , 2 7532. &> 3 2 "/" & 3 45

+* > 1) * M1 : 6 C 1 : ; & CC

!"#$# % & & !' (! !) * * + , * - . *
/ * ! * * * 5 , * 1 2 "/" & 3 45

-6 0 & * %G * 4 *5* 4D. *

+* >

!"#$# % & & !' (! !) * * + , * - . *
/ * ! * * * , 2 "/" & 3 45

!#)?! ! !- . * ! * * * , ! * * *
, 2 "/" & 3 45

-6 0 & * %G * 4 *5* 4D. *

+*

!"#$# % & & !' (! !) * * + , * - . *
/ * ! * * 45

!#)?! ! !- . * ! M * * ! * * 45

D. 2 3 C2 C 3

!"#$# % & & !' (! !) * * + , * - . *
/ * ! 5 + , 2 "/" & 3 45

!#)?! ! !- . * ! > 5 + ,! 5
+ , 2 "/" & 3 45

-6 0 & * %G * 4 *5* 4D. * O * * 6A 4* * 5* *@G 5* 4 8 *

D. 2 3

- Exercise 4.5.6 corrects the sophism discussed in section 2.2;
- Exercise 4.6.2 completes the section 3.5 coverage of SSA triangle congruence; and
- Exercise 4.8.3 derives the formula for the volume of a sphere mentioned in section 3.14.

Others introduce entirely new subjects: for example,

- Independence proofs for various axioms in section 4.1;
- Arithmetic, geometric, and harmonic means in 4.5 and 4.6; and
- Spherical trigonometry in section 4.9.

There are a few original exercises here. But for most, the author acknowledges debt to his predecessors. Only a few exercises, with clear heritage, are given footnote citations. Others are old gems, whose origin would be hard to establish. The author has used much of this exercise set for decades. Each year new, different, solutions appear. You're encouraged to dive in, try out ideas old and new, and make these problems *your* old friends.

Hints are provided for a few exercises. Sometimes they suggest an appropriate method. Don't feel constrained to follow that, however, since many problems yield to attack from several directions. Other hints are approximate solutions: for example, "the answer is about 14.14." These help you check the feasibility of your solution. But you need to find the *exact* answer, which isn't obvious, even given the approximation. Indeed, 14.14 is *about* the same as both $11 + \pi$ and $10\sqrt{2}$, but those numbers are different.

4.1 Exercises on the incidence and ruler axioms

Concepts

Models of axioms
Finite geometric systems
Independence of the incidence axioms
Four-dimensional simplices
Ruler placement along a line

The incidence axioms of section 3.1 entail only a small part of Euclidean geometry. It's impossible to single out many applications of that material alone, so the exercises in this section are theoretical.

Axiom I5 states roughly that there are enough points to do *some* three-dimensional geometry. It's worded awkwardly. One reason for that was to facilitate the proof of the more informative lemma 3.1.3 and theorem 3.1.4. Exercise 1 explores alternate wordings. Exercise 2 shows that you can't do without I5, because the remaining axioms can be interpreted to describe

!"#$# % & & !' (! !) * * + , * - . *
/ * ! 6 $ 5 . K (2 C M3 45

!#)?! ! !- . * ! 6 $ 5 .
K (! 6 $ 5 . K (2 C M3 45

-6 *

!"#$# % & & !' (! !) * * + , * - . *
/ * ! * * 5 ' , 5 45

!#)?! ! !- . * ! * * 5 ' , 5! * * 5
' , 5 45

-6 +* 1) * 1 : 6 1 : ;

!"#$# % & & !' (! !) * * + , * - . *
/ * !) & . , 2 "/" & 3 45

!#)?! ! !- . * ! C) & . ,!) &
. , 2 "/" & 3 45

-6 0 & 5

7

Curves, lengths, and curvatures

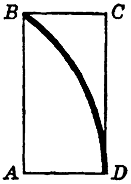

Figure 7.11

7.1 Curves and lengths One of the main purposes of this chapter may seem at first sight to be quite modest. We want to show that the length L of the circular arc of Figure 7.11, in which $ABCD$ is a rectangle and the arc is tangent to CD at D, satisfies the inequality

$$(7.12) \qquad |\overrightarrow{AB}| \leqq L \leqq |\overrightarrow{BC}| + |\overrightarrow{CD}|.$$

Before we ask our little sister to solve the problem, we should ask ourselves a question that shows us that the problem is not completely simple. What is a circular arc, and how do we know that it determines a number that can be called its length? It is clear that we need some definitions before we can do anything.

For an accurate treatment of matters relating to lengths of curves, we need more information about things that are sometimes called curves and are sometimes called oriented curves. A bumblebee can give us preliminary ideas by starting at A, flying to a rose at B, flying on to another rose at C, flying back over the same route to B, and then flying to D and on to E as in Figure 7.13. His total path is a curve, and we get the idea that a curve is not determined by a point set. We cannot know what the curve is until we know the order in which points on the curve were visited. For example, if the bumblebee flies from E back to A by traversing his route in reverse, his path is a new curve which can be called the negative of the original one. While the matter has psychological rather than logical importance to us now, we can feel quite confident that if the bumblebee makes a flight over the interval $t_1 \leqq t \leqq t_2$, and if we introduce an x, y, z coordinate system, then at each time in the interval he is surely someplace and that if we denote his coordinates by $x(t)$, $y(t)$, $z(t)$, then these coordinate functions are continuous functions of t. Of course, the curve does not uniquely determine the coordinate functions because the bumblebee can fly his course at different speeds, but any one set of appropriate coordinate functions does determine the intrinsic curve. The above brief discussion of curves leaves many unanswered questions. It will have served its purpose if it provides a hazy feeling that the following definition uses words in a reasonable way.

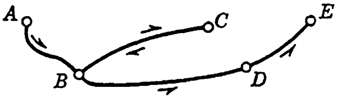

Figure 7.13

Definition 7.14 *If x, y, z are continuous functions of t over an interval, then the ordered set of points*

$$P(t) = P(x(t), y(t), z(t)) \tag{7.141}$$

for which t lies in the interval, and for which $P(t')$ is said to precede $P(t'')$ if $t' < t''$, is a curve (or oriented curve) C.

Professional creators of complicated curves can give an example of a curve in E_3 that is clever enough to "pass through" each point in a given cube or even in the whole E_3. Such curves are space-filling curves. It must not be presumed that all curves are complicated things, however. For example, when $z(t) = 0$ for each t, the curve lies in the xy plane and we omit the $z(t)$ when no confusion can result. We always have the possibility of setting $x(t) = t$ and replacing t by x. This shows that if f is continuous over $a \leqq x \leqq b$, then the set of points $(x, f(x))$ on the graph of $y = f(x)$ becomes a curve when we so order the points that $(x_1, f(x_1))$ precedes $(x_2, f(x_2))$ when $a \leqq x_1 < x_2 \leqq b$. Unless an explicit statement to the contrary is made, it is presumed that the points are ordered in this way whenever the graph of $y = f(x)$ is called a curve. Finally, such things as circles, ellipses, rectangles, and triangles become curves as soon as we think of them as being traversed once in the positive (counterclockwise) or negative (clockwise) direction.

Supposing that C is a given curve for which the interval appearing in Definition 7.14 is a closed interval $a \leqq t \leqq b$, we proceed to define (when it exists) a number $|C|$ which is called the length of C. Let P be a partition of the interval by partition points $t_0, t_1, \cdots, t_n$ for which

$$a = t_0 < t_1 < t_2 < \cdots < t_{n-1} < t_n = b. \tag{7.15}$$

For each $k = 0, 1, 2, \cdots, n$, let P_k be the point on C for which $t = t_k$. Figure 7.151 is a schematic figure that may be helpful, but it is much too simple to show how much care is needed to guarantee that the points are not scrambled in an inappropriate way. The polygon (or broken line) running from P_0 to P_1 to P_2 and so on to P_n is said to be *inscribed* in C. The number S_P defined by

$$S_P = \sum_{k=1}^{n} |\overrightarrow{P_{k-1}P_k}|$$

is the sum of the lengths of the sides of the inscribed polygon. Let S be the set of numbers S_P obtained by making partitions P of the interval $a \leqq t \leqq b$. In case there is no number M such that $S_P \leqq M$ whenever S_P is in S, it is said that C does not have length or that C does not have finite length or that C has infinite length or that $|C| = \infty$. In case there is a number M such that $S_P \leqq M$ whenever S_P is in S, then Theorem 5.46 guarantees existence of a least number $|C|$ such that $S_P \leqq |C|$ whenever S_P is in S. This number $|C|$ is then, by definition, the *length*† of C.

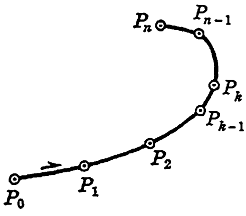

Figure 7.151

For future reference we state without proof the following theorem which involves notation we have been using.

Theorem 7.16 *The curve C has finite length $|C|$ if and only if*

$$|C| = \lim_{|P|\to 0} \sum_{k=1}^{n} |\overrightarrow{P_{k-1}P_k}|.$$

We now use the definition of length to prove the following theorem which gives, as a corollary, the desired result involving Figure 7.11.

Theorem 7.17 *Let f be continuous and monotone increasing (or monotone decreasing) over $a \leqq x \leqq b$. Let C be the graph of $y = f(x)$ so oriented*

---

† In old books particularly, a curve C having finite length is sometimes said to be *rectifiable;* the ancient idea is that if C is a string we could pull on the ends and straighten it out to get a straight string of finite length. We can all know enough about logic to know that if C is not a string, then little credence is placed upon consequences of the false assumption that C is a string.

that $(x_1, f(x_1))$ *precedes* $(x_2, f(x_2))$ *when* $x_1 < x_2$. *Then* C *has finite length* $|C|$ *and*

$$(7.171) \qquad |f(b) - f(a)| \leqq |C| \leqq |b - a| + |f(b) - f(a)|.$$

To prove this, let P be a partition of the interval $a \leqq x \leqq b$, and, for each k, let $P_k = (x_k, f(x_k))$. Let $\Delta x_k = x_k - x_{k-1}$ and let

$$\Delta y_k = f(x_k) - f(x_{k-1}).$$

Then

$$(7.172) \qquad |\overrightarrow{P_{k-1}P_k}| = \sqrt{\Delta x_k^2 + \Delta y_k^2}.$$

But

$$(7.173) \quad \Delta y_k^2 \leqq \Delta x_k^2 + \Delta y_k^2 \leqq |\Delta x_k|^2 + 2|\Delta x_k \, \Delta y_k| + |\Delta y_k|^2 = (|\Delta x_k| + |\Delta y_k|)^2.$$

Therefore,

$$(7.174) \qquad |\Delta y_k| \leqq |\overrightarrow{P_{k-1}P_k}| \leqq |\Delta x_k| + |\Delta y_k|.$$

But the numbers Δx_k are all positive and the numbers Δy_k are all negative (or all positive), and hence (why?) addition gives

$$(7.175) \quad |f(b) - f(a)| \leqq \sum_{k=1}^{n} |\overrightarrow{P_{k-1}P_k}| \leqq |b - a| + |f(b) - f(a)|.$$

Since this inequality holds for each partition P, the curve C has finite length $|C|$. Moreover, the least upper bound of the central sums cannot exceed the particular upper bound in (7.175), and no upper bound can be less than $|f(b) - f(a)|$. This proves Theorem 7.17.

We have been studying lengths of *curves*, that is, lengths of ordered sets of points of special types. There are other (and different) theories involving lengths of unordered sets S in E_3 and E_2 and E_1. We need not be authorities on these matters, but we should have at least a vague idea that, in elementary mathematics, the part of the set S of Figure 7.18 that lies "between" A and B has length L if there is a "natural ordering" of the points of the part that yields a curve C having length L.

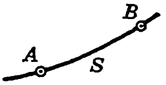

Figure 7.18

Problems 7.19

1 Tell how the length $|C|$ of a curve C is defined.

2 Show that if a curve C having finite length runs from A_1 through A_2 to A_3, then A_2 separates the curve into two parts each having finite length. It is not quite so easy to show that "the whole is equal to the sum of its parts," but it can be done. In particular, the length of a simple polygon is the sum of the lengths of its straight segments.

3 Show that the projection on the xy plane of a curve in E_3 is a curve in E_2.

4 Discover continuous functions x and y such that as t increases over the interval $0 \leqq t \leqq 4$, the point $(x(t), y(t))$ goes once in the positive direction around the square having vertices at the points (0,0), (0,1), (1,1), and (0,1). *Hint:* Write the equations which tell where a pencil point will be at time t if it traverses the square with unit speed. *One ans.:* $x(t) = t$ and $y(t) = 0$ when $0 \leqq t \leqq 1$; $x(t) = 1$ and $y(t) = t - 1$ when $1 \leqq t \leqq 2$; $x(t) = 3 - t$ and $y(t) = 1$ when $2 \leqq t \leqq 3$; $x(t) = 0$ and $y(t) = 4 - t$ when $3 \leqq t \leqq 4$.

5 Let L_1 be the length of the central path from A to B in Figure 7.191 and let L_2 be the length of the upper path. Find $L_2 - L_1$. *Ans.:* $(\pi - 2)h$.

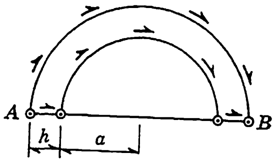

Figure 7.191

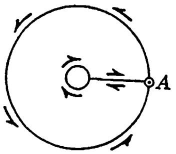

Figure 7.192

6 Supposing that the inner and outer circles of Figure 7.192 have radii ϵ and R, find the length of the curve which begins and ends at A.

7 Draw a square and from each corner draw a line segment toward the center but reaching only halfway to the center. Then insert arrows to specify a curve which forms the boundary of the inner region. Find the length of the curve.

8 This problem requires us to think about and calculate the lengths L_1 and L_2 of two of the paths by which an insect might crawl from the bottom B to the top T of the three-dimensional ring or anchor ring or torus of Figure 7.193. The first path, of length L_1, is a semicircle of radius $b + a$ which lies on the outer circumference of the torus. The second path, of length L_2, consists of a semicircle of radius a, and then a semicircle of radius $b - a$ which lies on the inner circumference of the torus, and finally another semicircle of a radius a. Try to guess which of the paths is shorter and then calculate L_1 and L_2. *Ans.:*

$$L_1 = \pi(a + b), \qquad L_2 = \pi a + \pi(b - a) + \pi a,$$

so the two paths have equal lengths. *Remark:* If a curve G lies in a set S and joins two points P_1 and P_2 of S, and if the length of G is less than or equal to the length of each other curve C in S which joins P_1 and P_2 then G is called a *geodesic* in S. As is easily imagined, the study of geodesics on a torus is an honorable part of an interesting subject.

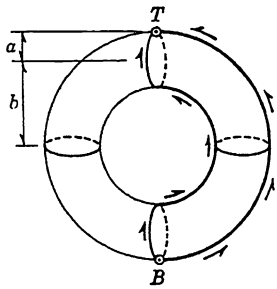

Figure 7.193

9 Creators of interesting tales for children say that, before the time when Columbus sailed across the ocean wet, everybody thought that the earth was flat. It has in fact been widely known for more than two thousand years that the earth is much like a spherical ball, but is not a spherical ball because mountains and valleys are quite noticeable around the Mediterranean sea and some other places. To show that a little information about lengths of arcs can have quite astonishing consequences, we look at the method said to have been used by the industrious Eratosthenes (*c.* 275 B.C. to *c.* 195 B.C.) to find what the radius (a number) of the earth would be if the earth were sandpapered to the shape of a spherical ball. Figure 7.194 shows O, the center of the earth, and a circular arc AB on the surface of the polished earth. The dotted vertical lines represent rays of light coming from the sun, and these rays are so nearly parallel that approximations can be based on a figure in which the rays are parallel to the line OA. Thus an observer at A finds that the sun is at his zenith. At B a pole BC of height h is erected in such a way that an observer at B thinks it is vertical, that is, the points O, B, C lie on the same line. In addition to h, two other lengths are obtained. In the first place, we find the length a of the circular arc DC which has its center at B (the base of the pole BC) and which has a shadow that just covers the pole. In the second place, the length b of the arc AB is obtained by more or less reliable surveyors. Let r be the radius of the earth. Since parallelism of the light rays implies that the angles DBC and AOB are equal, say to θ, we obtain the equation

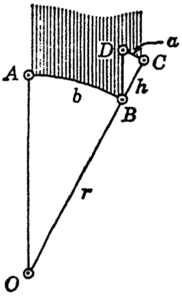

Figure 7.194

$$\frac{a}{h} = \frac{b}{r}$$

from the fact that each of the ratios is equal to θ. From this equation, r is easily calculated. It is not to be presumed that Eratosthenes spoke English and knew about miles, meters, and radians, but it is said that his calculations produced estimates of the radius and circumference of the earth that are almost as good as the estimates (radius 4000 miles, circumference 25,000 miles) that are ordinarily used for rough calculations. The fact that reliability of computed results depends upon accurate surveying is precisely the reason why two men named Mason and Dixon were commissioned to do some accurate surveying.

10 This problem, like some others, does not have a number for an answer. It requires us to think about connections between graphs and curves, and to learn some geometric terminology. The graph of a polynomial of degree n, that is, the graph of an equation of the form

$$y = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n, \tag{1}$$

in which $a_n \neq 0$, used to be (and occasionally still is) called a *parabola of degree n*. For example, a line is a parabola of degree 1 and a cubic is a parabola of degree 3. While this terminology is almost extinct, the graph of the equation

$$y^2 = x^3 \tag{2}$$

still is called a *semicubical parabola*, the old idea being that, when we solve for y, we get an exponent which is not an integer but is half of 3. According to this

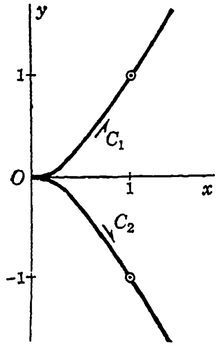

Figure 7.195

terminology, the semicubical parabola is a graph (not necessarily a curve) which is the point set shown in Figure 7.195. The graph has two *branches*, one being the upper branch consisting of points for which $y = x^{3/2}$, and the other being the lower branch consisting of points for which $y = -x^{3/2}$. Each of these branches contains the origin.

When the points on the upper branch are ordered in such a way that $P(x_1,y_1)$ precedes $P(x_2,y_2)$ when $x_1 < x_2$, the branch becomes a curve C_1 in the first quadrant having an end (or end point) at the origin. Similarly, when the points on the lower branch are ordered in such a way that $P(x_1,y_1)$ precedes $P(x_2,y_2)$ when $x_1 < x_2$. this branch becomes another curve C_2 having an end at the origin. With these orderings, the graph consists of the points on the two curves C_1 and C_2, the orderings being indicated by the arrows of Figure 7.195 which lie between the curves and the positive x axis. When $x > 0$, the curve C_1 has a tangent at the point $(x,x^{3/2})$ which has slope $\frac{3}{2}x^{1/2}$, and the curve C_2 has a tangent at the point $(x,-x^{3/2})$ which has slope $-\frac{3}{2}x^{1/2}$. We can observe that C_1 has a forward tangent at the origin whose slope is 0 and that C_2 has a forward tangent at the origin whose slope is 0. Although we have definitions involving tangents to graphs of equations of the form $y = f(x)$ and have definitions involving tangents to curves, we have no definitions which we can apply to decide whether the x axis is tangent to the semicubical parabola at the origin. In accordance with terminology other applications of which are not explained here, we say that the semicubical parabola has a *cusp* at the origin.

11 Prove that the curve C consisting of points P for which

$$\overrightarrow{OP} = t^2\mathbf{i} + t^3\mathbf{j} \tag{1}$$

coincides with the semicubical parabola having the equation $y^2 = x^3$. Sketch a new graph of the semicubical parabola and insert arrows which show how the points on the semicubical parabola are ordered to produce the single curve C. Put (1) in the equivalent form

$$\mathbf{r} = t^2\mathbf{i} + t^3\mathbf{j} \tag{2}$$

and, supposing that a particle moves along C in such a way that its displacement vector at time t is $\mathbf{r}$, find its velocity and acceleration at time t. See what happens when you try to write the equation of the line tangent to C at the point occupied by the particle at time $t = 0$.

12 The *folium of Descartes* is the graph of the parametric equations

$$x = \frac{3at}{1+t^3}, \qquad y = \frac{3at^2}{1+t^3}. \tag{1}$$

The vector $\mathbf{r}$ running from the origin to the point $P(x,y)$ on the folium is

$$\mathbf{r} = 3a\left[\frac{t}{1+t^3}\mathbf{i} + \frac{t^2}{1+t^3}\mathbf{j}\right], \tag{2}$$

there being one such point for each t for which $t \neq -1$. One who has considerable time to study the functions defined by (1) and to make some calculations can discover the manner in which P wanders as t increases. As t increases over the interval $-\infty < t < -1$, P runs over the curve C_1 extending from the origin (but not including the origin) down into the fourth quadrant. As t increases over the interval $-1 < t < \infty$, P traverses the curve C_2 which comes from the third quadrant to the origin and then runs around the loop toward the origin but does not again contain the origin. This problem requires that we obtain some more information without working so hard. Show that if (1) holds, then $y = tx$ and

$$x^3 + y^3 = 3axy. \tag{3}$$

Show that if (3) holds and $y = tx$, then (1) holds. Thus (3) is an equation of the folium. If $P(x,y)$ is a point on the folium, then (3) shows that $P(y,x)$ is also on the folium. Therefore, the line having the equation $y = x$ is an axis of symmetry of the folium. This suggests that we introduce the X, Y axes of Figure 7.196 which bear unit vectors $\mathbf{I}$ and $\mathbf{J}$. Use the formulas

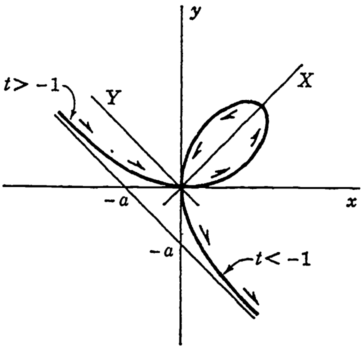

Figure 7.196

$$\mathbf{i} = \frac{1}{\sqrt{2}}(\mathbf{I} - \mathbf{J}), \qquad \mathbf{j} = \frac{1}{\sqrt{2}}(\mathbf{I} + \mathbf{J}) \tag{4}$$

with (2) to obtain the new equation

$$\mathbf{r} = \frac{3a}{\sqrt{2}}\left[\frac{t}{1 - t + t^2}\mathbf{I} + \frac{t(1-t)}{1+t^3}\mathbf{J}\right] \tag{5}$$

of the folium. Treating (5) as an equation of the form $\mathbf{r} = X\mathbf{I} + Y\mathbf{J}$, show that

$$X = \frac{3a}{\sqrt{2}}\frac{t}{1 - t + t^2}, \qquad \frac{dX}{dt} = \frac{3a}{\sqrt{2}}\frac{1 - t^2}{(1 - t + t^2)^2}$$

and use the result to find the minimum and maximum values of X and obtain more information about the folium.

13 Let f be continuous over the interval $0 \leqq x \leqq 1$ and let C be the set of points $(x, f(x))$ so ordered that $(x_1, f(x_1))$ precedes $(x_2, f(x_2))$ if $x_1 < x_2$. We may feel that C must have finite length, and we investigate the matter. Theorem 7.17 shows that C must have finite length if f is piecewise monotone, but we cannot be so sure when C is not piecewise monotone. When p and q are positive integers, the function $\sin^p x^q$ [or the function g for which $g(x) = \sin^p x^q$] is not piecewise monotone over the infinite interval $x > 1$ and the function $\sin^p (1/x^q)$ is not piecewise monotone over the interval $0 < x < 1$. Functions f for which $f(0) = 0$ and

$$f(x) = x^r \sin^p \frac{1}{x^q}$$

are, when $r > 0$, continuous but not piecewise monotone over $0 \leqq x \leqq 1$. Supposing that $p = q = r = 2$, show that $0 \leqq f(x) \leqq 1$ and f is differentiable as well as continuous over $0 \leqq x \leqq 1$. It can be proved that C does not have finite length.

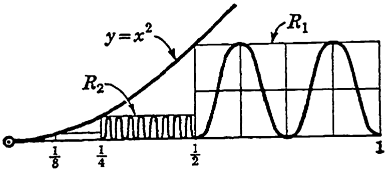

Figure 7.197

14 It is worthwhile to cultivate the ability to understand and even to construct simple examples of curves that have properties of various sorts. Figure 7.197 shows a part of the graph C of a function f, defined over $0 \leqq x \leqq 1$, for which $f'(x)$ exists when $0 \leqq x \leqq 1$ and, moreover, $f(x) = f'(x) = 0$ when $x = 0$, $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \cdots$. We do not bother to give formulas expressing $f(x)$ in terms of x, being content to describe the graph C. The rectangles have their upper left corners on the graph of the equation $y = x^2$. The part C_1 of the graph in the rectangle R_1 over the interval $\frac{1}{2} \leqq x \leqq 1$ consists of four monotone parts. Show that $1 \leqq |C_1| \leqq 2$. We now turn our attention to the rectangle R_2 which lies above the interval $\frac{1}{4} < x < \frac{1}{2}$ and has height $(\frac{1}{4})^2$. In this rectangle we construct another part C_2 of C, this part consisting of 4^2 monotone parts so that $1 \leqq |C_2| \leqq 2$. Tell how to continue the construction so that the total resulting curve C will not have finite length.

15 Let $x(t)$ and $y(t)$ be continuous over $a \leqq t \leqq b$ so that the set of points $P(t)$ for which $P(t) = (x(t), y(t))$ and $a \leqq t \leqq b$, so ordered that $P(t_1)$ precedes $P(t_2)$ when $t_1 < t_2$, constitutes a curve C in the xy plane. The curve C is said to be a *closed curve* if $P(a) = P(b)$, that is, if the curve ends where it starts. Give some examples of closed curves and some examples of curves that are not closed.

16 A set A in the xy plane is said to be *connected* if to each pair P_1 and P_2 of points in A there corresponds a curve C of the type described in Problem 15 such that $P(a) = P_1$, $P(b) = P_2$, and each point of C is a point of A. Let Γ (capital gamma) be the circle with center at P_0 and radius r. Prove that the set A_1 of points P for which $|\overrightarrow{P_0P}| < r$ (this set being the *interior* of Γ) is a connected set. Prove that the set A_2 of points P for which $|\overrightarrow{P_0P}| > r$ (this being the *exterior* of Γ) is a connected set. Prove finally that if a set A_3 contains a point P_1 for which $|\overrightarrow{P_0P_1}| < r$ and a point P_2 for which $|\overrightarrow{P_0P_2}| > r$ but contains no point of the circle Γ, then A_3 is not a connected set. *Proof of last part:* Suppose, intending to establish a contradiction, that the set A_3 is connected. Then there is a curve C, determined by continuous functions $x(t)$ and $y(t)$ as in Problem 15, such that $P(a) = P_1$, $P(b) = P_2$, and each point of C is a point of A_3. Let

$$f(t) = |\overrightarrow{P_0P(t)}| = \sqrt{(x_0 - x(t))^2 + (y_0 - y(t))^2}$$

so that $f(t)$ is the distance from P_0 to $P(t)$. Then $f(a) < r$ and $f(b) > r$. Since f is a continuous function, it follows from the intermediate-value theorem (Theorem 5.48) that there is a number t^* for which $a < t^* < b$ and $f(t^*) = r$. The point $P(t^*)$ is a point of the circle Γ, so $P(t^*)$ is not in A_3 and we have the required contradiction. *Remark:* The geometrical nature of a circle in a plane is so simple that basic facts involving its interior and exterior are easily described and easily proved. The remaining problems of this list involve more general situations.

17 Sketch some figures to obtain preliminary ideas about matters relating to the following definition. Let S be a given nonempty set in the xy plane and let P' be a point of S. Let S' be the set of points P'' in S that can be connected to P' by curves lying in S. This means that P'' is a point of S' if and only if there exist functions $x(t)$ and $y(t)$ continuous over an interval $a \leqq t \leqq b$ such that $P(x(a), y(a))$ is P', $P(x(b), y(b))$ is P'', and $P(x(t), y(t))$ is a point in S whenever $a \leqq t \leqq b$. The set S' is a connected set because if P_1 and P_2 lie in S', then P_1 and P_2 can be connected to P' by curves C_1 and C_2 lying in S' and these two curves can be combined to give a single curve C connecting P_1 to P_2 and lying in S'. This set S' is the *maximal connected subset* of S that contains P' or, briefly, the *component* of S that contains P'.

18 The curve C of Problem 15 is said to have a *multiple point* at the point Q if there exist two numbers t_1 and t_2 such that $a \leqq t_1 < t_2 < b$ (or $a < t_1 < t_2 \leqq b$) and the two points $P(t_1)$ and $P(t_2)$ coincide with Q. The curve is said to be *simple* (free from multiple points) if it has no multiple points. Give some examples of closed curves that have multiple points and some examples of simple closed curves.

19 The French mathematician Camille Jordan (1838–1922) was the first person to give serious attention to the difficult question whether each simple closed plane curve "separates the plane" into exactly two nonempty components one of which constitutes the outside of the curve and the other of which constitutes the inside. Such curves are called *Jordan curves.* The following theorem is known as the Jordan curve theorem.

Theorem *If C is a simple closed plane curve (or Jordan curve), then the set S of points of the plane which are not points of C contains exactly two (no more and no fewer) nonempty components S_1 and S_2.*

The author now owes his readers an explanation. It never has been and need not be expected that students of elementary calculus know anything about the Jordan curve theorem. Nevertheless, the author insists that each student should have opportunities to pick up ideas about mathematics. We can know that the Jordan curve theorem is so difficult that Jordan never succeeded in proving it and that we must learn more mathematics before we can undertake to understand proofs that have been given or to construct new proofs. We can know that many persons who never generated much interest in additions of fractions become intensely interested in problems involving sets and curves.

7.2 Lengths and integrals

Let x, y, and z be given functions having continuous derivatives over the closed interval $a \leqq t \leqq b$. Let

$$P(t) = P(x(t), y(t), z(t)) \tag{7.21}$$

so that, for each t, $P(t)$ is the point having coordinates displayed in (7.21). Let $\mathbf{r}(t)$ denote the vector running from the origin to $P(t)$ so that, for each t,

$$\mathbf{r}(t) = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k}. \tag{7.22}$$

We propose to set up an integral for the length $|C|$ of the curve C traversed by $P(t)$, or by the tip of the vector $\mathbf{r}(t)$, as t increases from a to b. It could be supposed that our main interest is in lengths of curves and that we should get our result as quickly as possible, but this is not true. Integrals are very important things, and we proceed slowly to learn as much as we can about them.

Our first step is to make a partition P of the interval $a \leqq t \leqq b$ having partition points $t_0, t_1, \cdots, t_n$ and to sketch schematic figures showing the partition and points $P_{k-1} = P(t_{k-1})$ and $P_k = P(t_k)$ on C. We then tell ourselves that if $|P|$, the norm of the partition, is small then $|\overrightarrow{P_{k-1}P_k}|$ should be a good approximation to the length of the part of C between P_{k-1} and P_k, so $\Sigma|\overrightarrow{P_{k-1}P_k}|$ should be a good approximation to $|C|$ and hence it should be true that

$$|C| = \lim \Sigma|\overrightarrow{P_{k-1}P_k}|. \tag{7.23}$$

We have employed the fundamental ideas about length given in Section 7.1. The next step is to put the right member into a more useful form. While introduction of the abbreviations may be unwise when acres of paper and blackboard space are available, we simplify our formulas by setting

$$\Delta t_k = t_k - t_{k-1}, \qquad \Delta x_k = x(t_k) - x(t_{k-1}), \qquad \Delta y_k = y(t_k) - y(t_{k-1}),$$
$$\Delta z_k = z(t_k) - z(t_{k-1}).$$

Then (7.23) gives

$$|C| = \lim \Sigma \sqrt{\Delta x_k^2 + \Delta y_k^2 + \Delta z_k^2}. \tag{7.231}$$

We can make this look much more like a limit of Riemann sums by introducing factors Δt_k in numerators and denominators to obtain

$$|C| = \lim \sum \sqrt{\left(\frac{\Delta x_k}{\Delta t_k}\right)^2 + \left(\frac{\Delta y_k}{\Delta t_k}\right)^2 + \left(\frac{\Delta z_k}{\Delta t_k}\right)^2}\, \Delta t_k. \tag{7.232}$$

The possibility of making further progress is provided by the hypothesis that x, y, and z have continuous derivatives.

We should be realistic and realize that there are different ways to proceed. We can cheerfully adopt the view that, however we choose t_k^* in the kth subinterval of our partition, the kth term in (7.232) should be closely approximated by $f(t_k^*)\,\Delta t_k$, where

$$f(t) = \sqrt{[x'(t)]^2 + [y'(t)]^2 + [z'(t)]^2}. \tag{7.233}$$

Thus we can expect that

$$|C| = \lim \Sigma f(t_k)\,\Delta t_k. \tag{7.234}$$

Since our hypothesis guarantees that f is continuous and hence integrable, the Riemann sums do have a limit and we are led to the formula

$$|C| = \int_a^b \sqrt{[x'(t)]^2 + [y'(t)]^2 + [z'(t)]^2}\, dt. \tag{7.24}$$

This is, in fact, a correct and very satisfying formula which treats all coordinates alike and enables us to calculate lengths of curves that wind around through E_3. In case C lies in the xy plane, everything is the same except that $z(t) = z'(t) = 0$ and the formula is thereby simplified. In applications, it often happens that $x(t) = t$ and the formula is written in one or another of the forms

$$(7.241) \quad |C| = \int_a^b \sqrt{1 + [y'(t)]^2 + [z'(t)]^2}\, dt = \int_a^b \sqrt{1 + [y'(x)]^2 + [z'(x)]^2}\, dx = \int_a^b \sqrt{1 + \left(\frac{dy}{dx}\right)^2 + \left(\frac{dz}{dx}\right)^2}\, dx$$

and, for plane curves, we have $z = 0$. It is not easy to be sure that the "cheerful derivation" of (7.24) is slovenly. A person can claim that he will not use the definition of length given in Section 7.1, that he is interested only in curves C for which $x'(t)$, $y'(t)$, and $z'(t)$ are continuous, and that all of the work that he has done merely motivates his definition whereby $|C|$ is *defined* by (7.24). His position is tenable and, so long as he stays in his own garden, it is even reasonable. In fact, many quantities in pure and applied mathematics are defined by integrals, and sometimes the definitions are so well motivated that readers (and even writers) fail to recognize the fundamental fact that it is utterly impossible to *prove* correctness of a formula for something that has not been defined.

We start again with the formula

$$(7.25) \qquad |C| = \lim \sum \sqrt{\left(\frac{\Delta x_k}{\Delta t_k}\right)^2 + \left(\frac{\Delta y_k}{\Delta t_k}\right)^2 + \left(\frac{\Delta z_k}{\Delta t_k}\right)^2}\, \Delta t_k,$$

which is (7.232) brought up where we can see it, but this time our situation is different. We suppose that $|C|$ is *defined* by (7.25), and we want to *prove* that the right side is an integral. For present purposes, the work of the preceding paragraph would be slovenly because nothing was proved. We could try to fix everything by trying to prove that there is a t_k^* such that, with the notation of the preceding paragraph, the kth term of (7.232) is exactly equal to $f(t_k^*)\, \Delta t_k$. We do not try this, however, because of the fear that the process would involve only uninformative hard work. Instead, we start by applying the mean-value theorem (Theorem 5.52) to (7.25). Since x' is continuous, there must be a t_k^* for which $t_{k-1} < t_k^* < t_k$ and

$$(7.251) \qquad \frac{\Delta x_k}{\Delta t_k} = \frac{x(t_k) - x(t_{k-1})}{t_k - t_{k-1}} = x'(t_k^*).$$

Similar formulas involving y and z end with $y'(t_k^{**})$ and $z'(t_k^{***})$, there being no reason for hope that the three numbers t_k^*, t_k^{**}, and t_k^{***} are the same. Thus

$$(7.252)\qquad |C| = \lim \Sigma \sqrt{[x'(t_k^*)]^2 + [y'(t_k^{**})]^2 + [z'(t_k^{***})]^2}\,\Delta t_k.$$

This looks much like a Riemann sum formed for the function f having values

$$(7.253)\qquad f(t) = \sqrt{[x'(t)]^2 + [y'(t)]^2 + [z'(t)]^2},$$

and there is no difficulty in the important special case in which $x(t) = t$ and $z(t) = 0$ for each t. Except in special cases, we must consider consequences of the fact that t_k^*, t_k^{**}, and t_k^{***} may be different. Since difficulties of this nature (and they were real difficulties in the old days) are usually associated with the name of Duhamel (1797–1872), we can call this a Duhamel difficulty. To treat all coordinates alike and to be precise about this matter, we let T_k be the center of the interval $t_{k-1} \leqq t \leqq t_k$. We can then put (7.252) in the form

$$(7.254)\qquad |C| = \lim_{|P|\to 0} \sum_{k=1}^{n} [f(T_k) + \delta_k]\,\Delta t_k,$$

where

$$(7.255)\qquad \delta_k = \sqrt{[x'(t_k^*)]^2 + [y'(t_k^{**})]^2 + [z'(t_k^{***})]^2} - \sqrt{[x'(T_k)]^2 + [y'(T_k)]^2 + [z'(T_k)]^2}.$$

To prove the desired result

$$(7.26)\qquad |C| = \int_a^b f(t)\,dt = \int_a^b \sqrt{[x'(t)]^2 + [y'(t)]^2 + [z'(t)]^2}\,dt,$$

it is therefore sufficient (and also necessary) to prove that

$$(7.261)\qquad \lim_{|P|\to 0} \sum_{k=1}^{n} \delta_k\,\Delta t_k = 0.$$

It is easy to originate the correct idea that the numbers t_k^*, t_k^{**}, and t_k^{***} are all near T_k, that continuity of the functions x', y', z' implies that the two terms on the right side of (7.255) are nearly equal, that the numbers δ_k are all near 0, and hence that the sum in (7.261) is near 0, whenever the norm $|P|$ of P is small. The easy way to make this precise is to use the fact that the first term in the right member of (7.255) is a continuous function of three variables. Using only functions of one variable, we can let $0 < \epsilon' < \epsilon$ and choose a number $\delta > 0$ such that the numbers α, β, and γ defined by

$$(7.262)\qquad [x'(u)]^2 - [x'(v)]^2 = \alpha, \qquad [y'(u)]^2 - [y'(v)]^2 = \beta, \qquad [z'(u)]^2 - [z'(v)]^2 = \gamma$$

all have absolute values less than ϵ' when $|u - v| < \delta$. When $|P| < \delta$, the right member of (7.255) then has the form

$$\delta_k = \sqrt{A_k \pm h} - \sqrt{A_k}, \tag{7.263}$$

where the quantities under the radicals are nonnegative and $0 \leqq h \leqq 3\epsilon'$. It follows that $|\delta_k| \leqq \sqrt{3\epsilon'}$ and hence

$$\left| \sum_{k=1}^{n} \delta_k \, \Delta t_k \right| \leqq \sum_{k=1}^{n} \sqrt{3\epsilon'} \, \Delta t_k = \sqrt{3\epsilon'}\,(b - a) < \epsilon \tag{7.264}$$

provided ϵ' is so chosen that the last inequality holds. This proves (7.261) and hence (7.26).

In order to introduce coordinates on curves, we suppose that x, y, and z are functions having continuous derivatives over an open interval $a < t < b$ and that C is the curve traversed by the point $P(t)$ having coordinates $(x(t), y(t), z(t))$ as t increases. To avoid difficulties, we suppose that the curve is *simple;* this means that $P(t_1) \neq P(t_2)$ when $t_1 \neq t_2$ so that $P(t)$ cannot be in the same place at two different times. Let t_0 be fixed such that $a < t_0 < b$. Then, as in Figure 7.27, we can assign coordinates to points on C by letting s be the length of the curve running from $P(t_0)$ to $P(t)$ when $t_0 \leqq t < b$ and letting $-s$ be the length of the curve running from $P(t)$ to $P(t_0)$ when $a < t \leqq t_0$. Then, whenever $a < t < b$, the coordinate $s(t)$ of P at time t is

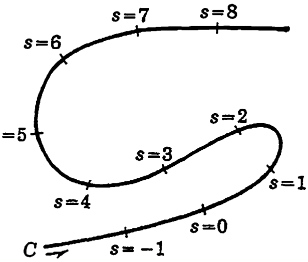

Figure 7.27

$$s(t) = \int_0^t \sqrt{[x'(\tau)]^2 + [y'(\tau)]^2 + [z'(\tau)]^2}\, d\tau. \tag{7.28}$$

In this formula τ (tau, the Greek t) is used as a dummy variable of integration. Since our hypothesis implies that the integrand is continuous, Theorem 4.35 enables us to differentiate with respect to t to obtain the formula

$$s'(t) = \sqrt{[x'(t)]^2 + [y'(t)]^2 + [z'(t)]^2}, \tag{7.281}$$

which is often written in the form

$$\frac{ds}{dt} = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2}. \tag{7.282}$$

This is related to the formulas we get when we set

$$\mathbf{r} = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k} \tag{7.283}$$

and calculate the velocity $\mathbf{v}$ and the speed $|\mathbf{v}|$ from the formulas

$$\mathbf{v} = x'(t)\mathbf{i} + y'(t)\mathbf{j} + z'(t)\mathbf{k} \tag{7.284}$$

and

$$|\mathbf{v}| = \sqrt{[x'(t)]^2 + [y'(t)]^2 + [z'(t)]^2}. \tag{7.285}$$

Since the right members of (7.282) and (7.285) are the same, this shows that, in appropriate circumstances, the speed determined so quickly by taking the absolute value of the velocity is the same as the "speed of the particle in its path" determined by use of coordinates on the path.

Problems 7.29

1 Find the length of the part of the helix

$$x = a\cos t, \qquad y = a\sin t, \qquad z = kt$$

traversed by a particle at $P(x,y,z)$ as t increases from 0 to 2π.

2 Using the first standard equation in

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \qquad x = a\cos\theta, \qquad y = b\sin\theta,$$

where $0 < b < a$, set up an integral for the length of the part of the ellipse lying in the first quadrant. Then use the parametric equations to set up another integral for the same length

Ans.:
$$\int_0^a \sqrt{\frac{a^2 - k^2x^2}{a^2 - x^2}}\,dx, \qquad a\int_0^{\pi/2} \sqrt{1 - k^2\cos^2\theta}\,d\theta,$$

where $k^2 = (a^2 - b^2)/a^2$. These are *elliptic integrals*, and others more or less like them are also called elliptic integrals. There is no easy way to obtain their exact numerical values.

3 Set up an integral for the length L of the curve traced by the point P having coordinates

$$x = a\theta - b\sin\theta, \qquad y = a - b\cos\theta$$

as θ increases from 0 to 2π, and show that the result can be put in the form

$$L = (a + b)\int_0^{2\pi} \sqrt{1 - \frac{4ab}{(a + b)^2}\cos^2\frac{\theta}{2}}\,d\theta.$$

When $b \neq a$, this is an elliptic integral. When $b = a$, the graph is, as Problem 16 of Section 6.5 shows, an ordinary cycloid. Sketch a figure and explain a simple geometric argument which shows that, when $b = a$ and the approximations $\pi = 3$ and $\pi^2 = 10$ are used, $\sqrt{52}\,a < L < 10a$. Finally, show that $L = 8a$.

4 Find the length of the part of the curve having the equations

$$x = 2t^3, \qquad y = \frac{t^2 - 3}{2}, \qquad z = \frac{2 - t^2}{2}$$

which lies between the points $(0, -\frac{3}{2}, 1)$ and $(2, -1, \frac{1}{2})$. *Ans.:* $L = \frac{\sqrt{2}}{54}[19^{3/2} - 1]$.

5 When $0 < b < a$ and a circle (or cog wheel) of radius b rolls without slipping inside a circle of radius a as in Figure 7.291, the path traced by a point on the small rolling circle is a hypocycloid (inside-cycloid) that appeared in one of Problems 6.59 and that even small dictionaries describe. It can be shown that when the center of the small circle has passed over an angle θ (that could be ωt), the point initially at $(a,0)$ has coordinates

$$x = (a - b)\cos\theta + b\cos\frac{a-b}{b}\theta$$

$$y = (a - b)\sin\theta - b\sin\frac{a-b}{b}\theta.$$

With the aid of Figure 7.291 find the amount by which θ must increase as P traverses one arch of the cycloid from the big circle back to the big circle. Then find the length of one arch of the cycloid. *Ans.:* $8b(a - b)/a$.

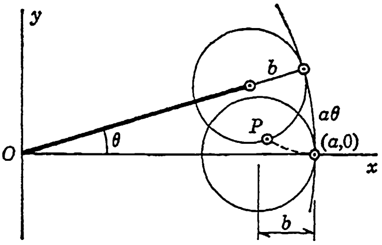

Figure 7.291

6 It is a remarkable fact that when $b = a/4$, so that the hypocycloid of the preceding problem has four cusps, the equations can be put in the form

$$x = \frac{a}{4}(3\cos\theta + \cos 3\theta) = a\cos^3\theta$$

$$y = \frac{a}{4}(3\sin\theta - \sin 3\theta) = a\sin^3\theta$$

so that

$$x^{2/3} + y^{2/3} = a^{2/3}.$$

Using this formula, find the length of the part of the hypocycloid in the first quadrant. *Ans.:* $3a/2$.

7 Let f and g have continuous derivatives over $a \leqq u \leqq b$. Let C be the curve which the point $P(f(u), g(u))$ traverses in an xy plane as u increases from a to b. Let C be regarded as a wire having linear density δ so that a piece of C of length L has mass δL. Let p be a nonnegative integer and let x_0 be a constant. Set up an integral for the pth moment of the wire about the line having the equation $x = x_0$.

8 With or without assistance from the preceding problem, set up an integral for the pth moment about a diameter of a circular wire having linear density δ.

9 Centroids of triangles, rectangles, and regular polygons coincide with the centroids of the regions that they bound. Draw some polygons for which there is violent departure from coincidence.

Figure 7.292

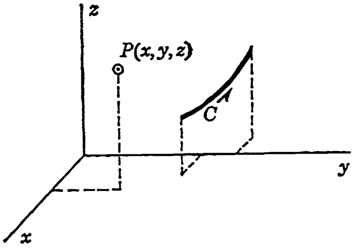

10 Figure 7.292 is intended to steer our thoughts toward a wire or cord concentrated on a curve C and to make us realize that we have not yet calculated the gravitational force $\mathbf{F}$ upon a particle P^* of mass m at $P(x,y,z)$ that is produced by the wire. Set up an integral for $\mathbf{F}$. *Outline of solution:* Since x, y, z have been preempted, we take the curve C to be the ordered set of points $P(t) = P(u,v,w)$ for which u, v, w are functions of t having continuous derivatives and

$$\overrightarrow{OP} = u(t)\mathbf{i} + v(t)\mathbf{j} + w(t)\mathbf{k} \qquad (a \leqq t \leqq b), \tag{1}$$

and suppose that the linear density (mass per unit length) of C at $P(t)$ is $\delta(t)$, where δ is a continuous function of t. Letting Q be a partition of the interval $a \leqq t \leqq b$ with partition points t_k and supposing that $t_{k-1} \leqq t_k^* \leqq t_k$, we take the number $\delta(t_k^*)\ |\overrightarrow{P_{k-1}P_k}|$ as an approximation to the mass of an element of the wire. The force $\Delta\mathbf{F}_k$ which this element produces upon the particle P^* of mass m at $P(x,y,z)$ should have approximate magnitude

$$\frac{Gm\delta(t_k^*)|\overrightarrow{P_{k-1}P_k}|}{|\overrightarrow{PP}(t_k^*)|^2}. \tag{2}$$

Introducing vectors and coordinates gives the approximation

$$\Delta\mathbf{F}_k = Gm\delta(t_k^*)\sqrt{\left(\frac{\Delta u_k}{\Delta t}\right)^2 + \left(\frac{\Delta v_k}{\Delta t}\right)^2 + \left(\frac{\Delta w_k}{\Delta t}\right)^2} \\ \frac{[u(t_k^*) - x]\mathbf{i} + [v(t_k^*) - y]\mathbf{j} + [w(t_k^*) - z)]\mathbf{k}}{\{[u(t_k^*) - x]^2 + [v(t_k^*) - y]^2 + [w(t_k^*) - z]^2\}^{3/2}}\Delta t. \tag{3}$$

The limit of the sum of these things should be $\mathbf{F}$, and we are led to the definition

$$\mathbf{F} = Gm\int_a^b \delta(t)\ \sqrt{[u'(t)]^2 + [v'(t)]^2 + [w'(t)]^2} \\ \frac{[u(t) - x]\mathbf{i} + [v(t) - y]\mathbf{j} + [w(t) - z]\mathbf{k}}{\{[u(t) - x]^2 + [v(t) - y]^2 + [w(t) - z]^2\}^{3/2}}\,dt. \tag{4}$$

This is our result. In case the wire coincides with a circle of radius a in the yz plane having its center at the origin, we can set $u(t) = 0$, $v(t) = a\cos t$, $w(t) = a\sin t$ and put (4) in the form

$$\mathbf{F} = Gma\int_0^{2\pi} \delta(t)\,\frac{-x\mathbf{i} + [a\cos t - y]\mathbf{j} + [a\sin t - z]\mathbf{k}}{\{x^2 + [a\cos t - y]^2 + [a\sin t - z]^2\}^{3/2}}\,dt. \tag{5}$$

In case P^* is on the axis of the wire so that $y = z = 0$, this takes the much simpler form

$$\mathbf{F} = Gma\int_0^{2\pi} \delta(t)\,\frac{-x\mathbf{i} + a\cos t\mathbf{j} + a\sin t\mathbf{k}}{(x^2 + a^2)^{3/2}}\,dt. \tag{6}$$

In case the wire is a uniform wire so that, for some constant δ_0 we have $\delta(t) = \delta_0$ for each t, the coefficients of $\mathbf{j}$ and $\mathbf{k}$ are zero because

$$\int_0^{2\pi} \cos t\,dt = \int_0^{2\pi} \sin t\,dt = 0 \tag{7}$$

(or, as is usually said, "because of symmetry") and (6) reduces to

$$\mathbf{F} = -\frac{GmMx}{(x^2 + a^2)^{3/2}}\mathbf{i}, \tag{8}$$

where $M = 2\pi a\delta_0$, the total mass of the wire. *Remark:* With slight differences in notation, (8) was used extensively in some of the Problems 4.79.

11 Let units of measurement be so employed that the potential at $P_0(x_0,y_0,z_0)$ due to a particle of mass m concentrated at a point $P(x,y,z)$ is $m/|\overline{P_0P}|^2$. Let x, y, z have continuous derivatives over $a \leqq u \leqq b$, and let $P(x,y,z)$ traverse a curve C as u increases from a to b. Let C have linear density δ. Let $P_0(x_0,y_0,z_0)$ be a point not on C. Remembering that potentials are scalars, and remembering or learning quickly that the potential due to a collection of particles is equal to the sum of the potentials due to the individual particles, set up an integral which is equal to the potential at P_0 due to C.

12 Set up an integral equal to the moment of inertia (second moment) about the z axis of the curve C of the preceding problem.

13 When a particle has fallen (or slid without friction) from height $G(b)$ to a lower height $G(u)$, its speed is $\sqrt{2g[G(b) - G(u)]}$, where g is the acceleration of gravity. We all know that if we travel a short distance with a constant speed, we get the time required for the trip by dividing the distance by the speed. With this basic information, we are prepared to attack a problem. Let F and G be functions having positive continuous derivatives over $0 \leqq u \leqq b$ and let the point $P(F(u), G(u))$ traverse a curve C from the origin to a point B in the first quadrant of an xy plane as u increases from 0 to b. Sketch a schematic figure showing the curve C. Set up an integral which gives the time T required for a particle starting from rest at B to slide without friction down the curve C to the origin.

14 The reflecting surface of a headlight is a part of a paraboloid, of depth 4 inches and diameter 12 inches, obtained by rotating a part of a parabola about its axis. Find its area. *Solution:* This is not an easy problem, since a basic difficulty lurks in the fact that areas of such surfaces have not been defined. We start by so determining k that the parabola having the equation $y = kx^2$ contains (or passes through) the point (6,4). This gives $y = \frac{1}{9}x^2$ for the equation of the generating parabola. With the possibility of setting $a = 0$, $b = 6$, and $f(x) = \frac{1}{9}x^2$, we suppose that f has a continuous derivative over an interval $a \leqq x \leqq b$. Let S be the surface generated by rotating, about the y axis, the part of the graph G of $y = f(x)$ for which $a \leqq x \leqq b$. Expecting to use some intuitive ideas about areas, we make a partition P of the interval $a \leqq x \leqq b$ and consider one particular subinterval $x_{k-1} \leqq x \leqq x_k$. As an approximation to the length L_k of the segment G_k of the graph G containing points x for which $x_{k-1} \leqq x \leqq x_k$, we may use the number

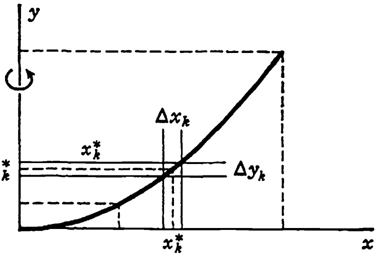

Figure 7.293

$$(1) \qquad \sqrt{\Delta x_k^2 + \Delta y_k^2} = \sqrt{1 + \left(\frac{\Delta y_k}{\Delta x_k}\right)^2}\,\Delta x_k = \sqrt{1 + [f'(x_k^*)]^2}\,\Delta x_k,$$

where $\Delta x_k = x_k - x_{k-1}$, $\Delta y_k = f(x_k) - f(x_{k-1})$, and x_k^* is a number for which $x_{k-1} < x_k^* < x_k$. Figure 7.293 is helpful. Even though areas are not yet defined. we can have a feeling that when the segment G_k is rotated about the y axis, its points all travel approximately the same distance $2\pi x_k$ and that the segment generates a surface essentially like a ribbon having width L_k and length $2\pi x_k$. This leads us to the intuitive idea that one term in the sum

$$|S| = \lim \Sigma 2\pi \sqrt{1 + [f'(x_k)]^2}\, x_k\, \Delta x_k \tag{2}$$

should, when the norm $|P|$ of P is small, be a good approximation to the area of the part of S generated by one segment of the graph G. The next step is to adopt the tentative intuitive conclusion that the sum is a good approximation to the total area $|S|$ of S when the norm of P is small or, in other words, that (2) should be valid. Our theory of Riemann sums and integrals assures us that if (2) is true, then

$$|S| = 2\pi \int_a^b \sqrt{1 + [f'(x)]^2}\, x\, dx. \tag{3}$$

If we have sufficient confidence in our calculations (it would, of course, be fatal to use an incorrect formula for the circumference of a circle of radius x_k) and in our intuitive ideas, we can install the formula (3) as one (not *the*) definition of area of surfaces of revolution. It does not make sense to claim that this definition is "correct," but experience shows that it is useful and this is all that we can expect from definitions. For the case in which $f(x) = kx^2$, $a = 0$, and $b = r$, (3) reduces to

$$|S| = 2\pi \int_0^r (1 + 4k^2x^2)^{1/2} x\, dx = \frac{\pi}{6k^2}[(1 + 4k^2r^2)^{3/2} - 1].$$

When $k = \frac{1}{8}$ and $r = 6$, this reduces to $|S| = 49\pi$.

15 We invest a little time to look at some *curve integrals* that are called *line integrals* by those who adhere to the notion that curves are lines. Let functions $x(t)$, $y(t)$, $z(t)$, the point $P(t)$, the vector $\mathbf{r}(t)$, the curve C, the partition P of $a \leqq t \leqq b$, and the numbers Δt_k, Δx_k, Δy_k, Δz_k be defined as in the part of this section preceding (7.231). Let a vector function $\mathbf{F}$ having scalar components f_1, f_2, f_3 be defined and continuous over a part of E_3 that contains the curve C. We consider $\mathbf{F}$ to be a force which operates upon a particle at $P(t)$ as the particle moves from $P(a)$ to $P(b)$. The schematic Figure 7.294 may be helpful. With the idea that the work done by $\mathbf{F}$ as the particle moves from $P(t_{k-1})$ to $P(t_k)$ is closely approximated by

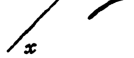

Figure 7.294

$$\mathbf{F}(x_k,y_k,z_k)\cdot\Delta\mathbf{r}_k \tag{1}$$

when the norm of the partition P is small, we can define the total work W done by $\mathbf{F}$ by the formula

$$W = \lim \Sigma \mathbf{F}(x_k,y_k,z_k)\cdot\Delta\mathbf{r}_k. \tag{2}$$

The right member of (2) is an example of a *curve integral*, and it is denoted by the symbol in

$$W = \int_C \mathbf{F}(\mathbf{r})\cdot d\mathbf{r}, \tag{3}$$

where $\mathbf{F}(\mathbf{r})$ is the vector function of the vector $\mathbf{r}$ defined by

$$\mathbf{F}(\mathbf{r}) = \mathbf{F}(x,y,z) \tag{4}$$

in which x, y, z are the numbers for which $\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$. The right member of (3) is read "the integral over C of $\mathbf{F}(\mathbf{r})$ dot $d\mathbf{r}$," the fundamental idea being that it is the curve C that is partitioned to produce the approximating sums. To learn something about this curve integral, we can use the formulas

$$\mathbf{F}(x,y,z) = f_1(x,y,z)\mathbf{i} + f_2(x,y,z)\mathbf{j} + f_3(x,y,z)\mathbf{k} \tag{5}$$

$$\Delta\mathbf{r}_k = \Delta x_k\mathbf{i} + \Delta y_k\mathbf{j} + \Delta z_k\mathbf{k} \tag{6}$$

to put (2) in the form

$$W = \lim \sum \left[f_1(x(t_k),\, y(t_k),\, z(t_k))\,\frac{\Delta x_k}{\Delta t_k} + f_2(x(t_k),\, y(t_k),\, z(t_k))\,\frac{\Delta y_k}{\Delta t_k} + f_3(x(t_k),\, y(t_k),\, z(t_k))\,\frac{\Delta z_k}{\Delta t_k}\right]\Delta t_k. \tag{7}$$

A proof very similar to the one centering around (7.255) enables us to show that

$$W = \int_a^b f_1(x(t),\, y(t),\, z(t))x'(t)\, dt + \int_a^b f_2(x(t),\, y(t),\, z(t))y'(t)\, dt + \int_a^b f_3(x(t),\, y(t),\, z(t))z'(t)\, dt. \tag{8}$$

This is a formula from which W can be calculated or approximated. Since we are not electronic computers, we do not (at least at the present time) make calculations, but we do point out that the integrals in (8) are abbreviated to those in the formula

$$W = \int_C [f_1(x,y,z)\, dx + f_2(x,y,z)\, dy + f_3(x,y,z)\, dz]. \tag{9}$$

The integrals in (9) are scalar curve integrals. Whenever we want to know what these things mean and how they can be evaluated, we should have the wits to check back to see what they abbreviate. While it may be possible to overemphasize the importance of the matter, we can observe that if we set

$$\mathbf{F}(\mathbf{r}) = f_1(x,y,z)\mathbf{i} + f_2(x,y,z)\mathbf{j} + f_3(x,y,z)\mathbf{k} \tag{10}$$

and make the pretense that $d\mathbf{r}$ is a vector for which

$$d\mathbf{r} = dx\,\mathbf{i} + dy\,\mathbf{j} + dz\,\mathbf{k} \tag{11}$$

and

$$\mathbf{F}(\mathbf{r})\cdot d\mathbf{r} = f_1(x,y,z)\, dx + f_2(x,y,z)\, dy + f_3(x,y,z)\, dz, \tag{12}$$

then we can substitute (12) into (3) to obtain (9). Of course, it should be thoroughly understood that we have done nothing but explain symbols. This would be a stupid waste of time if it were not for the fact the symbols are useful. They appear even in quite elementary physics and engineering, and mathematicians have a responsibility to tell what they mean.

7.3 *Center and radius of curvature* Courses and textbooks in "differential geometry" provide information about matters relating to curvature of curves that lie in E_3. In this section, we give most of our attention to curves that lie in an xy plane. Our approach to the subject sacrifices brevity to place emphasis upon elementary geometric ideas that can be of interest to everyone and are needed by engineers and others who study the bending of beams.

Let x and y, or $x(t)$ and $y(t)$, have continuous second derivatives over some open interval $a < t < b$ in which t and $t + \Delta t$ are always supposed to lie. Let $P(t)$ denote the point with coordinates $(x(t), y(t))$ and let $\mathbf{r}(t)$ be the vector running from the origin to $P(t)$ so that

$$\begin{aligned} &(7.311) & \mathbf{r}(t) &= x(t)\mathbf{i} + y(t)\mathbf{j} \\ &(7.312) & \mathbf{v}(t) &= x'(t)\mathbf{i} + y'(t)\mathbf{j} \\ &(7.313) & \mathbf{a}(t) &= x''(t)\mathbf{i} + y''(t)\mathbf{j}. \end{aligned}$$

Let C be the curve traversed by $P(t)$ and the tip of $\mathbf{r}(t)$ as t increases, so that $\mathbf{v}(t)$ is tangent to C at $P(t)$ when $\mathbf{v}(t) \neq 0$. Henceforth we consider only values of t for which

$$(7.32)\quad \mathbf{v}(t) \times \mathbf{a}(t) = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ x'(t) & y'(t) & 0 \\ x''(t) & y''(t) & 0 \end{vmatrix} = [x'(t)y''(t) - x''(t)y'(t)]\mathbf{k} \neq 0.$$

Since $|\mathbf{v} \times \mathbf{a}| = |\mathbf{v}|\,|\mathbf{a}|\,|\sin\theta|$, this means that $\mathbf{v}$ and $\mathbf{a}$ are nonzero vectors which do not lie on parallel lines. This and the first of the formulas

$$(7.321)\quad \lim_{\Delta t \to 0} \frac{\mathbf{v}(t + \Delta t) - \mathbf{v}(t)}{\Delta t} = \mathbf{a}(t), \qquad \mathbf{v}(t) \times \frac{\mathbf{v}(t + \Delta t) - \mathbf{v}(t)}{\Delta t} \neq 0$$

implies existence of a $\delta > 0$ such that the second holds when $0 < |\Delta t| < \delta$. Since $\mathbf{v}(t) \times \mathbf{v}(t) = 0$, we conclude that $\mathbf{v}(t) \times \mathbf{v}(t + \Delta t) \neq 0$ and hence that the tangents $\mathbf{v}(t)$ and $\mathbf{v}(t + \Delta t)$ do not lie on parallel lines when $0 < |\Delta t| < \delta$.

Figure 7.33

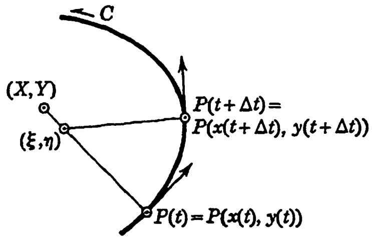

Supposing that $0 < |\Delta t| < \delta$, we construct Figure 7.33 and look at it. Since the tangents $\mathbf{v}(t)$ and $\mathbf{v}(t + \Delta t)$ to C at $P(t)$ and $P(t + \Delta t)$ do not lie on parallel lines, the normals to C at these points must intersect at a point (ξ,η) having coordinates ξ (xi) and η (eta). If C is a circle, then the most elementary of elementary geometry tells us that the points (ξ,η) obtained for different values of Δt would all lie at the center (X,Y) of the circle. This is no time to be devoid of ideas. We can have at least a vague feeling that, even when our curve C is not a circle, a small section of C in a neighborhood of $P(t)$ should be so much like a circle that there should be a point (X,Y) such that (ξ,η) is near (X,Y) whenever Δt is small. If all this happens, we should give the point (X,Y) a name and find formulas for X and Y. All this does happen. The point is called the *center of curvature* of C at $P(t)$. We shall find formu'as for X and Y and for the distance from (X,Y) to $P(t)$. This distance is called the *radius of curvature* ρ (rho) of C at $P(t)$. The circle with center at (X,Y) and radius ρ is called the *circle of curvature* of C at $P(t)$.

The coordinates ξ and η are determined by the system of equations

$$\begin{aligned}[]
[\xi - x(t)]x'(t) \quad &+ \quad [\eta - y(t)]y'(t) = 0\\
[\xi - x(t+\Delta t)]x'(t+\Delta t) &+ [\eta - y(t+\Delta t)]y'(t+\Delta t) = 0.
\end{aligned}$$

The left member of the first equation is the scalar product of the vector $\mathbf{v}(t)$ tangent to C at $P(t)$ and the vector running from $P(t)$ to (ξ,η), and the equation expresses the fact that the two vectors are orthogonal. Similarly, the second equation expresses the fact that (ξ,η) lies on the normal to C at $P(t + \Delta t)$. Replacing the quantities in brackets in the second equation by

$$[\xi - x(t) - x(t+\Delta t) + x(t)] \qquad \text{and} \qquad [\eta - y(t) - y(t+\Delta t) + y(t)]$$

and transposing a part of the result enables us to put the two equations in the form

$$[\xi - x(t)]x'(t) + [\eta - y(t)]y'(t) = 0 \tag{7.341}$$

$$[\xi - x(t)]x'(t+\Delta t) + [\eta - y(t)]y'(t+\Delta t) = Q, \tag{7.342}$$

where we have simplified matters by letting Q denote the quantity

$$Q = [x(t+\Delta t) - x(t)]x'(t+\Delta t) + [y(t+\Delta t) - y(t)]y'(t+\Delta t). \tag{7.343}$$

To eliminate η from the two equations (7.341) and (7.342), we multiply the first by $y'(t+\Delta t)$ and the second by $-y'(t)$ and add to obtain the first of the formulas

$$D[\xi - x(t)] = -Qy'(t), \qquad D[\eta - y(t)] = Qx'(t), \tag{7.35}$$

where D is the determinant

$$D = x'(t)y'(t+\Delta t) - x'(t+\Delta t)y'(t)$$

which can be put in the form

$$D = x'(t)[y'(t+\Delta t) - y'(t)] - [x'(t+\Delta t) - x'(t)]y'(t). \tag{7.351}$$

A similar procedure, in which the multipliers $-x'(t+\Delta t)$ and $x'(t)$ are used, gives the second of the formulas (7.35). Formulas for ξ and η are now obtained from (7.35) by dividing and transposing. As we shall see, formulas for X and Y are obtained by dividing (7.35) by Δt and taking limits as $\Delta t \to 0$. Since (7.343) and (7.351) show that

$$\lim_{\Delta t\to 0} \frac{Q}{\Delta t} = [x'(t)]^2 + [y'(t)]^2$$

$$\lim_{\Delta t\to 0} \frac{D}{\Delta t} = x'(t)y''(t) - x''(t)y'(t),$$

we find from (7.35) that

$$(7.361) \quad X - x(t) = \lim_{\Delta t\to 0} [\xi - x(t)] = -\frac{[x'(t)]^2 + [y'(t)]^2}{x'(t)y''(t) - x''(t)y'(t)}\, y'(t)$$

$$(7.362) \quad Y - y(t) = \lim_{\Delta t\to 0} [\eta - y(t)] = \frac{[x'(t)]^2 + [y'(t)]^2}{x'(t)y''(t) - x''(t)y'(t)}\, x'(t).$$

Transposing the terms $x(t)$ and $y(t)$ gives formulas

$$(7.363) \qquad X = x(t) - \frac{[x'(t)]^2 + [y'(t)]^2}{x'(t)y''(t) - x''(t)y'(t)}\, y'(t)$$

$$(7.364) \qquad Y = y(t) + \frac{[x'(t)]^2 + [y'(t)]^2}{x'(t)y''(t) - x''(t)y'(t)}\, x'(t)$$

for the coordinates (X,Y) of the center of curvature of C at the point $(x(t), y(t))$, but these new formulas are sometimes less useful than their parents. The definition of radius of curvature ρ implies that

$$\rho = \sqrt{[X - x(t)]^2 + [Y - y(t)]^2}$$

and hence

$$(7.37) \qquad \rho = \frac{\{[x'(t)]^2 + [y'(t)]^2\}^{3/2}}{|x'(t)y''(t) - x''(t)y'(t)|}.$$

In case $x(t) = t$ so $x'(t) = 1$ and $x''(t) = 0$, it is customary to replace t by x and write

$$(7.371) \qquad \rho = \frac{\{1 + [y'(x)]^2\}^{3/2}}{|y''(x)|} = \frac{\left[1 + \left(\frac{dy}{dx}\right)^2\right]^{3/2}}{\left|\frac{d^2y}{dx^2}\right|}.$$

To steer our thoughts toward another problem involving curvature, we consider the rate of change of direction of an automobile which traverses a level road that winds around swamps and between mountains. At each time t the rate depends upon the speed of the automobile and upon another number which is called the curvature of the road at the position of the automobile. To be more precise about this matter, let

x and y be functions of t having two derivatives each over some interval and let C be the curve traversed by the point $P(t)$, having coordinates $x(t)$, $y(t)$, as t increases over the interval. We suppose that there is no t for which $x'(t)$ and $y'(t)$ are both 0. Our results will be obtained with the aid of the three familiar vectors

$$\mathbf{r}(t) = x(t)\mathbf{i} + y(t)\mathbf{j} \tag{7.381}$$
$$\mathbf{v}(t) = x'(t)\mathbf{i} + y'(t)\mathbf{j} \tag{7.382}$$
$$\mathbf{a}(t) = x''(t)\mathbf{i} + y''(t)\mathbf{j}. \tag{7.383}$$

The vector $\mathbf{v}(t)$ having its tail at $P(t)$ is the forward tangent to C at $P(t)$. With each t we wish to associate an angle $\phi(t)$ that determines the direction of $\mathbf{v}(t)$, and this is a matter that must not be treated carelessly. To restrict $\phi(t)$ to an interval like $-\pi < \phi \leqq \pi$, so that ϕ would be discontinuous when the vector rotates from northwest to southwest, would defeat our purpose. To formulate general principles by which $\phi(t)$ can be calculated may be beyond our capabilities. Let us then avoid possible unforeseen topological difficulties by restricting attention to curves C for which it is clearly possible to determine $\phi(t)$ by the following procedure. Let t_0 be a particular t in the interval considered and let $\phi(t_0)$ be the angle for which $-\pi < \phi(t_0) \leqq \pi$ and $\phi(t_0)$ is the ordinary trigonometric angle having its initial side on the positive x axis and its terminal side on the vector through the origin parallel to $\mathbf{v}(t_0)$. Then, as t increases (or decreases) from t_0, let $\phi(t)$ vary with the vector $\mathbf{v}(t)$ in such a way that ϕ is continuous. Some curves are less complicated than that shown in Figure 7.384, and some are more complicated. Supposing that $\phi(t)$ is satisfactorily determined, we can put the formula (7.382) for $\mathbf{v}(t)$ in the form

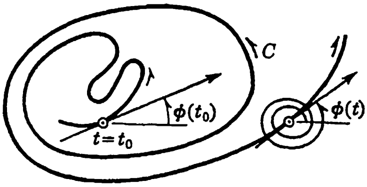

Figure 7.384

$$|\mathbf{v}(t)|[\cos\phi(t)\mathbf{i} + \sin\phi(t)\mathbf{j}] = \mathbf{v}(t), \tag{7.385}$$

where

$$x'(t) = |\mathbf{v}(t)|\cos\phi(t), \qquad y'(t) = |\mathbf{v}(t)|\sin\phi(t) \tag{7.3851}$$

and

$$|\mathbf{v}(t)| = \{[x'(t)]^2 + [y'(t)]^2\}^{1/2} \neq 0. \tag{7.3852}$$

Since we want a formula for $\phi'(t)$, it is reasonable to differentiate (7.385) with respect to t and to try to use the result. We divide by $|\mathbf{v}(t)|$ and then differentiate to obtain

$$[-\sin\phi(t)\mathbf{i} + \cos\phi(t)\mathbf{j}]\phi'(t) = \frac{|\mathbf{v}(t)|\mathbf{a}(t) - \mathbf{v}(t)\dfrac{d}{dt}|\mathbf{v}(t)|}{|\mathbf{v}(t)|^2}. \tag{7.386}$$

The coefficient of $\phi'(t)$ is the unit vector $\mathbf{n}$ for which

$$\mathbf{n} = [-\sin\phi(t)\mathbf{i} + \cos\phi(t)\mathbf{j}] = \frac{-y'(t)\mathbf{i} + x'(t)\mathbf{j}}{|\mathbf{v}(t)|}. \tag{7.3861}$$

Since $\mathbf{n}\cdot\mathbf{v}(t) = 0$, we can equate the scalar products of $\mathbf{n}$ and the members of (7.386) to obtain

$$\phi'(t) = \frac{|\mathbf{v}(t)|\mathbf{a}(t)\cdot\mathbf{n}(t)}{|\mathbf{v}(t)|^2} = \frac{[x''(t)\mathbf{i} + y''(t)\mathbf{j}]\cdot[-y'(t)\mathbf{i} + x'(t)\mathbf{j}]}{|\mathbf{v}(t)|^2}$$

and hence

$$\phi'(t) = \frac{x'(t)y''(t) - x''(t)y'(t)}{[x'(t)]^2 + [y'(t)]^2}. \tag{7.387}$$

For some applications of (7.387), we can set $x(t) = t$ so that $x'(t) = 1$ and $x''(t) = 0$. In such cases, we can replace t by x to obtain

$$\phi'(x) = \frac{y''(x)}{1 + [y\ (x)]^2}. \tag{7.3871}$$

In this, and in any other case in which $x'(t) \neq 0$, we can eliminate practically all of the work of this section by writing

$$\tan\phi = \frac{y'(t)}{x'(t)} \tag{7.3872}$$

with or without the aid of (7.3851) and then differentiating with respect to t to obtain (7.387).

Returning to (7.387), we note that $\phi'(t)$, which might be measured in radians per minute, gives the time rate of change of ϕ with respect to t. In case the point $P(t)$ traverses C with unit speed, which might be 1 kilometer per minute, $\phi'(t)$ becomes also a number of radians per unit distance measured along C, and this is called the *curvature* K or $K(t)$ of C at $P(t)$. To give curvature an additional leg upon which to stand, we introduce upon C a coordinate system like that in Figure 7.27 with the stipulation that the coordinate s of $P(t)$ increases as t increases. The curvature of C at $P(t)$ can then be defined by the formula

$$K = K(t) = \frac{d\phi}{ds}. \tag{7.388}$$

Since $d\phi/ds = \phi'(t)/s'(t) = \phi'(t)/|v(t)|$, the formulas (7.387) and (7.3852) yield the formula

$$K = \frac{x'(t)y''(t) - x''(t)y'(t)}{\{[x'(t)]^2 + [y'(t)]^2\}^{3/2}} \tag{7.389}$$

from which the curvature of C at $P(t)$ can be calculated without reference to other formulas. Perhaps we should take notice of the fact that a curve C is an ordered set of points, and that we can misunderstand (7.389) when we forget this fact. In particular, the sign in (7.389) will be wrong if we put the coordinate system on C backwards so that s decreases as t increases.

We conclude with a fundamental observation. As formulas (7.37) and (7.389) show, *the radius of curvature ρ and the absolute value $|K|$ of the curvature K* are reciprocals wherever we have defined both of them.

Problems 7.39

1 By use of the equations

$$x = a \cos t, \qquad y = a \sin t,$$

show that the curvature of the curve C consisting of a circle of radius a traced in the *positive* direction is identically $1/a$. Then by use of the equations

$$x = a \cos t, \qquad y = -a \sin t,$$

show that the curvature of the curve Γ (capital gamma) consisting of a circle of radius a traced in the *negative* direction is identically $-1/a$.

2 Hindsight can be very good. Look at (7.388). Then run with constant speed and in the positive direction around a circle of radius a and observe that ϕ increases at a constant rate. Then reverse the direction of the run and observe that ϕ decreases at the same constant rate.

3 Determine the radius of curvature of (that is, at points of) the parabola having the equation $y = kx^2$. Find the minimum radius of curvature. Sketch a graph for the case in which $k = 1$ and determine whether the answer seems to be correct.

4 Show that the normals to the graph of $y = x^2$ at the points (0,0) and (0.01, 0.0001) intersect at the point (0, 0.5001). Show that this intersection is at distance 0.0001 from the center of curvature of the graph at the point (0,0).

5 A glance at the graph of $y = \log x$ suggests that the absolute value of the curvature is greatest and that the radius of curvature is least when x is somewhere between $\frac{1}{2}$ and 2. Find the x for which the radius of curvature attains its minimum value. *Ans.:* $\sqrt{2}/2 = 0.707$.

6 When a point $P(x,y)$ moves along an arc or curve C having equations $x = x(t)$, $y = y(t)$ that satisfy appropriate conditions, the center (X,Y) of curvature moves along an arc that is called the *evolute* of C. The formulas (7.363) and (7.364), which we should be able to use but need not remember, show how X

and Y depend upon t. Supposing P moves along the graph of $y = kx^2$ in such a way that $x = t$ and $y = kt^2$, show that the equations of the evolute are

$$X = -4k^2t^3, \qquad Y = \frac{1}{2k} + 3kt^2.$$

7 This problem involves a little story. A string is wound in a clockwise direction around a circular spool of radius a with an end at the point $(a,0)$. When the string is unwound, being kept stretched during the process, the end of the string traces the spiral curve C shown in Figure 7.391. When $a\theta$ units of string have been unwound, this part of the string is tangent to the spool at the point $Q(X,Y)$ for which

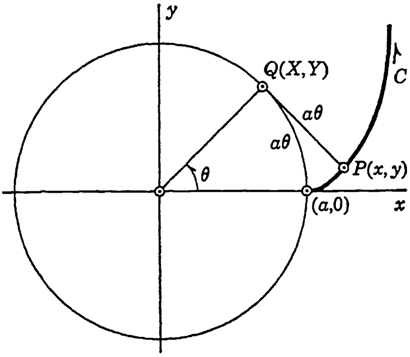

Figure 7.391

$$X = a\cos\theta, \qquad Y = a\sin\theta$$

and the end of the string is at the point $P(x,y)$ for which

$$\begin{aligned} x &= x(\theta) = a\cos\theta + a\theta\sin\theta \\ y &= y(\theta) = a\sin\theta - a\theta\cos\theta. \end{aligned}$$

It can be observed that Q moves around the circle just as rapidly as the distance from Q to P increases. It is not unreasonable to guess (or at least to consider the possibility) that the center and radius of curvature of C at P are Q and $a\theta$. On the other hand, a skeptic can be uncertain whether QP is perpendicular to the tangent to C at P. The situation demands clarification. Start with the equations of C and find the center and radius of curvature of C at P. *Ans.:* (X,Y) and $a\theta$. *Remark:* The curve C is called the *involute* of the spool. The spool is the evolute of its involute.

8 Determine the radius of curvature of the "standard ellipse"

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \qquad \text{or} \qquad x = a\cos\theta, \qquad y = b\sin\theta$$

by use of the first "standard equation" and then by use of the latter parametric equations.

9 Using the parametric equations of Problem 8, find the evolute of the standard ellipse.

10 As an alternative to (or in addition to) Problem 8, find the radius of curvature of the "standard hyperbola."

11 A glance at the hypocycloid having the equations

$$x = a\cos^3 t, \qquad y = a\sin^3 t$$

indicates that the radius of curvature ρ should be a maximum at points (x,y) for which $|x| = |y|$ and a minimum at points for which $xy = 0$. Can we believe our eyes? *Ans.:* No, because ρ has no minimum. When $t \neq n\pi/2$, calculations give $\rho = \frac{3a}{2}|\sin 2t|$. Thus ρ is a maximum when $2t = \left(n + \frac{1}{2}\right)\pi$ or $t = \left(\frac{n}{2} + \frac{1}{4}\right)\pi$ and hence when $|\sin t| = |\cos t|$ or $|x| = |y|$. We see that $\rho \to 0$ as $t \to 0$, but the curvature at the cusps is undefined.

12 When a flexible cord or chain (the Latin word for chain is *catena*) is suspended from its ends in a parallel force field, it hangs in a curve (or point set) called a *catenary*. Differential equations textbooks show that a rectangular coordinate system can be chosen in such a way that the equation of the catenary is

$$y = \frac{a}{2}(e^{x/a} + e^{-x/a}).$$

Find the radius of curvature of this catenary at the point (x,y). *Ans.:* y^2/a.

13 Find the radius of curvature of the cycloid having the equation

$$x = a(\theta - \cos\theta), \qquad y = a(1 - \cos\theta).$$

14 Find the evolute of the cycloid of Problem 13.

15 Persons who picnic beside lakes several miles long can wonder whether poor visibility instead of curvature of the earth is responsible for invisibility of distant boats and shores. This problem involving curvature can be solved very simply. Figure 7.392 shows the center of a spherical earth at C on the positive y axis, the x axis being tangent to the surface of the earth at O. The line from C to $Q(x,0)$ intersects the surface of the earth at P. The number $|\overrightarrow{PQ}|$ is the height h of an object which is visible from the point O on the earth x miles away. The simple calculation

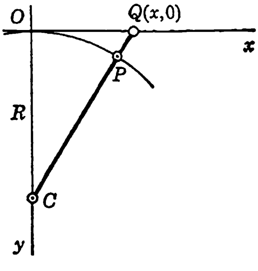

Figure 7.392

$$h = \sqrt{R^2 + x^2} - R = \frac{x^2}{\sqrt{R^2 + x^2} + R} \tag{1}$$

shows that the approximate formula

$$(2) \qquad h = \frac{x^2}{2R}$$

gives good results when x is small in comparison to R. It is a quite remarkable fact that if h and R are measured in miles and if the height H of the object is measured in feet so that $H = 5280h$, then we can put $R = 3960$ in (2) and multiply by 5280 to obtain

$$H = \tfrac{2}{3}x^2.$$

Putting $H = 6$ shows that only the hair on the top of the head of a man 6 feet tall is visible from a point on the earth 3 miles away. Putting $x = 30$ gives $H = 600$ and shows that less than the top half of the Empire State Building can be seen by persons on a ship 30 miles away.

16 It is sometimes useful to have formulas obtainable from (7.361) and (7.362). Letting $\mathbf{N}$ [where the N can make us think of a normal to C at $P(t)$] be the vector running from $P(t)$ to the center (X,Y) of curvature, we see that

$$(1) \qquad \mathbf{N} = \frac{[x'(t)]^2 + [y'(t)]^2}{x'(t)y''(t) - x''(t)y'(t)}\,[-y'(t)\mathbf{i} + x'(t)\mathbf{j}].$$

Let $\mathbf{b}$ (where the b can make us think of binormal or "second normal") be the unit vector in the direction of $\mathbf{v}(t) \times \mathbf{a}(t)$ so that, as (7.32) shows, $\mathbf{b}$ is the coordinate vector $\mathbf{k}$ in our work. Then, with the aid of (7.32) and the fact that $\mathbf{b}\cdot\mathbf{b} = 1$, we can put (1) in the intrinsic form

$$(2) \qquad \mathbf{N} = \frac{\mathbf{v}(t)\cdot\mathbf{v}(t)}{[\mathbf{v}(t) \times \mathbf{a}(t)]\cdot\mathbf{b}(t)}\,\mathbf{b}(t) \times \mathbf{v}(t)$$

in which coordinates do not appear. The formula (2) is valid when C is a curve in E_3 for which x,y,z are functions having continuous second derivatives and

$$\begin{aligned}
\mathbf{r}(t) &= x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k}\\
\mathbf{v}(t) &= x'(t)\mathbf{i} + y'(t)\mathbf{j} + z'(t)\mathbf{k}\\
\mathbf{a}(t) &= x''(t)\mathbf{i} + y''(t)\mathbf{j} + z''(t)\mathbf{k}
\end{aligned}$$

whenever t is such that $\mathbf{v}(t) \times \mathbf{a}(t) \neq 0$. We shall not prove this, but remark that the point (ξ,η,ζ) analogous to the point (ξ,η) of Figure 7.33 is the intersection of three planes and that (X,Y,Z) is the limit as $\Delta t \to 0$ of (ξ,η,ζ).

17 It is possible to obtain very informative formulas by considering the motion of a particle which moves along a plane curve C, endowed with coordinates as in Figure 7.27, in such a way that s increases as t increases. We assume existence and continuity of all the derivatives we want to use, and we assume that $dx/dt > 0$. Let $\mathbf{t}$ be the unit forward tangent vector to C at time t so that

$$(1) \qquad \mathbf{t} = \cos\phi\mathbf{i} + \sin\phi\mathbf{j},$$

where ϕ is, as in the discussion of Figure 7.384, an angle giving the direction of $\mathbf{t}$ at time t. Show that differentiating (1) gives

$$(2) \qquad \frac{d\mathbf{t}}{dt} = [-\sin\phi\mathbf{i} + \cos\phi\mathbf{j}]\frac{d\phi}{dt} = \mathbf{n}\frac{d\phi}{ds}\frac{ds}{dt} = \frac{ds}{dt}K\mathbf{n},$$

where $\mathbf{n}$ is the unit normal obtained by rotating $\mathbf{t}$ clockwise through the angle $\pi/2$, ds/dt is the speed of the particle, and K is the curvature $d\phi/ds$ of C. Show that writing the formula $\mathbf{v} = |\mathbf{v}|\mathbf{t}$ in the form

$$\mathbf{v} = \frac{ds}{dt}\mathbf{t} \tag{3}$$

and differentiating give

$$\mathbf{a} = \frac{d\mathbf{v}}{dt} = \frac{d^2s}{dt^2}\mathbf{t} + \frac{ds}{dt}\frac{d\mathbf{t}}{dt} \tag{4}$$

and that use of (2) then gives

$$\mathbf{a} = \frac{d^2s}{dt^2}\mathbf{t} + \left(\frac{ds}{dt}\right)^2 K\mathbf{n}. \tag{5}$$

Remark: This elegant formula gives the normal (or transverse) and tangential scalar components of the acceleration in terms of the speed and rate of change of the speed of the particle. The simplest applications involve the case in which the particle moves along the curve with constant speed σ so that $ds/dt = \sigma$ and $d^2s/dt^2 = 0$ at each time. In this case, (5) reduces to the simple but important formula

$$\mathbf{a} = \sigma^2 K\mathbf{n}. \tag{6}$$

If $K = 0$, then $\mathbf{a} = 0$. If $K \neq 0$, then $\mathbf{a}$ has magnitude $\sigma^2|K|$ or σ^2/ρ (where ρ is the radius of curvature) and is directed toward the center of curvature. Use ot (6) and the formula $\mathbf{F} = m\mathbf{a}$ gives the force required to propel a particle of mass m along a curve C with constant speed σ.

18 Let C be a simple closed convex curve which could, for example, be a famous "triangular roller" composed of the three vertices of an equilateral triangle together with three circular arcs of which each has its center at one vertex and contains the other two vertices. Let R be a rod (or line segment) whose length L is small enough to permit the rod to be "slid around" C in such a way that its two ends both remain on C. It is easy to presume, as has sometimes been done in "proofs" of the "Holditch theorem," that each point P on R must traverse a simple closed convex curve C_P as R slides around C. Persons having a compass, a straightedge like the edge of a sheet of paper upon which marks can be placed, and some spare time at their disposal can sketch interesting figures. Surprises await those who let C be a triangular roller, let the length L of the rod R be equal to or only a bit less than the distance between two vertices of the equilateral triangle, and let P be the mid-point of L.

8

Trigonometric functions

8.1 Trigonometric functions and their derivatives We are studying and perhaps even learning mathematics, and it may be amusing and perhaps even useful to see how an unscientific but highly critical Justice of a Supreme Court might be introduced to angles and trigonometry. We would teach him enough about numbers, E_2, continuous functions, and curves to make him realize that if we start with positive numbers a and h for which $0 < h < a$, then the ordered set of points $(x, \sqrt{a^2 - x^2})$ for which $h \leqq x \leqq a$, the point for which $x = x_2$ preceding the point for which $x = x_1$ when $a \leqq x_1 < x_2 \leqq h$, is a curve. Our curve lies on the circle with center at the origin and radius a, and we simplify (or complicate) matters by calling our curve "the arc, of the circle with center at the origin and radius a, which runs in the positive direction from the point $(a,0)$ to the point Q having coordinates $(h, \sqrt{a^2 - h^2})$" or, for temporary purposes, simply "the arc." To help us remember the meanings of our symbols, we can start constructing Figure 8.11 and amplify it as we proceed. Next we must teach our pupil a theory of length and prove to him that the arc has a length which we may denote by s. Knowing that our quantities depend upon h and a, we can define a number θ by the formula

$$(8.12) \quad \theta = \frac{s}{a} = \frac{\text{length of arc}}{\text{radius}}.$$

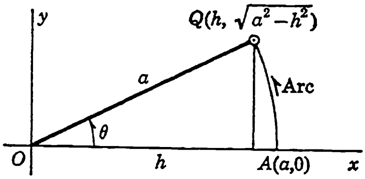

Figure 8.11

If we multiply each of a and h by the same positive constant λ (lambda), we get a new radius and a new arc, but the ratio s/a remains the same. To prove this, we must prove that the length of the new arc is the product of λ and the length of the original arc. Our theory of length enables us to do this because, by drawing radial lines from the origin, we can see that each polygon inscribed in the new arc determines (and is determined) by a polygon inscribed in the original arc and that the lengths of the straight segments of the polygons all differ by the same factor λ. All this shows that we get the same θ when we take another pair of values of a and h for which the point $(h, \sqrt{a^2 - h^2})$ lies on the same half-line extending from O through Q. Now we can again simplify (or complicate) matters by considering the number θ to be "a measure of the amount of rotation required to bring a line from the position OA to the position OQ" or "a measure of the opening between the lines OA and OQ." Perhaps to remind us where the number θ came from, or perhaps to indicate something of which θ is to be considered a measure, we complete Figure 8.11 by inserting the θ together with the curved arrow which shows the direction of our arc. It is the fashion to call θ an angle, but it should be permanently remembered that θ, like 5, is a number. The facts that we sometimes use θ to measure an amount of rotation and use 5 to measure a number of fingers do not imply that θ *is* a rotation and that 5 *is* a fistful of fingers, but we can nevertheless understand and even use the more or less convenient terminologies involving "angles" that have become a part of nonscientific as well as scientific attempts to convey information. If all this indicates that trigonometry is a subject much too difficult for inclusion in trigonometry textbooks, we do have one consolation. The hard work is done and the rest is easy.

Our theory of curves is sufficiently general to allow us to extend the above account of positive (that is, counterclockwise) arcs and angles to cover situations in which the arc is longer and Q lies in the second or third or fourth quadrant. Moreover, the arc can be so long that we must encircle the origin more than once to traverse it, and the number (or angle) θ is still defined by the same formula (8.12). In case the arc starts at A and is oriented in the negative (or clockwise) direction, everything is the same except that the directions of the arrows are reversed, a negative sign is prefixed to the middle and last members of

(8.12), and the number θ is negative. Even though we could pretend to be appalled by the idea that numbers have sides, we bow to conventions and agree that the way to find the six fundamental trigonometric functions of a given angle θ is to construct "the terminal side of θ" as in Figure 8.13, pick a point $P(x,y)$ on this terminal side, let r be the distance (positive, of course) from O to P, and use the numbers x, y, r in the usual way.

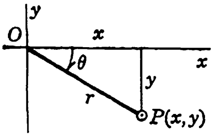

Figure 8.13

It is our purpose to present formulas for derivatives of trigonometric functions with derivations to which no logical objections can be raised. We could undertake to use the formula $A = \frac{1}{2}\theta r^2$ for the area of a circular sector of radius r which has central angle θ, but it would be immediately recognized that this formula has not been proved. Even if it be conceded that we know that the area of a whole circular disk of radius r is πr^2, it is only the docile acceptance of a crafty fraud that would allow us to "see at once" that the sector "obviously" has area $\frac{1}{2}\theta r^2$ because the area of the sector divided by the area of the circle is "obviously" $\theta/2\pi$. Nothing is obvious. Even a hazy understanding of the theory of area is enough to show that nothing can be proved without making substantial use of precise definitions and completeness of the real-number system or, what amounts to the same thing, consequences of these things. Since we do not now wish to be responsible for furnishing a complete treatment of areas of sectors, we shall base our work on the inequality (7.12). This inequality, the truth of which should seem thoroughly reasonable to a person at B who has the choice of walking two paths to D, has a virtue. It has been proved.

It was indicated in Problem 12 at the end of Section 3.2 that we will be able to derive the formulas for derivatives of trigonometric functions when we have proved the two formulas

$$(8.14)\qquad \lim_{\theta\to 0}\frac{\sin\theta}{\theta} = 1, \qquad \lim_{\theta\to 0}\frac{1-\cos\theta}{\theta} = 0.$$

Like a man who has drilled a hole and filled it with blasting powder to split a rock, we are ready to produce results. Figure 7.11 can be put into a coordinate system as in Figure 8.141, and we suppose that $0 < \theta < \pi/2$. In terms of the notation of this figure, the inequality (7.12) becomes

Figure 8.141

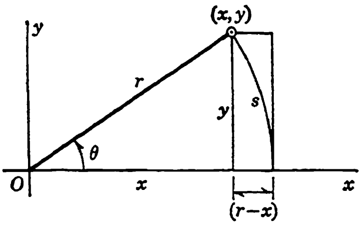

$$y \leqq s \leqq y + (r - x).$$

Dividing by r gives

$$(8.142)\qquad \sin\theta \leqq \theta \leqq \sin\theta + (1 - \cos\theta).$$

Since $1 - \cos\theta > 0$ and $1 + \cos\theta > 1$ and $\sin\theta < \theta$, the last term can be overestimated by the inequality

$$(8.143)\quad 0 < (1 - \cos\theta) < (1 - \cos\theta)(1 + \cos\theta) \\ = 1 - \cos^2\theta = \sin^2\theta < \theta\sin\theta.$$

Hence we can replace the last term in (8.142) by $\theta\sin\theta$ and divide by $\sin\theta$ to obtain the first and hence the second of the significant inequalities

$$(8.144)\qquad 1 \leqq \frac{\theta}{\sin\theta} < 1 + \theta, \qquad \frac{1}{1+\theta} < \frac{\sin\theta}{\theta} \leqq 1.$$

Dividing (8.143) by θ and using (8.142) gives

$$(8.145)\qquad 0 < \frac{1 - \cos\theta}{\theta} < \sin\theta < \theta.$$

The above inequalities have been proved to hold when $0 < \theta < \pi/2$. Since $|\theta| = \theta$ when $\theta > 0$, (8.144) and (8.145) imply that the inequalities

$$(8.146)\qquad \frac{1}{1+|\theta|} < \frac{\sin\theta}{\theta} \leqq 1, \qquad 0 < \frac{1-\cos\theta}{|\theta|} < |\theta|$$

hold when $0 < \theta < \pi/2$. Since the members of these inequalities are not changed when we replace θ by $-\theta$, we conclude that they hold when $|\theta| < \pi/2$ and $\theta \neq 0$. The desired formulas (8.14) follow from this and the sandwich theorem.

To derive the formulas for derivatives of sines and cosines, we use the formulas

$$\sin(x + \Delta x) = \sin x\cos\Delta x + \cos x\sin\Delta x$$
$$\cos(x + \Delta x) = \cos x\cos\Delta x - \sin x\sin\Delta x$$

to obtain

$$\sin(x + \Delta x) - \sin x = -\sin x(1 - \cos\Delta x) + \cos x\sin\Delta x$$
$$\cos(x + \Delta x) - \cos x = -\cos x\,(1 - \cos\Delta x) - \sin x\sin\Delta x.$$

After dividing by Δx, taking limits as $\Delta x \to 0$ of the resulting difference quotients gives, with the aid of (8.14), the fundamental formulas

$$(8.15)\qquad \frac{d}{dx}\sin x = \cos x, \qquad \frac{d}{dx}\cos x = -\sin x.$$

These formulas, which we have used many times, have at long last been proved. For values of x for which the functions are defined, we obtain

$$(8.151)\quad \frac{d}{dx}\tan x = \frac{d}{dx}\frac{\sin x}{\cos x} = \frac{\cos^2 x + \sin^2 x}{\cos^2 x} = \sec^2 x$$

$$(8.152)\quad \frac{d}{dx}\cot x = \frac{d}{dx}\frac{\cos x}{\sin x} = \frac{-\cos^2 x - \sin^2 x}{\sin^2 x} = -\csc^2 x$$

$$(8.153)\quad \frac{d}{dx}\sec x = \frac{d}{dx}\frac{1}{\cos x} = \frac{\sin x}{\cos^2 x} = \sec x \tan x$$

$$(8.154)\quad \frac{d}{dx}\csc x = \frac{d}{dx}\frac{1}{\sin x} = \frac{-\cos x}{\sin^2 x} = -\csc x \cot x.$$

The graphs of $\sin x$ and $\cos x$ are so important that we reproduce Figure 1.58 in Figure 8.16 and give further attention to the procedure by which reasonably accurate graphs are quickly sketched with or without use of graph paper. The trick is to sketch guide lines 1 unit above and 1 unit below the x axis, to hop 3 units and a bit more to the right of the origin to mark π, to make another such hop to mark 2π, and then sketch reasonably good copies of the figure. Each graph has slope 1 or -1 where it crosses an axis, and noticeable contradictions of this fact should not appear. In the problems at the end of the next section, we shall obtain formulas from which $\sin\theta$ and $\cos\theta$ can, for a given θ, be calculated as accurately as we wish. Meanwhile, we can be interested in Figure 8.17, which enables us to obtain reasonable estimates of $\sin\theta$ and $\cos\theta$ when $0 < \theta < \pi/2$. The circle has radius 1, and the radial lines make angles 0.1, 0.2, $\cdots$, 1.5 with the positive x axis. For example, the rough approximations

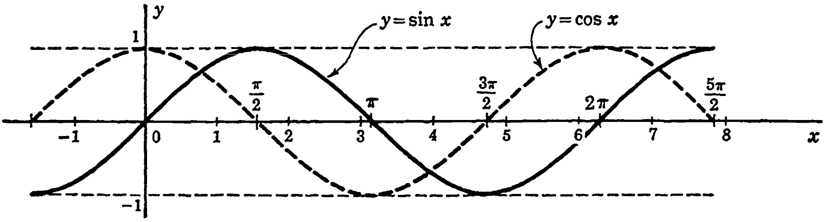

Figure 8.16

$$\sin 0.35 = \frac{0.34}{1} = 0.34, \qquad \cos 0.35 = \frac{0.94}{1} = 0.94,$$

$$\tan 0.35 = \frac{0.34}{0.94} = 0.36$$

can be read from the figure.

With the aid of information about derivatives, it is easy to see that the graph of the function t for which $t(x) = \tan x$ has the form shown in

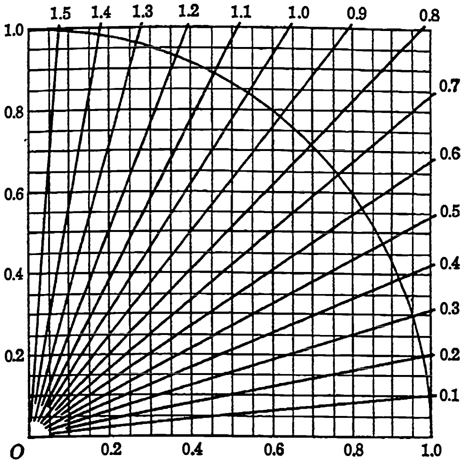

Figure 8.17

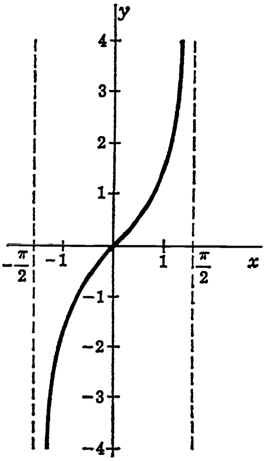

Figure 8.171

Figure 8.171. For each integer n, the interval $|x - n\pi| < \pi/2$ contains an exact copy of the graph in the interval $|x| < \pi/2$. The lines

$$x = n\pi \pm \frac{\pi}{2}$$

are all vertical asymptotes of the graph, and $\tan(n\pi \pm \pi/2)$ is undefined.

We should now be quite familiar with the fact that each formula for a derivative has a chain extension. The following list contains the chain extensions of formulas for derivatives of the six trigonometric functions and three additional formulas. All of these must be learned.

$$\frac{d}{dx}\sin u = \cos u\,\frac{du}{dx} \qquad \frac{d}{dx}\cot u = -\csc^2 u\,\frac{du}{dx}$$

$$\frac{d}{dx}\cos u = -\sin u\,\frac{du}{dx} \qquad \frac{d}{dx}\sec u = \sec u\tan u\,\frac{du}{dx}$$

$$\frac{d}{dx}\tan u = \sec^2 u\,\frac{du}{dx} \qquad \frac{d}{dx}\csc u = -\csc u\cot u\,\frac{du}{dx}$$

$$\frac{d}{dx}u^n = nu^{n-1}\frac{du}{dx}, \qquad \frac{d}{dx}e^u = e^u\frac{du}{dx}, \qquad \frac{d}{dx}\log u = \frac{1}{u}\frac{du}{dx}$$

Problems 8.19

1 With all graphs and tables out of sight, make the pretense that it has been forgotten whether the derivative with respect to x of $\sin x$ is $\sin x$ or $-\sin x$ or $\cos x$ or $-\cos x$. Sketch graphs of $\sin x$ and $\cos x$ and make these graphs give the correct answer.

2 Explain how you should modify the procedure for sketching the graph of

$y = \sin x$ to obtain a procedure for sketching the graph of $y = 3 \sin x$, and sketch the graph.

3 Equate the derivatives of the two members of the identity in the left column and show that obvious simplifications give the identity in the right column when

(*a*) $\sin 2x = 2 \sin x \cos x$	$\cos 2x = \cos^2 x - \sin^2 x$
(*b*) $\cos 2x = \cos^2 x - \sin^2 x$	$\sin 2x = 2 \sin x \cos x$
(*c*) $\sin^2 x + \cos^2 x = 1$	$0 = 0$
(*d*) $\sin (x + \phi)$ $= \sin x \cos \phi + \cos x \sin \phi$	$\cos (x + \phi)$ $= \cos x \cos \phi - \sin x \sin \phi$
(*e*) $\cos (x + \phi)$ $= \cos x \cos \phi - \sin x \sin \phi$	$\sin (x + \phi)$ $= \sin x \cos \phi + \cos x \sin \phi$
(*f*) $\cos^2 x = \dfrac{1 + \cos 2x}{2}$	$2 \sin x \cos x = \sin 2x$
(*g*) $\sin^2 x = \dfrac{1 - \cos 2x}{2}$	$2 \sin x \cos x = \sin 2x$
(*h*) $\sin \left(x + \dfrac{\pi}{2}\right) = \cos x$	$\cos \left(x + \dfrac{\pi}{2}\right) = -\sin x$
(*i*) $e^x e^x = e^{2x}$	$e^x e^x = e^{2x}$
(*j*) $\log x^2 = 2 \log x$	$\dfrac{2x}{x^2} = \dfrac{2}{x}$

4 Supposing that $0 \leqq x < \pi/2$, let $f(x) = \tan x$ and show that

$$\begin{aligned} f'(x) &= \sec^2 x \\ f''(x) &= 2 \sec^2 x \tan x \\ f'''(x) &= 2 \sec^4 x + 4 \sec^2 x \tan^2 x, \end{aligned}$$

and obtain the next derivative. Show that each derivative of higher order will also be a sum of terms of the form $A \sec^p x \tan^q x$, where A, p, q are nonnegative integers, and hence that $f^{(n)}(x) \geqq 0$ for each n and x.

5 Supposing that $0 < x < \pi/2$, calculate the first three derivatives of $\sec x$. Show that these and all derivatives of higher order are nonnegative.

6 Again supposing that $0 < x < \pi/2$, calculate some derivatives of $\sqrt{\tan x}$ and try to decide whether they are all nonnegative.

7 Prove the formula

$$\int \frac{\sin \theta}{1 + \sin \theta}\, d\theta = \sec \theta - \tan \theta + \theta + c$$

by differentiating the right side.

8 The length L of the longest beam that can be taken in a horizontal position around the corner of Figure 8.191 is the length L of the shortest line segment placed like the one in the figure. Find L. *Ans.:* $L = (a^{2/3} + b^{2/3})^{3/2}$. *Remark:* Putting the equation in the form $a^{2/3} + b^{2/3} = L^{2/3}$ can make us wonder why (a,b) should, for a fixed L, be a point on a hypocycloid.

Figure 8.191

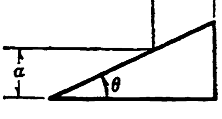

9 Find the length L of the shortest ladder that can reach from level ground to a high wall when it must go over a fence which is a feet high and b feet from the wall. *Ans.:* $L = (a^{2/3} + b^{2/3})^{3/2}$.

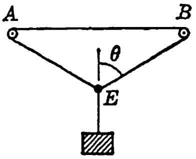

Figure 8.192

10 A heavy body is suspended from a rope, as in Figure 8.192, that runs up from the body to a pulley at E and thence (a wonderful word) over two stationary pulleys at A and B on the same horizontal level and back to E, where the pulley is tied to the other end of the rope. How should old man gravity select the θ of the figure in order to gratify his desire to bring the heavy body to its lowest possible position? *Ans.:* $\cos\theta = \frac{1}{2}$ and $\theta = \pi/3$ or $\theta = 60°$.

11 If $|P_n|$ is the length of a regular polygon of n sides which is inscribed in a circle of radius a, prove that $|P_n| = 2an \sin(\pi/n)$ and hence that $\lim\limits_{n\to\infty} |P_n| = 2\pi a$.

12 Supposing that A and B are constants not both 0 and that

$$y(x) = A \sin x + B \cos x,$$

calculate $y'(x) = y''(x)$ and show that

$$y''(x) = -y(x).$$

In terms of the graph of $y(x)$, tell precisely what it means to say that $y''(x) < 0$ when $y(x) > 0$ and $y''(x) > 0$ when $y(x) < 0$. Sketch graphs and verify your conclusion when $A = 0$ and when $B = 0$. It is now required that we learn a little trick that happens to be very important. Plot the point having coordinates (A,B) and, as in Figure 8.193, let ϕ be an angle having its terminal side on the line running from the origin to the point. Let $C = \sqrt{A^2 + B^2}$ and observe that

$$A = C \cos\phi \qquad \text{and} \qquad B = C \sin\phi$$

so

$$y(x) = C(\sin x \cos\phi + \cos x \sin\phi)$$

and hence

$$y(x) = C \sin(x + \phi).$$

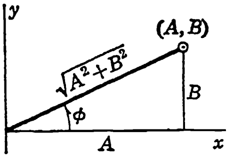

Figure 8.193

Finally, it is required that we learn some technical terminology by which this result can be remembered. Functions of the form $E \sin(\omega t + \phi)$ and $E \cos(\omega t + \phi)$ are called *sinusoids* (things like sines) of *angular frequency* ω. Since x can be ωt, we have proved that *the sum of two sinusoids having the same frequency ω is also a sinusoid having frequency ω.*

13 It is very easy to show that if k, A, B are constants for which $k > 0$ and if

$$y = A \sin kt + B \cos kt \tag{1}$$

then

$$\frac{d^2y}{dt^2} + k^2y = 0. \tag{2}$$

All we need to do is differentiate (1) twice and look at the result. The theory of differential equations contains theorems which imply that *if y is a function* *for which* (2) *holds, then there must be constants* A *and* B *for which* (1) *holds.* The latter and more difficult result is often needed by students who have not yet made reliable contacts with differential equations. Our solution of the more difficult problem will be postponed until we have solved the easy problem by a new method which involves only reversible steps. Starting with the formulas

$$(3)\qquad \begin{cases} y = A \sin kt + B \cos kt \\ \dfrac{dy}{dt} = Ak \cos kt - Bk \sin kt, \end{cases}$$

we eliminate B, and then eliminate A, from these equations to obtain

$$(4)\qquad \begin{cases} \cos kt \dfrac{dy}{dt} + (k \sin kt)y = Ak \\ \sin kt \dfrac{dy}{dt} - (k \cos kt)y = -Bk. \end{cases}$$

Therefore,

$$(5)\qquad \begin{cases} \dfrac{d}{dt}\left[\cos kt \dfrac{dy}{dt} + (k \sin kt)y\right] = 0 \\ \dfrac{d}{dt}\left[\sin kt \dfrac{dy}{dt} - (k \cos kt)y\right] = 0 \end{cases}$$

or

$$(6)\qquad \begin{cases} \cos kt \dfrac{d^2y}{dt^2} - k \sin kt \dfrac{dy}{dt} + k \sin kt \dfrac{dy}{dt} + (k^2 \cos kt)y = 0 \\ \sin kt \dfrac{d^2y}{dt^2} + k \cos kt \dfrac{dy}{dt} - k \cos kt \dfrac{dy}{dt} + (k^2 \sin kt)y = 0 \end{cases}$$

or

$$(7)\qquad \begin{cases} \cos kt \left[\dfrac{d^2y}{dt^2} + k^2y\right] = 0 \\ \sin kt \left[\dfrac{d^2y}{dt^2} + k^2y\right] = 0. \end{cases}$$

Since there is no t for which $\cos kt$ and $\sin kt$ are both 0, we conclude that

$$(8)\qquad \frac{d^2y}{dt^2} + k^2y = 0.$$

We are now ready to prove the more difficult result. Suppose that y is a given function for which (8) holds. Then we can multiply by $\cos kt$ and $\sin kt$ to obtain (7) and hence (6) and hence (5). Since the derivatives of the quantities in brackets are zero, these quantities must be constants which we can call Ak and $-Bk$ to obtain (4). Solving the equations (4) for y and dy/dt gives (3) and our problem is solved.

14 It is very easy to show that if k, A, B are constants for which $k > 0$ and if

$$(1)\qquad y = Ae^{kx} + Be^{-kx},$$

then

$$(2)\qquad \frac{d^2y}{dt^2} - k^2y = 0.$$

Do this. Then adapt the method of the preceding problem to show that if y is a function for which (2) holds, then there must be constants A and B for which (1) holds.

15 Prove that if y and u are functions of t having second derivatives and if

$$y = e^{-ht}u,$$

then

$$\frac{d^2y}{dt^2} + 2h\frac{dy}{dt} + k^2y = e^{-ht}\left[\frac{d^2u}{dt^2} + (k^2 - h^2)u\right].$$

16 A right circular cylinder (like a tomato can from which the top and bottom have been removed) has height h and base radius a. We examine the idea that good approximations to the area of this surface are obtained by triangulating it and calculating the sum of the areas of the triangles. The surface is partitioned into m strips each having height h/m like the one shown in Figure 8.194. On the top and bottom circular boundaries of each strip, we place the vertices of $2n$ isosceles triangles congruent to the triangle ABC of the figure. The side AB of the triangle subtends the angle $2\pi/n$ at the center O of the top circle. Of the $2n$ triangles in the top strip, n have two vertices on the top circle and the other n have two vertices on the lower circle. The total number of triangles congruent to the triangle ABC is $2mn$. The first factor in the product

$$(1)\qquad \left(a \sin\frac{\pi}{n}\right)\sqrt{\left(\frac{h}{m}\right)^2 + \left(a - a\cos\frac{\pi}{n}\right)^2}$$

is half the length of the base AB of the triangle. The last factor is the altitude, being the length of the hypotenuse of a triangle one leg of which runs from C to the right angle on the upper circle and the other leg of which runs (in the direction of O) from the right angle to the base AB. The sum S_{mn} of the areas of all of the triangles is the product of (1) and $2mn$. Therefore,

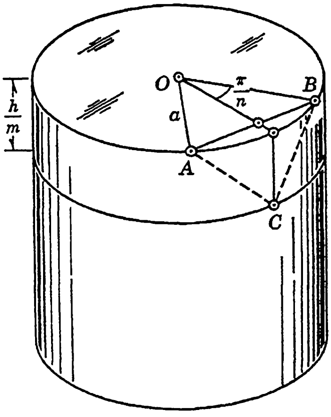

Figure 8.194

$$(2)\qquad S_{mn} = 2\pi ah\frac{\sin\frac{\pi}{n}}{\frac{\pi}{n}}\sqrt{1 + \left(\frac{\pi^2a}{2h}\right)^2\left(\frac{\sin\frac{\pi}{2n}}{\frac{\pi}{2n}}\right)^4\left(\frac{m}{n^2}\right)^2}.$$

When m and n are both large, the quotients having $\sin(\pi/n)$ in their numerators are near 1 and we have the approximation

$$(3)\qquad S_{mn} \sim 2\pi ah\sqrt{1 + \left(\frac{2\pi^2a}{h}\right)^2\left(\frac{m}{n^2}\right)^2}.$$

In case $m = n$, and in other cases where

$$\lim_{m,n\to\infty} \frac{m}{n^2} = 0, \tag{4}$$

we have

$$\lim_{m,n\to\infty} S_{mn} = 2\pi ah, \tag{5}$$

the right member being the number usually considered to be the correct area of the cylinder. In case $m = n^2$, we obtain

$$\lim_{m,n\to\infty} S_{mn} = 2\pi ah\sqrt{1 + \left(\frac{2\pi^2 a}{h}\right)^2}, \tag{6}$$

the right member being a number which is not usually considered to be the correct area of the cylinder. Other remarks can be made. The above calculations were made by a German mathematician Schwarz, and they constitute the *Schwarz paradox*. The paradox shows that the triangulation idea provides a precarious basis for definitions of area of curved surfaces. The theory of these areas is extremely complicated.

17 It is sometimes useful as well as interesting to have information about the things we see. A *polynomial* P in x and y is the sum of a finite set of terms of the form cx^jy^k, where c is a constant and j and k are nonnegative integers. The polynomial has *degree* n if $j + k = n$ for at least one term in the sum having a nonzero coefficient and $j + k \leqq n$ for each term in the sum having a nonzero coefficient. A polynomial in which the coefficients are all zero is said to be *trivial;* it does not have a degree. Thus the polynomials having values

$$x^4 + xy + y^4 - 34, \qquad (x^2 - y^2 - 1)(x^2 + y^2 - 4) \tag{1}$$

both have degree 4. A nontrivial polynomial is *irreducible* if it is not the product of two polynomials of lower degree. An *algebraic equation* of degree n is an equation of the form $P(x,y) = 0$, where P is a nontrivial polynomial of degree n in x and y. The graph of an algebraic equation is an *algebraic graph.* A function f of one variable x is said to be an *algebraic function* if there is a nontrivial polynomial $P(x,y)$ such that

$$P(x, f(x)) = 0 \tag{2}$$

for each x in the domain of f. Since each nontrivial polynomial in x and y can be put in the form

$$Q_0(x) + Q_1(x)y + Q_2(x)y^2 + \cdots + Q_n(x)y^n, \tag{3}$$

where $Q_0, Q_1, \cdots, Q_n$ are polynomials in x at least one of which is nontrivial, it follows that a function f is an algebraic function if and only if there exist polynomials $Q_0, Q_1, \cdots, Q_n$ in x such that Q_n is nontrivial and

$$Q_0(x) + Q_1(x)f(x) + Q_2(x)[f(x)]^2 + \cdots + Q_n(x)[f(x)]^n = 0 \tag{4}$$

for each x in the domain of f. Functions that are not algebraic functions are said to be *transcendental functions*, the old idea being that polynomials lie at the foundation of human experience and that things not closely related to polynomials are more ethereal. This matter can be of interest to us now because, even in quite elementary mathematics, the trigonometric functions are called transcendental functions. For example, the assertion that $\sin x$ is transcendental means that there do not exist polynomials $Q_0, Q_1, \cdots, Q_n$ in x such that Q_n is nontrivial and (4) is true for each x when $f(x) = \sin x$. Let us now think briefly about real numbers. A number x is said to be an *algebraic number* if there exist integers $k_0, k_1, k_2, \cdots, k_n$ not all zero such that

$$k_0 + k_1x + k_2x^2 + \cdots + k_nx^n = 0.$$

Thus a number x is algebraic if it is a zero of a nontrivial polynomial in x having integer coefficients. A number x which is not an algebraic number is said to be a *transcendental number*. The numbers π and e are transcendental. Nobody can prove these things unless he devotes very much time and energy to the operation, but nobody requires us to be so busy digging ditches that we never look at the stars. We can know that there exist theories of algebraic functions and algebraic numbers and that these theories invite the attention of persons who have completed studies of analytic geometry and calculus.

8.2 *Trigonometric integrands* Before introducing chain extensions and other modifications of the formulas, we systematically work out formulas for integrals of the six trigonometric functions. Even though the last three or four have minor importance, we must learn about them to be respectable. The first two are

$$\int \sin x\,dx = -\cos x + c, \qquad \int \cos x\,dx = \sin x + c. \tag{8.21}$$

They are immediate consequences of the formulas for derivatives of sines and cosines, since the formula

$$\int f(x)\,dx = F(x) + c$$

is valid over an interval if and only if $F'(x) = f(x)$ over the interval. It is necessary to keep negative signs in their places when we differentiate and integrate sines and cosines, and it happens that a foolish little trick enables us to permanently remember how the signs go. We can mentally write

$$\left\{\begin{array}{ll} \text{derivative} & \text{integral} \\ \text{sine} & \text{cosine} \end{array}\right. \tag{8.22}$$

in their natural orders and remember that "like things give plus," so differentiating sines and integrating cosines give plus signs, but that "unlike things give minus," so differentiating cosines and integrating sines give minus signs.

If an interval contains one of the points $n\pi \pm \pi/2$, there can be no F such that $\int \tan x\,dx = F(x) + c$ over that interval. When x is confined to an interval containing none of these points, we can use the fundamental formula

$$\int [u(x)]^{-1}u'(x)\,dx = \log|u(x)| + c$$

to obtain

$$(8.23)\qquad \int \tan x\,dx = -\int \frac{1}{\cos x}(-\sin x)\,dx = -\log|\cos x| + c$$
$$= \log|\sec x| + c.$$

In most applications of this formula, $|x| < \pi/2$, so $\cos x > 0$ and the absolute-value signs can be omitted. For $\cot x$, a miniature graph of which appears in Figure 8.232, we suppose that x is confined to an interval containing none of the points $x = n\pi$ and obtain

$$(8.231)\qquad \int \cot x\,dx = \int \frac{1}{\sin x}\cos x\,dx = \log|\sin x| + c.$$

In most applications of this formula, $0 < x < \pi$, so $\sin x > 0$ and the absolute-value signs can be omitted.

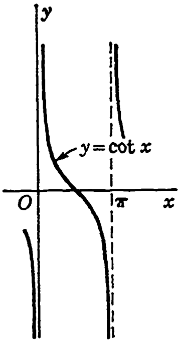

Figure 8.232

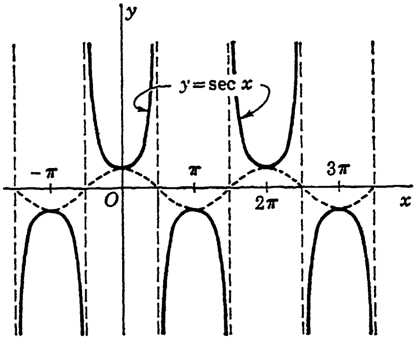

Figure 8.233

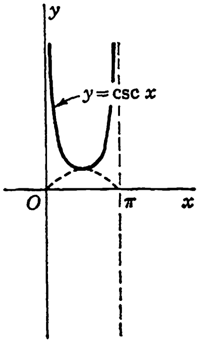

Figure 8.234

A sketch of the graph of $y = \sec x$ is most easily obtained by sketching a graph of $y = \cos x$ and estimating reciprocals. As the graph in Figure 8.233 indicates, $\sec x$ is undefined when x is an odd integer multiple of $\pi/2$. When we integrate $\sec x$ over an interval, we must suppose that the interval contains none of these points. We can then obtain the formula

$$(8.235)\qquad \int \frac{\sec x}{1}\,dx = \int \frac{1}{\sec x + \tan x}(\sec^2 x + \sec x \tan x)\,dx$$
$$= \log|\sec x + \tan x| + c$$

provided we happen to know that the result is obtained by multiplying the numerator and denominator of (sec x)/1 by the implausible factor sec x + tan x. We dispose of csc x as rapidly as possible by writing

$$(8.236)\quad \int \frac{\csc x}{1}\,dx = -\int \frac{1}{\csc x + \cot x}(-\csc^2 x - \csc x \cot x)\,dx = -\log|\csc x + \cot x| + c.$$

When $0 < x < \pi$ as in Figure 8.234, we can omit the absolute-value signs. The chain extensions of the six formulas are placed in a table at the end of this section where they are most available for reference. The formulas

$$(8.24)\quad \int \sin ax\,dx = \frac{1}{a}\int \sin ax\; a\,dx = -\frac{1}{a}\cos ax + c$$

$$(8.241)\quad \int \cos ax\,dx = \frac{1}{a}\int \cos ax\; a\,dx = \frac{1}{a}\sin ax + c,$$

which hold when $a \neq 0$, are by far the most important applications of the chain extensions.

We now consider some of the more or less important integrals that can be evaluated in terms of elementary functions. Adding and subtracting the two elementary formulas

$$(8.25)\quad \cos^2\theta + \sin^2\theta = 1$$

$$(8.251)\quad \cos^2\theta - \sin^2\theta = \cos 2\theta$$

gives the two formulas

$$(8.252)\quad \cos^2\theta = \frac{1+\cos 2\theta}{2}, \qquad \sin^2\theta = \frac{1-\cos 2\theta}{2}$$

that are so useful that we should either learn them or be able to work them out very quickly. Moreover, these formulas should come to mind when we are asked to evaluate the first integrals in the formulas

$$(8.253)\quad \int \cos^2 ax\,dx = \int \frac{1+\cos 2ax}{2}\,dx = \frac{1}{2}\left[x + \frac{1}{2a}\int \cos 2ax(2a)\,dx\right] = \frac{1}{2}\left[x + \frac{1}{2a}\sin 2ax\right] + c$$

and

$$(8.254)\quad \int \sin^2 ax\,dx = \int \frac{1-\cos 2ax}{2}\,dx = \frac{1}{2}\left[x - \frac{1}{2a}\int \cos 2ax(2a)\,dx\right] = \frac{1}{2}\left[x - \frac{1}{2a}\sin 2ax\right] + c.$$

Students are frequently called upon to realize that the formula

$$\cos^4 x = (\cos^2 x)^2 = \left(\frac{1 + \cos 2x}{2}\right)^2 = \frac{1 + 2 \cos 2x + \cos^2 2x}{4}$$

is useful when the left member must be integrated, and to know what to do to complete the problem.

The integral in

$$I = \int \sin mx \cos nx \, dx \tag{8.26}$$

looks forbidding until it is discovered or remembered that a useful formula for the integrand can be obtained by adding the formulas

$$\begin{aligned} \sin(\theta + \phi) &= \sin\theta \cos\phi + \cos\theta \sin\phi \\ \sin(\theta - \phi) &= \sin\theta \cos\phi - \cos\theta \sin\phi. \end{aligned}$$

Thus

$$\tfrac{1}{2}[\sin(mx + nx) + \sin(mx - nx)] = \sin mx \cos nx$$

and

$$\begin{aligned} I &= \frac{1}{2(m+n)} \int [\sin(m+n)x](m+n)\,dx \\ &\qquad + \frac{1}{2(m-n)} \int [\sin(m-n)x](m-n)\,dx \\ &= \frac{1}{2(m+n)} \cos(m+n)x - \frac{1}{2(m-n)} \cos(m-n)x + c \end{aligned} \tag{8.261}$$

except when $m = n$ or $m = -n$. Similarly, the integrals

$$\int \sin mx \sin nx \, dx, \qquad \int \cos mx \cos nx \, dx \tag{8.262}$$

can be evaluated by use of formulas obtained by adding and subtracting the formulas

$$\begin{aligned} \cos(\theta - \phi) &= \cos\theta \cos\phi + \sin\theta \sin\phi \\ \cos(\theta + \phi) &= \cos\theta \cos\phi - \sin\theta \sin\phi. \end{aligned}$$

Sometimes integrals can be evaluated by making quite direct use of the power formula and other integration formulas. For example,

(8.~~2~~63)

$$\begin{aligned} &\int \sin^p ax\ \cancel{\cos ax \, dx} \\ &\qquad = \frac{1}{a} \int (\sin ax)^p (a \cos ax)\,dx = \frac{1}{a}\frac{\sin^{p+1} ax}{p+1} + c \end{aligned}$$

when $a \neq 0$ and $p \neq -1$. Sometimes we need some ingenuity. For example, the integral

$$\int \sin^2 x \cos^3 x \, dx \tag{8.264}$$

is rendered manageable by replacing the factor $\cos^3 x$ by the last member of the formula

$$\cos^3 x = \cos^2 x \cos x = (1 - \sin^2 x) \cos x.$$

The integral in

$$(8.265) \qquad \int e^{\cos x} \sin x \, dx = -\int e^{\cos x}(-\sin x)\, dx = -e^{\cos x} + c$$

is evaluated by making the adjustment and compensation necessary to put it in the form $\int e^{u(x)}u'(x)\, dx$, and opportunities to make such adjustments and compensations should always be observed.

There are reasons why some of the integrals we have evaluated may be said to be so important that everyone should know how to evaluate them. There is an old and perhaps honorable tradition that requires students of the calculus to spend huge amounts of time cultivating "the technique" of "formal integration." The fact that we live in an age of electronic computers makes it much more important to learn fundamental theory than to acquire skill in formal integration. For this reason, the author requests that teachers join him in avoiding all but the simplest formal integration problems that are not likely to be encountered by undergraduates in courses other than calculus courses. "The student" who has not read dozens of calculus books and does not know what we are talking about is invited to look at the shiny example

$$(8.27) \qquad I = \int \sqrt{\sin x \tan x(1 + \sqrt{\cos x})}\, dx$$

of an integral that we shall not expect him to evaluate quickly. A person who has constructed this problem can easily feel very sure that the only sensible attack upon the problem lies in setting

$$(8.271) \qquad u(x) = 1 + (\cos x)^{1/2};$$

so, when $0 < x < \pi/2$,

$$(8.272) \qquad \begin{aligned} u'(x) &= \tfrac{1}{2}(\cos x)^{-1/2}(-\sin x) \\ &= -\tfrac{1}{2}\sqrt{\frac{\sin^2 x}{\cos x}} = -\tfrac{1}{2}\sqrt{\sin x \tan x} \end{aligned}$$

and hence

$$(8.273) \qquad \begin{aligned} I &= -2 \int [u(x)]^{1/2}[u'(x)]\, dx \\ &= -2 \cdot \frac{[u(x)]^{3/2}}{\frac{3}{2}} + c = -\tfrac{4}{3}[1 + \sqrt{\cos x}]^{3/2} + c. \end{aligned}$$

Instead of inviting attention to problems of this nature, we present problems more likely to promote scientific competence.

Table 8.28

$$\int \sin u \frac{du}{dx}\,dx = -\cos u + c \qquad \int \cos u \frac{du}{dx}\,dx = \sin u + c$$

$$\int \tan u \frac{du}{dx}\,dx = \log|\sec u| + c \qquad \int \cot u \frac{du}{dx}\,dx = \log|\sin u| + c$$

$$\int \sec u \frac{du}{dx}\,dx = \log|\sec u + \tan u| + c \qquad \int \csc u \frac{du}{dx}\,dx = \log|\csc u - \cot u| + c$$

$$\int u^n \frac{du}{dx}\,dx = \frac{u^{n+1}}{n+1} + c, \int \frac{1}{u}\frac{du}{dx}\,dx = |\log u| + c, \int e^u \frac{du}{dx}\,dx = e^u + c.$$

Problems 8.29

1 Make all of the calculations necessary to show that

(*a*) $\int_0^\pi \sin x\,dx = 2$ (*b*) $\int_0^{\pi/2} \cos x\,dx = 1$

(*c*) $\lim_{\omega\to\infty} \int_a^b \sin \omega x\,dx = 0$ (*d*) $\lim_{\omega\to\infty} \int_a^b \cos \omega x\,dx = 0$

(*e*) $\lim_{x\to\infty} \frac{1}{x}\int_0^x \sin^2 \omega t\,dt = \frac{1}{2}$ (*f*) $\lim_{x\to\infty} \frac{1}{x}\int_0^x \cos^2 \omega t\,dt = \frac{1}{2}$

(*g*) $\int_0^{\pi/4} \tan x\,dx = \frac{1}{2}\log 2$ (*h*) $\int_0^{\pi/2} \tan x\,dx = \infty$

2 Recall that, when u is a differentiable function of x and $u \neq 0$,

$$\frac{d}{dx}\log|u| = \frac{1}{u}\frac{du}{dx}; \tag{1}$$

the absolute-value signs, which are superfluous when $u > 0$, need not bother us. Supposing that x is not an odd multiple of $\pi/2$ and that

$$f(x) = \log|\sec x + \tan x|, \qquad g(x) = \log\left|\tan\left(\frac{x}{2} + \frac{\pi}{4}\right)\right|, \tag{2}$$

show that $f'(x) = \sec x$ and $g'(x) = \sec x$.

Remark: This proves that the two formulas

$$\int \sec x\,dx = \log|\sec x + \tan x| + c_1 \tag{3}$$

$$\int \sec x\,dx = \log\left|\tan\left(\frac{x}{2} + \frac{\pi}{4}\right)\right| + c_2 \tag{4}$$

are both correct. Some integral tables contain both of them. These things imply that, over each interval containing no odd multiple of $\pi/2$, $f(x) - g(x)$

must be a constant, but they do not imply that this constant must be zero. The identity

$$\tan\left(\frac{x}{2}+\frac{\pi}{4}\right)=\frac{1+\tan\frac{x}{2}}{1-\tan\frac{x}{2}}=\frac{\cos\frac{x}{2}+\sin\frac{x}{2}}{\cos\frac{x}{2}-\sin\frac{x}{2}}\,\frac{\cos\frac{x}{2}+\sin\frac{x}{2}}{\cos\frac{x}{2}+\sin\frac{x}{2}}$$

$$=\frac{1+2\sin\frac{x}{2}\cos\frac{x}{2}}{\cos^2\frac{x}{2}-\sin^2\frac{x}{2}}=\frac{1+\sin x}{\cos x}=\sec x+\tan x$$

clarifies the matter.

3 Prove that if x is not an even multiple of $\pi/2$ and

$$f(x)=\log|\csc x-\cot x|,\qquad g(x)=\log\left|\tan\frac{x}{2}\right|,$$

then $f'(x)=\csc x$ and $g'(x)=\csc x$. *Remark:* This proves that the two formulas

$$\int\csc x\,dx=\log|\csc x-\cot x|+c,\qquad \int\csc x\,dx=\log\left|\tan\frac{x}{2}\right|+c$$

are both correct.

4 When a steady (or constant) current I (measured in amperes) flows through a wire (or resistor) having resistance R (measured in ohms), the quantity Q (measured in watthours, or thousandths of kilowatthours) of energy converted into heat in Δt hours is calculated from the formula

$$Q=I^2R\,\Delta t. \tag{1}$$

With this basic information, show that if $I(t)$ is integrable (Riemann) over the time interval $t_1\leqq t\leqq t_2$, then the quantity $Q(t_1,t_2)$ of energy converged into heat between times t_1 and t_2 is (or should be)

$$Q(t_1,t_2)=R\int_{t_1}^{t_2}[I(t)]^2\,dt. \tag{2}$$

Suppose now that $I(t)$ is the sinusoidal (or alternating) current determined by

$$I(t)=I_0\sin(\omega t+\phi), \tag{3}$$

where I_0 is a constant "maximum current," ω is a constant "angular frequency," and ϕ is a "phase angle." Show that, in this case,

$$Q(t_1,t_2)=\frac{t_2-t_1}{2}I_0^2R-\frac{I_0^2R}{4\omega}[\sin(2\omega t_2+2\phi)-\sin(2\omega t_1+2\phi)]. \tag{4}$$

The last term is 0 whenever $2\omega t_2-2\omega t_1$ is an integer multiple of 2π, and in every case the absolute value of the last term cannot exceed $RI_0^2/2\omega$. In all ordinary applications of this formula, t_2-t_1 is so large in comparison to $1/\omega$ that the last term in (4) is insignificant. In such cases, (4) is always replaced by

$$Q(t_1,t_2)=\tfrac{1}{2}I_0^2R(t_2-t_1). \tag{5}$$

Putting (5) in the form

$$Q(t_1,t_2) = \left(\frac{I_0}{\sqrt{2}}\right)^2 R(t_2 - t_1) \tag{6}$$

shows that a sinusoidal current having maximum value I_0 produces heat (or dissipates energy) just as rapidly as a steady current of magnitude $I_0/\sqrt{2}$. For this reason, the number $I_0/\sqrt{2}$ is called the *effective value* of the sinusoidal current. For ordinary house current having "effective voltage" 120 volts, the maximum (or peak) voltage is not 120 volts but is 120 $\sqrt{2}$ volts.

5 Let, when L is a positive number and $n = 1, 2, 3, \cdots$,

$$\phi_n(x) = \sqrt{\frac{2}{L}} \sin \frac{n\pi x}{L}, \qquad \psi_0(x) = \frac{1}{\sqrt{L}}, \qquad \psi_n(x) = \sqrt{\frac{2}{L}} \cos \frac{n\pi x}{L}.$$

Show that the first of the formulas

$$\int_0^L \phi_m(x)\phi_n(x)\,dx = 0, \qquad \int_0^L [\phi_n(x)]^2\,dx = 1$$

holds when m and n are different positive integers and that the second holds when $n = 1, 2, 3, \cdots$. Show that the first of the formulas

$$\int_0^L \psi_m(x)\psi_n(x)\,dx = 0, \qquad \int_0^L [\psi_n(x)]^2\,dx = 1$$

holds when m and n are different nonnegative integers and that the second holds when $n = 0, 1, 2, \cdots$. Show that

$$\int_0^L \psi_0(x)\phi_n(x)\,dx = \sqrt{2}\,\frac{1 - \cos n\pi}{n\pi} \qquad (n = 1, 2, 3, \cdots).$$

Show that, when $m, n = 1, 2, 3, \cdots$,

$$\int_0^L \psi_m(x)\phi_n(x)\,dx = \frac{1 - \cos (m + n)\pi}{(m + n)\pi} + \frac{1 - \cos (m - n)\pi}{(m - n)\pi},$$

where the last term is to be omitted when $m = n$. *Remark:* While students in calculus courses have not yet heard about the matter, the above formulas are of great interest in the theory of orthonormal sets and Fourier series of the trigonometric variety. Our calculations and a little theory produce the interesting formulas

$$\frac{\pi}{4} = \sin \frac{\pi x}{L} + \frac{1}{3} \sin \frac{3\pi x}{L} + \frac{1}{5} \sin \frac{5\pi x}{L} + \cdots \qquad (0 < x < L)$$

$$\frac{\pi^2}{8} = 1 + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \cdots, \qquad \frac{\pi^2}{6} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots$$

and many others.

6 A problem in Section 8.4 will tell us about the formula

$$\int_0^{\pi/2} \sin^p \theta \cos^q \theta\,d\theta = \frac{\frac{p-1}{2}!\,\frac{q-1}{2}!}{2\,\frac{p+q}{2}!}$$

which is correct whenever $p > -1$ and $q > -1$. As usual, $0! = 1$, $1! = 1$, $2! = 1\cdot 2$, $3! = 1\cdot 2\cdot 3$, $4! = 1\cdot 2\cdot 3\cdot 4$, etcetera. The values of $x!$ when x is not an integer are more esoteric, but $x! > 0$ when $x > -1$. Put $p = q = 0$ and discover that $(-\frac{1}{2})! = \sqrt{\pi}$. Show that the formula is correct when $p = q = 1$. Put $p = q = 2$ and discover that $(\frac{1}{2})! = \sqrt{\pi}/2$. Put $p = 2x - 1$ and $q = 1$ and use the result to prove that $x!$ is the product of x and $(x - 1)!$.

7 With an assist from (8.144), which shows that

$$\int_0^1 \frac{1}{(1+\theta)^2}\,d\theta \leqq \int_0^1 \left(\frac{\sin\theta}{\theta}\right)^2 d\theta \leqq \int_0^1 1\,d\theta,$$

the middle integral being a Riemann-Cauchy integral because the integrand is undefined when $\theta = 0$, show that

$$\tfrac{1}{2} \leqq \int_0^1 \left(\frac{\sin\theta}{\theta}\right)^2 d\theta \leqq 1.$$

Then prove the first of the inequalities

$$0 \leqq \int_1^\infty \left(\frac{\sin\theta}{\theta}\right)^2 d\theta \leqq 1, \qquad \tfrac{1}{2} \leqq \int_0^\infty \left(\frac{\sin\theta}{\theta}\right)^2 d\theta \leqq 2$$

and use it to obtain the second.

8 Let a and b be constants for which $a = 0$ and $b = 1$ or $a = 1$ and $b = 0$ and let $x > 0$. Show that

$$-1 \leqq a\cos x + b\sin x \leqq 1.$$

Replace x by t and integrate over the interval $0 \leqq t \leqq x$ to obtain

$$-x \leqq a\sin x - b(\cos x - 1) \leqq x.$$

Replace x by t and integrate over the interval $0 \leqq t \leqq x$ to obtain

$$-\frac{x^2}{2} \leqq -a(\cos x - 1) - b(\sin x - x) \leqq \frac{x^2}{2}.$$

Repeat the process to obtain

$$-\frac{x^3}{3!} \leqq -a(\sin x - x) + b\left(\cos x - 1 + \frac{x^2}{2}\right) \leqq \frac{x^3}{3!}$$
$$-\frac{x^4}{4!} \leqq a\left(\cos x - 1 + \frac{x^2}{2!}\right) + b\left(\sin x - x + \frac{x^3}{3!}\right) \leqq \frac{x^4}{4!}$$
$$-\frac{x^5}{5!} \leqq a\left(\sin x - x + \frac{x^3}{3!}\right) - b\left(\cos x - 1 + \frac{x^2}{2!} - \frac{x^4}{4!}\right) \leqq \frac{x^5}{5!}.$$

Repeat the process two more times. With or without more attention to details, jump to the conclusion that

$$\left|\sin x - \left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \pm \frac{x^{2n+1}}{(2n+1)!}\right)\right| \leqq \frac{|x|^{2n+3}}{(2n+3)!}$$
$$\left|\cos x - \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots \pm \frac{x^{2n}}{(2n)!}\right)\right| \leqq \frac{|x|^{2n+2}}{(2n+2)!}$$

for each $n = 1, 2, 3, \cdots$. It is true (and is easy to surmise) that, for each x, $|x|^n/n! \to 0$ as $n \to \infty$. Therefore,

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$
$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots.$$

9 The formulas at the end of the preceding problem have been proved to be correct when $x > 0$. Use this fact to prove that the formulas are also correct when $x \leqq 0$. *Hint:* Look first at the case in which $x = 0$. Then use the facts that $\sin(-x) = -\sin x$ and $\cos(-x) = \cos x$.

10 Have a good look at the formulas at the end of Problem 8 and start learning them by obtaining some of the following results correct to four or more decimal places.

$$\begin{array}{ll} \sin 0.01 = 0.00999\,9833 & \cos 0.01 = 0.99995\,0000 \\ \sin 0.02 = 0.01999\,8667 & \cos 0.02 = 0.99980\,0007 \\ \sin 0.10 = 0.09983\,3417 & \cos 0.10 = 0.99500\,4165 \\ \sin 0.20 = 0.19866\,9331 & \cos 0.20 = 0.98006\,6587 \\ \sin 1.00 = 0.84147\,0985 & \cos 1.00 = 0.54030\,2306 \\ \sin 1.10 = 0.89120\,7360 & \cos 1.10 = 0.45359\,6121 \end{array}$$

11 Digest the following idea. We have a desk calculator and National Bureau of Standards Tables giving the values

$$\sin 0.2345 = 0.23235\,6699$$
$$\cos 0.2345 = 0.97263\,0641$$

If we want to find sin 0.23456 789 correct to eight decimal places, we can use the identity

$$\sin(x + 0.2345) = \sin x \cos 0.2345 + \cos x \sin 0.2345,$$

where $x = 0.00006789$. The values of $\sin x$ and $\cos x$ can easily be found correct to 10 decimal places from the formulas at the end of Problem 8.

8.3 Inverse trigonometric functions Before coming to the announced subject of this section, we think about a general situation applications of which appear in many branches of mathematics. Suppose we have an operator or transformer or mapper or function f that carries or transforms or maps each element x of a set D (the domain of f) into an element y of a set R (the range of f). In some cases the transformer transforms two or more different elements of D into the same element of R. For example, if f is one of the six trigonometric functions and $f(x)$ exists, then $f(x + 2\pi) = f(x)$ and hence more than one element of the domain is carried into the same element $f(x)$ of the range. When problems involving domains and ranges are involved, it is very often possible to eliminate confusion by singling out for special attention a subset D_1 of D such that to each y in R there corresponds exactly one x

in D_1 for which $f(x) = y$. Then we can introduce a function f_1, which is different from f because it has a different domain, for which $f_1(x) = f(x)$ when x is in D_1 and for which $f_1(x)$ is undefined when x is not in D_1. This function f_1 is called *the restriction of f to D_1*. We now have a general situation that may be easier to visualize than its applications to trigonometric functions. The schematic Figure 8.31 shows the domain and range of f_1, and the upper arc from x to y can make us think of a path along which f_1 might carry x to the y which is $f_1(x)$. The function f_1 sets up a one-to-one (or *schlicht*) correspondence between D_1 and R, that is, to each x in D_1 corresponds exactly one y in R for which $f(x) = y$ and to each y in R corresponds exactly one x in D_1 for which $f(x) = y$. The function with domain R and range D_1 which carries each y in R into the x in D_1 for which $f_1(x) = y$ is called the *inverse* of f_1 and is denoted by f_1^{-1}. As indicated by the figure, f_1^{-1} undoes what f_1 does. If $f_1(x) = y$, then $x = f_1^{-1}(y)$. Moreover, $f_1^{-1}(f_1(x)) = x$ when x is in D and

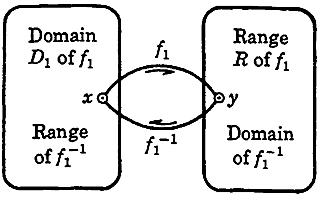

Figure 8.31

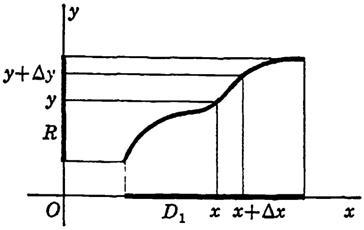

Figure 8.32

$$f_1(f_1^{-1}(y)) = y$$

when y is in R.

Figure 8.32 shows the graph of a particular function f to which the following theorem applies.

Theorem 8.33 *If f is continuous over an interval D_1 in E_1 and if f has a derivative for which $f'(x) > 0$ at each inner point x of D_1 [or $f'(x) < 0$ at each inner point x of D_1] then the restriction f_1 of f to D_1 has a range R which is an interval and has also an inverse f_1^{-1} which is differentiable at each inner point y of R. Moreover, the first of the formulas*

$$f_1^{-1\prime}(y) = \frac{1}{f_1'(f_1^{-1}(y))}, \qquad \frac{dx}{dy} = \frac{1}{\dfrac{dy}{dx}} \tag{8.331}$$

is valid when y is an inner point of R.

It seems to be possible to put the first formula of (8.331) in the second form without sacrifice of meaning. To prove the theorem, we use the hypotheses to conclude that f_1 is increasing (or decreasing) over D_1 and that (because of the intermediate-value theorem) R is an interval. Hence

f_1 has an inverse defined over R. Since we must *prove* that f_1^{-1} is differentiable, we introduce difference quotients for f_1^{-1}. Supposing that y and $y + \Delta y$ belong to the domain R of f_1^{-1}, we define x and $x + \Delta x$ by the formulas $x = f_1^{-1}(y)$ and $x + \Delta x = f_1^{-1}(y + \Delta y)$ and obtain $y = f_1(x)$, $y + \Delta y = f_1(x + \Delta x)$. Since f_1 has a positive (or negative) derivative, we conclude that $\Delta x \neq 0$ when $\Delta y \neq 0$ and that $\Delta x \to 0$ as $\Delta y \to 0$. Therefore,

$$\text{(8.332)} \quad \lim_{\Delta y \to 0} \frac{f_1^{-1}(y + \Delta y) - f_1^{-1}(y)}{\Delta y} = \lim_{\Delta y \to 0} \frac{\Delta x}{f_1(x + \Delta x) - f_1(x)}$$
$$= \lim_{\Delta y \to 0} \frac{1}{\dfrac{f_1(x_1 + \Delta x) - f_1(x)}{\Delta x}} = \frac{1}{f_1'(x)} = \frac{1}{f_1'(f_1^{-1}(y))}.$$

This proves Theorem 8.33. Of course it is possible to abbreviate the calculations by writing

$$\text{(8.333)} \qquad \frac{dx}{dy} = \lim_{\Delta y \to 0} \frac{\Delta x}{\Delta y} = \frac{1}{\lim\limits_{\Delta x \to 0} \dfrac{\Delta y}{\Delta x}} = \frac{1}{\dfrac{dy}{dx}},$$

but this one line is not, by itself, the equivalent of a theorem and proof which present conditions under which the formula is correct.

Our general discussion of inverse functions and Theorem 8.33 will now be used to guide us to six functions that are called inverse trigonometric functions even though they are in fact inverses of restrictions of trigonometric functions. It will turn out that the six functions will all have values between 0 and $\pi/2$, and that all of the information implied by Figures 8.341, 8.342, and 8.343 will be correct, provided $0 < x < 1$ in Figure 8.341, $x > 0$ in Figure 8.342, and $x > 1$ in Figure 8.343. The first figure shows, for example, that

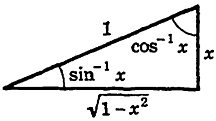

Figure 8.341

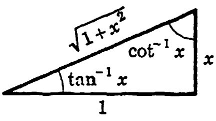

Figure 8.342

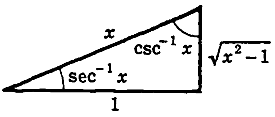

Figure 8.343

$$\sin(\sin^{-1} x) = x, \qquad \tan(\sin^{-1} x) = \frac{x}{\sqrt{1 - x^2}}$$

when $0 < x < 1$. These triangles can sometimes provide helpful information even when the functions appear in problems involving the misfortunes of gamblers who bet on horse races.

To begin, let f be the trigonometric function for which $f(x) = \sin x$, the domain D being the infinite interval $-\infty < x < \infty$ and the range

R being the closed interval $-1 \leqq y \leqq 1$. When we are called upon to select a domain D_1 to which we can apply Theorem 8.33, the graph of $y = \sin x$ can remind us of known properties of its derivative which show us that if we are going to be able to apply Theorem 8.33, we must let D_1 be the interval $-\pi/2 < x < \pi/2$ or some other interval lying $n\pi$ (n an integer) units to the right or left of it. Making the simplest choice, we let f_1 be the function, defined only when $-\pi/2 \leqq x \leqq \pi/2$, for which $f_1(x) = \sin x$. The inverse of f_1 is called the inverse sine. If $y = f_1(x)$, so that $y = \sin x$ and $-\pi/2 \leqq x \leqq \pi/2$ and $-1 \leqq y \leqq 1$, we write $x = \sin^{-1} y$. The graph of $x = \sin^{-1} y$ coincides with the part of the graph of $y = \sin x$ for which $-\pi/2 \leqq x \leqq \pi/2$, and the graph is shown in Figure 8.351. When we interchange x and y, which amounts to replacing each point (x,y) by its "image" (y,x) in the line $y = x$, we find that the graph of the equation $y = \sin^{-1} x$ is the same as the part of the graph of $x = \sin y$ for which $-\pi/2 \leqq y \leqq \pi/2$ and that the graph is shown in Figure 8.352. When $-1 < x < 1$ and $y = \sin^{-1} x$, we have $\sin y = x$, $-\pi/2 < y < \pi/2$, and $\cos y > 0$. Since Theorem 8.33, with x and y interchanged, implies that dy/dx exists, we can use the chain rule to obtain $\cos y(dy/dx) = 1$ and hence

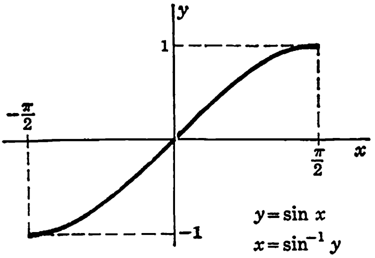

Figure 8.351

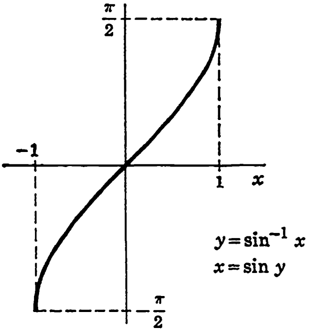

Figure 8.352

$$(8.353)\quad \frac{d}{dx}\sin^{-1} x = \frac{dy}{dx} = \frac{1}{\cos y} = \frac{1}{\sqrt{1-\sin^2 y}} = \frac{1}{\sqrt{1-x^2}}.$$

While we could not be expected to guess the result in this formula, we can look at Figure 8.352 and see that it seems to be a very reasonable answer. One who feels that inverse sines and their derivatives could never be useful should be informed at once that (8.353) gives the formula

$$(8.354)\qquad \int \frac{1}{\sqrt{1-x^2}}\,dx = \sin^{-1} x + c \qquad (|x| < 1)$$

which is useful because the integral appears in important problems.

The inverse cosine is the inverse of the function f_1, with domain $0 \leqq x \leqq \pi$ and range $-1 \leqq x \leqq 1$, for which $f_1(x) = \cos x$. Graphs of $x = \cos^{-1} y$, of $y = \cos^{-1} x$, and of the relevant parts of $y = \cos x$ and $x = \cos y$ appear in Figures 8.355 and 8.356. When $-1 < x < 1$ and $y = \cos^{-1} x$, then $\cos y = x$, $0 < y < \pi$, and $\sin y > 0$. Again Theorem 8.33 implies that dy/dx exists, so $(-\sin y)\,(dy/dx) = 1$ and

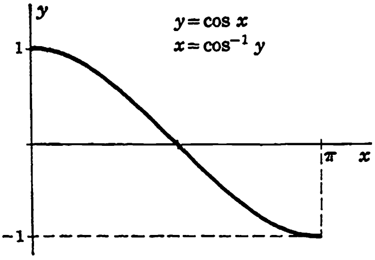

Figure 8.355

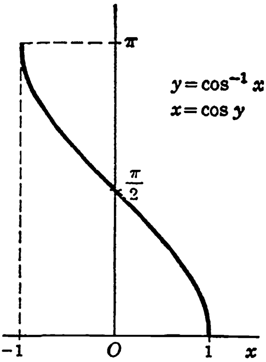

Figure 8.356

$$(8.357)\quad \frac{d}{dx}\cos^{-1} x = \frac{dy}{dx} = \frac{-1}{\sin y} = \frac{-1}{\sqrt{1-\cos^2 y}} = \frac{-1}{\sqrt{1-x^2}}.$$

This gives the integration formula

$$(8.358)\qquad \int \frac{1}{\sqrt{1-x^2}}\,dx = -\cos^{-1} x + c \qquad (|x| < 1)$$

which, because $\sin^{-1} x + \cos^{-1} x = \pi/2$, does not contradict (8.354).

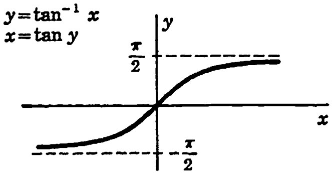

Figure 8.36

As Figure 8.171 may suggest, the inverse tangent is the inverse of the function f_1, with domain $-\pi/2 < x < \pi/2$ and range $-\infty < x < \infty$, for which $f_1(x) = \tan x$. The graph of $y = \tan^{-1} x$, which is the same as a part (sometimes called the principal part) of the graph of $x = \tan y$, is shown in Figure 8.36. The domain of the inverse tangent is the entire set of numbers, and if $y = \tan^{-1} x$, then $\tan y = x$ and $\sec^2 y(dy/dx) = 1$, so

$$(8.361)\quad \frac{d}{dx}\tan^{-1} x = \frac{dy}{dx} = \frac{1}{\sec^2 y} = \frac{1}{1+\tan^2 y} = \frac{1}{1+x^2}.$$

This gives the important formula

$$(8.362)\qquad \int \frac{1}{1+x^2}\,dx = \tan^{-1} x + c.$$

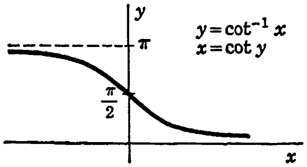

Figure 8.37

The inverse cotangent is the inverse of the function f_1, with domain $0 < x < \pi$ and range $-\infty < y < \infty$. A graph of $y = \cot^{-1} x$ is shown in Figure 8.37. The domain of the inverse cotangent is the entire set of real numbers and if $y = \cot^{-1} x$, then $\cot y = x$ and $-\csc^2 y(dy/dx) = 1$, so

$$(8.371) \qquad \frac{d}{dx}\cot^{-1} x = \frac{dy}{dx} = \frac{-1}{\csc^2 y} = \frac{-1}{1 + \cot^2 y} = \frac{-1}{1 + x^2}.$$

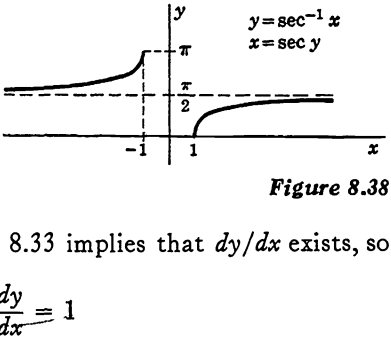

Figure 8.38

As Figure 8.233 indicates, the graph of $y = \sec x$ presents a difficulty that has not previously appeared. It is impossible to select an interval D_1 of values of x such that $\sec x$ is defined and continuous over D_1 and each y in the range of $\sec x$ is attained for some x in D_1. The best we can do is let D_1 be the interval $0 \leqq x \leqq \pi$ with the center point $x = \pi/2$ omitted. The inverse secant is then the inverse of the function f_1 with domain D_1 for which $f_1(x) = \sec x$. The graph of $y = \sec^{-1} x$, which coincides with a part of the graph of $x = \sec y$, is shown in Figure 8.38. If $x > 1$ or $x < -1$ and $y = \sec^{-1} x$, then $\sec y = x$, $0 < y < \pi$, $y \neq \pi/2$, and $\sec y \tan y > 0$. Again Theorem 8.33 implies that dy/dx exists, so

$$\sec y \tan y \frac{dy}{dx} = 1$$

and

$$(8.381) \qquad \frac{d}{dx}\sec^{-1} x = \frac{dy}{dx} = \frac{1}{\sec y \tan y} = \frac{1}{|\sec y|\,|\tan y|}$$
$$= \frac{1}{|\sec y|\sqrt{\sec^2 y - 1}} = \frac{1}{|x|\sqrt{x^2 - 1}}.$$

This gives the formula

$$(8.382) \qquad \int \frac{1}{|x|\sqrt{x^2 - 1}}\,dx = \sec^{-1} x + c,$$

which is valid when $x > 1$ and also when $x < -1$.

The graph of $y = \csc x$ presents the same difficulty that the graph of $y = \sec x$ presented. The best we can do is let D_1 be the interval $-\pi/2 \leqq x \leqq \pi/2$ with the center point $x = 0$ omitted. The inverse cosecant is then the inverse of the function f_1 with domain D_1 for which $f_1(x) = \csc x$. The graph of $y = \csc^{-1} x$, which coincides with a part of the graph of $x = \csc y$, is shown in Figure 8.383. If $x > 1$ or $x < -1$ and $y = \csc^{-1} x$, then $\csc y = x$, $-\pi/2 < y < \pi/2$, $y \neq 0$, and hence $\csc y \cot y > 0$. Again Theorem 8.33 implies that dy/dx exists, so

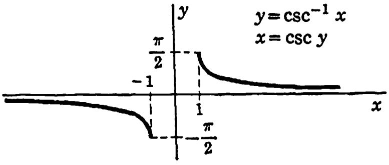

Figure 8.383

$$-\csc y \cot y \frac{dy}{dx} = 1$$

and

$$(8.384) \quad \frac{d}{dx} \csc^{-1} x = \frac{dy}{dx} = \frac{-1}{\csc y \cot y} = \frac{-1}{|\csc y|\,|\cot y|}$$
$$= \frac{-1}{|\csc y| \sqrt{\csc^2 y - 1}} = \frac{-1}{|x| \sqrt{x^2 - 1}}.$$

Problems 8.39

1 Supposing that $a > 0$ and $b^2 < 4ac$, show how the first of the formulas

$$(1) \quad \int \frac{1}{1 + u^2}\, du = \tan^{-1} u + c, \qquad \int \frac{1}{a^2 + u^2}\, du = \frac{1}{a} \tan^{-1} \frac{u}{a} + c$$

$$(2) \quad \int \frac{1}{ax^2 + bx + c}\, dx = \frac{2}{\sqrt{4ac - b^2}} \tan^{-1} \frac{2ax + b}{\sqrt{4ac - b^2}} + c$$

can be used to obtain the other two. *Remark:* Everyone should know how the last of these formulas and many similar ones are presented in standard integral tables. Let $a > 0$ and let

$$X = ax^2 + bx + c, \qquad q = b^2 - 4ac.$$

Then (2) takes the form

$$(3) \quad \int \frac{1}{X}\, dx = \frac{2}{\sqrt{-q}} \tan^{-1} \frac{2ax + b}{\sqrt{-q}} + c$$

when $q < 0$ and hence $-q > 0$. The derivation of (3) is based upon the identity

$$(4) \quad X = a\left[x^2 + \frac{b}{a}x + \frac{c}{a}\right] = a\left[\left(x^2 + \frac{b}{a}x + \frac{b^2}{4a^2}\right) + \left(\frac{c}{a} - \frac{b^2}{4a^2}\right)\right]$$
$$= a\left[\left(x + \frac{b}{2a}\right)^2 + \left(\frac{\sqrt{4ac - b^2}}{2a}\right)^2\right] = a\left[\left(x + \frac{b}{2a}\right)^2 + \left(\frac{\sqrt{-q}}{2a}\right)^2\right]$$

which involves completing a square. It may be time to renew the idea that normal persons who have forgotten (or never remembered) needed integral formulas either work them out to satisfy their vanities and preserve their mathematical powers or stoop to hunting them in a book containing a table of integrals.

2 Show that

$$\int_0^y \frac{1}{y^2 + x^2}\,dx = \frac{\pi}{4y}$$

when $y > 0$ and that

$$\int_1^a \left\{\int_0^y \frac{1}{y^2 + x^2}\,dx\right\} dy = \frac{\pi \log a}{4}$$

when $a > 0$.

3 Supposing that $-1 < x < 1$ and

$$T_n(x) = 2^{1-n} \cos (n \cos^{-1} x), \tag{1}$$

find formulas for $T_n'(x)$ and $T_n''(x)$. Then multiply $T_n(x)$ by n^2, multiply $T_n'(x)$ by $-x$, multiply $T_n''(x)$ by $(1 - x^2)$, and add the results to discover that

$$(1 - x^2)\frac{d^2y}{dx^2} - x\frac{dy}{dx} + n^2y = 0 \tag{2}$$

when $y = T_n(x)$. *Remark:* This problem involves much more than meets the eye. When n is a positive integer, T_n is the Tchebycheff polynomial of degree n and (2) is the Tchebycheff differential equation of order n. One who wishes a simple and brief discussion of these astonishing things can find it in the author's textbook "Differential Equations," 2d ed , McGraw-Hill Book Company, Inc., New York, 1960.

4 Supposing that $-1 < x < 1$ and $u = (\sin^{-1} x)^2$, calculate the first two derivatives of u and show that

$$(1 - x^2)\frac{d^2u}{dx^2} - x\frac{du}{dx} = 2.$$

5 The eye of a man is at a point E that is, as in Figure 8.391, x feet from a vertical wall bearing a picture the bottom and top of which are a and b feet above eye level. The angle θ which the picture subtends at the eye is surely small when x is small and when x is large. Derive the formulas

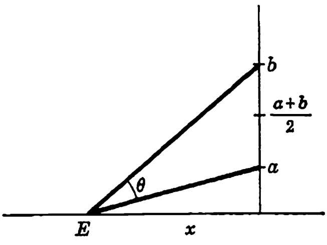

Figure 8.391

$$\theta = \tan^{-1}\frac{b}{x} - \tan^{-1}\frac{a}{x}, \qquad \frac{d\theta}{dx} = \frac{(b - a)(ab - x^2)}{(a^2 + x^2)(b^2 + x^2)}$$

and show that θ is a maximum when $x = \sqrt{ab}$. *Remark:* Addicts of elementary plane geometry can be pleased to see that the x which maximizes θ can be found in another way. As in Figure 8.392 let C be a circle which passes through the top and bottom of the picture and intersects the eye-level line L at two points E_1 and E_2. The angles which the picture subtends at E_1 and E_2 are equal. Of all such circles C that intersect L, the smallest one that produces the greatest θ is the circle C_0 for which E_1 and E_2 coincide, so C_0 is tangent to L. The radius of C_0 is $(b + a)/2$, and use of an appropriate right triangle shows that the distance from the wall to the center and to the critical point of tangency of C_0 is

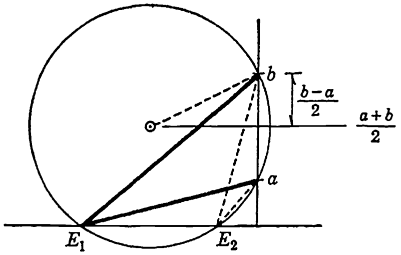

Figure 8.392

$$\sqrt{\left(\frac{b+a}{2}\right)^2 - \left(\frac{b-a}{2}\right)^2} \quad \text{or} \quad \sqrt{ab}.$$

6 Supposing that $a > 0$ and $|ax| < 1$, prove the formula

$$\tan^{-1} x + \tan^{-1} a = \tan^{-1} \frac{a+x}{1-ax}$$

by showing that the two members of the equation are equal when $x = 0$ and have equal derivatives when $|ax| < 1$.

7 With or without undertaking to prove the fundamental fact that $\cot^{-1} x = \tan^{-1} (1/x)$ when $x > 0$, suppose that $x \neq 0$ and fill in the omitted steps in the calculation

$$\frac{d}{dx} \cot^{-1} x = \frac{d}{dx} \tan^{-1} \frac{1}{x} = \cdots = \frac{-1}{1+x^2}.$$

8 With or without undertaking to prove the fundamental fact that $\sec^{-1} x = \cos^{-1} (1/x)$ when $|x| \geqq 1$, suppose that $|x| > 1$ and fill in the omitted steps in the calculation

$$\frac{d}{dx} \sec^{-1} x = \frac{d}{dx} \cos^{-1} \frac{1}{x} = \cdots = \frac{1}{|x| \sqrt{x^2 - 1}}.$$

Hint: Do not forget the chain rule and the fact that $x^2 = |x|\,|x|$.

9 With the aid of Figure 8.343 we can see that the formula

$$\sec^{-1} x = \sin^{-1} \frac{\sqrt{x^2-1}}{x} \tag{1}$$

is valid at least when $x > 1$. Differentiate the two members of (1) and determine the set of numbers x for which the derivatives are equal.

10 Letting N be a positive integer (which we shall not call "arbitrary"), manufacture and solve N problems of the nature of the preceding one.

11 Prove that

$$\int_0^1 \frac{1}{1+x^2}\,dx = \frac{\pi}{4}, \qquad \int_0^\infty \frac{1}{1+x^2}\,dx = \frac{\pi}{2}.$$

12 *Remark:* We should not be so busy that we never have time to look at our formulas and see that we can learn things from them. Supposing for simplicity that $x > 0$, we can look at the formula

$$\tan^{-1} x = \int_0^x \frac{1}{1+t^2}\,dt \tag{1}$$

and realize that each member is equal to the area of the shaded region in Figure 8.393. Any information we can obtain about $\tan^{-1} x$ or the area or the integral therefore provides information about all three. While there are much more complicated ways of getting the information, we can quickly learn some significant facts by starting with the formula

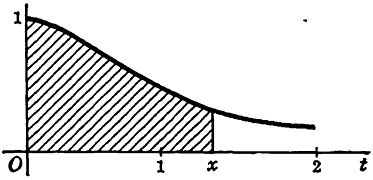

Figure 8.393

$$\frac{1}{1+t^2} = 1 - t^2 + t^4 - t^6 + \cdots + t^{4n} - \frac{t^{4n+2}}{1+t^2} \tag{2}$$

which can be obtained by long division and can be checked by multiplication by $(1 + t^2)$. Integrating this over the interval $0 \leqq t \leqq x$ gives the formula

$$\tan^{-1} x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots + \frac{x^{4n+1}}{4n+1} - R_n(x), \tag{3}$$

where

$$R_n(x) = \int_0^x \frac{t^{4n+2}}{1+t^2}\,dt. \tag{4}$$

Suppose now that $0 \leqq x \leqq 1$. Then $1 \leqq 1 + t^2 \leqq 2$ and therefore

$$\int_0^x \frac{t^{4n+2}}{2}\,dt \leqq R_n(x) \leqq \int_0^x t^{4n+2}\,dt \tag{5}$$

so

$$\frac{x^{4n+3}}{2(4n+3)} \leqq R_n(x) \leqq \frac{x^{4n+3}}{4n+3}. \tag{6}$$

This gives us an excellent estimate of $R_n(x)$ which implies the more crude estimate

$$0 \leqq R_n(x) \leqq \frac{1}{4n+3}. \tag{7}$$

This and the sandwich theorem imply that $R_n(x) \to 0$ as $n \to \infty$, and the definition involving (5.622) shows that

$$\tan^{-1} x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \frac{x^9}{9} - \frac{x^{11}}{11} + \cdots \tag{8}$$

when $0 < x \leqq 1$. Changing the sign of x changes the signs of both sides and shows that the formula is correct when $-1 \leqq x \leqq 1$. Putting $x = 1$ gives the famous Leibniz formula

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \frac{1}{11} + \cdots, \tag{9}$$

which is of particular interest to those who have not previously seen a formula from which π could be calculated. Putting $x = 1$ in (6) shows that we must take many terms of the series (8) to obtain a sum agreeing with $\pi/4$ to three or four decimal places, and we say that the series "converges slowly." The series in (8) converges more rapidly when x is nearer 0. We are not now in the computing business, but it is easy to verify that

$$\frac{\pi}{4} = \tan^{-1}\frac{1}{2} + \tan^{-1}\frac{1}{3}, \tag{10}$$

by taking tangents of both sides, and to use (8) to show that π is roughly 3.14. The formula of Machin (1680–1751)

$$\frac{\pi}{4} = 4\tan^{-1}\frac{1}{5} - \tan^{-1}\frac{1}{239}, \tag{11}$$

which is not so easily proved, is used by professional computers.

13 Find whether the function f for which $f(x) = x + \sin x$ has an inverse and, if so, whether the inverse is differentiable.

14 Let E be an interval of values of x; it could be the interval $x > 0$, and it could be the interval $-1 \leqq x \leqq 1$. Let D be the operator or transformer or differentiator which, when applied to a function f which is defined and differentiable over E, produces the function ϕ for which $\phi(x) = f'(x)$ when x is in E. The domain Δ of D is then the set of functions f that are defined and differentiable over E, and the range R of D is the set of functions ϕ that are defined over E and are derivatives of functions differentiable over E. Show that D does not possess an inverse.

15 Let E be the interval $-1 < x < 1$. Let D_1 be the restriction of the operator D of the preceding problem to the domain Δ_1 consisting of functions f defined over E for which $f(0) = 0$ and f' exists and is continuous over E. Show that D_1 has an inverse.

16 Let E be the interval $-1 < x < 1$. Let I_1 be the operator or transformer or integrator which, when applied to a function g which is continuous over E, produces the function G for which

$$G(x) = \int_0^x g(t)\,dt$$

when x is in E. Describe the domain and range of I_1 and show that I_1 has an inverse.

17 Try to understand and even prove the statement that each of the operators D_1 and I_1 (of the two preceding problems) is the inverse of the other.

8.4 Integration by trigonometric and other substitutions

We begin with a statement of the fact that there is an elementary function F whose derivative with respect to x is $\sqrt{a^2 - x^2}$. Letting F denote one such function, we try to learn about F by writing

$$F(x) = \int \sqrt{a^2 - x^2}\, dx \tag{8.41}$$

and searching for an idea. Once upon a time somebody discovered that if we substitute $x = a \sin \theta$, then the integrand will become $\sqrt{a^2 - a^2 \sin^2 \theta}$ or $\sqrt{a^2 \cos^2 \theta}$, and this is $a \cos \theta$ if $a \cos \theta > 0$. Thus a trigonometric substitution removes the radical and leaves $\pm a \cos \theta$, but we can still be unsure of the meaning of $\int \cos \theta\, dx$. Hence we must pause to make an observation.

The chain rule of Theorem 3.65 tells us that if

$$F'(x) = f(x) \tag{8.42}$$

and if u is a differentiable function whose range lies in the domain of F, then

$$\begin{aligned} \frac{d}{dt} F(u(t)) &= F'(u(t))u'(t) \\ &= f(u(t))u'(t). \end{aligned} \tag{8.421}$$

This gives the following important substitution theorem which shows how to replace x by $u(t)$ in integrals.

Theorem 8.43 *If*

$$F(x) = \int f(x)\, dx \tag{8.431}$$

and if u is a differentiable function whose range lies in the domain of f, then

$$F(u(t)) = \int f(u(t))u'(t)\, dt. \tag{8.432}$$

This theorem is used so often that it is worthwhile to be able to get from the right side of (8.431) to the right side of (8.432) in a purely formal way without thinking about the way in which the theorem was proved and the meanings of the formulas. We replace the old integrand $f(x)$ by the new integrand $f(u(t))u'(t)$ and the old dx by the new dt. This seems like a simple ritual, but the factor $u'(t)$ might be forgotten when problems are being solved. It seems to be safer and easier to think of $f(x)$ being replaced by $f(u(t))$ and dx being replaced by $u'(t)\, dt$. If we follow historical precedents, we become carried away by our own enthusiasm and try to eliminate all possibility of overlooking the factor $u'(t)$ by creating the fiction that dx is a number (a bunch of bananas would serve the same purpose) and $u'(t)\, dt$ is the same thing. There is a sense in which this whole business is utterly silly, but it really is a convenience to imagine that dx, $u'(t)$, and dt are three numbers for which $dx = u'(t)\,dt$. When $x = u(t)$, we can differentiate to get $\frac{dx}{dt} = u'(t)$, and the pretense that $\frac{dx}{dt}$ is the quotient of two numbers then enables us to multiply by dt to get the formula $dx = u'(t)\,dt$ that tells us to replace dx by $u'(t)\,dt$. It is possible to try to say more about these matters but, for present purposes, the important thing is that we do not forget the factor $u'(t)$.

Since it is not always convenient to introduce a new symbol, such as u, for the function which carries t into x, we restate the theorem in the following form which can be ignored by those who prefer the first version.

Theorem 8.433 *If*

$$F(x) = \int f(x)\,dx \tag{8.434}$$

and if x is a differentiable function whose range lies in the domain of f, then

$$F(x(t)) = \int f(x(t))x'(t)\,dt \tag{8.435}$$

or

$$F(x(t)) = \int f(x(t))\,\frac{dx}{dt}\,dt \tag{8.436}$$

or

$$F(x) = \int f(x)\,dx. \tag{8.437}$$

We give further attention to changes of variables in integrals by proving the following very useful variant of Theorem 8.433 which involves Riemann integrals.

Theorem 8.44 *If f is continuous over the interval $a \leqq x \leqq b$ and if u is a function which has a continuous derivative and is such that $u(t)$ increases from a to b as t increases (or decreases) from α to β, then*

$$\int_a^b f(x)\,dx = \int_\alpha^\beta f(u(t))u'(t)\,dt. \tag{8.441}$$

To prove this theorem, let F be a function whose derivative is f so that $\int f(x)\,dx = F(x) + c$ and

$$\int_a^b f(x)\,dx = F(x)\Big]_a^b = F(b) - F(a). \tag{8.442}$$

The chain formula for derivatives then implies that

$$\int_\alpha^\beta f(u(t))u'(t)\,dt = F(u(t))\Big]_\alpha^\beta = F(u(\beta)) - F(u(\alpha)). \tag{8.443}$$

The required conclusion (8.441) then follows from the fact that α and β are chosen such that $u(\alpha) = a$ and $u(\beta) = b$. Careful attention must be given to the manner in which the new limits of integration are determined. When we set $x = u(t)$, we must determine α such that x is a when t is α and we must determine β such that x is b when $t = \beta$.

We are now ready to attack (8.41) and, since this is the first time we have made a substitution (or changed the variable) in an integral, we proceed with great caution. Supposing that $a > 0$ and that $-a \leqq x \leqq a$, we set

$$F(x) = \int \sqrt{a^2 - x^2}\, dx \tag{8.45}$$

and let $x = a \sin \theta$, where $-\pi/2 \leqq \theta \leqq \pi/2$. Figure 8.451 always helps

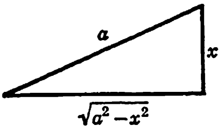

Figure 8.451

us to see what we are doing on occasions like this. Since $dx/d\theta = a \cos \theta$, we obtain

$$F(a \sin \theta) = \int \sqrt{a^2 \cos^2 \theta}\, a \cos \theta\, d\theta. \tag{8.452}$$

But $\cos \theta \geqq 0$ when $-\pi/2 \leqq \theta \leqq \pi/2$, so

$$\begin{aligned} F(a \sin \theta) &= a^2 \int \cos^2 \theta\, d\theta = \frac{a^2}{2} \int (1 + \cos 2\theta)\, d\theta \\ &= \frac{a^2}{2} [\theta + \tfrac{1}{2} \sin 2\theta] + c \\ &= \tfrac{1}{2}[a^2\theta + a^2 \sin \theta \cos \theta] + c. \end{aligned} \tag{8.453}$$

Since each x in the range $-1 \leqq x \leqq 1$ is obtained for a θ in the interval $-\pi/2 < \theta < \pi/2$, we can use Figure 8.451 and the fact that $\cos \theta \geqq 0$ to return to the original variable x so that

$$F(x) = \tfrac{1}{2}\left[a^2 \sin^{-1} \frac{x}{a} + x \sqrt{a^2 - x^2}\right] + c. \tag{8.454}$$

This gives the formula

$$\int \sqrt{a^2 - x^2}\, dx = \tfrac{1}{2}\left[a^2 \sin^{-1} \frac{x}{a} + x \sqrt{a^2 - x^2}\right] + c. \tag{8.455}$$

We are not always so careful about all of the details, and we do not always get correct answers either. There is a reason why we can sometimes be careless about quadrants in which angles lie. There is a "theory of analytic functions" that guarantees that, in many situations, a formula which is correct when angles lie in the first quadrant will be correct wherever we want to use it.

When we wish to make a substitution to evaluate the integral in

$$I = \int \frac{1}{\sqrt{a^2 - x^2}}\, dx, \tag{8.46}$$

we can clarify our work and save writing the integral by denoting it by I, or by I_1 or J if we wish to distinguish it from other integrals. We can then put $x = a \sin \theta$ and allow the variables to shift for themselves while we write

$$\begin{aligned} I &= \int \frac{1}{\sqrt{a^2 \cos^2 \theta}}\, a \cos \theta\, d\theta = \int 1\, d\theta \\ &= \theta + c = \sin^{-1} \frac{x}{a} + c. \end{aligned} \tag{8.461}$$

We have evaluated an integral that previously appeared in (8.354).

The identity $1 - \cos^2 \theta = \sin^2 \theta$ provides the reason why the substitution $x = a \sin \theta$ eliminated the square roots from the integrands of the above integrals. The identities

$$\sec^2 \theta - 1 = \tan^2 \theta, \qquad \tan^2 \theta + 1 = \sec^2 \theta, \tag{8.47}$$

which are obtainable from the one involving sines and cosines by dividing by $\cos^2 \theta$, are less familiar but are nevertheless important when we want to use them. Their uses are illustrated below.

To evaluate the integral in

$$J = \int \frac{1}{(x^2 + a^2)^{3/2}}\, dx, \tag{8.471}$$

we look at it and generate the idea that we should try setting $x = a \tan \theta$. Then $x' = a \sec^2 \theta$, so

$$\begin{aligned} J &= \int \frac{1}{(a^2 \sec^2 \theta)^{3/2}}\, a \sec^2 \theta\, d\theta = \frac{1}{a^2} \int \frac{1}{\sec \theta}\, d\theta \\ &= \frac{1}{a^2} \int \cos \theta\, d\theta = \frac{1}{a^2} \sin \theta + c = \frac{x}{a^2 \sqrt{x^2 + a^2}} + c, \end{aligned} \tag{8.472}$$

the last step being assisted by a figure showing θ, a, and x in the right way.

To evaluate the integral in

$$J_1 = \int \frac{1}{(x^2 - a^2)^{3/2}}\, dx, \tag{8.473}$$

we look at it and generate the idea that we should try setting $x = a \sec \theta$. Then $x' = a \sec \theta \tan \theta$, so

$$\begin{aligned} J_1 &= \int \frac{1}{(a^2 \tan^2 \theta)^{3/2}} a \sec \theta \tan \theta \, d\theta \\ &= \frac{1}{a^2} \int \frac{\sec \theta}{\tan^2 \theta} d\theta = \frac{1}{a^2} \int \frac{\cos \theta}{\sin^2 \theta} d\theta \\ &= \frac{1}{a^2} \int (\sin \theta)^{-2} \cos \theta \, d\theta = \frac{1}{a^2} \frac{(\sin \theta)^{-1}}{-1} + c \\ &= -\frac{1}{a^2} \frac{x}{\sqrt{x^2 - a^2}} + c. \end{aligned} \tag{8.474}$$

It is not implied that we know in advance that the integrand should be written in terms of sines and cosines, but trying this possibility and seeing that the power formula can be applied is all a part of the game.

Our last problem has a lengthy solution. We can bravely start to evaluate the integral in

$$J_2 = \int \sqrt{x^2 - a^2} \, dx \tag{8.48}$$

by setting $x = a \sec \theta$. Then $x' = a \sec \theta \tan \theta$ and

$$J_2 = \int \sqrt{a^2 \tan^2 \theta} \, a \sec \theta \tan \theta \, d\theta$$

so

$$J_2 = a^2 \int \tan^2 \theta \sec \theta \, d\theta. \tag{8.481}$$

It happens that this integral is an elementary function, but there is no simple direct way of discovering the facts. Authors and teachers, as well as students, can work for hours on this problem unless they remember how to solve it or (and this much better) have a guiding principle. The principle is the following. If we want to learn something about an integral and other methods fail to be helpful, we try integration by parts. By this we mean that we try to determine two functions u and v of the variable of integration in such a way that the integral we are studying will be the left member of the formula

$$\int u(\theta) v'(\theta) \, d\theta = u(\theta) v(\theta) - \int v(\theta) u'(\theta) \, d\theta \tag{8.482}$$

and, moreover, the "parts" on the right side enable us to make progress with our problem. Assuming that u and v have continuous derivatives over intervals where we use the formula, differentiation shows that the formula is correct. The trick is to use it effectively. Naturally, we want the functions u, v, u', v' to be as simple as possible, and it is quite easy to discover that the best way to convert J_2 into the left member of (8.482) is to set

$$\left\{ \begin{aligned} u(\theta) &= a^2 \tan \theta, & v'(\theta) &= \sec \theta \tan \theta \\ u'(\theta) &= a^2 \sec^2 \theta, & v(\theta) &= \sec \theta. \end{aligned} \right. \tag{8.483}$$

It is always a good idea to display u and v' in one line and u' and v in a lower line and to know that the formula (8.482) for integration by parts says that the integral of the product of the top two is equal to the product of u and v minus the integral of the product of the bottom two.[†] Thus

$$J_2 = a^2 \sec\theta \tan\theta - a^2\int \sec^3\theta\, d\theta. \tag{8.484}$$

The last integral seems to be quite as recalcitrant as J_2, and this really is true because

$$\begin{aligned} a^2\int \sec^3\theta\, d\theta &= a^2\int \sec^2\theta \sec\theta\, d\theta = a^2\int[1 + \tan^2\theta] \sec\theta\, d\theta \\ &= a^2\int \sec\theta\, d\theta + a^2\int \tan^2\theta \sec\theta\, d\theta \\ &= a^2 \log(\sec\theta + \tan\theta) + c + J_2. \end{aligned}$$

Instead of developing faint hearts by throwing everything into a wastebasket, we substitute our last result into (8.484) and get

$$J_2 = a^2 \sec\theta \tan\theta - a^2 \log(\sec\theta + \tan\theta) - c - J_2. \tag{8.485}$$

If we can now suddenly remember that we are trying to find J_2, we can transpose the term $-J_2$ (or add J_2 to both sides of our equation if transposing is onerous) and divide by 2 to obtain

$$J_2 = \tfrac{1}{2}a^2 \sec\theta \tan\theta - \tfrac{1}{2}a^2 \log(\sec\theta + \tan\theta) + c, \tag{8.486}$$

where the new constant c, like the old $-c/2$, can be any constant. With the aid of a figure showing how θ, a, and x are related, we obtain

$$\begin{aligned} J_2 = \int \sqrt{x^2 - a^2}\, dx &= \tfrac{1}{2}x\sqrt{x^2 - a^2} \\ &\quad - \tfrac{1}{2}a^2 \log \frac{x + \sqrt{x^2 - a^2}}{a} + c. \end{aligned} \tag{8.487}$$

If we wish, we can write the logarithm of the quotient as a difference of logarithms and combine the constant part with the c to obtain the formula

$$\begin{aligned} J_2 = \int \sqrt{x^2 - a^2}\, dx &= \tfrac{1}{2}x\sqrt{x^2 - a^2} \\ &\quad - \tfrac{1}{2}a^2 \log(x + \sqrt{x^2 - a^2}) + c, \end{aligned} \tag{8.488}$$

in which c has another value. The answers in (8.487) and (8.488) look different, but they are both correct. The last one may seem simpler, but the first one gives $c = 0$ if $J_2 = 0$ when $x = a$.

We should recognize the fact that all or nearly all of the integration formulas in the text and problems of this section appear in books of tables. We should hear more than once that books of tables are often used as labor-saving devices, but that many persons like to derive the formulas they use to preserve and improve their mathematical acumen.

[†] These matters are very important, and it is required that we think about them more than once. Our concentrated attack upon integration by parts appears in Section 9.5.

Problems 8.49

1 By making an appropriate figure and trigonometric substitution, fill in all of the missing details which show that

$$(a)\quad \int \frac{1}{\sqrt{x^2 - a^2}}\,dx = \int \sec\theta\,d\theta = \log\,(x + \sqrt{x^2 - a^2}) + c$$

$$(b)\quad \int \frac{1}{a^2 - x^2}\,dx = \frac{1}{a}\int \sec\theta\,d\theta$$
$$= \frac{1}{a}\log\frac{a + x}{\sqrt{a^2 - x^2}} + c = \frac{1}{2a}\log\frac{a + x}{a - x} + c \qquad (|x| < a)$$

$$(c)\quad \int \frac{1}{x^2 - a^2}\,dx = \frac{1}{a}\int \csc\theta\,d\theta$$
$$= -\frac{1}{a}\log\frac{x + a}{\sqrt{x^2 - a^2}} + c = \frac{1}{2a}\log\frac{x - a}{x + a} + c \qquad (x > a)$$

$$(d)\quad \int (a^2 - x^2)^{3/2}\,dx = a^4\int \cos^4\theta\,d\theta$$

2 When $x > 0$, the identity

$$(1)\qquad \sqrt{1 + \frac{1}{x^2}} = \frac{\sqrt{x^2 + 1}}{x} = \frac{x^2 + 1}{x\sqrt{1 + x^2}}$$

shows that

$$(2)\qquad \int \sqrt{1 + \frac{1}{x^2}}\,dx = \int (1 + x^2)^{-1/2}x\,dx + \int \frac{1}{x\sqrt{1 + x^2}}\,dx.$$

The first integral on the right is easily evaluated, and the second can be simplified by a trigonometric substitution. Use these ideas to derive the formula

$$(3)\qquad \int \sqrt{1 + \frac{1}{x^2}}\,dx = \sqrt{1 + x^2} + \log x - \log\,(1 + \sqrt{1 + x^2}) + c.$$

3 Sketch graphs of $y = e^x$ and $y = \log x$ in a single figure and note that each is the mirror image of the other in the line $y = x$. Observe that the arc C_1 on the graph of $y = e^x$ which joins two points (p_1,q_1) and (p_2,q_2) on the graph of $y = e^x$ is congruent to the arc C_2 on the graph of $y = \log x$ which joins the two points (q_1,p_1) and (q_2,p_2) on the graph of $y = \log x$. Letting L denote the length of C_1 and C_2, show that

$$(1)\qquad L = \int_{p_1}^{p_2} \sqrt{1 + e^{2x}}\,dx = \int_{e^{p_1}}^{e^{p_2}} \sqrt{1 + \frac{1}{x^2}}\,dx.$$

This gives the formula

$$(2)\qquad \int_{p_1}^{p_2} \sqrt{1 + e^{2t}}\,dt = \int_{e^{p_1}}^{e^{p_2}} \sqrt{1 + \frac{1}{x^2}}\,dx.$$

To show that errors and misprints have not led us to an incorrect formula, show that the substitution $t = \log x$ converts the first integral in (2) into the second integral in (2).

4 In Problem 2 of Problems 4.39, we called attention to the important nonelementary beta integral formula

$$(1) \qquad \int_0^1 x^p(1-x)^q\,dx\,\frac{p!q!}{(p+q+1)!}$$

which is correct when the integral is a Riemann integral and p and q are nonnegative; we can now report that it is correct when the integral is a Riemann-Cauchy integral and p and q exceed -1. Show that putting $x = \sin^2\theta$, where $0 \leqq \theta \leqq \pi/2$ when $0 \leqq x \leqq 1$, yields the formula

$$(2) \qquad \int_0^{\pi/2} \sin^{2p+1}\theta\cos^{2q+1}\theta\,d\theta = \frac{1}{2}\frac{p!q!}{(p+q+1)!}.$$

Be sure to observe the fact that the limits of integration are correctly determined; setting $x = 0$ in formulas involving x is equivalent to setting $\theta = 0$ in formulas involving θ, and setting $x = 1$ in formulas involving x is equivalent to setting $\theta = \pi/2$ in formulas involving θ. Finally, replace θ by x and let $2p + 1 = \alpha$, $2q + 1 = \beta$ to obtain the formula

$$(3) \qquad \int_0^{\pi/2} \sin^\alpha x\cos^\beta x\,dx = \tfrac{1}{2}\frac{\left(\dfrac{\alpha-1}{2}\right)!\left(\dfrac{\beta-1}{2}\right)!}{\left(\dfrac{\alpha+\beta}{2}\right)!},$$

which is valid when α and β exceed -1. *Remark:* If we happen to know that $(-\frac{1}{2})! = \sqrt{\pi}$, we can conclude from (3) that

$$(4) \qquad \int_0^{\pi/2} \sin^\alpha x\,dx = \sqrt{\frac{\pi}{2}}\frac{\left(\dfrac{\alpha-1}{2}\right)!}{\left(\dfrac{\alpha}{2}\right)!}.$$

5 Replace x by y^2 in the first formula of the preceding problem and then replace y by x to obtain the formula

$$\int_0^1 x^{2p+1}(1-x^2)^q\,dx = \frac{1}{2}\frac{p!q!}{(p+q+1)!}.$$

Tell why the limits of integration are correct.

6 When we proved the formulas for derivatives of trigonometric functions, we were unwilling to base the proofs upon the formula $A = \frac{1}{2}r^2\theta$ for the area of a sector of a circle having central angle θ and radius r. Supposing that $0 < \theta < \pi/2$, this problem requires that the formula be proved. With the aid of an appropriate figure, show that

$$A = \int_0^r f(x)\,dx$$

where

$$f(x) = x\tan\theta \qquad (0 \leqq x \leqq r\cos\theta)$$
$$f(x) = \sqrt{r^2 - x^2} \qquad (r\cos\theta \leqq x \leqq r).$$

Then find A with the aid of the formula (8.455) which has been proved.

7 It is easy to evaluate the integral

$$K_1 = \int x \sin x \, dx$$

with the aid of the formula

$$\int u(x)v'(x)\,dx = u(x)v(x) - \int v(x)u'(x)\,dx$$

for integration by parts. Do it by filling in the first and then the second row of the formulas

$$\begin{aligned} u(x) &= \qquad & v'(x) &= \\ u'(x) &= \qquad & v(x) &= \end{aligned}$$

in the most useful way. Check the result by differentiation.

8 Do not forget our guiding principle. If you have an idea other than integration by parts for evaluating the integral

$$K_2 = \int x^2 \sin x \, dx,$$

write the author a letter. Otherwise, develop a strong heart by solving the problem by integration by parts.

9 When we get good basic ideas we can understand and sometimes even originate modifications of them. If Mr. Watson asks us to evaluate the integral

$$L = \int \frac{1}{\sqrt{x+1}+1}\,dx,$$

we could eliminate the radical by setting $x = \tan^2 \theta$, but the result is unlovely. We can try to simplify matters by making a substitution $x = u(t)$ so cleverly devised that $\sqrt{x+1} = t$. The denominator will then be simply $t + 1$ and the other details may be of such a nature that we can find L. Show that

$$L = 2\sqrt{x+1} - 2\log(\sqrt{x+1}+1) + c.$$

Hint: The simple identity

$$\frac{t}{t+1} = \frac{t+1-1}{t+1} = 1 - \frac{1}{t+1}$$

enables us to integrate the first member in a hurry.

8.5 Integration by substituting $z = \tan \frac{\theta}{2}$

This section provides useful experience in changing variables by treating a particular substitution which is sometimes useful, but the section can be omitted without damaging understanding of the remainder of the book.

We begin by deriving formulas which enable us to use the substitution (or change of variable) $z = \tan \frac{\theta}{2}$ to convert integrals of quotients of polynomials in $\sin \theta$, $\cos \theta$, $\tan \theta$, $\cot \theta$, $\sec \theta$, and $\csc \theta$ into integrals of quotients of polynomials in z. We suppose that $-\pi < \theta < \pi$, so that $-\pi/2 < \theta/2 < \pi/2$, and define z in terms of θ by the formula

$$z = \tan\frac{\theta}{2}. \tag{8.51}$$

Differentiation gives

$$\frac{dz}{d\theta} = \frac{1}{2}\sec^2\frac{\theta}{2} = \frac{1}{2}\left(1 + \tan^2\frac{\theta}{2}\right) = \frac{1+z^2}{2},$$

and hence

$$d\theta = \frac{2}{1+z^2}\,dz. \tag{8.52}$$

Basic trigonometric identities and (8.51) give

$$\sin\theta = 2\sin\frac{\theta}{2}\cos\frac{\theta}{2} = \frac{2\tan\frac{\theta}{2}}{\sec^2\frac{\theta}{2}} = \frac{2z}{1+z^2}, \tag{8.531}$$

$$\begin{aligned}\cos\theta &= \cos^2\frac{\theta}{2} - \sin^2\frac{\theta}{2} = 2\cos^2\frac{\theta}{2} - 1\\ &= \frac{2}{\sec^2\frac{\theta}{2}} - 1 = \frac{2}{1+z^2} - 1 = \frac{1-z^2}{1+z^2}\end{aligned} \tag{8.532}$$

and, except when θ is $\pi/2$ or $-\pi/2$,

$$\tan\theta = \frac{2\tan\frac{\theta}{2}}{1 - \tan^2\frac{\theta}{2}} = \frac{2z}{1-z^2}. \tag{8.533}$$

In fact, Figures 8.541 and 8.542 enable us to write the six trigonometric functions of $\theta/2$ and θ in terms of z. Thus each of $\sin\theta$, $\cos\theta$, $\tan\theta$, $\cot\theta$, $\sec\theta$, $\csc\theta$, and $d\theta/dz$ is a quotient of polynomials in z. It follows from this that if P and Q are polynomials in the six trigonometric functions of x, then

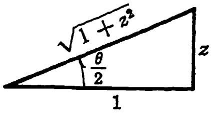

Figure 8.541

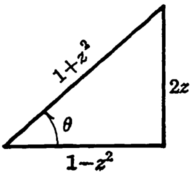

Figure 8.542

$$\int \frac{P}{Q}\,d\theta = \int \frac{P_1}{Q_1}\,dz,$$

where P_1 and Q_1 are polynomials in z. Several examples appear in the problems at the end of this section.

Except in special cases, evaluating integrals of the form

$$\int \frac{a_0 z^m + a_1 z^{m-1} + \cdots + a_{m-1} z + a_m}{b_0 z^n + b_1 z^{n-1} + \cdots + b_{n-1} z + b_n}\, dz$$

is very tedious business. When $n \leqq 2$, and in a few special additional cases, answers can be found in tables of integrals. Some information about the matter will appear in Section 9.4 of this book.

Problems 8.59

1 By use of the substitution $z = \tan \frac{\theta}{2}$, show that

(*a*) $\displaystyle\int \frac{1}{\sin\theta}\, d\theta = \int \frac{1}{z}\, dz = \log |z| + c = \log \left| \tan\frac{\theta}{2} \right| + c$

(*b*) $\displaystyle\int \frac{1}{\cos\theta}\, d\theta = \int \frac{2}{1 - z^2}\, dz = \int \frac{1 - z + 1 + z}{(1 - z)(1 + z)}\, dz$

$$= \int \left(\frac{1}{1 + z} + \frac{1}{1 - z} \right) dz = \log \left| \frac{1 + \tan\frac{\theta}{2}}{1 - \tan\frac{\theta}{2}} \right| + c$$

(*c*) $\displaystyle\int \frac{1}{a + b\sin\theta}\, d\theta = \int \frac{2}{az^2 + 2bz + a}\, dz$

(*d*) $\displaystyle\int \frac{1}{a + b\sin\theta + c\cos\theta}\, d\theta = \int \frac{2}{(a - c)z^2 + 2bz + (a + c)}\, dz$

(*e*) $\displaystyle\int \frac{\sin\theta + \cos\theta}{\tan\theta + \cot\theta}\, d\theta = 4 \int \frac{z(1 - z^2)(1 + 2z - z^2)}{(1 + z^2)^4}\, dz$

(*f*) $\displaystyle\frac{2z}{1 + z^2} + c = \sin\theta + c = \int \cos\theta\, d\theta = 2 \int \frac{1 - z^2}{(1 + z^2)^2}\, dz$

(*g*) $\displaystyle\log \frac{1 + z^2}{1 - z^2} + c = \log \sec\theta + c = \int \tan\theta\, d\theta = 4 \int \frac{z}{1 - z^4}\, dz$

(*h*) $\displaystyle\log \left(a + \frac{2bz}{1 + z^2} \right) + c = \log (a + b\sin\theta) + c$

$$= \int \frac{b\cos\theta}{a + b\sin\theta}\, d\theta$$

$$= 2b \int \frac{1 - z^2}{(az^2 + 2bz + a)(1 + z^2)}\, dz$$

(*i*) $\displaystyle\int \frac{\sin\theta}{1 + \sin\theta}\, d\theta = \int \frac{4z}{(1 + z^2)(1 + z)^2}\, dz$

(*j*) $\displaystyle\int \frac{\cos\theta}{1 + \cos\theta}\, d\theta = \int \frac{1 - z^2}{1 + z^2}\, dz$

$$= \int \frac{2 - 1 - z^2}{1 + z^2}\, dz = 2\tan^{-1}\left(\tan\frac{\theta}{2} \right) - \tan\frac{\theta}{2} + c$$

Chapter 6

Plane isometries and similarities

With this chapter you leave classical geometry, which concentrated on properties of figures constructed from points, lines, planes, circles, and spheres. You enter transformational geometry, whose fundamental objects are points and transformations. A transformation is a function that relates all points to others in a one-to-one way; it induces a correspondence between figures as well. You can use transformations to describe correspondences between various figures of classical geometry, to study positions of figures before and after motions, and to analyze the relationships between corresponding parts of symmetric figures. You can adapt transformational geometry to study aspects of a single figure that may differ when you use different coordinate systems. You'll investigate how figures change under various transformations; and —perhaps more important—you'll study figures and properties that *don't* change.

The transformations you'll study here—isometries and similarities—are closely related to the classical geometric notions of congruence and similarity. Isometries are transformations that preserve distance between points.[1] An isometry can change a triangle's location or its orientation in a plane. By the SSS congruence principle, though, an isometry cannot change a triangle's size or shape. In fact, two triangles are congruent *just in case* they're related by some isometry. This result doesn't lead to much new in the theory of *triangle* congruence, but does provide an easy way to extend the theory to figures of *any* type. You don't have to invent different congruence notions for each type, but can just call figures *congruent* if they're related by some isometry. The connection between the theory of similar figures and the study of similarity transformations, which multiply all distances by a constant ratio, is analogous.

---

[1] The word stems from the Greek words *isos* and *metron* for *equal* and *measure*.

Formulation of a general theory of congruent figures is the first of many uses of transformations in the foundations of geometry. They provide a tool for determining what's common between the geometric theories that result from various alternative axiom systems—for example, from assuming the negation of Euclid's parallel axiom. Those studies lie beyond the scope of this text, but not far![2]

The correspondence between locations of objects before and after a motion is clearly an isometry. This connection is so suggestive that many authors use the term *motion* instead of *isometry*. But isometry doesn't encompass everything involved in motion; it ignores what happens *during* the event. That requires additional apparatus not included in transformational geometry, so *motion* is a misnomer here.

The connection between transformations and symmetry is perhaps not so obvious. We exhibit *bilateral* symmetry because there's a close correspondence between points on our left and right sides. Some features are preserved; we can easily recognize an acquaintance's mirror image. Most familiar animals appear bilaterally symmetric. About a thousand years ago, an artist in the Mimbres region of New Mexico exploited that when she portrayed the dancer in figure 6.0.1, costumed to evoke simultaneously images of a deer, a man, and a bat. But many features of life are not symmetric. For example, it's hard for us to recognize the mirror image of text in figure 6.0.2. In fact, symmetry is defined in terms of isometry: a figure is symmetric if it's unchanged under some isometry. A plane figure is bilaterally symmetric if it's unchanged when you reverse the two sides of some line.

The symmetric frieze ornament in figure 6.0.3, a pottery design from the San Ildefonso Pueblo in New Mexico, is designed to be repeated as required, and remains unchanged as your gaze steps along it. A rotationally symmetric design remains the same if you turn it through a particular angle. For example, the *dharmachakra*, the 24-spoked wheel of right conduct on the flag of India (figure 6.0.4) symbolizes righteousness, truth, and recurrence.[3]

You've tasted transformational geometry if you've studied coordinate transformations in analytic geometry. That application needs a different approach; you're not interested in corresponding points, but rather in corresponding coordinate pairs or triples assigned to a single point by different coordinate systems. This chapter's theory can be adapted to that end. But that's not pursued, because the book's overall emphasis isn't on analytic

---

[2] For an introduction, consult Behnke et al. [1960] 1974, chapters 4 and 5.

[3] According to an official Indian government publication (Sivaramamurti 1966, 8, 14), the *chakra* represents "the Wheel of Law . . . whose message of righteousness was binding even on the greatest monarch. . . . The wheel of life is inexorable . . . . [It] moves on into a cycle of births and deaths, culminating in final liberation . . . . The wheel is thus the symbol of creation."

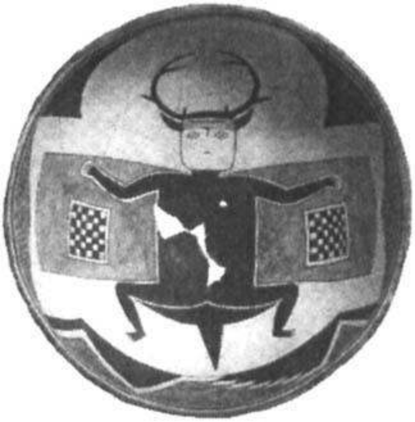

Figure 6.0.1
Bilateral symmetry[4]

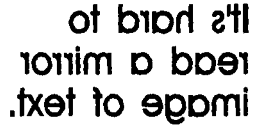

Figure 6.0.2
Bilateral asymmetry

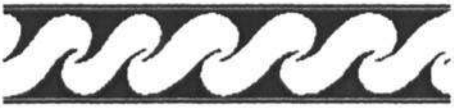

Figure 6.0.3
Frieze pattern[5]

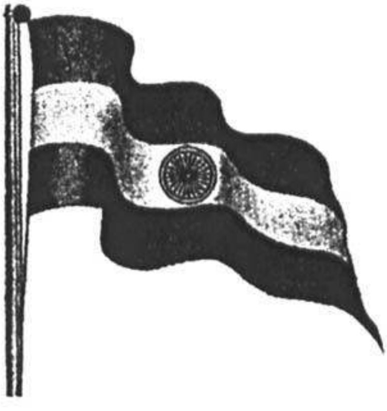

Figure 6.0.4
India's flag

---

[4] From Brody, Scott, and LeBlanc 1983, plate 5. Where did the artist intentionally *break* the overall symmetry? Like most surviving Mimbres pottery, this example was recovered from a grave. Mimbres funeral ceremonies evidently included punching a hole in the bottom of each bowl to be buried.

[5] Kenneth M. Chapman 1970, plate 124p. This example originated not far from the Mimbres, but is probably less than a century old.

methods. Coordinate transformation techniques are demonstrated in exercise 6.11.31.

What methods does this chapter use to study transformations? It must depart from the classical methods of earlier chapters. Many practical applications involve computation with coordinates. That's so cumbersome that it's generally avoided in formulating the theory. The computations involve vectors and matrices, so matrix algebra affords some simplification—you can handle matrices as whole entities instead of manipulating their entries. But even greater simplification results from employing the algebra of compositions and inverses of functions. These are the substance of much of modern algebra, particularly group theory. The more deeply you study transformational geometry, particularly its applications to symmetry, the more benefit you gain by using group-theoretic methods. This chapter is an introduction to those techniques. The theory is developed with minimal use of matrix algebra. The matrix computations typical in applications are demonstrated in the exercises that conclude the chapter.

The chapter begins by collecting in section 6.1 the relevant general information about functions, composition, and inverses. Since this material is fundamental to several areas of higher mathematics, you may have encountered much of it already in other courses. Section 6.2 introduces three types of plane isometries—translations, rotations, and reflections—and some of their properties. Sections 6.3 to 6.5 discuss them in greater detail. The chapter's principal result is the structure theorem in 6.6. That enables you to classify every plane isometry as one of several types. You'll see that the earlier introductory treatment omitted one type: glide reflections, which are handled in section 6.7. The connection with matrix algebra is established in section 6.8, and used in 6.9 to refine the classification.

This chapter overview has only once mentioned the other kind of transformation this book studies: similarities. They're introduced in section 6.10. Their study requires only one new technique. Otherwise, methods already introduced for isometries work as well for similarities; there's just more detail.

The chapter concludes with exercises. These pursue further most of the ideas introduced in the previous sections, and will give you much more experience with computation.

For simplicity, chapter 6 is limited to *plane* transformational geometry. But *all* of it extends to three dimensions. In fact, most applications of plane isometries and similarities were originally handled by specialized methods that were less algebraic, more classical, in flavor. The corresponding three-dimensional results are just enough more complicated that a new apparatus was required to handle them. Logically, the three-dimensional material should be integrated into this chapter, because we really live in three-dimensional space, and its complexities make more apparent the need for the new methods. But that would make the chapter too bulky, it wouldn't be so easy for you to discern its simple structure, and the text would become less flexible for instructors and students. So the three-dimensional results are presented together in chapter 7, with a separate set of exercises. Much of the theory in that chapter is identical to the corresponding plane theory, once the definitions are properly formulated. So chapter 7 is even more streamlined, with many details left to you.

Chapter 2 and occasional passages in other previous chapters have offered some details of the history of the material presented there. It's not too hard to find sources for that, because it was often controversial. Researchers reported the background for their ideas, and cited where they agreed or differed with their predecessors. With transformational geometry, the task isn't so simple. Its techniques evolved over a century and a half, beginning in the late 1700s. They were introduced to solve hard applications problems. Mathematicians concentrated on their successes and on new problems. They rarely dwelled on philosophical, organizational, or pedagogical aspects of their work. The order of development of the material from this chapter on was often opposite to its logical order. Many of the complex techniques for applying the theory were introduced early in this period. Leonhard Euler, in 1775, would have found familiar most of the computational exercises in these chapters.[6] Determinants came into use during the early 1800s, and transformation groups and matrices in the mid-1800s. The last part of this material to fall into place is the first part presented; the language and techniques for dealing with functions, their compositions, inverses, and image sets were refined and cast in the form of section 6.1 only in the early 1900s. The history of transformational geometry is thus blurred. When you try to find the origin of a specific concept or result, you often find traces of it in works all the way back to Euler's time, even though it wasn't stated explicitly until much later. You can't readily use history pedagogically, as punctuation for the presentation of the theory. So there are fewer historical references in these later chapters. Don't let their absence give you the impression that the theory was revealed all at once, or that its history is bland! It's involved and intriguing—a challenging field of study.

---

[6] For example, Euler ([1775] 1968) formulated equations for isometries, even though the matrix algebra used for that purpose in theorems 6.8.2 and 7.1.7 wasn't developed until the late 1800s.

6.1 Transformations

Concepts

Functions; domain, value, image, and range
Linear functions
Preserved and invariant properties of figures
Function composition
Associativity and noncommutativity
Surjections, injections, and bijections
Inverse functions
Transformations
Symmetric group, transformation groups
Geometric transformations

This section summarizes elementary material about functions, on which the rest of this chapter is based. Together with appendices A and C, it forms the algebraic foundation for the entire book. You may have met much of this material already in precalculus and calculus courses. But its application to transformational geometry has a flavor so different from those subjects that the fundamentals are presented here for emphasis. You can regard transformational geometry as a very sophisticated development of these few elementary ideas and some others from chapter 3. For that reason it's important that you understand in complete and precise detail these definitions, examples, and first results. The first four exercises in section 6.11 will help you.

As you study this material in other geometry texts or other application areas, you'll encounter conflicting terminology and notation. Its applications spread over all mathematics, and what's convenient and graceful for one area can appear somewhat inappropriate for another. Moreover, this mathematics is younger than most of that in earlier chapters, and its terminology and notation are still evolving. This book's conventions result from many compromises. Several footnotes indicate its differences from other usage, and some of the reasons for the divergence.

This section concludes by mentioning some topics to which this chapter naturally leads, but which lie just beyond its scope.

Functions

The function concept pervades mathematics. In this book, for example, it has appeared among the undefined concepts and axioms:

- The *distance function* assigns to each pair of points X, Y the distance XY.

- A *scale* for a line g assigns to each point P on g a coordinate $c(P)$.
- The *area* function assigns to each polygonal region Σ its area $a\Sigma$.

Functions also occur in later developments, either implicitly or explicitly. For example,

- A function assigns to each pair P, Q of distinct points the *line* $\overleftrightarrow{PQ}$.
- Each *coordinate system* has a function that assigns to each point X a triple $<x_1, x_2, x_3>$ of coordinates.
- An *angular scale* at a point O in a plane ε assigns to each real number a ray in ε with origin O.

Section 3.13 introduced angular scales to help define $\sin x$ and $\cos x$ as real-valued functions of real numbers x. You've met many functions of that type in algebra. The following examples of *exponential*, *power*, *polynomial*, and *linear* functions are considered in detail later in this section:

$$2^x \quad x^3 \quad x^{1/3} \quad x(x^2 - 1) \quad ax + b \text{ for specified constants } a, b.$$

Also, in two or three dimensions, given matrix and vector constants A and B, you've worked with the *linear* function that assigns to each vector X the vector value $AX + B$.[7]

Here's the formal definition that unifies these examples. A *function* f assigns to each element x of a set D a unique *value* $f(x)$ in a set S. This sentence is often abbreviated $f: D \to S$ and read *f maps D into S*.[8] The uniquely determined set D of x to which f applies is called the *domain* of f. But the set S is *not* uniquely determined; if $f: D \to S$, then f also maps D into *any* set that contains S. Often it's more convenient to refer to a formula such as 2^x for the value $f(x)$ than to refer to f by name. In a case like that, you can refer to *the function $x \to 2^x$ that maps x to* 2^x. Since that doesn't mention the set D or S, you must determine them from the context. In this example, D could be the set $\mathbb{R}$ of all real numbers x, and S could be any set that contains all $x > 0$. When there's no doubt which letter is the bound variable, you can use just the formula and refer to *the function* 2^x. Be careful, though—the phrase *function* $ax + b$ is ambiguous unless you know that a and b are constants.[9]

---

[7] Even though they still call $ax + b$ a linear function, most authors of linear algebra texts reserve the term *linear* in higher dimensions to functions of the form AX. They'd call $AX + B$ an *affine* function.

[8] Authors often use the terms *function* and *mapping* interchangeably.

[9] Many older books that apply advanced algebra to geometry use *postfix* functional notation; instead of $f(x)$ they write $(x)f$. Because it clashes with calculus notation, that practice is disappearing. In this text it survives in usages like *the function* $x \to x'$. But here, the symbol $'$ is not a function name, just part of a formula for the function value.

Two functions f and g are regarded as *equal* if they have the same domain D and $f(x) = g(x)$ for every $x \in D$.

If $f : D \to S$, then to each subset $A \subseteq D$ corresponds the *image set* $f[A] = \{f(x) : x \in A\}$. The image $f[D]$ of the domain is called the *range* of f.

Preserved and invariant properties

In many applications, a function maps its domain D into itself. For example, all the trigonometric and algebraic functions mentioned under the previous heading map the domain $D = \mathbb{R}$ into itself. A function $f : D \to D$ is said to *preserve* a property $\mathscr{P}$ of members of D if $f(x)$ has property $\mathscr{P}$ whenever x does. You can use an image set to describe that situation—if $A = \{x \in D : x \text{ has } \mathscr{P}\}$, then f preserves $\mathscr{P}$ just when $f[A] \subseteq A$. For example, $x \to 2^x$ preserves the property of being an integer, but not that of being an odd integer. $\mathscr{P}$ is called *invariant* under f if $f(x)$ has $\mathscr{P}$ *if and only if* x does: $\mathscr{P}$ is invariant under f just when $f[A] = A$. For example, being positive is invariant under $x \to x^3$ but not under $x \to 2^x$.

You can consider properties of pairs, or sequences $<x_1, x_2, \ldots>$ of any number of elements of D the same way. For example, 2^x and x^3 are called *strictly increasing* functions because they leave *order* invariant:

$$x_1 \le x_2 \leftrightarrow 2^{x_1} \le 2^{x_2} \leftrightarrow x_1^3 \le x_2^3 .$$

On the other hand, $\sin x$ doesn't preserve order; you can have $x_1 \le x_2$ but $\sin x_1 > \sin x_2$.

You can consider properties of subsets or sequences of subsets of D the same way. For example, the property *being a bounded open interval* is invariant under $f(x) = 2^x$; if $a < b$, then

$$f[\{x : a < x < b\}] = \{y : 2^a < y < 2^b\}$$

because $a < x < b$ if and only if $2^a < 2^x < 2^b$. But that property isn't preserved by $\sin x$; for example,

$$\sin[\{x : 0° < x < 360°\}] = \{y : -1 \le y \le 1\},$$

which is a bounded *closed* interval, not an open one.

Function composition

If $f : S \to T$ and $g : D \to S$, then the *composition* of f with g is the function $f \circ g : D \to T$ such that $(f \circ g)(x) = f(g(x))$ for all $x \in D$.[10] For example,

---

[10] Many authors used to define composition backward. They called $g \circ f : x \to f(g(x))$ the composition of $g : D \to S$ with $f : S \to T$, letting the left-to-right order emphasize that you perform g first, then f. As you see, that collides with prefix notation for function values. (continued...)

fg *usually* means their composition, and the equation $fg = gf$ is probably false!

Surjections, injections, and bijections

A function $f : D \to S$ is said to map D *onto* S *surjectively* if $S = f[D]$, the range of f.[12] For this phrase to make sense, you *must* specify S. For example, when $D = S = \mathbb{R}$, functions x^3 and $x(x^2 - 1)$ are surjective but 2^x and $\sin x$ are not. Should you change S to $\{x : x > 0\}$ or $\{x : -1 \le x \le 1\}$, then 2^x or $\sin x$, respectively, would become surjective.

f is said to map D into S *injectively* if $f(x_1) \ne f(x_2)$ whenever $x_1 \ne x_2$.[13] Functions 2^x and x^3 are injective but $x(x^2 - 1)$ and $\sin x$ are not. These functions display all four possible combinations of surjectivity and injectivity!

f is said to map D to S *bijectively* if it's both surjective and injective. The function x^3 is bijective; so is the linear function $ax + b$ unless $a = 0$. A bijective function $f : D \to S$ has an *inverse* $f^{-1} : S \to D$, which maps each $y \in S$ to the unique $x \in D$ for which $f(x) = y$. (Surjectivity means there's at *least* one such x; injectivity means there's at *most* one.) For these examples,

$$y = f(x) = x^3 \Rightarrow x = y^{1/3} \Rightarrow f^{-1}(y) = y^{1/3} \Rightarrow f^{-1}(x) = x^{1/3}$$

$$y = g(x) = ax + b \Rightarrow x = a^{-1}y - a^{-1}b$$
$$\Rightarrow f^{-1}(y) = a^{-1}y - a^{-1}b \Rightarrow f^{-1}(x) = a^{-1}x - a^{-1}b.$$

You derive the formula for the inverse function value by solving equation $y = f(x)$ for x in terms of y, then reversing x and y. Notice that *the inverse of a nonconstant linear function is linear*!

You should verify that the inverse of a bijection $f : D \to S$ is a bijection $f^{-1} : S \to D$, and *its* inverse is the original function f.[14]

---

[12] Authors often write merely *f maps D onto S*, omitting the adverb *surjectively*. Then the subtle distinction between *into* and *onto* becomes too important. Moreover, since no convenient adjective or noun matches the preposition *onto*, they must refer awkwardly to *onto* functions or to the property of *onto-ness*. *Surjective* and *surjectivity* are more graceful. The adjectives *surjective*, *injective*, and *bijective* were coined by a pseudonymous group of French mathematicians for use in Bourbaki 1939, the first installment of a multivolume encyclopedia that reformulated advanced mathematics. *Sur* means *on* in French, and the root *-ject-* comes from the past participle of the Latin verb *iacio*, which means *throw*.

[13] Authors often refer to *one-to-one functions*, employing an adverb where an adjective would be appropriate. Moreover, since no convenient noun matches *one-to-one*, they must use the awkward term *one-to-one-ness*. *Injective* and *injectivity* are more graceful.

[14] Many authors use the term *one-to-one correspondence* for *bijection*. The term *correspondence* emphasizes the symmetric nature of such a relation, whereas *bijection* stresses its asymmetric left-to-right nature. The left-to-right convention became predominant in this context after authors found that symmetric terminology obscures the distinction between a bijection and its inverse. For an example of this awkwardness, consult the research papers by

(continued...)

It's also easy to verify the following propositions about compositions:

- If $f: S \to T$ and $g: D \to S$ are both surjective, then so is $fg: D \to T$.
- If $f: S \to T$ and $g: D \to S$ are both injective, then so is $fg: D \to T$.
- If $f: S \to T$ and $g: D \to S$ are both bijective, then so is $fg: D \to T$, and $(fg)^{-1} = g^{-1}f^{-1}: T \to D$.

Be sure you understand the reversed order of the functions in the previous sentence. If $z = f(g(x))$, then $z = f(y)$, where $y = g(x)$, so that $y = f^{-1}(z)$ and $x = g^{-1}(y) = g^{-1}(f^{-1}(z))$. In general, to undo the result of two or more successive operations, you must undo them all *in reverse order*. Think of donning your stockings, shoes, and overshoes, then removing them!

Exercise 6.11.2 provides an interesting sidelight on the previous paragraph, as follows. Suppose $f: S \to T$ and $g: D \to S$. If fg is surjective, then f must be surjective, but not necessarily g. If fg is injective, then g must be injective, but not necessarily f. If fg is bijective, then f is surjective and g injective, but neither is necessarily bijective. To ensure that you understand the concepts under this heading, complete that exercise, proving the positive clauses and finding counterexamples to justify the others.

Suppose $f: D \to D$ bijectively and $\mathcal{P}$ is a property of elements, subsets, or sequences of elements or subsets of D. It's possible for f but not f^{-1} to preserve $\mathcal{P}$. For example, consider $f(x) = x^3$, $f^{-1}(x) = x^{1/3}$ and the property $x > 2$; that inequality implies $x^3 > 2$ but not $x^{1/3} > 2$. But if $\mathcal{P}$ is *invariant* under f, then it's also invariant under f^{-1}. For example, the property $x_1 > x_2$ is invariant under both x^3 and $x^{1/3}$.

Transformations; identity functions

A bijection $f: D \to D$ is called a *transformation* of D.[15] One bijection, in particular, is ubiquitous: the *identity* function $\iota_D : x \to x$. Note that *the identity function for* $D = \mathbb{R}$ *is linear*: $\iota_{\mathbb{R}}(x) = x = 1x + 0$. It's easy to verify the following general propositions about identity functions:

- If $f: D \to S$, then $f\iota_D = f = \iota_S f$. If f is bijective, then $f^{-1}f = \iota_D$ and $ff^{-1} = \iota_S$. If f is a transformation of D, then $f^{-1}f = \iota_D = ff^{-1}$.
- Suppose $f: D \to S$ bijectively and $g: D \to S$ bijectively. If $fg = \iota_D$ or $gf = \iota_S$, then $g = f^{-1}$ and $f = g^{-1}$.
- *Every* property of elements, subsets, or sequences of elements or subsets of D is invariant under ι_D.

---

[14] (...continued)
Hermann Wiener (1890–1893), which introduced the general approach taken in this chapter.

[15] Authors often use the term *permutation* instead of *transformation*, particularly when the domain is finite.

The symmetric group; transformation groups

The family $\mathcal{S}$ of *all* transformations of D is called the *symmetric group* of D. This chapter is concerned with subfamilies $\mathcal{T}$ of $\mathcal{S}$ that satisfy three requirements:

- the identity ι_D must belong to $\mathcal{T}$;
- $\mathcal{T}$ must contain the inverse of each of its members; and
- $\mathcal{T}$ must contain the composition of any two of its members.

If $\mathcal{T}$ satisfies these conditions, it's called a *transformation group on* D. Thus a subfamily of the symmetric group is a transformation group just in case it contains the identity and is *closed* under inversion and composition. D always has at least two transformation groups: the symmetric group itself, and the *trivial* group that has only one member ι_D. Each clearly satisfies the conditions.

If a transformation f belongs to a transformation group $\mathcal{T}$, so do any n-term compositions $f \circ \cdots \circ f$ and $f^{-1} \circ \cdots \circ f^{-1}$; they're called *powers* of f and written f^n and f^{-n}. f^1 and f^0 stand for f and the identity. You can verify that some laws analogous to elementary algebra hold for this concept. For example, $f^{m+n} = f^m f^n$, $f^{mn} = (f^m)^n$, and $(f^n)^{-1} = f^{-n}$.

When one transformation group is a subset of another, it's called a *subgroup*. For example, the trivial group is a subgroup of every transformation group, and every group is a subgroup of itself. A subgroup of a group $\mathcal{G}$ that's different from $\mathcal{G}$ is called a *proper* subgroup.

Transformation groups are often characterized by invariants: If $\mathcal{P}$ is a property of elements, subsets, or sequences of elements or subsets of D, then the family of all permutations of D under which $\mathcal{P}$ is invariant forms a transformation group on D. This family clearly contains the identity and is closed under inversion and composition.

Members of the symmetric group of two- or three-dimensional space $\mathbb{R}^2$ or $\mathbb{R}^3$ are called *geometric transformations.*[16]

Linear transformations

Among the most useful geometric transformations are the two- and three-dimensional linear functions $X \to AX + B$ for specified matrix and vector constants A and B with invertible A. *The identity function is linear*: $X \to X = IX + O$. Just as in the one-dimensional case, *the composition of linear maps is linear*: If $\varphi(X) = AX + B$ and $\chi(X) = CX + D$, then

---

[16] This term is also used more broadly, for analogous concepts in projective, non-Euclidean, and higher-dimensional geometry, and sometimes for permutations of sets of lines or other figures instead of just points.

$$\varphi\chi(X) = A(CX + D) + B = (AC)X + (AD + B)$$

and $(AC)^{-1} = C^{-1}A^{-1}$. Moreover, *the inverse of a linear function is linear*:

$$\varphi^{-1}(X) = A^{-1}X - A^{-1}B$$

and $(A^{-1})^{-1} = A$. Notice how matrix multiplication and inversion correspond to function composition and inversion! The case $B = D = O$ shows that the noncommutativity of matrix multiplication and of linear function composition are two aspects of the same phenomenon.

Invariance of betweenness

Betweenness is invariant under linear transformations. To verify that, consider a transformation $X \to X' = AX + B$ and points P, Q, and R. You must show that P-Q-R if and only if P'-Q'-R'. First, P-Q-R implies that $Q = tP + (1 - t)R$ for some scalar t such that $0 \le t \le 1$, hence

$$\begin{aligned} Q' &= AQ + B = A(tP + (1 - t)R) + B \\ &= t(AP + B) + (1 - t)(AR + B) = tP' + (1 - t)R', \end{aligned}$$

hence P'-Q'-R'. Thus $X \to X'$ preserves betweenness. Its inverse is also linear, hence preserves betweenness, so P'-Q'-R' implies P-Q-R also.

It follows that any property you can define in terms of betweenness is invariant under linear transformations $X \to X'$ (regardless of whether it was originally defined that way). For example, collinearity is invariant: For any points P, Q, and R,

P, Q, and R are collinear
 if and only if P-Q-R or Q-R-P or R-P-Q, hence
  if and only if P'-Q'-R' or Q'-R'-P' or R'-P'-Q', hence
   if and only if P', Q', and R' are collinear.

Similarly, all these properties are definable in terms of betweenness:

coplanarity
convexity

being
  a segment
  a ray
  a line

being
  a half plane
  a plane
  a half space

being
  an angle
  a triangle
  a tetrahedron

being the interior of
  an angle
  a triangle
  a tetrahedron

Therefore, they're all invariant under linear transformations.

The argument in the previous paragraph is more general than its context. It actually shows that if betweenness is invariant under *any* transformation

φ, then all properties definable in terms of betweenness are also invariant under φ.

Are linear transformations the *only* ones under which betweenness is invariant? The answer is *yes*, but its proof requires analysis methods beyond the scope of this book.[17]

Linear transformations don't necessarily preserve distance or even similarity. For example, under the transformation

$$X \to X' = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} X,$$

the image of ΔOUV is $\Delta O'U'V'$, where

$$\begin{array}{ll} O = <0,0> = O' & V = <0,1> \\ U = <1,0> = U' & V' = <2,1>. \end{array}$$

Notice that $OV \neq O'V'$; and these triangles aren't similar.

The rest of this chapter is a detailed, fairly complete study of those geometric transformations that do preserve distance—called *isometries*—and those that preserve similarity but not necessarily distance—called *similarities*. All isometries and similarities are linear, but not vice-versa. Detailed study of linear transformations that aren't isometries lies beyond the scope of this book.[11]

6.2 Isometries

Concepts

Definition
Physical motions and time are not geometric concepts
Translations
Rotations
Reflections
The isometry group
Noncommutativity
Invariance
Rigidity theorem
Uniqueness theorem

---

[17] Chapter 15 of Martin 1982 is an introduction to that theory in two dimensions. Virtually all the results in *Methods of geometry* about isometries and similarities have both two- and three-dimensional analogues in the theory of general linear transformations. But those are often cast in projective geometric language, and hard to find in the literature. You could start a search with the research paper Ellers 1979 and its references.

This section introduces plane isometries and some related concepts. It presents common examples, but gives few details. Those come later, in sections 6.3 to 6.5 and 6.7, which study in depth four types of isometries. This section and the next are arranged to provide a quick route to the classification theorem, which shows that *every* isometry is one of those four types.

The rest of chapter 6 is concerned with points in a single plane ε. It's always the same plane, so it's never again mentioned explicitly. For example, in the rest of this chapter, *point* means *point in* ε.

A plane isometry $\varphi: P \to P'$ is a transformation of the plane that preserves distance:

(1) to each point P corresponds another, its *image* P';
(2) distinct points $P \neq Q$ have distinct images $P' \neq Q'$;
(3) every point X is the image P' of some point P;
(4) the distance between any points P and Q is the same as that between their images: $PQ = P'Q'$.

Condition (2) is redundant: If $P \neq Q$, then $PQ \neq 0$, so $P'Q' \neq 0$ by (4), hence $P' \neq Q'$. Nevertheless, the definition is usually phrased this way to emphasize that an isometry (1) is a function, (2) is injective, (3) is surjective, and (4) preserves distance.

Many authors use the term *motion* in place of *isometry*, suggesting physical action. That idea isn't formally incorporated into this text because it involves *time*—the situations before, during, and after the action—and time isn't ordinarily regarded as a geometric concept at this level. Instead, this text emphasizes the bijective correspondence between points P and P'. It alludes to physical motion only occasionally, as an informal guide to your intuition.

Translations

A familiar isometry $\tau: X \to X'$, the *translation* shown in figure 6.2.1, can be described by its effect on a single point O. Once you know $V = O'$, you can define at once the image of any point X not on $g = \overleftrightarrow{OV}$; it's the point X' that makes $OVX'X$ a parallelogram. Determining the image of a point W *on* g takes two parallelograms, as shown, or you can just refer to a scale on g. Phrasing the definition like this is awkward. It's easier just to say

Figure 6.2.1
Translation
$\tau: X \to X' = X + V$

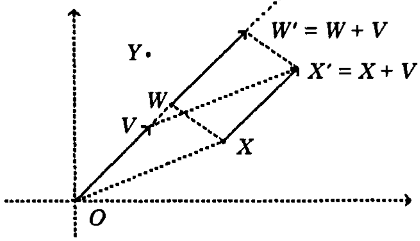

that if you introduce a coordinate system with origin O, then $X' = X + V$ for any X. By the parallelogram law, this gives the same result as the original definition, so it makes no difference which coordinate axes you use. You can regard the identity transformation as the translation τ with $V = O$.

Theorem 1. The translation $\tau : X \rightarrow X + V$ is an isometry.

Proof. It's surjective—any point Y is the image of $Y - V$. Find $Y - V$ in figure 6.2.1! And it preserves distance: $P' - Q' = (P + V) - (Q + V) = P - Q$, so $P'Q' = PQ$ by theorem 3.11.2. ♦

Using the word *translation* for this concept isn't as obscure as it might seem. The Latin prefix *trans-* means *across*, and *latus* is the past participle of the irregular verb *fero*, which means *carry*. So to *translate* X to X', you carry X across the parallelogram, just as you carried O to V.[18]

Rotations

You can conceive of a *rotation*, another kind of isometry, in much the same way. You've an intuition of rotation as a physical action. Hold this book in a vertical plane in front of you, your thumbs pressing on opposite spots on its covers, its spine toward you. Rotate it in that plane through 20° with your fingers, keeping your thumbs stationary. Which direction? Suppose friends on your right and left request the positive, counterclockwise direction. You try to comply, but one or the other always disagrees. Facing each other, they have opposite senses of direction. What one perceives as 20° the other sees as -20°. Perhaps intuition isn't entirely adequate! Your problem is to describe the rotation $X \rightarrow X'$ so that everyone agrees, and so that you can specify its effect on every point in all cases.

One solution is to specify not the center O and angle measure, but a particular angle, designating one ray as *initial* and the other as *terminal*. Actually, it's handier to mention just the rays; that lets you specify opposite rays for 180° and coincident rays for 0° or 360° rotations. To define the rotation $X \rightarrow X'$ about O that carries the initial ray $\overrightarrow{OA}$ to the terminal ray $\overrightarrow{OB}$, use an angular scale at O as in section 3.13. Suppose the rays correspond to angle parameters α and β, and let $\theta = \beta - \alpha$. Set $X' = O$ if $X = O$; otherwise find an angle parameter ξ for ray $\overrightarrow{OX}$, and let X' be the point such that $OX = OX'$ on the ray with angle parameter $\xi + \theta$. Figure 6.2.2 displays an example. This definition doesn't specify α, β, or ξ completely, but any other parameters α_1, β_1, and ξ_1 for the same rays would differ from them by multiples of 360°, so $\xi_1 + (\beta_1 - \alpha_1)$ would correspond to the same ray $\overrightarrow{OX'}$.

---

[18] *Translate* sometimes has this sense in English religious parlance: St. Mark's body was *translated* from Alexandria to Venice in A.D. 828. *Transfer* has the same etymology.

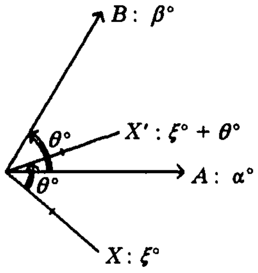

Figure 6.2.2 The rotation ρ_θ about O that maps $\vec{OA}$ to $\vec{OB}$

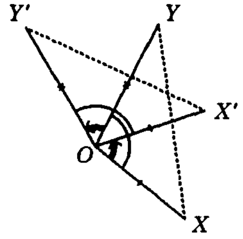

Figure 6.2.3 Rotation preserves distance: $XY = X'Y'$

Because the angle parameter θ figures so directly in the definition, the rotation is often denoted by ρ_θ. Since adding a multiple of 360° has no effect on ray $\vec{OX'}$, nor on the intuitive geometric notion of rotation, you don't have to restrict θ to the interval $0° < \theta < 180°$; it can take any value at all! If $\theta = 0°$, then $X' = X$ and ρ_θ is the identity. If $\theta = 180°$, ρ_θ is called the *half turn* σ_O about O.[19]

Theorem 2. For any θ, the rotation ρ_θ is an isometry.

Proof. Consider points X and Y noncollinear with O, with angle parameters ξ and η, as in figure 6.2.3. Then

$$\mathrm{m}\angle XOY = |(\xi - \eta) \bmod 360°|.$$

The images X' and Y' have angle parameters $\xi' = \xi + \theta$ and $\eta' = \eta + \theta$, so

$$\mathrm{m}\angle XOY = |(\xi' - \eta') \bmod 360°| = \mathrm{m}\angle X'OY'.$$

Triangles ΔXOY and $\Delta X'OY'$ are thus congruent, and $XY = X'Y'$. You should also prove this equation for the remaining case, where X and Y are collinear with O. ♦

Reflections

The reflections constitute a third type of isometry. They're perhaps less familiar, but much easier to describe. The *reflection* $X' = \sigma_g(X)$ of a point X across a line g is X itself if X is on g, else it's the point on the

---

[19] Some authors call σ_O a *point reflection.*

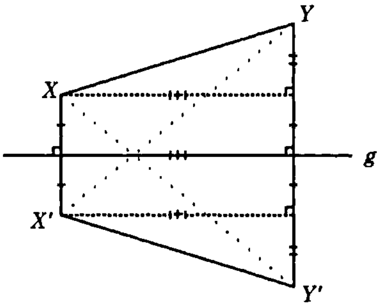

Figure 6.2.4 Reflection across g preserves distance: $XY = X'Y'$.

opposite side of g such that g is the perpendicular bisector of $\overline{XX'}$ —see figure 6.2.4.

Theorem 3. The reflection σ_g across line g is an isometry.

Proof. It's surjective because $X = \sigma_g(X') = \sigma_g(\sigma_g(X))$. Figure 6.2.4 sketches a proof that σ_g preserves the distance between points X and Y when they lie on the same side of g —that is, $XY = X'Y'$. If you show as well that $XY' = X'Y$, then interchange the labels Y and Y', you'll get a proof for X and Y on opposite sides of g. Finally, you can supply the even simpler argument that $XY = X'Y'$ when X and/or Y lies on g. ♦

The isometry group

There are other plane isometries besides the examples already given, but their details are postponed until section 6.7. The rest of this section concentrates on concepts that relate to *all* isometries.

Theorem 4. The isometries form a subgroup of the symmetric group of the plane.

Proof. This means that the identity transformation ι is an isometry, the inverse of an isometry φ is an isometry, and the composition of two isometries φ and χ is an isometry. That is, ι, φ^{-1}, and $\psi = \varphi\chi$ preserve distance. The first is trivial: ι preserves *everything*. To see that φ^{-1} preserves distance, consider two points P and Q. Let $X = \varphi^{-1}(P)$ and $Y = \varphi^{-1}(Q)$. Then $\varphi(X) = P$ and $\varphi(Y) = Q$. Moreover, $XY = \varphi(X)\varphi(Y)$ because φ is an isometry, so

$$\varphi^{-1}(P)\varphi^{-1}(Q) = XY = \varphi(X)\varphi(Y) = PQ.$$

Finally, to verify that $\psi(P)\psi(Q) = PQ$, note first that $\chi(P)\chi(Q) = PQ$ because χ is an isometry. Let $X = \chi(P)$ and $Y = \chi(Q)$. As before, $\varphi(X)\varphi(Y) = XY$, so

$$\psi(P)\psi(Q) = \varphi\chi(P)\varphi\chi(Q) = \varphi(X)\varphi(Y)$$
$$= XY = \chi(P)\chi(Q) = PQ. \blacklozenge$$

Theorem 4 is this book's first result of the form *the elements of a transformation group $\mathcal{T}$ that preserve ... form a subgroup of $\mathcal{T}$.* Here, $\mathcal{T}$ is the group of all transformations of the plane—its symmetric group—and the dots ... represent *distance.* You'll see many theorems that fit this pattern. They're all proved similarly.

Isometries don't necessarily commute. Figure 6.2.5 shows that $\varphi\chi \neq \chi\varphi$ for the reflections $\varphi = \sigma_g$ and $\chi = \sigma_h$ across selected lines g and h. You should investigate various other pairs φ and χ of isometries to get an idea of when their compositions $\varphi\chi$ and $\chi\varphi$ are or are not equal. Under the last heading of this section you'll see that any two *translations* commute. You'll study that question systematically later.

Invariance

A property of sets $\mathcal{X}$ or sequences $<X_1, X_2, \ldots>$ of points is called *invariant* under a transformation $X \rightarrow X'$ if $\mathcal{X}$ or $<X_1, X_2, \ldots>$ has that property when and only when its image $\{X' : X \in \mathcal{X}\}$ or $<X_1', X_2', \ldots>$ does. Distance, viewed as a property of pairs of points, is invariant under isometries by definition. Many other properties are invariant under isometries because you can define them in terms of distance. Theorem 5 presents a basic example.

Theorem 5. Betweenness is invariant under any isometry $X \rightarrow X'$.

Proof. If X_1, X_2, and X_3 are points, then

$$X_1\text{-}X_2\text{-}X_3 \leftrightarrow X_1X_2 + X_2X_3 = X_1X_3$$
$$\leftrightarrow X_1'X_2' + X_2'X_3' = X_1'X_3' \leftrightarrow X_1'\text{-}X_2'\text{-}X_3'. \blacklozenge$$

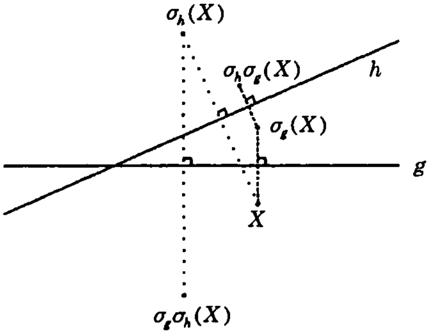

Figure 6.2.5
$\sigma_g\sigma_h(X) \neq \sigma_h\sigma_g(X)$

The critical idea of this proof is that you can define betweenness in terms of distance: X_1-X_2-X_3 $\leftrightarrow$ $X_1X_2 + X_2X_3 = X_1X_3$. You can use similar arguments to show that *any* property definable in terms of distance is invariant under isometries. At the end of section 6.1 you saw a long list of frequently used properties that are definable in terms of betweenness, hence in terms of distance. Invariance is frequently used implicitly in the rest of this chapter. For example, whenever you use a compound concept like the area of an isometric image of a triangle, you're alluding to two facts:

- The image of a triangle $\mathcal{T}$ is a triangle. (*Being a triangle* is in the list of properties definable in terms of betweenness.)
- It has the same area as $\mathcal{T}$. (You can define *area of a triangle* in terms of distance via Hero's formula, theorem 5.5.10.)

You can supply such arguments when they're needed.

Fixpoints; rigidity and uniqueness theorems

A *fixpoint* of a transformation φ is a point X that it leaves fixed: $\varphi(X) = X$. Often it's easier to analyze a transformation by considering its fixpoints than by studying how it changes other points. Translations, except for the identity, have *no* fixpoints. Rotations, except for the identity, have exactly *one*. The reflection in a line g has a whole *line* of fixpoints. And the fixpoints of the identity transformation constitute the whole *plane*. Are these the only possibilities for the fixpoint set of an isometry? The next two results—the *rigidity theorems*—answer that question.

Theorem 6. If an isometry $X \to X'$ has fixpoints $P \neq Q$, then every point X on $\overleftrightarrow{PQ}$ is fixed.

Proof. There are three cases: P-Q-X, Q-X-P, and X-P-Q. In the first case, where P-Q-X, theorem 5 implies P'-Q'-X', i.e. P-Q-X'. Thus X and X' are on the same side of Q in $\overleftrightarrow{PQ}$ and $QX = Q'X' = QX'$, so $X = X'$. You can prove the remaining cases similarly. ♦

Theorem 7. The only isometry with three noncollinear fixpoints is the identity.

Proof. Suppose an isometry $\varphi: X \to X'$ has noncollinear fixpoints A, B, and C. Then all edge points of ΔABC are fixed. For every point X you can find two distinct edge points P and Q with X on $\overleftrightarrow{PQ}$ —see figure 6.2.6. By theorem 6, X is fixed. Thus φ leaves every point fixed. ♦

Corollary 8 (Uniqueness theorem). If χ and ψ are isometries, and

$$\chi(A) = \psi(A) \qquad \chi(B) = \psi(B) \qquad \chi(C) = \psi(C)$$

for noncollinear points A, B, and C, then $\chi = \psi$.

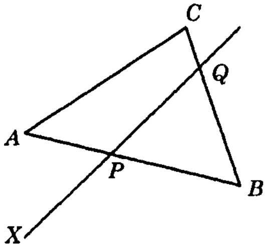

Figure 6.2.6
Proving theorem 7

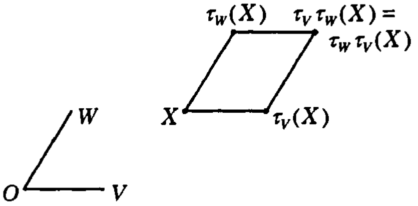

Figure 6.2.7
Translations commute

Proof. A, B, and C are fixpoints of $\varphi = \chi\psi^{-1}$, so φ is the identity. ♦

The uniqueness theorem is used often to prove that two isometries defined different ways are actually the same. Here's a typical example, which you can check by another method. Set up a coordinate system with origin O, and consider translations $\tau_V : X \to X + V$ and $\tau_W : X \to X + W$. Figure 6.2.7 shows that $\tau_V\tau_W(X) = \tau_W\tau_V(X)$ for the indicated point X. Often, such figures are valid only for a restricted set $\mathscr{X}$ of points X. But if $\mathscr{X}$ includes three noncollinear points, no matter how near each other, then by the uniqueness theorem, the equation holds for *all* points X. In this case, you can certainly pick three noncollinear points X for which the equation holds, so $\mathscr{X}$ is the entire plane and $\tau_V\tau_W = \tau_W\tau_V$. That is, any two translations commute. Of course, you could prove the same result algebraically:

$$\begin{aligned}\tau_V\tau_W(X) &= \tau_W(X) + V = (X + W) + V = (X + V) + W \\ &= \tau_V(X) + W = \tau_W\tau_V(X).\end{aligned}$$

But the uniqueness theorem is especially helpful when there are different cases to consider and no algebraic solution.

6.3 Reflections

Concepts
Mirror images
Triangle orientation
Reflections reverse orientation
Reflections in perpendicular lines
Half turns

Section 6.2 introduced the reflection $X' = \sigma_g(X)$ of a point X across a line g: If X lies on g, then $X' = X$, else $X' \neq X$ and g is the perpendicular bisector of $\overline{XX'}$. You saw that σ_g is a plane isometry. It's self-inverse: $\sigma_g^{-1} = \sigma_g$ because $\sigma_g\sigma_g(X) = X$ for each X. Why is it called a *reflection*? Stand with your eye at point Y on the same side of a mirror g as point X in figure 6.3.1. The mirror image of X appears to be at X', because light from X reflects from the mirror to Y so that the incidence and reflection angles $\angle XPO$ and $\angle YPQ$ are equal.[20]

The light ray reflected from X to Y strikes mirror g at the point P for which the path length $XP + PY$ is as short as possible. Figure 6.3.2 shows such a path for an arbitrary point P on g. Since $XP = X'P$, the shortest $XP + PY$ is the smallest sum $X'P + PY$, which occurs when P lies on $\overline{YX'}$ as in figure 6.3.1.

As a child you learned that a mirror reverses orientation; it reflects your right-hand gesture as one of your image's left hand. To learn some technological skills such as driving, you must develop intuition to process these reversed perceptions efficiently. This situation has a mathematical counterpart. Figure 6.3.3 should reinforce your intuition that if a figure $\mathcal{X}$ is oriented, then its image $\mathcal{X}' = \{X' : X \in \mathcal{X}\}$ under a reflection $\sigma_g : X \to X'$ has the opposite orientation.

What *is* an oriented figure? That concept is too general to formulate easily. In fact, it's more general than necessary; you need only consider triangles. Section 3.3 defined ΔABC as the *ordered* triple of its vertices. Thus, for example, ΔABC and ΔCBA are different, even though we use the same figure to represent them. Using figure 6.3.3, you can classify the six triangles with these vertices according to your intuitive notion of orientation:

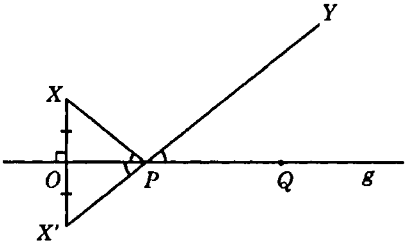

Figure 6.3.1 Mirror image X' of X as seen from Y

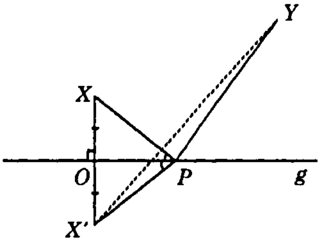

Figure 6.3.2 Path from X to Y via another point P on g

---

[20] The word *reflection* stems from the past participle of the Latin verb *reflexo*, which means *bend back*; the Greek σ in the traditional notation refers to the German word *Spiegel* for *mirror*.

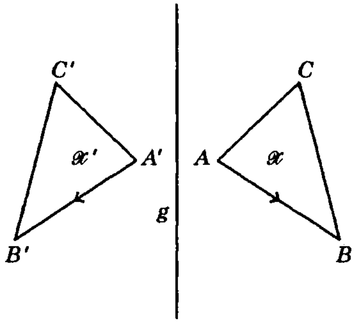

Figure 6.3.3
Reflection across g
reverses orientation.

Class (+): $\Delta ABC \quad \Delta BCA \quad \Delta CAB$
Class (-): $\Delta CBA \quad \Delta ACB \quad \Delta BAC$.

How can you distinguish these classes mathematically, without relying on figure 6.3.3 and your intuition? The determinant formula for triangle area derived in section 5.6 provides the tool. Choose a coordinate system, and regard the vertices as coordinate vectors. The triangles all have the same area; according to theorem 5.6.5, it's ± ½ times each of the determinants

$$\begin{array}{lll} \det[A-B,C-B] & \det[B-C,A-C] & \det[C-A,B-A] \\ \det[C-B,A-B] & \det[A-C,B-C] & \det[B-A,C-A]. \end{array}$$

Those in the first row are all equal. For example, by subtracting the right-hand column from the left, multiplying a column by -1, and interchanging columns, you get

$$\begin{aligned} \det[A-B,C-B] &= \det[(A-B)-(C-B),C-B] \\ &= \det[A-C,C-B] \\ &= -\det[A-C,B-C] \\ &= \det[B-C,A-C]. \end{aligned}$$

Moreover, each determinant in the second row is the negative of the one above it. Thus, all of the triangles in one row have the same *positive* determinant—they constitute class (+). The others have the same *negative* determinant—they constitute class (-). You can then call the triangles *positively* or *negatively* oriented, with respect to the chosen coordinate system. (The coordinate system for figure 6.3.3 evidently oriented ΔABC positively.)

To show that a reflection $\sigma_g : X \to X'$ reverses orientation, you could find equations for the coordinates of X' in terms of those of X, then show that $\det[A'-B', C'-B']$ and $\det[A-B,C-B]$ always differ in sign, so that $\Delta A'B'C'$ and ΔABC have opposite orientation. (In fact, these determinants always have the same magnitude; they differ *only* in sign!) Deriving the required equations for the reflection across an arbitrary line g is tedious. (See exercise 6.11.24.) Moreover, it's unnecessary, because a cleaner argument in section 6.9 will yield this general result. But you can check it easily for some special lines g. For example, if g is the first axis, then σ_g has equations $x_1' = x_1$, $x_2' = -x_2$, and you get $\det[A' - B', C' - B']$ from $\det[A - B, C - B]$ by changing the signs of the entries in its second row. If g is the second axis, change signs in the first row. If g is the diagonal line $x_1 = x_2$, then σ_g has equations $x_1' = x_2$, $x_2' = x_1$, and you get $\det[A' - B', C' - B']$ from $\det[A - B, C - B]$ by interchanging its rows. In each of these examples, the determinant merely changes sign.

The remainder of this section is concerned with the commutativity of reflections σ_g and σ_h across lines g and h. When does that occur? That is, for what lines g and h is $\sigma_g\sigma_h = \sigma_h\sigma_g$? Clearly, this equation holds if $g = h$. Figure 6.3.4 shows that if $g \perp h$, then $\sigma_g\sigma_h(P) = \sigma_h\sigma_g(P)$ for points P not on g or h. You can verify that it also holds for points P *on* the lines. Thus the reflections commute if the lines are equal or perpendicular. Conversely, suppose the reflections commute, but $g \neq h$. Find a point X on g but not h. Then

$$\sigma_g\sigma_h(X) = \sigma_h\sigma_g(X) = \sigma_h(X),$$

so $Y = \sigma_h(X)$ is a fixpoint of σ_g, and Y lies on g. Moreover, $X \neq Y$ because X isn't on h, so h is the perpendicular bisector of $\overline{XY}$ and $g = \overleftrightarrow{XY}$, hence $g \perp h$. This argument, and further inspection of figure 6.3.4, demonstrates

Theorem 1. σ_g and σ_h commute just when $g = h$ or $g \perp h$. If $g \perp h$, then $\sigma_g\sigma_h = \sigma_h\sigma_g$ is the half turn σ_O about the intersection $O = g \cap h$.

Corollary 2. A half turn σ_O commutes with a line reflection σ_h just when O lies on h. In that case, $\sigma_O\sigma_h = \sigma_h\sigma_O = \sigma_g$, the reflection across the line $g \perp h$ through O.

Proof. If O lies on h, then

$$\sigma_O\sigma_h = \sigma_g\sigma_h\sigma_h = \sigma_g = \sigma_h\sigma_h\sigma_g = \sigma_h\sigma_O$$

by theorem 1. Conversely,

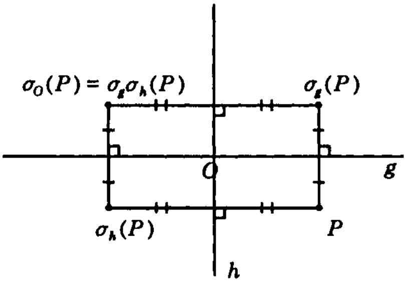

Figure 6.3.4 σ_g and σ_h commute if $g \perp h$.

$$\begin{aligned}
&\sigma_O\sigma_h = \sigma_h\sigma_O \\
&\quad\Rightarrow\ \sigma_O\sigma_h(O) = \sigma_h\sigma_O(O) = \sigma_h(O) \\
&\quad\Rightarrow\ \sigma_h(O) \text{ is a fixpoint of } \sigma_O \\
&\quad\Rightarrow\ \sigma_h(O) = O \text{ —the } \textit{only} \text{ fixpoint of } \sigma_O \text{ —} \\
&\quad\Rightarrow\ O \text{ is a fixpoint of } \sigma_h \\
&\quad\Rightarrow\ O \text{ lies on } h. \blacklozenge
\end{aligned}$$

You can complete this discussion of commutativity and reflections by verifying that two half turns σ_O and σ_P *never* commute unless $O = P$. In the next section, you'll see what kind of isometry $\sigma_O\sigma_P$ is in general.

6.4 Translations

Concepts

Definition of a translation
Equations for a translation
Compositions of translations and vector addition
Translation group
Compositions of reflections in parallel lines
Compositions of half turns
Translations preserve orientation

Section 6.2 presented two definitions of a translation: a purely geometric one using parallelograms and an equivalent formulation in terms of vector algebra. The latter is more convenient, but requires you to choose a coordinate system at the start of the discussion. Each translation is determined by the image V of the origin O; you can define the translation τ_V with vector V by setting $\tau_V(X) = X + V$ for each point X (regarded as a coordinate vector). This includes the case $\tau_V(O) = O + V = V$. As shown in figure 6.2.1, you can construct the point $X + V$ from V, O, and X using parallelograms, with no reference to coordinates. Therefore, the definition of τ_V depends only on V and O, not on other details of the coordinate system. Theorem 1 displays the *equations* for τ_V in terms of the coordinate system.

Theorem 1. The components of $X' = \tau_V(X)$ are

$$\begin{cases} x_1' = x_1 + v_1 \\ x_2' = x_2 + v_2 \,. \end{cases}$$

Not only do translations τ_V correspond one to one with their vectors V, but composition of translations corresponds to vector addition. Further, this analogy relates the inverse of a translation to the negative of its vector, and the identity, or *trivial*, translation corresponds to the zero vector:

Theorem 2. For any vectors V and W, $\tau_V \tau_W = \tau_{V+W} = \tau_W \tau_V$ and $\tau_V^{-1} = \tau_{-V}$. Moreover, τ_O is the identity transformation ι.

Proof. For all points X,

$$\begin{aligned}\tau_V \tau_W(X) &= \tau_V(X+W) = (X+W)+V = X+(W+V)\\ &= X+(V+W) = \tau_{V+W}(X)\\ \tau_O(X) &= X+O = X = \iota(X).\end{aligned}$$

Moreover,

$$\begin{aligned}&\tau_W \tau_V = \tau_{W+V} = \tau_{V+W} = \tau_V \tau_W\\ &\tau_V \tau_{-V} = \tau_{V+(-V)} = \tau_O = \iota = \tau_{(-V)+V} = \tau_{-V}\tau_V. \;\blacklozenge\end{aligned}$$

Corollary 3. The translations form a transformation group on the plane.

Proof. The composition $\tau_V \tau_W$ of two translations is the translation τ_{V+W}. The inverse τ_V^{-1} of a translation is the translation τ_{-V}. Finally, the identity transformation is a translation: $\iota = \tau_O$. ♦

The next results show how to build translations from reflections or half turns.

Theorem 4. The composition $\sigma_g \sigma_h$ of the reflections across equal or parallel lines g and h is a translation τ. With respect to any coordinate system, τ has vector $V = 2(P-Q)$, where P and Q are the intersections of g and h with a common perpendicular k.

Proof. If $g = h$, then V is the zero vector and $\sigma_g \sigma_h = \iota = \tau_V$. If $g \neq h$, as in figure 6.4.1, then

$$V = 2(P-Q) = \sigma_g \sigma_h(X) - X,$$

so $\tau_V(X) = \sigma_g \sigma_h(X)$ for any point X on the side of h opposite g. This equation holds for three noncollinear points—pick them near X in figure 6.4.1—so by the uniqueness theorem, $\sigma_g \sigma_h = \tau_V$. ♦

Corollary 5. The composition $\sigma_P \sigma_Q$ of half turns about points P and Q is a translation τ. With respect to any coordinate system, τ has vector $V = 2(P-Q)$.

Proof. Let g and h be the perpendiculars to $k = \overleftrightarrow{PQ}$ through P and Q. Then $\sigma_P \sigma_Q = \sigma_g \sigma_k \sigma_k \sigma_h = \sigma_g \sigma_h$. Now apply theorem 4. ♦

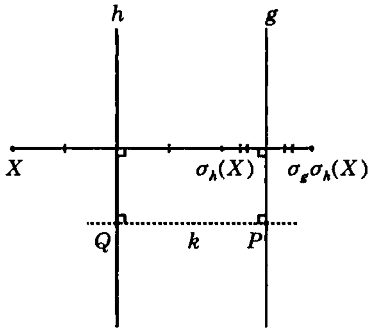

Figure 6.4.1 $\sigma_g\sigma_h$ is a translation when $g \parallel h$.

Given vector V and point P you can find a point Q, or given V and Q you can find P, such that $V = 2(P - Q)$. That yields

Corollary 6. Every translation τ is the composition $\sigma_P\sigma_Q$ of two half turns. You can choose P or Q arbitrarily; the other then depends on τ. If $P \neq Q$, then τ is the composition $\sigma_g\sigma_h$ of the reflections across the perpendiculars g and h to $\overleftrightarrow{PQ}$ through P and Q.

In section 6.5 you'll need to analyze all compositions of *three* reflections. The next result provides a start for that study.

Corollary 7. A composition $\sigma_g\sigma_h\sigma_k$ of reflections across parallel or equal lines g, h, and k is the reflection across a fourth line copencilar with these three.

Proof. A line l perpendicular to g, h, and k intersects them at points P, Q, and S. Choose a coordinate system, and find the point R on l such that $P - Q = R - S$. Let $V = 2(P - Q) = 2(R - S)$ and j be the perpendicular to l through R. By theorem 4, $\sigma_g\sigma_h = \tau_V = \sigma_j\sigma_k$, so $\tau_V(X) = \sigma_g\sigma_h(X)$ for any point X. ♦

Transformations correspond so closely to vectors that properties of vectors are often attributed to translations. For example, we refer to parallel translations or to the angle between them. By the *length* $|\tau|$ of a translation we mean the length of its vector: $|\tau| = P\tau(P)$ for any point P.

The last idea considered in this section is the effect of translation on orientation. Figure 6.4.2 should reinforce your intuition that any translation preserves orientation. You could deduce that from theorem 4 or corollary 5 if you'd shown already that every reflection reverses orientation or every half turn preserves it. However, neither of those results has been proved yet. Instead, you can use coordinates and determinants to construct the following argument, similar to the one devised for some special reflections in section 6.3.

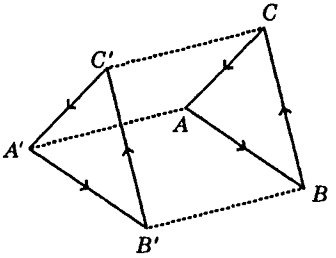

Figure 6.4.2 Translation preserves orientation.

Theorem 8. Suppose a translation τ_V maps ΔABC to $\Delta A'B'C'$. Then

$$\det[A - B, C - B] = \det[A' - B', C' - B'],$$

so τ_V preserves orientation.

Proof. $\det[A' - B', C' - B']$
$= \det[(A + V) - (B + V), (C + V) - (B + V)]$
$= \det[A - B, C - B]$. ♦

6.5 Rotations

Concepts

- Definition of a rotation
- Equation of a rotation about the origin
- Composition of rotations and angle addition
- Group of rotations about O
- Compositions of reflections in intersecting lines
- Compositions of rotations and translations

Section 6.2 defined the rotation $X \to X'$ about a point O that carries initial ray $\overrightarrow{OA}$ to terminal ray $\overrightarrow{OB}$ using an angular scale p at O as follows. Suppose p assigns those rays angle parameters α and β, and let $\theta = \alpha - \beta$. For any point $X \neq O$, find an angle parameter ξ corresponding to ray $\overrightarrow{OX}$, and let $X' = \rho_\theta(X)$ be the point on the ray with angle parameter $\xi + \theta$, such that $OX' = OX$. When you study a rotation in a broader context, you're often not free to choose the scale; some other aspect of your problem may have specified it already. Does it make any difference which angular scale you use? The answer is no. According to theorem 3.13.3, if p' is another scale at O, then there are constants ζ and n such that $n = \pm 1$ and for all ξ, $p(\xi) = p'(\zeta + n\xi)$. If you used scale p' to describe ρ, you'd find that the initial and terminal rays and ray $\overrightarrow{OX}$ correspond to parameters

$\alpha' = \zeta + n\alpha$, $\beta' = \zeta + n\beta$, and $\xi' = \zeta + n\xi$. To determine $\rho_\theta(X)$, you'd compute

$$\begin{aligned}\xi' + (\beta' - \alpha') &= (\zeta + n\xi) + (\zeta + n\beta) - (\zeta + n\alpha)\\ &= \zeta + n(\xi + \beta - \alpha) = \zeta + n(\xi + \theta),\end{aligned}$$

and you'd find that the corresponding ray is $p'(\zeta + n(\xi + \theta)) = p(\xi + \theta) = \vec{OX'}$, the same one you got with scale p.

Now choose a coordinate system with origin O. Use any angular scale at O. For any real number θ consider the rotation $\rho_\theta : X \rightarrow X'$ that carries the ray with parameter 0° to the one with parameter θ. Theorem 1 displays the *equations* for ρ_θ in terms of the coordinate system. Later, section 6.8 will derive equations for a rotation about a point different from the origin.

Theorem 1. The components of $X' = \rho_\theta(X)$ are

$$\begin{cases} x_1' = x_1 \cos\theta - x_2 \sin\theta \\ x_2' = x_1 \sin\theta + x_2 \cos\theta\,. \end{cases}$$

Proof. Suppose X has angle parameter ξ. Then $OX' = OX$ and X' has parameter $\xi + \theta$. Using the cosine and sine addition formulas (corollary 5.5.5) you get

$$\begin{aligned} x_1 &= OX \cos\xi \\ x_2 &= OX \sin\xi \\ x_1' &= OX' \cos(\xi + \theta) \\ &= OX[\cos\xi\cos\theta - \sin\xi\sin\theta] = x_1\cos\theta - x_2\sin\theta \\ x_2' &= OX'\sin(\xi + \theta) \\ &= OX[\cos\xi\sin\theta + \sin\xi\cos\theta] = x_1\sin\theta + x_2\cos\theta. \; \blacklozenge \end{aligned}$$

Not only do rotations ρ_θ about O correspond to angle parameters θ, but composition of these rotations corresponds to angle addition. Further, this analogy relates the inverse of a rotation to the negative of its parameter, and the identity rotation corresponds to parameter zero:

Theorem 2. $\rho_\eta\rho_\theta = \rho_{\eta+\theta} = \rho_\theta\rho_\eta$ and $\rho_\theta^{-1} = \rho_{-\theta}$ for any angle parameters η and θ. Moreover, ρ_{0° is the identity transformation ι.

Proof. First, $\rho_\eta\rho_\theta(O) = \rho_\eta(O) = O$. Consider a point $X \neq O$ for which ray $\vec{OX}$ has angle parameter ξ; let $\rho_\eta(X) = X'$ and $\rho_\theta(X') = X''$. Then $\vec{OX'}$ and $\vec{OX''}$ have parameters $\xi + \eta$ and $(\xi + \eta) + \theta = \xi + (\eta + \theta)$, so $\rho_{\eta+\theta}(X) = X''$. It follows that $\rho_\eta\rho_\theta = \rho_{\eta+\theta}$. Similarly, $\rho_\theta\rho_\eta = \rho_{\theta+\eta} = \rho_{\eta+\theta}$. Clearly, $\rho_{0^\circ} = \iota$. Finally, $\rho_\theta^{-1} = \rho_{-\theta}$ because $\rho_{-\theta}\rho_\theta = \rho_{-\theta+\theta} = \rho_{0^\circ} = \iota$. ♦

According to theorem 2, the composition of rotations about O, the inverse of any such rotation, and the identity transformation are all rotations about O:

Corollary 3. The rotations about O form a transformation group on the plane.

In section 6.4 you saw that a composition of reflections across parallel lines is a translation. The next result shows that if the lines intersect, the composition is a rotation. In fact, *every* rotation is such a composition of two reflections.

Theorem 4. If rays $\overrightarrow{OA}$ and $\overrightarrow{OB}$ have angle parameters α and β, then $\rho_{2(\alpha-\beta)}$ is the composition $\sigma_a\sigma_b$ of the reflections across lines $a = \overleftrightarrow{OA}$ and $b = \overleftrightarrow{OB}$.

Proof. If $a = b$, then $\rho_{2(\alpha-\beta)} = \iota = \sigma_a\sigma_b$. If $a \perp b$, the result is contained in theorem 6.3.1.

Now assume $\angle AOB$ is acute, as in figure 6.5.1. Select a point X on the side of b opposite A but on the same side of a as B, with $m\angle XOB < m\angle BOA$ as shown. Let $X' = \sigma_b(X)$ and $X'' = \sigma_a(X') = \sigma_a\sigma_b(X)$. Then rays $\overrightarrow{OX}$, $\overrightarrow{OB}$, $\overrightarrow{OX'}$, $\overrightarrow{OA}$, and $\overrightarrow{OX''}$ all fall in a single half plane as shown, and $m\angle XOX'' = 2m\angle BOA$, so $X'' = \rho_{2(\alpha-\beta)}(X)$. You can select three noncollinear points X that satisfy the same conditions, so $\sigma_a\sigma_b = \rho_{2(\alpha-\beta)}$ by the rigidity theorem.

If $\angle AOB$ is obtuse, replace A by a point $A' \neq O$ such that A-O-A'. Then $\angle A'OB$ is acute, $\overleftrightarrow{OA'} = a$, and $\overrightarrow{OA'}$ has parameter $\alpha + 180°$. By the previous paragraph, $\sigma_a\sigma_b = \rho_{2(\alpha+180°-\beta)} = \rho_{2(\alpha-\beta)+360°} = \rho_{2(\alpha-\beta)}$. ♦

Given an angle parameter θ and a ray $\overrightarrow{OA}$ with parameter α you can find a ray $\overrightarrow{OB}$ with parameter β, or given θ and $\overrightarrow{OB}$ you can find $\overrightarrow{OA}$, such that $\theta = 2(\alpha - \beta)$. Setting $a = \overleftrightarrow{OA}$ and $b = \overleftrightarrow{OB}$ yields

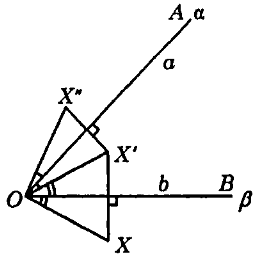

Figure 6.5.1
Proof of theorem 4

Corollary 5. Every rotation ρ about O is the composition $\sigma_a\sigma_b$ of two reflections across lines through O. You can choose either line a or b arbitrarily; the other then depends on ρ.

The next result continues the analysis of compositions of *three* reflections that you started with corollary 6.4.

Corollary 6. A composition $\sigma_g\sigma_h\sigma_k$ of reflections across lines through O is the reflection across a fourth line copencilar with these three.

Proof. Find points A, B, and $D \neq O$ on g, h, and k. Suppose rays $\vec{OA}$, $\vec{OB}$, and $\vec{OD}$ correspond to angle parameters α, β, and δ. Find the ray $\vec{OC}$ that corresponds to the parameter γ for which $\alpha - \beta = \gamma - \delta$, and let $j = \overleftrightarrow{OC}$. Then $\sigma_g\sigma_h = \rho_{2(\alpha-\beta)} = \rho_{2(\gamma-\delta)} = \sigma_j\sigma_k$, so $\sigma_g\sigma_h\sigma_k = \sigma_j$. ♦

Theorems 6.4.2 and 6.5.2 showed that the composition of two translations or of two rotations about the same center is a translation or rotation, respectively. Theorems 6.4.4 and 6.5.4 showed that the composition of two reflections is a translation or a rotation. The final two results in this section describe the composition of a rotation and a translation, and of rotations about different centers.

Corollary 7. The composition of a rotation $\rho \neq \iota$ and any translation τ is a rotation about some center.

Proof. By corollary 6.4.6, $\tau = \sigma_g\sigma_h$ for some line g through the center O of ρ and some line h parallel to g. By corollary 5, $\rho = \sigma_f\sigma_g$ for some line $f \neq g$ through O. Therefore $\rho\tau = \sigma_f\sigma_g\sigma_g\sigma_h = \sigma_f\sigma_h$, a rotation about the intersection P of f and h. You can supply a similar argument for $\tau\rho$. ♦

Corollary 8. Any composition of two rotations is a rotation or a translation.

Proof. By theorem 2 it's enough to consider rotations ρ and ρ' about distinct centers O and O'. Let $h = \overleftrightarrow{OO'}$. By corollary 5, $\rho = \sigma_g\sigma_h$ and $\rho' = \sigma_h\sigma_{g'}$ for some lines g and g' through O and O'. Therefore $\rho\rho' = \sigma_g\sigma_h\sigma_h\sigma_{g'} = \sigma_g\sigma_{g'}$. By theorem 6.4.4 or theorem 4, this composition is a translation or a rotation. ♦

Corollary 7 doesn't mention the relationship of the angle parameters of the constituent rotations to that of the composition. Corollary 8 doesn't show how the parameters of the constituent rotations determine whether the composition φ is a rotation or a translation, and in the former case how they are related to the angle parameter of φ. You can work out the connections in exercise 6.11.10 by analyzing the lines used in the proofs, or you can refer ahead to theorem 6.9.3, which derives these results through matrix algebra.

6.6 Structure theorem

Concepts

Congruent figures
Structure theorem
Even and odd isometries
Group of even isometries

Previous sections introduced the concept of a plane isometry, and described three familiar types in detail: reflections, translations, and rotations. You saw that all translations and rotations are compositions of two reflections. This section's main result is the structure theorem. It shows that *every* isometry is the identity transformation, or a reflection, or the composition of two reflections, *or the composition of three.* It's the fundamental tool for classifying plane isometries. Section 6.7 investigates the only type not yet introduced: the composition of three reflections. In section 6.8 the structure theorem plays a central role in connecting the study of isometries with matrix algebra.

The proof of the structure theorem proceeds via three preliminary results, which are arranged to make the argument as concise as possible. Lemma 2 and theorem 3 are hard to illustrate cleanly because each describes two or three different cases. You should work out separate figures for all cases.

Lemma 1. Suppose ΔABC and ΔABZ are congruent triangles. Then $C = Z$ or $\sigma_{\overleftrightarrow{AB}}(Z)$.

Proof. By theorem 3.12.5 case (2), the circles with centers A and B and radii $AC = AZ$ and $BC = BZ$ intersect at exactly two points Z and Z', and $\overleftrightarrow{AB}$ is the perpendicular bisector of $\overline{ZZ'}$. See figure 6.6.1. ♦

Lemma 2. Suppose ΔABC and ΔAYZ are congruent triangles. Then there exists an isometry χ such that $B = \chi(Y)$, $C = \chi(Z)$, and χ is the

(0) identity transformation,
(1) reflection across a line through A, or
(2) composition of the reflections across two lines through A.

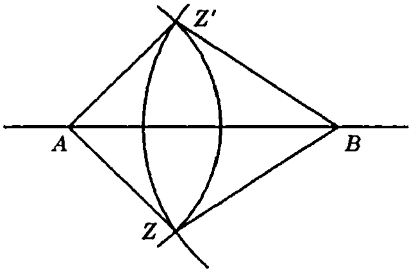

Figure 6.3.1
Proving lemma 1

Proof. If $B = Y$, lemma 1 yields the result with alternative (0) or (1). Otherwise, the perpendicular bisector h of $\overline{BY}$ passes through A because $AB = AY$. Then $B = \sigma_h(Y)$; let $Z' = \sigma_h(Z)$. By the SSS congruence principle, $\Delta ABZ' \cong \Delta AYZ \cong \Delta ABC$. By lemma 1 with Z' in place of Z, either

(1) $C = Z'$, or
(2) $C = \sigma_g(Z')$, where $g = \overleftrightarrow{AB}$.

In case (1), let $\chi = \sigma_h$; in case (2), $\chi = \sigma_g \sigma_h$. ♦

Theorem 3. Suppose ΔABC and ΔXYZ are congruent triangles. Then there exists an isometry φ such that $A = \varphi(X)$, $B = \varphi(Y)$, $C = \varphi(Z)$, and φ is

(0) the identity transformation,
(1) a reflection,
(2) the composition of reflections across two lines, or
(3) the composition of reflections across three lines.

Proof. If $A = X$, lemma 2 yields the result with alternative (0), (1), or (2). Otherwise, let k be the perpendicular bisector of $\overline{AX}$. Then $A = \sigma_k(X)$; let $Y' = \sigma_k(Y)$ and $Z' = \sigma_k(Z)$. By the SSS principle, $\Delta AY'Z' \cong \Delta XYZ \cong \Delta ABC$. By lemma 2 with Y' and Z' in place of Y and Z, there exists an isometry χ such that $B = \chi(Y')$, $C = \chi(Z')$, and χ is the

(1) identity transformation,
(2) reflection across a line through A, or
(3) composition of the reflections across two lines through A.

Let $\varphi = \chi \sigma_k$. ♦

According to theorem 3, congruent triangles are always related by some isometry. By the SSS congruence principle, triangles related by an isometry are always congruent. Thus, you should be able to *define* triangle congruence in terms of isometry in section 3.5. Some axiomatizations of geometry do exactly that. But this idea has more practical consequences, too. You can now easily *extend* the notion of congruence: Two figures $\mathcal{F}$ and $\mathcal{F}'$ are called *congruent*—in symbols, $\mathcal{F} \cong \mathcal{F}'$ —if $\mathcal{F}' = \varphi[\mathcal{F}]$ for some isometry φ. That's much more graceful than defining congruence separately for each interesting class of figure. You should check that this new definition is consistent with the section 3.5 definitions for segment and angle congruence, and with any other special cases you may have encountered earlier.

Theorem 4 **(*Structure theorem*).** The group of plane isometries consists of the identity, reflections, and compositions of two or three reflections.

Proof. Consider any isometry ψ, select any triangle ΔXYZ, and let $A = \psi(X)$, $B = \psi(Y)$, and $C = \psi(Z)$. By theorem 3, there exists an isometry φ such that $A = \varphi(X)$, $B = \varphi(Y)$, $C = \varphi(Z)$, and φ is

(0) the identity transformation,
(1) a reflection,
(2) the composition of reflections across two lines, or
(3) the composition of reflections across three lines.

By the uniqueness theorem, $\varphi = \psi$. ♦

To what extent do these isometry types overlap? The identity transformation ι is a composition of two reflections: $\iota = \sigma_g \sigma_g$ for any line g. Similarly, every reflection is a composition of three: $\sigma_g = \sigma_g \sigma_g \sigma_g$. You could write the structure theorem more tersely: Every plane isometry is a composition of two or three reflections. Is there any overlap between *these* two classes? The equation $\sigma_g \sigma_h = \sigma_j \sigma_k \sigma_l$ for lines g, h, j, k, and l leads to contradiction, as follows. Let $\varphi = \sigma_g \sigma_h$ and $\chi = \sigma_l \sigma_k$ so that $\varphi\chi = \sigma_j$. By theorems 6.4.4 and 6.5.4 and corollaries 6.5.7 and 6.5.8, each of φ and χ, and hence their composition, is a rotation or a translation. But the fixpoint set of a rotation or a translation is empty, a single point, or the entire plane, whereas that for a reflection is a line. Therefore, $\varphi\chi \neq \sigma_j$ —contradiction! *No* composition of two reflections is a composition of three.

Compositions of two reflections are called *even* isometries; compositions of three are called *odd*. By the previous paragraph, every isometry is even or odd, but none is both. The inverse of an even isometry—a translation or rotation—is also a translation or rotation, hence it's even. Summarizing these results about even isometries, here is

Theorem 5. The even isometries—the translations and rotations—constitute a group.

The notions of even and odd isometry are closely connected with that of orientation, mentioned several times in earlier sections. Since orientation was defined via coordinates and a determinant, that aspect of the theory is presented later, in section 6.8.

You've considered the reflections and the even isometries already in considerable detail. The next section investigates the remaining type: compositions of three reflections.

6.7 Glide reflections

Concepts

When do a translation and a reflection commute?
Glide reflections

In sections 6.4 and 6.5 you saw that the composition φ of reflections across three copencilar lines g, h, and k is a reflection across another line of the same pencil. This section investigates φ when g, h, and k aren't copencilar.

The first consideration seems unrelated, but it will provide the answer. When does a translation commute with a reflection?

Theorem 1. Let $\tau : P \to P'$ be a translation that's not the identity. Then τ and a reflection σ_g commute if and only if $\overleftrightarrow{PP'} \parallel g$ for some point P.

Proof. Figure 6.7.1 depicts the *if* argument. To verify the converse, suppose $\overleftrightarrow{PP'}$ and g intersect at a single point O. You'll find that $\tau\sigma_g(O)$ and $\sigma_g\tau(O)$ lie on different sides of g. ♦

When τ and σ_g commute, their composition φ is called a *glide reflection*. In that case, $\varphi\varphi = \tau\sigma_g\sigma_g\tau = \tau\tau$. This equation lets you determine τ if you know φ; for any point P, let $P'' = \varphi\varphi(P)$. Then $\tau(P)$ is the midpoint of $\overline{PP''}$. You can also determine g, since $\sigma_g = \varphi\tau^{-1}$. Thus you can call τ *the* corresponding translation and g *the* axis of φ. The vector corresponding to τ is often called the vector of φ.

When its translation is the identity, a glide reflection φ is just a reflection. Otherwise, you can see from figure 6.7.1 that φ has no fixpoint. In that case, φ is an odd motion, but not a reflection. Therefore, it's the composition of the reflections in three noncopencilar lines. Theorem 2 is the converse of this statement.

Theorem 2. If lines k, h, and g are not copencilar, then $\sigma_k\sigma_h\sigma_g$ is a glide reflection, but not a reflection.

Proof, case 1: k and h intersect at a point O, which may not lie on g. Find the line $h' \perp g$ through O; it intersects g at a point P. (See figure

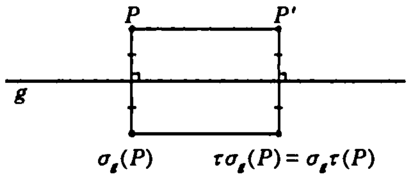

Figure 6.7.1
A commuting translation
τ and reflection σ_g

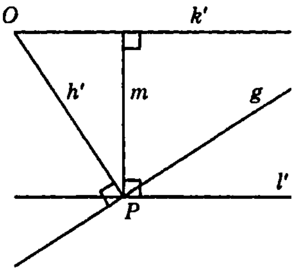

Figure 6.7.2
Case (1) of the proof
of theorem 2

6.7.2.) Let k' be the image of h' under the rotation $\sigma_k\sigma_h$, so that $\sigma_k\sigma_h = \sigma_{k'}\sigma_{h'}$. Find the line $m \perp k'$ through P, and the line $l' \perp m$ through P. Then $\sigma_{h'}\sigma_g = \sigma_P = \sigma_{l'}\sigma_m$ and $\iota \neq \sigma_{k'}\sigma_{l'} = \tau$, a translation. If $P' = \tau(P)$, then P' lies on m, hence $\tau\sigma_m$ is a glide reflection, but not a reflection. Finally,

$$\sigma_k\sigma_h\sigma_g = \sigma_{k'}\sigma_{h'}\sigma_g = \sigma_{k'}\sigma_P = \sigma_{k'}\sigma_{l'}\sigma_m = \tau\sigma_m .$$

Case 2: $k \nparallel h$. Then k and g intersect at a point. By case (1), $\sigma_g\sigma_h\sigma_k$ is a glide reflection $\tau\sigma_m$ with translation $\tau \neq \iota$ and axis m. Choose a point P, and set $O = \tau^{-1}(P)$ and $\tau(P) = Q$. Then $\overleftrightarrow{PO} = \overleftrightarrow{PQ}$, so τ^{-1} and σ_m commute and $\tau^{-1}\sigma_m$ is a glide reflection, but not a reflection. Finally,

$$\sigma_k\sigma_h\sigma_g = (\sigma_g\sigma_h\sigma_k)^{-1} = (\tau\sigma_m)^{-1} = \sigma_m^{-1}\tau^{-1} = \tau^{-1}\sigma_m . \blacklozenge$$

6.8 Isometries and orthogonal matrices

Concepts

- Orthogonal matrices A
- Inverses and products of orthogonal matrices are orthogonal
- Equation of a rotation about the origin O
- Equation of a reflection across a line through O
- Equations $X' = AX + V$ always represent isometries
- Every isometry has a unique equation $X' = AX + V$
- Angle parameters for arbitrary rotations
- Determinants of orthogonal matrices and parity of isometries

So far, your study of isometries has employed mostly synthetic methods. You've used coordinates only to formulate equations for translations, rotations about the origin, and a few special line reflections. And those have figured only in rudimentary discussions of orientation. It's time to use coordinates more seriously to derive equations for *all* plane isometries and to complete the discussion of orientation. Section 6.9 will then use equations to study various compositions of isometries.

For this section, use a fixed coordinate system with origin O, and the usual angular scale.

The most efficient way to formulate equations for arbitrary plane isometries uses matrix algebra. For example, you can restate as single matrix equations the pairs of equations derived in sections 6.5 and 6.3 for a rotation ρ_θ about O and a reflection σ_h across the first axis. For each of these isometries $X \to X'$ the components of X' are

$$\rho_\theta \begin{cases} x_1' = x_1 \cos\theta - x_2 \sin\theta \\ x_2' = x_1 \sin\theta + x_2 \cos\theta \end{cases} \qquad \sigma_h \begin{cases} x_1' = x_1 \\ x_2' = -x_2 \end{cases}$$

$$X' = A_\theta X \qquad\qquad X' = HX$$

$$A_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \qquad H = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Orthogonal matrices

In this process, *orthogonal* matrices play a central role: matrices whose transposes are their inverses. That is, a square matrix A is orthogonal if A^tA is the identity matrix I. For example, the matrices A_θ and H are orthogonal:

$$A_\theta^t A_\theta = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

$$= \begin{bmatrix} \cos^2\theta + \sin^2\theta & 0 \\ 0 & \sin^2\theta + \cos^2\theta \end{bmatrix} = I$$

$$H^t H = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & (-1)^2 \end{bmatrix} = I.$$

These examples show that *some* isometries that fix the origin O correspond to orthogonal matrices. In fact, *every* orthogonal matrix A represents an isometry $X \to X' = AX$ that fixes O. To prove that, note that $X \to X'$

bijectively because A is invertible. Moreover, $O' = AO = O$ and for any points P and Q,

$$\begin{aligned}(P'Q')^2 &= (P' - Q')\cdot(P' - Q')\\ &= (P' - Q')^t\,(P' - Q')\\ &= (AP - AQ)^t\,(AP - AQ)\\ &= \big(A(P - Q)\big)^t A(P - Q)\\ &= (P - Q)^t A^t A(P - Q)\\ &= (P - Q)^t\,(P - Q) = (PQ)^2.\end{aligned}$$

Thus $P'Q' = PQ$; the transformation $X \to X'$ preserves distance.

For any vector V, the transformation $X \to X'' = X' + V = AX + B$ is also an isometry:

$$\begin{aligned}(P''Q'')^2 &= (P'' - Q'')\cdot(P'' - Q'')\\ &= \big((P' + V) - (Q' + V)\big)\cdot\big((P' + V) - (Q' + V)\big)\\ &= (P' - Q')\cdot(P' - Q') = (P'Q')^2 = (PQ)^2.\end{aligned}$$

You'll soon see that you can represent *every* plane isometry $X \to X''$ this way; there's always an orthogonal matrix A and a vector V such that $X'' = AX + V$ for every point X.

First, here are some simple and useful facts about orthogonal matrices. Clearly, the identity matrix I is orthogonal. The inverse of an orthogonal matrix A is orthogonal, for if $A^tA = I$, then $A^t = A^{-1}$ and

$$(A^{-1})^t A^{-1} = (A^{-1})^t A^t = (AA^{-1})^t = I^t = I.$$

Moreover, the product of orthogonal matrices A and B is orthogonal:

$$(AB)^t(AB) = B^tA^tA\,B = B^tB = I.$$[21]

Now consider a line $g = \overrightarrow{OP}$, where ray $\overrightarrow{OP}$ has angle parameter θ. Find the ray $\overrightarrow{OB}$ with parameter $\frac{1}{2}\theta$, and let $b = \overleftrightarrow{OB}$, as in figure 6.8.1. Consider the rotation $\rho_\theta = \sigma_b\sigma_h = \sigma_g\sigma_b$. This equation implies

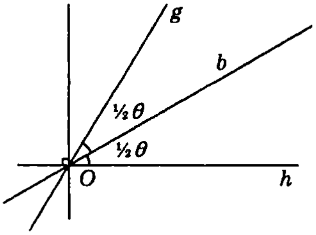

Figure 6.8.1
$\sigma_h\sigma_b = \sigma_b\sigma_g$

---

[21] This paragraph closely parallels the definition of a transformation group in section 6.1. Had this book defined the notion *matrix group*, it could speak in terms of the *group* of orthogonal matrices. But it avoids the more abstract group concepts because they're not so useful in this context.

$$\sigma_g = \sigma_b\sigma_h\sigma_b = \sigma_b\sigma_h\sigma_h\sigma_h\sigma_b = \rho_\theta\sigma_h\rho_\theta^{-1}.$$

Isometries ρ_θ and σ_h correspond to orthogonal matrices A_θ and H, so σ_g corresponds to the orthogonal matrix

$$\begin{aligned} A_\theta H A_\theta^{-1} &= A_\theta H A_{-\theta} \\ &= \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \cos(-\theta) & -\sin(-\theta) \\ \sin(-\theta) & \cos(-\theta) \end{bmatrix} \\ &= \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \\ &= \begin{bmatrix} \cos^2\theta - \sin^2\theta & 2\sin\theta\cos\theta \\ 2\cos\theta\sin\theta & -(\cos^2\theta - \sin^2\theta) \end{bmatrix} = \begin{bmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{bmatrix}. \end{aligned}$$

Theorem 1. Any isometry φ that fixes the origin has equation $\varphi(X) = AX$, for some orthogonal matrix A. Every orthogonal matrix A has the form $A = A_\theta$ or $A = A_\theta H A_\theta^{-1}$ for some angle parameter θ.

Proof. You've already seen that for any orthogonal matrix A the transformation $X \to AX$ is an isometry that fixes O. Now consider an arbitrary isometry φ that fixes O. According to sections 6.6 and 6.7, φ must be

(0) the identity,
(1) a reflection σ_g across a line g,
(2) a composition of two reflections, or
(3) a glide reflection.

A glide reflection with a fixpoint is a reflection, so case (3) is included in case (1). By sections 6.4 and 6.5, a composition φ of two reflections is

(2a) a translation, or
(2b) a rotation ρ.

A translation with a fixpoint is the identity, so case (2a) is included in case (0). Since ρ fixes only its center and σ_g fixes only the points on g, just these cases remain:

(0) φ is the identity, with orthogonal matrix I, or
(1) φ is a reflection σ_g across a line g through the origin, with orthogonal matrix $A_\theta H A_\theta^{-1}$, or
(2b) φ is a rotation ρ about the origin, with orthogonal matrix A_θ. ♦

Finally, consider an *arbitrary* isometry ψ. Let $B = \psi(O)$ and τ be the translation with vector $-B$. The isometry $\varphi = \tau\psi$ fixes the origin:

$$\varphi(O) = \tau\psi(O) = \tau(B) = B - B = O.$$

By the previous paragraph, there's an orthogonal matrix A such that $\varphi(X) = AX$ for every point X. Since τ^{-1} is the translation with vector $-(-B) = B$,

$$\psi(X) = \tau^{-1}\varphi(X) = \tau^{-1}(AX) = AX + B.$$

The transformation determines A and B uniquely; for if $AX + B = CX + D$ for every point X, then substituting $X = O$ yields $B = C$, and substituting the kth unit vector shows that the kth columns of A and C are equal. This discussion has proved

Theorem 2. The plane isometries are the transformations $\varphi : X \to X' = AX + B$ where A is an orthogonal matrix and B is an arbitrary vector. A and B are determined uniquely by φ.

A more concise argument for theorem 2 is presented in the proof of theorem 7.1.7. It doesn't refer to matrices A_θ and H, but uses the uniqueness theorem and some simpler matrix algebra. But that proof conveys little geometric insight; it's reserved until you've gained more experience.

A rotation $\varphi : X \to X'$ about a center $P \neq O$ has equation $X' = AX + B$ for some orthogonal matrix A. Let τ be the translation with vector $-B$. Then $\tau\varphi : X \to AX$ is an even isometry that fixes O —that is, a rotation ρ_θ about O with angle parameter θ. Thus $A = A_\theta$. Rotations with *any* center correspond to angle parameters just like the rotations about O!

Parity

If A is an orthogonal matrix, then

$$1 = \det I = \det A^tA = \det A^t \det A = (\det A)^2,$$

hence $\det A = \pm 1$. If φ is an isometry with matrix A, this determinant is called its *parity*, par φ. A rotation ρ_θ about O —an even isometry— has parity $\det A_\theta = \cos^2\theta + \sin^2\theta = 1$. In contrast, a reflection σ_g across a line g through O —an odd isometry—has parity $\det(A_\theta H A_\theta^{-1}) = (\det A_\theta)(\det H)(\det A_\theta)^{-1} = \det H = -1$. Theorem 3 shows that this pattern persists; parity provides an easy algebraic way to tell when φ is even or odd.

Theorem 3. φ is even if par $\varphi = +1$, and odd if par $\varphi = -1$.

Proof. A transformation $\varphi : X \to X' = AX + B$ preserves orientation of a triangle ΔPQR just in case $\det[P' - R', Q' - R']$ and $\det[P - R, Q - R]$ have the same sign. But

$$\begin{aligned}
&\det[P' - R', Q' - R'] \\
&\quad = \det[(AP + B) - (AR + B), (AQ + B) - (AR + B)] \\
&\quad = \det[AP - AR, AQ - AR] \\
&\quad = \det[A(P - R), A(Q - R)] \\
&\quad = \det A[P - R, Q - R] \\
&\quad = \det A \det[P - R, Q - R] \\
&\quad = \pm\det[P - R, Q - R]. \blacklozenge
\end{aligned}$$

Theorem 4. par $\varphi\chi$ = par φ par χ and par φ^{-1} = par φ for any isometries φ and χ.

6.9 Classifying isometries

Concepts

Compositions of reflections
Identity, reflections, translations, rotations, and glide reflections
Groups of even isometries
Rotations about arbitrary centers
Compositions of rotations and translations
Sets of fixpoints and fixed lines
Conjugacy
Transforming coordinate systems
Conjugacy classes

In previous sections you've seen that all plane isometries have simple matrix equations, and you've studied some particular isometries, their matrices, and their relationships in considerable detail. Parts of these analyses aren't yet complete. For example, although you've seen that the sets of all translations and of all rotations about a point O form groups, you haven't studied in detail the compositions of translations and rotations, nor of rotations about different centers. This kind of work often involves analyzing an isometry with a complicated definition, to fit it into the pattern of this chapter: Is it the identity, a translation, rotation, reflection, or glide reflection? What is its vector or center or axis? You've already started using the appropriate tools, without much attention to them. This section organizes them. They are employed to solve the example problems just mentioned, and to study in detail another kind of relationship—conjugacy—that's used to analyze symmetries in chapter 8.

Classification

The structure theorem in section 6.6 displayed a preliminary classification. A plane isometry is

the identity, or
a reflection, or
a composition of reflections across two lines g and h, or
a composition of reflections across three lines g, h, and k.

These classes aren't disjoint, since $\iota = \sigma_g \sigma_g$ and $\sigma_g = \sigma_g \sigma_g \sigma_g$ for any g. Clearly, $\sigma_g \sigma_h$ is the identity only when $g = h$. And by corollaries 6.4.7 and 6.5.6 and theorem 6.7.2, $\sigma_g \sigma_h \sigma_k$ is a reflection just when g, h, and k are copencilar. Therefore, a slight modification of the language just displayed yields a classification with disjoint classes:

Theorem 1. A plane isometry is

(0) the identity, or
(1) a reflection, or
(2) a composition of reflections across two distinct lines, or
(3) a composition of reflections across three noncopencilar lines.

For the rest of this heading, use a fixed coordinate system with origin O. Theorem 1 permits further refinement:

Corollary 2. A plane isometry is

(0) the identity ι, or
(1) a reflection σ_g across a line g, or
(2a) a translation τ_V with vector $V \neq O$, or
(2b) a rotation ρ_θ with center O and $\theta \bmod 360° \neq 0°$, or
(3) a glide reflection $\sigma_g \tau_V$ with $V \neq O$.

You can often use fixpoint information effectively to identify an isometry. This table shows the fixpoint sets of the isometry classes in corollary 2:

Class	*Isometry*		*Fixpoint set*
(0)	ι		the entire plane
(1)	σ_g		the line g
(2a)	τ_V	with $V \neq O$	empty
(2b)	ρ_θ	with $\theta \bmod 360° \neq 0°$	the center O
(3)	$\sigma_g \tau_V$	with $V \neq O$	empty

In exercise 6.11.12 you can work out an analogous description of the sets of fixed lines. You'll find several examples of the use of fixpoint information later in this section.

Classes (0) and (2) together form the group of *even* isometries. Those in (1) and (3) are called *odd.* By theorem 6.8.3, you can tell whether an isometry is even or odd by ascertaining whether its *parity*—the determinant of its matrix—is +1 or -1.

Rotations and translations

The group of even isometries has some interesting subgroups. Corollary 6.4.3 showed that the translations constitute a subgroup. In chapter 8 a yet smaller one is used to analyze ornamental patterns: the group of all translations τ_{cV}, where V is a specified nonzero vector. This set forms a group because it's closed under composition and inversion, and contains the identity:

$$\tau_{cV}\tau_{dV} = \tau_{(c+d)V} \qquad \tau_{cV}^{-1} = \tau_{-(cV)} = \tau_{(-c)V} \qquad \iota = \tau_O = \tau_{0V}.$$

You can allow c and d to be any scalars, or restrict them to be integers. You get two different groups, depending on your choice.

Theorem 6.5.3 showed that the rotations about a specified center also form a subgroup. For those about the origin, you've seen that composition corresponds to addition of angle parameters: $\rho_\eta\rho_\theta = \rho_{\eta+\theta}$. This is also true for other centers. If $\varphi : X \rightarrow A_\eta X + B$ and $\chi : X \rightarrow A_\theta X + C$ are rotations about a center P with angle parameters η and θ, then

$$\begin{aligned}\varphi\chi(X) &= \varphi(A_\theta X + C) = A_\eta(A_\theta X + C) + B\\ &= A_\eta A_\theta X + A_\eta C + B = A_{\eta+\theta}X + A_\eta C + B,\end{aligned}$$

so $\varphi\chi$ has angle parameter $\eta + \theta$. It follows that φ^{-1} has parameter $-\eta$.

You can use the matrix representation of a rotation $\varphi : X \rightarrow A_\eta X + B$ with angle parameter η to analyze compositions of φ with other isometries. For example, if $\tau : X \rightarrow X + V$ is any translation, then

$$\begin{aligned}\varphi\tau(X) &= \varphi(X + V) = A_\eta(X + V) + B = A_\eta X + A_\eta V + B\\ \tau\varphi(X) &= \tau(A_\eta X + B) = A_\eta X + B + V,\end{aligned}$$

so $\varphi\tau$ and $\tau\varphi$ are both rotations with the same angle parameter η, unless φ was the identity. For another rotation $\chi : X \rightarrow A_\theta X + C$ you get

$$\varphi\chi(X) = \varphi(A_\theta X + C) = A_\eta(A_\theta X + C) + B = A_{\eta+\theta}X + A_\eta C + B.$$

As in the previous paragraph, $\varphi\chi$ is a rotation with parameter $\eta + \theta$, unless $A_{\eta+\theta}$ is the identity matrix, in which case $\varphi\chi$ is a translation. This proves

Theorem 3. The composition of a rotation that's not the identity with any translation is a rotation with the same angle parameter. The composition of rotations with parameters χ and ψ is a rotation with parameter $\eta + \theta$ unless $(\eta + \theta) \bmod 360° = 0°$. In that case, it's a translation.

You can use theorem 3 to derive the equation of a rotation ψ with angle parameter θ but unknown center P if you can find $B = \psi(O)$. Consider $\rho = \tau_{-B}\psi$; since $\rho(O) = O$, ρ is a rotation about O with angle parameter θ. You know its equation, and from $\psi = \tau_B\rho$ you can derive the equation for ψ.

Conjugacy

In section 6.8, the isometry $\rho_\theta\sigma_h\rho_\theta^{-1}$ played a major role in deriving the matrix equation for a reflection across an arbitrary line g through O. This pattern occurs so commonly, particularly in studying symmetry, that it's studied in detail. For any transformations μ and φ, the composition $\varphi\mu\varphi^{-1}$ is called the *conjugate of* μ *by* φ. When λ is a conjugate of μ, we write $\lambda \sim \mu$. The next two theorems present some basic properties of conjugacy. They're used frequently, without mention, in later discussions.

Theorem 4. Conjugacy is an equivalence relation. That is, for any λ, μ, and ν,

$$\begin{array}{ll} \mu \sim \mu & \text{(reflexivity)} \\ \lambda \sim \mu \leftrightarrow \mu \sim \lambda & \text{(symmetry)} \\ \lambda \sim \mu \ \& \ \mu \sim \nu \Rightarrow \lambda \sim \nu & \text{(transitivity).} \end{array}$$

Conjugate isometries have the same parity.

Proof. μ is its own conjugate by ι. If λ is the conjugate of μ by φ, then μ is the conjugate of λ by φ^{-1}. If λ is the conjugate of μ by φ and μ is the conjugate of ν by χ, then λ is the conjugate of ν by $\varphi\chi$:

$$\lambda = \varphi\mu\varphi^{-1} = \varphi\chi\nu\chi^{-1}\varphi^{-1} = (\varphi\chi)\nu(\varphi\chi)^{-1}.$$

Finally, if $\lambda = \varphi\mu\varphi^{-1}$, then $\operatorname{par}\lambda = \operatorname{par}\varphi \operatorname{par}\mu \operatorname{par}(\varphi^{-1}) = (\operatorname{par}\varphi)^2 \operatorname{par}\mu = \operatorname{par}\mu$, by corollary 6.8.4. ♦

The equivalence classes generated by the conjugacy relation are called *conjugacy classes*.

Theorem 5. The only transformation that's conjugate with the identity ι is ι itself. Now consider any transformations φ, χ, and ψ, and let χ' and ψ' denote the conjugates of χ and ψ by φ. Then

(a) $\chi'\psi'$ is the conjugate of $\chi\psi$ by φ,
(b) $(\chi')^{-1}$ is the conjugate of χ^{-1} by φ, and
(c) χ' and ψ' commute just in case χ and ψ commute.

Proof. First, $\varphi\iota\varphi^{-1} = \varphi\varphi^{-1} = \iota$. Moreover,

(a) $\varphi\chi\psi\varphi^{-1} = \varphi\chi\iota\psi\varphi^{-1} = \varphi\chi\varphi^{-1}\varphi\psi\varphi^{-1}$.

(b) $(\chi')^{-1} = (\varphi\chi\varphi^{-1})^{-1} = (\varphi^{-1})^{-1}\chi^{-1}\varphi^{-1} = \varphi\chi^{-1}\varphi^{-1}$.

(c) $$\begin{aligned}\chi'\psi' = \psi'\chi' &\leftrightarrow \varphi\chi\varphi^{-1}\varphi\psi\varphi^{-1} = \varphi\psi\varphi^{-1}\varphi\chi\varphi^{-1}\\ &\leftrightarrow \varphi\chi\psi\varphi^{-1} = \varphi\psi\chi\varphi^{-1}\\ &\leftrightarrow \chi\psi = \varphi^{-1}\varphi\psi\chi\varphi^{-1}\varphi\\ &\leftrightarrow \chi\psi = \psi\chi.\ \blacklozenge\end{aligned}$$

Transforming coordinate systems

The conjugacy concept is even more closely connected to the equations of isometries than the section 6.8 examples suggest. The connection is based on the idea of transforming a coordinate system $\mathscr{C}$. All you need to describe $\mathscr{C}$ is the isosceles right triangle ΔOU_1U_2 in figure 6.9.1 that consists of the origin O and the axis points U_1 and U_2 with coordinate 1. An isometry φ transforms $\mathscr{C}$ into another system $\mathscr{D}$ defined by the image ΔPV_1V_2 of ΔOU_1U_2. What if you use *both* coordinate systems—for instance, to study different aspects of some geometric figure? How are a point's $\mathscr{C}$ and $\mathscr{D}$ coordinates related? Theorem 6 answers that question, and theorem 7 connects these considerations with conjugacy.

Theorem 6. If an isometry φ transforms coordinate system $\mathscr{C}$ into system $\mathscr{D}$, then the $\mathscr{C}$ coordinates of each point X are the same as the $\mathscr{D}$ coordinates of $Y = \varphi(X)$.

Proof. Consider figure 6.9.1. The figure formed by ΔOU_1U_2 and X is congruent to that formed by ΔPV_1V_2 and Y. ♦

Theorem 7. Isometries χ and ψ are conjugate if and only if there are coordinate systems under which they have the same matrix equation.

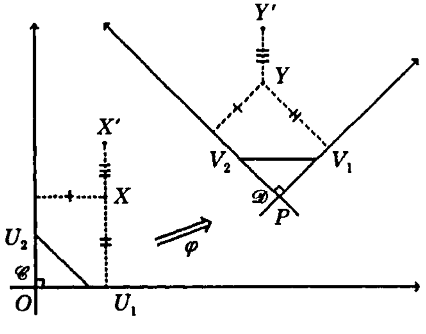

Figure 6.9.1
Transforming a coordinate system

Proof. Consider coordinate systems $\mathscr{C}$ and $\mathscr{D}$, points X and Y, and an isometry φ as in theorem 6. Suppose $\chi : X \to X'$ has equation $X' = AX + B$ with respect to $\mathscr{C}$. *Important: This equation treats X and X' as vectors of $\mathscr{C}$ coordinates!* By theorem 6, the $\mathscr{C}$ coordinates of X' are the same as the $\mathscr{D}$ coordinates of $Y' = \varphi(X')$. Thus

$$\begin{aligned} X' = AX + B &\leftrightarrow X' = \chi(X) \\ &\leftrightarrow \varphi^{-1}(Y') = \chi\varphi^{-1}(Y) \\ &\leftrightarrow Y' = \varphi\chi\varphi^{-1}(Y). \end{aligned}$$

If $\psi = \varphi\chi\varphi^{-1}$, then $X' = AX + B \leftrightarrow Y' = \psi(Y)$. Since the $\mathscr{D}$ coordinates of Y and Y' are the same as the $\mathscr{C}$ coordinates of X and X', you can conclude that

$$Y' = AY + B \leftrightarrow Y' = \psi(Y)$$

when you regard Y and Y' in the left-hand equation as vectors of $\mathscr{D}$ coordinates.

Conversely, suppose χ and ψ have the same equation with respect to coordinate systems $\mathscr{C}$ and $\mathscr{D}$. By theorem 6.6.3, $\mathscr{C}$ and $\mathscr{D}$ are related by an isometry φ as in figure 6.9.1. Since the $\mathscr{C}$ coordinates of X and X' are the same as the $\mathscr{D}$ coordinates of Y and Y',

$$\begin{array}{lclcl} X' = AX + B & \leftrightarrow & Y' = AY + B & \leftrightarrow & Y' = \psi(Y). \\ \text{with } X \text{ and } X' & & \text{with } Y \text{ and } Y' & & \\ \text{interpreted} & & \text{interpreted} & & \\ \text{as vectors of} & & \text{as vectors of} & & \\ \mathscr{C} \text{ coordinates} & & \mathscr{D} \text{ coordinates} & & \end{array}$$

But the first paragraph of this proof guaranteed $X' = AX + B \leftrightarrow Y' = \varphi\chi\varphi^{-1}(Y)$. Therefore $\psi(Y) = \varphi\chi\varphi^{-1}(Y)$ for all Y; having the same equation implies conjugacy. ♦

You now have *two* proofs of theorem 4; the word *same* in theorem 7 makes it *obvious* that conjugacy is an equivalence relation!

Conjugacy classes

By theorem 7, the half turns about any two points O and P are conjugate; with respect to coordinate systems with origins O and P, each has matrix form $X \to X' = -X$. By theorem 6.8.2, no other isometries have equations of this form, so the half turns constitute a single conjugacy class. But this result doesn't give complete information about the conjugate $\chi = \varphi\sigma_P\varphi^{-1}$ of σ_P by an isometry φ. Even without theorem 7, you can conclude that it's a half turn; χ is even and self-inverse. You can classify it completely by fixpoint analysis:

X is a fixpoint of χ
$\leftrightarrow \varphi\sigma_P\varphi^{-1}(X) = X$
$\leftrightarrow \sigma_P\varphi^{-1}(X) = \varphi^{-1}(X)$
$\leftrightarrow \varphi^{-1}(X)$ is a fixpoint of σ_P
$\leftrightarrow \varphi^{-1}(X) = P$
$\leftrightarrow X = \varphi(P)$.

Thus $\chi = \varphi\sigma_P\varphi^{-1} = \sigma_{\varphi(P)}$ because χ has exactly one fixpoint, $\varphi(P)$.

You can just as easily analyze reflections. By theorem 7, reflections across any two lines g and h are conjugate; with respect to coordinate systems with first axes g and h, each has the matrix form

$$X \to X' = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} X.$$

By theorem 6.8.2, no other isometries have equations of this form, so the reflections constitute a single conjugacy class. Again, this result doesn't give complete information about the conjugate $\psi = \varphi\sigma_g\varphi^{-1}$ of σ_g by an isometry φ. You can classify it completely by fixpoint analysis:

X is a fixpoint of ψ
$\leftrightarrow \varphi\sigma_g\varphi^{-1}(X) = X$
$\leftrightarrow \sigma_g\varphi^{-1}(X) = \varphi^{-1}(X)$
$\leftrightarrow \varphi^{-1}(X)$ is a fixpoint of σ_g
$\leftrightarrow \varphi^{-1}(X)$ lies on g
$\leftrightarrow X$ lies on the image set $\varphi[g]$.

The fixpoints of ψ constitute the line $\varphi[g]$, so ψ must be the reflection across that line. That is, $\varphi\sigma_g\varphi^{-1} = \sigma_{\varphi[g]}$. This paragraph and the previous one have proved

Theorem 8. The half turns constitute a single conjugacy class. Moreover, $\varphi\sigma_P\varphi^{-1} = \sigma_{\varphi(P)}$ for any point P and any isometry φ. Similarly, the reflections constitute a single conjugacy class, and $\varphi\sigma_g\varphi^{-1} = \sigma_{\varphi[g]}$ for any line g and any isometry φ.

Analyzing translations is a little more complicated. Consider a point O and translations τ and υ; let

$$T = \tau(O) \qquad t = OT$$
$$U = \upsilon(O) \qquad u = OU.$$

By theorem 7, τ and υ are conjugate just in case $t = u$. With respect to coordinate systems with origin O and with T and U on the positive rays of the first axes, they have matrix forms

$$X \to X' = X + \begin{bmatrix} t \\ 0 \end{bmatrix} \qquad X \to X' = X + \begin{bmatrix} u \\ 0 \end{bmatrix},$$

and these are the same just in case $t = u$. Theorem 9 provides complete information about the conjugate $\upsilon = \varphi\tau\varphi^{-1}$ of τ by an isometry φ:

Theorem 9. The conjugacy class of a translation τ whose vector has length t consists of all translations with vectors of that length. If τ and an isometry φ have equations $\tau: X \to X' = X + T$ and $\varphi: X \to X' = AX + B$ with respect to a single coordinate system, then $\upsilon = \varphi\tau\varphi^{-1}$ is the translation with vector AT.

Proof. The second sentence remains to be proved. Since $X = A^{-1}X' - A^{-1}B$, it follows that φ^{-1} has matrix form $X \to A^{-1}X - A^{-1}B$ and

$$\begin{aligned} \upsilon(X) &= \varphi\tau\varphi^{-1}(X) = \varphi\tau(A^{-1}X - A^{-1}B) = \varphi(A^{-1}X - A^{-1}B + T) \\ &= A(A^{-1}X - A^{-1}B + T) + B = X + AT. \blacklozenge \end{aligned}$$

The final results of this section provide analogous information for glide reflections and rotations. Their proofs differ. The argument for theorem 10 can use the concise techniques of function composition. But theorem 11 uses matrix computation, because that's convenient for handling angle parameters.

Theorem 10. Choose a coordinate system. The conjugacy class of a glide reflection χ with axis g and vector V consists of all glide reflections with vectors whose length is the same as that of V. The conjugate of χ by an isometry $\varphi: X \to X' = AX + B$ is the glide reflection with axis $\varphi[g]$ and vector AV.

Proof. Find points O and P on g with $V = P - O$, so that $\tau_V = \sigma_O\sigma_P$ and $\chi = \sigma_g\tau_V$. If φ is an isometry, then by theorems 8 and 9,

$$\varphi\chi\varphi^{-1} = \varphi\sigma_g\varphi^{-1}\varphi\tau_V\varphi^{-1} \qquad \varphi\sigma_g\varphi^{-1} = \sigma_{\varphi[g]} \qquad \varphi\tau_V\varphi^{-1} = \tau_{AV}.$$

Define $O' = \varphi(O)$, $P' = \varphi(P)$, and $V' = P' - O'$, so that O' and P' lie on $\varphi[g]$ and

$$\varphi\tau_V\varphi^{-1} = \varphi\sigma_O\varphi^{-1}\varphi\sigma_P\varphi^{-1} = \sigma_{O'}\sigma_{P'} = \tau_{V'}.$$

Thus $\varphi\chi\varphi^{-1}$ is the glide reflection with axis $\varphi[g]$ and vector V'.

Conversely, assuming that ψ is a glide reflection with axis g' whose vector V' has the same length as V, find points O' and P' on g' with $V' = P' - O'$, so that $\tau_{V'} = \sigma_{O'}\sigma_{P'}$ and $\psi = \sigma_{g'}\tau_{V'}$. By theorem 6.6.3, there's an isometry φ such that $\varphi(O) = O'$ and $\varphi(P) = P'$. It follows that $\varphi[g] = g'$. Then

$$\begin{aligned}
\sigma_{g'} &= \sigma_{\varphi[g]} = \varphi\sigma_g\varphi^{-1}\\
\tau_{V'} &= \sigma_{\varphi(O)}\sigma_{\varphi(P)} = \varphi\sigma_O\varphi^{-1}\varphi\sigma_P\varphi^{-1} = \varphi\sigma_O\sigma_P\varphi^{-1} = \varphi\tau_V\varphi^{-1}\\
\psi &= \varphi\sigma_g\varphi^{-1}\varphi\tau_V\varphi^{-1} = \varphi\sigma_g\tau_V\varphi^{-1} = \varphi\chi\varphi^{-1}. \ \blacklozenge
\end{aligned}$$

Theorem 11. Suppose χ is a rotation, not the identity, with angle parameter θ and center P. Its conjugate by an isometry φ is a rotation ψ with center $\varphi(P)$. Moreover, ψ has angle parameter θ or $-\theta$, depending on whether φ is even or odd.

Proof. $\psi = \varphi\chi\varphi^{-1}$ is a rotation with center $\varphi(P)$ because that's its only fixpoint:

$$\varphi\chi\varphi^{-1}(X) = X \leftrightarrow \chi\varphi^{-1}(X) = \varphi^{-1}(X) \leftrightarrow \varphi^{-1}(X) = P \leftrightarrow X = \varphi(P).$$

Suppose φ is even. From the matrix representations $\varphi : X \to A_\eta X + B$ and $\chi : X \to A_\theta X + C$, you get

$$\begin{aligned}
\varphi\chi\varphi^{-1}(X) &= A_\eta A_\theta A_\eta^{-1} X + \text{a constant vector}\\
&= A_\eta A_\theta A_{-\eta} X + \cdots\\
&= A_{\eta+\theta-\eta} X + \cdots\\
&= A_\theta X + \cdots,
\end{aligned}$$

so $\varphi\chi\varphi^{-1}$ is a rotation with parameter θ.

Now suppose ω is an odd isometry. Consider the even isometry $\varphi = \sigma_h\omega$, where h is the first axis. Then $\varphi\chi\varphi^{-1}$ is a rotation ρ with parameter θ:

$$\begin{aligned}
&\rho = (\sigma_h\omega)\chi(\sigma_h\omega)^{-1} = \sigma_h\omega\chi\omega^{-1}\sigma_h^{-1} = \sigma_h(\omega\chi\omega^{-1})\sigma_h\\
&\sigma_h\rho\sigma_h = \omega\chi\omega^{-1}.
\end{aligned}$$

From the matrix equation $\rho : X \to A_\theta X + B$, you see that

$$\omega\chi\omega^{-1}(X) = HA_\theta HX + \text{a constant vector,}$$

where H is the matrix corresponding to the reflection σ_h. Thus

$$H = \begin{bmatrix} 1 & 0\\ 0 & -1 \end{bmatrix}$$

$$\begin{aligned}
HA_\theta H &= \begin{bmatrix} 1 & 0\\ 0 & -1 \end{bmatrix}\begin{bmatrix} \cos\theta & -\sin\theta\\ \sin\theta & \cos\theta \end{bmatrix}\begin{bmatrix} 1 & 0\\ 0 & -1 \end{bmatrix}\\
&= \begin{bmatrix} \cos\theta & -\sin\theta\\ -\sin\theta & -\cos\theta \end{bmatrix}\begin{bmatrix} 1 & 0\\ 0 & -1 \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta\\ -\sin\theta & \cos\theta \end{bmatrix}\\
&= \begin{bmatrix} \cos(-\theta) & -\sin(-\theta)\\ \sin(-\theta) & \cos(-\theta) \end{bmatrix} = A_{-\theta}.
\end{aligned}$$

Therefore, $\omega\chi\omega^{-1}$ is a rotation with parameter $-\eta$. ♦

Corollary 12. Rotations with angle parameters η and θ are conjugate just in case $|\eta \bmod 360°| = |\theta \bmod 360°|$.

You can use theorem 11 to derive the equation of a rotation ψ with center P, as follows. Let ρ be the rotation about the origin with the same angle parameter, and let $\varphi = \tau_P$, so that $\psi = \varphi\rho\varphi^{-1}$. The equations of φ, ρ, and φ^{-1} are easy to find, and from those you derive the equation for ψ.

6.10 Similarities

Concepts

Similarities and their ratios
Contractions and dilations
Invariants
Similarity group
Uniqueness theorem
Classification theorem
Equation $X' = rAX + B$ of a similarity
Fixpoint theorem
Homothetic rotations and reflections
Even and odd similarities

Studying geometric transformations more general than isometries usually requires methods beyond those developed in this chapter. But one broader family of transformations, the similarities, needs little new. Let r be a positive real number. A *plane similarity with ratio* r is a plane transformation $\psi : P \to P'$ such that for all points P and Q,

$$P'Q' = r\,(PQ).$$

A similarity can't have two *different* ratios because $r\,(PQ) = r'\,(PQ)$ and $P \neq Q$ imply $r = r'$; so it makes sense to call r *the* ratio of ψ.

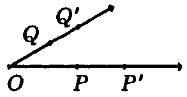

Figure 6.10.1 Dilation about O with ratio r

$$r = 1.7 \qquad OP' = r\,(OP)$$
$$OQ' = r\,(OQ)$$

Evidently, every isometry is a similarity with ratio 1. As another example, select any point O and consider the transformation $\varphi_{r,O} : P \rightarrow P'$ such that $O' = O$ and for $P \neq O$, P' is the point on $\overrightarrow{OP}$ with $OP' = r\,(OP)$. (See figure 6.10.1.) If O is the origin of a Cartesian coordinate system, so that you can regard points as vectors, then $P' = rP$ for every point P. If $r < 1$ or $r > 1$, then $\varphi_{r,O}$ is called a *contraction* or *dilation about* O.[22]

Theorem 1. $\varphi_{r,O}$ is a similarity with ratio r.

Proof. By theorem 3.11.2,

$$\begin{aligned}(P'Q')^2 &= (P' - Q')\cdot(P' - Q') = (rP - rQ)\cdot(rP - rQ)\\ &= r^2(P - Q)\cdot(P - Q) = r^2(PQ)^2. \blacklozenge\end{aligned}$$

You'll see soon, in the classification theorem, that *every* similarity is an isometry or the composition of an isometry with a dilation or contraction. That's why there's not much new in this section. First, though, theorems 2 and 3 record some general properties of similarities that you'll need for this analysis.

Theorem 2. A similarity $\psi : X \rightarrow X'$ with ratio r leaves angle measure and betweenness invariant. Thus any geometric property that you can define in terms of betweenness is invariant. In particular, ψ is a collineation. The image of a circle with radius t is a circle with radius rt. The image of a triangle with area s is a triangle with area $r^2 s$.

Proof. The equation $m\angle XYZ = m\angle X'Y'Z'$ follows from the SSS similarity principle. For betweenness, proceed as in the proof of theorem 6.2.5, taking into account ratio r:

$$\begin{aligned}X\text{-}Y\text{-}Z &\leftrightarrow XY + YZ = XZ\\ &\leftrightarrow r\,(XY) + r\,(YZ) = r\,(XZ)\\ &\leftrightarrow X'Y' + Y'Z' = X'Z'\\ &\leftrightarrow X'\text{-}Y'\text{-}Z'.\end{aligned}$$

If O is the center of circle Γ with radius t, then a point X lies on Γ if and only if $OX = t$. This condition is equivalent to $O'X' = rt$, so X lies on Γ if and only if X' lies on the circle with center O' and radius rt. The area result follows from theorem 3.9.7. ♦

Theorem 3. A composition of similarities with ratios r and s is a similarity with ratio rs. The inverse of a similarity with ratio r is a similarity with ratio r^{-1}. Therefore, the similarities form a transformation group that contains all isometries.

---

[22] Many authors also use the term *dilation* when $r < 1$. That jars, so this text uses *contraction* in that case. Don't confuse *dilation* with *dilatation*; the latter term has a different meaning —see the heading Dilatations in section 6.11.

Proof. If $P \to P'$ and $P' \to P''$ are similarities with ratios s and r, then for all points P and Q, $P''Q'' = r\,(P'Q') = rs\,(PQ)$, so the composition $P \to P''$ is a similarity with ratio rs. Moreover, $PQ = r^{-1}r(PQ) = r^{-1}(P'Q')$, so the inverse transformation $P' \to P$ is a similarity with ratio r^{-1}. ♦

Corollary 4 ***(Uniqueness theorem).*** Plane similarities χ and ψ coincide if $\chi(X) = \psi(X)$ for three noncollinear points X.

Proof. Suppose χ and ψ have ratios r and s, points X_1 and X_2 are distinct, $\chi(X_1) = \psi(X_1)$, and $\chi(X_2) = \psi(X_2)$. Then $r\,(X_1X_2) = \chi(X_1)\chi(X_2) = \psi(X_1)\,\psi(X_2) = s\,(X_1X_2)$, hence $r = s$. Since its ratio is $r/s = 1$, the similarity $\varphi = \chi\psi^{-1}$ is an isometry. Since φ fixes three noncollinear points, it's the identity, by theorem 6.2.7. ♦

Lemma 5 is a preliminary version of theorem 8. It's introduced now, to facilitate use of the equation of a similarity in the final version of theorem 8.

Lemma 5. Given any point O and any similarity ψ with ratio r, there's an isometry χ such that $\psi = \chi\varphi_{r,O}$.

Proof. The composition $\chi = \psi\varphi_{r,O}^{-1}$ is a similarity with ratio $rr^{-1} = 1$, hence it's an isometry. Moreover, $\psi = \chi\varphi_{r,O}$. ♦

Theorem 6. The similarities are the transformations $X \to X' = rAX + B$ where $r > 0$, A is an orthogonal matrix, and B an arbitrary vector. The transformation determines r, A, and B uniquely; r is the similarity ratio.

Proof. Every such equation defines a similarity: the composition of the isometry $X \to AX + B$ with $\varphi_{r,O}$. On the other hand, given a similarity $\psi : X \to X'$, lemma 5 provides an isometry χ such that $\psi = \chi\varphi_{r,O}$. By theorem 6.8.2 there's an orthogonal matrix A and a vector B such that $\chi(X) = AX + B$ for all X, so that ψ has the desired equation. If equations $X' = rA_1X + B_1$ and $X' = sA_2X + B_2$ represented the same transformation, it would be a similarity with ratio $r = s$. But then the transformation $X \to X'' = r^{-1}X'$ would be an isometry with equations $X'' = A_1X + r^{-1}B_1$ and $X'' = A_2X + r^{-1}B_2$, so $A_1 = A_2$ and $r^{-1}B_1 = r^{-1}B_2$ by theorem 6.8.2, hence $B_1 = B_2$. ♦

Theorem 7 ***(Fixpoint theorem).*** If $\psi : X \to X'$ is a similarity with ratio r, then $r = 1$ or ψ has exactly one fixpoint.

Proof. First consider the case $r < 1$. Select any point X_0 and define points $X_1, X_2, \ldots$ by the rule $X_{m+1} = X'_m$. Then

$$\begin{aligned} X_1X_2 &= X_0'X_1' = r\,(X_0X_1), \\ X_2X_3 &= X_1'X_2' = r\,(X_1X_2) = r^2\,(X_0X_1), \\ &\vdots \end{aligned}$$

so in general,

$$X_mX_{m+1} = r^m\,(X_0X_1).$$

Now let $m < n$ and consider the ith coordinates x_{mi} and x_{ni} of points X_m and X_n:

$$\begin{aligned} |x_{mi} - x_{ni}| &\le X_mX_n \le X_mX_{m+1} + X_{m+1}X_{m+2} + \cdots + X_{n-1}X_n \\ &= r^m\,(X_0X_1) + r^{m+1}\,(X_0X_1) + \cdots + r^n\,(X_0X_1) \\ &= r^m\,(1 + r + \cdots + r^{n-m})(X_0X_1). \end{aligned}$$

The sum in parentheses is the initial part of a convergent geometric series, so

$$|x_{mi} - x_{ni}| \le r^m\left(\sum_{n=0}^{\infty} r^n\right)(X_0X_1) = \frac{r^m\,(X_0X_1)}{1-r}.$$

Since $\lim_m r^m = 0$, you can make $|x_{mi} - x_{ni}|$ arbitrarily small by taking m sufficiently large and $n > m$. That is, $x_{0i}, x_{1i}, x_{2i}, \ldots$ is a Cauchy sequence; it has a limit x_i. Let X be the vector whose ith component, for each i, is x_i. Then X is a fixpoint. To see that, use the equation $X' = rAX + B$ provided by theorem 6:

$$\begin{aligned} i\text{th component of } X' &= i\text{th component of } rAX + B \\ &= (\textstyle\sum_j r a_{ij} x_j) + b_i \\ &= (\textstyle\sum_j r a_{ij} \lim_n x_{nj}) + b_i \\ &= (\lim_n \textstyle\sum_j r a_{ij} x_{nj}) + b_i \\ &= \lim_n ((\textstyle\sum_j r a_{ij} x_{nj}) + b_i) \\ &= \lim_n (i\text{th component of } rAX_n + B) \\ &= \lim_n (i\text{th component of } X_n') \\ &= \lim_n (i\text{th component of } X_{n+1}) \\ &= \lim_n x_{n+1,i} = x_i = i\text{th component of } X. \end{aligned}$$

Thus $X' = X$, as claimed.

Now consider the case $r > 1$. Since ψ^{-1} is a similarity with ratio < 1, it has a fixpoint X by the preceding paragraph. But then X is a fixpoint of ψ, too.

Finally, to prove uniqueness, suppose $r \ne 1$ and ψ has fixpoints X and Y. Then $X'Y' = r\,(XY) = r\,(X'Y')$, so $X'Y' = 0$, $X' = Y'$, and $X = Y$. ◆

The Cauchy sequence technique for proving the fixpoint theorem has a flavor different from that of the rest of this book. The only other part based on limits is the discussion of π in section 3.14. Cauchy sequences occur frequently in real analysis. In fact, similarities with ratio $r < 1$ are examples of the *contraction mappings* that you study in analysis: mappings such that $X'Y' \le r\,(XY)$ for all X and Y. And the proof of the fixpoint theorem is just an adaptation of a familiar proof of the contraction mapping theorem. That result[23] —*every* contraction mapping has a fixpoint—has several major applications. For example, it establishes the convergence of various iterative approximation methods such as the Newton–Raphson technique, and it implies the existence of solutions to first-order ordinary differential equations satisfying Lipschitz conditions. You can prove the fixpoint theorem without using analysis methods, but that proof will be less succinct, and may require different arguments for two and three dimensions.[24]

Compositions $\chi\varphi_{r,O}$ where χ is a rotation about O or a reflection across a line through O are called *homothetic rotations* or *reflections.*[25]

Theorem 8 ***(Classification theorem).*** A plane similarity is an isometry, a homothetic rotation, or a homothetic reflection.

Proof. By the fixpoint theorem, if a similarity ψ isn't an isometry, then it has a fixpoint O. Let r be its ratio, and define $\chi = \psi\varphi_{r^{-1},O}$, so that $\psi = \chi\varphi_{r,O}$. Then χ is an isometry with fixpoint O'. By the discussion following corollary 6.9.2, it's a rotation about O' or a reflection across a line g through O'. ♦

A similarity $\psi : X \to X' = rAX + B$ with $r > 0$ preserves orientation just in case the isometry $\chi : X \to AX + B$ does—that is, just in case $\det A = 1$. To see this, note that for any points P, Q, and R,

$$\begin{aligned}
&\det[P' - R', Q' - R'] \\
&\quad = \det[(rAP + B) - (rAR + B), (rAQ + B) - (rAR + B)] \\
&\quad = \det[rA(P - R), rA(Q - R)] \\
&\quad = (\det rA)\det[P - R, Q - R] \\
&\quad = r^2(\det A)\det[P - R, Q - R].
\end{aligned}$$

ψ is called *even* if orientation is preserved; *odd* if it's reversed.

---

[23] See Bartle 1964, 170 for the contraction mapping theorem.

[24] See Exercises 6.11.41–44 and Martin 1982, 141.

[25] *Homothetic* stems from the Greek prefix *homo-* and verb *tithenai,* which mean *same* and *put.* It's the opposite of *antithetic.*

6.11 Exercises

Concepts

- Curved mirrors
- Classifying isometries defined as compositions or by equations
- Constructions with isometries and similarities
- Billiards and dual billiards problems
- Equations of reflections
- Determining equations of isometries and similarities
- Coordinate transformations
- *Alias* and *alibi* interpretations of equations
- Reflection calculus
- Distance-preserving functions
- Dilatations

This section contains forty-five exercises related to earlier parts of the chapter. Like other exercises in this book, they're essential components of its development. A few—for example, exercises 1, 5, and 12—provide more experience with reasoning techniques introduced earlier. Others, particularly those on the classification and the equations of isometries, show you how to do the computations required to use the theorems derived in the text. The theory is streamlined into an efficient outline of the results important for applications. But those computations seem unwieldy at first. That would obscure the main points of the theory, so the computations are clustered here. Some may require awkward matrix arithmetic. There was no effort to avoid that. If you find a calculation too tedious, use mathematical software.[26]

This chapter's theory is often applied to analytic geometry problems. That work usually has a different flavor because it uses *coordinate* transformations. Those apply to the numbers used to locate points, not to the points themselves, like geometric transformations here. Exercise 31 shows how to make the transition. A few exercises introduce topics that could be developed much further. Those on reflection calculus, and on dilatations, for example, lead into deep studies in foundations of geometry. Chapter 7 extends the theory of isometries and similarities to three dimensions. Most of the exercises in this section can be extended as well. You'll see that in section 7.7.

Basic exercises

Exercises 1 and 2 present some important properties that apply to transformations in general, not just those in geometry. For instance, exercise 2, part 1, asks what you can say about functions φ and χ whose composition is surjective, and what you can *not* say, in general. To show that you can *not* claim that φ or χ has some property $\mathscr{P}$, you must provide a *counterexample*: specific functions whose composition is surjective, but for which $\mathscr{P}$ is not true. Searching for counterexamples for exercise 2, you should concentrate on functions whose domains and ranges are *very small* finite sets. You won't need more than that to make the required properties fail, and you'll avoid confusion by other phenomena that are really irrelevant. That kind of reasoning may seem nongeometric to you, but no—those tiny counterexamples are always contained in the larger ones that may seem closer to geometry, and they're much easier to understand.

---

[26] *X(Plore)* (Meredith 1993) is a good choice.

Exercise 1. Prove that two self-inverse transformations of a set D commute if and only if their composition is self-inverse.

Exercise 2. Consider functions $\varphi : S \to T$ and $\chi : D \to S$, so that $\varphi\chi : D \to T$.

Part 1. Prove that if $\varphi\chi$ is surjective, then φ must be surjective. Find an example where $\varphi\chi$ is surjective, but χ is neither surjective nor injective.

Part 2. Prove that if $\varphi\chi$ is injective, then χ must be injective. Find an example where $\varphi\chi$ is injective, but φ is neither injective nor surjective.

Part 3. Prove that if $\varphi\chi$ is bijective, then φ must be surjective and χ injective. Find an example where $\varphi\chi$ is bijective, but neither φ nor χ is.

Exercises 3 and 4 present some fundamental facts about linear functions. It wasn't necessary to state them in the text of chapter 6 or 7 because all linear functions considered there are isometries or similarities, and for those function classes these facts are evident. Had chapter 6 introduced isometries through their equations, however, the results in these exercises would have been incorporated in the text before that point. To prove them, use appendix C material about Gauss elimination, matrix invertibility, and determinants.

Exercise 3. Consider a matrix A, a vector B, and the linear function $X \to AX + B$ that maps $\mathbb{R}^n$ to itself, where $n = 2$ or 3. Show that these conditions are equivalent:

(a) A is invertible,
(b) $X \to AX + B$ is bijective,
(c) $X \to AX + B$ is injective,
(d) $X \to AX + B$ is surjective.

Exercise 4. Let $n = 2$ or 3. Show that the range of a linear function $X \to AX + B$ that maps $\mathbb{R}^n$ to itself is a point, line, plane, or the entire space. In which cases is φ bijective? *Suggestion*: First consider the function $\varphi : X \to AX$.

Theorem 6.6.3 shows that you can map any triangle onto any congruent one with an isometry; the uniqueness theorem, corollary 6.2.8, says that you can do that in only one way. The latter statement requires you to regard a triangle as an *ordered* triple of vertices. You'd lose the uniqueness if you defined a triangle as a mere *set* of vertices or as its boundary—the union of its edges—for there are several isometries that map the set of vertices and the boundary of an equilateral triangle onto themselves. You'll see much more detail on these considerations in chapter 8. In exercise 5 you can study the same question for a simpler example, a segment. As defined in section 3.2, a segment is the *set* of points between its vertices, *not* an ordered pair of vertices.

Exercise 5. Show that if Σ and Υ are segments of the same length, then there's a rotation ρ such that $\rho[\Sigma] = \Upsilon$. Could there be *another* isometry φ such that $\varphi[\Sigma] = \Upsilon$? Could there be another *rotation*?

At the beginning of section 6.3 you saw how to find the shortest path from one point to another via a point on a mirror line g by reflecting one of the points across g. And you saw how this relationship corresponds to the equality of incidence and reflection angles. Exercise 6 shows that the relationship holds for a smoothly curved mirror as well. You need differential calculus to work with the curve.

Exercise 6. Consider a smooth curve Γ and points A and B not on Γ. (See figure 6.11.1.) Suppose point O lies on Γ and

$$AO + OB \leq AX + XB$$

for every point X on Γ. Let g be the tangent to Γ at O and select points U and V on g such that U-O-V. Prove that $\angle UOA \cong \angle VOB$. *Suggestion*: Use a coordinate system with first axis g and origin O. Since Γ is smooth, there's a differentiable function f on an open interval $\mathscr{I}$ containing 0 such that for any $x_1 \in \mathscr{I}$, a point $X = \langle x_1, x_2 \rangle$ lies on Γ just in case $x_2 = f(x_1)$.

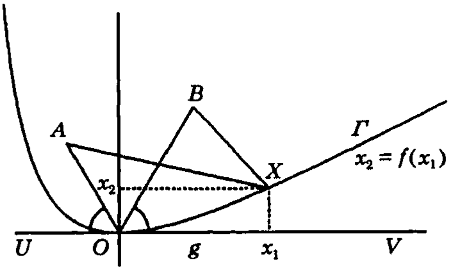

Figure 6.11.1
Exercise 6

The next exercise invites you to use a technique from section 6.3 to derive Fagnano's theorem, a morsel of triangle geometry mentioned in section 5.8.

Exercise 7. Show that of all triangles with vertices X, Y, and Z on the edges of an acute ΔABC, the orthic triangle has the shortest perimeter. *Suggestion*: Show first that for a given X on $\overline{BC}$ the shortest $XY + YZ + ZX$ is $\sigma_{\overleftrightarrow{AB}}(X)\,\sigma_{\overleftrightarrow{CA}}(X) = 2(AX)\sin \mathrm{m}\angle A$.

Classifying Isometries

Most exercises under this heading ask you to classify an isometry. That means decide whether it's the identity or one of these types of isometries, and provide the information in the right-hand column:

For a	*provide*
reflection	axis
translation	vector
rotation	center and an angle parameter
glide reflection	axis and vector.

Exercises 8 and 9 are aperitifs—you'll use this classification to verify some properties of isometries in general.

Exercise 8. Show that an isometry φ is even if and only if $\varphi = \chi\chi$ for some isometry χ.

Exercise 9. Show that every plane isometry is the composition of two self-inverse isometries.

Exercise 10. Prove a result like theorem 6.9.3 by continuing the proofs of corollaries 6.5.7 and 6.5.8 and analyzing the angles formed by the lines mentioned there, without referring to any equations of transformations. *Suggestion*: Be careful about the *signs* of the angle parameters. You don't need to get precisely the same result as the theorem. It was postponed from section 6.5 to section 6.9 because consistent assignment of angle parameters for different centers is so hard to describe.

Exercise 11 presents a nearly obvious result that you may use in several later exercises to find the axis of a glide reflection.

Exercise 11. Show that the axis of a glide reflection $\varphi : X \to X'$ consists of the midpoints of all segments $\overline{XX'}$.

Exercise 12. Construct a table analogous to the one following corollary 6.9.2, to show the possible sets of lines that are fixed by the various isometries. A fixed line might or might not consist entirely of fixed points. Lines that do are called *pointwise* fixed. What isometries have pointwise fixed lines? What isometries have fixed lines, *no* point of which is fixed?

The remaining exercises under this heading and some later ones require you to classify various isometries according to the scheme presented earlier. The main tools available are their fixpoints and parity—discussed in earlier sections—and exercises 11 and 12. *Use those tools!*

Exercise 13. Let $V_1V_2\cdots V_n$ be a convex polygon and $\theta_1, \theta_2, \ldots, \theta_n$ be the measures of the corresponding vertex angles. At each vertex V_i draw arc $\widehat{A_iB_i}$ with measure θ_i and radius $V_iA_i = V_iB_i$ small enough that no two of these arcs intersect. (See figure 6.11.2 for an example with $n = 4$.) Consider translations

τ_1 with vector $\overrightarrow{B_1A_2}$
τ_2 with vector $\overrightarrow{B_2A_3}$
$\vdots$ $\vdots$
τ_n with vector $\overrightarrow{B_nA_1}$

and for $i = 1, \ldots, n$ consider rotation ρ_i with center V_i and angle parameter θ_i. Classify the isometry $\tau_n\rho_n\tau_{n-1}\rho_{n-1}\cdots\tau_1\rho_1$.[27]

Exercise 14. Classify the composition of glide reflections with perpendicular axes g and h and vectors V and W.

Exercise 15. Classify the compositions of the reflection across the first coordinate axis and the translation with vector $<1, 2>$, in both orders.

Exercise 16. Let a, b, and c denote the lines containing the edges opposite vertices A, B, and C of ΔABC. Let D, E, and F denote the feet of

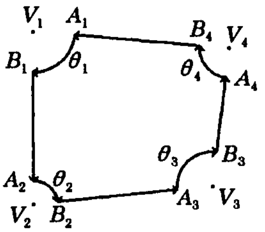

Figure 6.11.2
Exercise 13

---

[27] This exercise is due to Lawrence Levy (1970, 32).

the corresponding altitudes. Show that D and F lie on the axis of $\varphi = \sigma_a\sigma_b\sigma_c$. Classify φ when

(a) $m\angle A = m\angle B = 60°$ (c) $m\angle A = m\angle C = 45°$
(b) $m\angle A = m\angle B = 45°$ (d) $m\angle A = 30°$ and $m\angle C = 90°$.

Constructions

Following a more traditional approach to geometry, you'd have paid more attention to constructing the objects under study—probably with tools limited to straightedge and compasses. In fact, Euclid's axiomatization is sometimes viewed as a study of just those constructions, rather than of the space in which we live. This text has avoided them because they're no longer in the mathematical mainstream. The next seven exercises, however, merge transformational geometry into the older stream.[28] Each can be solved by traditional methods with no reference to transformations, but the isometries and similarities you've studied in this chapter shed more light on them. And if you seek the *simplest* constructions with straightedge and compasses, transformations may point the way.

Exercise 17. Suppose you want to carom ball A off the back, right, and front walls g, h, and k of a billiards table, to strike ball B, as in figure 6.11.3. How do you find the point P on g at which you must aim? *Suggestion*: Try more modest coups first.

Exercise 18. Consider nonparallel lines $g \nparallel h$ and three points P, Q, and R. Find a three-segment path beginning on g and ending on h, such that P, Q, and R are the midpoints of its segments. Can there be more than one such path? What if $g = h$ or $g \parallel h$?

Did your solution of exercise 17 require that the table be rectangular? Probably not. Could you extend it to apply to a convex polygonal table with any number of edges? Probably so—you can buy or build novelty billiards games like that. Exercise 18 didn't place any limitations on points

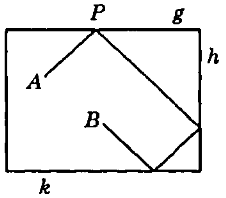

Figure 6.11.3
Exercise 17

---

[28] Several exercises under this heading are due to Max Jeger [1964] 1966.

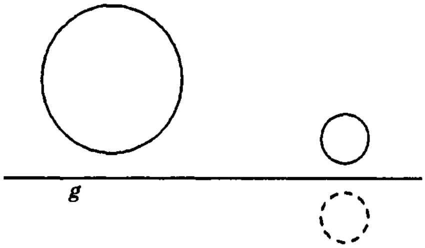

Figure 6.11.4
Exercise 19

P, Q, and R. Can you extend it to apply to any number of points? Exercises 34 and 35 consider *closed* polygonal paths of this sort.

Exercise 19. Consider two circles on the same side of line g. Find a point P on g such that tangents through P to the circles make equal angles with g. How many such points are there? *Suggestion*: Reflect one of the circles across g, as in figure 6.11.4.

Exercise 20. Consider two points A and B on the same side of line g. Find points O, P, and Q on g such that $2\text{m}\angle OPA = \text{m}\angle QPB$. *Suggestion*: Consider exercise 19.

Exercise 21. Consider a point P and concentric circles Γ and Δ. Look for points Q and R on Γ and Δ such that ΔPQR is equilateral. Under what condition on P, Γ, and Δ is a solution possible? How many solutions can there be? *Suggestion*: On which *two* circles must Q lie?

Exercise 22. Given ΔABC with $\text{m}\angle A < 90°$, find points C_1 and C_2 on the edge opposite C and points A_1 and B_1 on the edges opposite A and B such that $A_1B_1C_1C_2$ is a rectangle and $C_1C_2 = 2(B_1C_1)$.

Exercise 23. Consider a point B, two circles Γ and Δ, and any number $t > 0$. Find points C and D on Γ and Δ such that C-B-D and $BC/BD = t$. How many solutions could there be? Describe, as completely as you can, the set $\mathscr{B}$ of points B for which a solution is possible. *Suggestions*: On which *two* circles must C lie? Describing $\mathscr{B}$ is difficult; you may want to plot the points B corresponding to several thousand pairs of points C and D on Γ and Δ. The details of your description may vary depending on the situation of the circles.

Exercises on equations

Each exercise under this heading is concerned with the equation $X' = AX + B$ of a transformation $\varphi : X \to X'$. In exercises 24 to 26 the problem is simply to find the equation—that is, the entries of A and B.

Exercise 24. Two rotations about the point $Z = <1, 2>$ each map the line $2x_1 - x_2 = 0$ onto the line $x_1 - x_2 = -1$. Determine their equations, accurate to four decimal places.

Exercise 25. Consider a point $P = <p_1, p_2> \neq O$ and the line $g \perp \vec{OP}$ through O.

Part 1. Show that σ_g has matrix

$$A = I - \frac{2}{P^tP}K, \text{ where } K = PP^t = \begin{bmatrix} p_1^2 & p_1p_2 \\ p_2p_1 & p_2^2 \end{bmatrix}.$$

Part 2. Consider the line $h \perp \vec{OP}$ through a point Q. Show that $\sigma_h : X \rightarrow X'$ has equation

$$X' = AX + \frac{2}{P^tP}KQ.$$

Part 3. Find an equation for the reflection across the line h with equation $x_1 + 2x_2 = 3$.

Part 4. Find an equation for the glide reflection with axis h and vector $<2, -1>$.

Exercise 26. Consider points $A = <1, 0>$, $B = <2, 0>$, and $C = <0, 2>$, and the feet D and F of the altitudes of ΔABC through A and C. According to theorem 5.8.1, $\Delta ABC \sim \Delta DBF$.

Part 1. Find the equations of the similarity $\varphi : X \rightarrow X'$ that maps ΔABC onto ΔDBF.

Part 2. What's its ratio?

Part 3. Express φ as a composition of some or all of these: reflection, translation, rotation, glide reflection, dilation, and contraction.

Part 4. Find its fixpoint.

Suggestion: The entries of the matrix and vector for φ form the solution of a system of six linear equations in six unknowns. You might want to solve that for part (1), then work on (3). Or you could solve (3) first geometrically, then do (1).

Exercise 27 asks you to classify four transformations. First, show that they're isometries. Then you have the same tools at hand as for earlier classification exercises. Use them! The logic is the same, but now you compute with coordinates.

Exercise 27. Show that each of the following systems of equations represents an isometry φ, and classify it.

(a) $\begin{cases} x_1' = \frac{3}{5}x_1 + \frac{4}{5}x_2 \\ x_2' = -\frac{4}{5}x_1 + \frac{3}{5}x_2 \end{cases}$ (c) $\begin{cases} x_1' = \frac{3}{5}x_1 + \frac{4}{5}x_2 + 1 \\ x_2' = -\frac{4}{5}x_1 + \frac{3}{5}x_2 - 1 \end{cases}$

(b) $\begin{cases} x_1' = \frac{3}{5}x_1 - \frac{4}{5}x_2 \\ x_2' = -\frac{4}{5}x_1 - \frac{3}{5}x_2 \end{cases}$ (d) $\begin{cases} x_1' = \frac{3}{5}x_1 - \frac{4}{5}x_2 + 1 \\ x_2' = -\frac{4}{5}x_1 - \frac{3}{5}x_2 - 1 \end{cases}$

The next three exercises could have been placed under the earlier **Classifying isometries** heading. To solve them you don't really need to work with their equations. But that might be the simplest way.

Exercise 28. Classify the compositions of the 1° rotation about point <2, 3> and the translation with vector <4, 5>, in both orders.

Exercise 29. Classify the compositions of the 1° rotation about point <2, 3> and the 4° rotation about point <5, 6>, in both orders.

Exercise 30. Classify the compositions of the 10° rotation about the origin and the reflection across the line with equation $x_1 + x_2 = 1$, in both orders.

A geometric transformation φ associates with points of a figure those of a corresponding figure—its image—perhaps in another place. You could call this the *alibi* version of the concept—*alibi* is the Latin word for *elsewhere*. There's another version, often used in analytic geometry. Suppose the figure is a coordinate system; then so is its image. A point P with coordinates $<x_1, x_2>$ relative to the first system has coordinates $<x_1', x_2'>$ relative to the second. Associated with φ is a *coordinate transformation* that maps the coordinate pair $<x_1, x_2>$ to $<x_1', x_2'>$. Some texts regard *that* as the basic geometric transformation concept: Points of a figure Φ have two corresponding coordinate pairs. You could call it the *alias* version—*alias* is the Latin word for *otherwise*, to which we often add *known as*. Suppose you want to use two coordinate systems to model different aspects of some phenomenon. It's usually easy to find the isometry φ that transforms one into the other; it's part of the problem framework. Then you can derive an equation for φ. How do you use that to find equations that relate the corresponding coordinate pairs $<x_1, x_2>$ and $<x_1', x_2'>$ of a point P? (These considerations were introduced, with slightly different notation, under the heading **Transforming coordinate systems** in section 6.9.) Figures Φ in your model may be described by equations that relate the two coordinates of points P of Φ. How do you convert an equation that relates x_1 and x_2 to one for x_1' and x_2'? Exercise 31 provides some examples of this process. You'll have to decide when and how to use the equation of φ and that of its inverse. You'll see that how you transform equations that describe figures depends on whether they're single equations or pairs of parametric equations.

Exercise 31. Suppose that, in addition to your standard Cartesian coordinate system with origin O, you're using an auxiliary one with

- origin $O' = <3, 2>$,
- unit point U' for the first axis located below and left of O' on the line through O' with slope 1, and
- unit point V' for the second axis located above and left of O'.

A point $X = <x_1, x_2>$ has coordinates x_1' and x_2' relative to this coordinate system.

Part 1. Find formulas for x_1' and x_2' in terms of x_1 and x_2.

Part 2. Find the equation of the unit circle Γ with center O in terms of the auxiliary coordinate system: X must lie on Γ if and only if x_1' and x_2' satisfy the equation.

Part 3. $x_2 = x_1^2$ is the equation of a parabola Π. Find the equation of Π in terms of the auxiliary coordinate system.

Part 4. $x_1 = \cos t$ and $x_2 = 2 \sin t$ for $0 \le t < 2\pi$ are parametric equations for an ellipse Λ. Find parametric equations for Λ in terms of the auxiliary coordinate system.

Reflection calculus

During the 1920s Gerhard Thomsen discovered how to formulate many familiar properties of plane figures in terms of equations involving compositions of point and line reflections. The first few exercises under this heading introduce this method. You can solve them by manipulating equations involving reflections, half turns, translations, and occasionally vector algebra. You never have to use coordinates explicitly. You can call this method a *reflection calculus.*

Exercise 32. Show that for any points O, P, and Q, the composition $\sigma_O \sigma_P \sigma_Q$ is a half turn about point $R = O - P + Q$, and $OPRQ$ is a (possibly degenerate) parallelogram.

Exercise 33. Describe the figure formed by points A, B, C, and G for which $\sigma_A \sigma_G \sigma_B \sigma_G \sigma_C \sigma_G = \iota$. *Suggestion*: Reason with vectors.

Exercise 34. Construct a pentagon with given edge midpoints M_1 to M_5. Is it unique? Generalize your solution to apply to polygons with more edges.

Gerhard THOMSEN was born in Hamburg in 1899, the son of a physician. He served a year in World War I, then became one of the first students of the new university at Hamburg. Thomsen completed the Ph.D. in 1923 with a dissertation on differential geometry. He served as assistant in Karlsruhe and Hamburg, studied a year with Tullio Levi-Civita in Rome, then presented his Habilitationsschrift in Hamburg, on a problem in gravitational physics. In 1929 he became Ausserordentlicher Professor at Rostock. There he continued work on differential geometry, collaborating with Wilhelm Blaschke, and published as well in mathematical physics and foundations of geometry. In 1933, reacting to the Nazification turmoil in his university, he published an inflammatory lecture that seemed to support some Nazi aims but attacked Nazi suppression of education in the sciences. This evidently attracted the attention of the secret police. Thomsen died, an apparent suicide, on a Rostock railroad track in 1934.

Exercise 35. When can you construct a quadrilateral with given edge midpoints? When is it unique? How about a hexagon? *Suggestion*: What kind of figure is formed by the midpoints of the edges of a quadrilateral? Of a hexagon?

Exercises 34 and 35 are closely related to exercise 17 on billiards and to exercise 18. Can you imagine the pentagon and hexagon as *dual billiards* tables? You play on the outside, grazing the corners instead of rebounding from the sides!

A development of the reflection calculus and results obtained with it were published in Thomsen 1933a. Exercises 36 to 39 are selected from section 4 of that book.

Exercise 36. Consider points P and Q, lines g and h, and the isometry $\varphi = \sigma_g \sigma_P \sigma_h \sigma_Q$. Under what condition on the points and lines is $\varphi = \iota$? *Suggestion*: First, classify $\sigma_g \sigma_P$ and $\sigma_h \sigma_Q$.

Exercise 37. Prove that these conditions on lines g and h are equivalent:

(a) $g \parallel h$;
(b) for *every* point P, $\sigma_g \sigma_h \sigma_P \sigma_g \sigma_h \sigma_P = \iota$ but $\sigma_g \sigma_h \sigma_P \neq \iota$;
(c) for *some* point P, $\sigma_g \sigma_h \sigma_P \sigma_g \sigma_h \sigma_P = \iota$ but $\sigma_g \sigma_h \sigma_P \neq \iota$.

Exercise 38. Consider some lines a, b, and c and the isometry

$$\varphi = \sigma_c\sigma_a\sigma_b\sigma_a\sigma_b\sigma_a\sigma_c\sigma_b\sigma_a\sigma_c\sigma_a\sigma_c\sigma_a\sigma_b.$$

Prove that φ is the identity if and only if a, b, and c are copencilar or form an isosceles triangle with congruent edges on b and c. *Suggestion*: It's easy to show that copencilarity implies $\varphi = \iota$. When there's a triangle, consider the line g through the feet B^* and C^* of the altitudes perpendicular to b and c. When is $g \parallel a$? What isometry is $\gamma = \sigma_b\sigma_a\sigma_c$?

Exercise 39. Show that for any lines a, b, and c,

$$\sigma_a\sigma_b\sigma_c\sigma_a\sigma_b\sigma_c\sigma_b\sigma_c\sigma_a\sigma_b\sigma_c\sigma_a\sigma_c\sigma_b\sigma_a\sigma_c\sigma_b\sigma_c\sigma_b\sigma_a\sigma_c\sigma_b = \iota.$$[29]

Exercises on the definitions of isometry and similarity

The next two exercises are concerned with functions that map the plane to itself but satisfy (seemingly) weaker conditions than isometries or similarities.

Exercise 40. Show that any function $\varphi : X \to X'$ from the plane to itself that preserves distance is an isometry. That is, from the hypothesis that $X'Y' = XY$ for all X and Y, deduce that φ is injective and surjective. *Suggestion*: For surjectivity, given Y, you must find X such that $Y = X'$; select two points P and Q and observe the relationship of Y to P' and Q'.

Exercise 41. Let r be any positive number and φ be a function from the plane to itself such that $X'Y' = r\,(XY)$ for all X and Y. Prove that φ is a similarity with ratio r. *Suggestion*: Use exercise 40.

It's possible to weaken the definitions of isometry and similarity even further than these exercises suggest. For example, invariance of the property of separation by distance 1 —that is, the condition $XY = 1$ if and only if $X'Y' = 1$ —implies that φ is an isometry.[30]

---

[29] Responding to a question from Thomsen, Hellmuth Kneser (1931) showed that this equation is the shortest that's true for *all* lines a, b, and c, except for equations like $\sigma_a\sigma_a = \iota$ and others such as $\sigma_a\sigma_b\sigma_b\sigma_a = \iota$ that are derivable from it.

[30] Modenov and Parkhomenko [1961] 1965, appendix.

Dilatations

A *dilatation*[31] is a plane transformation $\varphi : X \rightsquigarrow X'$ such that the image $g' = \varphi[g]$ of any line g is a line parallel or equal to g. The last four exercises of this chapter develop the theory of dilatations, which plays a major role in some axiomatizations of geometry.[32] You should solve them without using any limit concepts from analysis. Your solutions will provide a proof of the fixpoint theorem for similarities, theorem 6.10.8, that, unlike the proof in the text, requires no analysis. In your solutions of exercises 43 to 45, you're expected to use the results in the preceding exercises.

Exercise 42. Prove that the dilatations form a transformation group. Of the various types of isometries and similarities considered in this chapter, which are dilatations and which not?

Exercise 43. Show that

(a) the only dilatation with two distinct fixpoints is the identity, and
(b) if φ and χ are dilatations and $\varphi(X) = \chi(X)$ for two distinct points X, then $\varphi = \chi$.

Suggestions: For (a) show first that a dilatation $\varphi : X \rightsquigarrow X'$ with distinct fixpoints P and Q fixes every point *not* on $\overleftrightarrow{PQ}$, then consider points *on* the line. Note that any line through a fixpoint is fixed and any point on two fixed lines is fixed.

Exercise 44. Show that every dilatation $\varphi : X \rightsquigarrow X'$ that's not a translation has a fixpoint. *Suggestion*: Find P and Q such that $PP'Q'Q$ is a quadrilateral, then show that it can't be a parallelogram.

Exercise 45. Prove that every dilatation that's not a translation is a contraction, a dilation, or the composition of a contraction or dilation with the half turn about its center.

---

[31] Be careful not to confuse the terms *dilation* and *dilatation*! The first stems from the Latin prefix *dis-*, meaning *apart*, with the past participle *latus* of the irregular Latin verb *fero*, which means *carry*. The second is from the related regular verb *dilato*, which means *broaden*.

[32] For example, Artin 1957, chapter 2.

This page intentionally left blank

Chapter

7

Three-dimensional isometries and similarities

Chapter 6 introduced plane transformational geometry. You studied plane isometries and similarities in depth. New concepts entered gradually, so that each new idea was related as closely as possible to basic geometry, where our visual intuition provides a clear guide. With that background, you've studied the effects of plane motions, and plane Cartesian coordinate transformations. You're prepared to analyze symmetries of plane ornamentation undertaken later, in chapter 8. And you're prepared to study more general types of plane transformations that fall beyond the scope of this book.

But you haven't many techniques yet to handle motions and coordinate transformations in three-dimensional space, where you live. Nor are you ready yet to analyze the three-dimensional symmetries of polyhedra and crystals. For that, this chapter provides a complete coverage of three-dimensional isometries and similarities. You'll see that it's a straightforward extension of the two-dimensional theory. Of course, there are more isometry types in three dimensions, but the algebra techniques you used in two dimensions apply without change in this chapter.

Because you've already met, in simpler form, most of the ideas discussed in this chapter, it's more streamlined. Concepts are introduced and studied in the order that leads most quickly to the main results of the theory. To that end, the chapter relies more heavily on algebra. Sometimes this is "brutally" efficient. For example, section 7.1 derives the matrix equation $X' = AX + B$ for an isometry $X \to X'$ directly from the uniqueness and rigidity theorems, before considering in detail *even one* particular type of isometry.

That plan makes various general techniques available for use earlier. But following it in chapter 6 would have prevented you from applying your geometric experience effectively in learning new concepts.

The greater reliance on algebra is also due to the greater complexity of three dimensions. Many problems to which we now apply algebraic plane transformational geometry methods could be solved without that theory. Algebra reveals its usefulness as it helps reduce the complexity in three dimensions.

7.1 Isometries

Concepts

- Definition
- The isometry group
- Invariance
- Rigidity theorem
- Uniqueness theorem
- Orthogonal matrices
- Equations of isometries
- Conjugacy
- Parity

This section extends to three dimensions the notion of isometry that you studied in chapter 6. It also presents various properties shared by all such isometries: the fixpoint, rigidity, and uniqueness theorems, representation of isometries $X \to X'$ by equations $X' = AX + B$ with orthogonal matrices A, the relationship of these equations to conjugacy, and the distinction between even and odd isometries. You'll study particular types of isometries in later sections.

The equations require a Cartesian coordinate system. Introduce a "standard" system $\mathscr{C}$ now, with origin O and unit vectors U_1 to U_3, and refer to it as necessary.

In the three-dimensional theory, an *isometry* is a distance-preserving transformation of the set of *all* points—not just those in some plane. As in two dimensions, you can show that the identity transformation ι is an isometry, the inverse φ^{-1} of an isometry φ is an isometry, and the composition $\varphi\chi$ of two isometries φ and χ is an isometry:

Theorem 1. The isometries form a subgroup of the symmetric group of the set of all points.

There's a whole catalog of properties of three-dimensional figures that are invariant under isometries: the invariant properties of plane figures that you've already considered—especially betweenness—plus

measure of a dihedral angle
volume of a tetrahedron
perpendicularity
parallelism, etc.

being a
- plane
- half space
- dihedral angle
- tetrahedron
- sphere

This text won't pursue those details. As necessary, you can supply invariance proofs like that of theorem 6.2.5, because you can define all these concepts in terms of distance (using betweenness if appropriate).

Rigidity and uniqueness theorems

The uniqueness theorem is a major tool in the plane theory. To see that an isometry under analysis is really a familiar one, you need only show that they agree at three noncollinear points. An analogous tool performs the same function in three dimensions. The plane version was derived from the *rigidity theorems* 6.2.6 and 6.2.7; its three-dimensional form requires analogs of those results. The first one is already valid in three dimensions:

Theorem 2. If an isometry $X \to X'$ has distinct fixpoints P and Q, then it fixes every point X on $\overleftrightarrow{PQ}$.

You can check that the original proof is valid, too. For the next result, you must edit the statement of theorem 6.2.7; its proof remains valid.

Theorem 3. If an isometry $X \to X'$ has noncollinear fixpoints P, Q, and R, then it fixes every point X on the plane PQR.

Three-dimensional geometry enters with a new rigidity result. Its proof, exercise 7.9.1, is merely a three-dimensional version of the proof of theorem 6.2.7.

Theorem 4. The only isometry with four noncoplanar fixpoints is the identity.

With this foundation, you can easily state the three-dimensional *uniqueness theorem*. Its proof is the same as that of the two-dimensional version, corollary 6.2.8.

Corollary 5 (Uniqueness theorem). If χ and ψ are isometries and $\chi(P) = \psi(P)$ for four noncoplanar points P, then $\chi = \psi$.

Orthogonal matrices

The entire analytic treatment of isometries with matrix algebra works in three dimensions just as it did in two. Appendix C presents an outline of the elementary features of vector and matrix algebra, independent of dimension. The notion of orthogonality, introduced in section 6.8, applies to square matrices A of *any* dimension: A is *orthogonal* if $A^tA = I$, i.e., its transpose is its inverse. As noted there, the identity matrix I is orthogonal, the inverse of an orthogonal matrix is orthogonal, and the product of two orthogonal matrices is orthogonal. That section also showed that for *any* orthogonal matrix A and vector B, the mapping $X \rightarrow AX + B$ is an isometry.

These analogies suggest that *every* three-dimensional isometry φ should have an equation $\varphi(X) = AX + B$ for some orthogonal matrix A and some vector B. The argument for the corresponding plane result, theorem 6.8.2, required the equations for two special types of isometries, including some rotations. You *could* take that approach in three dimensions, deriving analogs of those special equations. But that's difficult because equations of three-dimensional rotations are generally more complicated. The argument in the next proof, however, doesn't depend on individual equations nor on the dimension; so the desired result holds in three dimensions as well. The first step is to show that you can map your preselected coordinate system O, U_1, U_2, U_3 into *any* other one by applying a transformation of the form $X \rightarrow AX + B$ with an orthogonal matrix A.

Theorem 6. Given points B, V_1, V_2, V_3 such that $BV_j = 1$ for $j = 1$ to 3 and $\overrightarrow{BV}_1$ to $\overrightarrow{BV}_3$ are perpendicular, there's an orthogonal matrix A such that $AU_j + B = V$ for $j = 1$ to 3.

Proof. Construct the matrix A whose jth column is $A_j = V_j - B$. Then $AU_j + B = A_j + B = V_j$, as desired. Moreover, $A_j \cdot A_j = (BV_j)^2 = 1$. Now suppose $j \neq k$. Since $\overrightarrow{BV}_j \perp \overrightarrow{BV}_k$, Pythagoras' theorem implies

$$(OA_j)^2 + (OA_k)^2 = (BV_j)^2 + (BV_k)^2 = (V_jV_k)^2.$$

But $V_j - V_k = A_j - A_k$, so

$$(V_jV_k)^2 = (V_j - V_k)\cdot(V_j - V_k) = (A_j - A_k)\cdot(A_j - A_k) = (A_jA_k)^2.$$

The converse part of Pythagoras' theorem now yields $\overrightarrow{OA}_j \perp \overrightarrow{OA}_k$, and $A_j \cdot A_k = 0$ by theorem 3.11.10. It follows that A is orthogonal: $A^tA = I$. To see that, note that the j,kth entry of A^tA is the jth row of A^t times the kth column of A —that is, $A_j \cdot A_k$. According to the equations just derived, that's 1 if $j = k$ or 0 if $j \neq k$. It's the j,kth entry of I.[1] ♦

---

[1] The word *orthogonal*, a synonym for *perpendicular*, stems from the Greek words *orthos* and *gonia* for *upright* and *angle*. Matrices like A were termed orthogonal because their column vectors are perpendicular, as noted in this proof.

Theorem 7. The three-dimensional isometries are the mappings $X \to \varphi(X) = AX + B$, where A is an orthogonal matrix and B is an arbitrary vector. The isometry φ determines A and B uniquely.

Proof. You've already seen that this equation always defines an isometry. It remains to show that an arbitrary isometry φ has such an equation, and only one. First, φ maps your standard coordinate system O, U_1, U_2, U_3 into another system B, V_1, V_2, V_3. By theorem 6, there's an orthogonal matrix A such that $AU_j + B = V_j$ for $j = 1$ to 3. The isometry $X \to AX + B$ agrees with φ at four noncoplanar points O and U_1 to U_3. By the uniqueness theorem, the two isometries coincide: $\varphi(X) = AX + B$ for every X, as required. Finally, you can use the last paragraph of the proof of theorem 6.8.1 to show that φ determines A and B uniquely. ♦

The short algebraic argument in the proof of theorem 6 could have been used in chapter 6 to avoid some matrix calculations. But it has so little visual content that the longer argument was used instead. It's difficult to determine just how this material was originally discovered. In essence, it's already present in Euler [1775] 1968.

Conjugacy

Chapter 6 used the notion of conjugacy to classify plane isometries. It plays the same role in three dimensions. Its definition in section 6.9 applies to *any* transformations μ and φ; the composition $\varphi\mu\varphi^{-1}$ is the conjugate of μ by φ. We write $\lambda \sim \mu$ to indicate that a transformation λ is a conjugate of μ. The basic theorems 6.9.4 and 6.9.5 hold for three-dimensional isometries, too. In particular, conjugacy is an equivalence relation, and the equivalence class of an isometry is called its conjugacy class. Sections 7.2 to 7.5 use conjugacy to classify isometries.

The section 6.9 discussion of coordinate transformations, equations of isometries, and their conjugacy is essentially valid for three dimensions, too. An isometry φ transforms your standard coordinate system $\mathscr{C}$ into another system $\mathscr{D}$ defined by $P = \varphi(O)$ and $V_j = \varphi(U_j)$ for $j = 1$ to 3 as in figure 7.1.1. Theorem 9, the analog of theorem 6.9.6, shows how a point's $\mathscr{C}$ and $\mathscr{D}$ coordinates are related.

Theorem 9. If an isometry φ transforms coordinate system $\mathscr{C}$ into system $\mathscr{D}$, then the $\mathscr{C}$ coordinates of each point X are the same as the $\mathscr{D}$ coordinates of $Y = \varphi(X)$.

Proof. Consider figure 7.1.1. The figure formed by O, U_1 to U_3, and X is congruent to that formed by P, V_1 to V_3, and Y. ♦

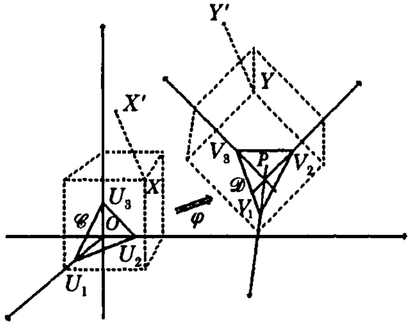

Figure 7.1.1
Transforming
a coordinate system

Theorem 10, the main result about conjugacy, is the analog of theorem 6.9.7; you can use its proof unchanged, if you cite theorem 6 of this section in place of theorem 6.6.3.

***Theorem 10* (*Conjugacy theorem*).** Two isometries are conjugate if and only if there are coordinate systems under which they have the same matrix equation.

Parity

As in two dimensions, the determinant par φ of the orthogonal matrix A corresponding to an isometry φ is called its *parity*. Since equation $A^tA = I$ implies $(\det A)^2 = (\det A^t)(\det A) = \det(A^tA) = \det I = 1$, the parity is ± 1. Isometries with parity 1 are called *even*. For example, the identity is an even isometry. The others are called *odd*. This section concludes with some simple facts about parity that you've already met in the plane theory. The connection between parity and orientation is discussed in the next section.

Theorem 11. par $\varphi\chi =$ par φ par χ and par $\varphi^{-1} =$ par φ for any isometries φ and χ.

Corollary 12. The even isometries constitute a subgroup of the isometry group.

Corollary 13. Conjugate isometries have the same parity.

Leonhard EULER was born in Basel in 1707. He was taught first by his father, a protestant minister, who had attended Jakob Bernouilli's lectures at the university there. In 1720 Euler entered the university. He studied with Johann Bernouilli, and received a degree in philosophy in 1723 with a comparison of Newton's and Descartes' ideas. He started theological studies, but dropped those for mathematics. Although he published some papers on mechanics, Euler could find no job at Basel. In 1727, however, Peter the Great's reign was ending, the Academy of Science at St. Petersburg was being formed, and the intelligentsia there was becoming dominated by Germans. On Bernouilli's recommendation, they offered Euler a job; he became Professor there in 1731. The Academy had scientific, educational, and technological assignments, and provided a very rich atmosphere. Euler produced about 90 papers on mathematics there during the period 1727–1741, and a treatise on mechanics. He worked on cartography, shipbuilding, and navigation as well.

Euler's brother also had an Academy position, as an artist. In 1733 Euler married the daughter of another Swiss artist at the Academy. Two sons were born in St. Petersburg. Later, in Berlin, they had another son and two daughters; eight other children died in infancy. In 1738, Euler contracted a disease that left him blind in one eye.

Political turmoil in Russia, particularly resentment against German influence, arose during the late 1730s, simultaneous with Frederick the Great's reorganization of the Academy of Science at Berlin. After prolonged negotiations, Euler moved there in 1741. However, he retained many connections with the St. Petersburg Academy, editing its mathematical publications and publishing there his own texts on calculus and navigation. Euler stayed in Berlin for 25 years. During that period he produced about 380 papers in pure and applied mathematics, and was elected to the Academies in London and Paris as well. Euler was very influential in the development and administration of the Berlin Academy, actually running it for many years. But he and the King were incompatible; Frederick wouldn't appoint him President.

In 1766, Euler returned to St. Petersburg. The long, stable reign of Catherine the Great had begun. One of Euler's sons became professor of physics there; another, an artillery officer. An illness in 1769–1771 led to Euler's total blindness. But his scientific output actually increased! Besides many research papers, he published texts on algebra, calculus, navigation, and life insurance, and treatises on optics and lunar theory. His first wife died in 1773; in 1776 he married her half-sister. Euler died of a brain hemorrhage in the midst of heavy work in 1783.

7.2 Reflections

Concepts

Reflections across planes
Conjugate reflections
Structure theorem
Even and odd isometries
Orientation

Reflections play a fundamental role in the theory of plane isometries. They provide components from which you can construct all other isometries, and they facilitate the distinction between even and odd isometries. Their three-dimensional counterparts are the *reflections across planes.* The reflection $X' = \sigma_\varepsilon(X)$ of a point X across a plane ε is X itself if X is on ε, else it's the point on the opposite side of ε such that ε is the perpendicular bisector of $\overline{XX'}$. The plane ε is called the *mirror* of σ_ε. To study these reflections, it's sometimes convenient to use the theory of plane isometries. Theorem 1 follows directly from the plane theory and the definition of σ_ε. (See figure 7.2.1.) You can supply the proof.

Theorem 1. σ_ε is self-inverse. Its fixpoints are the points in ε. It fixes any plane $\alpha \perp \varepsilon$, and its restriction to α is the (two-dimensional) reflection of the points on α across the intersection $g = \alpha \cap \varepsilon$.

Corollary 2. σ_ε is an isometry.

Proof. (See figure 7.2.1.) Any two points X and Y lie in some plane $\alpha \perp \varepsilon$, and σ_ε operates on X and Y just as the (two-dimensional) reflection of α across $g = \alpha \cap \varepsilon$. That preserves the distance XY. ♦

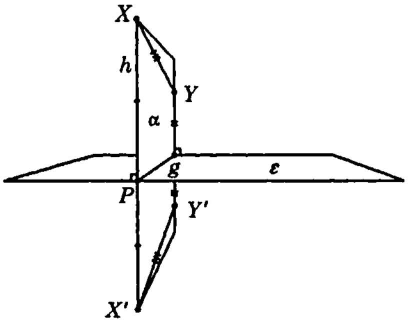

Figure 7.2.1
Reflection
$\sigma_\varepsilon: X \rightarrow X'$ across ε

Theorem 3. An isometry φ that leaves a plane ε pointwise fixed is the identity or σ_ε.

Proof. Select a point X not on ε, construct the line $h \perp \varepsilon$ through X, and let $P = h \cap \varepsilon$. (See figure 7.2.1.) Since φ fixes P and ε, it maps h to a line through P perpendicular to ε. That is, φ fixes h, so $\varphi(X)$ is one of the two points X and $\sigma_\varepsilon(X)$ on h whose distance from P is PX. Thus φ agrees either with ι or with σ_ε at four noncoplanar points: X and any three noncollinear points on ε. The uniqueness theorem yields $\varphi = \iota$ or σ_ε. ♦

Corollary 4. $\varphi\sigma_\varepsilon\varphi^{-1} = \sigma_{\varphi[\varepsilon]}$ for any isometry φ and any plane ε.

Proof. Use the argument under the section 6.9 heading **Conjugacy classes** to show that $\varphi\sigma_\varepsilon\varphi^{-1}$ leaves plane $\varphi[\varepsilon]$ pointwise fixed, but isn't the identity. ♦

The equation of the reflection $\sigma_\varepsilon : X \to X'$ across a plane ε is generally unwieldy—see exercise 7.9.20. But for some special planes it's very simple. For example, if ε is the plane through the first two coordinate axes, then σ_ε has equations

$$\begin{aligned} x_1' &= x_1 \\ x_2' &= x_2 \\ x_3' &= -x_3 \end{aligned} \qquad X' = AX \qquad A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$

This matrix A is orthogonal, with determinant -1.

Any plane ε contains the first two axes of some coordinate system, so any two reflections have the same equations with respect to suitably chosen coordinate systems. With the conjugacy theorem, corollary 4, and corollary 7.1.13, this implies

Theorem 5. The reflections constitute a single conjugacy class. All reflections are odd.

Structure theorem

According to the structure theorem in two-dimensional transformational geometry, every plane isometry is the identity, a reflection, or the composition of two or three reflections. That result is the basis for classifying plane isometries. An analogous one holds in the three-dimensional theory. You'll notice that its proof is organized just like that of theorem 6.6.4, the two-dimensional result.

Theorem 6 **(*Structure theorem*).** The three-dimensional isometry group consists of the identity, the reflections, and the compositions of two, three, and four reflections.

Proof. An isometry φ maps a tetrahedron Γ into a congruent tetrahedron Δ. If $\Gamma = \Delta$, then φ is the identity by the rigidity theorem. If Γ and Δ share three vertices but not four, then φ is the reflection across the plane δ through those vertices, by theorem 3.

If Γ and Δ share *two* vertices, then there's a plane γ through those vertices such that Γ and the tetrahedron $\sigma_\gamma[\Delta]$ share *three* vertices. (Make γ "bisect" a certain dihedral angle.) By the previous paragraph, $\sigma_\gamma \varphi$ is the identity or the reflection across a plane δ, so $\varphi = \sigma_\gamma$ or $\sigma_\gamma \sigma_\delta$.

If Γ and Δ share *one* vertex, then there's a plane β through it such that Γ and the tetrahedron $\sigma_\beta[\Delta]$ share *two*. By the previous paragraph, $\sigma_\beta \varphi = \sigma_\gamma$ or $\sigma_\gamma \sigma_\delta$ for some planes γ and δ, so $\varphi = \sigma_\beta \sigma_\gamma$ or $\sigma_\beta \sigma_\gamma \sigma_\delta$.

Even if Γ and Δ share *no* vertices, there's a plane α —the perpendicular bisector of any segment joining two corresponding vertices—such that Γ and the tetrahedron $\sigma_\alpha[\Delta]$ have at least *one* vertex in common. By the previous paragraph, $\sigma_\alpha \varphi = \sigma_\beta \sigma_\gamma$ or $\sigma_\beta \sigma_\gamma \sigma_\delta$ for planes β, γ, and δ, so φ is the composition of three or four reflections. ♦

Corollary *7.* An isometry is even or odd depending on whether it's a composition of an even or odd number of reflections.

Proof. Every isometry is a composition of reflections, which are odd. ♦

Orientation

You can distinguish even from odd isometries in three dimensions, as in two, by their effects on oriented figures. Informally, you can consider left- and right-hand gestures. Formally, consider tetrahedra—section 3.10 defined them as *ordered* quadruples of noncoplanar points. In figure 7.2.2, the vertices of tetrahedron *PQRS* are arranged as you might point with your right thumb, with P on the thumb, and ΔQRS oriented as your fingers are curled. The reflection across the plane maps them to points P' to S', arranged like the corresponding left-hand gesture. There are twenty-four tetrahedra with vertices P to S listed in various orders. Using the figure, you can classify them according to your intuitive notion of orientation:

Class $\mathscr{R}$: *PQRS*, *PRSQ*, and ten others;
Class $\mathscr{L}$: *PQSR*, *PRQS*, and ten others.

How can you distinguish these classes mathematically, without relying on the figure and your intuition? The determinant formula for tetrahedral

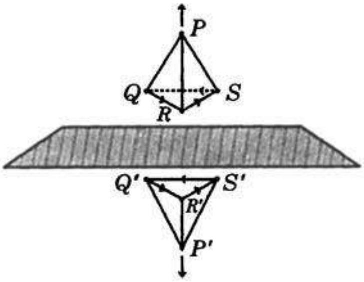

Figure 7.2.2
Reflection across a plane reverses orientation.

volume, theorem 5.6.10, provides the tool. These tetrahedra all have the same volume: ±1/6 times each of the determinants

$$\mathscr{R}\colon\ \det[P-R, Q-R, S-R],\ \det[P-S, R-S, Q-S], \ldots$$
$$\mathscr{L}\colon\ \det[P-S, Q-S, R-S],\ \det[P-Q, R-Q, S-Q], \ldots.$$

You can show that the determinants in the first class are all equal to the same value; and all those in the second, to its negative. Thus all those in one class have the same *positive* determinant; the others have the same *negative* determinant. Accordingly, you can call the tetrahedra *positively* or *negatively* oriented. (*Which* class of tetrahedra is positive depends on the standard coordinate system that you chose at the onset of this chapter. If you reverse the scale on any one axis, you'll interchange the positive and negative classes.) The following computation shows that an even isometry $X \to X' = AX + B$ preserves a tetrahedron's orientation, whereas an odd one reverses it:

$$\begin{aligned}
&\det[P'-R', Q'-R', S'-R'] \\
&\quad = \det[(AP+B)-(AR+B), (AQ+B)-(AR+B), \\
&\qquad\qquad (AS+B)-(AR+B)] \\
&\quad = \det[AP-AR,\ AQ-AR,\ AS-AR] \\
&\quad = \det[A(P-R), A(Q-R), A(S-R)] \\
&\quad = \det(A[P-R, Q-R, S-R]) \\
&\quad = (\det A)(\det[P-R, Q-R, S-R]) \\
&\quad = \pm\det[P-R, Q-R, S-R].
\end{aligned}$$

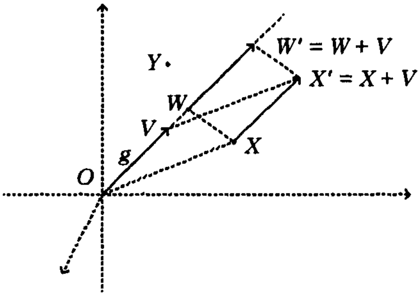

Figure 7.3.1
Translation
$\tau: X \to X' = X + V$

7.3 Translations and rotations

Concepts

- Translations
- Conjugate translations
- Rotations
- Angle parameters for rotations
- Conjugate rotations
- Commuting reflections
- Half turns

The previous section introduced three-dimensional reflections, and showed that each three-dimensional isometry is the identity, a reflection, or the composition of two, three, or four reflections. Section 7.2 studies the compositions of *two* reflections in detail. They're the translations and the rotations.

Translations

Translations are handled in three dimensions exactly as in the plane. The translation corresponding to a vector V is the transformation $\tau_V : X \rightarrow X' = X + V$ displayed in figure 7.3.1 on the preceding page. (This is just the corresponding two-dimensional figure 6.2.1 with a third axis inserted.) The following theorem summarizes some main facts about translations. You can prove it just as you proved the analogous planar results in sections 6.2 and 6.4.

Theorem 1. Composition of translations corresponds to vector addition: For any vectors V and W, $\tau_V \tau_W = \tau_{V+W}$ and $\tau_V^{-1} = \tau_{-V}$. Moreover, $\tau_O = \iota$. Therefore, the translations form a subgroup of the isometry group.

Just as in two dimensions, you can construct translations from reflections.

Theorem 2. The composition $\sigma_\delta \sigma_\varepsilon$ of reflections across parallel planes δ and ε is the translation τ with vector $2(P - Q)$, where P and Q are the intersections of δ and ε with any line k perpendicular to δ and ε.

Proof. Verify that $\sigma_\delta \sigma_\varepsilon(X) = \tau(X)$ for four noncoplanar points X, as in figure 7.3.2, then apply the uniqueness theorem. ♦

Figure 7.3.2 $\sigma_\delta\sigma_\varepsilon$ is a translation when $\delta \parallel \varepsilon$.

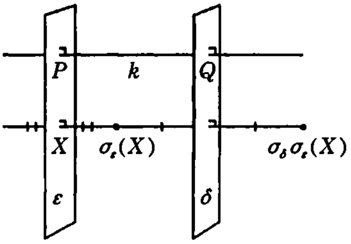

The conjugacy properties of three-dimensional translations are also analogous to the results for plane transformations. Theorem 3 is in fact *identical* to theorem 6.9.9.

Theorem 3. The conjugacy class of a translation τ whose vector has length t consists of all translations with vectors of length t. If τ and an isometry φ have equations $\tau: X \to X' = X + T$ and $\varphi: X \to X' = AX + B$ with respect to a single coordinate system, then $\varphi\tau\varphi^{-1}$ is the translation with vector AT.

Proof. The first sentence follows directly from the conjugacy theorem. The proof of the second is the same as that of theorem 6.9.9. ♦

Rotations

Pursuing the analogy with the theory of plane isometries, you'll next ask what's the composition $\sigma_\delta\sigma_\varepsilon$ of the reflections across two distinct *intersecting* planes δ and ε. It's called a *rotation* about its *axis* $g = \delta \cap \varepsilon$. The following theorem justifies this terminology. To prove it, apply theorem 7.2.1.

Theorem 4. Consider distinct planes δ and ε through a line g perpendicular to a plane α. (See figure 7.3.3.) Choose an angular scale on α with center $P = \alpha \cap g$. Choose half planes δ' and ε', each with edge g. If rays $\alpha \cap \delta'$ and $\alpha \cap \varepsilon'$ have angle parameters ζ and η, then the restriction of $\rho = \sigma_\delta\sigma_\varepsilon$ to α is the (two-dimensional) rotation of α about P with angle parameter $\theta = 2(\eta - \zeta)$. The fixpoints of ρ constitute its axis g.

Figure 7.3.3
$\sigma_\delta\sigma_\varepsilon$ is a rotation when δ intersects ε.

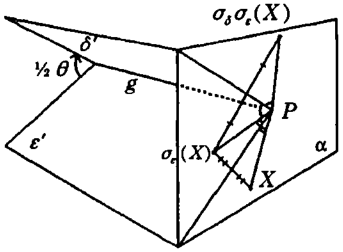

In theorem 4 you can choose δ' and ε' in four ways. Another choice might add 180° to ζ and/or η. Since angle parameters are determined only mod 360°, your choice has no effect on θ. Since θ depends only on the *difference* of two parameters, it depends only on ρ and the *orientation* of the angular scale. You can specify orientation by designating a ray on g as positive. (Informally, align it with your thumb when the knuckles and fingertips of your right fist fall along rays with zero and positive angle parameters.)

For theoretical considerations it's best to describe a rotation as a specific composition of plane reflections. For practical computations, however, it's often helpful to specify a positive axial ray and an angle parameter. For example, a rotation with angle parameter θ about one of the three coordinate axes fixes that coordinate of each point X, and operates on its other coordinates like the corresponding two-dimensional rotation about the origin. For the first coordinate axis, the rotation has equations

$$\begin{aligned} x_1' &= x_1 \\ x_2' &= x_2 \cos\theta - x_3 \sin\theta \\ x_3' &= x_2 \sin\theta + x_3 \cos\theta \end{aligned} \qquad X' = AX \qquad A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}.$$

The analogous rotations about the second and third axes have matrices

$$\begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix} \qquad \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

For other axes, the matrices have no such obvious form. In practice, if you must compute with the matrix of a rotation, you generally choose one of the coordinate axes as the rotational axis.

It's harder to phrase results about conjugacy and rotations in three dimensions than it was in the plane theory, because the angles are harder to describe. It seems easiest to use the next theorem to refer angle questions back to the planes whose reflections combine to form a rotation.

Theorem 5. Suppose line g is the intersection of planes δ and ε. Consider the rotation $\rho = \sigma_\delta \sigma_\varepsilon$ and an isometry φ. Then $\varphi\rho\varphi^{-1}$ is the rotation $\sigma_{\varphi[\delta]}\sigma_{\varphi[\varepsilon]}$ with axis $\varphi[g]$. A rotation is conjugate with ρ just in case it's the composition of the reflections across planes ζ and η that form a dihedral angle whose measure is the same as that of one of those formed by δ and ε.

Proof. $\varphi\rho\varphi^{-1} = \varphi\sigma_\delta\sigma_\varepsilon\varphi^{-1} = \varphi\sigma_\delta\varphi^{-1}\varphi\sigma_\varepsilon\varphi^{-1} = \sigma_{\varphi[\delta]}\sigma_{\varphi[\varepsilon]}$. You can check that the stated condition on ζ and η is equivalent to the existence of an isometry φ such that $\varphi[\delta] = \zeta$ and $\varphi[\varepsilon] = \eta$. ♦

When do two reflections commute? What's their composition in that case? To answer those questions, you can prove the following corollary of theorems 2 and 4. It's similar to the solution of the corresponding two-dimensional problem.

Theorem 6. Reflections σ_δ and σ_ε commute just when planes δ and ε are equal or perpendicular. In the former case, $\sigma_\delta \sigma_\varepsilon$ is the identity. In the latter, it's self-inverse but not the identity. Its restriction to a plane perpendicular to its axis is a (two-dimensional) half turn.

A self-inverse three-dimensional rotation that's not the identity is called the *half turn* about its axis g, and it's denoted by σ_g.[2] Several exercises in section 7.7 examine half turns in detail. Exercise 21 derives the equation of σ_g in terms of the equation of g.

7.4 Glide and rotary reflections

Concepts
- Commuting translations and reflections
- Glide reflections
- Commuting rotations and reflections
- Rotary reflections
- Reflections across points

Previous sections have analyzed reflections, and compositions of two reflections—that is, translations and rotations. What kind of isometry is a composition of three reflections? It's either the composition of a translation with a reflection, or of a rotation with a reflection. The next two headings describe these possibilities in turn. Each one asks first, when is the order of these components significant?

Glide reflections

Theorem 1. A translation τ_V commutes with the reflection σ_ε across a plane ε just when $V = O$ or $\overrightarrow{OV} \parallel \varepsilon$.

Proof. You can easily verify that $\sigma_\varepsilon \tau_V$ and $\tau_V \sigma_\varepsilon$ coincide under either condition. To show that they differ when neither condition holds, consider what they do to a point on ε. ♦

---

[2] Some authors call σ_g the *reflection across* g.

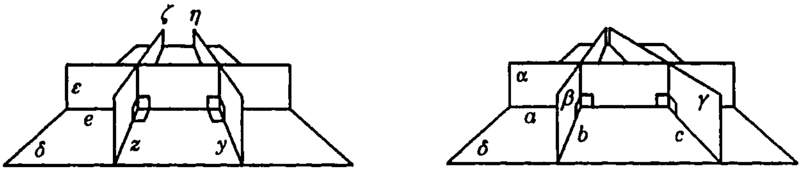

Figure 7.4.1 Proving theorem 2

When $V = O$ or $\overrightarrow{OV} \parallel \varepsilon$, the composition $\varphi = \sigma_\varepsilon \tau_V$ is called a *glide reflection*. The translational component is uniquely determined by φ because $\varphi\varphi = \tau_{2V}$. By decomposing the translational component into reflections across planes ζ and η, you can see that $\varphi = \sigma_\varepsilon \sigma_\zeta \sigma_\eta$ is a glide reflection when $\varepsilon \perp \zeta, \eta$ and $\zeta \parallel \eta$. (See the left-hand diagram of figure 7.4.1—ignore plane δ.) The reflectional component of φ is unique, too, because its mirror is the set of midpoints $\frac{1}{2}(X + \varphi(X))$ for all X. A glide reflection has no fixpoint unless it's a reflection.

Theorem 2. If planes α, β, and γ have a common perpendicular plane δ, then $\sigma_\alpha \sigma_\beta \sigma_\gamma$ is a glide reflection.

Proof. Let $a = \alpha \cap \delta$, $b = \beta \cap \delta$, and $c = \gamma \cap \delta$, as in the right-hand diagram of figure 7.4.1. By theorem 7.2.1, each of σ_α, σ_β, and σ_γ fixes δ, and they operate on a point X in δ in the same way as the two-dimensional reflections of δ across a, b, and c, respectively. Thus, $\varphi = \sigma_\alpha \sigma_\beta \sigma_\gamma$ affects X in the same way as the composition of three two-dimensional reflections. By the theory of plane isometries, that's a two-dimensional glide reflection. Therefore, there exist lines e, z, and y in δ such that $e \perp z$, $e \perp y$, and φ operates on points in δ in the same way as the composition of its two-dimensional reflections across e, z, and y. Construct planes ε, ζ, and η perpendicular to δ through e, z, and y, as in the left-hand diagram of figure 7.4.1. Each of σ_ε, σ_ζ, and σ_η fixes δ, and they affect points in δ in the same way as the two-dimensional reflections of δ across e, z, and y, respectively. That is, $\varphi(X) = \sigma_\varepsilon \sigma_\zeta \sigma_\eta(X)$ for all X in δ. Thus, the even isometry $\sigma_\eta \sigma_\zeta \sigma_\varepsilon \varphi$ fixes each point in δ. By theorem 7.2.3, it's the identity, so $\varphi = \sigma_\varepsilon \sigma_\zeta \sigma_\eta$. This is a glide reflection, since $\sigma_\zeta \sigma_\eta$ is a translation τ_V with $V = O$ or $\overrightarrow{OV} \parallel \varepsilon$. ♦

Rotary reflections and inversions

Next, consider compositions of *rotations* and reflections. When is the order of the components significant? The following theorem says when it's *not*. Its proof, similar to that of theorem 1, is exercise 7.9.2.

Figure 7.4.2
Proving theorem 4

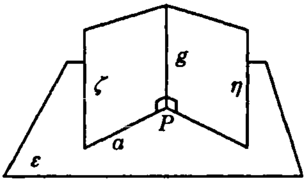

Theorem 3. A rotation ρ commutes with a reflection σ_ε just when $\rho = \iota$, or its axis g is perpendicular to ε, or it's a half turn about a line in ε.

When $\rho = \iota$ or ρ is a rotation with axis perpendicular to ε, the composition $\varphi = \sigma_\varepsilon \rho$ is called a *rotary reflection*. By decomposing the rotational component into reflections across planes ζ and η, you can see that $\varphi = \sigma_\varepsilon \sigma_\zeta \sigma_\eta$ is a rotary reflection when $\varepsilon \perp \zeta, \eta$ and ζ intersects η, as in figure 7.4.2. Clearly, φ has either a plane ε of fixpoints, or exactly one fixpoint P, called the *center* of φ. In the first case, $\varphi = \sigma_\varepsilon$. In the second, if φ isn't self-inverse, it has exactly one fixed plane ε —its *mirror*—which consists of all points of the form $\frac{1}{2}(X + \varphi(X))$. In that case, φ also determines its rotational component uniquely, since $\rho = \sigma_\varepsilon \varphi$.

However, a self-inverse rotary reflection has no unique mirror or rotational component. This situation occurs when $\zeta \perp \eta$ in figure 7.4.2. Then, φ is a composition of reflections across three mutually perpendicular planes. It's called the *reflection* σ_P *across center* $P = g \cap \varepsilon$.[3] The next result justifies that terminology—the point reflection depends only on P, not on g or ε. You can supply the proof. See exercises 7.7.7, 7.7.8, and 7.7.10 for a more detailed study of point reflections.

Theorem 4. The reflection σ_P across a point P is self-inverse, and $P = \frac{1}{2}(X + \varphi(X))$ for every point X, so $\varphi(X) = 2P - X$. If ε is any plane through P and g is the line through P perpendicular to ε, then $\sigma_P = \sigma_\varepsilon \sigma_g$.

The next result shows that a composition of three reflections is a rotary reflection, unless theorem 2 has classified it already as a glide reflection.

Theorem 5. If planes α, β, and γ have no common perpendicular plane, then $\sigma_\alpha \sigma_\beta \sigma_\gamma$ is a rotary reflection.

Proof. You can show that $h = \beta \cap \gamma$ is a line. Erect plane $\delta \perp \alpha$ through h, then find plane η through h such that $\sigma_\beta \sigma_\gamma = \sigma_\delta \sigma_\eta$ (by making two dihedral angles have equal measure). Erect plane $\varepsilon \perp \eta$ through the line $a = \alpha \cap \delta$, then find plane $\zeta \perp \varepsilon$ through a. (See figure 7.4.2.) It follows that $\sigma_\alpha \sigma_\beta \sigma_\gamma = \sigma_\alpha \sigma_\delta \sigma_\eta = \sigma_a \sigma_\eta = \sigma_\varepsilon \sigma_\zeta \sigma_\eta$. This isometry is a rotary reflection because $\sigma_\zeta \sigma_\eta$ is a rotation with axis $g = \zeta \cap \eta$ perpendicular to ε. ♦

---

[3] Some authors call σ_P the *inversion* with center P, but that usage conflicts with the notion of inversion with respect to a circle or sphere.

7.5 Classifying isometries

Concepts

Commuting translations and rotations
Screws
Classification

According to the structure theorem, each isometry is the identity, a reflection across a plane, or a composition of reflections across two, three, or four planes. In section 7.2 you studied reflections. Section 7.3 was devoted to compositions of two reflections: the translations and rotations. And section 7.4 described compositions of three: the glide and rotary reflections. This section covers the last case, compositions of four reflections. Then it summarizes the entire classification scheme.

Screws

A composition of four reflections could be a translation or a rotation. But there's one more type, whose description is facilitated by the next theorem. Its proof, which resembles those of theorems 7.4.2 and 7.4.4, is exercise 7.9.3.

Theorem 1. A translation τ_V commutes with a rotation ρ just when $V = O$, $\rho = \iota$, or $\vec{OV}$ is parallel to the axis g of ρ.

When τ_V commutes with ρ, the composition $\varphi = \rho\tau_V$ is called a *screw*. A screw determines its rotational and translational components ρ and τ_V uniquely. If it's neither a rotation nor a translation, then φ has no fixpoints but exactly one fixed line g, $V = \varphi(P) - P$ for any point P on g, and $\rho = \varphi\tau_V^{-1}$.

The next result seems like a special case, but Hermann Wiener showed in 1890[4] that it provides the key to identifying any composition of four reflections.

---

[4] Wiener 1890–1893, part II.

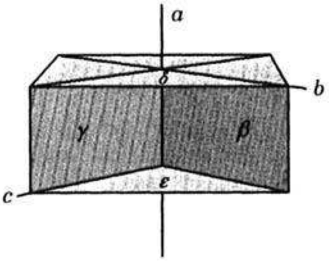

Figure 7.5.1
Proving lemma 2

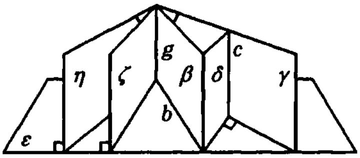

Figure 7.5.2
Proving theorem 3

Lemma 2. For any lines b and c, the composition $\varphi = \sigma_b \sigma_c$ is a screw.

Proof. This argument is designed to apply when b and c are skew (see figure 7.5.1). You can verify that it also works when they're coplanar. (You'll need additional figures for the parallel and intersecting cases.) First, find a line a perpendicular to b and c, plane β through a and b, and plane γ through a and c. Erect planes $\delta \perp \beta$ through b and $\varepsilon \perp \gamma$ through c. Then $\delta \perp \gamma$ also, so $\sigma_b \sigma_c = (\sigma_\beta \sigma_\delta)(\sigma_\gamma \sigma_\varepsilon) = (\sigma_\beta \sigma_\gamma)(\sigma_\delta \sigma_\varepsilon)$. This is a screw because $\sigma_\beta \sigma_\gamma$ is a rotation with axis a and $\sigma_\delta \sigma_\varepsilon$ is a translation parallel to a. ♦

Theorem 3. Every even isometry φ is a screw.

Proof. Suppose φ is neither a translation nor a rotation. Then there exist distinct points P and Q such that $\varphi(P) = Q$. Let δ be the perpendicular bisector of segment $\overline{PQ}$, so that $\varphi\sigma_\delta(Q) = Q$. Since $\varphi\sigma_\delta$ is odd, it's a composition of three reflections. Since it has a fixpoint, it's a rotary reflection, and $\varphi\sigma_\delta = \sigma_\varepsilon \sigma_\zeta \sigma_\eta$ for some planes ε, ζ, and η such that $\varepsilon \perp \zeta, \eta$. The intersection $g = \zeta \cap \eta$ is a line; otherwise, $\varphi\sigma_\delta = \sigma_\varepsilon$, and φ would equal $\sigma_\varepsilon \sigma_\delta$, a rotation or translation. Erect the plane $\gamma \perp \delta$ through g, as shown in figure 7.5.2. Find the plane β through g such that $\sigma_\zeta \sigma_\eta = \sigma_\gamma \sigma_\beta$ (by making two dihedral angles have equal measure). Then $\varepsilon \perp \beta$ also, hence $\varphi = \varphi\sigma_\delta\sigma_\delta = \sigma_\varepsilon \sigma_\zeta \sigma_\eta \sigma_\delta = \sigma_\varepsilon \sigma_\gamma \sigma_\beta \sigma_\delta = \sigma_c \sigma_b$, where $b = \beta \cap \delta$ and $c = \gamma \cap \varepsilon$. By lemma 2, this is a screw. ♦

Hermann WIENER was born in 1857 in Karlsruhe. His father Christian Wiener was Professor of Descriptive Geometry at the Technical Institute there. His younger brother Otto became a physicist, noted for work on the wave theory of light.

Wiener began university studies at Karlsruhe, but completed his Ph.D. at München in 1880 with a dissertation on plane curves. His training was influenced significantly by Felix Klein, a close collaborator at München with Wiener's cousin Alexander Brill. In 1880, Klein moved to Leipzig, and Wiener followed. After a year, though, he returned to Karlsruhe as assistant to his father. He continued research, and obtained a lecturer position in the university at Halle in 1885.

There Wiener wrote a series of research papers that clarified the relationships of several fundamental theorems and techniques in Euclidean and projective geometry, including Desargues' theorem and the use of self-inverse transformations. He made the first extensive use of calculations with reflections, like those in this and the preceding chapter. He regarded this calculus as an algebra of *geometric* objects, in contrast to the algebra of *numbers* used in coordinate geometry.

Wiener also pioneered the abstract axiomatic approach to geometry, emphasizing its usefulness, but he never built a complete theory. Hilbert's famous remark that he could regard geometry as a theory of tables, chairs, and bar stools just as well as one of points, lines, and planes stemmed from his excitement after hearing a lecture by Wiener in 1890. That took place in Halle at the first meeting of the German Mathematical Society. Wiener was one of the original members.

In 1894 Wiener became Professor of Mathematics at the technical university in Darmstadt, where he remained until he retired in 1927. Darmstadt lay outside the mainstream, and Wiener never continued his major research work. In 1911 he revised and marketed a set of geometric models originally developed by his father; they came into widespread use in secondary and technical school instruction. Wiener served his university well—he was elected Dean of his faculty for two terms. He died in 1939.

Wiener's results were extended and incorporated into major works on foundations of geometry for the next sixty years. In particular, crucial steps in the work of Hilbert, Thomsen, and the present author were established or proposed originally by Hermann Wiener.

Classification theorem

You've now completed the analysis of compositions of reflections. The following theorem refines and summarizes the classification begun with the structure theorem.

---

[5] Wiener 1890–1893.

[6] See Reid 1970, chapter 8.

***Theorem 4* (*Classification theorem*).** A three-dimensional isometry is

(0) the identity ι, or
(1) the reflection σ_ε across a plane ε, or
(2a) a translation τ_V with vector $V \neq O$, or
(2b) a rotation $\rho \neq \iota$, or
(3a) a glide reflection $\sigma_\varepsilon \tau_V$ with mirror ε and vector $V \neq O$, or
(3b) a reflection across a point P, or
(3c) a rotary reflection $\sigma_\varepsilon \rho$ with mirror ε and rotational component ρ that's not self-inverse, or
(4) a screw $\rho \tau_V$ with rotational component $\rho \neq \iota$ and vector $V \neq O$.

These classes are disjoint.

You might like to provide a table of fixed points, lines, and planes for each of the eight isometry types listed in the classification theorem, analogous to those after corollary 6.9.2 and in exercise 6.11.12.

Finding fixpoints, fixed lines, and fixed planes of an isometry $\varphi : X \to X'$ when you know its equation $X' = AX + B$ is a substantial exercise in solving possibly singular linear systems. You'll find examples under the heading **Exercises on equations** in section 7.7. The practical details of those computations tend to differ considerably from those of the theory presented so far. They're outlined in some paragraphs that precede the exercises.

Section 6.9 applied the classification of plane isometries to describe rotations and translations in more detail. It also pursued the notion of conjugacy beyond the level achieved by theorems 7.2.5, 7.3.3, and 7.3.5. For analogous three-dimensional studies, see the **Conjugacy exercises** heading in section 7.7.

7.6 Similarities

Concepts

Complete analogy with the two-dimensional theory
Hometheties and homothetic rotations
Classification theorem

You could really write this section on three-dimensional similarities yourself. Their theory is entirely analogous to its plane counterpart. Except for its main result, the classification theorem, this section just makes the necessary adjustments to the definitions and theorem statements. You can adapt every proof in section 6.10 to work in three dimensions.

A *similarity with ratio* r is a transformation $\psi : P \to P'$ of the set of all points, such that $P'Q' = r\,(PQ)$ for all P and Q. As an example, select any point Z and consider the transformation $\varphi_{r,Z} : P \to P'$ such that $Z' = Z$ and for $P \neq Z$, P' is the point on $\overrightarrow{ZP}$ with $ZP' = r\,(ZP)$. If $r = 1$, then $\varphi_{r,Z}$ is the identity; if $r < 1$ or $r > 1$, it's called a *contraction* or *dilation about* Z.

A similarity ψ: $X \to X'$ with ratio r leaves invariant angle measure, betweenness, and any geometric property that you can define in terms of them. In particular, ψ is a collineation, and the image of a plane is a plane. The image of a circle or sphere with radius t is a circle or sphere with radius rt. That of a triangle with area s is a similar triangle with area $r^2 s$. And the image of a tetrahedron with volume v is a similar tetrahedron with volume $r^3 v$.

A composition of similarities with ratios r and s is a similarity with ratio rs. The inverse of a similarity with ratio r is a similarity with ratio r^{-1}. Therefore, the similarities form a transformation group that contains the isometries.

The section 6.10 proof of the fixpoint theorem was chosen specifically to apply to the three-dimensional case as well: Unless a similarity ψ is an isometry, it has exactly one fixpoint.

The uniqueness theorem takes a slightly different form in three dimensions: Similarities χ and ψ coincide if $\chi(X) = \psi(X)$ for *four noncoplanar* points X. This follows from the fixpoint theorems, since $\chi\psi^{-1}$, a similarity with four noncoplanar fixpoints, must be an isometry, hence the identity.

A similar argument shows that every similarity $\psi : X \to X'$ is the composition of an isometry and a dilation or contraction with the same ratio r and arbitrarily chosen center, so that ψ has an equation $X' = rAX + B$ with an orthogonal matrix A. Moreover, ψ determines A and B uniquely, and every such equation defines a similarity. Since ψ preserves orientation just in case the isometry $X \to AX$ does, ψ is termed *even* or *odd* depending on whether $\det A = +1$ or -1.

The similarities $\varphi_{r,Z}$ and $\varphi_{r,Z}\,\sigma_Z$ are called *homotheties* with center Z. A composition of one of these with a rotation about an axis through Z is called a *homothetic rotation*. If ψ is a similarity but not an isometry, then it has a fixpoint Z. Define $\chi = \psi\varphi_{r^{-1},Z}$ or $\chi = \psi\sigma_Z\varphi_{r^{-1},Z}$ depending on whether ψ is even or odd, so that χ is an even isometry with fixpoint Z and $\psi = \chi\varphi_{r,Z}$ or $\chi\varphi_{r,Z}\sigma_Z$. By theorem 7.5.4, χ is a rotation about an axis through Z, so ψ is a homothetic rotation. This proves

Theorem 1 **(*Classification theorem*).** A similarity is an isometry or a homothetic rotation.

7.7 Exercises

Concepts

Distance-preserving functions
Dilatations
Compositions of commuting point and plane reflections and half turns
Conjugates
Compositions of reflections, translations, and rotations
Classifying isometries and similarities, given their equations
Finding fixpoints and fixed lines of isometries, given their equations
Reflection calculus
Three-reflections theorems

This section contains thirty-two exercises related to the material covered earlier in this chapter. The first three ask for some straightforward proofs omitted earlier. Several others, listed under the headings **Commuting reflections and half turns**, **Conjugacy exercises**, and **Classifying isometries**, extend the chapter's methods beyond the essentials covered earlier, and provide a nearly complete treatment of the interrelationships among reflections, translations, and rotations. The conjugacy exercises firmly establish results you'll often find used without mention in applications. The exercises under the heading **Exercises on equations** familiarize you with the algebraic and numerical computations often required to apply the theory. Some require tedious arithmetic to evaluate determinants and solve linear systems. There was no effort to avoid that. You're expected to use mathematical software for that work. Finally, exercises 5 and 6 and those under the heading **Reflection calculus** provide excursions into areas less central to the program of this book. You've already met their two-dimensional counterparts in chapter 6.

General exercises

Exercises 1 to 3 ask for several proofs omitted from earlier sections of this chapter.

Exercise 1. Prove theorem 7.1.4: The only isometry with four noncoplanar fixpoints is the identity. *Suggestion* : Imitate the proof of theorem 6.2.7.

Exercise 2. Prove theorem 7.4.3: A rotation ρ commutes with a reflection σ_ε just when $\rho = \iota$, or its axis g is perpendicular to the mirror ε, or it's a half turn about a line in ε. *Suggestion* : Imitate the proof of theorem 7.4.1.

Exercise 3. Prove theorem 7.5.1: A translation τ_V commutes with a rotation ρ just when $V = O$, $\rho = \iota$, or $\vec{OV}$ is parallel to the axis of ρ. *Suggestion*: Imitate the proofs of theorems 7.4.2 and 7.4.4.

Exercise 4 is the three-dimensional analog of exercise 6.11.8.

Exercise 4. Show that an isometry φ is even if and only if $\varphi = \chi\chi$ for some isometry χ.

Exercises 5 and 6 pursue excursions into less central themes, already begun in two-dimensional exercises under the section 6.11 headings **Exercises on the definitions of isometry and similarity** and **Dilatations**. If you've solved those in the most efficient way, you should need little to adapt your work to three dimensions.

Exercise 5. Formulate and prove three-dimensional versions of exercises 6.11.40–6.11.41 on the definitions of isometry and similarity.

Exercise 6. Formulate and prove three-dimensional analogs of exercises 6.11.42–6.11.45 on dilatations.

Commuting reflections and half turns

The conditions in plane geometry under which reflections and half turns commute—the nontrivial self-inverse isometries—were laid out at the end of section 6.3. Those ideas played essential roles in the rest of that chapter. In three-dimensional geometry there are *three* types of nontrivial self-inverse isometries. A complete treatment of their commutativity would have clogged the earlier sections of this chapter unacceptably, so they covered only a few aspects. In fact, the theory in this chapter was streamlined to allow that omission. But its *applications* require *all* the commutativity details, so they're presented here systematically, as exercises.

Exercise 7. Show that the composition $\sigma_P \sigma_Q$ of two point reflections is the translation with vector $2(P - Q)$, and they don't commute unless $P = Q$.

Exercise 8. Show that point and plane reflections σ_P and σ_ε commute just when P lies on ε, and in that case $\sigma_P \sigma_\varepsilon$ is the half turn about the line $g \perp \varepsilon$ through P.

Exercise 9. Show that

Part 1. a half turn σ_g commutes with a plane reflection σ_ε if and only if g lies in ε or $g \perp \varepsilon$;

Part 2. if g lies in ε, then $\sigma_g \sigma_\varepsilon$ is the reflection across the plane perpendicular to ε through g;

Part 3. if $g \perp \varepsilon$, then $\sigma_g \sigma_\varepsilon$ is the reflection across $g \cap \varepsilon$.

Exercise 10. Show that a point reflection σ_P and a half turn σ_g commute if and only if P lies on g, and in that case $\sigma_P \sigma_g$ is the reflection in the plane $\varepsilon \perp g$ through P.

Exercise 11. Show that half turns σ_g and σ_h about distinct lines g and h commute if and only if $g \perp h$, and in that case, $\sigma_g \sigma_h$ is the reflection across the intersection of g and h.

Conjugacy exercises

The conjugacy classes of two-dimensional isometries were described in complete detail under that heading in section 6.9. That information is important for applications, because authors often handle, for example, all 60° rotations interchangeably without mentioning explicitly the conjugacy relation that connects them. The same is true in three-dimensional applications. Earlier sections of chapter 7 contained the most essential conjugacy results: the general ideas near the end of section 7.1, and theorems 7.2.5, 7.3.3, and 7.3.5. Further results were omitted because their proofs are unwieldy and repetitive. They're offered here as exercises.

Exercise 12 shows that, like the plane reflections, the half turns and point reflections each constitute single conjugacy classes.

Exercise 12. Show that $\varphi \sigma_P \varphi^{-1} = \sigma_{\varphi(P)}$ and $\varphi \sigma_g \varphi^{-1} = \sigma_{\varphi[g]}$ for any point P, any line g, and any isometry φ. *Suggestion*: Consider fixpoints!

Exercise 13. Describe the conjugacy class of a glide reflection.

Exercise 14. Describe the conjugacy class of a rotary reflection.

Exercise 15. Describe the conjugacy class of a screw.

Classifying isometries

Earlier in this chapter, a glide reflection was defined as a composition of a translation and a commuting plane reflection, a rotary reflection as a composition of a rotation and a commuting plane reflection, and a screw as a composition of a translation and a commuting rotation. With these notions, you can classify *all* isometries. Exercises 16 to 18 give you some practice. What kinds of isometries are the analogous compositions whose components *don't* commute?

Exercise 16. What's the composition of a translation and a plane reflection when they don't commute?

Exercise 17. What's the composition of a rotation and a plane reflection when they don't commute?

Exercise 18. What's the composition of a translation and a rotation when they don't commute?

Exercises 16 to 18 don't cover the *whole* subject of commutativity. When do isometries of the various *other* types commute, and what are their compositions? Most of the combinations are too rare to warrant special treatment—except one: the composition of two rotations. In two dimensions its description is simple; see theorem 6.9.3. In three, that's not true unless the rotations commute. Exercise 19 investigates the situation.

Exercise 19. When do two rotations commute? What's their composition in that case and in general?

Exercises on equations

As in two-dimensional transformational geometry, the theory of isometries in three dimensions seems very different from the practice of computing with them. That's the effect of streamlining to present the essential details of the theory efficiently, in a form that's easy to remember. The exercises under this heading present the computational details. In the first two, closely related to two-dimensional exercise 6.11.25, you derive equations for plane reflections and half turns in terms of the equations of their mirrors and axes.

Exercise 20. Consider a point $P = \langle p_1, p_2, p_3 \rangle \neq O$ and the plane $\delta \perp \overset{\to}{OP}$ through O. Show that σ_δ has matrix

$$A = I - \frac{2}{P^tP}K,$$

where $K = PP^t$. Now consider the plane $\varepsilon \perp \overset{\to}{OP}$ through a point Q. Show that $\sigma_\varepsilon : X \to X'$ has equation

$$X' = AX + \frac{2}{P^tP}KQ.$$

Exercise 21. Find an equation for the half turn $X \to X'$ about the line g with parametric equations $X = tP + Q$. *Suggestion*: Use the matrix $K = PP^t$ introduced in exercise 20.

In the next two exercises, you start from geometric descriptions of certain isometries and find their equations. These involve very tedious computations. Use mathematical software wherever possible!

Exercise 22. Find equations for

(a) the reflection across the plane ε with equation $x_1 + 2x_2 + 3x_3 = 4$;

(b) the glide reflection with mirror ε and vector $<3,-2,1/9>$;

(c) the half turn about the line through points $<1,1,2>$ and $<2,1,1>$.

Exercise 23. Consider the plane ε in exercise 22, its intersections P, Q, and R with the coordinate axes, and the line $g \perp \varepsilon$ through Q. Find equations for

(a) the rotation $\rho : X \to X'$ about g that maps $\vec{QR}$ to $\vec{QP}$, with coefficients accurate to four decimal places;

(b) the rotary reflection $\rho\sigma_\varepsilon$.

Suggestion: The equation has the form $X' = AX + B$ for some matrix A and some vector B. Find four points X, no three of which are coplanar, for which you know X'. From the four equations $X' = AX + B$ construct a system of twelve linear equations for the twelve entries of A and B, then solve it.

The final two exercises under this heading ask you to classify an isometry or similarity, given its equation. To *classify* an isometry,

- identify its class among those listed in the classification theorem;
- find an equation for its mirror—if it has one;
- find its vector—if it has one;
- find an equation for its axis—if it has one—and specify a vector to indicate a positive ray, then find a corresponding angle parameter.

To complete a classification, you may need to find fixpoints or fixed lines of an isometry $\varphi : X \to X' = AX + B$. For a fixpoint, you'll have to solve the linear system $X = X'$ —that is, $X = AX + B$, or $(I - A)X = B$. If $I - A$ is invertible, there's just one fixpoint X. Otherwise, reduce the system to an equivalent one for which you can identify the solutions readily. There are four cases. You might

- derive an *inconsistent* equation equivalent to $0 = 1$ —there's *no* fixpoint;
- reduce the system to three *trivial* equations, each equivalent to $0 = 0$ —*every* point is fixed;

- reduce it to two trivial equations and a nontrivial one that identifies a *plane* of fixpoints, with equation $c_1x_1 + c_2x_2 + c_3x_3 = d$;
- reduce it to one trivial equation and two equations

  $$\begin{cases} x_i - c_i x_k = d_i \\ x_j - c_j x_k = d_j \,, \end{cases}$$

  where $\{i, j, k\} = \{1, 2, 3\}$ —you can convert those to parametric equations for a *line* of fixpoints:

  $$\begin{cases} x_i = t\,c_i + d_i \\ x_j = t\,c_j + d_j \\ x_k = t \cdot 1 + 0 \,. \end{cases}$$

To find fixed lines g of an isometry φ, note that φ maps points $tP + Q$ with $P \neq O$ to points

$$A(tP + Q) + B = tP^* + Q',$$

where $P^* = AP$ and $Q' = AQ + B$. Thus φ maps the line $g \parallel \vec{OP}$ through Q onto a line $g' \parallel \vec{OP}^*$ through Q'. It follows that $g \parallel g'$ just in case P^* is a scalar multiple of P. But φ preserves length, so $g \parallel g'$ just in case $P^* = \pm P$. The first step in finding fixed lines is to find all vectors P such that $P^* = \pm P$. Solve the linear systems $AP = \pm P$ —that is, $(A \pm I)P = O$. Second, find the points Q for which $g = g'$. For each P, both Q and Q' must lie on g, so

$$Q' - Q = AQ + B - Q = (A - I)Q + B = tP$$

for some t. You must solve the system

$$(I - A)Q + tP = B$$

of three linear equations in four unknowns: t and the components of Q.

That general procedure for finding fixed lines may be more involved than what you really need. For these classification exercises, your main fixed-line problem is probably to find the axis g of a screw $\varphi : X \to X' = AX + B$ that has no fixpoint. Here's a neater alternative for that special situation. Since φ affects points X on g like a translation with vector V,

$$\begin{aligned} &X' - X = V = (X')' - X', \\ &AX + B - X = [A(AX + B) + B] - [AX + B] \\ &(A^2 - 2A + I)X = (I - A)B \\ &(I - A)^2 X = (I - A)B. \end{aligned}$$

You can find the points X on g by solving that singular linear system!

You can recognize a similarity $X \to AX + B$ with ratio r by computing $AA^t = r^2 I$; then classify the corresponding isometry $X \to r^{-1}AX + r^{-1}B$.

These exercises are stated in terms of exact fractions. To get *exact* solutions, you'll need mathematical software that operates with exact fractions.[7] You *won't* be able to carry out the computations by hand or by calculator with decimal arithmetic.

You *could* convert each fraction in an exercise to a decimal approximation, and use decimal arithmetic. Do some exercises that way, too! Carry as many decimal digits as your calculator permits. You'll have to make certain decisions in the classification procedure by testing whether some computed number $d = 0$. With approximate decimal arithmetic, you'll never get $d = 0$ *exactly*. So you'll have to decide how close to zero is close enough. In practice, that can be a critical decision. It's important enough in some areas of computer graphics that it's worth covering in more detail *there*. But *these* exercises provide exact fractions so you can clearly relate theory to computation without worrying about approximation errors that depend on your equipment.

Exercise 24. For these matrices A verify that the equation $X' = AX$ defines an isometry or similarity $\varphi : X \to X'$, and classify it.

(a) $\begin{bmatrix} \frac{6}{7} & -\frac{2}{7} & -\frac{3}{7} \\ -\frac{2}{7} & \frac{3}{7} & -\frac{6}{7} \\ -\frac{3}{7} & -\frac{6}{7} & -\frac{2}{7} \end{bmatrix}$ (b) $\begin{bmatrix} \frac{12}{49} & \frac{31}{49} & -\frac{36}{49} \\ \frac{24}{49} & -\frac{36}{49} & -\frac{23}{49} \\ -\frac{41}{49} & -\frac{12}{49} & -\frac{24}{49} \end{bmatrix}$

(c) $\begin{bmatrix} -\frac{9518}{9555} & \frac{131}{1911} & \frac{526}{9555} \\ -\frac{142}{1911} & -\frac{1894}{1911} & -\frac{211}{1911} \\ -\frac{449}{9555} & \frac{218}{1911} & -\frac{9482}{9555} \end{bmatrix}$ (d) $\begin{bmatrix} -6 & 2 & 3 \\ 2 & -3 & 6 \\ 3 & 6 & 2 \end{bmatrix}$.

Exercise 25. For these matrices A and vectors B verify that the equation $X' = AX + B$ defines an isometry or similarity $\varphi : X \to X'$, and classify it.

(a) Same A as exercise 24(a), $B = <{}^4/_7, {}^8/_7, {}^{12}/_7>$;

(b) Same A as exercise 24(a), $B = <5, 6, 7>$;

(c) Same A as exercise 24(b), $B = <{}^4/_7, {}^8/_7, {}^{12}/_7>$;

(d) Same A as exercise 24(c), $B = <{}^{968}/_{195}, {}^{88}/_{39}, -{}^{616}/_{195}>$;

(e) $A = \begin{bmatrix} -\frac{9518}{9555} & -\frac{131}{1911} & \frac{526}{9555} \\ -\frac{142}{1911} & \frac{1894}{1911} & -\frac{211}{1911} \\ -\frac{449}{9555} & -\frac{218}{1911} & -\frac{9482}{9555} \end{bmatrix}$ $B = <{}^{968}/_{195}, {}^{88}/_{39}, -{}^{616}/_{195}>$;

---

[7] *X*(*Plore*) (Meredith 1993) is a good choice.

$$\text{(f) } A = \begin{bmatrix} -9518 & -655 & 526 \\ -710 & 9470 & -1055 \\ -449 & -1090 & -9482 \end{bmatrix} \qquad B = \langle 968, 440, -616 \rangle .$$

Reflection calculus

Just as in two dimensions, geometers have developed approaches to three-dimensional geometry that rely very heavily on algebraic computations with compositions of self-inverse isometries—plane and point reflections and half turns. Because of the greater variety, these approaches are not so elegant as their plane counterparts. The *three-reflections theorems* provide basic tools in both two and three dimensions. The two-dimensional results are theorems 6.4.7, 6.5.6, and 6.7.2, and exercise 6.11.32. Their analogs haven't appeared yet in this more streamlined chapter. They're presented as exercises 26, 27, and 29. The intervening exercise 28 is an extension of lemma 7.5.2 that you'll need for exercise 29.

Exercise 26. Show that for any points P, Q, and R, the composition $\sigma_P \sigma_Q \sigma_R$ is the reflection across point $S = P - Q + R$, and $PQRS$ is a (possibly degenerate) parallelogram.

Exercise 27. Show that the composition φ of the reflections across three planes is self-inverse in just three cases:

(a) they're all perpendicular to a single line g, or
(b) their intersection contains a line g, or
(c) they're mutually perpendicular.

Also show that in case

(a) φ is the reflection across a plane perpendicular to g, and
(b) φ is the reflection across a plane through g, and
(c) φ is the reflection across the intersection of the three planes.

Exercise 28. According to lemma 7.5.2, for any lines b and c the composition $\sigma_b \sigma_c$ is a screw. Fill in the blanks and prove this extension of the lemma:

- $\sigma_b \sigma_c = \iota$ if and only if … ;
- it's a translation if and only if … ;
- it's a rotation if and only if … ;
- if it's not a translation, then its axis is the line that's perpendicular to b and c, and its angle parameter is … ;
- if it's not a rotation, then its vector is … .

Exercise 29. Show that the composition φ of the half turns about three lines is self-inverse in just three cases:

(a) the lines are mutually perpendicular, or
(b) some pair of them is parallel, or
(c) they have a common perpendicular line m.

Show also that in case

(a) $\varphi = \iota$,
(b) φ is the half turn about a line l,
(c) φ is the half turn about a line $l \perp m$.

Exercise 30 is the three-dimensional analog of exercise 6.11.9.

Exercise 30. Show that every isometry φ is the composition of two self-inverse isometries. *Suggestion*: Consider separately the various types of isometry.

Working in the 1920s and early 1930s on the two-dimensional reflection calculus—see that heading in section 6.11—Gerhard Thomsen evidently considered analogous three-dimensional questions. Thomsen's Ph.D. student H. Boldt developed that material into a full theory, published as his dissertation in 1934. It isn't as clean as its two-dimensional counterpart. That's understandable; three-dimensional theorems are generally more complicated. As you've seen in previous exercises on commuting reflections and on the three-reflections theorems, enlarging the class of nontrivial self-inverse isometries from two types to three brings minor complications to almost every elementary theorem. Consequently, even only slightly less elementary results become unwieldy when stated solely in terms of reflections and half turns. Exercises 31 and 32 are taken from Boldt 1934. They're closely related to the two-dimensional reflection calculus exercises in section 6.11.

Exercise 31. Describe the figure formed by points A, B, C, and G for which $\sigma_A\sigma_G\sigma_B\sigma_G\sigma_C\sigma_G = \iota$. Find an analogous result with points A, B, C, D, and G. *Suggestion*: Reason with vectors.

Exercise 32, Part 1. Under what conditions on lines g and h is $\sigma_g\sigma_h\sigma_g\sigma_h = \iota$?

Part 2. Under what conditions on a line g and points P and Q is $\sigma_g\sigma_P\sigma_Q\sigma_g\sigma_Q\sigma_P = \iota$?

Part 3. Under what conditions on a line g and points P and Q is $\sigma_g\sigma_P\sigma_Q\sigma_g\sigma_P\sigma_Q = \iota$?

Part 4. Under what conditions on lines g and h do there exist points P and Q such that $\sigma_g \sigma_h = \sigma_P \sigma_Q$?

Part 5. Under what conditions on lines g and h do there exist points P and Q such that $\sigma_g \sigma_h \sigma_g \sigma_h = \sigma_P \sigma_Q$?

Suggestion: Exercise 28 is helpful.

3.2.6 Convex Functions

A function is called convex if any segment with endpoints on its graph lies above the graph itself. The picture you should have in mind is Figure 21. Formally, if D is an interval of the real axis, or more generally a convex subset of a vector space, then a function $f : D \to \mathbb{R}$ is called convex if

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y), \quad \text{for all } x, y \in D, \quad \lambda \in (0, 1).$$

Here we should remember that a set D is called convex if for any $x, y \in D$ and $\lambda \in (0, 1)$ the point $\lambda x + (1 - \lambda)y$ is also in D, which geometrically means that D is an intersection of half-spaces.

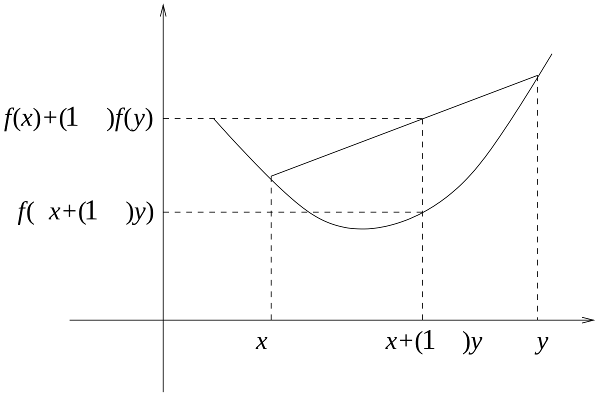

Figure 21

A function f is called concave if $-f$ is convex. If f is both convex and concave, then f is linear, i.e., $f(x) = ax + b$ for some constants a and b.

Proposition. *A twice-differentiable function on an interval is convex if and only if its second derivative is nonnegative.*

In general, a twice-differentiable function defined on a convex domain in $\mathbb{R}^n$ is convex if at any point its Hessian matrix is semipositive definite. This is a way of saying that modulo a local change of coordinates, around each point the function f is of the form

$$f(x_1, x_2, \ldots, x_n) = \phi(x_1, x_2, \ldots, x_n) + x_1^2 + x_2^2 + \cdots + x_k^2,$$

where $k \leq n$ and $\phi(x_1, x_2, \ldots, x_n)$ is linear.

As an application, we use convexity to prove Hölder's inequality.

Hölder's inequality. *If* $x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n, p$, *and* q *are positive numbers with* $\frac{1}{p} + \frac{1}{q} = 1$, *then*

$$\sum_{i=1}^{n} x_i y_i \le \left(\sum_{i=1}^{n} x_i^p\right)^{1/p} \left(\sum_{i=1}^{n} y_i^q\right)^{1/q},$$

with equality if and only if the two vectors $(x_1, x_2, \ldots, x_n)$ *and* $(y_1, y_2, \ldots, y_n)$ *are parallel.*

Proof. The second derivative of $f : (0, \infty) \to \mathbb{R}$, $f(x) = \ln x$, is $f''(x) = -\frac{1}{x^2}$, which is negative. So this function is concave. Setting $\lambda = \frac{1}{p}$, we obtain

$$\ln X^{1/p}Y^{1/q} = \frac{1}{p}\ln X + \frac{1}{q}\ln Y \le \ln\left(\frac{1}{p}X + \frac{1}{q}Y\right), \quad \text{for all } X, Y > 0;$$

hence

$$X^{1/p}Y^{1/q} \le \frac{1}{p}X + \frac{1}{q}Y.$$

Using this fact, if we let $X = \sum_i x_i^p$ and $Y = \sum_i y_i^q$, then

$$\frac{1}{X^{1/p}Y^{1/q}} \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} \left(\frac{x_i^p}{X}\right)^{1/p} \left(\frac{y_i^q}{Y}\right)^{1/q} \le \sum_{i=1}^{n} \left(\frac{1}{p} \cdot \frac{x_i^p}{X} + \frac{1}{q} \cdot \frac{y_i^q}{Y}\right)$$
$$= \left(\frac{1}{p} + \frac{1}{q}\right) = 1.$$

Hence

$$\sum_{i=1}^{n} x_i y_i \le X^{1/p}Y^{1/q} = \left(\sum_{i=1}^{n} x_i^p\right)^{1/p} \left(\sum_{i=1}^{n} y_i^q\right)^{1/q},$$

and the inequality is proved.

By analogy, a sequence $(a_n)_{n \ge 0}$ is called convex if

$$a_n \le \frac{a_{n+1} + a_{n-1}}{2}, \quad \text{for all } n \ge 1,$$

and concave if $(-a_n)_n$ is convex. Equivalently, a sequence is convex if its second difference (derivative) is nonnegative, and concave if its second difference is nonpositive. The following example motivates why convex sequences and functions should be studied together.

Example. Let $(a_n)_n$ be a bounded convex sequence. Prove that

$$\lim_{n \to \infty} (a_{n+1} - a_n) = 0.$$

Solution. A bounded convex function on $(0, \infty)$ has a horizontal asymptote, so its derivative tends to zero at infinity. Our problem is the discrete version of this result. The first derivative of the sequence is $b_n = a_{n+1} - a_n$, $n \ge 1$. The convexity condition can be written as $a_{n+1} - a_n \ge a_n - a_{n-1}$, which shows that $(b_n)_n$ is increasing. Since $(a_n)_n$ is bounded, $(b_n)_n$ is bounded too, and being monotonic, by the Weierstrass theorem it converges at a finite limit L. If $L > 0$, then b_n eventually becomes positive, so a_n becomes increasing because it has a positive derivative. Again by the Weierstrass theorem, a_n converges to some limit l, and then $L = l - l = 0$, a contradiction. A similar argument rules out the case $L < 0$. We are left with the only possibility $L = 0$.

And now some problems.

425. Let $x_1, x_2, \ldots, x_n$ be real numbers. Find the real numbers a that minimize the expression

$$|a - x_1| + |a - x_2| + \cdots + |a - x_n|.$$

426. Let $a, b > 0$ and $x, c > 1$. Prove that

$$x^{a^c} + x^{b^c} \ge 2x^{(ab)^{c/2}}.$$

427. A triangle has side lengths $a \ge b \ge c$ and vertices of measures A, B, and C, respectively. Prove that

$$Ab + Bc + Ca \ge Ac + Ba + Cb.$$

428. Show that if a function $f : [a, b] \to \mathbb{R}$ is convex, then it is continuous on (a, b).

429. Prove that a continuous function defined on a convex domain (for example, on an interval of the real axis) is convex if and only if

$$f\left(\frac{x+y}{2}\right) \le \frac{f(x) + f(y)}{2}, \quad \text{for all } x, y \in D.$$

430. Call a real-valued function *very convex* if

$$\frac{f(x) + f(y)}{2} \ge f\left(\frac{x+y}{2}\right) + |x - y|$$

holds for all real numbers x and y. Prove that no very convex function exists.

431. Let $f : [a, b] \to \mathbb{R}$ be a convex function. Prove that

$$f(x) + f(y) + f(z) + 3f\left(\frac{x+y+z}{3}\right) \ge 2\left[f\left(\frac{x+y}{2}\right) + f\left(\frac{y+z}{2}\right) + f\left(\frac{z+x}{2}\right)\right],$$

for all $x, y, z \in [a, b]$.

432. Prove that if a sequence of positive real numbers $(b_n)_n$ has the property that $(a^n b_n)_n$ is a convex sequence for all real numbers a, then the sequence $(\ln b_n)_n$ is also convex.

433. Find the largest constant C such that for every $n \geq 3$ and every positive concave sequence $(a_k)_{k=1}^n$,

$$\left(\sum_{k=1}^{n} a_k\right)^2 \geq C(n-1)\sum_{k=1}^{n} a_k^2.$$

A convex function on a closed interval attains its maximum at an endpoint of the interval. We illustrate how this fact can be useful with a problem from *Timişoara Mathematics Gazette*, proposed by V. Cârtoaje and M. Lascu.

Example. Let $a, b, c, d \in [1, 3]$. Prove that

$$(a+b+c+d)^2 \geq 3(a^2+b^2+c^2+d^2).$$

Solution. Divide by 2 and move everything to one side to obtain the equivalent inequality

$$a^2+b^2+c^2+d^2-2ab-2ac-2ad-2bc-2bd-2cd \leq 0.$$

Now we recognize the expression on the left to be a convex function in each variable. So the maximum is attained for some choice of $a, b, c, d = 1$ or 3. If k of these numbers are equal to 3, and $4-k$ are equal to 1, where k could be 1, 2, 3, or 4, then the original inequality becomes

$$(3k+4-k)^2 = 3(9k+4-k).$$

Dividing by 3, we obtain $k^2+4k+4 \geq 6k+3$, or $(k-1)^2 \geq 0$, which is clearly true. The inequality is proved. Equality occurs when one of the numbers a, b, c, d is equal to 3 and the other three are equal to 1.

Here are additional problems of this kind.

434. Let α, β, and γ be three fixed positive numbers and $[a, b]$ a given interval. Find x, y, z in $[a, b]$ for which the expression

$$E(x, y, z) = \alpha(x-y)^2 + \beta(y-z)^2 + \gamma(z-x)^2$$

has maximal value.

435. Let $0 < a < b$ and $t_i \geq 0$, $i = 1, 2, \ldots, n$. Prove that for any $x_1, x_2, \ldots, x_n \in [a, b]$,

$$\left(\sum_{i=1}^{n} t_i x_i\right)\left(\sum_{i=1}^{n} \frac{t_i}{x_i}\right) \leq \frac{(a+b)^2}{4ab}\left(\sum_{i=1}^{n} t_i\right)^2.$$

436. Prove that for any natural number $n \geq 2$ and any $|x| \leq 1$,

$$(1+x)^n + (1-x)^n \leq 2^n.$$

437. Prove that for any positive real numbers a, b, c the following inequality holds

$$\frac{a+b+c}{3} - \sqrt[3]{abc} \leq \max\{(\sqrt{a}-\sqrt{b})^2, (\sqrt{b}-\sqrt{c})^2, (\sqrt{c}-\sqrt{a})^2\}.$$

438. Let f be a real-valued continuous function on $\mathbb{R}$ satisfying

$$f(x) \leq \frac{1}{2h}\int_{x-h}^{x+h} f(y)dy, \quad \text{for all } x \in \mathbb{R} \text{ and } h > 0.$$

Prove that (a) the maximum of f on any closed interval is assumed at one of the endpoints, and (b) the function f is convex.

An important property of convex (respectively, concave) functions is known as Jensen's inequality.

Jensen's inequality. *For a convex function f let $x_1, x_2, \ldots, x_n$ be points in its domain and let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be positive numbers with $\lambda_1 + \lambda_2 + \cdots + \lambda_n = 1$. Then*

$$f(\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n) \leq \lambda_1 f(x_1) + \lambda_2 f(x_2) + \cdots + \lambda_n f(x_n).$$

If f is nowhere linear and the x_i's are not all equal, then the inequality is strict. The inequality is reversed for a concave function.

Proof. The proof is by induction on n. The base case is the definition of convexity. Let us assume that the inequality is true for any $n-1$ points x_i and any $n-1$ weights λ_i. Consider n points and weights, and let $\lambda = \lambda_1 + \cdots + \lambda_{n-1}$. Note that $\lambda + \lambda_n = 1$ and $\frac{\lambda_1}{\lambda} + \frac{\lambda_2}{\lambda} + \cdots + \frac{\lambda_{n-1}}{\lambda} = 1$. Using the base case and the inductive hypothesis we can write

$$\begin{aligned}
f(\lambda_1 x_1 + \cdots + \lambda_{n-1}x_{n-1} + \lambda_n x_n) &= f\left(\lambda\left(\frac{\lambda_1}{\lambda}x_1 + \cdots + \frac{\lambda_{n-1}}{\lambda}x_{n-1}\right) + \lambda_n x_n\right) \\
&\leq \lambda f\left(\frac{\lambda_1}{\lambda}x_1 + \cdots + \frac{\lambda_{n-1}}{\lambda}x_{n-1}\right) + \lambda_n f(x_n) \\
&\leq \lambda\left(\frac{\lambda_1}{\lambda}f(x_1) + \cdots + \frac{\lambda_{n-1}}{\lambda}f(x_{n-1})\right) + \lambda_n f(x_n) \\
&= \lambda_1 f(x_1) + \cdots + \lambda_{n-1}f(x_{n-1}) + \lambda_n f(x_n),
\end{aligned}$$

as desired. For the case of concave functions, reverse the inequalities.

As an application, we prove the following.

The generalized mean inequality. *Given the positive numbers* $x_1, x_2, \ldots, x_n$ *and the positive weights* $_1, \ _2, \ldots, \ _n$ *with* $_1 + \ _2 + \cdots + \ _n = 1$*, the following inequality holds:*

$$_1x_1 + \ _2x_2 + \cdots + \ _nx_n \quad x_1^{\ 1}x_2^{\ 2}\cdots x_n^{\ n}.$$

Solution. Simply write Jensen s inequality for the concave function $f(x) = \ln x$, then exponentiate.

For $_1 = \ _2 = \cdots = \ _n = \frac{1}{n}$ one obtains the AM GM inequality.

439. Show that if A, B, C are the angles of a triangle, then

$$\sin A + \sin B + \sin C \quad \frac{3\ \overline{3}}{2}.$$

440. Let a_i, $i = 1, 2, \ldots, n$, be nonnegative numbers with $\ {}_{i=1}^{n} a_i = 1$, and let $0 < x_i \quad 1$, $i = 1, 2, \ldots, n$. Prove that

$$\ {}_{i=1}^{n} \frac{a_i}{1 + x_i} \quad \frac{1}{1 + x_1^{a_1}x_2^{a_2}\cdots x_n^{a_n}}.$$

441. Prove that for any three positive real numbers a_1, a_2, a_3,

$$\frac{a_1^2 + a_2^2 + a_3^2}{a_1^3 + a_2^3 + a_3^3} \quad \frac{a_1^3 + a_2^3 + a_3^3}{a_1^4 + a_2^4 + a_3^4}.$$

442. Let $0 < x_i < \ $, $i = 1, 2, \ldots, n$, and set $x = \frac{x_1 + x_2 + \cdots + x_n}{n}$. Prove that

$$\ {}_{i=1}^{n} \frac{\sin x_i}{x_i} \quad \frac{\sin x}{x}^{n}.$$

443. Let $n > 1$ and $x_1, x_2, \ldots, x_n > 0$ be such that $x_1 + x_2 + \cdots + x_n = 1$. Prove that

$$\frac{x_1}{\overline{1 \quad x_1}} + \frac{x_2}{\overline{1 \quad x_2}} + \cdots + \frac{x_n}{\overline{1 \quad x_n}} \quad \frac{\overline{x_1} + \ \overline{x_2} + \cdots + \ \overline{x_n}}{\overline{n \quad 1}}.$$

3.2.7 Indefinite Integrals

Anyone who stops learning is old, whether at twenty or eighty. Anyone who keeps learning stays young. The greatest thing in life is to keep your mind young. Following this advice of Henry Ford, let us teach you some clever tricks for computing indefinite integrals.

518. Prove the Gaussian integral formula

$$\int_{-\infty}^{\infty} e^{-x^2}dx = \sqrt{\pi}.$$

519. Evaluate

$$\int_0^1\int_0^1\int_0^1 (1 + u^2 + v^2 + w^2)^{-2}dudvdw.$$

520. Let $D = \{(x, y) \in \mathbb{R}^2 \mid 0 \leq x \leq y \leq \pi\}$. Prove that

$$\iint_D \ln|\sin(x - y)|dxdy = -\frac{\pi^2}{2}\ln 2.$$

Our next topic is the continuous analogue of the change of the order of summation in a double sum.

Fubini's theorem. *Let* $f : \mathbb{R}^2 \to \mathbb{R}$ *be a piecewise continuous function such that*

$$\int_c^d\int_a^b |f(x, y)|dxdy < \infty.$$

Then

$$\int_c^d\int_a^b f(x, y)dxdy = \int_a^b\int_c^d f(x, y)dydx.$$

The matter of convergence can be bypassed for positive functions, in which case we have the following result.

Tonelli's theorem. *Let* $f : \mathbb{R}^2 \to \mathbb{R}$ *be a positive piecewise continuous function. Then*

$$\int_a^b\int_c^d f(x, y)dxdy = \int_c^d\int_a^b f(x, y)dydx.$$

The limits of integration can be finite or infinite. In the particular case that $f(x, y)$ is constant on the squares of an integer lattice, we recover the discrete version of Fubini's theorem, the change of order of summation in a double sum

$$\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} f(m, n) = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} f(m, n).$$

A slightly more general situation occurs when f is a step function in one of the variables. In this case we recover the formula for commuting the sum with the integral:

$$\int_a^b \sum_{n=0}^{\infty} f(n,x) = \sum_{n=0}^{\infty} \int_a^b f(n,x).$$

Here we are allowed to commute the sum and the integral if either f is a positive function, or if $\int_a^b \sum_{n=0}^{\infty} |f(n,x)|$ (or equivalently $\sum_{n=0}^{\infty} \int_a^b |f(n,x)|$) is finite. It is now time for an application.

Example. Compute the integral

$$I = \int_0^{\infty} \frac{1}{\sqrt{x}} e^{-x} dx.$$

Solution. We will replace $\frac{1}{\sqrt{x}}$ by a Gaussian integral. Note that for $x > 0$,

$$\int_{-\infty}^{\infty} e^{-xt^2} dt = \int_{-\infty}^{\infty} e^{-(\sqrt{x}t)^2} dt = \frac{1}{\sqrt{x}} \int_{-\infty}^{\infty} e^{-u^2} du = \frac{\sqrt{\pi}}{\sqrt{x}}.$$

Returning to the problem, we are integrating the positive function $\frac{1}{\sqrt{x}} e^{-x}$, which is integrable over the positive semiaxis because in a neighborhood of zero it is bounded from above by $\frac{1}{\sqrt{x}}$ and in a neighborhood of infinity it is bounded from above by $e^{-x/2}$.

Let us consider the two-variable function $f(x,y) = e^{-xt^2} e^{-x}$, which is positive and integrable over $\mathbb{R} \times (0, \infty)$. Using the above considerations and Tonelli's theorem, we can write

$$I = \int_0^{\infty} \frac{1}{\sqrt{x}} e^{-x} dx = \frac{1}{\sqrt{\pi}} \int_0^{\infty} \int_{-\infty}^{\infty} e^{-xt^2} e^{-x} dt dx = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \int_0^{\infty} e^{-(t^2+1)x} dx dt$$
$$= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \frac{1}{t^2+1} dt = \frac{\pi}{\sqrt{\pi}} = \sqrt{\pi}.$$

Hence the value of the integral in question is $I = \sqrt{\pi}$.

More applications are given below.

521. Let $a_1 \le a_2 \le \cdots \le a_n = m$ be positive integers. Denote by b_k the number of those a_i for which $a_i \ge k$. Prove that

$$a_1 + a_2 + \cdots + a_n = b_1 + b_2 + \cdots + b_m.$$

522. Show that for $s > 0$,

$$\int_0^{\infty} e^{-sx} x^{-1} \sin x dx = \arctan(s^{-1}).$$

523. Show that for $a, b > 0$,

$$\int_0^{\infty} \frac{e^{-ax} - e^{-bx}}{x}dx = \ln \frac{b}{a}.$$

524. Let $|x| < 1$. Prove that

$$\sum_{n=1}^{\infty} \frac{x^n}{n^2} = -\int_0^x \frac{1}{t}\ln(1-t)dt.$$

525. Let $F(x) = \sum_{n=1}^{\infty} \frac{1}{x^2+n^4}$, $x \in \mathbb{R}$. Compute $\int_0^{+\infty} F(t)dt$.

3.3.3 The Many Versions of Stokes' Theorem

We advise you that this is probably the most difficult section of the book. Yet Stokes' theorem plays such an important role in mathematics that it deserves an extensive treatment. As an encouragement, we offer you a quote by Marie Curie: "Nothing in life is to be feared. It is only to be understood."

In its general form, Stokes' theorem is known as

$$\int_M d\omega = \int_{\partial M} \omega,$$

where ω is a k-form, $d\omega$ its differential, and M a domain with boundary ∂M. The one-dimensional case is the most familiar; it is the Leibniz–Newton formula

$$\int_a^b f'(t)dt = f(b) - f(a).$$

Three versions of this result are of interest to us.

Green's theorem. *Let* D *be a domain in the plane with boundary* C *oriented such that* D *is to the left. If the vector field* $\vec{F}(x, y) = P(x, y)\vec{i} + Q(x, y)\vec{j}$ *is continuously differentiable on* D*, then*

$$\oint_C Pdx + Qdy = \iint_D \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dxdy.$$

Stokes' Theorem. *Let* S *be an oriented surface with normal vector* $\vec{n}$*, bounded by a closed, piecewise smooth curve* C *that is oriented such that if one travels on* C *with the upward direction* $\vec{n}$*, the surface is on the left. If* $\vec{F}$ *is a vector field that is continuously differentiable on* S*, then*

$$\oint_C \vec{F} \cdot d\vec{R} = \iint_S (\text{curl}\,\vec{F} \cdot \vec{n})dS,$$

where dS *is the area element on the surface.*

The Gauss Ostrogradsky (Divergence) Theorem. *Let* S *be a smooth, orientable surface that encloses a solid region* V *in space. If* F *is a continuously differentiable vector field on* V*, then*

$$\mathop{}_{S} F \cdot n\, dS = \mathop{}_{V} \operatorname{div} F\, dV,$$

where n *is the outward unit normal vector to the surface* S, dS *is the area element on the surface, and* dV *is the volume element inside of* V.

We recall that for a vector field $F = (F_1, F_2, F_3)$, the divergence is

$$\operatorname{div} F = \quad \cdot F = \frac{F_1}{x} + \frac{F_2}{y} + \frac{F_3}{z},$$

while the curl is

$$\operatorname{curl} F = \quad \times F = \begin{matrix} i & j & k \\ \frac{}{x} & \frac{}{y} & \frac{}{z} \\ F_1 & F_2 & F_3 \end{matrix}$$

$$= \quad \frac{F_3}{y} \quad \frac{F_2}{z} \quad i + \quad \frac{F_1}{z} \quad \frac{F_3}{x} \quad j + \quad \frac{F_2}{x} \quad \frac{F_1}{y} \quad k.$$

The quantity $\overset{++}{}_{S} F \cdot n\, dS$ is called the flux of F across the surface S.

Let us illustrate the use of these theorems with some examples. We start with an encouraging problem whose solution is based on Stokes theorem.

Example. Compute

$$\overset{5}{}_{C}\, y dx + z dy + x dz,$$

where C is the circle $x^2 + y^2 + z^2 = 1$, $x + y + z = 1$, oriented counterclockwise when seen from the positive side of the x-axis.

Solution. By Stokes theorem,

$$\overset{5}{}_{C}\, y dx + z dy + x dz = \mathop{}_{S} \operatorname{curl} F \cdot n\, dS,$$

where S is the disk that the circle bounds. It is straightforward that $\operatorname{curl} F = (\ 1,\ 1,\ 1)$, while n, the normal vector to the plane $x + y + z = 1$, is equal to $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Therefore,

$$\oint_C y\,dx + z\,dy + x\,dz = -A\sqrt{3},$$

where A is the area of the disk bounded by C. Observe that C is the circumcircle of the triangle with vertices $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$. The circumradius of this triangle is $\frac{\sqrt{6}}{3}$, so $A = \frac{2\pi}{3}$. The answer to the problem is therefore $-\frac{2\pi\sqrt{3}}{3}$.

Example. Orthogonal to each face of a polyhedron construct an outward vector with length numerically equal to the area of the face. Prove that the sum of all these vectors is equal to zero.

Solution. We exhibit first an elementary solution based on vector operations. Consider the particular case of a tetrahedron $ABCD$. The four vectors are $\frac{1}{2}BC \times BA$, $\frac{1}{2}BA \times BD$, $\frac{1}{2}BD \times BC$, and $\frac{1}{2}DA \times DC$. Indeed, the lengths of these vectors are numerically equal to the areas of the corresponding faces, and the cross-product of two vectors is perpendicular to the plane determined by the vectors, and it points outward because of the right-hand rule. We have

$$\begin{aligned}
&BC \times BA + BA \times BD + BD \times BC + DA \times DC\\
&\quad = BC \times BA + BA \times BD + BD \times BC + (BA - BD) \times (BC - BD)\\
&\quad = BC \times BA + BA \times BD + BD \times BC + BA \times BC - BA \times BD\\
&\qquad - BD \times BC + 0 = 0.
\end{aligned}$$

This proves that the four vectors add up to zero.

In the general case, dissect the polyhedron into tetrahedra cutting the faces into triangles by diagonals and then joining the centroid of the polyhedron with the vertices. Sum up all vectors perpendicular to the faces of these tetrahedra, and note that the vectors corresponding to internal walls cancel out.

The elegant solution uses integrals. Let S be the polyhedron and assume that its interior V is filled with gas at a (not necessarily constant) pressure p. The force that the gas exerts on S is $\iint_S p\,n\,dA$, where n is the outward normal vector to the surface of the polyhedron and dA is the area element. The divergence theorem implies that

$$\iint_S p\,n\,dA = \iiint_V \nabla p\,dV.$$

Here ∇p denotes the gradient of p. If the pressure p is constant, then the right-hand side is equal to zero. This is the case with our polyhedron, where $p = 1$. The double integral is exactly the sum of the vectors under discussion, these vectors being the forces exerted by pressure on the faces.

As a corollary, we obtain the well-known fact that a container filled with gas under pressure is at equilibrium; a balloon will never move as a result of internal pressure.

We conclude our series of examples with an application of Green s theorem, the proof given by D. Pompeiu to Cauchy s formula for holomorphic functions. First, let us introduce some notation for functions of a complex variable $f(z) = f(x + iy) = u(x, y) + iv(x, y)$. If u and v are continuously differentiable, define

$$\frac{f}{\bar{z}} = \frac{1}{2} \quad \frac{f}{x} + i\frac{f}{y} \quad = \frac{1}{2} \quad \frac{u}{x} \quad \frac{v}{y} \quad + i \quad \frac{u}{y} + \frac{v}{x} \quad .$$

The function f is called holomorphic if $\frac{f}{\bar{z}} = 0$. Examples are polynomials in z and any absolutely convergent power series in z. Also, let $dz = dx + idy$.

Cauchy s theorem. *Let be an oriented curve that bounds a region on its left, and let* a *. If* $f(z) = f(x + iy) = u(x, y) + iv(x, y)$ *is a holomorphic function on such that* u *and* v *are continuous on and continuously differentiable on , then*

$$f(a) = \frac{1}{2\ \ i} \overset{5}{\ } \quad \frac{f(z)}{z \quad a} dz.$$

Proof. The proof is based on Green s formula, applied on the domain obtained from by removing a disk of radius around a as described in Figure 23 to $P = F$ and $Q = iF$, where F is a holomorphic function to be specified later. Note that the boundary of the domain consists of two curves, and .

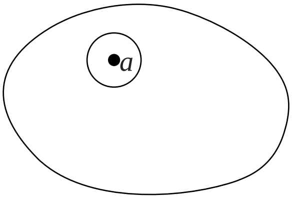

Figure 23

Green s formula reads

$$\overset{5}{\ } \quad F dz \quad \overset{5}{\ } \quad F dz = \overset{5}{\ } \quad F dx + iF dy \quad \overset{5}{\ } \quad F dx + iF dy$$

$$= \quad i \quad \frac{F}{x} + i\frac{F}{y} \quad dxdy = 2i \quad \frac{F}{\bar{z}} dxdy = 0.$$

Therefore,

$$\overset{5}{\ } \quad F(z)dz = \overset{5}{\ } \quad F(z)dz.$$

We apply this to

$$F(z) = \frac{f(z)}{z \quad a} = \frac{(u(x,y) + iv(x,y))(x \quad iy + \qquad i\ )}{(x + \ \ )^2 + (y + \ \ )^2},$$

where $a = \quad + i\ $. It is routine to check that F is holomorphic. We thus have

$$5\ \frac{f(z)}{z \quad a}dz = 5\ \frac{f(z)}{z \quad a}dz.$$

The change of variable $z = a + \ \ e^{it}$ on the right-hand side yields

$$5\ \frac{f(z)}{z \quad a}dz = \frac{f(a + \ \ e^{it})}{e^{it}} i\ \ e^{it}dt = i \qquad f(a + \ \ e^{it})dt.$$

When $\qquad 0$ this tends to $2\ \ if(a)$, and we obtain

$$5\ \frac{f(z)}{z \quad a}dz = 2\ \ if(a).$$

Hence the desired formula.

526. Assume that a curve $(x(t), y(t))$ runs counterclockwise around a region D. Prove that the area of D is given by the formula

$$A = \frac{1}{2}\ 5_{D}\ (xy \quad yx\ )dt.$$

527. Compute the flux of the vector field

$$F(x,y,z) = x(e^{xy} \quad e^{zx})\ i \ + y(e^{yz} \quad e^{xy})\ j \ + z(e^{zx} \quad e^{yz})\ k$$

across the upper hemisphere of the unit sphere.

528. Compute

$$5_{C}\ y^2dx + z^2dy + x^2dz,$$

where C is the Viviani curve, defined as the intersection of the sphere $x^2 + y^2 + z^2 = a^2$ with the cylinder $x^2 + y^2 = ax$.

529. Let $\ (x,y,z)$ and $\ (x,y,z)$ be twice continuously differentiable functions in the region $\{(x,y,z) \mid \frac{1}{2} < \overline{\ \ x^2 + y^2 + z^2} < 2\}$. Prove that

$$_{S}(\qquad \times \qquad)\cdot\ n\, dS = 0,$$

where S is the unit sphere centered at the origin, n is the normal unit vector to this sphere, and $\qquad$ denotes the gradient $\frac{\ }{x}\mathbf{i} + \frac{\ }{y}\mathbf{j} + \frac{\ }{z}\mathbf{k}$.

530. Let $f, g : \mathbb{R}^3 \to \mathbb{R}$ be twice continuously differentiable functions that are constant along the lines that pass through the origin. Prove that on the unit ball $B = \{(x, y, z) \mid x^2 + y^2 + z^2 \leq 1\}$,

$$\iiint_B f \nabla^2 g \, dV = \iiint_B g \nabla^2 f \, dV.$$

Here $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$ is the Laplacian.

531. Prove Gauss' law, which states that the total flux of the gravitational field through a closed surface equals $-4\pi G$ times the mass enclosed by the surface, where G is the constant of gravitation. The mathematical formulation of the law is

$$\iint_S \mathbf{F} \cdot \mathbf{n} \, dS = -4\pi MG.$$

532. Let

$$G(x, y) = \left(-\frac{y}{x^2 + 4y^2}, \frac{x}{x^2 + 4y^2}, 0 \right).$$

Prove or disprove that there is a vector field $F : \mathbb{R}^3 \to \mathbb{R}^3$,

$$F(x, y, z) = (M(x, y, z), N(x, y, z), P(x, y, z)),$$

with the following properties:

(i) M, N, P have continuous partial derivatives for all $(x, y, z) \neq (0, 0, 0)$;
(ii) curl $F = \mathbf{0}$, for all $(x, y, z) \neq (0, 0, 0)$;
(iii) $F(x, y, 0) = G(x, y)$.

533. Let $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$, $\mathbf{F}(x, y) = (F_1(x, y), F_2(x, y))$ be a vector field, and let $G : \mathbb{R}^3 \to \mathbb{R}$ be a smooth function whose first two variables are x and y, and the third is t, the time. Assume that for any rectangular surface D bounded by the curve C,

$$\frac{d}{dt} \iint_D G(x, y, t) dx dy = \oint_C \mathbf{F} \cdot d\mathbf{R}.$$

Prove that

$$\frac{\partial G}{\partial t} + \frac{\partial F_2}{\partial x} + \frac{\partial F_1}{\partial y} = 0.$$

534. For two disjoint oriented curves C_1 and C_2 in three-dimensional space, parametrized by $\vec{v}_1(s)$ and $\vec{v}_2(t)$, define the linking number

$$\mathrm{lk}(C_1, C_2) = \frac{1}{4\pi} \oint_{C_1} \oint_{C_2} \frac{\vec{v}_1 - \vec{v}_2}{\|\vec{v}_1 - \vec{v}_2\|^3} \cdot \left(\frac{d\vec{v}_1}{ds} \times \frac{d\vec{v}_2}{dt} \right) dt\,ds.$$

Prove that if the oriented curves C_1 and $-C_1'$ bound an oriented surface S such that S is to the left of each curve, and if the curve C_2 is disjoint from S, then $\mathrm{lk}(C_1, C_2) = \mathrm{lk}(C_1', C_2)$.

3.4 Equations with Functions as Unknowns

3.4.1 Functional Equations

We will now look at equations whose unknowns are functions. Here is a standard example that we found in B.J. Venkatachala, *Functional Equations: A Problem Solving Approach* (Prism Books PVT Ltd., 2002).

Example. Find all functions $f : \mathbb{R} \to \mathbb{R}$ satisfying the functional equation

$$f((x - y)^2) = f(x)^2 - 2xf(y) + y^2.$$

Solution. For $y = 0$, we obtain

$$f(x^2) = f(x)^2 - 2xf(0),$$

and for $x = 0$, we obtain

$$f(y^2) = f(0)^2 + y^2.$$

Setting $y = 0$ in the second equation, we find that $f(0) = 0$ or $f(0) = 1$. On the other hand, combining the two equalities, we obtain

$$f(x)^2 - 2xf(0) = f(0)^2 + x^2,$$

that is,

$$f(x)^2 = (x + f(0))^2.$$

Substituting this in the original equation yields

$$\begin{aligned} f(y) &= \frac{f(x)^2 - f((x - y)^2) + y^2}{2x} = \frac{(x + f(0))^2 - (x - y + f(0))^2 + y^2}{2x} \\ &= y + f(0). \end{aligned}$$

We conclude that the functional equation has the two solutions $f(x) = x$ and $f(x) = x + 1$.

(Gh. Bucur, E. Câmpu, S. Gaina, *Culegere de Probleme de Calcul Diferenţial şi Integral* (*Collection of Problems in Differential and Integral Calculus*), Editura Tehnica, Bucharest, 1967)

518. The method is similar to that for computing the Fresnel integrals, only simpler. If we denote the integral by I, then

$$I^2 = \quad e^{\ x^2}dx \quad e^{\ y^2}dy = \quad e^{\ (x^2+y^2)}dxdy.$$

Switching to polar coordinates, we obtain

$$I^2 = {}^{\ 2}_{0}\ {}_{0}\ e^{\ r^2}rdrd \ = {}^{\ 2}_{0}\ \frac{1}{2}\ e^{\ r^2}\ {}_{0}\ d \ = {}^{\ 2}_{0}\ \frac{1}{2}d \ = \ .$$

Hence the desired formula $I = \ \overline{}$.

519. Call the integral I. By symmetry, we may compute it over the domain $\{(u,v,w) \quad R^3 \mid 0 \quad v \quad u \quad 1\}$, then double the result. We substitute $u = r\cos \ $, $v = r\sin \ $, $w = \tan \ $, taking into account that the limits of integration become 0 , $\frac{}{4}$, and 0 r sec . We have

$$\begin{aligned} I &= 2\ {}^{\frac{}{4}}_{0}\ {}^{\frac{}{4}}_{0}\ {}^{\sec}_{0}\ \frac{r\sec^2}{(1 + r^2\cos^2 \ + r^2\sin^2 \ + \tan^2 \ )^2}drd \ d \\ &= 2\ {}^{\frac{}{4}}_{0}\ {}^{\frac{}{4}}_{0}\ {}^{\sec}_{0}\ \frac{r\sec^2}{(r^2 + \sec^2 \ )^2}drd \ d \\ &= 2\ {}^{\frac{}{4}}_{0}\ {}^{\frac{}{4}}_{0}\ \sec^2 \ \ \frac{1}{2(r^2 + \sec^2 \ )}\Bigg|^{r=\sec}_{r=0} d \ d \\ &= \ {}^{\frac{}{4}}_{0}\ {}^{\frac{}{4}}_{0}\ \frac{\sec^2}{\sec^2 \ + \sec^2}d \ d \ + \ \frac{}{4}^{\ 2}. \end{aligned}$$

But notice that this is the same as

$${}^{\frac{}{4}}_{0}\ {}^{\frac{}{4}}_{0}\ 1 \quad \frac{\sec^2}{\sec^2 \ + \sec^2} \ d \ d \ = {}^{\frac{}{4}}_{0}\ {}^{\frac{}{4}}_{0}\ \frac{\sec^2}{\sec^2 \ + \sec^2}d \ d \ .$$

If we exchange the roles of and in this last integral we see that

$${}^{\frac{}{4}}_{0}\ {}^{\frac{}{4}}_{0}\ \frac{\sec^2}{\sec^2 \ + \sec^2}d \ d \ + \ \frac{}{4}^{\ 2} = {}^{\frac{}{4}}_{0}\ {}^{\frac{}{4}}_{0}\ \frac{\sec^2}{\sec^2 \ + \sec^2}d \ d \ .$$

Hence

$${}^{\frac{}{4}}_{0}\ {}^{\frac{}{4}}_{0}\ \frac{\sec^2}{\sec^2 \ + \sec^2}d \ d \ = \frac{\ ^2}{32}.$$

Consequently, the integral we are computing is equal to $\frac{\pi^2}{32}$.

(*American Mathematical Monthly*, proposed by M. Hajja and P. Walker)

520. We have

$$\begin{aligned} I &= \iint_D \ln|\sin(x-y)|\,dx\,dy = \int_0^{\pi}\left(\int_0^{y}\ln|\sin(y-x)|\,dx\right)dy \\ &= \int_0^{\pi}\left(\int_0^{y}\ln\sin t\,dt\right)dy = y\int_0^{y}\ln\sin t\,dt\,\Big|_{y=0}^{y=\pi} - \int_0^{\pi} y\ln\sin y\,dy \\ &= \pi A - B, \end{aligned}$$

where $A = \int_0^{\pi}\ln\sin t\,dt$, $B = \int_0^{\pi} t\ln\sin t\,dt$. Note that here we used integration by parts! We compute further

$$\begin{aligned} A &= \int_0^{\frac{\pi}{2}}\ln\sin t\,dt + \int_{\frac{\pi}{2}}^{\pi}\ln\sin t\,dt = \int_0^{\frac{\pi}{2}}\ln\sin t\,dt + \int_0^{\frac{\pi}{2}}\ln\cos t\,dt \\ &= \int_0^{\frac{\pi}{2}}(\ln\sin 2t - \ln 2)\,dt = -\frac{\pi}{2}\ln 2 + \frac{1}{2}A. \end{aligned}$$

Hence $A = -\pi\ln 2$. For B we use the substitution $t = \pi - x$ to obtain

$$B = \int_0^{\pi}(\pi - x)\ln\sin x\,dx = \pi A - B.$$

Hence $B = \frac{\pi}{2}A$. Therefore, $I = \pi A - B = -\frac{\pi^2}{2}\ln 2$, and we are done.

Remark. The identity

$$\int_0^{\frac{\pi}{2}}\ln\sin t\,dt = -\frac{\pi}{2}\ln 2$$

belongs to Euler.

(S. Radulescu, M. Radulescu, *Teoreme și Probleme de Analiză Matematică* (*Theorems and Problems in Mathematical Analysis*), Editura Didactica și Pedagogica, Bucharest, 1982).

521. This problem applies the discrete version of Fubini's theorem. Define

$$f(i,j) = \begin{cases} 1 & \text{for } j \le a_i, \\ 0 & \text{for } j > a_i. \end{cases}$$

The left-hand side is equal to $\sum_{i=1}^{n}\sum_{j=1}^{m} f(i,j)$, while the right-hand side is equal to $\sum_{j=1}^{m}\sum_{i=1}^{n} f(i,j)$. The equality follows.

522. First, note that for $x > 0$,

$$e^{-sx}x^{-1}|\sin x| < e^{-sx},$$

so the integral that we are computing is finite.

Now consider the two-variable function

$$f(x,y) = e^{-sxy}\sin x.$$

We have

$$\int_0^\infty\int_1^\infty |f(x,y)|dydx = \int_0^\infty\int_1^\infty e^{-sxy}|\sin x|dydx = \frac{1}{s}\int_0^\infty e^{-sx}x^{-1}|\sin x|dx,$$

and we just saw that this is finite. Hence we can apply Fubini's theorem, to conclude that on the one hand,

$$\int_0^\infty\int_1^\infty f(x,y)dydx = \frac{1}{s}\int_0^\infty e^{-sx}x^{-1}\sin x dx,$$

and on the other hand,

$$\int_0^\infty\int_1^\infty f(x,y)dydx = \int_1^\infty \frac{1}{s^2y^2+1}dy.$$

Here of course we used the fact that

$$\int_0^\infty e^{-ax}\sin x dx = \frac{1}{a^2+1}, \quad a > 0,$$

a formula that can be proved by integrating by parts. Equating the two expressions that we obtained for the double integral, we obtain

$$\int_0^\infty e^{-sx}x^{-1}\sin x dx = \frac{\pi}{2} - \arctan s = \arctan(s^{-1}),$$

as desired.

(G.B. Folland, *Real Analysis, Modern Techniques and Their Applications*, Wiley, 1999)

523. Applying Tonelli's theorem to the function $f(x,y) = e^{-xy}$, we can write

$$\begin{aligned}\int_0^\infty \frac{e^{-ax}-e^{-bx}}{x}dx &= \int_0^\infty\int_a^b e^{-xy}dydx = \int_a^b\int_0^\infty e^{-xy}dxdy\\ &= \int_a^b \frac{1}{y}dy = \ln\frac{b}{a}.\end{aligned}$$

Remark. This is a particular case of integrals of the form $\int_0^{+\infty}\frac{f(ax)-f(bx)}{x}dx$, known as Froullani integrals. In general, if f is continuous and has finite limit at infinity, the value of the integral is $(f(0)-\lim_{x\to\infty} f(x))\ln\frac{b}{a}$.

524. We do the proof in the case $0<x<1$, since for $-1<x<0$ the proof is completely analogous, while for $x=0$ the property is obvious. The function $f:\mathbb{N}\times[0,x]\to\mathbb{R}$, $f(n,t)=t^{n-1}$ satisfies the hypothesis of Fubini's theorem. So integration commutes with summation:

$$\sum_{n=0}^{\infty}\int_0^x t^{n-1}dt=\int_0^x\frac{dt}{1-t}.$$

This implies

$$\sum_{n=1}^{\infty}\frac{x^n}{n}=-\ln(1-x).$$

Dividing by x, we obtain

$$\sum_{n=1}^{\infty}\frac{x^{n-1}}{n}=-\frac{1}{x}\ln(1-x).$$

The right-hand side extends continuously at 0, since $\lim_{x\to 0}\frac{1}{t}\ln(1-t)=-1$. Again we can apply Fubini's theorem to $f(n,t)=\frac{t^{n-1}}{n}$ on $\mathbb{N}\times[0,x]$ to obtain

$$\sum_{n=1}^{\infty}\frac{x^n}{n^2}=\sum_{n=1}^{\infty}\int_0^x\frac{t^{n-1}}{n}dt=\int_0^x\sum_{n=1}^{\infty}\frac{t^{n-1}}{n}dt=\int_0^x-\frac{1}{t}\ln(1-t)dt,$$

as desired.

525. We can apply Tonelli's theorem to the function $f(x,n)=\frac{1}{x^2+n^4}$. Integrating term by term, we obtain

$$\int_0^x F(t)dt=\int_0^x\sum_{n=1}^{\infty}f(t,n)dt=\sum_{n=1}^{\infty}\int_0^x\frac{dt}{t^2+n^4}=\sum_{n=1}^{\infty}\frac{1}{n^2}\arctan\frac{x}{n^2}.$$

This series is bounded from above by $\sum_{n=1}^{\infty}\frac{1}{n^2}=\frac{\pi^2}{6}$. Hence the summation commutes with the limit as x tends to infinity. We have

$$\int_0^{\infty}F(t)dt=\lim_{x\to\infty}\int_0^x F(t)dt=\lim_{x\to\infty}\sum_{n=1}^{\infty}\frac{1}{n^2}\arctan\frac{x}{n^2}=\sum_{n=1}^{\infty}\frac{1}{n^2}\cdot\frac{\pi}{2}.$$

Using the identity $_{n \quad 1}\frac{1}{n^2} = \frac{\ ^2}{6}$, we obtain

$$_{0}\ F(t)dt = \frac{\ ^3}{12}.$$

(Gh. Sireţchi, *Calcul Diferenţial şi Integral* (*Differential and Integral Calculus*), Editura Ştiinţifica şi Enciclopedica, Bucharest, 1985)

526. The integral from the statement can be written as

$$^{5}_{D}\ xdy \quad ydx.$$

Applying Green s theorem for $P(x,y) = \quad y$ and $Q(x,y) = x$, we obtain

$$^{5}_{D}\ xdy \quad ydx = \quad _{D}(1+1)dxdy,$$

which is twice the area of D. The conclusion follows.

527. It can be checked that div $F = 0$ (in fact, F is the curl of the vector field $e^{yz}\ i + e^{zx}\ j + e^{xy}\ k$). If S be the union of the upper hemisphere and the unit disk in the xy-plane, then by the divergence theorem $\overset{++}{}_{S}\ F \bullet n\, dS = 0$. And on the unit disk $F \bullet n = 0$, which means that the flux across the unit disk is zero. It follows that the flux across the upper hemisphere is zero as well.

528. We simplify the computation using Stokes theorem:

$$^{5}_{C}\ y^2dx + z^2dy + x^2dz = \quad 2 \quad _{S}\ ydxdy + zdydz + xdzdx,$$

where S is the portion of the sphere bounded by the Viviani curve. We have

$$2 \quad _{S}\ ydxdy + zdydz + xdzdx = \quad 2 \quad _{S}(z,x,y) \bullet n\, d \quad ,$$

where (z,x,y) denotes the three-dimensional vector with coordinates z, x, and y, while n denotes the unit vector normal to the sphere at the point of coordinates (x,y,z). We parametrize the portion of the sphere in question by the coordinates (x,y), which range inside the circle $x^2 + y^2 \quad ax = 0$. This circle is the projection of the Viviani curve onto the xy-plane.

The unit vector normal to the sphere is

$$n = \frac{x}{a}, \frac{y}{a}, \frac{z}{a} = \frac{x}{a}, \frac{y}{a}, \frac{\overline{a^2 \quad x^2 \quad y^2}}{a} \quad ,$$

while the area element is

$$d\ \ = \frac{1}{\cos\ }dxdy,$$

being the angle formed by the normal to the sphere with the xy-plane. It is easy to see that $\cos\ \ = \frac{z}{a} = \frac{\overline{a^2\ \ x^2\ \ y^2}}{a}$. Hence the integral is equal to

$$2\ \ _D\ z\frac{x}{a} + x\frac{y}{a} + y\frac{z}{a}\ \ \frac{a}{z}dxdy = \ 2\ \ _D\ x + y + \frac{xy}{\overline{a^2\ \ x^2\ \ y^2}}\ \ dxdy,$$

the domain of integration D being the disk $x^2 + y^2\ \ \ ax\ \ \ 0$. Split the integral as

$$2\ \ _D(x + y)dxdy\ \ \ 2\ \ _D\frac{xy}{\overline{a^2\ \ x^2\ \ y^2}}dxdy.$$

Because the domain of integration is symmetric with respect to the y-axis, the second double integral is zero. The first double integral can be computed using polar coordinates: $x = \frac{a}{2} + r\cos\ $, $y = r\sin\ $, $0\ \ \ r\ \ \ \frac{a}{2}$, $0\ \ \ \ \ 2\ $. Its value is $\ \frac{a^3}{4}$, which is the answer to the problem.

(D. Flondor, N. Donciu, *Algebră și Analiză Matematică* (*Algebra and Mathematical Analysis*), Editura Didactica și Pedagogica, Bucharest, 1965)

529. We will apply Stokes' theorem. We begin with

$$\frac{}{y}\frac{}{z}\ \ \ \ \frac{}{z}\frac{}{y} = \frac{}{y}\frac{}{z} + \frac{^2}{y\ z}\ \ \ \ \frac{}{z}\frac{}{y}\ \ \ \ \frac{^2}{z\ y}$$

$$= \frac{}{y}\ \ \ \frac{}{z}\ \ \ \ \frac{}{z}\ \ \ \frac{}{y}\ \ ,$$

which combined with the two other analogous computations gives

$$\times\ \ \ = \mathrm{curl}(\ \ \ ).$$

By Stokes' theorem, the integral of the curl of a vector field on a surface without boundary is zero.

(Soviet University Student Mathematical Competition, 1976)

530. For the solution, recall the following identity.

Green's first identity. *If* f *and* g *are twice-differentiable functions on the solid region* R *bounded by the closed surface* S*, then*

$$_R(f\ \ ^2g + \ \ f\bullet\ \ g)dV = \ \ _S\ f\frac{g}{n}dS,$$

where $\frac{g}{n}$ *is the derivative of* g *in the direction of the normal to the surface.*

Proof. For the sake of completeness we will prove Green's identity. Consider the vector field $F = f\nabla g$. Then

$$\begin{aligned}\operatorname{div} F &= \frac{\partial}{\partial x}\left(f\frac{\partial g}{\partial x}\right) + \frac{\partial}{\partial y}\left(f\frac{\partial g}{\partial y}\right) + \frac{\partial}{\partial z}\left(f\frac{\partial g}{\partial z}\right)\\ &= f\left(\frac{\partial^2 g}{\partial x^2} + \frac{\partial^2 g}{\partial y^2} + \frac{\partial^2 g}{\partial z^2}\right) + \left(\frac{\partial f}{\partial x}\frac{\partial g}{\partial x} + \frac{\partial f}{\partial y}\frac{\partial g}{\partial y} + \frac{\partial f}{\partial z}\frac{\partial g}{\partial z}\right).\end{aligned}$$

So the left-hand side is $\iiint_R \operatorname{div} F\, dV$. By the Gauss–Ostrogradski divergence theorem this is equal to

$$\iint_S (f\nabla g)\cdot n\, dS = \iint_S f(\nabla g \cdot n)dS = \iint_S f\frac{\partial g}{\partial n}dS.$$

Writing Green's first identity for the vector field $g\nabla f$ and then subtracting it from that of the vector field $f\nabla g$, we obtain Green's second identity

$$\iiint_R (f\nabla^2 g - g\nabla^2 f)dV = \iint_S \left(f\frac{\partial g}{\partial n} - g\frac{\partial f}{\partial n}\right)dS.$$

The fact that f and g are constant along the lines passing through the origin means that on the unit sphere, $\frac{\partial f}{\partial n} = \frac{\partial g}{\partial n} = 0$. Hence the conclusion.

531. Because F is obtained as an integral of the point-mass contributions of the masses distributed in space, it suffices to prove this equality for a mass M concentrated at one point, say the origin.

Newton's law says that the gravitational force between two masses m_1 and m_2 at distance r is equal to $\frac{m_1 m_2 G}{r^2}$. By Newton's law, a mass M located at the origin generates the gravitational field

$$F(x, y, z) = -MG\frac{1}{x^2+y^2+z^2}\cdot\frac{x\,i + y\,j + z\,k}{\sqrt{x^2+y^2+z^2}} = -MG\frac{x\,i + y\,j + z\,k}{(x^2+y^2+z^2)^{3/2}}.$$

One can easily check that the divergence of this field is zero. Consider a small sphere S_0 of radius r centered at the origin, and let V be the solid lying between S_0 and S. By the Gauss–Ostrogradski divergence theorem,

$$\iint_S F\cdot n\, dS - \iint_{S_0} F\cdot n\, dS = \iiint_V \operatorname{div} F\, dV = 0.$$

Hence it suffices to prove the Gauss law for the sphere S_0. On this sphere the flow $F\cdot n$ is constantly equal to $-\frac{GM}{r^2}$. Integrating it over the sphere gives $-4\pi MG$, proving the law.

532. The condition curl $\mathbf{F} = 0$ suggests the use of Stokes' theorem:

$$\iint_S \text{curl}\,\mathbf{F}\cdot\mathbf{n}\,dS = \oint_C \mathbf{F}\cdot d\mathbf{R}.$$

We expect the answer to the question to be no. All we need is to find a surface S whose boundary lies in the xy-plane and such that the integral of $\mathbf{G}(x, y)$ on ∂S is nonzero.

A simple example that comes to mind is the interior S of the ellipse $x^2 + 4y^2 = 4$. Parametrize the ellipse as $x = 2\cos\theta$, $y = \sin\theta$, $\theta \in [0, 2\pi)$. Then

$$\oint_{\partial S} \mathbf{G}\cdot d\mathbf{R} = \int_0^{2\pi}\left(-\frac{\sin\theta}{4}, \frac{2\cos\theta}{4}, 0\right)\cdot(-2\sin\theta, \cos\theta, 0)d\theta = \int_0^{2\pi}\frac{1}{2}d\theta = \pi.$$

By Stokes' theorem this should be equal to the integral of the curl of $\mathbf{F}$ over the interior of the ellipse. The curl of $\mathbf{F}$ is zero except at the origin, but we can fix that by adding a smooth tiny upward bump at the origin, which does not alter too much the above computation. The integral should on the one hand be close to 0, and on the other hand close to π, which is impossible. This proves that such a vector field $\mathbf{F}$ cannot exist.

(48th W.L. Putnam Mathematical Competition, 1987, solution from K. Kedlaya, B. Poonen, R. Vakil, *The William Lowell Putnam Mathematical Competition* 1985–2000, MAA, 2002)

533. Let $D = [a_1, b_1] \times [a_2, b_2]$ be a rectangle in the plane, and $a, b \in \mathbb{R}$, $a < b$. We consider the three-dimensional parallelepiped $V = D \times [a, b]$. Denote by $\mathbf{n}$ the outward normal vector field on the boundary ∂V of V (which is defined everywhere except on the edges). By the Leibniz–Newton fundamental theorem of calculus,

$$\begin{aligned}\int_a^b \frac{d}{dt}\iint_D G(x,y,t)dxdydt &= \int_a^b\iint_D \frac{\partial}{\partial t}G(x,y,t)dxdydt\\ &= \iint_D\int_a^b \frac{\partial}{\partial t}G(x,y,t)dtdxdy\\ &= \iint_D G(x,y,b)dxdy - \iint_D G(x,y,a)dxdy\\ &= \iint_{D\times\{b\}} G(x,y,t)\,\mathbf{k}\cdot d\mathbf{n} + \iint_{D\times\{a\}} G(x,y,t)\,\mathbf{k}\cdot d\mathbf{n},\end{aligned}$$

where $\mathbf{k}$ denotes the unit vector that points in the z-direction. With this in mind, we compute

$$0 = \int_a^b\left(\frac{d}{dt}\iint_D G(x,y,t)dxdy + \oint_C \mathbf{F}\cdot d\mathbf{R}\right)dt$$

$$= {}_{D\times\{b\}} G(x,y,t)\,k\cdot d\,n + {}_{D\times\{a\}} G(x,y,t)\,k\cdot d\,n$$

$$+ {}_a^b {}_{a_1}^{b_1} F_1(x,a_2)dx \qquad {}_a^b {}_{b_1}^{a_1} F_1(x,b_2)dx$$

$$+ {}_a^b {}_{a_2}^{b_2} F_2(b_1,y)dy \qquad {}_a^b {}_{b_2}^{a_2} F_2(a_1,y)dy.$$

If we introduce the vector field $H = F_2\, i + F_1\, j + G\, k$, this equation can be written simply as

$${}_V H \cdot n\, dS = 0.$$

By the divergence theorem,

$${}_V \operatorname{div} H\, dV = {}_V H \cdot n\, dS = 0.$$

Since this happens in every parallelepiped, div H must be identically equal to 0. Therefore,

$$\operatorname{div} H = \frac{F_2}{x} + \frac{F_1}{y} + \frac{G}{t} = 0,$$

and the relation is proved.

Remark. The interesting case occurs when F and G depend on spatial variables (spatial dimensions). Then G becomes a vector field B, or better a 2-form, called the magnetic flux, while F becomes the electric field strength E. The relation

$$\frac{d}{dt}\,{}_S B = {}_S E$$

is Faraday's law of induction. Introducing a fourth dimension (the time), and redoing mutatis mutandis the above computation gives rise to the first group of Maxwell's equations

$$\operatorname{div} B = 0, \qquad \frac{B}{t} = \operatorname{curl} E.$$

534. In the solution we ignore the factor $\frac{1}{4}$, which is there only to make the linking number an integer. We will use the more general form of Green's theorem applied to the curve $C = C_1 \quad C_1$ and surface S,

$$5_{C}\, P\,dx + Q dy + R dz = {}_{S}\left(\frac{Q}{x}\quad\frac{P}{y}\right) dx dy + \left(\frac{R}{y}\quad\frac{Q}{z}\right) dy dz$$

$$+\left(\frac{P}{z}\quad\frac{R}{x}\right) dz dx.$$

Writing the parametrization with coordinate functions $v_1(s) = (x(s), y(s), z(s))$, $v_2(t) = (x(t), y(t), z(t))$, the linking number of C_1 and C_2 (with the factor $\frac{1}{4}$ ignored) becomes

$$5_{C_1}\,5_{C_2} \frac{(x \quad x)(dz\, dy \quad dy\, dz) + (y \quad y)(dx\, dz \quad dz\, dx) + (z \quad z)(dy\, dx \quad dx\, dy)}{((x \quad x)^2 + (y \quad y)^2 + (z \quad z)^2)^{3/2}}$$

The 1-form $P dx + Q dy + R dz$, which we integrate on $C = C_1 \quad C_1$, is

$$5_{C_2} \frac{(x \quad x)(dz\, dy \quad dy\, dz) + (y \quad y)(dx\, dz \quad dz\, dx) + (z \quad z)(dy\, dx \quad dx\, dy)}{((x \quad x)^2 + (y \quad y)^2 + (z \quad z)^2)^{3/2}}.$$

Note that here we integrate against the variables x , y , z , so this expression depends only on x, y, and z. Explicitly,

$$P(x, y, z) = 5_{C_2} \frac{(y \quad y)dz + (z \quad z)dy}{((x \quad x)^2 + (y \quad y)^2 + (z \quad z)^2)^{3/2}},$$

$$Q(x, y, z) = 5_{C_2} \frac{(x \quad x)dz \quad (z \quad z)dx}{((x \quad x)^2 + (y \quad y)^2 + (z \quad z)^2)^{3/2}},$$

$$R(x, y, z) = 5_{C_2} \frac{(x \quad x)dy + (y \quad y)dx}{((x \quad x)^2 + (y \quad y)^2 + (z \quad z)^2)^{3/2}}.$$

By the general form of Green s theorem, lk$(C_1, C_2) = $ lk(C_1, C_2) if

$$\frac{Q}{x} \quad \frac{P}{y} = \frac{R}{y} \quad \frac{Q}{z} = \frac{P}{z} \quad \frac{R}{x} = 0.$$

We will verify only $\frac{Q}{x} \quad \frac{P}{y} = 0$, the other equalities having similar proofs. The part of it that contains dz is equal to

$$5_{C_2} \quad 2((x \quad x)^2 + (y \quad y)^2 + (z \quad z)^2) \;^{3/2}$$

$$+ 3(x \quad x)^2((x \quad x)^2 + (y \quad y)^2 + (z \quad z)^2) \;^{5/2}$$

$$+ 3(y \quad y)^2((x \quad x)^2 + (y \quad y)^2 + (z \quad z)^2) \;^{5/2} dz$$

$$= 5_{C_2} \quad ((x \quad x)^2 + (y \quad y)^2 + (z \quad z)^2) \;^{3/2}$$

$$+ 3(z \quad z)^2((x \quad x)^2 + (y \quad y)^2 + (z \quad z)^2) \;^{5/2} dz$$

$$= \overset{5}{}_{C_2} \frac{}{z}((x \quad x)^2 + (y \quad y)^2 + (z \quad z)^2)^{\quad 3/2} dz = 0,$$

where the last equality is a consequence of the fundamental theorem of calculus. Of the two, only $\frac{Q}{x}$ has a dx in it, and that part is

$$3 \overset{5}{}_{C_2} ((x \quad x\ )^2 + (y \quad y\ )^2 + (z \quad z\ )^2)^{\quad 5/2}(x \quad x\ )(z \quad z\ )dx$$

$$= \overset{5}{}_{C_2} \frac{}{x} \frac{z \quad z}{((x \quad x\ )^2 + (y \quad y\ )^2 + (z \quad z\ )^2)^{3/2}} dx = 0.$$

The term involving dy is treated similarly. The conclusion follows.

Remark. The linking number is, in fact, an integer, which measures the number of times the curves wind around each other. It was defined by C.F. Gauss, who used it to decide, based on astronomical observations, whether the orbits of certain asteroids were winding around the orbit of the earth.

535. Plugging in $x = y$, we find that $f(0) = 0$, and plugging in $x = \quad 1$, $y = 0$, we find that $f(1) = \quad f(\quad 1)$. Also, plugging in $x = a$, $y = 1$, and then $x = a$, $y = \quad 1$, we obtain

$$f(a^2 \quad 1) = (a \quad 1)(f(a) + f(1)),$$
$$f(a^2 \quad 1) = (a + 1)(f(a) \quad f(1)).$$

Equating the right-hand sides and solving for $f(a)$ gives $f(a) = f(1)a$ for all a.

So any such function is linear. Conversely, a function of the form $f(x) = kx$ clearly satisfies the equation.

(Korean Mathematical Olympiad, 2000)

536. Replace z by $1 \quad z$ to obtain

$$f(1 \quad z) + (1 \quad z)f(z) = 2 \quad z.$$

Combine this with $f(z) + zf(1 \quad z) = 1 + z$, and eliminate $f(1 \quad z)$ to obtain

$$(1 \quad z + z^2)f(z) = 1 \quad z + z^2.$$

Hence $f(z) = 1$ for all z except maybe for $z = e^{- \ i \ /3}$, when $1 \quad z + z^2 = 0$. For $\quad = e^{i \ /3}$, ﬂ $= \quad ^2 = 1 \quad$; hence $f(\ ) + \quad f($ﬂ$) = 1 + \quad$. We therefore have only one constraint, namely $f($ﬂ$) = [1 + \quad f(\ )]/ \quad =$ ﬂ $+ 1 \quad$ ﬂ $f(\ )$. Hence the solution to the functional equation is of the form

$$f(z) = 1 \quad \text{for } z = e^{-i \ /3},$$

4

Geometry and Trigonometry

Geometry is the oldest of the mathematical sciences. Its age-old theorems and the sharp logic of its proofs make you think of the words of Andrew Wiles, “Mathematics seems to have a permanence that nothing else has.”

This chapter is bound to take you away from the geometry of the ancients, with figures and pictorial intuition, and bring you to the science of numbers and equations that geometry has become today. In a dense exposition we have packed vectors and their applications, analytical geometry in the plane and in space, some applications of integral calculus to geometry, followed by a list of problems with Euclidean flavor but based on algebraic and combinatorial ideas. Special attention is given to conics and quadrics, for their study already contains the germs of differential and algebraic geometry.

Four subsections are devoted to geometry's little sister, trigonometry. We insist on trigonometric identities, repeated in subsequent sections from different perspectives: Euler's formula, trigonometric substitutions, and telescopic summation and multiplication.

Since geometry lies at the foundation of mathematics, its presence could already be felt in the sections on linear algebra and multivariable calculus. It will resurface again in the chapter on combinatorics.

4.1 Geometry

4.1.1 Vectors

This section is about vectors in two and three dimensions. Vectors are oriented segments identified under translation.

There are four operations defined for vectors: scalar multiplication v, addition $v + w$, dot product $v \cdot w$, and cross-product $v \times w$, the last being defined only in three dimensions. Scalar multiplication dilates or contracts a vector by a scalar. The sum of two vectors is computed with the parallelogram rule; it is the resultant of the vectors acting as forces on an object. The dot product of two vectors is a number equal to the product of the magnitudes of the vectors and the cosine of the angle between them. A dot product equal to zero tells us that the vectors are orthogonal. The cross-product of two vectors is a vector orthogonal to the two vectors and of magnitude equal to the area of the parallelogram they generate. The orientation of the cross-product is determined by the right-hand rule: place your hand so that you can bend your palm from the first vector to the second, and your thumb will point in the direction of the cross-product. A cross-product equal to zero tells us that the vectors are parallel (although they might point in opposite directions).

The dot and cross-products are distributive with respect to sum; the dot product is commutative, while the cross-product is not. For the three-dimensional vectors u, v, w, the number $u\cdot(v\times w)$ is the volume taken with sign of the parallelepiped constructed with the vectors as edges. The sign is positive if the three vectors determine a frame that is oriented the same way as the orthogonal frame of the three coordinate axes, and negative otherwise. Equivalently, $u\cdot(v\times w)$ is the determinant with the coordinates of the three vectors as rows.

A useful computational tool is the *cab-bac* identity:

$$a\times(b\times c)=(c\cdot a)b-(b\cdot a)c.$$

The quickest way to prove it is to check it for a, b, c chosen among the three unit vectors parallel to the coordinate axes i, j, and k, and then use the distributivity of the cross-product with respect to addition. Here is an easy application of this identity.

Example. Prove that for any vectors a, b, c, d,

$$(a\times b)\times(c\times d)=(a\cdot(b\times d))c-(a\cdot(b\times c))d.$$

Solution. We have

$$\begin{aligned}(a\times b)\times(c\times d)&=(d\cdot(a\times b))c-(c\cdot(a\times b))d\\&=(a\cdot(b\times d))c-(a\cdot(b\times c))d.\end{aligned}$$

In the computation we used the equality $u\cdot(v\times w)=w\cdot(u\times v)$, which is straightforward if we write these as determinants.

Let us briefly point out a fundamental algebraic property of the cross-product. Denote by so(3) the set of 3 × 3 matrices A satisfying $A+A^t=O_3$ endowed with the operation $[A,B]=AB-BA$.

Theorem. *The map*

$$(a_1,a_2,a_3)\mapsto\begin{pmatrix}0 & a_1 & a_2\\ -a_1 & 0 & a_3\\ -a_2 & -a_3 & 0\end{pmatrix}$$

establishes an isomorphism between $(\mathbb{R}^3, \times)$ *and* $(so(3), [\cdot,\cdot])$.

Proof. The proof is straightforward if we write the cross-product in coordinates. The result shows that the cross-product defines a Lie algebra structure on the set of three-dimensional vectors. Note that the isomorphism maps the sum of vectors to the sum of matrices, and the dot product of two vectors to the negative of half the trace of the product of the corresponding matrices.

And now the problems.

573. For any three-dimensional vectors u, v, w, prove the identity

$$u \times (v \times w) + v \times (w \times u) + w \times (u \times v) = 0.$$

574. Given three vectors a, b, c, define

$$u = (b \cdot c)a \quad (c \cdot a)b,$$
$$v = (a \cdot c)b \quad (a \cdot b)c,$$
$$w = (b \cdot a)c \quad (b \cdot c)a.$$

Prove that if a, b, c form a triangle, then u, v, w also form a triangle, and this triangle is similar to the first.

575. Let a, b, c be vectors such that b and c are perpendicular, but a and b are not. Let m be a real number. Solve the system

$$x \cdot a = m,$$
$$x \times b = c.$$

576. Consider three linearly independent vectors a, b, c in space, having the same origin. Prove that the plane determined by the endpoints of the vectors is perpendicular to the vector $a \times b + b \times c + c \times a$.

577. The vectors a, b, and c satisfy

$$a \times b = b \times c = c \times a = 0.$$

Prove that $a + b + c = 0$.

578. Find the vector-valued functions $u(t)$ satisfying the differential equation

$$u \times u = v,$$

where $v = v(t)$ is a twice-differentiable vector-valued function such that both v and v are never zero or parallel.

579. Does there exist a bijection f of (a) a plane with itself or (b) three-dimensional space with itself such that for any distinct points A, B the lines AB and $f(A)f(B)$ are perpendicular?

580. On so(3) we define the operation such that if A and B are matrices corresponding to the vectors $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3)$, then the ij entry of $A \quad B$ is equal to $(\quad 1)^{i+j} a_{4 \ j} b_{4 \ i}$. Prove the identity

$$CBA \quad BCA = (A \quad C)B \quad (A \quad B)C.$$

581. Prove that there is a bijection f from R^3 to the set su(2) of 2 × 2 matrices with complex entries that are skew symmetric and have trace equal to zero such that

$$f(v \times w) = [f(v), f(w)].$$

We present two applications of vector calculus to geometry, one with the dot product, one with the cross-product.

Example. Given two triangles ABC and $A B C$ such that the perpendiculars from A, B, C onto $B C$, $C A$, $A B$ intersect, show that the perpendiculars from A, B, C onto BC, CA, AB also intersect.

Solution. This is the *property of orthological triangles*. Denote by O the intersection of the first set of three perpendiculars, and by O the intersection of perpendiculars from A and B. Note that if the vector XY is orthogonal to a vector ZW, then for any point P in the plane,

$$(PX \quad PY) \cdot ZW = XY \cdot ZW = 0;$$

hence $PX \cdot ZW = PY \cdot ZW$. Using this fact we can write

$$O C \cdot OB = O A \cdot OB = O A \cdot OC = O B \cdot OC = O B \cdot OA = O C \cdot OA.$$

Therefore, $O C \cdot (OB \quad OA) = O C \cdot AB = 0$, which shows that $O C$ is perpendicular to AB. This proves that the second family of perpendiculars are concurrent.

Example. Let $ABCD$ be a convex quadrilateral, M, N on side AB and P, Q on side CD. Show that if $AM = NB$ and $CP = QD$, and if the quadrilaterals $AMQD$ and $BNPC$ have the same area, then AB is parallel to CD.

Solution. Throughout the solution we refer to Figure 24. We decompose the quadrilaterals into triangles, and then use the formula for the area in terms of the cross-product.

In general, the triangle determined by v_1 and v_2 has area equal to half the magnitude of $v_1 \times v_2$. Note also that $v_1 \times v_2$ is perpendicular to the plane of the triangle, so for

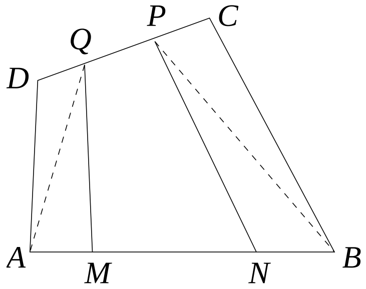

Figure 24

a problem in plane geometry there is no danger in identifying the areas with the cross-products, provided that we keep track of the orientation. The hypothesis of the problem implies that

$$\frac{1}{2}(DA \times DQ + AM \times AQ) = \frac{1}{2}(CP \times CB + BP \times BN).$$

Hence

$$DA \times DQ + AM \times (AD + DQ) = CP \times CB + (BC + CP) \times BN.$$

Because $BN = \quad AM$ and $CP = \quad DQ$, this equality can be rewritten as

$$(AM + DQ) \times (AD + CB) = 2DQ \times AM.$$

Using the fact that $AD + CB = AB + CD$ (which follows from $AB + BC + CD + DA = 0$), we obtain

$$AM \times CD + DQ \times AB = 2DQ \times AM.$$

From here we deduce that $AM \times QC = DQ \times MB$. These two cross-products point in opposite directions, so equality can hold only if both are equal to zero, i.e., if AB is parallel to CD.

More applications of the dot and cross-products to geometry can be found below.

582. Given two triangles ABC and $A\ B\ C$ with the same centroid, prove that one can construct a triangle with sides equal to the segments AA , BB , and CC .

583. Given a quadrilateral $ABCD$, consider the points A , B , C , D on the half-lines (i.e., rays) $|AB$, $|BC$, $|CD$, and $|DA$, respectively, such that $AB = BA$, $BC = CB$, $CD = DC$, $DA = AD$. Suppose now that we start with the quadrilateral $A\ B\ C\ D$. Using a straightedge and a compass only, reconstruct the quadrilateral $ABCD$.

584. On the sides of the triangle ABC construct in the exterior the rectangles ABB_1A_2, BCC_1B_2, CAA_1C_2. Prove that the perpendicular bisectors of A_1A_2, B_1B_2, and C_1C_2 intersect at one point.

585. Let $ABCD$ be a convex quadrilateral. The lines parallel to AD and CD through the orthocenter H of triangle ABC intersect AB and BC, respectively, at P and Q. Prove that the perpendicular through H to the line PQ passes through the orthocenter of triangle ACD.

586. Prove that if the four lines through the centroids of the four faces of a tetrahedron perpendicular to those faces are concurrent, then the four altitudes of the tetrahedron are also concurrent. Prove that the converse is also true.

587. Let ABC be a convex quadrilateral, M, N AB such that $AM = MN = NB$, and P, Q CD such that $CP = PQ = QD$. Let O be the intersection of AC and BD. Prove that the triangles MOP and NOQ have the same area.

588. Let ABC be a triangle, with D and E on the respective sides AC and AB. If M and N are the midpoints of BD and CE, prove that the area of the quadrilateral $BCDE$ is four times the area of the triangle AMN.

4.1.2 The Coordinate Geometry of Lines and Circles

Coordinate geometry was constructed by Descartes to translate Euclid's geometry into the language of algebra. In two dimensions one starts by fixing two intersecting coordinate axes and a unit on each of them. If the axes are perpendicular and the units are equal, the coordinates are called Cartesian (in the honor of Descartes); otherwise, they are called affine. A general affine change of coordinates has the form

$$\begin{matrix}x\\y\end{matrix} = \begin{matrix}a&b\\c&d\end{matrix}\;\begin{matrix}x\\y\end{matrix} + \begin{matrix}e\\f\end{matrix}, \quad \text{with} \quad \begin{matrix}a&b\\c&d\end{matrix} \quad \text{invertible.}$$

If the change is between Cartesian systems of coordinates, a so-called Euclidean change of coordinates, it is required additionally that the matrix

$$\begin{matrix}a&b\\c&d\end{matrix}$$

be orthogonal, meaning that its inverse is equal to the transpose.

Properties that can be formulated in the language of lines and ratios are invariant under affine changes of coordinates. Such are the properties of two lines being parallel or of a point to divide a segment in half. All geometric properties are invariant under Euclidean changes of coordinates. Therefore, problems about distances, circles, and angles should be modeled with Cartesian coordinates.

In this section we grouped problems that require only the knowledge of the theory of lines and circles. Recall that the general equation of a line (whether in a Cartesian or affine coordinate system) is $ax + by + c = 0$. That of a circle (in a Cartesian coordinate system) is $(x - h)^2 + (y - k)^2 = r^2$, where (h, k) is the center and r is the radius. Let us see two examples, one in affine and one in Cartesian coordinates. But before we do that let us recall that a complete quadrilateral is a quadrilateral in which the pairs of opposite sides have been extended until they meet. For that reason, a complete quadrilateral has six vertices and three diagonals.

Example. Prove that the midpoints of the three diagonals of a complete quadrilateral are collinear.

Solution. As said, we will work in affine coordinates. Choose the coordinate axes to be sides of the quadrilateral, as shown in Figure 25.

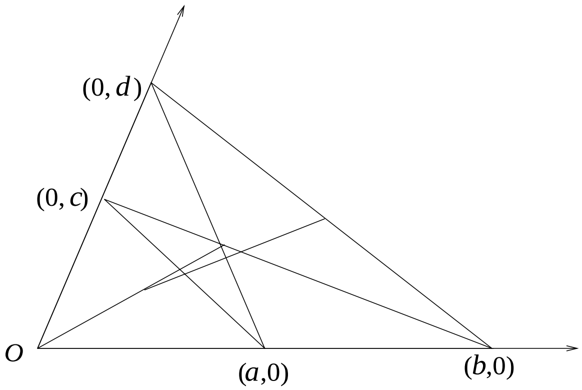

Figure 25

Five of the vertices have coordinates $(0, 0)$, $(a, 0)$, $(b, 0)$, $(0, c)$, and $(0, d)$, while the sixth is found as the intersection of the lines through $(a, 0)$ and $(0, d)$, respectively, $(0, c)$ and $(b, 0)$. For these two lines we know the x- and y-intercepts, so their equations are

$$\frac{1}{a}x + \frac{1}{d}y = 1 \quad \text{and} \quad \frac{1}{b}x + \frac{1}{c}y = 1.$$

The sixth vertex of the complete quadrilateral has therefore the coordinates

$$\left(\frac{ab(c - d)}{ac - bd}, \frac{cd(a - b)}{ac - bd}\right).$$

We find that the midpoints of the diagonals are

$$\left(\frac{a}{2}, \frac{c}{2}\right), \quad \left(\frac{b}{2}, \frac{d}{2}\right), \quad \left(\frac{ab(c - d)}{2(ac - bd)}, \frac{cd(a - b)}{2(ac - bd)}\right).$$

The condition that these three points be collinear translates to

$$\frac{1}{2(ac-bd)}\begin{vmatrix} a & c & 1 \\ b & d & 1 \\ ab(c-d) & cd(a-b) & ac-bd \end{vmatrix} = 0.$$

This is verified by direct computation.

Example. In a circle are inscribed a trapezoid with one side as diameter and a triangle with sides parallel to the sides of the trapezoid. Prove that the two have the same area.

Solution. We refer everything to Figure 26. Assume that the circle has radius 1, and the trapezoid has vertices $(1,0)$, (a,b), $(-a,b)$ and $(-1,0)$.

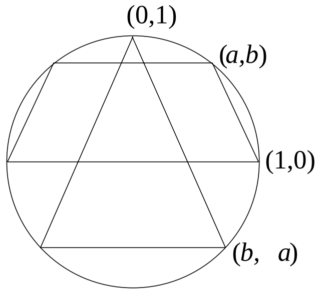

Figure 26

The triangle is isosceles and has one vertex at $(0,1)$. We need to determine the coordinates of the other two vertices. One of them lies where the parallel through $(0,1)$ to the line determined by $(1,0)$ and (a,b) intersects the circle. The equation of the line is

$$y = \frac{b}{a-1}x + 1.$$

The relation $a^2 + b^2 = 1$ yields $b^2 = (1-a)(1+a)$, or $\frac{b}{1-a} = \frac{1+a}{b}$. So the equation of the line can be rewritten as

$$y = -\frac{1+a}{b}x + 1.$$

Now it is easy to guess that the intersection of this line with the circle is $(b,-a)$ (note that this point satisfies the equation of the circle). The other vertex of the triangle is $(-b,-a)$, so the area is $\frac{1}{2}(2b)(1+a) = b + ab$. And the area of the trapezoid is $\frac{1}{2}(2a+2)b = b + ab$, the same number.

589. Prove that the midpoints of the sides of a quadrilateral form a parallelogram.

590. Let M be a point in the plane of triangle ABC. Prove that the centroids of the triangles MAB, MAC, and MCB form a triangle similar to triangle ABC.

591. Find the locus of points P in the interior of a triangle ABC such that the distances from P to the lines AB, BC, and CA are the side lengths of some triangle.

592. Let $A_1, A_2, \dots, A_n$ be distinct points in the plane, and let m be the number of midpoints of all the segments they determine. What is the smallest value that m can have?

593. Given an acute-angled triangle ABC with altitude AD, choose any point M on AD, and then draw BM and extend until it intersects AC in E, and draw CM and extend until it intersects AB in F. Prove that $\ ADE = \ ADF$.

594. In a planar Cartesian system of coordinates consider a fixed point $P(a, b)$ and a variable line through P. Let A be the intersection of the line with the x-axis. Connect A with the midpoint B of the segment OP (O being the origin), and through C, which is the point of intersection of this line with the y-axis, take the parallel to OP. This parallel intersects PA at M. Find the locus of M as the line varies.

595. Let $ABCD$ be a parallelogram with unequal sides. Let E be the foot of the perpendicular from B to AC. The perpendicular through E to BD intersects BC in F and AB in G. Show that $EF = EG$ if and only if $ABCD$ is a rectangle.

596. Find all pairs of real numbers (p, q) such that the inequality

$$\overline{1 \quad x^2} \quad px \quad q \qquad \frac{\overline{2} \quad 1}{2}$$

holds for every $x \quad [0, 1]$.

597. On the hyperbola $xy = 1$ consider four points whose x-coordinates are x_1, x_2, x_3, and x_4. Show that if these points lie on a circle, then $x_1x_2x_3x_4 = 1$.

The points of the plane can be represented as complex numbers. There are two instances in which complex coordinates come in handy: in problems involving nice angles (such as $\frac{}{4}, \frac{}{3}, \frac{}{2}$), and in problems about regular polygons.

In complex coordinates the line passing through the points z_1 and z_2 has the parametric equation $z = tz_1 + (1 \quad t)z_2$, $t \quad \mathbb{R}$. Also, the angle between the line passing through z_1 and z_2 and the line passing through z_3 and z_4 is the argument of the complex number $\frac{z_1 \quad z_2}{z_3 \quad z_4}$. The length of the segment determined by the points z_1 and z_2 is $|z_1 \quad z_2|$. The vertices of a regular n-gon can be chosen, up to a scaling factor, as $1, \ , \ ^2, \dots, \ ^{n \ 1}$, where $\ = e^{2 \ i/n} = \cos\frac{2}{n} + i \sin\frac{2}{n}$.

Example. Let ABC and BCD be two equilateral triangles sharing one side. A line passing through D intersects AC at M and AB at N. Prove that the angle between the lines BM and CN is $\frac{\pi}{3}$.

Solution. In the complex plane, let B have the coordinate 0, and C the coordinate 1. Then A and D have the coordinates $e^{i\pi/3}$ and $e^{-i\pi/3}$, respectively, and N has the coordinate $te^{i\pi/3}$ for some real number t.

The parametric equations of ND and AC are, respectively,

$$z = \alpha te^{i\pi/3} + (1-\alpha)e^{-i\pi/3} \quad\text{and}\quad z = \beta e^{i\pi/3} + (1-\beta), \quad \alpha, \beta \in \mathbb{R}.$$

To find their intersection we need to determine the real numbers α and β such that

$$\alpha te^{i\pi/3} + (1-\alpha)e^{-i\pi/3} = \beta e^{i\pi/3} + (1-\beta).$$

Explicitly, this equation is

$$\alpha t\frac{1+i\sqrt{3}}{2} + (1-\alpha)\frac{1-i\sqrt{3}}{2} = \beta\frac{1+i\sqrt{3}}{2} + (1-\beta).$$

Setting the real and imaginary parts equal, we obtain the system

$$\begin{aligned} \alpha t + (1-\alpha) &= \beta + 2(1-\beta), \\ \alpha t - (1-\alpha) &= \beta. \end{aligned}$$

By adding the two equations, we obtain $\alpha = \frac{1}{t}$. So the complex coordinate of M is $e^{i\pi/3} + (1-\frac{1}{t})e^{-i\pi/3}$.

The angle between the lines BM and CN is the argument of the complex number

$$\frac{e^{i\pi/3} + \left(1-\frac{1}{t}\right)e^{-i\pi/3}}{te^{i\pi/3}-1} = \frac{e^{i\pi/3} + e^{-i\pi/3} - \frac{1}{t}e^{-i\pi/3}}{te^{i\pi/3}-1} = \frac{1-\frac{1}{t}e^{-i\pi/3}}{te^{i\pi/3}-1} = \frac{1}{t}e^{-i\pi/3}.$$

The angle is therefore $\frac{\pi}{3}$, as claimed.

During the Mathematical Olympiad Summer Program of 2006, J. Bland discovered the following simpler solution:

Place the figure in the complex plane so that the coordinates of A, B, C, D are, respectively, $i\sqrt{3}$, -1, 1, and $-i\sqrt{3}$. Let MC have length $2t$, where t is a real parameter (positive if C is between A and M and negative otherwise). The triangles MCD and NBD have parallel sides, so they are similar. It follows that $BN = \frac{2}{t}$ (positive if B is between A and N and negative otherwise). The coordinates of M and N are

$$m = -\left(1+\frac{1}{t}\right) - \frac{1}{t}i\sqrt{3} \quad\text{and}\quad n = (t+1) - ti\sqrt{3}.$$

We compute

$$\frac{c \quad n}{b \quad m} = t\frac{2t + 1 + i \quad \overline{3}}{t \quad 2 + ti \quad \overline{3}} = \quad te^{i \frac{}{3}}.$$

It follows that the two lines form an angle of $\frac{}{3}$, as desired.

The second example comes from the 15th W.L. Putnam Mathematical Competition, 1955.

Example. Let $A_1A_2A_3 \ldots A_n$ be a regular polygon inscribed in the circle of center O and radius r. On the half-line $|OA_1$ choose the point P such that A_1 is between O and P. Prove that

$$\overset{n}{\underset{i=1}{}} PA_i = PO^n \quad r^n.$$

Solution. Place the vertices in the complex plane such that $A_i = r \ ^i$, $1 \quad i \quad n$, where is an nth root of unity. The coordinate of P is a real number rx, with $x > 1$. We have

$$\overset{n}{\underset{i=1}{}} PA_i = \overset{n}{\underset{i=1}{}} |rx \quad r \ ^i| = r^n \overset{n}{\underset{i=1}{}} |x \quad ^i| = r^n \overset{n}{\underset{i=1}{}} (x \quad ^i)$$
$$= r^n(x^n \quad 1) = (rx)^n \quad r^n = PO^n \quad r^n.$$

The identity is proved.

598. Let $ABCDEF$ be a hexagon inscribed in a circle of radius r. Show that if $AB = CD = EF = r$, then the midpoints of BC, DE, and FA are the vertices of an equilateral triangle.

599. Prove that in a triangle the orthocenter H, centroid G, and circumcenter O are collinear. Moreover, G lies between H and O, and $\frac{OG}{GH} = \frac{1}{2}$.

600. On the sides of a convex quadrilateral $ABCD$ one draws outside the equilateral triangles ABM and CDP and inside the equilateral triangles BCN and ADQ. Describe the shape of the quadrilateral $MNPQ$.

601. Let ABC be a triangle. The triangles PAB and QAC are constructed outside of the triangle ABC such that $AP = AB$, $AQ = AC$, and $\quad BAP = \quad CAQ = \quad$. The segments BQ and CP meet at R. Let O be the circumcenter of the triangle BCR. Prove that AO and PQ are orthogonal.

602. Let $A_1A_2 \ldots A_n$ be a regular polygon with circumradius equal to 1. Find the maximum value of $\overset{n}{\underset{k=1}{}} PA_k$ as P ranges over the circumcircle.

603. Let $A_0, A_1, \ldots, A_n$ be the vertices of a regular n-gon inscribed in the unit circle. Prove that

$$A_0A_1 \cdot A_0A_2 \cdots A_0A_{n\quad 1} = n.$$

4.1.3 Conics and Other Curves in the Plane

The general equation of a quadratic curve is

$$ax^2 + by^2 + cxy + dx + ey + f = 0.$$

Such a curve is called a conic because (except for the degenerate case of two parallel lines) it can be obtained by sectioning a circular cone by a plane.

The degenerate conics are pairs of (not necessarily distinct) lines, single points, the entire plane, and the empty set. We ignore them. There are three types of nondegenerate conics, which up to a change of Cartesian coordinates are described in Figure 27.

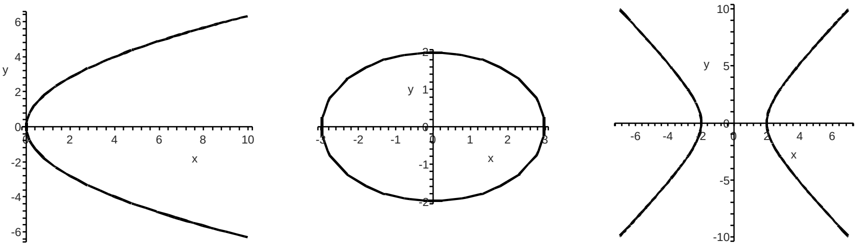

$y^2 = 4px$
parabola

$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$
ellipse

$\frac{x^2}{a^2} \quad \frac{y^2}{b^2} = 1$
hyperbola

Figure 27

The parabola is the locus of the points at equal distance from the point $(p, 0)$ (focus) and the line $x = \quad p$ (directrix). The ellipse is the locus of the points with the sum of distances to the foci $(c, 0)$ and $(\quad c, 0)$ constant, where $c = \quad \overline{|a^2 \quad b^2|}$. The hyperbola is the locus of the points with the difference of the distances to the foci $(c, 0)$ and $(\quad c, 0)$ constant, where $c = \quad \overline{a^2 + b^2}$.

Up to an affine change of coordinates, the equations of the parabola, ellipse, and hyperbola are, respectively, $y^2 = x$, $x^2 + y^2 = 1$, and $x^2 \quad y^2 = 1$. Sometimes it is more convenient to bring the hyperbola into the form $xy = 1$ by choosing its asymptotes as the coordinate axes.

As conic sections, these curves are obtained by sectioning the circular cone $z^2 = x^2 + y^2$ by the planes $z \quad x = 1$ (parabola), $z = 1$ (ellipse), and $y = 1$ (hyperbola). The vertex of the cone can be thought of as the viewpoint of a person. The projections through this viewpoint of one plane to another are called projective transformations. Up to a projective transformation there is only one nondegenerate conic—the circle. Any projectively invariant property that can be proved for the circle is true for any conic (and by passing to the limit, even for degenerate conics). Such is the case with Pascal's theorem: The opposite sides of a hexagon inscribed in a conic meet at three collinear points. Note that when the conic degenerates into two parallel lines, this becomes Pappus' theorem.

To conclude our discussion, let us recall that the equation of the tangent line to a conic at a point (x_0, y_0) is obtained by replacing in the general equation of the conic x^2 and y^2 by xx_0, respectively, yy_0, xy by $\frac{xy_0+yx_0}{2}$, and x and y in the linear terms by $\frac{x+x_0}{2}$, respectively, $\frac{y+y_0}{2}$.

We now proceed with an example from A. Myller's *Analytical Geometry* (3rd ed., Editura Didactica şi Pedagogica, Bucharest, 1972).

Example. Find the locus of the centers of the equilateral triangles inscribed in the parabola $y^2 = 4px$.

Solution. Let us determine first some algebraic conditions that the coordinates (x_i, y_i), $i = 1, 2, 3$, of the vertices of a triangle should satisfy in order for the triangle to be equilateral. The equation of the median from (x_3, y_3) is

$$\frac{y - y_3}{x - x_3} = \frac{y_1 + y_2 - 2y_3}{x_1 + x_2 - 2x_3}.$$

Requiring the median to be orthogonal to the side yields

$$\frac{y_1 + y_2 - 2y_3}{x_1 + x_2 - 2x_3} \cdot \frac{y_2 - y_1}{x_2 - x_1} = -1,$$

or

$$(x_1 - x_2)(x_1 + x_2 - 2x_3) + (y_1 - y_2)(y_1 + y_2 - 2y_3) = 0.$$

So this relation along with the two obtained by circular permutations of the indices are necessary and sufficient conditions for the triangle to be equilateral. Of course, the third condition is redundant. In the case of three points on the parabola, namely $(\frac{y_i^2}{4p}, y_i)$, $i = 1, 2, 3$, after dividing by $y_1 - y_2$, respectively, by $y_2 - y_3$ (which are both nonzero), we obtain

$$(y_1 + y_2)(y_1^2 + y_2^2 - 2y_3^2) + 16p^2(y_1 + y_2 - 2y_3) = 0,$$
$$(y_2 + y_3)(y_2^2 + y_3^2 - 2y_1^2) + 16p^2(y_2 + y_3 - 2y_1) = 0.$$

Subtracting the two gives

$$y_1^3 - y_3^3 + (y_1 - y_3)(y_2^2 - 2y_1y_3) + 48p^2(y_1 - y_3) = 0.$$

Divide this by $y_1 - y_3 \neq 0$ to transform it into

$$y_1^2 + y_2^2 + y_3^2 + 3(y_1y_2 + y_2y_3 + y_3y_1) + 48p^2 = 0.$$

This is the condition satisfied by the y-coordinates of the vertices of the triangle. Keeping in mind that the coordinates of the center of the triangle are

$$x = \frac{y_1^2 + y_2^2 + y_3^2}{12p}, \quad y = \frac{y_1 + y_2 + y_2}{3},$$

we rewrite the relation as

$$\frac{1}{2}(y_1^2 + y_2^2 + y_3^2) + \frac{3}{2}(y_1 + y_2 + y_3)^2 + 48p^2 = 0,$$

then substitute $12px = y_1^2 + y_2^2 + y_3^2$ and $3y = y_1 + y_2 + y_3$ to obtain the equation of the locus

$$-6px + \frac{27}{2}y^2 + 48p^2 = 0,$$

or

$$y^2 = \frac{4p}{9}(x - 8p).$$

This is a parabola with vertex at $(8p, 0)$ and focus at $((\frac{1}{9} + 8)p, 0)$.

The second problem was given at the 1977 Soviet Union University Student Mathematical Olympiad.

Example. Let P be a point on the hyperbola $xy = 4$, and Q a point on the ellipse $x^2 + 4y^2 = 4$. Prove that the distance from P to Q is greater than 1.

Solution. We will separate the conics by two parallel lines at a distance greater than 1. For symmetry reasons, it is natural to try the tangent to the hyperbola at the point $(2, 2)$. This line has the equation $y = 4 - x$.

Let us determine the point in the first quadrant where the tangent to the ellipse has slope -1. If (x_0, y_0) is a point on the ellipse, then the equation of the tangent at x is $xx_0 + 4yy_0 = 4$. Its slope is $-x_0/4y_0$. Setting $-x_0/4y_0 = -1$ and $x_0^2 + 4y_0^2 = 4$, we obtain $x_0 = 4/\sqrt{5}$ and $y_0 = 1/\sqrt{5}$. Consequently, the tangent to the ellipse is $y = \sqrt{5} - x$.

The distance between the lines $y = 4 - x$ and $y = \sqrt{5} - x$ is equal to $(4 - \sqrt{5})/\sqrt{2}$, which is greater than 1. Hence the distance between the arbitrary points P and Q is also greater than 1, and we are done.

604. Consider a circle of diameter AB and center O, and the tangent t at B. A variable tangent to the circle with contact point M intersects t at P. Find the locus of the point Q where the line OM intersects the parallel through P to the line AB.

605. On the axis of a parabola consider two fixed points at equal distance from the focus. Prove that the difference of the squares of the distances from these points to an arbitrary tangent to the parabola is constant.

606. With the chord PQ of a hyperbola as diagonal, construct a parallelogram whose sides are parallel to the asymptotes. Prove that the other diagonal of the parallelogram passes through the center of the hyperbola.

607. A straight line cuts the asymptotes of a hyperbola in points A and B and the hyperbola itself in P and Q. Prove that $AP = BQ$.

608. Consider the parabola $y^2 = 4px$. Find the locus of the points such that the tangents to the parabola from those points make a constant angle .

609. Let T_1, T_2, T_3 be points on a parabola, and t_1, t_2, t_3 the tangents to the parabola at these points. Compute the ratio of the area of triangle $T_1T_2T_3$ to the area of the triangle determined by the tangents.

610. Three points A, B, C are considered on a parabola. The tangents to the parabola at these points form a triangle MNP (NP being tangent at A, PM at B, and MN at C). The parallel through B to the symmetry axis of the parabola intersects AC at L.

(a) Show that $LMNP$ is a parallelogram.

(b) Show that the circumcircle of triangle MNP passes through the focus F of the parabola.

(c) Assuming that L is also on this circle, prove that N is on the directrix of the parabola.

(d) Find the locus of the points L if AC varies in such a way that it passes through F and is perpendicular to BF.

611. Find all regular polygons that can be inscribed in an ellipse with unequal semiaxes.

612. We are given the parabola $y^2 = 2px$ with focus F. For an integer n 3 consider a regular polygon $A_1A_2\ldots A_n$ whose center is F and such that none of its vertices is on the x-axis. The half-lines $|FA_1$, $|FA_2$, …, $|FA_n$ intersect the parabola at B_1, B_2, …, B_n. Prove that

$$FB_1 + FB_2 + \cdots + FB_n \quad np.$$

613. A *cevian* of a triangle is a line segment that joins a vertex to the line containing the opposite side. An *equicevian point* of a triangle ABC is a point P (not necessarily inside the triangle) such that the cevians on the lines AP, BP, and CP have equal lengths. Let SBC be an equilateral triangle, and let A be chosen in the interior of SBC, on the altitude dropped from S.

(a) Show that ABC has two equicevian points.
(b) Show that the common length of the cevians through either of the equicevian points is constant, independent of the choice of A.
(c) Show that the equicevian points divide the cevian through A in a constant ratio, which is independent of the choice of A.
(d) Find the locus of the equicevian points as A varies.
(e) Let S' be the reflection of S in the line BC. Show that (a), (b), and (c) hold if A varies on any ellipse with S and S' as its foci. Find the locus of the equicevian points as A varies on the ellipse.

A planar curve is called rational if it can be parametrized as $(x(t), y(t))$ with $x(t)$ and $y(t)$ rational functions of the real variable t. Here we have to pass to the closed real line, so t is allowed to be infinite, while the plane is understood as the projective plane, zero denominators giving rise to points on the line at infinity.

Theorem. *All conics are rational curves.*

Proof. The case of degenerate conics (i.e., pairs of lines) is trivial. The parabola $y^2 = 4px$ is parametrized by $(\frac{t^2}{4p}, t)$, the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ by $(a\frac{1-t^2}{1+t^2}, b\frac{2t}{1+t^2})$, and the hyperbola $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$ by $(a\frac{t+t^{-1}}{2}, b\frac{t-t^{-1}}{2})$. The general case follows from the fact that coordinate changes are rational (in fact, linear) transformations.

Compare the standard parametrization of the circle $(\cos x, \sin x)$ to the rational parametrization $(\frac{1-t^2}{1+t^2}, \frac{2t}{1+t^2})$. This gives rise to the trigonometric substitution $\tan \frac{x}{2} = t$ and explains why integrals of the form

$$\int R(\cos x, \sin x)dx,$$

with R a two-variable rational function, can be reduced to integrals of rational functions.

Let us change slightly our point of view and take a look at the conic

$$y^2 = ax^2 + bx + c.$$

If we fix a point (x_0, y_0) on this conic, the line $y - y_0 = t(x - x_0)$ intersects the conic in exactly one more point (x, y). Writing the conditions that this point is both on the line and on the conic and eliminating y, we obtain the equation

$$[y_0 + t(x - x_0)]^2 = ax^2 + bx + c.$$

A few algebraic computations yield

$$2y_0t + t^2(x - x_0) = a(x + x_0) + b.$$

This shows that x is a rational function of the slope t. The same is true for y. As t varies, (x, y) describes the whole conic. This is a rational parametrization of the conic, giving rise to Euler's substitutions. In their most general form, Euler's substitutions are

$$\sqrt{ax^2 + bx + c} - y_0 = t(x - x_0).$$

They are used for rationalizing integrals of the form

$$\int R(x, \sqrt{ax^2 + bx + c})dx,$$

where R is a two-variable rational function.

614. Compute the integral

$$\int \frac{dx}{a + b\cos x + c\sin x},$$

where a, b, c are real numbers, not all equal to zero.

615. Consider the system

$$x + y = z + u,$$
$$2xy = zu.$$

Find the greatest value of the real constant m such that $m \le \frac{x}{y}$ for any positive integer solution (x, y, z, u) of the system, with $x \ge y$.

We conclude our incursion into two-dimensional geometry with an overview of various famous planar curves. The first answers a question of G.W. Leibniz.

Example. What is the path of an object dragged by a string of constant length when the end of the string not joined to the object moves along a straight line?

Solution. Assume that the object is dragged by a string of length 1, that its initial coordinates are $(0, 1)$, and that it is dragged by a vehicle moving along the x-axis in the positive direction. Observe that the slope of the tangent to the curve at a point (x, y) points toward the vehicle, while the distance to the vehicle is always equal to 1. These two facts can be combined in the differential equation

$$\frac{dy}{dx} = -\frac{y}{\sqrt{1 - y^2}}.$$

Separate the variables

$$dx = \quad \frac{\overline{1 \quad y^2}}{y} dy,$$

and then integrate to obtain

$$x = \quad \overline{1 \quad y^2} \quad \ln y \quad \ln(1 + \quad \overline{1 \quad y^2}) + C.$$

The initial condition gives $C = 0$. The answer to the problem is therefore the curve

$$x = \quad \overline{1 \quad y^2} \quad \ln y \quad \ln(1 + \quad \overline{1 \quad y^2}),$$

depicted in Figure 28.

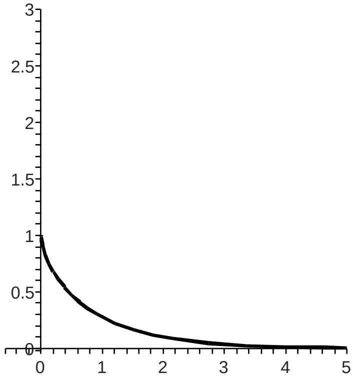

Figure 28

This curve is called a *tractrix*, a name given by Ch. Huygens. Clearly, it has the x-axis as an asymptote. E. Beltrami has shown that the surface of revolution of the tractrix around its asymptote provides a partial model for hyperbolic geometry. This surface has been used in recent years for the shape of loudspeakers.

A variety of other curves show up in the problems below. In some of the solutions, polar coordinates might be useful. Recall the formulas for changing between Cartesian and polar coordinates: $x = r \cos \quad$, $y = r \sin \quad$.

616. Find the points where the tangent to the cardioid $r = 1 + \cos \quad$ is vertical.

617. Given a circle of diameter AB, a variable secant through A intersects the circle at C and the tangent through B at D. On the half-line AC a point M is chosen such that $AM = CD$. Find the locus of M.

618. Find the locus of the projection of a fixed point on a circle onto the tangents to the circle.

619. On a circle of center O consider a fixed point A and a variable point M. The circle of center A and radius AM intersects the line OM at L. Find the locus of L as M varies on the circle.

620. The endpoints of a variable segment AB lie on two perpendicular lines that intersect at O. Find the locus of the projection of O onto AB, provided that the segment AB maintains a constant length.

621. From the center of a rectangular hyperbola a perpendicular is dropped to a variable tangent. Find the locus in polar coordinates of the foot of the perpendicular. (A hyperbola is called rectangular if its asymptotes are perpendicular.)

622. Find a transformation of the plane that maps the unit circle $x^2 + y^2 = 1$ into a cardioid. (Recall that the general equation of a cardioid is $r = 2a(1 + \cos\theta)$.)

623. Prove that the locus described by the equation $x^3 + 3xy + y^3 = 1$ contains precisely three noncollinear points A, B, C, equidistant to one another, and find the area of triangle ABC.

624. For n and p two positive integers consider the curve described by the parametric equations

$$x = a_1t^n + b_1t^p + c_1,$$
$$y = a_2t^n + b_2t^p + c_2,$$
$$z = a_3t^n + b_3t^p + c_3,$$

where t is a parameter. Prove that the curve is planar.

625. What is the equation that describes the shape of a hanging flexible chain with ends supported at the same height and acted on by its own weight?

4.1.4 Coordinate Geometry in Three and More Dimensions

In this section we emphasize quadrics. A quadric is a surface in space determined by a quadratic equation. The degenerate quadrics — linear varieties, cones, or cylinders over conics — add little to the picture from their two-dimensional counterparts, so we skip them. The nondegenerate quadrics are classified, up to an affine change of coordinates, as

$x^2 + y^2 + z^2 = 1$, ellipsoid;

$x^2 + y^2 - z^2 = 1$, hyperboloid of one sheet;

$x^2 - y^2 - z^2 = 1$, hyperboloid of two sheets;

$x^2 + y^2 = z$, elliptic paraboloid;

$x^2 - y^2 = z$, hyperbolic paraboloid.

In Cartesian coordinates, in these formulas there is a scaling factor in front of each term. For example, the standard form of an ellipsoid in Cartesian coordinates is

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1.$$

As in the case of conics, the equation of the tangent plane to a quadric at a point (x_0, y_0, z_0) is obtained by replacing in the equation of the quadric x^2, y^2, and z^2, respectively, by xx_0, yy_0, and zz_0; xy, xz, and yz, respectively, by $\frac{xy_0+yx_0}{2}$, $\frac{xz_0+zx_0}{2}$, and $\frac{yz_0+zy_0}{2}$; and x, y, and z in the linear terms, respectively, by $\frac{x+x_0}{2}$, $\frac{y+y_0}{2}$, and $\frac{z+z_0}{2}$.

Our first example comes from the 6th W.L. Putnam Mathematical Competition.

Example. Find the smallest volume bounded by the coordinate planes and by a tangent plane to the ellipsoid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1.$$

Solution. The tangent plane to the ellipsoid at (x_0, y_0, z_0) has the equation

$$\frac{xx_0}{a^2} + \frac{yy_0}{b^2} + \frac{zz_0}{c^2} = 1.$$

Its x, y, and z intercepts are, respectively, $\frac{a^2}{x_0}$, $\frac{b^2}{y_0}$, and $\frac{c^2}{z_0}$. The volume of the solid cut off by the tangent plane and the coordinate planes is therefore

$$V = \frac{1}{6} \frac{a^2b^2c^2}{x_0y_0z_0} .$$

We want to minimize this with the constraint that (x_0, y_0, z_0) lie on the ellipsoid. This amounts to maximizing the function $f(x, y, z) = xyz$ with the constraint

$$g(x, y, z) = \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1.$$

Because the ellipsoid is a closed bounded set, f has a maximum and a minimum on it. The maximum is positive, and the minimum is negative. The method of Lagrange multipliers yields the following system of equations in the unknowns x, y, z, and λ:

$$yz = 2 \ \frac{x}{a^2},$$
$$xz = 2 \ \frac{y}{b^2},$$
$$yz = 2 \ \frac{z}{c^2},$$
$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1.$$

Multiplying the first equation by x, the second by y, and the third by z, then summing up the three equations gives

$$3xyz = 2 \quad \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} \quad = 2 \ .$$

Hence $\ = \frac{3}{2}xyz$. Then multiplying the first three equations of the system together, we obtain

$$(xyz)^2 = 8 \ {}^3\frac{xyz}{a^2b^2c^2} = \frac{27(xyz)^4}{a^2b^2c^2}.$$

The solution $xyz = 0$ we exclude, since it does not yield a maximum or a minimum. Otherwise, $xyz = -\frac{abc}{\overline{27}}$. The equality with the plus sign is the maximum of f; the other is the minimum. Substituting in the formula for the volume, we find that the smallest volume is $\frac{\overline{3}}{2}abc$.

Example. Find the nature of the surface defined as the locus of the lines parallel to a given plane and intersecting two given skew lines, neither of which is parallel to the plane.

Solution. We will work in affine coordinates. Call the plane and the two skew lines l_1 and l_2. The x- and y-axes lie in and the z-axis is l_1. The x-axis passes through l_2 . The y-axis is chosen to make l_2 parallel to the yz-plane. Finally, the orientation and the units are such that l_2 is given by $x = 1$, $y = z$ (see Figure 29).

A line parallel to and intersecting l_1 and l_2 passes through $(1, s, s)$ and $(0, 0, s)$, where s is some real parameter playing the role of the height. Thus the locus consists of all points of the form $t(1, s, s) + (1 \quad t)(0, 0, s)$, where s and t are real parameters. The coordinates (X, Y, Z) of such a point satisfy $X = t$, $Y = ts$, $Z = s$. By elimination we obtain the equation $XZ = Y$, which is a hyperbolic paraboloid like the one from Figure 30. We stress once more that the type of a quadric is invariant under affine transformations.

A surface generated by a moving line is called a ruled surface. Ruled surfaces are easy to build in real life. This together with its structural resistance makes the hyperbolic paraboloid popular as a roof in modern architecture (see for example Felix Candela s roof

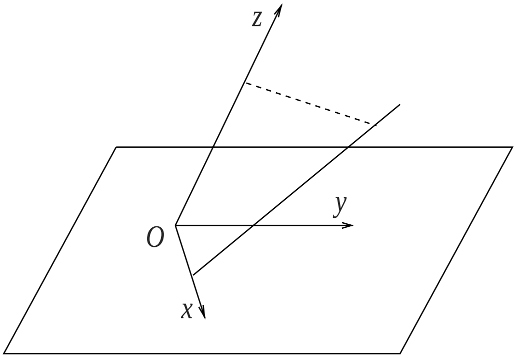

Figure 29

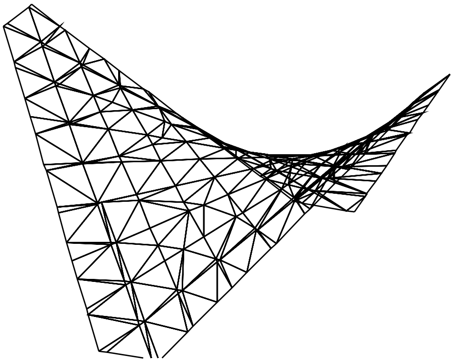

Figure 30

of the 1968 Olympic stadium in Mexico City). There is one more nondegenerate ruled quadric, which makes the object of the first problem in our list. And if you find some of the problems below too difficult, remember Winston Churchill s words: Success consists of going from failure to failure without loss of enthusiasm.

626. A cube is rotated about the main diagonal. What kind of surfaces do the edges describe?

627. Prove that the plane

$$\frac{x}{a}+\frac{y}{b}\quad\frac{z}{c}=1$$

is tangent to the hyperboloid of one sheet

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}\quad\frac{z^2}{c^2}=1.$$

628. Through a point M on the ellipsoid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$$

take planes perpendicular to the axes Ox, Oy, Oz. Let the areas of the planar sections thus obtained be S_x, S_y, respectively, S_z. Prove that the sum

$$aS_x + bS_y + cS_z$$

is independent of M.

629. Determine the radius of the largest circle that can lie on the ellipsoid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 \quad (a > b > c).$$

630. Let a, b, c be distinct positive numbers. Prove that through each point of the three-dimensional space pass three surfaces described by equations of the form

$$\frac{x^2}{a^2 \qquad} + \frac{y^2}{b^2 \qquad} + \frac{z^2}{c^2 \qquad} = 1.$$

Determine the nature of these surfaces and prove that they are pairwise orthogonal along their curves of intersection.

631. Show that the equations

$$\begin{aligned} x &= u + v + w, \\ y &= u^2 + v^2 + w^2, \\ z &= u^3 + v^3 + w^3, \end{aligned}$$

where the parameters u, v, w are subject to the constraint $uvw = 1$, define a cubic surface.

We conclude our discussion of coordinate geometry with some problems in n dimensions.

Example. Through a fixed point inside an n-dimensional sphere, n mutually perpendicular chords are drawn. Prove that the sum of the squares of the lengths of the chords does not depend on their directions.

Solution. We want to prove that the sum in question depends only on the radius of the sphere and the distance from the fixed point to the center of the sphere. Choose a coordinate system in which the chords are the n orthogonal axes and the radius of the sphere is $R > 0$. The fixed point, which we call P, becomes the origin. The endpoints of each chord have only one nonzero coordinate, and in the appropriate ordering, the kth coordinates of the endpoints X_k and Y_k of the kth chord are the nonzero numbers x_k and y_k, $k = 1, 2, \ldots, n$. The center of the sphere is then

$$O = \left(\frac{x_1 + y_1}{2}, \frac{x_2 + y_2}{2}, \ldots, \frac{x_n + y_n}{2}\right).$$

The conditions that the points X_k and Y_k lie on the sphere can be written as

$$\left(x_k - \frac{x_k + y_k}{2}\right)^2 + \sum_{j \neq k}\left(\frac{x_j + y_j}{2}\right)^2 = R^2,$$
$$\left(y_k - \frac{x_k + y_k}{2}\right)^2 + \sum_{j \neq k}\left(\frac{x_j + y_j}{2}\right)^2 = R^2,$$

with $k = 1, 2, \ldots, n$. This implies

$$\left(\frac{x_k - y_k}{2}\right)^2 = R^2 - \sum_{j \neq k}\left(\frac{x_j + y_j}{2}\right)^2, \quad k = 1, 2, \ldots, n.$$

The term on the left is one-fourth of the square of the length of X_kY_k. Multiplying by 4 and summing up all these relations, we obtain

$$\sum_{k=1}^{n} \|X_kY_k\|^2 = 4nR^2 - 4\sum_{k=1}^{n}\sum_{j \neq k}\left(\frac{x_j + y_j}{2}\right)^2 = 4nR^2 - 4(n-1)\sum_{k=1}^{n}\left(\frac{x_k + y_k}{2}\right)^2$$
$$= 4nR^2 - 4(n-1)\|PO\|^2.$$

Hence the conclusion.

632. Let n be a positive integer. Prove that if the vertices of a $(2n+1)$-dimensional cube have integer coordinates, then the length of the edge of the cube is an integer.

633. For a positive integer n denote by τ the permutation cycle $(n, \ldots, 2, 1)$. Consider the locus of points in $\mathbb{R}^n$ defined by the equation

$$\sum_{\sigma} \operatorname{sign}(\sigma) x_{\sigma(1)} x_{\tau(\sigma(2))} \cdots x_{\tau^{n-1}(\sigma(n))} = 0,$$

where the sum is over all possible permutations of $\{1, 2, \ldots, n\}$. Prove that this locus contains a plane.

634. Prove that the intersection of an n-dimensional cube centered at the origin and with edges parallel to the coordinate axes with the plane determined by the vectors

$$a = \left(\cos\frac{2\pi}{n}, \cos\frac{4\pi}{n}, \ldots, \cos\frac{2n\pi}{n}\right) \quad \text{and} \quad b = \left(\sin\frac{2\pi}{n}, \sin\frac{4\pi}{n}, \ldots, \sin\frac{2n\pi}{n}\right)$$

is a regular $2n$-gon.

635. Find the maximum number of points on a sphere of radius 1 in $\mathbb{R}^n$ such that the distance between any two points is strictly greater than $\sqrt{2}$.

4.1.5 Integrals in Geometry

We now present various applications of integral calculus to geometry problems. Here is a classic.

Example. A disk of radius R is covered by m rectangular strips of width 2. Prove that $m \geq R$.

Solution. Since the strips have different areas, depending on the distance to the center of the disk, a proof using areas will not work. However, if we move to three dimensions the problem becomes easy. The argument is based on the following property of the sphere.

Lemma. *The area of the surface cut from a sphere of radius* R *by two parallel planes at distance* d *from each other is equal to* $2\pi Rd$.

To prove this result, let us assume that the sphere is centered at the origin and the planes are perpendicular to the x-axis. The surface is obtained by rotating the graph of the function $f : [a, b] \to \mathbb{R}$, $f(x) = \sqrt{R^2 - x^2}$ about the x-axis, where $[a, b]$ is an interval of length d. The area of the surface is given by

$$\begin{aligned} 2\pi \int_a^b f(x)\sqrt{(f'(x))^2 + 1}\,dx &= 2\pi \int_a^b \sqrt{R^2 - x^2}\,\frac{R}{\sqrt{R^2 - x^2}}\,dx \\ &= 2\pi \int_a^b R\,dx = 2\pi Rd. \end{aligned}$$

Returning to the problem, the sphere has area $4\pi R^2$ and is covered by m surfaces, each having area $4\pi R$. The inequality $4\pi mR \geq 4\pi R^2$ implies that $m \geq R$, as desired.

The second example, suggested to us by Zh. Wang, is even more famous. We present the proof from H. Solomon, *Geometric Probability* (SIAM 1978).

Crofton s theorem. *Let* D *be a bounded convex domain in the plane. Through each point* $P(x,y)$ *outside* D *there pass two tangents to* D. *Let* t_1 *and* t_2 *be the lengths of the segments determined by* P *and the tangency points, and let be the angle between the tangents, all viewed as functions of* (x,y).[1] *Then*

$$_{P/D}\ \frac{\sin}{t_1t_2}dxdy = 2\ {}^{2}.$$

Proof. The proof becomes transparent once we examine the particular case in which D is the unit disk $x^2+y^2 < 1$. Each point outside the unit disk can be parametrized by the pair of angles $(\ _1,\ _2)$ where the tangents meet the unit circle S^1. Since there is an ambiguity in which tangent is considered first, the outside of the disk is in 1-to-2 correspondence with the set $S^1 \times S^1$. It so happens, and we will prove it in general, that on changing coordinates from (x,y) to $(\ _1,\ _2)$, the integral from the statement becomes $^{++}\ d\ _1 d\ _2$ (divided by 2 to take the ambiguity into account). The result follows.

In the general case we mimic the same argument, boosting your intuition with Figure 31. Fix a Cartesian coordinate system with the origin O inside D. For a point (x,y) denote by $(\ _1,\ _2)$ the angles formed by the perpendiculars from O onto the tangents with the positive semiaxis. This is another parametrization of the exterior of D, again with the ambiguity of which tangent is considered first. Let $A_i(\ _i,\ _i)$, $i = 1,2$, be the tangency points.

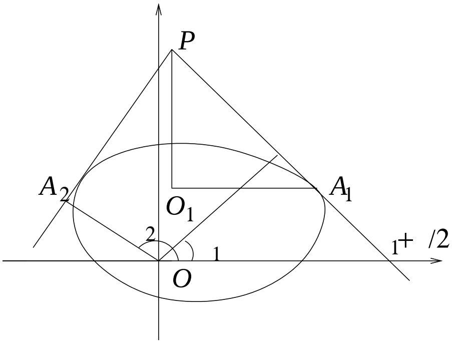

Figure 31

The main goal is to understand the change of coordinates $(x,y)\ \ (\ _1,\ _2)$ and in particular to write the Jacobian of this transformation. Writing the condition that the slope of the line A_1P is $\tan(\ _1 + \frac{}{2})$, we obtain

$$(x\ \ _1)\cos\ _1 + (y\ \ _1)\sin\ _1 = 0.$$

Taking the differential yields

---

[1] If the boundary of D has some edges, then there are points P for which t_1 and t_2 are not well defined, but the area of the set of these points is zero, so they can be neglected in the integral below.

$$\cos\ {}_1 dx \quad \cos\ {}_1 d\ {}_1 \quad (x \quad {}_1)\sin\ {}_1 d\ {}_1 + \sin\ {}_1 dy \quad \sin\ {}_1 d\ {}_1$$
$$+ (y \quad {}_1)\cos\ {}_1 d\ {}_1 = 0.$$

This expression can be simplified if we note that $\frac{d\ {}_1}{d\ {}_1}$ is the slope of the tangent, namely tan($_1$ + $\frac{}{2}$). Then cos $_1 d\ _1$ + sin $_1 d\ _1$ = 0, so

$$\cos\ {}_1 dx + \sin\ {}_1 dy = [(x \quad {}_1)\sin\ {}_1 \quad (y \quad {}_1)\cos\ {}_1] d\ {}_1.$$

And now a little Euclidean geometry. Consider the right triangle O_1A_1P with legs parallel to the axes. The altitude from O_1 determines on A_1P two segments of lengths $(x \quad {}_1)$ sin $_1$ and $(y \quad {}_1)$ cos $_1$ (you can see by examining the picture that the signs are right). This allows us to further transform the identity obtained above into

$$\cos\ {}_1 dx + \sin\ {}_1 dy = t_1 d\ {}_1.$$

The same argument shows that

$$\cos\ {}_2 dx + \sin\ {}_2 dy = t_2 d\ {}_2.$$

The Jacobian of the transformation is therefore the absolute value of

$$\frac{1}{t_1t_2}(\cos\ {}_1 \sin\ {}_2 \quad \sin\ {}_1 \cos\ {}_2) = \frac{1}{t_1t_2}\sin(\ {}_1 \quad {}_2).$$

And $_1$ $_2$ is, up to a sign, the supplement of . We obtain

$$2\ {}^2 = \frac{1}{2}\ {}_0^{2}\ {}_0^{2}\ d\ {}_1 d\ {}_2 = \ {}_{P/D} \frac{\sin}{t_1t_2} dxdy.$$

The theorem is proved.

636. A ring of height h is obtained by digging a cylindrical hole through the center of a sphere. Prove that the volume of the ring depends only on h and not on the radius of the sphere.

637. A polyhedron is circumscribed about a sphere. We call a face *big* if the projection of the sphere onto the plane of the face lies entirely within the face. Show that there are at most six big faces.

638. Let A and B be two finite sets of segments in three-dimensional space such that the sum of the lengths of the segments in A is larger than the sum of the lengths of the segments in B. Prove that there is a line in space with the property that the sum of the lengths of the projections of the segments in A onto that line is greater than the sum of the lengths of the projections of the segments in B.

639. Two convex polygons are placed one inside the other. Prove that the perimeter of the polygon that lies inside is smaller.

640. There are n line segments in the plane with the sum of the lengths equal to 1. Prove that there exists a straight line such that the sum of the lengths of the projections of the segments onto the line is equal to $\frac{2}{\ }$.

641. In a triangle ABC for a variable point P on BC with $PB = x$ let $t(x)$ be the measure of PAB. Compute

$$ {}_0^a \cos t(x)dx $$

in terms of the sides and angles of triangle ABC.

642. Let $f : [0, a] \quad R$ be a continuous and increasing function such that $f(0) = 0$. Define by R the region bounded by $f(x)$ and the lines $x = a$ and $y = 0$. Now consider the solid of revolution obtained when R is rotated around the y-axis as a sort of dish. Determine f such that the volume of water the dish can hold is equal to the volume of the dish itself, this happening for all a.

643. Consider a unit vector starting at the origin and pointing in the direction of the tangent vector to a continuously differentiable curve in three-dimensional space. The endpoint of the vector describes the spherical image of the curve (on the unit sphere). Show that if the curve is closed, then its spherical image intersects every great circle of the unit sphere.

644. With the hypothesis of the previous problem, if the curve is twice differentiable, then the length of the spherical image of the curve is called the total curvature. Prove that the total curvature of a closed curve is at least 2 .

645. A rectangle R is tiled by finitely many rectangles each of which has at least one side of integral length. Prove that R has at least one side of integral length.

4.1.6 Other Geometry Problems

We conclude with problems from elementary geometry. They are less in the spirit of Euclid, being based on algebraic or combinatorial considerations. Here imagination is more important than knowledge (A. Einstein).

Example. Find the maximal number of triangles of area 1 with disjoint interiors that can be included in a disk of radius 1. Describe all such configurations.

Solution. Let us first solve the following easier problem:

Find all triangles of area 1 *that can be placed inside a half-disk of radius* 1.

We will show that the only possible configuration is that in Figure 32. Consider a triangle that maximizes the area (such a triangle exists since the vertices vary on compact sets and the area depends continuously on the vertices). The vertices of this triangle must lie on the half-circle. If B lies between A and C, then A and C must be the endpoints of the diameter. Indeed, if say C is not an endpoint, then by moving it toward the closer endpoint of the diameter we increase both AC and the angle $\angle BAC$; hence we increase the area. Finally, among all triangles inscribed in a semicircle AC, the isosceles right triangle has maximal altitude, hence also maximal area. This triangle has area 1, and the claim is proved.

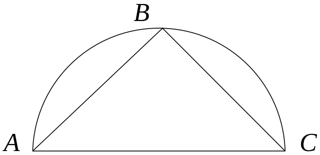

Figure 32

Returning to the problem, let us note that since the two triangles in question are convex sets, they can be separated by a line. That line cuts the disk into two regions, and one of them, containing one of the triangles, is included in a half-disk. By what we just proved, this region must itself be a half-disk. The only possible configuration consists of two isosceles triangles sharing the hypotenuse.

The next problem was published by the first author in the *Mathematics Magazine*.

Example. Let ABC be a right triangle ($\angle A = 90^\circ$). On the hypotenuse BC construct in the exterior the equilateral triangle BCD. Prove that the lengths of the segments AB, AC, and AD cannot all be rational.

Solution. We will find a relation between AB, AC, and AD by placing them in a triangle and using the law of cosines. For this, construct the equilateral triangle ACE in the exterior of ABC (Figure 33). We claim that $BE = AD$. This is a corollary of Napoleon's problem, and can be proved in the following way. Let M be the intersection of the circumcircles of BCD and ACE. Then $\angle AMC = 120^\circ$ and $\angle DMC = 60^\circ$; hence $M \in AD$. Similarly, $M \in BE$. Ptolemy's theorem applied to quadrilaterals $AMCE$ and $BMCD$ shows that $ME = AM + CM$ and $MD = BM + CM$; hence $AD = AM + BM + CM = BE$.

Applying the law of cosines in triangle ABE, we obtain $BE^2 = AB^2 + AE^2 + AB \cdot AE\sqrt{3}$, and since $BE = AD$ and $AE = AC$, it follows that

$$AD^2 = AB^2 + AC^2 + AB \cdot AC\sqrt{3}.$$

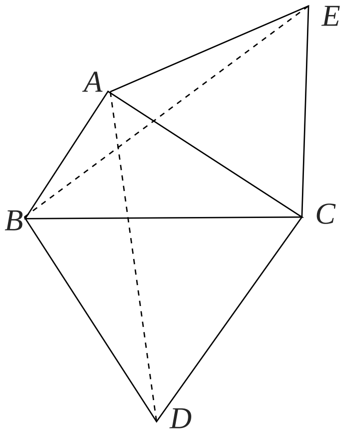

Figure 33

If all three segments AB, AC, and AD had rational lengths, this relation would imply that $\sqrt{3}$ is rational, which is not true. Hence at least one of these lengths is irrational.

646. Three lines passing through an interior point of a triangle and parallel to its sides determine three parallelograms and three triangles. If S is the area of the initial triangle and S_1, S_2, and S_3 are the areas of the newly formed triangles, prove that $S_1 + S_2 + S_3 \geq \frac{1}{3}S$.

647. Someone has drawn two squares of side 0.9 inside a disk of radius 1. Prove that the squares overlap.

648. A surface is generated by a segment whose midpoint rotates along the unit circle in the xy-plane such that for each $0 \leq \theta < 2\pi$, at the point of coordinates $(\cos\theta, \sin\theta)$ on the circle the segment is in the same plane with the z-axis and makes with it an angle of $\frac{\theta}{2}$. This surface, called a *Möbius band*, is depicted in Figure 34. What is the maximal length the segment can have so that the surface does not cross itself?

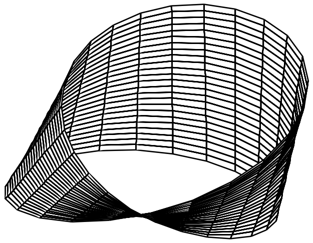

Figure 34

649. Let $ABCD$ be a convex quadrilateral and let O be the intersection of its diagonals. Given that the triangles OAB, OBC, OCD, and ODA have the same perimeter, prove that the quadrilateral is a rhombus. Does the property hold if O is some other point in the interior of the quadrilateral?

650. Prove that the plane cannot be covered by the interiors of finitely many parabolas.

651. Let ABC be a triangle with the largest angle at A. On line AB consider the point D such that A lies between B and D, and $AD = AB^3/AC^2$. Prove that $CD \leq \sqrt{3}BC^3/AC^2$.

652. Show that if all angles of an octagon are equal and all its sides have rational length, then the octagon has a center of symmetry.

653. Show that if each of the three main diagonals of a hexagon divides the hexagon into two parts with equal areas, then the three diagonals are concurrent.

654. Centered at every point with integer coordinates in the plane there is a disk with radius $\frac{1}{1000}$.
(a) Prove that there exists an equilateral triangle whose vertices lie in different disks.
(b) Prove that every equilateral triangle with vertices in different disks has side length greater than 96.

655. On a cylindrical surface of radius r, unbounded in both directions, consider n points and a surface S of area strictly less than 1. Prove that by rotating around the axis of the cylinder and then translating in the direction of the axis by at most $\frac{n}{4\pi r}$ units one can transform S into a surface that does not contain any of the n points.

4.2 Trigonometry

4.2.1 Trigonometric Identities

The beauty of trigonometry lies in its identities. There are two fundamental identities,

$$\sin^2 x + \cos^2 x = 1 \quad \text{and} \quad \cos(x \pm y) = \cos x \cos y \mp \sin x \sin y,$$

both with geometric origins, from which all the others can be derived. Our problems will make use of addition and subtraction formulas for two, three, even four angles, double- and triple-angle formulas, and product-to-sum formulas.

Example. Find all acute angles x satisfying the equation

$$2 \sin x \cos 40^\circ = \sin(x + 20^\circ).$$

Solution. Trying particular values we see that $x = 30^\circ$ is a solution. Are there other solutions? Use the addition formula for sine to rewrite the equation as

$$\tan x = \frac{\sin 20^\circ}{2\cos 40^\circ - \cos 20^\circ}.$$

The tangent function is one-to-one on the interval $(0, 90^\circ)$, which implies that the solution to the original equation is unique.

Example.

(a) Prove that if $\cos(\pi a) = \frac{1}{3}$ then a is an irrational number.
(b) Prove that a regular tetrahedron cannot be dissected into finitely many regular tetrahedra.

Solution. (a) Assume that a is rational, $a = \frac{m}{n}$. Then $\cos(n\pi a) = -1$. We will prove by induction that for all $k > 0$, $\cos(k\pi a) = \frac{m_k}{3^k}$, with m_k an integer that is not divisible by 3. This will then contradict the initial assumption.

The property is true for $k = 0$ and 1. The product-to-sum formula for cosines gives rise to the recurrence

$$\cos((k+1)\pi a) = 2\cos(\pi a)\cos(k\pi a) - \cos((k-1)\pi a), \quad k \geq 1.$$

Using the induction hypothesis, we obtain $\cos((k+1)\pi a) = \frac{m_{k+1}}{3^{k+1}}$, with $m_{k+1} = 2m_k - 3m_{k-1}$. Since m_k is not divisible by 3, neither is m_{k+1}, and the claim is proved.

Part (b) is just a consequence of (a). To see this, let us compute the cosine of the dihedral angle of two faces of a regular tetrahedron $ABCD$. If AH is an altitude of the tetrahedron and AE is an altitude of the face ABC, then $\angle AEH$ is the dihedral angle of the faces ABC and BCD (see Figure 35). In the right triangle HAE, $\cos\angle AEH = \frac{EH}{AD} = \frac{1}{3}$.

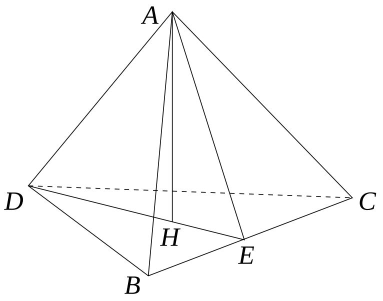

Figure 35

Now assume that there exists a dissection of a regular tetrahedron into regular tetrahedra. Several of these tetrahedra meet along a segment included in one of the faces of the initial tetrahedron. Their dihedral angles must add up to π, which implies that the dihedral angle of a regular tetrahedron is of the form $\frac{\pi}{n}$, for some integer n. This was shown above to be false. Hence no dissection of a regular tetrahedron into regular tetrahedra exists.

Remark. It is interesting to know that Leonardo da Vinci's manuscripts contain drawings of such decompositions. Later, however, Leonardo himself realized that the decompositions were impossible, and the drawings were mere optical illusions. Note also that Dehn's invariant mentioned in the first chapter provides an obstruction to the decomposition.

We conclude the introduction with a problem by the second author of the book.

Example. Let $a_0 = \sqrt{2} + \sqrt{3} + \sqrt{6}$ and let $a_{n+1} = \frac{a_n^2 - 5}{2(a_n+2)}$ for $n \geq 0$. Prove that

$$a_n = \cot\left(\frac{2^{n-3}\pi}{3}\right) - 2 \quad \text{for all } n.$$

Solution. We have

$$\cot\frac{\pi}{24} = \frac{\cos\frac{\pi}{24}}{\sin\frac{\pi}{24}} = \frac{2\cos^2\frac{\pi}{24}}{2\sin\frac{\pi}{24}\cos\frac{\pi}{24}} = \frac{1+\cos\frac{\pi}{12}}{\sin\frac{\pi}{12}} = \frac{1+\cos\left(\frac{\pi}{3}-\frac{\pi}{4}\right)}{\sin\left(\frac{\pi}{3}-\frac{\pi}{4}\right)}.$$

Using the subtraction formulas for sine and cosine we find that this is equal to

$$\frac{1+\frac{\sqrt{2}}{4}+\frac{\sqrt{6}}{4}}{\frac{\sqrt{6}}{4}-\frac{\sqrt{2}}{4}} = \frac{4+\sqrt{6}+\sqrt{2}}{\sqrt{6}-\sqrt{2}} = \frac{4(\sqrt{6}+\sqrt{2})+(\sqrt{6}+\sqrt{2})^2}{6-2}$$

$$= \frac{4(\sqrt{6}+\sqrt{2})+8+4\sqrt{3}}{4} = 2+\sqrt{2}+\sqrt{3}+\sqrt{6} = a_0+2.$$

Hence the equality $a_n = \cot\left(\frac{2^{n-3}\pi}{3}\right) - 2$ is true at least for $n = 0$.

To verify it in general, it suffices to prove that $b_n = \cot\left(\frac{2^{n-3}\pi}{3}\right)$, where $b_n = a_n + 2$, $n \geq 1$. The recurrence relation becomes

$$b_{n+1} - 2 = \frac{(b_n - 2)^2 - 5}{2b_n},$$

or $b_{n+1} = \frac{b_n^2 - 1}{2b_n}$. Assuming inductively that $b_k = \cot c_k$, where $c_k = \frac{2^{k-3}\pi}{3}$, and using the double-angle formula, we obtain

$$b_{k+1} = \frac{\cot^2 c_k - 1}{2\cot c_k} = \cot(2c_k) = \cot c_{k+1}.$$

This completes the proof.

656. Prove that

$$\sin 70^\circ \cos 50^\circ + \sin 260^\circ \cos 280^\circ = \frac{\sqrt{3}}{4}.$$

657. Show that the trigonometric equation

$$\sin(\cos x) = \cos(\sin x)$$

has no solutions.

658. Show that if the angles a and b satisfy

$$\tan^2 a \tan^2 b = 1 + \tan^2 a + \tan^2 b,$$

then

$$\sin a \sin b = - \sin 45\ .$$

659. Find the range of the function $f : R \quad\quad R$, $f(x) = (\sin x + 1)(\cos x + 1)$.

660. Prove that

$$\sec^{2n} x + \csc^{2n} x \quad\quad 2^{n+1},$$

for all integers $n \quad\quad 0$, and for all $x \quad\quad (0, \frac{}{2})$.

661. Compute the integral

$$\overline{\frac{1 \quad x}{1 + x}} dx, \quad\quad x \quad\quad (\quad 1, 1).$$

662. Find all integers k for which the two-variable function $f(x, y) = \cos(19x + 99y)$ can be written as a polynomial in $\cos x, \cos y, \cos(x + ky)$.

663. Let $a, b, c, d \quad\quad [0, \quad]$ be such that

$$2\cos a + 6\cos b + 7\cos c + 9\cos d = 0$$

and

$$2\sin a \quad\quad 6\sin b + 7\sin c \quad\quad 9\sin d = 0.$$

Prove that $3\cos(a + d) = 7\cos(b + c)$.

664. Let a be a real number. Prove that

$$5(\sin^3 a + \cos^3 a) + 3\sin a \cos a = 0.04$$

if and only if

$$5(\sin a + \cos a) + 2\sin a \cos a = 0.04.$$

665. Let $a_0, a_1, \ldots, a_n$ be numbers from the interval $(0, \frac{}{2})$ such that

$$\tan \quad a_0 \quad \frac{}{4} \quad + \tan \quad a_1 \quad \frac{}{4} \quad + \cdots + \tan \quad a_n \quad \frac{}{4} \quad\quad n \quad 1.$$

Prove that

$$\tan a_0 \tan a_1 \cdots \tan a_n \quad\quad n^{n+1}.$$

4.2.2 Euler's Formula

For a complex number z,

$$e^z = 1 + \frac{z}{1!} + \frac{z^2}{2!} + \cdots + \frac{z^n}{n!} + \cdots .$$

In particular, for an angle x,

$$e^{ix} = 1 + i\frac{x}{1!} - \frac{x^2}{2!} - i\frac{x^3}{3!} + \frac{x^4}{4!} + i\frac{x^5}{5!} - \frac{x^6}{6!} - i\frac{x^7}{7!} + \cdots .$$

The real part of e^{ix} is

$$1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots ,$$

while the imaginary part is

$$\frac{x}{1!} - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots .$$

These are the Taylor series of $\cos x$ and $\sin x$. We obtain Euler's formula

$$e^{ix} = \cos x + i \sin x.$$

Euler's formula gives rise to one of the most beautiful identities in mathematics: $e^{i\pi} = -1$, which relates the number e from real analysis, the imaginary unit i from algebra, and π from geometry.

The equality $e^{nz} = (e^z)^n$ holds at least for z a real number. Two power series are equal for all real numbers if and only if they are equal coefficient by coefficient (since coefficients are computed using the derivatives at 0). So equality for real numbers means equality for complex numbers. In particular, $e^{inx} = (e^{ix})^n$, from which we deduce the de Moivre formula

$$\cos nx + i \sin nx = (\cos x + i \sin x)^n.$$

We present an application of the de Moivre formula that we found in *Exercises and Problems in Algebra* by C. Nastasescu, C. Niţa, M. Brandiburu, and D. Joiţa (Editura Didactica şi Pedagogica, Bucharest, 1983).

Example. Prove the identity

$$\binom{n}{0} + \binom{n}{k} + \binom{n}{2k} + \cdots = \frac{2^n}{k}\sum_{j=1}^{k} \cos^n \frac{j\pi}{k} \cos \frac{nj\pi}{k}.$$

Solution. Let $\epsilon_1, \epsilon_2, \dots, \epsilon_k$ be the kth roots of unity, that is, $\epsilon_j = \cos\frac{2j\pi}{k} + i\sin\frac{2j\pi}{k}$, $j = 1, 2, \dots, k$. The sum

$$\epsilon_1^s + \epsilon_2^s + \cdots + \epsilon_k^s$$

is equal to k if k divides s, and to 0 if k does not divide s. We have

$$\sum_{j=1}^{k} (1 + \epsilon_j)^n = \sum_{s=0}^{n} \binom{n}{s} \sum_{j=1}^{k} \epsilon_j^s = k \sum_{j=0}^{\lfloor \frac{n}{k} \rfloor} \binom{n}{jk}.$$

Since

$$1 + \epsilon_j = 2\cos\frac{j\pi}{k}\left(\cos\frac{j\pi}{k} + i\sin\frac{j\pi}{k}\right),$$

it follows from the de Moivre formula that

$$\sum_{j=1}^{k} (1 + \epsilon_j)^n = \sum_{j=1}^{k} 2^n \cos^n\frac{j\pi}{k}\left(\cos\frac{nj\pi}{k} + i\sin\frac{nj\pi}{k}\right).$$

Therefore,

$$\binom{n}{0} + \binom{n}{k} + \binom{n}{2k} + \cdots = \frac{2^n}{k}\sum_{j=1}^{k} \cos^n\frac{j\pi}{k}\left(\cos\frac{nj\pi}{k} + i\sin\frac{nj\pi}{k}\right).$$

The left-hand side is real, so we can ignore the imaginary part and obtain the identity from the statement.

And now a problem given at an Indian Team Selection Test for the International Mathematical Olympiad in 2005, proposed by the first author of the book.

Example. For real numbers a, b, c, d not all equal to zero, let $f : \mathbb{R} \to \mathbb{R}$,

$$f(x) = a + b\cos 2x + c\sin 5x + d\cos 8x.$$

Suppose that $f(t) = 4a$ for some real number t. Prove that there exists a real number s such that $f(s) < 0$.

Solution. Let $g(x) = be^{2ix} - ice^{5ix} + de^{8ix}$. Then $f(x) = a + \operatorname{Re} g(x)$. Note that

$$g(x) + g\left(x + \frac{2\pi}{3}\right) + g\left(x + \frac{4\pi}{3}\right) = g(x)\left(1 + e^{2\pi i/3} + e^{4\pi i/3}\right) = 0.$$

Therefore,

$$f(x) + f \quad x + \frac{2}{3} \quad + f \quad x + \frac{4}{3} \quad = 3a.$$

If $a < 0$, then $s = t$ would work. If $a = 0$, then for some x one of the terms of the above sum is negative. This is because $f(x)$ is not identically zero, since its Fourier series is not trivial. If $a > 0$, substituting $x = t$ in the identity deduced above and using the fact that $f(t) = 4a$, we obtain

$$f \quad t + \frac{2}{3} \quad + f \quad t + \frac{4}{3} \quad = \quad a < 0.$$

Hence either $f(t + \frac{2}{3})$ or $f(t + \frac{4}{3})$ is negative. The problem is solved.

666. Prove the identity

$$\frac{1 + i \tan t}{1 \quad i \tan t}^{\,n} = \frac{1 + i \tan nt}{1 \quad i \tan nt}, \quad n \quad 1.$$

667. Prove the identity

$$1 \quad \genfrac{}{}{0pt}{}{n}{2} + \genfrac{}{}{0pt}{}{n}{4} \quad \genfrac{}{}{0pt}{}{n}{6} + \cdots = 2^{n/2} \cos \frac{n}{4}, \quad n \quad 1.$$

668. Compute the sum

$$\genfrac{}{}{0pt}{}{n}{1} \cos x + \genfrac{}{}{0pt}{}{n}{2} \cos 2x + \cdots + \genfrac{}{}{0pt}{}{n}{n} \cos nx.$$

669. Find the Taylor series expansion at 0 of the function

$$f(x) = e^{x \cos} \quad \cos(x \sin \quad),$$

where is a parameter.

670. Let z_1, z_2, z_3 be complex numbers of the same absolute value, none of which is real and all distinct. Prove that if $z_1 + z_2z_3$, $z_2 + z_3z_1$ and $z_3 + z_1z_2$ are all real, then $z_1z_2z_3 = 1$.

671. Let n be an odd positive integer and let be a real number such that − is irrational. Set $a_k = \tan(\quad + \frac{k}{n})$, $k = 1, 2, \ldots, n$. Prove that

$$\frac{a_1 + a_2 + \cdots + a_n}{a_1a_2 \cdots a_n}$$

is an integer and determine its value.

672. Find $(\cos \quad)(\cos 2 \quad)(\cos 3 \quad) \cdots (\cos 999 \quad)$ with $\quad = \frac{2}{1999}$.

673. For positive integers n define $F(n) = x^n \sin(nA) + y^n \sin(nB) + z^n \sin(nC)$, where x, y, z, A, B, C are real numbers and $A + B + C = k\pi$ for some integer k. Prove that if $F(1) = F(2) = 0$, then $F(n) = 0$ for all positive integers n.

674. The continuous real-valued function $\varphi(t)$ is defined for $t \geq 0$ and is absolutely integrable on every bounded interval. Define

$$P = \int_0^\infty e^{-(t+i\varphi(t))}dt \quad \text{and} \quad Q = \int_0^\infty e^{-2(t+i\varphi(t))}dt.$$

Prove that

$$|4P^2 - 2Q| \leq 3,$$

with equality if and only if $\varphi(t)$ is constant.

4.2.3 Trigonometric Substitutions

The fact that the circle $x^2 + y^2 = 1$ can be parametrized by trigonometric functions as $x = \cos t$ and $y = \sin t$ gives rise to the standard substitution $x = a\cos t$ (or $x = a\sin t$) in expressions of the form $\sqrt{a^2 - x^2}$. Our purpose is to emphasize less standard substitutions, usually suggested by the similarity between an algebraic expression and a trigonometric formula. Such is the case with the following problem from the 61st W.L. Putnam Mathematical Competition, 2000.

Example. Let $f : [-1, 1] \to \mathbb{R}$ be a continuous function such that $f(2x^2 - 1) = 2xf(x)$ for all $x \in [-1, 1]$. Show that f is identically equal to zero.

Solution. Here the expression $2x^2 - 1$ should remind us of the trigonometric formula $2\cos^2 t - 1 = \cos 2t$, suggesting the substitution $x = \cos t$, $t \in [0, \pi]$. The functional equation from the statement becomes $f(\cos 2t) = 2\cos t f(\cos t)$.

First, note that setting $x = 0$ and $x = 1$, we obtain $f(1) = f(-1) = 0$. Now let us define $g : \mathbb{R} \to \mathbb{R}$, $g(t) = \frac{f(\cos t)}{\sin t}$. Then for any t not a multiple of π,

$$g(2t) = \frac{f(2\cos^2 t - 1)}{\sin(2t)} = \frac{2\cos t f(\cos t)}{2\sin t\cos t} = \frac{f(\cos t)}{\sin t} = g(t).$$

Also, $g(t + 2\pi) = g(t)$. In particular, for any integers n and k,

$$g\left(1 + \frac{n\pi}{2^k}\right) = g(2^{k+1} + 2n\pi) = g(2^{k+1}) = g(1).$$

Because f is continuous, g is continuous everywhere except at multiples of π. The set $\{1 + \frac{n\pi}{2^k} \mid n, k \in \mathbf{Z}\}$ is dense on the real axis, and so g must be constant on its domain.

Then $f(\cos t) = c\sin t$ for some constant c and t in $(0, \pi)$, i.e., $f(x) = c\sqrt{1 - x^2}$ for all $x \in (-1, 1)$. It follows that f is an even function. But then in the equation from the statement $f(2x^2 - 1) = 2xf(x)$ the left-hand side is an even function while the right-hand side is an odd function. This can happen only if both sides are identically zero. Therefore, $f(x) = 0$ for $x \in [-1, 1]$ is the only solution to the functional equation.

We continue with a problem that was proposed by Belgium for the 26th International Mathematical Olympiad in 1985.

Example. Let x, y, z be real numbers such that $x + y + z = xyz$. Prove that

$$x(1 - y^2)(1 - z^2) + y(1 - z^2)(1 - x^2) + z(1 - x^2)(1 - y^2) = 4xyz.$$

Solution. The conclusion is immediate if $xyz = 0$, so we may assume that $x, y, z \neq 0$. Dividing through by $4xyz$ we transform the desired equality into

$$\frac{1 - y^2}{2y} \cdot \frac{1 - z^2}{2z} + \frac{1 - z^2}{2x} \cdot \frac{1 - x^2}{2x} + \frac{1 - x^2}{2x} \cdot \frac{1 - y^2}{2y} = 1.$$

This, along with the condition from the statement, makes us think about the substitutions $x = \tan A$, $y = \tan B$, $z = \tan C$, where A, B, C are the angles of a triangle. Using the double-angle formula

$$\frac{1 - \tan^2 u}{2\tan u} = \frac{1}{\tan 2u} = \cot 2u$$

we further transform the equality into

$$\cot 2B \cot 2C + \cot 2C \cot 2A + \cot 2A \cot 2B = 1.$$

But this is equivalent to

$$\tan 2A + \tan 2B + \tan 2C = \tan 2A \tan 2B \tan 2C,$$

which follows from $\tan(2A + 2B + 2C) = \tan 2\pi = 0$.

And now the problems.

675. Let $a, b, c \in [0, 1]$. Prove that

$$\sqrt{abc} + \sqrt{(1 - a)(1 - b)(1 - c)} \leq 1.$$

676. Solve the equation $x^3 - 3x = \sqrt{x + 2}$ in real numbers.

677. Find the maximum value of

$$S = (1 \quad x_1)(1 \quad y_1) + (1 \quad x_2)(1 \quad y_2)$$

if $x_1^2 + x_2^2 = y_1^2 + y_2^2 = c^2$, where c is some positive number.

678. Prove for all real numbers a, b, c the inequality

$$\frac{|a \quad b|}{\overline{1 + a^2}\ \ \overline{1 + b^2}} \qquad \frac{|a \quad c|}{\overline{1 + a^2}\ \ \overline{1 + c^2}} + \frac{|b \quad c|}{\overline{1 + b^2}\ \ \overline{1 + c^2}}.$$

679. Let a, b, c be real numbers. Prove that

$$(ab + bc + ca \quad 1)^2 \quad (a^2 + 1)(b^2 + 1)(c^2 + 1).$$

680. Prove that

$$\frac{x}{\overline{1 + x^2}} + \frac{y}{\overline{1 + y^2}} + \frac{z}{\overline{1 + z^2}} \qquad \frac{3\ \overline{3}}{2}$$

if the positive real numbers x, y, z satisfy $x + y + z = xyz$.

681. Prove that

$$\frac{x}{1 \quad x^2} + \frac{y}{1 \quad y^2} + \frac{z}{1 \quad z^2} \qquad \frac{3\ \overline{3}}{2}$$

if $0 < x, y, z < 1$ and $xy + yz + xz = 1$.

682. Solve the following system of equations in real numbers:

$$\frac{3x \quad y}{x \quad 3y} = x^2,$$
$$\frac{3y \quad z}{y \quad 3z} = y^2,$$
$$\frac{3z \quad x}{z \quad 3x} = z^2.$$

683. Let $a_0 = \quad \overline{2}$, $b_0 = 2$, and

$$a_{n+1} = \overline{2 \qquad \overline{4 \quad a_n^2}}, \qquad b_{n+1} = \frac{2b_n}{2 + \quad \overline{4 + b_n^2}}, \qquad n \quad 0.$$

(a) Prove that the sequences $(a_n)_n$ and $(b_n)_n$ are decreasing and converge to zero.

(b) Prove that the sequence $(2^n a_n)_n$ is increasing, the sequence $(2^n b_n)_n$ is decreasing, and these two sequences converge to the same limit.

(c) Prove there is a positive constant C such that one has $0 < b_n - a_n < \frac{C}{8^n}$ for all n.

684. Two real sequences $x_1, x_2, \ldots$, and $y_1, y_2, \ldots$ are defined in the following way:

$$x_1 = y_1 = \sqrt{3}, \quad x_{n+1} = x_n + \sqrt{1 + x_n^2}, \quad y_{n+1} = \frac{y_n}{1 + \sqrt{1 + y_n^2}}, \quad \text{for } n \geq 1.$$

Prove that $2 < x_n y_n < 3$ for all $n > 1$.

685. Let a, b, c be real numbers different from $-\frac{1}{\sqrt{3}}$. Prove that the equality $abc = a + b + c$ holds only if

$$\frac{3a - a^3}{3a^2 - 1} \cdot \frac{3b - b^3}{3b^2 - 1} \cdot \frac{3c - c^3}{3c^2 - 1} = \frac{3a - a^3}{3a^2 - 1} + \frac{3b - b^3}{3b^2 - 1} + \frac{3c - c^3}{3c^2 - 1}.$$

The parametrization of the hyperbola $x^2 - y^2 = 1$ by $x = \cosh t$, $y = \sinh t$ gives rise to the hyperbolic substitution $x = a \cosh t$ in expressions containing $\sqrt{a^2 - 1}$. We illustrate this with an example by the second author.

Example. Let $a_1 = a_2 = 97$ and

$$a_{n+1} = a_n a_{n-1} + \sqrt{(a_n^2 - 1)(a_{n-1}^2 - 1)}, \quad \text{for } n > 1.$$

Prove that

(a) $2 + 2a_n$ is a perfect square;
(b) $2 + \sqrt{2 + 2a_n}$ is a perfect square.

Solution. We are led to the substitution $a_n = \cosh t_n$ for some number t_n (which for the moment might be complex). The recurrence relation becomes

$$\cosh t_{n+1} = a_{n+1} = \cosh t_n \cosh t_{n-1} + \sinh t_n \sinh t_{n-1} = \cosh(t_n + t_{n-1}).$$

We deduce that the numbers t_n satisfy $t_0 = t_1$, and $t_{n+1} = t_n + t_{n-1}$ (in particular they are all real). And so $t_n = F_n t_0$, where $(F_n)_n$ is the Fibonacci sequence. Consequently, $a_n = \cosh(F_n t_0)$, $n \geq 1$.

Using the identity $2(\cosh t)^2 - 1 = \cosh 2t$, we obtain

$$2 + 2a_n = \left(2 \cosh F_n \frac{t_0}{2}\right)^2.$$

The recurrence relation

$$2 \cosh(k + 1)t = (2 \cosh t)(2 \cosh kt) - 2 \cosh(k - 1)t$$

allows us to prove inductively that $2\cosh k\frac{t_0}{2}$ is an integer once we show that $2\cosh\frac{t_0}{2}$ is an integer. It would then follow that $2\cosh F_n\frac{t_0}{2}$ is an integer as well. And indeed $2\cosh\frac{t_0}{2} = \sqrt{2+2a_1} = 14$. This completes the proof of part (a).

To prove (b), we obtain in the same manner

$$2+\sqrt{2+2a_n} = \left(2\cosh F_n\frac{t_0}{4}\right)^2,$$

and again we have to prove that $2\cosh\frac{t_0}{4}$ is an integer. We compute $2\cosh\frac{t_0}{4} = \sqrt{2+\sqrt{2+2a_n}} = \sqrt{2+14} = 4$. The conclusion follows.

686. Compute the integral

$$\int \frac{dx}{x+\sqrt{x^2-1}}.$$

687. Let $n > 1$ be an integer. Prove that there is no irrational number a such that the number

$$\sqrt[n]{a+\sqrt{a^2-1}}+\sqrt[n]{a-\sqrt{a^2-1}}$$

is rational.

4.2.4 Telescopic Sums and Products in Trigonometry

The philosophy of telescopic sums and products in trigonometry is the same as in the general case, just that here we have more identities at hand. Let us take a look at a slightly modified version of an identity of C.A. Laisant.

Example. Prove that

$$\sum_{k=0}^{n}\left(-\frac{1}{3}\right)^k\cos^3\left(3^{k-n}\pi\right) = \frac{3}{4}\left[\left(-\frac{1}{3}\right)^{n+1}+\cos\frac{\pi}{3^n}\right].$$

Solution. From the identity $\cos 3x = 4\cos^3 x - 3\cos x$, we obtain

$$\cos^3 x = \frac{1}{4}(\cos 3x + 3\cos x).$$

Then

$$\sum_{k=0}^{n}\left(-\frac{1}{3}\right)^k\cos^3\left(3^k a\right) = \frac{1}{4}\sum_{k=0}^{n}\left[\left(-\frac{1}{3}\right)^k\cos\left(3^{k+1}a\right)-\left(-\frac{1}{3}\right)^{k-1}\cos\left(3^k a\right)\right].$$

This telescopes to

$$\overset{2}{\frac{1}{4}}\qquad \frac{1}{3}^{\;n}\cos\ 3^{n+1}a \qquad\qquad \frac{1}{3}^{\;1}\overset{3}{\cos a}\ \ .$$

For $a = 3\ ^{n}$, we obtain the identity from the statement.

Test your skills against the following problems.

688. Prove that

$$27\sin^3 9\ \ + 9\sin^3 27\ \ + 3\sin^3 81\ \ + \sin^3 243\ \ = 20\sin 9\ \ .$$

689. Prove that

$$\frac{1}{\cot 9\qquad 3\tan 9} + \frac{3}{\cot 27\qquad 3\tan 27} + \frac{9}{\cot 81\qquad 3\tan 81} + \frac{27}{\cot 243\qquad 3\tan 243} = 10\tan 9\ \ .$$

690. Prove that

$$\frac{1}{\sin 45\ \ \sin 46} + \frac{1}{\sin 47\ \ \sin 48} + \cdots + \frac{1}{\sin 133\ \ \sin 134} = \frac{1}{\sin 1}.$$

691. Obtain explicit values for the following series:

(a) $\displaystyle \underset{n=1}{\quad}\ \arctan\frac{2}{n^2},$

(b) $\displaystyle \underset{n=1}{\quad}\ \arctan\frac{8n}{n^4\qquad 2n^2+5}.$

692. For $n\quad 0$ let

$$u_n = \arcsin\frac{\overline{n+1}\qquad \overline{n}}{\overline{n+2}\ \ \overline{n+1}}.$$

Prove that the series

$$S = u_0 + u_1 + u_2 + \cdots + u_n + \cdots$$

is convergent and find its limit.

Now we turn to telescopic products.

Example. Prove that

$$\,_{n=1}\frac{1}{1 \quad \tan^2 2^{\quad n}} = \tan 1.$$

Solution. The solution is based on the identity

$$\tan 2x = \frac{2\tan x}{1 \quad \tan^2 x}.$$

Using it we can write

$$\,_{n=1}^{N}\frac{1}{1 \quad \tan^2 2^{\quad n}} = \,_{n=1}^{N}\frac{\tan 2^{\quad n+1}}{2\tan 2^{\quad n}} = \frac{2^{\quad N}}{\tan 2^{\quad N}}\tan 1.$$

Since $\lim_{x \quad 0}\frac{\tan x}{x} = 1$, when letting N this becomes tan 1, as desired.

693. In a circle of radius 1 a square is inscribed. A circle is inscribed in the square and then a regular octagon in the circle. The procedure continues, doubling each time the number of sides of the polygon. Find the limit of the lengths of the radii of the circles.

694. Prove that

$$1 \quad \frac{\cos 61}{\cos 1} \quad 1 \quad \frac{\cos 62}{\cos 2} \quad \cdots \quad 1 \quad \frac{\cos 119}{\cos 59} = 1.$$

695. Evaluate the product

$$(1 \quad \cot 1\ )(1 \quad \cot 2\ )\cdots(1 \quad \cot 44\ ).$$

696. Compute the product

$$(\ \overline{3} + \tan 1\ )(\ \overline{3} + \tan 2\ )\cdots(\ \overline{3} + \tan 29\ ).$$

697. Prove the identities

(a) $\frac{1}{2} \quad \cos\frac{}{7} \quad \frac{1}{2} \quad \cos\frac{3}{7} \quad \frac{1}{2} \quad \cos\frac{9}{7} = \frac{1}{8},$

(b) $\frac{1}{2} + \cos\frac{}{20} \quad \frac{1}{2} + \cos\frac{3}{20} \quad \frac{1}{2} + \cos\frac{9}{20} \quad \frac{1}{2} + \cos\frac{27}{20} = \frac{1}{16}.$

698. Prove the identities

(a) $\,_{n=1}^{24} \sec(2^n) = 2^{24}\tan 2\ ,$

(b) $\,_{n=2}^{25} (2\cos(2^n) \quad \sec(2^n)\ ) = \quad 1.$

11.3 OMEGA TRIANGLES

Johann Friedrich Carl Gauss was born on April 30, 1777 in Brunswick in the German Duchy of Brunswick-Wolfenbüttel. He was born into a poor family. His paternal grandparents were originally small farmers but moved to Brunswick around 1740 in search of a better life. Gauss' parents, Gebhard and Dorothea, were hardworking but not well educated. Gebhard found that the local guilds did not welcome outsiders, so he struggled to find work. He would move from job to job, sometimes working as a gardener, a canal worker, a butcher, and sometimes as an accountant. This brought the family steady income, but acquiring wealth was very slow. This meant that the family goal of owning a home would take many years, but once achieved it would be a significant step up in social status (Bühler 1981, 5–6, 8; Burton 1985, 509–510).

After surviving a near drowning when he was around the age of 3, Gauss entered elementary school in 1784. In a class of about 50 students of various ages and abilities, Gauss stood out. He already knew how to read and write upon enrolling, abilities that he had taught himself. Supposedly, when he was 3 he could count and perform simple calculations, even to the point of correcting his father's ledgers. His masters at the school, J. G. Büttner and his assistant Martin Bartels, soon realized that they had an exceptional student on their hands. The details of the story differ among the various reports, but the essentials are that one day Gauss's class was given the task of adding the numbers from 1 to 100, either as simple busywork or as a punishment. Within moments Gauss had the answer, and it was the only correct one in the class! He had simply recognized the pattern in the sequence

$$1, 2, 3, \ldots, 98, 99, 100.$$

Since there are 50 pairs of numbers that add to 101, the sum must be 5050. Fortunately for the young Gauss, both Büttner and Bartels were good teachers, and they encouraged him in his studies. It was Büttner who aided Gauss in enrolling in secondary school in 1788. Gauss definitely benefited from his new surroundings. Not only was his academic life more structured, but he greatly improved his language skills. Up until this point, Gauss spoke only the local dialect of Brunswick. In secondary school he not only learned High German but also Latin (Bühler 1981, 7; Burton 1985, 510).

In 1791 an event happened to Gauss that would open more academic doors. Because of his excellent work in school, he was introduced to Duke Ferdinand of Brunswick. The duke was impressed with Gauss's achievements in school. As he had done with other promising students, Ferdinand granted Gauss the financial backing that would enable him to continue his education. This support was in the form of a stipend and would continue for many years. It eventually enabled Gauss to earn a doctorate and gave him the freedom to pursue some of his early mathematical research. Such backing was crucial for someone from his social standing at this time in Germany (Bühler 1981, 8; Burton 1985, 510).

The next stop in Gauss's academic career was Caroline College in 1792. He studied there until he was 18. This was a relatively new school, just 10 years old. Its curriculum was focused on the sciences, and it had a very good library. In it

Gauss found most of the great and contemporary works in mathematics. He studied the works of Newton and Archimedes, his two favorite mathematicians. It may have been in this library where Gauss began his work on the Parallel Postulate. As with many before him, he attempted to prove it indirectly, and as with many before him, his results were surprising but not contradictory. He would pursue this subject occasionally throughout the rest of his life (Bühler 1981, 8–9; Bonola 1912, 65).

Thus far, Gauss's education was within Brunswick-Wolfenbüttel. Ferdinand always appreciated it when those he supported attended his schools, but if they decided otherwise, he would still stand behind his students. This is what Gauss did in 1795. He packed up his bags and traveled 65 miles south to Göttingen to attend the university there. It had a solid reputation in the sciences and a good library. As with other institutions of higher learning at the time in Germany, the students were on their own. There were no mandatory classes or exams. The students were completely responsible for their education. Gauss took full advantage of this. He never did receive a Göttingen degree, but by the time that he left in 1798, he had laid the foundation for the mathematical research that would occupy much of his time over the next 25 years (Bühler 1981, 15, 17).

During his Göttingen years Gauss did not make many new friends. An exception was the Hungarian mathematician Farkas Bolyai, who he met in 1795. Bolyai was born on February 9, 1775 in the city of Bolya in Hungary. While he was a boy he attended the Lutheran and Calvinist College of Nagyenyed and showed a talent for mathematics. He was also skilled in music, drawing, and acting (Prékopa 2006, 11). While at the university, Gauss and Bolyai discovered that they shared an interest in the study of parallels. In June 1799, Bolyai left Göttingen to return to his home in Hungary, and for the next 54 years the two wrote frequently. In an example of their early correspondence from December 1799, Gauss discusses his attempts to prove the Parallel Postulate. He told Bolyai that he had tried to prove it by *reductio ad absurdum* arguments and had arrived at results that many would consider contradictory, but they were not. He wrote that because of this he had considered replacing the Parallel Postulate with another more intuitive one. He told Bolyai that "if it could be proved that a rectilinear triangle is possible with an area exceeding any given area," the Parallel Postulate would follow (Wolfe 1945, 128), but Gauss regarded such an assumption as no more likely to be true than the Postulate itself. As he had no experience with indefinitely producing two apparently parallel lines to see if they meet, Gauss had no experience with constructing arbitrarily large triangles. If the Parallel Postulate was to be replaced, any possible candidate must have at least some direct evidence in support of its truth (Bühler 1981, 16–17, 100–101; Bonola 1912, 65–66).

Gauss had wide and varying interests in addition to his mathematical studies. He was particularly fascinated with astronomy since his days at Caroline College. He enjoyed both the time at the telescope and the time crunching numbers, and he was good at both. He had the patience and attention to detail required of the discipline and the amazing ability for calculation that would bring him acclaim. The turn of the nineteenth century was an important time for astronomy. Uranus was discovered in 1781, and it would not be long until astronomers found Neptune in 1846. The instrumentation improved significantly, and with the aid that the newly discovered planets lent to orbital calculations, reliable sky maps were produced.

It was an event on January 1, 1801 that would set in motion a chain of events of which Gauss would take full advantage. The theory of the day stated that there should be a planet somewhere in the gap between Mars and Jupiter. On New Year's Day, Giuseppe Piazzi was observing the gap and thought that he had found the planet. He named it Ceres. We know it today as one of the larger minor planets in the asteroid belt. Piazzi successfully mapped approximately nine degrees of its orbit before it disappeared behind the "shadow of the sun." It was expected to reappear in late 1801 or early 1802, but since Piazzi had recorded only a little orbital data, where it would reappear was uncertain. Numerous astronomers made predications including Franz Xaver von Zach, one of Germany's top astronomers and the editor of the astronomical journal *Monatliche Correspondenz.* Gauss also made a prediction. When Ceres finally was sighted on December 7, 1801, it was Zach who found it. He published the results in the February issue of the *Monatliche Correspondenz*, proclaiming to the astronomical community that it was the mathematician Gauss whose prediction was closest. In 1809 Gauss published the full details of his technique in his famous *Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections* (Pannekoek 1961, 352–353; Bühler 1981, 43–44, 46).

Gauss was well on his way to becoming the greatest mathematician of his day and possibly of all time. With the Ceres calculation he earned the distinction of being among the top astronomers. Such a reputation had its benefits. In 1802, he began discussions with the Brunswick government about building a new observatory. In that same year he received an invitation from the academy at St. Petersburg to become the director of their observatory. When the astronomer Wilhelm Obers, who had become a good friend of Gauss, heard this, he contacted the university at Göttingen in an attempt to convince them to offer Gauss the directorship of their newly planned observatory. He did not want Gauss to leave Germany. Although initial reaction was restrained, the administration did express interest, so negotiations began with Obers taking the lead. While all of this was happening, Gauss continued to pursue his scientific interests. Then in 1805, Obers finally finished the deal with Göttingen. The university offered to appoint Gauss the head of the observatory. If he accepted he would be very independent from the university, with little classroom time and nearly no administrative duties. The Brunswick observatory never materialized, and the offer was too good to refuse. So, in 1807, Gauss left Brunswick and returned to Göttingen (Bühler 1981, 45–46, 50–51).

Gauss began work again on the parallels problem in 1813. He began to realize that the Parallel Postulate could not be proven from the other axioms. He accepted what Saccheri could not: that there might be a geometry different from Euclid's. This realization must have made all of his strange discoveries make sense. His previous attempts were within a framework of looking for a contradiction. Now he understood that there was no contradiction. The results were strange only because they were viewed from a Euclidean point of view. They needed to be viewed as part of a completely different geometry. Gauss initially called this new system **anti-Euclidean geometry**, but finally settled on calling his system **non-Euclidean**

geometry, the term we use today for any geometry in which the Parallel Postulate does not hold (Bonola 1912, 66–67).

Gauss was very busy both as a mathematician and an astronomer. He had worked hard all of his life, and he valued his reputation. He would not risk it harm by publishing anything that might set him up for ridicule. For this reason he did not want his discoveries concerning the new geometry made public. He did not even trust the mathematical community to be ready to accept such results. In any age, new ideas that go against the mainstream must struggle for acceptance, and in the early nineteenth century Euclid was still king in mathematics and even philosophy. Not wanting to deal with such matters, Gauss only wrote about his research to a few select confidants. One of his few trusted friends was the mathematician and astronomer Friedrich Bessel, to whom in 1829 he wrote:

> It may take a very long time before I make public my investigations on the issue. In fact, it may not happen during my lifetime, since I fear the scream of the Boeotians were I to completely express my views.

The Boeotians were an ancient Greek tribe who were not known for their intelligence. Despite his concern, in 1831, Gauss began to compose his *Meditations*, in which he planned to summarize his discoveries concerning non-Euclidean geometry. He did not wish for his research on the theory of parallels to die with him (Burton 1985, 551; Bonola 1912, 67).

The Area of a Triangle

Although Gauss was involved with many different types of projects during the course of his career, he would occasionally return to the question of parallels. One such time was on March 6, 1832. Bolyai had again written on the question, and Gauss responded. In his letter he included among other things a proof that the defect of a triangle is proportional to its area (Wolfe 1945, 128–130). Gauss had previously written to Bolyai in 1799 about the fact that the Parallel Postulate follows if it could be shown that there is no greatest triangle with respect to area. However, Gauss could not disprove the existence of such a triangle, so instead he assumed that there was one to see what would happen. He realized that the maximum triangle would be the limit of a sequence of non-Euclidean triangles found by letting the triangles increase in area as one progresses through the sequence. As this happens, the sides of the triangle pinch in and the angles decrease in measure until the vertices become ideal points (Section 9.2) and the measures of the angles become zero (Figure 11.4). In these triangles with three ideal vertices, all sides are parallel to each other.

■ DEFINITION 11.3.1

A triangle in which at least one of its vertices is an ideal point is called an **omega triangle**. An **ordinary triangle** has no ideal vertices.

Now assume **HA** and let M represent the greatest possible area of a triangle. Take $\overline{AB}$ and two lines $\overline{CD}$ and $\overline{EF}$ which are both parallel to $\overline{AB}$ and meet at a point P as in Figure 11.2. (Be careful to note that this is now Saccheri's definition of parallel;

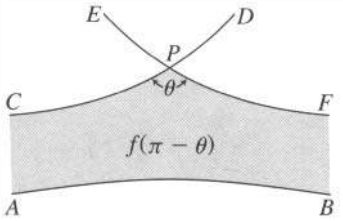

Figure 11.2 The area bounded by parallel lines.

see page 351.) Suppose that $m\sphericalangle CPF = \theta$. Define a function

$$f : [0, \pi] \to [0, M]$$

such that the area of the region bounded by these three lines equals $f(\pi - \theta)$. By the Saccheri-Legendre Theorem (11.2.7) we know that if $\theta = \pi$, the region can have no area. Therefore,

$$f(0) = 0.$$

If we take $\overline{CD}$ to be parallel to $\overline{EF}$ so that P is a point at infinity, then $\theta = 0$, and we have

$$f(\pi) = M.$$

Now imagine that the area transitions smoothly from 0 to M as θ ranges from π to 0. We see that it is not unreasonable to assume that f is not only continuous on $[0, \pi]$ but also differentiable. Furthermore, we infer from Figure 11.3 that

$$f(\pi - \theta) + f(\theta) = M$$

and

$$f(\theta) + f(\psi) + f(\pi - \theta - \psi) = M.$$

This last equation holds because

$$f(\pi - [\pi - \theta]) = f(\theta)$$

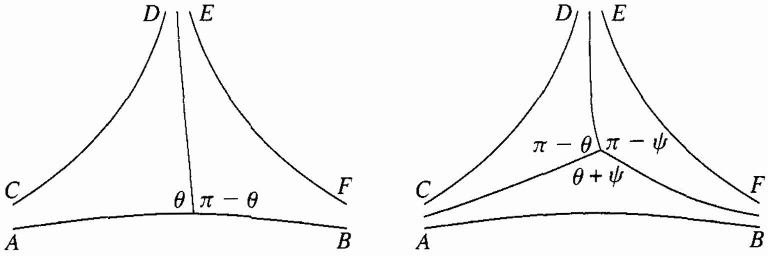

Figure 11.3 Omega triangles.

and

$$f(\pi - [\pi - \psi]) = f(\psi).$$

From this we conclude that

$$f(\psi) + f(\pi - \theta - \psi) = M - f(\theta) = f(\pi - \theta). \tag{11.2}$$

Next let $\gamma \geqslant 0$ be chosen such that $0 \leqslant \gamma \leqslant \pi$. Take $h \geqslant 0$, so that $\gamma + h \leqslant \pi$. Then there exists $\delta \geqslant 0$ such that $h = \pi - \delta - \gamma$. Notice that since $\delta = \pi - \gamma - h$, we have $0 \leqslant \delta \leqslant \pi$ and

$$\gamma + h = \gamma + \pi - \delta - \gamma = \pi - \delta. \tag{11.3}$$

From this we conclude that

$$f(\gamma + h) = f(\pi - \delta) = f(\gamma) + f(\pi - \delta - \gamma) = f(\gamma) + f(h),$$

where the first equation follows by Equation 11.3, the second by Equation 11.2, and the last by the definition of h. Hence for all α and β in the domain of f and $\alpha + h$, $\beta + h \leqslant \pi$,

$$\begin{aligned} f(\alpha + h) - f(\alpha) &= f(\alpha) + f(h) - f(\alpha) \\ &= f(\beta) + f(h) - f(\beta) = f(\beta + h) - f(\beta). \end{aligned}$$

This means that $f'(\alpha) = f'(\beta)$ for all α and β. It follows that there exists a constant $k > 0$ such that $f'(\theta) = k$, and since $f(0) = 0$ we see that for all θ, $f(\theta) = k\theta$. Because $f(\pi) = M$, we conclude that $M = k\pi$.

We are now ready to prove the result. Take a triangle with angles of measure α, β, and δ. Produce each of the sides as in Figure 11.4 and draw lines parallel to each of these extended sides producing a triangle with maximum area. Then

$$\text{area}(\triangle ABC) + f(\alpha) + f(\beta) + f(\delta) = M,$$

so

$$\text{area}(\triangle ABC) + k\alpha + k\beta + k\delta = k\pi.$$

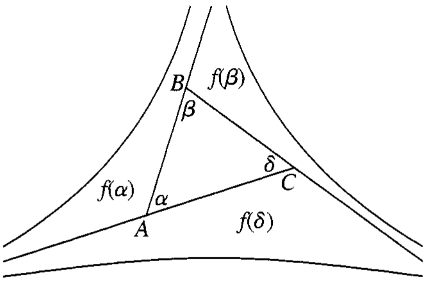

Figure 11.4 A triangle within a limiting maximum triangle.

From this we conclude that

$$\text{area}(\triangle ABC) = k(\pi - \alpha - \beta - \delta),$$

where $\pi - \alpha - \beta - \delta$ is the defect of the triangle. This shows that the area of a triangle is proportional to its defect assuming **HA**. This was a result also known to Lambert (Bonola 1912, 50).

Triangles with One Ideal Vertex

The omega triangle of Figure 11.4 has all of its vertices as ideal points. The triangles that we examine in this section will have only one ideal point as a vertex (see the figure for Theorem 11.3.4). These triangles share some of the same properties as regular triangles under **HA** (Wolfe 1945, 71–75).

Our first two results rely on an assumption that we have used before but now state explicitly. The postulate is named for Moritz Pasch, who in the late nineteenth century made significant contributions to the understanding of the logic behind geometry (Kline 1972, 1008). For these results, we say that a line **enters** a triangle if it contains a point on the interior of the triangle and either intersects a side or contains a vertex.

■ POSTULATE 11.3.2 [Pasch]

If a line enters an ordinary triangle through a side and does not contain one of its vertices, the line intersects exactly one of the other sides of the triangle.

The proof of the corollary is left for Exercise 2.

■ COROLLARY 11.3.3 [Pasch]

If a line enters an ordinary triangle through a given vertex, the line intersects the side of the triangle that is opposite the given vertex.

The next two theorems generalize these two propositions to omega triangles. We prove the first one here, but leave the second to Exercise 3.

■ THEOREM 11.3.4

[HA] If a line enters an omega triangle through a given vertex, the line will intersect the side opposite the given vertex.

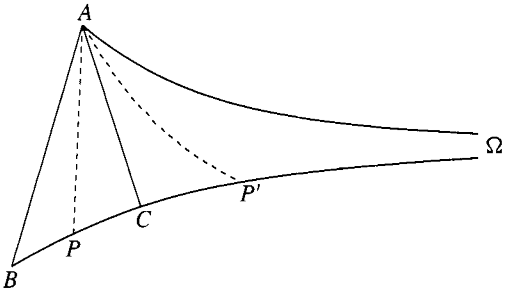

PROOF

Let $\overline{A\Omega}$ be parallel to $\overline{B\Omega}$. Find C on $\overline{B\Omega}$ such that $\overline{AC}$ is perpendicular to $\overline{B\Omega}$. If a line enters $\triangle AB\Omega$ through A, there are three options. The line could be on $\overline{AC}$. In this case it intersects $\overline{B\Omega}$ at C. It could lie on a line on the interior of $\measuredangle BAC$, but by Corollary 11.3.3 this line must intersect $\overline{BC}$ at some point P. Finally, if the line is on the interior of $\measuredangle CA\Omega$, it is below the parallel $\overline{A\Omega}$. Hence, it must intersect $\overline{C\Omega}$ at some P' by the definition of parallel line (Theorem 10.5.2 and page 351). ■

■ THEOREM 11.3.5

[HA] If a line enters an ideal triangle through a side and does not contain one of its vertices, the line intersects exactly one of the other sides of the triangle.

An important theorem regarding ordinary triangles states that an exterior angle is greater than either of its opposite interior angles (Theorem 5.3.9). The next theorem states that this also holds for triangles with an ideal vertex.

■ THEOREM 11.3.6

[HA] The exterior angles of $\triangle AB\Omega$ at A and B are greater than their respective opposite interior angles.

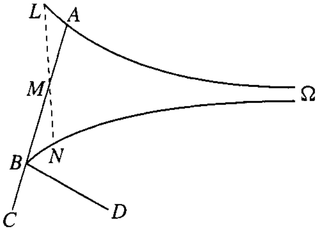

PROOF

In $\triangle AB\Omega$ produce $\overline{AB}$ to C. We will prove $m\measuredangle CB\Omega > m\measuredangle BA\Omega$. Copy $\measuredangle BA\Omega$ at B, producing $\measuredangle CBD$. First notice that $\overline{BD}$ will not intersect $\overline{A\Omega}$, else we would have a triangle in which one of its external angles is congruent to one of the opposite internal angles, and this would contradict Theorem 5.3.9. Therefore, $\overline{BD}$ is on the interior of $\measuredangle CB\Omega$ or D is on $\overline{B\Omega}$. Suppose the latter. This means that $\measuredangle CBD$ and $\measuredangle CB\Omega$ are the same angle. Let M be the midpoint of $\overline{AB}$ and find N on $\overline{B\Omega}$ such that $\overline{MN}$ is perpendicular to $\overline{B\Omega}$. Produce $\overline{A\Omega}$ to a point L through A such that $\overline{LA} \cong \overline{BN}$. Since our assumption on D implies that $\measuredangle CB\Omega \cong \measuredangle BA\Omega$, we have $\measuredangle LAM \cong \measuredangle NBM$. Thus, $\triangle MAL \cong \triangle MBN$ by SAS, and we may conclude that L, M, and N are collinear because $\measuredangle LMA \cong \measuredangle BMN$ (Theorem 5.3.8). From this we conclude that $\overline{LN}$ is perpendicular to both $\overline{A\Omega}$ and $\overline{B\Omega}$, but these two lines are parallel. This is a contradiction of **HA** (specifically, Theorem 10.5.2). Therefore, $\overline{BD}$ must lie on the interior of $\measuredangle CB\Omega$, which means that $m\measuredangle CB\Omega$ is greater than $m\measuredangle BA\Omega$. ■

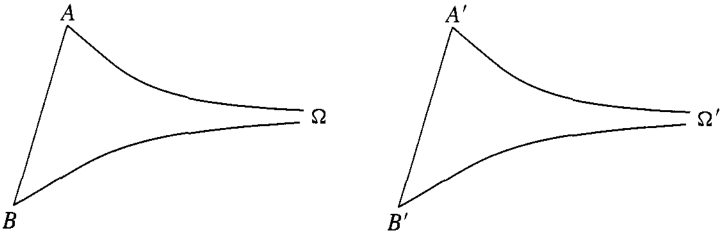

Figure 11.5 $\triangle AB\Omega \cong \triangle A'B'\Omega'$.

We finish the section by looking at what it means for two omega triangles with exactly one ideal vertex each to be congruent. Since any segment that has an ideal point as an endpoint is considered to be infinite in length, and since any angle that has an ideal point as a vertex is considered to have zero measure, we need to check only three conditions to show that these types of triangles are congruent (Figure 11.5).

■ DEFINITION 11.3.7

We say that $\triangle AB\Omega$ is **congruent** to $\triangle A'B'\Omega'$ if

- $\overline{AB} \cong \overline{A'B'}$
- $\sphericalangle AB\Omega \cong \sphericalangle A'B'\Omega'$.
- $\sphericalangle BA\Omega \cong \sphericalangle B'A'\Omega'$.

In this case we write $\triangle AB\Omega \cong \triangle A'B'\Omega'$.

We need to confirm only two of the three conditions to guarantee that a pair of such omega triangles are congruent.

■ THEOREM 11.3.8 [SA]

[HA] If $\overline{AB} \cong \overline{A'B'}$ and $\sphericalangle BA\Omega \cong \sphericalangle B'A'\Omega'$, then $\triangle AB\Omega \cong \triangle A'B'\Omega'$.

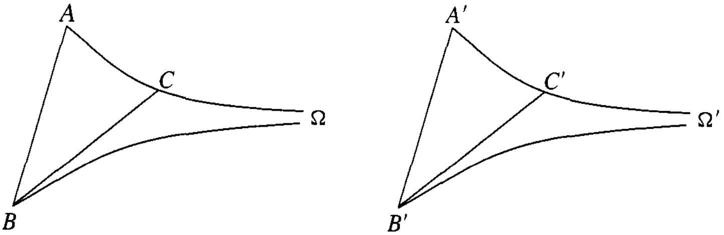

PROOF

Let $\overline{AB} \cong \overline{A'B'}$ and $\sphericalangle BA\Omega \cong \sphericalangle B'A'\Omega'$. We need to show that $\sphericalangle AB\Omega$ is congruent to $\sphericalangle A'B'\Omega'$. Suppose instead that $m\sphericalangle AB\Omega > m\sphericalangle A'B'\Omega'$. This means that there is a line on the interior of $\sphericalangle AB\Omega$ that forms an angle congruent to $\sphericalangle A'B'\Omega'$. This line must intersect $\overline{A\Omega}$ at some point C (Theorem 11.3.4). We may now write $\sphericalangle ABC \cong \sphericalangle A'B'\Omega'$. Next find C' on $\overline{A'\Omega'}$ such that $\overline{AC} \cong \overline{A'C'}$. We conclude that $\triangle BAC \cong \triangle B'A'C'$ by SAS. Therefore, $\sphericalangle ABC \cong \sphericalangle A'B'C'$. This is a contradiction because $\sphericalangle A'B'C' \cong \sphericalangle A'B'\Omega'$ by transitivity, but $\overline{B'C'}$ is on the interior of $\sphericalangle A'B'\Omega'$. The case when $m\sphericalangle AB\Omega$ is less than $m\sphericalangle A'B'\Omega'$ leads to a similar contradiction. ■

Our last theorem has a familiar name but an unfamiliar result. Compare this with Theorem 11.2.15.

■ THEOREM 11.3.9 [AA]

[HA] If $\sphericalangle BA\Omega \cong \sphericalangle B'A'\Omega'$ and $\sphericalangle AB\Omega \cong \sphericalangle A'B'\Omega$, then $\triangle AB\Omega$ is congruent to $\triangle A'B'\Omega'$.

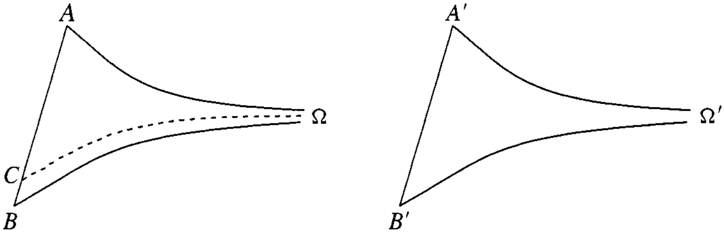

PROOF

Suppose that $\sphericalangle AB\Omega \cong \sphericalangle A'B'\Omega'$ and $\sphericalangle BA\Omega \cong \sphericalangle B'A'\Omega'$ in $\triangle AB\Omega$ and $\triangle A'B'\Omega'$, but $\overline{AB}$ is not congruent to $\overline{A'B'}$. This means either $AB > A'B'$ or $AB < A'B'$. It is enough to show that the first inequality leads to a contradiction, so suppose that we can find C on $\overline{AB}$ such that $\overline{AC} \cong \overline{A'B'}$. Thus, by SA, $\triangle AC\Omega \cong \triangle A'B'\Omega'$, which gives $\sphericalangle AC\Omega \cong \sphericalangle A'B'\Omega'$. Therefore, $\sphericalangle AC\Omega \cong \sphericalangle AB\Omega$ by transitivity. However, we know by Theorem 11.3.6 that $m\sphericalangle AC\Omega > m\sphericalangle AB\Omega$ because $\sphericalangle AC\Omega$ is exterior to $\triangle CB\Omega$. Therefore, $\overline{AB} \cong \overline{A'B'}$ and we conclude that $\triangle AB\Omega \cong \triangle A'B'\Omega'$. ■

Exercises

All problems except the first assume the Hypothesis of the Acute Angle.

1. Explain why Pasch's Postulate (11.3.2) is implied by the following statement: *If a line enters an ordinary triangle through a side and does not contain one of its vertices, the line intersects one of the other sides of the triangle.*

2. Prove Corollary 11.3.3.

3. Prove Theorem 11.3.5.

4. Let l intersect side $\overline{AB}$ of convex quadrilateral $ABCD$ and contain a point in the interior of $ABCD$. Prove that exactly one of the following holds: C is on l, D is on l, l intersects $\overline{BC}$, l intersects $\overline{CD}$, or l intersects $\overline{AD}$.

5. What proposition must be true in order for the area of a quadrilateral to be proportional to its defect? Assuming this proposition to be true, prove that the area of a quadrilateral is proportional to its defect.

6. Show that in $\triangle AB\Omega$ the sum of $m\sphericalangle AB\Omega$ and $m\sphericalangle BA\Omega$ is less than two right angles.

7. Let $\triangle AB\Omega$ be an omega triangle such that $\sphericalangle AB\Omega \cong \sphericalangle BA\Omega$. Let M be the midpoint of $\overline{AB}$ and let $\overline{M\Omega}$ be the line through M that is parallel to both $\overline{A\Omega}$ and $\overline{B\Omega}$. Prove that $\overline{M\Omega}$ is perpendicular to $\overline{AB}$.

8. Let A, B, and C be collinear with B between A and C. Let $\overline{AC}$ be perpendicular to $\overline{C\Omega}$. Demonstrate that $m\sphericalangle CA\Omega < m\sphericalangle CB\Omega$.

9. Show that the A in SA cannot be dropped by drawing two noncongruent omega triangles that have congruent corresponding sides.

10. Show that one of the A's in AA cannot be dropped by drawing two noncongruent omega triangles that have one pair of congruent corresponding angles

11. Let B be on $\overline{A\Omega}$ and C on $\overline{A\Omega'}$. Assume that $\overline{A\Omega}$ is perpendicular to $\overline{A\Omega'}$. Prove:
   (a) $\overline{B\Omega'}$ intersects $\overline{C\Omega}$ in a point D.
   (b) If $\overline{AB} \cong \overline{AC}$, then $\triangle BD\Omega \cong \triangle CD\Omega'$.

12. Let Ω and Ω' be on opposite sides of $\overline{AB}$. Suppose that $\overline{A\Omega}$ and $\overline{B\Omega'}$ are perpendicular to $\overline{AB}$. Prove that $\sphericalangle AB\Omega \cong \sphericalangle BA\Omega'$.

13. Does the Pythagorean Theorem hold when **HA** is assumed? Explain.

14. In $\triangle AB\Omega$ and $\triangle A'B'\Omega'$ we have $\sphericalangle AB\Omega \cong \sphericalangle A'B'\Omega'$ and $AB > A'B'$. Show that $m\sphericalangle BA\Omega < m\sphericalangle B'A'\Omega'$.

11.4 PURE REASON

Immanuel Kant was born in Königsberg on April 22, 1724 to Johann Georg and Ann Regina Kant. Immanuel was born the fourth child, but only an older sister was still alive at the time of his birth. Immanuel's father was a master harness maker and his mother was the daughter of a harness maker. Due to his father's expertise in his trade, the family lived well but not extravagantly and was very dependent on the local economy. As trade decreased in Königsberg, so did the family's fortunes, but they never reached a level that we would consider poor. Those early years with his family taught Kant the virtues of hard work, self-reliance, and honesty. Like his father before him, Kant determined that he would be a self-made man (Kuehn 2001, 26–31).

Kant's parents were very religious and greatly influenced by the Pietist movement within the Lutheran church. Pietism appeared in the late seventeenth century and was prominent in the first half of the eighteenth century. As a reaction to a church whose members appeared simply to be going through the motions, the Pietists attempted to bring back a living Christianity. Kant's parents wanted their son to receive a sound eduction, so they sent him to the Pietist *Collegium Fridericianum*. Little is known about the details of Kant's religious beliefs, but it is known that Kant hated the school. Even though it provided Kant with a good education and a chance for social advancement, he despised the school's attempts to convert him and the suffocating moralism of the faculty (Kuehn 2001, 34–35, 39–40, 45, 51–52, 54).

Not only did the Pietists run the *Collegium Fridericianum*, but they grew in influence at the University of Königsberg, where Kant began to study in 1740, when he was 17. The city of Königsberg was an important but remote city in northeastern Prussia near the Russian border. This remoteness had its disadvantages. It was difficult to recruit faculty, and the 44 full professors who were on staff earned low salaries. The remaining faculty were not paid by the university but instead were ***privatdozents***, faculty who were allowed to charge their students directly for attending their lectures. Kant would have attended both the free lectures given by those on the university's payroll and the paid ones. From these courses he learned subjects which included general philosophy, metaphysics, logic, ethics, natural law, physics, systematic theology, rhetoric, history, poetry, psychology, and mathematics. While at the *Collegium Fridericianum*, Kant became particularly interested in the classics, and he continued those studies at the university (Kuehn 2001, 45, 56, 63, 73, 75, 76).

A turning point came in 1744. Under the influence of faculty from both philosophy and physics, Kant turned to philosophy and began work on his first major philosophical piece. It was titled *Thoughts on the True Estimation of the Living Forces*. The work is a critique of both Newton and Leibnitz. Some also surmise that it contains a veiled attack on some of the Königsberg faculty. Eventually, this made Kant's life at Königsberg uncomfortable. 1744 also saw tragedy in the Kant family. Kant had lost his mother seven years earlier, and this year his father suffered a stroke that would lead to his death in 1746. This, combined with the turmoil at school and the fact that he had finished his formal studies, led to Kant's decision to leave Königsberg in 1748 and move to the surrounding countryside (Kuehn 2001, 80, 81, 86, 89–95).

There Kant became a private tutor. The position did not hold a high social status. It was something akin to a servant, and many who did not succeed at the university found themselves teaching the children of those wealthy enough to afford it. From the time he left his home to 1754, Kant worked for various families in Judtschea, Arnsdorf, and Rautenburg. He never did consider himself very good at teaching children, but his students and their parents thought otherwise and kept in contact with him long after his work was done (Kuehn 2001, 95–97).

It was apparently Kant's goal not to stay away from Königsberg forever. While working as a tutor he continued his research. In 1749 he published his *True Estimation of Living Forces*, and in 1754 he returned to Königsberg. On April 7 of the following year he turned in his dissertation. It was titled "Succinct Meditations on Fire," and in

June he made a public defense, the final step to his doctorate. At this point, however, Kant did not have the right to teach at the university. To earn this he would have to present another dissertation, called a ***Habilitationsschrift***, and then give a public lecture, the ***Habilitationsvortrag***. This occurred in September 1755 when he gave a talk entitled "A New Exposition of the First Principles of Metaphysics." For the next 15 years Kant taught students, conducted research, sought to be appointed to the position of full professor, and wrote. Some of his titles include *General History and Theory of the Heavens* (1755), *The Only Possible Argument in Support of a Demonstration of the Existence of God* (1763), *Observations on the Feeling of the Beautiful and Sublime* (1764), and his "prize essay" for the Prussian Academy of Sciences in Berlin entitled *Inquiry Concerning the Distinctness of the Principles of Natural Theology and Morality* (1764) (Kuehn 2001, xv–xvii, 98, 100).

Then in 1765, Kant received a letter that would begin a correspondence that would last for 12 years. It was from Lambert, who had read Kant's *The Only Possible Argument in Support of a Demonstration of the Existence of God* and found thoughts very similar to his own. Lambert was confident that if Kant would read his *Neues Organon* and *Anlage zur Architectonik*, Kant would share in this opinion. In writing, Lambert sought to exchange letters on subjects ranging from mathematics to metaphysics, with his main interest being Kant's views regarding space and time. Lambert wanted to prove that Euclidean geometry was true but had been unsuccessful. This caused him to recognize the logical possibility of an alternative geometry. Although Kant would reject Lambert's view of space, their letters greatly aided Kant's research and helped direct Kant's writing of his *Critique of Pure Reason* (Kant 1999, 77–79, 81–82; Kuehn 2001, xvii, 181, 196, 231).

The creation of the *Critique* began when the University of Königsberg appointed Kant to be professor of logic and metaphysics in 1770. Again a paper and a talk were required. Kant defended this inaugural dissertation on August 21. He sent a copy to Lambert a month later and wrote to him that he expected to quickly produce a major work from it by making only minor modifications. Kant later claimed that it took him four to five months of work to finish the *Critique*, but it was the result of 11 years of contemplation. In May 1781 the *Critique* was published, followed by a second edition in 1787. The fundamental question that Kant sought to answer was: "What can I know?" He concluded that traditional philosophy erred when it taught that one can discover truths about the world independent of experience. Instead, all of our knowledge of the world is obtained through the senses, but the structure that we put on those perceptions exists in our minds independent of any sensory input. The structure that we put onto space, for example, is described by Euclid's geometry. Metaphysics then becomes the study of the presuppositions of experience. We can only know the world as it appears to us, never as it actually is. Kant called his position **transcendental idealism** (Kuehn 2001, 231, 240–243, 246).

Kant intended to dedicate the *Critique of Pure Reason* to his good colleague Johann Lambert, but Lambert died on September 27, 1777 shortly before its publication. Lambert never saw the *Critique* in its final form, nor did he see the publication of his own work on geometry, entitled *Theory of Parallel Lines*. It was published in 1786 by Johann Bernoulli along with much of Lambert's correspondence, including his letters to Kant (Kuehn 2001, 231, 254; Kant 1999, 185–187; Boyer and Merzbach 1991, 462).

Describing Knowledge

Saccheri believed that the Parallel Postulate was true, so he went looking for a proof. He found one that he trusted, so he believed that the Postulate held. However, we know that his proof was not valid. We conclude that regardless of the truth of the Parallel Postulate, Saccheri did not know its truth value. This is because his reason for his belief was not sound. We base this conclusion on the definition of what it means to know that a statement is true.

■ **DEFINITION 11.4.1**

For a person to have **knowledge** that a proposition is true, three conditions must be satisfied:

- The person believes that the proposition is true.
- The statement is in fact true.
- The person has good reason to believe that the proposition is true.

The definition of knowledge and the discussion of this section is adapted from Halverson (1981, 25–63).

To clarify what it means to know that a proposition is true, let us examine a sentence that is simpler than the Parallel Postulate. Let p represent the claim that

there exist unicorns on Mars.

What would it mean for us to know p? With respect to the three conditions, it means:

- To know proposition p, we must first affirm it to be true. This is the bare minimum. If we believe that there are no unicorns on Mars, then whether there are unicorns or not, how could it be said that we know that they exist?
- Of course, simply believing p does not make it so. For us to know p, it must also be the case that p is true. If we believe that there are unicorns on Mars and there are none, we do not know that there are unicorns on Mars. Instead we believe a proposition that is false.
- Suppose that it is indeed the case that unicorns live on the red planet and we believe it. This is still not enough for us to be able to claim knowledge of this state of affairs. We must add to this a good reason to give credence to our belief. If we do not, we have no reason to have any confidence that our belief is accurate even though in reality it might be.

To have knowledge that p is true, we need the three listed conditions to hold, and it is crucial that we do not overlook the need for an adequate justification of our beliefs.

So what would be adequate justification for believing the Parallel Postulate? It depends on the type of proposition it is. Let us first compare it to p. Any justification of p requires information about the unicorns that is not included in the sentence. We need a photograph, a reliable eye witness, or some other kind of evidence that the words of the sentence do not provide. Nothing from the definition of a unicorn suggests that they live on Mars. Hence, the sentence p claims to give new information about its subject. Does the Parallel Postulate also claim to give new information about lines? The Postulate states that

> *if a straight line falling on two straight lines make the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which the angles are less than the two right angles.*

If there is something that the Parallel Postulate tells us about lines that is not included in the meaning of the word *line*, it is a sentence like p and requires evidence to determine its truth.

However, consider the assertion that

all bachelors are unmarried men.

For this statement no additional information is required to determine its truth. The definition of *bachelor* gives it to us. A bachelor is an unmarried man, and this is what the sentence states. It tells us nothing new about its subject. If the Parallel Postulate more closely resembles this type of assertion, observation is not needed to determine its truth. We need only to analyze the meanings of the words and possibly apply some logic. This and the previous case exhaust the possibilities, so we can make the following definitions.

■ **DEFINITION 11.4.2**

A **synthetic** proposition is one that gives new information about its subject.

■ **DEFINITION 11.4.3**

An **analytic** proposition is one that is not synthetic. It does not give new information about its subject.

The sentence *there exists unicorns on Mars* is synthetic. Either direct or indirect experience of the unicorns is required to prove it. If such evidence is found, we believe it, and if the sentence turns out to be true, we have gained knowledge of a particular type.

■ **DEFINITION 11.4.4**

Knowledge that is based on experience is called ***a posteriori***.

The term *a posteriori* refers to the knowledge coming *after* having experienced the state of affairs described by the sentence. If knowledge comes *before* experience,

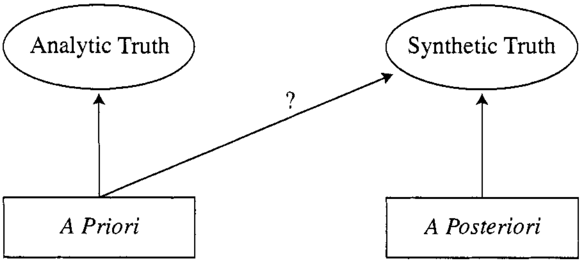

Figure 11.6 *A priori* and *a posteriori* knowledge.

such as that gained from the sentence *a bachelor is an unmarried man*, then it is not *a posteriori*. We do not need to meet a bachelor or have any other type of experience involving bachelors to know that the sentence is true. This represents the second type of knowledge.

■ **DEFINITION 11.4.5**

Knowledge that is not based on experience is called ***a priori***.

Based on our examples we may summarize by stating that we have *a posteriori* knowledge of some synthetic truths and we have *a priori* knowledge of some analytic truths. Can we push this further? Certainly all analytic propositions are known *a priori*, and all *a posteriori* knowledge is synthetic (Figure 11.6). However, is it possible to know a synthetic proposition *a priori*? Our answer to this question will put us into one of two mutually exclusive camps.

■ **DEFINITION 11.4.6**

A **rationalist** believes that there exists a true proposition about the world that can be known without either direct or indirect experience of it.

A rationalist does not deny that many propositions that describe the world accurately are known by observation. The claim is that there is at least one proposition that says something true about the world, yet we determine its truth simply by analyzing the terms of the sentence.

■ **DEFINITION 11.4.7**

An **empiricist** believes that the only way to know a true proposition about the world is by either direct or indirect experience of it.

The empiricist believes that there is no sentence that gives us new information about the world, yet we know it *a priori*. Only experiment or some other type of observation can lead to knowledge of the universe.

It is not uncommon for a scientist to speak as if the only way to learn about the universe is through the scientific method. All other modes of discovery are either analytic or superstition. It appears that scientists are empiricists. The rationalist will counter that this is a hasty generalization. There are synthetic *a priori* truths, and examples can be found in mathematics. This is what Kant believed.

Intuition

Mathematics served as Kant's main example of knowledge that is both *a priori* and synthetic. For example, consider Euclid's theorem regarding the sum of the interior angles of a triangle being equal to two right angles. Kant believed that this proposition is synthetic. Simply analyzing the meanings of the terms of the sentence will not yield whether the statement is true or not. We understand what *interior*, *angle*, and *triangle* mean, but none of their meanings involve *sum*, *two*, or *right*. We must arrive at the conclusion by some other way (Shabel 2006, 98).

The way to arrive at this conclusion is a method that Kant believed is unique to mathematics. Consider first this nonmathematical example. When we look at a tree, the light rays reflect off its surface and hit our eyes, which send signals along our optic nerves to our brains. We then carry the resulting image of the tree with us mentally. Kant had a name for this.

■ **DEFINITION 11.4.8**

An **intuition** is the immediate representation of an object in one's mind. An intuition that is caused by an object acting upon us in some way is called an **empirical intuition**, and one that is independent of any sensation is called a **pure intuition**.

The mental picture of the tree is an example of an empirical intuition. If we read the definition of a triangle, we see in our mind what it looks like based on the definition without needing a picture of the triangle. This is a pure intuition. From these intuitions we derive concepts.

■ **DEFINITION 11.4.9**

A **concept** is an abstract generalization of an intuition without the influence of the associated object.

We take our mental picture of the tree or the triangle and from that picture alone we create categories into which we place our intuitions. When we see other trees, new intuitions form in our minds, and the concepts that we have generalized from other intuitions allow us to group like intuitions under the same categories (Kant 1998, A19/B33, 155; cf. 36). This is illustrated in Figures 11.7 and 11.8.

Kant pulls this together to describe the very unique job of mathematicians. They arrive at truth by way of construction. According to Kant (1998, A713/B741, 630), "to **construct** a concept means to exhibit *a priori* the intuition corresponding to it." This means that if we take a mathematical concept, that of a triangle for instance, that

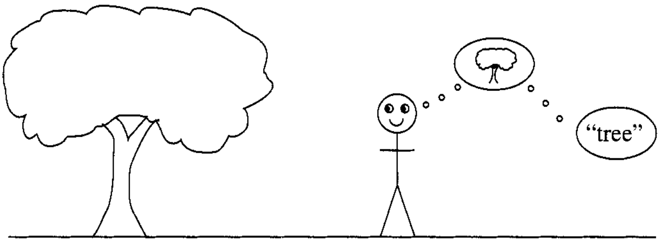

Figure 11.7 An empirical intuition leads to a concept.

is represented in our minds as an intuition, to construct that concept we must explain, illustrate, or describe the intuition in some way. The result is called a **construction**. Think of it as a picture that represents the concept. It can either be in one's mind or on a piece of paper. Either way, a construction is a generalization derived from the concept. For example, the construction of a triangle on paper is drawn in such a way that the lengths of its sides and the measures of its angles are irrelevant. What is important in the picture are the general properties of the triangle (Kant 1998, A 713–A 714/B 741–B 742, 630).

It is such constructions that make mathematical propositions synthetic. We see this in Kant's comparison between the philosopher and the mathematician. Consider the geometric proposition which states that the angle sum of a triangle equals two right angles. Philosophers will proceed by analysis. They will examine the meanings of *angle* and *side* and *vertex* and attempt to find within these terms a nugget of truth that will enable them to conclude the result about the angle sum. They will not find it. The mathematicians will find it because they will look beyond the definition. Their mental picture tells them that if through a vertex they draw a line parallel to the opposite side, they will see the angles of the triangle lying along the parallel at the chosen vertex. The truth about the angle sum is a synthetic one. The sentence gives us information about the triangle that goes beyond its definition, and it is the construction that shows this. Kant argued that this not only works for geometry but also for arithmetic. For example, the sentence *one plus one equals two* is synthetic because it tells us information about *one*, *two*, *plus*, and *equals* that go beyond their

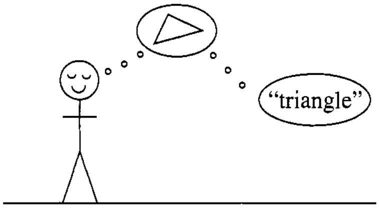

Figure 11.8 Pure intuition leads to a concept.

meanings. We then generalize to basic rules based on our intuition of numbers and addition. Because all of mathematics is either the study of the continuous (geometry) or the discrete (arithmetic), all mathematical propositions are synthetic (Shabel 2006, 101, 106–107; cf. Kant 1998, A 716/B 744, 631–632).

Now Kant asserts that mathematical knowledge is *a priori*. The example will again be the angle sum theorem for triangles. He claims that the proposition has two properties that make it clear that it entails *a priori* knowledge. They are the statement's necessity and universality. Kant claims that the angle sum result must be true because of its proof. The geometric demonstration does not depend on conclusions that are based on experimental evidence that later could be proven wrong. Instead, the proof transcends time and gives the geometer complete certainty of the proposition's truth. Necessity is not a quality that is determined *a posteriori*. We cannot determine whether a statement must be true by experiment alone. As for the fact that the sentence holds for all triangles at all times follows from the very act of construction. Indeed, the triangle constructed is a particular instance of the term, so the construction is particular, not universal. The algorithm that leads to the construction flows from the concept of triangle, and it is the same each time an arbitrary triangle is needed. This universality is also something that our senses cannot give us. When we try to generalize based on multiple observations, there is always at least some doubt that the next observation will not uphold our conclusion. Both necessity and universality are characteristics of *a priori* knowledge and not *a posteriori*. Because arithmetic propositions also have these two characteristics, Kant concluded that all of the true propositions of mathematics are instances of synthetic *a priori* knowledge (Shabel 2006, 107–109).

We may now define a **mathematical concept** as a synthetic definition that gives us the rule for describing the associated intuition. The definition of a triangle is an example. It tells us that to construct a triangle we need three noncollinear line segments joined at their endpoints. We now know how to construct it. This is an example of a specific type of mathematical concept called a **geometric concept**. Kant taught that geometric concepts give us a pattern for constructing space (Shabel 2006, 99). This means that since geometry is *a priori* and necessary, the same can be said about space. It is a pure intuition. It is not something that exists outside ourselves but originates in our minds as a fundamental framework to put meaning onto the input from our senses. Kant held this position against two prevailing views in which space exists in an objective sense either as the final container of objects or as an abstraction from the relations among the objects. He regarded the first view as untenable because it would make space into some kind of substance that was infinite in extent, and only God can be said to have such a quality. Moreover, Kant viewed God as existing outside space and time, and this appears to be impossible if the universe is infinite in size. Kant rejected the second view because it was not consistent with geometry being both necessarily true and an accurate description of the world. If space is abstracted from what we see, the structure of space is discovered by empirical means. This would mean that geometry is not *a priori* as Kant claimed.

To support his view, Kant notes that for us space is assumed by our minds when we use our senses. The standard example is that of a pair of hands. When we examine them, they are from our point of view essentially identical. To our eyes they look the same. If we measure the lengths of the fingers, the distance around the wrists, and so on, we find that they have the same dimensions. How, then, do we tell the left hand from the right? We are able to do this, according to Kant, because our spatial intuition comes preloaded with such concepts as up, down, in front, behind, and to the side. These *a priori* properties of space enable us to make the distinction. That this ability is universal in humans is further support for Kant's view. Since space is neither an object that exists independent of ourselves nor an abstraction based on observation, there are no competing realities when it comes to space. Everyone has the same intuition about its nature, and that intuition is described by Euclid's geometry. The mathematics that describe the universe is common to all people because everyone has the same intuition about space (Hatfield 2006, 73–78).

CHAPTER 12

NICOLAI LOBACHEVSKI AND JÁNOS BOLYAI

After he left Göttingen in 1799, Farkas Bolyai moved to the Hungarian city of Kolozsvár, where he began work as a tutor. He met Zsuzsanna Árkosi Benkő, and they soon were wed. The couple moved to Domáld but returned to Kolozsvár in the fall of 1802 expecting their first child. The baby was a boy, and they named him János (Prékopa 2006, 11).

János was a child prodigy. When he was 6, he taught himself to read. When he was 7 he learned German and how to play the violin. His father began to teach him mathematics when he was 9. He began formal schooling when he was 12, skipping the first three grades. János could easily solve calculus problems by the time he was 14. Passing his final exam in 1817 when he was 15 years old, he was ready to attend a university. The problem was that there were none in the local vicinity, and the closest ones, the University of Pest and the University of Vienna, did not have the caliber of mathematics faculty necessary to further János' education. The obvious solution was to send him to Gauss at Göttingen, so Farkas wrote his friend. At such a young age János would need proper supervision, so he inquired of Gauss whether his son might stay with him during his time at the university. Unfortunately, the letter could have been more tactfully written, for Farkas wanted to confirm that the Gauss household provided the proper environment for his son. It must be that Gauss was

Revolutions of Geometry. By Michael O'Leary
Copyright © 2010 John Wiley & Sons, Inc.

not "grumbling," his wife "exceptional," and his daughter not "dangerous." It is not surprising that Gauss did not answer the letter. It is a testament, however, to their friendship that this incident did not prevent future amicable correspondence (Prékopa 2006, 14–15).

So instead János attended the Vienna Academy of Military Engineers. Due to some financial problems, he enrolled a year later, in 1818. He excelled in his studies, and János was ranked first in his class by his professors. There was also a ranking based on student input in which he ranked second in his class. János did not enjoy drawing and his performance reflected this. This was evidently enough to keep him from first place. However, what he did enjoy was mathematics. Like his father he excelled at it, and like his father he was drawn to the study of parallel lines. Beginning in 1820 János researched the field thoroughly. His goal was to prove the Parallel Postulate once and for all. Learning of his son's intentions, Farkas wrote to him in a letter dated April 4, 1820 (Prékopa 2006, 15–16):

> You must not attempt this approach to parallels: I know this way to its very end. I have traversed this bottomless night, which extinguished all light and joy of my life. For God's sake! I entreat you leave parallels alone, abhor them like indecent talk, they may deprive you [just like me] from your time, health, tranquility and the happiness of your life. That bottomless darkness may devour a thousand tall towers of Newton and it will never bright up in the earth. ... I thought I would sacrifice myself for the sake of the truth. I was ready to become a martyr who would remove the flaw from geometry and return it purified to mankind. I accomplished monstrous, enormous labors; my creations are far better than those of others and yet I have not achieved complete satisfaction. ... I turned back when I saw that no man can reach the bottom of this night. I turned back unconsoled, pitying myself and all mankind. ... I have traveled past all reefs of this infernal Dead Sea and have always come back with broken mast and torn sail. The ruin of my disposition and my fall date back to this time. I thoughtlessly risked my life and happiness — *aut Caesar aut nihil.*

Evidently, Farkas's letter inspired János. If the Parallel Postulate could not be proven, then what? It should be replaced by something else, not with a statement that is equivalent to it because that would simply lead to all of the same problems that his father had encountered. Instead, the Parallel Postulate should be replaced by an assumption that would imply the Postulate's negation. He would continue his research. In 1822 János finished his studies at the Academy, but because he was one of their top two students, the school allowed him to stay an extra year. When he left Vienna after his extension, János was advanced to the position of sub-lieutenant and was assigned the position of Directorate of Fortification of Temesvár. It was there where things fell together. On November 3, 1823 he wrote to his father:

> I am resolved to publish a work on parallels as soon as I can complete and arrange the material, and the opportunity arises. At the moment I still do not clearly see my way through, but the path which I have followed is almost certain to lead me to my goal, provided it is at all possible. I have not quite reached it, but I have discovered things so wonderful that I was astounded and it would be an everlasting pity if these things were lost. When you, my dear father, see them, you will understand. All I can say at present is that out of nothing I have created a strange new world. All that I have sent you previously is like a house of cards in comparison with a tower. (Burton 1985, 554)

Possibly reflecting on the discovery of analytic geometry by Descartes and Fermat and that of the calculus by Newton and Leibnitz, Farkas wrote to János:

> It seems to me advisable, if you have actually succeeded in obtaining a solution of the problem, that, for a two-fold reason its publication be hastened: first, because ideas easily pass from one to another who, in that case, can publish them; secondly, because it seems to be true that many things have, as it were, an epoch in which they are discovered in several places simultaneously, just as the violets appear on all sides in the springtime. (Burton 1985, 554)

In 1825 János sent an abstract to his father, and then finally in 1829 he sent a copy of the work. Farkas was pleased but concerned about one point. An indeterminant constant appeared in his formula, and he did not understand why his son could not eliminate it. Nonetheless, the two came to agreement, and János's work would be appended to Farkas's *Essay on the Elements of Mathematics for Studious Youths*. The full title of János's text in Latin is *Appendix scientiam spatii absolute veram exhibens: a veritate aut falsitate Axiomatis XI. Euclidei, a priori haud unquam decidenda, independentem: adjecta ad casum falsitatis quadratura circuli geometrica*. This is shortened in English to *The Science of Absolute Space*, and it is often referenced simply as the *Appendix*. It was was sent to Gauss in 1831, but it was lost in transit. The *Appendix* was sent a second time in January 1832 (Bonola 1912, 99–100).

Dated March 6, 1832, the letter that Gauss sent in reply to Farkas is one of the most famous ever penned in the history of mathematics. As we have seen, Gauss studied the parallels question, and he came to many of the same results as others had. The difference was that he had kept his results private. Thus, we are not surprised by what Gauss wrote to his old colleague from Göttingen.

> If I commenced by saying that I *am unable to praise this work* (by Johann), you would certainly be surprised for a moment. But I cannot say otherwise. To praise it, would be to praise myself. Indeed the whole contents of the work, the path taken by your son, the results to which he is led, coincide almost entirely with my meditations, which have occupied my mind partly for the last thirty or thirty-five years. So I remained quite stupefied. So far as my own work is concerned, of which up till now I have put little on paper, my intention was not to let it be published during my lifetime. Indeed the majority of people have not clear ideas upon the questions of which we are speaking, and I have found very few people who could regard with any special interest what I communicated to them on this subject. To be able to take such an interest it is first of all necessary to have devoted careful thought to the real nature of what is wanted and upon this matter almost all are most uncertain. On the other hand it was my idea to write down all this later so that at least it should not perish with me. It is therefore a pleasant surprise for me that I am spared this trouble, and I am very glad that it is just the son of my old friend, who take the precedence of me in such a remarkable manner. (Bonola 1912, 100)

Farkas was pleased with the letter and viewed it as an honor. János, however, did not. He was devastated. He initially suspected that his father had secretly informed Gauss of his work and now Gauss was using that communication to upstage him. This could not have been further from the truth, because Farkas was very proud of his son and wanted only the best for him. This we have seen in Farkas's letters.

Nonetheless, the event strained the relationship between the two, and it never fully healed. It took nearly 10 years before János eventually accepted the truth of the situation, but he would never come to fully trust Gauss. Scholars dispute whether this distrust was justified. Gauss did work on the question of parallels, but whether his results rose to the level of János' work is left to various interpretations (Bonola 1912, 101; Gray 2004, 54–55).

János's acceptance of the facts did not mean that the letter did not have lasting effects. He would never publish again, although he did continue mathematical research. Making matters worse was the fact that the *Appendix* was ignored by the mathematical community, and then in 1848 János discovered a German edition of a treatise written by a Russian mathematician. Again he initially believed that the work must have been stolen since the results were so similar. They both used Saccheri's definition of parallel and proved the same basic results. They both examined spherical geometry, discovered the equation for the angle of parallelism, and constructed the same geometrical objects. There were some differences. The Russian tended to be more analytical, while Bolyai tended to focus on making clear the distinction between those theorems that followed from the negation of the Parallel Postulate and those that can be proven without any kind of parallel assumption (Burton 1985, 555; Prékopa 2006, 19).

The Russian's name was Nicolai Ivanovich Lobachevski. He was born on December 1, 1792 in the Russian city of Nizhny Novgorod. He was the son of Ivan Maksimovich Lobachevski, a government clerk, and his wife, Praskovia Aleksandrovna Lobachevskaya. Ivan died when the boy was only 7, so without an income, Praskovia moved to Kazan. The city was remote, near the Siberian border, and it had a large Muslim population. The benefit to moving to the city was that the Russian government supported Kazan both financially and educationally in an effort to spread Russian culture to the population. With the aid of a government scholarship, Lobachevski attended school. He was a bright student but not always interested in his studies. His teachers recognized in him an ability for mathematics, so they worked with him and helped him graduate.

With a goal of studying medicine, Lobachevski entered Kazan University in 1807. The university also benefited from the government's policies. They recruited professors to Kazan from Europe with the goal of creating a top center of Russian learning. One of these new faculty members was Martin Bartels from Caroline College. Lobachevski came into contact with him and soon abandoned medicine for mathematics. In 1811 he received his master's degree in mathematics and physics, and two years later he joined the university faculty teaching courses in his master's fields and astronomy. In 1816, at the age of 33, Lobachevski was promoted to the rank of full professor (Burton 1985, 557; Boyer and Merzbach 1991, 520).

Bartels and the other mathematics faculty at the University of Kazan educated Lobachevski in the German tradition. This meant that geometry played an important role in his education. He carried this influence into his professional life, often teaching the subject. One day in 1823 he wrote an outline for his students. He noted for his class that no one to date had found a proof of the Parallel Postulate. Such a note provides evidence that he was thinking about the subject, and as with every other interested mathematician, he must have thought about proving it. Three years later he presented a paper at his university in which he elaborated many of the results that follow when the Parallel Postulate is assumed to be false. Then in 1829 he did something that Gauss would not do. Convinced that it was impossible to prove the Parallel Postulate, he published his research on parallels as a new system of geometry with the Hypothesis of the Acute Angle as a postulate. It appeared in the *Kazan Messenger*, and it was titled "On the Principles of Geometry." He realized that the results were not in line with the mathematical community. Nonetheless, he believed that the theorems that flowed from the new postulate set formed a consistent logical system. Lobachevski called his new system **imaginary geometry** (Boyer and Merzbach 1991, 520–521).

Gauss appears to have been correct concerning the Boeotians. When Lobachevski submitted his 1829 paper, the referee, the person assigned to read the work and comment on its value and correctness, did not understand Lobachevski's ideas and commented negatively. Nonetheless, the paper still went to publication. Reviews were not favorable. One mathematician suggested that Lobachevski had titled his piece incorrectly. Instead of *On the Principles of Geometry*, it should have been named either *A Satire on Geometry* or *A Caricature of Geometry*. Lobachevski was not discouraged and continued to publish his research on his new imaginary geometry in 1835 with the *New Foundations of Geometry*, written in Russian. In 1840 he completed his *Geometrical Investigations on the Theory of Parallels*, written in German (Rosenfeld 1988, 208–209; Boyer and Merzbach 1991, 521). In it Lobachevski opens by writing:

> In geometry I find certain imperfections which I hold to be the reason why this science, apart from transition into analytics, can as yet make no advance from that state in which it has come to us from Euclid.
>
> As belonging to these imperfections, I consider the obscurity in the fundamental concepts of the geometrical magnitudes and in the manner and method of representing the measuring of these magnitudes, and finally the momentous gap in the theory of parallels, to fill which all efforts of mathematics have been so far in vain. (Lobachevski 1914, 11)

All of the efforts of the past had not solved the problem of parallels. Some were close, but none were able to take the last step. Lobachevski would be the first mathematician to publicly accept and detail results that would rival Euclid and revolutionize geometry.

12.1 PARALLEL FUNDAMENTALS

To begin our survey of the geometric work of Lobachevski and Bolyai, we start by proving some basic facts about parallel lines. Using Lobachevski's diagram (Figure 12.1), let $\overline{AN}$ be parallel to $\overline{BY}$ and let $\overline{AB}$ be perpendicular to $\overline{BY}$. By Theorem 10.5.2 acute $\sphericalangle BAN$ has the property that if $\overline{AK}$ is in the interior of $\sphericalangle BAN$, it intersects $\overline{BY}$, but if $\overline{AK}$ is in the interior of $\sphericalangle NAM$, it will not meet $\overline{BY}$. The point A is said to have the **characteristic of parallelism** for $\overline{AN}$ and $\overline{BY}$ because $\sphericalangle BAN$ has these properties. We should note that there is a mirror image of this

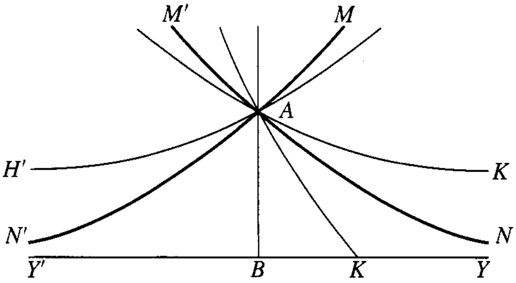

Figure 12.1 The parallel lines of Lobachevski.

situation on the other side of $\overline{AB}$ and A is not the only point on $\overline{AN}$ that has the characteristic of parallelism.

■ **THEOREM 12.1.1** [*Theory of Parallels* 17, *Appendix* 2]

[HA] If $\overline{AB}$ is parallel to $\overline{CD}$, every point on $\overline{AB}$ has the characteristic of parallelism.

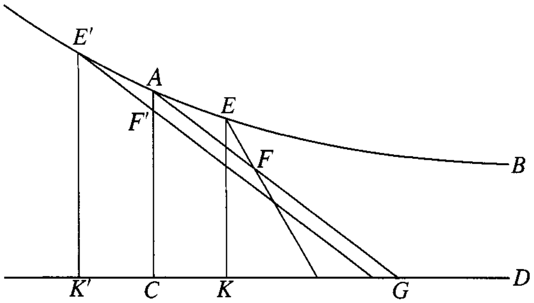

PROOF

Let $\overline{AB}$ be parallel to $\overline{CD}$ and let A have the characteristic of parallelism for $\overline{AB}$ and $\overline{CD}$. Take E to the right of A in the diagram making $\sphericalangle CAE$ acute. Find K on $\overline{CD}$ so that $\overline{EK}$ is perpendicular to $\overline{CD}$. Pick F in the interior of $\sphericalangle BEK$ and join both $\overline{EF}$ and $\overline{AF}$. Since $\overline{AF}$ is in the interior of m$\sphericalangle CAB$, we may conclude that $\overline{AF}$ will intersect $\overline{CD}$ at some G when extended. This forms $\triangle ACG$, so by Pasch's Postulate (11.3.2), $\overline{EF}$ produced must meet $\overline{CD}$.

Next, let E' be on $\overline{AB}$ but on the other side of $\overline{AC}$ and let K' be on the $\overline{CD}$ produced such that $\overline{E'K'}$ is perpendicular to $\overline{CD}$. Choose F' on $\overline{AC}$. Since A has the characteristic of parallelism, we can find G on $\overline{CD}$ such that $\sphericalangle BAG \cong \sphericalangle AE'F'$. We now have $\triangle ACG$, in which $\overline{E'F'}$ enters through $\overline{AC}$. If $\overline{E'F'}$ intersects $\overline{AG}$, we would have a triangle with angle sum greater than two right angles because $\sphericalangle E'AG$ and $\sphericalangle AE'F'$ are supplements. Hence, it must be the case that $\overline{E'F'}$ can be extended to meet $\overline{CD}$. ■

In the next two theorems we prove that parallel lines under the new definition share some common characteristics with Euclid's parallel lines. The first result states that lines are parallel is a symmetric relation. Lobachevski's proof relies on a continuity argument.

■ **THEOREM 12.1.2** [*Theory of Parallels* 18, *Appendix* 6]

[HA] If $\overline{AB}$ is parallel to $\overline{CD}$, then $\overline{CD}$ is parallel to $\overline{AB}$.

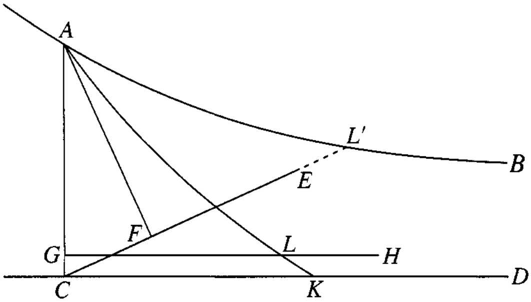

PROOF

Suppose that $\overline{AB}$ is parallel to $\overline{CD}$ with $\overline{AC}$ perpendicular to $\overline{CD}$. Because $\overline{AB}$ and $\overline{CD}$ do not intersect, to prove that $\overline{CD}$ is parallel to $\overline{AC}$, it suffices to show that any line in the interior of $\sphericalangle ACD$ must intersect $\overline{AB}$. To do this, let $\overline{CE}$ form the acute $\sphericalangle ACE$ in the interior of $\sphericalangle ACD$. Construct a perpendicular to $\overline{CE}$ through A. Let F be its point of intersection with $\overline{CE}$. The result is right $\triangle AFC$ with $AC > AF$. Thus, we can find G on $\overline{AC}$ so that $\overline{AG} \cong \overline{AF}$. Now construct $\overline{GH}$ perpendicular to $\overline{AC}$ and rotate $BAFE$ with pivot A so that $\overline{AF}$ and $\overline{AG}$ line up. This is equivalent to taking K on $\overline{CD}$ such that

$$\sphericalangle GAK \cong \sphericalangle FAB.$$

We can find such a K since $\overline{AB}$ is parallel to $\overline{CD}$. Let L be the intersection of $\overline{AK}$ and $\overline{GH}$. Find L' on $\overline{CE}$ (produced if necessary) such that $\overline{FL'} \cong \overline{GL}$. Because $\overline{AG} \cong \overline{AF}$, $\sphericalangle AGL \cong \sphericalangle AFL'$, and $\overline{GL} \cong \overline{FL'}$, we conclude by SAS that

$$\triangle AGL \cong \triangle AFL'.$$

Hence, $\sphericalangle FAL' \cong \sphericalangle GAL$. Since $\sphericalangle GAL$ and $\sphericalangle GAK$ are the same angle, we conclude that $\sphericalangle FAL' \cong \sphericalangle FAB$. The only way for this to happen is if L' is on $\overline{AB}$. Therefore, $\overline{CD}$ is parallel to $\overline{AB}$. ■

Being parallel is not transitive, as noted in Exercise 2. The next theorem is as close as we can get.

■ **THEOREM 12.1.3** [*Theory of Parallels* 25]

[HA] If two distinct lines are parallel to a third, they are parallel to each other.

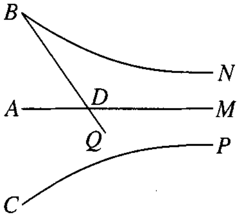

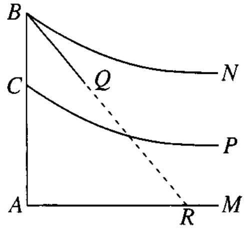

PROOF

Let both $\overline{BN}$ and $\overline{CP}$ be parallel to $\overline{AM}$ with C not on $\overline{BN}$. By Exercise 4, $\overline{BN}$ and $\overline{CP}$ do not intersect. We now have two cases to consider. First let $\overline{AM}$ be between $\overline{BN}$ and $\overline{CP}$. Take point Q and form $\sphericalangle NBQ$ toward $\overline{AM}$, so $\overline{BQ}$ and $\overline{AM}$ must intersect at some point D. Since $\overline{AM}$ is parallel to $\overline{CP}$ (Theorem 12.1.2), $\overline{DQ}$ will meet $\overline{CP}$ since all of the points of $\overline{AM}$ have the characteristic of parallelism (Theorem 12.1.1). Thus, $\overline{BN}$ is parallel to $\overline{CP}$.

Next let $\overline{AM}$ not be between the other two. Without loss of generalization we may assume that $\overline{CP}$ is between $\overline{BN}$ and $\overline{AM}$ and for clarification in the diagram assume that B, C, and A are collinear. Let $\overline{BQ}$ form an angle within $\sphericalangle NBA$. Since $\overline{BN}$ is parallel to $\overline{AM}$, we can produce $\overline{BQ}$ to intersect $\overline{AM}$ at some R. Now notice that $\overline{CP}$, $\overline{CA}$, and $\overline{AM}$ form an omega triangle (Definition 11.3.1). Since $\overline{BR}$ already intersects one of its sides, by Theorem 11.3.5, we conclude that $\overline{BQ}$ must intersect $\overline{CP}$. ■

The Angle of Parallelism

Let $\overline{AB}$ be perpendicular to $\overline{BY}$ and assume that $\overline{AN}$ is parallel to $\overline{BY}$ (Figure 12.2). Let $p = AB$. Since the measure of the angle of parallelism (page 351) is dependent only on the value of p, we will define

$$\Pi : [0, \infty) \to [0, \pi/2]$$

by

$$\Pi(p) = \begin{cases} \mathrm{m}\sphericalangle BAN, & p > 0 \\ \pi/2, & p = 0, \end{cases}$$

where we define $\Pi(0) = \pi/2$ because

$$\lim_{p \to 0+} \Pi(p) = \pi/2 \tag{12.1}$$

by Exercise 7. Furthermore, if we let H be a point and set $\alpha = \mathrm{m}\sphericalangle BAH$, we have:

- If $0 \leqslant \alpha < \Pi(p)$, the $\overline{AH}$ produced intersects $\overline{BY}$.
- If $\alpha = \Pi(p)$, then $\overline{AH}$ is parallel to $\overline{BY}$.
- If $\Pi(p) < \alpha \leqslant \pi/2$, then $\overline{AH}$ will not intersect $\overline{BY}$.

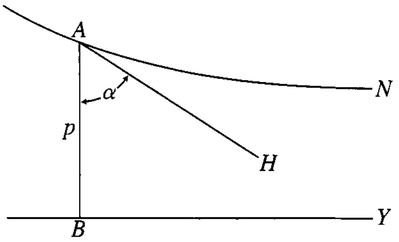

Figure 12.2 $\sphericalangle BAN$ is the angle of parallelism.

If $\alpha > \pi/2$, view the line as on the opposite side of $\overline{AB}$ and treat accordingly (Lobachevski 1914, 13–14). It is important to note that **HR** implies that $\Pi(p) = \pi/2$ for all p. However, because the work of Lobachevski and Bolyai focused only on the **HA** case, we assume this hypothesis for the rest of the chapter.

The function Π has many nice properties that we now prove. The proof of the first result uses omega triangles.

■ THEOREM 12.1.4

[**HA**] Π is a strictly decreasing function.

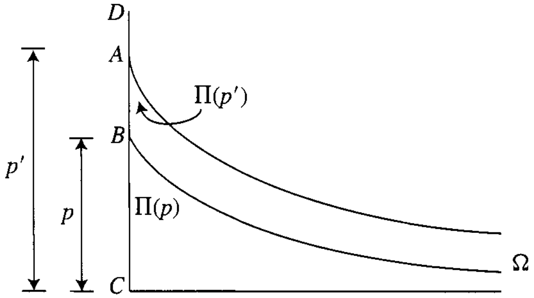

PROOF

Let $\overline{CD}$ be perpendicular to $\overline{C\Omega}$. Take A and B on $\overline{CD}$ and write $p = BC$ and $p' = AC$. To see that Π is strictly decreasing, suppose that $p < p'$. Remember that $\overline{A\Omega}$, $\overline{B\Omega}$, and $\overline{C\Omega}$ are all parallel to each other. Then $m\sphericalangle CB\Omega = \Pi(p)$ and $m\sphericalangle CA\Omega = \Pi(p')$. We observe that $\sphericalangle DA\Omega$ is exterior to $\triangle AB\Omega$. Therefore, by Theorem 11.3.6,

$$m\sphericalangle AB\Omega < m\sphericalangle DA\Omega.$$

This gives

$$\pi - m\sphericalangle AB\Omega > \pi - m\sphericalangle DA\Omega,$$

from which we see that $\Pi(p) > \Pi(p')$, so Π is strictly decreasing. ■

■ COROLLARY 12.1.5

[**HA**] Π is one-to-one.

By Theorem 10.5.2 we know that to every magnitude there is associated an angle of parallelism. The next theorem states that the converse is also true, so we now know that Π is a one-to-one correspondence.

■ THEOREM 12.1.6 [*Theory of Parallels* 23]

[**HA**] Π is onto.

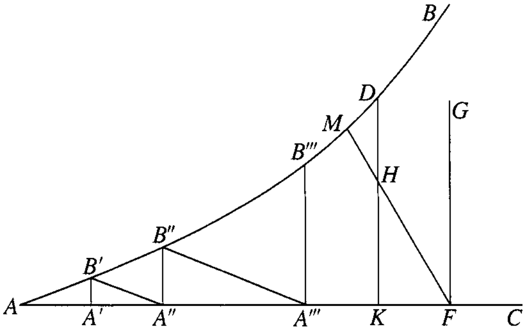

PROOF

Take α such that $0 \leqslant \alpha \leqslant \pi/2$. We must find p so that $\Pi(p) = \alpha$. Let $\text{m}\sphericalangle BAC = \alpha$. Take an arbitrary point B' on $\overline{AB}$ and drop a perpendicular to $\overline{AC}$, meeting it in A'. Find A'' on $\overline{A'C}$ such that $\overline{A'A''} \cong \overline{AA'}$. At A'' erect a perpendicular that intersects $\overline{AB}$ at B''. Continuing this process, we will find $\overline{FG}$ that is perpendicular to $\overline{AC}$ but does not intersect the $\overline{AB}$ produced. We will prove this fact indirectly, so assume that no such perpendicular exists. Appealing to Theorem 10.4.5, we know that

$$\text{d}(AB'A'') = \text{d}(AB'A') + \text{d}(A''B'A').$$

Because $\triangle AB'A' \cong \triangle A''B'A'$ by SAS, $\text{d}(AB'A'') = 2\text{d}(AB'A')$. Therefore,

$$\text{d}(AB''A'') > 2\text{d}(AB'A')$$

by Theorem 10.4.5. Next find A''' on $\overline{AC}$ and B''' on $\overline{AB}$ such that $\overline{A''A'''}$ is congruent to $\overline{AA''}$ and $\overline{A'''B'''}$ is perpendicular to $\overline{AC}$. As before we have $\triangle AB''A'' \cong \triangle A'''B''A''$, and we conclude that

$$\text{d}(AB'''A''') > 2^2\text{d}(AB'A').$$

We can repeat this construction and find for each positive integer n an ordinary triangle with defect greater than $2^n\text{d}(AA'B')$. That is, we can find a triangle with angle sum less than $\pi - 2^n\text{d}(AA'B')$ for each n, which means that we can find a triangle with angle sum less than or equal to zero, which is impossible.

Without loss of generality we may assume that $\overline{FG}$ is the first perpendicular that does not intersect $\overline{AB}$. Its existence is guaranteed by an argument similar to that for the existence of Saccheri's limit (Theorem 10.5.2 and page 351). To prove that $\overline{AB}$ is parallel to $\overline{FG}$, it suffices to show that $\overline{FG}$ is parallel to $\overline{AB}$ (Theorem 12.1.2). To do this take H in the interior of $\sphericalangle AFG$. Choose K on $\overline{AC}$ so that $\overline{HK}$ is perpendicular to $\overline{AC}$. Since $\overline{FG}$ is the first not to intersect $\overline{AB}$, there exists D on $\overline{AB}$ where the $\overline{KH}$ produced intersects $\overline{AB}$. We obtain $\triangle AKD$. Since $\overline{FH}$ enters this triangle at $\overline{DK}$, it must exit through $\overline{AD}$ at some point M by Pasch's Postulate (11.3.2). Thus, $\overline{FG}$ is a parallel to $\overline{AB}$, and we may set $p = AF$. ■

All of this allows us to conclude that Π is also continuous (Wolfe 1945, 76–77; Lobachevski 1914, 21). Assuming **HR**, Π is a constant function, so it is continuous. To prove the result under **HA**, suppose that Π is not continuous at p_0. This means that

$$\Pi(p_0) \neq \lim_{p \to p_0} \Pi(p),$$

and this means that one of the one-sided limits does not equal $\Pi(p_0)$. Suppose that it is $p \to p_0^-$. Then there exists $\epsilon > 0$ such that for all x where $0 < x < p_0$,

$$\Pi(x) - \Pi(p_0) > \epsilon.$$

From this we conclude that no angle measure θ such that $\Pi(p_0) < \theta < \Pi(p_0) + \epsilon$ corresponds to a segment length since Π is strictly decreasing. This contradicts Theorem 12.1.6. Therefore, we have:

■ **THEOREM 12.1.7**

Π is continuous.

Exercises

1. Let $\overline{AX}$ be parallel to $\overline{BY}$ and $\overline{BY}$ be parallel to $\overline{CZ}$. Is it possible for $\overline{AD}$ to not intersect $\overline{BY}$ but intersect $\overline{CZ}$. Explain.

2. Is it possible for l to be parallel to m, m parallel to n, but l not parallel to n?

3. Use omega triangles to prove Theorem 12.1.2 and reprove the first case of the proof for Theorem 12.1.3.

4. Prove that $\overline{BN}$ and $\overline{CP}$ do not intersect in the proof of Theorem 12.1.3.

5. Generalize the definition of two lines being parallel to the three dimensions.

6. Let l, m, and n be distinct lines in three dimensions. Assume that l and m are parallel to n in space. Using the definition from Exercise 5, demonstrate that n is parallel to both l and m and l is parallel to m.

7. Show $\lim_{p \to 0^+} \Pi(p) = \pi/2$.

8. Explain why AA (Theorem 11.3.8) implies that Π is not a constant function.

9. True or false: There exists a positive real number p such that $\Pi(p) = \pi/2$ if and only if **HR**. Explain your reasoning.

10. Let l be a line and P a point off l. True or false: There exists exactly one line n through P that is parallel to l if and only if **HR**. Explain your reasoning.

11. Let A not be on l. Let $\overline{AN}$ and $\overline{AN'}$ both be parallel to l as in Figure 12.1. Prove that the angle bisector of $\sphericalangle N'AN$ is perpendicular to l.

12. Suppose that X and Y are on the same side of $\overline{AB}$. Let $\sphericalangle BAX$ and $\sphericalangle ABY$ be supplementary. Prove that $\overline{AX}$ is not parallel to $\overline{BY}$ and that the two lines do not intersect.

12.2 HOROCYCLES

In 1804, Farkas Bolyai sent Gauss what he thought was a proof of the Parallel Postulate. Gauss showed him his error in a later letter, but Farkas was nonetheless still hopeful of finding a proof and continued to study the subject (Kline 1972, 878). One of his attempts was his proof that the Parallel Postulate follows from the assumption

$$\text{three points that are not collinear lie on a circle.} \tag{12.2}$$

To prove this, take $\overline{AB}$ and points A' and B' on the same side of $\overline{AB}$ such that $\overline{BB'}$ and $\overline{AB}$ are perpendicular and $\sphericalangle BAA'$ is acute (Figure 12.3). We must show that $\overline{AA'}$ and $\overline{BB'}$ intersect on the same side of $\overline{AB}$ as both A' and B'. Take any point M on $\overline{AB}$. Find M' on the opposite side of $\overline{AA'}$ such that $\overline{AA'}$ is the perpendicular bisector of $\overline{MM'}$. Similarly, let M'' be on $\overline{AB}$ produced so that $\overline{BB'}$ is the perpendicular bisector of $\overline{MM''}$. We observe that M, M', and M'' are not collinear. Hence, they lie on a circle by Proposition 12.2. Both $\overline{AA'}$ and $\overline{BB'}$ are concurrent with radii of the circle, so they must intersect at the circle's center (Bonola 1912, 61).

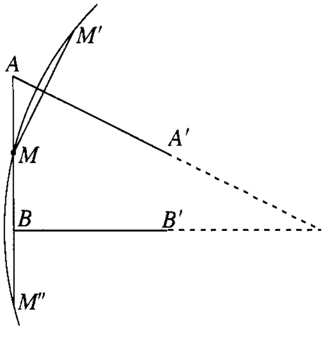

Figure 12.3 Farkas Bolyai's proof of the Parallel Postulate.

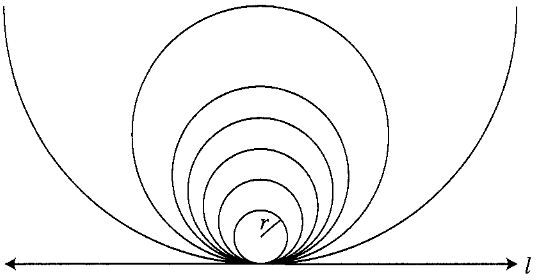

Figure 12.4 A circle with radius approaching infinity.

The problem with Farkas Bolyai's approach is that Proposition 12.2 is equivalent to the Parallel Postulate (Exercise 1). However, this result did inspire a student of Gauss by the name of Friedrich Ludwig Wachter. He also attempted to prove the Parallel Postulate. Despite believing otherwise, Wachter was not successful. In a letter written in 1816 to Gauss he described his efforts. He wrote concerning what he believed would happen when the radius of a sphere was allowed to approach infinity. Take a circle with radius r and let $r \to \infty$ as illustrated in Figure 12.4. The limit of the circles l is called the **circle with infinite radius**. If **HR** is assumed, l is a line; otherwise, l is a nonstraight curve. Now rotate the original circle about its diameter, forming a sphere. As $r \to \infty$ the sphere approaches the surface formed by revolving l about a point. When **HR** is assumed, the resulting surface is a plane, but under **HA** it will be a **sphere with infinite radius**. Wachter observed that even when working under the Hypothesis of the Acute Angle, **HR** held on the surface of rotation (Bonola 1912, 62–63). In this section we define a curve that is Wachter's circle with infinite radius. Later we consider his infinite sphere. Both of these played an important role in the work of Lobachevski and Bolyai.

Fundamentals

We begin with some required terminology. Take two parallel lines $\overline{A\Omega}$ and $\overline{B\Omega}$, where Ω is an ideal point (Figure 12.5). Join $\overline{AB}$. Construct the bisectors to $\sphericalangle BA\Omega$ and $\sphericalangle AB\Omega$. These two bisectors meet in a point E (Exercise 11). If we drop a perpendicular from E to $\overline{AB}$ intersecting it at F, then

$$\triangle AFE \cong \triangle ACE \text{ and } \triangle BFE \cong \triangle BDE$$

by AAS. Thus, $\overline{CE} \cong \overline{DE}$ because $\overline{CE} \cong \overline{FE}$ and $\overline{FE} \cong \overline{DE}$. Let $\overline{E\Omega}$ be parallel to both $\overline{A\Omega}$ and $\overline{B\Omega}$ (Exercise 10). This line is called the **line of symmetry** between the parallels $\overline{A\Omega}$ and $\overline{B\Omega}$. We note the following (Exercise 13):

- The line of symmetry bisects $\sphericalangle DEC$.
- The line of symmetry is equidistant between its two parallel lines.

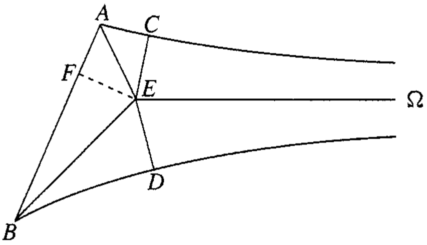

Figure 12.5 $\overline{E\Omega}$ is the line of symmetry between $\overline{A\Omega}$ and $\overline{B\Omega}$.

Furthermore, if we join C and D, the angles that $\overline{CD}$ makes with $\overline{A\Omega}$ and $\overline{B\Omega}$ are congruent (Exercise 12). We have a name for this (Wolfe 1945, 110–113).

■ **DEFINITION 12.2.1** [*Appendix* 8]

If $\overline{C\Omega}$ is parallel to $\overline{D\Omega}$ and $\sphericalangle CD\Omega$ is congruent to $\sphericalangle DC\Omega$, the point C **corresponds** to the point D.

We now prove two basic results about corresponding points. The first involves finding the line of symmetry. The second states that given a point on one of two parallel lines, there is a unique point on the second parallel that corresponds to the given point.

■ **THEOREM 12.2.2**

[HA] If two points on different parallel lines correspond, the line of symmetry between the parallel lines is the perpendicular bisector of the segment joining the two points.

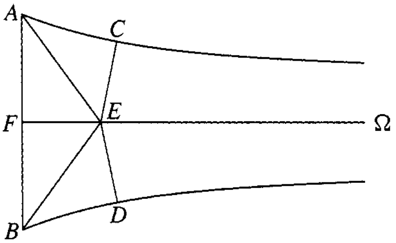

PROOF

Let $\overline{A\Omega}$ be parallel to $\overline{B\Omega}$ and suppose that A corresponds to B. Join $\overline{AB}$. Bisect $\sphericalangle BA\Omega$ and $\sphericalangle AB\Omega$ and let E be the intersection of the bisectors. Then $\overline{E\Omega}$ is the line of symmetry between $A\Omega$ and $B\Omega$. Extend it to intersect $\overline{AB}$ at F. Since A corresponds to B, $\sphericalangle BAE \cong \sphericalangle ABE$, and then $\overline{AE} \cong \overline{BE}$. Since $\overline{E\Omega}$ is the line of symmetry for the parallels, $\sphericalangle CE\Omega \cong \sphericalangle DE\Omega$. Thus, $\sphericalangle AEF \cong \sphericalangle BEF$. We now have $\triangle AFE \cong \triangle BFE$ by ASA. Therefore, $\overline{AF} \cong \overline{FB}$ and $\overline{F\Omega}$ is perpendicular to $\overline{AB}$ because $\sphericalangle AFE \cong \sphericalangle BFE$. ■

■ THEOREM 12.2.3

[HA] Given two parallel lines and any point on one of them, there exists exactly one point on the other line that corresponds to the first point.

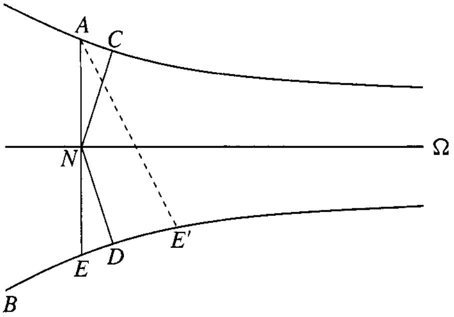

PROOF

Let $\overline{A\Omega}$ and $\overline{B\Omega}$ have line of symmetry $\overline{N\Omega}$. Let N be chosen so that $\overline{AN}$ is perpendicular to $\overline{N\Omega}$. From N construct $\overline{NC}$ perpendicular to $\overline{A\Omega}$ at C and $\overline{ND}$ perpendicular to $\overline{B\Omega}$ at D. Because $\overline{N\Omega}$ is the line of symmetry for $\overline{A\Omega}$ and $\overline{B\Omega}$, we have $\overline{NC} \cong \overline{ND}$. Finally, find E on $\overline{B\Omega}$ so that E is between B and D and $\overline{AC} \cong \overline{ED}$. Since $\triangle ACN \cong \triangle EDN$ by SAS, we see that A, N, and E are collinear (Theorem 5.3.8) and

$$\sphericalangle EAC \cong \sphericalangle AED.$$

That is, A corresponds to E.

To see that the point E is unique, suppose that A also corresponds to the point E' on $\overline{B\Omega}$. This means that

$$\sphericalangle AE'\Omega \cong \sphericalangle E'A\Omega.$$

We now examine the angle sum of $\triangle AEE'$. Since

$$\mathrm{m}\sphericalangle AE'E = \pi - \mathrm{m}\sphericalangle AE'\Omega$$

and

$$\mathrm{m}\sphericalangle AEE' = \mathrm{m}\sphericalangle EAE' + \mathrm{m}\sphericalangle E'A\Omega,$$

we find that

$$\mathrm{m}\sphericalangle EAE' + \mathrm{m}\sphericalangle AEE' + \mathrm{m}\sphericalangle AE'E = 2\mathrm{m}\sphericalangle EAE' + \pi.$$

This contradicts the Saccheri-Legendre Theorem (11.2.7). ■

We are now ready to formally define the sphere of infinite radius. Bolyai did not give this curve a name, so we use Lobachevski's terminology.

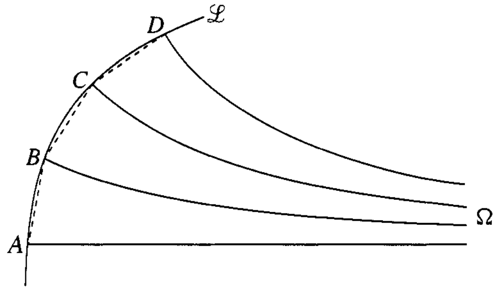

Figure 12.6 The horocycle $\mathscr{L}$ with corresponding points A, B, C, and D.

DEFINITION 12.2.4 [*Theory of Parallels* 31, *Appendix* 11]

Fix $\overline{A\Omega}$. The locus of all points B such that $\overline{B\Omega}$ is parallel to $\overline{A\Omega}$ and B corresponds to A is called a **horocycle** (or **boundary line**). The segment $\overline{A\Omega}$ is called the **axis** of the horocycle.

If we assume **HR**, the horocycle is a line. This is because when A and B are corresponding points on two parallel lines, they form interior angles on the same side that are congruent and supplementary. However, when we assume **HA**, the horocycle becomes a curve (Figure 12.6). This is the consequence of the next theorem.

THEOREM 12.2.5

[HA] If three points A, B, and C lie one on each of three mutually parallel lines such that A corresponds to B and B corresponds to C, the three points are not collinear.

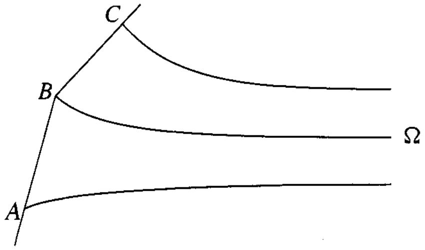

PROOF

Let A, B, and C be on distinct parallel lines through ideal point Ω. Let $\sphericalangle BA\Omega \cong \sphericalangle AB\Omega$ and $\sphericalangle CB\Omega \cong \sphericalangle BC\Omega$. Assume that A, B, and C are collinear. Then, $\sphericalangle AB\Omega$ and $\sphericalangle CB\Omega$ are supplementary. Hence, $\sphericalangle BA\Omega$ and $\sphericalangle BC\Omega$ are supplementary. This is impossible because we would have the angle sum of $\triangle AC\Omega$ at least as great as two right angles, so we conclude that A, B, and C cannot be collinear. ■

Not only is a horocycle $\mathcal{L}$ curved under **HA**, but any two points on $\mathcal{L}$ must correspond. This means that if a line drawn at a point on $\mathcal{L}$ is parallel to the defining axis of the horocycle, that line can be treated as the axis of the horocycle. For this reason we call every such line an **axis**. All of this follows from the next proposition.

■ **THEOREM 12.2.6** [*Theory of Parallels* **30**, *Appendix* **10**]

[HA] If three points A, B, and C lie one on each of three parallel lines, where A corresponds to C and B corresponds to C, A must correspond to B.

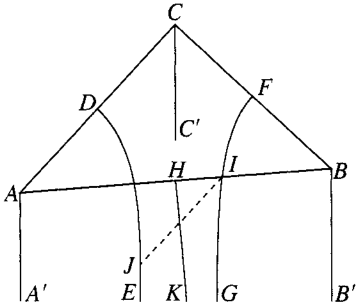

PROOF

Suppose that $\overline{AA'}$, $\overline{BB'}$, and $\overline{CC'}$ are parallel to each other with A and B corresponding with point C. Let $\overline{DE}$ be the line of symmetry between $\overline{AA'}$ and $\overline{CC'}$ and $\overline{FG}$ the line of symmetry between $\overline{CC'}$ and $\overline{BB'}$. Let $\overline{HK}$ be the perpendicular bisector of $\overline{AB}$ with H on $\overline{AB}$. We know that the two lines of symmetry are parallel to themselves and the three given lines (Theorem 12.1.3). Moreover, $\overline{HK}$ is parallel to $\overline{FG}$. To see this, let I be the intersection of $\overline{FG}$ and $\overline{AB}$ and let $\overline{IJ}$ be in the interior of $\sphericalangle HIG$. Since $\overline{DE}$ and $\overline{FG}$ are parallel, we may take J on $\overline{DE}$. Therefore, by Pasch's Postulate, $\overline{HK}$ must intersect $\overline{IJ}$. From this we may conclude that $\overline{HK}$ is parallel to both $\overline{AA'}$ and $\overline{BB'}$. Since $\overline{HK}$ is the perpendicular bisector to $\overline{AB}$, we have that $\sphericalangle A'AH$ is congruent to $m\sphericalangle B'BH$ because both have measure $\Pi(AH)$. Hence, the points A and B correspond to each other. ■

Concentric Horocycles

Let A be on horocycle $\mathcal{L}$ with axis $\overline{A\Omega}$ in a plane Λ. Let B be any other point on $\mathcal{L}$. By Definition 12.2.4, A corresponds with B on $\mathcal{L}$. Taking M to be the midpoint of $\overline{AB}$ implies that $\overline{M\Omega}$ is the line of symmetry between $\overline{A\Omega}$ and $\overline{B\Omega}$ and meets $\overline{AB}$ in a right angle (Theorem 12.2.2). Therefore, $m\sphericalangle MB\Omega = \Pi(AB/2)$, so we may describe $\mathcal{L}$ by

$$\mathcal{L} = \{X \in \Lambda : m\sphericalangle AX\Omega = \Pi(AX/2)\}. \tag{12.3}$$

Furthermore, since all of the points on a horocycle correspond to each other, any of the parallels at a given point along the curve could have been chosen as the axis upon which to define $\mathcal{L}$. This means that the curvature of the horocycle is constant at every point. In fact, since the horocycle is defined with Π, all horocycles have the same curvature, allowing us to conclude that any two are congruent. Therefore, if two horocycles in a plane have a common axis, they will share all of their axes and the two horocycles will be equidistant as measured along them. Such a pair of horocycles are called **concentric**. Suppose that $\mathscr{L}$ and $\mathscr{L}'$ are concentric with A on $\mathscr{L}$ and B on $\mathscr{L}'$. If B is between A and the ideal point Ω on a common axis, we say that $\mathscr{L}$ is the **outer horocycle** with respect to $\mathscr{L}'$.

In concentric horocycles a pair of axes will intercept arcs that have lengths that depend only on the distance that separates the horocycles. To prove this, we need two lemmas.

■ LEMMA 12.2.7

Let $\mathscr{L}$ and $\mathscr{L}'$ be concentric horocycles. Take A, B, and C on $\mathscr{L}$ and A', B', and C' on $\mathscr{L}'$ such that $\overline{AA'}$, $\overline{BB'}$, and $\overline{CC'}$ are common axes. If $\overset{\frown}{AB}$ and $\overset{\frown}{AC}$ are commensurable, then

$$\frac{\mathrm{m}\overset{\frown}{A'C'}}{\mathrm{m}\overset{\frown}{AC}} = \frac{\mathrm{m}\overset{\frown}{A'B'}}{\mathrm{m}\overset{\frown}{AB}}.$$

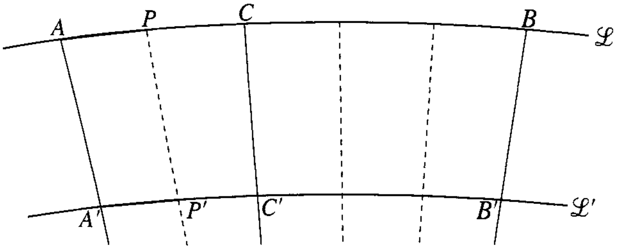

PROOF

We assume that the lengths of arcs of a horocycle are additive and we measure the arcs by measuring their intercepted line segments. Let r and s be positive integers such that

$$\mathrm{m}\overset{\frown}{AC} = r(\mathrm{m}\overset{\frown}{AP}) \text{ and } \mathrm{m}\overset{\frown}{AB} = s(\mathrm{m}\overset{\frown}{AP}),$$

where P is some point on $\mathscr{L}$. In the diagram, $r = 2$ and $s = 5$. Without loss of generality, P is on $\overset{\frown}{AC}$. Now join $\overline{AC}$ and $\overline{A'C'}$. Notice that P is the midpoint of $\overset{\frown}{AC}$. Draw the axis to $\mathscr{L}$ at P and let it intersect $\mathscr{L}'$ at P'. Therefore, $\overline{PP'}$ is perpendicular to $\overline{AC}$. Since A corresponds to C and A' corresponds to C', $\sphericalangle A'AC \cong \sphericalangle C'CA$ and $\sphericalangle AA'C' \cong \sphericalangle CC'A'$. Finally, $\overline{AA'} \cong \overline{CC'}$ since the horocycles are concentric. Hence, $APP'A' \cong CPP'C'$. From this we conclude that $\overset{\frown}{A'P'} \cong \overset{\frown}{P'C'}$, so $\mathrm{m}\overset{\frown}{A'P'} = \mathrm{m}\overset{\frown}{P'C'}$ and

$$\mathrm{m}\overset{\frown}{A'C'} = r(\mathrm{m}\overset{\frown}{A'P'}).$$

Similarly, $\mathrm{m}\overset{\frown}{A'B'} = s(\mathrm{m}\overset{\frown}{A'P'})$. Since this argument generalizes to all positive integers r and s, we conclude that $\mathrm{m}\overset{\frown}{A'C'}/\mathrm{m}\overset{\frown}{AC} = \mathrm{m}\overset{\frown}{A'B'}/\mathrm{m}\overset{\frown}{AB}$. ■

The second lemma is needed here and in a future application.

■ **LEMMA 12.2.8**

Let f be a continuous, positive-valued function such that $f(1) < 1$ and

$$f(nx) = [f(x)]^n$$

for every $x > 0$ and positive integer n. Then there exists $k > 0$ such that $f(x) = e^{-kx}$ for all $x > 0$.

PROOF

Let $x > 0$ and take positive integers r and s. Observe that

$$f\left(\frac{r}{s}x\right) = f\left(r\frac{x}{s}\right) = \left[f\left(\frac{x}{s}\right)\right]^r$$

and

$$\left[f\left(\frac{x}{s}\right)\right]^s = f\left(s\frac{x}{s}\right) = f(x).$$

Therefore,

$$f\left(\frac{r}{s}x\right) = \left[f\left(\frac{x}{s}\right)\right]^r = \left[f\left(\frac{x}{s}\right)^s\right]^{r/s} = [f(x)]^{r/s}. \tag{12.4}$$

Now let $a > 0$ be a real number. There exist positive rational numbers $q_1, q_2, q_3, \ldots$ such that

$$\lim_{i\to\infty} q_i = a.$$

By Equation 12.4 we know that

$$f(q_i x) = [f(x)]^{q_i}$$

for all $i = 1, 2, 3, \ldots$. Therefore, since f is continuous,

$$f(ax) = \lim_{i\to\infty} f(q_i x) = \lim_{i\to\infty} [f(x)]^{q_i} = [f(x)]^a.$$

Finally, choose $k > 0$ so that $f(1) = e^{-k}$. Then

$$f(x) = [f(1)]^x = e^{-kx}. \blacksquare$$

We can now prove the main result.

■ **THEOREM 12.2.9** [*Theory of Parallels* 33]

Given concentric horocycles $\mathcal{L}$ and $\mathcal{L}'$ such that $\mathcal{L}$ is outer with respect to $\mathcal{L}'$, there exists $k > 0$ such that for all $\overline{AA'}$ and $\overline{BB'}$ that are shared axes of $\mathcal{L}$ and $\mathcal{L}'$ with A and B on $\mathcal{L}$ and A' and B' on $\mathcal{L}'$,

$$\mathrm{m}\widehat{A'B'} = \mathrm{m}\widehat{AB} \cdot e^{-k(AA')}.$$

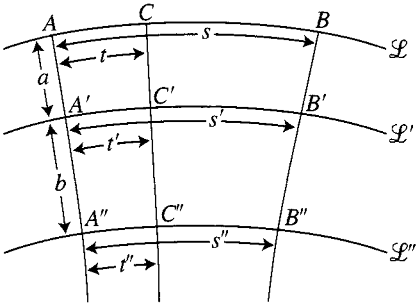

PROOF

Let $\mathcal{L}''$ be a third horocycle concentric with $\mathcal{L}$ and $\mathcal{L}''$. Take any C on $\widehat{AB}$ and let the axis through C intersect $\mathcal{L}'$ at C' and $\mathcal{L}''$ at C''. Make the following assignments:

$$s = \mathrm{m}\widehat{AB}, s' = \mathrm{m}\widehat{A'B'}, s'' = \mathrm{m}\widehat{A''B''}$$

and

$$t = \mathrm{m}\widehat{AC}, t' = \mathrm{m}\widehat{A'C'}, t'' = \mathrm{m}\widehat{A''C''}.$$

Assume that s and t are commensurable. Since the horocycles are concentric, we know by Lemma 12.2.7 that

$$\frac{t'}{t} = \frac{s'}{s}, \frac{t''}{t'} = \frac{s''}{s'}, \text{ and } \frac{t''}{t} = \frac{s''}{s}.$$

Since these ratios remain constant, provided that the distance between the horocycles remains constant, we conclude that we can define a function $f(x)$ that equals the value of the ratio of the smaller arc to the larger one when the distance between the horocycles is x. Based on the above, we have

$$f(a) = \frac{s'}{s}, f(b) = \frac{s''}{s'}, \text{ and } f(a+b) = \frac{s''}{s}.$$

Therefore,

$$f(a)f(b) = \frac{s'}{s}\frac{s''}{s'} = \frac{s''}{s} = f(a+b).$$

Since a and b were chosen arbitrarily, we conclude that for all $x > 0$ and positive integers n,

$$\begin{aligned} f(nx) &= f([n-1]x + x) \\ &= f([n-1]x)f(x) \\ &= f([n-2]x)[f(x)]^2 \\ &\vdots \\ &= [f(x)]^n. \end{aligned}$$

Since f is certainly a continuous function and $f(x) < 1$, Lemma 12.2.8 allows us to conclude that for some $k > 0$, $f(x) = e^{-kx}$ for every $x > 0$. Hence,

$$e^{-k(AA')} = f(AA') = \mathrm{m}\widehat{A'B'}/\mathrm{m}\widehat{AB}. \blacksquare$$

Exercises

1. Prove that the Parallel Postulate implies that three points determine a circle.

2. Show that Theorem 12.2.3 also holds for **HR**.

3. Prove: If $\overline{AB}$ is parallel to $\overline{CD}$, $\sphericalangle CAB \cong \sphericalangle ACD$, and $\overline{EF}$ is the perpendicular bisector of $\overline{AC}$, then $\overline{EF}$ is parallel to both $\overline{AB}$ and $\overline{CD}$.

4. Prove:
   (a) If perpendiculars are drawn at the midpoints of the sides of a triangle and two of them are parallel, all three are parallel.
   (b) If perpendiculars are drawn at the midpoints of the sides of a triangle such that no two are parallel but they do not intersect each other, they are all perpendicular to the same line.

5. In a plane, how many horocycles pass through two given points? Explain.

6. Prove that a horocycle is determined uniquely by three distinct points.

7. Describe a mechanical device that could be used to draw a horocycle.

8. Let $\overline{A\Omega}$ be an axis of horocycle $\mathscr{L}$. Show that if $\overline{AT}$ is perpendicular to $\overline{A\Omega}$, then $\overline{AT}$ is tangent to $\mathscr{L}$. Is the converse true?

9. If a horocycle is revolved about any two of its axes, explain why the resulting surfaces are congruent. Are the two surfaces the same surface?

10. Let $\overline{A\Omega}$ be parallel to $\overline{B\Omega}$. Let C be a point not on either of these two parallel lines. Prove that there exists a line through C parallel to both $\overline{A\Omega}$ and $\overline{B\Omega}$.

11. In Figure 12.5, prove that the bisectors of $\sphericalangle BA\Omega$ and $\sphericalangle AB\Omega$ intersect in E.

12. Show that C and D are corresponding points in Figure 12.5.

13. Let n be the line of symmetry between l and m. There are two ways to interpret that n is equidistant between l and m.
    (a) Let A be on l and find B on n such that $\overline{AB}$ is perpendicular to l. Next find C on m such that $\overline{BC}$ is perpendicular to m. Show that $\overline{AB} \cong \overline{BC}$.
    (b) Let B be on n. Find A on l and C on m such that $\overline{AB}$ and $\overline{BC}$ are perpendicular to n. Prove that $\overline{AB} \cong \overline{BC}$.

14. Prove that the line of symmetry between two parallel lines is unique.

15. Let $\mathscr{L}$ and $\mathscr{L}'$ be concentric horocycles. Let A and B be on $\mathscr{L}$ and A' and B' be on $\mathscr{L}'$ such that $\overline{AA'}$ and $\overline{BB'}$ are common axes. Prove that if $\mathscr{L}$ is the outer horocycle with respect to $\mathscr{L}'$, then $AB > A'B'$.

16. Show that the area enclosed by two horocycles and two common axes is proportional to the distance between the horocycles. (*Hint:* First determine how to measure this area.)

17. Assume **HR**. Let $\overset{\frown}{AB}$, $\overset{\frown}{A'B'}$, and $\overset{\frown}{A''B''}$ be on concentric circles such that $\overline{AA'A''}$ and $\overline{BB'B''}$ are common radii and $\overline{AA'} \cong \overline{A'A''}$. Show that

$$\frac{\mathrm{m}\overset{\frown}{AB}}{\mathrm{m}\overset{\frown}{A'B'}} \neq \frac{\mathrm{m}\overset{\frown}{A'B'}}{\mathrm{m}\overset{\frown}{A''B''}}.$$

18. Let f and g be real-valued, continuous functions. Answer the following and justify your responses.
   (a) If $f(n) = g(n)$ for all integers n, does $f = g$?
   (b) If $f(x) = g(x)$ for all rational numbers x, does $f = g$?
   (c) If $f(x) = g(x)$ for all irrational numbers x, does $f = g$?

19. Given $\overline{AB}$ and $k > 0$. The set of all points P such that the perpendicular distance from P to $\overline{AB}$ equals k is called an **equidistant curve**. $\overline{AB}$ is called the **baseline**, and any segment between the curve and the baseline that is perpendicular to the baseline is an **axis**.
   (a) Explain why each equidistant curve has two **branches**, one "below" the baseline and one "above."
   (b) Draw an equidistant curve assuming **HR** and then assuming **HA**.
   (c) Assuming **HA**, draw two equidistant curves to $\overline{AB}$, one with distance k_1 and the other with distance $k_2 > k_1$.

20. Prove the following about equidistant curves:
   (a) A line intersects an equidistant curve at most twice.
   (b) A line that intersects an equidistant curve at P and is perpendicular to the axis at P is tangent to the curve.

21. Let l be a line and $\mathscr{C}$ an equidistant curve to l. Let $\overline{AB}$ be an axis with A on $\mathscr{C}$ and B on l. Prove that for any two distinct points C and D on $\mathscr{C}$ such that $\overline{CA} \cong \overline{AD}$, there exist C' and D' on l such that $\sphericalangle C'CA \cong \sphericalangle D'DA$ and $\sphericalangle CC'B \cong \sphericalangle DD'B$. Can the construction be redone so that $\sphericalangle C'CA \cong \sphericalangle CC'B$?

22. Conjecture why the horocycle is viewed as a circle of infinite radius.

12.3 THE SURFACE OF A SPHERE

Thus far our introduction to non-Euclidean geometry has involved references to spheres. Lambert conjectured that his new geometry could be interpreted as existing on the surface of a sphere with an imaginary radius. Wachter wrote to Gauss about his sphere of infinite radius. In each of these cases, the surface of the sphere is what was important, so in this section we survey some of the results concerning this surface in Euclidean space. These results will later be generalized to help us understand the geometry of the Hypothesis of the Acute Angle.

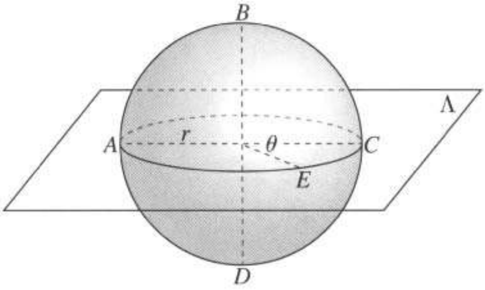

Figure 12.7 The sphere $ABCD$ cut by plane Λ forming a great circle.

The Area of a Triangle

Fix a sphere of radius r. The points of the geometry will be standard points on the surface of this sphere. The lines will be **great circles**. These are the circles formed by the intersection of the sphere with a plane that contains the sphere's center, as in Figure 12.7. A line segment in this geometry will be an arc of a great circle. Hence, the length of a segment is given by

$$\text{arc length} = \text{sphere's radius} \times \text{arc's central angle}, \tag{12.5}$$

where all angles are measured in radians. For example, in Figure 12.7 the length of $\overset{\frown}{EC}$ is $r\theta$.

As lines in the plane can be used to construct angles and triangles, the same is true for great circles on the surface of the sphere. Let Λ and K be two planes that cut the sphere, forming two great circles, as in Figure 12.8. Suppose that B is a point of intersection of the two great circles and assume that the plane A_1BC_1 is tangent to the sphere at B. We will denote the **spherical angle** formed by points A, B, and C by our usual angle notation, and its measure will be defined as the measure of $\sphericalangle A_1BC_1$. In some cases three great circles will meet to form a triangle. Such a figure is called a

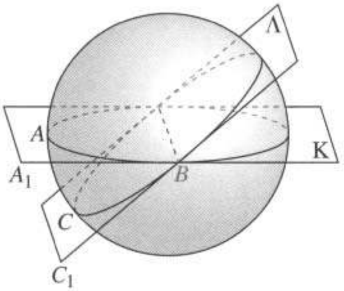

Figure 12.8 The measure of $\sphericalangle ABC$ equals the measure of $\sphericalangle A_1BC_1$.

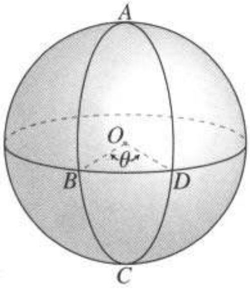

Figure 12.9 The lune $ABCD$.

spherical triangle, and again we will not change notation. The type of triangle will be clear from context. As opposed to the situation with triangles constructed under **HA**, triangles on the surface of a sphere have a positive angle excess. This means that for spherical $\triangle ABC$,

$$\text{angle excess} = (\mathrm{m}\sphericalangle A + \mathrm{m}\sphericalangle B + \mathrm{m}\sphericalangle C) - \pi > 0.$$

The proposition that allows us to conclude this is called **Girard's Theorem.** Albert Girard was Flemish and lived from 1590 to 1633. He is noted for his early promotion of algebraic symbolism, and he was the first to formulate the Fundamental Theorem of Algebra (Boyer and Merzbach 1991, 303, 319; Burton 1985, 512). His work in trigonometry led him to prove the theorem that bears his name. The proof given below, however, is due to Euler (Rosenfeld 1988, 31–32). He proved it using lunes. A **lune** is the region on the surface of a sphere that lies between two great circles. (Compare this definition with that of Exercise 3 in Section 4.3.) The great circles intersect at the endpoints of a diameter of the sphere. These two points are called **antipodal.**

Let us find the area of a lune on a sphere with center O and radius r. Suppose that A, B, and C lie on a great circle and A, D, and C lie on another great circle (Figure 12.9). Let $\theta = \mathrm{m}\sphericalangle BOD$. The ratio of the area of the lune swept out by θ to the surface area of the sphere is $\theta/2\pi$. Thus,

$$\text{area}(\text{lune } ABCD) = \frac{\theta}{2\pi} \cdot 4\pi r^2 = 2\theta r^2$$

because the surface area of the sphere is $4\pi r^2$. On the unit sphere the equation simplifies to

$$\text{area}(\text{lune } ABCD) = 2\theta.$$

This is what is needed in Euler's proof.

■ THEOREM 12.3.1 [Girard]

The area of a spherical triangle equals its angle excess.

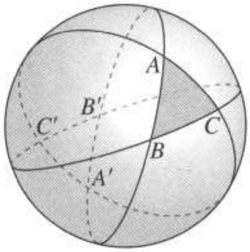

PROOF

Begin by choosing a sphere with unit radius. Let $\triangle ABC$ be on the sphere's surface. Let A', B', and C' be antipodal, respectively, to A, B, and C. Write

$$\alpha = \text{m}\sphericalangle BAC,\ \beta = \text{m}\sphericalangle ABC, \text{ and } \delta = \text{m}\sphericalangle ACB.$$

Then

$$\text{area}(ABA'C) = 2\alpha,\ \text{area}(BCB'A) = 2\beta, \text{ and area}(CAC'B) = 2\delta.$$

Notice that $S = \text{area}(\triangle ABC)$ is the area of the intersection of all three lunes. Therefore, the areas of the lunes without $\triangle ABC$ are $2\alpha - S$, $2\beta - S$, and $2\delta - S$. Observe that there are four pairs of congruent triangles on the sphere:

$$\begin{aligned}\triangle ABC &\cong \triangle A'B'C',\\ \triangle AB'C' &\cong \triangle A'BC,\\ \triangle BA'C' &\cong \triangle B'AC,\\ \triangle CA'B' &\cong \triangle C'AB.\end{aligned}$$

Moreover, these triangles cover the sphere, so their total sum must equal 4π. Hence,

$$2S + 4(\alpha + \beta + \delta) - 6S = 4\pi,$$

and then

$$S = (\alpha + \beta + \delta) - \pi.$$

The generalization to an arbitrary sphere is left to Exercise 6. ■

Identities

We shall use the convention that in $\triangle ABC$, the lowercase letters a, b, and c represent the lengths of the sides opposite the angles A, B, and C, respectively, and both McCleary (1994, 3–9) and Granville (1909, 193f) were used as sources. We begin with some fundamental identities.

THEOREM 12.3.2

If $\triangle ABC$ is on a sphere of radius r and $\sphericalangle C$ is the only right angle, then

- $\sin a/r = \sin c/r \sin A$.
- $\sin b/r = \sin c/r \sin B$.
- $\cos A = \cos a/r \sin B$.
- $\cos B = \cos b/r \sin A$.
- $\cos c/r = \cos a/r \cos b/r$.

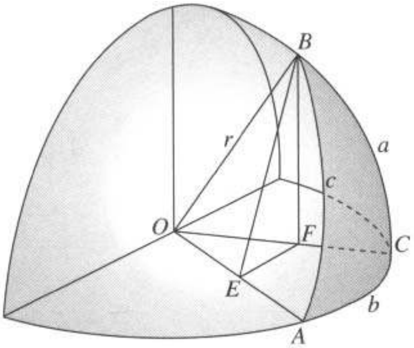

PROOF

We will prove the first and the last, leaving the others to Exercise 2. Let O be the center of the sphere and from B drop a perpendicular to $\overline{OC}$ at F and then find E on $\overline{OA}$ so that $\overline{EF}$ is perpendicular to $\overline{OA}$. First notice that

$$m\sphericalangle BOF = \frac{a}{r}, \sphericalangle EOF = \frac{b}{r}, \text{and } m\sphericalangle BOE = \frac{c}{r}.$$

Then

$$\sin\frac{a}{r} = \frac{BF}{OB}, \sin\frac{c}{r} = \frac{BE}{OB}, \text{ and } \sin A = \sin\sphericalangle BEF = \frac{BF}{BE}.$$

Hence, $\sin a/r = \sin c/r \sin A$. We also have

$$\cos\frac{c}{r} = \frac{OE}{OB}, \cos\frac{a}{r} = \frac{OF}{OB}, \text{ and } \cos\frac{b}{r} = \frac{OE}{OF}.$$

Thus, $\cos c/r = \cos a/r \cos b/r$. ■

The remaining theorems that we will prove concerning spherical triangles all have corresponding plane versions. We begin with the spherical analog to the Pythagorean Theorem.

■ **THEOREM 12.3.3 [Spherical Pythagorean Theorem]**

If $\sphericalangle C$ is the only right angle in $\triangle ABC$ on a sphere of radius r, then

$$\cos\frac{c}{r} = \cos\frac{a}{r}\cos\frac{b}{r}.$$

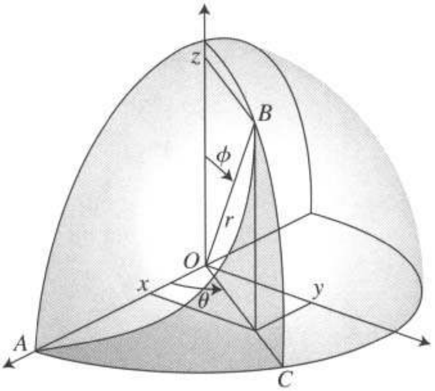

PROOF

We may assume by rotating the sphere, if necessary, that the coordinates of A are $(r,\ 0,\ 0)$ and C lies in the xy-plane. Recall that any point B on the sphere can be identified by the point $(r,\ \phi,\ \theta)$ in spherical coordinates. The point B can then be converted to Cartesian coordinates using the equations

$$x = r\sin\phi\cos\theta,\ y = r\sin\phi\sin\theta,\ \text{and}\ z = r\cos\phi,$$

where $\phi = \pi/2 - a/r$ and $\theta = b/r$. Therefore,

$$\begin{aligned} B &= \left(r\sin\left[\frac{\pi}{2}-\frac{a}{r}\right]\cos\frac{b}{r},\ r\sin\left[\frac{\pi}{2}-\frac{a}{r}\right]\sin\frac{b}{r},\ r\cos\left[\frac{\pi}{2}-\frac{a}{r}\right]\right) \\ &= \left(r\cos\frac{a}{r}\cos\frac{b}{r},\ r\cos\frac{a}{r}\sin\frac{b}{r},\ r\sin\frac{a}{r}\right) \end{aligned}$$

and

$$C = \left(r\cos\frac{b}{r},\ r\sin\frac{b}{r},\ 0\right).$$

Now apply some linear algebra and view $A(r,\ 0,\ 0)$ and $B(x,\ y,\ z)$ as vectors originating from the origin. Since the angle between these two vectors is $\sphericalangle AOB$ and the measures of this angle is c/r, we can apply the dot product to write

$$A \bullet B = \|A\|\,\|B\|\cos\frac{c}{r}.$$

Thus,

$$\cos\frac{c}{r} = \frac{A \bullet B}{\|A\|\,\|B\|} = \frac{r^2\cos a/r\cos b/r}{r^2} = \cos\frac{a}{r}\cos\frac{b}{r}. \blacksquare$$

The previous theorem is called the Spherical Pythagorean Theorem for two reasons. First, it gives a relationship between the sides of a spherical right triangle. Secondly, and more important, it is the Pythagorean Theorem when examined correctly. Using the Maclaurin series,

$$\cos x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!},$$

we see that the Spherical Pythagorean Theorem becomes

$$\sum_{n=0}^{\infty} (-1)^n \frac{(c/r)^{2n}}{(2n)!} = \left(\sum_{n=0}^{\infty} (-1)^n \frac{(a/r)^{2n}}{(2n)!}\right) \cdot \left(\sum_{n=0}^{\infty} (-1)^n \frac{(b/r)^{2n}}{(2n)!}\right).$$

By multiplying the series on the right, subtracting one from both sides, and multiplying through by $-2r^2$, we find that there exists infinite series $S(r)$ and $T(r)$ such that

$$c^2 + S(r) = a^2 + b^2 + T(r), \tag{12.6}$$

where both $S(r) \to 0$ and $T(r) \to 0$ as $r \to \infty$. We conclude that as the radius of the sphere increases without bound, the equation of the Spherical Pythagorean Theorem approaches that of the standard Pythagorean Theorem. In other words, as the volume of the sphere increases, the geometry on its surface becomes more Euclidean.

The Law of Sines for plane trigonometry states that

$$\frac{a}{\sin A} = \frac{b}{\sin B} = \frac{c}{\sin C}.$$

The following is the analogous theorem for the sphere.

■ THEOREM 12.3.4 [Spherical Law of Sines]

In $\triangle ABC$ on a sphere of radius r,

$$\frac{\sin(a/r)}{\sin A} = \frac{\sin(b/r)}{\sin B} = \frac{\sin(c/r)}{\sin C}.$$

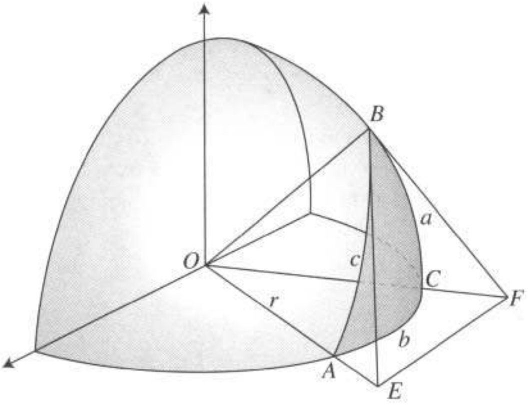

PROOF

We will prove the theorem for right triangles first and then generalize. Let $\sphericalangle C$ in $\triangle ABC$ be right on the surface of a sphere with center O and radius r. Let plane EBF be tangent to the sphere at B. Assume that $\overline{OA}$ is produced to E and $\overline{OC}$ produced to F. Since $\overline{EB}$ and $\overline{BF}$ are tangent to the sphere, $\sphericalangle OBE$ and $\sphericalangle OBF$ are right angles. Hence, appealing to Pythagoras we see that

$$OE^2 = BE^2 + OB^2$$

and

$$OF^2 = BF^2 + OB^2.$$

Observe that planes EBF and OBF are perpendicular, so $\sphericalangle OFE$ is right. We then conclude that

$$OE^2 = EF^2 + OF^2.$$

Therefore,

$$EF^2 = OE^2 - OF^2 = (BE^2 + OB^2) - (BF^2 + OB^2),$$

and then

$$BE^2 = EF^2 + BF^2.$$

Therefore, $\sphericalangle EFB$ is right, and we then have

$$\sin \sphericalangle AOC = \frac{EF}{OE} = \frac{EF}{BE} \cdot \frac{BE}{OE} = \sin B \sin \sphericalangle AOB.$$

Since $b/r = \mathrm{m}\sphericalangle AOC$,

$$\frac{\sin(b/r)}{\sin B} = \sin \sphericalangle AOB.$$

Similarly, we could redraw the plane parallel at A instead of at B and find that

$$\frac{\sin(a/r)}{\sin A} = \sin \sphericalangle AOB.$$

Because $\sin C = 1$ and $c/r = \mathrm{m}\sphericalangle AOB$, we have

$$\frac{\sin(a/r)}{\sin A} = \frac{\sin(b/r)}{\sin B} = \frac{\sin(c/r)}{\sin C}.$$

Now take $\triangle ABC$ to be an arbitrary spherical triangle. Drop $\overline{BD}$ perpendicular to $\overline{AC}$. By examining $\triangle ABD$ and $\triangle CBD$, the equation for right triangles yields

$$\sin \frac{a}{r} \sin C = \sin \frac{BD}{r} = \sin \frac{c}{r} \sin A. \tag{12.7}$$

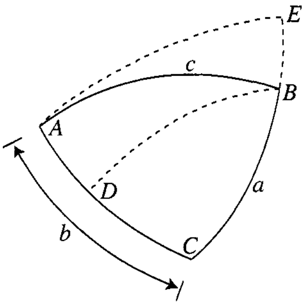

Extend $\overline{CB}$ to E such that $\overline{AE}$ joined is perpendicular to $\overline{EC}$. When we examine $\triangle ACE$ and $\triangle ABE$, we find that

$$\sin\frac{b}{r}\sin C = \sin\frac{AE}{r} = \sin\frac{c}{r}\sin(\pi - B) = \sin\frac{c}{r}\sin B. \tag{12.8}$$

These equations complete the proof. ■

The last theorem of this section is the spherical version of the Law of Cosines. Recall that the plane version states that in any $\triangle ABC$,

$$c^2 = a^2 + b^2 - 2ab\cos C.$$

When we examine the following result when r is small, we will find that it approximates the plane version (Exercise 11).

■ THEOREM 12.3.5 [Spherical Law of Cosines]

In $\triangle ABC$ on a sphere of radius r,

$$\cos\frac{a}{r} = \cos\frac{b}{r}\cos\frac{c}{r} + \sin\frac{b}{r}\sin\frac{c}{r}\cos A.$$

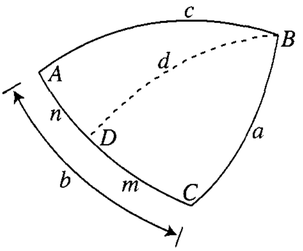

PROOF

Leaving the generalization to Exercise 10, we will prove the case when $r = 1$. Take spherical $\triangle ABC$ and drop an altitude from B to D on $\overline{AC}$. When we apply the Spherical Pythagorean Theorem (12.3.3) to $\triangle BCD$, letting $d = DB$, $m = DC$, and $n = AD$, we find that

$$\cos a = \cos d \cos m = \cos d \cos(b - n) = \cos d(\cos b \cos n + \sin b \sin n).$$

Therefore,

$$\cos a = \cos d \cos b \cos n + \cos d \sin b \sin n. \tag{12.9}$$

When we apply Theorem 12.3.3 to $\triangle BAD$, we get $\cos c = \cos d \cos n$, and substituting this into Equation 12.9 we see that

$$\cos a = \cos b \cos c + \cos d \sin b \sin n. \tag{12.10}$$

Furthermore, by Theorem 12.3.2 we have

$$\cos d = \frac{\cos c}{\cos n}.$$

Multiplying both sides of this equation by $\sin n$ we obtain

$$\cos d \sin n = \cos c \tan n. \tag{12.11}$$

However,

$$\cos A = \tan n \cot c \tag{12.12}$$

by Exercise 9, so $\tan n = \tan c \cos A$. We now substitute this into Equation 12.11 to find

$$\cos d \sin n = \cos c \tan c \cos A = \sin c \cos A,$$

and we may complete the proof by substituting into Equation 12.10. ■

Exercises

1. What would a triangle on the sphere look like if it had two right angles? Is it possible for a spherical triangle to have three right angles? Explain.

2. Let $\triangle ABC$ be on sphere of radius r with $\sphericalangle C$ the only right angle. Show the following identities.
   (a) $\sin b/r = \sin c/r \sin B$.
   (b) $\cos A = \cos a/r \sin B$.
   (c) $\cos B = \cos b/r \sin A$.

3. Let $\triangle ABC$ be on the unit sphere with $\sphericalangle C$ as the only right. Prove:
   (a) $\cos A = \tan b \cot c$.
   (b) $\sin b = \tan a \cot A$.
   (c) $\cos c = \cot A \cot B$
   (d) $\sec a = \sec A \sin B$

4. Find the remaining sides and angles (in degrees) of the spherical triangles given the following. Assume that C is the only right angle of the triangle.
   (a) $c = 110°, A = 80°$.
   (b) $c = 100°, b = 115°$.
   (c) $c = 120°, a = 20°$.
   (d) $B = 150°, b = 160°$.
   (e) $B = 150°, a = 110°$.

5. Prove that the following geometric results hold for $\triangle ABC$ on the unit sphere.
   (a) $a < b + c$.
   (b) If $a \neq b$, then $A \neq B$.
   (c) $a < b$ if and only if $A < B$.
   (d) $a + b + c < 2\pi$.
   (e) $\pi < A + B + C < 3\pi$.

6. Given (r, ϕ, θ) in spherical coordinates, show that the corresponding point in Cartesian coordinates is given by

$$\begin{aligned} x &= r \sin\phi \cos\theta, \\ y &= r \sin\phi \sin\theta, \\ z &= r \cos\phi. \end{aligned}$$

7. Prove Equation 12.12 in the proof of Theorem 12.3.5.

8. Generalize the proof of Theorem 12.3.5 to a sphere with any radius.

9. Use the Maclaurin series for sine and cosine to show that the Spherical Law of Cosines (12.3.5) yields an equation that approximates the plane Law of Cosines for small spherical triangles.

10. Prove Girard's Theorem (12.3.1) for a sphere of any radius.

11. Confirm Equations 12.7 and 12.8 from the proof of Theorem 12.3.4.

12. Verify for $\triangle ABC$ on the unit sphere:
    (a) $\cos A = \cos a \sin B \sin C - \cos B \cos C$.
    (b) $\cot A \sin C + \cos b \cos C = \cot a \sin b$.

12.4 HOROSPHERES

Both Bolyai and Lobachevski defined a three-dimensional analog to the horocycle. Bolyai did it by selecting a point A in space and an ideal point Ω (*Appendix* 11). He defined

$$\mathcal{F} = \{X : \sphericalangle AX\Omega \cong \sphericalangle XA\Omega\}.$$

The line $\overleftrightarrow{A\Omega}$ we will call the **axis** of $\mathcal{F}$. Like horocycles, at each point on $\mathcal{F}$ there is a line that is parallel to the axis. Each of these lines is parallel to each other (Exercise 6 in Section 12.1), so we can also view them as axes.

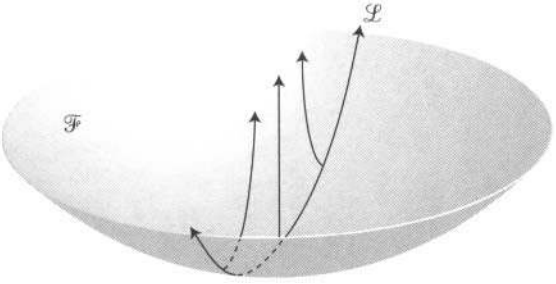

Figure 12.10 The horosphere $\mathcal{F}$ with generating horocycle $\mathcal{L}$.

If we let $B \in \mathcal{F}$, the plane Λ that contains A, B, and Ω intersects $\mathcal{F}$, resulting in a horocycle $\mathcal{L}$ as described in Equation 12.3. If we revolve Λ about $\overline{A\Omega}$, then we see that Bolyai's definition is equivalent to Lobachevski's.

■ **DEFINITION 12.4.1** [*Theory of Parallels* 34]

A **horosphere** (or **boundary surface**) is the surface $\mathcal{F}$ formed by revolving a horocycle about one of its axes.

Notice that due to the constant curvature of horocycles, if Λ instead makes an angle with $\overline{A\Omega}$, it will determine a circle on the surface of $\mathcal{F}$ (Exercise 3). Since a horocycle under **HR** is a line, the horosphere is a plane in Euclidean geometry. Under **HA** it looks more like Figure 12.10.

Induced Euclidean Geometry

Despite assuming the Hypothesis of the Acute Angle, there is a natural interpretation of the term *line* that yields a Euclidean geometry on the surface of a horosphere. We first need to prove a theorem.

■ **THEOREM 12.4.2** [*Theory of Parallels* 28]

[HA] If three planes intersect each other in parallel lines, the sum of the angles that the planes make with each other equals two right angles.

PROOF

Suppose that three planes intersect in $\overline{AA'}$, $\overline{BB'}$, and $\overline{CC'}$ and that the three lines are parallel. These planes make angles with each other of measure x at $\overline{AA'}$, y at $\overline{BB'}$, and z at $\overline{CC'}$ (not identified in the diagram on page 426). Take D on $\overline{BB'}$ and pass through it a plane containing A and C. This fourth plane intersects $ABB'A'$ in $\overline{AD}$ and $CBB'C'$ in $\overline{CD}$. Denote the angle of inclination of plane ACD with $ACC'A'$ by w. We now construct three spheres, one each

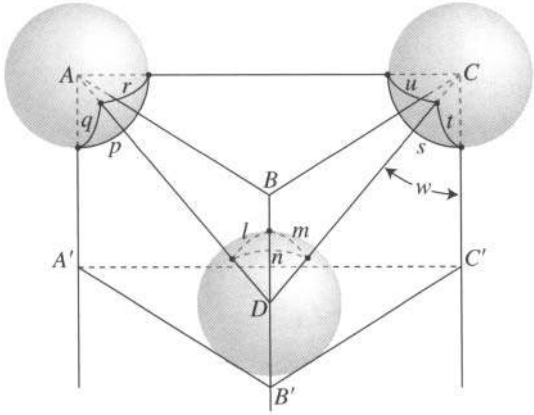

centered at A, C, and D. This results in three spherical triangles:

- The vertices of the spherical triangle at A are the intersections of $\overline{AA'}$, $\overline{AD}$, and $\overline{AC}$ with the sphere. Let its area be α and its sides have length p, q, and r. Observe that the measure of the angle opposite q is w, the angle opposite r is x, and by Girard's Theorem (12.3.1) the angle opposite p has measure $\alpha - w - x + \pi$. This is summarized in the following table:

length	opposite angle
p	$\alpha - w - x + \pi$
q	w
r	x

- Similarly, the sphere at C intersects $\overline{AC}$, $\overline{CD}$, and $\overline{CC'}$, forming a spherical triangle of area β:

length	opposite angle
s	$\beta - w - z + \pi$
t	w
u	z

- Let δ be the area of the spherical triangle formed by the sphere at D. We note that the angle opposite l in the diagram is supplementary to the angle opposite s in the sphere at C. This is because they both have one side in the plane ACD and their other side in the plane $BCC'B'$, but these sides are on opposite sides of ACD. Similarly, the angle opposite m and the angle opposite p are also supplementary. Therefore, we describe the spherical triangle at D by a table:

length	opposite angle
l	$w + z - \beta$
m	$w + x - \alpha$
n	y

Therefore,

$$\delta = (x + y + z - \pi) + 2w - \beta - \alpha$$

by Girard's Theorem. Now let $w \to 0$. We see that both $q \to 0$ and $t \to 0$, so $\alpha \to 0$ and $\beta \to 0$. Also, the sides l and m approach zero, which means that $\delta \to 0$. But the angles x, y, and z remain fixed. Hence, $x + y + z = \pi$. ■

On a horosphere $\mathcal{F}$ let the word *point* refer to a regular point and let the word *line* refer to a horocycle. These choices impart a geometric system upon $\mathcal{F}$ called the **induced Euclidean geometry**. Let A, B, and C be points on $\mathcal{F}$ but not all on the same horocycle. Call the three horocycles determined by these points AB, BC, and AC. Take three planes that each contain an axis of $\mathcal{F}$ and also intersect the horosphere in AB, BC, and AC, respectively. Define the angle between any two of these horocycles to be the angle between their corresponding planes. (This is similar to how we defined the measure of angles on a sphere.) By Theorem 12.4.2, the sum of the internal angles of $\triangle ABC$ on $\mathcal{F}$ is equal to two right angles. This means that the Parallel Postulate holds true in the induced geometry by Corollary 11.2.12.

■ **THEOREM 12.4.3** [*Theory of Parallels* 34, *Appendix* 21]

The induced geometry on a horosphere is Euclidean.

The last result means that inside a geometry that satisfies **HA**, there is a surface that can naturally be interpreted as a Euclidean plane. The German mathematician David Hilbert showed that there is no such copy of **HA** in Euclidean geometry (Bonola 1912, 145).

Spherical Trigonometry and HA

As we study the sphere with infinite radius under **HA**, it is natural to ask about the status of a sphere with finite radius under the same hypothesis. The answer will involve two technical results, the second requires two horospheres.

Let $\triangle ABC$ be such that $m\sphericalangle A = \Pi(\alpha)$, $m\sphericalangle B = \Pi(\beta)$, and $\sphericalangle C$ is a right angle (Figure 12.11). Find A' in space so that $\overline{AA'}$ is perpendicular to the plane ABC. In addition, find B' and C' so that $\overline{BB'}$ and $\overline{CC'}$ are parallel to $\overline{AA'}$. These lines are also mutually parallel and are the intersections of three planes that make angles with each other with the following measures:

$$\Pi(\alpha) \text{ at } \overline{AA'},\ \Pi(\alpha') \text{ at } \overline{BB'}, \text{ and } \pi/2 \text{ at } \overline{CC'},$$

where α' is chosen so that

$$\Pi(\alpha) + \Pi(\alpha') = \pi/2$$

by Theorem 12.4.2. At B construct a unit sphere. Let B_1, B_2, and B_3 be the intersections of the sphere with $\overline{BC}$, $\overline{BA}$, and $\overline{BB'}$, respectively. Then the measurements of spherical $\triangle B_1B_2B_3$ are

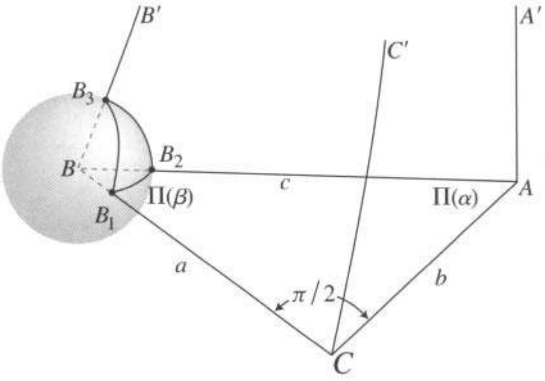

Figure 12.11 Deriving a spherical triangle from a plane triangle.

vertex	measure	opposite side
B_1	$\Pi(b)$	$\Pi(c)$
B_2	$\pi/2$	$\Pi(a)$
B_3	$\Pi(\alpha')$	$\Pi(\beta)$

We will return to this shortly, but first we prove the following.

THEOREM 12.4.4 [*Theory of Parallels* 35]

[HA] If m$\sphericalangle A = \Pi(\alpha)$, m$\sphericalangle B = \Pi(\beta)$, and $\sphericalangle C$ is right in $\triangle ABC$, then:

- $\sin \Pi(c) = \sin \Pi(a) \sin \Pi(b)$.
- $\sin \Pi(\beta) = \cos \Pi(\alpha) \sin \Pi(a)$.
- $\sin \Pi(\alpha) = \cos \Pi(\beta) \sin \Pi(b)$.
- $\cos \Pi(b) = \cos \Pi(c) \cos \Pi(\alpha)$.
- $\cos \Pi(a) = \cos \Pi(c) \cos \Pi(\beta)$.

PROOF

Let $\triangle ABC$ be as described in the hypothesis of the theorem. Let $\overline{AA'}$ be perpendicular to ABC at A and let $\overline{BB'}$ and $\overline{CC'}$ be parallel to $\overline{AA'}$. With axis $\overline{AA'}$ describe a horosphere that contains A and intersects $\overline{BB'}$ at B'' and $\overline{CC'}$ at C''. These three points are joined by horocycles. Each is the intersection of the horosphere with a plane that contains an axis of the horosphere. Since the angle between two of these horocycles is defined to be the angle between the two planes and letting $p = B''C''$, $q = AC''$, and $r = AB''$, we see that $\triangle AB''C''$ on the surface of the horosphere can be described by

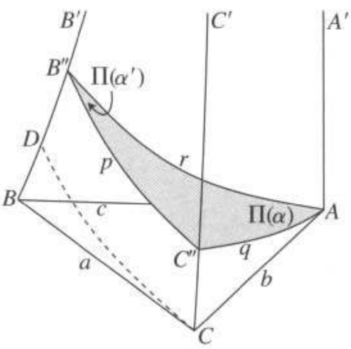

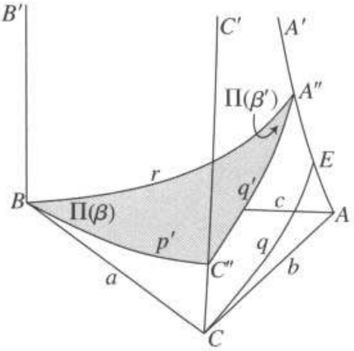

vertex	measure	opposite side
A	$\Pi(\alpha)$	p
B''	$\Pi(\alpha')$	q
C''	$\pi/2$	r

We may apply plane trigonometry because the induced geometry on the horosphere is Euclidean (Theorem 12.4.3) and write

$$q = r \cos \Pi(\alpha). \tag{12.13}$$

Now reverse the setup by erecting the perpendicular at B, but still let $\overline{AA'}$, $\overline{BB'}$, and $\overline{CC'}$ be parallel. Place a horosphere with axis $\overline{BB'}$ at B and let it intersect $\overline{AA'}$ at A'' and $\overline{CC'}$ at C''. Let $q' = C''A''$ and $p' = BC''$. Since all horocycles have the same curvature, $BA'' = r$. Place a horocycle at C so that it is concentric with $C''A''$. Its length is q. Since all horocycles are congruent, there exists a function f that measures the distance from the plane ABC to a horocycle along an axis. That is, $f(b) = AE$, $f(a) = CC'' = EA''$, and we may conclude that

$$f(c) = AA'' = AE + EA'' = f(a) + f(b). \tag{12.14}$$

Then by Theorem 12.2.9, $q = q'e^{kf(a)}$ for some $k > 0$, and again since the induced geometry on a horosphere is Euclidean, we have $q' = r \sin \Pi(\beta)$, so that

$$q = re^{kf(a)} \sin \Pi(\beta). \tag{12.15}$$

Therefore, setting Equation 12.13 equal to Equation 12.15 gives

$$\cos \Pi(\alpha) = e^{kf(a)} \sin \Pi(\beta). \tag{12.16}$$

A similar construction of $\triangle ABC$ with the two nonright angles having measures $\Pi(b')$ and $\Pi(c)$ will yield

$$\sin \Pi(b) = e^{kf(a)} \sin \Pi(c). \tag{12.17}$$

because $\cos \Pi(b') = \sin \Pi(b)$ (Exercise 5). Therefore, by multiplying both sides of Equation 12.17 by $e^{kf(b)}$, we have

$$e^{kf(b)} \sin \Pi(b) = e^{kf(c)} \sin \Pi(c)$$

by Equation 12.14. An argument similar to the above will yield

$$e^{kf(a)} \sin \Pi(a) = e^{kf(b)} \sin \Pi(b).$$

Finally, we note that the lengths of a and b may be chosen independently of each other. Hence, since $f(0) = 0$ and $\Pi(0) = \pi/2$, let $b = 0$ and conclude that

$$\sin \Pi(a) = e^{-kf(a)}.$$

Therefore, by Equation 12.17,

$$\sin \Pi(c) = \sin \Pi(a) \sin \Pi(b),$$

and then by Equation 12.16,

$$\sin \Pi(\beta) = \cos \Pi(\alpha) \sin \Pi(a).$$

The remaining three identities follow by a "mutation of the letters." ■

We close the section with the remarkable result that the geometry on the surface of a sphere where *line* is interpreted as *great circle* is independent of the hypothesis chosen. We accomplish this by examining the sphere described in Figure 12.11 and the corresponding table at the top of page 36.

■ **COROLLARY 12.4.5 [*Theory of Parallels* 35, *Appendix* 26]**

The geometry on the surface of a sphere is the same whether **HR** or **HA** is assumed.

PROOF

If we assign $a = \Pi(c)$, $b = \Pi(\beta)$, $c = \Pi(a)$, $A = \Pi(b)$, and $B = \Pi(\alpha')$ for the dimension of a triangle on a unit sphere, the identities of Theorem 12.4.4 are seen to be identical to the identities of Theorem 12.3.2. ■

Exercises

1. Let A, B, and C be points on a horosphere $\mathcal{F}$ but not all on the same horocycle. The three horocycles through these points form a triangle on $\mathcal{F}$. Show that if $\sphericalangle C$ is right, then $AB^2 = AC^2 + BC^2$.

2. Let $\mathcal{L}$ be a horocycle on the surface of a horosphere $\mathcal{F}$ with axis $\overline{A\Omega}$. Let B be on $\mathcal{F}$ but not on $\mathcal{L}$. Prove:
   (a) B corresponds with A.
   (b) B corresponds with every other point on $\mathcal{F}$.

3. Take A, B, and C to be points on the surface of a horosphere but not all on the same horocycle. Show that A, B, and C lie on a circle on the horosphere. (*Hint:* Find the intersection of the lines that are the perpendicular bisectors of the sides of the triangle.)

4. Redo Exercise 3 using another method.

5. Let $\triangle ABC$ be as in the proof of Theorem 12.4.4 but such that $\text{m}\sphericalangle A = \Pi(b')$, $\text{m}\sphericalangle B = \Pi(c)$, $\text{m}\sphericalangle C = \pi/2$, $BC = a$, $AC = \alpha$, and $AB = \beta$.
   (a) Show $\cos \Pi(b') = \sin \Pi(b)$.
   (b) Use $\triangle ABC$ to demonstrate Equation 12.17.

6. Suppose that spherical $\triangle B_1B_2B_3$ in Figure 12.11 was, instead, constructed at A. Provided that $\overline{AA'}$ is still perpendicular to plane ABC, find the measures of the angles and sides of $\triangle B_1B_2B_3$, where B_3 is the intersection of $\overline{AA'}$ with the sphere.

7. Let f be the function defined in the proof of Theorem 12.4.4. In the first diagram of that theorem, draw a horocycle through C that is concentric to the horocycle $B''C''$ and intersects $\overline{BB''}$ at D. Using B, B'', and D, show that $f(c) = f(a) + f(b)$.

8. Give the details for the "mutation of letters" from the proof of Theorem 12.4.4:
   (a) $\sin \Pi(\alpha) = \cos \Pi(\beta) \sin \Pi(b)$.
   (b) $\cos \Pi(b) = \cos \Pi(c) \cos \Pi(\alpha)$.
   (c) $\cos \Pi(a) = \cos \Pi(c) \cos \Pi(\beta)$.

9. Provide the details for the steps labeled as "similar" in the proof of Theorem 12.4.4.

12.5 EVALUATING THE PI FUNCTION

We have seen that Lobachevski's Π function has some very nice properties. In this section we derive a formula for it. The proof given is by Lobachevski with some modification by Manning (1901, 45–46, 58).

■ **LEMMA 12.5.1**

If n is a positive integer and

$$f(x) = \frac{f(2x)}{f(x)} = \frac{f(3x)}{f(2x)} = \cdots = \frac{f(nx)}{f([n-1]x)},$$

then $f(nx) = [f(x)]^n$.

PROOF

By hypothesis we conclude that

$$f(nx) = \frac{[f([n-1]x)]^2}{f([n-2]x)},\ f([n-1]x) = \frac{[f([n-2]x)]^2}{f([n-3]x)},\ \ldots,$$

$$f(3x) = \frac{[f(2x)]^2}{f(x)},\ f(2x) = [f(x)]^2.$$

Hence,

$$f(nx) = \frac{[f([n-1]x)]^2}{f([n-2]x)} = f([n-1]x)\frac{f([n-1]x)}{f([n-2]x)} = f([n-1]x)f(x),$$

with the last equality holding by the hypothesis. By a similar argument,

$$f([n-1]x) = f([n-2]x)f(x),$$

so

$$f(nx) = f([n-2]x)[f(x)]^2.$$

Continuing, we find that $f(nx) = [f(x)]^n$. ■

The next theorem contains the exact relationship between the length of a segment and its corresponding angle of parallelism. Lobachevski's formula did not include the constant k. Instead, he modified the units along the x-axis so that the constant could be taken to equal 1. Bolyai did not do this, so the constant remained an important part of his equation. His father was very much troubled that his son could never find a way to remove the constant. Farkas believed that the final description of the new geometry should be able to determine the value of k. János was content to leave it undetermined (Bonola 1912, 99).

■ **THEOREM 12.5.2** [*Theory of Parallels* **36**, *Appendix 29*]

[HA] There exists $k > 0$ such that for all $x \geqslant 0$,

$$\tan\frac{1}{2}\Pi(x) = e^{-kx}.$$

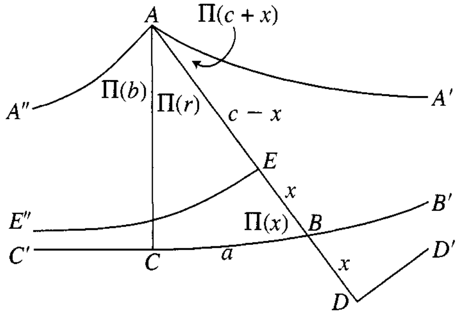

PROOF

We first note that if $x = 0$,

$$\tan\frac{1}{2}\Pi(0) = \tan\pi/4 = 1$$

because $\Pi(0) = \pi/2$, so the equation holds. So, let $x > 0$ and form $\triangle ABC$ with $m\sphericalangle ABC = \Pi(x)$ and $\sphericalangle ACB$ right. Further, since Π is onto (Theorem 12.1.6), there exists $r > 0$ such that $m\sphericalangle BAC = \Pi(r)$. To ease our calculations later, let $a = BC$, $b = AC$, and $c = AB$. Notice that the diagram assumes that $c > x$. Next extend $\overline{AB}$ to point D, where $BD = x$. Find D' so that $\overline{DD'}$ is perpendicular to $\overline{AD}$. Producing $\overline{CB}$ to B', we see that $\overline{BB'}$ is parallel to $\overline{DD'}$ because $m\sphericalangle B'BD = \Pi(x)$. Finally, draw $\overline{AA'}$ so it is parallel to $\overline{DD'}$. This gives us $m\sphericalangle A'AD = \Pi(c+x)$. Thus, by Theorem 12.1.1 and Theorem 12.1.3, $\overline{AA'}$ is also parallel to $\overline{CB'}$ and $m\sphericalangle A'AC = \Pi(b)$. Hence,

$$\Pi(b) = \Pi(r) + \Pi(c+x). \tag{12.18}$$

There are now two cases to consider. Either $c > x$ or it is not. Suppose that $c > x$ and find E on $\overline{AB}$ such that $AE = c - x$ and $EB = x$. At E construct $\overline{EE'}$ so that it meets $\overline{AB}$ in right angles and produce $\overline{BC}$ to C'. We see that $\overline{BC'}$ and $\overline{EE'}$ are parallel since $m\sphericalangle ABC = \Pi(x)$. Draw $\overline{AA''}$ parallel to $\overline{CC'}$. This means that we have $m\sphericalangle CAA' = \Pi(b)$. Therefore, $\overline{AA'}$ is parallel to $\overline{EE'}$ by Theorem 12.1.3. Thus, $m\sphericalangle EAA' = \Pi(c-x)$, and we conclude that

$$\Pi(b) = \Pi(c-x) - \Pi(r). \tag{12.19}$$

That this equation holds when $c \leqslant x$ is left to Exercise 2.

We may now combine Equations 12.18 and 12.19 to find that

$$\begin{aligned} 2\Pi(b) &= \Pi(c-x) + \Pi(c+x), \\ 2\Pi(r) &= \Pi(c-x) - \Pi(c+x). \end{aligned}$$

Therefore,

$$\frac{\cos\Pi(b)}{\cos\Pi(r)} = \frac{\cos\frac{1}{2}[\Pi(c-x)+\Pi(c+x)]}{\cos\frac{1}{2}[\Pi(c-x)-\Pi(c+x)]}.$$

By Theorem 12.4.4,

$$\cos\Pi(c) = \frac{\cos\Pi(b)}{\cos\Pi(r)},$$

and if we let $\theta = \frac{1}{2}\Pi(c-x)$ and $\phi = \frac{1}{2}\Pi(c+x)$, we find

$$\cos\Pi(c) = \frac{\cos(\theta+\phi)}{\cos(\theta-\phi)},$$

from which some trigonometric identities give

$$2\cos^2\tfrac{1}{2}\Pi(c) = \frac{\cos(\theta+\phi)+\cos(\theta-\phi)}{\cos(\theta-\phi)},$$

and then

$$\sec^2\tfrac{1}{2}\Pi(c) - 1 = \frac{\sin\theta\sin\phi}{\cos\theta\cos\phi}.$$

Therefore,

$$\tan^2\tfrac{1}{2}\Pi(c) = \tan\tfrac{1}{2}\Pi(c-x)\tan\tfrac{1}{2}\Pi(c+x).$$

For all positive integers n, let $y = c - nx$ so that

$$\tan^2 \tfrac{1}{2}\Pi(y + nx) = \tan \tfrac{1}{2}\Pi(y + (n-1)x) \tan \tfrac{1}{2}\Pi(y + (n+1)x),$$

and then

$$\frac{\tan \frac{1}{2}\Pi(y+x)}{\tan \frac{1}{2}\Pi(y)} = \frac{\tan \frac{1}{2}\Pi(y+2x)}{\tan \frac{1}{2}\Pi(y+x)} = \cdots = \frac{\tan \frac{1}{2}\Pi(y+nx)}{\tan \frac{1}{2}\Pi(y+(n-1)x)}.$$

Because $\tan \frac{1}{2}\Pi(0) = 1$, letting $y = 0$ yields

$$\tan \tfrac{1}{2}\Pi(x) = \frac{\tan \frac{1}{2}\Pi(2x)}{\tan \frac{1}{2}\Pi(x)} = \cdots = \frac{\tan \frac{1}{2}\Pi(nx)}{\tan \frac{1}{2}\Pi([n-1]x)}.$$

Hence, by Lemma 12.5.1,

$$\tan \tfrac{1}{2}\Pi(nx) = \tan^n \tfrac{1}{2}\Pi(x)$$

for all positive integers n. Since $\Pi(x) < \pi/2$, we have $\tan \frac{1}{2}\Pi(x) < 1$. Therefore, by Lemma 12.2.8 there exists $k > 0$ such that

$$\tan \tfrac{1}{2}\Pi(x) = e^{-kx}. \blacksquare$$

Since Π is undefined for negative numbers, Lobachevski extended the domain of Π to all real numbers by defining

$$\Pi(x) + \Pi(-x) = \pi. \tag{12.20}$$

With this definition we can show that $\tan \frac{1}{2}\Pi(x) = e^{-kx}$ for all real numbers x (Exercise 3).

The Graph of Pi

The Theorem 12.5.2 allows us to compute values for Π, for it follows quickly that

$$\Pi(x) = 2\tan^{-1} e^{-kx}.$$

The graph of this function is given in Figure 12.12 for $k = 1$. Notice that this confirms the basic properties of Π listed in Section 12.1.

We should now observe that there is a relationship between Π and the **hyperbolic functions**. For any real number x, define **hyperbolic cosine** and **hyperbolic sine** as

$$\cosh x = \frac{e^x + e^{-x}}{2},$$

$$\sinh x = \frac{e^x - e^{-x}}{2}.$$

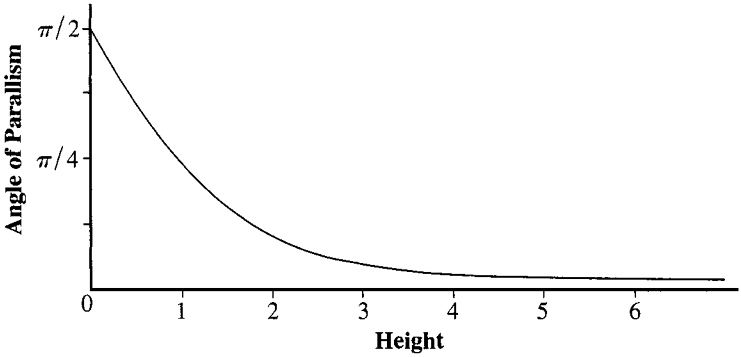

Figure 12.12 The graph of $\Pi(x) = 2\tan^{-1} e^{-x}$.

The hyperbolic versions of tangent, cotangent, secant, and cosecant are defined as their trigonometric counterparts. In addition, there are hyperbolic identities such as

$$1 - \tanh^2 x = \operatorname{sech}^2 x. \tag{12.21}$$

The hyperbolic functions can be used to help evaluate expressions that involve the composition of standard trigonometric functions and Π. For example, to evaluate $\cos \Pi(x)$, first note that

$$e^{-2kx} = \tan^2 \tfrac{1}{2}\Pi(x) = \frac{1 - \cos \Pi(x)}{1 + \cos \Pi(x)}.$$

Therefore,

$$\cos \Pi(x) + e^{-2kx} \cos \Pi(x) = 1 - e^{-2kx},$$

and then

$$\cos \Pi(x) = \frac{1 - e^{-2kx}}{1 + e^{-2kx}} = \frac{e^{kx} - e^{-kx}}{e^{kx} + e^{-kx}} = \frac{\sinh kx}{\cosh kx},$$

and we may conclude that

$$\cos \Pi(x) = \tanh kx. \tag{12.22}$$

We use this last equation to evaluate $\sin \Pi(x)$. From

$$\cos^2 \Pi(x) + \sin^2 \Pi(x) = 1$$

we may substitute and use Equation 12.21 to find

$$\sin^2 \Pi(x) = 1 - \cos^2 \Pi(x) = 1 - \tanh^2 kx = \operatorname{sech}^2 kx.$$

Therefore,

$$\sin \Pi(x) = \operatorname{sech} kx. \tag{12.23}$$

Pi in Space

Beginning in 1818 there was a large effort to map the entire Kingdom of Hanover. The interest in geodesy at the time was partly a theoretical one to use the methods of calculus to determine the shape of the globe and partly a practical one to create accurate maps to the benefit of commerce and the military. It was Gauss who was asked to lead the initial effort. He had previously made some geodetic studies after his return to Brunswick, so he was interested when he was offered the leadership position. Over a period of three summers, Gauss spent most of his time traveling. He led a small team of military men who took the measurements with a device called a heliotrope, which Gauss had invented (Bühler 1981, 95–98).

The work allowed Gauss an opportunity to conduct an experiment that would hopefully determine whether it was Euclidean or non-Euclidean geometry that was true. He decided that a check of the angle sum of a large triangle would provide a perfect test, so Gauss chose a triangle near Göttingen. Its vertices were mountain tops that had been surveyed previously, and its sides were approximately 40 miles long. Measured in degrees, the calculations yielded an angle sum for the triangle that was less than 2 seconds short of two right angles. This was well within the margin of error of the instruments that Gauss used, so he concluded that as far as measurement was concerned, it was safe to assume that the geometry near earth was Euclidean. However, since the defect did not calculate to zero, there was still some question (Burton 1985, 552).

Although not influenced directly by Gauss, Lobachevski shared his belief that empirical methods should determine the truth of the Parallel Postulate. Like Gauss, he rejected the prevalent Kantian philosophy of his day and believed that whether Euclid's geometry was true or not must be determined by experiment. Our senses must be allowed to inform our beliefs regarding mathematics. Our intuition and any notions that we may have pre-wired in our brains must not be trusted. To find his evidence, Lobachevski decided that he would need to examine large triangles. He did not use mountains; instead, Lobachevski found his large triangles in space (Rosenfeld 1988, 208).

Take the vertices of the triangle to be the points A, B, and C in Figure 12.13. Locate them at the centers of the earth, the sun, and a distant star, respectively. Assume that $\overline{AB}$ is perpendicular to $\overline{BC}$. With this setup Lobachevski used the

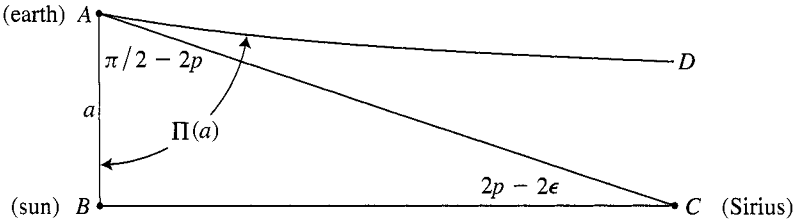

Figure 12.13 Lobachevski testing the geometry of space.

parallax angle, which is $\sphericalangle BCA$ in Figure 12.13, to measure how close to being Euclidean is space. Lobachevski chose Sirius for his calculations. Let $a = AB$,

$$\mathrm{m}\sphericalangle BAC = \frac{\pi}{2} - 2p, \text{ and } \mathrm{m}\sphericalangle BCA = 2p - 2\epsilon,$$

where $\epsilon \geqslant 0$ to allow for a possible positive defect in the triangle. Choose D so that $\sphericalangle BAD$ is the angle of parallelism. We conclude that

$$\Pi(a) > \frac{\pi}{2} - 2p,$$

and then

$$e^{-ka} = \tan\frac{1}{2}\Pi(a) > \tan\left(\frac{\pi}{4} - p\right) = \frac{1 - \tan p}{1 + \tan p}$$

for some constant k. Therefore, we conclude that

$$e^{ka} < \frac{1 + \tan p}{1 - \tan p}.$$

We know from calculus that if $|x| < 1$, then

$$\ln\frac{1+x}{1-x} = 2\left(x + \frac{x^3}{3} + \frac{x^5}{5} + \cdots + \frac{x^{2n+1}}{2n+1} + \cdots\right).$$

Since $0 < p < \pi/4$, we have $\tan p < 1$, and hence

$$ka < \ln\frac{1 + \tan p}{1 - \tan p} = 2\left(\tan p + \frac{\tan^3 p}{3} + \frac{\tan^5 p}{5} + \cdots + \frac{\tan^{2n+1} p}{2n+1} + \cdots\right).$$

Moreover,

$$2(\tan p + \tan^3 p + \tan^5 p + \cdots + \tan^{2n+1} p + \cdots) = \frac{2\tan p}{1 - \tan^2 p} = \tan 2p,$$

and we finally have that

$$ka < \tan 2p.$$

Lobachevski took the parallax of Sirius to be 1.24 arcseconds, but we know that it is closer to .375 arcseconds. Using the value of 93 million miles for the radius of earth's orbit, a little calculation yields

$$k < 1.95 \times 10^{-14}.$$

Therefore, Lobachevski drew the same conclusion as Gauss. Since $k = 0$ means that $\Pi(x) = \pi/2$ for all x, the very small k tells us that if the universe is not Euclidean, it is nearly so (Bonola 1912, 94–96).

The Imaginary

Lambert believed that his geometry would hold true on a sphere of imaginary radius (page 360). He did not explain how to visualize this, but if we take Lobachevski's equations and replace radius r with radius ir, we obtain familiar formulas from spherical geometry. To see how this works, we must make two notes. First, since the length of an arc of a circle with real radius r and central angle θ is $r\theta$,

> the length of an arc of a great circle on a sphere of imaginary radius is $ri\theta$.

Second, we need to extend the domain of Π to include any complex number. We already know that $\Pi(x) = 2\tan^{-1} e^{-kx}$. Without going into the details, to calculate $\Pi(a+bi)$, use the power series expansion for $2\tan^{-1} e^{-kx}$ and then substitute $a+bi$ in for x.

We now apply trigonometric functions to $\Pi(xi)$. We prove the first formula and leave the other two to Exercise 9. Using the power series for hyperbolic cosine, we have

$$\begin{aligned}\cosh xi &= 1 + \frac{(xi)^2}{2!} + \frac{(xi)^4}{4!} + \frac{(xi)^6}{6!} + \cdots \\ &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots \\ &= \cos x.\end{aligned}$$

Therefore, by Equation 12.23,

$$\sin\Pi(xi) = \sec kx.$$

The other two substitutions are

$$\begin{aligned}\cos\Pi(xi) &= i\tan kx, \\ \tan\Pi(xi) &= -i\csc kx.\end{aligned}$$

Lobachevski used these equations to translate his trigonometric equations into those of spherical trigonometry. He showed (*Theory of Parallels* 37) that

$$\sin A\tan\Pi(a) = \sin B\tan\Pi(b), \tag{12.24}$$

$$\cos A\cos\Pi(b)\cos\Pi(c) + \frac{\sin\Pi(b)\sin\Pi(c)}{\sin\Pi(a)} = 1, \tag{12.25}$$

$$\cot A\sin C\sin\Pi(b) + \cos C = \frac{\cos\Pi(b)}{\cos\Pi(a)}, \tag{12.26}$$

$$\cos A + \cos B\cos C = \frac{\sin B\sin C}{\sin\Pi(a)}. \tag{12.27}$$

Then on a sphere of imaginary radius the second equation becomes

$$\cos A \cos \Pi(bi) \cos \Pi(ci) + \frac{\sin \Pi(bi) \sin \Pi(ci)}{\sin \Pi(ai)} = 1.$$

Using the substitutions above we obtain

$$\cos A(i \tan kb)(i \tan kc) + \frac{\sec kb \sec kc}{\sec ka} = 1,$$

which gives

$$-\cos A \cdot \frac{\sin kb \sin kc}{\cos kb \cos kc} + \frac{\cos ka}{\cos kb \cos kc} = 1.$$

Multiplying through by $\cos kb \cos kc$ yields

$$-\cos A \sin kb \sin kc + \cos ka = \cos kb \cos kc.$$

Thus,

$$\cos ka = \cos kb \cos kc + \sin kb \sin kc \cos A,$$

and this is Theorem 12.3.5. It is left to Exercise 10 to show that similar work with Lobachevski's trigonometric formulas gives

$$\sin A \sin kb = \sin B \sin ka$$

and

$$\cos A = \cos ka \sin B \sin C - \cos B \cos C.$$

These equations are the spherical trigonometry identities found in Theorem 12.3.4 and Exercise 12 of Section 12.3. This is what it means for this new geometry to be found on the surface of a sphere of imaginary radius.

Exercises

1. Confirm the following using the equation for $\Pi(x)$ given by Theorem 12.5.2.
    (a) Equation 12.1
    (b) Π is a strictly decreasing continuous function.

2. Prove Equation 12.19 for the cases $c = x$ and $c < x$.

3. Use Lobachevski's extension of Π in Equation 12.20 to show that for all real numbers x, $\tan \frac{1}{2}\Pi(x) = e^{-kx}$.

4. Draw the following, paying careful attention to all measurements.
    (a) The graph of Π when $k = .1, k = .5, k = 2, k = 10$.
    (b) A 60 – 90 omega triangle ($k = 1$).
    (c) $\triangle ABC$ where $m\sphericalangle C = 90°$, $m\sphericalangle B = 60°$, and $AC = 1$ ($k = 1$).

5. The area of $\triangle ABC$ is $k^2(\pi - A - B - C)$. Use this to explain why the maximum area of a triangle πk^2?

6. Find the area of $\triangle ABC$ where $m\sphericalangle C = 90°$, $m\sphericalangle B = 60°$, and $AC = 1$ $(k = 1)$.
7. Give a formula for the area of a quadrilateral $ABCD$ and find its maximum area.
8. Use the value of 93 million miles for the radius of earth's orbit about the sun.
   (a) Confirm that if the parallax of Sirius is .375 arcseconds, $k < 1.95 \times 10^{-14}$.
   (b) Find the maximum value of k using Lobachevski's value of 1.24 arcseconds for the parallax of Sirius.
   (c) Calculate $\Pi(1)$ using both of these k values.
   (d) Find the distance to Sirius assuming Euclidean geometry and assuming that $k = 1.95 \times 10^{-14}$. Is the difference significant?
9. Confirm that the following hold.
   (a) $\cos \Pi(xi) = i \tan kx$.
   (b) $\tan \Pi(xi) = i \csc kx$.
10. Show that for triangles on a sphere of imaginary radius,
   (a) $\sin A \sin kb = \sin B \sin ka$.
   (b) $\cos A = \cos ka \sin B \sin C - \cos B \cos C$.
11. The circumference of a circle with radius r is

$$C = \frac{2\pi}{k} \sinh kr,$$

and its area is

$$A = \frac{4\pi}{k^2} \sinh^2 2kr.$$

Confirm that these equations approach the Euclidean formulas as $k \to 0$.

12. For very small values of x,

$$\cos \Pi(x) \approx kx, \sin \Pi(x) \approx 1 - \frac{k^2x^2}{2}, \text{ and } \cot \Pi(x) \approx kx.$$

   (a) Confirm that this is the case.
   (b) Use these approximations to show that Lobachevski's trigonometric equations (12.24–12.27) are equivalent to

$$b \sin A = a \sin B,$$

$$a^2 = b^2 + c^2 - 2bc \cos A,$$

$$a \sin(A + C) = b \sin A,$$

$$\cos A + \cos(B + C) = 0.$$

   (c) Use these results to show that the angle sum of very small triangles is approximately equal to two right angles, showing that in the geometry of Lobachevski and Bolyai, very small regions of the plane are essentially Euclidean.

CHAPTER 13

BERNHARD RIEMANN

Riemann was born in Breselenz in Hanover on September 17, 1826. His full name was Georg Friedrich Bernhard Riemann, and he was the son of a Lutheran pastor. Following his father's wishes, he moved to Göttingen in 1846 to study theology and philosophy. After attending various mathematics courses, he realized that he had an aptitude for the subject, so after receiving permission from his father, Riemann switched fields. Unfortunately, he found himself unhappy at the university. Despite Gauss holding a chair in mathematics, Riemann was disappointed in the level of academic rigor of the courses. Because of this he moved to Berlin and studied mathematics with such notables as Dirichlet, Eisenstein, and Jacobi (Burton 1985, 563; Boyer and Merzbach 1991, 545).

In 1848, revolution embroiled Berlin, and Riemann found himself in the middle of something that was quite foreign to his nature. Riemann was a frail, unassuming person, but evidently motivated by a sense of duty to his country, he joined a group of students who were loyal to the crown. His activities for the group included a two-day stint in which he stood guard at the royal palace. This proved to be too much excitement for Riemann. He decided that it was time to leave, and in 1849 he returned to Göttingen to finish his doctoral degree. With Gauss as his dissertation advisor, he worked on a topic that was new and significant. It involved surfaces

Revolutions of Geometry. By Michael O'Leary
Copyright © 2010 John Wiley & Sons, Inc.

over the complex plane that we now call **Riemann surfaces**. The dissertation also included equations that were discovered independently by Augustin-Louis Cauchy. They are called the **Cauchy-Riemann equations**, and the proposition states that if f is a complex differentiable function of the form

$$f(x, y) = u(x, y) + iv(x, y),$$

then

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \text{ and } \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}.$$

Gauss recognized in his student exceptional ability and was pleased with his results (Burton 1985, 563–564; Boyer and Merzbach 1991, 545).

Over the next few years Riemann continued his research. Besides his mathematical pursuits, he studied physics as an assistant to Wilhelm Weber. As a result of his work, Riemann published numerous papers. The topics included heat, light, gases, magnetism, gravitation, fluid dynamics, and acoustics. Riemann even conjectured that space was curved and that this was the reason that gravity works on objects at a distance (Kline 1972, 655–656; Gray 2007, 189).

Then in 1854 Riemann was ready to become a *privatdozent* at Göttingen. For his *Habilitationsschrift*, Riemann gave three topics to Gauss. The first involved using a trigonometric series to represent a function. The second dealt with quadratic equations. The last was titled *On the Hypotheses Which Lie at the Bases of Geometry*. Riemann was quoted as being surprised at Gauss going against the "usual convention" by picking his last suggestion, but Riemann should not have been surprised at the choice. The first two were too elementary for the presentation. Moreover, there is evidence that Riemann was both willing and ready for the topic (Laugwitz 1999, 219–220; Gray 2007, 189).

On June 10, 1854 Riemann gave his talk. Since mathematics was housed in the school of philosophy at Göttingen, the audience included both mathematicians and philosophers. For this reason his essay (Riemann 1854) and the accompanying presentation focused on big-picture issues and provided little detail. Despite the subject being geometry, Riemann included no diagrams and only two equations. Nonetheless, in this one presentation, Riemann gave a completely new definition of geometry. From the time of Euclid geometry had been the study of objects in space, and it is the choice of postulates that determines the behavior of these objects. Up until the time of Gauss, these postulates were viewed as self-evident truths. Gauss's work, along with that of Lobachevski and Bolyai, showed otherwise by proving that the Parallel Postulate may or may not describe space accurately. Its truth is now an empirical question, one Riemann, as Lobachevski before him (page 436), believed would be answered by astronomy (Burton 1985, 565).

The goal of Riemann's redefinition was to generalize the system of Euclid. He did not simply change the postulate set as did Lobachevski and Bolyai. In fact, it is doubtful that he knew their work. Instead, Riemann set his focus on the two different types of problems of the geometer. The first seeks to determine the intrinsic properties of a surface. An **intrinsic** property is one that requires no information about the encompassing space. Imagine little creatures living on a surface. They would not need to leave the surface to determine an intrinsic property. An example is the measure of angles. One of the creatures can easily take a little protractor and find the measure of any given angle without ever leaving home. The second type of geometric problem involves the examination of the various methods by which one surface can be mapped onto another. Sometimes these functions distort the surface by bending or warping it, but sometimes the maps simply superimpose one surface on top of the other. We do this, for example, when we study congruent figures. Congruence means that the figures look identical in size and shape so that one can be lined up on top of the other (Gray 2004, 85; Laugwitz 1999, 237).

Since geometry assumes that surfaces exist in three dimensions, Riemann needed a new geometric notion that could generalize the Euclidean concept of space. He accomplished this with what he called a **manifoldness** (Riemann 1854, I.1, 57). A manifoldness was a multidimensional space consisting of points. These points were identified by the use of coordinates, and the number of coordinates needed gave the number of dimensions of the space. Today we use the term *manifold* instead of *manifoldness*, and we define a **manifold** to be a topological space that has the property that every point is contained within a neighborhood that is essentially Euclidean. The primary example of a manifold is $\mathbb{R}^n$. Other examples include the surfaces of spheres, cylinders, and tori.

Riemann's manifoldness had the additional property that the length of any curve in the space was **rectifiable**; that is, its length could be measured. This was done using infinitesimally small rulers. Hence, instead of geometric postulates Riemann used the methods of calculus to study his manifolds. This field of mathematics that studies curves and surfaces using calculus is called **differential geometry** (Gray 2004, 85).

13.1 METRIC SPACES

Riemann taught that all that is needed for a geometry are its points and a function to compute distances. Since there are infinitely many such functions, there are infinitely many geometries. This implies the concept of a metric space (O'Leary 2002, 325–348; Marsden and Hoffman 1993; Simmons 1963). Although it came after Riemann's work, a metric space shares two main features with his manifoldness, namely, a set of points and a distance function.

■ DEFINITION 13.1.1

Let M be a nonempty set and $d : M \times M \to \mathbb{R}$ a function such that for all $x, y, z \in M$:

- $d[x, y] \geqslant 0$.
- $d[x, y] = 0$ if and only if $x = y$.
- $d[x, y] = d[y, x]$.
- $d[x, y] \leqslant d[x, z] + d[z, y]$.

The pair (M, d) is called a **metric space**. The function d is the distance function, and it is called a **metric**. The last condition is called the **Triangle Inequality**.

There are many examples of metric spaces. For example, $(\mathbb{R}, d)$ is a metric space where

$$d[x, y] = |x - y|.$$

The check of the first three conditions is left to Exercise 1. To see that the Triangle Inequality holds, notice that for all $a, b \in \mathbb{R}$,

$$-|a| \leqslant a \leqslant |a| \text{ and } -|b| \leqslant b \leqslant |b|.$$

Adding these inequalities yields

$$-(|a| + |b|) \leqslant a + b \leqslant |a| + |b|.$$

Therefore,

$$|a + b| \leqslant |a| + |b|. \tag{13.1}$$

Similarly, if we define the distance function d as

$$d[X, Y] = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

for all $X = (x_1, y_1)$ and $Y = (x_2, y_2)$, we see that $(\mathbb{R}^2, d)$ is a metric space (Exercise 1). This metric can be generalized to any dimension. For all

$$X = (x_1, x_2, \ldots, x_n),$$
$$Y = (y_1, y_2, \ldots, y_n),$$

define

$$d[X, Y] = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}.$$

Then $(\mathbb{R}^n, d)$ is a metric space. These distance functions are known as the **standard metrics** on $\mathbb{R}^n$, and unless otherwise noted, they will be the assumed metrics for these spaces.

For another example of a metric space, let $C[a, b]$ be the set of all functions continuous on $[a, b]$ and define for all $f, g \in C[0, 1]$,

$$d[f, g] = \int_0^1 |f(x) - g(x)|\, dx.$$

This is a natural definition for the distance between two functions because the closer two functions are to each other, the smaller the area between them will be (Figure 13.1). For instance, to measure the distance between $f(x) = x^2$ and $g(x) = x$ over the interval $[0, 1]$, we compute

$$d[f, g] = \int_0^1 |x^2 - x|\, dx = 1/6.$$

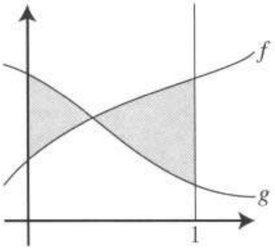

Figure 13.1 Measuring the distance between two functions using area.

Now check that $(C[0,\ 1],\ d)$ is a metric space. Since it is the most involved, we will check first that the Triangle Inequality works. Take $f,\ g,\ h \in C[0,\ 1]$ and calculate:

$$\begin{aligned}
d[f,\ g] &= \int_0^1 |f(x) - g(x)|\,dx \\
&= \int_0^1 |f(x) - h(x) + h(x) - g(x)|\,dx \\
&\leqslant \int_0^1 (|f(x) - h(x)| + |h(x) - g(x)|)\,dx \\
&= \int_0^1 |f(x) - h(x)|\,dx + \int_0^1 |h(x) - g(x)|\,dx \\
&= d[f,\ h] + d[h,\ g].
\end{aligned}$$

The inequality holds due to Equationl13.1. For the other three:

- Since $|f(x) - g(x)| \geqslant 0$ for all x,
$$d[f,\ g] = \int_0^1 |f(x) - g(x)|\,dx \geqslant 0.$$
- If $d[f,\ g] = 0$, this means that the area under the curve $y = |f(x) - g(x)|$ on the interval $[0,\ 1]$ equals 0. Since $|f(x) - g(x)|$ is never negative, the only way that this can happen is when $f = g$. Conversely, if $f = g$, then $|f(x) - g(x)| = 0$ for all x. Hence, $d[f,\ g] = 0$.
- Because $|f(x) - g(x)| = |g(x) - f(x)|$, we have that $d[f,\ g] = d[g,\ f]$.

Therefore, $(C[0,\ 1],\ d)$ is a metric space.

Open Sets

Manifolds have the property that the nearby region surrounding any point of the space resembles $\mathbb{R}^n$. In the one-dimensional case, this means that the region looks like an open interval. To generalize this to higher dimensions, we need our first definition.

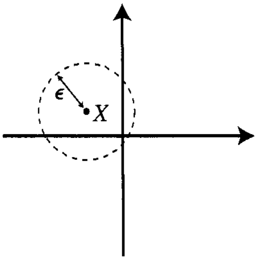

Figure 13.2 The open disk $D(x, \epsilon)$ in $\mathbb{R}^2$.

■ DEFINITION 13.1.2

Let (M, d) be a metric space and take $x \in M$. For all $\epsilon > 0$, define

$$D_M(x, \epsilon) = \{y \in M : d[x, y] < \epsilon\}.$$

This set is called an **open disk** about x in M.

If the metric space M is clear from context, we can write $D(x, \epsilon)$ for $D_M(x, \epsilon)$. For example, if $M = \mathbb{R}$ with the standard metric,

$$D(x, \epsilon) = (x - \epsilon, x + \epsilon),$$

and if $M = \mathbb{R}^2$ with the standard metric, $D(X, \epsilon)$ is the interior of a circle excluding its circumference. This is illustrated in Figure 13.2, where the dashed lines represent the exclusion of the circumference of the circle.

The open disk will allow us to make our main definition.

■ DEFINITION 13.1.3

Let M be a metric space. A subset U of M is called **open** in M if for all $x \in U$, there exists $\epsilon > 0$ such that $D_M(x, \epsilon) \subseteq U$.

We should note that for any metric space M, both M and $\varnothing$ are open. The is because $D_M(x, \epsilon)$ is a subset of M for any $x \in M$ and $\epsilon > 0$. As for the empty set, the proposition

$$\text{if } x \in \varnothing, \text{ then there exists } \epsilon > 0 \text{ such that } D_M(x, \epsilon) \subseteq \varnothing$$

is true for all x because $x \in \varnothing$ is always false (see page 38).

Here are some other examples of open sets.

- Let $a, b \in \mathbb{R}$. To prove that (a, b) is open, let $a < x < b$. We must isolate x inside an open interval within (a, b). To do this, find the distance from x to each endpoint and let ϵ be the minimum of the two distances. The notation that we will use for this is:

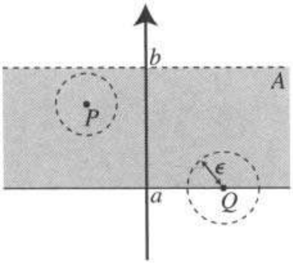

Figure 13.3 The set A is not open.

$$\epsilon = \min\{x - a,\ b - x\}.$$

Now let $y \in D(x,\ \epsilon)$. This means that $x - \epsilon < y < x + \epsilon$, and we may conclude that

$$a = x - (x - a) \leqslant x - \epsilon < y < x + \epsilon \leqslant x + (b - x) = b.$$

Therefore, $D(x,\ \epsilon) = (x - \epsilon,\ x + \epsilon) \subseteq (a,\ b)$.

- We can use the strategy of the previous example on other sets. Let A equal the set $(1,\ 2) \cup (4,\ 7)$. To show that A is open, pick a real number x such that $1 < x < 2$ or $4 < x < 7$. Define

  $$\epsilon = \min\{x - 1,\ 2 - x,\ x - 4,\ 7 - x\}.$$

  With this definition, $(x - \epsilon,\ x + \epsilon) \subseteq A$.

To show that a set A is not open in a metric space M, we must prove that there exists $x \in A$ such that $D(x,\ \epsilon)$ is not a subset of A for all $\epsilon > 0$. For this example, let $a,\ b \in \mathbb{R}$ so that $a < b$. The set

$$A = \{(x,\ y) \in \mathbb{R}^2 : a \leqslant y < b\}$$

is not open in $\mathbb{R}^2$ (Figure 13.3). Any point P that is on the interior of A can be surrounded by a disk that is a subset of A. However, this is not true for a point on the bottom border, so take Q to be $(3,\ a)$. Then for all $\epsilon > 0$,

$$D((3,\ a),\ \epsilon)\backslash A \neq \varnothing$$

because $(3,\ a - \epsilon/2)$ is a member of $D((a,\ 3),\ \epsilon)$ but not of A.

Now that we have seen examples of open and nonopen sets, we should prove that the disk is actually open and justify its name. The proof is a generalization of the fact that $(a,\ b)$ is open in $\mathbb{R}$.

■ THEOREM 13.1.4

$D(x,\ \epsilon)$ is open for any x in a metric space M.

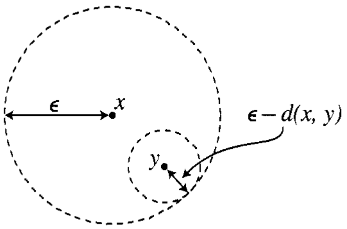

PROOF

We must show that we can isolate any element of $D(x, \epsilon)$ inside an open disk. So let $y \in D(x, \epsilon)$. The open disk about y that we want is

$$D(y, \epsilon - d[x, y]).$$

To show that this disk is included in $D(x, \epsilon)$, let $z \in D(y, \epsilon - d[x, y])$. This means that

$$d[y, z] < \epsilon - d[x, y],$$

and hence

$$d[x, z] \leqslant d[x, y] + d[y, z] < \epsilon$$

by the Triangle Inequality. We conclude that $z \in D(x, \epsilon)$. ■

An application of the definition is one way to prove that a set is open. Other ways include the next two theorems of the section.

■ THEOREM 13.1.5

Let M be a metric space.

- The union of a collection of open sets is open.
- The intersection of a finite collection of open sets is open.

PROOF

To prove the first part, let $\mathcal{F} = \{U_i : i \in I\}$ be a family of open sets. Take $x \in \bigcup \mathcal{F}$. This means that there exists $i \in I$ such that $x \in U_i$. Since U_i is open, we have $\epsilon > 0$, so that $D(x, \epsilon) \subseteq U_i$, but this implies that $D(x, \epsilon) \subseteq \bigcup \mathcal{F}$.

For the second part let $x \in U_1 \cap U_2 \cap \cdots \cap U_k$, where k is a positive integer ($k \in \mathbb{Z}^+$). By definition, x is an element of each U_i. Since each of these is open, there exists $\epsilon_i > 0$ so that $D(x, \epsilon_i)$ is a subset of U_i for $i = 1, 2, \ldots, k$. Define

$$\epsilon = \min\{\epsilon_1, \epsilon_2, \ldots, \epsilon_k\}.$$

Then $\epsilon \leqslant \epsilon_i$ for all i, and this means that $D(x, \epsilon) \subseteq D(x, \epsilon_i)$. Hence, $D(x, \epsilon)$ is a subset of the intersection. ■

To see that the finiteness condition is necessary for the intersection of open sets to be open, consider $\{(-1/n, 1/n) : n \in \mathbb{Z}^+\}$. This is a family of open sets in $\mathbb{R}$.

However,

$$\bigcap_{n=1}^{\infty}(-1/n,\ 1/n) = \{0\},$$

which is not open in $\mathbb{R}$.

The next definition relies on Theorem 13.1.5.

DEFINITION 13.1.6

If M is a metric space and A a subset of M, define

$$\text{int}(A) = \bigcup\{U : U \text{ is open in } M \text{ and } U \subseteq A\}.$$

This set is called the **interior** of A.

Since it is the union of open sets, the interior of a set A is open. Moreover, it must be the largest open subset of A (Exercise 14). As examples, $\text{int}([0,\ 1]) = (0,\ 1)$ and $\text{int}(\mathbb{Z}) = \varnothing$ in $\mathbb{R}$, because the empty set is the only open subset of $\mathbb{Z}$. In $\mathbb{R}^2$,

$$\text{int}([0,\ 1] \times [2,\ 4]) = (0,\ 1) \times (2,\ 4)$$

and

$$\text{int}(\{(x,\ y) \in \mathbb{R}^2 : xy \leqslant 1\}) = \{(x,\ y) \in \mathbb{R}^2 : xy < 1\}.$$

One way to identify the points in a set's interior is to use the next theorem.

THEOREM 13.1.7

If A is a subset of a metric space, then $x \in \text{int}(A)$ if and only if there exists $\epsilon > 0$ so that $D(x,\ \epsilon) \subseteq A$.

PROOF

First let $x \in \text{int}(A)$. This means that there exists an open subset U of A that contains x. So there exists $\epsilon > 0$ such that $D(x,\ \epsilon) \subseteq U \subseteq A$.

Conversely, if $D(x,\ \epsilon) \subseteq A$ for some $\epsilon > 0$, then x must be an element of the interior because $D(x,\ \epsilon)$ is open. ■

For instance, define $\mathbb{Z} = \{\ldots,\ -2,\ -1,\ 0,\ 1,\ 2,\ \ldots\}$ and view it as a metric space with the standard metric from $\mathbb{R}$. Let $A = \{2k : k \in \mathbb{Z}\}$, the set of even integers. The interior of A is A. To see this, let $n \in \mathbb{Z}$. Then

$$D_{\mathbb{Z}}(2n,\ 1) = \{m \in \mathbb{Z} : |2n - m| < 1\} = \{2n\}.$$

Therefore, $2n \in \text{int}(A)$, which implies that $A \subseteq \text{int}(A)$. Since we already know that $\text{int}(A) \subseteq A$, we conclude that $\text{int}(A) = A$. This means that A is open in $\mathbb{Z}$.

Closed Sets

The difference between an open interval and a closed one lies with the endpoints. Open intervals do not contain the endpoints. Instead, the endpoints belong to the complement of the open interval. A closed interval will contain its endpoints. We will take this idea and define what it means for a set to be closed in an arbitrary metric space and then prove some results.

DEFINITION 13.1.8

A subset A of a metric space M is **closed** in M if $M \backslash A$ is open in M.

Both M and $\varnothing$ are closed because $M \backslash M = \varnothing$ and $M \backslash \varnothing = M$ and both of these are open. Letting $A^c = M \backslash A$, here are some examples of closed sets.

- $[a, b]$ is closed in $\mathbb{R}$ for all real numbers $a < b$.
- Since $D(x, \epsilon)$ is open, $D(x, \epsilon)^c$ is closed.
- In $\mathbb{R}^2$ the set $\{(x, y) \in \mathbb{R}^2 : xy \geqslant 1\}$ is closed.

There are sets that are neither open nor closed. In $\mathbb{R}$, $(0, 1]$ provides a simple example. The set $A = \{(x, y) \in \mathbb{R}^2 : a \leqslant x < b\}$ is such an example in $\mathbb{R}^2$. Here is a more interesting one. Let $A = \{1/k : k \in \mathbb{Z}^+\}$.

- To see that A is not open, let n be a positive integer and consider
$$D(1/n, \epsilon) = (1/n - \epsilon, 1/n + \epsilon)$$
for any $\epsilon > 0$. Since there exists $m \in \mathbb{R}$ such that
$$\max\left\{\frac{1}{n+1}, \frac{1}{n} - \epsilon\right\} < m < \frac{1}{n},$$
$m \in D(1/n, \epsilon)$ but $m \notin A$. Hence, A is not open.
- To check that A is not closed, we show that A^c is not open. Suppose that $\epsilon > 0$. Then there exists a positive integer n such that $1/n < \epsilon$. Thus, $D(0, \epsilon) \cap A \neq \varnothing$, which means that $D(0, \epsilon)$ is not a subset of A^c even though $0 \in A^c$.

We shall see shortly that there is another way to show that $\{1/n : n \in \mathbb{Z}^+\}$ is not closed.

To emphasize the point of the examples, remember that although the complement of an open set is closed and the complement of a closed set is open, the terms are not negations of each other. There are sets that are both open and closed, and there are sets that are neither. Hence, we cannot show that a set is closed by proving that it is not open. Similarly, we cannot show that a set is open by showing that it is not closed. However, for the closed case the next theorems can help.

THEOREM 13.1.9

In any metric space:

- The intersection of a collection of closed sets is closed.
- The union of a finite collection of closed sets is closed.

PROOF

Let $\{C_i : i \in I\}$ be a family of closed sets. To prove that $\bigcap_{i\in I} C_i$ is closed, we must show that its complement is open. Since each C_i is closed, C_i^{c} is open. Therefore,

$$\left(\bigcap_{i\in I} C_i\right)^{\mathrm{c}} = \bigcup_{i\in I} C_i^{\mathrm{c}}$$

is open by Theorem 13.1.5.

Similarly, if I is a finite index set equal to $\{1, 2, \ldots, k\}$, then

$$C_1 \cup C_2 \cup \cdots \cup C_k$$

is closed because its complement

$$C_1^{\mathrm{c}} \cap C_2^{\mathrm{c}} \cap \cdots \cap C_k^{\mathrm{c}}$$

is open. ■

The largest open set within a set is its interior. Corresponding to this is the smallest closed set that contains a given set. This is our next definition.

■ DEFINITION 13.1.10

Let A be a subset of a metric space M. Define

$$\mathrm{cl}(A) = \bigcap\{F : F \text{ is closed in } M \text{ and } A \subseteq F\}$$

and call this set the **closure** of A.

In $\mathbb{R}$, $\mathrm{cl}((0, 1)) = [0, 1]$ and $\mathrm{cl}(\mathbb{Z}) = \mathbb{Z}$. In $\mathbb{R}^2$,

$$\mathrm{cl}((0, 1) \times (2, 4)) = [0, 1] \times [2, 4]$$

and

$$\mathrm{cl}(\{(x, y) \in \mathbb{R}^2 : xy < 1\}) = \{(x, y) \in \mathbb{R}^2 : xy \leqslant 1\}.$$

We note that the closure of a set is closed. To see this, examine the complement of $\mathrm{cl}(A)$:

$$\mathrm{cl}(A)^{\mathrm{c}} = \left(\bigcap\{F : F \text{ is closed and } A \subseteq F\}\right)^{\mathrm{c}} = \bigcup\{F^{\mathrm{c}} : F \text{ is closed and } A \subseteq F\}.$$

Since each F is closed, each F^{c} is open. Therefore, the union is open (Theorem 13.1.5). We should also note that a set A is closed in a metric space M if and only if $\mathrm{cl}(A) = A$. The proof of this is left as Exercise 15.

The way to picture a closed set is to think of it like a closed interval. A closed set contains all of the points of its interior plus all of the points that form a boundary around its interior. We therefore need a definition that will name all of these points.

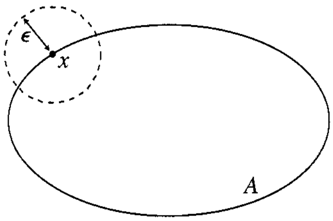

Figure 13.4 x is an accumulation point of A.

■ DEFINITION 13.1.11

Let A be a subset of a metric space M. An element $x \in M$ is an **accumulation point** of A if for all $\epsilon > 0$,

$$(D(x, \epsilon)\backslash\{x\}) \cap A \neq \varnothing.$$

Basically, an accumulation point of a set A has the property that there are points in A that are arbitrarily close to it (see Figure 13.4). We exclude the accumulation point from the criterion since it is always very close to itself!

Here is an example. Define $A = \{(x, y) \in \mathbb{R}^2 : 0 < x < 2\}$. The interior point $(1, 0)$ is an accumulation point. To prove that $(0, 0)$ is also an accumulation point, let $\epsilon > 0$ and take $a \in \mathbb{R}$ so that $0 < a < \min\{1, \epsilon\}$. Then,

$$(a, 0) \in [D((0, 0), \epsilon)\backslash\{(0, 0)\}] \cap A,$$

which means that the desired intersection is nonempty.

The last example shows that an accumulation point need not be an element of the set. In the metric space $\mathbb{R}$, we see that 0 is an accumulation point for both $(0, 1)$ and $[0, 1]$. For $\mathbb{R}^2$, $(1, 0)$ is an accumulation point for the open disk

$$\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}$$

and the line

$$\{(x, x - 1) : x \in \mathbb{R}\}.$$

On the other hand, a set need not have any accumulation points. Take the metric space $\mathbb{Z} \times \mathbb{Z}$ with the standard metric from $\mathbb{R}^2$. Then for all $a, b \in \mathbb{Z}$, (a, b) cannot be an accumulation point because

$$D_{\mathbb{Z}\times\mathbb{Z}}((a, b), 1/2)\backslash\{(a, b)\} = \varnothing.$$

We are now ready for our last test for closure.

■ THEOREM 13.1.12

A set is closed if and only if it contains its accumulation points.

PROOF

Let A be a subset of a metric space M. Assume that A is closed. Let x be an accumulation point of A. To obtain a contradiction, assume that $x \notin A$. This means that $x \in A^{\mathrm{C}}$, but this set is open. Hence, there is an $\epsilon > 0$ with the property that $D(x, \epsilon) \subseteq A^{\mathrm{C}}$. Therefore,

$$D(x, \epsilon) \cap A = \varnothing,$$

contradicting the fact that x is an accumulation point.

Now suppose that A contains all its accumulation points. We must show that A^{C} is open. Take $x \in A^{\mathrm{C}}$. By hypothesis, x cannot be an accumulation point of A. This means that there exists $\epsilon > 0$ such that $D(x, \epsilon) \cap A = \varnothing$. Therefore, $D(x, \epsilon)$ is a subset of A^{C}. ■

Exercises

Let (M, d) be a metric space.

1. Prove that $\mathbb{R}$, $\mathbb{R}^2$, and $\mathbb{R}^n$ when paired with the standard metrics are metric spaces.

2. Prove that for all $x, y \in M$ such that $x \neq y$, there exists $\epsilon > 0$ such that $D(x, \epsilon) \cap D(y, \epsilon) = \varnothing$.

3. Suppose that $\epsilon_1 \leqslant \epsilon_2$. Prove for any $x \in M$, $D(x, \epsilon_1) \subseteq D(x, \epsilon_2)$.

4. Define

$$d'(x, y) = \frac{d[x, y]}{1 + d[x, y]}.$$

Show that d' is a metric on M.

5. For all $x, y \in M$, let $d[x, y] = 0$ if and only if $x = y$, else $d[x, y] = 1$. Show that d is a metric on M.

6. Define the following metric on $\mathbb{R}^2$: For all $(x, y), (x', y') \in \mathbb{R}^2$,

$$d[(x, y), (x', y')] = |x - x'| + |y - y'|.$$

Let $N = \{(x, y) : x, y \in \mathbb{Z} \text{ and } x, y \geqslant 0\}$. Show that (N, d) is a metric space. (The function d is called the **taxicab** metric on N because N is viewed as the grid of a downtown area like Chicago. If a taxi is boarded at intersection $(1, 3)$ and ridden to intersection $(4, 1)$, the total number of blocks traveled is $|1 - 4| + |3 - 1| = 5$.)

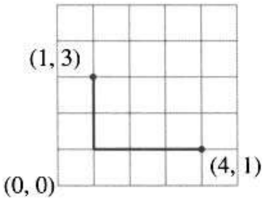

7. Sometimes a set will be open in one metric space but not open in another. Let A be the set of points that lie on the real line between 1 and 2, exclusive.
   (a) Explain why A is open in $\mathbb{R}$.
   (b) Describe A as a subset of $\mathbb{R}^2$.
   (c) Explain why A is not open in $\mathbb{R}^2$.

8. Prove that the following are open in the metric spaces indicated. Assume the usual metric for each space.
   (a) $\mathbb{R}\backslash\{0\}$ in $\mathbb{R}$
   (b) $\mathbb{R}^2\backslash\{(0, 0)\}$ in $\mathbb{R}^2$
   (c) $\mathbb{R}^2\backslash\{(x, y) \in \mathbb{R}^2 : y = 2x - 1\}$ in $\mathbb{R}^2$
   (d) $\{(x, y) \in \mathbb{R}^2 : x + y < 1\}$ in $\mathbb{R}^2$
   (e) $\{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 < 1\}$ in $\mathbb{R}^3$

9. Prove that $\{(x, y) \in \mathbb{R}^2 : a < x < b\}$ is open in $\mathbb{R}^2$.

10. Prove that every subset of M is open if and only if $\{a\}$ is open for all $a \in M$.

11. View $\mathbb{R}$ as a metric space with the standard metric. Verify Theorem 13.1.5 by directly proving the following propositions.
   (a) $\displaystyle\bigcup_{n=1}^{\infty} D(0, 1/n)$ is open.
   (b) $\displaystyle\bigcap_{n=1}^{k} D(0, 1/n)$ is open for all $k \in \mathbb{Z}^+$.
   (c) $\displaystyle\bigcap_{n=1}^{\infty} D(0, 1/n)$ is not open.

12. Describe the interior of each of the following subsets of $\mathbb{R}$: $\{0\}$, $(0, 2]$, $[0, 2]$, and $(0, 2) \cup \{-1, 3\}$.

13. Describe the interior of each of the given sets in $\mathbb{R}^2$:
   (a) $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}$
   (b) $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leqslant 1\}$
   (c) $\{(x, y) \in \mathbb{R}^2 : y \leqslant 1/x\}$

14. Let $A \subseteq M$. Show that $\text{int}(A)$ is the largest open subset of A by proving that if $U \subseteq A$ is open, then $U \subseteq \text{int}(A)$.

15. Let M be a metric space. Prove that $A \subseteq M$ is closed in M if and only if $\text{cl}(A) = A$.

16. Let $x \in M$. Prove $\{x\}$ is closed in M.

17. Show that the following are closed in the given metric spaces.
   (a) $\{0\}$ in $\mathbb{R}$
   (b) $\{(0, 0)\}$ in $\mathbb{R}^2$

(c) $\{(x, y) \in \mathbb{R}^2 : y = 2x - 1\}$ in $\mathbb{R}^2$
(d) $\{(x, y) \in \mathbb{R}^2 : 5 \leqslant x \leqslant 7\}$ in $\mathbb{R}^2$
(e) $\{(x, y) \in \mathbb{R}^2 : x + y \leqslant 1\}$ in $\mathbb{R}^2$
(f) $\{(x, y, 1) : x, y \in \mathbb{R}\}$ in $\mathbb{R}^3$

18. Let $A \subseteq M$. Show that $A \setminus \text{int}(A)$ is closed.

19. Take $a \in M$. Prove that $\{b \in M : d[a, b] \leqslant 1\}$ is closed using both the definition and Theorem 13.1.12.

20. Let $A \subseteq B \subseteq M$. Prove that if a is an accumulation point of A, it is an accumulation point of B.

21. Suppose that A and B are subsets of M. Prove the following about the closure:
   (a) $A \subseteq \text{cl}(A)$.
   (b) If $A \subseteq B$, then $\text{cl}(A) \subseteq \text{cl}(B)$.
   (c) $\text{cl}(\varnothing) = \varnothing$.
   (d) $\text{cl}(A) = \text{cl}(\text{cl}(A))$.
   (e) $\text{cl}(A)^{\text{C}} = \text{int}(A^{\text{C}})$.

22. Describe the set of accumulation points of each of the following as subsets of the given metric space:
   (a) $(0, 2)$ in $\mathbb{R}$
   (b) $(0, 2]$ in $\mathbb{R}$
   (c) $(0, 2) \cup \{-1, 3\}$ in $\mathbb{R}$
   (d) $\{(x, y) \in \mathbb{R}^2\ x^2 + y^2 < 1\}$ in $\mathbb{R}^2$

23. Let f be a continuous function on $[0, 1]$. Prove that $\{f \in C[0, 1] : \|f\| < 1\}$ is open in $C[0, 1]$ with the metric defined on page 446, and the **norm** $\|f\|$ defined by

$$\|f\| = \int_0^1 |f(x)|\, dx.$$

13.2 TOPOLOGICAL SPACES

Let us begin by recalling the definition of a continuous function from calculus. Suppose that $f : \mathbb{R} \to \mathbb{R}$ is continuous at $x = x_0$. This means that for all real numbers x and $\epsilon > 0$, there exists $\delta > 0$ such that

$$\text{if } |x - x_0| < \delta, \text{ then } |f(x) - f(x_0)| < \epsilon. \tag{13.2}$$

Let d be the standard metric on $\mathbb{R}$ and observe that we may write Sentence 13.2 as

$$\text{if } d[x, x_0] < \delta, \text{ then } d[f(x), f(x_0)] < \epsilon. \tag{13.3}$$

However, the x values that satisfy $d[x, x_0] < \delta$ are exactly the elements of $D(x_0, \delta)$. Therefore, we may rewrite Sentence 13.3 as

$$\text{if } x \in D(x_0, \delta), \text{ then } f(x) \in D(f(x_0), \epsilon). \tag{13.4}$$

These three sentences all describe continuity based on the standard metric on $\mathbb{R}$. To generalize the definition of continuity to other metric spaces, all that is needed is to drop the assumption that d is the standard metric. Sentences 13.3 and 13.4 are then definitions for continuity in any metric space. However, can we generalize further and define what it means for a function to be continuous without using a metric? Consider the following. Let $f : M \to N$ be a function where (M, d) and (N, d') are metric spaces. Assume that for all subsets V of N,

$$\text{if } V \text{ is an open in } N\text{, then } f^{-1}[V] \text{ is open in } M. \tag{13.5}$$

The set $f^{-1}[V] = \{x \in M : f(x) \in V\}$ is called the **inverse image** of V under f. We will now show that Sentence 13.5 is equivalent to Sentence 13.4.

- Assume Sentence 13.4 holds for all elements of M. Let V be open in N. To show that $f^{-1}[V]$ is open in M, take $x_0 \in f^{-1}[V]$. This means that $y_0 = f(x_0)$ is an element of V. Since V is open in N, there exists $\epsilon > 0$ such that $D_N(y_0, \epsilon) \subseteq V$. By hypothesis there exists $\delta > 0$ so that for all x,

  $$\text{if } x \in D_M(x_0, \delta)\text{, then } f(x) \in D_N(y_0, \epsilon).$$

  However, letting a be an element of $D_M(x_0, \delta)$, we find that $f(a)$ is an element of $D_N(y_0, \epsilon)$, which is a subset of V. Thus, $a \in f^{-1}[V]$, and this shows that we have $D_M(x_0, \delta) \subseteq f^{-1}[V]$.

- Now suppose that Sentence 13.5 is true. Let $x_0 \in M$ and $\epsilon > 0$. Write $y_0 = f(x_0)$. Since $D_N(y_0, \epsilon)$ is open in N, we know that $f^{-1}[D_N(y_0, \epsilon)]$ is open in M. Thus, we have $D_M(x_0, \delta) \subseteq f^{-1}[D_N(y_0, \epsilon)]$ for some $\delta > 0$. This is the δ that we want because if $x \in D_M(x_0, \delta)$, then $x \in f^{-1}[D_N(y_0, \epsilon)]$, and this means that $f(x) \in D_N(y_0, \epsilon)$.

Since the metric definition of continuity is equivalent to the definition given in Sentence 13.5, we may use the new sentence instead. Furthermore, notice that the new definition only requires an examination involving the open sets of the space. Therefore, a complete generalization of the concept of continuity will involve excluding the metric from the space in favor of an identification of the space's open sets. For this reason we make the following definition.

■ DEFINITION 13.2.1

Let X be a nonempty set. A collection $\mathcal{T}$ of subsets of X is called a **topology** on X if:

- $\varnothing$ and X are both elements of $\mathcal{T}$.
- the union of every collection of sets from $\mathcal{T}$ is an element of $\mathcal{T}$.
- the intersection of every finite collection of sets from $\mathcal{T}$ is again in $\mathcal{T}$.

The pair $(X, \mathcal{T})$ is called a **topological space**. The subsets found in $\mathcal{T}$ are defined to be the **open** sets of X, and the elements of X are called **points**.

The definition is chosen so that the metric space (M, d) can be viewed naturally as a topological space. Define $\mathcal{T}$ to be the open sets of M in the sense of Definition 13.1.3. Then by Theorem 13.1.5, $\mathcal{T}$ is a topology on M. A topology defined on a metric space M in this manner is called the **induced topology** on M.

Another example is the **discrete topology**. Let X be a set and let $\mathcal{T}$ be the set of subsets of X. Since the union and intersection of subsets of X is a subset of X, $\mathcal{T}$ is a topology on X. On the other extreme, if $\mathcal{T}' = \{\varnothing, X\}$, then $\mathcal{T}'$ is the smallest possible topology on X and is called the **trivial topology**. A more general example is the next theorem.

■ THEOREM 13.2.2

The intersection of two topologies on a set X is a topology on X.

PROOF

Let X be a set and $\mathcal{T}$ and $\mathcal{T}'$ be topologies on X. To show that $\mathcal{T} \cap \mathcal{T}'$ is a topology on X, we must check the three conditions.

- $X, \varnothing \in \mathcal{T} \cap \mathcal{T}'$ since $X, \varnothing \in \mathcal{T}$ and $X, \varnothing \in \mathcal{T}'$.
- Let $\mathcal{F} \subseteq \mathcal{T} \cap \mathcal{T}'$. Then, $\mathcal{F} \subseteq \mathcal{T}$ and $\mathcal{F} \subseteq \mathcal{T}'$. Therefore, $\bigcup \mathcal{F} \in \mathcal{T}$ and $\bigcup \mathcal{F} \in \mathcal{T}'$, and we conclude that
$$\bigcup \mathcal{F} \in \mathcal{T} \cap \mathcal{T}'.$$
- Let $\{U_1, U_2, \ldots, U_k\} \subseteq \mathcal{T} \cap \mathcal{T}'$. From this $U_i \subseteq \mathcal{T}$ and $U_i \subseteq \mathcal{T}'$ for all i. Therefore,
$$U_1 \cap U_2 \cap \cdots \cap U_k \in \mathcal{T},$$
and
$$U_1 \cap U_2 \cap \cdots \cap U_k \in \mathcal{T}'.$$
Hence,
$$U_1 \cap U_2 \cap \cdots \cap U_k \in \mathcal{T} \cap \mathcal{T}'. \blacksquare$$

Our generalization of continuity is complete with the next definition.

■ DEFINITION 13.2.3

Let X and Y be topological spaces. A function $f : X \to Y$ is called **continuous** when $f^{-1}[V]$ is open in X for all open subsets V of Y.

From our work above, any function between metric spaces that is continuous under the metric definition is continuous under the topological definition provided that the set has the induced topology.

For another example, let X be a set and let $f : X \to X$ be a function. View X as a topological space using the discrete topology. Let V be a subset of X. Under the discrete topology, since every subset of X is open, both V and $f^{-1}[V]$ are open. Therefore, in this case every function on X is continuous. This is also true when X

is given the trivial topology. Since only $\varnothing$ and X are the only open sets in X using this topology, $f^{-1}[X] = X$ and $f^{-1}[\varnothing] = \varnothing$ show that f is continuous.

We know from calculus that the composition of two continuous functions is continuous. The same result holds for the topological definition.

■ THEOREM 13.2.4

The composition of continuous functions is a continuous function.

PROOF

Let X, Y, and Z be topological spaces. Assume that $f : X \to Y$ and $g : Y \to Z$ are continuous functions. Let $V \subseteq Z$. We first observe that

$$(g \circ f)^{-1}[V] = f^{-1}[g^{-1}[V]].$$

To prove this, let $x \in (g \circ f)^{-1}[V]$. This means that there exists $z \in V$ such that $g(f(x)) = z$. Let $y = f(x)$ and write $z = g(y)$. Therefore, $y \in g^{-1}[V]$, and we conclude that $x \in f^{-1}[g^{-1}[V]]$. The other inclusion is left for Exercise 1.

To prove that the composition is continuous, let $V \subseteq Z$ be open. Since g is continuous, $g^{-1}[V]$ is open in Y, and since f is continuous, $f^{-1}[g^{-1}[V]]$ is open in X, so $(g \circ f)^{-1}[V]$ is open in X. Therefore, the composition is continuous. ■

We need one more test for continuity. We start by observing that the open disks of a metric space serve as the building block for all of the open sets of the metric space. To understand what this means in general, let X be a metric space and $\mathcal{B}$ the set of open disks in X. Let U be an open subset of X. For all $x \in X$, there exists $\epsilon_x > 0$ such that $D(x, \epsilon_x) \subseteq U$. Then

$$U = \bigcup_{x \in U} D(x, \epsilon_x). \tag{13.6}$$

Therefore, every open subset of X is the union of sets in $\mathcal{B}$. We now generalize this idea to topological sets.

■ DEFINITION 13.2.5

Let $(X, \mathcal{T})$ be a topological space. An **open base** for X is a set $\mathcal{B}$ with the property that every open set in $\mathcal{T}$ is the union of sets of $\mathcal{B}$. The sets of $\mathcal{B}$ are called **basic open sets**.

The reason that we need an open base is that it will make showing that a function is continuous more straightforward.

■ THEOREM 13.2.6

Let X and Y be topological spaces and $f : X \to Y$ a function. If $\mathcal{B}$ is an open base for Y, f is continuous if and only if the inverse image of each basic open set in $\mathcal{B}$ is open in X.

PROOF

Since sufficiency is clear, we shall let $\mathcal{B}$ be an open base for Y such that for all $B \in \mathcal{B}$, $f^{-1}[B]$ is open in X. To prove that f is continuous, take V to be an open subset of Y. Then there exists $\{B_i : i \in I\} \subseteq \mathcal{B}$ such that $V = \bigcup_{i\in I} B_i$. Thus,

$$f^{-1}[V] = f^{-1}\left[\bigcup_{i\in I} B_i\right] = \bigcup_{i\in I} f^{-1}[B_i].$$

Since the union of opens sets is open, $f^{-1}[V]$ is open. ■

Homeomorphisms

We should note that simply because $f : X \to Y$ is continuous and a one-to-one correspondence, it need not be the case that f^{-1} is continuous. If it is, by definition $f[U]$ is open for every open subset U of X. This is because $(f^{-1})^{-1} = f$. In this case we say that f **preserves** open sets and call f an **open mapping**.

■ **DEFINITION 13.2.7**

A **homeomorphism** is a continuous, one-to-one correspondence between two topological spaces that preserves open sets.

If there exists a homeomorphism between the topological spaces $(X, \mathcal{T})$ and $(Y, \mathcal{T}')$, we say that X and Y are **homeomorphic**. What this means is that the sets are the same size, or have the same **cardinality**, because there is a one-to-one correspondence between them, and there is a correspondence between the two spaces that associates the open sets of the one space with the open sets of the other. Essentially the two spaces look the same as topological spaces. For example, the function

$$f : (0, 1) \to \mathbb{R}$$

defined by $f(x) = \cot \pi x$ is a homeomorphism. We also have the following, whose proof we leave to Exercise 10.

■ **THEOREM 13.2.8**

The inverse of a homeomorphism is a homeomorphism. The composition of homeomorphisms is a homeomorphism.

We need an observation for our last definition. Let $(X, \mathcal{T})$ be a topological space and $Y \subseteq X$. Define

$$\mathcal{T}' = \{Y \cap U : U \in \mathcal{T}\}.$$

This set is a topology on Y. We will check the details:

- Since $Y = Y \cap X$ and $\varnothing = Y \cap \varnothing$, both Y and $\varnothing$ are elements of $\mathcal{T}'$.
- Let $\{U_i : i \in I\}$ be a collection of open sets from $\mathcal{T}$. Then because

$$\bigcup_{i\in I}(Y \cap U_i) = Y \cap \bigcup_{i\in I} U_i$$

and $\bigcup_{i\in I} U_i \in \mathcal{T}$, we have $\bigcup_{i\in I}(Y \cap U_i) \in \mathcal{T}'$.

- Now let $U_1, U_2, \ldots, U_k$ be sets in $\mathcal{T}$. Because
  $(Y \cap U_1) \cap (Y \cap U_2) \cap \cdots \cap (Y \cap U_k) = Y \cap (U_1 \cap U_2 \cap \cdots \cap U_k)$
  and $U_1 \cap U_2 \cap \cdots \cap U_k \in \mathcal{T}$, we may conclude that
  $$(Y \cap U_1) \cap (Y \cap U_2) \cap \cdots \cap (Y \cap U_k) \in \mathcal{T}'.$$

$\mathcal{T}'$ is called the **relative topology** on Y, and Y is said to be a **subspace** of X. Therefore, every subset of a topological space can itself be viewed as a topological space with the relative topology. Also, if $f : X \to Z$ is continuous, then $g : Y \to Z$ is also continuous when we define $g(x) = f(x)$ for all $x \in Y$. However, there is no reason to expect g to be a homeomorphism when f is.

We need the concept of a homeomorphism to make the last definition of the section.

■ DEFINITION 13.2.9

A topological space X is **Euclidean** if it is homeomorphic to a subspace of $\mathbb{R}^n$.

We may then define a topological space X to be **locally Euclidean** if for all $x \in X$, there exists a Euclidean open set containing x. A **manifold** is a locally Euclidean topological space. Examples of spaces that are manifolds include the conic sections and surfaces of spheres, tori, conoids, and spheroids.

Exercises

Unless otherwise noted, $(X, \mathcal{T})$ *and* $(Y, \mathcal{T}')$ *are topological spaces and* $\mathbb{R}^n$ *has the induced topology.*

1. Let X, Y, and Z be sets. If $f : X \to Y$ and $f : Y \to Z$ are functions, then for all subsets V of Z,
$$f^{-1}[g^{-1}[V]] \subseteq (g \circ f)^{-1}[V].$$

2. Let $\mathcal{T}$ and $\mathcal{T}'$ be topologies on X. Give an example which shows that $\mathcal{T} \cup \mathcal{T}'$ may not be a topology on X.

3. Take X to be $\{a, b, c\}$ and $\mathcal{T} = \{\varnothing, \{a\}, \{a, b\}, X\}$. Check that $(X, \mathcal{T})$ is a topological space and find all other topologies on X.

4. Let $Z \subseteq Y \subseteq X$. Let $\mathcal{S}$ be the topology of Z relative to X and $\mathcal{S}'$ the topology of Z relative to Y. Show that $\mathcal{S} = \mathcal{S}'$.

5. Confirm that $\mathcal{T} = \{U : \mathbb{R}\backslash U \text{ is finite}\} \cup \{\varnothing\}$ is a topology on $\mathbb{R}$.

6. Let $f, g : \mathbb{R} \to \mathbb{R}$ be two continuous functions. Prove that $f + g$ and fg are continuous using the metric definition of continuity. Can the topological definition be used for this problem?

7. Suppose that f, $g : \mathbb{R} \to \mathbb{R}$ are continuous functions. Prove using the metric definition of continuity that $f \circ g$ is continuous.

8. Prove Equation 13.6.

9. Let $\mathcal{B} \subseteq \mathbb{R}^2$ be the set of open squares, so

$$\mathcal{B} = \{(a, b) \times (c, d) : a < b, c < d\},$$

where × denotes the Cartesian product. Prove that $\mathcal{B}$ is an open base for $\mathbb{R}^2$.

10. Prove Theorem 13.2.8.

11. Find an example of a continuous one-to-one correspondence between topological spaces that is not a homeomorphism.

12. Let $\mathcal{B}$ be an open base for X. Prove that $f : X \to Y$ is an open mapping if and only if $f[U]$ is open for all $U \in \mathcal{B}$.

13. Suppose that $\mathcal{B} = \{U_i : i \in I\}$ and $\mathcal{B}' = \{U_i' : i \in I\}$ are open bases for X and Y, respectively. Prove that if f is a one-to-one correspondence such that $f[U_i] = U_i'$, then f is a homeomorphism.

14. Show that $\{(x, x^2) : x \in \mathbb{R}\}$ is homeomorphic to $\mathbb{R}$ by finding two bases and a homeomorphism as described in the previous problem.

15. Show that $(0, 1)$ is homeomorphic to both $(5, 6)$ and $\mathbb{R}$.

16. Let $\mathcal{C} = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 = 1\}$ be a subset of $\mathbb{R}^3$.
    (a) Give a geometric description for $\mathcal{C}$.
    (b) Show that $\mathcal{C} \setminus \{(1, 0, z) : z \in \mathbb{R}\}$ is homeomorphic to $(-1, 1) \times \mathbb{R}$.
    (c) Is $\mathcal{C}$ homeomorphic to $(-1, 1) \times \mathbb{R}$ or even $\mathbb{R}^2$? Explain.

17. Define a subset C of X to be **closed** in X if $X \setminus C$ is open. Prove:
    (a) X and $\varnothing$ are closed.
    (b) The intersection of closed sets is closed.
    (c) The union of finitely many closed sets is closed.

18. Let $f : X \to Y$ be a homeomorphism. Show:
    (a) If C is closed in X, then $f[C]$ is closed in Y.
    (b) If C is closed in Y, then $f^{-1}[C]$ is closed in X.

19. The **closure** of a subset A of X is the intersection of all closed subsets of X that contain A. This is the same definition as for metric spaces (Definition 13.1.10). Prove that A is closed if and only if $\text{cl}(A) = A$.

20. Let $f : X \to Y$ be a function. Prove that the following are equivalent:
    (a) f is continuous.
    (b) $f^{-1}[F]$ is closed in X for all F closed in Y.
    (c) $f[\text{cl}(A)] \subseteq \text{cl}(f[A])$ for each $A \subseteq X$.
    (d) $\text{cl}(f^{-1}[B]) \subseteq f^{-1}[\text{cl}(B)]$ for each $B \subseteq Y$.

13.3 STEREOGRAPHIC PROJECTION

Because of Exercise 16(b) of Section 13.2, we know that the surface of a cylinder is locally Euclidean. Therefore, it is a manifold. Like the cylinder, the surface of a sphere is also a locally Euclidean topological space, making it a manifold.

Take a sphere with center O and radius $\overline{ON}$. Construct $\overline{OU}$ and $\overline{OV}$ such that $\overline{ON}$, $\overline{OU}$, and $\overline{OV}$ are perpendicular to each other, setting up a coordinate system with O as the origin, $\overline{OU}$ as the x-axis, $\overline{OV}$ as the y-axis, and $\overline{ON}$ as the z-axis, as in Figure 13.5. With these axes the equation for the sphere is

$$x^2 + y^2 + z^2 = r^2, \tag{13.7}$$

where $r = ON$. Let $\mathcal{S}$ be the surface of the sphere without N and let $\mathcal{P}$ be the xy-plane. So we have

$$\mathcal{S} = \left\{ \left(x,\ y,\ \pm\sqrt{r^2 - x^2 - y^2} \right) : x \neq 0 \text{ or } y \neq 0 \right\} \cup \{(0,\ 0,\ -r)\}$$

and

$$\mathcal{P} = \{(x,\ y,\ 0) : x,\ y \in \mathbb{R}\}.$$

We wish to define a one-to-one correspondence between the points on the sphere (excluding N) and the points on the plane.

■ DEFINITION 13.3.1

A **stereographic projection** of a sphere $\mathcal{S}$ with **center of projection** N and **plane of projection** $\mathcal{P}$ is a function $\phi : \mathcal{S} \to \mathcal{P}$ that associates every point P on $\mathcal{S}$ with the point P' on $\mathcal{P}$ such that N, P, and P' are collinear.

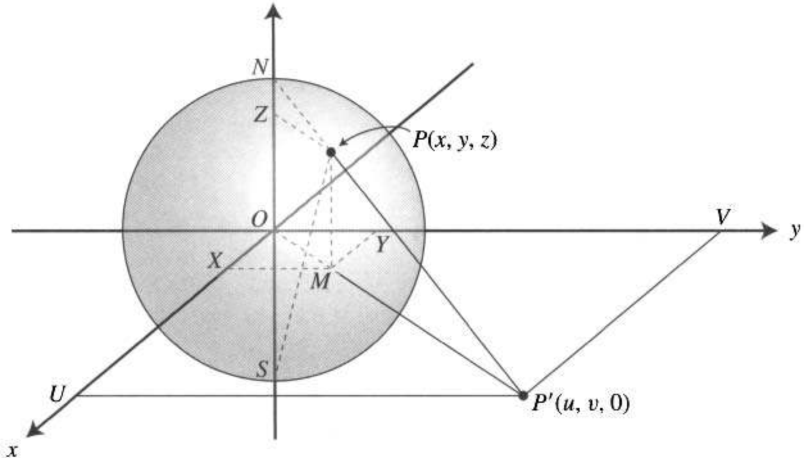

Figure 13.5 The stereographic projection maps P to P'.

Transformation Equations

We will follow Sommerville (1958, 172–174). Let N have coordinates $(0,\ 0,\ r)$ and S have coordinates $(0,\ 0,\ -r)$ using the labels of Figure 13.5. Let $\phi(P) = P'$, where P has coordinates $(x,\ y,\ z)$ and P' has $(u,\ v,\ 0)$. Because $\triangle OUP'$ is similar to $\triangle OXM$, we have

$$\frac{u}{x} = \frac{v}{y} = \frac{OP'}{OM},$$

and if we take Z on $\overline{NS}$ such that $\overline{ZP}$ is parallel to $\overline{OM}$,

$$\frac{OP'}{OM} = \frac{OP'}{ZP} = \frac{NP'}{NP} = \frac{NO}{NZ} = \frac{r}{r-z}$$

because $\triangle NOP' \sim \triangle NZP$ and $\overline{OM} \cong \overline{ZP}$. Observe that since $\sphericalangle NPS$ and $\sphericalangle NOP'$ are right angles, $\triangle NPS$ is similar to $\triangle NOP'$. Therefore,

$$NP \cdot NP' = NO \cdot NS = 2r^2.$$

Solving for NP' and substituting, we find that

$$\frac{NP'}{NP} = \frac{2r^2}{NP^2} = \frac{2r^2}{OM^2 + NZ^2} = \frac{2r^2}{x^2 + y^2 + (r-z)^2}.$$

Similarly, when we solve for NP we conclude that

$$\frac{NP'}{NP} = \frac{(NP')^2}{2r^2} = \frac{(OP')^2 + ON^2}{2r^2} = \frac{u^2 + v^2 + r^2}{2r^2}.$$

Therefore,

$$\frac{u}{x} = \frac{v}{y} = \frac{2r^2}{x^2 + y^2 + (r-z)^2} \tag{13.8}$$

and

$$\frac{u}{x} = \frac{v}{y} = \frac{r}{r-z} = \frac{u^2 + v^2 + r^2}{2r^2}, \tag{13.9}$$

These two equations can be used to give us parametric equations that will allow us to transform a point on the sphere to the projected point on the plane, and conversely. In particular, if $(x,\ y,\ z)$ is any point on the sphere other than N and $(u,\ v,\ 0)$ is its projection, then by solving for x, y, and z in Equation 13.9, we obtain

$$x = \frac{2r^2 u}{u^2 + v^2 + r^2}, \tag{13.10}$$

$$y = \frac{2r^2 v}{u^2 + v^2 + r^2}, \tag{13.11}$$

$$z = \frac{r(u^2 + v^2 + r^2) - 2r^3}{u^2 + v^2 + r^2}, \tag{13.12}$$

and by using Equation 13.8 we see that the stereographic projection ϕ is given by the equation

$$\phi(x, y, z) = \left(\frac{2r^2x}{x^2 + y^2 + (r - z)^2}, \frac{2r^2y}{x^2 + y^2 + (r - z)^2}, 0\right),$$

but since $x^2 + y^2 + (r - z)^2 = 2r^2 - 2rz$, we may simplify this formulation to

$$\phi(x, y, z) = \left(\frac{rx}{r - z}, \frac{ry}{r - z}, 0\right). \tag{13.13}$$

Images

We will find the image of circles on the surface of the sphere under the stereographic projection given by Equation 13.13. We want to do this for two reasons. First, the great circles of a sphere are considered to be the lines in the geometry found on its surface (page 415), so we would like to know their images. Second, we want to show that the sphere is locally Euclidean.

■ THEOREM 13.3.2

The stereographic projection maps a circle on the surface of the sphere onto

- a circle in the plane of projection if the circle does not contain the center of projection.
- a line in the plane of projection if the circle contains the center of projection.

PROOF

Circles on the sphere can be identified by slicing the surface with a plane. We will take the resulting circle, find its image in the plane of projection, and examine the equation that describes the image. Let

$$ax + by + cz = d$$

be a plane that intersects the sphere forming a circle and assume that (x, y, z) is on the surface of the sphere. Using the transformation Equations 13.10, 13.11, and 13.12, we conclude that d equals

$$a\left(\frac{2r^2u}{u^2 + v^2 + r^2}\right) + b\left(\frac{2r^2v}{u^2 + v^2 + r^2}\right) + c\left(\frac{r(u^2 + v^2 + r^2) - 2r^3}{u^2 + v^2 + r^2}\right).$$

Factoring out some common terms yields

$$(2r^2)(au + bv - cr) + cr(u^2 + v^2 + r^2) = d(u^2 + v^2 + r^2),$$

and from here we put the equation into the form of a general conic section with indeterminants u and v as

$$(cr - d)(u^2 + v^2) + (2r^2)(au + bv) - dr^2 - cr^3 = 0. \qquad (13.14)$$

If $cr - d \neq 0$, the coefficients on the two square terms are equal, and this conic is a circle. Its center and radius are found in Exercise 1. If $cr - d = 0$, then $(0, 0, r)$ is on the circle. In this case $(cr - d)(u^2 + v^2)$ drops out, and we are left with the equation for a line. ■

The function ϕ is a one-to-one correspondence because it has an inverse (Equations 13.10, 13.11, and 13.12). It is continuous because $f(x, z) = rx/(r - z)$ is continuous for $r \neq 0$. Since ϕ sends open disks to open disks as a result of Theorem 13.3.2, and since the set of open disks on the sphere serves as a base for its standard topology, ϕ is an open map (Theorem 13.2.6). We may then conclude the next two corollaries.

■ **COROLLARY 13.3.3**

The sphere without a point is homeomorphic to $\mathbb{R}^2$.

■ **COROLLARY 13.3.4**

The sphere is locally Euclidean.

Conformal Functions

A function that preserves angle measure is called **conformal**. We now show that the stereographic projection is a conformal map. What this means for the stereographic projection takes some explanation. Take, for instance, the planes Λ and K, and let them intersect the sphere in two great circles (Figure 13.6). To simplify our diagram, assume that the great circle that contains points A and B lies in the plane of projection K, so that it is projected onto itself. The great circle that contains C and D in Λ intersects the first great circle in B and E and is projected onto K, making the circle that contains C', D', F', and G'. The two projected circles meet in points B and E. We have two choices for the angle between the two great circles. We may choose

$$\theta = \mathrm{m}\sphericalangle ABC = \mathrm{m}\sphericalangle A_1B_1C_1$$

or its supplement ψ. Since both θ and ψ can be used to measure the the angle between planes Λ and K, they also measure the angle formed by the normal vectors of these two planes.

With this setup we can now examine the angles formed by the projected circles. Looking down from above N (Figure 13.7) we see two circles meeting in points B and E. Their centers are O and Q. We have two good options on how to measure the angle at B.

- Use $\sphericalangle OBQ$. This is equivalent to finding the angle between the vector normal to $\odot O$ at B and the vector normal to $\odot Q$ at B.

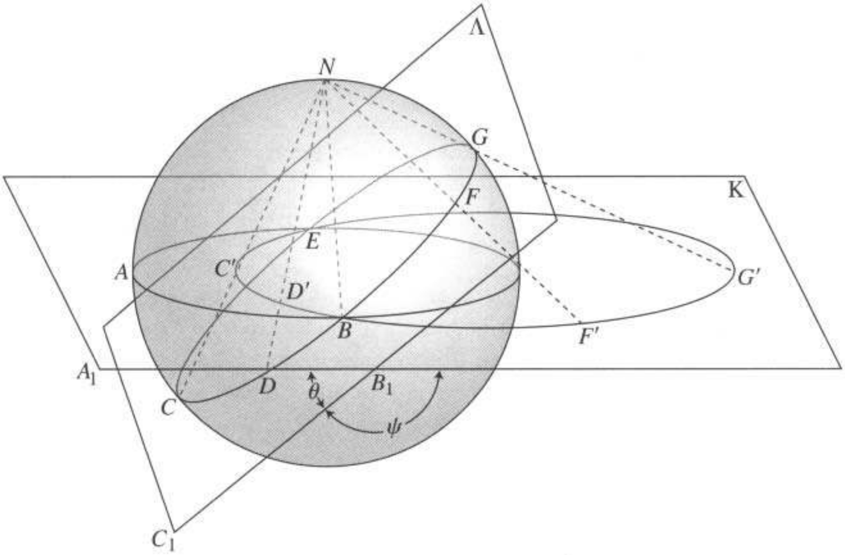

Figure 13.6 The stereographic projection of great circles onto the plane.

- Use $\measuredangle PBR$. This is the angle formed by the tangent to $\odot O$ at B and the tangent to $\odot Q$ at B. This angle is the supplement to $\measuredangle OBQ$ (Exercise 6).

The next theorem shows that $\theta = \text{m}\measuredangle OBQ$ and $\psi = \text{m}\measuredangle PBR$.

■ THEOREM 13.3.5

The stereographic projection maps angles between great circles on the surface of the sphere onto angles of equal measure in the plane of projection.

PROOF

To simplify our calculations, we will assume that the sphere has radius $r = 1$ and neither circle contains N. Take two planes that contain the center of the sphere in Figure 13.6. Assuming that the center also serves as the origin of a coordinate system, let the equations of the planes be

$$ax + by + cz = 0$$

and

$$dx + ey + fz = 0$$

and suppose that

$$a^2 + b^2 + c^2 = 1 \text{ and } d^2 + e^2 + f^2 = 1. \tag{13.15}$$

By Equations 13.15 these planes cut two great circles that form an angle of measure

$$\theta = \cos^{-1}(ad + be + cf).$$

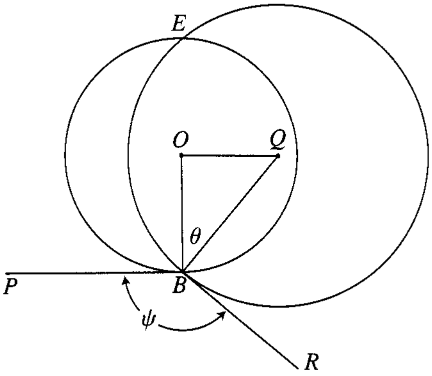

Figure 13.7 The top view of the images in the plane of two great circles.

By Equation 13.14 the projection of the two great circles onto the plane are the circles $\odot O$ and $\odot Q$ with

$$cu^2 + cv^2 + 2au + 2bv - c = 0 \tag{13.16}$$

and

$$fu^2 + fv^2 + 2du + 2ev - f = 0 \tag{13.17}$$

as their equations. Again due to Equations 13.15, we can complete the square on both of these last two equations to find

$$\left(u + \frac{a}{c}\right)^2 + \left(v + \frac{b}{c}\right)^2 = \frac{1}{c^2}$$

and

$$\left(u + \frac{d}{f}\right)^2 + \left(v + \frac{e}{f}\right)^2 = \frac{1}{f^2}.$$

Therefore, $OB = 1/c$ and $QB = 1/f$. We will obtain $\mathrm{m}\measuredangle OBQ$ by using the law of cosines. We first need to compute the distance between the two centers of the circles. It is

$$OQ = \sqrt{\left(\frac{a}{c} - \frac{d}{f}\right)^2 + \left(\frac{b}{c} - \frac{e}{f}\right)^2}.$$

By the law of cosines,

$$\left(\frac{a}{c} - \frac{d}{f}\right)^2 + \left(\frac{b}{c} - \frac{e}{f}\right)^2 = \frac{1}{c^2} + \frac{1}{f^2} - \frac{2\cos B}{cf}.$$

Square the binomials to find that

$$\frac{a^2}{c^2} - \frac{2ad}{cf} + \frac{d^2}{f^2} + \frac{b^2}{c^2} - \frac{2be}{cf} + \frac{e^2}{f^2} = \frac{1}{c^2} + \frac{1}{f^2} - \frac{2\cos B}{cf}.$$

When we gather terms on the left we see that

$$-\frac{2ad}{cf} - \frac{2be}{cf} + \frac{a^2 + b^2 - 1}{c^2} + \frac{d^2 + e^2 - 1}{f^2} = -\frac{2\cos B}{cf}.$$

Because of Equations 13.15, we obtain

$$-\frac{2ad}{cf} - \frac{2be}{cf} - 2 = \frac{-2\cos B}{cf},$$

which finally yields

$$ad + be + cf = \cos B.$$

Therefore, $m\sphericalangle OBQ = \theta$, and since $\sphericalangle OBQ$ and $\sphericalangle PBR$ are supplements, $m\sphericalangle PBR = \psi$. ■

Exercises

1. Find the center and radius of the circle given in Equation 13.14.

2. Suppose that the sphere in Figure 13.5 has unit radius. Find equations for the images of the given circles under ϕ.
   (a) $\{(x, 0, z) : x^2 + z^2 = 1\}$
   (b) $\{(x, y, 0) : x^2 + y^2 = 1\}$
   (c) $\{(x, y, 1/2) : x^2 + y^2 = 3/4\}$
   (d) $\{(x, y, -1/2) : x^2 + y^2 = 3/4\}$
   (e) The circle through $(2/3, 2/3, 1/3)$, $(-2/3, 2/3, 1/3)$, and $(0, \sqrt{3}/2, 1/2)$
   (f) The intersection of the plane $x + y + z = 0$ and the sphere

3. The points $A(2/3, 2/3, 1/3)$, $B(-2/3, 2/3, 1/3)$, and $C(0, \sqrt{3}/2, 1/2)$ also describe an angle on the surface of the unit sphere centered at the origin. Let A be the vertex of this angle.
   (a) Find the equations for the planes whose intersections with the unit sphere are the great circles that contain $\overset{\frown}{AB}$ and $\overset{\frown}{AC}$.
   (b) Find the lengths of $\overset{\frown}{AB}$ and $\overset{\frown}{AC}$.
   (c) Find the measure of spherical $\sphericalangle BAC$.
   (d) Find measure of the angle in $\mathbb{R}^2$ formed by $\phi[\overset{\frown}{AB}]$ and $\phi[\overset{\frown}{AC}]$.

4. What on the surface of the sphere projects onto the y-axis under the stereographic projection? What projects onto the circle $x^2 + y^2 = 4$?

5. What is the image of the upper hemisphere under ϕ, and what is the image of the front hemisphere?

6. Show that $\sphericalangle OBQ$ and $\sphericalangle PBR$ in Figure 13.7 are supplementary.

7. Find the angle formed by the circles $x^2 + y^2 - 2y = 0$ and $x^2 + y^2 - 2y - 5 = 0$.

8. Define ϕ' to be the stereographic projection of a sphere onto a plane that is tangent to the sphere at its south pole. Find parametric equations that describe ϕ'.

9. Reprove Theorem 13.3.5 assuming that at least one of the great circles contains N.

13.4 CONSISTENCY OF NON-EUCLIDEAN GEOMETRY

During the last five years of Gauss's life, his health deteriorated due to severe heart problems, yet there were periods when he fought his illness successfully and showed increased stamina. One such time was upon publication of Riemann's *Habilitationsschrift*. Although he expressed excitement about the results, his failing memory and trouble concentrating left him unable and unwilling for deep mathematical conversation. When able, he did involve himself with some simple problems, but he seems to have preferred reading newspapers and other popular literature. During this time, Gauss expressed a belief in God. He was not a Christian but would be better described as a deist who believed in an afterlife. He thought that when one died, one would perceive more clearly the geometric nature of the universe. Gauss evidently hoped that it would be non-Euclidean, since he liked the concept of an absolute measure of length. On February 23, 1855, Gauss died in his home. It is said that his watch stopped moments after his death (Bühler 1981, 152–155).

Two years later, in 1857, Riemann joined the regular faculty at Göttingen, and after another two years he was promoted to the chair of mathematics that had been held by Gauss. Unfortunately, Riemann had contacted tuberculosis. His doctor recommended a warmer climate, so he moved to Italy. He died there near Lake Maggiore on July 20, 1866 at the age of 39. Dying at such a young age, Riemann did not live to see what just a few years later would follow from his work (Gray 2007, 189; Bühler 1981, 152–155; Burton 1985, 564).

Beltrami Models

Lobachevski wrote his geometry using algebra, trigonometry, and calculus, so he believed that his new system was without contradiction. However, he did not provide a proof. The first to put the new geometry on a sound logical footing was Eugenio Beltrami. He was born in Cremona, Italy on November 16, 1835. Many of his relatives were talented in various forms of art, including the engraving of cameos and the painting of miniatures. Beltrami himself enjoyed music and would eventually study its relationship to mathematics. Beltrami was a student from 1853 to 1856 at the University of Pavia, but financial problems would take him away from his academic life to work as a secretary for the railroad in Verona. His new occupation would later take him to Milan, where in his free time he resumed his mathematical pursuits. His early interests were greatly influenced by both Gauss and Riemann and included differential geometry. Beltrami gained a solid reputation for mathematics and would eventually hold various positions in algebra, geometry, and mechanics at universities throughout Italy, including Bologna, Pisa, and Rome. No biography of Beltrami is complete, even one as short as this, without mentioning that it was he who was responsible for rediscovering Saccheri's *Euclides Vindicatus* (Struik 1970a, 599-600; Gray 2006, 79; Stillwell 1996, 3).

Building on Riemann's work, Beltrami was the first to prove that the Parallel Postulate was not provable. He did this by allowing the terms *point* and *line* to be reinterpreted, such as when great circles are considered to be the lines on a sphere. Any example that satisfies the postulates under the new interpretation will be called a **model** of the postulates. If we find such a model, we conclude that the postulates are consistent (page 324). Finding a model yields consistency because we believe that neither mathematics nor the physical world are self-contradictory, so any set of statements interpreted within either of these contexts could not lead to a contradiction.

Beltrami's first consistency proof is found in his "Essay on the Interpretation of Noneuclidean Geometry" (*Saggio di Interpretazione della Geometria Non-euclidea*). In this paper, Beltrami modeled the new geometry using geodesics on a surface of constant negative curvature. The notion of curvature was developed by Gauss in his *Disquisitiones Circa Superficies Curvas* of 1827. It is an intrinsic property of a surface and is measured with real numbers. Spheres have constant positive curvature, while planes and cylinders are considered flat, having zero curvature. The advantage of having any kind of constant curvature is that shapes can be moved on the surface without deforming. An example of a surface with constant negative curvature was studied by Gauss's student Ferdinand Minding in 1840. All of the trigonometric formulas of Lobachevski and Bolyai held true on this surface, so it was a logical candidate for interpreting non-Euclidean geometry (Stillwell 1996, 1–2; Boyer and Merzbach 1991, 505–506).

Minding's surface is brought about by revolving a particular curve about the y-axis. Attach a rope to a relatively heavy object sitting on the ground. While holding the rope, walk away from it until the rope is taught. Now start to pull the object by turning right and walking along the straight line that is perpendicular to the rope. As you do this, the object will be dragged along a path known as a **tractrix** (Figure 13.8). This curve was originally described by Leibnitz. Some have used the German word *hundkurve* to describe it because we can imagine the curve representing the path that a dog would take as it goes on a walk with a leash (Steinhaus 1999, 250–251).

To derive the equation for the tractrix, place the object along the x-axis at $x = a$. Let O be the origin and P an arbitrary point along the curve. Let T represent the point of intersection between the y-axis and the tangent to the tractrix at P. As the object is pulled (or the dog led), the rope is represented by that tangent, so $PT = a$. If the coordinates of P are (x, y) and $x \leqslant a$, then

$$OT = y + \sqrt{a^2 - x^2}$$

and

$$\frac{dy}{dx} = -\frac{\sqrt{a^2 - x^2}}{x}.$$

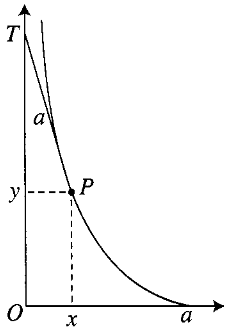

Figure 13.8 The tractrix.

To solve this differential equation we write

$$\int dy = -\int \frac{\sqrt{a^2 - x^2}}{x}\, dx.$$

Focusing our work on the right-hand side of the equation, we substitute $x = a \sin\theta$:

$$\int \frac{\sqrt{a^2 - x^2}}{x}\, dx = a \int \frac{\cos^2\theta}{\sin\theta}\, d\theta = a \int (\csc\theta - \sin\theta)\, d\theta.$$

Because

$$\int (\csc\theta - \sin\theta)\, d\theta = -\ln\left|\frac{a}{x} + \frac{\sqrt{a^2 - x^2}}{x}\right| + \frac{\sqrt{a^2 - x^2}}{a} + C$$

and $y = 0$ when $x = a$, we conclude that

$$y = a \ln\left(\frac{a + \sqrt{a^2 - x^2}}{x}\right) - \sqrt{a^2 - x^2}. \tag{13.18}$$

The original derivation of the equation for the tractrix is due to James Bernoulli (Kline 1972, 473–474).

Now take the tractrix and rotate it about its asymptote. The result is a solid called a **pseudosphere** (Figure 13.9). This surface is the simplest that has constant negative curvature. However, since the pseudosphere has a single point at infinity, it cannot represent the entire Lobachevskian plane. Instead, it models a portion that is bounded by two parallel lines and an arc of a horocycle. The surface that Beltrami used to show the consistency of **HA** was essentially a covering that wrapped around the pseudosphere infinitely many times. With minor modification this can be made to represent an arbitrarily large portion of the non-Euclidean plane (Stillwell 1996, 2; Sommerville 1958, 169–170).

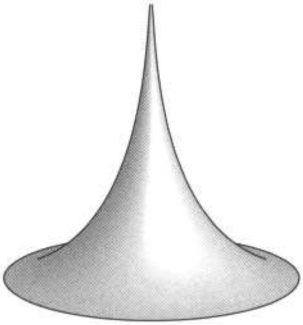

Figure 13.9 The pseudosphere.

In another of his papers, entitled "Fundamental Theory of Spaces of Constant Curvature" (*Teoria fondamentale degli spazii di curvatura constante*), Beltrami generalized his results to n dimensions. He also gave early versions of models that are now named after Henri Poincaré and Felix Klein and one that we will call Beltrami's **hemisphere model** (Figure 13.10). Take the hemisphere in $\mathbb{R}^3$ given by the equation

$$x^2 + y^2 + z^2 = r^2$$

and let $z > 0$. The surface of the hemisphere is interpreted to be the plane, *points* are regular points on the surface, and *lines* are the geodesics formed by slicing the model with vertical planes. The great circle $z = 0$ represents the line ω at infinity. To find the parallel lines of this model, let l be an arc formed by cutting the hemisphere with perpendicular plane Λ and let P be a point on the surface but not on l. Turn Λ until it intersects P by using the perpendicular to one of the endpoints of l as an axis. The resulting arc contains P and intersects l at a point on ω. This arc is considered parallel to l. If Λ had been turned using the other endpoint of l, the result would be the other parallel to l through P (Stillwell 1996, 35–37).

Both of Beltrami's models show that the geometries of Lobachevski and Bolyai are consistent by transferring the question to one within Euclidean geometry. If

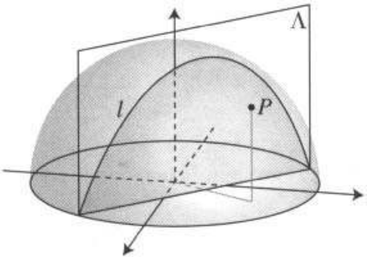

Figure 13.10 The hemisphere model.

there is an interpretation of non-Euclidean geometry that satisfies the negation of the Parallel Postulate, then non-Euclidean geometry is without contradiction if the system in which it is being interpreted has no contradictions. Since mathematicians were confident in the consistency of Euclid's system, they should now have confidence in the new geometry. This concept is known as **relative consistency**. All consistency proofs are actually proofs of relative consistency.

Poincaré Models

There are two models associated with Poincaré that can be obtained from Beltrami's hemisphere model. Obtain the first by stereographically projecting the hemisphere onto the plane $z = 0$ by completing the bottom half of the sphere and using the south pole as the center of projection (Definition 13.3.1). Let $\mathscr{C}$ be the circle with equation

$$x^2 + y^2 = r^2.$$

The images of points from the hemisphere lie within $\mathscr{C}$ and are the *points* of the model. The images of the arcs of the hemisphere will be projected onto either segments within $\mathscr{C}$ or half-circles within $\mathscr{C}$ that intersect the circle at right angles (Theorem 13.3.2). These are the *lines* of the model. Two lines are said to be parallel in this model if they intersect at $\mathscr{C}$. This means that $\mathscr{C}$ is interpreted as the line at infinity. The image of the hemisphere within the circle $\mathscr{C}$ is called the **Poincaré disk model** (Figure 13.11). Poincaré published this model in 1882 (Stillwell 1996, 35–36).

The second model is known as the **Poincaré half-plane model**. Take a plane Λ that is tangent to the Beltrami hemisphere at Q but perpendicular to the plane $z = 0$ (Figure 13.12). Let P be on the hemisphere opposite Q so that $\overline{PQ}$ is perpendicular to Λ. The image on the upper half of Λ of the hemisphere under the stereographic projection with center P is Poincaré's model. Points and lines are interpreted as in the Poincaré disk, and the intersection of Λ and $z = 0$ is the line at infinity (Stillwell 1996, 36–37).

The mathematician responsible for these two models earned a reputation that approached that of Gauss. Poincaré was born in Nancy, France on April 29, 1854.

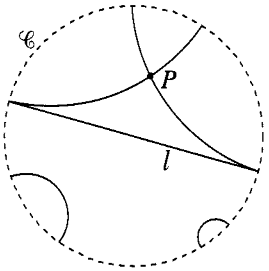

Figure 13.11 The Poincaré disk.

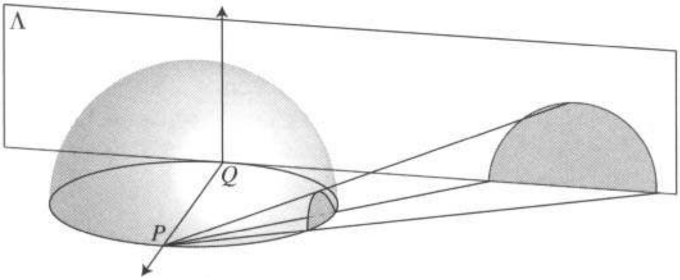

Figure 13.12 Stereographic projection of the hemisphere onto the Poincaré half-plane.

He was from an upper-middle-class family, and his relatives included physicians, professors, and politicians. He attended the École Polytechnique, and after graduating Poincaré worked at engineering while he studied for his doctorate degree in mathematics. He finished in 1879. Afterward he taught at the University of Caen and became a professor at the University of Paris in 1881. At the young age of 33, he was elected to the Académie des Sciences. He had wide mathematical interests that included function theory, number theory, algebra, and the foundations of mathematics. In addition to mathematics, Poincaré made important contributions to astronomy and physics (Dieudonné 1970, 51–61).

Klein Models

Felix Klein was another leading nineteenth century mathematician. He was born in Düsseldorf on April 25, 1849. His father worked in the government finance office and had very traditional Prussian views. This resulted in Klein attending the gymnasium in his hometown where he studied classical languages but very little mathematics. After graduation he attended the University of Bonn in 1866, where he met Julius Plücker. Klein was soon recruited by Plücker and became his assistant with the goal being to make Klein into a physicist. Klein had the talent for it, and it might have happened except that Plücker died in 1868. At the time Plücker had been working on various papers, including one on geometry. As his assistant, Klein prepared the unfinished works for publication and was particularly inspired by Plücker's *Neue Geometrie des Raumes*. By the end of the year, Klein finished his doctorate degree, moved to Göttingen for a few months, and then moved on to Berlin. This was followed by a trip to Paris that was interrupted by the Franco-Prussian War, and then a return to Göttingen, where he taught mathematics as a lecturer in 1871 (Yaglom 1988, 25–27; Burau and Schoenberg 1970, 396).

It was during this year that Klein's "On the So-called Noneuclidean Geometry" was published. It included a model of the geometry of Lobachevski and Bolyai. If Beltrami's hemisphere model is orthogonally projected onto the plane, the result is an open disk with the geodesics of the hemisphere mapped to chords of the circle. Viewing the plane as a copy of $\mathbb{R}^2$, the equation for the disk is $x^2 + y^2 < r^2$, where r

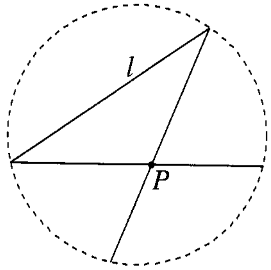

Figure 13.13 The Klein disk.

is the radius of the hemisphere. This is called the **Klein disk model** (Figure 13.13). The *points* of the model are regular points, and the *lines* are chords of the circle without the endpoints. Two lines are parallel if they meet at a point on the circle, and they are nonintersecting if the do not meet anywhere within or on the circle.

Klein also provided models that showed that the Hypothesis of the Obtuse Angle was possibly true. His motivation for this can be traced to Riemann. From Saccheri to Lobachevski and Bolyai, geometers did not accept geometries in which **HO** held. In such systems, Theorem 5.3.9 was false. Since this proposition was part of their data from Euclid that was not based on the Parallel Postulate, they consistently rejected **HO** as a possibility. This is surprising since it was a known fact that the angle sum of a triangle on the surface of a sphere had angle sum greater than two right angles. It was left to Riemann to accept that this surface yielded its own geometric system (Bonola 1912, 141).

Riemann observed that there are essentially three types of two-dimensional geometries, based on the type of surface on which it lies:

- A flat surface such as a plane or cylinder
- A surface of constant negative curvature such as a pseudosphere
- A surface of constant positive curvature

The first of these is Euclidean geometry and the second is Lobachevski's. The third can be interpreted as existing on the surface of a sphere, but to do this we must rethink our understanding of lines. According to Riemann:

> In the extension of space-construction to the infinitely great, we must distinguish between *unboundedness* and *infinite extent*; the former belongs to the extent relations, the latter to the measure-relations. That space is an unbounded threefold manifoldness, is an assumption which is developed by every conception of the outer world; according to which every instant the region of real perception is completed and the possible positions of a sought object are constructed, and which by these applications is forever confirming itself. The unboundedness of space possesses in this way a greater empirical certainty than any external experience. But its infinite extent by no means follows from this; on the other hand if we assume independence of bodies from position, and therefore ascribe

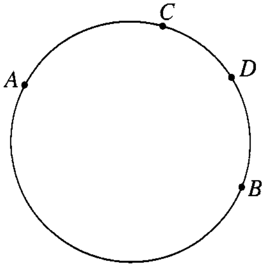

Figure 13.14 A finite, closed line.

> to space constant curvature, it must necessarily be finite provided this curvature has ever so small a positive value. If we prolong all the geodesics starting in a given surface-element, we should obtain an unbounded surface of constant curvature, *i.e.*, a surface which in a *flat* manifoldness of three dimensions would take the form of a sphere, and consequently be finite. (Riemann 1854, III.2, 67–68)

The result of Riemann's comments is two definitions that had previously not been applied to lines. We say that a line is **finite** if its length is finite. A line l is **closed** if given any two points A and B on l, there exist points C and D on l such that C and D are between A and B but B is between C and D (Figure 13.14).

This change requires a change in the postulates of Euclid. Since the days of the Greeks, geometers nearly universally assumed that lines were infinite. Even many of Euclid's proofs assumed this, despite this property not being found among the postulates. Therefore, this new geometry requires that Euclid's first and last postulates be replaced with the following two.

■ POSTULATES 13.4.1 [Riemann]

1. All lines are closed and finite.
2. All lines intersect.

In the development of this **Riemannian geometry** we may keep certain of Euclid's theorems up through *Elements* I.28 (Theorem 5.4.1), but others will have to be jettisoned. This includes the previously mentioned Theorem 5.3.9. We also cannot keep Theorem 5.3.12. The proofs of both of these propositions relied on lines being infinite, as indicated in Exercise 8.

Since the next theorem implies the existence of a triangle that has an angle sum greater than two right angles, we conclude by Theorem 10.3.1 that **HO** holds in this geometry.

■ THEOREM 13.4.2 [Riemann]

All lines perpendicular to a given line are concurrent.

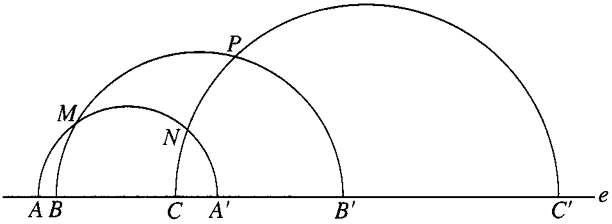

PROOF

Let e be a line and A a point on e. Let A' be chosen so that $\overline{AA'}$ is perpendicular to e. Since $\overline{AA'}$ is closed, it must intersect e again. The second time $\overline{AA'}$ intersects e is either back at A by looping back or by intersecting in some new point. Suppose it is the latter and that this new point is A'. Draw two more lines perpendicular to e at B and C and let the respective lines intersect e for a second time at B' and C'. By Postulate 13.4.1.2, the three pairs of lines intersect. Let M be the intersection of $\overline{AA'}$ and $\overline{BB'}$, N the intersection of $\overline{AA'}$ and $\overline{CC'}$, and P the intersection of $\overline{BB'}$ and $\overline{CC'}$. Since $\sphericalangle MAB$ and $\sphericalangle MBA$ are both right, $\triangle ABM$ is an isosceles triangle. Hence, $\overline{AM} \cong \overline{BM}$. Similarly, $\triangle BA'M$ is isosceles, so we conclude that $\overline{BM} \cong \overline{MA'}$. Thus, M is the midpoint of $\overline{AA'}$. A similar argument will show that N is also the midpoint of $\overline{AA'}$, from which we conclude that $M = N$. Since M and P are midpoints of $\overline{BB'}$, we see that they are in fact the same point. Hence, the three lines are concurrent. ■

The point N that exists on one side of the line e in the proof of Theorem 13.4.2 on which all perpendiculars to the line fall is called the **absolute pole** of e. The line e is the **absolute polar** of N. For any point P on e, $\overline{NP}$ is called a **quadrant**, and we say that N is **orthogonal** to P. All quadrants on the same side of e are congruent (Exercise 14).

Certainly on the other side of e we may repeat the construction above and find a point S that is contained by all perpendiculars. The question now is whether the points N and S are different. It turns out that both options are consistent with the postulates for this geometry. If N and S are different, the points are called **antipodal** and the geometry is called **double elliptic**. If this is the case, lines meet in two points, so Euclid's first postulate cannot hold. If, instead, the points are the same, we regain Euclid's assumption that two points determine a line. This geometry is called **single elliptic**. This terminology is due to Klein but from a different context. It has now become common to use the terms **elliptic** for geometries that satisfy **HO** and **hyperbolic** for the geometry of Lobachevski and Bolyai (Bonola 1912, 164; Wussing 1984, 177).

The two elliptic geometries were proven consistent by models given by Klein (Burton 1985, 568). He realized them on the surface of a sphere. Double elliptic is illustrated in Figure 13.15(a). *Lines* in the model are great circles of the sphere, and *points* are regular points on its surface. There is no line at infinity because there are no parallels. This model satisfies Postulates 13.4.1 and the three remaining postulates of Euclid.

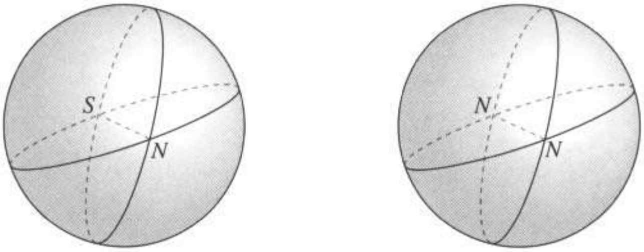

(a) Double elliptic (b) Single elliptic

Figure 13.15 Klein's models of elliptic geometry.

Single elliptic geometry is modeled in Figure 13.15(b). Here, *lines* are as in double elliptic; however, *point* is understood to refer to pairs of antipodal points on the surface of the sphere. Postulates 13.4.1 hold true in this model as well as all of Euclid's postulates except for the last postulate concerning parallels.

Exercises

1. Locate the point at infinity for the pseudosphere. Use this point to identify which objects on the surface of the pseudosphere should represent lines, parallel lines, and nonintersecting lines.

2. Assume $a = 1$ in the equation for the tractrix (Equation 13.18). This makes the equation equivalent to $y = \operatorname{sech}^{-1}x - \sqrt{1 - x^2}$.
   - (a) Confirm this result.
   - (b) Show that the surface area of the pseudosphere (the surface area of revolution of the tractrix) can be approximated by 12.56. What is the significance of this result?
   - (c) Show that the volume of the pseudosphere can be approximated by 2.09. What is the significance of this result?

3. A **pseudosquare** is a quadrilateral that is equilateral and equiangular. How can a pseudosquare be constructed in hyperbolic geometry?

4. Draw a picture representing parallel lines in Beltrami's hemisphere model.

5. Create a diagram for each of the Poincaré models and the Klein model that includes examples of points, lines, parallel lines, nonintersecting lines, and triangles.

6. The Poincaré models and the Klein model are intended to represent the non-Euclidean plane of Lobachevski and Bolyai. If this is the case, what is missing from the descriptions of these models in the section?

7. Draw horocycles in the Poincaré and Kline models.

8. Show that Theorems 5.3.9 and 5.3.12 are false under **HO**.

9. Let the open square $(0, 1) \times (0, 1)$ be the plane, *points* be regular points, and *lines* be line segments with endpoints on the border of the square. Is this geometry Euclidean? What is the angle sum of triangles? How do you harmonize the answers to these questions?

10. Prove that rectangles do not exist in elliptic geometry.

11. Explain why in single elliptic geometry a line does not separate the plane into two half planes. With this in mind, give an example of a "flat" model of double elliptic geometry.

12. Do lines divide the plane into two half planes in double elliptic geometry? Explain.

13. Does Pasch's Postulate (11.3.2) hold in either of the elliptic geometries?

14. Prove that all quadrants on the same side of e are congruent in Riemann's elliptic geometry.

15. What is the maximum measure of the third angle of a triangle with two right angles?

CHAPTER 14

JEAN-VICTOR PONCELET

Gaspard Monge was a significant figure in late eighteenth century French mathematics. In 1775, at the age of 29, he was appointed the Royal Professor of Mathematics and Physics at the École Militaire de Méziers. There he developed a method of designing forts that effectively defended the soldiers within from gunfire that originated from multiple directions. This led to Monge's creation of a field of mathematics known as **descriptive geometry**. Monge left Méziers in 1784 to become the head of naval education in France, and 10 years later he was appointed to another influential education position, one that led to the creation of École Polytechnique in June 1795. This was a two-year school designed to prepare students for the study of engineering. Its curriculum, which included both descriptive and differential geometry, was designed by mathematicians who believed that a study of pure mathematics was the best way to prepare their students for the various options they would have in their future studies. Besides his duties at École Polytechnique, Monge remained professionally active outside of the school. This included working to revive the French professional organizations, which had been neglected after the revolution, and accompanying Napoleon to Egypt, where he led the scientific team. When Monge returned a year later, he continued teaching at École Polytechnique until 1809. During his tenure there, Monge was a champion of synthetic geometry, and he conveyed this passion to

Revolutions of Geometry. By Michael O'Leary
Copyright © 2010 John Wiley & Sons, Inc.

one of his students: Jean-Victor Poncelet (Gray 2007, 4–6, 7; Gray 2004, 30; Boyer and Merzbach 1991, 468, 478).

The son of wealthy and politically influential parents, Poncelet was born on July 1, 1788 in the French town of Metz, a city that remained dear to him his entire life. At a very early age, Poncelet was sent away for his preliminary schooling and then returned to his family in 1804. Three years later he enrolled at École Polytechnique but fell behind due to poor health and took an extra year to complete his studies. After completing his work at École Polytechnique, Poncelet joined the corps of engineers in September 1810. A year and a half later he was assigned to the Dutch island of Walcheren, and in June 1812 he took part in the French invasion of Russia. Despite his role as an engineer and not as a fighting soldier, Poncelet was taken prisoner on November 18, 1812 during the retreat. He was held as a prisoner of war at Saratov and not released until June 1814. While in Russia, Poncelet spent his time writing. The result was a text on analytic geometry based on his studies at École Polytechnique. He would eventually have it published but not for 50 years. It was intended to be an introduction to his second work, the *Traité des propriétés projectives des figures*. This work would be published in 1822. Begun during the days of his imprisonment, in the *Traité* Poncelet aimed to write a geometry treatise using synthetic methods, yet with the generality of analytic geometry (Taton 1970, 76; Boyer and Merzbach 1991, 535).

The methods Poncelet used in the *Traité* were projective. For example, he used ideal points and introduced a new concept called **imaginary points**. These were defined in such a way that any line and circle would always intersect at one of these points (Boyer and Merzbach 1991, 535). Unfortunately, Poncelet had little familiarity with the work of Desargues and Pascal. The *Brouillon Project* of Desargues was lost soon after he wrote it, and it remained lost until 1847, when the handwritten copy made by Philippe de Lahire was discovered in a Paris library. This was 33 years after Poncelet's time in Russia. The result was that Poncelet spent unnecessary time reproving many of the early projective theorems. Nonetheless, Poncelet made further advancements in the subject, including developments regarding involutions and conic sections (Kline 1972, 840, 842; Boyer and Merzbach 1991, 361).

Poncelet viewed his projective geometry as a new branch of mathematics with its own distinct methods. One example of this is Poncelet's **principle of continuity**. Let us illustrate how it works by considering two theorems regarding secants and tangents to a circle. Let A and D be on a circle and C outside the circle such that they form secants $\overline{AC}$ and $\overline{DC}$ that intersect the circle in points B and D', as in Figure 14.1. We know by the Secant Theorem (5.5.4) that

$$AC \cdot BC = DC \cdot D'C.$$

Now let D move along the circumference of the circle so that it approaches D'. As this happens, the secant $\overline{DC}$ will eventually intersect the circle in only one point, so that it becomes a tangent, and this happens when D and D' coincide. Therefore, the equation obtained from the Secant Theorem becomes

$$AC \cdot BC = DC^2,$$

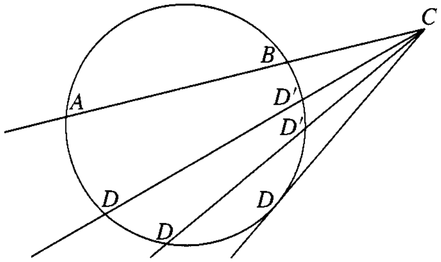

Figure 14.1 The Tangent-Secant Theorem and the principle of continuity.

which is what follows from the Tangent-Secant Theorem (Theorem 5.5.5). Poncelet's strategy is familiar to us. We have seen such continuity arguments used at various times. However, during the nineteenth century mathematicians were becoming more aware of the unstated assumptions that they had been using in the past. A champion of these new standards was Cauchy. He is known for his foundational work on the definitions of limit and continuity in calculus. He believed that it was important that these ideas have a rigorous foundation and was highly critical of Poncelet's principle (Burton 1985, 535; Kline 1972, 843, 950–953).

Poncelet took Cauchy's criticisms personally. He did not consider his principle of continuity a less rigorous form of mathematics. Quite the contrary, Poncelet viewed his principle as a postulate that allowed him to argue as he did. He rejected the notion that his projective geometry needed the analytic foundations that Cauchy apparently desired. Despite Cauchy's formidable reputation as one of the leading mathematicians of the time, Poncelet held his ground and set out to expand and clarify his geometric innovations. The result was an additional opening section to the *Traité*. This addition would make the *Traité* the major work of nineteenth-century projective geometry (Taton 1970, 77).

Partly due to the issues he had with Cauchy, and partly due to other controversies with which he found himself embroiled, Poncelet would eventually all but abandon geometry in favor of applied mechanics. In 1824 he became professor at École d'Application de l'Artillerie et du Génie at Metz. His appreciation of both the theoretical aspect of mechanics and its experimental side made him an excellent teacher. Eventually, the course on machine mechanics that he developed became very popular and widely used at other institutions. Poncelet's interest in machinery was not limited to the classroom, however. His fascination with mechanical devices led to his 1825 tour of France, Belgium, and Germany to visit various factories and study their machinery. Also in that year, Poncelet won a prize from the Académie des Sciences for his design of a more efficient undershot waterwheel. Its secret was in its curved paddles (Taton 1970, 77, 78).

As with many with a military background, Poncelet acquired an interest in politics. In 1830 he became a member of the city council of Metz. Seemingly destined to live the rest of his life in his hometown, plans changed in 1834, when Poncelet accepted an appointment to the school of mechanics at the Académie des Sciences in Paris. There he continued teaching, and later, in 1837, he created a course for the Faculty of Sciences of Paris. He married in 1842, and at this point he was no longer doing original research. He was happy to spend his time teaching and continuing his technical investigations. This all changed with the revolution of 1848. Poncelet was appointed brigadier general and assigned to be the commandant of École Polytechnique. While there, Poncelet was seemingly involved in every aspect of the institution: from reforming the school's curriculum to maintaining student order during the riots of mid-1848 (Taton 1970, 78–79).

Poncelet retired in 1850 and decided to edit and publish his collected works. He completed only four volumes. Two volumes were a second edition of the *Traité des propriétés projectives des figures*, and the other two were titled *Applications d'analyse et de géométrie*. This was the work on analytic geometry that he had written while a prisoner in Russia. Before he could complete the volumes on mechanics, Poncelet died in Paris, on December 22, 1867 (Taton 1970, 76, 78–79).

14.1 THE PROJECTIVE PLANE

We defined the projective plane in Section 9.2 when we studied Desargues. We defined it by adding points and a line at infinity to a Euclidean plane. The modern viewpoint turns this around and first defines the projective plane and then by removing elements constructs a Euclidean one.

Projective Axioms

The work in this section is based on Bennett (1995, 41–49) and the classic Veblen and Young (1910, 15–20). Let $\mathcal{P}$ and $\mathcal{L}$ be two nonempty sets. The elements of $\mathcal{P}$ will be called **points**, and the elements of $\mathcal{L}$ will be called **lines**. Assume that the points are fundamental and that a line is a set of points. We will use geometric terminology when we write sentences about the points and lines, but these should be interpreted with our new understanding. If we say that a point P **lies on** or is **incident** with a line l, this means that $P \in l$. Similarly, the line l **contains** or is **incident** with a point P if $P \in l$. Two points P and Q in $\mathcal{P}$ are **collinear** if there exists a line in $\mathcal{L}$ that contains both P and Q as elements, and two lines m and l are **concurrent** if there exists a point that is an element of both lines. Both collinear and concurrent can be generalized to any number of points and lines. Finally, if points P and Q chosen from $\mathcal{P}$ are the same, write $P = Q$; otherwise, they are **distinct** and we write $P \neq Q$. The pair $(\mathcal{P}, \mathcal{L})$ is called a **projective plane** when it follows these rules:

■ **POSTULATES 14.1.1**

1. For all distinct $P,\ Q \in \mathcal{P}$, there exists a unique line l in $\mathcal{L}$ such that P and Q are incident with l.

2. For all distinct $l,\ m \in \mathcal{L}$, there is a point incident with both l and m.

3. Every line contains at least three points.

4. There are at least two lines.

The geometric theory that satisfies these postulates is called a **projective geometry**. Notice that the second postulate states that projective geometry is elliptic.

Different authors will choose different sets of postulates. These are from Bennett. Veblen and Young use a different collection, which replaces some of the propositions of Postulates 14.1.1. These will serve as our first theorem, but its proof is left to Exercise 10.

■ THEOREM 14.1.2

Let $(\mathcal{P}, \mathcal{L})$ be a projective plane.

- There exists $P, Q \in \mathcal{P}$ such that $P \neq Q$ and P and Q are not collinear.
- For all $P, Q, R \in \mathcal{P}$, if P, Q, and R are not collinear, there exist points S, T, and U in $\mathcal{P}$ with $P \neq S$ such that P, R, and T are collinear, P, Q, and U are collinear, and S, T, and U are collinear.

Consistency

We now show that the projective geometry described by Postulates 14.1.1 is consistent by finding a model.

■ THEOREM 14.1.3

If $\mathcal{P} = \{0, 1, 2, 3, 4, 5, 6\}$ and $\mathcal{L}$ is the set

$$\{\{0, 1, 3\}, \{1, 2, 4\}, \{2, 3, 5\}, \{3, 4, 6\}, \{4, 5, 0\}, \{5, 6, 1\}, \{6, 0, 2\}\},$$

$(\mathcal{P}, \mathcal{L})$ is a projective plane.

PROOF

To prove this, we must systematically check that the four postulates hold under this definition of $\mathcal{P}$ and $\mathcal{L}$.

1. The table enumerates all possible pairs of distinct points and the unique line to which the pair belongs, so the first postulate is seen to be satisfied.

Points		Line	Points		Line	Points		Line
0	1	{0, 1, 3}	1	3	{0, 1, 3}	2	6	{6, 0, 2}
0	2	{6, 0, 2}	1	4	{1, 2, 4}	3	4	{3, 4, 6}
0	3	{0, 1, 3}	1	5	{5, 6, 1}	3	5	{2, 3, 5}
0	4	{4, 5, 0}	1	6	{5, 6, 1}	3	6	{3, 4, 6}
0	5	{4, 5, 0}	2	3	{2, 3, 5}	4	5	{4, 5, 0}
0	6	{6, 0, 2}	2	4	{1, 2, 4}	4	6	{3, 4, 6}
1	2	{1, 2, 4}	2	5	{2, 3, 5}	5	6	{5, 6, 1}

2. The next table shows that all pairs of distinct lines contain a common point.

Lines		Point	Lines		Point
{0, 1, 3}	{1, 2, 4}	1	{2, 3, 5}	{3, 4, 6}	3
{0, 1, 3}	{2, 3, 5}	3	{2, 3, 5}	{4, 5, 0}	5
{0, 1, 3}	{3, 4, 6}	3	{2, 3, 5}	{5, 6, 1}	5
{0, 1, 3}	{4, 5, 0}	0	{2, 3, 5}	{6, 0, 2}	2
{0, 1, 3}	{5, 6, 1}	1	{3, 4, 6}	{4, 5, 0}	4
{0, 1, 3}	{6, 0, 2}	0	{3, 4, 6}	{5, 6, 1}	6
{1, 2, 4}	{2, 3, 5}	2	{3, 4, 6}	{6, 0, 2}	6
{1, 2, 4}	{3, 4, 6}	4	{4, 5, 0}	{5, 6, 1}	5
{1, 2, 4}	{4, 5, 0}	4	{4, 5, 0}	{6, 0, 2}	0
{1, 2, 4}	{5, 6, 1}	1	{5, 6, 1}	{6, 0, 2}	6
{1, 2, 4}	{6, 0, 2}	2			

3. Every line has at least three points by definition.

4. There are clearly at least two lines. ■

There are **infinite models** of projective geometry. The Euclidean projective plane of Desargues (Section 9.2) is an example. Other models include those for single elliptic geometry and the **Riemann sphere**, the complex numbers joined with a point at infinity. The Riemann sphere is denoted by $\mathbb{C}^*$ or $\mathbb{C} \cup \{\infty\}$.

Isomorphic Models

Although one model is sufficient to show the consistency of a set of postulates, there are many different models for projective geometry. Our first was an example of a **finite model**. Here is another. It is named after Gino Fano (1871–1952). Fano was a respected projective and algebraic geometer who was a professor at the University of Turin until he lost his position under the Fascist laws of 1938. During World War II he taught Italian students at Lausanne, and after the war he continued his research and lecturing in both Italy and the United States (Struik 1970b, 522–523).

■ **THEOREM 14.1.4 [Fano Plane]**

If $\mathscr{P}_F$ is the set of dots illustrated by Figure 14.2 and

$$\mathscr{L}_F = \{\overline{EBA}, \overline{ADG}, \overline{EFG}, \overline{BCG}, \overline{ECD}, \overline{ACF}, \odot BDF\},$$

then $(\mathscr{P}_F, \mathscr{L}_F)$ is a projective plane.

We leave it to Exercise 8 to check that the interpretation satisfies Postulates 14.1.1 and that it is the smallest model to do so.

It should be noted that there is actually little difference between Theorem 14.1.3 and Theorem 14.1.4. The only variation between the two are the names of the elements. To see this, let $\mathscr{P}$ be from Theorem 14.1.3 and define $\phi : \mathscr{P}_F \to \mathscr{P}$ by

$$\phi(A) = 0, \phi(B) = 1, \phi(C) = 2, \phi(D) = 5, \phi(E) = 3, \phi(F) = 6, \text{and } \phi(G) = 4.$$

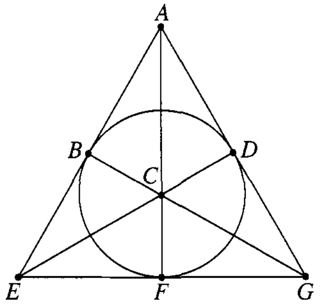

Figure 14.2 The Fano plane.

This function associates pairs of elements in such a way that if three points are on a line in one model, the corresponding points would be on a line in the other model. For instance, $\overline{EBA} = \{E, B, A\}$, $\phi(E) = 3$, $\phi(B) = 1$, and $\phi(A) = 0$. Thus,

$$\phi[\{E, B, A\}] = \{0, 1, 3\},$$

and for the rest of the lines we have

$$\begin{aligned}\phi[\{A, D, G\}] &= \{4, 5, 0\},\\ \phi[\{E, F, G\}] &= \{3, 4, 6\},\\ \phi[\{B, C, G\}] &= \{1, 2, 4\},\\ \phi[\{E, C, D\}] &= \{2, 3, 5\},\\ \phi[\{A, C, F\}] &= \{6, 0, 2\},\\ \phi[\{B, D, F\}] &= \{5, 6, 1\}.\end{aligned}$$

We now generalize this result.

■ DEFINITION 14.1.5

Let $(\mathcal{P}, \mathcal{L})$ and $(\mathcal{P}', \mathcal{L}')$ be projective planes. If $\phi : \mathcal{P} \to \mathcal{P}'$ is a one-to-one correspondence such that for all $P, Q, R \in \mathcal{P}$,

$$R \in \overline{PQ} \text{ if and only if } \phi(R) \in \overline{\phi(P)\phi(Q)},$$

then ϕ is a **collineation**.

If there is a collineation between two projective planes, the planes look identical except for the names put on their elements.

■ DEFINITION 14.1.6

The projective planes $(\mathcal{P}, \mathcal{L})$ and $(\mathcal{P}', \mathcal{L}')$ are **isomorphic** if there exists a collineation $\phi : \mathcal{P} \to \mathcal{P}'$.

Independence

Because of the work by Beltrami, Klein, and Poincaré in Section 13.4, we know that the Parallel Postulate is independent (page 324) of Euclid's other postulates. This means that the Parallel Postulate can neither be proven true nor proven false from Euclid's first four assumptions. It is important for any proposition in a set of postulates to have this property because we never want to assume a proposition that could have been proven from the others. We say that a set of propositions is **independent** if every proposition of the set is independent of the others. To prove that a set of propositions $\{p_1, p_2, \ldots, p_k\}$ is independent, for each i we must show that both

$$\{p_1, p_2, \ldots, p_i, \ldots, p_k\}$$

and

$$\{p_1, p_2, \ldots, \textit{not } p_i, \ldots, p_k\}$$

are consistent. In this way, neither *not* p_i nor p_i necessarily follow from the other postulates.

To prove that the set of projective postulates is independent, we first observe that the Fano plane shows that the postulates are consistent. It remains to show that the negation of any of the projective postulates plus the remaining three is a consistent set of propositions. The diagram of Figure 14.3(a) shows that it is possible that the first postulate is false while the others are true, and Figure 14.3(b) shows that it is possible for the third postulate to be false with the others true. The Euclidean plane is a model in which Postulates 1, 3, and 4 are satisfied but Postulate 2 is not because there are parallel lines in the Euclidean plane. The empty set is a model in which the last postulate is false because that postulate entails the existence of objects in the geometry. The others are **vacuously true** on the empty set. This refers to the fact that a universal sentence can be written as a conditional. For example, Postulate 14.1.1.1 can be stated as

if P and Q are distinct points in $\mathcal{P}$, there exists a unique line l in $\mathcal{L}$ such that P and Q are incident with l.

If $\mathcal{P} = \varnothing$, the antecedent of the conditional is false, and this implies that the conditional is true (page 38).

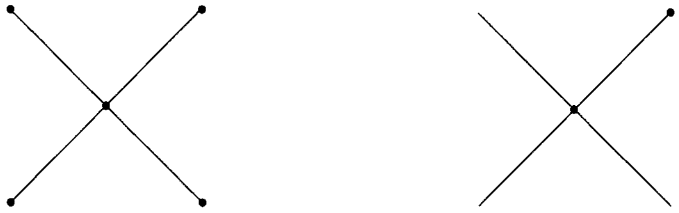

(a) Negation of Postulate 1 (b) Negation of Postulate 3

Figure 14.3 Independence of the projective postulates.

Exercises

1. Given the following set of postulates for **incidence geometry**:

   Postulate 1. There exist at least three points.
   Postulate 2. Two distinct points are on exactly one line.
   Postulate 3. There are at least two points on every line.
   Postulate 4. For every line l, there is a point not on l.

   (a) Show that this is a consistent set of postulates using a finite model.
   (b) Find an infinite model for this geometry.
   (c) Show that this is an independent set of postulates.
   (d) Determine whether incidence geometry is a projective geometry.

2. Given the following set of postulates for **three-point geometry**:

   Postulate 1. There exist exactly three points.
   Postulate 2. Two distinct points are on exactly one line.
   Postulate 3. Every two distinct lines are on at least one point.
   Postulate 4. For every line l, there is a point not on l.

   (a) Show that this is a consistent set of postulates.
   (b) Show that this is an independent set of postulates.
   (c) Determine whether three-point geometry is a projective geometry.

3. Prove the following theorems in three-point geometry.
   (a) Two distinct lines are on exactly one point.
   (b) There exist exactly three lines.
   (c) Lines cannot contain three distinct points.
   (d) There exists a set of two lines containing all of the points.
   (e) There exists a set of two points such that each line lies on at least one of the points.

4. Given the following set of postulates for **four-line geometry**:

   Postulate 1. There exist exactly four lines.
   Postulate 2. Two distinct lines intersect on exactly one point.
   Postulate 3. Every point is on exactly two lines.

   (a) Show that this is a consistent set of postulates.
   (b) Show that this is an independent set of postulates.
   (c) Determine whether four-line geometry is a projective geometry.

5. Given the following set of postulates for **Fano's geometry**:

   Postulate 1. There exists at least one line.
   Postulate 2. Every line contains exactly three points.
   Postulate 3. Not all points are on the same line.
   Postulate 4. Two distinct points are on exactly one line.
   Postulate 5. Two distinct lines intersect in at least one point.

(a) Show that this is a consistent set of postulates.
(b) Show that this is an independent set of postulates.
(c) Determine whether Fano's geometry is a projective geometry.

6. Prove the following theorems in Fano's geometry:
   (a) Every two lines have exactly one point in common.
   (b) There exist exactly seven points and seven lines.
   (c) Each point lies on exactly three lines.
   (d) Lines through any given point contain all of the points.
   (e) For any pair of points, there is exactly two lines not containing either point.
   (f) Given a set of three lines not all containing the same point, there exists exactly one point not on any of the three lines.

7. **Young's geometry** shares the first four postulates of Fano's geometry, but Postulate 5 is replaced with

> **Postulate 5.** If a point does not lie on a given line, there exists exactly one line on the point that does not intersect the given line.

   (a) Show that this is a consistent set of postulates.
   (b) Show that this is an independent set of postulates.
   (c) Determine whether Young's geometry is a projective geometry.

8. Show that the Fano plane is the smallest model of projective geometry.

9. Show that two lines in a projective plane intersect in exactly one point.

10. Prove Theorem 14.1.2.

11. Let $(\mathcal{P}, \mathcal{L})$, $(\mathcal{P}', \mathcal{L}')$, and $(\mathcal{P}'', \mathcal{L}'')$ be projective planes. Suppose that

$$\phi : \mathcal{P} \to \mathcal{P}'$$

and

$$\psi : \mathcal{P}' \to \mathcal{P}''$$

are collineations. Show that following:
   (a) ϕ^{-1} is a collineation.
   (b) $\psi \circ \phi$ is a collineation.

12. Do any of the five geometries defined in this exercise set have two nonisomorphic models? If so, which ones?

14.2 DUALITY

Throughout this section we assume that all points and lines are from the projective plane $(\mathcal{P}, \mathcal{L})$. We begin this section by proving a sequence of four theorems that follow quickly from the definition of a projective plane.

■ THEOREM 14.2.1

For all distinct l, $m \in \mathscr{L}$, there is a unique point $P \in \mathscr{P}$ so that P is on l and m.

PROOF

Let l, $m \in \mathscr{L}$. By Postulate 14.1.1.2 we know that there exists a point P that is incident with l and m. We must show that P is unique, so suppose that there exists another point Q incident with l and m that is different from P. This means that we have two distinct points on two distinct lines, which contradicts Postulate 14.1.1.1. ■

■ THEOREM 14.2.2

For all distinct P, $Q \in \mathscr{P}$, there is a line incident with both P and Q.

PROOF

This is just Postulate 14.1.1.1 without the uniqueness condition. ■

■ THEOREM 14.2.3

Every point is incident with at least three lines.

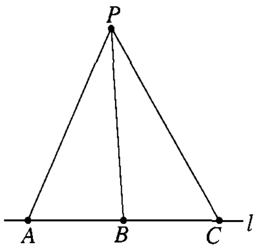

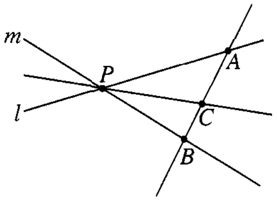

PROOF

Let P be a point and let l and m be the two lines guaranteed by Postulate 14.1.1.4. We have two cases to consider.

- Suppose that P is not on l. We know by Postulate 14.1.1.3 that l has three points on it. Call them A, B, and C. By Postulate 14.1.1.1 we may join $\overline{PA}$, $\overline{PB}$, and $\overline{PC}$. None of these lines equals l because they contain P but l does not. We also see that $\overline{PA}$, $\overline{PB}$, and $\overline{PC}$ are all distinct because if we suppose that any two of these lines are the same, there would be a pair of points on two distinct lines contradicting Postulate 14.1.1.1. For example, if $\overline{PA}$ is the same line as $\overline{PB}$, we conclude that A and B are on both $\overline{PA}$ and l.

- Now let P be on l. If P is not on m, we may construct three lines as we did in the first part, so suppose that P is also on m. We know that there is a point A on l and a B on m that are both different from P. Join $\overline{AB}$ by Postulate 14.1.1.1, and by Postulate 14.1.1.3 find another point C on

$\overline{AB}$. Join $\overline{PC}$. This is the third line through P because if $\overline{PC}$ were either equal to l or m, we would have two distinct points on two distinct lines. ■

■ **THEOREM 14.2.4**

There are at least two points.

PROOF

Postulate 14.1.1.4 tells us that there is a line, and by Postulate 14.1.1.3 there are at least three points on that line. Therefore, there are at least two points. ■

When we compare the four theorems with Postulates 14.1.1, we see that they are basically the same except that the words *point* and *line* have been interchanged. What this means is that every theorem about the projective plane has a corresponding theorem called its **dual** that is obtained from the original by exchanging every concept related to the word *point* with the corresponding concept related to the word *line*. This dual follows automatically without further proof once the original theorem is proven. This is a very important property of projective geometry called the **principle of duality**.

The principle of duality was discovered by Poncelet but not without controversy. In 1810 an artillery officer by the name of Joseph-Diez Gergonne published the first privately published journal of mathematics. He titled it the *Annales de Mathématiques Pures et Appliquées*. Gergonne was a fine mathematician and a strong supporter of analytic methods. Poncelet, on the other hand, was a staunch defender of the synthetic. These two soon became point men in the synthetic–analytic debate that had previously embroiled Klügel and Kästner with the English (page 355). At first their competition was civil, with both publishing papers espousing their points of view in Gergonne's *Annales* in 1818. Needless to say, the controversy was not settled, and years later, in 1826, matters heated up when both men claimed priority in the discovery of the principle of duality. Gergonne claimed that he had proven it using analytic methods, while Poncelet insisted that he had discovered duality during a study of the conic sections. The reality of the situation was that although the term *dual* was due to Gergonne, it was Poncelet who was first (Boyer and Merzbach 1991, 506–507, 536; Kline 1972, 845).

The Complete Quadrilateral

To illustrate the principle of duality, we consider two pairs of geometric figures. The first pair is the **triangle**,

three noncollinear points that join to form three sides,

and the **trilateral**,

three nonconcurrent lines that meet to form three vertices.

These definitions are duals of each other. One sentence is formed from the other by making the replacement:

point $\longleftrightarrow$ line
join $\longleftrightarrow$ meet
vertex $\longleftrightarrow$ side
collinear $\longleftrightarrow$ concurrent

When we draw both of these figures, we find that they are the same. We say that the figures are **self-dual**.

The second pair involves the complete quadrangle. We repeat its definition:

■ DEFINITION 14.2.5

Let A, B, C, and D be four points in a plane, no three of which are collinear. The configuration that consists of these four points and the six lines determined by all possible pairs of the points is called the **complete quadrangle** $ABCD$.

Since being *concurrent* is the dual of being *collinear*, we may define the dual of the complete quadrangle.

■ DEFINITION 14.2.6

Let a, b, c, and d be four lines in a plane, no three of which are concurrent. The configuration that consists of these four lines and the six points determined by all possible pairs of the lines is called the **complete quadrilateral** $abcd$.

Compare the following with the content on page 315. The lines of the complete quadrilateral $abcd$ are called its **sides**, and the points of $abcd$ are its **vertices**. The complete quadrilateral $abcd$ is illustrated in Figure 14.4. It consists of the four lines a, b, c, and d and the six vertices

$$ab,\ bc,\ cd,\ ad,\ ac,\ bd,$$

where ab represents the intersection of lines a and b. Two vertices are said to be **opposite** if they do not share a common side. There are three pairs of opposite

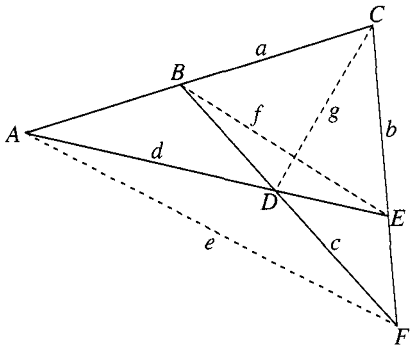

Figure 14.4 A complete quadrilateral $abcd$.

vertices. They lie on three lines called **diagonal lines**. The intersection of these lines form the **diagonal trilateral** of $abcd$. The diagonal trilateral of the quadrilateral in Figure 14.4 is Δefg.

Harmonic Sets

In Books III and IV of the *Conic Sections*, Apollonius determines necessary and sufficient conditions for a line through a point outside a conic section to be tangent to that section (Kline 1972, 96–97). Let A be a point external to a conic such that $\overline{AP}$ and $\overline{AQ}$ are tangent to the given conic section (Figure 14.5). Suppose that P and Q are the points of tangency. Join $\overline{PQ}$. This segment is called the **polar** of the **pole** A. Through any point C on $\overline{PQ}$, draw the line that joins A and C and let B and D be the points of intersection of this line with the conic. Apollonius showed that

$$\frac{AB}{AD} = \frac{CB}{CD}. \tag{14.1}$$

The phrase that is traditionally used to describe this proportion is:

A divides $\overline{BD}$ externally in the same ratio as C divides it internally.

Conversely, suppose that we have a point outside a conic and want to find the tangents to the conic that pass through the point. To do this, we again let A be off the conic and external to it. Draw a line through A that meets the conic at B and D. Draw another line through A that intersects the conic at B' and D' (Figure 14.5). On $\overline{BD}$ find C such that Equation 14.1 holds and find C' on $\overline{B'D'}$ so that

$$\frac{AB'}{AD'} = \frac{C'B'}{C'D'}.$$

Apollonius showed that the line through C and C' intersects the conic section at the points P and Q, respectively, so that $\overline{AP}$ and $\overline{AQ}$ are tangent lines.

Now let us change our perspective. Take quadrangular set Q$(ABC,\ EDF)$. By Theorem 9.4.5,

$$\frac{AB \cdot EB}{AD \cdot ED} = \frac{CB \cdot FB}{CD \cdot FD}.$$

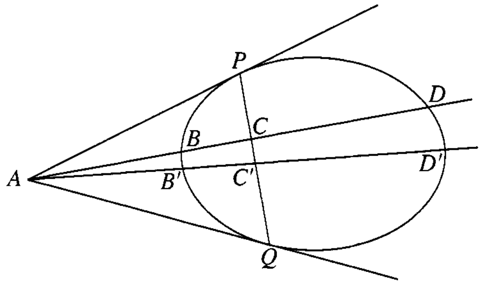

Figure 14.5 A pole and polar of a conic section.

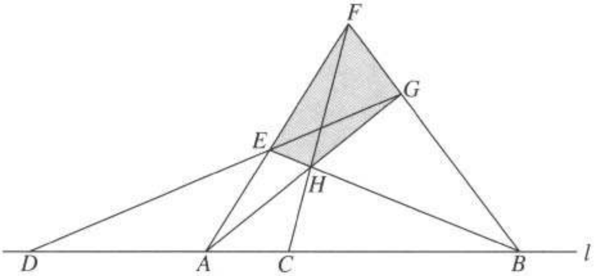

Figure 14.6 H(AB, CD).

Suppose that point A coincides with E and that C coincides with F. We then have

$$\frac{AB \cdot AB}{AD \cdot AD} = \frac{CB \cdot CB}{CD \cdot CD},$$

and this is equivalent to Equation 14.1. Desargues called the original points A, B, C, and D four points in involution (Exercise 4 of Section 9.4). We have a different name for this relation. It is defined in terms of a modified quadrangular set and illustrated in Figure 14.6.

■ DEFINITION 14.2.7

The four collinear points A, B, C, D form a **harmonic set of points** if there is a complete quadrangle such that points A and B are two of the vertices of its diagonal triangle and points C and D are on the sides through the third vertex of the diagonal triangle. The harmonic set of points is denoted by H(AB, CD).

Observe that

$$\mathsf{H}(AB, CD) \text{ if and only if } \mathsf{Q}(ABC, ABD).$$

Moreover, like the notation for involutions (page 313), there are permutations of the points of a harmonic set that immediately from the definition also form harmonic sets, but H(CD, AB) is not one of them. That H(AB, CD) implies H(CD, AB) requires proof.

■ THEOREM 14.2.8

For any four points A, B, C, D, if H(AB, CD), then H(CD, AB).

PROOF

Let A, B, C, and D be collinear such that H(AB, CD). This means there exists complete quadrangle $EFGH$ such that A is the intersection of $\overline{EF}$ and $\overline{HG}$, B is the intersection of $\overline{EH}$ and $\overline{FG}$, C is on $\overline{FH}$, and D is on $\overline{EG}$. To prove H(CD, AB) we must find a complete quadrangle such that C and D

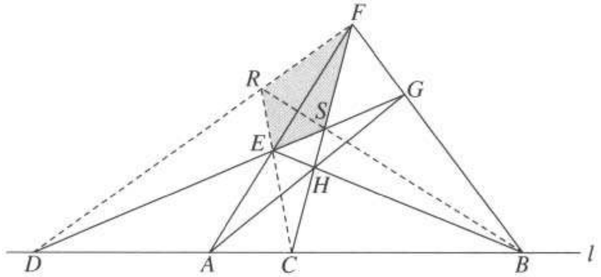

are vertices of its diagonal triangle and A and B are on lines through the third vertex of its diagonal triangle. To do this, join $\overline{DF}$ and $\overline{EC}$. Extend $\overline{EC}$ to meet $\overline{DF}$ at R. Let S be the intersection of $\overline{DG}$ and $\overline{CF}$. We want R, S, and B to be collinear. To obtain this, consider $\triangle REF$ and $\triangle SHG$. Notice that $\overline{RF}$ and $\overline{SG}$ meet at D, $\overline{EF}$ and $\overline{HG}$ meet at A, and $\overline{RE}$ and $\overline{SH}$ meet at C. This means that the points of intersection of pairs of corresponding sides of the two triangles are collinear. Therefore, the lines joining corresponding vertices are concurrent by Desargues' Theorem (9.3.2). Since we know that $\overline{EH}$ and $\overline{FG}$ already intersect at B, it must be the case that the $\overline{RS}$ produced will contain B. We conclude that $RFSE$ is the desired complete quadrangle because C and D are vertices of its diagonal triangle and A and B are on lines that intersect the diagonal triangle's third vertex. Therefore, $\mathsf{H}(CD,\ AB)$. ■

Choose three points A, B, and C on a line l. We can prove that there is a unique point D on l so that $\mathsf{H}(AB,\ CD)$. The point D is called the **harmonic conjugate** of C with respect to A and B.

■ THEOREM 14.2.9

Let A, B, and C be three points on a line l. There exists exactly one point D on l such that $\mathsf{H}(AB,\ CD)$.

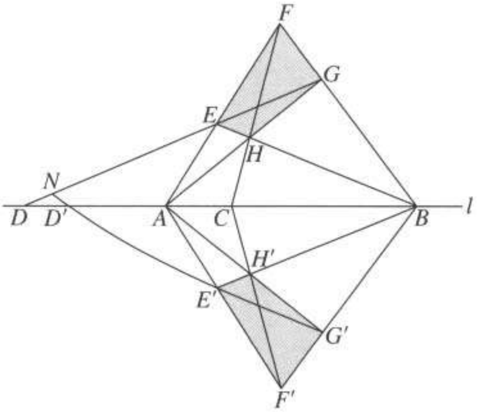

PROOF

Let l be a line and take A, B, and C to be points on l. We first show that the desired D exists. Take a point F not on l. After joining $\overline{FB}$, pick G on that segment. Join $\overline{AG}$ and $\overline{FC}$ and let these segments intersect at H (Theorem 9.2.8). Join $\overline{AF}$ and $\overline{BH}$, producing $\overline{BH}$ to meet $\overline{AF}$ at E. Take the point D to be the intersection of l and $\overline{GE}$ produced. Since $EFGH$ is a complete quadrangle, we see that $\mathsf{H}(AB,\ CD)$.

To show that D is unique, repeat the construction so that we have complete quadrangle $E'F'G'H'$ with $\overline{E'F'}$, $\overline{F'G'}$, and $\overline{H'F'}$ intersecting l at A, B, and C, respectively, but $\overline{E'G'}$ meeting l at D', which is different from D. By Theorem 9.2.8, there is a point N that is incident with both $\overline{GD}$ and $\overline{G'D'}$. First consider $\triangle EFH$ and $\triangle E'F'H'$. Since their corresponding sides intersect at A, C, and B on line l, we conclude by Desargues' Theorem (9.3.2) that $\overline{EE'}$, $\overline{HH'}$, and $\overline{FF'}$ are concurrent (not pictured). The same argument using $\triangle FGH$ and $\triangle F'G'H'$ allows us to conclude that $\overline{HH'}$, $\overline{FF'}$, and $\overline{GG'}$ are concurrent. Therefore, the lines joining corresponding vertices of $\triangle HEG$ and $\triangle H'E'G'$ are concurrent, so by another application of Desargues' Theorem (9.3.1), points N, A, and B are collinear. This means that N must be on l, so $D = D'$. ■

The dual of a harmonic set of points is a harmonic set of lines.

■ DEFINITION 14.2.10

The four concurrent lines a, b, c, d form a **harmonic set of lines** if there exists a complete quadrilateral such that a and b are two of the sides of its diagonal trilateral and c and d are on the points on the third side of the diagonal trilateral. When a, b, c, and d form a harmonic set in this fashion, we write $\mathsf{H}(ab,\ cd)$.

Exercises

1. Given ellipse $x^2 + 4y^2 = 16$, find
   (a) the polar of $(-7,\ 0)$.
   (b) the pole of the vertical polar at $x = 2$.
   (c) the pole of the segment that joins the points of intersection of $y = 2x + 3$ and the ellipse.

2. Write the duals of each of the following sentences.
   (a) *There exist at least two lines.*
   (b) *There are at least two lines on every point.*
   (c) *Not all points lie on the same line.*
   (d) *Two distinct lines intersect in a unique point.*
   (e) *The line intersects one of the sides of a complete quadrangle.*
   (f) *The line joining two points contains the third point.*
   (g) *All points lie on either the first or the second line.*

3. Brianchon's Theorem is the dual of Pascal's Theorem (9.3.5). State Brianchon's Theorem.

4. Let $\mathsf{H}(AB,\ CD)$. Write seven other orders of points A, B, C, and D that form harmonic sets.

5. Construct a complete quadrangle given its diagonal triangle and one vertex.

6. Let A, B, C, and D be collinear points. Find P and Q such that $\mathsf{H}(AB,\ PQ)$ and $\mathsf{H}(CD,\ PQ)$.

7. The **cross ratio** of the four collinear points A, B, C, and D is

$$(AB,\ CD) = \frac{AC \cdot BD}{AD \cdot BC}.$$

Show that $\mathsf{H}(AB,\ CD)$ if and only if $(AB,\ CD) = 1$.

8. Let A, B, C, and D be collinear. Show the following:
   (a) $(AB,\ CD) = (CD,\ AB)$.
   (b) $(AB,\ DC) = 1/(AB,\ CD)$.
   (c) Use these two results to confirm the findings of Exercise 4.

9. Assume that for any ordinary points A and B and ideal point Ω, $A\Omega = B\Omega$ because both distances are infinite. With this in mind, prove that if $\mathsf{H}(AB,\ CD)$ and C is the midpoint of $\overline{AB}$, then D is an ideal point.

10. Let A, B, C, D, E be collinear. Show that:
   (a) $(AB,\ CD) = (AB,\ CE) \cdot (AB,\ ED)$.
   (b) $(AB,\ CD) = (AE,\ CD) \cdot (EB,\ CD)$.
   (c) $(AB,\ DE) = (AB,\ CD)/(AB,\ CE)$

11. Let O, A, B, C, A', B', and C' be collinear and

$$OA \cdot OA' = OB \cdot OB' = OC \cdot OC'.$$

Prove that $\mathsf{H}(AB',\ BC)$ if and only if $\mathsf{H}(A'B,\ B'C')$.

12. Prove that $\mathsf{H}(AB,\ CD)$ if and only if

$$\frac{1}{AB} = \frac{1}{2}\left(\frac{1}{AC} + \frac{1}{AD}\right).$$

13. For the given sets of three points, find the Cartesian coordinates of the harmonic conjugate of C with respect to A and B.
   (a) $A(0,\ 0)$, $B(3,\ 0)$, $C(8,\ 0)$
   (b) $A(-1,\ 3)$, $B(1,\ 2)$, $C(5,\ 0)$

14. Write the duals for Theorems 14.2.8 and 14.2.9.

15. Let a, b, and c be concurrent lines. Explain the procedure for finding the line d that is the harmonic conjugate of the given lines.

16. Prove that given concurrent lines a, b, and c, there exists a unique line d such that $\mathsf{H}(ab, cd)$.

17. Suppose that l intersects concurrent lines a, b, c, and d in points A, B, C, and D. Prove that $\mathsf{H}(AB, CD)$ if and only if $\mathsf{H}(ab, cd)$.

14.3 PERSPECTIVITY

Poncelet made new contributions to the projection of Desargues. Let $(\mathcal{P}, \mathcal{L})$ be a projective plane, and take $l, m \in \mathcal{L}$, $O \in \mathcal{P}$ but O is concurrent with neither l nor m. Define $T : l \to m$ such that $T(A) = A'$ if and only if A and A' are collinear with O (Figure 14.7). The function T is called a **central perspectivity** and O is the **center of perspectivity**. We represent $T(A) = A'$ by

$$A \overset{O}{\barwedge} A'.$$

Read this as "A and A' are in **perspective** with respect to O."

The central perspectivity is the projection of the earlier projective geometers (Section 9.1), and it is an invertible function.

■ THEOREM 14.3.1

A central perspectivity is a one-to-one correspondence.

PROOF

Let $T : l \to m$ be a central perspectivity with center O. Let A and B be on l and assume that

$$T(A) = T(B) = A'.$$

If $A \neq B$, we conclude that there are two distinct lines that pass through O and A'. This contradicts Postulate 14.1.1.1. Hence, $A = B$, and we conclude that T is one-to-one. To prove that T is onto, let A' be incident with m and choose A to be the intersection of $\overline{OA'}$ with l. This means that $T(A) = A'$. ■

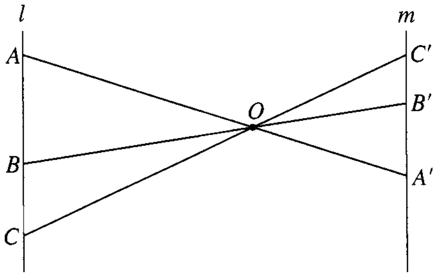

Figure 14.7 A central perspectivity.

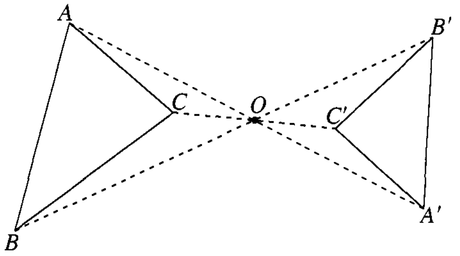

(a) $\triangle ABC$ and $\triangle A'B'C'$ are perspective from P

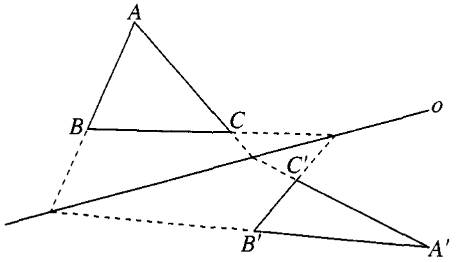

(b) $\triangle ABC$ and $\triangle A'B'C'$ are perspective from p

Figure 14.8 Desargues' Theorem.

Although a perspectivity is a function between two lines, there is a natural way to generalize it to plane figures (Figure 14.8).

■ DEFINITION 14.3.2

Two figures in the plane are **perspective from a point** if there exists a point O in the plane such that any line joining corresponding points from the figures contain O.

The dual of two figures perspective from a point is perspective from a line.

■ DEFINITION 14.3.3

Two figures in the plane are **perspective from a line** if there exists a line o in the plane such that all corresponding lines from the figures intersect at o. The line o is called the **axis of perspectivity**.

The new terminology allows us to restate Desargues' Theorem (9.3.1 and 9.3.2). Remember that it was required in the proof to pass to the third dimension. Since we are limiting ourselves to the projective plane, a proof would require an additional postulate justifying a key step (Coxeter 2003, 18–20). We will, nonetheless, still label it as a theorem.

THEOREM 14.3.4 [Desargues]

Two triangles are perspective from a point if and only if they are perspective from a line.

Projectivities

Since perspectivities are functions, we can compose them under the right conditions. This amounts to taking a sequence of points O_1, O_2, ..., O_k and lines l_1, l_2, ..., l_k, l_{k+1} such that O_i is not on l_i or l_{i+1} and defining perspectivities $T_i : l_i \to l_{i+1}$ with center of perspectivity O_i. The resulting function plays a central role in the study of projective geometry.

DEFINITION 14.3.5

The composition of central perspectivities is called a **projectivity**.

Since the composition of one-to-one and onto functions is one-to-one and onto, we conclude that projectivities are one-to-one correspondences and thus invertible.

As an example, let l and m be two lines with A, B, $C \in l$ and A', B', $C' \in m$ (Figure 14.9). We claim that we can find a projectivity T between l and m such that

$$T(A) = A', T(B) = B', T(C) = C',$$

and T is the composition of two central perspectivities. To do this, find C'' not on l or m. Join $\overline{AA'}$ and $\overline{CC''}$ and let O be their intersection. Let B'' be the intersection of $\overline{BO}$ and $\overline{C''A'}$. Now join $\overline{C''C'}$ and $\overline{B''B'}$ and let their intersection by O'. We now have two central perspectivities, one with center O and the other with center O'. Their composition is our desired T because

$$A \overset{O}{\barwedge} A' \overset{O'}{\barwedge} A', \; B \overset{O}{\barwedge} B'' \overset{O'}{\barwedge} B', \text{ and } C \overset{O}{\barwedge} C'' \overset{O'}{\barwedge} C'.$$

For a second example we write the identity map on a line l as the composition of three central perspectivities. Let A, B, and C be points on l (Figure 14.10). Take

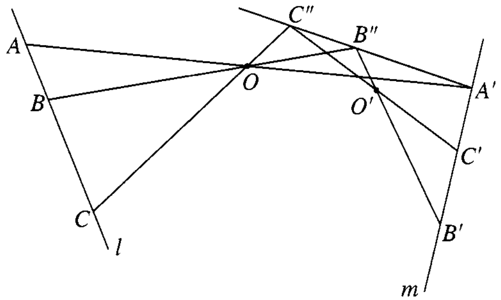

Figure 14.9 A projectivity that is the composition of two central perspectivities.

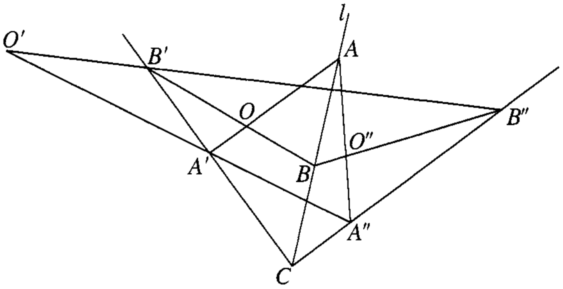

Figure 14.10 The identity map is the composition of three perspectivities.

B' and B'' not on l. Let A' be on $\overline{B'C}$ but not equal to either B' or C. Let O be the intersection of $\overline{A'A}$ with $\overline{B'B}$. Find a new point A'' on $\overline{CB''}$ and let O'' be the intersection of $\overline{AA''}$ and $\overline{BB''}$. Finally, let O' be the intersection of $\overline{A'A''}$ and $\overline{B'B''}$. This defines three perspectivities. Their centers are O, O', and O'', so we may write

$$A \overset{O}{\barwedge} A' \overset{O'}{\barwedge} A'' \overset{O''}{\barwedge} A$$

and similarly for B and C. These three points are invariants.

DEFINITION 14.3.6

A point that is mapped to itself under a projectivity is called **invariant**.

The question now is whether the composition of these three perspectivities yields the identity on l. If we pick a few random points from l and apply the projectivity (Exercise 14), we discover that the checked points are invariant. This, of course, is far from a proof.

Harmonic Nets

As Desargues was interested in what properties were preserved by projections, modern projective geometry asks what properties are preserved by projectivities. These are essentially the same question, so we should expect the next theorem. Compare this with Theorem 9.4.2. We state it without proof.

THEOREM 14.3.7

Let T be a projectivity with A, B, C, D in the domain of T. If $\mathsf{H}(AB,\ CD)$, then $\mathsf{H}(T(A)T(B),\ T(C)T(D))$.

Suppose that T is a projectivity with the line l as its domain. Take A, B, $C \in l$. By Theorem 14.2.9, there exists a unique $D \in l$ such that $\mathsf{H}(AB,\ CD)$. Therefore, by Theorems 14.3.7 and 14.2.9, $T(D)$ is the harmonic conjugate of $T(C)$ with respect to $T(A)$ and $T(B)$, and $T(D)$ could equal no other point. Therefore,

T is determined uniquely on $\{A, B, C, D\}$
because it is determined on $\{A, B, C\}$.

There also are unique points E and F in the domain of T and not among $\{A, B, C, D\}$ such that $\mathsf{H}(AC, BE)$ and $\mathsf{H}(BC, AF)$ (Exercise 15). This means that $T(E)$ and $T(F)$ are determined uniquely because projectivities preserve harmonic sets. Hence,

T is determined uniquely on $\{A, B, C, D, E, F\}$
because it is determined on $\{A, B, C\}$.

We may continue in this manner and either exhaust all of the points of l or find a countably infinite collection of points of l on which T is determined uniquely. Call the resulting set a **harmonic net** of l. All of this is restated as the next theorem.

■ THEOREM 14.3.8

There exists a unique projectivity that maps three distinct points of a harmonic net onto three distinct points in a given order.

The identity map I on l is a projectivity (Exercise 8) that leaves all points on a harmonic net invariant. If a projectivity T with domain l leaves three points invariant on the same harmonic net, we know that T and I must agree on all points of the harmonic net. This proves the corollary.

■ COROLLARY 14.3.9

A projectivity that leaves three points invariant on a harmonic net of points leaves all points invariant on the same harmonic net.

The corollary also implies Theorem 14.3.8 (Exercise 18), so the two statements are equivalent. Our ultimate goal is to generalize these equivalent statements so that they hold on the entire domain of a projectivity. If the domain is finite, the domain is a harmonic net, and the generalization follows easily. Otherwise, we are left in an impossible situation. It turns out that any attempt at generalizing Theorem 14.3.8 or its corollary requires knowing that the harmonic net is distributed on the domain of a projectivity in much the same way that the rational numbers are spread out on the real line. This seems to require the concepts of distance and convergence that have been deliberately left out of the projective postulates. Therefore, since we think both generalizations are reasonable, we will choose one of them as an assumption and prove that the other follows.

■ POSTULATE 14.3.10

If a projectivity leaves three points invariant, it is the identity.

With this we may prove the central theorem of Poncelet's new geometry.

■ THEOREM 14.3.11 [Fundamental Theorem of Projective Geometry]

A projectivity is determined uniquely by three points.

PROOF

Let T and T' be projectivities with domain l. Assume that $T(A) = T'(A) = A'$, $T(B) = T'(B) = B'$, and $T(C) = T'(C) = C'$, but $T(D) = D'$ and $T'(D) = D''$ with $D' \neq D''$. We know that T^{-1} exists (Theorem 14.3.1), so we check:

$$(T' \circ T^{-1})(A') = T'(T^{-1}(A')) = T'(A) = A'.$$

Similarly, B' and C' are invariant under $T' \circ T^{-1}$. But $(T' \circ T^{-1})(D') = D''$. This contradicts Postulate 14.3.10 because $T' \circ T^{-1}$ is a projectivity that holds three points invariant but is not the identity. ■

The Fundamental Theorem answers the question posed by the example on page 504 in the affirmative because the composition defined there and the identity are projectivities that agree on three points. The Fundamental Theorem also shows that the projectivity found in the first example on page 503 is unique (Exercise 17).

Exercises

1. Show that a central perspectivity has exactly one center if the perspectivity's domain and range are different.

2. Is it possible for a central perspectivity to have two centers if its domain and range share a point?

3. Prove that a central perspectivity is the identity map if the domain equals the range.

4. Let l and m be lines with $A,\ B \in l$ and $A',\ B' \in m$. Find a central perspectivity with center O such that

$$A \overset{O}{\barwedge} A' \text{ and } B \overset{O}{\barwedge} B'.$$

5. Show that a central perspectivity is determined uniquely by two points.

6. The dual of a central perspectivity is an **axial perspectivity**.
   (a) Define an axial perspectivity.
   (b) Without simply writing the dual of Theorem 14.3.1, show that an axial perspectivity is a one-to-one correspondence.
   (c) What is the dual of a projectivity?

7. Define central perspectivity and axial perspectivity using the terms *pencil* and *range*.

8. Show that the identity map on a line is a projectivity.

9. Prove that a projectivity $T : l \to l$ can be written as the composition of three central perspectivities.

10. Let T be a projectivity that leaves exactly one point invariant. Prove that T is a perspectivity.

11. Define $(ABCD) \barwedge (A'B'C'D')$ to mean that there exists a projectivity T such that $T(A) = A'$, $T(B) = B'$, $T(C) = C'$, and $T(D) = D'$. Show for any collinear A, B, C, and D, we have $(ABCD) \barwedge (BADC) \barwedge (CDAB) \barwedge (DCBA)$.

12. Prove that $(AECF) \barwedge (BDCF)$ if and only if $\mathsf{Q}(ABC,\ DEF)$.

13. Draw an example of a projectivity that leaves exactly two points invariant.

14. Choose a point on l in Figure 14.10 and confirm that it is invariant under the composition of the three perspectivities.

15. Assume $\mathsf{H}(AB,\ CD)$, where A, B, C, and D are on a line l. Show that there exists E and F on l and not among $\{A,\ B,\ C,\ D\}$ such that $\mathsf{H}(AC,\ BE)$ and $\mathsf{H}(BC,\ AF)$.

16. Prove that a harmonic net is determined uniquely by three points.

17. Draw a line l containing the points A, B, and C and another line containing A', B', and C'. Let T and T' be projectivities such that

$$T(A) = T'(A) = A',\ T(B) = T'(B) = B', \text{ and } T(C) = T'(C) = C'.$$

Draw T as the composition of two central perspectivities with centers O and O' as on page 503. Also, draw T' as the composition of two central perspectivities with centers P and P'. Pick a point D on l and confirm that $T(D) = T'(D)$.

18. Prove that Corollary 14.3.9 implies Theorem 14.3.8.

14.4 HOMOGENEOUS COORDINATES

To assign points using Cartesian coordinates, we fix a reference point called the origin, draw two perpendicular lines through it, and then by measuring distances horizontally and vertically along a grid of parallel lines, a point is labeled using an ordered pair consisting of those distances. But how do you assign coordinates in a system that does not allow for angle measure, has no parallel lines, and has no concept of distance? We use triangulation and the methods of linear algebra.

Points and Lines

Let E_1, E_2, and E_3 be points in a projective plane. Label E_1 as $(1,\ 0,\ 0)$, E_2 as $(0,\ 1,\ 0)$, and E_3 as $(0,\ 0,\ 1)$, as in Figure 14.11. Join $\overline{E_1E_2}$, $\overline{E_2E_3}$, and $\overline{E_1E_3}$ to form the **reference triangle**. Finding new points on these three lines requires the introduction of a fourth point E that we will label $(1,\ 1,\ 1)$ and place in the middle of the reference triangle. We call E the **unit point**. Joining $\overline{EE_1}$ results in a new point A on $\overline{E_2E_3}$ because any two lines in a projective plane intersect in a point (Postulate 14.1.1.2). Similarly, $\overline{EE_2}$ and $\overline{EE_3}$ result in points B and C on $\overline{E_1E_3}$ and $\overline{E_1E_2}$, respectively.

We know that given three points on a line, we can construct another point so that the four points are a harmonic set. Since the harmonic conjugate is unique, we have

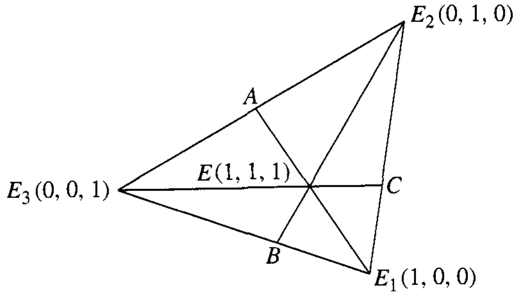

Figure 14.11 The reference triangle and the unit point.

a method, without a metric for either segments or angles, to generate points uniquely in a projective plane. It works like this. Start with three collinear points X, X_0, and X_1. If we choose X_2 such that $\mathsf{H}(XX_1, X_0X_2)$, X_2 will be between X_1 and X (Figure 14.12). Similarly, we may find X_3 between X_2 and X if $\mathsf{H}(XX_2, X_1X_3)$. We thus find a sequence

$$X_1,\ X_2,\ X_3,\ \ldots,\ X_n,\ \ldots$$

continuing indefinitely toward X provided that

$$\mathsf{H}(XX_{n-1},\ X_{n-2}X_n)$$

for $n > 2$. To find a point $X_{(m+n)/2}$ between a pair of points X_m and X_n, use

$$\mathsf{H}(XX_{(m+n)/2},\ X_mX_n).$$

If the process continues indefinitely, it will generate a harmonic net, which fills the line as the rational numbers fill $\mathbb{R}$. For this reason a harmonic net is often called a **net of rationality**.

Now that we know how to find points on a line in the projective plane, we define an analytic geometry so that we can write equations to identify the points and lines. For every point P in the plane, assign a set of ordered triples of the form

$$(kx_1,\ kx_2,\ kx_3),$$

where kx_1, kx_2, kx_3 are real numbers, not all zero. For example, the ordered triples $(1,\ 2,\ 3)$ and $(4,\ 8,\ 12)$ identify the same point. Because all nonzero multiples of $(x_1,\ x_2,\ x_3)$ represent the same point, we say that $(kx_1,\ kx_2,\ kx_3)$ is written using **homogeneous coordinates**. Lines in this system will be identified by equations of the form

$$a_1x_1 + a_2x_2 + a_3x_3 = 0,$$

where a_1, a_2, and a_3 are not all zero. This resembles the equation for a line using Cartesian coordinates except that the third term is not a constant.

To see how this works, let us find the equation for the line joining $(1, 2, 3)$ and $(2, 4, 1)$. To accomplish this, we must find real numbers a_1, a_2, and a_3 so that

$$\begin{aligned} a_1 + 2a_2 + 3a_3 &= 0, \\ 2a_1 + 4a_2 + a_3 &= 0. \end{aligned}$$

This is equivalent to forming the matrix

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 1 \end{bmatrix},$$

and putting it into row-reduced echelon form,

$$\begin{bmatrix} 1 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

From this we find that the coefficients for the equation of the line are

$$a_1 = -2k, a_2 = k, \text{and } a_3 = 0$$

for all nonzero k. We arbitrarily choose $k = -1$, so the equation for the line between $(1, 2, 3)$ and $(2, 4, 1)$ can be written as

$$2x_1 - x_2 = 0.$$

Any point (x_1, x_2, x_3) that satisfies this equation will lie on the line. Because of the principle of duality, this line should receive a name like the points. The natural choice is to use the coefficients, so call the line $[2, -1, 0]$. Because of the parameter k, each triple $[2k, -k, 0]$ names the same line for all possible values of $k \neq 0$. Evidently, the sides of the reference triangle are $[1, 0, 0]$, $[0, 1, 0]$, and $[0, 0, 1]$ [Exercise 3(b)].

Since every line has an equation, by duality every point also should have an equation. Take lines $[1, 2, 3]$ and $[2, 4, 1]$. The equations for these lines are

$$\begin{aligned} x_1 + 2x_2 + 3x_3 &= 0, \\ 2x_1 + 4x_2 + x_3 &= 0. \end{aligned}$$

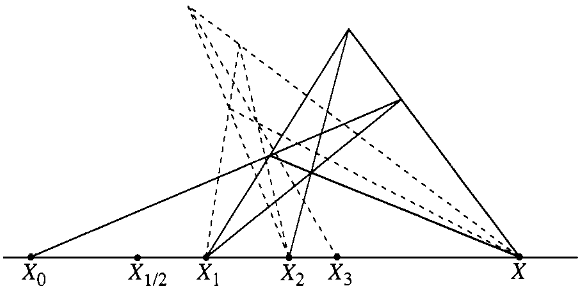

Figure 14.12 The points on a line form a harmonic net.

The solution to this linear system of equations is

$$x_1 = -2k, x_2 = k, \text{ and } x_3 = 0.$$

Therefore, the point of intersection can be taken to be $(2, -1, 0)$, and its equation is

$$2a_1 - a_2 = 0.$$

Any line $[a_1, a_2, a_3]$ that satisfies this equation will contain the point $(-2k, k, 0)$, provided that $k \neq 0$.

We may then combine these two concepts in the next definition.

■ DEFINITION 14.4.1

(x_1, x_2, x_3) is **incident** with $[a_1, a_2, a_3]$ or $[a_1, a_2, a_3]$ **contains** (x_1, x_2, x_3) if and only if

$$a_1x_1 + a_2x_2 + a_3x_3 = 0.$$

The definition makes clear why the ordered triple $(0, 0, 0)$ is excluded from our system of homogenous coordinates. Because $(0, 0, 0)$ satisfies the equation

$$a_1x_1 + a_2x_2 + a_3x_3 = 0$$

for all real numbers a_1, a_2, and a_3, the point that $(0, 0, 0)$ represents would be incident with every line. This is impossible in the projective plane (Exercise 1).

Incidence Results

Let $\mathcal{P} = \{P_1, P_2, \ldots, P_n\}$ be a set of points written using homogeneous coordinates. We define the set $\mathcal{P}$ to be **linearly independent** if it is linearly independent when viewed as a set of vectors. If $\mathcal{P}$ is not linearly independent, it is **linearly dependent**. Since $(x_1, x_2, \ldots, x_n)$ and $(kx_1, kx_2, \ldots, kx_n)$ with $k \neq 0$ represent the same point, two points in homogeneous coordinates are the same point if and only if they form a linearly dependent set. For three points, we have the following.

■ THEOREM 14.4.2

The set of points $\{(x_1, x_2, x_3), (y_1, y_2, y_3), (z_1, z_2, z_3)\}$ written in homogeneous coordinates is linearly dependent if and only if the points are collinear.

PROOF

Suppose that the three points are collinear. This means that there exists a line $[a_1, a_2, a_3]$ that is a solution to the equations for each of the three points:

$$\begin{array}{ccccccc} x_1a_1 & + & x_2a_2 & + & x_3a_3 & = & 0, \\ y_1a_1 & + & y_2a_2 & + & y_3a_3 & = & 0, \\ z_1a_1 & + & z_2a_2 & + & z_3a_3 & = & 0. \end{array}$$

The solution must be nontrivial in order for it to be a line, since $[0, 0, 0]$ does not represent a line. This can happen only if

$$\{(x_1, x_2, x_3), (y_1, y_2, y_3), (z_1, z_2, z_3)\}$$

is linearly dependent. Since these steps are reversible, we have proven the theorem. ■

A set of three vectors in $\mathbb{R}^3$ is linearly dependent if and only if the determinant of the matrix formed by the vectors of the set is zero.

■ COROLLARY 14.4.3

If (x_1, x_2, x_3), (y_1, y_2, y_3), (z_1, z_2, z_3) are in homogeneous coordinates,

$$\begin{vmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{vmatrix} = 0$$

if and only if the points are collinear.

Theorem 14.4.2 and its corollary also hold for lines in the projective plane, by the principle of duality.

Equivalence Classes

Let $\sim$ be an equivalence relation on $\mathcal{A}$ (page 51). If $a \in \mathcal{A}$, the **equivalence class** of a is defined as the set

$$[a] = \{x \in \mathcal{A} : x \sim a\}.$$

The set of all equivalence classes,

$$\mathcal{A}/\sim\, = \{[a] : a \in \mathcal{A}\},$$

is called a **quotient set**, and we read $\mathcal{A}/\sim$ as "A modulo tilde." It is always the case that

$$\mathcal{A} = \bigcup_{a \in \mathcal{A}} [a].$$

The equivalence relation associates elements of the set that are considered the same type of element with respect to the equivalence relation. This has the effect of dividing the set up into cells that do not overlap, yet exhaust the entire set.

■ DEFINITION 14.4.4

Let $\mathcal{A} \neq \varnothing$ be a set and $\mathcal{P}$ be a set of subsets of $\mathcal{A}$. The set $\mathcal{P}$ is a **partition** of $\mathcal{A}$ if and only if

- $\bigcup \mathcal{P} = \mathcal{A}$.
- For all $\mathcal{B}, \mathcal{C} \in \mathcal{P}$, either $\mathcal{B} = \mathcal{C}$ or $\mathcal{B} \cap \mathcal{C} = \varnothing$.

The second condition of the definition means that $\mathscr{P}$ is **pairwise disjoint**. To illustrate this definition, let $\mathscr{A} = \{1, 2, 3, 4, 5, 6, 7\}$ and define $\mathscr{A}_1 = \{1, 2, 5\}$, $\mathscr{A}_2 = \{3\}$, and $\mathscr{A}_3 = \{4, 6, 7\}$. The family $\mathscr{P} = \{\mathscr{A}_1, \mathscr{A}_2, \mathscr{A}_3\}$ is pairwise disjoint, and $\mathscr{A} = \mathscr{A}_1 \cup \mathscr{A}_2 \cup \mathscr{A}_3$. Hence, $\mathscr{P}$ is a partition of $\mathscr{A}$. Similarly, the quotient set of an arbitrary set under an equivalence relation will be a partition for the set.

■ THEOREM 14.4.5

If $\sim$ is an equivalence relation on $\mathscr{A}$, then $\mathscr{A}/{\sim}$ is a partition of $\mathscr{A}$.

PROOF

Take a set $\mathscr{A}$ with an equivalence relation $\sim$. We first prove that $\bigcup(\mathscr{A}/{\sim}) = \mathscr{A}$. Suppose that $x \in \bigcup(\mathscr{A}/{\sim})$. This means that there exists $a \in \mathscr{A}$ such that $x \in [a]$. Since $[a] \subseteq \mathscr{A}$, x is an element of $\mathscr{A}$. To prove the converse, let $x \in \mathscr{A}$. Because $x \sim x$, x is an element of $[x]$. This implies that $x \in \bigcup(\mathscr{A}/{\sim})$ since $[x] \in \mathscr{A}/{\sim}$.

To show that $\mathscr{A}/{\sim}$ is pairwise disjoint, let $[a], [b] \in \mathscr{A}/{\sim}$ and assume that $[a] \cap [b] \neq \varnothing$. We conclude that there exists $y \in [a] \cap [b]$. In other words, $y \sim a$ and $y \sim b$. Now take $x \in [a]$. This means that $x \sim a$. Since $a \sim y$ by symmetry, $x \sim y$ and then $x \sim b$, both by transitivity. Thus, $x \in [b]$, which shows that $[a] \subseteq [b]$. Similarly, $[b] \subseteq [a]$. Therefore, $[a] = [b]$. ■

We now define an equivalence relation on a set of homogeneous coordinates and examine the resulting partition. Let

$$\mathscr{H} = \{(x_1, x_2, x_3) : x_1, x_2, x_3 \in \mathbb{R} \text{ but not all zero}\}.$$

Since all multiples of a given ordered triple identify the same point, define

$$(x_1, x_2, x_3) \sim (y_1, y_2, y_3)$$

if and only if there exists $k \in \mathbb{R}$ such that

$$\begin{aligned} x_1 &= ky_1, \\ x_2 &= ky_2, \\ x_3 &= ky_3. \end{aligned}$$

To show that this is an equivalence relation, take (x_1, x_2, x_3), (y_1, y_2, y_3), and (z_1, z_2, z_3) in $\mathscr{H}$.

- $(x_1, x_2, x_3) \sim (x_1, x_2, x_3)$ because $x_1 = 1x_1$, $x_2 = 1x_2$, and $x_3 = 1x_3$.
- Assume that $(x_1, x_2, x_3) \sim (y_1, y_2, y_3)$. From this we conclude that there exists $k \in \mathbb{R}$ such that $x_1 = ky_1$, $x_2 = ky_2$, and $x_3 = ky_3$. Since not all of the coordinates are zero, k must be nonzero. Therefore,

  $$y_1 = (1/k)x_1,\ y_2 = (1/k)x_2, \text{ and } y_3 = (1/k)x_3,$$

  which means that $(y_1, y_2, y_3) \sim (x_1, x_2, x_3)$

- Let $(x_1, x_2, x_3) \sim (y_1, y_2, y_3)$ and $(y_1, y_2, y_3) \sim (z_1, z_2, z_3)$. Then there exists $j, k \in \mathbb{R}$ so that

  $$x_1 = jy_1, x_2 = jy_2, \text{and } x_3 = jy_3$$

  and

  $$y_1 = kz_1, y_2 = kz_2, \text{and } y_3 = kz_3.$$

  Hence,

  $$x_1 = (jk)z_1, x_2 = (jk)z_2, \text{and } x_3 = (jk)z_3.$$

  Therefore, $(x_1, x_2, x_3) \sim (z_1, z_2, z_3)$.

Since $\sim$ is an equivalence relation on $\mathcal{H}$, the quotient set

$$\mathbb{P}^2 = \mathcal{H}/\sim$$

is a partition of $\mathcal{H}$. If we take $(x_1, x_2, x_3) \in \mathcal{H}$, the points in $\mathcal{H}$ that are equivalent to it are exactly those ordered triples that are nonzero multiples of (x_1, x_2, x_3). We conclude that

$$[(x_1, x_2, x_3)] = \{(kx_1, kx_2, kx_3) : k \neq 0\}.$$

Although there is not a unique ordered triple for every point in the projective plane, there is a unique equivalence class. When we work with these classes, we typically choose a particular element of the class, called a **representative**, and label the point with it. The quotient set $\mathbb{P}^2$ is called the **real projective plane**. It is left to Exercise 2 to confirm that $\mathbb{P}^2$ satisfies the postulates of projective geometry.

Of course, everything just written about points holds also for lines, so we have developed a complete system to label the objects in a projective plane. Although homogeneous coordinates were created independently by several mathematicians, including August Möbius, the full system is due to the German mathematician Julius Plücker. He was born in Elberfeld in 1801 and lived for 67 years. He studied at various universities throughout Europe, earning his doctorate from the University of Marburg in 1824. His mathematical specialty was analytic geometry, which he then applied to the projective plane. Since most methods in projective geometry at the time were synthetic, as appears to be the nature of the two camps, controversy again arose over which of the two methods was best. Plücker sided with the analysts, a decision which at one point cost him a residency position at Berlin. Despite the controversy, Plücker became a top German mathematician with a reputation for elegant work. His other interests included physics, on which he had a lasting correspondence with Michael Faraday (Burau and Schoenberg 1970, 44, 46; Boyer and Merzbach 1991, 540–543).

Collineations

The functions that we have examined so far both synthetically and analytically in a projective plane have been functions between ranges (e.g., central perspectivities) or between pencils (e.g., axial perspectivities in Exercise 6 of Section 14.3). We now want to define a function whose domain is an entire projective plane but still resembles the other functions. We will do our work on $\mathbb{P}^2$, and the function will be a collineation (Definition 14.1.5).

Suppose that $\phi : \mathbb{P}^2 \longrightarrow \mathbb{P}^2$ is a collineation and write

$$\begin{aligned}\phi(1, 0, 0) &= (x_1, y_1, z_1),\\ \phi(0, 1, 0) &= (x_2, y_2, z_2),\\ \phi(0, 0, 1) &= (x_3, y_3, z_3),\\ \phi(1, 1, 1) &= (x_4, y_4, z_4).\end{aligned}$$

This function has the property that it maps all points on a line to all of the points of another line, so ϕ must be a linear transformation. Therefore, there exists a 3×3 matrix A such that

$$\phi(X) = AX,$$

where X is a point from $\mathbb{P}^2$. Since ϕ is a one-to-one function, the determinant of A is not equal to zero.

As an example, define $\phi : \mathbb{P}^2 \longrightarrow \mathbb{P}^2$ by

$$\begin{aligned}\phi(1, 0, 0) &= (1, -2, 3),\\ \phi(0, 1, 0) &= (5, 6, 5),\\ \phi(0, 0, 1) &= (1, 0, 4),\\ \phi(1, 1, 1) &= (3, 2, 1).\end{aligned}$$

Observe that

$$\phi(1, 1, 1) = \phi(1, 0, 0) + \frac{2}{3}\phi(0, 1, 0) - \frac{4}{3}\phi(0, 0, 1)$$

or

$$3\phi(1, 1, 1) = 3\phi(1, 0, 0) + 2\phi(0, 1, 0) - 4\phi(0, 0, 1).$$

Therefore,

$$3\begin{bmatrix} x_1' \\ x_2' \\ x_3' \end{bmatrix} = \begin{bmatrix} 3 & 10 & -4 \\ -6 & 12 & 0 \\ 9 & 10 & -16 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix},$$

which gives rise to this system of linear equations:

$$\begin{aligned} 3x_1' &= 3x_1 + 10x_2 - 4x_3,\\ 3x_2' &= -6x_1 + 12x_2,\\ 3x_3' &= 9x_1 + 10x_2 - 16x_3. \end{aligned}$$

We see that we can generalize this to the next proposition.

■ THEOREM 14.4.6

The collineation $\phi : \mathbb{P}^2 \to \mathbb{P}^2$, where $\phi(x_1, x_2, x_3) = (x_1', x_2', x_3')$, is defined by the equations

$$\begin{aligned} kx_1' &= a_{11}x_1 + a_{12}x_2 + a_{13}x_3, \\ kx_2' &= a_{21}x_1 + a_{22}x_2 + a_{23}x_3, \\ kx_3' &= a_{31}x_1 + a_{32}x_2 + a_{33}x_3 \end{aligned}$$

such that $k \neq 0$ and

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \neq 0.$$

Finally, let ϕ be a collineation on $\mathbb{P}^2$. Observe that if we take lines l and m in $\mathbb{P}^2$, their images $\phi[l]$ and $\phi[m]$ are also lines in $\mathbb{P}^2$. Let O be a point not on l or m, so $\phi(O)$ is not on $\phi[l]$ or on $\phi[m]$. If A is on m and B on m such that A, O, and B are collinear, $\phi(A)$ is on $\phi[l]$, $\phi(B)$ is on $\phi[m]$, and $\phi(A)$, $\phi(O)$, and $\phi(B)$ are collinear. All of this follows because ϕ is a collineation. Therefore, if

$$T : l \to m$$

is a central perspectivity,

$$\phi \circ T \circ \phi^{-1} : \phi[l] \to \phi[m]$$

is a central perspectivity. It is then reasonable to say that perspectivities are preserved under collineations. Similarly, since a projectivity is the composition of perspectivities, we can also say that projectivities are preserved under collineations. For this reason, a collineation on $\mathbb{P}^2$ is usually called a **projective collineation**.

Exercises

1. Explain why it is not possible that a point exists in the projective plane that is on every line.

2. Confirm that $\mathbb{P}^2$ is a projective plane.

3. Using Figure 14.11:
   (a) Find the coordinates of points A, B, and C.
   (b) Show that the coordinates for $\overline{E_2E_3}$, $\overline{E_1E_3}$, and $\overline{E_1E_2}$ are, respectively, $[1, 0, 0]$, $[0, 1, 0]$, and $[0, 0, 1]$.
   (c) Find the coordinates for $\overline{AE_1}$, $\overline{BE_2}$, and $\overline{CE_3}$.
   (d) Find two more points between B and E_1 and two more between E_3 and B.
   (e) Find the intersection of $\overline{AC}$ and $\overline{BE_2}$.

4. Find the homogeneous coordinates for the line between $(3, 0, 6)$ and $(0, 4, 1)$.

5. Find the equation for the intersection of $[6, -4, 2]$ and $[1, 1, 8]$.

6. Determine whether the following points are collinear.
   (a) $(3, 0, 0), (3, -4, 0), (1, 1, 1)$
   (b) $(9, 8, -5), (2, -8, 9), (2, -6, 9), (2, -8, 4)$

7. Find ordered triples (a_1, a_2, a_3), (b_1, b_2, b_3), and (c_1, c_2, c_3) in homogeneous coordinates that are multiples of the given points and such that

$$(a_1, a_2, a_3) = (b_1, b_2, b_3) + (c_1, c_2, c_3).$$

   (a) $(0, 0, 5), (1, 1, 1), (7, 7, 0)$
   (b) $(-2, 0, 1), (1, -1, 2), (4, -2, 3)$
   (c) $(1, 2, 3), (0, 1, 2), (-1, 0, 1)$

8. Using Figure 14.11:
   (a) Find the lines of the complete quadrangle with vertices E, C, E_1, and B.
   (b) Confirm that the diagonal points are not collinear.
   (c) Generalize this to show that the diagonal points of any complete quadrangle are not collinear.

9. For all nonzero real numbers a and b, let $a \sim b$ if and only if $ab > 0$.
   (a) Show that $\sim$ is an equivalence relation on $\mathbb{R}\backslash\{0\}$.
   (b) Find the equivalence classes $[1]$ and $[-3]$.

10. For all $a, b \in \mathbb{Z}$, let $a \sim b$ if and only if $|a| = |b|$.
   (a) Prove that $\sim$ is an equivalence relation on $\mathbb{Z}$.
   (b) Describe $\mathbb{Z}/\sim$.

11. For all $(a, b), (c, d) \in \mathbb{Z}\backslash\{0\} \times \mathbb{Z}\backslash\{0\}$, define $(a, b) \sim (c, d)$ if and only if $ad = bc$.
   (a) Show that $\sim$ is an equivalence relation on $\mathbb{Z}\backslash\{0\} \times \mathbb{Z}\backslash\{0\}$.
   (b) What is the equivalence class of $(1, 2)$?
   (c) Describe $(\mathbb{Z}\backslash\{0\} \times \mathbb{Z}\backslash\{0\})/\sim$.

12. Prove that the following are partitions of $\mathbb{R}^2$.
   (a) $\mathscr{E} = \{\{(a, b)\} : a, b \in \mathbb{R}\}$.
   (b) $\mathscr{F} = \{V_r : r \in \mathbb{R}\}$, where $V_r = \{(r, y) : y \in \mathbb{R}\}$ for all real numbers r.
   (c) $\mathscr{H} = \{H_n : n \in \mathbb{Z}\}$, where $H_n = \{\mathbb{R} \times (n, n+1] : n \in \mathbb{Z}\}$ for all $n \in \mathbb{Z}$.

13. Is $\{[n, n+1] \times (n, n+1) : n \in \mathbb{Z}\}$ a partition of $\mathbb{R}^2$? Explain.

14. Define the collineation $\phi : \mathbb{P}^2 \to \mathbb{P}^2$ by $\phi(X) = AX$, where

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 1 \\ 1 & 0 & 1 \end{bmatrix}.$$

   (a) Find $\phi(1, 2, 3)$ and $\phi[4, 5, 6]$.
   (b) Write the image of the unit point as a linear combination of the images of the vertices of the reference triangle.

(c) Find $\phi^{-1}(4, 0, 1)$.

(d) Find the intersection of $\phi^{-1}[-1, 4, 6]$ and $\phi^{-1}[0, 0, 2]$.

15. Let $\psi : \mathbb{P}^2 \to \mathbb{P}^2$ be a collineation such that $\psi(1, 0, 0) = (2, -2, 8)$, $\psi(0, 1, 0) = (2, 5, 1)$, $\psi(0, 0, 1) = (2, 2, 4)$, and $\psi(1, 1, 1) = (0, 1, -1)$.

(a) Find the matrix for ψ.

(b) Write ψ as a system of linear equations.

16. Let $\phi : \mathbb{P}^2 \to \mathbb{P}^2$ be a collineation such that $\phi(2, 1, 1) = (3, 1, -5)$, $\phi(2, -1, 1) = (1, 1, -3)$, $\phi(1, 2, 1) = (-1, 0, 2)$, $\phi(5, 2, 3) = (6, 4, -12)$.

(a) Find the matrix for $\psi_1 : \mathbb{P}^2 \to \mathbb{P}^2$ such that $\psi(1, 0, 0) = (2, 1, 1)$, $\psi(0, 1, 0) = (2, -1, 1)$, $\psi(0, 0, 1) = (1, 2, 1)$, $\psi(1, 1, 1) = (5, 2, 3)$.

(b) Find the matrix for $\psi_2 : \mathbb{P}^2 \to \mathbb{P}^2$ such that we have $\psi(1, 0, 0) = (3, 1, -5)$, $\psi(0, 1, 0) = (1, 1, -3)$, $\psi(0, 0, 1) = (-1, 0, 2)$, and $\psi(1, 1, 1) = (6, 4, -12)$.

(c) Use the matrices for ψ_1 and ψ_2 to find the matrix for ϕ.

(d) Write ϕ as a system of linear equations.

17. Let $E_1 = (1, 0, 0)$, $E_2 = (0, 1, 0)$, $E_3 = (0, 0, 1)$, and $E = (1, 1, 1)$. The point O at the intersection of $\overline{EE_3}$ and $\overline{E_1E_2}$ defines a central perspectivity between $\overline{E_1E_3}$ and $\overline{E_2E_3}$. After finding the homogeneous coordinates for A and B (Exercise 3), find an equation relating the homogeneous coordinates of A with those of B that must be satisfied in order for $A \overset{O}{\barwedge} B$ through O.

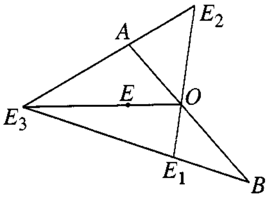

CHAPTER 15

FELIX KLEIN

After lecturing at Göttingen for a year, Klein joined the faculty at the University of Erlangen in 1872. The public lecture that earned him this position he titled *A Comparative Review of Recent Researches in Geometry* (Klein 1893). In it he introduced a new means of classifying geometries that became known as the **Erlangen Program**. It characterized a geometry based on the properties that were preserved under certain sets of functions. This is similar to what Poncelet had done in his *Traite*. There Poncelet looked for properties that were preserved under projectivities. One such property was that of being a harmonic set. Klein's scheme would begin with the projective transformations, and then by placing restrictions on these functions, find other collections that would yield their own types of invariants on other geometries. Each of the sets of functions were required to satisfy certain conditions, making them into what is called a group. Klein had learned of this algebra concept from Camille Jordan when he traveled to Paris with Sophus Lie (Yaglom 1988, 128–129; Burau and Schoenberg 1970, 396–397; Wussing 1984, 27).

Lie was from Norway and was seven years older than Klein. Lie was an interesting combination of outdoorsman and overall success in all of the academic subjects. When it came time for college, he had his choice of fields. Lie's father was a Lutheran pastor, so Lie considered studying theology, but decided instead on mathematics and

Revolutions of Geometry. By Michael O'Leary

Copyright © 2010 John Wiley & Sons, Inc.

the sciences at Christiania University. It was there where he read the works of Poncelet and Plücker. These two mathematicians had a lasting impact on Lie, and their work would inspire Lie's long research program on geometric transformations. Lie went to Berlin in 1870 to continue his mathematical education. It was in Berlin where Klein and Lie first met. They would become lifelong friends and work on many projects together (Yaglom 1988, 22; Burau and Schoenberg 1970, 396).

One of their first joint undertakings was to go to Paris and learn about current French mathematics. They arrived in late 1870. In Paris Klein and Lie studied geometry and learned about group theory. While there it was their intention to meet Camille Jordan, who had recently published *Commentaire sur le mémorie de Galois* in 1865 and *Commentaire sur Galois* in 1869. These were expositions on the work of Evariste Galois, a gifted mathematician who began a study of the subject when he was 15, developed an approach to deal with equations solvable by radicals, spent time in jail as a political prisoner, and was killed in a dual at the age of 21. Klein was particularly impressed with Jordan's *Traité des substitutions et des équations algébriques*. Inspired by what they learned, Klein and Lie coauthored two papers, one published in June 1870 and the other in March 1871. These would become the preliminary work for Klein's Erlangen lecture (Wussing 1984, 135–136, 180–184; Yaglom 1988, 22; Kline 1972, 755–756).

Klein and Lie left Berlin in July 1870, at the start of the Franco-Prussian War. Being German, Klein needed to leave the country. When he arrived home, he planned to join the Prussian army, but he caught typhus. When he recovered, Klein returned to Göttingen. All the while Lie remained in France to hike the French countryside, see the Alps, and then to see Italy. Although it was only for a month, he did spend some time in a French prison under suspicion of spying for Germany (Wussing 1984, 179; Yaglom 1988, 24, 27).

15.1 GROUP THEORY

So what is a group? Start with the set of integers $\mathbb{Z}$. If we add any two integers together, the result will be an integer, and this result will be unique to the original two integers. For example, the sum of 5 and 3 is 8 and always will be 8. The same thing goes for subtraction and multiplication. Division is a problem, however. We cannot divide by zero, and an integer divided by an integer is not guaranteed to be an integer. On the set of integers, addition, subtraction, and multiplication are examples of binary operations, whereas division is not.

■ **DEFINITION 15.1.1**

A **binary operation** on a set $\mathcal{S}$ is a function $* : \mathcal{S} \times \mathcal{S} \to \mathcal{S}$.

Since a binary operation is a generalization of the more concrete operations of addition and multiplication, we write

$$a * b$$

instead of $*(a, b)$. The following are examples of binary operations.

- $a * b = |ab|$ on $\mathbb{R}$.
- Addition, subtraction, and multiplication on $\mathbb{R}$.
- Division on $(-\infty, 0) \cup (0, \infty)$.
- $a * b = \sqrt{ab}$ on $[0, \infty)$.

To show that a rule on a set $\mathcal{S}$ is a binary operation, we must show that is satisfies two properties. First, it must be well defined as a function, and second, it must be true that $a * b \in \mathcal{S}$ for all $a, b \in \mathcal{S}$. When this second condition is satisfied, we say that $\mathcal{S}$ is **closed under** $*$.

Consider the following example. Denoting the set of positive integers by $\mathbb{Z}^+$, for all $n \in \mathbb{Z}^+$ and $a \in \mathbb{Z}$, define

$$[a]_n = \{a + kn : k \in \mathbb{Z}\}.$$

This is called the **congruence class** of a **modulo** n (see Exercise 12 from Section 2.3 and also page 511). For instance,

$$\begin{aligned}
[0]_4 &= \{\ldots, -4, 0, 4, 8, \ldots\},\\
[1]_4 &= \{\ldots, -3, 1, 5, 9, \ldots\},\\
[2]_4 &= \{\ldots, -2, 2, 6, 10, \ldots\},\\
[3]_4 &= \{\ldots, -1, 3, 7, 11, \ldots\}.
\end{aligned}$$

The collection cycles because $[4]_4 = [0]_4$. We now define

$$\mathbb{Z}_n = \{[a]_n : a \in \mathbb{Z}\},$$

and on this set specify $+$ by

$$[a]_n + [b]_n = [a + b]_n.$$

Be careful to note that the symbol $+$ is being used in two ways. The $+$ on the left is the new definition, but the $+$ on the right is standard addition.

We first show that this addition is well defined. This is particularly important to prove this in this case since each congruence class can be represented using infinitely many different integers. Take $[a]_n, [b]_n, [a']_n, [b']_n \in \mathbb{Z}_n$ and suppose that

$$[a]_n = [a']_n \text{ and } [b]_n = [b']_n.$$

This does *not* mean that $a = a'$ and $b = b'$. Instead, we conclude that

$$a = a' + jn \text{ and } b = b' + kn$$

for some $j, k \in \mathbb{Z}$. To show that the result of operating these two equal pairs of elements together yields the same answer, we must show that

$$[a + b]_n = [a' + b']_n.$$

To do this, let $x \in [a+b]_n$. This means that $x = a + b + ln$ for some $l \in \mathbb{Z}$. Therefore,

$$x = a' + jn + b' + kn + ln = a' + b' + (j + k + l)n,$$

from which follows that $x \in [a' + b']_n$ because $j + k + l \in \mathbb{Z}$. Hence,

$$[a + b]_n \subseteq [a' + b']_n.$$

The other inclusion follows similarly.

It is a simple matter to see that $\mathbb{Z}_n$ is closed under $+$. Let $[a]_n$, $[b]_n \in \mathbb{Z}_n$. This means that a and b are integers. Since we know that $\mathbb{Z}$ is closed under standard addition, $a + b \in \mathbb{Z}$. Therefore, $[a]_n + [b]_n = [a + b]_n \in \mathbb{Z}_n$.

Groups

Now consider the properties of addition as viewed as a binary operation on the set of integers. We know that addition is commutative and associative (page 32). There is a particular element of $\mathbb{Z}$ that is denoted by 0 that has the property

$$\text{for all } a \in \mathbb{Z}, 0 + a = a + 0 = a.$$

The number 0 is the additive identity of $\mathbb{Z}$. Finally, every integer a has an additive inverse. This inverse is denoted by $-a$ and has the property that

$$a + (-a) = (-a) + a = 0.$$

In other words, an integer added to its inverse yields the identity. Ignoring the commutative condition for the moment, we make our definition.

■ **DEFINITION 15.1.2**

Let $\mathscr{G}$ be a set and $*$ a binary operation with the following properties:

- $a * (b * c) = (a * b) * c$ for all $a, b, c \in \mathscr{G}$.
- There exists $e \in \mathscr{G}$ such that $e * a = a * e = a$ for all $a \in \mathscr{G}$. The element e is called the **identity** of $\mathscr{G}$.
- For all $a \in \mathscr{G}$, there exists $a' \in \mathscr{G}$ such that $a * a' = a' * a = e$. The element a' is called the **inverse** of a' in $\mathscr{G}$.

The pair $(\mathscr{G}, *)$ is called a **group**. When the group's binary operation is clear from context, the set $\mathscr{G}$ can be referred to as the group.

The term *group* was first used by Galois (Wussing 1984, 17). A group with a commutative binary operation (that $a * b = b * a$ for all a and b in the group) is known as an **abelian group.** It is named after the Norwegian mathematician Niels Henrik Abel.

$+$	$[0]$	$[1]$	$[2]$	$[3]$
$[0]$	$[0]$	$[1]$	$[2]$	$[3]$
$[1]$	$[1]$	$[2]$	$[3]$	$[0]$
$[2]$	$[2]$	$[3]$	$[0]$	$[1]$
$[3]$	$[3]$	$[0]$	$[1]$	$[2]$

(a) $\mathbb{Z}_4$

$*$	e	a	b	c
e	e	a	b	c
a	a	e	c	b
b	b	c	e	a
c	c	b	a	e

(b) Klein-4 group

Figure 15.1 The groups of order four.

The pairs $(\mathbb{Z}^+, +)$, $(\mathbb{Z}, \cdot)$, and $(\mathbb{R}, \cdot)$ are not groups, but $(\mathbb{Z}, +)$, $(\mathbb{Q}, +)$, $(\mathbb{R}, +)$, $(\mathbb{Q}\backslash\{0\}, \cdot)$, and $(\mathbb{R}\backslash\{0\}, \cdot)$ are. These are examples of infinite groups. When the group's set is finite, we use the term **order** to refer to its cardinality. The group $(\{e\}, *)$, where $*$ is any binary operation, is the only group of order 1. This means that any other group with one element, such as $(\{0\}, +)$ or $(\{1\}, \cdot)$, has the same structure as $(\{e\}, *)$. We say that these three groups are **isomorphic**. Any group of order 2 will be isomorphic to $\mathbb{Z}_2$, and any group of order 3 will be isomorphic to $\mathbb{Z}_3$. There are essentially two groups of order 4: those that are isomorphic to $\mathbb{Z}_4$ and those that are isomorphic to the **Klein-4** group. Their tables are given in Figure 15.1.

All of the examples given so far have been abelian. Define $\mathrm{GL}(n, \mathbb{R})$ to be the pair consisting of the set of all $n \times n$ invertible matrices with real entries and matrix multiplication. This is called the **general linear group**. We know that matrix multiplication is associative, that its identity is the identity matrix, which is invertible, and that every matrix in the set has an inverse by definition. However, we know that even for invertible matrices, multiplication is not commutative. Therefore, $\mathrm{GL}(n, \mathbb{R})$ is a nonabelian group.

The group definition states only that in a group there is an identity and there are inverses. Based on what we know about the integers, we should be able to prove more about these elements.

■ THEOREM 15.1.3

In a group there is exactly one identity and every element has a unique inverse.

PROOF

The uniqueness of the identity is left to Exercise 8. To show the uniqueness of inverses, let $\mathcal{G}$ be a group, and take $a \in \mathcal{G}$. Suppose that both a' and a'' are inverses of a in $\mathcal{G}$. Then

$$a' = a' * e = a' * (a * a'') = (a' * a) * a'' = e * a'' = a''. \blacksquare$$

The purpose behind a group is that it is the most basic algebraic structure in which simple equations can be solved. For example, to solve $x + 4 = 10$, we need the additive inverse of 4 to add to both sides of the equation. After this we need to be able to regroup so that we may add the inverse to 4 to obtain the identity. As a consequence of these three properties, we will discover that the solution found is the unique one for the equation. All of this is summarized in the next theorem.

■ **THEOREM 15.1.4 [Cancellation Law]**

Let $\mathcal{G}$ be a group and a' be the inverse of a. For all a, $b \in \mathcal{G}$:

- The unique solution to $a * x = b$ is $x = a' * b$.
- The unique solution to $x * a = b$ is $x = b * a'$.

PROOF

That $a' * b$ is a solution of $a * x = b$ simply requires a check:

$$a * (a' * b) = (a * a') * b = e * b = b.$$

To see that the solution is unique, suppose that both u and v are solutions to the equation. Again we calculate:

$$\begin{aligned} a * u &= a * v, \\ a' * (a * u) &= a' * (a * v), \\ (a' * a) * u &= (a' * a) * v, \\ e * u &= e * v, \\ u &= v. \end{aligned}$$

Therefore, $a' * b$ is the only solution to $a * x = b$. That the other equation has a unique solution as indicated is proven similarly. ■

Subgroups

Let a be an element of a group $\mathcal{G}$. For all positive integers n, define a^n to be the result of operating a with itself n times. That is,

$$a^1 = a,\ a^2 = a * a,\ a^3 = a * a * a,\ \ldots$$

and

$$a^m * a^n = a^{m+n}.$$

Further define $a^0 = e$ and a^{-1} to be the inverse of a. With this notation we observe that

$$(a * b)^{-1} = b^{-1} * a^{-1}$$

and

$$a^{-n} = (a^n)^{-1} = (a^{-1})^n.$$

We then gather all of these elements into a set,

$$\langle a \rangle = \{a^n : n \in \mathbb{Z}\},$$

and define the following.

■ DEFINITION 15.1.5

A group $\mathcal{G}$ is **cyclic** if there exists $a \in \mathcal{G}$ such that $\mathcal{G} = \langle a \rangle$. The element a is called a **generator** of $\mathcal{G}$.

For example, $(\mathbb{Z}, +)$ is a cyclic group. Both 1 and -1 are generators. However, $\mathbb{Q}$ and $\mathbb{R}$ are not cyclic groups under addition. As for finite groups, each $\mathbb{Z}_n$ is cyclic, generated by $[1]_n$, but the Klein-4 group is not cyclic because $a^2 = e$ for all a in the group.

An element a of a group may not generate the entire group, but since $e \in \langle a \rangle$ and both a^n and a^{-n} are elements of $\langle a \rangle$, the set generated by a is a group using the operation from $\mathcal{G}$.

■ DEFINITION 15.1.6

Let $\mathcal{G}$ be a group with binary operation $*$ and $\mathcal{H} \subseteq \mathcal{G}$. We say that $\mathcal{H}$ is a **subgroup** of $\mathcal{G}$ if $(\mathcal{H}, *)$ is a group.

Every group with at least two elements has at least two subgroups, itself (the **improper subgroup**) and $\{e\}$ (the **trivial subgroup**). A group that has as most these two subgroups is called **simple**. For example, $\mathbb{Z}_2$ and $\mathbb{Z}_3$ are simple groups, but $\mathbb{Z}_4$ is not. It has $\{[0]_4, [2]_4\}$ as a subgroup. Other examples of nonsimple groups are $(\mathbb{R}, +)$ because $(\mathbb{Z}, +)$ is one of its subgroups and $(\langle 2 \rangle, +)$ because $(\langle 6 \rangle, +)$ is one of its subgroups.

Since every element of a cyclic group can be written as a generator to a power, we can write equations for each of the elements of a cyclic group. This simplifies our work when we deal with these types of groups. The cyclic groups then pass this property to their subgroups. To prove this we need the following proposition, known as the **Division Algorithm**, that was presupposed by Euclid in his method for finding the greatest common divisor of two positive integers (Euclid 1925, VII.2). We omit the proof of the Division Algorithm, but it can be found in any introductory number theory text such as Rosen (1993, 37–38).

■ THEOREM 15.1.7 [Division Algorithm]

If $m > n > 0$ are integers, there exists unique integers q and r such that $0 \leqslant r < n$ and

$$m = nq + r.$$

The integer q is called the **quotient** and r is called the **remainder**. We need the theorem to prove the following.

■ THEOREM 15.1.8

A subgroup of a cyclic group is cyclic.

PROOF

Let $\mathcal{G}$ be a cyclic group with generator a. Let $\mathcal{H}$ be a subgroup of $\mathcal{G}$. If $\mathcal{H} = \{e\}$, the subgroup is cyclic with generator e. So suppose that $\mathcal{H}$ is not the trivial subgroup. Because $\mathcal{G}$ is cyclic, there exists a minimum positive integer n such that $a^n \in \mathcal{H}$. Suppose that this element is not a generator of $\mathcal{H}$. This means that there exists an integer $m > n$ such that $a^m \in \mathcal{H}$ but $a^m \notin \langle a^n \rangle$. This combined with the Division Algorithm yields unique integers q and r such that $m = nq + r$ and $0 < r < n$. Therefore,

$$a^m = a^{nq+r} = a^{nq} * a^r,$$

and from this we conclude that

$$a^r = a^m * a^{-nq}.$$

Since both a^m and a^{-nq} are elements of $\mathcal{H}$, a^r is an element of $\mathcal{H}$. This contradicts the minimality of n because $r < n$. Thus, a^n is a generator of $\mathcal{H}$. ■

To show that a set is a subgroup of a group, we do not need to show directly that the set satisfies the three group properties. We may check a different set of conditions instead.

■ THEOREM 15.1.9

Let $(\mathcal{G}, *)$ be a group with identity e. A subset $\mathcal{H}$ of $\mathcal{G}$ is a subgroup of $\mathcal{G}$ if and only if:

- $\mathcal{H}$ is closed under $*$.
- $e \in \mathcal{H}$.
- $a^{-1} \in \mathcal{H}$ for all $a \in \mathcal{H}$.

PROOF

Suppose that the three hypotheses of the theorem hold. We conclude the following:

- $*$ is a binary operation on $\mathcal{H}$ because $\mathcal{H}$ is closed under $*$.
- Since $*$ is associative on $\mathcal{G}$, it must be associative on $\mathcal{H}$.
- $e \in \mathcal{H}$, and every element of $\mathcal{H}$ has its inverse in $\mathcal{H}$.

Therefore, $(\mathcal{H}, *)$ is a group. Conversely, if $\mathcal{H}$ is a subgroup of $\mathcal{G}$, the three conditions clearly hold. ■

To illustrate the theorem, take a group $\mathcal{G}$ and a family of subgroups $\{\mathcal{H}_i : i \in I\}$. We claim that $\bigcap_{i \in I} \mathcal{H}_i$ is a subgroup of $\mathcal{G}$.

- Let $a, b \in \bigcap_{i \in I} \mathcal{H}_i$. This means that $a, b \in \mathcal{H}_i$ for all $i \in I$. Since each $\mathcal{H}_i$ is closed under the operation of $\mathcal{G}$, $a * b \in \mathcal{H}_i$ for all i. Hence,

$$a * b \in \bigcap_{i \in I} \mathcal{H}_i.$$

- Since e in each $\mathcal{H}_i$ for every $i \in I$, we must have $e \in \bigcap_{i \in I} \mathcal{H}_i$.
- Take $a \in \bigcap_{i \in I} \mathcal{H}_i$. Then because $a \in \mathcal{H}_i$ for all $i \in I$, we have $a^{-1} \in \mathcal{H}_i$, so $a^{-1} \in \bigcap_{i \in I} \mathcal{H}_i$.

We do not obtain the same result for union (Exercise 15).

Exercises

1. Which of the following are binary operations on the given sets?
   (a) Subtraction on $[0, \infty)$
   (b) Multiplication on $\mathbb{R}$
   (c) Division on $\mathbb{Z}^+$
   (d) $a * b = a^b$ on $\mathbb{Z}$
   (e) $a * b = \max\{a, b\}$ on $\mathbb{Z}$
   (f) $a * b = \sqrt{ab}$ on $\mathbb{R}$

2. What assumptions about $\mathbb{Z}$ must be made to show that $x * y = 2x - y$ is a binary operation on $\mathbb{Z}$?

3. Let $\mathcal{S}$ be a set and $P(\mathcal{S})$ the set of subsets of $\mathcal{S}$. The family of sets $P(\mathcal{S})$ is called the **power set** of $\mathcal{S}$.
   (a) Define $*$ such that for all $\mathcal{A}, \mathcal{B} \in \mathcal{S}$, $\mathcal{A} * \mathcal{B} = \mathcal{A} \cap \mathcal{B}$. Prove that $*$ is a binary operation on $\mathcal{S}$.
   (b) Is $(\mathcal{S}, *)$ a group? Explain.

4. Explain why $(\mathbb{Z}^+, +)$, $(\mathbb{Z}, \cdot)$, and $(\mathbb{R}, \cdot)$ are not groups.

5. Suppose that $*$ is defined by $x * y = x + y + 2$.
   (a) Identify the identity and the inverses for this group.
   (b) Prove that $(\mathbb{Z}, *)$ is a group.
   (c) Solve $8 * x = 10$ in $(\mathbb{Z}, *)$.

6. Let $\mathcal{F}$ be the set of **real-valued** functions. That is, each element of $\mathcal{F}$ is a function $f : \mathbb{R} \to \mathbb{R}$.
   (a) Prove that $(\mathcal{F}, +)$ is a group where

$$(f + g)(x) = f(x) + g(x).$$

   (b) Is $(\mathcal{F}, /)$ a group where $/$ is function division?
   (c) Is $(\mathcal{F}, \circ)$ a group where $\circ$ is composition of functions? Explain.

7. Let n be a positive integer and define

$$M_n = \left\{ \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{bmatrix} : a, b, c, d \in \mathbb{R} \right\}.$$

(a) Show that M_n under addition is a group.
(b) Solve the equation in M_2:

$$\begin{bmatrix} 1 & 4 \\ -3 & 0 \end{bmatrix} + \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} -3 & 8 \\ 0 & -6 \end{bmatrix}.$$

(c) Show that M_n under multiplication is not a group.

8. Prove that the identity of a group is unique.

9. Let $\mathcal{G}$ be a group. Define $\phi : \mathcal{G} \to \mathcal{G}$ by $\phi(a) = a^{-1}$. Prove that ϕ is a one-to-one correspondence.

10. Let $(\mathcal{G}, *)$ and $(\mathcal{G}', *')$ be two groups. Define the **Cartesian product** of $\mathcal{G}$ and $\mathcal{G}'$ as

$$\mathcal{G} \times \mathcal{G}' = \{(a, a') : a \in \mathcal{G} \text{ and } a' \in \mathcal{G}'\}.$$

Define $(a, a') \cdot (b, b') = (a * b, a' *' b')$.

(a) Confirm that $\cdot$ is a binary operation on $\mathcal{G} \times \mathcal{G}'$.
(b) Show that $(\mathcal{G} \times \mathcal{G}', \cdot)$ is a group. Prove that it is abelian if and only if both $\mathcal{G}$ and $\mathcal{G}'$ are abelian.

11. Let $\mathcal{G}$ be a group so that $(a * b)^2 = a^2 * b^2$ for all $a, b \in \mathcal{G}$. Show that $\mathcal{G}$ is abelian.

12. Identify all of the subgroups of the following groups.
(a) the Klein-4 group
(b) $\mathbb{Z}_5$
(c) $\mathbb{Z}_8$
(d) $\mathbb{Z}_2 \times \mathbb{Z}_3$
(e) $\mathbb{Z}_2 \times \mathbb{Z}_6$

13. Let $\mathcal{F}$ be the set of real-valued functions (Exercise 6). Show that

$$\{f \in \mathcal{F} : f(0) = 0\}$$

is a subgroup of $(\mathcal{F}, +)$.

14. Define $\mathcal{H} = \{A \in M_n : a_{1,1} + a_{2,2} + \cdots + a_{n,n} = 0\}$. Prove that $\mathcal{H}$ is a subgroup of $(M_n, +)$.

15. Let $\mathcal{G}$ be a group and $\{\mathcal{H}_i : i \in I\}$ a family of subgroups of $\mathcal{G}$.
(a) Give an example to show that $\bigcup_{i \in I} \mathcal{H}_i$ is not necessarily a subgroup of $\mathcal{G}$.
(b) Assume that for all $i, j \in I$, either $\mathcal{H}_i \subseteq \mathcal{H}_j$ or $\mathcal{H}_j \subseteq \mathcal{H}_i$. Show that $\bigcup_{i \in I} \mathcal{H}_i$ is a subgroup of $\mathcal{G}$.

16. Let $\mathcal{H}$ and $\mathcal{K}$ be two subgroups of an abelian group $\mathcal{G}$. Define

$$\mathcal{H}\mathcal{K} = \{a * b : a \in \mathcal{H} \text{ and } b \in \mathcal{K}\}.$$

Prove that $\mathcal{H}\mathcal{K}$ is a subgroup of $\mathcal{G}$. Is it abelian?

17. For any group $\mathscr{G}$, let $\mathscr{S} \subseteq \mathscr{G}$. Prove that

$$\mathscr{H} = \{a \in \mathscr{G} : a * b = b * a \text{ for all } b \in \mathscr{S}\}$$

is a subgroup of $\mathscr{G}$.

18. Let $s \in \mathscr{G}$. Define $\mathscr{G}_s = \{a \in \mathscr{G} : a * s = s\}$. Show that this set, called the **stabilizer** of s, is a subgroup of $\mathscr{G}$.

19. Let $\mathscr{H}$ and $\mathscr{H}'$ be subgroups of groups $\mathscr{G}$ and $\mathscr{G}'$, respectively. Show that $\mathscr{H} \times \mathscr{H}'$ is a subgroup of $\mathscr{G} \times \mathscr{G}'$ under the operation defined in Exercise 10.

20. Write four elements of each of the following cyclic subgroups.
   (a) $\langle 1/2 \rangle$ in $(\mathbb{Q}, +)$
   (b) $\langle 1/2 \rangle$ in $(\mathbb{R}\backslash\{0\}, \cdot)$
   (c) $\langle [3] \rangle$ in $\mathbb{Z}_8$
   (d) $\langle x^2 + 1 \rangle$ the group of real-valued functions under addition

21. Take a to be an element of a group with binary operation $*$. Let k be a positive integer. Prove that $\{a^{nk} : n \in \mathbb{Z}\}$ is a subgroup of $\langle a \rangle$.

22. Demonstrate that simple groups are cyclic.

15.2 TRANSFORMATION GROUPS

In this section we introduce the type of group that was central to Klein's Erlangen Program. We begin with an arbitrary set $\mathscr{S}$ and define a **transformation** of $\mathscr{S}$ to be a function $\phi : \mathscr{S} \to \mathscr{S}$ that is a one-to-one correspondence. Let Trans($\mathscr{S}$) be the set of all transformations of $\mathscr{S}$. We make four observations.

- If $\phi, \psi \in$ Trans($\mathscr{S}$), then $\phi \circ \psi \in$ Trans($\mathscr{S}$) since the composition of two one-to-one correspondences on $\mathscr{S}$ is a one-to-one correspondence on $\mathscr{S}$.
- Composition is associative, namely for all $\phi, \psi, \sigma \in$ Trans($\mathscr{S}$),
$$(\phi \circ \psi) \circ \sigma = \phi \circ (\psi \circ \sigma).$$
- The identity map is a one-to-one correspondence, so it is an element of Trans($\mathscr{S}$).
- If $\phi \in$ Trans($\mathscr{S}$), then $\phi^{-1} \in$ Trans($\mathscr{S}$) because the inverse of any one-to-one correspondence on $\mathscr{S}$ is a one-to-one correspondence on $\mathscr{S}$.

Therefore, (Trans($\mathscr{S}$), $\circ$) is a group. We will call any subgroup of Trans($\mathscr{S}$) a **transformation group**. An example is the set of collineations on $\mathbb{P}^2$. They form a subgroup **PR** of Trans($\mathbb{P}^2$) (Exercise 10). Harmonic sets are preserved by every element of **PR**, so we say that harmonic sets are an **invariant** of **PR**. On the other hand, distance and angle measure are not invariants for **PR**.

Klein used **PR** to categorize projective geometry, and he planned to use other groups, including subgroups of **PR**, to characterize other geometries by identifying their invariants under a given group. In this section we examine one such example. The texts of Bennett (1995, 18–33) and Eves (1965, 149–162), serve as general sources.

Affine Planes

Remember that Desargues defined the projective plane from the Euclidean plane by adding a line and points at infinity (Section 9.2). Let us now reverse that process and recoup the plane from $\mathbb{P}^2$. Let $\triangle E_1E_2E_3$ be the reference triangle (Figure 15.2). Let the ideal line ω be the line with equation $x_3 = 0$ and let all points on $x_3 = 0$ be the ideal points. These will be removed from the space but still used to identify parallel lines. Define lines l and m to be **parallel** if and only if l and m are different lines and there exists a point Ω on ω such that l and m both contain Ω. In Figure 15.2, both $\overline{AE_1}$ and $\overline{BE_1}$ are interpreted as being parallel to $\overline{E_1E_3}$.

When we remove the points with $x_3 = 0$, the remaining ordered triples in homogenous coordinates are of the form (kx_1, kx_2, kx_3), with both k and x_3 nonzero. However, since $x_3 \neq 0$, any of these points can be written as

$$(x_1/x_3, x_2/x_3, 1).$$

This means that we can uniquely identify each of these points by an ordered pair (x, y), where $x = x_1/x_3$ and $y = x_2/x_3$. These are called **nonhomogeneous coordinates**. Similarly, if we take a line that is represented using homogeneous coordinates $[a, b, c]$, its equation in homogeneous coordinates is

$$ax_1 + bx_2 + cx_3 = 0,$$

and letting $x = x_1/x_3$ and $y = x_2/x_3$, its equation in nonhomogeneous coordinates is

$$ax + by + c = 0. \tag{15.1}$$

Notice that if this line is parallel to

$$dx + ey + f = 0,$$

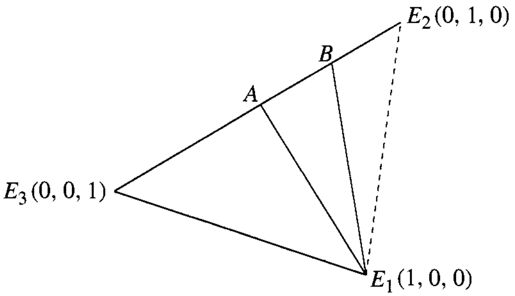

Figure 15.2 Interpreting $\overline{AE_1}$ and $\overline{AE_2}$ as parallel to $\overline{E_1E_2}$.

then

$$\begin{vmatrix} a & b \\ d & e \end{vmatrix} = 0.$$

If we convert these back to homogeneous coordinates, we obtain the system

$$\begin{array}{ccccccc} ax_1 & + & bx_2 & + & cx_3 & = & 0 \\ dx_1 & + & ex_2 & + & fx_3 & = & 0, \end{array}$$

which when put into a matrix and transform into row-reduced echelon form will look like either

$$\begin{bmatrix} 1 & * & * \\ 0 & 0 & 0 \end{bmatrix} \text{ or } \begin{bmatrix} 1 & * & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The first form tells us that the lines are the same, while the second means that the point in common lies on the line $x_3 = 0$. This confirms that in our new geometry parallel lines meet at the line at infinity.

We now define new planes using nonhomogeneous coordinates. Let $(\mathcal{P}, \mathcal{L})$ be a projective plane with its points and lines identified by homogeneous coordinates. Define

$$\mathbb{A}(\mathcal{P}) = \{(x, y) : (x, y, 1) \text{ represents a point in } \mathcal{P}\}$$

and

$$\mathbb{L}(\mathcal{L}) = \{\{(x, y) : ax + by + c = 0\} : [a, b, c] \text{ represents an ordinary line in } \mathcal{L}\}.$$

Notice that $[0, 0, 1]$ is not used to define $\mathbb{L}(\mathcal{L})$. When we examine the basic properties of these two sets, we find that they satisfy the following list of postulates.

■ POSTULATES 15.2.1

Let $\mathcal{P}_A$ be a set of objects that are called **points** and $\mathcal{L}_A$ be a set of objects that are called **lines**.

1. For all distinct $P, Q \in \mathcal{P}_A$, there exists a unique line l in $\mathcal{L}_A$ such that P and Q are incident with l.
2. Every line contains at least two points.
3. There exists at least three noncollinear points.
4. Given a line l and a point P not on l, there exists exactly one line through P that is parallel to l.

We call the pair $(\mathcal{P}_A, \mathcal{L}_A)$ an **affine plane** if it satisfies the four postulates. It is left to Exercise 1 to show that $(\mathbb{A}(\mathcal{P}), \mathbb{L}(\mathcal{L}))$ is an affine plane for all projective planes $(\mathcal{P}, \mathcal{L})$. Other examples of affine planes include $\mathbb{R}^2$ and $\mathbb{Q}^2$, where the points are ordered pairs and the lines are sets of points that satisfy equations of the form given in Equation 15.1. In fact, $\mathbb{R}^2$ is the affine plane obtained from $\mathbb{P}^2$. A finite example is

$$\mathcal{P}_A = \{A, B, C, D\}$$

and

$$\mathcal{L}_A = \{\{A, B\}, \{A, C\}, \{A, D\}, \{B, C\}, \{B, D\}, \{C, D\}\}.$$

This affine plane is obtained from the Fano projective plane (Exercise 4). That this is the smallest affine plane is the result of the next theorem.

■ THEOREM 15.2.2

An affine plane has at least four points and at least six lines.

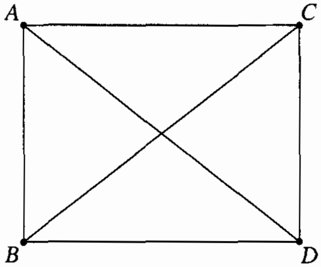

PROOF

By Postulate 3 of 15.2.1, there exists at least three noncollinear points in an affine plane. Let them be labeled A, B, and C. Join $\overline{AB}$, $\overline{AC}$, and $\overline{BC}$. These three lines are distinct. Since C is not on $\overline{AB}$, Postulate 4 guarantees a line on C that is parallel to $\overline{AB}$, and by Postulate 2 this line must contain a new point D. By the first postulate, join $\overline{BD}$ and $\overline{AD}$. The resulting diagram satisfies the postulates for an affine plane. ■

Notice that affine geometry feels a bit like projective geometry except that there are parallel lines. This is due not only to the construction of the affine plane from the projective plane but also because like projective geometry, affine geometry lacks a metric.

Affinities

A collineation on $\mathbb{R}^2$ is equivalent to a collineation on $\mathbb{P}^2$ that leaves the line $x_3 = 0$ fixed. No points can be mapped to or from it because it is not part of $\mathbb{R}^2$. By Theorem 14.4.6, the equations in homogeneous coordinates that describe these collineations are

$$\begin{bmatrix} kx_1' \\ kx_2' \\ kx_3' \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \tag{15.2}$$

where

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \neq 0.$$

If we use nonhomogeneous coordinates, Equation 15.2 becomes

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} a_{13} \\ a_{23} \end{bmatrix}. \tag{15.3}$$

■ DEFINITION 15.2.3

The function $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ is called an **affinity** if there exist an invertible 2×2 matrix A with real entries and vector $\mathbf{a}$ in $\mathbb{R}^2$ such that

$$\phi(\mathbf{x}) = A\mathbf{x} + \mathbf{a}.$$

Let **AF** denote the set of all affinities. We show that **AF** is a subgroup of **PR**. To do this, first note that by our work on the bottom of page 532, $\mathbf{AF} \subseteq \mathbf{PR}$. Next take $\phi, \psi \in \mathbf{AF}$. This means that we can write ϕ as in Definition 15.2.3 and ψ as

$$\psi(\mathbf{x}) = B\mathbf{x} + \mathbf{b},$$

where $\det(B) \neq 0$ and $\mathbf{b} \in \mathbb{R}^2$. We check the three conditions of Theorem 15.1.9.

- **AF** is closed under composition because
$$(\phi \circ \psi)(\mathbf{x}) = \phi(B\mathbf{x} + \mathbf{b}) = AB\mathbf{x} + A\mathbf{b} + \mathbf{a},$$
which is an element of **AF** because AB is invertible and $A\mathbf{b} + \mathbf{a} \in \mathbb{R}^2$.
- Certainly, the identity map is an element of **AF**.
- Define $\psi(x) = A^{-1}\mathbf{x} - A^{-1}\mathbf{a}$ and calculate:
$$(\phi \circ \psi)(\mathbf{x}) = \phi(A^{-1}\mathbf{x} - A^{-1}\mathbf{a}) = A(A^{-1}\mathbf{x} - A^{-1}\mathbf{a}) + \mathbf{a} = \mathbf{x}$$
and
$$(\psi \circ \phi)(\mathbf{x}) = \psi(A\mathbf{x} + \mathbf{a}) = A^{-1}(A\mathbf{x} + \mathbf{a}) - A^{-1}\mathbf{a} = \mathbf{x}.$$
Therefore, $\psi = \phi^{-1}$.

Affinities preserve neither distance nor angle measure. For example, take the affinity ϕ defined by

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 3 \\ 6 \end{bmatrix}. \tag{15.4}$$

Then, $\phi(0, 0) = (3, 6)$ and $\phi(1, 1) = (6, 15)$. We see that the distance between $(0, 0)$ and $(1, 1)$ is different from the distance between $(3, 6)$ and $(11, 29)$. Moreover, if we compare the angle of inclination α of the original segment with the angle of inclination β of the image, we compute that $\cos\alpha = 1/\sqrt{2}$ and $\cos\beta = 3/\sqrt{89}$, which imply that $\alpha \neq \beta$.

However, affinities do preserve parallel lines.

■ THEOREM 15.2.4

The image of two parallel lines under an affinity is two parallel lines.

PROOF

Let ϕ be an affinity of an affine plane, so we can view ϕ as a projective collineation that leaves the line at infinity fixed. Assume that lines l and m are parallel. We conclude that there exists an ideal point Ω that is incident with both l and m. Therefore, $\phi(\Omega)$ is an ideal point, and since ϕ is a collinearity, $\phi[l]$ and $\phi[m]$ must intersect at $\phi(\Omega)$. Hence, $\phi[l]$ and $\phi[m]$ are parallel. ■

Exercises

1. Prove that $(\mathbb{A}(\mathcal{P}), \mathbb{L}(\mathcal{L}))$ is an affine plane for all projective planes $(\mathcal{P}, \mathcal{L})$.

2. Can $\mathbb{Z}^2$ be considered an affine plane?

3. Let $\mathcal{P}_A = \{A, B, C, D\}$ and

$$\mathcal{L}_A = \{\{A, B\}, \{A, C\}, \{A, D\}, \{B, C\}, \{B, D\}, \{C, D\}\}.$$

   (a) Show that $(\mathcal{P}_A, \mathcal{L}_A)$ is an affine plane.
   (b) Find all affinities of $(\mathcal{P}_A, \mathcal{L}_A)$.

4. Can the Fano plane (Theorem 14.1.4) be considered an affine plane? Explain.

5. In Figure 14.2 of the Fano plane, let $A = (0, 1, 0)$, $E = (0, 0, 1)$, $G = (1, 0, 0)$, and $C = (1, 1, 1)$.
   (a) Find the homogeneous coordinates for B, D, and F.
   (b) Which line in Figure 14.2 should be considered the line at infinity?
   (c) If possible, find the nonhomogeneous coordinates for A, B, C, D, E, F, and G.
   (d) Let $\mathbb{A}$ represent the ordered pairs in nonhomogeneous coordinates from Exercise 5(c). Let $\mathbb{L}$ be the set of lines through the points of $\mathbb{A}$. Find $\mathbb{L}$.
   (e) Show that $(\mathbb{A}, \mathbb{L})$ is an affine plane.
   (f) Find an affinity $\phi : \mathbb{A} \to \mathbb{A}$ such that $\phi(E) = E$, $\phi(B) = F$, and $\phi(F) = B$. Write the function in matrix form (Equation 15.3).
   (g) Two affine planes are **isomorphic** if there exists a collineation between them (compare Definition 14.1.6). Using the affine plane $(\mathcal{P}_A, \mathcal{L}_A)$ from Exercise 3, show that $(\mathbb{A}, \mathbb{L})$ is isomorphic to $(\mathcal{P}_A, \mathcal{L}_A)$.

6. Let $(\mathcal{P}_A, \mathcal{L}_A)$ be an arbitrary affine plane. Prove the following.
   (a) Two distinct lines in $\mathcal{L}_A$ have at most one point in common.
   (b) For all $l, m, n \in \mathcal{L}_A$, if m and n intersect and l is parallel to m, then l intersects n.
   (c) If there exists a line in $\mathcal{L}_A$ that contains n points, all lines in $\mathcal{L}_A$ contains n points.
   (d) For all $l, m \in \mathcal{L}_A$, let $l \sim m$ if and only if $l = m$ or l is parallel to m. Show that $\sim$ is an equivalence relation.

7. Can an affine plane be constructed with 10 points? 16 points? Explain.

8. Which of Postulates 14.1.2 are implied by Postulates 15.2.1?

9. Prove that Postulates 15.2.1 are consistent and independent.

10. Confirm that the set of collineations form a subgroup of Trans$(\mathbb{P}^2)$.

11. Find the homogeneous coordinates for the point $(4, 10)$ given in nonhomogeneous coordinates, and give the nonhomogeneous coordinates for the homogeneous $(4, 10, 7)$.

12. Write the function given in Equation 15.4 using homogeneous coordinates.

13. Let $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ be an affinity such that $\phi(1, 1) = (4, 0)$ and $\phi(3, 2) = (2, 6)$. Give an equation for ϕ in the form $\phi(\mathbf{x}) = A\mathbf{x} + \mathbf{a}$.

14. A **dilatation** is a collineation $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ such that for all lines l, $\phi[l]$ is parallel to l.

   (a) Why is a dilatation an affinity?
   (b) Draw an example of a dilatation in $\mathbb{R}^2$ and give its equation in the form $\phi(\mathbf{x}) = A\mathbf{x} + \mathbf{a}$.
   (c) Is the set of dilatations a group under composition?
   (d) Prove that a dilatation that has two invariant points is the identity map.
   (e) A **stretch** is an affinity ψ so that $\psi(x, y) = (kx, ky)$ for some k. Show that a stretch is a dilatation.

15. An **affine translation** is the identity map or a dilatation with no fixed points. Show that the set of affine translations is a group under composition.

16. Let τ be an affine translation and ϕ an affinity. Prove that $\phi^{-1} \circ \tau \circ \phi$ is an affine translation.

17. An affinity ϕ such that $\phi(A) = B$ is a **shear** if it fixes every point on lines that are parallel but not equal to $\overline{AB}$.

   (a) Draw an example of a shear in $\mathbb{R}^2$ and give its equation in the form $\phi(\mathbf{x}) = A\mathbf{x} + \mathbf{a}$.
   (b) Does the set of shears form a group under composition? Explain.

18. An affinity ψ is an **affine reflection** if for some points A and B, $\psi(A) = B$, $\psi(B) = A$, and there exists a point C not on $\overline{AB}$ such that $\psi(C) = C$.

   (a) Draw an example of an affine reflection in $\mathbb{R}^2$ and give its equation in the form $\psi(\mathbf{x}) = A\mathbf{x} + \mathbf{a}$.
   (b) Does the set of affine reflections form a group under composition?
   (c) Let M be the midpoint of $\overline{AB}$. Show that ψ fixes every point on $\overline{CM}$.

15.3 THE PRINCIPAL GROUP

Jordan taught his students that transformation groups would one day play an important role in the development of geometry. This was realized by the work of Klein and Lie. Influenced by what they learned in Paris, Klein focused on discrete groups, while Lie made studying **continuous groups** a lifelong project. Examples of these uncountable groups include $(\mathbb{R}, +)$ and the groups of projective and affine transformations. Initially, however, Klein's and Lie's group theoretic interests coincided with the group of transformations of the Euclidean plane (Yaglom 1988, 95, 97).

As projective geometry can be defined as the study of the properties that remain invariant under the group of projectivities, Klein defined Euclidean geometry as the study of properties that are preserved under transformations of the Euclidean plane.

To understand the reason for his definition, we need to go back to Euclid. Although he used it reluctantly, Euclid applied superposition to prove SAS (page 138). He assumed that he could take a triangle and move it so that it aligned with any other triangle to which it was congruent. As the triangle moved, it did not deform. Its sides remained the same length and its angles kept their measure. Euclid also studied similar figures. Two figures are similar if there exists a transformation that changes only the size of the figure, not its angle measures. All of this means that Euclid indirectly applied transformations of the plane that preserved both distance and angle measure. With this in mind, Klein (1893, 218) made the following definition.

■ **DEFINITION 15.3.1**

The **principal group PG** is the transformation group on $\mathbb{R}^2$ that consists of the following functions:

- motions in the plane (translations, rotations)
- transformations into symmetrical figures (reflections)
- transformations into similar figures (dilations and contractions)
- any composition of the above

Isometries

All of the maps of the principal group are conformal, but the first two are also guaranteed to preserve distance. Throughout, let d represent the standard metric.

■ **DEFINITION 15.3.2**

A one-to-one correspondence $f : \mathbb{R}^2 \to \mathbb{R}^2$ is an **isometry** if

$$d[A, B] = d[f(A), f(B)]$$

for all $A, B \in \mathbb{R}^2$. In this case we say that f **preserves distance**.

We begin by proving three basic properties of isometries.

■ **THEOREM 15.3.3**

An isometry is a collineation.

PROOF

Let ϕ be an isometry and assume that $\phi(A) = A'$ and $\phi(B) = B'$. Take P on $\overline{AB}$. To obtain a contradiction, suppose that $\phi(P) = P'$ is not on $\overline{A'B'}$. By the Triangle Inequality (5.3.11),

$$A'B' < A'P' + P'B'.$$

However, this is impossible because

$$A'B' = AB = AP + PB = A'P' + P'B'.$$

Therefore, P' is on $\overline{A'B'}$. ■

■ THEOREM 15.3.4

Isometries are conformal.

PROOF

Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be an isometry and choose three points A, B, C in $\mathbb{R}^2$. Then

$$\begin{aligned} d[A, B] &= d[\phi(A), \phi(B)], \\ d[A, C] &= d[\phi(A), \phi(C)], \\ d[B, C] &= d[\phi(B), \phi(C)]. \end{aligned}$$

Thus, by SSS, $\triangle ABC \cong \triangle f(A)f(B)f(C)$, from which we conclude that $\sphericalangle ABC \cong \sphericalangle f(A)f(B)f(C)$. ■

■ THEOREM 15.3.5

An isometry is uniquely determined by three noncollinear points.

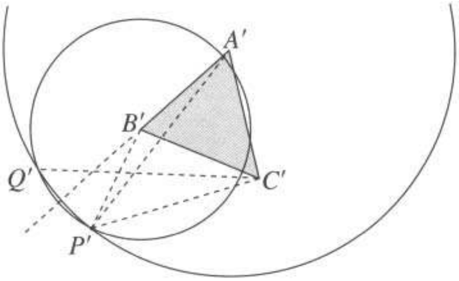

PROOF

Let A, B, and C be distinct noncollinear points. Let ϕ and ψ be two isometries and assume that

$$\phi(A) = \psi(A) = A', \ \phi(B) = \psi(B) = B', \text{ and } \phi(C) = \psi(C) = C'.$$

We need to show that $\phi = \psi$. To do this, take any point P in the plane and let $\phi(P) = P'$. We will show that P' is uniquely determined by its distance to each of the points A', B', and C', thus showing that $\phi(P) = \psi(P)$. Suppose that there is another point Q' that has the same distances from A', B', and C' as P'. We can then draw $\odot A'$ with radius $\overline{A'P'}$ and $\odot B'$ with radius $\overline{B'P'}$ such that Q' is also on the intersection of the two circles. Join $\overline{C'Q'}$ and $\overline{C'P'}$. The only way that these two segments can be congruent is for C' to lie on the perpendicular bisector of $\overline{Q'P'}$. However, this is not the case because the $\overline{A'B'}$ produced is that perpendicular bisector and C' is not collinear with A' and B'. Therefore, Q' does not exist, and P' is described uniquely by its distances to A', B', and C'. Since the distances from $\phi(A)$, $\phi(B)$, and $\phi(C)$ to P', respectively, equal the distances from $\psi(A)$, $\psi(B)$, and $\psi(C)$ to P', it must be the case that $\psi(P) = P'$ since ψ preserves distances. ■

The fact that three noncollinear points determine an isometry uniquely will enable us to show that all isometries are affinities.

■ THEOREM 15.3.6

A function $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ is an isometry if and only if there exists a 2×2 matrix A, a vector $\mathbf{b} \in \mathbb{R}^2$, and an angle θ such that either

$$A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \text{ or } A = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix}$$

and $\phi(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$.

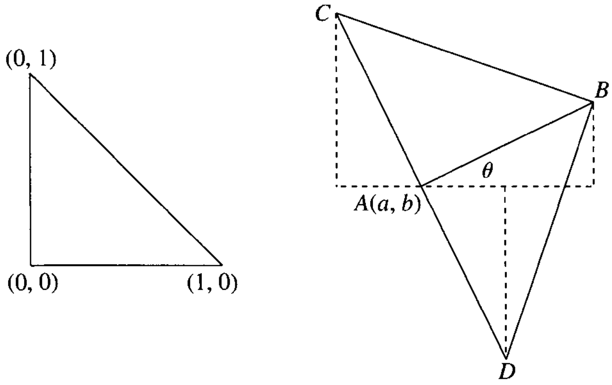

PROOF

We leave to Exercise 4 the proof that $\phi(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$ with A and $\mathbf{b}$ as described is an isometry. So to prove the converse, let ϕ be an isometry. First suppose that $\phi(0, 0) = (a, b) = A$. We have two cases to check, depending on whether ϕ preserves the orientation of the vertices of the triangle. Assume that it does. This means that $\phi(1, 0) = B$ and $\phi(0, 1) = C$, as in the diagram. Let θ be the angle of inclination of $\overline{AB}$. Then, since $AB = 1$,

$$\phi(1, 0) = (a + \cos\theta,\ b + \sin\theta),$$
$$\phi(0, 1) = (a - \sin\theta,\ b + \cos\theta).$$

Since ϕ is an isometry, we may conclude by Theorem 15.3.5 that

$$\phi(x, y) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} a \\ b \end{bmatrix}$$

because a function of this form is an isometry (Exercise 4). If the isometry does not preserve the order of the vertices, then $\phi(0, 1) = D$ and

$$\phi(0, 1) = (a + \cos(\theta - \pi/2), b + \sin(\theta - \pi/2)) = (a + \sin\theta, b - \cos\theta).$$

Therefore,

$$\phi(x, y) = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} a \\ b \end{bmatrix},$$

again due to Theorem 15.3.5. ■

Since both versions of A in the theorem have a nonzero determinant, we may conclude the following result.

■ COROLLARY 15.3.7

An isometry of $\mathbb{R}^2$ is an affinity.

Now to confirm the group structure of the set of isometries. We leave the proof of the first part of the lemma to Exercise 7.

■ LEMMA 15.3.8

- The composition of isometries is an isometry.
- The inverse of an isometry is an isometry.

PROOF

Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be an isometry. Take $X', Y' \in \mathbb{R}^2$ and let X and Y be in the plane so that $f^{-1}(X') = X$ and $f^{-1}(Y') = Y$, which means that $f(X) = X'$ and $f(Y) = Y'$. Therefore,

$$d[X', Y'] = d[f(X), f(Y)] = d[X, Y] = d[f^{-1}(X'), f^{-1}(Y')]. \blacksquare$$

Lemma 15.3.8 combined with the fact that the identity is an isometry yield the next theorem.

■ THEOREM 15.3.9

The set of isometries of $\mathbb{R}^2$ is a subgroup of the principal group.

Similarities

Isometries are conformal and preserve distances. If we require a transformation on $\mathbb{R}^2$ to only be conformal, we obtain another group of functions of the plane.

■ DEFINITION 15.3.10

The transformation $\sigma : \mathbb{R}^2 \to \mathbb{R}^2$ is a **similarity** if there exists $r > 0$ such that for all $A, B \in \mathbb{R}^2$, if $A' = \sigma(A)$ and $B' = \sigma(B)$, then $A'B' = rAB$. The real number r is called the **ratio** of the similarity.

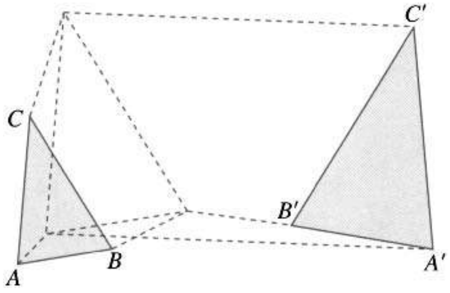

Figure 15.3 $\triangle A'B'C'$ is the image of $\triangle ABC$ under a similarity.

The similarity σ_0 defined by $\sigma_0(x, y) = (rx, ry)$ is an example of a **homothety** (Exercise 12). It leaves the origin fixed and either dilates or contracts all other points, depending on the value of r. To obtain formulas for the other similarities, take an isometry ϕ and compose it with σ_0 (Figure 15.3). That the resulting composition is a similarity is seen by

$$\|(\phi \circ \sigma_0)(x, y)\| = \|\phi(rx, ry)\| = \|(rx, ry)\| = r\|(x, y)\|,$$

where the second equality holds because ϕ is an isometry. This result can be used to prove the following characterization of similarities (Exercise 11).

■ THEOREM 15.3.11

If σ is a similarity with ratio r, there exists a 2×2 invertible matrix A and a vector $\mathbf{b} \in \mathbb{R}^2$ such that

$$\sigma(\mathbf{x}) = rA\mathbf{x} + \mathbf{b},$$

where

$$\phi(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$$

is an isometry.

■ COROLLARY 15.3.12

A similarity is an affinity.

Similarities do share some properties with isometries. The following is an example. Its proof is a modification of Theorems 15.3.3 and 15.3.4 (Exercise 16).

■ THEOREM 15.3.13

A similarity is a conformal collineation.

The similarities also form a subgroup of the principal group. To see this, we first show that the set of similarities is closed under composition.

■ LEMMA 15.3.14

The composition of similarities is a similarity.

PROOF

Let σ_1 and σ_2 be similarities with ratios r_1 and r_2. Then there are invertible matrices A_1 and A_2 with $\mathbf{b}_1$, $\mathbf{b}_2 \in \mathbb{R}^2$ so that $A_1\mathbf{x}+\mathbf{b}_1$ and $A_2\mathbf{x}+\mathbf{b}_2$ represent isometries and

$$\sigma_1(\mathbf{x}) = r_1A_1\mathbf{x} + \mathbf{b}_1 \text{ and } \sigma_2(\mathbf{x}) = r_2A_2\mathbf{x} + \mathbf{b}_2.$$

We now compute for all $\mathbf{x} \in \mathbb{R}^2$,

$$\begin{aligned}\|(\sigma_1 \circ \sigma_2)(\mathbf{x})\| &= \|\sigma_1(r_2A_2\mathbf{x} + \mathbf{b}_2)\| \\ &= \|r_1r_2A_1A_2\mathbf{x} + r_1A_1\mathbf{b}_2 + \mathbf{b}_1\| \\ &= r_1r_2\left\|A_1A_2\mathbf{x} + \frac{1}{r_2}A_1\mathbf{b}_2 + \frac{1}{r_1r_2}\mathbf{b}_1\right\| \\ &= r_1r_2\|\mathbf{x}\|.\end{aligned}$$

The last equality follows from the fact proven in Lemma 15.3.8 that the composition of isometries is an isometry (Exercise 8). Therefore, we conclude that $\sigma_1 \circ \sigma_2$ is a similarity. ■

The lemma, together with the facts that the identity map is a similarity with $r = 1$ and the inverse of a similarity with ratio r is a similarity with ratio $1/r$ (Exercise 13), gives us the next theorem.

■ THEOREM 15.3.15

- The set of similarities form a subgroup of the principal group.
- The set of isometries form a subgroup of the set of similarities.

Many more results concerning isometries and similarities, including applications, may be found in Dodge (1972, 48–130). That text served as the general reference for this section and the next.

Exercises

1. Define $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = x^3 + 1$.
   (a) Prove that f is a transformation.
   (b) Define what it would mean for f to be an isometry on $\mathbb{R}$. Based on the definition chosen, show that f not an isometry.
   (c) Find an example of an isometry on $\mathbb{R}$. Must it be linear?
2. Let $g : \mathbb{R}^2 \to \mathbb{R}^2$ be defined by $g(x, y) = (x + 2, y^5 - 1)$.
   (a) Prove that g is a transformation.
   (b) Is g an isometry? Explain.

3. Prove directly from the definition that $f(x, y) = (x - 7, y + 5)$ is an isometry.

4. Prove that $\phi(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$ is an isometry if either

$$A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \text{ or } A = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix}.$$

5. Prove that isometries preserve the areas of rectangles and triangles. Do isometries preserve the areas of all plane figures?

6. Prove that the image of a circle under an isometry is a circle.

7. Prove that the composition of isometries is an isometry (Lemma 15.3.8).

8. Let ϕ be an isometry. Show that for all $\mathbf{b} \in \mathbb{R}^2$, $\psi(\mathbf{x}) = \phi(\mathbf{x}) + \mathbf{b}$ is an isometry.

9. Is an isometry that fixes three noncollinear points the identity map? Explain.

10. Let $g(x, y) = (4x + 2, 4y - 8)$. Prove that g is a similarity.

11. Prove Theorem 15.3.11.

12. Given a point C called the **center** and $r > 0$, a **homothety** η is a transformation of the plane such that for all points A,

$$d[\eta(A), C] = rd[A, C].$$

- (a) Prove that a homothety has only one fixed point.
- (b) Show that the composition of two homotheties with the same center is a homothety. Identify the composition's ratio and center.
- (c) Let $\triangle ABC$ and $\triangle A'B'C'$ be similar but not congruent triangles such that corresponding sides are parallel. Find a homothety η such that the image of $\triangle ABC$ under η is $\triangle A'B'C'$.
- (d) Prove or disprove: The set of homotheties form a subgroup of the set of similarities.

13. Prove that the inverse of a similarity with ratio r is a similarity with ratio $1/r$.

14. Is the group of similarities abelian? Explain.

15. Let σ be a similarity with ratio r. If $r \geqslant 1$, then σ is a **dilation**, and if $0 < r < 1$, then σ is a **contraction**. Is the set of dilations a subgroup of the group of similarities? What about the set of contractions? Explain.

16. Prove Theorem 15.3.13

17. Let σ be an similarity with ratio r. Let $\mathcal{F}$ be a convex figure in the plane. True or false: The area of $\sigma[\mathcal{F}]$ equals r^2 times the area of $\mathcal{F}$. Give an example.

18. Prove the following for transformations of $\mathbb{R}^2$:
- (a) The composition of conformal maps is conformal.
- (b) The inverse of a conformal map is conformal.

19. Define $\phi(\mathbf{x}) = rA\mathbf{x} + \mathbf{b}$. For each A and $\mathbf{b}$ given, find ϕ^{-1}, the image of the triangle with vertices (0, 0), (1, 0), and (0, 1), and whether ϕ is an isometry or a similarity.

(a) $A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} -4 \\ 3 \end{bmatrix}$.

(b) $A = \begin{bmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ \sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

(c) $A = \begin{bmatrix} 2\sqrt{3} & -2 \\ 2 & 2\sqrt{3} \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$.

15.4 ISOMETRIES OF THE PLANE

There are four different types of isometries of the principal group. In this section we examine each and then prove that there are no other types.

Reflections

We begin our introduction to the transformations of the principal group by defining the function that is used to determine the symmetries of a plane figure. These functions also have the property that they map figures to congruent figures.

■ DEFINITION 15.4.1

The transformation $\mu : \mathbb{R}^2 \to \mathbb{R}^2$ is a **reflection** through a line l if for all $A \in \mathbb{R}^2$, if A is not on l, the line l is the perpendicular bisector of the segment joining A and $\mu(A)$, otherwise $\mu(A) = A$. We call l the **mirror** of the reflection.

The definition is illustrated in Figure 15.4. Although a reflection is an isometry and every reflection is its own inverse (Exercise 6), the diagram makes clear that the

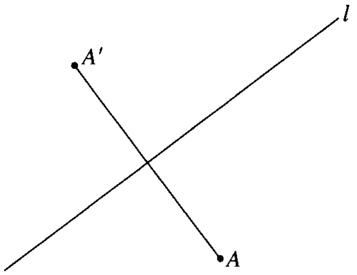

Figure 15.4 A reflection of A to image A' through l.

identity is not a reflection. Hence, the set of reflections do not form a subgroup of the principal group. Instead, they serve as the building blocks for the group of isometries.

■ THEOREM 15.4.2

Every isometry is the composition of at most three reflections.

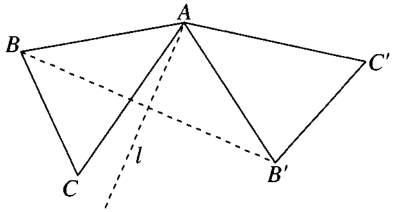

PROOF

Let ϕ be an isometry on $\mathbb{R}^2$. By Theorem 15.3.5 it suffices to examine ϕ on any three noncollinear points A, B, and C. We have four cases to consider.

- Let $\phi(A) = A$, $\phi(B) = B$, and $\phi(C) = C$. This means that ϕ is the identity, so let μ be any reflection. Then $\phi = \mu \circ \mu$.

- Let $\phi(A) = A$, $\phi(B) = B$, and $\phi(C) \neq C$. In this case ϕ is a reflection with mirror $\overline{AB}$ (Exercise 4).

- Let $\phi(A) = A$, $\phi(B) = B' \neq B$, and $\phi(C) = C' \neq C$. This is illustrated in the diagram for this theorem. Since $\triangle ABB'$ is isosceles, A is on the perpendicular bisector l of $\overline{BB'}$. Therefore, the reflection μ with mirror l has

  $$\mu(A) = A \text{ and } \mu(B') = B,$$

  and then

  $$(\mu \circ \phi)(A) = A \text{ and } (\mu \circ \phi)(B) = \mu(B') = B.$$

  By the previous two cases, $\mu \circ \phi$ is either the identity or a reflection. If $\mu \circ \phi$ is the identity, ϕ is a reflection. If $\mu \circ \phi$ is a reflection, ϕ is the composition of two reflections.

- The case where $\phi(A) \neq A$, $\phi(B) \neq B$, and $\phi(C) \neq C$ is also left to Exercise 4. ■

The natural question to ask next is whether all reflections are isometries. The next proposition gives the answer. Since it can be proven either along the lines of Theorem 15.3.6 or by using methods from later in this section, its proof is left to Exercise 5.

■ THEOREM 15.4.3

The equation for the reflection μ with a mirror that has an angle of inclination $\theta/2$ and passes through the point (a, b) is

$$\mu(x, y) = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix} \begin{bmatrix} x-a \\ y-b \end{bmatrix} + \begin{bmatrix} a \\ b \end{bmatrix}.$$

The consequence of the theorem is that reflections are isometries because they can be written in the form of Theorem 15.3.6.

■ COROLLARY 15.4.4

Reflections are isometries.

Translations

The most basic plane motion is when a figure is slid in the direction of a given vector.

■ DEFINITION 15.4.5

The transformation $\tau : \mathbb{R}^2 \to \mathbb{R}^2$ is a **translation** if there exists a vector $\langle a, b\rangle$ such that

$$\tau(x, y) = (x + a, y + b).$$

We say that $\langle a, b\rangle$ **describes** τ. This is illustrated in Figure 15.5.

Observe that the translation τ in the definition can be written in the form

$$\tau(x, y) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} a \\ b \end{bmatrix}. \tag{15.5}$$

Conversely, any function of the form given in Equation 15.5 is a translation. Thus, we have a proposition that quickly follows from the definition by Theorem 15.3.6.

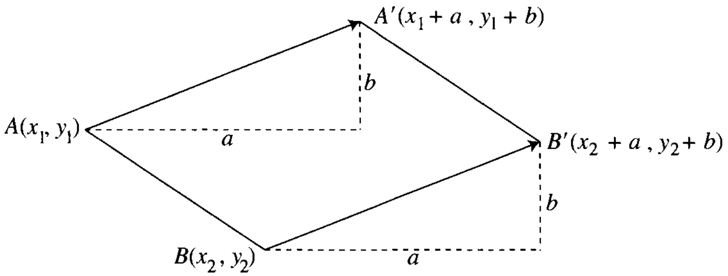

Figure 15.5 A translation of $\overline{AB}$ to image $\overline{A'B'}$.

THEOREM 15.4.6

A translation is an isometry.

As an isometry, a translation can be written as the composition of at most three reflections (Theorem 15.4.2). The next theorem improves on that.

THEOREM 15.4.7

A transformation is a translation if and only if it is the composition of two reflections with parallel mirrors.

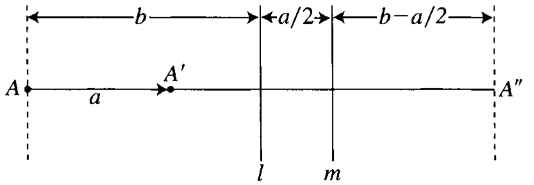

PROOF

We prove sufficiency. Take a vector of length a and let it describe a translation. Let l and m be two parallel lines that are perpendicular to the vector and are a distance $a/2$ apart. We claim that the translation is the composition of the reflection with mirror m and the reflection with mirror l. To prove this, let A be a point in the plane and let A' be its image under the translation. Let b be the distance from A to l. Take A'' to be the image of A by the reflection through l, so the distance from A to A'' is $2b$ and the distance from A'' to m is $b - a/2$, allowing for the possibility that this distance is signed. Now reflect A'' through m. Since $2b - 2(b - a/2) = a$, the image of this last reflection is A'. ■

We next want to show that the set of translations form a subgroup of the principal group under composition. For this we need some results.

LEMMA 15.4.8

The composition of translations is a translation.

PROOF

Let τ_1 and τ_2 be two translations of the plane. Let $\langle a_1, b_1\rangle$ be the vector that describes τ_1 and $\langle a_2, b_2\rangle$ be the vector that describes τ_2. When we compose the two functions, we find that

$$\begin{aligned}(\tau_1 \circ \tau_2)(x, y) &= \tau_1(\tau_2(x, y))\\ &= \tau_1(x + a_2, y + b_2)\\ &= ([x + a_2] + a_1, [x + b_2] + b_1)\\ &= (x + [a_2 + a_1], y + [b_2 + b_1]).\end{aligned}$$

The last ordered pair in this equation is that which defines the translation with vector $\langle a_2 + a_1, b_2 + b_1\rangle$. ■

Notice that although function composition is not generally commutative, it is in the case of translations. Namely, for any two translations τ_1 and τ_2,

$$\tau_1 \circ \tau_2 = \tau_2 \circ \tau_1.$$

This is because $\langle a_1 + a_2, b_1 + b_2 \rangle = \langle a_2 + a_1, b_2 + b_1 \rangle$.

■ LEMMA 15.4.9

If $\langle a, b \rangle$ describes the translation τ, $\langle -a, -b \rangle$ describes τ^{-1}.

PROOF

Let $\langle a, b \rangle$ describe τ and let $\langle -a, -b \rangle$ describe σ. We calculate:

$$(\tau \circ \sigma)(x, y) = \tau(x - a, y - b) = ([x - a] + a, [y - b] + b) = (x, y).$$

Therefore, $\tau \circ \sigma$ is the identity map. Since we know that composition is commutative for translations, we conclude that $\sigma \circ \tau$ is also the identity map. Thus, $\sigma = \tau^{-1}$. ■

The previous two lemmas, plus the fact that the identity map is a translation, allow us to draw our conclusion.

■ THEOREM 15.4.10

The set of translations of $\mathbb{R}^2$ is an abelian subgroup of the principal group.

Rotations

To rotate a point A around a center point O means to move A along the arc of a circle with center O and radius of length OA through a fixed angle measure θ. The result of the motion is A' in Figure 15.6. Algebraically, this means the following: Let (x, y) be a point in $\mathbb{R}^2$. Let ρ be the transformation that rotates (x, y) to (x', y') along an arc of measure θ radians around the origin. Let r be the radius, so

$$r = \sqrt{x^2 + y^2},$$

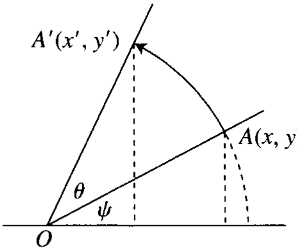

Figure 15.6 A rotation of A about the origin to image A'.

and let ψ be the angle that the line through the origin and (x, y) makes with the x-axis. This means that

$$x = r\cos\psi,\ y = r\sin\psi$$

and

$$x' = r\cos(\theta + \psi),\ y' = r\sin(\theta + \psi).$$

Using the addition formulas from trigonometry yields

$$x' = r\cos\theta\cos\psi - r\sin\theta\sin\psi = x\cos\theta - y\sin\theta$$

and

$$y' = r\sin\theta\cos\psi + r\cos\theta\sin\psi = x\sin\theta + y\cos\theta.$$

■ DEFINITION 15.4.11

The transformation $\rho : \mathbb{R}^2 \to \mathbb{R}^2$ is a **rotation** about the origin if there exists an angle measure θ such that

$$\rho(x, y) = (x\cos\theta - y\sin\theta,\ x\sin\theta + y\cos\theta).$$

We say that θ **describes** ρ.

We will not specifically define a function that represents a rotation about an arbitrary point. Instead, we compose translations with a rotation about the origin. Fix the point P with coordinates (a, b). Let τ be the translation described by $\langle -a, -b\rangle$ and ρ the rotation described by θ. The function

$$\tau^{-1} \circ \rho \circ \tau$$

is the rotation about P of measure θ (Figure 15.7). From this we can show that an arbitrary rotation with center (a, b) has the form

$$\rho(x, y) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x - a \\ y - b \end{bmatrix} + \begin{bmatrix} a \\ b \end{bmatrix}. \tag{15.6}$$

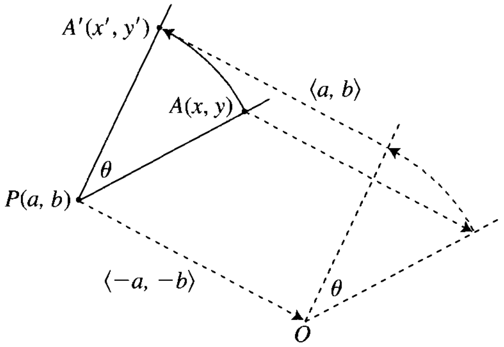

Figure 15.7 A rotation of A about P to image A'.

Moreover, any equation of the form given in Equation 15.6 is a rotation.

As with translations, we can write a rotation as a composition of two reflections.

■ THEOREM 15.4.12

A transformation is a rotation if and only if it is the composition of two reflections with intersecting mirrors.

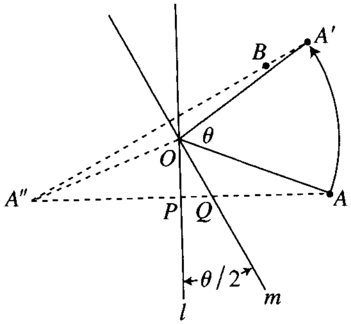

PROOF

We prove sufficiency. Let O be the center of the rotation and θ its angle. Let A be a point in the plane and A' its image under the rotation. Choose any line l containing O and let m also be through O, forming an angle of measure $\theta/2$ measured counterclockwise from l. Let A'' be the image of A under the reflection through mirror l and B the image of A'' through the mirror m. Then

$$\triangle A''PO \cong \triangle APO$$

by SAS. From this we conclude that $\overline{OA''} \cong \overline{OA}$, so A, A', and A'' lie on the same circle with center O. Moreover, letting Q be the intersection of m and $\overline{A''A}$, we see that $m\sphericalangle QA''B = \theta/2$ because this angle is the complement of $\sphericalangle OQP$. Therefore, the point B is actually the point A' because $\sphericalangle QA''A$ has the measure of an inscribed angle in $\odot O$ that intercepts $\overset{\frown}{AA'}$. Hence, A reflected through l and then through m results in A'. ■

■ COROLLARY 15.4.13

A rotation is an isometry.

We now identify the group structure of rotations. The first lemma involves the composition of two rotations. Supposing that they are both centered at the same point, the composition is another rotation about that point. Otherwise, we might obtain a translation.

■ LEMMA 15.4.14

The composition of two rotations is either a rotation or a translation.

PROOF

Let ρ_1 and ρ_2 be rotations about the points O and P, respectively. Let θ_1 be the angle the describes ρ_1 and θ_2 be the angle that describes ρ_2. Join $\overline{OP}$. As in the proof of Theorem 15.4.12, find l through O and a line m through P such that the measure of the angle from l to $\overline{OP}$ is $\theta_1/2$ and the measure of the angle from $\overline{OP}$ to m is $\theta_2/2$. That is,

$$\rho_1 = \mu_2 \circ \mu_1$$

and

$$\rho_2 = \mu_3 \circ \mu_2,$$

where μ_1 is the reflection with mirror l, μ_2 is the reflection with mirror $\overline{OP}$, and μ_3 is the reflection with mirror m. Thus,

$$\rho_2 \circ \rho_1 = \mu_3 \circ \mu_2 \circ \mu_2 \circ \mu_1 = \mu_3 \circ \mu_1.$$

If l and m are parallel, the composition is a translation (Theorem 15.4.7), otherwise it is a rotation (Theorem 15.4.12). ■

The next proposition states that the inverse of a rotation is the rotation in the opposite direction. Its proof is Exercise 14.

■ LEMMA 15.4.15

If θ describes the rotation ρ, then $-\theta$ describes ρ^{-1}.

We now have the pieces to prove that these transformations have a group structure under composition (Exercise 8).

■ THEOREM 15.4.16

The following are subgroups of the principal group. The first subgroup is abelian, but the second is not.

- The set of rotations about a given point.
- The set of translations and rotations.

There is one more transformation needed to complete our list of isometries. A **glide reflection** γ (Figure 15.8) is the composition of a reflection with mirror l and a translation described by a vector $\mathbf{v}$ that is parallel to l with angle of inclination θ. By Theorem 15.4.3 and Equation 15.5, the equation for a glide reflection with a mirror through a point (a, b) can be described by a function of the form

$$\gamma(x, y) = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix} \begin{bmatrix} x - a \\ y - b \end{bmatrix} + \begin{bmatrix} a + \|\mathbf{v}\| \cos\theta \\ b + \|\mathbf{v}\| \sin\theta \end{bmatrix} \tag{15.7}$$

Hence, when we examine the general form of an isometry given by Theorem 15.3.6 with Theorem 15.4.3 and Equations 15.5, 15.6, and 15.7, we can conclude that the translations of this section exhaust all possibilities for isometries in the plane. Therefore, we write the following result.

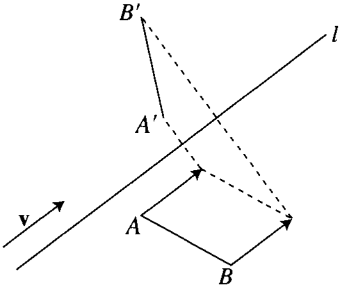

Figure 15.8 A glide reflection with vector **v** and mirror l.

■ THEOREM 15.4.17

The set of reflections, translations, rotations, and glide reflections of the plane is identical to the set of isometries.

Exercises

1. Find the image of the triangle with vertices $(0,\ 0)$, $(1,\ 0)$, and $(0,\ 1)$ under the following isometries.
   (a) a reflection with mirror $y = -4$
   (b) a reflection with mirror $2x + 5y = 1$
   (c) a translation described by vector $\langle 2,\ -3\rangle$
   (d) a translation $4\sqrt{2}$ units along the line $y = x$
   (e) a rotation about the origin of measure $\pi/6$
   (f) a rotation about the point $(-4,\ -3)$ of measure $3\pi/4$
   (g) a glide reflection with mirror $x = 3$ and vector $\langle 0,\ 4\rangle$
   (h) a glide reflection with mirror $x - 2y = 10$ and vector $\langle 2,\ 1\rangle$

2. For the transformations of Exercise 1.
   (a) Write each as a reflection or product of reflections.
   (b) Write each in the form $\phi(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$, where A is a 2×2 invertible matrix and $\mathbf{b}$ a vector in the plane.
   (c) Find the inverse of each transformation geometrically and then find it algebraically using the form $\phi(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$.

3. Identify all isometries that can map the given figure to itself.
   (a) equilateral triangle
   (b) square
   (c) regular hexagon
   (d) parallelogram

4. Prove the second and fourth cases in the proof of Theorem 15.4.2.

5. Prove Theorem 15.4.3.

6. Prove that a reflection is an isometry and a reflection is its own inverse.

7. Prove necessity for Theorems 15.4.7 and 15.4.12.

8. Prove Theorem 15.4.16.

9. How can a reflection be used to show that the base angles of an isosceles triangle are congruent?

10. Let A and B be distinct points in the plane. Prove that there exists exactly one reflection μ so that $\mu(A) = B$.

11. Show that the composition of three reflections is a reflection or a glide reflection.

12. Define $\mathscr{G}$ to be the set of translations described by a vector with integer coordinates. Show that $\mathscr{G}$ is a subgroup of the set of translations.

13. Show that the composition of two homotheties is a translation or a homothety.

14. Prove that if θ describes the rotation ρ, then $-\theta$ describes ρ^{-1}.

15. Prove that the composition of two rotations about the same point is another rotation about that point.

16. Let $n \in \mathbb{Z}$ and define $\mathscr{G}_n$ to be the set of rotations about the origin of measure $2k\pi/n$ $(k \in \mathbb{Z})$. Prove that $\mathscr{G}_n$ is a subgroup of the set of all rotations about the origin.

17. Does the set of glide reflections form a subgroup of the principal group?

18. What is the inverse of a glide reflection?

19. Give the equation for the image of the line $2x-y = 0$ under the given isometries.
    (a) a reflection with mirror $x + 5y = 3$
    (b) a translation described by the vector $\langle -3, 7\rangle$
    (c) a rotation of measure $\pi/6$ about the point $(1, 2)$
    (d) a glide reflection with mirror $x + y = 6$ and vector $\langle 3, -3\rangle$

20. Let $a, b \in \mathbb{R}$. Show for all $(x_1, y_1), (x_2, y_2) \in \mathbb{R}^2$,

$$d[(x_1, y_1), (x_2, y_2)] = d[(x_1 + a, y_1 + b), (x_2 + a, y_2 + b)]$$

and

$$d[(x_1, y_1), (x_1 + a, y_1 + b)] = d[(x_2, y_2), (x_2 + a, y_2 + b)].$$

21. We say that a plane figure $\mathscr{F}$ is **symmetric** with respect to a line l if given the reflection μ with mirror l, $\phi[\mathscr{F}] = \mathscr{F}$. In this case, l is called a **line of symmetry** for $\mathscr{F}$. Find the lines of symmetry of the given figures.
    (a) equilateral triangle
    (b) regular hexagon
    (c) parallelogram

22. Using the definition given in Exercise 21, show the each figure is symmetric through the given lines.
    (a) parallelogram through a diagonal
    (b) a triangle through a median
    (c) circle through a diameter
    (d) a regular pentagon through the line connecting a vertex with the midpoint of the opposite side

15.5 CONSISTENCY OF EUCLIDEAN GEOMETRY

We have seen three transformation groups connected with Klein's Erlangen Program,

$$\mathbf{PG} \subseteq \mathbf{AF} \subseteq \mathbf{PR},$$

and each of these groups leaves different sets of geometric properties invariant (Figure 15.9). Furthermore, the "larger" the group, the more general is its geometry. For instance, since **PG** is a subgroup of **AF**, a model of Euclidean geometry can be interpreted as an affine model but not necessarily conversely. Also, both of these geometries are contained within projective geometry when the line at infinity is included in the plane. We should note that not all geometries fit into Klein's classification scheme. Differential geometry is an example. However, the Erlangen Program directed research for about 50 years and provided a framework in which the relationships between many geometries could be studied (Kline 1972, 920).

Leipzig

After Klein introduced his program in 1872, he remained at the University of Erlangen for four years. Because of his outstanding work, the *Technische Hochschule* in

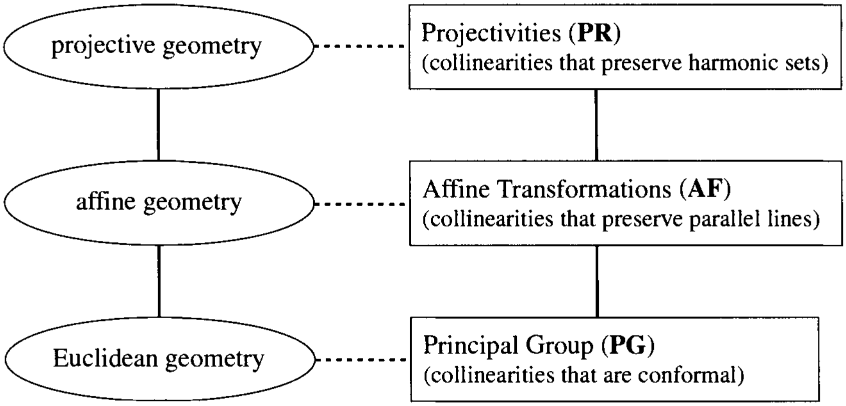
projective geometry
Projectivities (PR)
(collinearities that preserve harmonic sets)
affine geometry
Affine Transformations (AF)
(collinearities that preserve parallel lines)
Euclidean geometry
Principal Group (PG)
(collinearities that are conformal)

Figure 15.9 A hierarchy of geometries.

Munich offered him a position in the geometry department. As an institution with an outstanding scientific reputation, Klein was happy to accept. He spent several years there but moved on to Leipzig in 1880. At the university he began teaching geometry, set up a mathematical reading room, and taught students how to create mathematical drawings. However, his techniques were not well received by other faculty. Also during this time Klein began a correspondence with Poincaré concerning a mutual field of interest. In an attempt to beat him to results, Klein pushed himself to keep up with Poincaré. All of this stress led to a mental breakdown in 1882. Then in 1885, shortly after Klein had begun to recover from his rivalry with Poincaré, Klein met David Hilbert. At the time, Klein was 36 and Hilbert, 23 (Yaglom 1988, 125–126; Reid 1996, 18; Gray 2007, 225, 279).

Hilbert was born in Wehla near Kant's home town of Königsberg on January 23, 1862. His father was a judge and his mother had an interest in philosophy. Hilbert began school at the gymnasium when he was 8 years old. This was two years later than what was typical, probably because his mother home-schooled him for the first two years. Hilbert did not enjoy his school years because he thought that there was too much memorization. To him this was not interesting. What he did enjoy was mathematics. Hilbert thought the subject was easy because he did not have to memorize the answers. He could always simply figure it out. In the fall of 1880 he began his collegiate life at the University of Königsberg. Although his father wanted him to study law, Hilbert chose the only subject that truly interested him. It was traditional at the time in Germany to study at various schools, but Hilbert left Königsberg for only one semester. He much preferred to remain close to his home. After his coursework was completed, Lindemann, who was then at Königsberg, suggested Hilbert study invariants for his dissertation. He defended on December 11, 1884 (Reid 1996, 1, 3–4, 6–7, 9, 15).

After hearing Hilbert's first presentation at the *Technische Hochschule*, Klein realized that he would become a leading mathematician, so Klein suggested to Hilbert that he should travel to Paris as he had done 15 years earlier. Hilbert left in the early spring of 1886. While in Paris he met many of the leading French mathematicians and wrote back to Klein detailing his time there. He met Jordan, who he thought was a very pleasant person. He also met Poincaré; Hilbert's impression of the mathematician was that he was brilliant but quite shy. He also thought that Poincaré's tremendous output was due partly to his publishing even the smallest of results. Nonetheless, Hilbert benefited from the contacts he made in Paris and from the mathematics that he learned (Reid 1996, 19–20).

Göttingen

In 1886, Klein made the last move of his career. It was to the university of Gauss and Riemann at Göttingen. In June of that year Hilbert returned to Germany and made a stop at Klein's new home to make his last report in person. Hilbert did not stay long but returned to his home in Königsburg. By July he finished his *Habilitationsschrift*, which had been put on hold because of his travels. Hilbert was now a *privatdozent* at his hometown university, and he was pleased about his situation (Reid 1996, 25, 27).

Over the next several years the two mathematicians took different paths. Hilbert pushed ahead with his research. His main field of interest was invariant theory, a subject that Hilbert would eventually exhaust. Klein, on the other hand, had seen the last of his productive years as a mathematician. However, Klein was not the type to be inactive. He threw himself into other projects. For one, he was determined to make the university at Göttingen the leader in mathematical and scientific research. He began by changing the atmosphere of the university. He established a reading room on the third floor of the main hall as he had done in Leipzig. Students were encouraged to spend time there to discuss their work or freely take books from the shelves. In one of the hallways Klein displayed numerous mathematical models. This hall was famous among the students, so much so that they called it the "room of models," despite its being a simple hallway. To complete his plan, Klein then moved to recruit many of the top mathematicians to Göttingen. In 1894 a position opened and Klein made it clear that he wanted Hilbert to fill it. The search committee agreed and Hilbert accepted (Reid 1996, 45–46, 48; Yaglom 1988, 135).

At Göttingen Hilbert continued his mathematical pursuits and taught classes. It was here where the two brilliant mathematicians differentiated themselves from each other in a very public way. Hilbert was of the habit of preparing only broad outlines for his lectures and would then develop the mathematics in front of his students. He would often make errors when working the details, but he did not mind because he was showing his students how a mathematician worked. Hilbert's content was direct and to the point, and he would often repeat himself in an attempt to make sure that all of his students understood. Although some of his colleagues thought that his style was better suited for the gymnasium, most of his students appreciated the "beautiful insights" that they gleamed from the time in his classes.

Klein's approach was different. When it came to his lectures, Klein was a perfectionist. He would arrive at the room an hour early to mentally prepare his presentation. His assistant would have prepared for him a list of references. Klein would check these and use them all. By the time the lecture had started, he had prepared the entire lecture, including all of the formulas and diagrams that were required. By the end the chalkboard would be full of mathematics. Klein believed that he should present only an outline of the topic because he thought that students should spend a good four hours working on the mathematics outside his one-hour lecture (Reid 1996, 48–49).

New Postulates

During the 1898–1899 academic year, Hilbert surprised his students by giving a sequence of lectures on geometry. They had known him as an expert in invariant theory or his newfound interest of number theory, but not in geometry. Hilbert did have experience in the subject and had attended lectures on modern viewpoints that went beyond Lobachevski, Poncelet, and Riemann. Although inspired by their work, contemporary mathematicians such as Moritz Pasch had reduced the subject to pure logic, and the mathematician Giuseppe Peano took Pasch's work and translated it into his new symbolic notation. Essentially, in their hands geometry had become an abstract study not of points and lines but of sets of any kind that obeyed certain rules. Although Hilbert agreed with this new view, he realized that the new notation made it hard to grasp. Hilbert solved the problem by using the old Euclidean vocabulary when dealing with the new abstractions. The terms *point* and *line* simply became names for the objects and remain undefined. These objects could be anything as long as they could be interpreted to obey the postulates. Hilbert once clarified this viewpoint by stating that

> [o]ne must be able to say at all times — instead of points, straight lines, and planes — tables, chairs, and beer mugs.

Nonetheless, the familiar terms of *line* and *point* were preferred because they served as a mnemonic device for his students, making the new viewpoint easier to learn (Reid 1996, 57, 59, 61; Boyer and Merzbach 1991, 608).

Hilbert then took his geometric understanding to its logical conclusion. If the words *point* and *line* were going to be reexamined, why should not the same be done to Euclid's entire geometry? Despite his *Elements* remaining at the center of mathematics for over two millennia, Euclid's work was not without critiques. Gauss observed that Euclid had assumed betweenness properties in his proofs, and Leibnitz noted that he had assumed continuity. Neither of these are surprising when geometric proofs rely heavily on diagrams. Both betweenness and continuity are easily assumed from pictures. There is nothing wrong with these assumptions. The problem was that Euclid did not include them among his postulates.

To remedy matters, Hilbert introduced a new set of postulates and then proceeded to reprove the basic results of the first book of the *Elements*. As was standard, the number of postulates should be as few as possible, but they need not be self-evident because the terms *point* and *line* could represent anything. Nonetheless, the sentences chosen should use familiar terminology. More importantly, the postulate set must satisfy three criteria:

- The set must be independent.
- The set must be consistent.
- The set must be complete.

Although we saw the concepts of independence and consistency as early as Saccheri's work, Hilbert is responsible for making these criteria desirable for all modern axiom sets (Kline 1972, 102; Reid 1996, 63–64).

These were the ideas that Hilbert introduced to his students. His lectures became the basis for his *Foundations of Geometry*. It begins with these words,

> Geometry, like arithmetic, requires only a few and simple principles for its logical development. These principles are called the **axioms** of geometry. The establishment of the axioms of geometry and the investigation of their relationships is a problem which has been treated in many excellent works of the mathematical literature since the time of Euclid. This problem is equivalent to the logical analysis of our perception of space. (Hilbert 1992, 2)

Hilbert then begins to set up his system. He takes two types of objects. He called them **points** and **lines**. Denote points by uppercase Latin letters and lines by lowercase Latin letters. The points and lines exist in a **plane**.

When describing relationships among points and lines, it is common to use such words as **lie**, **between**, and **congruent**. For example,

- The point P *lies* on the line l.
- The point A is *between* the points B and C.
- The segments $\overline{AB}$ and $\overline{CD}$ are *congruent*.

These relationships will also be undefined, yet we will have some initial information about them because the axioms will codify our intuition concerning these relationships. The **incidence axioms** will explain *lie*, the **order axioms** will explain *between*, and the **congruence axioms** will explain *congruent*. There are two more sets of axioms: the singleton **parallel axiom**, which is a proposition equivalent to Euclid's Parallel Postulate, and the **continuity axioms**, which identify the principles that prevent Euclidean geometry from being discrete.

The incidence axioms determine what it means for one geometric object to lie on, pass through, intersect, or join another object (Hilbert 1992, 3–4). Compare them with the projective postulates of Sections 9.2 and 14.1.

■ **AXIOMS 15.5.1 [Incidence, *Foundations* §2]**

1. For every two points A and B, there exists a line l that contains each of the points A and B.
2. For every two points A and B, there exists no more than one line that contains each of the points A and B.
3. There exists at least two points on a line. There exist at least three points that do not lie on a line.

It is interesting that Hilbert stated the first two incident axioms as separate statements instead of combining them into one. However, it is reasonable to state the results as he did. Although we have not seen any geometry that does not satisfy the first, there is a model of the elliptic geometry of Riemann that does not satisfy the second (Section 13.4). However, both axioms together quickly yield the expected result.

■ **THEOREM 15.5.2**

For every two points A and B, there exists exactly one line that contains each of the points A and B.

Notice that $\mathbb{R}^2$ satisfies the Incidence Axioms, but there are other significantly smaller models that do the same thing. By Axiom 15.5.1.3 we arrive at a minimum of three points, A, B, and C. Axiom 15.5.1.1 guarantees that we have $\overline{AB}$, $\overline{AC}$, and $\overline{BC}$. If these are all the same line, then by Axiom 15.5.1.3, there exists at least three more points. However, it may be the case that $\overline{AB}$ and $\overline{AC}$ are different lines, so that the same axioms requires only two extra points, say D and E. Hence, only five points are required to satisfy Axioms 15.5.1.1 to 15.5.1.3. This is illustrated by the model

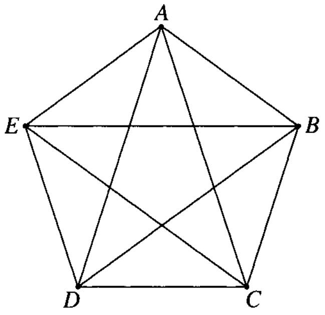

Figure 15.10 A finite model for the Incidence Axioms.

shown in Figure 15.10. Although the segments found inside the pentagon appear to intersect, they do not. Segments in this model intersect only at the points labeled.

The order axioms give meaning to the word *between*. We will write A–B–C to mean that B lies between A and C.

■ AXIOMS 15.5.3 [Order, *Foundations* §3]

1. If A–B–C, the points A, B, and C are three distinct points lying on a line, and C–B–A.
2. For two points A and B, there is a point C on $\overline{AB}$ produced such that A–B–C.
3. Of any three points on a line, there exists no more than one that lies between the other two.
4. **[Pasch]** In $\triangle ABC$, if a line intersects $\overline{AB}$ in a point D such that $A - D - B$, the line also intersects exactly one of the other sides in a point E such that either $A - E - C$ or $B - E - C$.

The order axioms allow us to conclude that points on lines behave as we expect

■ THEOREM 15.5.4 [*Foundations* §4.3]

For any two points A and C, there exists a point B on $\overline{AC}$ such that A–B–C.

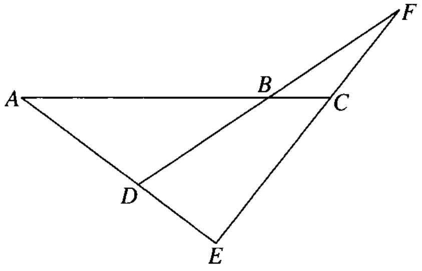

PROOF

Suppose that A and C are two points. Let D be a point not on $\overline{AC}$ by Axiom 15.5.1.3. Appealing to Axiom 15.5.3.2, we know that there is a point E such that A–D–E, and by the same axiom there is a point F such that E–C–F. If $\overline{DF}$ intersected $\overline{EC}$, Axiom 15.5.1.2 would be violated. Hence, $\overline{DF}$ intersects side $\overline{AC}$ at a point B by Pasch's Axiom (15.5.3.4). ■

If points A, B, and C are collinear, Axiom 15.5.3.3 prevents us from writing any two of A–B–C, B–A–C, or A–C–B at the same time, but none of the axioms state that one of these options must hold. Instead, this is a theorem. Its proof is important to read very carefully. Do not rely on results that seem to appear in the diagram but have yet to be proven.

■ THEOREM 15.5.5 [*Foundations* §4.4]

Of any three collinear points A, B, and C, there exists one that lies between the other two.

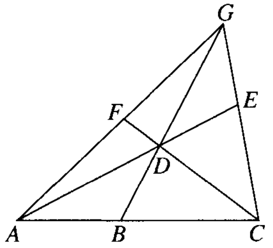

PROOF

Suppose that neither B–A–C nor A–C–B. Take D not on $\overline{AC}$ and find G such that B–D–G (Axioms 15.5.1.3 and 15.5.3.2). Using Pasch's Axiom (15.5.3.4) on $\triangle BCG$, it follows that $\overline{AD}$ intersects $\overline{GC}$ at E, where G–E–C. Notice that at the moment that we do not know if A–D–E. By a similar argument, $\overline{DC}$ intersects $\overline{AG}$ at F with A–F–G. Furthermore, it is still a question whether F–D–C.

Now apply Pasch's Axiom to $\triangle FGC$ using $\overline{AE}$. Since this line intersects $\overline{GC}$ at E and it is impossible for it to intersect $\overline{FG}$ (Axiom 15.5.1.2), we see that $\overline{AE}$ must intersect $\overline{FC}$. Since already $\overline{AE}$ meets $\overline{FC}$ at D, this point must be the point given by Pasch's Axiom. Hence, F–D–C. But, applying this same argument to $\triangle AFC$ with $\overline{DG}$ intersecting $\overline{FC}$ at D, we find that A–B–C. ■

Although the Incidence Axioms can be satisfied by a finite model, that is not the case for the Order Axioms. By a sequence of applications of Theorem 15.5.4, we may prove the next result.

■ THEOREM 15.5.6 [*Foundations* §4.6]

Between any two points on a line there exists a countably infinite number of points.

The next set of axioms deal with the familiar concept of congruence. Euclid wrote none of these as postulates in the *Elements*, but he used them all.

■ **AXIOMS 15.5.7 [Congruence, *Foundations* §5]**

1. If A and B are points on a line l, and A' is a point on the same or on another line l', there exists a point B' on a given side of the line that contains B' such that $\overline{AB} \cong \overline{A'B'}$.
2. If $\overline{A'B'} \cong \overline{AB}$ and $\overline{A''B''} \cong \overline{AB}$, then $\overline{A'B'} \cong \overline{A''B''}$.
3. On the line l, let $\overline{AB}$ and $\overline{BC}$ have no points in common other than B. On a line l' that may or may not be the same line as l, let $\overline{A'B'}$ and $\overline{B'C'}$ have no points in common other than B'. In this case, if $\overline{AB} \cong \overline{A'B'}$ and $\overline{BC} \cong \overline{B'C'}$, then $\overline{AC} \cong \overline{A'C'}$.
4. For any $\sphericalangle ABC$ and $\overline{B'A'}$, there exists a unique $\overline{B'C'}$ on a given side of $\overline{B'A'}$ such that $\sphericalangle ABC \cong \sphericalangle A'B'C'$. Moreover, every angle is congruent to itself.
5. In $\triangle ABC$ and $\triangle A'B'C'$, if
$$\overline{AB} \cong \overline{A'B'},\ \overline{AC} \cong \overline{A'C'}, \text{ and } \sphericalangle BAC \cong \sphericalangle B'A'C',$$
then $\sphericalangle ABC \cong \sphericalangle A'B'C'$.

Hilbert concluded his axiom set by including propositions that are required to make the Euclidean plane have the properties that we expect. That is, the plane should have parallel lines and be continuous. Hilbert accomplished this by including Archimedes' Postulate (6.2.1) and by returning to the definition of parallel used by Euclid. Instead of using the ancient version of the Parallel Postulate, Hilbert chose a modified version of Playfair's Postulate (10.1.1).

■ **DEFINITION 15.5.8**

Two lines are **parallel** if they do not intersect.

■ **AXIOM 15.5.9 [Euclid]**

If l is any line and P a point not on l, then there exists at most one line that passes through P and is parallel to l.

The Model

To show that Hilbert's axioms for Euclidean geometry are consistent, we must find a model in which the axioms can be interpreted as true statements. Because of Theorem 15.5.6, the model must be infinite. A suitable chosen set of rational numbers will suffice, but along with Hilbert we will choose the Cartesian plane $\mathbb{R}^2$.

We interpret the undefined terms *point*, *line*, *incident*, *between*, and *congruent* as follows:

- A *point* is an element of $\mathbb{R}^2$.
- A *line* is an element of $\mathscr{L} = \{[u, v, w] : u, v, w \in \mathbb{R}\}$.
- The point (x, y) is *incident* with line $[u, v, w]$ if and only if
$$ux + vy + w = 0.$$
- The point (x_2, y_2) is *between* (x_1, y_1) and (x_3, y_3) if and only if all three points are different and there exists a line on which all three points are incident and
$$\text{either } x_1 \leqslant x_2 \leqslant x_3 \text{ or } x_1 \geqslant x_2 \geqslant x_3, \text{ and}$$
$$\text{either } y_1 \leqslant y_2 \leqslant y_3 \text{ or } y_1 \geqslant y_2 \geqslant y_3.$$
- $\overline{AB}$ is *congruent* to $\overline{A'B'}$ if and only if there exists an isometry ϕ such that $\phi[\overline{AB}] = \overline{A'B'}$, where for every two points A and B,
$$\overline{AB} = \{C : A - C - B\}.$$
- $\sphericalangle ABC$ is *congruent* to $\sphericalangle A'B'C'$ if and only if there exists a similarity ψ such that $\psi[\sphericalangle ABC] = \sphericalangle A'B'C'$. An angle in the model is two line segments that share exactly one endpoint and are produced indefinitely in one direction.

Call $\mathbb{R}^2$ with these interpretations the **Cartesian model** of Euclidean geometry. To illustrate these interpretations, observe the following:

- $(0, 1)$ is incident with $[2, 4, -4]$ because $(0, 1)$ is a solution to $2x+4y-4 = 0$.
- $(0, 4)$ is between $(0, 1)$ and $(0, 5)$ because these points are on $[1, 0, 0]$, $0 \leqslant 0 \leqslant 0$, and $1 \leqslant 4 \leqslant 5$.
- The line segment with endpoints $(1, 2)$ and $(-3, 5)$ is congruent to the line segment with endpoints $(3, -1)$ and $(-1, 2)$ because the second segment is the image of the first under the translation described by $\langle 2, -3\rangle$.

We will check that some of the axioms hold true under this interpretation and leave the rest for Exercise 5. We begin with an incidence axiom.

> *For every two points A and B, there exists a line l that contains each of the points A and B.*

Let $A = (x_1, y_1)$ and $B = (x_2, y_2)$. Using some basic algebra, we find that
$$y - y_1 = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1).$$
Therefore,
$$(y_1 - y_2)x + (x_2 - x_1)y + x_1y_2 - x_2y_1 = 0,$$
which means that both A and B are on $[y_1 - y_2, x_2 - x_1, x_1y_2 - x_2y_1]$.

Next we examine an axiom of order.

> *For two points A and B, there is a point C on $\overline{AB}$ produced such that A–B–C.*

Let $A = (x_1, y_1)$ and $B = (x_2, y_2)$. We already know that these points are on $[y_1 - y_2, x_2 - x_1, x_1y_2 - x_2y_1]$. There are many cases to consider when finding C. For example, if $0 \leqslant x_1 \leqslant x_2$ and $0 \leqslant y_1 \leqslant y_2$, then

$$C = (2x_2 - x_1, 2y_2 - y_1)$$

is on the line and $A - B - C$ because

$$x_1 \leqslant x_2 \leqslant 2x_2 - x_1$$

and

$$y_1 \leqslant y_2 \leqslant 2y_2 - y_1.$$

Finally, we check a congruence axiom.

> *On the line l, let $\overline{AB}$ and $\overline{BC}$ have no points in common other than B. On a line l' that may or may not be the same line as l, let $\overline{A'B'}$ and $\overline{B'C'}$ have no points in common other than B'. In this case, if $\overline{AB} \cong \overline{A'B'}$ and $\overline{BC} \cong \overline{B'C'}$, then $\overline{AC} \cong \overline{A'C'}$.*

Let $A - B - C$ on l and $A' - B' - C'$ on l'. Assume that $\overline{AB} \cong \overline{A'B'}$ and $\overline{BC} \cong \overline{B'C'}$. Under our interpretation, this means that there exist isometries ϕ and ψ such that

$$\phi[\overline{AB}] = \overline{A'B'} \text{ and } \psi[\overline{BC}] = \overline{B'C'}.$$

Using these two functions, we define

$$\eta(X) = \begin{cases} \phi(X) & \text{if } A - X - B \text{ or } X = A \text{ or } X = B \\ \psi(X) & \text{if } B - X - C \text{ or } X = C. \end{cases}$$

This is the isometry desired.

The models of the hyperbolic and elliptic geometries show that these new systems are consistent if Euclidean geometry is consistent. By using the Cartesian plane as a model of his axioms, Hilbert showed that Euclidean geometry is consistent if arithmetic is consistent. Since mathematicians had little doubt concerning the consistency of arithmetic, Hilbert succeeded in putting all of the geometries on a solid logic footing.

Final Note

Lie, who had been a professor of mathematics at Christiania University, took Klein's place at Leipzig by Klein's recommendation. Three years later, in 1898, Lie received the first International Lobachevski Prize for his geometric work. The prize was established by the Kazan society in 1893 to celebrate the 100th anniversary of

Lobachevski's birth. Later winners would include Klein and Poincaré. In 1898, Lie returned to his home in Norway, and he died the following year on February 18, 1899 (Yaglom 1988, 126–130).

In 1872, Klein began assisting the editors of the journal *Mathematische Annalen* and became its chief editor in 1876. As was typical for Klein, he was instrumental in earning the journal its reputation as the leading mathematical journal. In 1898 he led the team whose task was to published Gauss's collected works. They finished in 1918. His other ventures included the publication of his own books and the lecture notes from the classes he taught at Göttingen (Yaglom 1988, 132–134).

Influenced by his childhood days at the Düsseldorf gymnasium, days that he did not remember with fondness, Klein led efforts to reform the teaching of mathematics. Himself known to be an excellent teacher, in 1898 he organized the International Commission on Mathematics Education. The commission recommended numerous changes that we see still today in the mathematics classroom:

- The use of visual aids
- The emphasis on the study of functions in algebra
- The application of geometric transformations in geometry
- A more unified approach to various mathematical subjects
- A closer connection between mathematics and science education

Such activities, combined with his academic accomplishments, led to his being the representative for Göttingen University in the upper house of the Prussian parliament (Yaglom 1988, 136–137).

Klein lived out his days in Göttingen and died on July 22, 1925 (Yaglom 1988, 137). Under Klein's leadership the university had risen from the mediocrity that Riemann encountered to being a preeminent institute for mathematics. He left the university in good hands. Not only did Hilbert remain to carry on the tradition, but there were other notable mathematicians, including Paul Bernays, Richard Courant, Edmund Landau, Emmy Noether, and Hermann Weyl, in addition to Max Born and James Franck in physics. In 1932, Hilbert celebrated his seventieth birthday, and the next year saw the beginning of the end of what Klein had established at Göttingen (Reid 1996, 191, 203–204).

Exercises

1. Show that the following axiom sets are independent.
   (a) Incidence axioms (15.5.1)
   (b) Order axioms (15.5.3)
   (c) Congruence axioms (15.5.7)
2. Which of Hilbert's axioms can be shown to be consistent with a finite model?
3. Identify all pairs of parallel lines in Figure 15.10.

4. Prove that Hilbert's geometry contains infinitely many lines.

5. Show that Hilbert's axioms hold in the Cartesian model.

6. Prove that congruence of line segments is an equivalence relation.

7. How would we show that a set of axioms is complete?

8. Use Hilbert's concept of betweenness to define the interior and exterior of an angle.

9. Define what it means for two points to be on opposite sides of a line segment. What would it mean for two points to be on the same side of a line segment?

10. Prove the following from Hilbert's *Foundations of Geometry*. Assume the standard definitions for all defined terms.

- (a) For any two points A and B, there exists C collinear with A and B such that C is between A and B.
- (b) Of any three distinct collinear points, there exists one the is between the other two.
- (c) The base angles of an isosceles triangle are congruent.
- (d) SAS
- (e) ASA
- (f) Let $\sphericalangle ABC$ be the supplement of $\sphericalangle CBD$ and $\sphericalangle A'B'C'$ the supplement of $\sphericalangle C'B'D'$. If $\sphericalangle ABC \cong \sphericalangle A'B'C'$, then $\sphericalangle CBD \cong \sphericalangle C'B'D'$.
- (g) Vertical angles are congruent.
- (h) Let $\overline{BD}$ be in the interior of $\sphericalangle ABC$ and $\overline{B'D'}$ be in the interior of $\sphericalangle A'B'C'$. If $\sphericalangle DBC \cong \sphericalangle D'B'C'$ and $\sphericalangle ABC \cong \sphericalangle A'B'C'$, then $\sphericalangle ABD \cong \sphericalangle A'B'D'$.
- (i) If two points A and C are put on different sides of $\overline{BD}$ and if $\overline{AD} \cong \overline{CD}$ and $\overline{BA} \cong \overline{BC}$, then $\sphericalangle ABD \cong \sphericalangle CBD$.
- (j) SSS
- (k) If $\sphericalangle A' \cong \sphericalangle A$ and $\sphericalangle A'' \cong \sphericalangle A$, then $\sphericalangle A' \cong \sphericalangle A''$.
- (l) All right angles are congruent to each other.
- (m) The exterior angle of a triangle is greater than any interior angle that is not adjacent to it.
- (n) In every triangle the greater angle lies opposite the greater side.
- (o) A triangle with two congruent angles is isosceles.
- (p) AAS
- (q) Every segment can be bisected.

Chapter 3 - Part 2

Volume-04---20211110

!"#' $ % % & ' ( )) * +) , -)
.) ( )) / !.! % 0 12

"(3 , -) ( )) ( ))
/ !.! % 0 12

*) 45 ( ) 6 5 7 8 5 7 9 5 % 66

!"#' $ % % & ' ( )) * +) , -)
.) ( )) 8+)+ 1 :)) + - + / !.! % 0 12

"(3 , -) ( )) 8+)+ 1 :
)) + - + ( )) 8+)+ 1 :)) + - + / !.! % 0 12

*) 5 ( ) 64 5 7 8 ; 5 % 45 < % 1) =

>- ; /6 0

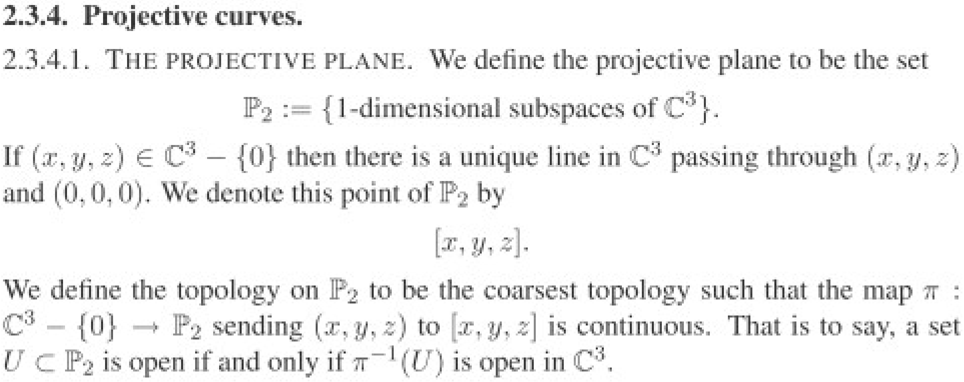

2.3.4. Projective curves.

2.3.4.1. THE PROJECTIVE PLANE. We define the projective plane to be the set

$$\mathbb{P}_2 := \{\text{1-dimensional subspaces of } \mathbb{C}^3\}.$$

If $(x, y, z) \in \mathbb{C}^3 - \{0\}$ then there is a unique line in $\mathbb{C}^3$ passing through (x, y, z) and $(0, 0, 0)$. We denote this point of $\mathbb{P}_2$ by

$$[x, y, z].$$

We define the topology on $\mathbb{P}_2$ to be the coarsest topology such that the map $\pi : \mathbb{C}^3 - \{0\} \to \mathbb{P}_2$ sending (x, y, z) to $[x, y, z]$ is continuous. That is to say, a set $U \subset \mathbb{P}_2$ is open if and only if $\pi^{-1}(U)$ is open in $\mathbb{C}^3$.

!"#' $ % % & ' ( )) * +) , -)
.) ( )) * +) , -) .) / !.! % 0 12

"(3 , -) ; ( )) * +) , -)
.) ( )) * +) , -) .) / !.! % 0 12

*) ; 5 ( ) =5 7 8 5 7 9 5 % ;5 < % 1) ;

!"#' $ % % & ' ( )) * +) , -)
.) # 1)) 2 $) #) / !.! % 0 12

"(3 , -) # 1)) 2 $) #
) # 1)) 2 $) #) / !.! % 0 12

*) 5 ( ) 5 7 8 ;5 7 9 5 % 5 1 < %

!"#' $ % % & ' ( )) * +) , -)
.) # : 8 ' , #) " 2 !) * + / !.! % 0 12

"(3 , -) 6 # : 8 ' , #) " 2
!) * + # : 8 ' , #) " 2 !) * + / !.! % 0 12

*) =5 ( ) 5 7 8 5 7 9 5 % 4= 5 1 < % %)

>- / ; 0 ( ) %)
) ?
>- 4 / 0

!"#" $ % % & ' ( )) * +) , -)
.) " - 8) @))) 2 / !.! % 0 12

"(3 , -) = " - 8) @))) 2
" - 8) @))) 2 / !.! % 0 12

,8 A %) BBB

>- / = 0 4 / 0 / 1 : $ 8 0

!"#" $ % % & ' ( )) * +) , -)
.) %)) + / !.! % 0 12

"(3 , -) 4 %)) + %)) +
/ !.! % 0 12

*) ;5 ( ) 4;5 7 8

,8
A % 2) %)2 C >) 2 D)

!"#" $ % % & ' ( )) * +) , -)
.)) 2 . 2 / !.! % 0 12

"(3 , -)) 2 . 2) 2
. 2 / !.! % 0 12

,8)BBB

!"#" $ % % & ' ( )) * +) , -)
.) E) @) . 1 " - 8) ! 1 ") + / !.! % 0 12

"(3 , -) E) @) . 1 " - 8)
! 1 ") + E) @) . 1 " - 8) ! 1 ") + / !.! % 0 12

,8)BBB

, %

"(3 , -) * +) 8 2- F % +
* +) 8 2- F % + / 0 12

,8 A % 2 BBB A % 2 8))) >

"(3 , -) #) -)GH 2) !))))
-) 2 2 8)) >) #) -)GH 2) !)))) -) 2 2 8)) >) 12

#? ; ;

Chapter 12

Euclidean Space Tangent at a Point

63. Euclidean tangent metric. Consider a Riemannian manifold defined as a manifold endowed with an arbitrary line element

$$ds^2 = g_{ij} du^i du^j. \tag{12.1}$$

We shall assume that the right-hand side of (12.1) is a positive definite differential form with coefficients g_{ij} of class C^1.

The metric defined by a Euclidean line element

$$d\sigma^2 = \gamma_{ij} du^i du^j \tag{12.2}$$

constructed with variables u^i is called a *Euclidean tangent to the metric* (12.1) at the point $A(u_0^1, \dots, u_0^n)$ if we have $\gamma_{ij} = g_{ij}$ for $u^i = u_0^i$.

Obviously, there exists an infinity of Euclidean metrics tangent to the metric (12.1) at a given point. In order to show this, it is sufficient, for example, to take

$$\gamma_{ij} = (g_{ij})_0.$$

On the other hand, the set of all these metrics does not depend on the choice of coordinates defining the Riemannian manifold analytically, since under the change of the variables u^i by certain other determined variables v^i, the new values at the point A of the coefficients of the line element only depend on their original values at A. Therefore, the equality $\gamma_{ij} = g_{ij}$ remains valid also for the new variables at the point A.

Rather than saying that we endow the manifold in question with a new (Euclidean) metric, we can say that we represent a Riemannian manifold in the Euclidean space in such a way that in this representation the line element of the Euclidean space becomes $d\sigma^2$. This Euclidean space is called the *Euclidean space tangent at the point A to the given Riemannian manifold.* This terminology is convenient since it provides a visualization.

We can say that there exists an infinity of Euclidean spaces tangent at the point A to a Riemannian manifold in the sense that the line element $d\sigma^2$ depends on an infinite number of arbitrary parameters. However, as we are going to consider in the sequel only geometric properties common to all these spaces, we may quite conveniently say, *the Euclidean space tangent at the point A to the Riemannian manifold.*

64. Tangent Euclidean space. The first geometric notion implied by consideration of the tangent Euclidean space is that of the distance between a point A and a point infinitesimally close to A. The distance is equal to $d\sigma$ or ds. It is the same notion on which the definition of a Riemannian manifold is based.

1°. *The angle between the two directions du^i and δu^i.* In the tangent Euclidean space, the value of the cosine of the angle between two vectors $\mathbf{X}$ and $\mathbf{Y}$ is obtained from the scalar product $\mathbf{X}\cdot\mathbf{Y}$ which coincides with the coefficient of 2λ in the decomposition of the square of the sum $\mathbf{X}+\lambda\mathbf{Y}$:

$$(\mathbf{X}+\lambda\mathbf{Y})^2 = \mathbf{X}^2 + 2\lambda\mathbf{X}\cdot\mathbf{Y} + \lambda^2\mathbf{Y}^2.$$

In the Riemannian manifold one must consider the product

$$\begin{aligned} g_{ij}\ &(du^i + \lambda\delta u^i)(du^j + \lambda\delta u^j) \\ &= g_{ij}(du^i du^j + \lambda^2 \delta u^i \delta u^j) + 2\lambda g_{ij} du^i \delta u^j. \end{aligned}$$

Thus, the required cosine of the angle between two vectors $\mathbf{X}$ and $\mathbf{Y}$ is

$$\cos\varphi = \frac{g_{ij}du^i du^j}{ds\,\delta s}. \tag{12.3}$$

We can be sure in advance that *the right-hand side of this equation does not depend on the choice of a coordinate system*, since the angle between two directions in the tangent Euclidean space with the metric $d\sigma^2$ is determined by the same formula.

Likewise one may define the angle between a p-dimensional plane element and a q-dimensional plane element, etc., if they have a common vertex.

2°. We apply the notion of the angle between directions at a point of a Riemannian manifold in order to obtain the *rectangular coordinates in an orthogonal frame.*

As is known, a positive definite form

$$ds^2 = g_{ij}du^i du^j$$

can be always represented as a sum of squares:

$$ds^2 = (\omega^1)^2 + (\omega^2)^2 + \ldots + (\omega^n)^2, \tag{12.4}$$

where

$$\left\{ \begin{array}{l} \omega^1 = a_1^1 du^1 + a_2^1 du^2 + \ldots + a_n^1 du^n, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \omega^n = a_1^n du^1 + a_2^n du^2 + \ldots + a_n^n du^n. \end{array} \right. \qquad \det(a_i^j) \neq 0,$$

Consider now vectors $\mathbf{I}_i$ in the tangent Euclidean space, whose coordinates X^k satisfy the equations

$$\text{for } \mathbf{I}_1: \quad \begin{cases} a_1^1 X^1 + \ldots + a_n^1 X^n = 1, \\ a_1^2 X^1 + \ldots + a_n^2 X^n = 0, \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ a_1^n X^1 + \ldots + a_n^n X^n = 0; \end{cases}$$

. . . .

$$\text{for } \mathbf{I}_n: \quad \begin{cases} a_1^1 X^1 + \ldots + a_n^1 X^n = 0, \\ a_1^2 X^1 + \ldots + a_n^2 X^n = 0, \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ a_1^n X^1 + \ldots + a_n^n X^n = 1. \end{cases}$$

The expression in the numerator of formula (12.3) is a bilinear form associated with the quadratic form in (12.1). After reduction, the quadratic form in (12.1) takes the form (12.4). Then the bilinear form in the numerator of (12.3) becomes

$$\omega^1(d)\,\omega^1(\delta) + \ldots + \omega^n(d)\,\omega^n(\delta), \tag{12.5}$$

where

$$\omega^i(d) = a_k^i du^k, \quad \omega^i(\delta) = a_k^i \delta u^k.$$

Consider two vectors $\mathbf{I}_1$ and $\mathbf{I}_2$. For the first vector $\mathbf{I}_1$, we have

$$\frac{\omega^1(d)}{dt} = 1, \quad \omega^2(d) = 0, \quad \ldots, \quad \omega^n(d) = 0,$$

and for the second, we have

$$\omega^i(\delta) = 0, \quad \frac{\omega^2(\delta)}{\delta t} = 1, \quad \ldots, \omega^2(\delta) = 0.$$

Substituting these values into bilinear form (12.5), we see that this form vanishes. Thus, the vectors $\mathbf{I}_1$ and $\mathbf{I}_2$ are orthogonal (cf. No. **38**). The remaining identities

$$\mathbf{I}_j \cdot \mathbf{I}_k = \delta_{jk}$$

can be proved in a similar way. Thus, the frame constructed on the n vectors $\mathbf{I}_i$ in the tangent Euclidean space is orthonormal. The components of translations of this frame are the same forms ω^i that were used in the frame definition.

3°. At the point A of the Euclidean tangent space, one can define *vectors* (in an oblique frame, covariant and contravariant), *bivectors, multivectors*, and generally speaking, arbitrary *tensors*. All these can be extended to a Riemannian manifold. Moreover, we can introduce notions concerning the entire curve or the entire surface.

Firstly, since the infinitesimal distance between two infinitesimally close points is known, by means of addition (integration), we obtain the arc length of any curve,

$$s = \int \sqrt{\sum_i (\omega^i)^2}$$

This implies immediately (an idea due to Riemann) the notion of a *straight line*, or *geodesic* as giving an extreme for the distance. Here we merely indicate this possibility in order to return to it later.

Secondly, as is known, the elementary volume of the manifold is defined by the formula

$$d\tau = \omega^1 \wedge \omega^2 \wedge \ldots \wedge \omega^n.$$

65. The main notions of vector analysis. Given a vector field in a Riemannian manifold, it is not yet possible to define the derivative tensor, since, for this, one needs to be able to compare tensors at different points of the manifold. Nevertheless, we can define the gradient of a function and the *curl* of a vector field. For example, consider a scalar field. Let V be a function of a point,

$$V = V(u^1, \ldots, u^n).$$

The differential dV depends linearly on $du^1, du^2, \ldots, du^n$. Thus, this differential can be represented in the form

$$dV = V_1\omega^1 + V_2\omega^2 + \ldots + V_n\omega^n. \tag{12.6}$$

The quantities V_i define the *gradient* of the function V.

Next, consider a field of vectors

$$(X_1, X_2, \ldots, X_n).$$

The tangent tensor form is

$$\widetilde{\omega} = X_1\omega^1 + \ldots + X_n\omega^n.$$

Compute its exterior differential. Since the differential of X is linearly expressed in terms of ω^i, we have

$$d\widetilde{\omega} = X_{ij}\omega^i \wedge \omega^j, \tag{12.7}$$

where X_{ij} is the skew-symmetric tensor called the *curl* of the vector field.

For example, for the expression

$$\widetilde{\omega} = X dx + Y dy + Z dz,$$

the exterior differential is

$$\begin{aligned}
&dX \wedge dx + dY \wedge dy + dZ \wedge dz \\
&= \frac{\partial X}{\partial y} dy \wedge dx + \frac{\partial X}{\partial z} dz \wedge dx + \frac{\partial Y}{\partial x} dx \wedge dy \\
&\quad + \frac{\partial Y}{\partial z} dz \wedge dy + \frac{\partial Z}{\partial x} dx \wedge dz + \frac{\partial Z}{\partial y} dy \wedge dz \\
&= \left(\frac{\partial Z}{\partial y} - \frac{\partial Y}{\partial z}\right) dy \wedge dz + \left(\frac{\partial X}{\partial z} - \frac{\partial Z}{\partial x}\right) dz \wedge dx + \left(\frac{\partial Y}{\partial x} - \frac{\partial X}{\partial y}\right) dx \wedge dy,
\end{aligned}$$

i.e., we obtain the usual form of the curl in the Euclidean space.

We also derive the *divergence*. The flux of a vector field defines the tensor form

$$\widetilde{\omega} = X_1 \omega^2 \wedge \omega^3 + X_2 \omega^3 \wedge \omega^1 + X_3 \omega^1 \wedge \omega^2,$$

where the exterior products $\omega^i \wedge \omega^j$ define certain areas.

The exterior differential has the form

$$d\widetilde{\omega} = a\, \omega^1 \wedge \omega^2 \wedge \omega^3.$$

Here, a is the *divergence* of the vector field.

If we take the flux of the gradient,

$$V_1 \omega^2 \wedge \omega^3 + V_2 \omega^3 \wedge \omega^1 + V_3 \omega^1 \wedge \omega^2, \tag{12.8}$$

and apply exterior differentiation, then we obtain the second differential parameter

$$\Delta_2 V \omega^1 \wedge \omega^2 \wedge \omega^3.$$

EXAMPLE. Consider the line element

$$ds^2 = z^2 dx^2 + z^2 dy^2 + dz^2.$$

Here we have

$$\omega^1 = z\, dx, \quad \omega^2 = z\, dy, \quad \omega^3 = dz.$$

How can we find the second differential parameter?

In order to find the gradient, we differentiate V:

$$\begin{aligned}
dV &= \frac{\partial V}{\partial x} dx + \frac{\partial V}{\partial y} dy + \frac{\partial V}{\partial z} dz \\
&= V_1 \omega^1 + V_2 \omega^2 + V_3 \omega^3 = V_1 z\, dx + V_2 z\, dy + V_3\, dz.
\end{aligned}$$

Thus,

$$V_1 = \frac{1}{z} \frac{\partial V}{\partial x}, \quad V_2 = \frac{1}{z} \frac{\partial V}{\partial y}, \quad V_3 = \frac{\partial V}{\partial z}.$$

These expressions are the components of the gradient. The elementary vector flux is

$$\begin{aligned} &V_1\,\omega^2\wedge\omega^3 + V_2\,\omega^3\wedge\omega^1 + V_3\,\omega^1\wedge\omega^2 \\ &= \frac{\partial V}{\partial x}\,dy\wedge dz + \frac{\partial V}{\partial y}\,dz\wedge dx + \left(z^2\frac{\partial V}{\partial z}\right) dx\wedge dy. \end{aligned}$$

Applying exterior differentiation, we find that

$$\begin{aligned} &\Delta_2 V\,\omega^1\wedge\omega^2\wedge\omega^3 = \Delta_2 V z^2 dx\wedge dy\wedge dz \\ &= \frac{\partial^2 V}{\partial x^2}\,dx\wedge dy\wedge dz + \frac{\partial^2 V}{\partial y^2}\,dy\wedge dz\wedge dx + \left(z^2\frac{\partial V}{\partial z}\right) dz\wedge dx\wedge dy, \end{aligned}$$

i.e., the second differential parameter is

$$\Delta_2 V = \frac{1}{z^2}\left(\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \left(z^2\frac{\partial V}{\partial z}\right)\right). \tag{12.9}$$

The function V is harmonic in this Riemannian manifold if $\Delta_2 V = 0$.

66. Three methods of introducing a connection. Although we were able to extend many notions from the tangent Euclidean space to a Riemannian manifold, there still remain many basic elementary notions that are missing: for example, we need to define the notion of an angle of rotation of a vector field when we move from one point of the manifold to another one.

In general, every geometric notion involving a *scalar* at any point can be generalized easily. The same can be said about a notion involving one vector or several vectors *provided that they have the same initial point.*

It seems as if the divergence of a vector field is an exception. However, in fact, *the elementary vector flux introduces a field at only one point.*

Summarizing, we see that until now a Riemannian manifold was for us a *collection of small pieces of the Euclidean space.* In some sense it was *amorphous,* since we have not yet connected these different pieces in a mutual orientation. Toward this end, we must define the rotation of a frame under a displacement of a point in the manifold. This can be done by different methods.

Riemann began with the postulate according to which a geodesic of the space, an extreme for the arc length, is not only the shortest line but also is the straightest line, i.e., it does not rotate a frame.

Another method is the same one that led us to the Euclidean tangent space: distinguishing an osculating Euclidean metric and extending its connection to a Riemannian manifold.

The third method is axiomatic: combining more or less natural requirements. This may be the most interesting method, since it admits a generalization to spaces with other fundamental groups.

Here we outline the second and the third methods.

67. Euclidean metric osculating at a point. A Euclidean metric osculating a given metric at a point $A(u_0^i)$ is defined by the line element

$$d\sigma^2 = \gamma_{ij} du^i du^j,$$

provided that the coefficients γ_{ij} and their first-order partial derivatives have the same numerical values at the point $A(u_0^i)$ as they have for the given line element.

The existence of osculating Euclidean metrics does not depend on the choice of coordinate system, since for a given change of variables, the new numerical values of the coefficients g_{ij} and their first-order partial derivatives are completely determined if the original values of the same quantities are known. Thus, rather than speak of an osculating Euclidean metric, we would rather say an *osculating Euclidean space.*

First, it is necessary to prove the existence of osculating Euclidean spaces at a given point A of Riemannian manifold.

From formulas (1.9), (1.10), and (5.4) for an oblique frame in the Euclidean space, we have

$$\begin{cases} d\mathbf{M} = \omega^i \mathbf{e}_i, \\ d\mathbf{e}_i = \omega_i^k \mathbf{e}_k, \quad \omega_i^k = \gamma_i{}^k{}_j \omega^j, \end{cases} \tag{12.10}$$

$$g_{ij} = \mathbf{e}_i \cdot \mathbf{e}_j, \quad dg_{ij} = \omega_i^k g_{kj} + \omega_j^k g_{ik}. \tag{12.11}$$

If we choose the so-called *natural frame*, i.e., if we set

$$\omega^i = du^i, \tag{12.12}$$

then

$$d\omega^i = 0,$$

and the first structure equation (II) of No. **26** implies that

$$du^j \wedge \omega_j^i = 0.$$

It follows from these equations that

$$\omega_j^i = \Gamma_j{}^i{}_k du^k, \quad \Gamma_j{}^i{}_k = \Gamma_k{}^i{}_j, \tag{12.13}$$

$$dg_{ij} = (\Gamma_j{}^i{}_h g_{kj} + \Gamma_j{}^i{}_h g_{ik}) du^k. \tag{12.14}$$

This implies immediately that

$$\frac{\partial g_{ij}}{\partial u^l} = \Gamma_i{}^k{}_l g_{kj} + \Gamma_j{}^k{}_l g_{ik}. \tag{12.15}$$

Thus, if at the point $u^i = u_0^i$, the quantities γ_{ij} coincide with $(g_{ij})_0$, and the coefficients $\Gamma_{j\ k}^{\ i}$ of the Euclidean frame at the initial point coincide with the corresponding values $(\Gamma_{j\ k}^{\ i})_0$ of the Riemannian metric for $u^i = u_0^i$, then not only the coefficients g_{ij} of the line element but also their partial derivatives $\frac{\partial g_{ij}}{\partial u^k}$ are respectively equal in both metrics.

This is not difficult to prove. Since from equations (12.10) and (12.13), it follows that

$$\frac{\partial \mathbf{M}}{\partial u^i} = \mathbf{e}_i, \quad \frac{\partial^2 \mathbf{M}}{\partial u^i \partial u^j} = \Gamma_{i\ j}^{\ k} \mathbf{e}_k, \tag{12.16}$$

we need only to select a parametrization of the Euclidean space in such a way that for $u^i = u_0^i$, equations (12.16) are satisfied.

At an arbitrary point O of the Euclidean space, construct an oblique Cartesian coordinate system $(\mathbf{e}_1, \dots, \mathbf{e}_n)$ such that for $u^i = u_0^i$, the equations

$$\mathbf{e}_i \cdot \mathbf{e}_j = (g_{ij})_0 \tag{12.17}$$

hold. With respect to this coordinate system, define the coordinates x^i of an arbitrary (moving) point of the Euclidean space by means of the equations

$$x^i = u^i - u_0^i + \frac{1}{2}\left(\Gamma_{r\ s}^{\ i}\right)_0 (u^r - u_0^r)(u^s - u_0^s), \quad i = 1, \dots, n. \tag{12.18}$$

The line element of the Euclidean space relative to this system of curvilinear coordinates enjoys the desired properties: the coincidence of the values of γ_{ij} and g_{ij} at the point $u^i = u_0^i$ follows from the construction of the initial frame (in particular, from formulas (12.17)), and the coincidence of the values $\Gamma_{j\ k}^{\ i}$ for the Euclidean space and the Riemannian manifold follows from formulas (12.16) by means of equations (12.18). Thus, at the point $u^i = u_0^i$, the numerical values of the functions γ_{ij}, $\frac{\partial \gamma_{ij}}{\partial u^k}$ and g_{ij}, $\frac{\partial g_{ij}}{\partial u^k}$ are equal.

Chapter 13

Osculating Euclidean Space

68. Absolute differentiation of vectors on a Riemannian manifold. All the geometric notions common to the different osculating Euclidean spaces at a point are obviously intrinsic (invariant) geometric properties of the Riemannian manifold. This is the case for all properties depending on numerical values at the point A of the coefficients g_{ij} and $\Gamma_{i\,j}^{\,k}$.

Let us imagine that in the osculating Euclidean space, the point A and the frame (R_0) attached to this point are fixed. Each point M in the Riemannian manifold is represented by a point $\overline{M}$ whose coordinates, with respect to the frame (R_0), are well defined to within an order of infinitesimals *greater than the second.* The frame attached to the point $\overline{M}$ in the Euclidean space, with respect to its magnitude and orientation, is well defined to within some order of infinitesimals *greater than the first.* Moreover, this frame is equal to the frame attached to the point M in the Riemannian manifold. Every vector of the Riemannian manifold attached to M is represented with the same degree of approximation by an equal vector.

In general, if M and N are any two points in a neighborhood of the point A in the Riemannian manifold and if the points representing M and N in the osculating Euclidean space lie inside a sphere centered at A and of radius r, then the scalar product of two vectors representing vectors attached to M and N are well defined to within infinitesimals of order of radius r inclusively.

In a Riemannian manifold, a vector $\mathbf{X}$ attached to a point A and a vector $\mathbf{X}'$ attached to an infinitesimally close point A' are represented in the osculating Euclidean space by two vectors, whose difference is well defined to within an order of infinitesimals greater than the first. This leads us to extend to an arbitrary Riemannian manifold the notion of the *absolute differential* (or the *covariant differential for the natural frame*) of a vector, or more generally, of a tensor. We shall continue for this absolute differential the same notation D which we used earlier.

From equations (7.17), the absolute differential is

$$DX^i = dX^i + X^k\omega_k^i \tag{13.1}$$

for a contravariant vector (X^i), and

$$DX_i = dX_i - X_k \omega_i^k \tag{13.2}$$

for a covariant vector (X_i). In an orthonormal frame, both definitions coincide, since then

$$X^k = X_k \text{ and } \omega_i^{\cdot k} = \omega_{ik} = \omega^i_{\cdot k} = -\omega_k^{\cdot i}.$$

The minus sign appears because of the change of the order of the indices.

It is important to note that in the Euclidean geometry, the absolute differential is the true differential, while in the geometry of Riemannian manifolds, one cannot be sure that this is the case.

In particular, denoting by $\mathbf{e}_i$ (see No. **4**) the coordinate vectors, we have

$$D\mathbf{e}_i = \omega_i^k \mathbf{e}_k.$$

DEFINITION. Two vectors attached to the two infinitesimally close points A and A' are said to be *equipollent*[1] if the absolute differential of the first vector (when this first vector is transported to the second vector) is equal to zero.

Thus, the conditions for equipollence are

$$dX^i + X^k \omega_k^i = 0, \tag{13.3}$$

or

$$dX_i - X_k \omega_i^k = 0. \tag{13.4}$$

69. Geodesics of a Riemannian manifold. It is not difficult now to define the *acceleration* of a moving point in the Riemannian manifold. If $\mathbf{v}$ is the velocity vector, then the acceleration vector is $\frac{D\mathbf{v}}{dt}$. For example, in order to write the contravariant components γ^i of the acceleration vector with coordinates (u^i) in the natural frame, note that from (12.12), (12.13), and (12.14), it follows that the velocity vector is

$$\mathbf{v} = \frac{D\mathbf{M}}{dt} = \frac{\omega^i}{dt}\mathbf{e}_i = \frac{du^i}{dt}\mathbf{e}_i, \quad v^i = \frac{du^i}{dt}.$$

Applying equations (13.1) to the components v^i and taking into account (12.13), we have

$$\frac{Dv^i}{dt} = \frac{d^2u^i}{dt^2} + \frac{du^i}{dt}\omega_k^i = \frac{d^2u^i}{dt^2} + \frac{du^k}{dt}\Gamma^i_{kh}\frac{du^h}{dt},$$

and hence

$$\gamma^i = \frac{d^2u^i}{dt^2} + \Gamma^i_{kh}\frac{du^k}{dt}\frac{du^h}{dt}. \tag{13.5}$$

A moving point whose acceleration is zero throughout has the velocity which is constantly equipollent to itself. The lines of a Riemannian manifold, whose tangent remains equipollent to itself throughout, are called the *geodesics* (the "straight lines") of the manifold. If a point moves along such a curve and has its velocity equal to 1, then the acceleration is equal to zero. It follows that the equations of geodesics in Riemannian manifolds have the form

---

[1] In modern terminology, the term *parallel* is used. (*Translator's note.*)

$$\frac{d^2u^i}{dt^2} + \Gamma^i_{kh}\frac{du^k}{dt}\frac{du^h}{dt} = 0. \tag{13.6}$$

These equations comprise Cauchy's system for the unknown functions u^i and the independent variable t. Under the hypotheses of Cauchy's theorem, with initial conditions

$$u^i = (u^i)_0, \quad \frac{du^k}{dt} = (v^k)_0 \ \text{ for } t = t_0, \ \ (u^i)_0, (v^k)_0 \ \text{ are constants,} \tag{13.7}$$

such a system has one and only one solution. If we note that the first condition of (13.7) defines an initial point, and the second defines an initial velocity, we obtain the following result.

Theorem 1. *Through a point of a domain of regularity of a Riemannian manifold and in any direction, there passes one and only one geodesic.*

70. Generalization of the Frenet formulas. Curvature and torsion. *The theory of curvature of lines* can be extended without modification from Euclidean space to any Riemannian manifold.

In fact, let M be a point on a curve, $\mathbf{T}$ be a unit tangent vector to the curve at the point M, and ds be the arc element of the curve. The vector $\frac{D\mathbf{T}}{ds}$ is normal to the vector $\mathbf{T}$. This can be seen immediately if one uses the Euclidean space osculating at the point M to the Riemannian manifold.

Suppose that

$$\frac{D\mathbf{T}}{ds} = \frac{1}{\rho}\mathbf{N},$$

where $\mathbf{N}$ is a unit vector (*principal normal*), and $\mathbf{B}$ is the unit vector (the *binormal*), which is orthogonal to $\mathbf{T}$ and $\mathbf{N}$ and forms a right trihedron with them. Suppose also that

$$\frac{D\mathbf{N}}{ds} = \alpha\mathbf{T} + \beta\mathbf{N} + \gamma\mathbf{B},$$

$$\frac{D\mathbf{B}}{ds} = \alpha'\mathbf{T} + \beta'\mathbf{N} + \gamma'\mathbf{B}.$$

The relations

$$\mathbf{T}\cdot\mathbf{N} = 0, \ \ \mathbf{T}\cdot\mathbf{B} = 0, \ \ \mathbf{N}\cdot\mathbf{B} = 0, \ \ \mathbf{N}^2 = 1, \ \ \mathbf{B}^2 = 1$$

on absolute differentiation, yield the following equations:

$$\frac{1}{\rho} + \alpha = 0, \ \ \alpha' = 0, \ \ \gamma + \beta' = 0, \ \ \beta = 0, \ \ \gamma' = 0.$$

Setting now

$$\gamma = -\beta' = \frac{1}{\tau},$$

we obtain the generalized Frenet formulas

$$\begin{cases} \dfrac{D\mathbf{T}}{ds} = \dfrac{1}{\rho}\mathbf{N}, \\ \dfrac{D\mathbf{N}}{ds} = -\dfrac{1}{\rho}\mathbf{T} + \dfrac{1}{\tau}\mathbf{B}, \\ \dfrac{D\mathbf{B}}{ds} = -\dfrac{1}{\tau}\mathbf{N}. \end{cases} \tag{13.8}$$

The quantities $\frac{1}{\rho}$ and $\frac{1}{\tau}$ in these formulas are respectively the *curvature* and *torsion* of the curve. The "straight lines" (geodesics) of the Riemannian manifold are curves of zero curvature. As for the curves of zero torsion, they are characterized by the following properties:

1°) The osculating plane element at a point M of the curve is parallel to the osculating plane element at an infinitesimally close point M'.

2°) The vectors

$$\mathbf{T} + D\mathbf{T}, \quad \mathbf{N} + D\mathbf{N},$$

that are the unit vectors taken along the tangent and the principal normal to the curve at the point M' become

$$\mathbf{T} + \frac{ds}{\rho}\mathbf{N}, \quad \mathbf{N} - \frac{ds}{\rho}\mathbf{T} + \frac{ds}{\tau}\mathbf{B}.$$

These new vectors lie in the osculating plane element at the point M if and only if *the torsion* $\frac{1}{\tau}$ vanishes.

If we represent the Riemannian manifold in the osculating Euclidean space at the point M, then the image of the given curve (C) is a certain curve (Γ) *having the same curvature at the point M as (C)*. Moreover, the vectors $\mathbf{T}, \mathbf{N}$, and $\mathbf{B}$ on the curve (Γ) are the images of the similar vectors on the curve (C). The following explanation makes this even more clear. *Whether we adapt the given metric or an osculating Euclidean metric at the point M, the curve (C) has the same tangent, principal vector, binormal, and the same curvature as (Γ). But (C) does not have the same torsion with the two metrics.* In particular, for an observer who would adapt an osculating Euclidean metric at M, every straight line (every geodesic) in the Riemannian manifold represents a point of inflection at M.

71. The theory of curvature of surfaces in a Riemannian manifold. The classical theory of curvature of surfaces can be extended to Riemannian manifolds quite easily.[2] The different curves on a surface (S) passing through a given point M of (S), have the same principal normal and even the same curvature in both metrics, the metric of the Riemannian manifold or the osculating Euclidean metric, adapted to M. It follows that the laws which govern the variation of the curvature of these curves, when their tangents at the point M are varied, are the same as for the Euclidean space.

All curves tangent to one another have the same *normal curvature.* This normal curvature is equal to $\frac{\cos V}{\rho}$, where V is the angle of the principal normal to the curve with the surface normal, and ρ is the radius of curvature of the curve. This result is *Meusnier's theorem.*

There exist two orthogonal directions tangent to the surface corresponding to the maximum and minimum normal curvature. These directions are called the *principal directions.* The normal curvatures corresponding to these directions are the *principal curvatures.*

The *curvature lines* are the lines which, at each of their points, are tangent to one of principal directions at that point. If θ is the angle formed by a curve with the curvature line of the first family, then we have for this curve

$$\frac{\cos V}{\rho} = \frac{\cos^2\theta}{R_1} + \frac{\sin^2\theta}{R_2}, \tag{13.9}$$

where $\dfrac{1}{R_1}$ and $\dfrac{1}{R_2}$ are the two principal curvatures.

The *asymptotic curves* are those whose normal curvature is zero. The osculating plane element of such curves is tangent to the surface. The straight line (the geodesic) of the Riemannian manifold tangent to an asymptotic line at a point M has a second-order tangency at this point with the surface.

The curvature lines can also be characterized by the following condition: for any given point M on the curve, the tangent to the curvature line at M, the surface normal at M, and the surface normal at an infinitesimally close point M' of this line, once transported by parallelism from M' to M, lie in the same plane element.

Denote by ν the unit vector of the surface normal, and by $\mathbf{T}_1$ and $\mathbf{T}_2$ the unit vectors of the tangents to the curvature lines. Then, for a displacement in the direction of the vector $\mathbf{T}_1$ with length ds_1, we have

$$D_1\nu = -\frac{ds_1}{R_1}\mathbf{T}_1.$$

Then, for a displacement in the direction of the vector $\mathbf{T}_2$ with length ds_2, we have

$$D_2\nu = -\frac{ds_2}{R_2}\mathbf{T}_2.$$

These formulas represent a generalization to the Riemannian manifolds of the classical *Olinde Rodriques equations.*

---

[2] The theory of curvature of surfaces in the Euclidean space is the subject of Chapter 23.

For a displacement with length ds in the direction making an angle θ with the principal direction $\mathbf{I}_1$, in general we have

$$\frac{D\nu}{ds} = -\frac{\cos\theta}{R_1}\mathbf{T}_1 - \frac{\sin\theta}{R_2}\mathbf{T}_2. \tag{13.10}$$

Finally, the total curvature $\frac{1}{R_1R_2}$ of the surface at a point M could be defined by the method applied by Gauss to the Euclidean space. Take a surface element $d\sigma$ enclosing the point M and the sphere of center M and of radius 1 in the Euclidean space tangent at M to the Riemannian manifold. Transport by parallelism to the point M the surface normals at the different points of the element $d\sigma$, and consider a small region of the sphere defined by the intersection with the sphere of the normals so transported. The limit of the ratio $\frac{d\omega}{d\sigma}$ of the area of this portion of the sphere to the given area $d\sigma$ is equal to the total curvature $\frac{1}{R_1R_2}$. We could the consider $d\omega$ as the solid angle of the cone at the vertex M obtained by transporting by parallelism at M the normals emanating from the different points of the element $d\sigma$.

72. Geodesic torsion. The Enneper theorem. The notion of the *geodesic torsion* of a curve on a surface can be extended to the most general Riemannian manifolds.

In fact, denote by $\mathbf{T}, \mathbf{N}$, and $\mathbf{B}$ the unit vectors along the tangent, principal normal, and binormal to the curve respectively, and keep the same notations $\mathbf{T}_1, \mathbf{T}_2$, and ν for the unit vectors along the tangents to the principal directions and the surface normal. Finally, denote by θ the angle formed by the tangent to the curve with the tangent to the first principal direction, and by V the angle of the principal normal to the curve and the surface normal. The direction cosines of the vectors $\mathbf{T}, \mathbf{N}$, and $\mathbf{B}$ with respect to the vectors $\mathbf{T}_1, \mathbf{T}_2$, and $\boldsymbol{\nu}$ are given in Table 13.1:

	$\mathbf{T}_1$	$\mathbf{T}_2$	$\boldsymbol{\nu}$
T	$\cos\theta$	$\sin\theta$	0
N	$-\sin\theta\sin V$	$\cos\theta\sin V$	$\cos\theta$
B	$\sin\theta\cos V$	$-\cos\theta\cos V$	$\sin\theta$

Table 13.1

Let us start with the generalized Frenet formulas

$$\begin{cases} \dfrac{D\mathbf{T}}{ds} = \dfrac{1}{\rho}\mathbf{N}, \\ \dfrac{D\mathbf{N}}{ds} = -\dfrac{1}{\rho}\mathbf{T} + \dfrac{1}{\tau}\mathbf{B}, \\ \dfrac{D\mathbf{B}}{ds} = -\dfrac{1}{\tau}\mathbf{N} \end{cases} \tag{13.11}$$

and the equation

$$\frac{D\nu}{ds} = -\frac{\cos\theta}{R_1}\mathbf{T}_1 - \frac{\sin\theta}{R_2}\mathbf{T}_2.$$

Differentiating the relations

$$\mathbf{T}\cdot\nu = 0, \quad \mathbf{N}\cdot\nu = \cos V, \quad \mathbf{B}\cdot\nu = \sin V,$$

we obtain two distinct equations:

$$\frac{\cos V}{\rho} = \frac{\cos^2\theta}{R_1} + \frac{\sin^2\theta}{R_2}, \tag{13.12}$$

$$\frac{dV}{ds} + \frac{1}{\tau} = \left(\frac{1}{R_1} - \frac{1}{R_2}\right)\sin\theta\cos\theta. \tag{13.13}$$

The left-hand side of equation (13.13) is a quantity called the *geodesic torsion* of the curve. The form of the right-hand side shows that all curves with the same tangent have the same geodesic torsion. The geodesic torsion vanishes for the curvature lines.

If we apply equation (13.13) to an asymptotic line which is not a straight line ($V = \frac{\pi}{2}$), then we obtain

$$\frac{1}{\tau} = \left(\frac{1}{R_1} - \frac{1}{R_2}\right)\sin\theta\cos\theta.$$

But for an asymptotic line, we have

$$\frac{\cos^2\theta}{R_1} + \frac{\sin^2\theta}{R_2} = 0,$$

and this implies that[3]

$$\frac{1}{\tau} = \pm\sqrt{-\frac{1}{R_1R_2}}.$$

---

[3] In fact, we have $\frac{1}{\tau} = \frac{\cos\theta}{R_1\sin\theta}$ or $\frac{1}{\tau^2} = -\frac{1}{R_1R_2}$. (*Translator's note.*)

This is the *Beltrami-Enneper theorem* for an arbitrary Riemannian manifold.

The asymptotic lines emanating from a point M of a surface have at that point torsions equal in absolute values but with opposite signs. The common absolute value of these torsions is equal to the square root of the total curvature of the surface taken with the sign changed.

In particular, the torsion of a doubly-asymptotic line is constantly zero.

It is known that, in Euclidean space, if two families of asymptotic lines coincide, then the asymptotic lines are straight lines. This theorem is no longer true in an arbitrary Riemannian manifold.[4] Nevertheless, by means of the Enneper theorem, we can affirm that they are curves with zero torsion.

73. Conjugate directions. The theory of *conjugate tangents* also generalizes to an arbitrary Riemannian manifold. If M and M' are two infinitesimally close points on a curve (C) traced on a surface (S), then the plane element tangent to the surface (S) at the point M', being transported by parallelism at the point M, has a certain direction, conjugate to the direction of (C) at M, in common with the plane element tangent to (S) at M. Two conjugate directions are harmonically conjugate with respect to the asymptotic directions. Finally, when we pass from M to M', the absolute differential of the unit vector of the surface normal is perpendicular to the conjugate direction of the curve. All these properties become trivial if we represent the Riemannian manifold in the osculating Euclidean space at the point M.

74. The Dupin theorem on a triply orthogonal system. We conclude with the *famous Dupin theorem.* As is known, a triply orthogonal system is a system formed by three one-parameter families of surfaces that intersect at right angles. The Dupin theorem states that the intersecting curves of each pair of surfaces from two different families is a curvature line for each of these surfaces.

The analytic proof that we are going to present can be applied just as well to Riemannian manifolds as to Euclidean spaces.

Since the cosines of the angles between the two intersecting coordinate curves is zero, the line element of our manifold is

$$ds^2 = g_{22}(du^1)^2 + g_{11}(du^2)^2 + g_{33}(du^3)^2.$$

For example, the condition, expressing the fact that the coordinate line (C_1) (u^1 a variable) is a curvature line for the surface (S_1) (u^2 a constant), is that the vector

$$\mathbf{e}_3 + D_1\mathbf{e}_3 du^1,$$

normal to (S_1) at the point M' infinitesimally close to the point M on the line (C_1), must belong to the same plane element as the vectors $\mathbf{e}_3$ and $\mathbf{e}_1$. In other words, the coefficient of $\mathbf{e}_2$ in the expression of $D_1\mathbf{e}_3$ is zero. But we have

---

[4] É. Cartan, *Sur les courbes de torsion nulle et les surfaces développables dans les espaces de Riemann*, Comptes Rendus Acad. Sc. **184** (1927), 138–141. (JFM **53**, p. 696.)

$$D\mathbf{e}_3 = \omega_3^1\,\mathbf{e}_1 + \omega_3^2\,\mathbf{e}_2 + \omega_3^3\,\mathbf{e}_3,$$

$$D_1\mathbf{e}_3 = \Gamma_3{}^1{}_1\mathbf{e}_1 + \Gamma_3{}^2{}_1\mathbf{e}_2 + \Gamma_3{}^3{}_1\mathbf{e}_3.$$

Thus, we need to prove that $\Gamma_3{}^2{}_1 = 0$. But taking equation (12.14) for $i = 1$, $j = 2$, $h = 3$ and applying equations $g_{23} = g_{31} = g_{12} = 0$, we come to the equations

$$\Gamma_1{}^2{}_3\, g_{22} + \Gamma_2{}^1{}_3\, g_{11} = 0,$$

$$\Gamma_2{}^3{}_1\, g_{33} + \Gamma_3{}^2{}_1\, g_{22} = 0,$$

$$\Gamma_3{}^1{}_2\, g_{11} + \Gamma_1{}^3{}_2\, g_{33} = 0,$$

where the last two equations are obtained from the first by cyclic permutation of the indices 1, 2, 3. By (12.13), the coefficients $\Gamma_i{}^k{}_j$ are symmetric with respect to the lower indices. Thus, adding the first two equations and subtracting the third, we obtain

$$2\Gamma_3{}^2{}_1\, g_{22} = 0, \quad g_{22} \neq 0,$$

and as a result, we have

$$\Gamma_3{}^2{}_1 = 0.$$

This page is intentionally left blank

Chapter 14

Euclidean Space of Conjugacy along a Line

75. Development of a Riemannian manifold in Euclidean space along a curve. The development of a Riemannian manifold in Euclidean space along a curve opens up new possibilities.

Suppose that coordinates of a varying point of a curve are expressed in terms of a parameter t which takes the value $t = 0$ at an initial point A_0. In a Euclidean space, take an origin O to which we attach a Cartesian frame R_0 determined in size and shape by the numerical values of the coefficients g_{ij} at the point A_0. With each point $M(t)$, we associate a point M' of the Euclidean space and a Cartesian frame $(\mathbf{e}'_i)$ attached to M'. We shall start with equations (I) of No. **26** and suppose that the frame $(\mathbf{e}'_i)$ is natural. Then we have

$$\begin{cases} d\mathbf{M}' = du^i \mathbf{e}'_i, \\ d\mathbf{e}'_i = \omega_i^j \mathbf{e}'_j \end{cases} \tag{14.1}$$

with the independent variable t and the initial conditions: for $t = 0$, the point M' coincides with the point O, and the vector $(\mathbf{e}'_i)$ coincides with the vector $(\mathbf{e}_i)$ of the frame (R_0). As in No. **28**, we can prove that in the frame $(R) = (\mathbf{e}'_i)$ attached to the point $M'(t)$, the scalar products $\mathbf{e}'_i \cdot \mathbf{e}'_j$ of its vectors are equal to the coefficients g_{ij}. Furthermore, if M and M_1 are arbitrary infinitesimally close points on a given curve, and M' and M'_1 are their images in the Euclidean space, then two equipollent vectors at the points M and M_1 have for their images two equipollent (simply parallel) vectors with at M' and M'_1.

Since a Cartesian frame (R) is attached to each point M', in fact, we obtain a development in Euclidean space not only of the given line but also of an entire infinitesimally small neighborhood of this line in the Riemannian manifold. In order to get the absolute geometric variation of a vector, whose initial point describes the arc of the given line in the Riemannian manifold, *it suffices to* *construct the ordinary geometric difference between the two vectors obtained in the preceding development.*

Thus, we have obtained a rigorous and precise realization of this absolute geometric variation. Now we also know precisely what is meant by "transport by the principle of equipollence[1] of a vector along a given path."

76. The constructed representation and the osculating Euclidean space. We prove now that the development obtained at each point of the curve enjoys the properties of an osculating Euclidean space. This means that we prove that along the curve, the constructed Euclidean line element with the variables u^i has the same numerical values of the coefficients g_{ij} and their first-order partial derivatives as those of the given line element.

Consider the operation of development in the Euclidean space in more detail. For simplicity, we take $n = 3$, and without loss of generality, assume that the line is defined by the equations $u^1 = 0$, $u^2 = 0$.

The development in the Euclidean space gives us the point M' and the vectors $\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3$ as functions of u^3. By means of equations (14.1), we obtain

$$\left\{\begin{aligned} &\frac{d\mathbf{M}'}{du^3} = \mathbf{e}'_3, \\ &\frac{d\mathbf{e}'_i}{du^3} = \Gamma_{i}{}^{k}{}_{3}\mathbf{e}'_k. \end{aligned}\right. \tag{14.2}$$

Next we find, in terms of functions u^1, u^2, and u^3, a point P of the Euclidean space satisfying the following initial conditions: for $u^1 = 0$, $u^2 = 0$,

$$\left\{\begin{gathered} \mathbf{P} = \mathbf{M}', \\ \frac{\partial \mathbf{P}}{\partial u^1} = \mathbf{e}'_1, \quad \frac{\partial \mathbf{P}}{\partial u^2} = \mathbf{e}'_2, \\ \frac{\partial^2 \mathbf{P}}{(\partial u^1)^2} = (\Gamma_{1}{}^{k}{}_{1})^* \mathbf{e}'_k, \quad \frac{\partial^2 \mathbf{P}}{\partial u^1 \partial u^2} = (\Gamma_{1}{}^{k}{}_{2})^* \mathbf{e}'_k, \quad \frac{\partial^2 \mathbf{P}}{(\partial u^2)^2} = (\Gamma_{2}{}^{k}{}_{2})^* \mathbf{e}'_k. \end{gathered}\right. \tag{14.3}$$

where $(\Gamma_{i}{}^{k}{}_{j})^*$ are the values of these coefficients for $u^1 = 0$, $u^2 = 0$.

Of course, these conditions are compatible with each other. It suffices, for example, to take

$$\begin{aligned} \mathbf{P} = \mathbf{M} + u^1\mathbf{e}'_1 + u^2\mathbf{e}'_2 + \frac{1}{2}(u^1)^2(\Gamma_{1}{}^{k}{}_{1})^*\mathbf{e}'_k \\ + u^1u^2(\Gamma_{1}{}^{k}{}_{2})^*\mathbf{e}'_k + \frac{1}{2}(u^2)^2(\Gamma_{2}{}^{k}{}_{2})^*\mathbf{e}'_k. \end{aligned}$$

In this formula, Euclidean space is related to a system of curvilinear coordinates u^1, u^2, u^3. For $u^1 = 0, u^2 = 0$, the natural frame attached to the point P (which reduces here to the point M') is determined by the vectors $\frac{\partial \mathbf{P}}{\partial u^i}$, which reduce here to the vectors $\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3$. The coefficients of the line element ds^2 in the Euclidean space are the coefficients of the given line element of the Riemannian manifold. As for the coefficients $\Gamma_{i\ j}^{\ k}$ of the Euclidean space, they are obtained if one compares the coefficients of $\mathbf{e}'_k$ in the expression $\frac{\partial^2 \mathbf{P}}{\partial u^i \partial u^j}$. If the indices i and j are both distinct from 3, then by applying (14.3), we find the coefficients $\Gamma_{i\ j}^{\ k}$ in the Riemannian manifold. If $i \neq 3$ and $j = 3$, then on account of (14.2) and for $u^1 = 0$, $u^2 = 0$, we find that

---

[1] This is "parallel transfer" or "parallel transport" in modern terminology. (*Translator's note.*)

$$\left(\frac{\partial^2 \mathbf{P}}{\partial u^i \partial u^3}\right)_0 = \frac{d}{du^3}\left(\frac{\partial \mathbf{P}}{\partial u^i}\right)_0 = \frac{d\mathbf{e}'_i}{du^3} = (\Gamma_{i\ 3}^{\ k})^* \mathbf{e}'_k,$$

and the conclusion is the same. Finally, if $i = j = 3$, then

$$\left(\frac{\partial^2 \mathbf{P}}{(\partial u^3)^2}\right)_0 = \frac{d\mathbf{e}'_3}{du^3} = (\Gamma_{3\ 3}^{\ k})^* \mathbf{e}'_k,$$

and again the conclusion is the same.

We see that *the determination of a Euclidean space of conjugacy along a given curve does not require integration, once the development of the curve on the Euclidean space is carried out.*

This result has numerous important consequences. First of all, *an observer, who moves only along the given line and is content to restrict his measurements to an immediate neighborhood of this line, would not be able to see that he is outside a Euclidean space as long as he neglects infinitesimals of an order greater than the first.*

Another important consequence is as follows: if, in a Riemannian manifold, one considers an arc of a curve infinitesimally close to the curve (C), then *the length of this arc, measured by the metric of the Riemannian manifold or by the Euclidean metric of conjugacy along the curve (C), is the same up to the second-order infinitesimals.* [2]

In fact, at a point infinitesimally close to the curve (C), the coefficients of the given line element and the line element of the Euclidean space of conjugacy are equal up to second-order infinitesimals. Thus, *the representation of the Riemannian manifold upon the Euclidean space of conjugacy preserves the distances measured in the neighborhood of the given line up to the same order.*

77. Geodesics. Parallel surfaces. When one develops a curve on a Euclidean space, at each of its points, the developed curve has the same curvature and the same torsion as the given curve, since in development, the absolute differential of a vector, whose initial point describes the curve, becomes the ordinary geometric variation of this vector. The generalized Frenet formulas become, in the Euclidean space of conjugacy, the *usual* Frenet formulas:

---

[2] It is necessary to assume than the two arcs of the curves in question correspond, point by point, in a way that at two corresponding points, the directions of the tangent elements are infinitesimally close.

$$\begin{cases} \dfrac{d\mathbf{T}}{ds} = \dfrac{1}{\rho}\mathbf{N}, \\ \dfrac{d\mathbf{N}}{ds} = -\dfrac{1}{\rho}\mathbf{T} + \dfrac{1}{\tau}\mathbf{B}, \\ \dfrac{d\mathbf{B}}{ds} = -\dfrac{1}{\tau}\mathbf{N}. \end{cases}$$

Thus, a curve with zero torsion is so developed along a plane curve, and a curve with zero curvature is developed along a straight line.

The straight lines (the geodesics) of a Riemannian manifold are therefore the curves whose development is along a straight line. This immediately implies that the straight lines (the geodesics) in our definition are equivalent to the straight lines in Riemann's definition.

In fact, if a curve (C) is developed along a straight line, and in a Riemannian manifold we take a curve (C') infinitesimally close to (C), which emanates from a given point A of (C), and ends at another given point B of (C), then the image, *a curve in the Euclidean space of conjugacy*, has the same length as (C), up to second-order infinitesimals. But in this Euclidean space, the line, which is the image of the curve (C'), results from the variation of a straight line segment. Thus, its length is *the same* as the length of the line segment up to second-order infinitesimals. Consequently, in the Riemannian manifold, the first variation of the length of a segment of a straight line (a geodesic) is identically zero, and the "straight line" (the geodesic) realizes the extremum of the distance conforming to the classical definition of Riemann.

This implies another *consequence*: if we consider a geodesic arc AB and an infinitesimally close geodesic arc $A'B'$, then as *in the Euclidean space*, up to second-order infinitesimals, we have

$$\text{arc } A'B' - \text{arc } AB = -AA' \cos(\widehat{AB, AA'}) - BB' \cos(\widehat{BA, BB'}).$$

In particular, taking a constant length R starting from the point A along the different geodesics emanating from A, we obtain a certain surface similar to the sphere as a region between the extremities. *The tangent plane element to this surface at a point M is normal to the geodesic emanating from A and ending at the point M.* The same applies if we consider a family of geodesics generating the surface from A only. Namely, if we take a constant length on each of these geodesics, then the line element tangent at any point M in the region between extremities is orthogonal to the geodesic emanating from A and ending at M.

One other consequence is the generalization of the property of *parallel surfaces*. If we construct normal geodesics at different points of a surface and take a constant length on each of them, then the region of the points so obtained is a surface which, at each of its points, is normal to the corresponding geodesic. It follows that if a two-parameter family of geodesics is normal to a surface, then it is normal to an infinity of surfaces.

78. Geodesics on a surface. In general, what we did above for a line is impossible to do for a surface. *In general, there is no Euclidean space of conjugacy along a surface.* The reason for this is very simple: in general, it is impossible to develop a surface in Euclidean space. Nevertheless, *if a development in Euclidean space is possible, then there exists a Euclidean space of conjugacy along a surface.*

Note the final application of the Euclidean space of conjugacy. In a three-dimensional Euclidean space, take a surface (S) and *a geodesic* (C) *of this surface*, i.e., a line realizing the minimum of distance *on a surface.* In the Euclidean space of conjugacy along the curve (C), the line (C) has a stationary length with respect to all lines on (S) having their ends at two given points of (C). Thus, by a classical property of geodesics on a surface of the Euclidean space, the osculating plane at any point of the line (C) is normal to the surface. *This property*, which is valid for the Euclidean space of conjugacy, *is also valid for a Riemannian manifold.*

This page is intentionally left blank

C. CURVATURE AND TORSION OF A MANIFOLD

This page is intentionally left blank

Chapter 15

Space with a Euclidean Connection

79. Determination of forms ω_i^j for given forms ω^i. In No. **66** we posed the problem of finding a connection of a Riemannian manifold and outlined certain ways to solve this problem. First, we discovered that this problem is equivalent to the problem of absolute differentiation of vectors and depends on the choice of coefficients $\Gamma_{i\ k}^{\ j}$ (sometimes called the object of connection[1] or the forms ω_i^j).

We made broad use of the notion of the metric osculating at a point and obtained a series of geometric results.

Now we want to investigate to what degree we need the postulate on extending a connection of an osculating metric into a Riemannian manifold.

In order not to complicate the situation by the choice of frame, we assume that an orthonormal frame is assigned to every point of the Riemannian manifold.

We saw (No. **64**) that it is sufficient to reduce a positive definite quadratic form to a sum of squares (this is always possible),

$$ds^2 = (\omega^1)^2 + (\omega^2)^2 + \ldots + (\omega^n)^2, \tag{15.1}$$

in order that the forms ω^i define an orthonormal frame.

Will this determine the forms ω_i^j ?

We saw that in the Euclidean space, the forms ω^i and ω_i^j satisfy structure equations (II) of No. **26**,

$$d\omega^i = \omega^k \wedge \omega_k^i, \tag{II$_1$}$$

$$d\omega_i^j = \omega_i^k \wedge \omega_k^j, \tag{II$_2$}$$

and of course, the orthonormality condition

$$\omega_i^j = -\omega_j^i. \tag{15.2}$$

---

[1] They are now called the connection coefficients. (*Translator's note.*)

Conditions (15.2) followed from our choice an orthonormal frame. So, we need to investigate two equations (II_1) and (II_2). If both of them are satisfied, then the space is Euclidean. Equation (II_2) does not contain the forms ω^i at all, but these forms determine the line element (15.1). Thus, it is natural to search for the forms ω_i^j satisfying the system of equations (II_1) and (15.2).

Theorem 1. *If the forms ω^i are given, then it is possible to find the $\frac{n(n-1)}{2}$ forms ω_i^j satisfying the system of equations* (II_1) *and* (15.2), *and such a solution of this system is unique.*

Proof. We assume that only one frame is associated with each point of the space. Hence the forms ω_i^j are linear combinations of the forms ω^i:

$$\omega_i^j = \gamma_i{}^j{}_k \omega^k. \tag{15.3}$$

We need to find the coefficients $\gamma_i{}^j{}_k$ in such a way that they satisfy the system of equations (II_1) and (15.2).

Exterior differentiation of the given forms ω^i gives us the exterior quadratic form with respect to the basis forms ω^k in the form [2]

$$d\omega^i = c^i_{[jk]}\omega^j \wedge \omega^k, \quad c^i_{[jk]} = c^i_{jk} - c^i_{kj}. \tag{15.4}$$

Then equations (II_1) take the form

$$c^i_{[jk]}\omega^j \wedge \omega^k = \omega^k \wedge \omega^i_k. \tag{15.5}$$

Substituting expressions (15.3) into equations (15.5) and (15.2) and comparing the coefficients in independent products $\omega^j \wedge \omega^k$, we obtain

$$c^i_{[jk]} = \gamma^i_{jk} - \gamma^i_{kj}, \tag{15.6}$$

$$0 = \gamma^i_{jk} + \gamma^j_{ik}. \tag{15.7}$$

Equations (15.6) and (15.7) determine γ^i_{jk} uniquely.

In fact, twice making the cyclic permutation of the indices i, j, k in equations (15.6) and applying relations (15.7), we find that

$$\left\{\begin{aligned} c^i_{[jk]} &= \gamma^i_{jk} + \gamma^k_{ij}, \\ c^j_{[ki]} &= \gamma^j_{ki} + \gamma^i_{jk}, \\ c^k_{[ij]} &= \gamma^k_{ij} + \gamma^j_{ki}. \end{aligned}\right.$$

---

[2] Here and in what follows, it is assumed that for the alternated coefficients $c^i_{[jk]}$ the summation is carried over all combinations (j,k) of indices j and k:

$$c^i_{[jk]}\omega^j \wedge \omega^k = \sum_{(j,k)} c^i_{jk}\omega^j \wedge \omega^k.$$

Adding the first two equations and subtracting the third, we obtain

$$\gamma^i_{jk} = \frac{c^i_{[jk]} + c^j_{[ki]} - c^k_{[ij]}}{2}. \tag{15.8}$$

80. Condition of invariance of line element. Suppose now that the quadratic form ds^2 is decomposed into a sum of squares in the most general way, i.e., with $\frac{n(n-1)}{2}$ parameters

$$v^1, v^2, \ldots, v^{\frac{n(n-1)}{2}}$$

(this is the same as the number of independent forms ω^j_i satisfying equations (15.2)). The difficulty of the problem depends on the fact that although the forms ω^i are linear combinations of $du^1, \ldots, du^n$, the coefficients of these linear combinations can depend on u^i and v^α. Thus, after exterior differentiation the differentials dv^α can occur. By hypothesis, the line element

$$ds^2 = (\omega^1)^2 + (\omega^2)^2 + \ldots + (\omega^n)^2$$

does not depend on v^α. We now discuss how to express this condition.

It is always possible to choose new forms $\widetilde{\omega}^i_k$, which can also depend also on v^α and dv^α, in such a way that

$$d\omega^i = \omega^k \wedge \widetilde{\omega}^i_k, \tag{15.9}$$

or in bilinear form,

$$d\,\omega^i(\delta) - \delta\,\omega^i(d) = \omega^k(d)\,\widetilde{\omega}^i_k(\delta) - \omega^k(\delta)\,\widetilde{\omega}^i_k(d). \tag{15.10}$$

Suppose that the symbol d of differentiation corresponds to differentiation with respect to the variables u^i, and the symbol δ corresponds to differentiation with respect to the variables v^α. Then, by hypothesis, we have

$$\omega^i(\delta) = 0.$$

Denote also

$$\widetilde{\omega}^i_k(\delta) = e^i_k, \quad \omega^i(d) = \omega^i.$$

Then equations (15.10) take the form

$$\delta\,\omega^i = -e^i_k\,\omega^k.$$

Differentiating ds^2 with respect to the parameter v^α, we find that

$$\delta\,(ds^2) = 2\sum_i \omega^i\,\delta\omega^i = -2e_{ki}\,\omega^i\,\omega^k.$$

By hypothesis, $\delta(ds^2) = 0$. Hence,

$$e_{ki}\,\omega^i\,\omega^k = 0.$$

As a result, for any choice of the forms ω^k, all coefficients of $(\omega^i)^2$ and $\omega^i\omega^j (i \neq j)$ vanish,

$$e_{ii} = 0 \;\; (i \text{ not summed}), \;\; e_{ij} + e_{ji} = 0.$$

This proves that for fixed u^i, we have

$$\widetilde{\omega}_i^j(\delta) + \widetilde{\omega}_j^i(\delta) = 0.$$

Therefore, for any change of secondary parameters v^α, this sum depends only on the differentials $du^1, du^2, \dots, du^n$, i.e., it depends on the forms ω^i. Thus,

$$\widetilde{\omega}_i^j(\delta) + \widetilde{\omega}_j^i(\delta) = a_{ik}^j \omega^k, \;\; a_{ik}^j = a_{jk}^i. \tag{15.11}$$

Now we will show how to write the form ω_i^j as a sum of the form $\widetilde{\omega}_i^j$, which can depend on dv^α, and a linear combination of the forms ω^i:

$$\omega_i^j = \widetilde{\omega}_i^j + \lambda_{ik}^j \omega^k, \tag{15.12}$$

where the coefficients λ_{ik}^j are unknown functions.

Equations (15.2) give now

$$\widetilde{\omega}_i^j + \widetilde{\omega}_j^i + (\lambda_{ik}^j + \lambda_{jk}^i)\omega^k = 0,$$

or by means of (15.11),

$$(a_{jk}^i + \lambda_{ik}^j + \lambda_{jk}^i)\, \omega^k = 0, \tag{15.13}$$

i.e.,

$$a_{jk}^i + \lambda_{ik}^j + \lambda_{jk}^i = 0. \tag{15.14}$$

The difference between equations (II_1) and (15.9) is

$$\omega^k \wedge (\omega_k^i - \widetilde{\omega}_k^i) = 0,$$

or by means of (15.12),

$$\lambda_{kh}^i \, \omega^k \wedge \omega^h = 0,$$

i.e.,

$$\sum_{(k,h)} (\lambda_{kh}^i - \lambda_{hk}^i)\, \omega^k \wedge \omega^h = 0.$$

Since the forms $\omega^k \wedge \omega^h$ are linearly independent, we find that

$$\lambda_{kh}^i - \lambda_{hk}^i = 0. \tag{15.15}$$

Making in (15.14) the cyclic permutation of the indices i, j, k, we obtain

$$\begin{cases} a_{ik}^j + \lambda_{ik}^j + \lambda_{jk}^i = 0, \\ a_{ji}^k + \lambda_{ji}^k + \lambda_{ki}^j = 0, \\ a_{kj}^i + \lambda_{kj}^i + \lambda_{ij}^k = 0. \end{cases} \tag{15.16}$$

Subtracting the sum of the first two equations from the third and applying (15.15), we get

$$\lambda^j_{ik} = \frac{a^i_{kj} - a^j_{ik} - a^k_{ji}}{2}. \tag{15.17}$$

Here, from (15.11), the coefficients a^j_{ik} are symmetric with respect to the indices i and j—this also can be seen from equations (15.16). The solution (15.17) satisfies the system of equations (15.14) and (15.15). Thus, it is always possible to find a solution of equations (15.1) and (15.2). Theorem 1 is proved. ☐

81. Axioms of equipollence of vectors. Can the same absolute differentiation $D\mathbf{X} = (dX^i + X^k\omega^i_k)\mathbf{e}_i$ correspond to two different decompositions of ds^2 into sums of squares? We saw that to any decomposition of ds^2 into sums of squares, there is a certain corresponding connection given by its components ω^j_i. These components, and only these components, satisfy the first group of structure equations (II_1). Since a connection defined by the osculating Euclidean space is also defined up to transformations of coordinate lines, and the components of this connection satisfy equations (II_1), the two connections coincide.

We proved above that this connection can be obtained as a solution of the system of equations (II_1) and (15.2). In order to generalize the notion of equipollence of vectors, we must use another approach.

Suppose that two infinitesimally close points M and N are given. Establish a correspondence of *equipollence* between vectors at M and N (Levi–Civita's parallelism) by the following five requirements:

1°. *If a vector* $\mathbf{X}'$ *is equipollent to a vector* $\mathbf{X}$, *then a vector* $m\mathbf{X}$ *is equipollent to a vector* $m\mathbf{X}'$, *where* m *is an arbitrary factor.*

These and the remaining requirements can be written as the formulas:

1°. If $\mathbf{X} \approx \mathbf{X}'$, then $m\mathbf{X} \approx m\mathbf{X}'$.

2°. If $\mathbf{X} \approx \mathbf{X}', \mathbf{Y} \approx \mathbf{Y}'$, then $\mathbf{X} + \mathbf{Y} \approx \mathbf{X}' + \mathbf{Y}'$.

3°. If $\mathbf{X} \approx \mathbf{X}'$, then $\mathbf{X}^2 \approx (\mathbf{X}')^2$, i.e., *equipollent vectors have equal lengths.*

4°. *If* $\mathbf{X} \approx \mathbf{X}'$, *then the difference of the corresponding components of these vectors is infinitesimally small.*

5°. This requirement will be given later.

Conditions 1° and 2° show that the vectors $\mathbf{X}$ and $\mathbf{X}'$ are linearly related:

$$X^{i'} = \alpha^i_k X^k.$$

From condition 4°, it follows that the last equation can be written in the form

$$X^{i'} = X^i + \omega^i_k X^k, \tag{15.18}$$

where ω^i_k are infinitesimally small.

Now we formulate the fifth condition:

5°. *The components ω^i_k are linear with respect to the differentials of independent variables.*

Condition 3° shows the equality of sums of squares of components of the vectors $\mathbf{X}$ and $\mathbf{X}'$:

$$\sum_i (X^i)^2 = \sum_i (X^{i'})^2.$$

Substituting (15.18) into the last equation and using the distributive law, we get

$$\begin{aligned}\sum_i (X^i)^2 &= \sum_i (X^i + \omega^i_k X^k)^2 \\ &= \sum_i (X^i)^2 + 2\sum_{i,k} X^i X^k + \sum_i (\omega^i_k X^k)^2.\end{aligned}$$

Cancelling out the first terms $\sum_i (X^i)^2$ and neglecting the last term (as infinitesimally small of second order), we get

$$\sum_{i,k} \omega^i_k X^i X^k = 0.$$

Summing in the last equation over the combinations (i,k) of i and k, we obtain

$$\sum_i \omega^i_i (X^i)^2 + \sum_{(i,k)} (\omega^i_k + \omega^k_i) X^i X^k = 0.$$

Since the vector X^i is arbitrary, we have

$$\omega^i_i = 0 \ \ (i \text{ not summed}), \ \ \omega^i_k + \omega^k_i = 0. \tag{15.19}$$

It follows that if

$$\mathbf{X}' = \mathbf{X} + d\mathbf{X},$$

then

$$dX^i = \omega^i_k X^k,$$

or

$$X^{i'} = X^i + \omega^i_k X^k. \tag{15.20}$$

This formula has the same form as earlier, but now all ω^j_i are arbitrary skew-symmetric forms.

We come to the following conclusion: the correspondence of equipollence is defined by an *arbitrary system of skew-symmetric forms ω^j_i*.

82. Space with Euclidean connection. Suppose now that a certain equipollence is given, i.e., there is given a law of correspondence of vectors at a point M and at a neighboring point N. Then the scalar product of vectors is preserved,

$$\mathbf{X} \cdot \mathbf{Y} = \mathbf{X}' \cdot \mathbf{Y}'.$$

In fact, from conditions 1° and 2°, it follows that

$$(\mathbf{X}+\lambda\mathbf{Y})^2 = \mathbf{X}^2 + \lambda^2\mathbf{Y}^2 + 2\lambda\mathbf{X}\cdot\mathbf{Y}.$$

But from condition 1°,

$$\mathbf{X}^2 = (\mathbf{X}')^2, \ \ \mathbf{Y}^2 = (\mathbf{Y}')^2.$$

As a result, we have

$$\mathbf{X}\cdot\mathbf{Y} = \mathbf{X}'\cdot\mathbf{Y}'.$$

Preservation of the scalar product (and preservation of norms of vectors) implies preservation of angles between corresponding vectors.

Note that we can define a position of the point N with respect to the frame associated with the point M by means of the components ω^i of translation. If we fix a law of equipollence, then we can represent the points M and N and vectors emanating from M and N in the same Euclidean space. Two "equipollent" vectors become *really equipollent* (parallel) in this Euclidean space.

Thus, we have defined an *Euclidean connection* in a "generalized space", i.e., as in the Euclidean space, we have connected neighborhoods of two infinitesimally close points M and N. In general, for arbitrary points M and N, this is impossible. However, two points A and B that are not infinitesimally close can be connected by a *certain* path AB.

Then we can represent a neighborhood of a point M_1, which is infinitesimally close to the point A on the line AB. Next we can move from M_1 to M_2, and so on, step-by-step by the same method. (Note that the correspondence obtained depends on the choice of a path AB.)

How is the passage to the limit realized effectively?

83. Euclidean space of conjugacy. Suppose that a point μ, representing a variable point M of the arc AB, moves along an arc $\alpha\beta$ representing the arc AB. Then for a frame at the point μ, we have

$$\left\{\begin{aligned} d\mu &= \omega^1\mathbf{i}_1 + \omega^2\mathbf{i}_2 + \ldots + \omega^n\mathbf{i}_n, \\ d\mathbf{i}_i &= \omega_i^k\mathbf{i}_k. \end{aligned}\right. \tag{15.21}$$

If the point M describes the arc AB in such a way that its coordinates are represented by continuously differentiable functions (of class C^1) of a parameter t, then

$$\left\{\begin{aligned} \omega^i &= p^i dt, \\ \omega_i^j &= p_i^j dt, \end{aligned}\right. \tag{15.22}$$

where for a fixed path AB, the quantities p^i and p_i^j are completely determined functions of the independent variable t. Substituting these values into equations

(15.21), we obtain the Cauchy system

$$\begin{cases} \dfrac{d\mu}{dt} = p^k\,\mathbf{I}_k, \\ \dfrac{d\mathbf{i}_i}{dt} = p_i^k\,\mathbf{i}_k. \end{cases} \tag{15.23}$$

If we integrate this system, then we come to the *Euclidean space of conjugacy* (see No. **75**).

Choosing a certain Euclidean connection, we can extend the tensor analysis to this connection defining the absolute differential of a vector by the formulas for the contravariant and covariant vectors (in an orthonormal frame these formulas coincide):

$$\begin{cases} DX^i = dX^i + X^k\omega_k^i, \\ DX_i = dX_i - X_k\omega_i^k. \end{cases} \tag{15.24}$$

84. Absolute exterior differential. The most important moment is the extension of the notion of absolute differentiation of differential forms to the space with Euclidean connection.

In a generalized manifold, one cannot add vectors with different initial points, if these points are not connected by certain paths.

Consider a point A inside the domain in question and take a family of paths connecting the point A with different points of the domain. Then we can consider integrals over surfaces, volumes, etc.

In fact, suppose that a vector $(\widetilde{\omega}^i)$ is associated with a point M of a surface S. This vector, after being transferred to the point A along a path MA by means of equipollence, becomes

$$\xi^i = \alpha_k^i\widetilde{\omega}^k.$$

The sum of these vectors on a p-dimensional surface S is represented by the integral

$$\underbrace{\int\cdots\int}_{(p)}\xi^i = \underbrace{\int\cdots\int}_{(p)}\alpha_k^i\widetilde{\omega}^k.$$

Applying the generalized Gauss theorem to this integral, we represent this integral in the form

$$\underbrace{\int\cdots\int}_{(p+1)}(\alpha_k^i d\widetilde{\omega}^k + d\alpha_k^i \wedge \widetilde{\omega}^k). \tag{15.25}$$

Now we restrict ourselves to a domain that is sufficiently small in order to be able to consider only an element of the integral.

We will see that under these conditions the result does not depend on the choice of paths.

In fact, the general formula

$$\xi^i = X^k \alpha^i_k$$

shows that at the point A, where $\xi^i = X^i$, all α^i_k, except α^i_i (i not summed), vanish,

$$\alpha^i_k = 0, \quad k \neq i, \quad \alpha^i_i = 1 \quad (i \text{ not summed}).$$

Thus the integrand in (15.15) reduces to the sum

$$D\tilde{\omega}^i = d\tilde{\omega}^i + d\alpha^i_k \wedge \tilde{\omega}^k.$$

At a point M, which is infinitesimally close to the point A, we have

$$\xi^i = X^i + d\alpha^i_k X^h.$$

On the other hand,

$$X^i = \xi^i + \omega^i_k \xi^k.$$

Hence

$$d\alpha^i_k = \omega^i_k$$

and

$$D\tilde{\omega}^i = d\tilde{\omega}^i + \omega^i_k \wedge \tilde{\omega}^k. \tag{15.26}$$

This generalizes the formula for the absolute exterior differential in the Euclidean space.

Similarly we have

$$D\tilde{\omega}^j_i = d\tilde{\omega}^j_i - \omega^k_i \wedge \tilde{\omega}^k_j + \omega^j_k \wedge \tilde{\omega}^k_i. \tag{15.27}$$

85. Torsion of the manifold. Take an infinitesimally close contour around the point A. Let M and M' be two infinitesimally close points on the contour. Associate with the point M an infinitesimally small vector $\overrightarrow{ds} = \overrightarrow{MM'}$. Its components are

$$\tilde{\omega}^i = \omega^i.$$

Take the sum of all these vectors. Because the contour is infinitesimally small, we can replace this sum by one vector. In Euclidean space, this sum vanishes. Here we have

$$D\omega^i = d\omega^i + \omega^i_k \wedge \omega^k = d\omega^i - \omega^k \wedge \omega^i_k = \Omega^i. \tag{15.28}$$

This sum can be equal to zero—*this is an intrinsic property of the space*, since for all decompositions of ds^2 into a sum of squares (i.e., for any choice of ω^i), this sum is zero.

In order to satisfy this property, we need to choose ω_i^j such that

$$d\omega^i = \omega^k \wedge \omega_k^i.$$

We saw that this is always possible to achieve, and in only one way.

This is one of the possible Euclidean connections for ds^2. In fact, this was Riemann's choice of connection. This choice implies that the sum of vectors of a closed contour around a point is equal to zero.

However, this connection is not the only possible connection.

If one considers the general formula

$$d\omega^i = \omega^k \wedge \omega_k^i + \Omega^i,$$

where ω_k^i depend on secondary parameters v^α (defining a decomposition of ds^2 into a sum of squares), then the quadratic form Ω^i does not depend on these parameters, and hence it does not depend on the forms ω_k^i. The form Ω^i defines the *torsion* of the manifold.

86. Structure equations of a space with Euclidean connection. Let us investigate now how the structure equations (II$_2$),

$$d\omega_i^j = \omega_i^k \wedge \omega_k^j, \tag{II$_2$}$$

will be changed in the new connection.

Associate with a point M of a cycle Γ a vector $\mathbf{X}$ and consider the integral

$$\int\limits_\Gamma D\mathbf{X}.$$

In the Euclidean space this integral vanishes.

For an arbitrary connection, the components of $D\mathbf{X}$ are

$$\widetilde{\omega}^i = dX^i + X^k \omega_k^i. \tag{15.29}$$

Applying formula (15.26) for the absolute exterior differential to form (15.29), we obtain

$$\begin{aligned} D\widetilde{\omega}^i &= dX^k \wedge \omega_k^i + X^k d\omega_k^i + \omega_k^i \wedge (dX^k + X^h \omega_h^k) \\ &= X^k (d\omega_k^i + \omega_h^i \wedge \omega_k^h). \end{aligned}$$

Set

$$d\omega_k^i - \omega_k^h \wedge \omega_h^i = \Omega_k^i.$$

Then

$$\int D\mathbf{X}^i = \iint X^k \Omega_k^i.$$

The quadratic forms Ω_k^i are skew-symmetric. They define a rotation which gives to the vector X the variation $\int D\mathbf{X}$. Thus, in general, we have

$$d\omega^i = \omega^k \wedge \omega_k^i + \Omega^i, \tag{15.30}$$

$$d\omega_i^j = \omega_i^k \wedge \omega_k^j + \Omega_i^j. \tag{15.31}$$

87. Translation and rotation associated with a cycle. As our paths AM, we take the arcs AM of the cycle in question and develop this cycle on the line $\alpha\mu\alpha'$ of the Euclidean space (Fig. 5).

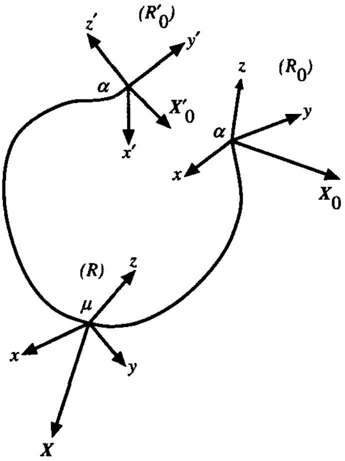

Fig. 5

A vector $\mathbf{X}$ attached to the point A is mapped into a vector $\mathbf{X}_0$ emanating from the point α and into a vector $\mathbf{X}'_0$ emanating from the point α'. When we move from M to M', we take $\mathbf{X}'$ at μ', $\mathbf{X}$ at μ, and compute

$$D\mathbf{X} = \mathbf{X}' - \mathbf{X},$$

i.e., in the Euclidean connection, we compute the ordinary difference $\mathbf{X}' - \mathbf{X}$, without moving to the point α. Thus,

$$\int D\mathbf{X},$$

is an ordinary integral for the path $\alpha\mu\alpha'$, i.e., it is the difference $\mathbf{X}'_0 - \mathbf{X}_0$. We see that this result depends on a field value at the point A (dX^k disappears in the computation).

Next we move from the vector $\mathbf{X}_0$ to the vector $\mathbf{X}'_0$ by means of a simple infinitesimal rotation which takes the frame (R_0) attached to the point α into the frame (R'_0) attached to the point α', i.e., it makes $\mathbf{X}_0$ parallel to $\mathbf{X}'_0$ (since the coordinates of $\mathbf{X}_0$ with respect to (R_0) are the same as the coordinates of $\mathbf{X}'_0$ with respect to (R'_0)).

The quantities Ω_i^j are the components in (R_0') of an infinitesimally small rotation making $\mathbf{R}_0$ parallel to $\mathbf{R}_0'$. Similarly, the quantities Ω^i are the components in (R_0') of an infinitesimally small translation transferring α to α'.

Thus, the displacement taking (R_0) to (R_0') has components Ω^i and Ω_i^j.

This displacement vanishes in the Euclidean space. The *inverse* displacement (i.e., the displacement taking (R_0') to (R_0)) is called the displacement *associated with the cycle.*

Consider a field of equipollent vectors. In the Euclidean connection, the vector $\mathbf{Y}(\alpha')$ is equipollent in an ordinary sense (equal) to the vector $\mathbf{X}_0$. At the same time an observer of a Riemannian manifold considers the vector $\mathbf{X}_0$, which is not equipollent to the vector $\mathbf{Y}$, as the initial vector. The variation is $\mathbf{Y} - \mathbf{X}_0'$, i.e., the *rotation* Ω_i^j *associated with the cycle.*

Thus, the rotation associated with a cycle is the rotation undergone by a vector which is transferred along a cycle equipollently.

The associated rotation defines the *curvature*, and the associated translation defines the *torsion* at the point A.

88. The Bianchi identities. Consider structure equations (15.30) and (15.31) of the space with Euclidean connection.

Here

$$\Omega^i = R^i_{[jk]}\omega^j \wedge \omega^k, \quad \Omega_i^j = R^j_{i[kh]}\omega^k \wedge \omega^h,$$

where

$$R^i_{[jk]} = R^i_{jk} - R^i_{kj}, \quad R^j_{i[kh]} = R^j_{ikh} - R^j_{ihk}.$$

In order to obtain the Bianchi identities, we apply exterior differentiation to equations (15.30) and (15.31):

$$0 = d\omega^k \wedge \omega_k^i - \omega^k \wedge \omega_k^i + d\Omega^i,$$
$$0 = d\omega_i^k \wedge \omega_k^j - \omega_i^k \wedge d\omega_k^i + d\Omega_i^j.$$

In ordinary Euclidean space, these equations are identically satisfied. This implies that if we substitute the values of $d\omega^i$ and $d\omega_i^j$ from (15.30) and (15.31) into the last equations, all the terms, except those containing the forms Ω^i and Ω_i^j, must cancel out:

$$0 = \Omega^k \wedge \omega_k^i - \omega^k \wedge \Omega_k^i + d\Omega^i, \tag{15.32}$$

$$0 = \Omega_i^k \wedge \omega_k^j - \omega_i^k \wedge \Omega_k^i + d\Omega_i^j. \tag{15.33}$$

This is the first group of the Bianchi identities written for the space with Euclidean connection.

89. Theorem of preservation of curvature and torsion. We turn now to the interpretation of equations (15.32) and (15.33). Start from equations (15.33). They show that the absolute differential of the form Ω_i^j is equal to zero.

From equations (15.26), the absolute differential of the form $\widetilde{\omega}^i$ is

$$D\widetilde{\omega}^i = d\widetilde{\omega}^i + \omega^i_k \wedge \widetilde{\omega}^k.$$

Similarly, the absolute differential of the form $\widetilde{\omega}^j_i$ is

$$D\widetilde{\omega}^j_i = d\widetilde{\omega}^j_i - \omega^k_i \wedge \widetilde{\omega}^j_k + \omega^j_k \wedge \widetilde{\omega}^k_i.$$

Applying the last formula to the form $\widetilde{\Omega}^j_i$ and using (15.33), we find that

$$D\Omega^j_i = d\Omega^j_i - \omega^k_i \wedge \Omega^j_k + \omega^j_k \wedge \Omega^k_i = 0.$$

In fact, in the last formula, the first sum is the same as in (15.33). As to the second sum, we have $\omega^j_k \wedge \Omega^k_i = \Omega^k_i \wedge \omega^j_k$, since the form Ω^k_i is of even order, and this permutation of factors does not change the sign in the product.

Thus, Ω^j_i is the sum of bivectors (skew-symmetric tensors). Take a volume V bounded by a surface S. Consider bivectors Ω^j_i applied to the surface S. Their sum is equal to zero. This follows from equations (15.33) by means of the Gauss theorem:

$$\iint\limits_S \Omega^j_i = \iiint\limits_V d\,\Omega^j_i.$$

Now we turn to equations (15.32). The absolute differential of the form Ω^i is

$$D\,\Omega^i = d\,\Omega^i + \Omega^k \wedge \omega^i_k.$$

From equations (15.32), it follows that

$$D\,\Omega^i = \omega^k \wedge \Omega^i_k.$$

Consider, for example a three-dimensional space and an oriented element of a surface (i.e., an element with a direction of an infinitesimally small cycle). To every cycle, there are the corresponding translation Ω^i and the rotation Ω^j_i with components

$$\Omega^1,\ \Omega^2,\ \Omega^3;\ \Omega^3_2,\ \Omega^1_3,\ \Omega^2_1.$$

The rotation Ω^j_i can be represented by a vector $(\widetilde{\omega}^1, \widetilde{\omega}^2, \widetilde{\omega}^3)$, and the translation Ω^i can be represented as the moment of a pair.

Thus, on a surface, a double system of vectors is formed. Consider the sum of all vectors and all pairs. We need to connect a point A with all points M of a surface by certain paths.

If (X^1, X^2, X^3) is a vector, then we can transfer it to the point A by the formula

$$\xi^i = a^i_k X^k.$$

The components of a pair can be found from the matrix

$$\begin{pmatrix} x & y & z \\ a^1_k\widetilde{\omega}^k & a^2_k\widetilde{\omega}^k & a^3_k\widetilde{\omega}^k \end{pmatrix},$$

and they are

$$
\begin{gathered}
ya^3_k\tilde{\omega}^k - za^2_k\tilde{\omega}^k,\\
za^1_k\tilde{\omega}^k - xa^3_k\tilde{\omega}^k,\\
xa^2_k\tilde{\omega}^k - ya^1_k\tilde{\omega}^k,
\end{gathered}
$$

where x, y, z are the coordinates of a point M with respect to the trihedron (R_0).

The sum of the vectors vanishes by means of equations (15.32). The sum of the pairs is

$$
\begin{aligned}
\iint\limits_S [a^1_1\Omega^1 &+ a^1_2\Omega^2 + a^1_3\Omega^3 + y(a^3_1\Omega^3_2 + a^3_2\Omega^1_3 + a^3_3\Omega^2_1)\\
&-z(a^2_1\Omega^3_2 + a^2_2\Omega^1_3 + a^2_3\Omega^2_1) + \dots].
\end{aligned}
$$

The integral along the entire surface reduces to the integral over the volume

$$
\begin{aligned}
\iiint\limits_V [a^1_1 d\Omega^1 &+ da^1_1 \wedge \Omega^1 + \dots + yd(a^3_1\Omega^3_2 + \dots)\\
&+dy \wedge (a^3_1\Omega^2_3 + \dots) - \dots].
\end{aligned}
$$

If a volume is very small, then by setting

$$a^1_1 = 1,\ \ a^2_1 = a^3_1 = 0,\ \ da^1_1 = 0,\ \ da^k_i = \omega^k_i,\ \ y = z = 0,\ \ dy = \omega^2,\ \ dz = \omega^3, \dots ,$$

we obtain

$$\iiint\limits_V (d\Omega^1 + \omega^1_k \wedge \Omega^k + \omega^2 \wedge \Omega^2_1 - \omega^3 \wedge \Omega^1_3).$$

But by means of the first equation of (15.32), we have

$$d\Omega^1 + \omega^1_k \wedge \Omega^k = d\Omega^1 + \omega^1_k \wedge \Omega^k - \omega^2 \wedge \Omega^1_2 - \omega^3 \wedge \Omega^1_3 = 0.$$

Thus, we have proved the following result.

Theorem 2. *In a Riemannian manifold with Euclidean connection, to a small volume there correspond both vectors and pairs, defining translations and rotations associated with a cycle on a surface. Geometrically, the sum of vectors and pairs is equal to zero.*

In mechanics one can find an interpretation in terms of the elastic forces acting on an area element.

Chapter 16

Riemannian Curvature of a Manifold

90. The Bianchi identities in a Riemannian manifold. The point of view of Riemann himself corresponds to a torsion-free connection,

$$\Omega^i = 0.$$

Now, the Bianchi identities take the form

$$\omega^k \wedge \Omega_k^i = 0, \tag{16.1}$$

$$\Omega_i^k \wedge \omega_k^j - \omega_i^k \wedge \Omega_k^j + d\Omega_i^j = 0. \tag{16.2}$$

The first of these identities show that there are no moments of pairs since the translation Ω^i is absent, and the second one shows that the sum of vectors vanishes (see No. **89**).

We decomposed ds^2 into a sum of squares and chose the unique forms ω_k^i in such a way that

$$d\omega^i = \omega^k \wedge \omega_k^i, \tag{16.3}$$

$$d\omega_i^j = \omega_i^k \wedge \omega_k^j + \Omega_i^j. \tag{16.4}$$

Is it possible to determine the forms ω_k^i independently of the choice of axes? This determination does not depend on the method of decomposition of ds^2 into a sum of squares. Take the most general decomposition of ds^2 into a sum of squares. The forms ω_i^j depend on the variables $u^1, u^2, \dots, u^n$ and on parameters $v^1, v^2, \dots, v^{\frac{n(n-1)}{2}}$ defining the method of decomposition. Obviously, Ω_{ij} can depend on both, ω^i and ω_i^j. We prove that Ω_{ij} depend only on ω^i. This follows from our Theorem 1 in No. **79**. We will prove this independently of this theorem.

From equations (15.31), we have

$$\Omega_i^j + \Omega_j^i = 0.$$

Denote the forms ω_{ij} by ω_α with Greek indices.

In general, the quadratic forms Ω_{ij} depend on both ω^i and ω_i^j. We can set

$$\Omega_{ki} = a_{ki[hl]}\, \omega^h \wedge \omega^l + b^\alpha_{kil}\, \omega^l \wedge \omega_\alpha + c^{[\alpha\beta]}_{ki}\, \omega_\alpha \wedge \omega_\beta,$$

where

$$a_{ki[hl]} = -a_{ki[lh]},$$
$$c^{[\alpha\beta]}_{ki} = -c^{[\beta\alpha]}_{ki}.$$

Substituting this expression for Ω_{ki} into equations (16.1), we find that

$$\omega^k \wedge \Omega_{ki} = a_{ki[hl]}\, \omega^k \wedge \omega^h \wedge \omega^l + b^\alpha_{kil}\, \omega^k \wedge \omega^l \wedge \omega_\alpha + c^{[\alpha\beta]}_{ki}\, \omega^k \wedge \omega_\alpha \wedge \omega_\beta.$$

Firstly, note that in the last sum, all terms contain different combinations of the basis forms $\omega^k, \omega_\alpha, \omega_\beta$, since the summation is carried over the combinations of α and β, and as a result, there are no similar terms. Thus,

$$c^{[\alpha\beta]}_{ki} = 0,$$

and the form Ω_{ki} does not contain terms with the products $\omega_\alpha \wedge \omega_\beta$.

Secondly, in the second sum for the triple of basis forms $\omega^k, \omega^l, \omega_\alpha$, two combinations

$$b^\alpha_{kil}\, \omega^k \wedge \omega^l \wedge \omega_\alpha + b^\alpha_{lik}\, \omega^l \wedge \omega^k \wedge \omega_\alpha$$

are possible. It follows that there is the symmetry in two lower indices:

$$b^\alpha_{kil} = b^\alpha_{lik}. \tag{16.5}$$

On the other hand, since the forms Ω_{ki} are skew-symmetric, we have

$$b^\alpha_{kil} = -b^\alpha_{ikl}. \tag{16.6}$$

Applying transformations (16.5) and (16.6) in turn, we obtain

$$b^\alpha_{kil} = b^\alpha_{lik} = -b^\alpha_{ilk} = -b^\alpha_{kli} = b^\alpha_{lki} = b^\alpha_{ikl} = -b^\alpha_{kil}.$$

It follows that

$$b^\alpha_{kil} = 0,$$

and the forms Ω_{ki} do not depend on the forms ω_α.

91. The Riemann–Christoffel tensor. From the previous section, we have

$$\Omega_i^j = R_i{}^j{}_{[kh]}\omega^k \wedge \omega^h. \tag{16.7}$$

The coefficients $R_i{}^j{}_{[kh]}$ constitute a tensor, since if we contract these quantities with the vectors ω^k and ω^h, we obtain the tensorial differential form Ω_{ij}. This tensor is called the *Riemann–Christoffel tensor.*

Structure equations (16.3) and (16.4) can be written now in the form

$$d\omega^i = \omega^k \wedge \omega^i_k, \tag{16.8}$$

$$d\omega^j_i = \omega^k_i \wedge \omega^j_k + R_i{}^j{}_{[kh]}\omega^k \wedge \omega^h. \tag{16.9}$$

If we substitute the values of the forms Ω_α from (16.7) into the first Bianchi identities (16.1), the latter give

$$R_k{}^i{}_{hl}\omega^k \wedge \omega^h \wedge \omega^l = 0.$$

In order to collect similar terms, we cycle these equations with respect to the three lower indices. Equating the coefficients obtained to zero, we find that

$$R_k{}^i{}_{hl} + R_h{}^i{}_{lk} + R_l{}^i{}_{kh} = 0,$$

or

$$(i) = R_{ikhl} + R_{ihlk} + R_{ilkh} = 0. \tag{16.10}$$

If we choose four arbitrary numbers i, k, h, l among $1, 2, \dots, n$, then there will be a few dependencies among the components R_{ikhl} with the chosen indices. These indices are split into two pairs, i, k and h, l.

1°. *Under permutation of indices of one pair, the quantity R_{ikhl} changes its sign.* For the first pair i, k, this follows from the skew-symmetry of the forms $\Omega_i{}^k$, and for the second pair h, l this follows from the accompanying permutation of factors ω^h, ω^l in the exterior product $\omega^h \wedge \omega^l$ and the invariance of the left-hand side Ω^i_k in (16.7).

2°. *Under permutation of pairs of indices, the quantity R_{ikhl} is not changed.* To prove this, consider equations (16.10) and equations obtained from these by the cyclic permutation of four indices i, k, h, l:

$$\begin{aligned}
(i) &= R_{ik,hl} + R_{ih,lk} + R_{il,kh} = 0,\\
(k) &= R_{kh,li} + R_{kl,ih} + R_{ki,hl} = 0,\\
(h) &= R_{hl,ik} + R_{hl,ik} + R_{hk,li} = 0,\\
(l) &= R_{li,kh} + R_{li,kh} + R_{lh,ik} = 0.
\end{aligned}$$

Then

$$\frac{1}{2}[(i) + (k) - (h) - (l)] = R_{khli} - R_{likh} = 0. \tag{16.11}$$

This proves property 2°.

We turn now to the second group (16.2) of the Bianchi identities:

$$d\Omega_i^j + \Omega_i^k \wedge \omega_k^j - \omega_i^k \wedge \Omega_k^j = 0.$$

Substituting Ω_i^j from (16.7) into these identities, we obtain

$$dR_i{}^j{}_{kh} \wedge \omega^k \wedge \omega^h + R_i{}^j{}_{kh}\, \omega^i \wedge \omega_i^k \wedge \omega^h - R_i{}^j{}_{kh}\, \omega^k \wedge \omega^l \wedge \omega_i^h$$
$$+ R_i{}^k{}_{lh}\, \omega^l \wedge \omega^h \wedge \omega_k^j - R_k{}^j{}_{lh}\, \omega_i^k \wedge \omega^l \wedge \omega^h = 0.$$

Changing the summation indices, we have

$$(dR_i{}^j{}_{kh} - R_i{}^j{}_{lh}\, \omega_k^l - R_i{}^j{}_{kl}\, \omega_h^i + R_i{}^l{}_{kh}\, \omega_l^j - R_l{}^j{}_{kh}\, \omega_i^l) \wedge \omega^k \wedge \omega^h = 0.$$

Using the notation D for the absolute differential, we get

$$DR_i{}^j{}_{kh} \wedge \omega^k \wedge \omega^h = 0,$$

or

$$R_i{}^j{}_{kh|l}\omega^l \wedge \omega^k \wedge \omega^h = 0.$$

Cycling three indices l, k, h and equating the coefficients to zero, we obtain the second group of conditions for the covariant derivatives of the Riemann–Christoffel tensor:

$$R_i{}^j{}_{kh|l} + R_i{}^j{}_{hl|k} + R_i{}^j{}_{lk|h} = 0. \qquad (16.12)$$

It follows that the number of all components of the Riemann–Christoffel tensor is:

$$\frac{n(n-1)}{2} \text{ of type } R_{ij,ij};$$
$$\frac{n(n-1)(n-2)}{2} \text{ of type } R_{ij,ik};$$
$$2\frac{n(n-1)(n-2)(n-3)}{4!} \text{ of type } R_{ij,kh};$$

altogether, we have

$$\frac{n^2(n^2-1)}{12}$$

different components.

It is always possible to find a Riemannian manifold corresponding to arbitrary values of these quantities.

92. Riemannian curvature. At a point A, consider a bivector defined by the coordinates $p^{ij} = -p^{ji}$. This is an oriented element of a surface. A certain rotation Ω_{ij} is associated with this element, and since

$$\omega^k \wedge \omega^h = p^{kh},$$

we have

$$q_{ij} = -R_{ij[kh]}p^{kh}.$$

Consider now the scalar product

$$p^{ij}q_{ij} = R_{ij,kh}p^{ij}p^{kh},$$

where $q_{ij} = -\Omega_{ij}$ are components of the associated rotation. The *Riemannian curvature* at the point A in the planar direction of the above bivector is defined by the formula

$$K = \frac{p^{ij}q_{ij}}{\sum\limits_{(i,j)}(p^{ij})^2} = -\frac{R_{ij,kh}p^{ij}p^{kh}}{\sum\limits_{(i,j)}(p^{ij})^2}. \tag{16.13}$$

93. The case $n = 2$. We have only one quadratic form, namely,

$$\Omega_1^2 = R_1{}^2{}_{12}\omega^1 \wedge \omega^2 = -K\omega^1 \wedge \omega^2. \tag{16.14}$$

This is the Riemannian curvature of the manifold.

We saw that a rotation is associated with a cycle. If $d\sigma$ is an area bounded by a cycle, then a vector $\mathbf{X}$, remaining equipollent to itself under traveling the cycle, will not return to the initial position after the cycle is complete. This vector takes new position $\mathbf{X}'$ forming an angle

$$K\,d\sigma$$

with the original position.

The structure equations are

$$\begin{cases} d\omega^1 = \omega^2 \wedge \omega_2^1, \\ d\omega^2 = \omega^1 \wedge \omega_1^2, \\ d\omega_1^2 = -K\,\omega^1 \wedge \omega^2. \end{cases}$$

In general, the Riemannian curvature is a function of a point.

The simplest case is the space of constant curvature,

$$K = \text{const}.$$

An example of such a space is a sphere in an ordinary space. Its curvature $K = \frac{1}{R^2}$, where R is the radius of the sphere.

In fact, let R be a radius of a sphere, and O be its center. Associate with the point O trihedrons O, x_1, x_2, x_3 depending on three parameters.

If the point O is fixed, then a trihedron rotates with rotation components

$$\varpi_{12},\quad \varpi_{23},\quad \varpi_{31}.$$

If M is the common point of the third axis of the trihedron and the sphere, then

$$\overrightarrow{OM} = R\mathbf{I}_3$$

is the position vector of the point M of the sphere.

For the trihedrons in question, we have the following structure equations and the equations of infinitesimal displacement of axes:

$$\begin{cases} d\varpi_{23} = \varpi_{21} \wedge \varpi_{13}, & d\mathbf{I}_1 = \varpi_{12}\mathbf{I}_2 + \varpi_{13}\mathbf{I}_3, \\ d\varpi_{31} = \varpi_{32} \wedge \varpi_{21}, & d\mathbf{I}_2 = \varpi_{21}\mathbf{I}_1 + \varpi_{23}\mathbf{I}_3, \\ d\varpi_{12} = \varpi_{13} \wedge \varpi_{32}, & d\mathbf{I}_3 = \varpi_{31}\mathbf{I}_1 + \varpi_{32}\mathbf{I}_2. \end{cases} \tag{16.15}$$

If R is constant, then the displacement of the point M is equal to $d\mathbf{I}_3$ multiplied by R:

$$d\mathbf{M} = R\,(\varpi_3^1\mathbf{I}_1 + \varpi_3^2\mathbf{I}_2).$$

Thus, for the sphere, we have

$$\omega^1 = R\varpi_3^1,\quad \omega^2 = R\varpi_3^2,\quad \omega_1^2 = \varpi_1^2.$$

The structure equations are

$$\begin{aligned} d\omega^1 &= R\,d\varpi_{31} = R\varpi_{32} \wedge \varpi_{21} = \omega^2 \wedge \varpi_{21}, \\ d\omega^2 &= \omega^1 \wedge \varpi_{12}, \\ d\omega_1^2 &= d\varpi_{12} = \varpi_{13} \wedge \varpi_{32} = -\frac{1}{R^2}\,\omega^1 \wedge \omega^2. \end{aligned}$$

It follows that the curvature $K = \frac{1}{R^2}$.

This can be generalized to arbitrary dimension. All of these results can be obtained by considering a line element of a sphere. It has the form

$$ds^2 = R^2\,d\theta^2 + R^2 \sin^2\theta\,d\varphi^2,$$

where $\frac{\pi}{2} - \theta$ is the latitude, and φ is the longitude. Thus,

$$\omega^1 = R\,d\theta,\quad \omega^2 = R\,\sin\theta\,d\varphi.$$

Applying exterior differentiation, we obtain

$$d\omega^1 = 0,\quad d\omega^2 = R\,\cos\theta\,d\theta \wedge d\varphi.$$

Since

$$d\omega^1 = \omega^2 \wedge \omega_2^1, \quad d\omega^2 = \omega^1 \wedge \omega_1^2,$$

and hence

$$\omega_1^2 \wedge d\varphi = 0, \quad \omega_1^2 \wedge d\theta = \cos\theta\, d\varphi \wedge d\theta,$$

we find that

$$\omega_1^2 = \cos\theta\, d\varphi.$$

Further, we have

$$d\omega_1^2 = -\sin\theta\, d\theta \wedge d\varphi = -\frac{1}{R^2}\,\omega^1 \wedge \omega^2,$$

and thus the curvature $K = \frac{1}{R^2}$.

94. The case $n = 3$. Now equations (16.7) have the form

$$\begin{cases} \Omega_{23} = R_{23,23}\,\omega^2 \wedge \omega^3 + R_{23,31}\,\omega^3 \wedge \omega^1 + R_{23,12}\,\omega^1 \wedge \omega^2, \\ \Omega_{31} = R_{31,23}\,\omega^2 \wedge \omega^3 + R_{31,31}\,\omega^3 \wedge \omega^1 + R_{31,12}\,\omega^1 \wedge \omega^2, \\ \Omega_{12} = R_{12,23}\,\omega^2 \wedge \omega^3 + R_{12,31}\,\omega^3 \wedge \omega^1 + R_{12,12}\,\omega^1 \wedge \omega^2. \end{cases}$$

For brevity, denote $R_{23,23}$ by $-K_{11}$, $R_{31,31}$ by $-K_{12}$, and so on, replacing in each group of indices a pair of indices by a complementary index. Then we have

$$\begin{cases} \Omega_{23} = -K_{11}\,\omega^2 \wedge \omega^3 - K_{12}\,\omega^3 \wedge \omega^1 - K_{13}\,\omega^1 \wedge \omega^2, \\ \Omega_{31} = -K_{21}\,\omega^2 \wedge \omega^3 - K_{22}\,\omega^3 \wedge \omega^1 - K_{23}\,\omega^1 \wedge \omega^2, \\ \Omega_{12} = -K_{31}\,\omega^2 \wedge \omega^3 - K_{32}\,\omega^3 \wedge \omega^1 - K_{33}\,\omega^1 \wedge \omega^2. \end{cases}$$

The property $R_{ij,kh} = R_{kh,ij}$, yields the symmetry of K_{ij}:

$$K_{21} = K_{12}, \quad K_{31} = K_{13}, \quad K_{32} = K_{23}, \tag{16.16}$$

and the condition (16.12),

$$R_{i\ kh|l}^{\ j} + R_{i\ hl|k}^{\ j} + R_{i\ lk|h}^{\ j} = 0$$

takes the form

$$\sum_k K_{ik|k} = 0. \tag{16.17}$$

It should be pointed out that in equations (16.17), each of the first two indices represent a pair of complementary indices, and the third index (written on the right of the vertical line), the index of covariant differentiation, indicates the form ω^k in which the covariant derivative stands for a coefficient.

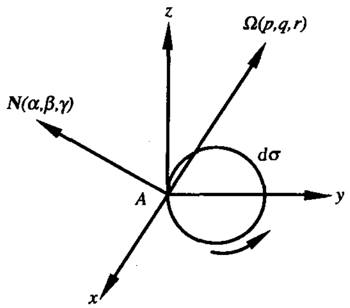

Fig. 6

95. Geometric theory of curvature of a three-dimensional Riemannian manifold. Suppose that a frame is connected with a point A and the element of surface $d\sigma$ passing through the point A (Fig. 6). Let $d\sigma$ be an oriented cycle, α, β, γ be the direction cosines of the normal, and $p d\sigma, q d\sigma, r d\sigma$ be the rotation associated with the cycle. We can always assume that $d\sigma$ is an infinitesimally small parallelogram with sides $\omega^i(d)$:

$$\begin{pmatrix} \omega^1(d) & \omega^2(d) & \omega^3(d) \\ \omega^1(\delta) & \omega^2(\delta) & \omega^3(\delta) \end{pmatrix}.$$

Then

$$\omega_2 \wedge \omega_3 = \alpha\, d\sigma.$$

Thus, if $\Omega(p, q, r)$ is the rotation associated with the cycle with components

$$p\, d\sigma, \quad q\, d\sigma, \quad r\, d\sigma,$$

then

$$\begin{cases} p\, d\sigma = K_{11}\, \alpha\, d\sigma + K_{12}\, \beta\, d\sigma + K_{13}\, \gamma\, d\sigma, \\ q\, d\sigma = K_{21}\, \alpha\, d\sigma + K_{22}\, \beta\, d\sigma + K_{23}\, \gamma\, d\sigma, \\ r\, d\sigma = K_{31}\, \alpha\, d\sigma + K_{32}\, \beta\, d\sigma + K_{33}\, \gamma\, d\sigma. \end{cases}$$

It follows that

$$\begin{cases} p = K_{11}\, \alpha + K_{12}\, \beta + K_{13}\, \gamma, \\ q = K_{21}\, \alpha + K_{22}\, \beta + K_{23}\, \gamma, \\ r = K_{31}\, \alpha + K_{32}\, \beta + K_{33}\, \gamma. \end{cases}$$

These formulas coincide with the formulas of the theory of elasticity which define the liquid pressure on a surface element.

Define the quadric-indicatrix (a surface of second order):

$$\Phi(x,y,z) = K_{11}x^2 + K_{22}y^2 + K_{33}z^2$$
$$+2K_{12}xy + 2K_{23}yz + 2K_{31}xz = 1.$$

The diametrical plane which is conjugate to the direction (α,β,γ) has the direction parameters

$$\frac{1}{2}\Phi'_\alpha,\ \frac{1}{2}\Phi'_\beta,\ \frac{1}{2}\Phi'_\gamma,\ \text{i.e., } p,\ q,\ r;$$
$$\Phi'_\alpha = \left.\frac{\partial\Phi}{\partial x}\right|_{x=\alpha,y=\beta,z=\gamma},\quad \Phi'_\beta = \left.\frac{\partial\Phi}{\partial y}\right|_{x=\alpha,y=\beta,z=\gamma},\quad \Phi'_\gamma = \left.\frac{\partial\Phi}{\partial z}\right|_{x=\alpha,y=\beta,z=\gamma}.$$

The equation of this plane is

$$x\Phi'_\alpha + y\Phi'_\beta + z\Phi'_\gamma = 0.$$

This implies the following result.

Theorem 1. *The vector representing the rotation associated with a surface element is normal to the diametrical plane which is conjugate to the direction of the surface normal. In order to find the magnitude of this rotation, it is sufficient to find its projection onto the direction* (α,β,γ).

Suppose that P is the common point of the quadric-indicatrix and the normal AP to a surface element. (A is a point at which the element is considered; this is the center of the quadric.)

We have

$$p\alpha + q\beta + r\gamma = \Phi(\alpha,\beta,\gamma) = \frac{1}{\overline{AP}^2}.$$

The *Riemannian curvature* at the point A in the direction of (α,β,γ) is equal to the projection of the rotation vector onto this direction: this is $\Phi(\alpha,\beta,\gamma)$ (Schur's definition).

The directions of the axes of the quadric-indicatrix are called the *principal directions* at the point A.

The rotation associated with an element normal to a principal direction takes place around this direction. If we take the coordinate axes as the quadric axes, then in the quadric equation, only squares remain. The coefficients in these squares are the *principal curvatures.* If the indicatrix is a sphere, then the principal curvatures are equal. In this case, the manifold is *isotropic* at a point.

96. Schur's theorem. *If a manifold is isotropic at each point, then the Riemannian curvature of the manifold is the same at all points.*

In fact, the vector Ω_{ij} is reduced to the form

$$-K\omega^i \wedge \omega^j.$$

The vanishing of its absolute exterior differential gives

$$dK \wedge \omega^i \wedge \omega^j = 0,$$

since

$$D\overline{\omega}_{ij} = d\overline{\omega}_{ij} + \omega_{ki} \wedge \overline{\omega}_{kj} + \omega_{kj} \wedge \overline{\omega}_{ik}.$$

All

$$D\overline{\omega}_{ij} = 0$$

by means of the second group of Bianchi's identities.

If

$$\overline{\omega}_{ij} = -K\,\omega^i \wedge \omega^j,$$

then

$$d\widetilde{\omega}_{ij} = -dK \wedge \omega^i \wedge \omega^j - K\,d\omega^i \wedge \omega^j + K\,d\omega^j \wedge \omega^i.$$

On the other hand, $D\overline{\omega}_{ij} = 0$ implies that

$$d\widetilde{\omega}_{ij} = -\omega_{ki} \wedge \overline{\omega}_{kj} - \omega_{kj} \wedge \overline{\omega}_{ik} = -K\,d\omega^i \wedge \omega^j + K\,d\omega^j \wedge \omega^i.$$

Thus,

$$dK \wedge \omega^i \wedge \omega^j = 0.$$

It follows from these equations that the differential dK depends linearly on ω^i and ω^j. But it does not depend on ω^i. Thus $dK = 0$, and $K = \text{const}$.

MECHANICAL INTERPRETATION. An ideal liquid in equilibrium, under action of internal forces only, is subject to a constant pressure at all points.

97. Example of a Riemannian space of constant curvature. Consider a rigid body rotating around a fixed point with velocity components p, q, r. If the ellipsoid of inertia is a sphere, then the kinetic energy has the form

$$2T = A\,(p^2 + q^2 + r^2),$$

and we have a Riemannian manifold with the line element

$$ds^2 = A\,(p^2 + q^2 + r^2)dt^2,$$

since

$$ds^2 = 2T\,dt^2.$$

Consider a coordinate trihedron invariantly connected with the rigid body in question. Suppose that $\widetilde{\omega}_{23}, \widetilde{\omega}_{31}, \widetilde{\omega}_{12}$ are the components of its rotation. We have

$$\widetilde{\omega}_{23} = p\,dt, \quad \widetilde{\omega}_{31} = q\,dt, \quad \widetilde{\omega}_{12} = r\,dt.$$

It follows that

$$ds^2 = A\,[(\widetilde{\omega}_{23})^2 + (\widetilde{\omega}_{31})^2 + (\widetilde{\omega}_{12})^2].$$

Since the rotation occurs in Euclidean space, we have

$$\begin{cases} d\omega_{23} = \omega_{21} \wedge \omega_{13}, \\ d\omega_{31} = \omega_{32} \wedge \omega_{21}, \\ d\omega_{12} = \omega_{13} \wedge \omega_{32}. \end{cases}$$

We can set

$$\begin{cases} \omega^1 = \sqrt{A}\,\omega_{23}, \\ \omega^2 = \sqrt{A}\,\omega_{31}, \\ \omega^3 = \sqrt{A}\,\omega_{12}. \end{cases}$$

This implies that

$$\begin{cases} d\omega^1 = -\dfrac{1}{\sqrt{A}}\,\omega^2 \wedge \omega^3, \\ d\omega^2 = -\dfrac{1}{\sqrt{A}}\,\omega^3 \wedge \omega^1, \\ d\omega^3 = -\dfrac{1}{\sqrt{A}}\,\omega^1 \wedge \omega^2. \end{cases}$$

Suppose that

$$\omega_{23} = \lambda\,\omega_1, \;\; \omega_{31} = \lambda\,\omega_2, \;\; \omega_{12} = \lambda\,\omega_3.$$

Then

$$d\omega^1 = -\omega^2 \wedge \omega_1^2 - \omega_3^1 \wedge \omega^3 = -2\lambda\,\omega^2 \wedge \omega^3.$$

Thus,

$$\lambda = \frac{1}{2\sqrt{A}}$$

and

$$\omega_{23} = \frac{1}{2\sqrt{A}}\,\omega_1, \;\; \omega_{31} = \frac{1}{2\sqrt{A}}\,\omega_2, \;\; \omega_{12} = \frac{1}{2\sqrt{A}}\,\omega_3.$$

It follows that

$$\begin{aligned} \Omega_{23} &= -d\omega_{23} - \omega_{21} \wedge \omega_{13}, \\ &= -\frac{1}{2A}\,\omega^2 \wedge \omega^3 + \frac{1}{4A}\,\omega^2 \wedge \omega^3 = -\frac{1}{4A}\,\omega^2 \wedge \omega^3. \end{aligned}$$

Finally, we have

$$K_{ij} = 0, \;\; i \neq j, \;\; K_{ii} = \frac{1}{4A} \;\; (i \text{ not summed}).$$

We have a space of constant Riemannian curvature $\frac{1}{4\sqrt{A}}$.

98. Determination of the Riemann–Christoffel tensor for a Riemannian curvature given for all planar directions.

Theorem 2. *If a Riemannian curvature is given for all planar directions emanating from a point A, then all components $R_{ij,kh}$ of the Riemann-Christoffel tensor can be found.*

Proof. Note that this is not obvious: the quantities p^{ij} are not arbitrary, since they are the components of a bivector and are connected by certain known relations. Thus, one cannot claim *a priori* that knowing the curvature

$$R_{ij,kh}p^{ij}p^{kh}$$

implies knowing the components $R_{ij,kh}$.

In order to prove the theorem, we need to show that if the equation

$$R_{ij,kh}p^{ij}p^{kh} = R'_{ij,kh}p^{ij}p^{kh}$$

holds for all p^{ij}, then this implies that

$$R_{ij,kh} = R'_{ij,kh}.$$

In other words, if we set

$$S_{ij,kh} = R_{ij,kh} - R'_{ij,kh},$$

then it is sufficient to prove that the equation

$$S_{ij,kh}\, p^{ij}p^{kh} = 0 \tag{16.18}$$

implies that

$$S_{ij,kh} = 0.$$

Introduce the vectors **X** and **Y** defining a bivector P with components

$$P^{ij} = X^iY^j - X^jY^i.$$

Then equation (16.18) takes the form

$$S_{ij,kh}\,(X^iY^j - X^jY^i)(X^kY^h - X^hY^k) = 0,$$

where all vectors X^i and Y^j are linearly independent. The last equation can be written as

$$S_{ij,kh}X^iY^jX^kY^h = 0.$$

In this sum (where the summation is carried out over all the indices i, j, k, and h), the term $(X^i)^2(Y^j)^2$ occurs only once. Thus, the coefficient of this term vanishes,

$$S_{ij,ij} = 0.$$

The term $(X^i)^2 Y^j Y^k$ occurs twice. Thus,

$$S_{ij,ih} + S_{ih,ij} = 0.$$

But using (16.11), we can easily prove that

$$S_{ij,ih} = S_{ih,ij}.$$

Hence,

$$S_{ij,ih} = S_{ih,ij} = 0.$$

Finally, the term $X^i X^k Y^j Y^h$ occurs four times and gives

$$S_{ij,kh} + S_{kj,ih} + S_{ih,kj} + S_{kh,ij} = 0,$$

or applying (16.11), we find that

$$2S_{ij,kh} + 2S_{ih,kj} = 0,$$

or

$$S_{ij,kh} = S_{ih,jk}.$$

Similarly,

$$S_{ik,hj} = S_{ij,kh}.$$

Thus, in the equations

$$S_{ij,kh} + S_{ik,hj} + S_{ih,jk} = 0,$$

which follows from (16.10), all terms are equal to one another. As a result, we have

$$S_{ij,kh} = 0.$$

99. Isotropic n-dimensional manifold. A manifold is called *isotropic* at a point A if the Riemannian curvature is the same in all directions, i.e., if

$$K = -\frac{R_{ij,kh}\, p^{ij} p^{kh}}{\sum\limits_{(i,j)} (p^{ij})^2}$$

and K does not depend on p^{ij}, or if

$$R_{ij,kh}\, p^{ij} p^{kh} + K \sum_{(i,j)} (p^{ij})^2 = 0$$

for all bivectors p^{ij}.

This reduces to Theorem 1 if we set

$$R'_{ij,ij} = -K, \;\; R'_{ij,kh} = 0.$$

As we proved, the components $R'_{ij,ij}$ must be equal to the corresponding components $R_{ij,ij}$.

Thus,

$$R_{ij,ij} = -K,$$

and all remaining components vanish,

$$R_{ij,kh} = 0.$$

Hence,

$$\Omega_{ik} = -K\omega^i \wedge \omega^j.$$

This is again Schur's theorem but now this theorem is extended to all values $n \geq 3$.

100. Curvature in two different two-dimensional planar directions. Let q_{ij} and q'_{ij} be rotations associated with two surface elements p^{ij} and p'^{ij}. We have

$$p^{ij} q'_{ij} = R_{ij,kh}\, p'^{kh} p^{ij}.$$

The number

$$-\frac{R_{ij,kh}\, p'^{kh} p^{ij}}{\sqrt{\sum\limits_{(i,j)} (p^{ij})^2}\sqrt{\sum\limits_{(i,j)} (p'^{ij})^2}}$$

is called the *mixed curvature* in two directions. The curvature K is a special case, when two directions coincide.

101. Riemannian curvature in a direction of arbitrary dimension. Until now we have considered only plane directions defined by bivectors.

Take now a three-dimensional element with which a system of trivectors

$$\Omega_{ijk} = \omega_i \wedge \Omega_{jk} + \omega_j \wedge \Omega_{ki} + \omega_k \wedge \Omega_{ij}$$

is associated. Consider a surface bounding this element, and attach a sliding (not free, as before) scalar bivector to each surface element.

The geometric sum of these bivectors gives a free trivector Ω_{ij}.

Consider the direction p^{ijk}, to which we attach a system of trivectors Ω_{ijk}, and consider the expression

$$K = -\frac{p^{ijk}\,\Omega_{ijk}}{\sqrt{\sum\limits_{(i,j,k)} (p^{ijk})^2}} = -\frac{p^{ijk}\, R_{ik,hl}\, p_i^{hl}}{\sqrt{\sum\limits_{(i,j,k)} (p^{ijk})^2}} = -\frac{R_{ik,hl}\, p^{ijk}\, p_i^{hl}}{\sqrt{\sum\limits_{(i,j,k)} (p^{ijk})^2}}.$$

This is the corresponding Riemannian curvature.

The Riemannian curvature in a four-dimensional direction is defined in a similar way:

$$K = -\frac{R_{ik,kh}\, p^{ij,\rho\sigma}\, p_{\rho\sigma}^{kh}}{\sqrt{\sum\limits_{(i,j,k)} (p^{ijk})^2}}.$$

102. Ricci tensor. Einstein's quadric. The important results are obtained for the dimension $n-1$.

An $(n-1)$-dimensional direction is defined by means of an $(n-1)$-vector, or by means of an orthogonal vector. If α_i are the direction cosines of the normal, we can set

$$p_{234\ldots n} = \alpha_1\, d\sigma, \quad p_{134\ldots n} = -\alpha_2\, d\sigma, \ldots \; .$$

Then, in the expression of the curvature, the numerator is

$$K_{[i,j]}\sigma^i\sigma^j,$$

where

$$K_{12} = R_1{}^k{}_{2k}, \quad K_{11} = -R_{23,23} - R_{34,34} - \ldots \; .$$

Define the *Ricci tensor*

$$R_{ij} = R_i{}^k{}_{jk}.$$

We have

$$K_{ij} = R_{ij}, \quad i \neq j,$$

$$K_{11} = R_{11} - R^{ij}_{\cdot\cdot,ij} = R_{11} - \frac{1}{2}\sum_i R_{ii}.$$

The tensor K_{ij} can be called the *Einstein tensor.*

We return to the notations of the case $n=3$. We can define the *Einstein quadric*

$$K_{ij}X^iX^j = 1$$

and the *isotropy of the second kind* (the Einstein quadric is a sphere). There is also a theorem similar to Schur's theorem.

The notion of a space of constant curvature of the second kind is more general (only $\frac{n(n-1)}{2}$ coefficients).

This page is intentionally left blank

Chapter 17

Spaces of Constant Curvature

103. Congruence of spaces of the same constant curvature. The structure equations of a space of constant curvature are

$$\begin{cases} d\omega^i = \omega^k \wedge \omega^i_k, \\ d\omega^j_i = \omega^k_i \wedge \omega^j_k - K\,\omega^i \wedge \omega^j. \end{cases} \tag{17.1}$$

All coefficients in these equations are constant for an arbitrary choice of an orthonormal frame.

Theorem 1. *If there exists a space of a given constant curvature K, then such a space is unique.*

This means the following: if there are two spaces with curvature K and line elements ds^2 and $d\sigma^2$, then in some domain by means of an appropriate change of variables, we can reduce ds^2 to coincide with $d\sigma^2$. Take a neighborhood of a point M_0 of one of these spaces and an arbitrary point M'_0 of the second space. Then we can superpose the neighborhood of the point M_0 on the second space in such a way that the point M_0 coincides with the point M'_0 and a frame (R_0) associated with the point M_0 coincides with a frame (R'_0) associated with the point M'_0.

Proof. Suppose that $u^1, \dots, u^N$, where $N = \frac{n(n+1)}{2}$, are parameters defining the most general frame at the point M_0, and $v^1, \dots, v^N$ are similar parameters at the point M'_0. At the point M_0, we have equations (17.1), and at the point M'_0, we have the similar equations

$$\begin{cases} d\overline{\omega}^i = \overline{\omega}^k \wedge \overline{\omega}^i_k, \\ d\overline{\omega}^j_i = \overline{\omega}^k_i \wedge \overline{\omega}^j_k - K\,\overline{\omega}^i \wedge \overline{\omega}^j. \end{cases} \tag{17.2}$$

Establish a correspondence between the points of two spaces by means of the equations

$$\begin{cases} \overline{\omega}^i(v, dv) = \omega^i(u, du), \\ \overline{\omega}_i^j(v, dv) = \omega_i^j(u, du), \end{cases} \tag{17.3}$$

where the v's must be considered as unknown functions of the u's. The system of equations (17.3) is completely integrable, since by equations (17.1) and (17.2), the equations

$$\begin{cases} d\overline{\omega}^i - d\omega^i = 0, \\ d\overline{\omega}_i^j - d\omega_i^j = 0 \end{cases} \tag{17.4}$$

are algebraic consequences of system (17.3).

In particular, system (17.4) admits solutions for which an initial system of values (u_0) corresponds to an arbitrary system of values (v_0), i.e., two arbitrary frames of the first and second family correspond to one another.

Since now

$$\begin{cases} v^1 = f^1(u^1, \dots, u^N), \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ v^N = f^N(u^1, \dots, u^N), \end{cases}$$

we have a point correspondence between the two spaces.

In fact, the system of equations $\omega^i = 0$, $i = 1, \dots, n$, is completely integrable. Take $u^1, \dots, u^n$ as first integrals of this system. For constant values (u^i), we have $\omega^i = 0$, and the point M_0 of the first space is fixed. Thus, u^i are curvilinear coordinates of the first space. Take $v^1, \dots, v^n$ as those values of the first integrals of the system $\overline{\omega}^i = 0$ that correspond to the values (u^i) obtained in integrating of system (17.3). Since

$$\overline{\omega}^i = \omega^i, \tag{17.5}$$

v^i depend on $u^1, \dots, u^n$ only. To a point of the first space, there is a corresponding point of the second.

Finally, equations (17.5) imply the coincidence of the line elements,

$$ds^2 = \sum_i (\omega^i)^2 \text{ and } d\sigma^2 = \sum_i (\overline{\omega}^i)^2.$$

Corollary 2. *A space of constant curvature is applicable onto itself in an infinite number of ways: this space admits infinitely many motions (transformations preserving the distances). As with the Euclidean space, a space of constant curvature admits* ∞^N *displacements, where* $N = \frac{n(n+1)}{2}$.

REMARK. The spaces of constant curvature are the most general among the spaces admitting the group of motions. In fact, since the Riemannian curvature is defined by a line element ds^2, for displacements preserving ds^2, the Riemannian curvature is also preserved. In particular, a motion with a fixed arbitrary point A is possible. Thus, at this point A, the curvature in all directions is the same. Hence, the space is isotropic and has a constant curvature.

104. Existence of spaces of constant curvature. We pose this problem in another, more general form. The system of equations (17.1) is a particular case of a more general system of equations

$$d\omega^i = c^i_{[jk]}\omega^j \wedge \omega^k, \tag{17.6}$$

where

$$c^i_{[jk]} = c^i_{jk} - c^i_{kj}$$

are constants. The question is: does system (17.1) admit a solution ω^i ?

We obtain a necessary condition of existence of a solution, which is similar to the Bianchi identities, if we apply exterior differentiation to system (17.6):

$$0 = c^i_{[jk]}\,(d\omega^j \wedge \omega^k - \omega^j \wedge d\omega^k).$$

Changing in the second sum the summation indices $\binom{j\,k}{k\,j}$ and taking into account the fact that the coefficients $c^i_{[jk]} = c^i_{jk} - c^i_{kj}$ are skew-symmetric and the condition $\omega^j \wedge d\omega^k = d\omega^k \wedge \omega^j$ ($d\omega^k$ is of degree two), we obtain

$$0 = c^i_{[jk]}\, d\omega^j \wedge \omega^k.$$

Insert here the expression of $d\omega^i$ from formula (17.6). This gives

$$c^i_{[jk]}\, c^j_{[hl]}\, \omega^h \wedge \omega^j \wedge \omega^k = 0.$$

To collect similar terms, we make the circular substitution of indices h, l, k. This gives

$$(c^i_{[jk]}c^j_{[hl]} + c^i_{[jh]}c^j_{[lk]} + c^i_{[jl]}c^j_{[kh]})\,\omega^h \wedge \omega^j \wedge \omega^k = 0,$$

or finally

$$c^i_{[jk]}c^j_{[hl]} + c^i_{[jh]}c^j_{[lk]} + c^i_{[jl]}c^j_{[kh]} = 0. \tag{17.7}$$

Equations (17.6) are called the *structure equations of the group of transformations,* and equations (17.7) are the *conditions for the structure constants* $c^i_{[jk]}$ *of the group.* The group transformations are obtained from the equations

$$\omega^i(u, du) = \omega^i(v, dv), \quad i = 1, 2, \dots, r,$$

which, by equations (17.6), are completely integrable. Here (u^i) are the original coordinates of a point, and (v^i) are the coordinates of the transformed point.

We prove that conditions (17.7) are sufficient for the existence of r independent Pfaffian forms satisfying system (17.6). In group theory this corresponds to the third Sophus Lie theorem on existence of the group of transformations with given structure coefficients $c^i_{[jk]}$ satisfying conditions (17.7).

105. Proof of Schur. Suppose that we have a solution

$$\omega^i = A^i_k du^k,$$

where A^i_k are functions of independent variables u^k. Since the system of forms ω^i is linearly independent, the determinant of its coefficients is necessarily different from zero at the initial point:

$$\det(A^i_k) \neq 0 \text{ for } u^k = 0. \tag{17.8}$$

Introduce a redundant number $r+1$ variables instead of former r variables by setting

$$u^k = v^k t.$$

Then

$$\omega^i = A^i_k \, v^k \, dt + tA^i_k \, dv^k.$$

For $t = 0$, all coefficients in the differential dt are reduced to r functions

$$A^i_k(0) \, v^k,$$

that are independent by condition (17.8).

Define the quantities

$$a^i = A^i_k \, v^k, \tag{17.9}$$

which can be considered as new variables. So, now we have

$$\omega^i = a^i \, dt + \overline{\omega}^i(a, t, da), \tag{17.10}$$

where $\overline{\omega}^i$ are linear with respect to da^k and vanish if $t = 0$. We prove that for system (17.6), it is possible to construct solutions having the form of equations (17.10).

Taking exterior differentials of equations (17.10), we find that

$$d\omega^i = da^i \wedge dt + dt \wedge \frac{\partial \overline{\omega}^i}{\partial t} + \text{terms not depending on } dt. \tag{17.11}$$

In fact, by hypothesis, the form $\overline{\omega}^i$ can be written as

$$\overline{\omega}^i = M^i_k \, da^k.$$

Thus,

$$\begin{aligned} d\overline{\omega}^i &= \left(\frac{\partial M^i_k}{\partial t} \, dt \right) \wedge da^k = \frac{\partial M^i_k}{\partial t} \, dt \wedge da^k \\ &= dt \wedge \left(\frac{\partial M^i_k}{\partial t} \, da^k \right) = dt \wedge \frac{\partial \overline{\omega}^i}{\partial t}. \end{aligned}$$

For a solution of equations (17.6), the terms containing dt on the right-hand side of (17.6) are

$$c^i_{[jh]}\, dt \wedge (a^j\, \overline{\omega}^h - a^h\, \overline{\omega}^j) = c^i_{jh}\, a^j\, dt \wedge \overline{\omega}^h, \tag{17.12}$$

where on the right-hand side the summation is carried out with respect to the indices j and h separately.

Since t is the independent variable, the differential dt is also completely arbitrary. Thus if we substitute expressions (17.11) and (17.12) into equations (17.6), then the coefficients of dt on the left- and right-hand side of the equations obtained must coincide. The terms with dt are

$$da^i \wedge dt + dt \wedge \frac{\partial \overline{\omega}^i}{\partial t} = c^i_{jk}\, a^j\, dt \wedge \overline{\omega}^k.$$

It follows that

$$\frac{\partial \overline{\omega}^i}{\partial t} - da^i = c^i_{jk}\, a^j\, \overline{\omega}^k. \tag{17.13}$$

Thus, the forms $\overline{\omega}^i$, considered as functions of t, satisfy a system of ordinary differential equations with constant coefficients. Therefore, there exists a unique solution of this system satisfying the initial conditions

$$\overline{\omega}^i = 0 \text{ for } t = 0.$$

Therefore, if there exists a solution, we can actually compute it.

106. The system is satisfied by the solution constructed. Suppose now that a solution $\overline{\omega}^i$ is obtained. We insert it into equations (17.10) and prove that system (17.6) is identically satisfied. In other words, if we set

$$d\omega^i = c^i_{[jk]}\, \omega^j \wedge \omega^k + \Omega^i, \tag{17.14}$$

we must obtain

$$\Omega^i = 0.$$

In fact, we apply exterior differentiation to equations (17.14) and use the fact that the terms not containing Ω^i vanish by the imposed conditions which we assume are fulfilled. As a result, we obtain

$$d\Omega^i + c^i_{jk}\, \Omega^j \wedge \omega^k = 0. \tag{17.15}$$

Introduce the symbol d for differentiation with respect to the variable t and the symbol δ_α for differentiation with respect to the variable a^α.

We prove that

$$\Omega^i(d, \delta_\alpha) = 0, \quad \Omega^i(\delta_\alpha, \delta_\beta) = 0. \tag{17.16}$$

To say that

$$\Omega^i(d, \delta_\alpha) = 0$$

is the same as to say that the terms containing dt are the same on the left- and right-hand sides of equations (17.1). But comparison of these exact terms produced equations (17.3) whose integration gave the forms $\overline{\omega}^i$.

In order to justify the second equation of (17.16), we apply equations (17.14) to compute the bilinear form $\Omega^i(\delta_\alpha, \delta_\beta)$ for $t = 0$. We have

$$\begin{aligned}\Omega^i(\delta_\alpha, \delta_\beta) = \delta_\alpha \omega^i(\delta_\beta) - \delta_\beta \omega^i(\delta_\alpha)\\ -c^i_{[jk]}\left[\omega^j(\delta_\alpha)\,\omega^k(\delta_\beta) - \omega^j(\delta_\beta)\,\omega^k(\delta_\alpha)\right] = 0,\end{aligned}$$

since for $t = 0$,

$$\omega^i(\delta_\alpha) = \omega^i(\delta_\beta) = 0.$$

Now the trilinear covariant with the symbols of differentiation $d, \delta_\alpha, \delta_\beta$ associated with equations (17.15) can be written as

$$\begin{aligned}d\,\Omega^i(\delta_\alpha, \delta_\beta) \ &+\delta_\alpha\,\Omega^i(\delta_\beta, d) + \delta_\beta\,\Omega^i(d, \delta_\alpha) + c^i_{[jk]}\left[\Omega^j(d, \delta_\alpha)\,\omega^k(\delta_\beta)\right.\\ &\left.+\Omega^j(\delta_\alpha, \delta_\beta)\,\omega^k(d) + \Omega^j(\delta_\beta, d)\,\omega^k(\delta_\alpha)\right] = 0.\end{aligned}$$

Here, by the first (already proved) identity (17.16), all forms Ω^i for two symbols of differentiation d and δ vanish.

We obtain

$$d\,\Omega^i(\delta_\alpha, \delta_\beta) + c^i_{jk}\,\Omega^j(\delta_\alpha, \delta_\beta)\,\omega^k(d) = 0.$$

If we substitute here the values of $\omega^k(d)$ taken from (17.10) and set

$$H^i = \Omega^i(\delta_\alpha, \delta_\beta),$$

then our equations become a system of ordinary differential equations:

$$\frac{dH^i}{dt} + c^i_j k\, H^j\, a^k = 0.$$

This is a linear homogeneous system. Thus, it admits the solution

$$H^i = 0.$$

But since we proved that for $t = 0$,

$$\Omega^i(\delta_\alpha, \delta_\beta) = 0,$$

it follows that the above solution is the only possible solution for any t.

Thus, we have proved the second identity of (17.16). This immediately implies the vanishing of the exterior quadratic form Ω^i.

Now we need only to prove that all the forms ω^i we have found are linearly independent.

Note that the variables a^i and t occur only as the products a^it. In fact, equations (17.3), from which we found the forms $\overline{\omega}^i$, admit the substitution

$$t = k\,t', \;\; a^i = \frac{\alpha^i}{k}, \;\; k = \text{const.}$$

In order to integrate system (17.1), we integrate system (17.3) with the initial conditions

$$\overline{\omega}^i = 0 \;\text{ for }\; t = 0.$$

We get the forms ω^i if we substitute $t = 1$ in the obtained solution $\overline{\omega}^i$: since

$$\overline{\omega}^i = A^i_k\,du^k, \;\; \overline{\omega}^i = tA^i_k\,dv^k, \;\;\; u^k = v^k t,$$

we have

$$\overline{\omega}^i = \omega^i \;\text{ for }\; t = 1.$$

Thus, we can make vanish those of a^i from (17.9) that correspond to the forms ω^j_i with two indices.

In fact, from equations (17.10), we have

$$\omega^i = a^i\,dt + \overline{\omega}^i, \;\; \overline{\omega}^j_i = \omega^j_i.$$

1°. For example, if $K = 0$ (the Euclidean space), the system of equations (17.1) becomes

$$d\omega^i = \omega^k \wedge \omega^i_k, \;\; d\omega^j_i = \omega^k_i \wedge \omega^j_k, \tag{17.17}$$

and system (17.13) takes the form

$$\left\{\begin{aligned} &\frac{d\overline{\omega}^i}{dt} = da^i + a^k\,\overline{\omega}^i_k, \\ &\frac{d\overline{\omega}^j_i}{dt} = 0 \end{aligned}\right. \tag{17.18}$$

In fact, compare equations (17.17) with system (17.6),

$$d\omega^i = c^i_{[jk]}\,\omega^j \wedge \omega^k,$$

and set

$$\omega^j_i = \omega^\alpha, \quad \alpha = n+1, \dots, N.$$

Note that

a) The coefficients $c^i_{[jk]}$ are equal to either 0 or 1.

b) For $i \leq n$ (the case of the first of equations (17.17)), the index j also satisfies the inequality $j \leq n$, and the index k satisfies the inequality $k \geq n+1$.

c) For $i > n$ (the case of the second of equations (17.17)), both indices j and k satisfy the inequality $j, k > n$, but for such large j, all $a^j = 0$.

Integrating system (17.18), we find that

$$\overline{\omega}_i^j = 0, \quad \overline{\omega}^i = t\, da^i,$$

and as a result, we have

$$\omega^i = da^i, \quad ds^2 = \sum_i (da^i)^2.$$

(This proves one more time that the forms ω^i are linearly independent.)

2°. If $K \neq 0$, then by (17.1), we obtain

$$\begin{cases} \dfrac{d\overline{\omega}^i}{dt} = da^i + a^k\, \overline{\omega}_k^i, \\ \dfrac{d\overline{\omega}_i^j}{dt} = -K\,(a^i \overline{\omega}^j - a^j\, \overline{\omega}^i). \end{cases} \tag{17.19}$$

Multiplying the first equation of (17.19) by a_i and summing with respect to the index i, we find that

$$a_i\, \frac{d\overline{\omega}^i}{dt} = a_i\, da^i, \tag{17.20}$$

since

$$a_i\, a^k\, \overline{\omega}_k^i = \sum_{(i,k)} (a_i\, a^k\, \overline{\omega}_k^i + a_k\, a^i\, \overline{\omega}_i^k) = \sum_{(i,k)} a_i\, a_k\, (\overline{\omega}_k^i + \overline{\omega}_i^k) = 0.$$

Integrating equation (17.20) with respect to t and taking into account that a^i does not depend on t, for the initial conditions $\omega^i = 0$ for $t = 0$, we obtain

$$a_i \overline{\omega}^i = t\, a_i\, da^i. \tag{17.21}$$

Now, differentiating the first equation of (17.19) with respect to t, we get

$$\frac{d^2\overline{\omega}^i}{dt^2} = a^k\, \frac{d\overline{\omega}_k^i}{dt} = -K \sum_{(i,k)} a^k\, (a_k\, \overline{\omega}^i - a_i\, \overline{\omega}^k)$$

$$= -K\, a^k a_k\, \overline{\omega}^i + K\, a_i\, t\, a_k\, da^k.$$

The first transformation was made by means of the second equation of (17.19), and the second by equations (17.21).

Thus,

$$\frac{d^2\overline{\omega}^i}{dt^2} = -K\, a^k a_k\, \overline{\omega}^i + \varphi(t),$$

and we can integrate with the independent variable t and the unknown functions $\overline{\omega}^i$.

Chapter 18

Geometric Construction of a Space of Constant Curvature

107. Spaces of constant positive curvature. We can obtain a space of constant curvature by considering a hypersphere in an $(n+1)$-dimensional Euclidean space. We prove now that a hypersphere of radius R has the constant curvature $\frac{1}{R^2}$.

In an $(n+1)$-dimensional Euclidean space, consider a rectangular Cartesian coordinate system with an origin O and coordinates $x^1, \ldots, x^{n+1}$. The line element of this space is

$$ds^2 = (dx^1)^2 + (dx^2)^2 + \ldots + (dx^{n+1})^2.$$

The equation of the hypersphere with center at the origin and radius R is

$$(x^1)^2 + (x^2)^2 + \ldots + (x^{n+1})^2 = R^2.$$

To the fixed point O, we attach an orthonormal frame with unit basis vectors $\mathbf{I}_1, \ldots, \mathbf{I}_{n+1}$. Denote by M the common point of the hypersphere and the position vector of the direction $\mathbf{I}_{n+1}$. Then

$$\overrightarrow{OM} = R\mathbf{I}_{n+1}. \tag{18.1}$$

If the origin of our frame is fixed, the frame displacement is determined by the forms ω_α^β, $\alpha, \beta, \gamma = 1, \ldots, n+1$, satisfying the structure equations

$$d\omega_\alpha^\beta = \omega_\alpha^\gamma \wedge \omega_\gamma^\beta.$$

In what follows, the Greek letters take the values $\alpha, \beta, \gamma = 1, \ldots, n+1$, and the Latin letters take the values $i, j, k = 1, \ldots, n$. The equations of infinitesimal displacement of the point $\mathbf{M}$ are obtained by differentiation of equation (18.1):

$$d\mathbf{M} = R\, d\mathbf{I}_{n+1} = R\,(\omega_{n+1}^1 \mathbf{I}_1 + \ldots + \omega_{n+1}^n \mathbf{I}_n).$$

Thus the components of infinitesimal displacement of the point **M** are

$$\omega^i = R\omega^i_{n+1}, \tag{18.2}$$

from which it follows that

$$\omega^i_{n+1} = \frac{1}{R}\omega^i. \tag{18.3}$$

Exterior differentiation of equations (18.2) gives the following exterior quadratic equations:

$$\frac{d\omega^i}{R} = \omega^k_{n+1} \wedge \omega^i_k = \frac{1}{R}\omega^k \wedge \omega^i_k.$$

Thus,

$$d\omega^i = \omega^k \wedge \omega^i_k. \tag{18.4}$$

Similarly,

$$d\omega^j_i = \omega^k_i \wedge \omega^j_k + \omega^{n+1}_i \wedge \omega^j_{n+1},$$

so from (18.3) and $\omega^{n+1}_i = -\omega^i_{n+1}$, we have

$$d\omega^j_i = \omega^k_i \wedge \omega^j_k - \frac{1}{R^2}\omega^i \wedge \omega^j. \tag{18.5}$$

Hence, the curvature of the space is

$$K = \frac{1}{R^2}.$$

With the help of (18.1) and (18.2), we can easily find that

$$d\mathbf{M} = \omega^i \mathbf{I}_i. \tag{18.6}$$

Other equations of infinitesimal displacement are

$$d\mathbf{I}_\alpha = \omega^\beta_\alpha \mathbf{I}_\beta.$$

In addition, again using equations (18.1) and (18.2), we find that

$$d\mathbf{I}_i = \omega^k_i \mathbf{I}_k + \omega^{n+1}_i \mathbf{I}_{n+1} = \omega^k_i \mathbf{I}_k - \frac{\omega^i}{R^2}\overrightarrow{OM},$$

or

$$d\mathbf{I}_i = \omega^k_i \mathbf{I}_k - K\omega^i \overrightarrow{OM}. \tag{18.7}$$

We will call an *analytic point* a set of coordinates

$$x^1, x^2, \dots, x^{n+1}.$$

The *scalar square of this point* is

$$(x^1)^2 + (x^2)^2 + \dots (x^{n+1})^2 = \mathbf{M}^2 = \frac{1}{K}.$$

Thus, we have the following *scalar products*:

$$\mathbf{M}^2 = \frac{1}{K}, \quad (\mathbf{I}_i)^2 = 1, \quad \overrightarrow{OM} \cdot \mathbf{I}_i = 0, \quad \mathbf{I}_i \cdot \mathbf{I}_j = 0. \tag{18.8}$$

The differentials $d\mathbf{M}$ and $d\mathbf{I}_i$ allow us to find the differentials of any vector considered as a point having a certain scalar square: $\mathbf{X} = X^i\mathbf{I}_i$.

108. Mapping onto an n-dimensional projective space. In order to interpret the concept of negative curvature, it is convenient to consider $n+1$ numbers as homogeneous coordinates of a point of an elliptic space of dimension n. Here the points $(x^1, x^2, \dots, x^{n+1})$ and $(-x^1, -x^2, \dots, -x^{n+1})$ are considered identical.[1] Thus, an elliptic space composes only half of a spherical space.

Change the homogeneous coordinates in the space and suppose that

$$(x^1)^2 + (x^2)^2 + \dots (x^{n+1})^2 = F\left(x^{1'}, x^{2'}, \dots, x^{(n+1)'}\right)$$

and

$$ds^2 = F\left(dx^{1'}, dx^{2'}, \dots, dx^{(n+1)'}\right),$$

where F is a homogeneous second-degree polynomial (a form).

As a result, we obtain the definition of the space by means of the quadratic form F satisfying the condition

$$F\left(x^{1'}, x^{2'}, \dots, x^{(n+1)'}\right) = \frac{1}{K}. \tag{18.9}$$

We started from a definite form, but the form F could be also indefinite provided that ds^2 is positive. In this way we arrive at a hyperbolic ($K < 0$) space. We prove that this is possible only if the form F is a sum of n squares with positive signs and one square with negative sign.

A form F must be positive for all values dx related by the condition

$$\frac{\partial F}{\partial x^1} dx^1 + \dots + \frac{\partial F}{\partial x^{n+1}} dx^{n+1} = 0. \tag{18.10}$$

This equation is obtained by differentiation of equation (18.9) and shows that the point $(dx^1, \dots, dx^{n+1})$ lies in the polar hyperplane of the point $(x^1, \dots, x^{n+1})$ with respect to the absolute

$$F = 0. \tag{18.11}$$

If the polar hyperplane intersects the absolute, then the differential form

$$ds^2 = F\left(dx^1, dx^2, \dots, dx^{n+1}\right) \tag{18.12}$$

---

[1] For $n = 2$, the elliptic plane is obtained as the projection of a sphere from its center onto a tangent plane.

changes sign.

In a three-dimensional space, the last requirement excludes the ruled surfaces and admits only two cases:

1°. *The absolute is an imaginary quadric.* The form F is positive definite. The curvature K is positive. The space is elliptic.

2°. *The absolute is a real nonruled quadric*, for example, an ellipsoid. In order for the polar plane of a point to have no common points with the quadric, one must take a point (pole) inside the ellipsoid. The curvature K of the space is negative. The space is hyperbolic. Thus, a hyperbolic space is represented by interior points of the absolute. The form F decomposes into three squares with plus signs and one square with a minus sign.

109. Hyperbolic space. Let us show that for an n-dimensional hyperbolic space, we can choose the form F composed of n squares with plus signs and one square with a minus sign.

Consider a point $\mathbf{P}(0,\dots,0,1)$. In order for the coordinates of this point to satisfy equation (18.9), it is sufficient to take F of the form

$$F = \frac{(x^{n+1})^2}{K} + 2a_i x^i x^{n+1} + a_{ij} x^i x^j.$$

If we change the coordinate x^{n+1} by taking a new coordinate

$$x^{n+1} + K a_i x^i,$$

then the coordinates of the point $\mathbf{P}$ are not changed. Assuming that we made this coordinate transformation, we have

$$F = \frac{(x^{n+1})^2}{K} + \Phi(x^1,\dots,x^n).$$

At the point $\mathbf{P}$, linear equation (18.10) reduces to one term

$$dx^{n+1} = 0,$$

and line element (18.12) at the point $\mathbf{P}$ is positive if

$$\Phi(dx^1,\dots,dx^n) > 0.$$

Assuming that the form Φ is positive definite, we arrive at elliptic geometry if $K > 0$ and at hyperbolic geometry if $K < 0$.

Moreover, the points of the hyperbolic space are represented only by those points of the Euclidean space whose coordinates satisfy the inequality

$$F = \frac{(x^{n+1})^2}{K} + \Phi(x^1,\dots,x^n) < 0,$$

i.e., by the interior points of the absolute.

110. Representation of vectors in hyperbolic geometry. Let Q be a hyperquadric of the absolute, and let $\mathbf{M}(x^i)$ and $\mathbf{M}'(x^i + dx^i)$ be two internal neighboring points.

We can consider dx^i as coordinates of a point satisfying equation (18.10) which shows that the point (dx^i) is located in the polar hyperplane of the point $\mathbf{M}$ with respect to the absolute.

We have the same result for the velocity vector $\frac{d\mathbf{X}}{dt}$. The square of the length of the vector X^i is

$$F(x^1, x^2, \dots, x^{n+1}).$$

This is the scalar square of the point representing the vector. Thus, *any vector is represented by a point situated in the polar hyperplane of the initial point of the vector with respect to the absolute. The scalar square of this point is equal to the scalar square of the vector.*

The scalar product of two vectors with a common initial point is equal to the scalar product of the points representing these vectors.

Two vectors are orthogonal if the points representing them are conjugate, i.e., if the directions of these two vectors are conjugate with respect to the absolute.

The case $n = 2$. There are infinitely many straight lines through a point that are parallel to a given straight line, i.e., there are infinitely many straight lines that do not have common points inside the absolute. Two perpendiculars to a straight line are parallel, and so on.

In the elliptic interpretation, for the vector

$$\mathbf{X} = X^i \mathbf{I}_i,$$

we have

$$\begin{aligned} d\mathbf{X} &= dx^i\,\mathbf{I}_i + x^i\,d\mathbf{I}_i = dx^i\,\mathbf{I}_i + x^i\,(-K\,\omega^i\,\overrightarrow{OM} + \omega_i^k\,\mathbf{I}_k) \\ &= -K\,\omega_i\,X^i\,\overrightarrow{OM} + \mathbf{I}_i\,(dX^i + \omega_k^i X^k) = -K\,\omega_i\,X^i\,\overrightarrow{OM} + \mathbf{I}_i\,DX^i. \end{aligned}$$

It follows that

$$d\mathbf{X} = D\mathbf{X} - K\,\omega_i\,X^i\,\overrightarrow{OM}. \tag{18.13}$$

111. Geodesics in Riemannian manifold. The *acceleration* is the absolute derivative of the velocity vector with respect to time.

As usual, denote by ω^i the components of infinitesimal displacement of a moving point, by t the time, and by v^i the components of the velocity vector. We have

$$\omega^i = v^i\,dt.$$

It follows that the components g^i of the acceleration vector are

$$g^i = \frac{Dv^i}{dt} = \frac{dv^i}{dt} + v^k\,\frac{\omega_k^i}{dt}.$$

Setting

$$\omega_i^j = \gamma_i{}^j{}_k \omega^k,$$

we find that

$$g^i = \frac{dv^i}{dt} + \gamma_k{}^i{}_h v^k v^h.$$

For a geodesic, we have $g^i = 0$. It follows that the equations of geodesics are

$$\begin{cases} \dfrac{dv^i}{dt} + \gamma_k{}^i{}_h \, v^k v^h = 0, \\ \omega^i = v^i \, ds; \end{cases} \tag{18.14}$$

here we set $ds = dt$.

REMARK. Consider a holonomic system with n parameters $q^1, \dots, q^n$. The kinetic energy T is defined by the equation

$$T \, dt^2 = g_{ij} \, dq^i \, dq^j. \tag{18.15}$$

We assume that none of the constraints depend on time.

It is easy to see that instantaneous motions of the system (without exterior forces) are represented by geodesics of the space $q^1, \dots, q^n$.

Form (18.15) was considered by Volterra.[2] He calls v^i the characteristics of the motion.

A state of the system at a given moment is defined by the quantities $(q^i, \dot{q}^i)$, where $\dot{q}^i = \dfrac{dq^i}{dt}$. Volterra replaces $\dot{q}^i$ by linear combinations of v^i.

In particular, Volterra considers the case

$$\sum_i (v^i)^2.$$

This is our point of view.

Volterra poses the following problem: is it possible to choose v^i in such a way that $\gamma_k{}^i{}_h$ are constants (a system with independent characteristics)? If this is possible, then the first group and the second group of equations (18.14) can be integrated separately.

EXAMPLE: the Euler equations (the rotation of a rigid body). It would be interesting to study Riemannian manifolds with constant $\gamma_k{}^i{}_h$.

112. Pseudoequipollent vectors: pseudoparallelism. Is it possible to choose a rectangular frame at every point of a Riemannian manifold in such a way that

$$g^i = \frac{dv^i}{dt}\ ?$$

If this is possible, then for each geodesic, we have

$$v^i = \text{const}.$$

---

[2] Volterra, Atti Torino, 1887.

Thus, the components of the unit tangent vector are constant along a geodesic.

We call two vectors $\overrightarrow{MX}$ and $\overrightarrow{M'X'}$ *pseudoequipollent* if they have the same components with respect to the frames associated with them.

The unit vectors tangent to a geodesic are pseudoequipollent.

Suppose that two vectors tangent to two geodesics are pseudoequipollent at two points. Then they are also pseudoequipollent everywhere. We call such geodesics *pseudoparallel.*

This implies the following:

1°. *Each geodesic is pseudoparallel to itself.*

2°. *Two geodesics pseudoparallel to a third geodesic are pseudoparallel to one another.*

3°. *Through any point, there passes one and only one geodesic which is pseudoparallel to a given geodesic.*

4°. *If two geodesics intersect each other, then two other geodesics, that are pseudoparallel at their common point to the first two geodesics, intersect each other at the same angle.*

We show that the converse is also true: if it is possible to define a pseudoparallelism satisfying the following two (less strong) conditions:

1°). Through any point, there passes one and only one geodesic corresponding to a given geodesic,

2°). Two geodesics corresponding to two intersecting geodesics intersect each other at the same angle as the first pair,

then it is possible to find a frame giving $v^i = \text{const}$.

Proof. At the point O, take an orthonormal frame. To each direction emanating from a point M, associate the components a_1, a_2, a_3 of the unit vector tangent to the corresponding geodesic at the point O. The angle of two such directions (a) and (b) is defined by the formula

$$\cos\varphi = a_1 b_1 + a_2 b_2 + a_3 b_3.$$

Thus, there exists a certain rectangular frame at the point M such that if the direction cosines of the geodesic are a_1, a_2, and a_3, then the direction cosines of the corresponding geodesic at the point O are also a_1, a_2, and a_3.

Therefore, to any geodesic emanating from the point O, there corresponds also a geodesic emanating from the point M. The components of the velocity vector of a moving point along the geodesic are also constants. It follows that

$$\gamma^i_{kh} v^k v^h = 0.$$

In fact, at an arbitrary point M_0, we have

$$\frac{d\overset{\circ}{v}{}^i}{dt} = 0.$$

Using (18.14), the last equations can be written as

$$\overset{\circ}{\gamma}{}^{i}_{kh}\overset{\circ}{v}{}^{k}\overset{\circ}{v}{}^{h} = 0.$$

It follows that two geodesics are pseudoparallel if the same geodesic emanating from the point O corresponds to them.

In Euclidean space we have an ordinary parallelism. We will see that in elliptic space, there are two types of pseudoparallelism.

Consider a rigid body moving with an angular velocity (p, q, r) and the moments $A = B = C = 1$. For the axes connected with this rigid body, we have

$$\frac{dp}{dt} = \frac{dq}{dt} = \frac{dr}{dt} = 0, \quad T = p^2 + q^2 + r^2.$$

It follows that the parallelism is "absolute".

The equations remain the same for fixed axes in the space.

THE FRENET FORMULAS FOR GEODESICS.

If $\mathbf{T}$ is the unit tangent vector at a point M, then

$$\frac{d\mathbf{M}}{ds} = \mathbf{T}, \quad \frac{d\mathbf{T}}{ds} = \mathbf{0}, \tag{18.16}$$

113. Geodesics in spaces of constant curvature. From (18.13), we have

$$d\mathbf{X} = D\mathbf{X} - K\,X^i\omega_i\,\overrightarrow{OM},$$

$$X^i\omega_i = \mathbf{X}\cdot d\mathbf{M}.$$

For $\mathbf{X} = \mathbf{T}$, by (18.16) and $\mathbf{T}\cdot d\mathbf{M} = ds$, we find that

$$d\mathbf{T} = -K\,ds\,\overrightarrow{OM}.$$

This implies that

$$\frac{d\mathbf{T}}{ds} = -K\,\overrightarrow{OM}.$$

Set $\overrightarrow{OM} = \mathbf{M}$. Excluding $\mathbf{T}$ by means of the first equation of (18.16), we obtain

$$\frac{d^2\mathbf{M}}{ds^2} + K\,\mathbf{M} = \mathbf{0}. \tag{18.17}$$

Integrating this equation as a linear ordinary differential equation with constant coefficients, we obtain for $K > 0$

$$\mathbf{M} = \mathbf{A}\,\cos\left(s\sqrt{K}\right) + \mathbf{B}\,\sin\left(s\sqrt{K}\right),$$

where $\mathbf{A}$ and $\mathbf{B}$ are constant points.

The point $\mathbf{M}$ describes the straight line AB.

Thus, the geodesics of an elliptic space are ordinary straight lines.

If $\mathbf{M}_0$ is an initial point of a straight line and $\mathbf{T}_0$ is the point representing an initial direction, then we easily obtain

$$\mathbf{M} = \mathbf{M}_0 \cos\bigl(s\sqrt{K}\bigr) + \frac{\mathbf{T}_0}{\sqrt{K}} \sin\bigl(s\sqrt{K}\bigr). \tag{18.18}$$

The denominator in the second term of the right-hand side is explained by the fact that according to formulas (18.8), the point $\mathbf{M}$ as all analytic points has the norm $\frac{1}{\sqrt{K}}$ while $\mathbf{T}_0$ is the unit vector.

THE TOTAL LENGTH OF A STRAIGHT LINE

1°. *Spherical space*: $K = \frac{1}{R^2}$.

$$S_T = \frac{2\pi}{\sqrt{K}} = 2\pi R. \tag{18.19}$$

Since the trigonometric functions in equation (18.18) are of period 2π, the argument $s\sqrt{K}$ can vary from 0 to 2π. Thus, the arc length s can vary from 0 to $\frac{2\pi}{\sqrt{K}}$.

2°. *Elliptic space*:

$$S_T = \frac{\pi}{\sqrt{K}}. \tag{18.20}$$

In an elliptic space, the points with coordinates of opposite signs are identified. Since if the argument is increased by π, both trigonometric functions in equation (18.18) change their signs, the period for the point $\mathbf{M}$ is π for $s\sqrt{K}$ and $\frac{\pi}{\sqrt{K}}$ for s.

We can multiply both members of equation (18.18) by $\mathbf{M}_0$. Since

$$\mathbf{M}_0 \cdot \mathbf{T}_0 = 0, \ (\mathbf{M}_0)^2 = \frac{1}{\sqrt{K}},$$

we obtain

$$\mathbf{M}_0 \cdot \mathbf{M} = \frac{1}{K} \cos\bigl(s\sqrt{K}\bigr) = R^2 \cos\frac{s}{R}.$$

It follows that for a fixed point $\mathbf{M}_0$, the period for the point $\mathbf{M}$ is $\frac{\pi}{R}$.

3°. *Hyperbolic space*: $K = -\frac{1}{R^2} < 0$. Now when we integrate equations (18.17), we obtain hyperbolic functions instead of trigonometric ones:

$$\begin{aligned} \mathbf{M} &= \mathbf{M}_0 \cosh\bigl(s\sqrt{-K}\bigr) + \frac{\mathbf{T}_0}{\sqrt{-K}} \sinh\bigl(s\sqrt{-K}\bigr) \\ &= \mathbf{M}_0 \cosh\Bigl(\frac{s}{R}\Bigr) + \mathbf{T}_0\, R \sinh\Bigl(\frac{s}{R}\Bigr). \end{aligned} \tag{18.21}$$

If we change the normalization of the analytic point $\mathbf{M}_0$, we can write (18.21) in the form

$$\mathbf{M} = \mathbf{M}_0 + \mathbf{T}_0 \tanh\Big(\frac{s}{R}\Big). \tag{18.22}$$

When s changes from 0 to ∞, the point $\mathbf{M}$ changes from $\mathbf{M}_0$ to $\mathbf{M}_0 + \mathbf{T}_0 \tanh(\frac{s}{R})$. The last point lies on the absolute since

$$(\mathbf{M}_0 + \mathbf{T}_0 R)^2 = (\mathbf{M}_0)^2 + R^2 = 0.$$

Thus, the absolute is the locus of the points at infinity of straight lines. Each straight line has two points at infinity, $\mathbf{P}(\infty)$ and $\mathbf{Q}(-\infty)$. When the argument s changes sign, the hyperbolic tangent changes sign too, but in a hyperbolic space there is no identification of points with opposite signs of their coordinates.

114. The Cayley metric. Cayley defined the projective distance M_0 by means of the cross-ratio

$$(\mathbf{M}, \mathbf{M}_0; \mathbf{Q}, \mathbf{P}).$$

Since parameters of these four points are proportional to the numbers

$$\tanh\Big(\frac{s}{R}\Big), 0, -1, +1,$$

we have

$$(\mathbf{M}, \mathbf{M}_0; \mathbf{Q}, \mathbf{P}) = \frac{1 + \tanh\big(\frac{s}{R}\big)}{1 - \tanh\big(\frac{s}{R}\big)} = \frac{\sinh\big(\frac{s}{R}\big) + \cosh\big(\frac{s}{R}\big)}{-\sinh\big(\frac{s}{R}\big) + \cosh\big(\frac{s}{R}\big)} = e^{\frac{2s}{R}}.$$

It follows that

$$s = \frac{R}{2} \ln(\mathbf{M}, \mathbf{M}_0; \mathbf{Q}, \mathbf{P}).$$

REMARK. In Euclidean space, take the Euclidean sphere of radius R as the absolute.

As a parameter of a point $\mathbf{M}$, we can take the Euclidean length $r = |OM|$. It is easy to derive that

$$r = \tanh\Big(\frac{s}{R}\Big)$$

and

$$\overrightarrow{MM_0} = -\frac{1}{R^2} \sinh\Big(\frac{s}{R}\Big).$$

D. THE THEORY OF GEODESIC LINES

This page is intentionally left blank

Chapter 19

Variational Problems for Geodesics

115. The field of geodesics. We consider a family of lines close to an arc AB of a curve. Suppose that α is a parameter of the family and that $\alpha = 0$ for the curve AB.

On each line of the family, the points are defined by a parameter t. Take the unit vector tangent to an arc of a curve of the family as the coordinate vector $\mathbf{I}_1$. Choose the following symbols of differentiation:

$$d = \frac{\partial}{\partial t} \cdot dt, \quad \delta = \frac{\partial}{\partial \alpha} \cdot d\alpha,$$

$$\omega^i(d) = \omega^i, \quad \omega^i(\delta) = e^i.$$

By our choice of axis $\mathbf{I}_1$, we have

$$\omega^1 = ds, \quad \omega^2 = \ldots = \omega^n = 0.$$

From structure equations (17.1), it follows that

$$\begin{cases} d\omega^1 = \omega^i \wedge \omega_i^1, \\ d\omega^i = \omega^1 \wedge \omega_1^i + \omega^k \wedge \omega_k^i, \quad i, j, k = 2, \ldots, n, \\ d\omega_1^i = \omega_1^k \wedge \omega_k^i + R_{1i,1k}\, \omega^1 \wedge \omega^k + R_{1i,jk}\, \omega^j \wedge \omega^k, \end{cases}$$

and this implies the following bilinear decompositions

$$\begin{cases} \delta\omega^1 = de^1 + e^i \omega_i^1, \\ \delta\omega^i = de^i + e^1 \omega_1^i - e_1^i \omega^1 + e^k \omega_k^i, \\ \delta\omega_1^i = de_1^i + e_1^k \omega_k^i - e_k^i \omega_1^k - R_{1i,1k}\, e^k \omega^i. \end{cases} \tag{19.1}$$

116. Stationary state of the arc length of a geodesic in the family of lines joining two points. *We prove that the length of the arc AB in the family of lines passing through the points A and B is stationary only for a geodesic.*

Since now all lines of the family pass through the points A and B, the variation $\omega^i(\delta), i = 1, \dots, n$, at these points vanishes:

$$e^i = 0 \text{ at the points } A \text{ and } B. \tag{19.2}$$

Since the differential of the arc length of a curve of the family is equal to ω^i, the length of this curve between the points A and B is expressed by the integral

$$\lambda = \int_A^B \omega^1.$$

It follows that under the passage from one curve of the family to another and for fixed ends A and B, the variation of this line is

$$\delta\lambda = \int_A^B \delta\omega^1.$$

By means of equations (19.1), we can write this variation as

$$\delta\lambda = \int_A^B (de^1 + e^i\omega_i^1) = \int_A^B de^1 + \int_A^B e^i\omega_i^1 = \int_A^B (e^2\omega_2^1 + \dots + e^n\omega_n^1), \tag{19.3}$$

since the first integral $\int_A^B de^1$ reduces to the difference of the values of e^1 at the points A and B, and conditions (19.2) show that this difference is equal to zero.

In order to have $\delta\lambda = 0$ for any variations e^i, $i = 2, \dots, n$, vanishing at the points A and B, it is necessary and sufficient that

$$\omega_1^i = 0, \quad i = 2, \dots, n. \tag{19.4}$$

Equations (19.4) show that the arc AB in question is a geodesic. In fact, applying (15.24), we find that the absolute differential of the tangent $\mathbf{I}_1$ is

$$D\mathbf{I}_1 = \omega_1^i \mathbf{I}_i,$$

and under conditions (19.4), the tangent preserves its direction. Thus, we made our definition of a geodesic to be consistent with that of Riemann (see No. **61**).

REMARK. We see that it is always possible to choose frames at points of a geodesic in such a way that

$$\omega^i = 0, \ \omega_1^i = 0, \quad i = 2, \dots, n.$$

117. The first variation of the arc length of a geodesic. *Compute the variation of the arc length of the geodesic AB, when AB is transferred to $A'B'$.*

Now we wave the requirement that all lines of the family (α) pass through the points A and B common to all of them. Instead of this requirement, we request that all lines of the family (α) are geodesics. Then condition (19.4) is extended to all curves of the family (α).

Now the variation (19.3) of the arc length of a geodesic takes the form

$$\delta\lambda = \int_A^B (de^1 + e^i\omega_i^1) = \int_A^B de^1 = (e^1)_B - (e^1)_A.$$

Here $(e^1)_A$ is the projection of the displacement AA' onto the tangent at the point A of the arc AB, and $(e^1)_B$ is the projection of the displacement BB' onto the tangent at the point B of the arc AB (Fig. 7).

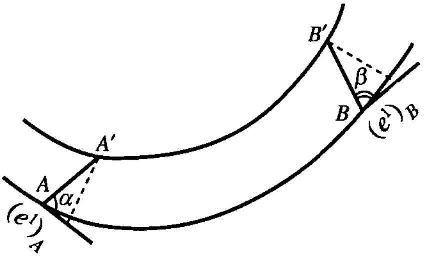

Fig. 7

Setting the angle between AA' and the tangent to AB at the point A equal to α and the angle between BB' and the tangent to AB at the point B equal to β,

$$\alpha = (\widehat{AA', AB}), \quad \beta = (\widehat{BB', AB}),$$

we obtain the desired variation of the arc length in the form

$$\delta\lambda = -AA'\cos\alpha - BB'\sin\beta.$$

COROLLARY. Consider a family of geodesics without common points. Suppose that their initial points A make up a line orthogonal to geodesics, and that all geodesics have the same length.

Then $\delta\lambda = 0$, since the length of a geodesic is not changed, and $\cos\alpha = 0$ because of orthogonality. Thus, $\cos\beta = 0$, and the trajectory BB' of the terminal point B is also orthogonal to the family of geodesics.

Thus, *if from the points of an arbitrary line AA', we draw geodesics orthogonal to the line AA' and plot the same segment on each geodesic, then the locus of these segments is also an orthogonal trajectory of the family of geodesics.*

As a result, we come to the notion of *parallel curves* AA' and BB'. The notion of *parallel surfaces* (of different dimensions) can be obtained in a similar way.

118. The second variation of the arc length of a geodesic. *Compute the second variation of the arc length of a geodesic for fixed points A and B.*

Consider a continuously differentiable family of lines (α) joining the points A and B and situated in a neighborhood of the geodesic $\alpha = 0$.

Now equation (19.3) takes the form

$$\delta\lambda = \int_A^B de^1 + \int_A^B e^i \omega_i^1. \tag{19.5}$$

We cannot assume that $\omega_1^i = 0$, since except for the line $\alpha = 0$, the lines of the family (α) are not geodesics. But for fixed ends A and B, we have

$$\omega^i(\delta) = e^i = 0 \ \text{ at the points } \ A \ \text{ and } \ B. \tag{19.6}$$

Therefore, the first integral on the right-hand side of (19.5) vanishes, and

$$\delta\lambda = \int_A^B e^i \omega_i^1.$$

For fixed ends A and B, we have

$$\delta^2\lambda = \int_A^B (\delta e^i \, \omega_i^1 + e^i \, \delta\omega_i^1).$$

But for the geodesic $\alpha = 0$, we have $\omega_i^1 = 0$, and

$$\delta^2\lambda = \int_A^B e^i \, \delta\omega_i^1. \tag{19.7}$$

Additional conditions. Suppose that at every point of the geodesic AB, a frame is obtained from an initial one at the point A by a *parallel transport.* This implies that along the geodesic AB, the forms ω_i^j vanish,

$$\omega_i^j = 0, \quad i, j = 1, \dots, n. \tag{19.8}$$

Thus, using the third equation of (19.1) and (19.8), we obtain

$$\delta\omega_1^i = de_1^i - R_{1i,1k} \, e^k \, \omega^1,$$

and as $\omega_i^1 = -\omega_1^i$, $\omega^1 = ds$, equation (19.7) takes the form

$$\delta^2\lambda = \int_A^B (-e_i \, de_1^i + R_{1i,1k} \, e^i \, e^k ds). \tag{19.9}$$

But by means of

$$\omega^1 = ds, \ \ \omega^2 = \ldots = \omega^n = 0, \ \ \delta\omega^i = 0,$$

and equations (19.8), the second equation of (19.1) gives

$$de^i = e^i_1 ds.$$

Now we can integrate by parts the first terms on the right-hand side of equation (19.9):

$$-\int_A^B e^i\, de^i_1 = -e^i e^i_1\Big|_A^B + \int_A^B (e^i_1)^2\, ds = \int_A^B (e^i_1)^2\, ds,$$

since $e_i = 0$ at the points A and B.

Hence, by equation (19.9), the second variation takes the form

$$\delta^2\lambda = \int_A^B \Big[\sum_i (e^i)^2 + R_{1i,1k}\, e^i\, e^k\Big]\, ds. \tag{19.10}$$

Consider now the Riemannian curvature in the planar direction defined by the geodesic AB and the vector $\overrightarrow{MM'}$, where M and M' correspond to the same value of t on two lines of the family. The matrix of components defining these directions is

$$\begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ e^1 & e^2 & e^3 & \ldots & e^n \end{pmatrix},$$

and the Riemannian curvature (16.13) is

$$K = -\frac{R_{1i,1k}\, e^i\, e^k}{\sum_i (e^i)^2}, \quad i \neq 1. \tag{19.11}$$

Equations (19.10) and (19.11) imply that

$$\delta^2\lambda = \int_A^B \Big[\sum_i (e^i_1)^2 - K \sum_i (e^i)^2\Big]\, ds. \tag{19.12}$$

If the curvature is negative, $K < 0$, then $\delta^2\lambda > 0$, and the arc of the geodesic has its minimum (this conclusion is not rigorous). If $K > 0$ and the arc AB is not too long, then again $\delta^2\lambda > 0$.

119. The minimum for the arc length of a geodesic (Darboux's proof). Consider a hypersurface through the point A which is normal to the arc AB and a family of parallel hypersurfaces passing through the points of the arc AB.

If the arc AB is not too long, then through every point near the arc AB there passes one and only one of these hypersurfaces. In order to define a point P lying on a hypersurface of the family passing through a point M of the line AB, as coordinates we take the parameter

$$u = AM$$

and the parameters $v, w, \ldots$ of the hypersurface (M) with line element $d\sigma^2$. Then for the line passing through the point P and joining the points A and B, we obtain the line element

$$ds^2 = du^2 + d\sigma^2.$$

Thus, we have

$$\int_0^{u_1} \sqrt{du^2 + d\sigma^2} > \int_0^{u_1} du,$$

i.e.,

$$AP > AM,$$

and the geodesic $\alpha = 0$ is the shortest line.

120. Family of geodesics of equal length intersecting the same geodesic at a constant angle. *Suppose that we have a family of geodesics of the same length a emanating from the points of the geodesic AB and making with AB the same angle φ. Let us find the difference between the length of AB and the length of the line formed by the ends of geodesics of the family, if the segment a is infinitesimally small.*

For $\alpha = 0$, taking $\delta = \dfrac{\partial}{\partial \alpha}$ and $\delta a = 1$, we have

$$e^1 = \cos\varphi, \tag{19.13}$$

and for an appropriate choice of frame, we also have

$$e^2 = \sin\varphi, \quad e^3 = \ldots = e^n = 0. \tag{19.14}$$

(Obviously, the choice of frame here is different from that in No. **118**.) For displacements along geodesics of our family (for increasing α), the variation of the arc length of AB is determined by the formula (19.5),

$$\delta\lambda = \int_A^B de^1 + \int_A^B e^i \omega_i^1, \tag{19.15}$$

or

$$\delta\lambda = (e^1)_B - (e^1)_A + \int_A^B e^i \omega_i^1. \tag{19.16}$$

The second variation is

$$\delta^2\lambda = (\delta e^1)_B - (\delta e^1)_A + \int_A^B e^i\,\delta\omega_i^1.$$

Applying formulas (19.13) and (19.14) and taking into account that for an orthonormal frame, $\omega_1^1 = 0$, we find that

$$\delta^2\lambda = (\delta e^1)_B - (\delta e^1)_A + \int_A^B e^2\,\delta\omega_2^1.$$

For the geodesic $\alpha = 0$, by means of (19.4), the form ω_1^n vanishes. Thus, since $\delta\omega_2^1 = -\delta\omega_1^2$, from the third equation of (19.1), we have

$$\delta\omega_2^1 = -de_1^2 - e_1^k\omega_k^2 + R_{12,1i}\,e^i\,\omega^1.$$

Thus,

$$\delta^2\lambda = (\delta e^1)_B - (\delta e^1)_A - \sin\varphi \int_A^B (de_1^2 + e_1^k\omega_k^2 - R_{12,1i}\,e^i\,\omega^1). \tag{19.17}$$

But for $i = 2$, the second equation of (19.1) gives

$$0 = -e_1^2\,ds,$$

since on the line $\alpha = 0$, we have $\omega^2 = 0$, $e^2 = \sin\varphi = \text{const.}$, $\omega_1^i = 0$, $\omega_k^i = 0$, and for $i > 2$, it gives

$$0 = -e_1^i\,ds + a_2^i\,\sin\varphi\,ds,$$

since for $i > 2$, we have $\delta\omega^i = 0$, $e^i = 0$, $\omega_2^i = a_2^i\,ds$. It follows that

$$e_1^2 = 0, \quad e_1^i = a_2^i\,\sin\varphi, \quad i > 2.$$

Thus, the second variation (19.17) takes the form

$$\begin{aligned}\delta^2\lambda &= (\delta e^1)_B - (\delta e^1)_A + \sin\varphi \int_A^B \Big(\sum_i (a_2^i)^2\,\sin\varphi\,ds + R_{12,12}\,\sin\varphi\,ds\Big)\\ &= (\delta e^1)_B - (\delta e^1)_A + \sin^2\varphi \int_A^B \Big(\sum_i (a_2^i)^2 - K\Big)\,ds,\end{aligned}$$

where K is the Riemannian curvature in the planar direction (AB, AA').

For our construction of a frame,

$$De^i = \delta e^i + e^k e_k^i = 0,$$

and since the line AA' is a geodesic, then $e_k^1 = 0$ and

$$\delta e^1 = 0.$$

Hence,

$$\delta^2\lambda = \sin^2\varphi \int\limits_A^B \Big(\sum_i (a_2^i)^2 - K\Big)\, ds. \tag{19.18}$$

But

$$d\mathbf{I}_2 = \omega_2^i\, \mathbf{I}_i = a_2^i\, \mathbf{I}_i\, ds,$$

$$|d\mathbf{I}_2| = ds\sqrt{\sum_i (a_2^i)^2}.$$

Thus,

$$\sum_i (a_2^i)^2 = \frac{1}{h^2},$$

where h is a reduced pitch of a helical motion moving AA' to a neighboring location.

Therefore,

$$\delta^2\lambda = \sin^2\varphi \int\limits_A^B \Big(\frac{1}{h^2} - K\Big)\, ds \approx \Big(\frac{1}{h^2} - K\Big)\, l, \tag{19.19}$$

where

$$l = \int\limits_A^B ds = AB.$$

Now decomposing the length

$$A'B' = \lambda$$

into a series with respect to powers of the infinitesimal length a plotted on geodesics of our family, we obtain

$$\lambda = l + a\,\delta\lambda + \frac{1}{2}a^2\,\delta^2\lambda + \dots\ . \tag{19.20}$$

Since for the geodesic $\alpha = 0$ the first variation $\delta\lambda = 0$, we have

$$\lambda = l + \frac{1}{2}a^2\, l\, \sin^2\varphi\,\Big(\frac{1}{h^2} - K\Big) \tag{19.21}$$

and

$$\lambda^2 = l^2 + a^2\, l^2\, \sin^2\varphi\,\Big(\frac{1}{h^2} - K\Big).$$

If we denote by S the area of the "parallelogram" $ABB'A'$ (Fig. 8), from the last equation we find that

$$\frac{l^2 - \lambda^2}{S^2} = K - \frac{1}{h^2}. \tag{19.22}$$

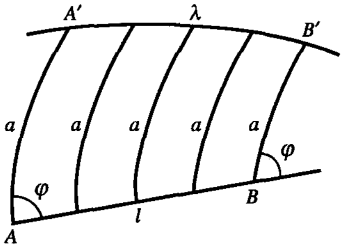

Fig. 8

1°. Suppose that $\frac{1}{h} = 0$. Then AA' and BB' are "parallel" in the sense of Levi-Civita. Levi-Civita called the "*parallelogramoid*" the figure obtained if one draws two "parallel" geodesics AA' and BB' of the same length through the ends A and B of the geodesic AB and join the points A' and B' by the arc $A'B'$ of another geodesic.

Our "parallelogram" is a similar figure for $h^{-1} = 0$. In this case,

$$K = \frac{l^2 - \lambda^2}{S^2}. \tag{19.23}$$

In Euclidean space, we have $l = \lambda$ while in elliptic space $\lambda < l$, and in hyperbolic space $\lambda > l$.

EXAMPLE. *Two-dimensional spherical space* $\left(n = 2,\ \varphi = \frac{\pi}{2}\right)$. We see that the arc of a small circle $A'B' < AB$. The quadrangle $ABB'A'$ is not a parallelogramoid.

We again obtain

$$K = \frac{1}{R^2} \text{ (R is the radius of the sphere).}$$

2°. Suppose that $\frac{1}{h} \neq 0$. If $K < 0$, then the difference

$$\lambda - l = \frac{\frac{1}{h^2} - K}{\lambda + l} S^2$$

is greater than the corresponding difference in Euclidean space. It can decrease in elliptic space, but it is possible to choose h in such a way that

$\lambda = l$ and even $\lambda > l$ (Fig. 9).

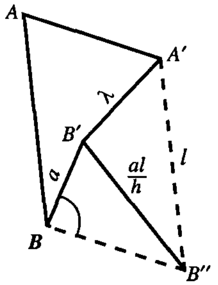
A
A′
λ
B′
a
$\frac{al}{h}$
l
B
B″

Fig. 9

EXAMPLE. $n = 3$, $K = 0$, $\varphi = \frac{\pi}{2}$. We again find that

$$\lambda^2 = l^2 + \frac{a^2\,l^2}{h^2}.$$

Chapter 20

Distribution of Geodesics near a Given Geodesic

121. Distance between neighboring geodesics and curvature of a manifold. We return to the frame chosen in No. **118**. There we transported by parallelism a frame along a geodesic. Along a geodesic, we have

$$\omega_i^j = 0. \tag{20.1}$$

Now applying formulas (19.1), we obtain

$$\left\{\begin{aligned} &\delta\omega^1 = de^1, \\ &0 = de^i - e_1^i\, ds, \\ &0 = de_1^i - R_{1i,1k}\, e^k\, ds. \end{aligned}\right. \tag{20.2}$$

The above equations give us the rule which we should follow when we go off the geodesic AB in order to find an infinitesimally close geodesic.

Differentiating the second equation of (20.2) with respect to s and applying the third equation of (20.2) to eliminate $\frac{de_1^i}{ds}$, we find that

$$\frac{d^2e^i}{ds^2} - R_1{}^i{}_{1k}e^k = 0, \quad i = 2, \ldots, n, \tag{20.3}$$

where e^i remains arbitrary.

Setting

$$R_1{}^i{}_{1k} = -A_k^i, \tag{20.4}$$

we write

$$\frac{d^2e^i}{ds^2} + A_k^i\, e^k = 0. \tag{20.5}$$

The curvature in the planar direction defined by the geodesic and the arc (e^i) is

$$K = \frac{A_{ij}\, e^i\, e^j}{\sum\limits_i (e^i)^2}. \tag{20.6}$$

Take a parallelogramoid $ABB'A'$ (Fig.10). Set

$$AB = l, \;\; AA' = a.$$

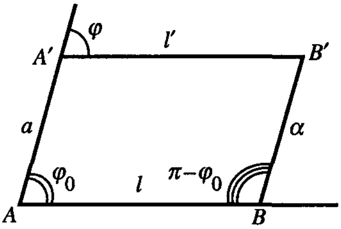

Fig. 10

We look for the upper base (a geodesic joining the points A' and B')

$$A'B' = l'.$$

Denote the angle at the point A by φ_0. Thus $\pi - \varphi_0$ is the angle at the point B.

Let λ be the length of the arc $A'B'$, the locus of points equidistant from the geodesic AB. Then from formula (19.23), we have

$$\frac{l^2 - \lambda^2}{S^2} = K. \tag{20.7}$$

The difference $\lambda - l'$ is negligible in our computations which only consider the terms up to $a^2 l$ (see (19.21)), since if α is the average distance between l and l', then $\lambda - l'$ is of order $\alpha^2\, l'$ or $\alpha^2\, l$, and α is of order $a\, l$. Thus, the difference $\lambda - l'$ is of order $a^2\, l^3$.

The computation of formula (20.7) is straightforward.

If we set $\delta = \dfrac{\partial}{\partial a}$, then from equation (19.15) we obtain

$$\begin{cases} \delta\lambda = (e^1)_B - (e^1)_A, \\ \delta^2\lambda = (\delta e^1)_B - (\delta e^1)_A. \end{cases} \tag{20.8}$$

But we have

$$\delta e^1 = e^i e^1_i,$$

and from (19.13) and (19.14), at the points A and B we have

$$e^1 = \cos\varphi_0, \quad e^2 = \sin\varphi_0, \quad e^3 = \ldots = e^n = 0. \tag{20.9}$$

Therefore, substituting these values into the second equation of (20.8), we obtain

$$\delta^2\lambda = (e_2 e_2')_B - (e_2 e_2')_A,$$

where by means of the second equation of (20.2),

$$e_i' = \frac{de^i}{ds} = e_1^i$$

and

$$\delta^2\lambda = \sin\varphi_0\,[(e_2')_B - (e_2')_A] = l\,\sin\varphi_0\,(e_2'')_C,$$

where C is an appropriately chosen point between A and B, and $e_2'' = \dfrac{d^2e}{ds^2}$.

For $\dfrac{1}{h} = 0$, by means of equation (19.19), we have

$$\delta^2\lambda = -K\,l\,\sin^2\varphi_0,$$

i.e., by means of (19.20),

$$l' = l - \frac{1}{2}\,K\,l\,\sin^2\varphi_0,$$

and we return to the formula

$$\frac{l^2 - \lambda^2}{S^2} = K.$$

122. The sum of the angles of a parallelogramoid. Denote the angle between the directions AA' and $A'B'$ by φ. From (20.9), at the point A, we have

$$-\sin\varphi_0\,\delta\varphi = \delta e^1 = \sin\varphi_0\,(e_2')_A.$$

Cancelling $\sin\varphi_0$, we obtain the equation

$$\delta\varphi = -(e_2')_A.$$

Integrating this equation under the condition $\delta a = 1$ (No. **120**), we find that

$$\varphi - \varphi_0 = -a(e_2')_A. \tag{20.10}$$

Since $AB = l$, at the point B, we have

$$e_2 = (e_2)_A + l\,(e_2')_A + \frac{1}{2}\,l^2\,(e_2')_A.$$

From this equation, using (19.14) and (20.9), we find that

$$\sin\varphi = \sin\varphi_0 + l\,(e_2')_A + \frac{1}{2}\,l^2\,(e_2')_A.$$

Since

$$\sin\varphi - \sin\varphi_0 = (\varphi - \varphi_0)\,\cos\varphi_1, \quad \varphi_1 = \varphi_0 + \theta\,(\varphi - \varphi_0), \quad 0 \le \theta \le 1,$$

we find that

$$-a\,\cos\varphi_1\,(e_2')_A = l\,(e_2')_A + \frac{1}{2}\,l^2\,(e_2')_A.$$

Assuming that $\frac{a}{l}$ tends to 0, we find from the last equation that

$$(e_2')_A = -\frac{1}{2}\,l\,(e_2')_A,$$

and hence by (20.3) and $R_1{}^2{}_{12} = -K$, we have[1]

$$(e_2')_A = \frac{1}{2}\,K\,l\sin\varphi_0. \tag{20.11}$$

Comparing equations (20.10) and (20.11), we find that

$$\varphi - \varphi_0 = -\frac{1}{2}\,K\,a\,l\sin\varphi_0,$$

i.e.,

$$\varphi = \varphi_0 - \frac{1}{2}\,K\,S,$$

where $S = a\,l\,\sin\varphi_0$ is the area of the parallelogramoid.

The sum of the interior angles A and A' is

$$\pi + \frac{1}{2}\,K\,S,$$

and the sum of the all angles of the parallelogramoid is

$$\pi + K\,S.$$

These are results of Saccheri.

123. Stability of a motion of a material system without external forces. *Suppose that a geodesic $A'B'$ is infinitesimally close at the point A' to a given geodesic AB. Are all the points of $A'B'$ also infinitesimally close to the points of AB?*

In this case, it is always possible to arrange that A_{ij} (see (20.4)) have the values

$$A_{ij} = 0, \quad i \neq j, \quad A_{ii} = K_i.$$

The basic equations

$$\frac{d^2e^i}{ds^2} + K_i\,e^i = 0 \quad (i \text{ not summed}) \tag{20.12}$$

---

[1] In the original, the sign "$-$" was erroneously written on the right-hand side of the formula (20.11). (*Translator's note.*)

are now differential equations with separable variables. If one of the components $K_i < 0$, then we have an instability. If all $K_i > 0$, then the motion is stable (Fig. 11).

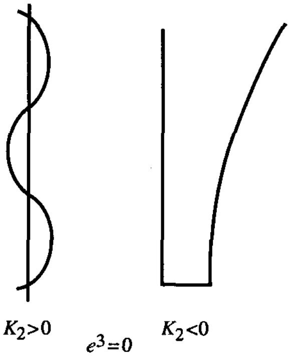

Fig. 11

Suppose that the dimension n of the manifold is three, $n = 3$. Develop a geodesic on the Euclidean space. The image is a plane curve in the preferred direction e^2 (or e^3). In other directions, the image is a space curve.

124. Investigation of the maximum and minimum for the length of a geodesic in the case A_{ij} = const. Suppose that for fixed points A and B, we have

$$\delta^2\lambda = \int_A^B \sum_i [(e_i')^2 - K_i\,(e_i)^2]\,ds.$$

Let us assume that the manifold is two-dimensional ($n = 2$) and of constant curvature. Then

$$\delta^2\lambda = \int_A^B [(e')^2 - K\,e^2]\,ds,$$

where e satisfies the equation

$$U'' + K\,U = 0; \quad U'' = \frac{d^2U}{ds^2}.$$

If the equation $U'' + K\,U = 0$ admits a nonvanishing solution between A and B, i.e., if there is a neighboring geodesic having no common points with AB, then

$\delta^2\lambda > 0$, and the minimum exists (provided that some additional conditions are satisfied). Suppose that $K > 0$. We can set

$$e = \eta U, \quad \eta = 0 \text{ at the points } A \text{ and } B.$$

Then

$$\delta^2\lambda = \int_A^B [U^2(\eta')^2 + 2UU'\eta\eta' + (U')^2\eta^2 - KU^2\eta^2]\,ds.$$

We have

$$\begin{aligned}\int_A^B 2UU'\eta\eta'\,ds &= \int_A^B UU'\,d(\eta)^2\,ds = [UU'\eta)^2]_A^B - \int_A^B \eta^2\,d(UU')\\ &= \int_A^B \eta^2\,[UU'' + (U')^2]\,ds\end{aligned}$$

and

$$\delta^2\lambda = \int_A^B [U^2(\eta')^2 - UU'\eta^2 - KU^2\eta^2]\,ds = \int_A^B U^2\,(\eta')^2\,ds > 0. \tag{20.13}$$

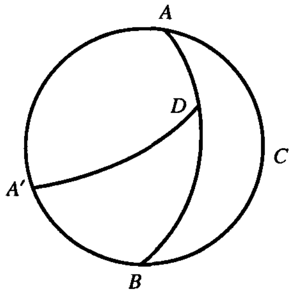

Fig. 12

This condition is satisfied if $K \leq 0$. If $K > 0$, then

$$U = a\,\sin(s - s_0)\,\sqrt{K}.$$

If the displacement of the length is less than $\dfrac{\pi}{\sqrt{K}}$, the condition $U \neq 0$ holds, and we have a minimum.

EXAMPLE OF A SPHERE. If $ACBA' > \pi$ (Fig. 12), there is no minimum for the length of a geodesic. In fact, consider a neighboring arc $ADB = ACB = \pi$, where D is an arbitrary point infinitesimally close to B. Take the arc of the great circle DA' through D. We have

$$AD + DA' < AD + DB + BA' < ACBA'.$$

It is always possible to eliminate the singular point D.

Return back to the general problem of arbitrary dimension n. A minimum is guaranteed if all $K_i \le 0$. Otherwise, we take $\widetilde{K}_i = \min K_i$. A minimum is guaranteed for a geodesic of length less than $\frac{\pi}{\sqrt{\widetilde{K}_i}}$.[2]

125. Symmetric vectors. Consider a point O of the Riemannian manifold. For a given point M, consider a geodesic OM and the arc $OM' = OM$ on the prolongation of this geodesic. The point M' is called *symmetric* to the point M with respect to the point O.

Consider a vector $\mathbf{V}$ emanating from the point M. This vector is the velocity of a moving point describing a certain line (C) for which a certain line (C') is symmetric with respect to the point O. The velocity $\mathbf{V}'$ $(V' = V)$ of a moving point describing the line (C') is the vector which is symmetric to the vector $\mathbf{V}$.

We transport by parallelism the vector $\mathbf{V}$ to a vector $\mathbf{V}_1$ with its initial point at M'.

Theorem 1. *The vector* $\mathbf{V}_1$ *is opposite to the vector* $\mathbf{V}'$ *to within second-order infinitesimals.*

Proof. We have a family of geodesics MOM' depending on the parameter α. Transport by parallelism the frames from O to every point M in such a way that

$$\omega_i^j = 0.$$

Then we have the Levi–Civita formulas (20.5),

$$\frac{d^2 e^i}{ds^2} + A_k^i\, e^k = 0.$$

For $s = 0$, all e^i vanish. Hence, all $e^{i'}$ also vanish for $s = 0$.

Set

$$e^{i'} = a^i.$$

Up to fourth-order infinitesimals, we have

$$e^i = a^i\, s - \frac{1}{6}\, s^3 A_k^i\, a^k.$$

Thus, at the point M,

$$e^i(+l) = a^i\, l - \frac{1}{6}\, l^3\, A_k^i\, a^k,$$

---

[2] Goursat, E., *Cours d'analyse; Tome 3: Intégrales infiniment voisines. Équations aux dérivées partielles du second ordre. Équations intégrales. Calcul des variations*, 3d edition, 1921. (JFM **48**, p. 491.)

and at the point M',

$$e^i(-l) = -a^i\,l + \frac{1}{6}\,l^3\,A^i_k\,a^k.$$

So, at the points M and M', e^i have the same absolute values but opposite signs.

Set also

$$a^i\,l = b^i.$$

Then

$$e^i(l) = b^i - \frac{1}{6}\,l^2\,A^i_k\,b^k,$$

$$e^i(-l) = -b^i + \frac{1}{6}\,l^2\,A^i_k\,b^k.$$

On the other hand,

$$e^1(l) = l' - l,$$

$$-e^1(-l) = l' - l,$$

where l' is the arc length of a geodesic close to OM. Therefore, by putting

$$e^1(l) = b^i,$$

we find that

$$e^1(-l) = -b^i$$

as desired.

For finite vectors $\mathbf{V}$ and $\mathbf{V}'$, all b^i are finite. This implies the theorem since the components of $\mathbf{V}'$ are the same as those of $\mathbf{V}$. ☐

126. Parallel transport by symmetry. Thus, in order to transport by parallelism a vector $\mathbf{V}$ from the point M to the point M', we construct at the point M' the vector $\mathbf{V}'$, which is symmetric to the vector $\mathbf{V}$, and take the vector $\mathbf{V}_1$ opposite to $\mathbf{V}'$. This construction is exact to within third-order infinitesimals.

The construction will be completely exact if the quantities e^i are odd functions, i.e., if the quantities A^i_k are even functions. This is the case, for example, if all A^i_k are constants (the case considered in No. **123** and No. **124**). The curvature in the direction of a plane element is preserved if one transports by parallelism this element tangentially to it. Then the symmetry preserves ds^2 (the symmetry is an *isometric transformation*).

This reduces to the fact that two moving points, that are constantly symmetric with respect to the point O, have the same velocity at all times. The latter is obvious from our previous discussion.

Conversely, suppose that the symmetry with respect to a point O is isometry. Then parallel transport preserves the curvature.

In fact, consider a bivector $(\mathbf{X}, \mathbf{Y})$ at a point M and transport by parallelism this bivector to a bivector $M'(\mathbf{X}', \mathbf{Y}')$. Let O be the midpoint of the geodesic MM', and the bivector $(\mathbf{X}_1, \mathbf{Y}_1)$ is symmetric to the bivector $(\mathbf{X}, \mathbf{Y})$. Since the symmetry is isometric, it preserves the curvature. Thus, the curvature of the area element $M'(\mathbf{X}_1, \mathbf{Y}_1)$ is equal to that of $M(\mathbf{X}, \mathbf{Y})$. But $(\mathbf{X}_1, \mathbf{Y}_1)$ coincides with $(\mathbf{X}', \mathbf{Y}')$ up to third-order infinitesimals.

Hence, the Riemannian curvature K is preserved up to third-order infinitesimals.

REMARK. *It is not necessary that a parallel transport is tangent to a plane element.*

PARTICULAR CASE: a space of constant curvature.

Corollary 2. *Take two points O and O' on the same geodesic and make consecutively two symmetries. Under such a transformation, the geodesic slides along itself. This transformation becomes isometry, if the symmetry possesses this property.*

127. Determination of three-dimensional manifolds, in which the parallel transport preserves the curvature. We put aside the case when the quadric-indicatrix Q of curvature is a sphere.

First, suppose that K_{11}, K_{22}, and K_{33} are mutually distinct. We prove that this case cannot take place. In fact, take at each point the trihedrons of principal directions. A principal direction remains principal under a parallel transport. Thus, the vector $\mathbf{I}_i$ is transported to the vector $\mathbf{I}'_i$, and as a result, $D\mathbf{I}_i = 0$. It follows that

$$\omega_i^j = 0.$$

Hence, the curvature vanishes, and this contradicts the hypothesis.

Only the case when Q is a surface of revolution remains to be considered. For example, we take

$$K_{11} = K_{22} = A, \quad K_{33} = C, \quad A \neq C.$$

We superpose the vector $\mathbf{I}_1$ with the axis of rotation of the indicatrix. The parallel transport moves the vector $\mathbf{I}_3$ to the vector $\mathbf{I}'_3$, but does not necessarily preserve $\mathbf{I}_1$ and $\mathbf{I}_2$.

We have only

$$D\mathbf{I}_3 = 0.$$

Therefore,

$$\omega_1^3 = 0, \quad \omega_2^3 = 0,$$

and also

$$A = \text{const.}, \quad C = \text{const.}$$

Conversely, suppose that the last four equations hold. Then the parallel transport preserves the curvature

$$A\,(\alpha^2 + \beta^2) + C\,\gamma^2,$$

since A, C, γ, and hence $\alpha^2 + \beta^2$, are preserved. So, we have the structure equations in the form

$$\begin{cases} d\omega^1 = \omega^2 \wedge \omega^1_2, \\ d\omega^2 = \omega^1 \wedge \omega^2_1, \\ d\omega^3 = 0, \\ d\omega^2_1 = -C\omega^1 \wedge \omega^2. \end{cases} \tag{20.14}$$

Moreover, we have

$$d\omega^3_1 = -A\omega^1 \wedge \omega^3.$$

Thus,

$$A = B = 0.$$

Do manifolds exist satisfying the system of equations (20.14)?

First of all, the third equation (20.14) implies

$$\omega^3 = dz.$$

Three other equations of (20.14) are the structure equations of a two-dimensional space of constant curvature $C \neq 0$.

Thus, all solutions are

$$ds^2 = dz^2 + d\sigma^2,$$

where $d\sigma^2 \neq 0$ is a line element of a two-dimensional space of constant curvature.

We can find all isometric transformations of this space. Such a transformation must preserve the basis vector $\mathbf{I}_3$, and hence also dz:

$$dz = dz'.$$

This implies that

$$z' = z + c,$$

where $c = \text{const}$.

In addition, there are two more families of isometric transformations $d\sigma^2$ (corresponding to the proper motions and the motions with symmetry).

Altogether, there are four families of transformations.

REMARK. There are important results for the manifolds of dimension higher than three. See these results in Chapter 11 "Symmetries and parallel transport. Symmetric spaces" of the book É. Cartan, *Les leçons sur la géométrie des espaces de Riemann*, Gauthier-Villars, Facs. 2, Paris, 1963. (MR **8**, 602; Zbl. **60**, 381.); English translation: É. Cartan, *Geometry of Riemannian spaces*, Math. Sci. Press, Brookline, MA, 1983; translated by James Glazebrook. (MR 85m:53001; Zbl. 603.53001.)

Chapter 21

Geodesic Surfaces

128. Geodesic surface at a point. Severi's method of parallel transport of a vector. If through a given point A in Riemannian manifold, we take the various geodesics (of the manifold) tangent at this point to the same plane element, we obtain a surface that is called *geodesic* at the point A. To see that a given surface is geodesic at one of its points, it suffices to consider the geodesics tangent to the surface at that point: they must lie entirely on the surface.

A geodesic surface at A has at this point vanishing principal curvatures, since the normal curvature is zero when passage is made from the point A to a point A' which is infinitesimally close to A on the surface. If we then transport by parallelism from A to A' a vector tangent to the surface at A, then it remains tangent to the surface at A' (up to infinitesimals of an order greater than the first).

It is possible to derive from this a geometric construction due to Severi[1] concerning parallel transport of a vector.

Theorem 1 (Severi). *To transport by equipollence a vector* $\mathbf{X}$ *emanating from a point* A *to a point* A'*, which is infinitesimally close to* A*, one must construct the joining geodesic* AA' *as well as the geodesic surface at* A *tangent to this geodesic and to the vector* $\mathbf{X}$ *at* A*. The required vector* $\mathbf{X}'$ *is tangent to this surface at* A' *and makes with the geodesic* AA'*, prolonged on the other side of* A'*, the same angle that the given vector* $\mathbf{X}$ *makes with the geodesic* AA'*.*

The last part of this theorem results from the fact that the direction of the geodesic AA' at the point A' is *parallel* to its direction at the point A and that parallel transport preserves the angle between two directions.

It is important to note that if we were to transport by parallelism the vector $\mathbf{X}$ along a *finite* arc of the geodesic AA' (sufficiently prolonged), the vector obtained would no longer be, in general, tangent to the considered geodesic surface, since there is no reason, *a priori*, to think that the surface, geodesic at A, is geodesic at the other points of the geodesic AA'.

---

[1] F. Severi, *Sulla curvature delle superficie a varietá*, Rend. Circ. matem. di Palermo, **42** (1917), pp. 227–229. (JFM **46**, p. 1068.)

If two surfaces, intersecting along a line (C), are geodesic at all points of (C), then they intersect at a constant angle, as the line (C) could be regarded as a curvature line on each of the surfaces.

In fact, if we develop the line (C) in Euclidean space (No. **67**), then at each point, this Euclidean space is tangent to both surfaces. The manifold of plane tangent elements of each of the surfaces forms a "curvature band" (in the sense of Blaschke). This means that after development, in a neighborhood of the developed line (C), the images of the tangent planes of both geodesic surfaces behave similarly to the tangent planes of two surfaces in Euclidean space, and the latter have a common curvature line as their intersection.

In particular, the Joachimstahl theorem on the intersection of two surfaces at a constant angle is preserved.

129. Totally geodesic surfaces. A surface which is geodesic at each of its points is called *totally geodesic.*[2] It enjoys the characteristic property that every geodesic which is tangent to this surface is completely contained within this surface.

Another characteristic property is that every geodesic that has two of its points (sufficiently close) on the surface is also completely contained within it. In fact, let A and A' be these two points. From the point A, there emanates an infinity of geodesics along the surface. They cover the whole surface within a sufficiently small neighborhood. Therefore, one of them passes through the point A', and since only one geodesic passes through two sufficiently close points in the manifold, it is precisely the geodesic in question. The converse can be proved in a similar way.

Thus, totally geodesic surfaces possess the characteristic properties of the *plane* in Euclidean space. We would reserve the name of *planes* for them. However, we will see that *the existence of such "planes" in a Riemannian manifold is an exceptional event.*

A totally geodesic surface enjoys the following three properties which are equivalent to one another:

First, *it has zero principal curvatures at each of its points.*

Second, *the unit vector normal to the surface remains normal under parallel transport along the surface.*

Third, *every vector tangent to the surface remains tangent under parallel transport along the surface.*

The second property is an immediate consequence of the first which in turn can be derived from the second. As for the second and the third, they are obviously equivalent to each other.

Let us show that the third property is *characteristic* of totally geodesic surfaces. This results in each of the first two properties being also characteristic of these surfaces.

Let us assume that every vector tangent to a given surface (S) remains tangent under parallel transport along (S), when its initial point describes an arbitrary curve on (S). This implies that the acceleration of a point describing an arbitrary curve on the surface (S) is always tangent to (S). If we express the coordinates of a point on the surface in terms of functions of two parameters α and β, it will suffice to have two second-order differential equations in α and β in order to show that the acceleration of the moving point is zero. *Therefore, there exists on the surface a geodesic* (of the manifold) *passing through an arbitrary point of this surface and having at this point an arbitrary direction tangent to the surface.* Consequently, the surface is totally geodesic. The proof by computation is not difficult.

---

[2] J. Hadamard, *Sur les éléments linéaires à plusieurs dimensions*, Darboux Bull. (2) **25** (1901), pp. 37–40. (JFM **32**, p. 614.)

130. Development of lines of a totally geodesic surface on a plane. Another important property of totally geodesic surfaces is the following: *If we develop in Euclidean space any one of the curves traced on a totally geodesic surface, a plane curve is obtained.*

In fact, since the absolute differential of the unit tangent vector is tangent to the surface, the principal normal is tangent to the surface. The unit vector of the binormal is then normal to the surface, and its absolute differential is zero. Thus, the torsion of the curve is zero. So, in the development, the curve with zero torsion becomes a plane curve.

The converse is also true. Suppose that for every curve on a surface, the torsion is zero, or what amounts to the same thing, that the velocity, the first and second order accelerations of a point moving along the surface in any manner are constantly coplanar. Suppose that the surface is defined by the equation $u^3 = 0$. It is always possible to imagine another point moving along the surface in such away that at any given instance, the quantities $\frac{du^i}{dt}$ and $\frac{d^2u^i}{dt^2}$ have the same numerical values as for the first moving point, whereas the quantities $\frac{d^3u^1}{dt^3}$ and $\frac{d^3u^i}{dt^3}$ vary from the first moving point to the second by arbitrary quantities α^1 and α^2, respectively. This implies that the arbitrary vector (α^1, α^2) tangent to the surface is coplanar with the velocity and acceleration of the first point. Consequently, this acceleration lies in the tangent plane to the surface, and all the lines on the surface are asymptotic. Thus, the surface is totally geodesic.

131. The Ricci theorem on orthogonal trajectories of totally geodesic surfaces. From this, we are going to deduce a remarkable theorem due to G. Ricci.[3]

Theorem 2. *If in a Riemannian manifold there exists a one-parameter family of planes, then their orthogonal trajectories establish an isometric point correspondence between the different planes of the family.*

Proof. In fact, let (P) be a plane in the family, and let (P') be an infinitesimally close plane. Let (C) be an arbitrary curve traced in the plane (P), and let (C') be the locus of the common points of (P') and the orthogonal trajectories taken through the different points of (C). Let us construct the representation in the Euclidean space of conjugacy (No. **67**) along the curve (C). In this representation, the curve (C) is *planar*, and the curve (C') is obtained from (C) by taking through each point of (C), an infinitesimal length normal to the plane of (C). Therefore, the first variation of any arc length of (C) is zero when going from (C) to (C'). On the other hand, since the length of (C') is preserved in this representation up to second-order infinitesimals, we see that in a Riemannian manifold, the first variation of the arc length of (C) is zero when going from the plane (P) to the infinitesimally close plane (P'). ☐

---

[3] G. Ricci, *Sulle superficie geodetiche in nun varietà qualunque e in particolare neue varietà a tre dimensioni.*, Rend. Acc. Lincei, **12** (1903), pp. 409–420. (JFM **34**, p. 658.)

An interesting consequence arises from this. In one of the planes (P_0) of the family, we take an arbitrary coordinate system u, v and denote by w a variable parameter defining the different planes of the family. Then, the line element of the manifold is of the form

$$ds^2 = d\sigma^2 + H(u, v, w)dw^2, \tag{21.1}$$

where $d\sigma^2$ is the line element of the plane (P_0):

$$ds^2 = E(u, v)du^2 + 2F(u, v)dudv + G(u, v)dv^2. \tag{21.2}$$

The line element (21.1) so obtained contains only one arbitrary function of three arguments (and functions of two arguments). It follows that Riemannian manifolds admitting a one-parameter family of planes are exceptional, since, as Riemann noticed, the most general line element ds^2 in three variables depends on three arbitrary functions of three variables (there are six arbitrary coefficients, but the possibility of arbitrary change in coordinates reduces the number of these functions to three).

Such a transformation must preserve the basis vector $\mathbf{I}_3$, and as a result, it must preserve dz:

$$dz = dz'.$$

This implies that

$$z' = z + c.$$

We have arrived at the conclusion of No. **127**.

4.2 Riemannian Curvature

Expression for Riemannian curvature: We are given two unit vectors p^i and q^i defined at a point P_0 of V_N. Consider the vector $t^i = p^i + q^i$; where and are parameters. This vector determines pencils of direction of p^i: Similarly, one and only one geodesic will pass through in the direction of q^i: These two geodesics through P_0 determine a two-dimensional geodesic surface through P_0 determined by the orientation of the unit vectors p^i and q^i. Call this surface by the name S.

The Gaussian curvature for the surface S at P_0 is called Riemannian curvature of V_N at P_0 determined by the orientation of p^i and q^i: Introduce Riemannian co-ordinates y^i with the origin at P_0: The equation of the surface S interms of y^i is

$$y^i = \quad p^i + \quad q^i \quad s; \tag{4.21}$$

where s denotes the arc length measured from P_0 to any point P along the geodesic Eq. (4.21) through P_0 in the direction of t^i: By taking $s = u^1$ and $s = u^2$, three parameters namely, ; and s can be reduced to two parameters u^1 and u^2: Here, u^1 and u^2 are taken as co-ordinates of any current point on the surface S defined by

$$y^i = u^1 p^i + u^2 q^i; \tag{4.22}$$

where p^i and q^i being fixed. Let $ds^2 = a_{ij}\, du^i du^j$; be the metric for the surface S. Let $\overset{k}{i\ j}\ {}_g$ and ${}_a$ be the Christoffel symbols of the second kind corresponding to the co-ordinates y^i and u^a, respectively, then

$$[\ ;\quad]_a = g_{km} \frac{\partial y^i}{\partial u} \frac{\partial y^j}{\partial u} \frac{\partial y^k}{\partial u} \quad \overset{m}{i\ j}\ {}_g \ : \tag{4.23}$$

Let R and R_{hijk} be the curvature tensors corresponding to the metrices $a \quad du \ du$ and $g_{ij}\, dy^i dy^j$, where ; ; ; take values 1 and 2; i and j take values from 1 to N. The number of independent components of $\overline{R}$ is $\frac{2^2(2^2 \quad 1)}{12} = 1$: We have,

$$\begin{aligned}
\overline{R}^0 &= \frac{\partial u}{\partial \bar u^1}\frac{\partial u}{\partial \bar u^2}\frac{\partial u}{\partial \bar u^1}\frac{\partial u}{\partial \bar u^2}\overline{R} = \frac{\partial u^1}{\partial \bar u^1}\frac{\partial u}{\partial \bar u^2}\frac{\partial u}{\partial \bar u^1}\frac{\partial u}{\partial \bar u^2}\overline{R}_1 + \frac{\partial u^2}{\partial \bar u^1}\frac{\partial u}{\partial \bar u^2}\frac{\partial u}{\partial \bar u^1}\frac{\partial u}{\partial \bar u^2}\overline{R}_2\\
&= \frac{\partial u^1}{\partial \bar u^1}\frac{\partial u^2}{\partial \bar u^2}\frac{\partial u}{\partial \bar u^1}\frac{\partial u}{\partial \bar u^2}\overline{R}_{12} + \frac{\partial u^2}{\partial \bar u^1}\frac{\partial u^1}{\partial \bar u^2}\frac{\partial u}{\partial \bar u^1}\frac{\partial u}{\partial \bar u^2}\overline{R}_{21}\\
&= \frac{\partial u^1}{\partial \bar u^1}\frac{\partial u^2}{\partial \bar u^2}\frac{\partial u^1}{\partial \bar u^1}\frac{\partial u^2}{\partial \bar u^2}\overline{R}_{1212} + \frac{\partial u^1}{\partial \bar u^1}\frac{\partial u^2}{\partial \bar u^2}\frac{\partial u^2}{\partial \bar u^1}\frac{\partial u^1}{\partial \bar u^2}\overline{R}_{1221}\\
&\quad + \frac{\partial u^2}{\partial \bar u^1}\frac{\partial u^1}{\partial \bar u^2}\frac{\partial u^1}{\partial \bar u^1}\frac{\partial u^2}{\partial \bar u^2}\overline{R}_{2121} + \frac{\partial u^2}{\partial \bar u^1}\frac{\partial u^1}{\partial \bar u^2}\frac{\partial u^2}{\partial \bar u^1}\frac{\partial u^1}{\partial \bar u^2}\overline{R}_{2121}
\end{aligned}$$

$$= \left[\left(\frac{\partial u^1}{\partial \bar u^1}\frac{\partial u^2}{\partial \bar u^2}\right)^2 \quad 2\frac{\partial u^1}{\partial \bar u^1}\frac{\partial u^2}{\partial \bar u^2}\frac{\partial u^2}{\partial \bar u^1}\frac{\partial u^1}{\partial \bar u^2} + \left(\frac{\partial u^2}{\partial \bar u^1}\frac{\partial u^1}{\partial \bar u^2}\right)^2 \right] \bar R_{1212}$$

$$= J^2 \bar R_{2121}; \text{ where } J = \begin{vmatrix} \frac{\partial u^1}{\partial \bar u^1} & \frac{\partial u^1}{\partial \bar u^2} \\ \frac{\partial u^2}{\partial \bar u^1} & \frac{\partial u^2}{\partial \bar u^2} \end{vmatrix} = \frac{\partial u}{\partial \bar u}: \tag{4.24}$$

Since a_{ij} is a covariant tensor, by Eq. (1.45), we get

$$\bar a \quad = \frac{\partial u}{\partial \bar u}\frac{\partial u}{\partial \bar u} a \quad \Rightarrow |\ \bar a \quad | = \left|\frac{\partial u}{\partial \bar u}\right| \left|\frac{\partial u}{\partial \bar u}\right| a \quad \Rightarrow \quad \bar a = J^2 a: \tag{4.25}$$

Dividing Eq. (4.24) by Eq. (4.25), we get

$$\frac{R'_{1212}}{\bar a} = \frac{R_{1212}}{\bar a} = \quad (\text{say}): \tag{4.26}$$

This , which is an invariant for transformation of co-ordinates, is defined as Gaussian curvature of the surface S at P_0: Hence, is the Riemannian curvature of S at P_0. Since Riemannian co-ordinate y^i with the origin at P_0 behave as geodesic co-ordinates with the pole at P_0: Thus, at P_0; ${k \atop i\ j}_g$ and $\quad_a = 0$: Therefore, we have

$$\bar R \quad = \frac{\partial}{\partial u}[\ ;\ ]_a + \frac{\partial}{\partial u}[\ ;\ ]_a \quad \text{at } P_0$$

or

$$\bar R_{1212} = \frac{\partial}{\partial \bar u^2}[21;1]_a + \frac{\partial}{\partial \bar u^1}[22;1]_a \text{ at } P_0: \tag{4.27}$$

or

$$= \frac{1}{a}\left(\frac{\partial}{\partial \bar u^2}[21;1]_a + \frac{\partial}{\partial \bar u^1}[22;1]_a\right); \text{ by Eq. } (4{:}26): \tag{4.28}$$

This is an expression for Riemannian curvature at P_0:

Formula for Riemannian curvature: Here, we have to derive a formula for Riemannian curvature interms of covariant curvature tensor of V_N. We know that

$$[\ ;\ ]_a = [ij;k]_g \frac{\partial y^i}{\partial u}\frac{\partial y^j}{\partial u}\frac{\partial y^k}{\partial u} + g_{ij}\frac{\partial^2 y^i}{\partial u u}\frac{\partial y^j}{\partial u}$$

$$= [ij;k]_g \frac{\partial y^i}{\partial u}\frac{\partial y^j}{\partial u}\frac{\partial y^k}{\partial u}; \text{ by Eq. } (4{:}22): \tag{4.29}$$

In particular, we have,

$$[21,1]_a = [ij,k]_g \frac{\partial y^i}{\partial u^2}\frac{\partial y^j}{\partial u^1}\frac{\partial y^k}{\partial u^1} = [ij,k]_g q^i p^j p^k$$

or

$$\frac{\partial}{\partial u^2}[21,1]_a = \frac{\partial}{\partial u^2}[ij,k]_g q^i p^j p^k = q^i p^j p^k q^h \frac{\partial}{\partial y^h}[ij,k]_g$$

$$= q^i p^j p^h q^k \frac{\partial}{\partial y^k}[ij,h]_g, \text{ interchanging } h \text{ and } k. \qquad (4.30)$$

By virtue of Eq. (4.29), we get

$$[22,1]_a = [ij,k]_g \frac{\partial y^i}{\partial u^2}\frac{\partial y^j}{\partial u^2}\frac{\partial y^k}{\partial u^1} = [ij,k]_g q^i q^j q^k$$

or

$$\frac{\partial}{\partial u^2}[22,1]_a = q^i q^j p^k p^h \frac{\partial}{\partial y^h}[ij,k]_g$$

$$= q^i q^k p^j p^h \frac{\partial}{\partial y^h}[ik,j]_g, \text{ interchanging } j \text{ and } k$$

$$= q^i q^k p^h p^j \frac{\partial}{\partial y^j}[ik,h]_g, \text{ interchanging } h \text{ and } j. \qquad (4.31)$$

Thus using Eqs. (4.30) and (4.31), from Eq. (4.27), the expression for $\overline{R}_{1212}$ at P_0 becomes

$$\overline{R}_{1212} = p^h q^i p^j q^k \left(-\frac{\partial}{\partial y^k}[ij,h]_g + \frac{\partial}{\partial y^j}[ik,h]_g\right) = p^h q^i p^j q^k R_{hijk}\,; \text{ at } P_0. \qquad (4.32)$$

Using the relation $a_{\alpha\beta} = g_{ij}\frac{\partial y^i}{\partial u^\alpha}\frac{\partial y^j}{\partial u^\beta}$, we get

$$a_{11} = g_{ij}\frac{\partial y^i}{\partial u^1}\frac{\partial y^j}{\partial u^1} = g_{ij}\,p^i p^j = g_{hj}\,p^h p^j.$$

$$a_{22} = g_{ij}\frac{\partial y^i}{\partial u^2}\frac{\partial y^j}{\partial u^2} = g_{ij}\,q^i q^j = g_{hk}\,q^i q^k.$$

$$a_{12} = g_{ij}\frac{\partial y^i}{\partial u^1}\frac{\partial y^j}{\partial u^2} = g_{ij}\,p^i p^j g_{hk}\,p^h q^k = a_{21}.$$

$$a = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = p^h q^i p^j q^k\,[g_{hj}\,g_{ik} - g_{ij}\,g_{kh}]. \qquad (4.33)$$

Dividing Eq. (4.32) by Eq. (4.33), we get

$$= \frac{\overline{R}_{1212}}{a} = \frac{p^h q^i p^j q^k R_{hijk}}{p^h q^i p^j q^k [g_{hj}\, g_{ik} \quad g_{ij}\, g_{kh}]} = \frac{H_{hijk}\, R_{hijk}}{G_{hijk}\, H_{hijk}}; \tag{4.34}$$

where the N^4 functions H_{hijk} and G_{hijk} are given by

$$H_{hijk} = \begin{matrix} p^h & p^i \\ q^h & q^i \end{matrix} \quad \begin{matrix} p^j & p^k \\ q^j & q^k \end{matrix} \tag{4.35}$$

and

$$G_{hijk} = g_{hj}\, g_{ik} \quad g_{ij}\, g_{kh}: \tag{4.36}$$

Equation (4.34) is the required expression for Riemannian curvature of V_N at P_0 determined by the orientation of the unit vectors p^i and q^j at P_0: If the Riemannian curvature at P_0 does not change with the orientation of a 2-flat through P_0, then the co-ordinates of P_0 is called isotropic. Thus, if $N = 2$; Eq. (4.34) reduces to

$$= \frac{R_{1212}}{g_{11} g_{22} \quad g_{12}^2} = \frac{R_{1212}}{g}$$

shows that all points of a two-dimensional Riemannian space are isotropic. Thus, at a given point in Riemannian 2-space, the curvature is determined by the g_{ij} and their derivatives, and is independent of the directions of P and Q.

Theorem 4.2.1 If at each point, the Riemannian curvature of a space is independent of the orientation chosen, then it is constant throughout the space.

Proof: Let be the Riemannian curvature of V_N at P_0 determined by the orientation of the unit vectors p^i and q^j, then it is given by

$$= \frac{\overline{R}_{1212}}{a} = \frac{p^h q^i p^j q^k R_{hijk}}{p^h q^i p^j q^k [g_{hj}\, g_{ik} \quad g_{ij}\, g_{hk}]}:$$

If is independent of orientation determined by p^i and q^i, then

$$= \frac{R_{hijk}}{(g_{hj}\, g_{ik} \quad g_{ij}\, g_{hk})} \;) \quad R_{hijk} = (\quad g_{ij}\, g_{hk} + g_{hj}\, g_{ik}): \tag{4.37}$$

Here is constant throughout the space V_N. If $N = 2$, then at any point of V_2 there is only one orientation, which is same at every point and the theorem is obvious.

So, we consider the case of V_N; where $N \quad 3$: Since g_{ij} are constants with respect to the covariant differentiation, so covariant differentiation of Eq. (4.37) yields

$$R_{hijk;l} = (\quad g_{ij}\, g_{hk} + g_{hj}\, g_{ik})_{;l}: \tag{4.38}$$

where $_{;l}$ denotes the partial derivative of with respect to xs (as is scalar). Taking the sum of Eq. (4.38) and two similar equations obtained by cyclic permutation of the suffixes $j; k$ and l

$$(g_{ij}g_{hk} + g_{hj}g_{ik})_{;l} + (g_{ik}g_{hl} + g_{il}g_{hk})_{;j} + (g_{il}g_{hj} + g_{ij}g_{hl})_{;k}$$

$$= R_{hijk;l} + R_{hikl;j} + R_{hilj;k} = 0; \text{ by Eq. } (4{:}20): \qquad (4.39)$$

Since N 3 and therefore we can give three distinct values to the suffixes $j; k$ and l. Multiplying Eq. (4.39) by g^{hj} and using the fact that $g^{hj}g_{hl} = {}^{j}_{l}$, we get

$$({}^{h}_{i}g_{hk} + Ng_{ik})_{;l} + (g_{ik}\,{}^{j}_{l} + g_{il}\,{}^{j}_{k})_{;j} + (Ng_{il} + {}^{h}_{i}g_{hl})_{;k} = 0$$

or

$$(N \quad 1)g_{ik\ ;l} + (1 \quad N)g_{il\ ;k} = 0; \text{ as } {}^{i}_{j} = 0 \text{ for } i \neq j$$

or

$$g_{ik\ ;l} \quad g_{il\ ;k} = 0; \text{ as } N \neq 1$$

or

$$g^{ik}g_{ik\ ;l} \quad g^{ik}g_{il\ ;k} = 0; \text{ multiplying } g^{ik}$$

or

$$N_{;l} \quad {}_{;k} = 0; \text{ as } g^{ik}g_{ik} = N \text{ and } g^{ik}g_{il} = {}^{k}_{j}$$

or

$$(N \quad 1)_{;l} = 0\ ) \quad {}_{;l} = \frac{@}{@x^l} = 0; \text{ as } N \neq 1:$$

This implies that the partial derivatives of with respect to xs are all zero. Hence, is a constant at P_0. But P_0 be an arbitrary point of V_N. Hence, the Riemannian curvature is constant throughout the space V_N. This is known as Schurt's theorem.

EXAMPLE 4.2.1 Calculate for the Riemannian metric

$$ds^2 = (x^1)^{\ 2}(dx^1)^2 \quad (x^1)^{\ 2}(dx^2)^2:$$

Solution: For the given metric, we have $g_{11} = (x^1)^{\ 2}$; $g_{12} = 0 = g_{21}$; $g_{22} = \ (x^1)^{\ 2}$: Since $g = \ (x^1)^{\ 4}$; so the reciprocal tensors g^{ij} are given by

$$g^{11} = (x^1)^2; \quad g^{12} = 0 = g^{21}; \quad g^{22} = \ (x^1)^2:$$

The non-vanishing Christoffel symbols of second kind are

$$\genfrac{}{}{0pt}{}{1}{1\ 1} = \frac{1}{x^1} = \genfrac{}{}{0pt}{}{1}{2\ 2}; \quad \genfrac{}{}{0pt}{}{2}{1\ 2} = \genfrac{}{}{0pt}{}{2}{2\ 1} = \frac{1}{x^1}:$$

Using Eq. (4.6), the Riemannian curvature tensors R_{ijk} can be written as

$$R^1_{212} = \frac{\partial}{\partial x^2}\begin{Bmatrix}1\\2\;1\end{Bmatrix} + \frac{\partial}{\partial x^1}\begin{Bmatrix}1\\2\;2\end{Bmatrix} + \begin{Bmatrix}1\\ \;\;1\end{Bmatrix}\begin{Bmatrix} \\2\;2\end{Bmatrix} \quad \begin{Bmatrix}1\\ \;\;2\end{Bmatrix}\begin{Bmatrix} \\2\;1\end{Bmatrix}.$$

$$= \frac{\partial}{\partial x^1}\begin{Bmatrix}1\\2\;2\end{Bmatrix} + \begin{Bmatrix}1\\1\;1\end{Bmatrix}\begin{Bmatrix}1\\2\;2\end{Bmatrix} \quad \begin{Bmatrix}1\\2\;2\end{Bmatrix}\begin{Bmatrix}2\\2\;1\end{Bmatrix}.$$

$$= \frac{1}{(x^1)^2} \quad \frac{1}{x^1} \quad \frac{1}{x^1} \qquad \frac{1}{x^1} \quad \frac{1}{x^1} = \frac{1}{(x^1)^2}.$$

$$R^2_{212} = \frac{\partial}{\partial x^2}\begin{Bmatrix}2\\2\;1\end{Bmatrix} + \frac{\partial}{\partial x^1}\begin{Bmatrix}2\\2\;2\end{Bmatrix} + \begin{Bmatrix}2\\ \;\;1\end{Bmatrix}\begin{Bmatrix} \\2\;2\end{Bmatrix} \quad \begin{Bmatrix}2\\ \;\;2\end{Bmatrix}\begin{Bmatrix} \\2\;1\end{Bmatrix} = 0.$$

Using Eq. (4.13), the only non-vanishing covariant curvature tensor is given by

$$R_{1212} = g_{1\;}R_{212} = g_{11}R^1_{212} + g_{12}R^2_{212} = g_{11}R^1_{212}.$$

Thus, the Riemannian curvature κ given by Eq. (4.34) as

$$\kappa = \frac{R_{1212}}{g} = \frac{g_{11}R^1_{212}}{g_{11}g_{22}} = \frac{R^1_{212}}{g_{22}} = \quad 1.$$

EXAMPLE 4.2.2 Evaluate the Riemannian curvature at any point (x^i) of Riemannian 3 space in the directions $P = (1; 0; 0)$ and $Q = (0; 1; 1)$ if the metric is given by

$$g_{11} = 1; \quad g_{22} = 2x^1; \quad g_{33} = 2x^2; \quad g_{ij} = 0; \text{ for } i \neq j.$$

Hence, calculate R_{ij}; R^i_j and R.

Solution: Since $g = 4x^1x^2$; so the reciprocal tensors g^{ij} are given by

$$g^{11} = 1; \quad g^{22} = \frac{1}{2x^1}; \quad g^{33} = \frac{1}{2x^2}; \quad g^{ij} = 0 = g^{ji}; i \neq j.$$

The non-vanishing Christoffel symbols of second kind are

$$\begin{Bmatrix}1\\2\;2\end{Bmatrix} = \quad 1; \quad \begin{Bmatrix}2\\1\;2\end{Bmatrix} = \begin{Bmatrix}2\\2\;1\end{Bmatrix} = \frac{1}{2x^1};$$

$$\begin{Bmatrix}2\\3\;3\end{Bmatrix} = \quad \frac{1}{2x^1}; \quad \begin{Bmatrix}3\\2\;3\end{Bmatrix} = \begin{Bmatrix}3\\3\;2\end{Bmatrix} = \frac{1}{2x^2}.$$

Since $N = 3$; only six components of the Riemannian tensor need to be considered: R_{1212}; R_{1313}; R_{2323}; R_{1213}; R_{1232}; R_{1323}: Using Eq. (4.6), the Riemannian curvature tensors R_{ijk} can be written as

$$R^1_{212} = \frac{\partial}{\partial x^2}\ {}^{1}_{2\,1} + \frac{\partial}{\partial x^1}\ {}^{1}_{2\,2} + \ {}^{1}_{\ 1}\ \ {}_{2\,2}\ \ {}^{1}_{\ 2}\ \ {}_{2\,1}$$

$$= \frac{\partial}{\partial x^1}\ {}^{1}_{2\,2}\quad {}^{1}_{2\,2}\quad {}^{2}_{2\,1} = \frac{1}{2x^1}$$

and similarly, we get

$$R^1_{313} = 0\,;\ \ R^2_{323} = \frac{1}{4x^1x^2};\ \ R^1_{213} = 0\,;\ \ R^2_{123} = 0\,;\ \ R^3_{132} = \frac{1}{4x^1x^2}:$$

Using Eq. (4.13), three non-zero terms of covariant curvature tensor R_{ijkl} of rank 4 are given by

$$R_{1212} = g_{11}R^1_{212} = \frac{1}{2x^1};\ \ R_{2323} = g_{22}R^2_{323} = \frac{1}{2x^2}$$

and

$$R_{3132} = g_{33}R^3_{132} = \frac{1}{2x^1}:$$

Since $g = (g_{ij})$ is diagonal, all the non-zero G_{ijkl} will be derivable from the relation,

$$G_{ijij} = g_{ii}\, g_{jj}\ ;\ \ i < j \ \text{ and no summation:}$$

Thus, the non-vanishing G_{ijkl} are given by

$$G_{1212} = g_{11}g_{22} = 2\,x^1;\ \ G_{1313} = g_{11}g_{33} = 2\,x^2;\ \ G_{2323} = g_{22}g_{33} = 4\,x^1x^2:$$

For $\begin{matrix} P \\ Q \end{matrix} = \begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{matrix}$, the coefficients H_{hijk} are given by Eq. (4.35). Therefore,

$$H_{1212} = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}^{\,2} = 1;\qquad H_{2323} = \begin{matrix} 0 & 0 \\ 1 & 1 \end{matrix}^{\,2} = 0;$$

$$H_{3132} = \begin{matrix} 0 & 1 \\ 1 & 0 \end{matrix}\ \begin{matrix} 0 & 0 \\ 1 & 1 \end{matrix} = 0;\quad H_{1313} = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}^{\,2} = 1:$$

Thus, the Riemannian curvature given by Eq. (4.34) as

$$= \frac{R_{1212}H_{1212} + R_{2323}H_{2323} + R_{3132}H_{3132} + R_{1332}H_{1332}}{G_{1212}H_{1212} + G_{2323}H_{2323} + G_{3132}H_{3132} + G_{1332}H_{1332}}$$

$$= \frac{1{=}2x^1\quad 1 + \ \ 1{=}2x^2\quad 0 + 2\ \ 1{=}2x^1\quad 0}{(2x^1)\quad 1 + (2\,x^2)\quad 0 + 2\,(4x^1x^2)\quad 0} = \frac{1}{4x^1(x^1 + x^2)}:$$

Using formula (4.10), the Ricci tensors of first kind are given by

$$R_{ij} = R_{ij} = R^1_{ij\,1} + R^2_{ij\,2} + R^3_{ij\,3}$$
$$= g^{11}R_{1ij\,1} + g^{22}R_{2ij\,2} + g^{33}R_{3ij\,3}:$$

Therefore, the non-zero Ricci tensors are given by

$$R_{11} = g^{22}R_{2112} = \frac{1}{4(x^1)^2};\ R_{22} = g^{11}R_{1221} + g^{33}R_{3223} = \frac{1}{2x^1} \quad \frac{1}{4(x^1)^2}$$
$$R_{33} = g^{22}R_{2332} = \frac{1}{4x^1x^2};\ R_{12} = g^{33}R_{3123} = \frac{1}{4x^1x^2} = g^{33}R_{3213} = R_{21}:$$

Using formula (4.11), the Ricci tensors of first kind are given by

$$R^i_j = g^{i}\ R_{\ j} = g^{i1}R_{1j} + g^{i2}R_{2j} + g^{i3}R_{3j}:$$

So

$$R^1_1 = g^{11}R_{11} + g^{12}R_{21} + g^{13}R_{31} = \frac{1}{4(x^1)^2}:$$
$$R^2_2 = g^{12}R_{12} + g^{22}R_{22} + g^{23}R_{32} = \frac{1}{2x^1} \quad \frac{1}{2x^1} \quad \frac{1}{4(x^1)^2} \ :$$
$$R^3_3 = g^{31}R_{13} + g^{32}R_{23} + g^{33}R_{33} = \frac{1}{2x^2} \quad \frac{1}{4x^1x^2} \ :$$

Therefore, the total curvature R is given by Eq. (4.12) as

$$R = R^i_j = R^1_1 + R^2_2 + R^3_3$$
$$= \quad \frac{1}{4(x^1)^2} \quad \frac{1}{4(x^1)^2} \quad \frac{1}{8x^1(x^2)^2} \quad \frac{1}{8x^1(x^2)^2} = \frac{x^1 + 2(x^2)^2}{4(x^1x^2)^2}:$$

Deduction 4.2.1 Isotropic points: If the Riemannian curvature at x does not change with the orientation of a 2-flat through x, then x is called isotropic. All points of a two-dimensional Riemannian space are isotropic.

EXAMPLE 4.2.3 Find the isotropic points in the Riemannian space R^3 with metric

$$g_{11} = 1;\ g_{22} = (x^1)^2 + 1 = g_{33};\ g_{ij} = 0;\ \text{for } i \neq j$$

and calculate the Riemannian curvature at that points.

Solution: Since $g = [(x^1)^2 + 1]^2$; so the reciprocal tensors g^{ij} are given by

$$g^{11} = 1;\quad g^{ij} = 0 = g^{ji};\ g^{22} = \frac{1}{(x^1)^2 + 1} = g^{33}:$$

The non-vanishing Christoffel symbols of second kind are,

$$\left\{{1 \atop 2\;2}\right\} = x^1 = \left\{{1 \atop 3\;3}\right\};\quad \left\{{2 \atop 2\;1}\right\} = \frac{x^1}{(x^1)^2+1} = \left\{{3 \atop 3\;1}\right\}:$$

Using Eq. (4.6), the Riemannian curvature tensors R_{ijk} can be written as

$$\begin{aligned}
R^1_{212} &= \frac{\partial}{\partial x^2}\left\{{1 \atop 2\;1}\right\} + \frac{\partial}{\partial x^1}\left\{{1 \atop 2\;2}\right\} + \left\{{1 \atop \;\;1}\right\}\left\{{\atop 2\;2}\right\} \quad \left\{{1 \atop 2}\right\}\left\{{\atop 2\;1}\right\}\\
&= \frac{\partial}{\partial x^1}\left\{{1 \atop 2\;2}\right\} \quad \left\{{1 \atop 2\;2}\right\}\left\{{2 \atop 2\;1}\right\} = 1 \quad \frac{x^1}{(x^1)^2+1}(\;x^1) = \frac{1}{(x^1)^2+1}:\\
R^1_{313} &= \frac{\partial}{\partial x^3}\left\{{1 \atop 3\;1}\right\} + \frac{\partial}{\partial x^1}\left\{{1 \atop 3\;3}\right\} + \left\{{1 \atop \;\;1}\right\}\left\{{\atop 3\;3}\right\} \quad \left\{{1 \atop 3}\right\}\left\{{\atop 3\;1}\right\}\\
&= \frac{\partial}{\partial x^1}\left\{{1 \atop 3\;3}\right\} \quad \left\{{1 \atop 3\;3}\right\}\left\{{3 \atop 3\;1}\right\} = 1 \quad \frac{x^1}{(x^1)^2+1}(\;x^1) = \frac{1}{(x^1)^2+1}:\\
R^2_{323} &= \left\{{1 \atop 3\;3}\right\}\left\{{2 \atop 1\;2}\right\} = \frac{(x^1)^2}{(x^1)^2+1}:\\
R^1_{213} &= R^2_{123} = R^3_{132} = 0:
\end{aligned}$$

Using Eq. (4.13), the non-vanishing covariant curvature tensor R_{hijk} are given by

$$R_{1212} = g_{11}R^1_{212} = \left((x^1)^2+1\right)^{1} = R_{1313} = g_{11}R^1_{313};\, R_{2323} = g_{22}R^2_{323} = (x^1)^2:$$

Using Eq. (4.36), the non-vanishing tensors G_{hijk} are given by

$$G_{1212} = g_{11}g_{22} = (\,x^1)^2+1 = G_{1313} = g_{11}g_{33};\, G_{2323} = g_{22}g_{33} = \left((x^1)^2+1\right)^2:$$

If the curvature is to be independent of H_{hijk}, defined in Eq. (4.35) (which vary with the direction of the 2 flat), then $(\;x^1)^2 = 1$; i.e. $x^1 = \;1$: Therefore, the isotropic points compose two surfaces, on which the curvature has the value given by

$$\begin{aligned}
&= \frac{\left((x^1)^2+1\right)^{1}H_{1212} \quad \left((x^1)^2+1\right)^{1}H_{1313} \quad (x^1)^2H_{2323}}{[(x^1)^2+1]\,H_{1212}+[(\,x^1)^2+1]\,H_{1313}+[(\,x^1)^2+1]^2\,H_{2323}}\\
&= \left((x^1)^2+1\right)^{2}\frac{H_{1212}+H_{1313}+(\,x^1)^2\left((x^1)^2+1\right)H_{2323}}{H_{1212}+H_{1313}+[(\,x^1)^2+1]\,H_{2323}}\\
&= [1+1]^{2}\;1 = \frac{1}{4}:
\end{aligned}$$

EXAMPLE 4.2.4 Show that every point of R^3 is isotropic for the metric

$$ds^2 = (x^1)^{2}(dx^1)^2 + (x^1)^{2}(dx^2)^2 + (x^1)^{2}(dx^3)^2:$$

Solution: For the given metric, we have $g_{11} = (x^1)^{2} = g_{22} = g_{33}$; $g_{ij} = 0 = g_{ji}$; for $i \neq j$: Since $g = (x^1)^{6}$; so the reciprocal tensors g^{ij} are given by

$$g^{11} = (x^1)^2 = g^{22} = g^{33}; \quad g^{ij} = 0 = g^{ji}; \quad \text{for } i \neq j:$$

The non-vanishing Christoffel symbols of second kind are

$$\begin{matrix}1\\1\;1\end{matrix} = \frac{1}{x^1} = \begin{matrix}2\\1\;2\end{matrix} = \begin{matrix}3\\1\;3\end{matrix}; \quad \begin{matrix}1\\2\;2\end{matrix} = \frac{1}{x^1} = \begin{matrix}1\\3\;3\end{matrix}:$$

Using Eq. (4.6), the Riemannian curvature tensors R_{ijk} can be written as

$$R^1_{212} = \frac{\partial}{\partial x^2}\begin{matrix}1\\2\;1\end{matrix} + \frac{\partial}{\partial x^1}\begin{matrix}1\\2\;2\end{matrix} + \begin{matrix}1\\1\end{matrix}\begin{matrix}\\2\;2\end{matrix} \qquad \begin{matrix}1\\2\end{matrix}\begin{matrix}\\2\;1\end{matrix}:$$

$$= \frac{\partial}{\partial x^1}\begin{matrix}1\\2\;2\end{matrix} \quad \begin{matrix}1\\2\;2\end{matrix}\begin{matrix}1\\1\;1\end{matrix} \quad \begin{matrix}2\\2\;1\end{matrix}\begin{matrix}1\\2\;2\end{matrix} = \frac{1}{(x^1)^2}:$$

$$R^1_{313} = \frac{\partial}{\partial x^3}\begin{matrix}1\\3\;1\end{matrix} + \frac{\partial}{\partial x^1}\begin{matrix}1\\3\;3\end{matrix} + \begin{matrix}1\\1\end{matrix}\begin{matrix}\\3\;3\end{matrix} \qquad \begin{matrix}1\\3\end{matrix}\begin{matrix}\\3\;1\end{matrix}:$$

$$= \frac{\partial}{\partial x^1}\begin{matrix}1\\3\;3\end{matrix} + \begin{matrix}1\\3\;3\end{matrix}\begin{matrix}1\\1\;1\end{matrix} \quad \begin{matrix}3\\3\;1\end{matrix}\begin{matrix}1\\3\;3\end{matrix} = \frac{1}{(x^1)^2}:$$

$$R^2_{323} = \begin{matrix}1\\3\;3\end{matrix}\begin{matrix}2\\1\;2\end{matrix} = \frac{1}{(x^1)^2}; \quad R^1_{213} = R^2_{123} = R^3_{132} = 0:$$

Using Eq. (4.13), the non-vanishing covariant curvature tensors R_{hijk} are given by

$$R_{1212} = R_{1313} = R_{2323} = (x^1)^{4}; \quad G_{1212} = G_{1313} = G_{2323} = (x^1)^{4}:$$

The Riemannian curvature is given by

$$= \frac{R_{1212}H_{1212} + R_{1313}H_{1313} + R_{2323}H_{2323}}{G_{1212}H_{1212} + G_{1313}H_{1313} + G_{2323}H_{2323}}$$

$$= \frac{(x^1)^{4}}{[(x^1)^{4}]}\,\frac{H_{1212} + H_{1313} + H_{2323}}{H_{1212} + H_{1313} + H_{2323}} = 1:$$

Thus, we see that this Riemannian space is more than just isotropic; it is a space of constant curvature.

4.2.1 Space of Constant Curvature

If the Riemannian curvature is constant in a space V_N, then that space V_N is said to be of constant curvature.

Let us consider an Euclidean space S_N of N dimensions. We have to show that $R_{hijk} = 0$. In the case of S_N, the metric tensor g_{ij} is either 1 or 0. Hence,

$$[ij;k] = 0 = \begin{matrix} k \\ i\ j \end{matrix}. \tag{4.40}$$

Using Eq. (4.17), the covariant curvature tensors R_{hijk} are given by

$$R_{hijk} = \begin{matrix} \frac{\partial}{\partial x^j} & \frac{\partial}{\partial x^k} \\ [ij;h] & [ik;h] \end{matrix} + \begin{matrix} i\ j & i\ k \\ [hj;\ ] & [hk;\ ] \end{matrix} = 0, \quad \text{by Eq. (4.40).}$$

Conversely, let $R_{hijk} = 0$ in a space V_N, we have to show that the space V_N is an Euclidean space S_N. The Riemannian curvature at any point P of V_N is given by

$$= \frac{p^h q^i p^j q^k R_{hijk}}{p^h q^i p^j q^k [g_{hj} g_{ik} \quad g_{ij} g_{hk}]} = 0, \text{ as } R_{hijk} = 0.$$

Consequently, geodesics through P are straight lines. But P is an arbitrary point of V_N. Hence, geodesics through every point of V_N are straight lines. This will happen only if V_N is S_N. Therefore, a necessary and sufficient condition that a space V_N be an Euclidean space of N dimensions is that the curvature tensor vanishes.

4.2.2 Zero Curvature

A fundamental question has run unanswered through preceding sections: How can one tell whether a given metrisation of R^N is an Euclidean or not?

Let us consider a specified co-ordinate system (x^i) in which a Riemannian metric $g = (g_{ij})$. Now, consider a co-ordinate system in which $\bar{g}_{ij} = {}_{ij}$, which is a rectangular system. Suppose that a rectangular system $(\bar{x}^i)$ does exist. Then $e = 0$, since all Christoffel symbols vanish in $(\bar{x}^i)$. But the Riemannian curvature is an invariant, so that $= 0$ in the original co-ordinate (x^i) as well. Moreover,

$$g_{ij} u^i u^j = \bar{u}^i \bar{u}^j \quad 0.$$

Thus, we see that a Riemannian metric (g_{ij}) is the Euclidean metric if the Riemannian curvature is zero at all points and the metric is positive definite. To prove the converse portion, we set up a system of first order partial differential equations for N

rectangular co-ordinates $\bar{x}^i$ as functions of the given co-ordinates x^j $(j = 1, 2, \ldots, N)$. The system that immediately comes to mind is $G = J^T J$; or,

$$\frac{\partial \bar{x}^k}{\partial x^i}\frac{\partial \bar{x}^k}{\partial x^j} = g_{ij}\left(x^1, x^2, \ldots, x^N\right). \tag{4.41}$$

But the partial differential equation Eq. (4.41) is generally intractable because of its nonlinearity. Instead, we select the linear system that results when barred and unbarred co-ordinates are interchanged in Eq. (3.10) and then the $\overline{\left\{ {i \atop j\;k} \right\}}$ are

$$\frac{\partial^2 \bar{x}^k}{\partial x^i \partial x^j} = \left\{ {r \atop i\;j} \right\} \frac{\partial \bar{x}^k}{\partial x^r}. \tag{4.42}$$

Setting $\omega \equiv \bar{x}^k$ and $u_i \equiv \frac{\partial \bar{x}^k}{\partial x^i}$ yields the desired first order system

$$\frac{\partial \omega}{\partial x^i} = u_i \text{ and } \frac{\partial u_i}{\partial x^j} = \left\{ {r \atop i\;j} \right\} u_r. \tag{4.43}$$

A Riemannian metric $g = (g_{ij})$, specified in a co-ordinate system (x^i), is the Euclidean metric, if, under some permissible co-ordinate transformation Eq. (1.2), $\bar{g} = (\delta_{ij})$.

EXAMPLE 4.2.5 Show that R^3 under the following metric is Euclidean.

$$ds^2 = [(x^1)^2 + (x^2)^2](dx^1)^2 + [(x^1)^2 + (x^2)^2](dx^2)^2 + (dx^3)^2.$$

Solution: For the given metric, we have $g_{11} = (x^1)^2 + (x^2)^2 = g_{22}$, $g_{33} = 1$; $g_{ij} = 0 = g_{ji}$; $j \neq i$. Since g_{33} = constant, and g_{11} and g_{22} independent of x^3, so $\left\{ {i \atop j\;k} \right\} = 0$, whenever i, j or k equals to 3. Consequently, of the six independent components of the Riemannian tensor, R_{1212} is possibly non-zero. Since $g = [(x^1)^2 + (x^2)^2]^2$, so the reciprocal tensors g^{ij} are given by

$$g^{11} = \frac{1}{(x^1)^2 + (x^2)^2};\quad g^{22} = \frac{1}{(x^1)^2 + (x^2)^2};\quad g^{33} = 1.$$

The non-vanishing Christoffel symbols of second kind are

$$\left\{ {1 \atop 1\;1} \right\} = \frac{x^1}{(x^1)^2 + (x^2)^2} = \left\{ {2 \atop 1\;2} \right\};\quad \left\{ {1 \atop 2\;2} \right\} = -\frac{x^1}{(x^1)^2 + (x^2)^2}$$

$$\left\{ {1 \atop 1\;2} \right\} = \frac{x^2}{(x^1)^2 + (x^2)^2} = \left\{ {2 \atop 2\;2} \right\};\quad \left\{ {2 \atop 1\;1} \right\} = -\frac{x^2}{(x^1)^2 + (x^2)^2}.$$

Using Eq. (4.6), the Riemannian curvature tensors R_{ijk} can be written as

$$R^1_{212} = \frac{\partial}{\partial x^2}\ \substack{1\\2\ 1} + \frac{\partial}{\partial x^1}\ \substack{1\\2\ 2} + \substack{1\\1}\ \substack{\\2\ 2}\qquad \substack{1\\2}\ \substack{\\2\ 1}\ :$$

$$= \frac{\sqrt{g} + x^1(2x^1)}{g}\quad \frac{\sqrt{g}\quad x^2(2x^2)}{g} + \frac{x^1}{\sqrt{g}}\ \frac{x^1}{\sqrt{g}} + \frac{x^2}{\sqrt{g}}\ \frac{x^2}{\sqrt{g}}\quad \frac{x^2}{\sqrt{g}}\ \frac{x^2}{\sqrt{g}}\quad \frac{x^1}{\sqrt{g}}\ \frac{x^1}{\sqrt{g}}$$

$$= 0:$$

Consequently, $R_{1212} = 0 =$: As the metric is clearly positive definite, so the space is Euclidean.

EXAMPLE 4.2.6 Consider the two-dimensional metric $g_{11} = 0;\ g_{12} = 0 = g_{21};\ g_{22} = (x^2)^2$: Find the general solution in $(\bar{x}^i)$ co-ordinate system.

Solution: Since $g = (x^2)^2$; so the reciprocal tensors g^{ij} are given by

$$g^{11} = 1;\quad g^{12} = 0 = g^{21};\quad g^{22} = (x^2)^2:$$

The only non-vanishing Christoffel symbol of second kind is $\substack{2\\2\ 2} = \frac{1}{x^2}$; so the metric is obviously positive definite. Using Eq. (4.6), the Riemannian curvature tensors R_{ijk} can be written as

$$R^1_{212} = \frac{\partial}{\partial x^2}\ \substack{1\\2\ 1} + \frac{\partial}{\partial x^1}\ \substack{1\\2\ 2} + \substack{1\\1}\ \substack{\\2\ 2}\qquad \substack{1\\2}\ \substack{\\2\ 1} = 0:$$

$$R^2_{212} = \frac{\partial}{\partial x^2}\ \substack{2\\2\ 1} + \frac{\partial}{\partial x^1}\ \substack{2\\2\ 2} + \substack{2\\1}\ \substack{\\2\ 2}\qquad \substack{2\\2}\ \substack{\\2\ 1} = 0:$$

Using Eq. (4.13), the covariant curvature tensor is given by

$$R_{1212} = g_1\ R_{212} = g_{11}R^1_{212} + g_{12}R^2_{212} = 0:$$

Thus, the Riemannian curvature is $= \frac{R_{1212}}{g} = 0$: Let us introduce,

$$f_1 = \frac{\partial \bar{x}^1}{\partial x^1};\ f_2 = \frac{\partial \bar{x}^1}{\partial x^2};\ f_3 = \frac{\partial \bar{x}^2}{\partial x^1};\ f_4 = \frac{\partial \bar{x}^2}{\partial x^2};$$

then the system of Eq. (4.41) becomes,

$$f_1^2 + f_3^2 = 1;\quad f_1 f_2 + f_3 f_4 = 0;\quad f_2^2 + f_4^2 = (x^2)^2$$

$$\Rightarrow\quad f_1 = f_1;\quad f_2 = x^2\sqrt{1\quad f_1^2};\quad f_3 = \sqrt{1\quad f_1^2};\quad f_4 = x^2 f_1:$$

Therefore, the system of equations becomes two simple first order systems in $\bar{x}^1$ alone and $\bar{x}^2$ alone

$$\mathrm{I}: \ \frac{\partial \bar{x}^1}{\partial x^1} = f_1, \qquad \frac{\partial \bar{x}^1}{\partial x^2} = x^2\sqrt{1 \quad f_1^2}$$

$$\mathrm{II}: \ \frac{\partial \bar{x}^2}{\partial x^1} = \quad \sqrt{1 \quad f_1^2}, \qquad \frac{\partial \bar{x}^2}{\partial x^2} = x^2 f_1.$$

The unknown function f_1 is determined by the requirements that the two conditions I and II be compatible

$$\frac{\partial f_1}{\partial x^2} = \frac{\partial}{\partial x^1} \quad x^2\sqrt{1 \quad f_1^2} \quad \text{and} \quad \frac{\partial}{\partial x^2} \quad \sqrt{1 \quad f_1^2} \quad = \frac{\partial}{\partial x^1} \quad x^2 f_1 \quad .$$

The only function satisfying these two compatibility condition is

$$f_1 = \text{ constant } = \cos$$

and I and II immediately integrate to give

$$\bar{x}^1 = x^1\cos \quad + \frac{1}{2}(x^2)^2\sin \quad + c$$

$$\bar{x}^2 = \quad x^1\sin \quad + \frac{1}{2}(x^2)^2\cos \quad + d,$$

where c and d are constants and we use, of course, free to set $= c = d = 0$. From Eq. (4.43) we have to solve,

$$\frac{\partial \omega}{\partial x^1} = u_1, \quad \frac{\partial \omega}{\partial x^2} = u_2$$

$$\frac{\partial u_1}{\partial x^1} = 0, \ \frac{\partial u_1}{\partial x^2} = 0, \quad \frac{\partial u_2}{\partial x^1} = 0, \ \frac{\partial u_2}{\partial x^2} = u_2 \quad {}^{2}_{2\,2} \quad = \frac{u_2}{x^2}.$$

$$\omega = a_1x^1 + a_2(x^2)^2 + a_3, \quad a_1, a_2, a_3 = \text{ constant}$$

$$\bar{x}^k = a^k_1x^1 + a^k_2(x^2)^2 + a^k_3, \quad a^k_1 = \text{ constant},$$

which is similar to the solution set $\bar{x}^1$ and $\bar{x}^2$ as above.

Deduction 4.2.2 The quasilinear first-order system,

$$\frac{\partial u}{\partial x^j} = F_j\,(u_0, u_1, \ldots, u_m;\ x^1, x^2, \ldots, x^n), \qquad = 0, 1, \ldots, m; j = 1, 2, \ldots, n$$

where the functions F_j are of the differentiability class C^1; has a nontrivial solution for the u , bounded over some region of R^n, if and only if

$$\frac{\partial F_j}{\partial u}F_k + \frac{\partial F_j}{\partial x^k} = \frac{\partial F_k}{\partial u}F_j + \frac{\partial F_k}{\partial x^j}, \qquad = 0, 1, \ldots, m; \quad 1 \quad j < k \quad n;$$

where summations run from 0 to m. The conditions $R_{hijk} = 0$ are the sufficient condition for compatibility.

4.2.3 Einstein Tensor

The covariant tensor

$$_{ij} = R_{ij} \quad \frac{1}{2} R g_{ij} \qquad (4.44)$$

is called Einstein tensor. The mixed tensor

$$^{i}_{j} = R^{i}_{j} \quad \frac{1}{2}\,^{i}_{j} R; \quad g^{i} \; R_{j} = R^{i}_{j} \text{ and } g^{ij} R_{ij} = R \qquad (4.45)$$

is called the Einstein tensor. This tensor is widely used in the general theory of relativity.

EXAMPLE 4.2.7 Prove that for any Riemannian metric, the divergence of the Einstein tensor is zero at all points.

Solution: The Einstein tensor $^{i}_{j}$ is given by

$$^{i}_{j} = R^{i}_{j} \quad \frac{1}{2} \quad ^{i}_{j} \; R;$$

where $g^{i} \; R_{j} = R^{i}_{j}$ and $g^{ij} R_{ij} = R$: Covariant differentiation of the Einstein tensor $^{i}_{j}$ with respect to x^{k} is

$$\begin{aligned} ^{i}_{j;k} &= R^{i}_{j;k} \quad \frac{1}{2}\,^{i}_{j;k} R \quad \frac{1}{2}\,^{i}_{j} R_{;k} \\ &= R^{i}_{j;k} \quad \frac{1}{2}\,^{i}_{j} R_{;k}; \text{ as } ^{i}_{j;k} = 0: \end{aligned}$$

Hence, the divergence of the Einstein tensor $^{i}_{j}$ is given by

$$\begin{aligned} ^{i}_{j;i} &= R^{i}_{j;i} \quad \frac{1}{2}\,^{i}_{j} R_{;i} \\ &= R^{i}_{j;i} \quad \frac{1}{2} R_{;j} = \frac{1}{2} R_{;j} \quad \frac{1}{2} R_{;j} = 0: \end{aligned}$$

Thus, Einstein tensor is divergence free. This equation plays an important role in the theory of relativity.

4.2.4 Flat Riemannian Spaces

We know, covariant differentiation of all vectors is commutative if and only if the Riemann{Christoffel curvature tensor R_{ijk} is identically zero. The covariant curvature tensors of fourth order R_{hijk} are zero, then all the components of the tensor R_{hijk} are zero and vice versa. A Riemannian space whose curvature tensor is identically zero is called a flat space.

If in a neighbourhood N of a point O of V_N, there exists a co-ordinate system with respect to which the metric tensor has constant values, then V_N is said to be locally flat at O. The Riemannian space V_N is said to be locally flat, if the metric tensor g_{ij} has constant values throughout V_N.

A space is Euclidean (or flat) if it is possible to find a Cartesian co-ordinate system everywhere in it; if this is not possible, the space is non-Euclidean (or curved).

Theorem 4.2.2 The necessary and sufficient condition that a space V_N be locally flat (or flat) in the neighbourhood of O is that the curvature tensor be identically zero.

Proof: First let V_N be locally flat in the neighbourhood of O, so that g_{ij} = constant for all i and j. Thus,

$$\frac{\partial g_{ij}}{\partial x^k} = 0 \Rightarrow \left\{ {k \atop i\;j} \right\} = 0;$$

in the neighbourhood of O. The Riemann–Christoffel symbol R^{α}_{ijk} is given by

$$\begin{aligned} R^{\alpha}_{ijk} &= -\frac{\partial}{\partial x^k}\left\{ {\alpha \atop i\;j} \right\} + \frac{\partial}{\partial x^j}\left\{ {\alpha \atop i\;k} \right\} - \left\{ {\beta \atop i\;j} \right\}\left\{ {\alpha \atop \beta\;k} \right\} + \left\{ {\beta \atop i\;k} \right\}\left\{ {\alpha \atop \beta\;j} \right\} \\ &= 0; \text{ in the neighbourhood of } O. \end{aligned}$$

Conversely, let $R^{\alpha}_{ijk} = 0$ in the neighbourhood of O. To prove that V_N is locally flat, it is enough to prove that g_{ij} = constant in the neighbourhood of O. Given co-ordinate system x^i, let us choose co-ordinate system $\bar{x}^i$ such that

$$\left\{ {\alpha \atop \beta\;\gamma} \right\} = \frac{\partial^2 \bar{x}^l}{\partial x^\beta \partial x^\gamma}\frac{\partial x^\alpha}{\partial \bar{x}^l}. \tag{4.46}$$

The tensor law of transformation for the Christoffel symbol is

$$\overline{\left\{ {k \atop i\;j} \right\}} = \left\{ {\gamma \atop \alpha\;\beta} \right\}\frac{\partial x^\alpha}{\partial \bar{x}^i}\frac{\partial x^\beta}{\partial \bar{x}^j}\frac{\partial \bar{x}^k}{\partial x^\gamma} + \frac{\partial^2 x^l}{\partial \bar{x}^i \partial \bar{x}^j}\frac{\partial \bar{x}^k}{\partial x^l}$$

or

$$\left\{ {k \atop i\;j} \right\} = \overline{\left\{ {\gamma \atop \alpha\;\beta} \right\}}\frac{\partial \bar{x}^\alpha}{\partial x^i}\frac{\partial \bar{x}^\beta}{\partial x^j}\frac{\partial x^k}{\partial \bar{x}^\gamma} + \frac{\partial^2 \bar{x}^l}{\partial x^i \partial x^j}\frac{\partial x^k}{\partial \bar{x}^l}$$

or

$$\frac{\partial^2 \bar{x}^l}{\partial x^i \partial x^j}\frac{\partial x^k}{\partial \bar{x}^l} = \left\{ {k \atop i\;j} \right\} - \overline{\left\{ {\gamma \atop \alpha\;\beta} \right\}}\frac{\partial \bar{x}^\alpha}{\partial x^i}\frac{\partial \bar{x}^\beta}{\partial x^j}\frac{\partial x^k}{\partial \bar{x}^\gamma}$$

or

$$\frac{\partial \bar{x}}{\partial x^i}\frac{\partial \bar{x}}{\partial x^j}\frac{\partial x^k}{\partial \bar{x}} = 0$$

or

$$_{l\ m}{}^{k} = 0;\ \text{multiplying}\ \frac{\partial x^i}{\partial \bar{x}^l}\frac{\partial x^j}{\partial \bar{x}^m}\frac{\partial \bar{x}^p}{\partial x^k}$$

or

$$l\ m\ = 0;\quad _l = \begin{matrix} 1; & \text{for} & = l \\ 0; & \text{for} & \neq l \end{matrix};\ \text{etc.}$$

or

$$\bar{R}^h_{ijk} = 0;\quad \text{i.e.}\ R^h_{ijk} = 0:$$

Since if a tensor vanishes in one co-ordinate system, then it vanishes in all systems. It follows that Eq. (4.46) is a solution of $R^h_{ijk} = 0$; which is given. Hence, co-ordinate system $\bar{x}^i$ exists, and

$$\frac{\partial \bar{g}_{ij}}{\partial \bar{x}^k} = 0 \Rightarrow \bar{g}_{ij} = \text{constant}.$$

Consequently, a co-ordinate system $\bar{x}^i$ exists relative to which $\bar{g}_{ij}$ is constant. Hence, V_N is locally flat.

EXAMPLE 4.2.8 Determine whether the following metric is flat and/or Euclidean:

$$ds^2 = (dx^1)^2\quad (x^2)^2(dx^2)^2;\quad N = 2:$$

Solution: For the given metric, we have $g_{11} = 1;\ g_{12} = 0 = g_{21};\ g_{22} = \ (x^2)^2$: Since $g = \ (x^2)^2$; so the reciprocal tensors g^{ij} are given by

$$g^{11} = 1;\quad g^{12} = 0 = g^{21};\quad g^{22} = \frac{1}{(x^2)^2}:$$

Since the metric is not positive definite, it cannot be Euclidean. The only non-vanishing Christoffel symbol of second kind is $\,^{2}_{2\,2} = (x^2)^3$: Using Eq. (4.6), the Riemannian curvature tensors R_{ijk} can be written as

$$R^1_{212} = \frac{\partial}{\partial x^2}\,{}^{1}_{2\,1} + \frac{\partial}{\partial x^1}\,{}^{1}_{2\,2} + \,{}^{1}_{\ 1}\ \,{}_{2\,2}\ \,{}^{1}_{\ 2}\ \,{}_{2\,1} = 0:$$

$$R^2_{212} = \frac{\partial}{\partial x^2}\,{}^{2}_{2\,1} + \frac{\partial}{\partial x^1}\,{}^{2}_{2\,2} + \,{}^{2}_{\ 1}\ \,{}_{2\,2}\ \,{}^{2}_{\ 2}\ \,{}_{2\,1} = 0:$$

Using Eq. (4.13), the covariant curvature tensor is given by

$$R_{1212} = g_{1} \ R_{212} = g_{11}R^{1}_{212} + g_{12}R^{2}_{212} = 0:$$

Hence, the given space is flat.

EXAMPLE 4.2.9 Consider the Riemannian metric

$$ds^2 = (dx^1)^2 + 4(x^2)^2(dx^2)^2 + 4(x^3)^2(dx^3)^2 \quad 4(dx^4)^2(dx^4)^2:$$

(a) Calculate the Riemannian curvature. (b) Find the solution $(\bar{x}^i)$, by considering that the space is flat.

Solution: (a) For the given metric, we have $g_{11} = 1;\ g_{22} = 4(x^2)^2; g_{33} = 4(x^3)^2$ and $g_{44} = \quad 4(x^4)^2$: Since $g = \quad 64(x^2x^3x^4)^2$; so the reciprocal tensors g^{ij} are given by

$$g^{11} = 1; \quad g^{22} = \frac{1}{4(x^2)^2}; \quad g^{33} = \frac{1}{4(x^3)^2}; \quad g^{44} = \quad \frac{1}{4(x^4)^2}:$$

The non-vanishing Christoffel symbols of second kind are

$$\genfrac{}{}{0pt}{}{2}{2\ 2} = \frac{1}{x^2}; \quad \genfrac{}{}{0pt}{}{3}{3\ 3} = \frac{1}{x^3} \text{ and } \genfrac{}{}{0pt}{}{4}{4\ 4} = \frac{1}{x^4}:$$

Because $\genfrac{}{}{0pt}{}{i}{j\ k} = 0$ unless $i = j = k$; the partial derivative terms drop out of Eq. (4.6), leaving

$$R_{ijk} = \genfrac{}{}{0pt}{}{r}{i\ k}\ \genfrac{}{}{0pt}{}{}{r\ j} \qquad \genfrac{}{}{0pt}{}{r}{i\ j}\ \genfrac{}{}{0pt}{}{}{r\ k}$$

$$= \qquad\qquad = 0; \text{ (not summed)},$$

which in turn implies that $R_{hijk} = 0$ and $\ = 0$:
(b) For the above calculated Christoffel symbols, system (4.43) becomes,

$$\frac{\partial u_1}{\partial x^1} = 0; \quad \frac{\partial u_2}{\partial x^2} = \frac{u_2}{x^2}; \quad \frac{\partial u_3}{\partial x^3} = \frac{u_3}{x^3}; \quad \frac{\partial u_4}{\partial x^4} = \frac{u_4}{x^4};$$

with $\frac{\partial u_i}{\partial x^j} = 0$ for $i \neq j$: Integrating,

$$u_1 = f_1\ x^2; x^3; x^4\ ;\ u_2 = x^2 f_2\ x^1; x^3; x^4\ ;$$

$$u_3 = x^3 f_3\ x^1; x^2; x^4\ ;\ u_4 = x^4 f_4\ x^1; x^2; x^3$$

for arbitrary functions f_i. But the remaining part of Eq. (4.43), $\frac{\partial \omega}{\partial x^i} = u_i$, gives rise to the compatibility relations $\frac{\partial u_i}{\partial x^j} = \frac{\partial u_j}{\partial x^i}$, which is satisfied only if $f_i = c_i$ = constant. Therefore,

$$\omega = a_1 x^1 + a_2 (x^2)^2 + a_3 (x^3)^2 + a_4 (x^4)^2 + a_5$$

and the transformation must be of the general form,

$$\bar{x}^k = a^k_1 x^1 + a^k_2 (x^2)^2 + a^k_3 (x^3)^2 + a^k_4 (x^4)^2 + a^k_5; \quad a^k_i \text{ constants.}$$

EXAMPLE 4.2.10 For the Euclidean space under the following metric:

$$ds^2 = [(x^1)^2 + (x^2)^2](dx^1)^2 + [(x^1)^2 + (x^2)^2](dx^2)^2 + (dx^3)^2$$

exhibit a transformation from the co-ordinate system (x^i) to a rectangular system $(\bar{x}^i)$.

Solution: For the given metric, we have $g_{11} = (x^1)^2 + (x^2)^2 = g_{22}$, $g_{33} = 1$; $g_{ij} = 0 = g_{ji}$; $j \neq i$. Since $g = [(x^1)^2 + (x^2)^2]^2$, so the reciprocal tensors g^{ij} are given by

$$g^{11} = \frac{1}{(x^1)^2 + (x^2)^2}; \quad g^{22} = \frac{1}{(x^1)^2 + (x^2)^2}; \quad g^{33} = 1.$$

The non-vanishing Christoffel symbols of second kind are

$$\left\{\begin{matrix}1\\1\;1\end{matrix}\right\} = \frac{x^1}{(x^1)^2 + (x^2)^2} = \left\{\begin{matrix}2\\1\;2\end{matrix}\right\}; \quad \left\{\begin{matrix}1\\2\;2\end{matrix}\right\} = -\frac{x^1}{(x^1)^2 + (x^2)^2}$$

$$\left\{\begin{matrix}1\\1\;2\end{matrix}\right\} = \frac{x^2}{(x^1)^2 + (x^2)^2} = \left\{\begin{matrix}2\\2\;2\end{matrix}\right\}; \quad \left\{\begin{matrix}2\\1\;1\end{matrix}\right\} = -\frac{x^2}{(x^1)^2 + (x^2)^2}.$$

For the above calculated Christoffel symbols, system Eq. (4.43) becomes,

$$\frac{\partial u_1}{\partial x^1} = \frac{x^1 u_1 - x^2 u_2}{\sqrt{g}}; \quad \frac{\partial u_1}{\partial x^2} = \frac{x^2 u_1 + x^1 u_2}{\sqrt{g}}; \quad \frac{\partial u_1}{\partial x^3} = 0; \tag{i}$$

$$\frac{\partial u_2}{\partial x^1} = \frac{x^2 u_1 + x^1 u_2}{\sqrt{g}}; \quad \frac{\partial u_2}{\partial x^2} = \frac{-x^1 u_1 + x^2 u_2}{\sqrt{g}}; \quad \frac{\partial u_2}{\partial x^3} = 0; \tag{ii}$$

$$\frac{\partial u_3}{\partial x^1} = 0; \quad \frac{\partial u_3}{\partial x^2} = 0; \quad \frac{\partial u_3}{\partial x^3} = 0. \tag{iii}$$

Thus, u_1 and u_2 are functions of x^1, x^2 alone, and u_3 = constant. Since the g_{ij} are all polynomials of degree 2 in x^1, x^2, use the method of undetermined coefficients, assuming polynomial forms

$$u_i = a_i (x^1)^2 + b_i x^1 x^2 + c_i (x^2)^2 + d_i x^1 + e_i x^2 + f_i; \quad i = 1, 2.$$

The compatibility relations $\frac{\partial u_1}{\partial x^2} = \frac{\partial u_2}{\partial x^1}$ implied by the second part of Eq. (i) and the first part of Eq. (ii) require,

$$b_1 = 2a_2; \quad 2c_1 = b_2; \quad e_1 = d_2.$$

Similarly, the compatibility relations $\frac{\partial u_1}{\partial x^1} = -\frac{\partial u_2}{\partial x^2}$ imply

$$2a_1 = -b_2; \quad b_1 = -2c_2; \quad d_1 = -e_2.$$

From the first part of Eq. (i), we get

$$a_1 = 0 = a_2; \quad c_1 = -b_2; \; b_1 = -c_2; \; d_1 = -e_2; \; f_1 = 0 = -f_2.$$
$$\Rightarrow \quad b_1 = b_2 = c_1 = c_2 = 0.$$

Thus, the solution can be written in the form

$$u_1 = ax^1 + bx^2; \quad u_2 = bx^1 - ax^2; \quad u_3 = c;$$

where we are redenoting the constants d_1 and e_1. The first part of Eq. (4.43) gives

$$\frac{\partial \phi}{\partial x^1} = ax^1 + bx^2; \quad \frac{\partial \phi}{\partial x^2} = bx^1 - ax^2; \quad \frac{\partial \phi}{\partial x^3} = c;$$

or

$$\phi = \frac{a}{2}(x^1)^2 + bx^1x^2 - \frac{a}{2}(x^2)^2 + cx^3 + d$$

or

$$\bar{x}^k = \frac{a^k}{2}(x^1)^2 + b^k x^1 x^2 - \frac{a^k}{2}(x^2)^2 + c^k x^3 \text{ with } d = 0.$$

It is clear that we take $c^1 = c^2 = 0 = a^3 = b^3$ and $c^3 = 1$;

$$\bar{x}^1 = \frac{1}{2}a^1(x^1)^2 + b^1x^1x^2 - \frac{1}{2}a^1(x^2)^2$$
$$\bar{x}^2 = \frac{1}{2}a^2(x^1)^2 + b^2x^1x^2 - \frac{1}{2}a^2(x^2)^2; \quad \bar{x}^3 = x^3.$$

Now, the Jacobian matrix J satisfies $J^TJ = G$. Here

$$J = \begin{bmatrix} a^1(x^1)^2 + b^1x^1x^2 & b^1x^1x^2 - \frac{1}{2}a^1(x^2)^2 & 0 \\ a^2(x^1)^2 + b^2x^1x^2 & b^2x^1x^2 - \frac{1}{2}a^2(x^2)^2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\Rightarrow \quad (a^1)^2 + (a^2)^2 = 1; \; a^1b^1 + a^2b^2 = 0; \; (b^1)^2 + (b^2)^2 = 1; \quad \text{as } J^TJ = G.$$

So, take $a^1 = 0; a^2 = 1; b^2 = 0; b^1 = 1.$ Finally, the transformation is given by

$$\bar{x}^1 = x^1x^2; \quad \bar{x}^2 = \frac{1}{2}[(x^1)^2 - (x^2)^2]; \quad \bar{x}^3 = x^3.$$

Deduction 4.2.3 Let the metric tensors g_{ij} are constants, then all the partial derivatives of g_{ij} are zero. Consequently, all Christoffel symbols will vanish and all $R_{ijkl} = 0$; i.e. $= 0$: Thus, by Theorem 4.2.2, the space is flat. If $\overline{x} = Ax$; then $J = A$ and

$$G = J^T \overline{G} J = A^T \overline{G} A:$$

However, since G is real and symmetric, its eigenvectors form an orthogonal matrix which we now choose as A, with

$$AGA^{-1} = AGA^T = D; \quad D = \text{diagonal matrix}$$

of eigenvalues of G. Hence, $\overline{G} = AGA^T = D$: Hence, if the metric tensor is constant, the space is flat and the transformation $\overline{x} = Ax$; where A is a rank N matrix of eigenvalues $G = (g_{ij})$, diagonalises the metric, i.e. $\overline{g}_{ij} = 0$ for $i \neq j$:

EXAMPLE 4.2.11 Find the signature of the flat metric

$$ds^2 = 4(dx^1)^2 + 5(dx^2)^2 - 2(dx^4)^2 - 4dx^2dx^3 - 4dx^2dx^4 - 10dx^3dx^4:$$

Solution: For the given metric, we have $g_{11} = 4$; $g_{22} = 5$; $g_{33} = -2$; $g_{44} = 2$, $g_{23} = -2$; $g_{24} = -2$ and $g_{34} = -5$: Now, we find the eigenvalues of $G = (g_{ij})$. The characteristic equation is

$$|G - \lambda I| = \begin{vmatrix} 4-\lambda & 0 & 0 & 0 \\ 0 & 5-\lambda & -2 & -2 \\ 0 & -2 & -2-\lambda & -5 \\ 0 & -2 & -5 & 2-\lambda \end{vmatrix} = 0$$

$$\Rightarrow \quad (4-\lambda)(5-\lambda)(37-\lambda^2) = 0$$

$$\Rightarrow \quad \lambda = 4; 5; \pm\sqrt{37}:$$

This means that there is a transformation which changes the metric into the form

$$\begin{aligned} ds^2 &= 4(dx^1)^2 + 5(dx^2)^2 + \sqrt{37}\,(dx^3)^2 - (dx^4)^2 \\ &= (d\overline{x}^1)^2 + (d\overline{x}^2)^2 + (d\overline{x}^3)^2 - (d\overline{x}^4)^2; \end{aligned}$$

with the obvious change of co-ordinates. Hence, the signature is $(+ + + -)$; or some permutation thereof.

4.2.5 Projective Curvature Tensor

The projective curvature tensor or Weyl tensor, denoted by W_{hijk}; is defined by

$$W_{hijk} = R_{hijk} + \frac{1}{1-N}\left(g_{kl} R_{hj} - g_{kh} R_{ij}\right); \qquad (4.47)$$

where R_{hijk} is the Riemannian curvature tensor and R_{ij} is the Ricci tensor.

Theorem 4.2.3 A necessary and sufficient condition that a Riemannian space V_N $(N > 3)$ to be of constant Riemannian curvature is that the Weyl tensor vanishes identically throughout V_N.

Proof: Let be the constant Riemannian curvature of V_N given by

$$= \frac{p^h q^i p^j q^k R_{hijk}}{p^h q^i p^j q^k [g_{hj} g_{ik} \quad g_{ij} g_{hk}]} = \text{constant}:$$

Since is constant, it is independent of the orientation determined by p^i and q^j and hence

$$= \frac{R_{hijk}}{g_{hj} g_{ik} \quad g_{ij} g_{hk}} \;) \; R_{hijk} = (\quad g_{ij} g_{hk} + g_{hj} g_{ik})$$

or

$$g^{hk} R_{hijk} = g^{hk} (\quad g_{ij} g_{hk} + g_{hj} g_{ik})$$

or

$$R_{ij} = g^{hk} (\quad N g_{ij} + {}^{k}_{j} g_{ik}) = (1 \quad N) g_{ij}$$

or

$$g^{ij} R_{ij} = (1 \quad N) g^{ij} g_{ij}$$

or

$$R = (1 \quad N) N \quad) \; R_{ij} = \frac{R}{N} g_{ij} : \tag{4.48}$$

Since is constant, Eq. (4.48) shows that R is constant. Now, the Weyl tensor W_{hijk} is given by

$$\begin{aligned} W_{hijk} &= R_{hijk} + \frac{1}{1 \quad N} (g_{ki} R_{hj} \quad g_{kh} R_{ij}) \\ &= R_{hijk} + \frac{1}{1 \quad N} \; g_{ki} \frac{1}{N} R g_{hj} \quad g_{kh} \frac{1}{N} R g_{ij} \\ &= R_{hijk} + \frac{R}{N (1 \quad N)} [g_{ki} g_{hj} \quad g_{hk} g_{ij}] \\ &= R_{hijk} + \frac{R}{N (1 \quad N)} \frac{1}{} R_{hijk} \\ &= R_{hijk} + \frac{1}{} R_{hijk} = 2 R_{hijk} = 0; \text{ as } \quad \text{is constant.} \end{aligned}$$

Thus, the condition is necessary. Conversely, let the Weyl tensor $W_{hijk} = 0$; we have to prove that $= \text{constant}$. Since $W_{hijk} = 0$; we have

$$R_{hijk} + \frac{1}{1 \quad N} [g_{ik} R_{hj} \quad g_{hk} R_{ij}] = 0$$

or

$$g^{hk}R_{hijk} + \frac{1}{1 \quad N}\,{}^{h}\, g^{hk}g_{ik}R_{hj} \qquad g^{hk}g_{hk}R_{ij}\,{}^{i} = 0$$

or

$$R_{ij} + \frac{1}{1 \quad N}\,{}^{h}\;{}^{h}_{i}R_{hj} \qquad NR_{ij}\,{}^{i} = 0$$

or

$$R_{ij} + \frac{1}{1 \quad N}[R_{ij} \qquad NR_{ij}] = 0$$

or

$$R_{ij} + \frac{1 \quad N}{1 \quad N}R_{ij} = 0\ ) \quad R_{ij} = 0$$

or

$$R_{ij} = g^{hk}R_{hijk} = 0\ ) \quad \text{either } g^{hk} = 0 \text{ or } R_{hijk} = 0:$$

If $R_{hijk} = 0$; then the Riemann curvature $= 0$: If $g^{hk} = 0$; then $g_{hk} = 0$ and hence,

$$= \frac{p^h q^i p^j q^k R_{hijk}}{p^h q^i p^j q^k [g_{hj}\, g_{ik} \qquad 0 g_{ij}]} = \frac{p^h q^i p^j q^k R_{hijk}}{p^h q^i p^j q^k g_{hj}\, g_{ik}}$$

$$= \frac{p^h q^i p^j q^k R_{hijk}}{(p^h p^j g_{hj})(q^i q^k g_{ik})}$$

$$= \frac{p^h q^i p^j q^k R_{hijk}}{p^2 q^2};\ p^2 = 1 = \ g_{hj}\, p^h p^j;\ \ q^2 = 1 = \ g_{ik}\, q^i q^k$$

or

$$= p^h q^i p^j q^k R_{hijk} = \text{constant as } R_{ij} = 0:$$

Since 0 is constant, in either case is a constant. Thus, the condition is necessary.

4.2.6 Uniform Vector Field

The construction of a field of parallel vectors is possible only when the Riemann{Christoffel curvature tensor vanishes. Here, we have to prove that, when Riemann{Christoffel tensor vanishes the differential equations

$$A_{;} = \frac{@A}{@x} \qquad {}^{a} \qquad A_a = 0$$

are integrable. Let A be a covariant vector, then by definition

$$\frac{@A}{@x} \quad A \qquad\qquad = A_{;} = 0; \tag{4.49}$$

where we suppose that $A_{;} = 0$: We want to show that Eq. (4.49) can be integrated only when curvature tensor $R^a_{ijk} = 0$: From Eq. (4.49) we get,

$$\frac{\partial A}{\partial x} dx = A \qquad dx$$

or

$$\int \frac{\partial A}{\partial x} dx = \int A \qquad dx$$

or

$$\int dA = A = \int A \qquad dx : \tag{4.50}$$

This means that Eq. (4.49) is integrable only when the right hand side of Eq. (4.50) is integrable, for which the condition is $A \qquad dx$ must be perfect differential. So, we can write,

$$A \qquad dx = dB = \frac{\partial B}{\partial x} dx$$

or

$$A \qquad \frac{\partial B}{\partial x} \quad dx = 0$$

or

$$A \qquad = \frac{\partial B}{\partial x}; \quad \text{as } dx \text{ is arbitrary.} \tag{4.51}$$

Differentiating Eq. (4.51) with respect to x, we get

$$\frac{\partial A}{\partial x} \qquad + A \frac{\partial}{\partial x} \qquad = \frac{\partial^2 B}{\partial x \partial x}: \tag{4.52}$$

Interchanging and we get

$$\frac{\partial A}{\partial x} \qquad + A \frac{\partial}{\partial x} \qquad = \frac{\partial^2 B}{\partial x \partial x}: \tag{4.53}$$

Subtracting Eq. (4.53) from Eq. (4.52), we get

$$\frac{\partial A}{\partial x} \qquad + A \frac{\partial}{\partial x} \qquad \frac{\partial A}{\partial x} \qquad A \frac{\partial}{\partial x} \qquad = 0$$

or

$$A_c \quad {}^{c} \quad + A \frac{@}{@x} \qquad\qquad A_c \quad {}^{c}$$

$$A_c \frac{@}{@x} \quad {}^{c} \qquad = 0 \text{ by Eq. } (4.49)$$

or

$$A_c \quad {}^{c} \quad + A_c \frac{@}{@x} \quad {}^{c} \qquad\qquad {}^{c} \quad A_c \quad A_c \frac{@}{@x} \quad {}^{c} \qquad = 0$$

or

$${}^{c} \quad + \frac{@}{@x} \quad {}^{c} \qquad\qquad {}^{c} \qquad \frac{@}{@x} \quad {}^{c} \qquad A_c = 0$$

or

$$\frac{@}{@x} \quad {}^{c} \quad + \frac{@}{@x} \quad {}^{c} \qquad\qquad {}^{c} \quad + \qquad {}^{c} \qquad A_c = 0$$

or

$$R^{c} \quad A_c = 0; \quad \text{i.e. } R^{c} \quad = 0; \text{ as } A_c \text{ is arbitrary}:$$

Thus, we see that the RHS of Eq. (4.50) is integrable only when the curvature tensor vanishes. Consequently, Eq. (4.49) is integrable only when $R^{c} \quad = 0$:

Now, we can carry the vector A to any point by parallel displacement. This gives a unique result independent of the path of transfer. If a vector is displaced in this way we obtain a uniform vector field. Thus, the construction of uniform vector field is only possible when the curvature tensor vanishes.

4.3 Einstein Space

Einstein space is defined as a space which is homogeneous with regard to the Ricci tensor R_{ij}. Thus, a space for which

$$R_{ij} = \ g_{ij}; \qquad = \text{invariant at all points} \tag{4.54}$$

at every point of the space, then that space is called Einstein space. Inner multiplication of g^{ij} shows that

$$g^{ij} R_{ij} = \ g^{ij} g_{ij} \ ) \quad R = \ N: \tag{4.55}$$

Thus for an Einstein space,

$$R_{ij} = \ g_{ij} = \frac{1}{N} R g_{ij}: \tag{4.56}$$

Hence, for an Einstein space Eq. (4.56) holds at every point of the space.

Theorem 4.3.1 A space of constant curvature is an Einstein space.

Proof: Let the Riemannian curvature κ at P of V_N for the orientation determined by p^i and q^i, is given in Eq. (4.34) as

$$\kappa = \frac{p^h q^i p^j q^k R_{hijk}}{p^h q^i p^j q^k \left[g_{hj} g_{ik} - g_{ij} g_{kh}\right]}.$$

Since κ is constant and independent of the orientation so,

$$\kappa = \frac{R_{hijk}}{g_{hj} g_{ik} - g_{ij} g_{kh}}$$

or

$$R_{hijk} = \kappa\left[g_{hj} g_{ik} - g_{ij} g_{kh}\right].$$

Multiplying by g^{hk}, we get

$$\kappa\, g^{hk}\left[g_{hj} g_{ik} - g_{ij} g_{kh}\right] = g^{hk} R_{hijk}$$

or

$$\kappa\left(\delta^h_i g_{hj} - N g_{ij}\right) = R_{ij}, \quad \text{since}, g^{hk} R_{hijk} = R_{ij}$$

or

$$\kappa\left(g_{ij} - N g_{ij}\right) = R_{ij}$$

or

$$\kappa(1-N)\, g_{ij}\, g^{ij} = R_{ij}\, g^{ij}; \text{ multiplying by } g^{ij}$$

or

$$\kappa N(1-N) = R, \quad \text{as} R_{ij}\, g^{ij} = R.$$

Therefore,

$$R_{ij} = \kappa(1-N) g_{ij} \frac{R}{N(1-N)} = \frac{R}{N} g_{ij}.$$

This is the necessary and sufficient condition for the space V_N to be Einstein space.

EXAMPLE 4.3.1 If $g^{ik} R_{kj} = R^i_j$ and $g^{ij} R_{ij} = R$, show that $R^i_{j;i} = \frac{1}{2}\frac{\partial R}{\partial x^j}$. Hence, show that for an Einstein space, the scalar curvature R is constant.

Solution: Here, we use the Bianchi identity, Eq. (4.20), $R_{pkjt;i} + R_{pkti;j} + R_{pkij;t} = 0$. The covariant differentiation of the relation $g^{ik} R_{kj} = R^i_j$ with respect to x^l is given by

$$R^i_{j;l} = g^{ik}_{;l} R_{kj} + g^{ik} R_{kj;l} = g^{ik} R_{kj;l}; \quad \text{as} \, g^{ik}_{;l} = 0.$$

Thus, the divergence of R^i_j is given by

$$R^i_{j;i} = g^{ik}R_{kj;i} = g^{ik}R^t_{kjt;i}\ ;\ \text{as } R^t_{kit} = R_{kj}$$
$$= g^{ik}\left(g^{pt}R_{pkjt}\right)_{;i}\ ;\quad \text{as } g^{pt}R_{pkjt} = R^t_{kjt}$$
$$= g^{ik}g^{pt}R_{pkjt;i}\ ;\quad \text{as } g^{pt}_{;i} = 0$$
$$= g^{ik}g^{pt}\left[R_{pkti;j} + R_{pkij;t}\right] = g^{pt}\left[g^{ik}R_{pkti;j} + g^{ik}R_{pkij;t}\right]$$
$$= g^{pt}\left[g^{ki}R_{kpti;j} + g^{ik}R_{kpji;t}\right]$$
$$\text{as } R_{kpti} = R_{pkti} \text{ and } R_{kpji} = R_{pkij}$$
$$= g^{pt}\left(R^i_{pti;j} + R^i_{pji;t}\right) = g^{pt}\left[R_{pt;j} + R_{pj;t}\right]$$
$$= \left(g^{pt}R_{pt}\right)_{;j}\quad g^{pt}R_{pj;t} = R_{;j}\quad R^t_{j;t} = R_{;j}\quad R^i_{j;i}$$
replacing the dummy index t by i

or

$$2R^i_{j;i} = R_{;j} = \frac{\partial R}{\partial x^j} \Rightarrow R^i_{j;i} = \frac{1}{2}\frac{\partial R}{\partial x^j}.$$

For an Einstein space, $R_{ij} = \frac{R}{N}g_{ij}$, therefore,

$$g^{ki}R_{ij} = \frac{R}{N}g_{ij}\,g^{ki} \Rightarrow R^k_j = \frac{R}{N}\delta^k_j$$

or

$$R^k_{j;k} = \frac{1}{N}R_{;k}\,\delta^k_j = \frac{1}{N}R_{;j}$$

or

$$\frac{1}{2}R_{;j} = \frac{1}{N}R_{;j} \Rightarrow R_{;j} = 0;\quad \text{for } N > 2.$$

Therefore, for an Einstein space, the scalar curvature R is constant.

EXAMPLE 4.3.2 For a V_2 referred to an orthogonal system of parametric curves $(g_{12} = 0)$, show that

$$R_{12} = 0;\quad R_{11}g_{22} = R_{22}g_{11} = R_{1221}$$
$$R = g^{ij}R_{ij} = \frac{2R_{1221}}{g_{11}g_{22}};\quad \text{consequently, } R_{ij} = \frac{R}{2}g_{ij}.$$

Hence, show that V_2 is an Einstein space.

Solution: Given that, in a V_2, $g_{12} = 0$ so that $g^{12} = 0$. Also,

$$g^{ij} = \frac{1}{g_{ij}} \Rightarrow g^{11} = \frac{1}{g_{11}}, \quad g^{22} = \frac{1}{g_{22}}.$$

The metric of V_2 is given by

$$\begin{aligned} ds^2 &= g_{ij}\, dx^i dx^j, \quad i, j = 1, 2 \\ &= g_{11}(dx^1)^2 + g_{22}(dx^2)^2, \quad \text{as } g_{12} = 0. \end{aligned}$$

Also, we know that $g^{hk}R_{hijk} = R_{ij}$ and

$$g = \begin{vmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{vmatrix} = \begin{vmatrix} g_{11} & 0 \\ 0 & g_{22} \end{vmatrix} = g_{11}g_{22}.$$

(a) Here, we have to show that $R_{12} = 0$. For this,

$$\begin{aligned} R_{12} &= g^{hk}R_{h12k} = g^{h1}R_{h121} + g^{h2}R_{h122} = g^{h1}R_{h121}, \quad \text{as } R_{h122} = 0 \\ &= g^{11}R_{1121} + g^{21}R_{2121} = 0, \quad \text{as } g^{12} = 0 \text{ and } R_{1121} = 0. \end{aligned}$$

(b) Using the relation $g^{hk}R_{hijk} = R_{ij}$ we get

$$R_{11} = g^{hk}R_{h11k} = g^{2k}R_{211k} = g^{22}R_{2112} = \frac{R_{2112}}{g_{22}}.$$

$$R_{22} = g^{hk}R_{h22k} = g^{1k}R_{122k} = g^{11}R_{1221} = \frac{R_{1221}}{g_{11}}.$$

Since $R_{2112} = R_{1221}$, it follows that, $R_{11}g_{22} = R_{1221} = R_{22}g_{11}$. Therefore, in a two-dimensional Riemannian space the components of a Ricci tensor are proportional to the components of a metric tensor.

(c) From the definition of R, we get

$$\begin{aligned} R &= g^{ij}R_{ij} = g^{i1}R_{i1} + g^{i2}R_{i2} \\ &= g^{11}R_{11} + g^{22}R_{22}, \quad \text{as } g^{12} = 0 \\ &= \frac{R_{11}}{g_{11}} + \frac{R_{22}}{g_{22}} = \frac{R_{1221}}{g_{22}g_{11}} + \frac{R_{1221}}{g_{11}g_{22}} \\ &= \frac{2R_{1221}}{g_{11}g_{22}}, \quad \text{as } R_{2112} = R_{1221}. \end{aligned}$$

(d) From (c), we see that

$$R = \frac{2R_{1221}}{g_{11}g_{22}} = \frac{2R_{1221}}{g}; \quad \text{as } g = g_{11}g_{22}$$
$$\Rightarrow \quad R_{1221} = \frac{1}{2}Rg.$$

Therefore, the relation in (b) can be written in the form

$$R_{11}g_{22} = \frac{1}{2}Rg = R_{22}g_{11}$$

from which it follows that

$$R_{11} = \frac{Rg}{2g_{22}} = \frac{Rg_{11}g_{22}}{2g_{22}} = \frac{1}{2}Rg_{11}$$
$$R_{22} = \frac{Rg}{2g_{11}} = \frac{Rg_{11}g_{22}}{2g_{11}} = \frac{1}{2}Rg_{22}$$
$$R_{12} = \frac{1}{2}Rg_{12} = 0; \quad \text{as } g_{12} = 0.$$

Consequently, we get $R_{ij} = \frac{R}{2}g_{ij}$. This shows that V_2 is an Einstein space.

4.4 Mean Curvature

Let $\lambda^i_{h|}$ be the components of unit vector in a direction at a point P of a V_N. Let $\lambda^i_{k|}$ be the components of unit vector forming an orthogonal ennuple. One vector out of N vectors $\lambda^i_{k|}$ is $\lambda^i_{h|}$ and hence, we can say that $\lambda^i_{k|}$ are components of $N-1$ vectors where $h \neq k$ and $k = 1, 2, \ldots, N$.

Let the Riemannian curvature at P of V_N for the orientation determined by $\lambda^i_{h|}$ and $\lambda^i_{k|}$ $(k \neq h)$ be denoted by K_{hk} and is given by

$$K_{hk} = \frac{\lambda^p_{h|}\lambda^q_{k|}\lambda^r_{h|}\lambda^s_{k|}R_{pqrs}}{\lambda^p_{h|}\lambda^q_{k|}\lambda^r_{h|}\lambda^s_{k|}(g_{pr}g_{qs} - g_{ps}g_{qr})}$$
$$= \frac{\lambda^p_{h|}\lambda^q_{k|}\lambda^r_{h|}\lambda^s_{k|}R_{pqrs}}{\left(\lambda^p_{h|}\lambda^r_{h|}g_{pr}\right)\left(\lambda^q_{k|}\lambda^s_{k|}g_{qs}\right) - \left(\lambda^p_{h|}\lambda^s_{k|}g_{ps}\right)\left(\lambda^q_{k|}\lambda^r_{h|}g_{qr}\right)}. \tag{4.57}$$

Since unit vectors $\lambda^i_{h|}$ and $\lambda^i_{k|}$ are orthogonal, so,

$$\lambda^p_{h|}\lambda^r_{h|}g_{pr} = 1 \text{ and } \lambda^p_{h|}\lambda^s_{k|}g_{ps} = 0.$$

Therefore, from Eq. (4.57), we get

$$K_{hk} = \frac{{}^{p}_{hj}\,{}^{q}_{kj}\,{}^{r}_{hj}\,{}^{s}_{kj}R_{pqrs}}{1{:}1 \quad 0{:}0} = \quad {}^{p}_{hj}\,{}^{q}_{kj}\,{}^{r}_{kj}\,{}^{s}_{kj}R_{pqrs}: \qquad (4.58)$$

Since the right-hand side of Eq. (4.58) vanishes for $h = k$, and so we say K_{hh} is equal to zero. Let $M_h = \sum_{k=1}^{N} K_{hk}$: Hence from Eq. (4.58) we have,

$$\sum_{k=1}^{N} K_{hk} = M_h = \sum_{k=1}^{N} {}^{p}_{hj}\,{}^{q}_{kj}\,{}^{r}_{hj}\,{}^{s}_{kj}R_{pqrs}$$

$$= {}^{p}_{hj}\,{}^{q}_{hj}\sum_{k=1}^{N} {}^{q}_{kj}\,{}^{s}_{kj}R_{pqrs}$$

$$= {}^{p}_{hj}\,{}^{r}_{hj}g^{qs}R_{pqrs}; \text{ as } \sum_{k=1}^{N} {}^{q}_{kj}\,{}^{s}_{kj} = g^{qs}$$

$$= \quad {}^{p}_{hj}\,{}^{r}_{hj}g^{qs}R_{qprs} = \quad {}^{p}_{hj}\,{}^{r}_{hj}R^{s}_{pqs} = \quad {}^{p}_{hj}\,{}^{r}_{hj}R_{pr}: \qquad (4.59)$$

This shows that M_h is independent of $(N \quad 1)$ orthogonal direction choosen to complete an orthogonal ennuple. Here, M_h is defined as mean curvature or Riccian curvature of V_N for the direction ${}^{p}_{hj}$: Summing Eq. (4.58) over the N mutually orthogonal directions, we get

$$\sum_{h=1}^{N} M_h = \sum_{h=1}^{N} {}^{p}_{hj}\,{}^{r}_{hj}R_{pr} = \quad g^{pr}R_{pr} = \quad R;$$

where R is the scalar curvature. This proves that the sum of mean curvatures for N mutual orthogonal directions is independent of the directions chosen to complete an orthogonal ennuple and has the value R:

4.4.1 Ricci's Principal Directions

Let us suppose that ${}^{i}_{hj}$ is not a unit vector and therefore, the mean curvature M_h in this case is given by

$$M_h = \frac{R_{ij}\,{}^{i}_{hj}\,{}^{j}_{hj}}{g_{ij}\,{}^{i}_{hj}\,{}^{j}_{hj}} \quad) \quad (R_{ij} + M_h g_{ij})\,{}^{i}_{hj}\,{}^{j}_{hj} = 0:$$

Differentiating this equation with respect to ${}^{i}_{hj}$; we get

$$\frac{@M_h}{@\,{}^{i}_{hj}} g_{jk}\,{}^{j}_{hj}\,{}^{k}_{hj} + 2(R_{ij} + M_h g_{ij})\,{}^{j}_{hj} = 0:$$

For extreme value of M_h, we must have $\frac{\partial M_h}{\partial\, {}^{i}_{hj}} = 0$; and so

$$(R_{ij} + M_h g_{ij})\ {}^{j}_{hj} = 0;$$

which are called Ricci's principal direction of the space and they are principal directions of Ricci tensor R_{ij}.

EXAMPLE 4.4.1 For a two-dimensional manifold, prove that $K_{pq} = \frac{R}{2}$:

Solution: In V_2, we know R_{1212} is the only non-vanishing Riemann–Christoffel tensor given by

$$R_{1212} = \frac{1}{2}Rg;\ R = \text{ total curvature.}$$

Let the Riemannian curvature at P of V_N for the orientation determined by ${}^{i}_{pj}$ and ${}^{i}_{qj}$ $(q \neq p)$ be denoted by K_{pq} and from Eq. (4.57) we get,

$$K_{pq} = \frac{{}^{h}_{pj}\ {}^{i}_{qj}\ {}^{j}_{pj}\ {}^{k}_{qj} R_{hijk}}{{}^{h}_{pj}\ {}^{i}_{qj}\ {}^{j}_{pj}\ {}^{k}_{qj} (g_{hj}\, g_{ik} \quad g_{hk}\, g_{ij})}$$

$$= \frac{{}^{1}_{pj}\ {}^{2}_{qj}\ {}^{1}_{pj}\ {}^{2}_{qj} R_{1212}}{{}^{1}_{pj}\ {}^{2}_{qj}\ {}^{1}_{pj}\ {}^{2}_{qj} (g_{11}g_{22} \quad g_{12}g_{21})} = \frac{R_{1212}}{g_{12}^2 \quad g_{11}g_{22}} = \frac{R_{1212}}{g} = \frac{1}{2}R:$$

4.5 Geodesics in a V_N

Consider a curve : $x^i = x^i(t)$ in a V_N, an N-dimensional Riemannian space, on it and let $x^i(t)$ be the co-ordinates of a general point P on it. Let the points A and B of the path correspond, respectively, to the value t_1 and t_2 of the parameter.

Consider two points on the path close to each other with co-ordinates x^i and $x^i + dx^i$. Let the parameter take values t and $t + dt$ corresponding to these two points, respectively. The interval ds between the two points is given by

$$ds^2 = g_{ij}\, dx^i dx^j = g_{ij} \frac{dx^i}{dt}\frac{dx^j}{dt} dt^2: \tag{4.60}$$

The total interval between A and B along the path therefore,

$$S = \int_{t_1}^{t_2} \sqrt{g_{ij} \frac{dx^i}{dt}\frac{dx^j}{dt}}\, dt$$

$$= \int_A^B F(x^i; \underline{x}^i)dt;\ \text{say; where } F(x^i; \underline{x}^i) = \sqrt{g_{ij}\, \underline{x}^i \underline{x}^j}: \tag{4.61}$$

In order C is a geodesic, integral Eq. (4.61) should be stationary. In order that integral Eq. (4.61) is stationary, we have the Euler-Lagrange condition,

$$\frac{\partial F}{\partial x^l} - \frac{d}{dt}\left(\frac{\partial F}{\partial \dot x^l}\right) = 0$$

or

$$\frac{1}{2\dot s}\frac{\partial g_{ij}}{\partial x^l}\dot x^i \dot x^j - \left[-\frac{\ddot s}{\dot s} g_{li}\dot x^i + \frac{1}{\dot s}\frac{\partial g_{li}}{\partial x^j}\dot x^j \dot x^i + \frac{1}{\dot s} g_{li}\ddot x^i\right] = 0$$

or

$$g_{li}\ddot x^i - \frac{\ddot s}{\dot s} g_{li}\dot x^i + [li;j]\dot x^i \dot x^j = 0. \tag{4.62}$$

Multiplying this relation by g^{lm} and perform summation over l, we get

$$\ddot x^m - \frac{\ddot s}{\dot s}\dot x^m + \left\{ {m \atop i\ j} \right\}\dot x^i \dot x^j = 0. \tag{4.63}$$

Equation (4.63) of geodesic can be further simplified by taking $s = t$, so that $\dot s = 1$; $\ddot s = 0$; and therefore, Eq. (4.63) becomes

$$\frac{d^2x^i}{ds^2} + \left\{ {i \atop j\ k} \right\}\frac{dx^j}{ds}\frac{dx^k}{ds} = 0; \tag{4.64}$$

where s itself is chosen as the parameter. This, of course, represents a set of N coupled differential equations. Equation (4.64) is known as geodesic equation in V_N.

EXAMPLE 4.5.1 Show that on the surface of a sphere all great circles are geodesics while no other circle is a geodesic.

Solution: The components of the metric tensor on the surface of a sphere of radius a are $g_{11} = a^2; g_{22} = a^2\sin^2\theta; g_{12} = g_{21} = 0$ and $g = a^4\sin^2\theta$ so that the metric is $ds^2 = a^2(d\theta)^2 + a^2\sin^2\theta\,(d\phi)^2$. The non-vanishing Christoffel symbols are

$$\left\{ {2 \atop 1\ 2} \right\} = \cot\theta = \left\{ {2 \atop 2\ 1} \right\}; \quad \left\{ {1 \atop 2\ 2} \right\} = -\sin\theta\cos\theta.$$

Thus, the geodesic equation reduces to (with $\theta = x^1$; $\phi = x^2$)

$$\frac{d^2\theta}{ds^2} - \sin\theta\cos\theta\left(\frac{d\phi}{ds}\right)^2 = 0 \tag{i}$$

$$\frac{d^2\phi}{ds^2} + 2\cot\theta\frac{d\theta}{ds}\frac{d\phi}{ds} = 0. \tag{ii}$$

(a) Consider a great circle on the surface of the sphere and choose the z-axis ($\theta = 0$) to be normal to the plane of the circle, so that the great circle is the equator. The parametric equation of the great circle is

$$\theta = \frac{\pi}{2}; \quad \phi = cs + d; \quad c \neq 0; \qquad \text{(iii)}$$

where c and d are independent of s, θ and ϕ. It is evident that Eq. (iii) satisfy Eqs. (i) and (ii), so that the great circle is a geodesic. Since the choice of the polar axis $\theta = 0$ is arbitrary, it follows that any great circle is a geodesic.

(b) Consider a circle on the sphere such that the plane of the circle does not pass through the centre of the sphere. Again, choose the direction $\theta = 0$ to be normal to the plane of the circle so that the parametric equation of the circle is

$$\theta = \theta_0; \quad \phi = ps + q; \quad \theta_0 \in (0, \pi) - \left\{\frac{\pi}{2}\right\}; \quad p \neq 0; \qquad \text{(iv)}$$

where p and q are independent of s, θ and ϕ. On substitution, it is seen that Eq. (ii) is satisfied but Eq. (i) is reduces to

$$p^2 \sin\theta_0 \cos\theta_0 = 0.$$

The conditions on θ_0 and p of Eq. (iv) show that this equation cannot be satisfied, proving that the circle is not a geodesic.

Now, we find the lines of shortest distance of the sphere. From Eq. (4.61), we have,

$$S = a \int_A^B F(x^i, \dot{x}^i)\,dt,$$

where

$$F(x^i, \dot{x}^i) = \sqrt{g_{ij}\dot{x}^i\dot{x}^j} = \sqrt{1 + \dot{\phi}^2 \sin^2\theta}.$$

In order ϕ is a geodesic, then Euler–Lagrange equation reduces to,

$$F_{\dot{\phi}} = \text{constant} \Rightarrow \frac{\dot{\phi}\sin^2\theta}{\sqrt{1 + \dot{\phi}^2 \sin^2\theta}} = c$$

$$\Rightarrow \dot{\phi}^2 = \frac{c^2}{\sin^4\theta - c^2\sin^2\theta} = \frac{c^2\operatorname{cosec}^4\theta}{(1 - c^2) - c^2\cot^2\theta}$$

$$\Rightarrow \dot{\phi} = \frac{k\operatorname{cosec}^2\theta}{\sqrt{1 - k^2\cot^2\theta}};\ k^2 = \frac{c^2}{1 - c^2}$$

$$\Rightarrow \phi = \alpha - \sin^{-1}(k\cot\theta); \ \text{i.e.}\ \sin(\alpha - \phi) = k\cot\theta.$$

Let us verify these curves are great circles. Multiplied last equation by sin , we get

$$(\ \sin\ \cos\ )\sin \qquad (\ \sin\ \sin\ )\cos\ = k\ \cos$$

or

$$x \sin \qquad y \cos\ = kz:$$

This represents a plane through the centre of the sphere, and the section of the sphere by this plane is a great circle.

4.5.1 Parallel Vector Fields

In an Euclidean space, a vector field $A_i(x^j)$ is said to be a field of parallel vectors if the components A_i are constants, i.e. $A_{i;j} = 0$; i.e. or, suppose that a vector A_i defined at a point x^j undergoes a displacement dx^j to the point $x^j + dx^j$, where, the vector becomes $A_i + \partial A_i$: We shall say that A_i undergoes a parallel displacement if $\frac{\partial A_i}{\partial x_j} = 0$: For example, a uniform electric field is a field of parallel vectors; the gravitational field of the earth in a sufficiently small region of space is a field of parallel vectors.

Here, our object is to generalise the concept of parallel displacement of surface vectors and tensors in Riemannian space, due to Levi Civita. Let us consider a space curve ; total of points, whose co-ordinates satisfy equation of the form

$$: \ x^i = x^i(t); \tag{4.65}$$

where x^i (belongs to the class C^1) are functions of a single parameter t. Let a vector A be localised at some point P of whose components A^i are functions of t. If we construct at every point of , a vector equal to A in magnitude and parallel to it in direction, we obtain what is known as a parallel field of vectors along the curve . We have shown in Chapter 3 that if A_i is a vector, its co-ordinate derivative $\frac{\partial A_i}{\partial x_j}$ is not a tensor, but the intrinsic derivative $\frac{A_i}{t}$ along the curve Eq. (4.65) is a tensor. To retain the tensor character, therefore, we must define the parallel displacement of a tensor in a Riemannian space differently. If A^i is a solution of the system of differential equation

$$\frac{A^i}{t} = \frac{dA^i}{dt} + \begin{matrix} i \\ j\ \ k \end{matrix}\ A^j \frac{dx^k}{ds} \tag{4.66}$$

then the vectors A^i are said to form a parallel vector field along the curve given by Eq. (4.65). We say that if the vector A is a parallel field along , then the vectors A do not change along the curve. Thus, the vector A suffers a parallel displacement along if the intrinsic derivative of the tensor along the curve vanishes, (Figure 4.1), i.e.

$$A^i_{;j}\frac{dx^i}{dt} = \frac{A^i}{t} = 0: \tag{4.67}$$

Conversely, we can show that every solution of system Eq. (4.67) yields a parallel vector field along Γ. Indeed, from the theory of differential equations it is known that

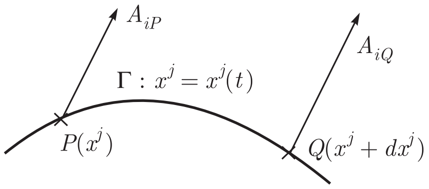

Figure 4.1: Parallel vectors A_{iP} and A_{iQ}.

this system of first order differential equations has a unique solution when the values of the components A^i are specified at a point of Γ. But the vector field formed by constructing a family of vectors of fixed lengths, parallel to a given vector, satisfies the system. Hence, every solution of Eq. (4.67) satisfies the initial conditions must form a parallel field along Γ.

In other words, we say that A^i is parallely displaced with respect to Riemannian space along the curve Γ, if Eq. (4.67) holds at all points of Γ. Thus, we have a field of parallel vectors, if,

$$A^i_{;j}\frac{dx^i}{dt} = 0 \Rightarrow g_{ik}\left(A^i_{;j}\frac{dx^j}{dt}\right) = 0 \Rightarrow \left(g_{ik}A^i\right)_{;j}\frac{dx^j}{dt} = 0$$

$$\Rightarrow A_{k;j}\frac{dx^j}{dt} = 0; \text{i.e. } A_{i;j}\frac{dx^j}{dt} = 0$$

or

$$\frac{dA^i}{dx^j} + \left\{\begin{matrix} i \\ j \end{matrix}\right\} A \frac{dx^j}{dt} = 0$$

or

$$\frac{dA^i}{dt} + \left\{\begin{matrix} i \\ j \end{matrix}\right\} A \frac{dx^j}{dt} = 0$$

or

$$dA^i = -\left\{\begin{matrix} i \\ j\ k \end{matrix}\right\} A^j dx^k \tag{4.68}$$

along the curve. The converse follows, as previously, for the existence and uniqueness of solution of such systems of differential equations. Similarly, the condition for a parallel displacement of a covariant vector A_i is

$$dA_i = \left\{\begin{matrix} j \\ i\ k \end{matrix}\right\} A_j dx^k. \tag{4.69}$$

Also, it follows from the observation that $A_{i;k} = g_i \ A_{;k}$ whenever $A_i = g_{ij} A^j$: Thus, the increments in A^i and A_i due to a displacement dx^j along the curve are given by Eqs. (4.68) and (4.69), respectively.

Result 4.5.1 Now, consider a point $P(x^i)$ and two displacements dx^i and $\ x^i$ from P giving rise to the points $Q(x^i + dx^i)$ and $R(x^i + \ x^i)$, respectively, as seen in Figure 4.2.

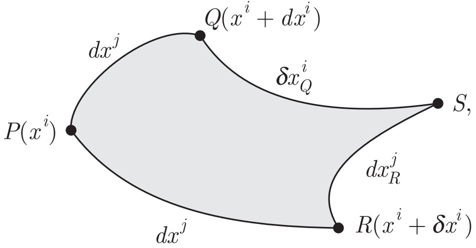

Figure 4.2: Generalised infinitesimal parallelogram.

Imagine the segment $PR = \ x^i$ being parallely displaced such that as the end point P goes to Q, the other end R goes to S. The displacement QS will be

$$QS \quad x^i_Q = \ x^i + d(\ x^i); \tag{4.70}$$

which makes the co-ordinates of the point S;

$$S \quad x^i + dx^i + \ x^i_Q \ = \ x^i + dx^i + \ x^i + d(\ x^i) \ ; \tag{4.71}$$

where by Eq. (4.68), we have

$$d(\ x^i) = \ \begin{matrix} i \\ j\ k \end{matrix}_P \ x_j \, dx^k : \tag{4.72}$$

The subscript P on the Christoffel symbol means is to be evaluated at P. Similarly, imagine the segment $PQ = dx^i$ being parallely displaced such that as the point P goes to R, the other end Q goes to S^0. The displacement RS^0 will be given by

$$RS = dx^i_R = \ x^i + d(\ x^i); \tag{4.73}$$

which makes the co-ordinates of the point S^0

$$S^0 \quad x^i + dx^i + \ x^i_R \ = \ x^i + \ x^i + dx^i + \ (dx^i) \ ; \tag{4.74}$$

$$\text{where,} \quad (dx^i) = \ \begin{matrix} i \\ j\ k \end{matrix}_P dx_j \ x^k = \ \begin{matrix} i \\ j\ k \end{matrix}_P dx_k \ x^j ; \tag{4.75}$$

which follows by interchanging j and k and using the symmetric property of the Christoffel symbol. Equations (4.72) and (4.75) show that $d(\ x^i) = \ (dx^i)$ indicating that the point S^0 coincides with S. The closed figure $PQRS$ can be regarded as the generalised infinitesimal parallelogram with opposite sides parallel to each other in the Riemannian space.

Result 4.5.2 (Curvature of Riemannian space V_N). Here we have introduced the curvature tensor by considering the parallel displacement of a vector along two different paths. Consider an infinitesimal parallelogram PQRS (Figure 4.3) as described with adjacent sides $PQ = dx^i$ and $PR = \delta x^i$. Imagine a constravariant vector A^i_p defined at the point P being parally displaced is the following two ways:

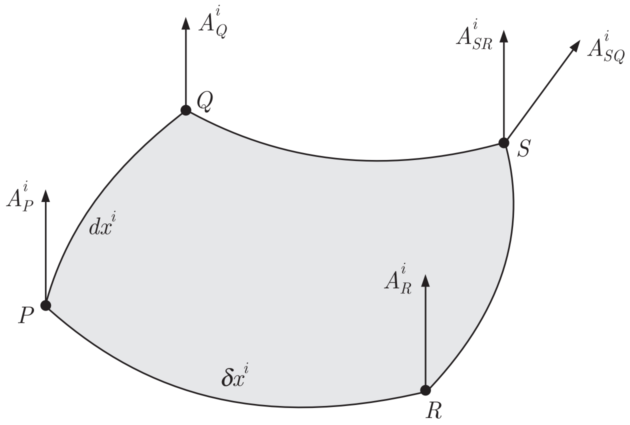

Figure 4.3: Curvature of Riemannian space.

(i) Displace A^i_p parallely from P to Q resulting A^i_Q and then displace A^i_Q parallely from Q to S giving the vector A^i_{SQ}.

(ii) Displace A^i_p parallely from P to R giving A^i_R parallely from R to S resulting the vector A^i_{SR}.

The vector A^i_Q obtained by a parallel displacement of A^i_P is given by

$$A^i_Q = A^i_P + dA^i_P; \quad \text{where } dA^i_P = -\left\{ {i \atop j\;k} \right\}_P A^j_P dx^k. \qquad \text{(i)}$$

This vector is parallely displaced from Q to S resulting the vector A^i_{SQ} which will be given by

$$A^i_{SQ} = A^i_Q + \delta A^i_Q = A^i_Q - \left\{ {i \atop l\;h} \right\}_Q A^l_Q\, \delta x^h. \qquad \text{(ii)}$$

The Christoffel symbol depends on the metric tensor which in turn is a function of co-ordinates. For a small displacement, we can therefore write

$$\left\{ {i \atop l\;h} \right\}_Q = \left\{ {i \atop l\;h} \right\}_P + \left\{ {i \atop l\;h} \right\}_{P;m} dx^m \qquad \text{(iii)}$$

where $\left\{ {i \atop l\ h} \right\}_{P;m} = \frac{\partial}{\partial x^m} \left\{ {i \atop l\ h} \right\}_{\text{at } P}$. Substituting Eqs. (iii) and (i) in Eq. (ii), we have

$$A^i_{SQ} = A^i_P - \left\{ {i \atop j\ k} \right\}_P A^j_P\, dx^k - \left[\left\{ {i \atop l\ h} \right\}_P + \left\{ {i \atop l\ h} \right\}_{P;m} dx^m \right]$$

$$\left(A^l_P - \left\{ {l \atop j\ k} \right\}_P A^j_P\, dx^k \right) \delta x^h$$

$$= A^i - \left\{ {i \atop j\ k} \right\} A^j\, dx^k - \left\{ {i \atop l\ h} \right\} A^l\, \delta x^h - \left\{ {i \atop l\ h} \right\}_{;m} A^l\, dx^m\, \delta x^h$$

$$+ \left\{ {i \atop l\ h} \right\} \left\{ {l \atop j\ k} \right\} A^j\, dx^k\, dx^h + O\left[(dx^k)^3 \right]; \qquad \text{(iv)}$$

where we have dropped the subscript P from the right hand side of (iv) because all the quantities are to be evaluated at P. The vector A^i_{SR} obtained by a parallel displacement along the path PRS will be simply given by an interchange of dx^i and δx^i in the expression for A^i_{SQ}:

$$A^i_{SR} = A^i - \left\{ {i \atop j\ k} \right\} A^j\, \delta x^k - \left\{ {i \atop l\ h} \right\} A^l dx^h - \left\{ {i \atop l\ h} \right\}_{;m} A^l\, \delta x^m\, dx^h$$

$$+ \left\{ {l \atop l\ h} \right\} \left\{ {l \atop j\ k} \right\} A^j\, \delta x^k\, dx^h + O\left[(dx^k)^3 \right]. \qquad \text{(v)}$$

Subtracting (iv) from (v) we get

$$A^i_{SR} - A^i_{SR} = \left\{ {i \atop l\ h} \right\}_{;m} A^l\, dx^m\, \delta x^h - \left\{ {i \atop l\ h} \right\}_{;m} A^l\, \delta x^m\, dx^h$$

$$+ \left\{ {i \atop l\ h} \right\} \left\{ {l \atop j\ k} \right\} A^j\, \delta x^k\, dx^h - \left\{ {i \atop l\ h} \right\} \left\{ {l \atop j\ k} \right\} A^j\, dx^k\, \delta x^h. \qquad \text{(vi)}$$

Replacing l by j and m by k in the first two terms on the right hand side of (vi) and interchanging k and h in the second and third terms, the Eq. (vi) can be written as

$$A^i_{SR} - A^i_{SQ} = R^i_{jhk}\, A^j\, dx^k\, \delta x^h; \qquad \text{(vii)}$$

where

$$R^i_{jkh} = \left\{ {i \atop j\ h} \right\}_{;k} - \left\{ {i \atop j\ k} \right\}_{;h} + \left\{ {i \atop l\ k} \right\} \left\{ {i \atop j\ h} \right\} - \left\{ {i \atop l\ h} \right\} \left\{ {l \atop j\ k} \right\} \qquad \text{(viii)}$$

Since A^j; dx^k and x^h are arbitrary vectors, if follows from quotient law that R^i_{jkh} is a tensor of rank four, known as the Riemann-Christoffel curvature tensor. It is independent of the vector A^i and depends only on the metric tensor and its first and second derivatives.

It follows that if a tensor is parallely displaced along a closed curve until we come back to the starting point, the resulting vector will not necessarily be the same as the original vector in a non-Euclidean space.

EXAMPLE 4.5.2 If A^i and B^i are two vectors of constant magnitudes and undergo parallel displacements along a curve then show that they are inclined at a constant angle.

Solution: Let the two vectors A^i and B^i be of constant magnitudes, so that their lengths do not change as we move along the curve . Let these vectors suffer parallel displacement along a curve , so by hypothesis

$$\frac{A^i}{t} = 0 = \frac{B^i}{t};$$

at each point of . The angle between the vectors A^i and B^i, is given by

$$\cos\ = g_{ij}A^iB^j:$$

The angle between the vectors A^i and B^i remains fixed as the parameter t is allowed to change. Here, $g_{ij}A^iB^j$ is an invariant. Differentiating intrinsically with respect to t, we get

$$\frac{}{t}(\cos\ ) = \frac{}{t}\ g_{ij}A^iB^j = \frac{d}{dt}\ g_{ij}A^iB^j$$

$$= g_{ij}\frac{A^i}{t}B^j + g_{ij}A^i\frac{B^j}{t} = g_{ij}\ 0\ B^j + g_{ij}A^i\ 0 = 0$$

or

$$\frac{d}{dt}(\cos\ ) = 0\ )\ \cos\ = \text{constant, i.e.}\ = \text{constant}:$$

Thus, we conclude that $g_{ij}A^iB^j$ is constant along . It follows directly from the result that, if $A^i = B^i$, then $g_{ij}A^iA^j = A^2$ is constant along and this implies = constant.

EXAMPLE 4.5.3 Prove that geodesics are autoparallel curves.

Solution: To prove the result, it is enough to prove that the unit tangent vector, say ^{i}, suffers a parallel displacement along a geodesic curve . Let $: x^i = x^i(t)$ be a geodesic in a Riemannian space. Then the co-ordinates of the points on the geodesic satisfy the geodesic equation,

$$\frac{d^2x^i}{dt^2} + \begin{matrix} i \\ j\ \ k \end{matrix} \quad \frac{dx^j}{dt}\frac{dx^k}{dt} = 0$$

or

$$\frac{d\ \ ^i}{dt} + \begin{matrix} i \\ j\ \ k \end{matrix} \quad {}^j \frac{dx^k}{dt} = 0$$

or

$$\frac{@\ \ ^i}{@x^k} + \begin{matrix} i \\ j\ \ k \end{matrix} \quad {}^j \ \frac{dx^k}{dt} = 0$$

or

$${}^i_{;k}\frac{dx^k}{dt} = 0; \ \text{i.e.} \ \ {}^i_{;j}\frac{dx^j}{dt} = 0;$$

at each point on . This result shows that the unit vectors form a system of parallel vectors along a geodesic. Hence, it follows from the definition of parallel displacement that the field of tangent vectors to a geodesic is a field of parallel vectors.

EXAMPLE 4.5.4 Any vector of constant magnitude which undergoes a parallel displacement along a geodesic is inclined at a constant angle to the curve.

Solution: Let a vector A^i of constant magnitude undergo a parallel displacement along a geodesic , so that,

$$\frac{A^{\ i}}{t} = A^i_{;j}\frac{dx^j}{dt} = 0$$

at each point of . Let $\ ^i$ be a unit tangent vector to the curve , so that,

$$\frac{\ \ ^i}{t} = \ {}^i_{;j}\frac{dx^j}{dt} = 0$$

at each point of , since geodesics are autoparallel curves. Let be the angle between the vector $\ ^i$ and a curve at any point of , i.e. is the angle between A^i and $\ ^i$, then,

$$A^i\ \ _i = A\ \ 1\ \ \cos$$

or

$$\frac{d}{dt}(A\cos\ \ ) = \frac{d}{dt}(A^i\ \ _i) = (\ A^i\ \ _i)_{;j}\frac{dx^i}{dt}$$

or

$$A\sin\ \ \frac{d}{dt} = A^i_{;j}\frac{dx^j}{dt}\ \ _i + A^i\ \ _{i;j}\frac{dx^j}{dt}$$

$$= 0t_1 + A^i 0 = 0:$$

or

$$\sin \frac{d}{dt} = 0; \quad \text{as } A \neq 0;$$

or

$$\sin \;= 0 \quad \text{or} \quad \frac{d}{dt} = 0 \Rightarrow \quad = 0 \quad \text{or} \quad = \text{constant}:$$

EXAMPLE 4.5.5 If the intrinsic derivative of a vector vanishes, then show that the magnitude of the vector is constant.

Solution: Let the intrinsic derivative of a vector A^i be zero so that

$$\frac{A^i}{t} = A^i_{;j}\frac{dx^j}{dt} = 0:$$

We have to show that $A^2 = A^iA_i$ is constant. Differentiating both sides with respect to a parameter t and noting that the intrinsic derivative of a scalar is ordinary derivative, we have

$$\frac{dA^2}{dt} = \frac{d}{dt}\;A^iA_i \;=\; A^iA_i\;_{;j}\frac{dx^j}{ds}$$

$$= A^i_{;j}\frac{dx^j}{dt}A_i + A^iA_{i;j}\frac{dx^j}{ds} = 0A_i + A^i0 = 0$$

or

$$A^2 = \text{constant} \Rightarrow A = \text{constant}:$$

Thus, if the derived vector of a given vector A in the direction of a given curve vanishes at all points of a curve, then the magnitude of the vector is constant. If A^i forms a field of parallel vectors along the curve $x^i = x(t)$, then we have

$$\frac{A^i}{t} = A^i_{;j}\frac{dx^j}{dt} = 0:$$

Then similarly, as above we can show that A = constant. Thus, the magnitude of a vector remain invariant under parallel displacement.

EXAMPLE 4.5.6 Show that the parallel displacement of a vector taken all around a circle on the surface of a sphere does not lead back to the same vector except when the circle is a great circle, that is, a geodesic.

Solution: Choose spherical polar co-ordinates $x^1 =$ and $x^2 =$ on the surface of the sphere. The non-vanishing Christoffel symbols are,

$$\genfrac{}{}{0pt}{}{2}{1\;2} = \cot \;= \genfrac{}{}{0pt}{}{2}{2\;1}; \quad \genfrac{}{}{0pt}{}{1}{2\;2} = \quad \sin \;\cos:$$

Choose a small circle normal to the polar axis so that its parametric equation is

$$\theta = \alpha; \quad \phi = ps + q; \quad p \neq 0 \text{ with } \frac{d\alpha}{dt} = 0.$$

Consider a vector $A^i = (A^1, A^2)$ at a point on the circle which may be taken to be $\phi = 0$. When the vector is displaced parallely along the circle, the change in its components is given, from Eq. (4.69), by

$$\frac{dA^1}{dt} = -\Gamma^1_{jk} A^j \frac{dx^k}{dt} = A^2 \sin\alpha \cos\alpha \frac{d\phi}{dt} \qquad \text{(i)}$$

$$\frac{dA^2}{dt} = -\Gamma^2_{jk} A^j \frac{dx^k}{dt} = -A^1 \cot\alpha \frac{d\phi}{dt}. \qquad \text{(ii)}$$

Now, for any function f of θ and ϕ through t, we have

$$\frac{df}{dt} = \frac{\partial f}{\partial \theta}\frac{d\theta}{dt} + \frac{\partial f}{\partial \phi}\frac{d\phi}{dt} = \frac{\partial f}{\partial \phi}\frac{d\phi}{dt},$$

which reduces Eqs. (i) and (ii) to

$$\frac{\partial A^1}{\partial \phi} = A^2 \sin\alpha \cos\alpha; \quad \frac{\partial A^2}{\partial \phi} = -A^1 \cot\alpha. \qquad \text{(iii)}$$

Differentiating once again with respect to ϕ and substituting one in the other, Eq. (iii) give rise to the decoupled equations

$$\frac{\partial^2 A^1}{\partial \phi^2} = -A^1 \cos^2\alpha; \quad \frac{\partial^2 A^2}{\partial \phi^2} = -A^2 \cos^2\alpha; \qquad \text{(iv)}$$

whose solution, consistent with Eq. (iii), is

$$A^1 = \sin\alpha\,[c\cos(\phi\cos\alpha) + d\sin(\phi\cos\alpha)]$$

$$A^2 = d\cos(\phi\cos\alpha) - c\sin(\phi\cos\alpha).$$

At $\phi = 0$, the components of the vector are given by

$$A^1(\phi = 0) = c\sin\alpha; \quad A^2(\phi = 0) = d; \qquad \text{(v)}$$

whereas, after parallel displacement once around the circle, the components become,

$$A^1(\phi = 2\pi) = \sin\alpha\,[c\cos(2\pi\cos\alpha) + d\sin(2\pi\cos\alpha)]$$

$$A^2(\phi = 2\pi) = d\cos(2\pi\cos\alpha) - c\sin(2\pi\cos\alpha);$$

which are not the same as those given by Eq. (v) except when $\alpha = \frac{\pi}{2}$; i.e. when the displacement is along the geodesic.

4.6 Exercises

1. Obtain the geodesic equations in cylindrical co-ordinates and spherical polar co-ordinates for the three-dimensional Euclidean space.
2. Consider a great circle on the surface of a sphere and choose the polar axis to lie in the plane of the great circle is a meridian. With this choice of co-ordinates, show that any meridian on a sphere is a geodesic.
3. Show that parallel displacement of a vector along any closed curve in an Euclidean space leads back to the original vector.
4. Prove that cyclic property of the covariant curvature tensor by setting a geodesic co-ordinate system.
5. Verify that, $R_{ijkl;m} + R_{ijlm;k} + R_{ijmk;l} = 0:$
6. In a two-dimensional Riemannian space, the components of the Ricci tensor are proportional to the components of the metric tensor.
7. Given the relation
$$A^{i}_{j} = R^{i}_{j} + {}^{i}_{j}(R + \quad)$$
where ; are constants. Show that for the value of $= \frac{1}{2}$, $A^{i}_{j;i} = 0:$
8. Let in a V_2 space the line element is $ds^2 = 2f(x^1;x^2)dx^1dx^2$. Find R_{1212} for the space. Prove that the space is flat, if
$$f\frac{@^2 f}{@x^1 @x^2} = \frac{@f}{@x^1}\frac{@f}{@x^2}:$$
Show also if f is independent of at least one of the two co-ordinates x^1 and x^2, then the space is flat.
9. Prove that a necessary and sufficient condition for a Riemannian V_N $(N > 2)$, to be of constant Riemannian curvature is that the Weyl tensor vanishes identically throughout $V_N:$
10. Prove that the mean curvature in the direction ${}_i$ at a point of a V_N is the sum of $N \quad 1$ Riemmanian curvatures along the direction pairs consisting of the direction and $N \quad 1$ other directions forming with this directions an orthogonal frame.
11. Prove that the sum of mean curvatures of a V_N for a mutually orthogonal directions at a point, is independent of the ennuple chosen. Obtain the value of this sum.
12. Prove that a Riemannian metric (g_{ij}) is the Euclidean metric if and only if the Riemannian curvature is zero at all points and the metric is positive definite.
13. (a) Calculate the Riemannian curvature for the metric
$$ds^2 = (\ dx^1)^2 \quad 2x^1(dx^2)^2:$$

(b) Show that the components R_{1212} of the curvature tensor for a two-dimensional space with metric $ds^2 = dx^2 + f(x, y)dy^2$ equals

$$\frac{1}{2}\frac{\partial^2 f}{\partial x^2} + \frac{1}{4f}\quad\frac{\partial f}{\partial x}\quad^{2}.$$

(c) A two-dimensional space have the metric

$$ds^2 = f(u, v)du^2 + h(u, v)dv^2.$$

Find the components R_{1212} of the curvature tensor for the space.

14. Show that the Riemannian curvature at any point (x^i) of Riemannian 3 space in the direction $P = (0, 1, 0)$ and $Q = (1, 1, 0)$ is $\frac{1}{4}(x^1)\;\;^{2}$, if the metric is given by

$$g_{11} = 1;\;\; g_{22} = 2x^1;\;\; g_{33} = 2x^2;\;\; g_{ij} = 0,\;\; \text{for } i \neq j.$$

15. Prove that, in a V_2,

$$R_{ijkl} = \quad \frac{R}{2}[g_{ik}g_{jl} \quad g_{il}g_{jk}].$$

16. Consider the metric $ds^2 = (dx^1)^2 \quad (x^1)^2[(dx^2)^2 + (dx^3)^2]$. Show that $R_{1212} = 2$ and that, therefore, the space is not flat.

17. Prove that a Riemannian space is an Einstein space if

$$R_{ijk} = \quad\;_j\, g_{jk} \quad\;_k\, g_{ij}\quad;$$

where is a constant. Hence, show that for an Einstein space, the scalar curvature R is constant.

18. Prove Schur's theorem. If all points in some neighbourhood in a Riemannian R^N are isotropic and $N \quad 3$, then is a constant throughout that neighbourhood.

19. Verify that $ds^2 = (dx^1)^2 + (x^1)^2(dx^2)^2$ represents the Euclidean metric in polar co-ordinates.

20. Determine whether the following metric is flat and/or Euclidean:

(a) $ds^2 = (dx^1)^2 \quad (x^1)^2(dx^2)^2, N = 2.$
(b) $ds^2 = (dx^1)^2 + (x^3)^2(dx^2)^2 + (dx^3)^2.$

21. Find the isotropic points for the Riemannian metric

$$ds^2 = (\log x^2)\quad (dx^1)^2 + (dx^2)^2 + (dx^3)^2\quad ;\;\; x^2 > 1;$$

and find the curvature at those points.

22. Show that R^3 under the metric

$$g_{11} = e^{x^2}; \quad g_{22} = 1; \quad g_{33} = e^{x^2}; \quad g_{ij} = 0; \text{ for } i \neq j$$

has constant Riemannian curvature with all points isotropic, find that curvature.

23. If ${}^{j}_{i}$ is the Einstein tensor, prove that

    (a) ${}^{j}_{i;j} = \left(R^{j}_{i} \quad \tfrac{1}{2}\ {}^{j}_{i} R \right)_{;j} = 0:$

    (b) The Einstein invariant $= \ {}^{i}_{i}$, vanishes if the space is flat.

24. For a two-dimensional space (V_2), show that

$$gR_{ij} = \quad g_{ij} R_{1212} \quad \text{and} \quad gR = \quad 2R_{1212}:$$

Hence, deduce that every V_2 is an Einstein space.

25. The metric of the V_2 formed by the surface of a sphere of radius r is

$$ds^2 = r^2 d\ ^2 + r^2 \sin^2\ d\ ^2$$

in spherical polar co-ordinates. Show that the surface of a sphere is a surface of constant curvature.

26. In an Euclidean V_4, prove that the hypersurface, $x^1 = c \sin u^1 \sin u^2 \sin u^3$, $x^2 = c \sin u^1 \sin u^2 \cos u^3$, $x^3 = c \sin u^1 \cos u^2$, $x^4 = c \cos u^1$ has a constant curvature $c\ ^2$:

27. Show that an Einstein space V_3 has constant curvature.

28. In a space in which the relation

$$R_{ij}\, g_{kl} \quad R_{il}\, g_{jk} + R_{jk}\, g_{il} \quad R_{kl}\, g_{ij} = 0$$

holds, show that the space is an Einstein space.

29. Find the lines of shortest distance of a sphere of radius a.

30. Show that the magnitude of a vector and angle between two vectors remain invariant under parallel displacement.

31. If two unit vectors are such that at all points of a given curve , their intrinsic derivatives along are zero, show that they are inclined at a constant angle.

32. Show that the field of tangent vectors to a geodesic is a field of parallel vectors.

Chapter 5

Geometry of Space Curve

For the study of a curve by the method of calculus, its parametric representation is covariant. In this chapter, three vectors (tangent, normal, binormal) of fundamental importance to curve theory are discussed. A curve is determined uniquely except for position by measures of two notions associated with it, called its *curvature* and *torsion*. For a given curve there is a set of important formulas dealing with its curvature and torsion. These formulas are called *Serret–Frenet formulae*, which will be established in the next theory.

5.1 Curve Theory

A curve C in E^3 is the image of a class C^3 mapping, r, from an interval of real numbers into E^3. The moving frame of the curve

$$C: r = r(s) = \left(x^1(s), x^2(s), x^3(s)\right) \tag{5.1}$$

with the arbitrary parameter s, has been defined as an image of the real number for the arc length parameterization. Let s be the arc length. Then,

$$\begin{aligned}
r' &= T = \frac{dr}{ds} = \left(x^{1\prime}(s), x^{2\prime}(s), x^{3\prime}(s)\right)\\
r'' &= \frac{d^2r}{ds^2} = \left(x^{1\prime\prime}(s), x^{2\prime\prime}(s), x^{3\prime\prime}(s)\right) = \frac{dT}{ds} = \kappa N\\
r''' &= \frac{d^3r}{ds^3} = \left(x^{1\prime\prime\prime}(s), x^{2\prime\prime\prime}(s), x^{3\prime\prime\prime}(s)\right)\\
&= \frac{d}{ds}(\kappa N) = \kappa' N + \kappa\frac{dN}{ds} = \kappa' N + \kappa(-\kappa T + \tau B).
\end{aligned}$$

We have at once for the unit tangent vector,

$$\hat{T} = \hat{r}' = \frac{1}{|r'|}r' = \frac{1}{|r'|}\left(x^{1\prime}(s), x^{2\prime}(s), x^{3\prime}(s)\right).$$

Since $r'' = \frac{dT}{ds} = \ \hat{N}$; where N is the normal vector, we have $|r''| = \ $; and so, the unit normal vector is given by

$$\hat{N} = \frac{1}{|r''|} r'' = \frac{1}{|r''|} \ \ x_1''(s); x_2''(s); x_3''(s) \ \ :$$

  is called the curvature. The unit vector $\hat{N}$ is perpendicular to $\hat{T}$ and in the plane of the tangents at P and a consecutive point. This plane, containing two consecutive tangents and three consecutive points at P, is called the plane of curvature or the osculating plane at P. If R is any point in this plane, then the equation of the osculating plane

$$[R \quad r; \hat{T}; \hat{N}] = 0; \text{i:e:} \ [R \quad r; r'; r''] = 0:$$

The unit vectors $\hat{T}$ and $\hat{N}$ are perpendicular to each other, and their plane is the plane of curvature. The straight line through P parallel to $\hat{N}$ is called the principal normal at P (Figure 5.1). If R be the current point on the line, its equation is

$$R = r + u\hat{N}:$$

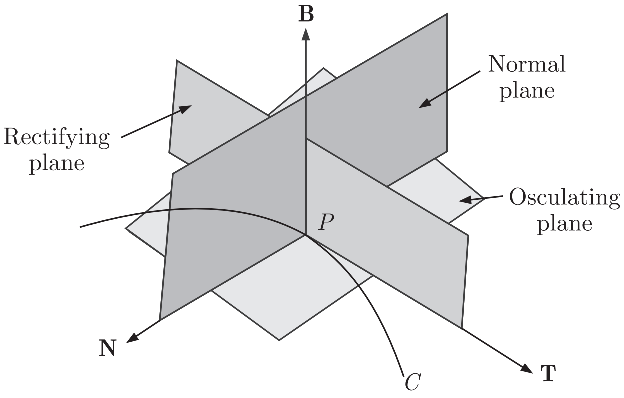

Figure 5.1: Planes.

The circle of curvature at P is the circle passing through three points on the curve ultimately coincident at P. Its radius P $= \frac{1}{\ }$ is called the radius of curvature, and its centre C the centre of curvature. This circle is clearly lies on the osculating plane at P, and its curvature is the same as that of the curve at P, for it has two consecutive tangents in common with the centre. The centre of curvature C lies on the principal normal, and

$$\overrightarrow{PC} = \ \hat{N} = \frac{1}{\ }\hat{N}$$

The direction cosines of $\hat{N}$ can be written as

$$\hat{N} = \frac{d\hat{T}}{ds} = \ r'':$$

Finally, for the binormal vector, we have

$$B = T \quad N = r^0 \quad \frac{1}{jr^{00}j} r^{00} = \frac{1}{jr^{00}j}\; r^0 \quad r^{00}:$$

The positive direction along the binormal is taken as that of B, just as the positive direction along the principal normal is that of $\hat{N}$. The equation of the binormal is

$$R = r + uB:$$

Since B is a vector of constant length, it follows that $\frac{dB}{ds}$ is perpendicular to B. Moreover differentiating the relation T.B = 0 with respect to s we find

$$kN \quad B + T \quad \frac{dB}{ds} = 0$$

) T $\frac{dB}{ds}$ = 0, as N is perpendicular to B. This equation shows that $\frac{dB}{ds}$ is perpendicular to T, which is perpendicular to B, and must therefore parallel to N. We may then write

$$\frac{dB}{ds} = \qquad N:$$

In this equation $\frac{dT}{ds}$ = N; the scalar measures the arc-rate of turning of the unit vector T, so here measures the arc-rate of turning of the unit vector B. This rate of turning of binormal is called the torsion of the curve at the point P. It is of course the rate of rotation of the osculating plane. The negative sign indicates that the torsion is regarded as positive when the rotation of the binormal as s increase is in the same sense as that of a right handed screw travelling in the direction of T. Now

$$\frac{dN}{ds} = \frac{d}{ds}(B \quad T) = \quad N \quad T + B \quad (\; N)$$

$$= \quad B \quad T:$$

The formulas for $\frac{dT}{ds}; \frac{dN}{ds}$ and $\frac{dB}{ds}$ gather together gives

$$\frac{dT}{ds} = \; N; \; \frac{dN}{ds} = \; B \quad T; \; \frac{dB}{ds} = \quad N:$$

These equations are the vector equivalents of the Serret{Frenet formulae for the derivatives of the direction cosines of the tangent, the principal normal and the binormal. Therefore

$$\frac{}{s}\begin{pmatrix} T \\ N \\ B \end{pmatrix} = \begin{pmatrix} 0 & & 0 \\ & 0 & \\ 0 & & 0 \end{pmatrix}\begin{pmatrix} T \\ N \\ B \end{pmatrix}$$

which is the Serret{Frenet formulae in the matrix form. The coefficient matrix is Skew-symmetric.

Now, we have to calculate $[r', r'', r''']$ as

$$r', r'', r''' = T;\ N;\ {}'N + (\ T + \ B)$$
$$= [T;\ N;\ B] = {}^2\ [T; N; B] = {}^2$$
$$) \qquad = \frac{1}{{}^2}\ r', r'', r''' = \frac{1}{|r''|^2}\ r', r'', r''':$$

An alternative formula giving the square of the torsion, may be deduced from the expression for r''' found above. On squaring this and diving throughout by ^{2} we obtain the result

$${}^2 = \frac{1}{{}^2}\, r'''^2 \qquad {}^2 \qquad \frac{'}{}^{2}:$$

By analogy with the relation that the radius of curvature is equal to the reciprocal of the curvature, it is customary to speak of the reciprocal of the torsion, and to denote it by . So $= \frac{1}{}$. But there is no circle of torsion or centre of torsion associated with the curve in the same way as the circle and centre of curvature.

Let the equation of the curve be

$$C : r = r(t) = x^1(t); x^2(t); x^3(t)\ ; \tag{5.2}$$

where t is a parameter. Therefore,

$$\dot r = \frac{dr}{dt} = \dot x^1(t); \dot x^2(t); \dot x^3(t) = \frac{dr}{ds}\frac{ds}{dt} = r'\dot s = \dot s T$$
$$r = \frac{d^2r}{dt^2} = x^1(t); x^2(t); x^3(t) = \frac{d}{dt}(\dot s T) = sT + \ \dot s^2 N$$
$$\dot r = \frac{d^3r}{dt^3} = \dot x^1(t); \dot x^2(t); \dot x^3(t) = \frac{d}{dt}\ sT + \ \dot s^2 N$$
$$= \frac{dT}{ds}\dot s s + T\dot s + \dot{\ }\dot s^2 N + (\ T + \ B)\,\dot s^3 + 2\ \dot s s N:$$

Since $T\ r = s$, so, $s = \frac{\dot r\ r}{|\dot r|}$: Therefore,

$$r = sT + \ \dot s^2 N = \frac{\dot r\ r}{|\dot r|}\frac{\dot r}{|\dot r|} + \ \dot s^2 N$$

or

$$N = \frac{1}{|\dot r|^2}\ r \qquad \frac{(r\ r)}{|\dot r|^2}\dot r$$

or

$$= \frac{1}{|\dot r|^2}\ |r| \qquad \frac{(r\ r)}{|\dot r|} = \frac{|\dot r\ \ r|}{|\dot r|^3}:$$

Therefore, the normal vector is given by

$$N = \frac{1}{\kappa|\dot{r}|^2}\left[\ddot{r} - \frac{(\dot{r}\cdot\ddot{r})}{|\dot{r}|^2}\dot{r}\right].$$

Now, we have to calculate $\left[\dot{r}, \ddot{r}, \dddot{r}\right]$ as

$$\left[\dot{r}, \ddot{r}, \dddot{r}\right] = \left[\dot{s}T, \kappa\dot{s}^2N, \kappa\tau\dot{s}^3B\right] = \kappa^2\tau\dot{s}^6[T, N, B] = \kappa^2\tau\dot{s}^6.$$

or

$$\tau = \frac{\left[\dot{r}, \ddot{r}, \dddot{r}\right]}{\kappa^2\dot{s}^6} = \frac{\left[\dot{r}, \ddot{r}, \dddot{r}\right]}{|\dot{r}\times\ddot{r}|^2}\frac{|\dot{r}|^6}{|\dot{r}|^6} = \frac{\left[\dot{r}, \ddot{r}, \dddot{r}\right]}{|\dot{r}\times\ddot{r}|^2},$$

which is the expression for torsion.

5.2 Plane Curve

When the curve $x^i = f^i(s)$ lies on the plane, then it is defined as a plane curve. Every planner curve has a principal normal vector. A necessary and sufficient condition for a curve $x^i = f^i(s)$ lies on a plane is

$$\begin{vmatrix} f^{1\prime} & f^{2\prime} & f^{3\prime} \\ f^{1\prime\prime} & f^{2\prime\prime} & f^{3\prime\prime} \\ f^{1\prime\prime\prime} & f^{2\prime\prime\prime} & f^{3\prime\prime\prime} \end{vmatrix} = 0.$$

Let the curve $x^i = f^i(s)$ lie on the plane

$$a_1x^1 + a_2x^2 + a_3x^3 + a_0 = 0$$

or

$$a_1f^1(s) + a_2f^2(s) + a_3f^3(s) + a_0 = 0.$$

Differentiate with respect to s thrice, we have

$$a_1f^{1\prime}(s) + a_2f^{2\prime}(s) + a_3f^{3\prime}(s) = 0$$

$$a_1f^{1\prime\prime}(s) + a_2f^{2\prime\prime}(s) + a_3f^{3\prime\prime}(s) = 0$$

$$a_1f^{1\prime\prime\prime}(s) + a_2f^{2\prime\prime\prime}(s) + a_3f^{3\prime\prime\prime}(s) = 0.$$

Eliminating a_1, a_2, a_3 from the above three equations, we get

$$\begin{vmatrix} f^{1\prime} & f^{2\prime} & f^{3\prime} \\ f^{1\prime\prime} & f^{2\prime\prime} & f^{3\prime\prime} \\ f^{1\prime\prime\prime} & f^{2\prime\prime\prime} & f^{3\prime\prime\prime} \end{vmatrix} = 0.$$

Therefore, the condition is necessary. To prove the sufficient condition, we have $= 0$. So the rank of may be $0, 1, 2$.

Case 1: Let rank $= 0$; i.e. all the elements of are zero, i.e.

$$f^{1'}(s) = f^{2'}(s) = f^{3'}(s) = 0$$

or

$$f^{1}(s) = f^{2}(s) = f^{3}(s) = \text{constant}$$

or

$$x^1(s) = k_1; x^2(s) = k_2; x^3(s) = k_3.$$

Thus, the curve $x^i = f^i(s)$ reduces to a point. Therefore, the given curve is a plane curve. (Trivial case).

Case 2: Let the rank of be 1, then any row will be proportional to any other row of , i.e.

$$\frac{f^{1''}}{f^{1'}} = \frac{f^{2''}}{f^{2'}} = \frac{f^{3''}}{f^{3'}}$$

or

$$\log f^{1'} = (s) + c_1 \Rightarrow f^{1'} = e^{(s)+c_1} = a_1 e^{(s)}$$

or

$$x^i = f^i(s) = a_i \int e^{(s)} ds + b_i = a_i (s) + b_i$$

or

$$\frac{x^1 - b_1}{a_1} = \frac{x^2 - b_2}{a_2} = \frac{x^3 - b_3}{a_3} = (s);$$

which is an equation of a line passing through b_i. Hence, the curves is a plane curve.

Case 3: Let the rank of be 2, then at least one of the minors of order two does not vanish. We assume that

$$b_1 = \begin{vmatrix} f^{2'} & f^{3'} \\ f^{2''} & f^{3''} \end{vmatrix}; b_2 = \begin{vmatrix} f^{3'} & f^{1'} \\ f^{3''} & f^{1''} \end{vmatrix}; b_3 = \begin{vmatrix} f^{1'} & f^{2'} \\ f^{1''} & f^{2''} \end{vmatrix}.$$

So, there exists three numbers $b_1; b_2; b_3$ such that

$$b_1 f^{1'}(s) + b_2 f^{2'}(s) + b_3 f^{3'}(s) = 0 \quad \text{(i)}$$

$$b_1 f^{1''}(s) + b_2 f^{2''}(s) + b_3 f^{3''}(s) = 0 \quad \text{(ii)}$$

$$b_1 f^{1'''}(s) + b_2 f^{2'''}(s) + b_3 f^{3'''}(s) = 0. \quad \text{(iii)}$$

Differentiating with respect to s, we get from Eq. (i),

$$b_i f^{i''}(s) + b_i' f^{i'}(s) = 0 \Rightarrow b_i' f^{i'}(s) = 0;\ \text{using Eq. (ii)}.$$

Similarly, differentiate Eq. (ii) with respect to s and using Eq. (iii), we get, $b_i' f^{i''}(s) = 0$. So, there exists a unique solution to the linear equations

$$b_1' f^{1'}(s) + b_2' f^{2'}(s) + b_3' f^{3'}(s) = 0$$

$$b_1' f^{1''}(s) + b_2' f^{2''}(s) + b_3' f^{3''}(s) = 0.$$

Solving, we get

$$\frac{b_1'}{\begin{vmatrix} f^{2'} & f^{3'} \\ f^{2''} & f^{3''} \end{vmatrix}} = \frac{b_2'}{\begin{vmatrix} f^{3'} & f^{1'} \\ f^{3''} & f^{1''} \end{vmatrix}} = \frac{b_3'}{\begin{vmatrix} f^{1'} & f^{2'} \\ f^{1''} & f^{2''} \end{vmatrix}}$$

$$\Rightarrow \frac{b_1'}{b_1} = \frac{b_2'}{b_2} = \frac{b_3'}{b_3} = \lambda'(s);\ \text{say},$$

or

$$\log b_i = \lambda(s) + c_i \Rightarrow b_i = e^{\lambda(s)+c_i} = a_i e^{\lambda(s)}.$$

But $b_i f^{i'} = 0$; from Eq. (i) replacing b_i, we get

$$a_i e^{\lambda(s)} f^{i'} = 0 \Rightarrow a_i f^{i'} = 0;\ \text{as } e^{\lambda(s)} \neq 0$$

or

$$a_1 f^{1'} + a_2 f^{2'} + a_3 f^{3'} = 0$$

or

$$a_1 f^1 + a_2 f^2 + a_3 f^3 + a_0 = 0.$$

Hence, the curve is a plane curve. When the curve is not plane, then it is called skew curve.

EXAMPLE 5.2.1 Determine $\phi(t)$ such that the curve $x^1 = a\cos t;\ x^2 = a\sin t;\ x^3 = \phi(t)$ is a plane curve and what is the nature of the conic?

Solution: Since the given curve is a plane curve, we have

$$\begin{vmatrix} x^{1'} & x^{2'} & x^{3'} \\ x^{1''} & x^{2''} & x^{3''} \\ x^{1'''} & x^{2'''} & x^{3'''} \end{vmatrix} = \begin{vmatrix} -a\sin t & a\cos t & \phi' \\ -a\cos t & -a\sin t & \phi'' \\ a\sin t & -a\cos t & \phi''' \end{vmatrix} = 0$$

or

$$\phi''' + \phi' = 0 \Rightarrow \phi(s) = a^1 + b\cos t + c\sin t.$$

The nature of the curve will be intersected between the cylinder $x^2 + y^2 = a^2$ and the plane $lx + my + nz = k$.

5.3 Twisted Curve

A curve in E^3 which does not lie on a plane is called *twisted* curve. Thus, the twisted curve is not a plane curve.

EXAMPLE 5.3.1 Find the curvature and the torsion of the circular helix

$$r = \quad a\cos\frac{s}{c}; a\sin\frac{s}{c}; \frac{bs}{c} \quad ; \; c = \sqrt{a^2+b^2};$$

where s is arc length.

Solution: By differentiation with respect to the arc length, the tangent vector is given by

$$T = r' = \quad \frac{a}{c}\sin\frac{s}{c}; \frac{a}{c}\cos\frac{s}{c}; \frac{s}{c} \quad :$$

The normal vector N is given by

$$N = r'' = \quad \cos\frac{s}{c}; \quad \sin\frac{s}{c}; 0 \quad \text{as } c = \sqrt{a^2+b^2}:$$

Finally, the binormal vector is given by

$$B = T \quad N = \begin{array}{ccc} i & j & k \\ \frac{a}{c}\sin\frac{s}{c} & \frac{a}{c}\cos\frac{s}{c} & \frac{s}{c} \\ \cos\frac{s}{c} & \sin\frac{s}{c} & 0 \end{array} = \quad \frac{b}{c}\sin\frac{s}{c}; \quad \frac{b}{c}\cos\frac{s}{c}; \frac{a}{c} \quad :$$

Therefore, the curvature and the torsion of the circular helix are given by,

$$= \frac{a}{c^2}\cos^2\frac{s}{c} + \frac{a}{c^2}\sin^2\frac{s}{c} + 0 = \frac{a}{c^2}$$

$$= \frac{b}{c^2}\cos^2\frac{s}{c} + \frac{b}{c^2}\sin^2\frac{s}{c} + 0 = \frac{b}{c^2}:$$

Thus the curvature and the torsion are both constant, and therefore their ratio is constant. The principal normal intersects the axis of the cylinder orthogonally and the tangent and binormal are inclined at constant angles to the fixed direction of the generators.

5.4 Space Curve Theory

In this section we describe the geometry of the space curve. Formulae analogous to those of Frenet also occur in the theory of surfaces. This section contains a set of three remarkable formulas, generally known as *Serret–Frenet formulae*, which characterise, in the small, all essential geometric properties of space curves.

5.4.1 Serret–Frenet Formulae

In this section we will investigate the theory of twisted curves. Let the curve Γ be given by the equation

$$\Gamma : \; x^i = x^i(s); \tag{5.3}$$

where the parameter s measures the arc distance along the curve. As arc length s along a curve remains invariant, we consider s as the parameter to study the geometry of the space curves. The square of the length of the elements of arc of Γ is given by

$$ds^2 = g_{ij}\, dx^i dx^j$$

or

$$1 = g_{ij} \frac{dx^i}{ds}\frac{dx^j}{ds}; \tag{5.4}$$

showing that $\frac{dx^i}{ds}$ is a unit vector. Let P be a given point with co-ordinates (x^i) and Q be a neighbouring point with co-ordinates $(x^i + dx^i)$ on Γ, corresponding to an increment ds in the arc (Figure 5.2). Then the vector $\lim_{Q\to P} \frac{\vec{PQ}}{ds}$ is called the tangent vector and we shall denote it by λ^i: Thus, Eq. (5.4) can be written as

$$1 = g_{ij}\lambda^i\lambda^j; \quad \text{where } \lambda^i = \frac{dx^i}{ds}: \tag{5.5}$$

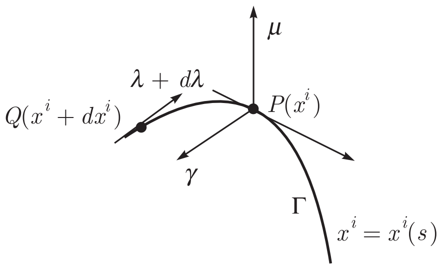

Figure 5.2: Tangent, normal, binormal.

These are precisely the direction cosines of the tangent vector to the curve Γ. We shall assume that curve Γ is of class C^2, so that it has continuously turning tangent at all points of Γ. Taking the intrinsic derivative of Eq. (5.5) with respect to s, we have

$$g_{ij} \frac{\delta\lambda^i}{\delta s}\lambda^j + g_{ij}\lambda^i\frac{\delta\lambda^j}{\delta s} = 0$$

or

$$g_{ij}\;{}^{i}\,\frac{{}^{j}}{s} = 0:$$

Thus, either $\frac{j}{s}$ vanishes or is orthogonal to the tangent vector at some point of a curve, is said to be anormal vector of that curve at that point. Hence, the condition that a vector ^{i} be normal to at P is

$$g_{ij}\;{}^{i}\;{}^{j} = 0: \tag{5.6}$$

Thus, if $\frac{i}{s}$ does not vanish, we denote the unit vector codirectional with $\frac{j}{s}$ by ^{j} and write it as

$${}^{j} = \frac{1}{\ }\frac{{}^{j}}{s}; \tag{5.7}$$

where > 0 is so chosen to make ^{j} a unit vector. The vector ^{i}, determined by Eq. (5.7) is called the principal normal vector to the curve at P and is called the curvature of the curve at that point P. The plane determined by the tangent and the normal is called the osculating plane to the curve at P.

Since is a unit vector, so the quadratic relation

$$g_{ij}\;{}^{i}\;{}^{j} = 1 \tag{5.8}$$

is satisfied. Differentiating Eq. (5.8) intrinsically, we get

$$g_{ij}\,\frac{{}^{i}}{s}\,{}^{j} + g_{ij}\,{}^{i}\,\frac{{}^{j}}{s} = 0 \quad) \quad g_{ij}\,{}^{i}\,\frac{{}^{j}}{s} = 0: \tag{5.9}$$

Taking the intrinsic derivative of the orthogonal relation Eq. (5.5) we get

$$g_{ij}\,\frac{{}^{i}}{s}\,{}^{j} + g_{ij}\,{}^{i}\,\frac{{}^{j}}{s} = 0$$

or

$$\begin{aligned} g_{ij}\,{}^{i}\,\frac{{}^{j}}{s} &= \quad g_{ij}\,\frac{{}^{i}}{s}\,{}^{j} = \quad g_{ij}\;{}^{i}\;{}^{j};\quad \text{from Eq. } (5{:}7)\\ &= \quad ;\quad \text{from Eq. } (5{:}8)\\ &= \quad g_{ij}\;{}^{i}\;{}^{j};\quad \text{from Eq. } (5{:}5) \end{aligned}$$

or

$$g_{ij}\;{}^{i}\quad \frac{{}^{j}}{s} + \quad {}^{j} \quad = 0: \tag{5.10}$$

Also, from Eq. (5.6) we can write $g_{ij}\ {}^{i}\ {}^{j} = 0$; as $g_{ij} = g_{ji}$; i.e.

$$g_{ij}\ {}^{i}\ {}^{j} = 0$$

or

$$g_{ij}\ {}^{i}\frac{{}^{j}}{s} + g_{ij}\ {}^{i}\ {}^{j} = 0$$

or

$$g_{ij}\ {}^{i}\ \frac{{}^{j}}{s} + \ {}^{j} = 0: \qquad (5.11)$$

From Eqs. (5.10) and (5.11) we see that, the vector $\frac{{}^{j}}{s} + \ {}^{j}$ is orthogonal to both the tangent vector ^{i} and the principal normal vector ^{i}: Hence, we choose a vector ^{i} such that

$${}^{i} = \frac{1}{}\ \frac{{}^{i}}{s} + \ {}^{i} \qquad (5.12)$$

and is chosen to make , a unit vector. The constant thus introduced is called the torsion of the curve. Note that, it may be positive or negative in contrast to the curvature which is essentially positive, being the magnitude of a vector.

The sign of is not always positive but we agree to choose the sign of in such a way that (; ;) form a right-hand system, i.e.

$$\varepsilon_{ijk}\ {}^{i}\ {}^{j}\ {}^{k} = +1; \qquad (5.13)$$

where

$$\varepsilon_{ijk} = \sqrt{g}e_{ijk}; \quad \text{and } g = \left|\frac{\partial y^{i}}{\partial x}\right|^{2}; \qquad (5.14)$$

so that the left-hand member of Eq. (5.13) is an invariant. The vector is called the binormal of at P. Also, using Eqs. (5.13) and (5.14), we get

$$\sqrt{g}e_{ijk}\ {}^{i}\ {}^{j}\ {}^{k} = 1$$

or

$$e^{pqr}\sqrt{g}e_{ijk}\ {}^{i}\ {}^{j}\ {}^{k} = e^{pqr}$$

or

$${}^{p}_{i}{}^{q}_{j}{}^{r}_{k}\sqrt{g}\ {}^{i}\ {}^{j}\ {}^{k} = e^{pqr}$$

or

$${}^{p}\ {}^{q}\ {}^{r} = \frac{1}{\sqrt{g}}e^{pqr} = \frac{1}{\sqrt{g}}e^{rpq}$$

or

$$g_{ip}\,{}^{i}g_{jq}\,{}^{j}\;{}^{p}\;{}^{q}\;{}^{r} = g_{ip}\,{}^{i}g_{jq}\,{}^{j}\,\varepsilon^{pqr}; \quad \text{by Eq. } (5{:}14)$$

or

$${}^{r} = \varepsilon^{pqr}\;{}_{p}\;{}_{q} \text{ i.e. } {}^{i} = \varepsilon^{ijk}\;{}_{j}\;{}_{k}. \qquad (5.15)$$

Differentiating Eq. (5.15) intrinsically, we get

$$\frac{{}^{i}}{s} = \varepsilon^{ijk}\frac{{}^{j}}{s}\;{}_{k} + \varepsilon^{ijk}\;{}_{j}\frac{{}^{k}}{s}$$
$$= \varepsilon^{ijk}\;{}_{j}\;{}_{k} + \varepsilon^{ijk}\;{}_{i}[\;{}_{k}\quad{}_{k}];$$

which reduces on account of skew symmetry of ε^{ijk} to

$$\frac{{}^{i}}{s} = \varepsilon^{ijk}\;{}_{j}\;{}_{k}.$$

But ${}_{i}$; ${}_{i}$ and ${}_{i}$ form a right-handed system of unit vectors, so $\varepsilon^{ijk}\;{}_{j}\;{}_{k} = {}^{i}$. Therefore,

$$\frac{{}^{i}}{s} = {}^{i}. \qquad (5.16)$$

Equations (5.7), (5.12) and (5.16) are called the *Frenet formulae* for space curve. On account of their importance in the theory of curves, we group these formulae together for convenient reference in the form

$$\left.\begin{array}{l} \frac{{}^{i}}{s} = {}^{i} \\ \frac{{}^{i}}{s} = {}^{i} + {}^{i} \\ \frac{{}^{i}}{s} = {}^{i} \end{array}\right\}. \qquad (5.17)$$

Equation (5.17) can be written in the matrix form as

$$\frac{}{s}\begin{pmatrix} {}^{i} \\ {}^{i} \\ {}^{i} \end{pmatrix} = \begin{pmatrix} 0 & & 0 \\ & 0 & \\ 0 & & 0 \end{pmatrix}\begin{pmatrix} {}^{i} \\ {}^{i} \\ {}^{i} \end{pmatrix}.$$

The coefficient matrix is skew symmetric. Formulae analogous to those of Frenet also occur in the theory of surfaces.

Note 5.4.1 If the curve lies on a plane, then it is called a *plane curve*, otherwise it is called a *skew curve*.

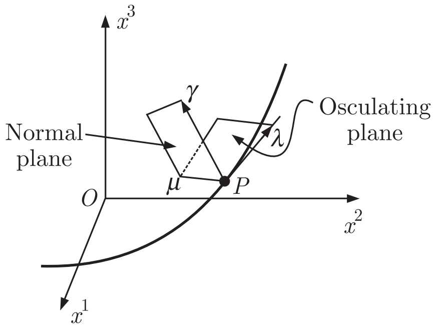

Figure 5.3: Osculating, normal rectifying plane.

Note 5.4.2 The plane determined by the tangent and principal normal is called the osculating plane at a point $P(x^i_0)$ of a space curve (Figure 5.3). It's equation is

$$g_{ij}\ \ x^i\ \ \ x^i_0\ \ \ {}^j = 0;$$

where x^i is any point on the plane. The plane determined by principal normal and binormal is called the normal plane at $P(x^i_0)$ and its equation is given by

$$g_{ij}\ \ x^i\ \ \ x^i_0\ \ \ {}^j = 0:$$

The plane determined by binormal and tangent is called the rectifying plane at $P(x^i_0)$ and its equation is given by (Figure 5.3)

$$g_{ij}\ \ x^i\ \ \ x^i_0\ \ \ {}^j = 0:$$

EXAMPLE 5.4.1 Show that

(i) $= g_{mn}\,\frac{\ \ m}{s}\,\frac{\ \ n}{s}\ {}^{1=2}$

(ii) $= "_{rst}\ \ {}^r\ \ {}^s\frac{\ \ t}{s}$

(iii) ${}_i = \frac{1}{\ }\ \ \frac{\ \ i}{s} + \ \ {}_i$

(iv) ${}_i = "_{ijk}\ \ {}^i\frac{\ \ k}{s}$

(v) ${}_i = \frac{1}{\ }\frac{\ \ i}{s};$

where the symbols have their usual meanings.

Solution:

(i) We know that

$$^{m} = \frac{1}{}\,\frac{{}^{m}}{s}; \quad \frac{{}^{m}}{s} = \quad {}^{m};$$

where > 0 is so chosen to make ^{m} a unit vector. Thus,

$$g_{mn}\frac{{}^{m}}{s}\,{}^{n} = g_{mn}\,{}^{m}\,{}^{n}$$

or

$$g_{mn}\frac{{}^{m}}{s}\,{}^{n} = \quad \text{as}\, g_{mn}\,{}^{m}\,{}^{n} = 1$$

or

$$= g_{mn}\frac{{}^{m}}{s}\frac{1}{}\frac{{}^{n}}{s}$$

or

$$^{2} = g_{mn}\frac{{}^{m}}{s}\frac{{}^{n}}{s}$$

or

$$= \quad g_{mn}\frac{{}^{m}}{s}\frac{{}^{n}}{s}\quad^{1=2}; \quad \text{as} > 0:$$

(ii) From the second Frenet formula we have

$$\frac{{}^{t}}{s} = \quad {}^{t} + \quad {}^{t}$$

or

$$"_{rst}\,{}^{r}\,{}^{s}\frac{{}^{t}}{s} = "_{rst}\,{}^{r}\,{}^{s}\quad{}^{t} + \quad{}^{t}$$

$$= "_{rst}\,{}^{r}\,{}^{s}\,{}^{t} + "_{rst}\,{}^{r}\,{}^{s}\,{}^{t}$$

$$= 0 + \quad {}^{t}\,{}_{t}; \ \text{as}\, "_{rst}\,{}^{r}\,{}^{t} = 0$$

$$= \quad; \ \text{as}\ {}^{t}\,{}_{t} = 1;$$

which is the required result.

(iii) Follows from Eq. (5.12).

(iv) Using relation Eq. (5.17), we get $\frac{{}^{k}}{s} = \quad {}^{k}$, i.e.

$$"_{ijk}\,{}^{j}\frac{{}^{k}}{s} = "_{ijk}\,{}^{j}\quad{}^{k} = "_{ijk}\,{}^{j}\,{}^{k} = \quad{}^{i}; \ \text{from Eq. (5:15):}$$

(v) We see that

$$\frac{1}{}\frac{{}^{i}}{s} = \frac{1}{}\frac{}{s}\; g_{ij} \;{}^{j} = \frac{1}{}g_{ij}\frac{{}^{j}}{s} = \frac{1}{}g_{ij} \;{}^{j} = \;{}_{i}:$$

EXAMPLE 5.4.2 Show that

$$g_{ij}\frac{{}^{i}}{s}\frac{{}^{j}}{s} = \;{}^{2} + \;{}^{2}:$$

Solution: Using Eq. (5.17), we get

$$g_{ij}\frac{{}^{i}}{s}\frac{{}^{j}}{s} = g_{ij} \quad {}^{i} + \quad {}^{i} \qquad {}^{j} + \quad {}^{j}$$

$$= g_{ij} \quad {}^{2}\; {}^{i}\; {}^{j} \qquad {}^{i}\; {}^{j} \qquad {}^{j}\; {}^{i} + \;{}^{2}\; {}^{i}\; {}^{j}$$

$$= \;{}^{2}g_{ij} \; {}^{i}\; {}^{j} \quad g_{ij} \; {}^{i}\; {}^{j} \quad g_{ij} \; {}^{j}\; {}^{i} + \;{}^{2}g_{ij} \; {}^{i}\; {}^{j}$$

$$= \;{}^{2}\; 1 \qquad 0 \qquad 0 + \;{}^{2}\; 1 = \;{}^{2} + \;{}^{2}:$$

EXAMPLE 5.4.3 Show that

(i) $\frac{{}^{2}\;{}^{i}}{s^{2}} = \frac{d}{ds}\;{}^{i} + \qquad {}^{i} \qquad {}^{i}$

(ii) $\frac{{}^{2}\;{}^{i}}{s^{2}} = \frac{d}{ds}\;{}^{i} \quad (\;{}^{2} + \;{}^{2})\;{}^{i} \quad \frac{d}{ds}\;{}^{i}$

(iii) $\frac{{}^{2}\;{}^{r}}{s^{2}} = (\quad {}^{r} \qquad {}^{r}) \quad \frac{d}{ds}\;{}^{r};$

where the symbols have their usual meanings.

Solution:

(i) Using Eq. (5.7), we get, $\frac{{}^{i}}{s} = \;{}^{i}$; where > 0: Differentiating intrinsically with respect to the parameter s we get

$$\frac{{}^{2}\;{}^{i}}{s^{2}} = \frac{}{s} \quad {}^{i} = \frac{d}{ds}\;{}^{i} + \quad {}^{j} \quad \begin{matrix} i \\ j\;\; k \end{matrix} \quad \frac{dx^{k}}{ds}$$

$$= \frac{d}{ds}\;{}^{i} + \frac{d\;{}^{i}}{ds} + \quad {}^{j} \quad \begin{matrix} i \\ j\;\; k \end{matrix} \quad \frac{dx^{k}}{ds}$$

$$= \frac{d}{ds}\;{}^{i} + \quad \frac{d\;{}^{i}}{ds} + \;{}^{j} \quad \begin{matrix} i \\ j\;\; k \end{matrix} \quad \frac{dx^{k}}{ds}$$

$$= \frac{d}{ds}\;{}^{i} + \frac{{}^{i}}{s} = \frac{d}{ds}\;{}^{i} + \qquad {}^{i} + \quad {}^{i} \quad ; \text{ from Eq. } (5{:}17)$$

$$\frac{{}^{2}\;{}^{i}}{s^{2}} = \frac{d}{ds}\;{}^{i} + \qquad {}^{i} \qquad {}^{i} \;;$$

which is the result as required.

(ii) Using Eq. (5.17), we get $\frac{\ {}^{i}}{s} = \ {}^{i} \quad \ {}^{i}$:

$$\frac{{}^{2}\ {}^{i}}{s^2} = \frac{}{s}\ {}^{i} \qquad \frac{}{s}\ {}^{i}$$

$$= \frac{d}{ds}\ {}^{i} + \ {}^{j} \begin{matrix} i \\ j\ k \end{matrix} \frac{dx^k}{ds} \qquad \frac{d}{ds}\ {}^{i} + \ {}^{j} \begin{matrix} i \\ j\ k \end{matrix} \frac{dx^k}{ds}$$

$$= \frac{d}{ds}\ {}^{i} + \frac{d\ {}^{i}}{ds} + \ {}^{j} \begin{matrix} i \\ j\ k \end{matrix} \frac{dx^k}{ds} \qquad \frac{d}{ds}\ {}^{i} \qquad \frac{d\ {}^{i}}{ds} + \ {}^{j} \begin{matrix} i \\ j\ k \end{matrix} \frac{dx^k}{ds}$$

$$= \frac{d}{ds}\ {}^{i} + \frac{\ {}^{i}}{s}\ \frac{d}{ds}\ {}^{i} \qquad \frac{\ {}^{i}}{s}$$

$$= \frac{d}{ds}\ {}^{i} + (\quad {}^{i})\ \frac{d}{ds}\ {}^{i} \qquad {}^{i} = \frac{d}{ds}\ {}^{i} \quad (\ {}^{2} + \ {}^{2})\ {}^{i} \quad \frac{d}{ds}\ {}^{i}:$$

(iii) Using Eq. (5.17), we get $\frac{\ {}^{i}}{s} = \quad {}^{i}$: Differentiating intrinsically with respect to the parameter s we get

$$\frac{{}^{2}\ {}^{i}}{s^2} = \frac{}{s}\ {}^{i} = \frac{d}{ds}\ {}^{i} \qquad {}^{j} \begin{matrix} i \\ j\ k \end{matrix} \frac{dx^k}{ds}$$

$$= \frac{d}{ds}\ {}^{i} \qquad \frac{d\ {}^{i}}{ds} + \ {}^{j} \begin{matrix} i \\ j\ k \end{matrix} \frac{dx^k}{ds}$$

$$= \frac{d}{ds}\ {}^{i} \qquad \frac{\ {}^{i}}{s}$$

$$= \frac{d}{ds}\ {}^{i} \qquad {}^{i} + \quad {}^{i}\ ; \text{ from Eq. } (5{:}17)$$

$$= \quad {}^{i} \quad {}^{i} \quad \frac{d}{ds}\ {}^{i}:$$

EXAMPLE 5.4.4 Show that

$$= \frac{1}{2}"_{ij\ k}\ {}^{i} \frac{\ {}^{i}}{s} \frac{{}^{2}\ {}^{k}}{s^2};$$

where the notations have their usual meaning.

Solution: Using Eq. (5.7), we get $\frac{\ {}^{i}}{s} = \quad {}^{i}$; where, $\ > \ 0$: Differentiating intrinsically with respect to the parameter s we get,

$$\frac{{}^{2}\ {}^{i}}{s^2} = \frac{}{s}\ {}^{i} + \frac{\ {}^{i}}{s} = \frac{}{s}\ {}^{i} + \qquad {}^{i} \qquad {}^{i}\ :$$

Using the Frenet formula [Eq. (5.17)], we get,

$$\frac{1}{2}"_{ijk}\ {}^{i}\frac{{}^{i}}{s}\frac{{}^{2\ \ k}}{s^{2}}$$

$$= \frac{1}{2}"_{ijk}\ {}^{i}\ \ {}^{j}\ \ \frac{}{s}\ {}^{k} + \ \ {}^{k}\ \ \ {}^{k}$$

$$= \frac{1}{}\ \ \frac{}{s}"_{ijk}\ {}^{i}\ {}^{j}\ {}^{k} + \ "_{ijk}\ {}^{i}\ {}^{j}\ {}^{k}\ \ \ {}^{2}"_{ijk}\ {}^{i}\ {}^{j}\ {}^{k}$$

$$= \frac{1}{}(0 + \ \ \ 0) = \ :$$

EXAMPLE 5.4.5 Find the curvature and torsion at any point of a given curve (circle) , given by

$$:\ x^{1} = a; x^{2} = t; x^{3} = 0;$$

where $ds^{2} = \ dx^{1}\ {}^{2} + (x^{1})^{2}\ dx^{2}\ {}^{2} + \ dx^{3}\ {}^{2}:$

Solution: The square of the element of arc in cylindrical co-ordinates as

$$ds^{2} = \ dx^{1}\ {}^{2} + (x^{1})^{2}\ dx^{2}\ {}^{2} + \ dx^{3}\ {}^{2}:$$

For the given metric, the non-vanishing Christoffel symbols are

$$\begin{matrix}1\\2\ 2\end{matrix} = \ x^{1};\quad \begin{matrix}2\\1\ 2\end{matrix} = \begin{matrix}2\\2\ 1\end{matrix} = \frac{1}{x^{1}};$$

and $g_{11} = 1 = g_{33}$ and $g_{22} = (x^{1})^{2} = a^{2}$: The tangent vector ^{i} is given by

$${}^{i} = \frac{dx^{i}}{ds} = \ \frac{dx^{1}}{ds};\frac{dx^{2}}{ds};\frac{dx^{3}}{ds}\ = \ 0;\frac{dt}{ds};0\ :$$

As is a unit vector, so $g_{ij}\ {}^{i}\ {}^{j} = 1$; i.e.

$$g_{11}\ {}^{1}\ {}^{1} + g_{22}\ {}^{2}\ {}^{2} + g_{33}\ {}^{3}\ {}^{3} = 1$$

or

$$a^{2}\ \frac{dt}{ds}\ {}^{2} = 1\ )\ \frac{dt}{ds} = \frac{1}{a}:$$

Therefore

$${}^{i} = \ 0;\frac{dt}{ds};0\ = \ 0;\frac{1}{a};0\ :$$

Now, from the Serret{Frenet formula [Eq. (5.7)], i.e. $^i = \frac{^i}{s}$; we get,

$$^1 = \frac{^1}{s} = \frac{d\ ^1}{ds} + \begin{matrix}1\\ j\ \ k\end{matrix}\ ^j\frac{dx^k}{ds}$$

$$= \frac{d\ ^1}{ds} + \begin{matrix}1\\ 2\ \ 2\end{matrix}\ ^2\frac{dx^2}{ds} = 0 + (\ \ x^1)\frac{1}{a}\frac{dt}{ds} = \ \ \frac{1}{a}:$$

$$^2 = \frac{^2}{s} = \frac{d\ ^2}{ds} + \begin{matrix}2\\ j\ \ k\end{matrix}\ ^j\frac{dx^k}{ds}$$

$$= 0 + \begin{matrix}2\\ 1\ \ 2\end{matrix}\ ^1\frac{dx^2}{ds} + \begin{matrix}2\\ 2\ \ 1\end{matrix}\ ^2\frac{dx^1}{ds}$$

$$= \frac{1}{x^1}0\ \ ^2 + \frac{1}{x^1}\frac{1}{a}\ \ ^1 = 0:$$

Similarly, $^3 = 0$: Since is a unit vector, so,

$$g_{ij}\ \ ^i\ \ ^j = 1;\quad \text{i.e.}\quad g_{11}\ \ ^1\ \ ^1 + g_{22}\ \ ^2\ \ ^2 + g_{33}\ \ ^3\ \ ^3 = 1$$

or

$$1\ \ \frac{1}{a}\ \ \frac{1}{a} + 0 + 0\ \ \ 0\ \ 0 = 1\ )\ \ \ = \frac{1}{a};\ \text{as}\ \ >\ \ 0:$$

Consequently,

$$^1 = \ \ \frac{1}{a}\ )\ \ \ ^1 = \ \ 1:$$

So, $^i = (\ \ 1; 0; 0)$: From Eq. (5.17), we get $\frac{^i}{s} = \ \ \ ^i + \ \ ^i; i = 1; 2; 3$; thus,

$$^1\ \ \ \ ^1 = \frac{^1}{s} = \frac{d\ ^1}{ds} + \begin{matrix}1\\ j\ \ k\end{matrix}\ ^j\frac{dx^k}{ds}$$

$$= 0 + \begin{matrix}1\\ 2\ \ 2\end{matrix}\ \ ^2\ \ ^2 = 0:$$

Therefore, $^1 = \ \ \ ^1 = 0$: Now

$$^2\ \ \ \ ^2 = \frac{^2}{s} = \frac{d\ ^2}{ds} + \begin{matrix}2\\ j\ \ k\end{matrix}\ ^j\frac{dx^k}{ds}$$

$$= 0 + \begin{matrix}2\\ 1\ \ 2\end{matrix}\ \ ^1\ \ ^2 + \begin{matrix}2\\ 2\ \ 1\end{matrix}\ \ ^2\ \ ^1$$

$$= \frac{1}{a}(\ \ 1)\frac{1}{a} + 0 = \ \ \frac{1}{a^2}$$

or

$$^2 = \quad ^2 \quad \frac{1}{a^2} = \frac{1}{a}\frac{1}{a} \quad \frac{1}{a^2} = 0:$$

and

$$^3 \quad\quad ^3 = \frac{\ ^3}{s} = \frac{d\ ^3}{ds} + \genfrac{}{}{0pt}{}{3}{j\ \ k} \quad ^j \frac{dx^k}{ds} = 0:$$

Therefore $^3 =$ ^{3}: Since is a unit binormal vector, so

$$^1 = \quad ^2 = \quad ^3 = 0 \ ) \quad = 0:$$

To determine the component of , using $^i = "^{ijk} \ _j \ _k$ we get

$$^1 = "^{1jk} \ _j \ _k = "^{123} \ _2 \ _3 + "^{132} \ _3 \ _2 = 0$$

$$^2 = "^{2jk} \ _j \ _k = "^{231} \ _3 \ _1 + "^{213} \ _1 \ _3 = 0$$

$$^3 = "^{3jk} \ _j \ _k = "^{312} \ _1 \ _2 + "^{321} \ _2 \ _1$$

$$= "^{321} \ _2 \ _1 = \frac{1}{g} e^{321} \ _2 \ _1; g = a^2:$$

Now, using the relation $_2 = g_{2m} \ ^m$ and $_1 = g_{1m} \ ^m$; we get

$$_2 = g_{2m} \ ^m = g_{22} \ ^2 = a^2 \ \frac{1}{a} = a$$

$$_1 = g_{1m} \ ^m = g_{11} \ ^1 = \quad 1:$$

Thus

$$^3 = \frac{1}{a}(\quad 1)a(\quad 1) = 1:$$

Hence, the component of the binormal $= (0; 0; 1)$.

Deduction 5.4.1 Let x^i are curvilinear co-ordinates and $\genfrac{}{}{0pt}{}{i}{j\ \ k}$ are Christoffel symbols corresponding to the metric tensor of E^3. According to definition (3.38), we have

$$\frac{\ ^i}{s} = \ ^i_{;k} \frac{dx^k}{ds};$$

where the comma (,) notation denotes covariant differentiation with respect to the metric tensor. Hence,

$$\frac{\ ^i}{s} = \frac{\ ^i}{x^k} + \ ^p \genfrac{}{}{0pt}{}{i}{p\ \ k} \frac{dx^k}{ds}$$

$$= \frac{d\ ^i}{ds} + \ ^p \genfrac{}{}{0pt}{}{i}{p\ \ k} \frac{dx^k}{ds} = \frac{d^2x^i}{ds^2} + \genfrac{}{}{0pt}{}{i}{p\ \ k} \frac{dx^k}{ds}\frac{dx^p}{ds}:$$

Thus, the formula $\frac{\ ^i}{s} = \quad ^i$ can be written as

$$\frac{d\ ^i}{ds} + \ ^p \quad \begin{matrix} i \\ j\ \ k \end{matrix} \quad \frac{dx^j}{ds} = \quad ^i$$

or

$$\frac{d^2x^i}{ds^2} + \quad \begin{matrix} i \\ j\ \ k \end{matrix} \quad \frac{dx^j}{ds}\frac{dx^k}{ds} = \quad ^i:$$

Similarly, we have,

$$\frac{\ ^i}{s} = \frac{d\ ^i}{ds} + \quad \begin{matrix} i \\ j\ \ k \end{matrix} \quad \ ^j \frac{dx^k}{ds}$$

and

$$\frac{\ ^i}{s} = \frac{d\ ^i}{ds} + \quad \begin{matrix} i \\ j\ \ k \end{matrix} \quad \ ^j \frac{dx^k}{ds}: \qquad (5.18)$$

These are the Serret{Frenet formulas in curvilinear co-ordinates E^3:

EXAMPLE 5.4.6 Find the equation of the straight line in curvilinear co-ordinate system E^3:

Solution: Let be a straight line in E^3 and (x^i) be the curvilinear co-ordinates of any point on it. Then the direction of the tangent vector of is fixed. Hence,

$$\frac{\ ^i}{s} = 0; \text{ i.e. } \quad ^i = 0:$$

Since $\ ^i$ is not zero, it follows that $= 0$: Hence, from Eq. (5.18) we get

$$\frac{d^2x^i}{ds^2} + \quad \begin{matrix} i \\ j\ \ k \end{matrix} \quad \frac{dx^j}{ds}\frac{dx^k}{ds} = 0:$$

Since this result holds for any point $P(x^i)$ on , it is the equation of in curvilinear co-ordinates E^3:

EXAMPLE 5.4.7 Show that a space curve is a straight line if and only if its curvature is zero at all points of it.

Solution: First, suppose that the curvature $= 0$ at all points of a space curve : Then by the Frenet formula,

$$\frac{\ ^i}{s} = \quad ^i\ ) \quad \frac{\ ^i}{s} = 0$$

for all s. This means that $\ ^i$ is a fixed direction. Hence is a straight line.

Conversely, let be a straight line, then the direction of the tangent vector of is fixed. Hence,

$$\frac{{}^{i}}{s} = 0 \ ) \qquad {}^{i} = 0:$$

Since ^{i} is not zero, it follows that the = 0: From this example, we can say that the curvature of a curve measures how much the curve differ from being a straight line.

EXAMPLE 5.4.8 Prove that a space curve is a plane curve if and only if its torsion is zero at all points of it.

Solution: First, suppose that a space curve is a plane curve. Then the osculating plane at every point is the plane of . Hence, the binormal ^{i} has a fixed direction, so $\frac{{}^{i}}{s} = 0$: Using Frenet formula we have

$$\frac{{}^{i}}{s} = \qquad {}^{i}\ ) \qquad = 0 \text{ as } {}^{i} \neq 0:$$

Next, we suppose that for a space curve , = 0 at every point of it. Then $\frac{{}^{i}}{s} = 0$; which means that the binormal vector ^{i} has a fixed direction. Hence, the curve is a plane curve. From this example, we can say that the torsion of a curve measures how far the curve departs from lying in a plane.

EXAMPLE 5.4.9 If (a ^{r} + b ^{r} + c ^{r}) forms a parallel vector field along , prove that

$$\frac{da}{ds} \quad b = 0; \ \frac{db}{ds} + \ a \quad c = 0; \ \frac{dc}{ds} + \ b = 0;$$

where the notations have their usual meanings.

Solution: Let $A^{r} = a\ {}^{r} + b\ {}^{r} + c\ {}^{r}$: This vector A^{r} forms a parallel vector field along , if,

$$\frac{A^{i}}{s} = A^{i}_{;j}\frac{dx^{j}}{ds} = 0$$

or

$$\frac{}{s}\ \ a\ {}^{i} + b\ {}^{i} + c\ {}^{i} \ \ = 0$$

or

$$\frac{da}{ds}\ {}^{i} + \frac{db}{ds}\ {}^{i} + \frac{dc}{ds}\ {}^{i} + a\frac{{}^{i}}{s} + b\frac{{}^{i}}{s} + c\frac{{}^{i}}{s} = 0$$

or

$$\frac{da}{ds}\ {}^{i} + \frac{db}{ds}\ {}^{i} + \frac{dc}{ds}\ {}^{i} + a\ \ {}^{i} + b\ \ {}^{i}\ \ \ {}^{i} + c(\ \ \ {}^{i}) = 0$$

or

$$\frac{da}{ds} \quad b \quad {}^{i} + \frac{db}{ds} + a \quad c \quad {}^{i} + \frac{dc}{ds} + b \quad {}^{i} = 0:$$

Since ^{i}; ^{i} and ^{i} are arbitrary, we get

$$\frac{da}{ds} \quad b = 0; \frac{db}{ds} + a \quad c = 0; \frac{dc}{ds} + b = 0:$$

Thus the curvature and the torsion are both constant, and therefore their ratio is constant. The principal normal intersects the axis of the cylinder orthogonally an the tangent and binormal are inclined at constant argles to the fixed direction of the generators.

5.5 Bertrand Curves

Saint{Venant proposed and Bertrand solved the problem of finding the curves whose principal normals are also the principal normals of another curve. If the two curves and $\overline{\ }$ given by (Figure 5.4)

$$: \ x^{i} = x^{i}(s) \text{ and } \overline{\ }: \ \overline{x}^{i} = \overline{x}^{i}(s); \tag{5.19}$$

have a common principal normal, at any of their points, the one is called theBertrand of another or one is called theBertrand associate of another.

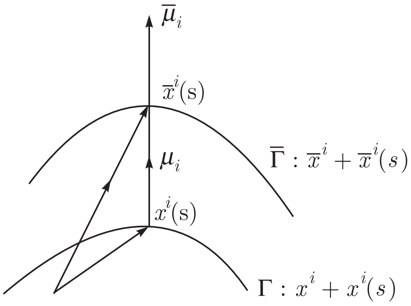

Figure 5.4: Bertrand curve.

EXAMPLE 5.5.1 If is a plane curve, find a curve such that and are Bertrand curves.

Solution: Let E be a plane evolute of , then all involutes of E have the same principal normal as the given curve , because by definition of the involute, these curves intersect orthogonally any tangent to E (Figure 5.5). and any of the curve are therefore Bertrand curves. Moreover, the distance measured along the common principal normal, between corresponding points of any two of these involutes is constant. Two plane curves are therefore called plane parallel curves if they have common principal normals.

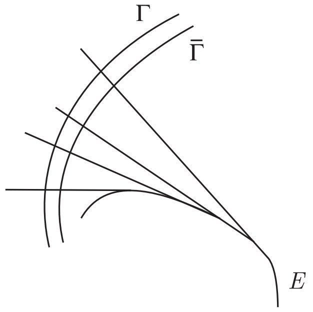

Figure 5.5: Plane curve.

EXAMPLE 5.5.2 Find the necessary and sufficient condition for two curves and are Betrand.

Solution: Let: : $x^i = x^i(s)$ and

$$= x^i = x^i(s) = x^i(s) + a(s) \quad {}^i \qquad (5.20)$$

Differentiating with respect to s intrinsically, we get

$$\frac{\bar{x}^i}{s} = \frac{x^i}{s} + a^1(s) \quad {}^i + a(s)\frac{{}^i}{s}$$

$$= {}^i + a^1(s) \quad {}^i + a(s) \quad {}^i_1 \qquad {}^i$$

or

$$\bar{}^i\frac{d\bar{s}}{ds} = {}^i + a^1(s) \quad {}^i + a(s) \quad {}^i_1 \qquad {}^i \;: \qquad (5.21)$$

Since the two curves are Bertrand, so $\bar{}^i$ is parallel to ^{i}: Taking the inner product of Eq. (5.21) with ^{i} we get

$$\bar{}^i\,\bar{}^i\frac{d\bar{s}}{ds} = {}^i\;{}^i + a^1(s) \quad {}^i\;{}^i + a(s) \quad {}^i_1\;{}^i \qquad {}^i\;{}^i$$

or

$$0 = 0 + a^1(s)1 + a(s)[\quad 0 \qquad 0]\;) \quad a(s) = \text{constant},$$

which is the necessary condition that the two curves are Bertrand curves. Also, from Eq. (5.21), we get

$$\bar{}^i\frac{d\bar{s}}{ds} = {}^i + a(s) \quad {}^i_1 \qquad {}^i = {}^i[1 \quad a \;] + a \quad {}^i:$$

Further, refer to the curve $\bar{}$, we have

$$\frac{d}{ds}(\;{}^i \;\bar{}^i) = \;{}^i\bar{}^i + \;{}^i(-\;{}^i)\frac{d\bar{s}}{ds} = 0 \;) \quad {}^i \;\bar{}^i = \text{constant}.$$

Thus, the tangents to the two curves are inclined at a constant angle. But the principal normals coincide, and therefore, the binormals of the two curves are inclined at the same constant angle. Let be the inclination of $\bar{}^{i}$ to ^{i}, then is constant. If the curve is not plane, i.e. $\neq 0$, then from Eq. (5.21), we get

$$\bar{}^{i} = {}^{i}[1 \quad a\ ]\frac{ds}{d\bar{s}} + a \ {}^{i}\frac{ds}{d\bar{s}}: \tag{5.22}$$

^{i} lies in the plane ^{i} ^{i}, i.e. on the rectifying plane. Moreover, it is obvious from the Figure (5.6) that

$$\bar{}^{i} = \cos \ {}^{i} + \sin \ {}^{i}: \tag{5.23}$$

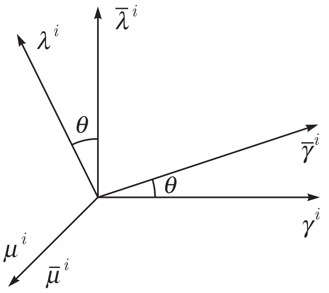

Figure 5.6: The triad.

Comparing Eqs. (5.22) and (5.23), we get

$$\frac{1 \quad a}{\cos} = \frac{a}{\sin} = \frac{d\bar{s}}{ds} \ ) \ 1 \quad a \ = a \ \cot \ : \tag{5.24}$$

Differentiating Eq. (5.23) with respect to s, we get

$$\begin{aligned} \bar{}\frac{d\bar{s}}{ds} &= \sin \ {}^{i}\frac{d}{ds} + \cos \ {}^{i} + \cos \ {}^{i}\frac{d}{ds} \ {}^{i}\sin \\ &= {}^{i}(\ \cos \quad \sin \ ) + \ {}^{i}\cos \quad {}^{i}\sin \ \frac{d}{ds}: \end{aligned}$$

Since the curves are Bertrand curve, so $\bar{}^{i}$ and ^{i} are parallel and equal. Therefore,

$$\begin{aligned} {}^{i}\cos \quad {}^{i}\sin \ \frac{d}{ds} &= 0 \\ ) \ \frac{d}{ds} = 0 \ ) \quad &= \text{constant} \\ ) \ 1 \quad a \ = a \ \cot \ &= b \ (\text{say}); \end{aligned}$$

where b = a cot = constant. Thus, we get a linear equation in and with constant coefficients as a + b = 1: This is the necessary condition that the two curves are Bertrand curves when the curves is not plane.

We shall now show that if X and Y satisfying a + b = 1, where a; b are constants, then is Bertrand of ¯. Differentiating Eq. (5.20) intrinsically with respect to s, we get

$$\bar{}^{i}\frac{d\bar{s}}{ds} = {}^{i}[1 \quad a\ ] + a\ {}^{i} = {}^{i}b + a\ {}^{i}$$

$$= \quad b\ {}^{i} + a\ {}^{i}\ ;\ \text{as}\ a + b = 1$$

or

$$\bar{}^{i} = \quad b\ {}^{i} + a\ {}^{i}\ \frac{ds}{d\bar{s}} = d\ b\ {}^{i} + a\ {}^{i}\ :$$

Since $\bar{}^{i}$ is unit vector, so

$$1 = a^2 d^2 + b^2 d^2 \ )\quad d = \frac{1}{\sqrt{a^2 + b^2}} = \text{constant}:$$

Therefore

$$\bar{}^{i} = p\ {}^{i} + q\ {}^{i};$$

where p and q are constants. Differentiating with respect to s, we get

$$\bar{}\frac{d\bar{s}}{ds} = p\ {}^{i} \quad q\ {}^{i} = {}^{i}(p \quad q\ )$$

$$)\ \bar{}^{i}\ \text{is parallel to}\ {}^{i}\ )\ \bar{}^{i} = {}^{i}:$$

Hence, the curves are Bertrand curves. Thus, the condition is sufficient.

Deduction 5.5.1 If the curve is a plane curve, then = 0 and so

$$\bar{}^{i}\frac{d\bar{s}}{ds} = {}^{i}[1 \quad a\ ]; \tag{5.25}$$

i.e. $\bar{}^{i}$ is parallel to ^{i}. Since $\frac{d\bar{s}}{ds}$ and (1 a) are constants, $\bar{}^{i}$ and ^{i} are both constant vectors and $\bar{}^{i}$ = ^{i}: Differentiating intrinsically, with respect to s, we get,

$$\frac{\bar{x}^{i}}{\bar{s}}\frac{d\bar{s}}{ds} = \frac{{}^{i}}{s}\ )\ \bar{}_{i}\frac{d\bar{s}}{ds} = {}^{i};$$

which shows that $\bar{}^{i}$ is parallel to ^{i}. Thus a = constant is the necessary and sufficient condition that the two curves and ¯ are Bertrand curves, if is a plane curve.

Deduction 5.5.2 There is a linear relation with constant coefficients between the curvature and torsion of . On comparing, we get

$$\cos \ = (1 \quad a\ )\frac{ds}{d\bar{s}};\ \sin \ = a\ \frac{ds}{d\bar{s}}:$$

Now, the relation between the curves and ¯ is clearly a reciprocal one. Hence,

$$\cos = (1 + a)\frac{d\bar{s}}{ds}; \ \sin = a \frac{d\bar{s}}{ds}:$$

On multiplying together of the above equations we obtain the relation

$$\bar{} = \frac{1}{a^2}\sin^2 :$$

This shows that the torsions of the two curves have the same sign, and their product is constant. This theorem is due to Schell. Also,

$$\cos^2 = (1 \quad a)(1 + a)$$

This shows that if $P; P_1$ are corresponding points on the two conjugate Bertrand curves, and $O; O_1$ their centres of curvature, the cross ratio of the range $(P\,OP_1O_1)$ is constant and equal to $\sec^2$. This theorem is due to Mannheim.

5.6 Helix

A space curve traced on the surface of a cylinder, and cutting the generators at a constant angle, is called a helix. Therefore, the tangent vector at every point to a helix makes a constant angle with the fixed direction. Let

$$: \ x^i = x^i(s) \tag{5.26}$$

be an equation of a helix. Let the tangent vector ^{i} to the helix makes a constant angle with a fixed direction a^i (parallel to the generators of the cylinder), where a^i is a unit vector (Figure 5.7). Then,

$$\cos = g_{ij} \ ^i a^j ; 0 < j \ j \quad \frac{1}{2} : \tag{5.27}$$

Differentiating Eq. (5.27) intrinsically with respect to s, we get

$$g_{ij} \frac{\ ^i}{s} a^j = 0$$

or

$$g_{ij} \ ^i a^j = 0; \ \text{as} \ \frac{\ ^i}{s} = \ ^i$$

or

$$g_{ij} \ ^i a^j = 0; \ \text{as} \ > 0:$$

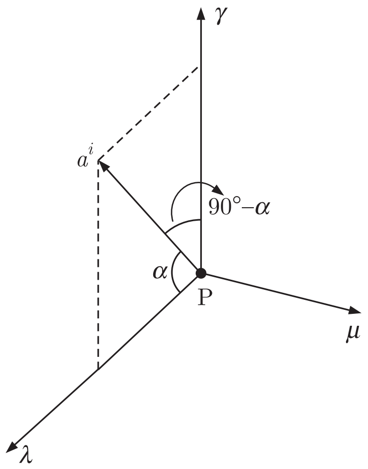

Figure 5.7: Helix.

So, a^i is orthogonal to ^{i}. Thus, since the curvature of the helix does not vanish, the principal normal is everywhere perpendicular to the generators. Hence, the fixed direction of the generators is parallel to the plane of ^{i} and ^{i}; and since it makes a constant angle with ^{i}, it makes a constant angle 90 with ^{i}:

An important property of all helices is that the curvature and torsion are in a constant ratio. Since the angle between a^i and ^{i} is 90 , so,

$$g_{ij}\ {}^i a^j = \cos(90\qquad) = \sin\ : \tag{5.28}$$

Differentiate the relation $g_{ij}\ {}^i a^j = 0$, we get

$$g_{ij}\qquad {}^i\qquad {}^i\ a^i = 0\,:$$

Thus, ^{i} is perpendicular to ^{i} ^{i}. Since a^i must lie in the plane of ^{i} and ^{i}, we can find scalars a and b such that

$$a^i = a\ {}^i + b\ {}^i \tag{5.29}$$

or

$$g_{ij}\ a^i\ {}^j = a g_{ij}\ {}^i\ {}^j + b g_j\ {}^i\ {}^j$$

or

$$\cos\quad = a;\ \text{as}\ g_{ij}\ {}^i\ {}^j = 1 \text{ and } g_{ij}\ {}^i\ {}^j = 0\,:$$

Similarly, from Eq. (5.29) we get,

$$g_{ij}\ a^i\ {}^j = a g_{ij}\ {}^i\ {}^j + b g_j\ {}^i\ {}^j$$

or

$$\sin\quad = b;\ \text{as}\ g_{ij}\ {}^i\ {}^j = 0 \text{ and } g_{ij}\ {}^i\ {}^j = 1\,:$$

Thus, Eq. (5.29) can be written as

$$a^i = {}^i\cos + {}^i\sin : \qquad (5.30)$$

Differentiating Eq. (5.30) intrinsically with respect to s we get

$$\frac{a^i}{s} = \frac{a^i}{s}\cos + \frac{a^i}{s}\sin \ ; \text{ as } = \text{ constant}$$

or

$$0 = {}^i\cos \quad {}^i\sin$$

or

$$\cos \quad \sin = 0; \text{ as } {}^i \neq 0:$$

or

$$- = \tan = \text{ constant.}$$

But a^i is parallel to the plane of ^{i} and ^{i}, and must therefore be parallel to the vector ${}^i +$ ^{i}, which is inclined to ^{i} at an angle $\tan^{1}-$: But this angle is constant. Therefore, the curvature and torsion are in a constant ratio.

Conversely, we may prove that a curve whose curvature and torsion are in a constant ratio is a helix. We suppose that for a space curve different from a straight line $-$ is a constant, i.e. $= c \ ; c =$ constant. Then using Frenet formula we get

$$\frac{{}^i}{s} = c\,{}^i = c\,{}^i = c\frac{{}^i}{s}$$

or

$$\frac{}{s}\ {}^i + c\,{}^i = 0:$$

Therefore, ${}^i + c\,{}^i$ is a constant vector, say a^i. Now,

$$g_{ij}\,a^i\,{}^j = g_{ij}\ {}^i + c\,{}^i\ {}^j$$

$$= g_{ij}\,{}^i\,{}^j + c g_{ij}\,{}^i\,{}^j = 0 + c = c:$$

This shows that the tangent vector ^{i} makes a constant angle with a fixed direction a^i, and the curve is therefore a general helix.

5.6.1 Cylindrical Helix

The cylindrical helix is a curve upon a cylinder which cuts the generator of the cylinder at a constant angle with a fixed direction. Let us take z axis (x^3) as the axis of the cylindrical. Then the equation of the cylinder

$$x^i = {}^i(t);\ i = 1;2:$$

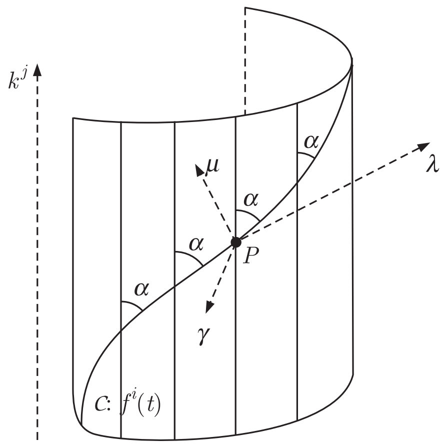

Figure 5.8: Cylindrical helix.

Any curve on the cylinder can be defined by C: $x^i = f^i(t)$; $i = 1, 2, 3$ (Figure 5.8). Let α be the angle at any point P on the curve made by the tangent at P and the generator of this cylinder passing through P. Let the unit vector along the generator is

$$\vec{k} = (0, 0, 1) = (k^1, k^2, k^3)$$

and the unit tangent to the curve at P is

$$\lambda^i = \frac{dx^i}{ds} = \frac{dx^i/dt}{ds/dt} = \frac{dx^i/dt}{|dx^i/dt|}$$

$$= \frac{\dot{f}^1(t), \dot{f}^2(t), \dot{f}^3(t)}{[\{\dot{f}^1(t)\}^2 + \{\dot{f}^2(t)\}^2 + \{\dot{f}^3(t)\}^2]^{1/2}}.$$

Hence, by given condition of the helix,

$$\cos\alpha = g_{ij}\lambda^i k^j = \frac{\dot{f}^3(t)}{\left[\{\dot{f}^1(t)\}^2 + \{\dot{f}^2(t)\}^2 + \{\dot{f}^3(t)\}^2\right]^{1/2}}$$

or

$$\{\dot{f}^3(t)\}^2 = \cos^2\alpha \left[\{\dot{f}^1(t)\}^2 + \{\dot{f}^2(t)\}^2 + \{\dot{f}^3(t)\}^2\right]$$

or

$$\{\dot{f}^3(t)\}^2 \sin^2\alpha = \cos^2\alpha \left[\{\dot{f}^1(t)\}^2 + \{\dot{f}^2(t)\}^2\right]$$

or

$$\dot{f}^3(t) = \cot\alpha \left[\{\dot{f}^1(t)\}^2 + \{\dot{f}^2(t)\}^2\right]^{1/2}$$

or

$$f^3(t) = \cot \int \left[\{\dot f^1(t)\}^2 + \{\dot f^2(t)\}^2 \right]^{1/2} + \text{constant}$$

or

$$f^3(t) = x^3(t) = \cot \int \{\dot x^1\}^2 + \{\dot x^2\}^2{}^{1/2} + \text{constant}:$$

Therefore, the equation of the cylindrical helix can be written as

$$\left.\begin{aligned} &x^1 = f^1(t);\ x^2 = f^2(t) \\ &x^3 = f^3(t) = \cot \int \{\dot x^1\}^2 + \{\dot x^2\}^2{}^{1/2} + \text{constant}: \end{aligned}\right\} \tag{5.31}$$

5.6.2 Circular Helix

If a space curve lies on a circular cylinder and the tangent at each point of it makes a constant angle with the axis of the cylinder, then the curve is a helix. Such a helix is called a circular helix. Thus, a particular case of the cylindrical helix is circular helix, which can be defined by $x^1 = a\cos t$ and $x^2 = a\sin t$. Hence, by Eq. (5.33), we get

$$\begin{aligned} x^3( ) &= \cot \int_0^t \sqrt{a^2(\cos^2 t + \sin^2 t)}\, d \\ &= a\cot\ t = bt;\ \text{where } b = a\cot\ = \text{constant.} \end{aligned}$$

Thus, the parametric equation of the circular helix can be written as

$$C : x^1 = a\cos t;\quad x^2 = a\sin t;\quad x^3 = bt;\quad b \neq 0: \tag{5.32}$$

This curve [Eq. (5.32)] can be defined by giving a constant rotation about the axis of the cylinder and a constant translation about the same axis (Figure 5.9).

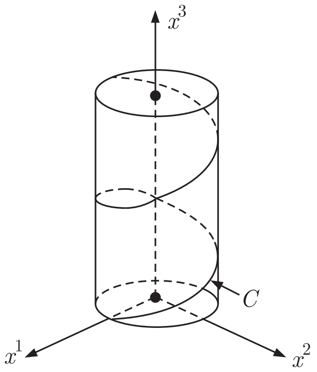

Figure 5.9: Circular Helix.

The circular helix is a twisted curve. The orthogonal projection of the helix into the x^1x^2-plane is the circle

$$(x^1)^2 + (x^2)^2 - a^2 = 0;\ x^3 = 0;$$

which is the intersection of this plane with the cylinder of revolution on which the helix lies. Projecting the helix orthogonally into the x^1x^3-plane, we obtain the sine curve

$$x^2 - a\sin\frac{x^3}{b} = 0;\ x^1 = 0.$$

A cosine curve is obtained by projecting the helix orthogonally into the x^1x^3-plane. Now

$$\frac{d\vec{r}}{dt} = (-a\sin t;\ a\cos t;\ b).$$

Therefore the arc $s(t)$ is given by

$$s(t) = \int_0^t \left(\frac{d\vec{r}}{dt}\cdot\frac{d\vec{r}}{dt}\right)^{1/2} dt = \int_0^t (a^2+b^2)^{1/2}dt$$
$$= \sqrt{a^2+b^2}\ t = wt; \quad \text{where } w = \sqrt{a^2+b^2}.$$

Thus the parametric representation of the circular helix with s as parameter is given by

$$\vec{r} = \left(a\cos\frac{s}{w};\ a\sin\frac{s}{w};\ \frac{b}{w}\right);\quad w = \sqrt{a^2+b^2}.$$

The unit tangent vector to this curve is of the form

$$\vec{t} = \frac{d\vec{r}}{ds} = \left(-\frac{a}{w}\sin t;\ \frac{a}{w}\cos t;\ \frac{b}{w}\right);\quad w = \sqrt{a^2+b^2}.$$

The osculating plane of the circular helix can be represented in the form

$$\begin{vmatrix} z_1 - a\cos t & -a\sin t & -a\cos t \\ z_2 - a\sin t & a\cos t & -a\sin t \\ z_3 - bt & b & 0 \end{vmatrix} = 0$$

which is equivalent to

$$z_1 b\sin t - z_2 b\cos t + (z_3 - bt)a = 0$$

The osculating plane passes through the tangent to the helix and passes through the straight line $\vec{z}(e) = (c\cos t;\ c\sin t;\ bt)$ where t is fixed and c is a real parameter. This line is parallel to x^1x^2-plane. It passes through the axis of the cylinder on which the helix lies as well as through a point P of the helix. Moreover, it is orthogonal to the tangent to the helix at P. Now the curvature and torsion of the circular helix are

$$= \frac{d\bar{r}}{ds} = \sqrt{\frac{d^2\bar{r}}{ds^2}\,\frac{d^2\bar{r}}{ds^2}} = \frac{a}{a^2+b^2}$$

$$= \frac{\left[\frac{d\bar{r}}{ds}; \frac{d^2\bar{r}}{ds^2}; \frac{d^3\bar{r}}{ds^3}\right]}{\frac{d\bar{r}}{ds}\,\frac{d\bar{r}}{ds}\quad\frac{d^2\bar{r}}{ds^2}\,\frac{d^2\bar{r}}{ds^2}\quad\frac{d\bar{r}}{ds}\,\frac{d^2\bar{r}}{ds^2}^{\,2}} = \frac{b}{a^2+b^2}.$$

The circular helix $r = (a\cos t; a\sin t; bt)$ is said to be right-handed (Figure 5.10) if $b > 0$, where the curvature $k = \frac{a}{a^2+b^2}$. As the torsion $= \frac{b}{a^2+b^2}$ so the curve is right-handed when > 0.

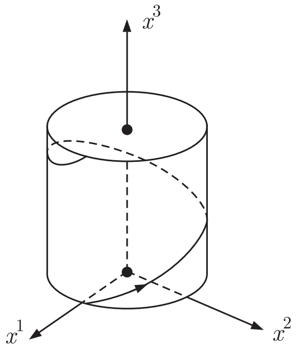

Figure 5.10: Circular Helix > 0.

Similarly, the circular helix $r = (a\cos t; a\sin t; bt)$ is said to be left-handed if $b < 0$, that is < 0 (Figure 5.11).

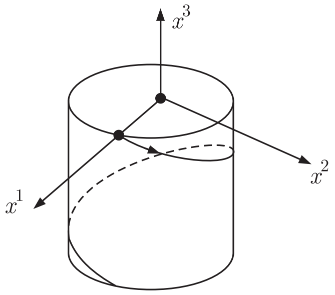

Figure 5.11: Circular Helix < 0.

If b tends to 0, the radius of curvature $= \frac{1}{}$ tends to the radius a of the circle of intersection between x^1x^2-plane and the cylinder of revolution on which the helix lies.

EXAMPLE 5.6.1 Prove that

$$"_{ijk} \frac{i}{s} \frac{2\ j}{s^2} \frac{3\ k}{s^3} = {}^5 \frac{d}{ds} \quad - :$$

Hence, a curve is a helix if and only if

Solution: Using Eq. (5.7), we get

$$"_{ijk} \frac{i}{s} \frac{2\ j}{s^2} \frac{3\ k}{s^3} = 0 :$$

$$\frac{i}{s} = {}^i; \text{ where } > 0:$$

Differentiating intrinsically with respect to the parameter s we get

$$\frac{2\ i}{s^2} = \frac{}{s}\ {}^i + \frac{i}{s} = \frac{}{s}\ {}^i + \quad {}^i \quad {}^i :$$

$$\frac{3\ i}{s^3} = \frac{2}{s^2}\ {}^i + \frac{}{s}\frac{i}{s} + \frac{}{s} \quad {}^i$$

$$+ \frac{}{s}\ {}^i + \frac{i}{s} \quad 2\frac{}{s} \quad {}^i \quad {}^2\frac{i}{s}$$

$$= \frac{2}{s^2}\ {}^i + \frac{}{s} \quad {}^i \quad {}^i + \frac{}{s} + \frac{}{s} \quad {}^i$$

$$+ \quad {}^i \quad 2\frac{}{s} \quad {}^i \quad {}^2 \quad {}^i$$

$$= \frac{2}{s^2} \quad {}^2 \quad {}^3 \quad {}^i + 2\frac{}{s} + \frac{}{s} \quad {}^i + 3\frac{}{s} \quad {}^i :$$

Therefore the given expression becomes,

$$"_{ijk} \frac{i}{s} \frac{2\ j}{s^2} \frac{3\ k}{s^3}$$

$$= "_{ijk} \quad {}^i \frac{}{s}\ {}^j + \quad {}^j \quad {}^j \quad \frac{2}{s^2} \quad {}^2 \quad {}^3 \quad {}^k$$

$$+ 2\frac{}{s} + \frac{}{s}\ {}^k + 3\frac{}{s} \quad {}^k$$

$$= \frac{}{s}"_{ijk} \quad {}^i\ {}^j + {}^2\ "_{ijk} \quad {}^i\ {}^j \quad 3"_{ijk} \quad {}^i\ {}^j$$

$$\frac{2}{s^2} \quad {}^2 \quad {}^3 \quad {}^k + 2\frac{}{s} + \frac{}{s}\ {}^k + 3\frac{}{s} \quad {}^k$$

$$= 0 + {}^2\ {}^k + {}^3\ {}^k \quad \frac{2}{s^2} \quad {}^2 \quad {}^3 \quad {}^k + 2\frac{}{s} + \frac{}{s} \quad {}^k$$

$$+ 3\frac{}{s} \quad {}^k$$

$$= \frac{2}{s^2} \quad 2 \quad 3 \quad 2 \quad k \; k + 3 \; k \; k + 2\frac{}{s} + \frac{}{s}$$

$$2 \quad k \; k + 3 \; k \; k \quad 3 \frac{}{s} \quad 2 \quad k \; k + 3 \; k \; k$$

$$= {}^{3} \; 2\frac{}{s} + \frac{}{s} \quad 3 \frac{}{s}({}^{2})$$

$$= {}^{3} \frac{}{s} \quad 3 \frac{}{s} = {}^{5}\frac{1}{2} \frac{}{s} \quad 3 \frac{}{s} = {}^{5}\frac{d}{ds} \; - :$$

Thus the result is established. Now,

$$"_{ijk} \frac{i}{s} \frac{2 \; j}{s^2} \frac{3 \; k}{s^3} = 0$$

or

$$^{i} + \quad ^{i} \quad ^{i}(\; ^{2} + \; ^{2}) = 0$$

or

$$^{i} + \quad ^{i} = 0; \text{ as } \; ^{2} + \; ^{2} \neq 0 \text{ and } \; ^{i} \text{ is arbitrary}$$

or

$$g_{ij} \quad ^{i} + \quad ^{i} = 0 :$$

If a^i is the fixed direction, then by hypothesis $g_{ij}\, a^i \; ^j = \cos$ and $g_{ij}\, a^i \; ^j = \sin$: Therefore,

$$g_{ij}\, a^i \quad ^{i} + \quad ^{i} = 0$$

or

$$\cos \; + \; \sin \; = 0$$

or

$$- = \; \tan \; = \text{ constant.}$$

Hence, the curve is a helix.

EXAMPLE 5.6.2 Prove that a circular helix is a Bertrand curve.

Solution: We know the circular helix is a twisted curve. The principal normals to a circular helix intersect the axis of rotation of the cylinder Z on which lies at right angles. The points of intersection of these normals with any cylinder $\bar{Z}$, coaxial with Z, determine a circular helix $\bar{}$ (Figure 5.12). Obviously and $\bar{}$ are Bertrand curves, and there exist infinitely many helices $\bar{}$ of this type which, together with , are Bertrand curves.

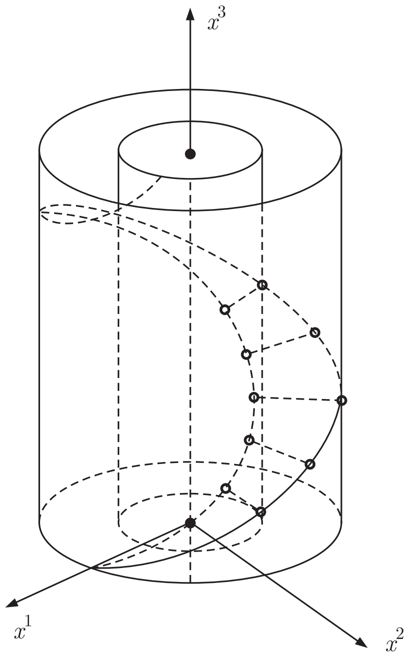

Figure 5.12: Coaxial circular helices < 0.

If two curves : $x^i(s)$; $k > 0$ and $\bar{}$: $\bar{x}^i(s)$ are Bertrand curves then, $\bar{}$ can be represented in the form

$$x^i(s) = x^i(s) + a(s)\ {}^i(s)$$

where ${}^i(s)$ is the unit principal normal vector to and the scalar $a(s)$ is the distance of a point P of $\bar{}$ from the corresponding point P of . Now $a(s)$ is constant, that is, is dependent of S. Hence the circular helix is a Bertrand curve.

EXAMPLE 5.6.3 If the tangent and binormal to a space curve make angles and with a xed direction, show that

$$- = \frac{\sin}{\sin}\frac{d}{d}:$$

Solution: Let the tangent vector with components ^{i} to a space curve C make an angle with a xed direction c^i, where the vector with components c^i is a unit vector and the binormal ^{i} makes an angle with c^i. Then,

$$g_{ij}\ {}^i c^j = \cos \quad \text{and}\ g_{ij}\ {}^i c^j = \cos\ :$$

Since and are invariants,

$$\frac{}{s} = \frac{d}{ds} \text{ and } \frac{}{s} = \frac{d}{ds}:$$

Taking intrinsic derivative of both the sides of the equation $g_{ij}\ ^{i}c^{j} = \cos$, we get

$$g_{ij}\quad ^{i}\ c^{j} = \frac{\ }{s}\cos\ = \frac{d}{ds}(\cos\ )$$

or

$$g_{ij}\ ^{i}c^{j} = \quad \sin\ \frac{d}{ds}:$$

Similarly, from the relation $g_{ij}\ ^{i}c^{j} = \cos$, we get

$$g_{ij}\ ^{i}c^{j} = \quad \sin\ \frac{d}{ds}:$$

Combining these two results, we get

$$\frac{\ }{\ } = \quad \frac{\sin}{\sin}\frac{d}{d}:$$

EXAMPLE 5.6.4 A curve is defined in cylindrical co-ordinates x^i as follows:

$$x^1 = a;\ x^2 = t;\ x^3 = bt;\ b \neq 0$$

where a and b are constants of which a is positive and t is a function of the natural parameter s. Find the curvature and torsion of .

Solution: The components of the tangent vector ^{i} are given by

$$^1 = \frac{dx^1}{ds} = 0;\quad ^2 = \frac{dx^2}{ds} = \frac{dt}{ds};\quad ^3 = \frac{dx^3}{ds} = b\frac{dt}{ds}:$$

Since $g_{ij}\ ^{i}\ ^{j} = 1$, we get

$$g_{11}\ ^{1}\ ^{1} + g_{22}\ ^{2}\ ^{2} + g_{33}\ ^{3}\ ^{3} = 1$$

or

$$0 + \ {x^1}^{2}\ {\frac{dt}{ds}}^{2} + b^2\ {\frac{dt}{ds}}^{2} = 1$$

or

$$(a^2 + b^2)\ {\frac{dt}{ds}}^{2} = 1\ )\quad {\frac{dt}{ds}}^{2} = \frac{1}{a^2 + b^2}:$$

From the relation

$$^i = \frac{d\ ^i}{ds} + \ \begin{matrix} i \\ j\ \ k \end{matrix}\ \ ^j\frac{dx^k}{ds};$$

we get

$$^{1} = \frac{d\ ^{1}}{ds} + \begin{matrix}1\\ 2\ \ 2\end{matrix}\ ^{2}\frac{dx^{2}}{ds} = \ x^{1}\ \frac{dt}{ds}\ ^{2} = \frac{a}{a^{2}+b^{2}}$$

$$^{2} = \frac{d\ ^{2}}{ds} + \begin{matrix}2\\ 2\ \ 1\end{matrix}\ ^{2}\frac{dx^{1}}{ds} + \begin{matrix}2\\ 1\ \ 2\end{matrix}\ ^{1}\frac{dx^{2}}{ds} = 0$$

$$^{3} = 0:$$

Using $g_{ij}\ ^{i}\ ^{j} = 1$, we get

$$g_{11}\ ^{1}\ ^{1} + g_{22}\ ^{2}\ ^{2} + g_{33}\ ^{3}\ ^{3} = 1$$

or

$$\frac{a}{(a^{2}+b^{2})}:\ \ \frac{a}{(a^{2}+b^{2})} = 1\ )\ \ = \frac{a}{a^{2}+b^{2}}:$$

So $^{1} = \ 1;\ \ ^{2} = 0;$ and $^{3} = 0:$ From the relation

$$\frac{d\ ^{i}}{ds} + \begin{matrix}i\\ j\ \ k\end{matrix}\ ^{j}\frac{dx^{k}}{ds} = \ ^{i} + \ ^{i};$$

we get

$$\frac{d\ ^{1}}{ds} + \begin{matrix}1\\ 2\ \ 2\end{matrix}\ ^{2}\frac{dx^{2}}{ds} = \ ^{1} + \ ^{2}$$

or

$$0 + 0 = \ ^{1}\ )\ \ ^{1} = 0:$$

Similarly, $^{2} = \frac{b^{2}}{a(a^{2}+b^{2})^{3=2}}$ and $^{3} = \frac{ab}{(a^{2}+b^{2})^{3=2}}:$ Using $g_{ij}\ ^{i}\ ^{j} = 1$, we get

$$g_{11}\ ^{1}\ ^{1} + g_{22}\ ^{2}\ ^{2} + g_{33}\ ^{3}\ ^{3} = 1$$

or

$$0 + a^{2}\frac{b^{2}}{a(a^{2}+b^{2})^{3=2}}\ \ \frac{b^{2}}{a(a^{2}+b^{2})^{3=2}} + \frac{ab}{(a^{2}+b^{2})^{3=2}}\ \ \frac{ab}{(a^{2}+b^{2})^{3=2}} = 1$$

or

$$^{2} = \frac{b^{2}(a^{2}+b^{2})}{(a^{2}+b^{2})^{3}} = \frac{b^{2}}{(a^{2}+b^{2})^{2}}\ )\ \ = \frac{b}{a^{2}+b^{2}}:$$

Thus, for the given helix, both the curvature and the torsion are constants. Next, we shall show that and are constants, thus, there exists a unique circular helix. Let $= l$ and $= m$ be constants. Choose

$$l = \frac{a}{a^{2}+b^{2}}\ \text{ and }\ m = \frac{b}{a^{2}+b^{2}};$$

where a and b are constants. Therefore,

$$\frac{l}{a} = \frac{m}{b} = \frac{1}{a^2 + b^2}) \quad a = l(a^2 + b^2);\ b = m(a^2 + b^2):$$

Also

$$l^2 + m^2 = \frac{a^2 + b^2}{(a^2 + b^2)^2} = \frac{1}{a^2 + b^2}$$

So

$$\frac{a}{l} = \frac{b}{m} = a^2 + b^2 = \frac{1}{l^2 + m^2}:$$

Let us form a curve

$$:\ x^i(t) = \frac{l}{l^2 + m^2}\cos t;\ \frac{m}{l^2 + m^2}\sin t;\ \frac{m}{l^2 + m^2}t \quad ;$$

which will satisfy all the necessary conditions for a circular helix.

5.7 Spherical Indicatrix

The locus of a point, whose position vector is equal to unit tangent ^{i} of a given curve, is called the spherical indicatrix of the tangent to the curve. Such a locus lies on the surface of a unit sphere; hence the name. Let the two curves and $\bar{}$ be given by

$$:\ x^i = x^i(s) \text{ and } \bar{}:\ \bar{x}^i = \bar{x}^i(s):$$

of class 2 with non-vanishing curvature. We assume that the tangent principal normal and the binormal vectors undergo a parallel displacement (Figure 5.13) and become bound at the origion O is space. Then the terminal points of these vectors (s); (s) are (s) lie on a unit sphere S with centre O and generate, is general, three curves on S which are called tangent indicatrix, the principal normal indicatrix and the binormal indicatrix, respectively, of the curve .

Thus, according to the definition of tangent indicatrix, $\bar{x}^i(s) = {}^i$; therefore,

$$\bar{}^i = \frac{d\bar{x}^i}{d\bar{s}} = \frac{d\ {}^i}{ds}\frac{ds}{d\bar{s}} = \quad {}^i\frac{ds}{d\bar{s}}$$

showing that the tangent to the spherical indicatrix is parallel to the principal normal at the corresponding point of the given curve.

The linear element $d\bar{s}_1; d\bar{s}_N$ and $d\bar{s}_B$ of these indicatrices or spherical images can be obtained by means of Eq. (5.17) for the Serret-Frenet formula. Since (s); (s) and

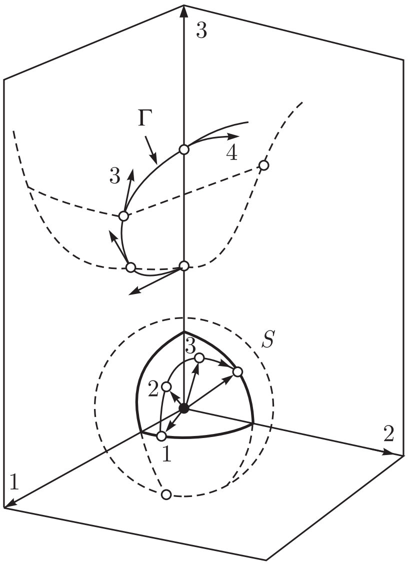

Figure 5.13: Curve and corresponding tangent indicatrix.

(s) are the vector functions representing these curves we find

$$ds_T^2 = \frac{d\ ^i}{ds}\frac{d\ ^j}{ds}ds^2 = k^2\ \ ^i\ \ ^j ds^2 = k^2 ds^2$$

$$ds_N^2 = \frac{d\ ^i}{ds}\frac{d\ ^j}{ds}ds^2 = (\ \ k\ \ ^i + \ \ ^i)(\ \ k\ \ ^i + \ \ ^i)ds^2 = (\ k^2 + \ \ ^2)ds^2$$

$$ds_B^2 = \frac{d\ ^i}{ds}\frac{d\ ^i}{ds}ds^2 = (\ \ ^i)(\ \ ^j)ds^2 = \ \ ^2 ds^2:$$

Moreover we obtain the Equation of Lancret $ds_N^2 = ds_T^2 + ds_B^2$. The expression $\sqrt{ds_T^2 + ds_B^2}$ is sometimes called the third and total curvature of a curve.

We may measure $\bar{s}$ so that $\bar{\ }^i = \ ^i$ and therefore $\frac{d\bar{s}}{ds} = \ $: For the curvature $\bar{\ }$ of the indicatrix, on differentiating the relation $\bar{\ }^i = \ ^i$, we find the formula

$$\bar{\ \ }^i = \frac{d\ ^i}{ds}\frac{ds}{d\bar{s}} = \frac{1}{\ }(\ \ \ ^i\ \ \ \ ^i):$$

Squaring both sides we obtain the result

$$\bar{\ }^2 = \frac{1}{\ ^2}(\ \ ^2 + \ \ ^2);$$

so that the curvature of the indicatrix is the ratio of the screw curvature to the circular curvature of the curve. The unit binormal of the indicatrix is

$$\bar{\ }^i = \text{"}^{ijk}\ \ _j\ \ _k = \frac{1}{\ }(\ \ ^i + \ \ ^i):$$

Using the result

$$"_{ijk}\,\frac{\ \ i}{s}\,\frac{2\ \ j}{s^2}\,\frac{3\ \ k}{s^3} = {}^3\quad\frac{}{s}\quad\frac{}{s}$$

the torsion can be obtained as

$$\overline{\ }^2\ \frac{d\bar{s}}{ds}^{6} = "_{ijk}\,\frac{\ \ i}{s}\,\frac{2\ \ j}{s^2}\,\frac{3\ \ k}{s^3} = {}^3\quad\frac{}{s}\quad\frac{}{s}$$

$$)\ \ \bar{} = \frac{1}{(\ ^2+\ ^2)}\quad\frac{}{s}\quad\frac{}{s}\ :$$

Similarly, the spherical indicatrix of the binormal of the given curve is the locus of a point whose position vector is $\bar{x}^i = {}^i$: Therefore,

$$\bar{}^i = \frac{d\ ^i}{ds}\frac{ds}{d\bar{s}} = \ {}^i\frac{ds}{d\bar{s}}:$$

We may measure $\bar{s}$ so that $\bar{}^i = {}^i$; and therefore $\frac{d\bar{s}}{ds} =$: To find the curvature differentiate the equation $\bar{}^i = {}^i$; we get

$$\bar{}'^i = \frac{d}{ds}(\ ^i)\frac{ds}{d\bar{s}} = \frac{1}{}(\ ^i\quad ^i);$$

giving the direction of the principal normal. On squaring this result we get

$$\bar{}^2 = \frac{1}{^2}(\ ^2+\ ^2):$$

Thus, the curvature of the indicatrix is the ratio of the screw curvature to the torsion of the given curve. The unit binormal is

$$\bar{}^i = \frac{1}{\bar{}}(\ ^i+\ ^i)$$

and the torsion, found as in the previous case, is equal to

$$\bar{} = \frac{1}{(\ ^2+\ ^2)}\quad\frac{}{s}+\frac{}{s}\ :$$

Different curves may have the same spherical images. Simple examples illustrating this fact are circles in the same plane, with arbitrary radius and centre, and also circular helices on coaxial cylinders (Figure 5.12) for which the ratio $a : b$ is the same.

5.8 Exercises

1. If the principal normal of a space curve vanishes identically, show that its equation in curvilinear co-ordinates x^i is given by

$$\frac{d^2x^i}{ds^2} + \begin{matrix} i \\ j \ \ k \end{matrix} \ \frac{dx^j}{ds}\frac{dx^k}{ds} = 0:$$

2. If $"^{ijk}$ is the covariant permutation tensor, show that
   (i) $"^{ijk} \ {}_i \ {}_j = \ {}^k = \ "^{ijk} \ {}_i \ {}_j$
   (ii) $"^{ijk} \ {}_i \ {}_j = \ {}^k = \ "^{ijk} \ {}_i \ {}_j$
   (iii) $"^{ijk} \ {}_i \ {}_j = \ {}^k = \ "^{ijk} \ {}_i \ {}_j:$

3. Show that
   (i) $g_{ij} \frac{i}{s}\frac{j}{s} =$
   (ii) $g_{ij} \frac{i}{s}\frac{j}{s} = \ {}^2 + \ {}^2$
   (iii) $g_{ij} \frac{i}{s}\frac{j}{s} = \ {}^2;$
   (iv) $g_{ij} \frac{i}{s}\frac{j}{s} = 0.$

4. Prove that,

$$"_{ijk} \frac{i}{s}\frac{2 \ j}{s^2}\frac{3 \ k}{s^3} = \ {}^5\frac{d}{ds} \quad - \ :$$

5. Prove that
   (i) $= g_{ij} \frac{i}{s} \ {}^j = \ g_{ij} \ {}^i \frac{j}{s}$
   (ii) $= g_{ij} \frac{i}{s} \ {}^j = \ g_{ij} \ {}^i \frac{j}{s}$
   (iii) $\frac{i}{s} = \ g_{ij} \frac{i}{s} \ {}^j \ \ {}^i + \ g_{ij} \frac{i}{s} \ {}^j \ \ {}^i + \ g_{ij} \frac{i}{s} \ {}^j \ \ {}^i$
   (iv) $g_{ij} \frac{i}{s} \ {}^j = 0;$ (v)$g_{ij} \ {}^i \frac{j}{s} = 0;$ (vi)$g_{ij} \frac{i}{s} \ {}^j = 0:$

6. If $a^i = \ {}^i + \ {}^j$, prove that

$$\frac{k}{s} = "^{ijk} a_i \ {}_j; \quad \frac{k}{s} = "^{ijk} a_i \ {}_j; \quad \frac{k}{s} = "^{ijk} a_i \ {}_j:$$

7. Prove that a space curve is helix if and only if

$$"_{ijk} \frac{i}{s} \frac{2\ j}{s^2} \frac{3\ k}{s^3} = 0:$$

8. Find the curvature and torsion at any point of a given curve , given by

$$: \ x^1 = a; x^2 = t; x^3 = ct;$$

where $ds^2 = dx^{1\ 2} + (x^1)^2\ dx^{2\ 2} + dx^{3\ 2}$; $a; c$ are non-zero constants. Also, show that the ratio of the curvature and torsion is a constant.

9. Show that the ratio of the curvature to the torsion of a space curve different from a straight line is a non-zero constant if and only if the curve is a helix.

10. If the curvature and torsion of a space curve are both constants, prove that the curve is a circular helix.

11. If and are the curvature and torsion of a circular helix given in cylindrical co-ordinates x^i by

$$x^1 = a; \ x^2 = t; \ x^3 = bt; b \neq 0;$$

where $a(> 0)$ and b are non-zero constants, show that
   (a) $a \ + b \ = 1:$
   (b) $a = \frac{}{2 + \ 2}; b = \frac{}{2 + \ 2}:$

12. Show that the curvature and torsion of a helix are in a constant ratio to the curvature of the plane section of the cylinder perpendicular to the generators.

13. Defining a Bertrand curve as a space curve for which $a \ + b \ = 1$; where a and b are non-zero constants with $a > 0$, prove that a circular helix is a Bertrand curve.

14. Prove that the principal normal of a cylindrical helix is everywhere perpendicular to the generators of the cylinder.

15. Prove that a circular helix C is the only twisted curve for which more than one curve C exists such that C and C are Bertrand curves.

16. Prove that the product of the torsions of Bertrand curves is constant.

17. Find the curves of which
   (a) the tangent indicatrix
   (b) the binormal indicatrix

   degenerates a point. What does it mean when a spherical image of a closed curve?

18. Investigate the spherical images of a circular helix.

Chapter 6

Intrinsic Geometry of Surface

We will investigate the differential geometry of surfaces by means of tensor calculus. A great advantage of this method lies in the fact that it can immediately be generalized to Riemannian spaces of higher dimension which have assumed increasing importance during the last few decades. In addition, many aspects of the theory of surfaces are simplified when treated with the aid of tensor calculus which thus leads to a better and deeper insight into several problems of differential geometry.

In this chapter we will study the geometric properties of surfaces imbedded in a three-dimensional Euclidean space by means of differential geometry. The reasoning which will lead us to a definition of a surface is similar to that which led us to the concept of a curve. It will be shown that certain of these properties can be phrased independently of the space in which the surface is immersed and that they are concerned solely with the structure of the differential quadratic form for the element of an arc of a curve drawn on a surface. All such properties of surfaces are termed as the intrinsic properties and the geometry based on the study of this differential quadratic form is called the intrinsic geometry of the surface.

6.1 Curvilinear Co-ordinates on a Surface

A surface often arises as the locus of a point P which satisfies some restrictions. We find it convenient to refer the space in which surface is imbedded to a set of orthogonal Cartesian axes Y and regarded the locus of points satisfying the equation

$$F\left(y^1, y^2, y^3\right) = 0 \tag{6.1}$$

as an analytical implicit or constraint definition of a surface S. We suppose that only two of the variables y^i in Eq. (6.1) are independent and that the specification of, say y^1 and y^2 in some region of the Y^1Y^2-plane determines uniquely a real number y^3 such that the $F\left(y^1, y^2, y^3\right)$ reduces to 0. If we suppose that $F\left(y^1, y^2, y^3\right)$, regarded as a function of three independent variables, is of class C^1 in some region R about the point $P_0\left(y_0^1, y_0^2, y_0^3\right)$ with $\left(\frac{\partial F}{\partial y^3}\right)_{P_0} \neq 0$ and $F\left(y_0^1, y_0^2, y_0^3\right) = 0$, then the fundamental theorem on implicit function guarantees the existence of a unique solution $y^3 = f\left(y^1; y^2\right)$, such that $y^3_0 = f\left(y^1_0; y^2_0\right)$.

We are now going to investigate the properties of a surface in its relation to the surrounding space. Consequently, we are dealing with two distinct system of co-ordinates namely, the three curvilinear co-ordinates for the surrounding space which we denote by $y^i;\ i = 1; 2; 3$ and the two curvilinear co-ordinates of the surface which we denote by $u^\alpha;\ \alpha = 1; 2$: A surface S is defined, in general, due to Gauss, as the set of points whose co-ordinates are functions of two independent parameters. Let us denote the equation of a surface S embedded in E^3 by

$$S : y^1 = y^1(u^1; u^2);\ y^2 = y^2(u^1; u^2);\ y^3 = y^3(u^1; u^2)$$

or

$$S : y^i = y^i\left(u^1; u^2\right); i = 1; 2; 3 \tag{6.2}$$

where $u^1_1 \le u^1 \le u^1_2$ and $u^2_1 \le u^2 \le u^2_2$ are parameters and the y^i are real functions of class C^1 in the region of definition of the independent parameters $u^1; u^2$. The function $y^i(u^1; u^2)$ are single-valued and continuous, and are here assumed to possess continuous partial derivatives of the rth order. In this case the surface is said to be of class r. In order to reconcile these two different definitions we shall require that the functions $y^i\left(u^1; u^2\right)$ be such that the Jacobian matrix

$$J = \begin{bmatrix} \frac{\partial y^1}{\partial u^1} & \frac{\partial y^2}{\partial u^1} & \frac{\partial y^3}{\partial u^1} \\ \frac{\partial y^1}{\partial u^2} & \frac{\partial y^2}{\partial u^2} & \frac{\partial y^3}{\partial u^2} \end{bmatrix} \tag{6.3}$$

be of rank 2, so that not all the determinants of the second order selected from this matrix vanish identically in the region of definition of parameters u^i. The three second order matrices

$$J_1 = \begin{bmatrix} \frac{\partial y^2}{\partial u^1} & \frac{\partial y^2}{\partial u^2} \\ \frac{\partial y^3}{\partial u^1} & \frac{\partial y^3}{\partial u^2} \end{bmatrix};\ J_2 = \begin{bmatrix} \frac{\partial y^3}{\partial u^1} & \frac{\partial y^3}{\partial u^2} \\ \frac{\partial y^1}{\partial u^1} & \frac{\partial y^1}{\partial u^2} \end{bmatrix};\ J_3 = \begin{bmatrix} \frac{\partial y^1}{\partial u^1} & \frac{\partial y^1}{\partial u^2} \\ \frac{\partial y^2}{\partial u^1} & \frac{\partial y^2}{\partial u^2} \end{bmatrix}$$

are called permuted submatrices of J. This requirement ensures that it is possible to solve two equations in Eq. (6.2) for u^1 and u^2 in terms of some pair of variables y^i, and the substitution of these solutions in the remaining equation leads to an equation of the form $y^3 = y^3\left(y^1; y^2\right)$. A representation of the form [Eq. (6.2)] is said to be an allowable representation, and parameters $u^1; u^2$ are frequently called curvilinear co-ordinates.

Note that, if any two determinants formed from the matrix [Eq. (6.3)] vanish identically, then the third one also vanishes, provided that the surface S is not a plane parallel to one of the co-ordinate planes.

6.1.1 Nature of Surface Co-ordinates

Since u^1 and u^2 are independent variables, the locus defined by Eq. (6.2) in two-dimensional, and these equations give the co-ordinates y^i of a point on the surface when u^1 and u^2 are assigned particular values. This point of view leads one to consider the surface as a two-dimensional manifold S imbedded in a three-dimensional enveloping space E_3. We can also show surface without reference to the surrounding space, and consider parameters u^1 and u^2 as co-ordinate points in the surface. A familiar example of this is the use of the latitude and longitude as co-ordinates of points on the surface of the earth.

The intersection of a pair of co-ordinate curves obtained by setting $u^1 = u^1_0; u^2 = u^2_0$ determines a point P_0. The variables $u^1; u^2$ determining the point of S are called the curvilinear or Gaussian co-ordinates of the surface (Figure 6.1).

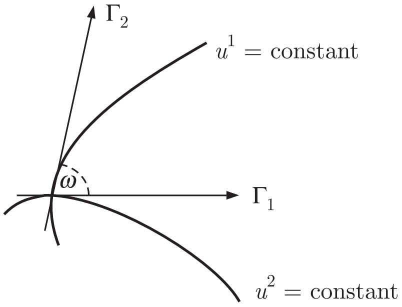

Figure 6.1: Curvilinear co-ordinates.

Obviously, the parametric representation of a surface is not unique and there are infinitely many curvilinear co-ordinate systems which can be used to locate points on a given surface S. For example,

$$x^1 = u^1 + u^2;\; x^2 = u^1 - u^2;\; x^3 = 4u^1u^2$$

and

$$x^1 = v^1 \cosh v^2;\; x^2 = v^1 \sinh v^2;\; x^3 = (v^1)^2$$

represent the same surface

$$(x^1)^2 - (x^2)^2 = x^3; \quad \text{hyperbolic paraboloid.}$$

The two representation may be related by the parametric transformation

$$v^1 = 2\sqrt{u^1u^2};\; v^2 = \frac{1}{2}\log\frac{u^1}{u^2}:$$

Let us examine the geometric significance of these co-ordinates $(u^1; u^2)$. If in Eq. (6.2) of the surface S we put, $u^1 = c$; a constant and let u^2 only vary, we get a locus given by a single variable parameter (one-dimensional manifold). Such a curve

$$S: y^i = y^i\,(c; u^2)\;; i = 1; 2; 3 \qquad (6.4)$$

is called a u^2 curve, characterised by the equation

$$u^2 \text{ curve: } u^1 = \text{ constant.}$$

For different values of u^1, we shall get a family of u^1 curves covering the whole surface. Similarly, for constant value of u^2 we get a family of curves, each called an u^1 curve. In this way we get a family of u^1 curves and a family of u^1 curves, each family covering the whole surface. Thus, u^1 curve is characterised by the equation

$$u^1 \text{ curve: } u^2 = \text{ constant.}$$

Both the families are called family of co-ordinate curves. Together they are called the co-ordinate net. For example, on the surface of revolution $x = u \cos\ ;$ $y = u \sin\ ; z = f(u)$ the parametric curves are

$$u \text{ curve: } x = u \cos c_1; y = u \sin c_1; z = f(u);\ c_1 = \text{ constant.}$$

$$\text{curve: } x = c_2 \cos\ ; y = c_2 \sin\ ; z = f(c_2);\ c_2 = \text{ constant.}$$

EXAMPLE 6.1.1 A surface S is generated by a curve rotating about a fixed straight line. Find a parametric representation of S.

Solution: A surface S generated by a curve rotating about a fixed straight line A is called a surface of revolution. A is called the axis of the surface S. A surface of revolution S; not a cylinder, can be represented in the form (Figure 6.2).

$$\begin{aligned} r &= x^1; x^2; x^3 = x^1\ u^1; u^2\ ; x^2\ u^1; u^2\ ; x^3\ u^1; u^2 \\ &= u^1 \cos u^2; u^1 \sin u^2; f(u^1)\ : \end{aligned}$$

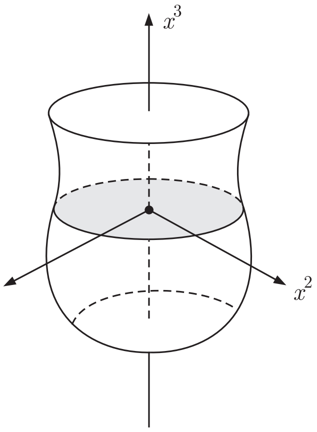

Figure 6.2: Surface of revolution.

This representation is allowable when $u^1 > 0$ and $\frac{df}{du^1} < 1$; i.e. at any point of the rotating curve C which does not lie on the axis A of S or at which the tangent to C is not parallel to A. The curves u^2 = constant are called the meridians, which are the curves of intersection of S and the planes passing through the x^3 axis and the curves u^1 = constant are called parallels, which are the circles parallel to the x^1x^2 plane.

EXAMPLE 6.1.2 Find a parametric representation of a right conoid.

Solution: A surface S is called a right conoid if it can be generated by a moving straight line G intersecting a fixed straight line G_0 so that G and G_0 are always orthogonal. Let us choose the co-ordinates in space so that the straight line G_0 coincides with x^3 axis. Then G is parallel to the x^1x^2 plane, and a right conoid S generated by G can be represented in the form (Figure 6.3)

$$\begin{aligned} r &= \left(x^1; x^2; x^3\right) = \left(x^1\left(u^1; u^2\right); x^2\left(u^1; u^2\right); x^3\left(u^1; u^2\right)\right) \\ &= \left(u^1 \cos u^2; u^1 \sin u^2; f\left(u^2\right)\right): \end{aligned}$$

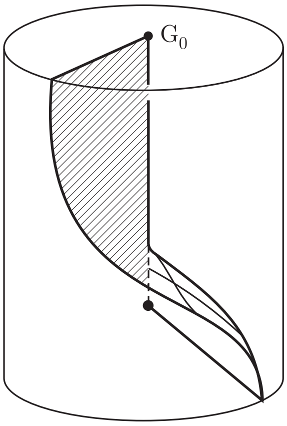

Figure 6.3: Right conoid.

This representation is allowable at any point at which $\frac{df}{du^2} < 1$ and $(u^1)^2 + \left(\frac{df}{du^2}\right)^2 > 0$: The co-ordinate curves u^2 = constant are called straight lines; each value of u^2 corresponds to a certain position of G in space. Any co-ordinate line u^1 = constant is the locus of all points of S which are at the same distance from G_0:

If in particular f is a linear function of u^2, say $f = cu^2 + d$, the corresponding right conoid is called a helicoid. In this case, the curves u^1 = constant are circular helices.

EXAMPLE 6.1.3 Find a parametric representation of a general helicoid.

Solution: A helicoid is a surface generated by the screw motion of a curve about a fixed line, the axis. The various positions of the generating curve thus obtained by first translating it through a distance a parallel to the axis and then rotating it through an angle u^2 about the axis, where $\frac{a}{u^2}$ has a constant value. The constant $\frac{2\pi a}{u^2}$ is the pitch of the helicoids, being the distance translated in one complete revolution. If a twisted curve C rotates about a fixed axis A and, at the same time, is displaced parallel to A so that the velocity of displacement is always proportional to the angular velocity of rotation, then C generates a surface S which is called general helicoid. The intersection M of a helicoid S and a plane passing through the axis A of S is called a meridian of S. The meridians are congruent curves. Consequently, any general helicoid can be generated by a plane curve. We choose as axis A the x^3 axis of the co-ordinate space. Then M can be represented in the form $x^3 = f(u^1)$; where u^1 is the distance of the points of M from A. We assume that at the start of the motion M lies in the x^1x^2 plane. Let u^2 denote the angle of rotation. The displacement of M has the direction of the x^3 axis and is proportional to u^2: We may thus represent S in the form

$$\begin{aligned} r &= \left(x^1; x^2; x^3\right) = \left(x^1\left(u^1; u^2\right); x^2\left(u^1; u^2\right); x^3\left(u^1; u^2\right)\right) \\ &= \left(u^1 \cos u^2; u^1 \sin u^2; cu^2 + f(u^1)\right): \end{aligned}$$

where C is a constant, and $2c$ is called the pitch of the helicoidal motion. If $c \neq 0$ this local representation is allowable at any point at which the tangent to M is not parallel to A. The curves u^1 = constant are circular helices; the curves u^2 = constant are the meridians of the helicoid.

EXAMPLE 6.1.4 Find the parametric representation of the sphere.

Solution: Here we are dealing with a surface of a sphere of radius, and that three mutually perpendicular diameters are chosen as co-ordinate axes. The sphere of radius r with centre at (0, 0, 0) can be represented in the form Eq. (6.1) by $\sum_{i=1}^{n} (x^i)^2 - r^2 = 0$. The latitude u^2 of a point P on the surface may be defined as the inclination of the radius through P to the x^1x^2 plane, and the longitude u^1 as the inclination of the plane containing P and the x^3-axis to the x^3x^1-plane. From this we can obtain a representation of the form (6.2)

$$x^3 = \pm\sqrt{r^2 - (x^1)^2 - (x^2)^2};$$

depending on the choice of the sign this is a representation of one of the two hemispheres $x^3 \geq 0$ and $x^3 \leq 0$. The parametric representation of the sphere under consideration is (Figure 6.4)

$$r = (x^1; x^2; x^3) = (r \cos u^2 \cos u^1; r \cos u^2 \sin u^1; r \sin u^2)$$

or, at length

$$x^1 = r\cos u^2 \cos u^1;\ x^2 = r\cos u^2 \sin u^1;\ x^3 = r\sin u^2$$

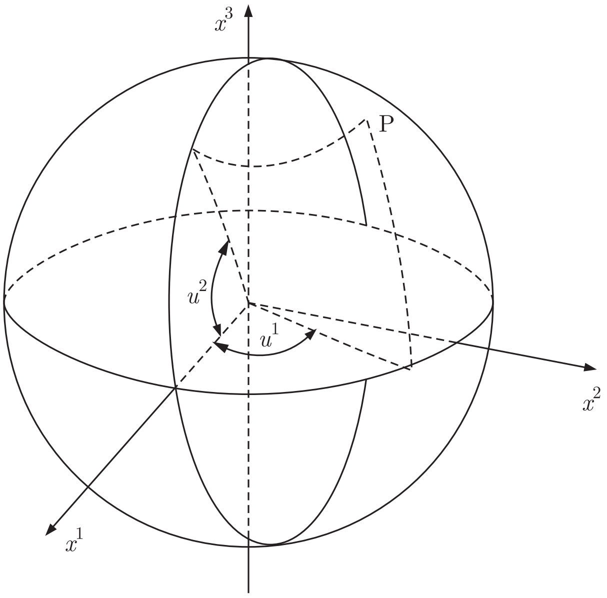

Figure 6.4: Parametric representation of sphere.

where 0 $u^1 < 2$; $\frac{1}{2}$ u^2 $\frac{1}{2}$ are polar angles. This co-ordinate system is used especially in geography for determining the latitude and longitude of points on the globe. Indeed, the co-ordinate curves u^1 = constant and u^2 = constant are the great circles 'meridians' and 'parallels', respectively. The 'equator' is given by $u^2 = 0$ and the 'poles' by $u^2 = \frac{1}{2}$. At the poles the corresponding matrix

$$J = \begin{array}{ccc} r\cos u^2 \sin u^1 & r\cos u^2 \cos u^1 & 0 \\ r\sin u^2 \cos u^1 & r \quad \sin u^2 \sin u^1 & r\cos u^2 \end{array}$$

is of rank 1, i.e. these points are singular with respect to the representation. Every co-ordinate curve u^1 = const. Passes through these points, are the curves $u^2 = \frac{1}{2}$ degenerateinto points. If these two systems of curves cut each other at right angles, we say the parametric curves are orthogonal.

EXAMPLE 6.1.5 Find the parametric representation of a cone of revolution.

Solution: A cone of revolution with apex at (0, 0, 0) and with x^3 axis as axis of revolution can be represented in the form

$$a^2 \ (x^1)^2 + (x^2)^2 \qquad (x^3)^2 = 0$$

The resulting representation of the form

$$x^3 = a\sqrt{(x^1)^2 + (x^2)^2}$$

represents one of the two portions $x^3 \quad 0$ and $x^3 \quad 0$ of this Cone, depending on the choice of the sign of the square root. Now,

$$r = (x^1; x^2; x^3) = (u^1 \cos u^2; u^1 \sin u^2; au^1)$$

is the parametric representation (Figure 6.5) of the cone.

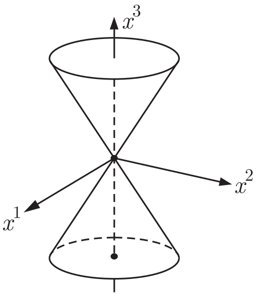

Figure 6.5: Cone of revolution.

The curve u^1 = constant are circles parallel to x_1x_2 plane while the curves u^2 = constant are the generating straight lines of the cone. The corresponding matrix

$$J = \begin{matrix} \cos u^2 & \sin u^2 & a \\ u^1 \sin u^2 & u^1 \cos u^2 & 0 \end{matrix}$$

is of rank 1 at $u^1 = 0$; the apex is a singular point of the cone.

EXAMPLE 6.1.6 Find the parametric representation of an elliptical helix.

Solution: An elliptical helix (Figure 6.6) is a helix lying on an elliptical cylinder $\frac{(x^1)^2}{a^2} + \frac{(x^2)^2}{b^2} = 1$ in the $x^1x^2x^3$ space. The parametric representation of the elliptical helix under consideration is given by

$$r = (x^1; x^2; x^3) = (a \cos t; b \sin t; ct)$$

where c is defined as pitch. At length

$$x^1 = a \cos t; x^2 b \sin t; x^3 = ct:$$

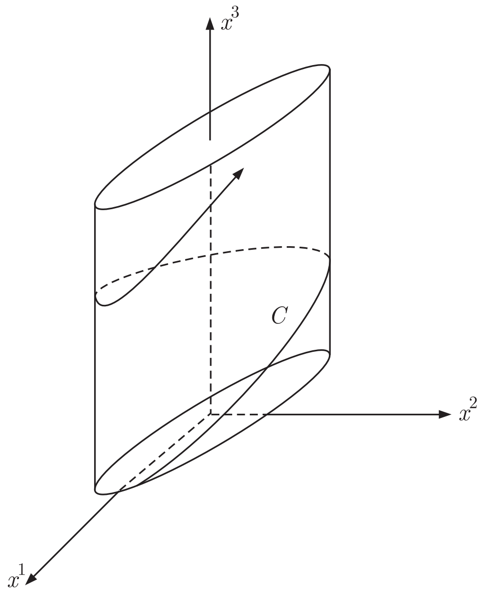

Figure 6.6: Elliptical helix.

6.1.2 Regular Surfaces

Surfaces are generally encountered in the calculus in the form $z = F(x, y)$; i.e. as graphs of two variable functions in three-dimensional space. A surface S in E^2 is the image of a C^3 vector function,

$$r(x^1, x^2) = [f(x^1, x^2), g(x^1, x^2), h(x^1, x^2)],$$

which maps some region V of E^2 into E^3. The co-ordinate breakdown of the mapping r,

$$x = f(x^1, x^2) \quad y = g(x^1, x^2) \quad z = h(x^1, x^2) \tag{6.5}$$

is called the Gaussian form or representation of S. Point P is a regular point of S if $\frac{\partial r}{\partial x^1} \times \frac{\partial r}{\partial x^2} \neq 0$ at P^0, otherwise, P is a singular point. If every point of S is a regular point, then S is a regular surface.

EXAMPLE 6.1.7 Show that $x^1 = u^1 \cos u^2$, $x^2 = u^1 \sin u^2$, $x^3 = u^1 u^2$ is a regular surface.

Solution: The condition which is necessary for

$$r = \left(x^1, x^2, x^3\right) = \left(x^1\left(u^1, u^2\right), x^2\left(u^1, u^2\right), x^3\left(u^1, u^2\right)\right)$$

to be a surface is that $r_{u^1} \times r_{u^2} \neq 0$. Here,

$$r = \left(x^1; x^2; x^3\right) = \left(u^1\cos u^2; u^1\sin u^2; u^1u^2\right):$$

Hence,
$$r_{u^1} = \left(\frac{\partial x^1}{\partial u^1}; \frac{\partial x^2}{\partial u^1}; \frac{\partial x^3}{\partial u^1}\right) = \left(\cos u^2; \sin u^2; u^2\right)$$

$$r_{u^2} = \left(\frac{\partial x^1}{\partial u^2}; \frac{\partial x^2}{\partial u^2}; \frac{\partial x^3}{\partial u^2}\right) = \left(-u^1\sin u^2; u^1\cos u^2; u^1\right)$$

$$\Rightarrow r_{u^1} \times r_{u^2} = \left(u^1\sin u^2 - u^1u^2\cos u^2; u^1u^2\sin u^2 + u^1\cos u^2; u^1\right) \neq 0:$$

Therefore, $x^1 = u^1\cos u^2$, $x^2 = u^1\sin u^2$, $x^3 = u^1u^2$ is a regular surface.

6.2 Intrinsic Geometry

In Section 6.1 we see that the properties of surfaces that can be described without reference to the space in which the surface is imbedded are termed as intrinsic properties. A study of intrinsic properties is made to depend on a certain quadratic differential form describing the metric character of the surface. A transformation T of space co-ordinate from one system x^i to another system $\bar{x}^i$ will be written as

$$T: x^i = x^i\left(\bar{x}^1; \bar{x}^2; \bar{x}^3\right) \tag{6.6}$$

a transformation of Gaussian surface co-ordinates will be

$$u^\alpha = u^\alpha\left(\bar{u}^1; \bar{u}^2\right): \tag{6.7}$$

We will now determine the element of arc of such a curve. Suppose that all functions appearing here are of class C^2 in the region of their definition. Let P and Q be two neighbouring points on the surface S with co-ordinates u^α and $u^\alpha + du^\alpha$, respectively, then from Eq. (6.2), we get

$$y^i = y^i\left(u^1; u^2\right) \text{ so that } dy^i = \frac{\partial y^i}{\partial u^\alpha}du^\alpha \tag{6.8}$$

where the y^i are the orthogonal Cartesian co-ordinates covering the space E^3 in which the surface S is imbedded, and a curve C on S defined by

$$C: u^\alpha = u^\alpha(t);\ t_1 \leq t \leq t_2;\ \text{so that, } du^\alpha = \frac{du^\alpha}{dt}dt; \tag{6.9}$$

where the u^αs are the Gaussian co-ordinates covering S. Viewed from the surrounding space, the curve defined by Eq. (6.9) is a curve in a three-dimensional Euclidean space and its element of arc is given by the formula

$$
\begin{aligned}
ds^2 = dy^i dy^i &= \frac{\partial y^i}{\partial u^\alpha}\frac{\partial y^i}{\partial u^\beta} du^\alpha du^\beta \\
&= a_{\alpha\beta}\, du^\alpha du^\beta \text{ (say); where } a_{\alpha\beta} = \frac{\partial y^i}{\partial u^\alpha}\frac{\partial y^i}{\partial u^\beta}. \qquad (6.10)
\end{aligned}
$$

The expression for ds^2, namely Eq. (6.10), is the square of the linear element of C lying on the surface S, and the right-hand member of Eq. (6.10), i.e. $a_{\alpha\beta}\, du^\alpha du^\beta$ is called the metric form or first fundamental quadratic form of the surface. $a_{\alpha\beta}$s are functions of u's and is called the coefficients of the first fundamental form. The first fundamental form of a regular surface is positive definite. This form (6.10) enable us to measure arc lengths, angles, and areas on a surface; so it defines a metric on the surface.

The length of arc of the curve defined by Eq. (6.7) is given by

$$
S = \int_{t_1}^{t_2} \sqrt{a_{\alpha\beta}\, \dot{u}^\alpha \dot{u}^\beta}\, dt; \text{ where } \dot{u}^\alpha = \frac{du^\alpha}{dt}. \qquad (6.11)
$$

Since in a non-trivial case $ds^2 > 0$, it follows at once from Eq. (6.10) upon setting. u^2 = constant and u^1 = constant in turn, that

$$
ds^2_{(1)} = a_{11}\left(du^1\right)^2 \text{ and } ds^2_{(2)} = a_{22}\left(du^2\right)^2.
$$

Thus, a_{11} and a_{22} are positive functions of u^1 and u^2.

Theorem 6.2.1 Show that $a_{\alpha\beta}$ is a symmetric covariant tensor of order two.

Proof: Consider a transformation of surface co-ordinates

$$
u^\alpha = u^\alpha\left(\bar{u}^1, \bar{u}^2\right) \text{ so that } du^\alpha = \frac{\partial u^\alpha}{\partial \bar{u}^\beta} d\bar{u}^\beta \qquad (6.12)
$$

with the Jacobian $J = \left|\frac{\partial u^\alpha}{\partial \bar{u}^\beta}\right| \neq 0$. The square of the linear element, i.e. ds^2 of C lying on the surface S, is given by the expression

$$
ds^2 = a_{\alpha\beta}\, du^\alpha du^\beta = \frac{\partial y^i}{\partial u^\alpha}\frac{\partial y^i}{\partial u^\beta} du^\alpha du^\beta.
$$

Also

$$
ds^2 = \bar{a}_{\gamma\delta}\, d\bar{u}^\gamma d\bar{u}^\delta = \frac{\partial y^i}{\partial \bar{u}^\gamma}\frac{\partial y^i}{\partial \bar{u}^\delta} d\bar{u}^\gamma d\bar{u}^\delta.
$$

Due to invariant property, we get

$$\frac{\partial y^i}{\partial u}\frac{\partial y^i}{\partial u} du\ du = \frac{\partial y^i}{\partial \bar u}\frac{\partial y^i}{\partial \bar u} d\bar u\ d\bar u$$

or

$$\frac{\partial y^i}{\partial u}\frac{\partial y^i}{\partial u} du\ du = \frac{\partial y^i}{\partial \bar u}\frac{\partial y^i}{\partial \bar u}\frac{\partial \bar u}{\partial u}\frac{\partial \bar u}{\partial u} du\ du$$

or

$$\frac{\partial y^i}{\partial u}\frac{\partial y^i}{\partial u} \quad \frac{\partial y^i}{\partial \bar u}\frac{\partial y^i}{\partial \bar u}\frac{\partial \bar u}{\partial u}\frac{\partial \bar u}{\partial u} \quad du\ du = 0$$

or

$$a \quad \bar a \ \frac{\partial \bar u}{\partial u}\frac{\partial \bar u}{\partial u} \ du\ du = 0$$

Since du ; du are arbitrary tensors, we have

$$a \quad \bar a \ \frac{\partial \bar u}{\partial u}\frac{\partial \bar u}{\partial u} = 0 \ ) \ \bar a = \frac{\partial u}{\partial \bar u}\frac{\partial u}{\partial \bar u} a \ ;$$

which shows that a is a tensor of order two. Now, we have to show that a is symmetric tensor. Let

$$a = \frac{1}{2}(a \quad + a \quad) + \frac{1}{2}(a \quad a \quad)$$

$$= A \quad + B \quad ; \text{ say}$$

where

$$A = \frac{1}{2}(a \quad a \quad) \text{ and } B = \frac{1}{2}(a \quad + a \quad) :$$

It is clear that A is symmetric and B is skew symmetric. Now,

$$a \ du\ du = A \ du\ du + B \ du\ du$$

or

$$(a \quad A \quad) du\ du = B \ du\ du \tag{6.13}$$

or

$$(a \quad A \quad) du\ du = \ B \ du\ du \ ; B \ \text{ is antisymmetric}$$

or

$$(a \quad A \quad) du\ du = \ B \ du\ du \ ;$$

interchanging the dummy indices ;

or

$$(a \quad A \quad) du\ du = \ B \ du\ du \ : \tag{6.14}$$

From Eqs. (6.13) and (6.14) we get

$$2B\ du\ du = 0 \Rightarrow B = 0$$

as du and du are arbitrary tensors. Therefore, $a = A$, where a is a symmetric tensor. Thus, we see that the set of quantities a represents a symmetric covariant tensor of rank 2, with respect to the admissible transformation Eq. (6.12) of surface co-ordinates. The tensor a is called the covariant metric tensor of the surface and are also called fundamental magnitudes of the first order.

Definition 6.2.1 Since the form Eq. (6.10) is positive definite, the determinant

$$a = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0$$

and we can define the reciprocal tensor a by

$$a = \frac{\text{cofactor of } a \text{ in } a}{a}$$

with the property $a\ a =\ $. Also, a is a contravariant tensor of order two. The contravariant tensor a is called the contravariant metric tensor. a is called the discriminant of the form.

Since our space admits of arbitrary contravariant vectors, let be any arbitrary contravariant vector defined at any general point on the surface. Then the equation

$$= a \qquad (6.15)$$

is uniquely solvable if $|a| \neq 0$. This shows that the correspondence is one to one. Since here is a covariant vector, we find that our space admits of arbitrary covariant vector. Taking inner product with a of Eq. (6.15) we get

$$a = a\ a = =$$

or

$$a = = |\ |^2. \qquad (6.16)$$

This relation shows that a is that contravariant tensor of second order. Interchanging the dummy indices and in Eq. (6.16) we get

$$a = |\ |^2.$$

Therefore, a is a symmetric tensor.

EXAMPLE 6.2.1 Find the first fundamental form of a plane with respect to polar co-ordinates.

Solution: Let $x^1 = u^1 \cos u^2, x^2 = u^1 \sin u^2$, where x^1, x^2 are Cartesian co-ordinates and u^1, u^2 are polar co-ordinates. Thus, the parametric representation is given by

$$x^1 = u^1 \cos u^2, x^2 = u^1 \sin u^2, x^3 = 0$$

i.e.

$$\mathbf{r} = (x^1, x^2, x^3) = \left(u^1 \cos u^2, u^1 \sin u^2, 0\right).$$

Thus, the coefficients $a_{\alpha\beta}$ of the first fundamental form are given by

$$a_{11} = \left(\frac{\partial x^1}{\partial u^1}\right)^2 + \left(\frac{\partial x^2}{\partial u^1}\right)^2 + \left(\frac{\partial x^3}{\partial u^1}\right)^2 = 1.$$

$$a_{22} = \left(\frac{\partial x^1}{\partial u^2}\right)^2 + \left(\frac{\partial x^2}{\partial u^2}\right)^2 + \left(\frac{\partial x^3}{\partial u^2}\right)^2 = (u^1)^2.$$

$$\begin{aligned} a_{12} &= \frac{\partial x^i}{\partial u^1}\frac{\partial x^i}{\partial u^2} = \frac{\partial x^1}{\partial u^1}\frac{\partial x^1}{\partial u^2} + \frac{\partial x^2}{\partial u^1}\frac{\partial x^2}{\partial u^2} + \frac{\partial x^3}{\partial u^1}\frac{\partial x^3}{\partial u^2} \\ &= \cos u^2 \left(-u^1 \sin u^2\right) + \sin u^2 (u^1 \cos u^2) + 0 = 0 = a_{21}. \end{aligned}$$

Therefore, the first fundamental form for the surface is given by

$$\begin{aligned} ds^2 &= a_{ij}\, du^i du^j = a_{11}(du^1)^2 + 2a_{12}du^1du^2 + a_{22}(du^2)^2 \\ &= (du^1)^2 + (u^1)^2(du^2)^2. \end{aligned}$$

EXAMPLE 6.2.2 Find the first fundamental form of the right helicoid.

Solution: The parametric representation of the right helicoid is given by

$$x^1 = u^1 \cos u^2, x^2 = u^1 \sin u^2, x^3 = cu^2$$

i.e.

$$\mathbf{r} = (x^1, x^2, x^3) = \left(u^1 \cos u^2, u^1 \sin u^2, cu^2\right),$$

where c is a constant. Now,

$$\frac{\partial x^1}{\partial u^1} = \cos u^2, \quad \frac{\partial x^2}{\partial u^1} = \sin u^2, \quad \frac{\partial x^3}{\partial u^1} = 0$$

$$\frac{\partial x^1}{\partial u^2} = -u^1 \sin u^2, \quad \frac{\partial x^2}{\partial u^2} = u^1 \cos u^2, \quad \frac{\partial x^3}{\partial u^2} = c.$$

Thus, the coefficients a_{ij} of the first fundamental form are given by

$$a_{11} = \left(\frac{\partial x^1}{\partial u^1}\right)^2 + \left(\frac{\partial x^2}{\partial u^1}\right)^2 + \left(\frac{\partial x^3}{\partial u^1}\right)^2 = 1:$$

$$a_{22} = \left(\frac{\partial x^1}{\partial u^2}\right)^2 + \left(\frac{\partial x^2}{\partial u^2}\right)^2 + \left(\frac{\partial x^3}{\partial u^2}\right)^2 = (u^1)^2 + c^2:$$

$$\begin{aligned} a_{12} &= \frac{\partial x^i}{\partial u^1}\frac{\partial x^i}{\partial u^2} = \frac{\partial x^1}{\partial u^1}\frac{\partial x^1}{\partial u^2} + \frac{\partial x^2}{\partial u^1}\frac{\partial x^2}{\partial u^2} + \frac{\partial x^3}{\partial u^1}\frac{\partial x^3}{\partial u^2} \\ &= \cos u^2(-u^1 \sin u^2) + \sin u^2(u^1 \cos u^2) + 0 \cdot c = 0 = a_{21}: \end{aligned}$$

Therefore, the first fundamental form for the surface is given by

$$\begin{aligned} ds^2 &= a_{ij}\, du^i du^j = a_{11}(du^1)^2 + 2a_{12}du^1du^2 + a_{22}(du^2)^2 \\ &= (du^1)^2 + [(u^1)^2 + c^2](du^2)^2: \end{aligned}$$

The metric for R^2 corresponding to the right helicoid is non-Euclidean. Now, the parameters $u^1; u^2$ which are actual polar co-ordinates in the x^1x^2 plane of E^3, (Figure 6.7) formerly keep that significance when the plane is considered abstractly as parameter space. This is an instance of the formal use of familiar co-ordinate system in non-Euclidean space.

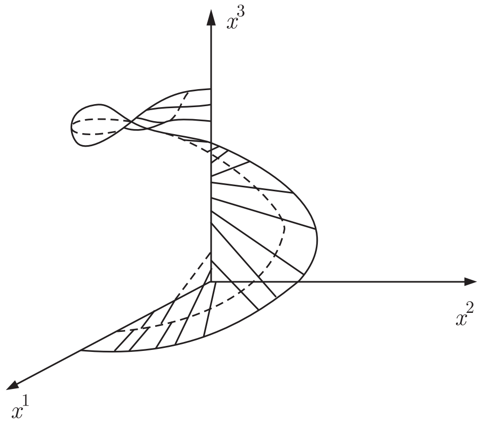

Figure 6.7: Right helicoid.

EXAMPLE 6.2.3 Find the first fundamental form for any surface of revolution, and specialise to a right circular cone (Figure 6.8).

Solution: The Gaussian form of a surface of revolution about z axis is

$$r(x^1; x^2) = \left(f(x^1)\cos x^2; f(x^1)\sin x^2; g(x^1)\right); \; f(x^1) > 0:$$

From this parametric representation, we get

$$\frac{\partial \mathbf{r}}{\partial x^1} = \left(f'(x^1)\cos x^2, f'(x^1)\sin x^2, g'(x^1)\right); \quad \frac{\partial \mathbf{r}}{\partial x^2} = \left(-f(x^1)\sin x^2, f(x^1)\cos x^2, 0\right).$$

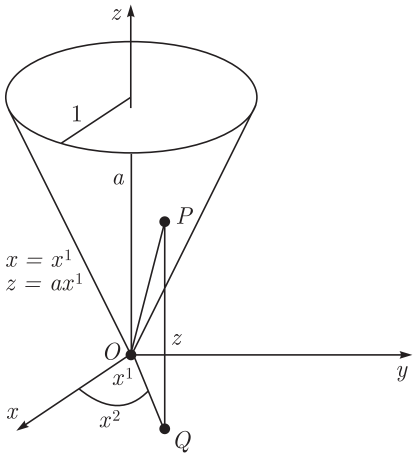

Figure 6.8: Right circular cone.

Thus, the coefficients $a_{\alpha\beta}$ of the first fundamental form are given by

$$a_{11} = \frac{\partial \mathbf{r}}{\partial x^1}\cdot\frac{\partial \mathbf{r}}{\partial x^1} = (f')^2 + (g')^2; \quad a_{22} = \frac{\partial \mathbf{r}}{\partial x^2}\cdot\frac{\partial \mathbf{r}}{\partial x^2} = f^2;$$

$$a_{12} = \frac{\partial \mathbf{r}}{\partial x^1}\cdot\frac{\partial \mathbf{r}}{\partial x^2} = 0 = a_{21}.$$

Therefore, the first fundamental form for the surface is given by

$$\begin{aligned} ds^2 &= a_{11}(dx^1)^2 + 2a_{12}dx^1dx^2 + a_{22}(dx^2)^2 \\ &= \left[(f')^2 + (g')^2\right](dx^1)^1 + f^2(dx^2)^2. \end{aligned}$$

For a right circular cone, $f(x^1) = x^1$ and $g(x^1) = ax^1$; hence,

$$ds^2 = \left[1 + a^2\right](dx^1)^1 + (x^1)^2(dx^2)^2.$$

EXAMPLE 6.2.4 Find the first fundamental form for the catenoid $x = a\cosh x^1$; $z = ax^1$ and compute the length of the curve $x^1 = t, x^2 = t, 0 \le t \le \log(1+\sqrt{2})$.

Solution: For that case, $f(x^1) = a\cosh x^1$; $g(x^1) = ax^1$. Therefore, the first fundamental form for the surface is given by

$$ds^2 = a^2\cosh^2 x^1\left[(dx^1)^2 + (dx^2)^2\right].$$

Thus, the required length of the arc, $0 \le t \le \log(1+\sqrt{2})$ is given by

$$L = a\sqrt{2}\int_0^{\log(1+\sqrt{2})} \cosh t\,dt = a\sqrt{2}\sinh[\log(1+\sqrt{2})] = a\sqrt{2} \text{ units.}$$

EXAMPLE 6.2.5 Find the first fundamental form of the surface is given in Monge's form $x^3 = f(u^1, u^2)$, where the co-ordinates u^1, u^2 may be taken as parameters.

Solution: The parametric representation of the right helicoid is given by

$$x^1 = u^1; \quad x^2 = u^2; \quad x^3 = f(u^1, u^2)$$

i.e.

$$r = (x^1, x^2, x^3) = \left(u^1, u^2, f(u^1, u^2)\right);$$

From this parametric representation, we get

$$\frac{\partial x^1}{\partial u^1} = 1; \quad \frac{\partial x^2}{\partial u^1} = 0; \quad \frac{\partial x^3}{\partial u^1} = \frac{\partial f}{\partial u^1} = f_1$$

$$\frac{\partial x^1}{\partial u^2} = 0; \quad \frac{\partial x^2}{\partial u^2} = 1; \quad \frac{\partial x^3}{\partial u^2} = \frac{\partial f}{\partial u^2} = f_2;$$

where $\frac{\partial f}{\partial u^i} = f_i$. Thus, the coefficients $a_{\alpha\beta}$ of the first fundamental form are given by

$$a_{11} = \frac{\partial x^i}{\partial u^1}\frac{\partial x^i}{\partial u^1} = \left(\frac{\partial x^1}{\partial u^1}\right)^2 + \left(\frac{\partial x^2}{\partial u^1}\right)^2 + \left(\frac{\partial x^3}{\partial u^1}\right)^2 = 1 + f_1^2.$$

$$a_{22} = \frac{\partial x^i}{\partial u^2}\frac{\partial x^i}{\partial u^2} = \left(\frac{\partial x^1}{\partial u^2}\right)^2 + \left(\frac{\partial x^2}{\partial u^2}\right)^2 + \left(\frac{\partial x^3}{\partial u^2}\right)^2 = 1 + f_2^2.$$

$$a_{12} = \frac{\partial x^i}{\partial u^1}\frac{\partial x^i}{\partial u^2} = \frac{\partial x^1}{\partial u^1}\frac{\partial x^1}{\partial u^2} + \frac{\partial x^2}{\partial u^1}\frac{\partial x^2}{\partial u^2} + \frac{\partial x^3}{\partial u^1}\frac{\partial x^3}{\partial u^2} = f_1 f_2 = a_{21}.$$

Therefore, the first fundamental form for the surface is given by

$$ds^2 = a_{ij}\,du^i du^j = a_{11}(du^1)^2 + 2a_{12}du^1du^2 + a_{22}(du^2)^2$$
$$= (1 + f_1^2)(du^1)^1 + 2f_1 f_2 du^1du^2 + [1 + f_2^2](du^2)^2.$$

Note 6.2.1 Along with the first fundamental form, tensor calculus enters the picture. For the intrinsic properties of a particular surface S in E^3 (the properties defined by measurements of distance on the surface) are all implicit in Eq. (6.10), which can be interpreted as a particular Riemannian metrication of the parameter plane. Thus, the study of intrinsic properties of surfaces becomes the tensor analysis of Riemannian matrices in R^2 and this may be conducted without any reference of E^3 whatever. Observe that the metrices under consideration will all be positive definite but not necessarily Euclidean. Accordingly, we shall drop the designation E^2 for the parameter plane, which shall henceforth be referred to general co-ordinates.

6.2.1 Angle Between Two Intersecting Curves

Here, we have to find the angle between two intersecting curves on a surface. The equation of a curve C drawn on the surface S can be written in the form

$$C: u^\alpha = u^\alpha(t); t_1 \leq t \leq t_2. \tag{6.17}$$

Since the $u^\alpha(t)$ is assumed to be of class C^2, the curve C has a continuously turning tangent. Let C_1 and C_2 be two such curves intersecting at the point P of S. We take the equations of S, referred to orthogonal Cartesian axes y^i, in the form

$$y^i = y^i(u_1; u_2); \ i = 1; 2; 3 \tag{6.18}$$

and denote the direction cosines of the tangent lies to C_1 and C_2 at P by ξ^i and η^i, respectively, then

$$\xi^i = \frac{\partial y^i}{\partial u^\alpha}\frac{du^\alpha}{dS_{(1)}} = \frac{dy^i}{dS_{(1)}} \text{ and } \eta^i = \frac{\partial y^i}{\partial u^\alpha}\frac{du^\alpha}{dS_{(2)}} = \frac{dy^i}{dS_{(2)}};$$

where the subscripts 1 and 2 refer to the elements of arc of C_1 and C_2, respectively. We can write the unit vectors in the directions of the tangents C_1 and C_2 as

$$\lambda^\alpha = \frac{du^\alpha}{dS_{(1)}}; \quad \mu^\alpha = \frac{du^\alpha}{dS_{(2)}} \text{ and } \xi^i = \frac{\partial y^i}{\partial u^\alpha}\lambda^\alpha; \quad \eta^i = \frac{\partial y^i}{\partial u^\alpha}\mu^\alpha. \tag{6.19}$$

The cosine of the angle between C_1 and C_2, calculated by a geometer in the enveloping space E^3, (Figure 6.9) is

$$\begin{aligned} \cos\theta &= \xi^i \eta^i = \frac{\partial y^i}{\partial u^\alpha}\frac{\partial y^i}{\partial u^\beta}\lambda^\alpha\mu^\beta \\ &= a_{\alpha\beta}\lambda^\alpha\mu^\beta = a_{\alpha\beta}\frac{du^\alpha}{dS_{(1)}}\frac{du^\beta}{dS_{(2)}}. \end{aligned} \tag{6.20}$$

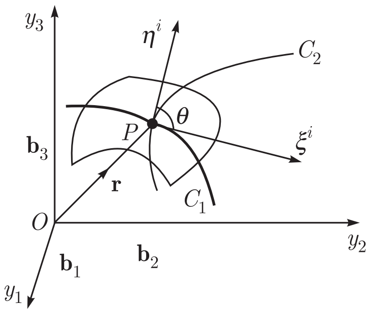

Figure 6.9: Angle between two intersecting curves.

In two-dimensional manifold, the skew-symmetric e-systems can be defined by formula (1.5) as $e_{11} = e_{22} = e^{11} = e^{22} = 0$; $e^{12} = \; e^{21} = e_{12} = \; e_{21} = 1$ and since these systems are relative tensors, the expressions

$$" \;\; = \sqrt{a}\, e \quad \text{and} \; " \;\; = \frac{1}{\sqrt{a}} e \qquad (6.21)$$

are absolute tensors. Let be the angle between two unit vectors ; , then,

$$\sin^2 \;\; = 1 \quad \cos^2 \;\; = \;\; {}^{i}\;{}_{i} \;\; {}^{i}\;{}_{i} \;\; {}^{i}\;{}_{i} \; {}^{2}$$

$$= \;\; {}^{1}_{2} \;\; {}^{1}_{2} \; {}^{2} + {}^{2}_{3} \;\; {}^{2}_{3} \; {}^{2} + {}^{3}_{1} \;\; {}^{3}_{1} \; {}^{2} \text{; by Lagrange identity}$$

$$= \begin{bmatrix} \frac{\partial y^1}{\partial u^1} & \frac{\partial y^1}{\partial u^2} \\ \frac{\partial y^2}{\partial u^1} & \frac{\partial y^2}{\partial u^2} \end{bmatrix}^2 + \begin{bmatrix} \frac{\partial y^2}{\partial u^1} & \frac{\partial y^2}{\partial u^2} \\ \frac{\partial y^3}{\partial u^1} & \frac{\partial y^3}{\partial u^2} \end{bmatrix}^2 + \begin{bmatrix} \frac{\partial y^3}{\partial u^1} & \frac{\partial y^3}{\partial u^2} \\ \frac{\partial y^1}{\partial u^1} & \frac{\partial y^1}{\partial u^2} \end{bmatrix}^2 \;\; {}^{1}_{2} \;\; {}^{1}_{2} \; {}^{2}$$

$$= J_1^2 + J_2^2 + J_3^2 \;\; {}^{1}\;{}^{2} \;\; {}^{2}\;{}^{1} \; {}^{2} = a \;\; {}^{1}\;{}^{2} \;\; {}^{2}\;{}^{1} \; {}^{2}$$

or

$$\sin \;\; = \sqrt{a} \;\; {}^{1}\;{}^{2} \;\; {}^{2}\;{}^{1} = \sqrt{a}\, e \;\;\;\; = " \;\;\;\; : \qquad (6.22)$$

Equation (6.22) " = sin ; is numerically equal to the area of the parallelogram constructed on the unit vectors and . It follows from this result that a necessary and sufficient condition for the orthonormality of two surface unit vectors and is " = 1:

If the curves C_1 and C_2 are orthogonal,

$$a \;\;\;\; = 0 \qquad (6.23)$$

Also, if the surface vectors and are taken along the co-ordinate curves

$$^{1} = \frac{1}{\sqrt{a_{11}}}; \;\; ^{2} = 0; \;\; ^{1} = 0; \text{ and } \; ^{2} = \frac{1}{\sqrt{a_{22}}};$$

then from Eq. (6.23) the co-ordinate curves will form an orthogonal net if and only if $a_{12} = 0$ at every point of the surface. If in particular, the vectors ; are taken along the parametric curves, then, for the u^0-curve, $du^2 = 0$. Consequently, Eq. (6.10) reduces to $dS^2_{(1)} = a_{11} \; du^1 \; ^2$ and hence, the unit vector $_{(1)}$ along u^1 curve is

$$_{(1)} = \frac{du^1}{dS_{(1)}}; \frac{du^2}{dS_{(1)}} \;\; = \;\; \frac{1}{\sqrt{a_{11}}}; 0 \;\; = \frac{1}{\sqrt{a_{11}}} \; _{(1)}:$$

Similarly, the unit vector $_{(2)}$ along u^2 curve is

$$_{(2)} = \frac{du^1}{dS_{(2)}}; \frac{du^2}{dS_{(2)}} \;\; = \;\; \frac{1}{\sqrt{a_{22}}}; 0 \;\; = \frac{1}{\sqrt{a_{22}}} \; _{(2)}:$$

EXAMPLE 6.2.6 Show that the arc length lying on the plane $z = 0$ is the same in both the systems.

Solution: Let P and Q be two points on the plane $z = 0$, whose co-ordinates are say $P\left(x^i\right) = P\left(x^1, x^2, x^3\right)$ and $Q\left(x^i + dx^i\right) = Q\left(x^1 + dx^1, x^2 + dx^2, x^3 + dx^3\right)$. Now, the symmetric tensors a_{ij} are given by

$$a_{11} = \frac{\partial x^i}{\partial u^1}\frac{\partial x^i}{\partial u^1} = \left(\frac{\partial x^1}{\partial u^1}\right)^2 + \left(\frac{\partial x^2}{\partial u^1}\right)^2 + \left(\frac{\partial x^3}{\partial u^1}\right)^2 = 1.$$

$$a_{12} = \frac{\partial x^i}{\partial u^1}\frac{\partial x^i}{\partial u^2} = \frac{\partial x^1}{\partial u^1}\frac{\partial x^1}{\partial u^2} + \frac{\partial x^2}{\partial u^1}\frac{\partial x^2}{\partial u^2} + \frac{\partial x^3}{\partial u^1}\frac{\partial x^3}{\partial u^2} = 0 = a_{21}.$$

$$a_{22} = \frac{\partial x^i}{\partial u^2}\frac{\partial x^i}{\partial u^2} = \left(\frac{\partial x^1}{\partial u^2}\right)^2 + \left(\frac{\partial x^2}{\partial u^2}\right)^2 + \left(\frac{\partial x^3}{\partial u^2}\right)^2 = 1.$$

Now, the square of the linear element is given by

$$\begin{aligned} ds^2 &= a_{\alpha\beta}\, du^\alpha du^\beta = a_{11}\left(du^1\right)^2 + 2a_{12}du^1du^2 + a_{22}\left(du^2\right)^2 \\ &= a_{11}\left(dx^1\right)^2 + 2a_{12}dx^1dx^2 + a_{22}\left(dx^2\right)^2; \\ &\qquad \text{as } u^1 = x^1;\, u^2 = x^2;\, u^3 = 0 \\ &= 1\left(dx^1\right)^2 + 1\left(dx^2\right)^2 = \left(dx^1\right)^2 + \left(dx^2\right)^2 = dx^i dx^i. \end{aligned}$$

Therefore, the arc length lying on the plane $z = 0$ is the same in both the systems.

EXAMPLE 6.2.7 If ω be the angle between the parametric curves lying on a surface, immersed in E_3, show that

$$\cos\omega = \frac{a_{12}}{\sqrt{a_{11}}\sqrt{a_{22}}} \text{ and } \sin\omega = \frac{\sqrt{a}}{\sqrt{a_{11}}\sqrt{a_{22}}}$$

and hence, show that the parametric curves on a surface are orthogonal if $a_{12} = 0$.

Solution: Since ω be the angle between the parametric curves, so,

$$\begin{aligned} \cos\omega &= a_{\alpha\beta}\lambda^\alpha_{(1)}\lambda^\beta_{(2)} = a_{11}\lambda^1_{(1)}\lambda^1_{(2)} + a_{22}\lambda^2_{(1)}\lambda^2_{(2)} + a_{12}\lambda^1_{(1)}\lambda^2_{(2)} + a_{21}\lambda^2_{(1)}\lambda^1_{(2)} \\ &= a_{11}\left(\frac{1}{\sqrt{a_{11}}}\right)(0) + a_{22}(0)\left(\frac{1}{\sqrt{a_{22}}}\right) + a_{12}\left(\frac{1}{\sqrt{a_{11}}}\right)\left(\frac{1}{\sqrt{a_{22}}}\right) + a_{21}(0)(0) \\ &= \frac{a_{12}}{\sqrt{a_{11}}\sqrt{a_{22}}}. \end{aligned}$$

Using the identity $\cos^2\theta = 1 - \sin^2\theta$ we get

$$\sin^2\theta = 1 - \cos^2\theta = 1 - \frac{a_{12}^2}{a_{11}a_{22}} = \frac{a_{11}a_{22} - a_{12}^2}{a_{11}a_{22}} = \frac{a}{a_{11}a_{22}}$$

$$\Rightarrow \sin\theta = \frac{\sqrt{a}}{\sqrt{a_{11}}\sqrt{a_{22}}}.$$

If the parametric curves on the surface S are orthogonal, i.e. if $\theta = \frac{\pi}{2}$; then,

$$\frac{\sqrt{a}}{\sqrt{a_{11}}\sqrt{a_{22}}} = \sin\frac{\pi}{2} = 1$$

so that a_{11} and $a_{22} \neq 0$ and so $a > 0$ and so,

$$\cos\frac{\pi}{2} = 0 = \frac{a_{12}}{\sqrt{a_{11}}\sqrt{a_{22}}} \Rightarrow a_{12} = 0.$$

EXAMPLE 6.2.8 Prove that the parametric curves on a surface given by $x^1 = a\sin u\cos v;\ x^2 = a\sin u\sin v;\ x^3 = a\cos u$ form an orthogonal system.

Solution: For the given surface, the symmetric covariant tensor $a_{\alpha\beta}$ of order two are given by

$$a_{11} = \left(\frac{\partial x^1}{\partial u}\right)^2 + \left(\frac{\partial x^2}{\partial u}\right)^2 + \left(\frac{\partial x^3}{\partial u}\right)^2$$

$$= (a\cos u\cos v)^2 + (a\cos u\sin v)^2 + (-a\sin u)^2 = a^2.$$

$$a_{22} = \left(\frac{\partial x^1}{\partial v}\right)^2 + \left(\frac{\partial x^2}{\partial v}\right)^2 + \left(\frac{\partial x^3}{\partial v}\right)^2$$

$$= (-a\sin u\sin v)^2 + (a\sin u\cos v)^2 + 0^2 = a^2\sin^2 u$$

$$a_{12} = \frac{\partial x^1}{\partial u}\frac{\partial x^1}{\partial v} + \frac{\partial x^2}{\partial u}\frac{\partial x^2}{\partial v} + \frac{\partial x^3}{\partial u}\frac{\partial x^3}{\partial v}$$

$$= a\cos u\cos v\,(-a\sin u\sin v) + a\cos u\sin v\,(a\sin u\cos v) + (-a\sin u)\,0 = 0 = a_{21}.$$

Since $a_{12} = 0$, so the co-ordinate curves are orthogonal. Let λ and μ be taken along the parametric curves. For the u-curve, $dv = 0$ and for the v-curve, $du = 0$. Thus, the unit vector $\lambda_{(1)}$ along the u-curve is

$$\lambda_{(1)} = \left(\frac{du}{ds_{(1)}}; \frac{dv}{ds_{(1)}}\right) = \left(\frac{1}{\sqrt{a_{11}}}; 0\right) = \left(\frac{1}{a}; 0\right)$$

and the unit vector $_{(2)}$ along the v-curve is

$$_{(2)} = \frac{du}{ds_{(2)}}; \frac{dv}{ds_{(2)}} = 0; p\frac{1}{\overline{a_{22}}} = 0; \frac{1}{a\sin u} :$$

If be the angle between the parametric curves, then

$$\cos = a \quad {}_{(1)} \quad {}_{(2)} = a_{11} \, {}^{1}_{(1)} \, {}^{1}_{(2)} + a_{22} \, {}^{2}_{(1)} 1 \, {}^{2}_{(2)} + a_{12} \, {}^{1}_{(1)} \, {}^{2}_{(2)}$$

$$= a^2 \quad \frac{1}{a} \quad 0 + a^2 \sin^2 u \quad 0 \quad \frac{1}{a\sin u} + 0 \quad \frac{1}{a} \quad \frac{1}{a\sin u} + 0 \quad 0 \quad 0$$

$$= 0 \ ) \quad = \frac{}{2}:$$

Thus, the parametric curves on the surface given by $x^1 = a\sin u\cos v; x^2 = a\sin u\sin v;$ $x^3 = a\cos u$ form an orthogonal system.

6.2.2 Element of Surface Area

Let r denotes the position vector of any point P on the surface S, and the b_i are the unit vectors directed along the orthogonal co-ordinate axes Y, then from Eq. (6.18) of the surface S can be written in the vector form as

$$r(u_1; u_2) = b_i y^i (u_1; u_2) :$$

According to the representation of the surface S, it follows that

$$ds^2 = dr \quad dr = \frac{@}{@u} \quad \frac{@}{@u} du \quad du$$

$$= a \quad du \quad du \ ; \quad \text{where}, a \quad = \frac{@}{@u} \quad \frac{@}{@u} :$$

Setting $\frac{@}{@u} = a$; where a_1 and a_2 are obviously tangent vectors to the co-ordinate curves, we see that

$$a_{11} = a_1 \quad a_1; \ a_{12} = a_1 \quad a_2; \ a_{22} = a_2 \quad a_2:$$

In the notation of Eq. (6.19) the space components of a_1 and a_2 are ^{i} and ^{i}, respectively. Let us define an element of area d of the surface S by the formula, (Figure 6.10)

$$d \quad = |a_1 \quad a_2| \, du^1 du^2$$

$$= q\overline{a_{11}a_{22} \quad a_{12}^2} \, du^1 du^2 = p\overline{a} \, du^1 du^2:$$

This formula has widely used to calculate the volume element.

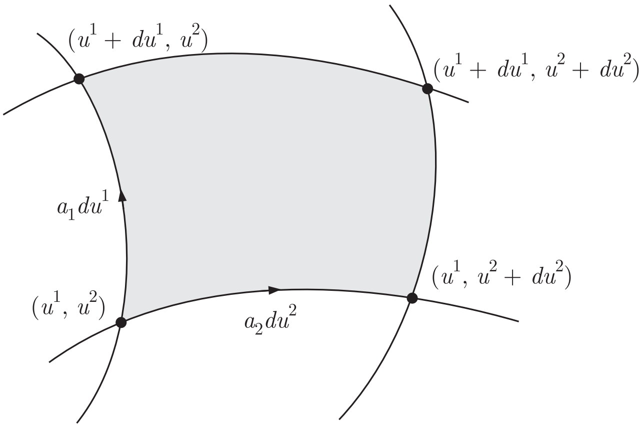

Figure 6.10: Element of area.

EXAMPLE 6.2.9 Find the element of area of the surface of radius r, if the equations of the surface are given in the form

$$y^1 = r\sin u^1\cos u^2;\ y^2 = r\sin u^1\sin u^2;\ y^3 = r\cos u^1;$$

where the y^i are orthogonal Cartesian co-ordinates.

Solution: The parametric representation of the surface of radius r is given by

$$y^1 = r\sin u^1\cos u^2;\ y^2 = r\sin u^1\sin u^2;\ y^3 = r\cos u^1$$

i.e.

$$\mathbf{r} = (y^1; y^2; y^3) = \left(r\sin u^1\cos u^2; r\sin u^1\sin u^2; r\cos u^1\right):$$

From this parametric representation, we get

$$\frac{\partial y^1}{\partial u^1} = r\cos u^1\cos u^2;\ \frac{\partial y^2}{\partial u^1} = r\cos u^1\sin u^2;\ \frac{\partial y^3}{\partial u^1} = -r\sin u^1$$

$$\frac{\partial y^1}{\partial u^2} = -r\sin u^1\sin u^2;\ \frac{\partial y^2}{\partial u^2} = r\sin u^1\cos u^2;\ \frac{\partial y^3}{\partial u^2} = 0:$$

Thus, the coefficients $a_{\alpha\beta}$ of the first fundamental form are given by

$$\begin{aligned} a_{11} &= \left(\frac{\partial y^1}{\partial u^1}\right)^2 + \left(\frac{\partial y^2}{\partial u^1}\right)^2 + \left(\frac{\partial y^3}{\partial u^1}\right)^2 \\ &= \left(r\cos u^1\cos u^2\right)^2 + \left(r\cos u^1\sin u^2\right)^2 + \left(-r\sin u^1\right)^2 = r^2: \\ a_{22} &= \left(\frac{\partial y^1}{\partial u^2}\right)^2 + \left(\frac{\partial y^2}{\partial u^2}\right)^2 + \left(\frac{\partial y^3}{\partial u^2}\right)^2 = r^2\sin^2 u^1: \end{aligned}$$

$$a_{12} = \frac{\partial y^1}{\partial u^1}\frac{\partial y^1}{\partial u^2} + \frac{\partial y^2}{\partial u^1}\frac{\partial y^2}{\partial u^2} + \frac{\partial y^3}{\partial u^1}\frac{\partial y^3}{\partial u^2}$$
$$= \left(r\cos u^1 \cos u^2\right)\left(-r\sin u^1 \sin u^2\right) + \left(r\cos u^1 \sin u^2\right)\left(r\sin u^1 \cos u^2\right)$$
$$+ \left(-r\sin u^1\right) 0 = 0 = a_{21}.$$

Therefore,

$$a = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = \begin{vmatrix} r^2 & 0 \\ 0 & r^2\sin^2 u^1 \end{vmatrix} = r^4\sin^2 u^1.$$

The element of area $d\sigma$ of the surface S of radius r by the formula,

$$d\sigma = \sqrt{a}\, du^1\, du^2 = r^2 \sin u^1 du^1 du^2.$$

6.3 Geodesic on a Surface

Here, we will discuss the problem of finding curves of minimum length or shortest arc joining a pair of given points on the surface. We will carry out our calculation for the case of the n-dimensional Riemannian manifolds, since our results will be of interest not only in connection with the geometry of surfaces but also in the study of dynamical trajectories. Obviously this is a problem of calculus of variation, first considered by J. Bernoulli.

Let metric properties of the n-dimensional manifold R^n be determined by

$$ds^2 = g_{ij}\, du^i du^j;\ ;\ i, j = 1, 2, \ldots, n; \tag{6.24}$$

where $g_{ij} = g_{ji}$ are specified functions on the variables u^i. Let us suppose that the form Eq. (6.24) is positive definite and the functions g_{ij} are of class C^2. Let C be a curve given by (Figure 6.11)

$$C: u^i = u^i(t);\ t_1 \le t \le t_2$$

and the length of the curve between two points P and Q on it be given by

$$S = \int_{t_1}^{t_2} \sqrt{g_{\alpha\beta}\, \dot{u}^\alpha \dot{u}^\beta}\, dt;\ ;\ \alpha, \beta = 1, 2, \cdots, n. \tag{6.25}$$

The extremals of the functional Eq. (6.25) will be termed as *geodesics* in R^n. Let $\overline{C}$ be any curve in the neighbourhood of C joining P and Q and let it be given by

$$\overline{C}: \overline{u}^i = u^i(t) + \varepsilon w^i(t);$$

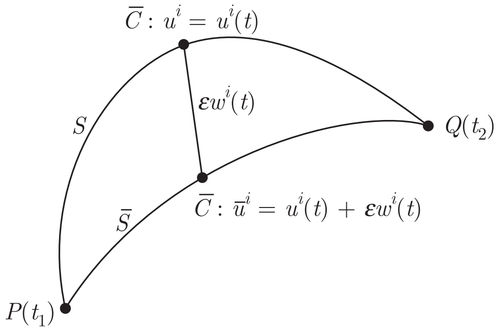
$\overline{C}: u^i = u^i(t)$
$\varepsilon w^i(t)$
S
$Q(t_2)$
$\bar{S}$
$\overline{C}: \bar{u}^i = u^i(t) + \varepsilon w^i(t)$
$P(t_1)$

Figure 6.11: Geodestic on a surface.

where w^i is a function of t such that $w^i(t) = 0$ at $P(t_1)$ and $Q(t_2)$ and ε is a number of infinitesimal order. The arc length between P and Q along $\overline{C}$ is given by

$$\bar{S} = \int_{t_1}^{t_2} \sqrt{g_{\alpha\beta}\dot{\bar{u}}^\alpha \dot{\bar{u}}^\beta}\, dt.$$

Let us consider the integral,

$$I = \int_{t_1}^{t_2} \phi(u^\alpha; \dot{u}^\alpha) dt;$$

where ϕ is a function and

$$\begin{aligned}\bar{I} &= \int_{t_1}^{t_2} \phi(u^\alpha + \varepsilon w^\alpha; \dot{u}^\alpha + \varepsilon \dot{w}^\alpha)\, dt\\ &= \int_{t_1}^{t_2} \phi(u^\alpha; \dot{u}^\alpha)\, dt + \varepsilon \int_{t_1}^{t_2} \left(w^\alpha \frac{\partial \phi}{\partial u^\alpha} + \dot{w}^\alpha \frac{\partial \phi}{\partial \dot{u}^\alpha}\right) dt;\end{aligned}$$

by Taylor's theorem, on neglecting the other terms. Therefore, the increment $\bar{I} - I$ is given by

$$\begin{aligned}\bar{I} - I &= \varepsilon \int_{t_1}^{t_2} \left(w^\alpha \frac{\partial \phi}{\partial u^\alpha} + \dot{w}^\alpha \frac{\partial \phi}{\partial \dot{u}^\alpha}\right) dt; \quad \alpha = 1; 2\\ &= \varepsilon \int_{t_1}^{t_2} w^\alpha \frac{\partial \phi}{\partial u^\alpha} dt + \varepsilon \int_{t_1}^{t_2} \frac{\partial \phi}{\partial \dot{u}^\alpha} \dot{w}^\alpha\, dt;\\ &= \varepsilon \int_{t_1}^{t_2} w^\alpha \frac{\partial \phi}{\partial u^\alpha}\, dt + \varepsilon \left[\frac{\partial \phi}{\partial \dot{u}^\alpha} w^\alpha\right]_{t_1}^{t_2} - \varepsilon \int_{t_1}^{t_2} \frac{d}{dt}\left(\frac{\partial \phi}{\partial \dot{u}^\alpha}\right) w^\alpha\, dt;\\ &= \varepsilon \int_{t_1}^{t_2} \left(\frac{\partial \phi}{\partial u^\alpha} - \frac{d}{dt}\frac{\partial \phi}{\partial \dot{u}^\alpha}\right) w^\alpha\, dt. \qquad (6.26)\end{aligned}$$

If C is a geodesic, then the increment I must be zero for all neighbouring curves through P and Q, i.e. Eq. (6.26) must vanish for all arbitrary values of the vector w along C. Thus,

$$\frac{d}{dt}\ \frac{\partial}{\partial \dot{u}}\qquad \frac{\partial}{\partial u} = 0; \tag{6.27}$$

which is called Euler's or Lagrange's equations for . In our case,

$$(u\ ;\dot{u}\ ) = \sqrt{g\ \ \dot{u}\ \dot{u}}$$

$$\Rightarrow\ \frac{\partial}{\partial u} = \frac{1}{2}\ g\ \ \dot{u}\ \dot{u}^{\ \ 1/2}\ \frac{\partial g}{\partial u}\dot{u}\ \dot{u}$$

and

$$\frac{\partial}{\partial \dot{u}} = \frac{1}{2}\ g\ \ \dot{u}\ \dot{u}^{\ \ 1/2}\ g\ \ \dot{u}\ :$$

Hence, the Euler's equation [Eq. (6.27)] gives,

$$\frac{d}{dt}\left[\frac{g\ \ \dot{u}}{\sqrt{g\ \ \dot{u}\ \dot{u}}}\qquad \frac{1}{2\sqrt{g\ \ \dot{u}\ \dot{u}}}\frac{\partial g}{\partial u}\dot{u}\ \dot{u}\right] = 0$$

or

$$\frac{d}{dt}\ \frac{g\ \ \dot{u}}{ds/dt}\qquad \frac{1}{2ds/dt}\frac{\partial g}{\partial u}\dot{u}\ \dot{u} = 0$$

or

$$g\ \ u\ \ + \frac{\partial g}{\partial u}\dot{u}\ \dot{u}\qquad \frac{1}{2}\frac{\partial g}{\partial u}\dot{u}\ \dot{u} = \frac{1}{ds/dt}g\ \ \dot{u}\ \frac{d^2s}{dt^2}$$

or

$$g\ \ u\ \ + \frac{1}{2}\ \ \frac{\partial g}{\partial u} + \frac{\partial g}{\partial u}\qquad \frac{\partial g}{\partial u}\ \ \dot{u}\ \dot{u} = \frac{1}{ds/dt}g\ \ \dot{u}\ \frac{d^2s}{dt^2}$$

or

$$g\ \ u\ \ + [\ ;\ \ ]\dot{u}\ \dot{u} = g\ \ \dot{u}\ \frac{d^2s/dt^2}{ds/dt}: \tag{6.28}$$

These are the desired equations of geodesics on an arbitrary surface the Eq. (6.28) of geodesics for the curves of minimal length will play a role similar to that of straight lines in a plane. Also, this can be written as

$$g\ \ g\ \ u\ \ + g\ \ [\ ;\ \ ]\dot{u}\ \dot{u} = g\ \ g\ \ \ \dot{u}\ \frac{d^2s/dt^2}{ds/dt}$$

or

$$u\ \ + \qquad\quad \dot{u}\ \dot{u} = \frac{d^2s/dt^2}{ds/dt}\dot{u}\ : \tag{6.29}$$

If we choose the parameter t to be the arc length s of the curve, i.e. if we get s = t, then,

$$\frac{ds}{dt} = \quad = \sqrt{g \quad \underline{u}\ \underline{u}} = 1 \text{ and } \frac{d^2s}{dt^2} = \frac{d}{dt} = 0$$

then Eq. (6.28) reduces to

$$g \quad u \ + [\ \ ;\ \ ]\underline{u}\ \underline{u} = 0; \tag{6.30}$$

where '.' denotes the differentiation with respect to the arc parameter s. Multiplying Eq. (6.30) by g and sum, we obtain a simple form of the equations of geodesics in R^n as

$$g \quad g \quad u \ + g \quad [\ \ ;\ \ ]\underline{u}\ \underline{u} = 0;$$

or

$$u \ + \qquad \underline{u}\ \underline{u} = 0$$

or

$$\frac{d^2u}{ds^2} + \qquad \frac{du}{ds}\frac{du}{ds} = 0;\ \ ;\ ;\ \ = 1;2;\ \ ;n; \tag{6.31}$$

which is the differential equation of a geodesic in R^n. Since Eq. (6.31) is an ordinary second order differential equation it possesses a unique solution when the values $u^i(s)$ and the first derivatives $\frac{du^i}{ds}$ are prescribed arbitrarily at a given point $u^i(s_0)$: If we regard a given surface S as a Riemannian two-dimensional manifold R^2, covered by Gaussian co-ordinates u , then Eq. (6.31) takes the form

$$\frac{d^2u}{ds^2} + \qquad \frac{du}{ds}\frac{du}{ds} = 0;\ \ ;\ ;\ \ = 1;2: \tag{6.32}$$

Hence, at each point of S there exists a unique geodesic with an arbitrarily prescribed direction $\quad = \frac{du}{ds}$. Thus, if there exists a unique solution u (s), passing through two given points on S, then the curve u (s) is the curve of shortest length joining these points. The solution of the system (6.32) of second order differential equations will define the geodesics $u^i = u^i(s)$:

Deduction 6.3.1 (Necessary and sufficient condition for a geodesic): Let us introduce

$$T(u \ ;\underline{u} \ ) = \frac{1}{2}g \quad \underline{u}\ \underline{u} \ ;U \quad \frac{d}{dt}\ \frac{@T}{@\underline{u}} \qquad \frac{@T}{@u} = \frac{1}{2T}\frac{dT}{dt}\frac{@T}{@\underline{u}}; \tag{6.33}$$

where ; = 1; 2. The expression U so defined are important in relation to any curve, whether it is a geodesic or not. Now,

$$\underline{u}\ U \ = \frac{d}{dt}\ \ \underline{u}\ \frac{@T}{@\underline{u}} \qquad u \ \frac{@T}{@\underline{u}} \quad \underline{u}\ \frac{@T}{@u} = \frac{d}{dt}(2T) \qquad \frac{dT}{dt} = \frac{dT}{dt}; \tag{6.34}$$

where T is a function of u $\dot{u}$ homogeneous of degree 2 in $\dot{u}$. Since also the expression for U on the right how side of Eq. (6.33) satisfy the same identity, i.e.

$$\dot{u}\quad \frac{1}{2T}\frac{dT}{dt}\frac{\partial T}{\partial \dot{u}} = \frac{1}{2T}\frac{dT}{dt}\quad \dot{u}\frac{dT}{d\dot{u}} = \frac{dT}{dt};$$

it follows that the equations in Eq. (6.33) are not independent, they are therefore equivalent to only one equation for the unknowns u (t). Eliminating $\frac{dT}{dt}$ in Eq. (6.33) we get

$$u \frac{\partial T}{\partial \dot{u}} \quad u \frac{dT}{\partial \dot{u}} = 0 : ; \quad = 1; 2: \tag{6.35}$$

This, then, is necessary for a geodesic. To prove that it is also sufficient, suppose that it is satisfied by functions u (t), whose first derivatives do not vanish simultaneously at any point. The $\frac{\partial T}{\partial \dot{u}}$ cannot vanish together. Hence

$$U = \frac{\partial T}{\partial \dot{u}}; \text{ for some } ;$$

and then from Eq. 6.34 we get

$$\frac{dT}{dt} = \quad \dot{u}\frac{\partial T}{\partial \dot{u}} = 2T \quad) \quad = \frac{1}{2T}\frac{dT}{dt}:$$

The function u (t) therefore satisfy Eq. (6.33).

Deduction 6.3.2 Intrinsic curvature: The intrinsic curvature of a curve C in the surface S is the function

$$e = \sqrt{g_{ij}\, b^i b^j}\,; \quad b^i = \frac{}{s}\frac{dx^i}{ds} \;;$$

where the intrinsic curvature vector b^i is given by

$$b^i = \frac{}{s}\frac{dx^i}{ds} = \frac{d^2x^i}{ds^2} + \begin{matrix} i \\ p\ q \end{matrix}\;\frac{dx^p}{ds}\frac{dx^q}{ds}:$$

Intrinsic curvature can be shown to be the instantaneous rate of change of the angle between the tangent vector of C and another vector in the tangent space that is transported parallelly along the curve. A curve on a surface is a geodesic if and only if its intrinsic curvature e is identically zero.

EXAMPLE 6.3.1 Prove that the curves of the family $\frac{v^3}{u^2}$ = constant are geodesics on the surface with metric $v^2du^2 \quad 2uvdudv + 2u^2dv^2; u > 0; v > 0$.

Solution: Let $\frac{v^3}{u^2} = c(> 0)$ so that the parametric form of the equation of the curve is $u = ct^3; v = ct^2$. Construct.

$$T = \frac{1}{2}\left(v^2\dot{u}^2 - 2uv\dot{u}\dot{v} + 2u^2\dot{v}^2\right)$$

Therefore,

$$\begin{aligned}
&\frac{\partial T}{\partial u} = -v\dot{u}\dot{v} + 2u\dot{v}^2 = 2c^3t^5; \quad \frac{\partial T}{\partial v} = v\dot{u}^2 - u\dot{u}\dot{v} = 3c^3t^6\\
&\frac{\partial T}{\partial \dot{u}} = v^2\dot{u} - uv\dot{v} = c^3t^6; \quad \frac{\partial T}{\partial \dot{v}} = -uv\dot{u} + 2u^2\dot{v} = c^3t^7\\
&U^1 = \frac{d}{dt}(c^3t^6) - 2c^3t^5 = 4c^3t^5; \quad V^2 = \frac{d}{dt}(c^3t^7) - 3c^3t^6 = 4c^3t^6
\end{aligned}$$

Since $V\frac{\partial T}{\partial \dot{u}} - U\frac{\partial T}{\partial \dot{v}} = 0$, i.e. the curve is a geodesic for every value of c.

EXAMPLE 6.3.2 Show that the geodesics are straight lines when the co-ordinates are Cartesian.

Solution: Consider the Euclidean space E^3 of 3 dimensions. In this case, the metric is $ds^2 = g_{ij}\, dx^i dx^j$, where the tensor g_{ij} is denoted by a_{ij} and $g_{ij} = a_{ij} = \delta^i_j$: Thus, the components of the Christofell symbols are given by

$$g_{ij} = \delta^i_j \Rightarrow \left\{ {k \atop i\ j} \right\} = 0 = [\,ij; k\,];$$

relative to E^n. The differential equation of a geodesic in E^n, with Euclidean co-ordinates x^i is

$$\frac{d^2x^i}{ds^2} + \left\{ {i \atop j\ k} \right\} \frac{dx^j}{ds}\frac{dx^k}{ds} = 0$$

or

$$\frac{d^2x^i}{ds^2} = 0 \Rightarrow x^i = a^i s + b^i; \quad s = \text{ arc length},$$

where a^i and b^i are constants with $g_{ij}a^i a^j = 1$: Thus, from each point $x = b$ of space there emanates a geodesic ray in every direction (unit vector) a: This equation can be written in the form

$$\frac{x^1 - b^1}{a^1} = \frac{x^1 - b^1}{a^1} = \frac{x^1 - b^1}{a^1}(= s):$$

The equation $x^i = a^i s + b^i$ represents the equation of straight lines in E^3. Thus in affine co-ordinates, where all g_{ij} are constants and all Christoffel symbols vanish, the geodesics are straight lines.

EXAMPLE 6.3.3 Find the differential equations for the geodesic in spherical co-ordinates.

Solution: The relation between Cartesian and spherical co-ordinates is

$$x^1 = u^1 \sin u^2 \cos u^3;\ x^2 = u^1 \sin u^2 \sin u^3;\ x^3 = u^1 \cos u^2.$$

The expression for the metric in spherical co-ordinates is given by

$$ds^2 = \left(du^1\right)^2 + \left(u^1\right)^2 \left(du^2\right)^2 + \left(u^1\right)^2 \sin^2 u^2 \left(du^3\right)^2.$$

Comparing the given metric, with $ds^2 = a_{ij}\, du^i du^j$, we get

$$a_{11} = 1;\ a_{22} = \left(u^1\right)^2;\ a_{33} = \left(u^1\right)^2 \sin^2 u^2 \text{ and } a_{ij} = 0 \text{ for } i \neq j.$$

$$\Rightarrow\ a = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = (u^1)^4 \sin^2 u^2.$$

Therefore, the reciprocal tensors are given by

$$a^{11} = \frac{\text{cofactor of } a_{11} \text{ in } a}{a} = 1;\ a^{22} = \frac{\text{cofactor of } a_{22} \text{ in } a}{a} = \frac{1}{(u^1)^2}.$$

$$a^{33} = \frac{\text{cofactor of } a_{33} \text{ in } a}{a} = \frac{1}{(u^1)^2 \sin^2 u^2}.$$

$$a^{12} = 0 = a^{21};\ a^{13} = 0 = a^{31};\ a^{23} = 0 = a^{32}.$$

The non-vanishing Christofell symbols of the second kind are

$$\left\{\begin{matrix}1\\2\ 2\end{matrix}\right\} = u^1;\ \left\{\begin{matrix}2\\1\ 2\end{matrix}\right\} = \frac{1}{(u^1)^2} = \left\{\begin{matrix}2\\2\ 1\end{matrix}\right\};\ \left\{\begin{matrix}2\\3\ 3\end{matrix}\right\} = \sin u^2 \cos u^2;$$

$$\left\{\begin{matrix}3\\1\ 3\end{matrix}\right\} = \frac{1}{u^1} = \left\{\begin{matrix}3\\3\ 1\end{matrix}\right\};\ \left\{\begin{matrix}1\\3\ 3\end{matrix}\right\} = u^1 \sin^2 u^2;\ \left\{\begin{matrix}3\\2\ 3\end{matrix}\right\} = \cot u^2 = \left\{\begin{matrix}3\\3\ 2\end{matrix}\right\}.$$

The differential equation of a geodesic for the surface is given by

$$\frac{d^2u^i}{ds^2} + \left\{\begin{matrix}i\\j\ k\end{matrix}\right\} \frac{du^j}{ds}\frac{du^k}{ds} = 0;\ i, j, k = 1, 2, 3. \tag{6.36}$$

For $i = 1$, the differential equation becomes

$$\frac{d^2u^1}{ds^2} + \left\{\begin{matrix}1\\j\ k\end{matrix}\right\} \frac{du^j}{ds}\frac{du^k}{ds} = 0$$

or

$$\frac{d^2u^1}{ds^2} + \genfrac{}{}{0pt}{}{1}{1\ \ k}\frac{du^1}{ds}\frac{du^k}{ds} + \genfrac{}{}{0pt}{}{1}{2\ \ k}\frac{du^2}{ds}\frac{du^k}{ds} + \genfrac{}{}{0pt}{}{1}{3\ \ k}\frac{du^3}{ds}\frac{du^k}{ds} = 0$$

or

$$\frac{d^2u^1}{ds^2} + \genfrac{}{}{0pt}{}{1}{1\ \ 1}{\frac{du^1}{ds}}^{2} + 2\genfrac{}{}{0pt}{}{1}{1\ \ 2}\frac{du^1}{ds}\frac{du^2}{ds} + 2\genfrac{}{}{0pt}{}{1}{1\ \ 3}\frac{du^1}{ds}\frac{du^3}{ds}$$

$$+ \genfrac{}{}{0pt}{}{1}{2\ \ 2}{\frac{du^2}{ds}}^{2} + 2\genfrac{}{}{0pt}{}{1}{2\ \ 3}\frac{du^2}{ds}\frac{du^3}{ds} + \genfrac{}{}{0pt}{}{1}{3\ \ 3}{\frac{du^3}{ds}}^{2} = 0$$

or

$$\frac{d^2u^1}{ds^2} \quad u^1 {\frac{du^2}{ds}}^{2} \quad u^1 \sin^2 u^2 {\frac{du^3}{ds}}^{2} = 0$$

For i = 2, the differential equation becomes

$$\frac{d^2u^2}{ds^2} + \genfrac{}{}{0pt}{}{2}{j\ \ k}\frac{du^j}{ds}\frac{du^k}{ds} = 0$$

or

$$\frac{d^2u^2}{ds^2} + \genfrac{}{}{0pt}{}{2}{1\ \ k}\frac{du^1}{ds}\frac{du^k}{ds} + \genfrac{}{}{0pt}{}{2}{2\ \ k}\frac{du^2}{ds}\frac{du^k}{ds} + \genfrac{}{}{0pt}{}{2}{3\ \ k}\frac{du^3}{ds}\frac{du^k}{ds} = 0$$

or

$$\frac{d^2u^2}{ds^2} + \genfrac{}{}{0pt}{}{2}{1\ \ 1}{\frac{du^1}{ds}}^{2} + 2\genfrac{}{}{0pt}{}{2}{1\ \ 2}\frac{du^1}{ds}\frac{du^2}{ds} + 2\genfrac{}{}{0pt}{}{2}{1\ \ 3}\frac{du^1}{ds}\frac{du^3}{ds}$$

$$+ \genfrac{}{}{0pt}{}{2}{2\ \ 2}{\frac{du^2}{ds}}^{2} + 2\genfrac{}{}{0pt}{}{2}{2\ \ 3}\frac{du^2}{ds}\frac{du^3}{ds} + \genfrac{}{}{0pt}{}{2}{3\ \ 3}{\frac{du^3}{ds}}^{2} = 0$$

or

$$\frac{d^2u^2}{ds^2} + \frac{2}{(u^1)^2}\frac{du^1}{ds}\frac{du^2}{ds} \quad \sin u^2 \cos u^2 {\frac{du^3}{ds}}^{2} = 0:$$

For i = 3, the differential equation becomes

$$\frac{d^2u^3}{ds^2} + \genfrac{}{}{0pt}{}{3}{j\ \ k}\frac{du^j}{ds}\frac{du^k}{ds} = 0$$

or

$$\frac{d^2u^3}{ds^2} + \genfrac{}{}{0pt}{}{3}{1\ \ k}\frac{du^1}{ds}\frac{du^k}{ds} + \genfrac{}{}{0pt}{}{3}{2\ \ k}\frac{du^2}{ds}\frac{du^k}{ds} + \genfrac{}{}{0pt}{}{3}{3\ \ k}\frac{du^3}{ds}\frac{du^k}{ds} = 0$$

or

$$\frac{d^2u^3}{ds^2}+\begin{Bmatrix}3\\1\;1\end{Bmatrix}\left(\frac{du^1}{ds}\right)^2+2\begin{Bmatrix}3\\1\;2\end{Bmatrix}\frac{du^1}{ds}\frac{du^2}{ds}+2\begin{Bmatrix}3\\1\;3\end{Bmatrix}\frac{du^1}{ds}\frac{du^3}{ds}$$
$$+\begin{Bmatrix}3\\2\;2\end{Bmatrix}\left(\frac{du^2}{ds}\right)^2+2\begin{Bmatrix}3\\2\;3\end{Bmatrix}\frac{du^2}{ds}\frac{du^3}{ds}+\begin{Bmatrix}3\\3\;3\end{Bmatrix}\left(\frac{du^3}{ds}\right)^2=0$$

or

$$\frac{d^2u^3}{ds^2}+\frac{2}{u^1}\frac{du^1}{ds}\frac{du^3}{ds}+\cot u^2\frac{du^2}{ds}\frac{du^3}{ds}=0.$$

EXAMPLE 6.3.4 Find the differential equations for the geodesic in cylindrical co-ordinates.

Solution: The relation between Cartesian and cylindrical co-ordinates is

$$x^1=u^1\cos u^2;\; x^2=u^1\sin u^2;\; x^3=u^3.$$

The expression for the metric in cylindrical co-ordinates is given by

$$ds^2=\left(du^1\right)^2+\left(u^1\right)^2\left(du^2\right)^2+\left(du^3\right)^2;$$

from which, the coefficients $a_{\alpha\beta}$ of the first fundamental form are given by

$$a_{11}=1;\; a_{22}=(u^1)^2;\; a_{33}=1 \text{ and } a_{ij}=0;\text{ for } i\neq j;$$

so that $a=(u^1)^2$. The reciprocal tensors $a^{\alpha\beta}$ are given by

$$a^{11}=1;\; a^{22}=\frac{1}{(u^1)^2};\; a^{33}=1$$
$$a^{12}=0=a^{21};\; a^{13}=0=a^{31};\; a^{23}=0=a^{32}.$$

The non-vanishing Christofell symbols of the second kind are

$$\begin{Bmatrix}1\\2\;2\end{Bmatrix}=-u^1;\quad \begin{Bmatrix}2\\1\;2\end{Bmatrix}=\frac{1}{u^1}=\begin{Bmatrix}2\\2\;1\end{Bmatrix}.$$

The differential equation of a geodesic for the surface is given by Eq. (6.36). For $i=1$, the differential equation becomes

$$\frac{d^2u^1}{ds^2}+\begin{Bmatrix}1\\j\;k\end{Bmatrix}\frac{du^j}{ds}\frac{du^k}{ds}=0$$

or

$$\frac{d^2u^1}{ds^2}+\begin{Bmatrix}1\\1\;k\end{Bmatrix}\frac{du^1}{ds}\frac{du^k}{ds}+\begin{Bmatrix}1\\2\;k\end{Bmatrix}\frac{du^2}{ds}\frac{du^k}{ds}+\begin{Bmatrix}1\\3\;k\end{Bmatrix}\frac{du^3}{ds}\frac{du^k}{ds}=0$$

or

$$\frac{d^2u^1}{ds^2}+\genfrac{}{}{0pt}{}{1}{1\ 1}\ \frac{du^1}{ds}^{\ 2}+2\ \genfrac{}{}{0pt}{}{1}{1\ 2}\ \frac{du^1}{ds}\frac{du^2}{ds}+2\ \genfrac{}{}{0pt}{}{1}{1\ 3}\ \frac{du^1}{ds}\frac{du^3}{ds}$$

$$+\ \genfrac{}{}{0pt}{}{1}{2\ 2}\ \frac{du^2}{ds}^{\ 2}+2\ \genfrac{}{}{0pt}{}{1}{2\ 3}\ \frac{du^2}{ds}\frac{du^3}{ds}+\ \genfrac{}{}{0pt}{}{1}{3\ 3}\ \frac{du^3}{ds}^{\ 2}=0$$

or

$$\frac{d^2u^1}{ds^2}\quad u^1\ \frac{du^2}{ds}^{\ 2}=0$$

For i = 2, the differential equation becomes

$$\frac{d^2u^2}{ds^2}+\ \genfrac{}{}{0pt}{}{2}{j\ \ k}\ \frac{du^j}{ds}\frac{du^k}{ds}=0$$

or

$$\frac{d^2u^2}{ds^2}+\ \genfrac{}{}{0pt}{}{2}{1\ \ k}\ \frac{du^1}{ds}\frac{du^k}{ds}+\ \genfrac{}{}{0pt}{}{2}{2\ \ k}\ \frac{du^2}{ds}\frac{du^k}{ds}+\ \genfrac{}{}{0pt}{}{2}{3\ \ k}\ \frac{du^3}{ds}\frac{du^k}{ds}=0$$

or

$$\frac{d^2u^2}{ds^2}+\genfrac{}{}{0pt}{}{2}{1\ 1}\ \frac{du^1}{ds}^{\ 2}+2\ \genfrac{}{}{0pt}{}{2}{1\ 2}\ \frac{du^1}{ds}\frac{du^2}{ds}+2\ \genfrac{}{}{0pt}{}{2}{1\ 3}\ \frac{du^1}{ds}\frac{du^3}{ds}$$

$$+\ \genfrac{}{}{0pt}{}{2}{2\ 2}\ \frac{du^2}{ds}^{\ 2}+2\ \genfrac{}{}{0pt}{}{2}{2\ 3}\ \frac{du^2}{ds}\frac{du^3}{ds}+\ \genfrac{}{}{0pt}{}{2}{3\ 3}\ \frac{du^3}{ds}^{\ 2}=0$$

or

$$\frac{d^2u^2}{ds^2}+\frac{2}{u^1}\frac{du^1}{ds}\frac{du^2}{ds}=0:$$

EXAMPLE 6.3.5 Find the differential equations of the geodesic for the metric

$$ds^2=(\ dx^1)^2+\ f(x^2)^2\quad (x^1)^2 g(dx^2)^2:$$

Solution: The square of the elementary arc length is given by

$$ds^2=(\ dx^1)^2+\ f(x^2)^2\quad (x^1)^2 g(dx^2)^2:$$

The coefficients a of the first fundamental form are given by

$$a_{11}=1;\ \ a_{22}=(\ x^2)^2\quad (x^1)^2 \text{ and } a_{12}=0=\ a_{21}:$$

so that $a = (x^2)^2 - (x^1)^2$. The reciprocal tensors a^{ij} are given by

$$a^{11} = 1;\ \ a^{22} = \frac{1}{(x^2)^2 - (x^1)^2};\ \ a^{12} = 0 = a^{21}.$$

The non-vanishing Christofell symbols of the second kind are

$$\left\{ {1 \atop 2\ 2} \right\} = x^1;\quad \left\{ {2 \atop 1\ 2} \right\} = -\frac{x^1}{(x^2)^2 - (x^1)^2};\quad \left\{ {2 \atop 2\ 2} \right\} = \frac{x^2}{(x^2)^2 - (x^1)^2}.$$

The differential equation of a geodesic for the surface is given by

$$\frac{d^2x^i}{ds^2} + \left\{ {i \atop j\ k} \right\} \frac{dx^j}{ds}\frac{dx^k}{ds} = 0;\ \ i, j, k = 1, 2. \tag{6.37}$$

For $i = 1$, the differential equation becomes

$$\frac{d^2x^1}{ds^2} + \left\{ {1 \atop j\ k} \right\} \frac{dx^j}{ds}\frac{dx^k}{ds} = 0$$

or

$$\frac{d^2x^1}{ds^2} + \left\{ {1 \atop 1\ k} \right\} \frac{dx^1}{ds}\frac{dx^k}{ds} + \left\{ {1 \atop 2\ k} \right\} \frac{dx^2}{ds}\frac{dx^k}{ds} = 0$$

or

$$\frac{d^2x^1}{ds^2} + \left\{ {1 \atop 1\ 1} \right\} \left(\frac{dx^1}{ds}\right)^2 + 2\left\{ {1 \atop 1\ 2} \right\} \frac{dx^1}{ds}\frac{dx^2}{ds} + \left\{ {1 \atop 2\ 2} \right\} \left(\frac{dx^2}{ds}\right)^2 = 0$$

or

$$\frac{d^2x^1}{ds^2} + x^1 \left(\frac{dx^2}{ds}\right)^2 = 0.$$

For $i = 2$, the differential equation becomes

$$\frac{d^2x^2}{ds^2} + \left\{ {2 \atop j\ k} \right\} \frac{dx^j}{ds}\frac{dx^k}{ds} = 0$$

or

$$\frac{d^2x^2}{ds^2} + \left\{ {2 \atop 1\ k} \right\} \frac{dx^1}{ds}\frac{dx^k}{ds} + \left\{ {2 \atop 2\ k} \right\} \frac{dx^2}{ds}\frac{dx^k}{ds} = 0$$

or

$$\frac{d^2x^2}{ds^2} + \left\{ {2 \atop 1\ 1} \right\} \left(\frac{dx^1}{ds}\right)^2 + 2\left\{ {2 \atop 1\ 2} \right\} \frac{dx^1}{ds}\frac{dx^2}{ds} + \left\{ {2 \atop 2\ 2} \right\} \left(\frac{dx^2}{ds}\right)^2 = 0$$

or

$$\frac{d^2x^2}{ds^2} - \frac{2x^1}{(x^2)^2 - (x^1)^2}\frac{dx^1}{ds}\frac{dx^2}{ds} + \frac{x^2}{(x^2)^2 - (x^1)^2}\left(\frac{dx^2}{ds}\right)^2 = 0;$$

which are the required equations of the geodesic.

EXAMPLE 6.3.6 Find the geodesics on the surface

$$x^1 = u^1 \cos u^2;\ x^2 = u^1 \sin u^2;\ x^3 = 0;$$

embedded in E^3, where x^i are orthogonal Cartesian co-ordinates.

Solution: The square of the elementary arc length is given by

$$\begin{aligned}(ds)^2 &= (dx^1)^2 + (dx^2)^2 + (dx^3)^2\\ &= \ du^1 \cos u^2 + u^1(\ \ \sin u^2)du^2\ ^2 + \ du^1 \sin u^2 + u^1 \cos u^2 du^2\ ^2 + 0\\ &= (du^1)^2 + (u^1)^2(du^2)^2:\end{aligned}$$

The coefficients a of the first fundamental form are given by

$$a_{11} = 1;\ \ a_{22} = (u^1)^2 \text{ and } a_{12} = 0 = a_{21};$$

so that $a = (u^1)^2$: The reciprocal tensors a are given by

$$a^{11} = 1;\ \ a^{22} = \frac{1}{(u^1)^2};\ \ a^{12} = 0 = a^{21}:$$

The non-vanishing Christofell symbols of the second kind are

$$\genfrac{}{}{0pt}{}{1}{2\ 2} = \ \ u^1;\quad \genfrac{}{}{0pt}{}{2}{1\ 2} = \frac{1}{u^1} = \genfrac{}{}{0pt}{}{2}{2\ 1}\ :$$

The differential equation of a geodesic for the surface is given by Eq. (6.37). For $i = 1$, the differential equation becomes

$$\frac{d^2u^1}{ds^2} + \genfrac{}{}{0pt}{}{1}{j\ k}\frac{du^j}{ds}\frac{du^k}{ds} = 0$$

or

$$\frac{d^2u^1}{ds^2} + \genfrac{}{}{0pt}{}{1}{1\ k}\frac{du^1}{ds}\frac{du^k}{ds} + \genfrac{}{}{0pt}{}{1}{2\ k}\frac{du^2}{ds}\frac{du^k}{ds} = 0$$

or

$$\frac{d^2u^1}{ds^2} + \genfrac{}{}{0pt}{}{1}{1\ 1}\ \frac{du^1}{ds}\ ^2 + 2\genfrac{}{}{0pt}{}{1}{1\ 2}\frac{du^1}{ds}\frac{du^2}{ds} + \genfrac{}{}{0pt}{}{1}{2\ 2}\ \frac{du^2}{ds}\ ^2 = 0$$

or

$$\frac{d^2u^1}{ds^2}\quad u^1\ \frac{du^2}{ds}\ ^2 = 0: \tag{i}$$

For $i = 2$, the differential equation becomes

$$\frac{d^2u^2}{ds^2} + \genfrac{}{}{0pt}{}{2}{j\ k}\frac{du^j}{ds}\frac{du^k}{ds} = 0$$

or

$$\frac{d^2u^2}{ds^2} + \genfrac{}{}{0pt}{}{2}{1\ k}\frac{du^1}{ds}\frac{du^k}{ds} + \genfrac{}{}{0pt}{}{2}{2\ k}\frac{du^2}{ds}\frac{du^k}{ds} = 0$$

or

$$\frac{d^2u^2}{ds^2} + \genfrac{}{}{0pt}{}{2}{1\ 1}\frac{du^1}{ds}\ ^{2} + 2\genfrac{}{}{0pt}{}{2}{1\ 2}\frac{du^1}{ds}\frac{du^2}{ds} + \genfrac{}{}{0pt}{}{2}{2\ 2}\frac{du^2}{ds}\ ^{2} = 0$$

or

$$\frac{d^2u^2}{ds^2} + \frac{2}{u^1}\frac{du^1}{ds}\frac{du^2}{ds} = 0;\ \text{i.e.}\ \frac{d^2u^2}{ds^2} = \frac{du^2}{ds} + \frac{2}{u^1}\frac{du^1}{ds} = 0$$

or

$$\log\ \frac{du^2}{ds}\ + 2\log u^1 = \log c;\ c = \text{constant}$$

or

$$\log\ (u^1)^2\frac{du^2}{ds}\ = \log c\,)\ \ \frac{du^2}{ds} = \frac{c}{(u^1)^2}: \tag{ii}$$

Therefore, from Eq. (i) we get

$$\frac{d^2u^1}{ds^2}\quad u^1\frac{c^2}{(u^1)^4} = 0;\ \text{i.e.}\ 2\frac{du^1}{ds}\frac{d^2u^1}{ds^2} = 2\frac{c^2}{(u^1)^3}\frac{du^1}{ds}$$

or

$$\frac{du^1}{ds}\ ^{2} = \quad \frac{c^2}{(u^1)^2} + c_1^2;\ c_1 = \text{constant}$$

$$= c_1^2\ \ 1\quad \frac{c^2}{c_1^2(u^1)^2}\ = c_1^2\frac{(u^1)^2\quad d^2}{(u^1)^2};\ d = \frac{c}{c_1}$$

or

$$\frac{u^1\ \ du^1}{(u^1)^2\quad d^2} = c_1 ds\,)\ \ (u^1)^2 = (c_1 s + e)^2 + d^2;$$

where e is an integration constant. Therefore, from Eq. (ii) we get

$$\frac{du^2}{ds} = \frac{c}{(u^1)^2} = \frac{c}{d^2}\frac{1}{1 + \ \frac{c_1 s + e}{d}\ ^{2}}$$

or

$$u^2 = \frac{c}{d^2}\frac{1}{c_1}\tan\ ^{1}\ \frac{c_1 s + e}{d}$$

or

$$u^2 = \frac{1}{p}\tan\ ^{1}\ \frac{c_1 s + e}{d}\ \ ;\ p = \frac{d^2 c_1}{c}$$

or

$$d\tan(pu^2) = c_1 s + e;$$

which is the required equation of the geodesic.

EXAMPLE 6.3.7 On a right helicoid of pitch $2\pi a$, a geodesic makes as angle α with a generator at a point distant c from the axis $(0 < \alpha < \frac{\pi}{2}; c > 0)$. Find the condition that the geodesics meets the axis.

Solution: The first fundamental form of the right helicoid is given by

$$ds^2 = du^2 + (u^2 + a^2)dv^2.$$

The non-vanishing Christofell symbols of the second kind are

$$\left\{ {1 \atop 2\;2} \right\} = -u; \quad \left\{ {2 \atop 2\;1} \right\} = \frac{u}{u^2 + a^2} = \left\{ {2 \atop 1\;2} \right\}.$$

The differential equation of geodesics are

$$\frac{d^2u}{ds^2} - u\left(\frac{dv}{ds}\right)^2 = 0 \quad \text{and} \quad \frac{d^2v}{ds^2} + \frac{du}{u^2 + a^2}\frac{du}{ds}\frac{dv}{ds} = 0.$$

Thus the first integral of the geodesic equations is

$$\frac{dv}{du} = \pm\frac{k}{\{(u^2 + a^2)(u^2 + a^2 - k^2)\}^{1/2}},$$

where k is an arbitrary positive constant. Further integration in general requires elliptic functions.

The given point is (c, 0) for a suitable choice of axes, and α is the angle between the direction (1, 0) and $\left(\frac{du}{ds}, \frac{dv}{ds}\right)$ at this point, then

$$\tan\alpha = (a_{11}a_{22} - a_{12}^2)^{1/2}\,\frac{dv/ds}{du/ds} = k(c^2 + a^2 - k^2)^{-1/2}$$

$$\Rightarrow \quad k = (c^2 + a^2)^{1/2}\sin\alpha.$$

There are two geodesics satisfying the given initial conditions, but it will be sufficient to consider the one for which $\frac{dv}{du} < 0$ initially. Now we consider the following three cases:

(i) $k^2 > a^2$, i.e. $c\tan\alpha > a$. Since which $\frac{dv}{du} < 0$ initially, u decreases as v increases until

$$u = (k^2 - a^2)^{1/2} = (c^2\sin^2\alpha - a^2\cos^2\alpha)^{1/2}.$$

As v continues to increase, the sign of which $\frac{dv}{du}$ changes and u increases indefinitely. The least distance from the axis is therefore $(c^2\sin^2\alpha - a^2\cos^2\alpha)^{1/2}$.

(ii) $k^2 < a^2$, i.e. $c \tan\ < a$. In this case which $\frac{dv}{du} < 0$ for all v, and u decreases indefinitely as v increases. There is a point on the curve at which u = 0, i.e. the curve meets the axis.

(iii) $k^2 = a^2$, i.e. $c \tan\ = a$. In this special case

$$\frac{dv}{du} = \frac{a}{u\sqrt{u^2 + a^2}} \Rightarrow v = \ + \sinh^{1} \frac{a}{u}$$

where $= \sinh^{1}(a/c)$, as v = 0, when u = c. The geodesic is therefore given by

$$u \sinh(v + \ ) = a; \quad = \sinh^{1}(a/c).$$

As v increases, the curve approaches the axis without reaching it. In the opposite sense, $u \to \infty$ as $v \to$, showing that the generator $v =$ is an asymptote.

6.4 Geodesic Co-ordinates

If a Riemannian space is Euclidean, a co-ordinate system exists in which the components of the metric tensors g_{ij} are constant throughout the space and hence, $\frac{\partial g_{ij}}{\partial x^k} = 0$; for all i, j, k. Consequently, the vanishing of the Christoffel symbols, as Eq. (3.1). This is not true for any Riemannian space. But if in a Riemannian space, there exists a co-ordinate system, in fact infinitely many, with respect to which the Christoffel symbols vanish at a given point, then, that system is called a geodesic co-ordinate system and the point is called the pole of the given system.

Let us consider a surface S whose curvilinear co-ordinates are u^1, u^2 and also consider a point $P(u_0^1, u_0^2)$ on S. If $v\ ;\ = 1, 2$ are the co-ordinates of some net on S, then we consider a transformation

$$u\ = u\ \left(v^1, v^2\right); \quad = 1, 2. \tag{6.38}$$

The second derivative formula yields the relation,

$$\frac{\partial^2 u}{\partial v \partial v} + \underset{u}{\ } \frac{\partial u}{\partial v}\frac{\partial u}{\partial v} = \underset{v}{\ } \frac{\partial u}{\partial v}, \tag{6.39}$$

where $\underset{u}{\ }$ and $\underset{v}{\ }$ are the Christoffel symbols in u co-ordinate system and v co-ordinate system, respectively. However, if there exists a transformation of co-ordinates Eq. (6.38) for which $\underset{v}{\ }$ vanish at P, then for that particular point

$$\frac{\partial^2 u}{\partial v \partial v} + \underset{u}{\ } \frac{\partial u}{\partial v}\frac{\partial u}{\partial v} = 0. \tag{6.40}$$

Next, we exhibit next a solution of the Eq. (6.40) yielding transformation Eq. (6.38) to a co-ordinate system v in which the Christoffel symbols vanish at $P(u_0^1; u_0^2)$ on S. Let us take a second degree polynomial

$$u = u_P + v \qquad {}_P\, v\ v\ ; \tag{6.41}$$

where u_P is the value of u at P. On differentiation, from Eq. (6.41), we get

$$\frac{@u}{@v} = \qquad {}_P\, v \text{ and } \frac{@^2 u}{@v@v} = \qquad {}_P$$

$$\left(\frac{@u}{@v}\right)_P = \quad \text{and} \quad \left(\frac{@^2u}{@v@v}\right)_P = \qquad {}_P : \tag{6.42}$$

Since the Jacobian determinant $\left(\frac{@u}{@v}\right)_P = j \quad j = 1 \neq 0$; so the given transformation is permissible in the neighbourhood of $P(u_0^1; u_0^2)$ on S. Using the formula,

$$u_{;} = \frac{@^2u}{@v@v} \qquad k \quad \frac{@u}{@v^k};$$

we see that at $P(u_0^1; u_0^2)$ on S,

$$\left(u_{;}\right)_P = \left(\frac{@^2u}{@v@v}\right)_P \qquad \left(k\right)_P \left(\frac{@u}{@v^k}\right)_P$$

$$= \qquad {}_P \quad \left(k\right)_P \ {}_k = 0 : \tag{6.43}$$

Therefore, the values in Eq. (6.42) satisfies Eq. (6.40) at P. From Eq. (6.41), we see that at P, the new co-ordinates, given by $v = 0$ are the geodesic co-ordinates.

EXAMPLE 6.4.1 Prove that the co-ordinate system u defined by

$$u = a_m^j (v^m \quad u_0^m) + \frac{1}{2} a_h^j \left\{ \begin{matrix} h \\ l\ m \end{matrix} \right\}_P (v^l \quad u_0^l)(v^m \quad u_0^m)$$

is a geodesic co-ordinate, where the coefficients a_m^j being constants and are such that their determinant do not vanish.

Solution: The necessary and sufficient condition that a system of co-ordinates be geodesic with pole at P are that their second covariant derivatives with respect to the metric of the space all vanish at P. From the given relation,

$$u = a_m^j (v^m \quad u_0^m) + \frac{1}{2} a_h^j \left\{ \begin{matrix} h \\ l\ m \end{matrix} \right\}_P (v^l \quad u_0^l)(v^m \quad u_0^m);$$

we get after differentiation,

$$\frac{\partial u^j}{\partial v^l} = a^j_m \delta^m_l + \frac{1}{2} a^j_h \left\{ \begin{matrix} h \\ l\ m \end{matrix} \right\}_P (v^m - u^m_0) + \frac{1}{2} a^j_h \left\{ \begin{matrix} h \\ l\ m \end{matrix} \right\}_P (v^l - u^l_0)\, \delta^m_l$$

$$= a^j_l + \frac{1}{2} a^j_h \left\{ \begin{matrix} h \\ l\ m \end{matrix} \right\}_P (v^m - u^m_0).$$

$$\Rightarrow \quad \left(\frac{\partial u^j}{\partial v^l} \right)_P = a^j_l .$$

Since the Jacobian determinant $\left| \frac{\partial u^j}{\partial v^l} \right|_P = |a^j_l| \neq 0$; so the given transformation is permissible in the neighbourhood of $P(u^1_0, u^2_0)$ on S. Again differentiating with respect to v^m, we get

$$\left(\frac{\partial^2 u^j}{\partial v^m \partial v^l} \right)_P = a^j_h \left\{ \begin{matrix} h \\ l\ m \end{matrix} \right\}_P .$$

Using the formula,

$$u^j_{;lm} = \frac{\partial^2 u^j}{\partial v^l \partial v^m} - \left\{ \begin{matrix} h \\ l\ m \end{matrix} \right\} \frac{\partial u^j}{\partial v^h},$$

we see that at $P(u^1_0, u^2_0)$ on S,

$$\left(u^j_{;lm} \right)_P = \left(\frac{\partial^2 u^j}{\partial v^l \partial v^m} \right)_P - \left\{ \begin{matrix} h \\ l\ m \end{matrix} \right\}_P \left(\frac{\partial u^j}{\partial v^h} \right)_P$$

$$= a^j_h \left\{ \begin{matrix} h \\ l\ m \end{matrix} \right\}_P - a^j_h \left\{ \begin{matrix} h \\ l\ m \end{matrix} \right\}_P = 0. \qquad (6.44)$$

This shows that the new co-ordinate system u defined in the given relation is a geodesic co-ordinate system with the pole at P.

6.5 Parallel Vector Fields on a Surface

The concept of parallel vector fields along a curve C imbedded in E^3 was generalised by Levi-Civita to curves imbedded in an n-dimensional Riemannian manifolds. As an illustration of the usefulness of the concept, consider a surface S immersed in E^3 and a curve C on S; whose equations are taken in the form

$$C : u^\alpha = u^\alpha(t);\ t_1 \leq t \leq t_2;\quad \alpha = 1, 2 \qquad (6.45)$$

where $u^1; u^2$ are the curvilinear co-ordinates covered by the surface S: Let the metric properties of the surface S are governed by the co-ordinates of the first fundamental form a . Let A be a surface vector field defined along C; then the surface intrinsic derivative of A along C is

$$\frac{A}{t} = \frac{dA}{dt} + \qquad A \frac{du}{dt}; \tag{6.46}$$

which is identical with the form of Eq. (3.39) defining the parallel vector field along a space curve. Thus, if $\frac{A}{t} = 0$, the differential Eq. (6.46) becomes

$$\frac{dA}{dt} + \qquad A \frac{du}{dt} = 0; \tag{6.47}$$

which determines a unique vector field when the components of the vector are specified at an arbitrary point of C as the definition of the parallel vector field along a curve C on the surface S: If the parameter t is chosen as the arc length s and if A is taken as the unit tangent vector to C, i.e. if we take

$$A = \frac{du}{ds} = \quad ; \text{ with } a \qquad = 1;$$

then Eq. (6.47) reduces to

$$\frac{d}{dt} \quad \frac{du}{ds} \quad + \qquad \frac{du}{ds}\frac{du}{ds} = 0$$

or

$$\frac{d^2u}{ds^2} + \qquad \frac{du}{ds}\frac{du}{ds} = 0; \tag{6.48}$$

which is the equation of a geodesic on the surface S: From uniqueness of the solution of Eq. (6.48), it follows that the property of tangency of a parallel vector field to a surface curve is both a necessary and sufficient condition for a geodesic. Also, Eq. (6.47) can be written in the form

$$\frac{d}{dt} \quad \frac{du}{ds} \quad + \qquad \frac{du}{ds}\frac{du}{ds} = 0$$

or

$$\frac{@}{@u} \quad \frac{du}{ds} \quad \frac{du}{ds} + \qquad \frac{du}{ds}\frac{du}{ds} = 0$$

or

$$\frac{@}{@u} \quad \frac{du}{ds} \quad + \qquad \frac{du}{ds} \quad \frac{du}{ds} = 0$$

or

$$\frac{du}{ds}{}_{;}\,\frac{du}{ds} = 0.$$

Therefore, the vector obtained by the parallel propagation of the tangent vector to a geodesic always remains tangent to the geodesic.

6.5.1 Geodesic Parallels

The ‘field of geodesics’ was introduced by Weierstrass. A one-parameter family of geodesics on a surface S is said to be a field of geodesics in a portion S^0 of S if through every point of S^0 there passes (just once) exactly one of those geodesics. For example, a family of parallel straight lines is a field of geodesics in a plane. The generating straight lines of a cylinder constitute a field of geodesics.

Suppose a family of geodesics is given, and that a parameter system is chosen so that the geodesics of the family are the curves v = constant and their orthogonal trajectories are the curves u = constant. Since the co-ordinates are orthogonal, we have $a_{12} = 0$. Since the curves v = constant are geodesics we have

$$(\kappa_g)_{v=\text{constant}} = 0 \Rightarrow \frac{1}{\sqrt{a_{22}}}\frac{\partial}{\partial v}\left(\log\sqrt{a_{11}}\right) = 0 \Rightarrow \frac{\partial a_{11}}{\partial v} = 0,$$

that is a_{11} depends on u only. The metric is of the form

$$ds^2 = a_{11}(u)\,du^2 + a_{22}(u,v)\,dv^2.$$

Consider the distance between any two orthogonal trajectories, say $u = a$ and $u = b$, measured only the geodesic $v = c$. Along $v = c$, $dv = 0$, so

$$ds = \sqrt{a_{11}(u)}\,du.$$

Therefore, the length of a geodesic intercepted between the trajectories $u = a$ and $u = b$ is

$$\int_a^b \sqrt{a_{11}(u)}\,du = c \text{ dependent}.$$

The distance is thus the same along whichever geodesic v = constant it is measured and is called geodesic distance between two curves. Because of this property, the orthogonal trajectories u = constant are called geodesic parallels.

Hence, if we take $\int\sqrt{a_{11}}\,du$ as a new parameter u, the first fundamental form is given by the expression

$$ds^2 = (du)^2 + a_{22}(u,v)(dv)^2 = (du)^2 + a(dv)^2,$$

which is characteristic for geodesic co-ordinates. In this metric the parameter u can be specialized by taking it to be the distance from some fixed parallel to the parallel determined by u, the distance being meased along any geodesic $v = c$, i.e. $a_{11} = 1$. Since a_{11} is now equal to unity, the length of an element of arc of a geodesic is du.

Consider the two points $P; Q$ in which a geodesic is cut by the parallels $u = a; u = b$. The length of the arc of the geodesic joining the two points is $(b \quad a)$. For any other curve joining them the length of arc is

$$\int_P^Q ds = \int_P^Q \sqrt{(du)^2 + a(dv)^2} > \int_a^b du;$$

since a is positive. Thus the distance is least in the case of geodesic. Hence, for any given family of geodesics, a parameter system can be chosen so that the metric takes the form $(du)^2 + a(dv)^2$. The given geodesics are the parametric curves $v =$ constant and their orthogonal trajectories are $u =$ constant, u being the measured along a geodesic from some fixed parallel.

EXAMPLE 6.5.1 Prove that the geodesic is an auto parallel curve.

Solution: An autoparallel curve is a curve whose tangent vector field constituted by the tangents at each point of the curve is a parallel vector field. The differential equation of the geodesic C is given by

$$\frac{d^2x}{ds^2} + \frac{dx}{ds}\frac{dx}{ds} \qquad = 0$$

or

$$\frac{@}{@s} + \quad \frac{dx}{ds} \qquad = 0$$

or

$$\frac{@}{@x} + \qquad \frac{dx}{ds} = 0 \;) \quad {}_{;} \quad = 0;$$

which shows that the curve C possesses the property that the tangents at all its points are parallel. Also, ${}_{;} \quad = 0$ shows that the unit tangent $\quad = \frac{dx}{ds}$ suffers a parallel displacement along a geodesic, i.e. a geodesic is an autoparallel curve.

6.6 Gaussian Curvature

The meaning of the curvature tensors will be explained here in connexion with parallel displacement and absolute differentiation.

Let us consider a surface S; embedded in E^3 whose fundamental tensors are a . On the surface S; where the metric is Eq. (6.10), with $u^1; u^2$ as the Gaussian coordinates for the surface S; the Riemann-Christoffel curvature tensor or simply the curvature tensor is given by

$$R \quad = \frac{\partial}{\partial u} \qquad\qquad \frac{\partial}{\partial u} \qquad\qquad +$$

$$= \quad \frac{\partial}{\partial u} \qquad \frac{\partial}{\partial u} \qquad + \tag{6.49}$$

and the associated tensor is given by

$$R \quad = a \quad R \quad ; \tag{6.50}$$

where R are the components of the Riemann curvature tensor for the surface S; given by

$$R \quad = \frac{\partial}{\partial u}[\ ;\ ] \quad \frac{\partial}{\partial u}[\ ;\ ] + \qquad [\ ;\ ] \qquad [\ ;\ ]: \tag{6.51}$$

Note that, this associated tensor R defined in Eq. (6.50) in skew symmetric in the first two and last two indices. Therefore,

$$R \quad = R \quad = 0;\ \text{i.e. } R_{1212} = \quad R_{2112} = \quad R_{1221} = R_{2121}:$$

Hence, every non-vanishing components of the Riemann curvature tensor for a surface is R_{1212} or to its negative. Let a surface S be immersed in E_3 with metric tensor a . Define a quantity by

$$= \frac{R_{1212}}{a} = \frac{1}{a} a_{1} \ R_{212}; \quad \text{where } a = |a \ | \neq 0 \tag{6.52}$$

$$= \frac{1}{a} a_{1} \quad \frac{\partial}{\partial u^{1}}\ _{2\,2} \quad \frac{\partial}{\partial u^{2}}\ _{2\,1} + \ _{2\,2} \quad ^{1} \quad _{2\,1} \quad ^{2} \quad :$$

Such a quantity is called the total curvature or the Gaussian curvature of the surface S. The Eq. (6.49) is a representation of the Gaussian curvature k in terms of the Christoffel symbols. Since only the metric tensor a and its derivatives are involved in expression (6.52) for , the properties described by are intrinsic properties of the surface S. Let us introduce an e-system e ; e defined by Eq. (1.5), and the permutation tensor by Eq. (6.21). With the help of ε tensor, Eq. (6.52) can also be written in the form

$$R_{1212} = a \quad = \sqrt{a}\ e_{12} \sqrt{a}\ e_{12} \quad = \varepsilon_{12}\varepsilon_{12}:$$

In general

$$R \quad = \text{"} \quad \text{"} \quad : \tag{6.53}$$

Also, we have

$$\text{"} \quad \text{"} \quad = \text{"}^1 \text{"}_1 + \text{"}^2 \text{"}_2 = \text{"}^{11}\text{"}_{11} + \text{"}^{12}\text{"}_{12} + \text{"}^{21}\text{"}_{21} + \text{"}^{22}\text{"}_{22}$$

$$= \sqrt{a}e_{11}\frac{e^{11}}{\sqrt{a}} + \sqrt{a}e_{12}\frac{e^{12}}{\sqrt{a}} + \sqrt{a}e_{21}\frac{e^{21}}{\sqrt{a}} + \sqrt{a}e_{22}\frac{e^{22}}{\sqrt{a}}$$

$$= 0 + 1 + (\quad 1) \quad (\quad 1) + 0 = 2 :$$

Therefore, from Eq. (6.53) we obtain

$$\text{"} \quad \text{"} \quad R \quad = \text{"} \quad \text{"} \quad \text{"} \quad \text{"} \quad = 4$$

or

$$= \frac{1}{4}R \quad \text{"} \quad \text{"} \tag{6.54}$$

from which it is evident that the Gaussian curvature is an invariant. Again, from (6.53), we get

$$a \quad a \quad R \quad = a^{11}a \quad R_{\;11} + a^{22}a \quad R_{\;22} + a^{12}a \quad R_{\;12} + a^{21}a \quad R_{\;21}$$

$$= a^{11}a^{22}R_{2112} + a^{12}a^{21}R_{2121} + a^{21}a^{12}R_{1212} + a^{22}a^{11}R_{1221}$$

$$= \quad a^{11}a^{22}R_{1212} + a^{12}a^{21}R_{1212} + a^{21}a^{12}R_{1212} \quad a^{22}a^{11}R_{1221}$$

$$= 2\,R_{1212} \; a^{12}a^{21} \quad a^{11}a^{22}$$

$$= \quad 2R_{1212} \left[\frac{a_{22}}{a}\frac{a_{11}}{a} \quad \frac{a_{12}}{a} \quad \frac{a_{12}}{a}\right]$$

$$= \quad 2R_{1212} \; \frac{a}{a^2} \; = \quad 2 \; ; \; a = |a \quad | = a_{11}a_{22} \quad a_{12}^2$$

or

$$= \quad \frac{1}{2} \; R; \quad \text{where}\, R = a \quad a \quad R \quad = a \quad R \quad : \tag{6.55}$$

R = R = a R is the Ricci tensor. The invariant R in Eq. (6.55) is sometimes defined as the scalar curvature or Einstein curvature of the surface S: A surface on which holds

$$R \quad = (a \; a \quad a \; a \;) \tag{6.56}$$

where κ is a scalar, is called a **surface of constant curvature**. In particular, if $\kappa = 0$, the surface is called a **flat surface**.

Geometrical properties which are expressible in terms of the first fundamental form, may be called **intrinsic Properties**. Since only metric coefficients $a_{\alpha\beta}$ are involved in the definition of κ in Eq. (6.55), the properties of κ are intrinsic properties of the surface.

EXAMPLE 6.6.1 If the co-ordinate system is orthogonal, show that

$$\kappa = -\frac{1}{2\sqrt{a}}\left[\frac{\partial}{\partial u^1}\left(\frac{1}{\sqrt{a}}\frac{\partial a_{22}}{\partial u^1}\right) + \frac{\partial}{\partial u^2}\left(\frac{1}{\sqrt{a}}\frac{\partial a_{11}}{\partial u^2}\right)\right].$$

Solution: If the system of co-ordinates is orthogonal, then $a_{12} = 0 = a_{21}$. Therefore,

$$a = \begin{vmatrix} a_{11} & 0 \\ 0 & a_{22} \end{vmatrix};\ a^{11} = \frac{1}{a_{11}};\ a^{22} = \frac{1}{a_{22}}.$$

The non-vanishing Christoffel symbols of the first kind are

$$[11;1] = \frac{1}{2}\frac{\partial a_{11}}{\partial u^1};\ [12;1] = \frac{1}{2}\frac{\partial a_{11}}{\partial u^2} = -[11;2]$$

$$[22;1] = -\frac{1}{2}\frac{\partial a_{22}}{\partial u^1} = -[12;2];\ [22;2] = \frac{1}{2}\frac{\partial a_{22}}{\partial u^2}.$$

The non-vanishing Christofell symbols of the second kind are

$$\left\{{1 \atop 1\ 1}\right\} = \frac{1}{2a_{11}}\frac{\partial a_{11}}{\partial u^1};\quad \left\{{1 \atop 1\ 2}\right\} = \frac{1}{2a_{11}}\frac{\partial a_{11}}{\partial u^2};\quad \left\{{1 \atop 2\ 2}\right\} = -\frac{1}{2a_{11}}\frac{\partial a_{22}}{\partial u^1};$$

$$\left\{{2 \atop 1\ 1}\right\} = -\frac{1}{2a_{22}}\frac{\partial a_{11}}{\partial u^2};\quad \left\{{2 \atop 1\ 2}\right\} = \frac{1}{2a_{22}}\frac{\partial a_{22}}{\partial u^1};\quad \left\{{2 \atop 2\ 2}\right\} = \frac{1}{2a_{22}}\frac{\partial a_{22}}{\partial u^2}.$$

Using formula (6.51), the Riemann tensor R_{1212} is given by

$$\begin{aligned} R_{1212} &= \frac{\partial}{\partial u^1}[22;1] - \frac{\partial}{\partial u^2}[21;1] + \left\{{\alpha \atop 2\ 1}\right\}[12;\alpha] - \left\{{\alpha \atop 2\ 2}\right\}[11;\alpha] \\ &= \frac{\partial}{\partial u^1}[22;1] - \frac{\partial}{\partial u^2}[21;1] + \left\{{1 \atop 2\ 1}\right\}[12;1] + \left\{{2 \atop 2\ 1}\right\}[12;2] \\ &\quad - \left\{{1 \atop 2\ 2}\right\}[11;1] - \left\{{2 \atop 2\ 2}\right\}[11;2] \end{aligned}$$

$$= \frac{1}{2}\frac{\partial}{\partial u^1}\left(\frac{\partial a_{22}}{\partial u^1}\right) \quad \frac{1}{2}\frac{\partial}{\partial u^2}\left(\frac{\partial a_{22}}{\partial u^2}\right) + \frac{1}{2a_{11}}\frac{\partial a_{11}}{\partial u^2}\cdot\frac{1}{2}\frac{\partial a_{11}}{\partial u^2}$$

$$+ \frac{1}{2a_{22}}\frac{\partial a_{22}}{\partial u^1} \quad \frac{1}{2}\frac{\partial a_{22}}{\partial u^1} + \frac{1}{2a_{11}}\frac{\partial a_{22}}{\partial u^1}\cdot\frac{1}{2}\frac{\partial a_{11}}{\partial u^1} + \frac{1}{2a_{22}}\frac{\partial a_{22}}{\partial u^2} \quad \frac{1}{2}\frac{\partial a_{11}}{\partial u^2}$$

$$= \frac{1}{2}\left(\frac{\partial^2 a_{11}}{(\partial u^2)^2} + \frac{\partial^2 a_{22}}{(\partial u^1)^2}\right) + \frac{1}{4a_{11}}\left[\left(\frac{\partial a_{11}}{\partial u^2}\right)^2 + \frac{\partial a_{11}}{\partial u^1}\frac{\partial a_{22}}{\partial u^1}\right]$$

$$+ \frac{1}{4a_{22}}\left[\left(\frac{\partial a_{22}}{\partial u^1}\right)^2 + \frac{\partial a_{11}}{\partial u^2}\frac{\partial a_{22}}{\partial u^2}\right]$$

$$= \frac{1}{2\sqrt{a_{11}a_{22}}}\left[\frac{\partial}{\partial u^1}\left(\frac{1}{\sqrt{a_{11}a_{22}}}\frac{\partial a_{22}}{\partial u^1}\right) + \frac{\partial}{\partial u^2}\left(\frac{1}{\sqrt{a_{11}a_{22}}}\frac{\partial a_{11}}{\partial u^2}\right)\right]$$

$$= \frac{1}{2\sqrt{a}}\left[\frac{\partial}{\partial u^1}\left(\frac{1}{\sqrt{a}}\frac{\partial a_{22}}{\partial u^1}\right) + \frac{\partial}{\partial u^2}\left(\frac{1}{\sqrt{a}}\frac{\partial a_{11}}{\partial u^2}\right)\right].$$

This is a representation of the Gaussian curvature in terms of the coefficients of the first fundamental form.

EXAMPLE 6.6.2 Calculate the Gaussian curvature for the surface with metric

$$ds^2 = a^2 \sin^2 u^1 (du^2)^2 + a^2 (du^1)^2.$$

Solution: Comparing the given metric with Eq. (6.10), we get $a_{11} = a^2, a_{22} = a^2 \sin^2 u^1$ and $a_{12} = 0 = a_{21}$, so that

$$\triangle = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a^4 \sin^2 u^1.$$

The non-vanishing Christoffel symbols of first kind are

$$[22, 1] = \frac{a^2}{2}\sin 2u^1; \quad [12, 2] = \frac{a^2}{2}\sin 2u^1 = [21, 2].$$

The reciprocal or conjugate tensors are

$$a^{11} = \frac{a^2 \sin^2 u^1}{a^4 \sin^2 u^1} = \frac{1}{a^2}, \quad a^{22} = \frac{a^2}{a^4 \sin^2 u^1} = \frac{1}{a^2 \sin^2 u^1}$$

and $a^{12} = 0 = a^{21}$. The non-vanishing Christoffel symbols of second kind are

$$\left\{ {1 \atop 2\;2} \right\} = a^{1k}[22, k] = a^{11}[22, 1] + a^{12}[22, 2] = \frac{1}{2}\sin 2u^1$$

$$\left\{ {1 \atop 2\;1} \right\} = a^{2k}[21, k] = a^{21}[21, 1] + a^{22}[21, 2] = \cot u^1 = \left\{ {1 \atop 1\;2} \right\}.$$

The Riemann tensor R_{1212} is given by

$$R_{1212} = a_1 \ R_{212} = a_{11}R^1_{212} + a_{12}R^2_{212} = a_{11}R^1_{212}$$

$$= a_{11} \ \frac{\partial}{\partial u^1} \ \genfrac{}{}{0pt}{}{1}{2\ 2} \quad \frac{\partial}{\partial u^2} \ \genfrac{}{}{0pt}{}{1}{2\ 1} \ + \ \genfrac{}{}{0pt}{}{}{2\ 2} \quad \genfrac{}{}{0pt}{}{1}{1} \quad \genfrac{}{}{0pt}{}{}{2\ 1} \quad \genfrac{}{}{0pt}{}{1}{2}$$

$$= a^2 \ \frac{\partial}{\partial u^1} \ \frac{1}{2}\sin 2u^1 \quad \frac{\partial}{\partial u^2} \ \cot u^1 \ + \ \genfrac{}{}{0pt}{}{1}{2\ 2} \ \genfrac{}{}{0pt}{}{1}{1\ 1} \ + \ \genfrac{}{}{0pt}{}{2}{2\ 2} \ \genfrac{}{}{0pt}{}{1}{2\ 1}$$

$$\genfrac{}{}{0pt}{}{1}{2\ 1} \ \genfrac{}{}{0pt}{}{1}{2\ 1} \quad \genfrac{}{}{0pt}{}{2}{2\ 1} \ \genfrac{}{}{0pt}{}{1}{2\ 2}$$

$$= a^2 \quad \cos 2u^1 \quad \cot u^1 \quad \frac{1}{2}\sin 2u^1 \quad = a^2 \sin^2 u^1.$$

The total curvature or the Gaussian curvature of the surface S is given by

$$= \frac{R_{1212}}{4} = \frac{a^2 \sin^2 u^1}{a^4 \sin^2 u^1} = \frac{1}{a^2}.$$

EXAMPLE 6.6.3 For a surface of revolution defined by $x^1 = u^1 \cos u^2$; $x^2 = u^1 \sin u^2$; $x^3 = f(u^1)$; where f is of class C^2, find the Gaussian curvature .

Solution: The parametric representation of the right helicoid is given by

$$x^1 = u^1 \cos u^2; \quad x^2 = u^1 \sin u^2; \quad x^3 = f(u^1)$$

i.e.

$$r = (x^1; x^2; x^3) = \quad u^1 \cos u^2; u^1 \sin u^2; f(u^1) \quad .$$

From this parametric representation, we get

$$\frac{\partial x^1}{\partial u^1} = \cos u^2; \ \frac{\partial x^2}{\partial u^1} = \sin u^2; \ \frac{\partial x^3}{\partial u^1} = \frac{\partial f}{\partial u^1} = f_1$$

$$\frac{\partial x^1}{\partial u^2} = \quad u^1 \sin u^2; \ \frac{\partial x^2}{\partial u^2} = u^1 \cos u^2; \ \frac{\partial x^3}{\partial u^2} = \frac{\partial f}{\partial u^2} = 0;$$

where $\frac{\partial f}{\partial u^1} = f_1$. Thus, the coefficients a of the first fundamental form are given by

$$a_{11} = \quad \frac{\partial x^1}{\partial u^1}^2 + \frac{\partial x^2}{\partial u^1}^2 + \frac{\partial x^3}{\partial u^1}^2 = 1 + f_1^2.$$

$$a_{22} = \quad \frac{\partial x^1}{\partial u^2}^2 + \frac{\partial x^2}{\partial u^2}^2 + \frac{\partial x^3}{\partial u^2}^2 = (u^1)^2.$$

$$a_{12} = \frac{\partial x^1}{\partial u^1}\frac{\partial x^1}{\partial u^2} + \frac{\partial x^2}{\partial u^1}\frac{\partial x^2}{\partial u^2} + \frac{\partial x^3}{\partial u^1}\frac{\partial x^3}{\partial u^2}$$

$$= \cos u^2 (\quad u^1 \sin u^2) + (u^1 \cos u^2) \sin u^2 + f_1 \quad 0 = 0 = a_{21}.$$

Therefore, $a = (1 + f_1^2)(u^1)^2$. The reciprocal tensors $a^{\alpha\beta}$ are given by

$$a^{11} = \frac{1}{1+f_1^2};\quad a^{22} = \frac{1}{(u^1)^2};\quad a^{12} = 0 = a^{21}.$$

The non-vanishing Christoffel symbols of the first kind are

$$[11;1] = \frac{1}{2}\frac{\partial a_{11}}{\partial u^1} = f_1 f_2;\quad [12;2] = \frac{1}{2}\frac{\partial a_{22}}{\partial u^1} = u^1 = [21;2]$$

$$[22;1] = -\frac{1}{2}\frac{\partial a_{22}}{\partial u^1} = -u^1;$$

where we use the notations $f_1 = \frac{df}{du^1}$ and $f_2 = \frac{d^2f}{(du^1)^2}$. The non-vanishing Christoffel symbols of the second kind are

$$\left\{ {1 \atop 1\,1} \right\} = a^{11}[11;1] = \frac{f_1 f_2}{1+f_1^2};\quad \left\{ {1 \atop 2\,2} \right\} = a^{11}[22;1] = -\frac{u^1}{1+f_1^2}$$

$$\left\{ {2 \atop 2\,1} \right\} = a^{22}[21;2] = \frac{1}{u^1} = \left\{ {2 \atop 1\,2} \right\}.$$

The Riemann tensor R_{1212} is given by

$$\begin{aligned}
R_{1212} &= a_{1\alpha} R^{\alpha}_{212} = a_{11}R^1_{212} + a_{12}R^2_{212} = a_{11}R^1_{212}\\
&= a_{11}\left[\frac{\partial}{\partial u^1}\left\{ {1 \atop 2\,2} \right\} - \frac{\partial}{\partial u^2}\left\{ {1 \atop 2\,1} \right\} + \left\{ {\alpha \atop 2\,2} \right\}\left\{ {1 \atop \alpha\,1} \right\} - \left\{ {\alpha \atop 2\,1} \right\}\left\{ {1 \atop \alpha\,2} \right\}\right]\\
&= a_{11}\left[\frac{\partial}{\partial u^1}\left(-\frac{u^1}{1+f_1^2}\right) + \left\{ {1 \atop 2\,2} \right\}\left\{ {1 \atop 1\,1} \right\} + \left\{ {2 \atop 2\,2} \right\}\left\{ {1 \atop 2\,1} \right\}\right.\\
&\qquad \left. - \left\{ {1 \atop 2\,1} \right\}\left\{ {1 \atop 2\,1} \right\} - \left\{ {2 \atop 2\,1} \right\}\left\{ {1 \atop 2\,2} \right\}\right]\\
&= (1+f_1^2)\left[-\frac{1}{1+f_1^2} + \frac{2u^1 f_1 f_2}{(1+f_1^2)^2} + \left(-\frac{u^1}{1+f_1^2}\right)\left(\frac{f_1 f_2}{1+f_1^2}\right) - \frac{1}{u^1}\left(-\frac{u^1}{1+f_1^2}\right)\right]\\
&= (1+f_1^2)\frac{u^1 f_1 f_2}{(1+f_1^2)^2} = \frac{u^1 f_1 f_2}{1+f_1^2}.
\end{aligned}$$

The total curvature K or the Gaussian curvature of the surface S is given by

$$K = \frac{R_{1212}}{4} = \frac{u^1 f_1 f_2}{1+f_1^2}\cdot\frac{1}{(1+f_1^2)(u^1)^2} = \frac{f_1 f_2}{u^1[1+(f_1)^2]^2}.$$

6.7 Isometry

The properties of surfaces (i.e. the lengths of curves, angle between intersecting curves), which have been already stated are expressed completely by means of first fundamental quadratic form ds^2 as given in Eq. (6.10). These properties constitute a body of what is known as the intrinsic geometry of surfaces. We have seen that intrinsic property of a surface depend on the metric tensor of the surface and its derivatives. But the metric of the surface is a local property and it may happen that two surfaces have the same metric.

Two surfaces S_1 and S_2 be such that there exists a co-ordinate system with respect to which the linear element on S_1 and S_2 are characterised by the same metric tensors a , then, they are said to be isometric, and the transformation of parameters is called an isometry. For example, the surfaces of the cylinder and cone are isometric with the Euclidean plane, since these surfaces can be rolled out, or developed, on the plane without changing the lengths of arc elements, and hence without altering the measurements of angles and areas.

The problem of finding surfaces isometric to a plane is of special importance. For this we now introduce the concept of a ruled surface and the developable surface. A surface is called a ruled surface (Figure 6.12) if it contains (at least) one one-parameter family of straight lines which can be chosen as co-ordinate curves on the surface. The straight lines are called generators of a ruled surface. One of the examples of a ruled surface is the hyperbolic paraboloid $\frac{(x^1)^2}{a^2} \quad \frac{(x^2)^2}{b^2} \quad x^3 = 0$ (Figure 6.13).

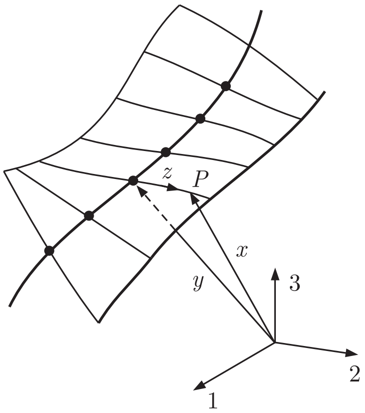

Figure 6.12: Portion of a ruled surface.

Consequently a ruled surface may be generated by a continuous motion of a straight line in space. Such a motion is completely determined if the path $y(s)$ (with arc length s) of a point of the moving line (Figure 6.12) is given and also the direction of the line for every values of s, for example, by a unit vector $z(s)$. A ruled surface may therefore

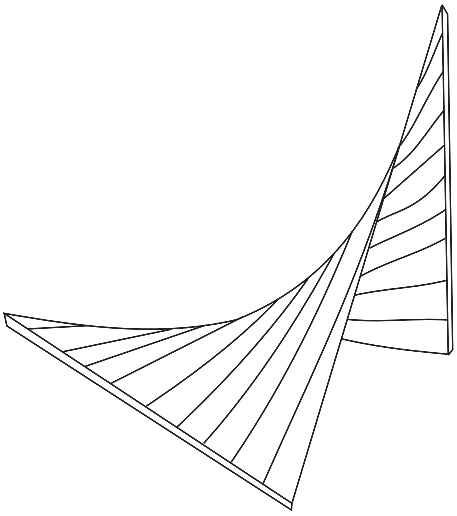

Figure 6.13: Hyperbolic paraboloid.

be represented in the form

$$\mathbf{x}(s;t) = \mathbf{y}(s) + t\mathbf{z}(s). \tag{6.57}$$

Obviously the co-ordinate t is the (directed) distance of the points of this surface Eq. (6.57) from the centre $\mathbf{y}(s)$, measured along the corresponding generator.

A surface which is isometric to a plane is called a *developable surface* or simply *developable*. A developable surface is a special ruled surface with the property that it has the same targent plane at all points on one and same generator. The tangent surface is always a developable surface. The principal normal surface and the binormal surface are developable surfaces if and only if the corresponding curve is plane. Developable surfaces are of particular importance because they are the only surfaces which can be mapped isometrically into a plane. The Gaussian curvature enables us to determine the circumstances under which a given surface is developable or not. The following theorem shows that, a developable surface is a surface which can be developed in a plane:

Theorem 6.7.1 *Prove the $K = 0$ is the necessary and sufficient condition for a surface to be a developable.*

Proof: The Gaussian curvature of the total curvature K of the surface S is given by

$$K = \frac{R_{1212}}{a}; \quad a = |a_{\alpha\beta}|;$$

where the tensor $R_{\alpha\beta\gamma\delta}$ is skew symmetric in the first two and last two indices. Now, when a surface S is isometric with the Euclidean plane, there exists on S a co-ordinate system with respect to which

$$a_{11} = a_{22} = 1;\quad a_{12} = 0 = a_{21} \Rightarrow a = 1 \neq 0.$$

In this case, the Riemann curvature tensor R = 0 for the surface S, in this particular co-ordinate system, and since R is a tensor, it must vanish in every co-ordinate system. Thus $R = 0$; and hence the Gaussian curvature = 0 for the surface S.

Conversely, let = 0, i.e. the Riemann tensors R vanishes at all points of the surface. A necessary and sufficient condition that a symmetric tensor a with $|a| \neq 0$ reduces under a suitable transformation of co-ordinates to a tensor h, where h are constants, is that the Riemann curvature tensor formed from the a s be a zero tensor.

Hence, it guarantees that there exists co-ordinate systems on the surface such that $a_{11} = a_{22} = 1;\ a_{12} = 0$. Hence, the surface is isometric with the Euclidean plane.

EXAMPLE 6.7.1 Prove that the following surfaces:

$$S_1 : y^1 = v^1 \cos v^2;\ y^2 = v^1 \sin v^2;\ y^3 = a \cosh^{-1} \frac{v^1}{a}$$

$$S_2 : y^1 = u^1 \cos u^2;\ y^2 = u^1 \sin u^2;\ y^3 = au^2$$

are isomorphic but non developable.

Solution: The first surface S_1 is the catenoid obtained by revolving the catenary $y^2 = \cosh \frac{y^3}{a}$ about the y^3 axis. Now,

$$a_{11} = \left(\frac{\partial y^i}{\partial v^1}\right)^2 = \left(\frac{\partial y^1}{\partial v^1}\right)^2 + \left(\frac{\partial y^2}{\partial v^1}\right)^2 + \left(\frac{\partial y^3}{\partial v^1}\right)^2$$

$$= (\cos v^2)^2 + (\sin v^2)^2 + \left(\frac{a}{\sqrt{(v^1/a)^2 - 1}}\,\frac{1}{a}\right)^2 = \frac{(v^1)^2}{(v^1)^2 - a^2}.$$

$$a_{22} = \left(\frac{\partial y^i}{\partial v^2}\right)^2 = \left(\frac{\partial y^1}{\partial v^2}\right)^2 + \left(\frac{\partial y^2}{\partial v^2}\right)^2 + \left(\frac{\partial y^3}{\partial v^2}\right)^2$$

$$= (-v^1 \sin v^2)^2 + (v^1 \cos v^2)^2 + 0 = (v^1)^2.$$

$$a_{12} = \frac{\partial y^1}{\partial v^1}\frac{\partial y^1}{\partial v^2} + \frac{\partial y^2}{\partial v^1}\frac{\partial y^2}{\partial v^2} + \frac{\partial y^3}{\partial v^1}\frac{\partial y^3}{\partial v^2}$$

$$= \cos v^2(-v^1 \sin v^2) + \sin v^2(v^1 \cos v^2) + \frac{a}{\sqrt{(v^1)^2 - a^2}}.0 = 0 = a_{21}.$$

Thus, the first fundamental quadratic form for the surface S_1 becomes,

$$\begin{aligned} ds^2 &= a_{\alpha\beta}\,dv^\alpha dv^\beta = a_{11}(dv^1)^2 + a_{22}(dv^2)^2 + 2a_{12}dv^1dv^2 \\ &= \frac{(v^1)^2}{(v^1)^2 - a^2}(dv^1)^2 + (v^1)^2(dv^2)^2 \\ &= \frac{(v^1)^2}{(v^1)^2 - a^2}(dv^1)^2 + [a^2 + \{(v^1)^2 - a^2\}](dv^2)^2. \end{aligned}$$

The second surface S_2 is a right helicoid. Now,

$$\begin{aligned} a_{11} &= \left(\frac{\partial y^i}{\partial u^1}\right)^2 = \left(\frac{\partial y^1}{\partial u^1}\right)^2 + \left(\frac{\partial y^2}{\partial u^1}\right)^2 + \left(\frac{\partial y^3}{\partial u^1}\right)^2 \\ &= (\cos u^2)^2 + (\sin u^2)^2 + 0 = 1\,; \\ a_{22} &= \left(\frac{\partial y^i}{\partial u^2}\right)^2 = \left(\frac{\partial y^1}{\partial u^2}\right)^2 + \left(\frac{\partial y^2}{\partial u^2}\right)^2 + \left(\frac{\partial y^3}{\partial u^2}\right)^2 \\ &= (-u^1\sin u^2)^2 + (u^1\cos u^2)^2 + a^2 = (u^1)^2 + a^2. \\ a_{12} &= \frac{\partial y^1}{\partial u^1}\frac{\partial y^1}{\partial u^2} + \frac{\partial y^2}{\partial u^1}\frac{\partial y^2}{\partial u^2} + \frac{\partial y^3}{\partial u^1}\frac{\partial y^3}{\partial u^2} \\ &= \cos u^2(-u^1\sin u^2) + \sin u^2(u^1\cos u^2) + 0\cdot a = 0. \end{aligned}$$

Thus the first fundamental quadratic form for the surface S_2 is given by

$$\begin{aligned} ds^2 &= a_{\alpha\beta}\,du^\alpha du^\beta = a_{11}(du^1)^2 + a_{22}(du^2)^2 + 2a_{12}du^1du^2 \\ &= (du^1)^2 + [a^2 + (u^1)^2](du^2)^2. \end{aligned}$$

Now, if we set $(v^1)^2 - a^2 = (u^1)^2$ and $v^2 = u^2$, then the two surfaces have the same metric. Thus, the surfaces S_1 and S_2 are isometric.

Now, we calculate the Gaussian curvature for the first surface S_1. The reciprocal tensors for the first fundamental quadratic form are

$$a^{11} = \frac{(v^1)^2 - a^2}{(v^1)^2};\; a^{22} = \frac{1}{(v^1)^2};\; a^{12} = 0 = a^{21}.$$

The non-vanishing Christoffel symbols of first kind are

$$\begin{aligned} [11;1] &= \frac{1}{2}\left(\frac{\partial a_{11}}{\partial v^1} + \frac{\partial a_{11}}{\partial v^1} - \frac{\partial a_{11}}{\partial v^1}\right) = \frac{1}{2}\frac{\partial a_{11}}{\partial v^1} = -\frac{a^2v^1}{[(v^1)^2 - a^2]^2} \\ [22;1] &= \frac{1}{2}\left(\frac{\partial a_{21}}{\partial v^2} + \frac{\partial a_{21}}{\partial v^2} - \frac{\partial a_{22}}{\partial v^1}\right) = -\frac{1}{2}\frac{\partial a_{22}}{\partial v^1} = -v^1 \\ [12;2] &= \frac{1}{2}\left(\frac{\partial a_{12}}{\partial v^2} + \frac{\partial a_{22}}{\partial v^1} - \frac{\partial a_{12}}{\partial v^2}\right) = \frac{1}{2}\frac{\partial a_{22}}{\partial v^1} = v^1 = [21;2]. \end{aligned}$$

The non-vanishing Christoffel symbols of second kind are

$$\left\{\begin{matrix}1\\1\;1\end{matrix}\right\} = a^{1k}[11;k] = a^{11}[11;1] + a^{12}[11;2]$$

$$= \frac{(v^1)^2 - a^2}{(v^1)^2}\left(-\frac{a^2v^1}{[(v^1)^2 - a^2]^2}\right) = -\frac{a^2}{v^1[(v^1)^2 - a^2]}$$

$$\left\{\begin{matrix}1\\2\;2\end{matrix}\right\} = a^{1k}[22;k] = a^{11}[22;1] + a^{12}[22;2]$$

$$= \frac{(v^1)^2 - a^2}{(v^1)^2}\left(-v^1\right) = -\frac{(v^1)^2 - a^2}{v^1}$$

$$\left\{\begin{matrix}2\\2\;1\end{matrix}\right\} = a^{2k}[21;k] = a^{21}[21;1] + a^{22}[21;2] = \frac{v^1}{(v^1)^2} = \frac{1}{v^1} = \left\{\begin{matrix}2\\1\;2\end{matrix}\right\}.$$

The Riemann tensor R_{1212} is given by

$$R_{1212} = a_{1\alpha}R^{\alpha}_{212} = a_{11}R^1_{212} + a_{12}R^2_{212} = a_{11}R^1_{212}$$

$$= a_{11}\left[\frac{\partial}{\partial v^1}\left\{\begin{matrix}1\\2\;2\end{matrix}\right\} - \frac{\partial}{\partial v^2}\left\{\begin{matrix}1\\2\;1\end{matrix}\right\} + \left\{\begin{matrix}\gamma\\2\;2\end{matrix}\right\}\left\{\begin{matrix}1\\\gamma\;1\end{matrix}\right\} - \left\{\begin{matrix}\gamma\\2\;1\end{matrix}\right\}\left\{\begin{matrix}1\\\gamma\;2\end{matrix}\right\}\right]$$

$$= a_{11}\left[\frac{\partial}{\partial v^1}\left(-\frac{(v^1)^2 - a^2}{(v^1)^2}\right) - 0 + \left\{\begin{matrix}1\\2\;2\end{matrix}\right\}\left\{\begin{matrix}1\\1\;1\end{matrix}\right\} + \left\{\begin{matrix}2\\2\;2\end{matrix}\right\}\left\{\begin{matrix}1\\2\;1\end{matrix}\right\}\right.$$

$$\left. - \left\{\begin{matrix}1\\2\;1\end{matrix}\right\}\left\{\begin{matrix}1\\2\;1\end{matrix}\right\} - \left\{\begin{matrix}2\\2\;1\end{matrix}\right\}\left\{\begin{matrix}1\\2\;2\end{matrix}\right\}\right]$$

$$= \frac{(v^1)^2}{(v^1)^2 - a^2}\left[-\frac{2a^2}{(v^1)^3} + \frac{a^2}{(v^1)^3} + \frac{(v^1)^2 - a^2}{(v^1)^3}\right] = \frac{(v^1)^2 - 2a^2}{v^1[(v^1)^2 - a^2]}.$$

The Gaussian curvature K of the surface S_1 is given by

$$K = \frac{R_{1212}}{a} = \frac{(v^1)^2 - 2a^2}{v^1[(v^1)^2 - a^2]}\,\frac{(v^1)^2 - a^2}{(v^1)^4} = \frac{(v^1)^2 - 2a^2}{(v^1)^5} \neq 0.$$

Thus, the surface S_1 is non-developable. Now, the reciprocal tensors for the first fundamental form to S_2 are

$$a^{11} = 1;\quad a^{22} = \frac{1}{a^2 + (u^1)^2};\quad a^{12} = 0 = a^{21}.$$

The non-vanishing Christoffel symbols of first and second kinds are

$$[22;1] = \frac{u^1}{[a^2 + (u^1)^2]^2};\quad [12;2] = -\frac{u^1}{[a^2 + (u^1)^2]^2} = [21;2].$$

$$\left\{\begin{matrix}1\\2\;2\end{matrix}\right\} = \frac{u^1}{[a^2 + (u^1)^2]^2};\quad \left\{\begin{matrix}2\\2\;1\end{matrix}\right\} = -\frac{u^1}{[a^2 + (u^1)^2]^3} = \left\{\begin{matrix}2\\1\;2\end{matrix}\right\}.$$

The Riemann tensor R_{1212} is given by

$$R_{1212} = a_{1}\ R_{212} = a_{11}R^{1}_{212} + a_{12}R^{2}_{212} = a_{11}R^{1}_{212}$$

$$= a_{11}\ \frac{@}{@u^1}\ \genfrac{}{}{0pt}{}{1}{2\ 2}\quad \frac{@}{@u^2}\ \genfrac{}{}{0pt}{}{1}{2\ 1} + \genfrac{}{}{0pt}{}{}{2\ 2}\quad \genfrac{}{}{0pt}{}{1}{1}\quad \genfrac{}{}{0pt}{}{}{2\ 1}\quad \genfrac{}{}{0pt}{}{1}{2}$$

$$= \frac{@}{@u^1}\ \frac{u^1}{[a^2+(u^1)^2]^2}\quad 0 + \genfrac{}{}{0pt}{}{1}{2\ 2}\ \genfrac{}{}{0pt}{}{1}{1\ 1} + \genfrac{}{}{0pt}{}{2}{2\ 2}\ \genfrac{}{}{0pt}{}{1}{2\ 1}$$

$$\genfrac{}{}{0pt}{}{1}{2\ 1}\ \genfrac{}{}{0pt}{}{1}{2\ 1}\quad \genfrac{}{}{0pt}{}{2}{2\ 1}\ \genfrac{}{}{0pt}{}{1}{2\ 2}$$

$$= \frac{a^2 \quad 3(u^1)^2}{[a^2+(u^1)^2]^3} + \frac{(u^1)^2}{[a^2+(u^1)^2]^5}:$$

The Gaussian curvature of the surface S_2 is given by

$$= \frac{R_{1212}}{a} = \frac{a^2 \quad 3(u^1)^2}{[a^2+(u^1)^2]^2} + \frac{(u^1)^2}{[a^2+(u^1)^2]^4} \neq 0:$$

Thus, the surface S_2 is non-developable.

EXAMPLE 6.7.2 Determine whether the surface with the metric $ds^2 = (u^2)^2(du^1)^2 + (u^1)^2(du^2)^2$ is a developable or not?

Solution: From the given metric, we see that $a_{11} = (u^2)^2; a_{22} = (u^1)^2; a_{12} = 0 = a_{21};$ so that $a = (u^1u^2)^2$: Now, the reciprocal tensors a are given by

$$a^{11} = \frac{1}{(u^2)^2};\ a^{22} = \frac{1}{(u^1)^2};\ a^{12} = 0 = a^{21}:$$

The non-vanishing Christoffel symbols of first and second kinds are

$$[11;2] = \quad u^2; [12;1] = u^2 = [21;1]; [12;2] = u^1 = [21;2]; [22;1] = \quad u^1:$$

$$\genfrac{}{}{0pt}{}{1}{2\ 1} = \frac{1}{u^2} = \genfrac{}{}{0pt}{}{1}{1\ 2};\quad \genfrac{}{}{0pt}{}{2}{2\ 1} = \frac{1}{u^1} = \genfrac{}{}{0pt}{}{2}{1\ 2};\quad \genfrac{}{}{0pt}{}{1}{2\ 2} = \frac{u^1}{(u^2)^2}:$$

The Riemann tensor R_{1212} is given by

$$R_{1212} = a_{1}\ R_{212} = a_{11}R^{1}_{212} + a_{12}R^{2}_{212} = a_{11}R^{1}_{212}$$

$$= a_{11}\ \frac{@}{@u^1}\ \genfrac{}{}{0pt}{}{1}{2\ 2}\quad \frac{@}{@u^2}\ \genfrac{}{}{0pt}{}{1}{2\ 1} + \genfrac{}{}{0pt}{}{}{2\ 2}\quad \genfrac{}{}{0pt}{}{1}{1}\quad \genfrac{}{}{0pt}{}{}{2\ 1}\quad \genfrac{}{}{0pt}{}{1}{2}$$

$$= (u^2)^2\ \frac{@}{@u^1}\ \frac{u^1}{(u^2)^2}\quad \frac{@}{@u^2}\ \frac{1}{u^2} + \genfrac{}{}{0pt}{}{1}{2\ 2}\ \genfrac{}{}{0pt}{}{1}{1\ 1} + \genfrac{}{}{0pt}{}{2}{2\ 2}\ \genfrac{}{}{0pt}{}{1}{2\ 1}$$

$$\genfrac{}{}{0pt}{}{1}{2\ 1}\ \genfrac{}{}{0pt}{}{1}{2\ 1}\quad \genfrac{}{}{0pt}{}{2}{2\ 1}\ \genfrac{}{}{0pt}{}{1}{2\ 2} = (u^2)^2 \quad 0 = 0:$$

The Gaussian curvature κ for the given surface is given by

$$\kappa = \frac{R_{1212}}{a} = \frac{0}{(u^1u^2)^2} = 0.$$

Thus, the surface is developable.

EXAMPLE 6.7.3 Show that the surface given by $x^1 = f_1(u^1)$; $x^2 = f_2(u^1)$; $x^3 = u^2$ is a developable, where f_1, f_2 are differentiable functions.

Solution: The first fundamental forms for the given surfaces are given by

$$a_{11} = \left(\frac{\partial x^i}{\partial u^1}\right)^2 = \left(\frac{\partial x^1}{\partial u^1}\right)^2 + \left(\frac{\partial x^2}{\partial u^1}\right)^2 + \left(\frac{\partial x^3}{\partial u^1}\right)^2 = (f_1')^2 + (f_2')^2.$$

$$a_{22} = \left(\frac{\partial x^i}{\partial u^2}\right)^2 = \left(\frac{\partial x^1}{\partial u^2}\right)^2 + \left(\frac{\partial x^2}{\partial u^2}\right)^2 + \left(\frac{\partial x^3}{\partial u^2}\right)^2 = 1.$$

$$a_{12} = \frac{\partial x^1}{\partial u^1}\frac{\partial x^1}{\partial u^2} + \frac{\partial x^2}{\partial u^1}\frac{\partial x^2}{\partial u^2} + \frac{\partial x^3}{\partial u^1}\frac{\partial x^3}{\partial u^2} = 0 = a_{21}.$$

Therefore, $a = (f_1')^2 + (f_2')^2$. The reciprocal tensors are given by

$$a^{11} = \frac{1}{(f_1')^2 + (f_2')^2};\ a^{22} = 1;\ a^{12} = 0 = a^{21}.$$

The only non-vanishing Christoffel symbols of first kind and second kinds are

$$[11, 1] = f_1'f_1'' + f_2'f_2'';\quad \left\{ {1 \atop 1\ 1} \right\} = \frac{f_1'f_1'' + f_2'f_2''}{(f_1')^2 + (f_2')^2}.$$

The Riemann tensor R_{1212} is given by

$$\begin{aligned} R_{1212} &= a_{1\alpha}R^{\alpha}_{212} = a_{11}R^1_{212} + a_{12}R^2_{212} = a_{11}R^1_{212} \\ &= a_{11}\left[\frac{\partial}{\partial u^1}\left\{ {1 \atop 2\ 2} \right\} - \frac{\partial}{\partial u^2}\left\{ {1 \atop 2\ 1} \right\} + \left\{ {\alpha \atop 2\ 2} \right\}\left\{ {1 \atop \alpha\ 1} \right\} - \left\{ {\alpha \atop 2\ 1} \right\}\left\{ {1 \atop \alpha\ 2} \right\}\right] \\ &= a_{11}\left[\frac{\partial}{\partial u^1}0 - \frac{\partial}{\partial u^2}0 + \left\{ {1 \atop 2\ 2} \right\}\left\{ {1 \atop 1\ 1} \right\} + \left\{ {2 \atop 2\ 2} \right\}\left\{ {1 \atop 2\ 1} \right\} \right. \\ &\quad \left. - \left\{ {1 \atop 2\ 1} \right\}\left\{ {1 \atop 1\ 2} \right\} - \left\{ {2 \atop 2\ 1} \right\}\left\{ {1 \atop 2\ 2} \right\}\right] = 0. \end{aligned}$$

The Gaussian curvature κ for the given surface is given by

$$\kappa = \frac{R_{1212}}{a} = \frac{0}{(f_1')^2 + (f_2')^2} = 0.$$

Thus, the surface is developable.

6.8 Geodesic Curvature

In this section we will describe the geodesic curvature of surface curves on a surface S of class 2. Here, using the intrinsic geometry of surfaces we will derive a formula describing the behaviour of the tangent vector to a surface curve. Let C be a surface curve given by a parametric representation

$$C: u \quad = u \quad (s); \qquad (6.58)$$

where s measures arc distance along the curve C. Then, at an arbitrary point P of the curve, we have the condition

$$a \quad \frac{du}{ds}\frac{du}{ds} = 1: \qquad (6.59)$$

The quantities $\frac{du^1}{ds}; \frac{du^2}{ds}$ obviously determine a tangent vector to C, and so writing the unit values for $\frac{du}{ds}$ we have from Eq. (6.59) that

$$a \quad \frac{du}{ds}\frac{du}{ds} = a \qquad = 1: \qquad (6.60)$$

Differentiate the quadratic relation (6.60) intrinsically with respect to s, we get

$$a \qquad \frac{\quad}{s} = 0$$

from which it follows that either the surface vector $\frac{\quad}{s}$ is orthogonal to or $\frac{\quad}{s} = 0$. If $\frac{\quad}{s} \neq 0$, we introduce a unit surface vector normal to (codirectional with $\frac{\quad}{s}$), (Figure 6.14), so that

$$\frac{\quad}{s} = {}_g \quad ; \qquad (6.61)$$

where $_g$ is a suitable scalar, positive or negative. We choose the sense of such that the rotation (;) is positive, i.e. = 1. The scalar $_g$ is called the geodesic curvature of the curve C on the surface. The geodesic curvature was introduced by Lionville.

(i) Now we give the geometrical significance of the geodesic curvature $_g$. Let C^0 be the orthogonal projection of C on the tangent plane to S at P (Figure 6.14). The geodesic curvature $_g$ of C at P is defined as the curvature of the projected curve C at P, taken with a suitable sign.

(ii) The sign of $_g$ is defined as follows: Suppose that the curvature of C^0 at P is not zero. Then $_g$ is positive sign if the centre of curvature C^0 at P lies in the direction of the unit vector = , where is the unit normal vector to S at P and is the tangent vector to C at P. $_g$ is negative sign if that centre of curvature lies in the direction opposite to the vector . The sign of $_g$ thus depends on the orientations of S and C. (In Figure 6.14 , $_g$ is negative).

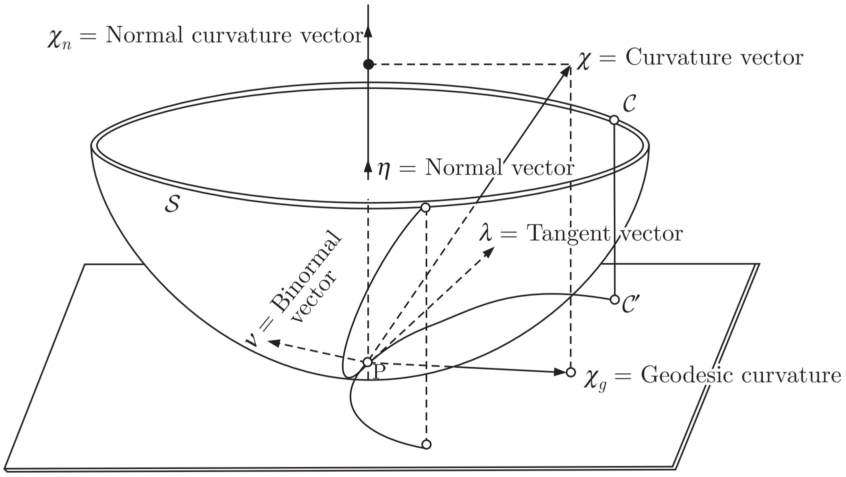
χ_n = Normal curvature vector
χ = Curvature vector
$\mathcal{C}$
η = Normal vector
$\mathcal{S}$
λ = Tangent vector
ν = Binormal vector
$\mathcal{C}'$
P
χ_g = Geodesic curvature

Figure 6.14: Geodesic curvature.

Since = " and = " , taking the intrinsic derivative, we get

$$\frac{\quad}{s} = " \; \frac{\quad}{s} = " \; \frac{\quad}{s} \; a$$

$$= " \; a \; \frac{\quad}{s} = " \; a \quad {}_g$$

or

$$\frac{\quad}{s} = {}_g" \qquad = {}_g" \qquad = {}_g$$

or

$$\frac{\quad}{s} = {}_g \; : \qquad (6.62)$$

We may refer this pair of Eqs. (6.61) and (6.62) as the Serret{Frenet formulae for the curve C relative to the surface. The transversion of $\frac{\quad}{s}$ = $_g$ by , we get

$${}_g = \frac{\quad}{s} = \frac{\quad}{s} \; " \qquad = " \qquad \frac{\quad}{s}$$

$$= \frac{1}{\sqrt{a}} e \qquad \frac{\quad}{s} = \frac{1}{\sqrt{a}} \; \begin{matrix} ^1 & ^2 \\ \frac{^1}{s} & \frac{^2}{s} \end{matrix} \; :$$

Alternatively, using $\frac{\quad}{s} = \ _g$; we get

$$ {}_g = " \quad \frac{\quad}{s} = \frac{1}{\sqrt{a}} \begin{array}{cc} {}^1 & {}^2 \\ {}^1 & {}^2 \\ \frac{\quad}{s} & \frac{\quad}{s} \end{array} $$

$$ = \sqrt{a} \begin{array}{cc} \frac{du^1}{ds} & \frac{du^2}{ds} \\ \frac{d^2u^1}{ds^2} + \ {}^1 \ \frac{du}{ds}\frac{du}{ds} & \frac{d^2u^2}{ds^2} + \ {}^2 \ \frac{du}{ds}\frac{du}{ds} \end{array} : \tag{6.63} $$

These expressions are known as Beltrami's formula for geodesics curvature.

Theorem 6.8.1 The necessary and sufficient condition for a curve on a surface to be a geodesic is that its geodesic curvature is zero.

Proof: The differential equation of the geodesic is

$$ \frac{d^2u}{ds^2} + \qquad \frac{du}{ds}\frac{du}{ds} = 0 $$

or

$$ \frac{d}{ds} \ \frac{du}{ds} \ + \qquad \frac{du}{ds}\frac{du}{ds} = 0 $$

or

$$ \frac{@}{@u} \ \frac{du}{ds} \ \frac{du}{ds} + \qquad \frac{du}{ds}\frac{du}{ds} = 0 $$

or

$$ \frac{@}{@u} \ \frac{du}{ds} \ + \qquad \frac{du}{ds} \ \frac{du}{ds} = 0 $$

or

$$ \frac{du}{ds} {}_{;} \ \frac{du}{ds} = 0 : $$

According to the intrinsic derivative of a tensor of type $(1;0)$ with component u , the above relation becomes

$$ \frac{\quad}{s} \ \frac{du}{ds} \ = 0 \ ) \ \frac{\quad}{s} \ = 0 $$

$$) \quad {}_g \ = 0; \text{ as } \frac{\quad}{s} = \ {}_g $$

$$) \quad {}_g = 0 \text{ as } \quad \text{is a unit vector.} $$

Hence, the curve on a surface is to be a geodesic if the geodesic curvature is zero.

Conversely, let ${}_g = 0$, identically, then we have

$$\frac{}{s}\frac{du}{ds} = \frac{}{s} = {}_g = 0$$

$$) \; \frac{du}{ds} = A \;) \; u = As + B:$$

This shows that, a geodesic of a surface is, therefore, analogous to a straight line in a plane. Therefore, a curve C on a surface S is a geodesic curve or geodesic if its geodesic curvature ${}_g$ vanishes identically. Straight lines on any surface are geodesics.

EXAMPLE 6.8.1 Find the geodesic curvature of the small circle

$$C: u^1 = \text{constant} = u^1_0 \, 6{=}\, 0; u^2 = u^2$$

on the surface of the sphere S.

Solution: On the surface of sphere

$$S: y^1 = a \cos u^1 \cos u^2; y^2 = a \cos u^1 \sin u^2; y^3 = a \sin u^1:$$

If the arc length s of C is measured from the plane $u_2 = 0$, we have $u^2 = \frac{s}{a \cos u^1_0}$, and the equation of C can be written in the form

$$u^1 = u^1_0; \quad u^2 = \frac{s}{a \cos u^1_0};$$

whose metric is $ds^2 = a^2 \; du^1 \;{}^2 + a^2 \; \cos u^1 \;{}^2 \; du^2 \;{}^2$; where the metric coefficients of s are $a_{11} = a^2; a_{12} = 0$ and $a_{22} = a^2 \cos^2 u^1$. The components of the unit tangent vector $= \frac{du}{ds}$ along C are given by

$$= {}^1; {}^2 = \frac{du^1}{ds}; \frac{du^2}{ds} = 0; \frac{1}{a \cos u^1_0} :$$

The only non-vanishing Christoffel symbols are

$$\begin{matrix}1\\2\;2\end{matrix} = a^{11}[22;1] = \frac{1}{a^{11}}\frac{@a_{22}}{@u^1} = \cos u^1_0 \sin u^1_0$$

$$\begin{matrix}2\\1\;2\end{matrix} = a^{22}[12;2] = \frac{1}{a^{22}}\frac{@a_{22}}{@u^1} = \tan u^1_0 = \begin{matrix}2\\2\;1\end{matrix} :$$

Using the definition of intrinsic derivative, we get

$$\frac{\delta\lambda^1}{\delta s} = \frac{d\lambda^1}{ds} + \Gamma^1_{\alpha\beta}\lambda^\alpha\frac{du^\beta}{ds} = \frac{d\lambda^1}{ds} + \Gamma^1_{11}\lambda^1\frac{du^1}{ds} + \Gamma^1_{12}\lambda^1\frac{du^2}{ds}$$

$$+\Gamma^1_{21}\lambda^2\frac{du^1}{ds} + \Gamma^1_{22}\lambda^2\frac{du^2}{ds} = \Gamma^1_{22}\lambda^2\lambda^2$$

$$\frac{\delta\lambda^2}{\delta s} = \frac{d\lambda^2}{ds} + \Gamma^2_{\alpha\beta}\lambda^\alpha\frac{du^\beta}{ds} = 0 + \Gamma^2_{22}\lambda^2\lambda^2$$

Using the formula (6.62) we get

$$\kappa_g\eta^1 = \frac{\delta\lambda^1}{\delta s} = \Gamma^1_{22}\lambda^2\lambda^2 = \cos u^1_0\sin u^1_0\left(\frac{1}{a\cos u^1_0}\right)^2 = \frac{1}{a^2}\tan u^1_0$$

$$\kappa_g\eta^2 = \Gamma^2_{22}\lambda^2\lambda^2 = 0.$$

Since C is not a geodesic, $\kappa_g \neq 0$ and we conclude that $\eta^2 = 0$. But η^2 is a unit vector so that $a_{\alpha\beta}\eta^\alpha\eta^\beta = 1$, i.e.

$$a_{11}\eta^1\eta^1 + 2a_{12}\eta^1\eta^2 + a_{22}\eta^2\eta^2 = 1$$

or

$$a^2\eta^1\eta^1 = 1 \Rightarrow \eta^1 = \frac{1}{a}.$$

Here κ_g is given by

$$\kappa_g = \frac{1}{\eta^1}\frac{1}{a^2}\tan u^1_0 = \frac{1}{a}\tan u^1_0.$$

Hence, the geodesic curvature is $\frac{1}{a}\tan u^1_0$.

EXAMPLE 6.8.2 Consider the surface of the right circular cone

$$C: x^1 = u^1\cos u^2;\ x^2 = u^1\sin u^2;\ x^3 = u^1;$$

where $u^1; u^2$ are the curvilinear co-ordinates on S and the curve C whose equation taken in the form $C: u^1 = a; u^2 = \frac{s}{a}$; where s is the arc parameter. Show that $\kappa_g = \frac{\sqrt{2}}{2\sqrt{a}}$.

Solution: Here, the metric is given by

$$ds^2 = 2\left(du^1\right)^2 + \left(u^1\right)^2\left(du^2\right)^2;$$

where the metric coefficients of S are $a_{11} = 2; a_{22} = \ u^1\ ^2; a_{33} = 0$ and $a_{12} = 0$ so that $a = 2\ u^1\ ^2$: The conjugate tensors are given by

$$a^{11} = \frac{1}{a_{11}} = 2; a^{22} = \frac{1}{(u^1)^2} \text{ and } a^{33} = 0;\ a^{12} = 0 = a^{21}:$$

The only non-vanishing Christoffel symbols of second kind are

$$\begin{matrix}1\\2\ 2\end{matrix} = \frac{1}{2a_{11}}\frac{@a_{22}}{@u^1} = \ u^1 = \ a:$$

Now, the components of the unit tangent vector $= \frac{du}{ds}$ along C are given by

$$= \ ^1;\ \ ^2 = \frac{du^1}{ds}; \frac{du^2}{ds} = \ 0; \frac{1}{a} \ :$$

Using the definition of intrinsic derivative, we get

$$\frac{\ ^1}{s} = \frac{d\ ^1}{ds} + \ \ 1 \ \ \frac{du}{ds}$$

$$= \frac{d\ ^1}{ds} + \begin{matrix}1\\1\ 1\end{matrix}\ {}_1\frac{du^1}{ds} + \begin{matrix}1\\2\ 2\end{matrix}\ {}_2\frac{du^2}{ds} + \begin{matrix}1\\2\ 1\end{matrix}\ {}_2\frac{du^1}{ds} + \begin{matrix}1\\1\ 2\end{matrix}\ {}_1\frac{du^2}{ds}$$

$$= \begin{matrix}1\\2\ 2\end{matrix}\ {}_2\frac{du^2}{ds} = \ \frac{1}{2a}$$

$$\frac{\ ^2}{s} = \frac{d\ ^2}{ds} + \ \ 2 \ \ \frac{du}{ds}$$

$$= \frac{d\ ^2}{ds} + \begin{matrix}2\\1\ 1\end{matrix}\ {}_1\frac{du^1}{ds} + \begin{matrix}2\\2\ 2\end{matrix}\ {}_2\frac{du^2}{ds} + \begin{matrix}2\\1\ 2\end{matrix}\ {}_2\frac{du^1}{ds} + \begin{matrix}2\\2\ 1\end{matrix}\ {}_1\frac{du^2}{ds} = 0:$$

Using formula (6.62), we get

$$_g\ ^1 = \frac{\ ^1}{s} = \ \frac{1}{2a} \text{ and } _g\ ^2 = \frac{\ ^2}{s} = 0:$$

Since C is not a geodesic $_g \neq 0$ and we conclude that $^2 = 0$. But ^{2} is a unit vector so that a $= 1$, i.e.

$$a_{11}\ ^1\ ^1 + a_{22}\ ^2\ ^2 + 2a_{12}\ ^1\ ^2 = 1$$

or

$$2\ \ ^1\ ^2 + \ u^1\ ^2\ \ ^2\ ^2 = 1 \ )\ \ ^1\ ^2 = \frac{1}{2}$$

$$)\ \ _g = \frac{1}{\ _1}\ \ \frac{1}{2a} = \ \frac{1}{\sqrt{2}}\frac{1}{2a} = \ \frac{1}{a2\sqrt{2}}:$$

EXAMPLE 6.8.3 Show that the geodesic curvature of the curve $u = c$ on a surface with metric $\lambda^2(du)^2 + \mu^2(dv)^2$ is $\frac{1}{\lambda\mu}\frac{\partial\mu}{\partial u}$.

Solution: For the given metric $ds^2 = \lambda^2(du)^2 + \mu^2(dv)^2$; the metric coefficients are $a_{11} = \lambda^2; a_{22} = \mu^2$ and $a_{12} = 0$ and so, $a = \lambda^2\mu^2$. Thus, the conjugate tensors are given by

$$a^{11} = \frac{a_{22}}{a} = \frac{1}{\lambda^2};\ a^{22} = \frac{a_{11}}{a} = \frac{1}{\mu^2};\ a^{12} = 0 = a^{21}.$$

The Christoffel symbols of first kind are

$$[11;1] = \frac{1}{2}\frac{\partial a_{11}}{\partial u^1} = \lambda\frac{d\lambda}{du};\ [22;2] = \frac{1}{2}\frac{\partial a_{22}}{\partial u^2} = \mu\frac{d\mu}{dv}$$

$$[11;2] = \frac{1}{2}\left(\frac{\partial a_{12}}{\partial u^1} + \frac{\partial a_{12}}{\partial u^1} - \frac{\partial a_{11}}{\partial u^2}\right) = -\lambda\frac{d\lambda}{dv}$$

$$[21;1] = \frac{1}{2}\left(\frac{\partial a_{21}}{\partial u^1} + \frac{\partial a_{11}}{\partial u^2} - \frac{\partial a_{21}}{\partial u^1}\right) = \frac{1}{2}\frac{\partial a_{11}}{\partial u^2} = \lambda\frac{d\lambda}{dv} = [12;1]$$

$$[22;1] = \frac{1}{2}\left(\frac{\partial a_{21}}{\partial u^2} + \frac{\partial a_{21}}{\partial u^2} - \frac{\partial a_{22}}{\partial u^1}\right) = -\frac{1}{2}\frac{\partial a_{22}}{\partial u^1} = -\mu\frac{d\mu}{du}$$

$$[12;2] = \frac{1}{2}\left(\frac{\partial a_{12}}{\partial u^2} + \frac{\partial a_{22}}{\partial u^1} - \frac{\partial a_{12}}{\partial u^2}\right) = \frac{1}{2}\frac{\partial a_{22}}{\partial u^1} = \mu\frac{d\mu}{du} = [21;2].$$

The Christoffel symbols of second kind are

$$\left\{ {1 \atop 1\ 1} \right\} = a^{1k}[11;k] = a^{11}[11;1] = \frac{1}{\lambda}\frac{d\lambda}{du}$$

$$\left\{ {1 \atop 2\ 1} \right\} = a^{1k}[21;k] = a^{11}[21;1] + a^{12}[21;2] = \frac{1}{\lambda}\frac{d\lambda}{dv} = \left\{ {1 \atop 1\ 2} \right\}$$

$$\left\{ {1 \atop 2\ 2} \right\} = a^{1k}[22;k] = a^{11}[22;1] + a^{12}[22;2] = -\frac{\mu}{\lambda^2}\frac{d\mu}{du}$$

$$\left\{ {2 \atop 2\ 1} \right\} = a^{2k}[21;k] = a^{21}[21;1] + a^{22}[21;2] = \frac{1}{\mu}\frac{d\mu}{du} = \left\{ {2 \atop 1\ 2} \right\}$$

$$\left\{ {2 \atop 2\ 2} \right\} = a^{2k}[22;k] = a^{21}[22;1] + a^{22}[22;2] = \frac{1}{\mu}\frac{d\mu}{dv}$$

$$\left\{ {2 \atop 1\ 1} \right\} = a^{2k}[11;k] = a^{21}[11;1] + a^{22}[11;2] = -\frac{\lambda}{\mu^2}\frac{d\lambda}{dv}$$

Let $_g$ be the geodesic curvature, then from Beltrami's formula (6.63), we have

$$
\begin{aligned}
{}_g &= \sqrt{a}\,\Bigg[\frac{du^1}{ds}\,\frac{d^2u}{ds^2} + \genfrac{}{}{0pt}{}{2}{}\,\frac{du}{ds}\,\frac{du}{ds}\qquad \frac{du^2}{ds}\,\frac{d^2u}{ds^2} + \genfrac{}{}{0pt}{}{1}{}\,\frac{du}{ds}\,\frac{du}{ds}\Bigg)\\
&= \sqrt{a}\,\frac{du}{ds}\,\frac{d^2v}{ds^2} + \genfrac{}{}{0pt}{}{2}{1\;1}\,\frac{du}{ds}^{2} + \genfrac{}{}{0pt}{}{2}{2\;2}\,\frac{dv}{ds}^{2} + 2\genfrac{}{}{0pt}{}{2}{1\;2}\,\frac{du}{ds}\frac{dv}{ds}\\
&\qquad \frac{dv}{ds}\,\frac{d^2u}{ds^2} + \genfrac{}{}{0pt}{}{1}{1\;1}\,\frac{du}{ds}^{2} + \genfrac{}{}{0pt}{}{1}{2\;2}\,\frac{dv}{ds}^{2} + 2\genfrac{}{}{0pt}{}{1}{1\;2}\,\frac{du}{ds}\frac{dv}{ds}\Bigg)\Bigg]
\end{aligned}
$$

Since for the v parametric curve, $u = c =$ constant, so,

$$
\begin{aligned}
{}_g &= \sqrt{a}\,\frac{dv}{ds}\,\genfrac{}{}{0pt}{}{1}{2\;2}\,\frac{dv}{ds}^{2} = \sqrt{a}\,\genfrac{}{}{0pt}{}{1}{2\;2}\,\frac{dv}{ds}^{3}\\
&= \sqrt{a}\,\genfrac{}{}{0pt}{}{1}{2\;2}\,\frac{1}{a_{22}\sqrt{a_{22}}};\ \text{as}\ \frac{dvdv}{ds^2} = \frac{dvdv}{a_{22}dvdv} = \frac{1}{a_{22}}\\
&= \frac{1}{2}\frac{\partial}{\partial u}(\ ^2)\qquad \frac{1}{\ ^3} = \frac{1}{\ }\frac{\partial}{\partial u}.
\end{aligned}
$$

EXAMPLE 6.8.4 Show that the condition that the parametric curves u^1 and u^2 on a portion of a surface be geodesic are respectively

$$\genfrac{}{}{0pt}{}{2}{1\;1} = 0 \ \text{and}\ \genfrac{}{}{0pt}{}{1}{2\;2} = 0.$$

Solution: Let $^{(1)}_g$ and $^{(2)}_g$ be the geodesic curvatures of the u^1 and u^2 parametric curves, respectively. Then from Beltrami's formula (6.63), we get

$$
\begin{aligned}
{}^{(1)}_g &= \sqrt{a}\,\Bigg[\frac{du^1}{ds}\,\frac{d^2u^2}{ds^2} + \genfrac{}{}{0pt}{}{2}{1\;1}\,\frac{du^1}{ds}^{2} + \genfrac{}{}{0pt}{}{2}{2\;2}\,\frac{du^2}{ds}^{2} + 2\genfrac{}{}{0pt}{}{2}{1\;2}\,\frac{du^1}{ds}\frac{du^2}{ds}\Bigg)\\
&\qquad \frac{du^2}{ds}\,\frac{d^2u^1}{ds^2} + \genfrac{}{}{0pt}{}{1}{1\;1}\,\frac{du^1}{ds}^{2} + \genfrac{}{}{0pt}{}{1}{2\;2}\,\frac{du^2}{ds}^{2} + 2\genfrac{}{}{0pt}{}{1}{1\;2}\,\frac{du^1}{ds}\frac{du^2}{ds}\Bigg)\Bigg] .
\end{aligned}
$$

Since for the u^1 parametric curve, $u^2 = c_1 =$ constant, so,

$$
\begin{aligned}
{}^{(1)}_g &= \sqrt{a}\,\frac{du^1}{ds}\,\genfrac{}{}{0pt}{}{2}{1\;1}\,\frac{du^1}{ds}^{2} = \sqrt{a}\,\genfrac{}{}{0pt}{}{2}{1\;1}\,\frac{du^1}{ds}^{3}\\
&= \sqrt{a}\,\genfrac{}{}{0pt}{}{2}{1\;1}\,\frac{1}{a_{11}\sqrt{a_{11}}};\ \text{as}\ \frac{du^1du^1}{ds^2} = \frac{du^1du^1}{a_{11}du^1du^1} = \frac{1}{a_{11}}.
\end{aligned}
$$

Similarly, for the u^2 parametric curve, $u^1 = c_2$ = constant, so,

$$\kappa_g^{(2)} = -\sqrt{a}\left\{\begin{matrix}1\\2\;2\end{matrix}\right\}\frac{1}{a_{22}\sqrt{a_{22}}}.$$

We can derive these results by using the formula

$$\kappa_g = \varepsilon_{\alpha\beta}\lambda^{\alpha}\frac{\delta\lambda^{\beta}}{\delta s};\quad \text{where } \lambda_{(1)} = \left(\frac{1}{\sqrt{a_{11}}}, 0\right) \text{ and } \lambda_{(2)} = \left(0, \frac{1}{\sqrt{a_{22}}}\right).$$

If the geodesic curvature vanishes identically, then the curve is a geodesic. Thus, from the relations

$$\kappa_g^{(1)} = \sqrt{a}\left\{\begin{matrix}2\\1\;1\end{matrix}\right\}\frac{1}{a_{11}\sqrt{a_{11}}} \text{ and } \kappa_g^{(2)} = -\sqrt{a}\left\{\begin{matrix}1\\2\;2\end{matrix}\right\}\frac{1}{a_{22}\sqrt{a_{22}}}$$

we get that parametric curves are geodesic if and only if

$$\left\{\begin{matrix}2\\1\;1\end{matrix}\right\} = 0 \text{ and } \left\{\begin{matrix}1\\2\;2\end{matrix}\right\} = 0;\ \text{as } a = |a_{\alpha\beta}| \neq 0.$$

EXAMPLE 6.8.5 Show that the condition that the parametric curves on a portion of a surface

$$\mathbf{r} = (c\cos u^2\cos u^1, c\cos u^2\sin u^1, c\sin u^2)$$

be geodesic.

Solution: In the representation of the sphere $\mathbf{r}(u^1, u^2)$, the co-ordinates u^1, u^2 are orthogonal. We have

$$a_{11} = c^2\cos^2 u^2;\ a_{22} = c^2;\ \frac{\partial a_{22}}{\partial u^1} = 0;\ \text{and } a_{12} = 0 = a_{21}.$$

Therefore, the Christoffel symbol $\left\{\begin{matrix}1\\2\;2\end{matrix}\right\}$ is given by

$$\begin{aligned}\left\{\begin{matrix}1\\2\;2\end{matrix}\right\} &= \frac{1}{2a}\left[a_{22}\left(2\frac{\partial a_{12}}{\partial u^2} - \frac{\partial a_{22}}{\partial u^1}\right) - a_{12}\frac{\partial a_{22}}{\partial u^2}\right]\\ &= \frac{1}{2c^4\cos^2 u^2}\left[c^2(2\cdot 0 - 0) - 0\cdot 0\right] = 0\end{aligned}$$

that is, the curves u^1 = constant are geodesics. Now,

$$\begin{aligned}\left\{\begin{matrix}2\\1\;1\end{matrix}\right\} &= \frac{1}{2a}\left[a_{11}\left(2\frac{\partial a_{12}}{\partial u^1} - \frac{\partial a_{11}}{\partial u^2}\right) - a_{12}\frac{\partial a_{11}}{\partial u^1}\right]\\ &= \frac{1}{2c^4\cos^2 u^2}\left[c^2\cos^2 u^2(2\cdot 0 + 2c^2\cos u^2\sin u^2) - 0\cdot 0\right]\\ &= \cos u^2\sin u^2 = 0;\ \text{if } u^2 = 0\end{aligned}$$

that is, along the equator which is the only geodesic of the curves u^2 = constant. At the poles $u^2 = \pm\frac{\pi}{2}$ and also $\Gamma^2_{11} = 0$ but

$$x^1 = c\cos u^2 \cos u^1;\ x^2 = c\cos u^2 \sin u^1;\ x^3 = c\sin u^2:$$

is not valid at these points. Furthermore we see from

$$(\kappa_g)_{u^1=\text{constant}} = \frac{1}{\sqrt{a_{11}}}\frac{\partial}{\partial u^1}(\log\sqrt{a_{22}})$$
$$(\kappa_g)_{u^2=\text{constant}} = -\frac{1}{\sqrt{a_{22}}}\frac{\partial}{\partial u^2}(\log\sqrt{a_{11}})$$

that the co-ordinate curves of orthogonal co-ordinates are geodesics if and only if a_{11} does not depend on u^2 and a_{22} does not depend on u^1.

6.9 Exercises

1. On the surface of revolution $r = (u\cos\theta; u\sin\theta; f(u))$ what are the parametric curves for u = constant and what are the curves θ = constant.
2. On the right helicoid given by $r = (u\cos\theta; u\sin\theta; c\theta)$; show that the parametric curves are circular helices and straight lines.
3. On the hyperboloid of one sheet

   $$\frac{x}{a} = \frac{\sigma+\tau}{1+\sigma\tau};\ \frac{y}{b} = \frac{1-\sigma\tau}{1+\sigma\tau};\ \frac{z}{c} = \frac{\sigma-\tau}{1+\sigma\tau};$$

   the parametric curves are the generators. What curves are represented by $\sigma = \tau$ and by $\sigma\tau$ = constant?
4. What types of surface are determined by the following representations:
   (a) $r = (x^1; x^2; x^3) = (0; u^1; u^2)$;
   (b) $r = (x^1; x^2; x^3) = (u^1 + u^2; u^1 + u^2; u^1)$;
   (c) $r = (x^1; x^2; x^3) = (a\cos u^1; a\sin u^1; u^2)$:
   Investigate the behaviour of the corresponding matrices and find representation of the form (6.1).
5. Find a parametric representation of the cylinder generated by a straight line G which moves along a curve.

   $$C: x(s) = (h_1(s); h_2(s); h_3(s))$$

   and is parallel to the x^3 axis; consider the corresponding matrix.

6. Find a representation of the form (6.1) of the following surfaces
   - (a) Ellipsoid: $r = (a\cos u^2 \cos u^1; b\cos u^2 \sin u^1; c\sin u^2)$:
   - (b) Elliptic paraboloid: $r = (au^1 \cos u^2; bu^1 \sin u^2; (u^1)^2))$:
   - (c) Hyperbolic paraboloid : $r = (au^1 \cosh u^2; bu^1 \sinh u^2; (u^1)^2)$:
   - (d) Hyperboloid of two sheets : $r = (a\sinh u^1 \cos u^2; b\sinh u^1 \sin u^2; c\cosh u^1)$:

   What kind of co-ordinate curves do you have in each case?

7. Show that the first fundamental form for
   - (a) the surface $x^1 = a\cos u^1; x^2 = a\sin u^1; x^3 = u^2$ is given by

     $$ds^2 = a^2(du^1)^2 + (du^2)^2:$$

   - (b) the surface $x^1 = u^1 \cos u^2;\ x^2 = u^1 \sin u^2;\ x^3 = 0$ is given by

     $$ds^2 = (du^1)^2 + (u^1)^2(du^2)^2:$$

   - (c) the surface $x^1 = c\cos u^1 \sin u^2;\ x^2 = c\sin u^1 \sin u^2;\ x^3 = c\cos u^2$ is given by

     $$ds^2 = c^2 \sin^2 u^2(du^1)^2 + c^2 \cos^2 u^2(du^2)^2:$$

   - (d) the surface of revolution $x^1 = u^1 \cos u^2;\ x^2 = u^1 \sin u^2;\ x^3 = f(u^1)$ is given by

     $$ds^2 = (1 + f_1^2)(du^1)^2 + (u^1)^2(du^2)^2:$$

   - (e) the paraboloid $x^1 = u^1; x^2 = u^2;\ x^3 = (u^1)^2 - (u^2)^2$ is given by

     $$ds^2 = [1 + 4(u^1)^2](du^1)^2 - 4u^1u^2du^1du^2 + [1 + 4(u^2)^2](du^2)^2:$$

   - (f) the sphere $x^1 = c\cos u^1 \cos u^2; x^2 = c\cos u^1 \sin u^2; x^3 = c\sin u^1$ is given by

     $$ds^2 = c^2(du^1)^2 + c^2 \cos^2 u^1(du^2)^2:$$

   - (g) a plane with respect to polar co-ordinates.

8. Prove that on the surface given by

   $$x^1 = a(u^1 + u^2);\ x^2 = b(u^1 - u^2);\ x^3 = u^1u^2$$

   the parametric curves are straight lines.

9. Determine the first fundamental form of the cylinder

$$r = (x^1; x^2; x^3) = h_1(u^1); h_2(u^1); u^2$$

and of a cylinder of revolution.

10. Prove that at regular points of a surface the first fundamental form is positive definite.

11. If be the angle between the parametric curves lying on a surface, immersed in E_3, show that

$$\sin = \frac{\sqrt{a}}{\sqrt{a_{11}}\sqrt{a_{22}}} \text{ and } \tan = \frac{\sqrt{a}}{a_{12}}$$

and hence, show that the parametric curves on a surface are orthogonal if and only if $a_{12} = 0$:

12. Show that the area of the anchor ring

$$r = ((b + a\cos u)\cos v; (b + a\cos u)\sin v; a\sin v)$$

where $0 < u < 2;\ 0 < v < 2$ is $4\,{}^2ab$:

13. Prove that a necessary and sufficient condition that a surface S be isometric with the Euclidean plane is that the Riemann tensor (or the Gaussian curvature of S) be identically zero.

14. (a) Prove that " and " can be obtained from the other by raising or lowering of indices with the help of the metric tensors.
    (b) If ; are two unit vectors the rotation ; is positive, show that

$$\sin = " \quad :$$

    (c) If ; are two parallel vector fields, show that " remains constant.

15. Prove that the vector obtained by parallel propagation of the tangent vector to a geodesic always remains tangent to the geodesic.

16. Find the differential equations of the geodesic in (i) Rectangular co-ordinates; (ii) Spherical co-ordinates; (iii) Cylindrical co-ordinates.

17. Show that an elliptic helix is not in general a geodesic on an elliptic cylinder.

18. Show that the differential equations of the geodesic for

(a) the metric $ds^2 = (dx^1)^2 + [(x^1)^2 + c^2](dx^2)^2$ are

$$\frac{d^2x^1}{ds^2} \quad x^1 \quad \frac{dx^2}{ds} \quad ^2 = 0; \ \frac{d^2x^2}{ds^2} + \frac{2x^1}{(x^1)^2 + c^2}\frac{dx^1}{ds}\frac{dx^2}{ds} = 0:$$

(b) the metric $ds^2 = (du)^2 + (\sin u)^2(dv)^2$ are

$$\frac{d^2u}{ds^2} \quad \sin u \cos u \quad \frac{dv}{ds} \quad ^2 = 0; \ \frac{d^2v}{ds^2} + 2 \cot u \frac{du}{ds}\frac{dv}{ds} = 0:$$

19. Obtain the differential equations of geodesics for metric
    (i) $ds^2 = f(x)dx^2 + dy^2 + dz^2 + \frac{1}{f(x)}dt^2$:
    (ii) $ds^2 = e^{2kt} \ dx^2 + dy^2 + dz^2 \ + dt^2$:

20. Show that the geodesics of a sphere of radius a is determined by the equation

$$y^1 = a \cos u^1 \cos u^2; \ y^2 = a \cos u^1 \sin u^2; \ y^3 = a \sin u^1;$$

are great circles, i.e. the great circles on sphere are geodesics.

21. Determine the components of curvature tensors
    (i) of a surface of revolution $r = (u^2 \cos u^1; u^2 \sin u^1; f(u^2))$;
    (ii) of a right conoid $r = (u^2 \cos u^1; u^2 \sin u^1; f(u^1))$;
    (iii) of a surface represented in the form $x_3 = f(x_1; x_2)$.

22. On the surface generated by the tangents to a twisted curve, find the differential equation of the curves which cut the generators at a constant angle .

23. If v are geodesic co-ordinates in the neighbourhood of a point if they are subjected to the transformation

$$u \ = v \ + \frac{1}{6}C \qquad v \ v \ v \ ;$$

where C's are constants, then show that u are geodesic co-ordinates in the neighbourhood of the origin O.

24. (a) If X is a vector field satisfying $\frac{X}{t} = 0$; show that a X X remains constant along the curve. Deduce that the magnitudes of the parallel vector fields are equal.
    (b) Prove that the unit tangent vector $= \frac{du}{ds}$ of a geodesic is a parallel vector field along the geodesic.

(c) Prove that the necessary and sufficient condition that a system of co-ordinates be geodesic with pole at P are that their second covariant derivatives with respect to the metric of the space all vanish at P.

(d) Prove that the geodesic curvature of a curve C on a surface S depends on the first fundamental form of S only.

25. Prove that the co-ordinate system u defined by

$$u = v \quad u_0 + \frac{1}{2} \quad {}_{P} (v \quad u_0)(v \quad u_0);$$

is a geodesic co-ordinate system with the pole at $P(u_0^1; u_0^2)$.

26. Prove that the catenoid,

$$ds^2 = a^2 \cosh^2 x^1 (dx^1)^2 + a^2 \cosh^2 x^2 (dx^2)^2$$

and the right helicoid

$$ds^2 = (dy^1)^2 + [(y^1)^2 + a^2](dy^2)^2$$

are locally isometric.

27. Prove that, if the system of co-ordinates is orthogonal, the non-vanishing Christofell symbols of the second kind are

$$\genfrac{}{}{0pt}{}{1}{1\ 1} = \frac{1}{2a_{11}}\frac{\partial a_{11}}{\partial u^1}; \quad \genfrac{}{}{0pt}{}{1}{1\ 2} = \frac{1}{2a_{11}}\frac{\partial a_{11}}{\partial u^2}; \quad \genfrac{}{}{0pt}{}{1}{2\ 2} = \frac{1}{2a_{11}}\frac{\partial a_{22}}{\partial u^1}$$

$$\genfrac{}{}{0pt}{}{2}{1\ 1} = \frac{1}{2a_{22}}\frac{\partial a_{11}}{\partial u^2}; \quad \genfrac{}{}{0pt}{}{2}{1\ 2} = \frac{1}{2a_{22}}\frac{\partial a_{22}}{\partial u^1}; \quad \genfrac{}{}{0pt}{}{2}{2\ 2} = \frac{1}{2a_{22}}\frac{\partial a_{22}}{\partial u^2}.$$

28. Prove that for a helicoids of non-zero pitch the sections by planes containing the axis are geodesics if and only if these sections are straight lines.

29. Prove that, on a general surface, a necessary and sufficient condition that the curve $v = c$ be a geodesic is $EE_2 + FE_1 \quad 2EF_1 = 0$ for all values of u, where $E = r_1^2; F = r_1 \cdot r_2; G = r_2^2$.

30. Prove that if a surface admits two orthogonal families of geodesics, it is isometric with the plane.

31. Show that, if the line element is of the form $ds^2 = (du^1)^2 + a_{22}(du^2)^2$, the Gaussian curvature is

$$= \frac{1}{\sqrt{a_{22}}}\frac{\partial^2}{(\partial u^1)^2}\sqrt{a_{22}}.$$

32. Prove that the Gaussian curvature is an invariant.

33. Find the Gaussian curvature of the surface given (in Monge form) by $z = f(x; y)$, where x and y are taken as parameters.

34. Show that the Gaussian curvature is $\frac{1}{}\frac{@^2}{@u^2}$ for a surface with the metric

$$ds^2 = (\ du)^2 + \quad {}^2(dv)^2$$

35. Show that for the sphere of radius c, with the equation of the form

$$x^1 = c \sin u^1 \cos u^2; x^2 = c \sin u^1 \sin u^2; \ x^3 = c \cos u^1;$$

where c is a constant, the total curvature is $\quad = \frac{1}{c^2}$.

36. Show that the surface given by

$$x^1 = f_1(u^1); \ x^2 = f_2(u^1); \ x^3 = u^2$$

is a developable, where $f_1; f_2$ are differentiable functions.

37. Determine whether the surface of a helicoid given by

$$x^1 = u^1 \cos u^2; \ x^2 = u^1 \sin u^2; x^3 = cu^1$$

is developable, where c is a constant and $u^1; u^2$ are curvilinear co-ordinates of the surface.

38. Show that the surface with the metric

    (a) $ds^2 = (\ du^1)^2 + [(\ du^1)^2 + \ c^2](du^2)^2$ is not developable.
    (b) $ds^2 = (\ u^2)^2(du^1)^2 + (\ u^1)^2(du^2)$ is developable.

39. Find the equations of the surface of revolution for which $ds^2 = du^2 + (\ a^2 \quad u^2)dv^2$.

40. Prove that a ruled surface $x(s; t) = y(s) + tz(s)$ is a developable surface if and only if $j[\underline{y;}\, z; \underline{z}]j = 0$, where a dot denotes the derivatives with respect to the arc length s of the curve $y(s)$ and [.,.,.] denotes the scalar triple product.

41. Find an analytic expression of the Gaussian curvature of a ruled surface and prove that a ruled surface is developable surface if and only if its Gaussian curvature vanishes.

42. Show that

    (i) $\quad_g = \quad " \qquad \frac{u}{ds} = " \qquad \frac{u}{ds}:$

(ii) $\,^2_g = a \quad \frac{}{s}\frac{}{s}$; and $\,_g = \varepsilon \quad \frac{}{s}$

(iii) $\frac{d^2u}{s^2} + \quad \frac{du}{s}\frac{du}{s} = \quad _g\varepsilon \quad a \quad \frac{du}{s}$:

43. Show that the geodesic curvature of the curve $u = c$ with the metric

$$ds^2 = \,^2(du)^2 + \,^2(dv)^2 \text{ is } \frac{1}{}\frac{\partial}{\partial u}:$$

44. When the parametric curves are orthogonal, prove that

$$^{(1)}_g = \frac{1}{2\sqrt{a_{22}}}\frac{\partial \log a_{11}}{\partial u^2} \text{ and } ^{(2)}_g = \frac{1}{2\sqrt{a_{11}}}\frac{\partial \log a_{22}}{\partial u^1}:$$

45. Deduce that

$$^{(1)}_g = \frac{1}{\sqrt{a}}\quad\frac{\partial}{\partial u^1}\quad\frac{a_{12}}{\sqrt{a_{11}}}\quad\frac{\partial\sqrt{a_{11}}}{\partial u^2}\quad ; \quad ^{(2)}_g = \frac{1}{\sqrt{a}}\quad\frac{\partial}{\partial u^2}\quad\frac{a_{12}}{\sqrt{a_{22}}} + \frac{\partial\sqrt{a_{22}}}{\partial u^1}\quad :$$

46. Prove that the geodesic curvatures for the co-ordinate curves are

$$^{(1)}_g = \frac{\sqrt{a}}{(a_{11})^{3/2}}\quad {}^{\;2}_{1\;1} \quad\text{and}\quad ^{(2)}_g = \frac{\sqrt{a}}{(a_{22})^{3/2}}\quad {}^{\;1}_{2\;2} \quad :$$

47. Prove that the necessary and sufficient condition for a curve on a surface to be a geodesic is that its geodesic curvature be zero.

48. Show that the formula for the Gaussian curvature can be written in the form

$$= \frac{1}{2\sqrt{a}}\quad\frac{\partial}{\partial u^1}\quad\frac{a_{12}}{a_{11}\sqrt{a}}\frac{\partial a_{11}}{\partial u^2}\quad\frac{1}{\sqrt{a}}\frac{\partial a_{22}}{\partial u^1}$$

$$+\frac{\partial}{\partial u^2}\quad\frac{2}{\sqrt{a}}\frac{\partial a_{12}}{\partial u^1}\quad\frac{1}{\sqrt{a}}\frac{\partial a_{11}}{\partial u^2}\quad\frac{a_{12}}{a_{11}\sqrt{a}}\frac{\partial a_{11}}{\partial u^1}\quad :$$

49. Prove that any circular helix C on a cylinder of revolution S is a geodesic on C.

50. Prove that if a geodesic on a surface of revolution cuts the meridians at a constant angle, the surface is a right cylinder.

51. Prove that the meridians of a ruled helicoid are geodesics.

52. Prove that the orthogonal trajectories of the helices on a helicoid are geodesics.

53. Show that if a geodesic on a surface of revolution cuts the medians at a constant angle, the surface is a right cylinder.

Chapter 7

Surfaces in Space

With the exception of occasional references to the surrounding space, our study of geometry of surfaces was carried out from the point of view of a two-dimensional being whose universe is determined by the surface parameters u^1 and u^2. The treatment of surfaces in space presented in this chapter was based entirely on the study of the first quadratic differential form. In this chapter, the geometric shape and properties of a surface have discussed in the neighbourhood of any of its points. Of curse, this problem is of fundamental importance.

7.1 The Tangent Vector

We are now going to investigate the properties of a surface in its relation to the surrounding space. Consequently, we are dealing with two distinct system of co-ordinates namely, the three orthogonal Cartesian co-ordinates for the surrounding space which we denote by $y^i; i = 1, 2, 3$ and the two curvilinear co-ordinates of the surface which we denote by $u^\alpha; \alpha = 1, 2$. Let us denote the equation of a surface S embedded in E^3 by

$$y^1 = y^1(u^1, u^2); \; y^2 = y^2(u^1, u^2); \; y^3 = y^3(u^1, u^2)$$

$$y^i = y^i(u^1, u^2); \; i = 1, 2, 3. \tag{7.1}$$

The line element of arc ds of a curve lying on S is determined by the formula,

$$ds^2 = a_{\alpha\beta}\, du^\alpha du^\beta; \text{ where } a_{\alpha\beta} = \frac{\partial y^i}{\partial u^\alpha}\frac{\partial y^i}{\partial u^\beta}. \tag{7.2}$$

The choice of Cartesian variables y^i in the space enveloping the surface is clearly not essential, and we could have equally well referred the points of E^3 to a curvilinear co-ordinate system X related to Y by the transformation

$$x^i = x^i(y^1, y^2, y^3).$$

Now, relative to the frame X , the line element in E^3 is given by

$$ds^2 = g_{mn}\,dx^m dx^n; \quad \text{where, } g_{mn} = \frac{\partial y^k}{\partial x^m}\frac{\partial y^k}{\partial x^n} \tag{7.3}$$

and the set of Eq. (7.1) for the surface S can be written as

$$S : x^i = x^i(u^1, u^2); \; i = 1, 2, 3. \tag{7.4}$$

If we take a small displacement du on the surface, the corresponding component of the displacement in space are given by

$$dx^i = \frac{\partial x^i}{\partial u}du \tag{7.5}$$

and hence expression (7.3) for the surface element of arc assumes the form

$$ds^2 = g_{ij}\,dx^i dx^j = g_{ij}\frac{\partial x^i}{\partial u}\frac{\partial x^j}{\partial u}\,du\;du \,: \tag{7.6}$$

A comparison of Eq. (7.6) with Eq. (7.2) leads to the conclusion that

$$a = g_{ij}\frac{\partial x^i}{\partial u}\frac{\partial x^j}{\partial u}; \; i, j = 1, 2, 3 \text{ and } ; \; = 1, 2. \tag{7.7}$$

Now, dx^i is a space vector and is surface invariant, i.e. its components are unaltered Gaussian co-ordinates alone one transformed. Similarly, du is a surface vector and is also a space invariant. Hence, if we regard [Eq. (7.5)] first from the point of view of a transformation of space co-ordinates and then from the point of view of a transformation of surface co-ordinates and then from the point of view of a transformation of surface co-ordinates, we see that $\frac{\partial x^i}{\partial u}$ is a contravariant space vector and also a covariant surface vector, so that, it may be represented by

$$dx^i = \frac{\partial x^i}{\partial u}du = x^i\,du \;; \quad \text{where, } x^i = \frac{\partial x^i}{\partial u}; \tag{7.8}$$

where the indices properly describe the tensor character of this set of quantities.

Let r be the position vector of an arbitrary point P on S. The point P is determined by a pair of Gaussian co-ordinates (u^1, u^2), or by a triplet of space co-ordinates (x^1, x^2, x^3). Accordingly, the vector r can be viewed as a function of the space variables x^i satisfying Eq. (7.4). Thus,

$$\frac{\partial}{\partial u} = \frac{\partial}{\partial x^i}\frac{\partial x^i}{\partial u}. \tag{7.9}$$

But $\frac{\partial \mathbf{r}}{\partial x^i}$ are the base vectors $\mathbf{b}_i$ at P, associated with the curvilinear co-ordinate system X, whereas $\frac{\partial \mathbf{r}}{\partial u^\alpha}$ are the base vectors $\mathbf{a}_\alpha$ at P relative to the Gaussian co-ordinate system U. Hence, Eq. (7.9) yields

$$\mathbf{a}_\alpha = \mathbf{b}_i \frac{\partial x^i}{\partial u^\alpha}. \tag{7.10}$$

It is clear from this representation that

$$\mathbf{a}_1 = \mathbf{b}_i \frac{\partial x^i}{\partial u^1} \text{ and } \mathbf{a}_2 = \mathbf{b}_i \frac{\partial x^i}{\partial u^2};$$

so that $\frac{\partial x^i}{\partial u^\alpha} \equiv x^i_\alpha$; $(\alpha = 1, 2)$, are the contravariant components of the surface base vectors $\mathbf{a}_\alpha$ referred to the base system $\mathbf{b}_i$. Thus, the set of quantities

$$x^i_1 : \left(\frac{\partial x^1}{\partial u^1}, \frac{\partial x^2}{\partial u^1}, \frac{\partial x^3}{\partial u^1} \right) \text{ and } x^i_2 : \left(\frac{\partial x^1}{\partial u^2}, \frac{\partial x^2}{\partial u^2}, \frac{\partial x^3}{\partial u^2} \right)$$

transform in a contravariant manner relative to the transformation of space co-ordinates x^i. Now the three surface vectors

$$x^1_\alpha : \left(\frac{\partial x^1}{\partial u^1}, \frac{\partial x^1}{\partial u^2} \right); x^2_\alpha : \left(\frac{\partial x^2}{\partial u^1}, \frac{\partial x^2}{\partial u^2} \right) \text{ and } x^3_\alpha : \left(\frac{\partial x^3}{\partial u^1}, \frac{\partial x^3}{\partial u^2} \right)$$

transform according to the covariant law with respect to the transformation of Gaussian surface co-ordinates u^α. Indeed, consider a transformation $u^\alpha = u^\alpha\left(\bar{u}^1, \bar{u}^2\right)$; then Eq. (7.4) of S go over into $x^i = x^i\left(\bar{u}^1, \bar{u}^2\right)$ and

$$\frac{\partial x^i}{\partial u^\alpha} = \frac{\partial x^i}{\partial \bar{u}^\beta} \frac{\partial \bar{u}^\beta}{\partial u^\alpha}. \tag{7.11}$$

But $\frac{\partial x^i}{\partial \bar{u}^\beta} = \bar{x}^i_\beta$; and Eq. (7.11) yields the covariant law

$$x^i_\alpha = \frac{\partial \bar{u}^\beta}{\partial u^\alpha} \bar{x}^i_\beta; \quad i = 1, 2, 3.$$

Let ds be an element of arc joining a pair of points $P(u^1, u^2)$ and $Q(u^1 + du^1, u^2 + du^2)$ on S. The direction of the line element ds is given by the direction parameters $\frac{du^\alpha}{ds} = \lambda^\alpha$. The same direction can be specified by an observer in the enveloping space by means of three parameters $\frac{dx^i}{ds} = \lambda^i$; and from the Eq. (7.5), we find that

$$\lambda^i = \frac{dx^i}{ds} = \frac{\partial x^i}{\partial u^\alpha} \frac{du^\alpha}{ds}; \text{ i.e. } \lambda^i = x^i_\alpha \lambda^\alpha. \tag{7.12}$$

This formula tells us that any surface vector A^α can be viewed as a space vector with components A^i determined by

$$A^i = x^i_\alpha A^\alpha. \tag{7.13}$$

We shall refer to a vector A^i determined by Eq. (7.13) as a *tangent vector to the surface* S.

7.2 The Normal Line

An entity that provides a characterisation of the shape of the surface as it appears from the enveloping surface is the normal line to the surface.

Let A and B be a pair of surface vectors drawn at some point P on the surface S, such that the rotation A ; B is positive. The unit normal vector to the surface S is orthogonal to both A and B and is so oriented that (A,B,) from a right-handed system (Figure 7.1), i.e.

$$"_{ijk} A^i B^j \;{}^k = 1 \qquad (7.14)$$

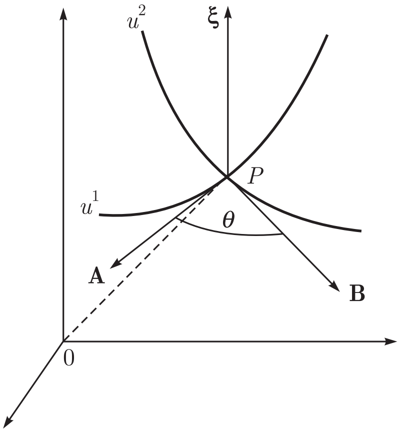

Figure 7.1: The normal lines to the surface.

where

$$= \frac{1}{jA \quad Bj}(A \quad B) = \frac{1}{AB\ \sin}(A \quad B)$$

where is the angle between A and B. We call the vector the unit normal vector to the surface S at P. Clearly, is a function of co-ordinates $u^1; u^2$, and as the point P $u^1; u^2$ is displaced to a new position Q $u^1 + du^1; u^2 + du^2$, the vector undergoes a change

$$d \;=\; \frac{@}{@u}du \;=\; \qquad du \;;\; \text{say;} \qquad (7.15)$$

whereas the position vector r is changed to the amount

$$dr = \frac{@}{@u}du \;=\; r \;\; du \;;\; \text{say.} \qquad (7.16)$$

The behaviour of the normal line as its foot displaced along the surface depends on the shape of the surface, and it occurred to Gauss to describe certain properties of surfaces with the aid of a quadratic form that depends in a fundamental way on the behaviour of the normal line. Let us form the scalar product

$$d\;\; dr = \frac{\partial}{\partial u}\;\;\frac{\partial}{\partial u} du\;\; du \;=\;\; r\;\; du\;\; du$$

$$=\;\; b\;\; du\;\; du\;\;;\;\; \text{say}; \tag{7.17}$$

if we define b as

$$b \;=\; \frac{1}{2}\;\;\frac{\partial}{\partial u}\;\;\frac{\partial}{\partial u} + \frac{\partial}{\partial u}\;\;\frac{\partial}{\partial u}$$

$$=\; \frac{1}{2}(\;\; r \;+\;\; r\;\;) \tag{7.18}$$

The left-hand side of Eq. (7.17), being the scalar product of two vectors, is an invariant. Thus, the second fundamental form is invariant with respect to any allowable co-ordinate transformation which preserves the sense of. Since the co-ordinate differentials du have a contravariant transformation behaviour and the second fundamental form is invariant, the coefficients b of this form are the components of a covariant tensor of second order with respect to co-ordinate transformations. From the quotient law, b is a covariant tensor of order (0,2) and from Eq. (7.18) it is symmetric with respect to and . The quadratic form

$$B\;\;\; b\;\; du\;\; du \tag{7.19}$$

introduced by Gauss, is called the second fundamental quadratic form of the surface. This differential form Eq. (7.19) plays an important part in the study of the geometry of the surface when they are viewed from the surrounding space.

Now, we shall write Eq. (7.14) in terms of the components x^i of the base vectors. The Eq. (7.13) tells us

$$A^i = x^i\, A\;\;;\;\; B^j = x^j\, B\;\;:$$

If be the angle between A^i and B^j, we can write

$$AB\,\sin\;\; = \varepsilon\;\;\; A\;\; B\;\;:$$

Let the contravariant components of the unit surface normal by ^{i} and the covariant components ${}_i$ is, then

$${}_i = \frac{\varepsilon_{ijk}\, x^j\, A\;\; x^k B}{\varepsilon\;\;\; A\;\; B}$$

or

$${}_i\,\varepsilon\;\;\;\; \varepsilon_{ijk}\, x^j\, x^k\;\;\; A\;\; B\;\; = 0:$$

This relation is valid for all surface vectors. Since, A and B are arbitrary, we have

$${}_i\epsilon \quad = \epsilon_{ijk}\, x^j\, x^k.$$

Multiplying this relation through by ϵ, and noting $\epsilon\ \epsilon = 2$, we get

$${}_i\epsilon\ \epsilon = \epsilon\ \epsilon_{ijk}\, x^j\, x^k$$

or

$${}_i\ 2 = \epsilon\ \epsilon_{ijk}\, x^j\, x^k$$

or

$${}_i = \frac{1}{2}\epsilon\ \epsilon_{ijk}\, x^j\, x^k.$$

It is clear from the structure of this formula that ${}_i$ is a space vector which does not depend on the choice of surface co-ordinates. This fact is also obvious from purely geometrical considerations.

EXAMPLE 7.2.1 Calculate the second fundamental form for the right helicoid given by $r = (u\cos v, u\sin v, cv)$.

Solution: Since $r = (u\cos v, u\sin v, cv)$, so

$$A = \frac{dr}{du} = (\cos v, \sin v, 0) \text{ and } B = \frac{dr}{dv} = (\quad u\sin v, u\cos v, c).$$

$$A \quad B = \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ \cos v & \sin v & 0 \\ u\sin v & u\cos v & c \end{vmatrix} = c\sin v\,\hat{i} \quad c\cos v\,\hat{j} + u\hat{k}.$$

Thus, the normal vector is given by

$$= \frac{1}{|A \quad B|}(A \quad B) = \frac{1}{\sqrt{c^2+u^2}}(c\sin v, \quad c\cos v, u).$$

The symmetric covariant tensor b are given by

$$b_{11} = \frac{1}{2}\left(\frac{\partial}{\partial u^1}\frac{\partial}{\partial u^1} + \frac{\partial}{\partial u^1}\frac{\partial}{\partial u^1}\right) = \frac{\partial}{\partial u^1}\frac{\partial}{\partial u^1} = \frac{\partial}{\partial u}\frac{\partial}{\partial u}$$

$$= \left(\frac{cu\sin v}{(c^2+u^2)^{\frac{3}{2}}}, \frac{cu\cos v}{(c^2+u^2)^{\frac{3}{2}}}, \frac{c^2}{(c^2+u^2)^{\frac{3}{2}}}\right) (\cos v, \sin v, 0) = 0$$

$$b_{22} = \frac{\partial}{\partial v}\cdot\frac{\partial}{\partial v} = \frac{1}{\sqrt{c^2+u^2}}(c\cos v; c\sin v; 0)\cdot(-u\sin v; u\cos u; c)$$

$$= \frac{1}{\sqrt{c^2+u^2}}(-uc\sin v\cos v + uc\sin v\cos v + 0) = 0$$

$$b_{12} = \frac{1}{2}\left(\frac{\partial}{\partial u^1}\cdot\frac{\partial}{\partial u^2} + \frac{\partial}{\partial u^1}\cdot\frac{\partial}{\partial u^2}\right) = \frac{1}{2}\left(\frac{\partial}{\partial u}\cdot\frac{\partial}{\partial v} + \frac{\partial}{\partial v}\cdot\frac{\partial}{\partial u}\right)$$

$$= \frac{1}{2}\left[\left(\frac{cu\sin v}{(u^2+c^2)^{\frac{3}{2}}}; \frac{cu\cos v}{(u^2+c^2)^{\frac{3}{2}}}; \frac{c^2}{(u^2+c^2)^{\frac{3}{2}}}\right)\cdot(-u\sin v; u\cos v; c)\right.$$

$$\left.+ \frac{1}{\sqrt{u^2+c^2}}(c\cos v; c\sin v; 0)\cdot(\cos v; \sin v; 0)\right]$$

$$= \frac{1}{2}\left[\frac{cu^2\sin^2 v + cu^2\cos^2 v + c^3}{(u^2+c^2)^{\frac{3}{2}}} + \frac{c\cos^2 v + c\sin^2 v}{\sqrt{u^2+c^2}}\right]$$

$$= \frac{1}{2}\left(\frac{c}{\sqrt{u^2+c^2}} + \frac{c}{\sqrt{u^2+c^2}}\right) = \frac{c}{(u^2+c^2)^{\frac{1}{2}}} = b_{21}.$$

Thus, the second fundamental quadratic form becomes

$$B \equiv b_{11}(du)^2 + 2b_{12}dudv + b_{22}(dv)^2 = 2b_{12}dudv \equiv \frac{2c}{(u^2+c^2)^{\frac{1}{2}}}dudv.$$

EXAMPLE 7.2.2 Calculate the second fundamental form for the paraboloid given by $r = (u; v; u^2 - v^2)$.

Solution: Since $r = (u; v; u^2 - v^2)$, so $A = (1; 0; 2u)$; $B = (0; 1; -2v)$ and

$$A \times B = -2u\hat{i} - 2v\hat{j} + \hat{k}.$$

Thus, the unit normal vector is given by

$$= \frac{1}{|A \times B|}(A \times B) = \frac{1}{\sqrt{4u^2+4v^2+1}}(-2u; -2v; 1).$$

The symmetric covariant tensors b are given by

$$b_{11} = \frac{\partial}{\partial u}\cdot\frac{\partial}{\partial u} = \frac{1}{(4u^2+4v^2+1)^{\frac{3}{2}}}\left(2(4v^2+1); -8uv; 4u\right)\cdot(1; 0; 2u)$$

$$= \frac{2(4u^2+1) + 8u^2}{(4u^2+4v^2+1)^{\frac{3}{2}}} = \frac{2}{\sqrt{4u^2+4v^2+1}}$$

$$b_{22} = \frac{\partial}{\partial v}\frac{\partial}{\partial v} = \frac{1}{(4u^2+4v^2+1)^{\frac{3}{2}}}\quad 8uv;\ \ 2(4u^2+1);4v\quad (0;1;\ \ 2v)$$

$$= \frac{2\ \ 4u^2+1\quad 8v^2}{(4u^2+4v^2+1)^{\frac{3}{2}}} = \quad \frac{2}{\sqrt{4u^2+4v^2+1}}$$

$$b_{12} = \frac{1}{2}\quad \frac{\partial}{\partial u}\frac{\partial}{\partial v} + \frac{\partial}{\partial v}\frac{\partial}{\partial u} = 0 = b_{21}:$$

Thus the second fundamental quadratic form becomes

$$B \quad b_{11}(du)^2 + b_{22}(dv)^2 + 2b_{12}dudv \quad \frac{2}{\sqrt{4u^2+4v^2+1}}\ (du)^2 \quad (dv)^2 \ :$$

EXAMPLE 7.2.3 Show a surface of revolution is regular and exhibit the unit surface normal.

Solution: The Gaussian form of a surface of revolution about z axis is

$$r(x^1;x^2) = \ f(x^1)\cos x^2; f(x^1)\sin x^2; g(x^1)\ ;\ f(x^1) > 0:$$

From this parametric representation, we get

$$A = \frac{\partial}{\partial x^1} = f'(x^1)\cos x^2; f'(x^1)\sin x^2; g'(x^1)\ ;$$

$$B = \frac{\partial}{\partial x^2} = \quad f(x^1)\sin x^2; f(x^1)\cos x^2; 0\ :$$

so

$$A \quad B = \quad fg'\cos x^2;\ \ fg'\sin x^2; ff'\ :$$

Therefore, the norm is $jA \quad Bj = f\sqrt{f'^2+g'^2}$: Now $f = f(x^1) \neq 0$; further, the generating curve is regular, which means that, with $t = x^1$, the tangent vector of the curve,

$$\frac{dx}{dt}; 0; \frac{dz}{dt} \quad = (f'; 0; g')$$

is non-null and $f'^2 + g'^2 \neq 0$: Therefore, $f\sqrt{f'^2+g'^2} \neq 0$ and the surface is regular. Thus, the unit surface normal vector is given by

$$= \frac{1}{jA \quad Bj}(A \quad B) = \frac{1}{f\sqrt{f'^2+g'^2}} \quad fg'\cos x^2;\ \ fg'\sin x^2; ff'$$

$$= \quad \frac{g'}{\sqrt{f'^2+g'^2}}\cos x^2;\ \ \frac{g'}{\sqrt{f'^2+g'^2}}\sin x^2; \frac{f'}{\sqrt{f'^2+g'^2}} \quad :$$

EXAMPLE 7.2.4 Prove that the tangent planes to a cone of revolution at the points of its generating straight lines coincide.

Solution: The parametric representation to a cone of revolution is

$$\vec{r} = (x^1, x^2, x^3) = (u^1 \cos u^2, u^1 \sin u^2, a u^1).$$

The first fundamental form of the surface, $a_{\alpha\beta}$ are given by

$$a_{11} = \left(\frac{\partial x^1}{\partial u^1}\right)^2 + \left(\frac{\partial x^2}{\partial u^1}\right)^2 + \left(\frac{\partial x^3}{\partial u^1}\right)^2 = 1 + a^2$$

$$a_{22} = \left(\frac{\partial x^1}{\partial u^2}\right)^2 + \left(\frac{\partial x^2}{\partial u^2}\right)^2 + \left(\frac{\partial x^3}{\partial u^2}\right)^2 = (u^1)^2$$

$$a_{12} = \frac{\partial x^1}{\partial u^1}\frac{\partial x^1}{\partial u^2} + \frac{\partial x^2}{\partial u^1}\frac{\partial x^2}{\partial u^2} + \frac{\partial x^3}{\partial u^1}\frac{\partial x^3}{\partial u^2} = 0.$$

Now $\vec{A}$ and $\vec{B}$ vectors are given by

$$\vec{A} = \frac{\partial \vec{r}}{\partial u^1} = (\cos u^2, \sin u^2, a)$$

$$\vec{B} = \frac{\partial \vec{r}}{\partial u^2} = (-u^1 \sin u^2, u^1 \cos u^2, 0).$$

Thus, the unit surface normal is given by

$$\vec{\xi} = \frac{1}{|\vec{A} \times \vec{B}|}(\vec{A} \times \vec{B}) = \left(-\frac{a \cos u^2}{\sqrt{1 + a^2}}, -\frac{a \sin u^2}{\sqrt{1 + a^2}}, \frac{1}{\sqrt{1 + a^2}}\right),$$

that is, the normal direction to the cone is independent of u^1. Since the curves $u^2 =$ constant are the generating straight lines of the cone, $\vec{\xi}$ is constant along any of those generators; hence the tangent planes to the cone at the points of any generating straight line coincide.

7.3 Tensor Derivative

In this section, we are introducing new tensors by differentiation of given vector fields. Consider a curve C lying on a given surface S. If A^i is a component of a space vector, defined along the curve C. If t is a parameter along C, the intrinsic derivative $\frac{\delta A^i}{\delta t}$ of A^i is given by

$$\frac{\delta A^i}{\delta t} = \frac{dA^i}{dt} + \left\{ {i \atop j\;k} \right\}_g A^j \frac{dx^k}{dt}. \tag{7.20}$$

The Christoffel symbols of second kind in Eq. (7.20) refer to the space co-ordinates x^i and are formed from the metric coefficients g_{ij}. This is indicated by the prefix g on the symbol. Again, if we consider a surface vector A, defined along the same curve C, we can form the intrinsic derivative with respect to the surface variables, namely,

$$\frac{A}{t} = \frac{dA}{dt} + {}_{a} \qquad A \frac{du}{dt}; \tag{7.21}$$

where the Christoffel symbols of second kind in Eq. (7.21) are formed from the metric coefficients a associated with the Gaussian surface co-ordinates u. The corresponding formulas for the intrinsic derivatives of the covariant vectors with components A_i and A are

$$\frac{A_i}{t} = \frac{dA_i}{dt} \quad {}_{g} \begin{matrix} k \\ i \ j \end{matrix} \quad A_k \frac{dx^i}{dt}$$

and

$$\frac{A}{t} = \frac{dA}{dt} \quad {}_{a} \qquad A \frac{du}{dt} \tag{7.22}$$

A geometric interpretation of these formulas is at hand where A^i and A are such that

$$\frac{A^i}{t} = 0 \text{ and } \frac{A}{t} = 0: \tag{7.23}$$

In the first case of Eq. (7.23) the vectors A^i form a parallel field with respect to C, considered as a space curve, where as in the later case of Eq. (7.23), the vectors A form a parallel field with respect to C regarded as a surface curve.

Consider a tensor field T^i, which is a contravariant vector with respect to a transformation of space co-ordinates x^i and a covariant vector relative to a transformation of the surface co-ordinates u. For example, a field of this type is a tensor x^i introduced in Eq. (7.8). If T^i is defined along a surface curve C, with t as parameter, then T^i is a function of t. We form an invariant

$$(t) = T^i A_i B \ ;$$

where a parallel vector field A_i along C, regarded as a space curve, and a parallel vector field B along C, viewed as a surface curve. The derivative of (t) with respect to the parameter t is given by

$$\frac{d\ (t)}{dt} = \frac{dT^i}{dt} A_i B + T^i \frac{dA_i}{dt} B + T^i A_i \frac{dB}{dt}; \tag{7.24}$$

which is obviously an invariant relative to both the space and surface co-ordinates. But since the fields $A_i(t)$ and B (t) are parallel,

$$\frac{dA_i}{dt} = {}_{g} \begin{matrix} k \\ i \ j \end{matrix} \quad A_k \frac{dx^i}{dt} \text{ and } \frac{dB}{dt} = {}_{a} \qquad B \frac{du}{dt}$$

therefore, Eq. (7.24) becomes

$$\frac{d}{dt} = \frac{dT^i}{dt} + \left\{ {i \atop j\;k} \right\} T^j \frac{dx^k}{dt} \quad a \qquad T^i \frac{du}{dt} \quad A_i B \ : \tag{7.25}$$

Since this is invariant for an arbitrary choice of parallel fields A_i and B , by quotient law, the expression in the bracket of Eq. (7.25) is a tensor of the same character as T^i with respect to the parameter t, and we write

$$\frac{T^i}{t} = \frac{dT^i}{dt} + \left\{ {i \atop j\;k} \right\} T^j \frac{dx^k}{dt} \quad a \qquad T^i \frac{du}{dt} : \tag{7.26}$$

If the field T^i is defined over the entire surface S, its components are functions of u s and we have a tensor field over the surface. Also, if C is any curve on the surface, we can argue that

$$\begin{aligned} \frac{T^i}{t} &= \frac{\partial T^i}{\partial u} + \left\{ {i \atop j\;k} \right\} T^j \frac{\partial x^k}{\partial u} \quad a \qquad T^i \ \frac{du}{dt} \\ &= \frac{\partial T^i}{\partial u} + \left\{ {i \atop j\;k} \right\} T^j x^k \quad a \qquad T^i \ \frac{du}{dt} \end{aligned}$$

is a tensor field. Since $\frac{du}{dt}$ is an arbitrary surface vector (for C is arbitrary) and since the right-hand side of this last equation is a tensor, we conclude that

$$T^i_{;} = \frac{\partial T^i}{\partial u} + \left\{ {i \atop j\;k} \right\} T^j x^k \quad a \qquad T^i \tag{7.27}$$

is a tensor which is singly contravariant in the space co-ordinates and doubly covariant in the surface co-ordinates. We shall call $T^i_{;}$, the tensor derivative of T^i with respect to u . Similarly, the tensor derivative of T^i with respect to u is given by

$$T^i_{;} = \frac{\partial T^i}{\partial u} + \left\{ {i \atop j\;k} \right\} T^j \ x^k \quad a \qquad T^i \quad a \qquad T^i \ : \tag{7.28}$$

If the surface co-ordinates at any point of the surface are geodesics and the space co-ordinates are orthogonal Cartesian, we see that at that point the tensor derivatives reduce to ordinary derivatives. This leads us to conclude that the operations of tensor differentiation of products and sums follow the usual rules and that the tensor derivatives of g_{ij} ; a ; " and their associated tensors vanish. Accordingly, they behave as constants in the tensor differentiation.

7.3.1 The Second Fundamental Form of a Surface

Before we introduce the new quadratic form, we must point out the relations, descriptions in the study of surfaces. Let us take the tensor derivative of the tensor x^i , representing the components of the surface base vectors , we find from Eq. (7.27) that

$$x^i_{;} = \frac{@^2 x^i}{@u \;\; @u} + {}_g \begin{matrix} i \\ j \;\; k \end{matrix} \; x^j x^k \quad {}_a \qquad x^i \tag{7.29}$$

from which it follows that $x^i_{;} = x^i_{;}$. Tensor differentiation of the relation a $= g_{mn} x^m x^n$ gives

$$g_{mn} x^m_{;} \; x^n + g_{mn} x^m x^n_{;} = 0: \tag{7.30}$$

Interchanging ; ; cyclically, we get

$$g_{mn} x^m_{;} \; x^n + g_{mn} x^m x^n_{;} = 0 \tag{7.31}$$

and

$$g_{mn} x^m_{;} \quad x^n + g_{mn} x^m x^n_{;} = 0: \tag{7.32}$$

Adding, Eqs. (7.31) and (7.32) and subtracting (7.30), we get

$$g_{mn} x^m_{;} \; x^n = 0; \;\; \text{as } x^m_{;} = x^m_{;} : \tag{7.33}$$

This orthogonality relation interpreted geometrically, states that $x^i_{;}$, from the point of view of the space co-ordinates is a space vector normal to the surface, and hence, it is codirectional with the normal vector ^{i}. Consequently, there must exist a set of functions such that

$$x^i_{;} = \quad {}^i; \tag{7.34}$$

where = b . The quantities b are the components of a symmetric surface tensor, and the differential quadratic form

$$B \quad b \;\; du \;\; du \tag{7.35}$$

is the desired second fundamental form. Equation (7.34) is known as Gauss's formula.

Deduction 7.3.1 Equivalence of the two definitions of the second fundamental quadratic form: Let us consider a surface S embedded in E^3 with the first fundamental form given by

$$ds^2 = a \quad du \;\; du$$

where (u^1, u^2) are the Gaussian co-ordinates for the surface S defined by

$$x^i = x^i(u^1, u^2), \quad i = 1, 2, 3,$$

x^i being the Cartesian co-ordinates of E^3. Let r be the position vector of any point P whose surface co-ordinates are (u^1, u^2) and space co-ordinates are (x^1, x^2, x^3). Now,

$$a = \frac{\partial}{\partial u}, \qquad = 1, 2,$$

are the surface base vectors and

$$b = \frac{\partial}{\partial x}, \qquad = 1, 2, 3,$$

are the space base vectors and they are usually related by

$$a = \frac{\partial}{\partial u} = \frac{\partial}{\partial x^i}\frac{\partial x^i}{\partial u} = x^i b_i. \qquad (7.36)$$

Let be the unit vector at P normal to the tangent plane, so, we get

$$= {}^i b_i \text{ and } = 1, \qquad a = 0. \qquad (7.37)$$

Since the vectors and a are orthogonal, so

$$\frac{\partial}{\partial u} = 0 \quad \text{and} \quad \frac{\partial}{\partial u} = 0.$$

Differentiating these two scalar products with respect to u and u, respectively, and then adding, we get

$$\frac{\partial}{\partial u}\ \frac{\partial}{\partial u} + \frac{\partial^2 r}{\partial u \partial u} + \frac{\partial}{\partial u}\ \frac{\partial}{\partial u} + \frac{\partial^2 r}{\partial u u} = 0$$

or

$$\frac{1}{2}\ \frac{\partial}{\partial u}\ \frac{\partial}{\partial u} + 2\ \frac{\partial^2 r}{\partial u \partial u} + \frac{\partial}{\partial u} . \frac{\partial}{\partial u} = 0$$

or

$$\frac{1}{2}\ \frac{\partial}{\partial u} . \frac{\partial}{\partial u} + \frac{\partial}{\partial u} . \frac{\partial}{\partial u} = .\frac{\partial^2 r}{\partial u \partial u}$$

or

$$b = .\frac{\partial^2 r}{\partial u \partial u} \quad \text{using Eq. (7.18).} \qquad (7.38)$$

Now, differentiating the relation $a = x^i b_i$ with respect to u; we get

$$\frac{\partial^2 r}{\partial u \partial u} = b_i \frac{\partial x^i}{\partial u} + \frac{\partial b_i}{\partial u} x^i = b_i \frac{\partial^2 x^i}{\partial u \partial u} + \frac{\partial b_i}{\partial x^j} x^i x^j$$

$$= b_h \quad \frac{\partial^2 x^h}{\partial u \partial u} + {}_g \genfrac{}{}{0pt}{}{h}{i\ j} \quad x^i x^j \quad ; \tag{7.39}$$

where in the last step, we made use of formula

$$\frac{\partial b_i}{\partial x^j} = {}_g \genfrac{}{}{0pt}{}{h}{i\ j} \quad b_h$$

for the derivatives of the base vector b_i: Using Eq. (7.29),

$$x^i_{;} = \frac{\partial^2 x^i}{\partial u \partial u} + {}_g \genfrac{}{}{0pt}{}{i}{j\ k} x^j x^k \quad {}_a \quad x^i$$

expression (7.39) reduces to

$$\frac{\partial^2 r}{\partial u \partial u} = x^h_{;} + {}_a \quad x^h \; b_h$$

$$= b \quad {}^h b_h + {}_a \quad x^h b_h; \quad \text{as } x^i_{;} = b \quad {}^i$$

$$= b \quad + {}_a \quad x^h a \;; \quad \text{using Eq. (7:37):} \tag{7.40}$$

Multiplying Eq. (7.39) scalarly by , we get

$$\frac{\partial^2 r}{\partial u \partial u} = b \quad + {}_a \quad x^h a$$

$$= b + 0 = b \;; \quad \text{as } a = 0 \tag{7.41}$$

This establishes the equivalence of the two definitions Eqs. [(7.38) and (7.41)] of the second fundamental quadratic form.

Deduction 7.3.2 The integrability conditions: In order to get insight into the significance of the tensor b, let us examine the Gauss's formula Eq. (7.34), as

$$x^i_{;} = b \quad {}^i;$$

where

$$x^i_{;} = \frac{\partial^2 x^i}{\partial u \partial u} + {}_g \genfrac{}{}{0pt}{}{i}{j\ k} x^j x^k \quad {}_a \quad x^i$$

$${}_i = \frac{1}{2}\varepsilon \quad \varepsilon_{ijk} x^j x^k \text{ and } x^i = \frac{\partial x^i}{\partial u}:$$

If we insert these expressions in Eq. (7.34), we obtain a set of second order partial differential equations, in which the dependent variables x^i are the functions of the surface co-ordinates u . The coefficients in the differential equations are functions of the metric coefficients g_{ij} of the manifold in which the surface S, defined by

$$x^i = x^i(u^1; u^2); \;\; i = 1; 2 \tag{7.42}$$

are immersed; they are also functions of

$$a \quad = g_{ij} \frac{@x^i}{@u} \frac{@x^j}{@u}$$

and b . If Eq. (7.42) is given, we can compute a and b , insert the approximate expressions in Eq. (7.34) and, of course, Eq. (7.34) will be satisfied identically.

Conversely, if the functions a and b are given, Eq. (7.34) will become the equation of conditions and in general they will have no solutions yielding Eq. (7.42) of the surface S. In order that the terms a and b are related to some surface, it is necessary that x^i satisfy the conditions of integrability

$$\frac{@^2 x^i}{@u @u} = \frac{@^2 x^i}{@u \; @u}; \tag{7.43}$$

whenever the function x_i are continuously differentiable of degree 2. Equation (7.43) is equivalent to

$$x^i_{;} \qquad x^i_{;} \quad = R \qquad x^i \;; \tag{7.44}$$

where R is the Riemann second kind curvature tensor, of type (1;3); formed with the coefficients a of the first fundamental quadratic form. Equation (7.44) involve third order partial derivatives of the co-ordinates x^i and we shall assume from now on that the functions entering in Eq. (7.42) are continuously thrice differentiable.

We shall see that the conditions of integrability Eq. (7.44) impose certain restrictions on the possible choice of functions b and a : These restrictive conditions are known as the equations of Gauss and Codazzi.

7.4 Structure Formulas for Surfaces

Two fundamental sets of relationships involve the parts of the moving triad $(A; B; \;)$ of a surface at any point P. Any vector bound at P can be represented as a linear combinations of those vectors. If the derive of those vectors exist, then the corresponding combinations are called the formulae of Frenet.

7.4.1 Equations of Weingarten

Weingarten's formula is an expression for the derivatives of the unit normal vector ^{i} to the surface S. Since ^{i}, the contravariant components of the normal vector is a unit vector, so,

$$g_{ij}\ ^i\ ^j = 1: \tag{7.45}$$

Here we have to show that the partial derivatives of this vector with respect to the co-ordinates u^1 and u^2 can be represented as a linear combination of vectors. Using the tensor derivative formula

$$^i_{;} = \frac{@^i}{@u} + {}_g \begin{matrix} i \\ j\ k \end{matrix}\ ^i x^k$$

we have from Eq. (7.45)

$$g_{ij}\ ^i_{;}\ ^j + g_{ij}\ ^i\ ^j_{;} = 0 \quad) \quad g_{ij}\ ^i\ ^j_{;} = 0: \tag{7.46}$$

Geometrically, Eq. (7.46) shows that $^j_{;}$; from the point of view of space co-ordinates, considered as a space vector, is orthogonal to the unit normal^i, and hence, it lies in the tangent plane to the surface. Accordingly, it can be expressed as a linear form in the base vectorsx^i , as

$$^i_{;} = c\ x^i\ ; \tag{7.47}$$

where c s are to be determined. Since ^{i} is normal to the surface, the tensor derivative of the orthogonality relation $g_{ij}\, x^i\ ^j = 0$ gives

$$g_{ij}\, x^i_{;}\ ^j + g_{ij}\, x^i\ ^j_{;} = 0$$

or

$$g_{ij}\, b\ ^i\ ^j + g_{ij}\, x^i\ ^j_{;} = 0;\ \text{as } x^i_{;} = b\ ^i$$

or

$$b + g_{ij}\, x^i\, c\ x^j = 0$$

or

$$b + a\ c = 0;\ \text{as } a = g_{ij}\, x^i\, x^j$$

or

$$a\ b + a\ a\ c = 0$$

or

$$c = a\ b\ ;\ \text{as } a\ a = : \tag{7.48}$$

Consequently, Eq. (7.47) reduces to

$$ {}^{i}_{;} = \quad a \quad b \quad x^{i} : \qquad (7.49) $$

These equations are known as Weingarten's formula, which can be used to derive the equations of Gauss and Codazzi. We have noticed that the formulae of Weingarten are the analogue of the formulae of Frenet. If we write

$$ C \quad = g_{ij} \ {}^{i}_{;} \ {}^{j}_{;} \ ; \qquad (7.50) $$

we see that C is symmetric tensor of order $(0;2)$ and we call the quadratic form

$$ C \quad C \quad du \ du \qquad (7.51) $$

the third fundamental form of the surface. Using Weingarten formula Eq. (7.49), we get

$$ \begin{aligned} C \quad &= g_{ij} \ {}^{i}_{;} \ {}^{j}_{;} = g_{ij} \quad a \quad b \quad x^{i} \quad a \quad b \quad x^{j} \\ &= a \quad a \quad b \quad a \quad b \quad = a \quad b \quad b \quad : \end{aligned} $$

This is a relation between three fundamental forms on a surface. This relation also shows that the third fundamental form is not actually a fundamental form, because this can be obtained from first and second fundamental forms.

Deduction 7.4.1 From the Weingarten formula Eq. (7.49), we get

$$ {}^{i}_{;} = \quad a \quad b \quad x^{i} = 0; \quad \text{if } b \quad = 0 : $$

Hence, ^{i} is a constant vector. Hence, the surface is a plane. Therefore, if $b \quad = 0$; then the normal vector to the surface is constant which means that the surface is a plane.

7.4.2 Equations of Gauss and Codazzi

The question arises whether, if functions $a \quad (u^1;u^2)$ and $b \quad (u^1;u^2)$ are given, there always exists a surface such that the given functions are the coefficients of the corresponding fundamental forms. In the case of functions of class 3 the answer to this problem is negative unless certain integrability conditions are satisfied which we will now derive. Let us consider the tensor derivative of Eq. (7.34), we get

$$ \begin{aligned} x^{i}_{;} \quad &= b_{;} \quad {}^{i} + b \quad {}^{i}_{;} \\ &= b_{;} \quad {}^{i} \quad b \quad a \quad b \quad x^{i} \ ; \ \text{using Eq. } (7{:}49) \end{aligned} $$

or

$$x^i_{;} \quad x^i_{;} = (b_{;} \quad b_{;})\ {}^i + b\ a\ b\ x^i \quad b\ a\ b\ x^i$$

$$= (b_{;} \quad b_{;})\ {}^i + a\ (b\ b \quad b\ b)x^i$$

or

$$(b_{;} \quad b_{;})\ {}^i \quad a\ (b\ b \quad b\ b)x^i \quad R \quad x^i\ ; \qquad (7.52)$$

if we write

$$x^i_{;} \quad x^i_{;} = R \quad x^i\ ;$$

where R is the Riemann curvature of the surface. Multiplying Eq. (7.52) by $_i$ and using the fact that $^i\ {}_i = 1$ and $x^i\ {}_i = 0$, we get

$$b_{;} \quad b_{;} = 0 : \qquad (7.53)$$

These equations are called the *Codazzi equations* of the surface, which will constitute all integrability conditions of the formulae of Weingarten. Again, multiplying Eq. (7.52) by $g_{ij}\ x^j$, we get

$$a\ R \quad = a\ b\ a\ b \quad a\ b\ a\ b$$

or

$$R \quad = b\ b \quad b\ b : \qquad (7.54)$$

These equations are called the *Gauss equations* of surface. Since $;$ assume values $1; 2$ and $b = b$; we see that there are two independent equations of Codazzi given by

$$b_{;} \quad b_{;} = 0; \quad \ne \qquad (7.55)$$

or

$$\frac{\partial b}{\partial u} \quad \frac{\partial b}{\partial u} \quad b \qquad + b \qquad = 0; \quad \ne$$

and only one independent equation of Gauss is

$$b = b_{11}b_{22} \quad b_{12}^2 = R_{1212}; \quad \text{where, } b = |b_{ij}| : \qquad (7.56)$$

This equation relates the coefficients b and a in the two fundamental quadratic forms. Thus, it follows from the theory of partial differential equation that if the tensors a and b are the fundamental tensors of the surface $S : x^i = x^i\ (u^1; u^2)$, Eqs. (7.53) and (7.56) are satisfied. Conversely, if the two sets of functions a and b satisfying Eqs. (7.53) and (7.56) are prescribed, and if $a\ du\ du$ is a positive definite form, then the surface S is determined (locally) to within a rigid body space.

Deduction 7.4.2 It is clear that, when = , we get identities from $b_{;}$ $b_{;}$ = 0 of the form 0 = 0: Again the equation with = p and = q is equivalent to the equation = q and = p. Hence, there are only two independent equations, as

$$b_{11;2} \quad b_{12;1} = 0 \text{ and } b_{21;2} \quad b_{22;1} = 0:$$

These two equations are equivalent to

$$b_{;} \quad b_{;} = 0; \quad 6= :$$

This shows that, $b_{;}$ $b_{;}$ = 0 is equivalent to $b_{;}$ $b_{;}$ = 0; 6= :

Deduction 7.4.3 From Codazzi equations, we get $b_{;}$ $b_{;}$ = 0; 6= : Let b = a ; then

$$b_{;} = {}_{;}\, a \quad \text{and } b_{;} = {}_{;}\, a$$

or

$${}_{;}\, a \qquad {}_{;}\, a = 0$$

or

$${}_{;}\, a \;\; a \qquad {}_{;}\, a \;\; a = 0$$

or

$$2 \; {}_{;} \qquad {}_{;} = 0 \;) \qquad {}_{;} = 0;$$

which implies that is a constant. Thus, if b = a ; where is a scalar, then is a constant.

7.4.3 Curvatures of a Surface

From Eq. (6.52), the total curvature of a surface is given by

$$= \frac{R_{1212}}{a}; \; a = a_{11}a_{22} \quad a_{12}^2 = ja_{ij}j: \tag{7.57}$$

Using the Gauss equation [Eq. (7.56)], the total curvature is given by

$$= \frac{b_{11}b_{22} \quad b_{12}^2}{a_{11}a_{22} \quad a_{12}^2} = \frac{b}{a}: \tag{7.58}$$

Thus, the Gaussian curvature is equal to the quotient of the determinants of the second and first fundamental quadratic forms. A surface S is said to be developable if = 0 identically at each point of the surface. A developable surface is a surface which can be developed into a plane. We now define another important invariant H , given by the formula

$$H \quad \frac{1}{2} a \quad b \; ; \tag{7.59}$$

which is called the mean curvature of the surface. The invariants ; H are connected with the ordinary curvatures of certain curves formed by taking normal sections of the surface (Section 8.2).

7.4.4 Minimal Surfaces

If at every point of a surface of class 2, the total curvature

$$H \quad \frac{1}{2} a \quad b \quad = 0; \tag{7.60}$$

then the surfaceS is called a minimal surface. Minimal surfaces derive their name from the fact that they are surfaces of minimum area satisfying given boundary conditions. They are illustrated by the shapes of thin soap films in equilibrium, with the air pressure the same on both sides. If this property of least area be taken as defining minimal surfaces, the use of the calculus of variations leads to the vanishing of the first curvature as an equivalent property. Later, we have to prove the following very important properties:

(i) Minimal surfaces cannot have elliptical points.
(ii) The lines of curvature on a minimal surfaces form an isometric system.
(iii) The asymptotic lines on a minimal surface form an orthogonal system.
(iv) Asymptotic lines on a minimal surface form on orthogonal system.

Theorem 7.4.1 (Theorema Egregium of Gauss): The Gaussian curvature is an intrinsic property of a surface, depending only on the first fundamental form and their derivatives but is independent of the second fundamental form.

Proof: The total curvature is given by

$$= \frac{b}{a} = \frac{b_{11}b_{22} \quad b_{12}^2}{a_{11}a_{22} \quad a_{12}^2}:$$

Thus, the total curvature, as obtained, depends upon the first fundamental form as well as the second fundamental form. But we have the equations of Gauss,

$$R \quad = b \quad b \quad b \quad b \quad ;$$

from which it follows that $R_{1212} = b_{11}b_{22} \quad b_{12}^2$: Therefore, $= \frac{R_{1212}}{a}$. But the Riemann curvature tensor depends only upon the first fundamental form a together with their derivatives. Hence, the theorem.

The practical question of whether inhabitants of a fog-enshrouded planet could, solely by measuring distances on the surface of the planet, determine its curvature, is answered in the affirmative by this theorem. Now,

(i) If two surfaces are locally isometric, then this theorem shows that, the Gaussian curvatures are identical.

(ii) Minding's theorem states that, if two surfaces are the same constant Gaussian curvature, then they are locally isometric.

(iii) In the case of constant Gaussian curvature, Beltrami's theorem tells us that there is a parameterisation for S for which the first fundamental form takes the form

Metric	Gaussian curvature
$ds^2 = a^2(dx^1)^2 + a^2 \sinh^2 x^1 (dx^2)^2$	$= \quad \frac{1}{a^2} < 0$
$ds^2 = (dx^1)^2 + (dx^2)^2$	$= 0$
$ds^2 = a^2(dx^1)^2 + a^2 \sin^2 x^1 (dx^2)^2$	$= \frac{1}{a^2} > 0$

(iv) From Eq. (7.54) we see that in case of surfaces of vanishing curvature (for example planes, cones, cylinders) all components of the curvature tensor are zero.

EXAMPLE 7.4.1 Show that

$$b \quad = g_{ij}\, x^{i}_{;} \quad {}^{j} = \frac{1}{2}" \quad "_{ijk}\, x^{i}_{;} \quad x^{j} x^{k}:$$

Solution: Equation (7.13) tells us $A^i = x^i \ A$; $B^j = x^j \ B$: If be the angle between the vectors A^i and B^j, we can write

$$AB \sin \quad = " \quad A \ B \ :$$

If is the unit surface normal, then

$${}_i = \frac{"_{ijk}\, x^j \ A \ \ x^k B}{" \ \ A \ B} \quad) \qquad {}_i" \qquad "_{ijk}\, x^j x^k \ \ A \ B \ = 0:$$

Since A and B are arbitrary, we have, ${}_i" \quad = "_{ijk}\, x^j x^k$ and therefore

$${}_i" \quad " \quad = " \quad "_{ijk}\, x^j x^k$$

or

$${}_i \ \ 2 = " \quad "_{ijk}\, x^j x^k \quad) \qquad {}_i = \frac{1}{2}" \quad "_{ijk}\, x^j x^k:$$

Using Eq. (7.34) as $x^i_{;} \quad = b \quad {}^i$, the Gauss's formula, we get

$$b \quad = x^i_{;} \quad {}_i = x^i_{;} \ g_{ij} \quad {}^j = g_{ij}\, x^i_{;} \quad {}^j:$$

Also, using the Eq. (7.34), we get

$$b \quad = g_{ij}\, x^i_{;} \quad {}_i = \frac{1}{2}"_{ijk}\, x^j x^k x^i_{;} \quad :$$

Combining these results, we get

$$b \quad = g_{ij}\, x^i_{;} \quad {}^j = \frac{1}{2}" \quad "_{ijk}\, x^i_{;} \quad x^j\, x^k:$$

EXAMPLE 7.4.2 Prove that C 2HB + A = 0; where the notations have their usual meaning.

Solution: The notations have the following expressions:

C : The third fundamental form of the surface c du du :
B : The second fundamental form of the surface b du du :
A : The first fundamental form of the surface a du du :
: The Gaussian curvature = $\frac{R_{1212}}{a}$:
H : The mean curvature of the surface = $\frac{1}{2}a \quad b$:
Let us take the Gauss's formula,

$$R \qquad = b \;\; b \qquad b \;\; b \quad \text{and}\; R \qquad = " \quad " \quad ;$$

we get

$$" \quad " \quad = b \;\; b \qquad b \;\; b \;\; :$$

Multiplying both sides by a , we get

$$" \quad " \quad a \quad = a \;\; b \;\; b \qquad a \;\; b \;\; b$$

or

$$a \qquad = a \;\; b \;\; b \qquad a \;\; b \;\; b \;\; ; \text{as}\, a \quad = \quad a \;\; " \quad "$$

or

$$a \qquad = 2Hb \qquad c \quad ; \;\text{as}\, c \quad = a \;\; b \;\; b$$

or

$$c \qquad 2Hb \quad + \; a \quad = 0$$

or

$$c \;\; du \;\; du \qquad 2Hb \;\; du \;\; du \; + \; a \quad du \;\; du \;\; = 0$$

or

$$C \quad 2HB + A = 0:$$

EXAMPLE 7.4.3 Show that

$$\text{(i)}\;\; a \quad c \quad = 4H^2 \quad 2: \qquad \text{(ii)}\;\; g_{ij} \;\; {}^i \; {}^j_{;} \quad = \quad c \;\; :$$

Solution: (i) Using the relation $a \quad = 2Hb \quad c$, we get

$$a \quad = 2Hb \quad c$$

or

$$= 2Hb \quad c \quad :$$

Also, using the relation $a \quad = 2Hb \quad c$, we get

$$c \; a \quad 2Hb \; a \; + a \; a \; = 0$$

or

$$a \; c \quad 2H \; 2H + \quad 2 = 0; \text{ as } a \; a \; = 2$$

or

$$c \; a \; = 4H^2 \quad 2 :$$

(ii) We know that, $g_{ij} \; {}^{i} \; {}^{j}_{;} = 0$: Taking the tensor derivative of both sides, we get

$$g_{ij} \; {}^{i}_{;} \; {}^{j}_{;} + g_{ij} \; {}^{i} \; {}^{j}_{;} \; = 0$$

or

$$g_{ij} \; {}^{i} \; {}^{j}_{;} \; = \; g_{ij} \; {}^{i}_{;} \; {}^{j}_{;} \; = \; c \; :$$

EXAMPLE 7.4.4 If b is the cofactor of b in $jb \; j$, divided by $jb \; j$, show that

$$x^i = \; a \; b \; {}^{i}_{;} :$$

Solution: From the definition,

$$b \; = \frac{\text{cofactor of } b \text{ in } jb \; j}{jb \; j}$$

with the property $b \; b \; = \;$: Hence,

$$a \; b \; x^{i}_{;} = \; a \; b \quad a \; b \; x^i \; = a \; a \; b \; b \; x^i$$

$$= a \; a \quad x^i = a \; a \; x^i = \quad x^i = x^i :$$

EXAMPLE 7.4.5 Find the Gaussian curvature at the point $(2; 0; 1)$ of the surface $(x^1)^2 \quad (x^2)^2 = 4x^3$; where $x^1; x^2; x^3$ are rectangular Cartesian co-ordinates.

Solution: Let the parametric representation of the surface is given by

$$x^1 = u^1 + u^2;\ x^2 = u^1 - u^2;\ x^3 = u^1u^2$$

i.e.

$$\mathbf{r} = (x^1; x^2; x^3) = (u^1 + u^2; u^1 - u^2; u^1u^2).$$

Therefore, at $(2; 0; 1)$, we have

$$u^1 = \frac{1}{2}(x^1 + x^2) = 1 \text{ and } u^2 = \frac{1}{2}(x^1 - x^2) = 1.$$

For the first fundamental form of the surface [Eq. (7.2)] the tensors $a_{\alpha\beta}$ are given by

$$a_{11} = \frac{\partial x^i}{\partial u^1}\frac{\partial x^i}{\partial u^1} = \left(\frac{\partial x^1}{\partial u^1}\right)^2 + \left(\frac{\partial x^2}{\partial u^1}\right)^2 + \left(\frac{\partial x^3}{\partial u^1}\right)^2$$
$$= 1^2 + 1^2 + (u^2)^2 = 2 + (u^2)^2.$$

$$a_{22} = \frac{\partial x^i}{\partial u^2}\frac{\partial x^i}{\partial u^2} = \left(\frac{\partial x^1}{\partial u^2}\right)^2 + \left(\frac{\partial x^2}{\partial u^2}\right)^2 + \left(\frac{\partial x^3}{\partial u^2}\right)^2$$
$$= 1^2 + (-1)^2 + (u^1)^2 = 2 + (u^1)^2.$$

$$a_{12} = \frac{\partial x^i}{\partial u^1}\frac{\partial x^i}{\partial u^2} = \frac{\partial x^1}{\partial u^1}\frac{\partial x^1}{\partial u^2} + \frac{\partial x^2}{\partial u^1}\frac{\partial x^2}{\partial u^2} + \frac{\partial x^3}{\partial u^1}\frac{\partial x^3}{\partial u^2}$$
$$= 1 \cdot 1 + 1 \cdot (-1) + u^2 \cdot u^1 = u^1u^2 = a_{21}.$$

Therefore,

$$a = \begin{vmatrix} 2 + (u^2)^2 & 0 \\ 0 & 2 + (u^2)^2 \end{vmatrix} = 2\{(u^1)^2 + (u^2)^2 + 2\}.$$

For the given parametric surface, $\mathbf{A} = (1; 1; u^2)$; $\mathbf{B} = (1; -1; u^1)$ and so

$$\mathbf{A} \times \mathbf{B} = (u^1 + u^2)\hat{\imath} + (u^2 - u^1)\hat{\jmath} - 2\hat{k}.$$

Therefore, the normal vector ξ is given by

$$\xi = \frac{1}{|\mathbf{A} \times \mathbf{B}|}(\mathbf{A} \times \mathbf{B}) = \frac{1}{\sqrt{2\{(u^1)^2 + (u^2)^2 + 2\}}}(u^1 + u^2; u^2 - u^1; -2).$$

Thus the covariant tensors $b_{\alpha\beta}$ for the second fundamental form are given by

$$b_{11} = \frac{1}{\sqrt{2}K^{\frac{3}{2}}}\ (u^2)^2 + 2 - u^1u^2;\ -\{(u^2)^2 + 2 + u^1u^2\};\ 2u^1 \cdot (1; 1; u^2);$$

[where $K = (u^1)^2 + (u^2)^2 + 1$];

$$= \frac{1}{\sqrt{2}K^{\frac{3}{2}}} \quad (u^2)^2 + 2 \quad u^1u^2 \quad (u^2)^2 \quad 2 \quad u^1u^2 + 2u^1u^2 \quad = 0.$$

$$b_{22} = \frac{1}{\sqrt{2}K^{\frac{3}{2}}} \quad (u^1)^2 + 2 \quad u^1u^2, \{(u^1)^2 + 2 + u^1u^2\}, 2u^2 \quad (1, \quad 1, u^1)$$

$$= \frac{1}{\sqrt{2}K^{\frac{3}{2}}} \quad (u^1)^2 + 2 \quad u^1u^2 \quad (u^1)^2 \quad 2 \quad u^1u^2 + 2u^1u^2 \quad = 0.$$

$$b_{12} = \frac{1}{2} \quad \frac{\partial}{\partial u^1} \quad \frac{\partial}{\partial u^2} + \frac{\partial}{\partial u^2} \quad \frac{\partial}{\partial u^1}$$

$$= \frac{2}{2\sqrt{2}} \left[\frac{(u^2)^2 + 1 + (u^1)^2}{\{(u^1)^2 + (u^2)^2 + 1\}^{\frac{3}{2}}} + \frac{(u^1)^2 + 1 + (u^2)^2}{\{(u^1)^2 + (u^2)^2 + 1\}^{\frac{3}{2}}} \right]$$

$$= \frac{2}{\sqrt{2\{(u^1)^2 + (u^2)^2 + 2\}}}.$$

Therefore,

$$b = \begin{vmatrix} 0 & \frac{2}{\sqrt{2\{(u^1)^2+(u^2)^2+2\}}} \\ \frac{2}{\sqrt{2\{(u^1)^2+(u^2)^2+2\}}} & 0 \end{vmatrix} = \frac{2}{(u^1)^2 + (u^2)^2 + 2}.$$

Thus, the Gaussian curvature is given by

$$= \frac{b}{a} = \frac{1}{[(u^1)^2 + (u^2)^2 + 2]^2} = \frac{1}{16}, \text{ at } (u^1, u^2) = (1, 1).$$

EXAMPLE 7.4.6 Find the Gaussian and mean curvature of the surface $x^1x^2 = x^3$.

Solution: Let the parametric representation of the surface (hyperbolic paraboloid) is given by, $x^1 = u^1, x^2 = u^2$ then, $x^3 = u^1u^2$. i.e.

$$r = (x^1, x^2, x^3) = (u^1, u^2, u^1u^2).$$

The symmetric covariant tensors a of the first fundamental form are given by

$$a_{11} = \left(\frac{\partial x^1}{\partial u^1}\right)^2 + \left(\frac{\partial x^2}{\partial u^1}\right)^2 + \left(\frac{\partial x^3}{\partial u^1}\right)^2 = 1 + (u^2)^2.$$

$$a_{22} = \left(\frac{\partial x^1}{\partial u^2}\right)^2 + \left(\frac{\partial x^2}{\partial u^2}\right)^2 + \left(\frac{\partial x^3}{\partial u^2}\right)^2 = 1 + (u^1)^2.$$

$$a_{12} = \frac{\partial x^1}{\partial u^1}\frac{\partial x^1}{\partial u^2} + \frac{\partial x^2}{\partial u^1}\frac{\partial x^2}{\partial u^2} + \frac{\partial x^3}{\partial u^1}\frac{\partial x^3}{\partial u^2} = u^1u^2 = a_{21}.$$

Therefore,

$$a = \begin{vmatrix} 1+(u^2)^2 & u^1u^2 \\ u^1u^2 & 1+(u^1)^2 \end{vmatrix} = 1+(u^1)^2+(u^2)^2.$$

The reciprocal tensors are given by

$$a^{11} = \frac{1+(u^1)^2}{1+(u^1)^2+(u^2)^2};\ a^{22} = \frac{1+(u^2)^2}{1+(u^1)^2+(u^2)^2};\ a^{12} = \frac{-u^1u^2}{1+(u^1)^2+(u^2)^2} = a^{21}.$$

Since $r = (x^1, x^2, x^3) = (u^1, u^2, u^1u^2)$, we have, $A = (1, 0, u^2)$; $B = (0, 1, u^1)$ and so

$$A \times B = -u^2\hat{i} - u^1\hat{j} + \hat{k}.$$

Therefore, the normal vector ν is given by

$$\nu = \frac{1}{|A \times B|}(A \times B) = \frac{1}{\sqrt{(u^1)^2+(u^2)^2+1}}(-u^2, -u^1, 1).$$

The symmetric covariant tensors $b_{\alpha\beta}$ are given by

$$b_{11} = -\frac{\partial \nu}{\partial u^1}\cdot\frac{\partial r}{\partial u^1} = -\frac{1}{\{(u^1)^2+(u^2)^2+1\}^{\frac{3}{2}}}\left(u^1u^2, -\{(u^2)^2+1\}, -u^1\right)\cdot(1, 0, u^2)$$

$$= -\frac{1}{\{(u^1)^2+(u^2)^2+1\}^{\frac{3}{2}}}\left(u^1u^2 - u^1u^2\right) = 0.$$

$$b_{22} = -\frac{\partial \nu}{\partial u^2}\cdot\frac{\partial r}{\partial u^2} = -\frac{1}{\{(u^1)^2+(u^2)^2+1\}^{\frac{3}{2}}}\left(-\{(u^1)^2+1\}, u^1u^2, -u^2\right)\cdot(0, 1, u^1)$$

$$= -\frac{1}{\{(u^1)^2+(u^2)^2+1\}^{\frac{3}{2}}}\left(u^1u^2 - u^1u^2\right) = 0.$$

$$b_{12} = -\frac{1}{2}\left(\frac{\partial \nu}{\partial u^1}\cdot\frac{\partial r}{\partial u^2} + \frac{\partial \nu}{\partial u^2}\cdot\frac{\partial r}{\partial u^1}\right)$$

$$= -\frac{1}{2}\left[\frac{-(u^2)^2-1-(u^1)^2}{\{(u^1)^2+(u^2)^2+1\}^{\frac{3}{2}}} - \frac{(u^1)^2+1+(u^2)^2}{\{(u^1)^2+(u^2)^2+1\}^{\frac{3}{2}}}\right]$$

$$= \frac{1}{\sqrt{(u^1)^2+(u^2)^2+1}} = b_{21}.$$

Therefore,

$$b = \begin{vmatrix} 0 & \frac{1}{\sqrt{(u^1)^2+(u^2)^2+1}} \\ \frac{1}{\sqrt{(u^1)^2+(u^2)^2+1}} & 0 \end{vmatrix} = -\frac{1}{(u^1)^2+(u^2)^2+1}.$$

The Gaussian curvature K is given by the formula,

$$K = \frac{b}{a} = -\frac{1}{\left[(u^1)^2 + (u^2)^2 + 1\right]^2} < 0.$$

Therefore, the mean curvature H is given by

$$H = \frac{1}{2}a^{\alpha\beta}b_{\alpha\beta} = \frac{1}{2}\left[a^{11}b_{11} + a^{22}b_{22} + 2a^{12}b_{12}\right]$$

$$= \frac{1}{2}\left[0 + 0 + \frac{u^1u^2}{1+(u^1)^2+(u^2)^2}\;\frac{1}{\sqrt{(u^1)^2+(u^2)^2+1}}\right]$$

$$= \frac{u^1u^2}{\left\{1+(u^1)^2+(u^2)^2\right\}^{3/2}}.$$

EXAMPLE 7.4.7 Find the mean curvature of the surface $\vec{r} = (u, v, u^2 - v^2)$.

Solution: For the first fundamental form of the surface [Eq. (7.2)], $a_{\alpha\beta}$ are given by

$$a_{11} = \left(\frac{\partial x^1}{\partial u}\right)^2 + \left(\frac{\partial x^2}{\partial u}\right)^2 + \left(\frac{\partial x^3}{\partial u}\right)^2 = 1 + 4u^2.$$

$$a_{22} = \left(\frac{\partial x^1}{\partial v}\right)^2 + \left(\frac{\partial x^2}{\partial v}\right)^2 + \left(\frac{\partial x^3}{\partial v}\right)^2 = 1 + 4v^2.$$

$$a_{12} = \frac{\partial x^1}{\partial u}\frac{\partial x^1}{\partial v} + \frac{\partial x^2}{\partial u}\frac{\partial x^2}{\partial v} + \frac{\partial x^3}{\partial u}\frac{\partial x^3}{\partial v} = -4uv = a_{21}.$$

Therefore,

$$a = (1+4u^2)(1+4v^2) - 16u^2v^2 = 1 + 4u^2 + 4v^2.$$

The reciprocal tensors are given by

$$a^{11} = \frac{1+4v^2}{1+4u^2+4v^2};\; a^{22} = \frac{1+4u^2}{1+4u^2+4v^2};\; a^{12} = \frac{4uv}{1+4u^2+4v^2} = a^{21}.$$

For the second fundamental form, the symmetric covariant tensors are given by

$$b_{11} = \frac{2}{\sqrt{1+4u^2+4v^2}};\; b_{22} = -\frac{2}{\sqrt{1+4u^2+4v^2}} \text{ and } b_{12} = 0 = b_{21}.$$

Therefore, the mean curvature H is given by

$$H = \frac{1}{2}a^{\alpha\beta}b_{\alpha\beta} = \frac{1}{2}\left[a^{11}b_{11} + a^{22}b_{22} + 2a^{12}b_{12}\right]$$

$$= \frac{1}{2}\left[\frac{2(1+4v^2)}{(1+4u^2+4v^2)^{3/2}} - \frac{2(1+4u^2)}{(1+4u^2+4v^2)^{3/2}} + 2\left(\frac{4uv}{1+4u^2+4v^2}\right)0\right]$$

$$= \frac{1+4v^2-1-4u^2}{(1+4u^2+4v^2)^{3/2}} = \frac{4(v^2-u^2)}{(1+4u^2+4v^2)^{3/2}}.$$

EXAMPLE 7.4.8 Show that the right helicoid given by

$$x^1 = u^1\cos u^2;\ x^2 = u^1\sin u^2;\ x^3 = cu^2$$

is a minimal surface.

Solution: For the first fundamental form of the surface [Eq. (7.2)], $a_{\alpha\beta}$ are given by

$$a_{11} = (\cos u^2)^2 + (\sin u^2)^2 + 0 = 1.$$

$$a_{22} = (-u^1\sin u^2)^2 + (u^1\cos u^2)^2 + (c)^2 = (u^1)^2 + c^2.$$

$$a_{12} = \frac{\partial x^1}{\partial u^1}\frac{\partial x^1}{\partial u^2} + \frac{\partial x^2}{\partial u^1}\frac{\partial x^2}{\partial u^2} + \frac{\partial x^3}{\partial u^1}\frac{\partial x^3}{\partial u^2}$$

$$= \cos u^2(-u^1\sin u^2) + \sin u^2(u^1\cos u^2) + 0 = 0 = a_{21}.$$

Therefore, $a = c^2 + (u^1)^2$. The reciprocal tensors are given by

$$a^{11} = \frac{c^2 + (u^1)^2}{c^2 + (u^1)^2} = 1;\ a^{22} = \frac{1}{c^2 + (u^1)^2};\ a^{12} = 0 = a^{21}.$$

Now, we calculate the tensors of second fundamental form. Since $\mathbf{r} = (x^1, x^2, x^3) = (u^1\cos u^2, u^1\sin u^2, cu^2)$; so,

$$\mathbf{A} = (\cos u^2, \sin u^2, 0) \text{ and } \mathbf{B} = (-u^1\sin u^2, u^1\cos u^2, c)$$

$$\mathbf{A}\times\mathbf{B} = \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ \cos u^2 & \sin u^2 & 0 \\ -u^1\sin u^2 & u^1\cos u^2 & c \end{vmatrix} = c\sin u^2\,\hat{i} - c\cos u^2\,\hat{j} + u^1\hat{k}$$

$$\mathbf{N} = \frac{1}{|\mathbf{A}\times\mathbf{B}|}(\mathbf{A}\times\mathbf{B}) = \frac{1}{\sqrt{c^2 + (u^1)^2}}(c\sin u^2, -c\cos u^2, u^1).$$

Thus, the covariant tensors $b_{\alpha\beta}$ for the second fundamental form are given by

$$b_{11} = 0;\ b_{22} = 0 \text{ and } b_{12} = -\frac{c}{\sqrt{c^2 + (u^1)^2}} = b_{21}.$$

Therefore, the mean curvature H is given by

$$H = \frac{1}{2}a^{\alpha\beta}b_{\alpha\beta} = \frac{1}{2}\left(a^{11}b_{11} + a^{22}b_{22} + 2a^{12}b_{12}\right)$$

$$= \frac{1}{2}\left[1\cdot 0 + \frac{1}{c^2 + (u^1)^2}\cdot 0 + 2\cdot 0\cdot\left(-\frac{c}{\sqrt{c^2 + (u^1)^2}}\right)\right] = 0.$$

Since the mean curvature $H = 0$, at every point of the given surface, so, it is a minimal surface.

EXAMPLE 7.4.9 Calculate the fundamental magnitudes for the conicoid

$$x^1 = u^1 \cos u^2;\ x^2 = u^1 \sin u^2;\ x^3 = f(u^2);$$

where $u^1; u^2$ are parameters. Find also Gaussian and mean curvature.

Solution: The parametric representation of the surface is given by

$$r = (x^1; x^2; x^3) = \left(u^1 \cos u^2;\ u^1 \sin u^2;\ f(u^2)\right);$$

The first fundamental magnitudes $a_{\alpha\beta}$ are given by

$$a_{11} = \left(\frac{\partial x^1}{\partial u^1}\right)^2 + \left(\frac{\partial x^2}{\partial u^1}\right)^2 + \left(\frac{\partial x^3}{\partial u^1}\right)^2 = (\cos u^2)^2 + (\sin u^2)^2 + 0^2 = 1;$$

$$a_{22} = \left(\frac{\partial x^1}{\partial u^2}\right)^2 + \left(\frac{\partial x^2}{\partial u^2}\right)^2 + \left(\frac{\partial x^3}{\partial u^2}\right)^2$$

$$= (-u^1 \sin u^2)^2 + (u^1 \cos u^2)^2 + f_1^2 = (u^1)^2 + f_1^2;$$

$$a_{12} = \frac{\partial x^1}{\partial u^1}\frac{\partial x^1}{\partial u^2} + \frac{\partial x^2}{\partial u^1}\frac{\partial x^2}{\partial u^2} + \frac{\partial x^3}{\partial u^1}\frac{\partial x^3}{\partial u^2} = 0 = a_{21};$$

where $f_1 = \frac{\partial f}{\partial u^2}$. Therefore,

$$a = \begin{vmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 0 & (u^1)^2 + f_1^2 \end{vmatrix} = (u^1)^2 + f_1^2.$$

The reciprocal tensors are given by

$$a^{11} = \frac{(u^1)^2 + f_1^2}{(u^1)^2 + f_1^2} = 1;\ a^{22} = \frac{1}{(u^1)^2 + f_1^2};\ a^{12} = \frac{0}{(u^1)^2 + f_1^2} = 0 = a^{21}.$$

Since $r = (x^1; x^2; x^3) = \left(u^1 \cos u^2;\ u^1 \sin u^2;\ f(u^2)\right)$, we have

$$A = (\cos u^2; \sin u^2; 0) \text{ and } B = (-u^1 \sin u^2; u^1 \cos u^2; f_1)$$

$$A \times B = \begin{vmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ \cos u^2 & \sin u^2 & 0 \\ -u^1 \sin u^2 & u^1 \cos u^2 & f_1 \end{vmatrix} = f_1 \sin u^2 \hat{\imath} + f_1 \cos u^2 \hat{\jmath} + u^1 \hat{k}$$

Therefore, the normal vector is given by

$$= \frac{1}{|A \times B|}(A \times B) = \frac{1}{\sqrt{(u^1)^2 + f_1^2}}\left(f_1 \sin u^2;\ -f_1 \cos u^2; u^1\right).$$

The symmetric covariant tensors b are given by

$$b_{11} = \frac{\partial}{\partial u^1}\frac{\partial}{\partial u^1} = (0\,;0;0)\cdot\frac{1}{\sqrt{(u^1)^2+f_1^2}}\left(f_1\sin u^2;\; f_1\cos u^2; u^1\right) = 0$$

$$b_{22} = \frac{\partial}{\partial u^2}\frac{\partial}{\partial u^2} = (\quad \sin u^2; \cos u^2; 0)\cdot\frac{1}{\sqrt{(u^1)^2+f_1^2}}\left(f_1\sin u^2;\; f_1\cos u^2; u^1\right)$$

$$= \frac{f_1}{\sqrt{(u^1)^2+f_1^2}}$$

$$b_{12} = \left(u^1\cos u^2;\; u^1\sin u^2; f_2\right)\cdot\frac{1}{\sqrt{(u^1)^2+f_1^2}}\left(f_1\sin u^2;\; f_1\cos u^2; u^1\right)$$

$$= \frac{u^1 f_2}{\sqrt{(u^1)^2+f_1^2}} = b_{21};$$

where $f_2 = \frac{\partial^2 f}{(\partial u^1)^2}$. Therefore,

$$b = \begin{vmatrix} b_{11} & b_{12} \\ b_{12} & b_{22} \end{vmatrix} = \begin{vmatrix} 0 & \frac{u^1 f_2}{\sqrt{(u^1)^2+f_1^2}} \\ \frac{u^1 f_2}{\sqrt{(u^1)^2+f_1^2}} & \frac{f_1}{\sqrt{(u^1)^2+f_1^2}} \end{vmatrix} = \frac{\left(u^1 f_2\right)^2}{(u^1)^2+f_1^2}.$$

The Gaussian curvature is given by the formula

$$= \frac{b}{a} = \frac{\left(u^1 f_2\right)^2}{\left((u^1)^2+f_1^2\right)^2} < 0.$$

Therefore, the mean curvature H is given by

$$H = \frac{1}{2}a\quad b = \frac{1}{2}\left(a^{11}b_{11} + a^{22}b_{22} + 2a^{12}b_{12}\right)$$

$$= \frac{1}{2}\left[0 + \frac{1}{(u^1)^2+f_1^2}\left(\frac{f_1}{\sqrt{(u^1)^2+f_1^2}}\right) + 0\right] = \frac{f_1}{2\{(u^1)^2+f_1^2\}^{3/2}}.$$

7.5 Exercises

1. Prove that the space vector perpendicular to ^{i} and tangent to the surface is

$$^i = x^i \quad ;$$

where ^{i} is the unit normal vector in the surface.

2. (i) Prove that for any curve on a surface $^{2} = {}^{2}_{g} + {}^{2}_{(n)}$:
   (ii) Prove that the covariant derivative $b_{;}$ can be written in the form

$$b_{;} = \frac{@b}{@u} \quad a \qquad b \quad a \qquad b \; :$$

3. Show that (i) $b = g_{ij} \, {}^{i}_{;} \, x^{j}$; (ii) $a \quad x^{r}_{;} = 2H \; {}^{r}$:
4. If the space co-ordinates are rectangular Cartesian, show that

$$b = \frac{1}{\sqrt{a}} "_{ijk} \frac{@^2 x^i}{@u@u} x^j_1 x^k_2 :$$

5. Show that a surface is a sphere if and only if its second fundamental form is a non-zero constant multiple of the first fundamental form.
6. Show that the second fundamental form for the
   (i) paraboloid $r = (u; v; u^2 \quad v^2)$ is given by

$$B \quad \frac{2}{\sqrt{4u^2 + 4v^2 + 1}}[(du)^2 \quad (dv)^2]:$$

   (ii) helicoid $r = (u \cos v; u \sin v; f(u) + cv)$ is given by

$$B \quad \frac{1}{\sqrt{u^2 + f_1^2 u^2 + c^2}}[uf_{11}(du)^2 \quad 2cdudv + f_1 u^2 (dv)^2]:$$

   (iii) the surface of revolution $r = (u \cos ; u \sin ; f(u))$ is given by

$$B \quad \frac{1}{\sqrt{1 + f_1^2}}[f_{11}(du)^2 + uf_1(d \;)^2]:$$

7. Determine the second fundamental form of a surface represented in the form $x_3 = f(x_1; x_2)$.
8. Prove by direct calculation that the coefficients of the second fundamental form are the components of a covariant tensor of second order with respect to co-ordinate transformations which preserve the sense of the unit normal vector.
9. Taking $x; y$ as parameters, calculate the fundamental magnitudes to the surface

$$2z = ax^2 + 2hxy + by^2$$

and show that the normal to the surface is given by

$$= \frac{1}{\sqrt{1 + (ax + hy)^2 + (hx + by)^2}}(\quad ax \quad hy; \quad hx \quad by; 1):$$

10. Show that the mean curvature for the

    (i) paraboloid $r = (u; v; u^2 \quad v^2)$ is given by

$$H = \frac{4(v^2 \quad u^2)}{[4u^2 + 4v^2 + 1]^{3=2}}:$$

    (ii) hyperbolic paraboloid $x^1 x^2 = x^3$ is given by

$$H = \quad \frac{x^3}{[(x^1)^2 + (x^2)^2 + 1]^{3=2}}:$$

    (iii) helicoid $r = (u \cos v; u \sin v; f(u) + cv)$ is given by

$$H = \frac{1}{2} \frac{(1 + f_1^2) u^2 f_1 + 2c^2 f_{11} + u f_{11} (u^2 + c^2)}{[u^2 + f_1^2 u^2 + c^2]^{3=2}}:$$

    (iv) right helicoid $r = (x^1; x^2; x^3) = \quad u^1 \cos u^2; u^1 \sin u^2; cu^2 \quad$ is zero.

11. Show that the Gaussian curvature for the

    (i) paraboloid $r = (u; v; u^2 \quad v^2)$ is given by

$$= \quad \frac{4}{[4u^2 + 4v^2 + 1]^2}:$$

    (ii) the surface of revolution $r = (u \cos \; ; u \quad \sin \; ; f \quad (u))$ is given by

$$= \frac{f_1 f_{11}}{u(1 + f_1^2)^2}:$$

    (iii) helicoid $r = (u \cos v; u \sin v; f(u) + cv)$ is given by

$$= \frac{u^3 f_1 f_{22} \quad c^2}{[u^2 + f_1^2 u^2 + c^2]^2}:$$

    (iv) right helicoid $r = (x^1; x^2; x^3) = \quad u^1 \cos u^2; u^1 \sin u^2; cu^2 \quad$ is given by

$$= \quad \frac{c^2}{(u^2 + c^2)^2}:$$

    (v) the anchor ring

$$r = (x^1; x^2; x^3) = \quad (b + a \cos u^1) \cos u^2; (b + a \cos u^1) \sin u^2; a \sin u^1$$

is given by $\quad = \dfrac{\cos u^1}{a(b + a \cos u^1)}:$

12. Calculate the fundamental magnitudes and unit normals for the
    (i) $x^1 = u^1 \cos u^2;\ x^2 = u^1 \sin u^2;\ x^3 = cu^2;$
    (ii) $x^1 = u^1 \cos u^2;\ x^2 = u^1 \sin u^2;\ x^3 = f(u^1),$
    (iii) $x^1 = u^1 \cos u^2;\ x^2 = u^1 \sin u^2;\ x^3 = f(u^1) + cu^2;$

    where $u^1; u^2$ are parameters. Find also Gaussian and mean curvature.
13. Determine the discriminant b of the second fundamental form of a sphere.
14. Prove that a surface for which at every point all coefficients of the second fundamental form vanish is a plane.
15. If two curves on a surface cut at right angles, show that the sum of their geodesic torsion is zero.
16. Prove that a surface, all of whose points are umbilics, is a sphere or a plane.
17. Prove that at each non-umbilical point of a surface, there exist two mutually orthogonal directions for which the normal curvature attains its extreme values.
18. Find the equation for the principal curvatures of the surface

$$x^1 = u^1;\ x^2 = u^2;\ x^3 = f(u^1; u^2);$$

    the x co-ordinates being rectangular Cartesian.
19. Show that the equation of the principal curvatures for
    (i) surface of revolution are $_{(1)} = \frac{f_{11}}{(1+f_1^2)^{3=2}}$ and $_{(2)} = \frac{f_1}{u(1+f_1^2)^{3=2}}$:
    (ii) surface $(x^1)^2 + (x^2)^2 + (x^3)^2 = 1$ are 1 and 1.
20. If two conjugate directions make angles $_1$ and $_2$ with a principal direction, show that

$$\tan\ _1 \tan\ _2 = \frac{1}{2}:$$

21. If a geodesic on a surface is a plane curve, show that it is also a line of curvature and conversely.
22. Prove that the lines of curvature on a surface are given by

$$"\quad a \quad a \quad du \ du = 0:$$

23. Show that the lines of curvature of the helicoid

$$x^1 = u^1 \cos u^2; x^2 = u^1 \sin u^2; x^3 = f(u^1) + cu^2$$

    are given by

$$c\ \ 1 + f_1^2 + u^1 f_1 f_{11}\ \ (du^1)^2 \quad c\ \ (u^1)^2 + c^2 + (u^1)^2 f_1^2\ \ (du^2)^2$$
$$+\ \ f(u^1)^2 + c^2 g u^1 f_{11} \quad (1 + f_1^2)(u^1)^2 f_1\ \ du^1 du^2 = 0:$$

24. Show that the lines of curvature for the surface catenoid, given by

$$x^1 = u^1 \cos u^2;\ x^2 = u^1 \sin u^2;\ x^3 = c \cosh^{1} \frac{u^1}{c};$$

where c is a constant, are the parametric curves.

25. If the co-ordinate curves are lines of curvature, prove that $a_{12} = 0 = b_{12}$ and conversely.

26. Calculate the Christoffel symbols of the second kind for the right helicoid, show that circular helices on the surface are geodesics.

27. Calculate $;H$ for two different parameterisation of the paraboloid $z = a(x^2 + y^2)$:

    (i) as the surface of revolution for which $f = x^1; g = a(x^1)^2$;
    (ii) as the surface $r = (y^1; y^2; a(y^1)^2 + a(y^2)^2)$:

    Interpret the results.

28. Find the values of the mean curvature H and Gaussian curvature , when parametric curves are asymptotic lines.

29. If be the angle between the asymptotic directions, find the value of $\tan\frac{}{2}$:

30. Prove that on a surface is given by

$$x^1 = a(u^1 + u^2); x^2 = b(u^1 \quad u^2);\ x^3 = u^1 u^2;$$

where a and b are constants, the parametric curves are asymptotic lines.

31. Show that the parametric curves are asymptotic lines of the surface

$$x^1 = u^1 \cos u^2; x^2 = u^1 \sin u^2; x^3 = a u^2:$$

Hints: For the surface we have to show that

$$b_{11} = 0; b_{22} = 0 \text{ and } b_{12} = \quad \frac{a}{\sqrt{a^2 + (u^1)^2}}:$$

32. Prove that for an asymptotic direction the normal curvature is zero.

33. Show that if every point of a surface S is parabolic, then S is developable.

Chapter 8

Curves on a Surface

In Chapter 7, we assumed the surface co-ordinates to be u and the space co-ordinates x^i, then the equation of a surface be given by

$$S : x^i = x^i(u^1; u^2): \qquad (8.1)$$

If the co-ordinates u are given as functions of a parameter then the point represented by these co-ordinates describes a curve on the surface as the parameter varies. We shall take the arc length s of the curve as the arc parameter and consequently, the equation of a smooth curve C lying on the surface S is given by

$$C: \quad u = u(s): \qquad (8.2)$$

We see from Eq. (8.1) that the space co-ordinates of any point of the curve are also functions of the arc parameter s and we obtain the space co-ordinates x^i in the form

$$S : x^i = x^i(s) = x^i \;\; u^1(s); u^2(s) \;\; ; \qquad (8.3)$$

which is the equation of C regarded as a space curve. The properties of C can then be studied with the aid of the Frenet{Serret formula, by analyzing the rates of change of the unit tangent vector , the unit principal normal and the unit binormal . Then its curvature and its torsion are connected with these vectors by the Frenet formulae [Eq. (5.17)] as

$$\begin{array}{l} \frac{\;^i}{s} = \;^i \\ \frac{\;^i}{s} = \;^i + \;^i \\ \frac{\;^i}{s} = \;^i \end{array} \qquad (8.4)$$

Considering the curve as a curve on the surface we shall denote its unit tangent vector by and its unit normal vector in the surface by :

8.1 Curves Theory on a Surface

If we regard C as a surface curve, defined by Eq. (8.2), the components of the unit tangent vector are related to the space components ^{i} of the same vector by the formulas

$$^{i} = \frac{dx^i}{ds} = \frac{@x^i}{@u}\frac{du}{ds} \quad x^i \quad ; \text{ where } \quad = \frac{du}{ds}: \qquad (8.5)$$

If is the unit normal to C in the tangent plane to the surface (Figure 8.1) and $_g$ is the geodesic curvature of C, then from Eq. (6.61) we get

$$\frac{}{s} = {}_g \quad :$$

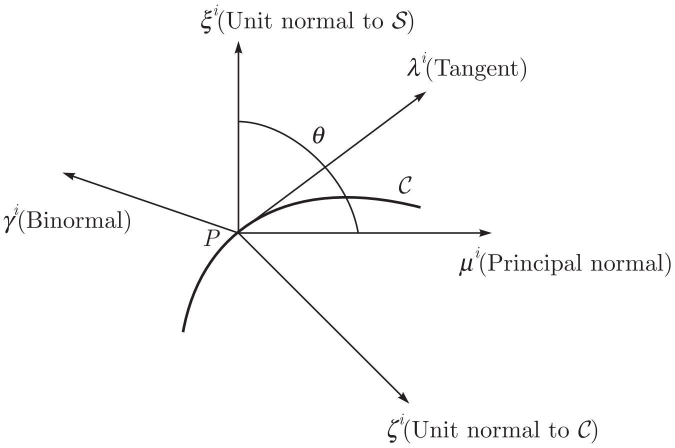

ξ^i(Unit normal to $\mathcal{S}$)
λ^i(Tangent)
θ
$\mathcal{C}$
γ^i(Binormal)
P
μ^i(Principal normal)
ζ^i(Unit normal to $\mathcal{C}$)

Figure 8.1: Tangent and normals on the surface.

If we differentiate Eq. (8.5) intrinsically with respect to s, we get

$$\frac{^i}{s} = x^i_{;} \quad \frac{du}{ds} + x^i \frac{}{s}$$

or

$$^i = x^i_{;} \quad + {}_g x^i \quad ;$$

where we use the Frenet formula $\frac{^i}{s} = \ {}^i$: Using the Gauss's formula $x^i_{;} \ = b \ \ ^i$ and the space components ^{i} of as $^i = x^i$; the Eq. (8.5) becomes

$$^i = b \qquad ^i + {}_g \ ^i; \qquad (8.6)$$

where ^{i} is the unit normal to the surface S. Formula (8.6) states that the principal normal to the curve C lies on the plane of the vectors and . Since ; and are orthogonal and = ; we have,

$$"_{ijk}\ ^{j}\ ^{k} = \ _{i}: \qquad (8.7)$$

Let be the angle between and : Since the tangent vector is orthogonal to the plane of and ; we have

$$= \sin\ ;\ \text{i.e.}\ "_{ijk}\ ^{j}\ ^{k} = \sin\ _{i}: \qquad (8.8)$$

On multiplying by , the foregoing Eq. (8.8) becomes

$$"_{ijk}\ ^{j}\ ^{k} = \sin\ _{i}$$

or

$$"_{ijk}\ b\ \ ^{j} + \ _{g}\ ^{j}\ \ ^{k} = \sin\ _{i};\ \text{using Eq. (8:6)}$$

or

$$_{g} = \sin\ ;\ \text{as}\ "_{ijk}\ ^{j}\ ^{k} = 0\ \text{and}\ "_{ijk}\ ^{j}\ ^{k} = \ _{i}: \qquad (8.9)$$

Also, as be the angle between the principal normal ^{i} and the surface normal ^{i}, then the angle between ^{i} and ^{i} is $\frac{}{2}$; therefore,

$$\cos = \ ^{i}\ ^{j}\ \text{and}\ \cos \frac{}{2} = \ ^{i}\ ^{j}:$$

Therefore, using Eq. (8.6), we get

$$^{i}\ ^{j} = b\ \ ^{i}\ ^{j} + \ _{g}\ ^{i}\ ^{j}$$

or

$$\cos = b\ \ ^{i}\ ^{j} + \ _{g}\ ^{i}\ ^{j} = b\ : \qquad (8.10)$$

From Eq. (8.10) we see that the quantity b is the same for all curves on the surface S which have the same tangent vector at P: In particular, it has the same value for the curve formed by the intersection of the normal plane containing and . But for every normal plane section the angle is either 0 or radians, so that for the normal plane section

$$\cos = \quad \text{or}$$

and consequently, cos is also the same for all such curves. Since b is an invariant, the value of cos for every curve C tangent to is equal to the curvature

$_{(n)}$ of the normal section in the direction . The curvature $_{(n)}$ is called the normal curvature of the surface S in the direction : Thus Eq. (8.10) can be written as

$$_{(n)} = \quad \cos \; = b \qquad : \tag{8.11}$$

Therefore, from Eq. (8.11), we state that, for all curves on a surface, which have the same tangent vector, the quantity cos has the same value, where is the angle between the surface normal and the principal normal, being the curvature of the curve. Accordingly, Eq. (8.6) can be written in the form

$$^{i} = \; _{(n)} \; ^{i} + \; _{g} \; ^{i} : \tag{8.12}$$

Equation (8.12) states that $_{(n)}$ and $_{g}$ are the components of the curvature ^{i} in the direction of the vectors ^{i} and ^{i}.

Result 8.1.1 The radius of curvature $= \frac{1}{}$ of any curve at a given point on the surface is equal to the product of the radius of curvature $_{(n)} = \frac{1}{_{(n)}}$ of the corresponding normal section at that point by cosine of the angle between the normal to the surface and the principal normal to the curve. This is known as Meusnier's theorem Mathematically, we have

$$= \; _{(n)} \cos : \tag{8.13}$$

If S is a sphere, every normal section is a great circle of the sphere, and C is any circle drawn on the sphere, then the result [Eq. (8.13)] becomes obvious from elementary geometric considerations.

If $ds^2 = a \quad du \; du$ and $= \frac{du}{ds}$; Eq. (8.11) can be written in the form

$$_{(n)} = b \qquad = b \; \frac{du}{ds}\frac{du}{ds}$$

$$= \frac{b \;\; du \; du}{ds^2} = \frac{b \;\; du \; du}{a \;\; du \; du} = \frac{B}{A} : \tag{8.14}$$

Result 8.1.2 If the surface is a plane, the normal curvature $_{(n)} = 0$ at all points of the plane, and if it is a sphere $_{(n)} = \frac{1}{R}$; where R is the radius of the sphere. Accordingly, we conclude from Eq. (8.14) that for the plane $b \quad = 0$: Also, for the sphere

$$b \quad du \; du \; = \frac{1}{R} a \quad du \; du \tag{8.15}$$

so that $a \quad = Rb$ at all points of the sphere and for all directions du . Therefore, for a sphere $b \quad = \frac{1}{R} a$, i.e. b and a are proportional.

Deduction 8.1.1 Since ^{i} is perpendicular to ^{i}, it lies in the plane containing ^{i} and ^{i}. Also, it is tangent to the surface and hence the angle between ^{i} and ^{i} is $\frac{}{2}$: If we multiply Eq. (8.6) by ^{i}, we get

$$\cos \frac{}{2} \quad = \ {}_g) \quad {}_g = \ \sin ; \tag{8.16}$$

which is the formula connecting the geodesic curvature of a surface curve with its curvature.

Deduction 8.1.2 From Eq. (8.14), we have

$$_{(n)} = \frac{b \ du \ du}{a \ du \ du} = \frac{B}{A}:$$

Squaring and adding Eqs. (8.11) and (8.16), we obtain

$$^2_{(n)} + \ ^2_g = (\ \cos \ )^2 + (\ \sin \ )^2 = \ ^2:$$

Also, from the Eq. (8.12), it follows that:

$$g_{ij} \quad ^i \ ^j = g_{ij} \, b \qquad ^i \ ^j + \ _g g_{ij} \ ^i \ ^j$$

or

$$\begin{aligned} &= b \qquad \cos \ + \ _g \cos \frac{}{2} \\ &= \ _{(n)} \cos \ + \ _g \sin \ : \end{aligned}$$

Deduction 8.1.3 On using relations (8.9) and (8.10), we get from Eq. (8.6)

$$^i = b \qquad ^i + \ _g \ ^i$$

or

$$^i = \ \cos \ ^i + \ \sin \ ^i$$

or

$$^i = \cos \ ^i + \sin \ ^i:$$

Deduction 8.1.4 If the surface is a plane, then any normal section at any point of the plane is a straight line and, therefore, its curvature is zero and hence $_{(n)} = 0:$ Therefore,

$$b \ du \ du \ = 0;$$

for all directions du , i.e. b $= 0:$ Thus, for a plane b $= 0:$

Deduction 8.1.5 The normal sections of a sphere are great circles, any other plane section is a small circle. At any point the circles of curvature are identical with the normal sections and therefore lie on the sphere.

EXAMPLE 8.1.1 Find the normal curvature of the right helicoid.

Solution: The parametric representation of the right helicoid is given by

$$x^1 = u^1 \cos u^2; \; x^2 = u^1 \sin u^2; \;\; x^3 = cu^2$$

i.e.

$$r = (x^1; x^2; x^3) = \; u^1 \cos u^2; \; u^1 \sin u^2; cu^2 \; :$$

The symmetric covariant second order tensors a for the first fundamental form of the surface are given by

$$a_{11} = 1; \quad a_{22} = (u^1)^2 + c^2 \text{ and } a_{12} = 0 = a_{21}:$$

$$) \, j \; a \; j = \begin{matrix} 1 & 0 \\ 0 & c^2 + (u^1)^2 \end{matrix} = c^2 + (u^1)^2 > 0:$$

The symmetric covariant tensors b for the second fundamental form are given by

$$b_{11} = 0; b_{22} = 0 \text{ and } b_{12} = \frac{c}{\sqrt{c^2 + (u^1)^2}}:$$

$$) \, j \; b \; j = \begin{matrix} 0 & \frac{c}{\sqrt{c^2+(u^1)^2}} \\ \frac{c}{\sqrt{c^2+(u^1)^2}} & 0 \end{matrix} = \frac{c^2}{c^2 + (u^1)^2}:$$

Therefore, the Gaussian curvature is given by

$$= \frac{jb \;\; j}{ja \;\; j} = \frac{c^2}{[c^2 + (u^1)^2]^2} < 0:$$

Thus, the given surface is a surface of negative curvature. Also, it is a minimal surface (as $H = 0$). Thus, the principal curvatures $_{(1)}$ and $_{(2)}$ are related as

$$_{(1)} + \; _{(2)} = 2H = 0 \text{ and } \; _{(1)} : \; _{(2)} = \; = \frac{c^2}{[c^2 + (u^1)^2]^2}:$$

Now, the normal curvature $_{(n)}$ of the given right helicoid is given by

$$_{(n)} = \frac{b \;\; du \;\; du}{a \;\; du \;\; du} = \frac{b_{11}(du^1)^2 + b_{22}(du^2)^2 + 2\,b_{12}du^1du^2}{a_{11}(du^1)^2 + a_{22}(du^2)^2 + 2\,a_{12}du^1du^2}$$

$$= \frac{\left[\,2c= \sqrt{c^2 + (\;u^1)^2}\,\right] du^1du^2}{(du^1)^2 + \{c^2 + (\;u^1)^2\}(du^2)^2}:$$

$$= \frac{1}{\sqrt{c^2 + (\;u^1)^2}}\,\frac{2cdu^1du^2}{[(du^1)^2 + \{c^2 + (\;u^1)^2\}(du^2)^2]}:$$

EXAMPLE 8.1.2 Prove that the normal curvatures in the directions of the co-ordinate curves are $\frac{b_{11}}{a_{11}}$ and $\frac{b_{22}}{a_{22}}$, respectively.

Solution: Let the co-ordinate curve or parametric curve be taken, then for u^1 curve; i.e. $du^2 = 0$, the normal curvature $_{(n1)}$ is given by

$$_{(n1)} = \frac{b \;\; du \;\; du}{a \;\; du \;\; du} = \frac{b_{11}(du^1)^2 + b_{22}(du^2)^2 + 2\,b_{12}du^1du^2}{a_{11}(du^1)^2 + a_{22}(du^2)^2 + 2\,a_{12}du^1du^2}$$

$$= \frac{b_{11}(du^1)^2}{a_{11}(du^1)^2} = \frac{b_{11}}{a_{11}}; \text{ as } du^1 \neq 0:$$

The normal curvature $_{(n2)}$ along the u^2 curve is given by

$$_{(n2)} = \frac{b \;\; du \;\; du}{a \;\; du \;\; du} = \frac{b_{11}(du^1)^2 + b_{22}(du^2)^2 + 2\,b_{12}du^1du^2}{a_{11}(du^1)^2 + a_{22}(du^2)^2 + 2\,a_{12}du^1du^2}$$

$$= \frac{b_{22}(du^2)^2}{a_{22}(du^2)^2} = \frac{b_{22}}{a_{22}}; \text{ as } du^2 \neq 0:$$

EXAMPLE 8.1.3 If a curve is a geodesic on the surface, prove that it is either a straight line or its principal normal is orthogonal to the surface at every point and conversely.

Solution: From the formula, $_g = \;\; \sin\;$; we get for a geodesic

$$_g = \;\; \sin \;\; = 0 \;) \;\; \text{either} \;\; = 0 \text{ or } \sin \;\; = 0:$$

Therefore, when $= 0$; the curve is a straight line. If $\sin \; = 0$; then $= 0$ or : If $= 0$ or , then the principal normal and the surface normal are colinear. Consequently, the principal normal is orthogonal to the surface at every point.

EXAMPLE 8.1.4 Given the paraboloid of revolution $S; x^3 = (\;x^1)^2 + (\;x^2)^2$: Determine the radius and centre of the circle of curvature of the normal section of S at a point $P : x^3 = x^3_0$ whose tangent at P is parallel to the x^1x^2-plane.

Solution: The intersection of the paraboloid S and the plane E : $x^3 = x_0^3$ is a circle (Figure 8.2) of radius $= \sqrt{x_0^3} = r_0$. This is not a normal section of S. Let be the angle between E and the normal to S, this can be obtained

$$\tan \frac{1}{2} \qquad = \tan \quad = 2r_0$$

or

$$\cos = \sin = \frac{\tan}{\sqrt{1+\tan^2}} = \frac{2r_0}{\sqrt{1+4r_0^2}}$$

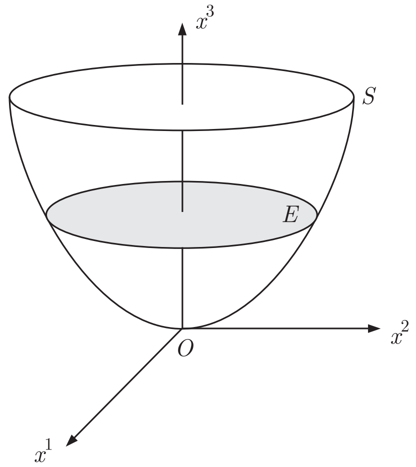

Figure 8.2: Paraboloid of revolution.

According to Meusnier's theorem we thus obtain

$${}_{(n)} = \frac{}{\cos} = \frac{1}{2}\sqrt{1+4r_0^2}$$

The centre of curvature of the normal section under consideration lies on the axis of revolution.

8.2 Principal Curvatures

The point of intersection of consecutive normals along a line of curvature at P is called a centre of curvature of the surface; and its distance from P, measured in the direction of the unit normal, is called a (principal) radius of curvature of the surface. The reciprocal of the principal radius of curvature is called a principal curvature.

Here, we will concern with the determination of the principal curvatures of a surface. In Section 8.1 we considered the normal curvature $_{(n)}$ at an arbitrary point

P of a surface as a function of the direction of the tangent to the normal sections at P as Eq. (8.14)

$$_{(n)} = \frac{b \quad du \ du}{ds^2} = \frac{b \quad du \ du}{a \quad du \ du} = \frac{B}{A}:$$

The normal curvature $_{(n)}$ at a point depends on the directions of $= \frac{du}{ds}$. We shall now nd out those directions $= \frac{du}{ds}$ (i.e. those values of $du^2 : du^1$) on the surface for which the normal curvature $_{(n)}$ has the extreme values. These directions are called the principal directions at the given point and the corresponding values for $_{(n)}$ are called principal normal curvatures of the surface at P. Let us denote $_{(n)}$ by $_{(p)}$ for the principal curvatures.

Since the vector is a unit vector, $_{(n)}$ has to be maximised subject to the constraint a $= 1$: The necessary condition for the extremum is

$$b \qquad + \#a \qquad = 0; \ \# = \text{ Lagrangian parameter}$$

or

$$b \qquad + \#a \qquad = 0$$

or

$$_{(p)} + \#1 = 0 \ ) \quad \# = \quad _{(p)}:$$

Thus, the equation, for the determination of directions yielding extreme values of $_{(n)}$ as $_{(p)}$; can be written as

$$b \qquad _{(p)}a \qquad = 0; \qquad = 1; 2: \tag{8.17}$$

The above set of homogeneous equations will possess non-trivial solutions for if and only if, the values of $_{(p)}$ are the roots of the determinant equation

$$b \qquad _{(p)}a \qquad = 0$$

or

$$\begin{matrix} b_{11} & _{(p)}a_{11} & b_{12} & _{(p)}a_{12} \\ b_{21} & _{(p)}a_{21} & b_{22} & _{(p)}a_{22} \end{matrix} = 0$$

or

$$b_{11}b_{22} \quad (a_{11}b_{22} + a_{22}b_{11}) \ _{(p)} + \ ^{2}_{(p)}a_{11}a_{22}$$
$$b_{12}b_{21} + (\ a_{12}b_{21} + \ a_{21}b_{12}) \ _{(p)} \qquad ^{2}_{(p)}a_{12}a_{21} = 0$$

or

$$(b_{11}b_{22} \quad b_{12}b_{21}) + {}^{2}_{(p)}(a_{11}a_{22} \quad a_{12}a_{21})$$
$$+ {}_{(p)}(a_{12}b_{21} + a_{21}b_{12} \quad a_{11}b_{22} \quad a_{22}b_{11}) = 0:$$

or

$$ {}^{2}_{(p)} \quad \frac{a_{11}b_{22} + b_{11}a_{22} \quad 2a_{12}b_{12}}{a} + \frac{b}{a} = 0; \quad b = jb \quad j; a = ja \quad j: \tag{8.18}$$

The roots of the quadratic Eq. (8.18) determine those directions for which the normal curvature ${}_{(n)}$ becomes extreme. Those directions are called the principal directions of normal curvature (or curvature directions) at the point P under consideration. The centres of curvature of the corresponding normal sections are called the centres of principal curvature of the surface S at P.

Also, since $a_{11} = aa^{22}; a_{12} = \quad aa^{21}; a_{21} = \quad aa^{12}$ and $a_{22} = aa^{11}$, so the above equation can be written in the form

$$a \, {}^{2}_{(p)} \quad (aa^{22}b_{22} + aa^{11}b_{11} + aa^{21}b_{21} + aa^{12}b_{12}) \, {}_{(p)} + b = 0$$

or

$$a \, {}^{2}_{(p)} \quad aa \quad b \quad {}_{(p)} + b = 0$$

or

$${}^{2}_{(p)} \quad a \quad b \quad {}_{(p)} + \frac{b}{a} = 0$$

or

$${}^{2}_{(p)} \quad 2H \, {}_{(p)} + \quad = 0; \tag{8.19}$$

where $H = \frac{1}{2}a \quad b$ is the mean curvature and $= \frac{b}{a}$ is the Gaussian curvature. We shall denote the two roots of ${}_{(p)}$ by ${}_{(1)}$ and ${}_{(2)}$ and call them the principal curvature of the surface and the directions corresponding to ${}_{(1)}$ and ${}_{(2)}$ are called principal directions of the surface. Note that, the principal directions on the surfaces are real.

Definition 8.2.1 Those portions of the surface on which the two principal curvatures ${}_{(1)}$ and ${}_{(2)}$ have the same sign are said to be synclasic. The surface of a sphere or of an ellipsoid as synclasic at all points. On the other hand if the principal curvatures have opposite signs on any part of the surface, this part is said to be anticlastic. The surface of a hyperbolic paraboloid is anticlastic at all points.

Result 8.2.1 The principal directions are real and orthogonal.

Proof: The reality of the principal directions is obvious geometrically. From Eq. (8.19) it is clear that, the principal curvatures ${}_{(1)}$ and ${}_{(2)}$ are related to the mean and Gaussian curvatures by the formulas

$${}_{(1)} + \;{}_{(2)} = 2H \text{ and } {}_{(1)}\;{}_{(2)} = \;: \qquad (8.20)$$

From Eq. (8.17), it follows that, the principal directions ${}_{(1)};\;{}_{(2)}$, say, corresponding to ${}_{(1)}$ and ${}_{(2)}$, respectively, are determined by

$$\left.\begin{array}{l} b \quad {}_{(1)}a \quad {}_{(1)} = 0 \\ b \quad {}_{(2)}a \quad {}_{(2)} = 0 \end{array}\right) \qquad (8.21)$$

Multiplying the first part of Eq. (8.21) by ${}_{(2)}$ and second part of Eq. (8.21) by ${}_{(2)}$ and subtracting, we get

$${}_{(2)} \quad {}_{(1)} \; a \quad {}_{(1)}\;{}_{(2)} = 0: \qquad (8.22)$$

A point at which ${}_{(1)} = \;{}_{(2)}$ is called an umbilic point or naval point. At an umbilic point, Eq. (8.19) has coincident roots, i.e. $H^2 = \;$; so that, it can be written in the form

$$\Big[\; a \quad b \;\Big]^2 = 4\frac{b}{a}$$

or

$$a^{11}b_{11} + a^{22}b_{22} + a^{21}b_{21} + a^{12}b_{12} \;{}^2 = 4\frac{b}{a}$$

or

$$4a(a_{11}b_{12} \quad a_{12}b_{11})^2 + [a_{11}(a_{11}b_{22} \quad a_{22}b_{11}) \quad 2a_{12}(a_{11}b_{12} \quad a_{12}b_{11})]^2 = 0:$$

Since $a_{ij}\,dx^i dx^j$ is positive definite, a is positive, and thus, we have

$$a_{11}b_{12} \quad a_{12}b_{11} = a_{11}b_{22} \quad a_{22}b_{11} = 0$$

$$) \quad \frac{b_{11}}{a_{11}} = \frac{b_{12}}{a_{12}} = \frac{b_{22}}{a_{22}}: \qquad (8.23)$$

Thus, at an umbilic we have the above condition Eq. (8.23). From Eq. (8.14), since

$${}_{(n)} = \frac{b \quad du \; du}{a \quad du \; du};$$

so, ${}_{(n)}$ is independent of the direction $\frac{du}{ds}$; i.e. at an umbilic, the normal curvature is the same in every directions. At all other points where ${}_{(1)} \neq \;{}_{(2)}$, we have

$$a \quad {}_{(1)}\;{}_{(2)} = 0: \qquad (8.24)$$

Therefore, at each non-umbilical point of a surface there exist two mutually orthogonal directions for which the normal curvature attains its extreme values.

If the principal curvatures at any point of a surface are equal in magnitude and opposite in sign and the indication is a rectangular hyperbola then it is a minimal surface.

EXAMPLE 8.2.1 Find the principal curvature of the surface defined by

$$x^1 = u^1, x^2 = u^2; \quad x^3 = f(u^1, u^2).$$

Solution: The parametric representation of the given curve is given by

$$x^1 = u^1;\ x^2 = u^2;\ x^3 = f(u^1, u^2)$$

i.e.

$$r = (x^1, x^2, x^3) = \left(u^1, u^2, f(u^1, u^2)\right).$$

The symmetric covariant second order tensors $a_{\alpha\beta}$ for the first fundamental form of the surface are given by

$$a_{11} = \left(\frac{\partial x^i}{\partial u^1}\right)^2 = 1^2 + 0^2 + \left(\frac{\partial f}{\partial u^1}\right)^2 = 1 + f_1^2.$$

$$a_{22} = \left(\frac{\partial x^i}{\partial u^2}\right)^2 = 0^2 + 1^2 + \left(\frac{\partial f}{\partial u^2}\right)^2 = 1 + f_2^2.$$

$$a_{12} = \frac{\partial x^i}{\partial u^1}\frac{\partial x^i}{\partial u^2} = 1.0 + 0.1 + \frac{\partial f}{\partial u^1}\frac{\partial f}{\partial u^2} = f_1 f_2 = a_{21}$$

where $f_1 = \frac{\partial f}{\partial u^1}$ and $f_2 = \frac{\partial f}{\partial u^2}$. Therefore,

$$a = |a_{\alpha\beta}| = \begin{vmatrix} 1 + f_1^2 & f_1 f_2 \\ f_1 f_2 & 1 + f_2^2 \end{vmatrix} = 1 + f_1^2 + f_2^2.$$

The reciprocal tensors are given by

$$a^{11} = \frac{1 + f_2^2}{1 + f_1^2 + f_2^2};\quad a^{22} = \frac{1 + f_1^2}{1 + f_1^2 + f_2^2};\quad a^{12} = \frac{-f_1 f_2}{1 + f_1^2 + f_2^2} = a^{21}.$$

Now, we calculate the tensors of second fundamental form. Since

$$r = (x^1, x^2, x^3) = \left(u^1, u^2, f(u^1, u^2)\right);$$

so

$$A=(1,0,f_1) \text{ and } B=(0,1,f_2);\quad A\times B=(-f_1,-f_2,1)$$

$$\xi=\frac{1}{|A\times B|}(A\times B)=\frac{1}{\sqrt{f_1^2+f_2^2+1}}(-f_1,-f_2,1).$$

Thus, the covariant tensors $b_{\alpha\beta}$ for the second fundamental form are given by

$$\begin{aligned}
b_{11}&=-\frac{\partial\xi}{\partial u^1}\cdot\frac{\partial X}{\partial u^1}=-(f_1^2+f_2^2+1)^{-1/2}[(-f_{11},-f_{12},0)\cdot(1,0,f_1)\\
&\quad-(f_1^2+f_2^2+1)(f_1f_{11}+f_2f_{12})(-f_1,-f_2,1)\cdot(1,0,f_1)]\\
&=\frac{f_{11}}{\sqrt{f_1^2+f_2^2+1}}.\\
b_{22}&=-\frac{\partial\xi}{\partial u^2}\cdot\frac{\partial X}{\partial u^2}=-(f_1^2+f_2^2+1)^{-1/2}[(-f_{21},-f_{22},0)\cdot(0,1,f_2)\\
&\quad-(f_1^2+f_2^2+1)(f_1f_{21}+f_2f_{22})(f_1,f_2,-1)\cdot(0,1,f_2)]\\
&=\frac{f_{22}}{\sqrt{f_1^2+f_2^2+1}}.
\end{aligned}$$

$$\begin{aligned}
b_{12}&=\frac{1}{2}\left(-\frac{\partial\xi}{\partial u^1}\cdot\frac{\partial X}{\partial u^2}+-\frac{\partial\xi}{\partial u^2}\cdot\frac{\partial X}{\partial u^1}\right)=-\frac{1}{2}(f_1^2+f_2^2+1)^{-1/2}[\{(-f_{11},-f_{12},0)\cdot(0,1,f_2)\\
&\quad+(f_1^2+f_2^2+1)(f_1f_{11}+f_2f_{12})(-f_1,-f_2,1)\cdot(0,1,f_2)\}\\
&\quad+\{(-f_{21},-f_{22},0)\cdot(1,0,f_1)+(f_1^2+f_2^2+1)(f_1f_{21}+f_2f_{22})(f_1,f_2,-1)\cdot(1,0,f_1)\}]\\
&=\frac{f_{12}}{\sqrt{f_1^2+f_2^2+1}}=b_{21}.
\end{aligned}$$

The symmetric covariant tensors $b_{\alpha\beta}$ for the second fundamental form are given by

$$b_{11}=\frac{f_{11}}{\sqrt{f_1^2+f_2^2+1}},\ b_{22}=\frac{f_{22}}{\sqrt{f_1^2+f_2^2+1}} \text{ and } b_{12}=\frac{f_{12}}{\sqrt{f_1^2+f_2^2+1}}=b_{21}.$$

$$\Rightarrow\ b=|b_{\alpha\beta}|=\begin{vmatrix}\frac{f_{11}}{P} & \frac{f_{12}}{P}\\ \frac{f_{12}}{P} & \frac{f_{22}}{P}\end{vmatrix}=\frac{f_{11}f_{22}-f_{12}^2}{1+f_1^2+f_2^2};\quad P=\sqrt{f_1^2+f_2^2+1};$$

where $f_i=\frac{\partial f}{\partial u^i}$ and $f_{ij}=\frac{\partial^2 f}{\partial u^i\partial u^j}$. The equation giving the principal curvatures is

$$\kappa^2-a^{\alpha\beta}b_{\alpha\beta}\kappa+\frac{b}{a}=0$$

or

$$\kappa^2 - \kappa\left[(f_{11}+f_{22}) - \frac{f_1^2 f_{11} + f_2^2 f_{22} + 2f_1 f_2 f_{12}}{1+f_1^2+f_2^2}\right] + f_{11}f_{22} - f_{12}^2 = 0.$$

This is a quadratic equation in κ which gives the two values of κ. At an umbilic, we have

$$\frac{b_{11}}{a_{11}} = \frac{b_{12}}{a_{12}} = \frac{b_{22}}{a_{22}} \Rightarrow \frac{f_{11}}{1+f_1^2} = \frac{f_{12}}{f_1 f_2} = \frac{f_{22}}{1+f_2^2}.$$

EXAMPLE 8.2.2 Find the principal curvature of the surface defined by

$$x^1 = u^1 \cos u^2;\ x^2 = u^1 \sin u^2;\ x^3 = f(u^1).$$

Find the condition that it is a minimal surface.

Solution: The parametric representation of the surface of revolution is given by

$$x^1 = u^1 \cos u^2; x^2 = u^1 \sin u^2;\ x^3 = f(u^1)$$

i.e.

$$\mathbf{r} = (x^1; x^2; x^3) = \left(u^1 \cos u^2; u^1 \sin u^2; f(u^1)\right).$$

The first fundamental magnitudes $a_{\alpha\beta}$ of the surface are given by

$$a_{11} = 1 + f_1^2;\ a_{22} = (u^1)^2 \text{ and } a_{12} = 0 = a_{21}.$$

$$\Rightarrow a = |a_{\alpha\beta}| = \begin{vmatrix} 1+f_1^2 & 0 \\ 0 & (u^1)^2 \end{vmatrix} = (u^1)^2(1+f_1^2);$$

where $f_1 = \frac{\partial f}{\partial u^1}$. The reciprocal tensors are given by

$$a^{11} = \frac{1}{1+f_1^2};\ a^{22} = \frac{1}{(u^1)^2};\ a^{12} = 0 = a^{21}.$$

The symmetric covariant tensors $b_{\alpha\beta}$ for the second fundamental form are given by

$$b_{11} = \frac{f_{11}}{\sqrt{1+f_1^2}}; b_{22} = \frac{u^1 f_1}{\sqrt{1+f_1^2}} \text{ and } b_{12} = 0 = b_{21}.$$

$$\Rightarrow |b_{\alpha\beta}| = \begin{vmatrix} \frac{f_{11}}{\sqrt{1+f_1^2}} & 0 \\ 0 & \frac{u^1 f_1}{\sqrt{1+f_1^2}} \end{vmatrix} = \frac{u^1 f_1 f_{11}}{1+f_1^2};$$

where $f_1 = \frac{\partial f}{\partial u^1}$ and $f_{11} = \frac{\partial^2 f}{\partial u^1 \partial u^1}$. The equation giving the principal curvatures is

$$\kappa^2 - a^{\alpha\beta} b_{\alpha\beta}\kappa + \frac{b}{a} = 0$$

or

$$\kappa^2 - \left[\frac{f_{11}}{(1+f_1^2)^{3/2}} + \frac{f_1}{u^1\sqrt{1+f_1^2}}\right]\kappa + \frac{f_1 f_{11}}{u^1(1+f_1^2)^2} = 0.$$

This is a quadratic equation in κ which gives the following two values of κ:

$$\kappa_{(1)} = \frac{f_{11}}{(1+f_1^2)^{3/2}}; \quad \kappa_{(2)} = \frac{f_1}{u^1(1+f_1^2)^{1/2}}.$$

If ρ_1 and ρ_2 be the corresponding radii of curvatures then

$$\rho_1 = \frac{1}{\kappa_{(1)}} = \frac{(1+f_1^2)^{3/2}}{f_{11}}; \quad \rho_2 = \frac{1}{\kappa_{(2)}} = \frac{u^1(1+f_1^2)^{1/2}}{f_1}.$$

The condition for the surface to be minimal is that

$$2H = \kappa_{(1)} + \kappa_{(2)} = 0$$

or

$$\frac{f_{11}}{(1+f_1^2)^{3/2}} + \frac{f_1}{u^1(1+f_1^2)^{1/2}} = 0$$

or

$$f_1(1+f_1^2) + u^1 f_{11} = 0.$$

EXAMPLE 8.2.3 Find the principal curvature of the surface defined by

$$x^1 = u^1\cos u^2; x^2 = u^1\sin u^2; \quad x^3 = a\log\left(u^1 + \sqrt{(u^1)^2 - a^2}\right).$$

Prove that it is a minimal surface and is the only minimal surface.

Solution: From example 8.2.2 we see that, the parametric representation of the surface of revolution is given by

$$x^1 = u^1\cos u^2; \; x^2 = u^1\sin u^2; \; x^3 = f(u^1);$$

where $f(u^1) = a\log\left(u^1 + \sqrt{(u^1)^2 - a^2}\right)$. Therefore,

$$f_1 = \frac{\partial f}{\partial u^1} = \frac{a}{\sqrt{(u^1)^2 - a^2}}; \; f_{11} = \frac{\partial^2 f}{(\partial u^1)^2} = -\frac{au^1}{[(u^1)^2 - a^2]^{3/2}}.$$

Therefore, $\kappa_{(1)}$ and $\kappa_{(2)}$ are given by

$$\kappa_{(1)} = \frac{f_{11}}{(1+f_1^2)^{3/2}} = -\frac{a}{(u^1)^2}; \quad \kappa_{(2)} = \frac{f_1}{u^1(1+f_1^2)^{1/2}} = \frac{a}{(u^1)^2}.$$

Since $2H = \kappa_{(1)} + \kappa_{(2)} = 0$; the surface is a minimal surface. Now, we will prove that the given surface of revolution is the minimal surface. Let the equation of any surface of revolution be

$$x^1 = u^1 \cos u^2;\ x^2 = u^1 \sin u^2;\quad x^3 = f(u^1).$$

As in example 8.2.2, the condition for the minimal surface is

$$f_1(1 + f_1^2) + u^1 f_{11} = 0$$

or

$$u^1 \frac{d^2f}{d(u^1)^2} + \frac{df}{du^1}\left[1 + \left(\frac{df}{du^1}\right)^2\right] = 0$$

or

$$\frac{dp}{p(1+p^2)} + \frac{du}{u} = 0;\quad \text{where, } p = \frac{df}{du^1}$$

or

$$\log p - \frac{1}{2}\log(1+p^2) + \log u^1 = \text{constant}$$

or

$$\frac{p^2(u^1)^2}{1+p^2} = a^2 \text{ (say)} \Rightarrow p = \frac{df}{du^1} = \frac{a}{\sqrt{(u^1)^2 - a^2}}$$

or

$$x^3 = f(u^1) = a\cosh^{-1}\frac{u^1}{a} + c_1 = a\log\left(u^1 + \sqrt{(u^1)^2 - a^2}\right) + c_1.$$

Choosing $x^3 = a \log a$ when $u^1 = a$ so that $c_1 = 0$, so,

$$x^3 = f(u^1) = a\log\left(u^1 + \sqrt{(u^1)^2 - a^2}\right)$$
$$= a\cosh^{-1}\frac{\sqrt{(x^1)^2 + (x^2)^2}}{a}$$

or

$$\sqrt{(x^1)^2 + (x^2)^2} = a\cosh\frac{x^3}{a}.$$

Thus, except for position in space, the only surface of revolution is

$$x^1 = u^1 \cos u^2;\ x^2 = u^1 \sin u^2;\quad x^3 = a\log\left(u^1 + \sqrt{(u^1)^2 - a^2}\right)$$

or $\sqrt{(x^1)^2 + (x^2)^2} = a\cosh\frac{x^3}{a}$ is the only minimal surface formed by revolution of catenary $x^1 = a\cosh\frac{x^3}{a}$ about x^1 axis, i.e. directrix.

EXAMPLE 8.2.4 Find the equation for the principal curvatures of the surface for the right helicoid.

Solution: The parametric representation of the right helicoid is given by

$$x^1 = u^1 \cos u^2;\ x^2 = u^1 \sin u^2;\ x^3 = cu^2$$

i.e.

$$r = (x^1; x^2; x^3) = \quad u^1 \cos u^2;\ u^1 \sin u^2; cu^2 \quad :$$

The symmetric covariant second order tensors a for the first fundamental form of the surface are given by

$$a_{11} = 1;\ a_{22} = (u^1)^2 + c^2 \text{ and } a_{12} = 0 = a_{21}:$$

$$) j \ a \ j = \begin{matrix} 1 & 0 \\ 0 & c^2 + (u^1)^2 \end{matrix} = c^2 + (u^1)^2:$$

The reciprocal tensors are given by

$$a^{11} = \frac{c^2 + (u^1)^2}{c^2 + (u^1)^2} = 1;\ a^{22} = \frac{1}{c^2 + (u^1)^2};\ a^{12} = 0 = a^{21}:$$

The symmetric covariant tensors b for the second fundamental form are given by

$$b_{11} = 0; b_{22} = 0 \text{ and } b_{12} = \text{p} \frac{c}{c^2 + (u^1)^2}:$$

$$) j \ b \ j = \begin{matrix} 0 & \text{p} \frac{c}{c^2+(u^1)^2} \\ \text{p} \frac{c}{c^2+(u^1)^2} & 0 \end{matrix} = \frac{c^2}{c^2 + (u^1)^2}:$$

The equation giving the principal curvatures is

$$^2 \quad a \ b \quad + \frac{b}{a} = 0$$

or

$$^2 \quad a^{11} b_{11} + a^{22} b_{22} + 2 a^{12} b_{12} \quad + \frac{b}{a} = 0$$

or

$$^2 \quad 1 \ 0 + \frac{1}{c^2 + (u^1)^2} \ 0 + 0 \quad \frac{c^2}{c^2 + (u^1)^2} \quad + \frac{b}{a} = 0$$

or

$$^2 \quad \frac{c^2}{[c^2 + (u^1)^2]^2} = 0 \ ) \quad = \frac{c}{c^2 + (u^1)^2}:$$

This is a quadratic equation in κ which gives the following two values of κ:

$$\kappa_{(1)} = \frac{c}{c^2 + (u^1)^2}; \quad \kappa_{(2)} = -\frac{c}{c^2 + (u^1)^2}.$$

Since $2H = \kappa_{(1)} + \kappa_{(2)} = 0$; the surface is a minimal surface.

EXAMPLE 8.2.5 Calculate the principal curvatures for the conicoid

$$x^1 = u^1 \cos u^2; \; x^2 = u^1 \sin u^2; \; x^3 = f(u^2);$$

where $u^1; u^2$ are parameters. Hence, find the Gaussian and the mean curvature.

Solution: The parametric representation of the surface is given by

$$\mathbf{r} = (x^1; x^2; x^3) = \left(u^1 \cos u^2; \; u^1 \sin u^2; \; f(u^2)\right).$$

The first fundamental magnitudes $a_{\alpha\beta}$ are given by

$$a_{11} = 1; \; a_{22} = (u^1)^2 + f_1^2; \; a_{12} = 0 = a_{21};$$

where $f_1 = \frac{\partial f}{\partial u^2}$. Therefore, $a = (u^1)^2 + f_1^2$. The reciprocal tensors are given by

$$a^{11} = \frac{(u^1)^2 + f_1^2}{(u^1)^2 + f_1^2} = 1; \quad a^{22} = \frac{1}{(u^1)^2 + f_1^2}; \quad a^{12} = \frac{0}{(u^1)^2 + f_1^2} = 0 = a^{21}.$$

The symmetric covariant tensors $b_{\alpha\beta}$ are given by

$$b_{11} = 0; \; b_{22} = \frac{f_1}{\sqrt{(u^1)^2 + f_1^2}}; \; b_{12} = -\frac{u^1 f_2}{\sqrt{(u^1)^2 + f_1^2}} = b_{21};$$

where $f_2 = \frac{\partial^2 f}{(\partial u^2)^2}$ so, $b = -\frac{(u^1 f_2)^2}{(u^1)^2 + f_1^2}$. The equation giving the principal curvatures is

$$\kappa^2 - a^{\alpha\beta} b_{\alpha\beta} \kappa + \frac{b}{a} = 0$$

or

$$\kappa^2 + \frac{f_1}{\{(u^1)^2 + f_1^2\}^{3/2}} \kappa - \left(\frac{u^1 f_2}{(u^1)^2 + f_1^2}\right)^2 = 0.$$

This is a quadratic equation in κ which gives the two values $\kappa_{(1)}$ and $\kappa_{(2)}$. The mean curvature H is given by

$$H = \frac{1}{2}[\kappa_{(1)} + \kappa_{(2)}] = -\frac{1}{2} \frac{f_1}{\{(u^1)^2 + f_1^2\}^{3/2}}$$

and the Gaussian curvature K is given by the formula,

$$K = \kappa_{(1)} \kappa_{(2)} = -\left(\frac{u^1 f_2}{(u^1)^2 + f_1^2}\right)^2.$$

EXAMPLE 8.2.6 Show that the surface $4c^2(x^3)^2 = ((x^1)^2 - 2c^2)((x^2)^2 - 2c^2)$ has a line of umbilic lying on the surface $(x^1)^2 + (x^2)^2 + (x^3)^2 = 4c^2.$

Solution: Let $(x^1)^2 - 2c^2 = (u^1)^2$ and $(x^2)^2 - 2c^2 = (u^2)^2$; then the parametric representation of the surface is given by

$$r = (x^1, x^2, x^3) = \left(\sqrt{2c^2 + (u^1)^2}, \sqrt{2c^2 + (u^2)^2}, \frac{u^1u^2}{2c}\right).$$

The first fundamental magnitudes $a_{\alpha\beta}$ are given by

$$a_{11} = \frac{(u^1)^2}{4c^2 + (u^1)^2} + \frac{(u^2)^2}{4c^2};\ a_{22} = \frac{(u^2)^2}{4c^2 + (u^2)^2} + \frac{(u^1)^2}{4c^2};\ a_{12} = \frac{u^1u^2}{4c^2} = a_{21},$$

where $f_1 = \frac{\partial f}{\partial u^1}$. The symmetric covariant tensors $b_{\alpha\beta}$ are given by

$$b_{11} = \frac{c(u^2)^2}{a(2c^2 + (u^1)^2)^{3/2}(2a^2 + (u^2)^2)^{1/2}},$$

$$b_{22} = \frac{c(u^1)^2}{a(2c^2 + (u^1)^2)^{1/2}(2a^2 + (u^2)^2)^{3/2}},$$

$$b_{12} = \frac{cu^2}{2ca\sqrt{(2c^2 + (u^1)^2)(2a^2 + (u^2)^2)}},$$

where $a = |a_{\alpha\beta}|$. At an umbilic, we have $\frac{b_{11}}{a_{11}} = \frac{b_{12}}{a_{12}} = \frac{b_{22}}{a_{22}}$, i.e. $b_{11}a_{22} = b_{22}a_{11}$, and so,

$$\frac{(u^1)^2}{2c^2 + (u^2)^2} + \frac{(u^2)^2}{4c^2}\ \frac{(u^1)^2}{2c^2 + (u^2)^2} = \frac{(u^2)^2}{2c^2 + (u^2)^2} + \frac{(u^1)^2}{4c^2}\ \frac{(u^2)^2}{2c^2 + (u^1)^2}$$

or

$$\frac{(u^1)^4 - (u^2)^4}{(2c^2 + (u^1)^2)(2c^2 + (u^2)^2)} + \frac{(u^1)^2(u^2)^2}{4c^2}\ \frac{1}{2c^2 + (u^2)^2} - \frac{1}{2c^2 + (u^1)^2} = 0$$

or

$$(u^1)^4 - (u^2)^4 + \frac{(u^1)^2(u^2)^2}{4c^2}\ (u^1)^2 - (u^2)^2 = 0$$

or

$$(u^1)^2 + (u^2)^2 + \frac{(u^1u^2)^2}{4c^2} = 0 \Rightarrow (x^1)^2 + (x^2)^2 + (x^3)^2 = 4c^2.$$

Thus, the umbilic lie on the surface $(x^1)^2 + (x^2)^2 + (x^3)^2 = 4c^2.$

8.2.1 Surfaces of Positive and Negative Curvatures

In order to determine the shape of a surface S in a neighbourhood of any of its points we first considered arbitrary curves on S. We enabled us to restrict of S and planes which are orthogonal to the tangent plane to S at a point P under consideration. That investigation led us to the introduction of the normal curvature [Eq. (8.14)]

$$_{(n)} = \frac{b \quad du \ du}{a \quad du \ du} = \frac{B}{A}:$$

$j\ _{(n)}j$ is the curvature of the normal section of S at P whose tangent direction is $du^2 : du^1$: We will now see that there are three different possible forms of S in a neighbourhood of a point at which the second fundamental form does not vanish identically. For this purpose we consider the behaviour of $_{(n)}$ as a function of the direction $du^2 : du^1$ of the tangent to the normal sections at P.

We know, the first fundamental form [Eq. (7.2)] is positive definite, hence the sign of $_{(n)}$ depends on the second fundamental form only.

(i) A surface is called a surface of positive curvature, if at all points, the Gaussian curvature > 0; i.e.

$$b = b_{11}b_{22} \quad b_{12}^2 > 0; \quad \text{as } a > 0 \text{ on the surface} \tag{8.25}$$

That is the second fundamental form is positive definite. In this case, $_{(n)}$ has the same sign for all possible directions of the normal sections at P, i.e. the centres of curvature of all normal sections lie on the same side of the surface S. The point P on the surface S is called elliptic. Figure 8.3 shows the shape of a surface S in the neighbourhood of an elliptic point.

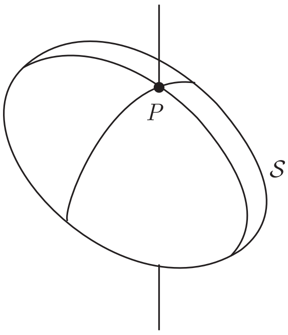

Figure 8.3: Elliptic point.

(ii) A surface is called a surface of negative curvature, if at all points, the Gaussian curvature < 0; i.e.

$$b = b_{11}b_{22} \quad b_{12}^2 < 0; \quad \text{as } a > 0 \text{ on the surface} \tag{8.26}$$

If b < 0 at a point P of a surface ${}_{(n)}$ does not maintain the same sign for all directions $du^2 : du^1$. More precisely, there exists two real asymptotic directions for which ${}_{(n)} = 0$. These directions separate the directions for which ${}_{(n)}$ is positive from which ${}_{(n)}$ is negative, P is then called a hyperbolic or saddle point of the surface S. Figure 8.4 shows the shape of a surface in the neighbourhood of a hyperbolic point.

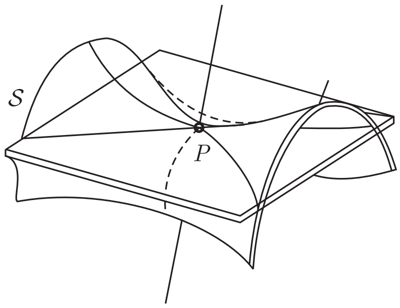

Figure 8.4: Hyperbolic point.

This figure shows also the tangent plane at that point.

(iii) If b = 0 at a point P of a surface S, ${}_{(n)}$ does not change sign, but there is exactly one direction where ${}_{(n)} = 0$; i.e. exactly one real asymptotic direction. P is then called a parabolic point or flat point or planner point of the surface S. Figure 8.5 shows the shape of a surface S in the neighbourhood of a parabolic point.

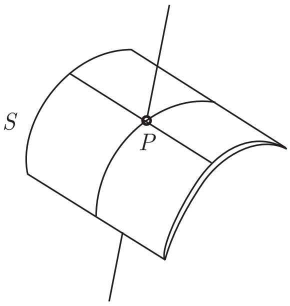

Figure 8.5: Parabolic point.

b of second fundamental form	Name	Number of real asymptotic directions	Figures
b > 0	Elliptic point	0	8.3
b = 0	Parabolic point	1	8.5
b < 0	Hyperbolic point	2	8.4

Since the conditions b >= 0 corresponds to a geometric property of the surface, they must be invariant with respect to any allowable co-ordinate transformation.

Deduction 8.2.1 Consider the section of the surface by a plane parallel and infinitely close to the tangent plane at the point P (Figure 8.6). Suppose the surface is synclastic in the neighbourhood of P. Then near P it lies entirely on one side of the tangent plane. Let the plane be taken on this (concave) side of the surface, parallel to the tangent plane at P, and at an infinitesimal distance from it, whose measure is h in the direction of the unit normal . Let the principal radii of curvature be ; , then

$$ {}_{(1)} = 1 \quad \text{and} \quad {}_{(2)} = 1: $$

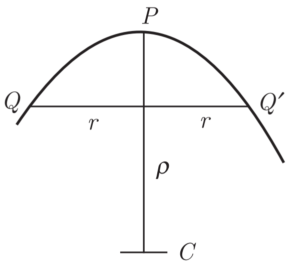

Figure 8.6: Dupin's indicatrix.

Thus, h has the same sign as and . Consider also any normal plane QPQ^0 through P, cutting the former plane in QQ^0. Then if is the radius of curvature of the normal section, and $2r$ the length of QQ^0, we have $r^2 = 2h$ to the first order. If is the inclination of the normal section to the principal direction, Euler's formula gives

$$ \frac{1}{}\cos^2 + \frac{1}{}\sin^2 = \frac{1}{} = \frac{2h}{r^2}; $$

If then we write $\& = r\cos$ and $= r\sin$, we have

$$ \frac{\&^2}{a} + \frac{{}^2}{b} = 2h: $$

Thus, the section of the surface by the plane parallel to the tangent plane at P, and indefinitely close to it, is similar and similarly situated to the ellipse

$$ \frac{\&^2}{j\ j} + \frac{{}^2}{j\ j} = 1; \tag{8.27} $$

whose axes are tangents to the lines of curvature at P. This ellipse is called the Dupin's indicatrix at the point P, and P is said to be an elliptic point. It is sometimes described as a point of positive curvature, because the second curvature is positive.

Let the principal radii, and , have opposite signs, and the surface lies partly on the other side of the tangent plane at P. In this case the Gaussian curvature is negative at P, so that the surface is anticlastic in the neighbourhood. Two planes parallel to this tangent plane, one on either side, and equidistant from it, cut the surface in the conjugate hyperbola

$$\frac{\&^2}{a} + \frac{^2}{b} = \quad 2h:$$

These are similar and similarly situated to the conjugate hyperbolas

$$\frac{\&^2}{} + \frac{^2}{} = \quad 1; \tag{8.28}$$

which constitute the indicatrix at P. The point P is then called a hyperbolic point, or a point of negative curvature. The normal curvature is zero in the directions of the asymptotes. When the Gaussian curvature is zero at the point P it is called a parabolic point. One of the principal curvatures is zero, and the indicatrix is a pair of parallel straight lines. Consequently, the principal directions bisect the angles between directions corresponding to the same normal curvature, in particular the angles between the asymptotic direction.

EXAMPLE 8.2.7 Show that the right helicoid is a surface of negative curvature.

Solution: The parametric representation of the right helicoid is given by

$$x^1 = u^1 \cos u^2;\ x^2 = u^1 \sin u^2;\ x^3 = cu^2$$

i.e.

$$r = (x^1; x^2; x^3) = \quad u^1 \cos u^2;\ u^1 \sin u^2;\ cu^2 \quad :$$

The symmetric covariant second order tensors a for the first fundamental form of the surface are given by

$$a_{11} = 1;\ \ a_{22} = (u^1)^2 + c^2 \text{ and } a_{12} = 0 = a_{21}:$$

$$\Rightarrow |\ a\ | = \begin{matrix} 1 & 0 \\ 0 & c^2 + (u^1)^2 \end{matrix} = c^2 + (u^1)^2 > 0:$$

The symmetric covariant tensors b for the second fundamental form are given by

$$b_{11} = 0;\ b_{22} = 0 \text{ and } b_{12} = \quad \frac{c}{\sqrt{c^2 + (u^1)^2}}:$$

$$\Rightarrow |\ b\ | = \begin{matrix} 0 & \frac{c}{\sqrt{c^2+(u^1)^2}} \\ \frac{c}{\sqrt{c^2+(u^1)^2}} & 0 \end{matrix} = \frac{c^2}{c^2 + (u^1)^2}:$$

Therefore, the Gaussian curvature is given by

$$= \frac{|b \quad|}{|a \quad|} = \frac{c^2}{[c^2 + (u^1)^2]^2} < 0:$$

Thus, the given surface is a surface of negative curvature.

EXAMPLE 8.2.8 Find the nature of a point on a unit sphere

$$(x^1)^2 + (x^2)^2 + (x^3)^2 = 1:$$

Solution: The parametric representation of a unit sphere is given by

$$x^1 = \cos u^1 \cos u^2; \; x^2 = \sin u^1 \cos u^2; \; x^3 = \sin u^2$$

i.e.

$$r = (x^1; x^2; x^3) = \quad \cos u^1 \cos u^2; \sin u^1 \cos u^2; \sin u^2 \quad :$$

The first fundamental magnitudes a of the surface are given by

$$a_{11} = \cos^2 u^2; \; a_{22} = 1 \text{ and } a_{12} = 0 = a_{21}:$$

$$) \quad a = |a \quad| = \begin{matrix} \cos^2 u^2 & 0 \\ 0 & 1 \end{matrix} = \cos^2 u^2 > 0:$$

The symmetric covariant tensors b for the second fundamental form are given by

$$b_{11} = \quad \cos^2 u^2; b_{12} = 0 = b_{21} \text{ and } b_{22} = \quad 1:$$

$$) \quad b = |b \quad| = \begin{matrix} \cos^2 u^2 & 0 \\ 0 & 1 \end{matrix} = \cos^2 u^2:$$

Therefore, the Gaussian curvature is given by

$$= \frac{b}{a} = \frac{\cos^2 u^2}{\cos^2 u^2} = 1 > 0:$$

Since > 0, it follows that the nature of the point on the sphere is elliptic.

8.2.2 Isometric Lines

Suppose that, in terms of the parameters $u; v$ the square of the linear element of the surface has the form

$$ds^2 = \quad (du^2 + dv^2); \tag{8.29}$$

where is a function of $u; v$ or a constant. Then the parameters curves are orthogonal because $a_{12} = 0$: Further, the lengths of elements of the parametric curves are $\sqrt{\quad} du$

and $\sqrt{\lambda}\,dv$; and these equal if $du = dv$. Thus, the parametric curves corresponding to the values $u, u + du, v, v + dv$ bound a small square provided $du = dv$. In this way the surface may be mapped out into small squares by means of the parametric curves, the sides of any one square corresponding to equal increments in u and v.

More generally, if the square of the linear element has the form

$$ds^2 = \lambda(U\,du^2 + V\,dv^2); \tag{8.30}$$

where U is a function of u only and V is a function of v only, we may change the parameters to θ, φ by the transformation

$$d\theta = \sqrt{U}\,du; \quad d\varphi = \sqrt{V}\,dv.$$

This does not alter the parametric curves; for the curves u = constant are identical with the curves θ = constant; and similarly the curves v = constant are identical with the curves φ = constant. Equation (8.30) then becomes

$$ds^2 = \lambda(d\theta^2 + d\varphi^2); \tag{8.31}$$

which is of the same form as Eq. (8.29). Whenever the square of the linear element has the form Eq. (8.30), so that, without alteration of the parametric curves, it may be reduced to the form Eq. (8.29), the parametric curves are called isometric lines, and the parameters isometric parameters.

In the form Eq. (8.29) the fundamental magnitudes a_{11} and a_{22} are equal; but in the more general form Eq. (8.30) they are such that

$$\frac{a_{11}}{a_{22}} = \frac{U}{V} \text{ and, therefore, } \frac{\partial^2}{\partial u \partial v} \log \frac{a_{11}}{a_{22}} = 0. \tag{8.32}$$

Either of these equations, in conjunction with $a_{12} = 0$, expresses the condition that the parametric variables may be isometric. If it is satisfied, ds^2 has the form Eq. (8.30) and may therefore be reduced to the form Eq. (8.29).

EXAMPLE 8.2.9 Prove that isometric curves is afforded by the meridians and parallels on a surface of revolution, given by

Solution: With the usual notation, let S is a surface of revolution, given by

$$x^1 = u\cos v; \; x^2 = u\sin v; \; x^3 = f(u); \text{ where } f(u) \in C^2.$$

The first order magnitudes are given by

$$a_{11} = 1 + f_1^2; \; a_{22} = u^2 \text{ and } a_{12} = 0 = a_{21};$$

where $f_1 = \frac{\partial f}{\partial u}$; In terms of the parameters $u; v$ the square of the linear element of the surface has the form

$$ds^2 = (1 + f_1^2)du^2 + u^2dv^2 = u^2 \quad \frac{1 + f_1^2}{u^2}du^2 + dv^2 \quad ;$$

which is of the form Eq. (8.30). The parametric curves are the meridians v = constant and the parallels u = constant. If we make the transformation

$$d\omega = \frac{1}{u}\sqrt{1 + f_1^2}\,du;$$

the curves ω = constant are the same as the parallels, and the square of the linear element becomes

$$ds^2 = u^2(d\omega^2 + dv^2);$$

which is of the form Eq. (8.29). Thus, the meridians and parallels of a surface of revolution are isometric lines.

8.3 Lines of Curvature

A curve drawn on a surface, and possessing the property that the normals to the surface at consecutive points intersect, is called a line of curvature. Thus, for a line of curvature, a curve on a surface such that the tangent line to it at every point is directed along a principal direction. The differential equation for which the lines of curvature on S are the integral curves from directly from Eq. (8.17), as

$$b \qquad = \quad {}_{(p)}a \qquad :$$

Therefore,

$$b_1 \quad = \quad {}_{(p)}a_1 \quad ; \text{ when } \quad = 1$$

$$b_2 \quad = \quad {}_{(p)}a_2 \quad ; \text{ when } \quad = 2:$$

If we eliminate ${}_{(p)}$ from these equations and set $= \frac{du}{ds}$; we get

$$\frac{b_1}{a_1}\frac{du}{du} = \frac{b_2}{a_2}\frac{du}{du}$$

or

$$(b_{11}a_{12} \quad b_{12}a_{11})(du^1)^2 + (b_{11}a_{22} \quad b_{22}a_{11})du^1du^2$$
$$+(b_{12}a_{22} \quad b_{22}a_{12})(du^2)^2 = 0: \tag{8.33}$$

This quadratic Eq. (8.33) represents the lines of curvature of the surface, showing that at every point of the surface there are two principal directions. Thus, at each point of a surface there exist two mutually orthogonal directions for which the normal curvature attains its extreme values.

It follows from the above that the direction of a line of curvature at any point is the principal direction at that point. Through each point on the surface pass two lines of curvature cutting each other at right angles; and on the surface there are two system of lines of curvature whose differential equation is Eq. (8.33).

Deduction 8.3.1 The existence of (real) asymptotic curves on a surface S depends on the geometric shape of S the lines of curvature are always real. At each point of S where either $b \quad du \quad du \neq 0$ or is not proportional to $a \quad du \quad du$, the Eq. (8.33) specifies two orthogonal directions

$$\frac{du^2}{du^1} = \quad u^1; u^2 \quad ; \quad = 1; 2; \tag{8.34}$$

which coincide with directions of the principal curvatures. Each equation in [Eq. (8.34)] determines a family of curves on S covering the surface without gaps. These two families of curves are orthogonal, and, if they are taken as a parametric net on S; the first fundamental form has the form

$$(ds)^2 = \bar{a}_{11}(d\bar{u}^1)^2 + \bar{a}_{22}(d\bar{u}^2)^2:$$

Accordingly, Eq. (8.33) in the co-ordinate system $\bar{u}$ takes the form

$$\bar{b}_{12}\bar{a}_{11}(d\bar{u}^1)^2 + (\bar{b}_{11}\bar{a}_{22} \quad \bar{b}_{22}\bar{a}_{11})d\bar{u}^1 d\bar{u}^2 + \bar{b}_{12}\bar{a}_{22}(d\bar{u}^2)^2 = 0$$

and its solutions are

$$\bar{u}^1 = \text{ constant, } \bar{u}^2 = \text{ constant.}$$

If we take $d\bar{u}^1 \neq 0$ and $d\bar{u}^2 = 0$; we see that $\bar{b}_{12} = 0$; since $\bar{a}_{11} = 0$: Thus, a necessary condition for the net of lines of curvature to be orthogonal is that $a_{12} = 0 = b_{12}$: Consequently we may always choose co-ordinates $u^1; u^2$ on S so that the lines of curvature are the co-ordinate curves of this system which is allowable at any point of S which is not umbilic. Conversely, if $a_{12} = 0 = b_{12}$, then Eq. (8.33) has the solutions u^1 = constant and u^2 = constant, so that the co-ordinate lines are the lines of curvature.

Note that, for every orthogonal net on a plane or a sphere $a_{12} = 0 = b_{12}$: Formula (8.14), for the normal curvature, when the co-ordinate system is taken to be the net of lines of curvature becomes

$$_{(n)} = \frac{b_{11}(du^1)^2 + b_{22}(du^2)^2}{a_{11}(du^1)^2 + a_{22}(du^2)^2}:$$

If we set $du^1 = 0; du^2 \neq 0$; and $du^2 = 0; du^1 \neq 0$, we get

$$_{(1)} = \frac{b_{11}}{a_{11}}; \quad _{(2)} = \frac{b_{22}}{a_{22}}$$

for the curvatures of the co-ordinate lines u^1 = constant and u^2 = constant. The lines of curvature on S should not be confused with the normal sections of S. The normal sections are necessarily plane curves, whereas the lines of curvature ordinarily are not plane curves. Thus the lines of curvature of any (real) surface S of class 3 are real curves.

Deduction 8.3.2 A point on a surface is called umbilic if at that point $b = a$, where is a scalar. In particular, if = 0, the point is called a planer point. Note that, at an umbilic point the normal curvature $_{(n)}$ is the same for all directions. If S has no umbilics the lines of curvature form an orthogonal net everywhere on S.

Theorem 8.3.1 (Rodrigue's formula): A line of curvature is characterised by

$$\frac{\partial^i}{\partial s} + {}_{(n)}\frac{dx^i}{ds} = 0;$$

where $_{(n)}$ is the principal curvature of the surface.

Proof: From Weingarten formula Eq. (7.49), we have

$$^i_{;} = a \quad b \quad x^i$$

or

$$^i_{;}\frac{du}{ds} = a \quad b \quad x^i \frac{du}{ds}$$

or

$$\frac{\partial^i}{\partial s} = a \quad b \quad x^i \quad : \tag{8.35}$$

Again, we know that for a line of curvature,

$$b = {}_{(n)}a \quad ; \text{i.e. } b = {}_{(n)}a \quad ;$$

as a and b are symmetric. Using result (8.35), we get

$$\begin{aligned}\frac{\partial^i}{\partial s} &= a \quad {}_{(n)}a \quad x^i \\ &= {}_{(n)} \quad x^i = {}_{(n)} \quad x^i \\ &= {}_{(n)}\frac{\partial u}{\partial s}\frac{\partial x^i}{\partial u} = {}_{(n)}\frac{dx^i}{ds}\end{aligned}$$

or

$$\frac{\partial^i}{\partial s} + {}_{(n)}\frac{dx^i}{ds} = 0:$$

Hence, the theorem is proved. Conversely, let the relation

$$\frac{\partial^i}{\partial s} + {}_{(n)}\frac{dx^i}{ds} = 0$$

holds. Using Weingarten formula Eq. (7.49), we have

$$\frac{\partial^i}{\partial s} = \quad a \quad b \quad x^i$$

or

$$a \quad b \quad x^i \quad + {}_{(n)}\frac{dx^i}{ds} = 0:$$

Taking the inner product with $g_{ik}x^k_p$, we get

$$a \quad b \quad g_{ik}x^k_p x^i \quad + {}_{(n)}g_{ik}x^k_p\frac{dx^i}{ds} = 0$$

or

$$a \quad b \quad a_p \quad + {}_{(n)}g_{ki}x^i_p\frac{dx^k}{ds} = 0:$$

Interchanging the dummy indices i and k in the second term, we get

$$b_p \quad + {}_{(n)}g_{ik}x^i_p\frac{\partial x^k}{\partial u}\frac{du}{ds} = 0$$

or

$$b_p \quad + {}_{(n)}g_{ik}x^i_p x^k \quad = 0$$

or

$$b_p \quad + {}_{(n)}a_p \quad = 0$$

or

$${}_{(n)}a_p \quad b_p \quad = 0;$$

which is the equation of the line of curvature. The Rodrigues formulae are characteristic for the lines of curvature.

EXAMPLE 8.3.1 Show that the parametric curves are the lines of curvature if and only if $a_{12} = 0 = b_{12}$:

Solution: Let us assume that the parametric curves be the lines of curvature. Then they form an orthogonal net and will satisfy Eq. (8.33). Now, for u^1 curve, $du^2 = 0$ and hence from Eq. (8.33), we get

$$b_{11}a_{12} - a_{11}b_{12} = 0; \quad \text{as } du^1 \neq 0$$

and for u^2 curve, $du^1 = 0$ and hence from Eq. (8.33), we get

$$b_{12}a_{22} - a_{12}b_{22} = 0; \quad \text{as } du^2 \neq 0.$$

Multiplying first equation by $-a_{22}$ and second by a_{11} and adding, we get

$$a_{12}(a_{22}b_{11} - a_{11}b_{22}) = 0$$

$$\Rightarrow \quad \text{either } a_{12} = 0 \text{ or } \frac{a_{11}}{b_{11}} = \frac{a_{22}}{b_{22}}.$$

For the parametric curve, $a_{22}b_{11} \neq a_{11}b_{22}$ and so $a_{12} = 0$. The condition $a_{12} = 0$ is that of orthogonality satisfied by all lines of curvature. Similarly, multiplying first equation by b_{22} and second by b_{11} and adding, we get

$$b_{12}(a_{22}b_{11} - a_{11}b_{22}) = 0 \Rightarrow b_{12} = 0.$$

Thus $b_{12} = 0$, is the necessary and sufficient condition that the parametric curves form a conjugate system. Conversely, let $a_{12} = 0 = b_{12}$, then Eq. (8.33) reduces to

$$(a_{22}b_{11} - a_{11}b_{22})\, du^1 du^2 = 0.$$

If the lines of curvature exist, then $a_{22}b_{11} - a_{11}b_{22} \neq 0$ and the curves are the solution of the equation $du^1 du^2 = 0$. Hence,

$$\text{when} \quad du^1 \neq 0; \quad du^2 = 0 \Rightarrow u^2 = \text{constant}.$$

$$\text{when} \quad du^2 \neq 0; \quad du^1 = 0 \Rightarrow u^1 = \text{constant}.$$

Hence, the parametric curves are the lines of curvature. Thus, if in particular the co-ordinates are chosen so that the co-ordinate curves are lines of curvature on the surface S then $a_{12} = b_{12} = 0$ and therefore the Rodrigues formulae becomes

$$\frac{\partial^i}{\partial S} = -\frac{b_i}{a_{ii}} \frac{dx^i}{ds} \text{(no summation)}; \; i = 1, 2.$$

EXAMPLE 8.3.2 Show that the lines of curvature on a minimal surface form an isometric systems.

Solution: If the parametric curves are the lines of curvature, then $a_{12} = 0 = b_{12}$. Now, for the minimal surface,

$$H = \frac{1}{2} a^{\alpha\beta} b_{\alpha\beta} = 0$$

$$\Rightarrow H = \frac{1}{2a}[a_{22}b_{11} + a_{11}b_{22}] = 0; \quad \text{as } a_{12} = 0 = b_{12}$$

$$\Rightarrow b_{11} = b_{22} = 0; \quad \text{as } a \neq 0$$

$$\Rightarrow b_{\alpha\beta} = 0.$$

Therefore, the surface is a plane, i.e. the surface is isomorphic with Euclidean plane.

Theorem 8.3.2 (Euler's theorem on normal curvature): If the lines of curvature are not indeterminant at a given point P on the surface and if θ is the angle between a given direction and a principal direction at P, then the normal curvature is given by the formula

$$\kappa_{(n)} = \kappa_{(1)} \cos^2\theta + \kappa_{(2)} \sin^2\theta.$$

Proof: We assume that P is not an umbilic. If the parametric curves are taken as the lines of curvature, then the principal curvatures are given by

$$\kappa_{(1)} = \frac{b_{11}}{a_{11}} \text{ and } \kappa_{(2)} = \frac{b_{22}}{a_{22}}.$$

Let θ be the angle between a given direction (u^1, u^2) and a principal direction at a given point P. Since the co-ordinate curves are orthogonal, we have

$$\cos\theta = \sqrt{a_{11}}\frac{du^1}{ds} \text{ and } \sin\theta = \sqrt{a_{22}}\frac{du^2}{ds}.$$

Also, let $\kappa_{(1)}, \kappa_{(2)}$ be the principal curvatures at P. If the parametric curves are taken as lines of curvature, then the normal curvature at P is given by

$$\kappa_{(n)} = \frac{b_{11}(du^1)^2 + b_{22}(du^2)^2}{ds^2} = b_{11}\left(\frac{du^1}{ds}\right)^2 + b_{22}\left(\frac{du^2}{ds}\right)^2$$

$$= \frac{b_{11}}{a_{11}}\cos^2\theta + \frac{b_{22}}{a_{22}}\sin^2\theta = \kappa_{(1)}\cos^2\theta + \kappa_{(2)}\sin^2\theta.$$

This is Euler's theorem on normal curvature. This theorem tells us that, the normal curvature corresponding to any direction can be simply represented in terms of the principal curvatures. The Gaussian and total curvature in this case is given by

$$K = \frac{b_{11}}{a_{11}}\frac{b_{22}}{a_{22}} \text{ and } H = \frac{1}{2}\left(\frac{b_{11}}{a_{11}} + \frac{b_{22}}{a_{22}}\right).$$

From this we conclude the lines of curvature on a minimal surface form an isometric system.

Deduction 8.3.3 Let $_{(n_1)}$ and $_{(n_2)}$ denote normal curvatures in two orthogonal directions on the surface and be the angle between the first direction $du^2 = 0$; thus, the angle between the second direction and the principal direction $du^1 = 0$ will be $\frac{}{2} +$. Thus, from Euler's theorem, we have

$$_{(n_1)} = \ _{(1)} \cos^2 \ + \ _{(2)} \sin^2 \ ;$$

$$_{(n_2)} = \ _{(1)} \cos^2 \ \frac{}{2} + \ \ + \ _{(2)} \sin^2 \ \frac{}{2} +$$

$$= \ _{(1)} \sin^2 \ + \ _{(2)} \cos^2 \ :$$

Now, adding, we get

$$_{(n_1)} + \ _{(n_2)} = \ _{(1)} + \ _{(2)} :$$

Thus, the sum of the normal curvatures in two orthogonal directions is equal to the sum of the principal curvatures at that point. This is known as Dupin's theorem.

Result: The theorem of enter and the theorem of Meusnier give complete information on the curvature of any curve on a surface.

EXAMPLE 8.3.3 Find the equations for the lines of curvature for the surface given by $r = (u \cos v; u \sin v; cv)$:

Solution: For the given surface, the symmetric covariant second order tensors a for the first fundamental form are given by

$$a_{11} = 1; \ a_{22} = u^2 + c^2; \ a_{12} = 0 = a_{21};$$

and the second fundamental magnitudes b are given by

$$b_{11} = 0; \ b_{22} = 0; \ b_{12} = \ \frac{c}{\sqrt{u^2 + c^2}} = b_{21}:$$

Since $a_{12} = 0 = b_{12}$, this given parametric curve has the line of curvature. The lines of curvature are given by

$$(b_{11}a_{12} \quad b_{12}a_{11})(du)^2 + (b_{11}a_{22} \quad b_{22}a_{11})du \ dv$$
$$+(b_{12}a_{22} \quad b_{22}a_{12})(dv)^2 = 0$$

or

$$b_{12}a_{11}(du)^2 + b_{12}a_{22})(dv)^2 = 0$$

or

$$\frac{c}{\sqrt{u^2+c^2}}(du)^2 - \frac{c}{\sqrt{u^2+c^2}}(u^2+c^2)(dv)^2 = 0$$

or

$$\frac{du}{\sqrt{u^2+c^2}} = \pm dv \Rightarrow v = \pm \sinh^{-1}\frac{u}{c} + k,$$

where k is a constant.

EXAMPLE 8.3.4 Given an ellipsoid of revolution, whose surface is determined by

$$x^1 = a\cos u^1 \sin u^2;\quad x^2 = a \sin u^1 \sin u^2; x^3 = c\cos u^2$$

where $a^2 > c^2$. Find $a_{\alpha\beta}$, $b_{\alpha\beta}$ and $\kappa = \kappa_{(1)}\kappa_{(2)}$. Discuss the lines of curvature.

Solution: The parametric representation of the ellipsoid of revolution is given by

$$x^1 = a\cos u^1 \sin u^2; x^2 = a\sin u^1 \sin u^2; x^3 = c\cos u^2$$

i.e.

$$\mathbf{r} = (x^1, x^2, x^3) = \left(a\cos u^1 \sin u^2, a\sin u^1 \sin u^2, c\cos u^2\right).$$

The symmetric covariant second order tensors $a_{\alpha\beta}$ for the first fundamental form of the surface of the ellipsoid of revolution are given by

$$a_{11} = \left(\frac{\partial x^1}{\partial u^1}\right)^2 + \left(\frac{\partial x^2}{\partial u^1}\right)^2 + \left(\frac{\partial x^3}{\partial u^1}\right)^2$$

$$= (-a\sin u^1 \sin u^2)^2 + (a\cos u^1 \sin u^2)^2 + 0^2 = a^2 \sin^2 u^2.$$

$$a_{22} = \left(\frac{\partial x^1}{\partial u^2}\right)^2 + \left(\frac{\partial x^2}{\partial u^2}\right)^2 + \left(\frac{\partial x^3}{\partial u^2}\right)^2$$

$$= (a\cos u^1 \cos u^2)^2 + (a\sin u^1 \cos u^2)^2 + (-c\sin u^2)^2$$

$$= a^2\cos^2 u^2 + c^2 \sin^2 u^2.$$

$$a_{12} = \frac{\partial x^1}{\partial u^1}\frac{\partial x^1}{\partial u^2} + \frac{\partial x^2}{\partial u^1}\frac{\partial x^2}{\partial u^2} + \frac{\partial x^3}{\partial u^1}\frac{\partial x^3}{\partial u^2} = 0 = a_{21}.$$

Therefore,

$$a = \begin{vmatrix} a^2\sin^2 u^2 & 0 \\ 0 & a^2\cos^2 u^2 + c^2\sin^2 u^2 \end{vmatrix} = a^2 \sin^2 u^2 (a^2\cos^2 u^2 + c^2 \sin^2 u^2).$$

To calculate the tensors of second fundamental form, we have

$$\mathbf{A} = (-a\sin u^1 \sin u^2, a\cos u^1 \sin u^2, 0) \text{ and}$$

$$\mathbf{B} = (a\cos u^1 \cos u^2, a\sin u^1 \cos u^2, -c\sin u^2)$$

$$\mathbf{A} \times \mathbf{B} = \left(-ac\cos u^1 \sin u^2, ac\sin u^1 \sin u^2, -a^2 \sin u^2 \cos u^2\right).$$

The normal vector is given by

$$= \frac{1}{\sqrt{a^2\cos^2 u^2 + c^2\sin^2 u^2}} \quad ac\cos u^1 \sin u^2; ac\sin u^1 \sin u^2; \quad a^2 \sin u^2 \cos u^2 \; :$$

Thus, the covariant tensors b for the second fundamental form are given by

$$b_{11} = \frac{ac\sin^2 u^2}{\sqrt{a^2\cos^2 u^2 + c^2\sin^2 u^2}}; b_{22} = \frac{ac}{\sqrt{a^2\cos^2 u^2 + c^2\sin^2 u^2}}; \; b_{12} = 0:$$

$$) \quad b = b_{11}b_{22} \quad b_{12}^2 = \frac{a^2c^2\sin^2 u^2}{a^2\cos^2 u^2 + c^2\sin^2 u^2}:$$

Therefore, the Gaussian curvature is given by

$$= \frac{b}{a} = \frac{a^2c^2\sin^2 u^2}{a^2\cos^2 u^2 + c^2\sin^2 u^2} \quad \frac{1}{a^2\sin^2 u^2(a^2\cos^2 u^2 + c^2\sin^2 u^2)}$$

$$= \frac{c^2}{a^2\cos^2 u^2 + c^2\sin^2 u^2 \;\; ^2}:$$

Since $a_{12} = 0; b_{12} = 0$; so the equation of the lines of curvature is given by

$$(b_{11}a_{22} \quad b_{22}a_{11})\, du^1 du^2 = 0;$$

$$(a^2 \quad c^2)\sin^4 u^2 du^1 du^2 = 0:$$

EXAMPLE 8.3.5 Show that any curve on a sphere is a line of curvature.

Solution: The parametric representation of the sphere of radius a is given by

$$x^1 = a\sin u^1 \cos u^2; \quad x^2 = a\sin u^1 \sin u^2; x^3 = a\cos u^2$$

i.e.

$$r = (x^1; x^2; x^3) = \quad a\sin u^1 \cos u^2; a\sin u^1 \sin u^2; a\cos u^2 \; :$$

For the given surface, the symmetric covariant second order tensors a for the first fundamental form are given by

$$a_{11} = a^2; \; a_{22} = a^2\sin^2 u^1; \; a_{12} = 0 = a_{21}$$

and the symmetric covariant second order tensors b for the second fundamental form are given by

$$b_{11} = \quad a; \; b_{22} = \quad a\sin^2 u^1; \; b_{12} = 0 = b_{21}:$$

Since $a_{12} = 0 = b_{12}$, this given parametric curve has the line of curvature. In fact,

$$\frac{b_{11}}{a_{11}} = \frac{b_{12}}{a_{12}} = \frac{b_{22}}{a_{22}} = -\frac{1}{a},$$

i.e. $b_{\alpha\beta}$ are proportional to $a_{\alpha\beta}$. Now,

$$\begin{aligned}&(b_{11}a_{12} - b_{12}a_{11})(du^1)^2 + (b_{11}a_{22} - b_{22}a_{11})du^1du^2\\ &\quad + (b_{12}a_{22} - b_{22}a_{12})(du^2)^2 = \left[-a^3\sin^2 u^1 + a^3\sin^2 u^1\right] du^1du^2 = 0.\end{aligned}$$

This shows that the equation of the lines of curvature reduces to an identity. Hence, any curve on a sphere is a line of curvature.

EXAMPLE 8.3.6 Given a surface of revolution S

$$x^1 = u^1\cos u^2,\; x^2 = u^1\sin u^2,\; x^3 = f(u^1)$$

with $f(u^1)$ of class C^2. Prove that the lines of curvature on S are the meridians $u^2 =$ constant and the parallels $u^1 =$ constant.

Solution: The parametric representation of the right helicoid is given by

$$x^1 = u^1\cos u^2,\; x^2 = u^1\sin u^2,\; x^3 = f(u^1)$$

i.e.

$$\vec{r} = (x^1, x^2, x^3) = \left(u^1\cos u^2,\; u^1\sin u^2,\; f(u^1)\right).$$

The symmetric covariant second order tensors $a_{\alpha\beta}$ for the first fundamental form of the surface are given by

$$a_{11} = 1 + f_1^2;\quad a_{22} = (u^1)^2 \text{ and } a_{12} = 0 = a_{21}$$

$$\Rightarrow\quad a = \begin{vmatrix} 1 + f_1^2 & 0 \\ 0 & (u^1)^2 \end{vmatrix} = (u^1)^2(1 + f_1^2),$$

where $f_1 = \frac{\partial f}{\partial u^1}$. The symmetric covariant tensors $b_{\alpha\beta}$ for the second fundamental form are given by

$$b_{11} = \frac{f_{11}}{\sqrt{1 + f_1^2}},\; b_{22} = \frac{u^1 f_1}{\sqrt{1 + f_1^2}} \text{ and } b_{12} = 0 = b_{21}.$$

$$\Rightarrow |b_{\alpha\beta}| = \begin{vmatrix} \frac{f_{11}}{\sqrt{1 + f_1^2}} & 0 \\ 0 & \frac{u^1 f_1}{\sqrt{1 + f_1^2}} \end{vmatrix} = \frac{u^1 f_1 f_{11}}{1 + f_1^2},$$

where $f_1 = \frac{\partial f}{\partial u^1}$ and $f_{11} = \frac{\partial^2 f}{\partial u^1 \partial u^1}$. Equation (8.33) giving the principal curvatures is

$$(b_{11}a_{12} - b_{12}a_{11})(du^1)^2 + (b_{11}a_{22} - b_{22}a_{11})du^1du^2$$
$$+(b_{12}a_{22} - b_{22}a_{12})(du^2)^2 = 0$$

or

$$\left[-\left(1 + f_1^2\right)\frac{u^1 f_1}{\sqrt{1 + f_1^2}} + \left(u^1\right)^2 \frac{f_{11}}{\sqrt{1 + f_1^2}}\right] du^1du^2 = 0$$

or

$$du^1du^2 = 0 \Rightarrow u^2 = \text{constant, or, } u^1 = \text{constant}.$$

EXAMPLE 8.3.7 Find the differential equation of lines of curvature of the helicoid

$$x^1 = u^1 \cos u^2;\; x^2 = u^1 \sin u^2;\; x^3 = f(u^1) + cu^2,$$

where u^1, u^2 are parameters.

Solution: The parametric representation of the surface is given by

$$\mathbf{r} = (x^1, x^2, x^3) = \left(u^1 \cos u^2,\; u^1 \sin u^2,\; f(u^1) + cu^2\right).$$

The first fundamental magnitudes $a_{\alpha\beta}$ are given by

$$a_{11} = \left(\frac{\partial x^1}{\partial u^1}\right)^2 + \left(\frac{\partial x^2}{\partial u^1}\right)^2 + \left(\frac{\partial x^3}{\partial u^1}\right)^2 = 1 + f_1^2;$$
$$a_{22} = \left(\frac{\partial x^1}{\partial u^2}\right)^2 + \left(\frac{\partial x^2}{\partial u^2}\right)^2 + \left(\frac{\partial x^3}{\partial u^2}\right)^2 = (u^1)^2 + c^2;$$
$$a_{12} = \frac{\partial x^1}{\partial u^1}\frac{\partial x^1}{\partial u^2} + \frac{\partial x^2}{\partial u^1}\frac{\partial x^2}{\partial u^2} + \frac{\partial x^3}{\partial u^1}\frac{\partial x^3}{\partial u^2} = cf_1 = a_{21},$$

where $f_1 = \frac{\partial f}{\partial u^1}$. Therefore,

$$a = \begin{vmatrix} 1 + f_1^2 & cf_1 \\ cf_1 & (u^1)^2 + c^2 \end{vmatrix} = (u^1)^2(1 + f_1^2) + c^2.$$

The reciprocal tensors are given by

$$a^{11} = \frac{(u^1)^2 + c^2}{(u^1)^2(1 + f_1^2) + c^2};\; a^{22} = \frac{1 + f_1^2}{(u^1)^2(1 + f_1^2) + c^2};$$
$$a^{12} = \frac{-cf_1}{(u^1)^2(1 + f_1^2) + c^2} = a^{21}.$$

Since $r = (x^1; x^2; x^3) = \quad u^1 \cos u^2; \quad u^1 \sin u^2; \quad f(u^1) + cu^2$, we have

$$A = (\cos u^2; \sin u^2; f_1) \text{ and } B = (\quad u^1 \sin u^2; u^1 \cos u^2; c)$$

$$A \quad B = (c \sin u^2 \quad u^1 f_1 \cos u^2)\hat{\imath} + (\quad c \cos u^2 \quad u^1 f_1 \sin u^2)\hat{\jmath} + u^1\hat{k}:$$

Therefore, the unit normal vector is given by

$$= \frac{1}{\sqrt{(u^1)^2(1+f_1^2)+c^2}} \quad c \sin u^2 \quad u^1 f_1 \cos u^2; \quad c \cos u^2 \quad u^1 f_1 \sin u^2; u^1 \quad :$$

The symmetric covariant tensors b are given by

$$b_{11} = \frac{u^1 f_2}{\sqrt{(u^1)^2(1+f_1^2)+c^2}}; b_{22} = \frac{(u^1)^2 f_1}{\sqrt{(u^1)^2(1+f_1^2)+c^2}}; b_{12} = \frac{c}{\sqrt{(u^1)^2(1+f_1^2)+c^2}};$$

where $f_2 = \frac{\partial^2 f}{(\partial u^1)^2}$: Therefore,

$$b = \begin{matrix} \frac{u^1 f_2}{\sqrt{(u^1)^2(1+f_1^2)+c^2}} & \frac{c}{\sqrt{(u^1)^2(1+f_1^2)+c^2}} \\ \frac{c}{\sqrt{(u^1)^2(1+f_1^2)+c^2}} & \frac{(u^1)^2 f_1}{\sqrt{(u^1)^2(1+f_1^2)+c^2}} \end{matrix} = \frac{(u^1)^3 f_1 f_2 \quad c^2}{(u^1)^2(1+f_1^2)+c^2}:$$

The equation giving the principal curvatures is

$${}^2 \quad a \quad b \qquad + \frac{b}{a} = 0$$

or

$${}^2 + \quad '' \frac{(1+f_1^2)(u^1)^2 f_1 + 2c^2 f_1 + ((u^1)^2 + c^2)u^1 f_2}{(u^1)^2(1+f_1^2)+c^2 \quad ^{3=2}} \# \qquad + \frac{(u^1)^3 f_1 f_2 \quad c^2}{(u^1)^2(1+f_1^2)+c^2 \quad ^2} = 0:$$

This is a quadratic equation in giving the two values $_{(1)}$ and $_{(2)}$. The mean curvature H is given by

$$H = \frac{1}{2}[\; _{(1)} + \; _{(2)}] = \frac{1}{2} \quad '' \frac{(1+f_1^2)(u^1)^2 f_1 + 2c^2 f_1 + ((u^1)^2 + c^2)u^1 f_2}{(u^1)^2(1+f_1^2)+c^2 \quad ^{3=2}} \#$$

and the Gaussian curvature is given by the formula,

$$= \; _{(1)} \; _{(2)} = \frac{(u^1)^3 f_1 f_2 \quad c^2}{(u^1)^2(1+f_1^2)+c^2 \quad ^2}:$$

From this relation, we see that, along the helix u^1 = constant, the value of is constant and is independent of parameter u^2. The differential equation giving the directions of the lines of curvatures is

$$(b_{11}a_{12} - b_{12}a_{11})(du^1)^2 + (b_{11}a_{22} - b_{22}a_{11})du^1du^2 + (b_{12}a_{22} - b_{22}a_{12})(du^2)^2 = 0$$

or

$$c[u^1 f_1 f_2 + 1 + f_1^2](du^1)^2 + [u^1 f_2 f(u^1)^2 + c^2 g - (u^1)^2 f_1(1 + f_1^2)]du^1du^2 - c[(u^1)^2(1 + f_1^2) + c^2](du^2)^2 = 0.$$

If the meridians are lines of curvature, i.e. curves u^2 = constant are lines of curvature then $du^2 = 0$ and so, we get

$$c[u^1 f_1 f_2 + 1 + f_1^2](du^1)^2 = 0$$

or

$$u^1 f_1 f_2 + 1 + f_1^2 = 0 \text{ as } c \neq 0;\ du^1 \neq 0;$$

which is the required condition for the meridians to be lines of curvatures.

EXAMPLE 8.3.8 Prove that the line of curvature of the paraboloid $x^1x^2 = ax^3$ lie on the surface

$$\sinh^{-1}\frac{x^1}{a} - \sinh^{-1}\frac{x^2}{a} = c;$$

where c is an constant. Hence find the equation of the cone passes through a line of curvature of the paraboloid $x^1x^2 = ax^3$.

Solution: If the parameter x^3 can be written in the form

$$x^3 = f(x^1; x^2) = \frac{x^1x^2}{a}$$

then the parametric equation of the paraboloid $x^1x^2 = ax^3$ can be written in the form

$$r = (x^1; x^2; x^3) = \{x^1; x^2; f(x^1; x^2)\}:$$

The first fundamental magnitudes a are given by

$$a_{11} = \left(\frac{\partial x^1}{\partial u^1}\right)^2 + \left(\frac{\partial x^2}{\partial u^1}\right)^2 + \left(\frac{\partial x^3}{\partial u^1}\right)^2 = 1 + f_1^2;$$

$$a_{22} = \left(\frac{\partial x^1}{\partial u^2}\right)^2 + \left(\frac{\partial x^2}{\partial u^2}\right)^2 + \left(\frac{\partial x^3}{\partial u^2}\right)^2 = 1 + f_2^2;$$

$$a_{12} = \frac{\partial x^1}{\partial u^1}\frac{\partial x^1}{\partial u^2} + \frac{\partial x^2}{\partial u^1}\frac{\partial x^2}{\partial u^2} + \frac{\partial x^3}{\partial u^1}\frac{\partial x^3}{\partial u^2} = f_1 f_2 = a_{21};$$

where $f_1 = \frac{\partial f}{\partial u^1}$ and $f_2 = \frac{\partial f}{\partial u^2}$. Therefore,

$$a = \begin{vmatrix} 1+f_1^2 & f_1 f_2 \\ f_1 f_2 & 1+f_2^2 \end{vmatrix} = 1 + f_1^2 + f_2^2.$$

Since $r = \left(x^1, x^2, f(x^1, x^2)\right)$, we have $A = (1, 0, f_1)$ and $B = (0, 1, f_2)$

$$A \times B = -f_1 \hat{i} - f_2 \hat{j} + \hat{k}.$$

Therefore, the unit normal vector ξ is given by

$$\xi = \frac{1}{\sqrt{1+f_1^2+f_2^2}}(-f_1, -f_2, 1).$$

The symmetric covariant tensors $b_{\alpha\beta}$ are given by

$$b_{11} = \frac{f_{11}}{\sqrt{1+f_1^2+f_2^2}}; \quad b_{22} = \frac{f_{22}}{\sqrt{1+f_1^2+f_2^2}}; \quad b_{12} = \frac{f_{12}}{\sqrt{1+f_1^2+f_2^2}};$$

where $f_{ij} = \frac{\partial^2 f}{\partial u^i \partial u^j}$. Therefore,

$$b = \begin{vmatrix} \frac{f_{11}}{\sqrt{1+f_1^2+f_2^2}} & \frac{f_{12}}{\sqrt{1+f_1^2+f_2^2}} \\ \frac{f_{12}}{\sqrt{1+f_1^2+f_2^2}} & \frac{f_{22}}{\sqrt{1+f_1^2+f_2^2}} \end{vmatrix} = \frac{f_{11} f_{22} - f_{12}^2}{1+f_1^2+f_2^2}.$$

The differential equation giving the directions of the lines of curvatures is

$$(b_{11}a_{12} - b_{12}a_{11})(dx^1)^2 + (b_{11}a_{22} - b_{22}a_{11})dx^1 dx^2 + (b_{12}a_{22} - b_{22}a_{12})(dx^2)^2 = 0$$

or

$$\frac{1}{a}\left(1 + \frac{(x^2)^2}{a^2}\right)(dx^1)^2 - \frac{1}{a}\left(1 + \frac{(x^1)^2}{a^2}\right)(dx^2)^2 = 0$$

or

$$\frac{dx^1}{\sqrt{a^2 + (x^1)^2}} - \frac{dx^2}{\sqrt{a^2 + (x^2)^2}} = 0$$

or

$$\sinh^{-1}\frac{x^1}{a} - \sinh^{-1}\frac{x^2}{a} = c;$$

where c is a constant. This equation can be written as

$$\sinh^{-1}\left[\frac{x^1}{a}\sqrt{1 + \frac{(x^2)^2}{a^2}} - \frac{x^2}{a}\sqrt{1 + \frac{(x^1)^2}{a^2}}\right] = c$$

or

$$\frac{x^1}{a}\sqrt{1+\frac{(x^2)^2}{a^2}} - \frac{x^2}{a}\sqrt{1+\frac{(x^1)^2}{a^2}} = \sinh c = p;\ \text{(say)}$$

or

$$\frac{x^3}{x^2}\sqrt{1+\frac{(x^3)^2}{(x^1)^2}} - \frac{x^3}{x^1}\sqrt{1+\frac{(x^3)^2}{(x^2)^2}} = p$$

or

$$x^3\left[\sqrt{(x^1)^2+(x^3)^2} - \sqrt{(x^2)^2+(x^3)^2}\right] = px^1x^2.$$

This is a homogeneous second degree equation and hence represents a cone passes through the paraboloid $x^1x^2 = ax^3$.

EXAMPLE 8.3.9 Find the principal directions and the principal curvatures on the surface

$$x^1 = a(u^1+u^2);\ x^2 = b(u^1-u^2);\ x^3 = u^1u^2.$$

Solution: The parametric representation of the surface is given by

$$\vec{r} = (x^1; x^2; x^3) = \left(a(u^1+u^2);\ b(u^1-u^2);\ u^1u^2\right).$$

The first fundamental magnitudes $a_{\alpha\beta}$ are given by

$$a_{11} = \left(\frac{\partial x^1}{\partial u^1}\right)^2 + \left(\frac{\partial x^2}{\partial u^1}\right)^2 + \left(\frac{\partial x^3}{\partial u^1}\right)^2 = a^2+b^2+\left(u^2\right)^2.$$

$$a_{22} = \left(\frac{\partial x^1}{\partial u^2}\right)^2 + \left(\frac{\partial x^2}{\partial u^2}\right)^2 + \left(\frac{\partial x^3}{\partial u^2}\right)^2 = a^2+b^2+\left(u^1\right)^2.$$

$$a_{12} = \frac{\partial x^1}{\partial u^1}\frac{\partial x^1}{\partial u^2} + \frac{\partial x^2}{\partial u^1}\frac{\partial x^2}{\partial u^2} + \frac{\partial x^3}{\partial u^1}\frac{\partial x^3}{\partial u^2} = a^2-b^2+u^1u^2 = a_{21}.$$

Therefore,

$$\Delta = 4a^2b^2 + a^2\left(u^1-u^2\right)^2 + b^2(u^1+u^2)^2.$$

The reciprocal tensors are given by

$$a^{11} = \frac{a^2+b^2+\left(u^1\right)^2}{\Delta};\quad a^{22} = \frac{a^2+b^2+\left(u^2\right)^2}{\Delta};\quad a^{12} = -\frac{a^2-b^2+u^1u^2}{\Delta} = a^{21}.$$

Since $\vec{r} = (x^1; x^2; x^3) = \left(a(u^1+u^2);\ b(u^1-u^2);\ u^1u^2\right)$, we have

$$\vec{A} = (a; b; u^2) \text{ and } \vec{B} = (a; -b; u^1)$$

$$\vec{A}\times\vec{B} = \left(b(u^1+u^2); a(u^2-u^1); -2ab\right).$$

Therefore, the unit normal vector   is given by

$$= \frac{1}{p\overline{4}} \quad b(u^1 + u^2);\; a(u^2 \quad u^1); \quad 2ab \;:$$

The symmetric covariant tensors b   are given by

$$b_{11} = \frac{@}{@u^1}\;\frac{@}{@u^1} = 0;\quad b_{22} = \frac{@}{@u^2}\;\frac{@}{@u^2} = 0;\quad b_{12} = \frac{2ab}{p\overline{4}}:$$

$$\text{Therefore, } b = \frac{4a^2b^2}{4}:$$

The equation giving the principal curvatures is

$$^2 \quad a \quad b \qquad + \frac{b}{4} = 0$$

or

$$^2 \quad \frac{4ab}{4^{3=2}}\; a^2 \quad b^2 + u^1u^2 \qquad \frac{4a^2b^2}{4^{\;2}} = 0:$$

This is a quadratic equation in   giving the two values $_{(1)}$ and $_{(2)}$. The rst curvature H is given by,

$$H = \frac{1}{2}[\;_{(1)} + \;_{(2)}] = \frac{2ab}{4^{3=2}}\; a^2 \quad b^2 + u^1u^2$$

and the speci c curvature   is given by the formula,

$$= \;_{(1)}\;_{(2)} = \frac{4a^2b^2}{4^{\;2}}:$$

The di erential equation giving the directions of the lines of curvatures is

$$(b_{11}a_{12} \quad b_{12}a_{11})(du^1)^2 + (\,b_{11}a_{22} \quad b_{22}a_{11})du^1du^2$$
$$+(\,b_{12}a_{22} \quad b_{22}a_{12})(du^2)^2 = 0$$

or

$$h\; a^2 + b^2 + \quad u^2\;{}^2\; i\; (du^1)^2 \qquad h\; a^2 + b^2 + \quad u^1\;{}^2\; i\; (du^2)^2 = 0$$

or

$$\frac{du^1}{q\overline{a^2 + b^2 + (\,u^1)^2}} = \frac{du^2}{q\overline{a^2 + b^2 + (\,u^2)^2}}$$

or

$$\log \quad u^1 + \quad q\overline{a^2 + b^2 + (\,u^1)^2} \quad = \quad \log \quad u^2 + \quad q\overline{a^2 + b^2 + (\,u^2)^2} \quad + C:$$

8.3.1 Null Lines

The null lines (or minimal curves) on a surface are defined as the curves of zero length. Therefore, they are imaginary on a real surface, and their importance is chiefly analytic. The differential equation of the null lines is obtained by equating to zero the square of the line element. It is therefore,

$$a_{11}du^2 + 2a_{12}dudv + a_{22}dv^2 = 0: \tag{8.36}$$

If the parametric curves are null lines, this equation must be equivalent to $dudv = 0$: Hence $a_{11} = 0; a_{12} = 0$ and $a_{12} \neq 0$: These are the necessary and sufficient conditions that the parametric curves be null lines. In this case the square of the line element has the form $ds^2 = \ dudv$; where is a function of $u; v$ or a constant; and the parameters $u; v$ are then said to be symmetric.

When the parametric curves are null lines, so that $a_{11} = 0; a_{22} = 0$, $a = \ a_{12}^2$; the differential equation of the lines of curvature is

$$b_{11}du^2 \quad b_{22}dv^2 = 0:$$

Therefore, the Gauss curvature and the mean curvature H are, respectively, given by

$$= \frac{b_{11}b_{22} \quad b_{12}^2}{a_{12}^2}; \; H = \frac{2b_{12}}{a_{12}}:$$

8.3.2 Conjugate Directions

Let Q be a point on the surface adjacent to P, and let PR be the line of intersection of the tangent plane at P and Q. Then, as Q tends to coincidence with P, the limiting directions of PQ and PR are said to be conjugate directions at P. Thus, the characteristic of the tangent plane, as the point of contact moves along a given curve, is the tangent line in the direction conjugate to that of the curve at a point of contact.

At a point $P \ 2 \ S$, let two directions $(du^1; du^2)$ and $(\ u^1; \ u^2)$ satisfy the relation

$$b_{11}du^1 \ u^1 + b_{12} \ du^1 \ u^2 + du^2 \ u^1 \ + b_{22}du^2 \ u^2 = 0;$$

or

$$b \ \ du \ \ u \ = 0; \tag{8.37}$$

then these are called conjugate directions at P. Thus, the necessary and sufficient condition that the direction $\frac{u^1}{u^2}$ be conjugate to the direction $\frac{du^1}{du^2}$ is

$$b_{11}\frac{du^1}{du^2} \ \frac{u^1}{u^2} + b_{12} \ \frac{du^1}{du^2} + \frac{u^1}{u^2} \ + b_{22} = 0 \tag{8.38}$$

and the symmetry of the relation shows that the property is a reciprocal one. Moreover, Eq. (8.38) is linear in each of the ratios $du^1 : du^2$ and $\delta u^1 : \delta u^2$; so that to a given direction there is one and only one conjugate direction.

EXAMPLE 8.3.10 Find a necessary condition so that the two directions determined by $p_{\alpha\beta}\, du^\alpha du^\beta = 0;\ p_{\alpha\beta} = p_{\beta\alpha}$ are conjugate.

Solution: Given that, $p_{\alpha\beta}\, du^\alpha du^\beta = 0$; i.e.

$$p_{11}(du^1)^2 + 2\,p_{12}du^1du^2 + p_{22}(du^2)^2 = 0.$$

Let $(du^1; du^2)$ and $(\delta u^1; \delta u^2)$ be the solutions, then

$$\frac{du^1}{du^2} + \frac{\delta u^1}{\delta u^2} = -2\frac{p_{12}}{p_{11}} \text{ and } \frac{du^1}{du^2}.\frac{\delta u^1}{\delta u^2} = \frac{p_{22}}{p_{11}}.$$

From Eq. (8.37), we get

$$b_{11}\frac{p_{22}}{p_{11}} + b_{12}\left(-2\frac{p_{12}}{p_{11}}\right) + b_{22} = 0$$

or

$$b_{11}p_{22} - 2b_{12}p_{12} + b_{22}p_{11} = 0$$

or

$$b^{\alpha\beta}p_{\alpha\beta} = 0;$$

where $b^{\alpha\beta}$ are given by

$$b^{\alpha\beta} = \frac{\text{cofactor of } b_{\alpha\beta} \text{ in } |b_{\alpha\beta}|}{|b_{\alpha\beta}|}.$$

Theorem 8.3.3 Parametric curves have conjugate directions if and only if $b_{12} = 0$.

Proof: Curve conjugate to u^1 curve is given by Eq. (8.37) as

$$b_{11}du^1\,\delta u^1 + b_{12}du^1\,\delta u^2 = 0; \text{ as } du^2 = 0.$$

Since the conjugate of u^1 curve is the u^2 curve, $\delta u^1 = 0$ but $\delta u^2 \neq 0$. Therefore, $b_{12} = 0$ which is the necessary condition.

Conversely, if $b_{12} = 0$; then the curve conjugate of u^1 curve is given by

$$b_{11}\,\delta u^1 = 0 \Rightarrow \delta u^1 = 0;$$

which shows that it is the u^2 curve. Hence, the principal directions at a point of the surface are conjugate directions. Thus the null lines on a minimal surface form a conjugate system.

Deduction 8.3.4 Let the lines of curvature be taken as parametric curves, so that $a_{12} = 0$ and $b_{12} = 0$. The directions $\frac{du^1}{du^2}$ and $\frac{\delta u^1}{\delta u^2}$ are inclined to the curve u^2 = constant at angles θ, θ', such that

$$\tan\theta = \sqrt{\frac{a_{22}}{a_{11}}}\frac{du^1}{du^2}; \quad \tan\theta' = \sqrt{\frac{a_{22}}{a_{11}}}\frac{\delta u^1}{\delta u^2}.$$

The condition that two directions be conjugate may be expressed

$$b_{11}\frac{du^1}{du^2}\frac{\delta u^1}{\delta u^2} + b_{12}\left(\frac{du^1}{du^2} + \frac{\delta u^1}{\delta u^2}\right) + b_{22} = 0$$

or

$$\tan\theta\tan\theta' = -\frac{b_{11}}{a_{11}}\frac{a_{22}}{b_{22}} = -\frac{\kappa_1}{\kappa_2}; \text{ as } b_{12} = 0;$$

that is to say, provided they are parallel to conjugate diameters of the indicatrix.

EXAMPLE 8.3.11 Show that the parametric curves are conjugate on the surface of revolution

$$x^1 = u^1\cos u^2; \; x^2 = u^1\sin u^2; \; x^3 = f(u^1)$$

with $f(u^1)$ of class C^2.

Solution: The parametric representation of the right helicoid is given by

$$x^1 = u^1\cos u^2; \; x^2 = u^1\sin u^2; \; x^3 = f(u^1)$$

i.e.

$$\mathbf{r} = (x^1; x^2; x^3) = \left(u^1\cos u^2; u^1\sin u^2; f(u^1)\right).$$

The symmetric covariant tensors $b_{\alpha\beta}$ for the second fundamental form are given by

$$b_{11} = \frac{f_{11}}{\sqrt{1+f_1^2}}; \; b_{22} = \frac{u^1 f_1}{\sqrt{1+f_1^2}} \text{ and } b_{12} = 0 = b_{21}.$$

where $f_1 = \frac{\partial f}{\partial u^1}$ and $f_{11} = \frac{\partial^2 f}{\partial u^1 \partial u^1}$. Since $b_{12} = 0$, the parametric curves on the given surface are conjugate.

8.3.3 Asymptotic Directions

The asymptotic directions at a point on the surface are the self-conjugate directions; and the asymptotic line is a curve whose direction at every point is self-conjugate.

Consequently, in Eq. (8.38) connecting conjugate directions we put $\frac{u^1}{u^2}$ equal to $\frac{du^1}{du^2}$, the directions on the surface given by

$$b_{11}(du^1)^2 + 2\,b_{12}du^1du^2 + b_{22}(du^2)^2 = 0$$

or

$$b \quad du \; du \;\; = b \qquad\quad = 0 \tag{8.39}$$

and are called the asymptotic directions. The curves whose tangents are asymptotic directions are called an asymptotic line. Thus, there are two asymptotic directions at a point. The asymptotic directions are

(i) real and distinct, when $b_{12}^2 \quad b_{11}b_{22}$ is positive, i.e. $\; < \; 0$;

(ii) imaginary, when $\; > \; 0$;

(iii) identical, when $\; = 0:$

Now, the normal curvature $_{(n)}$ is given in Eq. (8.14). In the case (iii) the surface is a developable, and the single asymptotic line through a point is a generator. Since for an asymptote direction Eq. (8.39) holds, it follows that, for an asymptotic direction the normal curvature is zero, i.e.

$$_{(n)} = 0 \;) \quad b \quad du \;\, du \;\; = 0$$

for the asymptotic directions. These directions are, therefore, the directions of the asymptotes of the indicatrix. They are at right angles when the indicatrix is a rectangular hyperbola, i.e. when the principal curvature are equal and opposite. Thus, the asymptotic lines are orthogonal when the surface is a minimal surface. The osculating plane at any point of an asymptotic line is the tangent plane to the surface.

Deduction 8.3.5 Let $(du^1; du^2)$ be the direction of asymptotic line at P then we know

$$b \quad du \;\, du \;\; = b_{11}(du^1)^2 + 2\,b_{12}du^1du^2 + b_{22}(du^2)^2 = 0:$$

Now, the normal curvature $_{(n)}$ is given by

$$_{(n)} = \frac{b_{11}(du^1)^2 + 2\,b_{12}du^1du^2 + b_{22}(du^2)^2}{a_{11}(du^1)^2 + 2\,a_{12}du^1du^2 + a_{22}(du^2)^2} = 0:$$

If $_a$ and $_b$ are principal curvatures at P, then, by Euler's theorem

$$_a\cos^2 \quad + \quad _b\sin^2 \quad = \quad _{(n)} = 0: \tag{8.40}$$

Now, if be the normal curvature in a direction perpendicular to asymptotic line then from Eq. (8.40), we get

$$= {}_a \cos^2 \frac{}{2} + \quad + {}_b \sin^2 \frac{}{2} +$$

or

$$= {}_a \sin^2 + {}_b \cos^2 : \tag{8.41}$$

Adding Eqs. (8.40) and (8.41), we get

$$= {}_a + {}_b = 2 \frac{a + b}{2}$$

$$= \text{twice of mean normal curvature at } P:$$

Thus, the normal curvature in a direction perpendicular to an asymptotic line is twice the mean normal curvature.

EXAMPLE 8.3.12 Show that a straight line on a surface is an asymptotic line.

Solution: We know, the normal curvature $_{(n)}$ is given by

$$_{(n)} = \cos ;$$

where is the curvature of the curve on the surface and is the angle between the principal normal and the surface normal. Since for a straight line $= 0$; so,

$$_{(n)} = \frac{b \quad du \ du}{a \quad du \ du} = 0$$

$$) \quad b \quad du \ du = b \qquad = 0:$$

Hence, any straight line on a surface is an asymptotic line. From this result, it follows that at each point of an asymptotic curve C, at which > 0, the principal normal of C lies in the tangent plane of the surface. Thus the generating straight lines of a cylinder or as cone are asymptotic curves on the surface. From this we obtain the following important property of asymptotic curves.

At any point of an asymptotic curve C for which the curvature > 0, the binormal of C and the normal to the surface coincide. Consequently, at any point of C, the osculating plane of C and the tangent plane to the surface then coincide.

Theorem 8.3.4 The parametric curves are asymptotic lines if and only if

$$b_{11} = b_{22} = 0:$$

Proof: First, let the parametric curves u^1 = constant and u^2 = constant be asymptotic lines. Then from Eq. (8.39), we get

$$b \quad du \quad du \;=\; b \qquad\quad = 0:$$

For u^1 curve, i.e. $du^2 = 0$; this equation reduces to

$$b \quad {}_{(1)}\;{}_{(1)} = b \quad \frac{1}{\sqrt{a_{11}}}\;{}_{(1)}\;\frac{1}{\sqrt{a_{11}}}\;{}_{(1)} = 0$$

$$\Rightarrow \frac{b_{11}}{a_{11}} = 0 \Rightarrow b_{11} = 0:$$

Similarly, for u^2 curve, i.e. $du^1 = 0$; i.e.

$$b \quad {}_{(2)}\;{}_{(2)} = b \quad \frac{1}{\sqrt{a_{22}}}\;{}_{(2)}\;\frac{1}{\sqrt{a_{22}}}\;{}_{(2)} = 0$$

$$\Rightarrow \frac{b_{22}}{a_{22}} = 0 \Rightarrow b_{22} = 0:$$

Conversely, let $b_{11} = b_{22} = 0$; then the differential Eq. (8.39) determining the asymptotic lines is

$$b_{12}du^1du^2 = 0 \Rightarrow du^1du^2 = 0; \quad \text{as } b_{12} \neq 0;$$

i.e. the curves are parametric. Of course, co-ordinates of this type can be introduced on a surface S if and only if at any point of S there are two (different real) asymptotic directions.

EXAMPLE 8.3.13 Prove that the parametric curves on a surface are asymptotic lines if and only if $b_{11} = b_{22} = 0$ and show that

$$= \quad \frac{b_{12}^2}{a}; \quad H = \quad \frac{a_{12}b_{12}}{a}; \quad \frac{a_{12}}{b_{12}} = \frac{H}{\ }:$$

Solution: We know, parametric curves are asymptotic lines if and only if $b_{11} = b_{22} = 0$: Now, the Gaussian curvature is given by

$$= \frac{b}{a} = \frac{b_{11}b_{22} \quad b_{12}^2}{a_{11}a_{22} \quad a_{12}^2} = \frac{b_{12}^2}{a_{11}a_{22} \quad a_{12}^2} = \quad \frac{b_{12}^2}{a}:$$

Since $b_{11} = b_{22} = 0$, therefore, the mean curvature H is given by

$$H = \frac{1}{2}a \quad b \quad = \frac{1}{2} \; a^{11}b_{11} + a^{22}b_{22} + 2\,a^{12}b_{12}$$

$$= a^{12}b_{12} = \quad \frac{a_{12}b_{12}}{a}:$$

These are the values of the mean curvature H and the Gaussian curvature , when parametric curves are asymptotic lines. Thus, the ratio is given by

$$\frac{H}{} = \frac{a_{12}b_{12}}{a} \quad \frac{a}{b_{12}^2} = \frac{a_{12}}{b_{12}}:$$

EXAMPLE 8.3.14 Show that the parametric curves are asymptotic lines to the surfaces

$$x^1 = u^1 \cos u^2;\ x^2 = u^1 \sin u^2;\ x^3 = cu^2:$$

Prove that the asymptotic lines consist of generators and the curves of intersection with coaxial right circular cylinders.

Solution: The parametric curves are asymptotic lines if $b_{11} = 0$; $b_{22} = 0$ and $b_{12} \neq 0$: For the given surface

$$x^1 = u^1 \cos u^2;\ x^2 = u^1 \sin u^2;\ x^3 = cu^2$$

or

$$r = (x^1; x^2; x^3) = \quad u^1 \cos u^2; u^1 \sin u^2; cu^2$$

the symmetric covariant tensors b for the second fundamental form are given by

$$b_{11} = 0; b_{22} = 0 \text{ and } b_{12} = \quad \frac{c}{\sqrt{c^2 + (u^1)^2}} \neq 0:$$

The differential equation giving the asymptotic lines is

$$b \quad du \ du \ = b_{11}(du^1)^2 + 2b_{12}du^1du^2 + b_{22}(du^2)^2 = 0$$

or

$$\frac{c}{\sqrt{c^2 + (u^1)^2}} du^1du^2 = 0$$

or

$$u^1 = \text{constant and } u^2 = \text{constant};$$

i.e. the parametric curves. Now, the curves u^2 = constant are generators, whereas the curves u^1 = constant are the curves of intersection of the given helicoid with the circular cylinders $(x^1)^2 + (x^2)^2 = (u^1)^2$: The axis of these coaxial cylinders is x^3 axis which is also the axis of the given helicoid.

EXAMPLE 8.3.15 Determine the asymptotic curves of a cylinder of revolution.

Solution: Let the parametric equation of the surface S is of the form

$$r = (x^1; x^2; x^3) = (a \cos u^1; a \sin u^1; u^2):$$

We know any straight line on a surface S of class r 2 is an asymptotic curve. Thus the generating straight lines of a cylinder (or a cone) S are asymptotic curves on these surfaces. We will prove that those curves are the only asymptotic curves of S. The second fundamental magnitudes are given by

$$b_{12} = b_{22} = 0 \text{ and } b_{11} = \ 1 \ 6= 0$$

Thus Eq. (8.39) takes the form

$$b_{11}(du^1)^2 + b_{22}(du^2)^2 + 2b_{12}du^1du^2 = 0$$
$$)\quad b_{11}(du^1)^2 = 0; \text{ where } b_{11} = \ 1 \ 6= 0$$
$$)\quad du^1 = 0; \text{ that is, } u^1 = \text{constant.}$$

Since this is the only solution of that equation the above straight lines are the only asymptotic curves on the cylinder.

EXAMPLE 8.3.16 Show that the two directions given by $h \ du \ du = 0$; are orthogonal if and only if $a \ h = 0$:

Solution: The given equation $h \ du \ du = 0$ can be written in the form

$$h_{11}(du^1)^2 + h_{12}du^1du^2 + h_{21}du^2du^1 + h_{22}(du^2)^2 = 0$$

or

$$h_{11}\ \frac{du^1}{du^2}^{\ 2} + (h_{12} + h_{21})\frac{du^1}{du^2} + h_{22} = 0:$$

Let $\frac{du^1}{du^2}$ and $\frac{u^1}{u^2}$ be the roots of the quadratic equation. Then

$$\frac{du^1}{du^2} + \frac{u^1}{u^2} = \frac{h_{12} + h_{21}}{h_{11}} \text{ and } \frac{du^1}{du^2}\frac{u^1}{u^2} = \frac{h_{22}}{h_{11}}:$$

The two directions $(du^1; du^2)$ and $(u^1; u^2)$ will be perpendicular if and only if

$$a_{11}du^1 u^1 + a_{12}du^1 u^2 + a_{21}du^2 u^1 + a_{22}du^2 u^2 = 0$$

or

$$a_{11}\frac{du^1}{du^2}\frac{u^1}{u^2} + a_{12}\ \frac{du^1}{du^2} + \frac{u^1}{u^2}\ + a_{22} = 0$$

or

$$a_{11}\frac{h_{22}}{h_{11}} \quad a_{12} \quad \frac{h_{12} + h_{21}}{h_{11}} \quad + a_{22} = 0$$

or

$$a_{11}h_{22} \quad a_{12}(h_{12} + h_{21}) + a_{22}h_{11} = 0$$

or

$$a \; a^{22}h_{22} + a^{12}h_{12} + a^{21}h_{21} + a^{11}h_{11} \; = 0$$

or

$$a \quad h \quad = 0; \quad \text{as } a \neq 0 :$$

EXAMPLE 8.3.17 The asymptotic lines are at right angles if and only if $H = 0$:

Solution: The asymptotic lines are given by the equation

$$b \quad du \; du \; = b \qquad = 0 :$$

Thus, the asymptotic lines are at right angles if and only if

$$a \quad b \quad = 0 \;) \quad H = \frac{1}{2}a \quad b \quad = 0 ;$$

where H is the mean curvature. Thus, the asymptotic lines are right angles if and only if the surface is a minimal surface. Hence the asymptotic lines form an orthogonal system, bisecting the angles between the lines of curvature.

EXAMPLE 8.3.18 Show that on a sphere there is no real asymptotic line.

Solution: The parametric representation of the sphere of radius a is given by

$$x^1 = a \sin u^1 \cos u^2; \quad x^2 = a \sin u^1 \sin u^2; x^3 = a \cos u^2$$

or

$$r = (x^1; x^2; x^3) = \quad a \sin u^1 \cos u^2; a \sin u^1 \sin u^2; a \cos u^2 \; :$$

For the given surface, the symmetric covariant second order tensor a for the first fundamental form are given by

$$a_{11} = a^2; \quad a_{22} = a^2 \sin^2 u^1; \quad a_{12} = 0 = a_{21}$$

and the symmetric covariant second order tensors b for the second fundamental form are given by

$$b_{11} = \quad a; \quad b_{22} = \quad a \sin^2 u^1; \quad b_{12} = 0 = b_{21}:$$

Since $a_{12} = 0 = b_{12}$, this given parametric curve has the line of curvature. In fact,

$$\frac{b_{11}}{a_{11}} = \frac{b_{12}}{a_{12}} = \frac{b_{22}}{a_{22}} = \quad \frac{1}{a};$$

i.e. b are proportional to a : Now, the asymptotic lines are given by

$$b \quad du \quad du \ = b_{11}(du^1)^2 + 2b_{12}du^1du^2 + b_{22}(du^2)^2 = 0$$

or

$$a(du^1)^2 \quad a\sin^2 u^1(du^2)^2 = 0$$

or

$$(du^1)^2 + \sin^2 u^1(du^2)^2 = 0:$$

Hence, there is no real asymptote.

Theorem 8.3.5 The torsion of an asymptotic line equals to p —; where is the Gaussian curvature of the surface.

Proof: From Eq. (8.6), we have, for an asymptotic curve,

$$^i = b \qquad ^i + \ _g\ ^i; = \ _g\ ^i; \quad \text{as}\, b \qquad = 0; \qquad (8.42)$$

where ^{i} is the principal normal and ^{i} is the space vector giving the same direction on the surface as ^{i}; the unit normal to the surface S. Therefore,

$$(\quad ^i)(\quad ^i) = (\ _g\ ^i)(\ _g\ ^i)$$

or

$$^2 = \ ^2_h\ ) \qquad = \qquad _h \qquad (8.43)$$

from which we conclude that the curvature and the geodesic curvature of an asymptotic line are equal in magnitude. Also,

$$^i = \ _g\ ^i\ ) \quad ^i = \quad ^i: \qquad (8.44)$$

We now derive the well known formula for torsion of an asymptotic line. From Eq. (8.44), it follows that the principal normal ^{i} of an asymptotic line is tangent to the surface and hence the binormal of an asymptotic line coincides with the surface normal ^{i}, i.e.

$$^i = \quad ^i: \qquad (8.45)$$

Taking intrinsic derivative of both the sides of Eq. (8.45), we get

$$\frac{\ \ ^{i}}{s} = \frac{\ \ ^{i}}{s} \Rightarrow \quad ^{i} = \ ^{i}_{;} \frac{du}{ds} = \ ^{i}_{;} \quad :$$

Therefore,

$$^{2} = g_{ij} (\quad ^{i})(\quad ^{j}) = g_{ij} \ ^{i}_{;} \ ^{j}_{;}$$

$$= c \qquad ; \quad \text{as}\, c \quad = g_{ij} \ ^{i}_{;} \ ^{j}_{;} :$$

Using the relation

$$c \qquad 2Hb \quad + \ a \qquad = 0;$$

where H is the mean curvature and is the Gaussian curvature, we get

$$c \qquad\qquad 2Hb \qquad\qquad + \ a \qquad\qquad = 0$$

or

$$^{2} \quad 0 + \ ^{2} = 0 \Rightarrow \quad = \sqrt{\quad} :$$

Therefore, the torsions of the two asymptotic lines through a point are equal in magnitude and opposite in sign; and the square of either is negative of the specific/Gaussian curvature of the surface. This theorem is formulated by Beltrami Enneper.

EXAMPLE 8.3.19 Find the asymptotic lines on the surface of revolution

$$x^1 = u^1 \cos u^2; \quad x^2 = u^1 \sin u^2; x^3 = f(u^1):$$

Find also the values of their torsions.

Solution: For the given surface

$$x^1 = u^1 \cos u^2; \; x^2 = u^1 \sin u^2; x^3 = f(u^1)$$

or

$$r = (x^1; x^2; x^3) = \quad u^1 \cos u^2; \; u^1 \sin u^2; f(u^1)$$

the symmetric covariant tensors b for the second fundamental form are given by

$$b_{11} = \frac{u^1 f_{11}}{\sqrt{1+f_1^2}}; b_{12} = 0 = b_{21} \text{ and } b_{22} = \frac{(u^1)^2 f_1}{\sqrt{1+f_1^2}}:$$

The differential equation giving the asymptotic lines is

$$b_{11}(du^1)^2 + 2b_{12}du^1du^2 + b_{22}(du^2)^2 = 0$$

or

$$f_{11}(du^1)^2 + u^1 f_1 (du^2)^2 = 0:$$

Using Beltrami Enneper's formula the torsion of an asymptotic line equals to $\sqrt{\ }$; where is the Gaussian curvature of the surface, as

$$= \sqrt{\ } = \frac{\sqrt{(u^1)^3 f_{11} f_1}}{(u^1)^2(1+f_1^2)}:$$

8.4 Conformal Mapping

Isometric mappings of a surface S on another surface S have been already mentioned. It is often convenient to consider mappings which are more general than isometrices-for example, it is useful to map parts of the earth's surface onto a flat atlas. Certain of these maps are particular cases of conformal mappings.

A surface S is said to be conformally mapped on a surface S if there is a differentiable homomorphism of S on S such that the angle between any two curves at an arbitrary point P on S is equal to the angle between the corresponding curves on S .

An isometric mappings preserves both distances and angles, whereas a conformal mapping just preserves angles. Let the fundamental forms of S, S be respectively

$$ds^2 = a_{11}du^2 + 2a_{12}dudv + a_{22}dv^2$$

$$ds^{2} = a_{11}du^2 + 2a_{12}dudv + a_{12}^{2}dv^2$$

the correspondence being such that corresponding points P; P have the same parametric values. Then if $\frac{ds}{ds}$ has the same value for all directions at the given point, we must have

$$\frac{a_{11}}{a_{11}} = \frac{a_{12}}{a_{12}} = \frac{a_{22}}{a_{22}} = \frac{ds^{2}}{ds^2} = {}^{2}; \qquad (8.46)$$

where is function of u and v or a constant. Equation (8.46) is evidently a necessary and sufficient condition for a differentiable homomorphism to be conformal. Therefore

$$ds = ds:$$

The quantity may be called the linear magnification. When = 1, for all points of the surface, ds = ds. The conformal representation is then said to be isometric, and the surfaces are said to be applicable. In this case, corresponding elements of the two surfaces are congruent. Thus every isometric mapping is conformal.

EXAMPLE 8.4.1 Prove that every point on a surface has a neighbourhood which can be mapped conformably on some neighbourhood of any other surface.

Solution: Let S be the given surface with metric

$$ds^2 = a_{11}du^2 + 2\,a_{12}dudv + a_{22}dv^2$$

in some co-ordinate domain. At any point P there are two (imaginary) directions such that $ds^2 = 0$. These are called isotropic directions at P, and since $a = a_{11}a_{22} \quad a_{12}^2 \neq 0$ if follows that these directions are always distinct.

When curves along these directions are chosen as parametric curves the metric assumes the form $ds^2 = \quad dudv$. The change of parameters

$$u = U + iV; v = U \quad iV;$$

where U and V are real, leads to a metric of the form

$$ds^2 = \quad {}^2[dU^2 + dV^2]:$$

If this is compared with the metric of the plane

$$ds^2 = du^{\,2} + dv^{\,2};$$

it is rapidly seen that the mapping $u = U; v = V$ gives a conformal mapping of a region of the given surface on a region of the plane. Hence the result follows.

EXAMPLE 8.4.2 Show that null lines on a surface correspond to null lines in the conformal representation.

Solution: We have deduce that $ds^{\,2} = r^2 ds^2$, if ds^2 vanishes along a curve S; $ds^{\,2}$ will vanish along the corresponding curve on S . Conversely, let the null lines be taken as parametric curves. Then

$$a_{11} = a_{22} = 0 \text{ and } a_{11} = a_{22} = 0$$

$$) \quad \frac{ds^{\,2}}{ds^2} = \frac{2a_{12}dudv}{2a_{12}dudv} = \frac{a_{12}}{a_{12}}:$$

Since the $\frac{ds}{ds}$ has the same value for all arcs through a given point, the representation is conformal. Thus, if null lines on S corresponds to null lines on S , the representation is conformal.

8.5 Spherical Image

We shall now consider briefly the spherical representation of a surface, in which each point or configuration on the surface has its representation on a unit sphere, whose centre may be taken as origin. If is the unit normal at the point P on the surface S, the point Q whose position vector is is said to correspond to P, or the image of P. Clearly, Q lies on the unit sphere; and if P moves in any curve on the surface, Q moves in the corresponding curve on the sphere. The mapping thus defined is called the Gaussian spherical mapping of the surface S. The set of all image points is called the spherical image of S. Since the position vector $\bar{r}$ of Q is given by $\bar{r} =$, it follows that

$$\frac{\partial}{\partial u} = \frac{\partial :}{\partial u} = \frac{1}{a^2}\,(a_{12}b_{12} \quad a_{22}b_{11})\,\frac{\partial}{\partial v} + (\,a_{12}b_{11} \quad a_{11}b_{12})\,\frac{\partial}{\partial v}$$

$$\frac{\partial}{\partial v} = \frac{\partial}{\partial v} = \frac{1}{a^2}\,(a_{12}b_{22} \quad a_{22}b_{11})\,\frac{\partial}{\partial u} + (a_{12}b_{12} \quad a_{11}b_{12})\,\frac{\partial}{\partial v}\,:$$

Consequently, if $a_{11}; a_{12}; a_{22}$ denote the fundamental magnitudes of the first order for the spherical image

$$\left.\begin{aligned} a_{11} &= \frac{1}{a^2}\; a_{11}b_{12}^2 \quad 2a_{12}b_{11}b_{12} + a_{22}b_{11}\;; \\ a_{12} &= \frac{1}{a^2}\; a_{11}b_{12}b_{22} \quad a_{12}b_{12}^2 \quad a_{12}b_{11}b_{22} + a_{22}b_{11}b_{12}\;; \\ a_{22} &= \frac{1}{a^2}\; a_{11}b_{22}^2 \quad 2a_{12}b_{12}b_{22} + a_{22}b_{12}^2\;; \end{aligned}\right\} \tag{8.47}$$

or, in terms of the first and second curvatures

$$a_{11} = Hb_{11} \quad ka_{11};\;\; a_{12} = Hb_{12} \quad ka_{12};\;\; a_{22} = Hb_{22} \quad ka_{22} \tag{8.48}$$

In virtue of Eq. (8.48) we may write the square of the linear element of the image.

$$ds^2 = H(b_{11}du^2 + 2b_{12}dudv + b_{22}dv^2) \quad k(a_{11}du^2 + 2a_{12}dudv + a_{22}dv^2);$$

or, if $_{(n)}$ is the normal curvature of the given surface in the direction of the arc-length

$$ds^2 = \;\; _{(n)}H \quad k \;\; ds^2:$$

If then $_{(1)}$ and $_{(2)}$ are the principal curvatures of the surface, we may write this in virtue of Euler's theorem

$$\begin{aligned} ds^2 &= \Big[\; _{(1)} + \; _{(2)} \quad _{(1)}\cos^2 \; + \; _{(2)}\sin^2 \quad _{(1)} \; _{(2)} \Big]\, ds^2 \\ &= \; _{(1)}\cos^2 \; + \; ^2_{(2)}\sin^2 \;\; ds^2: \end{aligned} \tag{8.49}$$

It is clear from either of these formulae that the value of the quotient $\frac{ds}{ds}$ depends upon the direction of the arc element. The fundamental magnitudes for the spherical representation, as given by Eq. (8.48), become in the case of a minimal surface

$$a_{11} = \quad a_{11}; \; a_{12} = \quad a_{12}; \; a_{22} = \quad a_{22}: \qquad (8.50)$$

From these relations several interesting general properties may be deduced.

EXAMPLE 8.5.1 Show that the spherical image is not a conformal representation in general.

Solution: The spherical image be conformal, however, if $_{(1)} = \quad _{(2)}$. When $_{(1)} = \quad _{(2)}$ at all points, the Gaussian curvature becomes

$$H = \frac{1}{2} \quad _{(1)} + \quad _{(2)} = 0$$

i.e. H vanishes indentically and the surface is a minimal surface. Thus, for a minimal surface

$$ds^2 = \quad ds^2:$$

Thus $\frac{ds}{ds}$ is independent of the direction of the arc-element through the point, and the representation is conformal. The magnification has the value $p\overline{\quad}$, the second curvature being essentially negative for a real minimal surface. Thus, the spherical representation of a minimal surface is conformal. Therefore, in general, the spherical image is not a conformal representation.

Moreover it follows from Eq. (8.49) that the turning values of $\frac{ds}{ds}$ are given by cos = 0 and sin = 0. Thus the greatest and least values of the magnification at a point are numerically equal to the principal curvatures. From this it follows that if the spherical image of a surface is a conformal representation, either the surface is minimal, or else its principal curvatures are equal at each point.

EXAMPLE 8.5.2 Prove that null lines on a minimal surface become both null lines and asymptotic lines in the spherical representation.

Solution: Let the null lines be taken as parametric curves, then $a_{11} = 0 = a_{22}$. Therefore

$$Hb_{11} \quad a_{11} = 0 \quad \text{and} \quad Hb_{22} \quad a_{22} = 0:$$

Thus the parametric curves in the spherical image are null lines. Again considering the second order magnitudes for the sphere, we have

$$b_{11} = 0 = b_{12}$$

and therefore, the parametric curves in the spherical image are also asymptotic lines. Conversely, let the null lines are parametric curves, then $a_{11} = 0 = a_{22}$; and since the parametric curves are also null lines in the spherical image

$$Hb_{11} \quad ka_{11} = 0;\ Hb_{22} \quad ka_{22} = 0$$

$$)\ \ Hb_{11} = 0 \quad \text{and} \quad Hb_{22} = 0:$$

Consequently, either $H = 0$ and the surface is minimal, or else $b_{11} = 0 = b_{22}$. In the later case, it follows that

$$H = \frac{2b_{12}}{a_{12}} \quad \text{and} \quad k = \frac{b_{12}^2}{a_{12}^2} \quad) \quad H^2 \quad 4 = 0;$$

which is the condition that the principal curvatures should be equal. Thus, if the null lines in the spherical representation, either the surface is minimal, or else its principal curvatures are equal.

EXAMPLE 8.5.3 Show that the isometric lines on a minimal surface are also isometric in the spherical representation.

Solution: If the isometric lines are taken as parametric curves, we have

$$a_{12} = 0 \quad \text{and} \quad \frac{a_{11}}{a_{22}} = \frac{U(u)}{V(v)};$$

where U is a function of u only and V is a function of v only. Therefore, it follows that

$$\frac{H}{} = \frac{a_{12}}{b_{12}};\ \frac{a_{12}}{a_{22}} = \frac{U(u)}{V(v)};$$

showing that the parametric curves in the spherical image are also isomorphic. Hence the theorem. This result is obvious from the fact that the spherical image on a minimal surface is conformal.

In particular, the spherical image of the lines of curvature on a minimal surface are isometric curves, for the lines of curvature on a minimal surface have been shown to be isometric.

EXAMPLE 8.5.4 Show that the lines of curvature on a surface are orthogonal in their spherical representation.

Solution: The lines of curvature on a surface are orthogonal in their spherical representation. For if they are taken as parametric curves, we have $a_{12} = 0 = b_{12}$, hence $a_{12} = 0$ which proves the statement. Further if $a_{12} = 0$ and $a_{12} = 0$ we must also have $b_{12} = 0$ unless the Gaussian curvature H vanishes identically. Thus, if the surface is not a minimal surface, the lines of curvature are the only orthogonal system whose spherical image is orthogonal.

Moreover, by Rodrigue’s formula along a line of curvature dr is parallel to d and therefore also to $d\bar{r}$. Hence the tangent to a line of curvature is parallel to the tangent to its spherical image at the corresponding point.

Conversely if dr is parallel to dr it is parallel to d . The vectors ; + d , dr are therefore coplanar, and the line is a line of curvature. Thus if the relation holds for a curve on a surface it must be a line of curvature.

EXAMPLE 8.5.5 Prove that if two directions are conjugate at a point on a given surface, each is perpendicular to the spherical image of the other at the corresponding point.

Solution: If dr and ds be two infinitesimal displacements on a given surface, and d the change in the unit normal due to the former, the directions of the displacements will be conjugate provided

$$d \quad ds = 0:$$

Conversely this relation holds if the directions are conjugate. But $d = dr$, where dr is the spherical image of dr. Consequently

$$dr \quad ds = 0:$$

Thus if two directions are conjugate at a point on a given surface, each is perpendicular to the spherical image of the other at the corresponding point. It follows that the inclination of two conjugate directions are equal, or supplementary, to that of their spherical representations.

Further, an asymptotic line is self-conjugate. Hence an asymptotic line on a surface is perpendicular to its spherical image at the corresponding point.

8.6 Surface of Revolution

A surface of revolution may be generated by the rotation of a plane curve about an axis in its plane. Let C: $x^3 = f(x^1)$ be a curve drawn on the plane $x^2 = 0$; where x^i are Cartesian co-ordinates of a point. Let $P(x^1; 0; x^3)$ be a point on the curve, then P traces a circle as the curve C is rotated about the x^3 axis. The surface generated by the curve is called the surface of revolution and its implicit equation is

$$x^3 = \sqrt{(x^1)^2 + (x^2)^2} : \qquad (8.51)$$

To obtain the parametric equation of the surface, let the generating curve be given by

$$x^1 = f_1(u^1); \quad x^2 = 0; \quad x^3 = f_2(u^2):$$

Let u^2 be the angle of rotation about the x^3 axis measured from the x^1 axis, then

$$x^1 = f_1(u^1)\cos u^2;\, x^2 = f_1(u^1)\sin u^2;\ \ x^3 = f_2(u^1).$$

This is the parametric representation of the surface of revolution obtained by rotating a curve about the x^3 axis. The curve C is called the generator of the curve.

Let the axis of rotation be x^3 and u^1 denote perpendicular distance from it, the co-ordinates of a point on the surface may be expressed

$$x^1 = u^1\cos u^2;\, x^2 = u^1\sin u^2;\ \ x^3 = f(u^1);$$

the longitude u^2 being the inclination of the axial plane through the given point to the x^3x^1 plane. The parametric curves u^2 = constant are the meridian lines, or intersections of the surface by the axial planes; the curves u^1 = constant are the parallels, or intersection of the surface by planes perpendicular to the axis.

EXAMPLE 8.6.1 Let S is a surface of revolution

$$x^1 = u^1\cos u^2;\, x^2 = u^1\sin u^2;\ \ x^3 = f(u^1);\ \ f(u^1) \in C^2.$$

(a) Find the surface of revolution that has a minimal surface.

(b) Find the surface of revolution that has a constant negative Gaussian curvature.

(c) Find the complete integral of the equations of geodesics.

Solution: With $u^1; u^2$ as parameters, the parametric representation of the right helicoid is given by

$$x^1 = u^1\cos u^2;\, x^2 = u^1\sin u^2;\ \ x^3 = f(u^1)$$

i.e.

$$\mathbf{r} = (x^1; x^2; x^3) = \left(u^1\cos u^2; u^1\sin u^2; f(u^1)\right).$$

Here we have to show that the catenoid is the only minimal surface of revolution. The first order magnitudes are given by

$$a_{11} = 1 + f_1^2;\ a_{22} = (u^1)^2 \text{ and } a_{12} = 0 = a_{21};$$

$$\Rightarrow\ a = |a_{\alpha\beta}| = \begin{vmatrix} 1 + f_1^2 & 0 \\ 0 & (u^1)^2 \end{vmatrix} = (u^1)^2(1 + f_1^2);$$

where $f_1 = \frac{\partial f}{\partial u^1}$. Since $a_{12} = 0$, it follows that the parallels cut the meridians orthogonally. The second order magnitudes are given by

$$b_{11} = \frac{f_{11}}{\sqrt{1 + f_1^2}};\, b_{22} = \frac{u^1 f_1}{\sqrt{1 + f_1^2}} \text{ and } b_{12} = 0 = b_{21};$$

$$\Rightarrow\ b = |b_{\alpha\beta}| = \begin{vmatrix} \frac{f_{11}}{\sqrt{1 + f_1^2}} & 0 \\ 0 & \frac{u^1 f_1}{\sqrt{1 + f_1^2}} \end{vmatrix} = \frac{u^1 f_1 f_{11}}{1 + f_1^2};$$

where $f_1 = \frac{\partial f}{\partial u^1}$ and $f_{11} = \frac{\partial^2 f}{\partial u^1 \partial u^1}$. Since $a_{12} = 0$ and $b_{12} = 0$, the parametric curves are the lines of curvature. The equation for the principal curvatures reduces to

$$u^1(1 + f_1^2)^2 \kappa^2 - \sqrt{1 + f_1^2}\,\{u^1 f_{11} + f_1(1 + f_1^2)\}\kappa + f_1 f_{11} = 0.$$

This is a quadratic equation in κ which gives the two values $\kappa_{(1)}$ and $\kappa_{(2)}$ given by

$$\kappa_{(1)} = \frac{f_{11}}{(1 + f_1^2)^{3/2}}; \quad \kappa_{(2)} = \frac{f_1}{u^1 \sqrt{1 + f_1^2}}.$$

The first of those is the curvature of the generating curve. The second is the reciprocal of the length of the normal intercepted between the curve and the axis of rotation.

(a) Now, the mean curvature H is given by

$$H = \frac{a_{11}b_{22} - 2a_{12}b_{12} + a_{22}b_{11}}{a} = \frac{f_1(1 + f_1^2) + u^1 f_{11}}{u^1(1 + f_1^2)^{3/2}}.$$

The surface will be a minimal surface if $H = 0$; i.e. if

$$f_1(1 + f_1^2) + u^1 f_{11} = 0$$

or

$$F(1 + F^2) + u^1 F_1 = 0; \quad F = f_1 = \frac{\partial f}{\partial u^1}$$

or

$$\frac{du^1}{u^1} + \frac{dF}{F(1 + F^2)} = 0$$

or

$$\log u^1 + \log F - \frac{1}{2}\log(1 + F^2) = \log c; \text{ integrating}$$

or

$$\frac{u^1 F}{\sqrt{1 + f_1^2}} = c \Rightarrow F = \frac{\partial f}{\partial u^1} = \pm\frac{c}{\sqrt{(u^1)^2 - c^2}}.$$

Take the positive sign (the other case is exactly similar) and integrating we get

$$f + c_1 = \cosh^{-1}\left(\frac{u^1}{c}\right) \Rightarrow u^1 = c\cosh(x^3 + c_1).$$

Hence, the required surface is the surface of revolution obtained by revolving the catenary $u^1 = c\cosh(x^3 + c_1)$ about x^3 axis. Thus the only minimal surface of revolution is that formed by the revolution of a catenary about its directrix.

(b) Now, the Gaussian curvature K is given by

$$K = \frac{b}{a} = \frac{f_1 f_{11}}{(u^1)^2(1 + f_1^2)}.$$

Given that, the surface of revolution that has a constant negative Gaussian curvature, say $K = -\frac{1}{c^2}$. Therefore,

$$\frac{f_1 f_{11}}{(u^1)^2(1+f_1^2)} = -\frac{1}{c^2}$$

or

$$\frac{dF}{F^2} = -\frac{2u^1du^1}{c^2};\quad F = 1+f_1^2 = 1+\left(\frac{\partial f}{\partial u^1}\right)^2$$

or

$$\frac{1}{F} = \frac{1}{1+f_1^2} = \frac{(u^1)^2}{c^2};\quad \text{integrating}$$

or

$$1+f_1^2 = \frac{c^2}{(u^1)^2} \Rightarrow f_1 = \pm\frac{\sqrt{c^2-(u^1)^2}}{u^1};$$

where we choose the constant of integration equal to 0. Take the positive sign (the other case is exactly similar) and integrating we get

$$f = x^3 = \sqrt{c^2-(u^1)^2} + \frac{c}{2}\log\frac{c-\sqrt{c^2-(u^1)^2}}{c+\sqrt{c^2-(u^1)^2}}.$$

Hence, the required surface is the surface of revolution obtained by revolving the tractrix

$$x^3 = \sqrt{c^2-(u^1)^2} + \frac{c}{2}\log\frac{c-\sqrt{c^2-(u^1)^2}}{c+\sqrt{c^2-(u^1)^2}}.$$

about x^3 axis, which is the axis of tractrix.

(c) The second equation for geodesics is of the form

$$\frac{d^2u^2}{ds^2} + \frac{2}{u^1}\frac{du^1}{ds}\frac{du^2}{ds} = 0.$$

On multiplication by u^2 this equation becomes exact, and has for its integral

$$(u^1)^2\frac{du^2}{ds} = h;\quad h = \text{constant}.$$

If ψ is the angle at which the geodesics cuts the meridian, we may write $u^1\sin\psi = h$, a theorem due to Clairant. This is a first integral of the equation of geodesics, involving one orbitrary constant h.

To obtain the complete integral we observe that, for any arc on the surface,

$$ds^2 = (1+f_1^2)(du^1)^2 + (u^1)^2(du^2)^2$$

and therefore, for the arc of a geodesic

$$(u^1)^4(du^2)^2 = h^2(1+f_1^2)(du^1)^2 + h^2(u^1)^2(du^2)^2$$

or

$$du^2 = \frac{h}{u^1}\sqrt{\frac{1+f_1^2}{(u^1)^2-h^2}}\,du^1$$

or

$$u^2 = c \pm h\int \frac{1}{u}\sqrt{\frac{1+f_1^2}{(u^1)^2-h^2}}\,du^1$$

involving the two arbitrary constants c and h, is the complete integral of the equation of geodesics on a surface of revolution.

8.7 Exercises

1. Find the principal curvatures of the surface S defined by
$$x^1 = u^1;\ x^2 = u^2;\quad x^3 = f(u^1;u^2).$$
2. Show that the principal curvatures are $-1;\ -1$ for the surface
$$(x^1)^2 + (x^2)^2 + (x^3)^2 = 1.$$
3. Find the equation for the principal curvatures, and the differential equation of the lines of curvature, for the surfaces
(i) $2z = \frac{x^2}{a} + \frac{y^2}{b}$; (ii) $3z = ax^3 + by^3$; (iii) $z = c\tan^{-1}\frac{y}{x}$.
4. Examine the curvature, and find the lines of curvature, on the surface $xyz = abc$.
5. Find the principal direction and principal curvatures on the surface
$$x^1 = a(u^1+u^2);\quad x^2 = b(u^1-u^2);\quad x^3 = u^1u^2.$$
Show also that on the surface the parametric curves are asymptotic.
6. Find equations for the principal radii, the lines of curvature, and the first and second curvatures of the following surfaces:
   - (i) the conoid $x^1 = u^1\cos u^2;\ x^2 = u^1\sin u^2;\ x^3 = f(u^1)$;
   - (ii) the catenoid $x^1 = u^1\cos u^2;\ x^2 = u^1\sin u^2;\ x^3 = c\log\left(u^1 + \sqrt{(u^1)^2-c^2}\right)$;
   - (iii) the cylindroid $x^3((x^1)^2 + (x^2)^2) = 2mx^1x^2$; find also the lines of curvature and the principal curvatures;
   - (iv) the surface $2z = ax^2 + 2hxy + by^2$;

(v) the surface

$$x^1 = 3u^1(1 + (u^2)^2) \quad (u^1)^3;$$
$$x^2 = 3u^2(1 + (u^1)^2) \quad (u^2)^3;$$
$$x^3 = 3 \quad (u^1)^2 \quad (u^2)^2 \quad ;$$

Also show that the asymptotic lines are $u \quad v$ = constant.

(vi) the surface

$$\frac{x}{a} = \frac{1 + uv}{u + v}; \quad \frac{y}{b} = \frac{u \quad v}{u + v}; \quad \frac{z}{c} = \frac{1 \quad uv}{u + v};$$

(vii) the surface $xyz = a^3$:

7. Find the principal curvatures and the lines of curvature on the surface
$$z^2(x^2 + y^2) = c^4:$$

8. Prove that along a line of curvature of a conoid, one principal radius varies as the cube of the other.

9. Prove that if a plane cuts a surface everywhere at the same angle, the section is a line of curvature on the surface.

10. Prove that the only developable surface which have isometric lines of curvature are either conical or cylindrical.

11. Prove that the co-ordinate surfaces of every triply orthogonal curvilinear co-ordinate system in E^3 intersect along the lines of curvature of co-ordinate surfaces.

    Hints: Consider the surface x^3 = constant and take $x^1 = u^1; x^2 = u^2$ as surface co-ordinates on it. Show that along the co-ordinate lines u^1 = constant, u^2 = constant, $b_{12} = 0$ if $a_{12} = 0$:

12. Find the components of the curvature tensors R and R on $x^3 = f(x^1; x^2)$:

13. Prove that the Gaussian curvature k depends only on the coefficients of the first fundamental form (and their derivatives) but not on the second fundamental form.

14. Given an ellipsoid of revolution, whose surface is given by

$$x^1 = a\cos u^1 \sin u^2; \quad x^2 = a\sin u^1 \sin u^2; \quad x^3 = c\cos u^2;$$

where a and c are constants satisfying $a^2 > c^2$ and $x^1; x^2; x^3$ are orthogonal Cartesian co-ordinates. Show that

$$a_{11} = a^2\sin^2 u^2; \; a_{12} = 0; \quad a_{22} = a^2\cos^2 u^2 + c^2\sin^2 u^2;$$

$$b_{11} = \frac{ac\sin^2 u^2}{\sqrt{a^2\cos^2 u^2 + c^2\sin^2 u^2}}; \; b_{12} = 0; \; b_{22} = \frac{ac}{\sqrt{a^2\cos^2 u^2 + c^2\sin^2 u^2}}$$

and the total curvature K is given by

$$K = \kappa_1\kappa_2 = \frac{c^2}{(a^2\cos^2 u^2 + c^2\sin^2 u^2)^2}.$$

Discuss the lines of curvature on this surface.

15. Show that the lines of curvatures for the surface given by

$$(x^1)^2 + (x^2)^2 + (x^3)^2 = 1$$

are the parametric curves.

16. Show that the lines of curvatures for the surface given by

$$\mathbf{r} = (u\cos v, u\sin v, cv)$$

are given by $v = \pm\sinh^{-1}\frac{u}{c} + \text{constant}$. Also, show that on the surface the parametric curves are asymptotic lines.

17. Show that a system of confocal ellipses and hyperbolas are isometric lines in the plane.

18. Show that the umbilics of the surface

$$\left(\frac{x^1}{a}\right)^{2/3} + \left(\frac{x^2}{b}\right)^{2/3} + \left(\frac{x^3}{c}\right)^{2/3} = 1$$

lie on a sphere of radius $abc/(ab + bc + ca)$ whose centre is the origin.

19. If $a > b > c$, the ellipsoid $\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$ has umbilici at the points

$$y = 0, \; x^2 = \frac{a^2(a^2 - b^2)}{a^2 - c^2}, \; z^2 = \frac{c^2(b^2 - c^2)}{a^2 - c^2}.$$

20. Show that the surface $4a^2z^2 = (x^2 - 2a^2)(y^2 - 2a^2)$ has a line of umbilics lying on the sphere $x^2 + y^2 + z^2 = 4a^2$.

21. Prove that, at any point of the surface, the sum of the radii of normal curvature in conjugate directions is constant.

22. Prove that the product of the radii of normal curvature in conjugate directions is a minimum for the lines of curvature.

23. Prove that the normal curvature in a direction perpendicular to an asymptotic line is twice the mean curvature.

24. Show that on the paraboloid $2z = \frac{x^2}{a^2} - \frac{y^2}{b^2}$ the asymptotic lines are $\frac{x}{a} \pm \frac{y}{b} = \text{constant}$.

25. Find the asymptotic lines on the surface

$$x = a(1+\cos u)\cot v; \ y = a(1+\cos u); \ z = \frac{a\cos u}{\sin v}:$$

26. Prove that the asymptotic lines of the right conoid

$$x^1 = u^1\cos u^2; x^2 = u^1\sin u^2; \ x^3 = f(u^2)$$

are given by u^2 = constant, $\frac{@f}{@u^2} = (u^{1})^2 c$: Hence, find the asymptotic lines of the cylindroid

$$x^1 = u^1\cos u^2; \ x^2 = u^1\sin u^2; \ x^3 = m\sin 2u^2:$$

Determine $f(u^2)$ so that on this conoid the parametric curves may be isometric lines.

27. Prove that the asymptotic lines of the catenoid $u^1 = c\cosh\frac{x^3}{c}$ lie on the cylinders

$$2u^1 = c \quad ae^{u^2} + \frac{1}{a}e^{\ \ u^2} \quad ; \text{where } a \text{ is arbitrary.}$$

Hints: Consider the parametric representation of the curve as

$$x^1 = u^1\cos u^2; x^2 = u^1\sin u^2; \ x^3 = f(u^1) = c\cosh^1 \ \frac{u^1}{c}:$$

28. Find the asymptotic lines on the surface $z = y\sin x$:

29. Prove that the asymptotic lines on the surface of revolution

$$x^1 = u^1\cos u^2; x^2 = u^1\sin u^2; \ x^3 = f(u^1)$$

are given by $f_{11}d(u^1)^2 + u^1 f_1 d(u^2)^2 = 0$: Write down the value of the torsions. Further if $u^1 = a\sin$ and $f(u^1) = a(\log\tan\frac{}{2} + \cos)$ the asymptotic lines are given by $du^2 = \frac{d}{\sin}$:

30. Find the asymptotic lines and torsions on the surface whose equation is

$$r = (x^1; x^2; x^3) = \quad x^1; \ x^2; \ f(x^1; x^2) \quad :$$

Hence, show that

(i) if $f(x^1; x^2) = x^2\sin x^1$, then the asymptotic line is $(x^2)^2\cos x^1 = c_1$:

(ii) if $f(x^1; x^2) = \frac{1}{2} \ \frac{(x^1)^2}{a^2} \quad \frac{(x^2)^2}{b^2}$, then the asymptotic line is $\frac{x^1}{a} \quad \frac{x^2}{b} = c_1$:

31. Prove that the generating straight lines of a cylinder or a cone are asymptotic curves on those surfaces.
32. Prove that the asymptotic lines on a minimal surface form an orthogonal system.
33. Prove that the indicatrix at a point of a surface $z = f(x; y)$ is a rectangular hyperbola if

$$(1 + p^2)t - 2pqs + (1 + q^2)r = 0:$$

Also, show that the asymptotic lines and torsions are

$$rdx^2 + 2sdxdy + tdy^2 = 0; \quad \tau = \pm\frac{\sqrt{s^2 - rt}}{1 + p^2 + q^2}:$$

34. Prove that the indicatrix at every point of a helicoid $z = c\tan^{-1}\frac{y}{x}$ is a rectangular hyperbola.
35. Show that the parametric curves are conjugate on the surface

$$x^3 = f(x^1) + f(x^2);$$

where x^1 and x^2 are parameters.
36. If a curve is a geodesic on the surface, prove that it is either a straight line or its principal normal is orthogonal to the surface at every point and conversely.
37. Prove that the normal curvature in a direction perpendicular to an asymptotic line is twice the mean normal curvature.
38. Prove that the sum of the normal curvatures in two directions at right angles is constant, and equal to the sum of the principal curvatures.
39. Prove that, at each non-umbilical point of a surface there exist two mutually orthogonal directions for which the normal curvature attains its extreme values.
40. Show that the helicoid

$$x^1 = u^1\cos u^2; \; x^2 = u^1\sin u^2; \; x^3 = au^2$$

is a surface of negative curvature.
41. If S is a surface of revolution

$$x^1 = u^1\cos u^2; \; x^2 = u^1\sin u^2; \; x^3 = f(u^1)$$

with $f(u^1)$ of class C^2; show that the points on a surface of revolution S for which $f'f'' > 0$ are elliptic; those for which $f'f'' < 0$ are hyperbolic; and if $f'' = 0$; then S is a cone.

42. Find the differential equation of the asymptotic curves of a surface of revolution

$$x^1 = u^1 \cos u^2;\ x^2 = u^1 \sin u^2;\ x^3 = f(u^1)$$

and determine the location of elliptic, parabolic, and hyperbolic points on this surface.

43. Prove that the co-ordinate curves of any allowable co-ordinate system on a surface S coincide with the lines of curvature if and only if for this system $a_{12} = 0 = b_{12}$ at any point where those co-ordinates are allowable.
44. Prove that at elliptic points the Gaussian curvature is positive, at parabolic ones it vanishes, and at hyperbolic ones it is negative.
45. Prove that if every point of a surface S which has a representation of class ≥ 3 is an umbilic, S is a plane or a sphere.
46. Show that on a cylinder $x^1 = c\cos u^1;\ x^2 = c\sin u^1;\ x^3 = u^2$ one distinct real asymptotic direction is given by u^1 curve.
47. Prove that the osculating plane to an asymptotic line which is not a straight line coincides with the tangent plane.
48. Prove that if every point of the surface S is parabolic, then S is developable.
49. Prove that any point of the following ellipsoid is elliptic

$$x^1 = a\cos u^1 \cos u^2;\ x^2 = b\sin u^1 \cos u^2;\ x^3 = c\sin u^2:$$

50. Prove that any point of a cylinder or of a cone is parabolic.
51. If S is a surface of revolution

$$x^1 = u^1 \cos u^2; x^2 = u^1 \sin u^2;\ x^3 = f(u^1)$$

with $f(u^1)$ of class C^2; show that a parallel surface $\overline{S}$ is also a surface of revolution.

52. Show that the meridians and parallels on a sphere form an isometric system, and determine the isometric parameters.
53. On the surface formed by the revolution of a parabola about its directrix, one principal curvature is double the other.
54. Find the surface of revolution for which

$$ds^2 = du^2 + (a^2 - u^2)dv^2:$$

55. If the surface of revolution is a minimal surface,

$$u\frac{d^2f}{du^2} + \frac{df}{du}\left[1 + \left(\frac{df}{du}\right)^2\right] = 0:$$

Hence, show that the only real minimal surface of revolution is that formed by the revolution of a catenary about its directrix.

56. Find the equations of the helicoid generated by a circle of radius a, whose plane passes through the axis; and determine the lines of curvature on the surface.
57. Prove that the null lines are conjugate if and only if the surface is minimal.
58. Show that the only minimal surface of the type $z = f(x) + F(y)$ is the surface $az = \log\cos(ax) - \log\cos(ay)$.
59. Show that the surface $\sin az = \sinh ax \sinh ay$ is minimal.
60. Show that the surfaces

$$x^1 = u^1\cos u^2;\ \ x^2 = u^1\sin u^2;\ \ x^3 = cu^2$$

and

$$x^1 = u^1\cos u^2;\ \ x^2 = u^1\sin u^2;\ \ x^3 = c\cosh^{-1}\frac{u^1}{c}$$

are applicable.
61. Find the first and second curvatures for the spherical image.
62. Prove that the osculating planes of a line of curvature and of its spherical image at corresponding points are parallel.
63. Show that the lines of curvature of a surface of revolution remains isometric in their spherical representation.
64. Show that the spherical images of the asymptotic lines on a minimal surface, as well as the asymptotic lines themselves, are an isometric system.
65. Show that the helicoids

$$x^1 = u^1\cos u^2;\ \ x^2 = u^1\sin u^2;\ \ x^3 = cu^2 + c\int\sqrt{\frac{(u^1)^2 + c^2}{(u^1)^2 - c^2}}\,\frac{du^1}{u^1}$$

is a minimal surface.

Chapter 9
Complex Numbers

9.1 Definition and Properties of Complex Numbers

9.1.1 Imaginary Numbers

The square of positive as well as negative real numbers is always a positive real number. For example, $3^2 = (\quad 3)^2 = 9$. The root of a positive number is therefore a positive or negative number. We now introduce a new type of number whose square always gives a negative real number: they are called *imaginary numbers*.

Definition The unit of imaginary numbers is the number j with the property that

$$\mathrm{j}^2 = \quad 1 \qquad (9.1)$$

The imaginary unit j corresponds to 1 for real numbers.

An arbitrary imaginary number is made up of the imaginary unit j and any real number y; thus $y\mathrm{j}$ is the general form for an imaginary number.

We know that we cannot extract the root of a negative number when dealing with real numbers; nevertheless, we can factorise the root of a negative number thus:

$$\overline{\quad 5} = \quad \overline{5(\quad 1)} = \quad \overline{5} \times \quad \overline{\quad 1}$$

Since $\mathrm{j}^2 = \quad 1$, it follows that $\mathrm{j} = \quad \overline{\quad 1}$; hence $\quad \overline{5} \times \quad \overline{\quad 1} = \mathrm{j} \quad \overline{5}$.

The root of a negative number is an imaginary number. Moreover, with $\mathrm{j}^2 = \quad 1$, we can simplify higher powers of j.

K. Weltner, W. J. Weber, J. Grosjean, P. Schuster, *Mathematics for Physicists and Engineers*
ISBN 978-3-642-00172-7 ' Springer 2009

The ordinary rules of algebra extend to imaginary numbers. In addition one must remember that $j^2 = -1$, i.e.

$$
\begin{aligned}
j^1 &= j\\
j^2 &= -1\\
j^3 &= j^2 \times j = -j\\
j^4 &= j^2 \times j^2 = 1
\end{aligned}
$$

9.1.2 Complex Numbers

The sum z of a real number x and an imaginary number jy is called a *complex number*, complex meaning "composed". Thus

$$z = x + jy$$

where x is the real part of z and y is the imaginary part of z.

If we replace j by $-j$, we obtain a different complex number z^* given by

$$z^* = x - jy$$

z^* is called the *complex conjugate* of z.

A complex number has the value zero only if both the real part and the imaginary part are zero.

9.1.3 Fields of Application

The most obvious property of imaginary numbers is that we can "extract" the root of a negative number, which means we obtain an expression which we can handle. This property enables us, in principle, to solve equations of any degree. Consider, for instance, the quadratic equation

$$ax^2 + bx + x = 0$$

We have seen that it is unsolvable in terms of real x if $b^2 < 4ac$, i.e. when the radicand $b^2 - 4ac$ becomes negative. The solution in terms of complex x is

$$x_{1,2} = \frac{1}{2a}\left(-b - j\sqrt{4ac - b^2}\right)$$

Complex numbers are important in the solution of differential equations which are dealt with in Chap. 10. They are also a useful concept in electrical engineering, and they are indispensable in the study of quantum physics.

9.1.4 Operations with Complex Numbers

When handling complex numbers there are two rules to remember:

The complex number $x + jy$ is zero if and only if $x = 0$ and $y = 0$.

Complex numbers obey the ordinary rules of algebra, in addition $j^2 = \quad 1$.

Addition and Subtraction of Complex Numbers

Rule The sum of complex numbers is obtained by adding real and imaginary parts separately.
The difference of complex numbers is obtained by subtracting real and imaginary parts separately.

Let	*Example*
$z_1 = x_1 + jy_1$	$z_1 = 6 + 7j$
$z_2 = x_2 + jy_2$	$z_2 = 3 + 4j$
Then their sum is	
$z_1 + z_2 = (x_1 + jy_1) + (x_2 + jy_2)$	$z_1 + z_2 = (6 + 7j) + (3 + 4j)$
$= (x_1 + x_2) + j(y_1 + y_2)$	$= 9 + 11j$
Their difference is	*Example*
$z_1 \quad z_2 = (x_1 \quad x_2) + (y_1 \quad y_2)j$	$z_1 \quad z_2 = 3 + 3j$

Product of Complex Numbers

Rule The product z_1, z_2 of two complex numbers is obtained by simple multiplication of the terms, taking into account $j^2 = \quad 1$.

General expression	*Example*
$z_1 z_2 = (x_1 + jy_1)(x_2 + jy_2)$	$z_1 z_2 = (6 + 7j)(3 + 4j)$
$= x_1 x_2 + jx_1 y_2 + jy_1 x_2 + j^2 y_1 y_2$	$= 18 + 24j + 21j \quad 28$
$= (x_1 x_2 \quad y_1 y_2) + j(x_1 y_2 + x_2 y_1)$	$= \quad 10 + 45j$

Division of Complex Numbers

Rule Division of a complex number by another complex number is carried out by multiplying numerator and denominator by the conjugate of the divisor to transform the latter into a real number. Note that the denominator then appears as the sum of two squares.

General expression

$$\frac{z_1}{z_2} = \frac{x_1 + jy_1}{x_2 + jy_2} = \frac{(x_1 + jy_1)(x_2 - jy_2)}{(x_2 + jy_2)(x_2 - jy_2)}$$

$$= \frac{(x_1x_2 + y_1y_2) - j(x_1y_2 - y_1x_2)}{x_2^2 + y_2^2}$$

The conjugate of z_2 is $z_2^* = x_2 - jy_2$

Example

$$\frac{z_1}{z_2} = \frac{6 + 7j}{3 + 4j} = \frac{(6 + 7j)(3 - 4j)}{(3 + 4j)(3 - 4j)}$$

$$= \frac{10 - 3j}{25} = \frac{2}{5} - \frac{3}{25}j$$

9.2 Graphical Representation of Complex Numbers

9.2.1 Gauss Complex Number Plane: Argand Diagram

The complex number $z = x + jy$ can be represented in an $x - y$ coordinate system by placing the real part along the x-axis and the imaginary part along the y-axis in a similar way to the components of a vector. Figure 9.1 shows this.

We obtain a point P(z) in this plane which corresponds to the complex number z. This plane is called the Gauss number plane, better known as the *Argand diagram*. In this way we have produced a geometrical picture of a complex number.

Example Where is the point in the Argand diagram corresponding to the complex number $z = 4 - 2j$? Figure 9.2 shows the answer.

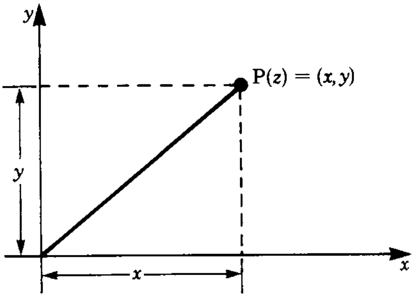

Fig. 9.1

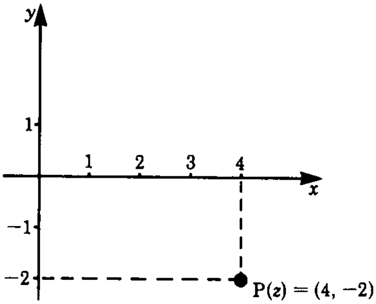

Fig. 9.2

Addition

$$z_3 = z_1 + z_2$$

If $z_1 = 6 + 3j$ and $z_2 = 2 + 5j$ we know that the sum $z_3 = 8 + 8j$. This addition may be represented in an Argand diagram, as shown in Fig. 9.3.

To obtain the complex number z_3, we first draw the two complex numbers z_1 and z_2 as vectors as shown. Then we shift vector z_2 parallel to itself in such a way that its tail is made to coincide with the tip of vector z_1. The tip of vector z_2 locates the tip of the required vector or complex number z_3 whose magnitude is given by the length OP. This construction is based on the parallelogram rule for the addition of vectors. The same result is obtained if we shift z_1 instead of z_2. We dealt with vectors in Chap. 1, so this construction should not be new to the reader.

To add more than two complex numbers we draw a polygon by joining the vectors tip to tail. The tip of the closing vector represents the required complex number.

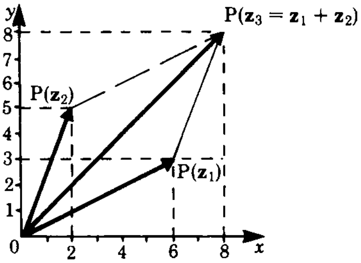

Fig. 9.3

Subtraction

$$z_3 = z_1 - z_2 = z_1 + (-z_2)$$

The problem of the subtraction of two complex numbers can be transformed into one of addition if to one complex number we add the negative of the other.

9.2.2 Polar Form of a Complex Number

Instead of specifying a complex number by means of the coordinates x and y, we could specify it by means of a distance r from the origin of the coordinates and an angle , as shown in Fig. 9.4 x and y are the Cartesian coordinates, and r and are the polar coordinates.

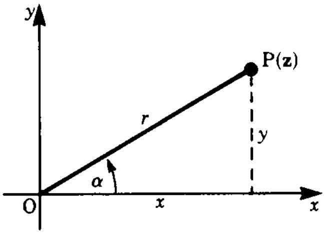

Fig. 9.4

From the figure we see that

$$x = r\cos$$
$$y = r\sin$$

Substituting in the expression for the complex number, $z = x + jy$, gives

$$z = r(\cos \quad + j\sin \quad) \tag{9.2}$$

This expresses a complex number in terms of the trigonometric functions. It follows that the conjugate z of the complex number z is

$$z \;= r(\cos \quad j\sin \quad)$$

If we know r and we can calculate x and y from the above equations. If on the other hand we know x and y and we want to express a complex number in polar form, it follows from Fig. 9.4 that

$$r = \sqrt{x^2 + y^2} \quad \text{taking the positive value only}$$
$$= \tan^{\;1} \frac{y}{x}$$

r is known as the *modulus* and as the *argument*.

Table 9.1 illustrates the values taken by according to the sign of x and y.

Table 9.1

x	y	tan	P(z) lies in	in the range
positive	positive	positive	1st quadrant	0 to $\frac{}{2}$
negative	positive	negative	2nd quadrant	$\frac{}{2}$ to
negative	negative	positive	3rd quadrant	to $\frac{3}{2}$
positive	negative	negative	4th quadrant	$\frac{3}{2}$ to 2

Example Express the complex number $z = 1 \quad j$ in polar form.
$z = 1 \quad j$ means that $x = 1$ and $y = \quad 1$.

Hence $$r = \overline{x^2 + y^2} = \overline{1+1} = \overline{2}$$
$$= \tan^{1} \frac{1}{1} = \tan^{1}(\quad 1) = \frac{3}{4} \quad \text{or} \quad \frac{7}{4}$$

We represent the complex number in the Argand diagram (Fig. 9.5) and find that it lies in the fourth quadrant. It follows then that

$$= \frac{7}{4}$$

Thus the complex number $z = 1 \quad j$ can be written in polar form:

$$z = \overline{2} \quad \cos\frac{7}{4} + j\sin\frac{7}{4}$$

Example Express the complex number $z = 6j$ in polar form.
In this case, $z = x + jy \quad 0 + 6j$, i.e. $x = 0$. Consequently $r = 6$

$$\tan \quad = \frac{y}{x} = \frac{6}{0} = d, \quad \text{i.e.} \quad = \frac{}{2} \quad \text{or} \quad \frac{3}{2}$$

To determine , the complex number is shown in Fig. 9.6. It follows, therefore, that $= \frac{}{2}$. In polar form the complex number is

$$z = 6 \quad \cos\frac{}{2} + j\sin\frac{}{2}$$

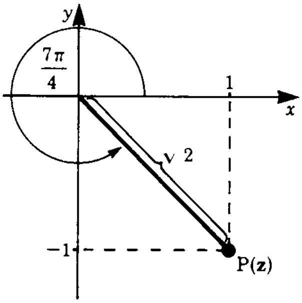

Fig. 9.5

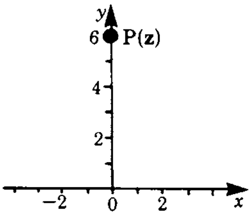

Fig. 9.6

9.3 Exponential Form of Complex Numbers

9.3.1 Euler's Formula

It is also possible to express a complex number in the form

$$z = re^{j} \tag{9.3}$$

As before, r is the modulus of z and is the argument. We can show that it is equivalent to the polar form of a complex number:

$$z = r(\cos \quad + j\sin \quad)$$

In other words we want to prove that

$$re^{j} \quad = r(\cos \quad + j\sin \quad)$$

Dividing by r gives

$$e^{j} \quad = \cos \quad + j\sin$$

In Chap. 8, Sect. 8.2, we showed that the expansion for e^x is

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

This expansion remains valid if we replace x by j . Remembering that $j^2 = \quad 1$, we obtain

$$e^{j} \quad = 1 + j \qquad \frac{{}^2}{2!} \quad j\frac{{}^3}{3!} + \frac{{}^4}{4!} + j\frac{{}^5}{5!}\cdots$$

We also showed that

$$\sin \quad = \qquad \frac{{}^3}{3!} + \frac{{}^5}{5!} \quad \cdots$$

Hence

$$j\sin \quad = j \quad j\frac{{}^3}{3!} + j\frac{{}^5}{5!} \quad \cdots$$

$$\cos \quad = 1 \quad \frac{{}^2}{2!} + \frac{{}^4}{4!} \quad \cdots$$

Comparing these expressions, we obtain

$$\cos \quad + j\sin \quad = 1 + j \qquad \frac{{}^2}{2!} \quad j\frac{{}^3}{3!} + \frac{{}^4}{4!} + j\frac{{}^5}{5!} \quad \cdots = e^{j}$$

Euler’s formula

$$e^{j} = \cos \quad + j \sin \qquad (9.4)$$

This enables us to express a complex number in a third way.

A table giving the various forms in which complex numbers can be expressed will be found in the appendix at the end of this chapter.

9.3.2 Exponential Form of the Sine and Cosine Functions

The conjugate complex number of e^{j} is obtained by replacing j by j, so that if $z = e^{j}$ then $z = e^{\ j}$.

Now we will try to express the sine and cosine function in terms of e^{j} .

According to Euler s formula, we have

$$e^{j} = \cos \quad + j \sin$$
$$e^{\ j} = \cos \quad j \sin$$

Adding both equations gives

$$\cos \quad = \frac{1}{2}(e^{j} + e^{\ j}) = \cosh j$$

Subtracting both equations gives

$$\sin \quad = \frac{1}{2j}(e^{j} \quad e^{\ j}) = \frac{1}{j}\sinh j$$

9.3.3 Complex Numbers as Powers

Given the complex number $z = x + jy$, we wish to calculate the modulus and argument of

$$w = e^{z}$$

Substituting for z, we have

$$w = e^{(x+jy)} = e^{x} e^{jy} \quad r e^{j}$$

Comparing the last two expressions, it follows that $r = e^{x}$ and $jy = j$. Hence the modulus of w is e^{x}, and the argument of w is y.

Example If $z = 2 + j \ /2$, calculate $w = e^{z}$.

$$w = e^{z} = e^{(2+j \ /2)} = e^{2} e^{j \ /2}$$

Hence the modulus is $r = e^2$ and the argument is $\ = \ /2$, and we have

$$w = e^2 \left(\cos \frac{}{2} + j \sin \frac{}{2} \right) = j e^2$$

The solution is shown in the Argand diagram (Fig. 9.7).

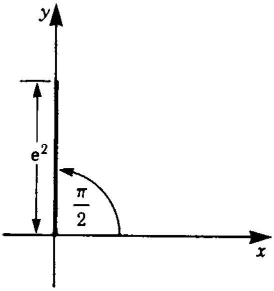

Fig. 9.7

Example If $z = \ 2 + j3 \ /4$, calculate $w = e^z$.

$$w = e^{\ 2 + j\frac{3}{4}} = e^{\ 2} e^{j\frac{3}{4}}$$

$$r = e^{\ 2} \quad \text{and} \quad = \frac{3}{4}$$

Hence $$w = e^{\ 2} \left(\cos \frac{3}{4} + j \sin \frac{3}{4} \right)$$

$$= e^{\ 2} \left(\frac{1}{\sqrt{2}} + j \frac{1}{\sqrt{2}} \right)$$

This solution is shown in the Argand diagram (Fig. 9.8).

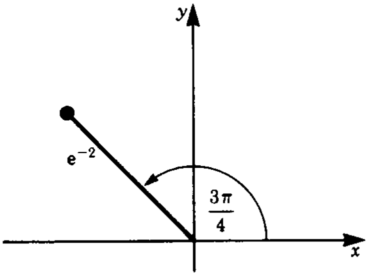

Fig. 9.8

Suppose now that z is a function of some parameter t. The simplest case is that of the linear function where a and b are constants.

$$z(t) = at + jbt$$

One important interpretation of t in practice is the time, so that $z(t)$ grows with time, i.e. the real part and the imaginary part grow with time.

Substituting for z in $w = e^z$ gives $w(t) = e^{(at+jbt)} = e^{at}e^{jbt}$

Using Euler s formula, we get $w(t) = e^{at}(\cos bt + j \sin bt)$

To examine the behaviour of this function, we can consider the real and the imaginary parts separately and represent each one graphically as a function of the time t.

The real part of $w(t)$ is $e^{at} \cos bt$. It is the product of an exponential function and a trigonometric function of period $p = 2\quad /b$.

If a is positive then $w = e^{at} \cos bt$ represents a vibration whose amplitude grows exponentially with time, as shown in Fig. 9.9.

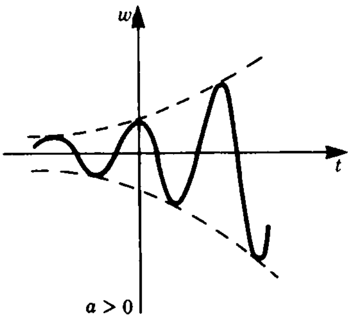

Fig. 9.9

If a is negative then $w = e^{at} \cos bt$ represents a vibration whose amplitude decreases exponentially with time; the vibration is said to be damped and is shown below in Fig. 9.10.

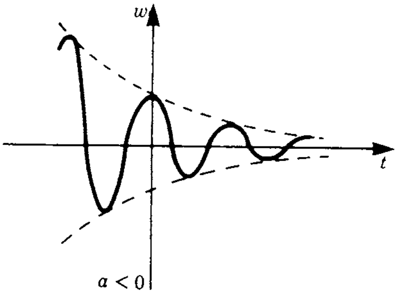

Fig. 9.10

The imaginary part of $w(t)$ is $e^{at}\sin bt$. It is also the product of an exponential function and a trigonometric function whose graphical representation is similar to Fig. 9.9 or Fig. 9.10, depending on the sign of a.

The mathematical solution of vibration problems is often simplified by means of complex numbers. Thus, in solving a practical problem, we start by considering the physical situation consisting of real quantities. We then perform all calculations using complex numbers and finally consider and interpret the results of the real and imaginary parts.

9.3.4 Multiplication and Division in Exponential Form

Addition and subtraction of complex numbers are best carried out using the form $z = x + jy$. Multiplication and division, on the other hand are best carried out by expressing the complex numbers either in exponential form or in polar form.

Consider two complex numbers

$$z_1 = r_1 e^{j\varphi_1} \quad \text{and} \quad z_2 = r_2 e^{j\varphi_2}$$

Multiplying gives

$$z = z_1 z_2 = r_1 e^{j\varphi_1} r_2 e^{j\varphi_2} = r_1 r_2 e^{j(\varphi_1 + \varphi_2)} \tag{9.5}$$

Here we use the power rule $a^n a^m = a^{n+m}$.

Dividing gives

$$z = \frac{z_1}{z_2} = \frac{r_1 e^{j\varphi_1}}{r_2 e^{j\varphi_2}} = \frac{r_1}{r_2} e^{j(\varphi_1 - \varphi_2)} \tag{9.6}$$

Here we use the power rule $a^n / a^m = a^{n-m}$.

Rule To multiply (or divide) complex numbers, we multiply (or divide) the moduli and add (or subtract) the arguments.

9.3.5 *Raising to a Power, Exponential Form*

We have

$$z^n = (re^{j\alpha})^n = r^n e^{jn\alpha} \tag{9.7}$$

Rule To raise a complex number to a given power we raise the modulus to that power and multiply the argument by that power.

Figure 9.11 shows the points z and z^2 in the complex plane with $r = 2$.

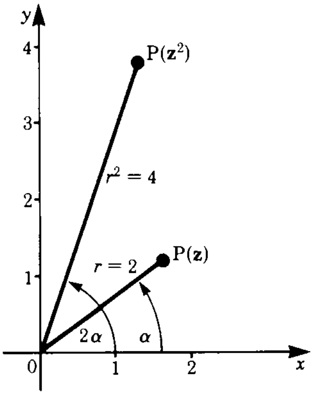

Fig. 9.11

In the case of $z^{1/n}$ we have

$$z^{1/n} = \sqrt[n]{z} = \sqrt[n]{re^{j\alpha}} = \sqrt[n]{r}\, e^{j\alpha/n} \tag{9.8}$$

Rule To extract the root of a complex number, we find the root of the modulus and divide the argument by the index.

9.3.6 *Periodicity of* $re^{j\alpha}$

We should like to mention the fact, perhaps surprising to the reader, that the complex number

$$z = re^{j\alpha}$$

is identical to

$$z = re^{j(\alpha+2\pi)}$$

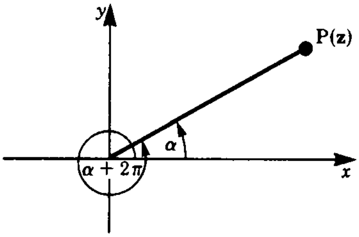

Fig. 9.12

If we examine Fig. 9.12 we can see that the same point P(z) is obtained whether the angle is α or $\alpha + 2\pi$. In fact we could equally take the angle to be $\alpha + 4\pi$, $\alpha + 6\pi$, $\alpha - 2\pi$, $\alpha - 4\pi$, etc. Hence, generally

$$re^{j\alpha} = re^{j(\alpha + 2\pi k)}$$

where $k = \pm 1, \pm 2, \pm 3$ etc.

In particular we have $1 = e^{j2\pi}$

9.3.7 *Transformation of a Complex Number From One Form into Another*

Transformation from the Algebraic Form into the Exponential Form

The transformation from $x + jy$ into $re^{j\alpha}$ is based on the relationships derived in Sect. 9.2.2, i.e.

$$r = \sqrt{x^2 + y^2}, \quad \tan\alpha = \frac{y}{x}$$

Example Convert the complex number $z = -\sqrt{5} + 2j$ to the exponential form.

$$r = \sqrt{(-\sqrt{5})^2 + 2^2} = 3$$

$$\tan\alpha = \frac{2}{-\sqrt{5}} = -0.894$$

The angle is in the second quadrant; therefore $\alpha = 138.19°$ or 0.768π radians.

Transformation of the Exponential Form into the Algebraic Form

Since $z = re^{j\alpha} = r(\cos\alpha + j\sin\alpha) = x + jy$

then $x = r\cos\alpha$

$y = r\sin\alpha$

Example Convert the expression $z = e^{(0.5+1.3j)}$ to the algebraic form.

$$\begin{aligned} z &= e^{(0.5+1.3j)} = e^{0.5}e^{1.3j} \\ x &= e^{(0.5)}\cos 1.3 = 0.441 \\ y &= e^{(0.5)}\sin 1.3 = 1.589 \end{aligned}$$

Hence $z = 0.441 + 1.589j$

9.4 Operations with Complex Numbers Expressed in Polar Form

In the previous section, the rules for operations with complex numbers have been derived using the exponential form. We will now show that the same rules are obtained if we express the complex numbers in polar form. We make use of the addition theorems for trigonometric functions given in Chap. 3, Sect. 3.6.6.

9.4.1 Multiplication and Division

Let z_1 and z_2 be expressed in polar form so that

$$\begin{aligned} z_1 &= r_1(\cos\ {}_1 + j\sin\ {}_1) \\ z_2 &= r_2(\cos\ {}_2 + j\sin\ {}_2) \end{aligned}$$

Multiplication

$$\begin{aligned} z_1 z_2 &= r_1 r_2(\cos\ {}_1 + j\sin\ {}_1)(\cos\ {}_2 + j\sin\ {}_2) \\ &= r_1 r_2[(\cos\ {}_1 \cos\ {}_2 \quad \sin\ {}_1 \sin\ {}_2) + j(\sin\ {}_1 \cos\ {}_2 + \cos\ {}_1 \sin\ {}_2)] \end{aligned}$$

Using the addition formulae for the sine and cosine functions,

$$\begin{aligned} z_1 z_2 &= r_1 r_2[\cos(\ {}_1 + \ {}_2) + j\sin(\ {}_1 + \ {}_2)] \\ &= r(\cos\ \ + j\sin\ \ ) \end{aligned} \tag{9.9}$$

Thus the modulus r of the product equals $r_1 r_2$, and the argument is $(\ {}_1 + \ {}_2)$.

This is exactly in accordance with the rule derived in the previous section: to multiply complex numbers we multiply the moduli and add the arguments.

Figure 9.13 illustrates geometrically the multiplication of complex numbers. Draw a triangle OPP_1 similar to the triangle OP_2Q (Q has coordinates 1, 0), then

$$\frac{z}{z_1} = \frac{z_2}{1}, \quad \text{hence} \quad z = z_1 z_2$$

Also = angle QOP = ${}_1 + {}_2$

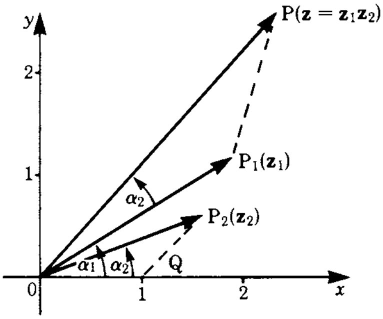
y
2
1
0
1
2
x
$P(\mathbf{z} = \mathbf{z}_1\mathbf{z}_2)$
$P_1(\mathbf{z}_1)$
$P_2(\mathbf{z}_2)$
Q
α_2
α_1
α_2

Fig. 9.13

Division

Both the numerator and the denominator are multiplied by the conjugate of the divisor.

$$\frac{z_1}{z_2} = \frac{r_1(\cos\ {}_1 + j\sin\ {}_1)}{r_2(\cos\ {}_2 + j\sin\ {}_2)}\frac{r_2(\cos\ {}_2 \quad j\sin\ {}_2)}{r_2(\cos\ {}_2 \quad j\sin\ {}_2)}$$

Using the addition formulae for the sine and cosine functions,

$$\frac{z_1}{z_2} = \frac{r_1}{r_2}[\cos(\ {}_1 \quad {}_2) + j\sin(\ {}_1 \quad {}_2)] \tag{9.10}$$

Thus the modulus of the quotient equals $\frac{r_1}{r_2}$, and the argument is $(\ {}_1 \quad {}_2)$.

Again this is in accordance with the rule derived in the previous section: to divide complex numbers we divide the moduli and subtract the arguments.

Figure 9.14 illustrates the division of complex numbers by a reasoning similar to that for multiplication.

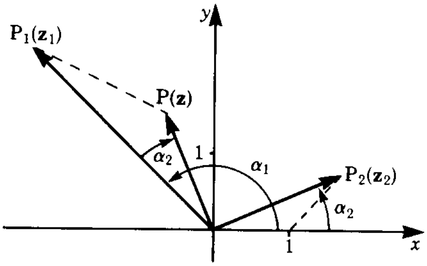
$P_1(\mathbf{z}_1)$
y
$P(\mathbf{z})$
α_2
1
α_1
$P_2(\mathbf{z}_2)$
α_2
1
x

Fig. 9.14

9.4.2 Raising to a Power

We saw earlier that $z_1 z_2 = r_1 r_2[\cos(\ _1 + \ _2) + j\sin(\ _1 + \ _2)]$

If we now let $z_1 = z_2 = z, r_1 = r_2 = r$ and $\ _1 = \ _2 = \ $ it follows that

$$z^2 = r^2(\cos 2\ \ + j\sin 2\ \ )$$

Similarly

$$z^3 = r^2(\cos 2\ \ + j\sin 2\ \ )r(\cos\ \ + j\sin\ \ )$$

$$= r^3(\cos 3\ \ + j\sin 3\ \ )$$

The general expression is

$$z^n < \ = r^n(\cos n\ \ + j\sin n\ \ ) \tag{9.11}$$

Rule To raise a complex number to a given power, we raise the modulus to that power and multiply the argument by that power.

By setting $r = 1$ we have

$$(\cos\ \ + j\sin\ \ )^n = \cos n\ \ + j\sin n$$

This is known as *De Moivre's theorem*.

9.4.3 Roots of a Complex Number

De Moivre s theorem holds true for *positive, negative* and *fractional powers*. We can, therefore, use this fact to determine all the distinct roots of any number.

Since $x + jy = r(\cos\ \ + j\sin\ \ )$, then, by De Moivre s theorem, it follows that

$$\sqrt[n]{x + jy} = \sqrt[n]{r}\ \ \cos\frac{\ }{n} + j\sin\frac{\ }{n}$$

However, using this equation, we obtain one root only. In order to obtain all the roots we must consider the fact that the cosine and sine functions are periodic functions of period $2\ $ radians or $360\ $. Thus we can write

$$(\cos\ \ + j\sin\ \ )^n = [\cos(\ \ + 2\ \ k) + j\sin(\ \ + 2\ \ k)]^n$$

$$= \cos(n\ \ + 2\ \ nk) + j\sin(n\ \ + 2\ \ nk)$$

$$\text{where}\quad k = 0, -1, -2, -3, \bullet\bullet\bullet$$

When raising a complex number to an integral power there is no ambiguity: the result is independent of periodicity. But extracting the roots of a complex number means raising it to a fractional power. Now periodicity becomes important, and we have

$$\sqrt[n]{x + jy} = \sqrt[n]{r}\ \cos\ \ \frac{}{n} + \frac{2}{n}k\ \ + j\sin\ \ \frac{k}{n} + \frac{2}{n}k$$

$$\text{where}\quad k = 0, -1, -2, \bullet\bullet\bullet \qquad (9.12)$$

Rule The nth roots of a complex number are obtained by extracting the nth root of the modulus and dividing the argument by n. Due to the periodicity of the trigonometric functions, there are n solutions.

By giving k the values $0, 1, 2, 3, \bullet\bullet\bullet, (n \quad 1)$, we obtain the n different roots of a complex number; for example,

$$k = 0, \quad z_1 = \sqrt[n]{r}\ \ \cos\frac{}{n} + j\sin\frac{}{n}$$

$$k = 1, \quad z_2 = \sqrt[n]{r}\ \ \cos\ \ \frac{}{n} + \frac{2}{n}\ \ + j\sin\ \ \frac{}{n} + \frac{2}{n}$$

$$k = 2 \quad, z_3 = \sqrt[n]{r}\ \ \cos\ \ \frac{}{n} + \frac{4}{n}\ \ + j\sin\ \ \frac{}{n} + \frac{4}{n}$$

and so on. With $k = n$, we would obtain the same value as with $k = 0$; also we would not obtain any new values for $k > n$ or $k = \quad 1, \quad 2, \quad 3, \bullet\bullet\bullet$. The root corresponding to $k = 0$ is called the *principal value*.

Example Calculate the four roots of

$$z^4 = \cos\frac{2}{3} + j\sin\frac{2}{3}$$

In this case $r = 1$ and $n = 4$; hence

$$z^4 = \cos\ \ \frac{}{6} + \frac{2}{4}k\ \ + j\sin\ \ \frac{}{6} + \frac{2}{4}k$$

$$= \cos\ \ \frac{}{6} + \frac{}{2}k\ \ + j\sin\ \ \frac{}{6} + \frac{}{2}k$$

The roots are

$$k = 0, \quad z_1 = \cos\frac{}{6} + j\sin\frac{}{6} = \frac{1}{2}\ \ \overline{3} + \frac{1}{2}j$$

$$k = 1, \quad z_2 = \cos\frac{4}{6} + j\sin\frac{4}{6} = \quad \frac{1}{2} + \frac{\overline{3}}{2}j$$

$$k = 2, \quad z_3 = \cos\frac{7}{6} + j\sin\frac{7}{6} = \quad \frac{\overline{3}}{2} \quad \frac{1}{2}j$$

$$k = 3, \quad z_4 = \cos\frac{10}{6} + j\sin\frac{10}{6} = \frac{1}{2} \quad \frac{\overline{3}}{2}j$$

The roots z_1 and z_3, as well as z_2 and z_4, are opposite to each other; moreover, we can see that the argument, starting from the principal value, is successively increased by $/2$. Figure 9.15 shows all the values on a circle of radius 1. They form a square.

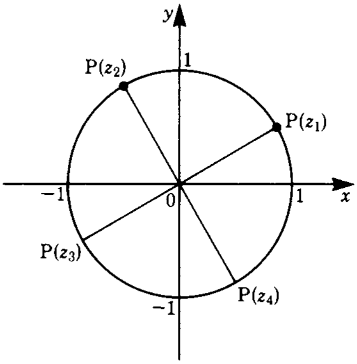

Fig. 9.15

All the nth roots of a complex number of modulus 1 have modulus 1. When depicted in an Argand diagram they form the vertices of a regular n-sided polygon inscribed in a unit circle.

Appendix

Summary of operations with complex numbers

Designation	Formulae
Imaginary unit j Imaginary number	$j^2 = \quad 1$ $\quad = jy \; (y \text{ real})$
Complex number z in arithmetic form	$z = x + jy \quad (x, y \text{ real})$ x = real part y = imaginary part
Complex conjugate	$z \; = x \quad jy$
Complex numbers in polar form	$z = r(\cos \quad + j \sin \quad)$
Transformation $(x, y) \quad (r, \quad)$	$x = r \cos \qquad r = \quad \overline{x^2 + y^2}$ $y = r \sin \qquad \tan \quad = y/x$
Complex number in exponential form Euler s formula	$z = r e^{j}$ $e^{j} \quad = \cos \quad + j \sin$
Exponential form for cosine and sine functions	$\cos \quad = \frac{1}{2}(e^{j} \quad + e^{\; j} \;) = \cosh j$ $\sin \quad = \frac{1}{2j}(e^{j} \qquad e^{\; j} \;) = \frac{1}{j} \sinh j$
Periodicity of complex numbers	$z \; = r e^{j}$ $\quad = r e^{j(\; + 2k \;)} \quad (k = -1, -2, -3, \bullet\bullet\bullet)$
Multiplication and division in exponential form	$z_1 = r_1 e^{j \; 1}, z_2 = r_2 e^{j \; 2}$ $z_1 z_2 = r_1 r_2 e^{j(\; 1 + \; 2)}$ $\frac{z_1}{z_2} = \frac{r_1}{r_2} e^{j(\; 1 \quad 2)}$
Raising to a power and extracting roots in exponential form	$z = r e^{j}$ $z^n = r^n e^{jn}$ ${}^n \overline{z} = {}^n \overline{r} e^{j[(\; + 2 \; k)/n]} (k = 0, -1, -2, \bullet\bullet\bullet)$
Multiplication and division in polar form	$z_1 = r_1(\cos \; {}_1 + j \sin \; {}_1)$ $z_2 = r_2(\cos \; {}_2 + j \sin \; {}_2)$ $z_1 z_2 = r_1 r_2[\cos(\; {}_1 + \; {}_2) + j \sin(\; {}_1 + \; {}_2)]$ $\frac{z_1}{z_2} = \frac{r_1}{r_2}[\cos(\; {}_1 \quad {}_2) + j \sin(\; {}_1 \quad {}_2)]$
Raising to a power and extracting roots in polar form	$z = r(\cos \quad + j \sin \quad)$ $z^n = r^n[\cos n \quad + j \sin n \quad]$ ${}^n \overline{z} = {}^n \overline{r} \quad \cos \quad \frac{}{n} + \frac{2 \; k}{n} \quad + j \sin \quad \frac{}{n} + \frac{2 \; k}{n}$ $(k = 0, -1, -2, \bullet\bullet\bullet)$

Exercises

9.1 Definition and Properties of Complex Numbers

1. Express the following in terms of j:

   (a) $\overline{4\quad 7}$ (b) $\overline{\quad 144}$

   (c) $\frac{\overline{5}}{\overline{\quad 4}}$ (d) $\overline{4(\quad 25)}$

2. Compute

   (a) j^8 (b) j^{15}

   (c) j^{45} (d) $(\quad j)^3$

3. Evaluate

   (a) $\overline{\quad 48} + \overline{\quad 75} \qquad \overline{\quad 27}$ (b) $\overline{\quad 12} \qquad \overline{\quad 8} + \overline{\quad 0.6}$

   (c) $\overline{\quad 3}\ \overline{\quad 3}$ (d) $\overline{\quad a}\ \overline{+b}$

   (e) $5j^3 2j^6$ (f) $(\quad j)^3 j^2$

   (g) $8j/2j$ (h) $1/j\ ^3$

   (i) $6j/j^7 \quad \overline{3}$ (j) $1/j^5 + 1/j^7$

   (k) $\overline{b\quad a}\ \overline{a\quad b}$ (l) $\frac{\overline{\quad 3}\ \overline{12}}{j\quad \overline{\quad a^2}}$

4. Determine the imaginary part of z:

   (a) $z = 3 + 7j$

   (b) $z = 15j \quad 4$

5. Determine the conjugate complex number z of z:

   (a) $z = 5 + 2j$

   (b) $z = \frac{1}{2} \qquad \overline{3}j$

6. Evaluate the (complex) roots of the following quadratic equations:

   (a) $x^2 + 4x + 13 = 0$

   (b) $x^2 + \frac{3}{2}x + \frac{25}{16} = 0$

7. Calculate the sum $z_1 + z_2$:

   (a) $z_1 = 3 \quad 2j$ (b) $z_1 = \frac{3}{4} + \frac{3}{4}j$

   $z_2 = 7 + 5j$ $z_2 = \frac{3}{4} \quad \frac{3}{4}j$

8. Compute $w = z_1 \quad z_2 + z_3$:

   (a) $z_1 = 5 \quad 2j$
   $z_2 = 2 \quad 3j$
   $z_3 = \quad 4 + 6j$

   (b) $z_1 = 4 \quad 3.5j$
   $z_2 = 3 + 2j$
   $z_3 = 7.5j$

9. Compute the product $w = z_1 z_2$:

   (a) $z_1 = 1 + j$
   $z_2 = 1 \quad j$

   (b) $z_1 = 3 \quad 2j$
   $z_2 = 5 + 4j$

10. Determine

   (a) $(16 + j \quad \overline{2})/2 \quad \overline{2}$

   (b) $(4 \quad j \quad \overline{3})/2j$

   (c) $(2 + 3j)/(2 \quad 4j)$

   (d) $1/(1 + j)$

   (e) $\dfrac{1 + j}{1 \quad j} \quad \dfrac{1 \quad j}{1 + j}$

   (f) $\dfrac{(5 + j \quad \overline{3})(5 \quad j \quad \overline{3})}{2 \quad j \quad \overline{3}}$

11. Convert the following sums into products:

   (a) $4x^2 + 9y^2$

   (b) $a + b$

9.2 Graphical Representation of Complex Numbers

12. Plot each point z_i and $\quad z_i$ in the complex number plane:

   (a) $z_1 = \quad 1 \quad j$

   (b) $z_2 = 3 + 2j$

   (c) $z_3 = 5 + 3j$

   (d) $z_4 = \dfrac{3}{2}j$

   (e) $z_5 = \quad 3 + \dfrac{1}{2}j$

   (f) $z_6 = \quad \overline{2}$

13. Using Fig. 9.16, determine the real and imaginary parts of each point $z_1, z_2, \cdots, z_6$.

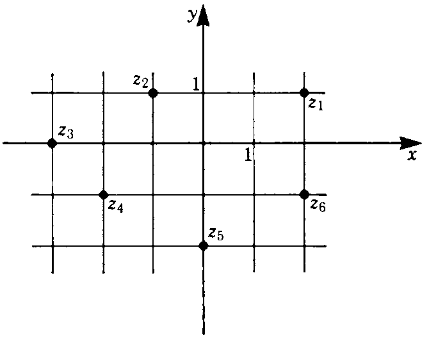

Fig. 9.16

14. Convert the complex number $z = x + jy$ to the polar form $z = r(\cos\ \ + j\sin\ \ )$:

(a) $z = j\ \ 1$

(b) $z = \ \ (1 + j)$

15. Transform the complex number $z = r(\cos\ \ + j\sin\ \ )$ into the form $z = x + jy$:

(a) $z = 5\ \ \cos\frac{\ }{3}\ \ \ j\sin\frac{\ }{3}$

(b) $z = 4(\cos 225\ \ + j\sin 225\ \ )$

16. Compute $z_1 z_2$:

(a) $z_1 = 2(\cos 15\ \ + j\sin 15\ \ )$
$z_2 = 3(\cos 45\ \ + j\sin 45\ \ )$

(b) $z_1 = \ \ \overline{5}(\cos 80\ \ + j\sin 80\ \ )$
$z_2 = \ \ \overline{5}(\cos 40\ \ + j\sin 40\ \ )$

17. Calculate z_1/z_2:

(a) $z_1 = \cos 70\ \ + j\sin 70$
$z_2 = \cos 25\ \ + j\sin 25$

(b) $z_1 = 4$
$z_2 = 4(\cos 30\ \ + j\sin 30\ \ )$
(*Hint:* $4 = 4(\cos 360\ \ + j\sin 360\ \ )$)

18. What is meant geometrically by the multiplication (or division) of a complex number by $\ \ j$?

19. Calculate

(a) $(1\ \ \ j)^5$

(b) $\frac{1}{2}\ \ \ j\frac{1}{2}\ \ \overline{3}\ \ ^{3}$

20. (a) Prove that $(\cos 50\ \ \ \ j\sin 50\ \ )^4 = \cos 200\ \ \ \ j\sin 200\ \ $.
(b) State De Moivre s theorem.

21. Calculate all the roots of

(a) $\overline{\ \ \ 5 + 12j}$

(b) $^4\overline{\cos 60\ \ + j\sin 60\ \ }$

9.3 Exponential Form of Complex Numbers

22. Using Euler s formula, compute cos and sin and convert to the algebraic form:

(a) $e^{j\ \ /2}$

(b) $e^{j\ \ /3}$

23. Let the values for e^{j} and $e^{\ \ j}$ be given. Compute the values of , cos and sin :

(a) $e^{j} = 1$
$e^{\ \ j} = 1$

(b) $e^{j} = \ \ 1$
$e^{\ \ j} = \ \ 1$

(c) $e^{j} = \ \ j$
$e^{\ \ j} = j$

(d) $e^{j} = \frac{1}{2}\ \ \overline{3} + \frac{j}{2}$
$e^{\ \ j} = \frac{1}{2}\ \ \overline{3}\ \ \ \frac{j}{2}$

24. Given the complex number $z = x + jy, w = e^{z}$ is then a new complex number. Put it in the form $w = re^{j}$ and compute r and if

(a) $z = 3 + 2j$ (b) $z = 2 \quad \frac{j}{2}$

25. Transform the complex number $w = e^{z}$ into the form $w = u + jv$ if

(a) $z = \frac{1}{2} + j$ (b) $z = \frac{3}{2} \quad j$

(c) $z = \quad 1 \quad j\frac{3}{2}$ (d) $z = 3 \quad j$

26. Let the complex quantity z be a linear function of the parameter t (for example time), i.e. $z(t) = at + jbt(0 \quad t \quad d)$. Given

(a) $z(t) = \quad t + j2 \quad t$,

(b) $z(t) = 2t \quad j\frac{3}{2}t$,

(i) what is the real part, Re[$w(t)$], of $w(t) = e^{z(t)}$?
(ii) what is the period of Re[$w(t)$]?
(iii) what is the amplitude of the function $w(t)$ at time $t = 2$?

27. Compute the product $z_1 z_2$:

(a) $z_1 = 2e^{j \quad /2}$ (b) $z_1 = \frac{1}{2}e^{j \quad /4}$

$z_2 = \frac{1}{2}e^{j \quad /2}$ $z_2 = \frac{3}{2}e^{\quad j3 \quad /4}$

28. Calculate for the pairs of numbers z_1, z_2 in the previous exercise the quotient $z_1 \quad / z_2$.

29. (a) Given $z = 2e^{j \quad /5}$, calculate z^{5}.
(b) Given $z = \frac{1}{2}e^{j \quad /4}$, calculate z^{3}.

30. (a) Given $z = 32e^{j10}$, calculate $z^{1/5}$.
(b) Given $z = \frac{1}{16}e^{j6}$, calculate $z^{1/4}$.

31. Given $z = re^{j}$, what in each of the following cases is the angle for which $0 \quad 2$?

(a) $z = 3e^{j7}$ (b) $z = \frac{1}{2}e^{j14 \quad /3}$

32. Put the following complex numbers into exponential form:

(a) $5 \quad 5j$ (b) $15 \quad 13j$

33. (a) Put the expression $z = 2.5e^{j43 \quad 30}$ into the form $x + jy$.
(b) Calculate $e^{j146} \quad e^{\quad j82}$ and express the result in the form $z = x + jy$.

9.4 Operations with Complex Numbers Expressed in Polar Form

34. (a) Determine the real and imaginary parts of $\dfrac{(1 + j)^2}{\overline{2}(1 \quad j)}$

    (b) What is the polar form of this complex number?

35. Put $z = \quad 2(\cos 30 \quad j \sin 30\ )$:

    (a) into the form $x + jy$,
    (b) into exponential form.

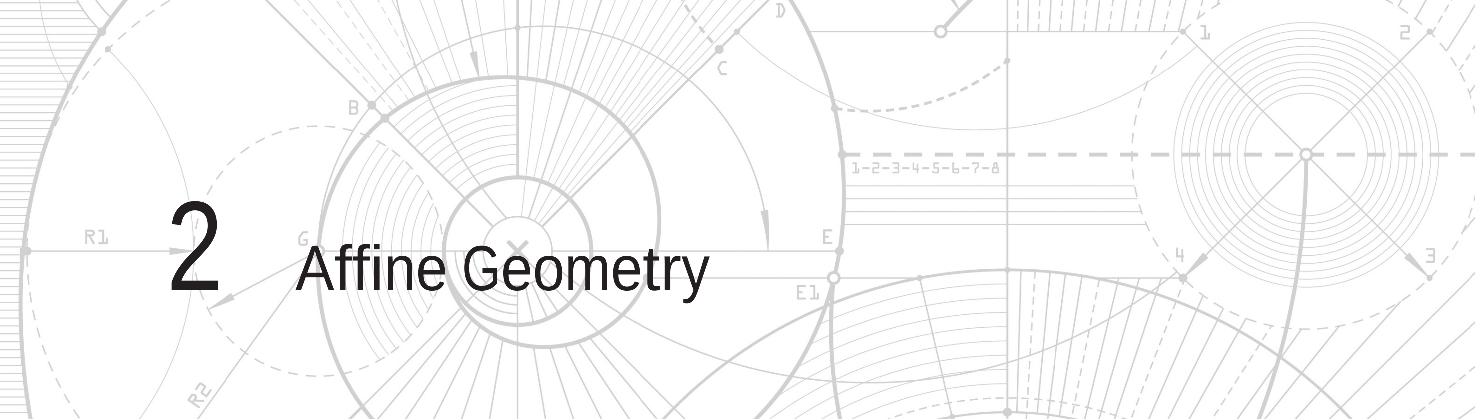

2 Affine Geometry

In Chapter 1 we studied conics in Euclidean geometry. In the rest of the book we prove a whole range of results about figures such as lines and conics, in geometries other than Euclidean geometry. In the process of doing this, we meet two particular features of our approach to geometry which may be new to you.

The first feature is the use of transformations in geometry to simplify problems and bring out their essential character. You may have met some of these transformations previously in courses on Group Theory or on Linear Algebra.

In this book, we shall often use matrices to simplify our work.

The second feature arises from the fact that the transformations we introduce form groups. Generally, we restrict our attention to geometry in the plane, $\mathbb{R}^2$, but even in this familiar setting there may be more than one group of transformations at our disposal. This leads to the exciting new idea that there are many different geometries!

Each geometry consists of a space, some properties possessed by figures in that space, and a group of transformations of the space that preserve these properties. For example, Euclidean plane geometry uses the space $\mathbb{R}^2$, and is concerned with those properties of figures that depend on the notion of distance. The group associated with Euclidean geometry is the group of isometries of the plane.

This is because isometries of the plane preserve distances.

This idea, that geometry can be thought of in terms of a space and a group acting on it, is called the *Kleinian view of geometry*, after the 19th-century German mathematician Felix Klein who proposed it first. It has the virtue of enabling us to generate many geometries, while seeing how they are related.

For instance, we can take $\mathbb{R}^2$ as our space and use the group of all transformations of the form $t(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{a}$, where $\mathbf{a} \in \mathbb{R}^2$ and $\mathbf{A}$ is a 2×2 invertible matrix. These are the so-called *affine transformations* of $\mathbb{R}^2$. But what properties of figures in $\mathbb{R}^2$ are preserved by such transformations, and what is the corresponding geometry? The geometry is called *affine geometry*, and it is the subject of this chapter. As you will see, it has some features in common with Euclidean geometry, but also some very different features.

In Section 2.1 we examine Euclidean geometry from the Kleinian point of view, and explain why geometries other than Euclidean geometry exist.

In Sections 2.2 and 2.3 we introduce affine geometry and consider its properties. In particular, we show that affine transformations map straight lines to straight lines, map parallel lines to parallel lines, and preserve ratios of lengths along a given line. We also discover that in affine geometry all triangles are congruent, in the sense that any triangle can be mapped onto any other triangle by an affine transformation. This result is known as the *Fundamental Theorem of Affine Geometry*.

In Section 2.4 we establish two important theorems, due to Ceva and Menelaus, which involve ratios of lengths along the sides of a triangle.

Finally, in Section 2.5 we investigate the effect of affine transformations on conics, and discover that we can use the methods of affine geometry to obtain very simple proofs of certain types of theorems about conics.

2.1 Geometry and Transformations

Before embarking on a study of various other geometries, it is useful first to look back at our familiar Euclidean geometry.

2.1.1 What is Euclidean Geometry?

To help us answer this question, we begin by considering the following well-known result.

Example 1 Let $\triangle ABC$ be a triangle in which $\angle ABC = \angle ACB$. Prove that $AB = AC$.

Solution First, reflect the triangle in the perpendicular bisector of BC, so that the points B and C change places and the point A moves to some point A', say. Since reflection preserves angles, it follows that $\angle A'BC = \angle ACB$.

Also, we are given that $\angle ACB = \angle ABC$, so

$$\angle A'BC = \angle ABC.$$

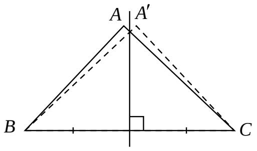

But this can happen only if A' lies on the line through A and B. Similarly,

$$\angle A'CB = \angle ABC = \angle ACB,$$

so A' must also lie on the line through A and C. This means that A and A' must coincide. Hence the line segment AB reflects to the line segment AC, and vice versa. Since reflection preserves lengths, it follows that $AB = AC$. □

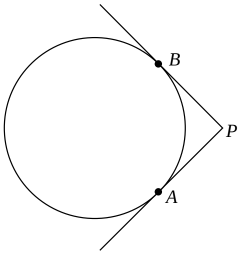

Problem 1 Let A and B be two points on a circle, and let the tangents to the circle at A and B meet at P. Prove that $AP = BP$.

Hint: Consider a reflection in the line which passes through P and the centre of the circle.

The result in Example 1 is concerned with the properties of length and angle associated with the triangle ABC. To investigate these properties, we introduced a reflection that enabled us to compare various lengths and angles. We were able to do this because reflections leave lengths and angles unchanged.

Of course, reflections are not the only transformations that preserve lengths and angles: other examples include rotations and translations. In general, any transformation that preserves lengths and angles can be used to tackle problems which involve these properties. In fact, we need worry only about leaving distances unchanged, since any transformation from $\mathbb{R}^2$ onto $\mathbb{R}^2$ that changes angles must also change lengths. Transformations that leave distances unchanged are called *isometries.*

This is because, once we know the lengths of the sides of a triangle, the angles are uniquely determined.

Definition An **isometry** of $\mathbb{R}^2$ is a function which maps $\mathbb{R}^2$ onto $\mathbb{R}^2$ and preserves distances.

In fact, every isometry has one of the following forms:

a translation along a line in $\mathbb{R}^2$;
a reflection in a line in $\mathbb{R}^2$;
a rotation about a point in $\mathbb{R}^2$;
a composite of translations, reflections and rotations in $\mathbb{R}^2$.

The identity isometry can be regarded as a rotation through an angle that is a multiple of 2 .

The composite of any two isometries is also an isometry, and so it is easy to verify that the set $S(\mathbb{R}^2)$ of isometries of $\mathbb{R}^2$ forms a group under composition of functions. These observations can be used to build up the transformations we need in order to prove Euclidean results.

Example 2 Prove that if ABC and DEF are two triangles such that

$$AB = DE, \quad AC = DF \quad \text{and} \quad BAC = EDF,$$

then $BC = EF$, $ABC = DEF$ and $ACB = DFE$.

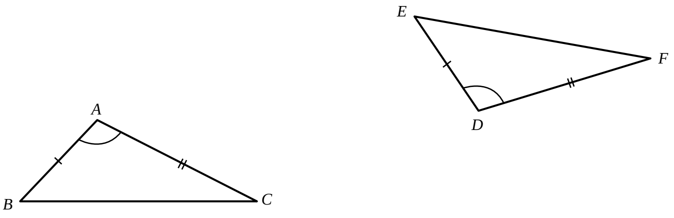

Solution It is sufficient to show that there is an isometry which maps ABC onto DEF. We construct this isometry in stages, starting with the translation which maps A to D. This translation maps $\triangle ABC$ onto $\triangle DB'C'$, where B' and C' are the images of B and C under the translation.

Of course, A and D may already coincide, in which case we omit the translation stage.

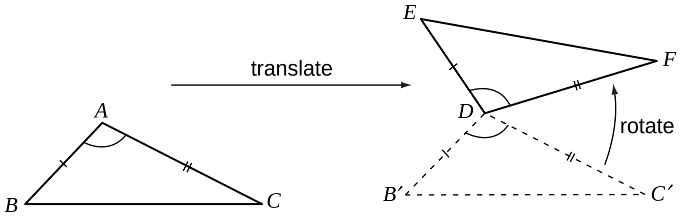

Since we are given that $DF = AC$, and since the translation maps AC onto DC', it follows that $DF = DC'$. We can therefore rotate the point C', about D, until it coincides with the point F. This rotation maps $\triangle DB'C'$ onto $\triangle DB''F$, as shown in the margin, where B'' is the image of B' under the rotation.

If C' already coincides with F, then we omit the rotation stage.

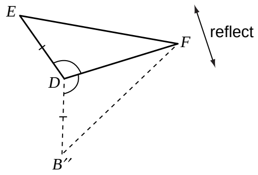

Finally, notice that

$$\begin{aligned} \angle FDE &= \angle CAB && \text{(given)} \\ &= \angle C'DB' && \text{(translation)} \\ &= \angle FDB'' && \text{(rotation)}, \end{aligned}$$

so either B'' lies on DE or the reflection of B'' in the line FD lies on DE. Also

$$\begin{aligned} DE &= AB && \text{(given)} \\ &= DB' && \text{(translation)} \\ &= DB'' && \text{(rotation)}. \end{aligned}$$

It follows that either B'' coincides with E or the reflection of B'' in the line FD coincides with E.

So, composing the translation, the rotation, and (if necessary) a reflection, we obtain the required isometry that maps $\triangle ABC$ onto $\triangle DEF$. Since isometries preserve length and angle, it follows that $BC = EF$, $\angle ABC = \angle DEF$ and $\angle ACB = \angle DFE$. □

Problem 2 Prove that if $\triangle ABC$ and $\triangle DEF$ are two triangles such that

$$AC = DF, \quad \angle BAC = \angle EDF \quad \text{and} \quad \angle ACB = \angle DFE,$$

then $BC = EF$, $AB = DE$ and $\angle ABC = \angle DEF$.

We can now answer the question 'What is Euclidean geometry?'. **Euclidean geometry** is the study of those properties of figures that are unchanged by the group of isometries. We call these properties **Euclidean properties**. Roughly speaking, a Euclidean property is one that is preserved by a rigid figure as it moves around the plane. Of course, these properties include distance and angle, but they also include other properties such as collinearity of points and concurrence of lines.

We consider only Euclidean geometry in the plane $\mathbb{R}^2$.

This idea, that geometry can be thought of in terms of a group of transformations acting on a space, is known as the *Kleinian view of geometry*. It enables us to generate many geometries, without losing sight of the relationship between them.

Here a space is simply a collection of points; for example, the plane is $\mathsf{R}^2 = \{(x, y) : x, y \text{ are real}\}$.

When we consider geometries in this way, it is often convenient to have an algebraic representation for the transformations involved. This not only enables us to solve problems in the geometry algebraically, but also provides us with formulas that can be used to compare different geometries.

In the case of Euclidean geometry, perhaps the easiest way to represent isometries algebraically is to use matrices. For example, the function defined by

$$t : \begin{matrix} x \\ y \end{matrix} \qquad \begin{matrix} \cos & \sin \\ \sin & \cos \end{matrix} \quad \begin{matrix} x \\ y \end{matrix} + \begin{matrix} e \\ f \end{matrix} \qquad (x, y) \quad \mathsf{R}^2 \tag{1}$$

is an isometry because it is the composite of an anticlockwise rotation through an angle about the origin, followed by a translation through the vector (e, f).

Similarly, the function

$$t : \begin{matrix} x \\ y \end{matrix} \qquad \begin{matrix} \cos & \sin \\ \sin & \cos \end{matrix} \quad \begin{matrix} x \\ y \end{matrix} + \begin{matrix} e \\ f \end{matrix} \qquad (x, y) \quad \mathsf{R}^2 \tag{2}$$

is an isometry because it is the composite of a reflection in the line through the origin that makes an angle $/2$ with the x-axis, followed by a translation through the vector (e, f).

Recall that the effect of the matrix multiplication in (1) and (2) can be interpreted geometrically by examining what happens to the vectors (1, 0) and (0, 1). For example, in (2) the matrix multiplication sends (1, 0) and (0, 1) to (cos , sin) and (sin , cos), respectively, so it corresponds to the reflection shown in the figure below.

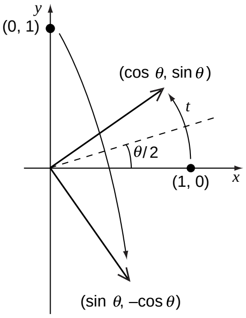

Remarkably, we can represent *any* isometry by one or other of the forms given in (1) and (2). To see this, notice that any isometry t can be written in the form

$$t(\mathbf{x}) = t_0(\mathbf{x}) + (e, f) \qquad \mathbf{x} \quad \mathsf{R}^2 , \tag{3}$$

where t_0 is an isometry which fixes the origin. Indeed, if we let $(e, f) = t(\mathbf{0})$, then we can let t_0 be the transformation defined by $t_0(\mathbf{x}) = t(\mathbf{x}) \quad (e, f)$. This is an isometry because it is the composite of the isometry t and the translation through the vector $\quad (e, f)$. It fixes the origin since $t_0(\mathbf{0}) = t(\mathbf{0}) \quad (e, f) = \mathbf{0}$.

Now an isometry that fixes the origin must be either a rotation about the origin, or a reflection in a line through the origin. If t_0 is a rotation about the origin, then (3) can be written in the matrix form given in (1), whereas if t_0 is a reflection in a line through the origin, then (3) can be written in the matrix form given in (2).

So together, equations (1) and (2) provide us with an algebraic representation of all possible isometries of the plane. The next problem indicates how we can obtain a more concise description of this algebraic representation by using orthogonal matrices to combine equations (1) and (2).

Recall that a matrix $\mathbf{U}$ is *orthogonal* if $\mathbf{U}^{\ 1} = \mathbf{U}^T$, that is, if $\mathbf{U}^T\mathbf{U} = \mathbf{I}$. This is equivalent to saying that the columns of $\mathbf{U}$ are *orthonormal*.

Problem 3 Show that both the matrices

$$\begin{matrix} \cos & \sin \\ \sin & \cos \end{matrix} \quad \text{and} \quad \begin{matrix} \cos & \sin \\ \sin & \cos \end{matrix} ,$$

which appear in (1) and (2), are orthogonal for each real number .

By applying the solution of Problem 3 to equations (1) and (2), we see that every isometry t has an algebraic representation of the form

$$t(\mathbf{x}) = \mathbf{U}\mathbf{x} + \mathbf{a},$$

where $\mathbf{U}$ is an orthogonal 2×2 matrix, and $\mathbf{a}$ is a vector in R^2.

This equation shows the matrix $\mathbf{U}$ acting on the vector $\mathbf{x} = (x, y)$. Strictly speaking, $\mathbf{U}$ acts on the coordinates $\begin{matrix} x \\ y \end{matrix}$ of $\mathbf{x}$ with respect to the standard basis of the vector space R^2 as in equations (1) and (2). However, since the numbers x and y are the same for the vector and its standard coordinates, no confusion should arise.

Definition A **Euclidean transformation** of R^2 is a function $t : \mathrm{R}^2 \quad \mathrm{R}^2$ of the form

$$t(\mathbf{x}) = \mathbf{U}\mathbf{x} + \mathbf{a},$$

where $\mathbf{U}$ is an orthogonal 2×2 matrix and $\mathbf{a} \quad \mathrm{R}^2$. The set of all Euclidean transformations of R^2 is denoted by $E(2)$.

We may summarize the discussion above by saying that every isometry of the plane is a Euclidean transformation of R^2.

In fact, the converse is also true, for if $\mathbf{U}$ is any orthogonal matrix, then its columns are orthonormal. In particular, its first and second columns have unit length and can therefore be written in the form $\begin{matrix} \cos \\ \sin \end{matrix}$ and $\begin{matrix} \cos \\ \sin \end{matrix}$, respectively, for some real $\ ,\ $. For these to be orthonormal, we must have $\cos \ \bullet \cos \ + \sin \ \bullet \sin \ = 0$, so that $\tan \ \bullet \tan \ = \ 1$ and hence $\ = \ - \frac{}{2}$. It follows that the second column must be

$$\begin{matrix} \cos(\ + \ /2) \\ \sin(\ + \ /2) \end{matrix} = \begin{matrix} \sin \\ \cos \end{matrix} \quad \text{or}$$

$$\begin{matrix} \cos(\quad /2) \\ \sin(\quad /2) \end{matrix} = \begin{matrix} \sin \\ \cos \end{matrix}.$$

So

$$\mathbf{U} = \begin{matrix} \cos & \sin \\ \sin & \cos \end{matrix} \quad \text{or} \quad \mathbf{U} = \begin{matrix} \cos & \sin \\ \sin & \cos \end{matrix}.$$

It follows that every Euclidean transformation $t(\mathbf{x}) = \mathbf{U}\mathbf{x} + \mathbf{a}$ of R^2 has one of the forms given in equations (1) and (2). Since both of these forms represent isometries of the plane,we have the following theorem.

Theorem 1 Every isometry of R^2 is a Euclidean transformation of R^2, and vice versa.

Now the set of all isometries of R^2 forms a group under composition of functions, so it follows from Theorem 1 that the same must be true of the set of all Euclidean transformations of R^2. We therefore have the following theorem.

It is not difficult to prove that this is a group, so we omit the proof.

Theorem 2 The set of Euclidean transformations of R^2 forms a group under the operation of composition of functions.

It is instructive to check the group axioms algebraically, for in the process of doing so we obtain formulas for the composites and inverses of Euclidean transformations.

We start by considering closure. Suppose that t_1 and t_2 are two Euclidean transformations given by

$$t_1(\mathbf{x}) = \mathbf{U}_1\mathbf{x} + \mathbf{a}_1 \quad \text{and} \quad t_2(\mathbf{x}) = \mathbf{U}_2\mathbf{x} + \mathbf{a}_2,$$

where $\mathbf{U}_1$ and $\mathbf{U}_2$ are orthogonal 2 × 2 matrices. Then the composite $t_1 \quad t_2$ is given by

$$\begin{aligned} t_1 \quad t_2(\mathbf{x}) &= t_1(\mathbf{U}_2\mathbf{x} + \mathbf{a}_2) \\ &= \mathbf{U}_1(\mathbf{U}_2\mathbf{x} + \mathbf{a}_2) + \mathbf{a}_1 \\ &= \mathbf{U}_1\mathbf{U}_2\mathbf{x} + (\mathbf{U}_1\mathbf{a}_2 + \mathbf{a}_1). \end{aligned}$$

This is a Euclidean transformation since $\mathbf{U}_1\mathbf{U}_2$ is orthogonal. Indeed,

$$(\mathbf{U}_1\mathbf{U}_2)^T = \mathbf{U}_2^T\mathbf{U}_1^T = \mathbf{U}_2^{\ 1}\mathbf{U}_1^{\ 1} = (\mathbf{U}_1\mathbf{U}_2)^{\ 1}.$$

Here we are using the result that $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$.

So the set of Euclidean transformations is closed under composition of functions.

Problem 4 Let the Euclidean transformations t_1 and t_2 of $\mathbb{R}^2$ be given by

$$t_1(\mathbf{x}) = \begin{matrix} \frac{3}{5} & \frac{4}{5} \\ \frac{4}{5} & \frac{3}{5} \end{matrix} \ \mathbf{x} + \begin{matrix} 1 \\ 2 \end{matrix}$$

and

$$t_2(\mathbf{x}) = \begin{matrix} \frac{4}{5} & \frac{3}{5} \\ \frac{3}{5} & \frac{4}{5} \end{matrix} \ \mathbf{x} + \begin{matrix} 2 \\ 1 \end{matrix} \ .$$

Determine $t_1 \quad t_2$ and $t_2 \quad t_1$.

Next recall that under composition of functions the identity is the transformation given by $i(\mathbf{x}) = \mathbf{x}$. This is a Euclidean transformation since it can be written in the form

$$i(\mathbf{x}) = \mathbf{I}\mathbf{x} + \mathbf{0},$$

where $\mathbf{I}$ is the 2 × 2 identity matrix, which is orthogonal.

The next problem asks you to show that inverses exist.

Problem 5 Prove that if t_1 is a Euclidean transformation of $\mathbb{R}^2$ given by

$$t_1(\mathbf{x}) = \mathbf{U}\mathbf{x} + \mathbf{a} \qquad (\mathbf{x} \quad \mathbb{R}^2),$$

then:

(a) the transformation of $\mathbb{R}^2$ given by

$$t_2(\mathbf{x}) = \mathbf{U}^{-1}\mathbf{x} - \mathbf{U}^{-1}\mathbf{a} \qquad (\mathbf{x} \in \mathbb{R}^2)$$

is also a Euclidean transformation;

(b) the transformation t_2 is the inverse of t_1.

The solution of Problem 5 shows that we can calculate the inverse of a Euclidean transformation by using the following result.

> The inverse of the Euclidean transformation $t(\mathbf{x}) = \mathbf{U}\mathbf{x} + \mathbf{a}$ is given by
>
> $$t^{-1}(\mathbf{x}) = \mathbf{U}^{-1}\mathbf{x} - \mathbf{U}^{-1}\mathbf{a}.$$

Problem 6 Determine the inverse of the Euclidean transformation given by

$$t(\mathbf{x}) = \begin{pmatrix} \frac{3}{5} & \frac{4}{5} \\ \frac{4}{5} & \frac{3}{5} \end{pmatrix} \mathbf{x} + \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

Finally, composition of functions is always associative. So all four group properties hold, as we expected.

Earlier, we described Euclidean geometry as the study of those properties of figures that are preserved by isometries. Having identified these isometries with the group of Euclidean transformations, we can now give the equivalent algebraic description of Euclidean geometry. **Euclidean geometry** is the study of those properties of figures that are preserved by Euclidean transformations of $\mathbb{R}^2$.

2.1.2 Euclidean-Congruence

In the solution to Example 2 we showed that if two triangles $\triangle ABC$ and $\triangle DEF$ are such that $AB = DE$, $AC = DF$ and $\angle BAC = \angle EDF$, then there is a Euclidean transformation which maps $\triangle ABC$ onto $\triangle DEF$.

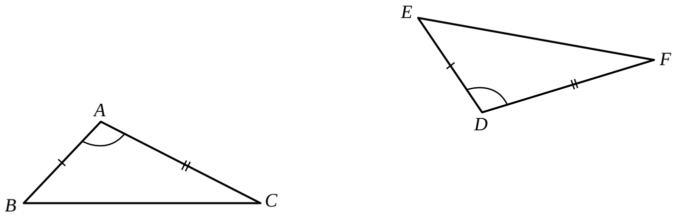

The existence of this transformation enabled us to deduce that both triangles have the same Euclidean properties. In particular, we were able to deduce that $BC = EF$, $\angle ABC = \angle DEF$ and $\angle ACB = \angle DFE$.

In order to formalize this way of relating two figures, we say that two figures are *congruent* if one can be moved to fill exactly the position of the other by means of a Euclidean transformation. Loosely speaking, two figures are congruent if they have the same size and shape.

Later we consider congruence with respect to other groups of transformations (that is, congruence in other geometries), so if there is any danger of confusion we sometimes say that two figures are *Euclidean-congruent*.

Definition A figure F_1 is **Euclidean-congruent** to a figure F_2 if there is a Euclidean transformation which maps F_1 onto F_2.

For example, any two circles of unit radius are Euclidean-congruent to each other because we can map one of the circles onto the other by means of a translation that makes their centres coincide.

Of course, there are many other Euclidean transformations which map one of the circles onto the other.

Problem 7 Which of the following sets consist of figures that are Euclidean-congruent to each other?

(a) The set of all ellipses
(b) The set of all line segments of length 1
(c) The set of all triangles
(d) The set of all squares that have sides of length 2

Earlier, we emphasized that the Euclidean transformations form a group. This is important because it ensures that Euclidean-congruence has the kind of properties that we should expect. For example, we should expect every figure to be congruent to itself. Also, if a figure F_1 is congruent to a figure F_2, then we should expect F_2 to be congruent to F_1. We can, in fact, establish the following result.

Theorem 3 Euclidean-congruence is an equivalence relation.

Proof We show that the three equivalence relation axioms E1, E2 and E3 hold.

E1 REFLEXIVE For all figures F in $\mathbb{R}^2$, the identity transformation maps F onto itself; so Euclidean-congruence is reflexive.

This uses the existence of an *identity* transformation.

E2 SYMMETRIC Let a figure F_1 in $\mathbb{R}^2$ be congruent to a figure F_2, and let t be a Euclidean transformation which maps F_1 onto F_2. Then the inverse Euclidean transformation t^{-1} maps F_2 onto F_1, so that F_2 is congruent to F_1. Thus Euclidean-congruence is symmetric.

This uses the existence of *inverse* transformations.

E3 TRANSITIVE Let a figure F_1 in $\mathbb{R}^2$ be congruent to a figure F_2, and let F_2 be congruent to a figure F_3. Then there exist Euclidean transformations t_1 mapping F_1 onto F_2 and t_2 mapping F_2 onto F_3. Thus the Euclidean transformation $t_2 \circ t_1$ maps F_1 onto F_3, so that F_1 is congruent to F_3. Hence Euclidean-congruence is transitive.

This uses the closure axiom for the group of Euclidean transformations.

It follows that Euclidean-congruence is an equivalence relation, because it satisfies the axioms El, E2 and E3.

Problem 8 Prove that if two figures in $\mathbb{R}^2$ are each Euclidean-congruent to a third figure, then they are Euclidean-congruent to each other.

Since Euclidean-congruence is an equivalence relation, it partitions the set of all figures into disjoint equivalence classes. Each class consists of figures which are Euclidean-congruent to each other, and hence share the same Euclidean properties (for example, one class consists of all circles of unit radius, another class consists of all equilateral triangles with sides of length 3, and so on). If we wish to show that two figures have the same Euclidean properties, then it is sufficient to show that they are Euclidean-congruent.

For example, to show that two triangles $\triangle ABC$ and $\triangle DEF$ have the same Euclidean properties, it is sufficient to show that $AB = DE$, $AC = DF$ and $\angle BAC = \angle EDF$, as you saw in Example 2. This congruence condition is frequently used in Euclidean geometry. It is known as the 'side angle side' (SAS) condition for congruence.

Now Euclidean geometry is just one of several different geometries. Each geometry is defined by a group G of transformations that act on a space. In general, we say that two figures are ***G*-congruent** if there is a transformation in G which maps one of the figures onto the other. Since the only properties used in the proof of Theorem 3 are the group properties of Euclidean transformations, the theorem holds also with 'G-congruent' in place of 'Euclidean-congruent'. Thus, like Euclidean-congruence, G-congruence is an equivalence relation that partitions the set of all figures into disjoint equivalence classes.

This idea of partitioning figures into equivalence classes is central to geometry. It enables us to distinguish between figures in different equivalence classes, without having to worry about the differences between figures in the same equivalence class. For example, if we are interested in whether a conic is an ellipse rather than a hyperbola or a parabola, but do not care about its shape (that is, the ratio of the lengths of its axes), we might choose to work with some geometry whose group of transformations makes all ellipses congruent to each other — but not congruent to any hyperbola or parabola. We describe a group of transformations which defines such a geometry in Section 2.2.

2.2 Affine Transformations and Parallel Projections

2.2.1 Affine Transformations

In Section 2.1 you met a new approach to Euclidean geometry in $\mathbb{R}^2$ — namely, the idea that Euclidean geometry of $\mathbb{R}^2$ can be interpreted as a space, $\mathbb{R}^2$, together with the group of Euclidean transformations which act on that space. Recall that a Euclidean transformation is a function $t : \mathbb{R}^2 \to \mathbb{R}^2$ of the form

$$t(\mathbf{x}) = \mathbf{Ux} + \mathbf{a} \quad (\mathbf{x} \in \mathbb{R}^2),$$

where $\mathbf{U}$ is an orthogonal 2×2 matrix. Euclidean properties of figures are those, like distance and angle, that are preserved by these transformations.

In this section we meet the first of our new geometries in $\mathbb{R}^2$ — *affine geometry*. This geometry consists of the space $\mathbb{R}^2$ together with a group of transformations, the *affine transformations*, acting on $\mathbb{R}^2$.

Affine geometry can be defined in $\mathbb{R}^n$, for any $n \geq 2$; we restrict our attention here to the case when $n = 2$.

Definition An **affine transformation** of $\mathbb{R}^2$ is a function $t : \mathbb{R}^2 \to \mathbb{R}^2$ of the form

$$t(\mathbf{x}) = \mathbf{Ax} + \mathbf{b},$$

where $\mathbf{A}$ is an invertible 2×2 matrix and $\mathbf{b} \in \mathbb{R}^2$. The set of all affine transformations of $\mathbb{R}^2$ is denoted by $A(2)$.

Remark

Note that every Euclidean transformation of $\mathbb{R}^2$ is an affine transformation of $\mathbb{R}^2$ since every orthogonal matrix is invertible. (In terms of groups, the group of Euclidean transformations of $\mathbb{R}^2$ is a proper subgroup of the group of affine transformations of $\mathbb{R}^2$.) This means that all properties of figures that are preserved by affine transformations must be preserved also by Euclidean transformations.

Problem 1 Determine whether or not each of the following transformations of $\mathbb{R}^2$ is an affine transformation.

(a) $t_1(\mathbf{x}) = \begin{pmatrix} 1 & 3 \\ 1 & 2 \end{pmatrix} \mathbf{x} + \begin{pmatrix} 4 \\ 2 \end{pmatrix}$ (b) $t_2(\mathbf{x}) = \begin{pmatrix} 6 & 5 \\ 3 & 2 \end{pmatrix} \mathbf{x} + \begin{pmatrix} 2 \\ 1 \end{pmatrix}$

(c) $t_3(\mathbf{x}) = \begin{pmatrix} 2 & 1 \\ 8 & 4 \end{pmatrix} \mathbf{x} + \begin{pmatrix} 1 \\ 3 \end{pmatrix}$ (d) $t_4(\mathbf{x}) = \begin{pmatrix} 5 & 3 \\ 2 & 2 \end{pmatrix} \mathbf{x}$

The algebra required to compose affine transformations is similar to the algebra that we used to compose Euclidean transformations.

Problem 2 For the transformations of $\mathbb{R}^2$ given in Problem 1, determine formulas for the following composites. In each case, state whether or not the composite is an affine transformation.

(a) $t_1 \circ t_2$ (b) $t_2 \circ t_4$

We now verify our assertion above that the set of affine transformations forms a group.

Theorem 1 The set of affine transformations $A(2)$ forms a group under the operation of composition of functions.

Proof We check that the four group axioms hold.

G1 CLOSURE Let t_1 and t_2 be affine transformations given by

$$t_1(\mathbf{x}) = \mathbf{A}_1\mathbf{x} + \mathbf{b}_1 \quad \text{and} \quad t_2(\mathbf{x}) = \mathbf{A}_2\mathbf{x} + \mathbf{b}_2,$$

where $\mathbf{A}_1$ and $\mathbf{A}_2$ are invertible 2 × 2 matrices. Then, for each $\mathbf{x} \in \mathbb{R}^2$,

$$\begin{aligned}(t_1 \circ t_2)(\mathbf{x}) &= t_1(\mathbf{A}_2\mathbf{x} + \mathbf{b}_2)\\ &= \mathbf{A}_1(\mathbf{A}_2\mathbf{x} + \mathbf{b}_2) + \mathbf{b}_1\\ &= (\mathbf{A}_1\mathbf{A}_2)\mathbf{x} + (\mathbf{A}_1\mathbf{b}_2 + \mathbf{b}_1).\end{aligned}$$

Since $\mathbf{A}_1$ and $\mathbf{A}_2$ are invertible, it follows that $\mathbf{A}_1\mathbf{A}_2$ is also invertible. So by definition $t_1 \circ t_2$ is an affine transformation.

G2 IDENTITY Let i be the affine transformation given by

$$i(\mathbf{x}) = \mathbf{I}\mathbf{x} + \mathbf{0} \quad (\mathbf{x} \in \mathbb{R}^2),$$

where $\mathbf{I}$ is the 2 × 2 identity matrix. If t is an affine transformation given by

$$t(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b} \quad (\mathbf{x} \in \mathbb{R}^2),$$

then, for each $\mathbf{x} \in \mathbb{R}^2$,

$$(t \circ i)(\mathbf{x}) = \mathbf{A}(\mathbf{I}\mathbf{x} + \mathbf{0}) + \mathbf{b} = \mathbf{A}\mathbf{x} + \mathbf{b} = t(\mathbf{x})$$

and

$$(i \circ t)(\mathbf{x}) = \mathbf{I}(\mathbf{A}\mathbf{x} + \mathbf{b}) + \mathbf{0} = \mathbf{A}\mathbf{x} + \mathbf{b} = t(\mathbf{x}).$$

Thus $t \circ i = i \circ t = t$. Hence i is the identity transformation.

G3 INVERSES If t is an arbitrary affine transformation given by

$$t(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b} \quad (\mathbf{x} \in \mathbb{R}^2),$$

then we can define another affine transformation t' by

$$t'(\mathbf{x}) = \mathbf{A}^{-1}\mathbf{x} - \mathbf{A}^{-1}\mathbf{b}.$$

Now for each $\mathbf{x} \in \mathbb{R}^2$, we have

$$\begin{aligned}(t \circ t')(\mathbf{x}) &= t(\mathbf{A}^{-1}\mathbf{x} - \mathbf{A}^{-1}\mathbf{b})\\ &= \mathbf{A}(\mathbf{A}^{-1}\mathbf{x} - \mathbf{A}^{-1}\mathbf{b}) + \mathbf{b}\\ &= (\mathbf{A}\mathbf{A}^{-1}\mathbf{x} - \mathbf{A}\mathbf{A}^{-1}\mathbf{b}) + \mathbf{b}\\ &= (\mathbf{x} - \mathbf{b}) + \mathbf{b}\\ &= \mathbf{x}.\end{aligned}$$

Also,

$$\begin{aligned}(t' \circ t)(\mathbf{x}) &= t'(\mathbf{Ax}+\mathbf{b})\\ &= \mathbf{A}^{-1}(\mathbf{Ax}+\mathbf{b}) - \mathbf{A}^{-1}\mathbf{b}\\ &= (\mathbf{A}^{-1}\mathbf{Ax}+\mathbf{A}^{-1}\mathbf{b}) - \mathbf{A}^{-1}\mathbf{b}\\ &= (\mathbf{x}+\mathbf{A}^{-1}\mathbf{b}) - \mathbf{A}^{-1}\mathbf{b}\\ &= \mathbf{x}.\end{aligned}$$

Thus $t \circ t' = t' \circ t = i$. Hence t' is an inverse for t.

G4 ASSOCIATIVITY Composition of functions is always associative.

It follows that the set of affine transformations $A(2)$ forms a group under composition of functions.

The above proof shows that we can calculate the inverse of an affine transformation by using the following result.

The inverse of the affine transformation $t(\mathbf{x}) = \mathbf{Ax} + \mathbf{b}$ is given by

$$t^{-1}(\mathbf{x}) = \mathbf{A}^{-1}\mathbf{x} - \mathbf{A}^{-1}\mathbf{b}.$$

Problem 3 Find the inverse of the affine transformation

$$t(\mathbf{x}) = \begin{pmatrix}1 & 3\\ 1 & 2\end{pmatrix}\mathbf{x} + \begin{pmatrix}4\\ 2\end{pmatrix}.$$

Having shown that the set of affine transformations forms a group under composition of functions, we now define **affine geometry** to be the study of those properties of figures in the plane $\mathbb{R}^2$ that are preserved by affine transformations. These are the so-called **affine properties** of figures. We begin our investigation of affine geometry by considering the three affine properties listed below.

Basic Properties of Affine Transformations

Affine transformations:

1. map straight lines to straight lines;
2. map parallel straight lines to parallel straight lines;
3. preserve ratios of lengths along a given straight line.

There are two approaches that we shall use to investigate these properties. One approach is to use the definition of an affine transformation to investigate the properties algebraically; we do this in Section 2.3. First, however, we investigate the properties geometrically. We begin to do this in the next subsection by introducing a special type of affine transformation for which there is a simple geometric interpretation.

2.2.2 Parallel Projections

A *parallel projection* is a one–one mapping from $\mathbb{R}^2$ onto itself, defined in the following way. First, we think of its domain and codomain as two separate copies of $\mathbb{R}^2$.

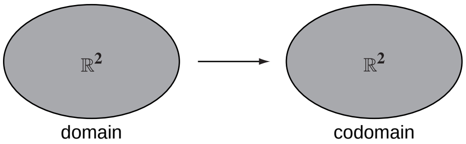

Geometrically, we can represent these copies of $\mathbb{R}^2$ by two separate planes, each equipped with a pair of rectangular axes.

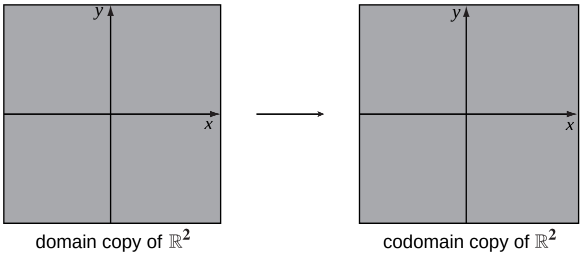
y
x
y
x
domain copy of $\mathbb{R}^2$
codomain copy of $\mathbb{R}^2$

Next we place these planes into three-dimensional space; we denote the domain plane by π_1 and the codomain plane by π_2.

Now imagine parallel rays of light shining through π_1 and π_2. Each point P in the plane π_1 has a (unique) ray passing through it, that also passes through a point P', say, in the plane π_2. This provides us with a one–one correspondence between points in the two planes π_1 and π_2. We call the function p which maps each point P in π_1 to the corresponding point P' in π_2 a **parallel projection from π_1 onto π_2.**

Of course, since π_1 and π_2 represent copies of $\mathbb{R}^2$, a parallel projection is really a function from $\mathbb{R}^2$ onto itself. In Subsection 2.2.3 we show that parallel projections are affine transformations.

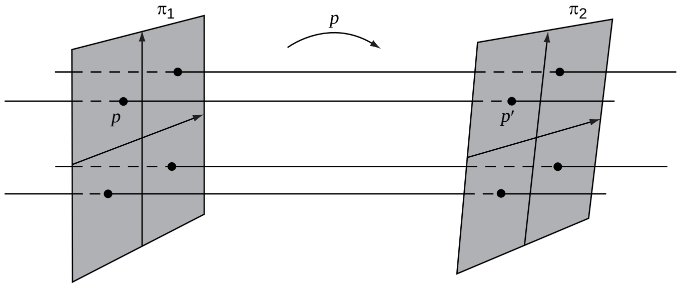

If the roles of the planes π_1 and π_2 are reversed, so that π_2 becomes the domain plane and π_1 becomes the codomain plane, then we obtain the inverse function $p\ ^{1}$ which maps points P in $_2$ back to the corresponding points P in $_1$. Clearly, $p\ ^{1}$ is a parallel projection of $_2$ onto $_1$.

Each choice of location for the domain plane $_1$, and the codomain plane $_2$, and each choice of direction for the rays of light, yields a parallel projection. The only constraint is that the rays of light must not be parallel to either plane.

If the planes $_1$ and $_2$ are parallel to each other, then any parallel projection p from $_1$ onto $_2$ is an isometry, since the distance between any two points is unaltered.

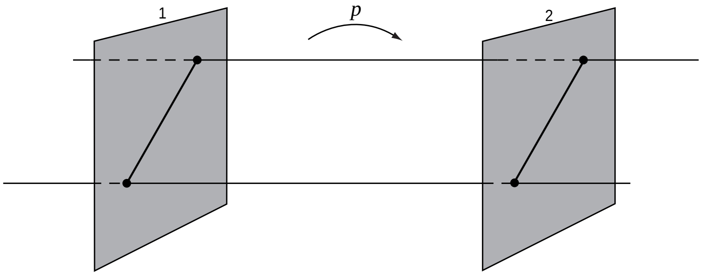

You can envisage the mapping p from $_1$ onto $_2$ as sliding $_1$ parallel to itself along the family of rays .

On the other hand, if the planes are not parallel to each other, then some distances are changed under the projection, and so the parallel projection is not an isometry; notice, however, that distances along the line of intersection of the planes $_1$ and $_2$ do remain unchanged by the parallel projection.

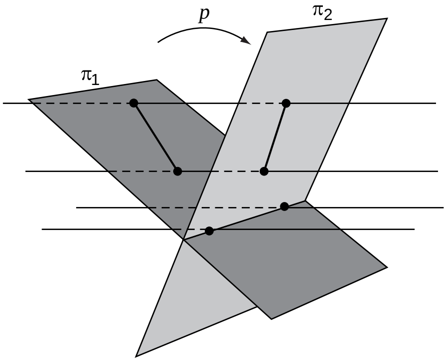

Although distances are not always preserved by a parallel projection, there are some basic properties that are preserved; three of these are listed below. As you will see, these are the same as the basic affine properties that we mentioned at the end of Subsection 2.2.1.

Basic Properties of Parallel Projections
Parallel projections:

1. map straight lines to straight lines;
2. map parallel straight lines to parallel straight lines;
3. preserve ratios of lengths along a given straight line.

Later, we will show that each basic affine property follows directly from the corresponding property for parallel projections. In anticipation of this, we first show that the properties hold for parallel projections.

Property 1 A parallel projection maps straight lines to straight lines.

Proof Let ℓ be a line in the plane π_1, and let p be a parallel projection mapping π_1 onto the plane π_2. Now consider all the rays associated with p that pass through ℓ. Since these rays are parallel, they must fill a plane. Call this plane π.

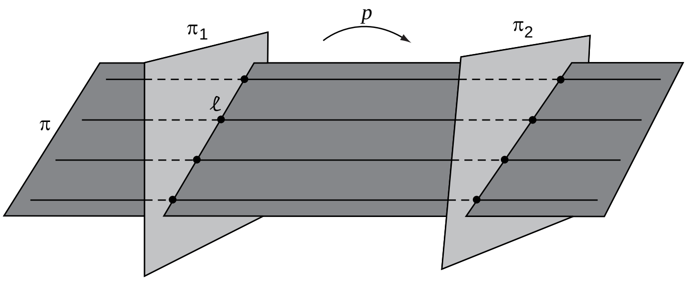

The image of ℓ under p consists of those points where the rays that pass through ℓ meet π_2. But these points are simply the points of intersection of π with π_2. Since any two intersecting planes in $\mathbb{R}^3$ meet in a line, it follows that the image of ℓ under p is a straight line.

Property 2 A parallel projection maps parallel straight lines to parallel straight lines.

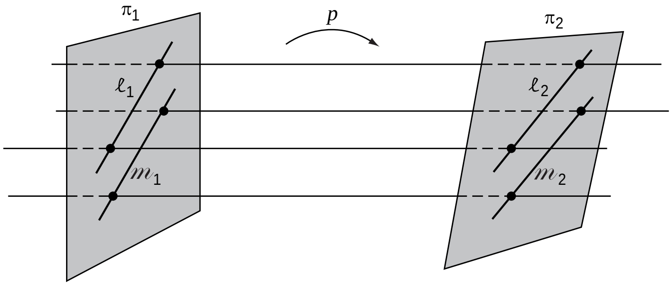

Proof Let ℓ_1 and m_1 be parallel lines in the plane π_1, and let p be a parallel projection mapping π_1 onto the plane π_2. Let ℓ_2 and m_2 be the lines in π_2 that are the images under p of ℓ_1 and m_1.

If ℓ_2 and m_2 are not parallel, they meet at some point, P_2 say. Let P_1 be the point of π_1 which maps to P_2. Then P_1 must lie on both ℓ_1 and m_1. Since ℓ_1 and m_1 are parallel, no such point of intersection can exist, which is a contradiction. It follows that ℓ_2 and m_2 must indeed be parallel.

P_1 is the point $p^{-1}(P_2)$.

Property 3 A parallel projection preserves ratios of lengths along a given straight line.

We shall give a slightly more general form of this property in Theorem 4 of Subsection 2.3.3.

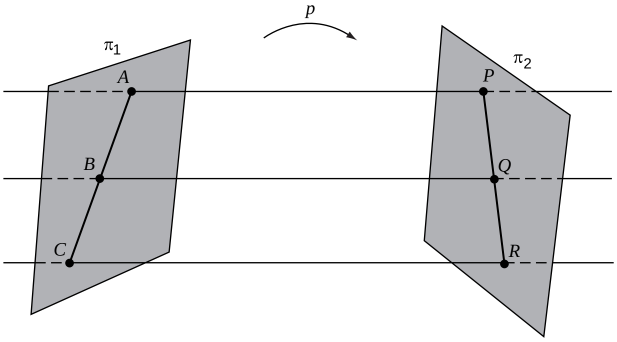

Proof Let A, B, C be three points on a line in the plane π_1, and let p be a parallel projection mapping π_1 onto the plane π_2. Let P, Q, R be the points in π_2 that are the images under p of A, B, C. We know from Property 1 that P, Q, R lie on a line; we have to show that the ratio $AB : AC$ is equal to the ratio $PQ : PR$.

If the planes π_1 and π_2 are parallel, then the parallel projection p is an isometry, and so the ratios $AB : AC$ and $PQ : PR$ are equal, as required. On the other hand, if π_1 and π_2 are not parallel, then we can construct a plane through the point P which is parallel to π_1, as shown in the margin. This plane intersects the ray through B and Q at some point B', and the ray through C and R at some point C'. So in this case the ratios $AB : AC$ and $PB' : PC'$ are equal.

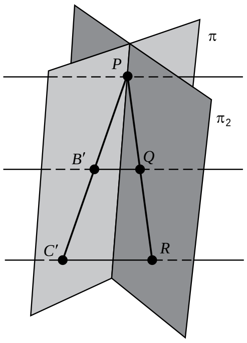

Now consider $\triangle PC'R$. The lines $B'Q$ and $C'R$ are parallel, since they are rays from the parallel projection. Hence $B'Q$ meets the sides PR and PC' in equal ratios. Thus $PQ : PR = PB' : PC'$. It follows that $PQ : PR = AB : AC$, as required.

Notice, in particular, that if a point is the midpoint of a line segment, then under a parallel projection the image of the point is the midpoint of the image of the line segment.

We make use of this fact in Subsection 2.2.3.

In Subsection 2.2.3 you will see why the basic properties of affine transformations and of parallel projections are the same, and you will meet some further properties of each.

2.2.3 Affine Geometry

In this subsection we explore further the ideas of affine geometry and of parallel projection in order to prove two attractive and unexpected results about ellipses. Also, we examine the relationship between affine transformations and parallel projections.

Two Results about Ellipses

First, starting with any chord of an ellipse, draw all the chords parallel to and construct their midpoints. We claim that these midpoints lie on a chord through the centre of the ellipse that is, on a *diameter* of the ellipse.

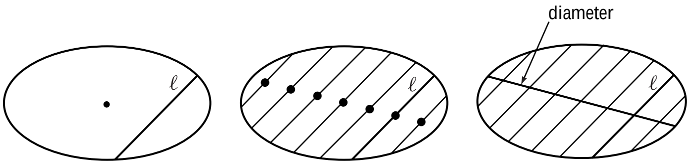

Theorem 2 Midpoint Theorem
Let be a chord of an ellipse. Then the midpoints of the chords parallel to lie on a diameter of the ellipse.

Next, start with any diameter of an ellipse and construct a second diameter m by following the construction used in Theorem 2, as shown below. Then repeat the construction starting this time with the diameter m; this might reasonably be expected to give us a third diameter of the ellipse but, surprisingly, it gives us the diameter with which we started.

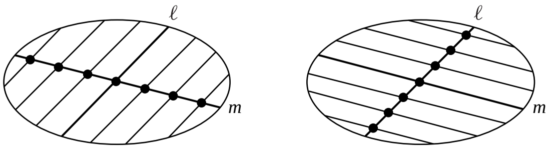

Theorem 3 Conjugate Diameters Theorem
Let be a diameter of an ellipse. Then there is another diameter m of the ellipse such that:

(a) the midpoints of all chords parallel to lie on m;
(b) the midpoints of all chords parallel to m lie on .

The directions of these two diameters are called *conjugate directions*, and the diameters are called *conjugate diameters*.

Proofs for the Special Case of a Circle

We now investigate these theorems for the special case when the ellipse is a circle. To prove the Midpoint Theorem in this case, start with a chord . If necessary, rotate the circle to ensure that is horizontal. It is then sufficient to prove that every horizontal chord is bisected by the vertical diameter, m.

Recall that a circle is an ellipse with eccentricity zero.

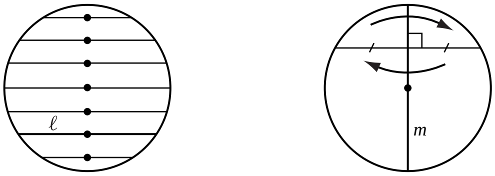

To do this note that the circle is symmetrical about m; so, reflection in m maps that part of every horizontal chord to the left of m exactly onto the part to the right of *m*. Since reflection preserves length, these two parts must be the same length; in other words, *m* bisects each horizontal chord, as required.

What about the Conjugate Diameters Theorem for the special case of the circle?

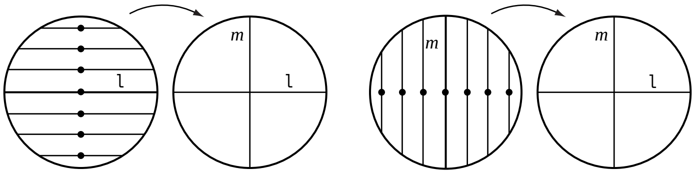

Start with the horizontal diameter , and carry out the construction of another diameter as in Theorem 2; this yields the vertical diameter m. If we then start with the vertical diameter *m* and repeat the construction, we obtain , the horizontal diameter of the circle. So Theorem 3 certainly holds when the ellipse is a circle.

Generalizing the Proof

We now investigate how the proofs of Theorems 2 and 3 for the circle can be turned into proofs for any kind of ellipse. The crucial fact is as follows.

Theorem 4 Given any ellipse, there is a parallel projection which maps the ellipse onto a circle.

An algebraic proof of a related theorem is given in Theorem 1 of Subsection 2.5.1.

A suitable parallel projection is illustrated below. Here the plane $_1$ (initially parallel to $_2$) has been tilted about the minor axis of the ellipse. Under the projection distances which are parallel to the minor axis remain unchanged, but distances parallel to the major axis are scaled by a factor which depends on the angle of tilt . By choosing just the right amount of tilt we can ensure that the image of the major axis is equal in length to the image of the minor axis, thereby ensuring that the image of the ellipse is a circle.

Algebraically, in terms of a suitable coordinate system, the mapping

$$x \qquad \frac{b}{a}x, \quad y \qquad y,$$

maps the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1,$$

to the circle $x^2 + y^2 = b^2$.

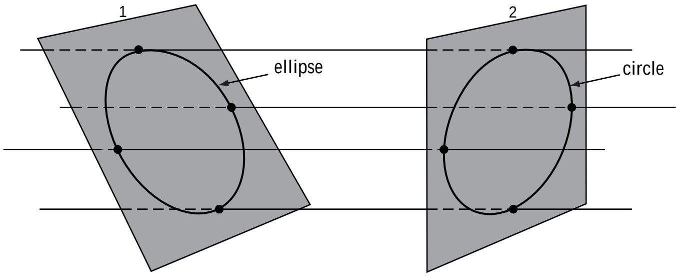

Both Theorems 2 and 3 may now be proved using the following technique. First, map the given ellipse onto a circle, using a suitable parallel projection *p*.

Since we have seen that the theorems hold in the case of the circle, we then map the circle back to the ellipse, using the inverse parallel projection p^{-1}. Now collinearity and parallelism are preserved under a parallel projection, as is the property of being the midpoint of a line segment, so the above two theorems, which hold for a circle, must hold also for the ellipse.

Here we are using Properties 1, 2 and 3 of parallel projections, in a crucial way.

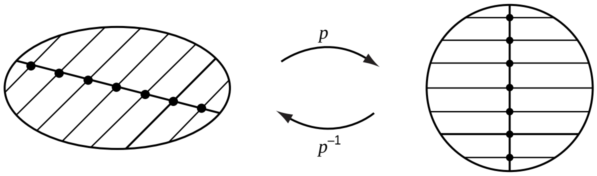

Notice that certain properties of figures, such as length and angle, are not preserved under a parallel projection. This is one difference between Euclidean geometry and affine geometry. The difference arises because the group of affine transformations is larger than the group of Euclidean transformations. In general, the larger the group that is used to define a geometry, the fewer properties the geometry has.

Affine Transformations and Parallel Projections

Earlier we mentioned that a parallel projection is a special type of affine transformation. We now show why this is indeed the case.

First, consider a parallel projection p of a plane π_1 onto a plane π_2. For the moment, suppose that the planes are aligned so that the origin in π_1 is mapped to the origin in π_2. Since ratios of lengths are preserved along a straight line, we must have, for any vector $\mathbf{v} \in \mathbb{R}^2$ and any $\lambda \in \mathbb{R}$,

Property 3, Subsection 2.2.2

$$p(\lambda \mathbf{v}) = \lambda p(\mathbf{v}). \qquad (1)$$

Next, let $\mathbf{v}$ and $\mathbf{w}$ be two position vectors in π_1. Their sum, $\mathbf{v} + \mathbf{w}$, is found from the Parallelogram Law for addition of vectors, as shown in the diagram below. The images under p in π_2 are $p(\mathbf{v})$ and $p(\mathbf{w})$, and the sum of these two vectors is $p(\mathbf{v}) + p(\mathbf{w})$.

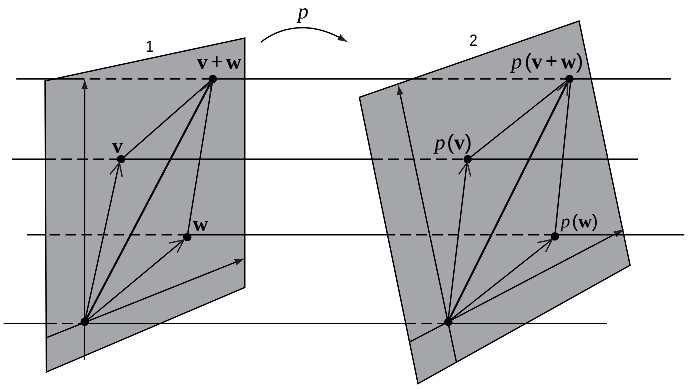

But a parallel projection maps parallel lines onto parallel lines, so it must map parallelograms onto parallelograms. Hence it must map the parallelogram in $_1$ onto the parallelogram in $_2$, and, in particular, it must map $\mathbf{v} + \mathbf{w}$ to $p(\mathbf{v}) + p(\mathbf{w})$. We may write this as

$$p(\mathbf{v} + \mathbf{w}) = p(\mathbf{v}) + p(\mathbf{w}). \qquad (2)$$

It follows from equations (1) and (2) that p must be a *linear* transformation of $\mathbb{R}^2$ onto itself.

Remember that $_1$ and $_2$ represent copies of $\mathbb{R}^2$.

Hence there exists some matrix $\mathbf{A}$ such that for each $\mathbf{v} \quad \mathbb{R}^2$,

$$p(\mathbf{v}) = \mathbf{A}\mathbf{v}. \qquad (3)$$

Since the linear transformation p is invertible, it follows that $\mathbf{A}$ is invertible.

Now suppose that the parallel projection maps the origin in $_1$ to some point B with position vector $\mathbf{b}$ in $_2$, as shown below. If we temporarily construct a new set of axes in $_2$ that are parallel to the original axes, but which intersect at the point B, then with respect to these new axes $p(\mathbf{v}) = \mathbf{A}\mathbf{v}$ for some invertible matrix $\mathbf{A}$, as before. To express $p(\mathbf{v})$ with respect to the original axes, we simply add on the vector $\mathbf{b}$ to obtain

$$p(\mathbf{v}) = \mathbf{A}\mathbf{v} + \mathbf{b} \qquad (4)$$

for some invertible 2 × 2 matrix $\mathbf{A}$.

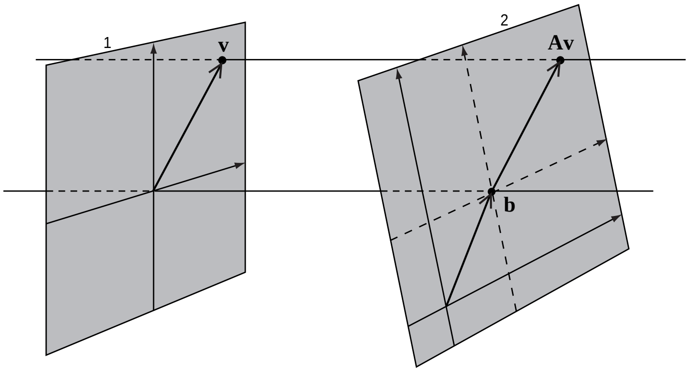

It follows from equation (4) that p must be an affine transformation.

Theorem 5 Each parallel projection is an affine transformation.

The converse is false, for it is *not* true that every affine transformation can be represented as a parallel projection.

For example, consider the so-called doubling map of $\mathbb{R}^2$ to itself given by

$$t(\mathbf{v}) = 2\mathbf{v} \quad (\mathbf{v} \quad \mathbb{R}^2). \qquad (5)$$

This is an affine transformation, since it can be written in the form $t(\mathbf{x}) = \mathbf{Ax}+\mathbf{b}$ with $\mathbf{A} = 2\mathbf{I}$ and $\mathbf{b} = \mathbf{0}$. However, a parallel projection is *either* between two parallel planes, in which case all lengths are unchanged, *or* between two intersecting planes, in which case distances along the line of intersection are unchanged. The doubling map has neither of these properties and so is not a parallel projection.

Observation An affine transformation is not necessarily a parallel projection.

Although the doubling map is not a parallel projection, it *is* possible to double lengths in $\mathbb{R}^2$ by following one parallel projection by another: the first doubles all horizontal lengths, and the second doubles all vertical lengths. Thus the doubling map (5) can be represented as the *composition* of two parallel projections.

We end this subsection by showing that *every* affine transformation can be expressed as a composition of two parallel projections.

Recall that any affine transformation $t : \mathbb{R}^2 \to \mathbb{R}^2$ has the form

$$t(\mathbf{x}) = \mathbf{Ax} + \mathbf{b} \quad (\mathbf{x} \in \mathbb{R}^2), \tag{6}$$

where $\mathbf{A}$ is an invertible 2×2 matrix. Now, t is not a linear transformation unless $\mathbf{b} = \mathbf{0}$, but we can use methods similar to those for linear transformations to determine $\mathbf{A}$ and $\mathbf{b}$.

First, it follows from equation (6) that $t(\mathbf{0}) = \mathbf{b}$; so $\mathbf{b}$ is the image of the origin under t. If we let e and f be the coordinates of $t(\mathbf{0})$, then we can write

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} e \\ f \end{pmatrix},$$

where a, b, c, d are real numbers that have yet to be found. It follows from equation (6) that the images under t of the points (1, 0) and (0, 1) are given by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix} = \begin{pmatrix} a \\ c \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}$$

and

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix} = \begin{pmatrix} b \\ d \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}.$$

So if, in addition to $t(\mathbf{0}) = (e, f)$, we know the points onto which (1, 0) and (0, 1) are mapped by t, then we can determine the values of a, b, c and d. Indeed, we have

$$(a, c) = t(1, 0) - (e, f) \quad \text{and} \quad (b, d) = t(0, 1) - (e, f).$$

Notice that for an affine transformation t, the images $t(1, 0)$, $t(0, 1)$ and $t(0, 0) = (e, f)$ cannot be collinear, for if they were, then (a, c) and (b, d) would be linearly dependent, and $\mathbf{A}$ would not be invertible.

It follows that an affine transformation is *uniquely* determined by its effect on the three non-collinear points (0, 0), (1, 0) and (0, 1). We shall return to this method of determining affine transformations in Section 2.3.

So suppose that a given affine transformation t maps the points (0, 0), (1, 0) and (0, 1) to three non-collinear points P, Q and R, respectively. In order to express t as the composition of two parallel projections p_1 and p_2, we need to define p_1 and p_2 in such a way that $p_2 \circ p_1$ has the same effect as t on (0, 0), (1, 0) and (0, 1). To do this, we first define p_1 so that it maps (0, 0) to P, (1, 0) to Q, and (0, 1) to some point X, say, and then define p_2 so that it maps X to R while leaving P and Q fixed.

Uniqueness then guarantees that $t = p_2 \circ p_1$.

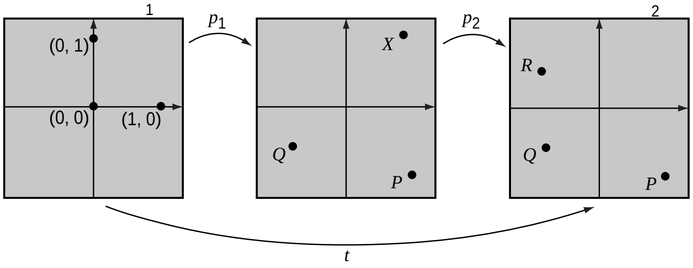

To construct p_1 we embed its domain plane π_1, and its codomain plane π, into $\mathbb{R}^3$ so that the point (0, 0) in π_1 coincides with the point P in π, as shown below. It does not matter how this is done, provided that (1, 0) does not lie in π. We then define p_1 by the family of rays that are parallel to the ray through the point (1, 0) in π_1 and the point Q in π. When defined in this way, p_1 maps (0, 0) to P, (1, 0) to Q, and (0, 1) to some point X, as required.

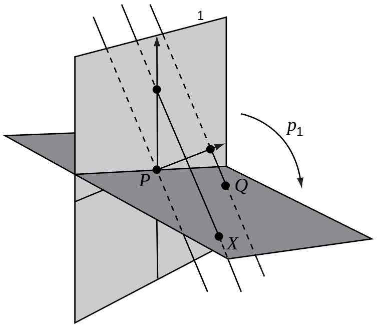

For clarity, we have omitted the axes from the plane π.

To construct p_2 we embed its domain plane π, and its codomain plane π_2, into $\mathbb{R}^3$ so that the points P and Q in π coincide with the points P and Q in π_2, as shown below. Again it does not matter how this is done, provided that X does not lie in π_2. We then define p_2 by the family of rays that are parallel to the ray through the point X in π and the point R in π_2. Then p_2 leaves P and Q fixed and maps X to R.

See the figure below.

Overall, the composite $p_2 \circ p_1$ of the two parallel projections maps (0, 0), (1, 0) and (0, 1) to P, Q and R, respectively. Now p_1 and p_2 are affine transformations, so $p_2 \circ p_1$ is also an affine transformation. Furthermore, $p_2 \circ p_1$ maps

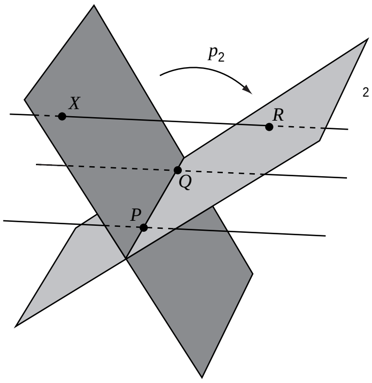

(0, 0), (1, 0) and (0, 1) to the same points as does t. Since such affine transformations are unique, it follows that $t = p_2 \circ p_1$. We have therefore demonstrated the following result.

Theorem 6 An affine transformation can be expressed as the composite of two parallel projections.

An important consequence of this theorem is that all properties of figures that are unchanged by parallel projections must also be unchanged by affine transformations. In particular, the three properties of parallel projections that we met in Subsection 2.2.2 must, in fact, be affine properties.

2.3 Properties of Affine Transformations

In the previous section you saw how parallel projections can be used to explore affine geometry from a visual point of view. In this section we explore some of the same ideas from an algebraic point of view.

2.3.1 Images of Sets Under Affine Transformations

We begin by describing how to find the image of a line under an affine transformation. To do this, recall that an affine transformation is a mapping $t : \mathbb{R}^2 \to \mathbb{R}^2$ given by a formula of the form

$$t(\mathbf{x}) = \mathbf{Ax} + \mathbf{b}, \tag{1}$$

where $\mathbf{A}$ is an invertible 2 × 2 matrix. The set of such transformations forms a group, in which the transformation inverse to t is given by

$$t^{-1}(\mathbf{x}) = \mathbf{A}^{-1}\mathbf{x} - \mathbf{A}^{-1}\mathbf{b}. \tag{2}$$

Subsection 2.2.1

When equations (1) and (2) are used to find images under t, it is easy to confuse points in the domain plane with points in the codomain plane, as both planes are copies of $\mathbb{R}^2$. To avoid such confusion, we often reserve the symbol $\mathbf{x}$ and the coordinates (x, y) for points in the domain of t, and use the symbol $\mathbf{x}'$ and the coordinates (x', y') to denote the image of $\mathbf{x}$ under t.

With this notation, we may rewrite equations (1) and (2) in the form

$$\mathbf{x}' = \mathbf{A}\mathbf{x} + \mathbf{b}, \tag{3}$$

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{x}' - \mathbf{A}^{-1}\mathbf{b}. \tag{4}$$

The next example illustrates how these equations can be used to find the image of a line under an affine transformation.

Example 1 Determine the image of the line $y = 2x$ under the affine transformation

$$t(\mathbf{x}) = \begin{pmatrix} 4 & 1 \\ 2 & 1 \end{pmatrix}\mathbf{x} + \begin{pmatrix} 2 \\ -1 \end{pmatrix} \quad (\mathbf{x} \in \mathbb{R}^2). \tag{5}$$

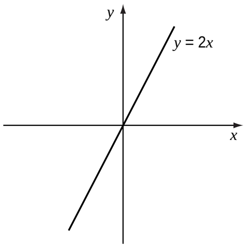

Solution Let (x, y) be an arbitrary point on the line $y = 2x$, and let (x', y') be the image of (x, y) under t. Then

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 4 & 1 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 2 \\ -1 \end{pmatrix}.$$

Next we use equation (4) to express (x, y) in terms of (x', y'). We have

$$\begin{pmatrix} 4 & 1 \\ 2 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -1 & 2 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -1 & 2 \end{pmatrix}\begin{pmatrix} 2 \\ -1 \end{pmatrix} = \begin{pmatrix} \frac{3}{2} \\ -4 \end{pmatrix},$$

Recall that the inverse of the invertible matrix $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is

$$\mathbf{A}^{-1} = \frac{1}{ad - bc} \times \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

so

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -1 & 2 \end{pmatrix}\begin{pmatrix} x' \\ y' \end{pmatrix} + \begin{pmatrix} -\frac{3}{2} \\ 4 \end{pmatrix}.$$

It follows that under the inverse mapping t^{-1} we have

$$x = \tfrac{1}{2}x' - \tfrac{1}{2}y' - \tfrac{3}{2} \quad \text{and} \quad y = -x' + 2y' + 4.$$

Since x and y are related by the equation $y = 2x$, it follows that x' and y' are related by the equation

$$-x' + 2y' + 4 = 2\left(\tfrac{1}{2}x' - \tfrac{1}{2}y' - \tfrac{3}{2}\right),$$

which simplifies to

$$2x' - 3y' = 7.$$

Dropping the dashes, we see that the image of the line $y = 2x$ under t is the line

$$2x - 3y = 7.$$ □

Problem 1 Determine the image of the line $3x \quad y + 1 = 0$ under the affine transformation

$$t(\mathbf{x}) = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 2 \end{pmatrix} \mathbf{x} + \begin{pmatrix} \frac{3}{2} \\ 4 \end{pmatrix} \qquad \mathbf{x} \quad \mathrm{R}^2 .$$

Problem 2 Determine the image of the circle $x^2 + y^2 = 1$ under the affine transformation

$$t(\mathbf{x}) = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 2 \end{pmatrix} \mathbf{x} + \begin{pmatrix} \frac{3}{2} \\ 4 \end{pmatrix} \qquad \mathbf{x} \quad \mathrm{R}^2 .$$

The same technique can be used to find the images of other types of figures, such as other conics. You will meet some examples of this in Section 2.5.

2.3.2 The Fundamental Theorem of Affine Geometry

The algebraic approach can also be used to investigate whether there is an affine transformation which maps one given figure onto another. Recall that if there is such a transformation, then the two figures are said to be **affine-congruent**. This concept of congruence is important because, as we explained in Section 2.1, figures that are affine-congruent to each other share the same affine properties.

In this subsection we prove the remarkable result that *all* triangles are affine-congruent and therefore share the same affine properties. In fact, since a triangle is completely determined by its three vertices, the congruence of triangles follows from the so-called *Fundamental Theorem of Affine Geometry* which states that any three non-collinear points can be mapped to any other three non-collinear points by an affine transformation.

This is very different to Euclidean geometry, where two triangles are congruent only if they have the same shape and size.

First, recall that in Subsection 2.2.3 we described how the points (0, 0), (1, 0) and (0, 1) in R^2 can be mapped to any three non-collinear points P, Q and R by an affine transformation. This transformation is unique in the sense that it is completely determined by the choice of P, Q and R. The following example should remind you of how such transformations are constructed.

There the mapping was constructed in a geometric manner. In this subsection we construct the mapping algebraically.

Example 2 Determine the affine transformation which maps the points (0, 0), (1, 0) and (0, 1) to the points (3, 2), (5, 8) and (7, 3), respectively.

Solution Let t be the affine transformation given by

$$t : \begin{pmatrix} x \\ y \end{pmatrix} \qquad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix} . \tag{6}$$

Since $t(0, 0) = (3, 2)$, it follows from (6) that $e = 3$ and $f = 2$.

Next, $t(1, 0) = (5, 8)$, so it follows from (6) that

$$\begin{pmatrix} 5 \\ 8 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} a \\ c \end{pmatrix} + \begin{pmatrix} 3 \\ 2 \end{pmatrix} .$$

The first column of the matrix for t is therefore

$$\begin{pmatrix} a \\ c \end{pmatrix} = \begin{pmatrix} 5 \\ 8 \end{pmatrix} - \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ 6 \end{pmatrix}.$$

Finally, $t(0,1) = (7,3)$, so that

$$\begin{pmatrix} 7 \\ 3 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} b \\ d \end{pmatrix} + \begin{pmatrix} 3 \\ 2 \end{pmatrix}.$$

The second column of the matrix for t is therefore

$$\begin{pmatrix} b \\ d \end{pmatrix} = \begin{pmatrix} 7 \\ 3 \end{pmatrix} - \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 4 \\ 1 \end{pmatrix}.$$

Hence the desired affine transformation is given by

$$t : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} 2 & 4 \\ 6 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 3 \\ 2 \end{pmatrix}. \qquad \square$$

In general, if we want to find an affine transformation t of the form

$$t : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix} \qquad (7)$$

which maps (0, 0) to **p**, (1, 0) to **q** and (0, 1) to **r**, then we must choose a, b, c, d, e and f so that

$$\begin{aligned} \mathbf{p} &= t(0,0) = (e, f), && \text{so } (e, f) = \mathbf{p}; \\ \mathbf{q} &= t(1,0) = (a,c) + (e, f), && \text{so } (a,c) = \mathbf{q} - \mathbf{p}; \\ \mathbf{r} &= t(0,1) = (b,d) + (e, f), && \text{so } (b,d) = \mathbf{r} - \mathbf{p}. \end{aligned}$$

Notice that any three points **p**, **q** and **r** uniquely determine a transformation t of the form (7), but t is *affine* only if the matrix

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

is invertible. Since the columns of **A** correspond to the vectors **q** − **p** and **r** − **p**, it follows that **A** is invertible only if the vectors **q** − **p** and **r** − **p** are linearly independent. That is, provided that **p**, **q** and **r** are *not* collinear.

So if **p**, **q** and **r** are *not* collinear, then we can use the following strategy to find an affine transformation which maps (0, 0) to **p**, (1, 0) to **q** and (0, 1) to **r**.

Strategy To determine the unique affine transformation $t(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ which maps (0, 0), (1, 0) and (0, 1) to the three non-collinear points **p, q** and **r**, respectively:

1. take $\mathbf{b} = \mathbf{p}$;
2. take **A** to be the matrix with columns given by **q** − **p** and **r** − **p**.

Problem 3 Use the strategy to determine the affine transformation which maps the points (0, 0), (1, 0) and (0, 1) to the points (2, 3), (1, 6) and (3, −1), respectively.

Problem 4 Use the strategy to determine the affine transformation which maps the points (0, 0), (1, 0) and (0, 1) to the points (1, −2), (2, 1) and (−3, 5), respectively.

We shall use the results of these two problems shortly, in Example 3.

Notice that the inverse of the transformation in Problem 3 is an affine transformation which maps the points (2, 3), (1, 6) and (3, −1) to the points (0, 0), (1, 0) and (0, 1), respectively. So if, after applying this inverse, we apply the affine transformation in Problem 4, then the overall effect is that of a composite affine transformation which sends the points (2, 3), (1, 6) and (3, −1) to the points (1, −2), (2, 1) and (−3, 5), respectively.

In a similar way, we can find an affine transformation which sends any three non-collinear points to any other three non-collinear points.

$(2,3)$	$(0,0)$	$(1,-2)$
$(1,6)$	$(1,0)$	$(2,1)$
$(3,-1)$	$(0,1)$	$(-3,5)$

Theorem 1 Fundamental Theorem of Affine Geometry

Let **p**, **q**, **r** and **p′**, **q′**, **r′** be two sets of three non-collinear points in $\mathbb{R}^2$. Then:

(a) there is an affine transformation t which maps **p**, **q** and **r** to **p′**, **q′** and **r′**, respectively;

(b) the affine transformation t is unique.

Proof

(a) Let t_1 be the affine transformation which maps (0, 0), (1, 0) and (0, 1) to the points **p**, **q** and **r**, respectively, and let t_2 be the affine transformation which maps (0, 0), (1, 0) and (0, 1) to the points **p′**, **q′** and **r′**, respectively. Then the composite $t = t_2 \circ t_1^{-1}$ is an affine transformation, and it maps **p**, **q** and **r** to **p′**, **q′** and **r′**, respectively.

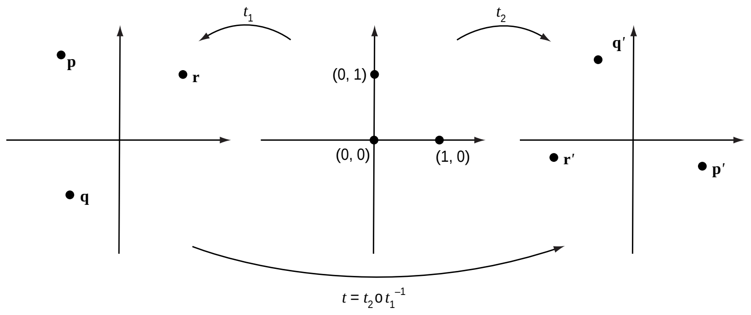

$$\mathbf{p} \xrightarrow{t_1^{-1}} (0,0) \xrightarrow{t_2} \mathbf{p}'$$
$$\mathbf{q} \quad (1,0) \quad \mathbf{q}'$$
$$\mathbf{r} \quad (0,1) \quad \mathbf{r}'$$

(b) Suppose that t and s are both affine transformations which map $\mathbf{p}$, $\mathbf{q}$ and $\mathbf{r}$ to $\mathbf{p}'$, $\mathbf{q}'$ and $\mathbf{r}'$, respectively, and let t_1 be the affine transformation defined in part (a). Then the composites $t \circ t_1$ and $s \circ t_1$ are both affine transformations which map (0, 0), (1, 0) and (0, 1) to $\mathbf{p}'$, $\mathbf{q}'$ and $\mathbf{r}'$, respectively. Since an affine transformation is uniquely determined by its effect on the points (0, 0), (1, 0) and (0, 1), it follows that $t \circ t_1 = s \circ t_1$.

If we then compose both $t \circ t_1$ and $s \circ t_1$ on the right with t_1^{-1}, it follows that $t = s$. Thus the mapping t constructed in part (a) is unique.

$$\begin{array}{lcccc} (0,0) & \xrightarrow{t_1} & \mathbf{p} & \xrightarrow{t \text{ or } s} & \mathbf{p}' \\ (1,0) & \to & \mathbf{q} & \to & \mathbf{q}' \\ (0,1) & \to & \mathbf{r} & \to & \mathbf{r}' \end{array}$$

Now suppose that we are given two arbitrary triangles $\triangle ABC$ and $\triangle DEF$. By the Fundamental Theorem there is an affine transformation which maps the vertices A, B, C to the vertices D, E, F, respectively. Since this transformation maps straight lines to straight lines, it must map the sides of $\triangle ABC$ to the sides of $\triangle DEF$, so we have the following important corollary. This will be used extensively in Section 2.4.

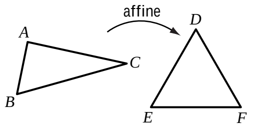

Corollary All triangles are affine-congruent.

In order to find the affine transformation which maps one triangle, vertex to vertex, onto another triangle, we follow the strategy used in part (a) of the proof of the Fundamental Theorem.

Strategy To determine the affine transformation t which maps three non-collinear points $\mathbf{p}$, $\mathbf{q}$ and $\mathbf{r}$ to another three non-collinear points $\mathbf{p}'$, $\mathbf{q}'$ and $\mathbf{r}'$, respectively:

1. determine the affine transformation t_1 which maps (0, 0), (1, 0) and (0, 1) to the points $\mathbf{p}$, $\mathbf{q}$ and $\mathbf{r}$, respectively;
2. determine the affine transformation t_2 which maps (0, 0), (1, 0) and (0, 1) to the points $\mathbf{p}'$, $\mathbf{q}'$ and $\mathbf{r}'$, respectively;
3. calculate the composite $t = t_2 \circ t_1^{-1}$.

Recall that the previous strategy explained how t_1 and t_2 can be determined.

Example 3 Determine the affine transformation which maps the points (2, 3), (1, 6) and (3, −1) to the points (1, −2), (2, 1) and (−3, 5), respectively.

Solution You have already seen in Problem 3 that the affine transformation t_1 which maps the points (0, 0), (1, 0) and (0, 1) to the points (2, 3), (1, 6) and (3, −1), respectively, is given by

$$t_1(\mathbf{x}) = \begin{pmatrix} -1 & 1 \\ 3 & -4 \end{pmatrix} \mathbf{x} + \begin{pmatrix} 2 \\ 3 \end{pmatrix}.$$

Also, in Problem 4 you saw that the affine transformation t_2 which maps the points (0, 0), (1, 0) and (0, 1) to the points (1, 2), (2, 1) and (3, 5), respectively, is given by

$$t_2(\mathbf{x}) = \begin{matrix} 1 & 4 \\ 3 & 7 \end{matrix} \mathbf{x} + \begin{matrix} 1 \\ 2 \end{matrix} .$$

Following the strategy, we need to find the inverse of t_1. We have

$$\begin{matrix} 1 & 1 \\ 3 & 4 \end{matrix}^{1} = \begin{matrix} 4 & 1 \\ 3 & 1 \end{matrix}$$

and

$$\begin{matrix} 4 & 1 \\ 3 & 1 \end{matrix} \begin{matrix} 2 \\ 3 \end{matrix} = \begin{matrix} 11 \\ 9 \end{matrix} ,$$

so that the inverse of t_1 is given by

$$t_1^{\;1}(\mathbf{x}) = \begin{matrix} 4 & 1 \\ 3 & 1 \end{matrix} \mathbf{x} + \begin{matrix} 11 \\ 9 \end{matrix} .$$

Thus the affine transformation which maps the points (2, 3), (1, 6) and (3, 1) to the points (1, 2), (2, 1) and (3, 5), respectively, is given by

$$\begin{aligned} t(\mathbf{x}) &= t_2 \quad t_1^{\;1}(\mathbf{x}) \\ &= t_2 \quad \begin{matrix} 4 & 1 \\ 3 & 1 \end{matrix} \mathbf{x} + \begin{matrix} 11 \\ 9 \end{matrix} \\ &= \begin{matrix} 1 & 4 \\ 3 & 7 \end{matrix} \quad \begin{matrix} 4 & 1 \\ 3 & 1 \end{matrix} \mathbf{x} + \begin{matrix} 11 \\ 9 \end{matrix} \quad + \begin{matrix} 1 \\ 2 \end{matrix} \\ &= \quad \begin{matrix} 8 & 3 \\ 33 & 10 \end{matrix} \mathbf{x} + \begin{matrix} 25 \\ 96 \end{matrix} \quad + \begin{matrix} 1 \\ 2 \end{matrix} \\ &= \quad \begin{matrix} 8 & 3 \\ 33 & 10 \end{matrix} \mathbf{x} + \begin{matrix} 24 \\ 94 \end{matrix} . \end{aligned}$$

□

Problem 5 Determine the affine transformation which maps the points (1, 1), (2, 2) and (3, 4) to the points (8, 13), (3, 4) and (0, 1), respectively.

2.3.3 Proofs of the Basic Properties of Affine Transformations

In Subsection 2.2.2 we used parallel projections to demonstrate that affine transformations have the following basic properties: they map straight lines to straight lines, they map parallel lines to parallel lines, and they preserve ratios of lengths along a given straight line. We now give algebraic proofs of these assertions.

Theorem 2 An affine transformation maps straight lines to straight lines.

Proof

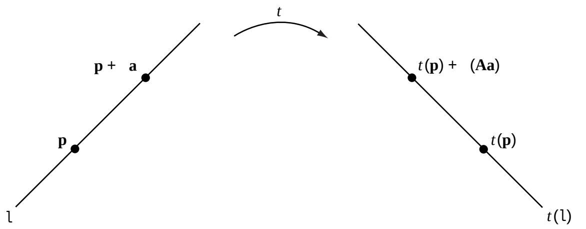
t
$\mathbf{p} + \lambda\mathbf{a}$
$\mathbf{p}$
ℓ
$t(\mathbf{p}) + \lambda(\mathbf{Aa})$
$t(\mathbf{p})$
$t(\ell)$

Let ℓ be a line through a point with position vector $\mathbf{p}$, and let the direction of ℓ be that of some vector $\mathbf{a}$. Then

$$\ell = \{\mathbf{p} + \lambda\mathbf{a} : \lambda \in \mathbb{R}\}.$$

Now let $t : \mathbb{R}^2 \to \mathbb{R}^2$ be an affine transformation given by

$$t(\mathbf{x}) = \mathbf{Ax} + \mathbf{b}.$$

We can find the image under t of an arbitrary point $\mathbf{p} + \lambda\mathbf{a}$ on ℓ as follows:

$$\begin{aligned} t(\mathbf{p} + \lambda\mathbf{a}) &= \mathbf{A}(\mathbf{p} + \lambda\mathbf{a}) + \mathbf{b} \\ &= (\mathbf{Ap} + \mathbf{b}) + \lambda\mathbf{Aa} \\ &= t(\mathbf{p}) + \lambda\mathbf{Aa}. \end{aligned}$$

So the image of ℓ is the set

$$t(\ell) = \{t(\mathbf{p}) + \lambda\mathbf{Aa} : \lambda \in \mathbb{R}\},$$

which is a line through $t(\mathbf{p})$ in the direction of the vector $\mathbf{Aa}$.

Theorem 3 An affine transformation maps parallel straight lines to parallel straight lines.

Proof

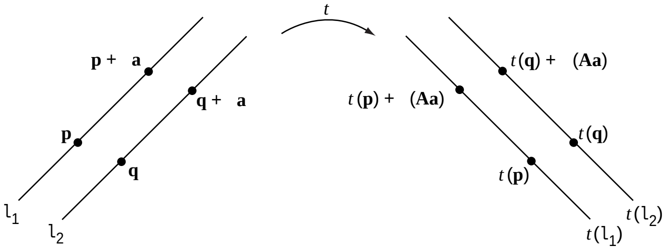
t
$\mathbf{p} + \lambda\mathbf{a}$
$\mathbf{q} + \lambda\mathbf{a}$
$\mathbf{p}$
$\mathbf{q}$
ℓ_1
ℓ_2
$t(\mathbf{q}) + \lambda(\mathbf{Aa})$
$t(\mathbf{p}) + \lambda(\mathbf{Aa})$
$t(\mathbf{q})$
$t(\mathbf{p})$
$t(\ell_2)$
$t(\ell_1)$

Let ℓ_1 and ℓ_2 be parallel lines through the points with position vectors $\mathbf{p}$ and $\mathbf{q}$, respectively, and let the direction of the lines be that of the vector $\mathbf{a}$. Then

$$\ell_1 = \{\mathbf{p} + \lambda\mathbf{a} : \lambda \in \mathbb{R}\} \quad \text{and} \quad \ell_2 = \{\mathbf{q} + \lambda\mathbf{a} : \lambda \in \mathbb{R}\}.$$

As in the proof of Theorem 2, the images of ℓ_1 and ℓ_2 under the affine transformation $t(\mathbf{x}) = \mathbf{Ax} + \mathbf{b}$ are the sets

$$t(\ell_1) = \{t(\mathbf{p}) + \lambda\mathbf{Aa} : \lambda \in \mathbb{R}\} \quad \text{and} \quad t(\ell_2) = \{t(\mathbf{q}) + \lambda\mathbf{Aa} : \lambda \in \mathbb{R}\}.$$

These sets are straight lines which pass through the image points $t(\mathbf{p})$ and $t(\mathbf{q})$, both in the same direction as that of the vector $\mathbf{Aa}$. Hence the two image lines under t are parallel, as claimed.

Rather than prove that affine transformations preserve ratios of lengths along a given straight line, as in Property 3 of Subsection 2.2.2, we prove the following more general result illustrated in the margin. The original result follows because any line is parallel to itself.

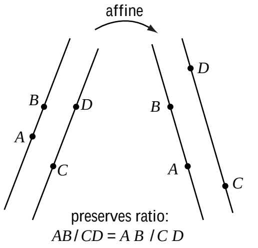

Theorem 4 An affine transformation preserves ratios of lengths along parallel straight lines.

Proof We begin by examining what happens to the length of a line segment under an affine transformation.

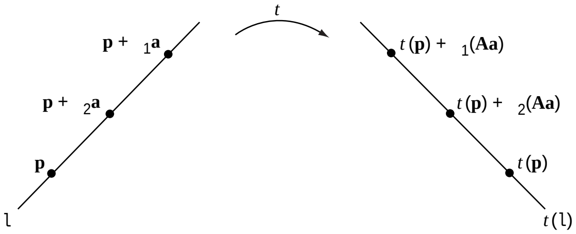

Let ℓ be a line through a point with position vector $\mathbf{p}$, and let the direction of ℓ be that of some *unit* vector $\mathbf{a}$. Then

$$\ell = \{\mathbf{p} + \lambda\mathbf{a} : \lambda \in \mathbb{R}\}.$$

As in the proof of Theorem 2, the image of ℓ under the affine transformation $t(\mathbf{x}) = \mathbf{Ax} + \mathbf{b}$ is the line

$$t(\ell) = \{t(\mathbf{p}) + \lambda\mathbf{Aa} : \lambda \in \mathbb{R}\}.$$

Now consider a segment of ℓ with endpoints $\mathbf{p} + \lambda_1\mathbf{a}$ and $\mathbf{p} + \lambda_2\mathbf{a}$. Since $\mathbf{a}$ is a unit vector, the length of the segment is

$$\|(\mathbf{p} + \lambda_2\mathbf{a}) - (\mathbf{p} + \lambda_1\mathbf{a})\| = |\lambda_2 - \lambda_1| \bullet \|\mathbf{a}\| = |\lambda_2 - \lambda_1|.$$

Recall that $\|\mathbf{a}\|$ means the length of $\mathbf{a}$.

The image of the segment has endpoints $t(\mathbf{p}) + \lambda_1\mathbf{Aa}$ and $t(\mathbf{p}) + \lambda_2\mathbf{Aa}$, so the image of the segment has length

$$\|(t(\mathbf{p}) + \lambda_2\mathbf{Aa}) - (t(\mathbf{p}) + \lambda_1\mathbf{Aa})\| = |\lambda_2 - \lambda_1| \bullet \|\mathbf{Aa}\|.$$

So, in the process of mapping segments along ℓ to segments along $t(\ell)$, lengths are stretched by the factor $\|\mathbf{Aa}\|$. Since this factor is the same for all segments which lie along lines parallel to $\mathbf{a}$, it follows that the *ratios* of lengths along parallel lines are unchanged by t.

2.4 Using the Fundamental Theorem of Affine Geometry

In this section we explain how the Fundamental Theorem of Affine Geometry can be used to deduce the fact that the medians of any triangle are concurrent from the special case that the medians of an equilateral triangle are concurrent. We then use similar methods to prove the classical theorems of Ceva and Menelaus.

These results are named after Giovanni Ceva (Italian mathematician, 1647/48–1734) and Menelaus of Alexandria (Greek geometer, 1st Century AD).

2.4.1 The Median Theorem

Let $\triangle ABC$ be an arbitrary triangle in the plane. If you join the midpoint of each side of the triangle to the opposite vertex (these lines are called the *medians* of the triangle), these three lines appear to pass through a single point. In fact, no matter what triangle you choose, you find that its medians meet in a single point.

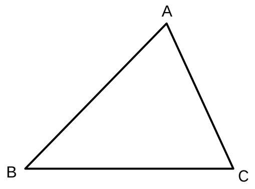

Theorem 1 Median Theorem
The medians of any triangle are concurrent.

We can get some evidence that this theorem holds in general by looking first at a special case where a proof of the theorem is straight-forward — namely, when the triangle is an equilateral triangle.

This technique of looking first to see whether a result holds in a special case is often useful.

To do this, consider an equilateral triangle $\triangle ABC$, with medians AP, BQ and CR. Since $\triangle ABC$ has sides of equal length, it must be symmetric about the line AP. Thus the point at which BQ meets CR must be symmetrically placed with respect to this line — that is, it must actually lie on the line AP. In other words, the lines AP, BQ and CR are concurrent if the triangle is equilateral.

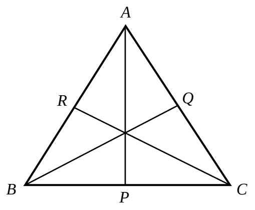

In order to show that the medians of an arbitrary triangle meet at a point, consider an arbitrary triangle $\triangle ABC$, and let P, Q and R be the midpoints of the sides BC, CA and AB, respectively. Next, choose a particular *equilateral* triangle $\triangle A'B'C'$, and let P', Q' and R' be the midpoints of the sides $B'C'$, $C'A'$ and $A'B'$, respectively.

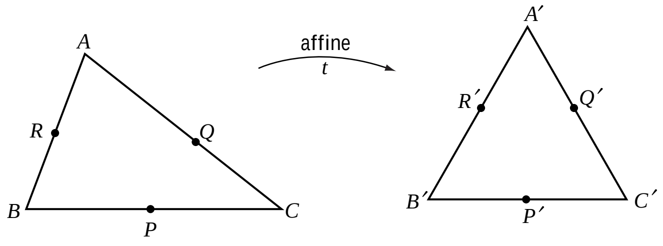

According to the Fundamental Theorem of Affine Geometry there is an affine transformation t which maps $\triangle ABC$ onto $\triangle A'B'C'$. Moreover, since affine transformations preserve ratios of lengths along lines it follows that t maps the mid-points P, Q and R to the mid-points P', Q' and R', respectively.

From the above discussion we know that the medians of any *equilateral* triangle meet at a point, so in particular we know that $A'P'$, $B'Q'$ and $C'R'$ meet at some point X', say, as shown on the right below.

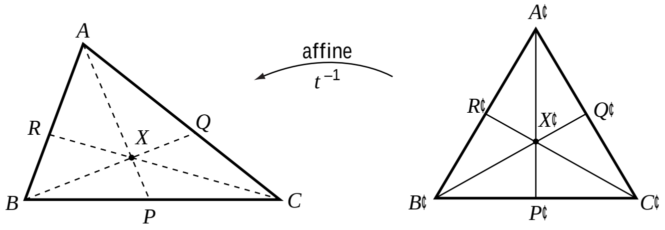

The trick now is to observe that t has an inverse t^{-1} which is also an affine transformation. This inverse maps the medians $A'P'$, $B'Q'$ and $C'R'$ back to the medians AP, BQ and CR of the original triangle $\triangle ABC$. Since X' lies on all three of the lines $A'P'$, $B'Q'$ and $C'R'$ it follows that t^{-1} maps X' to some point X which lies on all three of the lines AP, BQ and CR. In other words, the medians of $\triangle ABC$ are concurrent.

Since $\triangle ABC$ is an arbitrary triangle we have proved the Median Theorem.

The essence of the above proof is the fact that all triangles are affine-congruent. That powerful result enables us to prove theorems concerning the affine properties of triangles (such as collinearity, lines being parallel, and ratios of lengths along a given line) following a standard pattern. First, we choose a particular type of triangle for which it is easy to prove the result. Then, by asserting the existence of an affine transformation from that triangle to an arbitrary triangle, we deduce that the result holds for all triangles.

The basic affine properties were listed in Subsection 2.2.1.

This is the approach we shall use to prove the theorems of Ceva and Menelaus later in the section.

2.4.2 Ceva's Theorem

We now prove the following theorem due to Ceva.

Theorem 2 Ceva's Theorem
Let $\triangle ABC$ be a triangle, and let X be a point which does not lie on any of its (extended) sides. If AX meets BC at P, BX meets CA at Q and CX meets BA at R, then

$$\frac{AR}{RB} \cdot \frac{BP}{PC} \cdot \frac{CQ}{QA} = 1.$$

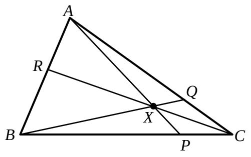

Proof According to the Fundamental Theorem of Affine Geometry there is an affine transformation t which maps the points A, B, C to the points $A' = (0,1)$, $B' = (0,0)$, $C' = (1,0)$, respectively. This transformation maps the triangle $\triangle ABC$ onto the right-angled triangle $\triangle A'B'C'$, and it maps the point X to some point $X' = (u, v)$.

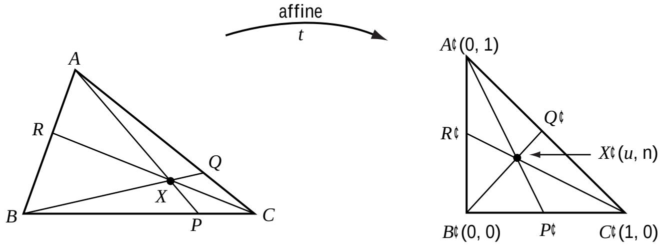

Using coordinate geometry we can calculate the equations of the lines $A\ X$, $B\ X$, $C\ X$ and hence find the coordinates of the point P where $A\ X$ meets $B\ C$, of the point Q where $B\ X$ meets $A\ C$, and of the point R where $C\ X$ meets $A\ B$.

Starting with the point P, we note that the line $B\ C$ has equation $y = 0$.

Also, the line $A\ X$ has slope $\frac{1}{0\quad u}$, so its equation is $y \quad 1 = \frac{1}{0\quad u}(x \quad 0)$. Hence, at the point P where the two lines meet, we must have $y = 0$ and $y \quad 1 = \frac{1}{0\quad u}(x \quad 0)$, so

$$P = \frac{u}{1}, 0 \quad .$$

Similarly, at the point R we have $x = 0$, and $y \quad 0 = \frac{0}{1\quad u}(x \quad 1)$, so

$$R = 0, \frac{}{1 \quad u} \quad .$$

Finally, at Q we have $x + y = 1$ and $y = \frac{}{u}x$, so $x = \frac{u}{u+}$ and $y = \frac{}{u+}$. Hence

$$Q = \frac{u}{u+}, \frac{}{u+} \quad .$$

Thus, using the coordinate formulas for calculating ratios we obtain

These formulas are given at the beginning of Appendix 2 just above the Section Formula.

$$\frac{A\ R}{R\ B} = \frac{y_R \quad y_A}{y_B \quad y_R} = \frac{\frac{}{1\ u} \quad 1}{0 \quad \frac{}{1\ u}} = \frac{u + \quad 1}{},$$

$$\frac{B\ P}{P\ C} = \frac{x_P \quad x_B}{x_C \quad x_P} = \frac{\frac{u}{1} \quad 0}{1 \quad \frac{u}{1}} = \frac{u}{1 \quad u},$$

and

$$\frac{C\ Q}{Q\ A} = \frac{y_Q \quad y_C}{y_A \quad y_Q} = \frac{\frac{}{u+} \quad 0}{1 \quad \frac{}{u+}} = \frac{}{u}.$$

Hence

$$\frac{A\ R}{R\ B} \bullet \frac{B\ P}{P\ C} \bullet \frac{C\ Q}{Q\ A} = 1.$$

Since $t\ ^{1}$ is an affine transformation, it preserves ratios along a line. It must therefore map P, Q, R back to the points P, Q, R in such a way that

$$\frac{AR}{RB} \bullet \frac{BP}{PC} \bullet \frac{CQ}{QA} = 1,$$

as required.

The next example illustrates how we can use Ceva's Theorem to calculate certain unknown distances along the sides of a triangle. For the method to work correctly, it is important to remember that all the ratios in Ceva's Theorem are *signed* ratios. Thus, if X lies inside the triangle, as in part (a) of the example, then all the ratios are positive. But if X lies outside the triangle, as in part (b), then two of the ratios will be negative.

Example 1

(a) In the figure on the left below, $AR = 1$, $RB = 2$, $BP = 3$, $CQ = 2$ and $QA = 2$. Calculate the distance PC.
(b) For the figure on the right, $AR = 1$, $AB = 3$, $PC = 1$, $CQ = 2$ and $QA = 2$. Calculate the distance BC.

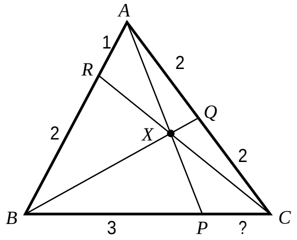

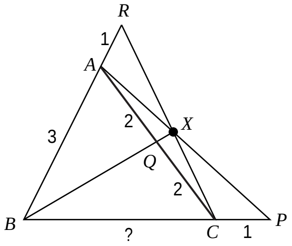

Solution

(a) By Ceva's Theorem, we have

$$\frac{AR}{RB} \cdot \frac{BP}{PC} \cdot \frac{CQ}{QA} = 1;$$

so,

$$\frac{1}{2} \cdot \frac{3}{PC} \cdot \frac{2}{2} = 1.$$

It follows that $PC = \frac{3}{2}$.

(b) By Ceva's Theorem, we have

$$\frac{AR}{RB} \cdot \frac{BP}{PC} \cdot \frac{CQ}{QA} = 1;$$

so,

$$\frac{1}{4} \cdot \quad \frac{BC + 1}{1} \quad \cdot \frac{2}{2} = 1.$$

It follows that $BC = 3$. □

Problem 1

(a) Determine the ratio $\frac{BP}{PC}$ in the left diagram below, given that

$$\frac{AR}{RB} = \frac{AQ}{QC} = \frac{3}{2}.$$

(b) Determine the ratio $\frac{CQ}{QA}$ in the middle diagram below, given that

$$\frac{AR}{RB} = \frac{1}{2} \quad \text{and} \quad \frac{BP}{PC} = \frac{2}{7}.$$

(c) Determine the ratio $\frac{AR}{RB}$ in the right diagram below, given that

$$\frac{BP}{PC} = \frac{5}{7} \quad \text{and} \quad \frac{CQ}{QA} = 7.$$

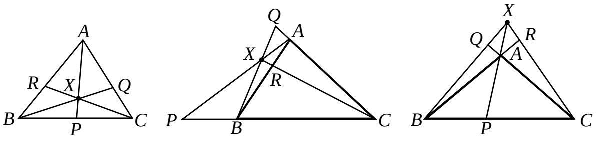

Ceva's Theorem has the following converse, which can be regarded as a generalization of the Median Theorem to configurations where P, Q, R are not all midpoints of sides.

Theorem 3 Converse to Ceva's Theorem
Let P, Q and R be points, other than vertices, on the (possibly extended) sides BC, CA and AB of a triangle $\triangle ABC$, such that

$$\frac{AR}{RB} \cdot \frac{BP}{PC} \cdot \frac{CQ}{QA} = 1. \tag{1}$$

Then the lines AP, BQ and CR are concurrent.

In the Median Theorem,

$$\frac{AR}{RB} = 1, \quad \frac{BP}{PC} = 1, \quad \frac{CQ}{QA} = 1,$$

so Theorem 3 generalizes the Median Theorem.

Proof Let the lines BQ and CR intersect at a point X, and let the line AX meet BC at some point P'. It is sufficient to prove that $P = P'$.

It follows from Ceva's Theorem that

$$\frac{AR}{RB} \cdot \frac{BP'}{P'C} \cdot \frac{CQ}{QA} = 1. \tag{2}$$

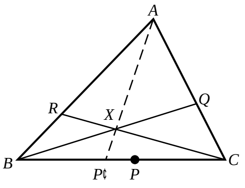

Hence, from equations (1) and (2), we have

$$\frac{BP}{PC} = \frac{BP'}{P'C},$$

so that P and P' must indeed be the same point.

Example 2 The triangle $\triangle ABC$ has vertices $A(1,3)$, $B(-1,0)$ and $C(4,0)$, and the points $P(0,0)$, $Q\left(\frac{8}{3}, \frac{4}{3}\right)$ and $R\left(-\frac{2}{3}, \frac{1}{2}\right)$ lie on BC, CA and AB, respectively.

(a) Determine the ratios in which P, Q and R divide the sides of the triangle.
(b) Determine whether the lines AP, BQ and CR are concurrent.

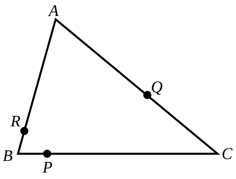

Solution

(a) Using the coordinate formulas for calculating ratios, we obtain

$$\frac{AR}{RB} = \frac{x_R - x_A}{x_B - x_R} = \frac{\frac{2}{3} - 1}{1 + \frac{2}{3}} = 5, \quad \frac{BP}{PC} = \frac{x_P - x_B}{x_C - x_P} = \frac{0 + 1}{4 - 0} = \frac{1}{4},$$

$$\frac{CQ}{QA} = \frac{x_Q - x_C}{x_A - x_Q} = \frac{\frac{8}{3} - 4}{1 - \frac{8}{3}} = \frac{4}{5}, \tag{3}$$

so that P divides BC in the ratio $1:4$, Q divides CA in the ratio $4:5$ and R divides AB in the ratio $5:1$.

(b) It follows from (3) that the product

$$\frac{AR}{RB} \cdot \frac{BP}{PC} \cdot \frac{CQ}{QA} = 5 \cdot \frac{1}{4} \cdot \frac{4}{5} = 1;$$

so by the converse to Ceva's Theorem the lines AP, BQ and CR must be concurrent. □

Problem 2 The triangle △ABC has vertices $A(-1,1)$, $B(2,-1)$ and $C(3,2)$, and the points $P\left(\frac{8}{3},1\right)$, $Q\left(2,\frac{7}{4}\right)$ and $R\left(\frac{4}{5},-\frac{1}{5}\right)$ lie on BC, CA and AB, respectively.

(a) Determine the ratios in which P, Q and R divide the sides of the triangle.

(b) Determine whether the lines AP, BQ and CR are concurrent.

2.4.3 Menelaus' Theorem

Ceva's theorem is concerned with lines through the vertices of a triangle that meet at a point. We now use the Fundamental Theorem of Affine Geometry to prove an analogous theorem due to Menelaus which is concerned with points on the sides of a triangle that are collinear.

Theorem 4 Menelaus' Theorem

Let △ABC be a triangle, and let ℓ be a line that crosses the sides BC, CA, AB at three distinct points P, Q, R, respectively. Then

$$\frac{AR}{RB} \cdot \frac{BP}{PC} \cdot \frac{CQ}{QA} = -1.$$

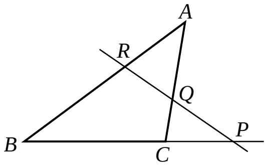

Proof According to the Fundamental Theorem of Affine Geometry there is an affine transformation t which maps the points A, B, C to the points $A'(0,1)$, $B'(0,0)$, $C'(1,0)$, respectively. This transformation maps the triangle △ABC

onto the right-angled triangle $A'B'C'$, and it maps the line ℓ to some line ℓ'. Let the equation of ℓ' be $y = mx + c$.

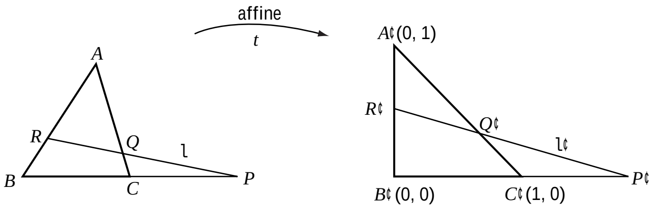

We now calculate the coordinates of the points P', Q' and R' where ℓ' meets the sides $B'C'$, $C'A'$ and $A'B'$, respectively.

At P' we have $y = 0$ and $y = mx + c$. This implies that $x = -\frac{c}{m}$, and hence

$$P' = \left(-\frac{c}{m}, 0\right).$$

At R' we have $x = 0$ and $y = mx + c$. This implies that $y = c$, and hence

$$R' = (0, c).$$

At Q' we have $x + y = 1$ and $y = mx + c$. This implies that $1 - x = mx + c$ so that $x = \frac{1-c}{m+1}$; also $y = m(1 - y) + c$, so that $y = \frac{m+c}{m+1}$; and hence

$$Q' = \left(\frac{1-c}{m+1}, \frac{m+c}{m+1}\right).$$

Using the coordinate formulas for calculating ratios we obtain

$$\frac{A'R'}{R'B'} = \frac{y_{R'} - y_{A'}}{y_{B'} - y_{R'}} = \frac{c-1}{0-c} = -\frac{c-1}{c},$$

$$\frac{B'P'}{P'C'} = \frac{x_{P'} - x_{B'}}{x_{C'} - x_{P'}} = \frac{-\frac{c}{m} - 0}{1 + \frac{c}{m}} = -\frac{c}{m+c},$$

and

$$\frac{C'Q'}{Q'A'} = \frac{x_{Q'} - x_{C'}}{x_{A'} - x_{Q'}} = \frac{\frac{1-c}{m+1} - 1}{0 - \frac{1-c}{m+1}} = \frac{-(m+c)}{c-1}.$$

Hence,

$$\frac{A'R'}{R'B'} \cdot \frac{B'P'}{P'C'} \cdot \frac{C'Q'}{Q'A'} = -1.$$

Since t^{-1} is an affine transformation, it preserves ratios along a line. It must therefore map P', Q', R' back to the points P, Q, R in such a way that

$$\frac{AR}{RB} \cdot \frac{BP}{PC} \cdot \frac{CQ}{QA} = -1,$$

as required.

Remark

As for Ceva's Theorem, it is important to remember that all the ratios in Menelaus' Theorem are *signed* ratios. In fact if ℓ passes through the interior of the triangle, then precisely one of the ratios is negative; otherwise all the ratios are negative.

Example 3

(a) In the figure on the left below: $AR = 1$, $RB = 2$, $BC = 2$, $CQ = 1$ and $QA = 1$. Calculate the distance PC.

(b) In the figure on the right below: $AR = 2$, $AB = 1$, $BC = 2$, $CA = 2$ and $BP = 2$. Calculate the distance QA.

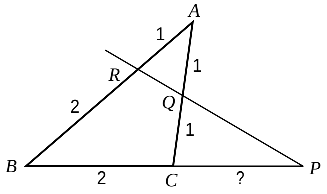

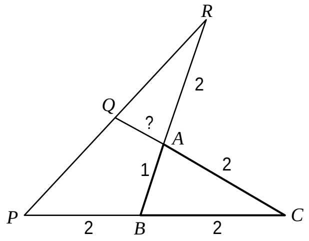

Solution

(a) By Menelaus' Theorem, we have

$$\frac{AR}{RB} \bullet \frac{BP}{PC} \bullet \frac{CQ}{QA} = 1.$$

So

$$\frac{1}{2} \bullet \frac{2+PC}{PC} \bullet \frac{1}{1} = 1.$$

It follows that $2 + PC = 2PC$, and hence $PC = 2$.

(b) By Menelaus' Theorem, we have

$$\frac{AR}{RB} \bullet \frac{BP}{PC} \bullet \frac{CQ}{QA} = 1.$$

So

$$\frac{2}{3} \cdot \frac{2}{4} \cdot \frac{2+QA}{QA} = 1.$$

It follows that $2 + QA = 3QA$, and hence $QA = 1$. □

Problem 3

(a) Determine the ratio $\frac{CQ}{QA}$ in the left diagram below, given that

$$\frac{AR}{RB} = 2 \quad \text{and} \quad \frac{BP}{PC} = 2.$$

(b) Determine the ratio $\frac{CQ}{QA}$ in the right diagram below, given that

$$\frac{AR}{RB} = \frac{1}{4} \quad \text{and} \quad \frac{BP}{PC} = 2.$$

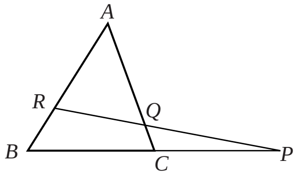

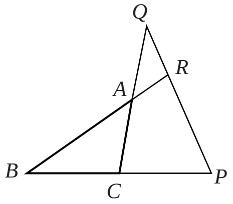

Menelaus Theorem has a converse that enables us to check whether points on the three sides of a triangle are collinear.

Theorem 5 Converse to Menelaus' Theorem
Let *P*, *Q* and *R* be points other than vertices on the (possibly extended) sides *BC*, *CA* and *AB* of a triangle *ABC*, such that

$$\frac{AR}{RB}\bullet\frac{BP}{PC}\bullet\frac{CQ}{QA} = 1. \tag{3}$$

Then the points *P*, *Q* and *R* are collinear.

Proof Let the line that passes through *Q* and *R* meet *BC* at some point *P* . It is sufficient to prove that $P = P$.

It follows from Menelaus Theorem that

$$\frac{AR}{RB}\bullet\frac{BP}{P\ C}\bullet\frac{CQ}{QA} = 1. \tag{4}$$

Hence, from equations (3) and (4) we deduce that

$$\frac{BP}{PC} = \frac{BP}{P\ C}.$$

It follows that *P* and *P* must indeed be the same point.

The strategy of the proof is the same as that of Theorem 3.

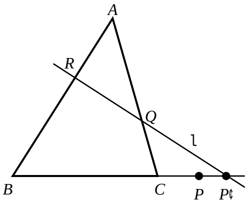

Problem 4 The triangle *ABC* has vertices $A(2,4)$, $B(\ 2,0)$ and $C(1,0)$, and the points $P\ \frac{5}{2},0$, $Q\ \frac{3}{2},2$ and $R(1,3)$ lie on *BC*, *CA* and *AB*, respectively.

(a) Determine the ratios in which *P*, *Q* and *R* divide the sides of the triangle.
(b) Hence determine whether the points *P*, *Q* and *R* are collinear.

We end this subsection with two revision problems.

Problem 5 Let *ABC* be a triangle, and let *X* be a point which does not lie on any of its (extended) sides. Also, let *AX* meet *BC* at *P*, *BX* meet *CA* at *Q* and *CX* meet *BA* at *R*; and let *QR* and *BC* meet at *T*.

Given that $\frac{BP}{PC} = k$, determine $\frac{BT}{TC}$ in terms of *k*.

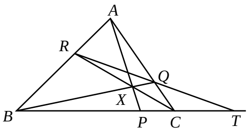

Problem 6 Suppose that P and Q are the midpoints of the sides AB and BC of a parallelogram $ABCD$, and that the lines DP and AQ meet at R.

(a) Determine the image of B under the affine transformation t which maps A, D and C to (0, 1), (0, 0) and (1, 0), respectively.

(b) By considering the image of $ABCD$ under t, determine the ratios $PR : RD$ and $AR : RQ$.

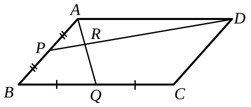

2.4.4 Barycentric Coordinates

In this subsection we introduce a new coordinate system, of *barycentric coordinates* with respect to a *triangle of reference*, for points in the plane, which can simplify some calculations. Rather than use two perpendicular axes to determine the coordinates of an arbitrary point in the plane, we use a weighted sum of the coordinates of three non-collinear points the vertices of the triangle of reference.

Barycentric coordinates were introduced by Möbius in 1827.

Definitions Let $A = (a_1, a_2)$, $B = (b_1, b_2)$ and $C = (c_1, c_2)$ be three non-collinear points in the plane $\mathbb{R}^2$; we will call ABC the **triangle of reference**. Then a point (x, y) in the plane has **barycentric coordinates** $(\ , \ , \ )$ with respect to ABC if

$$\begin{aligned} x &= \ a_1 + \ b_1 + \ c_1, \\ y &= \ a_2 + \ b_2 + \ c_2, \text{ and} \\ 1 &= \quad + \quad + \quad . \end{aligned} \tag{5}$$

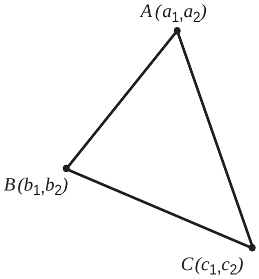

Remark

In particular, notice that the barycentric coordinates of the vertices A, B and C of the triangle of reference ABC are (1,0,0), (0,1,0) and (0,0,1), respectively.

Thus, for example, if the triangle of reference ABC has vertices $A = (1, 2)$, $B = (\ 3, 0)$ and $C = (\ 2, 4)$, then the point with barycentric coordinates $\frac{1}{2}, \frac{3}{4}, \ \frac{1}{4}$ has cartesian coordinates

Cartesian coordinates are the standard Euclidean coordinates in the plane.

$$\begin{aligned} &\tfrac{1}{2} \bullet (1) + \tfrac{3}{4} \bullet (\ 3) \quad \tfrac{1}{4} \bullet (\ 2), \tfrac{1}{2} \bullet (2) + \tfrac{3}{4} \bullet (0) \quad \tfrac{1}{4} \bullet 4 \\ &= \frac{2 \quad 9 + 2}{4}, \frac{4 + 0 \quad 4}{4} \\ &= \ \tfrac{5}{4}, 0 \ . \end{aligned}$$

We may express the formula (5) for barycentric coordinates in terms of matrices in the form

$$\begin{matrix} x \\ y \\ 1 \end{matrix} = \mathbf{M} \qquad , \quad \text{where } \mathbf{M} = \begin{matrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ 1 & 1 & 1 \end{matrix} . \tag{6}$$

Since $A = (a_1, a_2)$, $B = (b_1, b_2)$ and $C = (c_1, c_2)$ are non-collinear, the determinant $\det \mathbf{M}$ is non-zero. Then, since $\mathbf{M}$ is non-singular, we can reformulate the representation (6) as

We do not prove that $\det \mathbf{M} = 0$ here.

$$= \mathbf{M}^{\ 1} \begin{matrix} x \\ y \\ 1 \end{matrix} . \tag{7}$$

Example 4 Determine barycentric coordinates for the point (2,1) with respect to the triangle of reference ABC where $A = (1, 0)$, $B = (1,\ \ 1)$ and $C = (\ \ 1, 1)$.

Solution The matrix $\mathbf{M}$ for the triangle of reference ABC is

$$\mathbf{M} = \begin{matrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{matrix} ,$$

whose inverse is

We omit the details of the calculation of this inverse.

$$\begin{matrix} 1 & 1 & 0 \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{matrix} .$$

It follows from the representation (7) that the point (2,1) has barycentric coordinates with respect to the triangle of reference ABC given by

$$\begin{matrix} 1 & 1 & 0 \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{matrix} \begin{matrix} 2 \\ 1 \\ 1 \end{matrix} = \begin{matrix} 3 \\ \frac{3}{2} \\ \frac{1}{2} \end{matrix} ;$$

namely, barycentric coordinates $3,\ \ \frac{3}{2},\ \ \frac{1}{2}$. □

Problem 7 Determine barycentric coordinates for the point $(\ \ 1, 1)$ with respect to the triangle of reference ABC where $A = (1, 1)$, $B = (2, 2)$ and $C = (1, 2)$.

Next, we give barycentric versions of the condition for collinearity of three points in the plane and of the equation of a line in the plane.

Theorem 6 The points P, Q and R with barycentric coordinates $({}_1, {}_1, {}_1)$, $({}_2, {}_2, {}_2)$ and $({}_3, {}_3, {}_3)$ are collinear if and only if

$$\begin{matrix} {}_1 & {}_2 & {}_3 \\ {}_1 & {}_2 & {}_2 \\ {}_1 & {}_2 & {}_3 \end{matrix} = 0.$$

Proof Let the points P, Q and R have cartesian coordinates (x_1, y_1), (x_2, y_2) and (x_3, y_3), respectively. It follows that, if the triangle of reference ABC has vertices $A = (a_1, a_2)$, $B = (b_1, b_2)$ and $C = (c_1, c_2)$, then we may apply the formula (5) to each of P, Q and R in turn to obtain

$$x_1 = a_1 {}_1 + b_1 {}_1 + c_1 {}_1 \text{ and } y_1 = a_2 {}_1 + b_2 {}_1 + c_2 {}_1,$$

$$x_2 = a_1 {}_2 + b_1 {}_2 + c_1 {}_2 \text{ and } y_2 = a_2 {}_2 + b_2 {}_2 + c_2 {}_2,$$

$$x_3 = a_1 {}_3 + b_1 {}_3 + c_1 {}_3 \text{ and } y_3 = a_2 {}_3 + b_2 {}_3 + c_2 {}_3.$$

We may write these simultaneous equations in matrix form as

$$\begin{matrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{matrix} = \begin{matrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ 1 & 1 & 1 \end{matrix} \begin{matrix} {}_1 & {}_2 & {}_3 \\ {}_1 & {}_2 & {}_3 \\ {}_1 & {}_2 & {}_3 \end{matrix}$$

$$= \mathbf{M} \begin{matrix} {}_1 & {}_2 & {}_3 \\ {}_1 & {}_2 & {}_3 \\ {}_1 & {}_2 & {}_3 \end{matrix}, \qquad (8)$$

The results for P, Q and R are set out as columns 1, 2 and 3, respectively, of the left-hand matrix.

where $\mathbf{M} = \begin{matrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ 1 & 1 & 1 \end{matrix}$.

Now, **M** is non-singular since the points A, B and C are non-collinear. Then it follows from equation (8) that

$$\begin{matrix} {}_1 & {}_2 & {}_3 \\ {}_1 & {}_2 & {}_3 \\ {}_1 & {}_2 & {}_3 \end{matrix} = 0 \text{ if and only if } \begin{matrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{matrix} = 0.$$

But the second determinant equation here is the condition for P, Q and R to be collinear. The desired result then follows immediately.

Corollary The line through the points with barycentric coordinates $({}_1, {}_1, {}_1)$ and $({}_2, {}_2, {}_2)$ has equation

$$\begin{matrix} {}_1 & {}_2 \\ {}_1 & {}_2 \\ {}_1 & {}_2 \end{matrix} = 0.$$

That is, an equation in terms of barycentric coordinates.

For example, if the triangle of reference is *ABC*, the equation of the line *AB* in terms of barycentric coordinates is

$$\begin{matrix} 1 & 0 & \\ 0 & 1 & \\ 0 & 0 & \end{matrix} \quad = 0;$$

which simplifies to the equation $= 0$.

Similarly, *BC* and CA have equations $= 0$ and $= 0$.

Problem 8 Determine which of the following sets of points, described by their barycentric coordinates with respect to an unspecified triangle of reference, are collinear; for those that are collinear, determine the equation of the line on which they lie.

(a) $(1, 1, \quad 1), (4, \quad 2, \quad 1), \quad \frac{1}{2}, 2, \quad \frac{3}{2}$

(b) $(1, 1, \quad 1), (2, \quad 2, 1), (\quad 1, 7, \quad 5)$

Next, we meet a version of the Section Formula in terms of barycentric coordinates.

The Section Formula is given in Appendix 2.

Theorem 7 Section Formula

The point *R* that divides the line joining the points *P* and *Q* with barycentric coordinates $({}_1, {}_1, {}_1)$ and $({}_2, {}_2, {}_2)$ in the ratio $(1 \quad) : \quad$ has barycentric coordinates

$$(\ , \ , \ ) = \ ({}_1, {}_1, {}_1) + (1 \quad)({}_2, {}_2, {}_2).$$

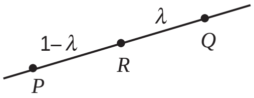

Proof Let *P*, *Q* and *R* have cartesian coordinates (x_1, y_1), (x_2, y_2) and (x, y), respectively. It follows from the cartesian form of the Section Formula that

$$(x, y) = \ (x_1, y_1) + (1 \quad)(x_2, y_2), \quad \text{for some real number} \quad ;$$

we can rewrite this equation in matrix form as

$$\begin{matrix} x \\ y \\ 1 \end{matrix} = \ \begin{matrix} x_1 \\ y_1 \\ 1 \end{matrix} + (1 \quad) \begin{matrix} x_2 \\ y_2 \\ 1 \end{matrix} .$$

Multiplying both sides of this equation on the left by the matrix $\mathbf{M}^{\ 1}$ and using the formula (8), it follows that

$$= \ \begin{matrix} {}_1 \\ {}_1 \\ {}_1 \end{matrix} + (1 \quad) \begin{matrix} {}_2 \\ {}_2 \\ {}_2 \end{matrix} .$$

This is the desired result.

We can now use barycentric coordinate methods to give further proofs of Menelaus s Theorem and Ceva s Theorem.

Theorem 8 Menelaus' Theorem
Let ABC be a triangle, and let be a line that crosses the sides *BC*, *CA* and *AB* at three distinct points *P*, *Q*, *R*, respectively. Then

$$\frac{AR}{RB} \cdot \frac{BP}{PC} \cdot \frac{CQ}{QA} = \quad 1.$$

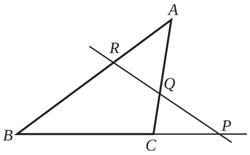

Proof Define , μ and as follows:

$$\frac{BP}{PC} = \frac{1 \quad}{\quad}, \quad \frac{CQ}{QA} = \frac{1 \quad \mu}{\mu}, \quad \frac{AR}{RB} = \frac{1 \quad}{\quad}.$$

Hence, by the Section Formula, P has barycentric coordinates

$$\begin{aligned} P &= \quad (0,1,0) + (1 \quad\ )(0,0,1) \\ &= (0, \ , 1 \quad\ ). \end{aligned}$$

For, P divides BC in the ratio (1) : .

Similarly, Q and R have barycentric coordinates $(1 \quad \mu, 0, \mu)$ and $(\ , 1 \ , 0)$, respectively.

Then, by Theorem 6 the points P, Q and R are collinear if and only if

We omit the details of these calculations.

$$\begin{matrix} 0 & 1 \quad \mu & \\ & 0 & 1 \\ 1 & \mu & 0 \end{matrix} = 0.$$

Expanding this determinant, we have that P, Q and R are collinear if and only if

$$(1 \quad \mu) \begin{matrix} & 1 \\ 1 & 0 \end{matrix} + \begin{matrix} & 0 \\ 1 & \mu \end{matrix} = 0;$$

that is, if and only if

$$(1 \quad\ )(1 \quad \mu)(1 \quad\ ) + \quad \mu \quad = 0,$$

or

$$\frac{1 \quad}{\quad} \cdot \frac{1 \quad \mu}{\mu} \cdot \frac{1 \quad}{\quad} = \quad 1.$$

From the original definition of , μ and , it follows that P, Q and R are collinear if and only if

$$\frac{AR}{RB} \cdot \frac{BP}{PC} \cdot \frac{CQ}{QA} = \quad 1.$$

From our assumption that P, Q and R are collinear, the desired result follows.

Notice that we have also obtained a proof of the converse of Menelaus Theorem, since this equality is an if and only if result!

Theorem 9 Ceva's Theorem

Let *ABC* be a triangle, and let *X* be a point which does not lie on any of its (extended) sides. If *AX* meets *BC* at *P*, *BX* meets *CA* at Q and *CX* meets *BA* at *R*, then

$$\frac{AR}{RB} \bullet \frac{BP}{PC} \bullet \frac{CQ}{QA} = 1.$$

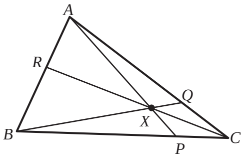

Proof We use the same notation , μ and as in the previous proof, so that again *P*, *Q* and *R* have homogeneous coordinates $(0,\ ,1\quad)$, $(1\quad\mu,0,\mu)$ and $(\ ,1\quad,0)$, respectively.

Then, using the Corollary to Theorem 6 the equation of *AP* is

$$\begin{matrix} 1 & & 0 \\ 0 & & \\ 0 & 1 & \end{matrix} = 0,$$

which simplifies to

$$\frac{\quad}{1\quad} = 0.$$

This gives that the equation of *AP* is $\ (1\quad)\ = 0$, or

$$= \frac{1}{\quad}.$$

Similarly, we find that the equation of *BQ* is $= \frac{1\ \ \mu}{\mu}$.

Since *AP* and *BQ* meet at *X*, it follows that the barycentric coordinates $(\ ,\ ,\ )$ of *X* must satisfy the equations $= \frac{1}{\quad}$ and $= \frac{1\ \ \mu}{\mu}$; hence its barycentric coordinates are

We omit the details of this calculation.

$$\frac{1}{\quad} \bullet \frac{1\quad\mu}{\mu}\ ,\ ,\ \frac{1}{\quad}\ ,\quad \text{for some}\ = 0.$$

Then *C*, *X* and *R* are collinear if and only if

$$\begin{matrix} 0 & \frac{1}{\quad} \bullet \frac{1\quad\mu}{\mu} & \\ 0 & & 1 \\ 1 & \frac{1}{\quad} & 0 \end{matrix} = 0;$$

this simplifies to

$$\frac{1}{\quad} \bullet \frac{1\quad\mu}{\mu} \bullet \frac{1}{\quad} = 1.$$

It follows that C, X and R are collinear — that is, that AP, BQ and CR are concurrent — if and only if

$$\frac{AR}{RB} \cdot \frac{BP}{PC} \cdot \frac{CQ}{QA} = 1.$$

From our assumption that AP, BQ and CR are concurrent, the desired result follows.

Notice that we have also obtained a proof of the converse of Ceva's Theorem, since this equality is an 'if and only if' result!

2.5 Affine Transformations and Conics

2.5.1 Classifying Non-Degenerate Conics in Affine Geometry

In Section 2.2 you saw that under an affine transformation a straight line maps to a straight line. Indeed, it follows from the Fundamental Theorem of Affine Geometry that any straight line can be mapped to any other straight line by some affine transformation. We now explore the corresponding situation for conics.

We discussed conics in Chapter 1.

Recall that a conic is a set in $\mathbb{R}^2$ given by an equation of the form

$$Ax^2 + Bxy + Cy^2 + Fx + Gy + H = 0, \qquad (1)$$

where A, B, C, F, G and H are real numbers, and A, B and C are not all zero. The three types of non-degenerate conic are ellipses, parabolas and hyperbolas. A non-degenerate conic is an ellipse if $B^2 - 4AC < 0$, a parabola if $B^2 - 4AC = 0$, and a hyperbola if $B^2 - 4AC > 0$.

You met the $B^2 - 4AC$ test for conics in Theorem 3 of Section 1.3.

First, consider the case where equation (1) represents an ellipse, as illustrated on the left of the figure below. We can apply a translation to move the centre of the ellipse to the origin, and then a rotation to align its major and minor axes with the directions of the x-axis and y-axis, respectively. After we have applied these two Euclidean transformations, the equation of the ellipse becomes

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \quad a \geq b > 0. \qquad (2)$$

If we now apply the affine transformation $t_1 : (x, y) \mapsto (x', y')$, where

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 1/a & 0 \\ 0 & 1/b \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix},$$

then $x' = x/a$ and $y' = y/b$, so equation (2) becomes

$$(x')^2 + (y')^2 = 1.$$

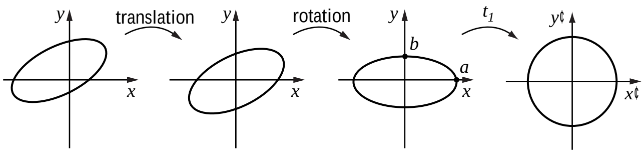

Since the translation, the rotation and the transformation t_1 are all affine, their composite must also be affine. Overall, this shows that each ellipse can be mapped onto the unit circle by an affine transformation. We therefore have the following theorem.

Theorem 1 Every ellipse is affine-congruent to the unit circle with equation $x^2 + y^2 = 1$.

Secondly, consider the case where equation (1) represents a hyperbola, as illustrated on the left of the figure below. Again, we can apply a translation to move the centre of the hyperbola to the origin, and then a rotation to align its major and minor axes with the directions of the x-axis and y-axis, respectively. After we have applied these two transformations, the equation of the hyperbola becomes

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1. \tag{3}$$

Under the affine transformation t_1 defined above, equation (3) becomes

$$(x')^2 - (y')^2 = 1,$$

that is,

$$(x' - y')(x' + y') = 1. \tag{4}$$

Finally, if we apply the affine transformation $t_2 : (x', y') \mapsto (x'', y'')$, where

$$\begin{pmatrix} x'' \\ y'' \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x' \\ y' \end{pmatrix},$$

then equation (4) becomes

$$x''y'' = 1.$$

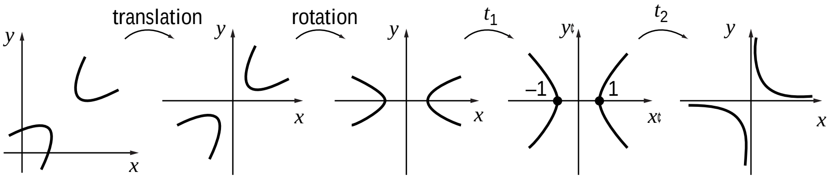

Dropping the dashes from the equation $x''y'' = 1$, we obtain the following theorem.

Theorem 2 Every hyperbola is affine-congruent to the rectangular hyperbola with equation $xy = 1$.

Recall that 'rectangular' means that the asymptotes of the hyperbola are at right angles to each other.

Finally, consider the case where equation (1) represents a parabola, as illustrated on the left of the figure below. We can apply a translation to move the vertex of the parabola to the origin, and then a rotation to align its axis with the (positive) x-axis. After we have applied these two Euclidean transformations, the equation of the parabola becomes

$$y^2 = ax, \tag{5}$$

where a is some positive number which depends on the coefficients in equation (1).

Here we omit the details of the particular transformations involved, and concentrate instead on the principles underlying the successive mappings which are used.

Next, if we apply the affine transformation $t_3 : (x, y) \mapsto (x', y')$, where

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 1/a & 0 \\ 0 & 1/a \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix},$$

then $x' = x/a$ and $y' = y/a$, so equation (5) becomes $(y'a)^2 = a(x'a)$, or

$$(y')^2 = x'.$$

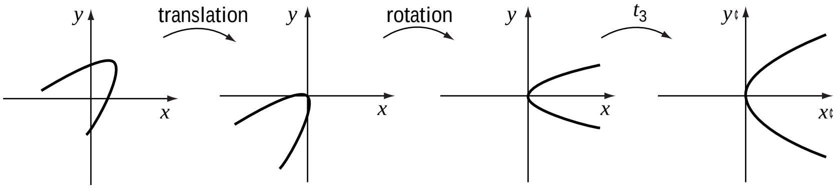

Dropping the dashes, we obtain the following theorem.

Theorem 3 Every parabola is affine-congruent to the parabola with equation $y^2 = x$.

Since all parabolas are affine-congruent to $y^2 = x$, they must be affine-congruent to each other. Similarly, by Theorem 1, all ellipses must be affine-congruent to each other; and, by Theorem 2, all hyperbolas must be affine-congruent to each other.

This raises the question as to whether it is possible for one type of conic (such as an ellipse) to be affine-congruent to another type of conic (such as a hyperbola). The next theorem shows that this cannot happen. In fact, since an affine transformation can be expressed as the composite of two parallel projections, this should not surprise you. After all, no parallel projection can change a bounded curve (such as an ellipse) into an unbounded one (such as a parabola or a hyperbola); nor can it change a curve with two branches (a hyperbola) into a curve with just one branch (an ellipse or a parabola).

Theorem 4 Affine transformations map ellipses to ellipses, parabolas to parabolas, and hyperbolas to hyperbolas.

Remember that a circle is a special type of ellipse.

Proof Consider the non-degenerate conic with equation

$$Ax^2 + Bxy + Cy^2 + Fx + Gy + H = 0, \tag{6}$$

and its image under an affine transformation $t : \mathbf{x} \mapsto \mathbf{x}'$ given by

$$\mathbf{x}' = \mathbf{A}\mathbf{x} + \mathbf{b},$$

where $\mathbf{A}$ is an invertible 2 × 2 matrix.

The inverse affine transformation $t^{-1} : \mathbf{x}' \mapsto \mathbf{x}$ is given by

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{x}' - \mathbf{A}^{-1}\mathbf{b},$$

You met this formula for the inverse in Subsection 2.2.1.

which we may write in the form

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} p & q \\ r & s \end{pmatrix} \begin{pmatrix} x' \\ y' \end{pmatrix} + \begin{pmatrix} u \\ v \end{pmatrix},$$

for some real numbers p, q, r, s, u and v. It follows that

$$x = px' + qy' + u \quad \text{and} \quad y = rx' + sy' + v. \tag{7}$$

If we now substitute these expressions for x and y into equation (6), then the resulting equation is a second-degree equation in x' and y', so the image of the conic under the affine transformation t must be another conic.

We omit the details of these calculations, as they are complicated and uninformative.

Next we show that this image conic cannot be degenerate. A degenerate image would consist of a pair of lines, a single line, a point, or the empty set. Since the affine transformation t^{-1} maps lines to lines, it would map the degenerate image to another degenerate conic. But this cannot happen since t^{-1} maps the image back to the original non-degenerate conic (6). It follows that the image of (6) cannot be degenerate.

Theorem 2 of Section 2.3

Finally, if we substitute for x and y from equations (7) into equation (6), and keep careful track of the algebra involved, it turns out that the discriminant of the image conic is just

$$(ps - rq)^2(B^2 - 4AC).$$

Recall that the sign of the discriminant of a non-degenerate conic determines the type of the conic.

Here $B^2 - 4AC$ is the discriminant of the original conic. Since $(ps - rq)^2 > 0$, the sign of the discriminant is not changed by an affine transformation of a conic. Hence the type of the conic is also unchanged.

Here $ps - rq = 0$ since $\mathbf{A}$ is invertible.

We can combine the results of Theorems 1–4 to obtain the following corollary.

Corollary In affine geometry:

(a) all ellipses are congruent to each other;
(b) all hyperbolas are congruent to each other;
(c) all parabolas are congruent to each other.

Non-degenerate conics are congruent only to non-degenerate conics of the same type.

The corollary shows that affine-congruence partitions the set of non-degenerate conics into three disjoint equivalence classes. One class consists of all the ellipses, another class consists of all the hyperbolas, and the third consists of all the parabolas. Each class contains one of the so-called *standard* conics $x^2 + y^2 = 1$, $xy = 1$ and $y^2 = x$.

Just as the Fundamental Theorem of Affine Geometry enables us to deduce a given result about an arbitrary triangle by showing that the result holds for an equilateral triangle, so the corollary enables us to deduce a given result about an arbitrary ellipse, hyperbola or parabola by showing that the result holds for the corresponding standard conic. Of course, this works only if the result is concerned with the affine properties of the conic, so we need to be able to recognize such properties.

The following theorem shows that one such property is the property of being the centre of an ellipse or hyperbola.

Theorem 5 Let t be an affine transformation, and let C be an ellipse or hyperbola with centre R. Then $t(C)$ has centre $t(R)$.

Recall that a parabola does not have a centre.

Proof Let C' and R' be the images of C and R under t. If P' is any point on C', then it must be the image of some point P on C. Since R is the centre of C, we can rotate P about R through an angle π to a point Q which must also lie on C. Hence $Q' = t(Q)$ is a point on C'.

This uses the definition of centre given in Chapter 1.

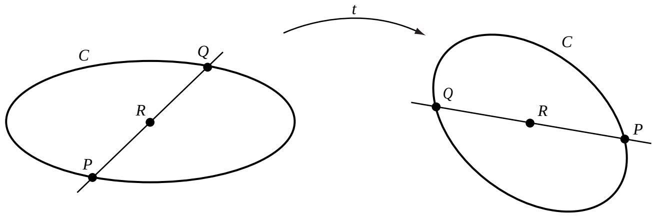

This figure illustrates the proof for an ellipse C, but the proof works equally well for a hyperbola.

Now t preserves ratios of lengths along lines, so the line segment PRQ maps onto the line segment $P'R'Q'$ with $P'R' = R'Q'$. Thus if we rotate P' about R' through an angle π, it must go to Q' on C'. Now, as our choice for P' as a point on C' varies, so do $P = t^{-1}(P')$ and Q, but the point R is always the same point. It follows that the midpoint of $P'Q'$ is always the same point $R' = t(R)$. Hence $R' = t(R)$ is the centre of C', as required.

For since $PR/RQ = 1$ it follows that $P'R'/R'Q' = 1$.

Another affine property is the property of being an asymptote of a hyperbola.

Theorem 6 Let t be an affine transformation, and let H be a hyperbola with asymptotes ℓ_1 and ℓ_2. Then $t(H)$ has asymptotes $t(\ell_1)$ and $t(\ell_2)$.

The figure below illustrates that this theorem is plausible for parallel projections.

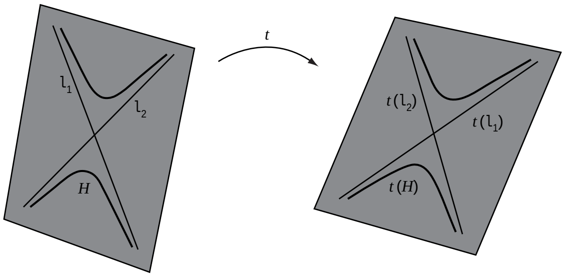

Proof The hyperbola H possesses exactly two (distinct) families of parallel lines each of which fills the plane, with each member of each family meeting H exactly once apart from one line in each family that is an asymptote of H, and so does not meet H.

The image of H under the affine transformation t is also a hyperbola, $t(H)$. The images under t of the two families of parallel lines are also (distinct) families of parallel lines; within each family, a line that meets H once is mapped onto a line that meets $t(H)$ once, and the single line that does not meet H maps onto a line that does not meet $t(H)$. So the two exceptional lines in the image families must be the asymptotes of the hyperbola $t(H)$.

It follows that the asymptotes of H are mapped by t to the asymptotes of $t(H)$, as required.

Many of the problems concerning conics which are particularly amenable to solution using the methods of affine geometry involve tangents.

This is due to the following theorem, which asserts that tangency is an affine property.

Theorem 7 Let t be an affine transformation, and let be a tangent to a conic C. Then $t(\ )$ is a tangent to the conic $t(C)$.

The figure below illustrates the theorem for parallel projections.

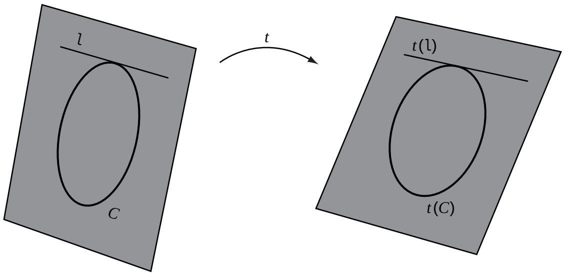

Solution We shall use the fact that a tangent to a conic (whether it is an ellipse, a hyperbola or a parabola) intersects the conic at exactly one point.

However we have to be a little careful. For example, any line parallel to its axis meets a parabola in exactly one point.

First, the image of an ellipse E under an affine transformation t is an ellipse. A tangent to E is a line that intersects E in exactly one point. These properties remain unchanged under an affine projection; hence the image of a tangent to E under an affine transformation t must be a tangent to $t(E)$.

This characterizes tangents to ellipses.

Next, the image of a hyperbola H under an affine transformation t is a hyperbola. A tangent to H is a member of a family of parallel lines that fill the plane such that there are lines in the family that meet H twice, once and not at all; there are exactly two lines in the family that meet H exactly once, and these are tangents to H. The image of the family of lines under t is again a family of parallel lines that fill the plane; it contains lines that meet the parabola $t(H)$ twice and not at all, and exactly two lines that meet H exactly once. These lines are the images of the original tangents to H, and must themselves be tangents to $t(H)$. Hence, the image of a tangent to H under an affine transformation t must be a tangent to $t(H)$.

This characterizes tangents to hyperbolas.

Finally, the image of a parabola P under an affine transformation t is a parabola. A tangent to P is a member of a family of parallel lines that fill the plane such that there are lines in the family that meet P twice, once and not at all; the tangent is the unique member of the family that meets P exactly once. The image of the family of lines under t is again a family of parallel lines that fill the plane; it contains lines that meet the parabola $t(P)$ twice and not at all, and a single line that meets P exactly once. This line is the image of the original tangent to P, and must itself be a tangent to $t(P)$. Hence, the image of a tangent to P under an affine transformation t must be a tangent to $t(P)$.

This characterizes tangents to parabolas.

This completes the proof.

In applications we often use the following facts that you met earlier.

Theorem 2, Subsection 1.2.1.

Tangents to Conics in Standard Form The equation of the tangent to a standard conic at the point (x_1, y_1) is as follows.

Conic	*Tangent*
Unit circle $x^2 + y^2 = 1$	$xx_1 + yy_1 = 1$
Rectangular hyperbola $xy = 1$	$xy_1 + yx_1 = 2$
Parabola $y^2 = x$	$2yy_1 = x + x_1$

2.5.2 Applying Affine Geometry to Conics

We are now in a position to apply the methods of affine geometry to the solution of problems involving conics. Of course, affine geometry can be helpful in this task only if the property being investigated is one which is preserved under affine transformations. The underlying idea is that we use an affine transformation to map the original conic onto one of our standard conics, tackle the problem in hand there, and then map back to the original conic.

We used these techniques in Subsection 2.2.3 to prove the Conjugate Diameters Theorem for the ellipse.

Example 1 *AB* is a diameter of an ellipse. Prove that the tangents to the ellipse at *A* and *B* are parallel to the diameter conjugate to *AB*.

Recall that the diameter conjugate to *AB* is the set of midpoints of all the chords parallel to *AB* (see Subsection 2.2.3).

Solution First, map the ellipse onto the unit circle, by an affine transformation *t*. Since the centre *O* of the ellipse maps to the centre O' of the circle, the image of the diameter *AB* is a diameter $A'B'$ of the unit circle.

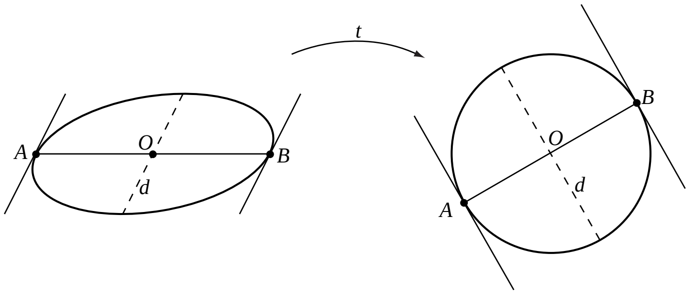

All chords of the circle that are parallel to the tangents at A' and B' are bisected by $A'B'$, and so the diameter through O' is the diameter conjugate to $A'B'$. Since parallel lines map to parallel lines and ratios along parallel lines are preserved under the inverse affine transformation t^{-1}, it follows that all chords of the ellipse that are parallel to the tangents at *A* and *B* are bisected by *AB*, and so the diameter through *O* that is parallel to the tangents at *A* and *B* is the diameter conjugate to *AB*. □

Problem 1 An ellipse touches the sides *BC*, *CA* and *AB* of $\triangle ABC$ at the points *P*, *Q* and *R*, respectively. Prove that

$$\frac{AR}{RB} \cdot \frac{BP}{PC} \cdot \frac{CQ}{QA} = 1,$$

and deduce that the lines *AP*, *BQ* and *CR* are concurrent.

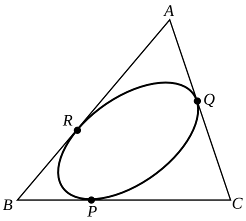

Problem 2 The tangents to an ellipse at two points *A* and *B* meet at a point *T*. Prove that the line joining *T* to the centre *O* of the ellipse bisects the chord *AB*.

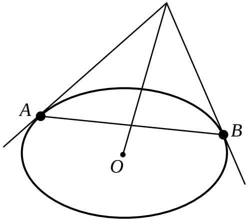

The rectangular hyperbola $H = \{(x, y) : xy = 1\}$ does not possess as much symmetry as does the unit circle; so the fact that every hyperbola is affine-congruent to *H* may not be sufficient to simplify a given problem. Fortunately, however, we can also arrange for any given point on the original hyperbola to map to the point (1, 1) on *H*.

To see this, note that for any non-zero number a, the affine transformation

$$t_a : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} a & 0 \\ 0 & 1/a \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

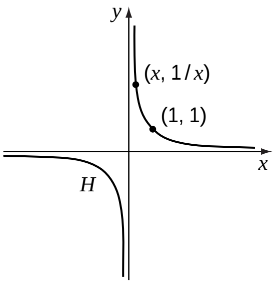

maps H to itself. For, an arbitrary point on H has coordinates of the form $(x, 1/x)$, $x \neq 0$, and under t_a this is mapped to the point $(ax, 1/ax)$, which also lies on H. As x varies through $\mathbb{R} - \{0\}$, its image $(ax, 1/ax)$ varies over the whole of H, so the image of H under t_a is the whole of H.

So if we start with a given hyperbola and a point P on it, we can map the hyperbola to H by some affine transformation s. The point $s(P)$ will then have coordinates $(b, 1/b)$ for some number $b \in \mathbb{R} - \{0\}$; so if we choose $a = 1/b$, then the affine transformation t_a will map $s(P)$ to (1,1). Overall, the composite $t = t_a \circ s$ is an affine transformation which maps the given hyperbola to H, and maps P to (1, 1). We now state this as a corollary to Theorem 2.

Corollary Given any hyperbola and a point P on it, there is an affine transformation which maps the hyperbola onto the rectangular hyperbola $xy = 1$, and the point P to (1, 1).

Example 2 The tangent at the point P on a hyperbola meets the asymptotes at the points A and B. Prove that $PA = PB$.

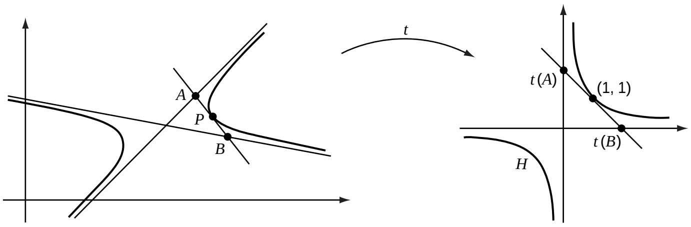

Solution Let t be an affine transformation which maps the hyperbola onto the rectangular hyperbola $H = \{(x, y) : xy = 1\}$ in such a way that $t(P) = (1, 1)$. Then, by Theorem 6 of Subsection 2.5.1, the asymptotes of the hyperbola map to the asymptotes of H; and, by Theorem 7 of Subsection 2.5.1, the tangent at P maps to the tangent at (1,1).

By symmetry, (1, 1) is the midpoint of the line segment from $t(A)$ to $t(B)$. Since midpoints are preserved under the affine transformation t^{-1}, it follows that P is the midpoint of AB. □

Problem 3 P is a point on a hyperbola H with centre O. Prove that there exists a line ℓ through O such that all chords of the hyperbola which are parallel to ℓ are bisected by OP.

This result is an analogue for the hyperbola of the Conjugate Diameters Theorem for the ellipse (Theorem 3 of Subsection 2.2.3).

2.6 Exercises

Section 2.1

1. Let ABC be a triangle in which $AB = AC$. Prove that

$$ABC = \quad ACB.$$

Hint: Consider a reflection in the bisector of BAC.

2. Determine which of the following transformations $t : \mathbb{R}^2 \quad \mathbb{R}^2$ are Euclidean transformations.

(a) $t(\mathbf{x}) = \begin{matrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{matrix} \mathbf{x} + \begin{matrix} 3 \\ 1 \end{matrix}$

(b) $t(\mathbf{x}) = \begin{matrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} \end{matrix} \mathbf{x} + \begin{matrix} 3 \\ 2 \end{matrix}$

(c) $t(\mathbf{x}) = \begin{matrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{matrix} \mathbf{x} + \begin{matrix} 2 \\ 3 \end{matrix}$

3. The Euclidean transformations t_1 and t_2 are given by

$$t_1(\mathbf{x}) = \begin{matrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{matrix} \mathbf{x} + \begin{matrix} 1 \\ 1 \end{matrix}$$

and

$$t_2(\mathbf{x}) = \begin{matrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{matrix} \mathbf{x} + \begin{matrix} 2 \\ 1 \end{matrix}.$$

Determine the composites $t_1 \quad t_2$ and $t_2 \quad t_1$.

4. Determine the inverse of each of the following Euclidean transformations.

(a) $t(\mathbf{x}) = \begin{matrix} \frac{5}{13} & \frac{12}{13} \\ \frac{12}{13} & \frac{5}{13} \end{matrix} \mathbf{x} + \begin{matrix} 4 \\ 5 \end{matrix}$

(b) $t(\mathbf{x}) = \begin{matrix} \frac{12}{13} & \frac{5}{13} \\ \frac{5}{13} & \frac{12}{13} \end{matrix} \mathbf{x} + \begin{matrix} 1 \\ 1 \end{matrix}$

5. The Euclidean transformations t_1 and t_2 are given by

$$t_1(\mathbf{x}) = \begin{matrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{matrix} \mathbf{x} + \begin{matrix} 1 \\ 1 \end{matrix}$$

and

$$t_2(\mathbf{x}) = \begin{matrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{matrix} \mathbf{x} + \begin{matrix} 1 \\ 1 \end{matrix}.$$

Determine the composite $t_2{}^{1} \quad t_1$.

Section 2.2

1. Determine whether or not each of the following transformations $t : \mathbb{R}^2$ $\mathbb{R}^2$ is an affine transformation.
   (a) $t(\mathbf{x}) = \begin{matrix} 2 & 2 \\ 3 & 3 \end{matrix} \mathbf{x} + \begin{matrix} 2 \\ 1 \end{matrix}$
   (b) $t(\mathbf{x}) = \begin{matrix} 5 & 2 \\ 2 & 5 \end{matrix} \mathbf{x} + \begin{matrix} 3 \\ 1 \end{matrix}$
   (c) $t(\mathbf{x}) = \begin{matrix} 1 & 1 \\ 1 & 2 \end{matrix} \mathbf{x}$
2. Write down an example (if one exists) of each type of transformation $t : \mathbb{R}^2 \quad \mathbb{R}^2$ described below. In each case, justify your answer.
   (a) An affine transformation t which is not a Euclidean transformation
   (b) A Euclidean transformation t which is not an affine transformation
   (c) A transformation t which is both Euclidean and affine
   (d) A transformation t which is one one, but is neither Euclidean nor affine
3. The affine transformations t_1 and t_2 are given by

$$t_1(\mathbf{x}) = \begin{matrix} 2 & 3 \\ 1 & 1 \end{matrix} \mathbf{x} + \begin{matrix} 1 \\ 1 \end{matrix}$$

   and

$$t_2(\mathbf{x}) = \begin{matrix} 1 & 2 \\ 1 & 1 \end{matrix} \mathbf{x} + \begin{matrix} 1 \\ 1 \end{matrix}.$$

   Determine the following composites.
   (a) $t_1 \quad t_2$ (b) $t_2 \quad t_1$ (c) $t_1 \quad t_1$
4. Determine the inverse of each of the following affine transformations.
   (a) $t(\mathbf{x}) = \begin{matrix} 2 & 3 \\ 3 & 5 \end{matrix} \mathbf{x} + \begin{matrix} 2 \\ 4 \end{matrix}$ (b) $t(\mathbf{x}) = \begin{matrix} 3 & 2 \\ 4 & 2 \end{matrix} \mathbf{x} + \begin{matrix} 1 \\ 2 \end{matrix}$
5. Prove that the transformation

$$t(\mathbf{x}) = 3\mathbf{x} \ (\mathbf{x} \quad \mathbb{R}^2)$$

   is an affine transformation, but not a parallel projection.
6. Which of the following are affine properties?
   (a) distance (b) collinearity (c) circularity
   (d) magnitude of angle (e) midpoint of line segment

Section 2.3

1. The affine transformation $t : \mathbb{R}^2 \quad \mathbb{R}^2$ is given by

$$t(\mathbf{x}) = \begin{matrix} 1 & 1 \\ 2 & 3 \end{matrix} \mathbf{x} + \begin{matrix} 2 \\ 4 \end{matrix}.$$

   Determine the image under t of each of the following lines.
   (a) $y = \quad 2x$ (b) $2y = 3x \quad 1$
2. The affine transformation $t : \mathbb{R}^2 \quad \mathbb{R}^2$ is given by

$$t(\mathbf{x}) = \begin{matrix} 4 & 5 \\ 1 & 1 \end{matrix} \mathbf{x} + \begin{matrix} 1 \\ 1 \end{matrix}.$$

Determine the image under t of each of the following lines.
(a) $2x - 5y + 3 = 0$ (b) $3x + y - 4 = 0$

3. Determine the affine transformation which maps the points (0, 0), (1, 0) and (0, 1) to the points:
   (a) (0, −1), (1, 1) and (−1, 1), respectively;
   (b) (−4, −5), (1, 7) and (2, −9), respectively.
4. Determine the affine transformation which maps the points (1, 1), (3, 2) and (4, 1) to the points (0, 1), (1, 2) and (3, 7), respectively.
5. Determine the affine transformation which maps the points (1, −1), (5, −4) and (−2, 1) to the points (1, 1), (4, 0) and (0, 2), respectively.
6. Prove that the affine transformation t for which
$$t(\mathbf{x}) = \begin{pmatrix} -1 & 2 \\ 3 & -2 \end{pmatrix} \mathbf{x}$$
maps each point of the line $y = x$ in $\mathbb{R}^2$ onto itself.
7. Determine the matrices **A** and **b** for the affine transformation
$$t(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b},$$
where **A** and **b** are 2 × 2 and 2 × 1 matrices, respectively, given that t maps each point of the line $y = 0$ onto itself and (0,1) onto (2,3). Prove also that t is a parallel projection of $\mathbb{R}^2$ onto itself.

Section 2.4

1. The points P, Q, R and S lie on a line, in that order; the distances between them are 4 units, 2 units and 3 units, respectively. Determine the ratios *PR* : *RS* and *PS* : *SQ*.
2. A point X lies inside a triangle $\triangle ABC$, and the lines *AX*, *BX* and *CX* meet the opposite sides of the triangle at P, Q and R, respectively. The ratios *AR* : *AB* and *BP* : *BC* are 1 : 5 and 3 : 7, respectively. Determine the ratio *AC* : *AQ*.
3. Let be a line that crosses the sides *BC*, *CA* and *AB* of a triangle $\triangle ABC$ at three distinct points *P*, *Q* and *R*, respectively. The ratios *BC* : *CP* and *CQ* : *QA* are 3 : 2 and 1 : 3, respectively. Determine the ratio *AR* : *RB*.
4. *ABCD* is a parallelogram, and the point P divides *AB* in the ratio 2 : 1; the lines *AC* and *DP* meet at *Q*, and the lines *BQ* and *AD* meet at *R*.
   (a) Determine the images of P, Q and R under the affine transformation t which maps *A*, *D* and *C* to (0, 1), (0, 0) and (1, 0), respectively.
   (b) By considering the image of *ABCD* under t, determine the ratios *BQ* : *QR* and *AR* : *RD*.
5. The triangle $\triangle ABC$ has vertices $A(-1,2)$, $B(-3,-1)$ and $C(3,1)$, and the points $P\left(1,\frac{1}{3}\right)$, $Q\left(1,\frac{3}{2}\right)$ and $R\left(-\frac{5}{3},1\right)$ lie on *BC*, *CA* and *AB*, respectively.
   (a) Determine the ratios in which P, Q and R divide the sides of the triangle.
   (b) Determine whether or not the lines *AP*, *BQ* and *CR* are concurrent.

6. The triangle ABC has vertices $A(2,0)$, $B(\;3,0)$ and $C(3,\;3)$, and the points $P(\;1,\;1)$, $Q(1,3)$ and $R\;\frac{1}{4},0$ lie on BC, CA and AB, respectively.
   (a) Determine the ratios in which P, Q and R divide the sides of the triangle.
   (b) Determine whether or not the points P, Q and R are collinear.
7. ABC is a triangle, and X a point which does not lie on any of its (extended) sides. Also, AX meets BC at P, BX meets CA at Q and CX meets BA at R. Prove that
$$\frac{AX}{XP} = \frac{AR}{RB} + \frac{AQ}{QC}.$$
   (This result is often known as van Aubel s Theorem.)

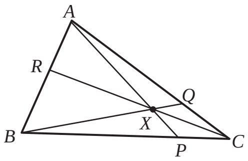

8. ABC is a triangle, and X a point which does not lie on any of its (extended) sides. Next, AX meets BC at P, BX meets CA at Q and CX meets BA at R. Also, RQ meets BC at L, PR meets CA at M and PQ meets BA at N. Prove that L, M and N are collinear.
   Hint: Apply the result of Problem 5 in Subsection 2.4.3 to ABC and points L, M and N in turn. Then evaluate the product $\frac{BL}{LC} \bullet \frac{CM}{MA} \bullet \frac{AN}{NB}$.

Direct calculation using coordinates is also possible, but very tedious!

9. Three disjoint circles of unequal radii lie in the plane, their centres being non-collinear. Pairs of tangents are drawn to each pair of circles such that the point of intersection of the two tangents to each pair of circles lies beyond the two circles. Prove that the three intersection points are collinear.

Section 2.5

1. An ellipse touches the sides AB, BC, CD, DA of a parallelogram $ABCD$ at the points P, Q, R, S, respectively. Prove that the lengths CQ, QB, BP and CR satisfy the equation
$$\frac{CQ}{QB} = \frac{CR}{BP}.$$

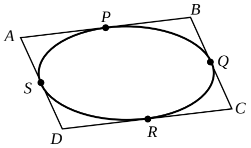

2. Determine the equation of the image of the parabola P with equation $y = x^2$ under the affine transformation $t : \mathbb{R}^2 \quad \mathbb{R}^2$ given by
$$t(\mathbf{x}) = \begin{matrix} 1 & 0 \\ 2 & 1 \end{matrix} \mathbf{x}.$$
   Show that the image of the vertex of P is not the vertex of $t(P)$.

This proves that the property of being a vertex of a parabola is not an affine property.

3. Prove that for any triangle ABC there exists an ellipse that touches the sides AB, BC and CA at their midpoints.
4. Let $P(a\cos\;, b\sin\;)$, where is not a multiple of $/2$, be a point on the ellipse $C : \frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$, where $a \quad b > 0$; and $P\,(a\cos\;, a\sin\;)$ the corresponding point on the auxiliary circle $C : x^2 + y^2 = a^2$. Prove that the tangents at P to C and at P to C meet on the x-axis.
   Hint: Write down an affine transformation that maps C to C and P to P, and that maps each point of the x-axis to itself.
5. Given any two points P and P on ellipses E and E, respectively, show that there exists an affine transformation that maps E to E and P to P.

This result is analogous to that for hyperbolas in the Corollary in Subsection 2.5.2.

6. Find the endpoints of the chord AB of the hyperbola H with equation $xy = 1$ that is bisected by the point $P(2,1)$.

7. E is the ellipse with equation $\frac{x^2}{9} + \frac{y^2}{4} = 1$, and $P \quad \frac{3}{5}, \frac{2}{5}$ is a point inside E. AB is a chord of E through P, and O is the centre of E. Find the maximum value of $\frac{AP}{PB}$ as A varies on E.

Summary of Chapter 2

Section 2.1: Geometry and transformations

1. An **isometry** of $\mathbb{R}^2$ is a function which maps $\mathbb{R}^2$ onto $\mathbb{R}^2$ and preserves distances.
   Every isometry of $\mathbb{R}^2$ has one of the following forms:
   a translation along a line in $\mathbb{R}^2$;
   a reflection in a line in $\mathbb{R}^2$;
   a rotation about a point in $\mathbb{R}^2$;
   a composite of translations, reflections and rotations in $\mathbb{R}^2$.
   The set of all isometries of $\mathbb{R}^2$ forms a group under composition of functions; in particular, the composite of two isometries is an isometry.
2. **Euclidean geometry** is the study of those properties of figures that are unchanged by the group of isometries.
   These properties are called **Euclidean properties**, and include distance, angle, collinearity of points and concurrence of lines.
3. The **Kleinian view of geometry** is the idea that geometry can be thought of in terms of a group of transformations acting on a space.
4. The transformations of $\mathbb{R}^2$ given by

$$\begin{array}{c} x \\ y \end{array} \quad \begin{array}{cc} \cos & \sin \\ \sin & \cos \end{array} \quad \begin{array}{c} x \\ y \end{array} + \begin{array}{c} e \\ f \end{array}, \text{ and}$$

$$\begin{array}{c} x \\ y \end{array} \quad \begin{array}{cc} \cos & \sin \\ \sin & \cos \end{array} \quad \begin{array}{c} x \\ y \end{array} + \begin{array}{c} e \\ f \end{array}$$

   are isometries; they represent, respectively, anticlockwise rotation about the origin through an angle followed by a translation by (e, f), and reflection in a line through the origin that makes an angle $/2$ with the x-axis followed by a translation by (e, f).
   Every isometry of $\mathbb{R}^2$ is of one or other of these two forms
5. A **Euclidean transformation** of $\mathbb{R}^2$ is a function $t : \mathbb{R}^2 \quad \mathbb{R}^2$ of the form $t(\mathbf{x}) = \mathbf{Ux} + \mathbf{a}$, where $\mathbf{U}$ is an orthogonal 2×2 matrix and $\mathbf{a} \quad \mathbb{R}^2$. The set of all Euclidean transformations of $\mathbb{R}^2$ is denoted by $E(2)$.
6. Every isometry of $\mathbb{R}^2$ is a Euclidean transformation of $\mathbb{R}^2$, and vice versa.
7. The set of Euclidean transformations of $\mathbb{R}^2$ forms a group under the operation of composition of functions.
8. The inverse of the Euclidean transformation $t(\mathbf{x}) = \mathbf{Ux} + \mathbf{a}$ is given by $t^{\ 1}(\mathbf{x}) = \mathbf{U}^{\ 1}\mathbf{x} \quad \mathbf{U}^{\ 1}\mathbf{a}$.

9. **Euclidean geometry** is the study of those properties of figures that are preserved by Euclidean transformations of $\mathbb{R}^2$.
10. A figure F_1 is **Euclidean-congruent** to a figure F_2 if there is a Euclidean transformation which maps F_1 onto F_2. Loosely speaking, two figures are congruent if they have the same size and shape.

    Euclidean congruence is an equivalence relation.

    A figure F_1 is ***G*-congruent** to a figure F_2 in some geometry defined by a group G of transformations acting on the space of the geometry if there is a transformation in G which maps F_1 onto F_2. G-congruence is an equivalence relation.

Section 2.2: Affine Transformations and Parallel Projections

1. An **affine transformation** of $\mathbb{R}^2$ is a function $t : \mathbb{R}^2 \to \mathbb{R}^2$ of the form $t(\mathbf{x}) = \mathbf{Ax} + \mathbf{b}$, where $\mathbf{A}$ is an invertible 2×2 matrix and $\mathbf{b} \in \mathbb{R}^2$. The set of all affine transformations of $\mathbb{R}^2$ is denoted by $A(2)$.

    Every Euclidean transformation of $\mathbb{R}^2$ is an affine transformation.
2. The set of affine transformations $A(2)$ forms a group under the operation of composition of functions.

    The inverse of the affine transformation $t(\mathbf{x}) = \mathbf{Ax} + \mathbf{b}$ is $t^{-1}(\mathbf{x}) = \mathbf{A}^{-1}\mathbf{x} - \mathbf{A}^{-1}\mathbf{b}$.
3. **Affine geometry** is the study of those properties (called **affine properties**) of figures in the plane $\mathbb{R}^2$ that are preserved by affine transformations.

    Basic properties of affine transformations

    Affine transformations:

    1. map straight lines to straight lines;
    2. map parallel straight lines to parallel straight lines;
    3. preserve ratios of lengths along a given straight line.
4. A **parallel projection** is a one-one mapping of $\mathbb{R}^2$ to itself defined in the following way. Let π_1 and π_2 be planes in $\mathbb{R}^3$, with parallel rays of light shining through them; then the function p which maps each point P in π_1 to the corresponding point P' in π_2 is a parallel projection from π_1 onto π_2.

    If π_1 and π_2 are parallel, then the parallel projection from π_1 onto π_2 is an isometry.
5. **Basic properties of parallel projections**

    Parallel projections:

    1. map straight lines to straight lines;
    2. map parallel straight lines to parallel straight lines;
    3. preserve ratios of lengths along a given straight line.
6. A **diameter** of an ellipse is a chord of the ellipse that passes through its centre.

    Midpoint Theorem Let ℓ be a chord of an ellipse. Then the midpoints of the chords parallel to ℓ lie on a diameter of the ellipse.
7. **Conjugate Diameters Theorem** Let ℓ be a diameter of an ellipse. Then there is another diameter m of the ellipse such that

(a) the midpoints of all chords parallel to ℓ lie on m;
(b) the midpoints of all chords parallel to m lie on ℓ.

The directions of these two diameters are called **conjugate directions**, and the diameters are called **conjugate diameters**.

8. Given any ellipse, there is a parallel projection which maps the ellipse onto a circle.
9. Under parallel projection certain properties of figures, such as length and angle, are not necessarily preserved. This is one difference between Euclidean geometry and affine geometry.
10. Each parallel projection is an affine transformation.

    However, an affine transformation is not necessarily a parallel projection. For example, the **doubling map** of $\mathbb{R}^2$ to itself given by $t(\mathbf{v}) = 2\mathbf{v}$ is an affine transformation but cannot be modelled by a parallel projection.
11. An affine transformation t is **completely determined** by its effect on the three non-collinear points (0,0), (1,0) and (0,1).

    Hence if we know the points onto which (0,0), (1,0) and (0,1) are mapped by t, we can determine $\mathbf{A}$ and $\mathbf{b}$ in the formula $t(\mathbf{x}) = \mathbf{Ax} + \mathbf{b}$, $\mathbf{x} \in \mathbb{R}^2$.
12. An affine transformation can be expressed as the composite of two parallel projections.

Section 2.3: Properties of Affine Transformations

1. **Strategy** To determine the image of a line or conic in $\mathbb{R}^2$ under an affine transformation $t : \mathbb{R}^2 \to \mathbb{R}^2$ given by $t(\mathbf{x}) = \mathbf{Ax} + \mathbf{b}$, let $\mathbf{x}$ and coordinates (x, y) denote points in the domain copy of $\mathbb{R}^2$, and $\mathbf{x}'$ and coordinates (x', y') denote points in the codomain copy. Then:
    1. express the relationship between $\mathbf{x}$ and $\mathbf{x}'$ in the form $\mathbf{x} = \mathbf{A}^{-1}\mathbf{x}' - \mathbf{A}^{-1}\mathbf{b}$;
    2. determine formulas for x and y in terms of x' and y';
    3. substitute for x and y in the equation of the line or conic;
    4. drop the dashes from x' and y'.

    The resulting equation describes the image under t.
2. **Strategy** To determine the unique affine transformation $t(\mathbf{x}) = \mathbf{Ax} + \mathbf{b}$ which maps (0,0), (1,0) and (0,1) to the three non-collinear points $\mathbf{p}$, $\mathbf{q}$ and $\mathbf{r}$, respectively:
    1. take $\mathbf{b} = \mathbf{p}$;
    2. take $\mathbf{A}$ to be the matrix with columns given by $\mathbf{q} - \mathbf{p}$ and $\mathbf{r} - \mathbf{p}$.

    Warning This Strategy requires that $\mathbf{p}$, $\mathbf{q}$ and $\mathbf{r}$ are non-collinear. If they are collinear, the matrix $\mathbf{A}$ described in the Strategy is non-invertible, and hence the transformation given by the procedure in the Strategy is not an affine transformation.
3. **Fundamental Theorem of Affine Geometry** Let $\mathbf{p}, \mathbf{q}, \mathbf{r}$ and $\mathbf{p}', \mathbf{q}', \mathbf{r}'$ be two sets of three non-collinear points in $\mathbb{R}^2$. Then:
    (a) there is an affine transformation t which maps $\mathbf{p}, \mathbf{q}$ and $\mathbf{r}$ to $\mathbf{p}', \mathbf{q}'$ and $\mathbf{r}'$, respectively;
    (b) the affine transformation t is unique.

Strategy To determine the affine transformation t which maps three non-collinear points **p, q** and **r** to another three non-collinear points **p** , **q** and **r** , respectively:

1. determine the affine transformation t_1 which maps (0,0), (1,0) and (0,1) to the points **p, q** and **r**, respectively;
2. determine the affine transformation t_2 which maps (0,0), (1,0) and (0,1) to the points **p** , **q** and **r** , respectively;
3. calculate the composite $t = t_2 \quad t_1^{\ 1}$.

4. Two figures are **affine-congruent** if there is an affine transformation which maps one onto the other.

   All triangles are affine-congruent.

5. An affine transformation preserves ratios of lengths along parallel straight lines.

Section 2.4: Using the Fundamental Theorem of Affine Geometry

1. **Median Theorem** The medians of any triangle are concurrent.
2. If a point R divides a line segment PQ in the **ratio** $(1 \quad) : \quad$, then $\frac{PR}{RQ} = \frac{1}{\quad}$. The **magnitude** of the ratio equals the length of PR divided by the length of RQ, and the ratio is **positive** if $0 < \quad < 1$ (when PR and RQ lie in the same direction) and **negative** if $\quad < 0$ or $\quad > 1$ (when PR and RQ lie in opposite directions).

   If P, Q and R have coordinates (x_P, y_P), (x_Q, y_Q) and (x_R, y_R), respectively, then $\frac{PR}{RQ} = \frac{x_R \quad x_P}{x_Q \quad x_R}$ and $\frac{PR}{RQ} = \frac{y_R \quad y_P}{y_Q \quad y_R}$. (If the denominator of one of these fractions vanishes, then use the other to determine the ratio.)
3. **Ceva's Theorem** Let ABC be a triangle, and let X be a point which does not lie on any of its (extended) sides. If AX meets BC at P, BX meets CA at Q and CX meets BA at R, then

$$\frac{AR}{RB} \cdot \frac{BP}{PC} \cdot \frac{CQ}{QA} = 1.$$

4. **Converse to Ceva's Theorem** Let P, Q and R be points other than vertices on the (possibly extended) sides BC, CA and AB of a triangle ABC, such that $\frac{AR}{RB} \cdot \frac{BP}{PC} \cdot \frac{CQ}{QA} = 1$. Then the lines AP, BQ and CR are concurrent.
5. **Menelaus's Theorem** Let ABC be a triangle, and let be a line that crosses the sides BC, CA and AB at three distinct points P, Q and R, respectively. Then $\frac{AR}{RB} \cdot \frac{BP}{PC} \cdot \frac{CQ}{QA} = \quad 1$.
6. **Converse to Menelaus' Theorem** Let P, Q and R be points other than vertices on the (possibly extended) sides BC, CA and AB of a triangle ABC, such that $\frac{AR}{RB} \cdot \frac{BP}{PC} \cdot \frac{CQ}{QA} = \quad 1$. Then the points P, Q and R are collinear.

7. Many results about properties of triangles (such as collinearity, lines being parallel, and ratios of lengths along a given line) are preserved under affine transformations, are proved following a standard pattern.

   First, we choose a particular type of triangle for which it is easy to prove the result. Then, by asserting the existence of an affine transformation from that triangle to an arbitrary triangle, we may deduce that the result holds for all triangles.
8. Let $A = (a_1, a_2)$, $B = (b_1, b_2)$ and $C = (c_1, c_2)$ be three non-collinear points in the plane $\mathbb{R}^2$; we call ABC the **triangle of reference**. Then a point (x, y) in the plane has **barycentric coordinates** $(\; , \; , \;)$ with respect to ABC if

$$x = \; a_1 + \; b_1 + \; c_1,$$
$$y = \; a_2 + \; b_2 + \; c_2, \text{ and}$$
$$1 = \; + \; + \; .$$

   If we set $\mathbf{M} = \begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ 1 & 1 & 1 \end{pmatrix}$, then $\mathbf{M}$ is invertible. Also $\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \mathbf{M}$ and $= \mathbf{M}^{1} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}$.
9. The points P, Q and R with barycentric coodinates $({}_1, {}_1, {}_1)$, $({}_2, {}_2, {}_2)$ and $({}_3, {}_3, {}_3)$ are collinear if and only if $\begin{vmatrix} {}_1 & {}_2 & {}_3 \\ {}_1 & {}_2 & {}_3 \\ {}_1 & {}_2 & {}_3 \end{vmatrix} = 0$.

   The line through the points with barycentric coordinates $({}_1, {}_1, {}_1)$ and $({}_2, {}_2, {}_2)$ has equation $\begin{vmatrix} {}_1 & {}_2 & \\ {}_1 & {}_2 & \\ {}_1 & {}_2 & \end{vmatrix} = 0$.
10. **Section Formula** The point R that divides the line joining the points P and Q with barycentric coordinates $({}_1, {}_1, {}_1)$ and $({}_2, {}_2, {}_2)$ in the ratio $(1 \quad) : \;$ has barycentric coordinates

$$(\; , \; , \;) = \; ({}_1, {}_1, {}_1) + (1 \quad)({}_2, {}_2, {}_2).$$

Section 2.5: Affine Transformations and Conics

1. Every ellipse is affine-congruent to the unit circle with equation $x^2 + y^2 = 1$.

   Every hyperbola is affine-congruent to the rectangular hyperbola with equation $xy = 1$.

   Every parabola is affine-congruent to the parabola with equation $y^2 = x$.
2. Affine transformations map ellipses to ellipses, parabolas to parabolas, and hyperbolas to hyperbolas.

   In Affine Geometry:

   (a) all ellipses are congruent to each other;

(b) all hyperbolas are congruent to each other;
(c) all parabolas are congruent to each other.

Non-degenerate conics are congruent only to non-degenerate conics of the same type.

3. Let t be an affine transformation, and let C be an ellipse or hyperbola with centre R. Then $t(C)$ has centre $t(R)$.
4. Let t be an affine transformation, and let H be a hyperbola with asymptotes $_1$ and $_2$. Then $t(H)$ has asymptotes $t(\ _1)$ and $t(\ _2)$.
5. Let t be an affine transformation, and let be a tangent to a conic C. Then $t(\ )$ is a tangent to the conic $t(C)$.
6. Affine geometry can be used to tackle problems involving conics when the property being investigated is an affine property. We first use an affine transformation to map the original conic onto one of our standard conics, tackle the problem in hand there, and then map back to the original conic.
7. Given any hyperbola and a point P on it, there is an affine transformation which maps the hyperbola onto the rectangular hyperbola $xy = 1$, and the point P to (1,1).

3 Projective Geometry: Lines

Geometry is one branch of mathematics that has an obvious relevance to the 'real world'. Earlier, we studied some results in Euclidean geometry and we described the group of Euclidean transformations, the isometries. We saw that the Euclidean transformations preserve distances and angles, and have a definite physical significance.

Chapters 1 and 2.

In this chapter we study *projective geometry*, a very different type of geometry, that has important but less obvious applications. It was discovered through artists' attempts over many centuries to paint realistic-looking pictures of scenes composed of objects situated at differing distances from the eye. How can three-dimensional scenes be represented on a two-dimensional canvas? Projective geometry explains how an eye perceives 'the real world', and so explains how artists can achieve realism in their work.

For example, in Computer Graphics and in Art.

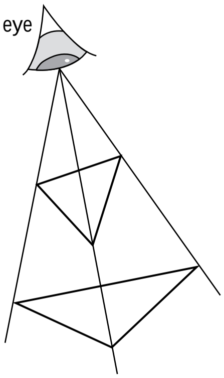

In Section 3.1, we look at the development of perspective in Art and explain the concept of a *perspectivity*. We describe Desargues' Theorem, which concerns a curious property of two triangles whose vertices are in perspective from a single point, and so explain that perspective can play a key role in the statement and the proof of theorems in mathematics.

In Section 3.2, we define the term *projective point* (or *Point*) and call the set of all such Points the *projective plane*, which we denote by RP^2. We also define a *projective line* (or *Line*). To enable us to tackle problems in projective geometry algebraically, we introduce *homogeneous coordinates* to specify the Points in RP^2.

In Section 3.3, we define the projective transformations of RP^2 and use them to define projective geometry. We also prove the *Fundamental Theorem of Projective Geometry*, which states that given two sets of four Points there is a unique projective transformation of RP^2 that maps the Points in one set to the corresponding Points in the other set. This crucial result enables us to apply a preliminary transformation to many geometric problems, thereby simplifying their solution by reducing the arithmetic involved. It turns out that there is a close connection between the idea of perspective in R^3 and projective transformations.

We also require that no three Points in either set lie on a Line.

In Section 3.4, we use the Fundamental Theorem of Projective Geometry to prove several results, including Desargues' Theorem. We also introduce the concept of *duality*, which involves a remarkable relationship between Points and Lines.

Finally, in Section 3.5, we note that the ideas of distance and ratio of distances along a line have no immediate analogues in RP^2; nevertheless, we are able to define a related quantity called the *cross-ratio* of four collinear Points in RP^2. This quantity is very useful in proving various mathematical results, and it has real life applications such as in aerial photography.

3.1 Perspective

3.1.1 Perspective in Art

The first pictures were probably Cave Art wall paintings: for example, depictions of animals and hunters. Up to the Middle Ages, most pictures were drawn on walls, floors or ceilings of buildings and were intended to convey messages rather than to be accurate illustrations of what an eye might see. For example, Christian religious art portrayed Christ and the Saints, the Bayeux tapestry outlined events such as the Norman Conquest and the Battle of Hastings, and so on.

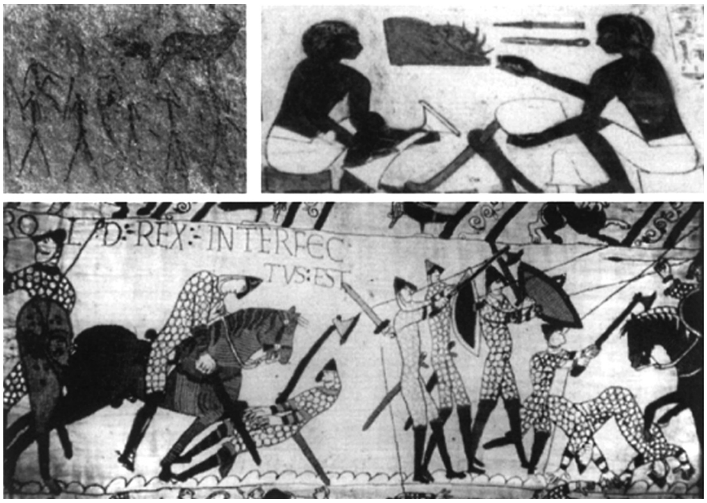

(Clockwise from left) Hunters below antelopes. Bambata cave, Zimbabwe c M. Jelliffe; Tomb of Rekhmare, Thebes. 1500 BC c Ronald Sheridan; Bayeux Tapestry: The death of Harold.
These prints are reproduced by kind permission of A.A. & A. Ancient Art and Architecture Collection.

To the modern eye, the people and animals in these pictures appear to be rather stylized, and the whole scene seems very two-dimensional. The events illustrated do not appear to be properly integrated into the background, even if this is included.

Towards the end of the 13th century, early Renaissance artists began to attempt to portray real situations in a realistic way. For example, people at the back of a group would be drawn higher up than those at the front a technique known as *terraced perspective.*

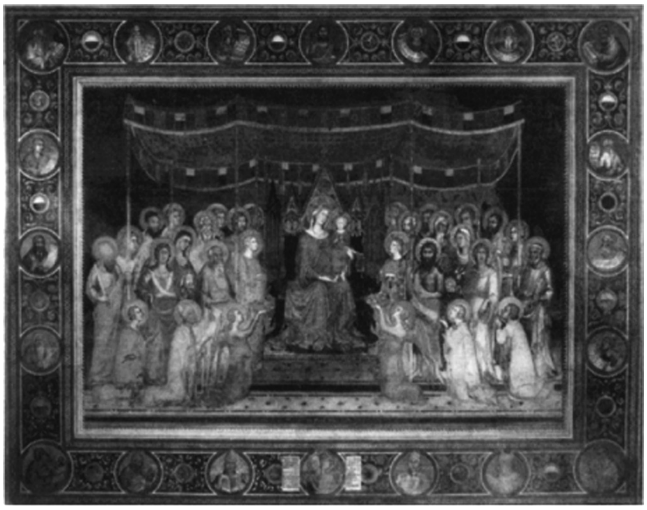

Simone Martini Maestà Palazzo Pubblico, Sala del Mappamodo, Siena (su concessione del commune di Siena). Foto LENSINI Siena.

As artists struggled to find better techniques to improve the realism of their work, the idea of *vertical perspective* was developed by the Italian school of artists (including Duccio (1255–1318) and Giotto (1266–1337)). To create an impression of depth in a scene, the artist would represent pairs of parallel lines that are symmetrically placed either side of the scene by lines that meet on the centre line of the picture. The method is not totally realistic, since objects do not appear to recede into the distance in the way that might be expected. The problem of depicting 'distant objects looking smaller', with a properly integrated foreground and background, was tackled by many artists, including notably Ambrogio Lorenzetti (c. 1290–1348).

Giotto is sometimes called the 'Father of Modern Painting'.

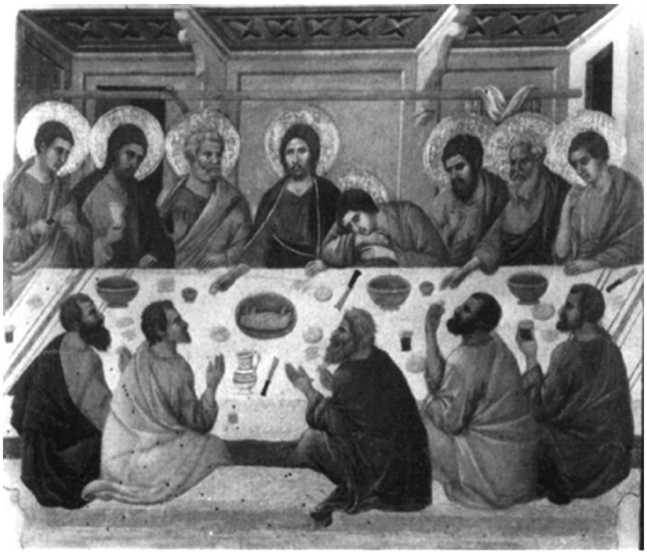

'Last Supper' painted by Duccio; Opera del Duomo, Siena. Foto LENSINI Siena.

The modern system of *focused perspective* was discovered around 1425 by the sculptor and architect Brunelleschi (1377–1446), developed by the painter and architect Leone Battista Alberti (1404 1472), and finally perfected by Leonardo da Vinci (1452 1519).

Alberti wrote that the first necessity for a painter is to know geometry .

These artists realized that what the eye actually sees of a scene are the various rays of light travelling from each point in the scene to the eye. An effective way of deciding how to depict a three-dimensional scene on a two-dimensional canvas so as to create a realistic impression is therefore as follows. Imagine a glass screen placed between the eye and the three-dimensional scene. Each line joining the eye to a point of the scene pierces the glass screen at some point. The set of all such points forms an image on the screen known as a *cross-section*. Since the eye cannot distinguish between light rays coming from the points of the actual scene and light rays coming from the corresponding points of the cross-section (since these are in exactly the same direction), the cross-section produces the same impression as the original scene. In other words, the cross-section gives a realistic two-dimensional representation of the three-dimensional scene.

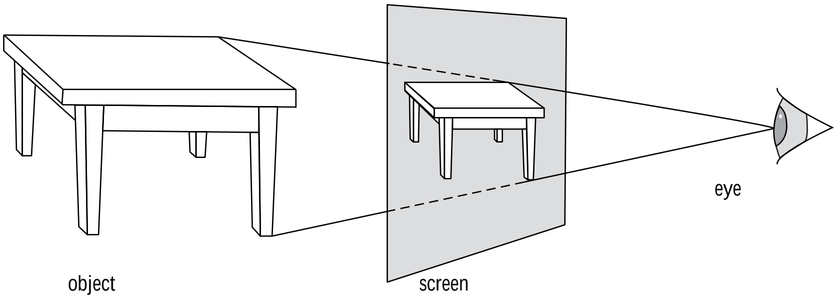

The German artist Albrecht Dürer (1471 1528) introduced the term *perspective* (from the Latin verb meaning to see through) to describe this technique, and illustrated it by a series of well-known woodcuts in his book *Underweysung der Messung mit dem Zyrkel und Rychtsscheyed* (1525). The Dürer woodcut below shows an artist peering through a grid on a glass screen to study perspective and the effect of foreshortening.

In English: *Instruction on measuring with compass and straight edge.*
We discuss foreshortening in Subsection 3.1.2.

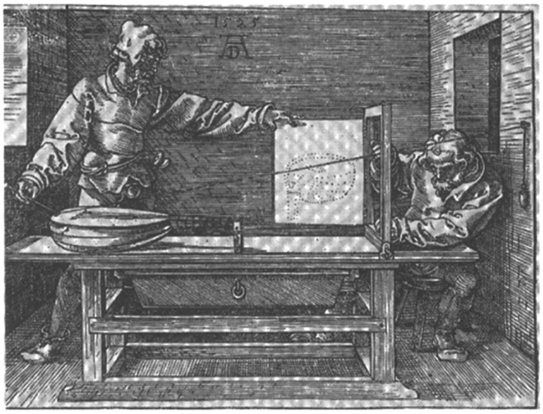

By permission of The British Library. c The British Library Board C.119.h.7(l).

Of course, the picture displayed on the screen is just one representation of the scene. If the screen is placed closer to, or further away from, the eye, the size of the cross-section changes. Also, the screen may be placed at a different angle for a given position of the eye, or the eye itself may be moved to a different position. In each case, a different cross-section is obtained, though they are all related to each other.

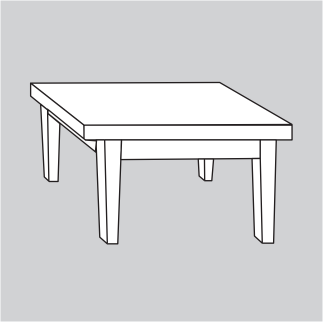

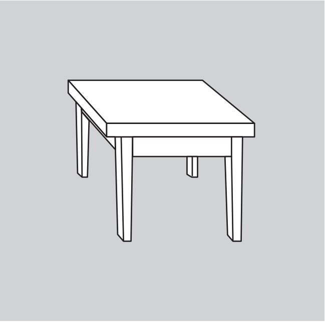

3.1.2 Mathematical Perspective

To help us understand the relationship between different representations of a scene, we now look at perspective from a mathematical point of view. In place of an eye and light rays travelling to it, we use the family of all lines in $\mathbb{R}^3$ through a given point. For convenience, this point will often be the origin O. The glass screen is replaced by a plane in $\mathbb{R}^3$ that does not pass through the origin.

In order to compare the cross-sections that appear on different screens, we consider two planes and that do not pass through O. A point P in and a point Q in are said to be *in perspective from O* if there is a straight line through O, P and Q. A *perspectivity from to* centred at O is a function that maps a point P of to a point Q of whenever P and Q are in perspective from O. Notice that the planes and may lie on the same side of O as shown on the left below, or they may lie on opposite sides of O as shown on the right.

In terms of O representing an eye, the figure on the right corresponds to the observer having the ability to look simultaneously both forwards and backwards!

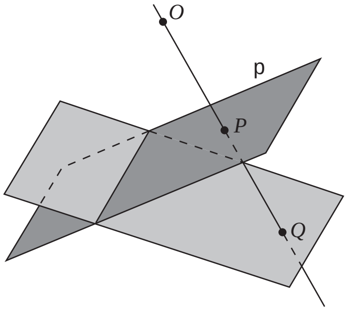

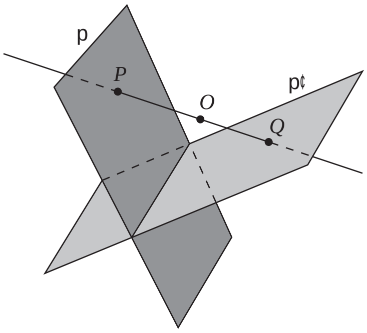

One complication with the above definition of a perspectivity is that the domain of the perspectivity is not necessarily the whole of . Indeed, if *P* is any point of such that *OP* is parallel to , as shown in the margin, then *P* cannot have an image in , and cannot therefore belong to the domain of the perspectivity. From a mathematical point of view, this need to exclude such exceptional points from the domain of a perspectivity turns out to be rather a nuisance. In Subsection 3.2.3 we shall therefore reformulate the definition of a perspectivity in such a way that these exceptional points can be included in the domain.

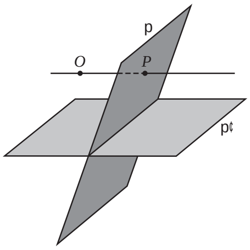

Even with only the preliminary definition of perspectivity given above, it is clear that some features of figures are preserved under a perspectivity, while others are not. For example, the figure on the left below illustrates a particular perspectivity in which a line segment in one plane maps onto a line segment in another plane. This suggests that collinearity is preserved by a perspectivity. On the other hand, the figure on the right illustrates a perspectivity in which a circle in one plane appears to map to a parabolic shape in another plane, which suggests that circularity is not preserved.

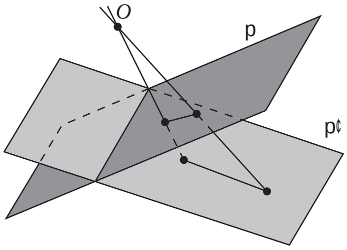

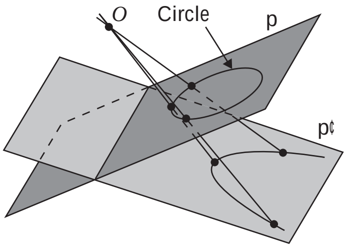

One of our main tasks is to study the images of standard configurations such as lines and conics under perspectivities. This chapter deals with lines; the next chapter deals with conics.

Consider a perspectivity with centre *O* that maps points in a plane to points in a plane . A convenient way to visualize the image of a line under the perspectivity is to consider an arbitrary point *P* on . As *P* moves along , the line *OP* sweeps out a plane. The line where this plane intersects is the image of .

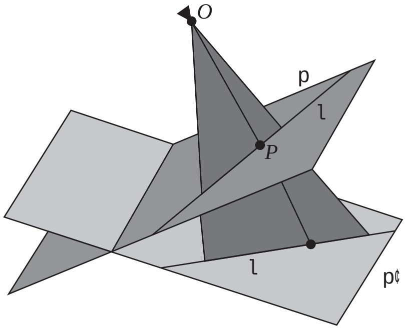

To be specific, consider the perspectivity p with centre O that maps points in a horizontal plane to points in a vertical plane , and let L be the line where and intersect. Under p, every line in that is parallel to L maps to a horizontal line in . In particular, L maps to itself. The only exception is the line h that passes through the foot of the perpendicular from O to . This line does not have an image in since the lines joining points of h to O are parallel to .

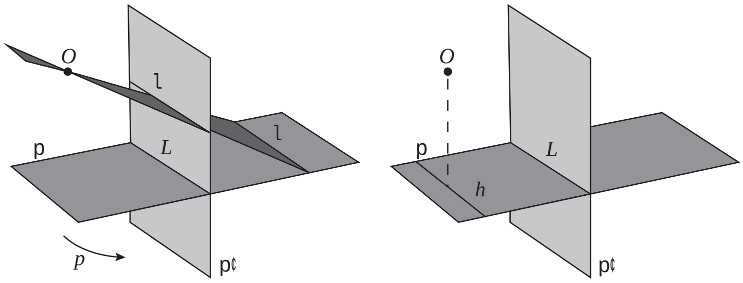

Next, consider the image under the same perspectivity p of a line in that is perpendicular to L. To do this, let P denote the foot of the perpendicular from O to the plane . Although P is not the image of any point of , the plane through O and meets in some line that passes through P. It follows that the image of under p is some line through P, with the point P itself omitted.

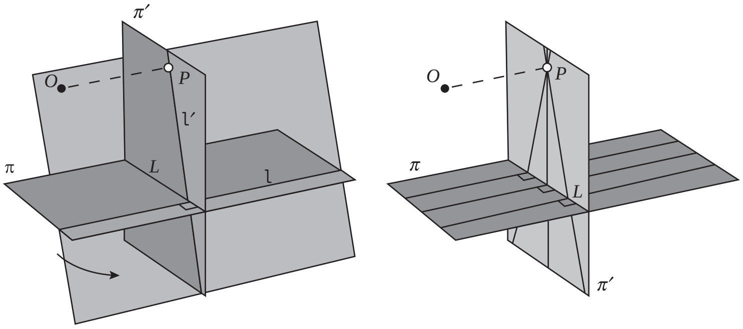

The above argument works for *any* line in that is perpendicular to L. All such lines are mapped by the perspectivity p to lines in that pass through P, and that omit the point P itself.

We may combine our observations concerning lines in that are parallel to L or perpendicular to L in the following way. Let $ABCD$ be a rectangle in on the opposite side of L from O, with sides AB and CD that lie on lines $_1$ and $_2$, perpendicular to L. Then AD and BC both map onto horizontal lines in between L and P. As the side BC recedes from L, its image $B\ C$ under the perspectivity p moves further up towards P, becoming shorter as it moves.

Artists describe this shortening of the image on of lines of equal length in as *foreshortening*.

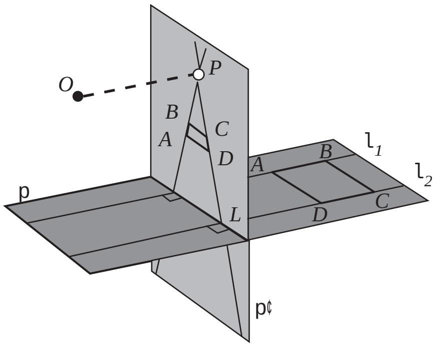

To an observer whose eye is located at O, the lines $_1$ and $_2$ *appear* to meet at infinity, and this corresponds to their images under p *appearing* to meet at P. The point P is called the *principal vanishing point* of the perspectivity p because the images in of all lines in perpendicular to L appear to vanish there.

You can think of $_1$ and $_2$ as a pair of railroad lines disappearing into the distance.

In fact, a perspectivity has many vanishing points. For instance, let be any line in that intersects L at an angle of /4. Now let h be the horizontal line in through P, and let D be the point on h such that OD is parallel to . Then the plane through O and meets in some line that passes through D. It follows that the image of under p is a line through D, with the point D itself omitted.

Here the symbol is being used in two different ways: as a label for the embedding plane, and as an angle.

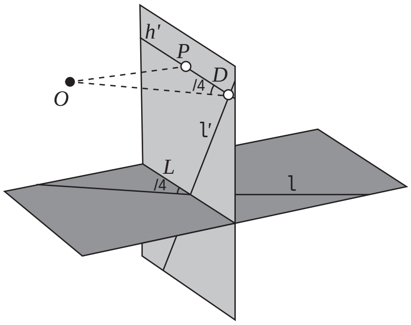

The point D is called a *diagonal vanishing point* of the perspectivity. All lines in the plane that are parallel to the given line have images in that are lines through D, with the point D itself omitted.

That is, the images appear to *vanish* at D.

In the same way, each point of the horizontal line h in through P is a *vanishing point* for the images of all lines in in some direction; hence the line h

is called the *vanishing line*. It corresponds to the horizon in the plane in other words, to the points at infinity towards which an observer s eye is pointing when looking in a horizontal direction.

3.1.3 Desargues Theorem

The idea that information in three dimensions can be related to information in two dimensions, and vice versa, plays an important role in mathematics just as it does in Art. For example, consider the following three-dimensional figure that consists of two triangles ABC and $A B C$ which are in perspective from a point U. For the moment we shall assume that no pair of corresponding sides BC and $B C$, CA and $C A$, and AB and $A B$, are parallel.

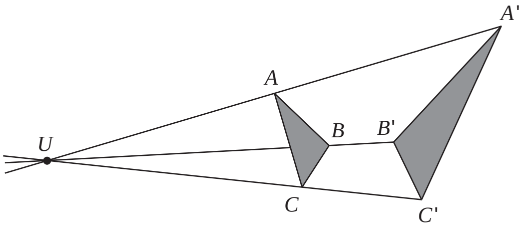

We shall show that this three-dimensional figure has the property that BC and $B C$, CA and $C A$, AB and $A B$ meet at P, Q, R, respectively, where P, Q and R are collinear. This will enable us to formulate an equivalent two-dimensional result, known as *Desargues Theorem.*

Girard Desargues (1593 1662) was a French engineer and architect.

To prove the three-dimensional result, observe that both BC and $B C$ lie in the plane that passes through the points U, B and C. Since BC and $B C$ are coplanar but not parallel, they must meet at some point P.

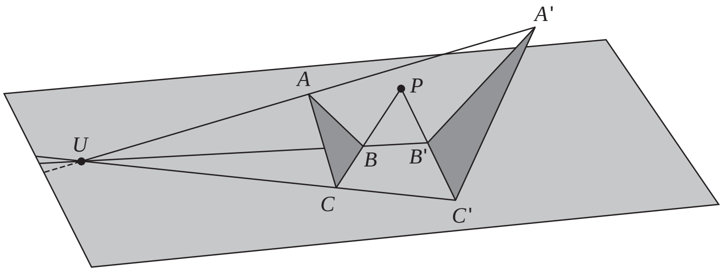

Similarly, the sides CA and $C A$ meet at some point Q, and the sides AB and $A B$ meet at some point R.

Since the points P, Q and R lie both on the plane which contains the triangle ABC and on the plane which contains the triangle $A B C$, they must lie on the line where the two planes meet. It follows that P, Q and R are collinear.

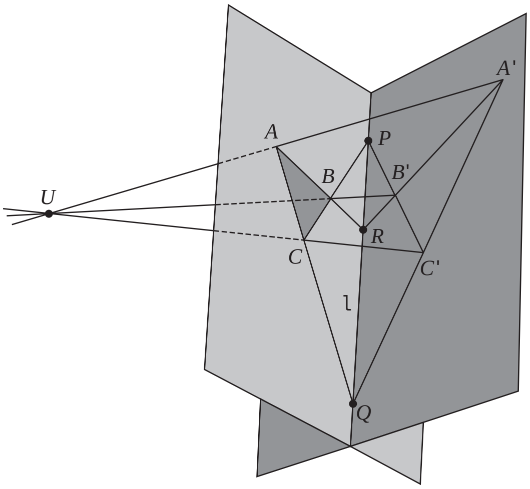

To obtain the equivalent two-dimensional result, imagine that you are viewing the three-dimensional configuration through a transparent screen. Since this viewing process will not alter the collinearity of points or the coincidence of lines, we may reinterpret the three-dimensional result in terms of the image on the screen to obtain the following theorem.

We give a rigorous proof of Desargues Theorem in Theorem 1, Subsection 3.4.1.

Theorem 1 Desargues' Theorem
Let ABC and $A B C$ be triangles in $\mathbb{R}^2$ such that the lines AA, BB and CC meet at a point U. Let BC and $B C$ meet at P, CA and $C A$ meet at Q, and AB and $A B$ meet at R. Then P, Q and R are collinear.

Strictly speaking, we have not proved this theorem since it is not immediately obvious that ABC and $A B C$ can be obtained as images of triangles in $\mathbb{R}^3$ which have corresponding sides that are not parallel. Nevertheless, the above argument does provide reasonably convincing evidence that the theorem is true.

One remarkable feature of the above argument is the way in which the geometry of the figure on the transparent screen is characterized by the rays of light that enter an eye. Thus a point on the screen corresponds to a single ray of light that enters the eye, a line on the screen corresponds to a plane of rays of light that enter the eye, and so on. The geometry of the figure can be investigated entirely in terms of these rays of light. The screen is needed only to interpret the result in terms of a two-dimensional figure.

In the rest of Chapters 3 and 4, we introduce a geometry known as *projective geometry* that enables us to work with figures on a plane (a screen) as if they correspond to rays of light that enter an eye in the way described above.

3.2 The Projective Plane $\mathbb{RP}^2$

You have already met the Kleinian view that a geometry consists of a group of transformations acting on a space of points. In this section we begin our discussion of projective geometry by investigating its space of points. The group of transformations is discussed in Section 3.3.

Introductary remarks to Chapter 2

3.2.1 Projective Points

Imagine an eye situated at the origin of $\mathbb{R}^3$ looking at a fixed screen. As we mentioned in Subsection 3.1.1, each point of the screen corresponds to the ray of light that enters the eye from the point. This correspondence between points of the screen and rays of light through the origin is the clue that we need to define a space of points for our new geometry.

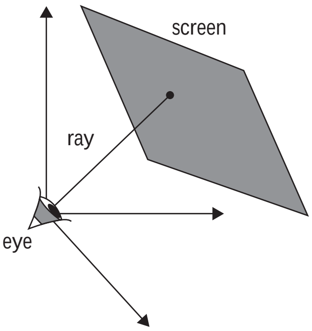

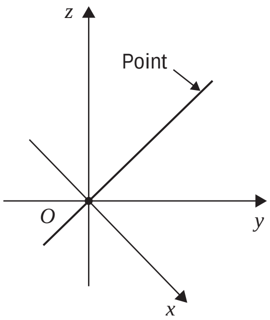

Rather than use the points of the screen directly, we use the rays of light that enable an eye to see the points from the origin. We can express this idea mathematically by defining a *projective point* to be a Euclidean line in $\mathbb{R}^3$ that passes through the origin. In order to avoid confusion with Euclidean points of $\mathbb{R}^3$, we write Point with a capital P whenever we mean a projective point.

It is important that you use the capital letter P in Point.

Definitions **A Point** (or **projective point)** is a line in $\mathbb{R}^3$ that passes through the origin of $\mathbb{R}^3$. The **real projective plane** $\mathbb{RP}^2$ is the set of all such Points.

In order to prove results in projective geometry algebraically, we need to have an algebraic notation that can be used to specify the Points of $\mathbb{RP}^2$. To do this, we use the fact that a line through the origin O in $\mathbb{R}^3$ is uniquely determined once we have specified a Euclidean point (other than O) that lies on . For example, there is a unique line in $\mathbb{R}^3$ through O and the point with Euclidean coordinates (4, 2, 6), so we can use these coordinates to specify a projective point. When doing this we write the coordinates in the form [4, 2, 6], with square brackets to indicate that the coordinates refer to a projective point.

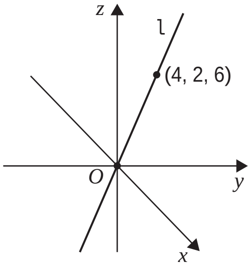

Definition The expression $[a, b, c]$, in which the numbers a, b, c are not all zero, represents the Point P in $\mathbb{RP}^2$ which consists of the unique line in $\mathbb{R}^3$ that passes through (0, 0, 0) and (a, b, c). We refer to $[a, b, c]$ as **homogeneous coordinates** of P. If (a, b, c) has position vector $\mathbf{v}$, then we often denote P by $[\mathbf{v}]$ and we say that P can be **represented** by $\mathbf{v}$.

Note that [0, 0, 0] is *not* defined.

Remark

Often we abuse our notation slightly, by talking about the Point $[a, b, c]$ when strictly speaking we should say the Point with homogeneous coordinates $[a, b, c]$.

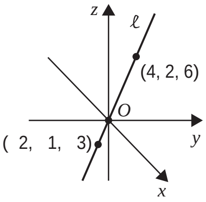

Notice that the homogeneous coordinates of a Point are not unique. For example, the Point with homogeneous coordinates [4, 2, 6] consists of a line that passes through (0, 0, 0) and (4, 2, 6). But this line also passes through (2, 1, 3), so [4, 2, 6] and [2, 1, 3] both represent the same Point.

In general, if (a, b, c) is any point on a line through the origin, and is any real number, then $(\ a,\ b,\ c)$ also lies on the line. Moreover, if (a, b, c) is not at the origin and $= 0$, then $(\ a,\ b,\ c)$ is not at the origin either. It follows that $[a, b, c]$ and $[\ a,\ b,\ c]$ both represent the same Point, for any $= 0$. We express this by writing

$$[a,b,c] = [\ a,\ b,\ c], \qquad \text{for any} \ = 0. \tag{1}$$

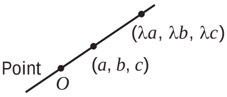

Conversely, if there is no non-zero real number such that

$$(a\ ,b\ ,c\ ) = (\ a,\ b,\ c),$$

then (a, b, c) and $(a\ ,b\ ,c\ )$ cannot lie on the same line through the origin, and so the homogeneous coordinates $[a, b, c]$ and $[a\ ,b\ ,c\ ]$ must represent different Points in $\mathbb{RP}^2$.

Example 1 Which of the following homogeneous coordinates represent the same Point in $\mathbb{RP}^2$ as [6, 3, 2]?

(a) [18, 9, 6] (b) [12, 6, 4] (c) $1, \frac{1}{2}, \frac{1}{3}$ (d) [1, 2, 3]

Solution

Throughout the solution we use equation (1):

$$[a, b, c] = [\ a,\ b,\ c],$$

for any $\ = 0$.

(a) This represents the same Point as [6, 3, 2], for if $\ = 3$, then

$$[18, 9, 6] = [6\ , 3\ , 2\ ] = [6, 3, 2].$$

(b) This represents a Point different from [6, 3, 2], for there is no that satisfies the simultaneous equations

$$12 = 6\ ,\quad 6 = 3\ , 4 = 2\ .$$

(c) This represents the same Point as [6, 3, 2], for if $\ = \frac{1}{6}$, then

$$1, \tfrac{1}{2}, \tfrac{1}{3}\quad = [6\ , 3\ , 2\ ] = [6, 3, 2].$$

(d) This represents a Point different from [6, 3, 2], for there is no that satisfies the simultaneous equations

$$1 = 6\ ,\quad 2 = 3\ ,\quad 3 = 2\ .$$ ☐

Problem 1 Which of the following homogeneous coordinates represent the same Point in $\mathbb{RP}^2$ as [1, 2, 3]?

(a) [2, 4, 6] (b) $[1, 2,\ \ 3]$ (c) $[\ \ 1,\ \ 2,\ \ 3]$ (d) [11, 12, 13]

At first sight it may seem rather unsatisfactory that the coordinates of a Point are not unique. However, this ambiguity can often be turned to our advantage. For example, if a calculation yields a Point of $\mathbb{RP}^2$ with fractional homogeneous coordinates such as $1, \frac{1}{2}, \frac{1}{3}$, then the rest of the calculation may be simpler if we clear the fractions and represent the Point by the integer homogeneous coordinates [6, 3, 2] instead.

Problem 2 For each of the following homogeneous coordinates, find integer homogeneous coordinates which represent the same Point.

(a) $\frac{3}{4}, \frac{1}{2},\ \ \frac{1}{8}$ (b) $0, 4, \frac{2}{3}$ (c) $\frac{1}{6},\ \ \frac{1}{3},\ \ \frac{1}{2}$

Given a collection of homogeneous coordinates, it is not always easy to spot those that represent the same Point. In such cases it is sometimes possible to rewrite the coordinates in a form that makes the comparison easier.

Example 2 Determine homogeneous coordinates of the form [*a*, *b*, 1] for the Points

$$[2,\ \ 1, 4],\qquad [4, 2, 8],\qquad [2\ \ ,\qquad , 4\ \ ],$$

$$[200, 100, 400],\qquad \frac{1}{2},\ \ \frac{1}{4},\ \ 1\ \ ,\qquad [6,\ \ 9,\ \ 12].$$

Hence decide which homogeneous coordinates represent the same Points.

Solution According to equation (1), a Point of $\mathbb{RP}^2$ is unchanged if its homogeneous coordinates are multiplied (or divided) by any non-zero real number. Since the third coordinate of each Point is non-zero, we may divide by this third coordinate to obtain homogeneous coordinates of the form [*a*, *b*, 1] as follows:

For, dividing by a non-zero number is equivalent to multiplying by the non-zero number 1 .

$$[2,\quad 1,4] = \tfrac{1}{2},\quad \tfrac{1}{4},1\ ; \qquad [4,2,8] = \tfrac{1}{2},\tfrac{1}{4},1\ ;$$

$$[2\ \ ,\qquad ,4\ \ ] = \tfrac{1}{2},\quad \tfrac{1}{4},1\ ; \qquad [200,100,400] = \tfrac{1}{2},\tfrac{1}{4},1\ ;$$

$$\tfrac{1}{2},\quad \tfrac{1}{4},\quad 1\ = \tfrac{1}{2},\tfrac{1}{4},1\ ; \qquad [6,\quad 9,\quad 12] = \tfrac{1}{2},\tfrac{3}{4},1\ .$$

Since $[a,b,1] = [a\ ,b\ ,1]$ if and only if $a = a$ and $b = b$, it follows that:

$[2,\quad 1,4]$ and $[2\ \ ,\qquad ,4\ \ ]$ represent the same Point;

$[4,2,8]$, $[200,100,400]$ and $\tfrac{1}{2},\quad \tfrac{1}{4},\quad 1$ represent the same Point;

$[6,\quad 9,\quad 12]$ represents none of the other Points. □

Notice that the method used in Example 2 works only if the third coordinates of all the Points are non-zero. If this is not the case, then you may still be able to apply the technique using the first or second coordinates.

Problem 3 Determine homogeneous coordinates of the form [1, *b*, *c*] for the Points

$$[2,3,\quad 5],\qquad [\quad 8,\quad 12,20],\qquad \overline{2},\quad \overline{3},\qquad \overline{5}\ ,$$

$$[4,\quad 6,10],\qquad [\quad 20,\quad 30,50],\qquad [74,148,0].$$

Hence decide which homogeneous coordinates represent the same Points.

Having defined projective points, we are now in a position to define a *projective figure.* Just as a figure in Euclidean geometry is defined to be a subset of $\mathbb{R}^2$, so figures in projective geometry are defined to be subsets of $\mathbb{RP}^2$.

Definition **A projective figure** is a subset of $\mathbb{RP}^2$.

Projective figures are just sets of lines in $\mathbb{R}^3$ that pass through the origin. Thus a double cone with a vertex at *O*, and a double square pyramid with a vertex at *O*, are both examples of projective figures, for they can both be formed from sets of lines that pass through the origin of $\mathbb{R}^3$.

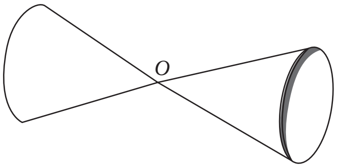

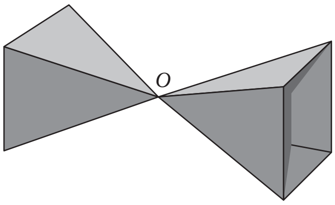

3.2.2 Projective Lines

A particularly simple type of projective figure is a plane through the origin. Such a plane is a projective figure because it can be formed from the set of all Points (lines through the origin of R^3) that lie on the plane. Since all but one of these Points can be thought of as rays of light that come from a line on a screen, it seems reasonable to define any plane through the origin to be a *projective line*.

The exception is the ray of light parallel to the screen. We shall discuss the significance of this ray later, in Subsection 3.2.3.

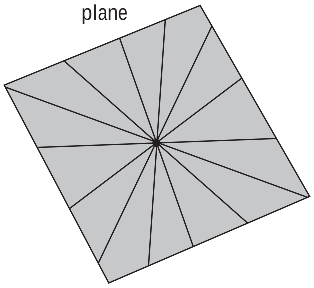

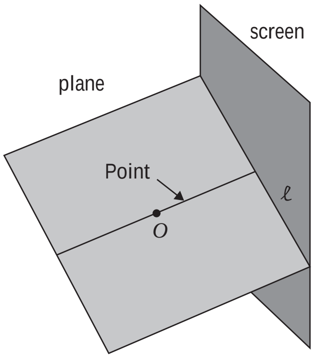

Just as we use ‘Point’ to refer to a ‘projective point’, so we use ‘Line’ to refer to a ‘projective line’. The use of a capital L avoids any confusion with lines in R^3.

Definitions **A Line** (or **projective line)** in RP^2 is a plane in R^3 that passes through the origin. Points in RP^2 are **collinear** if they lie on a Line.

Since a Line in RP^2 is simply a plane in R^3 that passes through the origin, it must consist of the set of Euclidean points (x, y, z) that satisfy an equation of the form

$$ax + by + cz = 0,$$

where a, b and c are real and not all zero. We can interpret this fact in terms of RP^2 as follows.

Theorem 1 The general equation of a Line in RP^2 is

$$ax + by + cz = 0, \tag{2}$$

where a, b, c are real and not all zero.

Remark

1. The equation of a Line is not unique, for, if $\lambda \neq 0$, then $\lambda ax + \lambda by + \lambda cz = 0$ is also an equation for the Line. We can use this fact to 'clear fractions' from the coefficients just as we did for the homogeneous coordinates of a Point.
2. From the figure in the margin it is clear that a Point lies on a Line, or a Line passes through a Point, if and only if the Point has homogeneous coordinates $[x, y, z]$ which satisfy the equation of the Line. For example, $[1, -1, 1]$ lies on the Line $3x + y - 2z = 0$, but $[0, 1, 3]$ does not.

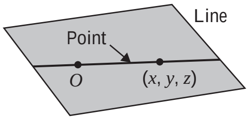

In Euclidean geometry there is a unique line that passes through any two distinct points, as illustrated on the left of the figure below. Similarly, in projective geometry two distinct Points (lines through the origin) lie on a unique Line (plane through the origin).

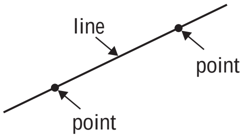

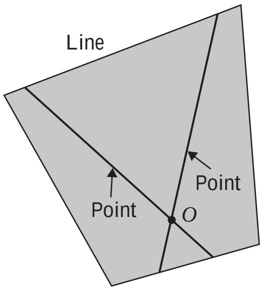

We express this observation in the form of a theorem, as follows.

Theorem 2 Collinearity Property of $\mathbb{RP}^2$
Any two distinct Points of $\mathbb{RP}^2$ lie on a unique Line.

It is sometimes possible to find an equation for the Line that passes through two distinct Points of $\mathbb{RP}^2$ simply by spotting an equation of the form (2) that is satisfied by the homogeneous coordinates of both Points.

Example 3 For each of the following pairs of Points, write down an equation for the Line that passes through them.

(a) $[3, 2, 0]$ and $[3, 4, 0]$ (b) $[1, 2, 1]$ and $[3, 0, 3]$
(c) $[1, 0, 0]$ and $[0, 0, 1]$

Solution

(a) Both the Points have a z-coordinate equal to 0, so the homogeneous coordinates must satisfy the equation $z = 0$. This equation is of the form (2) with $a = 0, b = 0$ and $c = 1$, so it must be the required equation for the Line.

The equation $x = 3$ is not of the form (2), and so is not the equation of a Line.

(b) The homogeneous coordinates of both Points satisfy $x = z$. This equation is of the form (2) with $a = 1$, $b = 0$ and $c = -1$. It must therefore be the required equation for the Line.

(c) The homogeneous coordinates of both Points satisfy $y = 0$. This equation is of the form (2) with $a = 0$, $b = 1$ and $c = 0$, so it must be the required equation for the Line. □

Problem 4 For each of the following pairs of Points, write down an equation for the Line that passes through them.

(a) [0, 1, 0] and [0, 0, 1] (b) [2, 2, 3] and [3, 3, 7]

But how do we find an equation for a Line through two given Points in cases where it cannot be found by inspection? As an example, consider the Points $[2, -1, 4]$ and $[1, -1, 1]$. We could certainly substitute the values $x = 2$, $y = -1$, $z = 4$ and $x = 1$, $y = -1$, $z = 1$ into equation (2), to obtain the pair of simultaneous equations

$$2a - b + 4c = 0,$$
$$a - b + c = 0.$$

Then subtracting twice the second equation from the first, we obtain $b = -2c$. So from the second equation it follows that $a = -3c$. If we set $c = -1$, say, then $a = 3$ and $b = 2$, so an equation for the Line is

$$3x + 2y - z = 0.$$

Of course, we could set c to have any non-zero value, but $c = -1$ keeps the calculation simple.

In this case the calculations are fairly straightforward, but there is an alternative method that is often simpler. Notice that the Line in $\mathbb{RP}^2$ through the Points $[2, -1, 4]$ and $[1, -1, 1]$ is the Euclidean plane in $\mathbb{R}^3$ that contains the position vectors of the points $(2, -1, 4)$ and $(1, -1, 1)$ in $\mathbb{R}^3$. A point (x, y, z) lies in this plane if and only if the vector (x, y, z) is a linear combination of the vectors $(2, -1, 4)$ and $(1, -1, 1)$; in other words, if and only if the vectors (x, y, z), $(2, -1, 4)$ and $(1, -1, 1)$ are linearly dependent.

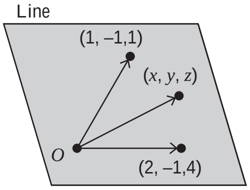

But three vectors in $\mathbb{R}^3$ are linearly dependent if and only if the 3×3 determinant that has these vectors as its rows is zero. It follows that (x, y, z) lies in the plane containing the position vectors $(2, -1, 4)$ and $(1, -1, 1)$ if and only if

$$\begin{vmatrix} x & y & z \\ 2 & -1 & 4 \\ 1 & -1 & 1 \end{vmatrix} = 0.$$

Translating this statement back into a statement concerning $\mathbb{RP}^2$, we deduce that the Point $[x, y, z]$ lies on the Line through the Points $[2, -1, 4]$ and $[1, -1, 1]$ if and only if

$$\begin{vmatrix} x & y & z \\ 2 & -1 & 4 \\ 1 & -1 & 1 \end{vmatrix} = 0.$$

Expanding this determinant in terms of the entries in its first row, we obtain

$$\begin{vmatrix} x & y & z \\ 2 & -1 & 4 \\ 1 & -1 & 1 \end{vmatrix} = x\begin{vmatrix} -1 & 4 \\ -1 & 1 \end{vmatrix} - y\begin{vmatrix} 2 & 4 \\ 1 & 1 \end{vmatrix} + z\begin{vmatrix} 2 & -1 \\ 1 & -1 \end{vmatrix}$$

$$= 3x + 2y - z.$$

Hence an equation for the required Line in $\mathbb{RP}^2$ is

$$3x + 2y - z = 0. \tag{3}$$

Remark

It is always sensible to check your arithmetic by checking that the two given Points actually lie on the Line that you have found. For instance, the answer above is correct, since equation (3) is a homogeneous linear equation in x, y and z, and the equation is satisfied by $x = 2$, $y = -1$, $z = 4$ and by $x = 1$, $y = -1$, $z = 1$.

We may summarize the above method in the form of a strategy, as follows.

Strategy To determine an equation for the Line in $\mathbb{RP}^2$ through the Points $[d, e, f]$ and $[g, h, k]$:

1. write down the equation

$$\begin{vmatrix} x & y & z \\ d & e & f \\ g & h & k \end{vmatrix} = 0;$$

2. expand the determinant in terms of the entries in its first row to obtain the required equation in the form $ax + by + cz = 0$.

Example 4 Find an equation for the Line that passes through the Points $[1, 2, 3]$ and $[2, -1, 4]$.

Solution An equation for the Line is

$$\begin{vmatrix} x & y & z \\ 1 & 2 & 3 \\ 2 & -1 & 4 \end{vmatrix} = 0.$$

Now

$$\begin{vmatrix} x & y & z \\ 1 & 2 & 3 \\ 2 & -1 & 4 \end{vmatrix} = x\begin{vmatrix} 2 & 3 \\ -1 & 4 \end{vmatrix} - y\begin{vmatrix} 1 & 3 \\ 2 & 4 \end{vmatrix} + z\begin{vmatrix} 1 & 2 \\ 2 & -1 \end{vmatrix}$$

$$= 11x + 2y - 5z.$$

An equation for the Line is therefore

$$11x + 2y \quad 5z = 0. \qquad \square$$

You can easily check that the Points [1, 2, 3] and [2, 1, 4] lie on this Line.

Problem 5 Determine an equation for each of the following Lines in RP^2:

(a) the Line through the Points [2, 5, 4] and [3, 1, 7];
(b) the Line through the Points [2, 4, 5] and [3, 2, 4].

A similar technique can be used to check whether three given Points are collinear. Indeed, three Points $[a, b, c]$, $[d, e, f]$, $[g, h, k]$ are collinear if and only if the position vectors of the points (a, b, c), (d, e, f), (g, h, k) are linearly dependent; that is, if and only if

$$\begin{matrix} a & b & c \\ d & e & f \\ g & h & k \end{matrix} = 0.$$

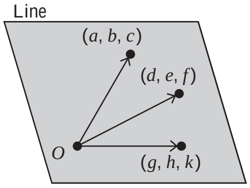

Example 5 Determine whether the Points [2, 1, 3], [1, 2, 1] and [1, 4, 3] are collinear.

Solution We have

$$\begin{aligned}\begin{matrix} 2 & 1 & 3 \\ 1 & 2 & 1 \\ 1 & 4 & 3 \end{matrix} &= 2\begin{matrix} 2 & 1 \\ 4 & 3 \end{matrix} \quad 1\begin{matrix} 1 & 1 \\ 1 & 3 \end{matrix} + 3\begin{matrix} 1 & 2 \\ 1 & 4 \end{matrix} \\ &= 2(\; 6 \quad 4) \quad (\; 3 + 1) + 3(4 + 2) \\ &= \quad 20 + 2 + 18 \\ &= 0.\end{aligned}$$

Since this is zero it follows that [2, 1, 3], [1, 2, 1] and [1, 4, 3] are collinear. □

We summarize the method of Example 5 in the following strategy.

Strategy To determine whether three Points $[a, b, c]$, $[d, e, f]$, $[g, h, k]$ are collinear:

1. evaluate the determinant $\begin{matrix} a & b & c \\ d & e & f \\ g & h & k \end{matrix}$;
2. the Points $[a, b, c]$, $[d, e, f]$, $[g, h, k]$ are collinear if and only if this determinant is zero.

Problem 6 Determine whether the following sets of Points are collinear.

(a) $[1,2,3], [1,1,-2], [2,1,-9]$ (b) $[1,2,-1], [2,1,0], [0,-1,3]$

Before rushing to solve a problem using determinants, you should always stop to see if you can solve the problem more easily by inspection. For example, suppose that you are asked to check whether the Points [1, 0, 0], [0, 1, 0], [1, 1, 1] are collinear. Clearly, [1, 0, 0] and [0, 1, 0] lie on the Line $z = 0$, whereas [1, 1, 1] does not, so the Points are not collinear.

Problem 7 Verify that no three of the Points [1, 0, 0], [0, 1, 0], [0, 0, 1] and [1, 1, 1] are collinear.

The Points that you considered in Problem 7 play an important part in our development of the theory of projective geometry, so we give them special names.

Definitions The Points [1, 0, 0], [0, 1, 0], [0, 0, 1] are known as the **triangle of reference.** The Point [1, 1, 1] is called the **unit Point.**

Next, observe that any two distinct Lines necessarily meet at a unique Point. Indeed, a Line in $\mathbb{RP}^2$ is simply a plane in $\mathbb{R}^3$ that passes through the origin, and two distinct planes through the origin of $\mathbb{R}^3$ must intersect in a unique Euclidean line through the origin; that is, in a Point. This is very different to the situation in Euclidean geometry where parallel lines do not meet.

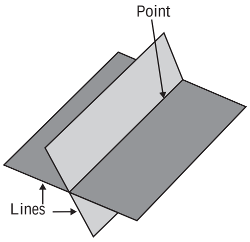

Theorem 3 Incidence Property of $\mathbb{RP}^2$
Any two distinct Lines in $\mathbb{RP}^2$ intersect in a unique Point of $\mathbb{RP}^2$.

This result neatly complements Theorem 2, the *Collinearity Property* of $\mathbb{RP}^2$.

We can determine the Point of intersection of two Lines simply by solving the equations of the two Lines as a pair of simultaneous equations.

Example 6 Determine the Point of intersection of the Lines in $\mathbb{RP}^2$ with equations $x + 6y - 5z = 0$ and $x - 2y + z = 0$.

We know that there is a unique Point of intersection, by Theorem 3.

Solution At the Point of intersection $[x, y, z]$ of the two Lines, we have

$$x + 6y - 5z = 0,$$
$$x - 2y + z = 0.$$

Subtracting the second equation from the first, we obtain

$$8y - 6z = 0,$$

so that $y = \frac{3}{4}z$. Substituting this into the second equation, we obtain $x = \frac{1}{2}z$.

It follows that the Point of intersection has homogeneous coordinates $\frac{1}{2}z, \frac{3}{4}z, z$ which we can rewrite in the form $\frac{1}{2}, \frac{3}{4}, 1$ or [2, 3, 4]. □

Note that $z = 0$, since [0, 0, 0] are not allowed as homogeneous coordinates.

Problem 8 Determine the Point of intersection of each of the following pairs of Lines in RP^2:

(a) the Lines with equations $x \quad y \quad z = 0$ and $x + 5y + 2z = 0$;
(b) the Lines with equations $x + 2y \quad z = 0$ and $2x + y \quad 4z = 0$.

Problem 9 Determine the Point of RP^2 at which the Line through the Points $[1, 2, \quad 3]$ and $[2, \quad 1, 0]$ meets the Line through the Points $[1, 0, \quad 1]$ and [1, 1, 1].

In some cases we can write down the Point at which two Lines intersect without having to solve any equations at all. For example, the Lines with equations $x = 0$ and $y = 0$ clearly meet at the Point [0, 0, 1].

Problem 10 Determine the Point of RP^2 at which the Line through the Points [1, 0, 0] and [0, 1, 0] meets the Line through the Points [0, 0, 1] and [1, 1, 1].

3.2.3 Embedding Planes

So far we have used three-dimensional space to develop the theory of projective geometry. In practice, however, we want to use projective geometry to study two-dimensional figures in a plane. In order to do this, we now investigate a way of associating figures in a plane with figures in RP^2, and vice versa.

Suppose that a plane contains a figure F. We can place into R^3, making sure that it does not pass through the origin, and then construct a corresponding projective figure by drawing in all the Points of RP^2 that pass through the points of F. For example, if F is the triangle shown on the left below, then the corresponding projective figure is a double triangular pyramid. Note that if we change the position of in R^3, we obtain a different projective figure corresponding to F.

It is a *double* pyramid because the Points which make up the pyramid are lines that emerge from the origin in *both* directions.

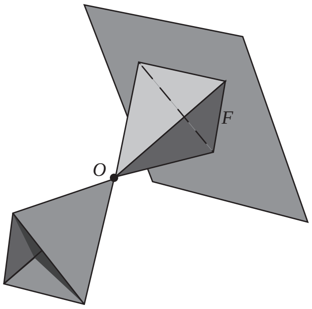

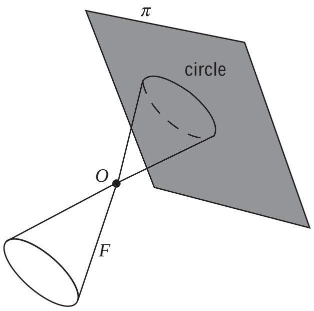

Conversely, suppose that we start with a projective figure F. The corresponding Euclidean figure in consists of the Euclidean points where the Points of F pierce . For example, if F is a double cone whose axis is at right angles to the embedding plane, as shown on the right above, then the corresponding Euclidean figure is a circle. Note that if we change the position of in $\mathbb{R}^3$, we obtain a different plane figure corresponding to F.

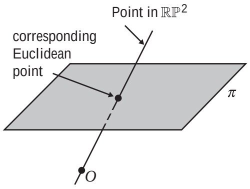

This correspondence between projective figures and Euclidean figures works well provided that each Point of the projective figure pierces the plane , as shown in the margin. Unfortunately, any Point of $\mathbb{RP}^2$ that consists of a line through the origin *parallel* to does not pierce , and so cannot be associated with a point of . Such a Point is called an **ideal Point** for .

All the ideal Points for lie on a plane through O parallel to . This plane is a projective line known as the **ideal Line** for .

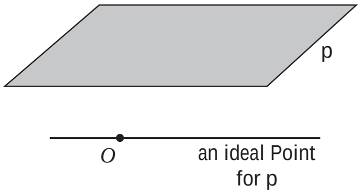

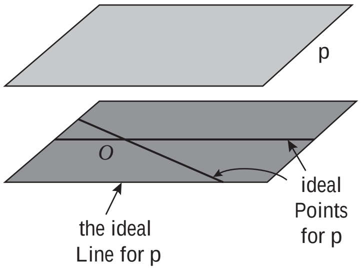

How can we represent a projective figure on if the figure includes some of the ideal Points for ? As a simple example, consider the Line illustrated in the margin. This is a projective figure which intersects in a line . Every Point of the Line pierces the embedding plane at a point of except for the ideal Point P which cannot be represented on . In order to represent the Line completely, we need not only the line but also the ideal Point P. In other words, the Line is represented by $\{P\}$.

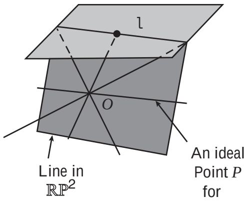

In general, a projective figure can be represented by a figure in provided that we are prepared to include a subset of Points taken from the ideal Line for . In order to allow for these additional ideal Points, we introduce the concept of an *embedding plane.*

Definitions An **embedding plane** is a plane, , which does not pass through the origin, together with the set of all ideal Points for . The plane in $\mathbb{R}^3$ with equation $z = 1$ is called the **standard embedding plane**. The mapping of $\mathbb{RP}^2$ into the standard embedding plane is called the **standard embedding** of $\mathbb{RP}^2$.

We frequently use to denote both a plane and an embedding plane, but no confusion should arise.

We may summarize the above discussion by saying that for a given embedding plane, every projective figure in $\mathbb{RP}^2$ corresponds to a figure in the embedding plane, and vice versa. The figure in the embedding plane may include some ideal Points but is otherwise a Euclidean figure.

If two embedding planes are parallel to each other, the same Points of $\mathbb{RP}^2$ correspond to ideal Points of the embeddings; whereas, if the embedding planes are not parallel, different Points of $\mathbb{RP}^2$ correspond to ideal Points of the two embedding planes.

Once we have represented a projective figure in an embedding plane, we can investigate the relationship between its Points and Lines without having to refer to three-dimensional space at all. For example, consider the representation of the triangle of reference and unit Point on the embedding plane $x + y + z = 1$, shown on the left below. If we extract the embedding plane from $\mathbb{R}^3$, as shown on the right, we can use the algebraic theory developed earlier to write down an equation for the Line through any two given Points, without reference to $\mathbb{R}^3$.

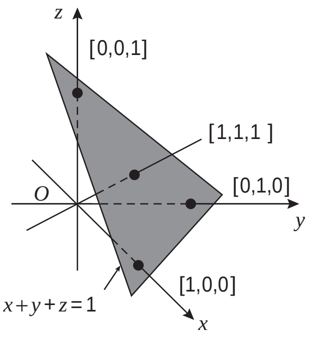

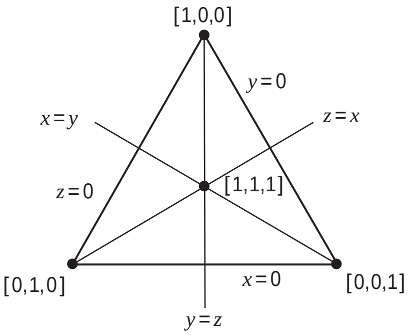

Similarly, we can use the algebraic techniques to calculate the homogeneous coordinates of the Point of intersection of any two given Lines.

> **Problem 11** On the right-hand diagram above, insert the homogeneous coordinates of the Points where the Lines through [1, 1, 1] meet the sides of the triangle of reference.

Any plane may be used as an embedding plane provided that it does not pass through the origin. For example, if we take to be the plane $z =$ 1, then the ideal Line for has equation $z = 0$, and the ideal Points are Points of the form $[a, b, 0]$, where a and b are not both zero. Any other Point $[a, b, c]$ has $c = 0$ and can therefore be represented in by the Euclidean point(a/c, b/c, 1).

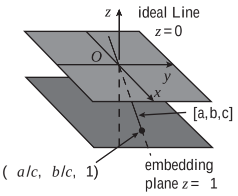

> **Problem 12** Let be the embedding plane $y =$ 1. Describe the ideal Points for , and specify the Euclidean point of which represents the Point [2, 4, 6].

Although we can choose any embedding plane to represent figures of $\mathbb{RP}^2$, the representation does depend on the choice. For example, suppose that $_1$ is the embedding plane $y = 1$, and that $_2$ is the embedding plane $z = 1$. Now consider the projective figure which consists of two Lines $_1$ and $_2$ with equations $x = z$ and $x = z$, respectively. These Lines intersect at the Point [0, 1, 0].

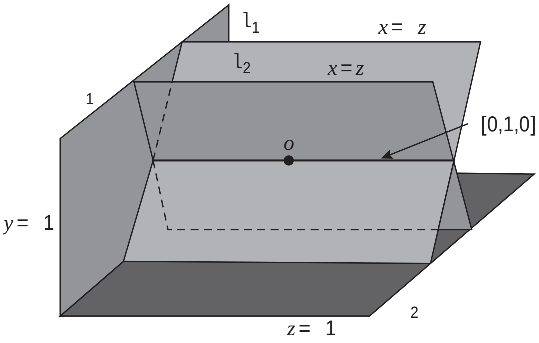

On the embedding plane $_1$ the Lines $_1$ and $_2$ are represented by two lines that can be seen to meet at the point corresponding to [0, 1, 0]. However, on the embedding plane $_2$ the Point of intersection [0, 1, 0] is an ideal Point and so the Lines $_1$ and $_2$ are represented by parallel lines that do not appear to meet. The contrast between the two representations of $_1$ and $_2$ is particularly striking if we extract the two embedding planes from $\mathbb{R}^3$ and lay them side by side, as follows.

This is the mathematical fact which explains why artists sometimes draw parallel lines (such as railroad lines) as intersecting lines.

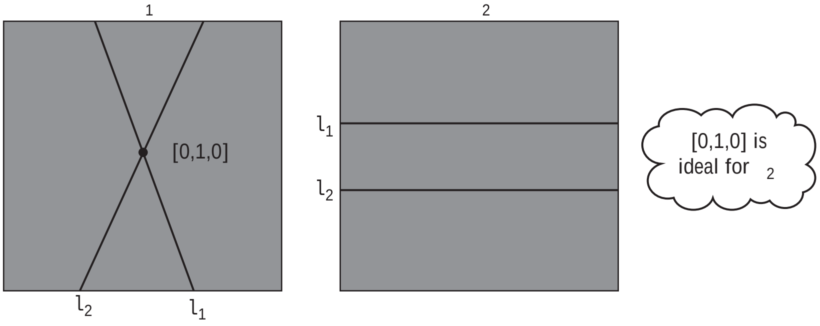

This example illustrates that Lines which appear to be parallel in one embedding plane may not appear to be parallel in another embedding plane. In the next section you will see that the transformations of projective geometry are chosen so as to ensure that the projective properties of a figure are unaffected by the choice of embedding plane. Since parallelism does depend on the choice of embedding plane, it cannot be a projective property, so the concept of parallel Lines is meaningless in projective geometry.

3.2.4 An equivalent definition of Projective Geometry

In our work on projective geometry, we have used Euclidean points in a plane in $\mathbb{R}^3$ to construct the projective points (Points) of the geometry $\mathbb{RP}^2$, homogeneous coordinates for those Points, and projective lines (Lines).

Equivalently, we could have defined $\mathbb{RP}^2$ as the set of ordered triples $[a, b, c]$, where a, b, c are real and not all zero, with the convention that we regard [a, b, c] and $[a, b, c]$ (where $= 0$) as the same Point in the geometry. We would then have defined projective lines (Lines) as the set of points $[x, y, z]$ in $\mathbb{RP}^2$ that satisfy an equation of the form $ax + by + cz = 0$, where a, b, c are real and not all zero, Then we would continue to develop the theory of projective geometry in the same way as we have done here.

However, we chose to start our work by looking at a model of $\mathbb{RP}^2$ obtained by using an embedding plane in $\mathbb{R}^3$ that does not pass through the origin. We modeled the projective points $[a, b, c]$ by the Euclidean lines through the origin and the corresponding Euclidean points (a, b, c), plus points at infinity (the ideal Points); and we modeled the projective lines by Euclidean planes through the origin, For convenience, we chose often to use Euclidean points (a, b, c) on a given embedding plane to describe the Euclidean model.

Any plane that does not pass through the origin in $\mathbb{R}^3$ will serve as an embedding plane.

The formal method of defining projective geometry, though, is less intuitive than the description motivated by the $\mathbb{R}^3$ model!

3.3 Projective Transformations

3.3.1 The Group of Projective Transformations

By now you should be familiar with the idea that a geometry consists of a space of points together with a group of transformations which act on that space.

Having introduced the space of projective points $\mathbb{RP}^2$ in Section 3.2, we are now in a position to describe the transformations of $\mathbb{RP}^2$. First we shall define the transformations algebraically, then we give a geometrical interpretation of the transformations using the ideas of perspectivity introduced in Section 3.1, and finally meet the Fundamental Theorem of Projective Geometry.

Recall that a point of $\mathbb{R}^3$ (other than the origin) on an embedding plane (that does not pass through the origin) has coordinates $\mathbf{x} = (x, y, z)$ with respect to the standard basis of $\mathbb{R}^3$, and homogeneous coordinates of the corresponding Point $[\mathbf{x}]$ in $\mathbb{RP}^2$ (which represents the points $\{\ \mathbf{x} : \quad \mathbb{R}\}$) are [x, y, z] for some real $= 0$. Since the Points of $\mathbb{RP}^2$ are just lines through the origin of $\mathbb{R}^3$, we need a group of transformations that map the lines through the origin of $\mathbb{R}^3$ onto the lines through the origin of $\mathbb{R}^3$. Suitable transformations of $\mathbb{R}^3$ that do this are the invertible linear transformations.

Subsection 3.2.1

If **A** is the matrix of an invertible linear transformation of $\mathbb{R}^3$ to itself, the transformation maps points $\mathbf{x} = (x, y, z)$ of $\mathbb{R}^3$ to points $\mathbf{Ax}$ of $\mathbb{R}^3$; then the *projective transformation* with matrix **A** maps Points $[\mathbf{x}]$ of $\mathbb{RP}^2$ to Points $[\mathbf{Ax}]$ of $\mathbb{RP}^2$. This suggests that we define the transformations of projective geometry as follows.

In fact, any continuous transformation of $\mathbb{RP}^2$ to itself that maps Lines to Lines and that preserves incidences of Lines corresponds to an invertible linear transformation of $\mathbb{R}^3$. We omit a proof of this fact.

Definitions A **projective transformation** of $\mathbb{RP}^2$ is a function $t : \mathbb{RP}^2 \to \mathbb{RP}^2$ of the form

$$t : [\mathbf{x}] \mapsto [\mathbf{Ax}],$$

where **A** is an invertible 3×3 matrix. We say that **A** is a matrix **associated** with t. The set of all projective transformations of $\mathbb{RP}^2$ is denoted by $P(2)$.

Example 1 Show that the function $t : \mathbb{RP}^2 \to \mathbb{RP}^2$ defined by

$$t : [x, y, z] \mapsto [2x + z, -x + 2y - 3z, x - y + 5z]$$

is a projective transformation, and find the image of $[1, 2, 3]$ under t.

Solution The transformation t has the form $t : [\mathbf{x}] \mapsto [\mathbf{Ax}]$, where $\mathbf{x} = (x, y, z)$ and

$$\mathbf{A} = \begin{pmatrix} 2 & 0 & 1 \\ -1 & 2 & -3 \\ 1 & -1 & 5 \end{pmatrix}.$$

Now

$$\begin{aligned} \det \mathbf{A} &= \begin{vmatrix} 2 & 0 & 1 \\ -1 & 2 & -3 \\ 1 & -1 & 5 \end{vmatrix} \\ &= 2(10 - 3) - 0 + (1 - 2) \\ &= 13 \neq 0. \end{aligned}$$

So **A** is invertible. It follows that t is a projective transformation.

We have

$$t([1, 2, 3]) = [2 + 3, -1 + 4 - 9, 1 - 2 + 15] = [5, -6, 14]. \qquad \square$$

Problem 1 Decide which of the following functions t from $\mathbb{RP}^2$ to itself are projective transformations. For those that are projective transformations, write down a matrix associated with t.

(a) $t : [x, y, z] \mapsto [-2y + 3z, -x + 5y - z, -3x]$
(b) $t : [x, y, z] \mapsto [x - 7y + 4z, -x + 5y - z, x - 9y + 7z]$
(c) $t : [x, y, z] \mapsto [x - 1 + z, 2y - 4z + 5, 2x]$

Problem 2 Let t be the projective transformation associated with the matrix.

$$\mathbf{A} = \begin{matrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 4 & 3 & 4 \end{matrix}.$$

Determine the image under t of each of the following Points.

(a) $[1, 2, \quad 1]$ (b) $[1, 0, 0]$ (c) $[0, 1, 0]$
(d) $[0, 0, 1]$ (e) $[1, 1, 1]$

Since we can multiply the homogeneous coordinates of Points in $\mathbb{RP}^2$ by any non-zero real number without altering the Point itself, it follows that if **A** is a matrix associated with a particular projective transformation then so is the matrix **A**, provided that $= 0$. For example, another matrix associated with the transformation in Example 1 is

$$\mathbf{B} = \begin{matrix} 4 & 0 & 2 \\ 2 & 4 & 6 \\ 2 & 2 & 10 \end{matrix},$$

for we have $\mathbf{B} = \quad 2\mathbf{A}$.

Problem 3 Write down a matrix with top left-hand entry $\frac{1}{2}$ which is associated with the transformation in Example 1.

Before we can use the projective transformations to define projective geometry, we must first check that they form a group.

Theorem 1 The set of projective transformations $P(2)$ forms a group under the operation of composition of functions.

Recall that a similar result holds for affine transformations.

Proof We check that the four group axioms hold.

G1 CLOSURE Let t_1 and t_2 be projective transformations defined by

$$t_1 : [\mathbf{x}] \quad [\mathbf{A_1x}] \quad \text{and} \quad t_2 : [\mathbf{x}] \quad [\mathbf{A_2x}],$$

where $\mathbf{A}_1$ and $\mathbf{A}_2$ are invertible 3×3 matrices. Then

$$\begin{aligned} t_1 \quad t_2([\mathbf{x}]) &= t_1(t_2([\mathbf{x}])) \\ &= t_1([\mathbf{A_2x}]) \\ &= [(\mathbf{A_1A_2})\mathbf{x}]. \end{aligned}$$

Since $\mathbf{A}_1$ and $\mathbf{A}_2$ are invertible, it follows that $\mathbf{A_1A_2}$ is invertible. So by definition $t_1 \quad t_2$ is a projective transformation.

G2 IDENTITY Let $i : \mathbb{RP}^2 \quad \mathbb{RP}^2$ be the transformation defined by

$$i : [\mathbf{x}] \quad [\mathbf{Ix}],$$

where **I** is the 3×3 identity matrix; this is a projective transformation, since **I** is invertible.

Let t: $\mathbb{RP}^2 \to \mathbb{RP}^2$ be an arbitrary projective transformation, defined by $t:[\mathbf{x}] \mapsto [\mathbf{Ax}]$, for some invertible 3 × 3 matrix $\mathbf{A}$. Then for any $[\mathbf{x}] \in \mathbb{RP}^2$,

$$t \circ i([\mathbf{x}]) = [\mathbf{A}(\mathbf{Ix})] = [\mathbf{Ax}]$$

and

$$i \circ t([\mathbf{x}]) = [\mathbf{I}(\mathbf{Ax})] = [\mathbf{Ax}].$$

Thus $t \circ i = i \circ t = t$. Hence i is the identity transformation.

G3 INVERSES Let $t : \mathbb{RP}^2 \to \mathbb{RP}^2$ be an arbitrary projective transformation defined by

$$t: [\mathbf{x}] \mapsto [\mathbf{Ax}],$$

for some invertible 3 × 3 matrix $\mathbf{A}$. Then we can define another projective transformation t' : $\mathbb{RP}^2 \to \mathbb{RP}^2$ by

$$t' : [\mathbf{x}] \mapsto [\mathbf{A}^{-1}\mathbf{x}].$$

Now, for each $[\mathbf{x}] \in \mathbb{RP}^2$, we have

$$t \circ t'([\mathbf{x}]) = \mathbf{t}([\mathbf{A}^{-1}\mathbf{x}]) = [\mathbf{A}(\mathbf{A}^{-1}\mathbf{x})] = [\mathbf{x}]$$

and

$$t' \circ t([\mathbf{x}]) = t'([\mathbf{Ax}]) = [\mathbf{A}^{-1}(\mathbf{Ax})] = [\mathbf{x}].$$

Thus t' is an inverse for t.

G4 ASSOCIATIVITY Composition of functions is always associative.

It follows that the set of projective transformations $P(2)$ forms a group.

The above proof shows that if t_1 and t_2 are projective transformations with associated matrices $\mathbf{A}_1$ and $\mathbf{A}_2$, respectively, then $t_1 \circ t_2$ is a projective transformation with an associated matrix $\mathbf{A}_1\mathbf{A}_2$. We therefore have the following strategy for composing projective transformations.

Strategy To compose two projective transformations t_1 and t_2:

1. write down matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ associated with t_1 and t_2;
2. calculate $\mathbf{A}_1\mathbf{A}_2$;
3. write down the composite $t_1 \circ t_2$ with which $\mathbf{A}_1\mathbf{A}_2$ is associated.

The proof also shows that if t is a projective transformation with an associated matrix $\mathbf{A}$, then t^{-1} is a projective transformation with associated matrix $\mathbf{A}^{-1}$. We therefore have the following strategy for calculating the inverse of a projective transformation.

See Appendix 2 for one method to calculate $\mathbf{A}^{-1}$.

Strategy To find the inverse of a projective transformation t:

1. write down a matrix $\mathbf{A}$ associated with t;
2. calculate $\mathbf{A}^{-1}$;
3. write down the inverse t^{-1} with which $\mathbf{A}^{-1}$ is associated.

Example 2 Let t_1 and t_2 be projective transformations defined by

$$t_1 : [x, y, z] \mapsto [x + z, x + y + 3z, -2x + z],$$
$$t_2 : [x, y, z] \mapsto [2x, x + y + z, 4x + 2y].$$

Determine the projective transformations $t_2 \circ t_1$ and t_1^{-1}.

Solution The transformations t_1 and t_2 have associated matrices

$$\mathbf{A}_1 = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 3 \\ -2 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{A}_2 = \begin{pmatrix} 2 & 0 & 0 \\ 1 & 1 & 1 \\ 4 & 2 & 0 \end{pmatrix},$$

respectively. It follows that $t_2 \circ t_1$ has an associated matrix

$$\mathbf{A}_2\mathbf{A}_1 = \begin{pmatrix} 2 & 0 & 0 \\ 1 & 1 & 1 \\ 4 & 2 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 3 \\ -2 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 2 \\ 0 & 1 & 5 \\ 6 & 2 & 10 \end{pmatrix},$$

so

$$t_2 \circ t_1 : [x, y, z] \mapsto [2x + 2z, y + 5z, 6x + 2y + 10z].$$

Next, t_1^{-1} has an associated matrix $\mathbf{A}_1^{-1}$ given by

$$\mathbf{A}_1^{-1} = \begin{pmatrix} \frac{1}{3} & 0 & -\frac{1}{3} \\ -\frac{7}{3} & 1 & -\frac{2}{3} \\ \frac{2}{3} & 0 & \frac{1}{3} \end{pmatrix};$$

a simpler matrix associated with t_1^{-1} is then

$$\begin{pmatrix} 1 & 0 & -1 \\ -7 & 3 & -2 \\ 2 & 0 & 1 \end{pmatrix},$$

so

$$t_1^{-1} : [x, y, z] \mapsto [x - z, -7x + 3y - 2z, 2x + z].$$ □

Problem 4 Let t_1 and t_2 be projective transformations defined by

$$t_1 : [x, y, z] \mapsto [2x + y, -x + z, y + z],$$
$$t_2 : [x, y, z] \mapsto [5x + 8y, 3x + 5y, 2z].$$

Determine the projective transformations $t_1 \circ t_2$ and t_1^{-1}.

Having shown that the set of projective transformations forms a group under composition of functions, we can now define **projective geometry** to be the study of those properties of figures in $\mathbb{RP}^2$ that are preserved by projective transformations. Those properties that are preserved by projective transformations are known as **projective properties.**

3.3.2 Some Properties of Projective Transformations

We now check two important properties of projective transformations, namely, that they preserve collinearity and incidence.

A Line in $\mathbb{RP}^2$ is a plane in $\mathbb{R}^3$ that passes through the origin. It therefore consists of the set of points (x, y, z) of $\mathbb{R}^3$ that satisfy an equation of the form

$$ax + by + cz = 0,$$

where a, b and c are not all zero. We can write this condition equivalently in the matrix form $\mathbf{Lx} = 0$, where $\mathbf{L}$ is the non-zero row matrix $(a\ b\ c)$ and $\mathbf{x} = (x\ y\ z)^T$.

Now let t be a projective transformation defined by $t : [\mathbf{x}] \mapsto [\mathbf{Ax}]$, where $\mathbf{A}$ is an invertible 3×3 matrix, and let $[\mathbf{x}]$ be an arbitrary Point on the Line $\mathbf{Lx} = 0$. Then the image of $[\mathbf{x}]$ under t is a Point $[\mathbf{x}']$ where $\mathbf{x}' = \mathbf{Ax}$. Since $\mathbf{x}$ satisfies the equation $\mathbf{Lx} = 0$, it follows that $\mathbf{x}'$ satisfies $\mathbf{L}(\mathbf{A}^{-1}\mathbf{x}') = 0$, or $(\mathbf{LA}^{-1})\mathbf{x}' = 0$. Dropping the dash, we conclude that the image of the Line $\mathbf{Lx} = 0$ under t is the Line with equation

$$\left(\mathbf{LA}^{-1}\right)\mathbf{x} = 0.$$

Here, $\mathbf{LA}^{-1}$ is non-zero, for if

$$\mathbf{LA}^{-1} = \mathbf{0},$$

then

$$\mathbf{0} = \left(\mathbf{LA}^{-1}\right)\mathbf{A}$$

$$= \mathbf{L}\left(\mathbf{A}^{-1}\mathbf{A}\right) = \mathbf{L},$$

which is not the case.

Since the image of a Line in $\mathbb{RP}^2$ is a Line, it follows that collinearity is preserved under a projective transformation.

Notice that if $\mathbf{B}$ is any matrix associated with t^{-1}, then $\mathbf{B} = \lambda\mathbf{A}^{-1}$ for some non-zero real number λ, and so $\left(\mathbf{LA}^{-1}\right)\mathbf{x} = 0$ if and only if $(\mathbf{LB})\mathbf{x} = 0$. It follows that the image of the Line can equally well be written as $(\mathbf{LB})\mathbf{x} = 0$. (For instance, since $\mathbf{A}^{-1} = \operatorname{adj}(\mathbf{A}) / \det(\mathbf{A})$ so that t^{-1} also has $\operatorname{adj}(\mathbf{A})$ as an associated matrix, we can express the image of the Line as $(\mathbf{L}\operatorname{adj}(\mathbf{A}))\mathbf{x} = 0$.)

We therefore summarize the above discussion in the form of a strategy, as follows.

Strategy To find the image of a Line

$$ax + by + cz = 0$$

under a projective transformation $t : [\mathbf{x}] \mapsto [\mathbf{AX}]$:

1. write the equation of the Line in the form $\mathbf{Lx} = 0$, where $\mathbf{L}$ is the matrix $(a\ b\ c)$;
2. find a matrix $\mathbf{B}$ associated with t^{-1};
3. write down the equation of the image as $(\mathbf{LB})\mathbf{x} = 0$.

Example 3 Find the image of the Line $2x + y - 3z = 0$ under the projective transformation t_1 defined by

$$t_1 : [x, y, z] \qquad [x + z, x + y + 3z, \quad 2x + z].$$

Solution The equation of the Line can be written in the form $\mathbf{L}\mathbf{x} = 0$, where

$$\mathbf{L} = (2 \quad 1 \qquad 3).$$

In Example 2 we showed that $t_1^{\ \ 1}$ has an associated matrix

$$\mathbf{B} = \begin{matrix} 1 & 0 & 1 \\ 7 & 3 & 2 \\ 2 & 0 & 1 \end{matrix} .$$

So

$$\mathbf{LB} = (2 \quad 1 \quad 3) \begin{matrix} 1 & 0 & 1 \\ 7 & 3 & 2 \\ 2 & 0 & 1 \end{matrix} = (\quad 11 \quad 3 \quad 7).$$

It follows that the required image has equation

$$11x + 3y \quad 7z = 0.$$ □

Problem 5 Find the image of the Line $x + 2y \quad z = 0$ under the projective transformation t_1 defined by

$$t_1 : [x, y, z] \qquad [2x + y, \quad x + z, y + z].$$

Next, we consider the incidence property. If two Lines intersect at the Point P, then P lies on both Lines. So if t is a projective transformation, then $t(P)$ lies on the images of both Lines. It follows that the image under t of the Point of intersection of the two Lines is the Point of intersection of the images of the two Lines. In other words, incidence is also preserved under a projective transformation.

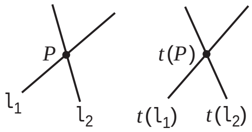

Theorem 2 Collinearity and incidence are both projective properties.

3.3.3 Fundamental Theorem of Projective Geometry

In Chapter 2 we discussed the Fundamental Theorem of Affine Geometry which states that given any two sets of three non-collinear points of $\mathbb{R}^2$ there is a unique affine transformation which maps the points in one set to the corresponding points in the other set. So an affine transformation is uniquely determined by its effect on any given triangle.

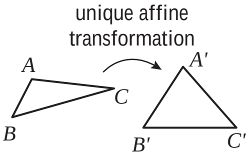

In this subsection we explore an analogous result for projective geometry known as the *Fundamental Theorem of Projective Geometry*. We begin by asking you to tackle the following problem.

Problem 6 Let t_1 and t_2 be the projective transformations with associated matrices

$$\mathbf{A}_1 = \begin{matrix} 4 & 1 & 1 \\ 3 & 2 & 1 \\ 4 & 2 & 1 \end{matrix} \quad \text{and} \quad \mathbf{A}_2 = \begin{matrix} 8 & 6 & 2 \\ 3 & 4 & 7 \\ 6 & 0 & 4 \end{matrix} ,$$

respectively. Find the images of the Points $[1, -1, 1]$, $[1, -2, 2]$ and $[-1, 2, -1]$ under t_1 and t_2.

You should have found that both of the projective transformations t_1 and t_2 map the Points $[1, -1, 1]$, $[1, -2, 2]$ and $[-1, 2, -1]$ to the Points $[-2, 0, 1]$, $[0, 3, -2]$ and $[1, -2, 1]$, respectively. Notice, however, that t_1 and t_2 are not the same projective transformation, since their matrices are not multiples of each other. It follows that, unlike affine transformations, projective transformations are not uniquely determined by their effect on three (non-collinear) Points.

This raises the question as to whether it is possible to specify how many Points *are* required to determine a projective transformation. According to the Fundamental Theorem of Projective Geometry, the answer is four. In fact the theorem states that given any two sets of four Points, no three of which are collinear, there is a *unique* projective transformation that maps the Points in one set to the corresponding Points in the second set. Thus, in projective geometry a transformation is uniquely determined by its effect on a quadrilateral.

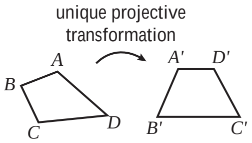

To understand why a triangle is insufficient to determine a projective transformation uniquely, consider what happens when we look for a projective transformation that maps the triangle of reference to three given non-collinear Points.

Recall that the Points [1, 0, 0], [0, 1, 0] and [0, 0, 1] are known as the *triangle of reference.*

Example 4 Find a projective transformation t that maps the Points [1, 0, 0], [0, 1, 0] and [0, 0, 1] to the non-collinear Points $[1, -1, 1]$, $[1, -2, 2]$ and $[-1, 2, -1]$, respectively.

Solution Let $\mathbf{A}$ be a matrix associated with t, and let the first column of $\mathbf{A}$ be $\begin{pmatrix} a \\ b \\ c \end{pmatrix}$. Then since

$$\begin{pmatrix} a & * & * \\ b & * & * \\ c & * & * \end{pmatrix}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix},$$

Here the asterisks * denote unspecified numbers.

it follows that we may take $\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$ as the first column of $\mathbf{A}$.

Similarly, since

$$\begin{pmatrix} * & d & * \\ * & e & * \\ * & f & * \end{pmatrix}\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} d \\ e \\ f \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix}$$

and

$$\begin{array}{cc} g & 0 \\ h & 0 \\ k & 1 \end{array} = \begin{array}{c} g \\ h \\ k \end{array} = \begin{array}{c} 1 \\ 2 \\ 1 \end{array},$$

it follows that a suitable transformation is given by $t : [\mathbf{x}] \quad [\mathbf{A}x]$ where

$$\mathbf{A} = \begin{array}{ccc} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 1 \end{array}.$$

□

Notice that because the Points [1, 1, 1], [1, 2, 2] and [1, 2, 1] are not collinear it follows that the columns of **A** are linearly independent, so that **A** is invertible.

This example illustrates the fact that we can always find a projective transformation $t : [\mathbf{x}] \quad [\mathbf{Ax}]$ which maps the triangle of reference to three non-collinear Points simply by writing the homogeneous coordinates of the Points as the columns of **A**. Notice, however, that the transformation we obtain is not unique. Indeed, if the Points [1, 1, 1], [1, 2, 2] and [1, 2, 1] in Example 4 are rewritten in the form $[u, \ u, u]$, [, 2 , 2] and [w, 2w, w], for some non-zero real numbers u, , w, then the matrix becomes

$$\mathbf{A} = \begin{array}{ccc} u & & w \\ u & 2 & 2w \\ u & 2 & w \end{array}.$$

The corresponding transformation $t : [\mathbf{x}] \quad [\mathbf{Ax}]$ still maps the triangle of reference to the Points [1, 1, 1], [1, 2, 2] and [1, 2, 1], as required, but the effect that t has on the other Points of $\mathbb{RP}^2$ depends on the numbers u, v and w.

So if we wish to specify t uniquely we need to assign particular values to u, v and w. We can do this by specifying the effect that t has on a fourth Point [1, 1, 1].

Recall that the Point [1, 1, 1] is known as the *unit Point.*

Example 5 Find the projective transformation t which maps the Points [1, 0, 0], [0, 1, 0], [0, 0, 1] and [1, 1, 1] to the Points [1, 1, 1], [1, 2, 2], [1, 2, 1] and [0, 1, 2], respectively.

Solution If **A** is the matrix associated with t, then its columns must be multiples of the homogeneous coordinates [1, 1, 1], [1, 2, 2], [1, 2, 1]; that is,

$$\mathbf{A} = \begin{array}{ccc} u & & w \\ u & 2 & 2w \\ u & 2 & w \end{array}.$$

Also, to ensure that t maps [1, 1, 1] to [0, 1, 2] we must choose u, v and w so that

$$\begin{pmatrix} u & v & -w \\ -u & -2v & 2w \\ u & 2v & -w \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}.$$

We can do this by solving the equations

$$\begin{aligned} u + v - w &= 0, \\ -u - 2v + 2w &= 1, \\ u + 2v - w &= 2. \end{aligned}$$

Adding the second and third equations we obtain $w = 3$. If we then subtract the first equation from the third we obtain $v = 2$. Finally, if we substitute v and w into the first equation we obtain $u = 1$. The required projective transformation is therefore given by $t : [\mathbf{x}] \mapsto [\mathbf{Ax}]$, where

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & -3 \\ -1 & -4 & 6 \\ 1 & 4 & -3 \end{pmatrix}.$$

□

The columns of **A** are still linearly independent because they are non-zero multiples of the linearly independent vectors $(1, -1, 1)$, $(1, -2, 2)$ and $(-1, 2, -1)$.

It is natural to ask whether the method used in this example can be adapted to find a projective transformation which maps the triangle of reference and unit Point to *any* four given Points. The answer is usually yes, but since collinearity is a projective property, and since no three of the Points [1, 0, 0], [0, 1, 0], [0, 0, 1], [1, 1, 1] are collinear, the method must fail if three of the four given Points lie on a Line. Provided we exclude this possibility, the answer is yes!

Strategy To find the projective transformation which maps

$[1, 0, 0]$ to $[a_1, a_2, a_3]$,
$[0, 1, 0]$ to $[b_1, b_2, b_3]$,
$[0, 0, 1]$ to $[c_1, c_2, c_3]$,
$[1, 1, 1]$ to $[d_1, d_2, d_3]$,

where no three of $[a_1, a_2, a_3], [b_1, b_2, b_3], [c_1, c_2, c_3], [d_1, d_2, d_3]$ are collinear:

1. Find u, v, w such that

$$\begin{pmatrix} a_1u & b_1v & c_1w \\ a_2u & b_2v & c_2w \\ a_3u & b_3v & c_3w \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix};$$

We explain why the method works in the Remark that follows the strategy.

2. write down the required projective transformation in the form $t : [\mathbf{x}]$ $[\mathbf{Ax}]$, where $\mathbf{A}$ is any non-zero real multiple of the matrix

$$\begin{matrix} a_1 u & b_1 v & c_1 w \\ a_2 u & b_2 v & c_2 w \\ a_3 u & b_3 v & c_3 w \end{matrix} .$$

The non-zero multiple can be used to clear fractions from the entries of the matrix **A**.

Remark

To see why this strategy always works, notice that we can rewrite the equation from Step 1 in the form

$$u \begin{matrix} a_1 \\ a_2 \\ a_3 \end{matrix} + v \begin{matrix} b_1 \\ b_2 \\ b_3 \end{matrix} + w \begin{matrix} c_1 \\ c_2 \\ c_3 \end{matrix} = \begin{matrix} d_1 \\ d_2 \\ d_3 \end{matrix} .$$

From this we can make the following observations.

(a) The equation in Step 1 must have a unique solution for u, v, w because the required values of u, v and w are simply the coordinates of (d_1, d_2, d_3) with respect to the basis of $\mathbb{R}^3$ formed from the three linearly independent vectors $(a_1, a_2, a_3), (b_1, b_2, b_3), (c_1, c_2, c_3)$.

(b) The values *of* u, v and w must all be non-zero, because otherwise three of the vectors $(a_1, a_2, a_3), (b_1, b_2, b_3), (c_1, c_2, c_3), (d_1, d_2, d_3)$ would be linearly dependent.

(c) Since the columns of $\mathbf{A}$ are non-zero, multiples of the linearly independent vectors $(a_1, a_2, a_3), (b_1, b_2, b_3), (c_1, c_2, c_3)$ it follows that $\mathbf{A}$ is invertible, and hence that t is a projective transformation.

Since no three of the Points $[a_1, a_2, a_3]$, $[b_1, b_2, b_3]$, $[c_1, c_2, c_3]$, $[d_1, d_2, d_3]$ are collinear, it follows that *any three* of the vectors (a_1, a_2, a_3), (b_1, b_2, b_3), (c_1, c_2, c_3), (d_1, d_2, d_3) must be linearly independent.

There is no need to check whether any three of the four given Points are collinear because any failure of this condition will emerge in the process of applying the strategy. Indeed, if the equation in Step 1 fails to yield unique non-zero values for u, v and w, then it must be because three of the Points $(a_1, a_2, a_3), (b_1, b_2, b_3), (c_1, c_2, c_3), (d_1, d_2, d_3)$ lie on a Line.

Problem 7 Use the above strategy to find the projective transformation which maps the Points [1, 0, 0], [0, 1, 0], [0, 0, 1] and [1, 1, 1] to the Points:

(a) [1, 0, 0], [3, 2, 0], [2, 0, 4] and [1, 2, 5], respectively;
(b) [1, 0, 0], [0, 0, 1], [0, 1, 0] and [3, 4, 5], respectively;
(c) [2, 1, 0], [1, 0, 1], [0, 3, 1] and [3, 1, 2], respectively.

Now consider the transformation t_1 in Problem 7(a). The inverse of this, $t_1{}^{1}$, is a projective transformation which maps the Points [1, 0, 0], [3, 2, 0], [2, 0, 4] and [1, 2, 5] back to the triangle of reference and unit Point. So if, after applying this inverse, we apply the projective transformation t_2 in Problem 7(c), then the overall effect of the composite $t_2 \circ t_1^{-1}$ is that of a projective transformation which sends the Points $[-1, 0, 0]$, $[-3, 2, 0]$, $[2, 0, 4]$ and $[1, 2, -5]$ directly to the Points $[2, 1, 0]$, $[1, 0, -1]$, $[0, 3, -1]$ and $[3, -1, 2]$, respectively.

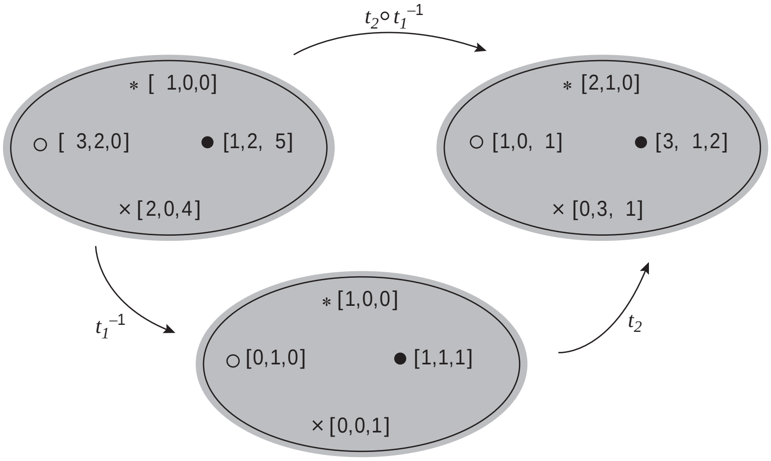

In a similar way we can find a projective transformation which maps any set of four Points to any other set of four Points. The only constraint is that no three of the Points in either set can be collinear. In the following statement of the Fundamental Theorem we express this constraint by requiring that each of the four sets of Points lie at the vertices of some quadrilateral, where a *quadrilateral* is defined as follows. A *quadrilateral* is a set of four Points A, B, C and D (no three of which are collinear), together with the Lines AB, BC, CD and DA.

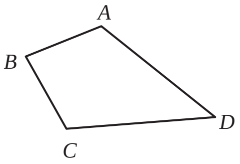

Theorem 3 The Fundamental Theorem of Projective Geometry
Let $ABCD$ and $A'B'C'D'$ be two quadrilaterals in $\mathbb{RP}^2$. Then:

(a) there is a projective transformation t which maps

$$A \text{ to } A',\ B \text{ to } B',\ C \text{ to } C',\ D \text{ to } D';$$

(b) the projective transformation t is unique.

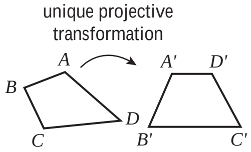

Proof According to the strategy above, there is a projective transformation t_1 which maps the Points $[1, 0, 0]$, $[0, 1, 0]$, $[0, 0, 1]$, $[1, 1, 1]$ to the Points A, B, C, D, respectively. Similarly, there is a projective transformation t_2 which maps the Points $[1, 0, 0]$, $[0, 1, 0]$, $[0, 0, 1]$, $[1, 1, 1]$ to the Points A', B', C', D', respectively.

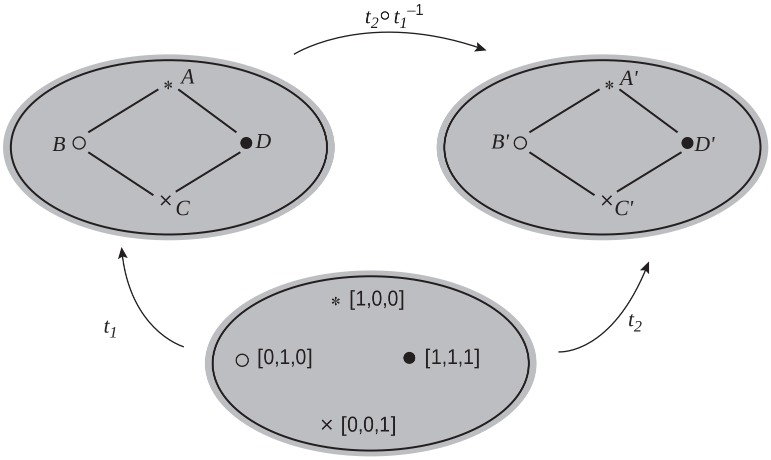

(a) The composite $t = t_2 \circ t_1^{-1}$ is then a projective transformation which maps A to A', B to B', C to C', D to D'.

(b) To check uniqueness of t, we first check that the identity transformation is the only projective transformation which maps each of the Points [1, 0, 0], [0, 1, 0], [0, 0, 1], [1, 1, 1] to themselves. In fact any projective transformation with this property must have an associated matrix which is some non-zero multiple of the matrix

$$\begin{pmatrix} u & 0 & 0 \\ 0 & v & 0 \\ 0 & 0 & w \end{pmatrix}, \quad \text{where} \quad \begin{pmatrix} u & 0 & 0 \\ 0 & v & 0 \\ 0 & 0 & w \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Such a matrix must be (a non-zero multiple of) the identity matrix, and so the transformation must indeed be the identity.

This follows from the discussion leading to the strategy above.

Next suppose that t and t' are two projective transformations which satisfy the conditions of the theorem. Then the composites $t_2^{-1} \circ t \circ t_1$ and $t_2^{-1} \circ t' \circ t_1$ must both be projective transformations which map each of the Points [1, 0, 0], [0, 1, 0], [0, 0, 1], [1, 1, 1] to themselves. Since this implies that both composites are equal to the identity, we deduce that

$$t_2^{-1} \circ t \circ t_1 = t_2^{-1} \circ t' \circ t_1.$$

If we now compose both sides of this equation with t_2 on the left and with t_1^{-1} on the right, then we obtain $t = t'$, as required.

The Fundamental Theorem tells us that there is a projective transformation which maps any given quadrilateral onto any other given quadrilateral. So we have the following corollary.

Corollary All quadrilaterals are projective-congruent.

By *projective-congruent* we mean that there is a projective transformation that maps any quadrilateral onto any other quadrilateral.

If we actually need to find the projective transformation which maps one given quadrilateral onto another given quadrilateral, we simply follow the strategy used to prove part (a) of the Fundamental Theorem.

Strategy To determine the projective transformation t which maps the vertices of the quadrilateral $ABCD$ to the corresponding vertices of the quadrilateral $A\ B\ C\ D$:

1. find the projective transformation t_1 which maps the triangle of reference and unit Point to the Points A, B, C, D, respectively;
2. find the projective transformation t_2 which maps the triangle of reference and unit Point to the Points $A\ , B\ , C\ , D$, respectively;
3. calculate $t = t_2 \quad t_1^{\ \ 1}$.

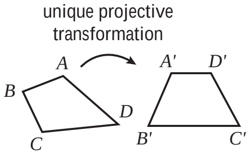

Example 6 Find the projective transformation t which maps the Points [1, 1, 2], [1, 2, 1], [5, 1, 2], [1, 0, 1] to the Points [1, 3, 2], [3, 7, 5], [2, 5, 4], [3, 8, 5], respectively.

Solution We follow the steps in the above strategy.

(a) Any matrix associated with the projective transformation t_1 which maps the Points [1, 0, 0], [0, 1, 0], [0, 0, 1], [1, 1, 1] to the Points [1, 1, 2], [1, 2, 1], [5, 1, 2], [1, 0, 1], respectively, must be a multiple of the matrix

$$\begin{matrix} u & v & 5w \\ u & 2v & w \\ 2u & v & 2w \end{matrix} \ , \quad \text{where} \quad \begin{matrix} u & v & 5w \\ u & 2v & w \\ 2u & v & 2w \end{matrix} \quad \begin{matrix} 1 \\ 1 \\ 1 \end{matrix} \ = \ \begin{matrix} 1 \\ 0 \\ 1 \end{matrix} \ .$$

Solving the equations

$$u + v + 5w = 1,$$
$$u \quad 2v \quad w = 0,$$
$$2u + v + 2w = 1,$$

we obtain $u = \frac{1}{2}, v = \quad \frac{1}{3}, w = \frac{1}{6}$. So a suitable choice of matrix for t_1 is

$$\begin{matrix} \frac{1}{2} & \frac{1}{3} & \frac{5}{6} \\ \frac{1}{2} & \frac{2}{3} & \frac{1}{6} \\ 1 & \frac{1}{3} & \frac{1}{3} \end{matrix} \ , \quad \text{or more simply } \mathbf{A}_1 = \begin{matrix} 3 & 2 & 5 \\ 3 & 4 & 1 \\ 6 & 2 & 2 \end{matrix} \ .$$

It is simpler to multiply the first matrix by 6 to obtain integer entries. This does not alter the projective transformation t_1 with which the matrix is associated.

(b) Any matrix associated with the projective transformation t_2 which maps the Points [1, 0, 0], [0, 1, 0], [0, 0, 1], [1, 1, 1] to the Points [1, 3, 2], [3, 7, 5], [2, 5, 4], [3, 8, 5], respectively, must be a multiple of the matrix

$$\begin{matrix} u & 3v & 2w \\ 3u & 7v & 5w \\ 2u & 5v & 4w \end{matrix} \ , \ \text{where} \quad \begin{matrix} u & 3v & 2w \\ 3u & 7v & 5w \\ 2u & 5v & 4w \end{matrix} \quad \begin{matrix} 1 \\ 1 \\ 1 \end{matrix} \ = \ \begin{matrix} 3 \\ 8 \\ 5 \end{matrix} \ .$$

Solving the equations

$$\begin{aligned} u \quad 3v + 2w &= \quad 3,\\ 3u + 7v \quad 5w &= 8,\\ 2u \quad 5v + 4w &= \quad 5, \end{aligned}$$

we obtain $u = 2, v = 1, w = 1$. So a suitable choice of matrix for t_2 is

$$\mathbf{A}_2 = \begin{matrix} 2 & 3 & 2\\ 6 & 7 & 5\\ 4 & 5 & 4 \end{matrix}.$$

(c) A matrix associated with the inverse, $t_1^{\ 1}$, of t_1 is $\mathbf{A}_1^{\ 1}$, which we can calculate to be

$$\mathbf{A}_1^{\ 1} = \begin{matrix} \frac{1}{12} & \frac{1}{12} & \frac{1}{4}\\ 0 & \frac{1}{3} & \frac{1}{6}\\ \frac{1}{4} & \frac{1}{12} & \frac{1}{12} \end{matrix};$$

then a simpler matrix associated with $t_1^{\ 1}$ is

$$\mathbf{B} = \begin{matrix} 1 & 1 & 3\\ 0 & 4 & 2\\ 3 & 1 & 1 \end{matrix}.$$

The required projective transformation is therefore $t : [\mathbf{x}] \quad [\mathbf{A}x]$, where

$$\begin{aligned} \mathbf{A} = \mathbf{A}_2\mathbf{B} &= \begin{matrix} 2 & 3 & 2\\ 6 & 7 & 5\\ 4 & 5 & 4 \end{matrix} \quad \begin{matrix} 1 & 1 & 3\\ 0 & 4 & 2\\ 3 & 1 & 1 \end{matrix}\\ &= \begin{matrix} 8 & 12 & 14\\ 21 & 29 & 37\\ 16 & 20 & 26 \end{matrix}. \end{aligned}$$

□

Problem 8 Find the projective transformation t that maps the Points [1, 0, 0], [3, 2, 0], [2, 0, 4], [1, 2, 5] to the Points [2, 1, 0], [1, 0, 1], [0, 3, 1], [3, 1, 2], respectively.

Problem 9 Find the projective transformation t that maps the Points [1, 0, 3], [1, 1, 2], [3, 3, 5], [6, 4, 13] to the Points [3, 5, 3], [$\frac{1}{2}$, 1, 0], [3, 5, 6], [8, 13, 12], respectively.

3.3.4 Geometrical Interpretation of Projective Transformations

You may omit this subsection at a first reading, as it is quite hard going.

In this subsection we discuss the relationship between projective transformations and the perspectivities introduced in Section 3.1.

Starting from the geometric definition of perspective in Subsection 3.1.1, we will define the term *perspective transformation*, show how a perspective transformation may be interpreted as a projective transformation, and finally prove that any projective transformation can be expressed as the composite of at most three perspective transformations.

Recall that a *projective transformation* of $\mathbb{R}P^2$ is a map $[\mathbf{x}] \quad [\mathbf{Ax}]$, where $\mathbf{A}$ is an invertible 3×3 matrix.

So, let and be two embedding planes in $\mathbb{R}^3$ that do not pass through the origin O in $\mathbb{R}^3$, and let C $(= \; O)$ be another point in $\mathbb{R}^3$ such that OC is not parallel to either or . Let C have position vector $\mathbf{c}$ (based at O). Also, let denote an arbitrary perspectivity from the point C that maps the plane to the plane .

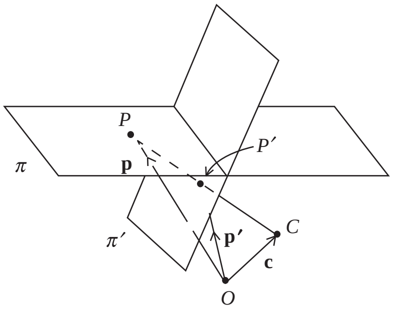

Note that O, C, P and P are coplanar.

Now, the perspectivity will map any point P (with position vector $\mathbf{p}$) in onto some point P (with position vector $\mathbf{p}$) in , so long as the vector $\mathbf{p} \quad \mathbf{c}$ is not parallel to the plane . We then define the perspective transformation associated with to be the mapping of $\mathbb{R}^3$ to itself that maps the line $[\mathbf{p} \quad \mathbf{c}]$ onto the line $[\mathbf{p} \quad \mathbf{c}]$. But since C, P and P are collinear, it follows that the vectors $\mathbf{p} \quad \mathbf{c}$ and $\mathbf{p} \quad \mathbf{c}$ must be parallel (equivalently, that $[\mathbf{p} \quad \mathbf{c}] = [\mathbf{p} \quad \mathbf{c}]$); hence there is some real number t such that

$$\mathbf{p} \quad \mathbf{c} = t(\mathbf{p} \quad \mathbf{c}).$$

Recall that a line through the origin in $\mathbb{R}^2$ is a 1-dimensional vector space.

The value of t will depend on the particular point P under discussion.

We can then rewrite this formula in the form

$$\mathbf{p} \; = t(\mathbf{p} \quad \mathbf{c}) + \mathbf{c},$$

or

$$\mathbf{p} \; = t\mathbf{p} + (1 \quad t)\mathbf{c},$$

so that

$$[\mathbf{p}] \qquad [t\mathbf{p} + (1 \quad t)\mathbf{c}].$$

In this way, the perspectivity gives a one-one mapping from onto , except that it is not defined on the line in where cuts the plane through C parallel to ; also, there are no points of that map onto points of on the line where cuts the plane through C parallel to . We use the fact that the ideal Points for a plane correspond to the directions of lines in the plane, rather than actual points in the plane, to extend our definition of the map.

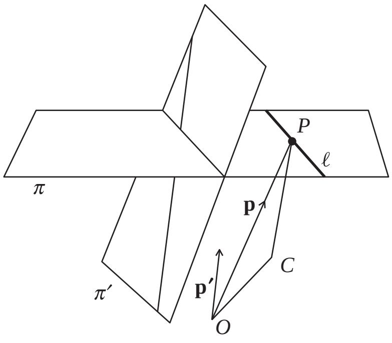

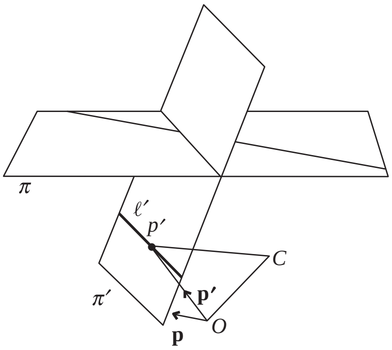

So, first, let P be a point of that lies on the line ; then the points O, C, P are not collinear, since OC is not parallel to the plane . Denote the position vector OP by $\mathbf{p}$. Let the plane through O, C and P meet the plane in a line, and let $\mathbf{p}$ be a position vector based at O that is parallel to this line. Then we specify that our (extended) map maps the line $[\mathbf{p}]$ through O onto the line $[\mathbf{p}\ ]$ through O.

Similarly, let P be a point of that lies on the line ; then the points O, C, P are not collinear, since OC is not parallel to the plane . Denote the position vector OP by $\mathbf{p}$. Let the plane through O, C and P meet the plane in a line, and let $\mathbf{p}$ be a position vector based at O that is parallel to this line. Then we specify that our (further extended) map maps the line $[\mathbf{p}]$ through O onto the line $[\mathbf{p}\ ]$ through O.

Finally, we specify that the extended map maps the line through O that is parallel to the line of intersection of and onto itself.

In this way, we have constructed a transformation that is a one-one mapping of $\{$the ideal Points for $\}$ onto the ideal Points for associated in a natural way with the given perspectivity , and we call it the associated **perspective transformation**. This maps the family of Euclidean lines through O onto itself, in other words RP^2 onto itself.

We now explain why we can think of this perspective transformation as a projective transformation.

First, we consider a perspectivity in R^2 as this will prove useful later in our discussion. So, let and be two lines in R^2 that do not pass through the origin O in R^2, let L be the common point of the two lines, and let $C(= O)$ be another point in R^2 such that OC is not parallel to either or . Consider any perspectivity , with centre some point C, say, that maps onto .

In our discussion we shall omit discussion of the exceptional points , for simplicity.

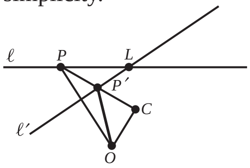

We are interested in the map that sends lines through O to lines through O that is obtained from as follows. Let map the point P on onto the point P on . Then we define $(OP) = OP$.

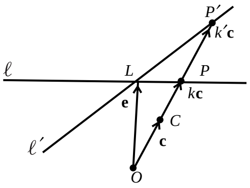

It is clear that $(OL) = OL$, because L is fixed by .

It is also clear that $(OC) = OC$, because if the line OC meets at P and the line at P then O, C, P and P are collinear and so $(OC) = OC$.

We therefore choose to take as basis vectors in R^2 the vector $\mathbf{e} = OL$ and the vector $\mathbf{c} = OC$.

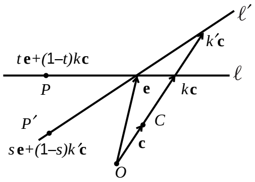

We now find the effect of the map on a line through the origin O. We shall suppose that the line OC meets the line at the point with position vector $k\mathbf{c}$ and the line at the point with position vector $k\ \mathbf{c}$. Any point P on the line then has position vector $t\mathbf{e} + (1\quad t)k\mathbf{c}$, for some real number t, and any point P on the line has position vector $s\mathbf{e} + (1\quad s)k\ \mathbf{c}$, for some real number s.

Next, the line OP consists of the points with coordinates $u(\mathbf{e} + m\mathbf{c}) = u\mathbf{e} + mu\mathbf{c}$ for a fixed value of m and varying values of u. This gives two (equivalent) expressions for the position of the point P, namely $t\mathbf{e}+(1\quad t)k\mathbf{c}$ and $u\mathbf{e}+mu\mathbf{c}$. It follows that we must have

$$u = t \quad \text{and} \quad mu = (1\quad t)k.$$

Dividing the second equation by the first, we get

$$m = \frac{1\quad t}{t}k$$

so that

$$tm = k\quad kt,$$

which yields the formula

$$t = \frac{k}{m + k}.$$

Similarly, the line OP consisting of points $u\ (\mathbf{e} + m\ \mathbf{c}) = u\ \mathbf{e} + m\ u\ \mathbf{c}$, for a fixed value of m and varying values of u , meets the line at the point P where

$$s = \frac{k}{m\ + k}.$$

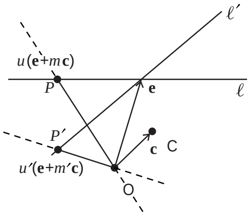

Now, we have $(OP) = OP$ if and only if the points C, P and P are collinear; that is, if and only if there is a real number r such that

$$r(t\mathbf{e} + (1\quad t)k\mathbf{c}\quad \mathbf{c}) = s\mathbf{e} + (1\quad s)k\ \mathbf{c}\quad \mathbf{c}.$$

This is the case if and only if

$$rt = s \quad \text{and} \quad r((1\quad t)k\quad 1) = (1\quad s)k\quad 1.$$

We can eliminate r by dividing the second equation by the first equation, so obtaining

$$\frac{(1\quad t)k\quad 1}{t} = \frac{(1\quad s)k\quad 1}{s}.$$

Then, if we substitute $t = \frac{k}{m+k}$ and $s = \frac{k}{m\ +k}$ into this equation, after some manipulation we obtain the remarkable result that

$$m\ = m\frac{(k\quad 1)k}{(k\quad 1)k}.$$

It follows that, in terms of the basis elements **e** and **c**, the map of the family of lines through O to itself given by

$$\begin{matrix} 1 \\ m \end{matrix} \qquad \begin{matrix} 1 \\ m \end{matrix}$$

can be represented by the matrix

$$\begin{matrix} 1 & 0 \\ 0 & \frac{(k \quad 1)k}{(k \quad 1)k} \end{matrix} .$$

For convenience, we write this matrix in the form

$$\begin{matrix} 1 & 0 \\ 0 & r \end{matrix} ,$$

where r is a fixed number that depends only on the geometry of the two lines and and on our choice of the point C. Furthermore, $r = 0$ since $k = 1$ (for $k = 1$ implies that C lies on) and $k = 0$ (for $k = 0$ implies that O lies on).

Now we consider the situation in R^3, in relation to RP^2. Consider an arbitrary perspectivity , with centre some point C, say, that maps a plane onto a plane , with neither O nor C lying on or . As before, we shall suppose that the line OC meets the plane at a point with position vector $k\mathbf{c}$ and the plane at a point with position vector $k\ \mathbf{c}$.

Once again, we are interested in the map that sends lines (in R^3) through O to lines through O that is obtained from as follows. Let map a point P in to a point P in . Then we define $(OP) = OP$.

Let E_1 and E_2 be any two points on the common line of the two planes. Then, clearly, $(OE_1) = OE_1$ and $(OE_2) = OE_2$ because every point of is fixed by .

It is also clear that $(OC) = OC$, because if the line OC meets the plane at Q and the plane at Q , then O, C, Q and Q are collinear, and so $(OC) = OC$.

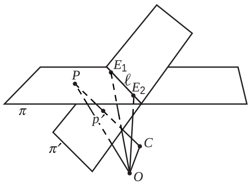

We therefore now take as basis vectors in R^3 the vectors $\mathbf{e}_1 = OE_1$, $\mathbf{e}_2 = OE_2$, and $\mathbf{c} = OC$.

We now find the effect of on a line in R^3 through the origin O. We can simplify our task by observing that the lines OC, OP and OP all lie in a plane, say; let this plane meet the line common to the given planes and at a point E. We then define $\mathbf{e} = OE$.

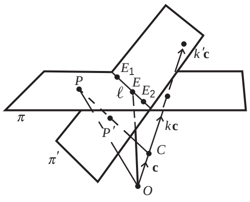

We can now apply our earlier discussion of the planar case to the restriction of the mapping to the plane . If we denote the lines OP and OP by the parametrizations

$$OP = u(\mathbf{e} + m\mathbf{c}) = u\mathbf{e} + um\mathbf{c}, \quad \text{where } m \text{ is fixed and } u \text{ varies,}$$

$$OP \ = u\ (\mathbf{e} + m\ \mathbf{c}) = u\ \mathbf{e} + u\ m\ \mathbf{c}, \quad \text{where } m \text{ is fixed and } u \text{ varies,}$$

it follows from our earlier discussion that m can be expressed in the form rm, where r is a fixed number that depends only on the geometry of the two planes and and on our choice of the point C. Hence we can rewrite the parametrization of OP as

$$OP = u (\mathbf{e} + rm\mathbf{c}) = u \mathbf{e} + u rm\mathbf{c}, \quad \text{where } r, m \text{ are fixed and } u \text{ varies.}$$

Next, we can express the position vector $\mathbf{e}$ of the point E in the form

$$\mathbf{e} = t\mathbf{e}_1 + (1 \quad t)\mathbf{e}_2, \quad \text{for some number } t.$$

Then we have that the mapping maps the line

$$OP = u(t\mathbf{e}_1 + (1 \quad t)\mathbf{e}_2 + m\mathbf{c}) = ut\mathbf{e}_1 + u(1 \quad t)\mathbf{e}_2 + um\mathbf{c}$$

onto the line

$$OP = u (t\mathbf{e}_1 + (1 \quad t)\mathbf{e}_2 + rm\mathbf{c}) = u t\mathbf{e}_1 + u (1 \quad t)\mathbf{e}_2 + u rm\mathbf{c}.$$

It follows analogously to the two-dimensional situation that, in terms of the basis elements $\mathbf{e}_1, \mathbf{e}_2, \mathbf{c}$, the mapping of the family of lines in R^3 to itself can be represented by the matrix

$$\begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & r \end{matrix},$$

Notice that this maps the line represented by the points $[t, 1 \quad t, 0]$ to itself pointwise, as it should.

where the (non-zero) constant r depends only on the position of the planes and . Since this is an invertible 3 × 3 matrix, this is our required description of the perspective transformation as a projective transformation.

We now go the other way round, and obtain a projective transformation as a sequence of three perspective transformations. We have a lot of freedom, because we can choose the centres of perspectivities and the planes.

Suppose we are given a projective transformation and two embedding planes and . Let $[\mathbf{a}], [\mathbf{b}], [\mathbf{c}], [\mathbf{d}]$ be any four non-collinear Points, and $[\mathbf{a}], [\mathbf{b}], [\mathbf{c}], [\mathbf{d}]$ be any other four non-collinear Points; we can represent these Points by (Euclidean) points A, B, C, D in and A , B , C , D in , respectively.

We will use in our discussion the existence and uniqueness parts of the Fundamental Theorem of Projective Geometry: namely, that there is one and only one projective transformation mapping four non-collinear Points to any other four non-collinear Points. This means that if we can find a composite of three perspective transformations that maps $[\mathbf{a}], [\mathbf{b}], [\mathbf{c}], [\mathbf{d}]$ to $[\mathbf{a}], [\mathbf{b}], [\mathbf{c}], [\mathbf{d}]$, respectively, then this composite must be a projective transformation (by the existence part) and it must equal the given projective transformation (by the uniqueness part).

Subsection 3.3.3, Theorem 3

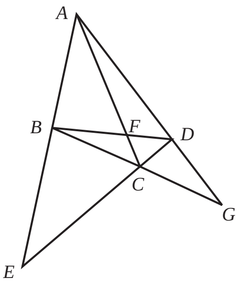

We now exhibit a sequence of three perspectivities the composite of which maps A, B, C, D in to A , B , C , D in . In our discussion it is convenient to let AB and CD meet at E, AC and BD meet at F, and AD and BC meet at G, with analogous definitions of E , F and G . (You may find the figures below helpful to follow through the argument; though, for simplicity, we have omitted the initial plane and the points F , F and F .)

1. The first perspectivity is from the plane to a plane that passes through A . The centre of this perspectivity is an arbitrary point P_1 on the line AA (if $A = A$, then P_1 can be chosen to lie anywhere not on or). This perspectivity maps A to A and B, C, D, E, F, G to B , C , D , E , F , G , say, respectively. We can assume, by suitably varying , that $B B$ and $E E$ are not parallel.

A A

B B , C C ,
D D , E E ,
F F , G G

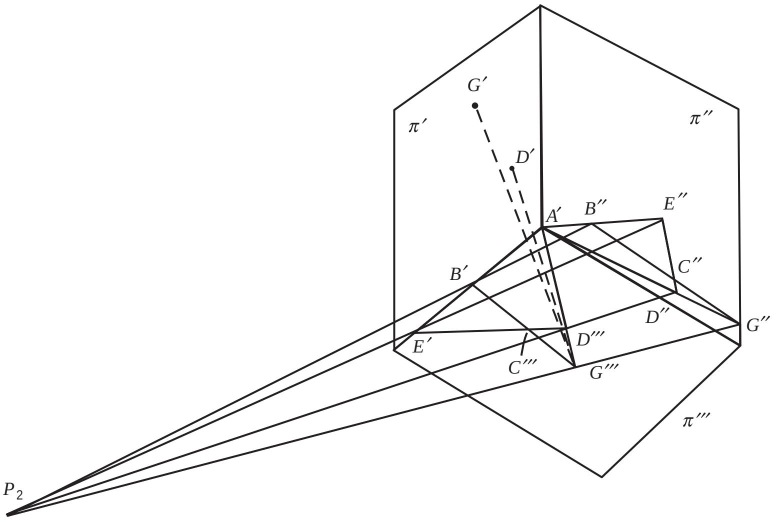

2. Now, by the definition of E the points A , B , E lie on a line through A , and since the points A, B, E are collinear the points A , B , E lie on a line through A ; so these five points lie in a plane, and because the lines $B B$ and $E E$ are not parallel they meet in a point, P_2 say. We then pick a plane through the line $A B E$ and map the plane onto by the perspectivity with centre P_2. This sends A , B , E to A , B , E , respectively, and the points C , D , F , G to, say, C , D , F , G , respectively. As before, we can assume that $D D$ and $G G$ are not parallel.

A fixed;
B B , E E

C C , D D ,
F F , G G

3. Now, since A, D, G are collinear, the points A , D , G lie on a line through A and the points A , D , G lie on a line through A , so these five points lie in a plane; and because the lines $D D$ and $G G$ are not parallel they meet in a point, P_3 say. We then map the plane onto the plane by the perspectivity with centre P_3. This map sends A , B , D , E , G to A , B , D , E , G , respectively.

   This third perspectivity sends the line $B G$ to the line $B G$ and the line $D E$ to the line $D E$, so it maps the point C to the point C .

Refer here to the figure below.

A , B , E fixed

D D , G G

Thus the composite of these three perspectivities maps the points A, B, C, D in to A , B , C , D in , as required.

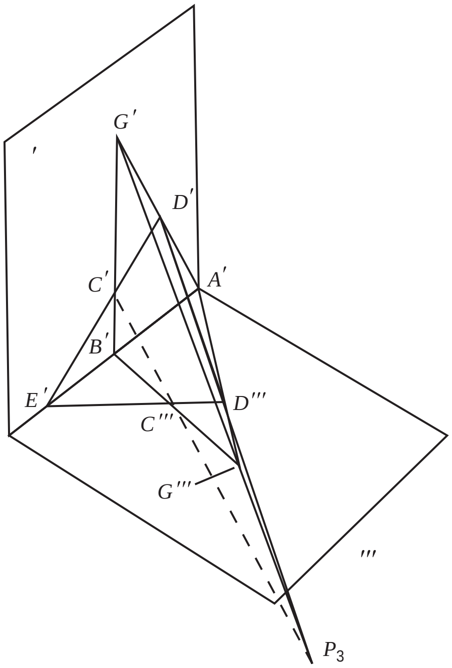

Theorem 4 Perspectivity Theorem
Every projective transformation can be expressed as the composite of three perspective transformations.

3.4 Using the Fundamental Theorem of Projective Geometry

In Section 3.2 we described how an embedding plane can be used to represent projective space $\mathbb{R}P^2$. The Points of $\mathbb{R}P^2$ are represented by Euclidean points in and the Lines of $\mathbb{R}P^2$ are represented by Euclidean lines in .

In general, any Euclidean figure in an embedding plane corresponds to a projective figure in $\mathbb{R}P^2$, and visa versa. This correspondence enables us to compare Euclidean theorems about a figure in an embedding plane with projective theorems about the corresponding projective figure. Provided that the theorems are concerned exclusively with projective properties, such as collinearity and incidence, then a Euclidean theorem will hold if and only if the corresponding projective theorem holds.

The Euclidean figure may have some ideal Points attached to it.

3.4.1 Desargues Theorem and Pappus Theorem

The advantage of interpreting a Euclidean theorem as a projective theorem in this way is that we can often obtain a much simpler proof of the theorem than would be possible using Euclidean geometry directly. We illustrate this by using projective geometry to prove the theorem of Desargues.

We introduced Desargues Theorem in Subsection 3.1.3.

Theorem 1 Desargues' Theorem
Let ABC and $A\ B\ C$ be triangles in $\mathbb{R}^2$ such that the lines AA , BB and CC meet at a point U. Let BC and $B\ C$ meet at P, CA and $C\ A$ meet at Q, and AB and $A\ B$ meet at R. Then P, Q and R are collinear.

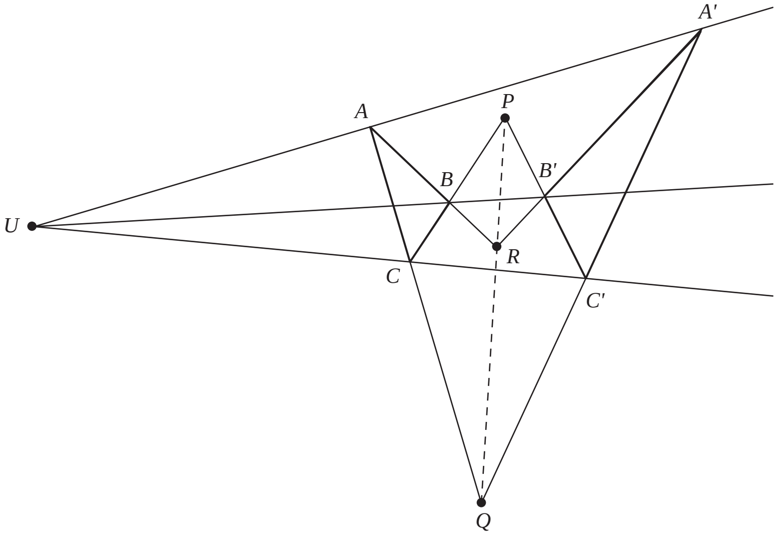

Proof Because this theorem is concerned exclusively with the projective properties of collinearity and incidence we can interpret it as a projective theorem in $\mathbb{RP}^2$. Moreover, by the Fundamental Theorem of Projective Geometry we know that any configuration of the theorem is projective-congruent to a configuration of the theorem in which $A = [1, 0, 0]$, $B = [0, 1, 0]$, $C = [0, 0, 1]$ and $U = [1, 1, 1]$. If we can prove the theorem in this special case then we can use the fact that projective-congruence preserves projective properties to deduce that the theorem holds in general.

To prove the special case we use the algebraic techniques described in Section 3.2. First observe that the Line AU passes through the Points [1, 0, 0] and [1, 1, 1], and therefore has equation $y = z$. Since A is a Point on AU, it must have homogeneous coordinates of the form $[a, b, b]$, for some real numbers a and b. Now, $b = 0$, since $A = A$; so we may write the homogeneous coordinates of A in the form $[p, 1, 1]$ (where $p = a/b$).

For $[a, 0, 0] = [1, 0, 0]$.

Similarly, the homogeneous coordinates of the Points B and C may be written in the form $[1, q, 1]$ and $[1, 1, r]$, respectively, for some real numbers q and r.

We omit the details of the calculations.

We now find the Point P where BC and $B\ C$ intersect. The Line BC has equation $x = 0$. Since the Line $B\ C$ passes through the Points $B = [1, q, 1]$ and $C = [1, 1, r]$, it must have equation

$$\begin{matrix} x & y & z \\ 1 & q & 1 \\ 1 & 1 & r \end{matrix} = 0,$$

which we may rewrite in the form

$$(qr - 1)x - (r - 1)y + (1 - q)z = 0.$$

It follows that at the Point P of intersection of the Lines BC and $B'C'$ we must have $x = 0$ and $(r - 1)y = (1 - q)z$, so that P has homogeneous coordinates $[0, 1 - q, r - 1]$.

Similarly, the Points Q and R have homogeneous coordinates $[1 - p, 0, r - 1]$ and $[1 - p, q - 1, 0]$, respectively.

We omit the details of the calculations.

Now, the Points P, Q and R are collinear if

$$\begin{vmatrix} 0 & 1-q & r-1 \\ 1-p & 0 & r-1 \\ 1-p & q-1 & 0 \end{vmatrix} = 0.$$

But

$$\begin{aligned} &\begin{vmatrix} 0 & 1-q & r-1 \\ 1-p & 0 & r-1 \\ 1-p & q-1 & 0 \end{vmatrix} \\ &= -(1-q)\begin{vmatrix} 1-p & r-1 \\ 1-p & 0 \end{vmatrix} + (r-1)\begin{vmatrix} 1-p & 0 \\ 1-p & q-1 \end{vmatrix} \\ &= -(1-q)(1-p)(1-r) + (r-1)(1-p)(q-1) \\ &= 0. \end{aligned}$$

It follows that P, Q and R are collinear, as asserted. The general result now holds, by projective-congruence.

When using the Fundamental Theorem to simplify proofs of results in projective geometry, we do not usually refer to projective-congruence. Instead, *so long as the properties involved are projective properties*, we content ourselves with an initial remark of the type: ‘By the Fundamental Theorem of Projective Geometry, we may choose the four Points…, no three of which are collinear, to be the triangle of reference and the unit Point; that is, to have homogeneous coordinates [1, 0, 0], [0, 1, 0], [0, 0, 1] and [1, 1, 1], respectively’.

Problem 1 Let $\triangle ABC$ be a triangle in $\mathbb{R}^2$, and let U be any point of $\mathbb{R}^2$ that is not collinear with any two of the points A, B and C. Let the lines AU, BU and CU meet the lines BC, CA and AB at the points A', B' and C', respectively. Next, let the lines BC and $B'C'$ meet at P, AC and $A'C'$ meet at Q, and AB and $A'B'$ meet at R. Prove that P, Q and R are collinear.

Hint: Let A, B, C be the vertices of the triangle of reference, and let U be the unit Point. Then determine the homogeneous coordinates of the Points A', B' and C'.

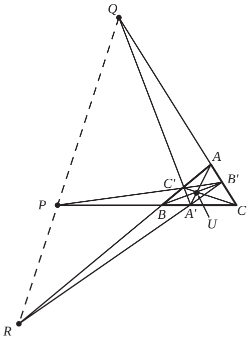

Next we use the Fundamental Theorem of Projective Geometry to prove Pappus Theorem.

Theorem 2 Pappus’ Theorem
Let A, B and C be three points on a line in $\mathbb{R}^2$, and let A , B and C be three points on another line. Let BC and $B C$ meet at P, CA and $C A$ meet at Q, and AB and $A B$ meet at R. Then P, Q, R are collinear.

This theorem is named after Pappus, a Greek mathematician who discovered it in the 3rd century AD.

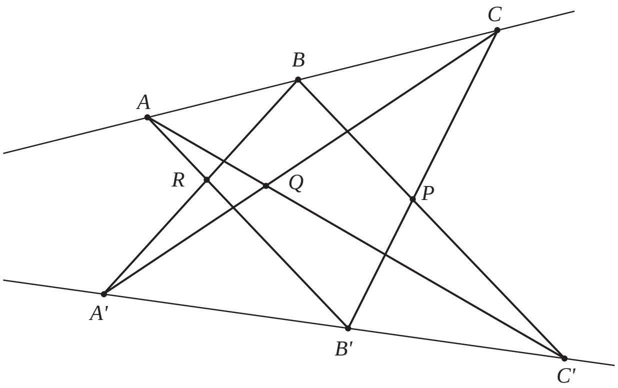

Proof We interpret the theorem as a projective theorem, so: by the Fundamental Theorem of Projective Geometry we may choose the four Points A, A , P, R, no three of which are collinear, to be the triangle of reference and the unit Point; that is, to have homogeneous coordinates [1, 0, 0], [0, 1, 0], [0, 0, 1] and [1, 1, 1], respectively.

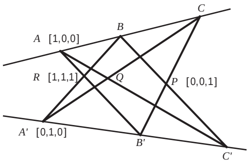

First observe that the Line AR passes through the Points [1, 0, 0] and [1, 1, 1], and must therefore have equation $y = z$. Since B is a Point on AR, it must have homogeneous coordinates of the form $[a, b, b]$ for some real numbers a and b. Now, $b = 0$ since $A = B$, so we may write the homogeneous coordinates of B in the form $[r, 1, 1]$ (where $r = a/b$).

Similarly, the Point B lies on the Line $x = z$ through the Points $A = [0, 1, 0]$ and $R = [1, 1, 1]$, so it must have homogeneous coordinates of the form $[1, s, 1]$.

Next we find the Point C where the Line AB intersects the Line $B P$. Since the Line AB passes through the Points $A = [1, 0, 0]$ and $B = [1, s, 1]$, it must have equation $y = sz$. Also since the Line $B P$ passes through the Points $B = [r, 1, 1]$ and $P = [0, 0, 1]$ it must have equation $x = ry$. At the Point C where AB meets $B P$ we have $y = sz$ and $x = ry$, so $C = [rs, s, 1]$.

Similarly, C is the point where the Line BP intersects the Line $A B$. Since $B = [1, s, 1]$ and $P = [0, 0, 1]$, BP has equation $y = sx$; and, since $A = [0, 1, 0]$ and $B = [r, 1, 1]$, $A B$ has equation $x = rz$. It follows that $C = [r, rs, 1]$.

Finally we find the point Q where AC intersects $A C$. Since the Line AC passes through the Points $A = [1, 0, 0]$ and $C = [r, rs, 1]$ it must have equation $y = rsz$. Also the Line $A C$ passes through the Points $A = [0, 1, 0]$ and $C = [rs, s, 1]$ so it must have equation $x = rsz$. At the Point Q where AC intersects A C we have $y = rsz$ and $x = rsz$, so $Q = [rs, rs, 1]$.

To complete the proof we simply observe that the Points $R = [1, 1, 1]$, $Q = [rs, rs, 1]$ and $P = [0, 0, 1]$ all lie on the Line $x = y$. It follows that P, Q and R are collinear.

A fortunate choice of Points for the triangle of reference and unit Point meant that we did not have to use the determinant criterion for collinearity at the final stage of the argument.

Although we can sometimes simplify the proof of a Euclidean theorem by using projective geometry, there is another more subtle reason for interpreting a Euclidean theorem as a projective theorem. By doing so we can often avoid having to make special provision for exceptional cases, such as when two lines are parallel. In projective geometry, Lines which correspond to a pair of parallel lines in an embedding plane actually meet and are therefore no different to any other Lines.

As an example, consider the diagram in the margin. This illustrates the situation that occurs in Pappus Theorem when the Point of intersection R of $A B$ and AB is an ideal Point for the embedding plane. The above proof of Pappus Theorem is able to cope with this situation because it uses arguments from $\mathbb{RP}^2$! Our interpretation of the theorem on an embedding plane in this situation is that the Points P and Q must be collinear with the ideal Point R at which $A B$ and AB meet. That is, PQ must be parallel to both $A B$ and AB .

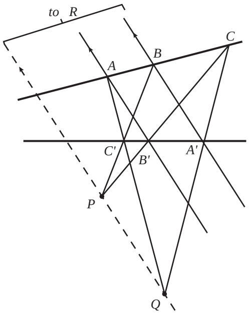

Problem 2 Give a Euclidean interpretation of Desargues Theorem on an embedding plane in the case where Q is an ideal Point for .

3.4.2 Duality

Recall that two key projective properties that we have met so far have a certain symmetry between them.

Collinearity Property	Incidence Property
Any two distinct *Points* lie on a unique *Line.*	Any two distinct *Lines* meet in a unique *Point.*

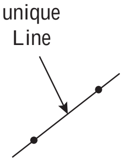

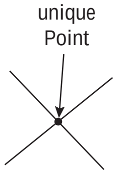

We can obtain one property from the other simply by interchanging the words Point and Line , and making whatever other changes are needed to ensure that the sentence makes sense. We say that this interchanging process *dualizes* one statement into the other, and that each statement is the *dual* of the other.

For example, a family of Points on a Line becomes a family of Lines through a Point under dualization. Similarly, a triangle or a family of three non-collinear Points and the three Lines joining them dualizes to a family of three non-concurrent Lines and the three Points where they meet , which is again a triangle. Since a triangle is thus dual to a triangle, we say that triangles are *self-dual* figures.

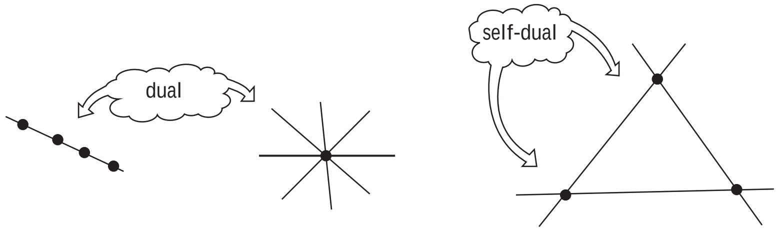

The dualization process is particularly interesting when applied to theorems. We shall illustrate this in the context of Pappus' Theorem. In order to do this, it is helpful to rephrase Pappus' Theorem using the term *hexagon.* As you would expect, a **hexagon** in $\mathbb{RP}^2$ consists of six Points joined by six Lines. The figure below illustrates (Euclidean) hexagons in an embedding plane; the corresponding hexagons in $\mathbb{RP}^2$ are the corresponding six Points and six Lines—that we model as six lines and six planes in $\mathbb{R}^3$.

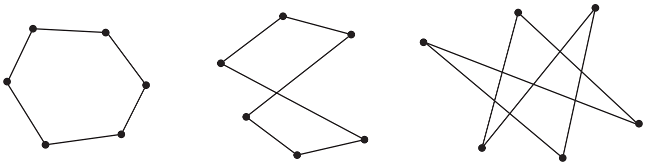

We can now rephrase Pappus' Theorem in the following form.

Theorem 3 Pappus' Theorem (rephrased)
Let the vertices A, B', C, A', B and C' of a hexagon lie alternately on two different Lines. Then the Points of intersection of opposite sides $B'C$ and BC', CA' and $C'A$, AB' and $A'B$, are collinear.

You should compare this formulation with that in Subsection 3.4.1.

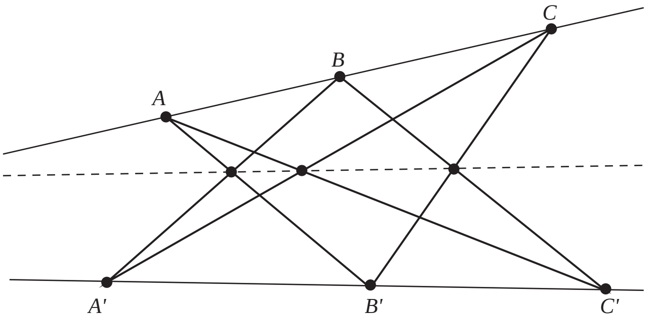

If we dualize this theorem, we obtain the following theorem.

Theorem 4 Brianchon's Theorem
Let the sides AB', $B'C$, CA', $A'B$, BC', $C'A$ of a hexagon pass alternately through two (different) Points P and Q in $\mathbb{RP}^2$. Then the Lines joining opposite vertices A and A′, B and B', C and C', are concurrent.

Charles J. Brianchon (1785–1864) was one of many distinguished French geometers who studied under Gaspard Monge (1746–1818).

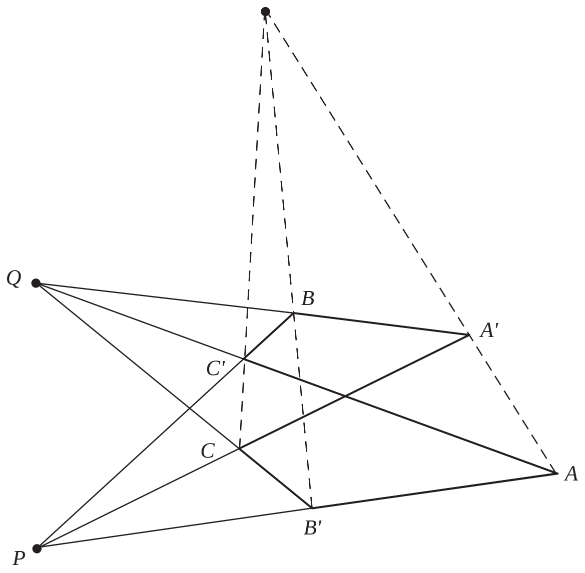

Problem 3 Prove Brianchon's Theorem.

Hint: Let P, C, Q, C be the Points [1, 0, 0], [0, 1, 0], [0, 0, 1], [1, 1, 1], respectively.

It turns out that the dual of any true statement concerning Points, Lines and their projective properties remains true after dualization; that is, if we dualize any theorem in projective geometry, then the statement that we obtain is itself a theorem.

We do not prove this assertion, as it would take us beyond the scope of this book.

Problem 4 Earlier you saw that three Points $[a, b, c]$, $[d, e, f]$, $[g, h, k]$ are collinear if and only if $\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & k \end{vmatrix} = 0$. Write down the dual result of this statement.

Subsection 3.2.2, Strategy

We end this subsection by forming the dual of Desargues' Theorem, as follows.

Desargues' Theorem	Dual Theorem
Let two *triangles* be such that the *Lines* joining corresponding *vertices* meet at a *Point*. Then the *Points* of intersection of the corresponding *sides* of the two *triangles* are *collinear*.	Let two *triangles* be such that the *Points* through which corresponding *sides* pass are *collinear*. Then the *Lines* through the corresponding *vertices* of the two *triangles* are *concurrent*.

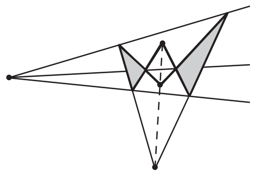

Note that the dual theorem is simply the converse result for Desargues' Theorem! Thus the Principle of Duality enables us to deduce that the converse of Desargues' Theorem holds.

3.5 Cross-Ratio

3.5.1 Another Projective Property

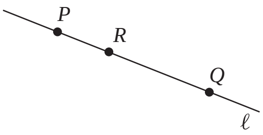

Earlier, in Subsection 2.2.1, we noted that *ratio of lengths along a line* is an affine property. Thus, in affine geometry, if we are given two points P and Q on a line , then we can locate the position of a third point R along by specifying the ratio $PR : RQ$. In particular, it is possible to talk about the point midway between P and Q.

In projective geometry it is meaningless to talk about the Point midway between two other Points. In one embedding plane a Point R may appear to be midway between the Points P and Q, whereas in another embedding plane the ratio $PR : RQ$ may be very different.

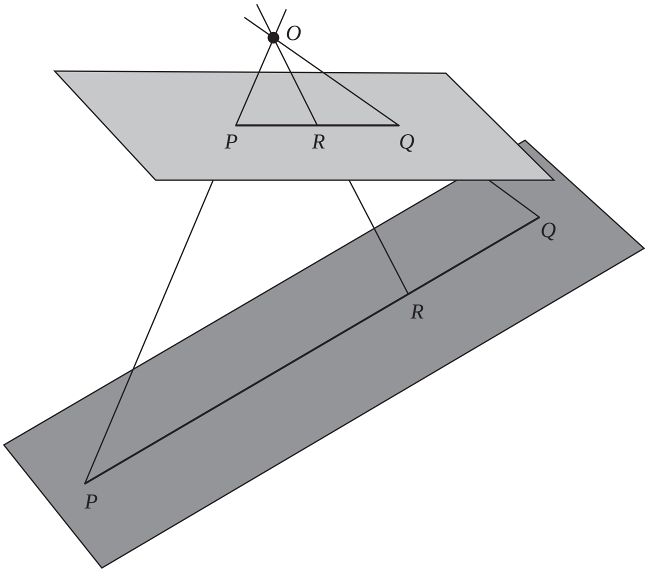

This ambiguity arises from the fact that perspectivities do not preserve the ratio of lengths along a line, so: *ratio of lengths along a line* is not a projective property.

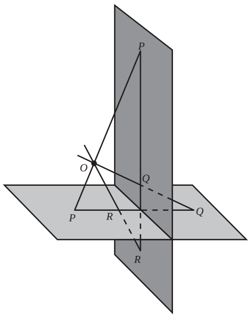

In some embedding planes, such as the plane illustrated in the margin, the Point R does not even appear to lie between P and Q, so *betweenness* is not a projective property either!

Fortunately, there is a quantity, known as *cross-ratio*, that is preserved under all projective transformations. To see how this is defined, consider four collinear Points $A = [\mathbf{a}]$, $B = [\mathbf{b}]$, $C = [\mathbf{c}]$, $D = [\mathbf{d}]$ in RP^2. We can express the fact that A, B, C, D are collinear by writing $\mathbf{c}$ and $\mathbf{d}$ as linear combinations of $\mathbf{a}$ and $\mathbf{b}$. Thus we can write

$$\mathbf{c} = \quad \mathbf{a} + \quad \mathbf{b} \quad \text{and} \quad \mathbf{d} = \quad \mathbf{a} + \quad \mathbf{b},$$

for suitable real numbers , , , .

The cross-ratio is then defined to be the ratio of the ratios − and −.

Definition Let A, B, C, D be four collinear Points in $\mathbb{RP}^2$ represented by position vectors **a**, **b**, **c**, **d**, and let

$$\mathbf{c} = \quad \mathbf{a} + \quad \mathbf{b} \quad \text{and} \quad \mathbf{d} = \quad \mathbf{a} + \quad \mathbf{b}.$$

Then the **cross-ratio** of A, B, C, D is

$$(ABCD) = - \quad -.$$

Equivalently, we can write *(ABCD)* = —.

Of course, before we can be sure that this definition makes sense, we must ensure that it does not depend on the particular choice of position vectors **a**, **b**, **c**, **d** that are used to represent the Points A, B, C, D. We shall check this shortly, but first we illustrate how cross-ratios are calculated.

Example 1 Let $A = [1, 2, 3]$, $B = [1, 1, 2]$, $C = [3, 5, 8]$, $D = [1, \quad 1, 0]$ be Points of $\mathbb{RP}^2$. Calculate the cross-ratio *(ABCD)*.

Solution First, we have to find real numbers and such that the following vector equation holds:

$$(3, 5, 8) = \quad (1, 2, 3) + \quad (1, 1, 2).$$

Note that we have not verified that A, B, C, D are collinear; but if they were not, the equations for , , , could not be solved.

Comparing corresponding coordinates on both sides of this vector equation, we deduce that

$$3 = \quad + \quad, \quad 5 = 2 \quad + \quad \text{and} \quad 8 = 3 \quad + 2 \quad .$$

Solving these equations gives $= 2$, $= 1$.

Next, we find real numbers and such that the vector equation

$$(1, \quad 1, 0) = \quad (1, 2, 3) + \quad (1, 1, 2)$$

holds. Comparing corresponding coordinates on both sides of this vector equation, we deduce that

$$1 = \quad + \quad, \quad 1 = 2 \quad + \quad \text{and} \quad 0 = 3 \quad + 2 \quad .$$

Solving these equations gives $= \quad 2$, $= 3$.

It follows from the definition of cross-ratio that

$$(ABCD) = - \quad - = \frac{1}{2} \quad \frac{3}{2} = \quad \frac{1}{3}.$$

□

Problem 1 Calculate the cross-ratio *(ABCD)* for each of the following sets of collinear Points in $\mathbb{RP}^2$.

(a) $A = [1, \quad 1, \quad 1]$, $B = [1, 3, \quad 2]$, $C = [3, 5, \quad 5]$, $D = [1, \quad 5, 0]$
(b) $A = [1, 2, 3]$, $B = [2, 2, 4]$, $C = [\quad 3, \quad 5, \quad 8]$, $D = [3, \quad 3, 0]$

You may have noticed that the Points A, B, C, D in Problem l (b) are the same as those which appear in Example 1. The only difference is that different homogeneous coordinates are used to represent the Points in each case. As we mentioned after the definition of cross-ratio, the value of the cross-ratio *(ABCD)* does not depend on the homogeneous coordinates that are used to represent A, B, C, D, so it is not surprising that the cross-ratio turned out to have the value $\frac{1}{3}$ in both cases.

Theorem 1 The cross-ratio *(ABCD)* is independent of the homogeneous coordinates that are used to represent the collinear Points A, B, C, D.

Proof Suppose that $A = [\mathbf{a}]$, $B = [\mathbf{b}]$, $C = [\mathbf{c}]$, $D = [\mathbf{d}]$, and let

$$\mathbf{c} = \quad \mathbf{a} + \quad \mathbf{b} \quad \text{and} \quad \mathbf{d} = \quad \mathbf{a} + \quad \mathbf{b}. \tag{1}$$

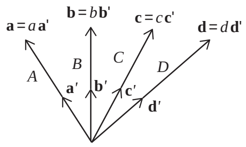

Now suppose that $A = [\mathbf{a}\ ]$, $B = [\mathbf{b}\ ]$, $C = [\mathbf{c}\ ]$, $D = [\mathbf{d}\ ]$. Then

$$\mathbf{a} = a\mathbf{a}\ , \quad \mathbf{b} = b\mathbf{b}\ , \quad \mathbf{c} = c\mathbf{c}\ , \quad \mathbf{d} = d\mathbf{d}\ ,$$

where a, b, c, d are some non-zero real numbers.

By substituting these expressions into the equations (1), we obtain

$$c\mathbf{c} = \quad a\mathbf{a} + \quad b\mathbf{b} \quad \text{and} \quad d\mathbf{d} = \quad a\mathbf{a} + \quad b\mathbf{b}\ ,$$

which we can rewrite in the form

$$\mathbf{c} = \quad \mathbf{a} + \quad \mathbf{b} \quad \text{and} \quad \mathbf{d} = \quad \mathbf{a} + \quad \mathbf{b}\ , \tag{2}$$

where $\quad = a/c, \quad = b/c, \quad = a/d, \quad = b/d.$

We can now check that equations (1) and (2) yield the same value for the cross-ratio:

$$\begin{aligned} \text{—} \quad \text{—} &= \frac{b/c}{a/c} \quad \frac{b/d}{a/d} \\ &= \frac{b}{a} \quad \frac{b}{a} \\ &= \text{–} \quad \text{–}. \end{aligned}$$

So, as expected, the cross-ratio is independent of the choice of homogeneous coordinates.

The next problem illustrates that although the value of the cross-ratio *(ABCD)* is independent of the choice of homogeneous coordinates that are used to represent A, B, C, D, the value of the cross-ratio does depend on the order in which the Points A, B, C, D appear.

Problem 2 Calculate the cross-ratios *(BACD)* and *(ACBD)* for the four Points used in Problem l(a).

When answering Problem 2 you may have noticed that *(BACD)* is the reciprocal of the value which we obtained for *(ABCD)* in Problem 1(a). Also, *(ACBD)* is equal to 1 *(ABCD)*. The next result shows that this is not simply chance!

Note the way in which the Points are changed from their original ordering in the various cross-ratios. We take the reciprocal when swapping the first or last pair of Points, and we subtract from 1 when swapping the inner or outer pair of Points.

Theorem 2 Let A, B, C, D be four distinct collinear Points in $\mathbb{RP}^2$, and let $(ABCD) = k$. Then

$$(BACD) = (ABDC) = 1/k,$$

$$(ACBD) = (DBCA) = 1 \quad k.$$

Proof Let **a**, **b**, **c**, **d** be any position vectors in $\mathbb{R}^3$ in the directions of the Points A, B, C, D, respectively, of $\mathbb{RP}^2$, and let , , , be real numbers such that

$$\mathbf{c} = \ \mathbf{a} + \ \mathbf{b} \quad \text{and} \quad \mathbf{d} = \ \mathbf{a} + \ \mathbf{b}.$$

Then, by definition of cross-ratio, the cross-ratio $(ABCD)$ of the four Points A, B, C, D is the quantity

$$(ABCD) = \frac{\ }{\ } \quad \frac{\ }{\ } = \frac{\quad}{\quad} = k, \text{say}.$$

To determine $(BACD)$, we interchange the roles of A and B in the evaluation of $ABCD$ above; it follows that, since

$$\mathbf{c} = \ \mathbf{b} + \ \mathbf{a} \quad \text{and} \quad \mathbf{d} = \ \mathbf{d} + \ \mathbf{a},$$

the cross-ratio $(BACD)$ is the quantity

$$(BACD) = \frac{\ }{\ } \quad \frac{\ }{\ } = \frac{\quad}{\quad} = \frac{1}{k}.$$

To determine $(ABDC)$, we interchange the roles of C and D in the evaluation of $(ABCD)$ above; it follows that, since

$$\mathbf{d} = \ \mathbf{a} + \ \mathbf{b} \quad \text{and} \quad \mathbf{c} = \ \mathbf{a} + \ \mathbf{b},$$

the cross-ratio $(ABDC)$ is the quantity

$$(ABDC) = \frac{\ }{\ } \quad \frac{\ }{\ } = \frac{\quad}{\quad} = \frac{1}{k}.$$

This completes the first part of the proof.

To evaluate $(ACBD)$, we use the equations

$$\mathbf{c} = \ \mathbf{a} + \ \mathbf{b} \quad \text{and} \quad \mathbf{d} = \ \mathbf{a} + \ \mathbf{b} \tag{3}$$

to express **b** and **d** in terms of **a** and **c**, as follows.

From the first equation in (3) we have

$$\begin{aligned}\mathbf{b} &= (\mathbf{c} \qquad \mathbf{a})/ \\ &= (\quad / \quad)\mathbf{a} + (1/ \quad)\mathbf{c}. \end{aligned}\tag{4}$$

cannot be zero, for if it were then we would have **c** = **a**; this cannot happen since A and C are distinct Points.

If we then substitute this expression for **b** into the second equation in (3), we obtain

$$\begin{aligned}\mathbf{d} &= \quad \mathbf{a} + \quad ((\quad / \quad)\mathbf{a} + (1/ \quad)\mathbf{c}) \\ &= ((\qquad\qquad)/ \quad)\mathbf{a} + (\quad / \quad)\mathbf{c}. \end{aligned}\tag{5}$$

It follows from the coefficients of **a** and **c** in equations (4) and (5) that

$$\begin{aligned}(ACBD) &= \frac{1/}{\quad /} \quad \frac{/}{(\qquad\qquad)/} \\ &= \qquad \frac{}{\qquad\qquad} \\ &= 1 \quad \frac{}{\quad} \\ &= 1 \quad k. \end{aligned}$$

Finally, we can use the previous parts of the proof to evaluate $(DBCA)$, as follows:

This avoids the algebra involved in expressing **c** and **a** in terms of **d** and **b**.

$$\begin{aligned}(DBCA) &= 1/(BDCA) && \text{(swap first two Points)} \\ &= (BDAC) && \text{(swap last two Points)} \\ &= 1 \quad (BADC) && \text{(swap middle two Points)} \\ &= 1 \quad 1/(ABDC) && \text{(swap first two Points)} \\ &= 1 \quad (ABCD) && \text{(swap last two Points)} \\ &= 1 \quad k. \end{aligned}$$

Example 1

Earlier, we showed that the cross-ratio $(ABCD)$ of the four collinear Points $A = [1, 2, 3]$, $B = [1, 1, 2]$, $C = [3, 5, 8]$, $D = [1, \quad 1, 0]$ in $\mathbb{RP}^2$ is $\quad \frac{1}{3}$. Theorem 2 enables us to deduce that

$$\begin{aligned}(BACD) &= \quad 3, & (ABDC) &= \quad 3, \\ (ACBD) &= \tfrac{4}{3}, & (DBCA) &= \tfrac{4}{3}. \end{aligned}$$

Problem 3 Let the Points $A = [1, \quad 1, \quad 1]$, $B = [1, 3, \quad 2]$, $C = [3, 5, \quad 5]$, $D = [1, \quad 5, 0]$ be collinear Points of $\mathbb{RP}^2$. By applying Theorem 2 to the solution of Problem l(a), determine the values of the cross-ratios $(ABDC)$, $(DBCA)$ and $(ACBD)$.

The next theorem confirms that cross-ratio is preserved by projective transformations.

Theorem 3 Let t be a projective transformation, and let A, B, C, D be any four collinear Points in RP^2. If $A = t(A), B = t(B), C = t(C), D = t(D)$, then

$$(ABCD) = (A\ B\ C\ D).$$

Proof Let t be the projective transformation $t : [\mathbf{x}] \qquad [\mathbf{Ax}]$, where **A** is an invertible 3 × 3 matrix. If $A = [\mathbf{a}], B = [\mathbf{b}], C = [\mathbf{c}], D = [\mathbf{d}]$, and

$$\mathbf{a} = \mathbf{Aa},\ \mathbf{b} = \mathbf{Ab},\ \mathbf{c} = \mathbf{Ac},\ \mathbf{d} = \mathbf{Ad},$$

then $A = [\mathbf{a}],\ B = [\mathbf{b}],\ C = [\mathbf{c}],\ D = [\mathbf{d}]$.

Since A, B, C, D are collinear, we can write

$$\mathbf{c} = \quad \mathbf{a} + \quad \mathbf{b} \quad \text{and} \quad \mathbf{d} = \quad \mathbf{a} + \quad \mathbf{b}, \tag{6}$$

so

$$(ABCD) = - \quad -.$$

Multiplying each equation in (6) through by **A**, we obtain

$$\mathbf{c} = \quad \mathbf{a} + \quad \mathbf{b} \quad \text{and} \quad \mathbf{d} = \quad \mathbf{a} + \quad \mathbf{b},$$

so that

$$(A\ B\ C\ D) = - \quad -.$$

It follows that

$$(A\ B\ C\ D) = (ABCD).$$

We now use Theorem 3 to prove that if four distinct Points on a Line are in perspective with four distinct Points on another Line, then the cross-ratios of the four Points on each Line are equal.

Theorem 4 Let A, B, C, D be four distinct Points on a Line, and let A, B, C, D be four distinct Points on another Line such that AA, BB, CC, DD all meet at a Point U. Then

$$(ABCD) = (A\ B\ C\ D).$$

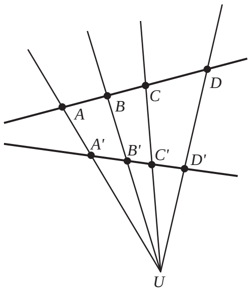

Proof By the Fundamental Theorem of Projective Geometry, there is a unique projective transformation t which maps B to B, C to C, B to B, and C to C. We shall show that $t(A) = A$ and $t(D) = D$, and hence by Theorem 3 it follows that $(ABCD) = (A\ B\ C\ D)$.

First observe that the composite $t \quad t$ fixes the Points B, C, B and C. By the Fundamental Theorem of Projective Geometry, the only projective transformation which does this is the identity transformation, so $t \quad t = i$ and $t = t^{\ 1}$.

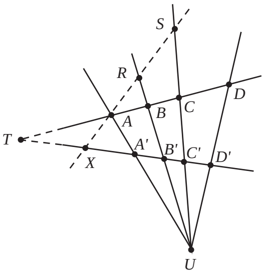

Next observe that t maps the Line BC onto the Line $B\ C$, and vice versa; so the Point T at which BC and $B\ C$ intersect must be fixed by t. Also, t maps the Lines BB and CC onto themselves, so their Point of intersection U must be fixed by t.

Now let X be the image of A under t. Then X lies on $B\ C$. We want to show that $X = A$.

Suppose that $X = A$; then AX cannot pass through U so it must intersect BB at R and CC at S, where R, S and U are distinct Points.

Since t is self-inverse, it maps X back to A and therefore maps AX onto itself. But this implies that t fixes the four Points R, S, T, U; so by the Fundamental Theorem of Projective Geometry t must be the identity transformation. This is a contradiction with the hypothesis that the Lines $ABCD$ and $A\ B\ C\ D$ are different. It follows that we must conclude that $X = A$, that is, $t(A) = A$. A similar argument shows that $t(D) = D$.

Finally, it follows by Theorem 3 that $(ABCD) = (A\ B\ C\ D\ )$, as required.

In affine geometry, if we are given two points A and B, then the ratio AC/CB uniquely determines a third point C on the line AB. We now explore the analogous result for projective geometry, namely that if we are given any three collinear Points A, B, C in $\mathbb{RP}^2$, then the value of the cross-ratio $(ABCD)$ uniquely determines a fourth Point D.

Theorem 5 Unique Fourth Point Theorem

Let A, B, C, X, Y be collinear Points in $\mathbb{RP}^2$ such that

$$(ABCX) = (ABCY).$$

Then $X = Y$.

Proof Let $A = [\mathbf{a}]$, $B = [\mathbf{b}]$, $C = [\mathbf{c}]$, $X = [\mathbf{x}]$, $Y = [\mathbf{y}]$. Since A, B, C, X, Y are collinear, it follows that there are real numbers , , , , μ such that

$$\mathbf{c} = \ \mathbf{a} + \ \mathbf{b}, \quad \mathbf{x} = \ \mathbf{a} + \ \mathbf{b} \quad \text{and} \quad \mathbf{y} = \ \mathbf{a} + \mu\mathbf{b}. \tag{7}$$

Then

$$(ABCX) = \frac{\ }{\ } \quad \text{and} \quad (ABCY) = \frac{\ }{\mu}.$$

Since $(ABCX) = (ABCY)$, it follows that

This is the hypothesis of the theorem.

$$\frac{\ }{\ } = \frac{\ }{\mu},$$

so $= \mu/$. If we substitute this value of into the expression for $\mathbf{y}$ in equation (7), we obtain

$$\mathbf{y} = (\ \mu/\ )\mathbf{a} + \mu\mathbf{b} = (\mu/\ )(\ \mathbf{a} + \ \mathbf{b}) = (\mu/\ )\mathbf{x}.$$

Since $\mathbf{y}$ is a scalar multiple of $\mathbf{x}$, it follows that $X = Y$, as required.

In Theorem 4 we showed that the cross-ratios $(ABCD)$ and $(A\ B\ C\ D\ )$ are equal if the Points $A\ , B\ , C\ , D$ are in perspective with the Points A, B, C, D. Our next result is a partial converse of this result.

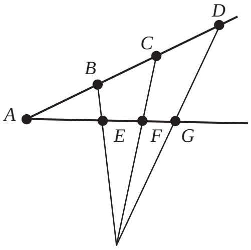

Theorem 6 Let A, B, C, D and A, E, F, G be two sets of collinear Points (on different Lines in $\mathbb{RP}^2$) such that the cross-ratios $(ABCD)$ and $(AEFG)$ are equal. Then the Lines BE, CF and DG are concurrent.

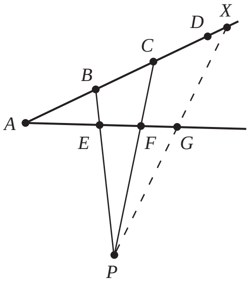

Proof Let P be the Point at which the Lines BE and CF meet, and let X be the Point at which the Line PG meets the Line $ABCD$. Then the Points A, B, C and X are in perspective from P with the Points A, E, F and G, so that

$$(ABCX) = (AEFG).$$

Since we know that $(AEFG) = (ABCD)$, it follows that

$$(ABCX) = (ABCD).$$

By Theorem 5, we must therefore have $X = D$. Hence the Points A, B, C, D and the Points A, E, F, G are in perspective from P.

We can now use Theorem 6 together with the other properties of cross-ratio to give a second proof of Pappus Theorem.

Theorem 7 Pappus' Theorem
Let A, B and C be three Points on a Line in $\mathbb{RP}^2$, and let $A\ , B$ and C be three Points on another Line. Let BC and $B\ C$ meet at P, CA and $C\ A$ meet at Q, and AB and $A\ B$ meet at R. Then P, Q and R are collinear.

You met this Theorem earlier, in Subsection 3.4.1.

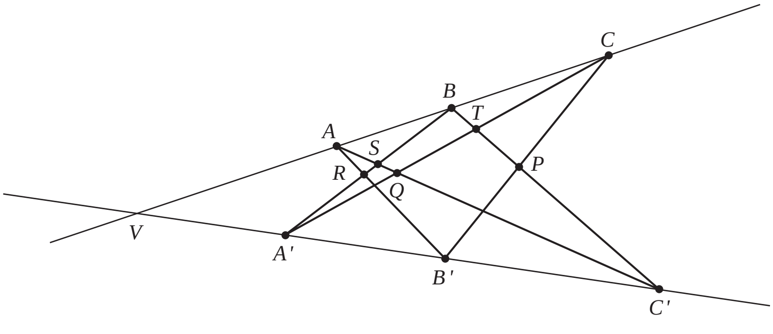

Proof Let V be the Point of intersection of the two given Lines. Also let the Lines BA and AC meet at the Point S, and the Lines BC and CA meet at the Point T.

Now, the Points $V, A\ , B\ , C$ are in perspective from A with the Points B, $A\ , R, S$, so that

$$(VA\ B\ C\ ) = (BA\ RS). \tag{8}$$

By Theorem 4

Similarly, the Points V, A', B', C' are in perspective from C with the Points B, T, P, C', so that

$$(VA'B'C') = (BTPC'). \tag{9}$$

It follows from equations (8) and (9) that

$$(BA'RS) = (BTPC'),$$

so that by Theorem 6 the Lines $A'T, RP, SC'$ are concurrent.

We may rephrase this statement as follows: the Line RP passes through the Point where $A'T$ meets SC'; that is, the Line RP passes through Q. In other words, P, Q and R are collinear.

3.5.2 Cross-Ratio on Embedding Planes

So far, we have calculated a given cross-ratio $(ABCD)$ by applying the definition of cross-ratio directly to the Points A, B, C, D. However, it is sometimes convenient to evaluate the cross-ratio by examining the representation of the Points on some embedding plane.

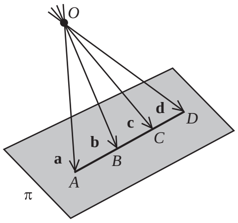

See Appendix 2.

Suppose that four collinear Points of $\mathbb{RP}^2$ pierce an embedding plane π at the points A, B, C, D with position vectors **a**, **b**, **c**, **d**, respectively.

According to the Section Formula, we can write **c** and **d** in the form

$$\mathbf{c} = \lambda\mathbf{a} + (1-\lambda)\mathbf{b} \quad \text{and} \quad \mathbf{d} = \mu\mathbf{a} + (1-\mu)\mathbf{b},$$

where $(1-\lambda) : \lambda$ is the ratio $AC : CB$, and $(1-\mu) : \mu$ is the ratio $AD : DB$. Then from the definition of cross-ratio

$$(ABCD) = \frac{1-\lambda}{\lambda} \div \frac{1-\mu}{\mu},$$

so

$$(ABCD) = \frac{AC}{CB} \div \frac{AD}{DB}. \tag{10}$$

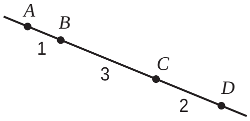

Example 2 In an embedding plane, the points A, B, C, D lie in order along a line with the distances AB, BC, CD being 1 unit, 3 units and 2 units, respectively. Determine the cross-ratios $(ABCD)$, $(BACD)$ and $(ACBD)$.

Solution Using equation (10) and the sign convention for ratios, we have

$$(ABCD) = \frac{AC}{CB} \div \frac{AD}{DB} = \frac{4}{3} \div -\frac{6}{5} = -\frac{10}{9},$$

$$(BACD) = \frac{BC}{CA} \div \frac{BD}{DA} = -\frac{3}{4} \div -\frac{5}{6} = \frac{9}{10}$$

and

$$(ACBD) = \frac{AB}{BC} \quad \frac{AD}{DC} = \frac{1}{3} \qquad \frac{6}{2} = \frac{1}{9}. \qquad \square$$

Problem 4 The points A, B, C, D lie in order along a line with the distances AB, BC, CD being 2 units, 1 unit and 3 units, respectively. Determine the cross-ratios $(ABCD)$ and $(DBCA)$.

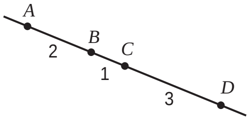

Sometimes one of the Points whose cross-ratio we are trying to find turns out to be an ideal Point for the embedding plane. In such cases, formula (10) cannot be used since some of the distances in the formula will not be defined.

To be specific, suppose that the Points A, B, C, D are collinear, but that A is an ideal Point for the embedding plane , as shown in the margin. As before, we can let **b**, **c**, **d** be the position vectors of the points B, C, D on , but we take **a** to be a unit vector along A. Then

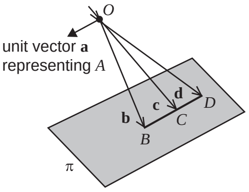

$$\mathbf{c} = \quad (CB)\mathbf{a} + \mathbf{b} \quad \text{and} \quad \mathbf{d} = \quad (DB)\mathbf{a} + \mathbf{b}.$$

From the definition of cross-ratio, it follows that

$$(ABCD) = \frac{1}{CB} \quad \frac{1}{DB} = \frac{DB}{CB}. \tag{11}$$

We can now obtain the corresponding formulas for the cases where B, C or D is an ideal Point, by applying Theorem 2. For example, if B is an ideal Point, then

$$\begin{aligned}(ABCD) &= \frac{1}{(BACD)} && \text{(swap first two terms)}\\ &= (BADC) && \text{(swap last two terms)}\\ &= \frac{CA}{DA} && \text{by equation (11).}\end{aligned}$$

Problem 5 Use Theorem 2 To prove that:

(a) $(ABCD) = \frac{AC}{BC}$ if D is an ideal Point;
(b) $(ABCD) = \frac{BD}{AD}$ if C is an ideal Point.

We now summarize the various formulas for cross-ratio in the form of a strategy, as follows.

Strategy To use an embedding plane to calculate the cross-ratio of four collinear Points:

1. if the four Points pierce the embedding plane at A, B, C, D, then

$$(ABCD) = \frac{AC}{CB} \quad \frac{AD}{DB};$$

2. if one of the Points is an ideal Point for the embedding plane, then

$$(ABCD) = \frac{DB}{CB} \text{ if } A \text{ is ideal,}$$

$$(ABCD) = \frac{CA}{DA} \text{ if } B \text{ is ideal,}$$

$$(ABCD) = \frac{BD}{AD} \text{ if } C \text{ is ideal,}$$

$$(ABCD) = \frac{AC}{BC} \text{ if } D \text{ is ideal.}$$

Example 3 Determine $(ABCD)$ for the collinear points A, B, C, D illustrated in the margin, where C is an ideal Point.

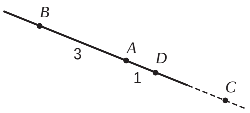

Solution Since C is an ideal Point, we have

$$(ABCD) = \frac{BD}{AD} = \frac{4}{1} = 4.$$

□

Problem 6 Determine $(ABCD)$ for the collinear points A, B, C, D illustrated in the margin, where B is an ideal Point.

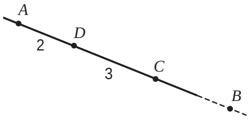

3.5.3 An Application of Cross-Ratio

Earlier, we described how projective geometry can be used to obtain two-dimensional representations of three-dimensional scenes. We now describe how cross-ratios can be used to obtain information about a three-dimensional scene from a two-dimensional representation of the scene. We do this in the context of aerial photography.

Subsection 3.1.1

For simplicity, consider an aerial camera that takes pictures on a flat film behind its lens, L, of features on a flat piece of land in front of L. Since a point on the ground lies on the same line through L as its image on the film, we can regard the process of taking a photograph as a perspectivity centred at L.

Since collinearity is invariant under a perspectivity, the image of any line on the ground is a line on the film. Moreover, the cross-ratio of any four collinear points is invariant under a perspectivity, so the cross-ratio of any four points on must be equal to the cross-ratio of their images on the film.

Section 3.3, Theorem 2

Theorem 3

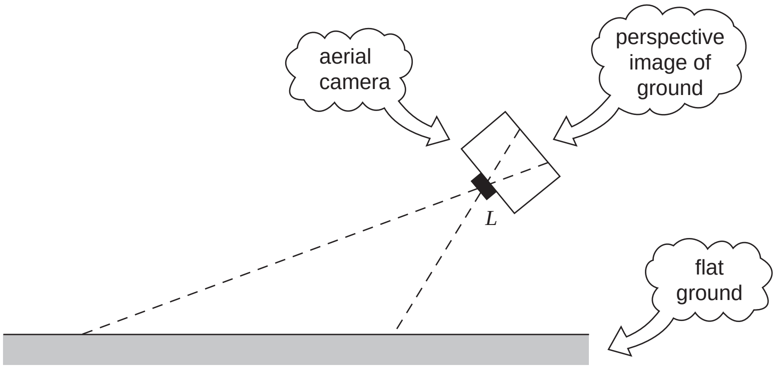
aerial camera
perspective image of ground
L
flat ground

Example 4 An aerial camera photographs a car travelling along a straight road on flat ground towards a junction. Before the junction there are two warning signs at distances of 4 km and 2 km from the junction. On the film the signs are 1 cm and 3 cm from the junction, and the car is $\frac{3}{7}$ cm from the junction. How far is the car from the junction on the ground?

Strictly speaking, the car is not in line with the two signposts. Consequently, the distances marked on the photograph are approximations measured along the line of the left-hand kerb of the road.

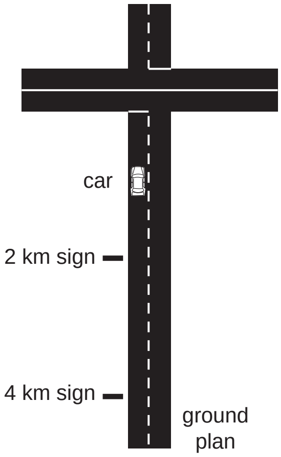
car
2 km sign
4 km sign
ground plan

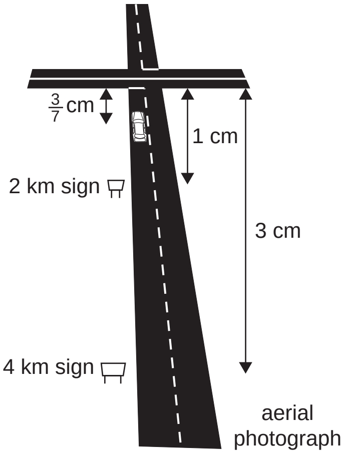
$\frac{3}{7}$ cm
1 cm
2 km sign
3 cm
4 km sign
aerial photograph

Solution Let A and B denote the signs, C denote the car, and D denote the junction, and let A', B', C', D' be their images on the film. Then

$$\begin{aligned}(A'B'C'D') &= \frac{A'C'}{C'B'} \div \frac{A'D'}{D'B'} \\ &= \frac{18/7}{4/7} \div \frac{3}{1} \\ &= \frac{3}{2}.\end{aligned}$$

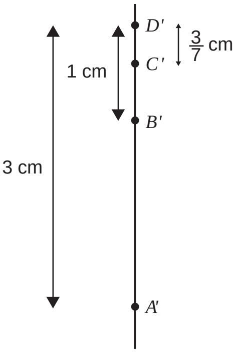
D′
C′
B′
A′
$\frac{3}{7}$ cm
1 cm
3 cm

Now let the car be n km from the junction. Then

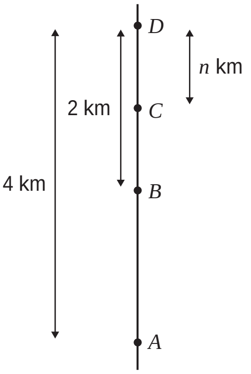

$$\begin{aligned}(ABCD) &= \frac{AC}{CB} \div \frac{AD}{DB}\\ &= \frac{4-n}{2-n} \div \frac{4}{2}\\ &= \frac{4-n}{2(2-n)}.\end{aligned}$$

Since $(ABCD)$ and $(A'B'C'D')$ must be equal, it follows that

$$\frac{4-n}{2(2-n)} = \frac{3}{2}.$$

Hence

$$4-n = 3(2-n).$$

and so $n = 1$. That is, the car is 1 km from the junction. □

Problem 7 An aerial camera photographs a car travelling along a straight road on flat ground towards a junction. Before the junction there are two warning signs, at distances of 2 km and 3 km from the junction. On the film the signs are 4 cm and 6 cm from the junction, and the car is 1 cm from the junction. How far is the car from the junction on the ground?

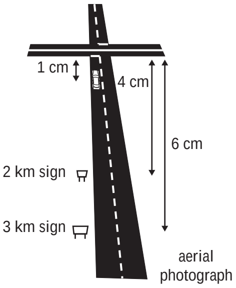

If two lines that are known to be parallel on the ground appear to meet on the film, then the point of intersection on the film corresponds to the ideal Point where the ‘parallel lines meet’. We can therefore use the above technique even when one of the Points is ideal, for we can use the second part of the strategy in Subsection 3.5.2 to calculate the cross-ratio whenever one of the Points is ideal.

Problem 8 An aerial camera photographs a train travelling between two stations along a straight track on flat ground. The stations are 50 km apart. When the film is inspected, the stations are 4 cm apart, the train is midway between the stations, and the rails appear to meet (or vanish) 4 cm beyond the station towards which the train is travelling. How far has the train to travel to the next station?

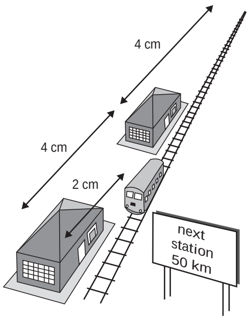

3.6 Exercises

Section 3.2

1. (a) Write down numbers a, b, c and d such that

$$[1, a, b] = \quad \tfrac{1}{2}, 3, 4 \qquad \text{and} \qquad [c, d, 2] = [3, 0, 1].$$

   (b) Which of the following homogeneous coordinates represent the same Point of $\mathbb{RP}^2$ as $[4, \quad 8, 2]$?

$$\text{(i) } [1, 4, \quad 2] \qquad \text{(ii) } \quad \tfrac{1}{4}, \quad \tfrac{1}{2}, \tfrac{1}{8} \qquad \text{(iii) } \quad \tfrac{1}{2}, \quad 2, 1$$

$$\text{(iv) } [\quad 2, 4, \quad 1] \qquad \text{(v) } \quad \tfrac{1}{8}, \quad \tfrac{1}{2}, \tfrac{1}{4}$$

2. Determine an equation for each of the following Lines in $\mathbb{RP}^2$:
   (a) the Line through the Points [1, 2, 3] and [3, 0, 2];
   (b) the Line through the Points [1, 1, 1] and [2, 1, 3].
3. Determine whether each of the following sets of Points are collinear:
   (a) [1, 1, 0], [1, 0, 1] and [2, 1, 1];
   (b) [1, 0, 1], [0, 1, 2] and [1, 2, 3].
4. Determine the Point of intersection of each of the following pairs of Lines in $\mathbb{RP}^2$:
   (a) the Lines with equations $x \quad 2y + z = 0$ and $x \quad y \quad z = 0$;
   (b) the Lines with equations $x + 2y + 5z = 0$ and $3x \quad y + z = 0$.
5. Determine the Point of $\mathbb{RP}^2$ at which the Line through the Points [8, 1, 2] and [1, 2, 1] meets the Line through the Points [0, 1, 1] and [2, 3, 1].
6. Determine the Point of $\mathbb{RP}^2$ at which the Line through the Points [1, 2, 2] and [2, 3, 3] meets the Line through the Points [0, 1, 2] and [0, 1, 3].

Section 3.3

In these exercises, you may find the following list of matrices and their inverses useful.

$$\mathbf{A}: \quad \begin{matrix} 2 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{matrix} \qquad \begin{matrix} 2 & 0 & 1 \\ 0 & 3 & 2 \\ 1 & 3 & 1 \end{matrix} \qquad \begin{matrix} 0 & 3 & 1 \\ 2 & 0 & 1 \\ 0 & 0 & 1 \end{matrix} \qquad \begin{matrix} 0 & 3 & 4 \\ 1 & 3 & 2 \\ 3 & 3 & 3 \end{matrix}$$

$$\mathbf{A}^{\ 1}: \quad \begin{matrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 1 \end{matrix} \qquad \begin{matrix} 1 & 1 & 1 \\ \frac{2}{3} & 1 & \frac{4}{3} \\ 1 & 2 & 2 \end{matrix} \qquad \begin{matrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 \end{matrix} \qquad \begin{matrix} 5 & 7 & 2 \\ 3 & 4 & \frac{4}{3} \\ 2 & 3 & 1 \end{matrix}$$

1. Determine which of the following transformations t of $\mathbb{RP}^2$ are projective transformations. For those that are projective transformations, write down a matrix associated with t.

(a) $t : [x, y, z] \quad [2x, y + 3z, 1]$
(b) $t : [x, y, z] \quad [x, x \ \ y + 3z, x + y]$
(c) $t : [x, y, z] \quad [2y, y \ \ 4z, x]$
(d) $t : [x, y, z] \quad [x + y \ \ z, y + 3z, x + 2y + 2z]$

2. Determine the images of the Points [1, 2, 3], [0, 1, 0] and [1, $\ $ 1, 1] under the projective transformation t associated with the matrix

$$\mathbf{A} = \begin{matrix} 2 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{matrix} .$$

3. Let

$$t_1 : [x, y, z] \quad [2x + y, \ \ x + z, y + z],$$
$$t_2 : [x, y, z] \quad [x + y, 3x \ \ z, 4y \ \ 2z]$$

be projective transformations from $\mathbb{R}P^2$ to $\mathbb{R}P^2$.
(a) Write down matrices associated with each of t_1 and t_2.
(b) Determine formulas for $t_2 \quad t_1$ and $t_2 \quad t_1^{\ 1}$.

4. Find the image of the Line $x + 2y + 3z = 0$ under the projective transformation t_1 defined in Exercise 3.

5. Determine matrices for the projective transformations which map the Points [1, 0, 0], [0, 1, 0], [0, 0, 1] and [1, 1, 1] onto the following Points:
(a) [$\ $ 2, 0, 1], [0, 1, $\ $ 1], [$\ $ 1, 2, $\ $ 1] and [$\ $ 1, 1, $\ $ 1];
(b) [0, 1, 0], [1, 0, 0], [$\ $ 1, $\ $ 1, 1] and [2, 1, 1];
(c) [0, 1, $\ $ 3], [1, 1, $\ $ 1], [4, 2, 3] and [7, 4, 3].

6. Use the results of Exercise 5 to determine the projective transformations that map:
(a) the Points

$$[\ \ 2, 0, 1],\ [0, 1, \ \ 1],\ [\ \ 1, 2, \ \ 1],\ [\ \ 1, 1, \ \ 1]$$

to the Points

$$[0, 1, 0],\ [1, 0, 0],\ [\ \ 1, \ \ 1, 1],\ [2, 1, 1],$$

respectively;
(b) the Points

$$[0, 1, 0],\ [1, 0, 0],\ [\ \ 1, \ \ 1, 1],\ [2, 1, 1]$$

to the Points

$$[0, 1, \ \ 3],\ [1, 1, \ \ 1],\ [4, 2, 3],\ [7, 4, 3],$$

respectively;
(c) the Points

$$[0, 1, \ \ 3],\ [1, 1, \ \ 1],\ [4, 2, 3],\ [7, 4, 3]$$

to the Points

$$[\ \ 2, 0, 1],\ [0, 1, \ \ 1],\ [\ \ 1, 2, \ \ 1],\ [\ \ 1, 1, \ \ 1],$$

respectively.

Section 3.4

1. For which of the following configurations of Points A, B, C and D in $\mathbb{RP}^2$ is there a projective transformation sending A, B, C to the triangle of reference and D to the unit Point?

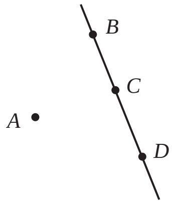

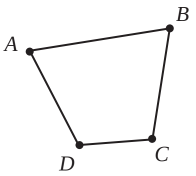

2. Let ABC be a triangle in $\mathbb{R}^2$, and let U be any point of $\mathbb{R}^2$ that is not collinear with any two of the points A, B, C. Let the Lines BC and AU meet at P, CA and BU meet at Q, and AB and CU meet at R. Prove that P, Q, R cannot be collinear.

Section 3.5

1. For each of the following sets of Points A, B, C, D, calculate the cross-ratio $(ABCD)$.
   (a) $A = [2, 1, 3],\ B = [1, 2, 3],\ C = [8, 1, 9],\ D = [4, \ 1, 3]$
   (b) $A = [2, 1, 1],\ B = [\ 1, 1, \ 1],\ C = [1, 2, 0],\ D = [\ 1, 4, \ 2]$
   (c) $A = [\ 1, 1, 1],\ B = [0, 0, 2],\ C = [5, \ 5, 3],\ D = [\ 3, 3, 7]$
2. For the Points A, B, C, D in Exercise 1(a), determine the cross-ratios $(BACD)$, $(BDCA)$ and $(ADBC)$.
3. For each set of collinear points A, B, C, D illustrated below, calculate the cross-ratio $(ABCD)$.

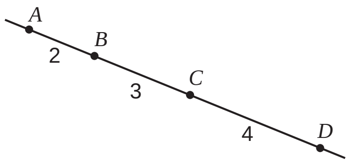

(a)

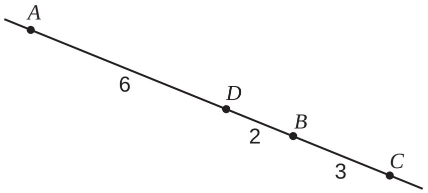

(b)

4. Calculate the cross-ratio $(ABCD)$ for the collinear points A, B, C, D illustrated below, where D is an ideal Point.

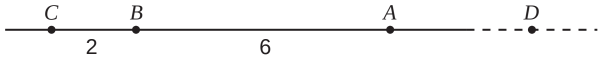

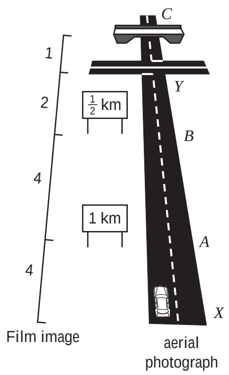

5. The diagram in the margin represents an aerial photograph of a straight road on flat ground. At A there is a sign Junction 1 km , at B a sign Junction

$\frac{1}{2}$ km , and C is the road junction. Also, a police patrol car is at X, and a bridge is at Y. The distances marked on the left of the diagram are measured in cm from the photograph.

Calculate the actual distances (in km) of the patrol car and the bridge from the junction.

Summary of Chapter 3

Section 3.1: Perspective

1. Renaissance artists used **terraced perspective** and later **vertical perspective** in an attempt to portray real scenes in a realistic way. The modern system of **focused perspective** was discovered by Brunelleschi and finally perfected by Leonardo da Vinci; it is well illustrated by the woodcuts of Albrecht Dürer.

   The family of lines joining an eye to each point of a scene meets a screen in front of the eye, and the image on the screen is called a **cross-section** (or **section**). The cross-section gives a realistic two-dimensional representation of the three-dimensional scene.
2. For two planes and that do not pass through the origin O in $\mathbb{R}^3$, points P in and Q in are **in perspective from *O*** if there is a straight line through O, P and Q.

   A **perspectivity from to** centred at O is a function that maps a point P of to a point Q of whenever P and Q are in perspective from O. (The planes and may lie on the same or on opposite sides of O.)
3. The domain of a perspectivity may not be the whole of ; for, if P is any point of such that OP is parallel to to , then P cannot have an image in .

   The image of a line under a perspectivity is another line, possibly minus one point.

   Foreshortening is the effect under a perspectivity of lines of equal lengths at different distances from a screen corresponding to different lengths on the screen.
4. Two parallel lines in a horizontal plane appear to an observer to meet at a **vanishing point** on a vertical screen ; this is the **principal vanishing point** if the lines are perpendicular to the line of intersection of and , and a **diagonal vanishing point** otherwise.

   The family of vanishing points is a line, the **vanishing line**, that corresponds to the horizon line in a picture.
5. **Desargues' Theorem** Let ABC and $A\ B\ C$ be triangles in $\mathbb{R}^2$ such that the lines AA , BB and CC meet at a point U. Let BC and $B\ C$ meet at P, CA and $C\ A$ meet at Q, and AB and $A\ B$ meet at R. Then P, Q and R are collinear.

Section 3.2: The Projective Plane $\mathbb{RP}^2$

1. A **Point** (or **projective point**) is a line in $\mathbb{R}^3$ that passes through the origin of $\mathbb{R}^3$.

   The **real projective plane** $\mathbb{RP}^2$ is the set of all such Points.
2. The expression $[a,b,c]$, in which the numbers a, b and c are not all zero, represents the Point P of $\mathbb{RP}^2$ which consists of the unique line in $\mathbb{R}^3$ that passes through (0,0,0) and (a,b,c). We refer to $[a,b,c]$ as **homogeneous coordinates** of P.

   If (a,b,c) has position vector **v**, then we often denote P by [**v**] and we say that P can be **represented** by **v**.

   It makes NO sense to write the expression [0,0,0], since not all of a, b and c can be zero.
3. The homogeneous coordinates $[a,b,c]$ and $[\lambda a, \lambda b, \lambda c]$ (where $\lambda \neq 0$) represent the same Point of $\mathbb{RP}^2$; that is, $[a,b,c] = [\lambda a, \lambda b, \lambda c]$, for any $\lambda \neq 0$.

   If there is no non-zero real number λ such that $[a,b,c] = \left[\lambda a', \lambda b', \lambda c'\right]$, then the homogeneous coordinates $[a,b,c]$ and $\left[a', b', c'\right]$ represent different Points of $\mathbb{RP}^2$.

   Further, $\left[a', b', 1\right] = \left[a'', b'', 1\right]$ if and only if $a' = a''$ and $b' = b''$.
4. A **projective figure** is a subset of $\mathbb{RP}^2$.
5. A **Line** (or **projective line**) in $\mathbb{RP}^2$ is a plane in $\mathbb{R}^3$ that passes through the origin. Points of $\mathbb{RP}^2$ are **collinear** if they lie on a Line.
6. The general equation of a Line in $\mathbb{RP}^2$ is $ax + by + cz = 0$, where a, b, c are real and not all zero.
7. **Collinearity Property of** $\mathbb{RP}^2$ Any two distinct Points of $\mathbb{RP}^2$ lie on a unique Line.

   Strategy To determine an equation for the Line in $\mathbb{RP}^2$ through the Points $\left[d,e,f\right]$ and $\left[g,h,k\right]$:

   1. write down the equation $\begin{vmatrix} x & y & z \\ d & e & f \\ g & h & k \end{vmatrix} = 0$;
   2. expand the determinant in terms of the entries in its first row to obtain the required equation in the form $ax + by + cz = 0$.

   Sometimes it is possible to 'spot' the equation of a Line through two Points without using the determinant.
8. **Strategy** To determine whether three Points $[a,b,c]$, $[d,e,f]$ and $[g,h,k]$ are collinear:

   1. evaluate the determinant $\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & k \end{vmatrix}$;
   2. the Points $[a,b,c]$, $[d,e,f]$ and $[g,h,k]$ are collinear if and only if this determinant is zero.
9. The Points [1,0,0], [0,1,0], [0,0,1] are known as the **triangle of reference**. The Point [1,1,1] is called the **unit Point**.

10. **Incidence Property of** $\mathbb{RP}^2$ Any two distinct Lines in $\mathbb{RP}^2$ intersect in a unique Point of $\mathbb{RP}^2$.
11. Let be any plane in $\mathbb{R}^3$ that does not pass through the origin O. Then there is a one-one correspondence between the points of and those Points of $\mathbb{RP}^2$ which pierce . Those Points of $\mathbb{RP}^2$ which do not pierce are called **ideal Points** for .

    The set of ideal Points for is a plane through O parallel to , called the **ideal Line** for .
12. An **embedding plane** is a plane, , which does not pass through the origin, together with the set of all ideal Points for .

    The plane in $\mathbb{R}^3$ with equation $z = 1$ is called the **standard embedding plane**. The mapping of $\mathbb{RP}^2$ into the standard embedding plane is called the **standard embedding** of $\mathbb{RP}^2$.
13. Parallelism is not a projective property.

Section 3.3: Projective Transformations

1. A **projective transformation** of $\mathbb{RP}^2$ is a function t: $\mathbb{RP}^2$ $\mathbb{RP}^2$ of the form $t : [\mathbf{x}]$ $[\mathbf{Ax}]$, where $\mathbf{A}$ is an invertible 3 × 3 matrix. We say that $\mathbf{A}$ is a matrix **associated** with t. The set of all projective transformations is denoted by $P(2)$.

   If $\mathbf{A}$ is a matrix **associated** with t, then so is $\mathbf{A}$ for any non-zero number .
2. The set of projective transformations $P(2)$ forms a group under the operation of composition of functions. In particular, if t_1 and t_2 are projective transformations with associated matrices $\mathbf{A}_1$ and $\mathbf{A}_2$, respectively, then $t_1 \ t_2$ and $t_1^{\ 1}$ are projective transformations with associated matrices $\mathbf{A}_1\mathbf{A}_2$ and $\mathbf{A}_1^{\ 1}$.

   Strategy To compose two projective transformations t_1 and t_2:
   1. write down matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ associated with t_1 and t_2;
   2. calculate $\mathbf{A}_1\mathbf{A}_2$;
   3. write down the composite $t_1 \ t_2$ with which $\mathbf{A}_1\mathbf{A}_2$ is associated.

   Strategy To find the inverse of a projective transformation t:
   1. write down a matrix $\mathbf{A}$ associated with t;
   2. calculate $\mathbf{A}^{\ 1}$;
   3. write down the inverse $t^{\ 1}$ with which $\mathbf{A}^{\ 1}$ is associated.
3. **Strategy** To find the image of a Line $ax + by + cz = 0$ under a projective transformation $t : [\mathbf{x}]$ $[\mathbf{Ax}]$:
   1. write the equation of the Line in the form $\mathbf{Lx} = 0$, where $\mathbf{L}$ is the matrix $(a\,b\,c)$;
   2. find a matrix $\mathbf{B}$ associated with $t^{\ 1}$;
   3. write down the equation of the image as $(\mathbf{LB})\,\mathbf{x} = 0$.
4. **Projective geometry** is the study of those properties of figures in $\mathbb{RP}^2$ that are preserved by projective transformations. Collinearity and incidence are both projective properties.

5. A **quadrilateral** is a set of four Points A, B, C and D (no three of which are collinear), together with the Lines AB, BC, CD and DA.
   All quadrilaterals are projective-congruent.
6. **Strategy** To find the projective transformation which maps [1,0,0] to $[a_1, a_2, a_3]$, [0,1,0] to $[b_1, b_2, b_3]$, [0,0,1] to $[c_1, c_2, c_3]$, [1,1,1] to $[d_1, d_2, d_3]$, where no three of $[a_1, a_2, a_3]$, $[b_1, b_2, b_3]$, $[c_1, c_2, c_3]$ and $[d_1, d_2, d_3]$ are collinear:
   1. find u, v, w for which $\begin{matrix} a_1 u & b_1 \mathsf{v} & c_1 \mathsf{w} \\ a_2 u & b_2 \mathsf{v} & c_2 \mathsf{w} \\ a_3 u & b_3 \mathsf{v} & c_3 \mathsf{w} \end{matrix} \quad \begin{matrix} 1 \\ 1 \\ 1 \end{matrix} = \begin{matrix} d_1 \\ d_2 \\ d_3 \end{matrix}$;
   2. write down the required projective transformation in the form t: $[\mathbf{x}] \quad [\mathbf{Ax}]$, where $\mathbf{A}$ is any non-zero real multiple of the matrix $\begin{matrix} a_1 u & b_1 \mathsf{v} & c_1 \mathsf{w} \\ a_2 u & b_2 \mathsf{v} & c_2 \mathsf{w} \\ a_3 u & b_3 \mathsf{v} & c_3 \mathsf{w} \end{matrix}$.
7. **Fundamental Theorem of Projective Geometry** Let $ABCD$ and $A\ B\ C\ D$ be two quadrilaterals in $\mathbb{RP}^2$. Then:
   (a) there is a projective transformation t which maps A to A , B to B , C to C , D to D ;
   (a) the projective transformation t is unique.
8. **Strategy** To determine the projective transformation t which maps the vertices of the quadrilateral $ABCD$ to the corresponding vertices of the quadrilateral $A\ B\ C\ D$:
   1. find the projective transformation t_1 which maps the triangle of reference and unit Point to the Points A, B, C, D, respectively;
   2. find the projective transformation t_2 which maps the triangle of reference and unit Point to the Points A , B , C , D , respectively;
   3. calculate $t = t_2 \quad t_1{}^{1}$.
9. With any given perspectivity we can construct an associated **perspective transformation** that is a one-one mapping of {the ideal Points for } onto the ideal Points for . This maps $\mathbb{RP}^2$ onto itself.
   Every projective transformation can be expressed as the composite of three perspective transformations.

Section 3.4: Using the Fundamental Theorem of Projective Geometry

1. **Desargues' Theorem** Let ABC and $A\ B\ C$ be triangles in $\mathbb{R}^2$ such that the lines AA , BB and CC meet at a point U. Let BC and $B\ C$ meet at P, CA and $C\ A$ meet at Q, and AB and $A\ B$ meet at R. Then P, Q and R are collinear.
2. The Fundamental Theorem is often used to simplify proofs of results in projective geometry, where the properties involved are projective properties. Generally, we do not explicitly refer to the corresponding auxiliary projective transformation t concerned, but simply comment that By the

Fundamental Theorem of Projective Geometry, we may choose the four Points . . . (no three of which are collinear) to be the triangle of reference and the unit Point; that is, to have homogeneous coordinates [1,0,0], [0,1,0], [0,0,1] and [1,1,1], respectively.

3. **Pappus' Theorem** Let A, B and C be three points on a line in $\mathbb{R}^2$, and let A , B and C be three points on another line. Let BC and $B\ C$ meet at P, CA and $C\ A$ meet at Q, and AB and $A\ B$ meet at R. Then P, Q and R are collinear.
4. Any Euclidean figure in an embedding plane corresponds to a projective figure in $\mathbb{RP}^2$. It follows that a Euclidean theorem concerned with projective properties (such as collinearity and coincidence) holds if and only if the corresponding projective theorem holds.
5. The **dual** of a statement about the projective properties of some figure in $\mathbb{RP}^2$ is the corresponding statement about $\mathbb{RP}^2$ in which the terms Point and Line are interchanged, and such other changes are made that ensure that the sentence makes sense.

   A triangle (three non-collinear Points and the Lines joining them) is self-dual.
6. A **hexagon** in $\mathbb{RP}^2$ consists of six Points joined by the six Lines joining them in turn.

   Pappus' Theorem (rephrased) Let the vertices A, B , C, A , B and C of a hexagon lie alternately on two different Lines. Then the Points of intersection of opposite sides $B\ C$ and BC , CA and $C\ A$, and AB and $A\ B$, are collinear.
7. **Brianchon's Theorem** (the dual of Pappus Theorem) Let the sides AB , $B\ C, CA$, $A\ B, BC$, $C\ A$ of a hexagon pass alternately through two (different) Points P and Q in $\mathbb{RP}^2$. Then the Lines joining opposite vertices A and A , B and B , C and C , are concurrent.
8. **Converse of Desargues' Theorem** (also its dual) Let two triangles be such that the Points through which corresponding sides pass are collinear. Then the Lines through the corresponding vertices of the two triangles are concurrent.

Section 3.5: Cross-Ratio

1. Let A, B, C, D be four collinear Points in $\mathbb{RP}^2$ represented by the position vectors $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, $\mathbf{d}$, and let $\mathbf{c} = \ \mathbf{a} + \ \mathbf{b}$% and $\mathbf{d} = \ \mathbf{a} + \ \mathbf{b}$. Then the **cross-ratio** of A, B, C, D is $(ABCD) = -\ \ -$.

   The cross-ratio $(ABCD)$ is independent of the homogeneous coordinates that are used to represent the collinear Points A, B, C, D.
2. Let A, B, C, D be four distinct collinear Points in $\mathbb{RP}^2$, and let $(ABCD) = k$. Then $(BACD) = (ABDC) = 1\ \ k$ and $(ACBD) = (DBCA) = 1\ \ \ k$.
3. Let t be a projective transformation, and let A, B, C, D be any four collinear Points in $\mathbb{RP}^2$. If $A\ \ = t(A)$, $B\ \ = t(B)$, $C\ \ = t(C)$, $D\ \ = t(D)$, then $(ABCD) = \ A\ B\ C\ D\ \ $.

4. Let A, B, C, D be four distinct Points on a Line, and let $A\ , B\ , C\ , D$ be four distinct Points on another Line such that $AA\ , BB\ , CC\ , DD$ all meet at a Point U. Then $(ABCD) = \ A\ B\ C\ D\ $.
5. **Unique Fourth Point Theorem** Let A, B, C, X, Y be collinear Points in $\mathbb{RP}^2$ such that $(ABCX) = (ABCY)$. Then $X = Y$.
6. Let A, B, C, D and A, E, F, G be two sets of collinear Points (on different Lines in $\mathbb{RP}^2$) such that the cross-ratios $(ABCD)$ and $(AEFG)$ are equal. Then the Lines BE, CF and DG are concurrent.
7. Let four collinear Points of $\mathbb{RP}^2$ pierce an embedding plane at the points A, B, C, D with position vectors **a**, **b**, **c**, **d**, respectively. Then, if we can write **c** and **d** in the form $\mathbf{c} = \ \mathbf{a} + (1 \quad)\,\mathbf{b}$ and $\mathbf{d} = \mu\mathbf{a} + (1 \quad \mu)\,\mathbf{b}$, we have
$$(ABCD) = \frac{1 \qquad}{\qquad} \quad \frac{1 \quad \mu}{\mu} = \frac{AC}{CB} \quad \frac{AD}{DB}.$$
8. **Strategy** To use an embedding plane to calculate the cross-ratio of four collinear Points:
    1. if the four Points pierce the embedding plane at A, B, C, D, then $(ABCD) = \frac{AC}{CB} \% \frac{AD}{DB}$;
    2. if one of the Points is an ideal Point for the embedding plane, then
$$\begin{aligned}(ABCD) &= \frac{DB}{CB} &&\text{if } A \text{ is ideal,}\\ (ABCD) &= \frac{CA}{DA} &&\text{if } B \text{ is ideal,}\\ (ABCD) &= \frac{BD}{AD} &&\text{if } C \text{ is ideal,}\\ (ABCD) &= \frac{AC}{BC} &&\text{if } D \text{ is ideal.}\end{aligned}$$
9. Cross-ratios can be used to measure distances on the ground from aerial photographs, since the cross-ratio of any four points on a line on the ground equals the cross-ratio of their images on the film.

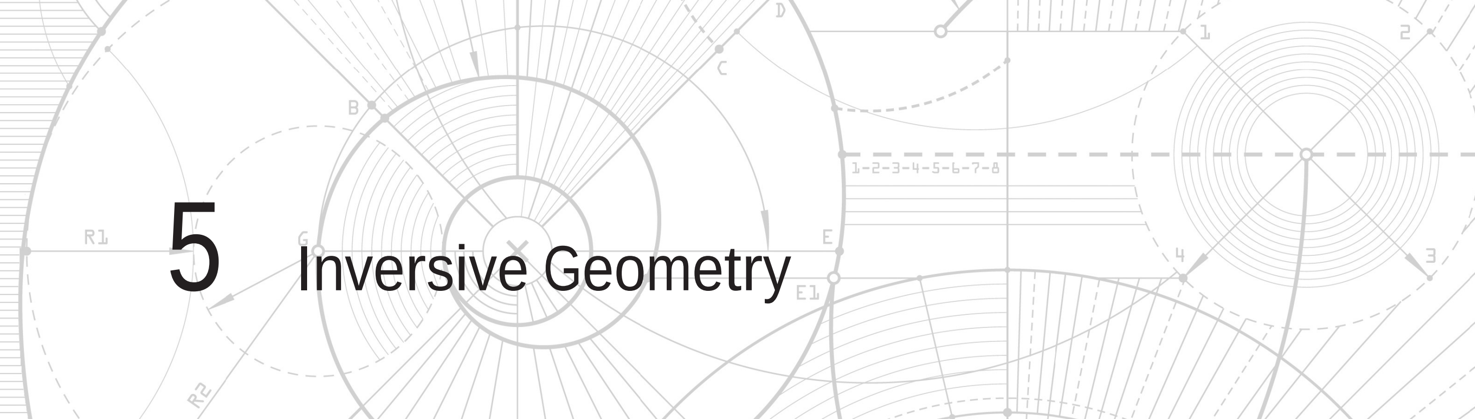

5 Inversive Geometry

In this chapter we introduce a geometry known as *inversive geometry* and use it to investigate circles and lines.

We also prove the *Apollonian Circles Theorem*: namely, that if A and B are two given points in the plane, then all the points P in the plane such that $PA : PB$ is a fixed ratio $k : 1$, for some positive real number $k = 1$, lie on a circle. Notice that if $k = 1$, the locus of points P such that $PA : PB = 1 : 1$ is simply the set of points equidistant from A and B, namely the perpendicular bisector of the segment AB.

This surprising fact was proved by the Greek geometer Apollonius of Perga (2nd century BC).

So for each pair of points A, B there is a family of circles known as the *circles of Apollonius*. Small values of k yield circles close to A, whereas large values yield circles close to B. It is sometimes helpful to regard the points A and B as point circles, corresponding to k being zero and infinity, respectively.

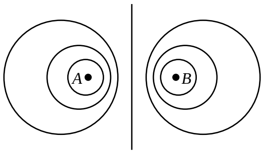

Apollonian circles

The key tool in our work is *inversion*, which is a generalization of the notion of reflection of points in a line. Just as reflection in a line maps points on one side of the line to points on the other, so inversion in a circle maps points inside the circle to points outside, and vice versa.

One difficulty that arises when we tackle problems like the Apollonian Circles Theorem is the presence of a line amongst what would otherwise be a family of circles. Clearly, it would be convenient if we could tackle such problems without having to treat the line as a special case. An ingenious way to do this is to think of the line as a circle of infinite radius in which the ends of the line have been joined by an additional point at infinity. This interpretation enables us to introduce the term *generalized circle* to mean either a line or a circle.

To make the idea of a point at infinity precise, we introduce a space known as the *extended plane*. This consists of the ordinary plane $\mathbb{R}^2$ together with an additional point that we define to be the *point at infinity*. This extended plane is the space in which we study inversive geometry. The transformations of this geometry, known as *inversive transformations*, are defined to be the composites of inversions.

Inversive transformations have the remarkable properties that they preserve the magnitude of angles between intersecting curves, and they map generalized circles onto generalized circles. These properties enable us to use inversive geometry to prove results like the Apollonian Circles Theorem.

In Section 5.1, we define inversion and study its basic properties.

In Section 5.2, we introduce the *extended plane*, and show how certain transformations of this plane can be represented using complex numbers. We then introduce the idea of a *generalized circle*, and show that every inversion preserves the magnitude of angles, and maps generalized circles onto generalized circles. We end the section by giving an interpretation of the extended plane in terms of the so-called *Riemann sphere*. On this sphere all generalized circles actually look like ordinary circles.

In Section 5.3, we formally define what is meant by an *inversive transformation* and *inversive geometry*. The group of all inversive transformations has a subgroup which consists of the so-called *Möbius transformations*. This subgroup is analogous to the subgroup of direct isometries in Euclidean geometry.

In Section 5.4, we prove the *Fundamental Theorem of Inversive Geometry:* namely, that any three points in the extended plane can be mapped onto any other three points by an inversive transformation. We use this result to show that, in inversive geometry, all generalized circles are congruent.

Finally, in Section 5.5 we use inversive geometry to prove the Apollonian Circles Theorem and to study various families of circles.

5.1 Inversion

5.1.1 Reflection and Inversion

We begin our exploration of inversive geometry by introducing a type of transformation known as *inversion*. These transformations will be used in Section 5.3 to define inversive geometry.

Roughly speaking, an inversion is a transformation of the plane that generalizes reflection in a line. Instead of mapping points from one side of a line to the other, an inversion maps points inside a circle to points outside the circle, and vice versa.

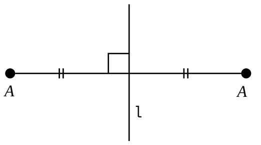

Recall that under reflection in a line ℓ, a point A is mapped to an image point A' that lies an equal distance from ℓ, but on the opposite side of ℓ. In order to generalize this notion of a reflection, we shall reformulate it in a way that provides us with a sensible analogue when the line ℓ is replaced by a circle C.

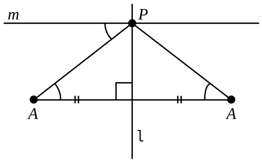

To do this, let m be a line parallel to AA' that crosses ℓ at some point P. Under reflection in ℓ, $\angle PAA'$ maps to $\angle PA'A$, so these two angles must be equal. But since the lines m and AA' are parallel, the angle $\angle PA'A$ is equal to the angle between PA' and m, so $\angle PAA'$ is equal to the angle between PA' and m. This is the clue that we need to generalize reflection.

By a stretch of the imagination, we can think of ℓ as an infinitely large circle with m lying along the radius through P. If we replace ℓ by a circle C of finite radius, and we replace m by a radial line that meets C at some point P, then, by analogy with reflection, we can define the image of A to be the point A' on the line segment OA for which $\angle OPA'$ is equal to $\angle PAO$. We say that A' is the point *inverse to* A *with respect to the circle* C.

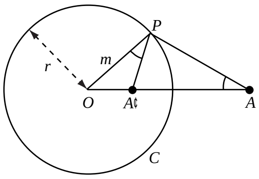

Of course, for this definition to work, we must check that for a given point A, the position of the point A' is independent of P. To do this, observe that the triangles $\triangle POA'$ and $\triangle AOP$ are similar, for they have a common angle at O and $\angle OPA' = \angle OAP$. It follows that

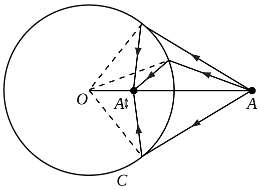

$$\frac{OA'}{OP} = \frac{OP}{OA},$$

so

$$OA \cdot OA' = OP^2.$$

But OP is equal to the radius r of C, so

$$OA \cdot OA' = r^2. \tag{1}$$

Since there is only one point A' on the line segment OA that satisfies equation (1), and since the equation does not depend on P, it follows that the position of A' does not depend on the choice of P.

Although the above construction illustrates how inversion can be defined as a generalization of reflection, it is worth noting that equation (1) is all that we need to determine the point A' that corresponds to a given point A. For simplicity, we shall therefore use equation (1) as our formal definition of inversion.

Definitions Let C be a circle with centre O and radius r, and let A be any point other than O. If A' is the point on the line OA that lies on the same side of O as A and satisfies the equation

$$OA \cdot OA' = r^2, \tag{2}$$

then we call A' the **inverse** of A with respect to the circle C. The point O is called the **centre of inversion**, and C is called the **circle of inversion**. The transformation t defined by

$$t(A) = A' \quad A \in \mathbb{R}^2 - \{O\}$$

is known as **inversion in** C.

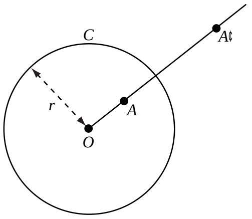

$OA \cdot OA' = r^2$

Remark

Since $OA \cdot OA' = r^2$ is non-zero, neither OA nor OA' can be zero, and so neither A nor A' can coincide with O. It is for this reason that O is excluded from the domain of the transformation t since there is no point to which O can be mapped. Likewise, there is no point that maps *to* O.

We can sometimes write down the inverse of a point directly from the above definition. For example, if C is the unit circle $\left\{(x,y) : x^2 + y^2 = 1\right\}$, then the inverse of (0, 2) with respect to C is the point $\left(0, \frac{1}{2}\right)$, and the inverse of $\left(-\frac{1}{3}, 0\right)$ is the point $(-3, 0)$.

Problem 1 Write down the inverse of each of the following points with respect to the circle of unit radius, centred at the origin.

(a) (4, 0) (b) (0, 1) (c) $\left(0, -\frac{1}{3}\right)$ (d) $\left(\frac{1}{4}, 0\right)$

Inversion distorts the plane considerably, for it maps points inside a circle C to points outside C, and vice versa. Indeed, if $OA < r$, then $OA' = r^2/OA > r$, whereas, if $OA > r$, then $OA' = r^2/OA < r$. Any point that lies on C maps to itself.

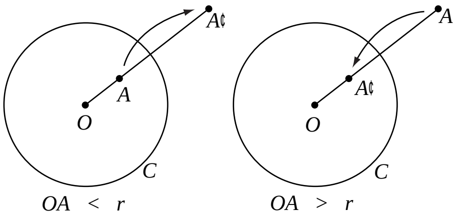

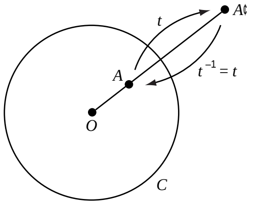

Note that if A' is the inverse of A, then A must be the inverse of A', for if $OA \cdot OA' = r^2$ then clearly $OA' \cdot OA = r^2$; we say that A and A' are **inverse points** with respect to C. In this sense, inversion is like reflection; if we reflect a point in a line and then reflect the reflection, we obtain the original point back again. Any transformation t that has this property is said to be *self-inverse* because it shows that t^{-1} exists and is equal to t.

Theorem 1 Inversion in a circle is a self-inverse transformation.

Since any transformation that has an inverse is one–one, it follows that every inversion is a one–one transformation. Remember, however, that since O is excluded from its domain, an inversion is a one–one transformation of $\mathbb{R}^2 - \{O\}$ onto itself.

Earlier, we mentioned that inversion can be regarded as a generalization of reflection. To see how this follows from the definition of inversion, we examine what happens to the inverse of a point A as we increase the radius of the circle.

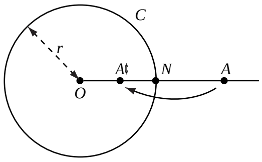

Let A be a point outside a circle C with centre O and radius r, let A' be the inverse of A with respect to C, and let the line segment AA' meet the circle at N. Then $OA = r + AN$ and $OA' = r - A'N$, so the equation $OA \cdot OA' = r^2$ becomes

$$(r + AN)(r - A'N) = r^2.$$

After expanding the brackets, cancelling the r^2 terms, and solving for $A'N$, we obtain

$$A'N = \frac{AN \bullet r}{r + AN} = \frac{AN}{1 + AN/r}.$$

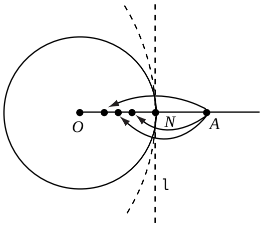

Now fix A and N, and let the radius r of the circle tend to infinity. As it does so, it follows from the above equation that the length of $A'N$ tends to the length of AN. In other words, reflection in a line can be regarded as the limiting case of inversion in circles of increasing radii. For this reason, we adopt the following useful convention.

Convention We use the term **inversion** to mean either reflection in a line or inversion in a circle.

The following gives a geometric method for constructing inverse points with respect to a given circle.

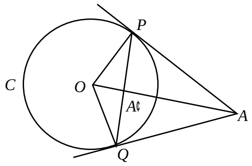

Let A be a point outside a circle C with centre O and radius r, let AP and AQ be the two tangents from A to C, and let A' be the point of intersection of OA and PQ. Then A and A' are inverse with respect to C.

For, in the triangles $\triangle OPA'$ and $\triangle OAP$, the angles at O are equal and the two angles $\angle OA'P$ and $\angle OPA$ are equal, since they are both right angles; hence all the angles in the two triangles are equal. Thus the triangles are similar; hence, in particular, we must have that

$$\frac{OA}{OP} = \frac{OP}{OA'},$$

so that

$$OA \bullet OA' = OP^2 = r^2.$$

Thus A and A' are inverse with respect to C, as claimed.

We end this subsection by showing that an inversion in a circle is a different type of transformation from those that arise in Euclidean and affine geometry.

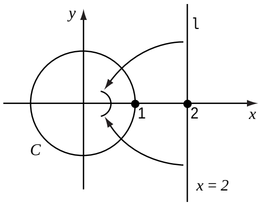

First, an inversion in a circle is not an affine transformation since it does not map lines to lines. For example, let C be the unit circle and ℓ be the line with equation $x = 2$. All the points on ℓ lie outside the unit circle, and so inversion in C must map all the points of ℓ to points that lie inside the unit circle. In particular, the image of ℓ cannot be a line. Thus, since the image of a line under any affine transformation is itself a line, it follows that inversion cannot be an affine transformation.

An inversion in a circle is not a Euclidean transformation either, for like affine transformations, Euclidean transformations map lines to lines. Indeed, a Euclidean transformation is just a special type of affine transformation.

5.1.2 The Effect of Inversion on Lines and Circles

In the previous subsection you saw how to construct the point A' that is the inverse of a given point A with respect to a given circle C. However, in order to study many of the properties of inversion, we require an *algebraic* formula that relates the coordinates of A and A'.

For our present purposes it is sufficient to derive the formula for the case where $\mathscr{C}$ is the unit circle $\{(x, y) : x^2 + y^2 = 1\}$. This circle will occur so frequently in our work that we denote it by the special symbol C.

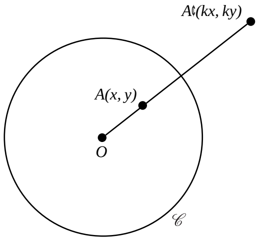

Let A be a point $(x, y) \in \mathbb{R}^2 - \{O\}$, and let A' be its image under inversion in the unit circle C. Since A' lies on the same half-line from the origin as does A, it follows that A' must have coordinates (kx, ky), for some positive number k.

Since the radius of C is 1, we must have $OA \cdot OA' = 1$. Thus $OA^2 \cdot OA'^2 = 1$, and so

$$\left(x^2 + y^2\right)\left(k^2x^2 + k^2y^2\right) = 1.$$

It follows that

$$k^2 = \frac{1}{\left(x^2 + y^2\right)^2},$$

and hence that

$$k = \frac{1}{x^2 + y^2}.$$

Thus A' is the point $\left(\frac{x}{x^2+y^2}, \frac{y}{x^2+y^2}\right)$. We therefore have the following algebraic description of inversion in the unit circle.

Theorem 2 Inversion in the unit circle C is the function

$$t : (x, y) \mapsto \left(\frac{x}{x^2 + y^2}, \frac{y}{x^2 + y^2}\right) \quad \left((x, y) \in \mathbb{R}^2 - \{O\}\right).$$

We may use this theorem to find the image of any non-zero point of $\mathbb{R}^2$ under inversion in the unit circle C. For example, the image of $(3, -2)$ is the point

$$\left(\frac{3}{3^2 + (-2)^2}, \frac{-2}{3^2 + (-2)^2}\right) = \left(\tfrac{3}{13}, -\tfrac{2}{13}\right).$$

Problem 2 Determine the image of each of the following points under inversion in C.

(a) (4,1) (b) $\left(\frac{1}{2}, -\frac{1}{4}\right)$

Now let (x', y') be the image of (x, y) under inversion in C. Since inversion is self-inverse, it follows that

$$(x, y) = \left(\frac{x'}{(x')^2 + (y')^2}, \frac{y'}{(x')^2 + (y')^2}\right).$$

Using this relationship between (x, y) and (x', y'), we can find the image of a curve under inversion in $\mathscr{C}$ in much the same way as we found the image of a curve in affine geometry. The main difference is that we must be careful to remember that inversion is not defined at the origin (the centre of inversion) and that no point is mapped to the origin.

We shall return to the question of what happens at the origin later, in Subsection 5.2.3.

Example 1 Determine the image under inversion in $\mathscr{C}$ of the line $2x + 4y = 1$.

Solution Let (x, y) be an arbitrary point on the line $2x + 4y = 1$, and let (x', y') be the image of (x, y) under inversion in $\mathscr{C}$. Then

$$(x, y) = \left(\frac{x'}{(x')^2 + (y')^2}, \frac{y'}{(x')^2 + (y')^2}\right).$$

Since x and y are related by the equation $2x + 4y = 1$, it follows that x' and y' are related by the equation

$$\frac{2x'}{(x')^2 + (y')^2} + \frac{4y'}{(x')^2 + (y')^2} = 1.$$

Multiplying by $(x')^2 + (y')^2$, we obtain $2x' + 4y' = (x')^2 + (y')^2$, and by completing the square we may write this as

$$(x' - 1)^2 + (y' - 2)^2 = 5.$$

Dropping the dashes, we see that points on the image must have (x, y)-coordinates which satisfy the equation

$$(x - 1)^2 + (y - 2)^2 = 5.$$

This is the equation of a circle with centre (1, 2) and radius $\sqrt{5}$. Since this passes through the origin, and since the origin cannot be part of the image under the inversion, the required image must be the circle with the origin removed. □

For every point on $(x - 1)^2 + (y - 1)^2 = 5$ other than (0, 0) is the image of some point on $2x + 4y = 1$.

In this example we used dashes to distinguish the coordinates of a point (x, y) from the coordinates of its image (x', y'). However, once you have understood the method you may prefer to adopt the strategy below. This uses the same method as in the example, but it drops the dashes throughout.

Strategy To determine an equation for the image of a curve under inversion in the unit circle $\mathscr{C}$:

1. write down an equation that relates the x- and y-coordinates of points on the curve;
2. replace x by $\frac{x}{x^2+y^2}$ and y by $\frac{y}{x^2+y^2}$, and simplify the resulting equation.

When using this strategy to find the image of a curve that passes through the origin, we must first remove the origin from the curve.

We may also have to exclude the origin from the image.

Example 2 Determine the image under inversion in C of the line $y = x$ (with the origin removed).

Solution Replacing x by $\frac{x}{x^2+y^2}$ and y by $\frac{y}{x^2+y^2}$, we obtain

$$\frac{y}{x^2 + y^2} = \frac{x}{x^2 + y^2}.$$

Hence

$$y = x.$$

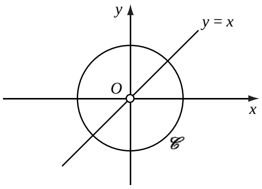

This is the line we started with. Just as the origin had to be excluded from that line before we could find its image, so the origin has to be excluded from the image. Hence the line $y = x$ (with the origin removed) maps to itself under inversion in C. □

We shall use the term **punctured** to describe a line or circle that has one of its points removed. More specifically, if A is the point removed from a line or circle, then the line or circle is said to be **punctured at** A.

Problem 3 Determine the image under inversion in C of the line $y = 2x$ punctured at the origin.

Problem 4 Let be the line with equation $x + y = 1$.

(a) Determine the image of under inversion in C.
(b) Explain why the points (1, 0), (0, 1) lie both on and on its image.
(c) Sketch and its image on a single diagram.

The solutions to these problems and the preceding examples illustrate the following general result.

Theorem 3 Images of Lines under Inversion
Under inversion in a circle with centre O:

(a) a line that does not pass through O maps onto a circle punctured at O;
(b) a line punctured at O maps onto itself.

Proof First choose a pair of coordinate axes with origin at O, and choose a unit of length equal to the radius of the circle in which we are inverting. Then the circle in which we are inverting becomes the unit circle C.

(a) If is a line that does not pass through the origin, then it has an equation of the form

$$ax + by + c = 0,$$

where c is non-zero. Using the above strategy, we know that the image of ℓ under inversion in $\mathscr{C}$ has equation

$$\frac{ax}{x^2+y^2} + \frac{by}{x^2+y^2} + c = 0.$$

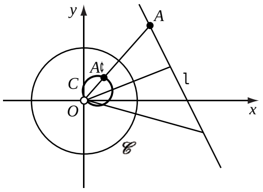

Since c is non-zero, we may rewrite this in the form

$$x^2 + y^2 + (a/c)x + (b/c)y = 0.$$

This is the equation of a circle C through the origin. If the origin is removed from this circle, then each remaining point A' is the image of the point A at which OA' intersects ℓ. It follows that the image of ℓ is the whole of the punctured circle $C - \{O\}$.

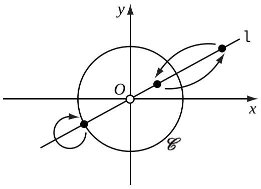

(b) Although we could use the strategy to prove this part as well, it is easier to work directly from the definition of inversion. Indeed, if ℓ is a line punctured at O, then each point of ℓ inside $\mathscr{C}$ is the image of a point of ℓ outside $\mathscr{C}$, and each point of ℓ outside $\mathscr{C}$ is the image of a point of ℓ inside $\mathscr{C}$. Points of ℓ that lie on $\mathscr{C}$ map to themselves. It follows that the punctured line ℓ maps onto itself, as required.

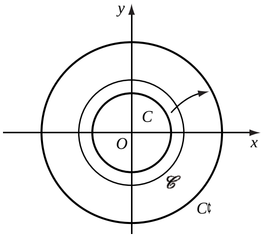

Next, we consider the images of circles under inversion. Since points on the circle of inversion map to themselves, the image of that circle is also a circle. Also, any circle C with centre the centre of inversion O must map onto another circle C' with centre O, for, by symmetry, every point of C is mapped an equal distance along a radial line.

This raises the question as to whether an inversion always maps circles to circles.

Example 3 Use the above strategy to determine the image under inversion in $\mathscr{C}$ of the circle C with centre (2, 0) and radius 1.

Solution The circle C has equation $(x-2)^2+y^2=1$, which we may rewrite in the form

$$x^2+y^2-4x+3=0.$$

Using the above strategy, we deduce that the image of C under inversion in $\mathscr{C}$ has equation

Note that the origin does not lie on C.

$$\left(\frac{x}{x^2+y^2}\right)^2 + \left(\frac{y}{x^2+y^2}\right)^2 - \frac{4x}{x^2+y^2} + 3 = 0.$$

We may add together the first two terms of this equation to obtain

$$\frac{1}{x^2+y^2} - \frac{4x}{x^2+y^2} + 3 = 0,$$

which we may rearrange in the form

$$x^2 + y^2 - \tfrac{4}{3}x + \tfrac{1}{3} = 0.$$

By completing the square we obtain

$$\left(x - \tfrac{2}{3}\right)^2 + y^2 = \left(\tfrac{1}{3}\right)^2.$$

This is the equation of a circle with centre $\left(\tfrac{2}{3}, 0\right)$ and radius $\tfrac{1}{3}$. □

So in this example the circle C does indeed map to another circle. Notice, however, that the centre (2, 0) of C maps to $(\tfrac{1}{2}, 0)$ which is *not* the centre of the image of C. It follows that even if an inversion maps one circle onto another, it may not map the centres to each other.

Problem 5 Determine the image under inversion in $\mathscr{C}$ of the circle with centre (2, 2) and radius 1.

The next example illustrates what happens when we use the strategy to find the image of a circle that passes through the origin.

Example 4 Let C be the circle with centre $(-2, 0)$ and radius 2, punctured at the origin. Determine the image of C under inversion in $\mathscr{C}$.

Solution The circle C has equation $(x + 2)^2 + y^2 = 2^2$, which we may rewrite in the form

$$x^2 + y^2 + 4x = 0.$$

Using the above strategy, we deduce that the image of C under inversion in $\mathscr{C}$ has equation

$$\left(\frac{x}{x^2 + y^2}\right)^2 + \left(\frac{y}{x^2 + y^2}\right)^2 + \frac{4x}{x^2 + y^2} = 0.$$

Adding together the first two terms of this equation, we obtain

$$\frac{1}{x^2 + y^2} + \frac{4x}{x^2 + y^2} = 0,$$

which we may rearrange in the form $1 + 4x = 0$.

It follows that the image of the punctured circle C is the line ℓ with equation $x = -\tfrac{1}{4}$. From the figure in the margin it is clear that every point of ℓ is the image of some point on C, so there is no need to puncture ℓ. □

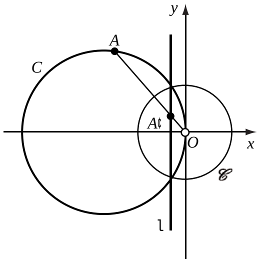

Problem 6 Let C be the circle with centre $\left(0, -\tfrac{1}{4}\right)$ and radius $\tfrac{1}{4}$, punctured at the origin. Determine the image of C under inversion in $\mathscr{C}$.

The conclusions of Examples 3 and 4, and of Problems 5 and 6, suggest the following result.

Theorem 4 Images of Circles under Inversion
Under inversion in a circle with centre O:

(a) a circle that does not pass through O maps onto a circle;
(b) a circle punctured at O maps onto a line that does not pass through O.

Proof As for Theorem 3, we choose a pair of coordinate axes that makes the circle in which we are inverting the unit circle $\mathcal{C}$.

Now let C be an arbitrary circle with centre (a, b) and radius r. This has equation

$$(x - a)^2 + (y - b)^2 = r^2,$$

which we may rewrite in the form

$$x^2 + y^2 - 2ax - 2by + c = 0,$$

Note that this passes through O if and only if $c = 0$.

where $c = a^2 + b^2 - r^2$. Using the strategy for determining the images of curves under inversion, we deduce that the image of C under inversion has equation

$$\left(\frac{x}{x^2+y^2}\right)^2 + \left(\frac{y}{x^2+y^2}\right)^2 - \frac{2ax}{x^2+y^2} - \frac{2by}{x^2+y^2} + c = 0.$$

Adding together the first two terms of this equation, we obtain

$$\frac{1}{x^2+y^2} - \frac{2ax}{x^2+y^2} - \frac{2by}{x^2+y^2} + c = 0.$$

By multiplying this equation by $\left(x^2 + y^2\right)$, we may rearrange it in the form

$$1 - 2ax - 2by + c\left(x^2 + y^2\right) = 0. \qquad (3)$$

This is the equation of either a line or a circle, depending on whether or not C passes through O.

(a) If C does not pass through O, then c is non-zero. We can therefore divide equation (3) by c to obtain

$$x^2 + y^2 - 2\tfrac{a}{c}x - 2\tfrac{b}{c}y + \tfrac{1}{c} = 0.$$

This is the equation of a circle on which the image of C must lie.

(b) If C does pass through O, then $c = 0$, so equation (3) becomes

$$1 \quad 2ax \quad 2by = 0.$$

This is the equation of a line that does not pass through O.

The following box summaries the results of Theorem 3 and 4.

Under inversion with respect to a circle with centre O:

a line punctured at O	*maps onto*	the same line punctured at O;
a line not through O	*maps onto*	a circle punctured at O;
a circle punctured at O	*maps onto*	a line not through O;
a circle not through O	*maps onto*	a circle not through O.

If the circle of inversion is the unit circle C, then O is the origin.

There is no need to remember the details of this summary since its predictions can be recalled intuitively as follows. First, points on a line or circle through the origin can be chosen arbitrarily close to the origin. The images of such points can therefore be chosen arbitrarily far from the origin, and must therefore lie on a line. Secondly, points on a line can be chosen arbitrarily far from the origin. The images of these points can be chosen arbitrarily close to the origin and must therefore lie on a circle or line punctured at the origin.

With a little practice it is easy to use the summary to simplify the work needed to determine the image of a circle or line under an inversion.

Example 5 Determine the image of each of the following under inversion in the unit circle C:

(a) the line with equation $x = 2$;
(b) the circle C with centre (0, 2) and radius 1.

Solution

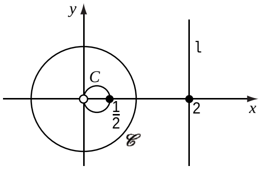

(a) From the summary, we know that maps to a circle C punctured at the origin. This circle passes through the point $\frac{1}{2}, 0$, since $\frac{1}{2}, 0$ is the image of the point (2, 0) on the line. Since is symmetrical about the x-axis, it follows that the image circle must also be symmetrical about the x-axis. The only circle C that fulfils all these criteria is the circle with radius $\frac{1}{4}$ and centre $\frac{1}{4}, 0$.

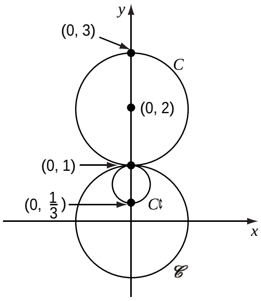

(b) From the summary, the image C of C is a circle that does not pass through the origin. It must be symmetrical about the y-axis (because C is), and it must pass through the points $0, \frac{1}{3}$ and (0, 1) (the images of (0, 3) and (0, 1), respectively). The only circle C that fulfils all these criteria is the circle with radius $\frac{1}{3}$ and centre $0, \frac{2}{3}$. □

Problem 7 Determine the image of each of the following under inversion in the unit circle $\mathcal{C}$:

(a) the line ℓ with equation $y = 1$;
(b) the circle C with centre (0, 1) and radius 1 (punctured at the origin).

5.1.3 The Effect of Inversion on Angles

In this subsection we shall show that inversion preserves the magnitude of the angle at which two curves meet. First, however, we must clarify what it means to measure the angle between two intersecting curves.

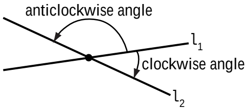

In the case of two intersecting lines ℓ_1 and ℓ_2 there are two ways in which we can measure the angle from ℓ_1 to ℓ_2. We can measure it either in a clockwise direction, or in an anticlockwise direction, as shown in the margin. Clearly, the magnitude of the angle depends on the direction we choose, so when specifying an angle we must give both its magnitude and direction.

In the case of two intersecting *curves*, we define the angle between the curves by using tangents, as follows.

Definitions Let c_1 and c_2 be two curves that intersect at the point A, and let the tangents to the curves at A be ℓ_1 and ℓ_2, respectively. Then the **anticlockwise angle** from c_1 to c_2 at A is the anticlockwise angle from ℓ_1 to ℓ_2, and the **clockwise angle** from c_1 to c_2 at A is the clockwise angle from ℓ_1 to ℓ_2.

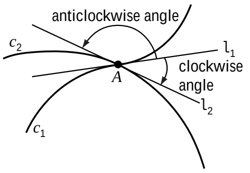

To examine what happens to the angles between two curves under an inversion, it is sufficient to examine what happens to the angles between the corresponding tangents.

For the moment let us concentrate on what happens to a single line ℓ under inversion in a circle centred at O. We know that ℓ maps onto a circle C punctured at O. But we can say more, for if m is the line through O that is perpendicular to ℓ, then ℓ is symmetrical about m. It follows that the circle C is symmetrical about m, and so ℓ is parallel to the tangent to C at O. We state this result as the Symmetry Lemma.

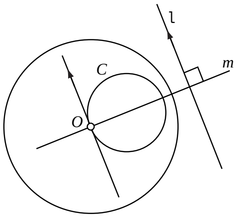

Lemma 1 Symmetry Lemma
Let ℓ be a line that does not pass through the point O. Then under inversion in a circle with centre O, ℓ maps to a circle C (punctured at O), and the tangent to C at O is parallel to ℓ.

Now consider what happens to the angle between two lines ℓ_1 and ℓ_2 which intersect at some point A other than O, as shown below. For the moment we shall assume that neither line passes through O, so that under the inversion the lines map to punctured circles C_1 and C_2, respectively. These punctured circles meet at the point A', where A' is the image of A. We want to compare

We ask you to investigate what happens when one of the lines passes through O in Problem 8.

the angle from ℓ_1 to ℓ_2 at A' with the angle from C_1 to C_2 at A. The trick we use is to compare both angles with the angle from C_1 to C_2 at O.

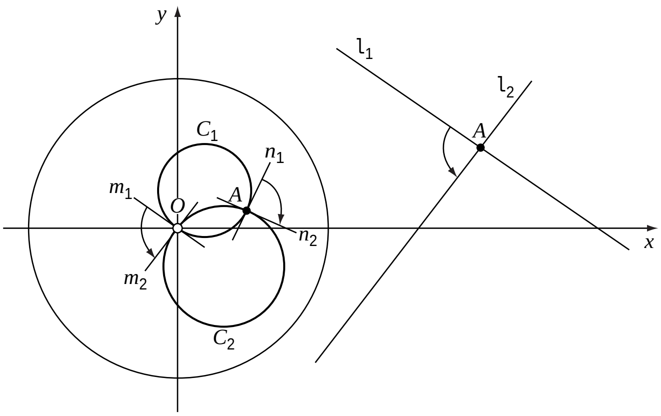

By the Symmetry Lemma, ℓ_1 is parallel to the tangent m_1 to C_1 at O, and ℓ_2 is parallel to the tangent m_2 to C_2 at O. It follows that the angle from ℓ_1 to ℓ_2 must be equal in magnitude and direction to the angle we have shown from m_1 to m_2.

Next observe that the reflection in the line through the centres of C_1 and C_2 sends the tangents m_1, m_2 at O to the tangents n_1, n_2 at A. Since the reflection preserves the magnitude of an angle but changes its orientation, we conclude that the angle from n_1 to n_2 at A must be equal in magnitude but *opposite* in orientation to the angle from m_1 to m_2 at O.

Overall, we have shown that the angle from C_1 to C_2 at A must be equal in magnitude but opposite in orientation to the angle from ℓ_1 to ℓ_2 at A'.

We sometimes abbreviate this by saying that the angle at A' is *equal but opposite* to the angle at A.

Theorem 5 Angle Theorem
An inversion in any circle preserves the magnitude of angles between curves but reverses their orientation.

The next problem asks you to complete our proof of this theorem.

Problem 8 Prove the Angle Theorem in the case where one of the lines passes through the centre of inversion.

The Angle Theorem provides us with a very powerful tool for locating the images of two or more circles or lines under inversion.

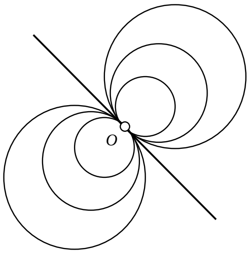

Example 6 A family of circles shares a common tangent at the origin O, as shown in the margin. Describe the effect of inverting the family of circles (all punctured at O) in the unit circle C. Illustrate your answer with a sketch.

Solution Let be the common tangent to the circles, as shown on the left below. If d is the punctured line through O perpendicular to , then each of the circles crosses d at right angles. Since all the circles are punctured at O, their images under the inversion must be straight lines, as shown on the right. By the Angle Theorem, these straight lines must cross the image of d at right angles. But d maps onto itself under the inversion, so the punctured circles map to a family of parallel lines perpendicular to d.

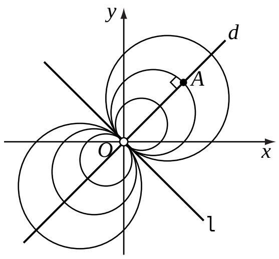

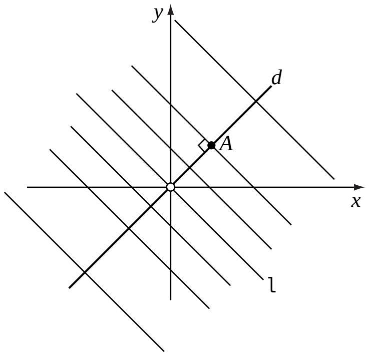

In fact, apart from , every line perpendicular to d must be the image of a punctured circle in the family. For if a perpendicular line meets d at some point A , then it must be the image of the punctured circle that passes through the point A which is mapped to A under the inversion. □

Problem 9 One family of circles touches the x-axis at the origin O, and another family of circles touches the y-axis at O. Describe the effect of inverting the two families of circles (all punctured at O) in the unit circle C. Illustrate your answer with a sketch.

Problem 10 A family of circles intersects at the origin O and at another point A, as shown in the margin. Describe the effect of inverting the family of circles (all punctured at O) in the unit circle C. Illustrate your answer with a sketch.

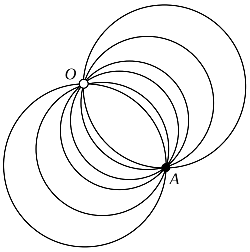

Problem 11 Let C_1, C_2 and C_3 be circles in the plane such that C_1 and C_2 touch at the origin O, C_3 and C_1 touch at another point A, and C_2 and C_3 touch at the further point B. Describe the effect of inverting the circles C_1, C_2 (both punctured at O) and C_3 in the unit circle C. Illustrate your answer with a sketch.

These problems illustrate that, by carefully choosing centres of inversion, we can transform some of the circles in a figure to straight lines. Since straight lines are often easier to deal with than are circles, we can use such transformed figures to investigate those properties of the original figures that are preserved by inversions. In essence, this is the idea that underlies inversive geometry, and you will see several example of its use in Section 5.5.

5.2 Extending the Plane

In order to define a geometry in which we can study the properties of circles, lines and angles, we need a group of transformations that preserve these properties. Among the transformations that do this are the Euclidean transformations, and the inversions introduced in the previous section. In this section we describe how such transformations can be represented in terms of complex numbers. This will enable us to manipulate the transformations by using the algebra of complex numbers.

5.2.1 Transformations of the Complex Plane

We begin by reminding you of some facts concerning complex numbers.

First, there is a one one correspondence between points (x, y) in the plane $\mathbb{R}^2$ and complex numbers $z = x + iy$ in the complex plane $\mathbb{C}$; we call x and y the **real part** and the **imaginary part** of the complex number z, and denote them by the symbols Re z and Im z, respectively. All the arithmetic operations may be carried out in $\mathbb{C}$ as for real numbers, except that we replace i^2 by 1 wherever i^2 occurs.

Recall that $x + iy$ is the *Cartesian* form of the complex number.

If z is the complex number $x + iy$, then its **conjugate** $\bar{z}$ is defined by

$$\bar{z} = x \quad iy,$$

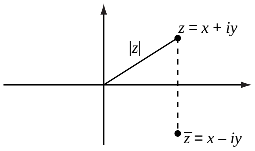

and its **modulus** $|z|$ is defined by

$$|z| = \sqrt{x^2 + y^2}.$$

Recall that $|z|^2 = z\bar{z}$.

Problem 1 Let $z_1 = 2 \quad 3i$ and $z_2 = \quad 3 + 4i$.

(a) Determine each of the following complex numbers in Cartesian form.

(i) $z_1 + z_2$ (ii) $z_1 \quad z_2$ (iii) $z_1 z_2$
(iv) z_1 / z_2 (v) $\overline{z_1}$ (vi) $\overline{z_2}$

(b) Determine $|z_1|$ and $|z_2|$.

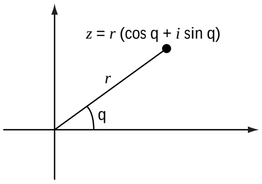

If a non-zero complex number $z = x + iy$ has modulus r, and if the position vector of the point (x, y) lies at an angle to the positive x-axis, then we can express z in the form

$$r(\cos \quad + i \sin \quad).$$

Such an expression is known as a **polar form** of z, and the angle θ is known as an **argument** of z, written $\arg z$. Polar forms and arguments are not unique, since the same z could equally well be expressed in the form

$$r(\cos(\theta + 2\pi n) + i\sin(\theta + 2\pi n))$$

for any integer n. The **principal argument** of z, written $\operatorname{Arg} z$, is the unique value of $\arg z$ that lies in the interval $(-\pi, \pi]$.

We can obtain the Cartesian form of a complex number from any of its polar forms by using the equations

$$x = r\cos\theta \quad \text{and} \quad y = r\sin\theta,$$

and we can obtain a polar form of a complex number from its Cartesian form by using the equations

$$r = \sqrt{x^2 + y^2}, \quad \cos\theta = \frac{x}{r} \quad \text{and} \quad \sin\theta = \frac{y}{r}.$$

Problem 2 Determine the polar forms of the complex numbers $z_1 = 1 - i$ and $z_2 = -\sqrt{3} + i$ in terms of their principal arguments.

The following strategy can be used to multiply and divide complex numbers given in polar form.

Strategy To multiply two complex numbers given in polar form, multiply their moduli and add their arguments.

To divide two complex numbers given in polar form, divide their moduli and subtract their arguments.

Although this strategy gives an argument for a product or quotient, you may require the principal argument, in which case you will need to adjust the argument by adding an appropriate multiple of 2π.

Problem 3 Let $z_1 = 1 - i$ and $z_2 = -\sqrt{3} + i$. Determine the polar forms of the following complex numbers in terms of their principal arguments.

(a) $z_1 z_2$ (b) z_1/z_2

We can now use the algebra of complex numbers described above to represent some of the basic Euclidean transformations of the plane.

Translations

First, consider the transformation

$$t(z) = z + c \quad (z \in \mathbb{C}), \tag{1}$$

where $c = a + ib$. This maps an arbitrary point $x + iy \in \mathbb{C}$ to the point

$$(x + a) + i(y + b),$$

and therefore corresponds to a translation through the vector (a, b). Clearly, such transformations preserve angles, and map circles and lines to circles and lines.

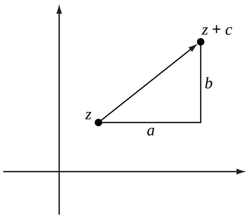

Reflection in the x-Axis

Next, consider the transformation

$$t(z) = \bar{z} \quad (z \in \mathbb{C}). \tag{2}$$

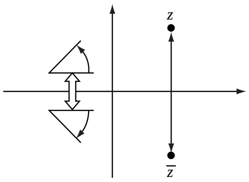

This maps an arbitrary point $x + iy \in \mathbb{C}$ to the point $x - iy$, and therefore corresponds to a *reflection* in the x-axis. It maps circles and lines to circles and lines, and it preserves the magnitude of angles; however, the orientation of angles is reversed.

Rotation about the Origin

Now, consider the transformation

$$t(z) = az \quad (z \in \mathbb{C}), \tag{3}$$

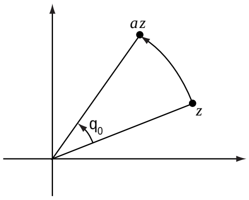

where $|a| = 1$. Since $|a| = 1$ we can write $a = \cos\theta_0 + i\sin\theta_0$, where $\theta_0 = \operatorname{Arg} a$. From the above strategy, t maps an arbitrary point $r(\cos\theta + i\sin\theta)$ in $\mathbb{C}$ to the point

$$r(\cos(\theta + \theta_0) + i\sin(\theta + \theta_0)),$$

and therefore corresponds to a rotation through the angle $\theta_0 = \operatorname{Arg} a$ about the origin. The rotation is clockwise if $\operatorname{Arg} a < 0$ and anticlockwise if $\operatorname{Arg} a > 0$.

Arbitrary Isometries

All the other isometries can be represented in the complex plane as composites of the basic transformations described above. In fact, we have the following result.

Theorem 1 Each isometry t of the plane can be represented in the complex plane by one of the functions

$$t(z) = az + b \quad \text{or} \quad t(z) = a\bar{z} + b,$$

where $a, b \in \mathbb{C}$, $|a| = 1$. Conversely, all such functions represent isometries.

Proof The converse is easy to prove, because every function of the type described is a composite of the basic isometries described above.

So let t be an isometry of the complex plane, and let $t(0) = b$, $t(1) = c$. If we denote $c - b$ by a, then $|a|$ is the distance between $t(0)$ and $t(1)$, and since t is an isometry it follows that $|a| = 1$. Now let s be the isometry defined by $s(z) = az + b$; then

$$s(0) = b = t(0) \quad \text{and} \quad s(1) = a + b = c = t(1).$$

Thus $s^{-1} \circ t$ is an isometry that fixes 0 and 1, and so, since it is an isometry, it must fix each point of the x-axis.

Then using the fact that $s^{-1} \circ t$ is an isometry (that is, it does not alter distances) that leaves each point of the x-axis unaltered, we may deduce that $s^{-1} \circ t$ is EITHER the identity transformation OR a reflection in the x-axis. If it is the identity, then $s^{-1} \circ t(z) = z$, in which case $t(z) = s(z) = az + b$. If it is a reflection, then $s^{-1} \circ t(z) = \bar{z}$, in which case $t(z) = s(\bar{z}) = a\bar{z} + b$.

Theorem 1 may be interpreted as saying that every isometry of the complex plane can be obtained as a rotation (through an angle Arg a) followed by a translation (through the vector (Re b, Im b)), possibly all preceded by a reflection in the real axis.

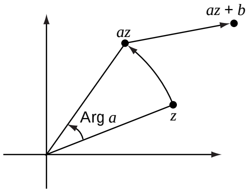

Slightly more surprising is the fact that every isometry of the plane can be expressed as a composite of reflections alone. This is because each rotation and each translation can be expressed as a composite of two reflections.

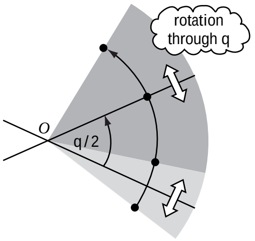

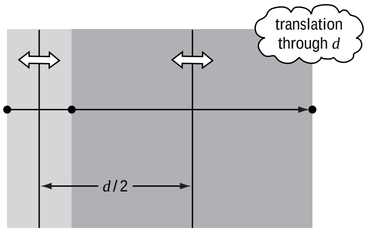

Thus a rotation about a point O can be expressed as a reflection in any line through O followed by a reflection in a second line through O, where the angle from the first line to the second line is half the desired angle rotation.

Similarly, a translation can be expressed as a reflection in any line perpendicular to the direction of the translation followed by a reflection in a second line that is parallel to the first and in the direction of the translation relative to the first line. The distance between the lines is half the distance of the translation.

So we have the following result.

Theorem 2 Every isometry can be expressed as a composite of reflections.

Example 1 Let t be the isometry defined by

$$t(z) = i\bar{z} + 4 + 2i \quad (z \in \mathbb{C}).$$

(a) Show that t represents an isometry.
(b) Interpret t as the composite of a reflection, a rotation and a translation.
(c) Interpret t as a composite of reflections.

Solution

(a) The coefficient of $\bar{z}$ is i, which has modulus $|i| = 1$. By Theorem 1, it follows that t is an isometry.

(b) In the formula that defines t, the conjugation corresponds to a reflection in the x-axis, the multiplication by i corresponds to an anticlockwise rotation through $\pi/2$, and the addition of $4+2i$ corresponds to a translation through the vector (4, 2).

(c) First, let r be the reflection in the x-axis that corresponds to the conjugation.

Next, observe that the anticlockwise rotation through $\pi/2$ can be interpreted as the composite $r_1 \circ r$, where r is the reflection in the x-axis again, and r_1 is the reflection in the line $y = x$ through the origin that makes an angle $\pi/4$ with the x-axis.

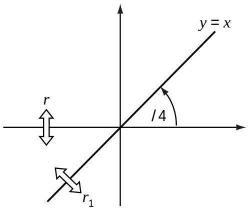

Finally, observe that the translation through the vector (4, 2) can be interpreted as the composite $r_3 \circ r_2$, where r_2 is the reflection in the line $4x + 2y = 0$ through the origin that is perpendicular to the vector (4, 2), and r_3 is the reflection in the parallel line $4x + 2y = 10$ that passes through $\frac{1}{2}(4,2) = (2,1)$.

Overall, we have $t = r_3 \circ r_2 \circ r_1 \circ r \circ r$ or, since r is its own inverse, $t = r_3 \circ r_2 \circ r_1$. □

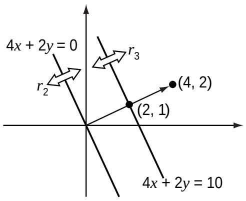

The decompositions illustrated above generalize as follows.

Problem 4 Let t be the transformation defined by

$$t(z) = -i\bar{z} + 6 - 4i \quad (z \in \mathbb{C}).$$

(a) Show that t represents an isometry.
(b) Interpret t as the composite of a rotation and a translation.
(c) Interpret t as a composite of reflections.
Hint: Use the decompositions described in the margin.

To *rotate* the plane $\mathbb{R}^2$ through an angle θ: first reflect in the line $\{z : \operatorname{Arg} z = 0\}$, then reflect in the line $\{z : \operatorname{Arg} z = \frac{1}{2}\theta\}$.

To *translate* the plane $\mathbb{R}^2$ through a vector (a, b): first reflect in the line $\{(x, y) : ax + by = 0\}$, then reflect in the line $\{(x, y) : ax + by = \frac{1}{2}(a^2 + b^2)\}$.

Having discussed isometries, we now turn our attention to two other transformations of the complex plane that preserve the magnitude of angles and map circles and lines to circles and lines.

Scalings

The transformation defined by

$$t(z) = kz \quad (z \in \mathbb{C}), \tag{4}$$

where k is real and positive, multiplies the modulus of each complex number by a factor k but leaves its argument unchanged. It is therefore a *scaling* by the factor k. Clearly, scalings preserve angles, and map circles and lines to circles and lines.

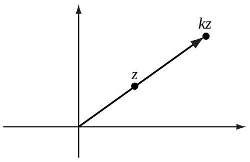

Inversions

The primary reason for introducing complex numbers into our discussion is that they provide a particularly convenient way in which to represent the effect of an inversion on points in the plane.

Theorem 3 An inversion in a circle C of radius r with centre (a, b) may be represented in the complex plane by the transformation

$$t(z) = \frac{r^2}{\overline{z - c}} + c \quad (z \in \mathbb{C} - \{c\}),$$

where $c = a + ib$.

Proof We first consider the case where C is the unit circle $\mathcal{C}$. The image under inversion in $\mathcal{C}$ of the point (x, y) is the point $\left(\frac{x}{x^2+y^2}, \frac{y}{x^2+y^2}\right)$. We may reformulate this expression in terms of complex numbers, by using the fact that the modulus $|z|$ of a complex number $z = x + iy$ satisfies the identity

By Theorem 2, Subsection 5.1.2

$$x^2 + y^2 = |z|^2 = z\bar{z}.$$

Thus the image under inversion of the point $z = x + iy$ is the point

$$\frac{x}{x^2 + y^2} + i\frac{y}{x^2 + y^2} = \frac{x + iy}{x^2 + y^2} = \frac{z}{z\bar{z}} = \frac{1}{\bar{z}}.$$

Next we consider the general case where C is a circle of radius r with centre (a, b). In this case the inversion in C can be expressed as the composite $t = t_3 \circ t_2 \circ t_1$, where

$t_1(z) = \frac{z-c}{r}$ is a translation and a scaling that sends C to the unit circle;

$t_2(z) = 1/\bar{z}$ is the inversion in the unit circle;

$t_3(z) = rz + c$ is the inverse of t_1 and sends the unit circle back to C.

Then

$$t(z) = t_3 \circ t_2 \circ t_1(z) = \frac{r^2}{\overline{z - c}} + c,$$

as required.

The representation of inversion provided by Theorem 3 has a particularly simple form in the case where C is the unit circle $\mathcal{C}$. In that case, $r = 1$ and $c = 0$ so the inversion is represented by the transformation

$$t(z) = \frac{1}{\bar{z}} \quad (z \in \mathbb{C} - \{O\}).$$

You met this formula in the proof of Theorem 3 above.

For example, the image of the point $1 \quad i$ under inversion in the unit circle C is the point

$$\frac{1}{\overline{1 \quad i}} = \frac{1}{1+i} = \frac{1 \quad i}{(1+i)(1 \quad i)} = \tfrac{1}{2}(1 \quad i).$$

Note that the technique of multiplying both numerator and denominator by the conjugate of the denominator is often useful.

Problem 5 Determine, in Cartesian form, the image under inversion in the unit circle C of each of the following points.

(a) $\overline{3} + i$ (b) $3 \quad 4i$

Problem 6 Let C be the circle of radius 2 with centre at the origin. Write down the inversion in C as a transformation of $\mathsf{C} \quad \{O\}$.

We know that inversions preserve the magnitude of angles. But what about their effect on circles and lines? Unfortunately, it is not strictly correct to say that an inversion maps circles and lines to circles and lines since some of the circles and lines may have to be punctured before the map can be carried out. We describe how to overcome this complication later, in Subsection 5.2.3.

By Theorem, Subsection 5.1.3

5.2.2 Linear and Reciprocal Functions

We can use the basic complex functions described in the previous subsection to give a geometric interpretation of many other complex functions. We illustrate this for the so-called *linear* and *reciprocal functions*.

Definition A **linear function** is a function of the form

$$t(z) = az + b \quad (z \quad \mathsf{C}),$$

where a, b C and $a = 0$.

Every linear function $t(z) = az + b$ can be decomposed into a composite $t_2 \quad t_1$ where

t_1 is the scaling $t_1(z) = |a|z$,
t_2 is the isometry $t_2(z) = (a/|a|)z + b$.

The geometrical interpretation of the linear function t depends on how we choose to interpret the isometry t_2. We can say that the linear function is a scaling by the factor $|a|$, followed by a rotation through the angle $\operatorname{Arg}(a/|a|)$, followed by a translation through the vector $(\operatorname{Re} b, \operatorname{Im} b)$. Alternatively, we can use Theorem 2 and say that the linear function is a scaling composed with a number of reflections.

Either way, since both the scaling and the isometry preserve angles and map circles and lines to circles and lines, it follows that the same must be true of all linear functions.

Next, we consider the *reciprocal function*.

Definition The **reciprocal function** is defined by

$$t(z) = \frac{1}{z} \quad (z \in \mathbb{C} - \{O\}).$$

This function can be decomposed into the composite $t_2 \circ t_1$ where

t_1 is the inversion $t_1(z) = \frac{1}{\bar{z}}$,
t_2 is the conjugation $t_2(z) = \bar{z}$.

It follows that, geometrically, the reciprocal function can be interpreted as an inversion in the unit circle $\mathcal{C}$ followed by a reflection in the real axis, as shown below.

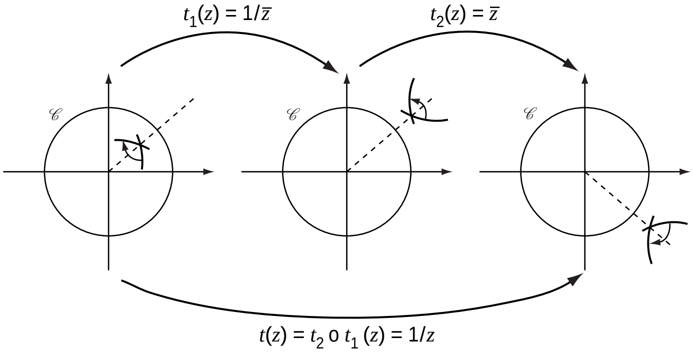

Since the inversion and the conjugation are both one-one functions that preserve the magnitude of angles and reverse their orientation, it follows that the reciprocal function must be a one-one function that preserves both the magnitude and the orientation of angles.

As in the case of inversions, it is not strictly correct to say that the reciprocal function maps circles and lines to circles and lines, because some of the circles and lines may have to be punctured before the reciprocal function can be applied. We next show how we can overcome this complication by extending the complex plane.

5.2.3 The Extended Plane

So far, our discussion of the effect of inversion on lines and circles has been complicated by the need to puncture those lines and circles that pass through the centre of inversion. Also, the need to distinguish between those images that are circles and those that are lines makes the description of the inversion process somewhat cumbersome. We can deal with both these difficulties in a very elegant way by adding an additional point to the plane to obtain the so-called *extended complex plane*.

To illustrate the ideas involved, consider the line with equation $x = 2$. Recall that inversion in the unit circle C maps to a circle C with radius $\frac{1}{4}$ and centre $(\frac{1}{4}, 0)$, punctured at the origin.

Example 5, part (a), Subsection 5.1.2

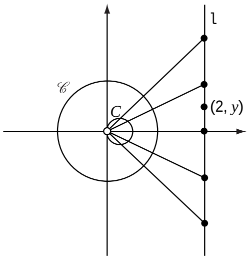

The point (2, 0) is mapped to $(\frac{1}{2}, 0)$, (2, 1) is mapped to $(\frac{2}{5}, \frac{1}{5})$, (2, 2) is mapped to $(\frac{1}{4}, \frac{1}{4})$, (2, 3) is mapped to $(\frac{2}{13}, \frac{3}{13})$ and, in general, the point $(2, y)$ is mapped to the point $(\frac{2}{4+y^2}, \frac{y}{4+y^2})$ on C. As the point $(2, y)$ moves from (2, 0) *up* the line , its image under inversion moves from $(\frac{1}{2}, 0)$ around the circle C towards the origin in an *anticlockwise* direction. Similarly, as the point $(2, y)$ moves from (2, 0) *down* the line , its image under inversion moves from $(\frac{1}{2}, 0)$ around the circle C towards the origin in a *clockwise* direction. The gap in the circle C at the origin arises because there is no point on , or indeed any point in R^2, that is inverted to the origin. To fill the gap we attach an additional *point at infinity* to the plane.

Definitions The **extended plane** is the union of the Euclidean plane R^2 and one extra point, the **point at infinity**, denoted by the symbol . When we wish to consider the plane as the complex plane C, then we call the extended plane the **extended complex plane** and denote it by the symbol C; thus $\mathsf{C} = \mathsf{C} \quad \{ \quad \}$.

C is read as C hat .

Remarks

1. The *extended plane* and the *extended complex plane* both consist of the ordinary plane together with the point . We use both terms interchangeably, depending on whether we wish to think of the ordinary points in the plane as being represented by pairs of real numbers or by complex numbers.
2. The symbol does *not* represent a complex number and so it should *not* be used in association with arithmetic operations that act on complex numbers. For example, $+ 3i$ is a meaningless expression.

When we wish to think of points as complex numbers, we often denote them by lower-case letters such as z or a rather than the upper case letters usually used for points.

Having extended the plane in this way, we now extend the definition of inversion in a circle with centre O. We simply define the image of O to be , and the image of to be O. Other points are mapped as specified by the definition of inversion given in Subsection 5.1.1.

A more formal definition of inversion in C is given below.

With this extended definition of inversion, the point O corresponds to , and so the circle C in the above discussion now corresponds under inversion in C to the set $\{ \}$. This certainly fills the gap in C, but how can the set $\{ \}$ be interpreted?

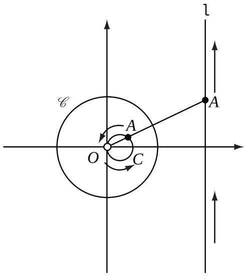

As a point A on the circle C moves anticlockwise towards O, its image A under inversion moves up . When A reaches O, A reaches . As A continues around C below O, its image returns up from below. You can think of the point as linking the two ends of the line , thereby enabling points to travel round and round the line. With some stretch of your imagination, you can therefore think of the line as a circle of infinite radius, where the gap at infinity has been filled by the point .

Any line in the plane may have its gap at infinity filled by forming the set $\{ \}$. Such a set is called an **extended line**. Since an extended line can be thought of as a circle of infinite radius, we make the following definition.

Definition A **generalized circle** in the extended plane is a set that is either a circle or an extended line.

Remark

With this definition, you can think of an extended line as a generalized circle which passes through , and you can think of an ordinary line as a generalized circle which has been punctured at .

Recall that in Subsection 5.1.1 we adopted the convention that an inversion is either a reflection in a line, or an inversion with respect to a circle. We can now regard both of these as inversions in generalized circles.

Definition Let C be a generalized circle in the extended complex plane. Then an **inversion of the extended plane with respect to** C is a function t defined by one of the following rules:

(a) if C is a circle of radius r with centre O, then

$$t(A) = \begin{cases} \text{the inverse of } A \text{ with respect to } C, & \text{if } A \quad \mathbb{C} \quad \{O\},\\ \quad , & \text{if } A = O,\\ O, & \text{if } A = \quad ; \end{cases}$$

(b) if C is an extended line $\{ \}$, then

$$t(A) = \begin{cases} \text{the reflection of } A \text{ in } \quad , & \text{if } A \quad \mathbb{C},\\ \quad , & \text{if } A = \quad . \end{cases}$$

Remember that A' is the inverse of A with respect to C if it lies on the same radial half-line from O as A, and $OA \cdot OA' = r^2$.

Remark

Note that any inversion in an extended line fixes the point at infinity. Conversely, every inversion that fixes must be an inversion in an extended line.

The above definition of inversion ensures that generalized circles map to generalized circles. Indeed, we already know from Subsection 5.1.2 that this is true if we allow the circles to be punctured, but we still need to check that the gap in a punctured circle is mapped to the gap in its image circle.

In the case of inversion with respect to a circle with centre O, we know that a circle or line punctured at O maps to a line, and this is consistent with the fact that O maps to . Also, a line maps to a circle or line that is punctured at O, and this is consistent with the fact that maps to O.

In the case of inversion with respect to an extended line, we know that lines reflect to lines, which is consistent with the fact that maps to . Ordinary circles are not a problem since they reflect onto circles.

We therefore have the following important result.

Theorem 4 Inversions of the extended plane map generalized circles onto generalized circles.

We shall sometimes find it convenient to write inversions of the extended plane as inversions of $\mathbb{C}$. For example, by Theorem 3, we can write the inversion of the extended plane with respect to the unit circle $\mathbb{C}$ in the form

Subsection 5.2.1

$$t(z) = \begin{cases} \frac{1}{\bar{z}}, & \text{if } z \quad \mathbb{C} \quad \{O\}, \\ 0, & \text{if } z = \quad , \\ \quad , & \text{if } z = 0. \end{cases}$$

Problem 7 Write down each of the following inversions of the extended plane as a transformation of $\mathbb{C}$:

(a) the inversion with respect to the circle of radius 2 with centre the origin;

(b) the inversion with respect to the extended real axis.

The inversion of the extended plane that we asked you to write down in Problem 7, part (b), is particularly important because it provides us with a natural way of extending the conjugation function from $\mathbb{C}$ to $\mathbb{C}$.

Definition The function $t : \mathbb{C} \quad \mathbb{C}$ defined by

$$t(z) = \begin{cases} \bar{z}, & \text{if } z \quad \mathbb{C}, \\ \quad , & \text{if } z = \quad , \end{cases}$$

is called the **extended conjugation function**.

This function occurs so frequently that we shall introduce a notation for the images of the points in its domain. Since we already have the notation $\bar{z}$ for the conjugate of a complex number z, we simply adopt the convention that $\overline{} = $.

Inversions are not the only transformations that can be extended to $\mathbb{C}$. Indeed, most of the transformations discussed at the beginning of this section can be extended in a natural way to $\mathbb{C}$.

Definitions

(a) The function $t : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$ defined by

$$t(z) = \begin{cases} \dfrac{1}{z}, & \text{if } z \in \mathbb{C} - \{0\}, \\ \infty, & \text{if } z = 0, \\ 0, & \text{if } z = \infty, \end{cases}$$

is called the **extended reciprocal function**.

(b) A function $t : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$ of the form

$$t(z) = \begin{cases} az + b, & \text{if } z \in \mathbb{C}, \\ \infty, & \text{if } z = \infty, \end{cases}$$

where $a, b \in \mathbb{C}$ and $a \neq 0$, is called an **extended linear function**.

Since scalings, rotations and translations are all special types of linear function, they too are extended to $\hat{\mathbb{C}}$ by this definition.

The solution to the next example shows that the extended reciprocal function can be expressed as a composite of two inversions.

Example 2 Find the composite $t = t_2 \circ t_1$ where

t_1 is the inversion in the unit circle $\mathcal{C}$,
t_2 is the extended conjugation function.

Solution Since

$$t_1(z) = \begin{cases} \dfrac{1}{\bar{z}}, & \text{if } z \in \mathbb{C} - \{0\}, \\ \infty, & \text{if } z = 0, \\ 0, & \text{if } z = \infty, \end{cases} \quad \text{and} \quad t_2(z) = \begin{cases} \bar{z}, & \text{if } z \in \mathbb{C}, \\ \infty, & \text{if } z = \infty, \end{cases}$$

we have

$$t(\infty) = t_2 \circ t_1(\infty) = t_2(0) = 0,$$
$$t(0) = t_2 \circ t_1(0) = t_2(\infty) = \infty.$$

For the remaining values of $z \in \mathbb{C} - \{0\}$ we have

$$t(z) = t_2 \circ t_1(z) = t_2\left(\frac{1}{\bar{z}}\right) = \overline{\left(\frac{1}{\bar{z}}\right)} = \frac{1}{z}.$$

It follows that $t = t_2 \circ t_1$ is the extended reciprocal function. □

Next we describe how an extended linear function can be expressed as a composite of inversions. In preparation for this, we ask you to tackle the following problem.

Problem 8 Find the composite $t = t_2\ \ t_1$ where

t_1 is the inversion with respect to the unit circle C ,
t_2 is the inversion with respect to the circle of radius 2 with centre 0.

Give an interpretation of the composite function t. Also describe what would happen if the circle of radius 2 were replaced by a circle of radius $\overline{k}$ with centre 0.

Earlier, we noted that every linear function is a composite $t\ \ s$ of a scaling s followed by an isometry t. Since every isometry t is a composite of reflections, it follows that every linear function is a composite $r_n\ \ \ldots\ \ r_2\ \ r_1\ \ s$ of a scaling s followed by a number of reflections $r_1, r_2, \ldots, r_n$.

Now the only difference between a linear function and an *extended* linear function is that the latter contains an additional point at infinity in its domain. Since this additional point maps to itself, it follows that an extended linear function is a composite of a scaling that also fixes , followed by a number of reflections that also fix .

But a reflection that fixes is just an inversion in an extended line. Also, by the solution to Problem 8, a scaling that also fixes is a composite of two inversions. It follows that every extended linear function is a composite of inversions.

Combining this with the result of Example 2 above we have the following theorem.

Theorem 5 The extended reciprocal function and the extended linear functions can be decomposed into composites of inversions.

Since inversions of the extended plane map generalized circles onto generalized circles, we have the following corollary of Theorem 5.

Corollary The extended linear functions and the extended reciprocal function map generalized circles onto generalized circles.

Example 3 Let t be the extended linear function defined by

$$t(z) = \begin{cases} 2\ \ \ 1+\ \overline{3}i\ \ z + (4\ \ \ 2i), & \text{if } z\ \ \ \mathsf{C},\\ & \text{if } z = \ \ . \end{cases}$$

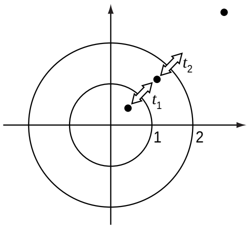

Express t as a composite of inversions of the extended complex plane.

Solution In addition to mapping to , the transformation t *scales* the complex plane by the factor $|2(\ \ 1+\ \ \overline{3}i)| = 4$, *rotates* it through the angle $\text{Arg}\,(2(\ \ 1+\ \ \overline{3}i)) = \frac{2}{3}$, and then *translates* it through the vector $(4,\ \ 2)$.

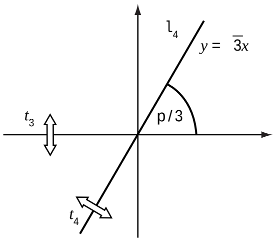

The *scaling* by the factor 4 can be decomposed into the composite $t_2 \circ t_1$, where

t_1 is the inversion in the unit circle $\mathbb{C}$,
t_2 is the inversion in the circle of radius $\sqrt{4} = 2$ centred at the origin.

The *rotation* through the angle $2\pi/3$ can be decomposed into the composite $t_4 \circ t_3$, where

t_3 is the inversion in the extended real axis,
t_4 is the inversion in the extended line $\ell_4 \cup \{\infty\}$, where ℓ_4 is the line $y = \sqrt{3}x$.

(Here ℓ_4 is the line through the origin that makes an angle $\pi/3$ with the x-axis.)

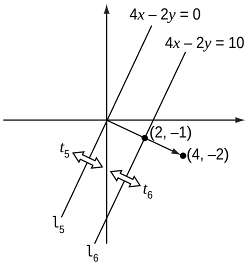

The *translation* through the vector $(4, -2)$ can be decomposed into the composite $t_6 \circ t_5$, where

t_5 is the inversion in the extended line $\ell_5 \cup \{\infty\}$, where ℓ_5 is the line $4x - 2y = 0$,
t_6 is the inversion in the extended line $\ell_6 \cup \{\infty\}$, where ℓ_6 is the line $4x - 2y = 10$.

(Note that ℓ_5 is the line through the origin that is perpendicular to the vector $(4, -2)$, and ℓ_6 is the line that passes through $(2, -1)$ parallel to ℓ_5.)

Since t_6, t_5, t_4, t_3 and $t_2 \circ t_1$ all map ∞ to itself, it follows that

$$t = t_6 \circ t_5 \circ t_4 \circ t_3 \circ t_2 \circ t_1. \qquad \square$$

Problem 9 Let t be the extended linear function defined by

$$t(z) = \begin{cases} 9z + (6 - 10i), & \text{if } z \in \mathbb{C}, \\ \infty, & \text{if } z = \infty. \end{cases}$$

Express t as a composite of inversions of the extended complex plane.
Hint: Use the decompositions described in the margin.

The decompositions illustrated above generalize as follows.

To *scale* $\mathbb{C}$ by a factor k: first invert in the circle $\{z : |z| = 1\}$, then invert in the circle $\{z : |z| = \sqrt{k}\}$.

To *rotate* $\mathbb{C}$ through an angle θ: first invert in the extended line $\{z : \operatorname{Arg} z = 0\} \cup \{\infty\}$, then invert in the extended line $\{z : \operatorname{Arg} z = \frac{1}{2}\theta\} \cup \{\infty\}$.

To *translate* $\mathbb{C}$ through a vector (a, b): first invert in the extended line $\{(x, y) : ax + by = 0\} \cup \{\infty\}$, then invert in the extended line $\{(x, y) : ax + by = \frac{1}{2}(a^2 + b^2)\} \cup \{\infty\}$.

5.2.4 The Riemann Sphere

In the previous subsection, we introduced the point at infinity and the extended complex plane in order to provide a simplified explanation of the effect of inversion on lines and circles. However, it may seem unsatisfactory to have to visualize a straight line as a (generalized) circle, since the 'ends' of the line appear to be infinitely far apart. Also, it is difficult to visualize where the point at infinity should be placed relative to $\mathbb{C}$, other than to think of it as smeared in some vague way around the 'outer edge' of $\mathbb{C}$.

Fortunately, there is a model of the extended (complex) plane in which the point at infinity appears as an actual point! Consider the complex plane $\mathbb{C}$ as lying in $\mathbb{R}^3$ with the real and imaginary axes aligned along the x-axis and y-axis, respectively. Then each complex number $x + iy$ may be represented by the point $(x, y, 0)$ in the (x, y)-plane.

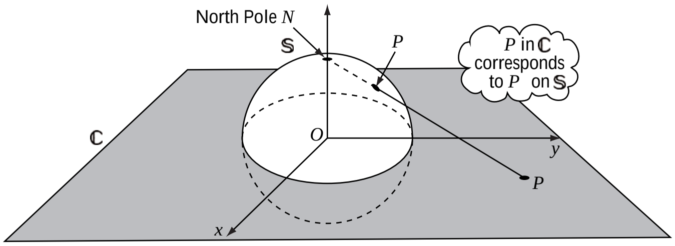

Next, draw the sphere S with centre at the origin and with radius 1; we call this the **Riemann sphere**. By analogy with the Earth, we refer to the point $N = (0, 0, 1)$ at the top of the sphere as the **North Pole** of S, and the point $S = (0, 0, \quad 1)$ at the bottom of the sphere as the **South Pole** of S.

Each line joining a point P in the complex plane to the North Pole intersects the Riemann sphere at some point P , and vice versa. In this way, we obtain a one one correspondence between all points P in the complex plane and all but one of the points P on the sphere. The one point on the sphere that cannot be associated with a point in the complex plane is the North Pole N.

As the point P on the sphere moves closer to N, the corresponding point P in the plane moves further away from the origin O. This suggests that we associate the North Pole N with the point in the extended complex plane.

The function $: \mathsf{S} \quad \mathsf{C}$, which maps points on the Riemann sphere to the associated points in the extended complex plane, is called **stereographic projection**. Since is one one and onto, it follows that we can use the Riemann sphere as a convenient visualization of the extended complex plane C.

In fact, we can find an explicit formula for stereographic projection of the point (X, Y, Z) onto the point $z = x + iy$, and vice versa.

We use capital letters to denote coordinates of points in R^3, and small letters to denote points or coordinates in the complex plane.

Theorem 6 Let denote the mapping of the Riemann sphere S onto C given by stereographic projection. Then the stereographic projection of the point (X, Y, Z) of S onto the point $z = x + iy$ of C is given by

$$(X, Y, Z) = \frac{X}{1 \quad Z} + i\frac{Y}{1 \quad Z}.$$

Also, the inverse mapping is given by

$$^{1}(x + iy) = \frac{2x}{x^2 + y^2 + 1}, \frac{2y}{x^2 + y^2 + 1}, \frac{x^2 + y^2 \quad 1}{x^2 + y^2 + 1} \quad .$$

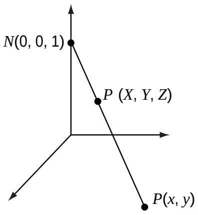

Proof Let the point $P\ (X, Y, Z)$ be a point on S (other than N), and $P(x, y)$ the point in the plane that corresponds to P under stereographic projection. Project the line $NP\ P$ perpendicularly onto the (Y, Z)-plane, with P and P projecting onto the points Q and Q; then Q and Q have coordinates $(0, Y, Z)$ and $(0, y, 0)$ in R^3. Draw the perpendicular $Q\ R$ from Q to the Z-axis.

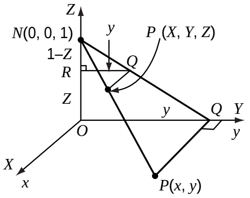

Since RQ is parallel to OQ, the triangles NRQ and NOQ are similar. It follows that $NR : RQ = NO : OQ$, or

$$\frac{1-Z}{Y} = \frac{1}{y},$$

from which we obtain that

$$y = \frac{Y}{1-Z}.$$

By projecting the line $NP P$ onto the (X, Z)-plane and using a similar argument, we can show that

We omit the details.

$$x = \frac{X}{1-Z}.$$

It then follows that the mapping is given by

$$(X, Y, Z) = x + iy = \frac{X}{1-Z} + i\frac{Y}{1-Z}.$$

Clearly this formula also holds when N is mapped to .

For, here $Z = 1$.

To find the formula for ^{1}, we use the fact that $X^2 + Y^2 + Z^2 = 1$. Substituting the values of X and Y from the above formulas into the equation $X^2 + Y^2 + Z^2 = 1$ and doing some manipulation, we deduce that

We omit the details.

$$x^2 + y^2 + 1 = \frac{2}{1-Z}. \tag{5}$$

It follows from equation (5) and the earlier equations for X and Y that

$$X = x(1-Z) = \frac{2x}{x^2 + y^2 + 1}$$

and

$$Y = y(1-Z) = \frac{2y}{x^2 + y^2 + 1}.$$

Also, we may rearrange equation (5) in the form

$$Z = \frac{x^2 + y^2 - 1}{x^2 + y^2 + 1}.$$

We can then combine these formulas to give the required formula for ^{1}:

$$\begin{aligned} {}^{1}(x + iy) &= (X, Y, Z) \\ &= \left(\frac{2x}{x^2 + y^2 + 1}, \frac{2y}{x^2 + y^2 + 1}, \frac{x^2 + y^2 - 1}{x^2 + y^2 + 1}\right). \end{aligned}$$

Notice that stereographic projection distorts the distances between points. For example, two points that are close together on S may project onto points that are close together in C, or onto points that are far from each other in C.

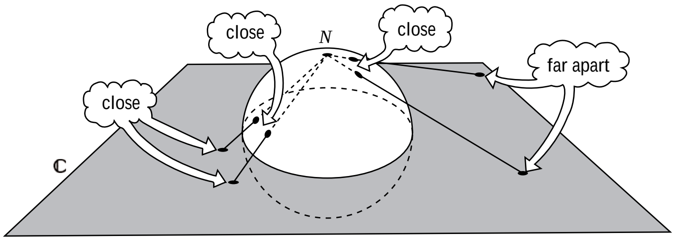

We can use the Riemann sphere to visualize extended lines as generalized circles in a very natural way, as follows. Consider a point P on a line in ℂ.

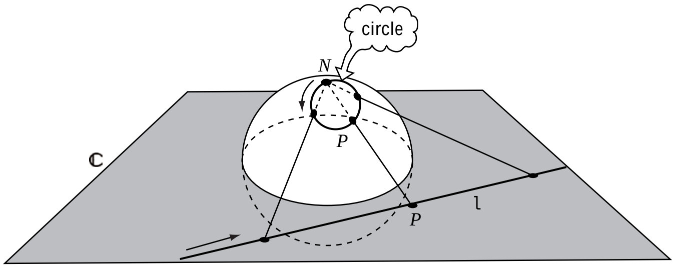

As P moves along , the line that joins P to the North Pole sweeps out the plane through and N; this intersects the sphere in a circle. Hence, as P moves along , the corresponding point P on the sphere traces out a circle. As P moves further out along the line, the point P moves towards the North Pole. The point P never actually reaches the North Pole, since the North Pole corresponds to the *gap* in the line that we mentioned earlier.

Subsection 5.2.3

We filled the gap in by attaching to to obtain the extended line { } in ℂ. On the sphere, the point N fills the corresponding gap in the circle, so that the extended line { } corresponds to an *actual* circle on the sphere in fact, a circle through N.

This fact is a particular instance of the following general result.

Theorem 7 Under stereographic projection, circles on the Riemann sphere map onto generalized circles in ℂ.

In particular, circles on the sphere that pass through N map onto extended lines in ℂ, and circles on the sphere that do not pass through N map onto ordinary circles in ℂ.

Proof A circle on the sphere is the intersection of the sphere $X^2 + Y^2 + Z^2 = 1$ with some plane $aX + bY + cZ + d = 0$, where a, b and c are not all zero. It follows from substituting the expressions

$$X = \frac{2x}{x^2 + y^2 + 1}, \quad Y = \frac{2y}{x^2 + y^2 + 1} \quad \text{and} \quad Z = \frac{x^2 + y^2 \quad 1}{x^2 + y^2 + 1},$$

You met these equations in the proof of Theorem 6 above.

for X, Y and Z into the equation $X^2 + Y^2 + Z^2 = 1$ that

$$\frac{2ax + 2by + c \quad x^2 + y^2 \quad 1}{x^2 + y^2 + 1} + d = 0.$$

This equation may be rewritten in the form

$$2ax + 2by + c \quad x^2 + y^2 \quad 1 \quad + d \quad x^2 + y^2 + 1 \quad = 0,$$

or

$$(c + d)x^2 + (c + d)y^2 + 2ax + 2by + (d \quad c) = 0.$$

It follows that the image of a circle on the sphere is a generalized circle in the (extended) plane.

For its equation is that of a circle or of an (extended) line.

The circle on the sphere passes through $N(0, 0, 1)$ if the plane $aX + bY + cZ + d = 0$ passes through N, and so if $c + d = 0$.

It follows that if the circle on the sphere passes through N, its image in the extended plane has an equation of the form

$$2ax + 2by + (d \quad c) = 0,$$

and is an extended line.

On the other hand, if the circle on the sphere does not pass through N, then $c + d = 0$ and its image is an ordinary circle in the plane.

Problem 10 Determine the images under stereographic projection onto $\mathbb{C}$ of the following circles on $\mathbb{S}$.

(a) The circle $\{(X, Y, Z) : X^2 + Y^2 + Z^2 = 1, X = \frac{1}{2}\}$

(b) The circle of intersection of $\mathbb{S}$ with the plane $3X + 2Y + Z = 1$

As an illustration of Theorem 7, we describe what happens to the lines of latitude and longitude of the Riemann sphere under stereographic projection.

A line of longitude on the Riemann sphere is a circle on $\mathbb{S}$ that passes through the North Pole and the South Pole of $\mathbb{S}$; this projects onto a line through the origin in the plane. Similarly, a line of latitude on the Riemann sphere is a circle at a constant height above the (x, y)-plane, so, by symmetry, this projects onto a circle centred at the origin in the plane.

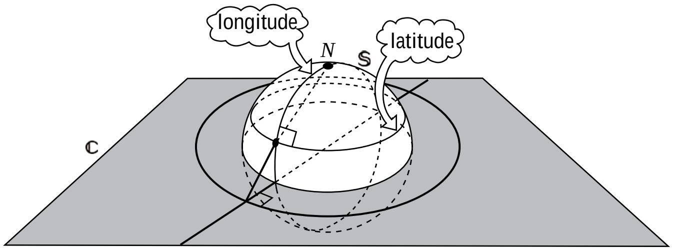

Just as the lines of latitude and longitude meet at right angles on the sphere, so do their projections. This is a consequence of the following remarkable result.

Theorem 8 Stereographic projection preserves the magnitude of angles.

Recall that inversion has a similar property (see Theorem 5, Subsection 5.1.3).

Proof First, we make the following crucial observation. Recall that if a point P moves along a line ℓ in the complex plane, then the corresponding point P' on the Riemann sphere moves round a circle through N. Indeed, as P moves out towards ∞, P' approaches N; and the line NP' (in $\mathbb{R}^3$) approaches the tangent at N to the circle on the sphere. It follows that the tangent at N to the circle on the sphere must be parallel to the original line ℓ.

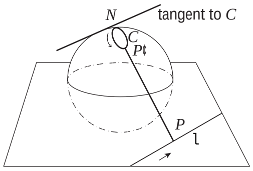

Let C_1 and C_2 be curves in the complex plane that intersect at some point P, and let ℓ_1 and ℓ_2 be the tangents to C_1 and C_2 at P. (Recall that we define the angle between C_1 and C_2 at P to be the angle between ℓ_1 and ℓ_2 at P.)

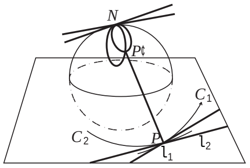

The curves on the Riemann sphere that correspond to ℓ_1 and ℓ_2 are circles through the North Pole N and the point P' on the sphere that corresponds to the point P in the plane. We have to show that the angle between these circles at P' (that is, the angle between their tangents at P') is equal to the angle between the lines ℓ_1 and ℓ_2 at P.

But, by symmetry, the angle between the circles at P' is equal to the angle between the same circles when they meet again at N. And, as we saw above, the lines ℓ_1 and ℓ_2 through P are parallel to the corresponding tangent lines through N. Thus the angle between the circles at N must be equal to the original angle between ℓ_1 and ℓ_2 at P.

Hence, stereographic projection does preserve the magnitude of angles, as asserted.

Earlier, we showed that inversion preserves the magnitude of angles but reverses their orientation. Having introduced the Riemann sphere, we can now show that there is a sense in which this is true even for angles which have a vertex at the centre of inversion, or at ∞.

To see this, let ℓ_1 and ℓ_2 be two lines that intersect at the centre of inversion A. The corresponding extended lines $\ell_1 \cup \{\infty\}$ and $\ell_2 \cup \{\infty\}$ intersect again at the point ∞. On the Riemann sphere these extended lines become circles C_1 and C_2 that intersect at the North Pole and at the point A' corresponding to A.

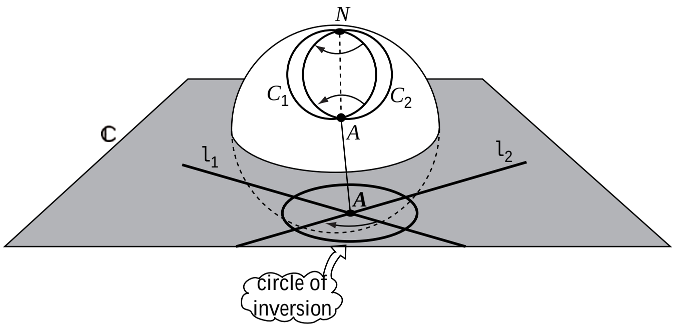

Under the inversion, each extended line maps to itself, so C_1 and C_2 also map to themselves. Notice, however, that the points A and N swap over, and so the angles at A and N swap over as well. By symmetry, these angles are equal in magnitude but opposite in direction. So, provided we interpret the angle between curves at as the corresponding angle on the Riemann sphere, we can conclude that the fact that inversion preserves the magnitude of angles holds throughout the whole of $\mathbb{C}$.

5.3 Inversive Geometry

5.3.1 Inversive Geometry

We now come to the main purpose of this chapter, which is to introduce a geometry, known as *inversive geometry*, that we can use to study the properties of circles, lines and angles.

Recall that according to Klein, a geometry has the following ingredients:

Introduction to Chapter 2, before Subsection 2.1.1

a space consisting of a set of points;
a group of transformations that act on the space.

Each geometry is used to study those properties of figures in its space that are preserved by its transformations. For example, Euclidean geometry is used to study those properties of figures in $\mathbb{R}^2$, such as angle and distance, that are preserved by the isometries of $\mathbb{R}^2$.

Since each isometry of $\mathbb{R}^2$ can be decomposed into a composite of reflections, we can think of the group associated with Euclidean geometry as the group of all possible composites of reflections.

This provides the clue that we need to define the transformations of our new geometry. Rather than consider composites of reflections, we consider composites of inversions.

Definition A transformation $t : \mathbb{C} \quad \mathbb{C}$ is an **inversive transformation** if it can be expressed as a composite of inversions.

It is important to distinguish between inversions and inversive transformations.

For example, the extended reciprocal function is an inversive transformation because it can be expressed as a composite $t_2 \quad t_1$ where t_1 is the inversion in the unit circle and t_2 is the inversion in the real axis. Similarly, all extended linear functions can be expressed as composites of inversions, so they too are inversive transformations.

Theorem 1 The extended reciprocal function and the extended linear functions are inversive transformations.

Since every inversion preserves the magnitude of angles and maps generalized circles to generalized circles, the same must be true of all composites of inversions. We therefore have the following result.

Theorem 2 Inversive transformations preserve the magnitude of angles, and map generalized circles to generalized circles.

Before we can use the inversive transformations to define a geometry, we must first check that they form a group.

Theorem 3 The set of inversive transformations forms a group under the operation of composition of functions.

Proof We check that the four group axioms hold.

G1 CLOSURE Let r and s be inversive transformations. Then we can write

$$r = t_1 \circ t_2 \circ \ldots \circ t_k$$

and

$$s = t_{k+1} \circ t_{k+2} \circ \ldots \circ t_n,$$

where $t_1, t_2, \ldots, t_n$ are inversions. Thus

$$r \circ s = (t_1 \circ t_2 \circ \ldots \circ t_k) \circ (t_{k+1} \circ t_{k+2} \circ \ldots \circ t_n)$$

is a composite of inversions, and is therefore an inversive transformation.

G2 IDENTITY The identity for composition of functions is the identity transformation given by

$$t(z) = z \quad (z \in \mathbb{C}).$$

This is an inversive transformation since $t = s \circ s$, where s is the inversion in the unit circle.

Here we could have chosen any other inversion for s since all inversions are self-inverse.

G3 INVERSES If t is an inversive transformation, then we can write

$$t = t_1 \circ t_2 \circ \ldots \circ t_n,$$

where $t_1, t_2, \ldots t_n$ are inversions. It follows that t has inverse

$$t^{-1} = t_n^{-1} \circ t_{n-1}^{-1} \circ \ldots \circ t_1^{-1} = t_n \circ t_{n-1} \circ \ldots \circ t_1,$$

which is an inversive transformation.

G4 ASSOCIATIVITY Composition of functions is always associative.

Since all four group properties hold, it follows that the set of inversive transformations forms a group under composition of functions.

Having shown that the inversive transformations form a group, we can use them to define a geometry. But what space should we use for the geometry? Since the inversive transformations act on $\mathbb{C}$, the space will have to be $\mathbb{C}$. But what do figures look like in this space? Just as figures in $\mathbb{R}^2$ are defined to be subsets of $\mathbb{R}^2$, so **figures in $\mathbb{C}$** are defined to be subsets of $\mathbb{C}$. Thus circles and extended lines are both examples of figures in $\mathbb{C}$. We can now make the following definition.

Definition **Inversive geometry** is the study of those properties of figures in $\mathbb{C}$ that are preserved by inversive transformations.

We have already met the following inversive properties: generalized circles, magnitude of angles, and tangency (since zero angles are preserved).

5.3.2 Relationship with Other Geometries

Although the point at infinity is crucial to the theory of inversive geometry, when we come to interpret results about figures we often choose to confine our attention to points in $\mathbb{C}$. By ignoring the point ∞ in this way, the remaining points in $\mathbb{C}$ can be interpreted as points of $\mathbb{R}^2$ and any extended lines in the figure can be interpreted as ordinary lines.

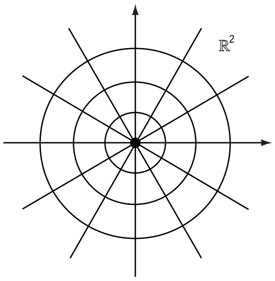

In this sense figures in $\mathbb{R}^2$ can be interpreted as figures in $\mathbb{C}$, and vice versa. For example, the figure in the margin is a figure in $\mathbb{R}^2$. It consists of a family of circles centred at the origin, and a family of lines that intersect at the origin. If we add the point ∞ to the figure, then all the lines become generalized circles and the figure becomes a figure in $\hat{\mathbb{C}}$.

Earlier, you saw that every Euclidean transformation can be expressed as a composite of reflections. Since a reflection can be interpreted as an inversion that fixes ∞, it follows that we can interpret every Euclidean transformation as a composite of inversions, and hence as an inversive transformation.

We could make this interpretation precise by adopting the convention that a Euclidean transformation maps ∞ to ∞ when it acts on $\hat{\mathbb{C}}$.

An immediate consequence of the above observations is that when figures in $\mathbb{C}$ are interpreted as figures in $\mathbb{R}^2$, all their inversive properties become Euclidean properties. This is because any property that is preserved by all inversive transformations must also be preserved by all Euclidean transformations. For example, the magnitude of angles is both an inversive and a Euclidean property.

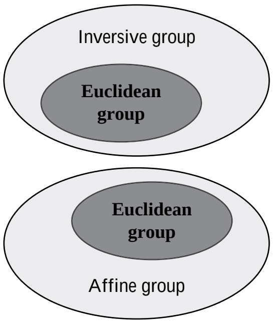

A pictorial representation of the relationship between inversive and Euclidean geometry is given in the margin. Provided that we ignore the point ∞, the group of Euclidean transformations can be regarded as a subgroup of the group of inversive transformations. Because the Euclidean group is smaller than the inversive group, it follows that Euclidean geometry has more properties than does inversive geometry. For example, length is a Euclidean property but it is not an inversive property.

How does affine geometry fit into the scheme? In Subsection 2.2.1 we showed that every Euclidean transformation is an affine transformation. But what is the relationship between affine transformations and inversive transformations?

Certainly, some affine transformations are inversive without being Euclidean. For example, earlier you saw that the doubling map t,

See Subsection 2.2.3

$$t(z) = 2z \quad (z \in \mathbb{C}),$$

is an affine transformation. This transformation is not a Euclidean transformation, and yet it can be decomposed into the composite $t_2 \quad t_1$, where t_1 is the inversion in the unit circle C, and t_2 is the inversion in the circle $\{|z| = \quad \overline{2} : z \quad \mathrm{C}\}$. So, with the usual proviso about , t is an inversive transformation.

You saw how a scaling can be decomposed into a composite of inversions in Problem 8, Subsection 5.2.3.

However, not all affine transformations are inversive transformations. For example, the transformation

$$t(\mathbf{x}) = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \mathbf{x} \qquad \mathbf{x} \quad \mathrm{R}^2$$

represents a horizontal shear. This cannot be an inversive transformation, since it does not preserve angles.

It follows that the affine group of transformations contains the Euclidean group but overlaps the inversive group, as illustrated in the margin.

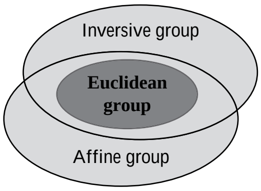

5.3.3 Möbius Transformations

To enable us to tackle problems in inversive geometry algebraically, we need an algebraic representation of the inversive transformations.

We shall show that each inversive transformation has either the form $t(z) = M(z)$, or the form $t(z) = M(\bar{z})$, where M is a so-called *Möbius transformation*.

Definition A **Möbius transformation** is a function $M : \mathrm{C} \quad \mathrm{C}$ of the form

$$M(z) = \frac{az + b}{cz + d},$$

where $a, b, c, d \quad \mathrm{C}$ and $ad \quad bc = 0$.

If $c = 0$, then we adopt the convention that $M(\quad) = \quad$; otherwise, we adopt the convention that $M(\quad d/c) = \quad$ and $M(\quad) = a/c$.

Remarks

1. Every (extended) linear function is a Möbius transformation, as can be seen by setting $c = 0$, $d = 1$. Also, the (extended) reciprocal function is a Möbius transformation with $a = d = 0$ and $b = c = 1$.
2. If $c = 0$, then the formula for M reduces to $M(z) = (a/d)z + (b/d)$. This defines an extended linear function, because the condition $ad \quad bc = 0$ ensures that both a and d are non-zero, and so a/d is also non-zero. Also, the convention that $M(\quad) = \quad$ complies with the definition of an extended linear function.
3. The condition $ad \quad bc = 0$ is equivalent to the statement that the ratios $a : c$ and $b : d$ are unequal. This is necessary to ensure that the numerator is not simply a multiple of the denominator, for if it were, M would be a constant function.

For example,

$6i \times 2 \quad 4i \times 3 = 0,$

so that

$\frac{6iz + 4i}{3z + 2} = 2i$, a constant.

Problem 1 Which of the following formulas define a Möbius transformation?

(a) $M(z) = \frac{5}{z}$ (b) $M(z) = \frac{z + 2i}{3z \quad 4i}$

(c) $M(z) = \quad 3z + \frac{i}{z}$ (d) $M(z) = 1 + \frac{5}{z + 2i}$

For those formulas that do define a Möbius transformation, state the image of under M.

Before we look at the properties of Möbius transformations, we first verify that a Möbius transformation is indeed an inversive transformation.

Theorem 4 Every Möbius transformation is an inversive transformation.

Proof Let M be the Möbius transformation defined by the formula

$$M(z) = \frac{az + b}{cz + d}.$$

If $c = 0$, then M is an extended linear function, and is therefore an inversive transformation.

If $c = 0$, then for $z \quad \mathbb{C} \quad \{ \quad d/c\}$ we can write

$$\begin{aligned} M(z) &= \frac{ad + bc + a(cz + d)}{c(cz + d)} \\ &= \quad \frac{ad \quad bc}{c} \bullet \frac{1}{cz + d} \quad + \frac{a}{c}. \end{aligned}$$

It follows that M may be expressed as the composite $t_3 \quad t_2 \quad t_1$, where t_2 is the extended reciprocal function, and t_1 and t_3 are the extended linear functions

$$t_1(z) = \begin{cases} cz + d, & \text{if } z = \quad , \\ \quad , & \text{if } z = \quad , \end{cases}$$

and

$$t_3(z) = \begin{cases} ((ad \quad bc)/c)z + (a/c), & \text{if } z = \quad , \\ \quad , & \text{if } z = \quad . \end{cases}$$

Next, we check that the transformations $t_3 \quad t_2 \quad t_1$ and M agree also at the exceptional points and $\quad d/c$, as follows:

$$t_3 \quad t_2 \quad t_1(\quad) = t_3 \quad t_2(\quad) = t_3(0) = a/c = M(\quad);$$

$$t_3 \quad t_2 \quad t_1(\quad d/c) = t_3 \quad t_2(0) = t_3(\quad) = \quad = M(\quad d/c).$$

Since the extended reciprocal function and the extended linear functions are inversive transformations, it follows that M is an inversive transformation.

Since Möbius transformations are inversive transformations, they must preserve the magnitude of angles and map generalized circles to generalized circles. In fact, we can say slightly more than this, for in the proof of Theorem 4 we showed that every Möbius transformation is either an extended linear function or a composite of two extended linear functions and an extended reciprocal function. Since the extended linear functions and the extended reciprocal function preserve both the magnitude and *orientation* of angles, the same must be true of Möbius transformations.

Theorem 5 Möbius transformations preserve the magnitude and orientation of angles, and map generalized circles onto generalized circles.

5.3.4 Matrix Representation of Möbius Transformations

In order to explore some of the other properties of Möbius transformations, it is helpful to establish a correspondence between Möbius transformations and matrices.

Recall that, if $a, b, c, d, e, f, g, h \in \mathbb{C}$, then 2×2 matrices have the following properties:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \bullet \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} ae + bg & af + bh \\ ce + dg & cf + dh \end{pmatrix},$$

and

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}, \quad \text{if } ad - bc \neq 0.$$

Here, the condition $ad - bc \neq 0$ for the matrix to be invertible is reminiscent of the condition $ad - bc \neq 0$ that appears in the definition of a Möbius transformation. This suggests the following fruitful connection between Möbius transformations and 2×2 invertible matrices.

Definition Let M be a Möbius transformation defined by

$$M(z) = \frac{az + b}{cz + d}, \tag{1}$$

where $a, b, c, d \in \mathbb{C}$. Then

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

is a matrix **associated** with M.

Remark

1. Every matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ associated with a Möbius transformation $M(z) = \frac{az+b}{cz+d}$ is invertible because $ad - bc \neq 0$.

2. A matrix associated with a Möbius transformation M is not unique, because we can multiply the numerator and denominator of the formula (1) by the same non-zero constant without altering the transformation. For example, both
$$M(z) = \frac{2z + i}{3z + 2i} \quad \text{and} \quad M(z) = \frac{2iz \quad 1}{3iz \quad 2}$$
specify the same Möbius transformation M, and so both the matrices
$$\begin{matrix} 2 & i \\ 3 & 2i \end{matrix} \quad \text{and} \quad \begin{matrix} 2i & 1 \\ 3i & 2 \end{matrix}$$
are associated with M.

   In general, if $\mathbf{A}$ is a matrix associated with a Möbius transformation M, then for any non-zero $c \quad \mathbb{C}$, $c\mathbf{A}$ is also a matrix associated with M. In fact, every matrix associated with M has the form $c\mathbf{A}$ for some $c \quad \mathbb{C} \quad \{0\}$.

Example 1 Decide which, if any, of the matrices
$$\mathbf{A}_1 = \begin{matrix} 0 & 4 \\ i & 0 \end{matrix}, \quad \mathbf{A}_2 = \begin{matrix} 8 & 0 \\ 2i & 8 \end{matrix}, \quad \mathbf{A}_3 = \begin{matrix} 4i & 0 \\ 0 & 1 \end{matrix}$$
are associated with each of the following Möbius transformations M.

(a) $M(z) = \dfrac{3iz + 2}{z \quad 3i}$ (b) $M(z) = \dfrac{4i}{z}$ (c) $M(z) = \dfrac{4iz}{z \quad 4i}$

Solution

(a) Every matrix associated with this M is a non-zero multiple of the matrix
$$\begin{matrix} 3i & 2 \\ 1 & 3i \end{matrix}.$$
Hence none of the three given matrices is associated with M.

(b) Every matrix associated with this M is a non-zero multiple of the matrix
$$\mathbf{A} = \begin{matrix} 0 & 4i \\ 1 & 0 \end{matrix}.$$
Since $\mathbf{A}_1 = i\mathbf{A}$, it follows that $\mathbf{A}_1$ is a matrix associated with M.

(c) Every matrix associated with this M is a non-zero multiple of the matrix
$$\mathbf{A} = \begin{matrix} 4i & 0 \\ 1 & 4i \end{matrix}.$$
Since $\mathbf{A}_2 = \quad 2i\mathbf{A}$, it follows that $\mathbf{A}_2$ is a matrix associated with M. □

Problem 2 Decide which, if any, of the matrices
$$\mathbf{A}_1 = \begin{matrix} 0 & 1 \\ \frac{1}{2}i & 0 \end{matrix}, \quad \mathbf{A}_2 = \begin{matrix} 0 & 2 \\ 1 & 2i \end{matrix}, \quad \mathbf{A}_3 = \begin{matrix} 2i & 0 \\ 1 & i \end{matrix}$$

are associated with each of the following Möbius transformations M.

(a) $M_1(z) = \dfrac{2i}{iz+2}$ (b) $M_2(z) = \dfrac{2i}{z}$ (c) $M_3(z) = \dfrac{iz+2}{2z-i}$

A particularly important transformation of $\mathbb{C}$ is the *identity* function defined by

$$t(z) = z \quad (z \in \mathbb{C}).$$

This is a Möbius transformation because it can be written in the form

$$t(z) = \frac{z+0}{0z+1}.$$

A matrix associated with this Möbius transformation is the identity matrix

$$\mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

5.3.5 Composing Möbius Transformations

The next example illustrates what happens when two Möbius transformations are composed.

Example 2 Determine the composite $M_1 \circ M_2$, where M_1 and M_2 are the Möbius transformations defined by

$$M_1(z) = \frac{iz+1}{2z-2} \quad \text{and} \quad M_2(z) = \frac{z+i}{2z-1}.$$

Solution Since M_1 and M_2 are one-one mappings of $\mathbb{C}$ onto $\mathbb{C}$, the same must be true of $M_1 \circ M_2$. A formula for $M_1 \circ M_2(z)$ is

$$\begin{aligned} M_1 \circ M_2(z) &= M_1\left(\frac{z+i}{2z-1}\right) \\ &= \frac{i\left(\frac{z+i}{2z-1}\right)+1}{2\left(\frac{z+i}{2z-1}\right)-2} \\ &= \frac{i(z+i)+(2z-1)}{2(z+i)-2(2z-1)} \\ &= \frac{(2+i)z-2}{-2z+(2+2i)}. \end{aligned}$$

We can check that this formula defines a Möbius transformation by noting that

$$(2+i)(2+2i) - (-2)(-2) = -2+6i \neq 0.$$

Strictly speaking, we should check our convention for the exceptional points separately. For example, by convention $M_2(\infty) = \frac{1}{2}$, so that

$$\begin{aligned} M_1 \circ M_2(\infty) &= M_1\left(\tfrac{1}{2}\right) \\ &= \frac{\frac{1}{2}i+1}{1-2} \\ &= \frac{2+i}{-2}. \end{aligned}$$

This agrees with the value obtained when we apply our convention to the formula for $M_1 \circ M_2$.

Thus $M_1 \circ M_2$ is a Möbius transformation given by

$$M_1 \circ M_2(z) = \frac{(2+i)z - 2}{2z + (2+2i)}. \qquad \square$$

Problem 3 Write down matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ associated with the Möbius transformations M_1 and M_2 defined in Example 2. Calculate the product $\mathbf{A}_1\mathbf{A}_2$, and compare it with the Möbius transformation $M_1 \circ M_2$.

In Example 2, the composite of two Möbius transformations M_1 and M_2 turned out to be another Möbius transformation. The solution to Problem 3 demonstrates that the product of matrices associated with M_1 and M_2 is a matrix associated with $M_1 \circ M_2$. The following theorem confirms that this is always the case.

Theorem 6 Composition of Möbius Transformations

Let M_1 and M_2 be Möbius transformations with associated matrices $\mathbf{A}_1$ and $\mathbf{A}_2$, respectively. Then $M_1 \circ M_2$ is a Möbius transformation with an associated matrix $\mathbf{A}_1\mathbf{A}_2$.

Proof Let M_1 and M_2 be defined by

$$M_1(z) = \frac{az+b}{cz+d} \quad \text{and} \quad M_2(z) = \frac{ez+f}{gz+h}.$$

Since M_1 and M_2 are one–one mappings of $\hat{\mathbb{C}}$ onto $\hat{\mathbb{C}}$, the same must be true of $M_1 \circ M_2$. We can obtain a formula for $M_1 \circ M_2(z)$ as follows:

$$\begin{aligned}
M_1 \circ M_2(z) &= M_1\left(\frac{ez+f}{gz+h}\right)\\
&= \frac{a\left(\frac{ez+f}{gz+h}\right) + b}{c\left(\frac{ez+f}{gz+h}\right) + d}\\
&= \frac{a(ez+f) + b(gz+h)}{c(ez+f) + d(gz+h)}\\
&= \frac{(ae+bg)z + (af+bh)}{(ce+dg)z + (cf+dh)}.
\end{aligned}$$

Strictly speaking, we should check the exceptional cases separately. However, we shall omit this checking from here onwards.

Since $M_1 \circ M_2$ is one–one, this formula cannot remain constant as z varies, so $(ae+bg)(cf+dh) - (af+bh)(ce+dg) \neq 0$. The formula must therefore define a Möbius transformation.

A matrix associated with $M_1 \circ M_2$ is

$$\begin{pmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{pmatrix},$$

which is the product of the matrices

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} e & f \\ g & h \end{pmatrix}$$

associated with M_1 and M_2, respectively.

Theorem 6 enables us to calculate the composite of two Möbius transformations by using the following strategy.

Strategy To compose two Möbius transformations M_1 and M_2:

1. write down matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ associated with M_1 and M_2;
2. calculate $\mathbf{A}_1\mathbf{A}_2$;
3. write down the Möbius transformation $M_1 \circ M_2$ with which $\mathbf{A}_1\mathbf{A}_2$ is associated.

Example 3 Use the strategy to determine the composite $M_1 \circ M_2$ of the Möbius transformations

$$M_1(z) = \frac{3z+1}{iz-2} \quad \text{and} \quad M_2(z) = \frac{2iz+3}{z-2}.$$

Solution The Möbius transformations M_1 and M_2 have associated matrices

$$\mathbf{A}_1 = \begin{pmatrix} 3 & 1 \\ i & -2 \end{pmatrix} \quad \text{and} \quad \mathbf{A}_2 = \begin{pmatrix} 2i & 3 \\ 1 & -2 \end{pmatrix},$$

respectively. It follows that a matrix associated with $M_1 \circ M_2$ is

$$\mathbf{A}_1\mathbf{A}_2 = \begin{pmatrix} 3 & 1 \\ i & -2 \end{pmatrix} \begin{pmatrix} 2i & 3 \\ 1 & -2 \end{pmatrix} = \begin{pmatrix} 1+6i & 7 \\ -4 & 4+3i \end{pmatrix}.$$

The composite $M_1 \circ M_2$ is therefore the Möbius transformation defined by

$$M_1 \circ M_2(z) = \frac{(1+6i)z+7}{-4z+(4+3i)}.$$ □

Problem 4 Let M_1 and M_2 be the Möbius transformations defined by

$$M_1(z) = \frac{3z+1}{iz-2} \quad \text{and} \quad M_2(z) = \frac{2iz+3}{z-2}.$$

Use the strategy to determine each of the following composites.

(a) $M_2 \circ M_1$ (b) $M_1 \circ M_1$

Problem 5 Let M_1 and M_2 be the Möbius transformations defined by

$$M_1(z) = \frac{z - i}{iz + 2} \quad \text{and} \quad M_2(z) = \frac{2z + i}{-iz + 1}.$$

Use the strategy to determine the composite $M_1 \circ M_2$.

5.3.6 Inverting Möbius Transformations

In the solution to Problem 5 you saw that the composite $M_1 \circ M_2$ of the Möbius transformations

$$M_1(z) = \frac{z - i}{iz + 2} \quad \text{and} \quad M_2(z) = \frac{2z + i}{-iz + 1}$$

is the identity function on $\mathbb{C}$. This shows that M_2 is the inverse function of M_1. In terms of matrices, this is equivalent to saying that there are matrices associated with M_1 and M_2 whose product is the identity matrix $\mathbf{I}$. For example,

$$\begin{pmatrix} 1 & -i \\ i & 2 \end{pmatrix} \begin{pmatrix} 2 & i \\ -i & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We can use this idea to find the inverse of any given Möbius transformation M. The inverse function M^{-1} certainly exists, since M is a one-one transformation from $\mathbb{C}$ onto $\mathbb{C}$. To find the inverse, let $\mathbf{A}$ be a matrix associated with M. Since $\mathbf{A}$ is invertible,

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I} = \mathbf{A}^{-1}\mathbf{A},$$

so the Möbius transformation associated with the matrix $\mathbf{A}^{-1}$ must be the inverse function M^{-1}.

Now, if $M(z) = \frac{az+b}{cz+d}$, we can take $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, so that

$$\mathbf{A}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

But any non-zero multiple of $\mathbf{A}^{-1}$ is also a matrix associated with M^{-1}, so we shall usually use the matrix $\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$ as a matrix for M^{-1}. Then

$$M^{-1}(z) = \frac{dz - b}{-cz + a}.$$

We summarize the result of this discussion in the following theorem.

Theorem 7 Inverse of a Möbius Transformation
The inverse of the Möbius transformation

$$M(z) = \frac{az + b}{cz + d}$$

is also a Möbius transformation, and it may be written in the form

$$M^{-1}(z) = \frac{dz - b}{-cz + a}.$$

For example, the inverse of the Möbius transformation

$$M(z) = \frac{iz + 1}{2z - 2}$$

is given by

$$M^{-1}(z) = \frac{-2z - 1}{-2z + i} = \frac{2z + 1}{2z - i}.$$

Problem 6 Determine the inverse of each of the following Möbius transformations.

(a) $M_1(z) = \frac{3iz+2}{z-3i}$ (b) $M_2(z) = \frac{4i}{z}$ (c) $M_3(z) = \frac{4iz}{z-4i}$

5.3.7 The Inversive Group

We now have all the information we need to prove the following theorem.

Theorem 8 The set of all Möbius transformations forms a group under composition of functions.

Proof We show that the four group axioms hold.

G1 CLOSURE By Theorem 6, the composite of two Möbius transformations is itself a Möbius transformation.
G2 IDENTITY The identity is the Möbius transformation given by $M(z) = \frac{1z+0}{0z+1}$.
G3 INVERSES By Theorem 7, every Möbius transformation has an inverse.
G4 ASSOCIATIVITY Composition of functions is always associative.

It follows that the set of all Möbius transformations forms a group under composition of functions.

Having shown that the set of Möbius transformations forms a group, we now investigate its relationship with the group of all inversive transformations. We know from Theorem 4 that every Möbius transformation is an inversive transformation, but is every inversive transformation a Möbius transformation? Clearly, the answer is no. For example, an inversion cannot be a Möbius transformation since it *reverses* the orientation of angles, whereas all Möbius transformations *preserve* the orientation of angles.

Although inversions are not Möbius transformations, there is a close connection between inversions and Möbius transformations.

Theorem 9 Every inversion t has the form $t(z) = M(\bar{z})$, where M is a Möbius transformation.

Proof If t is an inversion of $\mathbb{C}$ in an extended line, then by Theorem 1 of Section 5.2 it must have the form $t(z) = a\bar{z} + b$, with $t(\infty) = \infty$. It follows that $t(z) = M(\bar{z})$, where M is the Möbius transformation

$$M(z) = \frac{az + b}{0z + 1}.$$

It cannot have the form $t(z) = az + b$, since t reverses the orientation of angles.

On the other hand, if t is an inversion of $\mathbb{C}$ in a circle of radius r with centre c, then by Theorem 3 of Section 5.2,

$$t(z) = \frac{r^2}{\overline{z - c}} + c = \frac{r^2 + c(\bar{z} - \bar{c})}{\bar{z} - \bar{c}} = \frac{c\bar{z} + (r^2 - c\bar{c})}{\bar{z} - \bar{c}}.$$

So once again t has the form $t(z) = M(\bar{z})$, where M is the Möbius transformation

$$M(z) = \frac{cz + r^2 - c\bar{c}}{z - \bar{c}}.$$

Notice that

$$c \cdot (-\bar{c}) - (r^2 - c\bar{c}) \cdot 1 = -r^2 \neq 0.$$

We can now show that every inversive transformation t has the form $t(z) = M(z)$ or $t(z) = M(\bar{z})$, where M is a Möbius transformation.

Theorem 10 Every inversive transformation t can be represented in $\mathbb{C}$ by one of the formulas

$$t(z) = \frac{az + b}{cz + d} \quad \text{or} \quad t(z) = \frac{a\bar{z} + b}{c\bar{z} + d},$$

where $a, b, c, d \in \mathbb{C}$ and $ad - bc \neq 0$.

Proof We first show that the composite of two inversions t_1 and t_2 is a Möbius transformation. By Theorem 9 above, we can write $t_1(z) = M_1(\bar{z})$ and $t_2(z) = M_2(\bar{z})$, where M_1 and M_2 are Möbius transformations. Thus

$$t_1 \circ t_2(z) = t_1(M_2(\bar{z})) = M_1(\overline{M_2(\bar{z})}).$$

But if

$$M_2(z) = \frac{az + b}{cz + d},$$

then we can define a Möbius transformation M_3 by

$$M_3(z) = \overline{M_2(\bar{z})} = \overline{\frac{a\bar{z} + b}{c\bar{z} + d}} = \frac{\bar{a}z + \bar{b}}{\bar{c}z + \bar{d}}.$$

Since $t_1 \circ t_2(z) = M_1(\overline{M_2(\bar{z})}) = M_1(M_3(z))$, it follows that $t_1 \circ t_2$ is a composite of two Möbius transformations, and is therefore a Möbius transformation.

Next, let t be an arbitrary inversive transformation, and write

$$t = t_1 \circ t_2 \circ \ldots \circ t_n,$$

where $t_1, t_2, \ldots, t_n$ are inversions. If n is even, then we can rewrite t as a composite of Möbius transformations by pairing together the inversions in the form

$$t = (t_1 \circ t_2) \circ (t_3 \circ t_4) \circ \ldots \circ (t_{n-1} \circ t_n).$$

It follows that t is a Möbius transformation, M say, so we can write $t(z) = M(z)$.

To deal with the case where n is odd, let r be the extended conjugation function. Since r is its own inverse, we can write

$$t = (t_1 \circ t_2 \circ \ldots \circ t_n \circ r) \circ r.$$

Here the composite in the bracket involves an even number of inversions, so it must be a Möbius transformation, M say. Hence $t(z) = M \circ r(z) = M(\bar{z})$.

Theorem 10 provides us with an insight into the structure of the group G of all inversive transformations. Those inversive transformations that can be written in the form

$$t(z) = \frac{az + b}{cz + d}$$

are Möbius transformations, and we know from Theorem 8 that they form a subgroup of G. We refer to these transformations as **direct** inversive transformations because they preserve the orientation of angles.

The remaining inversive transformations are of the form

$$t(z) = \frac{a\bar{z} + b}{c\bar{z} + d}.$$

These are Möbius transformations composed with the extended conjugation function, and so they reverse the orientation of angles. For this reason we shall refer to them as **indirect** inversive transformations.

5.3.8 Images of Generalized Circles

We have shown that an inversive transformation maps generalized circles onto generalized circles. However, apart from the special case of inversion in the unit circle, we have not yet given a strategy for finding the image of a generalized circle under an inversive transformation.

One such strategy is based on the fact that every generalized circle in C is uniquely determined by any three points lying on it. Indeed, if the three points lie in C, and are non-collinear, then they determine a unique circle in C. On the other hand, if the three points are collinear, or if one of the points is , then they determine a unique (extended) line.

Strategy To determine the image of a generalized circle C under an inversive transformation t:

1. write down three points z_1, z_2, z_3 on C;
2. determine the images $t(z_1), t(z_2), t(z_3)$;
3. the image $t(C)$ is the (unique) generalized circle through $t(z_1)$, $t(z_2)$, $t(z_3)$.

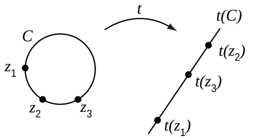

Example 4 Use the strategy to find the image of the unit circle C under the inversive transformation defined by

$$t(z) = \frac{\bar{z} + i}{\bar{z}\quad 1}.$$

Solution We first pick three distinct points on the unit circle. There is no definite rule about which points should be chosen, so to keep the calculations simple, we pick the points 1, i and 1, as shown below. Now

$$t(1) = \quad,\ t(i) = \frac{\bar{i} + i}{\bar{i}\quad 1} = 0 \quad \text{and}$$

$$t(\quad 1) = \frac{\overline{\quad 1} + i}{\overline{\quad 1}\quad 1} = \frac{1}{2}(1 \quad i).$$

If you can pick a point that maps to then you know immediately that the image is an extended line.

So the image of C is the generalized circle through the points , 0 and $\frac{1}{2}(1 \quad i)$. This is an extended line through the origin with slope 1. □

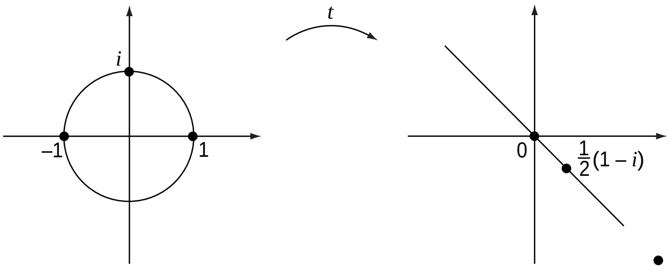

Problem 7 Let t be the inversive transformation defined by

$$t(z) = \frac{z - 2i}{z + 2}.$$

Use the strategy to determine the image of each of the following generalized circles under t:

(a) the extended line $L \cup \{\infty\}$, where L is the line with equation $x + y = 2$;
(b) the circle with equation $(x + 1)^2 + y^2 = 1$.

Here $x + iy = z$, as usual.

5.4 Fundamental Theorem of Inversive Geometry

5.4.1 Comparison with Affine Geometry

In Subsection 2.3.2 we introduced the Fundamental Theorem of Affine Geometry, which states that, given any three non-collinear points in the plane, there is always an affine transformation that maps the points to another three given non-collinear points. Since a triangle is uniquely determined by its three (non-collinear) vertices, and since affine transformations map triangles to triangles, we were able to use the Fundamental Theorem of Affine Geometry to show that all triangles are affine-congruent.

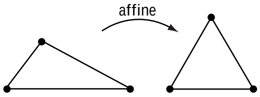

In this section we prove an analogous *Fundamental Theorem of Inversive Geometry* which states that given any three points in the extended complex plane $\hat{\mathbb{C}}$ there is always an inversive transformation (in fact a Möbius transformation) that maps them to another three given points in the extended complex plane. Since a generalized circle is uniquely determined by any three of its points, and since inversive transformations map generalized circles to generalized circles, we can show that all generalized circles are inversive-congruent. That is, given any two generalized circles C_1 and C_2 in the extended plane, it is always possible to find an inversive transformation (in fact a Möbius transformation) that maps C_1 onto C_2.

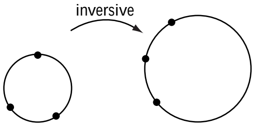

5.4.2 Mapping Three Points to Three Points

The proof of the Fundamental Theorem of Inverse Geometry is very similar to the proof of the Fundamental Theorem of Affine Geometry which we gave in Subsection 2.3.2. There we described how to construct an affine transformation which maps one given set of three (non-collinear) points in $\mathbb{R}^2$ onto another. We did this by forming a composite of two affine transformations which map the points via the auxiliary points (0, 0), (1, 0) and (0, 1).

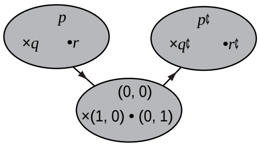

A similar idea works for the Fundamental Theorem of Inversive Geometry. In this geometry the space is $\hat{\mathbb{C}}$ rather than $\mathbb{R}^2$, and we map via the auxiliary points 0, 1 and ∞.

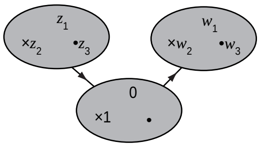

So let us start by considering how to construct a Möbius transformation that maps three given points z_1, z_2, z_3 in $\hat{\mathbb{C}}$ to the points $0, 1, \infty$, respectively.

First observe that if z_1, z_2, z_3 belong to C, then any Möbius transformation of the form

$$M(z) = K\frac{z \quad z_1}{z \quad z_3}$$

Here K can be any complex number.

maps z_1 to 0 and z_3 to . We can therefore obtain the required Möbius transformation by choosing K so that $M(z_2) = 1$. The following example illustrates how this is done. It also shows how the method can be modified to deal with cases where one of z_1, z_2, z_3 is .

Example 1 For each set of three points given below, determine a Möbius transformation M that maps the points to 0, 1, , respectively.

(a) $\frac{1}{2}$, 1, 3 (b) , i, 2 (c) i, , 3 (d) 5, 2,

Solution

(a) To ensure that $M\ \frac{1}{2}\ = 0$ and $M(3) =$ we let M have the form

$$M(z) = K\frac{z \quad \frac{1}{2}}{z \quad 3},$$

for some complex number K. Since $M(\ \ 1) = 1$, we must have

$$1 = K\frac{1 \quad \frac{1}{2}}{1 \quad 3},$$

This has the form

$$M(z) = \frac{az + b}{cz + d}$$

with $a = K$, $b = \ \ \frac{1}{2}K$, $c = 1$ and $d = \ \ 3$.

so that $K = \frac{8}{3}$. The required Möbius transformation is therefore

$$M(z) = \frac{8z \quad 4}{3z \quad 9}.$$

(b) To ensure that $M(\ \ ) = 0$ and $M(2) =$ we let M have the form

$$M(z) = \frac{K}{z \quad 2}.$$

Since $M(i) = 1$, we must have

$$1 = \frac{K}{i \quad 2},$$

so that $K = i \quad 2$. It follows that

$$M(z) = \frac{i \quad 2}{z \quad 2}.$$

This has the form

$$M(z) = \frac{az + b}{cz + d}$$

with $a = 0$, $b = K$, $c = 1$ and $d = \ \ 2$. Since $c = 0$, it follows from the definition of a Möbius transformation that $M(\ \ ) = a/c = 0$, as required.

(c) To ensure that $M(i) = 0$ and $M(3) =$ we let M have the form

$$M(z) = \frac{z \quad i}{z \quad 3}.$$

In this case there is no need to include a constant K because $M(\ \ )$ is already equal to 1.

This has the form

$$M(z) = \frac{az+b}{cz+d}$$

with $a = 1$, $b = \ \ i$, $c = 1$ and $d = \ \ 3$. Since $c = 0$, we have $M(\ \ ) = a/c = 1$, as required.

(d) Here we require $M(5) = 0$ and $M(\quad) = \quad$, so M is an extended linear function of the form

$$M(z) = K(z \quad 5).$$

Since $M(2) = 1$, we must have

$$1 = K(2 \quad 5),$$

so that $K = \quad \frac{1}{3}$. It follows that

$$M(z) = \quad \frac{1}{3}(z \quad 5). \qquad \square$$

This has the form

$$M(z) = \frac{az+b}{cz+d}$$

with $a = K$, $b = \quad 5K$, $c = 0$ and $d = 1$. Since $c = 0$, it follows from the definition of a Möbius transformation that $M(\quad) = \quad$.

Guided by the above example we can summarize all the possible cases to obtain the following general strategy.

Strategy To determine the Möbius transformation M which maps three given points z_1, z_2, z_3 onto the points $0, 1, \quad$ respectively:

1. choose the appropriate form of mapping from the following formulas for M:
   for $z_1, z_2, z_3 \quad 0, 1, \quad$ use $M(z) = K\frac{z \quad z_1}{z \quad z_3}$;
   for $\quad, z_2, z_3 \quad 0, 1, \quad$ use $M(z) = \frac{K}{z \quad z_3}$;
   for $z_1, \quad, z_3 \quad 0, 1, \quad$ use $M(z) = \frac{z \quad z_1}{z \quad z_3}$;
   for $z_1, z_2, \quad 0, 1, \quad$ use $M(z) = K(z \quad z_1)$;
2. find the complex number K for which $M(z_2) = 1$.

Problem 1 For each set of three points given below, determine a Möbius transformation M that maps the points to $0, 1, \quad$, respectively.

(a) $\quad 1, \quad 3, 0$ (b) $\frac{3}{2}, 2, 1$ (c) $\quad, \quad 3, 2$ (d) $\frac{3}{2}, 2, \quad$

We are now in a position to prove the main theorem of this section.

Theorem 1 The Fundamental Theorem of Inversive Geometry
Let z_1, z_2, z_3 and w_1, w_2, w_3 be two sets of three points in the extended complex plane C. Then there is a unique Möbius transformation M which maps z_1 to w_1, z_2 to w_2, and z_3 to w_3.

Proof According to the above strategy there is a Möbius transformation M_1 which maps the points z_1, z_2, z_3 to the points $0, 1, \quad$, respectively. Similarly, there is a Möbius transformation M_2 which maps the points w_1, w_2, w_3 to the points $0, 1, \quad$, respectively.

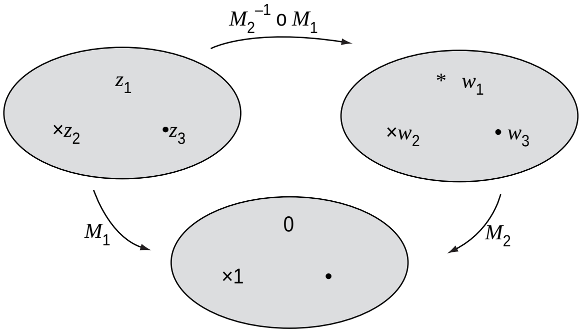
$M_2^{-1} \circ M_1$
z_1
$\times z_2$
$\bullet z_3$
$*$ w_1
$\times w_2$
$\bullet w_3$
M_1
0
$\times 1$
M_2

The composite $M = M_2^{-1} \circ M_1$ is therefore the required Möbius transformation which maps z_1, z_2, z_3 to the points w_1, w_2, w_3, respectively.

To check uniqueness we first observe that the identity is the only Möbius transformation which maps each of the points $0, 1, \infty$ to themselves. Indeed, if

$$M(z) = \frac{az + b}{cz + d}$$

is a Möbius transformation which maps ∞ to itself, then $c = 0$. Also, if M maps 0 to itself, then $M(0) = b/d = 0$, so $b = 0$. It follows that $M(z) = (a/d)z$. But if M maps 1 to itself, then $M(1) = a/d = 1$, which implies that $a = d$. Putting all this together we conclude that M is the identity $M(z) = z$, as required.

Next suppose that M and M' are two Möbius transformations which satisfy the conditions of the theorem.

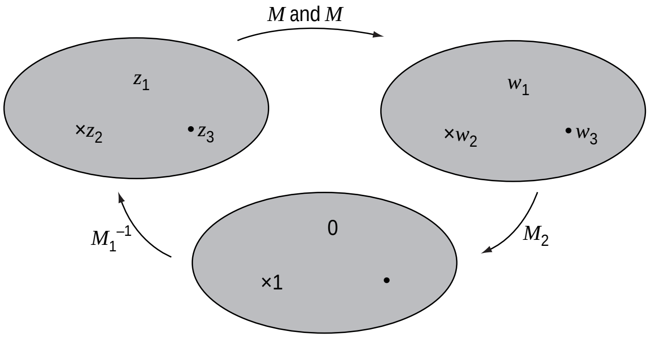
M and M'
z_1
$\times z_2$
$\bullet z_3$
w_1
$\times w_2$
$\bullet w_3$
M_1^{-1}
0
$\times 1$
M_2

Then the composites $M_2 \circ M \circ M_1^{-1}$ and $M_2 \circ M' \circ M_1^{-1}$ must both be Möbius transformations which map each of the points $0, 1, \infty$ to themselves. Since this implies that both composites are equal to the identity, we can write

$$M_2 \circ M \circ M_1^{-1} = M_2 \circ M' \circ M_1^{-1}.$$

If we now compose both sides of this equation with M_2^{-1} on the left and with M_1 on the right, then we obtain $M = M'$, as required.

If we actually need to find the Möbius transformation which maps one set of three points onto another set of three points we simply follow the strategy used to prove part (a) of the Fundamental Theorem.

Strategy To determine the Möbius transformation M which maps the points z_1, z_2, z_3 to the points w_1, w_2, w_3, respectively:

1. find the Möbius transformation M_1 which maps the points z_1, z_2, z_3 to the points 0, 1, , respectively;
2. find the Möbius transformation M_2 which maps the points w_1, w_2, w_3 to the points 0, 1, , respectively;
3. calculate $M = M_2^{\ 1} \quad M_1$.

The following example illustrates how to implement this strategy.

Example 2 Find the Möbius transformation M which maps the points i, , 3 to the points $\frac{1}{2}$, 1, 3, respectively.

Solution We follow the steps in the above strategy.

1. From part (c) of Example 1, we know that the Möbius transformation M_1 which maps the points i, , 3 to the points 0, 1, , respectively, is given by

$$M_1(z) = \frac{z \quad i}{z \quad 3}.$$

2. Also, from part (a) of Example 1, we know that the Möbius transformation M_2 which maps the points $\frac{1}{2}$, 1, 3 to the points 0, 1, , respectively, is given by

$$M_2(z) = \frac{8z \quad 4}{3z \quad 9}.$$

3. Matrices associated with M_1 and M_2 are

$$\begin{matrix} 1 & i \\ 1 & 3 \end{matrix} \quad \text{and} \quad \begin{matrix} 8 & 4 \\ 3 & 9 \end{matrix} \ ;$$

also, by Theorem 7 of Subsection 5.3.6, a matrix associated with the inverse of M_2 is

$$\begin{matrix} 9 & 4 \\ 3 & 8 \end{matrix} \ .$$

Hence, a matrix associated with $M = M_2^{\ l} \quad M_1$ is given by

$$\begin{matrix} 9 & 4 \\ 3 & 8 \end{matrix} \quad \begin{matrix} 1 & i \\ 1 & 3 \end{matrix} = \begin{matrix} 5 & 12 + 9i \\ 5 & 24 + 3i \end{matrix} \ .$$

The required Möbius transformation is therefore

$$M(z) = \frac{-5z - 12 + 9i}{5z - 24 + 3i}. \qquad \square$$

As a check, you can verify that $M(i) = \frac{1}{2}$, $M(\infty) = -1$ and $M(3) = 3$.

Problem 2

(a) Find the Möbius transformation which maps the points $-1, i, 1$ to the points $-1, -3, 0$, respectively.

(b) Find the Möbius transformation which maps the points $3, \infty, -2$ to the points $3, \frac{7}{3}, 1$, respectively.

5.4.3 Circles through Four Points

We now explore an application of the Fundamental Theorem which enables us to determine whether four given points z_1, z_2, z_3 and z_4 lie on some generalized circle.

If there is a generalized circle that passes through z_1, z_2, z_3 and z_4, then it must be the unique generalized circle C that passes through z_1, z_2 and z_3. We have to decide whether C also passes through z_4. To do this, consider the Möbius transformation M that maps z_1, z_2, z_3 to $0, 1, \infty$, respectively. Under M the image of C is the extended real axis, for this is the only generalized circle that passes through 0, 1 and ∞. If z_4 lies on C, then its image under M must lie on the real axis; whereas if z_4 does not lie on C, then its image under M cannot lie on the real axis.

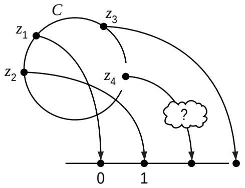

Example 3 Determine whether the four points i, $1 + 4i$, 3, $4 + 3i$ lie on a generalized circle.

Solution First, we determine the Möbius transformation M that maps i, $1 + 4i$, 3 to $0, 1, \infty$, respectively. Following the strategy in the previous subsection, we observe that this transformation must be of the form

$$M(z) = K\frac{z - i}{z - 3}$$

for some complex number K. Since $M(1 + 4i) = 1$, we must have

$$1 = K\frac{(1 + 4i) - i}{(1 + 4i) - 3} = K\frac{1 + 3i}{-2 + 4i},$$

so

$$K = \frac{-2 + 4i}{1 + 3i} = \frac{(-2 + 4i)(1 - 3i)}{10} = 1 + i.$$

You met this method of simplifying a quotient by multiplying both numerator and denominator by the conjugate of the denominator earlier, in Subsection 5.2.1.

Thus the transformation M is given by

$$M(z) = \frac{(z - i)(1 + i)}{z - 3}.$$

It follows that

$$M(4+3i)=\frac{(4+2i)(1+i)}{1+3i}=\frac{2+6i}{1+3i}=2.$$

Since this is real, it follows that i, $1+4i$, 3, $4+3i$ do all lie on a generalized circle. □

In this example we showed that the points i, $1+4i$, 3, $4+3i$ all lie on a generalized circle; but suppose we want to show that the points lie on an *ordinary* circle. To do this we need to check that the generalized circle that passes through the points is not an extended line. If it were an extended line, then it would pass through , and so $M(\ \ )$ would be real. Since $M(\ \ )=1+i$ is not real, it follows that the points do indeed lie on an ordinary circle.

$M(\ \ )$ cannot be , as M is one one and $M(3)=$.

Strategy To determine whether z_1, z_2, z_3 and z_4 lie on a circle:

1. find the Möbius transformation M which maps z_1, z_2, z_3 to 0, 1, , respectively;
2. the points z_1, z_2, z_3, z_4, lie on a generalized circle if and only if $M(z_4)$ is real;
3. the generalized circle in Step 2 is a circle provided that $M(\ \ )$ is not real.

If $M(z_4)$ and $M(\ \ )$ are both real, then z_1, z_2, z_3, z_4 lie on a line.

Problem 3 Determine whether each of the following sets of four points lies on a circle.

(a) 0, 4, 2i, 1 3i (b) 1, i, i, 2 i

5.4.4 Congruence of Generalized Circles

We end this section by stating an important consequence of the Fundamental Theorem of Inversive Geometry.

Theorem 2 Let C_1 and C_2 be generalized circles in the extended complex plane. Then there is a Möbius transformation that maps C_1 onto C_2.

Proof Let a, b, c be any three points on C_1 and let d, e, f be any three points on C_2. By the Fundamental Theorem of Inversive Geometry, there is a Möbius transformation M that maps a, b, c to d, e, f, respectively.

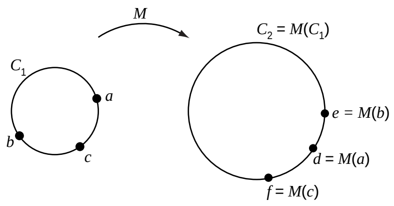

Since M maps generalized circles to generalized circles, it must map C_1 to a generalized circle through d, e and f. But a generalized circle is uniquely determined by any three of its points, so C_2 is the only generalized circle that passes through d, e and f. It follows that M maps C_1 onto C_2.

One of the remarkable consequences of Theorem 2 is that all generalized circles are inversive-congruent. In particular, an ordinary circle is inversive-congruent to any given (extended) line. Since lines are often easier to investigate than circles, we can sometimes simplify a problem that involves the inversive properties of a figure by using an inversive transformation to map one or more of the circles in a figure onto (extended) lines. You will see several examples of this in the next section.

5.5 Coaxal Families of Circles

We now prove some lovely theorems about circles, where the beauty of the final results is matched by the beauty of the proofs themselves. Inversion is our key tool.

Sadly this is not always the case in Mathematics!

5.5.1 Apollonian Circles Theorem

We start with the proof of the Apollonian Circles Theorem, which we stated at the beginning of the chapter.

Theorem 1 Apollonian Circles Theorem
Let A and B be two distinct points in the plane, and let k be a positive real number other than 1. Then the locus of points P that satisfy $PA : PB = k : 1$ is a circle whose centre lies on the line through A and B.

Recall that if $k = 1$, then P must lie on the perpendicular bisector of AB.

We give two proofs of this result. The first uses methods from Euclidean geometry, and has the advantage of providing us with equations for the circles in terms of k. The second proof uses the methods of inversive geometry, and has the advantage of providing us with a deeper insight into the geometry of the circles.

The second proof is given later.

First Proof

To keep the algebra simple, we introduce x- and y-axes into the plane such that A and B have coordinates $(-a, 0)$ and $(a, 0)$, respectively, where $a > 0$.

Now fix a value of $k > 0$, $k \neq 1$, and let C be the locus of points P that satisfy $PA : PB = k : 1$. Then a point $P(x, y)$ belongs to C if and only if $PA = k \bullet PB$, and since $k > 0$ this is equivalent to the equation

$$PA^2 = k^2 \bullet PB^2.$$

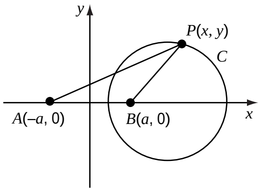

Using the Euclidean formula for distance between points, we see that this last equation holds if and only if

$$(x + a)^2 + y^2 = k^2((x - a)^2 + y^2).$$

Multiplying out the brackets and collecting terms in x^2, y^2 and x, we see that this is equivalent to

$$x^2(1 \quad k^2) + y^2(1 \quad k^2) + 2ax\,(1 + k^2) + a^2(1 \quad k^2) = 0.$$

Since $k = 1$, this holds if and only if

$$x^2 + y^2 + 2a \quad \frac{1 + k^2}{1 \quad k^2} \quad x + a^2 = 0.$$

It follows that the locus C is a circle with centre c and radius r where

$$c = \quad a \quad \frac{1 + k^2}{1 \quad k^2} \quad , 0$$

and

$$r = \overline{\quad a^2 \quad \frac{1 + k^2}{1 \quad k^2}^{\,2} \quad a^2} = \frac{2ak}{|1 \quad k^2|}.$$

Here we are using the formulas for centre and radius given in Theorem 2, Subsection 1.1.2.

In particular, the centre lies on the x-axis, which by our choice of axes is the line through A and B.

Notice that the centre of the circle is *not* at either of the points A or B.

If $k = 1$, then the locus of points that satisfy $PA = k \bullet PB$ is the line that bisects AB at right angles. If we adopt the convention that this locus includes the point , then we can think of this locus as the extended line $\{ \}$. With this convention, every positive value of k gives rise to a generalized circle known as a **circle of Apollonius**. The family of all such circles is known as the **Apollonian family of circles** defined by the points A and B.

A circle of Apollonius is often referred to as an Apollonian circle .

As k increases through the interval $(0, \quad)$, the corresponding circles range through the Apollonian family.

When $0 < k < 1$, we have that $PA < PB$, and so P is closer to A than it is to B. The circles that correspond to these values of k therefore lie on the same side of as does A. For values of k close to 0, $PA = k \bullet PB$ is small, and so the corresponding circles are close to A. As k tends to 1, the circles grow and become ever closer to .

When $k = 1$, the corresponding circle is the extended line $\{ \}$.

When $k > 1$, we have that $PA > PB$, so that P is closer to B than it is to A. The circles that correspond to these values of k therefore lie on the same side of as does B. As k increases from 1, $PB = (1/k) \bullet PA$ becomes smaller, and the corresponding circles shrink and become ever closer to B.

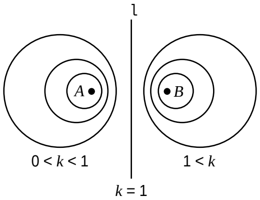

In light of the above discussion, it is sometimes convenient to refer to the points A and B as **point circles** corresponding to the cases $k = 0$ and $k = \quad$, respectively. With this convention, every point of the plane belongs to a unique circle associated with the Apollonian family.

Example 1 For the Apollonian family defined by the point circles $(\quad 1, 0)$ and $(1, 0)$, determine the equation of the circle in the family that passes through the point $(2, 1)$.

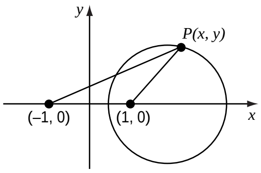

Solution Let P be a point (x, y) in the plane whose distance from the point $(-1, 0)$ is k times its distance from the point $(1, 0)$. Then, if we use the Euclidean formula for distance between points in the plane, we obtain

$$(x+1)^2 + y^2 = k^2((x-1)^2 + y^2).$$

For each value of k this yields an equation for the corresponding Apollonian circle. The point $(2, 1)$ lies on the Apollonian circle for which

$$(2+1)^2 + 1^2 = k^2((2-1)^2 + 1^2),$$

that is, $k = \sqrt{5}$.

The equation of the Apollonian circle through the point $(2, 1)$ is therefore

$$(x+1)^2 + y^2 = 5((x-1)^2 + y^2),$$

which simplifies to

$$x^2 + y^2 - 3x + 1 = 0. \qquad \square$$

Problem 1 For the Apollonian family defined by the point circles $(0, -1)$ and $(0, 2)$, determine the equation of the circle in the family that passes through the point $(1, 1)$.

Problem 2 The circles C_1 and C_2 in an Apollonian family of circles have the segments $[-18, -2]$ and $[3, 12]$, respectively, as diameter. Determine the point circles in the family, and hence the equation of the Apollonian circle in the family that passes through the point $(6, 9)$.

We now give a second proof of the Apollonian Circles Theorem, using inversive geometry.

Theorem 1 Apollonian Circles Theorem
Let A and B be two given points in the plane, and let k be a positive real number other than 1. Then the locus of points P that satisfy $PA : PB = k : 1$ is a circle whose centre lies on the line through A and B.

Second Proof

Let C be the locus of points P that satisfy $PA = k \cdot PB$, and let t be the inversion in the circle with centre A and radius 1. We shall show that $C' = t(C)$ is a circle, and hence that $C = t^{-1}(C')$ is a generalized circle.

We specify radius 1 here simply for definiteness; any positive radius would serve equally well.

Let P be an arbitrary point in the extended plane, and let $t(B) = B'$, $t(P) = P'$. Then by the definition of inversion,

$$AB \cdot AB' = 1 \quad \text{and} \quad AP \cdot AP' = 1, \tag{1}$$

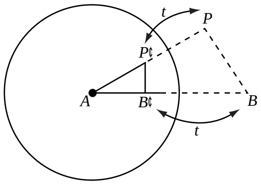

so that

$$\frac{AB}{AP} = \frac{AP}{AB}.$$

This shows that the sides AB and AP of APB are proportional to the sides AP and AB of $AB\ P$. Since BAP is the same as $P\ AB$, it follows that APB is similar to $AB\ P$. Hence

$$\frac{B\ P}{PB} = \frac{AP}{AB}.$$

But, from equation (1), $AP = 1/AP$, so P is related to P by the equation

$$B\ P = \frac{PB}{AB \bullet AP}. \qquad (2)$$

Now if P lies on the locus C, then $AP\ (= PA) = k \bullet PB$, so that

$$B\ P = \frac{1}{k \bullet AB}.$$

Thus P lies on a circle C of radius $1/(k \bullet AB)$ and with centre B.

Conversely, if P lies on C, then $B\ P = 1/(k \bullet AB)$, so from equation (2),

$$\frac{1}{k \bullet AB} = \frac{PB}{AB \bullet AP}.$$

Thus $AP = k \bullet PB$, and so P lies on the locus C.

It follows that the inversion t maps the locus C onto the circle C. But $t^{\ 1} = t$ maps generalized circles to generalized circles, so $C = t^{\ 1}(C)$ is a generalized circle.

This proof sets up a one one correspondence between the family of Apollonian circles and a family of concentric circles. Each Apollonian circle corresponds to a value of k, and this in turn corresponds to a circle of radius $1/(k \bullet AB)$ with centre $B = t(B)$.

In particular, the extended line in the Apollonian family corresponds to $k = 1$, which corresponds to the circle of radius $l/AB = AB$, with centre B. As you would expect, this circle passes through the centre A of the inversion t.

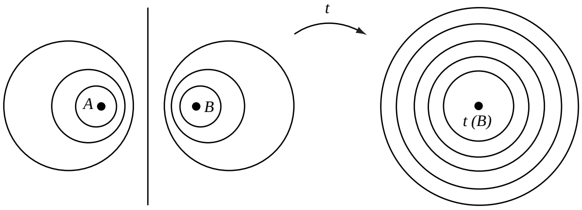

The importance of this correspondence is that it enables us to characterize Apollonian families of circles in terms of inversive transformations. Such a characterization is ideal for tackling problems in inversive geometry so we state it in the form of a theorem.

Theorem 2 Let A and B be distinct points in the plane, and let t be the inversion in the circle with centre A and radius 1. Then the Apollonian family of circles defined by the point circles A and B is mapped by t to the family of all concentric circles with centre $t(B)$; and the family of all concentric circles with centre $t(B)$ is mapped by t to the Apollonian family of circles defined by the point circles A and B.

For t is self-inverse.

A remarkable consequence of this theorem is that in inversive geometry all Apollonian and concentric families of circles are congruent to each other. Indeed the theorem shows that every Apollonian family of circles is congruent to a family of concentric circles, so it is sufficient to show that all families of concentric circles are congruent to each other. This is easily achieved by noting that any family of concentric circles can be mapped onto any other family of concentric circles by a translation that makes their centres coincide.

Recall that a translation is an inversive transformation provided we adopt the convention that maps to .

5.5.2 Families of Circles

Any two (distinct) generalized circles in the extended plane are related in precisely one of the following ways:

1. they may not intersect;
2. they may intersect in precisely *one* point;
3. they may intersect in *two* distinct points.

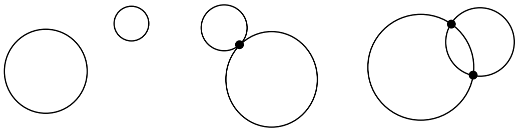

Case 1 Case 2 Case 3

In the previous subsection you met Apollonian families of circles, in which no two circles intersect; in the following subsection we will show that any two non-intersecting circles determine an Apollonian family of circles. We now consider the two other families, namely families of generalized circles that intersect at precisely one point and families of generalized circles that intersect at precisely two points.

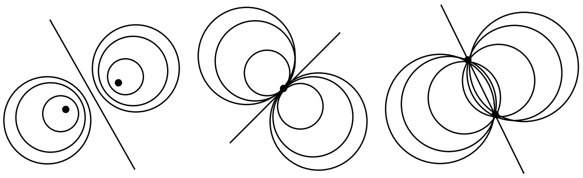

Apollonian circles Circles through one point Circles through two points

Each of these three families contains precisely one extended line known as the *radical axis* of the family. The centres of the circles in each family lie on a line, or *axis*, that is perpendicular to the radical axis. Since all the circles in each family are symmetrical about the axis through their centres, the families are known as *coaxal families*.

Definition A **coaxal family of circles** in the plane is a family of (generalized) circles of one of the following types:

1. an Apollonian family, with particular point circles;
2. a family that intersect at one particular point;
3. a family that intersect at two particular points.

The extended line in each family is called the **radical axis** of the family.

Given the Apollonian family of circles F defined by point circles A and B, there is a corresponding coaxal family of circles G that pass through A and B. Conversely, given a coaxal family of circles G that pass through distinct points A and B, there is a corresponding Apollonian family of circles F defined by the point circles A and B.

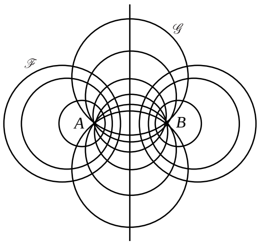

The following theorem describes a remarkable relationship between the two families of circles, F and G.

Theorem 3 Coaxal Circles Theorem
Let A and B be distinct points in the plane. Let F be the Apollonian family defined by the point circles A and B, and let G be the family of all generalized circles through A and B. Then every member of F is orthogonal to every member of G.

Two circles are *orthogonal* if they meet at right angles.

Proof Let t be the inversion in the unit circle with centre A. By Theorem 2, the Apollonian family F is mapped by t to the family of concentric circles with centre $t(B)$.

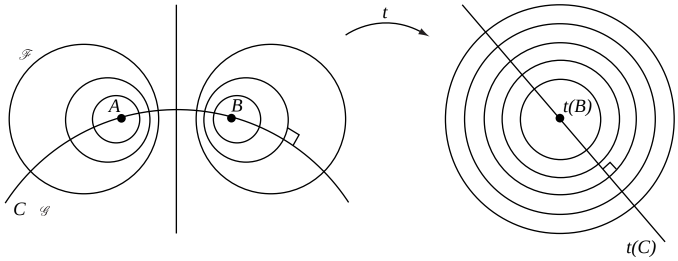

Now let C be an arbitrary member of G. Since C is a generalized circle that passes through A and B, it follows that $t(C)$ is a generalized circle that passes through $t(A) =$ and $t(B)$. In other words, $t(C)$ is an extended line through $t(B)$. Clearly, this line $t(C)$ intersects each of the circles with centre $t(B)$ at right angles. Since inversion preserves the magnitude of angles, it follows that C meets each of the Apollonian circles in F at right angles.

For an Apollonian family of circles defined by point circles A and B, the Coaxal Circles Theorem states that every generalized circle through A and B meets each of the Apollonian circles at right angles. The next theorem enables us to use this fact to deduce that A and B are inverse points with respect to each of the Apollonian circles.

Theorem 4 Two points A and B in the extended complex plane are inverse points with respect to a generalized circle C if and only if every generalized circle through A and B meets C at right angles.

Proof We first show that A and B are inverse points if every generalized circle through A and B meets C at right angles.

Let C_1 be any generalized circle through A and B that meets C at right angles, at points R and S. Now invert the figure C_1 in the generalized circle C. Since R and S remain fixed, C_1 is mapped to a generalized circle that meets C at right angles at R and S. It follows that C_1 is mapped to itself; for the fact that the image passes through R and S means that it must be a circle, and the radii of this circle through R and S must be along the tangents to C at R and S.

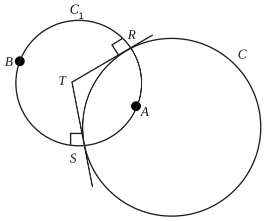

In fact, by a similar argument, every generalized circle through A and B that meets C at right angles maps to itself. The only way this can happen is if A inverts to B, and vice versa. Hence A and B are inverse points.

Next we show that if A and B are inverse points, then every generalized circle through A and B meets C at right angles.

Let A and B be inverse points, and let C_1 be a generalized circle through A and B that meets C at R. Under inversion in C, the points A and B swap over and R remains fixed, so C_1 maps to itself. It follows that both the angles that C_1 makes with the circle C must be equal. These angles must therefore be right angles, which is what we want to prove.

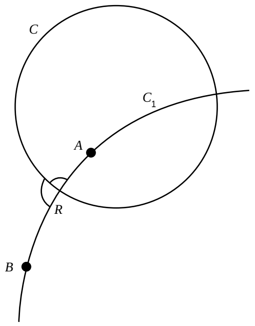

Theorem 4 has two important corollaries. The first follows directly from the Coaxal Circles Theorem.

Corollary 1 If C is an Apollonian circle defined by the point circles A and B, then A and B are inverse points with respect to C.

Proof Let C be an Apollonian circle with respect to the point circles A and B. Then by the Coaxal Circles Theorem, every generalized circle through A and B meets C at right angles. By Theorem 4, A and B are inverse points with respect to C.

The second corollary asserts that inverse points are preserved by inversive transformations.

Corollary 2 Let A and B be inverse points with respect to a generalized circle C, and let t be an inversive transformation. Then $t(A)$ and $t(B)$ are inverse points with respect to the generalized circle $t(C)$.

Proof

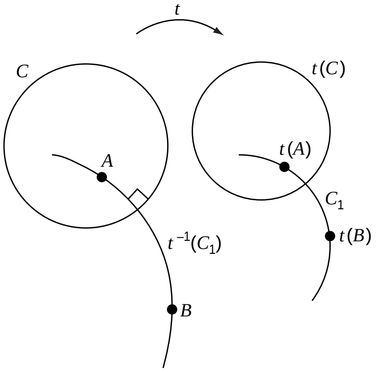

Let C_1 be an arbitrary generalized circle through $t(A)$ and $t(B)$. Then $t^{-1}(C_1)$ is a generalized circle that passes through A and B. By Theorem 4, $t^{-1}(C_1)$ meets C at right angles. But t preserves the magnitude of angles, so C_1 meets $t(C)$ at right angles. Since C_1 is an arbitrary generalized circle through $t(A)$ and $t(B)$, it follows from Theorem 4 that $t(A)$ and $t(B)$ are inverse points with respect to $t(C)$.

An immediate consequence of Corollary 2 is that if F is the family of *all* generalized circles that have two given points A and B as inverse points, then under an inversive transformation t the family F maps onto the family of *all* generalized circles that have $t(A)$ and $t(B)$ as inverse points.

At first sight this observation appears to be little more than a restatement of Corollary 2; however the restatement has a particular significance, as the following theorem shows.

Theorem 5 Let F be the family of all generalized circles that have A and B as inverse points. Then F is either a concentric family of circles with centres A or B, or the Apollonian family of circles with point circles A and B.

Proof First, suppose that either A or B is the point ; to be definite, assume $A =$. Then each circle in F has and B as inverse points. But this can happen only if B is the centre of each circle in F; in other words F is the family of concentric circles with centre B. A similar argument applies if $B =$.

Next suppose that neither A nor B is the point , and let t be the inversion in the circle of unit radius with centre A. By Theorem 2, t maps the family F to the family of all generalized circles with inverse points at $t(A) =$ and $t(B)$, namely the family of concentric circles with centre $t(B)$. By Theorem 5, F is the Apollonian family of circles defined by the point circles A and B.

Roughly speaking, you can think of the family of concentric circles with centre A as an Apollonian family of circles defined by point circles A and . Then an Apollonian family of circles defined by the point circles A and B maps to an Apollonian family of circles defined by the point circles $t(A)$ and $t(B)$.

In light of the remarks that precede this theorem, it follows that under an inversive transformation t an Apollonian family of circles defined by point circles A and B

EITHER maps to an Apollonian family of circles defined by the point circles $t(A)$ and $t(B)$,
OR maps to a concentric family of circles with centre $t(A)$ or $t(B)$.

Similarly a concentric family of circles with centre A either maps to the Apollonian family of circles defined by the point circles $t(A)$ and $t(\ )$, or it maps to a concentric family of circles with centre $t(A)$ or $t(\ )$.

So far, we have used inversive geometry to prove results about Apollonian families of circles, but similar methods can be used to prove results about other plane figures. The technique is to use an inversive transformation (often an inversion) to map one or more circles in the figure to extended straight lines. The additional symmetry that results from this transformation is then used to establish the required result.

For future reference, we also note here a fact that we met in the proof of Theorem 4 above.

Corollary 3 Let C and C_1 be circles that meet at right angles at two points R and S. Then the centre of C_1 is the point T of intersection of the tangents to C at R and S; and C_1 maps to itself under inversion in C.

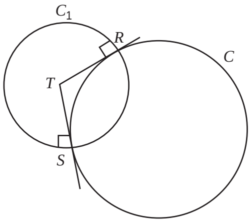

Example 2 Let C_1, C_2 and C_3 be circles in the plane such that C_2 and C_3 touch at a point P, C_1 and C_3 touch at a point Q, and C_1 and C_2 touch at a point R. Let C be the circle that passes through P, Q and R. Prove that C cuts C_1, C_2 and C_3 at right angles.

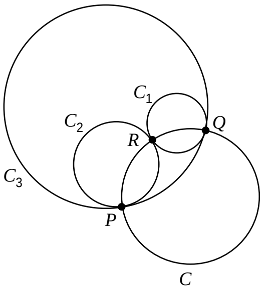

Solution Let t be an inversion in a circle with centre R. Then $t(C_1)$ and $t(C_2)$ are straight lines. Moreover, since C_1 and C_2 do not meet at any point other than P, it follows that $t(C_1)$ and $t(C_2)$ cannot meet in $\mathbb{C}$ and must therefore be parallel.

Now C_3 is tangential to C_1 and C_2, and does not pass through R, so $t(C_3)$ must be a circle that touches $t(C_1)$ and $t(C_2)$ tangentially at $t(Q)$ and $t(P)$, respectively. Next observe that C passes through R, so $t(C)$ is the extended line through $t(Q)$ and $t(P)$. Since the transformed figure has reflectional symmetry about $t(C)$, it follows that $t(C)$ cuts $t(C_1)$, $t(C_2)$ and $t(C_3)$ at right angles at $t(P)$ and $t(Q)$. But t preserves angles, so C cuts C_1, C_2 and C_3 at right angles at P and Q.

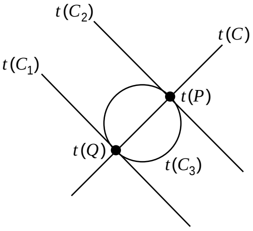

A similar argument with t replaced by inversion in a circle with centre P (or Q) shows that C cuts C_1 and C_2 at right angles at R. □

5.5.3 Two Circles Determine a Coaxal Family

It is clear that if we are given two circles in a coaxal family of mutually tangential circles, this determines the whole family.

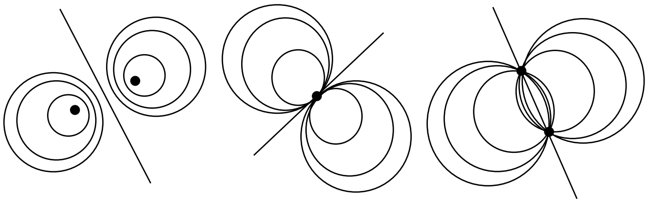

Similarly if we are given two circles in a coaxal family of circles through two fixed points, this determines the whole family. This raises the question as to whether two circles in an Apollonian family are sufficient to determine the whole family. We now prove that this is indeed the case. The crucial tool that we need is the following result.

Theorem 6 The Concentricity Theorem
Let C_1 and C_2 be any two non-intersecting circles in the plane. Then there is a inversive transformation that maps C_1 and C_2 onto a pair of concentric circles.

Proof If the circles are concentric then there is nothing to prove, so we may assume they are not concentric.

Step 1 Pick a point O on the circle C_1, and invert both circles in the circle of unit radius with centre O. Under this inversion C_1 maps to a line, C_1, and C_2 maps to a circle, C_2.

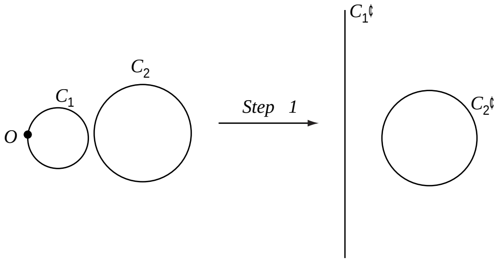

Step 2 Pick a point T on C_2 that does not lie on the perpendicular from the centre of C_2 to the line C_1, and let the tangent to C_2 at T intersect C_1 at the point U, say. Draw the circle C_3 with centre U and radius UT. This circle is perpendicular to C_2 at T, because a tangent to a circle is perpendicular to the radius at the point of contact. The circle C_3 is perpendicular to the line C_1 because its centre lies on the line C_1.

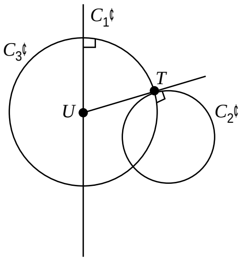

Step 2

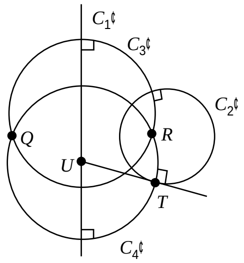

Step 3

Step 3 Repeat the construct in Step 2 starting with another position for the point T, to obtain a circle C_4 perpendicular to both C_1 and C_2 at the respective points T. Let C_3 and C_4 meet at Q and R, say.

Step 4 Invert the figure again, this time in the unit circle with centre Q. Then the line C_1 maps to a circle C_1, and the circle C_2 maps to a circle C_2; the circles C_3 and C_4 pass through Q, so they invert to straight lines at right angles to C_1 and C_2. These lines are therefore diameters of the circles C_1 and C_2, and so the point where they meet (the image R of R) must be the centre of both circles.

As in the (second) proof of Theorem 1, the choice of 1 for the radius is not significant.

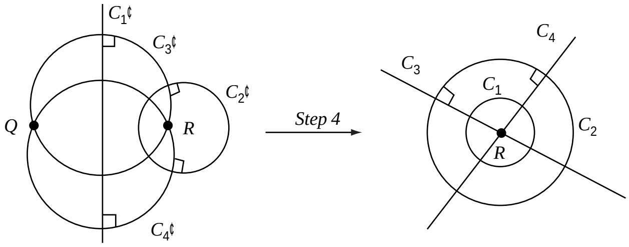

Step 5 Composing the inversion in Step 1 with the inversion in Step 4, we obtain an inversive transformation that maps C_1 and C_2 onto the pair of concentric circles C_1 and C_2.

We can now prove the following beautiful result.

Theorem 7 Two Apollonian Circles Theorem

Let C_1 and C_2 be two non-concentric circles that do not intersect. Then there is a unique Apollonian family of circles that contains C_1 and C_2.

Proof By the Concentricity Theorem there is an inversive transformation t that maps C_1 and C_2 onto a pair of concentric circles $t(C_1)$ and $t(C_2)$. Let O be the common centre of these circles, and let G be the family of concentric circles with centre O. Under the inversive transformation t^{1} the family G maps to the family of all generalized circles that have $t^{1}(O)$ and $t^{1}(\ )$ as inverse points; we denote this family by F. Now F cannot be a concentric family of circles, since it contains C_1 and C_2; so, by Theorem 5, it must be an Apollonian family of circles.

Theorem 6 above

To complete the proof we must show that $\mathcal{F}$ is the only Apollonian family of circles that contains C_1 and C_2. To do this suppose that $\mathcal{F}'$ is any Apollonian family of circles that contains C_1 and C_2. Then $\mathcal{F}'$ is mapped by t either to an Apollonian family of circles, or to a concentric family of circles. In fact the image of $\mathcal{F}'$ under t contains the concentric circles $t(C_1)$ and $t(C_2)$ and must therefore be the family $\mathcal{G}$ of concentric circles with centre O. It follows that $\mathcal{F}$ and $\mathcal{F}'$ are the same family of circles, for they are both mapped to $\mathcal{G}$ by t.

5.5.4 Some Applications of Inversion

As we mentioned earlier, inversion can often be used to prove results about plane figures, using a suitable inversion to map one or more circles in the figures to extended lines — the additional symmetry then being used to establish the desired result.

Our first application of inversion to prove a beautiful result is known as Steiner's Porism. A *porism* is a mathematical construction problem that has a surprising answer: either the construction cannot be carried out or it has infinitely many solutions.

Jakob Steiner (1796–1863) was a 19th-century Swiss geometer.

Theorem 8 Steiner's Porism
Let C_1 and C_2 be non-intersecting circles, with C_1 inside C_2. Then:

EITHER it is impossible to fit a chain of circles between C_1 and C_2, with each circle touching C_1 and C_2 and two other circles in the chain;
OR it is possible to construct such a chain, and the first circle in the chain can be placed in any convenient position.

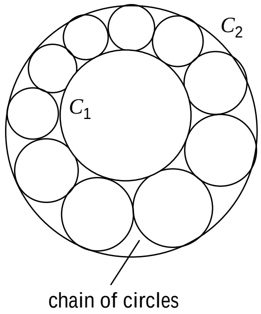

Proof There are two possibilities.

EITHER It is impossible to fit a chain between C_1 and C_2 with the properties described, in which case the porism is established.
OR There is at least one chain F that fits between C_1 and C_2. To establish the porism in this case we must show that a chain also exists for each choice of a first circle C between C_1 and C_2.

By the Concentricity Theorem there is an inversive transformation t that maps C_1 and C_2 onto concentric circles C_1' and C_2'. Under t the chain F maps onto a chain of circles F' between C_1' and C_2', and C maps onto a circle C' which touches C_1' and C_2'. Since C_1' and C_2' have a common centre O, we can rotate the chain F' about O until its first circle is superimposed on C'. If we denote the rotated chain by G', then the inversive transformation t^{-1} maps G' back to a chain G between C_1 and C_2. Moreover, the first circle of G is C as required.

Recall that t^{-1} must be an inversive transformation, since t is.

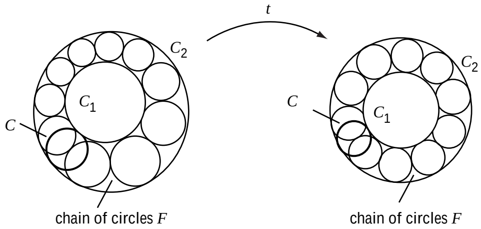

Another famous porism is **Poncelet's Porism**: If C_1 and C_2 are any two plane conics for which (for any given $n > 2$) it is possible to find one n-sided polygon which is simultaneously inscribed in C_1 and circumscribed around C_2, then it is possible to find infinitely many of them.

We omit a proof, as it is beyond the scope of this book.

Next, we prove a classical result known as Ptolemy's Theorem — but using inversion rather than using Ptolemy's Pythagorean geometry methods! A key tool in our proof is the following useful result.

Claudius Ptolemaus (c. 85–165 AD) was a geometer, astronomer and geographer in Alexandria, Egypt.

Lemma 1 Let a and b be points in $\mathbb{C} - \{O\}$. Then, under inversion in a circle C with centre O and radius r the distance between the images of a and b is $\frac{r^2}{|a|\bullet|b|}|a - b|$.

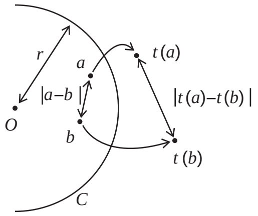

Proof Inversion in C may be represented in the complex plane by the transformation

$$t(z) = \frac{r^2}{\bar{z}}, \quad \text{where } z \in \mathbb{C} - \{O\}.$$

This is a special case of a general formula for inversion in $\mathbb{C}$ that you met in Theorem 3, Subsection 5.2.1.

It follows from this formula that

$$\begin{aligned} t(a) - t(b) &= \frac{r^2}{\bar{a}} - \frac{r^2}{\bar{b}} \\ &= \frac{r^2}{\bar{a}\bullet\bar{b}}(\bar{b} - \bar{a}), \end{aligned}$$

from which we may deduce that

$$|t(a) - t(b)| = \frac{r^2}{|a|\bullet|b|}|b - a|,$$

which is equivalent to the desired result.

The formula for distances in Lemma 1 plays a crucial role in our proof of Ptolemy's Theorem that the sum of the product of the lengths of opposite sides of a cyclic quadrilateral equals the product of the lengths of the diagonals.

Theorem 9 Ptolemy's Theorem
Let A, B, C, D be the vertices (in order round a circle) of a quadrilateral $ABCD$ inscribed in a circle. Then

$$AD \bullet BC + AB \bullet CD = AC \bullet BD.$$

Proof Let the transformation t be inversion of the figure $ABCD$ in the circle C with centre A and radius 1. This inversion maps A to , the circle C to an extended line (which we will denote by), and the points B, C, D on C to points B , C , D on .

Since B, C, D are in order on C, it follows that B , C , D are in order on . Consequently,

$$B C + C D = B D. \tag{3}$$

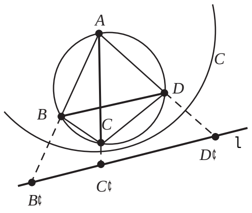

Since t is self-inverse, the images of B , C , D under inversion in C are B, C, D. Hence, if we apply the result of Lemma 1 in turn to the distances $B C$, $C D$, $B D$ under the inversion t, it follows from equation (3) that

$$\frac{BC}{AB \bullet AC} + \frac{CD}{AC \bullet AD} = \frac{BD}{AB \bullet AD}.$$

Multiplying both sides of this equation by the product $AB \bullet AC \bullet AD$, we obtain the desired result

$$AD \bullet BC + AB \bullet CD = AC \bullet BD.$$

Our final application of inversion is particularly attractive visually.

Theorem 10 Shoemaker's Knife
Let D be a region in the upper half-plane, whose boundary consists of three semicircles S_1, S_2 and S_3 that meet at three points A, B and C on the x-axis, as shown. Then it is possible to fit a chain of circles into D between S_1 and S_2, with the first circle touching S_i, S_2 and S_3, and with each successive circle touching S_1, S_2 and the previous and following circles in the chain.

The name *Shoemaker s Knife* (or o , *Arbelos*) is due to Archimedes.

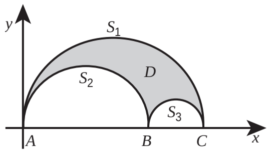

Proof Let t denote inversion in the circle C with centre A and radius 1.

Under t, the positive x-axis maps to itself, and points in the upper half-plane (the set of points above the x-axis) map to points in the upper half-plane.

The circle of which S_3 is a portion is mapped by t to another circle, which crosses the x-axis at right angles, at points B and C say. The image of S_3 is therefore another semicircle in the upper half-plane, with endpoints B and C .

Under t, the origin A maps to , so the semicircles S_1 and S_2 map to portions of (generalized) circles in the upper half-plane that join to B and C , respectively. These portions do not intersect in the (ordinary) plane C, and

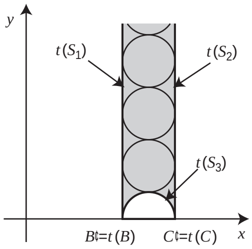

they meet the x-axis at right angles. Hence the images of S_1 and S_2 are two vertical half-lines with endpoints B' and C'.

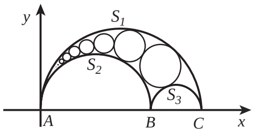

Now, it is obviously possible to fit a chain of circles into the vertical strip bounded by the semicircle $t(S_3)$ and the vertical lines $t(S_1)$ and $t(S_2)$, with the first circle touching $t(S_3)$ and with each successive circle touching $t(S_1)$, $t(S_2)$ and the previous and following circles in the chain.

We can transform this new configuration back into the original configuration by applying the inverse transformation t^{-1}. Since t^{-1} preserves circles and their tangencies, the required result now follows.

Problem 3 Prove that the height above the x-axis of the centre of the n-th circle, C_n say, in the chain of circles constructed in Theorem 10 is n times the diameter of that circle.
Hint: Invert the figure in a circle with centre A that intersects C_n at right angles.

Problem 4 In the figure for Theorem 10, let A, B and C have coordinates (0,0), $(b, 0)$ and $(c, 0)$, respectively. Show that the centres of the circles in the chain constructed in Theorem 10 all lie on an ellipse with foci $F\left(\frac{1}{2}b, 0\right)$ and $F'\left(\frac{1}{2}c, 0\right)$, and major axis $\left[(0, 0), \left(\frac{1}{2}(b + c), 0\right)\right]$.
Hint: Use the Sum of Focal Distances of the Ellipse theorem.

Theorem 5, Subsection 1.1.4

5.6 Exercises

Section 5.1

1. Determine the image under inversion in the unit circle C of each of the following points.
   (a) $(3, -4)$ (b) $(-1, 1)$ (c) $(9, 0)$ (d) $\left(\frac{1}{2}, -\frac{\sqrt{3}}{2}\right)$
2. Determine the image under inversion in the unit circle C of the following circles C (with the origin removed if it belongs to C):
   (a) the circle with centre $(3, -4)$ and radius 5;
   (b) the circle with centre (1, 2) and radius 3.
3. Determine the image under inversion in the unit circle C of the following lines ℓ (with the origin removed if it belongs to ℓ):
   (a) the line $y + 3x = 5$;
   (b) the line $y + 2x = 0$.

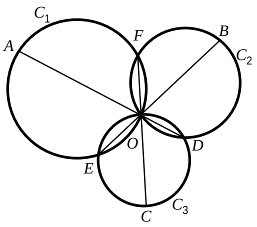

4. Three circles C_1, C_2, C_3 pass through the origin O, and meet at three other distinct points D, E and F, as shown. The following sets of points are collinear: A, O and D; B, O and E; C, O and F. OA and OB are diameters of C_1 and C_2, respectively. Prove that OC is a diameter of C_3.
   Hint: Invert the figure in a unit circle with centre O, and then use the fact that the altitudes of a triangle are concurrent.

5. By means of a specific example, show that the centre of the image circle $t(C_2)$ of a circle C_2 under inversion t in another circle C_1 may not be the image of the centre of C_2 under t.
6. Let the origin O, P and Q be distinct points; and let P' and Q', respectively, be the images of P and Q under inversion in the circle with centre O and radius r. Prove that

$$P'Q' = \frac{r^2}{OP \cdot OQ} PQ.$$

(For simplicity, you may assume that O, Q', P' are points inside the circle of inversion and occur in this order along the radius outwards from O. The result holds in all other cases too, by a similar argument.)

7. Let P and Q be distinct points in $\mathbb{C} - \{O\}$ with the origin O, P and Q not being collinear; and let P' and Q', respectively, be their images under inversion in the circle with centre O and radius r. Let ON and ON' be the perpendiculars from O to PQ and $P'Q'$ (extended, if necessary), respectively.
   (a) Prove that

$$P'Q' = \frac{r^2}{OP \cdot OQ} PQ.$$

   (b) Prove that

$$\frac{ON}{PQ} = \frac{ON'}{P'Q'}.$$

8. Let AB be a diameter of a circle C, and let chords AD and BE (extended, if necessary) of C intersect at a point F. Prove that the circle DEF is orthogonal to the circle C.

Section 5.2

1. Let t be the transformation defined by

$$t(z) = \tfrac{1}{2}(1 + \sqrt{3}i)z + 2i \quad (z \in \mathbb{C}).$$

   (a) Show that t represents an isometry.
   (b) Interpret t as the composite of a rotation and a translation.
   (c) Interpret t as a composite of reflections.
2. Determine the image under inversion in the unit circle $\mathcal{C}$ of each of the following points.
   (a) $-3 + 4i$ (b) $5 - 12i$
3. Let C be the circle of radius 2 with centre $1 + i$.
   (a) Write down the inversion in C of the extended complex plane, as a transformation of $\mathbb{C}$.
   (b) Determine the image of i under the inversion in C.

4. Let t be the extended linear function defined by

$$t(z) = \begin{array}{ll} 5iz + (2+6i), & \text{if } z \quad \mathbb{C}, \\ \quad, & \text{if } z = \quad . \end{array}$$

Express t as a composite of inversions of the extended complex plane.

5. Prove that two points z and z in the complex plane correspond to two diametrically-opposite points on the Riemann sphere if and only if $z \bullet \overline{(z\ )} = \quad 1$.

Section 5.3

1. Which of the following formulas define a Möbius transformation? For each formula that does define a Möbius transformation, state the image of under M.
(a) $M(z) = \quad 2\frac{z+1}{z\quad 2}$ (b) $M(z) = \quad 2z + \frac{1}{z\quad 2}$ (c) $M(z) = \frac{2iz\quad 2}{z+i}$

2. Decide which, if any, of the following matrices

$$\mathbf{A}_1 = \begin{matrix} 1 & i \\ i & 1 \end{matrix}, \quad \mathbf{A}_2 = \begin{matrix} 1 & i \\ i & 1 \end{matrix}, \quad \mathbf{A}_3 = \begin{matrix} i & i \\ 1 & i \end{matrix},$$

are associated with each of the following Möbius transformations.
(a) $M_1(z) = \frac{z\quad 1}{iz+1}$ (b) $M_2(z) = \frac{iz\quad 1}{z\quad i}$ (c) $M_3(z) = \frac{z\quad 1}{z+i}$

3. Determine a formula for each of the following Möbius transformations, where M_1 and M_2 are the Möbius transformations defined in Exercise 2.
(a) $M_1 \quad M_1(z)$ (b) $M_1 \quad M_2(z)$ (c) $M_2 \quad M_1^{\quad 1}(z)$

4. State which of the following transformations of $\mathbb{C}$ onto itself are inversive transformations, giving a brief reason in each case.
(a) $t(z) = \frac{4i\quad 2}{3i\quad 2}$ (b) $t(z) = 5$ (c) $t(z) = \frac{iz+2}{z\quad 2i}$

5. Let t be the inversive transformation defined by

$$t(z) = \frac{z+i}{z\quad i}.$$

Determine the image of each of the following generalized circles under t:
(a) the extended line $\quad \{\quad\}$, where $\quad$ is the line with equation $y = x$;
(b) the unit circle $\mathbb{C}$.

Section 5.4

1. For each set of three points below, determine the Möbius transformation that maps the three points to 0, 1, $\quad$, respectively.
(a) i, $\quad i$, $\quad$ (b) $\quad$, 1, i (c) $2i$, $\quad$, 3 (d) 1, 2, 3

2. Using the results of Exercise 1, determine the Möbius transformations that map:
(a) 1, 2, 3 to $2i$, $\quad$, 3; (b) $\quad$, 1, i to i, $\quad i$, $\quad$;
(c) $2i$, $\quad$, 3 to $\quad$, 1, i; (d) i, $\quad i$, $\quad$ to 1, 2, 3.

3. Determine whether each of the following sets of points are collinear:
(a) $\frac{3}{2} + i$, $2i$, $\quad 6 + 6i$ (b) $1 + 2i$, $4 \quad 5i$, $10 \quad 20i$

4. Determine whether each of the following sets of points lies on a circle:
   (a) i, 3, $1 + 2i$, $2 + 4i$ (b) $1 \quad 2i$, $4 + i$, $4 + i$, $1 + 6i$

Section 5.5

1. An Apollonian family of circles F includes (6,0) as a point circle and also the circle C with equation $x^2 + y^2 = 4$. Determine the other point circle in the family.
2. The circles C_1 and C_2 in an Apollonian family of circles have the segments [0,8] and $\#1, \frac{7}{2}\$$, respectively, of the x-axis as diameters.
   (a) Determine the point circles of the family.
   (b) Hence determine the equation of the Apollonian circle in the family that passes through the point (1,1).
3. Determine the images under inversion in a unit circle with centre A of the following families F of coaxal circles.
   (a) F is the family of all generalized circles tangential to the y-axis at the origin, and $A = (0, 0)$.
   (b) F is the family of all generalized circles through the points $A = (\quad 1, 0)$ and $B = (1, 0)$.
   (c) F is the Apollonian family of all generalized circles with point circles $A = (\quad 1, 0)$ and $B = (1, 0)$.
4. Let C_1 be the (extended) y-axis in the plane, and C_2 the circle with centre (2, 0) and radius 1. Find a sequence of inversions whose composite inversive transformation maps C_1 and C_2 onto concentric circles.
5. Let C_1 and C_2 be circles that touch at a point P, apart from which C_2 lies inside C_1. Prove that it is possible to fit a chain of circles between C_1 and C_2, with each circle touching C_1 and C_2 and two other circles in the chain, and with the first circle in the chain placed in any convenient position.
6. Prove that, if the Möbius transformation M maps C, the unit circle with centre the origin, to itself, then M is necessarily of the form

$$M(z) = K\frac{z \quad}{\bar{\ }z \quad 1}, \quad \text{where } |K| = 1 \quad \text{and} \quad |\ \ | = 1.$$

7. Let F be an Apollonian family of circles with point circles A and B and radical axis . Prove that, for any given point P on the radical axis, all the tangents from P to circles in F are of equal length.
8. Let D be the region in the upper half-plane whose boundary consists of three semicircles S_1, S_2 and S_3 with diameter the segments [1, 0], [0, 1] and [1, 1], respectively, of the x-axis, as shown.
   (a) Prove that it is possible to fit a chain of circles into D between S_1 and S_2, with the first circle touching S_1, S_2 and S_3, and with each successive circle touching S_1, S_2 and the previous and following circles in the chain.

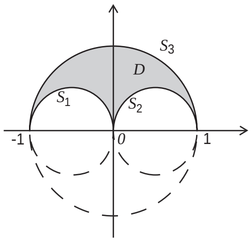

(b) Prove that the n^{th} circle in the chain has its centre at a height $\frac{2n}{4n^2-1}$ above the x-axis, and has radius $\frac{1}{4n^2-1}$.

9. Let C_1, C_2, C_3 be three circles, external to each other, where the pairs C_1 and C_2, C_2 and C_3, and C_3 and C_1 each touch at a point. Prove that there are exactly two circles that touch all of C_1, C_2, C_3.

   These circles are sometimes called *kissing circles* or *Soddy circles*, and were discovered by René Descartes (1643), then rediscovered by Jakob Steiner (1826) and Frederick Soddy (1936). Soddy published his discovery as a poem titled *The Kiss Precise*, in the scientific journal *Nature*; it starts:

   For pairs of lips to kiss maybe
   Involves no trigonometry.
   'Tis not so when four circles kiss
   Each one the other three.
   To bring this off the four must be
   As three in one or one in three.
   If one in three, beyond a doubt
   Each gets three kisses from without.
   If three in one, then is that one
   Thrice kissed internally.

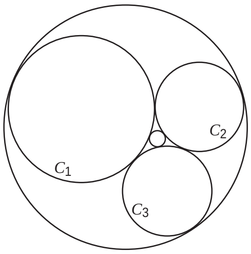

Frederick Soddy (1877–1956) won the 1921 Nobel Prize in chemistry for research into radioactive decay and the theory of isotopes.

Summary of Chapter 5

Section 5.1: Inversion

1. Under **reflection in a line** ℓ in the plane, a point A is mapped to an image point A' that lies an equal distance from ℓ, but on the opposite side of ℓ.
2. Let C be a circle in the plane, with centre O and radius r, and let A be any point other than O. If A' is the point on the line OA that lies on the same side of O as A and satisfies the equation $OA \cdot OA' = r^2$, then we call A' the **inverse** of A with respect to (or *in*) the circle C. The point O is called the **centre of inversion**, and C the **circle of inversion**. The transformation t defined by $t(A) = A'$ $(A \in \mathbb{R}^2 - \{O\})$ is known as **inversion in** C.

   There is no point to which O is mapped by the inversion, and no point that is mapped to O by the inversion.

   Inversion in a circle is a generalization of reflection in a line.

   We often use the term **inversion** to mean EITHER reflection in a line OR inversion in a circle.
3. If A is a point outside a circle C with centre O, AB and AC are the two tangents from A to C, and A' is the point of intersection of OA and BC, then A and A' are inverse points with respect to C.
4. Inversion in a circle C maps points outside C to points inside C, and vice-versa; points on C map to themselves.

Inversion in a circle is a one-one transformation of the plane minus the centre of inversion onto itself.

Inversion is a self-inverse transformation.

5. Inversion is not a Euclidean transformation; for example, it does not preserve lengths.

   Inversion is not an affine transformation; for example, it does not preserve straight lines.

6. Inversion in the unit circle $\mathcal{C}$ is the function $t : (x, y) \mapsto \left(\frac{x}{x^2+y^2}, \frac{y}{x^2+y^2}\right)$, where $(x, y) \in \mathbb{R}^2 - \{O\}$.

   Strategy To determine an equation for the image of a curve under inversion in the unit circle $\mathcal{C}$:

   1. write down an equation that relates the x- and y-coordinates of the points on the curve;
   2. replace x by $x/(x^2 + y^2)$ and y by $y/(x^2 + y^2)$, and simplify the resulting equation.

   If the curve passes through the origin we must first remove the origin from the curve. If we remove a point A from a curve, the curve is said to be **punctured** at A.

7. **Images of lines under inversion** Under inversion in a circle with centre O:

   (a) a line that does not pass through O maps onto a circle punctured at O;
   (b) a line punctured at O maps onto itself.

   Images of circles under inversion Under inversion in a circle with centre O:

   (a) a circle that does not pass through O maps onto a circle;
   (b) a circle punctured at O maps onto a line that does not pass through O.

   Even if inversion maps one circle onto another, it may not map centre to centre.

8. Let c_1 and c_2 be two curves that intersect at the point A, and let the tangents to the curves at A be ℓ_1 and ℓ_2, respectively. Then the **anti-clockwise angle** from c_1 to c_2 at A is the anti-clockwise angle from ℓ_1 to ℓ_2, and the **clockwise angle** from c_1 to c_2 at A is the clockwise angle from ℓ_1 to ℓ_2.

9. **Symmetry Lemma** Let ℓ be a line that does not pass through the point O. Then under inversion in a circle with centre O, ℓ maps to a circle C (punctured at O), and the tangent to C at O is parallel to ℓ.

10. **Angle Theorem** An inversion in any circle preserves the magnitude of angles between curves but reverses their orientation.

Section 5.2: Extending the Plane

1. The transformation $t(z) = z + c$ $(z \in \mathbb{C})$, where $c = a + ib$, is a **translation** through the vector (a, b).

   The transformation $t(z) = \overline{z}$ $(z \in \mathbb{C})$ is a **reflection** in the x-axis.

   The transformation $t(z) = az$ $(z \in \mathbb{C})$, where $|a| = 1$, $a = \cos\theta_0 + i\sin\theta_0$, is a **rotation** through an anti-clockwise angle θ_0 about the origin.

Each isometry t **of the plane** can be represented in the complex plane by one of the functions $t(z) = az + b$ or $t(z) = a\overline{z} + b$, where $a, b \in \mathbb{C}$, $|a| = 1$. Conversely, all such functions represent isometries.

Every isometry can be expressed as a composite of reflections.

2. The transformation $t(z) = kz$ $(z \in \mathbb{C})$, where k is real and positive, is a **scaling** by a **factor** k.
3. An inversion in a circle C of radius r with centre (a, b) may be represented in the complex plane by the transformation $t(z) = \frac{r^2}{\overline{z} - \overline{c}} + c$ $(z \in \mathbb{C} - \{c\})$, where $c = a + ib$.

   In particular, inversion in the unit circle $\mathcal{C}$ may be represented by the transformation $t(z) = \frac{1}{\overline{z}}$ $(z \in \mathbb{C} - \{O\})$.
4. A **linear function** is a function of the form $t(z) = az + b$ $(z \in \mathbb{C})$, where $a, b \in \mathbb{C}$ and $a \neq 0$.

   It may be decomposed into a composite $t_2 \circ t_1$, where t_1 is the scaling $t_1(z) = |a|z$ and t_2 is the isometry $t_2(z) = (a/|a|)z + b$. This can be described geometrically as a scaling by the factor $|a|$, followed by a rotation through the angle $\operatorname{Arg}(a/|a|)$, followed by a translation through the vector $(\operatorname{Re} b, \operatorname{Im} b)$.

   Linear functions preserve angles, and map circles and lines to circles and lines.
5. The **reciprocal function** is the function of the form $t(z) = \frac{1}{z}$ $(z \in \mathbb{C} - \{O\})$.

   It may be decomposed into a composite $t_2 \circ t_1$, where t_1 is the inversion $t_1(z) = 1/\overline{z}$ and t_2 is the conjugation $t_2(z) = \overline{z}$.

   The reciprocal function preserves angles.
6. The **extended (complex) plane** $\hat{\mathbb{C}}$ is the union of the complex plane $\mathbb{C}$ and one extra point, the **point at infinity**, denoted by the symbol ∞.
7. An **extended line** is any line ℓ in the plane together with the point ∞. An extended line can be thought of as a circle of infinite radius.

   A **generalized circle** in the extended plane is a set that is either a circle or an extended line.
8. Let C be a generalized circle in the extended complex plane. Then an **inversion of the extended plane with respect to** C is a function t defined by one of the following rules.
   (a) If C is a circle of radius r with centre O, then

$$t(A) = \begin{cases} \text{the inverse of } A \text{ with respect to } C, & \\ & \text{if } A \in \mathbb{C} - \{O\}, \\ \infty, & \text{if } A = O, \\ O, & \text{if } A = \infty. \end{cases}$$

   (b) If C is an extended line $\ell \cup \{\infty\}$, then

$$t(A) = \begin{cases} \text{the reflection of } A \text{ in } \ell, & \text{if } A \in \mathbb{C}, \\ \infty, & \text{if } A = \infty. \end{cases}$$

Inversions of the extended plane map generalized circles onto generalized circles.

9. The **extended conjugation function**, **extended reciprocal function** and **extended linear functions** may be defined in the natural way as mappings of C to itself.
10. The extended reciprocal function and the extended linear functions can be decomposed into a composite of inversions, and map generalized circles onto generalized circles.
11. **Strategy** To scale C by a factor k:
    1. invert in the circle $\{z : |z| = 1\}$, then
    2. invert in the circle $\{z : |z| = \ \overline{k}\}$.

    Strategy To rotate C through an angle :
    1. invert in the line $\{z : \operatorname{Arg} z = 0\}\ \{\ \}$, then
    2. invert in the line $z : \operatorname{Arg} z = \frac{1}{2}\quad \{\ \}$.

    Strategy To translate C through a vector (a, b):
    1. invert in the line $\{(x, y) : ax + by = 0\}\quad \{\ \}$, then
    2. invert in the line $\{(x, y) : ax + by = \frac{1}{2}(a^2 + b^2)\}\quad \{\ \}$.
12. The **Riemann sphere** S is the sphere in $\mathbb{R}^3$ with centre the origin and radius 1. The points $N(0, 0, 1)$ and $S = (0, 0,\ 1)$ are called the **North Pole** and the **South Pole** of S, respectively.

    Lines through the North Pole intersecting S at points P and the (x, y)-plane at points P give a one one correspondence between points on S and points in the plane. (We associate the North Pole with the point in the extended complex plane.)

    The one-one onto function : S C which maps points on S to the associated points in the extended complex plane is called **stereographic projection**. If (X, Y, Z) S and $z = x + iy$ C, then

$$(X, Y, Z) = \frac{X}{1\quad Z} + i\,\frac{Y}{1\quad Z}$$

    and

$$^{1}(x + iy) = \quad \frac{2x}{x^2 + y^2 + 1}, \frac{2y}{x^2 + y^2 + 1}, \frac{x^2 + y^2\quad 1}{x^2 + y^2 + 1}\quad .$$

13. Under stereographic projection, circles on the Riemann sphere map onto generalized circles in C. In particular, circles on the sphere that pass through N map onto extended lines in C, and circles on the sphere that do not pass through N map onto ordinary circles in C.

    Stereographic projection preserves the magnitude of angles.

Section 5.3: Inversive Geometry

1. A transformation t : C C is an **inversive transformation** if it can be expressed as a composite of inversions.

    Inversive geometry is the study of those properties of figures in C that are preserved by inversive transformations.

The extended reciprocal function and the extended linear functions are inversive transformations.

2. Inversive transformations preserve the magnitude of angles, map generalized circles to generalized circles, and preserve tangency.

   The set of inversive transformations forms a group under the operation of composition of functions, called the **inversive group**.

3. Every Euclidean transformation is also an inversive transformation, and every Euclidean property is also an inversive property.

   The doubling map $t(z) = 2z$ $(z \in \mathbb{C})$ is an affine transformation and an inversive transformation, but is not a Euclidean transformation.

   The transformation of $\mathbb{R}^2$ to itself given by $t(\mathbf{x}) = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \mathbf{x}$ represents a horizontal shear; it is an affine transformation, but is not an inversive transformation.

4. A **Möbius transformation** is a function $t : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$ of the form $M(z) = \frac{az+b}{cz+d}$, where $a, b, c, d \in \mathbb{C}$, and $ad - bc \neq 0$.

   If $c = 0$, we adopt the convention that $M(\infty) = \infty$; otherwise we adopt the convention that $M(-d/c) = \infty$ and $M(\infty) = a/c$.

   The extended linear functions and the extended reciprocal function are Möbius transformations.

   Every Möbius transformation is an inversive transformation.

5. Möbius transformations preserve the magnitude and orientation of angles, and map generalized circles onto generalized circles.

6. Let M be a Möbius transformation defined by $M(z) = \frac{az+b}{cz+d}$, where a, b, c, $d \in \mathbb{C}$, and $ad - bc \neq 0$; then $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is a matrix **associated** with M.

   A matrix associated with a Möbius transformation is invertible.

   If $\mathbf{A}$ is a matrix associated with a Möbius transformation M, then so is $c\mathbf{A}$ for any non-zero $c \in \mathbb{C}$. Every matrix associated with M has the form $c\mathbf{A}$ for some $c \in \mathbb{C} - \{O\}$.

7. Let M_1 and M_2 be Möbius transformations with associated matrices $\mathbf{A}_1$ and $\mathbf{A}_2$, respectively. Then $M_1 \circ M_2$ is a Möbius transformation with associated matrix $\mathbf{A}_1\mathbf{A}_2$.

   Strategy To compose two Möbius transformations M_1 and M_2:

   1. write down matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ associated with M_1 and M_2;
   2. calculate $\mathbf{A}_1\mathbf{A}_2$;
   3. write down the Möbius transformation $M_1 \circ M_2$ with which $\mathbf{A}_1\mathbf{A}_2$ is associated.

8. The inverse of the Möbius transformation $M(z) = \frac{az+b}{cz+d}$ is also a Möbius transformation, and it may be written in the form $M^{-1}(z) = \frac{dz-b}{-cz+a}$.

9. The set of all Möbius transformations forms a group under composition of functions.

10. Every inversion t has the form $t(z) = M(\bar{z})$, where M is a Möbius transformation.

Every inversive transformation t can be represented in C by one of the formulas $t(z) = \frac{az+b}{cz+d}$ or $t(z) = \frac{a\bar{z}+b}{c\bar{z}+d}$, where $a, b, c, d \quad$ C, and $ad \quad bc = 0$.

The composite of an even number of inversions is a Möbius transformation and is called a **direct** inversive transformation (since it preserves the orientation of angles); the composite of an odd number of inversions is called an **indirect** inversive transformation (since it reverses the orientation of angles).

11. **Strategy** To determine the image of a generalized circle C under an inversive transformation t:
    1. write down three points z_1, z_2, z_3 on C;
    2. determine the images $t(z_1), t(z_2), t(z_3)$;
    3. the image $t(C)$ is the (unique) generalized circle through $t(z_1)$, $t(z_2)$, $t(z_3)$.

Section 5.4: Fundamental Theorem of Inversive Geometry

1. **Fundamental Theorem of Inversive Geometry**
   Let z_1, z_2, z_3 and w_1, w_2, w_3 be two sets of three points in the extended complex plane C. Then there is a unique Möbius transformation M which maps z_1 to w_1, z_2 to w_2, and z_3 to w_3.
2. **Strategy** To determine the Möbius transformation M which maps three given points z_1, z_2, z_3 onto the points 0, 1 and , respectively:
   1. choose the appropriate form of mapping from the following formulas for M:

mapping	form of M
$z_1, z_2, z_3 \qquad 0, 1,$	$K\frac{z \quad z_1}{z \quad z_3}$
$, z_2, z_3 \qquad 0, 1,$	$\frac{K}{z \quad z_3}$
$z_1, \quad , z_3 \qquad 0, 1,$	$\frac{z \quad z_1}{z \quad z_3}$
$z_1, z_2, \qquad 0, 1,$	$K(z \quad z_1)$

   2. find the complex number K for which $M(z_2) = 1$.
3. **Strategy** To determine the Möbius transformation M which maps the points z_1, z_2, z_3 to the points w_1, w_2, w_3, respectively:
   1. find the Möbius transformation M_1 which maps the points z_1, z_2, z_3 to the points 0, 1, , respectively;
   2. find the Möbius transformation M_2 which maps the points w_1, w_2, w_3 to the points 0, 1, , respectively;
   3. calculate $M = M_2^{\ 1} \quad M_1$.
4. **Strategy** To determine whether z_1, z_2, z_3 and z_4 lie on a circle:
   1. find the Möbius transformation M which maps z_1, z_2, z_3 to 0, 1, , respectively;
   2. the points z_1, z_2, z_3, z_4 lie on a generalized circle if and only if $M(z_4)$ is real;

   3. the generalized circle in Step 2 is a circle provided that $M(\infty)$ is not real.

   If $M(z_4)$ and $M(\infty)$ are both real, then z_1, z_2, z_3, z_4 lie on a line.

5. Let C_1 and C_2 be generalized circles in the extended complex plane. Then there is a Möbius transformation that maps C_1 onto C_2.

   All generalized circles are inversive-congruent.

Section 5.5: Coaxal Families of Circles

1. **Apollonian Circles Theorem** Let A and B be two distinct points in the plane, and let k be a positive real number other than 1. Then the locus of points P that satisfy $PA : PB = k : 1$ is a circle whose centre lies on the line through A and B.

   The centre of the circle is not at either A or B. If $A = (-a, 0)$ and $B = (a, 0)$, then the circle has centre c and radius r where $c = \left(-a\frac{1+k^2}{1-k^2}, 0\right)$ and $r = \frac{2ak}{|1-k^2|}$.

   If $k = 1$, then P lies on the perpendicular bisector of AB.

   Every positive value of k gives rise to a generalized circle, known as a **circle of Apollonius**; the family of all such circles is known as the **Apollonian family of circles** defined by the points A and B.

   If $0 < k < 1$, the Apollonian circle surrounds A; if $k > 1$, the Apollonian circle surrounds B. The points A and B are called **point circles**, corresponding to the cases $k = 0$ and $k = \infty$, respectively. Every point in the plane lies on precisely one circle (or generalized circle) in each Apollonian family.
2. Let A and B be distinct points in the plane, and let t be the inversion in the circle with centre A and radius 1. Then the Apollonian family of circles defined by the point circles A and B is mapped by t to the family of all concentric circles with centre $t(B)$; and the family of all concentric circles with centre $t(B)$ is mapped by t to the Apollonian family of circles defined by the point circles A and B.
3. A **coaxal family of circles** in the plane is a family of (generalized) circles of one of the following types:
   1. an Apollonian family, with particular point circles;
   2. a family that intersect at one particular point;
   3. a family that intersect at two particular points.

   The extended line in each family is called the **radical axis** of the family.

   Coaxal Circles Theorem Let A and B be distinct points in the plane. Let $\mathcal{F}$ be the Apollonian family defined by the point circles A and B, and let $\mathcal{G}$ be the family of all generalized circles through A and B. Then every member of $\mathcal{F}$ is orthogonal to every member of $\mathcal{G}$.
4. Two points A and B in the extended complex plane are inverse points with respect to a generalized circle C if and only if every generalized circle through A and B meets C at right angles.

If C is an Apollonian circle defined by the point circles A and B, then A and B are inverse points with respect to C.

Let A and B be inverse points with respect to a generalized circle C, and let t be an inversive transformation. Then $t(A)$ and $t(B)$ are inverse points with respect to the generalized circle $t(C)$.

Let $\mathcal{F}$ be the family of all generalized circles that have A and B as inverse points. Then $\mathcal{F}$ is either a concentric family of circles with centre A or B, or an Apollonian family of circles with point circles A and B.

Let C and C_1 be circles that meet at right angles at two points R and S. Then the centre of C_1 is the point T of intersection of the tangents to C at R and S; and C_1 maps to itself under inversion in C.

5. **Concentricity Theorem** Let C_1 and C_2 be any two non-intersecting circles in the plane. Then there is an inversive transformation that maps C_1 and C_2 onto a pair of concentric circles.
6. **Two Apollonian Circles Theorem** Let C_1 and C_2 be two non-concentric circles that do not intersect. Then there is a unique Apollonian family of circles that contains C_1 and C_2.
7. A **porism** is a mathematical construction problem that has a surprising answer: either the construction cannot be carried out or it has infinitely many solutions.

   Steiner's Porism Let C_1 and C_2 be non-intersecting circles, with C_1 inside C_2. Then EITHER it is impossible to fit a chain of circles between C_1 and C_2, with each circle touching C_1 and C_2 and two other circles in the chain OR it is possible to construct such a chain, and the first circle in the chain can be placed in any convenient position.

   Poncelet's Porism If C_1 and C_2 are any two plane conics for which (for any given $n > 2$) it is possible to find one n-sided polygon which is simultaneously inscribed in C_1 and circumscribed around C_2, then it is possible to find infinitely many of them.
7. Let a and b be points in $\mathbb{C} - \{O\}$. Then, under inversion in a circle C with centre O and radius r the distance between the images of a and b is $\frac{r^2}{|a|\cdot|b|}|a - b|$.

   Ptolemy's Theorem Let A,B,C,D be the vertices (in order) of a quadrilateral inscribed in a circle. Then $AD \cdot BC + AB \cdot CD = AC \cdot BD$.
8. **Shoemaker's Knife/Arbelos** Let D be a region in the upper half-plane, whose boundary consists of three semicircles S_1, S_2 and S_3 that meet at three points A, B and C on the x-axis. Then it is possible to fit a chain of circles into D between S_1 and S_2, with the first circle touching S_1, S_2 and S_3, and with each successive circle touching S_1, S_2 and the previous and following circles in the chain.

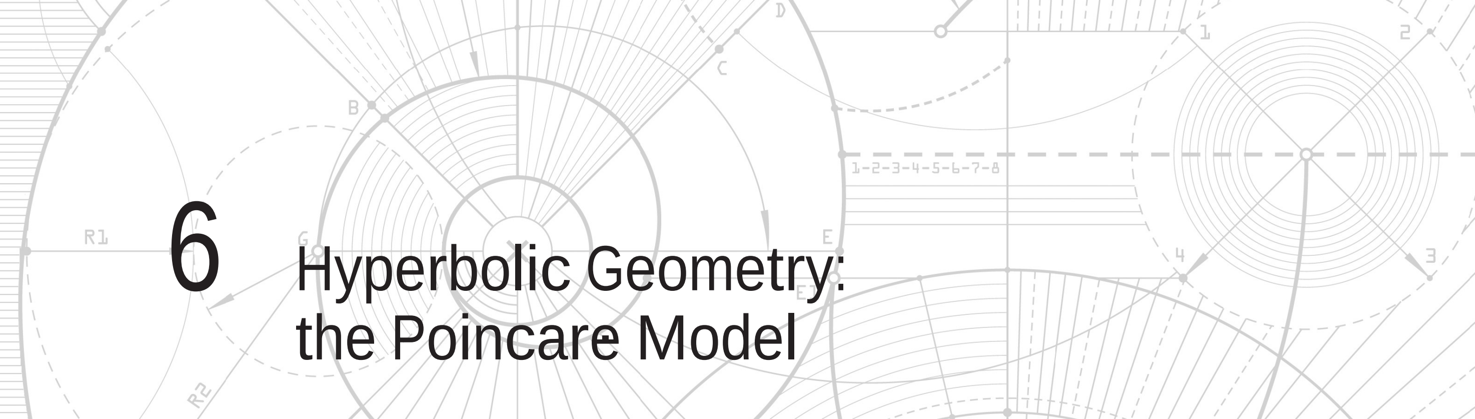

6

Hyperbolic Geometry: the Poincare Model

The book, Euclid's *Elements*, which underlay most geometrical teaching in Western Europe for over 2000 years, gave definitions of the basic terms in geometry and rules (called postulates) for their use. Many of Euclid's assumptions seem entirely uncontroversial, such as the assertions that through any two distinct points in a plane or in space there passes a unique line, and that a line can be extended indefinitely in both directions. On these foundations Euclid gave rigorous proofs of theorems in elementary geometry, which could be accepted as true because of the way they had been established.

Among the postulates for Euclidean geometry is one about parallel lines which is equivalent to the following statement.

The (Euclidean) Parallel Postulate Given any line ℓ and a point P not on ℓ, there is a unique line m in the same plane as P and ℓ which passes through P and does not meet ℓ.

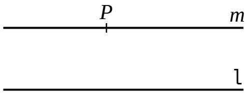

Note that the parallel postulate makes two assertions: first that the parallel line exists, and second that it is unique.

Historically it was felt that the Parallel Postulate as it stands is not obvious. It implies that every line through P other than m eventually meets ℓ, but plainly this meeting place can be a long way away. How much confidence would you place in an assumption about lines meeting somewhere in the Virgo star cluster, some 100 million light years away?

Many Greek, Arab and later Western geometers had felt that the answer was 'not much', and had tried to delete the Euclidean Parallel Postulate from the list and instead derive it as a theorem. They all failed, and in the end the reason was laid bare: the Parallel Postulate cannot be made into a theorem in this way because there are internally consistent models of geometries which obey all the Euclidean postulates except the Parallel Postulate.

But how do we obtain geometries other than Euclidean geometry with its Parallel Postulate? One way to do this might be to insist that any two lines intersect. The geometry to which this gives rise is called *Elliptic Geometry*. One model of elliptic geometry is *Spherical Geometry* (for example, on the surface of the Earth), and we shall investigate this model in Chapter 7.

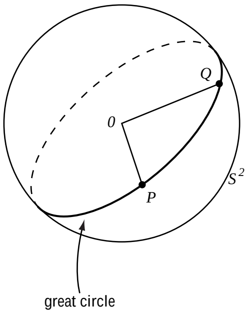

A line PQ on the surface of a sphere (which we call S^2), such as on the Earth s surface, can be defined to be the *great circle* (that is, the intersection of the plane through the centre O of the sphere with the surface of the sphere) through the two points P and Q. Any two distinct great circles meet in two (diametrically opposite) points, such as P and P shown in the margin, so any two lines in spherical geometry meet in two points.

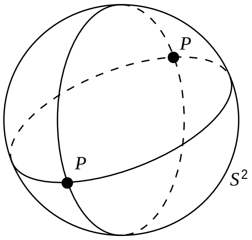

Generalizing this fact, in *Elliptic Geometry* we keep all the other postulates that collectively describe Euclidean geometry, remove the Parallel Postulate, and replace it with the following analogue.

The (Elliptic) Parallel Postulate Given any line and a point P not on , all lines through P meet . (That is, there are no lines through P that are parallel to .)

Another way to obtain a non-Euclidean geometry is to insist that the line through P that does not meet is not unique. We could again keep all the other postulates that collectively describe Euclidean geometry, remove the Parallel Postulate, and replace it with the following.

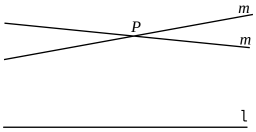

The (Hyperbolic) Parallel Postulate Given any line and a point P not on , there are at least two lines m through P that do not meet . (That is, there are at least two lines through P that are parallel to .)

The geometry whose postulates are those of Euclidean geometry but with this variant of the Euclidean parallel postulate is called *Hyperbolic Geometry*. The term *non-Euclidean Geometry* is often used to describe elliptic and hyperbolic geometries together.

In fact there are still other geometries (and other models of both hyperbolic and elliptic geometries), that lie beyond the scope of this book.

Non-Euclidean geometry was discovered independently in the late 1820s by the distinguished Russian mathematician Nicolai Ivanovich Lobachevskii (1792 1856), working at the University of Kazan, and the Hungarian JÆnos Bolyai (1802 1860), an Army officer whose father, Farkas (or Wolfgang) (1775 1856), was also a mathematician. Many of the same ideas were also known to Carl Friedrich Gauss (1777 1855), but he did not publish them. However, non-Euclidean geometry was not accepted until after these three were dead, when the German Bernhard Riemann (1826 1866) and the Italian Eugenio Beltrami (1835 1900) published their ideas about geometry in the late 1860s. Riemann gave the appropriate general setting, Beltrami a specific account of non-Euclidean geometry. Only then did the new geometry become rapidly accepted. Amongst other things, it spurred Felix Klein (1849 1925) to propose his view of geometry.

Non-Euclidean geometry was one of the most momentous mathematical discoveries of the l9th century. It had several revolutionary implications, because it provided a physically plausible description of space that differed markedly from Euclid s. It became possible to imagine that the universe is not Euclidean, and in the 19th century some mathematicians and astronomers, such as Gauss and his friend Bessel), entertained the idea quite seriously.

Friedrich Wilhelm Bessel (1784 1846) was a German mathematician and astronomer, after whom Bessel functions are named.

The theorems in non-Euclidean geometries often differ markedly from their Euclidean equivalents. For example, in Euclidean geometry the angles of any triangle sum to ; in elliptic geometry the angle sum of any triangle turns out to be strictly greater than , whereas in hyperbolic geometry the angle sum of any triangle turns out to be strictly less than .

In this chapter we shall concentrate on a model of hyperbolic geometry due to the French mathematician Henri Poincaré. In this model (which we will generally call hyperbolic geometry for simplicity), the space of points is the interior of the unit disc $\mathbb{D} = \{z : |z| < 1\}$. All figures in this geometry will be drawn as they appear in this disc. In Section 6.1 we define *hyperbolic lines* and *hyperbolic angles*. We then introduce *hyperbolic reflections* in these lines, and obtain the *group of hyperbolic transformations*: a hyperbolic transformation is a composite of hyperbolic reflections. In Section 6.2 we see that every hyperbolic transformation can be described as either a Möbius transformation (of a certain form) or a Möbius transformation composed with complex conjugation.

You met Möbius transformations in Subsection 5.3.3.

The properties of hyperbolic geometry include angle and length, and so we can define *hyperbolic circles*, which we do in Section 6.3. We also obtain a formula relating Euclidean lengths and *hyperbolic lengths* (at least in special cases), and in Section 6.4 we give useful formulas for the study of hyperbolic triangles. In Section 6.5 we look at *area* in hyperbolic geometry, and at hyperbolic tilings, or *tessellations*, such as the one displayed on the cover of this book. Finally, in Section 6.6 we briefly discuss the half-plane model of hyperbolic geometry.

6.1 Hyperbolic Geometry: the Disc Model

6.1.1 What is Hyperbolic Geometry?

The **points** or ***d*-points** (where d stands for disc) in (Poincaré s version of) hyperbolic geometry consist of the points in the unit disc

$$\mathbb{D} = \{z : |z| < 1\} = \left\{(x, y) : x^2 + y^2 < 1\right\}.$$

Note that we shall use the alternative notations z and (x, y) freely to describe points in $\mathbb{D}$, according to convenience.

In sketches we usually illustrate $\mathbb{D}$ by drawing in its boundary

$$\mathbb{C} = \{z : |z| = 1\} = \left\{(x, y) : x^2 + y^2 = 1\right\},$$

but you should note that points on $\mathbb{C}$ are *not* points which belong to the geometry.

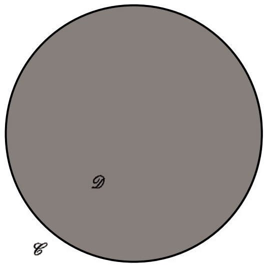

A crucial feature of this geometry is the concept of a *d-line*, or a *line in hyperbolic geometry.*

Definition A ***d*-line** is that part of a (Euclidean) generalized circle which meets C at right angles and which lies in D.

A generalized circle is either a circle or a (Euclidean) line; see Subsection 5.2.3.

Every d-line is part of a generalized (Euclidean) circle which meets the boundary circle in two points. We call these two points the **boundary points** of the d-line. Notice that the boundary points of a d-line are not d-points for they are not points in D.

The sketch below illustrates various d-lines. Some of these d-lines are arcs of (Euclidean) circles; some are (Euclidean) line segments, in fact, diameters of D.

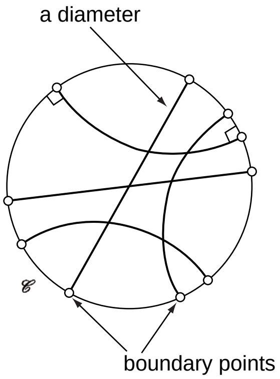

You may find the following figure a helpful reminder for the shapes of d-lines:

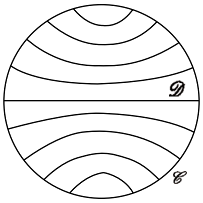

In fact, any d-line that is part of a Euclidean line must be a diameter of D, since otherwise it cannot meet C at right angles.

Conversely, any d-line that is an arc of a Euclidean circle cannot pass through the origin. For if meets C at a point P, then the tangent to at P is a radius of C and so passes through the origin O, which is the centre of C. But then is a line touching the Euclidean circle at P, and because circles lie entirely on one side of their tangents, cannot meet again. Thus cannot pass through O.

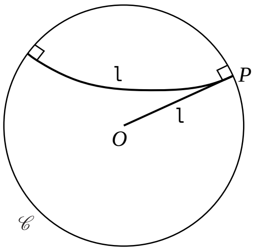

Indeed, if is a d-line that is an arc of a Euclidean circle, then there is a diameter of D that does not meet it. For, if meets C at the points P and Q, then the tangents to at P and Q are radii of C and so pass through the origin. They divide the disc D into four regions, one of which contains . Clearly, a diameter can be drawn that lies entirely in two of the other regions in fact, you can see that there are infinitely many such diameters. Each of these divides D into two regions, one which contains and one which does not.

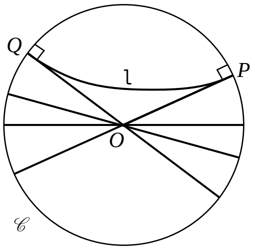

Notice also that if a d-line is part of a Euclidean circle meeting C at points P and Q, then its centre is the point R where the tangents to C at P and Q meet. So the point R lies outside the circle C.

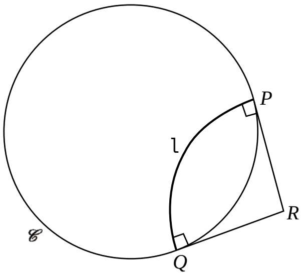

Sometimes it is clear from a (reasonably accurate) sketch whether or not part of a generalized circle is a d-line. For example, consider

$$\ell = \left\{(x, y) : x^2 + y^2 - 4y = 0\right\} \cap \mathbb{D}.$$

Since

$$x^2 + y^2 - 4y = 0 \Leftrightarrow x^2 + (y - 2)^2 = 4,$$

we can sketch ℓ, and deduce that it is not a d-line since it clearly does not meet $\mathcal{C}$ right angles. (The tangents to ℓ at its boundary points on $\mathcal{C}$ do not pass through the origin.)

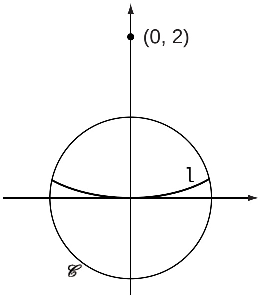

Problem 1 Sketch the following parts of generalized circles, and determine which of them are d-lines.

$$\ell_1 = \{(x, y) : y = 3x\} \cap \mathbb{D}$$
$$\ell_2 = \{(x, y) : 3x + y = 1\} \cap \mathbb{D}$$
$$\ell_3 = \left\{(x, y) : x^2 + y^2 + 2x + 2y + 1 = 0\right\} \cap \mathbb{D}$$

Now, it is rather easy to be mislead by a sketch into believing that something is true in general whereas it is not! So, we now give an analytic criterion for identifying d-lines.

Lemma 1 The equation of a d-line ℓ is of one of the following forms:

$ax + by = 0$, where a and b are not both zero;

$x^2 + y^2 + fx + gy + 1 = 0$, where $f^2 + g^2 > 4$.

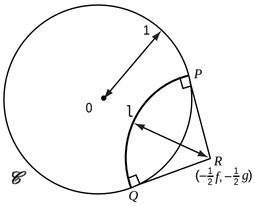

Proof If a d-line ℓ is (part of) a Euclidean line through the origin, then certainly its equation is of the form $ax + by = 0$, where a and b are not both zero.

The other possibility is that the d-line ℓ is (part of) a Euclidean circle C that intersects the boundary $\mathcal{C}$ of $\mathbb{D}$ at right angles. So, let C have equation

$$x^2 + y^2 + fx + gy + h = 0, \text{ for some real numbers } f, g \text{ and } h.$$

First, it follows from the Orthogonality Test that any circle with this equation intersects the unit circle $\mathcal{C}$ with equation

$$x^2 + y^2 - 1 = 0$$

Subsection 1.1.2, Theorem 3

if and only if

$$f \cdot 0 + g \cdot 0 = 2 \cdot (h + (-1));$$

that is, if and only if $h = 1$. So the equation of C must be of the form

$$x^2 + y^2 + fx + gy + 1 = 0, \text{ for some } f \text{ and } g.$$

Next, the centre $\left(-\frac{1}{2}f, -\frac{1}{2}g\right)$ of C must be outside $\mathbb{C}$, so that $\left(-\frac{1}{2}f\right)^2 + \left(-\frac{1}{2}g\right)^2 > 1$, which we can equivalently write in the form $f^2 + g^2 > 4$.

Finally, we need the circle C to meet $\mathbb{D}$. This happens if and only if

$$1 + \text{radius of } C > \text{distance of the centre of } C \text{ from } O;$$

since $h = 1$ we can write this requirement in the form

$$1 + \sqrt{\frac{1}{4}f^2 + \frac{1}{4}g^2 - 1} > \sqrt{\frac{1}{4}f^2 + \frac{1}{4}g^2}. \qquad (1)$$

Since we already know that $\sqrt{\frac{1}{4}f^2 + \frac{1}{4}g^2} > 1$, the inequality (1) follows at once if we can prove that

$$1 + \sqrt{t - 1} > \sqrt{t}, \quad \text{whenever } t > 1. \qquad (2)$$

Here we are writing t in place of $\frac{1}{4}f^2 + \frac{1}{4}g^2$.

So, let $F(t) = 1 + \sqrt{t-1} - \sqrt{t}$. Then

$$\begin{aligned} F(1) &= 1 + \sqrt{0} - \sqrt{1} \\ &= 0, \end{aligned}$$

and

$$\begin{aligned} F'(t) &= \frac{1}{2\sqrt{t-1}} - \frac{1}{2\sqrt{t}} \\ &= \frac{\sqrt{t} - \sqrt{t-1}}{2\sqrt{t-1}\sqrt{t}} > 0. \end{aligned}$$

It follows at once that $F(t) > F(1) = 0$ for $t > 1$, and so that the inequality (2) holds, as required.

Thus, for example, let C be the circle in $\mathbb{R}^2$ with equation

$$x^2 + y^2 + 3x - 2y + 1 = 0,$$

and let be the part of C that lies in $\mathbb{D}$. Then is a d-line since its equation is of the form required by Lemma 1, with $f = 3$ and $g = -2$, and $f^2 + g^2 = 3^2 + (-2)^2 = 13 > 4$.

In fact, C is the circle with centre $\left(-\frac{3}{2}, 1\right)$ and radius $\sqrt{\left(-\frac{3}{2}\right)^2 + (-1)^2 - 1} = \frac{3}{2}$.

Note the following useful result that we obtain from putting $a = -\frac{1}{2}f$, $b = -\frac{1}{2}g$ into Lemma 1.

Corollary 1 Let $= a + ib$ be a point outside the unit circle $\mathbb{C}$. Then the d-line that is part of a Euclidean circle with centre has equation

$$x^2 + y^2 - 2ax - 2by + 1 = 0.$$

We can now turn to the question of parallelism, the notion that gave rise to the discovery of non-Euclidean geometries in the first place.

Definition Two d-lines that do not meet in $\mathbb{D}$ are:
parallel if the generalized Euclidean circles of which they are parts meet at a point on $\mathbb{C}$;
ultra-parallel if the generalized Euclidean circles of which they are parts do not meet on $\mathbb{C}$.

Recall that the boundary $\mathbb{C}$ is *not* part of the unit disc $\mathbb{D}$.

Remarks

1. It follows that given any d-line ℓ and any point P in $\mathbb{D}$ which is not on ℓ, there exist exactly two d-lines through P which are parallel to ℓ, as the next figure shows. A rigorous proof that such d-lines exist will be found below (Problem 5).

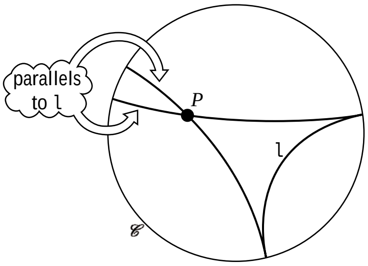

2. Similarly, corresponding to any given d-line ℓ and any point P in $\mathbb{D}$ which is not on ℓ, there exist infinitely many d-lines through P which are ultra-parallel to ℓ.

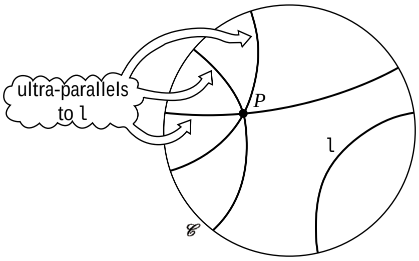

Problem 2 The following sets represent d-lines:

$$\ell_1 = \{(x,y) : y = x\} \cap \mathbb{D};$$
$$\ell_2 = \left\{(x,y) : x^2 + y^2 - 4x + 1 = 0\right\} \cap \mathbb{D};$$
$$\ell_3 = \left\{(x,y) : x^2 + y^2 - 2\sqrt{2}x + 1 = 0\right\} \cap \mathbb{D};$$
$$\ell_4 = \left\{(x,y) : x^2 + y^2 + 2x + 2y + 1 = 0\right\} \cap \mathbb{D}.$$

Sketch them, and decide from your figure which of the d-lines intersect each other, which are parallel, and which are ultra-parallel.

Problem 3 Sketch three d-lines ℓ_1, ℓ_2 and ℓ_3 with the property that ℓ_1 is parallel to ℓ_2, and ℓ_2 is parallel to ℓ_3, but ℓ_1 is not parallel to ℓ_3.

It follows that the relation is parallel to is not an equivalence relation in hyperbolic geometry.

In Euclidean geometry and in inversive geometry, reflection and inversion played important roles; this is also the case in hyperbolic geometry. Obviously, reflection of the unit disc $\mathcal{D}$ in a diameter maps $\mathcal{D}$ onto itself; in fact, so does inversion of $\mathcal{D}$ in a d-line that does not pass through the origin.

Theorem 1 Let ℓ be a d-line that is part of a Euclidean circle C. Then inversion in C maps $\mathcal{C}$ onto $\mathcal{C}$, and $\mathcal{D}$ onto $\mathcal{D}$.

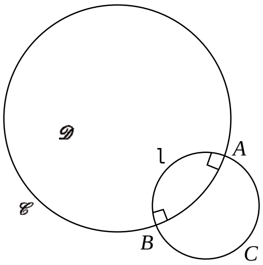

Proof Let C meet $\mathcal{C}$ at the points A and B.

Under inversion in C, A and B map to themselves. Since inversion preserves angles, the circle $\mathcal{C}$ maps onto some circle that meets C at right angles at the points A and B. There is only one such circle, $\mathcal{C}$ itself; it follows that inversion in C maps $\mathcal{C}$ onto itself.

Subsection 5.1.3, Theorem 5

Hence the inside $\mathcal{D}$ of $\mathcal{C}$ must map either to the inside or to the outside of the image of $\mathcal{C}$, namely $\mathcal{C}$ itself. But the points of ℓ map to themselves under this inversion, so the image of $\mathcal{D}$ must be $\mathcal{D}$ itself.

Inversion in C maps the shorter arc AB of $\mathcal{C}$ onto the longer arc AB of $\mathcal{C}$, and vice versa.

Inversion in C maps the region bounded by ℓ and the shorter arc AB of $\mathcal{C}$ onto the region bounded by ℓ and the longer arc AB of $\mathcal{C}$, and vice versa.

We now know that inversion in a d-line maps the disc $\mathcal{D}$ onto itself. We also know that the composition of an inversion with itself is the identity map. Because the analogous properties are true of Euclidean reflections, we make the following definition.

Definition A **hyperbolic reflection** in a d-line ℓ is the restriction to $\mathcal{D}$ of the inversion in the generalized circle of which ℓ is part.

Recall that we use the term inversion to mean *either* reflection in a line *or* inversion in a circle (see Subsection 5.1.1, Convention).

Remarks

1. If the d-line ℓ is part of a Euclidean circle C meeting the boundary circle $\mathcal{C}$ at points P and Q, then the point R where the tangents to $\mathcal{C}$ at P and Q meet is the centre of the inversion.
2. Notice that Theorem 1 may be reformulated as follows: hyperbolic reflection in any d-line maps the unit disc $\mathcal{D}$ onto itself.

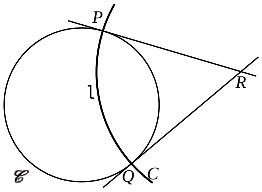

Hyperbolic reflections are the building blocks of transformations in hyperbolic geometry. Composites of a finite number of hyperbolic reflections are called **hyperbolic transformations**. The set of all such transformations under the operation of composition of functions forms a group, called the **hyperbolic group**, $G_{\mathcal{D}}$.

A proof of the fact that these form a group will be found as an Exercise in Section 6.7; it consists simply of checking the axioms for a group.

Definition **Hyperbolic geometry** consists of the unit disc, $\mathcal{D}$, together with the group $G_{\mathcal{D}}$ of hyperbolic transformations.

Remarks

1. Notice, in particular, that the identity mapping of $\mathbb{D}$ to itself belongs to $G_{\mathbb{D}}$, since it may be expressed as the finite composition $r \circ r^{-1}$ for any hyperbolic reflection r. Also, if $t_1 \circ t_2 \circ \ldots \circ t_n$ is any hyperbolic transformation, then its inverse is another hyperbolic transformation $t_n^{-1} \circ \ldots \circ t_2^{-1} \circ t_1^{-1}$.
2. Let the generalized circle C of which a d-line ℓ is part meet the boundary circle $\mathbb{C}$ in points P and Q, and let r be the hyperbolic reflection in ℓ. Then although r has domain the unit disc $\mathbb{D}$, and $P, Q \notin \mathbb{D}$, we sometimes find it convenient to call P and Q the boundary points of ℓ and to write the images of P and Q under inversion in C (of which r is a restriction) as $r(P)$ and $r(Q)$. The same convention is extended to the case where r is a hyperbolic transformation.

Next, we define the idea of angle in hyperbolic geometry.

Definition The **(hyperbolic) angle between two curves** (for example, two d-lines) through a given point A in $\mathbb{D}$ is the Euclidean angle between their (Euclidean) tangents at A.

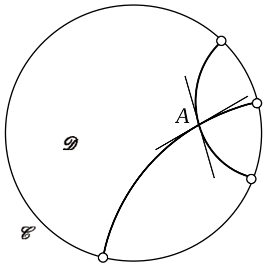

Now, Euclidean reflections and inversions both preserve the magnitudes of angles. It follows that hyperbolic transformations also preserve magnitudes of angles. Also, Euclidean reflections and inversions map generalized circles onto generalized circles. Combining this with the angle-preservation property, we deduce that hyperbolic reflections and inversions map d-lines onto d-lines and so compositions of a finite number of such transformations also have this property. We summarize these facts in the following result.

Theorem 2 Hyperbolic transformations map d-lines onto d-lines, and preserve the magnitudes of angles.

In the remainder of Chapter 6 we study the properties of various figures under hyperbolic transformations, and obtain surprising (and sometimes beautiful) results!

6.1.2 Existence of *d*-lines

Through a typical point A of $\mathbb{D}$ there is at least one d-line: namely, the diameter through the origin O and A. But through the origin there are infinitely many d-lines: the diameters of $\mathbb{D}$. Is there more than one d-line through an arbitrary point A of $\mathbb{D}$?

The first step towards answering this question is the following useful result, which shows that there is a hyperbolic transformation which maps A to O.

Lemma 2 Origin Lemma
Let A be a point of $\mathbb{D}$ other than the origin O. Then there exists a d-line ℓ such that hyperbolic reflection in ℓ maps A to O.

Proof We seek a d-line which is part of a Euclidean circle with centre R, say, such that inversion in this circle maps A to O. Suppose that this circle meets $\mathcal{C}$ at the point P.

The condition that this inversion maps A to O is

$$RO \bullet RA = RP^2, \tag{3}$$

since RP is a radius of the circle we seek.

The condition that part of this circle is a d-line is that triangle RPO is right-angled at P. By Pythagoras Theorem, this implies that

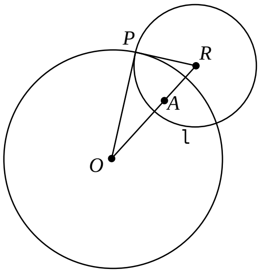

$$RP^2 + PO^2 = RO^2,$$

which, since $OP = 1$, is equivalent to

$$RP^2 = RO^2 - 1. \tag{4}$$

Eliminating the radius RP^2 from equations (3) and (4), we deduce that

$$RO \bullet RA = RO^2 - 1.$$

This is equivalent to

$$RO^2 - RO \bullet RA = 1$$

or

$$RO \bullet (RO - RA) = 1.$$

But $RO - RA = AO$, so we deduce that

$$RO \bullet AO = 1$$

which is equivalent to

$$OA \bullet OR = 1.$$

This tells us that the circle we seek has for its centre the point R which is found by inverting the given point A in the boundary circle $\mathcal{C}$ an unexpectedly memorable result!

The great value of the Origin Lemma is that it enables us to study any problem in hyperbolic geometry by mapping a suitably chosen point to the origin, thereby yielding a (frequently) simpler picture than before yet without losing any generality. We shall use this method often in this chapter.

We can, for example, use this approach to answer the question of how many d-lines pass through a given point of $\mathcal{D}$.

Theorem 3 Let A be a point of $\mathcal{D}$. Then there exist infinitely many d-lines through A.

Proof Let A be the origin. As we said above, each of the infinitely many diameters of $\mathcal{D}$ passes through the origin, and each of these diameters is a d-line.

If A is not the origin, then by the Origin Lemma, there is a hyperbolic transformation, r say, that maps A to the origin O, through which pass infinitely many d-lines all the diameters of $\mathcal{D}$.

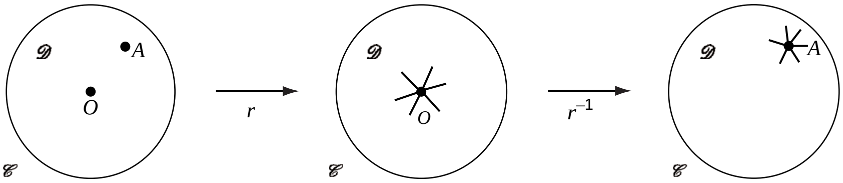

Since r ^{1} is also a hyperbolic transformation, it follows that the images of these diameters are also d-lines and they pass through A.

For, hyperbolic transformations map d-lines to d-lines, by Theorem 2.

An important result in Euclidean geometry is that there is exactly one line through any two given points. There is an analogous result in hyperbolic geometry.

Theorem 4 Let A and B be any two distinct points of D. Then there exists a unique d-line through A and B.

Proof (Existence) By the Origin Lemma, there is a hyperbolic transformation r that maps A to the origin O; let the image of B under r be the point B of D.

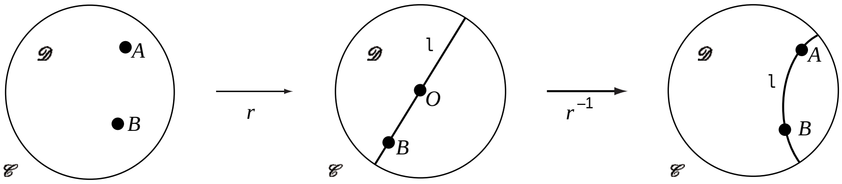

Then there is a unique d-line (a diameter of D) that passes through O and B . Since r ^{1} is also a hyperbolic transformation, it follows that $= r$ $^{1}($ $)$ is also a d-line (by Theorem 2) and it passes through A and B.

(Uniqueness) Suppose that $_1$ is another d-line through A and B. Then $r($ $_1)$ is a d-line that passes through O and B . It follows that $r($ $_1) =$, so $_1$ must be the same as r $^{1}($ $) =$. This proves the uniqueness of the d-line through A and B.

Problem 4 Let A_1 and A_2 be any two points of D. Use the Origin Lemma to prove that there is a hyperbolic transformation that maps A_1 to A_2.

In fact, we can establish a result stronger than that of Problem 4: we can determine a hyperbolic transformation which maps A_1 to A_2 *and* maps any given d-line through A_1 to any given d-line through A_2.

Theorem 5 Let A_1 and A_2 be any two points of D, and let $_1$ and $_2$ be d-lines through A_1 and A_2, respectively. Then there is a hyperbolic transformation which maps A_1 to A_2 and $_1$ to $_2$.

Proof By the Origin Lemma, there are hyperbolic transformations r_1 and r_2 which map A_1 and A_2, respectively, to the origin O. Let the images of $_1$ and $_2$ under r_1 and r_2 be the d-lines $_1$ and $_2$, respectively. Let r_3 be the hyperbolic transformation which rotates D about O so that $_1$ maps onto $_2$.

We shall verify in Subsection 6.2.1 that any rotation of D about O is a hyperbolic transformation.

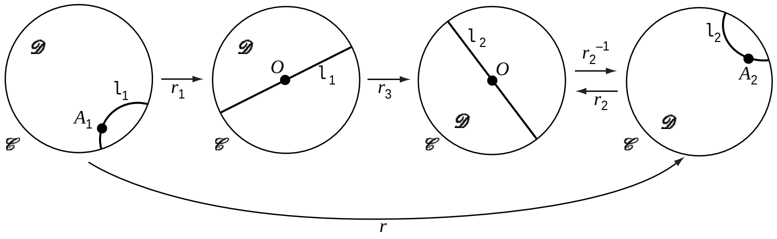

Now let r be the hyperbolic transformation given by

$$r = r_2{}^{1} \quad r_3 \quad r_1.$$

Then r maps A_1 to A_2 and $_1$ to $_2$, as required.

Remark

We can choose the rotation r_3 to map the two parts of $_1$ O onto the two parts of $_2$ O in whichever way we please. Hence we can arrange for r to map the two parts of $_1$ A_1 onto the two parts of $_2$ A_2 in whichever way we please.

In the following example we show that parallel d-lines map to parallel d-lines under a hyperbolic transformation.

Example 1 Let A, B and C be points of C such that the d-lines AB and BC are parallel. Let r be a hyperbolic transformation under which the images of AB and BC are the d-lines $A\ B$ and $B\ C$, respectively, where A, B and C are points of C. Show that $A\ B$ and $B\ C$ are parallel d-lines.

Solution First, we show that $A\ B$ and $B\ C$ do not meet in D. If they do, let them meet at the point P, say. Then the point P corresponds to a point P, say, on the d-lines AB and BC; that is, $r\ {}^{1}(P) = P$. But there is no such point, since AB and BC are parallel; it follows that $A\ B$ and $B\ C$ do not meet in D.

Next, the generalized circles of which $A\ B$ and $B\ C$ are parts meet at a point B on C. So we conclude that the images of $A\ B$ and $B\ C$ are parallel d-lines. □

Note the convenient use of the label AB for a d-line even though A and B do not lie on that d-line (they lie on C). There is no ambiguity since there is only one circle through A and B with centre the point where the tangents to C at A and B meet. Analogously, we shall say that a d-line passes through its boundary points.

Problem 5 Show that, given a d-line and a point P not on , there are exactly two d-lines through P which are parallel to .
Hint: Use the Origin Lemma.

6.1.3 Inversion Preserves Inverse Points

Earlier, we proved the following result.

Subsection 5.5.2, Corollary 2

> Let A and B be inverse points with respect to a generalized circle C, and let t be an inversive transformation. Then $t(A)$ and $t(B)$ are inverse points with respect to the generalized circle $t(C)$.

We now consider a special case of this result, which we can interpret in terms of hyperbolic geometry.

We take C to be a generalized circle of which part (a d-line) lies inside the unit disc $\mathbb{D}$. Let A be a point in $\mathbb{D}$; then its inverse with respect to , B say, is also in $\mathbb{D}$. We take t to be an inversion for which the generalized circle of inversion C^* is such that part (a d-line *) lies inside $\mathbb{D}$. Then, by the above result, $A \; = t(A)$ and $B \; = t(B)$ are inverse points in $\mathbb{D}$ with respect to the generalized circle $C \; = t(C)$, part of which lies in $\mathbb{D}$, namely the d-line $\; = t(\;)$.

See the diagram below.

Every inversion is an inversive transformation.

We now interpret this special case in terms of hyperbolic geometry.

Theorem 6 Let $A, B \quad \mathbb{D}$ be inverse points with respect to the d-line , and let A , B and be the images of A, B and under inversion (hyperbolic reflection) in another d-line *. Then A and B are inverse points with respect to inversion in .

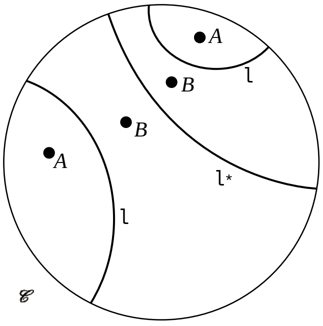

Remark

Note the convenient use of the term inversion to mean hyperbolic reflection . This is consistent with the use of this term in Chapter 5.

We shall make use of this theorem in Section 6.5, when we study the design on the cover of the book.

We finish this section with an illustration of the use of Theorem 6.

Example 2 Let be a d-line and P be a point in $\mathbb{D}$. Let Q be the image of P under inversion in the circle C of which is part. Let be the d-line through P and Q, meeting at the point S.

(a) Show that inversion in C maps to itself.

(b) Deduce that the d-lines and meet at right angles.

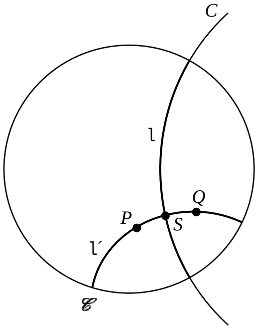

(c) Use Theorems 5 and 6 to show that there is a hyperbolic transformation mapping the figure to one in which the d-line is the x-axis and P and Q are complex conjugates.

Solution

(a) Inversion in C exchanges P and Q. So it maps , the unique d-line through P and Q, to itself.
(b) Inversion in C exchanges the angles makes with , and maps angles to angles of equal magnitude. So the angles must be equal, and since their sum is , each must be $/2$. Thus and meet at right angles.
(c) We use Theorem 5 to map the point S to O and the d-line to the x-axis. This makes the d-line part of the y-axis. By Theorem 6, P and Q are mapped to points that are inverse with respect to the x-axis, and so are complex conjugates of each other. □

6.2 Hyperbolic Transformations

6.2.1 Hyperbolic Transformations and Möbius Transformations

Each element of the group G_D of hyperbolic transformations is the composite of a finite number of reflections in d-lines. In this subsection we shall establish an explicit formula for a hyperbolic transformation in terms of Möbius transformations.

Subsection 5.3.3

We start by considering reflection in a d-line which is part of a circle C with centre the point R and radius r. Suppose that the d-line has boundary points P and Q. We let the coordinates of the point R be (a, b), and write the complex number $a + ib$ as , so that $| \; | > 1$ (which follows because is a d-line). We shall need the following observation: because the d-line is part of the circle with centre and radius r, the triangle RPO is right angled at P. So, by Pythagoras Theorem, $RP^2 + PO^2 = RO^2$. This implies that

We shall call this d-line the d-line obtained from .

$$r^2 + 1 = a^2 + b^2.$$

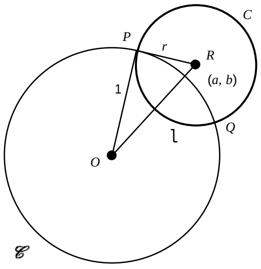

Using the fact that $a^2 + b^2 =$ fl, we may rewrite this equation as

$$r^2 \quad \text{fl} = \quad 1. \tag{1}$$

We now use equation (1), and the fact that inversion t in the circle C has the form

$$t(z) = \frac{r^2}{\overline{\overline{z}}} + \qquad (z \quad \mathrm{C} \quad \{ \; \}),$$

to obtain the form of the hyperbolic reflection in the d-line .

This follows from Subsection 5.2.1, Theorem 3.

Since is part of C, is given by

$$\begin{aligned} (z) &= \frac{r^2}{\overline{z}} + \\ &= \frac{r^2 + \bar{z} \quad \bar{}}{\bar{z} \quad \bar{}} \\ &= \frac{\bar{z} \quad 1}{\bar{z} \quad \bar{}} \quad (z \quad \mathbb{D}), \end{aligned}$$

by equation (1). Thus we have proved the following lemma.

Lemma 1 The hyperbolic reflection in the d-line that is part of a Euclidean circle with centre is given by the hyperbolic transformation

$$(z) = \frac{\bar{z} \quad 1}{\bar{z} \quad \bar{}} \qquad (z \quad \mathbb{D}).$$

Recall that $| \; | > 1$.

Notice that we may write

$$(z) = (M \quad B)(z), \qquad \text{for } z \quad \mathbb{D},$$

where $M(z) = \frac{z \quad 1}{z \quad \bar{}}$ is a Möbius transformation, and $B(z) = \bar{z}$ is complex conjugation. You know from Chapter 5 that Möbius transformations are important in inversive geometry: they preserve the magnitudes of angles, and they map generalized circles to generalized circles. Since angles and generalized circles are important in hyperbolic geometry, it is not surprising that particular sorts of Möbius transformations turn up here too.

Subsection 5.3.3, Theorem 5

Problem 1 Find the point which has image 0 under reflection in the d-line obtained from . Hence obtain a second proof of the Origin Lemma.

Reflection in a d-line which is a diameter of $\mathbb{D}$ is simply (Euclidean) reflection in that line. Recall that

Subsection 5.2.1

$$t_1(z) = \bar{z}$$

is reflection in the x-axis, and

$$t_2(z) = \quad z,$$

where $= \cos \quad + i \sin$ with $= \operatorname{Arg}$, is rotation about the origin through the angle . It follows that the composite

$$t_2 \quad t_1 \quad t_2^{\ 1} \quad (z) = \quad (\bar{z}) = \quad ^2\bar{z}$$

For $t_2^{\ 1}(z) = \bar{}z$

is reflection in the line with equation $y = x \tan$.

Thus hyperbolic reflection in the diameter on which $y = x \tan$ is given by

$$(z) = \quad ^2\bar{z} \qquad (z \quad \mathbb{D}),$$

where $= \cos \quad + i \sin$.

Example 1 Find the composite $_2 \quad _1$ of the hyperbolic reflections

$$_1(z) = {}^2\text{ﬂ}z \quad \text{and} \quad _2(z) = {}^2\text{ﬂ}z,$$

where $= \cos\ _1 + i \sin\ _1$ and $= \cos\ _2 + i \sin\ _2$, and interpret the transformation $_2 \quad _1$ geometrically.

Solution We have

$$(\ _2 \quad _1)(z) = {}^2\overline{(\ {}^2\text{ﬂ}z)} = {}^2\text{ﬂ}\,{}^2 z,$$

which is a Euclidean rotation about the origin of D. □

The composite of reflections and in two d-lines that are obtained from and , respectively, is found as follows. By Lemma 1, we have

$$(z) = \frac{\text{ﬂ}z \quad 1}{\text{ﬂ}z \quad \text{ﬂ}} = (M \quad B)(z)$$

Recall that $M(z) = \frac{z \quad 1}{z \quad \overline{\ }}$ and $B(z) = \text{ﬂ}z$.

and

$$(z) = \frac{\text{ﬂ}z \quad 1}{\text{ﬂ}z \quad \text{ﬂ}} = (M \quad B)(z),$$

where $M\ (z) = \frac{z \quad 1}{z \quad \text{ﬂ}}$ is a Möbius transformation. So

$$(\qquad)(z) = (M \quad B \quad M \quad B)(z).$$

But

$$(B \quad M)(z) = \frac{\overline{\ z} \quad 1}{\text{ﬂ}z} = (M \quad B)(z),$$

where $M(z)$ is the Möbius transformation $M(z) = \frac{\text{ﬂ}z \quad 1}{z}$. So

$$\begin{aligned}(\qquad)(z) &= (M \quad B \quad M \quad B)(z)\\ &= (M \quad M \quad B \quad B)(z) = (M \quad M)(z),\end{aligned}$$

because $(B \quad B)(z) = z$.

Now matrices associated with M and M are $\begin{pmatrix} & 1 \\ 1 & \text{ﬂ}\end{pmatrix}$ and $\begin{pmatrix}\text{ﬂ} & 1 \\ 1 & \end{pmatrix}$, respectively. Hence a matrix associated with the composite $M \quad M$ is

Here we are using the strategy in Subsection 5.3.5 for finding composites.

$$\begin{pmatrix} & 1 \\ 1 & \text{ﬂ}\end{pmatrix}\begin{pmatrix}\text{ﬂ} & 1 \\ 1 & \end{pmatrix} = \begin{pmatrix}\text{ﬂ} & 1 & \\ \text{ﬂ} & \text{ﬂ} & \text{ﬂ} \quad 1\end{pmatrix},$$

so that the composite of the hyperbolic reflections and is of the form

$$(\qquad)(z) = \frac{az + b}{\text{ﬂ}bz + \text{ﬂ}a},$$

where $a = \text{ﬂ} \quad 1, b = \quad$.

For ease of reference, we record this last result as a theorem.

Theorem 1 The composite of the hyperbolic reflections

$$(z) = \frac{\bar{z} \quad 1}{\bar{z} \quad \bar{\ }} \quad \text{and} \quad (z) = \frac{\bar{z} \quad 1}{\bar{z} \quad \bar{\ }} \qquad (z \quad D)$$

is the hyperbolic transformation

$$(\qquad)(z) = \frac{(\bar{\ } \qquad 1)z + \qquad}{(\bar{\ } \qquad \bar{\ })z + \quad \bar{\ } \quad 1} \qquad (z \quad D).$$

Example 2 Show that the hyperbolic reflection is its own inverse, by showing that $(\ (z)) = z$.

Solution From Theorem 1, we know that the composite of the reflection with itself is given by

$$(\qquad)(z) = \frac{(\bar{\ } \qquad 1)z + \qquad}{(\bar{\ } \quad \bar{\ })z + \quad \bar{\ } \quad 1} = \frac{(\bar{\ } \qquad 1)z}{\bar{\ } \quad 1} = z,$$

as required. □

$\bar{\ } = 1$ since $|\ | > 1$.

Example 3 Show that the composite $M_2 \quad M_1$ of the Möbius transformations

$$M_1(z) = \frac{az+b}{\bar{b}z+\bar{a}} \quad \text{and} \quad M_2(z) = \frac{cz+d}{\bar{d}z+\bar{c}}$$

is a Möbius transformation of the same form.

That is, of the form $\frac{ez+f}{\bar{f}z+\bar{e}}$, for some $e, f \quad$ C.

Solution By the strategy for finding composites of Möbius transformations, a matrix associated with the composite $M_2 \quad M_1$ is the matrix product

$$\begin{matrix} c & d \\ \bar{d} & \bar{c} \end{matrix} \qquad \begin{matrix} a & b \\ \bar{b} & \bar{a} \end{matrix} \quad = \quad \begin{matrix} ca+d\bar{b} & cb+d\bar{a} \\ \bar{d}a+\bar{c}\bar{b} & \bar{d}b+\bar{c}\bar{a} \end{matrix} \quad .$$

This product is of the required form, because

$$\overline{(ca+d\bar{b})} = \bar{d}b+\bar{c}\bar{a} \quad \text{and} \quad \overline{(cb+d\bar{a})} = \bar{d}a+\bar{c}\bar{b}.$$ □

It follows from this example that every composite of an even number of hyperbolic reflections in d-lines can be represented as a Möbius transformation, restricted to D, of the form

$$M(z) = \frac{az+b}{\bar{b}z+\bar{a}}, \quad z \quad D.$$

Similarly, it can be shown that every composite of an odd number of reflections can be represented as a Möbius transformation of this form composed with complex conjugation.

We omit the details of a proof of this.

It follows from the above discussion that any composite of a finite number of hyperbolic reflections, that is, any hyperbolic transformation, can be expressed in one of the forms

$$z \mapsto M(z) \quad \text{or} \quad z \mapsto M(\bar{z}) \quad (z \in \mathbb{D}),$$

where M is the Möbius transformation

$$M(z) = \frac{az+b}{\bar{b}z+\bar{a}}, \quad z \in \mathbb{D}.$$

Moreover, since $M(0) = b/\bar{a}$ and the image of 0 under a hyperbolic transformation must be in $\mathbb{D}$, we require that the above Möbius transformation must be such that

$$|b| < |a|.$$

The remaining question is: Do all such Möbius transformations represent hyperbolic transformations? In fact the answer is YES, as the following theorem shows.

Theorem 2 The restriction to $\mathbb{D}$ of every Möbius transformation of the form $M(z) = \frac{az+b}{\bar{b}z+\bar{a}}$ with $|b| < |a|$ is a composite of two hyperbolic reflections, and is therefore a hyperbolic transformation.

When the context is clear, we often omit to say that the Möbius transformation we are considering is restricted to $\mathbb{D}$.

The proof is a little devious (and no less elegant for that). We first prove it for the special case when the Möbius transformation maps the origin to itself. Using this special case, we then prove the general case.

Proof

Case 1 The Möbius transformation $M(z) = \frac{az+b}{\bar{b}z+\bar{a}}$ maps the origin to itself.

We shall show that M is the composite of two reflections.

The condition $M(0) = 0$ implies that $b = 0$, so that the Möbius transformation is simply $M(z) = \frac{a}{\bar{a}}z$ — a rotation about the origin.

Now, if we let $a = re^{i\theta}$, then

$$\frac{a}{\bar{a}} = e^{2i\theta}.$$

So, if we let σ_1 and σ_2 be the reflections in diameters of $\mathbb{D}$ given by

$$\sigma_1(z) = \bar{z}, \quad z \in \mathbb{D}, \quad \text{and} \quad \sigma_2(z) = e^{2i\theta}\,\bar{z}, \quad z \in \mathbb{D},$$

then

$$\sigma_2 \circ \sigma_1(z) = e^{2i\theta}\,\overline{(\bar{z})} = e^{2i\theta}z,$$

so that $M = \sigma_2 \circ \sigma_1$, as required.

Recall that reflection in the diameter $y = x\tan\theta$ of $\mathbb{D}$ is a mapping of the form $z \mapsto \lambda^2\bar{z}$, where $\lambda = e^{i\theta} = \cos\theta + i\sin\theta$ and $z \in \mathbb{D}$.

Case 2 The Möbius transformation $M(z) = \frac{az+b}{\bar{b}z+\bar{a}}$ does not map the origin to itself.

We shall again show that M is the composite of two hyperbolic reflections.

Consider a hyperbolic transformation φ given by $\varphi(z) = (M \circ B)(z)$, where $M\ (z) = \frac{z \quad 1}{z \quad \bar{\ }}$ is a Möbius transformation, and $B(z) = \bar{z}$ is complex conjugation. We shall choose so as to make φ and $M \circ \varphi$ hyperbolic reflections. This will show that $M = (M \circ \varphi) \circ \varphi^{-1}$ is a composite of two hyperbolic reflections.

Since M and M_λ are associated with the matrices

$$\begin{pmatrix} a & b \\ \bar{b} & \bar{a} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \lambda & -1 \\ 1 & -\bar{\lambda} \end{pmatrix},$$

respectively, the composite $M \circ M_\lambda$ is associated with the matrix

$$\begin{pmatrix} a & b \\ \bar{b} & \bar{a} \end{pmatrix} \begin{pmatrix} \lambda & -1 \\ 1 & -\bar{\lambda} \end{pmatrix} = \begin{pmatrix} a\lambda + b & -(\bar{\lambda} b + a) \\ \lambda\bar{b} + \bar{a} & -\bar{b} - \bar{a}\bar{\lambda} \end{pmatrix},$$

so we deduce that

$$(M \circ M_\lambda \circ B)(z) = \frac{(a\lambda + b)\bar{z} - (\bar{\lambda} b + a)}{(\lambda\bar{b} + \bar{a})\bar{z} - \bar{b} - \bar{a}\bar{\lambda}}.$$

Now since $M(0) \neq 0$, it follows that $b \neq 0$. Hence we may choose $\lambda = -\overline{(a/b)}$, so that

$$\bar{\lambda} b + a = -(a/b)b + a = 0.$$

For, σ is of the form required by Lemma 1.

Next, since $|b| < |a|$, we deduce that $|\lambda| = \frac{|a|}{|b|} > 1$, and so the transformation $\sigma = M_\lambda \circ B$ is a hyperbolic reflection as required. Moreover,

$$(M \circ \sigma)(z) = (M \circ M_\lambda \circ B)(z) = \frac{(a\lambda + b)\bar{z}}{-\bar{b} - \bar{a}\bar{\lambda}}$$

is of the form $-\frac{\mu}{\bar{\mu}}\bar{z}$, where $\mu = a\lambda + b$, and so the transformation $M \circ \sigma$ is a hyperbolic reflection. It follows that the transformation M is a composite of two hyperbolic reflections, $M \circ \sigma$ and σ^{-1}.

A similar argument shows that the more general transformation

$$z \mapsto \frac{a\bar{z} + b}{\bar{b}\bar{z} + \bar{a}}$$

is a composite of three reflections. For it is the composite $(M \circ B)(z)$ of the Möbius transformation

$$M(z) = \frac{az + b}{\bar{b}z + \bar{a}}$$

with complex conjugation $B(z) = \bar{z}$. Now, the Möbius transformation M is a composite of two reflections, and complex conjugation B is reflection in the x-axis. So $M \circ B$ is a composite of three reflections, as claimed.

Putting together all the above results, we have shown the following.

Theorem 3 Every hyperbolic transformation can be written as a composite of at most three hyperbolic reflections.

Example 4 Show that the composite $\sigma_2 \circ \sigma_1$ of the two reflections in lines through the origin given by $\sigma_1(z) = \alpha^2\bar{z}$ and $\sigma_2(z) = \beta^2\bar{z}$ can be written in the form $M(z) = \frac{az+b}{\bar{b}z+\bar{a}}$.

Solution From Example 1, we know that

$$(\ {}_2\ \ {}_1)(z) = {}^2\text{fl}^2 z.$$

Since $|\ | = 1$ and $|\ | = 1$,

$${}^1 = \text{fl} \quad \text{and} \quad {}^1 = \text{fl},$$

and so

$$(\ {}_2\ \ {}_1)(z) = M(z),$$

where

$$M(z) = \frac{az + b}{\text{fl}bz + \text{fl}a} \quad \text{with } a = \ \text{fl} \text{ and } b = 0. \qquad \square$$

Problem 2 Let $M(z) = \frac{az+b}{\text{fl}bz+\text{fl}a}$ be a hyperbolic transformation mapping the origin to itself. Show that $b = 0$.

Hyperbolic Rotations and Translations

It follows from the Angle Theorem of inversive geometry that a single hyperbolic reflection reverses the orientation of angles between d-lines. So a composite of two such transformations leaves the orientation unchanged, while a composite of three reverses it again. We call a hyperbolic transformation that leaves orientation unchanged a **direct** hyperbolic transformation, and one that reverses orientation **indirect.** So, by Theorem 3 above, we deduce that

Subsection 5.1.3, Theorem 5

> a direct hyperbolic transformation can be written as a composite of at most two hyperbolic reflections

Theorem 1 gives the form of a composite of two reflections.

and

> an indirect hyperbolic transformation can be written as a composite of at most three hyperbolic reflections.

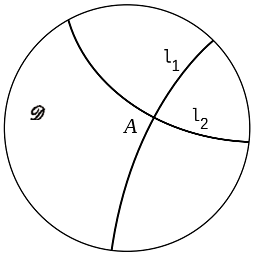

It is possible to say more about the direct transformations. Let r_1 and r_2 be reflections in the d-lines ${}_1$ and ${}_2$, respectively.

First, suppose that ${}_1$ and ${}_2$ intersect at some point A. (Certainly they cannot intersect at more than one point if they are distinct d-lines, since by Theorem 4 of Subsection 6.1.2 there is a unique d-line through any two points of D.) Then the composition $r_2\ \ r_1$ leaves the point A fixed, moves the points of D around A, and does not alter the orientation of D. Such a hyperbolic transformation is called a **(hyperbolic) rotation,** and has exactly one fixed point in D.

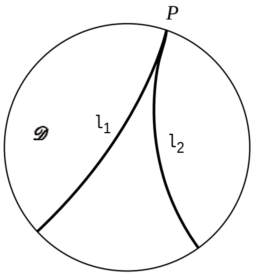

Next, suppose that the d-lines ${}_1$ and ${}_2$ are parallel, that is, the generalized circles of which they are parts do not meet in D but do meet at some point P on C. Then the composite $r_2\ \ r_1$ moves the points of D around P, and does not alter the orientation of D; it has no fixed point in D, but it leaves all the parallel lines with P as their common boundary point as parallel lines ending at P. We can regard this as the limiting case of a rotation (about a point of $\mathcal{C}$), and so call it a **(hyperbolic) limit rotation.**

Finally, suppose that the d-lines ℓ_1 and ℓ_2 are ultra-parallel, that is, the generalized circles of which they are parts do not meet in $\mathcal{D}$ or on $\mathcal{C}$. Then the composite $r_2 \circ r_1$ moves all the points of $\mathcal{D}$ in one general direction, and does not alter the orientation of $\mathcal{D}$; but no point of $\mathcal{D}$ (or $\mathcal{C}$) remains fixed. Such a hyperbolic transformation is called a **(hyperbolic) translation.**

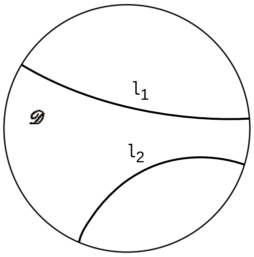

The analogy between Euclidean geometry and hyperbolic geometry is not exact, however. For instance, the composite of two Euclidean translations is independent of the order in which they are applied, whereas the composite of two hyperbolic translations may not be independent of the order in which they are applied.

6.2.2 The Canonical Form of a Hyperbolic Transformation

In this subsection we show how to write a hyperbolic transformation in the most suitable form for applications. We shall do this only for a direct transformation of the form $M(z) = \frac{az+b}{\bar{b}z+\bar{a}}$, where $|b| < |a|$. Our result is the following theorem.

Theorem 4 A direct hyperbolic transformation M can be written in the form

$$M(z) = K\frac{z-m}{1-\bar{m}z},$$

where K and m are complex numbers with $|K| = 1$ and $m \in \mathcal{D}$.

Remark

We call this form the **canonical form** of a direct hyperbolic transformation. It has the great advantage of showing that the transformation M maps the point m to the origin.

Proof We know from the previous subsection that a direct hyperbolic transformation can always be written in the form

$$M(z) = \frac{az+b}{\bar{b}z+\bar{a}}, \quad \text{with } |b| < |a|.$$

See the discussion before Theorem 2.

Indeed, on dividing the expression for $M(z)$ above and below by $\bar{a}$, we can write it as

$$M(z) = \frac{\frac{a}{\bar{a}}z + \frac{b}{\bar{a}}}{1-\left(-\frac{\bar{b}}{\bar{a}}\right)z} = \frac{a}{\bar{a}}\left(\frac{z-\left(-\frac{b}{a}\right)}{1-\left(-\frac{\bar{b}}{\bar{a}}\right)z}\right),$$

which is of the required form, with $K = a/\bar{a}$ and $m = -b/a$. Since $|a| = |\bar{a}|$, it follows that $|K| = 1$, as required. Since $|b| < |a|$, it follows that $|m| < 1$, as required.

We can now find the form of every direct hyperbolic transformation that maps a point m of D to the origin. We know from Theorem 4 that it must be of the form

$$M(z) = K\frac{z \quad m}{1 \quad \overline{m\ } z}, \qquad \text{where } |K| = 1 \text{ and } |m\ | < 1.$$

The condition that $M(m) = 0$ implies that $m\ = m\ $, so we deduce that in fact *every* direct hyperbolic transformation mapping the point m to the origin is of the form

$$M(z) = K\frac{z \quad m}{1 \quad \overline{m}z}.$$

The direct hyperbolic transformations that map the origin to itself are therefore those for which $m\ = 0$, so they are of the form $M(z) = Kz$ with $|K| = 1$. These are (Euclidean) rotations of the disc D about the origin through an angle , where $K = \cos \quad + i \sin$. In general, if all that is required is just one direct hyperbolic transformation sending a given point m to the origin, then we may set $K = 1$, and use the transformation $M(z) = \frac{z \quad m}{1 \quad \overline{m}z}$.

$K = 1$ is the most convenient value to choose.

Example 5

(a) Find the general form of a direct hyperbolic transformation that maps the point $\frac{1}{2}i$ to the origin.
(b) Find one direct hyperbolic transformation that maps $\frac{3}{4}$ to the origin.

Solution

(a) The general form is

$$M(z) = K\frac{z \qquad \frac{1}{2}i}{1 \qquad \frac{1}{2}i \quad z} = K\frac{2z \quad i}{iz + 2}, \qquad \text{where } |K| = 1.$$

(b) Taking $K = 1$ in the formula of Theorem 4, as suggested above, one such transformation is

$$M(z) = \frac{z \quad \frac{3}{4}}{1 \quad \frac{3}{4}z}. \qquad \square$$

Problem 3 Determine all the direct hyperbolic transformations that map the origin to the origin and the line $y = x/\quad \overline{3}$ to the line $y = \quad \overline{3}x$.

Problem 4 For each of the following points, determine all the direct hyperbolic transformations that map it to the origin.
(a) $\frac{1}{4}i$ (b) $\quad \frac{1}{3} + \frac{2}{3}i$

The general form of the inverse of a direct hyperbolic transformation M mapping the point m to the origin is a direct hyperbolic transformation sending the origin to m. To find this inverse explicitly, we write

$$\begin{aligned} M(z) &= K\frac{z - m}{1 - \overline{m}z} \quad \text{(by Theorem 4)} \\ &= \frac{Kz - Km}{-\overline{m}z + 1}, \end{aligned}$$

and make use of the fact that M is the restriction to $\mathbb{D}$ of a Möbius transformation to write

$$M^{-1}(z) = \frac{z + Km}{\overline{m}z + K}, \tag{2}$$

where $|K| = 1$.

By restriction to $\mathbb{D}$ we mean that we only consider those z in $\mathbb{D}$.

This form for M^{-1} follows immediately from Subsection 5.3.6, Theorem 7.

Problem 5 Determine the general form of the inverse of the direct hyperbolic transformation M which maps $\frac{3}{4}$ to the origin.

Problem 6 Prove that any direct hyperbolic transformation M that maps the diameter $(-1, 1)$ onto itself must be of the form

$$M(z) = -\frac{z - m}{1 - mz}, \qquad \text{where } m \in (-1, 1).$$

Since an *indirect* hyperbolic transformation can be written as the composite of at most three reflections, it follows that the general form of such a transformation is

$$z \mapsto \frac{a\overline{z} + b}{\overline{b}\overline{z} + \overline{a}}.$$

See the discussion preceding Theorem 2.

An argument similar to that for direct transformations shows that this transformation can always be written in the form

$$z \mapsto K\frac{\overline{z} - m}{1 - \overline{m}\overline{z}}, \qquad \text{where } |k| = 1 \text{ and } |m| < 1.$$

We call this form the **canonical form** of an indirect hyperbolic transformation.

Theorem 4 gives the general form of a direct hyperbolic transformation that maps a given point m in $\mathbb{D}$ to the origin. By two applications of the theorem, therefore, we can obtain the general form of a direct hyperbolic transformation that maps any given point in $\mathbb{D}$ to any other given point in $\mathbb{D}$.

If we require a particular direct transformation that maps p to q, then we may choose a particular M_1 and a particular $\mathbf{A}_1$.

The constant arising in Step 1 means that it is sufficient to use any particular transformation in Step 2, since the final general transformation requires only one constant.

Strategy To determine the general form of the direct hyperbolic transformation that maps one point p in $\mathbb{D}$ to another point q in $\mathbb{D}$:

1. write down the general form of the direct transformation M_1 that maps p to 0, and a matrix $\mathbf{A}_1$ associated with M_1;
2. write down a direct transformation M_2 that maps q to 0, and a matrix $\mathbf{A}_2$ associated with M_2;

3. form the matrix product $\mathbf{A}_2^{-1}\mathbf{A}_1$ associated with the direct transformation $M_2^{-1} \circ M_1$, and hence write down the general form of the required direct transformation $M_2^{-1} \circ M_1$.

The following figure illustrates this strategy.

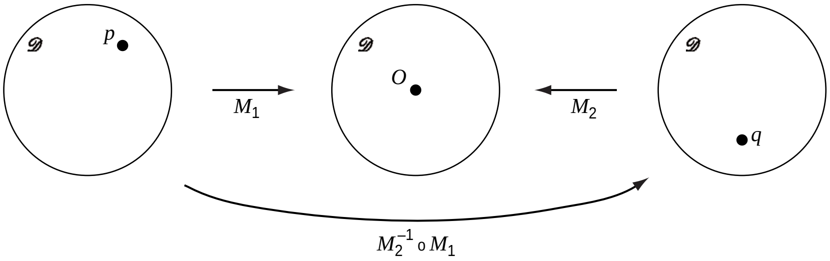

Example 6 Determine the general form of the direct hyperbolic transformation that maps $\frac{1}{2}i$ to $\frac{3}{4}$.

Solution We have already seen that the general form of the direct hyperbolic transformation M_1 that maps $\frac{1}{2}i$ to 0 is

Example 5, part (a).

$$M_1(z) = K\frac{2z - i}{iz + 2}, \qquad \text{where } |K| = 1;$$

a matrix associated with M_1 is

$$\mathbf{A}_1 = \begin{pmatrix} 2K & -iK \\ i & 2 \end{pmatrix}.$$

Also, the direct transformation

$$M_2(z) = \frac{z - \frac{3}{4}}{1 - \frac{3}{4}z}$$

Example 5, part (b).

maps $\frac{3}{4}$ to 0; a matrix associated with M_2 is

$$\mathbf{A}_2 = \begin{pmatrix} 1 & -\frac{3}{4} \\ -\frac{3}{4} & 1 \end{pmatrix}.$$

The inverse of $\mathbf{A}_2$ is

$$\mathbf{A}_2^{-1} = \tfrac{16}{7}\begin{pmatrix} 1 & \frac{3}{4} \\ \frac{3}{4} & 1 \end{pmatrix}.$$

So a matrix associated with the required direct transformation is

$$\begin{aligned} \mathbf{A}_2^{-1}\mathbf{A}_1 &= \tfrac{16}{7}\begin{pmatrix} 1 & \frac{3}{4} \\ \frac{3}{4} & 1 \end{pmatrix}\begin{pmatrix} 2K & -iK \\ i & 2 \end{pmatrix} \\ &= \tfrac{4}{7}\begin{pmatrix} 8K + 3i & -4iK + 6 \\ 6K + 4i & -3iK + 8 \end{pmatrix}. \end{aligned}$$

Hence any direct transformation that maps $\frac{1}{2}i$ to $\frac{3}{4}$ may be written in the form

The fraction $\frac{4}{7}$ disappears at this point, as it is simply a multiple of the whole matrix.

$$M(z) = \frac{(8K + 3i)z + (-4iK + 6)}{(6K + 4i)z + (-3iK + 8)}, \qquad \text{where } |K| = 1. \qquad \square$$

Problem 7

(a) Determine the general form of the direct hyperbolic transformation that maps $\frac{1}{3}i$ to $\frac{2}{3}$.
(b) Determine the direct hyperbolic transformation that also maps i to 1.

Problem 8 Determine the general form of the direct hyperbolic transformation that maps $\frac{1}{2}$ to $\frac{1}{2}$.

These functions fix the point $\frac{1}{2}$.

Problem 9 For each of the following pairs of points, either find a direct hyperbolic transformation mapping the first point to 0 and the second point to $\frac{1}{2}$, or prove that no such transformation exists.

(a) $\frac{1}{2}i$ and 0 (b) $\frac{1}{2}$ and $\frac{2}{3}$ (c) $\frac{1}{3}(1+i)$ and $\frac{1}{3}(1-i)$.

6.3 Distance in Hyperbolic Geometry

In the Euclidean geometry of the plane, corresponding to any two points there exists a non-negative number called the distance between the points. This is given by the formula

$$d(z_1, z_2) = |z_1 - z_2|, \quad z_1, z_2 \in \mathbb{C}.$$

So $d(z_1, z_2) = d(z_2, z_1)$.

In this section, we introduce an analogous formula for the distance in hyperbolic geometry between any two points of the unit disc $\mathbb{D}$.

6.3.1 The Distance Formula

We begin by looking at various properties that we would expect any distance function d to have in any geometry whose points lie in the complex plane $\mathbb{C}$. Ordinary Euclidean distance, for example, clearly possesses the following four properties.

Properties of a Distance Function *d*

1. $d(z_1, z_2) \geq 0$ for all z_1 and z_2;
   $d(z_1, z_2) = 0$ if and only if $z_1 = z_2$.
2. $d(z_1, z_2) = d(z_2, z_1)$ for all z_1 and z_2.
3. $d(z_1, z_3) + d(z_3, z_2) \geq d(z_1, z_2)$ for all z_1, z_2 and z_3.
4. $d(z_1, z_3) + d(z_3, z_2) = d(z_1, z_2)$ if and only if z_1, z_3 and z_2 lie in this order on a line.

Property 3 is known as the *Triangle Inequality.*

Property 1 asserts that the distance between any two points in the geometry is always positive, unless the two points coincide — in which case the distance between them must be zero.

Property 2 asserts that the distance from z_1 to z_2 is the same as the distance from z_2 to z_1.

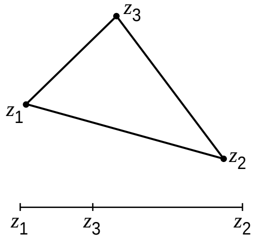

Property 3 asserts that the distance from z_1 to z_2 is always less than (or equal to) the distance from z_1 to another point z_3 plus the distance from z_3 to z_2. In Euclidean geometry this property may be rewritten in the form

$$|z_1 - z_3| + |z_3 - z_2| \geq |z_1 - z_2|.$$

Property 4 asserts that distance along a line is additive.

However, there are some additional properties that we wish the distance function in hyperbolic geometry to possess.

Additional Properties of the Distance Function *d* in Hyperbolic Geometry

5. $d(z_1, z_2) = d(\bar{z}_1, \bar{z}_2)$ for all z_1 and z_2 in $\mathbb{D}$.
6. $d(z_1, z_2) = d(M(z_1), M(z_2))$ for all z_1 and z_2 in $\mathbb{D}$ and all direct hyperbolic transformations M in $G_{\mathbb{D}}$.

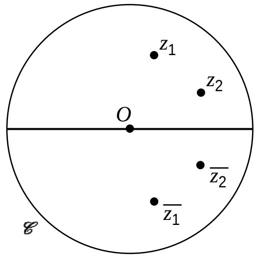

Properties 5 and 6 together assert that hyperbolic transformations of the unit disc $\mathbb{D}$ do not alter distances between points, since a hyperbolic transformation is either a direct transformation of the unit disc to itself (as in Property 6) or the composite of such a function with the conjugation function $z \mapsto \bar{z}$ (which, by Property 5, preserves distance).

These additional properties enable us to make some useful observations about the form of the distance function. First, the direct hyperbolic transformation

$$M : z \mapsto \frac{z - z_1}{1 - \bar{z}_1 z}, \quad \text{where } z_1 \in \mathbb{D},$$

maps z_1 to 0 and z_2 to $\frac{z_2 - z_1}{1 - \bar{z}_1 z_2}$. If R is the rotation of the unit disc $\mathbb{D}$ about the origin which sends $\frac{z_2 - z_1}{1 - \bar{z}_1 z_2}$ to the point $\left|\frac{z_2 - z_1}{1 - \bar{z}_1 z_2}\right|$, then overall the composite $R \circ M$ is a direct hyperbolic transformation which sends z_1 to 0 and z_2 to $\left|\frac{z_2 - z_1}{1 - \bar{z}_1 z_2}\right|$.

Recall that any rotation of $\mathbb{D}$ about the origin is a direct hyperbolic transformation (Subsection 6.2.1).

By Property 6 it follows that

$$d(z_1, z_2) = d\left(0, \left|\frac{z_2 - z_1}{1 - \bar{z}_1 z_2}\right|\right), \quad \text{for all } z_1, z_2 \in \mathbb{D}.$$

This shows that the distance $d(z_1, z_2)$ must be some function of the quantity $\left|\frac{z_2 - z_1}{1 - \bar{z}_1 z_2}\right|$ alone. But what function ensures that d has the Properties 1-6 above? Indeed, does such a function exist? In fact, there is essentially only one well-behaved function which yields these properties, and it turns out to be the function $\tanh^{-1}$.

We verify this assertion in Subsection 6.3.5.

Definition The **hyperbolic distance** $d(z_1, z_2)$ between the points z_1 and z_2 in the unit disc $\mathbb{D}$ is defined by

$$d(z_1, z_2) = \tanh^{-1}\left(\left|\frac{z_2 - z_1}{1 - \bar{z}_1 z_2}\right|\right).$$

Of course before we can be sure that this is a reasonable definition of distance we need to check that d satisfies all the Properties 1–6. Clearly d satisfies Properties 1, 2, 5, and we shall prove the other properties later in this section; for the moment we assume that all the properties hold, and we use them to explore some of the practical consequences of the definition.

First observe that the formula for the distance between two given points is particularly simple in the case where one of the points lies at the origin. Indeed we have

Definition The **hyperbolic distance** $d(0,z)$ between the points 0 and z in the unit disc $\mathbb{D}$ is

$$d(0,z) = \tanh^{-1}(|z|).$$

Since this formula is simpler to remember, we shall tend to use it more often than the formula for $d(z_1, z_2)$!

Since $|z|$ is just the Euclidean distance of z from the origin, this equation tells us that we can obtain the hyperbolic distance of a point z from the origin by applying the inverse tanh function to the Euclidean distance of z from the origin. The graph of the inverse tanh function in the margin reveals two important characteristics of hyperbolic distance.

1. Near the origin the graph is nearly a straight line with slope 1, so for a point z near the origin the hyperbolic distance of z from 0 is approximately equal to the Euclidean distance of z from 0. (This is analogous to the situation on the surface of the Earth; small portions of the Earth's surface *look* flat and distances between its points are *approximately* Euclidean.)
2. As the point z approaches the boundary of the disc $\mathbb{D}$, the hyperbolic distance of z from the origin increases without bound. Indeed, as the Euclidean distance $|z|$ tends to 1, the hyperbolic distance $\tanh^{-1}(|z|)$ tends to ∞. From the point of view of someone living in the hyperbolic geometry, the boundary points appear to be infinitely far away—an observation that is consistent with the idea that parallel lines 'meet' on the boundary.

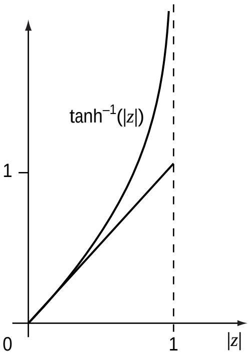

With the help of a calculator we can easily use the formula

$$d(0,z) = \tanh^{-1}(|z|) \tag{1}$$

to calculate the hyperbolic distance of a point from the origin.

At the end of this subsection we prove the alternative formula

$$d(0,z) = \tfrac{1}{2}\log_e\left(\frac{1+|z|}{1-|z|}\right)$$

which you can use if your calculator does not have a tanh function.

For example, the hyperbolic distance of $\frac{1}{2}$ from the origin is given by

$$d\left(0,\tfrac{1}{2}\right) = \tanh^{-1}\left(\frac{1}{2}\right) = \tanh^{-1}(0.5) \simeq 0.549.$$

But we can also use equation (1) to calculate the hyperbolic distance between any two points which lie on a diameter of the disc $\mathbb{D}$.

Example 1 Find the hyperbolic distance between the points 0.1 and 0.2.

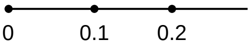

Solution Here we use Property 4, which states that distances along a (hyperbolic) line are additive. In particular, this implies that

$$d(0.1, 0.2) = d(0, 0.2) \quad d(0, 0.1),$$

so we can write

$$\begin{aligned} d(0.1, 0.2) &= d(0, 0.2) \quad d(0, 0.1) \\ &= \tanh\ ^{1}(0.2) \quad \tanh\ ^{1}(0.1) \\ & \quad 0.203 \quad 0.100 \\ & \quad 0.102. \end{aligned}$$

□

Throughout Section 6.3 we shall work to the full accuracy of our calculator, but we shall record our results only to 3 or 4 decimal places.

Problem 1 Determine the hyperbolic distances $d \quad 0, \tfrac{1}{3}i$ and $d(0.8, 0.9)$.

By rearranging equation (1) we obtain the formula

$$|z| = \tanh d(0, z). \tag{2}$$

This can be used to locate a point, given its hyperbolic distance along a radius of the disc D. For example, the point a that is at a hyperbolic distance 0.1 from the origin along the positive real axis is given by $a = \tanh 0.1 \quad 0.0997$.

Problem 2 The following table (plotted in the margin) illustrates how points bunch up towards the boundary of D as their hyperbolic distance from 0 doubles.

Find the two missing entries in the table.

$d(0, a), a > 0$	0.2	0.4	0.8	1.6	3.2
a	0.197 . . .	0.380 . . .	0.664 . . .		

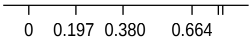

If we have to calculate the distance between two points that do not lie on a diameter of D, then we can use the distance formula given in the definition.

Example 2 Find the hyperbolic distance between the points $\tfrac{1}{2}$ and $\tfrac{1}{3}i$.

Solution From the definition of hyperbolic distance we have

$$\begin{aligned} d \quad \tfrac{1}{2}, \tfrac{1}{3}i &= \tanh\ ^{1} \quad \frac{\tfrac{1}{2} \quad \tfrac{1}{3}i}{1 \quad \overline{\tfrac{1}{3}i} \bullet \tfrac{1}{2}} \\ &= \tanh\ ^{1} \quad \frac{3 \quad 2i}{6 + i} \\ &= \tanh\ ^{1} \quad \overline{\tfrac{13}{37}} \\ & \quad 0.6819. \end{aligned}$$

□

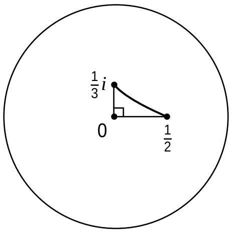

We can use the distances that we have calculated above to demonstrate that Pythagoras' Theorem fails to hold in hyperbolic geometry, at least without some rewording. For, if we consider the hyperbolic right-angled triangle with vertices at 0, $\frac{1}{3}i$ and $\frac{1}{2}$, then the square of the hypotenuse is

$$\left(d\left(\tfrac{1}{2}, \tfrac{1}{3}i\right)\right)^2 \simeq 0.6819^2 \simeq 0.4650,$$

whereas the sum of the squares of the other two sides is

$$\left(d\left(0, \tfrac{1}{2}\right)\right)^2 + \left(d\left(0, \tfrac{1}{3}i\right)\right)^2 \simeq 0.5493^2 + 0.3466^2 \simeq 0.4219.$$

We evaluated $d\left(0, \frac{1}{2}\right)$ before Example 1, and $d\left(0, \frac{1}{3}i\right)$ in Problem 1.

We shall see later on how Pythagoras' Theorem can be reformulated in hyperbolic geometry.

Subsection 6.4.3, Theorem 8

We end this subsection by giving an alternative formula for the hyperbolic distance of a point z from the origin. First, observe that

$$\begin{aligned} |z| &= \tanh d(0,z) \\ &= \frac{\sinh d(0,z)}{\cosh d(0,z)} \\ &= \frac{e^{d(0,z)} - e^{-d(0,z)}}{e^{d(0,z)} + e^{-d(0,z)}} \\ &= \frac{e^{2d(0,z)} - 1}{e^{2d(0,z)} + 1}. \end{aligned}$$

We can solve this equation to obtain an expression for $d(0,z)$ in terms of $|z|$. By cross-multiplication, we obtain

$$e^{2d(0,z)} - 1 = |z|\left(e^{2d(0,z)} + 1\right),$$

which implies that

$$e^{2d(0,z)}(1 - |z|) = 1 + |z|,$$

which is equivalent to

$$e^{2d(0,z)} = \frac{1 + |z|}{1 - |z|}.$$

Taking natural logarithms of both sides, we deduce that

$$2d(0,z) = \log_e\left(\frac{1 + |z|}{1 - |z|}\right),$$

and so

$$d(0,z) = \frac{1}{2}\log_e\left(\frac{1 + |z|}{1 - |z|}\right). \qquad (3)$$

This formula is useful if your calculator does not have a tanh function.

The formula for $d(0,z)$ is sometimes quoted in this form.

6.3.2 Hyperbolic Midpoints

Having described how to calculate hyperbolic distances we now introduce the idea of the hyperbolic midpoint of a hyperbolic line segment.

Definition A point m is the **hyperbolic midpoint** of the hyperbolic line segment joining a and b if m lies on this segment and

$$d(a,m) = d(m,b) = \frac{1}{2}d(a,b).$$

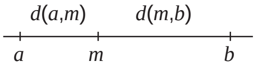

For simplicity, we confine our attention to midpoints of line segments that lie along a diameter of $\mathbb{D}$. The method depends on whether the endpoints of the segment lie on the same side of the origin O, or on opposite sides of O.

Example 3 Find the hyperbolic midpoint m of the line segment which joins each of the following pairs of points.

(a) $\frac{1}{4}i$ and $\frac{3}{4}i$ (b) $\frac{1}{4}i$ and $\frac{3}{4}i$

Solution First observe that

$$d\ \ 0, \tfrac{1}{4}i \ \ = \tanh^{\ 1}\ \ \tfrac{1}{4}i \ \ = \tanh^{\ 1}(0.25) = 0.255\ldots,$$

and

$$d\ \ 0, \tfrac{3}{4}i \ \ = \tanh^{\ 1}\ \ \tfrac{3}{4}i \ \ = \tanh^{\ 1}(0.75) = 0.973\ldots.$$

(a) Here $\frac{3}{4}i$ and $\frac{1}{4}i$ lie on the same side of 0, so the midpoint must lie at a hyperbolic distance

$$\tfrac{1}{2}\ \ d\ \ 0, \tfrac{3}{4}i \ \ + d\ \ 0, \tfrac{1}{4}i \qquad \tfrac{1}{2}(0.973 + 0.255) = 0.614$$

from 0 on the radius through $\frac{3}{4}i$. That is, $m \quad (\tanh 0.614)i \quad 0.547i$. (Notice that this is further from 0 than the Euclidean midpoint $0.5i$.)

(b) Here $\frac{3}{4}i$ and $\frac{1}{4}i$ lie on opposite sides of 0, so the midpoint lies at a hyperbolic distance

$$\tfrac{1}{2}\ \ d\ \ 0, \quad \tfrac{3}{4}i \qquad d\ \ 0, \tfrac{1}{4}i \qquad \tfrac{1}{2}(0.973 \quad 0.255) = 0.359$$

from 0 on the radius through $\frac{3}{4}i$. That is, $m \quad (\tanh 0.359)i \quad 0.344i$. □

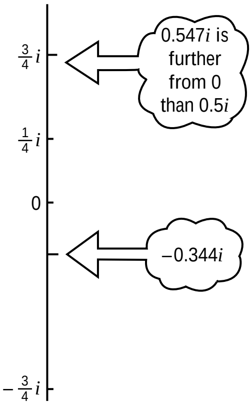

The following strategy generalizes the method used in Example 3.

Strategy To find the hyperbolic midpoint of a line segment joining two points p, q on a diameter of $\mathbb{D}$, where $|p| > |q|$:

1. calculate $d(0, p)$ and $d(0, q)$;
2. if p and q lie on opposite sides of 0, then calculate

$$d = \tfrac{1}{2}(d(0,p) \quad d(0,q)),$$

otherwise calculate

$$d = \tfrac{1}{2}(d(0,p) + d(0,q));$$

3. then the hyperbolic midpoint is the point m on the radius through 0 and p at a Euclidean distance $\tanh(d)$ from 0.

Problem 3 Find the hyperbolic midpoint m of the line segment which joins each of the following pairs of points:

0.5 and 0.8; 0.2 and 0.8.

Remark

To calculate the hyperbolic midpoint of a line segment which joins two *arbitrary* points p, q in $\mathbb{D}$ we would use a Möbius transformation M to map p to 0 and q to $M(q)$. After calculating the hyperbolic midpoint m' of the segment from 0 to $M(q)$, we would then obtain the midpoint of the original segment by calculating $M^{-1}(m')$.

6.3.3 Hyperbolic Circles

By analogy with Euclidean circles, we define a hyperbolic circle to be the locus of points which are a fixed hyperbolic distance from a fixed point.

Definition The **hyperbolic circle** with **hyperbolic radius** r and **hyperbolic centre** at c is the set defined by

$$\{z : d(c,z) = r, \quad z \in \mathbb{D}\}.$$

It is natural to ask what a hyperbolic circle looks like to the Euclidean eye. In the case of a hyperbolic circle with radius r centred at the origin, it is just the set of points given by

$$\{z : d(0,z) = r, \quad z \in \mathbb{D}\};$$

that is, the set

$$\{z : \tanh^{-1}(|z|) = r, \quad z \in \mathbb{D}\},$$

or

$$\{z : |z| = \tanh r, \quad z \in \mathbb{D}\}.$$

This is a Euclidean circle with radius $\tanh r$ centred at the origin! Even more remarkable is the fact that *every* hyperbolic circle is a Euclidean circle.

Theorem 1 Every hyperbolic circle is a Euclidean circle in $\mathbb{D}$, and vice versa.

Proof We have already established the result for circles centred at 0, so let C be any hyperbolic circle with hyperbolic centre $m \neq 0$. Let the diameter of $\mathbb{D}$ through m meet C at the points a and b, and let K be the Euclidean circle which has ab as a diameter and (Euclidean) centre p.

For the steps in the argument before Theorem 1 can be reversed.

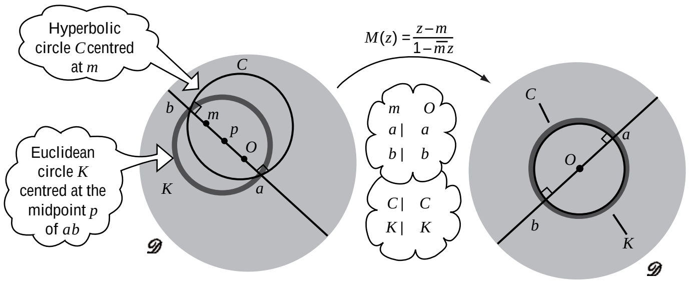

We will show that C is a Euclidean circle by showing that it coincides with K. To do this let M be the hyperbolic transformation defined by

$$M(z) = \frac{z-m}{1-\overline{m}z}, \quad z \in \mathbb{D},$$

and let a' and b' be the images of a and b under M.

Since M preserves hyperbolic distances, and maps m to O, it must map C to a hyperbolic circle C' with centre O. But we already know that such a circle is also a Euclidean circle. Moreover, since the line segment ab passes through m, its image $a'b'$ passes through O and is therefore a diameter of C'.

Also, since M is angle-preserving, and since ab meets K at right angles, it follows that M maps K to a circle K' which meets $a'b'$ at right angles at the points a' and b'. This implies that $a'b'$ is also a diameter of K'. It follows that C' coincides with K', and hence that C coincides with K.

Theorem 2, Subsection 6.1.1

Conversely, let K be any Euclidean circle in $\mathbb{D}$ with Euclidean centre $p \neq 0$, and let the diameter of $\mathbb{D}$ through p meet K at the points a and b. If C is the hyperbolic circle through a and b with hyperbolic centre the hyperbolic midpoint m of ab, then we can use the same argument as before to deduce that K coincides with C, and so we conclude that K is a hyperbolic circle. ■

The above proof shows that the Euclidean and hyperbolic centres of a circle lie on the same line through O. This observation enables us to write down the following strategy for finding the Euclidean centre and radius of a hyperbolic circle.

Strategy To find the Euclidean centre and radius of a hyperbolic circle C with hyperbolic centre m:

1. find the points a, b where Om meets C;
2. the Euclidean centre of C is the Euclidean midpoint of ab;
3. the Euclidean radius of C is $\frac{1}{2}|a \quad b|$.

Example 4 Find the Euclidean centre and radius of the hyperbolic circle

$$C = \& \; z : d \quad z, \tfrac{1}{2} \quad = \tfrac{1}{2} \; ’ .$$

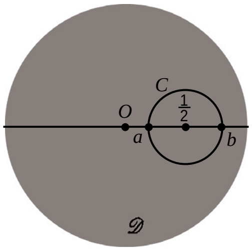

Solution Here the hyperbolic centre m is the point $\frac{1}{2}$, and so

$$d(0, m) = \tanh^{\;1} \; \tfrac{1}{2} \qquad 0.549.$$

Since the hyperbolic radius of C is equal to $\frac{1}{2}$, it follows that Om meets C at the points a, b, where

$$d(0, a) \quad 0.549 \quad 0.5 = 0.049,$$
$$d(0, b) \quad 0.549 + 0.5 = 1.049.$$

So

$$a \quad \tanh 0.049 \quad 0.049,$$
$$b \quad \tanh 1.049 \quad 0.782.$$

Since a and b both lie on the same side of 0, we have

$$\text{Euclidean centre} = \tfrac{1}{2}(a + b) \quad 0.415,$$
$$\text{Euclidean radius} = \tfrac{1}{2}|a \quad b| \quad 0.366.$$ □

Problem 4 Determine the Euclidean centre and radius of the hyperbolic circle

$$C = \& \; z : d \quad z, \quad \tfrac{1}{4} \quad = \tfrac{1}{2} \; ’ .$$

A similar strategy can be used to find the hyperbolic centre and radius of a Euclidean circle.

Strategy To find the hyperbolic centre and radius of a Euclidean circle K with Euclidean centre p:

1. find the points a, b where Op meets K;
2. the hyperbolic centre of K is the hyperbolic midpoint of ab;
3. the hyperbolic radius of K is $\frac{1}{2}d(a, b)$.

Example 5 Find the hyperbolic centre and radius of the Euclidean circle

$$K = \left\{ z : \left| z + \tfrac{1}{2} i \right| = \tfrac{1}{4} \right\}.$$

Solution Here the Euclidean centre p is the point $-\frac{1}{2}i$, and the Euclidean radius is $\frac{1}{4}$, so Op meets K at the points $a = -\frac{1}{4}i$ and $b = -\frac{3}{4}i$. Thus

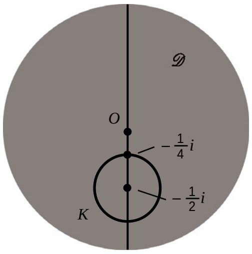

$$d(0,a) = \tanh^{-1}\left|-\tfrac{1}{4}i\right| = \tanh^{-1}(0.25) \approx 0.255\ldots;$$

$$d(0,a) = \tanh^{-1}\left|-\tfrac{3}{4}i\right| = \tanh^{-1}(0.75) \approx 0.973\ldots.$$

Since a and b both lie on the same side of O, the hyperbolic centre m of K is given by

$$d(0,m) = \tfrac{1}{2}(d(0,a) + d(0,b)) \approx 0.614,$$

so that

$$\text{hyperbolic centre } m \approx -i \bullet \tanh 0.614 \approx -0.547i;$$

$$\text{hyperbolic radius} = \tfrac{1}{2}|d(0,a) - d(0,b)| \approx 0.359. \qquad \square$$

Problem 5 Determine the hyperbolic centre and radius of the Euclidean circle

$$K = \left\{ z : \left| z - \tfrac{1}{4} \right| = \tfrac{1}{2} \right\}.$$

We now use the fact that hyperbolic circles are also Euclidean circles to prove the Triangle Inequality property of hyperbolic distance.

Property 3 of a distance function, Subsection 6.3.1.

Theorem 2 For all z_1, z_2, z_3 in $\mathbb{D}$:

$$d(z_1,z_3) + d(z_3,z_2) \geq d(z_1,z_2).$$

Proof First let M be a hyperbolic transformation that maps z_1 to 0, and let $M(z_2) = b$ and $M(z_3) = c$. Then Ob is a straight line.

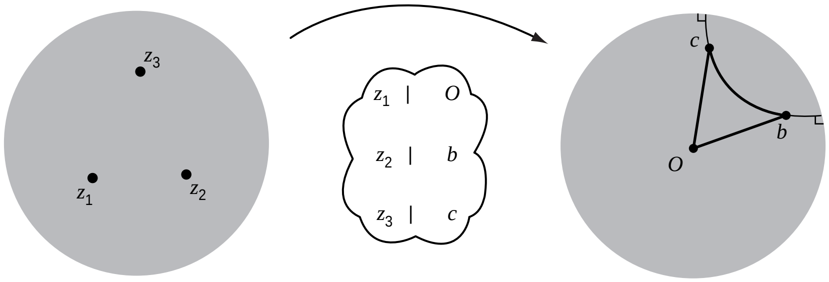

Taking O as centre we draw a hyperbolic circle C_2 through c to meet Ob at the point q; then $d(O,q) = d(O,c)$. Similarly, taking b as centre we draw a hyperbolic circle C_1 through c to meet Ob at the point p, where $d(b,p) = d(b,c)$.

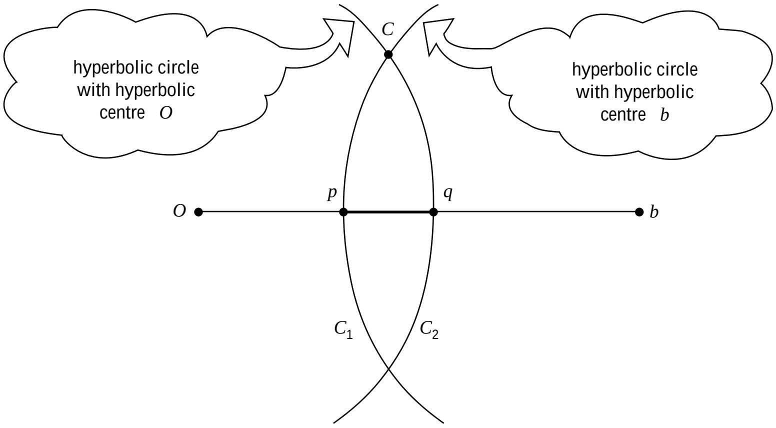
hyperbolic circle with hyperbolic centre O
c
hyperbolic circle with hyperbolic centre b
p
q
O
b
C_1
C_2

We have drawn this figure in the case that p and q lie between O and b. The argument also holds if q lies beyond b or p beyond O.

Since both circles are also Euclidean circles with centres on Ob, and since the circles intersect, it follows that

$$\begin{aligned} d(0,c) + d(c,b) &= d(0,q) + d(p,b) \\ & \quad d(0,q) + d(q,b) \\ &= d(0,b). \end{aligned}$$

Recall that hyperbolic circles are Euclidean circles, by Theorem 1.

Since M preserves hyperbolic distances, we deduce that

$$d(z_1,z_3) + d(z_3,z_2) \quad d(z_1,z_2),$$

as required.

Property 6, Subsection 6.3.1

We can deduce from the Triangle Inequality (Property 3) that in hyperbolic geometry the curve of shortest length or geodesic between two points of $\mathbb{D}$ is the segment of the d-line that joins them. Rather loosely, we can express this fact as follows: Distances are measured along d-lines in hyperbolic geometry.

We omit a formal definition of length of a curve , and so a proof of this claim.

6.3.4 Reflected Points in Hyperbolic Geometry

In Euclidean geometry the image A of a point A under reflection in a line is an equal distance from the line and on the opposite side of the line, so that is the perpendicular bisector of the line segment AA. We now prove that the same result also holds in hyperbolic geometry, where the distances and lines are hyperbolic ones.

To state the result, we need the concept of *hyperbolic line segment*, or *d-line segment*; this is just that part of the d-line through two points which lies between those points.

Theorem 3 Let A and A be points in the unit disc D that are images of each other under reflection in a d-line . Then is the hyperbolic perpendicular bisector of the hyperbolic line segment AA.

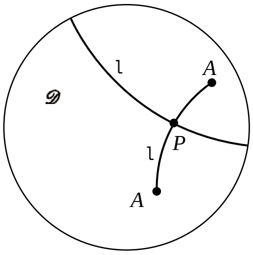

Proof Let the d-line through A and A meet at P.

Now A and A map to each other under hyperbolic reflection in , and P remains invariant. But hyperbolic reflection maps d-lines to d-lines, and there is exactly one d-line through A and A. Thus must map onto itself under the hyperbolic reflection in , and the hyperbolic line segment PA must map onto the hyperbolic line segment PA.

But hyperbolic reflection preserves angles and lengths. Since lengths are preserved, it follows that PA and PA must be of equal hyperbolic length. Also, since the angles that PA and PA make with the same part of at P map onto each other, they must be equal too; since they must add up to a total angle at P (because APA is a hyperbolic line segment through P), it follows that PA and PA both meet at right angles.

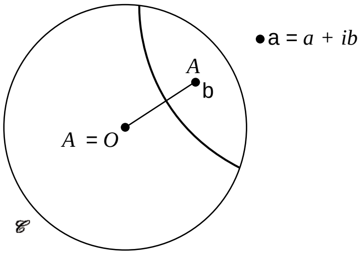

We already know one case where we can find the hyperbolic perpendicular bisector of a hyperbolic line segment AA. This is the case where one of the vertices, say A, coincides with O, the centre of the disc D. In that case, the reflection that sends A (represented by the complex number) to O is given by $M(z) = \frac{\text{ﬂ}z \;\; 1}{\text{ﬂ}z \;\; \text{ﬂ}}$, for some $= a + ib$, with $| \; | > 1$.

The equation of the d-line obtained from in which the reflection takes place is, as we saw earlier,

By Corollary 1, Subsection 6.1.1

$$x^2 + y^2 \quad 2ax \quad 2by + 1 = 0.$$

Since $M(\;) = 0$, it follows from the formula for M that $\text{ﬂ} \quad 1 = 0$.

More generally, suppose we want to find a hyperbolic reflection that exchanges the points p and q in D. Again, the hyperbolic reflection that does this must be of the form

$$M(z) = \frac{\text{ﬂ}z \quad 1}{\text{ﬂ}z \quad \text{ﬂ}}.$$

Subsection 6.2.1, Lemma 1

The condition $M(p) = q$ implies

$$q = \frac{\text{ﬂ}p \quad 1}{\text{ﬂ}p \quad \text{ﬂ}},$$

which is equivalent to

$$1 + \text{ﬂ}pq = \text{ﬂ}q + \quad \text{ﬂ}p. \tag{4}$$

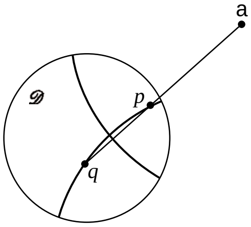

The condition $M(q) = p$ implies

$$p = \frac{\text{ﬂ}q \quad 1}{\text{ﬂ}q \quad \text{ﬂ}},$$

which is equivalent to

$$1 + \text{ﬂ}qp = \text{ﬂ}p + \quad \text{ﬂ}q. \tag{5}$$

Equations (4) and (5) can be solved for provided that $p\bar{p} \quad q\bar{q} = 0$. Indeed, if we subtract q times equation (5) from p times equation (4), and then divide by $p\bar{p} \quad q\bar{q}$, we obtain

$$= \frac{p \quad q + pq(\bar{p} \quad \bar{q})}{p\bar{p} \quad q\bar{q}}.$$

Notice that in the special case when $q = 0$, this formula gives $= 1/\bar{p}$, as it should.

If $p\bar{p} \quad q\bar{q} = 0$, then $|p| = |q|$, in which case the sought-for reflection is simply Euclidean reflection in the diameter of D which bisects the angle pOq.

We summarize this discussion in the following lemma.

Lemma 1 Reflection Lemma

Let p and q be points in the unit disc D. If $|p| = |q|$, then the hyperbolic reflection that maps p and q onto each other is given by

$$M(z) = \frac{\bar{z} \quad 1}{z \quad \bar{}}, \quad \text{where} \quad = \frac{p \quad q + pq(\bar{p} \quad \bar{q})}{p\bar{p} \quad q\bar{q}}.$$

The d-line in which this reflection takes place has equation

$$x^2 + y^2 \quad 2ax \quad 2by + 1 = 0, \quad \text{where} \quad = a + ib.$$

By Lemma 1, Subsection 6.1.1

Example 6 Determine the equation of the hyperbolic perpendicular bisector of [0.2, 0.9], the line segment from 0.2 to 0.9.

Solution Using the Reflection Lemma, with $p = 0.9$ and $q = 0.2$, we find that

$$= \frac{0.7 + 0.9 \bullet 0.2 \bullet 0.7}{0.81 \quad 0.04} \quad 1.0727.$$

So the equation of the d-line which is the (hyperbolic) perpendicular bisector of [0.2, 0.9] is

$$x^2 + y^2 \quad 2ax + 1 = 0, \quad \text{where } a \quad 1.0727.$$ □

Problem 6

(a) Determine the equation of the hyperbolic perpendicular bisector of the line segment $[0.5, 0.8]$.

(b) Deduce the equations of the hyperbolic perpendicular bisectors of the line segments $[\quad 0.8, \quad 0.5]$ and $[0.5i, 0.8i]$.

We often write $[z_1, z_2]$ to mean the Euclidean line segment in C with endpoints z_1, z_2.

Problem 7

(a) Let A be a point in D, a d-line through A, and $0 < \quad < \quad$. Prove that there are exactly two d-lines through A that make an angle with .

(b) Prove that, if ABC is a d-triangle, then there is a unique d-line that bisects BAC, and that reflection in maps the d-lines containing BA and CA onto each other.

6.3.5 Proofs

In Subsection 6.3.1 we asserted that the hyperbolic distance between two points is invariant under hyperbolic transformations. We now supply a proof of this fact.

Theorem 4 The formula for hyperbolic distance

$$d(z_1, z_2) = \tanh^{1} \frac{z_2 \quad z_1}{1 \quad \bar{z}_1 z_2}, \quad \text{where } z_1, z_2 \quad \mathsf{D},$$

satisfies Property 6; that is,

$$d(z_1, z_2) = d(M(z_1), M(z_2))$$

for all z_1 and z_2 in D and all direct hyperbolic transformations M in G_{D}.

See Subsection 6.3.1.

Proof First, we define the expression $R(z_1, z_2)$ to be

$$R(z_1, z_2) = \frac{z_2 \quad z_1}{1 \quad \bar{z}_1 z_2}, \quad \text{where } z_1, z_2 \quad \mathsf{D}. \tag{6}$$

It follows from the definition of $d(z_1, z_2)$ above that

$$d(z_1, z_2) = \tanh^{1}(R(z_1, z_2));$$

We omit the details of this, for brevity.

hence d possesses Property 6 if we can prove that

$$R(z_1, z_2) = R(M(z_1), M(z_2)) \tag{7}$$

for all $z_1, z_2 \quad \mathsf{D}$ and all $M \quad G_{\mathsf{D}}$.

So, let $z_1, z_2 \quad \mathsf{D}$ and $M \quad G_{\mathsf{D}}$. Then the direct hyperbolic transformation M_1 in G_{D} given by

$$M_1 : z \quad \frac{z \quad z_1}{1 \quad \overline{z_1} z}, \quad z \quad \mathsf{D},$$

maps z_1 to 0, and z_2 to $\frac{z_2 \quad z_1}{1 \quad \bar{z}_1 z_2}$.

Also, the direct hyperbolic transformation M_2 in G_{D} given by

$$M_2 : z \quad \frac{z \quad M(z_1)}{1 \quad \overline{M(z_1)} z}, \quad z \quad \mathsf{D},$$

where $M \quad G_{\mathsf{D}}$ so that $|M(z_1)| < 1$, maps $M(z_1)$ to 0, and $M(z_2)$ to $\frac{M(z_2) \quad M(z_1)}{1 \quad \overline{M(z_1)} M(z_2)}$.

Now, the composite mapping $M_2 \circ M \circ M_1^{-1}$ is also a transformation in $G_{\mathbb{D}}$. It maps $\mathbb{D}$ to $\mathbb{D}$, 0 to 0, and

$$\frac{z_2 - z_1}{1 - \overline{z}_1 z_2} \quad \text{to} \quad \frac{M(z_2) - M(z_1)}{1 - \overline{M(z_1)}M(z_2)}. \tag{8}$$

Note that M_1^{-1} maps 0 to z_1, and $\frac{z_2 - z_1}{1 - \overline{z}_1 z_2}$ to z_2.

But any direct transformation in $G_{\mathbb{D}}$ that maps 0 to 0 is simply a (Euclidean) rotation; so it follows from equation (8) that

$$\left|\frac{z_2 - z_1}{1 - \overline{z}_1 z_2}\right| = \left|\frac{M(z_2) - M(z_1)}{1 - \overline{M(z_1)}M(z_2)}\right|. \tag{9}$$

See Case 1 of Theorem 2, Subsection 6.2.1.

It follows from the definition of R that we can rewrite equation (9) in the desired form

$$R(z_1, z_2) = R(M(z_1), M(z_2)).$$

We defined R in equation (6).

Next, we stated earlier that the function d that we used to define the hyperbolic distance function in $\mathbb{D}$ was *essentially* the only *well-behaved* function with Properties 1-6 of Subsection 6.3.1. We now explain why this is so.

Theorem 5 Let $d(z_1, z_2)$ be any *well-behaved* function defined for all $z_1, z_2 \in \mathbb{D}$ that satisfies Properties 1-6 of Subsection 6.3.1. Then

$$d(z_1, z_2) = K \tanh^{-1}\left(\left|\frac{z_2 - z_1}{1 - \overline{z_1} z_2}\right|\right), \text{ for some } K > 0. \tag{10}$$

Proof Let $z_1, z_2 \in \mathbb{D}$. Since hyperbolic distances are invariant under Möbius transformations, $d(z_1, z_2)$ is unchanged if we map z_1 to the origin by the direct hyperbolic transformation (a Möbius transformation)

$$z \mapsto \frac{z - z_1}{1 - \overline{z_1} z}.$$

By Property 6

Next, hyperbolic distances are invariant under rotations of $\mathbb{D}$, since such rotations are hyperbolic transformations. So, we may also assume that z_2 lies on the positive real axis.

Hence, in order to prove the formula (10) it is sufficient to prove that

$$d(0, z) = K \tanh^{-1}(z), \text{ for some } K > 0 \text{ and all } z \in (0, 1). \tag{11}$$

For simplicity, we will now simply write $d(z)$ in place of $d(0, z)$ whenever the context means that no confusion will arise.

Let $0 < a, c < 1$. Then the direct hyperbolic transformation (a Möbius transformation)

$$M(z) = \frac{z + a}{1 + az}, \quad z \in \mathbb{D}$$

maps 0 to a, and c onto $b = \frac{c+a}{1+ac}$. Since M maps $[0, 1)$ one-one to $[a, 1)$, we must have $a < b < 1$.

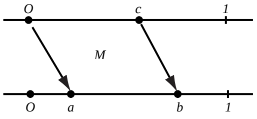

Now, hyperbolic distances are additive along a line, so that

$$d(b) = d(a) + d(a, b). \tag{12}$$

By Property 4

But, since hyperbolic distances are invariant under hyperbolic transformations, we have

By Property 6

$$\begin{aligned} d(a,b) &= d(0,c) \\ &= d(c), \end{aligned}$$

so that it follows from equation (12) that

In this equation, a and c are independent variables.

$$d\left(\frac{c+a}{1+ac}\right) = d(a) + d(c). \qquad (13)$$

Now we will assume that the function d is differentiable on $[0,1)$. If we then differentiate both sides of equation (13) with respect to the variable c, we get

This is where we use the hypothesis that the distance function is well-behaved.

$$d'\left(\frac{c+a}{1+ac}\right) \cdot \frac{(1+ac)\cdot 1 - (c+a)\cdot a}{(1+ac)^2} = 0 + d'(c);$$

and we may rewrite this equation in the form

$$d'\left(\frac{c+a}{1+ac}\right) = \frac{(1+ac)^2}{1-a^2} \cdot d'(c).$$

If we substitute 0 for c into this equation, we get

Here $d_R'(0)$ is the right derivative of d at 0.

$$\begin{aligned} d'(a) &= \frac{d_R'(0)}{1-a^2} \\ &= \frac{K}{1-a^2}, \text{ for some real } K. \end{aligned}$$

Since d must be an increasing function on $[0,1)$, we must have $K \geq 0$. But we cannot have $K = 0$, as it would then follow from the above equation that $d'(a) = 0$ for all $a \in [0,1)$ - so that $d(0,a)$ takes the same value (which would have to be 0) for all $a \in [0,1)$. Hence

This follows from Property 1.

$$d'(a) = \frac{K}{1-a^2}, \text{ for some } K > 0, \text{ and all } a \in (0,1).$$

If we then integrate both sides of this formula from 0 to z, we get

$$d(z) = K\tanh^{-1}(z), \text{ for some } K > 0, \text{ and all } z \in [0,1).$$

Remarks

1. For any (complex) point z near the origin, equation (10) gives
$$\begin{aligned} d(0,z) &= K\tanh^{-1}(|z|) \\ &\simeq K|z|, \end{aligned}$$
so it is natural to make the choice $K = 1$ for the definition of distance in hyperbolic geometry. Then $d(0,z) = \tanh^{-1}(|z|)$, for all $z \in \mathbb{D}$.
2. When $d(0,z) = \tanh^{-1}(|z|)$, $z \in \mathbb{D}$, we can reformulate the equation $d'(z) = \frac{1}{1-|z|^2}$ in the form
$$\mathrm{d}s = \frac{|\mathrm{d}z|}{1-|z|^2},$$
relating infinitesimal hyperbolic distances ds to infinitesimal Euclidean distances $|\mathrm{d}z|$ in $\mathbb{D}$ near any point $z \in \mathbb{D}$.

In advanced texts, the formula for hyperbolic distance is often used in this infinitesimal form.

6.4 Geometrical Theorems

6.4.1 Triangles

A **triangle** in hyperbolic geometry, or a ***d*-triangle,** consists of three points in the unit disc D that do not lie on a single *d*-line, together with the segments of the three *d*-lines joining them. One or two of the sides may be segments of diameters of D; but, in general, the sides are parts of Euclidean circles.

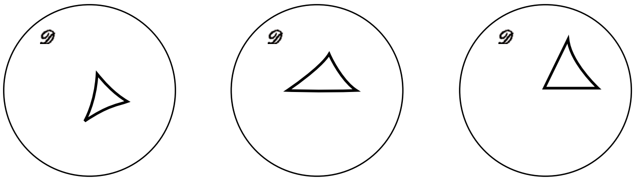

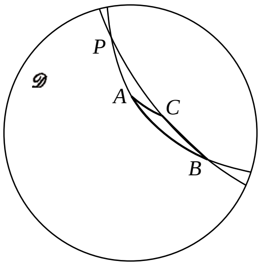

Notice that if ABC is a *d*-triangle, then the extended sides can meet only at A, B and C. For if the *d*-lines through *A* and *B* and through *B* and C meet at a point *P* distinct from *B*, then the two *d*-lines meet at *B* and *P*. But we saw earlier that there is a unique *d*-line through any two points of D; so it would follow that *AB* and *BC* are part of a single *d*-line, a possibility that we have already excluded. So the sides of a *d*-triangle cannot overlap .

Subsection 6.1.2,
Theorem 4

One of the basic results in Euclidean geometry is that the sum of the angles of a (Euclidean) triangle is . This is *not* true for *d*-triangles in hyperbolic geometry!

Theorem 1 The sum of the angles of a *d*-triangle is less than .

Proof By the Origin Lemma, we can map the *d*-triangle ABC onto a *d*-triangle OB C by any hyperbolic transformation that sends *A* to the origin *O*. Since hyperbolic transformations preserve angles, the sums of the angles of the two *d*-triangles are the same.

Subsection 6.1.2,
Lemma 2

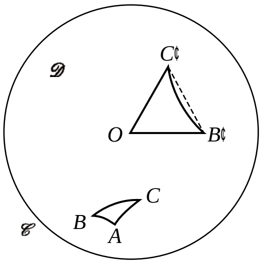

Then *OB* and *OC* are parts of Euclidean lines, and *B C* is part of a Euclidean circle that bends towards the origin. Thus the angles at *B* and *C* of the *d*-triangle OB C are less than the corresponding angles of the Euclidean triangle OB C , and the angles of both triangles at *O* are the same.

Since the sum of the angles of the Euclidean triangle OB C is , it follows that the sum of the angles of the *d*-triangle OB C is (strictly) less than . The result then follows.

Problem 1 Prove that each external angle of a *d*-triangle is greater than the sum of the opposite two internal angles. (See the diagram on the left below.)

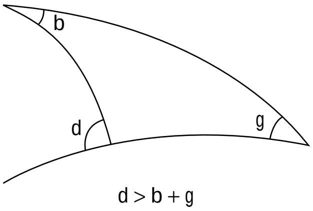

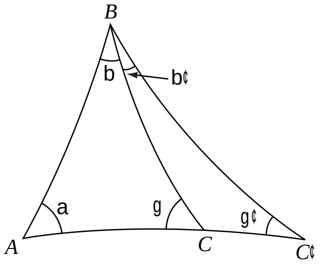

Problem 2 Let ABC and ABC be two hyperbolic triangles such that A, C and C lie on a d-line in this order. Let the angles at A, B, C in the first triangle and at A, B, C in the second triangle be , , and , + , , respectively. Show that + + > + + + . (See the diagram on the right, above.)

It will be seen from Problem 2 that the larger triangle, ABC, has the smaller angle sum. This is true in general, not only when one triangle fits neatly inside another. In fact, it can be shown that the area of a hyperbolic triangle with angles , and is proportional to (+ +), and we shall prove this in Subsection 6.5.1. You may for the moment use the following result as a useful memory aid:

> small triangles have angle sums close to (but less than) , and triangles with large areas have angle sums close to zero.

In Euclidean geometry, we can also prove that the sum of the angles of a (Euclidean) quadrilateral equals 2 . Now, a **quadrilateral** $ABCD$ in hyperbolic geometry, or a ***d*-quadrilateral,** consists of four points A, B, C, D in D (no three of which lie on a single d-line), together with the segments AB, BC, CD and DA of the four d-lines joining them. We require also that no two of these segments meet except at one of the points A, B, C or D.

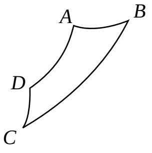

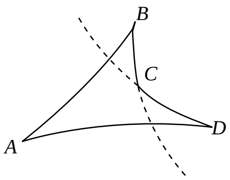

How large can the sum of the angles of a d-quadrilateral be?

Theorem 2 The sum of the angles of a d-quadrilateral is less than 2 .

Proof Any d-quadrilateral can be divided into two (non-overlapping) d-triangles by one or other of the d-lines joining alternate vertices.

We shall assume this fact without proof; see the figures below.

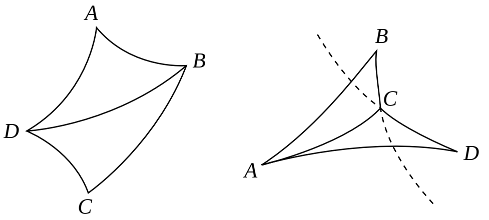

The angles of each d-triangle sum to less than π. The angles of the d-quadrilateral belong to one or other d-triangle, or partly to one and partly to the other. It follows that the angles of the d-quadrilateral sum to less than 2π.

We shall use this fact later.

In hyperbolic geometry there are many theorems about d-triangles that are analogues of the corresponding theorems about Euclidean triangles, such as the following.

Theorem 3 Let $\triangle ABC$ be a d-triangle in which $\angle ABC = \angle ACB$. Then the sides AB and AC are of equal length.

This is the hyperbolic analogue of the Euclidean result in Example 1, Subsection 2.1.1.

Proof Let D be the midpoint of the d-line segment BC. By applying the Origin Lemma, if necessary, we may assume that D coincides with O, the centre of the disc $\mathbb{D}$. (Although this is not strictly necessary for the proof, it simplifies the picture.) Then BC is part of a diameter of $\mathbb{D}$.

Let the d-line ℓ be the perpendicular bisector of BC; it is the diameter of $\mathbb{D}$ perpendicular to BC. Reflect (in both the Euclidean and hyperbolic senses) the triangle $\triangle ABC$ in the d-line ℓ. Because reflections preserve length, and $DB = DC$, it follows that B and C change places. Suppose that A moves to some point A'. Since reflection preserves angles, it follows that $\angle A'BC = \angle ACB$. Also, recall that we are given that $\angle ACB = \angle ABC$, so $\angle A'BC = \angle ABC$. But this can happen only if A' lies on the d-line through A and B. Similarly, $\angle A'CB = \angle ABC = \angle ACB$, so A' must also lie on the d-line through A and C. This means that A and A' must coincide. Hence the d-line segment BA reflects to the d-line segment CA, and vice versa; so these d-line segments have the same length.

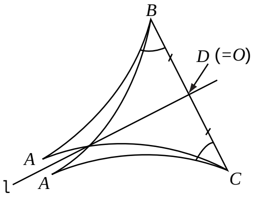

Problem 3 Let $\triangle ABC$ be a d-triangle in which the sides AB and AC have equal hyperbolic length. Prove that $\angle ABC = \angle ACB$.
Hint: Consider reflection in the d-line that bisects angle $\angle BAC$.

The notion of mapping one figure onto another *exactly* by a transformation in the geometry (for example, the reflection in the d-line ℓ in the proof of Theorem 3) is one that we met previously: namely, congruence.

Definition Two figures in the unit disc $\mathbb{D}$ are ***d*-congruent** if there is a hyperbolic transformation that maps one onto the other.

For example, if A, B, C and D are four points in $\mathbb{D}$ such that the hyperbolic distances $d(A, B)$ and $d(C, D)$ are equal, then the d-line segment AB is congruent to the d-line segment CD. For, as we saw earlier, there is a hyperbolic transformation t that maps A to C and the d-line through A and B onto the d-line through C and D, in such a way that D and $t(B)$ lie on the same side of C along $t(\ )$. Then, since B and D are the same hyperbolic distances from A and C, respectively, it follows that t must map B onto D. Thus AB and CD are congruent.

Subsection 6.1.2, Theorem 5, and the Remark following that theorem.

Now, since hyperbolic transformations preserve angles, it follows that if two d-triangles do not have corresponding angles equal (that is, they are not **similar),** then they certainly cannot be d-congruent.

Recall that two Euclidean figures are *similar* if one is a scale copy of the other; in particular, two Euclidean triangles are similar if they have corresponding angles equal.

However, the following result is still very surprising, because the analogous result is *false* in Euclidean geometry.

Theorem 4 Similar d-triangles are d-congruent.

Proof We have to prove that if the d-triangles ABC and PQR have the angles at A, B and C and the angles at P, Q and R equal, respectively, then the two d-triangles are d-congruent.

We may apply a hyperbolic transformation to map A to the origin O; this does not change the angles of the triangle. To avoid complicated notation, we shall still denote this image d-triangle by ABC.

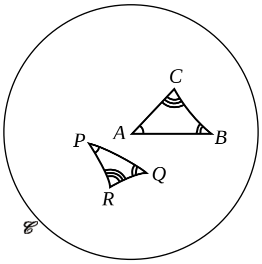

We can also apply a hyperbolic transformation to map P to the origin and the radius on which Q lies to the radius on which B lies. By reflecting in the d-line through O and B, if necessary, we can deduce from the fact that the angles at A and P are equal that the image of R lies on the same radius as (the image of) C. Again, we shall denote the image d-triangle still by the notation PQR, for simplicity.

The following figure shows the result of the transformations described above.

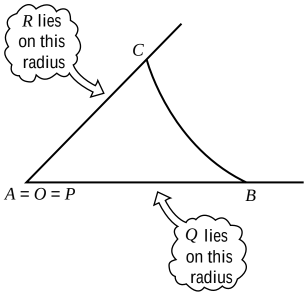

To prove the result, we have to show (in terms of the points obtained by the above preliminary mappings) that $B = Q$ and $C = R$. We proceed by considering the various possible situations that would arise if the result were false.

(a) *The d-line segment BC might lie between the origin and the d-line segment QR.* (The argument is similar if QR lies between O and BC.)

In this case,

$$RCB = OCB \quad \text{and} \quad CBQ = OBC,$$

so the angle sum of the d-quadrilateral $CBQR$ is

$$(\quad OCB) + (\quad OBC) + OQR + ORQ = 2\ ,$$

which is impossible, by Theorem 2.

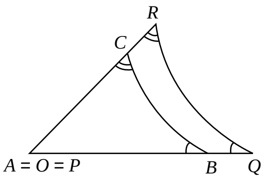

(b) *B and Q may coincide, but not C and R.* (The argument is similar if R lies between O and C, or if C and R coincide, but not B and Q.)

In this case, if C lies between O and R, the external angle at C of the d-triangle BCR is less than or equal to the sum of the opposite two internal angles, which is impossible by the result of Problem 1.

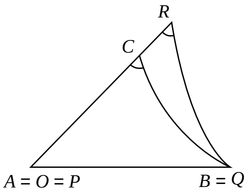

(c) *The d-line segments BC and QR may cross*.

Let the point of intersection of BC and QR be X. In this case, the external angle at B of the d-triangle XBQ is less than or equal to the sum of the opposite two internal angles, which is impossible by the result of Problem 1.

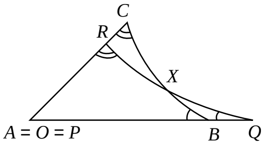

It follows that the only possibility is that $B = Q$ and $C = R$, as required.

In the following diagram, all the d-triangles have the same angles at their vertices, and so are similar. It follows from Theorem 4 that they are d-congruent to each other.

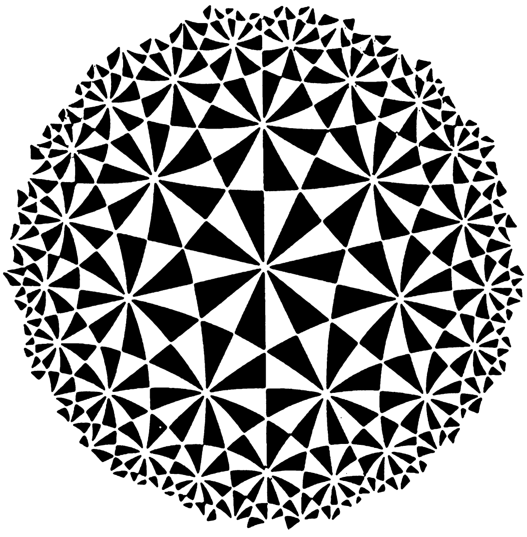

This looks very unlikely, since the triangles *seem* to be getting smaller as they move away from the origin towards C; however, it is a consequence of the way in which we defined both hyperbolic transformations and hyperbolic distance that all the triangles have sides of the same (hyperbolic) lengths as well as angles of the same sizes. This results from the fact that equal Euclidean distances on a ruler correspond to increasing hyperbolic distances as the ruler is moved outwards.

Asymptotic Triangles

The vertices of d-triangles lie in the unit disc D. However, it is often useful to talk about figures in D with three sides that are d-lines but where one or more of the Euclidean circles or lines of which they are part meet on C rather than in D. In this sense, we say that they are triangles with one or more vertices on C ; such triangles are called **asymptotic triangles**.

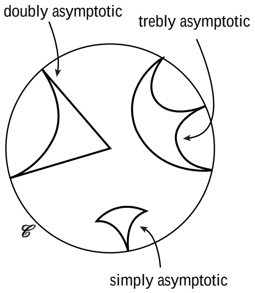

If an asymptotic triangle has one vertex on C is said to be **simply asymptotic**, if it has two vertices on C **doubly asymptotic**, and three vertices on C **trebly asymptotic**.

There is no hyperbolic transformation that maps points of C to points of D, or vice-versa; so asymptotic triangles are *essentially* different from ordinary d-triangles. However, it turns out that many of the results that hold for d-triangles hold also for asymptotic triangles, and that their proofs are similar.

Theorem 5 The angle sum of an asymptotic triangle is less than . The angle sum of a trebly asymptotic triangle is zero.

We omit the proof of this result. The proof of the first assertion is similar to that of Theorem 1.

Problem 4 Prove that two doubly asymptotic triangles are d-congruent if and only if they have the same angle at their vertex in D.

In hyperbolic geometry you should be careful not to assume that results are valid simply because they hold in Euclidean geometry. Sometimes asymptotic triangles are useful in constructing counter-examples.

We shall do exactly this at the end of Subsection 6.4.2.

6.4.2 Perpendicular Lines

In Euclidean geometry, two given lines and have a common perpendicular if and only if they are parallel to each other. In hyperbolic geometry, the situation is somewhat different, as you will see. First, we must define the term *common perpendicular* in hyperbolic geometry.

Definition Let and be two d-lines, and suppose that there exist points A on and A on such that the d-line segment AA meets and at right angles. Then AA is a **common perpendicular** to and .

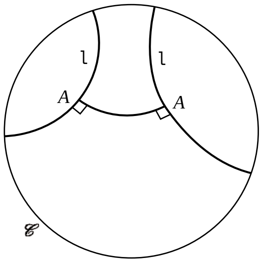

Theorem 6 Common Perpendiculars Theorem
Two d-lines have a common perpendicular if and only they are ultra-parallel. This common perpendicular is unique.

Proof First we show that if two d-lines and have a common perpendicular, then they are ultra-parallel. Let the common perpendicular be the d-line segment AA, where A is on and A is on . By the Origin Lemma, we can find a transformation r G_{D} which maps the point A to the origin, O. Then the d-line $r(\ )$ is a diameter of the disc D, as shown on the left below. The

d-line $r(\;)$ is part of a Euclidean circle whose centre R lies somewhere outside the disc D on the continuation of the diameter joining O to $r(A\;)$. The radius of this Euclidean circle is less than RO, and RO is perpendicular to $r(\;)$, so this circle cannot meet $r(\;)$, and hence $r(\;)$ and $r(\;)$ are ultra-parallel. It follows that the d-lines and are ultra-parallel.

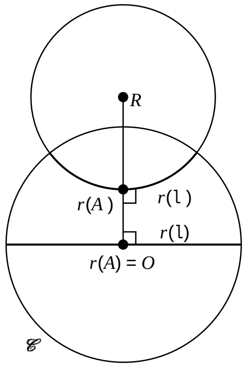

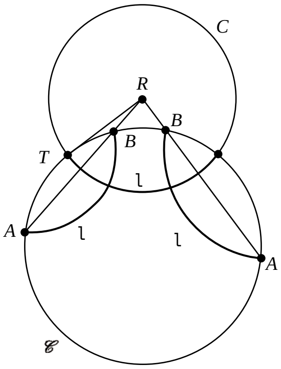

Secondly, we show that if two d-lines and are ultra-parallel, then they have a common perpendicular. As shown in the figure on the right above, let the boundary points of be A and B, and the boundary points of be A and B. Consider the *Euclidean* lines AB and $A\ B$.

If the Euclidean lines AB and $A\ B$ are not parallel, then they meet at a point R outside D, as shown on the right above. Let RT be a tangent from R to the boundary circle C. Consider the circle C with centre R and radius RT. This circle meets C at right angles, so the part of it in D is a d-line, say.

The Euclidean triangles RTB and RAT are similar, since $TRB = ART$ (being the same angle) and $RTB = RAT$ (since the exterior angle equals the interior opposite angle for a tangent and chord of a circle). Hence, in particular, $RB/RT = RT/RA$ so that $RA \bullet RB = RT^2$. In other words, A and B are inverse points with respect to the circle with centre R and radius RT.

Inversion in C exchanges A and B, and exchanges A and B. So it maps the d-line to itself, and the d-line to itself. It follows that intersects and at right angles, and so a common perpendicular to the d-lines and exists.

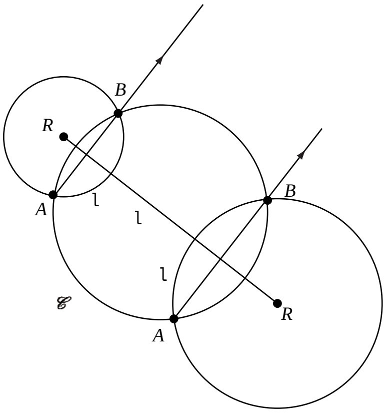

If the Euclidean lines *AB* and *A B* are parallel, then there is a diameter of D that is the perpendicular bisector of *AB* and *A B* . It therefore passes through the centres *R* and *R* of the Euclidean circles of which and are parts, and so it is perpendicular to both and . Since is a diameter of D it is a *d*-line, and the common perpendicular we seek exists.

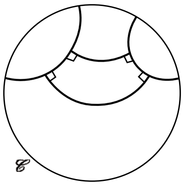

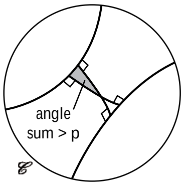

Finally, the common perpendicular to two ultra-parallel lines is unique. For, if there were two, and they were disjoint (as shown on the top illustration alongside), they and the two ultra-parallel lines would form four sides of a hyperbolic quadrilateral all of whose angles were right angles. But then the angle sum of such a quadrilateral would be 2 , which is impossible. Similarly, if there were two common perpendiculars that intersected each other (as shown in the lower illustration alongside), we should have a *d*-triangle whose angle sum was , which is impossible.

It follows that the common perpendicular to two ultra-parallel lines is unique.

In Euclidean geometry, the altitudes of a triangle play an interesting role; this is also the case in hyperbolic geometry.

Before we can talk about altitudes, we need to prove that through any point *P* not on a given *d*-line there is necessarily another (unique) *d*-line *p* such that *p* meets at right angles.

This fact is needed in order to prove that altitudes of a given *d*-triangle *exist*.

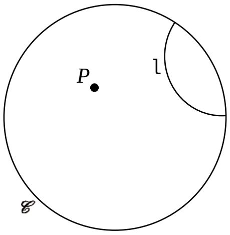

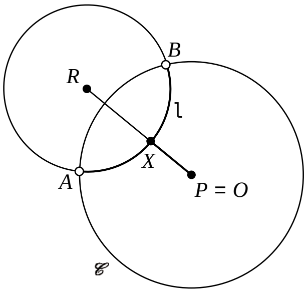

By a preliminary hyperbolic transformation, if necessary, we may assume that *P* is the origin. Let have boundary points *A* and *B*, and let *R* be the centre of the Euclidean circle of which is a part. Let the (Euclidean) line *OR* meet at X; *OR* and intersect at right angles at X, so that the d-line segment *OX* is a perpendicular from O to . The diameter which includes *OX* meets at right angles, as required. It is unique because no other line through O can meet at right angles.

Thus it is possible to drop a perpendicular from a point to a d-line, and so it makes sense to talk about *the* altitude through a vertex of a d-triangle.

Definition Let be a d-line which passes through one vertex A of a d-triangle *ABC* and which is perpendicular to the side *BC* at the point D. The d-line segment *AD* is an **altitude** of the triangle *ABC*.

Theorem 7 Altitude Theorem
Let the sides *AB* and *AC* of a d-triangle *ABC* be of equal hyperbolic length, and let the angle at A be . Then the hyperbolic length of the altitude through A of the triangle is less than some number that depends only on .

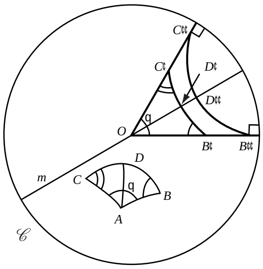

Proof Map the d-triangle *ABC* onto a d-triangle *OB C* by any hyperbolic transformation that sends A to the origin O. Since hyperbolic transformations preserve (hyperbolic) lengths and angles, *OB* and *OC* are of equal length and *B OC* = .

Let the diameter m of D that bisects the angle *B OC* meet the d-line segment *B C* at D. It is clear, by symmetry, that the d-line segment *OD* must be perpendicular to the d-line segment *B C*, so that *OD* (which is part of a Euclidean line) is the altitude of the d-triangle *OB C* through O. Since d-triangles have a unique altitude through each vertex, the altitude *AD* of the d-triangle *ABC* must map onto the altitude *OD*, and so they are of equal hyperbolic length.

Then the length of *OD* is less than the length of *OD*, where D lies on m and the d-line joining the boundary points B and C of the lines *OB* and *OC* (as shown). It follows that the hyperbolic length of the original altitude *AD* is less than the hyperbolic length of *OD*, which clearly depends only on .

Problem 5 Prove that the d-triangles *OB D* and *OC D* in the proof of Theorem 7 are d-congruent.

Problem 6 Determine a (numerical) upper bound for the hyperbolic length of the altitude *AD* of an isosceles d-triangle *ABC* in which *AB* and *AC* are of equal hyperbolic length and the angle at A is a right angle. *Hint:* This situation is a special case of that in Theorem 7: = /2, so B and C in the proof of that theorem are 1 and i, respectively.

In Euclidean geometry the altitudes of a triangle are concurrent; but this is not true in general in hyperbolic geometry. For example, let A, B and C be the points 1, 0 and $\frac{1+i}{2}$. Then the altitudes through A and C of the d-triangle ABC are the d-lines with endpoints 1, i and $\frac{1+i}{2}$, $\frac{1\ \ i}{2}$, respectively. These d-lines do not meet anywhere in D or on C. So certainly the altitudes of a doubly asymptotic triangle are not concurrent.

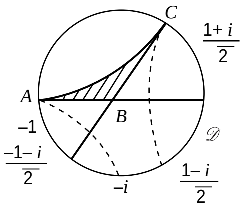

Finally, if we choose a (non-asymptotic) d-triangle with one vertex at O and the other two vertices in D but very close to A and C, it is clear that the altitudes of *this* d-triangle cannot be concurrent.

6.4.3 Right-Angled *d*-Triangles

In Euclidean geometry, Pythagoras Theorem plays a central role in many calculations. As we have seen, however, the theorem does not hold in hyperbolic geometry if we simply replace Euclidean lines by d-lines and Euclidean distances by hyperbolic distances in its statement.

Paragraph following Example 2 of Subsection 6.3.1.

However it seems reasonable to imagine that hyperbolic functions might play some role in a version of Pythagoras Theorem in hyperbolic geometry, and indeed this is the case.

Theorem 8 Pythagoras' Theorem
Let ABC be a d-triangle in which the angle at C is a right angle. If a, b and c are the hyperbolic lengths of BC, CA and AB, then

$$\cosh 2c = \cosh 2a \times \cosh 2b.$$

We prove Theorem 8 later in this subsection.

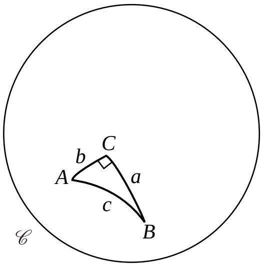

Problem 7 Use Pythagoras Theorem to determine the hyperbolic lengths of the sides of the d-triangle with vertices $0, (1+i)/2\ \ 2$ and $(\ \ 1+i)/5\ \ 2$, which has a right angle at O.

Problem 8 Use Pythagoras Theorem to determine the hyperbolic lengths of the sides of the d-triangles with vertices at the following points:

(a) $0, \frac{3}{4}$ and $\frac{3}{4}i$; (b) $0, r$ and ir (where $0 < r < 1$).

We can use these ideas to provide another solution to Problem 6 above.

Example 1 Prove that, if ABC is an isosceles d-triangle in which AB and AC are of equal hyperbolic lengths and the angle at A is a right angle, then the hyperbolic length of the altitude AD is less than $\frac{1}{2}\cosh^{1}\ \ 2\ \ \ 0.4407$ which is $\frac{1}{2}\log_e\ \ 1+\ \ 2$.

In Problem 6 we found that this upper bound $\frac{1}{2}\cosh^{1}\ \ 2$ was equal to $\tanh^{1}\ \ 2\ \ 1$; in fact the two numbers are equal. Such unexpected identities involving hyperbolic functions occur surprisingly often!

Solution The given triangle is d-congruent to the d-triangle OBC with vertices at $0, r$ and ir, for some r with $0 < r < 1$. Let OD be an altitude of this triangle. Let the hyperbolic lengths of the sides OB and OC be denoted by b, of side BC by a, and of the altitude OD by d.

It follows from Problem 5 (or by symmetry) that D is the Euclidean midpoint and the hyperbolic midpoint of the d-line segment BC, so that the d-line segments BD and CD both have hyperbolic length $\frac{1}{2}a$.

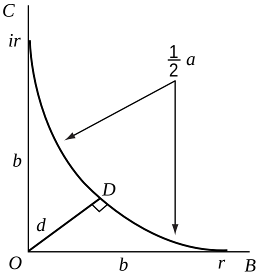

By applying Pythagoras' Theorem to the d-triangle $\triangle OBD$, we find that

$$\begin{aligned}\cosh 2b &= (\cosh 2d) \times \cosh 2\left(\tfrac{1}{2}a\right)\\ &= (\cosh 2d) \times \cosh a. \qquad (1)\end{aligned}$$

Next, by applying Pythagoras' Theorem to the d-triangle $\triangle OBC$, we find that

$$\cosh 2a = \cosh^2 2b. \qquad (2)$$

Now, we may rewrite equation (2) in the form

$$2\cosh^2 a - 1 = \cosh^2 2b; \qquad (3)$$

Here we use the identity $\cosh^2 2x = 2\cosh^2 x - 1$, for $x \in \mathbb{R}$.

and if we then substitute for $\cosh a$ from equation (1) into equation (3), we obtain

$$2\left(\frac{\cosh 2b}{\cosh 2d}\right)^2 - 1 = \cosh^2 2b.$$

We may rearrange this equation in the form

$$\left(\frac{\cosh 2d}{\cosh 2b}\right)^2 = \frac{2}{1 + \cosh^2 2b},$$

so that

$$\begin{aligned}\cosh 2d &= \frac{\sqrt{2}\cosh 2b}{\sqrt{(1 + \cosh^2 2b)}}\\ &= \frac{\sqrt{2}}{\sqrt{1 + \frac{1}{\cosh^2 2b}}}\\ &< \sqrt{2}.\end{aligned}$$

It follows that $d < \frac{1}{2}\cosh^{-1}\left(\sqrt{2}\right) \simeq 0.4407$, as claimed. □

Pythagoras' Theorem takes a simple, if unexpected, form in hyperbolic geometry. Moreover, the relationship closely approximates the Euclidean one if the sides of the triangle are very small, which is what you would expect because small hyperbolic triangles have angle sums nearly equal to π and so are nearly Euclidean themselves. To see the relationship, we expand the cosh terms as power series, using the formula

$$\cosh 2c = 1 + \frac{(2c)^2}{2!} + \text{ higher powers of } 2c.$$

If we ignore the higher powers of $2c$, as we may when $2c$ is very small, we find

$$\cosh 2c \simeq 1 + 2c^2$$

Recall that

$$\begin{aligned}\cosh x &= 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \ldots\\ &= \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!}\end{aligned}$$

for all $x \in \mathbb{R}$.

and

$$\cosh 2a \times \cosh 2b \qquad 1 + 2a^2 \quad 1 + 2b^2$$

$$= 1 + 2a^2 + 2b^2 + 4a^2b^2.$$

If we ignore the term $4a^2b^2$ on the grounds that it is as small as other terms we have already dropped, then we are left with

$$1 + 2c^2 \quad 1 + 2a^2 + 2b^2,$$

which reduces to

$$c^2 \quad a^2 + b^2.$$

So for small d-triangles the hyperbolic version of Pythagoras Theorem is essentially the same as the Euclidean version.

In Euclidean geometry there is a simple connection between the lengths of the sides of a right-angled triangle and its angles. An analogous result exists in hyperbolic geometry also.

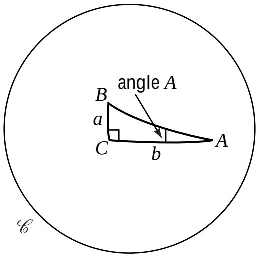

Theorem 9 Lobachevskii’s Formula
Let the d-triangle ABC have a right angle at C, and let the hyperbolic lengths of sides AC and BC be b and a, respectively. Then

$$\tan A = \frac{\tanh 2a}{\sinh 2b}.$$

We prove this result later in the subsection.

For example, let the points A, B, C be $\frac{1}{2}, \frac{1}{2}i$ and 0, respectively. Then it follows from Lobachevskii s Formula that

$$\tan A = \frac{\tanh 2d \quad 0, \frac{1}{2}i}{\sinh 2d \quad 0, \frac{1}{2}}$$

$$= \frac{\tanh 2d \quad 0, \frac{1}{2}}{\sinh 2d \quad 0, \frac{1}{2}}$$

$$= \frac{1}{\cosh 2d \quad 0, \frac{1}{2}}$$

$$\frac{1}{1.6667} \qquad 0.6,$$

so that $A \quad \tan^{\ 1} 0.6 \quad 0.5404$ radians.

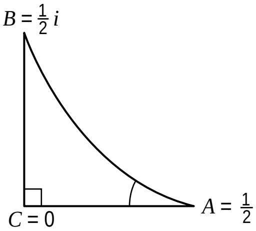

Remark

It can be shown that when the triangle is very small, and so nearly Euclidean, $\tanh 2a$ is approximately $2a$ and $\sinh 2b$ is approximately $2b$, so that $\tan A \quad a/b$, which is the Euclidean value.

Problem 9 Determine the angle at A in the d-triangle ABC with vertices $\frac{3}{5}i$, $\frac{4}{5}$ and 0, respectively.

We can also use these ideas to answer another question that loomed very large in the discovery of hyperbolic geometry.

Let be the d-line from a point P that is perpendicular to a d-line not through P, and let p be the length of the segment of from P to . What is the angle between and a d-line through P which is parallel to ?

The angle is called the angle of parallelism .

Let Q be the foot of the perpendicular from P to , and R a point on different from Q. As the point R moves away from Q towards C, the hyperbolic length of QR, $d(Q, R)$, tends to infinity; it follows that tanh $2d(Q, R)$ tends to 1 as R approaches C.

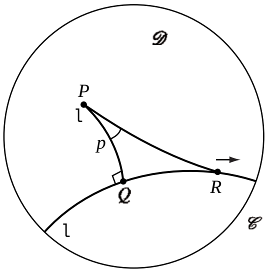

From Lobachevskii s Formula,

$$\tan \quad QPR = \frac{\tanh 2d(Q, R)}{\sinh 2p},$$

and so, if we let R approach C along , it follows that

$$\tan \quad = \frac{1}{\sinh 2p}.$$

We have proved the following result.

Corollary 1 Angle of Parallelism

Let be a d-line, and P a point of D that does not lie on . Then the angle between the perpendicular from P to (of hyperbolic length p) and either d-line through P that is parallel to is given by

$$\tan \quad = \frac{1}{\sinh 2p}.$$

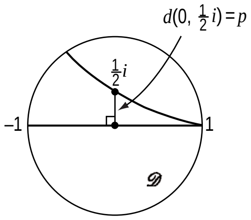

For example, the hyperbolic distance of the point $\frac{1}{2}i$ from the diameter $= (\quad 1, 1)$ is $\tanh^{1} 0.5 \quad 0.5493$. It follows that the angle between the perpendicular from $\frac{1}{2}i$ to and either parallel to through $\frac{1}{2}i$ is given by

$$\tan \quad = \frac{1}{\sinh\ 2\tanh^{1} 0.5}$$
$$\frac{1}{\sinh 1.0986};$$

hence, 0.6435 radians.

In fact, the remarkable formula

$$\sinh\ 2\tanh^{1} x \quad = \frac{2x}{1 \quad x^2}$$

gives tan $= \frac{3}{4}$ exactly.

Problem 10 Determine the angle between the perpendicular from the point $\frac{3}{4}$ to the d-line with equation $x = 0$ and either parallel to through $\frac{3}{4}$.

Finally we give, without proof, another useful formula.

Theorem 10 Sine Formula
Let ABC be a d-triangle right-angled at C. Let a and c be the hyperbolic lengths of BC and AB. Then

$$\sin A = \frac{\sinh 2a}{\sinh 2c}.$$

A proof of this result will be found as an Exercise in Section 6.7.

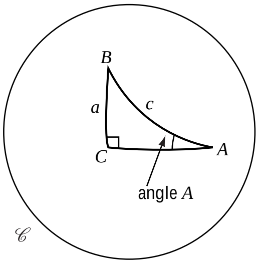

Remark

This result agrees with the expected one when the triangle is small, for in such cases $\sinh 2a \quad 2a$ and $\sinh 2c \quad 2c$, so $\sin A \quad a/c$, which is the Euclidean result.

Problem 11 Use the Sine Formula to find the length a of the sides of a hyperbolic equilateral triangle ABC with angle , as a function of .
Hint: Apply the Sine Formula to the triangle obtained by dropping the perpendicular AD from A to BC.

Proofs

We now supply proofs of two results that you met earlier in this subsection.

The formulas in Theorems 8 and 9 depend on the following remarkable formula:

$$\cosh \quad 2\tanh \quad ^{1} x \quad = \frac{1+x^2}{1 \quad x^2}. \tag{4}$$

The substitution $x = \tanh t$ shows that equation (4) holds if and only if

$$\cosh 2t = \frac{1+\tanh^2 t}{1 \quad \tanh^2 t},$$

and this formula follows from the definitions of cosh and tanh, as you can check.

For future convenience, we now list various expressions that we have now met for the inverse tanh function.

For $x \quad (\quad 1, 1)$, the following are all equal to $\tanh \quad ^{1} x$:

$$\frac{1}{2}\log_e \frac{1+x}{1 \quad x}, \ \frac{1}{2}\sinh \quad ^{1} \frac{2x}{1 \quad x^2}, \ \frac{1}{2}\cosh \quad ^{1} \frac{1+x^2}{1 \quad x^2}.$$

Theorem 8 Pythagoras' Theorem
Let ABC be a d-triangle in which the angle at C is a right angle. If a, b and c are the hyperbolic lengths of BC, CA and AB, then

$$\cosh 2c = \cosh 2a \times \cosh 2b.$$

Proof To prove Pythagoras Theorem in hyperbolic geometry, by the Origin Lemma we may assume that C is at the centre of D, the point A is at the point a on the horizontal diameter and the point B is at ib on the vertical diameter, where a and b are real and positive.

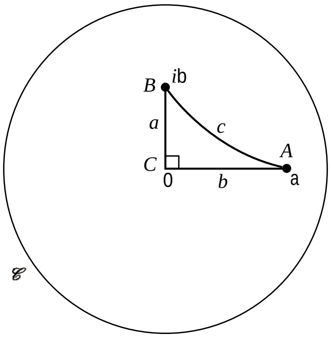

Here a, b, c are real because they are the hyperbolic lengths of the sides of the d-triangle ABC.

From the Distance Formula, it follows that

$$a = \tanh^{1} b \quad \text{and} \quad b = \tanh^{1} a . \tag{5}$$

We now map D to itself by the hyperbolic transformation

$$M(z) = \frac{z \quad a}{1 \quad a z}.$$

Under M, A goes to the origin (which we shall also call A, for clarity), B goes to the point B, with complex coordinates

$$b = \frac{ib \quad a}{1 \quad ia b},$$

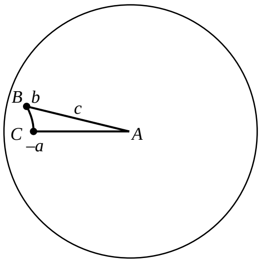

and C goes to the point C, with coordinates a.

Because the hyperbolic transformation preserves lengths, the hyperbolic length c of AB is equal to the hyperbolic length of $A B$.

To find c, the hyperbolic length of AB, we first find the modulus of b. We calculate its square:

$$|b|^2 = \frac{|ib \quad a|^2}{|1 \quad ia b|^2} = \frac{a^2 + b^2}{1 + a^2 b^2}. \tag{6}$$

Now we observe from formula (4) and equation (5) above that

$$\cosh 2a = \cosh \quad 2 \tanh^{1} b \quad = \frac{1 + b^2}{1 \quad b^2},$$

$$\cosh 2b = \cosh \quad 2 \tanh^{1} a \quad = \frac{1 + a^2}{1 \quad a^2};$$

also, since $c = \tanh^{1} |b|$, we get from the formula (4) that

$$\cosh 2c = \cosh \quad 2 \tanh^{1} |b| \quad = \frac{1 + |b|^2}{1 \quad |b|^2}.$$

But, it then follows from equation (6) that

$$\cosh 2c = \frac{1+|b|^2}{1-|b|^2} = \frac{1+\frac{a^2+b^2}{1+a^2b^2}}{1-\frac{a^2+b^2}{1+a^2b^2}}$$

$$= \frac{1+a^2+b^2+a^2b^2}{1-a^2-b^2+a^2b^2}$$

$$= \frac{1+b^2}{1-b^2} \times \frac{1+a^2}{1-a^2},$$

$$= \cosh\left(2\tanh^{-1} b\right) \times \cosh\left(2\tanh^{-1} a\right)$$

$$= \cosh 2a \times \cosh 2b.$$

This proves the analogue of Pythagoras' Theorem in hyperbolic geometry.

Theorem 9 Lobachevskii's Formula
Let the d-triangle ABC have a right angle at C, and let the hyperbolic lengths of sides AC and BC be b and a, respectively. Then

$$\tan A = \frac{\tanh 2a}{\sinh 2b}.$$

Proof We proceed as in the proof of Theorem 8, by mapping the d-triangle ABC with vertices at a, ib and 0, respectively, by the hyperbolic transformation $M(z) = \frac{z-a}{1-az}$ onto the d-triangle with vertices at 0,

$$b = \frac{ib-a}{1-iab} = \frac{-a\left(1+b^2\right)+ib\left(1-a^2\right)}{1+a^2b^2} \qquad (7)$$

We obtain the second fraction in equation (7) by multiplying the numerator and denominator of the first fraction by $1 + iab$.

and $-a$, respectively, where $b = \tanh^{-1} a$ and $a = \tanh^{-1} b$, so that

$$a = \tanh b \quad \text{and} \quad b = \tanh a.$$

The mapping also sends the angle A in ABC onto an angle of equal size (which we also denote by A by a slight abuse of notation) at the origin. Since the triangle with vertices at $0, b$ and $-a$ is right-angled at $-a$, it follows from equation (7) that

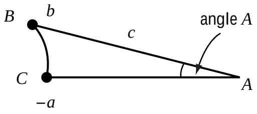

$$\tan A = \frac{b\left(1-a^2\right)}{a\left(1+b^2\right)}. \qquad (8)$$

Now, since $a = \tanh b$, we can write

$$\frac{1\quad a^2}{a} = \frac{1\quad \tanh^2 b}{\tanh b}$$

$$= \frac{\cosh^2 b\quad \sinh^2 b}{\sinh b \bullet \cosh b} = \frac{2}{\sinh 2b}.$$

Here we use the identities

$$\cosh^2 x\quad \sinh^2 x = 1$$

and

$$\sinh 2x = 2\sinh x \cosh x$$

for x R.

Similarly, we can verify that

$$\frac{b}{1+b^2} = \tfrac{1}{2}\tanh 2a.$$

Hence, if we substitute these expressions for $\frac{1\ \ a^2}{a}$ and $\frac{b}{1+b^2}$ into equation (8), we find that

$$\tan A = \frac{\tanh 2a}{\sinh 2b}.$$

6.4.4 Equidistant Curves to *d*-Lines

In the Euclidean plane R^2, the set of points at a fixed distance from a given line is a pair of straight lines parallel to . In the hyperbolic disc D there are two curves equidistant from a given *d*-line, though neither is itself a *d*-line.

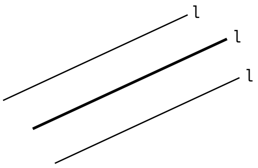

Theorem 11 Let be a *d*-line that ends at points *A* and *B* on C, and let be that part in D of a (Euclidean) circle through *A* and *B*, and let be the angle (in D) between and . Then all points on are the same hyperbolic distance *d* from , where

$$d = \tfrac{1}{2}\log_e \quad \tan \quad \frac{\ }{2} + \frac{\ }{4} \quad .$$

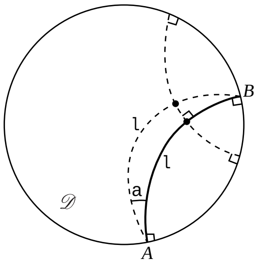

Proof First, map C and D to themselves by a Möbius transformation *M* that maps *A* and *B* to 1 and 1 (not necessarily respectively) and maps to a curve in the upper half of D. Since Möbius transformations preserve angles and generalized circles, it follows that the image of the *d*-line *AB* is the diameter [1, 1], and the image of is an arc of a circle in the upper half of D through −1 making an angle (in D) with [1, 1].

It is not hard to verify that such an *M* exists.

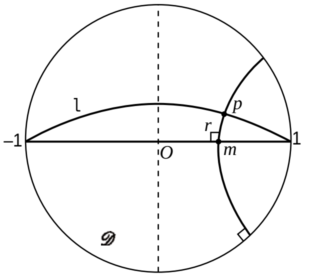

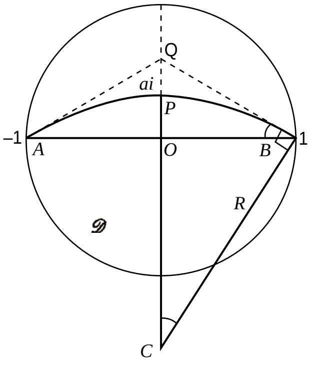

In view of the above remarks, it is sufficient to assume that the d-line is the diameter $(\quad 1, 1)$ of D. Let p be a point on , and m the foot of the perpendicular r from p to $(\quad 1, 1)$. Let M_1 be the Möbius transformation defined by

Similarly, we shall refer below to rather than to $M(\quad)$.

$$M_1(z) = \frac{z \quad m}{1 \quad mz}.$$

Then M_1 maps D to itself, and $(\quad 1, 1)$ to itself. It maps m to 0, to an arc of a (Euclidean) circle, with radius R and centre C (say), through -1; and it maps r to a portion of a d-line through O that is perpendicular to $(\quad 1, 1)$ and so to a portion of the imaginary axis. It follows that the point $P = M_1(p)$ is of the form ai, for some real number $a \quad (\quad 1, 1)$. Since M_1 is a hyperbolic transformation it does not alter hyperbolic distances, so that

Property 6, Subsection 6.3.1

$$\begin{aligned} d(p, m) = d(0, ai) &= d(0, a) \\ &= d, \text{ say.} \end{aligned}$$

Now, if Q is the point where the perpendicular to CB at B meets OC, the triangles CBQ and BOQ are similar, since $CBQ = \frac{}{2} = BOQ$ and $BCQ = OQB$. It follows that $BCQ = OBQ =$. Also, from OBC we see that $R \sin \quad = OB = 1$. Hence

Here we use the identities

$$\cos 2x = 1 \quad 2\sin^2 x$$

and

$$\sin 2x = 2 \sin x \cos x$$

for x R.

$$\begin{aligned} a &= R \quad R\cos \\ &= \frac{1 \quad \cos}{\sin} \\ &= \frac{2\sin^2 \frac{}{2}}{2\sin \frac{}{2} \cos \frac{}{2}} = \tan \frac{}{2} \; . \end{aligned}$$

It follows that

Recall from equation (3) of Subsection 6.3.1 that

$$\tanh^{\;1} k = \tfrac{1}{2} \log_e \frac{1 + k}{1 \quad k}$$

for any $k \quad 0$.

$$\begin{aligned} d = \tanh^{\;1} a &= \tanh^{\;1} \tan \frac{}{2} \\ &= \tfrac{1}{2} \log_e \frac{1 + \tan \frac{}{2}}{1 \quad \tan \frac{}{2}} \\ &= \tfrac{1}{2} \log_e \tan \frac{}{2} + \frac{}{4} \quad . \end{aligned}$$

This completes the proof of the theorem.

For example, the portion in the upper half of D of the Euclidean circle through -1 that makes an angle of $/4$ with the segment $[\quad 1, 1]$ is a set of points equidistant in D from $[\quad 1, 1]$, with the distance of separation being

$$d = \frac{1}{2} \log_e \tan \frac{}{8} + \frac{}{4} = \frac{1}{2} \log_e \tan \frac{3}{8} \qquad 0.4407.$$

6.5 Area

6.5.1 Area of a *d*-triangle

In the Euclidean plane $\mathbb{R}^2$, the area of a set is defined in terms of the union of areas of small rectangles. In the hyperbolic disc $\mathbb{D}$ the situation is somewhat different; *d*-quadrilaterals cannot have four right angles and so do not conveniently exhaust areas in an analogous way.

Theorem 2, Subsection 6.4.1

We want our definition of area in hyperbolic geometry to have such properties as the following:

The area of a *d*-triangle is non-negative, and is zero only if its Euclidean area is zero;
d-congruent *d*-triangles (and *d*-congruent asymptotic *d*-triangles) should have the same area;
if one *d*-triangle can be fitted inside another, it should have smaller area;
area should be additive.

Then we can determine the area of any figure in $\mathbb{D}$ that can be divided up into *d*-triangles.

Throughout this subsection, all lines and line segments will be *d*-lines and *d*-line segments, but to avoid constant repetition we will simply drop the *d* .

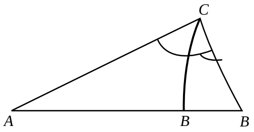

Notice first that as *d*-triangles grow in size their angle sum decreases. For example, let *ABC* be contained in *AB C*, where *A*, *B* and *B* are collinear, as shown. Denote by , , and , , the angles of the two *d*-triangles, as shown. Consider the *d*-triangle *BCB* . The angles at *B* and *C* in this are and . Then, since the sum of the angles of a *d*-triangle is less than , it follows that

Theorem 1, Subsection 6.4.1

$$(\quad) + (\quad) + \quad < \quad ,$$

so that $+ \; > \; +$; hence $+ \; + \; > \; + \; +$, as asserted.

Problem 1 Prove that a trebly asymptotic *d*-triangle can be divided into two doubly asymptotic right-angled *d*-triangles.
Hint: Consider a new vertex sliding along one edge; how do the angles vary?

Now, since very small *d*-triangles are *d*-congruent to small *d*-triangles near the origin, where *d*-lines are approximately Euclidean lines, the sum of the angles of very small *d*-triangles is close to .

At the opposite end of the size scale, we can use the result of Problem 1 above to prove the surprising fact that all trebly asymptotic *d*-triangles are *d*-congruent to each other.

Theorem 1 below

Theorem 1 All trebly asymptotic *d*-triangles are *d*-congruent to each other.

Proof It is sufficient to prove that trebly asymptotic d-triangles are d-congruent to the trebly asymptotic d-triangle with vertices -1, 1 and i.

So, let $\triangle ABC$ be any trebly asymptotic d-triangle, where (for convenience) we assume that the vertices A, B, C occur in clockwise order on $\mathbb{C}$. Let ℓ denote the line AB, and let D be any point of ℓ in $\mathbb{D}$.

Then, by the Origin Lemma, we can map $\mathbb{D}$ to itself and D onto the origin O by a Möbius transformation, M_1 say. It follows that M_1 maps AB onto a diameter of $\mathbb{D}$.

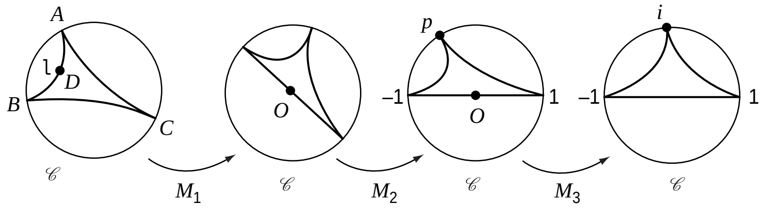

We can then map $\mathbb{D}$ to itself by another Möbius transformation, M_2 say, that rotates the disc about the origin and maps A and B onto -1 and 1, respectively. Since Möbius transformations are direct hyperbolic transformations, they preserve the orientation of points; it follows that the image of C under $M_2 \circ M_1$ is some point p, say, on the upper half of $\mathbb{C}$.

Now, for any real number $m \in (-1, 1)$, the Möbius transformation

$$M_3(z) = \frac{z - m}{1 - mz}$$

maps $\mathbb{D}$ to $\mathbb{D}$, $\mathbb{C}$ to $\mathbb{C}$, -1 to -1, 1 to 1, and the upper halves of $\mathbb{D}$ and $\mathbb{C}$ to themselves. It also maps p to i if m is chosen so that

$$i = \frac{p - m}{1 - mp},$$

which we can rewrite in the form

$$m = \frac{p - i}{1 - pi}. \tag{1}$$

Then, using the facts that $|p| = 1$ and $\operatorname{Im} p > 0$, we can check that equation (1) indeed provides a value of m that is real and strictly between -1 and 1.

We omit the details.

Composing the three Möbius transformations that we have met, the composite Möbius transformation $M_3 \circ M_2 \circ M_1$ maps the trebly asymptotic d-triangle $\triangle ABC$ onto the trebly asymptotic d-triangle with vertices -1, 1 and i, as required.

Trebly asymptotic d-triangles will play a crucial role in our discussion of hyperbolic area. First, we see that every d-triangle can be fitted inside some trebly asymptotic d-triangle.

Theorem 2 Let ABC be a d-triangle. Then there exists a trebly asymptotic d-triangle DEF that contains ABC.

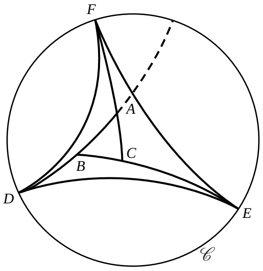

Proof First, we construct a trebly asymptotic d-triangle DEF as follows: extend the segment AB beyond B to meet C at D, the segment BC beyond C to meet C at E, and the segment CA beyond A to meet C at F. Then draw the trebly asymptotic d-triangle DEF.

Now, the d-line containing the points A, B, D cuts off a minor arc of C that contains F. It follows that the d-line through D and F lies on that side of the d-line through A, B, D, and (in particular) that the d-triangles DAF and ABC do not overlap.

Similarly, the d-triangles EBD and ABC do not overlap, and also the d-triangles FCE and ABC do not overlap. Also, none of the d-triangles DAF, EBD, FCE overlap.

It follows that ABC is contained within DEF, as required.

Then, since every d-triangle can be fitted inside some trebly asymptotic d-triangle, the following completely surprising result (which holds whatever our definition of hyperbolic area is) proves the existence of an upper bound for the area of all d-triangles.

Theorem 3 All trebly asymptotic d-triangles have the same (hyperbolic) area, and this is a finite number.

Note that the area must be finite, whatever definition of hyperbolic area we choose to use.

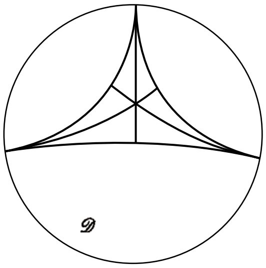

Proof Using the result of Problem 1, we can divide a given trebly asymptotic d-triangle into six singly asymptotic d-triangles, as shown in the margin diagram. Therefore it is sufficient to prove that the area of a singly asymptotic d-triangle is finite.

Let ABC be a singly-asymptotic d-triangle, with C on C (as in the left figure below). First, extend AB to meet C at a point C , and bisect BAC by AD, where D is a point on CC . Next, (hyperbolically) reflect BAC in AD; since AB has finite hyperbolic length and AC has infinite hyperbolic length (since C lies on C), this reflection maps B to some point A_1, say, on AC, as shown in the right diagram below. Also, the reflection must map BC onto A_1C ; and, by symmetry, BC and A_1C must meet at some point, B_1 say, on AD.

Next, bisect C BC by BD_0, where D_0 is on C C, and bisect C A_1C by A_1D_1, where D_1 is on C C, as shown below.

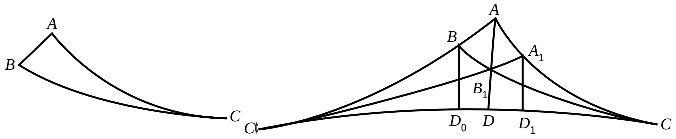

Our strategy of proof is now as follows. The d-pentagon $AA_1D_1D_0B$ and its boundary all lie in D, and so at a minimum positive distance from C, the boundary of D. Also, by adding additional line segments B_1D_0 and B_1D_1 we can divide the d-pentagon into eight d-triangles; each d-triangle has finite area, and so the d-pentagon must also have finite area. We will show that BAC can be cut up and reassembled (iteratively) inside this d-pentagon, and so has an area less than or equal to the area of the d-pentagon; in particular, it must have finite area.

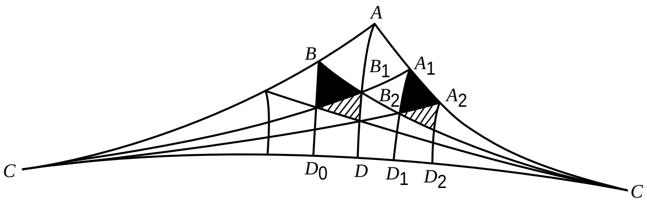

We now do to B_1A_1C what we did previously to BAC. Reflect (in a hyperbolic way) B_1A_1C in A_1D_1 : B_1 maps to some point A_2 on A_1C and B_1C maps to A_2C , which meets B_1C at some point B_2 on A_1D_1.

$B_2A_1A_2$ is congruent to several d-triangles in the d-pentagon; for convenience we choose the black one on the left in the figure above. We then repeat the process for B_2A_2C. We bisect B_2A_2C with A_2D_2, where D_2 is some point on $C C$. The cross-hatched d-triangles in the above figure are d-congruent one is in BAC and one is in the d-pentagon.

Continuing indefinitely, we do for $B_{n+1}A_{n+1}C$ what we have just done for B_nA_nC. With each pair of hyperbolic reflections we get two new d-triangles in BAC and two new d-triangles in the d-pentagon. These d-triangles are d-congruent by means of a sequence of hyperbolic reflections in the d-lines $A_{n+1}D_{n+1}, A_nD_n, \ldots, A_1D_1$ and AD.

In this way, we reassemble the asymptotic d-triangle *inside* the d-pentagon; hence the asymptotic d-triangle must have finite area.

It follows that all trebly asymptotic d-triangles have finite area, and so (as we noted before the theorem that all trebly asymptotic d-triangles are d-congruent) they have the same area.

Problem 2

(a) Prove that a doubly asymptotic d-triangle is a subset of a suitably chosen trebly asymptotic d-triangle, and so is finite.

(b) Prove that the area of a doubly asymptotic d-triangle depends only on the angle at its vertex in D.

We can now give Gauss magnificent proof that the area of each d-triangle is proportional to the quantity

$$\qquad \text{(angle sum)}.$$

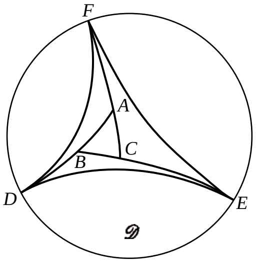

Problem 2, part (b)

Theorem 4 The area of a d-triangle with angles , , is given by $K($ $(\ + \ + \ ))$, where K is the same constant for all d-triangles.

Proof Let ABC be a d-triangle with angles , , at vertices A, B, C, respectively. Extend the line segment AB to a point D on C, the segment BC to a point E on C, and the segment CA to a point F on C. Since DEF is trebly asymptotic, it has finite area, k say, by Theorem 3.

Now, we have seen that any doubly asymptotic d-triangle has an area which depends only on the angle at its vertex in D, and so only on the exterior angle of the triangle at its vertex in D. So, we define a function f such that the area of any doubly asymptotic d-triangle with exterior angle , say, in D is $f(\ )$.

Then, it is clear from the margin figure above that in FAD this exterior angle equals ; so, the (hyperbolic) area of FAD is $f(\ )$.

If the segment FAC extended meets C at X (as in the left figure below), then DXF is trebly asymptotic and is composed of the two doubly asymptotic d-triangles FAD and DAX. Then, if we denote the area of each trebly-asymptotic d-triangle (such as DXF) by k, say, and notice that (from our definition of f) the area of DAX is $f(\qquad)$ since its exterior angle at A is , it follows that

$$k = f(\ ) + f(\qquad). \qquad (2)$$

Here we use the fact that (hyperbolic) area is additive.

Our next objective is to use equation (2) to determine an explicit formula for f.

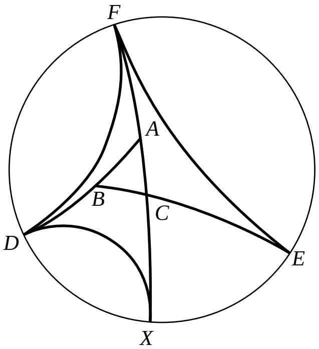

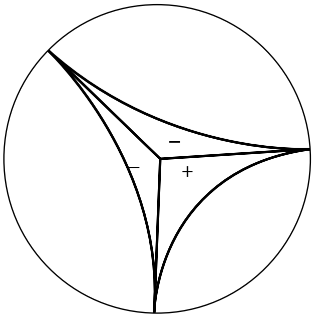

So, consider any trebly asymptotic d-triangle, as shown in the figure on the right above. By drawing radii from the center of D to its vertices on C, we can divide this d-triangle into three doubly asymptotic d-triangles. If we denote the angles at O of these three d-triangles by , and + , as shown above, we have

$$k = f(\ ) + f(\ ) + f(\qquad (\ + \ ))$$

By equation (2), $f(\quad (\ +\ )) = k \quad f(\ +\ )$, so that

$$f(\ ) + f(\ ) + k \quad f(\ +\ ) = k,$$

that is,

$$f(\ ) + f(\ ) = f(\ +\ ). \qquad (3)$$

It follows from equation (3) that f is linear, so that $f(\ ) = \quad$ for some constant .

We omit the details of this fact, which uses a continuity argument.

Using this formula for f, we deduce from equation (2) that

$$(\ ) + \ (\qquad) = k,$$

so that $= \frac{k}{\ }$. Hence $f(\ ) = \frac{k}{\ }$.

Returning to the original d-triangle ABC in the margin figure above, the area of ABC is equal to the area of DEF minus the sum of the areas of the three doubly asymptotic d-triangles FAD, DBE and CAF. Since the latter four areas are k, $f(\ )$, $f(\ )$ and $f(\ )$, respectively, it follows that

We again use the fact that (hyperbolic) area is additive.

$$\text{area of } ABC = k \quad (f(\ ) + f(\ ) + f(\ )).$$

Using the formula for f that we have just found, we can write this equation as

$$\text{area of } ABC = k \quad \frac{k}{\ } + \frac{k}{\ } + \frac{k}{\ } \quad ,$$

and so as

$$\text{area of } ABC = K(\quad (\ +\ +\ )),$$

where $K = k$.

Note that here K is arbitrary.

It is usual to take $K = 1$ (which is equivalent to taking the area of a trebly asymptotic d-triangle to be). Then

$$\text{area of } ABC = \quad (\ +\ +\ ).$$

The expression $(\ +\ +\ )$ is called the *angular defect* of ABC.

Now this result is all very well, but how do we know whether, given any positive real numbers , , with $+\ +\ <$, there actually is a d-triangle whose angles are , , ?

Recall that, if , , are positive real numbers with $+\ +\ =$, then such a Euclidean triangle exists.

A key tool in showing that such a d-triangle exists, is the following result.

Lemma 1 Let P be a point on a d-line of D, and let $0 < \quad < \frac{}{2}$. Then there are exactly two d-lines through P that make an acute angle with , and each is a (hyperbolic) reflection of the other in .

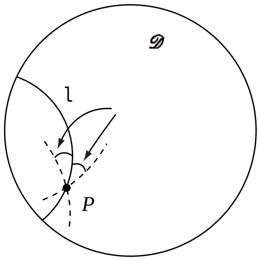

Proof Since all the properties in the statement of the Lemma are hyperbolic properties, and there is a hyperbolic transformation (for example, a Möbius transformation) that maps the point P to the origin O, it is sufficient to prove the result when P is the origin.

Then there are (by Euclidean geometry) exactly two diameters of D (and diameters are *d*-lines) through the origin that make an acute angle with the given diameter , and each is a reflection of the other in . This completes the proof.

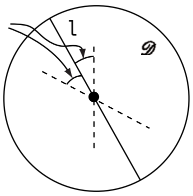

We are now in a position to prove the final result in this subsection.

Theorem 5 Let , , be positive real numbers with $0 < \; + \; + \; < \;$. Then there is a *d*-triangle in D with angles , , .

Proof Since $0 < \; + \; + \; < \;$, at least one of the numbers lies strictly between 0 and $\frac{}{2}$; let us suppose that this number is . Next, let *A*, *P*, *Q* be the points in D with coordinates 0, 1, cos + *i* sin , and let be a variable *d*-line in D whose initial position is as the *d*-line with endpoints *P* and *Q*. Then the angles of the doubly-asymptotic *d*-triangle *APQ* are , 0 and 0.

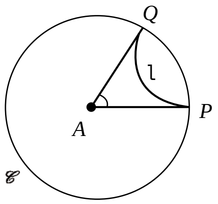

The first step is to keep fixed the endpoint of that is originally at *Q*, and let the endpoint that is originally at *P* move round C away from *Q*. As this endpoint moves away from *P*, the angle between and the radius *AP* varies continuously, starting at 0. By the time that the moving endpoint reaches *Q* , the image of *Q* under Euclidean reflection in the radius *AP*, the angle has reached the value $\frac{}{2}$. It follows, by continuity and the intermediate value property, that there is some point, *R* say, on *AP* such that for the *d*-line through *P* and *R* the corresponding value of the angle is .

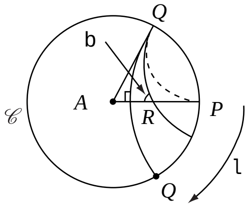

Now, it follows from Lemma 1 that through every point *S* between *A* and *R* there is a *d*-line that makes an angle with the radius *AP*; also, if we assume that this angle lies above the radius *AP* and contains part of the second quadrant of D in its interior, this *d*-line is unique.

We now let *S* move along *AP* from *R* towards *A*, taking with it the corresponding *d*-line . Denote by *T* the point where the variable *d*-line intersects the radius *AQ*. Then, as *S* moves along *AP* from *R* towards *A*, *T* moves along *AQ* from *Q* towards *A*.

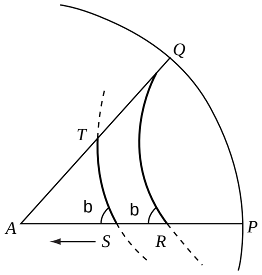

The *d*-lines $_1$ and $_2$ corresponding to successive positions S_1 and S_2 of *S* do not meet in D. For, if they met at a point *X* in D, then the external angle at S_2 of the *d*-triangle XS_1S_2 would be less than or equal to the sum of the opposite two internal angles, which is impossible by the result of Problem 1 in Subsection 6.4.1. (The following figure shows the situation when *X* is in the upper/lower half of D, respectively.)

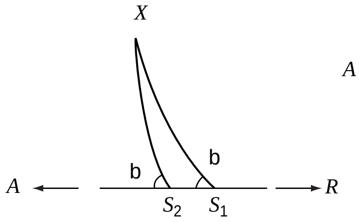

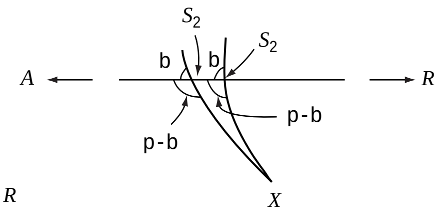

Hence as S moves monotonically along AP from R towards A, T moves monotonically along AQ from Q towards A. As S approaches A, the curvature of the d-line decreases, and approaches the diameter d of D that makes an angle with the diameter containing AP.

As T moves monotonically along AQ from Q towards A, the angle at T between the radial line AT and starts at 0, and approaches the value (+) as T approaches A. It follows, by continuity and the intermediate value property, that there is some point, C say, on AQ such that at C the d-line makes an angle with the radius through C.

If we denote by B the point where the d-line through C meets AP, then the triangle ABC has the desired angles , , .

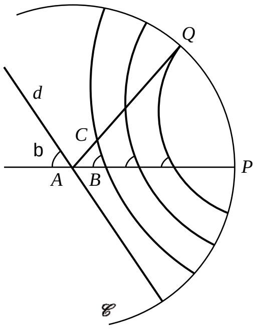

6.5.2 Tessellations

The term *tessellate* means fit together exactly a number of identical shapes, leaving no spaces , as in the making of a mosaic. For example, we may tessellate the Euclidean plane using rectangles of a given size in several ways, and we may tessellate it in other ways as well as illustrated below. In these examples, we may move a given occurrence of any shape onto any other occurrence by a Euclidean transformation, and so by a combination of Euclidean reflections.

You saw in Theorem 2 of Subsection 5.2.1 that every Euclidean transformation is a composite of reflections.

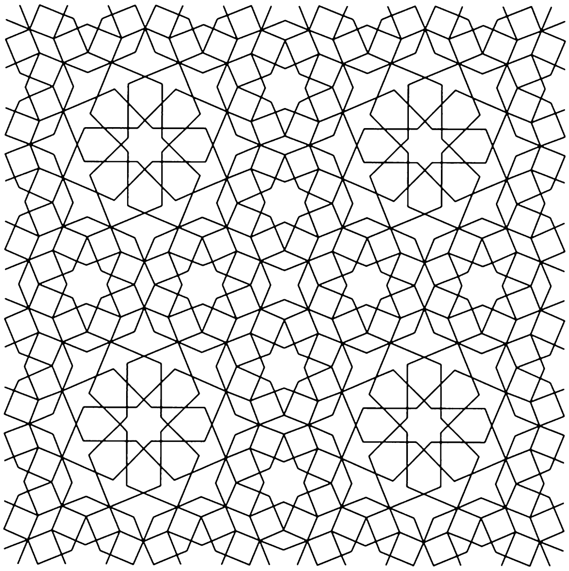

Euclidean tessellations using a single figure are sometimes called wallpaper patterns , for obvious reasons! With each such figure we can associate its group of symmetries; it is an interesting piece of mathematics that there are only 17 different wallpaper patterns, in the sense that the symmetry group of any wallpaper pattern is isomorphic to one of only 17 non-isomorphic groups.

Patterns like those in the figure above are found among the tilings of the Alhambra Palace in Granada (Spain), and elsewhere in Moorish art.

Tessellations occur in hyperbolic geometry also. For example, we may start with a symmetric figure, a symmetric d-triangle ABC, in D bounded by three d-lines (as in the following figure), reflect ABC in each d-line in turn to obtain 3 copies each d-congruent to ABC, and so on. This produces a hyperbolic tessellation of D, as shown in the figure below; for clarity we have shaded all the copies that can be obtained from ABC by direct hyperbolic transformations. All the copies are d-congruent to each other, even though to a Euclidean eye they seem to be of various different shapes and sizes!

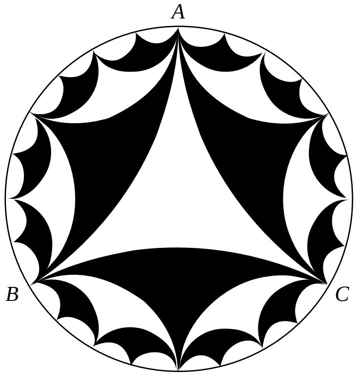

Another tessellation of D was suggested to the graphic artist M. C. Escher by the geometer H. S. M. Coxeter, and resulted in the picture below, *Circle Limit IV*. The angels and devils rest on a tessellation of D by d-triangles.

Harold Scott MacDonald Coxeter (1907–2003) was born in London, guided towards mathematics by Bertrand Russell, and spent most of his life in Canada. His work on regular polytopes led to the discovery of Carbon 60, and to the famous geodesic dome of Buckminster Fuller.

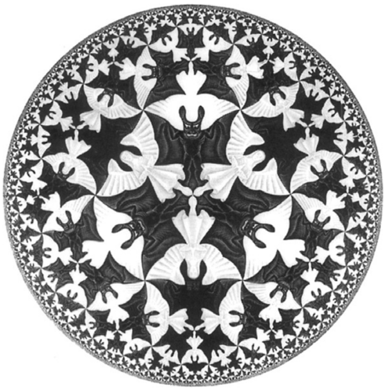

Maurits Cornelis Escher (1898–1972) received his graphic art training in Arnhem and Haarlem, and worked in Italy, Switzerland, Belgium and Holland.

M. C. Escher's Circle Limit IV c 2010. The M. C. Escher Company-Holland. All rights reserved. www.mcescher.com

6.5.3 Kaleidoscopes

We now explore the ideas underlying the (Euclidean) kaleidoscope, and explain how hyperbolic tessellations can be constructed.

A kaleidoscope is a machine that produces wallpaper patterns using reflections in mirrors.

A Euclidean Kaleidoscope

If two mirrors meet at right angles, an angle of $/2$, any point between the mirrors has an image in each mirror; then, each image itself has an image in the mirrors giving a total of four points in all, lying on a circle centred at the vertex where the mirrors meet.

If the mirrors meet at an angle of $/3$, there are 6 points in all; and, in general, if the mirrors meet at an angle $/k$ there are $2k$ points in all.

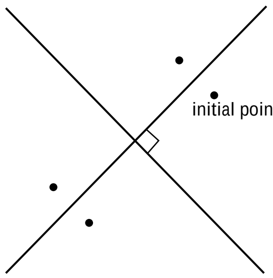

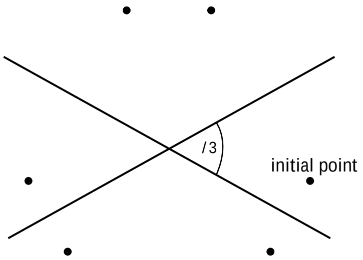

A real kaleidoscope has three mirrors, each pair meeting at an angle $/k$, where k is some integer. For example, consider the case where the mirrors meet at angles of $/2$, $/3$ and $/6$. Corresponding to these vertices there are 4 points, 6 points and 12 points making a total (in the first instance) of 17 points. But each of these points in turn has images in the mirrors, and so on . . .. The whole pattern of images spreads through the plane!

The total number is 17 rather than 22, since some of them are repeated.

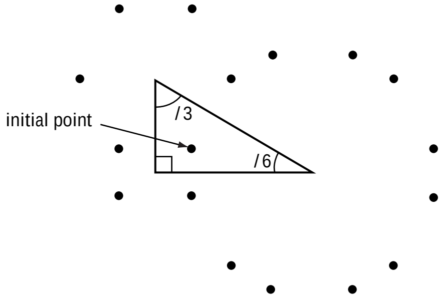

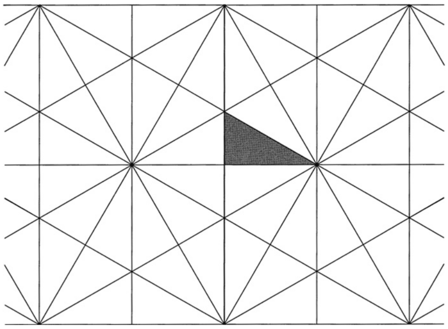

Since reflection is a Euclidean isometry, the process of reflections fills the whole plane with congruent shapes.

A ‘Hyperbolic Kaleidoscope’

Hyperbolic reflections and transformations are hyperbolic isometries; that is, they map figures without altering the hyperbolic distances between their points even though the images may look (to a Euclidean eye) very different in size and shape from the originals. In particular, they map d-triangles to d-congruent d-triangles.

If two mirrors along d-lines meet at an angle of $/2$, any point between the mirrors has an image under hyperbolic reflection in each mirror; then, each image itself has an image in the mirrors. Again, we obtain a total of four points in all, lying on a hyperbolic circle centred at the vertex.

Similarly, if the mirrors along d-lines meet at an angle of $/3$, there are 6 points in all; and, in general, if the mirrors meet at an angle of $/k$, there are $2k$ points in all.

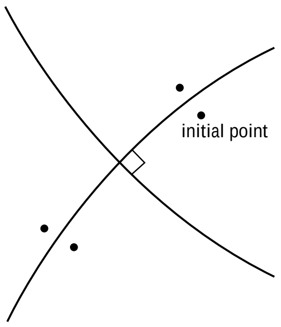

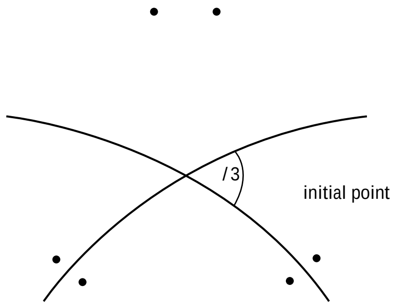

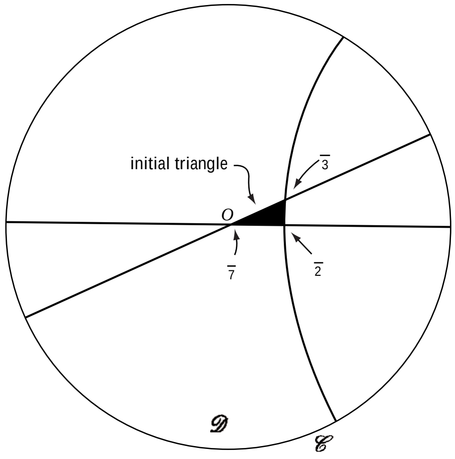

Now construct a hyperbolic analogue of the Euclidean kaleidoscope, with three mirrors along d-lines in D: two of them lying along d-lines through the origin that meet at O at an angle of $\pi/7$, and the third lying along a d-line that meets the first two at angles of $\pi/2$ and $\pi/3$. We reflect the triangle in each side in turn (using hyperbolic reflections), and continue to reflect indefinitely in the three d-lines.

It is not hard to verify that such a d-line must exist.

Note that $\frac{\pi}{7} + \frac{\pi}{2} + \frac{\pi}{3} < \pi$.

Since hyperbolic reflection is a hyperbolic isometry, the process of reflections fills the whole unit disc with d-congruent d-triangles. The figure that we end up with is just the design on the front of the book and in Subsection 6.4.1.

6.6 Hyperbolic Geometry: the Half-Plane Model

In previous sections we described the disc model of hyperbolic geometry, in which the unit disc $\mathbb{D}$ is the space of points, and the transformations of this space as being of the form

$$z \mapsto M(z) = \frac{az + b}{\bar{b}z + \bar{a}} \quad \text{or} \quad z \mapsto M(\bar{z}) = \frac{a\bar{z} + b}{\bar{b}\bar{z} + \bar{a}}, \quad \text{where } |b| < |a|.$$

However there is a completely *equivalent* hyperbolic geometry, in which the space (the set of points) is the upper half-plane

$$H = \{z : z \in \mathbb{C}, \quad \operatorname{Im} z > 0\},$$

That is, there is a one–one correspondence between points of the two spaces and the transformations in the two geometries.

and the h-lines in H (the analogues of d-lines in $\mathbb{D}$) are the restriction to H of vertical lines in the plane together with the restriction to H of circles with centre on the real axis in $\mathbb{C}$. Notice that the h-lines are all arcs of generalized circles that meet the boundary of H at right angles, just as d-lines are all arcs of generalized circles that meet the boundary $\mathbb{C}$ of $\mathbb{D}$ at right angles.

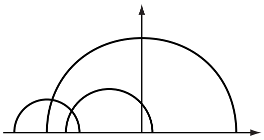

The Möbius transformation $z \mapsto i\frac{1+z}{1-z}$ maps $\mathbb{D}$ one–one onto H, and provides a convenient method for 'transferring' the geometric properties in $\mathbb{D}$ to corresponding properties in H. For example, it maps the d-lines in $\mathbb{D}$ to the h-lines in H.

The hyperbolic transformations in H are the mappings of H to itself given by

$$z \mapsto M(z) = \frac{az + b}{cz + d} \quad \text{or} \quad z \mapsto M(-\bar{z}) = \frac{a(-\bar{z}) + b}{c(-\bar{z}) + d},$$

where a, b, c and d are real, and $ad - bc > 0$. These are the composites of reflections in h-lines, which are just inversions in the Euclidean circles of which the h-lines form a part.

The geometry in H has many properties similar to those in $\mathbb{D}$:

the sum of the angles of an h-triangle is less than π;
two h-triangles are h-congruent if their angles are equal in pairs;
the area of an h-triangle with angles α, β, γ is $K(\pi - (\alpha + \beta + \gamma))$, for some $K > 0$;
h-circles are Euclidean circles in H, and vice-versa.

Generally we take $K = 1$, as in Subsection 6.5.1.

Infinitesimal hyperbolic distances ds in H are related to their Euclidean lengths by the formula $ds = \frac{|dz|}{y}$, from which it can be shown that geodesics in H are the h-lines.

Recall that a *geodesic* is the curve of shortest length joining two points in the space.

Finally, we comment that the hyperbolic geometry of H turns out to be of great importance in many unexpected branches of mathematics such as Number Theory, in which tessellations of H such as the following occur naturally.

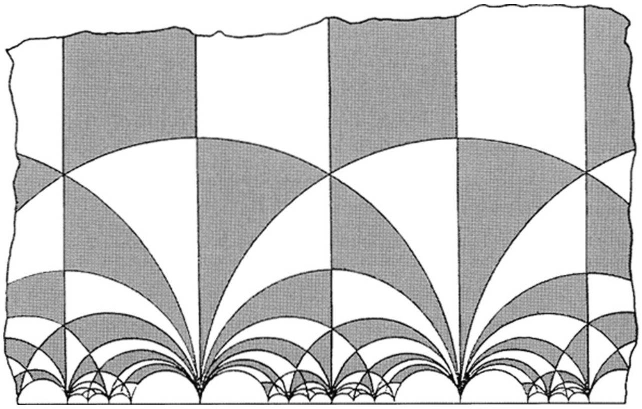

6.7 Exercises

Section 6.1

1. Sketch the following d-lines:

$$_1 = \{(x, y) : x + y = 0\} \quad D;$$

$$_2 = \{(x, y) : x^2 + y^2 + 2\ \overline{2}y + 1 = 0\} \quad D;$$

$$_3 = \{(x, y) : x^2 + y^2 \quad \overline{5}y + 1 = 0\} \quad D;$$

$$_4 = \{(x, y) : x^2 + y^2 \quad 3x + 1 = 0\} \quad D.$$

Hence decide which of the d-lines intersect, which are parallel and which are ultra-parallel.

2. Sketch three d-lines $_1$, $_2$ and $_3$ with the following properties: $_1$ and $_3$ are parallel, $_2$ intersects $_1$ and $_3$, and the angles at which $_2$ crosses $_1$ and $_3$ are not equal.

(This shows that the *Corresponding Angles Theorem* in Euclidean geometry does not hold in hyperbolic geometry.)

3. Verify that the points $\frac{1}{\overline{2}}, \frac{1}{2}$ and $\frac{1}{\overline{2}}, \frac{1}{2}$ lie on the circle C with equation

$$x^2 + y^2 + 2\ \overline{2}x + 1 = 0.$$

Hence write down the equations of the diameters of $\mathbb{D}$ which are parallel to the d-line which is part of C.

4. Use the Origin Lemma to show that, given two d-lines meeting at right angles, there is a hyperbolic reflection mapping them to two perpendicular diameters of $\mathbb{D}$.
5. Let $_1$, $_2$ and $_3$ be any three d-lines such that hyperbolic reflection in any one exchanges the other two.
   (a) Prove that there is some point P of $\mathbb{D}$ which lies on each of the three d-lines.
   (b) Use the Origin Lemma to prove that the acute angle between any two of the d-lines is $\frac{}{3}$.
6. Verify that the following equations are the equations of generalized circles of which d-lines are a part, and sketch the corresponding d-lines.
   (a) $x^2 + y^2 + 4x + 1 = 0$
   (b) $x^2 + y^2 + 2x \quad 2y + 1 = 0$
7. Use the fact that the circle with radius a and centre $(1, a)$, where $a > 0$, touches the x-axis to write down the equation of a d-line through the point $\frac{1}{2}i$ that is parallel to the d-line with equation $y = 0$.

Section 6.2

1. Determine the general form of the direct hyperbolic transformations that map 0 to 0 and the diameter $y = x$ of $\mathbb{D}$ to the diameter $y = 0$ of $\mathbb{D}$.
2. (a) Determine the general form of the direct hyperbolic transformations that map each of the following points to 0:

$$\tfrac{1}{2}, \quad \tfrac{1}{2} \quad \tfrac{1}{2}i.$$

   (b) Determine a direct hyperbolic transformation that maps $\tfrac{1}{2}$ to $\tfrac{1}{2} \quad \tfrac{1}{2}i$.
   (c) Determine the general form of the direct hyperbolic transformation that maps $\tfrac{1}{2} \quad \tfrac{1}{2}i$ to $\tfrac{1}{2}$.
3. By drawing a sketch, or in some other way, decide whether the effect of successive hyperbolic reflections in d-lines $_1$ and then $_2$ given by the following equations is a (hyperbolic) rotation, a limit rotation or a translation.
   (a) $_1 : x^2 + y^2 \quad 2x \quad 2y + 1 = 0;$ $_2 : x^2 + y^2 + 2x + 2y + 1 = 0.$
   (b) $_1 : x^2 + y^2 \quad 2x \quad 2y + 1 = 0;$ $_2 : x^2 + y^2 \quad 2x + 2y + 1 = 0.$
4. Let M be a direct hyperbolic transformation mapping the point p in $\mathbb{D}$ to the point q. Let M_q be a direct hyperbolic transformation mapping the point q to 0.
   (a) Show that the composite transformation $M = M_q \quad M$ maps p to 0.
   (b) Deduce that M is the composite of a hyperbolic transformation mapping p to 0 followed by a hyperbolic transformation mapping 0 to q.

5. (a) Prove that the set of all direct hyperbolic transformations forms a group under the operation of composition of functions.
   (b) Prove that the set of all hyperbolic transformations forms a group under the operation of composition of functions.

Section 6.3

1. Determine $d \quad \frac{1}{5}i, 0$ and $d(0.9, 0.99)$.
2. Determine the positive number a for which $d(0, a) = 2$, and the negative number b for which $d(0, b) = 10$.
3. Determined $d \quad \frac{1}{2}, \frac{1}{2}i$.
4. (a) Determine the hyperbolic midpoints of the d-line segments $[0.1, 0.9]$ and $[\quad 0.1, 0.9]$.
   (b) Determine the hyperbolic centre and radius of the circle $C = \{z : |z \quad 0.5| = 0.4\}$.
   (c) Determine the equation of the hyperbolic perpendicular bisector of $[0.1, 0.9]$.
5. Determine the Euclidean centre and radius of the hyperbolic circle $C = \quad z : d \quad \frac{1}{2}, z \quad = 1$.
6. Let $0 < a < b < 1$, and let the d-line with equation $x^2 + y^2 \quad 2cx + 1 = 0$, where $c > 1$, intersect the radius $[0, 1)$ between $A(a, 0)$ and $B(b, 0)$.
   (a) Prove that intersects AB at the point $c \quad \overline{c^2 \quad 1}, 0$.

   Now assume that is the hyperbolic perpendicular bisector of AB.

   (b) Use the Reflection Lemma to prove that

$$c = \frac{1 + ab}{a + b}.$$

   (c) Deduce from parts (a) and (b) that the hyperbolic midpoint of AB is

$$\frac{1 + ab \qquad \overline{(1 \quad a^2)(1 \quad b^2)}}{a + b}, 0 \quad .$$

7. Let be a d-line that is not a diameter of D and whose endpoints on C subtend an angle , $0 < \quad <$, at the centre of D. Prove that the hyperbolic length of the d-line segment from the origin to that is perpendicular to is $\tanh^{\quad 1} \quad \cot \frac{1}{4}(\quad + \quad)$.

Section 6.4

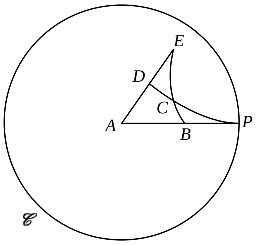

1. $ABCD$ is a d-quadrilateral. The d-lines through A and B and through D and C are parallel, with a common boundary point P on C; the d-lines through D and A and through C and B meet at a point E in D, as in the figure in the margin (in which A is at the origin).
   (a) Prove that it is impossible that $EAB = \quad EDC$.
   Hint: Consider the Euclidean triangle ADP.

(b) Prove that it is impossible that $EBA = ECD$.
Hint: Consider the Euclidean triangle PBC, with B at the centre of D.

(c) Prove that the following combination is impossible:

$$EAB = ECD \quad \text{and} \quad EBA = EDC.$$

(d) Prove that the following combination is impossible:

$$EAB = EBA \quad \text{and} \quad EDC = ECD.$$

2. Let $_1$ and $_2$ be a pair of perpendicular diameters of D. Sketch $_1$ and $_2$, and some of the ultra-parallels to $_1$ for which $_2$ is their common perpendicular with $_1$.
3. $ABCD$ is a d-quadrilateral such that the hyperbolic length of each side is 1. The d-line segments AC and BD are of equal hyperbolic length and intersect at right angles at the origin O, their common midpoint, and OE is an altitude of the d-triangle OAB. Also, P is the hyperbolic midpoint of OE.

   Determine the hyperbolic length of OE, the Euclidean length of OP, the angle ABC, the hyperbolic length of OB, the Euclidean length of BD, and the angle between PE and either d-line through P that is parallel to the d-line through A and B.

ABCD is a hyperbolic square of side-length 1.

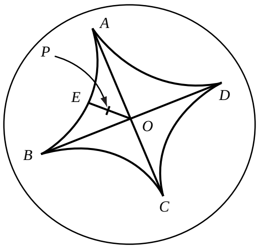

Proof of the Sine Formula (Subsection 6.4.3, Theorem 10).

4. Let ABC be a d-triangle with $ABC = /3$, and let the sides BA and BC have the same hyperbolic length c. Let BD be the altitude from B to the side AC. Show that the hyperbolic length b of AC is greater than c.
Hint: Apply the Sine Formula to triangle BDA.
5. Show that if ABC is a d-triangle right-angled at C, then

$$\sin A = \frac{\sinh 2a}{\sinh 2c},$$

   where a and c are the hyperbolic lengths of the sides BC and AB.
Hint: If $\tan \ = t$, then $\sin^2 \ = \frac{t^2}{1+t^2}$.
6. Determine the hyperbolic lengths of the sides of the d-triangle with angles $\frac{}{2}$, $\frac{}{6}$ and $\frac{}{6}$.

Section 6.5

1. Determine the hyperbolic area of the d-triangle ABC with vertices at the points $A(0,0)$, $B\ \frac{1}{2}, 0$ and $C\ 0, \frac{1}{2}$.
2. Let A, B, C and D be the points (0,0), (1,0), (0,1) and $(\ 1, 0)$, respectively, and let be the d-line BC.

(a) Determine the vertices of the d-triangle that is the image of the d-triangle ABC under hyperbolic reflection in .

(b) Determine the vertices of the d-triangle that is the image of the d-triangle ACD under hyperbolic reflection in .

(c) Sketch the d-triangles ABC and ACD, and their images under hyperbolic reflection in .

(d) Determine the areas of the d-triangles ABC and ACD, and of their images under hyperbolic reflection in .

Hint: You may assume that Theorem 4 of Subsection 6.5.1 holds for asymptotic triangles, with $K = 1$.

Summary of Chapter 6

Introduction

1. **The (Euclidean) Parallel Postulate** Given any line and a point P not on , there is a unique line m in the same plane as P and which passes through P and does not meet .
2. **The (Elliptic) Parallel Postulate** Given any line and a point P not on , all lines through P meet . (That is, there are no lines through P that are parallel to .)
3. **The (Hyperbolic) Parallel Postulate** Given any line and a point P not on , there are at least two lines m through P that do not meet . (That is, there are at least two lines through P that are parallel to .)

Section 6.1: Hyperbolic Geometry: the Disc Model

1. In (Poincarø s model of) hyperbolic geometry, **points** (or ***d*-points**, where d stands for disc) consists of the points in the unit disc $\mathbb{D} = \{z : |z| < 1\} = \{(x, y) : x^2 + y^2 < 1\}$.

   The points on the boundary $\mathbb{C} = \{z : |z| = 1\} = \{(x, y) : x^2 + y^2 = 1\}$ of $\mathbb{D}$ are not points that belong to the geometry.
2. A ***d*-line** (or **line in hyperbolic geometry**) is that part of a (Euclidean) generalized circle which meets $\mathbb{C}$ at right angles and which lies in $\mathbb{D}$. A d-line may be an arc of a (Euclidean) circle or a Euclidean line segment that is a diameter of $\mathbb{D}$. The points where the corresponding generalised circle meets $\mathbb{C}$ are called the **boundary points** of the d-line.

   d-lines that are arcs of Euclidean circles all lie in one-half of the disc $\mathbb{D}$, and all curve away from the diameter bounding the half-disc.

   The equation of a d-line is of the form $ax + by = 0$, where a and b are not both zero, or $x^2 + y^2 + fx + gy + 1 = 0$, where $f^2 + g^2 > 4$.
3. Two d-lines that do not meet in $\mathbb{D}$ are **parallel** if the generalized Euclidean circles of which they are parts meet at a point on $\mathbb{C}$, or **ultra-parallel** if the generalized Euclidean circles of which they are parts do not meet on $\mathbb{C}$.

   Corresponding to any given d-line and any point P in $\mathbb{D}$ that is not on , there exist exactly two d-lines through P that are parallel to , and infinitely many d-lines through P that are ultra-parallel to .
4. Let be a d-line that is part of a Euclidean circle C. Then inversion in C maps $\mathbb{C}$ onto $\mathbb{C}$, and $\mathbb{D}$ onto $\mathbb{D}$.

5. A **hyperbolic reflection** in a d-line is the restriction to D of the inversion in the generalized circle C of which is part.

   The centre of such a circle C is the point of intersection of the tangents to C where C and C intersect.
6. **Hyperbolic transformations** are the composites of a finite number of hyperbolic reflections, and form a group called the **hyperbolic group**, G_{D}, under the operation of composition of functions.

   Hyperbolic geometry consists of the unit disc D together with the group G_{D} of hyperbolic transformations.
7. The (**hyperbolic**) **angle between two curves** through a given point A in D is the Euclidean angle between their (Euclidean) tangents at A.
8. Hyperbolic transformations map d-lines onto d-lines, and preserve the magnitudes of angles.
9. **Origin Lemma** Let A be a point of D other than the origin O. Then there exists a d-line such that hyperbolic reflection in maps A to O.
10. Let A be a point of D. Then there exist infinitely many d-lines through A.

    Let A and B be two distinct points of D. Then there exists a unique d-line through A and B.
11. Let A_1 and A_2 be any two points of D, and let $_1$ and $_2$ be d-lines through A_1 and A_2, respectively. Then there is a hyperbolic transformation which maps A_1 to A_2 and $_1$ to $_2$.
12. Let A, B D be inverse points with respect to the d-line , and let A , B and be the images of A, B and under hyperbolic reflection in another d-line . Then A and B are inverse points with respect to inversion in .

Section 6.2: Hyperbolic Transformations

1. Hyperbolic reflection in the d-line which is part of a Euclidean circle C with centre is given by the formula

   $$(z) = \frac{\overline{z} \quad 1}{\overline{z} \quad \overline{}}, \; z \quad \mathsf{D}.$$

   Hyperbolic reflection is self-inverse.

   Hyperbolic reflection in the diameter of D on which $y = x \tan$ is given by the formula $(z) = {}^2\overline{z}$, z D, where $= \cos + i \sin$.
2. The composite of the hyperbolic reflections $(z) = \frac{\overline{z} \; 1}{\overline{z} \; \overline{}}$ and $(z) = \frac{\overline{z} \; 1}{\overline{z} \; \overline{}}$ is $(\quad)(z) = \frac{(\overline{} \quad 1)\overline{z} +}{(\overline{} \quad \overline{})\overline{z} + \quad \overline{} \quad 1}$.
3. The restriction to D of every Möbius transformation of the form $M(z) = \frac{az+b}{\overline{b}z+\overline{a}}$, with $|b| < |a|$, is a composite of two hyperbolic reflections, and is therefore a hyperbolic transformation.
4. Every hyperbolic transformation can be expressed as a composite of at most three hyperbolic reflections.
5. The composition of an even number of hyperbolic reflections does not change the orientation of the unit disc, and is called a **direct** hyperbolic

transformation. A direct hyperbolic transformation can be expressed as a composite of at most two hyperbolic reflections.

Any direct hyperbolic transformation is given by $z \mapsto M(z) = \frac{az+b}{\overline{b}z+\overline{a}}$, $z \in \mathbb{D}$, for some a, b with $|b| < |a|$.

The composition of an odd number of hyperbolic reflections changes the orientation of the unit disc, and is called an **indirect** hyperbolic transformation.

Any indirect hyperbolic transformation is given by $z \mapsto M(\overline{z})$, where $M(z) = \frac{az+b}{\overline{b}z+\overline{a}}$, $z \in \mathbb{D}$, for some a,b with $|b| < |a|$.

6. The composite of hyperbolic reflections in d-lines ℓ_1 and ℓ_2 that intersect at a point in $\mathbb{D}$ is called a (**hyperbolic**) **rotation** and has exactly one fixed point in $\mathbb{D}$.

   The composite of hyperbolic reflections in d-lines ℓ_1 and ℓ_2 that do not meet in $\mathbb{D}$ but do meet at a point on $\mathbb{C}$ is called a (**hyperbolic**) **limit rotation**, and has no fixed points in $\mathbb{D}$ but leaves one point of $\mathbb{C}$ fixed.

   The composite of hyperbolic reflections in d-lines ℓ_1 and ℓ_2 that do not meet in $\mathbb{D}$ or on $\mathbb{C}$ is called a (**hyperbolic**) **translation**, and has no fixed points in $\mathbb{D}$ or on $\mathbb{C}$.

   The composite of two hyperbolic translations may depend on the order in which they are applied.

7. A direct hyperbolic transformation M can be written in the (canonical) form $M(z) = K\frac{z-m}{1-\overline{m}z}$, where K and m are complex numbers with $|K| = 1$ and $m \in \mathbb{D}$.

   Conversely, every direct hyperbolic transformation M that maps a point m of $\mathbb{D}$ to the origin is of this form.

   A direct hyperbolic transformation M that maps the diameter $(-1,1)$ onto itself is of the form $M(z) = \pm\frac{z-m}{1-mz}$, where $m \in (-1,1)$.

   An indirect hyperbolic transformation M can be written in the (canonical) form $z \mapsto K\frac{\overline{z}-m}{1-\overline{m}\overline{z}}$, where $|K| = 1$ and $m \in \mathbb{D}$.

8. **Strategy** To determine the general form of the direct hyperbolic transformation that maps one point p in $\mathbb{D}$ to another point q in $\mathbb{D}$:
   1. write down the general form of the direct hyperbolic transformation M_1 that maps p to 0, and a matrix $\mathbf{A}_1$ associated with M_1;
   2. write down a direct hyperbolic transformation M_2 that maps q to 0, and a matrix $\mathbf{A}_2$ associated with M_2;
   3. form the matrix product $\mathbf{A}_2^{-1}\mathbf{A}_1$ associated with the direct transformation $M_2^{-1} \circ M_1$, and hence write down the general form of the transformation.

Section 6.3: Distance in Hyperbolic Geometry

1. **Properties of a Distance Function d**
   1. $d(z_1, z_2) \geq 0$ for all z_1 and z_2;
      $d(z_1, z_2) = 0$ if and only if $z_1 = z_2$.
   2. $d(z_1, z_2) = d(z_2, z_1)$ for all z_1 and z_2.

3. $d(z_1, z_3) + d(z_3, z_2) \geq d(z_1, z_2)$ for all z_1, z_2 and z_3. (Triangle Inequality)
4. $d(z_1, z_3) + d(z_3, z_2) = d(z_1, z_2)$ if and only if z_1, z_2 and z_3 lie in this order on a line.

Additional Properties of the Distance Function *d* in Hyperbolic Geometry

5. $d(z_1, z_2) = d(\overline{z_1}, \overline{z_2})$ for all z_1 and z_2 in $\mathbb{D}$.
6. $d(z_1, z_2) = d(M(z_1), M(z_2))$ for all z_1 and z_2 in $\mathbb{D}$ and all direct hyperbolic transformations M in $G_{\mathbb{D}}$.

2. The **hyperbolic distance** $d(0, z)$ between the origin and any point $z \in \mathbb{D}$ is given by $d(0, z) = \tanh^{-1}(|z|) = \frac{1}{2}\log_e\left(\frac{1+|z|}{1-|z|}\right)$. Also, $|z| = \left(e^{2d(0,z)} - 1\right) / \left(e^{2d(0,z)} + 1\right)$.

   In general, the hyperbolic distance between two arbitrary points z_1 and z_2 in $\mathbb{D}$ is given by $d(z_1, z_2) = \tanh^{-1}\left(\left|\frac{z_2 - z_1}{1 - \overline{z_1}z_2}\right|\right)$.
3. For small z, $d(0, z) \simeq |z|$; as $|z| \to 1$, $d(0, z) \to +\infty$; for $z \neq 0$, $d(0, z) > |z|$.
4. A point m in $\mathbb{D}$ is the **hyperbolic midpoint** of the hyperbolic line segment joining a and b in $\mathbb{D}$ if m lies on this segment and $d(a, m) = d(m, b) = \frac{1}{2}d(a, b)$.
5. **Strategy** To find the hyperbolic midpoint of a hyperbolic line segment joining two points p, q on a diameter of $\mathbb{D}$, where $|p| > |q|$:
   1. calculate $d(0, p)$ and $d(0, q)$;
   2. if p and q lie on opposite sides of O, then calculate $d = \frac{1}{2}(d(0, p) - d(0, q))$; otherwise calculate $d = \frac{1}{2}(d(0, p) + d(0, q))$;
   3. then the hyperbolic midpoint is the point m on the radius through O and p at a Euclidean distance $\tanh(d)$ from O.
6. The **hyperbolic circle** with **hyperbolic radius** r and **hyperbolic centre** c is the set $\{z : d(c, z) = r,\ z \in \mathbb{D}\}$.
7. Every hyperbolic circle in $\mathbb{D}$ is a Euclidean circle, and vice versa.

   Strategy To find the Euclidean centre and radius of a hyperbolic circle C with hyperbolic centre m:
   1. find the points a, b where Om meets C;
   2. the Euclidean centre of C is the Euclidean mid-point of ab;
   3. the Euclidean radius of C is $\frac{1}{2}|a - b|$.

   Strategy To find the hyperbolic centre and radius of a Euclidean circle K with Euclidean centre p:
   1. find the points a, b where Op meets K;
   2. the hyperbolic centre of K is the hyperbolic mid-point of ab;
   3. the hyperbolic radius of K is $\frac{1}{2}d(a, b)$.
8. Let A and A' be points in the unit disc $\mathbb{D}$ that are images of each other under reflection in a d-line ℓ. Then ℓ is the (hyperbolic) perpendicular bisector of the hyperbolic line segment AA'.
9. **Reflection Lemma** Let p and q be points of $\mathbb{D}$. If $|p| \neq |q|$, then the hyperbolic reflection that maps p and q onto each other is given by

$$M(z) = (\ \bar{z} \quad 1) \quad (\bar{z} \quad \bar{\ }), \text{ where}$$

$$= (p \quad q + pq(\bar{p} \quad \bar{q})) \quad (p\bar{p} \quad q\bar{q}).$$

The *d*-line in which this reflection takes place has equation

$$x^2 + y^2 \quad 2ax \quad 2by + 1 = 0, \text{ where } \quad = a + ib.$$

If $|p| = |q|$, then the reflection that maps p and q onto each other is reflection in the diameter of D that bisects the angle pOq.

Section 6.4: Geometrical theorems

1. A ***d*-triangle** consists of three points in the unit disc D that do not lie on a single *d*-line, together with the segments of the three *d*-lines joining them.
2. The sum of the angles of a *d*-triangle is less than .

   Each external angle of a *d*-triangle is greater than the sum of the opposite two internal angles.

   Small triangles have angle sums close to (but less than) , and triangles with large areas have angle sums close to zero.
3. A ***d*-quadrilateral** *ABCD* consists of four points A, B, C, D in D (no three of which lie on a single *d*-line), together with the segments *AB*, *BC*, *CD* and *DA* of the four *d*-lines joining them. We also require that no two of these segments meet except at one of the points A, B, C or D.

   The sum of the angles of a *d*-quadrilateral is less than 2 .
4. In a *d*-triangle *ABC* the angles *ABC* and *ACB* are equal if and only if the sides *AB* and *AC* are of equal (hyperbolic) length.
5. Two figures in the unit disc D are ***d*-congruent** if there is a hyperbolic transformation that maps one onto the other.

   Similar *d*-triangles (that is, *d*-triangles with corresponding angles equal) are *d*-congruent.
6. An **asymptotic triangle** is a *d*-triangle, except that one or more of its vertices lie on C rather than in D. Simply asymptotic, doubly asymptotic and trebly asymptotic triangles have 1, 2 and 3 vertices on C, respectively.

   The angle sum of an asymptotic triangle is less than . The angle sum of a trebly asymptotic triangle is zero.
7. Let and be two *d*-lines, and suppose that there exist points *A* on and *A* on such that the *d*-line segment *AA* meets and at right angles. Then *AA* is a **common perpendicular** to and .

   Common Perpendiculars Theorem Two *d*-lines have a common perpendicular if and only if they are ultra-parallel. This common perpendicular is unique.
8. Through any point *P* of D not on a *d*-line , there is a unique *d*-line that meets at right angles.

   Let be a *d*-line which passes through one vertex *A* of a *d*-triangle *ABC* and which is perpendicular to the side *BC* at the point *D*. The *d*-line segment *AD* is an **altitude** of the triangle *ABC*.

   Altitude Theorem Let the sides *AB* and *AC* of a *d*-triangle *ABC* be of equal hyperbolic length, and let the angle at *A* be . Then the hyperbolic length of the altitude through A of the triangle is less than some number that depends only on .

9. **Pythagoras' Theorem** Let ABC be a d-triangle in which the angle at C is a right angle. If a, b and c are the hyperbolic lengths of BC, CA and AB, then $\cosh 2c = \cosh 2a \times \cosh 2b$.
10. **Lobachevskii's Formula** Let the d-triangle ABC have a right angle at C, and let the hyperbolic lengths of AC and BC be b and a, respectively. Then $\tan A = \tanh 2a \quad \sinh 2b$.
11. **Angle of Parallelism** Let be a d-line, and P a point of $\mathbb{D}$ that does not lie on . Then the angle between the perpendicular from P to (of hyperbolic length p) and either d-line through P that is parallel to is given by $\tan \quad = 1 \quad \sinh 2p$.
12. **Sine Formula** Let ABC be a d-triangle right-angled at C. Let a and c be the hyperbolic lengths of BC and AB, respectively. Then $\sin A = \sinh 2a \quad \sinh 2c$.
13. Let be a d-line that ends at points A and B on $\mathbb{C}$, and let be that part in $\mathbb{D}$ of a (Euclidean) circle through A and B, and let be the angle (in $\mathbb{D}$) between and . Then all points on are the same distance $\frac{1}{2}\log_e \ \tan \ \frac{}{2} + \frac{}{4}$ from .
14. For $x \quad (\ 1, 1)$, the following are all equal to $\tanh \ ^1 x$: $\frac{1}{2}\log_e \frac{1+x}{1 \ x}$, $\frac{1}{2}\sinh \ ^1 \frac{2x}{1 \ x^2}$, $\frac{1}{2}\cosh \ ^1 \frac{1+x^2}{1 \ x^2}$.

Section 6.5: Area

1. **Area** in hyperbolic geometry has the following properties:
    the area of a d-triangle is non-negative, and is zero only if its Euclidean area is zero;
    d-congruent d-triangles (and d-congruent asymptotic d-triangles) have the same area;
    if one d-triangle can be fitted inside another, it has a smaller area;
    area is additive.
2. All trebly asymptotic d-triangles are d-congruent to each other.
3. Let ABC be a d-triangle. Then there exists a trebly asymptotic d-triangle DEF that contains ABC.
4. All trebly asymptotic d-triangles have the same (hyperbolic) area, and this is a finite number.
5. The area of a d-triangle with angles , , is given by $K(\quad (\ + \ + \ ))$, where K is the same constant for all d-triangles. It is usual to take $K = 1$, so that the area of a d-triangle is its **angular defect**.
    Let , , be positive real numbers with $0 < \quad + \quad + \quad < \quad$. Then there is a d-triangle in $\mathbb{D}$ with angles , , .
6. Let P be a point on a d-line of $\mathbb{D}$, and let $0 < \quad < \frac{}{2}$. Then there are exactly two d-lines through P that make an acute angle with , and each is a (hyperbolic) reflection of the other in .

7. **Tessellate** means fit together exactly a number of identical shapes, leaving no spaces .

   Tessellations of the Euclidean plane using a single figure are sometimes called **wallpaper patterns**. There are only 17 different wallpaper patterns, in the sense that the symmetry group of any wallpaper pattern is isomorphic to one of only 17 non-isomorphic groups.
8. A **kaleidoscope** is a machine that produces wallpaper patterns using reflections in mirrors.

   Successive reflections of a point in two plane mirrors that meet at an angle π/k give $2k$ points in all.

   Successive (hyperbolic) reflections of a point in two mirrors along d-lines in $\mathbb{D}$ that meet at an angle π/k give $2k$ points in all. We may obtain a **hyperbolic tessellation** of $\mathbb{D}$ (with d-congruent d-triangles) by successive hyperbolic reflections starting from a d-triangle in $\mathbb{D}$ formed by three mirrors along d-lines at angles of $\pi/7$ (at the origin), $\pi/2$ and $\pi/3$.

Section 6.6: Hyperbolic Geometry: the Half-Plane Model

1. In the half-plane model, the set of points is the upper half-plane $H = \{z : z \in \mathbb{C}, \operatorname{Im} z > 0\}$; h-lines are the restriction to H of vertical lines in the plane together with the restriction to H of circles with centre on the real axis in $\mathbb{C}$; and the hyperbolic transformations are the mappings of H to itself given by $z \mapsto (az + b)/(cz + d)$ or $z \mapsto (-a\bar{z} + b)/(-c\bar{z} + d)$, where a, b, c, d are real and $ad - bc > 0$.

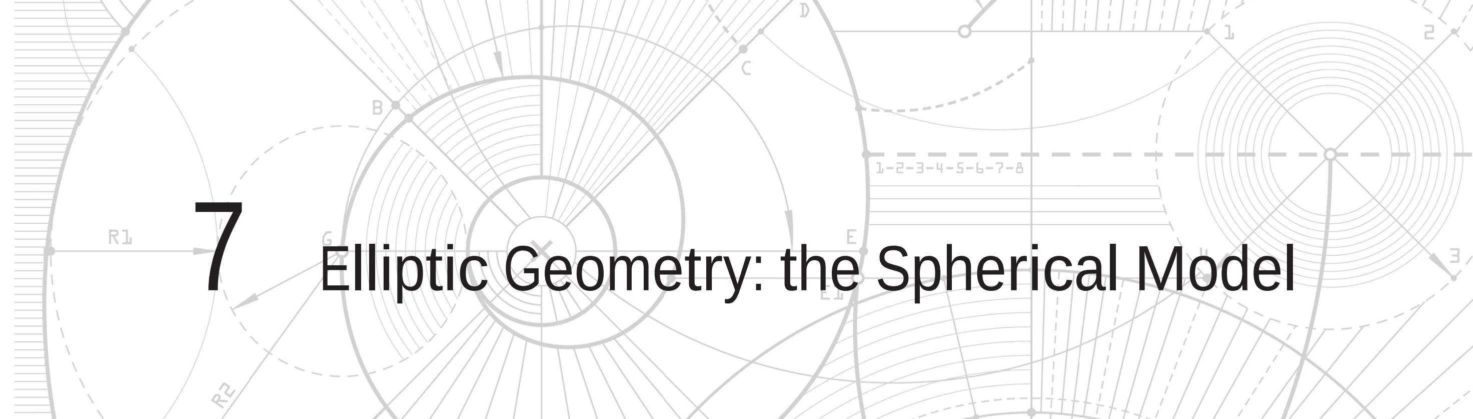

7 Elliptic Geometry: the Spherical Model

In the introduction to Chapter 6, we saw that in Elliptic Geometry all the postulates of Euclidean Geometry hold except that the Euclidean Parallel Postulate is replaced by the following.

We usually take the sphere to have radius 1.

The (Elliptic) Parallel Postulate Given any line and a point P not on , all lines through P meet . (That is, there are no lines through P that are parallel to .)

In this chapter we study the geometry of the sphere (which we call S^2), such as the surface of the Earth, as a model of elliptic geometry.

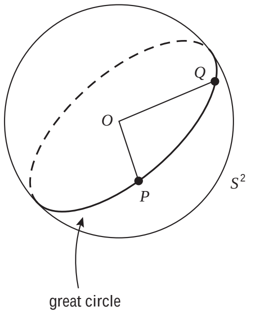

A line joining two points P and Q on the surface of a sphere is the curve of shortest length (that is, a *geodesic*) on the sphere between the points: it is actually the *great circle* (that is, the intersection of the plane through the centre O of the sphere with the surface of the sphere) through the two points. Any two distinct great circles meet in two (diametrically opposite) points, so any two lines in spherical geometry meet in two points.

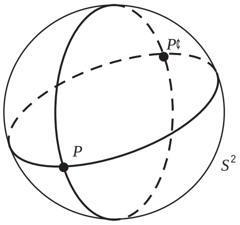

The group of transformations in spherical geometry is the group of isometries of the sphere. Clearly a rotation of the sphere is an isometry, and we shall see that the group of isometries is composed of rotations of the sphere together with reflections in great circles. It turns out that any orientation-preserving isometry of the sphere is in fact a single rotation.

Recall that an *isometry* is a one one mapping that preserves distances.

In Section 7.1, we establish a system of coordinates for the sphere; these are essentially the familiar coordinates of *latitude* and *longitude*. Then, in Section 7.2, we obtain matrix representations for the isometries of the sphere: every isometry can be written as a 3 × 3 matrix.

Next, in Section 7.3, we define a *triangle* on the sphere, find a formula expressing its area in terms of the sum of its angles, and show that every spherical triangle has a *dual triangle* whose sides and angles have the magnitudes of the angles and sides, respectively, of the original triangle.

There is an attractive formula for the distance between two points on the sphere in terms of their coordinates. There is also an attractive generalization of Pythagoras Theorem in spherical geometry, and in proving it we are led to the trigonometric formulas connecting the sides and angles of a spherical triangle. These establish that of the six measurements that describe a spherical triangle (three side lengths, three angles) any set of three determines the other three. This is not the case in Euclidean geometry (where similar, non-congruent figures exist), but it is also true in hyperbolic geometry. Indeed, there are closer analogies between spherical geometry and hyperbolic geometry than there are between either geometry and Euclidean geometry.

Many aspects of spherical geometry may be studied more easily if we first map the sphere onto a plane. This can be done in many ways, although, none of the mappings (or *projections*) is an isometry of the sphere onto the plane. In Section 7.4, we revisit *stereographic projection*, which maps the sphere onto the plane in a way that preserves angles and generalized circles. We also show how to obtain the group of isometries of the stereographic image of the sphere on the plane; and revisit coaxal circles. Finally, in Section 7.5, we look briefly at the problem of mapping the spherical Earth onto a flat map.

7.1 Spherical Space

7.1.1 Spherical Geometry

Spherical geometry is the study of geometry on the surface of a sphere. It is simplest to choose this sphere to have unit radius; we therefore concentrate on what we call the *unit sphere* in Euclidean 3-dimensional space $\mathbb{R}^3$, namely the sphere whose centre is at the origin and which has radius 1. In terms of the usual x-, y-, z-Cartesian coordinates in $\mathbb{R}^3$ with origin O at the centre of the sphere, the unit sphere has the equation $x^2 + y^2 + z^2 = 1$. We shall denote it by the symbol S^2.

It is usually straightforward to modify our results to the case of a sphere of arbitrary radius, R say.

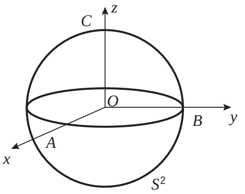

The positive x-, y-, z-axes meet the sphere S^2 at the points $A(1,0,0)$, $B(0, 1, 0)$, $C(0, 0, 1)$, respectively. For obvious reasons the point C is sometimes called the *North Pole* of the sphere, and denoted by N. Every point (x,y,z) on the sphere S^2 has a diametrically opposite or *antipodal* point $(-x, -y, -z)$; for example, the antipodal point to the North Pole is the *South Pole* $S(0,0,-1)$. The plane through O perpendicular to the line ON cuts the sphere in a circle called the *equator*.

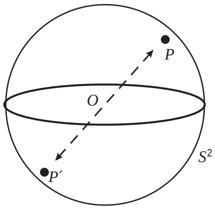

Antipodal points

Example 1 Prove that any plane that cuts the sphere S^2 in more than one point cuts it in a circle.

Solution Let OT be a radius of the sphere perpendicular to the given plane. A rotation of the sphere about this axis maps the plane to itself, and the sphere to itself; so it maps their intersection to itself. Each point on the intersection traces out a circle (of the same radius and centre) under the rotation, so the intersection is a circle. □

The analogue in spherical geometry of a straight line in the plane is a curve on S^2 called a great circle.

Definition A **great circle** is the circle cut out on the sphere S^2 by a plane through the centre of the sphere.

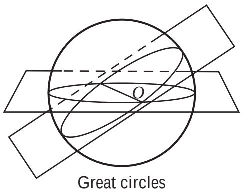

Great circles

Note that, by Example 1, a great circle is indeed a circle. Moreover, it is a circle of maximal radius 1 on the sphere, which gives rise to the name great circle.

Definitions Any great circle through N is called a **circle of longitude**. The part of the circle of longitude through $A(1,0,0)$ which has positive x-coordinates we call the **Greenwich Meridian**. The curve of intersection of S^2 with a plane parallel to the equator is a circle called a **circle of latitude**. The curve of intersection of S^2 with a plane that is not a great circle is called a **little circle**.

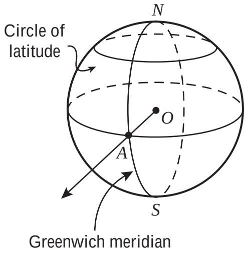

The equator is a circle of latitude; circles of latitude other than the equator are little circles.

We define distance in spherical geometry (on S^2) in the natural way.

Definitions Let P and Q be any two points on S^2, and C a great circle through P and Q. Then either arc with endpoints P and Q is said to be a **line** on S^2 joining P and Q, or a **line** PQ. The **length** of a given line on S^2 is defined to be the Euclidean length of the corresponding circular arc; the **distance** (or **spherical distance**) between two points P and Q on S^2 is the length of the shorter of the two lines PQ, and is often denoted simply as PQ.

Let P, Q and R be any three points on S^2 that do not lie on a single great circle. Then a line PQ, a line QR and a line RP form a **triangle** PQR.

Thus the line PQ is not unique.

Thus there are several triangles PQR.

A property of great circles that we state without proof is that *great circles are geodesics on* S^2; that is, curves of shortest length between any of their points. This is intuitively clear for the circles of longitude from the North Pole, for instance. Note that little circles are not geodesics on S^2.

The distance between any two points on S^2 is a portion of a great circle through the points. This circle has radius 1, and so has circumference 2 the same as the angle that the great circle subtends at its center (the centre of the sphere). It follows, by proportion, that the distance between any two points on S^2 is equal to the angle that they subtend at the centre of S^2.

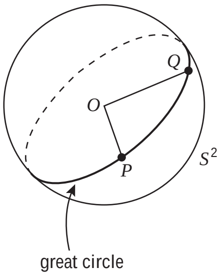

Theorem 1 The distance between any two points P and Q on S^2 is equal to the angle that they subtend at the centre of S^2.

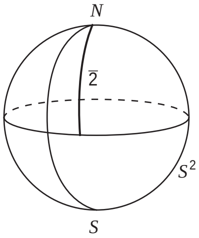

Thus, for example, the distance from the North Pole to any point on the equator is $\pi/2$, and the distance from the North Pole to the South Pole is π.

With the above definition, spherical distance possesses all the "natural" properties that a distance function should possess. (We listed these in Subsection 6.3.1.)

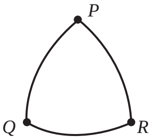

For example, it satisfies the Triangle Inequality; that is, if P, Q and R are any three points on S^2, then the side lengths of the triangle $\triangle PQR$ satisfy the inequality $PQ + QR \geq PR$. This can be proved in a similar way to the Triangle Inequality in hyperbolic geometry (Theorem 2, Subsection 6.3.3).

Great circles in spherical geometry play many of the roles that lines play in Euclidean plane geometry. There is always a great circle joining any two points on S^2, because there is always a plane passing through them and the origin. The plane and the corresponding great circle are unique unless the points are diametrically opposite, in which case there are infinitely many great circles joining the points.

However, one way in which spherical geometry differs markedly from Euclidean plane geometry comes from observing that any two great circles meet not in one, but in two (diametrically opposite) points. They do so because the planes that define them have a common point, the origin, and therefore a common line in $\mathbb{R}^3$, which cuts S^2 in the given points. So although great circles are analogous to the straight lines of the plane, the analogy breaks down almost at once. There are certainly no "parallel lines" on the sphere, because any two great circles meet in two points.

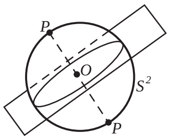

The antipodal points P and P' have a common equator

It is sometimes helpful to observe that corresponding to each point P on S^2 there is a unique great circle of points $\pi/2$ distant from P, which plays the role of the equator with P as a "pole". We shall call this great circle the **great circle associated with the point** P. Conversely, given any great circle there are two points of S^2 which play the role of its "poles".

Example 2 Let P and Q be any two points on S^2. Show that there is a rotation of S^2 that maps P to Q.

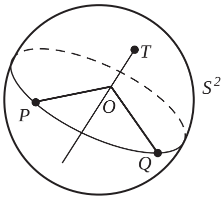

Solution Consider the great circle through P and Q, and let T be one of its poles. The rotation of S^2 about the axis OT and through the angle $\angle POQ$ (in the appropriate direction) sends P to Q, as required. □

Problem 1 Determine whether rotations of S^2 exist with the following properties.

(a) $(1,0,0) \to (0,-1,0)$
(b) $(1,0,0) \to (0,1,0)$ and $(0,1,0) \to (0,0,1)$
(c) $(1,0,0) \to (0,-1,0)$ and $(0,0,1) \to (0,1,0)$

7.1.2 Coordinates on S^2

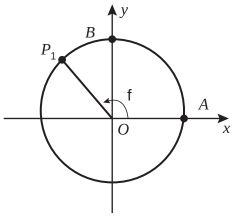

Equatorial plane of S^2

We now set up a system of *spherical polar coordinates* on S^2. We start by defining the direction along the equator from $A(1,0,0)$ to $B(0,1,0)$ to be the positive direction on the equator. Then for any point P_1 on the equator let ϕ be the angle $\angle AOP_1$ measured in the positive direction along the equator; then we assign the coordinates $(\cos\phi, \sin\phi, 0)$ to P_1, where $0 \le \phi < 2\pi$.

Next, let P be any point on S^2 (other than the North and South Poles) that does not lie on the equator. Let P_1 be the point on the equator that lies on the arc NS of the great circle through N and P.

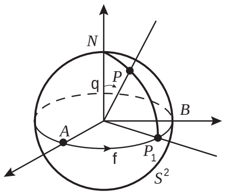

P has spherical coordinates $(\cos\phi\sin\theta, \sin\phi\sin\theta, \cos\theta)$.

Then let ϕ be the angle $\angle AOP_1$ measured in the positive direction along the equator $(0 \le \phi < 2\pi)$, and let θ be the angle $\angle NOP$ $(0 < \theta < \pi)$; we assign the coordinates $(\cos\phi\sin\theta, \sin\phi\sin\theta, \cos\theta)$ to P.

For example, the point of S^2 with spherical polar coordinates $\left(\cos\frac{\pi}{3}\sin\frac{\pi}{4}, \sin\frac{\pi}{3}\sin\frac{\pi}{4}, \cos\frac{\pi}{4}\right)$ is the point $\left(\frac{1}{2\sqrt{2}}, \frac{\sqrt{3}}{2\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$.

Problem 2 Determine the angles ϕ and θ in the spherical polar representation $(\cos\phi\sin\theta, \sin\phi\sin\theta, \cos\theta)$ of the points $\left(0, -\frac{\sqrt{3}}{2}, \frac{1}{2}\right)$ and $\left(-\frac{1}{\sqrt{14}}, -\frac{\sqrt{2}}{\sqrt{14}}, -\frac{3}{\sqrt{14}}\right)$.

In this way every point on S^2 other than the North and South Poles has been given spherical polar coordinates involving appropriate values of ϕ and θ. At the North Pole we take θ to be zero and let ϕ take any value in the interval $[0, 2\pi)$, and at the South Pole we take θ to be π and let ϕ take any value in the interval $[0, 2\pi)$; the non-uniqueness of coordinates at the Poles will cause us no problems.

Definitions Let P be a point on S^2 with spherical polar coordinates $(\cos\phi\sin\theta, \sin\phi\sin\theta, \cos\theta)$, where $0 \le \phi < 2\pi$ and $0 \le \theta \le \pi$. The angle θ of P is called its **colatitude**; the quantity $\lambda = \frac{\pi}{2} - \theta$ is called its **latitude** (note that $-\frac{\pi}{2} \le \lambda \le \frac{\pi}{2}$). The angle ϕ of P is called its **longitude**.

Sometimes we shall find it convenient to regard ϕ as being defined on $\mathbb{R}$ rather than on $[0, 2\pi)$, as a periodic function with period 2π.

Notice, for example, that the distance of a point P with colatitude θ on S^2 from the North Pole is simply θ.

Problem 3 Determine the latitude, colatitude and longitude of the point $\left(\frac{1}{2\sqrt{2}}, \frac{\sqrt{3}}{2\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ on S^2.
Hint: You met this point just before Problem 2 above.

Problem 4 Determine the spherical polar coordinates of the point 45° W of the Greenwich meridian on S^2 and 30° S.

Note that 30° S means '30° below the equatorial plane'. Also, sometimes we take the phrase 'of the Greenwich meridian' as understood when describing longitude.

Notice also that the set of points on the sphere with a common colatitude θ forms a circle, C_θ say; this is a little circle if $\theta \neq \pi/2$, and a great circle (the equator) if $\theta = \pi/2$. Since θ measures the distance of each point of the circle C_θ from the North Pole N, we shall say that C_θ is a *circle on the sphere* *with centre N*. Since any point on S^2 may be chosen as the North Pole (with a suitable choice of coordinate axes), it follows that a circle on S^2, defined as the set of points equidistant from a given point, is either a great circle or a little circle.

Problem 5 Show that, given any two great circles on S^2, there are two great circles bisecting the angles between them, each at right angles to the other.
Hint: The angle between two curves on S^2 at one of their points of intersection is the angle between their tangents at that point.

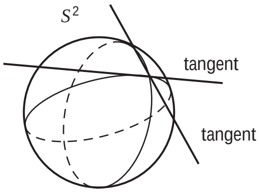

7.2 Spherical Transformations

7.2.1 Isometries of S^2

An isometry of S^2 is a mapping of the unit sphere to itself that preserves distances between points. In fact the isometries of a sphere form a group, since

the composition of two isometries is also an isometry;
the identity mapping is an isometry, and it is the identity element of the group;
the inverse of any isometry is an isometry;
the composition of isometries is associative, since the composition of functions is always associative.

We have therefore established the following theorem.

Theorem 1 The isometries of S^2 form a group.

We call this group the *group of spherical isometries*, and denote it by the symbol $S(2)$.

We shall see at the end of this section that the rotations of S^2 form a group, which is identical to the group of direct isometries of the sphere. A reflection of S^2 in a plane through the centre of the unit sphere is not a direct isometry of S^2, but every composition of an even number of such reflections is a direct isometry of S^2.

A transformation is *direct* if it preserves the *orientation* of angles, and *indirect* if it reverses them.

We now consider the three *elementary rotations* of S^2 (that is, its rotations about the coordinates axes) and their composites.

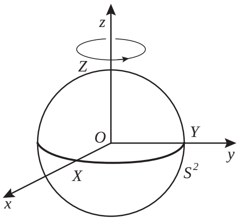

The elementary rotation about OZ leaves the z-axis fixed, and rotates the (x, y)-plane. The positive direction of this rotation is anticlockwise as seen from above (that is, as viewed from a point of the positive z-axis), from the positive x-axis towards the positive y-axis. So a rotation of S^2 about OZ through an angle , which we denote by $R(Z, \ )$, is given by

$$R(Z, \ ): S^2 \quad S^2$$
$$\mathbf{x} \quad \mathbf{Ax}$$

where $\mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$ and $\mathbf{A}$ is the matrix $\begin{pmatrix} \cos & \sin & 0 \\ \sin & \cos & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

We say that **A** is the matrix *associated with* (or *of*) the elementary rotation $R(Z, \ )$.

Next, the elementary rotation about OX leaves the x-axis fixed, and rotates the (y, z)-plane. The positive direction of this rotation is taken to be anticlockwise as seen from above (that is, as viewed from a point of the positive x-axis), from the positive y-axis towards the positive z-axis. So a rotation of S^2 about OX through an angle , which we denote by $R(X, \ )$, is given by

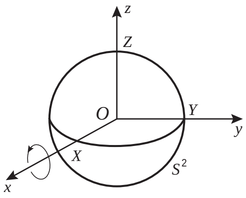

$$R(x, \ ): S^2 \quad S^2$$
$$\mathbf{x} \quad \mathbf{Ax}$$

where $\mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$ and $\mathbf{A}$ is the matrix $\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos & \sin \\ 0 & \sin & \cos \end{pmatrix}$.

Here **A** is the matrix *associated with* (or *of*) the elementary rotation $R(X, \ )$.

Finally, the elementary rotation about OY leaves the y-axis fixed, and rotates the (x, z)-plane. The positive direction of this rotation is taken to be anticlockwise as seen from above (that is, as viewed from a point of the positive y-axis), from the positive z-axis towards the positive x-axis. So a rotation of S^2 about OY through an angle , which we denote by $R(Y, \ )$, is given by

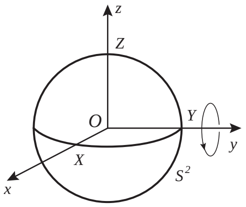

$$R(Y, \ ): S^2 \quad S^2$$
$$\mathbf{x} \quad \mathbf{Ax}$$

where $\mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$ and $\mathbf{A}$ is the matrix $\begin{pmatrix} \cos & 0 & \sin \\ 0 & 1 & 0 \\ \sin & 0 & \cos \end{pmatrix}$.

Here **A** is the matrix *associated with* (or *of*) the elementary rotation $R(y, \ )$.

It is easy to verify that the determinants of the elementary rotations $R(X, \ )$, $R(Y, \ )$ and $R(Z, \ )$ are all equal to 1. Since any rotation can be expressed as a finite composition of these rotations, it follows that the determinant of the matrix of any rotation of S^2 also has determinant 1.

We will prove this fact in Theorem 7, Subsection 7.2.3.

Problem 1 Determine the images of the points $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$ and $\frac{1}{14}, \ \frac{2}{14}, \frac{3}{14}$ under the transformation $R \ \ Y, \frac{}{4}$.

Notice that the simple cases of these rotations when , or $=$ $/2$ correspond to rotations of one axis onto another. For example,

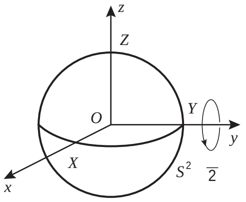

$$R \ \ Y, \frac{}{2} = \begin{pmatrix} \cos\frac{}{2} & 0 & \sin\frac{}{2} \\ 0 & 1 & 0 \\ \sin\frac{}{2} & 0 & \cos\frac{}{2} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix};$$

under this mapping, $(1, 0, 0) \quad (0, 0, \ 1)$, $(0, 1, 0) \quad (0, 1, 0)$ and $(0, 0, 1) \quad (1, 0, 0)$.

Example 1 Show that $R\ \ X, \frac{}{2} \ \bullet R\ \ Y, \frac{}{2} \ = R\ \ Y, \frac{}{2} \ \bullet R\ \ X, \frac{}{2}$.

This shows that multiplication of the matrices of elementary rotations (and hence composition of elementary rotations) is not commutative.

Solution Here

$$R\ \ X, \frac{}{2} \ \bullet R\ \ Y, \frac{}{2} \ = \begin{matrix} 1 & 0 & 0 \\ 0 & \cos\frac{}{2} & \sin\frac{}{2} \\ 0 & \sin\frac{}{2} & \cos\frac{}{2} \end{matrix}$$

$$\times \begin{matrix} \cos\frac{}{2} & 0 & \sin\frac{}{2} \\ 0 & 1 & 0 \\ \sin\frac{}{2} & 0 & \cos\frac{}{2} \end{matrix}$$

$$= \begin{matrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{matrix} \quad \begin{matrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{matrix}$$

$$= \begin{matrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{matrix} \ ,$$

and

$$R\ \ Y, \frac{}{2} \ \bullet R\ \ X, \frac{}{2} \ = \begin{matrix} \cos\frac{}{2} & 0 & \sin\frac{}{2} \\ 0 & 1 & 0 \\ \sin\frac{}{2} & 0 & \cos\frac{}{2} \end{matrix}$$

$$\times \begin{matrix} 1 & 0 & 0 \\ 0 & \cos\frac{}{2} & \sin\frac{}{2} \\ 0 & \sin\frac{}{2} & \cos\frac{}{2} \end{matrix}$$

$$= \begin{matrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{matrix} \quad \begin{matrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{matrix}$$

$$= \begin{matrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{matrix} \ .$$

Thus $R\ \ X, \frac{}{2} \ \bullet R\ \ Y, \frac{}{2} \ = R\ \ Y, \frac{}{2} \ \bullet R\ \ X, \frac{}{2}$, as required. □

Problem 2 Calculate $R(Z,\ ) \bullet R(Y,\ )$ and $R(Y,\ ) \bullet R(Z,\ )$.

Problem 3 For the elementary rotation $R(X,\ )$, verify that the transpose of its matrix is equal to the matrix of the elementary rotation $R(X,\quad)$. State and prove the corresponding results for the elementary rotations $R(Y,\ )$ and $R(Z,\ )$.

Notice that the inverse of an elementary rotation through a given angle is an elementary rotation through the same angle but in the opposite direction;

that is,

$$R(X,\ )^{\ 1} = R(X,\ \ ),$$
$$R(Y,\ )^{\ 1} = R(Y,\ \ )$$
$$\text{and}\qquad R(Z,\ )^{\ 1} = R(Z,\ \ ).$$

Combining this observation with the result of Problem 3, we obtain the agreeable consequence that the inverse of a matrix representing an elementary rotation is equal to the transpose of that matrix. For example the matrix **A** of the elementary rotation $R(Y,\ )$ is

$$\mathbf{A} = \begin{matrix} \cos & 0 & \sin \\ 0 & 1 & 0 \\ \sin & 0 & \cos \end{matrix}\ ;$$

then $\mathbf{A}^{\ 1} = \mathbf{A}^T$.

Recall that any square matrix **A** with the property that $\mathbf{A}^{\ 1} = \mathbf{A}^T$ or $\mathbf{A}\mathbf{A}^T = \mathbf{I}$ is an *orthogonal* matrix.

We saw earlier that we can rotate S^2 in such a way as to map any particular point onto any other. We now see how we can determine the matrix of such a rotation explicitly.

Subsection 7.1.1, Example 2

First, we see how to rotate S^2 so as to send the point $A(1,0,0)$ to any point $P(\cos\ \ \sin\ ,\sin\ \ \sin\ ,\cos\ )$ of S^2. First we rotate S^2 to send A to the point $P_1(\cos\ ,\sin\ ,0)$; then we rotate S^2 to send P_1 to P.

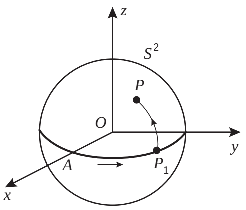

From A to P_1, and from P_1 to P by rotations

The first rotation is the elementary rotation $R(Z,\ )$; but what is the second rotation? It is a rotation of S^2 towards the North Pole through an angle $\ = \frac{}{2}\ \ $, so it is something like $R(Y,\ \ \ )$ or $R\ \ Y,\ \ \ \frac{}{2}$. But it is not a rotation about the y-axis; instead, it is a rotation of S^2 about the direction where the first rotation sent the y-axis. We may perform such a rotation in several stages: rotate the sphere first by the reverse transformation $R(Z,\ \ )$, then by the elementary rotation $R\ \ Y,\ \ \ \frac{}{2}$, and then finally by the transformation $R(Z,\ )$.

For $\ = \frac{}{2}\ \ $, and $R(Z,\ )^{\ 1} = R(Z,\ \ )$.

Thus the second rotation is the composition mapping

$$R(Z,\ )R(Y,\ \ \ )R(Z,\ \ \ ) = R(Z,\ )R(Y,\ \ \ )R(Z,\ )^{\ 1}.$$

It follows that the final transformation that sends A to P is given by

$$R(Z,\ )R(Y,\ \ \ )R(Z,\ )^{\ 1}\ \ R(Z,\ ) = R(Z,\ )R(Y,\ \ \ ).$$

Theorem 2 A rotation of S^2 that maps $A(1,0,0)$ to $P(\cos\ \ \sin\ ,$ $\sin\ \ \sin\ ,\cos\ )$ is given by the composition $R(Z,\ )R(Y,\ \ \ )$, where $\ = \frac{}{2}\ \ \ $.

Note that there are many such rotations. *Any* composition of this particular rotation with a subsequent rotation about OP has the required properties.

Problem 4 Determine the matrix **A** of a rotation **x** $\ \ $ **Ax** of S^2 that maps $A(1,0,0)$ to $P\ \ \frac{1}{2\ 2}, \frac{3}{2\ 2}, \frac{1}{2}\ \ $. Verify your answer by direct calculation.

Hint: You met this point P in Problem 3 of Subsection 7.1.2.

Example 2 Determine a rotation that maps the point $P(\cos\ \ \sin\ \ ,$ $\sin\ \ \sin\ \ , \cos\ \ )$ of S^2 to $N(0,0,1)$, in terms of elementary rotations.

Solution By Theorem 2, one rotation that sends A to P is given by the composition $R(Z,\ \ )R(Y,\ \ \ \ )$, where $\ \ = \overline{\ \ 2}\ \ \ \ $. It follows that the mapping

$$R(Z,\ \ )R(Y,\ \ \ \ )^{\ 1} = R(Y,\ \ \ \ )^{\ 1}R(Z,\ \ )^{\ 1}$$
$$= R(Y,\ \ )R(Z,\ \ \ \ )$$

sends P to A.

Now at the North Pole $\ \ = 0$ and $\ \ = 0$, so that $\ \ = \overline{\ \ 2}$. Hence one rotation that sends A to N is $R\ \ Y,\ \ \overline{\ \ 2}\ \ $. It follows that one rotation that sends P to N is

$$R\ \ Y,\ \ \frac{\ \ }{2}\ \ \ \ R\ \ Y,\ \ \ \ R(Z,\ \ \ \ )$$
$$= R\ \ Y,\ \ \ \ \ \ \frac{\ \ }{2}\ \ R(Z,\ \ \ \ )$$
$$= R(Y,\ \ \ \ )R(Z,\ \ \ \ ).$$ □

For
$R(Y,\ \ _1)R(Y,\ \ _2)$
$= R(Y,\ \ _1 +\ \ _2)$,
for any angles $\ _1,\ \ _2$.

Problem 5 Determine the matrix **A** of a rotation **x** $\ \ \ \ $ **Ax** of S^2 that maps $P\ \ \frac{1}{2\ \overline{2}}, \frac{\overline{3}}{2\ \overline{2}}, \frac{1}{\overline{2}}\ \ $ to N. Verify your answer by direct calculation.
Hint: You met this point P in Problem 4.

Problem 6 Determine the matrix **A** of a rotation **x** $\ \ \ \ $ **Ax** of S^2 that maps $Q\ \ \frac{1}{2},\ \ \frac{1}{2},\ \ \ \frac{1}{\overline{2}}$ to $P\ \ \frac{1}{2\ \overline{2}}, \frac{\overline{3}}{2\ \overline{2}}, \frac{1}{\overline{2}}\ \ $. Verify your answer by direct calculation.

7.2.2 Reflections in Great Circles

Reflection in the *equatorial plane* of S^2 (that is, in the plane $z = 0$ that contains the equator of S^2) is the mapping **x** $\ \ \ \ $ **Ax** that sends the point $\mathbf{x} = (x,y,z)$ to the point $(x,y,\ \ z)$, so $\mathbf{A} = \begin{matrix} 1 & 0 & \ \ 0 \\ 0 & 1 & \ \ 0 \\ 0 & 0 & \ \ 1 \end{matrix}$. Each point lies vertically above or below its image, unless it lies in the equatorial plane in which case it coincides with its image.

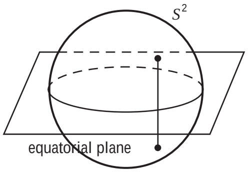

Problem 7 Evaluate the determinant of the above matrix **A**.

Next, let $\ \ $ be any plane that passes through the centre O of the sphere S^2. The equation of $\ \ $ must be of the form $ax + by + cz = 0$, since it passes through O; if we define the vectors **a** and **x** to be (a,b,c) and (x,y,z), respectively, then we can express the equation of $\ \ $ in the form

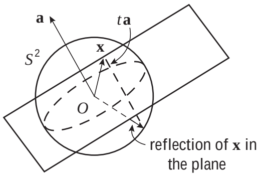

$$\mathbf{a} \bullet \mathbf{x} = 0.$$

This defines a plane Π with the vector $\mathbf{a}$ as a normal to it; without loss of generality, we shall assume that a, b and c have been chosen so that $\mathbf{a}$ has length 1.

Let r be the reflection of S^2 in the plane Π; we shall find a matrix for this mapping r of S^2 to itself.

We shall interchangeably call this *reflection in* Π and *reflection in the great circle* in which Π intersects S^2.

The image of a point $\mathbf{x}$ of S^2 is the point $r(\mathbf{x})$. Now the line joining the points $\mathbf{x}$ and $r(\mathbf{x})$ is a line parallel to the normal vector $\mathbf{a}$, so we must have that

$$r(\mathbf{x}) = \mathbf{x} + t\mathbf{a}, \qquad \text{for some value of the parameter } t.$$

For $r(\mathbf{x}) - \mathbf{x} = t\mathbf{a}$, for some t.

We can find this value of t by observing that since $\mathbf{x}$ lies on S^2 so does its image $r(\mathbf{x})$, and so

$$(\mathbf{x} + t\mathbf{a}) \bullet (\mathbf{x} + t\mathbf{a}) = 1.$$

This implies that

$$\mathbf{x} \bullet \mathbf{x} + 2t\,\mathbf{x} \bullet \mathbf{a} + t^2\mathbf{a} \bullet \mathbf{a} = 1,$$

so that

$$2t\,\mathbf{x} \bullet \mathbf{a} + t^2 = 0.$$

For $\mathbf{x} \bullet \mathbf{x} = 1$ and $\mathbf{a} \bullet \mathbf{a} = 1$.

Now, if $\mathbf{x} \bullet \mathbf{a} = 0$, then the point $\mathbf{x}$ lies in the plane Π in which we are reflecting, and then as you would expect $t = 0$ and $r(\mathbf{x}) = \mathbf{x}$.

For then clearly $r(\mathbf{x}) = \mathbf{x}$.

If $\mathbf{x} \bullet \mathbf{a} \neq 0$, then we may exclude the possibility that $t = 0$. It follows from the equation $2t\mathbf{x} \bullet \mathbf{a} + t^2 = 0$ that $t = -2\mathbf{x} \bullet \mathbf{a}$, and so

$$r(\mathbf{x}) = \mathbf{x} - 2(\mathbf{x} \bullet \mathbf{a})\mathbf{a}.$$

To represent this mapping via a matrix, we express this equation in terms of coordinates and group the terms carefully:

$$\begin{aligned} r(x, y, z) &= (x, y, z) - 2(ax + by + cz)(a, b, c) \\ &= \Big(\left(1 - 2a^2\right)x - 2aby - 2acz,\ -2abx + \left(1 - 2b^2\right)y \\ &\qquad - 2bcz,\ -2acx - 2bcy + \left(1 - 2c^2\right)z \Big), \end{aligned}$$

so that it can be expressed in matrix form as

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} 1 - 2a^2 & -2ab & -2ac \\ -2ab & 1 - 2b^2 & -2bc \\ -2ac & -2bc & 1 - 2c^2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

Theorem 3 Reflection of S^2 in the plane Π with equation $ax + by + cz = 0$, where $a^2 + b^2 + c^2 = 1$, is given by the mapping $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ where

$$\mathbf{A} = \begin{pmatrix} 1 - 2a^2 & -2ab & -2ac \\ -2ab & 1 - 2b^2 & -2bc \\ -2ac & -2bc & 1 - 2c^2 \end{pmatrix}.$$

Notice that in the particular case when the plane is the equatorial plane, with equation $z = 0$ (so that $a = b = 0$ and $c = 1$), Theorem 3 gives the matrix for the mapping $\mathbf{x}$ $\mathbf{Ax}$ to be

$$\mathbf{A} = \begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{matrix}$$

that you met above.

> **Problem 8** Determine the matrix $\mathbf{A}$ such that the mapping $\mathbf{x}$ $\mathbf{Ax}$ represents reflection of S^2 in the plane with equation $3x + 4y \quad 5z = 0$.

Remark

There is another way to find the matrix representing a reflection in a great circle. Let the point P be one of the polar points of the plane through O that meets S^2 in the given great circle. Let R be a rotation taking P to the North Pole N, and let r be reflection in the equatorial plane of S^2. Then the composition mapping $R\ ^{1}rR$ represents reflection of S^2 in .

We omit the details of this calculation.

We now establish a result relating products of reflections of S^2 to rotations of S^2. It is the exact analogue of a result in Euclidean geometry. The result holds also for general reflections and rotations of R^3, so we prove it in that more general setting.

Subsection 5.2.1, margin note alongside Problem 4

> **Theorem 4** The product of any two reflections of R^3 in planes through O that meet in a common line is a rotation about that common line.

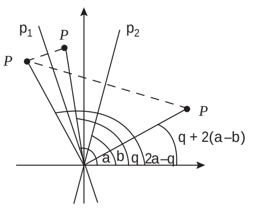

Proof By choosing coordinate axes suitably, we may arrange that the common line is the z-axis. The planes are therefore vertical, and it follows that each reflection leaves the z-coordinate of each point unaltered. It is therefore sufficient to look at the effect of each reflection and composite on planes perpendicular to the z-axis; we shall discuss the (x, y)-plane, but a similar argument applies to all planes parallel to this one.

Let the first plane $_1$ make an angle with the positive x-axis and the second plane $_2$ make an angle with the positive x-axis. Let a point P in the (x, y)-plane have plane polar coordinates $(r, \ )$. Then the image of P under reflection in $_1$ is the point P with polar coordinates $(r, 2 \quad)$. The image of P under the reflection in $_2$ is the point P with polar coordinates $(r, 2 \quad (2 \quad)) = (r, \ + 2(\quad))$. This is the image of P under rotation about the common line of $_1$ and $_2$, through twice the angle between the planes.

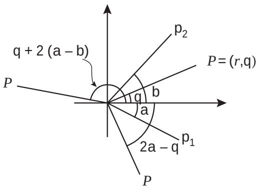

We deduce from this theorem that the composite of reflections in two great circles on S^2 is a rotation of S^2 about the line common to the planes defining the great circles, and through twice the angle between them. We can also deduce that every rotation is a product of two reflections; these reflections are in planes that have the axis of rotation as their common line, and are separated by half the angle of the rotation (but they are otherwise arbitrary).

Example 3 Prove that the image of a circle C on S^2 under an isometry t of S^2 to itself is a circle.

Solution Let the circle C be cut out by a plane , and let be the line through the origin perpendicular to . Let meet S^2 at the points P and P. Then all points of C are the same distance from P.

Since t is an isometry, the points of $t(C)$ are all the same distance from $t(P)$. It follows that they lie on a circle, which must be the circle cut out on S^2 by the plane $t(\ )$. □

We now establish the following important result.

Theorem 5 The Product Theorem
Every isometry of S^2 is a product of at most three reflections in great circles.

Proof The key idea in this proof is the observation that *an isometry of S^2 is known completely when its effect on three points (which do not lie on a common great circle) is known*. We shall prove this observation in the course of proving the theorem.

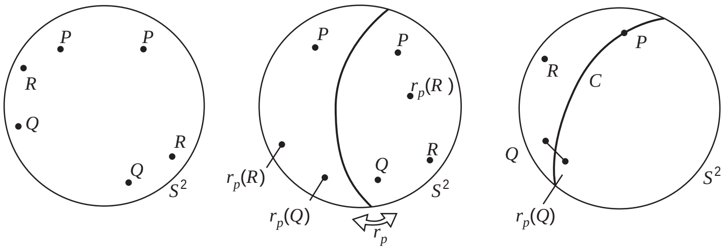

Let P, Q and R be three points not lying on the same great circle, and let P, Q and R be their images under an isometry t of S^2 to itself. Let r_P be a reflection of S^2 in a great circle that sends P to P. Consider the angle in $\mathbb{R}^3$ with arms $P\ r_P(Q)$ and $P\ Q$, and let C be the great circle that bisects this angle and the segment joining $r_P(Q)$ and Q. Reflection in C fixes P and sends $r_p(Q)$ to Q. Suppose that this reflection sends $r_P(R)$ to R.

See left figure above.
See centre figure above.
See right figure above.

The distance of R from P is the same as the distance RP (call it d_P), and the distance of R from Q is the same as the distance QR (call it d_Q), since both reflections are isometries. So the point R lies on the intersection of the circle with centre P and radius d_P and the circle with centre Q and radius d_Q; so too does the point R.

Then it may be that R and R coincide, in which case we have found a product of two reflections sending R to R. Or it may be that R and R do not coincide but a reflection in the great circle through P and Q maps R to R, and so three reflections have been used.

We now claim that any isometry fixing three points on S^2 fixes every point of S^2. Let P, Q and R be the points, let M be any other point on S^2, and let t be the isometry fixing the points P, Q and R. Then $t(M)$ lies on the circle with centre P and radius PM and on the circle with centre Q and radius QM. Of the two points satisfying this condition, only one also lies on the circle with centre R and radius RM, and that is the point M itself; so the image of M under t is M, and therefore t fixes every point of S^2. It follows that if two isometries t_1 and t_2 agree on three points, then they agree on every point, since the composite t_2^{1} t_1 satisfies the conditions of our last remark.

We deduce that the isometry mapping the points P, Q and R to the points P, Q and R must be the product of the two or three reflections that we described above. This completes the proof.

It does not follow that these reflections are uniquely determined recall the case of the rotation discussed above.

Finally in this subsection we show that the Isosceles Triangle Theorem in spherical geometry can be established using a reflection in a great circle.

Theorem 6 Isosceles Triangle Theorem
Let a triangle PQR on S^2 have sides PQ of length r, QR of length p, and RP of length q, and let $p = q$. Also, let the angles at the vertices P, Q and R be , and , respectively. Then = .

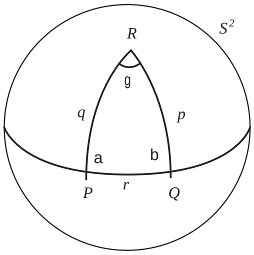

Proof Consider the reflection in the great circle that is the internal bisector of the angle PRQ.

It maps the great circle through R and P to the great circle through R and Q; and, since it is an isometry and $p = q$, it therefore maps P to Q, and Q to P. It maps the great circle through P and Q to itself, and so it interchanges the angles and , which shows that = .

7.2.3 Direct Isometries

The Product Theorem allows us to deduce that every isometry is either a reflection, a rotation or the composite of a rotation and a reflection. For if it is a composite of two reflections, then it is a rotation; and if it is a composite of three reflections, then the composite of the first two is a rotation.

Theorem 5

It is clear that every rotation of S^2 is an direct isometry of S^2.

We shall now see that every direct isometry of S^2 is a product of rotations. For each reflection reverses the orientation of angles, and so the product of two reflections preserves the orientation of angles, while the product of three reflections reverses it again. So a direct isometry must be the product of an *even* number of reflections and therefore be a rotation.

For the product of any pair of reflections is a rotation.

It is instructive to give a direct proof of this result that does not use the Product Theorem, because reflections are not easy to handle geometrically or algebraically.

First, any direct isometry of S^2 that fixes the North Pole N must map each circle of latitude to itself, and so must be a rotation of that circle. Let D and E be points on different circles of latitude; they remain the same distance apart, so the rotation of their two circles of latitude must be through the same angle. It follows that every circle of latitude on S^2 is rotated equally, and so the isometry fixing N must be a rotation about ON.

Next, let M be a direct isometry that sends a point P to P and a point Q to Q, and let R be a rotation that sends P to P. Then $M \quad R^{\ 1}$ is an isometry that fixes P. We may assume that $P = N$ (if need be by additionally composing with a rotation that sends P to N). Then the isometry $M \quad R^{\ 1} = R(Z, \ )$, by the remarks in the previous paragraph; so the direct isometry $M = R(Z, \ ) \quad R$ is a product of two rotations. We deduce that *every* direct isometry is a product of rotations.

We shall now see that every direct isometry of S^2 is a single rotation.

Since every 3 × 3 matrix with real entries (such as a product of rotations must be) must have a real eigenspace, then any such 3 × 3 matrix describing an isometry of S^2 to itself must leave the points where its eigenspace meets S^2 fixed. Since it is also a direct symmetry, it must therefore be a rotation about those points. It follows that every direct isometry of S^2 to itself is simply a rotation.

See Appendix 2.

We can deduce from this also the fact that the composite of two rotations of S^2 is also a rotation of S^2, since the composite is itself a direct isometry.

We summarize these facts as follows.

Theorem 7

(a) Every isometry of S^2 is a reflection, a rotation, or the composite of a reflection and a rotation.
(b) Every direct isometry of S^2 is a rotation, and vice versa.

Problem 9 Prove that the composite of an even number of reflections of S^2 in great circles is a rotation of S^2.

7.3 Spherical Trigonometry

7.3.1 Spherical Triangles

Recall that a triangle on S^2, or a *spherical triangle*, consists of the arcs of three great circles joining three points on S^2. We first establish that spherical triangles exist with any given angles; more precisely, we establish the following result.

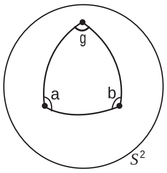

Theorem 1 Given any three angles , and with $+ \quad + \quad >$, there exists a spherical triangle with those angles.

In fact we also require that $+ \quad + \quad < 5$ in view of Theorem 3 below but we shall not go into this here.

Proof We use the idea of stereographic projection, under which circles on S^2 correspond to generalized circles (that is, lines or circles) in the extended plane.

Theorem 7 of Subsection 5.2.4 (in which stereographic projection was introduced).

By a rotation of S^2, if necessary, we may assume that one of the vertices of a triangle on the unit sphere is at the South Pole, S.

Under stereographic projection a triangle on S^2 with one vertex at the South Pole S corresponds to a circular-arc triangle in the plane (that is, a triangle in $\mathbb{C}$ whose sides are arcs of generalized circles) with one vertex at O, the centre of the unit circle $\mathbb{C}$, and two straight sides. It follows that to prove the theorem it is enough to construct a triangle of this form in $\mathbb{C}$ that has the specified angles. This we now do; in our construction we assume that and lie between 0 and $\frac{}{2}$, for simplicity.

Recall that stereographic projection is angle-preserving (Theorem 8 of Subsection 5.2.4).

Consider the circular-arc triangle ABP in $\mathbb{C}$ that has angles $\frac{}{2}$ at A, $\frac{}{2}$ at B, and at P, where PA and PB are straight lines (so that P is the centre of the circle through A and B). Let D be the point on PB such that the angle BAD equals , let C be the point on PA such that the angle ABC equals , let AD and BC meet at O, and let the angle AOB equal x (as shown). Our first task is to evaluate x.

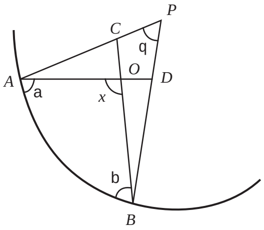

Both the angles AOC and BOD are equal to x. By considering the triangles AOC and BOD in turn, we deduce that

$$PCO = (\quad x) + \frac{}{2} \quad = \frac{3}{2} \quad x \quad ,$$

and

$$PDO = (\quad x) + \frac{}{2} \quad = \frac{3}{2} \quad x \quad .$$

Then the sum of the angles of the quadrilateral $PCOD$ is

$$2 \quad = \quad + \frac{3}{2} \quad x \quad + x + \frac{3}{2} \quad x \quad ,$$

so that

$$x = \quad + \quad .$$

It follows that $x =$ if and only if

$$= \quad + \quad + \quad .$$

Hence if we repeat the construction with the value $+ \quad +$ for , the circular-arc triangle ABO has the desired angles , and . It follows from the earlier discussion that by stereographic projection we can then construct the required spherical triangle on S^2 with these angles.

This is possible, since by hypothesis $+ \quad + \quad > \quad .$

A modification of the above argument gives the following important result, which we state without proof.

Theorem 2 The sum of the angles of any spherical triangle is greater than π.

In view of Theorems 1 and 2, it is useful to assign a name to the difference between the sum of the angles of a spherical triangle and π: this measures how far the triangle departs from being Euclidean.

Definition The **angular excess** of a spherical triangle is the difference between the sum of the angles of the triangle and π.

Problem 1 Construct spherical triangles with angular excess of

(a) 0.01 radians, (b) 3.14 radians.

Subsection 7.1.1

Recall that earlier when we defined a distance PQ between two given points P and Q on S^2, we noted that there were two 'line segments' PQ. Now, for definiteness, we call the shorter of these two the **strict line segment** PQ. (If both are of length π, we choose either in some explicit way.)

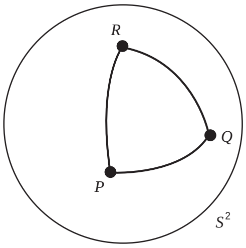

Similarly, when we defined a triangle △PQR with three given vertices P, Q and R on S^2, we noted that there were several such triangles. For definiteness, we now call the triangle whose sides are strict line segments the **strict triangle** △PQR. Then it is possible to rotate the sphere in such a way that the whole of the strict triangle can be seen at one time. (You should convince yourself that this is the case.) The **inside** of the strict triangle △PQR is then defined to be the region of S^2 bounded by the three strict line segments. The **outside** is the complement of the inside.

We now investigate the regions into which the corresponding great circles defining the strict triangle △PQR divide S^2. Notice first that the equatorial plane defining each of the relevant great circles cuts S^2 into two regions. A second great circle (and equatorial plane) divides each of these regions into two, giving four regions. Finally a third great circle (and equatorial plane) divides each of these regions into two, giving eight regions: so there are eight spherical triangles with given vertices P, Q and R on S^2.

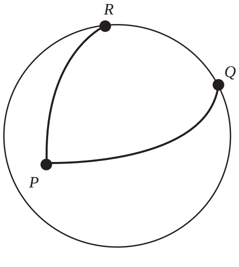

Now let us look at the strict triangle △PQR more closely. For definiteness, we will assume that the vertices P, Q and R are chosen in clockwise order as viewed by an observer at the centre of the sphere. Rotate the sphere to move R to the North Pole, and Q to some point in the (y, z)-plane. By our choice of the orientation of △PQR and the fact that the line segments PR and PQ are strict line segments, it follows that this rotation leaves P on 'the front' of the sphere, where we can see it.

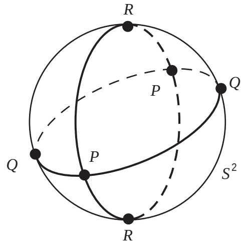

The great circle through R and Q is the 'rim' of S^2 that we can see. So the great circle RP meets it again at R', say, the diametrically opposite point to R; and the great circle QP meets it again at Q', say, the diametrically opposite point to Q. Similarly the great circles PR and PQ meet again at P', say, the diametrically opposite point to P.

So, associated with the strict triangle PQR in this particular position on S^2 there are three great circles, each meeting in two diametrically opposite points; a total of six points in all. The great circles define eight triangles:

the strict triangle PQR and the triangles $PQ\ R$, $PQ\ R$ and $PR\ Q$ (all visible on the front of S^2), and
four triangles $P\ QR$, $P\ Q\ R$, $P\ Q\ R$ and $P\ R\ Q$ round the back .

In fact, each of the eight triangles has a diametrically opposite triangle to which it is congruent (under the map (x, y, z) $(\ x,\ y,\ z)$). For example, the triangle PQR is diametrically opposite to the triangle $P\ Q\ R$.

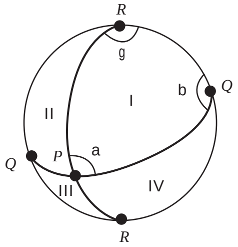

In the figure in the margin we have marked the four visible triangles PQR, $PQ\ R$, $PQ\ R$ and $PR\ Q$ with Roman numerals I, II, III and IV, respectively. Similarly, we denote their diametrically opposite triangles $P\ Q\ R$, $P\ QR$, $P\ QR$ and $P\ RQ$ with Roman numerals I , II , III and IV , respectively.

In order to label equal angles conveniently, let us denote by the angle QPR in the strict triangle PQR, by the angle PQR, and by the angle QRP. A useful observation is that two great circles meet at the same angles at each of their points of intersection; so, for example, the angles QPR and $Q\ P\ R$ are equal.

Example 1 Show that the triangles PQR and $P\ Q\ R$ have their angles equal in pairs.

Solution The angle QPR at P in PQR is formed by two great circles that meet again at P , forming the angle $Q\ P\ R$ at P in $P\ Q\ R$, so they must be equal.

A similar argument shows that the angles at vertices Q and Q , and the angles at vertices R and R , are also equal. □

Problem 2 Determine the angles in each of triangles II, III, and IV, in terms of , and .

Area of a Triangle

We next find the area of a spherical triangle in terms of its angles. In this discussion, we introduce the term **lune** to denote a two-sided polygon defined by two great circles, choosing arbitrarily which of the four such regions on the sphere we mean.

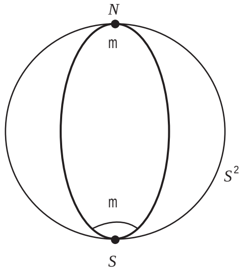

A lune with vertices at N and S, and angle μ.

The area of a lune is clearly proportional to the corresponding angle between the two great circles, the *angle of the lune.* Now when the lune has angle $\frac{\ }{2}$, its area is (for the lune is then one-quarter of S^2, which has area 4); so by proportion the area of a lune of angle μ, say, is 2μ.

The front of S^2 as shown in the margin consists of two lunes with vertices R and R , each formed of two strict triangles, I + IV and II + III (using the above notation for triangles). The lune I + IV has area 2 , since it has angle . With a slight abuse of notation we express this fact about areas in the form

$$\text{I} + \text{IV} = 2 \; .$$

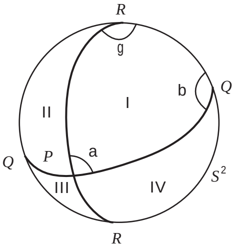

A similar argument involving the lune with vertices Q and Q yields the equation

$$\text{I} + \text{II} = 2 \; ,$$

and then with the lune with vertices P and P yields the equation

$$\text{I} + \text{III} = 2 \; .$$

Since triangles III and III are congruent, we may rewrite this last equation as

$$\text{I} + \text{III} = 2 \; .$$

Adding the three equations, we obtain

$$3 \bullet \text{I} + \text{II} + \text{III} + \text{IV} = 2(\; + \; + \;).$$

Also, the area of the front of S^2 is

$$\text{I} + \text{II} + \text{III} + \text{IV} = 2 \; .$$

It follows that

$$2 \bullet \text{I} + 2 \; = 2(\; + \; + \;)$$

or

$$\text{I} = (\; + \; + \;) \qquad .$$

This completes the proof of the following theorem for triangles, in the special case that they are strict triangles.

The general case is proved by dividing an arbitrary triangle into strict triangles, and applying the special case to each; we omit the details.

Theorem 3 The area of a spherical triangle is equal to its angular excess.

Problem 3 Construct a spherical triangle of area $\frac{3}{4}$.

7.3.2 Dual Triangles

The somewhat oddly nature of the definition of a strict triangle pays off handsomely in a feature of spherical geometry that was often exploited in the 18th century by Euler and Lagrange, two of the pioneer figures in spherical geometry.

We required that the sides of a strict triangle be less than or equal to in length; for simplicity in what follows, we shall ignore the case of equality.

Let A, B and C be the vertices of a strict triangle, with angles , and at the vertices, respectively, and opposite sides of length a, b and c, respectively. Each side and angle of the triangle ABC has magnitude between 0 and .

We now construct a new spherical triangle $A B C$, called the **dual** of the original triangle ABC, with vertices A, B and C, as follows. Its angles at A, B and C are a, b and c, respectively; and its sides are $A B$ of length , $B C$ of length , and $C A$ of length .

One way of constructing such a triangle is as follows. Extend the segments AB and AC to segments AD and AD, each of length $\frac{}{2}$, and draw the great circle through D and D (this is the equator for A). Similarly, extend the segments BC and BA to segments BE and BE, each of length $\frac{}{2}$, and draw the great circle through E and E (this is the equator for B); and extend the segments CA and CB to segments CF and CF, each of length $\frac{}{2}$, and draw the great circle through F and F (this is the equator for C). Then either of the strict triangles defined by the great circles DD, EE and FF is a spherical triangle with the desired properties.

For example, choose the one that overlaps ABC, and denote its vertex on EE and FF by A, its vertex on FF and DD by B, and its vertex on DD and EE by C. We now check that the triangle $A B C$ is indeed dual to the original triangle ABC.

The lengths of CF and CB are both $\frac{}{2}$, since $A FF B$ is the equator for C. But B is on the equator $B DD C$ for A, so $B A = \frac{}{2}$. Thus A and C are on the equator for B, so that the great circle $FACD$ defined by A and C must be the equator for B. This means that D lies on the equator for B, so $B D = \frac{}{2}$. Similarly $DC = \frac{}{2}$.

Now consider the equator $B DD C$ for A, and denote by O the centre of the sphere. The angle DOD is equal to $BAC =$ (since both DOD and BAC are just the angle between the planes OAB and OAC); thus DD lies on a (great) circle of radius 1, and so the length of DD is also (since DD is equal to DOD). It follows that the length of $B C$ is

$$B C = B D + DC \quad DD = \frac{}{2} + \frac{}{2} \quad = \quad .$$

Similarly the length of $C A$ is , and the length of $A B$ is .

Next, by construction $CF = \frac{}{2}$ and $BE = \frac{}{2}$. It follows that the length of $F E$ is

$$F E = F C + BE \quad BC = \frac{}{2} + \frac{}{2} \quad a = \quad a.$$

Hence the angle $F OE$ equals a (since $F BCE$ is a circle of radius 1 with centre O and hence $F OE$ equals the length of $F E$), so that the angle $F A E$ also equals a (since both are just the angle between the planes $OA F$ and $OA E$). It follows that the angle $B A C$ equals a, as required. Similarly, $A B C =$ b and $B C A =$ c.

We have therefore established the following theorem.

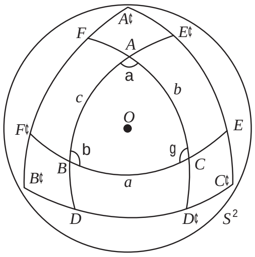

$A B C$ is dual to ABC.

Theorem 4 Dual Triangles Theorem

With every strict spherical triangle there is associated another triangle , called its *dual*, whose angles are the complements of the sides of and whose sides are the complements of the angles of .

Here by the complement of a quantity x we mean the quantity x.

Here is an example of how the theorem proves useful. Let ABC be a strict triangle with $ABC = ACB$. Then, if $A\ B\ C$ is its dual (which exists by Theorem 4), we have $A\ C = A\ B$; it then follows from the Isosceles Triangle Theorem that $A\ B\ C = A\ C\ B$. It then follows by duality that $AC = AB$. This is the converse of the Isosceles Triangle Theorem, in the case of strict triangles.

Section 7.2.2, Theorem 6

We ask you to prove the general case of the converse in Section 7.6.

Problem 4 Let A, B and N be the points (1, 0, 0), (0, 1, 0) and (0, 0, 1) on S^2. Prove that the triangle ABN is congruent to any of its duals.

7.3.3 Distance between Two Points

In real life we are often interested in the (spherical) distance between points on the surface E of the Earth, whose radius is about 4000 miles. The comparable problem for S^2 is to find a formula for the distance between two points of S^2. Then a scaling map

$$\begin{aligned} t : S^2 &\qquad E \\ \mathbf{x} &\qquad 4000\mathbf{x} \end{aligned}$$

provides an answer to the original question concerning E.

In metric units, the radius of the Earth is about 6378 km.

Here we take the unit of measurement for the radius S^2 to be 1 mile.

Let P and Q be points on S^2 with Cartesian coordinates (p_1, p_2, p_3) and (q_1, q_2, q_3), respectively, whose distance apart on S^2 is d. It follows that, if O denotes the centre of S^2, the angle POQ also equals d.

Then the dot product of the vectors OP and OQ is

$$OP \bullet OQ = p_1q_1 + p_2q_2 + p_3q_3 = 1 \bullet 1 \bullet \cos d,$$

so that $\cos d = p_1q_1 + p_2q_2 + p_3q_3$.

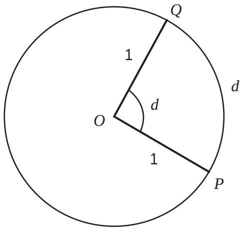

great circle

Theorem 5 The distance d on S^2 between the points (p_1, p_2, p_3) and (q_1, q_2, q_3) is given by

$$\cos\ d = p_1q_1 + p_2q_2 + p_3q_3.$$

In the expression $\cos d$, d must be taken in radians, as usual.

In applications to E, we must remember that if a point has latitude and longitude then it has colatitude $= \frac{\ }{2}$ and Cartesian coordinates (cos sin , sin sin , cos).

The latitudes and longitudes of various cities are approximately as follows.

City	Latitude	Longitude
New York	41	74
Rio de Janeiro	23	43
Sydney	34	151
Tokyo	36	140

Remember that (geographers) longitude West of the Greenwich meridian is *negative*.

Example 2 Estimate the distance between New York and Sydney, taking the radius of the Earth as 4000 miles.

Solution The colatitudes of New York and Sydney are $90 \quad 41 = 49$ and $90 + 34 = 124$, so that the coordinates of the corresponding points P and Q on S^2 are

$$\begin{aligned}&(\cos(\quad 74)\sin(49), \sin(\quad 74)\sin(49), \cos(49))\\ &\quad (0.2756 \bullet 0.7547, \quad 0.9613 \bullet 0.7547, 0.6561)\\ &\quad (0.208, \quad 0.725, 0.656),\end{aligned}$$

and

$$\begin{aligned}&(\cos(151)\sin(124), \sin(151)\sin(124), \cos(124))\\ &\quad (\quad 0.8746 \bullet 0.8290, 0.4848 \bullet 0.8290, \quad 0.5592)\\ &\quad (\quad 0.725, 0.402, \quad 0.559).\end{aligned}$$

Hence the distance between New York and Sydney is approximately $4000d$ miles, where

$$\begin{aligned}\cos d &\quad 0.208 \bullet 0.725 \quad 0.725 \bullet 0.402 \quad 0.656 \bullet 0.559\\ &\quad 0.151 \quad 0.291 \quad 0.367 = \quad 0.809.\end{aligned}$$

Thus $d \quad \cos^{\ 1}(\quad 0.809) \quad 2.513$ radians, and the required distance $4000d$ $\quad 4000 \bullet 2.51 \quad 10\,050$ miles. □

Problem 5 Estimate the distance between Tokyo and Rio de Janeiro, taking the radius of the Earth as 4000 miles.

Right-Angled Triangles

It only makes sense to discuss Pythagoras Theorem in those geometries with a *distance.* We have already seen Pythagoras Theorem in hyperbolic geometry, so what is the analogue in spherical geometry?

Theorem 8 of Subsection 6.4.3

Theorem 6 Pythagoras' Theorem
Let ABC be a triangle on S^2 in which the angle at C is a right angle. If a, b and c are the lengths of BC, CA and AB, then

$$\cos c = \cos a \times \cos b.$$

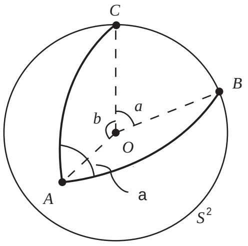

Proof We may rotate S^2 so that C is the North Pole $N \ (0, 0, 1)$ of S^2, A lies on the Greenwich meridian and has coordinates $(\sin b, 0, \cos b)$, and B has coordinates $(0, \sin a, \cos a)$.

We then apply the rotation $R(Y, \quad b)$ to S^2. This maps A to C, C to some point C , and B to some point B .

Since the sphere has radius 1, at A, $\ = b$ and $\ = 0$; and at B, $\ = \frac{}{2}$ and $\ = a$.

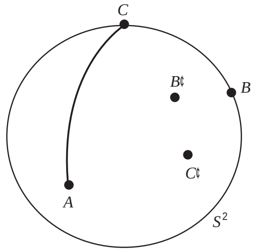

Let denote the angle CAB. Then the rotation $R(Y,\ \ b)$ maps the angle CAB to the angle $C\ CB$, so that the angle $C\ CB\ =\ $; it follows that the angle $ACB\ =\ $ since $ACB\ +\ \ B\ CC\ =\ $. Hence the longitude of B is . Also, the colatitude of B is

$$\text{length of CB} = \text{length of AB} = c.$$

So the coordinates of B are

$$\begin{aligned}&(\cos(\quad)\sin c, \sin(\quad)\sin c, \cos c)\\ &= (\ \ \cos\ \ \sin c,\ \sin\ \ \sin c, \cos c).\end{aligned} \qquad (1)$$

But we also have as coordinates of B

$$R(Y,\ \ b)(B) = \begin{matrix}\cos b & 0 & \sin b\\ 0 & 1 & 0\\ \sin b & 0 & \cos b\end{matrix} \qquad \begin{matrix}0\\ \sin a\\ \cos a\end{matrix}$$

$$= \begin{matrix}\sin b\cos a\\ \sin a\\ \cos b\cos a\end{matrix} \quad \bullet \qquad (2)$$

Here we use the formulas for coordinates under isometries given in Subsection 7.2.1.

Equating the third coordinates in (1) and (2), we get the formula

$$\cos c = \cos a \times \cos b,$$

as required.

For example, if $a = 1, b = 2$ and $ACB = \frac{}{2}$, then

$$\cos c = \cos 1 \times \cos 2 \quad 0.5403 \bullet (\ \ 0.4161) \quad 0.2248,$$

so that c 1.798.

Problem 6 PQR is a triangle on S^2 in which the angle at R is a right angle and the lengths of PQ and PR are 1.7 and 1.9, respectively. Estimate the length of QR.

Remark

For all values of x, $\cos x = 1 \quad \frac{x^2}{2!} +$ higher powers of x; so, when x is small, $\cos x \quad 1 \quad \frac{1}{2}x^2$. It follows that, if we are dealing with small triangles on S^2, then the conclusion of Theorem 6 may be interpreted in the form

$$1 \quad \tfrac{1}{2}c^2 \qquad 1 \quad \tfrac{1}{2}a^2 \bullet 1 \quad \tfrac{1}{2}b^2 \qquad 1 \quad \tfrac{1}{2}a^2 \quad \tfrac{1}{2}b^2,$$

Here we drop the term in a^2b^2, since it is smaller than the remaining terms.

so that $c^2 \quad a^2 + b^2$. Thus Theorem 6 is a close approximation to the Euclidean version of Pythagoras Theorem in $\mathbb{R}^2$.

The area formula for S^2 suggests that the angles of a spherical triangle determine its size as well as its shape. In fact, any three pieces of the six possible

Theorem 3

pieces of information about a triangle (its three angles and the lengths of its three sides) determine the remaining three.

We prove this fact here for right angled triangles; a proof for general triangles can then be constructed by dividing a general triangle into two right angled triangles using an altitude.

Returning to the formula in the proof of Theorem 6 for the coordinates of B and equating the first and second coordinates, we obtain

$$\cos\ \sin c = \sin b \cos a \quad \text{and} \quad \sin\ \sin c = \sin a.$$

Dividing the second of these by the first, we get $\tan\ = \tan a / \sin b$.

Re-arranging the above formulas to make them easier to remember, we have the following result.

Theorem 7 Let ABC be a triangle on S^2 in which the angle at C is a right angle. If a, b and c are the lengths of BC, CA and AB, and denotes the angle CAB, then

$$\cos c = \cos a \times \cos b,$$

$$\sin\ = \frac{\sin a}{\sin c}, \quad \text{and} \quad \tan\ = \frac{\tan a}{\sin b}.$$

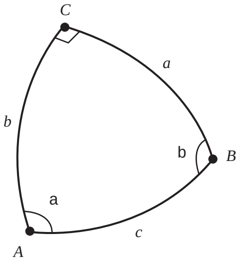

In the proof of Theorem 6 and before Theorem 7

A similar discussion to that above gives that

$$\cos\ \sin c = \sin a \cos b \quad \text{and} \quad \sin\ \sin c = \sin b,$$

so that

$$\sin\ = \frac{\sin b}{\sin c} \quad \text{and} \quad \tan\ = \frac{\tan b}{\sin a}.$$

From these formulas and those in Theorem 7 above, it follows that in a right-angled triangle two further pieces of information about the triangle enable us to determine all its angles and sides.

> **Problem 7** PQR is a triangle on S^2 in which the angle at R is a right angle and the lengths of PQ and PR are 1.7 and 1.9, respectively. Estimate the angle QPR.

7.3.4 Spherical Trigonometry

In this subsection we deduce various formulas of spherical trigonometry for strict triangles by dropping perpendiculars from a vertex of a spherical triangle to the opposite side, giving two right-angled ones for which we already have formulas. First, we verify that we can always drop a perpendicular from a point to a line. It follows that a spherical triangle has 3 altitudes.

Given a strict segment AB and a point C on S^2 that form a triangle ABC, there is always a great circle through C perpendicular to AB. This great circle meets the great circle of which AB is part in two points, which we call D and D. Since D and D are antipodal points, they are a distance apart, so one of three things must happen.

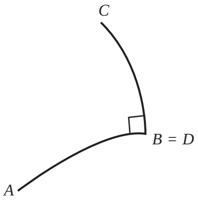

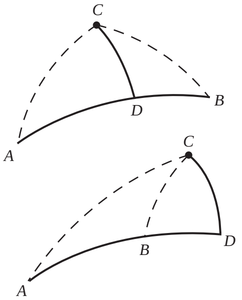

1. One of D and D coincides with A or B. The corresponding strict line segment is the **perpendicular** from C to AB. This case will not lead us to any new results, so we do not consider it further.
2. One of D and D, say D, lies between A and B. Then we call CD the **perpendicular** from C to AB; it lies inside ABC.
3. One of D and D, say D, is such that the distance AD is less than and B lies between A and D. Then we again call CD the **perpendicular** from C to AB; in this case it lies outside ABC.

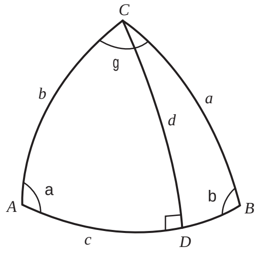

Now consider a strict spherical triangle ABC, in which the sides AB, BC and CA have lengths c, a and b, respectively, and the angles at A, B and C are , and , respectively. Drop the perpendicular CD from C to AB, and let d denote the length of CD.

If we apply the second formula in Theorem 7 to triangle ADC we get

$$\sin \quad = \frac{\sin d}{\sin b},$$

so that $\sin d = \sin \quad \times \sin b$. Applying a similar argument to triangle BDC we get $\sin d = \sin \quad \times \sin a$.

Thus, $\sin \quad \times \sin b = \sin \quad \times \sin a$, which is more memorable in the form

$$\frac{\sin}{\sin a} = \frac{\sin}{\sin b}.$$

Then, by a similar argument involving the perpendicular from A to BC, we obtain the following result.

Theorem 8 Sine Rule for a Strict Spherical Triangle
Let ABC be a strict triangle on S^2, in which the sides AB, BC and CA have lengths c, a and b, respectively, and the angles at A, B and C are , and respectively. Then

$$\frac{\sin}{\sin a} = \frac{\sin}{\sin b} = \frac{\sin}{\sin c}.$$

We can use an argument similar to that in the proof of Pythagoras Theorem to find further results.

Theorem 6

Let ABC be a strict triangle on S^2. We may rotate S^2 so that C is the North Pole $N(0, 0, 1)$, A lies on the Greenwich meridian and has coordinates $(\sin b, 0, \cos b)$, and B has coordinates $(\cos \quad \sin a, \sin \quad \sin a, \cos a)$.

We then apply the rotation $R(Y, \quad b)$ to S^2. This maps A to C, C to some point C, and B to some point B. As we saw earlier, the coordinates of B are $(\quad \cos \quad \sin c, \sin \quad \sin c, \cos c)$.

In the proof of Theorem 6

But we also have as coordinates of B

$$R(Y,\quad b)(B) = \begin{pmatrix} \cos b & 0 & \sin b \\ 0 & 1 & 0 \\ \sin b & 0 & \cos b \end{pmatrix} \begin{pmatrix} \cos \quad \sin a \\ \sin \quad \sin a \\ \cos a \end{pmatrix}$$

$$= \begin{pmatrix} \cos b \cos \quad \sin a \quad \sin b \cos a \\ \sin \quad \sin a \\ \sin b \cos \quad \sin a + \cos b \cos a \end{pmatrix}.$$

Equating the third coordinates in the two expressions for B we get

$$\cos c = \sin b \cos \quad \sin a + \cos b \cos a.$$

Theorem 9 Cosine Rule for Sides of a Strict Spherical Triangle
Let ABC be a strict triangle on S^2, in which the sides AB, BC and CA have lengths c, a and b, respectively, and the angles at A, B and C are , and , respectively. Then

$$\cos c = \cos a \cos b + \sin a \sin b \cos \quad .$$

Notice that this formula reduces to Pythagoras Theorem when is a right-angle.

Example 3 Let ABC be a strict isosceles triangle on S^2, in which the sides AB, BC and CA have lengths c, a and a, respectively, and the angles at A, B and C are , and , respectively. Prove that

(a) $\cos c = \cos^2 a + \sin^2 a \cos \quad$; (b) $\cos \quad = \dfrac{\tan \frac{1}{2}c}{\tan a}$.

Solution Putting $b = a$ in the conclusion of Theorem 9, we obtain the formula (a).

Next, by applying Theorem 9 to find $\cos a$, we get

$$\cos a = \cos a \cos c + \sin a \sin c \cos \quad .$$

We may then rearrange this formula to get

$$\cos \quad = \frac{\cos a (1 \quad \cos c)}{\sin a \sin c}$$

$$= \frac{\cos a \bullet 2 \sin^2 \frac{1}{2}c}{\sin a \bullet 2 \sin \frac{1}{2}c \cos \frac{1}{2}c} = \frac{\tan \frac{1}{2}c}{\tan a},$$

as required. □

Here we use the identities

$$\cos 2x = 1 \quad 2 \sin^2 x$$

and

$$\sin 2x = 2 \sin x \cos x$$

for $x \quad$ R.

Problem 8 Let ABC be a strict equilateral triangle on S^2, in which the sides have length a and the angles are equal to . Prove that

$$\cos \quad = \frac{\cos a}{1 + \cos a}.$$

We can now use any dual triangle $A\ B\ C$ of ABC to prove a new trigonometric result. The dual triangle has sides of length , and , and has angles of magnitude a, b and c. By applying Theorem 9 to the dual triangle and using the formulas $\sin(\ \ x) = \sin x$ and $\cos(\ \ x) = \ \cos x$, we get

You met dual triangles in Subsection 7.3.2.

$$\cos\ \ = \cos\ \ \cos\ \ \ \ \sin\ \ \sin\ \ \cos c;$$

and we can reformulate this formula as follows.

Theorem 10 Cosine Rule for Angles of a Strict Spherical Triangle
Let ABC be a strict triangle on S^2, in which the sides AB, BC and CA have lengths c, a and b, respectively, and the angles at A, B and C are , and , respectively. Then

$$\cos\ \ = \sin\ \ \sin\ \ \cos c\ \ \ \ \cos\ \ \cos\ \ .$$

7.4 Spherical Geometry and the Extended Complex Plane

Earlier we studied the Riemann sphere and stereographic projection of the sphere onto the extended complex plane $\mathbb{C} = \mathbb{C}\ \ \{\ \ \}$.

Subsection 5.2.4

Here we denote the Riemann sphere by S^2 rather than $\mathbb{S}$.

7.4.1 Stereographic Projection

In this section we regard stereographic projection as a mapping of the sphere S^2 onto an extended plane $\mathbb{R}^2\ \ \{\ \ \}$, under which the point (X, Y, Z) maps onto the point $(X, Y, Z) = \ \frac{X}{1\ \ Z}, \frac{Y}{1\ \ Z}$; the inverse mapping is given by

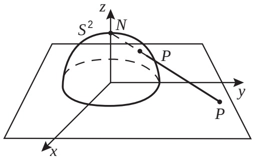

$$^{1}(x, y) = \ \frac{2x}{x^2 + y^2 + 1}, \frac{2y}{x^2 + y^2 + 1}, \frac{x^2 + y^2\ \ \ 1}{x^2 + y^2 + 1}\ \ .$$

Example 1 Let C denote a great circle on S^2 that is the intersection of S^2 with the plane $aX + bY + cZ = 0$, where $a^2 + b^2 + c^2 = 1, c = 0$.

Since $c = 0$, C does not pass through the North Pole N of S^2.

(a) Determine the point N that is the image of N under reflection in C.
(b) Determine the images of N and C in $\mathbb{R}^2$ under stereographic projection .
(c) Verify that $(N\ )$ is the centre of the circle (C).

Solution

(a) The point N has coordinates

We follow the approach of Subsection 7.2.2.

$$\begin{aligned}&(0,0,1)\ \ \ 2\{(0,0,1) \bullet (a,b,c)\}(a,b,c)\\ &= (0,0,1)\ \ \ 2c(a,b,c) = \ \ \ 2ac,\ \ \ 2bc, 1\ \ \ 2c^2\ \ .\end{aligned}$$

(b) From the formula for , the image of N under is

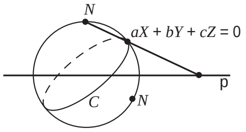

$$(\ \ 2ac,\ \ 2bc, 1\ \ \ 2c^2) = \frac{2ac}{1\ \ \ (1\ \ \ 2c^2)}, \frac{2bc}{1\ \ \ (1\ \ \ 2c^2)}$$

$$= \quad \frac{a}{c}, \ \ \frac{b}{c} \quad .$$

Here we use the fact $c = 0$.

From the formula for ^{1} the image (C) of C has equation

$$a\frac{2x}{x^2+y^2+1} + b\frac{2y}{x^2+y^2+1} + c\frac{x^2+y^2\ \ \ 1}{x^2+y^2+1} = 0;$$

we may rewrite this in the form

$$2ax + 2by + c\ \ \ x^2 + y^2\ \ \ \ 1\ \ \ = 0,$$

or

$$x^2 + y^2 + 2\frac{a}{c}x + 2\frac{b}{c}y\ \ \ \ 1 = 0.$$

Again we use the fact $c = 0$.

(c) Since the equation of the circle (C) may be written in the form

$$x + \frac{a}{c}\ \ ^2 + \ \ \ y + \frac{b}{c}\ \ \ ^2 = 1 + \frac{a^2+b^2}{c^2} = \frac{1}{c^2},$$

the centre of (C) is the point $\frac{a}{c}, \ \ \frac{b}{c}$, that is, the point $(N\ )$. □

Problem 1 Let $P = (x, y)$ be a point in $\mathbb{R}^2$ other than the origin. Let P be the point on S^2 that corresponds to P under stereographic projection, P the reflection of P in the (X, Y)-plane, and Q the point in $\mathbb{R}^2$ that corresponds to P under stereographic projection.

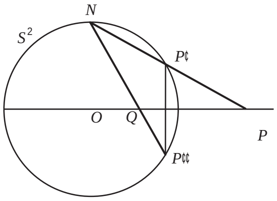

(a) Determine the coordinates of P , P and Q.

(b) Prove that P and Q are inverse points with respect to the unit circle $x^2 + y^2 = 1$ in $\mathbb{R}^2$.

There is a simple and useful expression for the image under stereographic projection of a point on S^2 with given colatitude and longitude.

Example 2 The point P on S^2 with colatitude and longitude has coordinates (cos sin , sin sin , cos). Prove that the image P of P under stereographic projection has coordinates $\frac{\cos}{\tan\ \frac{1}{2}}, \frac{\sin}{\tan\ \frac{1}{2}}$.

Recall that the latitude of P is $= \ \overline{2}$.

Solution From the formula for , the image P of P under is the point

$$\frac{\cos\quad\sin}{1\quad\cos}, \frac{\sin\quad\sin}{1\quad\cos}$$

$$= \frac{\cos\quad \bullet 2\sin\ \frac{1}{2}\quad \cos\ \frac{1}{2}}{2\sin^2\ \frac{1}{2}}, \frac{\sin\quad \bullet 2\sin\ \frac{1}{2}\quad \cos\ \frac{1}{2}}{2\sin^2\ \frac{1}{2}}$$

$$= \frac{\cos}{\tan\ \frac{1}{2}}, \frac{\sin}{\tan\ \frac{1}{2}}\quad ,$$

as required. □

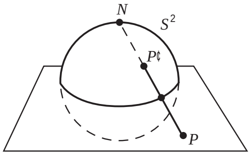

Next, we obtain a neat formula relating the distance on S^2 between the South Pole S and a point P on S^2 to the distance of the corresponding point $P = \quad (P\ )$ from O in $\mathbb{R}^2$.

Let P be a point $(x, 0)$ in $\mathbb{R}^2, x > 0$. The corresponding point $P\ = \quad ^1(P)$ on S^2 has coordinates $\frac{2x}{1+x^2}, 0, \frac{x^2\ 1}{x^2+1}$; it follows from the definition of spherical polar coordinates on S^2 as $(\cos\ \sin\ , \sin\ \sin\ , \cos\ )$, where is colatitude, that the colatitude of P satisfies the equation

$$\tan\quad = \frac{2x}{x^2\quad 1}. \tag{1}$$

Then if the spherical distance of P from S is d, the colatitude of P also satisfies the equation $\quad + d = \quad$, so that

$$\tan\quad = \quad \tan d. \tag{2}$$

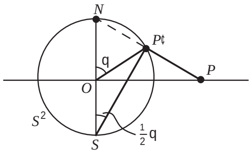

Then, if we denote by T the quantity $\tan\ \frac{1}{2}d$, it follows from equations (1) and (2) that $\frac{2x}{x^2\ 1} = \quad \frac{2T}{1\ T^2}$. We can write this equation in the form $(T\quad x)(1 + xT) = 0$; and, since both x and T are positive it follows that $T = x$. In other words $\tan\ ^1 x = \tan\ ^1 T = \frac{1}{2}d$, so that $d = 2\tan\ ^1 x$.

Since a rotation of S^2 and $\mathbb{R}^2$ about the Z-axis does not change any of the distances in the above argument, we deduce the following result.

Theorem 1 Let a point P on S^2 map under stereographic projection to a point P in $\mathbb{R}^2$ with coordinates $(r\cos\ , r\sin\ )$. Then the spherical distance of P from the South Pole S is $2\tan\ ^1(r)$.

Here Q does not represent colatitude.

Problem 2 Let P and Q be the points $(2, 2)$ and $(3, 3)$ in $\mathbb{R}^2$, and P and Q the corresponding points on S^2 under the mapping $\quad ^1$. Use Theorem 1 to determine the spherical distance $P\ Q$.

7.4.2 Conjugate Transformations

We have seen a correspondence between points of the sphere S^2 and points of the extended plane. We now investigate the correspondence between spherical transformations and Möbius transformations.

Subsection 7.4.1

Earlier we saw that a spherical transformation is the composition of at most three reflections in great circles, and that the composition of any two reflections is a rotation. Here we shall see that reflection in a great circle corresponds to inversion in the image of that circle under stereographic projection, and that a rotation of S^2 corresponds to a particular kind of Möbius transformation. Thus we have two equivalent descriptions of spherical transformations, one on the sphere using 3 × 3 matrices, and one in the extended plane using Möbius transformations.

Theorems 5 and 4 of Subsection 7.2.2

Firstly, we establish the following pleasant result.

Theorem 2 Let P and Q be points on S^2 that are mirror images under reflection in a great circle C, and let the images of P, Q and C under stereographic projection be P, Q and C, respectively. Then P and Q are inverse points with respect to C.

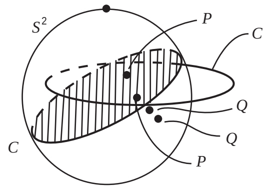

Proof The great circle lies in a plane, say, through the origin. consists of those points in $\mathbb{R}^3$ that are equidistant from the points P and Q. So any sphere through P and Q has its centre somewhere in , and so any circle on S^2 through P and Q also has its centre in . Thus any circle C on S^2 through P and Q meets the great circle C at right angles.

Under stereographic projection , the circles C and C map to generalized circles; since stereographic projection is angle-preserving, these images meet at right angles.

Now let C be any generalized circle through P and Q. The image of C under ^{1} is a circle on S^2 through P and Q; so $^1(C)$ is some circle C on S^2 that meets C at right angles. Then, since stereographic projection is angle-preserving, it follows that C meets C at right angles.

But C is any circle through P and Q, so any circle through P and Q meets C at right angles. It follows that P and Q are inverse points with respect to C, as required.

Here we are using Theorem 4 of Subsection 5.5.2.

Notice that, if C_1 and C_2 are a great circle and a little circle on S^2, and C_1 and C_2 are their images in the extended plane under , then there exists a

Möbius transformation that maps C_1 onto C_2, but there cannot be a rotation of S^2 that maps C_1 onto C_2. Hence there are Möbius transformations of the extended plane that do not correspond to rotations of the sphere.

Now, inversion in a generalized circle in the extended plane is necessarily a Möbius transformation composed with a complex conjugation. It follows from Theorem 2 that any spherical transformation corresponds to a Möbius transformation, composed (if need be) with a complex conjugation.

By Theorem 10, Subsection 5.3.7

So each rotation R of S^2 corresponds to some Möbius transformation of the extended plane by means of stereographic projection . If P is a point in the extended plane, it follows that $^{1}(P)$ is a point on S^2; R $^{1}(P)$ is the image of that point after the rotation, and R $^{1}(P)$ is the projection of that image onto the extended plane.

For both are orientation-preserving mappings.

So, R ^{1} is a mapping of the extended plane to itself. We therefore make the following definition.

Definition Let R denote a rotation of S^2, and stereographic projection. Then the transformation R ^{1} of the extended plane is called the **conjugate transformation** of the rotation R.

It follows from the above discussion that the image of the group of rotations of S^2 under is a group of Möbius transformations.

Lemma The conjugate transformation of a rotation of S^2 is a Möbius transformation of the extended plane.

This was proved in the earlier discussion.

Indeed it can be proved that the mapping R R ^{1} maps the group of rotations of S^2 isomorphically onto a group of Möbius transformations of the extended plane, but that it is not the whole group of Möbius transformations. We now characterize the Möbius transformations that are conjugate to rotations of S^2.

We omit a proof of this fact.

We saw this in the discussion above.

Theorem 3 A Möbius transformation M that is conjugate to a rotation of S^2 is of the form $M(z) = \frac{az+b}{\bar{b}z+\bar{a}}$, where a, b C.

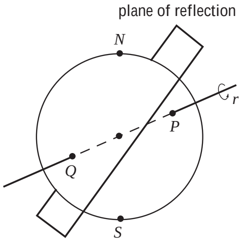

Proof There is one family of rotations of S^2 that we can easily relate to Möbius transformations, namely rotations around the Z-axis. Their conjugate transformations are simply rotations about the origin through the same angle. Our strategy will be to use reflections of S^2 to reduce a given rotation of S^2 to one about the Z-axis.

Let R denote a rotation of S^2 about the (diametrically opposite) points P and Q of S^2, through an angle , and let r denote the reflection in a suitable great circle of S^2 that maps P to the North Pole N; so that $r(P) = N$. Then since r is a reflection, it is its own inverse, so that $r(N) = P$. Similarly $r(Q) = S$, the South Pole, and $r(S) = Q$.

So we shall often use r and r ^{1} interchangeably below.

The spherical transformation $r^{-1}Rr$ is a direct isometry of S^2, and so is a rotation of S^2. Since

$$\begin{aligned} r^{-1}Rr\;(N) &= r^{-1}R\;(r(N))\\ &= r^{-1}R\;(P)\\ &= r^{-1}R(P) = r^{-1}(P) = N, \end{aligned}$$

the point N is a fixed point of the rotation; similarly S is also a fixed point. So $r^{-1}Rr$ is a rotation of S^2 about the Z-axis.

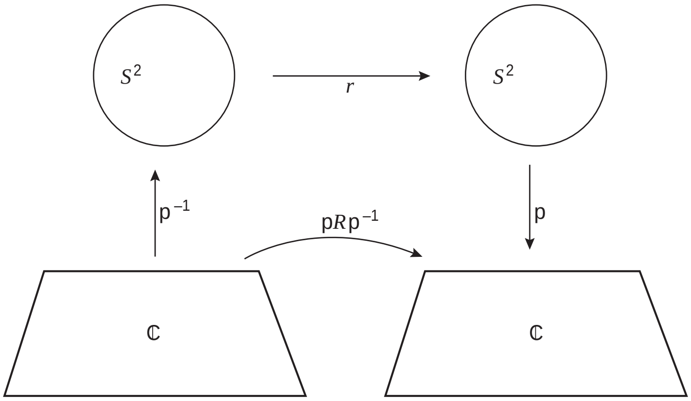

It follows that the transformation $(r^{-1}Rr)^{-1}$ is a rotation of the (x, y)-plane about the origin. Denoting this by M, it follows that $M(z) = e^{i}z$, for some real number .

Now

$$\begin{aligned} M &= r^{-1}Rr^{-1}\\ &= r^{-1\;-1}R^{-1}r^{-1}\\ &= r^{-1}R^{-1}r^{-1}; \end{aligned}$$

For $^{-1}$ is the identity map of S^2 to itself.

so that, since we know M, in order to know R^{-1} it is now sufficient to find r^{-1}.

Since N and P map to each other under the reflection r in a suitable great circle of S^2, the mapping r^{-1} must be inversion in a circle in the plane.

By Theorem 2 and $r^{-1} = r$.

Now the great circle of S^2 is the intersection of the sphere $X^2 + Y^2 + Z^2 = 1$ with some plane $aX + bY + cZ = 0$; we may assume that $a^2 + b^2 + c^2 = 1$. Since the plane cannot be vertical, $c = 0$.

This assumption will simplify our calculations.

By applying the formula for $^{-1}$ to the equation of the plane, we see that the image of the great circle is a circle C in the (x, y)-plane with equation

$$2ax + 2by + c\;x^2 + y^2 - 1 = 0,$$

or

$$x^2 + y^2 + 2\frac{a}{c}x + 2\frac{b}{c}y \quad 1 = 0.$$

The circle C has centre $\quad \frac{a}{c}, \quad \frac{b}{c}$, and the square of its radius is

For, $a^2 + b^2 + c^2 = 1$

$$1 + \quad \frac{a}{c} \quad ^2 + \quad \frac{b}{c} \quad ^2 = \frac{1}{c^2}\bullet$$

Next we express things in terms of complex numbers. Let $\quad = \frac{a}{c} \quad i\frac{b}{c}$, so that $\text{ﬂ} = |\quad|^2 = \frac{a^2+b^2}{c^2}$. Now inversion in C is a mapping of the form

is the centre of the circle in the plane.

$$\begin{aligned} z \quad & \frac{(1/c)^2}{\overline{\overline{z}}} + \\ &= \frac{\text{ﬂ}z \quad |\quad|^2 + (1/c)^2}{\text{ﬂ}z \quad \text{ﬂ}} \\ &= \frac{\text{ﬂ}z + 1}{\text{ﬂ}z \quad \text{ﬂ}}, \end{aligned}$$

You met this formula in Theorem 3, Subsection 5.2.1.

so that

For, $\quad |\quad|^2 + (1/c)^2 = 1$

$$r \quad ^1(z) = \frac{\text{ﬂ}z + 1}{\text{ﬂ}z \quad \text{ﬂ}}\bullet$$

Then

$$M(\quad r \quad ^1)(z) = e^{i} \quad \bullet \frac{\text{ﬂ}z + 1}{\text{ﬂ}z \quad \text{ﬂ}},$$

so that

$$\begin{aligned} R \quad ^1(z) &= \frac{e^{\quad i} \quad \bullet \frac{\text{ﬂ}z+1}{z} \quad + 1}{e^{\quad i} \quad \bullet \frac{\text{ﬂ}z+1}{z} \quad \text{ﬂ}} \\ &= \frac{e^{\quad i}\ (\text{ﬂ}z + 1) + (z \quad\quad)}{e^{\quad i}\ (\text{ﬂ}z + 1) \quad \text{ﬂ}(z \quad\quad)} \\ &= \frac{z \ |\quad|^2 e^{\quad i} \quad + 1 \quad + (\quad e^{\quad i} \quad\quad)}{z(\text{ﬂ}e^{\quad i} \quad \text{ﬂ}) + (e^{\quad i} \quad + |\quad|^2)} \\ &= \frac{z \ |\quad|^2 e^{\quad i\quad /2} + e^{i\quad /2} \quad + \quad e^{\quad i\quad /2} \quad\quad e^{i\quad /2}}{z \quad \text{ﬂ}e^{\quad i\quad /2} \quad \text{ﬂ}e^{i\quad /2} \quad + \quad e^{\quad i\quad /2} + |\quad|^2 e^{i\quad /2}}; \end{aligned}$$

thus $\quad R \quad ^1(z) = \frac{az+b}{\bar{b}z+\bar{a}}$, where $a = |\quad|^2 e^{\quad i\quad /2} + e^{i\quad /2}$ and $b = \quad e^{\quad i\quad /2} \quad e^{i\quad /2}$, as required.

Remark

Notice that the Möbius transformations $M(z) = \frac{az+b}{\bar{b}z+\bar{a}}$ conjugate to a rotation of S^2 are rather like the Möbius transformations representing direct hyperbolic transformations $M(z) = \frac{az+b}{\bar{b}z+\bar{a}}$, with $|b| < |a|$.

You met these Möbius transformations in Theorem 2, Subsection 6.2.1.

Problem 3 The two fixed points of a rotation of S^2 are distinct from N and S. Show that they map under stereographic projection to points in $\mathbb{C}$ of the form s and $-1/\bar{s}$.

Example 3 Prove that a Möbius transformation M that is conjugate to a rotation $R(Y, \theta)$ of S^2 is of the form $M(z) = \frac{az-b}{bz+a}$ where a, b are real.

Solution A rotation $R(Y, \theta)$ of S^2 fixes the points $(0,1,0)$ and $(0,-1,0)$, which map under stereographic projection to the points i and $-i$, respectively.

By Theorem 3, a conjugate Möbius transformation must be of the form

$$M(z) = \frac{cz+d}{-\bar{d}z+\bar{c}};$$

We use c and d to avoid using the letters a and b for several different things.

this fixes i and $-i$ if

$$i = \frac{ci+d}{-\bar{d}i+\bar{c}} \quad \text{and} \quad -i = \frac{-ci+d}{\bar{d}i+\bar{c}},$$

which we can rewrite in the form

$$\bar{d} + \bar{c}i = ci + d \quad \text{and} \quad \bar{d} - \bar{c}i = -ci + d.$$

Adding these equations gives $2\bar{d} = 2d$, so that d is real, and subtracting the equations gives $2\bar{c}i = 2ci$, so that c is real.

It follows that M has the desired form $M(z) = \frac{az-b}{bz+a}$, with a in place of c and $-b$ in place of d. □

Problem 4 Prove that a Möbius transformation M that is conjugate to a rotation $R(X, \theta)$ of S^2 is of the form $M(z) = \frac{az+ib}{ibz+a}$, where a, b are real.

We end with some interesting observations concerning Möbius transformations. A *fixed point* of a Möbius transformation $M(z) = \frac{az+b}{cz+d}$, $ad - bc \neq 0$, is a point z for which $M(z) = z$, that is, $\frac{az+b}{cz+d} = z$ or $cz^2 + (d-a)z - b = 0$. Since a quadratic equation can have at most two roots, it follows that a Möbius transformation can have at most two fixed points.

Problem 5 Determine the fixed points in $\mathbb{C}$ and in $\hat{\mathbb{C}}$ of the following Möbius transformations:

$$z \mapsto z+1; \quad z \mapsto -\frac{4}{z+4}; \quad z \mapsto -\frac{4}{z+5}.$$

Example 4 Prove that a Möbius transformation M with distinct fixed points a and b in $\mathbb{C}$ maps the family $\mathcal{A}$ of Apollonian circles defined by the point circles a and b to itself.

Solution Let B be the family of generalized circles through a and b.

Since M has fixed points a and b, it maps the family B onto itself. Also, since a Möbius transformation preserves angles, M thus maps the family of generalized circles that is orthogonal to those in B to itself. This family is the family A of Apollonian circles defined by the point circles a and b. □

A Möbius transformation maps generalized circles onto generalized circles.

By Theorem 3, Subsection 5.5.1

Example 5 Prove that a Möbius transformation M with distinct fixed points a and b in C *either* maps each circle in the family A of Apollonian circles defined by the point circles a and b to itself *or* it maps no circle in A to itself.

Solution By a suitable Möbius transformation f, we can map the point a to 0 and b to . Then f maps the family A of Apollonian circles defined by the point circles a and b to the family B of concentric circles with centre 0.

Then $f \quad M \quad f^{\ 1}$ is a Möbius transformation, and it maps 0 to 0 and to . It follows that $f \quad M \quad f^{\ 1}$ is of the form $z \qquad Kz$, for some non-zero complex number K.

If $|K| = 1$, then $f \quad M \quad f^{\ 1}$ maps each concentric circle in B to itself; but if $|K| = 1$ then $f \quad M \quad f^{\ 1}$ maps no circle in B to itself.

If we then apply the mapping $f^{\ 1}$ to return to the original situation, the required result follows at once. □

It is easy to verify that a Möbius transformation of C that has 0 and as its fixed points is necessarily of the form $z \qquad Kz$.

7.4.3 Coaxal Circles

Stereographic projection enables us to interpret the theory of coaxal circles, including all its special cases, very elegantly on the sphere S^2.

You met coaxal circles in Subsection 5.5.2.

Earlier you saw that, if C is the circle cut out on S^2 by the plane with equation $aX + bY + cZ + d = 0$, then stereographic projection maps C to:

In the proof of Theorem 7 of Subsection 5.2.4

(a) the circle with equation

$$x^2 + y^2 + 2\frac{a}{c+d}x + 2\frac{b}{c+d}y \quad \frac{c \quad d}{c+d} = 0, \quad \text{if } c = \quad d;$$

(b) the straight line with equation

$$ax + by \quad \frac{c \quad d}{2} = 0, \quad \text{if } c = \quad d.$$

If $c = \quad d$, the circle C passes through N.

Problem 6 Prove that a circle in R^2 with equation

$$x^2 + y^2 + 2 \quad x + 2 \quad y + \quad = 0$$

corresponds under stereographic projection to the circle cut out on S^2 by the plane with equation

$$2 \quad X + 2 \quad Y + (1 \qquad)Z + (1 + \quad) = 0.$$

Definition The circle in $\mathbb{R}^2$ with equation

$$x^2 + y^2 + 2 \quad x + 2 \quad y + \quad = 0$$

and the plane with equation

$$2 \quad X + 2 \quad Y + (1 \qquad)Z + (1 + \quad) = 0$$

are said to be **associated**.

That is, they correspond to each other under stereographic mapping.

We now use the idea of associated planes and circles to revisit the theory of coaxal circles.

Let C_1 and C_2 be distinct circles in $\mathbb{R}^2$ with equations

We allow C_1 and C_2 to be point circles.

$$x^2 + y^2 + 2 \quad {}_1 x + 2 \quad {}_1 y + \quad {}_1 = 0$$

and

$$x^2 + y^2 + 2 \quad {}_2 x + 2 \quad {}_2 y + \quad {}_2 = 0,$$

respectively; then their corresponding associated planes ${}_1$ and ${}_2$ have equations

$$2 \quad {}_1 X + 2 \quad {}_1 Y + (1 \qquad {}_1)Z + (1 + \quad {}_1) = 0$$

and

$$2 \quad {}_2 X + 2 \quad {}_2 Y + (1 \qquad {}_2)Z + (1 + \quad {}_2) = 0.$$

Now, either ${}_1$ and ${}_2$ are parallel or they are not parallel; they cannot be the same plane, since the associated circles are distinct.

First, suppose that ${}_1$ and ${}_2$ are not parallel, so that they have a common line, say.

For arbitrary p and q, $p \quad = \quad q$, let C_{pq} be the generalized circle with equation

$$\frac{p}{p+q} \quad x^2 + y^2 + 2 \quad {}_1 x + 2 \quad {}_1 y + \quad {}_1$$
$$+ \frac{q}{p+q} \quad x^2 + y^2 + 2 \quad {}_2 x + 2 \quad {}_2 y + \quad {}_2 \quad = 0,$$

and let ${}_{pq}$ be its associated plane, with equation

$$\frac{p}{p+q}(2 \quad {}_1 X + 2 \quad {}_1 Y + (1 \qquad {}_1)\, Z + (1 + \quad {}_1))$$
$$+ \frac{q}{p+q}(2 \quad {}_2 X + 2 \quad {}_2 Y + (1 \qquad {}_2)\, Z + (1 + \quad {}_2)) = 0.$$

The associated plane ${}_{pq}$ passes through the line ; so, as p and q vary, ${}_{pq}$ gives a family of planes through . There are three cases to consider.

Case 1: The line does not meet S^2.

Here the family of planes cuts out a family of circles in S^2 consisting of two point circles P and Q and a set of disjoint circles in between these.

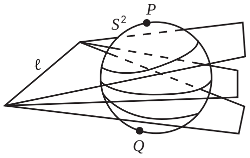

Under stereographic projection from N, the associated circles in the plane form an Apollonian family: a family of coaxal circles with two point circles.

Case 2: The line touches S^2 (at the point P, say).
Here, every plane of the family meets S^2. The plane tangent to the sphere at the point P meets it in a single point, P. All the other planes cut S^2 in a circle. All these circles have a common tangent at P, which is the line .

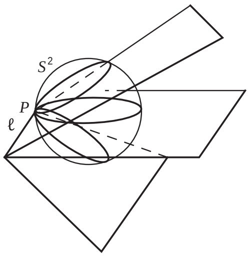

Under stereographic projection from N, the associated circles in the plane form a family of coaxal circles having a common tangent. If N and P are distinct, the image circles are true circles except for one straight line; if N coincides with P, the images are parallel lines.

Case 3: The line cuts S^2 in two points, say P and Q.
Here, every plane of the family meets S^2. Each plane cuts the sphere in a circle passing through the points P and Q.

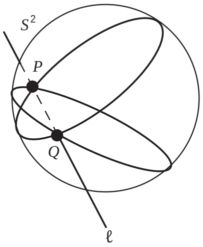

Under stereographic projection from N, the associated circles in the plane form a family of coaxal circles with two common points. If N, P and Q are distinct, the image circles are true circles except for one straight line; if N coincides with either P or Q, say P, the images are straight lines passing through the image of Q.

On the other hand, if the associated planes $_1$ and $_2$ are parallel then the family of planes $_{pq}$ consists of the planes parallel to $_1$ and $_2$, and the circles they cut out on the sphere consist of two point circles and a set of disjoint circles in between these.

This is similar to the situation in Case 1 above.

7.5 Planar Maps

We use maps of the Earth on a daily basis: for example. street maps of towns, road maps of countries, geological maps of continents, and for land, sea and air travel. Most of our maps are on flat sheets of paper, rather than globes; and we have to reduce the actual dimensions by a convenient *scale factor*, such as 1:12,000,000 for a continent, 1:100,000 for a map of a large city, or 1:25,000 for a detailed neighbourhood map.

Here, scale factor = length on map divided by length on Earth s surface.

We will assume in this section that the Earth is a sphere. In fact, the shape of the Earth is closer to that of an *oblate spheroid*, a sphere flattened near each pole and with a bulge around the equator. This bulge results from the rotation of the Earth, and causes its equatorial diameter to be over 26 miles (43 km) greater than its pole-to-pole diameter

Now an ideal planar map would represent shapes on the Earth s surface in such a way that relative distances, relative areas, shape, and angles are preserved under the operation of mapping. Unfortunately, no ideal map is possible!

For example, let ABC be a triangle on S^2, and let P, Q, R be the points on the planar map that correspond to A, B, C, respectively. Then the sides of ABC are line segments on S^2, that is, parts of great circles; and the sum of its angles exceeds . The sides of PQR are line segments in the plane, and the sum of its angles is . So we cannot simultaneously have geodesics on S^2

Subsection 7.3.1, Theorem 2

mapping onto geodesics in the plane AND preserve angles. This shows that an ideal planar map cannot exist.

Recall that ‘geodesics’ are curves of shortest length between two points.

Furthermore, Euler showed in 1775 that it is not possible to map a portion of the Earth's surface onto a planar map in such a way that relative distances (that is, distances up to a scaling factor) are preserved.

Leonhard Euler (1707–1783) was a prolific Swiss analyst, who worked for Frederick the Great (Prussia) and Catherine the Great (Russia) even after becoming blind.

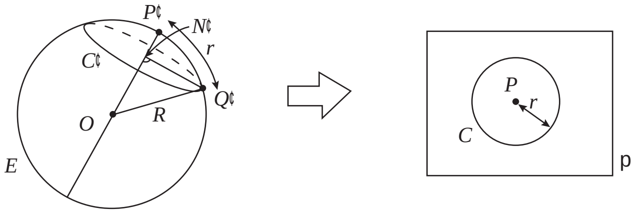

To see this, let E denote the surface of the Earth, a sphere with centre O and radius R; and let $P' \in E$ be the centre of a circle C' on E of radius r (we will suppose that $r < \frac{1}{2}\pi R$, so that C' lies in a hemisphere on E with P' as pole). Suppose that some ideal mapping represents the hemisphere onto a flat sheet, Π say, and let P, C on Π represent P', C' on E.

Here r is a distance measured along the surface.

Since the mapping scales all distances by the same scale factor, c say, all points on E at a distance r from P' map to points at a distance cr from P; hence C' is represented on Π by a circle C of radius cr, so that the length of C is $2\pi cr$.

Next, let Q' be any point on C', and N' the foot of the perpendicular from Q' to OP'. Then $\angle P'OQ' = \frac{r}{R}$, so that $Q'N' = R\sin(\angle P'OQ') = R\sin\left(\frac{r}{R}\right)$. But the circle C' is a circle in $\mathbb{R}^3$ with radius $Q'N'$, so that its length is $2\pi R\sin\left(\frac{r}{R}\right)$. Hence its representation C on Π must have length $c \times 2\pi R\sin\left(\frac{r}{R}\right)$.

We deduce from the last two paragraphs that

$$2\pi cr = 2\pi cR\sin\left(\frac{r}{R}\right),$$

so that

$$\frac{r}{R} = \sin\left(\frac{r}{R}\right).$$

This equality is impossible, since $\frac{\sin x}{x} < 1$ for all x with $0 < x < \frac{\pi}{2}$.

This is a standard result in Calculus.

So we have proved that it is not possible to represent a portion of the Earth's surface on a planar map in such a way that relative distances are preserved. It follows from this, too, that no (precisely accurate) planar map of a portion of E can have a fixed scale factor for the whole map.

In fact, there are three principal types of *map projections* for representing the spherical Earth, E, on a flat sheet of paper, Π. In their work, cartographers often use the term *meridians* to denote lines of latitude, and *parallels* to denote lines of longitude: we will follow this convention here.

In addition there are various purely mathematical approaches, for special purposes.

Azimuthal Projections

Azimuthal projections of E onto are ones where we project E onto a sheet that is tangent to E at some point, often the South Pole S. We can think of the projection mapping in the following way: a light source placed at some point (the *centre* of the projection) on the perpendicular to at the point of contact projects each point P on E onto a corresponding point P on . Wherever we place the source, meridians map to radial lines from the point of contact S, parallels to concentric circles with centre S, and directions from the point of contact are preserved.

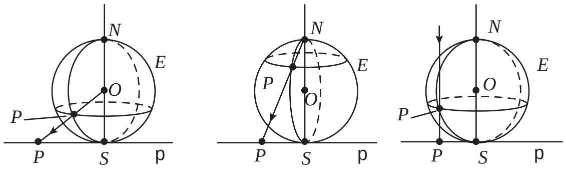

If the source is placed at the centre of E, the projection is called the *gnomic projection*, and we obtain a map of the lower hemisphere of E. Since any plane through O meets E in a great circle and in a line, each line in comes from a great circle on E. Hence geodesics on E are mapped onto straight lines on .

By making the tangent point N, we could obtain a map of the upper hemisphere.

If the light source is placed at the North Pole N of E, we obtain a map of the whole of E $\{N\}$ on . This projection is called the *stereographic projection* of E onto , and is the stereographic projection that you met earlier. It preserves angles, and so is called a *conformal* projection.

Subsection 5.2.4

If the light source is placed at infinity along the direction SON, and we consider only the upper hemisphere or the lower hemisphere of E, then we obtain a map of one hemisphere of E onto that is actually circular. This projection is called the *orthogonal* or *orthographic projection*. It preserves distances along parallels.

Cylindrical projections

Cylindrical projections of E onto are ones where we project E onto a sheet that is rolled in the form of a cylinder round E, touching it at its Equator; then we slit the sheet and flatten it.

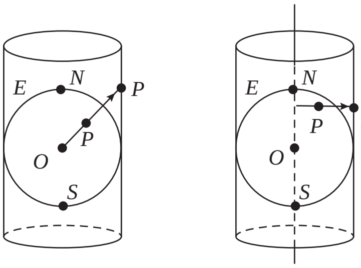

In one cylindrical projection, a light source placed at the centre O of E, the *centre* of the projection, projects each point P on E onto a corresponding point P on . Parallels map to horizontal circles that become horizontal lines when we unwrap the sheet, and meridians to vertical lines. The equator is mapped to itself, and areas near the North and South Poles become greatly enlarged.

Another cylindrical projection projects each point P on E horizontally onto the nearest point P on . Again, parallels map to horizontal circles that become horizontal lines when we unwrap the sheet, and meridians to vertical lines. This projection preserves the relative sizes of areas, so it is called an *equal area projection*.

The well-known *Mercator projection* is a cylindrical map where the parallels are horizontal and the meridians vertical, but the spacing of the parallels is chose cunningly so that *paths of constant compass bearing* (that is, paths making the same angle with every meridian they cross) are shown as straight lines. This projection is invaluable in navigation, since sailors and aircrew using it do not need to constantly change the compass bearing of the course they follow.

Gerardus Mercator (Latinized form of Gerard Kremer) (1512 1594) was a Flemish map maker who devised this projection in 1569, and introduced the term atlas .

Conical projections

Conical projections of E onto are ones where we project E onto a sheet that is rolled in the form of a cone round E, touching it at some parallel; then we slit the sheet along a meridian and flatten it, taking on the shape of a sector of a circle.

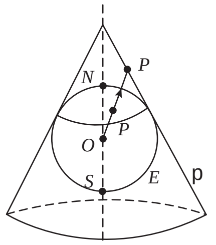

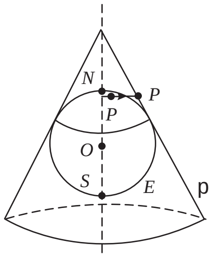

We may, for example, project each point P on E to the corresponding point P on radially out from the centre of E, or horizontally out from the axis of the cone to the nearest point on . Parallels map to arcs of circles whose centres are at the vertex V of the flattened cone, and meridians to straight lines radially out from V. Conic projections are particularly useful for mapping large areas in the mid-latitudes that are wide from east to west (such as Canada).

There are many variations of these fundamental types, all designed for special purposes; but a full discussion is beyond the scope of this book. Map-making is a fascinating subject!

You should look at a good Atlas now to see the various types of projections that it uses.

7.6 Exercises

Section 7.1

1. (a) Determine the equation of the great circle on S^2 that passes through the points $\frac{1}{6}, \quad \frac{1}{6}, \frac{2}{6}$ and $\frac{1}{6}, \frac{2}{6}, \frac{1}{6}$.
   (b) Determine the equation of the little circle on S^2 that passes through the points $\frac{1}{6}, \frac{1}{6}, \frac{2}{6}$, $(0, 0, 1)$ and $\frac{1}{6}, \frac{2}{6}, \frac{1}{6}$.
2. Determine the spherical polar representation of the point $\frac{3}{2\ 2}, \quad \frac{3}{2\ 2}, \quad \frac{1}{2}$ on S^2. Hence state the latitude, colatitude, and longitude of this point.
3. Determine the coordinates on S^2 of San Francisco (38 N, 123 W) and Brisbane (27 S, 153 E).

Section 7.2

1. Determine the images of the point $\frac{3}{5\ 2}, \frac{4}{5\ 2}, \quad \frac{1}{2}$ under the mappings $R(X, \frac{}{3})$, $R(Y, \frac{3}{4})$ and $R(Z, \frac{5}{4})$.
2. Determine the matrix **A** of a rotation $\mathbf{x} \quad \mathbf{Ax}$ of S^2 that maps $\frac{3}{5\ 2}, \frac{4}{5\ 2}, \quad \frac{1}{2}$ to $\frac{4}{5\ 2}, \quad \frac{3}{5\ 2}, \frac{1}{2}$. Verify your answer by direct calculation.
3. (a) Determine the matrix **A** such that the mapping $r : \mathbf{x} \quad \mathbf{Ax}$ represents reflection of S^2 in the plane with equation $x \quad 2y + z = 0$.
   (b) Hence determine the image under r of the point $P \quad \frac{1}{14}, \quad \frac{2}{14}, \frac{3}{14}$.
   (c) Prove that the midpoint of P and $r(P)$ lies in the plane .
4. Let a triangle PQR on S^2 have sides PQ of length r, QR of length p, and RP of length q, and angles , and at the vertices P, Q and R, respectively; let = . Prove that $p = q$.
5. Express the rotation $R(X, \frac{}{4})$ of S^2 as the composition of two reflections of S^2 in great circles.

This is the converse of Theorem 6 of Subsection 7.2.2.

Section 7.3

1. Construct the following figures on S^2:
   (a) a spherical triangle of area $(2 \quad 0.1)$;
   (b) a lune of area $/3$.
2. An equilateral strict spherical triangle has area $\frac{}{12}$. Determine:
   (a) the magnitudes of its angles;
   (b) the lengths of its sides;
   (c) the area of a dual triangle.
3. Estimate the distance between Beijing (40 N, 117 E) and Lima (12 S, 77 W), taking the radius of the Earth as 4000 miles.
4. Let ABC be a triangle on S^2 in which the angle at C is a right angle. If a, b and c denote the lengths of BC, CA and AB, and and denote the angles CAB and CBA, prove that:

(a) $\cos c = \cot \quad \times \cot \quad ;$
(b) $\cos \quad = \tan a \times \cot c = \cos b \times \sin \quad ;$
(c) $\sin a = \tan b \times \cot \quad = \sin \quad \times \sin c$.

5. Let the spherical triangle ABC have sides AB, BC and CA of lengths 1, 0.6 and 1.3, respectively. Estimate the length of the perpendicular from B to CA.
6. Let the figure $ABCD$ on S^2 be a strict spherical square that is, have four sides of the same length r and four angles of the same magnitude . Determine r in terms of .
7. Prove that the distance d on S^2 between the points with spherical polar coordinates $(\cos{}_1 \sin{}_1, \sin{}_1 \sin{}_1, \cos{}_1)$ and $(\cos{}_2 \sin{}_2, \sin{}_2 \sin{}_2, \cos{}_2)$ is given by.

$$\cos d = \sin{}_1 \sin{}_2 \cos({}_1 \quad {}_2) + \cos{}_1 \cos{}_2.$$

A strict spherical square is one whose whole can be seen at one time after a suitable rotation of S^2.

Sometimes this formula can save you a bit of arithmetic.

Section 7.4

1. Let P and Q be the points $\frac{1}{2}, \frac{1}{2}$ and $\frac{1}{4}, \frac{1}{4}$ in R^2, and P and Q be the corresponding points on S^2 under the mapping ^{1}. Use Theorem 1 of Subsection 7.4.1 to determine the spherical distance $P\ Q$.
2. Determine the Möbius transformation that is conjugate to the rotation $R\ \ _{Y, \frac{}{4}}$ of S^2.
3. (a) Prove that a Möbius transformation that maps 0 to and to 0 is of the form $z \quad \frac{K}{z}$, for some non-zero complex number K.
   (b) Prove that a Möbius transformation M that maps a to b and b to a (where $a = b$) maps the family A of Apollonian circles defined by the point circles a and b to itself, and maps exactly one circle in A to itself.

Summary of Chapter 7

Introduction

The (Elliptic) Parallel Postulate Given any line and a point P not on , all lines through P meet . (That is, there are no lines through P that are parallel to .)

Section 7.1: Spherical Space

1. The **unit sphere** in R^3 is $S^2 = \ (x, y, z) : x^2 + y^2 + z^2 = 1\ $. Each point (x, y, z) on S^2 has an **antipodal point** $(\ x, \ y, \ z)$. The **North Pole** of S^2 is $N(0, 0, 1)$, and the **South Pole** is $S(0, 0, \ 1)$. The **equator** is the intersection of S^2 with the horizontal plane through the origin.
2. A **great circle** is the circle cut out on S^2 by a plane through the centre of S^2.

A **circle of longitude** is a great circle through N. The **Greenwich Meridian** is the part of the circle of longitude through $A(1,0,0)$ which has positive x-coordinates.

A **circle of latitude** is the curve of intersection of S^2 with a plane parallel to the equator. A **little circle** is the curve of intersection of S^2 with a plane that is not a great circle.

3. A **line** on S^2 joining two points P and Q, or a **line** PQ, is either arc with endpoints P and Q of a great circle through P and Q.
4. The **length** of a line on S^2 is the Euclidean length of the corresponding circular arc; the **distance** (or **spherical distance**) between two points P and Q on S^2 is the length of the shorter of the two lines PQ, and is often denoted simply as PQ.

   For any three points P, Q and R on S^2 that do not lie on a single great circle, the lines PQ, QR and RP form a **triangle** PQR.
5. Great circles are **geodesics** on S^2; that is, curves of shortest distance between points.
6. The distance between any two points P and Q on S^2 is equal to the angle that they subtend at the centre of S^2. There is always a great circle joining any two points on S^2.

   Spherical distance possesses all the natural properties that a distance function should possess.
7. Any two lines on S^2 meet in two points.
8. Corresponding to each point P on S^2 the **great circle associated with *P*** is the great circle of points 2 distant from P. Conversely, for each great circle there are two points of S^2 which play the role of its poles .
9. For any two points P and Q on S^2, there is a rotation of S^2 that maps P to Q.
10. Each point P on S^2 has **spherical polar coordinates** $(\cos \ \sin \ , \sin \ \sin \ , \cos \ )$, where $0 \quad < 2$ and 0 .

    The **colatitude** of P is the angle ; the **latitude** of P is $= \frac{}{2}$. The **longitude** of P is the angle .

    The distance of P from the North Pole is its colatitude .
11. A **circle** on S^2 is the set of points on S^2 that are equidistant from a given point of S^2; it is either a little circle or a great circle.

Section 7.2: Sperical transformations

1. An **isometry** of S^2 is a mapping of S^2 to itself that preserves distances between points. The isometries of S^2 form a group, the **group of spherical isometries**, denoted by the symbol $S(2)$.

   A direct isometry of S^2 preserves the orientation of angles on S^2; an **indirect isometry** of S^2 reverses them. Reflection of S^2 in a plane through the centre of S^2 is an **indirect isometry** of S^2.
2. The three elementary rotations of S^2 are its rotations about the coordinate axes.

The transformation

$$\begin{matrix} R(Z,\ ): & S^2 & S^2 \\ & \mathbf{x} & \mathbf{Ax} \end{matrix}$$

where $\mathbf{x} = \begin{matrix} x \\ y \\ z \end{matrix}$ and $\mathbf{A} = \begin{matrix} \cos & \sin & 0 \\ \sin & \cos & 0 \\ 0 & 0 & 1 \end{matrix}$ is a rotation of S^2 about the positive z-axis through an angle .

The transformation

$$\begin{matrix} R(X,\ ): & S^2 & S^2 \\ & \mathbf{x} & \mathbf{Ax} \end{matrix}$$

where $\mathbf{x} = \begin{matrix} x \\ y \\ z \end{matrix}$ and $\mathbf{A} = \begin{matrix} 1 & 0 & 0 \\ 0 & \cos & \sin \\ 0 & \sin & \cos \end{matrix}$ is a rotation of S^2 about the positive x-axis through an angle .

The transformation

$$\begin{matrix} R(Y,\ ): & S^2 & S^2 \\ & \mathbf{x} & \mathbf{Ax} \end{matrix}$$

where $\mathbf{x} = \begin{matrix} x \\ y \\ z \end{matrix}$ and $\mathbf{A} = \begin{matrix} \cos & 0 & \sin \\ 0 & 1 & 0 \\ \sin & 0 & \cos \end{matrix}$ is a rotation of S^2 about the positive y-axis through an angle .

3. The **inverse of an elementary rotation** through a given angle is an elementary rotation through the same angle but in the opposite direction. Its matrix is the transpose of the matrix of the original elementary rotation.
4. A rotation of S^2 that maps $A(1,0,0)$ to $P(\cos\ \ \sin\ ,\sin\ \ \sin\ ,\cos\ )$ is given by the composition $R(Z,\ )\,R\ \ Y,\ $, where $=\frac{\ }{2}$.
5. The **equatorial plane** of S^2 is the plane $z = 0$ that contains the equator of S^2. **Reflection in the equatorial plane** of S^2 is the mapping $\mathbf{x}\quad \mathbf{Ax}$ that sends the point $\mathbf{x} = (x,y,z)$ to the point $(x,y,\ z)$; it has matrix $\mathbf{A} = \begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{matrix}$.
6. Reflection of S^2 in the plane with equation $ax + by + cz = 0$, where $a^2 + b^2 + c^2 = 1$, is given by the mapping $\mathbf{x}\quad \mathbf{Ax}$ where $\mathbf{A} = \begin{matrix} 1\ \ 2a^2 & 2ab & 2ac \\ 2ab & 1\ \ 2b^2 & 2bc \\ 2ac & 2bc & 1\ \ 2c^2 \end{matrix}$.
7. The product of any two reflections of $\mathbb{R}^3$ in planes through O that meet in a common line is a rotation about that common line.
8. Every rotation is a product of two reflections; these reflections are in planes that have the axis of rotation as their common line, and are separated by half the angle of the rotation.
9. The image of a circle on S^2 under an isometry of S^2 to itself is a circle.
10. **The Product Theorem** Every isometry of S^2 is a product of at most three reflections in great circles.

    An isometry of S^2 is known completely when its effect on three points (which do not lie on a common great circle) is known.

11. **Isosceles Triangle Theorem** Let a triangle PQR on S^2 have sides PQ of length r, QR of length p, and RP of length q, and let $p = q$. Also, let the angles at the vertices P, Q and R be , and , respectively. Then = .
12. Every isometry of S^2 is a reflection, a rotation, or the composite of a reflection and a rotation.

    Every direct isometry of S^2 is a rotation, and vice versa.

Section 7.3: Spherical Trigonometry

1. Given any three angles , and with (5 >) + + > , there exists a spherical triangle with those angles.
2. The sum of the angles of any spherical triangle is greater than . The **angular excess** of a spherical triangle is the difference between the sum of the angles of the triangle and .
3. The **strict line segment** PQ is the shorter of the two line segments PQ. (If both are of length , we choose either in some explicit way.)

    The **strict triangle** PQR is that triangle PQR whose sides are strict line segments, and its inside is the region of S^2 that they bound; we can rotate S^2 so that the whole of the strict triangle can be seen at one time.
4. A **lune** is a 2-sided polygon bounded by two great circles
5. The area of a spherical triangle is equal to its angular excess.
6. **Dual Triangles Theorem** With every strict spherical triangle there is associated another triangle , called its *dual*, whose angles are the complements of the sides of and whose sides are the complements of the angles of . (The complement of a quantity x is the quantity x.)
7. The distance d on S^2 between the points (p_1, p_2, p_3) and (q_1, q_2, q_3) is given by $\cos d = p_1q_1 + p_2q_2 + p_3q_3$. The distance between corresponding points on the surface of the Earth is then given (in miles) by the scaling map **x** 4000**x**.
8. **Pythagoras' Theorem** Let ABC be a triangle on S^2 in which the angle at C is a right angle. If a, b and c are the lengths of BC, CA and AB, then $\cos c = \cos a \times \cos b$.

    Also $\sin \ = \sin a \ \ \sin c$ and $\tan \ = \tan a \ \ \sin b$.
9. We can always drop a perpendicular from a point on S^2 to a line on S^2.
10. Let ABC be a **strict triangle** on S^2, in which the sides AB, BC and CA have lengths c, a and b, respectively, and the angles at A, B and C are , and , respectively. Then

    Sine Rule $\frac{\sin}{\sin a} = \frac{\sin}{\sin b} = \frac{\sin}{\sin c}$;

    Cosine Rule for Sides $\cos c = \cos a \cos b + \sin a \sin b \cos$;

    Cosine Rule for Angles $\cos \ = \sin \ \sin \ \cos c \ \ \cos \ \cos$.

Section 7.4: Spherical Geometry and the Extended Complex Plane

1. The image of the point P (cos sin %sin sin , cos)% under stereographic projection has coordinates cos tan $\frac{1}{2}$, sin tan $\frac{1}{2}$.

2. Let a point P on S^2 map under stereographic projection to a point P in $\mathbb{R}^2$ with coordinates $(r\cos\ , r\sin\ )$. Then the spherical distance of P from the South Pole S is $2\tan\ ^{1}(r)$.
3. Let P and Q be points on S^2 that are mirror images under reflection in a great circle C , and let the images of P , Q and C under stereographic projection be P, Q and C, respectively. Then P and Q are inverse points with respect to C.
4. Let R denote a rotation of S^2, and stereographic projection. Then the transformation $R\ ^{1}$ of the extended plane is the **conjugate transformation** of the rotation R.

   The conjugate transformation of a rotation of S^2 is a Möbius transformation of the extended plane.
5. A Möbius transformation M that is conjugate to a rotation of S^2 is of the form $M(z) = (az+b) \quad \bar{b}z + \bar{a}$, where $a, b \quad \mathbb{C}$.

   A Möbius transformation can have at most two fixed points.
6. The circle in $\mathbb{R}^2$ with equation $x^2 + y^2 + 2\ x + 2\ y + \ = 0$ and the plane with equation $2\ X + 2\ Y + (1\ )Z + (1 + \ ) = 0$ are said to be **associated**.

Section 7.5: Planar Maps

1. It is not possible to map a portion of the Earth s surface onto a planar map in such a way that relative distances are preserved. There is no ideal planar map.
2. **Azimuthal projections** of the Earth E onto a flat sheet of paper are ones where we project E onto placed tangent to E at some point, often the South Pole S.

   The **gnomic projection** uses the centre of E as the centre of projection, stereographic projection uses the North Pole of E as the centre of projection, and *orthogonal* or *orthographic projection* uses rays from infinity parallel to the polar axis.
3. **Cylindrical projections** of E onto are ones where we project E onto a sheet that is rolled in the form of a cylinder round E, touching it at its Equator; then we slit the sheet along a meridian and flatten it.
4. **Conical projections** of E onto are ones where we project E onto a sheet that is rolled in the form of a cone round E, touching it at some parallel; then we slit the sheet along a meridian and flatten it, taking on the shape of a sector of a circle.

8 The Kleinian View of Geometry

We now describe in detail what we have called the Kleinian view of geometry. As we said earlier, this is a powerful way in which the geometries we have presented can be unified and seen as different aspects of the same idea.

Chapter 0

8.1 Affine Geometry

Klein argued that in any geometry there are properties of figures that one discusses, but these properties vary from one geometry to another. In Euclidean geometry, segments have lengths, lines cross at angles one can measure; it makes sense to ask if a given triangle is equilateral, and so on. But none of these properties makes sense in affine geometry. There one may speak only of ratios of lengths along parallel lines, and all triangles are (affine) congruent. Klein interpreted this as follows: in each geometry we are studying the same space of points, the familiar plane $\mathbb{R}^2$. In Euclidean geometry we admit only the transformations that preserve length. These are transformations of the form

$$t(\mathbf{x}) = \mathbf{U}\mathbf{x} + \mathbf{a},$$

where $\mathbf{U}$ is an orthogonal 2×2 matrix and $\mathbf{a} \in \mathbb{R}^2$.

Because every transformation in Euclidean geometry preserves lengths, and therefore angles, it makes sense to speak of distance in this geometry.

In affine geometry, however, we admit all transformations of the form

$$t(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b},$$

where $\mathbf{A}$ is an invertible 2×2 matrix and $\mathbf{b} \in \mathbb{R}^2$. Because every transformation in affine geometry preserves ratios of lengths along parallel lines, it makes sense to speak of such ratios in this geometry. But because there are affine transformations that do not preserve length, it does not make sense to speak of distance in this geometry.

Klein observed that rather more is true. In both geometries, the allowable transformations form a group. This was a novel observation at the time, for Group Theory as such was only just being discovered. Moreover, the group of Euclidean transformations is a subgroup of the group of affine transformations.

This observation enabled Klein to unify the two geometries. He said that both geometries concern the same space $\mathbb{R}^2$, that each geometry involves a group of transformations which act on this space, and that the Euclidean group is a subgroup of the affine group. In this sense Euclidean geometry is a *subgeometry* of affine geometry.

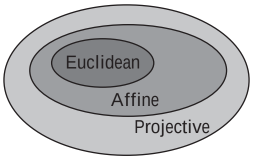

The hard work comes with the remaining geometries. How can affine geometry and projective geometry be unified? Projective geometry is not about the plane $\mathbb{R}^2$ but about projective space $\mathbb{RP}^2$. Its transformations are written as 3×3 homogeneous matrices, not as 2×2 matrices. To accomplish the unification, we must first establish that the space $\mathbb{R}^2$ can be regarded as a subset of $\mathbb{RP}^2$.

Here, *homogeneous* means that the matrix can be multiplied by any non-zero real number without affecting the transformation.

To do this, we consider the subset of Points in $\mathbb{RP}^2$ that pierce the standard embedding plane $z = 1$. We shall temporarily denote this subset by $\mathbb{RP}^2_0$, so that

$$\mathbb{RP}^2_0 = \{[x, y, z] : z = 0\}.$$

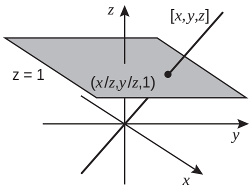

Now each point (x, y) in $\mathbb{R}^2$ can be identified with the Point $[x, y, 1]$ in $\mathbb{RP}^2_0$; and, conversely, each Point $[x, y, z] = [x/z, y/z, 1]$ in $\mathbb{RP}^2_0$ can be identified with the point $(x/z, y/z)$ in $\mathbb{R}^2$. This provides us with a one–one correspondence between $\mathbb{R}^2$ and $\mathbb{RP}^2_0$ which enables us to regard $\mathbb{R}^2$ as a subset of $\mathbb{RP}^2$.

Next (following Klein) we prove a result which shows that the affine group can be considered as a subgroup of the projective group.

Theorem 1 The group of affine transformations can be identified with the subgroup of projective transformations that map $\mathbb{RP}^2_0$ onto itself.

Proof A projective transformation which maps $\mathbb{RP}^2_0$ onto itself is also one that maps the ideal Line $z = 0$ onto itself. It is therefore sufficient to characterize the projective transformations which map ideal Points to ideal Points.

It is clear that the set of projective transformations which map the Line $z = 0$ onto itself is a group. For, if t is a transformation which maps ideal Points to ideal Points, then so is its inverse t^{-1}; and, if t and s are projective transformations which map ideal Points to ideal Points, then so is $t \circ s$. We denote this group by $P_0(2)$.

Recall that the group of all projective transformations is denoted by $P(2)$.

Let $\mathbf{A} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & k \end{pmatrix}$ be a matrix associated with a projective transformation t. Then t maps ideal Points to ideal Points if and only if

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & k \end{pmatrix} \begin{pmatrix} x \\ y \\ 0 \end{pmatrix} = \begin{pmatrix} ? \\ ? \\ 0 \end{pmatrix},$$

where we do not care what numbers the ? s stand for, so long as they are not both zero. It follows that t maps ideal Points to ideal Points if and only if

$$gx + hy + 0 = 0.$$

This must hold for all $x = 0$ and all $y = 0$, so $g = h = 0$. Since $\mathbf{A}$ is invertible, it follows that $\det \mathbf{A} = k(ae \quad db) = 0$, so $k = 0$ and $(ae \quad db) = 0$. Moreover, since the matrices associated with a projective transformation are determined only up to a non-zero real multiple, we can divide $\mathbf{A}$ by k. If we then rewrite a/k as a, b/k as b, and so on, we see that the transformations in $P_0(2)$ are precisely those projective transformations which have an associated matrix of the form

$$\begin{matrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{matrix}\ , \quad \text{where } ae \quad db = 0.$$

Let us now consider the effect that such a transformation has on Points in $\mathbb{RP}^2_0$. If $^{\#}\ x\ , y\ , 1\ ^{\$}$ is the image of $[x, y, 1]$ under the transformation, then

$$\begin{matrix} x \\ y \\ 1 \end{matrix} = \begin{matrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{matrix} \quad \begin{matrix} x \\ y \\ 1 \end{matrix} = \begin{matrix} ax + by + c \\ dx + ey + f \\ 1 \end{matrix} .$$

This equation provides us with a relationship between x, y and $x\ , y$ which we can rewrite in the form

$$\begin{matrix} x \\ y \end{matrix} = \begin{matrix} a & b \\ d & e \end{matrix} \quad \begin{matrix} x \\ y \end{matrix} + \begin{matrix} c \\ f \end{matrix} .$$

Since $ae \quad db = 0$ it follows that we can map the point (x, y) associated with the Point $[x, y, 1]$ to the point $(x\ , y\ )$ associated with the Point $[x\ , y\ , 1]$ by an affine transformation. We conclude that the group $P_0(2)$ may be identified with the affine group $A(2)$.

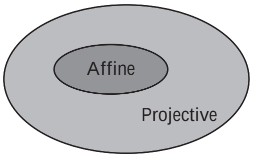

It follows that the space $\mathbb{R}^2$ of affine geometry can be considered as a subset of projective space $\mathbb{RP}^2$, and that the group $A(2)$ of affine geometry can be considered as a subgroup of the projective group $P(2)$. In other words, affine geometry is a *subgeometry* of projective geometry.

It is important to realize that in the process of extracting affine space from projective space we lose all the ideal Points on the ideal Line, and that this omission can have a significant effect on figures. For example, in projective geometry any two Lines meet at a Point. But if the Point of intersection is an ideal Point, then the corresponding affine figure consists of two lines that do not meet. This gives rise to the affine concept of parallel lines, which is meaningless in projective geometry. The fact that parallelism is an affine property can be checked easily by recalling that an affine transformation corresponds to a projective transformation which maps ideal Points to ideal Points. Under such a transformation, Lines which intersect at an ideal Point map to Lines which intersect at an ideal Point. This is an alternative proof of the following theorem that we first proved in Chapter 2.

Theorem 2 An affine transformation maps parallel straight lines to parallel straight lines.

Subsection 2.3.3, Theorem 3

Parallelism is not the only new property to arise from the omission of ideal Points. In affine geometry, three distinct types of (non-degenerate) conic arise because there are three ways in which a projective conic can meet the ideal Line. In the first figure below, the projective conic and the Line cross at two Points; the corresponding picture in $\mathbb{R}^2$ is that of a hyperbola. In the second figure, the projective conic and the Line touch at one Point; the corresponding picture in $\mathbb{R}^2$ is that of a parabola. In the third figure, the projective conic and the Line do not meet at any Point; the corresponding picture in $\mathbb{R}^2$ is that of an ellipse.

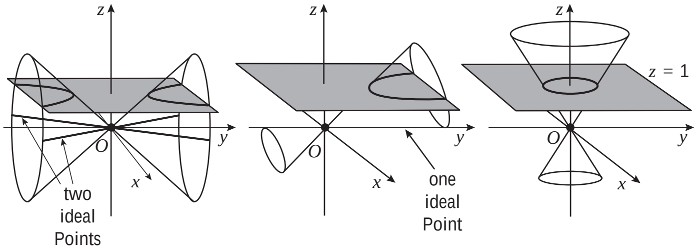

So hyperbolas, parabolas and ellipses are plane conics which correspond to projective conics with two, one and no ideal Points missing, respectively. Since affine transformations correspond to projective transformations which map ideal Points to ideal Points, they cannot change the number of ideal Points on a projective conic, and hence cannot change the type of the plane conic. This is an alternative proof of the following theorem first proved in Chapter 2.

Subsection 2.5.1, Theorem 4

Theorem 3 Affine transformations map ellipses to ellipses, parabolas to parabolas, and hyperbolas to hyperbolas.

So far, we have discussed some affine properties that do not make sense in projective geometry. By contrast, any property in projective geometry must also be a property in affine geometry. Indeed, *any property that is preserved by the group of all projective transformations must also be preserved by the subgroup of all affine transformations*.

For example, tangency is a property of projective conics, so every projective transformation maps tangents to tangents. In particular, a projective transformation which corresponds to an affine transformation must map a tangent whose Point of contact is not an ideal Point to a tangent whose Point of contact is not an ideal Point. Interpreting this observation as an affine result, we obtain the following theorem which we first met in Chapter 2.

Theorem 4 Let t be an affine transformation, and let be a tangent to a conic C. Then $t(\ )$ is a tangent to the conic $t(C)$.

Subsection 2.5.1, Theorem 7

Notice that in this proof of the theorem we did not have to consider tangents whose Point of contact is an ideal Point. But to what does such a tangent correspond in R^2? The answer depends on how many ideal Points lie on the projective conic.

A projective conic which passes through one ideal Point corresponds to a parabola in R^2. Since such a projective conic meets the ideal Line at precisely one Point, the tangent at that Point must be the ideal Line $z = 0$. This tangent does not correspond to anything in R^2!

Recall that a Line is a *tangent* to a projective conic if it meets the conic at precisely one Point.

More interesting are those projective conics which pass through two ideal Points. Such a projective conic corresponds to a hyperbola in R^2. The tangents which touch the projective conic at its two ideal Points correspond to lines in R^2, and these lines are defined to be the asymptotes of the hyperbola.

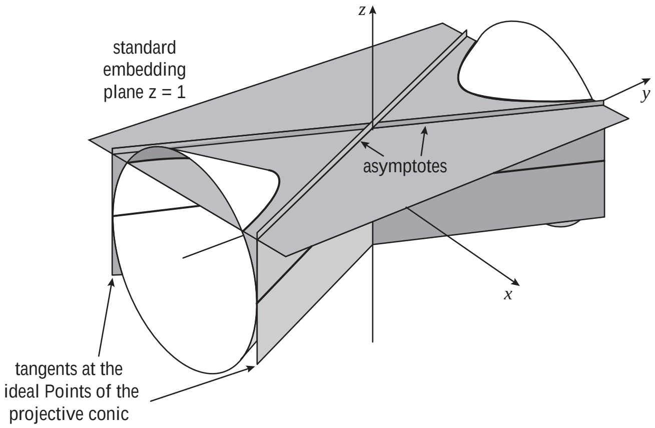
z
standard embedding plane z = 1
y
asymptotes
x
tangents at the ideal Points of the projective conic

Since affine transformations correspond to those projective transformations which map ideal Points to ideal Points, it follows that tangents at ideal Points map to tangents at ideal Points, and hence that asymptotes map to asymptotes. This proves the following theorem which you met in Chapter 2.

Subsection 2.5.1, Theorem 6

Theorem 5 Let t be an affine transformation, and let H be a hyperbola with asymptotes $\ _1$ and $\ _2$. Then $t(H)$ has asymptotes $t(\ _1)$ and $t(\ _2)$.

We now summarize what we have done so far.

Geometry	Space	Transformations	Invariants
Projective	$\mathbb{R}P^2 = \mathbb{R}^2$ + ideal Line	$t : [\mathbf{x}] \quad [\mathbf{Ax}]$, where $\mathbf{A}$ is a 3×3 invertible matrix	incidence, collinearity, cross-ratio
Affine	$\mathbb{R}^2$	$t : \mathbf{x} \quad \mathbf{Ax} + \mathbf{b}$, where $\mathbf{A}$ is a 2×2 invertible matrix	ratios along parallel lines
Euclidean	$\mathbb{R}^2$	$t : \mathbf{x} \quad \mathbf{Ux} + \mathbf{a}$, where $\mathbf{U}$ is a 2×2 orthogonal matrix	length, angle

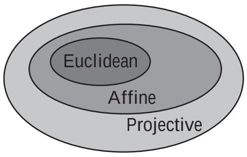

We can think of the geometries in this table as forming a hierarchy of geometries with projective geometry at the top. Any property of a geometry in the hierarchy is also a property of the geometries further down the hierarchy. Also, theorems that are valid in one geometry in the hierarchy also hold in geometries further down.

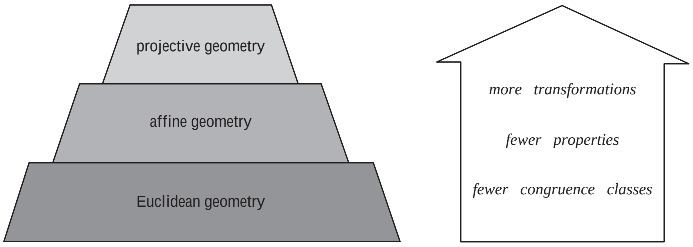

As we move up the hierarchy, the group of transformations becomes larger and so the congruence classes of figures merge together as more and more figures become congruent to each other. In Euclidean geometry, two triangles are congruent if and only if they have the same side lengths, and two conics are congruent if and only if they are of the same type (hyperbola, parabola or ellipse) and have the same shape and size. In affine geometry, any two triangles are congruent, and conics are congruent if and only if they are of the same type. In projective geometry, any two triangles are congruent and any two projective conics are congruent.

8.2 Projective Reflections

The cross-ratio enables us to introduce a type of projective transformation that generalizes reflection in Euclidean geometry.

Definitions Let F be a Point that does not lie on a Line . Then a **projective reflection** in with **centre** F and non-zero **parameter** k is a mapping

$$\begin{aligned} r : \mathbb{RP}^2 &\to \mathbb{RP}^2 \\ P &\mapsto Q \end{aligned}$$

where

(a) $Q = P$, if P lies on ℓ;
(b) the cross-ratio $(FRPQ) = k$, where R is the Point of intersection of FP and ℓ.

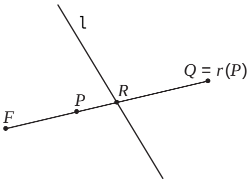

Remark

Provided that P does not lie on ℓ, then by varying the parameter k we can arrange that the Point Q is any Point on the Line FP other than the Points F, R and P.

We now indicate why this map is called a projective reflection.

Let ℓ' be the representation of a Line ℓ in an embedding plane π, and consider all the lines in π that are perpendicular to ℓ'. Since these lines are parallel, they must correspond to Lines in $\mathbb{RP}^2$ which meet at an ideal Point, F say. Let r denote the projective reflection in ℓ with centre F and parameter $k = -1$. Then, in the embedding plane π the effect of r is to map each (non-ideal) point P' to its Euclidean reflection in ℓ'.

r fixes the Points of ℓ so that its effect in π is to fix the points of ℓ'.

Next consider a Point P that does not lie on ℓ, and let $Q = r(P)$. Then PQ passes through F and meets ℓ at the Point R, where $(FRPQ) = -1$. For the embedding plane π, F is an ideal Point, so that

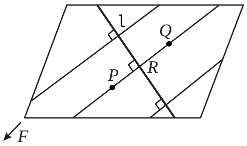

$$(FRPQ) = \frac{Q'R'}{P'R'} = -1$$

(where R', P' and Q' are the representations of R, P and Q in π). The point R' is therefore the midpoint of $P'Q'$. Moreover, since PQ passes through F, $P'Q'$ must be perpendicular to ℓ', so Q' is the Euclidean reflection of P' in ℓ', as suggested.

We now relate projective reflections to projective transformations.

Theorem 1 A projective reflection is a projective transformation.

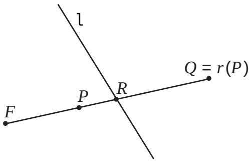

Proof Consider the projective reflection r in the Line ℓ with centre F and parameter k, and let $Q = r(P)$ where P is a Point of $\mathbb{RP}^2$ that does not belong to $\ell \cup \{F\}$. Then $(FRPQ) = k$, where R is the Point of intersection of PQ and ℓ.

By the Fundamental Theorem of Projective Geometry there is a projective transformation t which maps F to $F' = [0, 0, 1]$ and which maps the Line ℓ to the Line with equation $z = 0$. (This can be achieved by mapping a pair of Points on to $[1, 0, 0]$ and $[0, 1, 0]$.) If $P = t(P)$, $Q = t(Q)$ and $R = t(R)$, then, since t preserves cross-ratio, it follows that $(F\,R\,P\,Q\,) = (FRPQ) = k$.

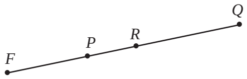

Now, if $P = [a, b, c]$, then the Line through P and F has equation $bx = ay$, and so it meets the Line $z = 0$ at $R = [a, b, 0]$. As Q also lies on the Line $bx = ay$, we can write $Q = [a, b, d]$, for some d; here $d = 0$, since Q does not lie on .

Since

$$(a, b, c) = 1 \bullet (a, b, 0) + c \bullet (0, 0, 1)$$

and

$$(a, b, d) = 1 \bullet (a, b, 0) + d \bullet (0, 0, 1),$$

if follows that $k = (F\,R\,P\,Q\,) = c/d$, so that $d = c/k$. Hence the mapping $s : \mathbb{RP}^2 \quad \mathbb{RP}^2$ which sends P to Q is the projective transformation defined by

$$s([a, b, c]) = [a, b, c/k].$$

Also, s fixes the Points of $\{F\,\}$. Thus the transformation r which maps P to Q must be equal to the composite $t^{\ 1} \quad s \quad t$ and so must be a projective transformation, as required.

Remarks

1. If the parameter $k = 1$, then s and hence $r = t \quad s \quad t^{\ 1}$ are the identity map.
2. If $k = 1$, then the only Points fixed by r are F and the Points of ; for, if
$$s([a, b, c]) = [a, b, c/k] = [a, b, c],$$
then either $c = 0$ (in which case $[a, b, c]$ lies on) or $a = b = 0$ (in which case $[a, b, c] = F\,$).
3. If $k = 0$, then s, and hence r, is not a projective transformation.

We showed earlier that every projective transformation can be expressed as a composite of perspectivity transformations. In fact, a similar result holds for projective reflections.

Theorem 4, Subsection 3.3.4

Theorem 2 Every projective transformation can be expressed as a composite of a finite number of projective reflections.

We omit a proof, since we shall not use this result.

8.3 Hyperbolic Geometry and Projective Geometry

First we introduce a new way of representing lines in hyperbolic geometry that will be useful in explaining how hyperbolic geometry is related to projective geometry. In this representation lines appear straight, but angles are no longer represented accurately. The realization that this could be done was one of the crucial discoveries made by Klein around the time he was preparing his Erlangen Program.

To obtain the new representation we use two mappings, stereographic projection and orthogonal projection. Stereographic projection from the North Pole of S^2 maps the Southern hemisphere onto the interior of the unit disc $\mathbb{D}$, and the equator to the boundary of $\mathbb{D}$. Its inverse therefore maps $\mathbb{D}$ onto the Southern hemisphere. It is an angle-preserving map, so it sends an arc of a generalized circle in $\mathbb{D}$, which is perpendicular to the equator, to an arc of a circle on the sphere perpendicular to the equator such an image may be thought of as a circular arc hanging vertically downwards.

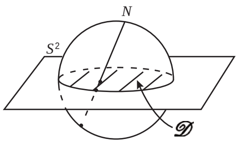

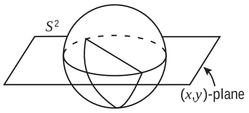

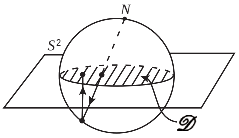

Orthogonal projection maps each point (x, y, z) in $\mathbb{R}^3$ along a line parallel to the z-axis to the point $(x, y, 0)$ in the (x, y)-plane. It maps the Southern hemisphere and the equator back onto $\mathbb{D}$. It maps a circular arc on the sphere perpendicular to the equator onto a segment of a straight line.

Definition The **linearization map** L of the unit disc $\mathbb{D}$ onto itself consists of the inverse of the stereographic projection map followed by the orthogonal projection map.

Then L fixes every point of the boundary of the unit disc $\mathbb{D}$, and maps arcs of generalized circles perpendicular to the boundary of $\mathbb{D}$ to segments of straight lines.

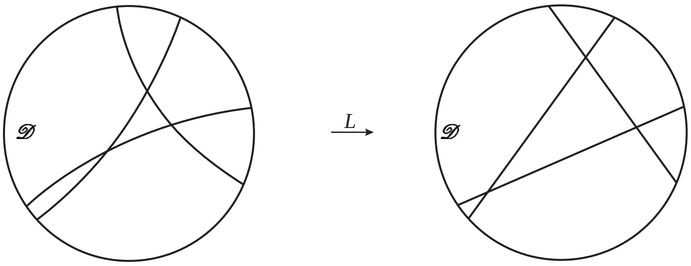

Three hyperbolic lines in the Poincaré model

Their images under L

The linearization map transforms the Poincaré model of hyperbolic geometry into a new model of hyperbolic geometry in which d-lines are represented by straight line segments. We call this new model of hyperbolic geometry the *Beltrami Klein model*, after the Italian mathematician Eugenio Beltrami (1833 1900) who discovered it, and the German mathematician Felix Klein, who connected it to projective geometry.

It is important to realize that although d-lines look straight in this model, angles are not represented accurately.

By considering the definition of L as the composite of the inverse of stereographic projection followed by orthogonal projection, it is clear that the image of the d-line joining two points on the boundary of $\mathbb{D}$ is the straight line segment joining the points. To find the image of a point P of $\mathbb{D}$ under L, consider

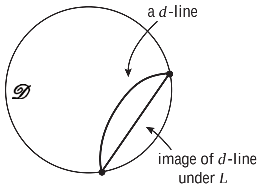

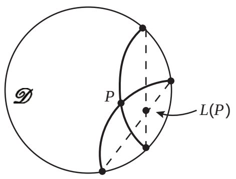

two d-lines through P and locate their end-points on the boundary of $\mathbb{D}$. Then $L(P)$ is the point where the corresponding chords intersect. Notice that, if it helps, we can always assume that one of the d-lines is a diameter of $\mathbb{D}$.

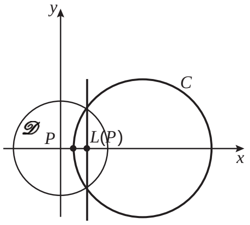

Example 1 Find the image under L of the point $P(r, 0)$ in $\mathbb{D}$.

Solution Notice first that it is sufficient to consider the image under L of the x-axis and the d-line through P at right angles to the x-axis.

The transformation L maps the x-axis to itself (as a line, not pointwise). The d-line through P perpendicular to the x-axis is part of the circle C with equation

$$x^2 + y^2 - 2ax + 1 = 0. \qquad (1)$$

The centre of C is on the x-axis, so there is no y-term in equation (l).

Since $P(r, 0)$ lies on C, it follows from equation (1) that $r^2 - 2ar + 1 = 0$, so that $a = \frac{r^2+1}{2r}$.

The circle C with equation (1) meets the boundary of $\mathbb{D}$ where

$$x^2 + y^2 - 2ax + 1 = 0 \quad \text{and} \quad x^2 + y^2 = 1.$$

The chord of $\mathbb{D}$ through the two points of intersection of C and the boundary of $\mathbb{D}$ is found by subtracting these two equations, and so it has equation

$$2 - 2ax = 0, \quad \text{or} \quad ax = 1.$$

It follows that the image $L(P)$ of P is the point $\left(\frac{1}{a}, 0\right) = \left(\frac{2r}{1+r^2}, 0\right)$. □

According to the Kleinian view of geometry, to define a geometry we must specify a space of points and a group of transformations acting on that space. Our new Beltrami–Klein model of hyperbolic geometry agrees with the (earlier) Poincaré model as far as the space of points is concerned: the space is the unit disc $\mathbb{D}$. The group of transformations is different, although isomorphic to the original group; the new group consists of all transformations of $\mathbb{D}$ to itself of the form LtL^{-1}, where t is a hyperbolic transformation. Clearly, if t is a hyperbolic transformation mapping $\mathbb{D}$ to itself and d-lines to d-lines, then LtL^{-1} is a transformation mapping $\mathbb{D}$ to itself and chords to chords.

To connect transformations of the form LtL^{-1} with projective geometry, we shall show that every such transformation LtL^{-1} is a projective transformation which maps $\mathbb{C}$ to itself and $\mathbb{D}$ to itself. It is sufficient to prove this in the case

$\mathbb{C} = \{(x, y) : x^2 + y^2 = 1\}$

that t is a hyperbolic reflection; for, if $t = t_1 \circ t_2 \circ \dots \circ t_n$ is a decomposition of t into hyperbolic reflections then it follows that

$$LtL^{-1} = \left(L \circ t_1 \circ L^{-1}\right) \circ \left(L \circ t_2 \circ L^{-1}\right) \circ \dots \circ \left(L \circ t_n \circ L^{-1}\right)$$

is a decomposition of LtL^{-1} into projective transformations with the required properties; then it follows that LtL^{-1} is a projective transformation with the required properties.

Theorem 1 Let t be the hyperbolic reflection in the d-line d joining the boundary points A and B of $\mathcal{D}$, and let F be the pole of the (Euclidean) line ℓ through A and B. Then LtL^{-1} is the projective reflection r in ℓ with centre F and parameter -1.

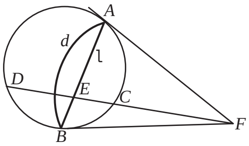

Proof First we show that LtL^{-1} agrees with r on $\mathcal{C}$. To do this, observe that lines through F are perpendicular to d, so the hyperbolic reflection r maps a given point C of $\mathcal{C}$ to the other point D of $\mathcal{C}$ that lies on FC. Since L maps $\mathcal{C}$ to itself, it follows that $LtL^{-1}(C) = D$.

Now let E be the point of intersection of FC with ℓ. Then, since F is the polar of ℓ with respect to $\mathcal{C}$, it follows that the cross-ratio $(FECD) = -1$. Hence

$$r(C) = D = LtL^{-1}(C),$$

as required.

By the result of Exercise 7 on Subsection 4.3, in Subsection 4.6.

Next we show that LtL^{-1} agrees with r on $\mathcal{D}$. To do this, let P be an arbitrary point of $\mathcal{D}$. Then let AP meet $\mathcal{C}$ at D, and let AC meet FP at Q. We will verify that $r(P) = Q$.

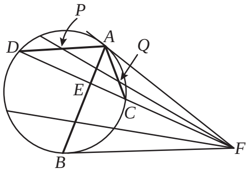

We know that $r(A) = A$ and that $r(D) = C$, so that r must map AD to AC. We also know that r maps the line FP to itself, so that it must map P (the point of intersection of AD and FP) to the point of intersection of AC and FP), namely the point Q. Thus $r(P) = Q$.

Now we consider the transformation LtL^{-1}. It certainly maps chords of $\mathcal{D}$ to chords; so, to find $LtL^{-1}(P)$ we may take AD and FP as two Lines through P, and find their images. The images of AD and FP are AC and FP, so that the image of P is Q. That is, $LtL^{-1}(P) = Q$.

Thus LtL^{-1} and r agree on both $\mathcal{C}$ and $\mathcal{D}$.

We have thus shown that every hyperbolic transformation t corresponds to a projective transformation LtL^{-1} which fixes $\mathcal{C}$ and maps its interior $\mathcal{D}$ to itself. In a similar way we can verify that every projective transformation that fixes $\mathcal{C}$ and maps $\mathcal{D}$ to itself corresponds to a hyperbolic transformation.

Here we think of $\mathcal{C}$ as being embedded in a suitable embedding plane.

We omit the details of this.

This shows that hyperbolic geometry can be considered as the sub-geometry of projective geometry in which the space of points is the set of points inside the fixed projective conic $\mathcal{C}$, and its group of transformations corresponds to those projective transformations mapping $\mathcal{C}$ to itself and the inside $\mathcal{D}$ of $\mathcal{C}$ to itself.

We defined the *inside* of a projective conic in Subsection 4.1.2.

In fact, since all non-degenerate conics are projectively equivalent, we can replace C in the above discussion by any non-degenerate conic E. This may be achieved by replacing the linearization map L by the composite map $s \circ L \circ s^{-1}$, where s is a projective transformation that maps C onto E.

We omit the details.

Defining Hyperbolic Distance in the Projective Setting

The fact that hyperbolic geometry may be considered a sub-geometry of projective geometry enables us to give a new method for defining distance in hyperbolic geometry. Projective transformations preserve cross-ratio, which involves four Points; distance involves only two points. So we shall have a ready-made sense of distance in the Beltrami–Klein model of projective geometry as soon as we can connect two interior Points of a projective conic (for simplicity, we shall take C as our projective conic) with a set of four collinear Points. This is easy — we join the two given Points P and Q by a Line that meets the projective conic in the Points A and B.

Historically, it was this discovery that excited Klein.

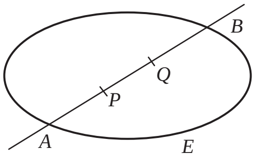

We know that the cross-ratio $(ABPQ)$ is unaltered by any projective transformation, and so in particular by the projective transformations that map the interior of the projective conic E to itself.

We now investigate whether we can sensibly define the hyperbolic distance d from P to Q by the formula $d(P,Q) = (ABPQ)$. The first thing we should want is that distance along a line is additive: that is, if R also lies on the d-line through P and Q, and R lies between P and Q, then is it true that

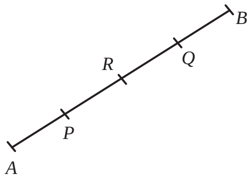

$$d(P,Q) = d(P,R) + d(R,Q)?$$

It follows from the definition of cross-ratio that

$$d(P,Q) = (ABPQ) = \frac{AP}{PB} \div \frac{AQ}{QB},$$

$$d(P,R) = (ABPR) = \frac{AP}{PB} \div \frac{AR}{RB},$$

and

$$d(R,Q) = (ABRQ) = \frac{AR}{RB} \div \frac{AQ}{QB}.$$

In this discussion we shall not bother to distinguish between Points and the corresponding points in a suitable embedding plane, for simplicity.

Hence, if we use the definition $d(P,Q) = (ABPQ)$, then $d(P,Q) \neq d(P,R) + d(R,Q)$ but instead $d(P,Q) = d(P,R) \bullet d(R,Q)$. This suggests that in the Beltrami–Klein model we should instead define distance between arbitrary points P and Q of D by the formula

$$d(P,Q) = k \log_e(ABPQ), \quad \text{for some number } k,$$

Recall that $\log_e(xy) = \log_e x + \log_e y$, for any positive numbers x and y.

since then distance d is additive. (In fact, the arbitrariness in the choice of the constant k is analogous to the freedom we have to choose the radius of the sphere in spherical geometry. However, we choose the sign of k so that distances are non-negative.)

To define a distance function d for the Poincaré model, let the conic E be the unit circle C and define the distance d between two points P and Q in D to be the distance d' between the corresponding points $P' = L(P)$ and $Q' = L(Q)$ in the Beltrami–Klein model. That is, we define

Notice that we use d' here, temporarily, to avoid confusion.

$$d(P,Q) = d'(P',Q').$$

To verify that this definition agrees with that given when we introduced hyperbolic geometry earlier, let P be the origin and let Q be an arbitrary point $(r,0)$ on the positive x-axis. Then, by Example 1, $P' = L(P)$ is also at the origin and $Q' = L(Q) = (s,0)$, where $s = \frac{2r}{1+r^2}$. Since the x-axis meets $\mathcal{C}$ at the points $A(-1,0)$ and $B(1,0)$, it follows that

Subsection 6.3.1

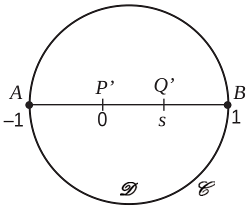

$$\begin{aligned} d(P,Q) &= d'(P',Q') \\ &= k\log_e(ABP'Q') \\ &= k\log_e\left(\frac{1}{-1} \div \frac{1+s}{-(1-s)}\right) \\ &= k\log_e\left(\frac{1-\frac{2r}{1+r^2}}{1+\frac{2r}{1+r^2}}\right) \\ &= k\log_e\left(\frac{1-2r+r^2}{1+2r+r^2}\right) \\ &= -k\log_e\left(\left(\frac{1+r}{1-r}\right)^2\right) \\ &= -2k\log_e\left(\frac{1+r}{1-r}\right). \end{aligned}$$

Here we use that

$$\log_e\left(\frac{1}{x}\right) = -\log_e(x)$$

and

$$\log_e\left(x^2\right) = 2\log_e(x).$$

By choosing $k = -\frac{1}{4}$ and representing P and Q by the complex numbers 0 and $z = r + 0i$, respectively, we obtain

$$d(0,z) = \frac{1}{2}\log_e\left(\frac{1+|z|}{1-|z|}\right).$$

This is the logarithm formula for hyperbolic distance in $\mathbb{D}$ that you met earlier.

Subsection 6.3.1

8.4 Elliptic Geometry: the Spherical Model

To bring (the spherical model of) elliptic geometry into the Kleinian fold, we need to connect its space S^2 to the space $\mathbb{RP}^2$ of projective points and its group to the group of projective transformations.

With each point of S^2 we can associate a line through the origin, the line that joins the given point to the origin; this map assigns the same line to a pair of antipodal points. Conversely, given a projective point we can associate it with the pair of diametrically opposite points that it cuts out on S^2. In this way, we see that there is a 2–1 map of S^2 to $\mathbb{RP}^2$ that sends each pair of antipodal points on S^2 to the line through the origin (the projective point) that they define.

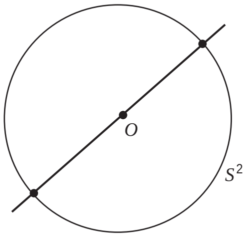

Since a spherical transformation sends antipodal points to antipodal points, it maps projective points to projective points. Also, every spherical transformation maps great circles to great circles, and so maps projective lines to projective lines. So it is clear that spherical transformations are projective transformations, and the group of spherical transformations is a subgroup of the projective group.

Projective Geometry and Spherical Geometry

Spherical geometry is a subgeometry of projective geometry which preserves the Euclidean angle between two lines through the origin in $\mathbb{R}^3$; the exact description of the subgroup is quite subtle, however, because the map from points of S^2 to Points in $\mathbb{RP}^2$ is 2–1.

The description lies beyond the scope of this book.

One method of representing a projective point by a single point on S^2 is to keep exactly half of S^2, in some sensible way. For example, let E^2 be the subset of S^2 defined by

Here we give an informal description of the representation.

$$\begin{aligned} E^2 = {} & \text{northern hemisphere}, \{(x,y,z) : z > 0\} \\ & \cup\ \text{half the equator}, \{(x,y,z) : z = 0, x > 0\} \cup \{B(0,1,0)\}. \end{aligned}$$

Then every line through the origin meets E^2 in exactly one point, and every projective line meets E^2 in half a great circle.

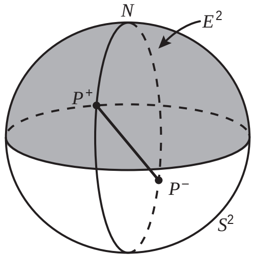

The 2–1 map from S^2 to $\mathbb{RP}^2$ therefore gives a 1–1 map from S^2 to E^2: every point of $\mathbb{RP}^2$ corresponds to a unique point of E^2, and every point of E^2 defines a unique Point of $\mathbb{RP}^2$. So E^2 provides 'a complete picture' of projective space $\mathbb{RP}^2$.

To see what's odd about E^2, consider the Points corresponding to the points $(\sin\theta, 0, \cos\theta)$ on S^2 as they trace out the Line $y = 0$ in $\mathbb{RP}^2$. Each such Point gives rise to two points on S^2:

$$P^+ = (\sin\theta, 0, \cos\theta) \quad \text{and} \quad P^- = (-\sin\theta, 0, -\cos\theta).$$

As θ goes from 0 to $\pi/2$, the point P^+ lies in E^2. But when $\theta > \pi/2$, the point P^- lies in E^2, rather than the point P^+, and this continues as θ goes from $\pi/2$ to $3\pi/2$ (including $\theta = 3\pi/2$). Then as θ goes from $3\pi/2$ to 2π, it is again the point P^+ that lies in E^2.

Here we discuss the points of S^2 that correspond to the complete Line $y = 0$.

So in E^2 the image of the Line is one-half of a great circle, described twice. There is an awkward switch in E^2 from the Euclidean point corresponding to the projective point being P^+ to being P^-. We cope with this by imagining the pairs of antipodal points to be 'joined up'; that is, we agree to regard antipodal points in the equator of S^2 as 'being the same point'.

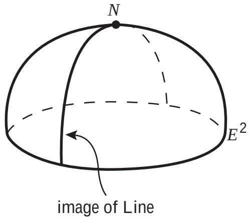

Conics appear equally odd. For example, let E denote a projective conic made up of Euclidean lines through the origin of $\mathbb{R}^3$ that form a right circular (Euclidean) cone in $\mathbb{R}^3$. This cone meets S^2 in two *little circles*.

Suppose first that the points P^+ all lie in E^2, so that the cone does not meet the equatorial plane in S^2. Then the points in E^2 that correspond to E form a little circle on S^2.

Suppose next that the cone cuts the equatorial plane of S^2. Then we must switch from P^+ to P^- as the Point describes the Line, and the points in E^2 that correspond to E form two curves (parts of little circles) on S^2.

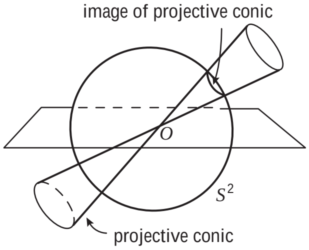

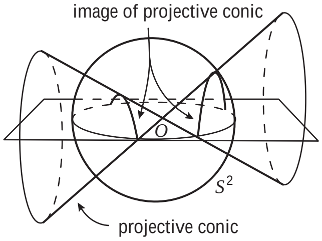

To see what s really odd about this, thicken up the Line until it is a strip bounded by two little circles in S^2. The boundaries of the strip can be taken to be two projective conics, one on either side of the Line. Since the image of the Line meets the equator in S^2, we know that the images of the (associated) projective conics also meet the equator of S^2.

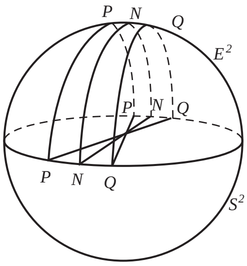

Now trace the image of the Line on S^2. As you trace the Line from N to N , the curve PP (which is part of a projective conic $PP\ P$) is to your right, and the curve QQ (which is part of a projective conic $QQ\ Q$) is to your left.

When you reach the equator, the equatorial points N and N are identified, and you can complete your journey round S^2 to return to N. But the point P is identified with Q , so that the moving point that started from P returns to Q; similarly, the point Q is identified with P , so that the moving point that started from Q returns to P.

The curve $P\ \ PP$ that was on your right is now on your left, and the curve $Q\ \ QQ$ that was on your left is now on your right. The image of the strip is a strip in which the ends have been joined after receiving a half-twist. The image is a *Möbius band*, which illustrates a significant mathematical point: 2-dimensional real projective space $\mathbb{RP}^2$ is not *orientable* that is, it cannot be given an everywhere-consistent sense of left and right.

Möbius bands are beloved of conjurers!

That is, of *orientation*.

Inversive Geometry and Spherical Geometry

In inversive geometry, we study transformations which map generalized circles to generalized circles and preserve the magnitude of angles. Every transformation in inversive geometry is composed of a sequence of inversions.

In spherical geometry, we study transformations which map great circles to great circles and preserve the magnitude of angles. Every transformation in spherical geometry is composed of a sequence of reflections in great circles, which are therefore inversions. It follows that spherical geometry is a subgeometry of inversive geometry, in exact analogy with the way that Euclidean geometry is a subgeometry of affine geometry.

We omit the full details.

8.5 Euclidean Geometry

Spherical geometry appears easier to understand than hyperbolic geometry because we can produce spheres S^2 as a model of the geometry, but there seems to be no surface in Euclidean 3-dimensional space which is a good model of hyperbolic geometry. One may exhibit pieces of hyperbolic geometry, but it was shown by David Hilbert around 1900 that *no* surface embedded in 3-dimensional space is a complete model of hyperbolic 2-dimensional space.

David Hilbert (1862 1943), a professor at Göttingen University, was a mathematical universalist, influential in the foundations of mathematics; he is best known for his 1900 23 Problems collection that profoundly influenced the course of 20th-century mathematics.

But it can be argued that space might be hyperbolic (indeed, some would say that, according to Einstein s theory of General Relativity, it is *more nearly* hyperbolic than Euclidean). This raises the question: Is Euclidean geometry a subgeometry of hyperbolic geometry? We sketch a discussion of the answer to this interesting question.

Remarkably, the answer to the question is YES: Euclidean geometry *is* a subgeometry of hyperbolic geometry

It lies beyond our scope to deal with the matter fully.

To see that this is so, we need first to go up a dimension . Hyperbolic 3-dimensional space can be modelled on the interior of the *unit ball*: a hyperbolic point is a point (x, y, z) such that $x^2 + y^2 + z^2 < 1$, and a d-line is an arc of a generalized circle perpendicular to S^2. All the other definitions in hyperbolic geometry generalize in the obvious way, including that of inversion and hyperbolic reflection. In particular, the transformations of hyperbolic geometry correspond to composites of inversions in spheres perpendicular to S^2.

$S^2 = (x, y, z)$: $x^2 + y^2 + z^2 = 1$.

In the unit ball *horospheres* are spheres that touch S^2 at a single point. For simplicity, we now fix our attention on the horosphere H_N that touches S^2 at N, and restrict our attention to the subgroup G_N, say, of hyperbolic transformations that map H_N to itself. Then we have the following remarkable result.

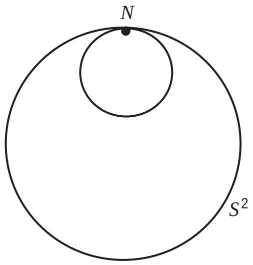

Theorem 1 The geometry on H_N is Euclidean geometry, and the group G_N acts on H_N as Euclidean isometries.

Proof Invert the whole figure in a sphere with centre N. The point N is mapped to infinity ; S^2 maps to a plane : $z = a$, say, and H_N maps to a plane parallel to , the plane : $z = b$, say. An element of G_N is an inversion in a generalized sphere through N and perpendicular to H_N; following composition with the first inversion, the element of G_N becomes an inversion in an unbounded generalized sphere perpendicular to the plane . These generalized spheres must in fact be planes perpendicular to the plane .

Now, inversion in a plane is Euclidean reflection in that plane. So the subgroup mapping the image of H_N to itself is the group generated by all Euclidean reflections, which is the group of Euclidean isometries. It follows that the geometry on the plane is Euclidean geometry. The desired result then follows by applying the inverse of the original inversion.

If indeed we had grown up believing that we lived in a hyperbolic 3-dimensional space, we should therefore be very familiar with spheres and horospheres!

Summary of Chapter 8

Section 8.1: Affine Geometry

1. The group of affine transformations can be identified with the subgroup of projective transformations that map $\mathbb{RP}^2_0$ onto itself.
2. Euclidean geometry is a subgeometry of affine geometry, which is itself a subgeometry of projective geometry; the three form a hierarchy. The higher up the geometry in the hierarchy, it has more transformations, fewer properties and fewer congruence classes.

Section 8.2: Projective Reflections

1. Let F be a Point that does not lie on a Line ℓ. Then a **projective reflection** r in ℓ with centre F and non-zero parameter k is a mapping of $\mathbb{RP}^2$ to itself, with P mapping to Q, where (i) $Q = P$ if P lies on ℓ and (ii) the cross-ratio $(FRPQ) = k$, where R is the Point of intersection of FP and ℓ.
2. A projective reflection is a projective transformation. Every projective transformation can be expressed as a composite of a finite number of projective reflections.

Section 8.3: Hyperbolic Geometry and Projective Geometry

1. **Orthogonal projection** maps $(x, y, z) \in \mathbb{R}^3$ to $(x, y, 0) \in \mathbb{R}^3$ or equivalently, to $(x, y) \in \mathbb{R}^2$.
2. The **linearization map** L of the unit disc $\mathbb{D}$ onto itself consists of the inverse of the stereographic projection map followed by the orthogonal projection map.
3. In the **Beltrami–Klein model** of hyperbolic geometry the space is $\mathbb{D}$ and the d-lines are straight-line segments.
4. Let t be the hyperbolic reflection in the d-line d joining the boundary points A and B of $\mathbb{D}$, and let F be the pole of the (Euclidean) line ℓ through A and B. Then LtL^{-1} is the projective reflection r in ℓ with centre F and parameter -1.
5. Every projective transformation that fixes $\mathcal{C}$ and maps $\mathbb{D}$ to itself corresponds to a hyperbolic transformation.

   Hyperbolic geometry can be considered as a subgeometry of projective geometry, where the points are points inside a fixed projective conic $\mathcal{C}$ and the group of transformations corresponds to those projective transformations mapping $\mathcal{C}$ onto itself and the inside of $\mathcal{C}$ to itself.
6. In the Beltrami–Klein model, we can define the **distance** between two points P and Q of $\mathbb{D}$ (where PQ meets $\mathcal{C}$ at A and B) by $d(P, Q) = k \log_e (ABPQ)$, where k is chosen so that distances are non-negative. This is equivalent to the definition of distance in the Poincaré model.

Section 8.4: Elliptic Geometry: the Spherical Model

1. There is a 2 1 map of S^2 to $\mathbb{RP}^2$ that sends each pair of antipodal points on S^2 to the line through the origin that they define. There is a 1-1 map of $E^2 = \{(x, y, z) : z > 0\} \quad \{(x, y, z) : z = 0, x > 0\} \quad \{(0, 1, 0)\}$ to $\mathbb{RP}^2$.
2. Spherical transformations are projective transformations; the group of spherical transformations is a subgroup of the projective group.
3. Spherical geometry is a subgeometry of inversive geometry.

Section 8.4: Euclidean Geometry

1. Euclidean geometry is a subgeometry of hyperbolic geometry
2. A point in 3-dimensional hyperbolic space is a point (x, y, z) with $x^2 + y^2 + z^2 < 1$, and a d-line is an arc of a generalized circle perpendicular to S^2.
3. A **horosphere** is a sphere in $\mathbb{R}^3$ that touches S^2 at a single point.

   Let H_N be the horosphere that touches S^2 at N, and G_N the subgroup of hyperbolic transformations that map H_N to itself. Then the geometry on H_N is Euclidean geometry, and the group G_N acts on H_N as Euclidean isometries.

Chapter 4

Volume-05---20211111

! " # " $ " ! ! " ! %& ' # (!")
* + ! (!"
& ! ,! " " $!$ $! ! " ' !

(-& .- - -/ ' ! - / 0 1 ! ! "
2 1 ! 3 4 0# & !" % 5 6-/ 0 1 ! ! "
2 1 ! 3 4 0# & !" % 5 6 $"
7' 5%) 6 8 $! %9 : (" 6

(-& .- - -/ ' ! - / 0 1 ! ! "
2 1 ! 3 4 0# & !" % 5 6- / 0 1 !
! " 2 1 ! 3 4 0# & !" % 5 6 $"

It may be mentioned that a formula similar to (3.9) can also be obtained from (3.5) in replacing there $f(x_3)$ by $f'(\xi')(x_3 - \zeta)$, y_1 by $(x_1 - \zeta)f'(\xi_1)$, y_2 by $(x_2 - \zeta)f'(\xi_2)$, and $(y_2 - y_1)/(x_2 - x_1)$ by $f'(\xi_3)$. Then we get

$$\zeta - x_3 = \left[-\frac{f''(\xi)f'(\xi_1)f'(\xi_2)}{2f'(\xi')f'(\xi_3)^2}\right](\zeta - x_1)(\zeta - x_2).$$

Here the bracketed coefficient is less easy to handle than the corresponding coefficient in (3.9), since in the case $\xi' = \xi_3 = \xi$ the expression $f''(\xi)/f'(\xi)^3$ can be better estimated. On the other hand, the last formula also gives the inequality (3.10).

GEOMETRIC INTERPRETATION (FOURIER'S CONDITIONS)

6. We consider now the case where $y = f(x)$ has a graph as indicated below. Taking x_0 and x_1 as our initial approximations, we

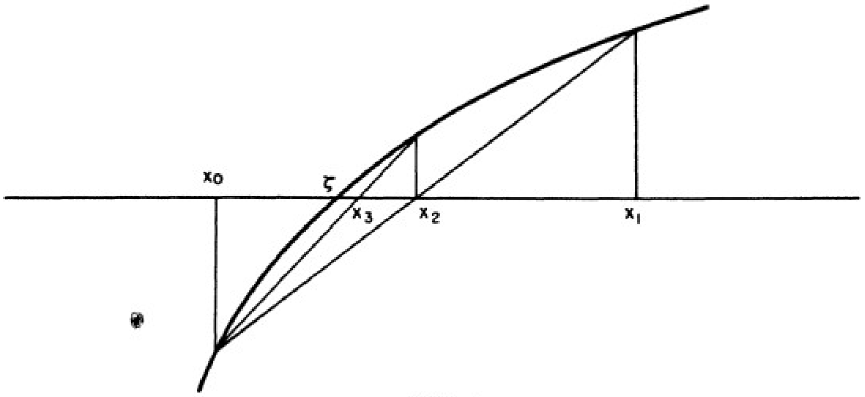

Fig. 1

obtain x_2. Taking now x_2 and x_0 as approximations, we obtain x_3. Continuing this process, we obtain a sequence of points $x_1, x_2, x_3, \ldots$, where generally

$$x_{\nu+1} = \frac{x_0 f(x_\nu) - x_\nu f(x_0)}{f(x_\nu) - f(x_0)} \qquad (\nu = 1, 2, \ldots). \tag{3.12}$$

Does x_ν $(\nu = 1, 2, \ldots)$ converge?

7. If the geometric situation is as we have drawn in Fig. 1, then x_ν converges indeed. For $x_1, x_2, x_3, \ldots$ lie on the concave side of the arc and cannot go beyond ζ; i.e., we have a monotonic decreasing sequence which is bounded from below by ζ and hence must converge to a limit ζ_0.

Is ζ_0 the root ζ of $f(x) = 0$? Subtracting ζ_0 from both sides of (3.12) and taking the limit as $\nu \to \infty$, we obtain

$$0 = \frac{x_0 f(\zeta_0) - \zeta_0 f(x_0) - \zeta_0 [f(\zeta_0) - f(x_0)]}{f(\zeta_0) - f(x_0)} = f(\zeta_0) \frac{x_0 - \zeta_0}{f(\zeta_0) - f(x_0)} \cdot$$

Now $x_0 \neq \zeta_0$; hence, $f(\zeta_0) = 0$ and ζ_0 is a zero of $f(x)$ in J_x and therefore equal to ζ.

To use the argument outlined above, the curve must have no points of inflection in the considered domain; for instance, it is sufficent to assume that $f''(x) \neq 0$ throughout J_x. As to x_0, we must take it in such a way that $f(x_0)f''(x_0) > 0$.

If the conditions listed above (the so-called *Fourier conditions*) are not satisfied, e.g., if x_0 is a point for which $\operatorname{sgn} f(x_0) \neq \operatorname{sgn} f''(x_0)$, we may get a sequence which still converges to ζ but oscillates about ζ. Such a case will be considered in Chapter 5. See also Example 1 in Section 17.

ITERATION WITH SUCCESSIVE ADJACENT POINTS

8. The convergence of the method of false position is considerably improved in the long run if, instead of constantly using x_0 as one of the interpolation points, we use at each step the two last points of the sequence:

$$x_{\nu+1} = \frac{x_{\nu-1} f(x_\nu) - x_\nu f(x_{\nu-1})}{f(x_\nu) - f(x_{\nu-1})} \qquad (\nu = 1, 2, \ldots). \tag{3.13}$$

Applying (3.9) to (3.12), we replace x_1, x_2, x_3 by x_0, x_ν, $x_{\nu+1}$ and obtain

$$\zeta - x_{\nu+1} = \left[-\frac{f''(\xi)}{2f'(\xi)^3} f'(\xi_1) f'(\xi_2)(\zeta - x_0) \right] (\zeta - x_\nu), \tag{3.14}$$

$$\xi \prec (\zeta, x_0, x_\nu), \quad \xi_1 \prec (x_0, \zeta), \quad \xi_2 \prec (x_\nu, \zeta).$$

If the modulus of the bracketed expression is always $\leqq \rho$, $\rho < 1$, and does not tend to zero, we have what is called *linear convergence.* From (3.14) it follows then in the case of the formula (3.12)

$$|\zeta - x_\nu| \leqq \rho^{\nu-1} |\zeta - x_1|, \qquad \rho < 1. \tag{3.15}$$

7' ; %<;6 =! " > " 3 !$!
(-& .- - -/ ' ! - / 0
1 ! ! " 2 1 ! 3 4 0# & !" % 5 6- / 0
1 ! ! " 2 1 ! 3 4 0# & !" % 5 6 $"

10

Dandelin Bounds for Newton-Raphson Iteration

1. We assume as in Chapter 9 replacing there y_0 by z_0 that in the interval J_0: $x_0 \leqq x \leqq z_0$, the Fourier conditions and in particular the conditions (9.0) are satisfied. Let x_0 and z_0 be the starting points, where the sequence x_ν $(\nu = 0, 1, \ldots)$ is defined by the Newton-Raphson formula (6.3) and the sequence z_ν is obtained by applying the *regula falsi* to $x_{\nu-1}$ and $z_{\nu-1}$; i.e.,

$$z_{\nu+1} = z_\nu - f(z_\nu)\,\frac{z_\nu - x_\nu}{f(z_\nu) - f(x_\nu)} \qquad (\nu = 0, 1, \ldots). \qquad (10.1)$$

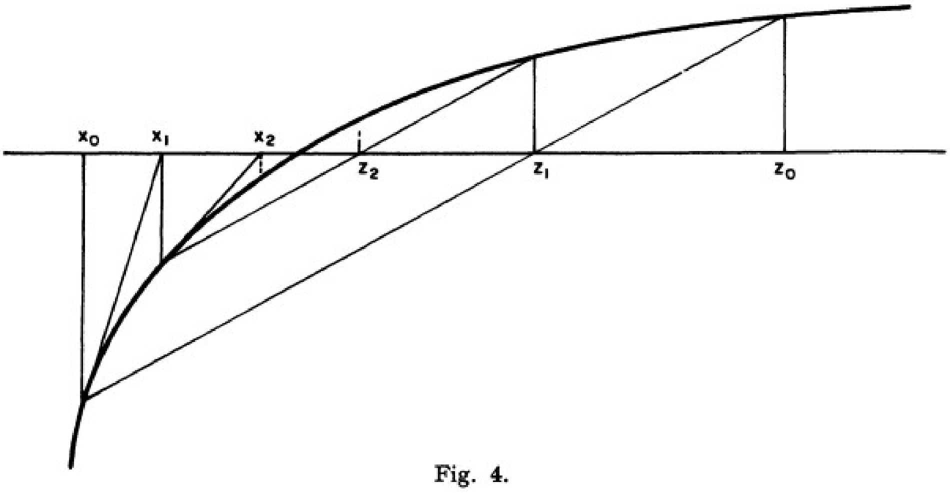

Fig. 4.

We see from Fig. 4 immediately that the z_ν decrease monotonically to a certain number η and it follows now from (10.1) that

$$\frac{f(z_\nu)}{[f(z_\nu) - f(x_\nu)]/(z_\nu - x_\nu)}$$

tends to zero. But here the denominator remains between two positive bounds $[\min f'(x), \max f'(x)]$, and therefore $f(\eta) = 0$, $\eta = \zeta$, $z_\nu \downarrow \zeta$. x_ν and z_ν are *Dandelin's bounds* for ζ.

We introduce μ_ν by

$$\mu_\nu = \frac{z_\nu - \zeta}{\zeta - x_\nu} \qquad (\nu = 0, 1, \ldots) \tag{10.2}$$

and put $J_1 = \langle x_0, y_0 \rangle$ for a fixed $y_0 > \zeta$. Then we will prove

Theorem 10.1. *If $f^{(3)}(x)$ is continuous in J_0, the sequence μ_ν tends to a limit $\Lambda(z_0)$ with $\nu \to \infty$, uniformly for $\zeta < z_0 \leqq y_0$, where $\Lambda(z_0)$ is a continuous function of z_0, and decreases monotonically to 0 as z_0 tends to ζ.*

2. Proof. We have identically

$$\frac{z_{\nu+1} - \zeta}{z_\nu - \zeta} = 1 - \frac{f(z_\nu) - f(\zeta)}{z_\nu - \zeta} \frac{z_\nu - x_\nu}{f(z_\nu) - f(x_\nu)} \tag{10.3}$$

and can write

$$\frac{z_{\nu+1} - \zeta}{z_\nu - \zeta} = \frac{N}{D},$$

$$N = \frac{f(x_\nu) - f(z_\nu)}{x_\nu - z_\nu} - \frac{f(\zeta) - f(z_\nu)}{\zeta - z_\nu}, \qquad D = \frac{f(z_\nu) - f(x_\nu)}{z_\nu - x_\nu}. \tag{10.4}$$

The formula (9.7) is an identity true for all $\zeta \neq y_\nu$; hence, we can write

$$\frac{f(a) - f(b)}{a - b} = \int_0^1 f'[b + t(a - b)]\, dt = \int_0^1 f'[a + t(b - a)]\, dt \qquad (a \neq b). \tag{10.5}$$

Applying (10.5) to D and N in (10.4), we have

$$D = \int_0^1 f'[x_\nu + t(z_\nu - x_\nu)]\, dt, \tag{10.6}$$

$$N = \int_0^1 \{f'[z_\nu + t(x_\nu - z_\nu)] - f'[z_\nu + t(\zeta - z_\nu)]\}\, dt. \tag{10.7}$$

Now apply (10.5) to the integrand of (10.7); we obtain

$$N = -(\zeta - x_\nu)\int_0^1\!\!\int_0^1 f''[z_\nu + t(x_\nu - z_\nu) + ut(\zeta - x_\nu)]t\,dt\,du. \tag{10.8}$$

3. Now put

$$\frac{z_{\nu+1} - \zeta}{(z_\nu - \zeta)(\zeta - x_\nu)} = \frac{N/(\zeta - x_\nu)}{D} = \frac{N^*}{D}, \qquad N^* = \frac{N}{\zeta - x_\nu} \tag{10.9}$$

and consider

$$S = \int_0^1\!\!\int_0^1 f''(x_\nu + uth)t\,du\,dt, \tag{10.10}$$

where $h = \zeta - x_\nu$. We introduce, instead of u, a new variable of integration, y, by $y = x_\nu + uth$, and have

$$hS = \int_0^1 \int_{x_\nu}^{x_\nu + th} f''(y)\,dy\,dt = \int_0^1 [f'(x_\nu + th) - f'(x_\nu)]\,dt,$$

$$h^2S = \int_0^1 f'(x_\nu + th)h\,dt - hf'(x_\nu).$$

Since the integrand is $d[f(x_\nu + th)]/dt$, and $x_\nu + h = \zeta$, we have

$$h^2S = f(\zeta) - f(x_\nu) - hf'(x_\nu) = -[f(x_\nu) + hf'(x_\nu)],$$

$$-\frac{h^2S}{f'(x_\nu)} = \frac{f(x_\nu)}{f'(x_\nu)} + h,$$

$$-\frac{h^2S}{f'(x_\nu)} = x_\nu - x_{\nu+1} + \zeta - x_\nu = \zeta - x_{\nu+1},$$

$$\frac{-S}{f'(x\ )} = \frac{\zeta - x_{\nu+1}}{(\zeta - x_\nu)^2}. \tag{10.11}$$

Dividing (10.9) by (10.11), we have from (10.2)

$$\frac{(z_{\nu+1}-\zeta)/(\zeta-x_{\nu+1})}{(z_\nu-\zeta)/(\zeta-x_\nu)} = -\frac{N^*}{S}\frac{f'(x_\nu)}{D} = \frac{\mu_{\nu+1}}{\mu_\nu}. \qquad (10.12)$$

4. We will have to show that μ_ν tends to a limit. First we state two well-known theorems on infinite products which we will need but shall not prove here.

Theorem A. *If* $\Sigma_{\nu=0}^{\infty}|c_\nu - 1|$ *converges and no* c_ν *is* $= 0$, *then* $\Pi_{\nu=0}^{\infty} c_\nu$ *converges to a limit* $\neq 0$.

Theorem B. *If the factors* c_ν *are continuous functions of a point and if* $\Sigma_{\nu=0}^{\infty}|c_\nu - 1|$ *converges uniformly for all points in a domain, then* $\Pi_{\nu=0}^{\infty} c_\nu$ *converges to a continuous function in this domain.*

We write now

$$\mu_n = \mu_0 \prod_{\nu=0}^{n-1} \frac{\mu_{\nu+1}}{\mu_\nu} = \mu_0 \frac{\prod_{\nu=0}^{n-1} Q_\nu}{\prod_{\nu=0}^{n-1} q_\nu}, \qquad (10.13)$$

where by (10.12)

$$Q_\nu = -\frac{N^*}{S}, \qquad q_\nu = \frac{D}{f'(x_\nu)} \qquad (\nu = 0, 1, \ldots). \qquad (10.14)$$

5. Consider first

$$Q_\nu - 1 = \frac{-N^* - S}{S}. \qquad (10.15)$$

From (10.8), (10.9), and (10.10) follows

$$-N^* - S$$
$$= \int_0^1\int_0^1 \{f''[z_\nu + t(x_\nu - z_\nu) + ut(\zeta - x_\nu)] - f''[x_\nu + ut(\zeta - x_\nu)]\}\, t\, dt\, du.$$

Applying (10.5) to the integrand, we have

$$-N^* - S$$
$$= (z_\nu - x_\nu)\int_0^1\int_0^1\int_0^1 t(1-t) f^{(3)}[x_\nu + ut(\zeta - x_\nu) + w(1-t)(z_\nu - x_\nu)]\, dw\, dt\, du.$$

Since $t(1-t)$ is nonnegative in $\langle 0, 1\rangle$, we apply the generalized mean value theorem of the integral calculus and obtain

$$-N^*-S=(z_\nu-x_\nu)f^{(3)}(\xi)\int_0^1\int_0^1\int_0^1(1-t)t\,dt\,dw\,du,\qquad \xi\prec(x_\nu, z_\nu).$$

Integrating, we have

$$-N^*-S=\frac{f^{(3)}(\xi)}{6}(z_\nu-x_\nu),\qquad \xi\prec(x_\nu, z_\nu).\tag{10.16}$$

From (10.10) the generalized mean value theorem leads to

$$S=f''(\eta)\int_0^1\int_0^1 t\,dt\,du,\qquad S=\frac{f''(\eta)}{2},\qquad \eta\prec(x_\nu, z_\nu).\tag{10.17}$$

From (10.15), (10.16), and (10.17) follows

$$Q_\nu-1=\frac{z_\nu-x_\nu}{3}\frac{f^{(3)}(\xi)}{f''(\eta)},\qquad \xi,\eta\prec(x_\nu, z_\nu).\tag{10.18}$$

6. On the other hand, by (10.14) and (10.6),

$$q_\nu-1=\frac{D-f'(x_\nu)}{f'(x_\nu)},\tag{10.19}$$

$$D-f'(x_\nu)=\int_0^1\{f'[x_\nu+t(z_\nu-x_\nu)]-f'(x_\nu)\}\,dt.\tag{10.20}$$

Applying (10.5) to (10.20), we obtain by the generalized mean value theorem

$$D-f'(x_\nu)=(z_\nu-x_\nu)\int_0^1\int_0^1 f''[x_\nu+ut(z_\nu-x_\nu)]t\,dt\,du$$

$$=(z_\nu-x_\nu)f''(\xi_1)\int_0^1\int_0^1 t\,dt\,du,\qquad \xi_1\prec(x_\nu, z_\nu).$$

$$D-f'(x_\nu)=\frac{z_\nu-x_\nu}{2}f''(\xi_1),\qquad \xi_1\prec(x_\nu, z_\nu),$$

and from (10.19)

$$q_\nu - 1 = \frac{z_\nu - x_\nu}{2} \frac{f''(\xi_1)}{f'(x_\nu)}, \qquad \xi_1 \prec (x_\nu, z_\nu). \tag{10.21}$$

7. We introduce M_k and m_k by

$$\max_{\langle x_0, y_0 \rangle} |f^{(k)}(x)| = M_k, \qquad \min_{\langle x_0, y_0 \rangle} |f^{(k)}(x)| = m_k \qquad (k = 1, 2, 3). \tag{10.22}$$

From (10.18) and (10.21) we have then

$$|Q_\nu - 1| \leqq |z_\nu - x_\nu| K, \qquad |q_\nu - 1| \leqq |z_\nu - x_\nu| K, \tag{10.23}$$

where

$$K = \max\left(\frac{1}{3}\frac{M_3}{m_2}, \frac{1}{2}\frac{M_2}{m_1}\right). \tag{10.24}$$

8. But, on the other hand, from (6.9) and (9.17), we have

$$\frac{\zeta - x_{\nu+1}}{(\zeta - x_\nu)^2} \to \kappa, \qquad \frac{\zeta - x_{\nu+1}}{\zeta - x_\nu} \sim \kappa(\zeta - x_\nu) \to 0.$$

Hence $\Sigma_{\nu=0}^{\infty} |\zeta - x_\nu|$ is convergent.

Similarly, from (3.11) and (9.17) we have

$$\frac{\zeta - z_{\nu+1}}{(\zeta - z_\nu)(\zeta - x_\nu)} \to \kappa, \qquad \frac{\zeta - z_{\nu+1}}{\zeta - z_\nu} \sim \kappa(\zeta - x_\nu) \to 0,$$

and $\Sigma_{\nu=0}^{\infty} |\zeta - z_\nu|$ is convergent. From

$$z_\nu - x_\nu = z_\nu - \zeta + \zeta - x_\nu, \qquad |z_\nu - x_\nu| \leqq |\zeta - x_\nu| + |\zeta - z_\nu|$$

the convergence of $\Sigma_{\nu=0}^{\infty} |z_\nu - x_\nu|$ now follows.

9. We now show that $\Sigma_{\nu=0}^{\infty} |z_\nu - x_\nu|$ is *uniformly convergent* as a function of z_0 if, for a fixed y_0, $\zeta < z_0 \leqq y_0$. Notice that the points x_ν do not depend on z_ν. We will keep x_0 fixed and discuss what happens as $z_0 \downarrow \zeta$. Now all z_ν are strictly increasing functions of z_0. The initial value of z_0 is $y_0 = z_0'$; denote corresponding values of z_ν by z_ν'. Then

$$z_\nu \leqq z_\nu' \qquad (\nu = 0, 1, \ldots),$$

and hence

$$|z_\nu - x_\nu| \leqq |z_\nu' - x_\nu|.$$

Therefore by (10.23)

$$|Q_\nu - 1| \leqq K\,|z_\nu' - x_\nu|, \qquad |q_\nu - 1| \leqq K|z_\nu' - x_\nu|.$$

This holds for $\zeta < z_0 \leqq y_0$. Hence $\Sigma_{\nu=0}^{\infty}|Q_\nu - 1|$, $\Sigma_{\nu=0}^{\infty}|q_\nu - 1|$ converge *uniformly* if z_0 varies between ζ and y_0. We see by Theorem B that both products $\Pi_{\nu=0}^{\infty}Q_\nu$ and $\Pi_{\nu=0}^{\infty}q_\nu$ converge uniformly for these z_0, and from (10.13) it follows now that $\lim_{\nu\to\infty}\mu_\nu$ exists uniformly for all z_0 between ζ and y_0. Now the z_ν are rational and continuous functions of z_0. By (10.2) the μ_ν are also continuous functions of z_0 and we see that $\Lambda(z_0)$ is a continuous function of z_0 for $\zeta < z_0 \leqq y_0$.

10. If $z_0 = \zeta$, then each $z_\nu = \zeta$ $(\nu = 0, 1, \ldots)$ and it follows that for $z_0 \downarrow \zeta$ each $\mu_\nu(z_0)$ tends to 0. As each $\mu_\nu(z_0)$ is a monotonically increasing function of z_0, so is $\Lambda(z_0)$ in the interval $\zeta < z_0 \leqq y_0$. Therefore, as $\Lambda(z_0) \geqq 0$, it tends to limit Λ_0 as $z_0 \downarrow \zeta$. On the other hand, for any positive ε there exists a $\mu_k(z_0)$ such that we have

$$|\Lambda(z_0) - \mu_k(z_0)| \leqq \varepsilon \qquad (\zeta < z_0 \leqq y_0).$$

We have therefore

$$\Lambda(z_0) \leqq \varepsilon + \mu_k(z_0)$$

and, as $z_0 \downarrow \zeta$, $\Lambda_0 \leqq \varepsilon$. We see that $\Lambda_0 = 0$, i.e., $\Lambda(z_0) \downarrow 0$ $(z_0 \downarrow \zeta)$. It follows that the error from the z_ν side can be made an arbitrarily small part of the error from the x_ν side if we start with a sufficiently close point z_0.

(-& .- - -/ ' ! - ! 0 ! * & $ 2 ! "
('2 $ ' !'- ! 0 ! * & $ 2 ! " ('2 $ '
!' $"
7' %)<6

(-& .- - -/ ' ! - ! 0 ! *
& $ 2 ! " ('2 $ ' !'- ! 0 ! * & $ 2 ! "
('2 $ ' !' $"

Gauge theories can be formulated in the *background field approach*, as introduced by B. de Witt and others (cf. [20]). In this approach, the quantum field theory depends on a background field, but not on any one in particular, and the theory enjoys background gauge invariance.

Is it enough to have quantum gravity formulated in such a way?[1]

It can be argued that the only vacuum expectation value consistent with diffeomorphisms invariance is

$$\langle 0|g_{\alpha\beta}|0\rangle = 0 \tag{2}$$

in which case the answer to the above question ought to be in the negative, because this is a singular background and curvature invariants do not make sense. It all boils down as to whether the ground state of the theory is diffeomorphism invariant or not. There is an example, namely three-dimensional gravity in which invariant quantization can be performed [70]. In this case at least, the ensuing theory is almost topological.

In all attempts of a canonical quantization of the gravitational field, one always ends up with an (constraint) equation corresponding physically to the fact that the total hamiltonian of a parametrization invariant theory should vanish. When expressed in the Schrödinger picture, this equation is often dubbed the *Wheeler-de Witt equation*. This equation is plagued by operator ordering and all other sorts of ambiguities. It is curious to notice that in ordinary quantum field theory there also exists a Schrödinger representation, which came recently to be controlled well enough as to be able to perform lattice computations [42].

Gauge theories can be expressed in terms of gauge invariant operators, such as Wilson loops . They obey a complicated set of equations, the loop equations, which close in the large N limit as has been shown by Makeenko and Migdal [43]. These equations can be properly regularized, e.g. in the lattice. Their explicit solution is one of the outstanding challenges in theoretical physics. Although many conjectures have been advanced in this direction, no definitive result is available.

3 The Canonical Approach

It is widely acknowledged that there is a certain tension between a $(3+1)$ decomposition implicit in any canonical approach, privileging a particular notion of time, and the beautiful geometrical structure of general relativity, with its invariance under general coordinate transformations.

Let us now nevertheless explore how far we can go on this road, following the still very much worth reading work of De Witt [20].

---

[1] This was, incidentally, the way G. Hooft and M. Veltman did the first complete one-loop calculation [65].

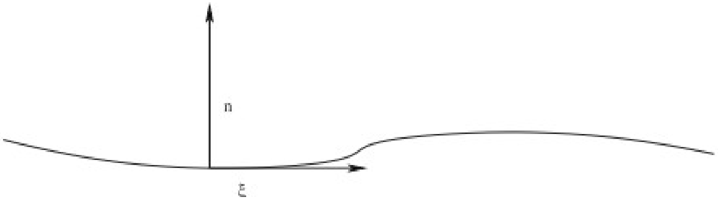

Fig. 1. Spacelike surface of codimension one

If (Fig. 1) we are given a spacelike surface (which will represent physically all spacetime events to which it will be assigned a fixed time), say

$$y^\alpha = f^\alpha(x^i)$$

The tangent vector to the surface are

$$\xi_i \equiv \partial_i f^\alpha \partial_\alpha$$

and the induced metric (that is, the pull-back to the surface of the spacetime metric) is

$$h_{ij} \equiv g_{\alpha\beta}\xi_i^\alpha \xi_j^\beta$$

The unit normal is then defined as;

$$\begin{aligned} &g_{\alpha\beta} n^\alpha \xi_i^\beta = 0 \\ &n^2 \equiv g_{\alpha\beta} n^\alpha n^\beta = 1 \end{aligned}$$

We are interested now in a set of such surfaces which covers all spacetime; that is, a foliation of (a portion of) the spacetime; namely a one-parameter family of spacelike disjoint hypersurfaces

$$\Sigma_t \equiv \{y^\alpha = f^\alpha(x^i, t)\}$$

In a classical analysis Arnowitt, Deser and Misner (ADM) [1] characterized the embedding via two functions: the *lapse* and the *shift* : we first define the vector (see Fig. 2)

$$N^\alpha \equiv \frac{\partial f^\alpha}{\partial t}$$

in terms of which the lapse, N, is just the projection in the direction of the normal, and the *shift*, N_i the (three) projections tangent to the hypersurface.

$$N^\alpha = N n^\alpha + N^i \xi_i^\alpha$$

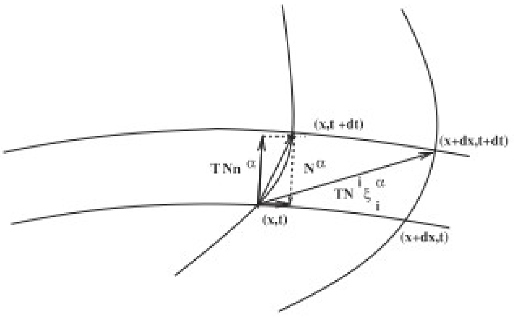

Fig. 2. ADM lapse and shift variables

All this amounts to a particular splitting of the full spacetime metric:

$$\begin{aligned} ds^2 &= g_{\alpha\beta} dx^\alpha dx^\beta = g_{\alpha\beta} df^\alpha df^\beta \\ &= g_{\alpha\beta}(N^\alpha dt + \xi_i^\alpha dx^i)(N^\beta dt + \xi_j^\beta dx^j) \\ &= N^2 dt^2 + h_{jk}(N^j dt + dx^j)(N^k dt + dx^k) \end{aligned}$$

or, what is the same,

$$g_{\mu\nu} = h^{ij}\xi_{i\mu}\xi_{j\nu} + n_\mu n_\nu$$

All surfaces which are equivalent from the intrinsic point of view, can be however embedded differently; the extrinsic curvature discriminates between them:

$$K_{ij} = -\xi_i^\alpha \nabla_\rho n_\alpha \xi_j^\rho$$

The Gauss-Codazzi equations relate intrinsic curvatures associated with the intrinsic geometry in the hypersurface with spacetime curvatures precisely through the extrinsic curvature:

$$R[h]_{lijk} = R[g]_{\alpha\beta\sigma\rho}\xi_l^\alpha \xi_i^\beta \xi_j^\sigma \xi_k^\rho - K_{ij}K_{lk} + K_{ik}K_{lj} \tag{3}$$

and

$$\nabla[h]_k K_{ij} - \nabla[h]_j K_{ik} = R_{\alpha\beta\sigma\rho} n^\alpha \xi_i^\beta \xi_j^\rho \xi_k^\sigma \tag{4}$$

whereas the curvature scalar is given by

$$R = R^{\alpha\beta}{}_{\alpha\beta} = 2R^{ni}{}_{ni} + R^{ij}{}_{ij} \tag{5}$$

In terms of this splitting, the Einstein-Hilbert action reads:

$$L_{EH} \equiv \sqrt{g} R[g] = N\sqrt{h}(R[h] + K_{ij}K^{ij} - K^2) - \partial_\alpha V^\alpha \tag{6}$$

with

$$V^\alpha = 2\sqrt{g}(n^\beta \nabla_\beta n^\alpha - n^\alpha \nabla_\beta n^\beta)$$

Primary constraints appear when defining canonical momenta:

$$p^\mu \equiv \frac{\partial L}{\partial \dot{N}_\mu} \sim 0$$

the momenta conjugate to the spatial part of the metric is:

$$\pi^{ij} \equiv \frac{\delta L}{\delta \dot{h}_{ij}} = -h^{1/2}(K^{ij} - K h^{ij})$$

The canonical commutation relations yield:

$$\{\pi^{ij}(\boldsymbol{x}), h_{kl}(\boldsymbol{x}')\} = -\delta(\boldsymbol{x}, \boldsymbol{x}')\frac{1}{2}(\delta^i_k \delta^j_l + \delta^j_k \delta^i_l)$$

The total Hamiltonian reads

$$H \equiv \int d^3x(\pi_\mu \dot{N}^\mu + \pi^{ij}\dot{h}_{ij} - L) = \int d^3x(N\mathcal{H} + N^i \mathcal{H}^i)$$

where

$$\mathcal{H}(h, \pi) = h^{-1/2}\left(\pi_{ij}\pi^{ij} - \frac{1}{2}\pi^2\right) - h^{1/2}R[h] \tag{7}$$

and

$$\mathcal{H}_i(h, \pi) = -2h_{ik}\partial_j \pi^{kj} - (2\partial_j h_{ki} - \partial_i h_{kj})\pi^{kj} = -2\nabla[h]_j \pi_i{}^j \tag{8}$$

The system of constraints is now consistent (that is, that the classical time evolution of the constraints is still a linear combination of constraints):

$$\dot{p}^\mu = \{p^\mu, H\} = (\mathcal{H}, \mathcal{H}_i) \sim 0$$

Second class constraints

$$N^\mu = f^\mu$$

can now be imposed. The whole hamiltonian analysis boils down to the two constraint equations

$$\mathcal{H} = 0$$
$$\mathcal{H}_i = 0$$

Much of the preceding analysis is actually quite generic for generally covariant systems. The full set of constraints obeys the Dirac-Schwinger algebra

$$\begin{aligned}
&\{\mathcal{H}(\boldsymbol{x}), \mathcal{H}(\boldsymbol{y})\} = [\mathcal{H}^i(\boldsymbol{x}) + \mathcal{H}^i(\boldsymbol{y})]\partial_i \delta(\boldsymbol{x}, \boldsymbol{y})\\
&\{\mathcal{H}_i(\boldsymbol{x}), \mathcal{H}(\boldsymbol{y})\} = \mathcal{H}(\boldsymbol{x})\partial_i \delta(\boldsymbol{x}, \boldsymbol{y})\\
&\{\mathcal{H}_i(\boldsymbol{x}), \mathcal{H}_j(\boldsymbol{y})\} = \mathcal{H}_i(\boldsymbol{y})\partial_j \delta(\boldsymbol{x}, \boldsymbol{y}) + \mathcal{H}_j(\boldsymbol{x})\partial_i \delta(\boldsymbol{x}, \boldsymbol{y})
\end{aligned} \tag{9}$$

which is nothing else than the Σ-projected algebra of the Diff(M) group.

Usually no reduction is made on the dynamical variables of the system, which amounts to keep h_{ij}, π^{ij} as (redundant) quantum variables. It is not clear how singular metrics can be avoided, because it is not easy to impose the condition that the metric is a positive definite operator.

Physical states in the Hilbert space are provisionally defined à la Dirac

$$\hat{\mathcal{H}}|\psi\rangle = 0$$
$$\hat{\mathcal{H}}_i|\psi\rangle = 0$$

It has been realized since long that this whole approach suffers from the *frozen time problem*, i.e., the Hamiltonian reads

$$H \equiv \int d^3x (N\mathcal{H} + N^i\mathcal{H}^i)$$

so that acting on physical states

$$\hat{H}|\psi\rangle = 0 \tag{10}$$

in such a way that Schrödinger's equation

$$i\frac{\partial}{\partial t}|\psi\rangle = \hat{H}|\psi\rangle \tag{11}$$

seemingly forbids any time dependence.

There are many unsolved problems in this approach, which has been kept at a formal level. The first one is an obvious operator ordering ambiguity owing to the nonlinearity. In the same vein, it is not clear whether it is possible to make the constraints hermitian. Besides, it is not clear that one recovers the full Diff invariance from the Dirac-Schwinger algebra. Actually, it is not known whether this is necessary; that is, what is the full symmetry of the quantum theory.

We can proceed further, still formally[2], using the Schrödinger representation defined in such a way that

$$(\hat{h}_{ij}\psi)[h] \equiv h_{ij}(x)\psi[h] \tag{12}$$

and

$$(\hat{\pi}^{ij}\psi)[h] \equiv -i\hbar\frac{\delta\psi}{\delta h_{ij}(x)}[h] \tag{13}$$

If we assume that diffeomorphisms act on wave functionals as:

$$\psi[f^*h] = \psi[h] \tag{14}$$

then the whole setup for the quantum dynamics of the gravitational field lies in Wheeler's *superspace* (nothing to do with supersymmetry) which is the set of three-dimensional metrics modulo three-dimensional diffs: $Riem(\Sigma)/Diff(\Sigma)$.

---

[2] It is bound to be formal as long as the problem of the infinities is not fully addressed. We know from the analysis of this representation for gauge theories in the lattice that those are the most difficult problems to solve.

The Hamiltonian constraint then implies the famous *Wheeler-De Witt equation.*

$$-\hbar^2 2\kappa^2 G_{ijkl}\frac{\delta^2\psi}{\delta h_{ik}\delta h_{jl}}[h] - \frac{\hbar}{2\kappa^2}R^{(3)}[h]\psi[h] = 0 \tag{15}$$

where the De Witt tensor is:

$$G_{ijkl} \equiv \frac{1}{\sqrt{h}}\left(h_{ij}h_{kl} - \frac{1}{2}h_{ik}h_{jl}\right) \tag{16}$$

Needless to say, this equation, suggestive as it is, is plagued with ambiguities. The manifold of positive definite metrics has been studied by De Witt. He showed that it has signature $(-1, +1^5)$, where the timelike coordinate is given by the breathing mode of the metric:

$$\zeta = \sqrt{\frac{32}{3}}h^{1/4} \tag{17}$$

and in terms of other five coordinates ζ^a orthogonal to the timelike coordinate, the full metric reads

$$ds^2 = -d\zeta^2 + \frac{3}{32}\zeta^2 g_{ab}d\zeta^a d\zeta^b \tag{18}$$

with

$$g_{ab} = tr\, h^{-1}\partial_a h h^{-1}\partial h \tag{19}$$

The five dimensional submanifold with metric g_{ab} is the coset space

$$SL(3,\mathbb{R})/SO(3) \tag{20}$$

It has been much speculated whether the timelike character of the dilatations lies at the root of the concept of time. The Wheeler-De Witt equation can be written in a form quite similar to the Klein-Gordon equation:

$$\left(-\frac{\partial^2}{\partial\zeta^2} + \frac{32}{3\zeta^2}g^{ab}\partial_a\partial_b + \frac{3}{32}\zeta^2 R^{(3)}\right)\Psi = 0 \tag{21}$$

The analogy goes further in the sense that also here there is a naturally defined scalar product which is not positive definite:

$$(\psi,\chi) \equiv \int_\Sigma \psi^* d\Sigma^{ij}G_{ijkl}\frac{\delta\chi}{i\delta h_{kl}} - \chi^* d\Sigma^{ij}G_{ijkl}\frac{\delta\psi}{i\delta h_{kl}} \tag{22}$$

(-& .- - -/ ' ! - ? ! & ' @ !
!$$! - ;) 5< $"
7' <% ;56 / ' ! 2 ! A) %)) 6 /@ ? + " BBB
(-& .- - -/ ' ! - ? ! & ' @ !
!$$! - ;) 5< $"

8

Orthogonality

In Section 5.3 we introduced the dot product in $\mathbb{R}^n$ and extended the basic geometric notions of length and distance. A set $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m\}$ of nonzero vectors in $\mathbb{R}^n$ was called an **orthogonal set** if $\mathbf{f}_i \cdot \mathbf{f}_j = 0$ for all $i \neq j$, and it was proved that every orthogonal set is independent. In particular, it was observed that the expansion of a vector as a linear combination of orthogonal basis vectors is easy to obtain because formulas exist for the coefficients. Hence the orthogonal bases are the "nice" bases, and much of this chapter is devoted to extending results about bases to orthogonal bases. This leads to some very powerful methods and theorems. Our first task is to show that every subspace of $\mathbb{R}^n$ *has* an orthogonal basis.

SECTION 8.1 Orthogonal Complements and Projections

If $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$ is linearly independent in a general vector space, and if $\mathbf{v}_{m+1}$ is not in $\text{span}\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$, then $\{\mathbf{v}_1, \dots, \mathbf{v}_m, \mathbf{v}_{m+1}\}$ is independent (Lemma 1 Section 6.4). Here is the analog for *orthogonal* sets in $\mathbb{R}^n$.

Lemma 1

Orthogonal Lemma

Let $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m\}$ be an orthogonal set in $\mathbb{R}^n$. Given $\mathbf{x}$ in $\mathbb{R}^n$, write

$$\mathbf{f}_{m+1} = \mathbf{x} - \frac{\mathbf{x}\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 - \frac{\mathbf{x}\cdot\mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 - \cdots - \frac{\mathbf{x}\cdot\mathbf{f}_m}{\|\mathbf{f}_m\|^2}\mathbf{f}_m$$

Then:

1. *$\mathbf{f}_{m+1} \cdot \mathbf{f}_k = 0$ for $k = 1, 2, \dots, m$.*
2. *If $\mathbf{x}$ is not in $\text{span}\{\mathbf{f}_1, \dots, \mathbf{f}_m\}$, then $\mathbf{f}_{m+1} \neq \mathbf{0}$ and $\{\mathbf{f}_1, \dots, \mathbf{f}_m, \mathbf{f}_{m+1}\}$ is an orthogonal set.*

PROOF

For convenience, write $t_i = (\mathbf{x}\cdot\mathbf{f}_i)/\|\mathbf{f}_i\|^2$ for each i. Given $1 \le k \le m$:

$$\begin{aligned}\mathbf{f}_{m+1}\cdot\mathbf{f}_k &= (\mathbf{x} - t_1\mathbf{f}_1 - \cdots - t_k\mathbf{f}_k - \cdots - t_m\mathbf{f}_m)\cdot\mathbf{f}_k \\ &= \mathbf{x}\cdot\mathbf{f}_k - t_1(\mathbf{f}_1\cdot\mathbf{f}_k) - \cdots - t_k(\mathbf{f}_k\cdot\mathbf{f}_k) - \cdots - t_m(\mathbf{f}_m\cdot\mathbf{f}_k) \\ &= \mathbf{x}\cdot\mathbf{f}_k - t_k\,\|\mathbf{f}_k\|^2 \\ &= 0\end{aligned}$$

This proves (1), and (2) follows because $\mathbf{f}_{m+1} \neq \mathbf{0}$ if $\mathbf{x}$ is not in $\text{span}\{\mathbf{f}_1, \dots, \mathbf{f}_m\}$.

)
(-& .- - -/ ' ! -) * + ($! ! 8 2
(+ ! " ! $ ' % ;6-* + ($! ! 8 2 (+ ! " !
$ ' % ;6 $" 7' % 6
(-& .- - -/ ' ! -) * + ($! !
8 2 (+ ! " ! $ ' % ;6-) * + ($! ! 8 2
(+ ! " ! $ ' % ;6 $"

1

Arrangements

Efi Fogel, Dan Halperin*, Lutz Kettner, Monique Teillaud, Ron Wein, and Nicola Wolpert

1.1 Introduction

Arrangements of geometric objects have been intensively studied in combinatorial and computational geometry for several decades. Given a finite collection S of geometric objects (such as lines, planes, or spheres) the *arrangement* $\mathcal{A}(S)$ is the *subdivision* of the space where these objects reside into cells as induced by the objects in S. Figure 1.1 illustrates a planar arrangement of circles, which consists of vertices, edges, and faces: a *vertex* is an intersection point of two (or more) circles, an *edge* is a maximal portion of a circle not containing any vertex, and a *face* is a maximal region of the plane not containing any vertex or edge. For convenience we also introduce two (artificial) vertices in each circle at the x-extreme points splitting the circle into two x-monotone arcs (thus each edge of the arrangement now has two distinct endpoints).

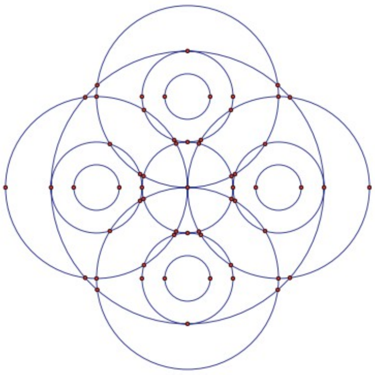

Fig. 1.1. An arrangement of 14 circles in the plane, with 53 faces (one of which is unbounded), 106 edges, and 59 vertices

Arrangements are defined and have been investigated for general families of geometric objects. One of the best studied type of arrangements is that of lines in the plane. This means that arrangements may have unbounded edges and faces. Furthermore, arrangements are defined in any dimension. There are, for example, naturally defined and useful arrangements of hypersurfaces in six-dimensional space, arising in the study of rigid motion of bodies in three-dimensional space.

---

* Chapter coordinator

Written evidence of the study of arrangements goes back to the nineteenth century (see [249]). We notice three periods in the *computational* study of arrangements. From the inception of computational geometry in the seventies till the mid eighties the focus was almost exclusively on the theoretical study of arrangements of unbounded linear objects, of hyperplanes. Many of the results obtained during this period are summarized in the book by Edelsbrunner [130]. The central role of arrangements in computational geometry has fortified in the following period, from the mid-eighties to the mid-nineties, where the theoretical focus has shifted toward arrangements of curves and surfaces. Many of the results obtained in those years are summarized in the book by Sharir and Agarwal [312] and in the survey papers [15] and [196]. From the mid nineties till the time of writing this chapter, a new more practical aspect of the study of arrangements has strengthened, emphasizing implementation and usage. The goal of robustly implementing algorithms for arrangements continues to raise numerous challenging technological and scientific problems. This chapter is devoted to the developments in this applied direction, with an emphasis on curves and surfaces. It should be noted however, that the appearance of a new topic of study in arrangements never replaced previous trends, and to this very day the theoretical study of arrangements of hyperplanes or of arrangements of algebraic curves is a thriving and fruitful domain.

Besides being interesting in their own right, arrangements are useful in a variety of applications. They have been used in solving problems in robot motion planning, computer vision, GIS (geographic information systems), computer-assisted surgery, statistics, and molecular biology, to mention just a few of the application domains. What makes arrangements such a useful structure is that they enable the accurate discretization of continuous problems, without compromising the exactness or completeness of the solution.

When coming to solve a problem using arrangements, we first need to cast the problem in “arrangement terms”. To this end, there is a large arsenal of techniques, most of which are rather simple. They include duality transforms (used to cast problems on point configurations as problems on arrangements of hyperplanes), Plücker coordinates [287], and the so-called locus method that analyzes criticalities in a problem and transforms the criticalities into hypersurfaces in the space where the problem is studied.

The next stage is to understand the combinatorial complexity of the relevant arrangement or of a portion thereof. Quite often, one does not need to construct the entire arrangement in order to solve a problem, and having only a substructure is sufficient (e.g., a single connected component of the ambient space often suffices in solving motion planning problems). This analysis gives a lower bound on the resources required by the algorithms and data structures that will be used in the solution.

Choosing or devising efficient algorithms and data structures to build the required (sub)arrangement is the next step of solving a problem using arrangements. This and the above two steps have been studied for decades.

Then comes the stage of effective implementation of the solution, which has various aspects to it. Sometimes asymptotically efficient algorithms are not necessarily practically the best. Precision and robustness of the solution is a central issue and raises questions such as: (i) how to cope with degeneracies, which are typically ignored in theory under the *general position* assumption; (ii) is the ready-made computer arithmetic sufficient or do we need to use more sophisticated machinery? These are the topics that this chapter focuses on.

The next section of the chapter gives a brief overview of the state-of-the-art in constructing arrangements. In Sect. 1.3 we survey the main advances in exact construction of planar arrangements, focusing mostly on the sweep-line approach, but explaining also what is needed for incremental construction. In Sect. 1.4 we review the software implementation details of these methods. The more recent work on the three-dimensional case is discussed in Sect. 1.5. Stepping away from exact computing is the topic of Sect. 1.6. A tour of implemented applications of curves and surfaces is given in Sect. 1.7. We conclude in Sect. 1.8 with suggestions for further reading and open problems.

1.2 Chronicles

In 1995, CGAL, the *Computational Geometry Algorithms Library*, was founded as a research project with the goal of making the large body of geometric algorithms developed in the field of computational geometry available for industrial applications with correct and efficient implementations [2, 222, 156]. It has since then evolved to an Open Source Project and is now representing the state-of-art in implementing computational geometry software in many areas. CGAL contains an elaborate and efficient implementation of arrangements that supports general types of curves.

The recent ECG project, which stands for *Effective Computational Geometry for Curves and Surfaces*, running from 2001 to 2004, extended the scope of implementation research towards curved objects.[1] Arrangements of curves and surfaces were an important theme in this project, and several different approaches to the topic were taken. This body of work is now collected and presented in a uniform manner in the following sections of the current chapter. In this brief section however, we present how the different results evolved.

The first branch of research, undertaken at Tel-Aviv University, Israel, is founded on the CGAL arrangement computation. Originally the CGAL arrangements package supported line segments, circular arcs and restricted types of parabolas. Wein [331] extended the CGAL implementation to ellipses and

---

[1] ⟨`http://www-sop.inria.fr/prisme/ECG/`⟩

7' <<% 6 7' 5 %)6
&$ = !' ! !

(-& .- - -/ ' ! -) * + ($! ! 8 2
(+ ! " ! $ ' % ;6-) * + ($! ! 8 2 (+ ! "
! $ ' % ;6 $"

2.4.3 Apollonius Diagrams

In this section, we present diagrams that are closely related to Möbius diagrams: namely, the Euclidean Voronoi diagrams of hyperspheres, also called Apollonius or Johnson-Mehl diagrams. Contrary to Möbius and anisotropic diagrams, the bisectors of Apollonius diagrams are not algebraic hypersurfaces since the bisector between two hyperspheres is only one sheet of a hyperboloid. As a consequence, Apollonius diagrams cannot be linearized in the same way as Möbius and anisotropic diagrams. Nevertheless, another linearization scheme can be applied, leading to interesting combinatorial and algorithmic results.

Definition of Apollonius Diagrams

Let us consider a finite set of weighted points $\mathcal{S} = \{\sigma_0, \sigma_1, \ldots, \sigma_n\}$ where $\sigma_i = (p_i, r_i)$, $p_i \in \mathbb{R}^d$ and $r_i \in \mathbb{R}$. We define the distance from x to σ_i as

$$\delta_i(x) = \|x - p_i\| - r_i.$$

This distance is also called the *additively weighted* distance from x to the weighted point σ_i. The minimization diagram of the distance functions δ_i, $i = 1, \ldots, n$, is called the additively weighted Voronoi diagram, or the Apollonius diagram of $\mathcal{S}$. We denote it by $\mathrm{Apo}(\mathcal{S})$ (see Fig. 2.8).

The Apollonius region $A(\sigma_i)$ of σ_i is defined as

$$A(\sigma_i) = \{x \in \mathbb{R}^d, \delta_i(x) \leq \delta_j(x)\}.$$

It is easy to see that $A(\sigma_i)$ is either empty or star-shaped from p_i. The boundary of $A(\sigma_i)$ may have a complicated structure. In fact, as we will see, the boundary of $A(\sigma_i)$ has the same combinatorial structure as a Möbius diagram in $\mathbb{R}^{d-1}$.

Since the diagram is not changed if we replace all r_i by $r_i + r$ for any $r \in \mathbb{R}$, we can assume, without loss of generality, that all r_i are non negative. The weighted points are then hyperspheres and the distance to a weighted point is the signed Euclidean distance to the corresponding hypersphere, counted positively outside the hypersphere and negatively inside the hypersphere.

Observe that, in the plane, a vertex of an Apollonius diagram is the center of a circle tangent to three circles of $\mathcal{S}$ (assuming all r_i non negative). Computing such a point is known as Apollonius' Tenth Problem, hence the name of the diagram.

Apollonius Diagrams and Power Diagrams

The graph of the distance function $\delta_i(x)$ is the half-cone of revolution $\mathcal{C}_i$ of equation

$$\mathcal{C}_i : x_{d+1} = \|x - p_i\| - r_i, \quad x_{d+1} + r_i \geq 0$$

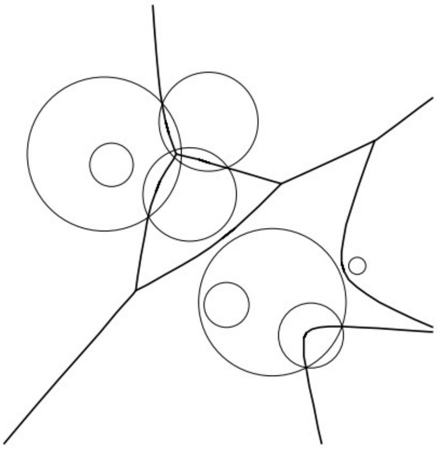

Fig. 2.8. The Apollonius diagram of a set of circles. Compare with the power diagram of the same set of circles in Fig. 2.6

The bisector of two hyperspheres of $\mathcal{S}$ is thus the projection of the intersection of two half-cones. This intersection is a quadratic hypersurface (in fact, a sheet of a two sheet hyperboloid) contained in a hyperplane. Indeed, we have

$$\begin{aligned} \mathcal{C}_1 &: (x_{d+1}+r_1)^2 = (x-p_1)^2, \;\; x_{d+1}+r_1 > 0, \\ \mathcal{C}_2 &: (x_{d+1}+r_2)^2 = (x-p_2)^2, \;\; x_{d+1}+r_2 > 0. \end{aligned}$$

The intersection of the two half-cones is contained in the hyperplane h_{12} whose equation is obtained by subtracting the two sides of the above equations:

$$h_{12} \quad : \quad -2(p_1-p_2)\cdot x - 2(r_1-r_2)x_{d+1} + p_1^2 - r_1^2 - p_2^2 + r_2^2 = 0.$$

This shows that there exists a correspondence between the diagram $\mathrm{Apo}(\mathcal{S})$ and the power diagram of the hyperspheres Σ_i in $\mathbb{R}^{d+1}$ $(i = 1, \ldots, n)$, where Σ_i is centered at (p_i, r_i) and has radius $r_i\sqrt{2}$. More precisely, $A(\sigma_i)$ is the projection of the intersection of the half-cone $\mathcal{C}_i$ with the power region $L(\Sigma_i)$. Indeed, x is in $A(\sigma_i)$ if and only if the projection X_i of x onto $\mathcal{C}_i$ has a smaller x_{d+1}-coordinate than the projections of x onto the other half-cones $\mathcal{C}_j$, $j \neq i$. In other words, the coordinates (x, x_{d+1}) of X_i must obey

$$\begin{aligned} &(x_{d+1}+r_i)^2 = (x-p_i)^2 \\ &(x_{d+1}+r_j)^2 \leq (x-p_j)^2 \quad \text{for any} \;\; j \neq i, \end{aligned}$$

and by subtracting both sides, it follows that $\Sigma_i(X_i) \leq \Sigma_j(X_i)$ for all j.

Algorithm 2 Construction of Apollonius diagrams

```
INPUT: a set of hyperspheres $\mathcal{S}$
1. Compute $\Sigma_i$, for $i = 1, \ldots, n$;
2. Compute the power diagram of the $\Sigma_i$'s;
3. For all $i = 1, \ldots, n$, project vertically the intersection of the power region $L(\Sigma_i)$
   with the half-cone $\mathcal{C}_i$.
OUTPUT: the Apollonius diagram of $\mathcal{S}$.
```

The Apollonius diagram of $\mathcal{S}$ can be computed using the following algorithm:

The power diagram of the Σ_i can be computed in time $O(n^{\lfloor \frac{d}{2} \rfloor+1} \log n)$. The intersection involved in Step 3 can be computed in time proportional to the number of faces of the power diagram of the Σ_i's, which is $O(n^{\lfloor \frac{d}{2} \rfloor+1})$. We have thus proved the following theorem due to Aurenhammer [35]:

Theorem 10. *The Apollonius diagram of a set of n hyperspheres in $\mathbb{R}^d$ has complexity $O(n^{\lfloor \frac{d}{2} \rfloor+1})$ and can be computed in time $O(n^{\lfloor \frac{d}{2} \rfloor+1} \log n)$.*

This result is optimal in odd dimensions, since the bounds above coincide with the corresponding bounds for the Voronoi diagram of points under the Euclidean distance. It is not optimal in dimension 2 (see Exercise 20). We also conjecture that it is not optimal in any even dimension.

Computing a Single Apollonius Region

We now establish a correspondence, due to Boissonnat and Karavelas [63], between a single Apollonius region and a Möbius diagram on a hypersphere.

To give the intuition behind the result, we consider first the case where one of the hyperspheres, say σ_0, is a hyperplane, i.e. a hypersphere of infinite radius. We take for σ_0 the hyperplane $x_d = 0$, and assume that all the other hyperspheres lie the half-space $x_d > 0$. The distance $\delta_0(x)$ from a point $x \in \mathbb{R}^d$ to σ_0 is defined as the Euclidean distance.

The points that are at equal distance from σ_0 and σ_i, $i > 0$, belong to a paraboloid of revolution with vertical axis. Consider such a paraboloid as the graph of a $(d-1)$-variate function ϑ_i defined over $\mathbb{R}^{d-1}$. If follows from Sect. 2.4.1 that the minimization diagram of the ϑ_i, $i = 1, \ldots, n$, is a Möbius diagram (see Fig. 2.9).

Easy computations give the associated weighted points. Write $p_i = (p'_i, p''_i)$, $p'_i \in \mathbb{R}^{d-1}$, $p''_i \in \mathbb{R}$, $i > 0$ and let $\Omega = \{\omega_1, \ldots, \omega_n\}$ be the set of Möbius sites of $\mathbb{R}^d$ where $\omega_i = \{p'_i, \lambda_i, \mu_i\}$, and

$$\lambda_i = \frac{1}{r_i + p''_i}, \qquad \mu_i = r_i - p''_i, \qquad i > 0.$$

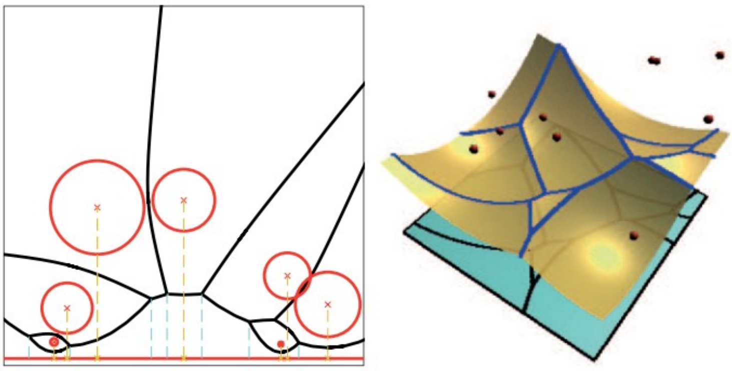

Fig. 2.9. A cell in an Apollonius diagram of hyperspheres projects vertically onto a Möbius diagram in σ_0

We let as an exercise to verify that the vertical projection of the boundary of the Apollonius region $A(\sigma_0)$ of σ_0 onto σ_0 is the Möbius diagram of ω.

We have assumed that one of the hyperspheres was a hyperplane. We now consider the case of hyperspheres of finite radii. The crucial observation is that the radial projection of $A(\sigma_0) \cap A(\sigma_i) \cap A(\sigma_j)$ onto σ_0, if not empty, is a hypersphere. It follows that the radial projection of the boundary of $A(\sigma_0)$ onto σ_0 is a Möbius diagram on σ_0.

Such a Möbius diagram on σ_0 can be computed by constructing the restriction of the power diagram of n hyperspheres of $\mathbb{R}^d$ with the hypersphere σ_0 (see Exercise 14).

Theorem 11. *Let $\mathcal{S}$ be a set of n hyperspheres in $\mathbb{R}^d$. The worst-case complexity of a single Apollonius region in the diagram of n hyperspheres of $\mathbb{R}^d$ is $\Theta(n^{\lfloor \frac{d+1}{2} \rfloor})$. Such a region can be computed in optimal time $\Theta(n \log n + n^{\lfloor \frac{d+1}{2} \rfloor})$.*

Exercise 17. Show that the cell of hypersphere σ_i in the Apollonius diagram of $\mathcal{S}$ is empty if and only if σ_i is inside another hypersphere σ_j.

Exercise 18. The predicates required to construct an Apollonius region are multivariate polynomials of degree at most 8 and 16 when $d = 2$ and 3 respectively. Detail these predicates [62].

Exercise 19. Show that the convex hull of a finite number of hyperspheres can be deduced from the restriction of a power diagram to a unit hypersphere [62].

Exercise 20. Prove that the combinatorial complexity of the Apollonius diagram of n circles in the plane has linear size.

Exercise 21 (Open problem). Give a tight bound on the combinatorial complexity of the Apollonius diagram of n hyperspheres of $\mathbb{R}^d$ when d is even.

2.5 Linearization

In this section, we introduce abstract diagrams, which are diagrams defined in terms of their bisectors. We impose suitable conditions on these bisectors so that any abstract diagram can be built as the minimization diagram of some distance functions, thus showing that the class of abstract diagrams is the same as the class of Voronoi diagrams.

Given a class of bisectors, such as affine or spherical bisectors, we then consider the inverse problem of determining a small class of distance functions that allows to build any diagram having such bisectors. We use a linearization technique to study this question.

2.5.1 Abstract Diagrams

Voronoi diagrams have been defined (see Sect. 2.2) as the minimization diagram of a finite set of continuous functions $\{\delta_1, \ldots, \delta_n\}$. It is convenient to interpret each δ_i as the distance function to an abstract object o_i, $i = 1, \ldots, n$. We define the bisector of two objects o_i and o_j of $\mathcal{O} = \{o_1, \ldots, o_n\}$ as

$$b_{ij} = \{x \in \mathbb{R}^d, \delta_i(x) = \delta_j(x)\}.$$

The bisector b_{ij} subdivides $\mathbb{R}^d$ into two open regions: one, b_{ij}^i, consisting of the points of $\mathbb{R}^d$ that are closer to o_i than to o_j, and the other one, b_{ij}^j, consisting of the points of $\mathbb{R}^d$ that are closer to o_j than to o_i. We can then define the Voronoi region of o_i as the intersection of the regions b_{ij}^i for all $j \neq i$. The union of the closures of these Voronoi regions covers $\mathbb{R}^d$. Furthermore, if we assume that the bisectors are $(d-1)$-manifolds, the Voronoi regions then have disjoint interiors and we can define the closed region associated to b_{ij}^i as $\bar{b}_{ij}^i = b_{ij}^i \cup b_{ij}$.

In a way similar to Klein [230], we now define diagrams in terms of bisectors instead of distance functions. Let $B = \{b_{ij}, i \neq j\}$ be a set of closed $(d-1)$-manifolds without boundary. We always assume in the following that $b_{ij} = b_{ji}$ for all $i \neq j$. We assume further that, for all distinct i, j, k, the following incidence condition (I.C.) holds:

$$b_{ij} \cap b_{jk} = b_{jk} \cap b_{ki} \quad (I.C.)$$

This incidence condition is obviously needed for B to be the set of bisectors of some distance functions.

By Jordan's theorem, each element of B subdivides $\mathbb{R}^d$ into at least two connected components and crossing a bisector b_{ij} implies moving into another

7') % <6 7' % ;6 ; C $ '2 " $ (+

7' % <6 7' ;5% 5 6) = ! 8 2 = !

(-& .- - -/ ' ! -) * + ($! ! 8 2
(+ ! " ! $ ' % ;6-) * + ($! ! 8 2 (+ ! "
! $ ' % ;6 $"

(-D - 5 ; - " . ! ! -C ! >! 0 . ! ! E
& " ((% <6 $"

(-& .- - -/ ' ! - C ! >! 0 . ! !
& " ((% <6-C ! >! 0 . ! ! & " ((
% <6 $"

7' % 6 7' ;% <6 C F " ' 1 ! '2 *

(-& .- - -/ ' ! - C ! >! 0 . ! !
& " ((% <6- C ! >! 0 . ! ! & "
((% <6 $"

cities a few days later, then once a few days have passed we know that one of those two predictions was wrong. If we thought of the probability distribution as a physical thing, it would seem insane to suddenly change it in Galveston, because we see a storm hitting New Orleans. Bayes' rule of conditional probability tells us this is exactly the right thing to do.

Bayes' rule is not in general applicable in QM, because probabilities only add when interference terms are negligible. If we want to understand *why* the interference terms suddenly become negligible when we put a detector near one of the slits, then we have to understand more about the properties of the detector than Feynman's discussion supplies. We will see in Chapter 10 that the crucial property is that the detector is a complicated system, made of many atoms, with a huge number of microscopic atomic states corresponding to each position of the "needle on the detector's dial." As a consequence, the interference terms in the probability distribution for seeing the detector's dial either register a hit or not, are unimaginably small, and Bayes' rule is approximately valid. This is a general lesson: *the quantum rules for the probabilities that refer to macroscopic objects differ from the rules of a classical statistical theory, which satisfies Bayes' rule, by amounts which are of order* e^{-cN} *where* N *is the number of atoms out of which the object is composed.* These differences are so small that they are impossible to measure, even in principle.

The fact that the rule for computing probabilities of future events depends on whether the detector is there or not is not mysterious. Indeed, we are describing very different experiments in the two cases. In the first, we ask for a prediction of the probability of getting a hit on the screen. In the second, we are asking for probabilities of hits, *conditioned on the detector's behaving a certain way at an intermediate time.* As we will see, the fact that the violation of Bayes' rule for observations done on the macroscopic detector is vanishingly small, completely accounts for the difference in the predictions.

4.6 A WORKED EXERCISE

Exercise 4.6 As we will see in the next chapter, the generalization of the Schrödinger equation to particles moving under the influence of a force is simply to add the potential energy $V(x)$ to the Hamiltonian operator. In the following exercises, we will consider potentials, which are locally constant, so that we can use free particle solutions to solve the equation. To begin, solve the Schrödinger equation for a step function potential $V(x) = V_0\theta(x+a)\theta(a-x)$. The real line divides into three intervals, in each of which the equation reduces to that of a free particle. By taking the integral of the Schrödinger equation in small intervals around the points of discontinuity, you can figure out how to join the solutions in the three regions and impose the condition of normalizability. For one sign of V_0, you should find only delta function normalizable eigenstates, whereas for the other, you will find both delta function normalizable and truly normalizable eigenfunctions. Show that the spectrum of the latter is discrete and describe how it depends on V_0 and a. The normalizable wave functions are called bound states. Show that the bound state spectrum depends only on the parameter $s_0 \equiv \frac{a}{\hbar}\sqrt{2mV_0}$. By examining the equation for bound states graphically, show that the number of such states is finite, but goes to infinity as $s_0 \to \infty$. For small s_0, show that there is only a single bound state. Now find the eigenfunctions for $E > 0$. There are two linearly independent solutions, which can be characterized by saying that the associated time-dependent solution moves toward or away from the center, near $+\infty$. Alternatively, we can characterize the solutions as incoming or outgoing at $-\infty$. There must be two linear relations between these four solutions, since only two of them can be linearly independent. Compute the 2×2 *Scattering matrix* or S-matrix, which expresses the two outgoing solutions in terms of the two incoming solutions.

Exercise 4.7 Consider the square well potential of Exercise 4.6. For the case of the normalizable wave functions, show that they decrease exponentially for large $|x|$. Think about classical mechanics with the same potential, and show that motion in the region where the quantum wave function is exponentially decreasing is simply not allowed. The fact that there is nonzero probability for the particle to be found in this region is called *quantum tunneling through a barrier*. We will study it more extensively in Chapter 17 on the JWKB approximation. Show that the equation for bound states depends only on the parameter s_0.

Exercise 4.8 For the continuum (i.e., scattering) wave functions in Exercise 4.6, we have two possible behaviors $e^{\pm ikx}$ for positive k in each of the asymptotic regions $x \to \pm\infty$. By considering the time-dependent Schrödinger equation show that the choices of the sign of k correspond to left and right moving waves in the two regions. Argue that the two different signs in a given asymptotic region, correspond to two linearly independent solutions of the second order ordinary differential equation for fixed k^2. Argue that there are exactly two linearly independent solutions. Thus, there must be linear relations between the four possible asymptotic behaviors. Using your exact solutions, find these linear relations. Your tool for finding these relations is continuity of the wave function and its derivative at the points where the potential, and thus the *second* derivative of ψ is discontinuous.

Answer to Exercises 4.6–4.8: The solution to the Schrödinger equation with energy E is

$$\psi_E = e^{\pm\sqrt{\frac{2m}{\hbar^2}(V(x)-E)}x}, \tag{4.34}$$

where $V(x) = V_0\theta(a-x)\theta(x+a)$. Integrating the Schrödinger equation near the discontinuities in $V(x)$, we find that the wave function and its first derivative must be continuous there Figure 4.2. If $E > 0$, we have delta function normalizable solutions, while if $E < 0$, we can get normalizable bound state solutions by picking only the falling exponential as $x \to \pm\infty$. The problem is invariant under reflections, so for the bound states, we can choose to look at even and odd solutions. We can impose the continuity conditions only at $x = a$ and they

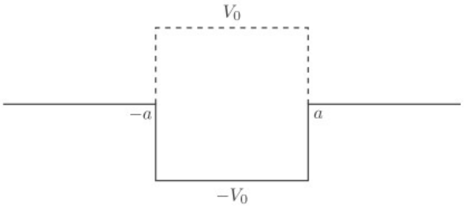

Figure 4.2 The potential for Exercises 4.6–4.8.

will automatically be satisfied for negative x. In the region near the origin, the wave function depends on the crucial quantity

$$ik_0 = \frac{\sqrt{2m(V_0 - E)}}{\hbar}, \tag{4.35}$$

which gives k_0 imaginary. The imaginary part of ak_0 is equal to the classical action in units of $\hbar$ of a particle traveling from $-a$ to a, if $E - V_0 > 0$. When the opposite inequality holds, k_0 is imaginary. In either case, we define k_0 with the positive (imaginary) square root. We define

$$ik = \frac{\sqrt{-2mE}}{\hbar}. \tag{4.36}$$

In addition, define $r \equiv \frac{k_0}{k}$. In the region $|x| > a$, the wave function grows or falls exponentially if $E < 0$. We must choose the falling solution at both positive and negative values of x, to obtain a normalizable wave function. The continuity conditions for even and odd wave functions at $x = a$ are:

$$A\cosh(|k_0|a) = Be^{-|k|a}, \quad Ar\sinh(|k_0|a) = -Be^{-|k|a}, \tag{4.37}$$

$$A\sinh(|k_0|a) = Be^{-|k|a}, \quad Ar\cosh(|k_0|a) = -Be^{-|k|a}, \tag{4.38}$$

when k_0 is imaginary. The hyperbolic sine and cosine are both positive, and these equations have no solution. Thus, we must have the intuitively obvious condition $E - V_0 > 0$, which implies a potential well rather than a barrier and asymptotic energy less than the depth of the well, in order to have a bound state. For k_0 real, the matching conditions become

$$A\cos(|k_0|a) = Be^{-|k|a}, \quad Ar\sin(|k_0|a) = Be^{-|k|a}, \tag{4.39}$$

$$A\sin(|k_0|a) = Be^{-|k|a}, \quad Ar\cos(|k_0|a) = -Be^{-|k|a}. \tag{4.40}$$

Motivated by the bound on E, we define $E = yV_0$, with $0 \leq y \leq 1$. The conditions become

$$\sqrt{\frac{y}{1-y}} = -\tan(s_0\sqrt{1-y}). \tag{4.41}$$

$$\sqrt{\frac{y}{1-y}} = -\cot(s_0\sqrt{1-y}). \tag{4.42}$$

These depend only on a single parameter $s_0 \equiv a\sqrt{\frac{-2mV_0}{\hbar^2}}$. When s_0 is small, we can expand the trigonometric functions. The equation for odd solutions becomes $y = 1/s_0^2$, which is inconsistent with $0 \leq y \leq 1$. The even solution is $y = s_0^2$ and is consistent. There is a single bound state. When s_0 is large, we can satisfy the equation with $s_0\sqrt{1-y} \sim \frac{n\pi}{2}$, where n is odd for even solutions and vice versa, as long as $\frac{n\pi}{2s_0} < 1$. The point is that the tangent and cotangent take on any large value in the vicinity of their poles, so that as long as y is sufficiently close to 1 we can match. These explicit formulae are only valid for large n, but there are bound states for every value of n satisfying the inequality. As s_0 goes to infinity, we get an infinite number of states and the explicit formula becomes more exact. In this limit, after adding a constant to make all the energies positive, the spectrum approaches that of the infinite square well. Note, however, that s_0 can be large for a shallow, but very broad, well also. In that situation, we would still have continuum eigenstates.

Turning now to $E > 0$, we define

$$\psi^+ = A^+_{out}e^{ikx} + A^+_{in}e^{-ikx}, \quad x > a, \tag{4.43}$$

$$\psi^+ = A^-_{out}e^{-ikx} + A^-_{in}e^{ikx}, \quad x < -a. \tag{4.44}$$

The subscripts *in* and *out* refer to the fact that, when multiplied by $e^{-i\frac{\hbar^2k^2}{2m}t}$, the relevant part of the solution becomes a traveling incoming or outgoing wave. Note that the sign of the exponential changes when we compare an incoming (outgoing) wave on the left and right. To write the continuity conditions at $x = \pm a$ compactly, it is convenient to define $z_0 = e^{ik_0a}$, $z = e^{ika}$, and $r^\pm = 1 \pm r$. Then we have

$$A^-_{in}z^{-1} + A^-_{out}z = Az_0^{-1} + Bz_0, \tag{4.45}$$

$$A^-_{in}z^{-1} - A^-_{out}z = r(Az_0^{-1} - Bz_0), \tag{4.46}$$

$$A^+_{out}z + A^+_{in}z^{-1} = Az_0 + Bz_0^{-1}, \tag{4.47}$$

$$A^+_{out}z - A^+_{in}z^{-1} = r(Az_0^{-1} - Bz_0). \tag{4.48}$$

These are equivalent to a pair of 2×2 matrix equations expressing the vectors

$$A_{in/out} \equiv \begin{pmatrix} A^-_{in/out} \\ A^+_{in/out} \end{pmatrix}$$

in terms of the vector

$$A_m \equiv \begin{pmatrix} A \\ B \end{pmatrix}$$

$$A_{in} = \frac{z}{2}[\frac{r_+}{z_0} + r_- z_0 \sigma_1] A_m, \tag{4.49}$$

$$A_{out} = \frac{1}{2z}[\frac{r_-}{z_0} + r_+ z_0 \sigma_1] A_m. \tag{4.50}$$

The *Scattering matrix* or S-matrix is the transformation that takes A_{in} into A_{out}, so we have

$$S = \frac{1}{2z}[\frac{r_-}{z_0} + r_+ z_0 \sigma_1] \frac{2}{z}[\frac{r_+}{z_0} + r_- z_0 \sigma_1]^{-1}, \tag{4.51}$$

These equations reflect the fact that there are only two linearly independent solutions of the Schrödinger equation. The *in* and *out* states represent two independent bases for the Hilbert space. These equations tell us how to transform between them.

Note that the two factors in the formula for the S-matrix commute with each other. Thus, since $S = S_+ S_-^{-1}$, we have $S^\dagger S = S_+^\dagger S_+ [S_-^\dagger S_-]^{-1}$. Now compute

$$S_+^\dagger S_+ = \frac{1}{4}[r_- z_0 + r_+ z_0^{-1} \sigma_1][\frac{r_-}{z_0} + r_+ z_0 \sigma_1] = \frac{1}{4}[r_+^2 + r_-^2 + r_+ r_- (z_0^2 + z_0^{-2})], \tag{4.52}$$

$$S_- S_-^\dagger = \frac{1}{4}[\frac{r_+}{z_0} + r_- z_0 \sigma_1][\frac{r_+}{z_0} + r_- z_0^{-1} \sigma_1] = \frac{1}{4}[r_+^2 + r_-^2 + r_+ r_- (z_0^2 + z_0^{-2})]. \tag{4.53}$$

Therefore, $S^\dagger S = 1$. The S-matrix is a unitary transformation between two bases of the Hilbert space. Of course, the states of fixed k are not normalizable states in the Hilbert space, but this calculation guarantees that the mapping of normalizable incoming wave packet states into normalizable outgoing wave packet states, defined as superpositions of the fixed k states, is a genuine unitary operator in the Hilbert space.

We can write the eigenvalues of the unitary S-matrix as $e^{2i\delta_\pm(k)}$ where $\delta_\pm$ are called the phase shifts.

$$e^{2i\delta_\pm} = z^2 \frac{(r_- z_0^{-1} \pm r_+ z_0)}{(r_- z_0^{-1} \pm r_- z_0)}. \tag{4.54}$$

They are analytic functions of k, which will have poles whenever the denominator has a linear zero. The equation for the poles is

$$\frac{r_+}{r_-} = -z_0^2. \tag{4.55}$$

For real k, there are no solutions of this equation because the left-hand side is a real number greater than 1 and the right-hand side is a phase. For imaginary k, the left-hand side is also a phase. If $k = il$ the equation for a pole is

$$r_- = \pm z_0^2 r_+. \tag{4.56}$$

Some simple algebra shows that this is identical to the equation for bound states, and has solutions only when $V_0 < 0$. The fact that the S-matrix is an analytic function of energy, with poles at the bound state energies, is a very general result of immense importance.

4.7 FURTHER EXERCISES

4.4 Consider the Fourier transform of a smooth function $f(x)$ (see Appendix B for a definition of smooth).

$$\tilde{f}(k) = \int dx\ f(x)e^{ikx},$$

for large values of k. Show that smoothness implies that the Fourier transform falls off faster than any power of k. Compute the Fourier transform of $(x^2 + a^2)^{-b}$ for those values of b for which the function is square integrable. How does the falloff in wave number space depend on b?

4.5 Compute the Fourier transform of a step function, $\theta(a - x)\theta(a + x)$, where $\theta(x)$ is 1 for $x > 0$ and 0 for negative x, and demonstrate the relation between the width of the interval over which the function is nonzero and the falloff in wave number space. Use this to calculate the time evolution of a wave function that is initially confined in a fixed interval.

4.9 Consider the plane wave normalizable solutions for the potential of Exercises 4.6–4.8, in the case where there is no left moving or incoming wave in the region of large positive x. Thus, $\psi = A^+_{out}e^{ikx}$ in this region. For $x < -a$ we have

$$\psi = A^-_{in}e^{ikx} + A^-_{out}e^{-ikx}.$$

A^-_{in} is called the incident amplitude. It is a wave moving toward the barrier from the left. A^-_{out} is called the reflected amplitude. A^+_{out} is the transmitted amplitude. The *transmission coefficient*,

$$T = |\frac{A^+_{out}}{A^-_{in}}|^2,$$

gives us the probability of transmission through the barrier. Calculate it, for all values of V_0, E, and a. Also *define* and calculate the *reflection coefficient* R and show that $R+T = 1$. Show that T oscillates as a function of E and that there are values of E for which it is equal to 1. That is, there are discrete energies for which the barrier becomes completely transparent! This is called the Ramsauer–Townsend effect, and has been observed in the laboratory.

4.10 Consider the potential $V = -a^2\theta(x)$. Consider a particle of mass m and energy E traveling to the right from large negative x. Assume that for $x > 0$ there is no left

(-& .- - -/ ' ! - * + ($! ! 8 2
(+ ! " ! $ ' % ;6-* + ($! ! 8 2 (+ ! " !
$ ' % ;6 $"
7' % 6 7' %)6 ; " ! &G ! " ? * + $

axis. Hence the problem of approximating the medial axis of Ω boils down to sampling the boundary of Ω, a problem that is closely related to mesh generation (see Chap. 5). Other informations on the medial axis can be found in Chap. 6.

2.7.1 Medial Axis and Lower Envelope

The *medial axis* of an open set Ω, denoted by $\mathcal{M}(\Omega)$, is defined as the the set of points of Ω that have more than one nearest neighbor on the boundary of Ω. Nearest refers in this section to the Euclidean distance although the results may be extended to other distance functions. A *medial sphere* σ is a sphere centered at a point c of the medial axis and passing through the nearest neighbors of c on $\partial\Omega$. Those points where σ is tangent (in the sense that σ does not enclose any point of $\partial\Omega$) to $\partial\Omega$ are called the *contact points* of σ.

The concept of medial axis can be considered as an extension of the notion of Voronoi diagram to infinite sets. Let o be a point of the boundary of Ω and δ_o be the distance function to o defined over Ω

$$\forall x \in \Omega : \ \delta_o(x) = \|x - o\|.$$

The *lower envelope* of the infinite set of functions δ_o is defined as

$$\Delta^- = \inf_{o \in \partial\Omega} \delta_o.$$

Following what we did for Voronoi diagrams (Sect. 2.2), we define, for any point $x \in \Omega$, its index set $I(x)$ as the set of all o such that $\Delta^-(x) = \delta_o(x)$. The set of points x such that $|I(x)| > 1$ constitutes the medial axis of Ω.

Computing the medial axis is difficult in general. If Ω is defined as a semi-algebraic set, i.e. a finite collection of algebraic equations and inequalities, $\mathcal{M}(\Omega)$ is also a semi-algebraic set that can therefore be computed using techniques from real algebraic geometry [32, 44]. This general approach, however, leads to algorithms of very high complexity. Theorem 1 can also be used but, still, working out the algebraic issues is a formidable task. Effective implementations are currently limited to simple objects. If Ω is a planar domain bounded by line segments and circular arcs, one can apply the results of Sect. 2.6. Further results can be found in [159].

An alternative and more practical approach consists in departing from the requirement to compute the medial axis exactly. In Sect. 2.7.2, we describe a method that approximates the medial axis of an object by first sampling its boundary, and then computing and pruning the Voronoi diagram of the sample.

2.7.2 Approximation of the Medial Axis

Approximating the medial axis of a set is a non trivial issue since sets that are close for the Hausdorff distance may have very different medial axes. This is illustrated in Fig. 2.13. Let S be a closed curve, $\Omega = \mathbb{R}^2 \setminus S$ and $\mathcal{P}$ be a finite set of points approximating S. As can be seen on the figure, the skeleton of the Voronoi diagram of $\mathcal{P}$ is far from the medial axis of Ω since there are many long branches with no counterpart in $\mathcal{M}(\Omega)$. These branches are Voronoi edges whose dual Delaunay edges are small (their lengths tend to 0 when the sampling density increases). In other words, the medial axis is not continuous under the Hausdorff distance. Notice however that if we remove the long branches in Fig. 2.13, we obtain a good approximation of the medial axis of S. This observation will be made precise in Lemma 10 below. It leads to an approximate algorithm that first sample S and extract from the Euclidean Voronoi diagram of the sample a sub-complex that approximates the medial axis of S.

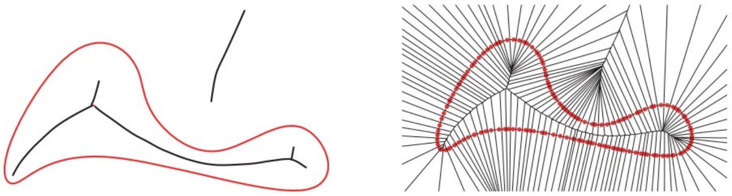

Fig. 2.13. On the left side, a closed curve S and the medial axis of $\Omega = \mathbb{R}^2 \setminus S$. On the right side, a dense sample E of points and its Voronoi diagram. The medial axis of $\mathbb{R}^2 \setminus E$ (which is the 1-skeleton of $\mathrm{Vor}(E)$) is very different from the medial axis of Ω

Given an open set Ω and a point x on the medial axis of Ω, we define $D(x)$ as the diameter of the smallest closed ball containing the contact points of the medial sphere centered at x. We define the λ-medial axis of Ω, denoted $\mathcal{M}_\lambda(\Omega)$ as the subset of the medial axis of Ω consisting of points x such that $D(x) \geq \lambda$.

Let Ω and Ω' be two open sets and let S and S' denote their boundaries. We assume that S and S' are compact and that their Hausdorff distance $d_H(S, S')$ is at most ε: any point of S is at distance at most ε from a point of S' and vice versa. Notice that we do not specify S nor S' to be finite or infinite point sets. For convenience, we rename medial axis of S (resp. S') and write $\mathcal{M}(S)$ (resp. $\mathcal{M}(S')$) the medial axis of $\mathbb{R}^3 \setminus S$. Similarly, we rename medial axis of S' and write $\mathcal{M}(S')$) the medial axis of $\mathbb{R}^3 \setminus S$.

The following lemma says that the λ-medial axis of S is close to the medial axis of S', provided that λ is sufficiently large. It should be emphasized that close here refers to the one-sided Hausdorff distance: the medial axis of S' is not necessarily close to the λ-medial axis of S although, by exchanging the roles of S and S', the lemma states that the λ'-medial axis of S' is close to the medial axis of S for a sufficiently large λ'. We will go back to this point later.

We say that a ball is S-empty if its interior does not intersect S. The sphere bounding a S-empty ball is called a S-empty sphere.

Lemma 10. *Let σ be a S-empty sphere centered at c, of radius r, intersecting S in two points x and y. If $\varepsilon < \frac{r}{2}$ and $l \overset{\text{def}}{=} \frac{\|x-y\|}{4} \geq \sqrt{\varepsilon r(1-\frac{\varepsilon}{r})}$, there exists an S'-empty sphere tangent to S' in two points whose center c' and radius r' are such that $|1-\frac{r'}{r}|$ and $\frac{\|c'-c\|}{r}$ are at most $\delta = \frac{\varepsilon r}{l^2-\varepsilon r+\varepsilon^2}$.*

Proof. Let σ'' be the maximal S'-empty sphere centered at c and let r'' be its radius (see Fig. 2.14). The Hausdorff distance between S and S' being at most ε, we have $|r-r''| \leq \varepsilon$. Let y' be a point of $\sigma'' \cap S'$.

Let σ' be the maximal S'-empty sphere tangent to σ'' at y'. σ' is tangent to S' at at least two points. Let c' be its center and r' its radius.

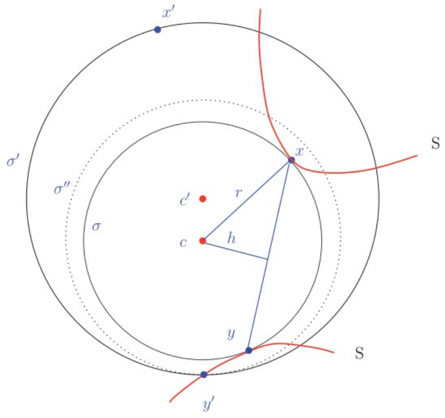

Fig. 2.14. S is the continuous curve. x' and y' belong to S'

Noting $h = \sqrt{r^2 - \frac{1}{4}\|x-y\|^2} = \sqrt{r^2-4l^2}$ the distance from c to line xy, we have

$$\begin{aligned} d(c', S) &\leq \min(\|c'-x\|, \|c'-y\|) \\ &\leq \sqrt{(\|c-c'\|+h)^2 + \frac{1}{4}\|x-y\|^2} \\ &= \sqrt{r^2 + \|c-c'\|^2 + 2h\|c-c'\|}. \end{aligned}$$

On the other hand,

$$d(c', S') = \|c' - y'\| = \|c - c'\| + r'' \geq \|c - c'\| + r - \varepsilon.$$

From these two inequalities, we deduce

$$\|c - c'\| + r - \varepsilon \leq d(c', S') \leq d(c', S) + \varepsilon \leq \sqrt{r^2 + \|c - c'\|^2 + 2h\|c - c'\|} + \varepsilon.$$

and, since, by assumption, $r > 2\varepsilon$, we get $\|c - c'\|\,(r - 2\varepsilon - h) \leq 2\varepsilon(r - \varepsilon)$. Moreover, by assumption, $l \geq \sqrt{\varepsilon r(1 - \frac{\varepsilon}{r})}$ which implies $r - 2\varepsilon - h \geq 0$. Indeed,

$$h = \sqrt{r^2 - 4l^2} \leq \sqrt{r^2 - 4\varepsilon r + 4\varepsilon^2} = r - 2\varepsilon.$$

We then deduce

$$\|c - c'\| \leq \frac{2\varepsilon(r - \varepsilon)}{r - h - 2\varepsilon} \leq \frac{2\varepsilon r}{r - \sqrt{r^2 - 4l^2} - 2\varepsilon}.$$

We then get

$$\frac{\|c - c'\|}{r} \leq \frac{2\varepsilon(r + \sqrt{r^2 - 4l^2} - 2\varepsilon)}{(r - 2\varepsilon)^2 - (r^2 - 4l^2)} \leq \frac{\varepsilon r}{l^2 - \varepsilon r + \varepsilon^2}.$$

The same bound plainly holds for $|1 - \frac{r'}{r}|$.

We consider the case where S is a surface of $\mathbb{R}^3$ and S' is a finite set of points on S at Hausdorff distance at most ε from S. To avoid confusion, we rename S' as $\mathcal{P}$. As already noticed, $\mathcal{M}(\mathcal{P})$ is the 1-skeleton of the Voronoi diagram $\mathrm{Vor}(\mathcal{P})$, called simply the Voronoi diagram of $\mathcal{P}$ in this section. The set of contact points of a medial sphere centered at a point $c \in \mathcal{M}(\mathcal{P})$ is the set of closest points of c in $\mathcal{P}$. Any point in the relative interior of a face of $\mathrm{Vor}(\mathcal{P})$ has the same closest points in $\mathcal{P}$. It follows that the λ-medial axis of $\mathcal{P}$ is the subset of the faces of $\mathrm{Vor}(\mathcal{P})$ whose contact points cannot be enclosed in a ball of diameter λ.

Lemma 10 then says that any Delaunay sphere of $\mathrm{Del}(\mathcal{P})$ passing through two sample points that are sufficiently far apart, is close to a medial sphere of S (for the Hausdorff distance). We have therefore bounded the *one-sided* Hausdorff distance from the λ-medial axis (Voronoi diagram) of an ε-sample $\mathcal{P}$ of S to the medial axis of S, when λ is sufficiently large with respect to ε. If we apply the lemma the other way around, we see that, for sufficiently large λ', $\mathcal{M}_{\lambda'}(S)$ is close to $\mathcal{M}(\mathcal{P})$. However, as observed above (Fig. 2.13), we cannot hope to bound the two-sided Hausdorff distance between $\mathcal{M}(S)$ and $\mathcal{M}(\mathcal{P})$.

The above lemma can be strengthened as recently shown by Chazal and Lieutier [83, 32]. They proved that the λ-medial axis of S is close to the λ'-medial axis of $\mathcal{P}$ for a sufficiently large λ and some positive λ' that depends on λ and ε. More precisely, let D be the diameter of S and k'' a positive constant. They showed that there exist three functions of ε, $k(\varepsilon) = 15\sqrt{2}\,\sqrt[4]{D^3\varepsilon}$, $k'(\varepsilon) = 10\sqrt{3}\,\sqrt[4]{D^3\varepsilon}$ and $k''(\varepsilon) = k''\,\sqrt[4]{D^3\varepsilon}$, such that

$$\mathcal{M}_{k(\varepsilon)}(S) \subset \mathcal{M}_{k'(\varepsilon)}(\mathcal{P}) \oplus B_{2\sqrt{D\varepsilon}} \subset \mathcal{M}_{k''(\varepsilon)}(S) \oplus B_{4\sqrt{D\varepsilon}}.$$

Here, B_r denotes the ball centered at the origin of radius r, and $\oplus$ the Minkowski sum.

Consider now a family of point sets $\mathcal{P}_\varepsilon$ parametrized by ε such that $d_H(S, \mathcal{P}_\varepsilon) \leq \varepsilon$ and let ε tends to 0. Because $\mathcal{M}_\eta(S)$ tends to $\mathcal{M}(S)$ when η tends to 0, we deduce from the above inequalities, that

$$\lim_{\varepsilon \to 0} d_H(\mathcal{M}(S), \mathcal{M}_{10\sqrt[4]{9D^3\varepsilon}}(\mathcal{P}_\varepsilon)) = 0.$$

The above discussion provides an algorithm to approximate the medial axis of S within any specified error (see Algorithm 5).

Algorithm 5 Approximation of the Medial Axis

```
INPUT: A surface $S$ and a positive real $\varepsilon$
1. Sample $S$ so as to obtain a sample $\mathcal{P}$ such that $d_H(S, \mathcal{P}) \leq \varepsilon$;
2. Construct the Voronoi diagram of $\mathcal{P}$;
3. Remove from the diagram the faces for which the diameter of the set of contact
   points is smaller than $10\sqrt[4]{9D^3\varepsilon}$.
OUTPUT: A PL approximation of $\mathcal{M}(S)$
```

The main issue is therefore to compute a sample of points on S (step 1). If S is a surface of $\mathbb{R}^3$, one can use a surface mesh generator to mesh S and take for $\mathcal{P}$ the vertices of the mesh. Various algorithms can be found in Chap. 5 and we refer to that chapter for a thorough description and analysis of these algorithms. Especially attractive in the context of medial axis approximation, are the algorithms that are based on the 3-dimensional Delaunay triangulation since we get the Voronoi diagram of the sample points (step 2) at no additional cost. An example obtained with the surface mesh generator of Boissonnat and Oudot [65] is shown in Fig. 2.15.

Exercise 31. Let O be a bounded open set. Show that $\mathcal{M}(O)$ is a retract of O (and therefore has the same homotopy type as O) [243].

2.8 Voronoi Diagrams in CGAL

The Computational Geometry Algorithms Library CGAL [2] offers severals packages to compute Voronoi diagrams. Euclidean Voronoi diagrams of points and power diagrams are represented through their dual Delaunay and regular triangulations. CGAL provides Delaunay and regular triangulations in $\mathbb{R}^2$ and $\mathbb{R}^3$. The implementation is based on a randomized incremental algorithm

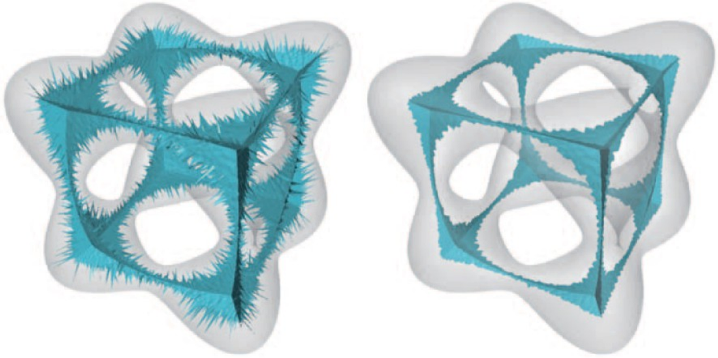

Fig. 2.15. Two λ-medial axes of the same shape, with λ increasing from left to right, computed as a subset of the Voronoi diagram of a sample of the boundary (courtesy of Steve Oudot)

using a variant of the Voronoi hierarchy described in Sect. 2.6. Delaunay triangulations are also provided in higher dimensions.

The library also contains packages to compute Voronoi diagrams of line segments [215] and Apollonius diagrams in $\mathbb{R}^2$ [216]. Those packages implement the incremental algorithm described in Sect. 2.6. A prototype implementation of Möbius diagrams in $\mathbb{R}^2$ also exists. This prototype computes the Möbius diagram as the projection of the intersection of a 3-dimensional power diagram with a paraboloid, as described in Sect. 2.4.1. This prototype also serves as the basis for the developement of a CGAL package for 3-dimensional Apollonius diagrams, where the boundary of each cell is computed as a 2-dimensional Möbius diagram, following the results of Sect. 2.4.3 [62]. See Fig. 2.8.

2.9 Applications

Euclidean and affine Voronoi diagrams have numerous applications we do not discuss here. The interested reader can consult other chapters of the book, most notably Chap. 5 on surface meshing and Chap. 6 on reconstruction. Other applications can be found in the surveys and the textbooks mentionned in the introduction.

Additively and multiplicatively weighted distances arise when modeling growing processes and have important applications in biology, ecology and other fields. Consider a number of crystals, all growing at the same rate, and all starting at the same time : one gets a number of growing circles. As these circles meet, they draw a Euclidean Voronoi diagram. In reality, crystals start

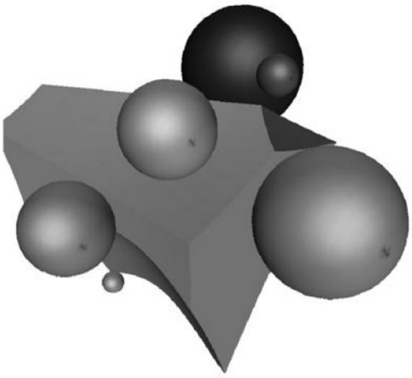

Fig. 2.16. A cell in an Apollonius diagram of spheres

growing at different times. If they still grow at the same rate, they will meet along an Apollonius diagram. This growth model is known as the Johnson-Mehl model in cell biology. In other contexts, all the crystals start at the same time, but grow at different rates. Now we get what is called the multiplicatively weighted Voronoi diagram, a special case of Möbius diagrams.

Spheres are common models for a variety of objects such as particles, atoms or beads. Hence, Apollonius diagrams have been used in physics, material sciences, molecular biology and chemistry [245, 339, 227, 228]. They have also been used for sphere packing [246] and shortest paths computations [256].

Euclidean Voronoi diagrams of non punctual objects find applications in robot motion planning [237, 197]. Medial axes are used for shape analysis [160], for computing offsets in Computer-Aided Design [118], and for mesh generation [290, 289, 316]. Medial axes are also used in character recognition, road network detection in geographic information systems, and other applications.

Acknowledgments

We thank D. Attali, C. Delage and M. Karavelas with whom part of the research reported in this chapter has been conducted. We also thank F. Chazal and A. Lieutier for fruitful discussions on the approximation of the medial axis.

;

(-& .- - -/ ' ! - ; = ! 9 "! ! ! "
&$$! - 5 ; $"

7' 5% 6 9 ' > ! 2 G ! ! ! +! + & ! $!
! 2 ! " ! ! ! '
!' ! " ' !' ! % * 6 %!6 ! + ! +! + %@6 *G ! +
! +! + % 6 % 6 ! +! + ! ! ! @! H I %! ! " @6# ! @! H
I % 6 & $ @2 &C ($2 ' + " @2 $ ' !$

9 ' %!6 ; $!

(-& .- - -/ ' ! - ; = ! 9 "! ! ! "
&$$! - ; 5 ; $"

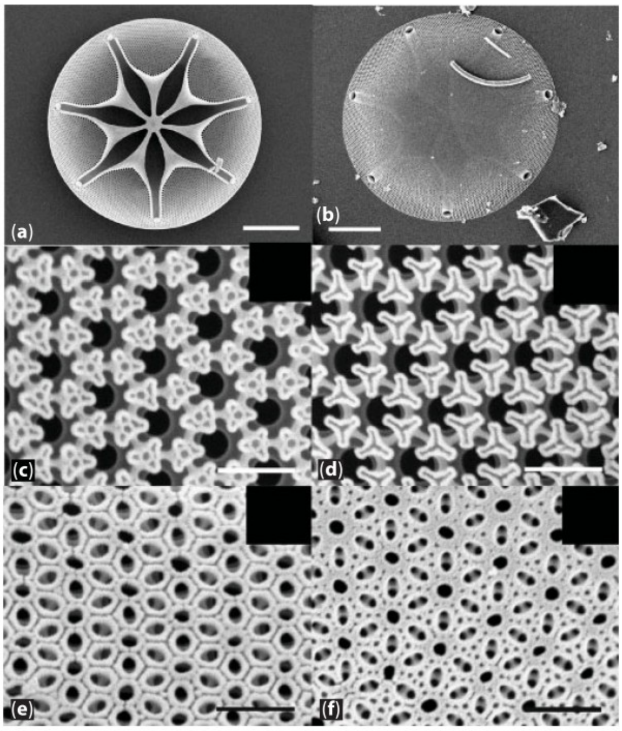

Figure 3.5 Beauty of the external surface of immature valves of *Asterolampra marylandica* at increasing stages and high magnification (SEM). (a) Internal view of whole mature valve. (b) External view of immature valve. (c)-(f) Surface of valve as it matures. Scale bars = 20 μm (a and b), scale bars = 1 μm (c-f). All photos by MAT. Copyright reserved by the photographer.

3.6 Jamin-Lebedeff Interference Contrast Microscopy

Using specialized microscopic illumination techniques, spectacularly beautiful photographs of diatoms may be made. Combining darkfield with polarized light techniques can reveal beautiful but subtle interference colors at medium power magnification. Phase contrast images of diatoms may be manipulated using digital editing programs to bring out latent colors which are the product of refraction and interference, but no technique really compares to images which can be created using Jamin-Lebedeff interference contrast equipment with white light. A number of English and German microscope manufacturers made this equipment early in the 20th century, but the equipment is rare, often mishandled and thus unusable, and at best it is difficult to use. The polarized illuminating beam of light from the condenser is split into two parallel light bundles by a calcite plate in the condenser, one passing adjacent to the diatom and the other passing through the frustule. The two beams are recombined in the objective, which is fitted with an identical but opposite calcite plate, and the resulting image can be manipulated to generate interference colors which vary according to the optical path length of the part of the diatom being illuminated. The second author of this article has made a specialty of imaging diatom using this

<

(-& .- - -/ ' ! - < ($ G & ! 2 C 8
$ C ! ! ! & ! & ! %)6-($ G & ! 2 C 8
$ C ! ! ! & ! & ! %)6 $"

7') % 6 7' %;)6

7') %)6) C ! A

(-& .- - -/ ' ! - < ($ G & ! 2 C 8
$ C ! ! ! & ! & ! %)6- < ($ G & ! 2 C
8 $ C ! ! ! & ! & ! %)6 $"

2. If g is holomorphic, then calculate that

$$\triangle(f \circ g) = (\triangle f \circ g) \cdot |g'|^2.$$

3. If f is holomorphic, then calculate that

$$\triangle(f \circ g) = (f' \circ g) \triangle g + (f'' \circ g)[(D_x g)^2 + (D_y g)^2].$$

3. Isometries

In any mathematical subject there are morphisms: functions which preserve the relevant properties being studied. In linear algebra these are invertible linear maps, in Euclidean geometry these are rigid motions, and in Riemannian geometry these are "isometries." We now define the concept of isometry.

Definition 1. Let Ω_1 and Ω_2 be planar domains and let

$$f : \Omega_1 \to \Omega_2$$

be a continuously differentiable mapping with isolated zeros. Assume that Ω_2 is equipped with a metric ρ. We define the *pullback* of the metric ρ under the map f to be the metric on Ω_1 given by

$$f^*\rho(z) = \rho(f(z)) \cdot \left|\frac{\partial f}{\partial z}\right|.$$

Remark. The particular form that we use to define the pullback is motivated by the way that f induces mappings on tangent and cotangent vectors, but this motivation is irrelevant for us here.

It should be noted that the pullback of any metric under a conjugate holomorphic f will be the zero metric. Thus we have designed our definition of pullback so that holomorphic pullbacks will be the ones of greatest interest. This assertion will be made substantive in Proposition 3 below. ∎

Definition 2. Let Ω_1, Ω_2 be planar domains equipped with metrics ρ_1 and ρ_2, respectively. Let

$$f : \Omega_1 \to \Omega_2$$

be an onto, holomorphic mapping. If

$$f^*\rho_2(z) = \rho_1(z)$$

for all $z \in \Omega_1$ then f is called an *isometry* of the pair (Ω_1, ρ_1) with the pair (Ω_2, ρ_2).

The differential definition of isometry (Definition 2) is very natural from the point of view of differential geometry, but it is not intuitive. The next proposition relates the notion of isometry to more familiar ideas.

Proposition 3. *Let Ω_1, Ω_2 be domains and ρ_1, ρ_2 be metrics on these respective domains. If*

$$f : \Omega_1 \to \Omega_2$$

is an isometry of (Ω_1, ρ_1) to (Ω_2, ρ_2), then the following three properties hold:

(a) *If $\gamma : [a, b] \to \Omega_1$ is a continuously differentiable curve, then so is the push-forward $f_*\gamma \equiv f \circ \gamma$ and*

$$\ell_{\rho_1}(\gamma) = \ell_{\rho_2}(f_*\gamma).$$

(b) *If P, Q are elements of Ω_1 then*

$$d_{\rho_1}(P, Q) = d_{\rho_2}(f(P), f(Q)).$$

(c) *Part (b) implies that the isometry f is one-to-one. Then f^{-1} is well defined and f^{-1} is also an isometry.*

Proof. Assertion (b) is an immediate consequence of (a). Also (c) is a formal exercise in definition chasing. Therefore we shall prove (a).

By definition,

$$\begin{aligned}\ell_{\rho_2}(f_*\gamma) &= \int_a^b \left\|(f_*\gamma)'(t)\right\|_{\rho_2, f_*\gamma(t)}\, dt \\ &= \int_a^b \left\|\frac{\partial f}{\partial z}(\gamma(t)) \cdot \dot{\gamma}(t)\right\|_{\rho_2, f_*\gamma(t)} dt.\end{aligned}$$

With elementary manipulations, we see that the integrand equals

$$\begin{aligned}\left|\frac{\partial f}{\partial z}(\gamma(t))\right| \cdot \|\dot{\gamma}(t)\|_{\rho_2, f_*\gamma(t)} &= \left|\frac{\partial f}{\partial z}(\gamma(t))\right| \cdot |\dot{\gamma}(t)| \cdot \rho_2(f(\gamma(t))) \\ &= \|\dot{\gamma}(t)\|_{f^*\rho_2, \gamma(t)} \\ &= \|\dot{\gamma}(t)\|_{\rho_1, \gamma(t)},\end{aligned}$$

since f is an isometry. Substituting this back into the formula for the length of $f_*\gamma$ gives

$$\ell_{\rho_2}(f_*\gamma) = \int_a^b \|\dot{\gamma}(t)\|_{\rho_1, \gamma(t)}\, dt = \ell_{\rho_1}(\gamma).$$

This ends the proof. ■

It is not difficult to see that any positively oriented rigid motion of the Euclidean plane—rotations, translations, and their compositions (but *not* reflections, because they are conjugate holomorphic)—is an isometry of the Euclidean metric. In the next section we shall consider the isometries of the Poincaré metric on the disc.

If you have previously studied metric space theory or Banach space theory then you may have already encountered the term “isometry.” The essential notion is that an isometry should preserve distance. In fact, you can prove as an exercise that if f is a holomorphic mapping of (Ω_1, ρ_1) onto (Ω_2, ρ_2) that preserves distance then f is an isometry according to Definition 2. (Hint: compose f with a curve and differentiate.)

We close this section by noting an important technical fact about isometries:

Proposition 4. *Let ρ_j be metrics on the domains Ω_j, $j = 1, 2, 3$, respectively. Let $f : \Omega_1 \to \Omega_2$ and $g : \Omega_2 \to \Omega_3$ be isometries. Then $g \circ f$ is an isometry of the metric ρ_1 to the metric ρ_3.*

Proof. We calculate that

$$\rho_3(g([f(z)]) \cdot |g'(f(z))| = \rho_2(f(z))$$

hence

$$\rho_3(g([f(z)]) \cdot |g'(f(z))| \cdot |f'(z)| = \rho_2(f(z)) \cdot |f'(z)| = \rho_1(z).$$

In other words,

$$\rho_3(g \circ f(z)) \cdot |(g \circ f)'(z)| = \rho_1(z),$$

as required by the definition of an isometry. ∎

4. The Poincaré Metric

The Poincaré metric on the disc has occurred in many of the examples in previous sections. This metric is the paradigm for much of what we want to do in this book, and we want to treat it in some detail here. The Poincaré metric on the disc D is given by

$$\rho(z) = \frac{1}{1 - |z|^2}.$$

(For the record, we note that there is no agreement in the literature as to what constant goes in the numerator; many references use a factor of $\sqrt{2}$ or $1/\sqrt{2}$.)

In this and succeeding sections, we shall use the phrase "conformal map" to refer to a holomorphic mapping of one planar region to another which is both one-to-one and onto.

Proposition 1. *Let ρ be the Poincaré metric on the disc D. Let $h : D \to D$ be a conformal self-map of the disc. Then h is an isometry of the pair (D, ρ) with the pair (D, ρ).*

Proof. We have that

$$h^*\rho(z) = \rho(h(z)) \cdot |h'(z)|.$$

Because of Theorem 3 of Section 0.2 (and using Proposition 4 of Section 1.3), we now have two cases:

(i) If h is a rotation, then $h(z) = \mu \cdot z$ for some unimodular constant $\mu \in \mathbb{C}$. So $|h'(z)| = 1$ and

$$h^*\rho(z) = \rho(h(z)) = \rho(\mu z) = \frac{1}{1 - |\mu z|^2} = \frac{1}{1 - |z|^2} = \rho(z)$$

as desired.

(ii) If h is a Möbius transformation (these were defined in Section 0.2, following Theorem 2), then

$$h(z) = \frac{z - a}{1 - \overline{a}z}, \qquad \text{some constant } a \in D.$$

But then

$$|h'(z)| = \frac{1 - |a|^2}{|1 - \overline{a}z|^2}$$

and

$$\begin{aligned} h^*\rho(z) &= \rho\left(\frac{z-a}{1-\overline{a}z}\right) \cdot |h'(z)| \\ &= \frac{1}{1 - \left|\frac{z-a}{1-\overline{a}z}\right|^2} \cdot \frac{1 - |a|^2}{|1 - \overline{a}z|^2} \\ &= \frac{1 - |a|^2}{|1 - \overline{a}z|^2 - |z - a|^2} \\ &= \frac{1 - |a|^2}{1 - |z|^2 - |a|^2 + |a|^2|z|^2} \\ &= \frac{1}{1 - |z|^2} \\ &= \rho(z). \end{aligned}$$

Since any conformal self-map of D is a composition of maps of the form (i) or (ii), the proposition is proved. ■

We know from this result and from Proposition 3 in Section 3 that conformal self-maps of the disc preserve Poincaré distance. To understand what this means, consider the Möbius transformation

$$\phi(z) = \frac{z+a}{1+\overline{a}z}.$$

Then ϕ maps the disc to the disc conformally. It does not preserve Euclidean distance, but it *does* preserve Poincaré distance. See Figure 1.

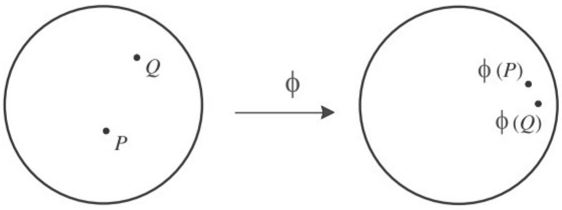

Figure 1.

We can use what we have learned so far to calculate the Poincaré metric explicitly.

Proposition 2. *If P and Q are points of the disc D, then the Poincaré distance of P to Q is*

$$d_\rho(P,Q) = \frac{1}{2}\log\left(\frac{1+\left|\frac{P-Q}{1-\overline{P}Q}\right|}{1-\left|\frac{P-Q}{1-\overline{P}Q}\right|}\right).$$

Proof. In case $P = 0$ and $Q = R + i0$, the result was already noted in Section 1. In the general case, note that we may define

$$\phi(z) = \frac{z-P}{1-\overline{P}z},$$

a Möbius transformation of the disc. Then, by Proposition 1,

$$d_\rho(P, Q) = d_\rho(\phi(P), \phi(Q)) = d_\rho(0, \phi(Q)).$$

Next we have

$$d_\rho(0, \phi(Q)) = d_\rho(0, |\phi(Q)|) \tag{$*$}$$

since there is a rotation of the disc taking $\phi(Q)$ to $|\phi(Q)| + i0$. Finally,

$$|\phi(Q)| = \left| \frac{P - Q}{1 - \overline{P}Q} \right|,$$

so that $(*)$ together with the special case treated in the the first sentence gives the result. ■

One of the reasons that the Poincaré metric is so useful is that it induces the same topology as the usual Euclidean metric topology. That is our next result.

Proposition 3. *The topology induced on the disc by the Poincaré metric is the usual planar topology.*

Proof. A neighborhood basis for the topology of the Poincaré metric at the origin is given by the balls

$$\mathbf{B}(0, r) = \{z : d_\rho(0, z) < r\}.$$

However, a calculation using Proposition 2 yields that these balls are the same as the Euclidean discs

$$\left\{ z : |z| < \frac{e^{2r} - 1}{e^{2r} + 1} \right\}.$$

These discs form a neighborhood basis for the origin in the Euclidean topology. Thus we find that the two topologies are the same at the origin. Now the origin can be moved to any other point $a \in D$ by the Möbius transformation

$$z \longmapsto \frac{z+a}{1+\overline{a}z}.$$

Since the Poincaré metric is invariant under Möbius transformations, and since Möbius transformations take circles to circles (after all, they are linear fractional), the two topologies are the same at every point.

■

One of the most striking facts about the Poincaré metric on the disc is that it turns the disc into a *complete* metric space. How could this be? The boundary is missing! The reason that the disc is complete in the Poincaré metric is the same as the reason that the plane is complete in the Euclidean metric—the boundary is infinitely far away. We now prove this assertion.

Proposition 4. *The unit disc D, when equipped with the Poincaré metric, is a complete metric space.*

Proof. Let p_j be a sequence in D that is Cauchy in the Poincaré metric. Then the sequence is bounded in that metric. So there is a positive, finite M such that

$$d_\rho(0, p_j) \leq M, \quad \text{all } j.$$

With Proposition 2 this translates to

$$\frac{1}{2}\log\left(\frac{1+|p_j|}{1-|p_j|}\right) \leq M.$$

Solving for $|p_j|$ gives

$$|p_j| \leq \frac{e^{2M}-1}{e^{2M}+1} < 1.$$

Thus our sequence is contained in a relatively compact subset of the disc. A similar calculation yields that the sequence must in fact be Cauchy in the Euclidean metric. Therefore it converges to a limit point in the disc, as required for completeness. ■

In the proof of Proposition 2 and in the subsequent remarks we used the fact, whose verification was sketched earlier, that the curve of least length (in the Poincaré metric) connecting 0 to a point of the form $R+i0$ is in fact a Euclidean segment. More generally, the shortest path from 0 to any point w is a rotation of the shortest path from 0 to $|w|+i0$, which is a segment. Let us now calculate the "curve of least length" connecting any two given points P and Q in the disc.

Proposition 5. *Let* P, Q *be elements of the unit disc. The "curve of least Poincaré length" connecting* P *to* Q *is*

$$\gamma_{P,Q}(t) = \frac{t\frac{Q-P}{1-Q\overline{P}} + P}{1 + t\overline{P}\cdot\frac{Q-P}{1-Q\overline{P}}}, \qquad 0 \leq t \leq 1.$$

Proof. Define the Möbius transformation

$$\phi(z) = \frac{z-P}{1-\overline{P}z}.$$

By what we already know about shortest paths emanating from the origin, the curve $\tau(t) \equiv t\cdot\phi(Q)$ is the shortest curve from $\phi(P)=0$ to $\phi(Q)$, $0 \leq t \leq 1$. Applying the isometry

$$\phi^{-1}(z) = \frac{z+P}{1+\overline{P}z}$$

to τ we obtain that

$$\phi^{-1}\circ\tau(t) = \frac{(t\phi(Q)+P)}{(1+\overline{P}t\phi(Q))}$$

is the shortest path from P to Q. Since $\phi^{-1}\circ\tau = \gamma_{P,Q}$, we are done.

■

Let us analyze the curves discovered in Proposition 5. First notice that, since the curve $\gamma_{P,Q}$ is the image under a linear fractional transformation of a part of a line, the trace of $\gamma_{P,Q}$ is, therefore, either a line segment or an arc of a circle. In fact, if P and Q are collinear with 0, then the formula for $\gamma_{P,Q}$ quickly reduces to that for a segment; otherwise $\gamma_{P,Q}$ traces an arc of a Euclidean circle. Which circle is it?

Matters are simplest if we let t range over the entire real line and look for the whole circle. We find that $t = \infty$ corresponds to the point $1/\overline{P}$. It is now a simple, but tedious, calculation to determine the Euclidean center and radius of the circle determined by the three points P, Q, $1/\overline{P}$ (note that, by symmetry, the circle will also pass through $1/\overline{Q}$). The circle is depicted in Figure 2. Notice that since the segment $\{t + i0 : -1 \leq t \leq 1\}$ is orthogonal to ∂D at the endpoints 1 and -1, conformality dictates that the circular arcs of least length provided by Proposition 5 are orthogonal to ∂D at the points of intersection. Some of these "geodesic arcs" are exhibited in Figure 3.

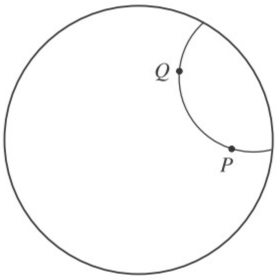

Figure 2.

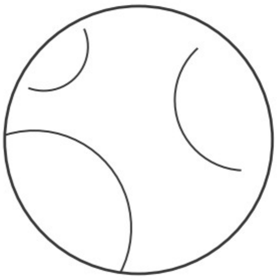

Figure 3.

A final note on this matter is that geodesics are particularly easy to calculate in the upper-half-plane realization $\mathcal{U}$ of the disc. For the geodesic circles will then have their centers on the boundary (i.e., the real line). If $\widetilde{P}$, $\widetilde{Q}$ are points of $\mathcal{U}$, then the perpendicular bisector of the segment connecting these points will intersect the real axis at the center C of the circular arc that forms the geodesic through $\widetilde{P}$ and $\widetilde{Q}$. See Figure 4.

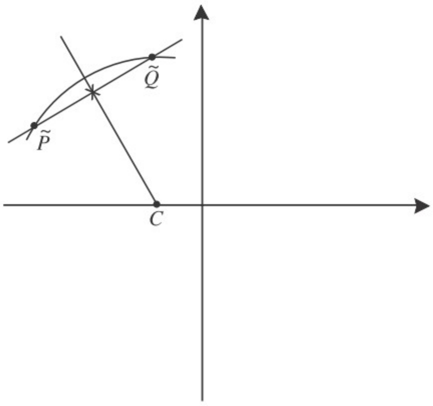

Figure 4.

We next see that the Poincaré metric is characterized by its property of invariance under conformal maps.

Proposition 6. *If* $\widetilde{\rho}(z)$ *is a metric on* D *which is such that every conformal map of the disc is an isometry of the pair* $(D, \widetilde{\rho})$ *with the pair* $(D, \widetilde{\rho})$*, then* $\widetilde{\rho}$ *is a constant multiple of the Poincaré metric* ρ*.*

Proof. The hypothesis guarantees that if $z_0 \in D$ is fixed and

$$h(z) = \frac{z + z_0}{1 + \overline{z}_0 z}$$

then

$$h^*\widetilde{\rho}(0) = \widetilde{\rho}(0).$$

Writing out the left-hand side gives

$$|h'(0)|\widetilde{\rho}(h(0)) = \widetilde{\rho}(0)$$

or

$$\widetilde{\rho}(z_0) = \frac{1}{1 - |z_0|^2} \cdot \widetilde{\rho}(0) = \widetilde{\rho}(0) \cdot \rho(z_0).$$

Thus we have exhibited $\widetilde{\rho}$ as the constant $\widetilde{\rho}(0)$ times ρ. ■

Now that we know that the Poincaré metric is the right metric for complex analysis on the disc, a natural next question is to determine which other maps preserve the Poincaré metric.

Proposition 7. *Let $f : D \to D$ be continuously differentiable and let ρ be the Poincaré metric. If f pulls back the pair (D, ρ) to the pair (D, ρ), then f is holomorphic and is one-to-one. We may conclude then that f is the composition of a Möbius transformation and a rotation (see Theorem 3 of Section 0.2).*

Proof. First suppose that $f(0) = 0$. For $R > 0$ let C_R be the set of points in D that have Poincaré distance R from 0. Since the Poincaré metric is invariant under rotations (after all, rotations are holomorphic self-maps), it follows that C_R is a Euclidean circle (however this circle will have Euclidean radius $(e^{2R}-1)/(e^{2R}+1)$, not R). Since $f(0) = 0$ and f preserves the metric, it follows that $f(C_R) = C_R$. The fact that f is distance-preserving then shows that f is one-to-one on each C_R. As a result, f is globally one-to-one.

Let $P \in C_R$. Then

$$\frac{|f(P) - f(0)|}{|P - 0|} = \frac{|f(P)|}{|P|} = 1.$$

Letting $R \to 0^+$ we conclude that f is conformal at the origin (that is, f preserves length infinitesimally). Since f pulls back the metric ρ, we can be sure that $\partial\rho/\partial z \neq 0$ at the origin.

Now we drop the special hypothesis that $f(0) = 0$. Pick an arbitrary $z_0 \in D$ and set $w_0 = f(z_0)$. Define

$$\phi(z) = \frac{z + z_0}{1 + \overline{z}_0 z}, \qquad \psi(z) = \frac{z - w_0}{1 - \overline{w}_0 z}.$$

Also define

$$g = \psi \circ f \circ \phi.$$

Then $g(0) = 0$ and g is an isometry (since ψ, f, and ϕ are); therefore the argument in the last paragraph applies and g is conformal at the origin. It follows as before that $\partial g/\partial z \neq 0$ at the origin.

We conclude that f is conformal at every point—since ψ and ϕ are—and $\partial f/\partial z \neq 0$ at every point. Therefore f is holomorphic. ■

Propositions 1 and 7 demonstrate that there are plenty of isometries of the unit disc in the Poincaré metric. On the other hand, isometries are very rigid objects. They are completely determined by their first-order behavior at just one point. While a proof of this assertion in general is beyond us at this point, we can certainly prove the result for the Poincaré metric on the disc.

Proposition 8. *Let ρ be the Poincaré metric on the disc. Let f be an isometry of the pair (D, ρ) with the pair (D, ρ). If $f(0) = 0$ and $\partial f/\partial z(0) = 1$ then $f(z) \equiv z$.*

Proof. By Proposition 7, f must be holomorphic. Since f preserves the origin and is one-to-one and onto, f must be a rotation. Since $f'(0) = 1$, it follows that f is the identity. ■

Corollary 8.1. *Let f and g be isometries of the pair (D, ρ) with the pair (D, ρ). Let $z_0 \in D$ and suppose that $f(z_0) = g(z_0)$ and $(\partial f/\partial z)(z_0) = (\partial g/\partial z)(z_0)$. Then $f(z) \equiv g(z)$.*

Proof. We noted in Section 2 that g^{-1} is an isometry. If ψ is a Möbius transformation that takes 0 to z_0, then $\psi^{-1} \circ g^{-1} \circ f \circ \psi$ satisfies the hypothesis of the Proposition. As a result $\psi^{-1} \circ g^{-1} \circ f \circ \psi(z) \equiv z$ or $g(z) \equiv f(z)$. ■

5. The Schwarz Lemma

One of the important facts about the Poincaré metric is that it can be used to study not just conformal maps (as in Section 1.4) but all holomorphic maps of the disc. The key to this assertion is the classical Schwarz lemma. We begin with an elegant geometric interpretation of the Schwarz–Pick lemma (see Section 0.2).

Proposition 1. *Let $f : D \to D$ be holomorphic. Then f is distance-decreasing in the Poincaré metric ρ. That is, for any $z \in D$,*

$$f^*\rho(z) \leq \rho(z).$$

The integrated form of this assertion is that if $\gamma : [0, 1] \to D$ is a continuously differentiable curve then

$$\ell_\rho(f_*\gamma) \leq \ell_\rho(\gamma).$$

Therefore, if P and Q are elements of D, we may conclude that

$$d_\rho(f(P), f(Q)) \leq d_\rho(P, Q).$$

Proof. Now

$$f^*\rho(z) \equiv |f'(z)|\rho(f(z)) = |f'(z)| \cdot \frac{1}{1 - |f(z)|^2}$$

and

$$\rho(z) = \frac{1}{1 - |z|^2}$$

5

(-& .- - -/ ' ! - 5 (! ! ! (+ "
$! C (! @ "' J + 2 % 6-(! ! ! (+ " $! C
(! @ "' J + 2 % 6 $"

/@ ? + " BBB

7')% 6 (!$ $! $ '

7' 5%;;6 (!$) $! 9 !

(-& .- - -/ ' ! - 5 (! ! ! (+ "
$! C (! @ "' J + 2 % 6-(! ! ! (+ " $! C
(! @ "' J + 2 % 6 $"

!'
(-& .- - -/ ' ! - 5 (! ! ! (+ "
$! C (! @ "' J + 2 % 6- 5 (! ! ! (+ " $!
C (! @ "' J + 2 % 6 $"

(-& .- - -/ ' ! - C K** &"+! " ($ +
! ! - C K** &"+! " ($ + ! ! $"

/@ ? + " BBB

7' 5 5 5

(!$ C = ! 8 2

(-& .- - -/ ' ! - C K** &"+! " ($ +
! ! - C K** &"+! " ($ + ! ! $"

data 19.07.2019
S T Q Q [✓ S] S D

Source: Comprehensive Mathematics for JEE/
(20190719.02) Advanced Comprehensive Mathematics.pdf
Pag. 915 (item 20.1)
three Dimensional Geometry
Coordinates.

Distance Formula

$$AB = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2}$$

Section Formulae

The coordinates of the point P which divides the join of $A(x_1, y_1, z_1)$ and $B(x_2, y_2, z_2)$ in the ratio $m:n$ are

$$\left(\frac{mx_2+nx_1}{m+n}, \frac{my_2+ny_1}{m+n}, \frac{mz_2+nz_1}{m+n}\right)$$

by taking $m=n$, we find the coordinates of the mid-point of AB as

$$\left(\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2}, \frac{z_1+z_2}{2}\right)$$

Coordinates of any point on the join of two points $A(x_1, y_1, z_1)$ and $B(x_2, y_2, z_2)$ are

$$\left(\frac{\lambda x_2+x_1}{\lambda+1}, \frac{\lambda y_2+y_1}{\lambda+1}, \frac{\lambda z_2+z_1}{\lambda+1}\right)$$ where

$\lambda \neq -1$. If λ takes only positive real values, we get the coordinates of the points on the segment AB (except A and B).

Jandaia

20

Three Dimensional Geometry

20.1 COORDINATES

Distance Formula

Distance between $A(x_1, y_1, z_1)$ and $B(x_2, y_2, z_2)$ is given by

$$AB = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}.$$

Section Formulae

The coordinates of the point P which divides the join of $A(x_1, y_1, z_1)$ and $B(x_2, y_2, z_2)$ in the ratio $m : n$ are

$$\left(\frac{mx_2 + nx_1}{m+n}, \frac{my_2 + ny_1}{m+n}, \frac{mz_2 + nz_1}{m+n}\right)$$

By taking $m = n$, we find the coordinates of the mid-point of AB as

$$\left(\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2}, \frac{z_1 + z_2}{2}\right)$$

Coordinates of any point on the join of two points $A(x_1,y_1,z_1)$ and $B(x_2,y_2,z_2)$ are $\left(\frac{\lambda x_2 + x_1}{\lambda + 1}, \frac{\lambda y_2 + y_1}{\lambda + 1}, \frac{\lambda z_2 + z_1}{\lambda + 1}\right)$ where $\lambda \neq -1$. If λ takes only positive real values, we get the coordinates of the points on the segment AB (except A and B).

Volume of a Tetrahedron

If $A_i(x_i, y_i, z_i)$, $i = 1, 2, 3, 4$ are the vertices of a tetrahedron, its volume is equal to

$$\frac{1}{6}\left|\begin{vmatrix} x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \\ x_4 & y_4 & z_4 & 1 \end{vmatrix}\right|$$

Direction-Cosines

If α, β, γ are the angles that a given directed line makes with the positive directions $X'OX$, $Y'OY$, $Z'OZ$ of the coordinate axes, then $\cos \alpha$, $\cos \beta$, $\cos \gamma$ are called the *direction-cosines* of the line.

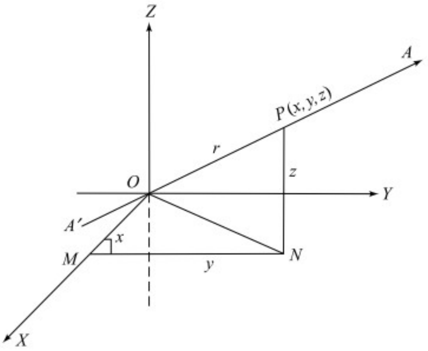

Fig. 20.1

Let $P(x, y, z)$ be any point on $A'OA$, and let the measure of OP be r. Let PN be the perpendicular from P on the plane XOY, and NM be the perpendicular from N on OX. Then $PN = z$, $NM = y$, $OM = x$

$$x = r \cos \alpha,\ y = r \cos \beta,\ z = r \cos \gamma.$$

and $$\cos^2 \alpha + \cos^2 \beta + \cos^2 \gamma = 1.$$

Direction-Ratios

If l, m, n be the direction-cosines of a line OP and a, b, c be numbers proportional to l, m, n respectively, then a, b, c are called the direction ratios of OP. If l, m, n be the direction-cosines of a line and k ($\neq 0$) be any number, then kl, km, kn are the direction-ratios of OP.

If a, b, c are the direction-ratios of a line, the actual direction-cosines $\cos \alpha$, $\cos \beta$, $\cos \gamma$ are obtained from the relations

$$\frac{\cos \alpha}{a} = \frac{\cos \beta}{b} = \frac{\cos \gamma}{c}$$

$$= \pm\frac{\sqrt{\cos^2 \alpha + \cos^2 \beta + \cos^2 \gamma}}{\sqrt{a^2 + b^2 + c^2}} = \frac{\pm 1}{\sqrt{a^2 + b^2 + c^2}}$$

If P be the point (a, b, c), and $\cos\alpha$, $\cos\beta$, $\cos\gamma$ are the direction cosines of the directed line OP, then

$$\cos\alpha = \frac{a}{\sqrt{(a^2+b^2+c^2)}},\ \cos\beta = \frac{b}{\sqrt{(a^2+b^2+c^2)}},$$

$$\cos\gamma = \frac{c}{\sqrt{(a^2+b^2+c^2)}}$$

The direction-cosines of PO are

$$-\frac{a}{\sqrt{(a^2+b^2+c^2)}}, -\frac{b}{\sqrt{(a^2+b^2+c^2)}}, -\frac{c}{\sqrt{(a^2+b^2+c^2)}}$$

Illustration 1

Find the direction cosines of the line whose direction ratios are 12, 4, – 18.

Solution: Direction cosines of the given line are

$$\frac{12}{\sqrt{(12)^2+(4)^2+(-18)^2}}, \frac{4}{\sqrt{(12)^2+(4)^2+(-18)^2}}, \frac{-18}{\sqrt{(12)^2+(4)^2+(-18)^2}}$$

$$\text{or } \frac{12}{22}, \frac{4}{22}, \frac{-18}{22}$$

$$\text{or } \frac{6}{11}, \frac{2}{11}, \frac{-9}{11}.$$

Angle Between Two Lines

(a) If θ is an angle between two lines whose direction-cosines are (l_1, m_1, n_1) and (l_2, m_2, n_2) then

$$\cos\theta = l_1l_2 + m_1m_2 + n_1n_2.$$

The expressions for $\sin\theta$, $\tan\theta$ are given below:

$$\sin^2\theta = \begin{vmatrix} m_1 & n_1 \\ m_2 & n_2 \end{vmatrix}^2 + \begin{vmatrix} n_1 & l_1 \\ n_2 & l_2 \end{vmatrix}^2 + \begin{vmatrix} l_1 & m_1 \\ l_2 & m_2 \end{vmatrix}^2$$

and $\tan\theta = \pm\,[\Sigma\,(l_1 m_2 - l_2 m_1)^2]^{1/2}/(l_1l_2 + m_1m_2 + n_1n_2)$

The lines are parallel to each other if and only if

$$l_1/l_2 = m_1/m_2 = n_1/n_2.$$

Also the lines are perpendicular to each other provided

$$l_1l_2 + m_1m_2 + n_1n_2 = 0.$$

(b) Angle θ between two lines whose direction-ratios are a_1, b_1, c_1 and a_2, b_2, c_2 is given by

$$\cos\theta = \pm\frac{a_1a_2 + b_1b_2 + c_1c_2}{\sqrt{a_1^2+b_1^2+c_1^2}\,\sqrt{a_2^2+b_2^2+c_2^2}}$$

$$\sin\theta = \pm\frac{[(a_1b_2-a_2b_1)^2+(b_1c_2-b_2c_1)^2+(c_1a_2-c_2a_1)^2]^{1/2}}{\sqrt{a_1^2+b_1^2+c_1^2}\,\sqrt{a_2^2+b_2^2+c_2^2}}$$

$$\tan\theta = \pm\frac{[(a_1b_2-a_2b_1)^2+(b_1c_2-b_2c_1)^2+(c_1a_2-c_2a_1)^2]^{1/2}}{a_1a_2+b_1b_2+c_1c_2}$$

The lines are parallel to each other if and only if

$$\frac{a_1}{a_2} = \frac{b_1}{b_2} = \frac{c_1}{c_2}$$

Also, the lines are perpendicular to each other if and only if

$$a_1a_2 + b_1b_2 + c_1c_2 = 0.$$

Illustration 2

Direction ratios of a line L_1, are 1, –1, 1 and of L_2 are 1, 0, λ. Find the value of λ so that L_1 and L_2 are perpendicular to each other.

Solution: The line L_1 and L_2 are perpendicular if the angle θ between them is 90°

$$\Rightarrow \cos\theta = \frac{1(1)+(-1)(0)+(1)(\lambda)}{\sqrt{(1)^2+(-1)^2+(1)^2}\,\sqrt{1^2+0+\lambda^2}} = \cos 90° = 0$$

$$\Rightarrow \lambda = -1$$

20.2 THE PLANE

Equation of a Plane

A plane is represented by an equation of the first degree i.e. by $ax + by + cz + d = 0$. Conversely, every equation of the first degree in x, y, z represents a plane.

(a) Equation of the yz-plane is $x = 0$.

(b) Equation of the zx-plane is $y = 0$.

(c) Equation of the xy-plane is $z = 0$.

(d) If l, m, n be the direction cosines of the normal to a plane and p be the length of the perpendicular from the origin on the plane, then an equation of the plane is $lx + my + nz = p$.

Illustration 3

Find the direction cosines of the normal to the plane $4x - 3y + 5z = 25$ and the length of the perpendicular from the origin on this plane

Solution: Direction ratios of the normal to the plane are 4, –3, 5. So direction cosines are

$$l = \frac{4}{\sqrt{4^2+(-3)^2+5^2}},\ m = \frac{-3}{\sqrt{4^2+(-3)^2+5^2}},\ n = \frac{5}{\sqrt{4^2+(-3)^2+5^2}}$$

$$\text{or } l = \frac{4}{5\sqrt{2}},\ m = \frac{-3}{5\sqrt{2}},\ n = \frac{5}{5\sqrt{2}}$$

We can write the equation of the plane as

$$\frac{4}{5\sqrt{2}}x - \frac{3}{5\sqrt{2}}y + \frac{5}{5\sqrt{2}}z = \frac{25}{5\sqrt{2}}$$

or $lx + my + nz = p$ where $p = \dfrac{25}{5\sqrt{2}} = \dfrac{5}{\sqrt{2}}$

is the required distance of the plane from the origin.

(e) If a plane makes intercepts a, b, c on the axes of x, y, z respectively its equation is

$$\frac{x}{a} + \frac{y}{b} + \frac{z}{c} = 1.$$

Systems of Planes

An equation of a plane contains three independent constants, as such a plane is uniquely determined by three independent conditions. If we consider a plane satisfying just two given conditions, its equation will contain one arbitrary constant. If we consider a plane satisfying one given condition, its equation will contain two arbitrary constants.

We give below the equations of some well-known systems of planes:

(i) The equation $ax + by + cz + k = 0$ represents a system of planes parallel to the plane $ax + by + cz + d = 0$, k being a parameter.

(ii) The equation $ax + by + cz + k = 0$, represents a system of planes perpendicular to the line $\dfrac{x}{a} = \dfrac{y}{b} = \dfrac{z}{c}$.

(iii) The equation $(a_1x + b_1y + c_1z + d_1) + k\,(a_2x + b_2y + c_2z + d_2) = 0$ represents a system of planes passing through the intersection of the planes $a_1x + b_1y + c_1z + d_1 = 0$ and $a_2x + b_2y + c_2z + d_2 = 0$, k being a parameter.

Illustration 4

Find the equation of the plane passing through the intersection of the planes $3x + 4y + 7z + 5 = 0$ and $2x - 3y + 4z - 6 = 0$ and making an intercept of 1 unit on the axis of x.

Solution: Equation of any plane through the intersection of the given planes is

$3x + 4y + 7z + 5 + \lambda\,(2x - 3y + 4z - 6) = 0$

Intercept on x-axis $= \dfrac{6\lambda - 5}{3 + 2\lambda} = 1 \Rightarrow \lambda = 2$

So the required equation of the plane is

$7x - 2y + 15z - 7 = 0$

(iv) The equation $A(x - x_1) + B(y - y_1) + C(z - z_1) = 0$. represents a system of planes passing through the point (x_1, y_1, z_1), the ratios of A, B, C being the two parameters.

It may be noted that (i)–(iii) above are examples of one-parameter family of planes and (iv) is an example of a two-parameter family of planes.

Angle Between Two Planes

Angle θ between the planes $a_1x + b_1y + c_1z + d_1 = 0$, $a_2x + b_2y + c_2z + d_2 = 0$, is given by

$$\cos\theta = \pm\frac{a_1a_2 + b_1b_2 + c_1c_2}{\sqrt{(a_1^2 + b_1^2 + c_1^2)}\,\sqrt{(a_2^2 + b_2^2 + c_2^2)}}$$

Parallelism and Perpendicularity of Two Planes

The planes $a_1x + b_1y + c_1z + d_1 = 0$, $a_2x + b_2y + c_2z + d_2 = 0$ are parallel if and only if

$$\frac{a_1}{a_2} = \frac{b_1}{b_2} = \frac{c_1}{c_2};$$

and perpendicular if and only if

$$a_1a_2 + b_1b_2 + c_1c_2 = 0.$$

Two Sides of a Plane

Two points $P(x_1, y_1, z_1)$ and $Q(x_2, y_2, z_2)$ lie on the same or different sides of the plane $ax + by + cz + d = 0$ according as $ax_1 + by_1 + cz_1 + d$ and $ax_2 + by_2 + cz_2 + d$ are of the same or of different signs.

Length of the Perpendicular from a Point to a Plane

The perpendicular distance of the point (x_1, y_1, z_1) from the plane $lx + my + nz = p$ is $|p - lx_1 - my_1 - nz_1|$, where l, m, n are direction cosines of the normal to the plane and p is the length of the perpendicular to the plane from the origin.

The perpendicular distance of the point (x_1, y_1, z_1) from the plane $ax + by + cz + d = 0$ is

$$\frac{|\,ax_1 + by_1 + cz_1 + d\,|}{\sqrt{(a^2 + b^2 + c^2)}}$$

Bisectors of the Angles between Two Planes

(A) Let $a_1x + b_1y + c_1z + d_1 = 0$ and $a_2x + b_2y + c_2z + d_2 = 0$ be the equations of two planes, written in such a way that d_1 and d_2 are both positive. Then

$$\frac{a_1x + b_1y + c_1z + d_1}{\sqrt{(a_1^2 + b_1^2 + c_1^2)}} = \frac{a_2x + b_2y + c_2z + d_2}{\sqrt{(a_2^2 + b_2^2 + c_2^2)}}$$

is the equation of the plane bisecting that angle between the given planes which contains the origin, and

$$\frac{a_1x + b_1y + c_1z + d_1}{\sqrt{(a_1^2 + b_1^2 + c_1^2)}} = -\frac{a_2x + b_2y + c_2z + d_2}{\sqrt{(a_2^2 + b_2^2 + c_2^2)}}$$

is the equation of the plane bisecting that angle between the given planes which does not contain the origin.

Illustration 5

Find the angle between the planes $x + y + z + 1 = 0$ and $x - y + z - 1 = 0$. Also find the equation of the plane bisecting the angle between these planes which contain the origin.

Solution: If θ is the angle between these planes, then

$$\Rightarrow \cos\theta = \frac{(1)\times(1)+(1)(-1)+(1)(1)}{\sqrt{1^2+1^2+1^2}\sqrt{(1)^2+(-1)^2+(1)^2}} = \frac{1}{3}$$

$$\Rightarrow \theta = \cos^{-1}(1/3)$$

Equation of the required plane bisecting the angle between the given planes is

$$\frac{x+y+z+1}{\sqrt{1^2+1^2+1^2}} = \frac{-x+y-z+1}{\sqrt{(1)^2+(-1)^2+(1)^2}}$$

$$\Rightarrow x+z=0$$

(B) The equations of the planes bisecting the angles between the planes $a_1x + b_1y + c_1z + d_1 = 0$ and $a_2x + b_2y + c_2z + d_2 = 0$ are

$$\frac{a_1x+b_1y+c_1z+d_1}{\sqrt{\left(a_1^2+b_1^2+c_1^2\right)}} = \pm\frac{a_2x+b_2y+c_2z+d_2}{\sqrt{\left(a_2^2+b_2^2+c_2^2\right)}}$$

Of these two planes one makes angle less than 45° with both the given planes and the other makes angles greater than 45° with both the given planes.

The bisector of the acute angle between the given planes is the one which makes angles less than 45° with both the given planes.

20.3 THE STRAIGHT LINE

Equations of a Line

An equation of the first degree represents a plane. Two equations of the first degree are satisfied by the coordinates of every point on the line of intersection of the planes represented by them. Therefore, the two equations together represent that line. Therefore $ax + by + cz + d = 0$ and $a'x + b'y + c'z + d' = 0$ together represent a straight line.

The Equations to the Axes of Coordinates

(a) Equations to the x-axis are $y = 0, z = 0$.
(b) Equations to the y-axis are $x = 0, z = 0$.
(c) Equations to the z-axis are $x = 0, y = 0$.

Symmetrical form of the Equation of a Line

If a straight line passes through a given point (x_1, y_1, z_1) and has direction cosines l, m, n, then the coordinates of any point on it satisfy the equations

$$\frac{x-x_1}{l} = \frac{y-y_1}{m} = \frac{z-z_1}{n} \; (= r)$$

These equations enable us to write down the coordinates of any point (x, y, z) on the line in terms of its distance r from the point (x_1, y_1, z_1) in the form:

$$x = x_1 + lr,\; y = y_1 + mr,\; z = z_1 + nr.$$

This form is called the *distance form* of the equation of a line.

Illustration 6

P and Q are two points on a line whose direction ratios are 3, 5, 1. If the coordinates of P are $(1, -4, 2)$ and $(PQ)^2 = 35$. Find the coordinates of Q.

Solution: Direction cosines of the line are

$$\frac{3}{\sqrt{3^2+5^2+1^2}},\quad \frac{5}{\sqrt{3^2+5^2+1^2}},\quad \frac{1}{\sqrt{3^2+5^2+1^2}}$$

or $\frac{3}{\sqrt{35}}, \frac{5}{\sqrt{35}}, \frac{1}{\sqrt{35}}$

So the equation of the line can be written as

$$\frac{x-1}{3/\sqrt{35}} = \frac{y+4}{5/\sqrt{35}} = \frac{z-2}{1/\sqrt{35}} = r = \sqrt{35} = PQ\,.$$

and the coordinates of Q are.

$(3 + 1, 5 - 4, 1 + 2) = (4, 1, 3)$

If l, m, n are only direction-ratios of the line and not the actual direction cosines, even then the equations represent a straight line, but in this case r is only a parameter, it is *not* the distance of (x, y, z) from (x_1, y_1, z_1).

Conversely, the equations

$$\frac{x-a}{l} = \frac{y-b}{m} = \frac{z-c}{n},$$

represent a straight line passing through the point (a, b, c) and having direction ratios l, m, n.

Equation of Line through two Given Points

If $A(x_1, y_1, z_1)$, $B(x_2, y_2, z_2)$ be two given points, an equation of the line AB is

$$\frac{x-x_1}{x_2-x_1} = \frac{y-y_1}{y_2-y_1} = \frac{z-z_1}{z_2-z_1}$$

The coordinates of a variable point on AB can be expressed in terms of a parameter λ in the form

$$x = \frac{\lambda x_2 + x_1}{\lambda+1},\; y = \frac{\lambda y_2 + y_1}{\lambda+1},\; z = \frac{\lambda z_2 + z_1}{\lambda+1},$$

λ being any real number different from -1. In fact, (x, y, z) are the coordinates of the point which divides the join of A and B in the ratio $\lambda : 1$.

Changing unsymmetrical form to symmetrical form

The unsymmetrical form of a line

$$ax + by + cz + d = 0,$$

$$a'x + b'y + c'z + d' = 0.$$

can be changed to symmetrical form as follows:

$$\frac{x - \dfrac{bd' - b'd}{ab' - a'b}}{bc' - b'c} = \frac{y - \dfrac{da' - d'a}{ab' - a'b}}{ca' - c'a} = \frac{z}{ab' - a'b}$$

Illustration 7

Find the direction ratios of the line $x + 2y - z = 3$, $2x - y + z = 1$ and its equation in a symmetrical form.

Solution: Direction ratios of the line are

$$\begin{vmatrix} 2 & -1 \\ -1 & 1 \end{vmatrix}, \; -\begin{vmatrix} 1 & -1 \\ 2 & 1 \end{vmatrix}, \; \begin{vmatrix} 1 & 2 \\ 2 & -1 \end{vmatrix}$$

or 1, –3, –5

For any point on the line, let $x = 0$, then $x = y = 1$ and the equation of the line in the symmetrical form is

$$\frac{x-1}{1} = \frac{y-1}{-3} = \frac{z-0}{-5} \qquad (1)$$

Verification Any point on the line is

$r + 1, -3r + 1, -5r$

$r + 1 + 2(-3r + 1) - (-5r) = 3$ for all r

and $2(r + 1) - (-3r + 1) + (-5r) = 1$ for all r.

showing that each point on the line satisfies the given equations of the line. (1)

Number of Constants in the Equation of a Line

The equation $\dfrac{x-a}{l} = \dfrac{y-b}{m} = \dfrac{z-c}{n}$,

of a line can be written as $x = (l/n)\, z + a - (lc/n)$,

$y = (m/n)\, z + b - (mc/n)$,

which are of the form $x = Az + B,\ y = Cz + D$

Therefore the *equation of a line contains four arbitrary constants.*

A Plane and a Straight Line

Let the equations

$$ax + by + cz + d = 0,\ \frac{x-\alpha}{l} = \frac{y-\beta}{m} = \frac{z-\gamma}{n},$$

represent a given plane and a straight line respectively.

(i) The line is perpendicular to the plane if and only if

$$\frac{a}{l} = \frac{b}{m} = \frac{c}{n},$$

(ii) The line is parallel to the plane if and only if $al + bm + cn = 0$.

(iii) The line lies in the plane if and only if

$$al + bm + cn = 0 \text{ and } a\alpha + b\beta + c\gamma + d = 0.$$

Angle between a Line and a Plane

The angle θ between the line

$$\frac{x-\alpha}{l} = \frac{y-\beta}{m} = \frac{z-\gamma}{n},$$

and the plane $ax + by + cz + d = 0$, is given by

$$\sin\theta = \frac{al + bm + cn}{\sqrt{(a^2 + b^2 + c^2)}\sqrt{(l^2 + m^2 + n^2)}}$$

Coplanar Lines

The lines $\dfrac{x-\alpha}{l} = \dfrac{y-\beta}{m} = \dfrac{z-\gamma}{n}$,

and $\dfrac{x-\alpha'}{l'} = \dfrac{y-\beta'}{m'} = \dfrac{z-\gamma'}{n'}$

are coplanar if and only if

$$\begin{vmatrix} \alpha - \alpha' & \beta - \beta' & \gamma - \gamma' \\ l & m & n \\ l' & m' & n' \end{vmatrix} = 0$$

In case the above condition is satisfied, the equation of the plane containing the lines is

$$\begin{vmatrix} x - \alpha & y - \beta & z - \gamma \\ l & m & n \\ l' & m' & n' \end{vmatrix} = 0.$$

Illustration 8

Find the value of λ, so that the lines

$\dfrac{x-1}{1} = \dfrac{y-2}{2} = \dfrac{z+\lambda}{3}$ and

$\dfrac{x+1}{2} = \dfrac{y-1}{3} = \dfrac{z-3}{1}$ are coplanar

Also find the equation of the plane containing them.

Solution: Given lines are coplanar if

$$\begin{vmatrix} 1+1 & 2-1 & -\lambda-3 \\ 1 & 2 & 3 \\ 2 & 3 & 1 \end{vmatrix} = 0 \Rightarrow \lambda = 6$$

and the equation of the plane containing the coplanar lines is

$$\begin{vmatrix} x-1 & y-2 & z+6 \\ 1 & 2 & 3 \\ 2 & 3 & 1 \end{vmatrix} = 0$$

or $7x - 5y + z + 9 = 0$

General Equation of a Plane Containing a Line

The general equation of the plane containing the line

$$\frac{x-\alpha}{l} = \frac{y-\beta}{m} = \frac{z-\gamma}{n},$$

is $$A(x-\alpha)+B(y-\beta)+C(z-\gamma)=0,$$
where $Al+Bm+Cn=0.$

Length of the Perpendicular from a Point to a Line

The length p of the perpendicular from a given point $P(x_1, y_1, z_1)$ to a given line $\dfrac{x-\alpha}{l}=\dfrac{y-\beta}{m}=\dfrac{z-\gamma}{n}$ (l, m, n are direction cosines of the line), is given by $p^2=(x_1-\alpha)^2+(y_1-\beta)^2+(z_1-\gamma)^2-[l(x_1-\alpha)+m(y_1-\beta)+n(z_1-\gamma)]^2$. Another expression for p^2 is given by

$$p^2=\begin{vmatrix} x_1-\alpha & y_1-\beta \\ l & m \end{vmatrix}^2+\begin{vmatrix} y_1-\beta & z_1-\gamma \\ m & n \end{vmatrix}^2+\begin{vmatrix} z_1-\gamma & x_1-\alpha \\ n & l \end{vmatrix}^2$$

Illustration 9

Find the perpendicular distance of the point (2, 1, 3) from the line
$$\frac{x-1}{3}=\frac{y-3}{1}=\frac{z-4}{-5}$$

Solution: Let $P(2, 1, 3)$ be the given point $A(1, 3, 4)$ is a point on the given line $p=PM$, be the perpendicular from P on the line.

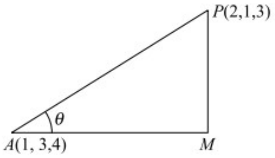

Fig. 20.2

Let $\lfloor\underline{PAM}=\theta$ then
$p^2=(PM)^2=(AP)^2-(AM)^2=(AP)^2-(AP\cos\theta)^2$
$\Rightarrow p^2=(AP)^2\sin^2\theta$
Direction ratios of AP are $2-1, 1-3, 3-4$
or $1, -2, -1$ and of the given line are $3, 1, -5$

$$\text{so } \cos\theta=\frac{(1)(3)+(-2)(1)+(-1)(-5)}{\sqrt{(1)^2+(-2)^2+(-1)^2}\sqrt{(3)^2+(1)^2+(-5)^2}}$$

$$=\frac{6}{\sqrt{6}\sqrt{35}}=\sqrt{\frac{6}{35}}$$

$(AP)^2=(2-1)^2+(1-3)^2+(3-4)^2=6$

Hence $p^2=6\times\dfrac{29}{35}\Rightarrow p=\sqrt{\dfrac{174}{35}}$

Equation of a Line through the Intersection of Given Lines

If $L_1=0=L_2$ and $L_3=0=L_4$, where $L_K=a_kx+b_ky+c_kz+d_k$, $k=1, 2, 3, 4$; are equations of two intersecting lines then for all values of λ and μ the equation $L_1+\lambda L_2=0=L_3+\mu L_4$ represents a line through their intersection.

20.4 VECTORIAL EQUATIONS

1. Parametric vectorial equation of the line through a point with position vector **a** and parallel to the vector **b** is
   $\mathbf{r}=\mathbf{a}+t\mathbf{b}$
   where t, is a scalar parameter.
2. Parametric vectorial equation of the line through two points with position vectors **a** and **b** is
   $\mathbf{r}=(1-\text{t})\mathbf{a}+t\mathbf{b}$
   where t, is a scalar parameter.
3. Parametric vectorial equation of a plane which passes through the point with position vector **a** and which is parallel to the vectors **b** and **c** is
   $\mathbf{r}=\mathbf{a}+t\mathbf{b}+p\mathbf{c}$
   where, t and p are scalar parameters.
4. Parametric vectorial equation of a plane passing through two given points with position vectors **a** and **b** and parallel to the vectors **c** is
   $\mathbf{r}=\mathbf{a}+t(\mathbf{b}-\mathbf{a})+p\mathbf{c}=(1-t)\,\mathbf{a}+t\mathbf{b}+p\mathbf{c}$
   where t and p are scalar parameters.
5. Equation of the plane passing through three points with position vectors **a**, **b**, **c** is
   $\text{r}\cdot(\mathbf{b}\times\mathbf{c}+\mathbf{c}\times\mathbf{a}+\mathbf{a}\times\mathbf{b})=[\mathbf{a}\ \mathbf{b}\ \mathbf{c}].$
6. *Normal form of the vector equation of a plane* $\mathbf{r}\cdot\mathbf{n}=p$ is vector equation of a plane, such that **n** is the unit vector normal to the plane and p is the length of the perpendicular from the origin to the plane.
7. $(\mathbf{r}-\mathbf{a})\cdot\mathbf{N}=0$ is vector equation of a plane normal to the vector **N** and passing through a point with position vector **a**.

Illustration 10

Find the vectorial equation of the plane passing through the point (2, 5, –3) and perpendicular to the line with direction ratios 3, 7, 5.

Solution: Let $\mathbf{a}=2\mathbf{i}+5\mathbf{j}-3\mathbf{k}$

and $\mathbf{N}=3\mathbf{i}+7\mathbf{j}+5\mathbf{k}$

Required equation of the plane is

$(\mathbf{r}-\mathbf{a})\cdot\mathbf{N}=0$

or $[\mathbf{r} - (2\mathbf{i} + 5\mathbf{j} - 3\mathbf{k})]\,(3\mathbf{i} + 7\mathbf{j} + 5\mathbf{k}) = 0$

8. *Angle between two planes* $\mathbf{r} \cdot \mathbf{N}_1 = p_1$ and $\mathbf{r} \cdot \mathbf{N}_2 = p_2$ is $\cos^{-1} \dfrac{\mathbf{N}_1 \cdot \mathbf{N}_2}{|\mathbf{N}_1||\mathbf{N}_2|}$
9. *Angle between a line and a plane* whose vectorial equations are $\mathbf{r} = \mathbf{a} + \mathbf{b}t$ and $\mathbf{r} \cdot \mathbf{N} = q$ respectively is $\sin^{-1} \dfrac{\mathbf{N} \cdot \mathbf{b}}{|\mathbf{N}||\mathbf{b}|}$.
10. *The perpendicular distance* of a point with position vector **a** from the plane $\mathbf{r} \cdot \mathbf{N} = q$ is $\dfrac{|q - \mathbf{a} \cdot \mathbf{N}|}{|\mathbf{N}|}$.
11. Equation of the plane containing two coplanar lines $\mathbf{r} = \mathbf{a} + t\mathbf{b}$ and $\mathbf{r} = \mathbf{c} + p\mathbf{d}$ is
    $(\mathbf{r} - \mathbf{a}) \cdot \mathbf{b} \times \mathbf{d} = 0$
    $\Rightarrow \quad \mathbf{r} \cdot \mathbf{b} \times \mathbf{d} = \mathbf{a} \cdot \mathbf{b} \times \mathbf{d}$.
12. Condition for the lines $\mathbf{r} = \mathbf{a} + t\mathbf{b}$ and $\mathbf{r} = \mathbf{c} + p\mathbf{d}$ to be *coplanar* is $[\mathbf{cbd}] = [\mathbf{abd}]$.
13. The line LM of shortest distance between the lines $\mathbf{r} = \mathbf{a} + t\mathbf{b}$ and $\mathbf{r} = \mathbf{c} + p\mathbf{d}$ is parallel to the vector $\mathbf{b} \times \mathbf{d}$. Length of LM is the projection of the line joining the points with position vectors **a** and **c** on **LM**.

    Thus $LM = \dfrac{|(\mathbf{c} - \mathbf{a}) \cdot (\mathbf{b} \times \mathbf{d})|}{|\mathbf{b} \times \mathbf{d}|} = \dfrac{|[\mathbf{cbd}] - [\mathbf{abd}]|}{|\mathbf{b} \times \mathbf{d}|}$

    Also, the line of shortest distance is the line of intersection of the two planes through the given lines and the line LM. The equations of these planes are
    $(\mathbf{r} - \mathbf{a}) \cdot \mathbf{b} \times (\mathbf{b} \times \mathbf{d}) = 0$,
    $(\mathbf{r} - \mathbf{c}) \cdot \mathbf{d} \times (\mathbf{b} \times \mathbf{d}) = 0$
    The line of intersection of these planes is the required line of shortest distance between the given lines.

Illustration 11

Find an equation of the line of shortest distance between the lines

$\mathbf{r} = \lambda(\mathbf{i} - \mathbf{j} + \mathbf{k})$ and $\mathbf{r} = (\mathbf{i} - \mathbf{j}) + \mu(-2\mathbf{j} + \mathbf{k})$

In the vectorial notation and Cartesian notation. Also find the shortest distance between them.

Solution: $\mathbf{a} = \mathbf{0}, \quad \mathbf{b} = \mathbf{i} - \mathbf{j} + \mathbf{k}$

$\mathbf{c} = \mathbf{i} - \mathbf{j}, \quad \mathbf{d} = -2\mathbf{j} + \mathbf{k}$

Then equation of the line of shortest distance is

$(\mathbf{r} - \mathbf{a}) \cdot \mathbf{b} \times (\mathbf{b} \times \mathbf{d}) = 0, (\mathbf{r} - \mathbf{c}) \cdot \mathbf{d} \times (\mathbf{b} \times \mathbf{d}) = 0$

$(\mathbf{b} \times \mathbf{d}) = \mathbf{i} - \mathbf{j} - 2\mathbf{k}$

$\mathbf{b} \times (\mathbf{b} \times \mathbf{d}) = 3\mathbf{i} + 3\mathbf{j}$

$\mathbf{d} \times (\mathbf{b} \times \mathbf{d}) = 5\mathbf{i} + \mathbf{j} + 2\mathbf{k}, \mathbf{c} \cdot [\mathbf{d} \times (\mathbf{b} \times \mathbf{d})] = 4$

So equation of the required line is

$\mathbf{r} \cdot (3\mathbf{i} + 3\mathbf{j}) = 0, \mathbf{r} \cdot (5\mathbf{i} + \mathbf{j} + 2\mathbf{k}) = 4$

$\mathbf{r} = (\mathbf{i} - \mathbf{j})$. satisfies both the equations.

So the line of shortest distance is

$\mathbf{r} = (\mathbf{i} - \mathbf{j}) + \lambda(\mathbf{i} - \mathbf{j} - 2\mathbf{k})$

Or $\dfrac{x-1}{1} = \dfrac{y+1}{-1} = \dfrac{z}{-2}$

Required shortest distance is $\dfrac{|(\mathbf{c} - \mathbf{a}) \cdot (\mathbf{b} \times \mathbf{d})|}{|\mathbf{b} \times \mathbf{d}|}$

$$= \frac{|(\mathbf{i} - \mathbf{j}) \cdot (\mathbf{i} - \mathbf{j} - 2\mathbf{k})|}{\sqrt{(1)^2 + (-1)^2 + (-2)^2}} = \frac{2}{\sqrt{6}}$$

SOLVED EXAMPLES

SINGLE CORRECT ANSWER TYPE QUESTIONS

Example 1 Let P be the image of the point (3, 1, 7) with respect to the plane $x - y + z = 3$. Then equation of the plane through P and containing the straight line $\dfrac{x}{1} = \dfrac{y}{2} = \dfrac{z}{1}$ is

(a) $x + y - 3z = 0$ (b) $3x + z = 0$
(c) $x - 4y + 7z = 0$ (d) $2x - y = 0$

Ans. (c)

Solution: An equation of normal to plane $x - y + z = 3$ and passing through $A(3, 1, 7)$ is

$\dfrac{x-3}{1} = \dfrac{y-1}{-1} = \dfrac{z-7}{1} = r$ (say)

Let the coordinates of P be $(r + 3, -r + 1, r + 7)$

Mid-point M of AP is $\left(\dfrac{r}{2}+3, -\dfrac{r}{2}+1, \dfrac{r}{2}+7\right)$ and lies on $x - y + z = 3$

$$\Rightarrow \quad \frac{r}{2}+3+\frac{r}{2}-1+\frac{r}{2}+7 = 3 \Rightarrow r = -4$$

Thus coordinates of P are $(-1, 5, 3)$

Now equation of the plane through P and containing the line $\dfrac{x}{1} = \dfrac{y}{2} = \dfrac{z}{1}$ is $\begin{vmatrix} x & y & z \\ -1 & 5 & 3 \\ 1 & 2 & 1 \end{vmatrix} = 0$

$\Rightarrow x - 4y + 7z = 0$

!))

(-& .- - -/ ' ! - ! $ '! (2
L! % 6- ! $ '! (2 L! % 6 $"

(-& .- - -/ ' ! - ! $ '! (2
L! % 6- ! $ '! (2 L! % 6 $"

7' 5% < 6))% < 6

> " ' $ 2!

data 19.07.2019 X

11) Inorganic Chemistry 4th Ed

Source: 20190719-03/Inorganic Chemistry, 4th edition.. / Catherine Housecroft, " Inorganic Chemistry...pdf

Pág. 140 (182a)

sp Hybridization: a scheme for linear species

The notation sp means that s atomic orbital and one p atomic orbital mix to form a set of two hybrid orbitals with different directional properties.

Off: Fig. 5.1 e 5.2 → shape of the H_2O molecule (interessante)

The model of sp hybridization can be used to describe the σ-bonding in a linear molecule such as $BeCl_2$ in which the Be-Cl bond are of equal length.

The ground state eletronic configuration of Be is $[He]2s^2$ and the valence shell contains the 2s atomic orbital and three 2p atomic orbitals (Fig. 5.3)

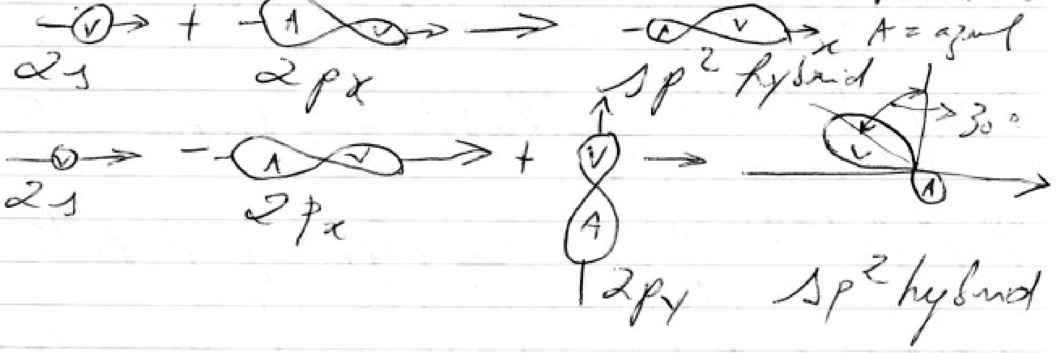

→

estudar!!!

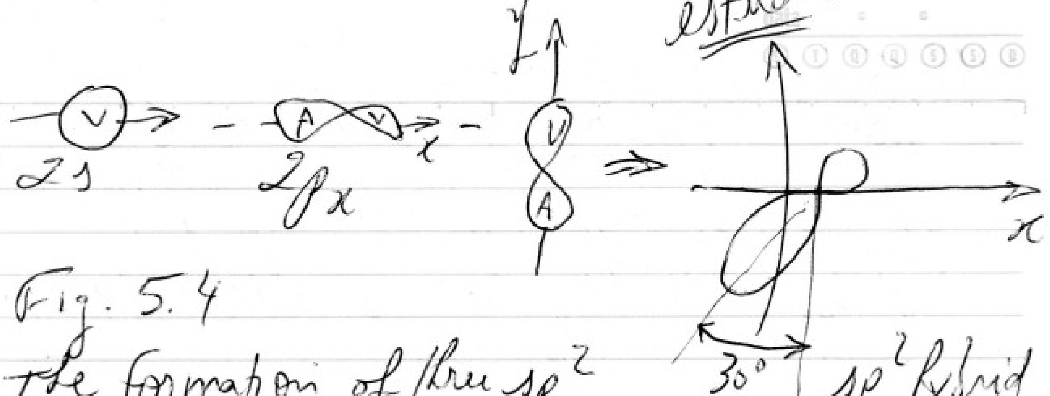

Fig. 5.4

the formation of three sp² hybrid orbitals from one 2s atomic orbitals and two 2p atomic orbitals.

obs: the choice of px and py is arbitrary. (If we started with 2px and 2pz atomic orbitals, the hybrid would lie in the ~~xy~~ xz plane; using the 2py and 2pz atomic orbitals gives hybrid orbitals in the yz plane).

- wave functions (5.3 – 5.5) Pag. 141 [183]

↳ sp² hybridization: a scheme for trigonal planar species

sp² → the notation sp² means that one s and two p atomic orbitals mix to form of three orbitals with different directional properties.

$$\psi_{sp^2 \text{ hybrid}} = \frac{1}{\sqrt{3}}\psi_{2s} + \sqrt{\frac{2}{3}}\psi_{2p_x} \quad (5.3)$$

$$\psi_{sp^2 \text{ hybrid}} = \frac{1}{\sqrt{3}}\psi_{2s} - \frac{1}{\sqrt{6}}\psi_{2p_x} + \frac{1}{\sqrt{2}}\psi_{2p_y} \quad (5.4)$$

$$\psi_{sp^2 \text{ hybrid}} = \frac{1}{\sqrt{3}}\psi_{2s} - \frac{1}{\sqrt{6}}\psi_{2p_x} - \frac{1}{\sqrt{2}}\psi_{2p_y} \quad (5.5)$$

Topics

- Hybridization of atomic orbitals
- Molecular orbital theory: polyatomics
- Application of character tables
- Delocalized bonding

5

Bonding in polyatomic molecules

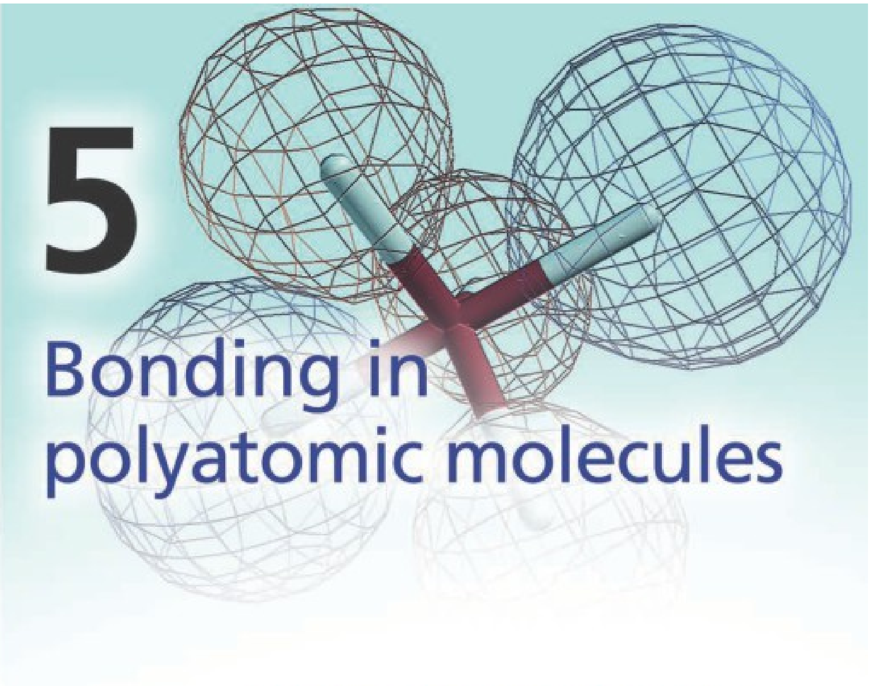

5.1 Introduction

In Chapter 2, we considered three approaches to the bonding in diatomic molecules:

- Lewis structures;
- valence bond (VB) theory;
- molecular orbital (MO) theory.

In this chapter we extend the discussion to polyatomic molecules (i.e. those containing three or more atoms). Within the valence bond model, treatment of a molecule XY_n ($n \geq 2$) raises the question of compatibility (or not) between the positions of the Y atoms and the directionalities of the atomic orbitals on the central atom X. Although an s atomic orbital is spherically symmetric, other atomic orbitals possess directional properties (see Section 1.6). Consider H_2O: Fig. 5.1 illustrates that, if the atoms of the H_2O molecule lie in (for example) the yz plane, the directionalities of the $2p_y$ and $2p_z$ atomic orbital of oxygen are not compatible with the directionalities of the two O–H bonds. Although we could define the z axis to coincide with one O–H bond, the y axis could not (at the same time) coincide with the other O–H bond. Hence, there is a problem in trying to derive a localized bonding scheme in terms of an atomic orbital basis set (see Section 2.3). The next section describes a bonding model within valence bond (VB) theory that overcomes this problem. After we have considered how VB theory views the bonding in a range of XY_n species, we move on to the problems of applying molecular orbital theory to polyatomic species.

A ***polyatomic species*** contains three or more atoms.

5.2 Valence bond theory: hybridization of atomic orbitals

What is orbital hybridization?

The word 'hybridization' means 'mixing' and when used in the context of atomic orbitals, it describes a way of deriving *spatially directed orbitals* which may be used within VB theory. Like all bonding theories, *orbital hybridization is a model*, and should *not* be taken to be a real phenomenon.

Hybrid orbitals may be formed by mixing the characters of atomic orbitals that are close in energy. The character of a hybrid orbital depends on the atomic orbitals involved and their percentage contributions. The labels given to hybrid orbitals reflect the contributing atomic orbitals, e.g. an sp hybrid possesses equal amounts of s and p orbital character.

Hybrid orbitals are generated by mixing the characters of atomic orbitals.

The reason for creating a set of hybrid orbitals is to produce a convenient bonding scheme for a particular molecular species. An individual hybrid orbital points along a given internuclear axis within the framework of the molecule under consideration, and use of a set of hybrid orbitals provides a bonding picture in terms of *localized σ-bonds*. In working through the rest of this section, notice that each hybridization scheme for an atom X in a molecule XY_n is appropriate only for a particular shape, the shape being defined by the number of attached groups and any lone pairs.

A set of hybrid orbitals provides a bonding picture for a molecule in terms of ***localized σ-bonds***.

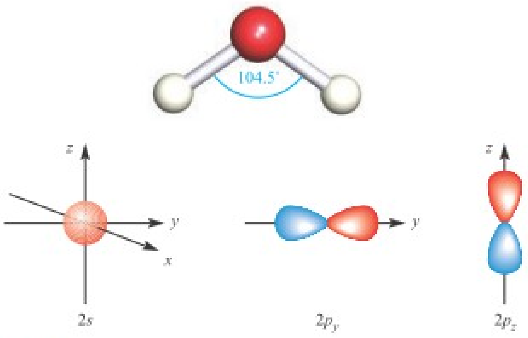

Fig. 5.1 A comparison of the shape of the H_2O molecule (the framework of which is taken as lying in the yz plane) with the spatial properties of the $2s$, $2p_y$ and $2p_z$ atomic orbitals of oxygen.

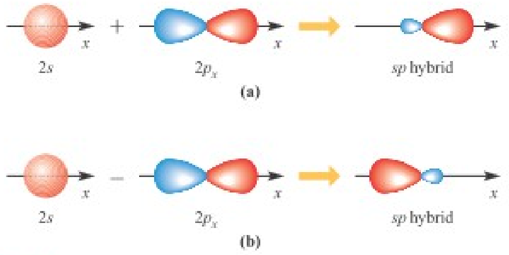

Fig. 5.2 The formation of two sp hybrid orbitals from one $2s$ atomic orbital and one $2p$ atomic orbital.

sp Hybridization: a scheme for linear species

The notation ***sp*** means that one s atomic orbital and one p atomic orbital mix to form a set of two hybrid orbitals with different directional properties.

One possible combination of a $2s$ atomic orbital and $2p_x$ atomic orbital is shown in Fig. 5.2a. In the figure, the colour of the orbital lobe corresponds to a particular phase (see Section 1.6) and the addition of the $2s$ component reinforces one lobe of the $2p_x$ atomic orbital but diminishes the other. Equation 5.1 represents the combination mathematically. The wavefunction $\psi_{sp\,\text{hybrid}}$ describes a normalized (see Section 2.2) sp hybrid orbital which possesses 50% s and 50% p character. Although eq. 5.1 and Fig. 5.2a refer to the combination of $2s$ and $2p_x$ atomic orbitals, this could just as well be $2s$ with $2p_y$ or $2p_z$, or $3s$ with $3p_x$, and so on.

$$\psi_{sp\,\text{hybrid}} = \frac{1}{\sqrt{2}}(\psi_{2s} + \psi_{2p_x}) \tag{5.1}$$

Now comes an important general rule: *if we begin with n atomic orbitals, we must end up with n orbitals after hybridization.* Figure 5.2b and eq. 5.2 show the second possibility for the combination of a $2s$ and a $2p_x$ atomic orbital. The sign change for the combination changes the phase of the $2p_x$ orbital and so the resultant hybrid points in the opposite direction to the one shown in Fig. 5.2a. (Remember that p atomic orbitals have vector properties.)

$$\psi_{sp\,\text{hybrid}} = \frac{1}{\sqrt{2}}(\psi_{2s} - \psi_{2p_x}) \tag{5.2}$$

Equations 5.1 and 5.2 represent two wavefunctions which are equivalent in every respect *except for their directionalities* with respect to the x axis. Although the orbital energies of the initial $2s$ and $2p_x$ atomic orbitals were different, mixing leads to two hybrid orbitals of equal energy.

The model of sp hybridization can be used to describe the σ-bonding in a linear molecule such as $BeCl_2$ in which the Be−Cl bonds are of equal length. The ground state electronic configuration of Be is $[He]2s^2$ and the valence shell contains the $2s$ atomic orbital and three $2p$ atomic orbitals (Fig. 5.3). If we use two of these atomic orbitals,

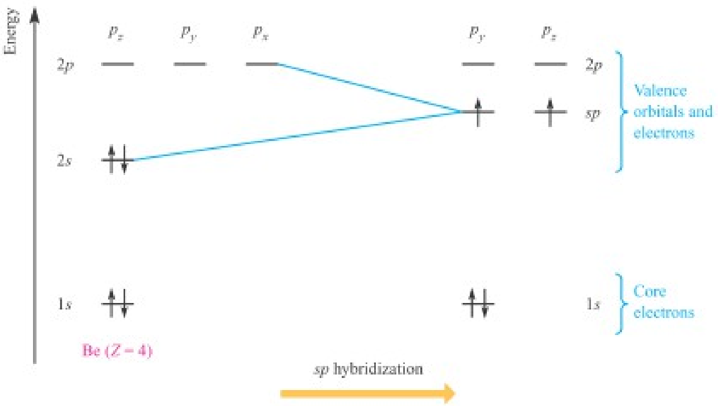

Fig. 5.3 Scheme to show the formation of the sp hybridized valence state of a beryllium atom from its ground state. This is a formalism and is not a 'real' observation, e.g. the valence state *cannot* be observed by spectroscopic techniques. The choice of using the $2p_x$ orbital for hybridization is arbitrary.

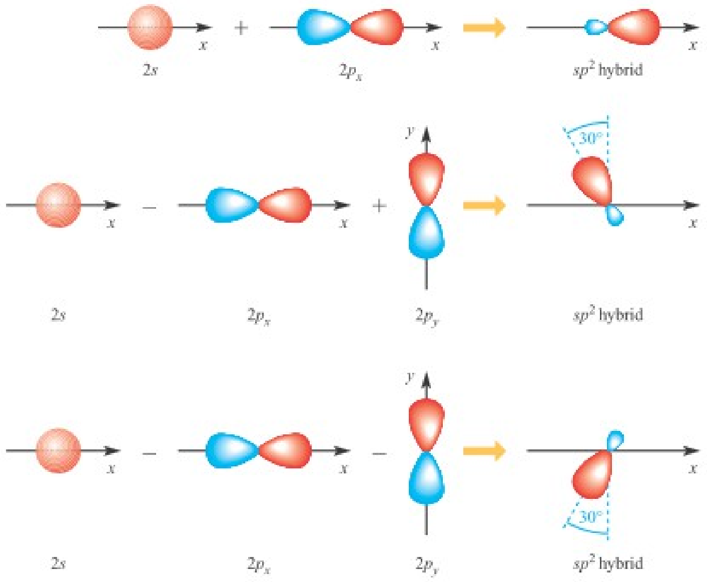

Fig. 5.4 The formation of three sp^2 hybrid orbitals from one 2*s* atomic orbital and two 2*p* atomic orbitals. The choice of p_x and p_y is arbitrary. (If we started with $2p_x$ and $2p_z$ atomic orbitals, the hybrids would lie in the *xz* plane; using the $2p_y$ and $2p_z$ atomic orbitals gives hybrid orbitals in the *yz* plane.) The directionalities of the hybrid orbitals follow from the relative contributions of the atomic orbitals (see eqs. 5.3–5.5).

treating them separately, to form two localized Be−Cl bonds, we cannot rationalize the bond equivalence. However, if we take the 2*s* atomic orbital and one 2*p* atomic orbital, mix their characters to form *sp* hybrids, and use one hybrid orbital to form one Be−Cl interaction and the other hybrid orbital for the second interaction, then the equivalence of the Be−Cl interactions is a natural consequence of the bonding picture. Effectively, we are representing the valence state of Be in a linear molecule as consisting of two degenerate *sp* hybrids, each containing one electron; this is represented by the notation $(sp)^2$. Figure 5.3 represents the change from the ground state electronic configuration of Be to an *sp* valence state. This is a *theoretical state* which can be used to describe σ-bonding in a linear molecule.

sp^2 Hybridization: a scheme for trigonal planar species

The notation $\boldsymbol{sp^2}$ means that one *s* and two *p* atomic orbitals mix to form a set of three hybrid orbitals with different directional properties.

Let us consider the combination of 2*s*, $2p_x$ and $2p_y$ atomic orbitals. The final hybrid orbitals must be equivalent in every way except for their directional properties; sp^2 hybrids must contain the same amount of *s* character as each other and the same amount of *p* character as one another. We begin by giving one-third of the 2*s* character to each sp^2 hybrid orbital. The remaining two-thirds of each hybrid orbital consists of 2*p* character, and the normalized wavefunctions are given in eqs. 5.3 to 5.5.

$$\psi_{sp^2\ \text{hybrid}} = \frac{1}{\sqrt{3}}\psi_{2s} + \sqrt{\frac{2}{3}}\psi_{2p_x} \tag{5.3}$$

$$\psi_{sp^2\ \text{hybrid}} = \frac{1}{\sqrt{3}}\psi_{2s} - \frac{1}{\sqrt{6}}\psi_{2p_x} + \frac{1}{\sqrt{2}}\psi_{2p_y} \tag{5.4}$$

$$\psi_{sp^2\ \text{hybrid}} = \frac{1}{\sqrt{3}}\psi_{2s} - \frac{1}{\sqrt{6}}\psi_{2p_x} - \frac{1}{\sqrt{2}}\psi_{2p_y} \tag{5.5}$$

Figure 5.4 gives a pictorial representation of the way in which the three sp^2 hybrid orbitals are constructed. Remember that a change in sign for the atomic wavefunction means a change in phase. The resultant directions of the lower two hybrid orbitals in Fig. 5.4 are determined by resolving the vectors associated with the $2p_x$ and $2p_y$ atomic orbitals.

The model of sp^2 hybridization can be used to describe the σ-bonding in trigonal planar molecules such as BH_3. The valence state of the B atom is $(sp^2)^3$ (i.e. three sp^2 hybrid orbitals, each with one electron) and the equivalence of the B−H interactions follows by considering that each interaction is formed by the overlap of one B sp^2 hybrid orbital with the 1*s* atomic orbital of an H atom (Fig. 5.5). Each H atom contributes one electron to the bonding scheme and, so, each B−H σ-bond is a localized 2c-2e interaction (see Section 2.2). A diagram similar to that shown in Fig. 5.3 can be constructed to show the formation of a valence state for the trigonal planar B atom.

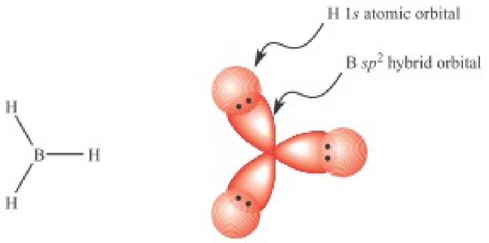

Fig. 5.5 The bonding in trigonal planar BH_3 can be conveniently described in terms of the interactions between a set of sp^2 hybrid orbitals centred on the B atom and three H 1s atomic orbitals. Three pairs of electrons are available (three electrons from B and one from each H) to give three 2c-2e σ-bonds.

sp^3 Hybridization: a scheme for tetrahedral and related species

The notation sp^3 means that one s and three p atomic orbitals mix to form a set of four hybrid orbitals with different directional properties.

A similar scheme to those described above can be derived to generate four sp^3 hybrid orbitals from one 2s and three 2p atomic orbitals. The sp^3 hybrid orbitals are described by the normalized wavefunctions in eqs. 5.6–5.9 and are shown pictorially in Fig. 5.6a. Each sp^3 hybrid orbital possesses 25% s character and 75% p character, and the set of four equivalent orbitals defines a tetrahedral framework.

$$\psi_{sp^3\text{ hybrid}} = \tfrac{1}{2}(\psi_{2s} + \psi_{2p_x} + \psi_{2p_y} + \psi_{2p_z}) \qquad (5.6)$$

$$\psi_{sp^3\text{ hybrid}} = \tfrac{1}{2}(\psi_{2s} + \psi_{2p_x} - \psi_{2p_y} - \psi_{2p_z}) \qquad (5.7)$$

$$\psi_{sp^3\text{ hybrid}} = \tfrac{1}{2}(\psi_{2s} - \psi_{2p_x} + \psi_{2p_y} - \psi_{2p_z}) \qquad (5.8)$$

$$\psi_{sp^3\text{ hybrid}} = \tfrac{1}{2}(\psi_{2s} - \psi_{2p_x} - \psi_{2p_y} + \psi_{2p_z}) \qquad (5.9)$$

Figure 5.6b illustrates how the tetrahedral structure of CH_4 relates to a cubic framework. This relationship is important because it allows you to describe a tetrahedron in terms of a Cartesian axis set. Within valence bond theory, the bonding in CH_4 can conveniently be described in terms of an sp^3 valence state for C, i.e. four degenerate orbitals, each containing one electron. Each hybrid orbital overlaps with the 1s atomic orbital of one H atom to generate one of four equivalent, localized 2c-2e C–H σ-interactions.

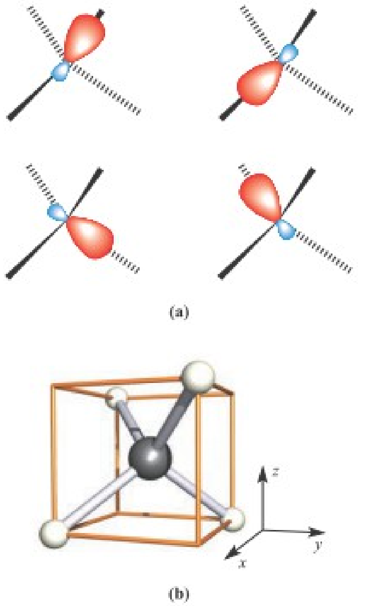

Fig. 5.6 (a) The directions of the orbitals that make up a set of four sp^3 hybrid orbitals correspond to a tetrahedral array. (b) The relationship between a tetrahedron and a cube; in CH_4, the four H atoms occupy alternate corners of a cube, and the cube is easily related to a Cartesian axis set.

Worked example 5.1 Hybridization scheme for the nitrogen atom in NH_3

Use the VSEPR model to account for the structure of NH_3, and suggest an appropriate hybridization scheme for the N atom.

The ground state electronic configuration of N is $[He]2s^22p^3$.

Three of the five valence electrons are used to form three N–H single bonds, leaving one lone pair.

The structure is trigonal pyramidal, derived from a tetrahedral arrangement of electron pairs:

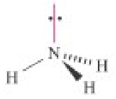

The N atom has four valence atomic orbitals: 2s, $2p_x$, $2p_y$ and $2p_z$. An sp^3 hybridization scheme gives a tetrahedral arrangement of hybrid orbitals, appropriate for accommodating the four pairs of electrons:

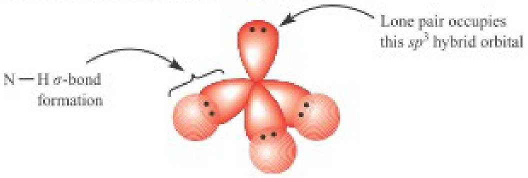

Self-study exercises

1. Use the VSEPR model to account for the tetrahedral structure of $[NH_4]^+$.

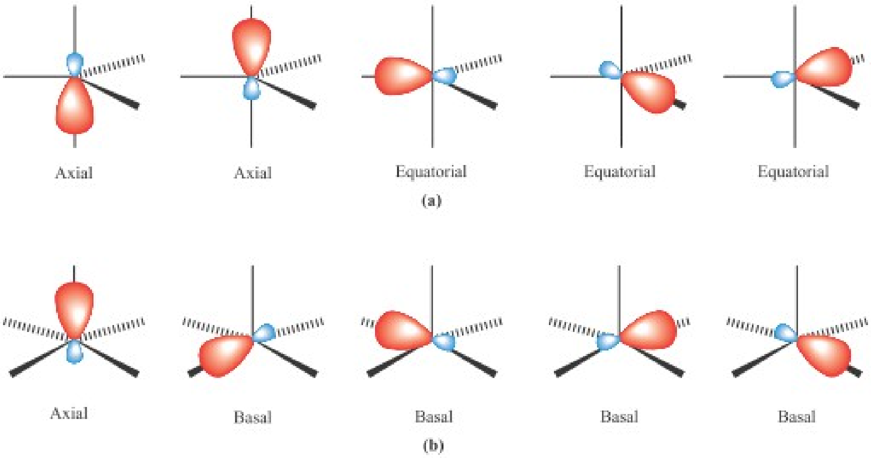

Fig. 5.7 A schematic representation of sp^3d hybridization. (a) A combination of s, p_x, p_y, p_z and d_{z^2} atomic orbitals gives a set of five sp^3d hybrid orbitals corresponding to a trigonal bipyramidal arrangement; the axial sp^3d hybrid orbitals are directed along the z axis. (b) A combination of s, p_x, p_y, p_z and $d_{x^2-y^2}$ atomic orbitals gives a set of five sp^3d hybrid orbitals corresponding to a square-based pyramidal arrangement; the axial sp^3d hybrid orbital is directed along the z axis.

2. Rationalize why H_2O is bent but XeF_2 is linear. Explain why an sp^3 hybridization scheme can be applied to H_2O, but not to XeF_2.
3. Give a suitable hybridization scheme for the central atom in each of the following: (a) $[NH_4]^+$; (b) H_2S; (c) BBr_3; (d) NF_3; (e) $[H_3O]^+$.
   [*Ans.* (a) sp^3; (b) sp^3; (c) sp^2; (d) sp^3; (e) sp^3]

Other hybridization schemes

For molecular species with other than linear, trigonal planar or tetrahedral-based structures, it is usual to involve d orbitals within valence bond theory. We shall see later that this is not necessarily the case within molecular orbital theory. We shall also see in Chapters 15 and 16 that the bonding in so-called *hypervalent compounds* such as PF_5 and SF_6, can be described without invoking the use of d-orbitals. One should therefore be cautious about using sp^nd^m hybridization schemes in compounds of p-block elements with apparently expanded octets around the central atom. *Real* molecules do not have to conform to simple theories of valence, nor must they conform to the sp^nd^m schemes that we consider in this book. Nevertheless, it is convenient to visualize the bonding in molecules in terms of a range of simple hybridization schemes.

The mixing of s, p_x, p_y, p_z and d_{z^2} atomic orbitals gives a set of five sp^3d hybrid orbitals, the mutual orientations of which correspond to a trigonal bipyramidal arrangement (Fig. 5.7a). The five sp^3d hybrid orbitals are *not* equivalent and divide into sets of two axial and three equatorial orbitals; the axial orbital lobes lie along the z axis.[†] The model of sp^3d hybridization can be used to describe the σ-bonding in 5-coordinate species such as $[Ni(CN)_5]^{3-}$ (see Section 21.11).

The σ-bonding framework in a square-pyramidal species may also be described in terms of an sp^3d hybridization scheme. The change in spatial disposition of the five hybrid orbitals from trigonal bipyramidal to square-based pyramidal is a consequence of the participation of a different d orbital. Hybridization of s, p_x, p_y, p_z and $d_{x^2-y^2}$ atomic orbitals generates a set of five sp^3d hybrid orbitals (Fig. 5.7b).

Hybridization of s, p_x, p_y, p_z, d_{z^2} and $d_{x^2-y^2}$ atomic orbitals gives six sp^3d^2 hybrid orbitals corresponding to an octahedral arrangement. The bonding in MoF_6 can be described in terms of sp^3d^2 hybridization of the central atom. If we remove the z-components from this set (i.e. p_z and d_{z^2}) and hybridize only the s, p_x, p_y and $d_{x^2-y^2}$ atomic orbitals, the resultant set of four sp^2d hybrid orbitals corresponds to a square planar arrangement, e.g. $[PtCl_4]^{2-}$.

Each set of hybrid orbitals is associated with a particular shape, although this may not coincide with the molecular shape if lone pairs also have to be accommodated:

• sp	linear
• sp^2	trigonal planar
• sp^3	tetrahedral
• sp^3d (d_{z^2})	trigonal bipyramidal
• sp^3d ($d_{x^2-y^2}$)	square-based pyramidal
• sp^3d^2	octahedral
• sp^2d	square planar

[†] Choice of coincidence between the z axis and the axial lobes is convenient and tends to be conventional.

(-& .- - -/ ' ! - 5<) . ! C 2
C '-5<) . ! C 2 C ' $"

7' M M% 6 MM% 6 "

! "

(-& .- - -/ ' ! - 5<) . ! C 2
C '- 5<) . ! C 2 C ' $"

7' 5%)6 C ' ! " J ! 2 $

(-& .- - -/ ' ! - 5<) . ! C 2
C '- 5<) . ! C 2 C ' $"

7' 5 % 6 5 % 6

" * " ! ' !

Introduction

Quantum tunneling is a microscopic phenomenon where a particle can penetrate and in most cases pass through a potential barrier. The maximum height of the barrier is assumed to be higher than the kinetic energy of the particle, therefore such a motion is not allowed by the laws of classical dynamics. The simplest problems in quantum tunneling are one-dimensional and most of the research is done on these problems. But the extension of one-dimensional tunneling to higher dimensions is not straightforward. In addition there are certain characteristics that appear in two- or three-dimensional tunneling problems which do not show up in the one-dimensional motion.

In most of the one-dimensional systems we study the motion of a particle in a potential $V(x)$, where $V(x)$ has a finite (or infinite) number of maxima. As long as the height of the barriers remain finite, the motion of the particle will not be restricted and we may choose the energy of the particle E to be greater than the asymptotic value of the potential say at $x = -\infty$. Then the simplest case will be that of a particle with energy $E > V(-\infty)$ approaching the barrier from the left and then penetrating the barrier. Now depending on whether $E > V(\infty)$ or $E < V(\infty)$ the particle can pass through the potential or be reflected back and move to $x = -\infty$. The value of E is arbitrary as long as these conditions or inequalities are met.

The other possibility is the one where the potential is finite on one side and tends to infinity on the other, e.g. $V(-\infty) = \infty$ and $V(\infty)$ is finite. In addition the potential has at least one local maximum say at $x = a$. Then depending on the boundary conditions of the problem we can have two different possibilities:

(i) - If the particle moves from $x = \infty$ in the direction of $x = -\infty$, then there is the possibility that by tunneling the particle can enter the region between a and $-\infty$ and stays there for a finite time (a metastable state).

(ii) - If the initial condition states that the particle is in the region $-\infty < x < a$, then in the course of time by the process of tunneling the particle passes through the barrier and goes to $x = \infty$. This initial state has also a finite lifetime. In both of these cases there are characteristic energies for which the tunneling probability is large, whereas for other energies it is small.

(iii) -The third possibility appears in the case where the potential $V(x)$ is a confining potential, i.e. $V(x)$ tends to infinity on both sides of the central maximum (or maxima), $V(x) \to \infty$ as $x \to \pm\infty$. In this case the motion of the particle will be restricted to a part of the x-axis. Depending on the number of maxima of $V(x)$, we can have the eigenvalue problem for a double- or multi-well potential. The motion of the wave packet for double-wells, due to their importance in applied physics have received extensive treatment and will be discussed at length in this book.

A double-well can be symmetric or asymmetric. The motion of a wave packet which represents the particle in a symmetric double-well, under certain conditions, can be obtained from the superposition of the eigenfunctions of the two lowest states of the system. When this happens then the wave packet oscillates between the two wells with a well defined frequency, and it also preserves its shape after successive back and forth tunneling. This very important case is called quantum coherence. For an asymmetric double-well, if the motion from one well to the other takes place by tunneling, then we have a situation which we call quantum hopping.

In two or three dimensions when the barrier is only a function of the distance from the origin of the coordinate system, i.e. $\rho = \sqrt{x^2 + y^2}$ in cylindrical coordinates or $r = \sqrt{x^2 + y^2 + z^2}$, in spherical polar coordinates we can separate the variables in the Schrödinger equation and thus reduce the problem to a one-dimensional motion but now with the boundary conditions imposed at $\rho = 0$ or $r = 0$ and at ρ or r going to infinity. For instance in three dimensions if we assume that $V(r \to \infty) \to 0$ and that $V(r)$ has a maximum at $r = a$, then if the particle is originally confined in the region $0 \leq r < a$, it can tunnel through the barrier and go to infinity. These special cases of two- or three-dimensional tunneling can be regarded as one-dimensional but the potentials $V(\rho)$ and $V(r)$ are replaced by $V_{eff}(\rho)$ and $V_{eff}(r)$, with the boundary condition that the reduced radial wave function must vanish at the origin.

A Brief Review of the Contents of This Book — In the first chapter of this book we present a brief history of the subject of quantum tunneling and the role that a number of pioneers played in its development. In the second chapter we discuss the physics of tunneling and the solution to the problem of local kinetic energy provided by the uncertainty principle. Furthermore we show that the principles of wave mechanics imply that, in general, the decay of the system, either because of tunneling or by some other mechanism, is non-exponential. Following this argument, we consider a special solvable problem to show that the exponential decay law is a very good approximation except for very short initial time and also after a very long time, and only at these extremes there are departures from the exponential decay. In the next chapter (Chapter 3) we solve a number of simple problems for which analytic solutions are known, and we determine, in appropriate cases, the lifetime of the quasi-stationary states, and or the motion of the wave packets. Of special interest is the resonant tunneling from either two identical barriers or a group of barriers.

The time-dependence of the wave function in one-dimensional tunneling is

2.1 Tunneling and the Uncertainty Principle

At the first sight the tunneling of a particle looks like a paradoxical problem, since if the height of the barrier is greater than the total energy of the particle,

$$E = \frac{1}{2m}p^2 + V(x) \tag{2.1}$$

then in the range b where $V(x) > E$, the kinetic energy $\frac{1}{2m}p^2$ is negative and p is imaginary, but this simple picture is not correct.

At the root of this paradox is our assumption that at each instant we know both the kinetic and the potential energy separately, or in other words we can assign values to the coordinate x and the momentum p simultaneously and this is in violation of the uncertainty principle. Here we want to know whether it is possible to determine the position of the particle when it is moving under the barrier or not. For this we observe that the particle can be at the point x where $E < V(x)$ but then according to the uncertainty principle its momentum is uncertain by an amount $\sqrt{\overline{\Delta p^2}}$. Thus if we know the position of the particle to be x, then its total energy cannot be E.

Position-Momentum Uncertainty in Tunneling — Since the transmission amplitude during tunneling in WKB approximation is proportional to (see Chapters 5)

$$\exp\left[-\frac{1}{\hbar}\int_{x_0}^{x}\sqrt{2m(V(x)-E)}dx\right], \tag{2.2}$$

where x_0 is the classical turning point, then the probability of finding the particle which is coming from the left to be on the right of the barrier, i.e. $x_0 + b$ is proportional to the square of this amplitude or to the factor

$$\exp\left[-\frac{2}{\hbar}\int_{x_0}^{x_0+b}\sqrt{2m(V(x)-E)}dx\right]. \tag{2.3}$$

Now if we want a non-negligible probability then we must have

$$2\sqrt{2m(V_m-E)}\,b \approx \hbar. \tag{2.4}$$

where V_m is the maximum height of the potential. To find the position of the particle inside the barrier, we have to measure its coordinate with an accuracy $\Delta x < b$, therefore the uncertainty in momentum for this measurement is

$$\overline{\Delta p^2} = \frac{\hbar^2}{4(\Delta x)^2} = \frac{\hbar^2}{4b^2}. \tag{2.5}$$

By substituting b from (2.4) in (2.5) we find

$$\frac{\overline{\Delta p^2}}{2m} = V_m - E. \tag{2.6}$$

Thus the kinetic energy of the particle must be greater than the difference between the height of the barrier V_m and the total energy E [1].

Time-Energy Uncertainty in Tunneling — A result similar to (2.5) and (2.6) can be obtained from the time-energy uncertainty relation [11] i.e. from

$$\Delta E \Delta t \approx \frac{\hbar}{2}. \tag{2.7}$$

Again let us denote the energy of the incident particle by E. For a very short time Δt, the uncertainty in the energy is ΔE, and for sufficiently small Δt, the energy of the particle $E+\Delta E$ is greater than the height of the barrier V_m. Tunneling takes place if in the time Δt the particle can traverse the barrier. For a rectangular barrier of width b this time Δt is given by

$$\Delta t = \frac{b}{\sqrt{\left(\frac{2}{m}\right)\left(E+\Delta E - V_m\right)}}. \tag{2.8}$$

From Eqs. (2.7) and (2.8) we find ΔE to be the solution of the quadratic equation

$$(\Delta E)^2 - \frac{\hbar^2}{2mb^2}\Delta E + \frac{\hbar^2}{2mb^2}\left(V_m - E\right) = 0, \tag{2.9}$$

and the condition for ΔE to be real is given by

$$\frac{\hbar^2}{8mb^2} > V_m - E, \tag{2.10}$$

which is the same as Eq. (2.6).

The uncertainty principle can also be used to estimate the tunneling time, i.e. the time that takes a particle to tunnel through a barrier. In order for a particle moving with the energy E to surmount a square barrier of height V_m, $(V_m > E)$ it must acquire an amount of energy $V_m - E$, and this according to the uncertainty principle can be achieved in a time Δt where

$$\Delta t \approx \frac{\hbar}{V_m - E}, \tag{2.11}$$

The time Δt is a crude upper bound for the tunneling time. For a rectangular barrier of width b, the effective speed of the particle according to (2.11) is of the order $b\frac{(V_m - E)}{\hbar}$ [12]. Different definitions of quantum mechanical tunneling times and their meanings will be discussed in detail later (Chapters 18-20).

A similar argument based on the time-energy uncertainty can be found in ref. [13].

7' 5 % 6

where in this relation $\frac{\omega_0}{2\pi}$ is the frequency of the oscillation of the particle between $r = 0$ and $r = a$

$$\frac{\omega_0}{2\pi} = \frac{K_0}{2ma} = \frac{1}{T_0}, \tag{13.33}$$

and $|p(r)| = \sqrt{v(r) - K_0^2}$. This is similar to the result that we found in Chapters 6 and 11.

The formulation of path integral presented in the preceding section can be applied to one-dimensional problems or separable problems in two or three dimensions. However its usefulness in solving nonseparable multidimensional tunnelings has not been investigated except for a very limited number of systems [16].

13.2 Method of Euclidean Path Integral

In the part of space where tunneling takes place, the momentum of the particle is imaginary, but if we choose the time variable as an imaginary variable, (called Euclidean time), then the momentum becomes real and the motion, in the sense of classical dynamics, is possible.

The advantage of introducing "imaginary time" is that it allows us to determine the action and the escape path using the laws of motion of classical mechanics. In introducing this fictitious time parameter along the path we have to be careful about the interpretation of the results concerning the motion of the particle. For instance, since pure imaginary time is somehow orthogonal to the real time, one may be inclined to say that the tunneling happens instantaneously. Here nothing is said about the quantum observation and or coupling to a measuring apparatus, and thus it does not provide us with a correct account of the tunneling time [17]. The question of tunneling time will be discussed at length in Chapters 19 and 20.

Imaginary Time Formulation of the Path Integral — As an introduction to this technique, we try to solve a one-dimensional problem where we can write for the path integral in the form of [18]–[20]

$$\left\langle x_f \left| \exp\left(-\frac{H\mathcal{T}}{\hbar}\right) \right| x_i \right\rangle = \mathcal{N} \int [\mathcal{D}(x)] \exp\left(-\frac{S}{\hbar}\right). \tag{13.34}$$

Here we have used $t = -i\mathcal{T}$ ($\mathcal{T}$ is a positive number called Euclidean time), H is the Hamiltonian operator of the particle, and $|x_i\rangle$ and $|x_f\rangle$ are the initial and final states of the particle. Now we expand $|x_i\rangle$ and $|x_f\rangle$ in terms of the energy eigenfunctions $|n\rangle$ of H and these are defined by the eigenvalue equation

$$H|n\rangle = E_n|n\rangle. \tag{13.35}$$

We also expand $|x_i\rangle$ and $|x_f\rangle$ in terms the eigenfunctions $|n\rangle$;

$$|x_i\rangle = \sum_n |n\rangle\langle n|x_i\rangle. \tag{13.36}$$

With the help of this expansion we write the left-hand side of (13.34) as

$$\left\langle x_f \left| \exp\left(-\frac{H\mathcal{T}}{\hbar}\right) \right| x_i \right\rangle = \sum_n \exp\left(-\frac{E_n\mathcal{T}}{\hbar}\right) \langle x_f\,|n\rangle\,\langle n|\,x_i\rangle\,. \tag{13.37}$$

We observe that for large $\mathcal{T}$ in the right-hand side of (13.37) only the lowest energy state and its corresponding wave function will make the major contribution. Now let us study the integral in Eq. (13.34) in which $\mathcal{N}$ is the normalization constant and S is the Euclidean action

$$S = \int_{-\frac{\mathcal{T}}{2}}^{\frac{\mathcal{T}}{2}} \left[\frac{1}{2}\left(\frac{dx}{d\tau}\right)^2 + V(x)\right] d\tau, \tag{13.38}$$

where we have assumed that the particle has a unit mass. In Eq. (13.34), $[\mathcal{D}(x)]$ shows that the integral is over all paths $x(t)$ which satisfy the initial and final conditions

$$x\left(-\frac{\mathcal{T}}{2}\right) = x_i, \quad \text{and} \quad x\left(\frac{\mathcal{T}}{2}\right) = x_f. \tag{13.39}$$

To make this idea clearer, we can define the paths in the following way: Let us assume that $\bar{x}$ is a function of t which satisfies the conditions (13.39). Then a general set of functions which satisfies (13.39) can be written as

$$x(\tau) = \bar{x}(\tau) + \sum_n c_n x_n(\tau) \tag{13.40}$$

where $\{x_n\}$ s form a complete set of orthonormal states which vanish at the initial and final times, $\tau = -\frac{\mathcal{T}}{2}$ and $\tau = \frac{\mathcal{T}}{2}$,

$$\int_{-\frac{\mathcal{T}}{2}}^{\frac{\mathcal{T}}{2}} x_m(\tau)x_n(\tau)d\tau = \delta_{nm}, \tag{13.41}$$

and

$$x_n\left(\pm\frac{\mathcal{T}}{2}\right) = 0. \tag{13.42}$$

Now we define $[\mathcal{D}(x)]$ in the following way

$$[\mathcal{D}(x)] = \prod_n \frac{1}{\sqrt{2\pi\hbar}}\, dc_n. \tag{13.43}$$

Again we observe that we can calculate the right-hand side of (13.34) in the semiclassical limit of $\hbar \to 0$. In this case, as we have seen earlier, the major contribution to the path integral comes from the path(s) for which S is minimum. For simplicity we assume that there is a single path and we denote it by $\bar{x}$. The minimum of S with respect to the variation of path is found from the equation of motion (for imaginary time)

$$\frac{\delta S}{\delta \bar{x}} = -\frac{d^2\bar{x}}{d\tau^2} + V'(\bar{x}) = 0, \tag{13.44}$$

in which $V'(x) = \frac{dV(x)}{dx}$. In addition we take $\{x_n\}$ s to be the eigenfunctions of the second derivative of S with respect to path, i.e. $\frac{\delta^2 S}{\delta x^2}$ for $x = \bar{x}$,

$$-\frac{d^2 x_n}{d\tau^2} + V''(\bar{x})x_n = \lambda_n x_n. \tag{13.45}$$

With these assumptions, in the limit of $\hbar \to 0$, the path integral is expressible as a product of Gaussian integrals

$$\begin{aligned}\left\langle x_f \left| \exp\left(-\frac{H\mathcal{T}}{\hbar}\right)\right| x_i \right\rangle &= \mathcal{N}\exp\left[-\frac{S(\bar{x})}{\hbar}\right]\prod_n \frac{1}{\sqrt{\lambda_n}}\left(1+\mathcal{O}(\hbar)\right)\\ &= \mathcal{N}\exp\left[-\frac{S(\bar{x})}{\hbar}\right]\left\{\det\left[-\frac{\partial^2}{\partial\tau^2}+V''(\bar{x})\right]\right\}^{-\frac{1}{2}}\\ &\quad\times\left(1+\mathcal{O}(\hbar)\right).\end{aligned} \tag{13.46}$$

For this derivation we have chosen all the eigenstates to be positive. Those cases where some of the eigenvalues are negative can be dealt with in a similar manner [20]. If there are a number of roots (or paths) for the equation $\frac{\delta S}{\delta x} = 0$ then we have to sum S over all of these paths.

Equation (13.45) is the equation of motion of a particle of unit mass in the potential $-V(x)$ for which the first integral is

$$E = \frac{1}{2}\left(\frac{d\bar{x}}{d\tau}\right)^2 - V(\bar{x}), \tag{13.47}$$

where E is a constant.

Euclidean Path Integral for a Potential with a Single Minimum — Let us apply these ideas to the motion where the potential is of the form shown in Fig. 13.3, and let us take $x_i = x_f = d$. Then $\bar{x} = 0$ is the only solution which satisfies the initial and the final conditions and for this solution $S = 0$. If we denote $V''(d) = \omega^2$, we want to show that

$$\mathcal{N}\left[\det\left(-\frac{\partial^2}{\partial\tau^2}+\omega^2\right)\right]^{-\frac{1}{2}} = \sqrt{\frac{\omega}{\pi\hbar}}\exp\left(-\frac{\omega\mathcal{T}}{2}\right). \tag{13.48}$$

To prove this relation we write Eq. (13.45) in the following way:

$$\left(-\frac{d^2}{d\tau^2}+\omega^2\right)x_n = \lambda_n x_n, \tag{13.49}$$

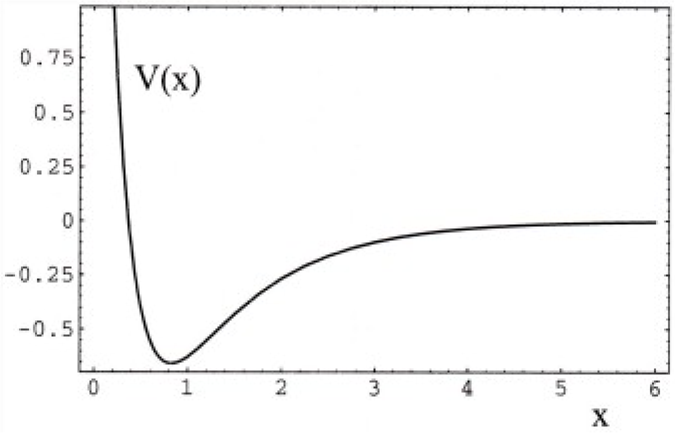

Figure 13.3: A potential with a single minimum at $x = d$.

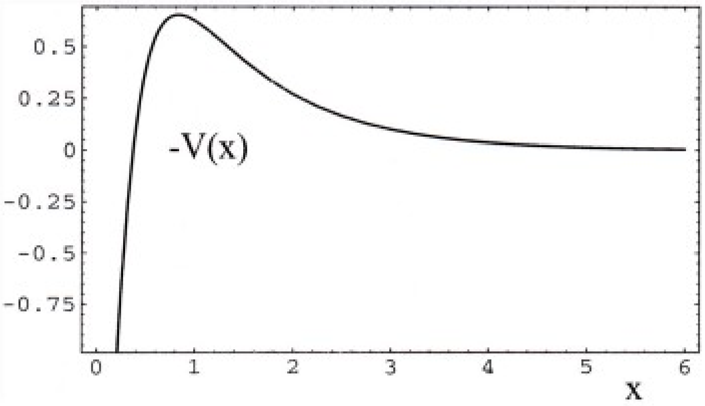

Figure 13.4: The negative of the potential shown in the previous figure.

where λ_n s are the characteristic values. The initial and final conditions for the path x_n is given by (13.42). From Eqs. (13.42) and (13.49) we conclude that

$$x_n = A_n \sin\left(\sqrt{\lambda_n - \omega^2}\,\tau\right), \tag{13.50}$$

in which

$$\lambda_n = \omega^2 + \left(\frac{n\pi}{\mathcal{T}}\right)^2. \tag{13.51}$$

Next we consider the product [7]

$$\mathcal{N} \lim_{N\to\infty} \prod_{n=1}^{N} \frac{1}{\sqrt{\omega^2 + \left(\frac{n\pi}{\mathcal{T}}\right)^2}}, \tag{13.52}$$

and write it as

$$\mathcal{N} \lim_{N\to\infty} \prod_{n=1}^{N} \frac{1}{\sqrt{\left(\frac{n\pi}{\mathcal{T}}\right)^2}} \prod_{n=1}^{N} \frac{1}{\sqrt{1+\left(\frac{\omega\mathcal{T}}{n\pi}\right)^2}}. \tag{13.53}$$

But the first product is independent of ω, therefore we choose $\mathcal{N}$ in such a way that

$$\mathcal{N} \lim_{N\to\infty} \prod_{n=1}^{N} \frac{1}{\sqrt{\left(\frac{n\pi}{\mathcal{T}}\right)^2}} = \frac{1}{\sqrt{2\pi\hbar\mathcal{T}}}, \tag{13.54}$$

Now we need to calculate the product

$$\mathcal{J} = \lim_{N\to\infty} \prod_{n=1}^{N} \frac{1}{\sqrt{1+\left(\frac{\omega\mathcal{T}}{n\pi}\right)^2}}. \tag{13.55}$$

To determine this infinite product, we start with another infinite product which can be written in closed form [23]

$$\left(1+\frac{z^2}{1^2}\right)\left(1+\frac{z^2}{2^2}\right)\left(1+\frac{z^2}{3^2}\right)\cdots = \frac{1}{\pi z}\sinh(\pi z). \tag{13.56}$$

By raising the two sides of (13.56) to the power of $(-\frac{1}{2})$ we find

$$\prod_{n=1}^{\infty} \frac{1}{\sqrt{1+\frac{z^2}{n^2}}} = \sqrt{\frac{\pi z}{\sinh(\pi z)}}. \tag{13.57}$$

Now we choose z to be $\frac{\omega\mathcal{T}}{\pi}$, then we get

$$\mathcal{J} = \sqrt{\frac{\omega\mathcal{T}}{\sinh(\omega\mathcal{T})}}, \tag{13.58}$$

therefore

$$\mathcal{N} \lim_{N\to\infty} \prod_{n=1}^{N} \frac{1}{\sqrt{\omega^2+\left(\frac{n\pi}{\mathcal{T}}\right)^2}} = \left(\frac{\omega}{2\pi\hbar}\right)^{\frac{1}{2}} \frac{1}{\sqrt{\sinh\omega\mathcal{T}}}. \tag{13.59}$$

For $\omega\mathcal{T} \gg 1$, this relation reduces to (13.48). The normalization constant $\mathcal{N}$ in (13.59) in the limit of $\omega \to 0$ is the same as the normalization constant for the motion of a free particle [7]. From Eqs. (13.46) and (13.48) it follows that the lowest energy eigenvalue is

$$E_0 = \frac{1}{2}\hbar\omega\left(1+\mathcal{O}(\hbar)\right). \tag{13.60}$$

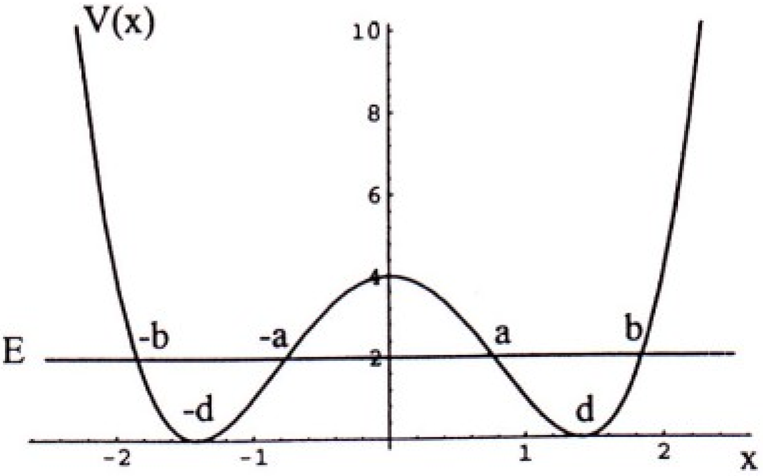

Figure 13.5: A symmetric double-well potential with minima at $x = -d$ and at $x = d$. For the energy E lower than $V(0)$, there are four turning points $-b$, $-a$, a and b.

13.3 Other Applications of the Path Integral Method in Tunneling

In Chapter 5, we studied tunneling in a symmetric double-well potential, and we found the splitting of levels for such a well as an application of WKB approximation. Now we apply the method of path integration to find the energy splitting between the two lowest states [20],[24].

Let us consider a symmetric double-well $V(x)$ shown in Fig. 13.5. The potential $V(x)$ is an even function of x, i.e. $V(x) = V(-x)$, and has two minima at $x = \pm d$. It is convenient to choose $V(x)$ so that $V(\pm d) = 0$. Here again we denote $V''(\pm d)$ by ω^2. First we want to calculate the two expectation values

$$\left\langle -d \left| \exp\left(-\frac{H\mathcal{T}}{\hbar}\right)\right| -d \right\rangle = \left\langle d \left| \exp\left(-\frac{H\mathcal{T}}{\hbar}\right)\right| d \right\rangle, \tag{13.61}$$

and

$$\left\langle d \left| \exp\left(-\frac{H\mathcal{T}}{\hbar}\right)\right| -d \right\rangle = \left\langle -d \left| \exp\left(-\frac{H\mathcal{T}}{\hbar}\right)\right| d \right\rangle. \tag{13.62}$$

To this end we find these path integrals in their semiclassical limit of $\hbar \to 0$, exactly as we found Eq. (13.44). For this we need to solve Eq. (13.44) with proper initial and final conditions. The two solutions of this problem are obtained by placing the particle on one of the two maxima in the potential $-V(x)$ of Fig. 13.6. But there is another interesting solution for this problem, and that is when the particle starts its motion at the time $\tau = -\frac{1}{2}\mathcal{T}$ at $x = -d$ and reaches $x = d$ at $\tau = \frac{1}{2}\mathcal{T}$. Since at the end we want to consider the limit

data 19.07.2019

12) Quantum theory of Tunneling

Source: 20190719_04/ Quantum theory of Tunneling/ jaret.it_981452506.pdf
Pág. XX(21)

In two or three dimensions when the barrier is only a function of the distance from the origin of the coordinate system, i.e. $\rho = \sqrt{x^2+y^2}$ in cylindrical coordinates or $r = \sqrt{x^2+y^2+y^2}$ in spherical polar coordinates we can separate the variables in the Schröndiger equation and thus reduce the problem to a one-dimensional motion but now with the boundary conditions imposed at $\rho = 0$ or $r = 0$ and at ρ or r going to infinity.

* For instance in three dimensions if we assume that $V(r \rightarrow \infty) \rightarrow 0$ and that $V(r)$ has a maximum at $r = a$, then if the particle is originally confined in the region $0 \leq r < a$, it can tunnel through the barrier and go to infinity.

These special cases of two- or three-dimensional tunneling can be regarded as one-dimensional but the potentials $V(\rho)$ and $V(r)$ are replaced by $V_{eff}(\rho)$ and $V_{eff}(r)$, with the boundary condition that the reduced radial wavefunction must vanish at the origin.

Important: Pág. 9 (34) →

2.1. Tunneling and the Uncertainty Principle

At the first sight the tunneling of a particle looks like a paradoxical problem, since if the height of the barrier is greater that (than) the total energy of the particle

$$E = \frac{1}{2m} p^2 + V(x) \qquad (2.1)$$

then in the range b where $V(x) > E$, the kinetic energy

$\frac{1}{2m} p^2$ is negative and p is imaginary,

but this simple picture is not correct.

∴ particle can be at the point x where $E < V(x)$ but then according to the uncertainty principle its momentum is uncertain by an amount $\sqrt{\Delta p^2}$.

Thus if we know the position of the particle to be x, then its total energy cannot be E.

Position-Momentum Uncertainty in Tunneling (WKB approximation - Chapter 5)

$$\exp\left[-\frac{1}{\hbar}\int_{x_0}^{x}\sqrt{2m(V(x)-E)}\,dx\right] \qquad (2.2)$$

where x_0 is the classical turning point ...

data 19 07 2019

13) Exploratory Galois theory
Source: 20190719-04/ Exploratory Galois theory/ Swallow J... (2004).pdf
Pág. 3 (17) Introduction

i is a certain solution to the polynomial equation $x^2 = -1$
obs: π → to see as an infinite sum.

i is a root of polynomial with rational coefficients, and the fact that π is not the root of a polynomial with rational coefficients – in other words, the fact that π is transcendental – means that the methods of studying i are very likely not going to be especially useful in studying π.

Approaching Galois theory, we choose to consider only those numbers that are roots of polynomials with rational number coefficients.

$1/7$ is root of $7x - 1$;
$\sqrt{2}$ is root of $x^2 - 2$;
$\sqrt[3]{-5}$ is a root of $x^3 + 5$;
$\sqrt[4]{20}$ is a root of $x^4 - 20$, and
$11 + 13/\sqrt{17}$ is a root of $x^2 - 22x + (1888/17)$.
We call a root of a polynomial with rational coefficients an algebraic number.

7' % ;6

(-& .- - -/ ' ! - ! K *G$! 2 8! C 2
%)6- ! K *G$! 2 8! C 2 %)6 $"

)

= $! ! # !$! "! ! N ! " ! 1 + # O!# ! " ! 1 + 7
! !, !"
/@
C " ! 1 + " + ! P " Q

(-& .- - -/ ' ! -) C , ! ! ! C * '
$! ' C 2 "' % 56 $"

7' % ;6

(-& .- - -/ ' ! - ! /$! 9 " 9 "! !
! " &$$! R " @ ' ($! 2 % 6 $"

7' %))6

(-& .- - -/ ' ! - ! " (! 9
L! % 5 <6 "O+

7' % 56

;

(-& .- - -/ ' ! - ; ! ! L! "@ 0 !
! " ! ! ((% 6 $"

7' ; % 6

<

(-& .- - -/ ' ! - < $! " ! !
$ C 2# " '# ! " ! R 2 % <6 $"

7' ;%) 6

(-& .- - -/ ' ! - < ! /$! 9 " 9 "! !
! " &$$! R " @ ' ($! 2 % 6 $"

7' %) 6

5

!6 (-& .- - -/ ' ! - 5
? ! E& ' @ !EE ESE&$$ "E $"

7' %)<6 55%) 6 (!$) $!

7') %);<6 / ' ! 2 % ! 6 O !

@6 (-& .- - -/ ' ! - 5 E " ! !
!+! ,!" $! ! ' $"

7' % 6 ;% 6 * $! + ! 2 $!" !

!6 8 ! &G 7 ! 8 ! $! ! ! # T" ! 2 ! ! T !

=! !E (E ; (! E $! 5E' ! $"

7')

<) * C ! " 7' ! % " !UV 6

7' < ?! ! " *

7') C !

C " $! $ 2 + " ! ! ' !

(-& .- - -/ ' ! - 5E' ! !G ! !E' !
$! ! ! " ! 2 ! ! ! $"

/@ ? + " BBB

@6 + ! & 3W K! &@ <
* 3 ; 5

7' % 6

(-& .- - -/ ' ! - <) > $"

6 & !UP 7 ! $!U " !"

5 ; * $!U E! ' ! <) > $"

! !

(-& .- - -/ ' ! - <) > $"

"6 CQ$ " L Q ! "! ! 7 ! ! !+A " @ ! A

5 ; * $!U E! ' ! $"

! ! (-& .- - -/ ' ! - $"

7' T $ O UP

7') ! ! %? @ ,6

7' ; UP ' A ! %? @ ,6

!6 (-& .- - -/ ' ! - C $ E8 ! $"

7' ; C ! " = ! ' !$$

/@ ? + " BBB

@6 (-& .- - -/ ' ! - 5E' !
!G ! !E' ! $! ! ! " ! 2 ! ! ! $"

C 8 T! &G 7 ! % $! 6

7')) C ! C ! " L @ '
* C ! " = ! ' $ " " ! ,! " W ! 2 G +!
C ! " !$$ 2 !G ! " !

X !$$! $! ! = ! ' ! Y %8 L @ '# ! ! & ! # % 5 6 $$
; 6

/

6 (-& .- - -/ ' ! - " ! ' @ " $"

& (! "2 9 ! , ' O + 8 2 (1 = ! ' C

"6 (-& .- - -/ ' ! - 55 5 5 $"

& $ L @ ': C & (

(-& .- - -/ ' ! - &"+! " * ' ' * !'
% " *" 6 K R 2 ! " % 6 $"

7' <)% 6 $ ' ! " @2 !$$ " !' " %9 ' 6

($! ! "

(-& .- - -/ ' ! - 9 "! !+
* ' ' R 2 *** % 6 $"

/@ ? + " N ! " $

(-& .- - -/ ' ! -
($ E8 !$ E! "E8 E " 'E% 6 $"

)

(-& .- - -/ ' ! -) & " 8 ! ! + 2
! " ('2 (! @ "' J + 2 % 6 $"

(-& .- - -/ ' ! - (! 9 "
K/ =&3 $"

7' <)%)6 <;% ;6 $! 7' ! 55 8 ' 2: ! 9 !

(* : !

(-& .- - -/ ' ! - K! 1 "! ! !
' 'E ! @ 0) 5 /G " J + 2 E% <6 $"

7' 5 % 56 < 3 Z ? @ ,# ;

;

(-& .- - -/ ' ! - ; < ; @)) $"

= !UP " W $ %[6 $ " ! "
1 !" !"! " "N = ! $ %[6 @ ' 1 ! " !

<

(-& .- - -/ ' ! - < & ' ! 1 ! $"

5

(-& .- - -/ ' ! - 5 (! ' 2 0 (!$
! ! (' @ ! % 6 $"

$! BBB

(-& .- - -/ ' ! - 5< < ;) $"

* " ! = ' C 2

7' 5% ;6 (@! !

!6 (-& .- - -/ ' ! - ! & # R L % " 6
! ! ! " " ' ! $ # R 2 % 5 6 "O+

@6 (-& .- - -/ ' ! - ! & # R L % " 6
! ! ! " " ' ! $ # R 2 % 5 ;6 "O+

7' %)<6

. !" ! C 2

6 (-& .- - -/ ' ! - L > ! . !" ! 2
& ! ! ! ! 2 % 6 $"

. !" ! C 2 (!$ 8! ! . !" !

7')5% <6 5% <6 ! T BBB

---

5 ;)

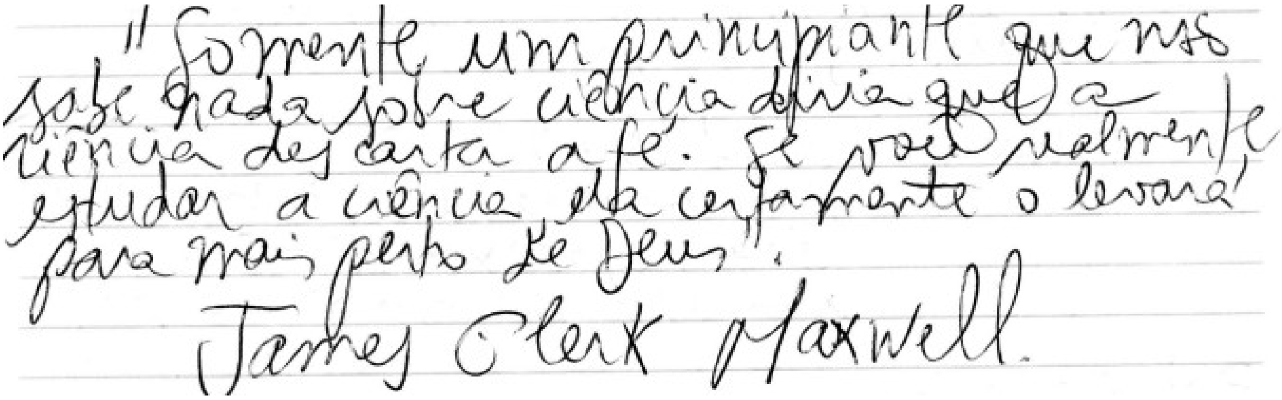

"Somente um principiante que não sabe nada sobre ciência diria que a ciência descarta a fé. Se você realmente estudar a ciência, ela certamente o levará para mais perto de Deus".

James Clerk Maxwell.

(-& .- - -/ ' ! - & ! " @ " "
(! ! E =2 ! ! 2 ((% 6 $"

7' % <6 9 ' \ $!

)

(-& .- - -/ ' ! -) " (! ! E 9
C 2 &$$! E " ' $"

7' % 6 9 ' 5

9 ' 5 ?!2 G ! ! " :! 2 " ! ": ' !

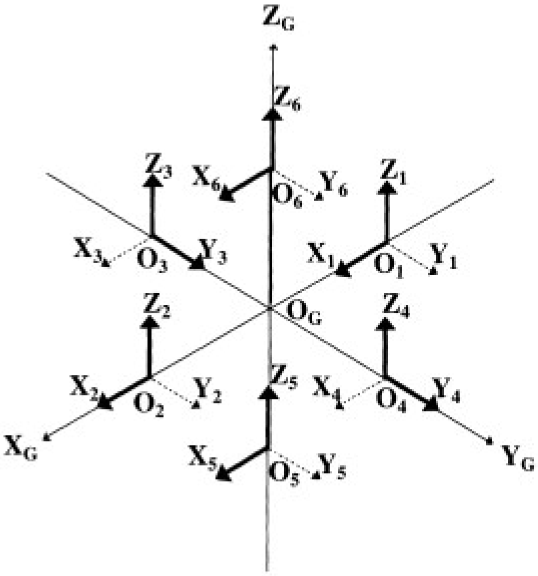
Z_G
Z_6
Z_3
X_6
Y_6
Z_1
O_6
X_3
O_3
Y_3
X_1
O_1
Y_1
Z_2
O_G
Z_4
X_2
Y_2
Z_5
X_4
O_4
Y_4
O_2
X_G
Y_G
X_5
O_5
Y_5

(-& .- - -/ ' ! - ! R @ : ! = ! 2 $"

7' % 6 (! (" ! 2

(-& .- - -/ ' ! - C "2 ! *! ! " ! ! 2
$ ' !' > L " @ ' % <6 $"

7')%) 6 + " @ !

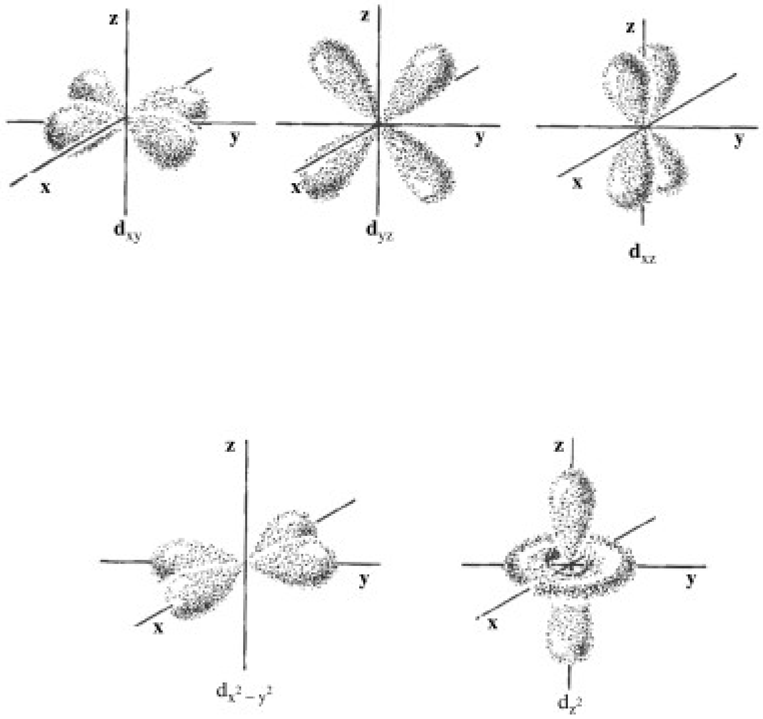
z
y
x
d_{xy}
z
y
x
d_{yz}
z
y
x
d_{xz}
z
y
x
$d_{x^2-y^2}$
z
y
x
d_{z^2}

9 ' 5 C $! ! ! + " @ ! ! ! ' ! 2 " ' ! ! $
! 2 2 +] " , @ ! ! @ ' ! " @2 ! !@
, !G 9 / ' % 5 6

Fig. 1.9 The spatial orientations of the five d-orbitals that are energetically degenerate in a spherically symmetric environment; the form of the d_{z^2} orbital can be generated by rotation about the z axis. From Orgel (1966)

;

(-& .- - -/ ' ! - ; & ! ! 8 " ? ' +
2 ! " & ! 2 $ ' ! ! @ ' % 6 $"

7' 5 % 6

(' (! ' ' & " *! 8 2
! " ! & '

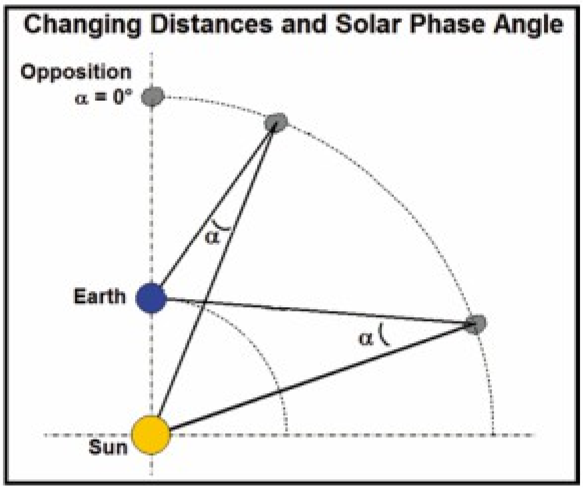
Changing Distances and Solar Phase Angle
Opposition
α = 0°
α
Earth
α
Sun

$$V_{t_0} = H + 5\log_{10}(r\Delta) - 2.5\log_{10}\left[(1-G)\phi_1 + G\phi_2\right]$$

$$\phi_1 = e^{\left[-3.33\left(\left(\tan\frac{\alpha}{2}\right)^{0.63}\right)\right]}$$

$$\phi_2 = e^{\left[-1.87\left(\left(\tan\frac{\alpha}{2}\right)^{1.22}\right)\right]}$$

<

(-& .- - -/ ' ! - < 9 ! 3 %! 6 ? ! & ' @ !
($! ! ! " * ' ' $ ' ! ! @ ' % 6 $"

7' %) 6

($ G 3 @

5

8 (!$)

; 5 ? ! & ' @ ! ($! ! M $"
7' % 6

(UP
/ " Q A

5 ; ;

(-& .- - -/ ' ! - 5 M $"

7' % 6 8

)

(-& .- - -/ ' ! -) 9) & ' @ ! ! " 8 2E
" ? ! & ' @ ! S * ' ' ! % ;6 $"

/@ ? + " BBB

)

!6 (-& .- - -/ ' ! -) 4 " ' 2
! ' % 6 $"

7' % <6 (!$

C * ? ! " 3 $ (

/@ J ! ! & ' >

$! T J ,! !$$: C

@6 (-& .- - -/ ' ! -) (! E =E);)) $"

7' %)6 C & ' >

Theorem 1.3.2 (Angle Bisectors). *In any triangle, the lines which bisect the three interior angles meet in a common point.*

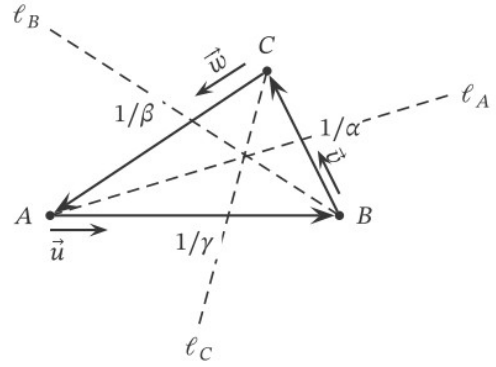

Figure 1.18. Theorem 1.3.2

Suppose the position vectors of the vertices A, B, and C are $\vec{a}$, $\vec{b}$, and $\vec{c}$, respectively.

(a) Show that the *unit* vectors pointing *counterclockwise* along the edges of the triangle (see Figure 1.18) are as follows:

$$\vec{u} = \gamma\vec{b} - \gamma\vec{a}, \quad \vec{v} = \alpha\vec{c} - \alpha\vec{b}, \quad \vec{w} = \beta\vec{a} - \beta\vec{c},$$

where

$$\alpha = \frac{1}{|BC|}, \quad \beta = \frac{1}{|AC|}, \text{ and } \gamma = \frac{1}{|AB|}$$

are the reciprocals of the lengths of the sides (each length is labelled by the Greek analogue of the name of the opposite vertex).

(b) Show that the line ℓ_A bisecting the angle $\angle A$ can be given as

$$\vec{p_A}(r) = (1 - r\beta - r\gamma)\vec{a} + r\gamma\vec{b} + r\beta\vec{c}$$

and the corresponding bisectors of $\angle B$ and $\angle C$ are

$$\vec{p_B}(s) = s\gamma\vec{a} + (1 - s\alpha - s\gamma)\vec{b} + s\alpha\vec{c}$$

$$\vec{p_C}(t) = t\beta\vec{a} + t\alpha\vec{b} + (1 - t\alpha - t\beta)\vec{c}.$$

(c) Show that the intersection of ℓ_A and ℓ_B is given by

$$r = \frac{\alpha}{\alpha\beta + \beta\gamma + \gamma\alpha}, \quad s = \frac{\beta}{\alpha\beta + \beta\gamma + \gamma\alpha}.$$

(d) Show that the intersection of ℓ_B and ℓ_C is given by

$$s = \frac{\beta}{\alpha\beta + \beta\gamma + \gamma\alpha}, \quad t = \frac{\gamma}{\alpha\beta + \beta\gamma + \gamma\alpha}.$$

(e) Conclude that all three lines meet at the point given by

$$\overrightarrow{p_A}\left(\frac{\alpha}{\alpha\beta+\beta\gamma+\gamma\alpha}\right) = \overrightarrow{p_B}\left(\frac{\beta}{\alpha\beta+\beta\gamma+\gamma\alpha}\right) = \overrightarrow{p_C}\left(\frac{\gamma}{\alpha\beta+\beta\gamma+\gamma\alpha}\right) = \frac{1}{\alpha\beta+\beta\gamma+\gamma\alpha}\left(\beta\gamma\vec{a}+\gamma\alpha\vec{b}+\alpha\beta\vec{c}\right).$$

(9) **Barycentric Coordinates:** Show that if $\vec{a}$, $\vec{b}$, and $\vec{c}$ are the position vectors of the vertices of a triangle $\triangle ABC$ in $\mathbb{R}^3$, then the position vector $\vec{p}$ of every point P in that triangle (lying in the plane determined by the vertices) can be expressed as a linear combination of $\vec{a}$, $\vec{b}$ and $\vec{c}$

$$v' = \lambda_1\vec{a} + \lambda_2\vec{b} + \lambda_3\vec{c} \qquad \text{with } 0 \le \lambda_i \le 1 \text{ for } i = 1,2,3 \text{ and } \lambda_1+\lambda_2+\lambda_3 = 1.$$

(*Hint:* (see Figure 1.19) Draw a line from vertex A through P, and observe where it meets the opposite side; call this point D. Use Remark 1.3.1 to show that the position vector $\vec{d}$ of D is a linear combination of $\vec{b}$ and $\vec{c}$, with coefficients between zero and one and summing to 1. Then use Remark 1.3.1 again to show that $\vec{p}$ is a linear combination of $\vec{d}$ and $\vec{a}$.)

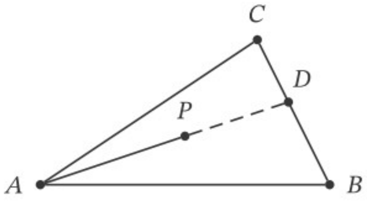

Figure 1.19. Barycentric Coordinates

The numbers λ_i are called the **barycentric coordinates** of P with respect to A, B, and C. Show that P lies on an edge of the triangle precisely if one of its barycentric coordinates is zero.

Barycentric coordinates were introduced (in a slightly different form) by August Möbius (1790-1860) in his book *Barycentrische Calcul* (1827). His name is more commonly associated with “Möbius transformations” in complex analysis and with the “Möbius band” (the one-sided surface that results from joining the ends of a band after making a half-twist) in topology.[12]

(10) Find a line that lies entirely in the set defined by the equation $x^2+y^2-z^2=1$.

History note:

(11) Heath [28, pp. 85-88] points out that the proof of Proposition 4, Book IV of the *Elements* contains the following implicit proof of Theorem 1.3.2 (see Figure 1.20). This was proved by vector methods in Exercise 8.

---

[12] The “Möbius band” was independently formulated by Johann Listing (1808-1882) at about the same time—in 1858, when Möbius was 68 years old. These two are often credited with beginning the study of topology. [32, p. 1165]

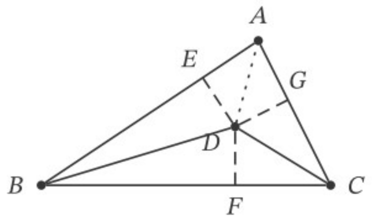

Figure 1.20. Euclid's proof of Theorem 1.3.2

(a) The lines bisecting $\angle B$ and $\angle C$ intersect at a point D above BC because of Book I, Postulate 5 (known as the **Parallel Postulate**):

> That, if a straight line falling on two straight lines make the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which are the angles less than the two right angles.

Why do the interior angles between BC and the two angle bisectors add up to less than a right angle? (*Hint:* What do you know about the angles of a triangle?)

(b) Drop perpendiculars from D to each edge of the triangle, meeting the edges at E, F, and G.
Show that the triangles $\triangle BFD$ and $\triangle BED$ are congruent. (*Hint:* ASA—angle, side, angle!)

(c) Similarly, show that the triangles $\triangle CFD$ and $\triangle CGD$ are congruent.

(d) Use this to show that $|DE| = |DF| = |DG|$.

(e) Now draw the line DA. Show that the triangles $\triangle AGD$ and $\triangle AED$ are congruent. (*Hint:* Both are right triangles; compare one pair of legs and the hypotenuse.)

(f) Conclude that $\angle EAD = \angle GAD$—which means that AD bisects $\angle A$. Thus D lies on all three angle bisectors.

1.4 Projection of Vectors; Dot Products

Suppose a weight is set on a ramp which is inclined θ radians from the horizontal (Figure 1.21). The gravitational force $\vec{g}$ on the weight is directed downward, and some of this is countered by the structure holding up the ramp. The effective force on the weight can be found by expressing $\vec{g}$ as a sum of two (vector) forces, $\overrightarrow{g_{\perp}}$ *perpendicular* to the ramp, and $\overrightarrow{g_{\|}}$ *parallel* to the ramp. Then $\overrightarrow{g_{\perp}}$ is cancelled by the structural forces in the ramp, and the net unopposed force is $\overrightarrow{g_{\|}}$, the *projection* of $\vec{g}$ in the direction of the ramp.

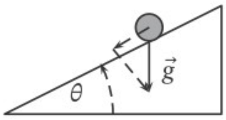

Figure 1.21. A weight on a ramp

)

(-& .- - -/ ' ! -) E ! ! !
" ! E " ! " " ! " ! " ! ! $"

7' % 6 9 '

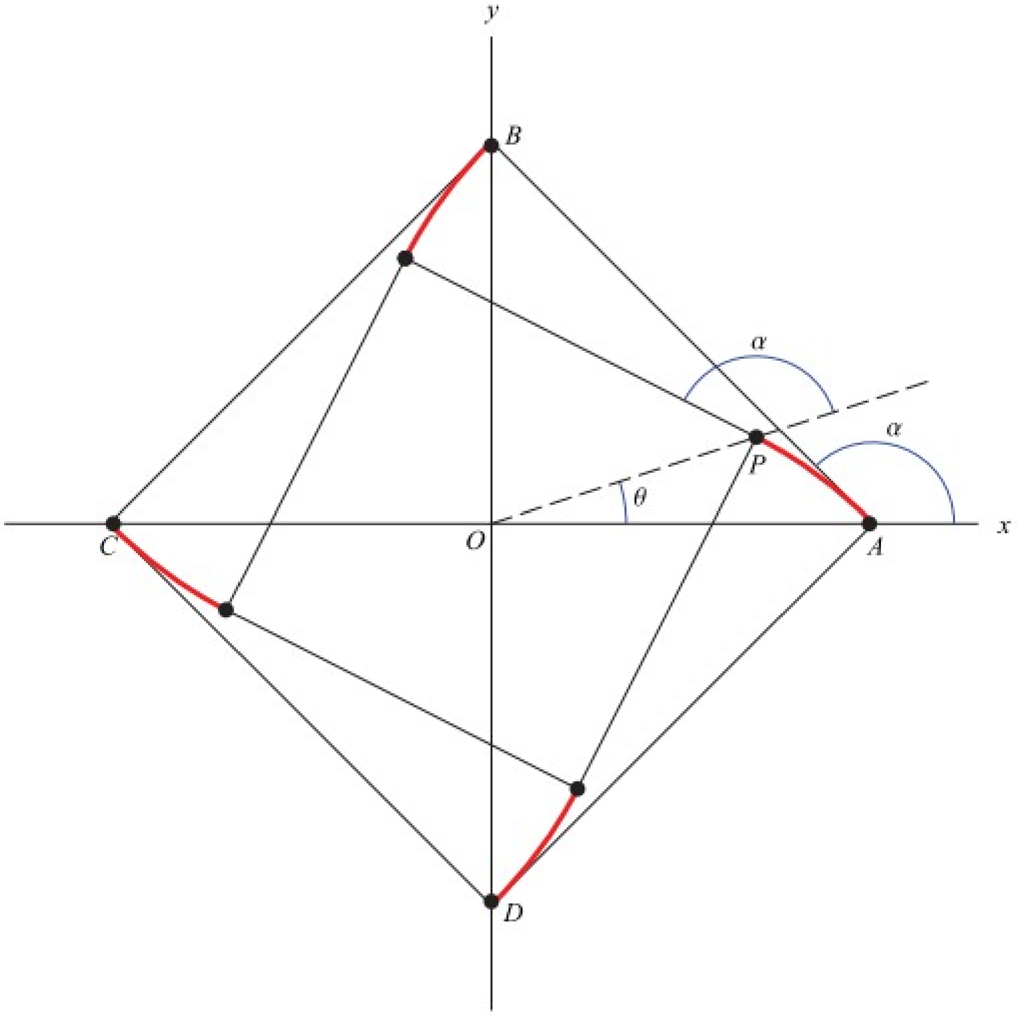
y
B
α
α
P
θ
O
C
A
x
D

9 5

)
;)

) $ + ! ' $ '

!
E3 + ! 1 + E 5 * $ E(+ * $ E(+ $"

7' % <6

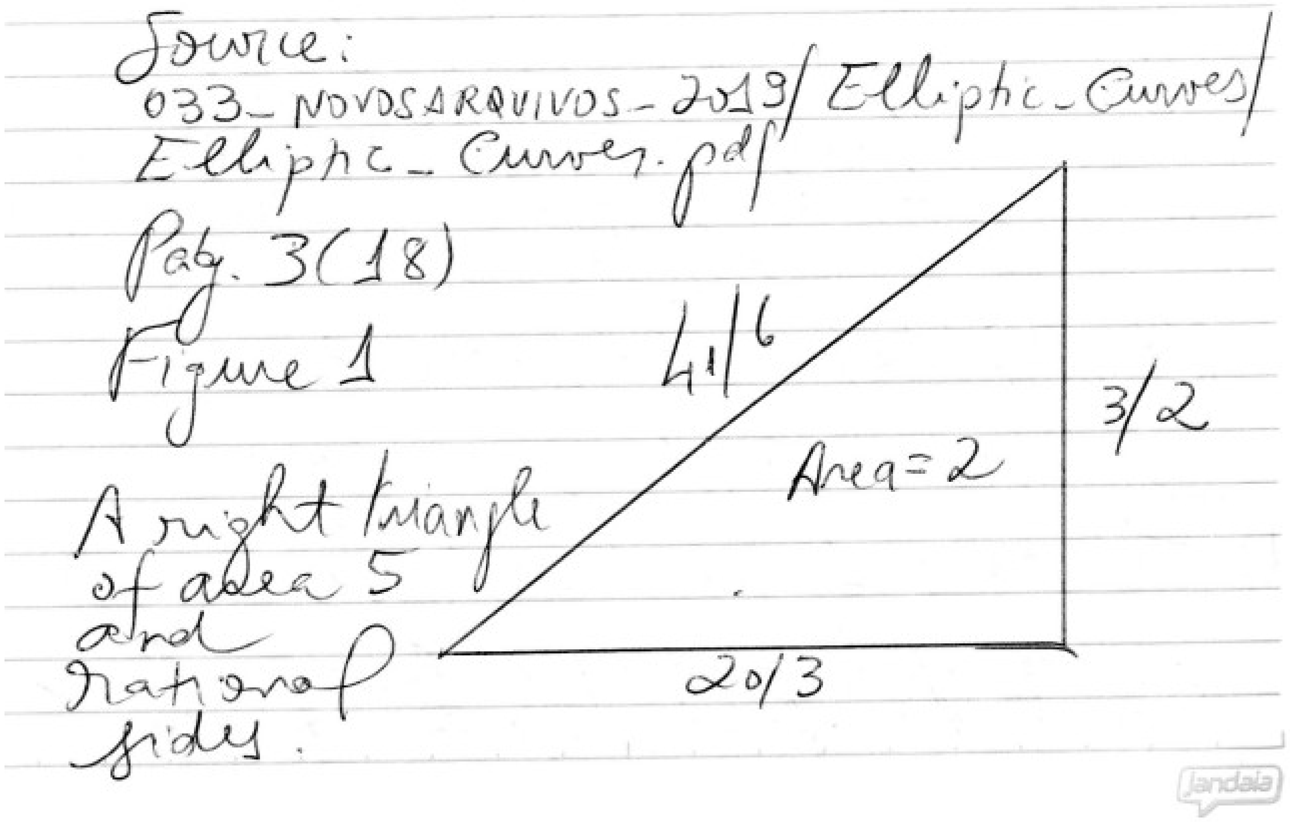
Source:
033_NOVOSARQUIVOS-2019/Elliptic_Curves/Elliptic_Curves.pdf

Pag. 3(18)

Figure 1

A right triangle of area 5 and rational sides.

41/6
3/2
Area=2
20/3

))

(-=&C& 3 + + -> 0 - -& + ! &
8 2-& + ! & 8 2 & ! ! ! ! 2 % 55 6 "O+

(-& .- - -/ ' ! -)) & + ! & 8 2
& ! ! ! ! 2 % 55 6 "O+

7' 5% ;6 % 6

; C ! ' ! ! $! $ O + +

) ^^^

!6 (-& .- - -/ ' ! -) ! " ' $"

7') % 5 6) % 5)6 A " " ! A $ 1

9 '

@6 (-& .- - -/ ' ! -) ! " ' $"

6 (-& .- - -/ ' ! -) ! " ' $"

"6 (-& .- - -/ ' ! -) ! " ' $! $"

6 (-=&C& 3 + + -> 0 - -) 8 ! " 3 &"+! "
! ! ! 2 (! @ "' J + 2 % 55;6-) 8 ! " 3 &"+! " ! !

!   2 (! @   "' J +   2   % 55;6 $"

(-& .-   -   -/ ' ! -)   8 ! " 3   &"+!   " ! !
!   2 (! @   "' J +   2   % 55;6 $"

7')% 56   !" H ;   5 ;<

! "   A

7'   5 %))6   @ _   <)
!
7'   < %) 6   @ _   <

O! 7'   %)<6   ! 7' !   "   ! (A
/@   *   "! ! A ! $7' !

)

!6 (-& .-   -   -/ ' ! -)
`L   #E>   ]E\E   #ER   '! 'aE=   ! E8   2 $"

7'   5% 6   3   !   !   2 ' $!   /   !   ' $!

9 '   C $ O $!   +

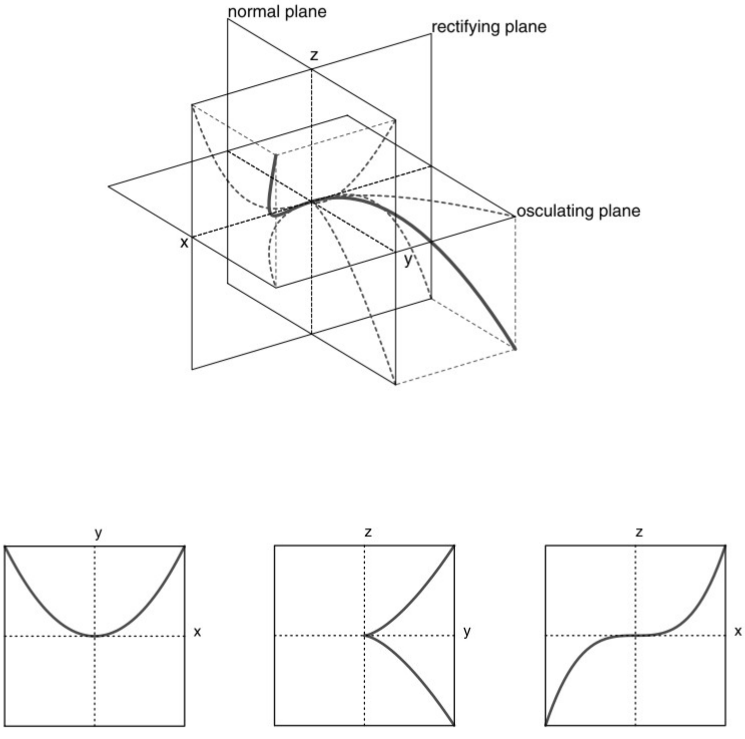
normal plane
rectifying plane
z
osculating plane
x
y
y
x
z
y
z
x

(a) Osculating plane

(b) Normal plane

(c) Rectifying plane

Figure 2.5. Three projections of the space curve $xe_1 + ye_2 + ze_3$

@6 (-& .- - -/ ' ! -) = ! 8 2 (+ ! "
! $ ' ! ! @ ' % 6 $"

);

(-& .- - -/ ' ! -); # C ! &$$ " !
! ' @ ! ! " ! G ! ! 2 $ ' % <6 $"

7' ;;% < 6 8 ! & $! "! " $!

)<

(-& .- - -/ ' ! -)< \ ! ,# + 8 ' = !
1 ! C 2# 1 # ! " $! 8 ! L % ;6 "O+

)5

(-& .- - -/ ' ! -)5
`8 ' E9 E aE= ! E*1 ! E E&$$! $"

/@ ? + " BBB

7' % <6 9 ')

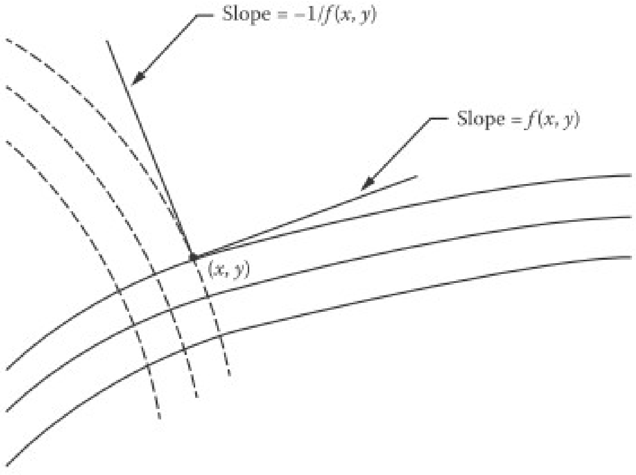
Slope = $-1/f(x, y)$
Slope = $f(x, y)$
(x, y)

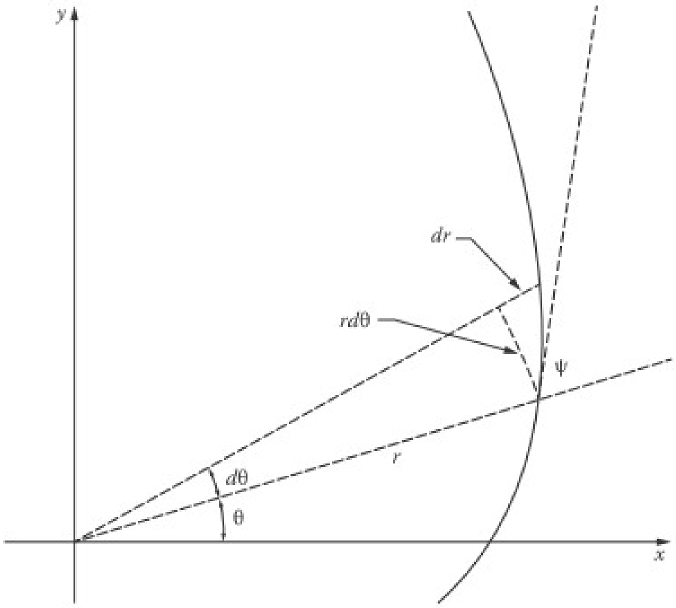
y
dr
rdθ
ψ
r
dθ
θ
x

9 '

(-& .- - -/ ' ! - 4! ! $"

7' % 6 9 '

& ! 2 + " 2 " $

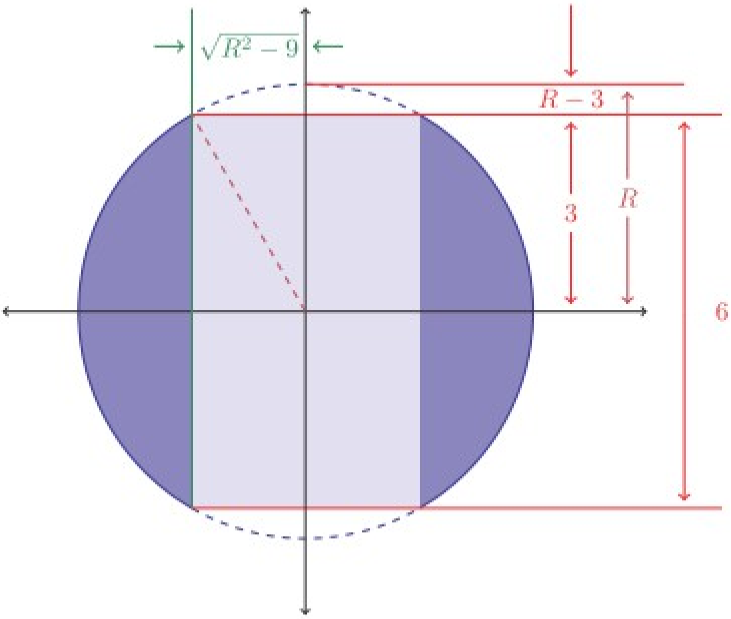
$\sqrt{R^2-9}$
$R-3$
3
R
6

9 ' & ! 2 + " 2 " $
! 2 " ! # @ ! $
!2B

7' % 6

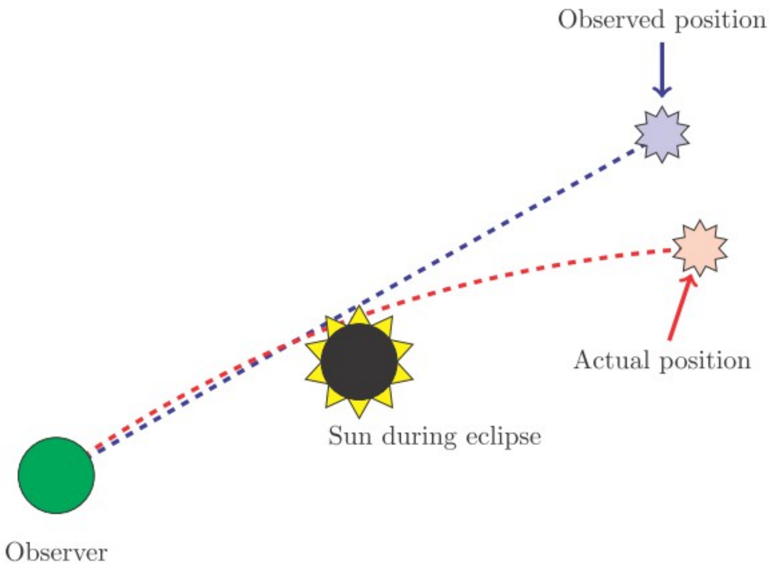
Observed position
Actual position
Sun during eclipse
Observer

9 ' 8 !+ ! ! '

(-& .- - -/ ' ! - 4 ! ! $"

7' % 6

C @! Z ! ! 2 b

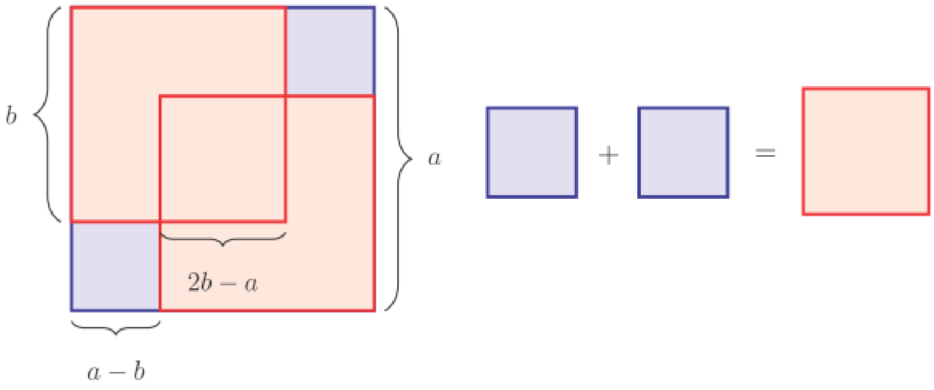
b
a
2b − a
a − b
+
=

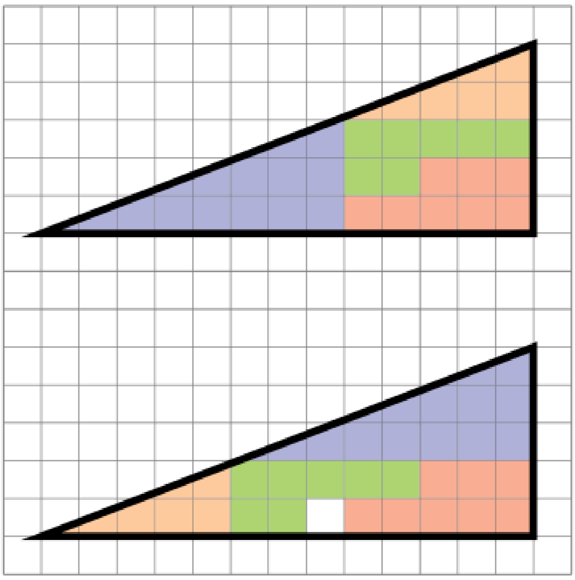

9 ' & !" ' X$ @2 $ Y C ! ' ! '
!+ ! ! !# @ ! ' ! ! '
$! ! ' ! ' 2 " ! ! ! C # H

(-& .- - -/ ' ! - C 8/3/ *C & * 3J * /
(/ ?*M/ = $"

7'

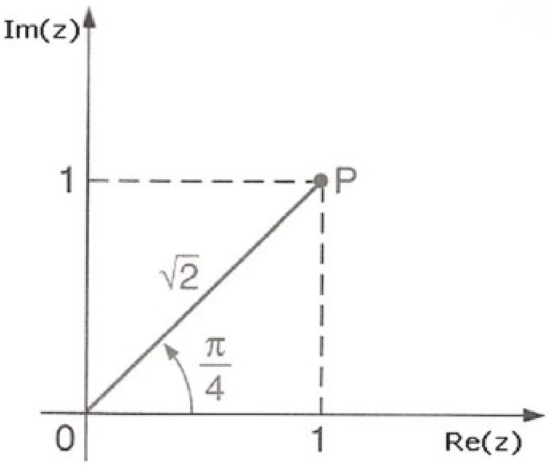
Im(z)
1
P
√2
π/4
0
1
Re(z)

9 ' ! $!UP ' A ! " ,H c

(-& .- - -/ ' ! - (" E C 8 2 3 !
& ' $ ' !' > L " @ ' % 6 $"

7') %; 6

data 31.07.2019

Source: 20190801/Condition* .jpe
Geometry/.../Condition...pdf

Pág. 45(73)

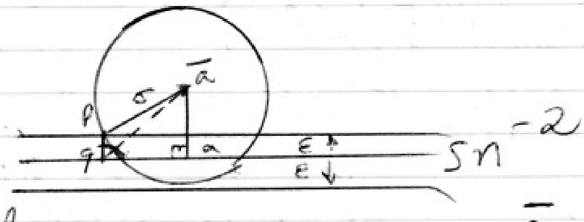
ā
σ
p
q
a
ε
ε
S^{n-2}

Veja ângulo:

ā e q̄

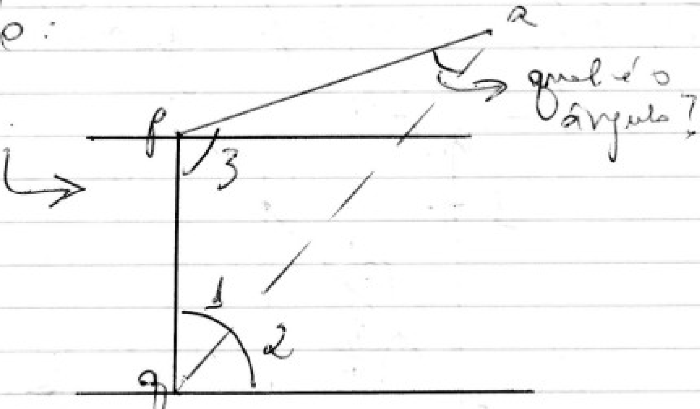
ā
qual é o ângulo?
p
3
1
2
q

— Quais são os ângulos 1 e 2?
Calcular!!!

— Veja qual é o ângulo do ponto P(3)?

— Calcular a área do triângulo.

ā p q

— Faça simulação com a linha de baixo (calcule os ângulos e as áreas)

→

data 31.07.2019
X

Simulação: 2 hipóteses

A)

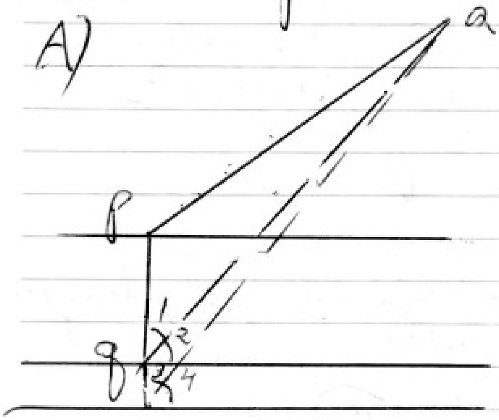

Primeira simulação:

use os pontos p e q.

B)

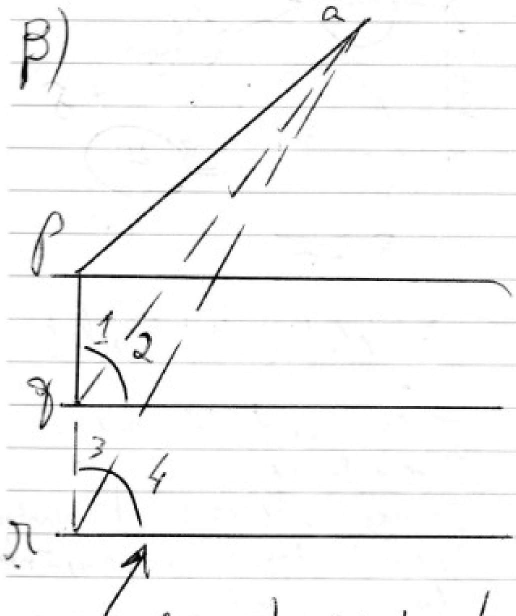

Segunda simulação:

use os pontos p, q e r, sendo que no ponto r a linha reta está tangenciando o círculo cujo centro $\bar{a}$ e posteriormente, com a reta afastada do círculo.

Simule esta linha tangenciando o círculo com centro $\bar{a}$ e com a reta afastada do círculo.

Lemma 2.37 Let $0 < \varepsilon, \sigma < 1$, $n > 2$ and $\bar{a} \in S^{n-1}$. Fix also an embedding $S^{n-2} \hookrightarrow S^{n-1}$.

(veja demonstração depois)

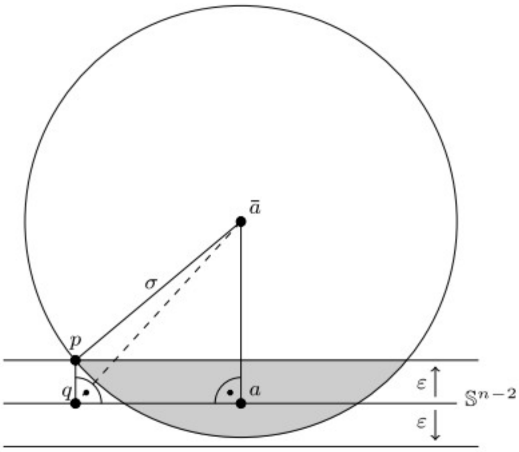
$\bar{a}$
σ
p
q
a
ε
$\mathbb{S}^{n-2}$
ε

Lemma 2.37 *Let* $0 < \varepsilon, \sigma \leq 1$, $n > 2$, *and* $\bar{a} \in \mathbb{S}^{n-1}$. *Fix also an embedding* $\mathbb{S}^{n-2} \hookrightarrow \mathbb{S}^{n-1}$. *Then*

$$\frac{\operatorname{vol}(T(\mathbb{S}^{n-2}, \varepsilon) \cap B(\bar{a}, \sigma))}{\operatorname{vol} B(\bar{a}, \sigma)} \leq 2(n-1)\left(1 + \frac{\varepsilon}{\sigma}\right)^{n-2} \frac{\varepsilon}{\sigma}.$$

)

(-& .- - -/ ' ! -) 5<)< $"

7' ;% 6 9 '

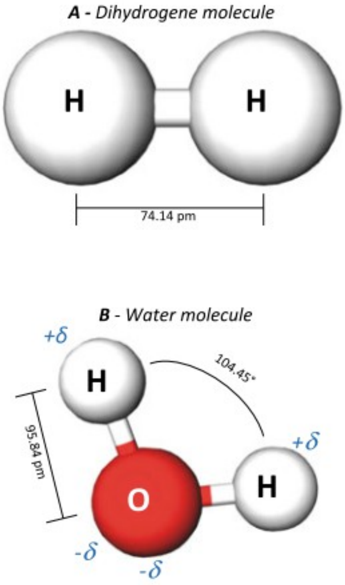
A - Dihydrogene molecule
H
H
74.14 pm
B - Water molecule
+δ
H
104.45°
+δ
95.84 pm
O
H
-δ
-δ

(-& .- - -/ ' ! - " (" ! (2
% '! (2E & C G @ 0 6 % 6 $"

7' %) 6

> "" " = C "2 ! !@ 2

(-& .- - -/ ' ! - 5<)< $"

!' < % 5 6)

= $ > ! !

$$\varepsilon = \varepsilon' - j\varepsilon'' = \varepsilon_0\varepsilon_r - j\frac{\sigma}{\omega}$$

In this complex expression, $j = \sqrt{-1}$ and ω is the angular frequency. The real part $\varepsilon' = \varepsilon_0\varepsilon_r$ accounts for the permittivity of the material (where ε_0 is the permittivity of vacuum and ε_r is the dielectric constant of the material, also designated as the relative permittivity). The permittivity measures the induced polarization of the material by units of electrical field, i.e. the extent to which the bound charge distributions can be distorted and polarized under the influence of the field, and the consequent energy stored within the material. For organic material, these charge distributions are mainly associated with double layer membrane proteins, with the surface of solvated macromolecules or with other polar molecules [129].

7' % <6

; 9 " = !' 9

$$\overrightarrow{F}_d = 6\pi\eta R_p\left(\overrightarrow{v}_f - \overrightarrow{v}_p\right)f_D$$

$$f_D = \left[1 - \frac{9}{16}\left(\frac{R_p}{R_p+z}\right) + \frac{1}{8}\left(\frac{R_p}{R_p+z}\right)^3 - \frac{45}{256}\left(\frac{R_p}{R_p+z}\right)^4 - \frac{1}{16}\left(\frac{R_p}{R_p+z}\right)^5\right]^{-1}$$

data 05.08.2019

Estudar esta equação (Pag. 323)

$$f_D = \left[1 - \frac{9}{16}\left(\frac{R_p}{R_p+z}\right) + \frac{1}{8}\left(\frac{R_p}{R_p+z}\right)^{\textcircled{3}} - \frac{45}{256}\left(\frac{R_p}{R_p+z}\right)^{\textcircled{4}} - \frac{1}{16}\left(\frac{R_p}{R_p+z}\right)^{\textcircled{5}}\right]^{\textcircled{-1}} \quad (7.26$$

estudar os motivos que estabeleceram essas razões.

analisar as suas relações.

Obs.:

R_p is the radius of the particle,

$\vec{V}_p$ its the velocity,

n and $\vec{V}_f$ respectively the viscosity of the external medium and the velocity of the fluid

f_D is the drag (arrasto) coefficient of the particle that takes into account the influence of a solid wall in the vicinity of the particle,

z being the distance of the particle to the wall

Obs.: f_D varies between 1, for particle distant from the wall to 3 for

05.08.2019

particle in contact with micro-channel wall (z=0).

57) Conic Sections

Source: 20190805_02/Important/Linear Algebra through Geometry - Springer...&.pdf

Pág. 95(105)

Fig. 2.52

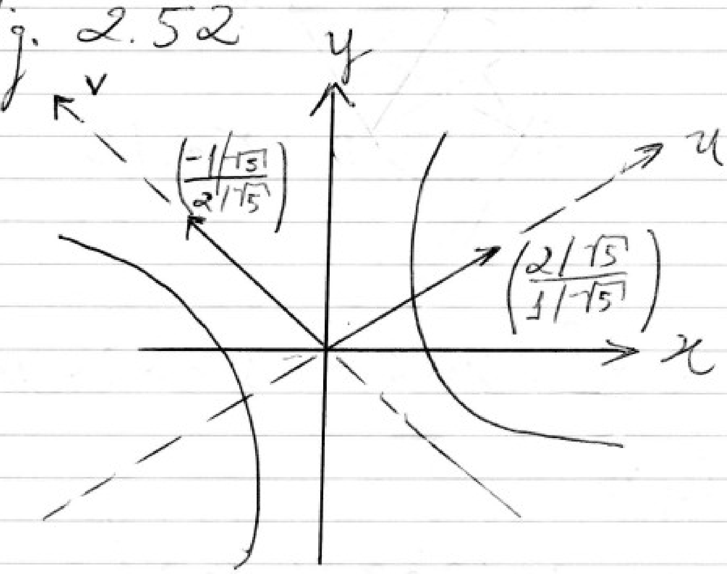

v
y
u
$\begin{pmatrix} -1/\sqrt{5} \\ 2/\sqrt{5} \end{pmatrix}$
$\begin{pmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \end{pmatrix}$
x

58) 20190806

Traité de géométrie précédé...

Source: 20190806_03/Courbes/Traité_de_géométrie_précédé...bpt6K920521x.pdf.

Pág. 52(63) →

;

(-& .- - -/ ' ! - ; ? ! & ' @ ! C ' 8 2
$ ' !' 3 4 0 % 55 6 $"

7' 5<% 6 9 '

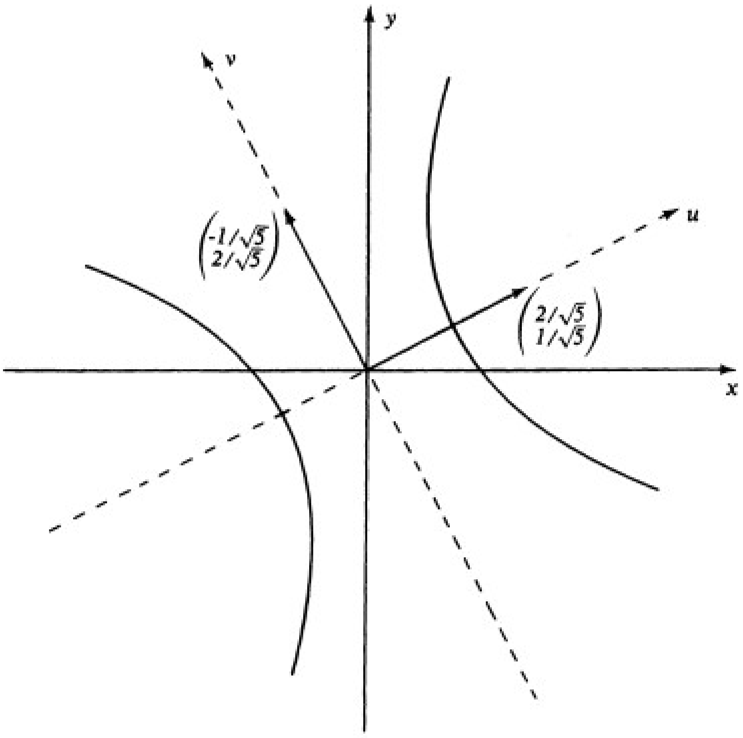
y
v
u
x
$\begin{pmatrix}-1/\sqrt{5}\\ 2/\sqrt{5}\end{pmatrix}$
$\begin{pmatrix}2/\sqrt{5}\\ 1/\sqrt{5}\end{pmatrix}$

+ ' 7' 5 % 6

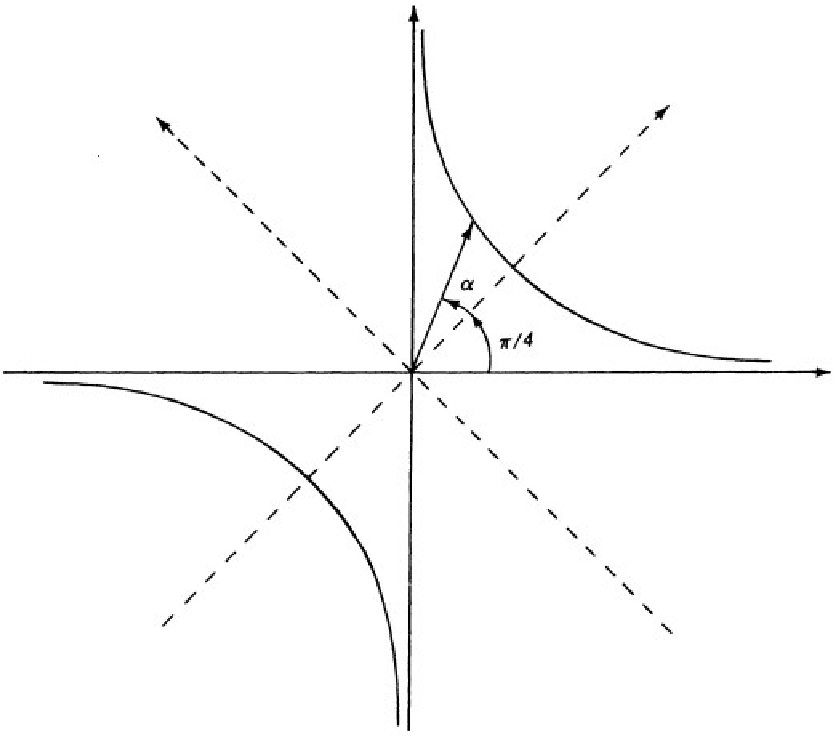

<

(-& .- - -/ ' ! - <
C ! AE" E'A A E! ! 2 1 E$ A A"AE` a> E& E@$ 05 G $"

7' % 6 " + ! ": ! '

Fig. 15.

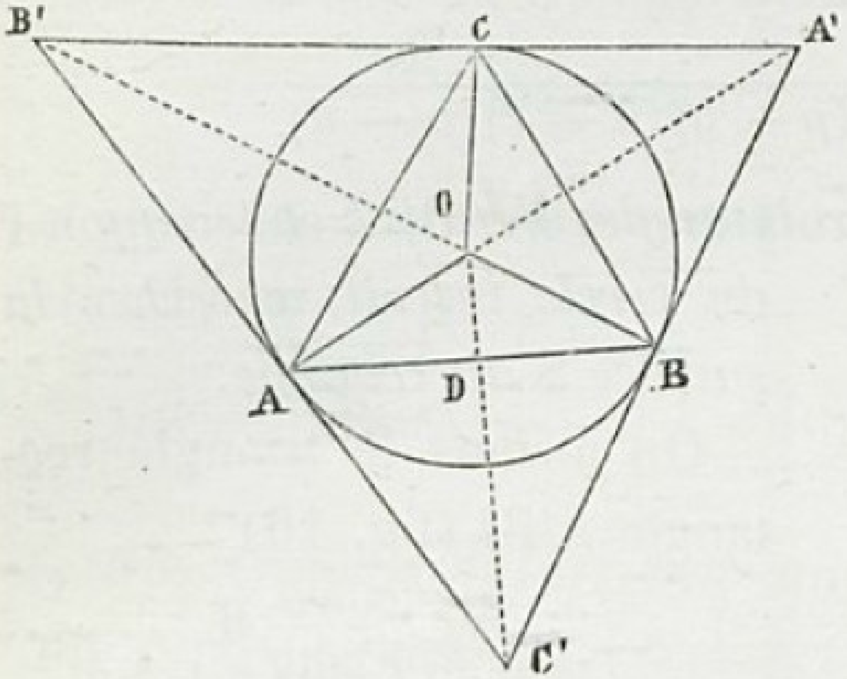

En multipliant par ordre (1) et (2), il viendra :

$$ab:(a+x)(b+x)=\sin\alpha\sin\gamma:\sin(\alpha+\beta)\sin(\beta+\gamma) \quad (3);$$

d'où $$x^2+(a+b)x+ab=\frac{ab\sin(\alpha+\beta)\sin(\beta+\gamma)}{\sin\alpha\sin\gamma}$$

et $$x=-\frac{1}{2}(a+b)\pm\frac{1}{2}\sqrt{(a-b)^2+\frac{4ab\sin(\alpha+\beta)\sin(\beta+\gamma)}{\sin\alpha\sin\gamma}}.$$

En posant $$\frac{4ab\sin(\alpha+\beta)\sin(\beta+\gamma)}{\sin\alpha\sin\gamma}=(a-b)^2\operatorname{tg}^2\omega' \quad (4),$$

on aura : $$x=-\frac{1}{2}(a+b)\pm\frac{1}{2}(a-b)\sqrt{1+\operatorname{tg}^2\omega'}$$

et enfin $$x=-\frac{a+b}{2}\pm\frac{(a-b)}{2\cos\omega'}.$$

L'angle auxiliaire ω' est déterminé par l'équation (4). On ne prendra des deux valeurs de x que celle qui est positive. Si l'une des deux autres distances a et b, a par exemple, était inconnue, on la regarderait comme telle et les deux autres x et b comme étant connues ; comme il est facile de le voir, l'équation serait du 1er degré.

Surface d'un triangle.

51. Soit ABC un triangle quelconque (fig. 8).

On sait, par la Géométrie, que la surface S de ce triangle est $S=\frac{AB\times CD}{2}$, CD étant la hauteur du triangle. Le triangle rectangle ACD donne :

$$CD=\frac{b\sin A}{R}; \quad \text{d'où} \quad S=\frac{bc\sin A}{2R}.$$

Connaissant les trois angles A, B, C et le rayon R du cercle circonscrit, on a aussi pour la surface S de ce triangle :

$$S=\overline{2R}^2\sin A\sin B\sin C.$$

Fig. 15.

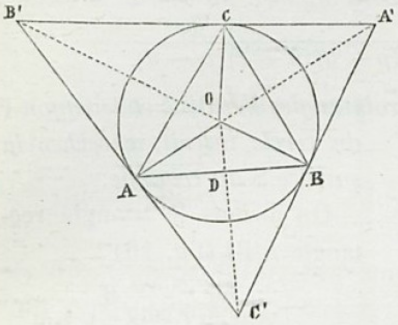

En effet, soit O le centre du cercle circonscrit; désignons par R le rayon du cercle circonscrit, il viendra :

$$AB = c = 2R \sin C,$$

et pour la perpendiculaire abaissée du centre sur AB: $OD = R \cos C$; de sorte que la surface du triangle $AOB = \frac{R^2}{2} \sin 2C$; celle de $AOC = \frac{R^2}{2} \sin 2B$, et celle de $BOC = R^2 \sin A \cos A$. La surface totale

$$ABC = \frac{R^2}{2}(\sin 2A + \sin 2B + \sin 2C).$$

Mais on a vu (15) que

$$\sin 2A + \sin 2B + \sin 2C = 4 \sin A \sin B \sin C;$$

la surface cherchée est donc :

$$S = 2R^2 \sin A \sin B \sin C.$$

52. *Connaissant les quatre côtés* a, b, c, d *d'un quadrilatère inscrit* ABCD, *trouver la surface* Q *de ce quadrilatère.*

On a : $$Q = \left(\frac{ad + bc}{2}\right) \sin A,$$

a, b, c, d étant les côtés AB, BC, CD, AD.

Les deux triangles ABD, BCD donnent pour le côté BD:

$$\overline{BD}^2 = a^2 + d^2 - 2ad \cos A = b^2 + c^2 + 2bc \cos A;$$

d'où $$\cos A = \frac{a^2 + d^2 - b^2 - c^2}{2(ad + bc)}$$

et $$\sin A = \sqrt{1 - \frac{(a^2 + d^2 - b^2 - c^2)^2}{4(ad + bc)^2}}.$$

En remplaçant cette valeur de sin A dans l'expression de Q, il vient :

$$Q = \frac{ad + bc}{2} \sqrt{1 - \frac{(a^2 + d^2 - b^2 - c^2)^2}{4(ad + bc)^2}}$$

et $Q = \frac{1}{4}\sqrt{(b+c+d-a)(a+c+d-b)(a+b+d-c)(a+b+c-d)}$;

en désignant par $2p$ le périmètre du quadrilatère inscrit, on a :

$$Q = \sqrt{(p-a)(p-b)(p-c)(p-d)}.$$

53. *Connaissant les trois angles* A, B, C *et le rayon* r *du cercle inscrit, chercher la surface* S *du triangle.*

Fig. 16.

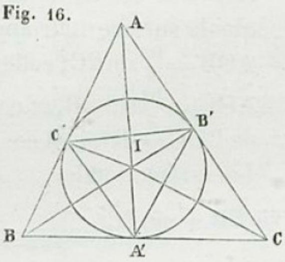

On a dans le triangle rectangle A'BI (fig. 16) :

$$A'B = r \operatorname{cotg}\frac{B}{2},$$

$$A'C = r \operatorname{cotg}\frac{C}{2};$$

le côté $BC = a = r\left(\operatorname{cotg}\frac{B}{2} + \operatorname{cotg}\frac{C}{2}\right) = r\frac{\cos\frac{A}{2}}{\sin\frac{B}{2}\sin\frac{C}{2}}.$

La surface du triangle BIC est donc :

$$BIC = \frac{r^2}{2}\left(\operatorname{cotg}\frac{B}{2} + \operatorname{cotg}\frac{C}{2}\right);$$

celle du triangle total ABC est :

$$S = r^2\left(\operatorname{cotg}\frac{A}{2} + \operatorname{cotg}\frac{B}{2} + \operatorname{cotg}\frac{C}{2}\right).$$

Mais on a vu que

$$\operatorname{cotg}\frac{A}{2} + \operatorname{cotg}\frac{B}{2} + \operatorname{cotg}\frac{C}{2} = \operatorname{cotg}\frac{A}{2}\operatorname{cotg}\frac{B}{2}\operatorname{cotg}\frac{C}{2};$$

d'où $$S = r^2 \operatorname{cotg}\frac{A}{2}\operatorname{cotg}\frac{B}{2}\operatorname{cotg}\frac{C}{2}.$$

54. *On connait les trois angles d'un triangle et le rayon* r *du cercle inscrit. Chercher l'aire du triangle qui joint les trois points de contact.*

Soient A', B', C' les trois points de contact et I le centre du cercle inscrit (fig. 16). L'angle au centre B'IC' est le supplément de l'angle A du triangle; de sorte que l'on a pour la surface du triangle B'IC' :

$$\text{surface } B'IC' = \frac{r^2}{2} \sin A,$$

et ainsi des autres. En ajoutant, on obtient pour la surface cherchée S :

$$S = \frac{r^2}{2}(\sin A + \sin B + \sin C) = 2r^2 \cos\frac{A}{2}\cos\frac{B}{2}\cos\frac{C}{2}.$$

55. *Connaissant les trois angles* A, B, C *et le rayon* R *du cercle circonscrit, trouver la surface* S *du triangle* A'B'C', *compris par les tangentes aux sommets du triangle* (fig. 15).

Le triangle rectangle A'BO donne :

$$A'B = R \operatorname{tg} A.$$

On a de même dans le triangle rectangle C'BO :

$$C'B = R \operatorname{tg} C;$$

de sorte que le côté $A'C' = R(\operatorname{tg} A + \operatorname{tg} C)$,

$$A'C' = \frac{R \sin B}{\cos A \cos C}.$$

L'aire du triangle $A'OC' = \frac{R^2}{2}(\operatorname{tg} A + \operatorname{tg} C)$.

La surface du triangle A'B'C' est :

$$S = R^2(\operatorname{tg} A + \operatorname{tg} B + \operatorname{tg} C),$$

$$S = R^2 \operatorname{tg} A \operatorname{tg} B \operatorname{tg} C,$$

puisque $\operatorname{tg} A + \operatorname{tg} B + \operatorname{tg} C = \operatorname{tg} A \operatorname{tg} B \operatorname{tg} C$.

56. Soient α, β, γ les rayons des trois cercles ex-inscrits au triangle ABC; R, r les rayons des cercles circonscrit et inscrit, a, b, c les trois côtés et $2p$ le périmètre du triangle, I, I' étant les centres des cercles inscrit et ex-inscrit au côté $BC = a$; on aura (fig. 17) :

$$BH = \alpha \operatorname{tg}\frac{B}{2}$$

Fig. 17.

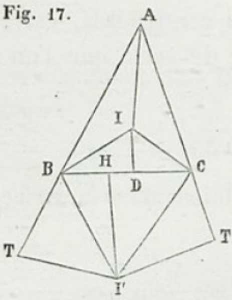

et $$CH = \alpha \operatorname{tg}\frac{C}{2};$$

$$a = \alpha\left(\operatorname{tg}\frac{B}{2} + \operatorname{tg}\frac{C}{2}\right).$$

En remplaçant $\operatorname{tg}\frac{B}{2}$ et $\operatorname{tg}\frac{C}{2}$ par leurs valeurs en fonction des côtés a, b, c du triangle, il viendra :

$$a = \alpha\left[\sqrt{\frac{(p-a)(p-c)}{p(p-b)}} + \sqrt{\frac{(p-a)(p-b)}{p(p-c)}}\right],$$

$$a = \alpha\left[\frac{(p-a)(p-c+p-b)}{\sqrt{p(p-a)(p-b)(p-c)}}\right] = \frac{\alpha(p-a)a}{\sqrt{p(p-a)(p-b)(p-c)}};$$

d'où $$\sqrt{p(p-a)(p-b)(p-c)} = \alpha(p-a);$$

donc, $$S = \alpha(p-a).$$

On aura de même :

$$S = \beta(p-b), \quad S = \gamma(p-c);$$

d'ailleurs, $$S = pr.$$

En ajoutant les valeurs inverses de α, β, γ, on obtiendra :

$$\frac{1}{\alpha} + \frac{1}{\beta} + \frac{1}{\gamma} = \frac{p-a+p-b+p-c}{S} = \frac{p}{S} = \frac{1}{r},$$

et, en multipliant :

$$S^2 = r\alpha\beta\gamma; \quad \text{d'où} \quad S = \sqrt{r\alpha\beta\gamma}.$$

En ajoutant les valeurs de α, β, γ et en retranchant de cette somme le rayon r, il vient :

$$\alpha + \beta + \gamma - r = \frac{abc}{S} \quad \text{ou} \quad \alpha + \beta + \gamma - r = 4R,$$

puisque l'on a :

$$abc = 4RS.$$

Ainsi, *la somme des rayons des cercles ex-inscrits est* *égale à quatre fois le rayon du cercle circonscrit plus le rayon du cercle inscrit.*

57. *Connaissant la hauteur* h, *la base* a *et l'angle* A *opposé à celle-ci, résoudre le triangle* (fig. 18).

Fig. 18.

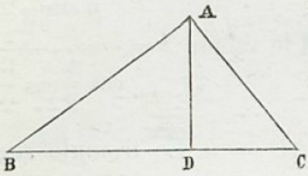

Représentons DC par y et l'angle DAC par x, on aura :

$$\mathrm{BD} = a - y.$$

Les triangles rectangles BDA, ADC donnent :

$$y = h \operatorname{tg} x \quad (1), \quad a - y = h \operatorname{tg}(\mathrm{A} - x) \quad (2).$$

On tire de (1) :

$$\operatorname{tg} x = \frac{y}{h}; \quad \text{d'où} \quad \frac{a-y}{h} = \frac{\operatorname{tg}\mathrm{A} - \frac{y}{h}}{1 + \frac{y}{h}\operatorname{tg}\mathrm{A}}.$$

En résolvant cette équation, il vient :

$$y = \frac{a}{2} \pm \sqrt{\frac{a^2}{4} + h^2 - \frac{ah}{\operatorname{tg}\mathrm{A}}}.$$

Cherchons un carré $v^2 = \frac{\rho\, ah}{\operatorname{tg}\mathrm{A}}$, ρ étant le rayon des tables. On aura :

$$y = \frac{a}{2} \pm \sqrt{\frac{a^2}{4} + h^2 - v^2}.$$

On procédera de la même manière pour connaitre les autres côtés dans les triangles rectangles.

58. *Connaissant les trois angles et le périmètre* 2p, *résoudre le triangle et trouver sa surface.*

On a :
$$\frac{\sin\mathrm{A}}{a} = \frac{\sin\mathrm{B}}{b} = \frac{\sin\mathrm{C}}{c}$$

et
$$\frac{\sin\mathrm{A} + \sin\mathrm{B} + \sin\mathrm{C}}{2p} = \frac{\sin\mathrm{A}}{a}.$$

Mais on sait que

$$\sin A + \sin B + \sin C = 4 \cos \frac{A}{2} \cos \frac{B}{2} \cos \frac{C}{2}.$$

La proportion

$$\frac{\sin A}{\sin B} = \frac{a}{b} = \frac{p \sin \frac{A}{2}}{b \cos \frac{B}{2} \cos \frac{C}{2}} \text{ fournit la valeur } b = \frac{p \sin \frac{B}{2}}{\cos \frac{A}{2} \cos \frac{C}{2}}.$$

Menons la perpendiculaire AD $= h$. Le triangle rectangle ADC donne

$$h = b \sin C; \quad \text{d'où} \quad h = \frac{2p \sin \frac{B}{2} \sin \frac{C}{2}}{\cos \frac{A}{2}}.$$

La surface ABC $= \frac{bh}{2}$. On aura donc :

$$\text{surf. ABC} = \frac{p^2 \sin \frac{A}{2} \sin \frac{B}{2} \sin \frac{C}{2}}{\cos \frac{A}{2} \cos \frac{B}{2} \cos \frac{C}{2}} \text{ ou surf. ABC} = p^2 \operatorname{tg} \frac{A}{2} \operatorname{tg} \frac{B}{2} \operatorname{tg} \frac{C}{2}.$$

En rendant cette formule homogène, elle devient :

$$\text{surface ABC} = \frac{p^2 \operatorname{tg} \frac{A}{2} \operatorname{tg} \frac{B}{2} \operatorname{tg} \frac{C}{2}}{\rho^3}.$$

59. *Résoudre le triangle et trouver sa surface, connaissant les trois angles et la bissectrice* δ *de l'un d'eux* (fig. 19).

Fig. 19.

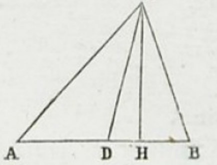

On aura :

$$\frac{\sin A}{\sin CDA} = \frac{\delta}{b};$$

mais

$$\sin CDA = \sin \left(A + \frac{C}{2}\right):$$

donc
$$\frac{\sin A}{\sin\left(A+\frac{C}{2}\right)}=\frac{\delta}{b}.$$

On tire de cette expression :
$$b=\frac{\delta\sin\left(A+\frac{C}{2}\right)}{\sin A}.$$

La proportion $\frac{\sin C}{\sin B}=\frac{c}{b}$ donne :
$$c=\frac{\delta\sin C\sin\left(A+\frac{C}{2}\right)}{\sin A\sin B}.$$

On sait que l'on a :
$$h=\delta\sin\left(A+\frac{C}{2}\right)\quad\text{et}\quad \text{surf. ABC}=\frac{ch}{2}:$$
$$\text{surface ABC}=\frac{\delta^2\sin^2\left(A+\frac{C}{2}\right)\sin C}{2\sin A\sin B}.$$

60. *Connaissant l'aire* t, *la hauteur* h *et un angle* A *à la base, résoudre le triangle.*

On a : $\frac{bh}{2}=t$; d'où $b=\frac{2t}{h}$.

De l'expression $h=c\sin A$, on tire :
$$c=\frac{h}{\sin A}.$$

Les côtés b et c sont donc connus. Il vient :
$$A+B+C=180°;\quad\text{d'où}\quad B+C=180°-A$$
et
$$\frac{1}{2}(B+C)=90°-\frac{A}{2}.$$

On connaîtra $\operatorname{tg}\frac{1}{2}(B-C)$ par la proportion
$$\frac{b+c}{b-c}=\frac{\operatorname{cotg}\frac{1}{2}A}{\operatorname{tg}\frac{1}{2}(B-C)};$$
par suite, on aura la valeur de $\frac{1}{2}(B-C)$.

On a déjà obtenu $\frac{1}{2}(B+C)$; connaissant la demi-somme et la demi-différence de deux angles, on sait trouver chacun d'eux. La valeur de B et de C étant déterminée, on cherchera a au moyen de

$$\frac{\sin A}{\sin B} = \frac{a}{b}.$$

61. *Résoudre le triangle connaissant deux côtés* a *et* b *et la différence* 2d *des angles opposés.*

On donne a et b, $A - B = 2d$;

d'où $\frac{1}{2}(A - B) = d$.

Il vient : $$\frac{a+b}{a-b} = \frac{\operatorname{cotg} \frac{1}{2} C}{\operatorname{tg} \frac{1}{2}(A-B)} = \frac{\operatorname{cotg} \frac{C}{2}}{\operatorname{tg} d};$$

d'où $$\operatorname{cotg} \frac{C}{2} = \frac{(a+b) \operatorname{tg} d}{a-b} = \operatorname{tg} \tfrac{1}{2}(A+B).$$

On connaît ainsi $\frac{1}{2}(A+B)$. D'un autre côté,

$$\tfrac{1}{2}(A - B) = d.$$

Au moyen de la demi-somme et de la demi-différence de deux angles, on peut déterminer chacun d'eux. On trouvera la valeur de c par la formule $\frac{\sin A}{\sin C} = \frac{a}{c}$, sin A, sin C, a étant connus.

62. *Connaissant un angle* A, *le côté opposé* a *et le rectangle* K^2 *des deux autres côtés, résoudre le triangle.*

On a les équations :

$$bc = K^2, \quad a^2 = b^2 + c^2 - 2bc \cos A;$$

d'où $$b^2 + c^2 = a^2 + 2K^2 \cos A.$$

En complétant le carré, il vient :

$$b + c = \sqrt{a^2 + 4K^2 \cos^2 \frac{A}{2}}.$$

On a de même :

$$b - c = \sqrt{a^2 - 4K^2 \sin^2 \frac{A}{2}}.$$

On fera $\frac{K^2 \cos^2 \frac{A}{2}}{\rho^2} = u^2, \qquad \frac{K^2 \sin^2 \frac{A}{2}}{\rho^2} = v^2 :$

les côtés b et c seront déterminés.

63. *Résoudre le triangle connaissant l'angle* B, *le côté* b *et la somme* a + c = S *des deux autres côtés* (fig. 20).

Fig. 20.

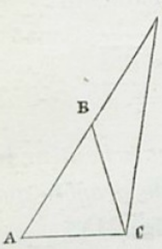

Prolongeons le côté c d'une quantité égale à a. On aura $AI = S$, $AC = b$. Puisque $IBC = 180° - B$, $BIC + ICB = B$. Le triangle IBC étant isoscèle donne :

$$ICB = BIC = \frac{B}{2}.$$

Dans AIC on connaît b, S et l'angle I;

il vient : $\sin \frac{B}{2} : b = \sin \left(C + \frac{B}{2}\right) : S;$

d'où $\sin \left(C + \frac{B}{2}\right) = \frac{S \sin \frac{B}{2}}{b}.$

Dans cette équation, il n'y a d'autre inconnue que l'angle C, la valeur de $\frac{B}{2}$ permettant de trouver B.

64. *Résoudre le triangle connaissant la base* a, *la hauteur* h *et la différence* d *des angles à la base.*

On a $B = d + C.$

Il vient (fig. 18)

$$BD = h \operatorname{cotg} B, \quad DC = h \operatorname{cotg} C$$

et $\frac{a}{h} = \operatorname{cotg} B + \operatorname{cotg} C.$

Or,
$$\text{cotg B} = \text{cotg}\,(d + \text{C});$$

d'où
$$\frac{a}{h} = \frac{\text{cotg}\,d\ \text{cotg C} - 1}{\text{cotg}\,d + \text{cotg C}} + \text{cotg C} \quad . \quad . \quad . \quad (1),$$

équation qui donne cotg C et par suite C.

En représentant BD par x, DC est égal à $a - x$. Les valeurs de x et de $a - x$ fournissent :

$$\text{cotg B} = \frac{x}{h}, \quad \text{cotg C} = \frac{a - x}{h}.$$

Remplaçant cotg B et cotg C dans (1), on obtient :

$$x = \frac{a \sin d + 2h \cos d \pm \sqrt{4h^2 + a^2 \sin^2 d}}{2 \sin d}.$$

65. *Connaissant les trois angles* A, B, C *et la hauteur* h, *trouver l'aire et le périmètre tant du triangle proposé que de celui qui joint les pieds de ses hauteurs* (fig. 21).

Dans le triangle rectangle ABF, on a :

$$\text{AF} = h\ \text{cotg A}.$$

Le triangle rectangle BFC donne :

$$\text{FC} = h\ \text{cotg C},$$

$$b = h(\text{cotg A} + \text{cotg C}) = \frac{h \sin \text{B}}{\sin \text{A} \sin \text{C}}.$$

Fig. 21.

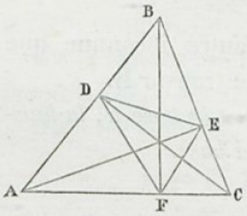

La hauteur h étant donnée, il vient :

$$\text{surf. ABC} = \frac{h^2 \sin \text{B}}{2 \sin \text{A} \sin \text{C}}.$$

On connait les trois angles, ainsi que b ; par conséquent,

$$\frac{\sin \text{A} + \sin \text{B} + \sin \text{C}}{2p} = \frac{\sin \text{B}}{b} = \frac{\sin \text{A} \sin \text{C}}{h};$$

d'où
$$2p = \frac{h \cos \frac{B}{2}}{\sin \frac{A}{2} \sin \frac{C}{2}}.$$

Passons au triangle qui joint les pieds des trois hauteurs.

BCD donne $h' = a \sin B = \frac{h \sin B}{\sin C}$;

de même, $h'' = \frac{h \sin B}{\sin A}$. Il vient : $AF = h \operatorname{cotg} A$.

Triangle ADF fournit :

$$DF = \frac{h \cos A}{\sin C}, \quad DE = \frac{h \sin B \cos B}{\sin C \sin A}, \quad EF = \frac{h \cos C}{\sin A}.$$

Le périmètre de DEF est :

$$2p = \frac{h \cos A}{\sin C} + \frac{h \cos C}{\sin A} + \frac{h \sin B \cos B}{\sin C \sin A},$$

$$2p = h \left[\frac{2 \sin A \cos A + 2 \sin C \cos C + 2 \sin B \cos B}{2 \sin A \sin C}\right],$$

$$2p = \frac{h}{2} \left[\frac{\sin 2A + \sin 2B + \sin 2C}{\sin A \sin C}\right] = 2h \sin B.$$

Connaissant les trois côtés, on déterminera les trois angles. Il sera facile de trouver la hauteur de DEF et, par suite, sa surface S. On trouve :

$$S = h^2 \operatorname{cotg} A \operatorname{cotg} C \cos B \sin B.$$

66. *Résoudre le triangle connaissant la hauteur* h, *le rayon du cercle circonscrit* R *et celui du cercle inscrit* r.

a, b, c étant les trois côtés du triangle cherché, on a :

$$pr = \sqrt{p(p-a)(p-b)(p-c)};$$

d'où
$$p^2r^2 = p(p-a)(p-b)(p-c),$$

le périmètre $2p$ étant égal à $a + b + c$.

Divisant par p^2, on obtient :

$$r^2 = -p^2 + a(2p - a) + 2Rh - 4Rr.$$

En remplaçant p et p^2 par leurs valeurs $\frac{ah}{2r}$ et $\frac{a^2h^2}{4r^2}$, et en substituant $2Rh$ au lieu de bc, il vient :

$$a^2(h - 2r)^2 = 4r^2(2Rh - r^2 - 4Rr);$$

d'où
$$a = \frac{2r}{h - 2r}\sqrt{2Rh - r(r + 4R)}.$$

67. *Résoudre le triangle dans lequel on donne : l'angle* A, *la hauteur* h *menée du sommet de cet angle et l'excès* d *de la somme des côtés de cet angle sur le troisième. Chercher la relation qui doit exister entre les données de la question pour que le triangle soit isoscèle.*

A, B, C étant les trois angles de ce triangle et x, y, z les côtés opposés à ces angles, on a les trois équations :

$$y + z = d + x \quad \ldots\ldots \quad (1),$$

$$yz = \frac{hx}{\sin A} \quad \ldots\ldots \quad (2),$$

$$y^2 + z^2 - 2yz\cos A = x^2 \quad \ldots\ldots \quad (3).$$

En ajoutant les deux dernières, après avoir multiplié l'équation (2) par 2, on obtient :

$$(y + z)^2 = x^2 + 2hx\,\mathrm{cotg}\,\frac{A}{2}$$

et, en ayant égard à l'équation (1),

$$x = \frac{d^2}{2\left(h\,\mathrm{cotg}\,\frac{A}{2} - d\right)},$$

valeur que l'on peut rendre calculable par logarithmes, au moyen d'un angle auxiliaire.

De ces équations, on tire :

$$y+z=\frac{d\left[2h\operatorname{cotg}\frac{A}{2}-d\right]}{2\left[h\operatorname{cotg}\frac{A}{2}-d\right]},$$

$$yz=\frac{hd^2}{2\sin A\left(h\operatorname{cotg}\frac{A}{2}-d\right)}.$$

On aura donc pour les valeurs de y et de z, l'équation

$$u^2-\frac{d\left(2h\operatorname{cotg}\frac{A}{2}-d\right)}{2\left(h\operatorname{cotg}\frac{A}{2}-d\right)}u+\frac{hd^2}{2\sin A\left(h\operatorname{cotg}\frac{A}{2}-d\right)}=0.$$

La condition pour que les racines de cette équation soient égales, c'est-à-dire, pour que le triangle soit isoscèle, est :

$$d^2+4hd\operatorname{tg}\frac{A}{2}=4h^2.$$

68. *Trois parallèles* LM, NP, QR *étant données, chercher les côtés d'un triangle dont les angles sont connus et dont les sommets sont situés sur ces parallèles* (fig. 22).

Fig. 22.

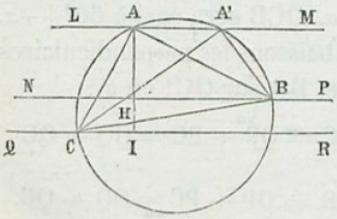

Soit A B C le triangle demandé; représentons par x, y, z, les côtés de ce triangle opposés aux angles A, B, C. Du sommet A, abaissons la perpendiculaire AHI sur NP.

Posons $AH=a$, $AI=b$. La perpendiculaire AHI divise l'angle A en deux parties : $BAH=\alpha$, $CAH=\beta$.

On a dans les triangles rectangles BAH, CAI :

$$a=z\cos\alpha,\quad b=y\cos\beta$$

et $$\cos A = \cos\alpha \cos\beta - \sin\alpha \sin\beta;$$

en remplaçant α et β par leurs valeurs, il vient :

$$\cos A = \frac{ab}{yz} - \sqrt{\left(1 - \frac{a^2}{z^2}\right)\left(1 - \frac{b^2}{y^2}\right)}$$

et, en faisant disparaître le radical :

$$y^2z^2 \sin^2 A = a^2y^2 + b^2z^2 - 2abyz \cos A.$$

On a : $$x : y = \sin A : \sin B,$$
$$x : z = \sin A : \sin C.$$

En substituant dans l'équation précédente, on obtient, après les simplifications :

$$x^2 = \frac{a^2}{\sin^2 C} + \frac{b^2}{\sin^2 B} - \frac{2ab \cos A}{\sin B \sin C}.$$

Il est facile de construire le triangle ; à cette fin, par un point A′ de la droite LM, faisons les angles A′BH = B et A′CR = C : les points B et C détermineront la longueur BC = x.

69. *Étant donnés trois cercles concentriques de rayon* R, r, r′, *chercher le côté du triangle équilatéral dont les sommets sont situés sur chaque circonférence.*

Soit ABC le triangle cherché (fig. 23).

Fig. 23.

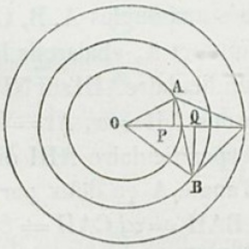

Menons les rayons OA, OB, OC. Si nous représentons ACO par α, OCB sera égal à $60^0 - \alpha$.

Abaissons les perpendiculaires AP et BQ sur OC ; on a :

$$OC = OP + PC = OQ + QC$$

ou

$$R = OP + PC = OQ + QC.$$

Mais on sait que $\overline{OP}^2 = r'^2 - \overline{AP}^2$ et $AP = x \sin\alpha$;

donc $OP = \sqrt{r'^2 - x^2 \sin^2\alpha}$ et $PC = x \cos\alpha$.

Il vient : $$R = x\cos\alpha + \sqrt{r'^2 - x^2\sin^2\alpha} \quad \ldots \quad (1).$$

On a aussi : $$OQ = \sqrt{r^2 - x^2\sin^2(60^0 - \alpha)}$$

et $$QC = x\cos(60^0 - \alpha);$$

d'où $$R = x\cos(60^0 - \alpha) + \sqrt{r^2 - x^2\sin^2(60^0 - \alpha)} \quad . \quad (2).$$

On obtient : $$x^2 - 2Rx\cos\alpha = r'^2 - R^2 \quad \ldots \quad (1),$$

$$x^2 - 2Rx\cos(60^0 - \alpha) = r^2 - R^2 \quad \ldots \quad (2).$$

En nous rappelant que $\sin 60^\circ = \frac{1}{2}\sqrt{3}$ et $\cos 60^\circ = \frac{1}{2}$, ce système devient :

$$x^2 - 2Rx\cos\alpha = r'^2 - R^2 \quad \ldots \quad (1),$$

$$x^2 - Rx\cos\alpha - \sqrt{3}Rx\sin\alpha = r^2 - R^2 \quad . \quad . \quad (2).$$

De (1), on tire : $$\cos\alpha = \frac{x^2 + R^2 - r'^2}{2Rx} \quad \ldots \quad (4).$$

Cette valeur, remplacée dans (2), donne :

$$\sin\alpha = \frac{x^2 + R^2 + r'^2 - 2r^2}{2\sqrt{3}Rx} \quad \ldots \quad (5).$$

On sait que $\sin^2\alpha + \cos^2\alpha = 1$. Élevant (4) et (5) au carré et additionnant membre à membre, il vient :

$$\frac{(x^2 + R^2 - r'^2)^2}{4R^2x^2} + \frac{(x^2 + R^2 + r'^2 - 2r^2)^2}{12R^2x^2} = 1,$$

équation bi-carrée qui donnera la valeur du côté cherché.

Exercices.

I. *On connaît : 1° le périmètre* 2p; *2° la hauteur* h *et 3° l'angle* B *à la base d'un triangle quelconque* ABC. *Calculer les trois côtés et les deux angles.*

II. *On connaît* : 1° *la médiane* m; 2° *la hauteur* h *partant du même sommet et 3° l'angle* B *à la base d'un triangle quelconque* ABC; *calculer les trois côtés et les deux angles.*

III. *On connaît le rayon* R *du cercle circonscrit, le rayon* r *du cercle inscrit et l'un des angles du triangle; calculer les trois côtés et les deux autres angles du triangle.*

IV. *Connaissant la surface* S *d'un triangle quelconque, un angle et le rayon du cercle inscrit, calculer les côtés et les angles du triangle.*

V. *Calculer les trois côtés du triangle quelconque* ABC *connaissant :* 1° *le périmètre* $2p$; 2° *le produit ou rectangle* K^2 *des deux côtés de l'angle* A ; 3° *la médiane* $AM = m$.

VI. *Trouver l'aire du triangle équilatéral, connaissant, soit les distances de ses sommets à une droite extérieure, soit les projections des côtés sur cette droite.*

VII. *Dans un triangle quelconque on donne les deux hauteurs partant des deux extrémités d'un même côté, et l'angle qu'elles font entre elles. Déterminer les trois côtés et les trois angles.*

VIII. *Dans un triangle quelconque on connaît la hauteur, la bissectrice et la médiane partant d'un même sommet; calculer les trois côtés et les trois angles du triangle.*

IX. *Étant donnés les angles, la surface et le périmètre d'un trapèze, trouver les quatre côtés.*

X. *Une circonférence et deux tangentes étant données, mener une troisième tangente telle que sa partie interceptée, soit par les deux premières, soit par les prolongements de deux diamètres, à angles droits, ait une longueur donnée* c.

XI. *Un point* M *et un cercle de centre* C *étant donnés, inscrire dans ce cercle un triangle dont un côté* AB *passe par* M, *et dont les deux autres côtés sont parallèles à deux droites données de position.*

XII. *Dans un triangle quelconque* ABC, *on donne :* 1° *l'angle* A ; 2° *les deux médianes qui partent des sommets* B *et* C *du triangle. Calculer les trois côtés et les deux angles.*

/ @ 0 % $! 6

!6 (-& .- - -/ ' ! - <
C ! AE" E'A A E! ! 2 1 E% @ E` a ! E8 ' E@$ 05; ; $"

@6 (-& .- - -/ ' ! - <
C ! AE" E'A A E $A EEE` a(! E E@$ 05< 5 $"

5

(-& .- - -/ ' ! - 5 + ' + ! @ $"

7' <% 6 9 ')

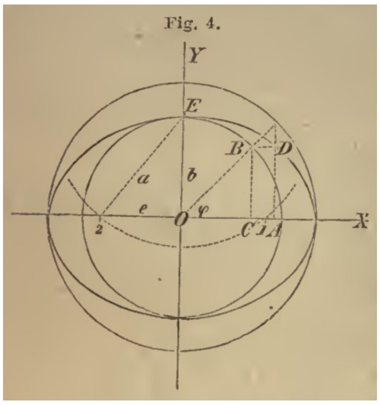
Fig. 4.
Y
E
B
D
a
b
e
O
φ
2
C
A
X

7' 5% ;6

$$r^2 = \frac{1}{\frac{\cos^2 \varphi}{a^2} - \frac{\sin^2 \varphi}{b^2}};$$

(-& .- - -/ ' ! - &"+! & ! 8 2
) $ ' ! ! @ ' % 6 $"

7' < %5 6 9 '

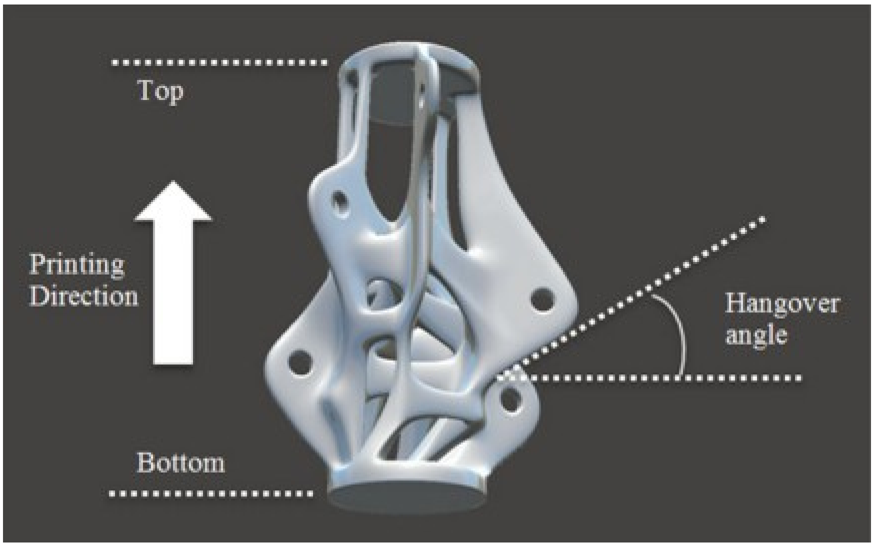
Top
Printing Direction
Hangover angle
Bottom

(-& .- - -/ ' ! -); ;) < E ! $"
7' % 56 *1 !

orbital. The next orbital up in energy ψ_2, is different from those we have met so far. Its symmetry demands that the node be in the middle; but this time the centre of the conjugated system is occupied by an *atom* and not by a *bond*. Having a node in the middle means having a zero coefficient c_2 on C-2, and hence the coefficients on C-l and C-3 in this orbital must be $\pm 1/\sqrt{2}$, if, squared and summed, they are to equal one. The atomic orbitals in ψ_2 are so far apart in space that their repulsive interaction does not, to a first approximation, raise the energy of this molecular orbital relative to that of an isolated p orbital. In consequence, whether filled or not, it does not contribute to the overall bonding. If the sum of the squares of the three orbitals on C-2 is also to equal one, then the coefficients on C-2 in ψ_1 and ψ_3* must also be $\pm 1/\sqrt{2}$. Finally, since symmetry requires that the coefficients on C-1 and C-3 in ψ_1 and ψ_3* have the same absolute magnitude, and the sum of their squares must equal $1-(1/\sqrt{2})^2$, we can deduce the unique set of *c*-values shown in Fig. 1.30. A pattern present in the allyl system because of its symmetry is seen with other symmetrical conjugated systems: the $|c|$ values are reflected across a mirror plane placed horizontally, half way up the set of orbitals, between ψ_1 and ψ_3*, and also across a mirror plane placed vertically, through C-2. It is only necessary therefore to calculate four of the nine numbers in Fig. 1.30, and deduce the rest from the symmetry.

In this picture of the bonding, we get no immediate appreciation of the energies of these orbitals relative to those of ethylene. The nonbonding orbital ψ_2 is clearly on the α level, that of a p orbital on carbon, and ψ_1 is lowered by the extra π bonding and ψ_3* is raised. To assess the energies, there is a simple geometrical device that works for linear conjugated systems. The conjugated system, *including the dummy atoms at the ends of the sine curves*, is inscribed vertically inside a circle of radius 2β, following the convention that one π bond in ethylene defines β. This is shown for ethylene and the allyl system in Fig. 1.31, where the dummy atoms are marked as dots at the top and bottom of the circle. The energies *E* of the π orbitals can then be calculated using Equation 1.10:

$$E = 2\beta \cos\frac{k\pi}{n+1} \qquad 1.10$$

where *k* is the number of the atom along the sequence of *n* atoms. This is simply an expression based on the trigonometry of Fig. 1.31, where, for example, the π orbital of ethylene is placed on the first atom ($k = 1$) of the sequence of two ($n = 2$) reading anticlockwise from the bottom. Thus the energies of the π orbitals in the allyl system are 1.414β below the α level and 1.414β above the α level.

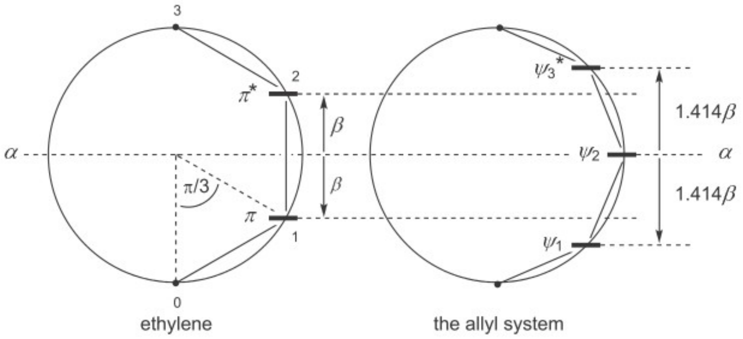

Fig. 1.31 Energies of π molecular orbitals in ethylene and the allyl system

We can gain further insight by building the picture of the π orbitals of the allyl system in another way. Instead of mixing together three p orbitals on carbon, we can combine two of them in a π bond first, as in Fig. 1.26, and then work out the consequences of having a third p orbital held within bonding distance of

$$E = 2\beta \cos \frac{k\pi}{n+1}$$

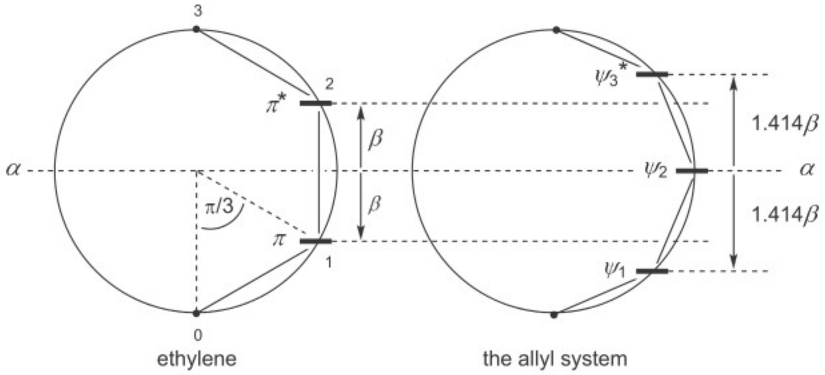
3
2
1
0
π^*
π
α
$\pi/3$
β
β
ψ_3^*
ψ_2
ψ_1
1.414β
1.414β
α
ethylene
the allyl system

9 ' * ' $! @ ! 2 ! " ! 2 2

(-& .- - -/ ' ! - (! ! " & ! 2 * " !
$! $ ' ! ! @ ' % 6 $"

7' ;% ;6 ; (! ' ! !@

6

$$a = \frac{-1+\sqrt{5}}{2} = 0.618033988\ldots.$$

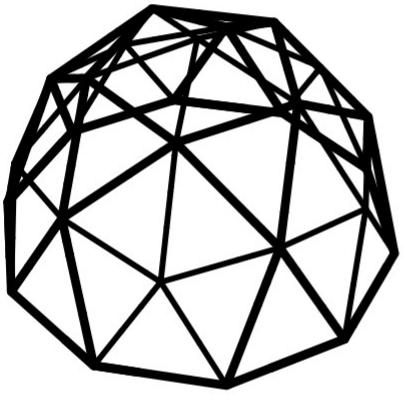

Figure 6.44. Geodesic dome

6.7.18. Figure 6.44 shows a geodesic dome with 5-fold vertices and 6-fold vertices. (A *geodesic* of the sphere is a great circle.) Figure 6.45 shows a bird's-eye view of the dome. The thinner edges emanate from the 5-vertices, while four of the six edges emanating from each 6-vertex are thicker. The five triangles that meet at a 5-vertex are isosceles, while two of the six triangles that meet at a 6-vertex are equilateral. This exercise uses vector algebra and the spherical coordinate system to work out the lengths and angles of the dome. Integration and the change of variable theorem play no role in this exercise.

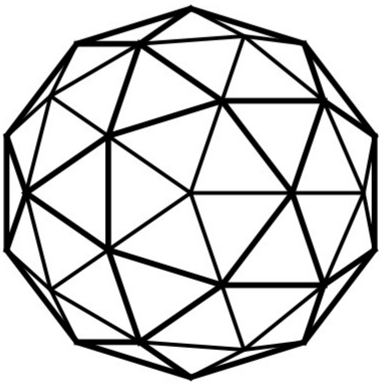

Figure 6.45. Geodesic dome, bird's-eye view

(a) Take all vertices to lie on a sphere of radius 1. The ten thick edges around the equator form a regular 10-gon. Show that consequently the thick edges have length

$$a = 2\sin(\pi/10) = 2\cos(2\pi/5).$$

This famous number from geometry goes back to Euclid. Note that $a = \zeta_5 + \zeta_5^{-1}$ where $\zeta_5 = e^{2\pi i/5} = \cos(2\pi/5) + i\sin(2\pi/5)$ is the fifth root of unity one-fifth of the way counterclockwise around the complex unit circle. Thus $a^2 + a - 1 = \zeta_5^2 + 2 + \zeta_5^3 + \zeta_5 + \zeta_5^4 - 1$, and the right side is 0 by the finite geometric sum formula. That is,

$$a^2 + a - 1 = 0, \quad a > 0,$$

and so the length of the thick edges is

$$a = \frac{-1+\sqrt{5}}{2} = 0.618033988\ldots.$$

This number is a variant of the so-called golden ratio.

(b) The dome has a point at the north pole $(0,0,1)$; then a layer of five points p_0 through p_4 around the north pole at some colatitude φ; then a layer of ten points q_0 through q_9, five at colatitude 2φ and the other five at some second colatitude ϕ; and finally the layer of ten equatorial points r_0 through r_9. The colatitude φ must be such that the triangle with vertices n, q_0, and q_2 is equilateral. These vertices may be taken to be

$$\begin{aligned} n &= (0,0,1), \\ q_0, q_2 &= (\cos(\pi/5)\sin(2\varphi), \mp\sin(\pi/5)\sin(2\varphi), \cos(2\varphi)). \end{aligned}$$

Show that the equilateral condition $|n - q_0|^2 = |q_2 - q_0|^2$ gives the condition $\cos^2(\varphi) = 1/(2-a)$, then $\sin^2(\varphi) = (1-a)/(2-a)$, then $\tan^2(\varphi) = a^2$, so that the colatitude of the five points about the north pole is

$$\varphi = \arctan(a) = 31.7174\ldots^\circ.$$

Use the cross-sectional triangle having vertices $\mathbf{0}$, n, p_0 and the law of cosines to show that the shorter segments have length

$$b = \sqrt{2(1 - 1/\sqrt{2-a})} = 0.546533057\ldots.$$

(Alternatively, one can find φ and b using the triangle with vertices n, p_0, p_1.) For reference in part (e), show that

$$2\varphi = \arctan(2) = 63.4349\ldots^\circ.$$

(c) Show that the angle of an isosceles triangle where its equal sides meet at a 5-vertex is

$$\alpha = 2\arcsin(a/(2b)) = 68.8619\ldots^\circ,$$

and the angles where its unequal sides meet at 6-vertices are

$$\beta = \arccos(a/(2b)) = 55.5690\ldots^\circ\,.$$

(d) Show that the angle where two a-segments meet along a geodesic is $180^\circ - 36^\circ$. Show that the angle where two b-segments meet along a geodesic (this happens at the 6-vertices but not at the 5-vertices) is $180^\circ - \varphi$.

(e) To find the colatitude ϕ of $q_1, q_3, \ldots, q_9$, take q_9 and q_1 to be

$$q_9, q_1 = (\cos(\pi/5)\sin(2\varphi), \mp\sin(\pi/5)\sin(2\varphi), \cos(2\varphi)),$$

and consider the geodesic containing them. Their cross product is normal to the plane of the geodesic. Show that this cross product is

$$q_9 \times q_1 = 2\sin(\pi/5)\sin(2\varphi)(-\cos(2\varphi), 0, \cos(\pi/5)\sin(2\varphi)).$$

Show by illustration that the latitude of this cross product is the colatitude ϕ of q_9 and q_1. Show that $\phi = \arctan(\sqrt{2+a})$. Show further that $(a+1)^2 = 2+a$, so that in fact,

$$\phi = \arctan(a+1) = 58.2825\ldots^\circ\,.$$

6.8 Topological Preliminaries for the Change of Variable Theorem

This section establishes some topological results to prepare for proving the change of variable theorem (Theorem 6.7.1), and then the next section gives the proof. Both sections are technical, and so the reader is invited to skim as feels appropriate. For instance, one might focus on the discussion, the statements, and the figures, but go light on the proofs.

In preparation for proving the change of variable theorem, we review its statement. The statement includes the terms *boundary* and *interior*, which we have considered only informally so far, but we soon will discuss them more carefully. The statement also includes the term *open*, and the reader is reminded that a set is called open if its complement is closed; we soon will review the definition of a closed set. The statement includes the term $\mathcal{C}^1$*-mapping*, meaning a mapping such that all partial derivatives of all of its component functions exist and are continuous. And the statement includes the notation K° for the interior of a set K. The theorem says:

> *Let $K \subset \mathbb{R}^n$ be a compact and connected set having boundary of volume zero. Let A be an open superset of K, and let*
>
> $$\Phi : A \longrightarrow \mathbb{R}^n$$
>
> *be a $\mathcal{C}^1$-mapping such that*

6 (-& .- - -/ ' ! - Kd ' & " ! "#
$ ' % 6 $"
7')%;<6

$$U_{U3} = \frac{2U_{E3}}{\sqrt{2+\sqrt{2+\sqrt{2}}}} \tag{4.61}$$

where U_{Un} and U_{En} refer to the perimeters of circumscribed and inscribed polygons with 2^{n+1} sides. Kabus then made the numerically interesting observation that the perimeters of the circumscribed polygons were always twice as far from the circumference of the circle as those of the inscribed 2^n-sided polygons. He proved

$$\lim_{n\to\infty} \frac{\frac{1}{2}U_{Un} - \pi}{\pi - \frac{1}{2}U_{En}} = 2 \tag{4.62}$$

From the two observations, Kabus then derived the following iterative algorithm for π:

Algorithm (4.63) Kabus (alg 4.63)

```
Initialise:
    $a_0 := \sqrt{\frac{1}{2} + \frac{1}{4}\sqrt{2}}$
Iterate: $(n = 0, 1, 2, \ldots K - 1)$
    $a_{n+1} := \sqrt{\frac{1}{2} + \frac{1}{2}a_n}$
Then:
    $p_K = \frac{2^{K+2}}{3} \frac{\sqrt{2-2a_K}+\sqrt{2-2a_{K+1}}}{\sqrt{2+2a_{K+1}}} \xrightarrow{1} \pi$
```

Each iteration improves the π approximation by 1.2 places, whereas, for example, Archimedes's method only produced 0.6 places each time the number of sides of the polygon was doubled. Adding various cunning improvements in the arithmetic and using a PASCAL program, Kabus arrived in this way at 5000 decimal places of π.

4.4 On continued fractions

Continued fractions are mentioned a number of times in this book and we would therefore like to say a few words about them at this point.

Continued fractions are the "lost sons" of mathematics classroom teaching. They are regarded as too advanced for secondary schools and too elementary for universities, as a result of which they generally fall between the cracks of both syllabuses [18, p. 129]. Their historical roots go back to the 17th century mathematicians Cataldi, Wallis and Huygens, and their theory traces its origins to Leonhard Euler, who presented them in 1748 in his *Introductio in analysin infinitorum* [52, pages 303 ff]. The classic textbook on this interesting topic was written by Oskar Perron [90], and a more recent textbook is that by C. D. Olds [87].

A continued fraction is a fraction whose numerator is an integer and whose denominator is the sum of an integer and a fraction which in turn possesses the same form.

$$b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3 + \ddots \cfrac{a_{n-1}}{b_{n-1} + \cfrac{a_n}{b_n}}}}} \tag{4.64}$$

$$= b_0 + a_1/\,(b_1 + a_2/\,(b_2 + a_3/(b_3 + \cdots + a_n/b_n))) \tag{4.65}$$

This representation, which rapidly grows to occupy a large physical area, is often replaced by

$$b_0 + \frac{a_1|}{|b_1} + \frac{a_2|}{|b_2} + \frac{a_3|}{|b_3} + \cdots + \frac{a_n|}{|b_n} \tag{4.66}$$

A. Pringsheim, 1898

or by

$$b_0 + \frac{a_1}{b_1 +} \frac{a_2}{b_2 +} \frac{a_3}{b_3 +} \cdots \frac{a_n}{b_n} \tag{4.67}$$

L.J. Rogers, 1907

Regular (Perron) or *simple* (Olds) continued fractions [90][87] are ones in which all the numerators $a_i = 1$. They can be written even more concisely like this:

$$[b_0, b_1, b_2, \dots b_n] \tag{4.68}$$

The elements a_i and b_i of a continued fraction are also known, respectively, as "partial numerator" and "partial denominator".

Every real number can be represented uniquely by a simple continued fraction. A number is rational if its continued fraction is finite and irrational if its continued fraction is infinite.

The second theorem immediately proves, for example, that all the numbers contained in the following list are irrational because their continued fractions are infinite. The proof of the irrationality of π which Johann Heinrich Lambert wrote in 1766 is also based on this theorem. Lambert showed that the continued fraction of $\arctan 1 =$

$\pi/4$ is infinite and therefore $\pi/4$ and hence π must be irrational (see page 192).

The transformation of a number from its decimal representation into its simple continued fraction form, and vice versa, is quite simple. It can be performed with any pocket calculator which has an inverse operation $1/x$. The two algorithms look like this:

```
// cf[] : array with the elements of the continued fraction

procedure NumberToCf(number, n, cf[0..n-1])
{
    for k:=0 to n-1
    {
       x      := floor(number)
       cf[k]  := x
       number:= 1/(number-x)
    }
}
```

and the reverse procedure for calculating a number from the elements of its simple continued fraction goes like this:

```
function CfToNumber(n, cf[0..n-1])
{
   number := cf[n-1]

   for k:=n-2 to 0 step -1
   {
      number := 1/number + cf[k]
   }
   return number
}
```

The simple continued fractions of some prominent constants are shown below[3]:

$$\begin{aligned}\phi &= \frac{\sqrt{5}+1}{2} \quad \text{(Golden ratio)} \qquad (4.69)\\ &= 1.61803\,39887\,49894\,84820\,\ldots\\ &= [1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,\ldots]\end{aligned}$$

$$\begin{aligned}\sqrt{2} &= 1.41421\,35623\,73095\,04880\,\ldots \qquad (4.70)\\ &= [1,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,\ldots]\end{aligned}$$

$$\begin{aligned}\sqrt{3} &= 1.73205\,08075\,68877\,29352\,\ldots \qquad (4.71)\\ &= [1,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,\ldots]\end{aligned}$$

---

[3] As this book is concerned almost exclusively with numbers that are transcendental and, as a very minimum, are irrational, we are naturally only interested in *infinite* continued fractions.

$$\begin{aligned} e &= 2.71828\,18284\,59045\,23536\,\ldots \qquad (4.72)\\ &= [2,1,2,1,1,4,1,1,6,1,1,8,1,1,10,1,1,12,1,1,14,1,\ldots] \end{aligned}$$

Euler, 1737

$$\begin{aligned} e^2 &= 7.38905\,60989\,30650\,22723\,\ldots \qquad (4.73)\\ &= [7,2,1,1,3,18,5,1,1,6,30,8,1,1,9,42,11,1,1,12,54,\ldots] \end{aligned}$$

Stieltjes, c. 1890 [90]

$$\sqrt{e} = 1.64872\,12707\,00128\,14684\,\ldots \qquad (4.74)$$

Sundman, 1895

$$= [1,1,1,1,5,1,1,9,1,1,13,1,1,17,1,1,21,1,1,25,1,1,\ldots]$$

$$\begin{aligned} \pi &= 3.14159\,26535\,89793\,23846\,\ldots \qquad (4.75)\\ &= [3,7,15,1,292,1,1,1,2,1,3,1,14,2,1,1,2,2,2,2,1,84,\ldots] \end{aligned}$$

$$\begin{aligned} \sqrt{\pi} &= 1.77245\,38509\,05516\,02729\,\ldots \qquad (4.76)\\ &= [1,1,3,2,1,1,6,1,28,13,1,1,2,18,1,1,1,83,1,4,\ldots] \end{aligned}$$

$$\begin{aligned} \pi^e &= 22.45915\,77183\,61045\,47343\,\ldots \qquad (4.77)\\ &= [22,2,5,1,1,1,1,1,3,2,1,1,3,9,15,25,1,1,5,4,1,\ldots] \end{aligned}$$

$$\begin{aligned} \sqrt[3]{2} &= 1.25992\,10498\,94873\,16476\,\ldots \qquad (4.78)\\ &= [1,3,1,5,1,1,4,1,1,8,1,14,1,10,2,1,4,12,2,3,2,1,\ldots] \end{aligned}$$

$$\begin{aligned} \gamma &= \lim_{n\to\infty} \sum_{k=1}^{n} \frac{1}{k} - \ln(n+1) \quad \text{(Euler's constant)} \qquad (4.79)\\ &= 0.57721\,56649\,01532\,86060\,\ldots\\ &= [0,1,1,2,1,2,1,4,3,13,5,1,1,8,1,2,4,1,1,40,1,11,3,\ldots] \end{aligned}$$

Euler, 1734

On closer examination we find there is a recurring pattern in the continued fractions of the first 6 constants, e.g. in $\sqrt{e}$ the pattern $1, 1, 4n+1$, whereas this does not occur in the second five cases beginning with π.

We cannot conclude from this that the simple continued fractions of many or even of the majority of common constants possess a regular pattern. In reality, this applies only to a minority, namely to the so-called quadratic irrationalities, one example of which is the "Golden ratio" ($\phi = 1 + \sqrt{5}/2$), also to a few well-known transcendental numbers, such as the constant e itself and individual algebraic expressions containing e, e^2 or $e^{1/q}$, where q is an integer.

Interestingly, in almost all irrational or transcendental numbers, including π, the geometric mean of their first n continued fraction elements converges towards a fixed limit and this limit is the same in all such numbers. This astonishing discovery was made in 1935 by Alexander Khintchine (1894–1959) [74]:

$$\lim_{n\to\infty} \sqrt[n]{b_1 \cdot b_2 \cdot \ldots \cdot b_n} = 2.68545\,20010\ldots = K_0 \tag{4.80}$$

The limit K_0 is defined as $\prod_{k=1}^{\infty}[1 + \frac{1}{k(k+2)}]^{\log_2 k}$ and became known as the *Khintchine constant*. The number is difficult to calculate, and only a modest 110,000 decimal places are available.

For the number π, the geometric mean of the first $17,001,303$ elements of its simple continued fraction come to 2.686393 [12], which is fairly close to the Khintchine constant. On the other hand, the geometric mean for the number e, the big exception, moves away as n increases, while in the case of another exception, the constant ϕ (Golden ratio), whose continued fraction consists of pure ones, the geometric mean does converge, but towards 1 rather than K_0.

If one looks at the table with the approximations for π on page 54 in this light, one could form the opinion that continued fractions produce a more compact representation of π than the decimal representation. For the table shows, for example, that π can be just as well approximated with 3 continued fraction elements as with 6 decimal digits. However, this is only the case at particular places, and these are precisely the ones at which it is especially beneficial to terminate a continued fraction. Further on in both representations, the advantage of the continued fraction always disappears, and generally the simple continued fraction of an irrational number requires just as many elements as its equally accurate decimal fraction possesses decimal places. The quotient from the number of accurate decimal places of the approximation and the number of continued fraction elements converges towards a fixed limit which is close to 1, as Alexander Khintchine once again proved in 1935. The exact value $1.03064\ldots = \pi^2/(6 \cdot \ln 2 \cdot \ln 10)$ was stated somewhat later, in 1937, by Paul Lévy.

(-& .- - -/ ' ! - & " 8 2 $ '
% 6 $"

7' %; 6 9 ' ; C * (

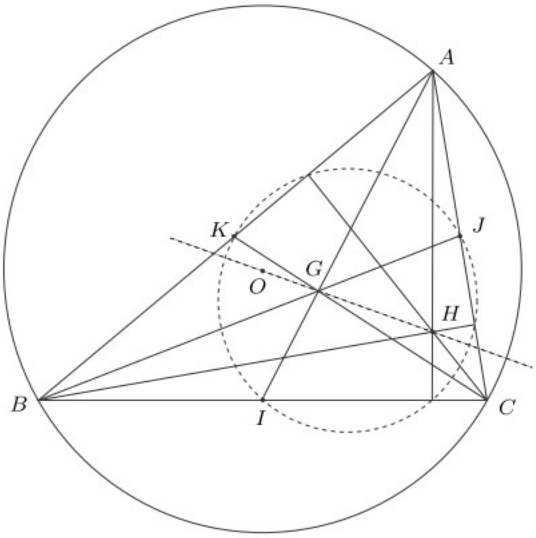

7'# % 56

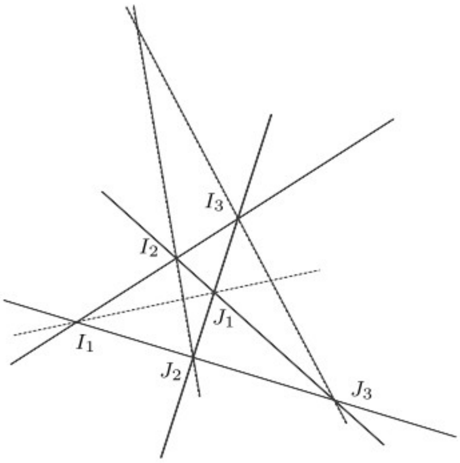
I_3
I_2
J_1
I_1
J_2
J_3

)

(-& .- - -/ ' ! -) 9 " C 2 ! " (! !
@ $"

7' %) 6

C ' & '

irrationality of $\sqrt[3]{2}$. (The irrationality of $\sqrt[3]{2}$ may be established very simply as in Problem 2 of the previous section.) ∎

The basic idea of this section will be slightly generalized in the next section, where it will yield the solution of the angle trisection problem.

PROBLEMS

1. Can the cube be "tripled"?
2. Can the cube be "quadrupled"?

Section 1.3 Trisecting the Angle

The classical problem involving angle trisection is: **Given an arbitrary angle, is it possible to trisect it?** (As usual, we are restricting ourselves to the use of a compass and an unmarked straightedge.) The answer is "no", but one must be careful to understand what this answer means. There are certainly some particular angles that can be trisected, as, for example, the angle of 90°. (Why?) However, it is not possible to trisect an *arbitrary* angle. To put it another way, there exists at least one angle that *cannot* be trisected. We shall actually produce one such angle, in particular the innocuous-looking angle of 60°.

Since a 60° angle can be constructed, if it could also be trisected, then a 20° angle could be constructed. But then, as illustrated in Figure 5, the number cos 20° would be constructible. It will be shown, however, that the number cos 20° is a root of a cubic equation which has no constructible roots, thus leading to a contradiction.

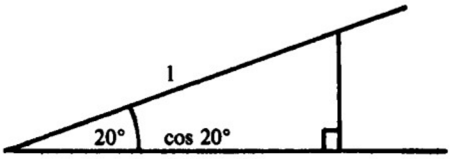

Mark off a unit length on one side and drop a perpendicular to the other.

Figure 5

From elementary trigonometric formulas, we have

$$\begin{aligned}\cos 3\theta &= \cos(2\theta + \theta)\\ &= \cos 2\theta \cos \theta - \sin 2\theta \sin \theta\\ &= (\cos^2\theta - \sin^2\theta)\cos \theta - (2 \sin \theta \cos \theta)\sin \theta\\ &= \cos^3\theta - 3 \sin^2\theta \cos \theta\\ &= \cos^3\theta - 3(1 - \cos^2\theta)\cos \theta\\ &= 4 \cos^3\theta - 3 \cos \theta.\end{aligned}$$

When $\theta = 20°$, $\cos 3\theta = \frac{1}{2}$, so the number cos 20° must be a root of the cubic equation

$$8u^3 - 6u - 1 = 0.$$

This equation is made a little simpler by the change of variable $x = 2u$, which yields

$$x^3 - 3x - 1 = 0.$$

If cos 20° were constructible, then it would be possible to construct a root of this equation. As the first step in the proof that this is impossible, we make an observation which is very similar to that of Lemma 2:

LEMMA 3. *Let* $\mathbf{F}(\sqrt{k})$ *be a quadratic extension of a field* $\mathbf{F}$. *If the equation* $x^3 - 3x - 1 = 0$ *has a root in* $\mathbf{F}(\sqrt{k})$, *then it has a root in* $\mathbf{F}$ *itself.*

Proof. If the equation has a root in $\mathbf{F}(\sqrt{k})$, we may denote it by $a + b\sqrt{k}$, where a, b, and k are in $\mathbf{F}$. If $b = 0$ we are done, for $a \in \mathbf{F}$ is a root. If $b \neq 0$, it will be shown that $-2a$ is a root, and this is obviously in $\mathbf{F}$.

Substituting the given root into the equation, we obtain:

$$(a + b\sqrt{k})^3 - 3(a + b\sqrt{k}) - 1 = 0$$

$$a^3 + 3a^2b\sqrt{k} + 3ab^2k + b^3k\sqrt{k} - 3a - 3b\sqrt{k} - 1 = 0$$

$$[a^3 + 3ab^2k - 3a - 1] + [3a^2b + b^3k - 3b]\sqrt{k} = 0.$$

But now we can conclude that $[3a^2b + b^3k - 3b] = 0$, for otherwise we could solve for $\sqrt{k}$ as an element of $\mathbf{F}$. But this then also requires that $[a^3 + 3ab^2k - 3a - 1] = 0$. After dividing the first of these two equations by b (which is not 0), we have

$$3a^2 + b^2k - 3 = 0$$

$$a^3 + 3ab^2k - 3a - 1 = 0.$$

Solving the first for b^2k and substituting the result into the second, we obtain

$$a^3 + 3a(3 - 3a^2) - 3a - 1 = 0$$

$$-8a^3 + 6a - 1 = 0$$

$$(-2a)^3 - 3(-2a) - 1 = 0$$

from which we see that $-2a$, which is in $\mathbf{F}$, is a solution of the original equation. ∎

Now the main result is easily accessible.

THEOREM 3. *It is not possible to trisect an arbitrary angle.*

Proof. If it were possible to trisect an arbitrary angle, it would be possible to trisect, say, a 60° angle, which is itself constructible. But, as noted earlier, this would imply the constructibility of cos 20° and hence the constructibility of a root of the cubic equation $x^3 - 3x - 1 = 0$. By Theorem 1, such a root would have to belong to a field $\mathbf{F}_N$ situated in a sequence $\mathbf{Q} = \mathbf{F}_0 \subset \mathbf{F}_1 \subset \mathbf{F}_2 \subset \cdots \subset \mathbf{F}_N$ in which each field is a quadratic extension of the previous one. By a repeated application of Lemma 3, then, there would have to be a *rational* root of $x^3 - 3x - 1 = 0$. But this will now be shown to be impossible.

If a rational number M/N, expressed in lowest terms, were a root of the equation, we would have

$$\frac{M^3}{N^3} - \frac{3M}{N} - 1 = 0,$$

and so

$$M^3 - 3MN^2 - N^3 = 0.$$

Thus we could write both

$$M^3 = N[3MN + N^2]$$

and

$$N^3 = M[M^2 - 3N^2].$$

From the first we conclude that if N had a prime factor p, then p would divide M^3 and hence M. From the second we conclude that if M had a prime factor q, then q would divide N^3 and hence N. Since M and N were assumed to have no factors in common, the only possible values for M and N are ±1, and so the only possible rational roots are ±1. But by inspection, neither of these actually turns out to be a root. ∎

It is important to realize that there is a difference between the questions:

(i) Is it possible to construct the angle θ?

(ii) Given the angle 3θ, is it possible to construct the angle θ?

The second question is the one addressed by the angle trisection problem, and it only reduces to the first question in cases where the angle 3θ is itself constructible. It happened to be convenient to use such a case in the proof of Theorem 3, but the reader should know that there are angles for which the answer to question (i) is "no" but the answer to question (ii) is "yes". To analyze these two different questions, it is useful to convert them into more algebraic terms. Remembering that in the presence of a unit segment we have available an angle if and only if we have available a segment whose length is its cosine, we can rephrase these questions as follows:

(i′) Is $\cos\theta$ constructible?

(ii′) Is $\cos\theta$ constructible if, in addition to a unit segment, we are also given a segment of length $\cos 3\theta$?

In Problem 4, the reader is asked to show the existence of an angle for which these questions have different answers.

PROBLEMS

1. It is obvious that by repeated bisection it is possible to divide an arbitrary angle into 4 equal parts. Show

how this may also be deduced algebraically from the equation relating cos 4θ and cos θ.

2. Discuss the validity of the following proof of Theorem 3: "By the continuity of cos θ, there exists an angle θ_0 such that cos $\theta_0 = (\sqrt[3]{2})/2$. By Theorem 2, cos θ_0 is not constructible. Hence $3\theta_0$ is an angle that cannot be trisected."

*3. Prove that it is not possible, in general, to quintisect an arbitrary angle, i.e., to divide it into five equal parts.

4. Is the difficulty in Problem 2 real or imagined? Show that there are trisectible angles 3θ such that θ is not constructible.

Section 1.4. Squaring the Circle

The third classical construction problem, that of squaring the circle, turns out to be much more difficult than the previous two. The question is: **Given a circle, is it possible to construct a square with the same area as the circle?** If we take as our unit length the radius of the circle, its area is π, and so the construction of the appropriate square amounts to the construction of a segment of length $\sqrt{\pi}$, which would necessitate the constructibility of π itself. It will be shown, however, that π is not constructible, and thus that it is not possible to square the circle. The proof that π is not constructible will consist of two main parts: first, that every constructible number is the root of some polynomial equation with rational coefficients; second, that π cannot be a root of any polynomial equation with rational coefficients. The first part, which is fairly simple, will be covered in this section. The second part, which is much more difficult, will be proved in Section 1.7, the intervening sections being used to introduce some ideas needed there as well as later on.

!' %)<6 1 ! ' (

how this may also be deduced algebraically from the equation relating cos 4θ and cos θ.

2. Discuss the validity of the following proof of Theorem 3: "By the continuity of cos θ, there exists an angle θ_0 such that cos $\theta_0 = (\sqrt[3]{2})/2$. By Theorem 2, cos θ_0 is not constructible. Hence $3\theta_0$ is an angle that cannot be trisected."

*3. Prove that it is not possible, in general, to quintisect an arbitrary angle, i.e., to divide it into five equal parts.

4. Is the difficulty in Problem 2 real or imagined? Show that there are trisectible angles 3θ such that θ is not constructible.

Section 1.4. Squaring the Circle

The third classical construction problem, that of squaring the circle, turns out to be much more difficult than the previous two. The question is: **Given a circle, is it possible to construct a square with the same area as the circle?** If we take as our unit length the radius of the circle, its area is π, and so the construction of the appropriate square amounts to the construction of a segment of length $\sqrt{\pi}$, which would necessitate the constructibility of π itself. It will be shown, however, that π is not constructible, and thus that it is not possible to square the circle. The proof that π is not constructible will consist of two main parts: first, that every constructible number is the root of some polynomial equation with rational coefficients; second, that π cannot be a root of any polynomial equation with rational coefficients. The first part, which is fairly simple, will be covered in this section. The second part, which is much more difficult, will be proved in Section 1.7, the intervening sections being used to introduce some ideas needed there as well as later on.

We shall make use of the following simple lemma:

LEMMA 4. *Let* $\mathbf{F}(\sqrt{k}\,)$ *be a quadratic extension of a field* $\mathbf{F}$. *If* a *is a root of a polynomial equation of degree* n *with coefficients in* $\mathbf{F}(\sqrt{k}\,)$, *then* a *is a root of a polynomial equation of degree* $2n$ *with coefficients in* $\mathbf{F}$.

Proof. By hypothesis, a is a root of an equation of the form

$$\left(a_n + b_n\sqrt{k}\,\right)x^n + \left(a_{n-1} + b_{n-1}\sqrt{k}\,\right)x^{n-1} + \cdots + \left(a_0 + b_0\sqrt{k}\,\right) = 0,$$

where for all i, a_i and $b_i \in \mathbf{F}$, and furthermore, a_n and b_n are not both zero. Transferring the terms involving $\sqrt{k}$ to the other side of the equation, we obtain

$$a_nx^n + a_{n-1}x^{n-1} + \cdots + a_0 = -\sqrt{k}\left(b_nx^n + b_{n-1}x^{n-1} + \cdots + b_0\right).$$

Upon squaring both sides and collecting all terms together again on the left, we have a polynomial equation of which a is also a root. Its degree is $2n$, because the coefficient of x^{2n}, the highest power present, is $a_n^2 - kb_n^2$, which is not 0 by the assumption that $\sqrt{k} \notin \mathbf{F}$. ∎

Using this lemma, we are now able to prove:

THEOREM 4. *Every constructible number is the root of some polynomial equation with rational coefficients.*

Proof. If the number a is constructible, then by Theorem 1 there is a sequence of fields $\mathbf{Q} = \mathbf{F}_0 \subset \mathbf{F}_1 \subset \cdots \subset \mathbf{F}_N$ such that $a \in \mathbf{F}_N$ and for each j, $0 \leqslant j$

$\leqslant N-1$, $\mathbf{F}_{j+1}$ is a quadratic extension of $\mathbf{F}_j$. Now, a satisfies a polynomial equation of degree 1 with coefficients in $\mathbf{F}_N$, namely, the equation $x - a = 0$. By the repeated application of Lemma 4, a satisfies a polynomial equation of degree 2^i with coefficients in $\mathbf{F}_{N-i}$. Hence for $i = N$, we see that a satisfies a polynomial equation of degree 2^N with coefficients in $\mathbf{Q}$. ∎

In view of Theorem 4, once it has been shown (in Section 1.7) that π is not the root of any polynomial equation with rational coefficients, we shall have a complete proof of the result stated earlier, namely:

THEOREM 5. *It is not possible to "square the circle".*

PROBLEMS

1. The number $\sqrt{2}+\sqrt{3}$, being constructible, has associated with it a sequence of quadratic field extensions $\mathbf{Q} = \mathbf{F}_0 \subset \mathbf{F}_1 \subset \cdots \subset \mathbf{F}_N$ such that $\sqrt{2}+\sqrt{3} \in \mathbf{F}_N$. Determine the smallest such number N. Using this value, what is the degree of the polynomial whose existence is proved in Theorem 4? Following the procedure in the proof of Theorem 4, determine this and any intermediate polynomials.
2. Does there exist a polynomial equation with rational coefficients and of degree less than that determined in Problem 1 such that $\sqrt{2}+\sqrt{3}$ is a root?

Section 1.5. Polynomials and Their Roots

It now becomes essential for us to expand our considerations to the set $\mathbf{C}$ of complex numbers. These

(-& .- - -/ ' ! - ! $"

7' %)6 9 '

& 8 & !

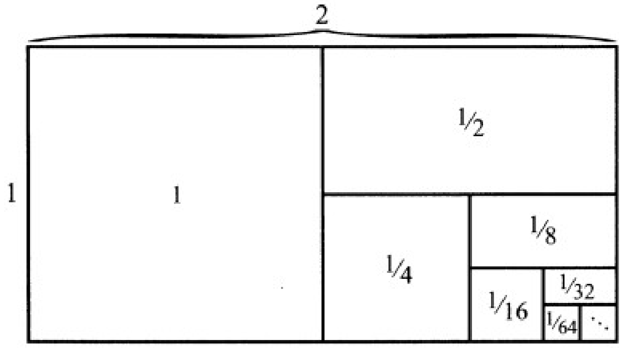
2
1
1
1⁄2
1⁄4
1⁄8
1⁄16
1⁄32
1⁄64
⋱

!6 (-& .- - -/ ' ! - K 4 ! '
$ @ E C $ ' C ! ! ((% <6 $"

@6 (-& .- - -/ ' ! - K 4 ! '
$ @ E C $ ' C ! ! % 6 $"

7' % 6 C " ! 1 !

This is how I first got interested in mathematics, and irrational numbers in particular. Getting to "know" the infinite decimal for $\sqrt{2}$ may be too discouraging for some—mathematicians generally prefer a problem that gives them some chance of success—but it is a reminder of how much remains to be understood. And since the infinite decimal is so completely baffling, we are encouraged to look at different approaches to $\sqrt{2}$, in the hope of finding enlightenment elsewhere. Surely $\sqrt{2}$ is a simple irrational number, so there should be a simple way of looking at it.

1.6 The Continued Fraction for $\sqrt{2}$

A more enlightening process than the infinite decimal algorithm would relate $\sqrt{2}$ to a *predictable* infinite sequence of numbers, and such a process actually exists: the famous *Euclidean algorithm.* The algorithm is named after Euclid because its first known appearance is in the *Elements*, Book VII, but there is evidence for its earlier use in the investigation of irrationals.

The object of the Euclidean algorithm is to find a *common measure* of two magnitudes a and b. We say that c is a common measure of a and b if

$$a = mc \quad \text{and} \quad b = nc \quad \text{for some natural numbers } m \text{ and } n.$$

The key to the algorithm is the fact that

$$a - b = (m - n)c,$$

so *any common measure c of a and b is also a measure of their difference.* If a is greater than b, we can therefore replace the problem of finding a common measure of a and b by the "smaller" problem of finding a common measure of $a - b$ and b. This is precisely what the algorithm does. The algorithm seeks a common measure by "continually subtracting the lesser magnitude from the greater" (as Euclid put it).

When a common measure exists the Euclidean algorithm will find it in a finite number of steps. In particular, when a and b are positive natural numbers the algorithm gives what is called the *greatest common divisor* of a and b. This concept is very important in the theory of natural numbers, and we pursue it further in Chapter 7.

Right now, we are interested in the opposite situation, where a and b have no common measure. In this case the algorithm necessarily runs forever, but perhaps in a way that enables to us to understand *why* it runs forever. We are in luck when $a = \sqrt{2}$ and $b = 1$. Here, the Euclidean algorithm runs forever because it becomes *periodic*: it returns to a situation it has been in before. Unlike

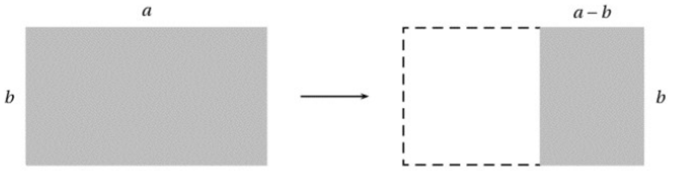

Figure 1.9. Subtracting the lesser from the greater of two magnitudes.

the infinite decimal process for $\sqrt{2}$, the Euclidean algorithm on $\sqrt{2}$ and 1 has infinite behavior of the simplest possible kind.

To show periodic behavior as clearly as possible, I apply the Euclidean algorithm instead to the numbers $a = \sqrt{2}+1$ and $b = 1$, where periodicity kicks in a little earlier. Also, I present the algorithm in a geometric form. Two magnitudes a and b will be represented by the greater and lesser sides of a rectangle, and the lesser b is subtracted from the greater a by cutting off a square of side b (Figure 1.9).

When $a = \sqrt{2}+1$ and $b = 1$ we can subtract b from a twice before getting a magnitude less than b. The result is a rectangle with greater side 1 and lesser side $\sqrt{2}-1$, as shown in Figure 1.10. I claim *this is a repetition of the original situation* because *the new rectangle is the same shape as the old.*

To see why, we calculate the ratio $\frac{\text{long side}}{\text{short side}}$, which represents the shape of the rectangle. For the old rectangle,

$$\frac{\text{long side}}{\text{short side}} = \frac{\sqrt{2}+1}{1} = \sqrt{2}+1.$$

For the new rectangle,

$$\begin{aligned}
\frac{\text{long side}}{\text{short side}} &= \frac{1}{\sqrt{2}-1} \\
&= \frac{1}{\sqrt{2}-1}\,\frac{\sqrt{2}+1}{\sqrt{2}+1} \quad \text{multiplying top and bottom by } \sqrt{2}+1 \\
&= \frac{\sqrt{2}+1}{(\sqrt{2})^2-1^2} \quad \text{because } (\sqrt{2}-1)(\sqrt{2}+1) = (\sqrt{2})^2-1^2, \\
&\qquad \text{by the “difference of two squares” from Section 1.4} \\
&= \frac{\sqrt{2}+1}{2-1} = \sqrt{2}+1.
\end{aligned}$$

Thus the new rectangle has the same shape as the old, so as we continue to run the Euclidean algorithm the same thing will happen over and over: two

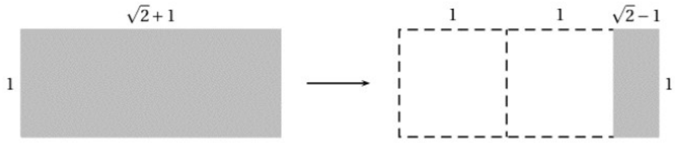

Figure 1.10. Subtracting 1 twice from $\sqrt{2}+1$.

squares will be cut off the new rectangle, producing yet another rectangle of the same shape, and so on. The algorithm will not terminate because it keeps reproducing the same geometric situation.

It is also clear that the Euclidean algorithm will not terminate if it starts with the pair $\sqrt{2}$ and 1. The only change is that just one square is cut off the initial rectangle. After that, rectangles with sides in the ratio $\sqrt{2}+1:1$ recur forever.

Since the Greeks knew that $\sqrt{2}$ is irrational, they knew that the Euclidean algorithm does not terminate when applied to $\sqrt{2}$ and 1. Indeed, the *Elements*, Book X, Proposition 2 actually states nontermination of the algorithm as a criterion for irrationality. What is not so clear is whether the Greeks observed the *periodicity* of the algorithm on $\sqrt{2}$ and 1. Many people believe they did, but the *Elements* does not mention it, and the other likely sources of this information are lost.

One thing is clear. It is a lot easier to display the periodicity of $\sqrt{2}$ using notation for fractions, which the Greeks did not have. Again, it is better to use $\sqrt{2}+1$ for purposes of demonstration, because the periodicity starts earlier. We start by subtracting 1 twice from $\sqrt{2}+1$, as the algorithm requires. This produces the remainder $\sqrt{2}-1$, which equals $\frac{1}{\sqrt{2}+1}$ because $(\sqrt{2}-1)(\sqrt{2}+1)=1$, as found above. Thus

$$\sqrt{2}+1=2+(\sqrt{2}-1)=2+\frac{1}{\sqrt{2}+1}.$$

This says that the $\sqrt{2}+1$ in the bottom of the fraction on the right side equals the *whole* right side. We can therefore replace $\sqrt{2}+1$ by $2+\frac{1}{\sqrt{2}+1}$, and keep on doing this to our hearts' content:

$$\sqrt{2}+1=2+\frac{1}{\sqrt{2}+1}=2+\frac{1}{2+\frac{1}{\sqrt{2}+1}}=2+\frac{1}{2+\frac{1}{2+\frac{1}{\sqrt{2}+1}}}\quad\cdots$$

The logical conclusion of this process is an infinite periodic fraction for $\sqrt{2}+1$, called its *continued fraction*:

$$\sqrt{2}+1=2+\cfrac{1}{2+\cfrac{1}{2+\cfrac{1}{2+\cfrac{1}{2+\cfrac{1}{2+\ddots}}}}}$$

Finally, we get the continued fraction for $\sqrt{2}$ by subtracting 1 from both sides:

$$\sqrt{2}=1+\cfrac{1}{2+\cfrac{1}{2+\cfrac{1}{2+\cfrac{1}{2+\cfrac{1}{2+\ddots}}}}}$$

The infinite recurrence of 2s produced by the Euclidean algorithm is reflected by the infinite sequence of 2s in the continued fraction.

Periodicity in Continued Fractions and Decimals

You may be wondering whether it is valid to expel $\sqrt{2}$ from the right side of the equations as we did above, by letting the bottom line of the fraction descend to infinity. As soon as we accept that the continued fraction is meaningful, we can verify that its value is $\sqrt{2}+1$ as follows. Let

$$x=2+\cfrac{1}{2+\cfrac{1}{2+\cfrac{1}{2+\cfrac{1}{2+\cfrac{1}{2+\ddots}}}}}$$

Then x is a positive number (greater than 2 in fact), and since the term under the 1 on the right side equals the whole right side, we also have

$$x=2+\frac{1}{x}.$$

Multiplying both sides by x, and rearranging, we get the equation

$$x^2-2x-1=0.$$

This quadratic equation has two solutions, $1+\sqrt{2}$ and $1-\sqrt{2}$, but only the first one is positive, hence it is the value of the continued fraction.

A similar argument shows that any periodic continued fraction is the solution of a quadratic equation, which means periodicity is quite rare. In infinite decimals, periodicity is even more special: it occurs only for rational numbers. The reason for this can be seen by taking a random example, such as

$$x = 0.235717171717171\cdots$$

First we shift the decimal point three places (the length of the segment 235 that comes before the periodic part) by multiplying by 1,000:

$$1000x = 235.717171717171\cdots$$

Then we shift by another two places (the length of the "period" 71) by multiplying by 100:

$$100000x = 23571.717171717171\cdots$$

Finally we subtract $1000x$ from $100000x$, to cancel the parts after the decimal point:

$$100000x - 1000x = 23571 - 235, \quad \text{that is,} \quad 99000x = 23336,$$

so x is the rational number $\frac{23336}{99000}$.

1.7 Equal Temperament

Now it is time to return to the music question raised in Section 1.1: can a sum of fifths equal a sum of octaves? The answer is no, because summing m fifths corresponds to multiplying frequency by $(3/2)^m$, while summing n octaves corresponds to multiplying frequency by 2^n. If the two sums of intervals are equal, then

$$\left(\frac{3}{2}\right)^m = 2^n,$$

and therefore, multiplying both sides by 2^m,

$$3^m = 2^m \times 2^n.$$

But this is impossible, because the left side is odd (being a product of odd numbers) and the right side is even (being a product of even numbers). Thus the octave and the fifth have no "common measure" and the Pythagorean attempt to divide the octave into natural steps using rational numbers was doomed from the start—though the Pythagoreans may never have noticed.

If one wants a scale with the sweetest harmonies, given by the octave and the fifth, then it is impossible to include both by dividing the octave into equal steps. An *approximation* to such a scale is possible, since the sum of 12 fifths is close to seven octaves. But (as mentioned in Section 1.1) the difference between 12 fifths and seven octaves is a noticeable fraction of a step, so some compromise is necessary. If the perfect fifth is a note in the scale, then some steps in the scale that are supposed to be equal will not be. If we insist on equal steps, then the fifth note of the scale will be imperfect—its frequency ratio with the first note will not be exactly 3:2.

This problem bedevilled music in both the East and West. In China, music was generally based on a different scale (dividing the octave into five steps instead of seven—corresponding to the black keys on the piano), but the Chinese also tried to build their scale from a cycle of 12 perfect fifths, and ran into the same problem. Amazingly, the compromise solution that builds the scale from 12 equal parts of an octave ("semitones") was proposed in the East and West almost simultaneously: by Zhu Zaiyu in China and by Simon Stevin in Holland. Their discoveries are generally given the dates 1584 and 1585, respectively, but neither is known exactly, so it seems fair to consider the discoveries independent. (An additional fact, which makes coincidence almost unbelievable, is that the 1580s were also the decade when Jesuit missionaries arrived in China, bringing western mathematics to the region for the first time.)

Both Zhu and Stevin realized that adding 12 equal semitones to form an octave amounts to doing 12 multiplications, by the same number, to achieve multiplication by 2. Therefore, one semitone corresponds to multiplication by the 12th root of 2, or $2^{1/12}$ as Stevin wrote it. In both eastern and western scales the crucial note is seven semitones above the first, hence at a frequency ratio of

$$\left(2^{1/12}\right)^7 = 2^{7/12} = 1.49831\cdots$$

This is certainly close to the perfect fifth ratio of 1.5, though some people can hear the difference. The advantage of this imperfect fifth, however, is that 12 of them make exactly seven octaves, and a cycle of 12 fifths hits all notes of the new scale exactly.

The system of equal semitones, or *equal temperament*, obviously has great simplicity and mathematical beauty. However, it was not immediately popular with musicians. In western music it did not become widespread until the nineteenth century. Bach's *Well-Tempered Clavier* of 1722 is often thought to be an advertisement for equal temperament, but more likely it was intended to show off an older system. There are many systems of "just temperament" that retain perfect fifths while coming closer to equal semitones than the Pythagorean scale.

For a modern mathematician, of course, equal temperament is interesting because the basic ratio $2^{1/12}$ is *irrational*. It must be, because its sixth power is the irrational number $\sqrt{2}$, while powers of rational numbers are rational. Thus *equal temperament rejects all integer ratios in music except the ratio 2:1 for the octave.* For a long time, the Chinese do not seem to have been aware of irrational numbers. According to Joseph Needham's *Science and Civilisation in China* [40, vol. III, p. 90]:

> The Chinese mathematicians ... seem to have been neither attracted nor perplexed by irrationals, if indeed they appreciated their separate existence.

Zhu Zaiyu may be an exception to this characterization. In 1604, he wrote a *New Explanation of the Theory of Calculation* in which he derived values of the roots of 2 needed for equal temperament. He was so attracted to $\sqrt{2}$ that he used nine abacuses to compute it to 25-digit accuracy!

As we know from Section 1.3, Stevin appreciated the existence of irrationals, and aggressively asserted their equal rights with rational numbers. In his *Theory of the Art of Singing* he seized the fresh opportunity to mock the unbelievers in irrational numbers. He even claimed that the equal-tempered fifth must *sound* sweet because $2^{7/12}$ is such a wonderful number!

> Now someone might wonder, according to the ancient view, how the sweet sound of the fifth could consist in so unspeakable, irrational, and inappropriate a number. To this we might give a detailed answer. However, ... it is not our intention to teach those with the inexpressible irrationality and inappropriateness of such a misunderstanding the expressibility, rationality, appropriateness, and natural wonderful perfection of these numbers ... .[1]

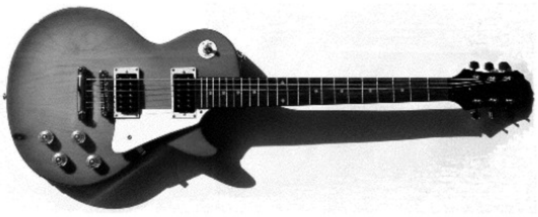

Figure 1.11. Frets on a guitar.

---

[1] Translation by Adriaan Fokker in *The Principal Works of Simon Stevin*, vol. V, p. 441. (Modified by the author with the assistance of Hendrik Lenstra.)

This is Pythagoreanism reborn! “All is number” again, but the meaning of “number” has been expanded to include irrationals.

In conclusion, we should mention that both Zhu and Stevin saw musical ratios the way Pythagoras did—as ratios of string lengths—because they were writing a few decades before Beeckman discovered that pitch corresponds to frequency of vibration. Thus the most direct realization of their idea today is seen in the placement of frets on an instrument such as a guitar (see Figure 1.11). Sliding the finger from one fret to the next changes the length of the vibrating string by a factor of $2^{1/12}$.

data 14.08.2019

impossible... / Stillwell ... (2006).pdf *

Pág. 10 (30) -

1.6. The Continued Fraction for $\sqrt{2}$.

67) 20190815

Riemann Hypothesis (Proof)

Sir Michael Francis Atiyah

www.heidelberg-laureate-forum.org/
blog/video/lecture-monday-
september-24-2018-sir-michael-
francis-atiyah/

Documents:

1) The Fine Structure Constant
17 pages

2) The Riemann Hypothesis
5 pages.

Source:

20190815
20190815-02

data 15.08.2019

S T Q Q S S D ×

other ref.:

a) https://arxiv.org/pdf/math/0307136
Geometric Proof of Riemann
Hypothesis

b) https://phys.org/news/2018-09-
math-greatest-mysteries-
riemann-hypothesis.html

c) www.scientificamerican.com/
article/are-mathematicians-
finall/

d) www.aps.org/publications/
apsnews/201311/physicshistory.cfm

e) www.gizmodo.com/mathematicians-
skeptical-of-supposed-million-
dollar-pro-1829301425

f) www.preposterousuniverse.com/
blog/2018/09/25/atiyah-and-the-
fine-structure-constant/

(20190326).
email
sent

(15)
data 18·08·2019
S T Q [✓ Q] S S D

per Sean Carroll
... thanks to Weinberg and Salam that the weak and electromagnetic forces are unified into the electroweak theory.

~~68~~) Abstract: heidelberg-laureate-forum.org

The Riemann Hypothesis is a famous unsolved problem dating from 1859.
A simple proof using a radically new approach.
It is based on work of

von Neumann (1936),
Hilzebruch (1954) and
Dirac (1928).

---

68) Fernando J. Corbató
July 1, 1926.
July 12, 2019.
Book: Spheroidal Wave functions: including Tables of Separation Constants and Coefficients → 1956.

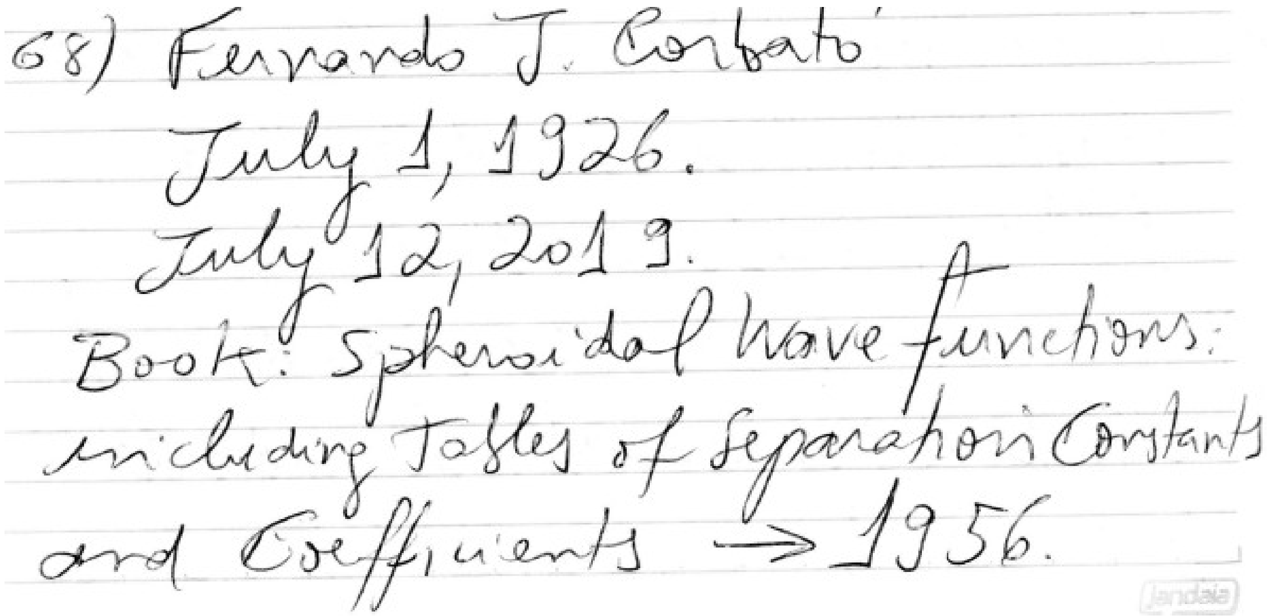

68) Fernando J. Corbató
July 1, 1926.
July 12, 2019.
Book: Spheroidal Wave functions: including Tables of Separation Constants and Coefficients → 1956.

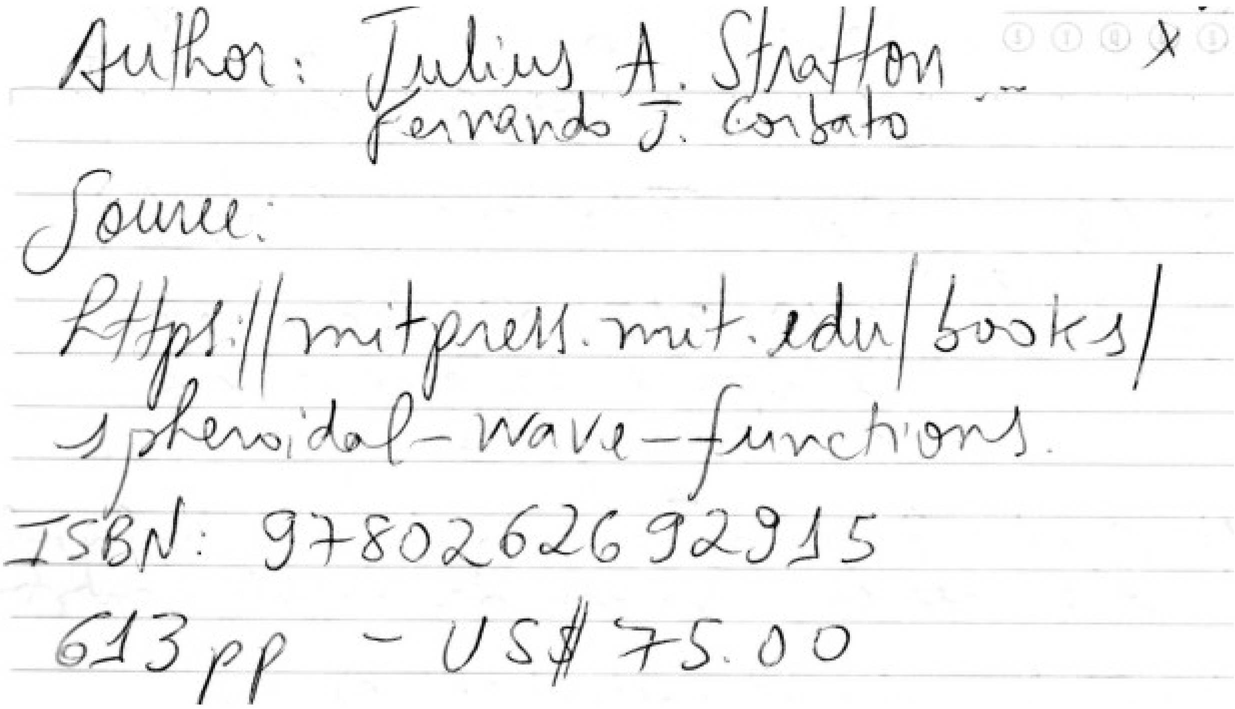

Author: Julius A. Stratton ...
Fernando J. Corbato
Source:
https://mitpress.mit.edu/books/spheroidal-wave-functions.
ISBN: 9780262692915
613 pp - US$ 75.00

@6 (-& .- - -/ ' ! - 5 K &"! !
* !' C 2 % 5) 6 $"

6 (-& .- - -/ ' ! - 5 K &"! !
* !' C 2 % *** * !' R!+ C 26 R 2 ***
% ;6 $"

"6 (-& .- - -/ ' ! - 5 ? R ? $ "! $"

;

(-& .- - -/ ' ! - ; &) < $"

? " ' ! " ! !

;

!6 (-& .- - -/ ' ! - ; 20 ! 0 @ $"

@6 (-& .- - -/ ' ! - ; 20 ! 0 @ $"

7' ;%) 6 < " + 9 ' ! <5

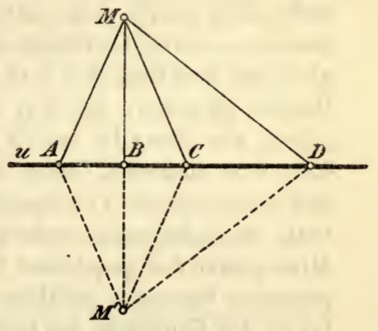

7' 5%) 6 9 ' ! 5

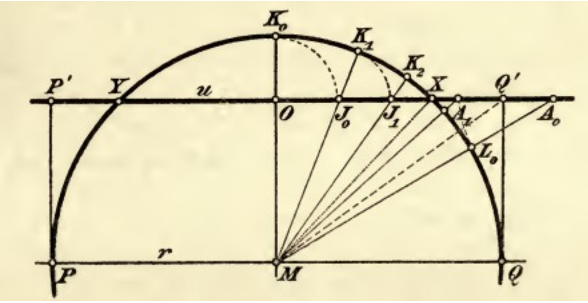
K₀
K₁
K₂
P'
Y
u
X
Q'
O
J₀
J₁
A₁
A₀
L₀
r
P
M
Q

6 (-& .- - -/ ' ! - ; 20 ! 0 @ $"

;

(-& .- - -/ ' ! - ; K (! 2 8 2 !
= !@ $ (! @ "' J + 2 % 6 $"

7' % ;6

C) 5 C ! ' ! 2 $ " @
$ G$ @ ! G$! " ! @
!

7' % 56

; <# ?! @ $ " " ! $ + "2 2$ @ '
! ' 0 * ! " > %?! @ ; <6 ! " + $!"
@2 ! % ; ; 6 ! !" "2 ' @

Suppose $\Pi(AP) > \Pi(AR)$, then there is a line through P, say $\overrightarrow{PT}$ with $\angle RPT \cong \angle ARS$. Then $\angle APT < \angle APQ$ and $\overrightarrow{PT}$ meets $\overleftrightarrow{AB}$, and hence $\overleftrightarrow{RS}$. By the Alternate Interior Angles Theorem (Proposition I.27), however, $\overleftrightarrow{PT}$ does not meet $\overleftrightarrow{RS}$.

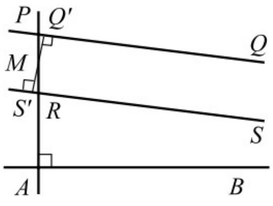

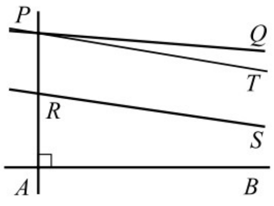

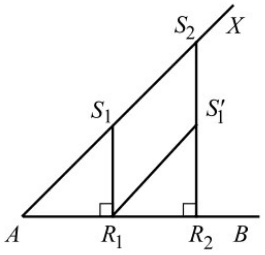

(3) In fact, any acute angle is the angle of parallelism for some length. To see this, fix the angle $\angle BAX$. Suppose R_1S_1 is perpendicular to $\overleftrightarrow{AB}$ at R_1 and S_1 lies on $\overrightarrow{AX}$. The triangle $\triangle AR_1S_1$ has some angle defect, say $\delta > 0$. Mark R_2 along $\overleftrightarrow{AB}$ with $R_1R_2 \cong AR_1$. If the perpendicular to $\overleftrightarrow{AB}$ from R_2 meets $\overrightarrow{AX}$ at S_2, then $R_2S_2 > R_1S_1$ and we can construct the point S_1' along R_2S_2 such that $R_2S_1' \cong R_1S_1$. It follows that $\triangle AR_2S_2$ contains two copies of $\triangle AR_1S_1$ and so the angle defect of $\triangle AR_2S_2$ is at least 2δ (see Exercise 3.2). If every perpendicular to $\overleftrightarrow{AB}$ within $\angle BAX$ meets $\overrightarrow{AX}$, then we can iterate this procedure to construct a triangle with angle defect greater than $n\delta$ for any n and so greater than two right angles, which is impossible. Hence, some perpendicular to $\overleftrightarrow{AB}$ does not meet AX and $\angle BAX$ is the angle of parallelism for some length.

The hyperbolic plane

Without the millstone of proving Postulate V around their necks, Gauss, J. Bolyai, and Lobachevskiĭ developed the properties of the plane with the new definition of parallelism and **HAA**. Auxiliary constructions and curves led to new tools and results, especially when developed in space.

We introduce properties and objects that do not depend on the choice of **HAA** or **HRA**. We call such a property or object **absolute**. The following collections of lines are absolute. A **pencil of lines** is a set of lines satisfying one of three properties: There is the pencil of all lines through a given point P, denoted $\mathbf{P}_P$; the pencil of all lines parallel to a given line l, denoted $\mathbf{P}_l$; and the pencil of all lines perpendicular to a given line, m, denoted $\mathbf{P}_m^{\perp}$.

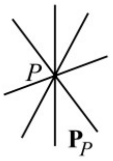

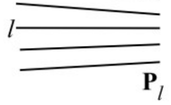

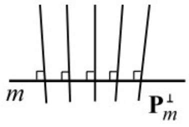

The pencil of lines $\mathbf{P}_l$ parallel to a given line l can be thought of as a limit of pencils of lines $\mathbf{P}_X$ through a point X, as X travels along l to “infinity.” Notice that every point in the plane lies on a unique line in the pencils of parallels or perpendiculars, and every point other than the given point P lies on a unique line in $\mathbf{P}_P$.

A choice of three distinct points determines each sort of pencil of lines.

Theorem 4.4. *Given distinct points X, Y, and Z, the lines l, m, and n that are the perpendicular bisectors of the segments XY, YZ, and XZ, respectively, lie in a pencil of lines.*

Proof (Shirikov 1964): Consider the triangle $\triangle XYZ$. If l and m meet, then they meet at the center of a circle circumscribed around the triangle. Furthermore, the line n passes through the same point of intersection. We leave it to the reader to prove this.

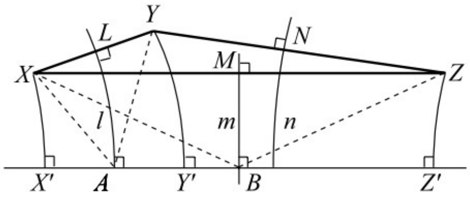

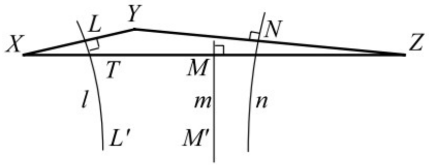

Suppose that l and m have a common perpendicular $\overleftrightarrow{AB}$. Let $L = l \cap XY$, $M = m \cap XZ$, and $N = n \cap YZ$, the midpoints of the segments XY, XZ, and YZ. Construct segments XX', YY', and ZZ' perpendicular to $\overleftrightarrow{AB}$ through X, Y, and Z, respectively. Join X to B and Z to B. Then by Side–Angle–Side, $\triangle BMZ \cong \triangle BMX$. It follows from $\angle XBM \cong \angle ZBM$ that $\angle XBX' \cong \angle ZBZ'$. By Angle–Angle–Side we find $\triangle XBX' \cong \triangle ZBZ'$ and so $XX' \cong ZZ'$. By a similar construction joining X to A and Y to A we find $XX' \cong YY'$. Thus $Y'YZZ'$ is a Saccheri quadrilateral and the line n perpendicular to YZ through the midpoint of YZ meets the side $\overleftrightarrow{Y'Z'} = \overleftrightarrow{AB}$ in a right angle. Thus l, m, and n lie in $\mathbf{P}^{\perp}_{\overleftrightarrow{AB}}$, the pencil of lines perpendicular to $\overleftrightarrow{AB}$.

Finally, if l and m are parallel, then n cannot meet either l or m nor will n share a common perpendicular with l or m. It remains to show that l, m, and n are parallel in the same direction. By Pasch’s Axiom, a line entering a triangle along a side must exit through another side. The midpoint on the longest side separates the other lines from each other and so all the lines pass through the longest side. Let $l = \overleftrightarrow{LL'}$ meet XZ at T. Since $\angle LTZ$ is exterior to the right triangle $\triangle TLX$, this angle is obtuse by the Exterior Angle Theorem. Therefore the supplement $\angle L'TZ$ is acute. Since l is parallel to m and m is perpendicular to XZ, $\angle L'TZ = \Pi(TM)$ and so l and m are parallel away from XZ in the direction of L' and M'. The analogous argument shows n parallel to m in the same direction. Thus l, m, and n are elements of $\mathbf{P}_l$. ■

Definition 4.5 (Gauss 1831). *Given a pencil of lines* **P**, *two points X and Y are said* **to correspond with respect to P** *if the line in the pencil that passes through* *the midpoint of segment XY is perpendicular to XY. When the pencil of lines is the set of lines parallel to a given line, the set of all points corresponding to a given point X is called a* **horocycle** *through X (*Parazykel *in Gauss).*

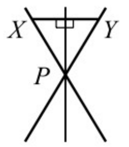

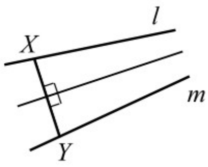

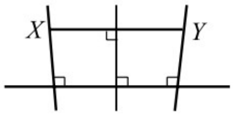

Points that correspond are related by a reflection of the plane across a line in the pencil. For a pencil of lines through a given point, a direct consequence of the proof of Euclid's Proposition I.6 (the converse of *Pons Asinorum*) is that the set of points corresponding to a given point forms a circle. From this point of view, a horocycle may be thought of as a "circle with center at infinity." The set of points corresponding to a given point with respect to $\mathbf{P}_m^{\perp}$ is the curve of points equidistant to the line m. The properties of the Saccheri quadrilateral may be used to explore the properties of this curve.

We leave the proof of the following proposition to the reader.

Proposition 4.6. *A horocycle is a line if and only if* **HRA** *holds.*

Thus, a new feature of the plane with **HAA** is the set of horocycles in the plane. An immediate consequence of Theorem 4.4 is the independence of the choice of the point X in the definition of a horocycle through X.

Corollary 4.7. *If Y and Z lie on the horocycle through X determined by $\mathbf{P}_l$, then X and Y lie on the horocycle through Z determined by $\mathbf{P}_l$.*

The points on a horocycle can be characterized by the angle of parallelism: If X and Y are on the horocycle, the perpendicular bisector m of the segment XY lies in the pencil of parallel lines that determine the horocycle. The corresponding angles at X and Y have measure $\Pi(XM) = \Pi(YM)$, where M is the midpoint of XY.

If m meets the horocycle at a point Q, then the tangent to the horocycle at Q is perpendicular to m as can be seen by letting the chord XY shrink to Q. Thus, as in the case of radii to a circle, the lines in the pencil $\mathbf{P}_l$ meet the horocycle at right angles.

We say that two horocycles are **concentric** if they are determined by the same pencil of parallel lines. Segments of horocycles have lengths that are, as usual, limits of polygonal approximations.

Lemma 4.8.

(1) *To equal chords there correspond equal arcs of the horocycle. To the longer chord there corresponds a longer arc.*

(2) *Segments of lines in the pencil bounded by two concentric horocycles are congruent.*
(3) *The arcs of concentric horocycles bounded by two lines in the pencil decrease in the direction of parallelism.*

PROOF:

(1) Suppose X, Y and X', Y' lie on a horocycle determined by $\mathbf{P}_l$ and that $XY \cong X'Y'$. Denote the arc along a horocycle between points X, and Y by $\overset{\frown}{XY}$. The goal is to show that the horocycle segment $\overset{\frown}{XY}$ is congruent to $\overset{\frown}{X'Y'}$. Let M be the midpoint of XY and m the line in $\mathbf{P}_l$ through M which is perpendicular to XY and suppose m meets the horocycle at Q. By construction $\triangle XMQ \cong \triangle YMQ$ and $XQ \cong YQ$. The angle $\angle XQM$ is congruent to $\Pi(XQ/2)$ because X and Q correspond.

Let M' be the midpoint of $X'Y'$ and m' the line in $\mathbf{P}_l$ through M' perpendicular to $X'Y'$. On m' let Q' be away from the direction of parallelism with $M'Q' \cong MQ$. The triangles $\triangle XMQ$ and $\triangle X'M'Q'$ are congruent so $XQ \cong X'Q'$ and $\angle X'Q'M' \cong \Pi(XQ/2) \cong \Pi(X'Q'/2)$ and so the line through the midpoint of $X'Q'$ perpendicular to $X'Q'$ lies in $\mathbf{P}_l$. Thus Q' corresponds to X' and Q' lies on the horocycle. The arc $\overset{\frown}{XY}$ is approximated by the line segments XQ and QY and $\overset{\frown}{X'Y'}$ by the congruent $X'Q'$ and $Q'Y'$. To complete the proof, take the midpoints of the approximating segments and make the comparisons. Iterated halving gives the arcs in the limit.

(2) Suppose $\overleftrightarrow{XA}$ and $\overleftrightarrow{YB}$ are in $\mathbf{P}_l$ and X and Y lie on one horocycle and A and B on a concentric horocycle determined by $\mathbf{P}_l$. We want to show that $XA \cong YB$. Let m be the line through the midpoint M of XY and perpendicular to XY. Suppose m meets AB at a point N. Join X to N and Y to N to form triangles $\triangle XMN$ and $\triangle YMN$. These triangles are congruent by Side–Angle–Side and so $XN \cong YN$. It follows that:

$$\angle NXA \cong \Pi(XM) - \angle MXN \cong \Pi(YM) - \angle NYB \cong \angle NYB.$$

It is also the case that $\angle NAX \cong \pi - \Pi(AB/2) \cong \angle NBY$. By the congruence criterion Angle–Angle–Side, triangles $\triangle NAX$ and $\triangle NBY$ are congruent and so $XA \cong YB$. It also follows that m is the perpendicular bisector of AB.

(3) This statement follows from the approximations in part 1 and the fact that the angle of parallelism decreases with increasing length (Property 2 of $\Pi(x)$). ■

With these ideas we arrive at the following result.

Theorem 4.9. *The ratio of the lengths of concentric arcs of two horocycles intercepted between two lines in the pencil is expressible in terms of an exponential function of the distance between these arcs.*

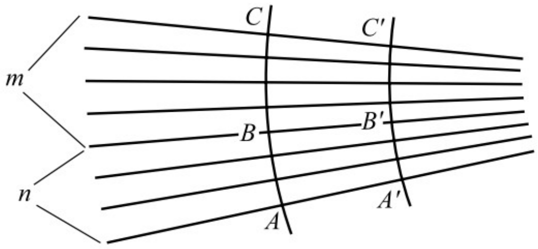

PROOF: We first show that the ratio depends only on the distance. Denote the length of an arc along the horocycle between points A, B by $l(\overset{\frown}{AB})$. First suppose $l(\overset{\frown}{BC}) = (m/n)l(\overset{\frown}{AB})$. Then the congruent segments along the parallels AA', BB', and CC' induce equal subdivisions of $\overset{\frown}{A'B'}$ and $\overset{\frown}{B'C'}$. We then have $l(\overset{\frown}{B'C'}) = (m/n)l(\overset{\frown}{A'B'})$, and so:

$$\frac{l(\overset{\frown}{AB})}{l(\overset{\frown}{A'B'})} = \frac{(m/n)l(\overset{\frown}{BC})}{(m/n)l(\overset{\frown}{B'C'})} = \frac{l(\overset{\frown}{BC})}{l(\overset{\frown}{B'C'})};$$

therefore, the ratio $l(\overset{\frown}{AB})/l(\overset{\frown}{A'B'})$ depends only on the length of AA'. If $l(\overset{\frown}{AB})$ and $l(\overset{\frown}{BC})$ are related by an irrational number, a limiting process obtains the desired ratio.

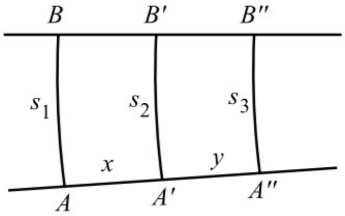

Suppose $s_1 = l(\overset{\frown}{AB})$, $s_2 = l(\overset{\frown}{A'B'})$, and $s_3 = l(\overset{\frown}{A''B''})$. We can write $s_1/s_2 = f(x)$ and $s_2/s_3 = f(y)$, where x and y denote the lengths of the line segments AA' and $A'A''$, respectively. It follows that $s_1/s_3 = f(x+y)$ and we obtain the equation:

$$f(x+y) = f(x)\cdot f(y).$$

Assuming the differentiability of $f(x)$ for the moment we get

$$f'(x) = \lim_{\Delta x\to 0}\frac{f(x+\Delta x)-f(x)}{\Delta x} = f(x)\cdot\lim_{\Delta x\to 0}\frac{f(\Delta x)-1}{\Delta x}.$$

By the definition of $f(x)$, as Δx goes to 0, $f(\Delta x)$ goes to 1, and we write, for some value of k,

$$\lim_{\Delta x\to 0}\frac{f(\Delta x)-1}{\Delta x} = \frac{1}{k}.$$

Then $f'(x) = (1/k)f(x)$ and so $f(x) = e^{x/k}$ or $l(\widehat{AB}) = l(\widehat{A'B'})e^{l(AA')/k}$. ∎

The value $1/k$ is chosen to anticipate the more general development in Chapter 12 where it shows up again. Theorem 4.9 offers us a first step into the analytic properties of the plane under **HAA**. The factor of the exponential function will lead us to the *hyperbolic trigonometric functions*, the elementary properties of which we review next.

In the study of the integral giving the area under a semicircle, $\int \sqrt{a^2 - x^2}\,dx$, the familiar trigonometric functions $\sin(x)$, $\cos(x)$, and $\tan(x)$ play a role as parametrizing functions. Analogously, the functions $\sinh(x)$, $\cosh(x)$, and $\tanh(x)$ play the same roles in the study of the integral giving the area under a hyperbola, $\int \sqrt{x^2 - a^2}\,dx$. In 1768, Lambert produced a comprehensive study of the hyperbolic trigonometric functions containing the work of Euler and the Bernoullis (Lambert 1768). Similar developments were made by V. Riccati (1701–75) with Saladini for studying the solutions to cubics.

To define the hyperbolic trigonometric functions, formally introduce a complex variable into the Taylor series for e^x to get:

$$\begin{aligned} e^{iy} &= 1 + iy - \frac{y^2}{2} - \frac{iy^3}{3} + \frac{y^4}{4} - \cdots \\ &= \left(1 - \frac{y^2}{2} + \frac{y^4}{4} - \cdots\right) + i\left(y - \frac{y^3}{3} + \frac{y^5}{5} - \cdots\right) \\ &= \cos y + i \sin y. \end{aligned}$$

It follows from these formulas that:

$$\cos y = \frac{e^{iy} + e^{-iy}}{2}, \quad \sin y = \frac{e^{iy} - e^{-iy}}{2i} \quad \text{and} \quad \tan y = \frac{\sin y}{\cos y} = \frac{1}{i}\frac{e^{iy} - e^{-iy}}{e^{iy} + e^{-iy}}.$$

By analogy with these formulas, we define the real-valued functions:

$$\cosh x = \frac{e^x + e^{-x}}{2}, \quad \sinh x = \frac{e^x - e^{-x}}{2}, \quad \text{and} \quad \tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}}.$$

From the power series, notice that $\cos(iy) = \cosh y$, and $\sin(iy) = i \sinh(y)$. The fundamental relation for the hyperbolic trigonometric functions is given by:

$$\cosh^2 x - \sinh^2 x = 1.$$

The convergence theorems that support our formal manipulations are classical and can be found in any good complex analysis book, for example, Ahlfors (1979).

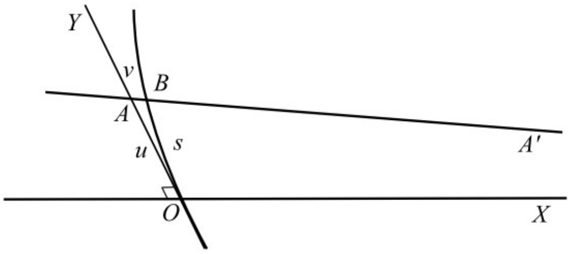

Under **HAA**, a horocycle differs from a straight line. We next seek to measure how much it differs. Fix some line segment RR_1 as a unit length with which we can assign a positive real number length to all other line segments. Let $\overleftrightarrow{OX} = l$ denote a line and consider the horocycle through O determined by the pencil of parallels $\mathbf{P}_l$ in the direction $\overrightarrow{OX}$. Let $\overleftrightarrow{OY}$ denote the line perpendicular to l through O. If we measure a length u along OY to a point A, then there is a line $\overleftrightarrow{AA'}$ in $\mathbf{P}_l$ with A' toward X and a point B on the line $\overleftrightarrow{AA'}$ and on the horocycle through O. Denote the length of AB by v. We apply Theorem 4.9 to relate u and v.

Lemma 4.10. *In the notation of the previous paragraph,* $e^{v/k} = \cosh(u/k)$.

Proof: In order to apply Theorem 4.9, we need a family of strategically chosen concentric horocycles. To construct them we first obtain a reference line: Among the properties of the angle of parallelism, it was shown that any acute angle is $\Pi(x)$ for some x. Let OC' denote a segment of length x_0 for which $\Pi(x_0) = \pi/4$. Then there is a line $\overleftrightarrow{TT'}$ that is parallel to the perpendicular lines $\overleftrightarrow{OX}$ and $\overleftrightarrow{OY}$. Also $\overleftrightarrow{TT'}$ is in $\mathbf{P}_l$. The horocycle through O determined by $\mathbf{P}_l$ meets $\overleftrightarrow{TT'}$ at a point C. Denote the length of $\overset{\frown}{OC}$ by σ.

Consider the diagram in which we use reflections across lines to construct various comparable angles and figures.

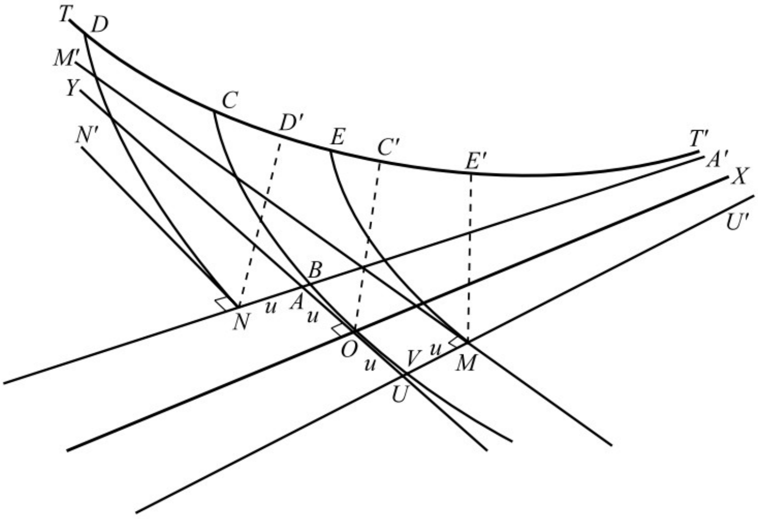

Let $\overleftrightarrow{NN'}$ be perpendicular to $\overleftrightarrow{AA'}$ and parallel to $\overrightarrow{OY}$, that is, choose N with $\angle NAY = \Pi(AN)$. Since $\angle OAA' = \Pi(OA) = \Pi(u)$ and $\angle NAY = \angle OAA'$ we have $AN \cong OA$. Similarly let U be on $\overrightarrow{YO}$ with $OU \cong OA$, $U \neq A$, and $\overleftrightarrow{UU'}$ the parallel to $\overrightarrow{OX}$ through U in the direction of X. Let UM have length u and construct $\overleftrightarrow{MM'}$ perpendicular to $\overleftrightarrow{UU'}$ with $\overleftrightarrow{MM'}$ parallel to $\overrightarrow{OY}$. The segments AN and UM determine concentric horocycles $\overset{\frown}{ND}$, $\overset{\frown}{OC}$, and $\overset{\frown}{ME}$. The perpendiculars to $\overleftrightarrow{TT'}$ from N, O, and M, determine points D', C', and E', respectively. By the construction we see that $\overleftrightarrow{TT'}$ is parallel to $\overleftrightarrow{AA'}$ and $\overleftrightarrow{NN'}$ and so $\angle D'NA' \cong \angle D'NN' \cong \Pi(x_0) = \pi/4$. Similarly, $ND' \cong OC' \cong ME'$ are all of length x_0. It follows from Lemma 4.8 that $\overset{\frown}{ND} \cong \overset{\frown}{OC} \cong \overset{\frown}{ME}$. Their common length is σ.

We can apply the formula in Theorem 4.9, that is, the ratio of the lengths of concentric arcs of two horocycles is given by $l(\overset{\frown}{ZW}) = l(\overset{\frown}{Z'W'})e^{-t/k}$, where t is the distance between the arcs. If B denotes the point along $\overleftrightarrow{AA'}$ where $\overset{\frown}{OC}$ crosses, then let V denote the point corresponding to B along $\overleftrightarrow{UU'}$. Then $l(BN) = u + v$. Let $s = l(\overset{\frown}{OB})$, the length between O and B along the horocycle; we then have:

$$l(\overset{\frown}{ND}) = l(\overset{\frown}{OC}) = \sigma = (\sigma - s)e^{(u+v)/k} \quad \text{or} \quad \frac{\sigma - s}{\sigma} = e^{-(u+v)/k}.$$

Horocycle arcs $\overset{\frown}{VC}$ and $\overset{\frown}{ME}$ are concentric and $l(VM) = u - v$. Since $l(\overset{\frown}{VC}) = \sigma + s$ and $l(\overset{\frown}{ME}) = \sigma$, then we also have $\sigma + s = \sigma e^{(u-v)/k}$, or $(\sigma + s)/\sigma = e^{(u-v)/k}$. Thus,

$$2 = \frac{\sigma + s}{\sigma} + \frac{\sigma - s}{\sigma} = e^{(u-v)/k} + e^{-(u+v)/k}.$$

Multiplying through by $e^{v/k}$ gives $e^{v/k} = \dfrac{e^{u/k} + e^{-u/k}}{2} = \cosh(u/k)$. This proves the lemma. ■

Corollary 4.11. $s = \sigma \tanh(u/k)$.

PROOF: By the proof of Lemma 4.10 we see that:

$$2\frac{s}{\sigma} = \frac{\sigma + s}{\sigma} - \frac{\sigma - s}{\sigma} = e^{(u-v)/k} - e^{-(u+v)/k} = (e^{u/k} - e^{-u/k})e^{-v/k}.$$

Therefore,

$$\frac{s}{\sigma}\cosh(u/k) = \frac{s}{\sigma}e^{v/k} = \frac{e^{u/k} - e^{-u/k}}{2} = \sinh(u/k),$$

and so $s = \sigma \tanh(u/k)$. ■

These analytical relations will play a key role in building non-Euclidean trigonometry. To develop further tools we introduce three dimensions.

Digression: Neutral space

In Chapter 1 we saw that the trigonometric formulas for triangles on the sphere could be proved by using the fact that the sphere is embedded in three dimensional space. We adopt this approach more generally by considering three dimensional space, where we find points, lines, and planes for which Hilbert's Axioms of Incidence hold, together with the Postulates I–IV of Euclid (Book XI of *The Elements*). We call such a space *neutral space*.

Without using Postulate V, define *parallel lines* in space to be two lines ℓ_1, ℓ_2 that are coplanar, and, in a chosen direction on that plane, satisfy Definition 4.1.

In order to develop the geometry of neutral space we need to explore certain relations between pairs of lines, lines and planes, and pairs of planes. According to the assumptions of an incidence geometry (Definition 2.1), if we are given three noncollinear points, P, Q, and R, or a line l and a point P not on l or two intersecting lines l and m, then each group uniquely determines a plane. We denote this relation by $T = \text{Plane}(P,Q,R)$ or $T = \text{Plane}(P,l)$ or $T = \text{Plane}(l,m)$.

Our goal in this digression is to prove Theorem 4.25, an analogue of Playfair's Axiom but in a higher dimension. For Playfair's Axiom we are given a point P not on a line m, from which we obtain a unique line m' through P with m' parallel to m. One dimension higher, the role of the point is played by a line l, which is parallel to a plane T from which we obtain a unique plane T' containing l and parallel to T. To prove this assertion we generalize a special case of Euclid's proof of Playfair's Axiom, Proposition I.31: First construct the unique perpendicular PA through P to the given line l (Proposition I.12), then construct the unique perpendicular m to PA through P. The line m is parallel to l under the assumption of Postulate V.

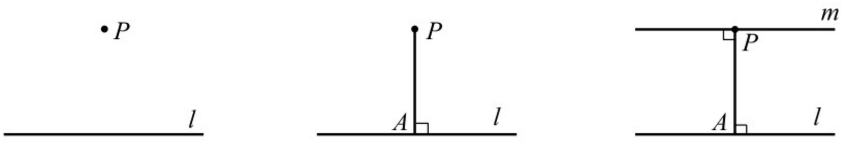

What is remarkable about neutral space is that the same construction, appropriately construed, leads to the analogous result for a line l parallel to a given plane T.

In order to make these constructions in space, we begin by exploring parallelism of lines. Our presentation is inspired by the relevant sections of the book by Borsuk and Smielew (1960).

Lemma 4.12. *Suppose l_1 and l_2 are two lines parallel in space and T_1 and T_2 are two planes with l_1 on T_1, l_2 on T_2, and $T_1 \neq T_2$. If m is the line of intersection of T_1 and T_2, and $l_1 \neq m \neq l_2$, then m is parallel to l_1 and to l_2.*

7' %) 6) % ;6 " G

! 2 # !+ " $! ! !] 2 " ! 0 !
! $!] " $ + 2 2 ! # # ! " $!
" 2 ! O 2 2 C ' 2 !" ! ! $! ! "
3 ! ! " + @ ' !] + !$$2
! ! ! !@ + ' 2
9 >/?4& % 3 & ?*CC* C/ K >/?4& 6

@6 (-& .- - -/ ' ! - ; ?! & ($ G & ! 2
%C " *" 6 8 ! L EE* ' 'EE ! % 5;56 $"

7' <%)6) ! :

6 (-& .- - -/ ' ! - ; ?! & ($ G ! ! 2
8 ! L E* ' 'E ! % 5;56 "O+

;

(-& .- - -/ ' ! - ; E 5 < ? ! "
* $ '

;)
/$ 1

(-& .- - -/ ' ! - ;) < $"

7' % ; 6

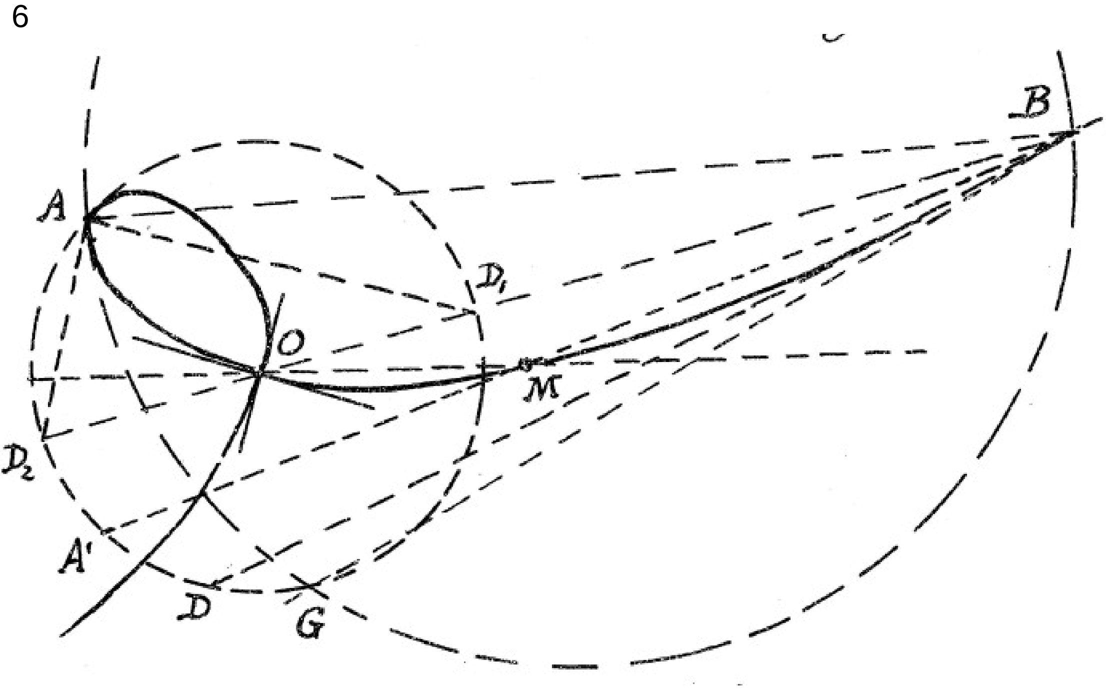
B
A
D₁
O
M
D₂
A'
D
G

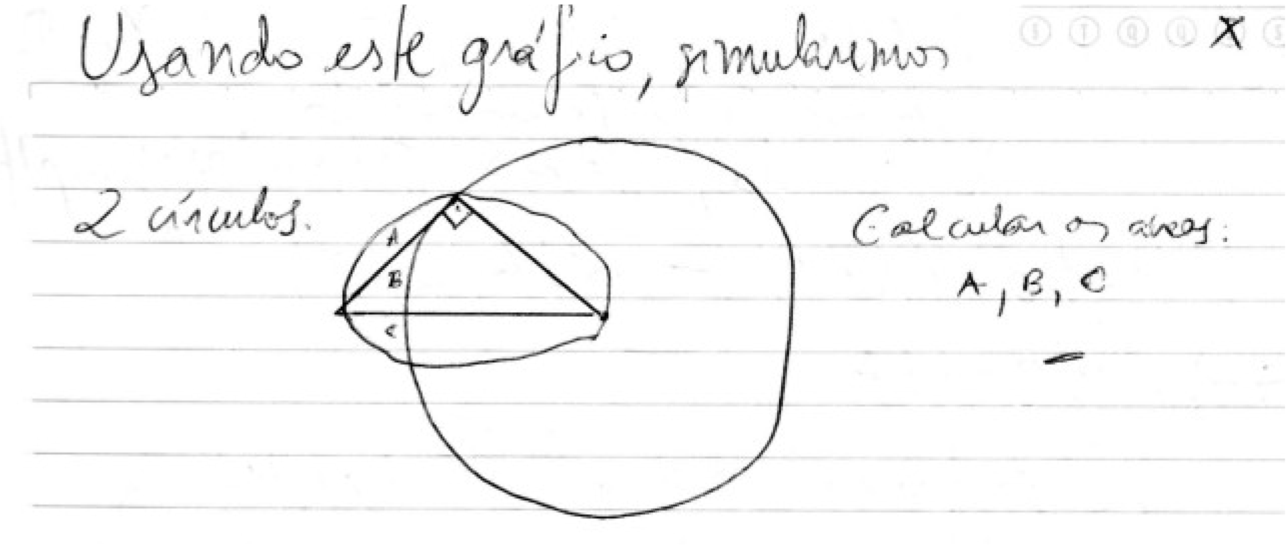
Usando este gráfico, simularemos 2 círculos.

Calcular as áreas: A, B, C

;

(-& .- - -/ ' ! - ; 9 "! ! 2 E ! #
! + 2# ! " C "2 ! 4! J + 2 %)6 $"

7' % ;6 % <6

9 ') C ! + & ! &G c O&2 !

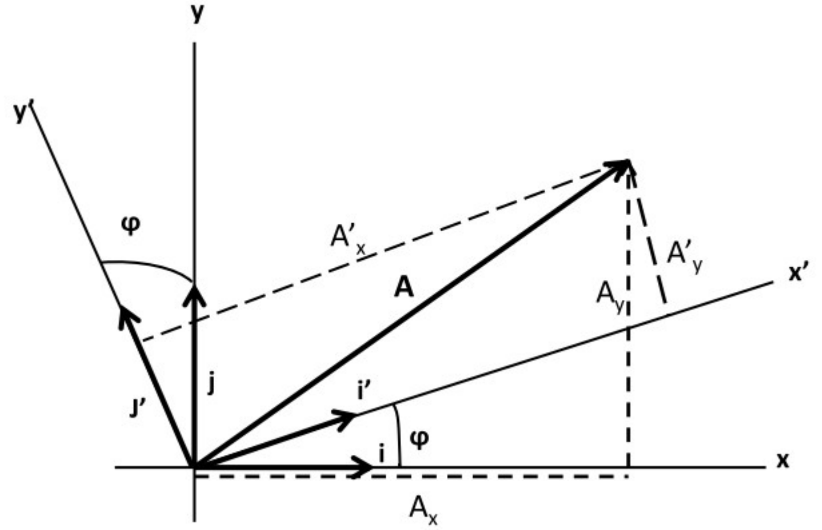

. ^

Figure 2.4 The same vector **A** is written as $\mathbf{i}A_x + \mathbf{j}A_y$ in one frame and as $\mathbf{i}'A'_x + \mathbf{j}'A'_y$ in the other. The dotted lines indicate the components in the two frames.

First we express $\mathbf{i}'$ and $\mathbf{j}'$ in terms of $\mathbf{i}$ and $\mathbf{j}$ using the figure:

$$\mathbf{i}' = \mathbf{i}\cos\phi + \mathbf{j}\sin\phi$$
$$\mathbf{j}' = \mathbf{j}\cos\phi - \mathbf{i}\sin\phi.$$

;

(-& .-   -   -/ ' ! - ;   "! 2   !   !
(!   &"! !   &'! ! > ! !   > ! !   > ! !   > ! !   %   56 $"

7'   %   6   % ;6

($ G 3 @   ! " . !" !   *1 !

/@   ? +   " BBB

;;

* '   / @ ! & ' !

(-& .-   -   -/ ' ! - ;;   )< $"

7' ; %5 6

. !   !   / @ ! & ' !   % 7' ; %5 66

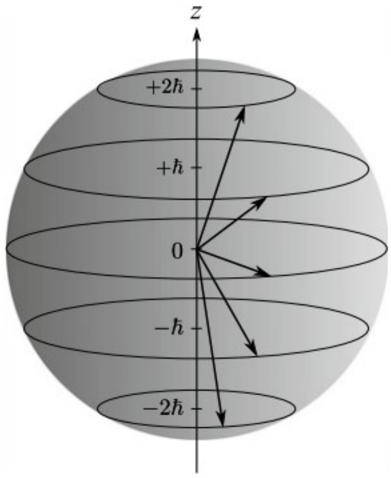

Fig. 1.23 Quantisation of angular momentum for the case of $l = 2$. In this case, there are five angular momentum eigenstates in the multiplet. The vectors from top to bottom illustrate respectively the angular momentum with $L_z = +2\hbar,\ +\hbar,\ 0,\ -\hbar,\ -2\hbar$. The magnitude of the angular momentum is given by $L = \sqrt{2(2+1)}\hbar = \sqrt{6}\hbar$. Adapted from Auletta and Wang (2014, p. 215)

7' ; %5;6   9 '   )

Fig. 1.24 Relation between Cartesian (rectangular) coordinates (x, y, z) and spherical coordinates (r, θ, ϕ) of a point P in three-dimensional space. The versor $\mathbf{e}_r$ makes an angle θ with the Cartesian versor $\mathbf{e}_z$. Adapted from Auletta and Wang (2014, p. 217)

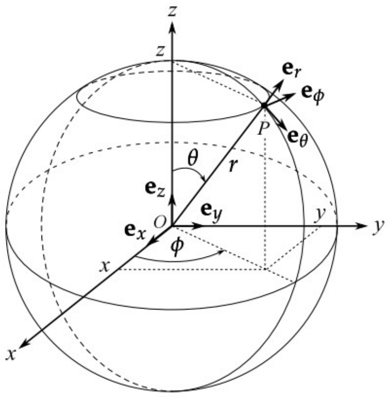
z
z
$\mathbf{e}_r$
$\mathbf{e}_\phi$
P
$\mathbf{e}_\theta$
θ
r
$\mathbf{e}_z$
$\mathbf{e}_y$
O
$\mathbf{e}_x$
y
y
ϕ
x
x

$$r = \sqrt{x^2 + y^2 + z^2}, \quad \theta = \arctan \frac{\sqrt{x^2 + y^2}}{z}, \quad \phi = \arctan \frac{y}{x},$$

;<

N !

data 18.08.2019

78) a) Cercle de Brocard

fr.wikipedia.org/wiki/Cercle_de_Brocard

Obs.: Figures de Brocard, Points de Brocard

b) Théorème de Neuberg

fr.wikipedia.org/wiki/Théorème_de_Neuberg

Obs.: Point de Vecten, Théorème de van Aubel

c) Musselman's Theorem

en.wikipedia.org/wiki/Musselman's_theorem.

d) Courbes (Curvophilie)

web4.ensiie.fr/~gacogne/courbes.pdf

Faz referência ao livro: Courbes Mathématiques
J. Brette.
↳ Source: 20190818/Courbes-Mathématiques,
numéro spécial... djvu

e) List of unsolved in mathematics

www.wikiwand.com/en/List_of_unsolved_problems_in_mathematics

Obs.: In the Bloch Sphere

wikiwand.com/en/Bloch_sphere

data 18.08.2019

... In the Bloch Sphere representation of a qubit, a SIC-POVM (symmetric, informationally complete, positive operator-value measure) - is a special case of a generalized measurement on a Hilbert space, used in the field of quantum mechanics.

form a regular tetrahedron.

Zauner conjecture that analogous structures exist in complex Hilbert spaces of all finite dimensions.

f) Spherical Bernstein's problem

Look: Bernstein's problem

g) Pi

wikiwand.com/en/Pi

Obs: Look

Fourier transform and Heisenberg uncertainty principle.

→ Heisenberg uncertainty principle also contains the number π.

Veja: wikiwand.com/en/Uncertainty_principle

;5 ^^^ C! !

data 19.08.2019

79) Musich Transforms

(11) Descrição dos passos:

1) Point A: Intersection of x Axis and y Axis

$A = (0, 0)$

2) Point B: Point on x Axis

$B = (2, 0)$

3) Circle c: Circle through B with center A.

$c: x^2 + y^2 = 4$

4) Point C: Point on x Axis

$C = (1, 0)$

5) Point D: $D = (1, 1)$

6) Segment f: Segment A, C

$f = 1$

7) Segment g: Segment C, D

$g = 1$

8) Segment h: Segment A, D

$h = 1.4142135623731 0 95$ $(\sqrt{2})$

data 19.08.2019

9) Circle d: Circle through A with center C

$d: (x-1)^2 + y^2 = 1$

10) Point E

$E = (8, -8)$

11) Circle e: Circle through A with center E

$e: (x-8)^2 + (y+8)^2 = 128$

Obs:

Fazer estas construções geométricas por meio de equações matemáticas. (Evite usar simuladores ou geradores gráficos pois eles apresentam uma imprecisão inerente ao problema da Quadratura 2D.)

O Circle e foi obtido pela execução de vários Triângulos com catetos iguais (mesmo valor) e cuja hipotenusa foi projetada num ângulo de 90° (noventa graus) da ponta do triângulo criado seguindo-se os quadrantes 1, 2, 3 e 4 até o 8 (oitavo) círculo, cujos catetos são 8 unidade de medida.

Vejamos a seguir:

data 19.08.2019

No exemplo em questão:

O triângulo ΔACD, tem as seguintes medidas:

$\overline{AC} = 1$, $\overline{CD} = 1$ e $\overline{AD} = \sqrt{2}$

Deste triângulo básico, projetando-se a $\sqrt{2}$ no ponto D com ângulo de 90° (noventa graus), teremos assim, 2 (dois) catetos com o mesmo tamanho cujo valor é $\sqrt{2}$, gerando assim, uma hipotenusa de valor 2.

Desta forma teremos o seguinte:

a) $1 \quad 1 \quad \sqrt{2}$

b) $\sqrt{2} \quad \sqrt{2} \quad 2$

c) $2 \quad 2 \quad 2\sqrt{2}$

d) $2\sqrt{2} \quad 2\sqrt{2} \quad 4$

e) $4 \quad 4 \quad 4\sqrt{2}$

f) $4\sqrt{2} \quad 4\sqrt{2} \quad 8$

g) $8\sqrt{2} \quad 8\sqrt{2} \quad 16$

h) $16 \quad 16 \quad 16\sqrt{2}$

Finalmente, usando-se o ponto E

→

data 19 08 2019

$E = (8, -8)$

Sobre o ponto E trace um círculo cujo raio passe sobre o ponto $A = (0, 0)$

→ Circle through A with center E

equation:

$$(x-8)^2 + (y+8)^2 = 128$$

O gráfico desenhado ficará assim:

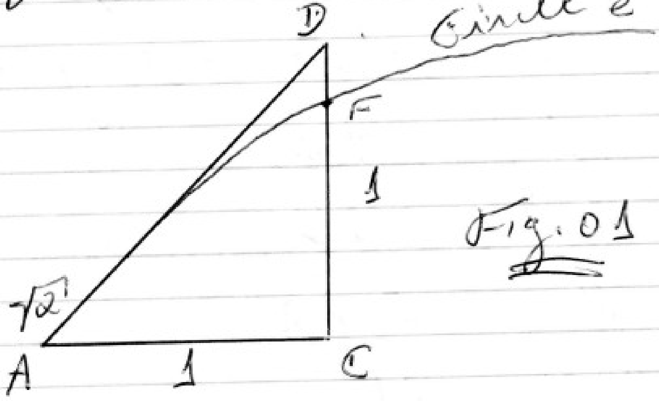

Fig. 01

Obs.:

$F \cong (1, 0.888194417315591)$

Comando:

F = Intersect(e, g, 1)

Important:

A diferença entre o ponto D e F

data 19.08.2019

decorre do problema da quadratura 2D e do valor infinitesimal do π (Pi).
Valor este irracional e transcendental
(o pi não é uma solução de uma equação polinomial com coeficientes racionais). Portanto, o pi transcende a classe dos números algébricos.

The Transcendence of π
Steve Mayer
November 2006

Source: 20190819_03/Pi-Transcendental/
pitrans.pdf

obs.: Earlier Liouville (1844) had actually constructed transcendentals, for example,

$$\sum_{n=1}^{\infty} 10^{-n!}$$ is transcendental.

Source: http://rutherglen.ics.mq.edu.au/math334s106/ma2334.Dioph.Liouville.pdf (não disponível)

Desigualdade interessante:

e^{pi} ge pi^{e} _

19.08.2019

Source: https://www.ime.unicamp.br/~samuel/irracionais-e-transcendentais-π-e/

Interessante:

Space-time fractional reaction-diffusion equations associated with a generalized Riemann-Liouville fractional derivative
R. K. Saxena

2014-09-05

Source: http://arxiv.org/
http://arxiv.org/abs/1409.1817
archive.org/details/arxiv-1409.1817

Source: 20190819_03/Pi-Transcendental/1409.1817.pdf

On intermediate level sets of two-dimensional discrete Gaussian Free Field

Source: http://arxiv.org/abs/1612.01424

19.08.2019

www.archive.org/details/arxiv-1612.01424

Source: 20190819_03/Pi-Transcendental/
1612.01424.pdf

Return:

Na Fig. 01 observamos a discrepância decorrente do π.

Fazendo uma simulação na qual iremos fazer uma distorção do space-time (2D) criaremos um valor para converter 2D que se enquadre na quadratura do círculo.

Este valor é **MUS**. ($\pi\!\!\!\pi$)

π $\pi\!\!\!\pi$

Fazer cálculo com equações:

Usar o livro:

20190819_03/Complex Analysis and Applications by Hemant Kumar Pathak/9811397333.pdf

Pág. 39(60) →

Fig. 1.17 Straight line joining two points

z_1 and z_2.

data 19.08.2019

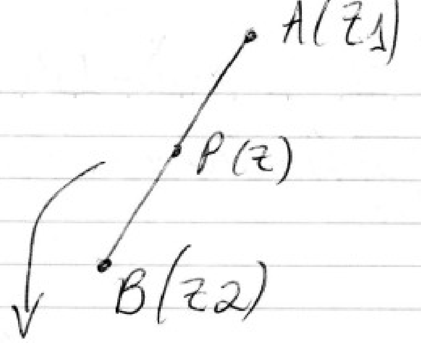

Veja também:
Fig. 1.16

1.8 Equations of Straight Lines

Theorem 1.6

Proof Let z_1, z_2 be two given points A, B on the Argand plane.
Let z be any point P on the line AB as shown in Fig. 1.17.
Then, we have

$$\arg\left(\frac{z - z_1}{z_1 - z_2}\right) = 0 \text{ or } \pi$$

80) cont. (2019 08 20)

Use esta pesquisa na simulação.

a) Hyperbolic lattice appears in a coplanar waveguide array. Belle Dumé Author: Alicia Kollár. 11 Jul 2019. Nature 10.1038/s41586-019-1348-3

data 19.08.2019

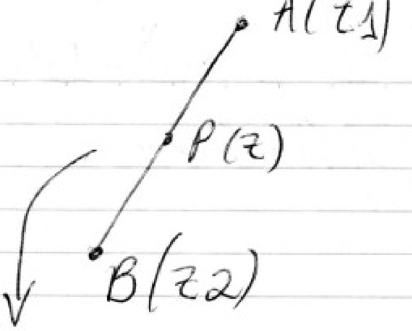

Veja também:

Fig. 1.16

1.8 Equations of Straight Lines

Theorem 1.6

Proof Let z_1, z_2 be two given points A, B on the Argand plane.

Let z be any point P on the line AB as shown in Fig. 1.17.

Then, we have

$$\arg\left(\frac{z - z_1}{z_1 - z_2}\right) = 0 \text{ or } \pi$$

80) cont. (2019 0820)

Use esta pesquisa na simulação.

a) Hyperbolic lattice appears in a coplanar waveguide array. Belle) Durré Author: Alicia Kollár. 11 Jul 2019. Nature 10.1038/s41586-019-1348-3

20 08 2019

Source: https://physicsworld.com/a/
hyperbolic-lattice-appears-in-a-
coplanar-waveguide-array/

Research is detailed Nature 10.1038/
s41586-019-1348-3

Source: 2018 08 20/ Nature_Magazine/
~~s41586-019-1348-3.pdf~~ .png

Obs.: The heptagon-Kagome lattice drawn in a particularly nice projection know as the Poincaré disc-model.

Import.: "Curved-space versions of this lattice occur if we use pentagon and five-pointed stars or heptagons and seven-pointed stars instead of hexagons and six-pointed stars".

Important:

5) Hyperbolic lattices in circuit quantum electrodynamics.
Hollar

Source: 1) nature.com/articles/s41586-019-1348-3/

2) 20190820/Nature_Magazine/
s41586-019-1348-3.pdf

Voltando aos desenhos do item 79, deveremos fazer um aprofundamento da simulação para o referido enquadramento do espaço euclidiano à sua nova realidade.

20.08.2019

Devemos estar atentos a seguinte área:

Fig. I

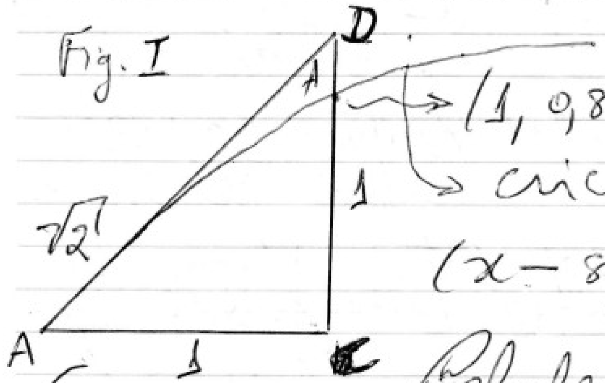

$\rightarrow (1, 0{,}88819441731559 1)$

$\rightarrow$ circle

$$(x-8)^2+(y+8)^2=128$$

Calcular a área A

Compare a discrepância dos cálculos em relação ao cálculo

$$(x+2)^2+(y-2)^2=8$$

Points A e F(−2,2)

Circle K: Circle through A with center F.

Continuando a análise:

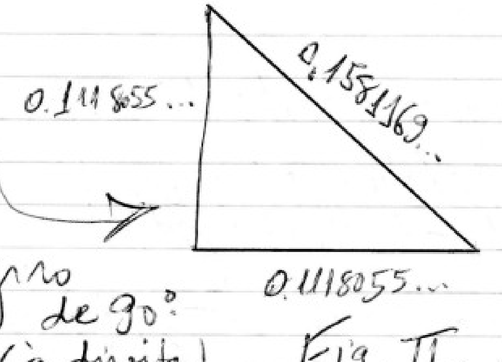
0.1118055...
0.1581169...
0.1118055...

giro de 90° (à direita) Rosário

Fig. II

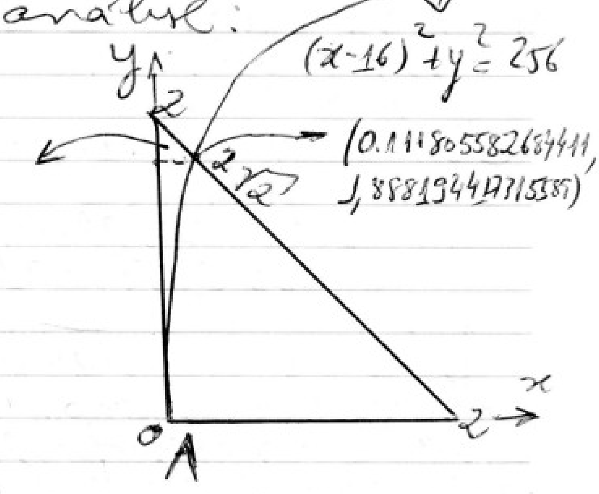

$(x-16)^2+y^2=256$

(0.11180558268441 1, 1,8881944173155 85)

20.08.2019

Posicionando as figuras I e II no triângulo maior:

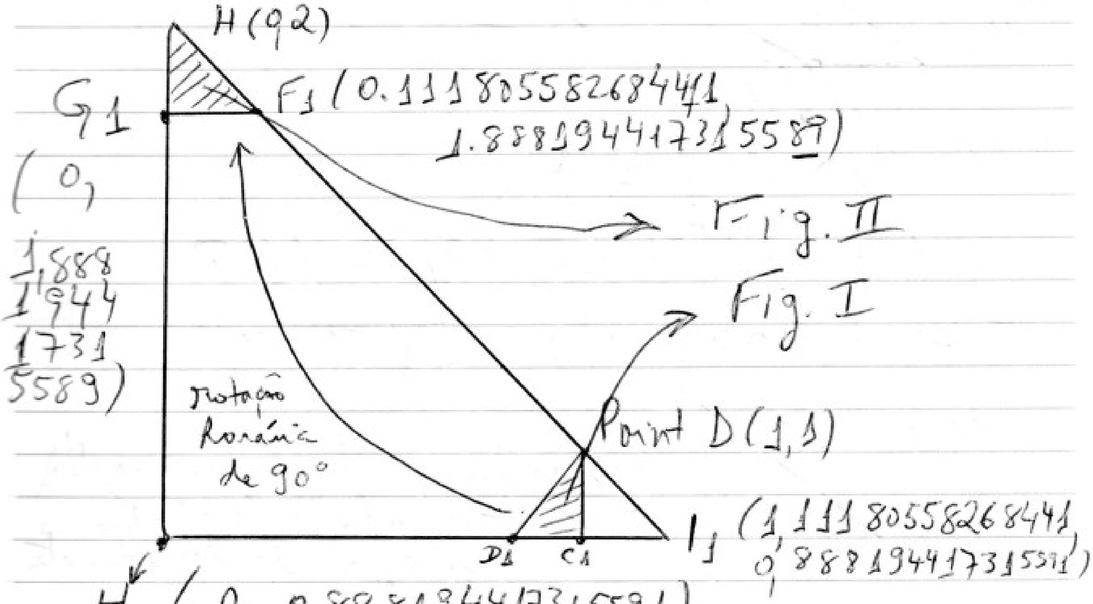
H (0,2)
G1
(0,
1,888
1944
1731
5589)
F1 (0.11180558268441,
1.88819441731558 9)
Fig. II
Fig. I
rotação
horária
de 90°
Point D (1,1)
D1
C1
I1 (1,11180558268441,
0,88819441731559 1)

H_1 (0, 0.888194417315591)

C_1 (1, 0.888194417315591)

D_1 (0.888194417315591,
0.888194417315591)

Distâncias:

$\overline{H_1I_1} = 1.11180558268441$

$\overline{HH_1} = 1.11180558268441$

$\overline{HI_1} = 1.572330533754414$

Important: Os pontos C_1 e F_1 estão intimamente relacionados com os círculos

20.08.2019

que eu vou explicar em seguida:

a)

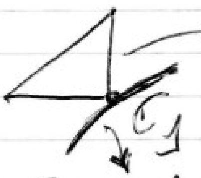

→ Fig. I

Circle (E, A):

$$(x-8)^2+(y+8)^2=128$$

Point (origem): E(8, -8)

Obs.:

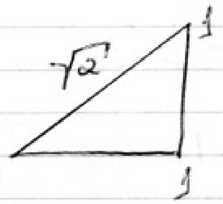

Partindo-se deste triângulo retângulo e fazendo a rotação de quadrante (anti-horário). Q1, Q2, Q3 e Q4, usando perpendiculares com a mesma medidas da hipotenusa por 7 vezes, obteremos este círculo C_1.

$\underbrace{\text{(sete) 7}}_{}$ vezes excluindo o padrão 1 x 1

guarde este valor.

Obs.: O Ponto C_1 é obtido quando o círculo (E, A)

$$(x-8)^2+(y+8)^2=128$$

faz a interseção com a perpendicular que passa centro do círculo que tem como raio 1 unidade (diâmetro 2).

Para a obtenção dos triângulos, nós deveremos fazer o seguinte:

20.08.2019

No Ponto C_1 trace uma perpendicular que passe por ele e que intercepta a hipotenusa ($\sqrt{2}$), cujos catetos são igual a 1. A intersecção na hipotenusa criaremos o ponto D_1. A distância entre $D_1 C_1$ é $\cong$

0,111805582684409...

Pegando-se este valor como base use-o como divisor do Pi (π) teremos:

$$\pi \div 0,11180.... = \underline{28},09871008....$$

Lembra-se do número 7 que foi armazenado anteriormente, faça a divisão usando-o como divisor.

$$28 \div 7 = \underline{\underline{4}}$$

Próximo círculo é o $\underline{8}$ (oitavo).

$$28 \div 8 = 3,5$$

$$(x-16)^2 + y^2 = 256$$

Obs: Este círculo se aproxima do eixo $\underline{y}$.

Point for creation E_1 (16,0) e faça círculo com o point A(0,0).

Important: O círculo 7 intercepta o centro do círculo 8 no ponto (16,0).

20.08.2019

b) Fig. II

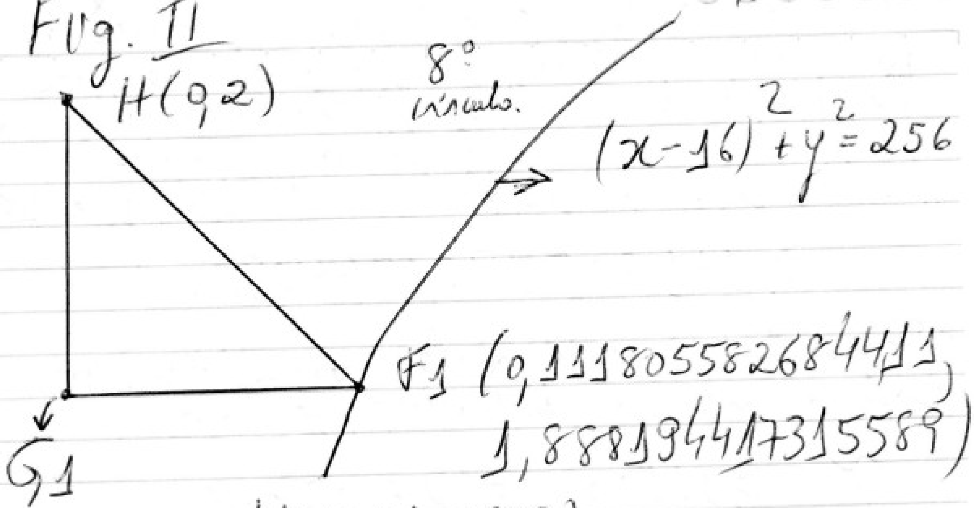
H(0,2)
8° círculo
$(x-16)^2+y^2=256$
F1 (0,111805582684411,
1,888194417315589)
G1

(0, 1.888194417315589).

Complementando-se a análise:

Contando-se 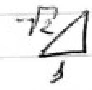

são 9 círculos.

$28 \div 9 = 3,\overline{1}$.

(muito próximo do π).

Obs.: Poderemos usar este valor quando da simulação. (distorção space-time)

Utilizar o livro:

The Nonlinear Schrödinger Equation Singular Solutions and Optical Collapse

Source: 20190807_02/The Nonlinear Schrödinger Equation Singular.../The Nonlinear... pdf

20.08.2019

Source: arquivo.ggb

ARQ/Geogebra_org_2017_versão_6_0_
4112_02/
20190820_Padrão_007_003.ggb

Obs: Contém Fig. I e II.

Original file com todos os passos:

20190818_011_V003_004_padrão_002.ggb.

File: 20190820_Padrão_006_002.ggb

Outro file:

20190818_011_V003_004_padrão_002.ggb

Distância $\overline{VU}$:

$V = (0.732050807568878,$
$1.267949192431122)$

$U = (1.062740668062227,$
$0.937253931193773)$

Distância: 0.467673720625947

Obs: Criei outro arquivo:
20190818-001_V003_padrão_002_
distancia_VU_001.ggb.

20.08.2019

$\sqrt{U} \times 3 = 1.40301961877841^{x}$

$< \sqrt{2}$

81) cont. (20190821):

File: 20190818_009.ggb
(compare neste)
com 20190818_009_padrão.ggb

1) Criar um círculo de raio 4 unidades (diâmetro 8 unidades) cujo centro é A, faça outro círculo cujo centro é a metade do raio do primeiro círculo.

Todos os círculos estarão um dentro do outro e todos são tangentes ao mesmo ponto^{D} que está fixado do outro lado da circunferência alinhado com o centro da primeira circunferência.

2) Ao todo utilizaremos 5 círculos

Descrição:

a) Point A(0,0)
b) Point B(4,0)
c) Point C(-2,0)
d) Point E(-3,0)
e) Point F(-3,5, 0)

Point G(-3.75, 0)

obs.: O círculo com centro no ponto E(-3,0) e raio de 1 unidade no ponto C(-2,0) tem uma área = Pi (π).

Fazendo um estudo sobre perpendiculares que

! "# # $ " %

& $ ' $! ! !

( ) * ! + , $ $- , .

& $ / 0/

( ) * ! + , $ $- - - .

& $"

( ) * ! + , $ - ! 1 - .

2 , $ $- - "- /-3 ! 1 .

& $/

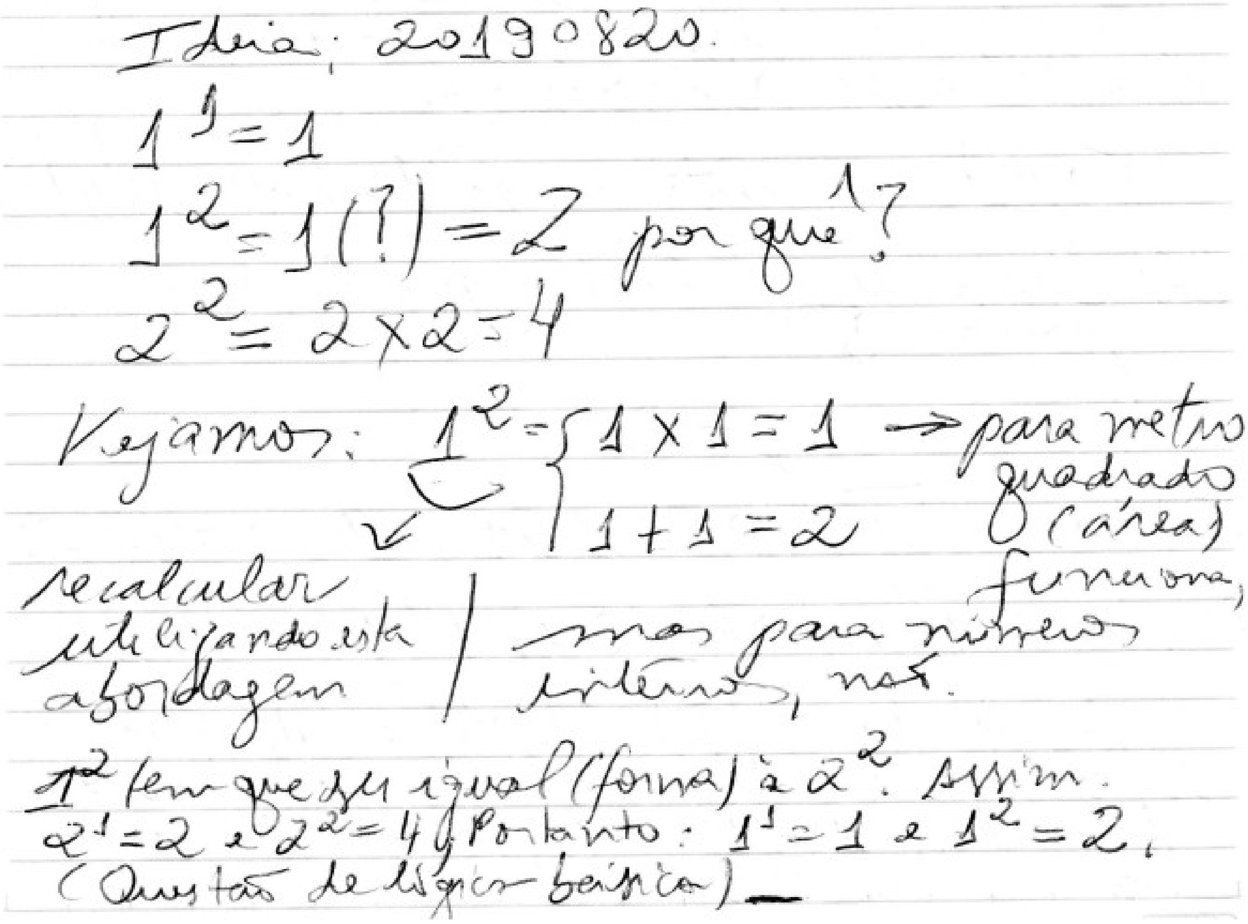

Ideia: 20190820.

$1^1 = 1$

$1^2 = 1(?) = 2$ por que?

$2^2 = 2 \times 2 = 4$

Vejamos: $1^2 = \begin{cases} 1 \times 1 = 1 \\ 1 + 1 = 2 \end{cases}$ → para metro quadrado (área) funciona, mas para números inteiros, não.

recalcular utilizando esta abordagem

1^2 tem que ser igual (forma) à 2^2. Assim: $2^1 = 2$ e $2^2 = 4$ (Portanto: $1^1 = 1$ e $1^2 = 2$. (Questão de lógica básica) —

& $

23.08.2019

85) 20190823

Solução para a quadratura do círculo.

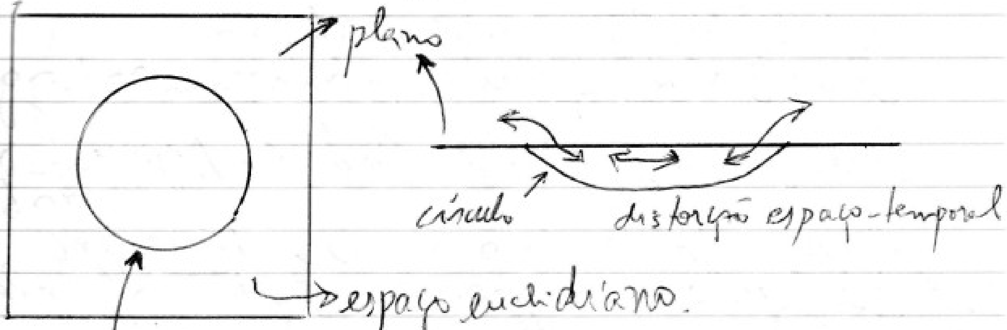
plano
círculo
distorção espaço-temporal
espaço euclidiano.

valor $\frac{1}{\pi}$ (distorcido) < π e igual quadratura.
(ajustamento)

(fator conversão de universos diferentes mas coexistentes.

Obs.: Como Teoria Clássica e a Teoria Quântica.

π é o fator de convergência entre este dois universos com as suas complexidades que são inerentes à cada sistema e agora serão intercambiáveis.

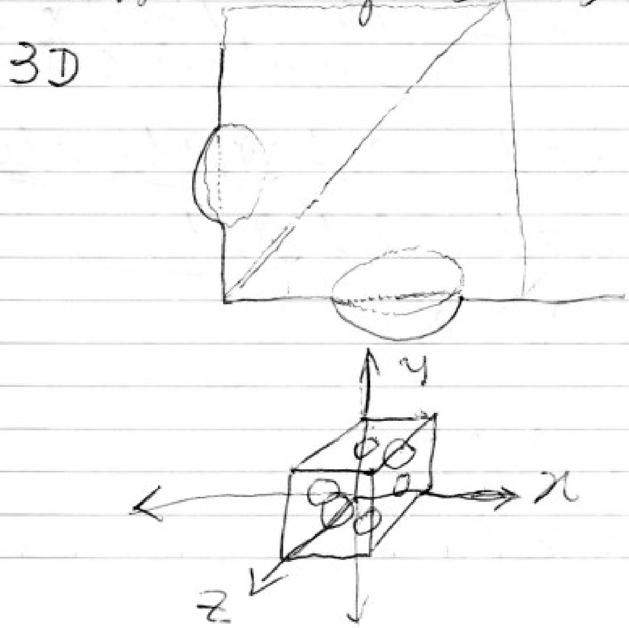

data 24.08.2019

85) Cont. 84. (20190824)

Para simular este cubo sem a distorção criei os arquivos:

20190823_Cube_001.ggb
20190823_Cube_montagem_001.ggb
20190823_Cube_Padrão_Ajustado_001.ggb

Agora falta colocar a distorção e calcular os pontos entre os planos em 3D. (nova trigonometria !!!)

Vejamos:

(base)
20190823_Cube_Padrão_Ajustado_001.ggb
versão salva: 002.ggb

Passos:

1) Slider
a = 1 (Min: 0, Max: 10 → é melhor simulação
Increment: 1)
b = 1 (0, 10, +1) → 10
c = 1 (0, 10, 1)

2) ~~Point~~ A = (0, 0, 0)
coloque direto ↑

B = (a, 0, 0)

C = (b, c, 0)

24.08.2019

D = (0, b, 0)

3) Polygon [A, B, C, D]
↳ parênteses

4) Entre nas opções do programa: ≡
(lado direito da tela)

a) Perspectives

b) 3D Graphics

c) Vá na opção 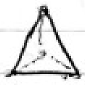
↳ Extrude to ~~Pyramid or Cone~~ Prism or Cylinder
(4.ª opção)
Após, clique no Polygon (A, B, C, D)
e dê altitude $\underline{\underline{1}}$ → OK
Obs: Criou-se um Cubo.

d) Opção do Programa: ≡
Volte para a visualização
Geometry

5) Crie novo controle deslizante:
Slide e = 1 (0, 1, 0.01)
increment

data 24.08.2019

D S T Q Q [✓ S] S

6) Commando do programa: ≡

↳ clique: △

↳ escolha a 9ª opção: NET.

↳ clique no Cubo.

obs: O controle deslizante cria-se automaticamente.

A velocidade de execução da demonstração está no canto superior direito do controle deslizante (~~Graphing~~)

Algebra → local onde os comandos são inseridos

→ Possibilita-se na opção VIEW → marque ☑ Algebra.

—

obs:

Velocidade aumenta-se 0.5 em 0.5

<< 2x >> fica em cima do comando play.

▶

(Para facilitar: esta nova versão do programa, permite que você use a extensão e em seguida você executar a NET (simulação)

→ não funcionou.

—

& $%

( ) * , $ "- . - ! -(4 ! - " .

& $0

( ) * , $ /- 51 -6 ! - .

& $$ 777

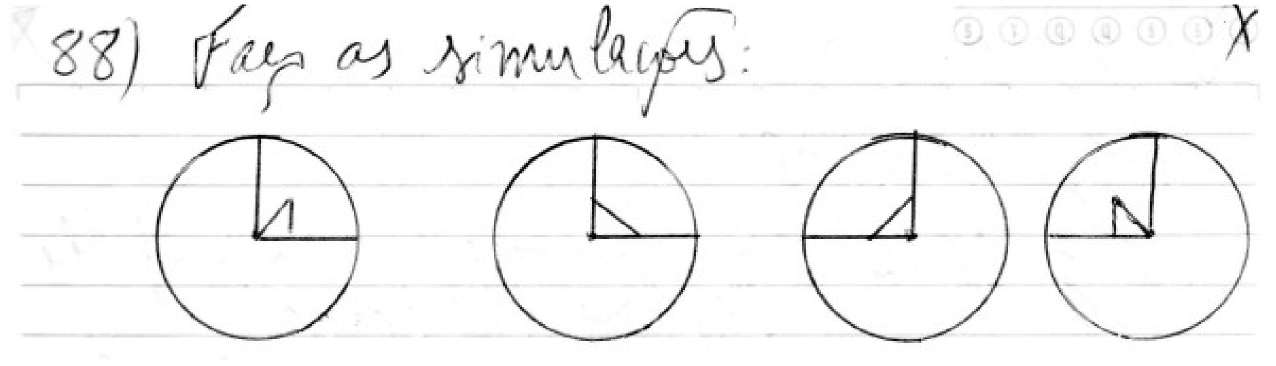

& $,

8 .

9(:69 9 92 9 $, . 3 *
; 3 0 3!

2. < ! * ===

' / > ? ;

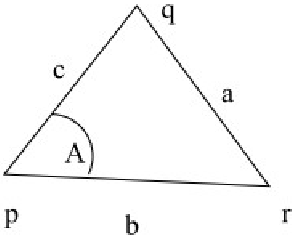

Fig. 1.1. A triangle in $\mathbb{E}^2$

In the Euclidean plane $\mathbb{E}^2$ we have distances between points and angles between smooth curves wherever the curves meet. The equation

$$a^2 = b^2 + c^2 - 2bc\cos A$$

should be kept in mind, where a, b, c are the pairwise distances between three points, as in figure 1.1 and A is the angle opposite the side of length a. The triangle inequality in $\mathbb{E}^2$:

$$a \leq b + c$$

@) ; 3 . @ 3 .
) ! A) . ;@ ! ! ;
! ; < B ! . ;
3 @ ; ; ! ! ! @

C 3 B < B
; ; ! @ D C)
* ;! ; @ 3 * ! ;@ ;@ @

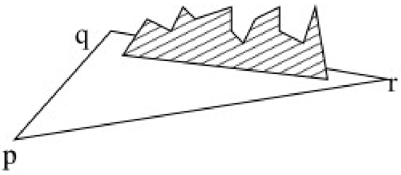

E 2 ! !

9(:69 9 92 9 $, . 3 *
; 3 0 3!
F *

' $ % /

E

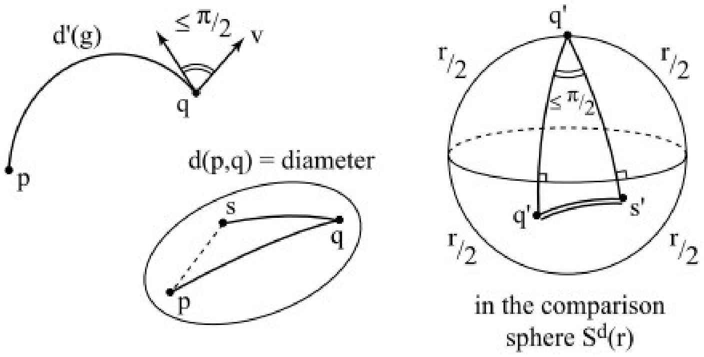
d'(g)
≤ π/2
v
q
p
d(p,q) = diameter
s
q
p
q'
r/2
r/2
≤ π/2
q'
s'
r/2
r/2
in the comparison
sphere S^d(r)

E * 3

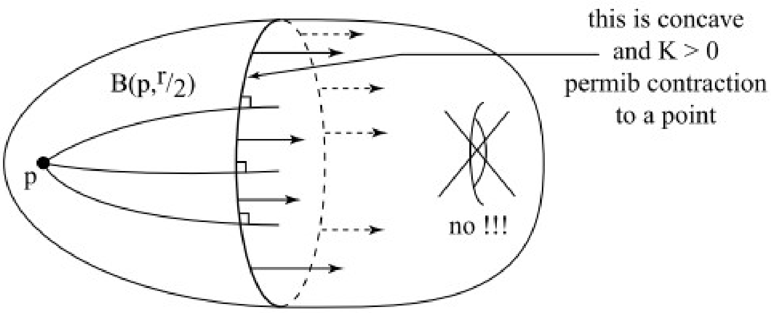
this is concave
and K > 0
permib contraction
to a point
B(p,r/2)
p
no !!!

E " ? *G 3 3

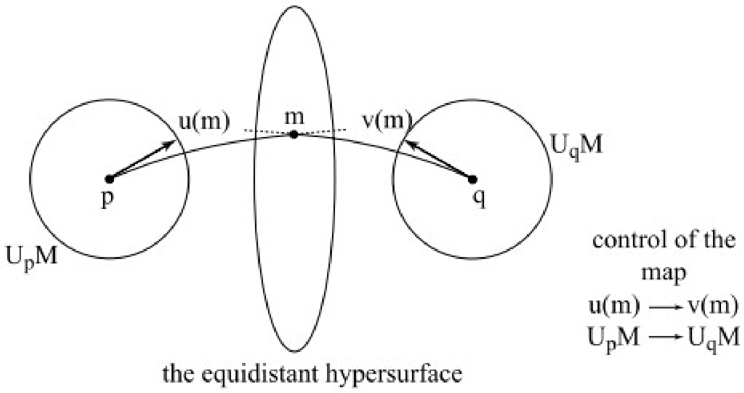
u(m)
m
v(m)
U_qM
p
q
U_pM
control of the map
$u(m) \longrightarrow v(m)$
$U_pM \longrightarrow U_qM$
the equidistant hypersurface

E ? G 3 3

' ," %

< " 0

Theorem 307 (Durumeric 1987 [466]) *For a suitable ε, any compact Riemannian manifold of diameter larger than $\pi\sqrt{\delta}/2-\varepsilon$ with sectional curvature $K \geq 1$ is homeomorphic to a spherical space form manifold or a simply connected manifold whose cohomology ring is generated by a single element (which is the case for the $\mathbb{KP}^n$ but the converse at the topology level is not true).*

> D " $ < ! 3 !
. ! 3
< ! D * ! E H 3 @ B * ;
3 ! & ! ! @ D /
3

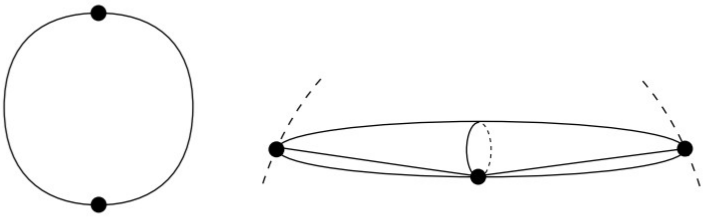

radius = diameter

radius close to half the diameter

E / : ! I ! . : ! !

< * ! ; ; ! (!
* * J4 " 3 % " !

. 9(:69 9 92 9 $, K ? > ? ;
; ! * E ! > E - E A !
: 3!

' $" < L ?

' , / * ! .

$$r = \sqrt{x^2 + y^2 + z^2}, \quad \phi = \arctan\frac{y}{x}, \quad \theta = \arctan\frac{\sqrt{x^2 + y^2}}{z}.$$

& * * . .; * H
* * ! . !
< @ @ ! 3 ! ; .

$$x = r\sin\theta\cos\phi, \quad y = r\sin\theta\sin\phi, \quad z = r\cos d\theta.$$

This is typical when one knows the R^i_{kl}'s and R^i_{jlm}'s but not the ω^k and ω^r_s. Sometimes, one refers to the $d\omega^i - \omega^j \wedge \omega^i_j$ and the $d\omega^j_i - \omega^k_i \wedge \omega^j_k$ themselves as the torsion and the curvature. Since computations with the explicit form that ω^j and ω^i_j take in the bundle would be laborious in general, one actually computes in sections of the bundle, a concept which we are about to introduce. The Ω^i and Ω^i_j may be viewed as differential form components of a $(1,0)-$tensor and $(1,1)-$tensor respectively. But this is only just one way of looking at curvature and torsion. Impatient readers can refer to sections 5.9 and 8.13 for their Lie algebra valuedness.

Let us return to affine space. The equations of structure of affine space are then written as

$$\Omega^i = 0, \qquad \Omega^i_j = 0. \tag{3.8}$$

One then says that affine spaces are particular cases of differentiable manifolds endowed with affine connections whose torsion and curvature are zero.

An interesting exercise at this point with the equations of affine space is the following. Substitute Eqs. (2.7) in (3.5). Use $A^j_i B^k_j = \delta^k_i$ and reach two trivially looking identities, $0 = 0$. In generalized spaces, they will take interesting forms. We shall then use this trivial identity to simplify calculations.

5.4 Curvilinear coordinates: holonomic bases

A section of the affine frame bundle is a continuous map $S : Af^n \to F^n$ such that $\mathbf{P} \circ S$ is the identity. In simpler terms, it is a map that picks one and only one point from each fiber. Since those points are frames, the section is a frame field, which we shall assume to be not only continuous but also differentiable to whichever order may be needed. The concept is equally applicable to the bundles of generalizations of affine and Euclidean spaces.

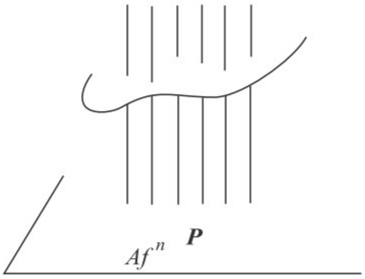

Figure 4: Section of F^n

As we shall see in the next paragraph, sections emerge naturally when one introduces in the base manifold (i.e. affine space in this chapter) curvilinear coordinates through an invertible transformation: $x^i = x^i(A^1 \ldots A^n)$, $i = 1 \ldots n$. Needless to say that we may define new curvilinear coordinates in terms of other curvilinear coordinates rather than directly from the A^i. The x^i are no longer rectilinear if the equations of the transformation are not linear. Inexperienced readers should be aware of the inconsistency in terminology from one area of mathematics to another. One should use the term affine equations to refer to what are usually called linear equations. Linear transformation yields the special linear equations whose independent terms are all zero.

Curvilinear coordinate systems define point-dependent bases of vectors $\mathbf{e}_i$ by means of

$$\mathbf{e}_i \equiv \frac{\partial \mathbf{P}}{\partial x^i}. \tag{4.1}$$

We use the same term $\mathbf{e}_i$ as before since these frame fields are sections of the bundle of frames. One says that the $\mathbf{e}_i$ of (4.1) are pull-backs (to sections) of the $\mathbf{e}_i$ of (2.3). Curvilinear coordinates are not components of $\mathbf{P}$ or some other vector upon some basis $\{\mathbf{b}_i\}$ and, in fact, one first defines those coordinates and then a basis field associated with them.

The basis vectors $\mathbf{e}_i$ and $\mathbf{a}_j$ are related by:

$$\mathbf{e}_i = \frac{\partial \mathbf{P}}{\partial A^j}\frac{\partial A^j}{\partial x^i} = \frac{\partial A^j}{\partial x^i}\mathbf{a}_j. \tag{4.2}$$

The partial derivatives are not constants but functions of the coordinates, unless we were dealing with a transformation between two rectilinear coordinate systems. To be precise, we are dealing here with composite functions, since the A^i are themselves functions that assign numbers to the points of the manifold (Refer to point (c) of section 6 of chapter 1). Because of its importance, we repeat that, when dealing with coordinates, we use the same symbol for coordinate functions and for the value that those functions take. Whether we are considering the ones or the others will be obvious from the context. The nature of the coordinates as functions on the manifold started with the definition of $\mathbf{P}$ as a projection map on Af^n. It follows that rectilinear coordinates are also functions on Af^n. Curvilinear coordinates being functions of the rectilinear coordinates, they also are (composite) functions on Af^n.

The $(\mathbf{e}_i)$ vary from one point to another. Each coordinate system determines a vector basis at each point of the region of affine space that it covers (No basis is determined at the origin of the system of polar coordinates, since this system does not assign a unique pair of numbers to it; the definition (4.2) fails in any case). We are thus dealing here with fields of bases, not just one basis as was the case with $(\mathbf{a}_i)$. But $(\mathbf{a}_i)$ can now be viewed as a constant basis field. This becomes obvious when we obtain a system of rectilinear coordinates from another one in the same way as for curvilinear coordinates.

The vector basis fields generated in the preceding way by the system of curvilinear coordinates is referred to as holonomic or coordinate basis field. In that change from rectilinear to rectilinear, the new basis field $(\mathbf{b}_i)$ appears as a particular case of curvilinear basis field that happens to be the same everywhere in terms of the $(\mathbf{a}_i)$, and thus equal to itself from one point to another. The argument can be reversed, meaning that we could have started the whole argument with the basis ($\mathbf{b}_i$) and have obtained the basis ($\mathbf{a}_i$) as a constant frame field also denoted ($\mathbf{a}_i$).

Non-holonomic basis fields will be introduced in the next section. We leave for then the push-forward (a concept that amounts here to a recovery) of the invariant forms in the bundle from the dx^k and ω_i^j just considered.

The use of a particular coordinate system does not force us to use corresponding holonomic basis fields (unlike in the most common version of the tensor calculus, version where bases are not explicitly used).

From the definition (4.1) of $\mathbf{e}_i$, we have:

$$d\mathbf{P} = dx^i \mathbf{e}_i. \tag{4.3}$$

Along each coordinate line, all the dx^i's but one are zero. Equation (4.3) then shows that the vector is tangent to the coordinate line x^i. In a similar way to the obtaining of Eq. (4.2), we would get

$$\mathbf{a}_i = \frac{\partial \mathbf{P}}{\partial x^j}\frac{\partial x^j}{\partial A^i} = \frac{\partial x^j}{\partial A^i}\mathbf{e}_j. \tag{4.4}$$

This shows that the vectors $(\mathbf{e}_j)_P$ constitute a basis since any vector can be expressed as a linear combination of the $\mathbf{e}_i$'s through the $\mathbf{a}_i$'s (and the number of vectors $\mathbf{e}_j$ is the same as the dimension of the vector space).

Let us practice some of these ideas using polar coordinates in the plane

$$x = \rho\ \cos\phi, \qquad y = \rho\ \sin\phi. \tag{4.5}$$

We use (4.2) with (x, y) and (ρ, ϕ) replacing the A^i's and x^i's respectively. We thus have

$$\mathbf{e}_\rho = \cos\phi \mathbf{a}_1 + \sin\phi \mathbf{a}_2 \qquad \mathbf{e}_\phi = -\rho\sin\phi \mathbf{a}_1 + \rho\cos\phi \mathbf{a}_2 \tag{4.6}$$

where we have used $\mathbf{a}_1$ and $\mathbf{a}_2$ for $\mathbf{i}$ and $\mathbf{j}$. Notice that, with this definition of $\mathbf{e}_j$, $d\mathbf{P}$ is not given by the familiar $d\mathbf{P} = d\rho\mathbf{e}_\rho + \rho d\phi\mathbf{e}_\phi$, which corresponds to orthonormal bases, but rather by

$$d\mathbf{P} = d\rho\mathbf{e}_\rho + d\phi\mathbf{e}_\phi, \tag{4.7}$$

in accordance with Eq. (4.3). The concept of orthonormal bases does not even enter here since a dot product or a metric has not been defined. Thus, although ρ and ϕ have metric connotation, we ignore it here and consider those functions of (x, y) as any other functions without such connotation.

From (4.6), we read $A_1^1 = \cos\ \phi$, $A_1^2 = \sin\phi$, $A_2^1 = -\rho\sin\ \phi$, $A_2^2 = \rho\cos\phi$ for the vector basis (on the section) that is above the point (ρ,ϕ) of Af^2.

Exercise. Obtain $\mathbf{e}_i$ in terms of $\boldsymbol{a}_j$, and vice versa, for the spherical coordinates:

$$r = \sqrt{x^2 + y^2 + z^2}, \quad \phi = \arctan\frac{y}{x}, \quad \theta = \arctan\frac{\sqrt{x^2 + y^2}}{z}. \tag{4.8}$$

It is not convenient to invert the change of basis by inverting the matrix that gives this change if the inverse coordinate transformation is remembered. Thus, in this case, we would proceed in the same way but starting from the transformations

$$x = r\sin\theta\cos\phi, \quad y = r\sin\theta\sin\phi, \quad z = r\cos d\theta. \tag{4.9}$$

Notice that we do not need to draw figures to relate different vector bases, which may not be easy to do for unusual coordinate systems.

We next summarize much of what has been done in this section. We arbitrarily chose a point of affine space and introduced an arbitrary basis $(\mathbf{a}_i)$. Rectilinear coordinates (A^i) were obtained as components of $\mathbf{P}$ relative to that basis. This was followed by the introduction of curvilinear coordinates as functions of the rectilinear ones. We then generated basis fields associated with the coordinate system. We completed the process by finally viewing the rectilinear coordinates as the particular cases of curvilinear coordinates that generate constant vector basis fields $\{\mathbf{a}_i\}$.

The use of sections —thus of a specific vector basis at each point of affine space— has the implication that the ω_i^j's are now linear combinations

$$\omega_i^j = \Gamma_{ik}^j dx^k \tag{4.10}$$

of the differentials of the coordinates in affine space. We thus have

$$d\mathbf{P} = dx^i \mathbf{e}_i \tag{4.11a}$$

$$d\mathbf{e}_i = \omega_i^j \mathbf{e}_j = \Gamma_{ik}^j \mathbf{e}_j dx^k. \tag{4.11b}$$

Consistently with concepts already advanced, these equations pertain to quantities (on the left) that are said to be pull-backs to sections of quantities in the bundle.

One can show that the pull-backs of exterior derivatives and exterior products are the exterior derivatives and exterior products of the pull-backs of the quantities being exterior differentiated and exterior multiplied. We leave for section 6 the pull-back of the equations of structure to sections of the bundle, whether holonomic sections like those we have just seen, or anholonomic ones, about to be considered.

Finally, the basis fields of two different coordinate patches can be related in a way similar to (4.4). Let (x^i) and (x'^i) be two sets of coordinates. We shall denote by $(\mathbf{e}'^i)$ the basis field for the coordinate patch (x'^i). A change of coordinate vector basis field is then given by

$$\mathbf{e}_i = \frac{\partial \mathbf{P}}{\partial x'^j}\frac{\partial x'^j}{\partial x^i} = \frac{\partial x'^j}{\partial x^i}\mathbf{e}'_j \tag{4.12a}$$

$$\mathbf{e}'_i = \frac{\partial \mathbf{P}}{\partial x^j}\frac{\partial x^j}{\partial x'^i} = \frac{\partial x^j}{\partial x'^i}\mathbf{e}_j. \tag{4.12b}$$

' , $ A ! L ? ;

' , : ! 3 M : 3

9(:69 9 92 9 $, ? ; :
@ . ! N * ; 3!

9(:69 9 92 9 $, ? ; : !
< ! 3 ,@ & 8 3!

! 9(:69 9 92 9 $,
OC -P ; @-K -8 BQ-A -A 3!

' /

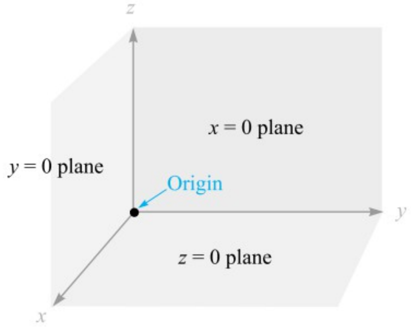
z
x = 0 plane
y = 0 plane
Origin
y
z = 0 plane
x

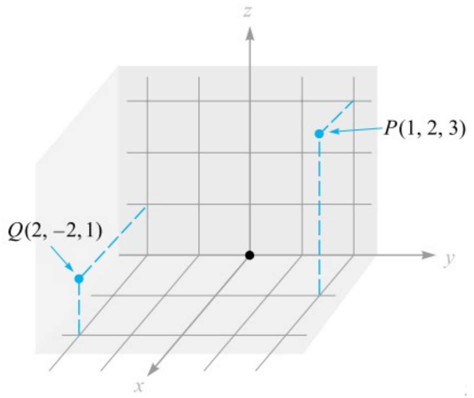
z
P(1, 2, 3)
Q(2, −2, 1)
y
x

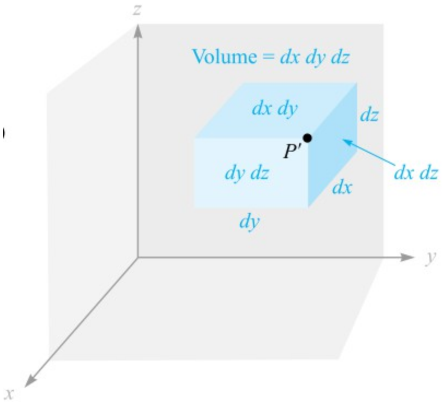
z
Volume = dx dy dz
dx dy
dz
P'
dy dz
dx dz
dx
dy
y
x

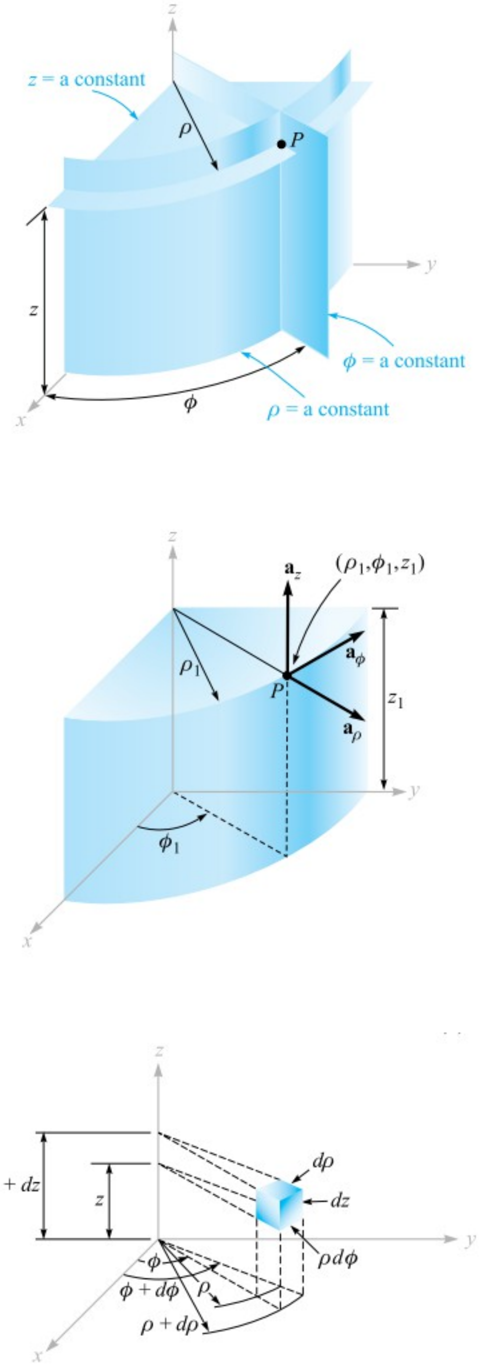
z
z = a constant
ρ
P
y
z
ϕ = a constant
x
ϕ
ρ = a constant
z
a_z
(ρ₁,ϕ₁,z₁)
a_ϕ
ρ₁
P
z₁
a_ρ
y
ϕ₁
x
z
dρ
z + dz
z
dz
y
ρdϕ
ϕ
ϕ + dϕ
ρ
ρ + dρ
x

Figure 1.6 (*a*) The three mutually perpendicular surfaces of the circular cylindrical coordinate system. (*b*) The three unit vectors of the circular cylindrical coordinate system. (*c*) The differential volume unit in the circular cylindrical coordinate system; $d\rho$, $\rho d\phi$, and dz are all elements of length.

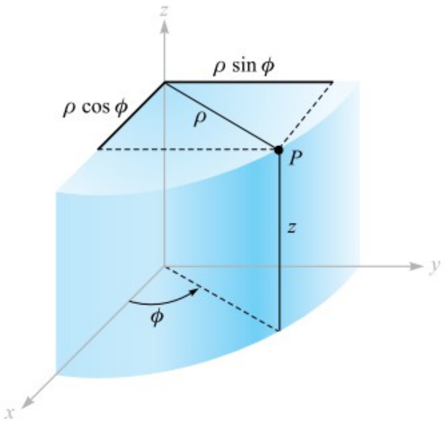

Figure 1.7 The relationship between the rectangular variables x, y, z and the cylindrical coordinate variables ρ, ϕ, z. There is no change in the variable z between the two systems.

The variables of the rectangular and cylindrical coordinate systems are easily related to each other. Referring to Figure 1.7, we see that

$$\begin{aligned} x &= \rho \cos \phi \\ y &= \rho \sin \phi \\ z &= z \end{aligned} \tag{10}$$

From the other viewpoint, we may express the cylindrical variables in terms of x, y, and z:

$$\begin{aligned} \rho &= \sqrt{x^2 + y^2} \quad (\rho \geq 0) \\ \phi &= \tan^{-1} \frac{y}{x} \\ z &= z \end{aligned} \tag{11}$$

2. F * ! ===

9(:69 9 92 9 $, ; ;- : 3 < ;
! & (33 - & P R : C /@ 8 B S 8 3!
777777777

9(:69 9 92 9 $, ?
TUVWXYYUXYZWXYU[\]^_Z`ZXaa]UbXcU\Y_Zc\ZaXdcUX]Z^U[[VdVYcUX]ZVefXcU\Y_ghUdijkf_Vd

' %0 0"

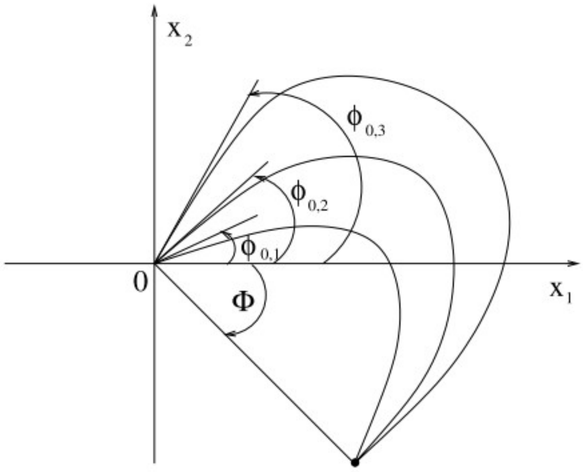

E , < 3 4 ! H 3 !

A r3 ' ! ! 3 , /

< A F ;

$$
\begin{aligned}
\theta &= \frac{\partial L}{\partial \dot{t}} = \dot{t} + 2r^2\dot{\phi},\\
\eta &= \frac{\partial L}{\partial \dot{\phi}} = r^2\dot{\phi} + 2\dot{t}r^2 + 4r^4\dot{\phi},\\
\rho &= \frac{\partial L}{\partial \dot{r}} = \dot{r}.
\end{aligned}
$$

9.4.1.1 The Euler–Lagrange system

The momenta are

$$\begin{aligned}\theta &= \frac{\partial L}{\partial \dot t} = \dot t + 2r^2\dot\phi,\\ \eta &= \frac{\partial L}{\partial \dot\phi} = r^2\dot\phi + 2\dot t r^2 + 4r^4\dot\phi,\\ \rho &= \frac{\partial L}{\partial \dot r} = \dot r.\end{aligned}$$

As the Lagrangian L does not depend on t and ϕ, the Euler–Lagrange equations yield

$$\theta = \text{constant}, \qquad \eta = \text{constant}.$$

The momentum η can be written in terms of θ as

$$\eta = r^2(\dot\phi + 2\dot t + 4r^2\dot\phi)s = r^2(\dot\phi + 2\theta).$$

The Euler–Lagrange equation $\dot\rho = \dfrac{\partial L}{\partial r}$ becomes

$$\begin{aligned}\ddot r &= r\dot\phi^2 + 4\dot t r\dot\phi + 8r^3\dot\phi^2\\ &= r\dot\phi^2 + 4r\dot\phi(\dot t + 2r^2\dot\phi)\\ &= r\dot\phi^2 + 4r\dot\phi\theta\\ &= r\dot\phi(\dot\phi + 4\theta).\end{aligned}$$

Hence the Euler–Lagrange system is

$$\begin{cases}\ddot r & = r\dot\phi(\dot\phi + 4\theta),\\ r^2(\dot\phi + 2\theta) & = \eta,\\ \dot t + 2r^2\dot\phi & = \theta\\ \theta & = \text{constant},\\ \eta & = \text{constant}.\end{cases} \tag{9.4.17}$$

It suffices to study only the geodesics from the origin, because of the Heisenberg translation. In this case $r(0) = 0$ and hence $\eta = 0$. It follows that $\dot{\phi} = -2\theta$ and the system (9.4.17) becomes

$$\begin{cases} \ddot{r} &= -4\theta^2 r, \\ \dot{\phi} &= -2\theta, \\ \dot{t} &= \theta - 2r^2\dot{\phi} = \theta(1+4r^2), \\ \theta &= \text{constant} \end{cases} \tag{9.4.18}$$

with the boundary conditions

$$r(0) = 0, \quad \phi(0) = \phi_0, \qquad t(0) = t_0 = 0, \tag{9.4.19}$$

$$r(\tau) = \mathbf{r}, \quad \phi(\tau) = \Phi, \qquad t(\tau) = \mathbf{t}. \tag{9.4.20}$$

We shall show in the following that the system (9.4.18) has solutions if and only if some compatibility of the above boundary conditions holds. The solutions are

$$r(s) = \frac{\sin(2\theta s)}{\sin(2\theta\tau)}\mathbf{r}, \tag{9.4.21}$$

$$\phi(s) = -2\theta s + \phi_0. \tag{9.4.22}$$

The boundary condition $\phi(\tau) = \Phi$ yields

$$\theta = \frac{1}{2\tau}(\phi_0 - \Phi). \tag{9.4.23}$$

Integrating in (9.4.21) yields

$$\begin{aligned} t(s) &= \theta \int_0^s (1 + 4r^2(u))\,du = \theta\Big(s + 4\int_0^s r^2\Big) \\ &= \theta\Big(s + \frac{4\mathbf{r}^2}{\sin^2(2\theta\tau)}\int_0^s \sin^2(2\theta u)\,du\Big) \\ &= \theta\Big[s + \frac{4\mathbf{r}^2}{2\theta\sin^2(2\theta\tau)}\Big(\frac{1}{2}(2\theta s) - \frac{1}{4}\sin(4\theta s)\Big)\Big] \\ &= \theta s + \frac{\mathbf{r}^2}{\sin^2(2\theta\tau)}\Big(2\theta s - \frac{1}{2}\sin(4\theta s)\Big). \end{aligned} \tag{9.4.24}$$

The boundary condition $t(\tau) = \mathbf{t}$ yields

$$\begin{aligned} \mathbf{t} &= \theta\tau + \frac{\mathbf{r}^2}{\sin^2(2\theta\tau)}\Big(2\theta\tau - \sin(2\theta\tau)\cos(2\theta\tau)\Big) \\ &= \theta\tau + \mathbf{r}^2\Big(\frac{2\theta\tau}{\sin^2(2\theta\tau)} - \cot(2\theta\tau)\Big). \end{aligned} \tag{9.4.25}$$

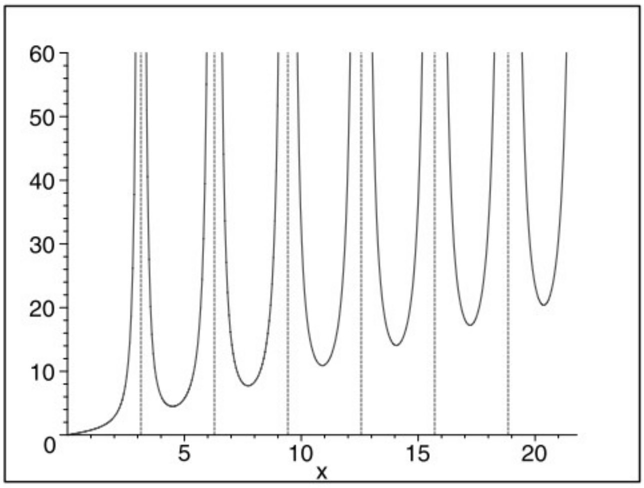

Figure 9.3: The graph of $\mu(x)$.

Let

$$\mu(x) = \frac{x}{\sin^2 x} - \cot x. \tag{9.4.26}$$

The graph of μ for $x > 0$ is sketched in Figure 9.3. It suffices to study only the case $\theta > 0$. The case $\theta < 0$ can be obtained from the previous one changing $t \to -t$ and $\phi \to -\phi$. This follows from the relation $\theta = \dot{t} + 2r^2\dot{\phi}$. Then (9.4.25) becomes

$$\mathbf{t} = \theta\tau + \mathbf{r}^2\mu(2\theta\tau). \tag{9.4.27}$$

In order to understand the exact number of geodesics, which join the origin with any given point, we need the following lemma, see Beals, Gaveau and Greiner [37].

Lemma 9.20 *μ is a monotone increasing diffeomorphism of the interval $(-\pi, \pi)$ onto $\mathbb{R}$. On each interval $(m\pi, (m+1)\pi)$, $m = 1, 2, \ldots$, μ has a unique critical point x_m. On this interval μ decreases strictly from $+\infty$ to $\mu(x_m)$ and then increases strictly from $\mu(x_m)$ to $+\infty$. Moreover*

$$\mu(x_m) + \pi < \mu(x_{m+1}), \quad m = 1, 2, \ldots, \tag{9.4.28}$$

$$0 < \Big(m + \frac{1}{2}\Big)\pi - x_m < \frac{1}{m\pi}. \tag{9.4.29}$$

Proof. As μ is an odd function, it suffices to show that it is a monotone increasing diffeomorphism of the interval $(0, \pi)$ onto $(0, +\infty)$. We note that $\sin x - x\cos x$ vanishes at $x = 0$ and it is increasing in $(0, \pi)$. Then

$$\frac{1}{2}\mu'(x) = \frac{\sin x - x\cos x}{\sin^3 x} = \begin{cases} = 1/3, & x = 0, \\ > 1/3, & x \in (0, \pi). \end{cases}$$

The first identity holds as an application of the l'Hospital rule:

$$\lim_{x\to 0}\frac{\sin x - x\cos x}{\sin^3 x} = \lim_{x\to 0}\frac{\cos x - \cos x + x\sin x}{3\sin^2 x}$$
$$= \frac{1}{3}\lim_{x\to 0}\frac{x}{\sin x} = \frac{1}{3}.$$

The second inequality holds because

$$\frac{1}{2}\mu''(x) = \frac{x + 2x\cos^2 x - 3\cos x\sin x}{\sin^4 x} > 0.$$

The numerator vanishes at $x = 0$, and its derivative is

$$4\sin x(\sin x - x\cos x) > 0, \quad x \in (0, \pi).$$

Therefore μ is a diffeomorphism of the interval $(0, \pi)$ onto $(0, \infty)$. In the interval $(m\pi, (m+1)\pi)$ μ approaches $+\infty$ at the endpoints. In order to find the critical points, we set

$$\frac{1}{2}\mu'(x) = \frac{\sin x - x\cos x}{\sin^3 x} = \frac{1 - x\cot x}{\sin^4 x} = 0.$$

Hence the critical point x_m is the solution of the equation $x = \tan x$ on the interval $(m\pi, (m+1)\pi)$. Note that

$$\begin{aligned}\mu(x+\pi) &= \frac{x+\pi}{\sin^2(x+\pi)} - \cot(x+\pi)\\ &= \frac{x}{\sin^2(x+\pi)} - \cot(x+\pi) + \frac{\pi}{\sin^2 x}\\ &= \mu(x) + \frac{\pi}{\sin^2 x},\end{aligned}$$

so the successive minimum values increase by more than π. From Figure 9.4 we have

$$m\pi < x_m < m\pi + \frac{\pi}{2} = (m + \frac{1}{2})\pi. \tag{9.4.30}$$

Using $x_m = \tan x_m$ yields

$$\cot x_m = \frac{1}{x_m} < \frac{1}{m\pi}. \tag{9.4.31}$$

Let $f(x) = \cot x$. As $f'(x) = -\frac{1}{\sin^2 x} < -1$, there is a ξ between x and y such that

$$f(x) - f(y) = f'(\xi)(x - y) < -(x - y).$$

Hence $x - y < f(y) - f(x)$. Choosing $x = m\pi + \frac{\pi}{2}$, $y = x_m$ and using

$$f(m\pi + \frac{\pi}{2}) = \frac{\cos(m\pi + \frac{\pi}{2})}{\sin(m\pi + \frac{\pi}{2})} = 0,$$

and (9.4.31) yields

$$0 < (m + \frac{1}{2})\pi - x_m < \cot x_m < \frac{1}{m\pi}.$$

■

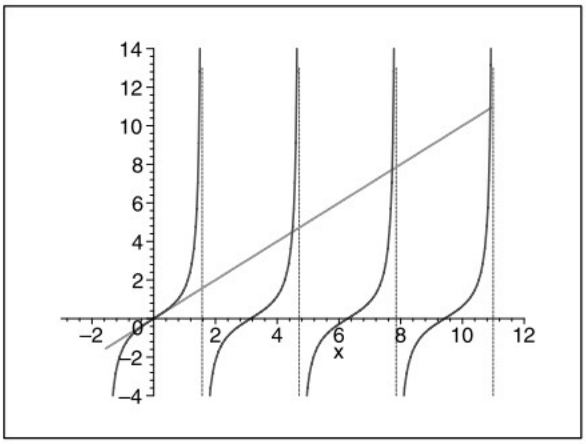

Figure 9.4: Critical points of μ are solutions of $\tan x = x$.

The number of geodesics that join the origin with an arbitrary given point is given in the following theorems.

Theorem 9.21. *(i) Given a point* $P(\mathbf{x}, \mathbf{t})$, $\mathbf{r} = |\mathbf{x}| \neq 0$, $\mathbf{t} > 0$, *there are finitely many geodesics between the origin and* P. *Let* $0 < \zeta_1 < \cdots < \zeta_N$ *be the solutions of*

$$\mathbf{t} - \frac{1}{2}\zeta = \mathbf{r}^2\mu(\zeta). \tag{9.4.32}$$

Then, with $\theta_m = \dfrac{\zeta_m}{2\tau}$, *the geodesic equations are*

$$\begin{aligned} r_m(s) &= \frac{\sin(2\theta_m s)}{\sin(2\theta_m \tau)}\mathbf{r}, \\ \phi_m(s) &= -2\theta_m s + \phi_0, \\ t_m(s) &= \theta_m s + \frac{\mathbf{r}^2}{\sin^2(2\theta_m \tau)}\Big(2\theta_m \tau - \frac{1}{2}\sin(4\theta_m s)\Big), \quad m = 1, 2, \dots N. \end{aligned}$$

(ii) The compatibility condition for the boundary conditions is

$$\zeta_m = \phi_0 - \Phi, \quad m = 1, 2, \dots, N. \tag{9.4.33}$$

Given the point $P(\mathbf{x}, \mathbf{t})$, *let* $\Phi = \arctan(|\mathbf{x}|)$ *be the final argument. Then the initial arguments of the geodesics joining the origin and* P *are*

$$\phi_{0,m} = \zeta_m - \Phi, \quad m = 1, 2, \dots, N. \tag{9.4.34}$$

Proof. (i) It is obvious that equation (9.4.32) has finitely many solutions, see Figure 9.3. For each solution of (9.4.27), substitute θ in the equations (9.4.21), (9.4.22) and (9.4.24).
(ii) It follows from (i) and condition (9.4.23). See Figure 9.5. ■

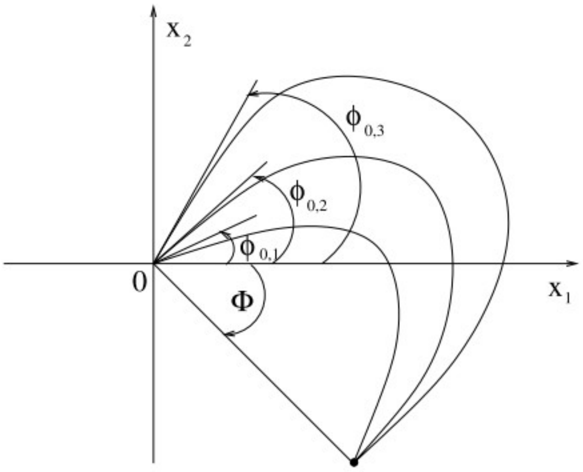

Figure 9.5: The projections of the geodesics on an x-plane start with different arguments.

Remark 9.22 *A similar theorem works for the case* $\mathbf{t} < 0$.

It is well known that locally, there is only one geodesic joining the origin and the point P. The size and the shape of the neighborhood is given by the following result.

Theorem 9.23. *Given a point* $P(\mathbf{x}, \mathbf{t})$*, with* $|\mathbf{t}| < \big(\frac{1}{2} + |\mathbf{x}|^2\big)\pi$ *and* $|\mathbf{x}| \neq 0$*, there is a unique geodesic joining the origin and the point* P.

Proof. We shall discuss the following cases: $0 < \mathbf{t} < \big(\frac{1}{2} + |\mathbf{x}|^2\big)\pi$, $\mathbf{t} = 0$ and $-\big(\frac{1}{2} + |\mathbf{x}|^2\big)\pi < \mathbf{t} < 0$. The third case can be treated in a similar way as the first case.
Case $0 < \mathbf{t} < \big(\frac{1}{2} + |\mathbf{x}|^2\big)\pi$.
We shall show that equation (9.4.32) has only one solution $\zeta > 0$. Consider the function $\varphi(\zeta) = \dfrac{1}{|\mathbf{x}|^2}(\mathbf{t} - \dfrac{1}{2}\zeta)$. We shall show that the solutions of the equation $\varphi(\zeta) = \mu(\zeta)$ are only in the interval $(0, \pi)$. It suffices to show that

$$\varphi(\zeta) < \mu(\zeta), \qquad \text{for} \quad \pi < \zeta. \tag{9.4.35}$$

Let $x_1 \in (\pi, 2\pi)$ be the first critical point of μ. Using Lemma 9.20, the monotonicity of μ and convexity of μ yields

$$\varphi(\zeta) < \varphi(\pi) = \frac{1}{|\mathbf{x}|^2}(\mathbf{t} - \frac{1}{2}\pi) < \pi < \mu(x_1) = \min_{\pi<\zeta} \mu(\zeta),$$

which yields (9.4.35). Then there are no solutions on $(\pi, +\infty)$. As φ is decreasing and μ is increasing on $(0, \pi)$, there is only one solution for the equation $\varphi(\zeta) = \mu(\zeta)$, see Figure 9.6

Case $\mathbf{t} = 0$.
If $\mathbf{t} = 0$, then $-\dfrac{1}{2}\zeta = |\mathbf{x}|\mu(\zeta)$ yields only the solution $\zeta = 0$. Then $\theta = \dfrac{\zeta}{2\tau} = 0$. Theorem 9.21 yields $\phi_0 = \Phi$, $t(s) = 0$. $r(s)$ satisfies $\ddot{r} = 0$, with solution $r(s)|\mathbf{x}|s$. There is a unique solution, which is a straight line from the origin to P, in the x-plane. ■

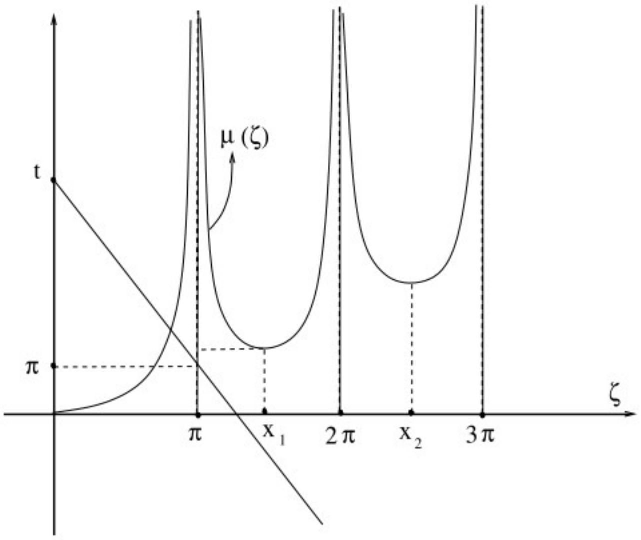

Figure 9.6: The case when $\phi(\zeta) = \mu(\zeta)$ has a unique solution.

Corollary 9.24 *Given a point $P(\mathbf{x}, 0)$, $|\mathbf{x}| \neq 0$, there is a unique geodesic between the origin and P. The geodesic is given by the equations $r(s) = |\mathbf{x}|s$, $\phi(s) = \Phi$, and $t(s) = 0$, i.e., it is a straight segment in the x-plane.*

In Theorems 9.21 and 9.23 we assumed $|\mathbf{x}| \neq 0$. In the following we shall cover the case when $|x| = 0$.

Theorem 9.25. *Given a point $P(0, \mathbf{t})$ on the t-axis, there is a unique geodesic between the origin and the point P.*

Proof. If $|\mathbf{x}| = \mathbf{r} = 0$, from (9.4.21) we get $r(s) = 0$. Using (9.4.24) yields $t(s) = \theta s$, with $\theta = \mathbf{t}/\tau$. The geodesic is along the t-axis. ■

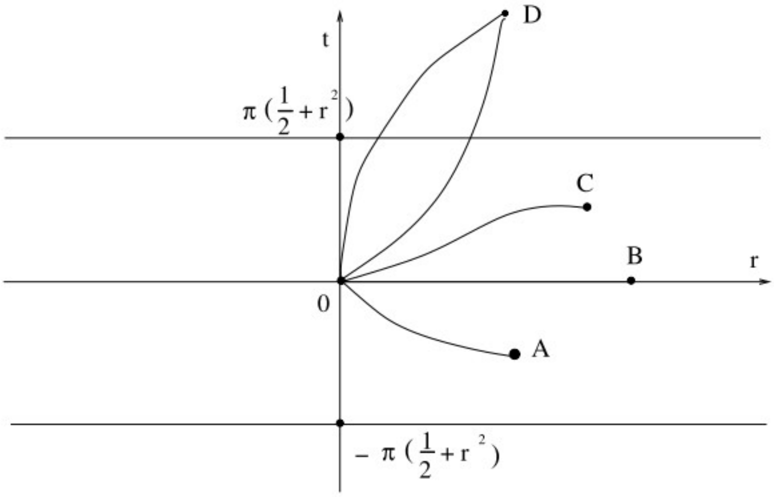

Figure 9.7: There is a unique geodesic in the strip $|\mathbf{t}| < \pi(\frac{1}{2} + \mathbf{r}^2)$ between O and (x, t).

Remark 9.26 *Theorem* 9.23 *works also in the case* $|\mathbf{x}| = 0$.

9.4.2 The classical action

In a strip like in Theorem 9.23 the geodesic is unique. Let θ denote the unique solution. The Lagrangian along the geodesic is

$$\begin{aligned}
L &= \frac{1}{2}(\dot r^2 + r^2\dot\phi^2 + \dot t^2) + 2(r^2\dot\phi + \dot t)r^2\dot\phi \\
&= \frac{1}{2}(\dot r^2 + r^2\dot\phi^2 + \dot t^2) + 2(\theta - r^2\dot\phi)r^2\dot\phi \\
&= \frac{1}{2}(\dot r^2 + r^2\dot\phi^2 + \dot t^2) + 2\theta r^2\dot\phi - 2r^4\dot\phi^2 \\
&= \frac{1}{2}\dot r^2 + \frac{1}{2}r^2\dot\phi^2 + \frac{1}{2}(\theta - 2r^2\dot\phi)^2 + 2\theta r^2\dot\phi - 2r^4\dot\phi^2 \\
&= \frac{1}{2}\dot r^2 + \frac{1}{2}r^2\, 4\theta^2 + \frac{1}{2}\theta^2 \\
&= \frac{1}{2}\dot r^2 + \frac{1}{2}\theta^2(1 + 4r^2) \\
&= \frac{1}{2}\dot r^2 + \frac{1}{2}\theta\dot t.
\end{aligned}$$

The classical action is obtained by integrating the Lagrangian along the geodesic

$$\begin{aligned}
S(\tau) = S(\mathbf{x}, \mathbf{y}, \tau) &= \int_0^\tau L\, ds = \int_0^\tau \left(\frac{1}{2}\dot r^2 + \frac{1}{2}\theta\dot t\right) ds \\
&= \frac{1}{2}\int_0^\tau \dot r^2(s)\, ds + \frac{1}{2}\theta\big(t(\tau) - t(0)\big). \qquad (9.4.36)
\end{aligned}$$

Integrating the first term yields

$$\begin{aligned}\int_0^\tau \dot{r}(s)^2\,ds &= \frac{4\theta^2\mathbf{r}^2}{\sin^2(2\theta\tau)}\int_0^\tau \cos^2(2\theta s)\,ds = \frac{2\theta\mathbf{r}^2}{\sin^2(2\theta\tau)}\int_0^{2\theta\tau}\cos^2 v\,dv\\
&= \frac{2\theta\mathbf{r}^2}{\sin^2(2\theta\tau)}\Big[\theta\tau + \frac{1}{4}\sin(4\theta\tau)\Big]\\
&= \frac{\theta\mathbf{r}^2}{\sin^2(2\theta\tau)}\Big(2\theta\tau + \sin(2\theta\tau)\cos(2\theta\tau)\Big)\\
&= \theta\mathbf{r}^2\Big[\frac{2\theta\tau}{\sin^2(2\theta\tau)} + \cot(2\theta\tau)\Big] = \theta\mathbf{r}^2\widetilde{\mu}(2\theta\tau),\end{aligned} \tag{9.4.37}$$

where

$$\widetilde{\mu}(x) = \frac{x}{\sin^2 x} + \cot x \tag{9.4.38}$$

Proposition 9.27 *The classical action starting at the origin is*

$$\begin{aligned}S(\mathbf{x},\mathbf{t},\tau) &= \theta^2|\mathbf{x}|^2\Big(\frac{1}{2} + \frac{2|\mathbf{x}|^2}{\sin^2(2\theta\tau)}\Big)\\
&= \theta\mathbf{t} - \frac{1}{2}\theta^2\tau + \theta|\mathbf{x}|^2\cot(2\theta\tau).\end{aligned}$$

Proof. Using $t(0) = 0$, substituting (9.4.37) in equation (9.4.36) yields

$$\begin{aligned}S(\mathbf{x},\mathbf{t},\tau) &= \frac{1}{2}\theta\mathbf{r}^2\widetilde{\mu}(2\theta\tau) + \frac{1}{2}\theta\mathbf{t}\\
&= \frac{1}{2}\theta\mathbf{r}^2\widetilde{\mu}(2\theta\tau) + \frac{1}{2}\theta\big(\theta\tau + \mathbf{r}^2\mu(2\theta\tau)\big)\\
&= \frac{1}{2}\theta\mathbf{r}^2\widetilde{\mu}(2\theta\tau) + \frac{1}{2}\theta^2\tau + \frac{1}{2}\theta\mathbf{r}^2\mu(2\theta\tau)\\
&= \frac{1}{2}\theta\mathbf{r}^2\big[\widetilde{\mu}(2\theta\tau) + \mu(2\theta\tau)\big] + \frac{1}{2}\theta^2\tau\\
&= \theta\mathbf{r}^2\frac{2\theta\tau}{\sin^2(2\theta\tau)} + \frac{1}{2}\theta^2\tau\\
&= \theta^2\mathbf{r}\Big(\frac{1}{2} + \frac{2\mathbf{r}^2}{\sin^2(2\theta\tau)}\Big).\end{aligned}$$

For the second identity, using $\mu(x) = \widetilde{\mu}(x) + 2\cot(x)$, we have

$$\begin{aligned}S(\mathbf{x},\mathbf{t},\tau) &= \frac{1}{2}\theta\mathbf{r}^2\Big(\mu(2\theta\tau) + 2\cot(2\theta\tau)\Big) + \frac{1}{2}\theta\mathbf{t}\\
&= \frac{1}{2}\theta\mathbf{r}^2\mu(2\theta\tau) + \theta\mathbf{r}^2\cot(2\theta\tau) + \frac{1}{2}\theta\mathbf{t}\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\theta(\mathbf{t} - \theta\tau) + \theta\mathbf{r}^2\cot(2\theta\tau) + \frac{1}{2}\theta\mathbf{t} \\
&= \frac{1}{2}\theta\mathbf{t} - \frac{1}{2}\theta^2\tau + \theta\mathbf{r}^2\cot(2\theta\tau) + \frac{1}{2}\theta\mathbf{t} \\
&= \theta\mathbf{t} - \frac{1}{2}\theta^2\tau + \theta\mathbf{r}^2\cot(2\theta\tau).
\end{aligned}
$$

Replacing $\mathbf{r}$ by $|\mathbf{x}|$ we obtain the desired formulas. ■

9.4.3 The complex action

The space $(\mathbb{R}^3, h)$ with h given by (9.4.15) is *not a radially symmetric space*. The reason is the fact that the momentum $\theta = \theta(\mathbf{x}, \mathbf{t}, \tau)$, which appears in the classical action given by Proposition 9.27, is a solution of the equation $\mathbf{t} = \theta\tau + |\mathbf{x}|^2\mu(2\theta\tau)$, and hence depends on the boundary conditions $\mathbf{t}$ and $\mathbf{x}$ in a complicated manner. Therefore we do not expect $\Delta_{Cas}S(\mathbf{x}, \mathbf{t}, \tau)$ to be a function that depends just on τ.

However, we can fix the situation. In the next chapter, when computing the heat kernels, we need an action function, which satisfies the Hamilton–Jacobi equation. We define the *complex action* for our problem to be the function obtained by substituting $\theta = -i$ in the classical action. Let $S_{\mathbb{C}}$ denote the complex action. Using the properties $\sin(-ix) = -i\sinh(x)$ and $\cos(-ix) = i\cosh(x)$ yields

$$
S_{\mathbb{C}} = -i\mathbf{t} + \frac{1}{2}\tau + (\mathbf{x}_1^2 + \mathbf{x}_2^2)\coth(2\tau). \tag{9.4.39}
$$

Proposition 9.28 *The complex action* (9.4.39) *satisfies the Hamilton–Jacobi equation*

$$
\frac{\partial S_{\mathbb{C}}}{\partial\tau} + \frac{1}{2}\Big(X_1 S_{\mathbb{C}}\Big)^2 + \frac{1}{2}\Big(X_2 S_{\mathbb{C}}\Big)^2 + \frac{1}{2}\Big(T S_{\mathbb{C}}\Big)^2 = 0. \tag{9.4.40}
$$

Proof. A computation provides

$$
\begin{aligned}
\partial_t S_{\mathbb{C}} &= -i, \\
\partial_{x_1} S_{\mathbb{C}} &= 2\mathbf{x}_1\coth(2\tau), \\
\partial_{x_2} S_{\mathbb{C}} &= 2\mathbf{x}_2\coth(2\tau).
\end{aligned}
$$

$$
\begin{aligned}
2H(\nabla S_{\mathbb{C}}) := {} & \big(X_1 S_{\mathbb{C}}\big)^2 + \big(X_2 S_{\mathbb{C}}\big)^2 + \big(T S_{\mathbb{C}}\big)^2 \\
= {} & \big(\partial_{x_1} S_{\mathbb{C}} + 2\mathbf{x}_2\partial_t S_{\mathbb{C}}\big)^2 + \big(\partial_{x_2} S_{\mathbb{C}} - 2\mathbf{x}_1\partial_t S_{\mathbb{C}}\big)^2 + \big(\partial_t S_{\mathbb{C}}\big)^2 \\
= {} & \big(2\mathbf{x}_1\coth(2\tau) - 2i\mathbf{x}_2\big)^2 + \big(2\mathbf{x}_2\coth(2\tau) + 2i\mathbf{x}_1\big)^2 + (-i)^2 \\
= {} & 4\mathbf{x}_1^2\coth^2(2\tau) - 4\mathbf{x}_2^2 - 8i\mathbf{x}_1\mathbf{x}_2\coth(2\tau) \\
& +4\mathbf{x}_2^2\coth^2(2\tau) - 4\mathbf{x}_1^2 + 8i\mathbf{x}_1\mathbf{x}_2\coth(2\tau) - 1 \\
= {} & 4|\mathbf{x}|^2\coth^2(2\tau) - 4|\mathbf{x}|^2 - 1.
\end{aligned} \tag{9.4.41}
$$

. ! ! 3 4 5s

9(:69 9 92 9 $, (& ! :
? ;- C (33 ! : * ; 3 & . 3!

2. F * ! ===

' " "" 3 / *

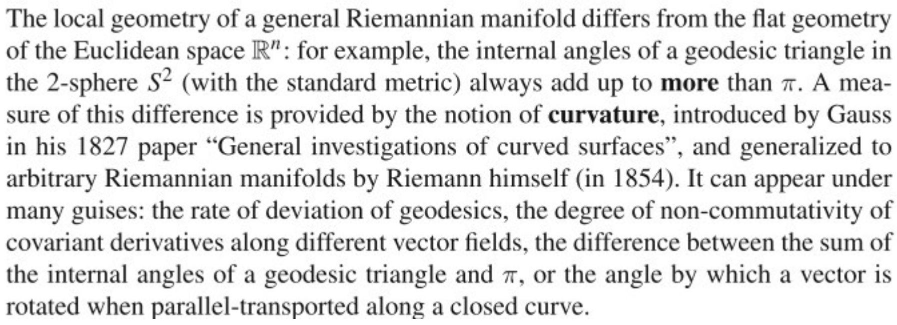

The local geometry of a general Riemannian manifold differs from the flat geometry of the Euclidean space $\mathbb{R}^n$: for example, the internal angles of a geodesic triangle in the 2-sphere S^2 (with the standard metric) always add up to **more** than π. A measure of this difference is provided by the notion of **curvature**, introduced by Gauss in his 1827 paper "General investigations of curved surfaces", and generalized to arbitrary Riemannian manifolds by Riemann himself (in 1854). It can appear under many guises: the rate of deviation of geodesics, the degree of non-commutativity of covariant derivatives along different vector fields, the difference between the sum of the internal angles of a geodesic triangle and π, or the angle by which a vector is rotated when parallel-transported along a closed curve.

9(:69 9 92 9 $, P !. B E
(33 @ 8 < ; > ? ; , 3!

' "" : F * *

9(:69 9 92 9 $, P !. B E
(33 @ E > A) > ? ; , 3!

9(:69 9 92 9 $, > A) !
! A C ! . 3 ; % 3!

9(:69 9 92 9 $, E H ! < ; >

2. 3

& 3 F * < ! === 77777

9(:69 9 92 9 $, A (t2 - u<:& 2 3!

2. F * ! ===

4 9(:69 9 92 9 $, K 2H ; ! < ; @
2H ! N * ; @ N (3!

' " " E $ "

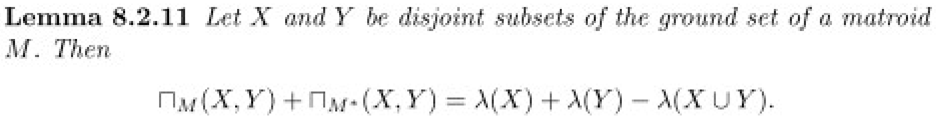

Lemma 8.2.11 *Let X and Y be disjoint subsets of the ground set of a matroid M. Then*

$$\sqcap_M(X,Y) + \sqcap_{M^*}(X,Y) = \lambda(X) + \lambda(Y) - \lambda(X \cup Y).$$

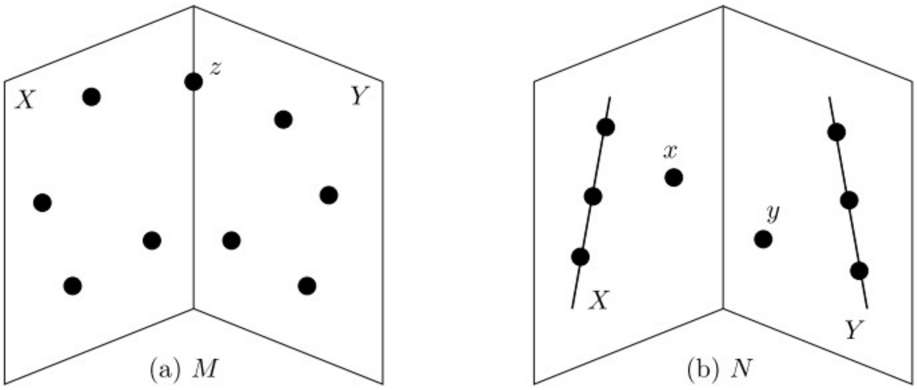

(a) M (b) N

FIG. 8.3. $z \in \mathrm{cl}_M(X) \cap \mathrm{cl}_M(Y)$ and $y \in \mathrm{cl}^*{}_N(X) \cap \mathrm{cl}^*{}_N(Y \cup x)$.

Proof The result follows easily by substitution using the facts that X and Y are disjoint and $\lambda(Z) = r_M(Z) + r_{M^*}(Z) - |Z|$. □

In a matroid M, a set X or a pair $(X, E - X)$ is *exactly k-separating* if $\lambda_M(X) = k - 1$. Similarly, a k-separation (X, Y) is *exact* if $\lambda_M(X) = k - 1$.

Example 8.2.12 Consider the rank-4 matroid M for which a geometric representation is shown in Figure 8.3(a). Let z be as labelled, let X be the set of four points on the left-hand plane other than z, and let Y be the corresponding set of four points on the right-hand plane. It is not difficult to verify that M has no 1- or 2-separations and that $(X, Y \cup z)$ and $(X \cup z, Y)$ are exact 3-separations. Thus the element z can be moved from one side of the 3-separation to the other. We observe that $z \in \mathrm{cl}(X) \cap \mathrm{cl}(Y)$.

Next let N be the rank-4 matroid shown in Figure 8.3(b), where X and Y are 3-point lines as indicated. Then N has no 1- or 2-separations and has each of $(X, Y \cup \{x, y\}), (X \cup x, Y \cup y)$, and $(X \cup \{x, y\}, Y)$ as an exact 3-separation. Moreover, although it is perhaps less obvious from the figure, $(X \cup y, Y \cup x)$ is also an exact 3-separation. In this case, we have, for example, that $y \in \mathrm{cl}^*(X) \cap \mathrm{cl}^*(Y \cup x)$, where we recall that $e \in \mathrm{cl}^*(S) - S$ if and only if e is a coloop of $M \backslash S$. □

The phenomenon that was just exemplified can be generalized.

& 3 777

N F $ 3 3 3

N v ! 3 w3 4 x * . 3

B A * 1 ! r

, $ - "#? : * ;#

B ? : * ;@ ; ! (3 ; y @

B 9(:69 9 92 9 $, z;* ! ? { @ .4 P * B
A | ; * ; @ 3 3!

B" 9(:69 9 92 9 $,
OC -: ! Q-: * ;-- 3 @- @- 3!

' % $ 3 $

* ! 3 ! . ! * ;

& 3

, $ # ? ;

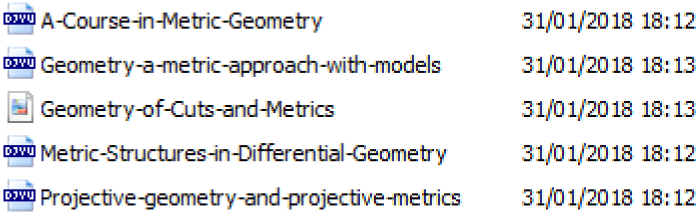

A-Course-in-Metric-Geometry	31/01/2018 18:12
Geometry-a-metric-approach-with-models	31/01/2018 18:13
Geometry-of-Cuts-and-Metrics	31/01/2018 18:13
Metric-Structures-in-Differential-Geometry	31/01/2018 18:12
Projective-geometry-and-projective-metrics	31/01/2018 18:12

9(:69 9 92 9 $, ? ; 3

B/ 9(:69 9 92 9 $, (? ; !4*

B 9(:69 9 92 9 $, ? ; 33
! !4*

B% 9(:69 9 92 9 $, ? ; ! 3!

B0 9(:69 9 92 9 $, >
? ; !4*

B$ 9(:69 9 92 9 $, 4 * ; ! 3 4 *
!4*

E E

9(:69 9 92 9 $, E E - < ; ! (33
"@ 8 B S 8 3!

' 0 "

Hence $\langle \varphi_i, \varphi_j \rangle = 0$ for all $i \neq j$. Thus, $(\varphi_i)_{i=1}^M$ is an orthonormal system which is complete by (ii), and (iii) is proved.

(iv) If $(\varphi_i)_{i=1}^M$ is a frame, by (ii), it is also a spanning set for $\mathscr{H}^N$. Towards a contradiction, assume that $(\varphi_i)_{i=1}^M$ is linearly dependent. Then there exist some $i_0 \in \{1, \dots, M\}$ and values λ_i, $i \in I := \{1, \dots, M\} \setminus \{i_0\}$ such that

$$\varphi_{i_0} = \sum_{i \in I} \lambda_i \varphi_i .$$

This implies that $(\varphi_i)_{i \in I}$ is also a frame, thus contradicting exactness of the frame. □

Before presenting some insightful basic results in frame theory, we first discuss some examples of frames to develop an intuitive understanding.

1.3.2 Examples

By Lemma 1.2 (iii), orthonormal bases are unit norm Parseval frames (and vice versa). However, applications typically require *redundant* Parseval frames. One basic way to approach this construction problem is to build redundant Parseval frames using orthonormal bases, and we will present several examples in the sequel. Since the associated proofs are straightforward, we leave them to the interested reader.

Example 1.1 Let $(e_i)_{i=1}^N$ be an orthonormal basis for $\mathscr{H}^N$.

(1) The system

$$(e_1, 0, e_2, 0, \dots, e_N, 0)$$

is a Parseval frame for $\mathscr{H}^N$. This example indicates that a Parseval frame can indeed contain zero vectors.

(2) The system

$$\left(e_1, \frac{e_2}{\sqrt{2}}, \frac{e_2}{\sqrt{2}}, \frac{e_3}{\sqrt{3}}, \frac{e_3}{\sqrt{3}}, \frac{e_3}{\sqrt{3}}, \dots, \frac{e_N}{\sqrt{N}}, \dots, \frac{e_N}{\sqrt{N}} \right)$$

is a Parseval frame for $\mathscr{H}^N$. This example indicates two important issues. First, a Parseval frame can have multiple copies of a single vector. Second, the norms of vectors of an (infinite) Parseval frame can converge to zero.

We next consider a series of examples of non-Parseval frames.

Example 1.2 Let $(e_i)_{i=1}^N$ be an orthonormal basis for $\mathscr{H}^N$.

(1) The system

$$(e_1, e_1, \dots, e_1, e_2, e_3, \dots, e_N)$$

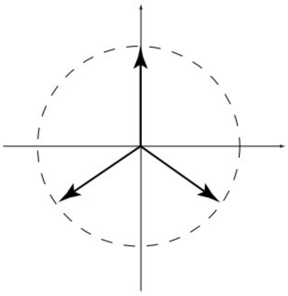

Fig. 1.1 Mercedes-Benz frame

with the vector e_1 appearing $N+1$ times, is a frame for $\mathscr{H}^N$ with frame bounds 1 and $N+1$.

(2) The system

$$(e_1, e_1, e_2, e_2, e_3, e_3, \ldots, e_N)$$

is a 2-tight frame for $\mathscr{H}^N$.

(3) The union of L orthonormal bases of $\mathscr{H}^N$ is a unit norm L-tight frame for $\mathscr{H}^N$, generalizing (2).

A particularly interesting example is the smallest truly redundant Parseval frame for $\mathbb{R}^2$, which is typically coined the *Mercedes-Benz frame*. The reason for this naming becomes evident in Fig. 1.1.

Example 1.3 The *Mercedes-Benz frame* for $\mathbb{R}^2$ is the equal norm tight frame for $\mathbb{R}^2$ given by:

$$\left(\sqrt{\tfrac{2}{3}}\begin{pmatrix}0\\1\end{pmatrix}, \sqrt{\tfrac{2}{3}}\begin{pmatrix}\frac{\sqrt{3}}{2}\\-\frac{1}{2}\end{pmatrix}, \sqrt{\tfrac{2}{3}}\begin{pmatrix}-\frac{\sqrt{3}}{2}\\-\frac{1}{2}\end{pmatrix}\right)$$

Note that this frame is also equiangular.

For more information on the theoretical aspects of equiangular frames we refer to [60, 91, 120, 139]. A selection of their applications is reconstruction without phase [5, 6], erasure-resilient transmission [15, 102], and coding [136]. We also refer to the Chaps. 4, 5 in this book for more details on equiangular frames.

Another standard class of examples can be derived from the *discrete Fourier transform* (*DFT*) *matrix*.

Example 1.4 Given $M \in \mathbb{N}$, we let $\omega = \exp(\frac{2\pi i}{M})$. Then the DFT matrix in $\mathbb{C}^{M\times M}$ is defined by

$$D_M = \frac{1}{\sqrt{M}}\left(\omega^{jk}\right)_{j,k=0}^{M-1}.$$

This matrix is a unitary operator on $\mathbb{C}^M$. Later (see Corollary 1.11) it will be seen that the selection of any N rows from D_M yields a Parseval frame for $\mathbb{C}^N$ by taking the associated M column vectors.

There also exist particularly interesting classes of frames such as Gabor frames utilized primarily for audio processing. Among the results on various aspects of Gabor frames are uncertainty considerations [113], linear independence [119], group-related properties [89], optimality analysis [127], and applications [67, 74, 75, 87, 88]. Chapter 6 provides a survey on this class of frames. Another example is the class of group frames, for which various constructions [24, 101, 147], classifications [64], and intriguing symmetry properties [146, 148] have been studied. A comprehensive presentation can be found in Chap. 5.

1.4 Frames and Operators

For the rest of this introduction we set $\ell_2^M := \ell_2(\{1, \dots, M\})$. Note that this space in fact coincides with $\mathbb{R}^M$ or $\mathbb{C}^M$, endowed with the standard inner product and the associated Euclidean norm.

The analysis, synthesis, and frame operators determine the operation of a frame when analyzing and reconstructing a signal. The Gramian operator is perhaps not that well known, yet it crucially illuminates the behavior of a frame $(\varphi_i)_{i=1}^M$ embedded as an N-dimensional subspace in the high-dimensional space ℓ_2^M.

1.4.1 Analysis and Synthesis Operators

Two of the main operators associated with a frame are the analysis and synthesis operators. The analysis operator—as the name suggests—analyzes a signal in terms of the frame by computing its frame coefficients. We start by formalizing this notion.

Definition 1.15 Let $(\varphi_i)_{i=1}^M$ be a family of vectors in $\mathscr{H}^N$. Then the associated *analysis operator* $T : \mathscr{H}^N \to \ell_2^M$ is defined by

$$Tx := \big(\langle x, \varphi_i \rangle\big)_{i=1}^M, \quad x \in \mathscr{H}^N.$$

In the following lemma we derive two basic properties of the analysis operator.

Lemma 1.3 *Let $(\varphi_i)_{i=1}^M$ be a sequence of vectors in $\mathscr{H}^N$ with associated analysis operator T.*

(i) *We have*

$$\|Tx\|^2 = \sum_{i=1}^M \big|\langle x, \varphi_i \rangle\big|^2 \quad \textit{for all } x \in \mathscr{H}^N.$$

Hence, $(\varphi_i)_{i=1}^M$ is a frame for $\mathscr{H}^N$ if and only if T is injective.

R: A ; 3 !

"/-R *) * - ,#

2 r3

9(:69 9 92 9 $, < R: A ; 3 !
3 R 3!

9(:69 9 92 9 $, < R: A ; 3 !
3 N ,0 3!

& ,

9(:69 9 92 9 , H < E *
C P E 3!

' , 0 (33)

A geometric approach

Exercise 9 on page 40 considered the problem of finding the roots of $f(x) = 1 - 2x^2$. A bit of algebra confirms that $\sqrt{2}/2$ is a root—i.e., $f(\sqrt{2}/2) = 0$. The question is: what is the *numerical value* of $\sqrt{2}/2$?

We'll answer this question by constructing a sequence of approximations that add digits, one at a time, to an estimate for $\sqrt{2}/2$. Since the root lies at the point where the graph of f crosses the x-axis, we just magnify the graph at this point over and over again, "trapping" the point between x values that can be made arbitrarily close together.

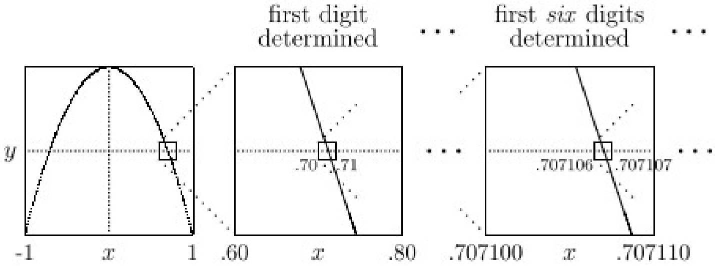

The graph of $y = 1 - 2x^2$ under successive magnifications

If we make each stage a ten-fold magnification over the previous one, then, as we zoom in on the next smaller interval that contains the root, one more digit in our estimate will be stabilized. The first six stages are described in the table below. They tell us $\sqrt{2}/2 = .707106\ldots$ to six decimal places accuracy.

Since this method of finding roots requires only that we be able to plot successive magnifications of the graph of f on a computer screen, the method can be applied to any function that can be entered into a computer.

The positive root of $1 - 2x^2$

when the root lies between:		the decimal expansion of the root begins with
lower value	upper value	
.70	.80	.7
.700	.710	.70
.7070	.7080	.707
.70710	.70720	.7071
.707100	.707110	.70710
.7071060	.7071070	.707106
⋮	⋮	⋮

Copyright 1994, 2008 Five Colleges, Inc.
DVI file created at 11:39, 1 February 2008

In the limit process we take $\Delta x \to 0$ because Δx is the width of the field of view. Since five digits of $\Delta y/\Delta x$ have stabilized, we can write

$$f'(27) = -.27235\ldots.$$

To find $f'(x)$ at some other point x, proceed the same way. Magnify the graph at that point repeatedly, until the value of the slope stabilizes. The method is very powerful. In the exercises you will have an opportunity to use it with other functions.

By using a microscope of arbitrarily high power, we have obtained further insights about rates and slopes. In fact, with these insights we can now state definitively what we mean by the slope of a curved graph and the rate of change of a function.

Definition. The **slope** of a graph at a point is the *limit* of the slopes seen in a microscope at that point, as the field of view shrinks to zero.

Definition. The **rate of change** of a function at a point is the slope of its graph at that point. Thus the rate of change is also a limit.

To calculate the value of the slope of the graph of $f(x)$ when $x = a$, we have to carry out a limit process. We can break down the process into these four steps:

1. Magnify the graph at the point $(a, f(a))$ until it appears straight.
2. Calculate the slope of the magnified segment.
3. Repeat steps 1 and 2 with successively higher magnifications.
4. Take the limit of the succession of slopes produced in step 3.

Local Linearity

A microscope gives a local view

The crucial property of a microscope is that it allows us to look at a graph **locally**, that is, in a small neighborhood of a particular point. The two functions we have been studying in this section have curved graphs—like most functions. But *locally*, their graphs are straight—or nearly so. This is

Copyright 1994, 2008 Five Colleges, Inc.
DVI file created at 11:39, 1 February 2008

a remarkable property, and we give it a name. We say these functions are **locally linear**. In other words, a locally linear function looks like a linear function, locally.

The graph of a linear function has a definite slope at every point, and so does a *locally* linear function. For a linear function, the slope is easy to calculate, and it has the same value at every point. For a locally linear function, the slope is harder to calculate; it involves a limit process. The slope also varies from point to point.

How common is local linearity? All the standard functions you already deal with are locally linear almost everywhere. To see why we use the qualifying phrase "almost everywhere," look at what happens when we view the graph of $y = f(x) = x^{2/3}$ with a microscope. At any point other than the origin, the graph is locally linear. For instance, if we view this graph over the interval from 0 to 2 and then zoom in on the point $(1, 1)$ by two successive powers of 10, here's what we see:

All the standard functions are locally linear at almost all points

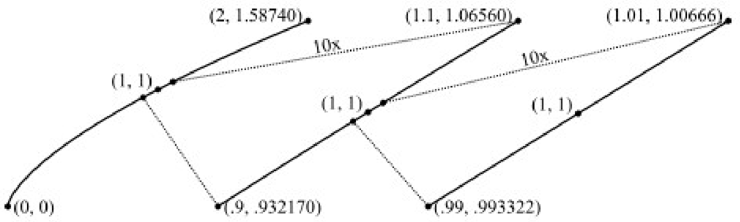

As the field of view shrinks, the graph looks more and more like a straight line. Using the highest magnification given, we estimate the slope of the graph—and hence the rate of change of the function—to be

$$f'(1) \approx \frac{\Delta y}{\Delta x} = \frac{1.006656 - .993322}{1.01 - .99} = \frac{.013334}{.02} = .6667.$$

Similarly, if we zoom in on the point $(.001, .01)$ we get:

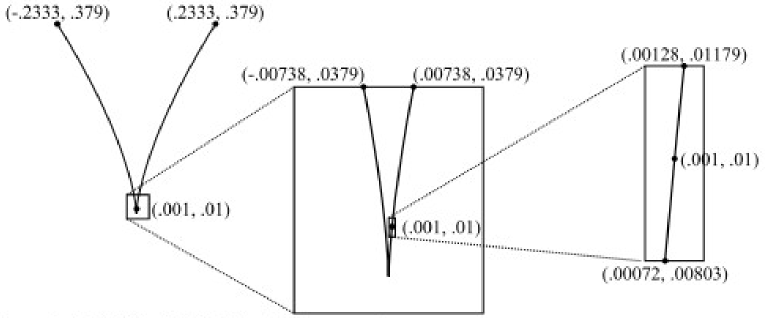

Copyright 1994, 2008 Five Colleges, Inc.
DVI file created at 11:39, 1 February 2008

In the last window the graph looks like a line of slope

$$f'(.001) \approx \frac{.0118 - .0082}{.00128181 - .00074254} = \frac{.0036}{.00053937} = 6.674.$$

At the origin, though, something quite different happens:

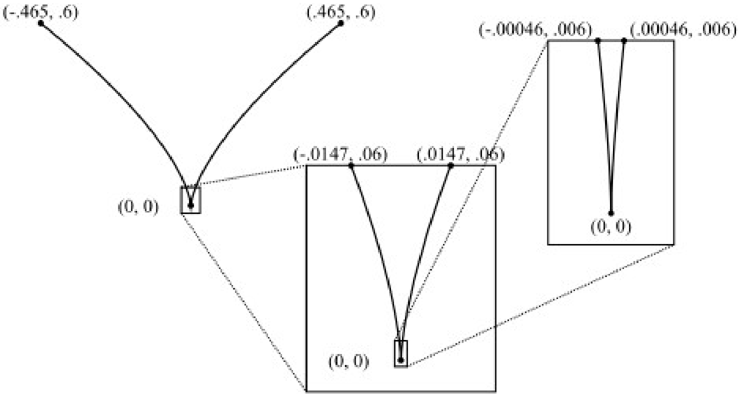

The graph simply looks more and more sharply pointed the closer we zoom in to the origin—it never looks like a straight line. However, the origin turns out to be the only point where the graph does not eventually look like a straight line.

Fractals are locally *non*-linear objects

In spite of these examples, it is important to realize that local linearity is a very special property. There are some functions that fail to be locally linear anywhere! Such functions are called **fractals**. No matter how much you magnify the graph of a fractal at any point, it continues to look non-linear—bent and "pointy" in various ways. In recent years fractals have been used in problems where the more common (locally linear) functions are inadequate. For instance, they describe irregular shapes like coastlines and clouds, and they model the way molecules are knocked about in a fluid (this is called *Brownian motion*). However, calculus does not deal with such functions. On the contrary,

Calculus studies functions that are locally linear almost everywhere.

Copyright 1994, 2008 Five Colleges, Inc.
DVI file created at 11:39, 1 February 2008

. 9(:69 9 92 9 , P ! !| ; (
! | 3 (;@ A! > * 3 3!

' /, % AH $

AH $ & 3 3
3 @ ! !
3 3

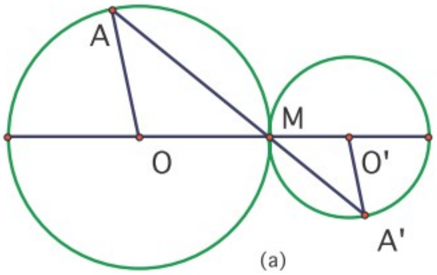

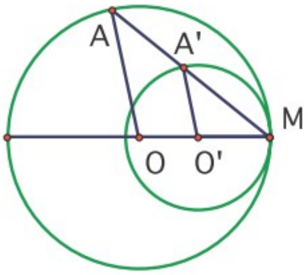

E $

Solution. If line AA' passes through the point M of tangency of the circles O and O' (Figures t58a and 58b), then $\widehat{OAM} = \widehat{OMA} = \widehat{O'MA'} = \widehat{O'A'M}$, so that OA is parallel to $O'A'$.

Notes. This is another good problem for students to cut their teeth on. If they have solved Exercise 57, they may be able to recognize the need to consider both internally and externally tangent circles.

AH , < ;
! ! ! ! .;) ;@
< H ; @
@ ! ! .;) ; ! ! !
.;) ;

Solution. The circles remain tangent because the conditions $OO' = |R - R'|$ (for internally tangent circles) or $OO' = R + R'$ (for externally tangent circles) remains true when the radii are changed as described in the problem statement. See **71** in the text.

Notes. This problem complements Exercise 57, as it involves the same conditions for tangency for two circles.

AH % E ! !3 ! 3
DH ! 3

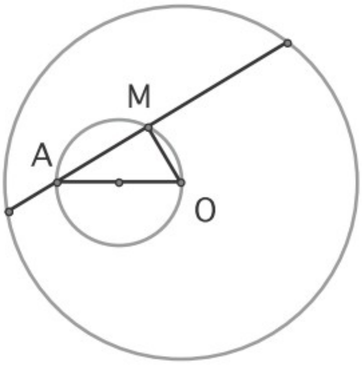

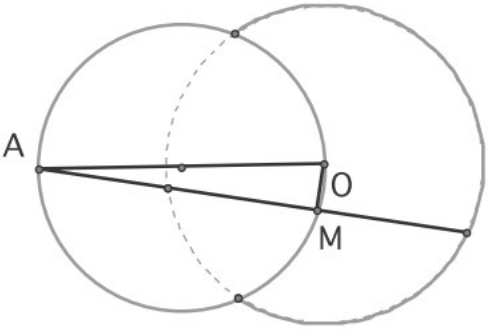

E %

& 2 * ! (* 3 @
2(. ! !3 ; ! 3
(<) ! ! 2((!
* E % @ @ (! *
E % .

R < DH ! 3 !
H3 @ ! . . ;3 3
! 3 3 B . !; B <
3 . @ H3 ! .; ! @ .
33) B ; ! B
. 3 ! DH ! 3
!3 ! & ! @ 3
3 @ * ; .
< DH ! 3 ! ! !
! ; ! G B H ! ! ;
w3 x 3 ! (3 !)
B

' % E %

AH % 2 ! ; @ @
) ! !3 ! DH ! !
E ! !3 . !

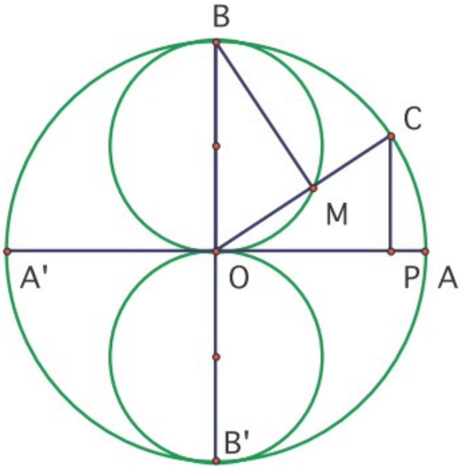

E %

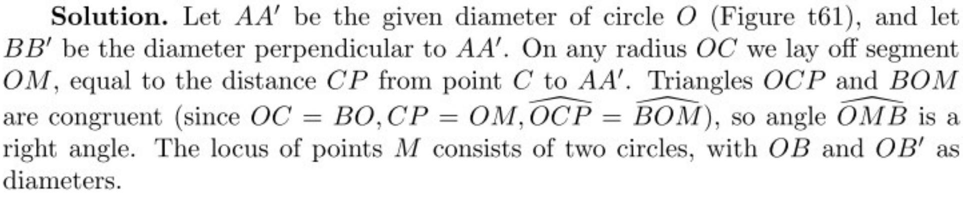

Solution. Let AA' be the given diameter of circle O (Figure t61), and let BB' be the diameter perpendicular to AA'. On any radius OC we lay off segment OM, equal to the distance CP from point C to AA'. Triangles OCP and BOM are congruent (since $OC = BO, CP = OM, \widehat{OCP} = \widehat{BOM}$), so angle $\widehat{OMB}$ is a right angle. The locus of points M consists of two circles, with OB and OB' as diameters.

& ! ; 3 .
. B ! @ 3
@ ! 3 !3 !
! 3 ! . B ! 3 ! !
R (!; B @) ! ! @
! ! ! 3 ! .; !
DH ! < ; ; D ! * . H3 !
) ! * ! @ .
! ! !
H3

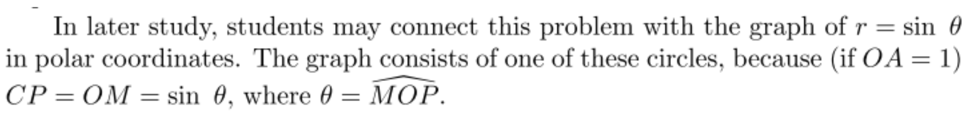

In later study, students may connect this problem with the graph of $r = \sin\ \theta$ in polar coordinates. The graph consists of one of these circles, because (if $OA = 1$) $CP = OM = \sin\ \theta$, where $\theta = \widehat{MOP}$.

. 9(:69 9 92 9 , P ! ! <;3 E
> A) @ & ! &) 3 0 3!

." 9(:69 9 92 9 , K) P ! ! F 5 !
v v v v & v v 3 $,$@ (! 3!

./ 9(:69 9 92 9 , K) P ! ! F 5 !
v v v v && v v ! | 3 , @ (! 3!

. 9(:69 9 92 9 , B P ! !|
? ; (; 3!

9(:69 9 92 9 , R ; 3
8 P ! . / 3!

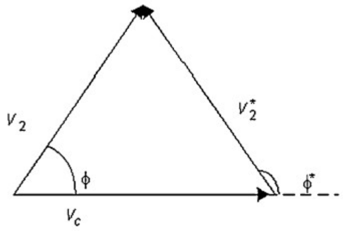

Fig. 1.5 Velocity triangle for the recoil particle

1.2.6 Limits on the Scattering Angle θ

Case (i) $m_2 > m_1$, or $\gamma < 1$; i.e. $v_1^* > v_c$.

In Fig. 1.6, the circle is drawn with O as the centre and radius $OP = v_1^*$. A is a point within the circle such that $AO = v_c$, and the line AOB represents the incident direction. As before, the lab velocity of m_1 is v_1 which is obtained by compounding v_1^* and v_c vectorially. The lab angle $\theta =$ angle PAO and the CM angle $\theta^* =$ angle POB. As the point P moves counterclockwise on the circumference, θ^* increases and so does θ. When P approaches P', $\theta^* = \theta = \pi$. Thus, θ increases monotonically from 0 to π, and in this case there is no restriction on the scattering angle in the LS. In other words, m_1 can be scattered in completely backward direction.

Case (ii) $m_2 = m_1$, or $\gamma = 1$; i.e. $v_1^* = v_c$.

Here A lies on the circumference of the circle (Fig. 1.7). As θ^* increases, θ also increases. But when P approaches A, PA becomes tangential at A and so $\theta \rightarrow \frac{\pi}{2}$. θ varies from 0 to π. Thus, in this case $m_1$can be scattered up to a maximum angle of $\pi/2$ but not beyond. In other words, backward scattering in the LS is not permissible.

Case (iii) $m_2 < m_1$, or $\gamma > 1$, i.e. $v_1^* < v_c$.

Here A lies outside the circle (Fig. 1.8). There are two positions P and P^l for which the same scattering angle θ is obtained for two different values of θ^*. As P moves back on the circumference, θ increases. The maximum angle θ_m is reached when AP becomes tangent to the circle (Fig. 1.9). In that case

$$\sin\theta_m = \frac{OP}{AO} = \frac{v_1^*}{v_c} = \frac{m_2}{m_1} \quad \text{or}$$

$$\theta_m = \sin^{-1}(m_2/m_1)$$

Thus, there is a limitation on the scattering angle when $m_2 < m_1$. θ first increases from 0 to a maximum value $\sin^{-1}(1/\gamma)$ which is less than $\pi/2$, as θ^* increases from 0 to $\cos^{-1}(-1/\gamma)$. θ then decreases to 0 as θ^* further increases to π. At a given angle θ between 0 and $\sin^{-1}(1/\gamma)$, there will be two groups of particles associated with different velocities corresponding to the two values of θ^*.

Fig. 1.6 Limits on the scattering angle for $m_2 > m_1$, θ max $= \pi$

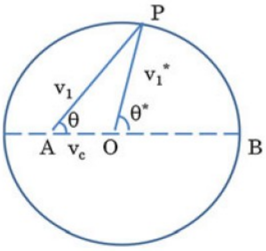

Fig. 1.7 Limits on the scattering angle θ for $m_1 = m_2$, θ max $= \pi/2$

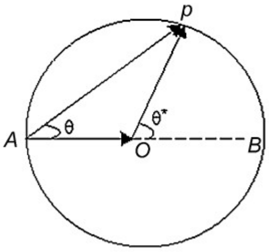

Fig. 1.8 Limits on the scattering angle θ. For $m_2 < m_1$, $\theta(\max) = \sin^{-1}(\frac{m_2}{m_1})$

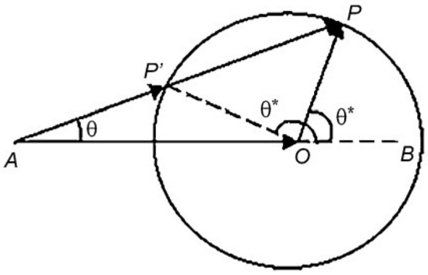

Fig. 1.9 $\theta_{\max}$ is reached when AP is a tangent to the circle

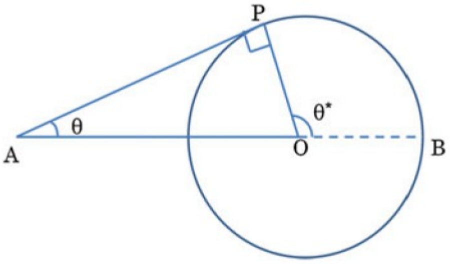

1.2.7 Limits on the Recoil Angle ϕ

Since $v_2^* = v_c$ (always), $\phi = \frac{1}{2}\phi^*$ by (1.15). Since the maximum angle of ϕ^* is π, the maximum angle ϕ_m is $\frac{1}{2}\pi$. In other words, the target particle cannot recoil in the backward hemisphere.

Fig. 1.10 Azimuth angle β is measured with respect to the positive axis in the xy plane $\perp$ to the direction of incidence

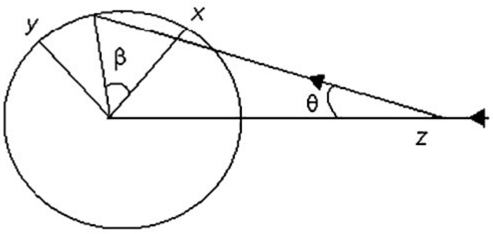

Fig. 1.11 Element of solid angle $d\Omega = ds/r^2$

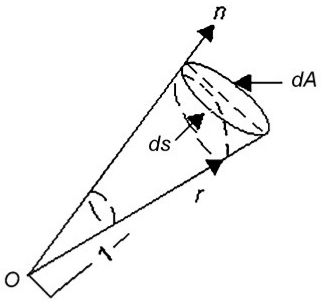

1.2.8 *Scattering in Three Dimensions*

Since scattering is described under central forces, a particle which is incident on a target particle and initially moves in a certain plane would be necessarily confined to this plane after the scattering because of the conservation of angular momentum. Thus, a single scattering event is completely described in two dimensions. However, in practice one is concerned with a flux of particles incident say along the z-axis on a target material. Since various particles proceed in different planes, the scattering on the whole will be in three dimensions. In order to fix the orientation of the plane of scattering, we need to introduce the azimuth angle β which is measured with respect to the positive x-axis in the xy plane, Fig. 1.10. We must also consider the element of solid angle into which the particles are scattered. This is illustrated in Fig. 1.11.

Let dA denote an element of surface area and connect all points on the boundary of dA to O so as to form a cone. Let ds be the area of that portion of a sphere with O as the centre and radius r which is cut out by this cone. The solid angle subtended by dA at O is defined as $d\Omega = ds/r^2$ and is numerically equal to the area cut out by a sphere with centre O and unit radius. From Fig. 1.12, it is seen that $ds = r^2 \sin\theta d\theta d\beta$ so that $d\Omega = \sin\theta d\theta d\beta = 2\pi \sin\theta d\theta$, where we have integrated over $d\beta$. When the scattering is independent of the azimuth angle then the area subtended at O is due to the entire circular strip, $ds = 2\pi r^2 \sin\theta d\theta$ as in Fig. 1.13, so that the element of solid angle $d\Omega = 2\pi \sin\theta d\theta$. Observe that the maximum solid angle is 4π since it is given by the entire surface area of a sphere $(4\pi r^2)$ divided by r^2.

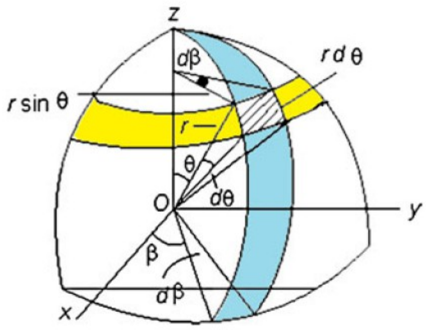

Fig. 1.12 Elements of solid angle for a general case

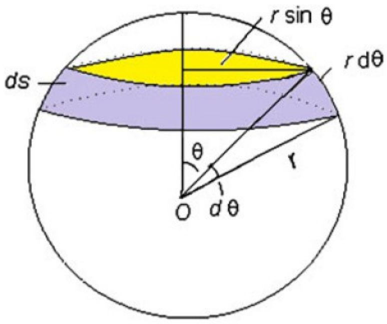

Fig. 1.13 Element of solid angle for an azimuthal symmetry

1.2.9 Scattering Cross-Section

In order to describe the angular distribution of particles scattered by target particles which are initially stationary, the concept of cross-section is introduced. Let a uniform parallel flux of N_0 particles be incident per unit area normal to the direction per unit time on a group of n scattering centres. Let N particles be scattered per unit time into a small solid angle $d\Omega$ centred towards a direction which has polar angle θ and azimuth angle β with respect to the incident direction as polar axis. N will be proportional to N_0, n and $d\Omega$ provided the flux is small enough to ensure that the incident particles do not interfere with one another, that there is no appreciable decrease in the number of scattering centres on account of their being knocked out due to collisions and that the incident particles are far enough apart so that each collision is made only by one of them.

The number of incident particles that emerge per unit time in $d\Omega$ can be written as:

$$N = nN_0\sigma(\theta,\beta)d\Omega \tag{1.16}$$

where the proportionality factor $\sigma(\theta,\beta)$ is called the differential scattering cross-section. The quantity $\sigma(\theta,\beta)$ is a measure of the probability of scattering in a given direction (θ,β), per unit solid angle from the given nucleus.

' @ , 0 E

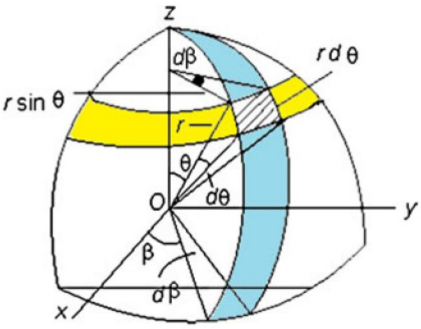
z
dβ
r d θ
r sin θ
r
θ
dθ
O
y
β
dβ
x

E "

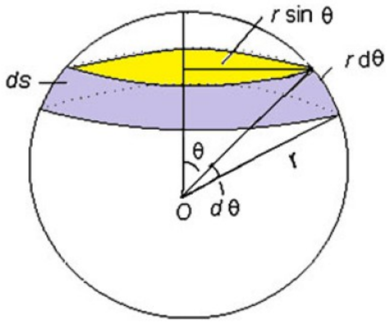
r sin θ
r dθ
ds
θ
r
O
d θ

9(:69 9 92 9 ,
P * ;- --< -E -(- - 3 -< ;-8 3!

' , 0 E < 3

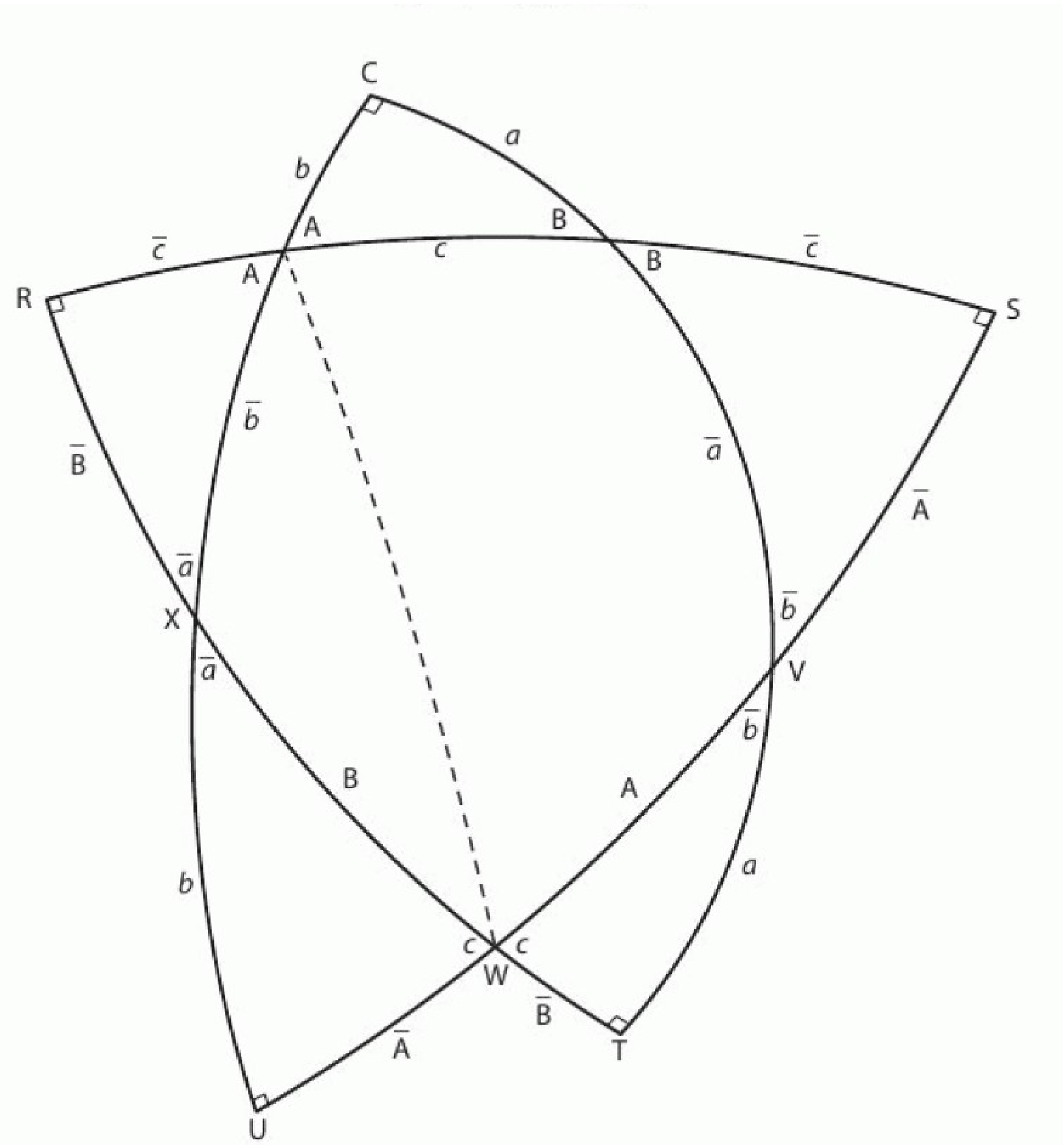
C
a
b
A
B
$\bar{c}$
c
$\bar{c}$
A
B
R
S
$\bar{b}$
$\bar{a}$
$\bar{B}$
$\bar{A}$
$\bar{a}$
X
$\bar{b}$
$\bar{a}$
V
$\bar{b}$
B
A
a
b
c
c
W
$\bar{B}$
$\bar{A}$
T
U

' / , E %%

< 3 > . | ;

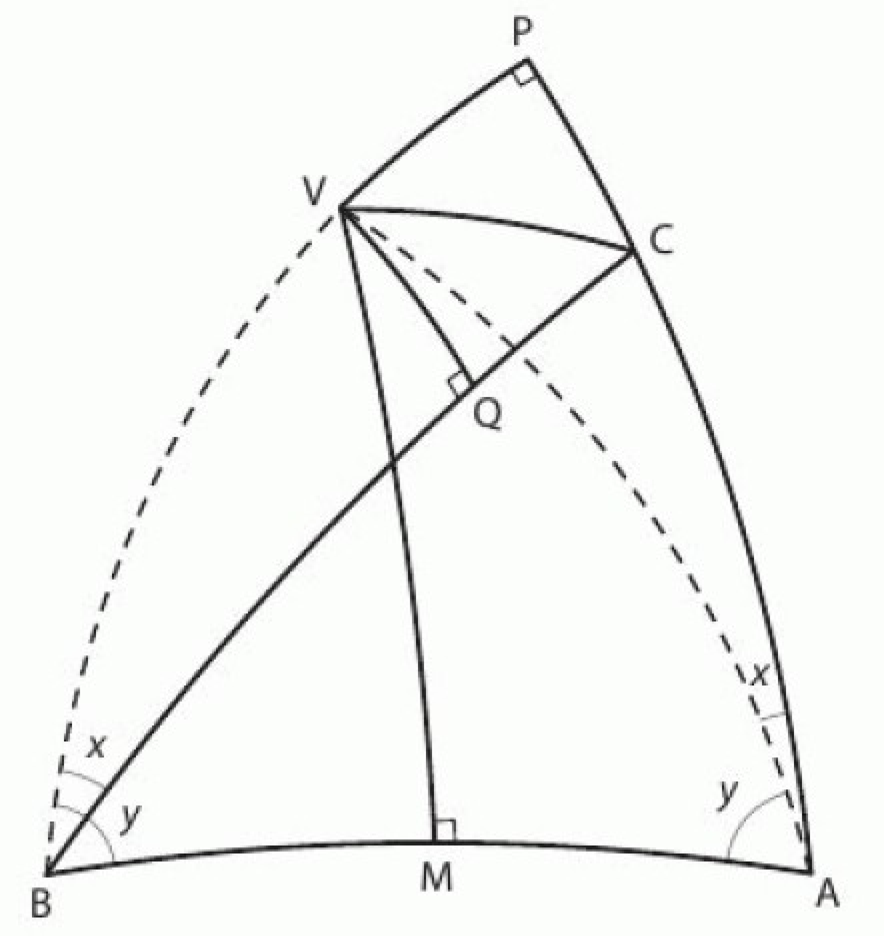
P
V
C
Q
x
y
x
y
B
M
A

? 3

care, a navigator might compute a distance that would leave us hundreds of miles short of our destination!

The mathematical face of this ambiguity appears immediately when we use the Law of Sines to find $\measuredangle A$: $\sin A = \sin a \sin B / \sin b = 0.9985$, so $A = 86.84^\circ$ or $180^\circ - 86.84^\circ = 93.16^\circ$. We can tell from our diagram that we want the smaller angle 86.84°, but if the ambiguity had passed without notice, we might have run out of fuel at X, with no land in sight.

We now know a, b, A, and B; but finding c and C proves surprisingly awkward. The Law of Sines doesn't help since there is no way to apply it without leaving two unknowns in the equation. It is possible to use the Law of Cosines (rearranging the letters in the diagram appropriately), but only by changing $\sin c$ into $\sqrt{1-\cos^2 c}$, eventually requiring us to solve a quadratic equation for $\cos c$. A cleaner approach would be nice. Fortunately, there are several possibilities.

Delambre's and Napier's Analogies

It's surprisingly common, and rather eerie, when mathematical discoveries are made almost simultaneously by two or more people working on their own. The invention of calculus by Isaac Newton and Gottfried Wilhelm Leibniz is the most famous episode of this kind, followed by the birth of non-Euclidean geometry in the works of János Bolyai, Nicolai Lobachevsky, and Carl Friedrich Gauss in the early 19th century. Bolyai's father wrote to Gauss of his son's breakthrough, only to receive the disheartening reply from an impressed Gauss that he had done it all already, but had not bothered to publish it.

This incident was not the first of Gauss's experiences with simultaneous discovery. The theorems we are about to describe were known first as *Gauss's formulas*, appearing in his 1809 *Theoria motus corporum coelestium*, a monumental two-volume treatise on the motions of celestial bodies. But this time it was Gauss who had been scooped, by mere months; and not once, but twice. In the previous year German scientist Karl Brandon Mollweide had published the same results in Leipzig, and he in turn referred to their appearance in a book by Antonio Cagnoli. However, in a rare case when a name actually changed when a prior discovery was verified, the theorems are now known as *Delambre's analogies* after the astronomer Jean-Baptiste Joseph Delambre, who came up briefly in chapter 5 for maligning Napier's Rules. Delambre likely would prefer to be remembered for his contributions to celestial mechanics, his work on determining the length of the meter, and his books on the history of astronomy. But his discovery and publication of the analogies in 1807 in the long-running French astronomical journal *Connaissance des Temps* (oddly the volume was dated 1809, leading to some of the confusion) grant him a measure of additional fame by crossing the finish line only months before his unwitting rivals.

> →Delambre's analogies are usually demonstrated disappointingly through algebraic manipulation of various known identities. It is not easy to approach them geometrically, but at least one textbook (Isaac Todhunter's classic) gives it a go. The argument, based on Delambre's original demonstration, begins with ΔABC in figure 6.6. We begin by bisecting $\overset{\frown}{AB}$ at M and drawing a perpendicular upward. Next bisect $\measuredangle BCP$ to form $\overset{\frown}{CV}$. Drop arcs perpendicular to $\overset{\frown}{CP}$ and $\overset{\frown}{CB}$, defining P and Q respectively. Finally, join $\overset{\frown}{AV}$ and $\overset{\frown}{BV}$. We show first that ΔAVP and ΔBVQ are equal, by matching three elements of the triangles. They both contain a right angle; since $\overset{\frown}{MV}$ is perpendicular to $\overset{\frown}{AB}$ we know that $\overset{\frown}{AV} = \overset{\frown}{BV}$; finally, since $\measuredangle BCP$ was bisected by $\overset{\frown}{CV}$ we know that $\overset{\frown}{PV} = \overset{\frown}{QV}$. So the two triangles are equal.*
>
> This fact allows us to label the angles at A and B as we have done in the diagram. The angles at the base of the original triangle are $\measuredangle A = y + x$ and $\measuredangle B = y - x$; a little algebra yields $x = (A - B)/2$ and $y = (A + B)/2$. We are in a similar algebraic situation with respect to sides a and b of the original triangle: $a = \overset{\frown}{BQ} + \overset{\frown}{CQ}$ and $b = \overset{\frown}{AP} - \overset{\frown}{CP}$, but $\overset{\frown}{BQ} = \overset{\frown}{AP}$ (since triangles AVP and BVQ are equal) and $\overset{\frown}{CQ} = \overset{\frown}{CP}$, so $\overset{\frown}{BQ} = (a + b)/2$ and $\overset{\frown}{CQ} = (a - b)/2$. Turning to the angles at the top of the diagram, we know that $\measuredangle AVB = \measuredangle PVQ$ since both are composed of $\measuredangle AVQ$ and equal angles. Cutting these angles in half, we have $\measuredangle AVM = \measuredangle PVC$.

---

* As my students gleefully pointed out, there is an error in the argument here. The two triangles are asserted to be equal by side-side-angle equivalence, but side-side-angle is the ambiguous case. The fact that the angle is right seems to allow the argument to escape unscathed from the ambiguity, but this turns out not to be quite true on the sphere. I leave it to my eagle-eyed students and readers to explore the flaw and discover how to patch it.

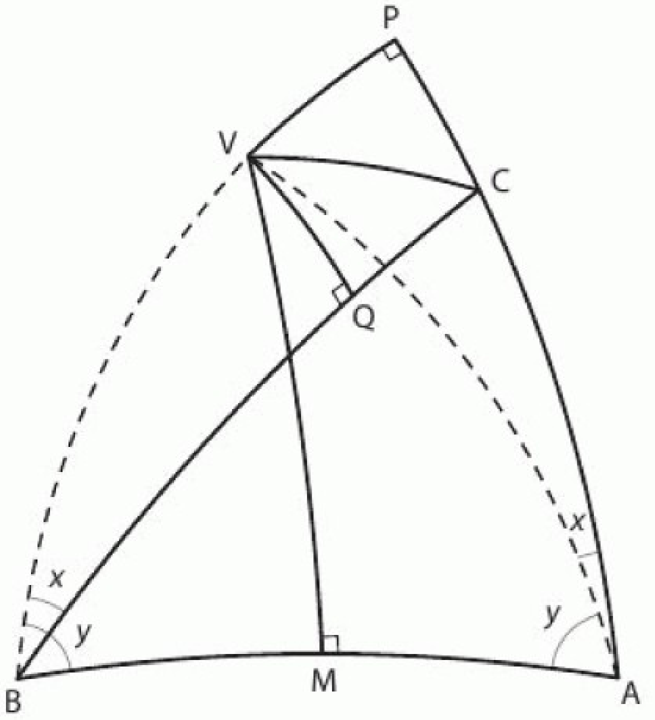

Figure 6.6. The proof of Delambre's first analogy.

We're in the home stretch of the proof. Consider the two right-angled triangles ΔAVM and ΔPVC, with equal angles at V. Apply Geber's Theorem to both and bring the results together:

$$\sin\measuredangle VAM \cos\widehat{AM} = \cos\measuredangle AVM = \sin\measuredangle PCV \cos\widehat{CP}.$$

Substituting each of the angles for their values from the elements of the original triangle and rearranging the terms, we finally arrive at *Delambre's first analogy*: →

$$\frac{\sin\frac{1}{2}(A+B)}{\cos\frac{1}{2}C} = \frac{\cos\frac{1}{2}(a-b)}{\cos\frac{1}{2}c}.$$

Admittedly this argument, as clever as it is, is not much of an advertisement for geometric over algebraic proofs. We shall explore an algebraic derivation in the exercises. Delambre's other three analogies,

$$\frac{\sin\frac{1}{2}(A-B)}{\cos\frac{1}{2}C} = \frac{\sin\frac{1}{2}(a-b)}{\sin\frac{1}{2}c},$$

$$\frac{\cos\frac{1}{2}(A+B)}{\sin\frac{1}{2}C} = \frac{\cos\frac{1}{2}(a+b)}{\cos\frac{1}{2}c}, \text{ and}$$

$$\frac{\cos\frac{1}{2}(A-B)}{\sin\frac{1}{2}C} = \frac{\sin\frac{1}{2}(a+b)}{\sin\frac{1}{2}c},$$

may be demonstrated similarly.

The word "analogy" might seem strange here. We use it in what is now an obsolete mathematical sense that goes back to the original Greek meaning of the word: a ratio, or an equality between ratios. Even today, one might think of a ratio as a comparison, or an "analogy," between two quantities.

It has taken some work to secure these identities, apparently for the purpose of dealing with oblique triangles when a, b, A, and B are known. But each of these identities refers to *all six* of the triangle's elements, and so can be used only if five elements are known, not four! In fact, in most textbooks Delambre's identities are used only as a tool to check the correctness of completed solutions to triangles. Has our effort gone to waste?

Thankfully not. We are on the verge of another set of analogies, named after John Napier (although Napier's friend and successor Henry Briggs actually contributed two of the four). Several terms appear more than once in Delambre's analogies; this suggests that we might combine the analogies by eliminating the common terms. For instance, the first and third of Delambre's analogies both contain $\cos\frac{1}{2}c$, so if we divide one by the other, we get

$$\frac{\tan\frac{1}{2}(A+B)}{\cot\frac{1}{2}C} = \frac{\cos\frac{1}{2}(a-b)}{\cos\frac{1}{2}(a+b)}.$$

We arrive at three other identities in a similar fashion:

$$\frac{\tan\frac{1}{2}(A-B)}{\cot\frac{1}{2}C} = \frac{\sin\frac{1}{2}(a-b)}{\sin\frac{1}{2}(a+b)},$$

$$\frac{\tan\frac{1}{2}(a+b)}{\tan\frac{1}{2}c} = \frac{\cos\frac{1}{2}(A-B)}{\cos\frac{1}{2}(A+B)}, \text{ and}$$

$$\frac{\tan\frac{1}{2}(a-b)}{\tan\frac{1}{2}c} = \frac{\sin\frac{1}{2}(A-B)}{\sin\frac{1}{2}(A+B)}.$$

Napier's analogies each contain only five triangle elements, not six. So for our ambiguous triangle, once we know a, b, A, and B, any one of these four identities will give us either c or C. For instance, for our Honolulu to Tokyo trip, we substitute into Napier's third analogy to get

$$\frac{\tan\frac{1}{2}(68.69^\circ + 54.3^\circ)}{\tan\frac{1}{2}c} = \frac{\cos\frac{1}{2}(86.84^\circ - 60.5^\circ)}{\cos\frac{1}{2}(86.84^\circ + 60.5^\circ)},$$

from which we have $\tan\frac{1}{2}c = 0.5317$, and so $c = 28^\circ$, equal to 1680 nautical miles or 1933 statute miles. And if we're really interested, we can calculate the last unknown C using almost any of the identities we've seen in this chapter.

There is plenty more to spherical trigonometry than we've seen so far, but this concludes the basic theory needed for solving triangles. In the remaining three chapters we'll see some special topics and applications.

Exercises

1. Solve the following triangles:
   (a) $a = 135^\circ$, $b = 120^\circ$, $c = 45^\circ$
   (b) $A = 68.72^\circ$, $B = 104.35^\circ$, $C = 47.62^\circ$
   (c) $b = 48.62^\circ$, $c = 78.85^\circ$, $C = 128.77^\circ$.
   [Brink 1942, 26 # 1 and 4, 43 # 3]
2. Find the perimeter of a spherical triangle with angles 69°, 84°, 100°, upon a sphere whose radius is 10 inches. [Rothrock 1911, 136]
3. Find the length of the shortest air route between Cape Town (33.93° S, 18.47° E) and Dakar (14.67° N, 17.42° W). What is the bearing of this journey as the plane leaves Cape Town?
4. A ship sailing on a great circle from Ceylon to Madagascar crosses the meridian of 79° east longitude bearing S 50° W. After sailing 2060 nautical miles farther, it crosses the meridian of 52° east longitude. Find its latitude and the bearing of its course at this point and its latitude at the first point. [Brink 1942, 41]
5. Charles A. Lindbergh flew his plane *The Spirit of St. Louis* on the great circle route from New York (40.75° N, 73.97° W) to Paris (48.83° N, 2.33° E). He left New York at 7:52 AM (Eastern Standard Time) on May 20, 1927, and arrived at Paris the next day at 5:24 PM (Eastern Standard Time). What was his average speed for the flight? [Rosenbach/Whitman/Moskovitz 1937, 332]
6. We are at the southwest corner of an open field. We walk across the field to the northeast corner, departing with a heading 49° east of north and traveling exactly one nautical mile to the northeast corner. At the end of our journey, our heading is now 49.01° east of north. What is our terrestrial latitude? (*Warning:* use as many digits of precision as your calculator allows.)
7. Given the latitudes and longitudes of two places on the earth's surface, show how to find the shortest distance between them. [Anderegg/Roe 1896, 107]

8. From Napier's third analogy,

$$\frac{\cos\frac{1}{2}(A-B)}{\cos\frac{1}{2}(A+B)}=\frac{\tan\frac{1}{2}(a+b)}{\tan\frac{1}{2}c},$$

show that in any spherical triangle, one-half the sum of two angles is in the same quadrant as one-half the sum of the opposite sides, that is, $\frac{1}{2}(a+b)$ and $\frac{1}{2}(A+B)$ are in the same quadrant. [Kells/Kern/Bland 1942, 81 # 4]
9. We are going to develop an alternate solution to an oblique spherical triangle in the side-side-angle (SSA) case using right-angled triangles. Say we are given sides a, b, and the angle A. Drop a perpendicular from C to a point D on side c. (Note: there are two different possible configurations. You may just solve the case where the perpendicular falls on the side.) Solve the right-angled triangles and show how this will give the solution to the original triangle. [Casey 1889, 60]

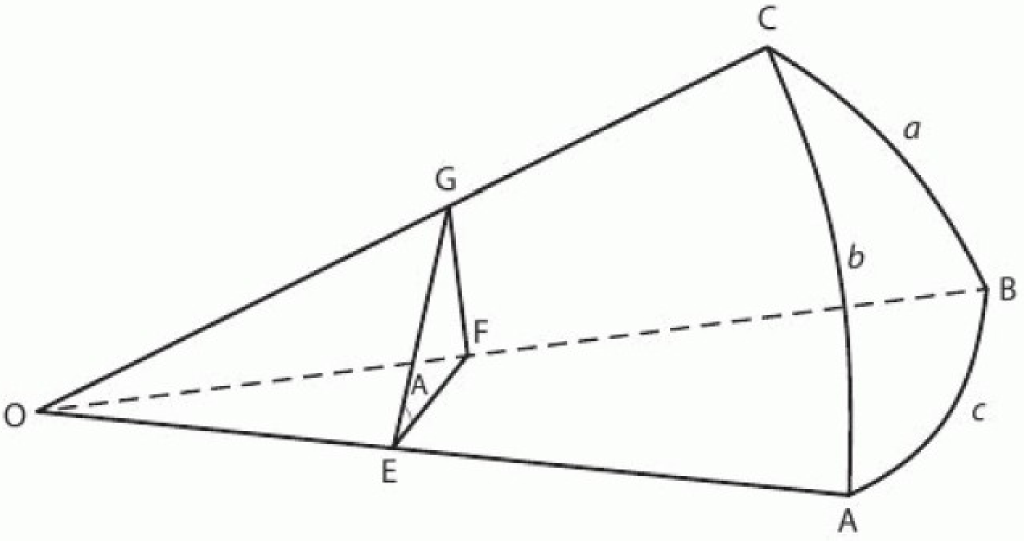

Figure E-6.10.

10. Develop an alternate proof of the Law of Cosines using figure E-6.10, where $\measuredangle OEF$ and $\measuredangle OEG$ are right angles by construction (but $\measuredangle EFG$ is not necessarily a right angle). *Hint:* Apply the planar Law of Cosines to triangles OGF and GFE to solve for their common side and combine the two statements. Simplify the new statement using the Pythagorean theorem on OFE and OEG. Then solve for $\cos a$. [Moritz 1913, 38–39 # 5]
11. Derive the Law of Sines algebraically from the Law of Cosines. (*Hint:* Solve for $\cos A$ in the equation $\cos a = \cos b \cos c + \sin b \sin c \cos A$, form $\sin^2 A$, and reduce the numerator to a form involving cosines only. Then show that $\sin^2 A/\sin^2 a$ is symmetrical in a, b, c.) [Kells/Kern/Bland 1942, 73 # 4]

12. When the polar duality theorem is applied to one of Delambre's analogies, another of Delambre's analogies results. Which pair? What happens if you apply the polar duality theorem to the other two analogies?
13. Napier's analogies were sometimes used as a method of solving geographical problems. Suppose that we know the latitudes of Edmonton (53.6° N) and Vancouver (49.3° N), as well as their difference in longitude (9.6°). The Law of Cosines is the obvious choice here, but instead use Napier's analogies to determine the headings of a great circle path between the cities, and then use some other identity to determine the distance.
14. Here we shall construct an algebraic proof of Napier's first analogy. First, notice from the Law of Sines that $\frac{\sin A}{\sin a} = \frac{\sin B}{\sin b} = \frac{\sin A + \sin B}{\sin a + \sin b}$. Call this ratio m.
(a) Use the Law of Cosines for Angles, twice (once expressed for $\cos A$, once for $\cos B$) to derive the expression

$$(\cos A + \cos B)(1 + \cos C) = m \sin C \sin(a + b).$$

(b) Divide the Law of Sines expression above by your result from (a), to get

$$\frac{\sin A + \sin B}{\cos A + \cos B} = \frac{\sin a + \sin b}{\sin(a+b)} \cdot \frac{1+\cos C}{\sin C}.$$

(c) Transform each of the terms in your expression from (b) using the following identities from plane trigonometry:

$$\sin x + \sin y = 2 \sin \tfrac{1}{2}(x + y) \cos \tfrac{1}{2}(x - y)$$

$$\cos x + \cos y = 2 \cos \tfrac{1}{2}(x + y) \cos \tfrac{1}{2}(x - y)$$

$$\sin 2x = 2 \sin x \cos x$$

$$1 + \cos x = 2 \cos^2 (x/2).$$

(d) Simplify. [Clough-Smith 1978, 76]
15. Although it is possible to derive Delambre's analogies algebraically on their own, it is easier to derive them algebraically from Napier's analogies.
(a) Square Napier's first analogy; solve for the tangent-squared term, and add one to both sides. This should leave you with

$$\sec^2 \tfrac{1}{2}(A + B) = \frac{\cos^2 \tfrac{1}{2}(a-b) \cos^2 \tfrac{1}{2}C + \cos^2 \tfrac{1}{2}(a+b) \sin^2 \frac{1}{2}C}{\cos^2 \tfrac{1}{2}(a+b) \sin^2 \frac{1}{2}C}.$$

(b) Apply the cosine half-angle formula $\cos^2 \frac{\theta}{2} = \frac{1}{2}(1 + \cos\theta)$ to two of the terms in the numerator. Then continue to simplify the numerator, until you arrive at $\frac{1}{2}(1 + \cos a \cos b + \sin a \sin b \cos C)$.

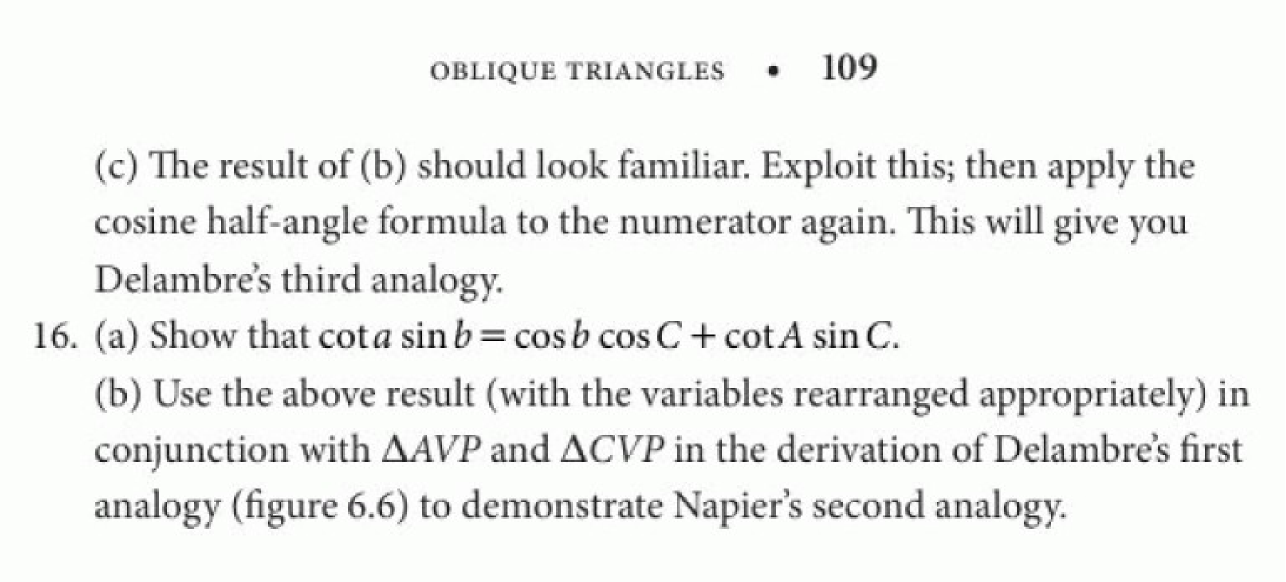

OBLIQUE TRIANGLES • 109

(c) The result of (b) should look familiar. Exploit this; then apply the cosine half-angle formula to the numerator again. This will give you Delambre's third analogy.

16. (a) Show that $\cot a \sin b = \cos b \cos C + \cot A \sin C$.

(b) Use the above result (with the variables rearranged appropriately) in conjunction with ΔAVP and ΔCVP in the derivation of Delambre's first analogy (figure 6.6) to demonstrate Napier's second analogy.

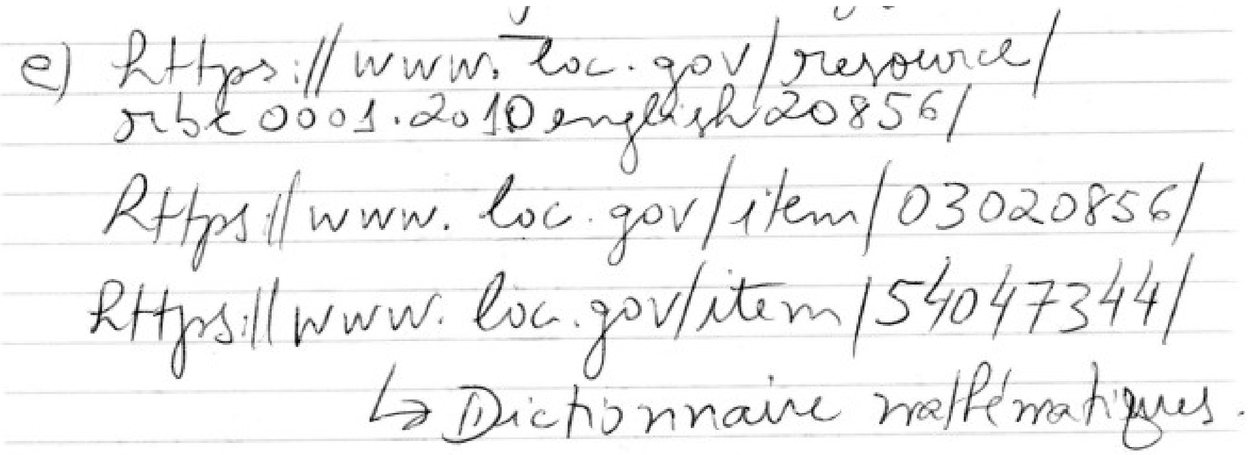

e) https://www.loc.gov/resource/
rbc0001.2010english20856/
https://www.loc.gov/item/03020856/
https://www.loc.gov/item/54047344/
↳ Dictionnaire mathématiques.

9(:69 9 92 9 , * . .
$ % $ % 3!

3 ## *# # . $ %#

& ,

9(:69 9 92 9 , &* * < !
6 E ! @ 3 3!

' 0 " " * * E

3.3 CENTRAL CONSERVATIVE FORCES

*A force **F** is said to be central if is directed towards or away from a fixed point*. In this course we shall mainly deal with central forces acting according to the law of inverse squares:

$$F=-\frac{C}{r^2}. \tag{3.17}$$

The electrostatic and gravitational forces, acting between two immobile particles has such form. If one of the particles is positioned at the origin of the co-ordinate system and the other at a distance r, these central forces are always directed along the radius-vector **r**:

$$\mathbf{F}=-\frac{C}{r^2}\mathbf{r}^0=-\frac{C}{r^2}\frac{\mathbf{r}}{r}. \tag{3.18}$$

For the gravitational force between two particles of masses m_1 and m_2, we have $C=G_0m_1m_2$, where $G_0=6,67.10^{-11}\ \text{N.m}^2/\text{kg}^2$ is the gravitational constant, and for the electrostatic force between two charges $\mathbf{q_1}$ and $\mathbf{q_2}$, $C=\mathbf{q_1}\mathbf{q_2}/4\pi\varepsilon_0$, where ε_0 is the permittivity of the vacuum. These forces are conservative and can be expressed through the potential energy

$$\mathbf{F}=-\frac{\partial U(r)}{\partial \mathbf{r}},\quad U(r)=-\frac{C}{r}. \tag{3.19}$$

We have chosen the arbitrary constant of U so that $U(\infty)=0$. It is preferable to deal with the scalar $U(r)$ than with the vector $\mathbf{F}(\mathbf{r})$.

We shall note that $C>0$ corresponds to attractive forces (a gravitational force or an electrostatic one between charges of opposite signs) and $C<0$ corresponds to repulsive forces (an electrostatic force between charges of equal signs). In atomic physics another central force plays a basic role – the nuclear force. From experiments on scattering of elementary particles, it is known that the nuclear force acts on a small distance between two nucleons (protons or neutrons) with a potential energy

$$U(r)=-\frac{C\exp(-r/r_0)}{r}. \tag{3.20}$$

Here $r_0 \approx 2.10^{-15}$ m, $C \approx 10^{-27}$ J.m. It is easy to see evaluate that at $r \le r_0$ the nuclear force between two protons dominates the Coulomb force of their repulsion. We shall note that (3.20) is an approximation.

Since central forces are always collinear to the radius-vector **r**, their torque is zero

$$\mathbf{N} = \mathbf{r} \times \mathbf{F} = 0. \tag{3.21}$$

According to (2.39) this means that the angular momentum is constant

$$\mathbf{L} = \mathbf{r} \times \tilde{m}\dot{\mathbf{r}} = \text{const}, \tag{3.22}$$

where $\tilde{m}$ is the reduced mass of the particles interacting with the force (3.18). This is the law of conservation of angular momentum and it contains two statements: that its direction is constant and that its magnitude is constant. The angular momentum is perpendicular to the plane of motion - the vector **L** is along the normal to the plane of **r** and $\dot{\mathbf{r}}$. Hence the statement that the direction of **L** is constant implies that **r** and $\dot{\mathbf{r}}$ must always lie in a fixed plane, which is perpendicular to **L**. From here it is evident that the orbit of the particle lies entirely in one plane. The description of such motion requires only two variables, which may be taken as the polar co-ordinates r and φ. Then the Lagrangian function can be written as

$$\mathscr{L} = \frac{\tilde{m}}{2}\left(\dot{\mathbf{r}}^2 + \mathbf{r}^2\dot{\varphi}^2\right) - U(r). \tag{3.23}$$

The co-ordinate φ is a cyclic co-ordinate (see Section 2.1) and the generalized momentum $P_\varphi = \partial\mathscr{L}/\partial\dot{\varphi}$ is a constant of the motion:

$$P_\varphi = \tilde{m}r^2\dot{\varphi} = \text{const}. \tag{3.24}$$

We shall verify that this quantity is equal to the angular momentum. The velocity $\dot{\mathbf{r}}$ (3.22) has components in the directions of the unit vectors $\mathbf{r}^0$ and $\boldsymbol{\varphi}^0$. Indeed, as it is shown in Fig. 3-5 the displacement $d\mathbf{r}$ of the particle, due to the increments of the elements dr and $d\varphi$, can be represented as

$$d\mathbf{r} = dr\mathbf{r}^0 + rd\varphi\,\boldsymbol{\varphi}^0 \tag{3.25}$$

and dividing by dt, we obtain

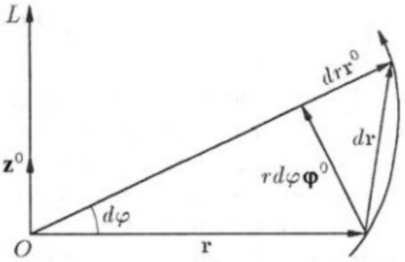

Figure 3-5. The motion of a particle of a reduced mass $\tilde{m}$ in a central field with a centre O is performed in a plane, normal to **L** and is described with the polar co-ordinates r and φ ($\mathbf{r}^0$ and $\boldsymbol{\varphi}^0$ are corresponding unit vectors). At the elementary change of the co-ordinates with dr and $d\varphi$ the particle displaces with $d\mathbf{r}$.

$$\mathbf{v} = \frac{d\mathbf{r}}{dt} = \dot{r}\mathbf{r}^0 + r\dot{\varphi}\,\boldsymbol{\varphi}^0. \tag{3.26}$$

The substitution of (3.26) into the expression for **L** (3.22) yields

$$\mathbf{L} = \mathbf{r}\times\tilde{m}\mathbf{v} = r\,\mathbf{r}^0\times\tilde{m}\left(\dot{r}\,\mathbf{r}^0 + r\dot{\varphi}\boldsymbol{\varphi}^0\right) = 0 + \tilde{m}r^2\dot{\varphi}\left(\mathbf{r}^0\times\boldsymbol{\varphi}^0\right) = \tilde{m}r^2\dot{\varphi}\mathbf{z}^0. \tag{3.27}$$

The magnitude of **L** is $L = L_z = \tilde{m}r^2\dot{\varphi}.$ We proved that under the motion in central-force field the generalized momentum P_φ has a meaning of angular momentum:

$$L = L_z = P_\varphi = \tilde{m}r^2\dot{\varphi}. \tag{3.28}$$

The law of conservation of angular momentum has an interesting geometrical interpretation. When the angle φ changes by an amount $d\varphi$ the position vector **r** sweeps out an area dS (Fig. 3-5):

$$dS = r^2 d\varphi/2. \tag{3.29}$$

The rate of change of this area, which is called an area velocity or a sector velocity, is

$$\frac{dS}{dt} = \frac{1}{2}r^2\dot{\varphi} = \frac{L}{2\tilde{m}} = \text{const.} \tag{3.30}$$

*For equal intervals of time the radius-vector sweeps equal surfa*ce. This law is known as *second Kepler's law* of the motion of the planets. As we saw, it is valid for any central force, not only for a gravitational one.

The forces acting according to the inverse-square law are central and conservative ones. Besides the law of angular-momentum conservation, the law of the conservation of energy is also valid:

$$E = \frac{\tilde{m}}{2}\left(\dot{r}^2 + r^2\dot{\varphi}^2\right) + U\left(r\right). \tag{3.31}$$

We have found two constants of motion (3.28) and (3.31), which express the laws of conservation of angular momentum and energy. (These constants of motion are very often called first integrals of the motion, as they contain first derivatives of the co-ordinates.) Important information can be obtained directly from these equations, without solving them to determine r and φ as functions of time. Eliminating $\dot{\varphi}$ in (3.31) with the help of (3.28), we obtain an equation for r and $\dot{r}$ only:

$$E = \frac{1}{2}\tilde{m}\dot{r}^2 + \frac{L^2}{2\tilde{m}r^2} + U(r). \tag{3.32}$$

Often this is called a radial equation, underlying that it indicates the energy balance only under the change of the distance r to the centre. Comparing this equation with (3.3), we see that it has the same form as the equation of energy for one-dimensional motion with "potential energy" $\tilde{U}(r)$:

$$\tilde{U}\left(r\right) = \frac{L^2}{2\tilde{m}r^2} + U(r). \tag{3.33}$$

This energy is called effective potential energy.

It is easy to explain the physical meaning of the additional term $L^2/2\tilde{m}r^2$ in $\tilde{U}\left(r\right)$. Taking the value of L from (3.28), it can be represented as $\tilde{m}r^2\dot{\varphi}/2$. This term is different from zero only for motion along φ, i.e. rotation. This is the reason to call the additional energy $L^2/2\tilde{m}r^2$ a centrifugal energy. Despite that in (3.32) the motion is represented as one-dimensional (radial), the motion along φ, i.e. the rotation, is present non-explicitly in the term $\propto L^2$. The angular momentum is a constant, but it exists as a consequence of the motion along φ, i.e., of the fact that $\dot{\varphi} \neq 0$. The effective potential energy can be used for qualitative analysis, just as $U(x)$ can be used in a one-dimensional motion. When

$$\tilde{U}(r)=\frac{L^2}{2\tilde{m}r^2}+U(r)\le E\ . \tag{3.34}$$

and $\tilde{U}(r')=\tilde{U}(r")=E$ the motion is finite and is bounded in the region $r'\le r\le r"$. Both solutions of the equation determine the minimal and maximal distance from the centre, and the orbit lies between them. Under the conditions (3.34) and $\tilde{U}(r')=E$, the motion is semi-infinite in the region $r\ge r'$ with the turning point r'. If $\tilde{U}(r)$ has a minimum and E is equal to $\tilde{U}_{\min}$, then $\dot{r}$ is always zero (see (3.32)) and r is fixed at the position of the minimum. In this case the particle moves along a circle around the centre.

3.4* ORBITS OF THE MOTION

In order to determine the character of motion we have to know the form of the function $\tilde{U}(r)$ in (3.33). The centrifugal energy is always positive. In the case of a repulsive force, the function of the potential $U(r)=-C/r$ $(\mathrm{C}<0)$ is positive for any r. Then the effective potential energy $\tilde{U}(r)$ decreases monotonically from $+\infty$ at $r=0$ to 0 at $r=\infty$ (Fig. 3-6). There is no minimum and the circular motion is not possible. For any (positive) value of E there is a minimal distance $r=r_{\min}$ (the distance of closest approach) which is the only solution of the equation $\tilde{U}(r)=E$, but there is no maximal distance. If the radial velocity is such that the particle approaches the centre, it will move with an orbit in which the value radius-vector r will decreased to $r_{\min}$ and after that it will increases to infinity. It is known (and later on will be proved) that such an orbit is a hyperbola (Fig. 3-6b). If the force is attractive $(C>0)$ the effective potential energy takes the form

$$\tilde{U}(r)=\frac{L^2}{2\tilde{m}r^2}+U(r)=\frac{L^2}{2\tilde{m}r^2}-\frac{C}{r}. \tag{3.35}$$

From the summation of the potential and centrifugal energy we obtain the curve $\tilde{U}(r)$, which possesses a minimum (Fig. 3-7). Equating the derivative $d\tilde{U}(r)/dr$ to zero, we determine r_0, at which $\tilde{U}(r)$ is minimal

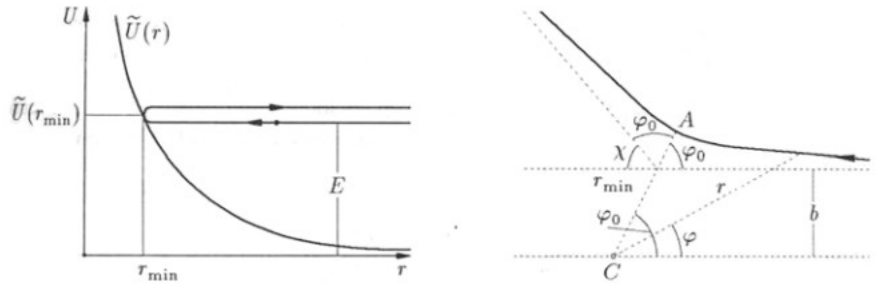

a b

Figure 3-6. Repulsive central conservative force:

a) the effective potential energy is monotonically decreasing function;

b) the orbit of the particle is hyperbola; *b* is the impact factor; the angle χ, determining the deflection from the initial particle direction, is a scattering angle, $r_{\min}$ is the minimal distance of the centre to the orbit, which is symmetrical with respect to the line defining $r_{\min}$.

$$\frac{d\tilde{U}(r)}{dr} = -\frac{L^2}{2\tilde{m}r^3} + \frac{C}{r^2} = 0, \quad r_0 = \frac{L^2}{C\tilde{m}}, \quad \tilde{U}_{\min} = -\frac{C^2\tilde{m}}{2L^2} = \frac{U(r_0)}{2}. \qquad (3.36)$$

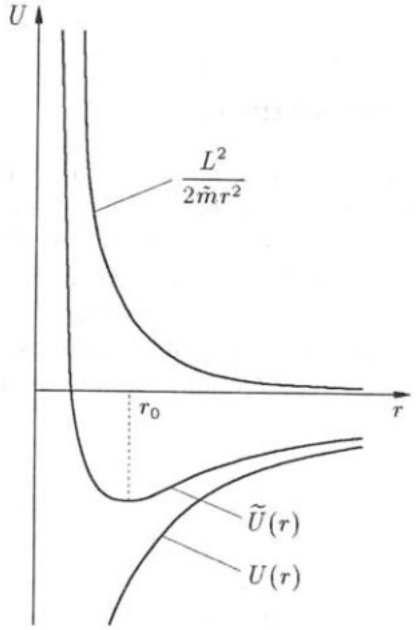

Figure 3-7. The effective potential energy $U(r)$ is a sum of the potential energy $U(r)$ and the centrifugal energy $L^2/2\tilde{m}r^2$, and it possesses a minimum $\tilde{U}_{\min} = U(r_0)/2$ at $r = r_0$.

The form of the curve $\tilde{U}(r)$ allows the determination of the character of the motion. We shall note that the curve is symmetrical with respect to the force centre *O* (Fig. 3-8a).

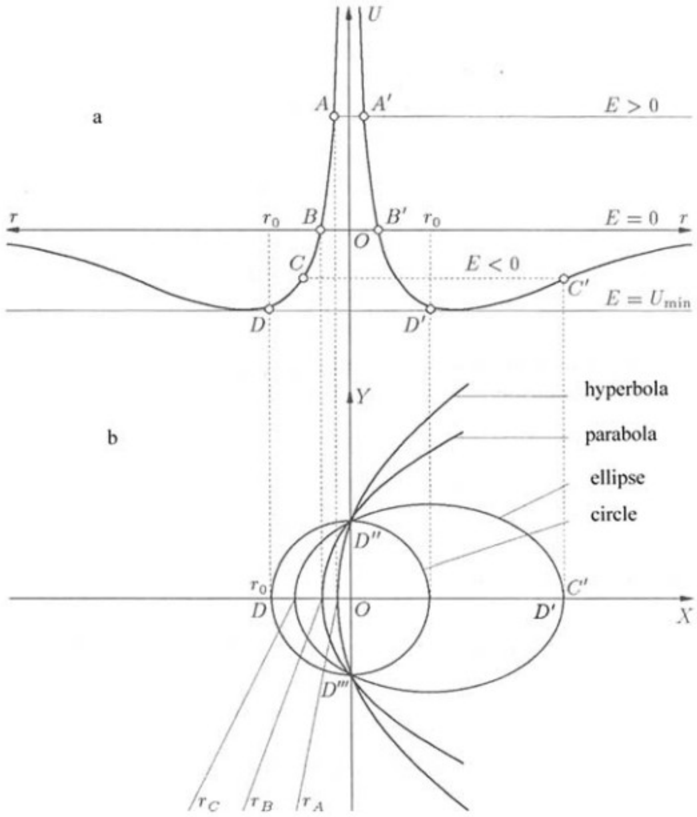
U
A
A'
E > 0
a
r
r₀
B
B'
r₀
E = 0
r
O
C
E < 0
C'
E = U_min
D
D'
Y
b
hyperbola
parabola
ellipse
circle
D''
r₀
D
O
D'
C'
X
D'''
r_C
r_B
r_A

Figure 3-8. a) View of the effective potential energy $\tilde{U}(r)$ for bodies with equal reduced masses and angular moment **L**, but different energies *E*. The corresponding orbits are: $E > 0$ - a hyperbola, E=0 - a parabola, E<0 - an ellipse, $E=\tilde{U}_{\min}$ - a circle. All orbits go through the points D'' and D'''.

At $E > 0$ the motion is unbounded and the minimal distance to the centre *O* is $r_{\min} = r_A$. The orbit is a hyperbola.

At $E = 0$ the motion is unbounded and the minimal distance to the centre *O* is $r_{\min} = r_B$. The orbit is a parabola.

At $E < 0$ the motion is finite and holds in the region $r_{\min} \le r \le r_{\max}$, where $r_{\min} = r_C$ and $r_{\max} = r_{C'}$. The orbit is an ellipse.

At $E=\tilde{U}_{\min}$ the motion is finite with a constant minimal distance $r=r_0$. The orbit is a circle.

And now, let us show that the orbits are indeed such. In order to find the equation of the orbit, we have to eliminate the time from the constants of motion (3.28) and (3.32), and obtain the equation which relates r and φ. From (3.32) at $U(r)=-C/r$ we readily find

$$\dot{r}\equiv\frac{dr}{dt}=\sqrt{\frac{2}{\tilde{m}}\left(E+\frac{C}{r}-\frac{L^2}{2\tilde{m}r^2}\right)}. \tag{3.37}$$

We rewrite (3.28) as $dt=d\varphi\tilde{m}r^2/L$. Substituting it into the last equation, we get

$$\varphi=\int\frac{L/r^2}{\sqrt{2\tilde{m}\left(E+\frac{C}{r}\right)-\frac{L^2}{r^2}}}dr+\text{const.} \tag{3.38}$$

This equation relates r and φ, i.e., it is the equation of the orbit. We rewrite it in the form

$$\varphi=\int\frac{L/r^2}{\sqrt{2\tilde{m}E+\left(\frac{C\tilde{m}}{L}\right)^2-\left(-\frac{C\tilde{m}}{L}+\frac{L}{r}\right)^2}}dr+\text{const.} \tag{3.39}$$

and use the substitutions

$$2\tilde{m}E+\left(\frac{C\tilde{m}}{L}\right)^2=a^2,\quad \left(-\frac{C\tilde{m}}{L}+\frac{L}{r}\right)=u,\quad \frac{-Ldr}{r^2}=du. \tag{3.40}$$

This yields

$$\varphi=-\int\frac{du}{\sqrt{a^2-u^2}}=\arccos\frac{u}{a}+\text{const.} \tag{3.41}$$

Returning to the initial variable r we represent the relation between r and φ in the following form

$$\varphi = \arccos\frac{-\frac{C\tilde{m}}{L}+\frac{L}{r}}{\sqrt{2\tilde{m}E+\left(\frac{C\tilde{m}}{L}\right)^2}} = \arccos\frac{-1+\frac{L^2}{C\tilde{m}}\frac{1}{r}}{\sqrt{1+\left(\frac{2El^2}{C^2\tilde{m}}\right)^2}} = \arccos\frac{-1+\frac{p}{r}}{e}, \quad (3.42)$$

where we substituted $L^2/C\tilde{m} = r_0 \equiv p$ (see (3.36)) and $\sqrt{1+2L^2E/C_2\tilde{m}} = e$. We set the initial angle $\varphi_0 = 0$ (it determines the orientation of the orbit). The final form of the equation of the orbit is

$$p/r = 1 + e\cos\varphi. \quad (3.43)$$

This is the conventional representation of a conic section with the focus in the centre of the field; the quantity p is called the parameter of the orbit and *e* is called eccentricity. The parameter determines the size of the conic section and the eccentricity - its form. Depending on the values of *e* we obtain the four types of orbits:

$e > 1$ - a hyperbola,
$e = 1$ - a parabola,
$0 < e < 1$ - an ellipse,
$e = 0$ - a circle.

Taking into account (3.36) we can write for the eccentricity

$$e = \sqrt{1-\frac{E}{\tilde{U}_{\min}}} = \sqrt{1+\frac{E}{\left|\tilde{U}_{\min}\right|}}, \quad (3.44)$$

from where it is evident that the conditions for *e*, which determine different types of orbits precisely coincide with the conditions for *E* in Fig. 3-8.

We shall note that the obtained orbits are the orbits of a body moving around another immobile body. In order to find the real orbits around the centre of mass (Fig. 3-9), we have to return to the equations (3.10).

The radius-vector value of the reduced mass changes as (see (3.43))

$$r = \frac{p}{1+e\cos\varphi}. \quad (3.45)$$

According to (3.10), $\mathbf{r}_1$ has the same direction and its magnitude is

$$r_1 = \frac{p\,m_2}{m_1+m_2}\frac{1}{1+e\cos\varphi} = \frac{p_1}{1+e\cos\varphi}\,. \tag{3.46}$$

The particle moves along an ellipse and the direction of its radius-vector $\mathbf{r}_1$ coincides with the direction of **r.**

In any moment the vector $\mathbf{r}_2$ is opposite to **r** (3.10). Therefore, when **r** is oriented at an angle φ, the vector $\mathbf{r}_2$ is oriented at an angle $\varphi_2 = \pi + \varphi$. Then from (3.10) and (3.45) we get

$$r_2 = \frac{m_2}{m_1+m_2}\frac{1}{1+e\cos(\pi+\varphi)} = \frac{p_2}{1-e\cos\varphi}\,. \tag{3.47}$$

The orbit is also an ellipse and m_2 moves so that both particles are on the line passing through the centre of mass at any moment (Fig. 3-9).

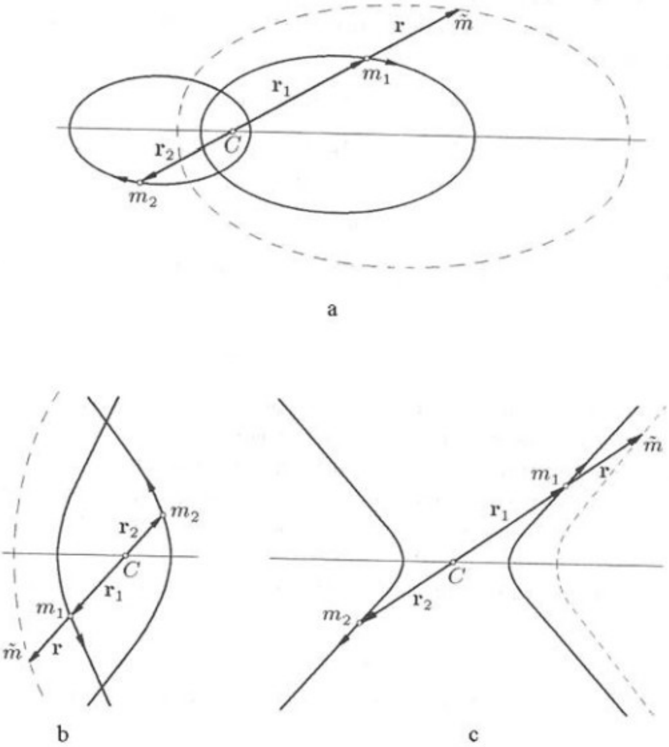

Figure 3-9. The orbits of particles with masses $\tilde{m}, m_1$ and m_2 $(m_1 : m_2 = 2:3)$ in CM-frame with an origin at the point *C*: a) ellipses (if $m_1 << m_2$, the ellipses are contained in one another!); b) hyperbolas (an attractive force); c) hyperbolas (a repulsive force). $\mathbf{r}_1$ and $\mathbf{r}_2$ are radius-vectors of the particles m_1 and m_2, **r** is a radius-vector of the reduced mass $\tilde{m}$.

2. * 51

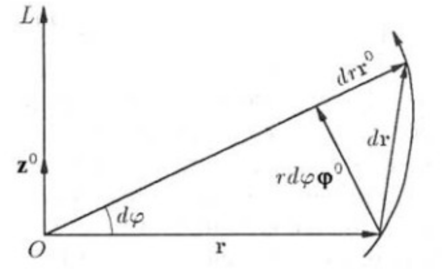
L
$dr\mathbf{r}^0$
$d\mathbf{r}$
$\mathbf{z}^0$
$rd\varphi\boldsymbol{\varphi}^0$
$d\varphi$
O
$\mathbf{r}$

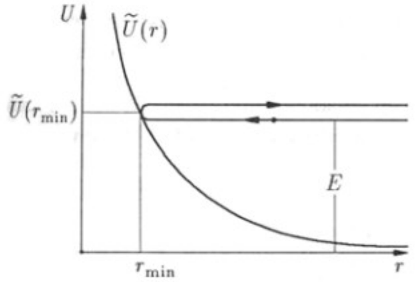
U
$\widetilde{U}(r)$
$\widetilde{U}(r_{\min})$
E
$r_{\min}$
r

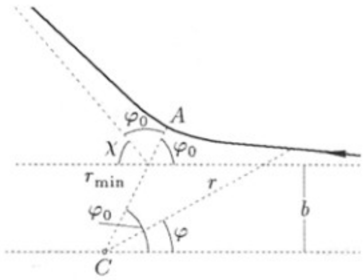
φ_0
A
χ
φ_0
$r_{\min}$
r
φ_0
b
φ
C

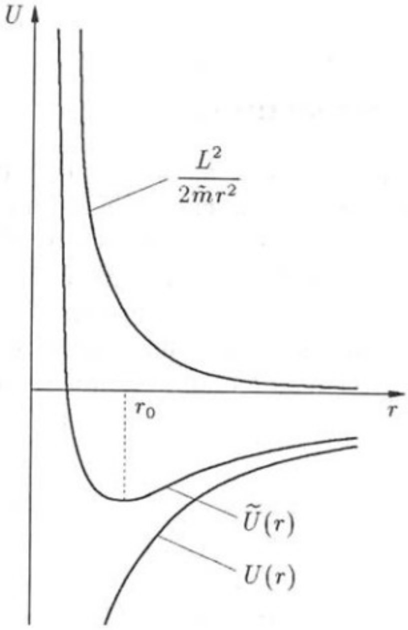
U
$\frac{L^2}{2\tilde{m}r^2}$
r_0
r
$\widetilde{U}(r)$
$U(r)$

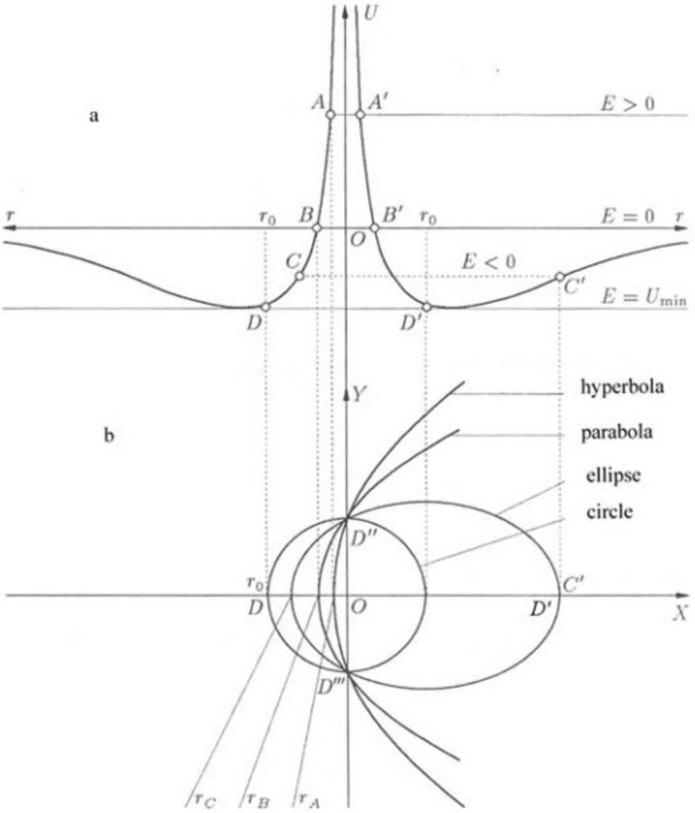
U
a
A
A'
E > 0
r
r_0
B
B'
r_0
E = 0
r
O
C
E < 0
C'
E = U_min
D
D'
b
Y
hyperbola
parabola
ellipse
circle
D''
r_0
C'
D
O
D'
X
D'''
r_C
r_B
r_A

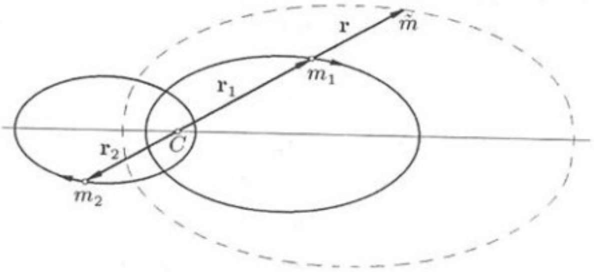
r
m̃
m_1
r_1
r_2
C
m_2

2.
N ' 4 33 | <

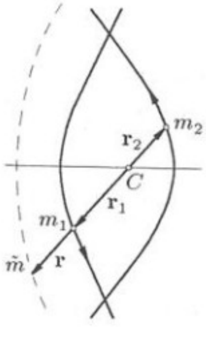
m_2
r_2
C
r_1
m_1
m̃
r

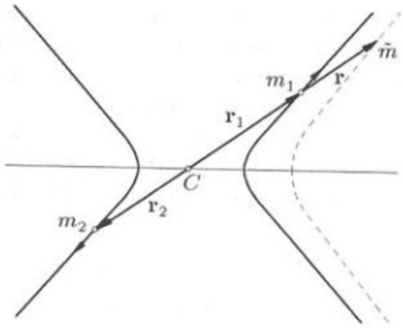
$\tilde{m}$
m_1
$\mathbf{r}$
$\mathbf{r}_1$
C
$\mathbf{r}_2$
m_2

. 9(:69 9 92 9 , E !
? : * ; ! ? < ; ,$ @ C ; 3!

AH 3

9(:69 9 92 9 , E ! ?
: * ; ! ? < ; ,$ @ C ; $" 3!

2. < : * < ' %0

' $/ " /

9(:69 9 92 9 , E ! ?
: * ; ! ? < ; ,$ @ C ; $/ / 3!

3 " < A E ! A)

9(:69 9 92 9 , R 4 @ ! (
; @ : 3!

' "

D. Squeezed states

Squeezed states are important for precision measurements because they have the potential to get precision better than that set by the uncertainty principle. They are defined as quantum states which have **imbalanced** uncertainties between two conjugate variables, such as P and X in a harmonic oscillator. They must, of course, satisfy the uncertainty principle $\Delta P\,\Delta X \geq \hbar/2$. But if the signal is encoded in the X variable, one gains in signal-to-noise ratio (SNR) by having a smaller ΔX (if one is limited by quantum noise), while the increased uncertainty in P does not affect the measurement.

A **balanced** distribution of uncertainties means that the quantities ΔP and ΔX remain constant in time. In a harmonic oscillator this occurs when

$$\Delta P = \omega \Delta X$$

If the oscillator is prepared with a different distribution, then ΔP and ΔX will **oscillate**. For example, if ΔP is small and ΔX is big, the potential will cause the portions of the wavepacket at large X to accelerate so that ΔP will increase dramatically during the next quarter period. If one thinks *classically* of an ensemble of oscillators prepared at the bottom of the potential with a statistical distribution having a narrow ΔX and large ΔP, it is clear that they will rapidly spread due to their high initial velocity distribution. A quarter period later they will all have stopped, so ΔP will be a minimum; however, ΔX will be a maximum. This out-of-phase oscillation of uncertainties for conjugate variables is a hallmark of squeezed states; **note that it occurs at** 2ω.

1. Hyperbolic transform of the HO Hamiltonian

We now present a mathematical treatment of squeezing in a harmonic oscillator. The basic idea is to make a hyperbolic rotation of the operators a and $a^\dagger$ to b and $b^\dagger$, a transformation which distorts the $\widehat{X}\widehat{Y}$ phase space while preserving the commutator. A new Hamiltonian with the same form as the old one is obtained, which has number states which are called **generalized number** states. Investigation of the generalized vacuum* will show that this new state has squeezed uncertainties in the old quadrature variables, $\widehat{X}$ and $\widehat{Y}$.

The transformation to the generalized raising and lowering operators is

$$b \equiv a \cosh r - a^\dagger e^{2i\theta} \sinh r$$
$$b^\dagger \equiv a^\dagger \cosh r - a e^{-2i\theta} \sinh r$$

---

*The ground state is often called the vacuum state even for non-electromagnetic harmonic oscillators.

where r is the distortion parameter, stretching by a factor e^r along an axis at angle θ to the $\widehat{X}$ axis, and shrinking by a factor of e^{-r} perpendicular to this axis, while preserving the area. This is called a Bogoliubov transformation and corresponds closely to a rotation with hyperbolic trigonometric functions. It can be verified that the transformation preserves the commutator, $[b, b^\dagger] = 1$. Defining the unitary squeezing operator $S(r, \theta)$ which does this transformation, i.e. $SaS^{-1} = b$ or $Sa^\dagger S^{-1} = b^\dagger$, one finds

$$S = \exp\left[\frac{r}{2}\left(e^{-2i\theta}a^2 - e^{2i\theta}a^{\dagger 2}\right)\right]$$

Hence the transformed Hamiltonian $H' = SHS^{-1}$ is

$$H' = \hbar\omega(b^\dagger b + 1/2)$$

which has the same form as the original Hamiltonian H. The corresponding quanta are not eigensolutions of H, but correspond to a superposition state that reforms itself after one period, and which has squeezed variances that oscillate twice each period.

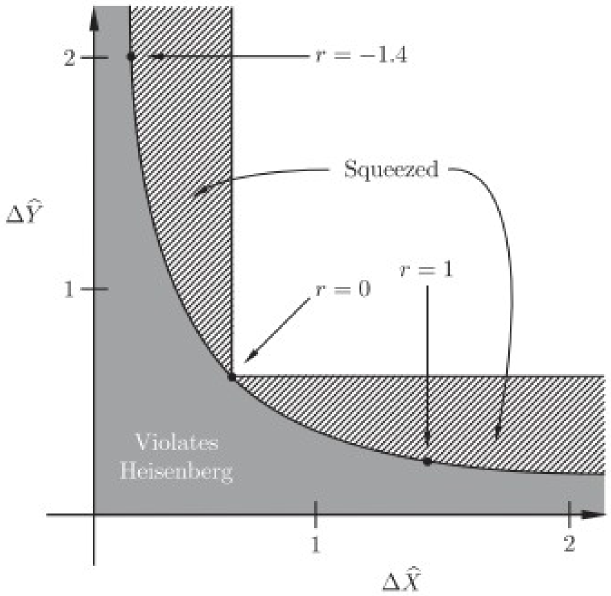

Figure 2.7: The solid line is the ideal squeezing achievable for various values of the squeeze parameter r (with $\theta = 0$). The shaded regions have significant but not ideal squeezing. The region with $\Delta\widehat{X} > 1/2$ and $\Delta\widehat{Y} > 1/2$ is unsqueezed.

In the following we specialize to $\theta = 0$ for clarity. The generalized quadrature operators are defined analogously to $\widehat{X}$ and $\widehat{Y}$

$$\widehat{X}_b = (b^\dagger + b)/2 = \widehat{X}e^{-r}$$
$$\widehat{Y}_b = i(b^\dagger + b)/2 = \widehat{Y}e^{r}$$

showing that the stretching occurs along the $\widehat{X}$ and $\widehat{Y}$ axes. The Hamiltonian then becomes

$$H' = \hbar\omega(\widehat{X}_b^2 + \widehat{Y}_b^2) = \hbar\omega(\widehat{X}^2 e^{-2r} + \widehat{Y}^2 e^{2r})$$

which may be interpreted in two ways.

(i) It may be written using the original position and momentum operators as

$$H' = \frac{P^2}{2me^{-2r}} + \frac{1}{2}me^{-2r}\omega^2 X^2$$

which is an oscillator with the same frequency but mass smaller by e^{2r}.

(ii) Alternatively, one can write

$$H' = \hbar\omega\left[e^{-2r}(\widehat{X}^2 + \widehat{Y}^2) + 2\widehat{Y}^2 \sinh 2r\right]$$

which is the HO Hamiltonian plus a nonlinear perturbation in $\widehat{Y}$.

2. Squeezed vacuum

Since both the commutator and the transformed Hamiltonian, when expressed in terms of b and $b^\dagger$, have exactly the same mathematical form as the HO Hamiltonian when a and $a^\dagger$ were used, the eigensolutions are the same. So we can immediately write down $\Delta\widehat{X}_b = \Delta\widehat{Y}_b = 1/2$. Hence

$$\langle 0_b|\Delta\widehat{X}|0_b\rangle = e^r/2$$
$$\langle 0_b|\Delta\widehat{Y}|0_b\rangle = e^{-r}/2$$

showing squeezing **but maintaining the uncertainty product**. Therefore the generalized vacuum state $|0_b\rangle$ represents a squeezed state with respect to the original oscillator.

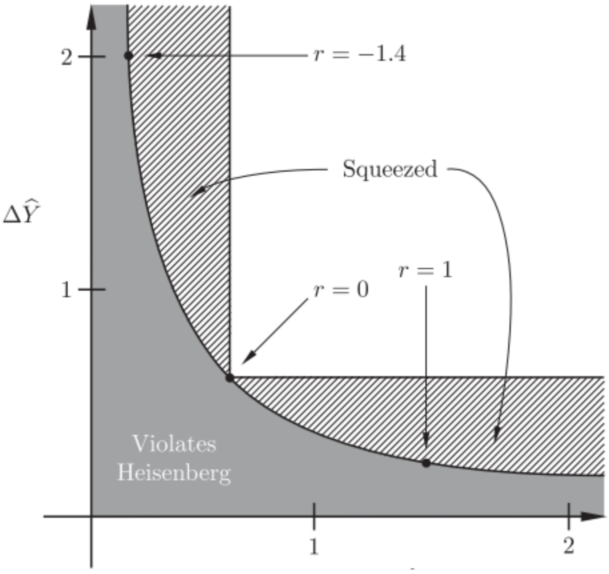

Figure 2.7: The solid line is the ideal squeezing achievable for various values of the squeeze parameter r (with $\theta = 0$). The shaded regions have significant but not ideal squeezing. The region with $\Delta\widehat{X} > 1/2$ and $\Delta\widehat{Y} > 1/2$ is unsqueezed.

9(:69 9 92 9 , A | ; * ;-
! 33 ; 0@ 3 3!

' %/ 3 0 *

' %, < : * <

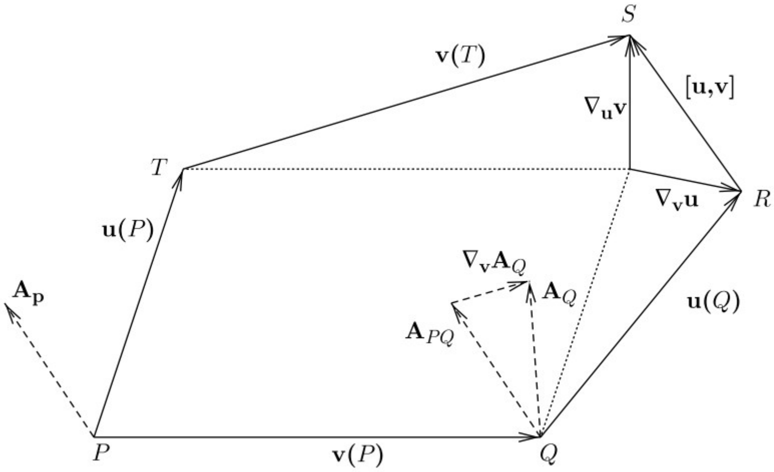
S
v(T)
[u,v]
$\nabla_{\mathbf{u}}\mathbf{v}$
T
$\nabla_{\mathbf{v}}\mathbf{u}$
R
u(P)
$\nabla_{\mathbf{v}}\mathbf{A}_Q$
$\mathbf{A_p}$
$\mathbf{A}_Q$
$\mathbf{A}_{PQ}$
u(Q)
P
v(P)
Q

Figure 7.3: The vector $\mathbf{A}_P$ parallel transported around the polygon $PQRSTP$.

' % 0 0 / AH ! & * 77777

' %/ 0$ 0 <) ! ! *

' %% $ 0 % 3 *

' 00 , A&R <A&RG E&AF> A6N(<&2R

' 0, , $ E R G * A G
D !)

9(:69 9 92 9 , (! 4 . > < ? < ;
: * ; (AH3 @ 3 3!

' ""

3
< ! : 3 < ! /

AH

9(:69 9 92 9 , (! 4 . > < ? < ;
: * ; (AH3 @ 3 ' $ 3!

' "0 % 3 $
< 3 ! A H L ? ! A)

AH

9(:69 9 92 9 , (! 4 . > < ? < ;
: * ; (AH3 @ 3 ' "0 %0 3!

' % 0 %/ (33 ! H /
; E 3 < ! w E E x

9(:69 9 92 9 , (! 4 . > < ? < ;
: * ; (AH3 @ 3 ' % 0 % / 3!

9(:69 9 92 9 , ! ! < 3
< ; : * ;- E ! @ < ;@ @ (33 %@ 3 3!

' $ / , H G A) ! C * A)

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0}, \qquad \nabla \cdot \mathbf{B} = 0, \qquad \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \qquad \nabla \times \mathbf{B} = \varepsilon_0 \mu_0 \frac{\partial \mathbf{E}}{\partial t} + \mu_0 \mathbf{J},$$

similar to those of the experiment of Fizeau, that a shift will be observed in the fringe pattern by $2\Delta k$ fringes, where

$$\Delta k = \frac{2l}{\lambda}\frac{V}{c}(n^2 - 1). \tag{1.28}$$

Such a displacement would be easily observable given the speed of the Earth on its orbit, $V = 30$ km/s. For $\lambda = 400\,\text{nm}$, $l = 1$ m and $n = 1.333$, it would have been equal to $2\Delta k = 777$ fringes. No displacement of the fringe pattern was observed, however, during the rotation of the apparatus.

If we consider that the aether is partially dragged by the water in the pipe AB, the fact that a shift in the fringes does not occur on rotating the interferometer implies that the drag coefficient of the aether for the particular medium, f, is equal to

$$f = 1 - \frac{1}{n^2}. \tag{1.29}$$

This is exactly the relation suggested by Fresnel in the interpretation of Arago's observations. Hoek's experiment verifies Fresnel's formula, at least to an accuracy of first order in the ratio V/c.

The experiments just described are interpreted by the hypothesis that the aether is dragged by optical media. However, the dragging has the strange property of depending on the refractive index and therefore on the wavelength (color) of the light, something that is very difficult to justify. The question arises of whether it is possible to detect the motion of the Earth through the aether. Michelson, and later Michelson and Morley, tried to answer this question with their very important experiments. These experiments we will describe below.

1.9 Maxwell's Equations and the Wave Equation

The laws of electromagnetism are summarized in four equations, known as Maxwell's equations, which, in vacuum, take the form

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0}, \qquad \nabla \cdot \mathbf{B} = 0, \qquad \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \qquad \nabla \times \mathbf{B} = \varepsilon_0 \mu_0 \frac{\partial \mathbf{E}}{\partial t} + \mu_0 \mathbf{J}, \tag{1.30}$$

where **E** and **B** are the electric and the magnetic fields, respectively, ρ is the volume charge density, **J** the current density, and ε_0 and μ_0 are the electric and the magnetic constants of vacuum, respectively. For more information on the notation, the reader is referred to Appendix 4 or to a book on vector analysis. The physical meaning of these equations is as follows:

$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0}$	This equation is derived from Coulomb's law and is a formulation of Gauss' law in differential form. The expression on the left is the flux of the electric field per unit of infinitesimal volume, which is shown to be proportional to the charge enclosed by this volume
$\nabla \cdot \mathbf{B} = 0$	This is Gauss' law for the magnetic field. Its flux per unit of infinitesimal volume is equated to zero due to the non-existence of magnetic monopoles, which would correspond to the positive and negative electric charges
$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$	This is the differential form of Faraday's law of electromagnetic induction or the fact that a magnetic field varying with time produces an electric field
$\nabla \times \mathbf{B} = \varepsilon_0 \mu_0 \frac{\partial \mathbf{E}}{\partial t} + \mu_0 \mathbf{J}$	The complete form of Ampère's law, according to which an electric field varying with time or an electric current, produce a magnetic field. The first term on the right is due to the displacement current, suggested by Maxwell

In vacuum and without free charges, $\rho = 0$, or currents, $\mathbf{J} = 0$, the equations simplify to

$$\nabla \cdot \mathbf{E} = 0 \qquad \nabla \cdot \mathbf{B} = 0 \qquad \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \qquad \nabla \times \mathbf{B} = \varepsilon_0 \mu_0 \frac{\partial \mathbf{E}}{\partial t} \tag{1.31}$$

which may be combined to give an equation for the electric field **E**

$$\frac{\partial^2 \mathbf{E}}{\partial x^2} + \frac{\partial^2 \mathbf{E}}{\partial y^2} + \frac{\partial^2 \mathbf{E}}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2 \mathbf{E}}{\partial t^2} \qquad \text{or} \qquad \nabla^2 \mathbf{E} = \frac{1}{c^2}\frac{\partial^2 \mathbf{E}}{\partial t^2}, \tag{1.32}$$

and an equation for the magnetic field **B**

$$\frac{\partial^2 \mathbf{B}}{\partial x^2} + \frac{\partial^2 \mathbf{B}}{\partial y^2} + \frac{\partial^2 \mathbf{B}}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2 \mathbf{B}}{\partial t^2} \qquad \text{or} \qquad \nabla^2 \mathbf{B} = \frac{1}{c^2}\frac{\partial^2 \mathbf{B}}{\partial t^2}. \tag{1.33}$$

These equations describe an electromagnetic wave moving with a speed of

$$c = \frac{1}{\sqrt{\varepsilon_0 \mu_0}} = 3 \times 10^8 \text{ m/s}, \tag{1.34}$$

which, it should be noted, does not depend on the speeds of the source of the wave or of the observer.

The flow of electromagnetic energy per unit time and per unit area, at a point, is given by the Poynting vector:

$$\mathbf{S} = \frac{1}{\mu_0}\mathbf{E} \times \mathbf{B}. \tag{1.35}$$

Necessary for the propagation of electromagnetic waves was considered to be the existence of the aether, although its properties were impossible to explain using mechanical models.

If, however, the ignorance concerning the structure of aether could be tolerated, there was another more serious problem with the laws of electromagnetism, which could not be ignored. Up till then, it was taken for granted that the laws of Physics satisfied the *Galilean principle of (mechanical) relativity*, i.e. they had the same form in all inertial frames of reference. As we have already demonstrated in the case of Newton's second law of motion, the transition from one inertial frame of reference to another, based on the Galilean transformation, leaves the law invariant. This is not true, however, for the laws of electromagnetism. Maxwell's equations and, thus, the wave equation also, are not invariant under the Galilean transformation. We will return to this subject in Chap. 9. For the moment, it is enough to mention that the transformation of the wave equation, Eq. (1.32), from one inertial frame of reference $S(x, y, z, t)$ to another, $S'(x', y', z', t')$, which is moving with a constant velocity $\mathbf{V}=V\hat{\mathbf{x}}$ relative to S, based on the Galilean transformation given by Eqs. (1.1), transforms the wave equation for the electric field to

$$\nabla'^2\mathbf{E} = \frac{1}{c^2}\frac{\partial^2\mathbf{E}}{\partial t'^2} + \frac{1}{c^2}\left(V^2\frac{\partial^2\mathbf{E}}{\partial x'^2} - 2V\frac{\partial^2\mathbf{E}}{\partial x'\partial t'}\right). \tag{1.36}$$

Obviously, it is not possible for the wave equation to have the form of Eq. (1.32) in one inertial frame of reference and the form of Eq. (1.36) in another, unless we accept that the frame of reference of the Earth is privileged, as it is from experiments in this frame of reference that the laws leading to the 'simple' form of Eq. (1.32) for the propagation of electromagnetic waves were found. Lorentz played the leading role in the adoption of the view that there is a stationary aether, which defines a privileged frame of reference in which the laws of electromagnetism assume their 'simpler' form. This would define Newton's absolute space, in which no fictitious forces appear. The question immediately arises of whether it is possible to detect experimentally the motion of a frame of reference, such as the Earth, relative to the aether.

1.10 The Experiment of Michelson and Morley

The most serious investigation of whether it is possible to detect the motion of the Earth relative to the aether, was performed by a series experiments by Michelson in 1881 [15] and, in an improved form, by Michelson and Morley in 1887 [16].

' "%, /< : * ; < !F

The reader must have noticed that, in the figures, the length of the train was taken as being different for the two observers: L_0 for the observer on the train and L for the observer on the platform. This is not accidental, as we will see in what follows. However, our ignorance, at present, of the exact lengths, does not influence the arguments developed above.

2.4 The Relativity of Time and Length

The invariance of the speed of light has very important consequences on time and length, as we will see in this section. We will begin by examining a quantity which remains invariant—any transverse dimension of a moving body. We will then examine the changes that have to be made to our understanding of time and length. We will return to these subjects in the next chapter, for a more general examination which will be based on the Lorentz transformation.

2.4.1 *The Invariance of the Dimensions of a Body Which Are Perpendicular to Its Velocity*

We examine two hollow cylinders, A and B (Fig. 2.6), which, when at rest, are identical, with radius r_0 and length L_0. Let the two cylinders be in relative motion with relative speed V, along their common axis. We will assume that to an observer a moving cylinder shrinks in directions perpendicular to its velocity, so that the radius of the moving cylinder is $r(V) < r_0$. The arguments that will follow would still hold had we assumed that the moving cylinder expands in directions perpendicular to its velocity. It will not affect our arguments, but we may assume that the

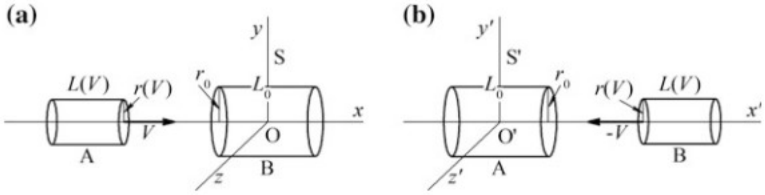

Fig. 2.6 Two identical *hollow cylinders*, A and B, in relative motion to each other, with their axes being coincident. The two cylinders, when at rest, have a radius equal to r_0. We will assume that when a cylinder is moving with speed V along its axis, its radius changes to $r(V)$ and its length to $L(V)$. Without loss of generality, we will take $r(V) < r_0$. In the reference frame of B, S, cylinder A moves with a velocity $\mathbf{V} = V\hat{\mathbf{x}}$ and its radius shrinks to $r(V)$. Cylinder A will, therefore, pass through cylinder B. In the reference frame of A, S′, cylinder B moves with a velocity $\mathbf{V} = -V\hat{\mathbf{x}}$ and its radius shrinks to $r(V)$. In the frame of reference S′, cylinder B will be expected to pass through cylinder A

length of the moving cylinder changes to $L(V)$. In the frame of reference S of cylinder B, cylinder A moves with a velocity $\mathbf{V} = V\hat{\mathbf{x}}$ and its radius is $r(V) < r_0$. Cylinder A will, therefore, pass through cylinder B (Fig. 2.6a).

In the frame of reference S′ of cylinder A, cylinder B moves with a velocity $\mathbf{V} = -V\hat{\mathbf{x}}$. Since space is isotropic, the radius of the moving cylinder will depend on its speed and so the radius of cylinder B will be $r(V) < r_0$. In this frame of reference, cylinder B is expected to pass through cylinder A (Fig. 2.6b).

Since it is impossible for both these predictions to be true, we must conclude that the transverse dimensions of a moving body do not change. This conclusion is purely a consequence of the non-existence of a privileged inertial frame of reference and the equivalence of all the inertial frames of reference.

2.4.2 The Relativity of Time

We will examine the rates of counting time by two clocks in relative motion to each other with a constant speed. The clocks we will consider work in the following way: Two mirrors, A and B in Fig. 2.7a, are at a distance D from each other. A very short light pulse (or a photon!) travels between the two mirrors. An observer, who is at rest relative to such a clock, will consider one unit of time, as measured by the clock, to be the time $T_0 = 2D/c$ needed for the pulse to travel from A to B and back to A, Fig. 2.7a.

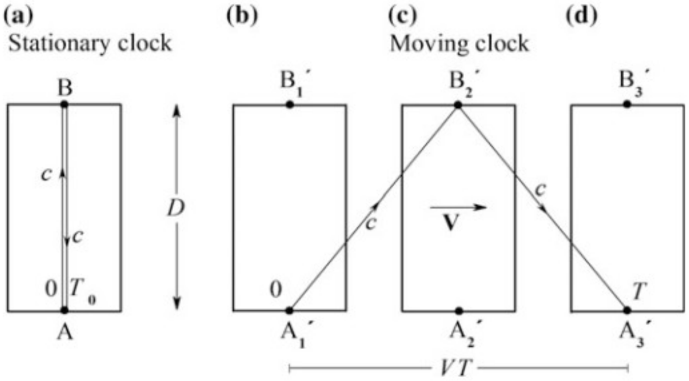

Fig. 2.7 *The dilation of time*. Two clocks, one at rest relative to an observer and one moving with a speed V relative to him. The operation of the clocks is based on a light pulse traveling between two mirrors, A and B (**a**). A unit of time (between two successive beats) is measured by the clock when the light pulse travels from A to B and then back to A. The moving clock moves in a direction perpendicular to the line joining the two mirrors of the clock at rest. The observer sees that the light in the moving clock has to cover a greater distance between successive beats than the light in the clock at rest relative to him (**b–d**). He therefore observes that more time passes between the successive beats of the moving clock as compared to the corresponding time in the clock at rest relative to him

Let now the observer who is at rest relative to a clock, observe another identical clock moving with speed V relative to himself in a direction perpendicular to the line AB, as shown in Figs. 2.7b, c, d. If for this observer the light needs a time T to travel from A to B and then back to A, the observer will see the moving clock shift by a distance $\mathrm{A}_1'\mathrm{A}_3' = VT$. The distance this observer will see as travelled by the light, always with a speed c, will be

$$cT = \mathrm{A}_1'\mathrm{B}_2'\mathrm{A}_3' = 2\sqrt{D^2 + (\mathrm{A}_1'\mathrm{A}_2')^2} = 2\sqrt{D^2 + (VT/2)^2}, \tag{2.1}$$

where we have taken into account the fact we have already proved, that the distance D does not change due to the motion of the clock. But $D = cT_0/2$ and, hence, substituting, we have

$$c^2T^2 = c^2T_0^2 + V^2T^2, \tag{2.2}$$

from which we find that

$$T = \frac{T_0}{\sqrt{1 - V^2/c^2}}, \tag{2.3}$$

which is a time greater than T_0.

It follows that the observer sees the moving clock count time at a slower rate than that of the clock at rest relative to himself. A clock at rest registers that the time needed for light to travel the path ABA, from one of its mirrors to the other and back (*proper time*), is smaller by a factor of $\sqrt{1 - V^2/c^2}$ compared to the time that the clock at rest registers as needed for the light to travel from one mirror of the moving clock to the other and back. This phenomenon is known as *dilation of time*. We will return to it in the next chapter and in the one after, where we will also examine its experimental verification.

As a numerical example, consider two identical clocks, one at rest in a frame of reference S and one at rest in a frame of reference S′. S′ moves relative to S with a speed $V = (4/5)c$, for which it is $\sqrt{1 - V^2/c^2} = 3/5$. Let the clock in S′ count a time interval $T_0 = 1$ unit of time, as its light pulse moves from mirror A to mirror B and back. To an observer relative to which this clock is moving, the time taken for this process to be completed in the moving clock will be $T = T_0\Big/\sqrt{1 - V^2/c^2} = 1/(3/5) = 5/3 = 1.67$ units of time. The observer in S sees that the moving clock needs 1.67 units of time to complete a procedure that corresponds to the counting of 1 unit of time. He, therefore, concludes that the moving clock is slow in counting time.

One may claim that the dilation of time was proved to apply only to a clock in which the light travels in a direction normal to the clock's velocity **V**. We will now show that the same result holds for a clock at any orientation relative to **V**. In Fig. 2.8 we have drawn two identical clocks, at rest in the same frame of reference,

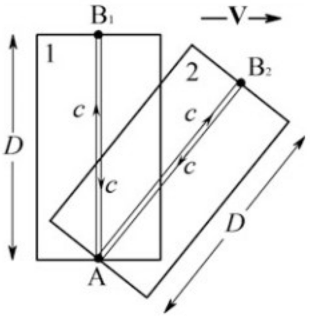

Fig. 2.8 Two identical clocks, similar to those of Fig. 2.7. The clocks are at rest in the same frame of reference and are oriented one (*1*) normal to the velocity **V** of an outside observer and the other (*2*) at some other angle. The two clocks emit their pulses of light at a common point, *A*

one (1) with its axis normal to **V** and the other (2) inclined at another angle. The two clocks have a common point of emission of their photons, A.

In the frame of reference of the clocks, if two pulses start simultaneously in the two clocks from point A, they will return to the same point simultaneously, because they have the same distance to travel, $2D$, and the speed of light is the same in all directions. In the frame of reference of the clocks, point A and the two pulses coincide initially and they will also coincide again when the two pulses return. It is obvious that the two pulses will also coincide at point A when viewed from any other frame of reference. If this did not happen, we might have a phenomenon, such as an interaction, in one frame of reference and not in the other. This would naturally be physically unacceptable. The two pulses will therefore coincide at point A in all other frames of reference as well. The rate at which the two clocks measure time will be the same and what is true for clock 1 will also be true for clock 2. The dilation of time is, therefore, the same, independent of the clock's orientation relative to the velocity **V**.

2.4.3 The Relativity of Length

We will now consider two clocks, similar to those of the previous section, one of which is at rest relative to some observer, and another that is moving with speed V, in a direction which is now parallel to the straight line AB between the two mirrors, as shown in Fig. 2.9. We must leave open the possibility that the length of the moving clock in the direction of its motion may differ from that of a stationary clock. If the distance measured by the observer between the two mirrors is L_0 for the clock at rest, let it be equal to L for the clock that moves relative to the observer (Fig. 2.9b, c, d).

In the clock which is at rest relative to the observer, light travels the distance ABA in time $T_0 = 2L_0/c$. We have already proved that the observer will measure

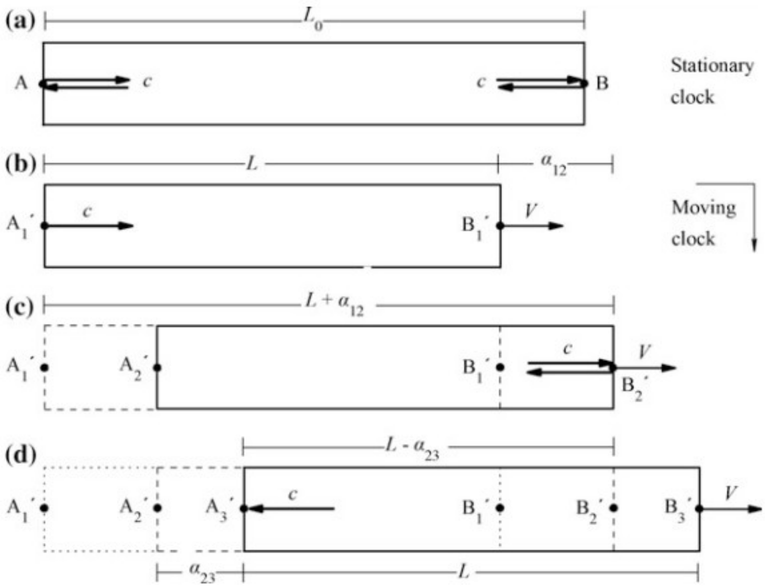
(a)
L_0
A
c
c
B
Stationary clock
(b)
L
α_{12}
A_1'
c
B_1'
V
Moving clock
(c)
$L + \alpha_{12}$
A_1'
A_2'
B_1'
c
V
B_2'
(d)
$L - \alpha_{23}$
A_1'
A_2'
A_3'
c
B_1'
B_2'
B_3'
V
α_{23}
L

Fig. 2.9 *The contraction of length*. Two clocks, one at rest relative to an observer and one moving with a speed *V* relative to him. The operation of the clocks is based on a light pulse traveling between two mirrors, A and B. The moving clock moves in a direction parallel to the line AB. The distances travelled by the light in the one direction and in the other, as measured by the observer, are different in the moving clock

that for the moving clock the time needed for the light to travel from one mirror to the other and back is

$$T = \frac{T_0}{\sqrt{1 - V^2/c^2}}. \tag{2.4}$$

Alternatively, the observer sees light travel in the moving clock from mirror A to mirror B in time ΔT_+, covering a distance $\mathrm{A}_1'\mathrm{B}_2'$ (Fig. 2.9b, c) which, due to the motion of the clock with speed *V*, will be equal to

$$\mathrm{A}_1'\mathrm{B}_2' = c\Delta T_+ = L + \alpha_{12} = L + V\Delta T_+, \tag{2.5}$$

from which we find,

$$\Delta T_+ = \frac{L}{c - V}. \tag{2.6}$$

In the motion from mirror B to mirror A, light travels in time ΔT_- the distance $B_2'A_3'$ (Fig. 2.9c, d) which is equal to

$$B_2'A_3' = c\Delta T_- = L - \alpha_{23} = L - V\Delta T_-, \tag{2.7}$$

from which we find,

$$\Delta T_- = \frac{L}{c+V}. \tag{2.8}$$

The total time needed for light in the moving clock to travel from mirror A to mirror B and back is, according to the external observer,

$$T = \Delta T_+ + \Delta T_- = \frac{L}{c-V} + \frac{L}{c+V} = \frac{2cL}{c^2 - V^2} = \frac{2L/c}{1 - V^2/c^2}. \tag{2.9}$$

It follows from Eqs. (2.4) and (2.9) that

$$T = \frac{T_0}{\sqrt{1 - V^2/c^2}} = \frac{2L/c}{1 - V^2/c^2}. \tag{2.10}$$

However, it is $T_0 = 2L_0/c$, so that

$$\frac{2L_0/c}{\sqrt{1 - V^2/c^2}} = \frac{2L/c}{1 - V^2/c^2} \tag{2.11}$$

and, finally,

$$L = L_0\sqrt{1 - V^2/c^2}. \tag{2.12}$$

This equation describes the phenomenon of the *contraction of length*, according to which an observer finds that the length of a body that is moving with a speed V relative to him is, in the direction of motion of the body, smaller by a factor $\sqrt{1 - V^2/c^2}$ compared to the length of the same body in its own frame of reference, in which it is stationary (*rest length* or *proper length*). We will return to this effect in the next chapter.

2.5 The Inevitability of the Special Theory of Relativity

It is natural that someone learning about the Special Theory of Relativity for the first time will react to the proposition that he or she should accept the consequences of the theory. The conclusions of the theory are in direct contrast to our everyday experiences, given the limited means we have for acquiring them with our senses.

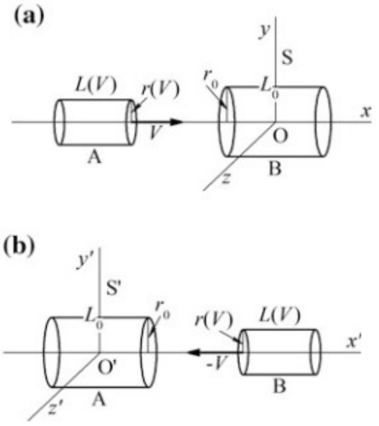
(a)
y
S
L(V)
r(V)
r0
L0
x
V
O
A
z
B
(b)
y′
S′
L0
r0
r(V)
L(V)
x′
-V
O′
z′
A
B

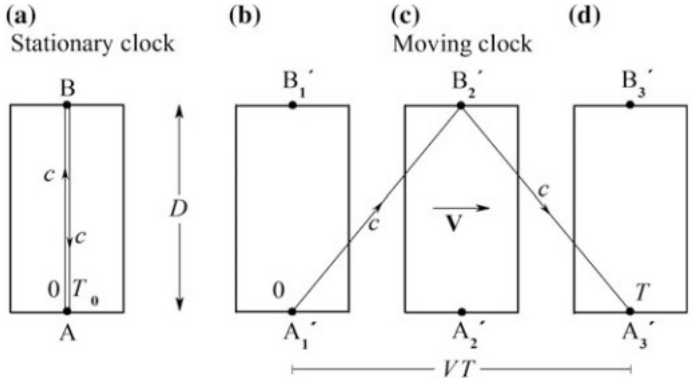
(a)
Stationary clock
(b)
(c)
Moving clock
(d)
B
c
c
0
T0
A
D
B1′
0
A1′
B2′
c
V
c
A2′
B3′
T
A3′
VT

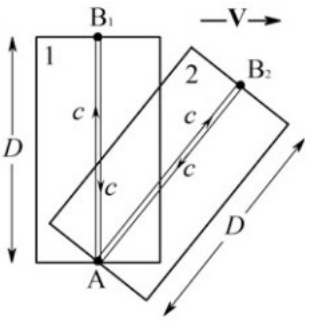
B1
—V→
1
2
B2
c
c
D
c
c
A
D

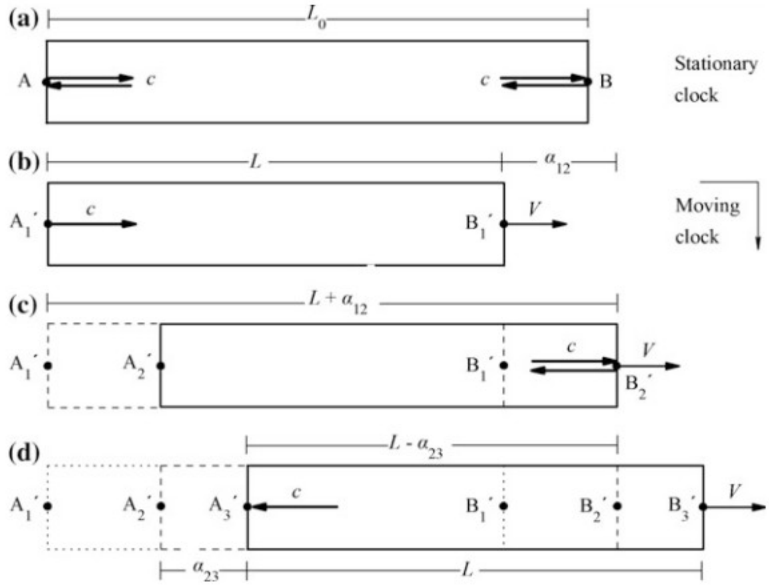
(a)
L_0
A
c
c
B
Stationary clock
(b)
L
α_{12}
A_1'
c
B_1'
V
Moving clock
(c)
$L + \alpha_{12}$
A_1'
A_2'
B_1'
c
V
B_2'
(d)
$L - \alpha_{23}$
A_1'
A_2'
A_3'
c
B_1'
B_2'
B_3'
V
α_{23}
L

' % 0% 3 "
: * L

AH

9(:69 9 92 9 , ! ! < 3
< ; : * ;- E ! @ < ;@ @ (33 %@ 3 ' %
, 3!

9(:69 9 92 9 , 3 : * ; ! 6
< ;- (3 v ? 3 ,$$@ 3 R ! 3!

' "

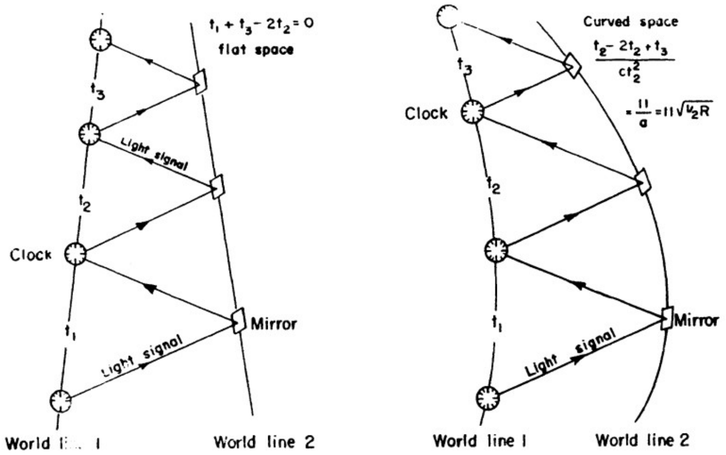
$t_1 + t_3 - 2t_2 = 0$
flat space
t_3
Light signal
t_2
Clock
t_1
Light signal
Mirror
World line 1
World line 2
Curved space
$\frac{t_2 - 2t_2 + t_3}{ct_2^2}$
$= \frac{11}{a} = 11\sqrt{1/2R}$
t_3
Clock
t_2
t_1
Light signal
Mirror
World line 1
World line 2

&

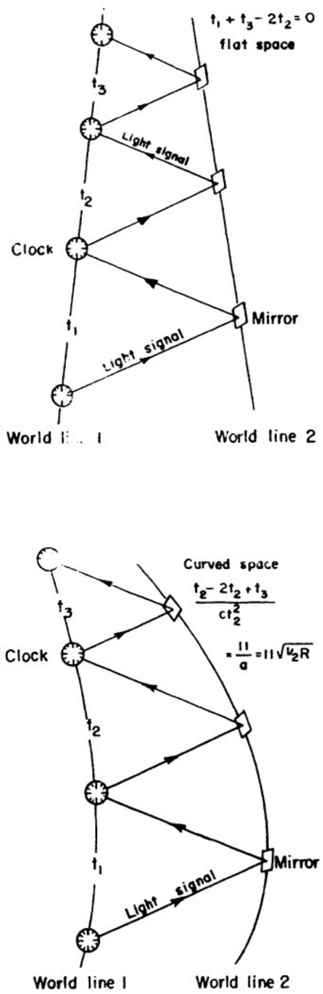
t₁ + t₃ − 2t₂ = 0
flat space
t₃
Light signal
t₂
Clock
Mirror
t₁
Light signal
World line 1
World line 2
Curved space
t₂ − 2t₂ + t₃ / ct₂²
= 11/a = 11√½R
t₃
Clock
t₂
t₁
Mirror
Light signal
World line I
World line 2

FIG. 5. Direct measurement of the curvature by means of a clock and mirror. Only one space-like dimension is considered and the curvature assumed to be constant within the space-time region contained in the figure. The explanation is given in Appendix V.

E&? >
*
.;
! B ! 2 ;
3 B !
! ! !
* !
.
3

! <
H3 *
(33 ! H

' % , : ; ;

nonzero mass. For these, the $2S+1$ directions of polarization follow from the invariance of the theory with respect to proper Lorentz transformations. In particular, if the particle is at rest, the spin will have different orientations with respect to coordinate systems which have different orientations in space. Thus, the existence of all the states of polarization follow from the existence of one, if only the theory is invariant with respect to proper Lorentz transformations. For particles with zero rest-mass, there are only two states of polarization, and even the existence of the second one can be inferred only on the basis of reflection symmetry.

REFLECTION SYMMETRY

The problem and existence of reflection symmetry have been furthered in a brilliant way by recent theoretical and experimental research. There is nothing essential that can be added at present to the remarks and conjectures of Lee, Yang, and Oehme, and all that follows has been said, or at least implied, by Salam, Lee, Yang, and Oehme.[5] The sharpness of the break with past concepts is perhaps best illustrated by the cobalt experiment of Wu, Ambler, Hayward, Hoppes, and Hudson.

The ring current—this may be a permanent current in a superconductor—creates a magnetic field. The Co source is in the plane of the current and emits β particles (Fig. 3). The whole experimental arrangement, as shown in Fig. 3, has a symmetry plane and, if the principle of sufficient cause is valid, the symmetry plane should remain valid throughout the further fate of the system. In other words, since the right and left sides of the plane had originally identical properties, there is no sufficient reason for any difference in their properties at a later time. Nevertheless, the intensity of the β radiation is larger on one side of the plane than the other side. The situation is paradoxical no matter what the mechanism of the effect is—in fact, it is most paradoxical if one disregards its mechanism and theory entirely. If the experimental circumstances can be idealized as indicated, even the principle of sufficient cause seems to be violated.

It is natural to look for an interpretation of the experiment which avoids this very far-reaching conclusion and, indeed, there is such an interpretation.[6a] It is good to reiterate, however, that no matter what interpretation is adopted, we have to admit that the symmetry of the real world is smaller than we had thought. However, the symmetry may still include reflections.

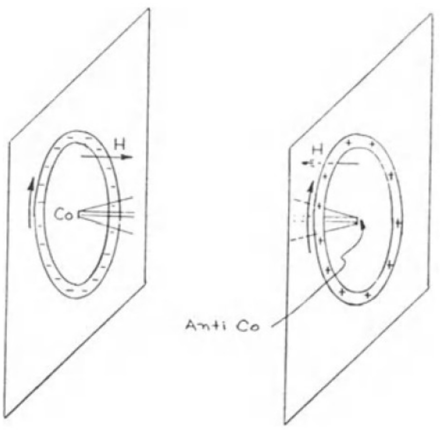

FIG. 3. The right side is the mirror image of the left side, according to the interpretation of the parity experiments[6a] which maintains the reflection as a symmetry element of all physical laws. It must be assumed that the reflection transforms matter into antimatter: the electronic ring current becomes a positronic ring current, the radioactive cobalt is replaced by radioactive anticobalt.

If it is true that a symmetry plane always remains a symmetry plane, the initial state of the Co experiment could not have contained a symmetry plane. This would not be the case if the magnetic vector were polar—in which case the electric vector would be axial. The charge density, the divergence of the electric vector, would then become a pseudoscalar rather than a simple scalar as in current theory. The mirror image of a negative charge would be positive, the mirror image of an electron a positron, and conversely. The mirror image of matter would be antimatter. The Co experiment, viewed through a mirror, would not present a picture contrary to established fact: it would present an experiment carried out with antimatter. The right side of Fig. 3 shows the mirror image of the left side. Thus, the principle of sufficient cause, and the validity of symmetry planes, need not be abandoned if one is willing to admit that the mirror image of matter is antimatter.

The possibility just envisaged would be technically described as the elimination of the operations of reflection and charge conjugation, as presently defined, as true symmetry operations. Their product would still be assumed to be a symmetry operation and proposed to be named, simply, reflection. A few further technical remarks are contained in Appendix II. The proposition just made has two aspects: a very appealing one, and a very alarming one.

Let us look first at the appealing aspect. Dirac has said that the number of elementary particles shows an alarming tendency of increasing. One is tempted to add to this that the number of invariance properties also showed a similar tendency. It is not equally alarming because, while the increase in the number of elementary particles complicates our picture of nature, that of the symmetry properties on the whole simplifies it. Nevertheless the clear correspondence between the invariance properties of the laws of nature, and the symmetry properties of space-time, was most clearly breached by the operation of charge conjugation. This postulated that the laws of nature remain the same if all positive charges are replaced by negative charges and vice versa, or more generally, if all particles are replaced by antiparticles. Reasonable as this postulate appears to us, it corresponds to no symmetry of the space-time continuum. If the preceding interpretation of the Co experiments should be sustained, the correspondence between the natural symmetry elements of space-time, and the invariance properties of the laws of nature, would be restored. It is true that the role of the planes of reflection would not be that to which we are accustomed—the mirror image of an electron would become a positron—but the mirror image of a sequence of events would still be a possible sequence of events. This possible sequence of events would be more difficult to realize in the actual physical world than what we had thought, but it would still be possible.

---

[5] Lee, Yang, and Oehme, Phys. Rev. **106**, 340 (1957).

[6a] The interpretation referred to has been proposed independently by numerous authors, including A. Salam, Nuovo cimento **5**, 229 (1957); L. Landau, Nuclear Phys. **3**, 127 (1957); H. D. Smyth and L. Biedenharn (personal communication). Dr. S. Deser has pointed out that the "perturbing possibility" was raised already by Wick, Wightman, and Wigner [Phys. Rev. **88**, 101 (1952)] but was held "remote at that time." Naturally, the apparent unanimity of opinion does not prove its correctness.

The restoration of the correspondence between the natural symmetry properties of space-time on one hand, and the laws of nature on the other hand, is the appealing feature of the proposition. It has, actually, two alarming features. The first of these is that a symmetry operation is, physically, so complicated. If it should turn out that the operation of time inversion, as we now conceive it, is not a valid symmetry operation (e.g., if one of the experiments proposed by Treiman and Wyld gave a positive result) we could still maintain the validity of this symmetry operation by reinterpreting it. We could postulate, for instance, that time inversion transforms matter into *meta*-matter which will be discovered later when higher energy accelerators will become available. Thus, maintaining the validity of symmetry planes forces us to a more artificial view of the concept of symmetry and of the invariance of the laws of physics.

The other alarming feature of our new knowledge is that we have been misled for such a long time to believe in more symmetry elements than actually exist. There was ample reason for this and there was ample experimental evidence to believe that the mirror image of a possible event is again a possible event with electrons being the mirror images of electrons and not of positrons. Let us recall in this connection first how the concept of parity, resulting from the beautiful though almost forgotten experiments of Laporte,[6] appeared to be a perfectly valid concept in spectroscopy and in nuclear physics. This concept could be explained very naturally as a result of the reflection symmetry of space-time, the mirror image of electrons being electrons and not positrons. We are now forced to believe that this symmetry is only approximate and the concept of parity, as used in spectroscopy and nuclear physics, is also only approximate. Even more fundamentally, there is a vast body of experimental information in the chemistry of optically active substances which are mirror images of each other and which have optical activities of opposite direction but exactly equal strength. There is the fact that molecules which have symmetry planes are optically inactive; there is the fact of symmetry planes in crystals.[7] All these facts relate properties of right-handed matter to left-handed *matter*, not of right-handed matter to left-handed *antimatter*. The new experiments leave no doubt that the symmetry plane in this sense is not valid for all phenomena, in particular not valid for β decay, that if the concept of symmetry plane is at all valid for all phenomena, it can be valid only in the sense of converting matter into antimatter.

Furthermore, the old-fashioned type of symmetry plane is not the only symmetry concept that is only approximately valid. Charge conjugation was mentioned before, and we are remainded also of isotopic spin, of the exchange character, that is multiplet system, for electrons and also of nuclei which latter holds so accurately that, in practice, parahydrogen molecules can be converted into orthohydrogen molecules only by first destroying them.[8] This approximate validity of laws of symmetry is, therefore, a very general phenomenon—it may be *the* general phenomenon. We are reminded of Mach's axiom that the laws of nature depend on the physical content of the universe, and the physical content of the universe certainly shows no symmetry. This suggests—and this may also be the spirit of the ideas of Yang and Lee—that all symmetry properties are only approximate. The weakest interaction, the gravitational force, is the basis of the distinction between inertial and accelerated coordinate systems, the second weakest known interaction, that leading to β decay, leads to the distinction between matter and antimatter. Let me conclude this subject by expressing the conviction that the discoveries of Wu, Ambler, Hayward, Hoppes, and Hudson,[9] and of Garwin, Lederman, and Weinreich[10] will not remain isolated discoveries. More likely, they herald a revision of our concept of invariance and possibly of other concepts which are even more taken for granted.

---

[6] O. Laporte, Z. Physik **23**, 135 (1924). For the interpretation of Laporte's rule in terms of the quantum-mechanical operation of inversion, see the writer's *Gruppentheorie und ihre Anwendungen auf die Quantenmechanik der Atomspektren* (Friedrich Vieweg und Sohn, Braunschweig, 1931), Chap. XVIII.

[7] For the role of the space and time inversion operators in classical theory, see H. Zocher and C. Török, Proc. Natl. Acad. Sci. U.S. **39**, 681 (1953) and literature quoted there.

[8] See A. Farkas, *Orthohydrogen, Parahydrogen and Heavy Hydrogen* (Cambridge University Press, New York, 1935).

[9] Wu, Ambler, Hayward, Hoppes, and Hudson, Phys. Rev. **105**, 1413(L) (1957).

[10] Garwin, Lederman, and Weinreich, Phys. Rev. **105**, 1415(L) (1957); also, J. L. Friedman and V. L. Telegdi, *ibid.* **105**, 1681(L) (1957).

QUANTUM LIMITATIONS OF THE CONCEPTS OF GENERAL RELATIVITY

The last remarks naturally bring us to a discussion of the general theory of relativity. The main premise of this theory is that coordinates are only labels to specify space-time points. Their values have no particular significance unless the coordinate system is somehow anchored to events in space-time.

Let us look at the question of how the equations of the general theory of relativity could be verified. The purpose of these equations, as of all equations of physics, is to calculate, from the knowledge of the present, the state of affairs that will prevail in the future. The quantities describing the present state are called initial conditions; the ways these quantities change are called the equations of motion. In relativity theory, the state is described by the metric which consists of a network of points in space-time, that is a network of events, and the distances between these events. If we wish to translate these general statements into something concrete, we must decide what events are, and how we measure distances between *events*. The metric in the general theory of relativity is a metric in space-time, its elements are distances between space-time points, not between points in ordinary space.

The events of the general theory of relativity are coincidences, that is, collisions between particles. The founder of the theory, when he created this concept, had evidently macroscopic bodies in mind. Coincidences, that is, collisions between such bodies, are immediately observable. This is not the case for elementary particles; a collision between these is something much more evanescent. In fact, the point of a collision between two elementary particles can be closely localized in space-time only in case of high-energy collisions. (See Appendix III.) This shows that the establishment of a close network of points in space-time requires a reasonable energy density, a dense forest of world lines wherever the network is to be established. However, it is not necessary to discuss this in detail because the measurement of the distances between the points of the network gives more stringent requirements than the establishment of the network.

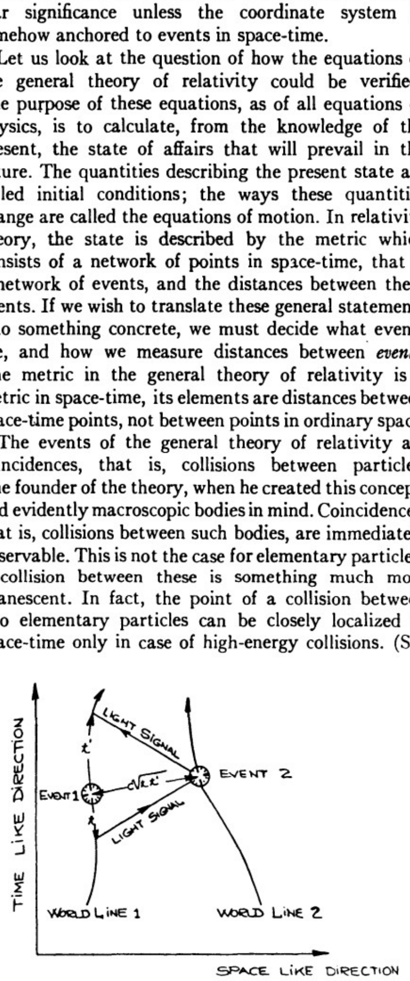

FIG. 4. Measurement of space-like distances by means of a clock. It is assumed that the metric tensor is essentially constant within the space-time region contained in the figure. The space-like distance between events 1 and 2 is measured by means of the light signals which pass through event 2 and a geodesic which goes through event 1. Explanation in Appendix IV.

It is often said that the distances between events must be measured by yardsticks and rods. We found that measurements with a yardstick are rather difficult to describe and that their use would involve a great deal of unnecessary complications. The yardstick gives the distance between events correctly only if its marks coincide with the two events simultaneously from the point of view of the rest-system of the yardstick. Furthermore, it is hard to image yardsticks as anything but macroscopic objects. It is desirable, therefore, to reduce all measurements in space-time to measurements by clocks. Naturally, one can measure by clocks directly only the distances of points which are in time-like relation to each other. The distances of events which are in space-like relation, and which would be measured more naturally by yardsticks, will have to be measured, therefore, indirectly.

It appears, thus, that the simplest framework in space-time, and the one which is most nearly microscopic, is a set of clocks, which are only slowly moving with respect to each other, that is, with world lines which are approximately parallel. These clocks tick off periods and these ticks form the network of events which we wanted to establish. This, at the same time, establishes the distance of those adjacent points which are on the same world line.

Figure 4 shows two world lines and also shows an event, that is, a tick of the clock, on each. The figure shows an artifice which enables one to measure the distance of space-like events: a light signal is sent out from the first clock which strikes the second clock at event 2. This clock, in turn, sends out a light signal which strikes the first clock at time t' after the event 1. If the first light signal had to be sent out at time t before the first event, the calculation given in Appendix IV shows that the space-like distance of events 1 and 2 is the geometric average of the two measured time-like distances t and t'. This is then a way to measure distances between space-like events by clocks instead of yardsticks.

It is interesting to consider the quantum limitations on the accuracy of the conversion of time-like measurements into space-like measurements, which is illustrated in Fig. 4. Naturally, the times t and t' will be well defined only if the light signal is a short pulse. This implies that it is composed of many frequencies and, hence, that its energy spectrum has a corresponding width. As a result, it will give an indeterminate recoil to the second clock, thus further increasing the uncertainty of its momentum. All this is closely related to Heisenberg's uncertainty principle. A more detailed calculation[1] shows that the added uncertainty is of the same order of magnitude as the uncertainty inherent in the nature of the best clock that we could think of, so that the conversion of time-like measurements into space-like measurements is essentially free.

We finally come to the discussion of one of the principal problems—the limitations on the accuracy of the clock. It led us to the conclusion that the inherent limitations on the accuracy of a clock of given weight and size, which should run for a period of a certain length, are quite severe. In fact, the result in summary is that a clock is an essentially nonmicroscopic object. In particular, what we vaguely call an atomic clock, a single atom which ticks off its periods, is surely an idealization which is in conflict with fundamental concepts of measurability. This part of our conclusions can be considered to be well established. On the other hand, the actual formula which will be given for the limitation of the accuracy of time measurement, a sort of uncertainty principle, should be considered as the best present estimate.

Let us state the requirements as follows. The watch shall run T seconds, shall measure time with an accuracy of $T/n=t$, its linear extension shall not exceed l, its mass shall be below m. Since the pointer of the watch must be able to assume n different positions, the system will have to run, in the course of the time T, over at least n orthogonal states. Its state must, therefore, be the superposition of at least n stationary states. It is clear, furthermore, that unless its total energy is at least $\hbar/t$, it cannot measure a time interval which is smaller than t. This is equivalent to the usual uncertainty principle. These two requirements follow directly from the basic principles of quantum theory; they are also the requirements which could well have been anticipated. A clock which conforms with these postulates is, for instance, an oscillator, with a period which is equal to the running time of the clock, if it is with equal probabilty in any of the first n quantum states. Its energy is about n times the energy of the first excited state. This corresponds to the uncertainty principle with the accuracy t as time uncertainty. Broadly speaking, the clock is a very soft oscillator, the oscillating particle moving very slowly and with a rather large amplitude. The pointer of the clock is the position of the oscillating particle.

The clock of the preceding paragraph is still very light. Let us consider, however, the requirement that the linear dimensions of the clock be limited. Since there is little point in dealing with the question in great generality, it may as well be assumed here that the linear dimension shall correspond to the accuracy in time. The requirement $l\approx ct$ increases the mass of the clock by n^3 which may be a very large factor indeed:

$$m>n^3\hbar t/l^2\approx n^3\hbar/c^2t. \tag{2}$$

For example, a clock, with a running time of a day and an accuracy of 10^{-8} second, must weigh almost a gram—for reasons stemming solely from uncertainty principles and similar considerations.

So far, we have paid attention only to the physical dimension of the clock and the requirement that it be able to distinguish between events which are only a distance t apart on the time scale. In order to make it usable as part of the framework which was described before, it is necessary to *read* the clock and to start it. As part of the framework to map out the metric of space-time, it must either register the readings at which it receives impulses, or transmit these readings to a part of space outside the region to be mapped out. This point was already noted by Schrödinger.[11] However, we found it reassuring that, in the most interesting case in which $l=ct$, that is, if space and time inaccuracies are about equal, the reading requirement introduces only an insignificant numerical factor but does not change the form of the expression for the minimum mass of the clock.

The arrangement to map the metric might consist, therefore, of a lattice of clocks, all more or less at rest with respect to each other. All these clocks can emit light signals and receive them. They can also transmit their reading at the time of the receipt of the light signal to the outside. The clocks may resemble oscillators, well in the nonrelativistic region. In fact, the velocity of the oscillating particle is about n times smaller than the velocity of light where n is the ratio of the *error* in the time measurement, to the *duration* of the whole interval to be measured. This last quantity is the spacing of the events on the time axis, it is also the distance of the clocks from each other, divided by the light velocity. The world lines of the clocks from the dense forest which was mentioned before. Its branches suffuse the region of space-time in which the metric is to be mapped out.

We are not absolutely convinced that our clocks are the best possible. Our principal concern is that we have considered only one space-like dimension. One consequence of this was that the oscillator had to be a one-dimensional oscillator. It is possible that the size limitation does not increase the necessary mass of the clock to the same extent if use is made of all three spatial dimensions.

The curvature tensor can be obtained from the metric in the conventional way, if the metric is measured with sufficient accuracy. It may be of interest, nevertheless, the describe a more direct method for measuring the curvature of space. It involves an arrangement, illustrated in Fig. 5, which is similar to that used for obtaining the metric. There is a clock, and a mirror, at such a distance from each other that the curvature of space can be assumed to be constant in the interven-

---

[11] E. Schrödinger, Ber. Preuss. Akad. Wiss. phys.-math. Kl. 1931, 238.

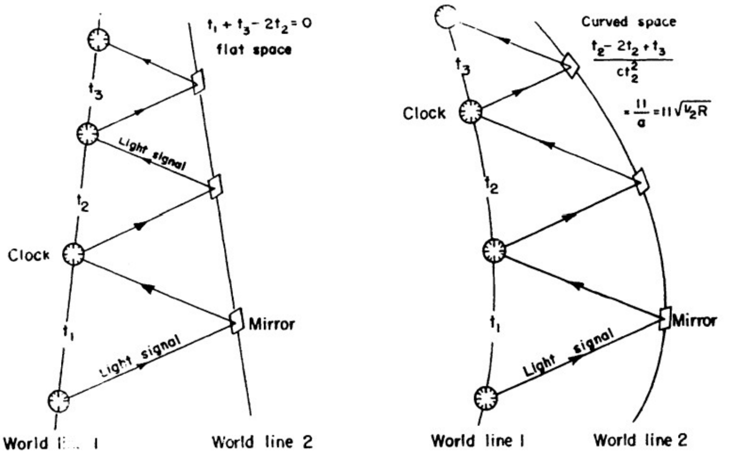

FIG. 5. Direct measurement of the curvature by means of a clock and mirror. Only one space-like dimension is considered and the curvature assumed to be constant within the space-time region contained in the figure. The explanation is given in Appendix V.

ing region. The two clocks need not be at rest with respect to each other, in fact, such a requirement would involve additional measurements to verify it. If the space is flat, the world lines of the clocks can be drawn straight. In order to measure the curvature, a light signal is emitted by the clock, and this is reflected by the mirror. The time of return is read on the clock—it is t_1—and the light signal returned to the mirror. The time which the light signal takes on its second trip to return to the clock is denoted by t_2. The process is repeated a third time, the duration of the last roundtrip denoted by t_3. As shown in Appendix V, the radius of curvature, a, and the relevant component R_{0101} of the Riemann tensor are given by

$$\frac{t_1 - 2t_2 + t_3}{t_2^2} = \frac{11}{a} = 11(\tfrac{1}{2}R_{0101})^{\frac{1}{2}}. \tag{2}$$

If classical theory would be valid also in the microscopic domain, there would be no limit on the accuracy of the measurement indicated in Fig. 5. If h is infinitely small, the time intervals t_1, t_2, t_3 can all be measured with arbitrary accuracy with an infinitely light clock. Similarly, the light signals between clock and mirror, however short, need carry only an infinitesimal amount of momentum and thus deflect clock and mirror arbitrarily little from their geodesic paths. The quantum phenomena considered before force us, however, to use a clock with a minimum mass if the measurement of the time intervals is to have a given accuracy. In the present case, this accuracy must be relatively high unless the time intervals t_1, t_2, t_3 are of the same order of magnitude as the curvature of space. Similarly, the deflection of clock and mirror from their geodesic paths must be very small if the result of the measurement is to be meaningful. This gives an effective limit for the accuracy with which the curvature can be measured. The result is, as could be anticipated, that the curvature at a *point* in space-time cannot be measured at all; only the average curvature over a finite region of space-time can be obtained. The error of the measurement[1] is inversely proportional to the two-thirds power of the area available in space-time, that is, the area around which a vector is carried, always parallel to itself, in the customary definition of the curvature. The error is also proportional to the cube root of the Compton wavelength of the clock. Our principal hesitation in considering this result as definitive is again its being based on the consideration of only one space-like dimension. The possibilities of measuring devices, as well as the problems, may be substantially different in three-dimensional space.

Whether or not this is the case, the essentially nonmicroscopic nature of the general relativistic concepts seems to us inescapable. If we look at this first from a practical point of view, the situation is rather reassuring. We can note first, that the measurement of electric and magnetic fields, as discussed by Bohr and Rosenfeld,[12] also requires macroscopic, in fact *very* macroscopic, equipment and that this does not render the electromagnetic field concepts useless for the purposes of quantum electrodynamics. It is true that the measurement of space-time curvature requires a finite region of space and there is a minimum for the mass, and even the mass uncertainty, of the measuring equipment. However, numerically, the situation is by no means alarming. Even in interstellar space, it should be possible to measure the curvature in a volume of a light second or so. Furthermore, the mass of the clocks which one will wish to employ for such a measurement is of the order of several micrograms so that the finite mass of elementary particles does not cause any difficulty. The clocks will contain many particles and there is no need, and there is not even an incentive, to employ clocks which are lighter than the elementary particles. This is hardly surprising since the mass which can be derived from the gravitational constant, light velocity, and Planck's constant, is about 20 micrograms.

---

[12] N. Bohr and L. Rosenfeld, Kgl. Danske Videnskab. Selskab Mat.-fys. Medd. 12, No. 8 (1933). See also further literature quoted in L. Rosenfeld's article in *Niels Bohr and the Development of Physics* (Pergamon Press, London, 1955).

It is well to repeat, however, that the situation is less satisfactory from a more fundamental point of view. It remains true that we consider, in ordinary quantum theory, position operators as observables without specifying what the coordinates mean. The concepts of quantum field theories are even more weird from the point of view of the basic observation that only coincidences are meaningful. This again is hardly surprising because even a 20-microgram clock is too large for the measurement of atomic times or distances. If we analyze the way in which we "get away" with the use of an absolute space concept, we simply find that we do not. In our experiments we surround the microscopic objects with a very macroscopic framework and observe *coincidences* between the particles emanating from the microscopic system, and parts of the framework. This gives the collision matrix, which is observable, and observable in terms of macroscopic coincidences. However, the so-called observables of the microscopic system are not only not observed, they do not even appear to be meaningful. There is, therefore, a boundary in our experiments between the region in which we use the quantum concepts without worrying about their meaning in face of the fundamental observation of the general theory of relativity, and the surrounding region in which we use concepts which are meaningful also in the face of the basic observation of the general theory of relativity but which cannot be described by means of quantum theory. This appears most unsatisfactory from a strictly logical standpoint.

APPENDIX I

It will be necessary, in this appendix, to compare various states of the same physical system. These states will be generated by looking at the same state—the standard state—from various coordinate systems. Hence every Lorentz frame of reference will define a state of the system—the state as which the standard state appears from the point of view of this coordinate system. In order to define the standard state, we choose an arbitrary but fixed Lorentz frame of reference and stipulate that, in this frame of reference, the particle in the standard state be at rest and its spin (if any) have the direction of the z axis. Thus, if we wish to have a particle moving with a velocity v in the z direction and with a spin also directed along this axis, we look at the particle in the standard state from a coordinate system moving with the velocity v in the $-z$ direction. If we wish to have a particle at rest but with its spin in the yz plane, including an angle α with the z axis, we look at the standard state from a coordinate system the y and z axes of which include an angle α with the y and z axes of the coordinate system in which the standard state was defined. In order to obtain a state in which both velocity and spin have the aforementioned direction (i.e., a direction in the yz plane, including the angles α and $\frac{1}{2}\pi-\alpha$ with the y and z axes), we look at the standard state from the point of view of a coordinate system in which the spin of the standard state is described as this direction and which is moving in the opposite direction.

Two states of the system will be identical only if the Lorentz frames of reference which define them are identical. Under this definition, the relations which will be obtained will be valid independently of the properties of the particle, such as spin or mass (as long as the mass in nonzero so that the standard state exists). Two states will be approximately the same if the two Lorentz frames of reference which define them can be obtained from each other by a very small Lorentz transformation, that is, one which is near the identity. Naturally, all states of a particle which can be compared in this way are related to each other inasmuch as they represent the same standard state viewed from various coordinate systems. However, we shall have to compare only these states.

Let us denote by $A(0,\varphi)$ the matrix of the transformation in which the transformed coordinate system moves with the velocity $-v$ in the z direction where $v=c\tanh\varphi$

$$A(0,\varphi)=\begin{Vmatrix} 1 & 0 & 0 \\ 0 & \cosh\varphi & \sinh\varphi \\ 0 & \sinh\varphi & \cosh\varphi \end{Vmatrix}. \tag{1.1}$$

Since the x axis will play no role in the following consideration, it is suppressed in (1.1) and the three rows and the three columns of this matrix refer to the y', z', ct' and to the y, z, ct axes, respectively. The matrix (1.1) characterizes the state in which the particle moves with a velocity v in the direction of the z axis and its spin is parallel to this axis.

Let us further denote the matrix of the rotation by an angle φ in the yz plane by

$$R(\vartheta)=\begin{Vmatrix} \cos\vartheta & \sin\vartheta & 0 \\ -\sin\vartheta & \cos\vartheta & 0 \\ 0 & 0 & 1 \end{Vmatrix}. \tag{1.2}$$

We refer to the direction in the yz plane which lies between the y and z axes and includes an angle ϑ with the z axis as the direction ϑ. The coordinate system which moves with the velocity $-v$ in the ϑ direction is obtained by the transformation

$$A(\vartheta,\varphi)=R(\vartheta)A(0,\varphi)R(-\vartheta). \tag{1.3}$$

In order to obtain a particle which moves in the direction ϑ and is polarized in this direction, we first rotate the coordinate system counterclockwise by ϑ (to have the particle polarized in the proper direction) and impart it then a velocity $-v$ in the ϑ direction. Hence, it is the transformation

$$T(\vartheta,\varphi)=A(\vartheta,\varphi)R(\vartheta) = \left\|\begin{matrix} \cos\vartheta & \sin\vartheta\cosh\varphi & \sin\vartheta\sinh\varphi \\ -\sin\vartheta & \cos\vartheta\cosh\varphi & \cos\vartheta\sinh\varphi \\ 0 & \sinh\varphi & \cosh\varphi \end{matrix}\right\| \tag{1.4}$$

which characterizes the aforementioned state of the particle. It follows from (1.3) that

$$T(\vartheta,\varphi)=R(\vartheta)A(0,\varphi)=R(\vartheta)T(0,\varphi) \tag{1.5}$$

so that the same state can be obtained also by viewing the state characterized by (1.1) from a coordinate system that is rotated by ϑ. It follows that the statement "velocity and spin are parallel" is invariant under rotations. This had to be expected.

If the state generated by $A(0,\varphi)=T(0,\varphi)$ is viewed from a coordinate system which is moving with the velocity u in the direction of the z axis, the particle will still appear to move in the z direction and its spin will remain parallel to its direction of motion, unless $u>v$ in which case the two directions will become antiparallel, or unless $u=v$ in which case the statement becomes meaningless, the particle appearing to be at rest. Similarly, the other states in which spin and velocity are parallel, i.e., the states generated by the transformations $T(\vartheta,\varphi)$, remain such states if viewed from a coordinate system moving in the direction of the particle's velocity, as long as the coordinate system is not moving faster than the particle. This also had to be expected. However, if the state generated by $T(0,\varphi)$ is viewed from a coordinate system moving with velocity $v'=c\tanh\varphi'$ in the $-y$ direction, spin and velocity will *not* appear parallel any more, *provided the velocity v of the particle is not close to light velocity.* This last proviso is the essential one; it means that the high velocity states of a particle for which spin and velocity are parallel (i.e., the states generated by (1.4) with a large φ) are states of this same nature if viewed from a coordinate system which is not moving too fast in the direction of motion of the particle itself. In the limiting case of the particle moving with light velocity, the aforementioned states become invariant under *all* Lorentz transformations.

Let us first convince ourselves that if the state (1.1) is viewed from a coordinate system moving in the $-y$ direction, its spin and velocity no longer appear parallel. The state in question is generated from the normal state by the transformation

$$A(\tfrac{1}{2}\pi,\varphi')A(0,\varphi) = \left\|\begin{matrix} \cosh\varphi' & \sinh\varphi\sinh\varphi' & \cosh\varphi\sinh\varphi' \\ 0 & \cosh\varphi & \sinh\varphi \\ \sinh\varphi' & \sinh\varphi\cosh\varphi' & \cosh\varphi\cosh\varphi' \end{matrix}\right\|. \tag{1.6}$$

This transformation does not have the form (1.4). In order to bring it into that form, it has to be multiplied on the right by $R(\epsilon)$, i.e., one has to rotate the spin ahead of time. The angle ϵ is given by the equation

$$\tan\epsilon=\frac{\tanh\varphi'}{\sinh\varphi}=\frac{v'}{v}(1-v^2/c^2)^{\frac{1}{2}} \tag{1.7}$$

and is called the angle between spin and velocity. For $v\ll c$, it becomes equal to the angle which the ordinary resultant of two perpendicular velocities, v and v', includes with the first of these. However, ϵ becomes very small if v is close to c; in this case it is hardly necessary to rotate the spin away from the z axis before giving it a velocity in the z direction. These statements express the identity

$$A(\tfrac{1}{2}\pi,\varphi')A(0,\varphi)R(\epsilon)=T(\vartheta,\varphi'') \tag{1.8}$$

which can be verified by direct calculation. The right side represents a particle with parallel spin and velocity, the magnitude and direction of the latter being given by the well-known equations

$$v''=c\tanh\varphi''=(v^2+v'^2-v^2v'^2/c^2)^{\frac{1}{2}} \tag{1.8a}$$

and

$$\tan\vartheta=\frac{\sinh\varphi'}{\tanh\varphi}=\frac{v'}{v(1-v'^2/c^2)^{\frac{1}{2}}}. \tag{1.8b}$$

Equation (1) given in the text follows from (1.7) and (1.8b) for $v\sim c$.

The fact that the states $T(\vartheta,\varphi)\psi_0$ (where ψ_0 is the standard state and $\varphi\gg 1$) are approximately invariant under all Lorentz transformations is expressed mathematically by the equations,

$$R(\vartheta)\cdot T(0,\varphi)\psi_0=T(\vartheta,\varphi)\psi_0, \tag{1.5a}$$

$$A(0,\varphi')\cdot T(0,\varphi)\psi_0=T(0,\varphi'+\varphi)\psi_0, \tag{1.9a}$$

and

$$A(\tfrac{1}{2}\pi,\varphi')\cdot T(0,\varphi)\psi_0\rightarrow T(\vartheta,\varphi'')\psi_0, \tag{1.9b}$$

which give the wave function of the state $T(0,\varphi)\psi_0$, as viewed from other Lorentz frames of reference. Naturally, similar equations apply to all $T(\alpha,\varphi)\psi_0$. In particular, (1.5a) shows that the states in question are invariant under rotations of the coordinate system, (1.9a) that they are invariant with respect to Lorentz transformations with a velocity not too high *in* the direction of motion (so that $\varphi'+\varphi\gg 0$, i.e., φ' not too large a negative number). Finally, in order to prove (1.9b), we calculate the transition probability between the states $A(\frac{1}{2}\pi,\varphi')\cdot T(0,\varphi)\psi_0$ and $T(\vartheta,\varphi'')\psi_0$ where ϑ and φ'' are given by (1.8a) and (1.8b). For this, (1.8) gives

$$\begin{aligned}(A(\tfrac{1}{2}\pi,\varphi')\cdot T(0,\varphi)\psi_0,\ T(\vartheta,\varphi'')\psi_0) &= (T(\vartheta,\varphi'')R(\epsilon)^{-1}\psi_0,\ T(\vartheta,\varphi'')\psi_0) \\ &= (R(\epsilon)^{-1}\psi_0,\psi_0)\rightarrow(\psi_0,\psi_0).\end{aligned}$$

The second line follows because $T(\vartheta,\varphi'')$ represents a coordinate transformation and is, therefore, unitary. The last member follows because $\epsilon\rightarrow 0$ as $\varphi\rightarrow\infty$ as can be seen from (1.7) and $R(0)=1$.

The preceding consideration is not fundamentally new. It is an elaboration of the facts (a) that the subgroup of the Lorentz group which leaves a null-vector invariant is different from the subgroup which leaves a time like vector invariant[4] and (b) that the representations of the latter subgroup decompose into one dimensional representations if this subgroup is "contracted" into the subgroup which leaves a null-vector invariant.[13]

APPENDIX II

Before the hypothesis of Lee and Yang[14] was put forward, it was commonly assumed that there are, in addition to the symmetry operations of the proper Poincaré group, three further independent symmetry operations. The proper Poincaré group consists of all Lorentz transformations which can be continuously obtained from unity and all translations in space-like and time-like directions, as well as the products of all these transformations. It is a continuous group; the Lorentz transformations contained in it do not change the direction of the time axis and their determinant is 1. The three independent further operations which were considered to be rigorously valid, were

Space inversion I, that is, the transformation $x, y, z\rightarrow -x, -y, -z$, without changing particles into antiparticles.

Time inversion T, more appropriately described by Lüders[15] as *Umkehr der Bewegungsrichtung*, which replaces every velocity by the opposite velocity so that the position of the particles at $+t$ becomes the same as it was, without time inversion, at $-t$. The time inversion T (also called time inversion of the first kind by Lüders[16]) does not convert particles into antiparticles either.

Charge conjugation C, that is, the replacement of positive charges by negative charges and more generally of particles by antiparticles, without changing either the position or the velocity of these particles.[17] The quantum-mechanical expressions for the symmetry operations I and C are unitary, that for T is antiunitary.

The three operations I, T, C, together with their products TC (Lüders' time inversion of the second kind), IC, IT, ITC and the unit operation form a group and the products of the elements of this group with those of the proper Poincaré group were considered to be the symmetry operations of all laws of physics. The suggestion given in the text amounts to eliminating the operations I and C separately while continuing to postulate their product IC as symmetry operation. The discrete symmetry group then reduces to the unit operation plus

$$IC,\ T,\ \text{and}\ ICT, \tag{2.1}$$

and the total symmetry group of the laws of physics becomes the proper Poincaré group plus its products with the elements (2.1). This group is isomorphic (essentially identical) with the unrestricted Poincaré group, i.e., the product of *all* Lorentz transformations with all the displacements in space and time. The quantum mechanical expressions for the operations of the proper Lorentz group and its product with IC are unitary, those for T and ICT (as well as for their products with the elements of the proper Poincaré group) antiunitary. Lüders[16] has pointed out that, under certain very natural conditions, ICT belongs to the symmetry group of every *local* field theory.

APPENDIX III

Let us consider, first, the collision of two particles of equal mass m in the coordinate system in which the average of the sum of their momenta is zero. Let us assume that, at a given time, the wave function of both particles is confined to a distance l in the direction of their average velocity with respect to each other. If we consider only this space-like direction, and the time axis, the area in space-time in which the two wave functions will substantially overlap is [see Fig. 6(a)]

$$a=l^2/2v_{\min}, \tag{3.1}$$

where $v_{\min}$ is the lowest velocity which occurs with substantial probability in the wave packets of the colliding particles. Denoting the average momentum by $\bar{p}$ (this has the same value for both particles) the half-width of the momentum distribution by δ, then $v_{\min}=(\bar{p}-\delta)(m^2+(\bar{p}-\delta)^2/c^2)^{-\frac{1}{2}}$. Since l cannot be below h/δ, the area (3.1) is at least

$$\frac{\hbar^2}{2\delta^2}\,\frac{(m^2+(\bar{p}-\delta)^2/c^2)^{\frac{1}{2}}}{\bar{p}-\delta}. \tag{3.1a}$$

(Note that the area becomes infinite if $\delta>\bar{p}$.) The

---

[13] E. Inonu and E. P. Wigner, Proc. Natl. Acad. Sci. U.S. **39**, 510 (1953).

[14] T. D. Lee and C. N. Yang, Phys. Rev. **104**, 254 (1956). See also E. M. Purcell and N. F. Ramsey, Phys. Rev. **78**, 807 (1950).

[15] G. Lüders, Z. Physik **133**, 325 (1952).

[16] G. Lüders, Kgl. Danske Videnskab. Selskab Mat.-fys. Medd. **28**, No. 5 (1954).

[17] All three symmetry operations were first discussed in detail by J. Schwinger, Phys. Rev. **74**, 1439 (1948). See also H. A. Kramers, Proc. Acad. Sci. Amsterdam **40**, 814 (1937) and W. Pauli's article in *Niels Bohr and the Development of Physics* (Pergamon Press, London, 1955). The significance of the first two symmetry operations (and their connection with the concepts of parity and the Kramers degeneracy respectively), were first pointed out by the present writer, Z. Physik 43, 624 (1927) and Nachr. Akad. Wiss. Göttingen, Math.-physik. **1932**, 546. See also T. D. Newton and E. P. Wigner, Revs. Modern Phys. **21**, 400 (1949); S. Watanabe, Revs. Modern Phys. **27**, 26 (1945). The concept of charge conjugation is based on the observation of W. Furry, Phys. Rev. **51**, 125 (1937).

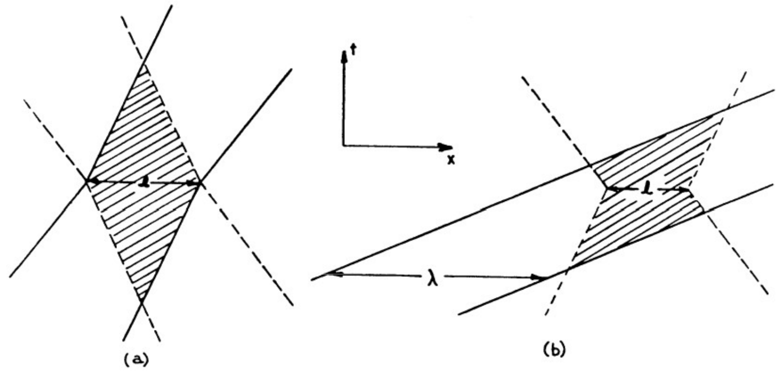

FIG. 6. (a) Localization of a collision of two particles of equal mass. The full lines indicate the effective boundaries of the wave packet of the particle traveling to the right, the broken lines the effective boundaries of the wave packet of the particle traveling to the left. The collision can take place in the shaded area of space-time. (b) Localization of a collision between a particle with finite mass and a particle with zero rest-mass. The full lines, at a distance λ apart in the x direction, indicate the boundary of the particle with zero rest-mass, the broken lines apply to the wave packet of the particle with nonzero rest-mass. The collision can take place in the shaded area.

minimum of (3.1a) is, apart from a numerical factor

$$a_{\min}\approx\frac{\hbar^2}{\bar{p}^3}(m^2+\bar{p}^2/c^2)^{\frac{1}{2}}\approx\frac{\hbar^2c}{E^{\frac{3}{2}}(E+mc^2)^{\frac{1}{2}}},\qquad(3.2)$$

where E is the kinetic energy (total energy minus rest-energy) of the particles.

The kinetic energy E permits the contraction of the wave functions of the colliding particles also in directions perpendicular to the average relative velocity, to an area $\hbar^2c^2/E(E+2mc^2)$. Hence, again apart from a numerical factor, the volume to which the collision can be confined in four dimensional space-time becomes

$$V_{\min}=\frac{\hbar^4c^3}{E^{\frac{5}{2}}(E+mc^2)^{\frac{3}{2}}}.\qquad(3.3)$$

E is the average kinetic energy of the particles in the coordinate system in which their center of mass is, on the average, at rest. Equation (3.3) is valid apart from a numerical constant of unit order of magnitude but this constant depends on E/mc^2.

Let us consider now the opposite limiting case, the collision of a particle with finite rest-mass m with a particle with zero rest-mass. The collision is viewed again in the coordinate system in which the average linear momentum is zero. In this case, one will wish to confine the wave function of the particle with finite rest-mass to a narrower region l than that of the particle with zero rest-mass. If the latter is confined to a region of thickness λ, [see Fig. 6(b)], its momentum and energy uncertainties will be at least $\hbar/\lambda$ and $\hbar c/\lambda$ and these expressions will also give, apart from a numerical factor, the average values of these quantities. Hence $\bar{p}\approx\hbar/\lambda$. The kinetic energy of the particle with finite restmass will be of the order of magnitude

$$\tfrac{1}{2}(m^2c^4+(\bar{p}+\hbar/l)^2c^2)^{\frac{1}{2}}+\tfrac{1}{2}(m^2c^4+(\bar{p}-\hbar/l)^2c^2)^{\frac{1}{2}}-mc^2,\qquad(3.4)$$

since $\hbar/l$ is the momentum uncertainty. Since $l\leq\lambda$, one can neglect $\bar{p}$ in (3.4) if one is interested only in the order of magnitude. This gives for the total kinetic energy,

$$E\approx\hbar c/\lambda+(m^2c^4+\hbar^2c^2/l^2)^{\frac{1}{2}}-mc^2,\qquad(3.5)$$

while the area in Fig. 6(b) is of the order of magnitude

$$a=(\lambda/c)(l+\Delta v\lambda/c),\qquad(3.6)$$

where Δv is the uncertainty in the velocity of the second particle

$$\Delta v=\frac{\bar{p}+\hbar/l}{(m^2+(\bar{p}+\hbar/l)^2/c^2)^{\frac{1}{2}}}-\frac{\bar{p}-\hbar/l}{(m^2+(\bar{p}-\hbar/l)^2/c^2)^{\frac{1}{2}}}.\qquad(3.6a)$$

This can again be replaced by $(\hbar/l)(m^2+\hbar^2/l^2c^2)^{-\frac{1}{2}}$.

For given E, the minimum value of a is assumed if the *kinetic energies* of the two particles are of the same order of magnitude. The two terms of (3.6) then become about equal and $l/\lambda\approx(E/(m+E))^{\frac{1}{2}}$. The minimum value of a, as far as order of magnitude is concerned, is again given by (3.2). Similarly, (3.3) also remains valid if one of the two particles has zero rest-mass.

The two-dimensional case becomes simplest if both particles have zero rest-mass. In this case the wave packets do not spread at all and (3.2) can be immediately seen to be valid. In the four-dimensional case, (3.3) again holds. However, its proof by means of explicitly constructed wave packets (rather than reference to the uncertainty relations) is by no means simple. It requires wave packets which are confined in every direction, do not spread too fast and progress essentially only into one half space (one particle going toward the right, the other toward the left). The construction of such wave packets will not be given in detail. They are necessary to prove (3.2) and (3.3) more rigorously also in the case of finite masses; the preceding proofs, based on the uncertainty relations show only that a and v cannot be *smaller* than the right sides of the corresponding equations. It is clear, in fact, that the limits given by (3.2) and (3.3) would be very difficult to realize, except in the two-dimensional case and for the collision of two particles with zero restmass. In all other cases, the relatively low values of $a_{\min}$ and $V_{\min}$ are predicated on the assumption that the wave packets of the colliding particles are so constituted that they assume a minimum size at the time of the collision. At any rate, (3.2) and (3.3) show that only collisions with a relatively high collision energy, and high energy uncertainty, can be closely localized in space-time.

APPENDIX IV

Let us denote the components of the vector from event 1 to event 2 by x_i, the components of the unit vector along the world line of the first clock at event 1 by e_i. The components of the first light signal are x_i+te_i, that of the second light signal $x_i-t'e_i$. Hence

$$g^{ik}(x_i+te_i)(x_k+te_k)=0 \tag{4.1}$$

$$g^{ik}(x_i-t'e_i)(x_k-t'e_k)=0. \tag{4.2}$$

Elimination of the linear terms in t and t' by multiplication of (4.1) with t' and (4.2) with t and addition gives

$$2g^{ik}x_ix_k+2tt'g^{ik}e_ie_k=0. \tag{4.3}$$

Since e is a unit vector $g^{ik}e_ie_k=1$ and (4.3) shows that the space-like distance between points 1 and 2 is $(tt')^{\frac{1}{2}}$.

APPENDIX V

Since the measurement of the curvature, described in the text, presupposes *constant curvature* over the space-time domain in which the measurement takes place, we use a space with constant curvature, or, rather, part of a space with constant curvature, to carry out the calculation. We consider only one spatial dimension, i.e., a two-dimensional deSitter space. This will be embedded, in the usual way, in a three-dimensional space[18] with coordinates x, y, τ. The points of the deSitter space then form the hyperboloid

$$x^2+y^2-\tau^2=a^2, \tag{5.1}$$

where a is the "radius of the universe." As coordinates of a point we use x and y, or rather the corresponding

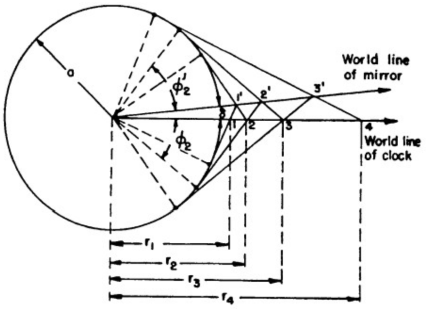

FIG. 7. Analysis of the experiment of Fig. 5. The figure represents a view of the hyperboloid of deSitter space, viewed along its axis. Every point of the plane which is outside the circle corresponds to two points of the deSitter world with one spatial dimension, those with oppositely equal times. The first light signal is emitted at 1, reaches the mirror at 1′, and returns to the clock at 2. The paths of the second and third light signals are 22′3 and 33′4.

polar angles r, ϕ. The metric form in terms of these is

$$ds^2=\frac{a^2}{r^2-a^2}dr^2-r^2d\phi^2. \tag{5.2}$$

Two points of deSitter space correspond to every pair r, ϕ (except $r=a$): those with positive and negative $\tau=(r^2-a^2)^{\frac{1}{2}}$. This will not lead to any confusion as all events take place at positive τ. The null lines (paths of light signals) are the tangents to the $r=a$ circle.

The experiment described in the text can be analyzed by means of Fig. 7. For the sake of simplicity, the clock and mirror are assumed to be "at rest," i.e., their world lines have constant polar angles which will be assumed as 0 and δ, respectively. The first light signal travels from 1 to 1′ and back to 2, the second from 2 to 2′ and back to 3, the third from 3 to 3′ and back to 4. The polar angle of the radius vector which is perpendicular to the first part 22′ of the world line of the second light signal is denoted by ϕ_2. The construction of Fig. 7 shows that angle ϕ_2' which the world line of the mirror includes with the radius vector perpendicular to the second part 2′3 of the second light signal's world line is

$$\phi_2'=\phi_2+\delta. \tag{5.3}$$

The angles ϕ_1, ϕ_1', ϕ_3, ϕ_3' have similar meanings; they are not indicated in the figure in order to avoid overcrowding. For reasons similar to those leading to (5.3), we have

$$\phi_3=\phi_2'+\delta=\phi_2+2\delta \tag{5.3a}$$

$$\phi_1=\phi_2-2\delta \tag{5.3b}$$

$$\phi_4=\phi_3+2\delta=\phi_2+4\delta. \tag{5.3c}$$

---

[18] See, e.g., H. P. Robertson, Revs. Modern Phys. 5, 62 (1933).

The radial coordinates of the points 1, 2, 3, 4 are denoted by r_1, r_2, r_3, r_4

$$r_i = a/\cos\phi_i. \tag{5.4}$$

The proper time t, registered by the clock, can be obtained by integrating the metric form (5.2) along the world line $\phi=0$ of the clock

$$t = a\ln[r+(r^2-a^2)^{\frac{1}{2}}]. \tag{5.5}$$

Hence, the traveling time t_2 of the second light signal becomes

$$t_2 = a\ln\frac{r_3+(r_3^2-a^2)^{\frac{1}{2}}}{r_2+(r_2^2-a^2)^{\frac{1}{2}}} = a\ln\frac{\cos\phi_2(1+\sin\phi_3)}{\cos\phi_3(1+\sin\phi_3)}. \tag{5.6}$$

Similar expressions apply for the traveling times of the first and third light signals; all ϕ can be expressed by means of (5.3a), (5.3b), (5.3c) in terms of ϕ_2 and δ. This allows the calculation of the expression (3). For small δ, one obtains

$$\frac{t_1-2t_2+t_3}{t_2^2} \approx \frac{11}{a}, \tag{5.7}$$

and Riemann's invariant $R=2/a^2$ is proportional to the square of (5.7). In particular, it vanishes if the expression (3) is zero.

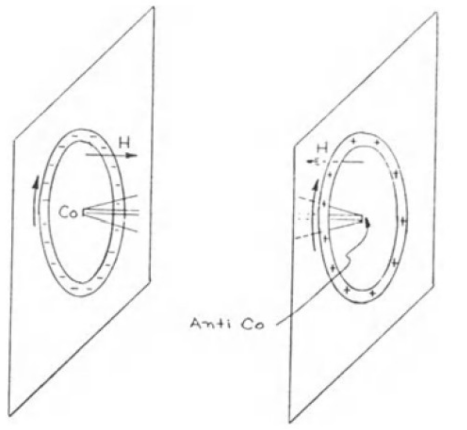
H
Co
H
Anti Co

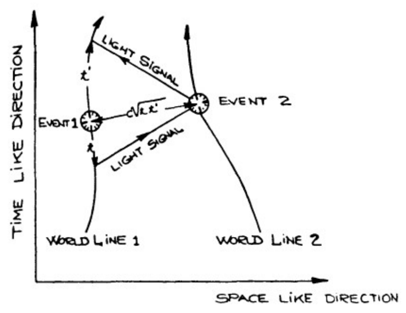
TIME LIKE DIRECTION
LIGHT SIGNAL
t'
EVENT 1
c√tt'
EVENT 2
t
LIGHT SIGNAL
WORLD LINE 1
WORLD LINE 2
SPACE LIKE DIRECTION

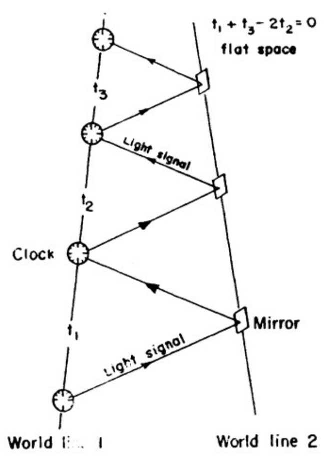
t₁ + t₃ − 2t₂ = 0
flat space
t₃
Light signal
t₂
Clock
Mirror
t₁
Light signal
World line 1
World line 2

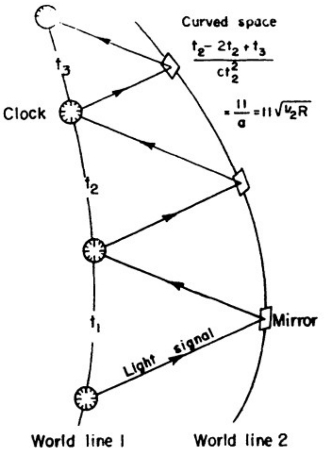
Curved space
$\frac{t_2 - 2t_2 + t_3}{ct_2^2}$
$= \frac{11}{a} = 11\sqrt{\frac{1}{2}R}$
t₃
Clock
t₂
t₁
Light signal
Mirror
World line 1
World line 2

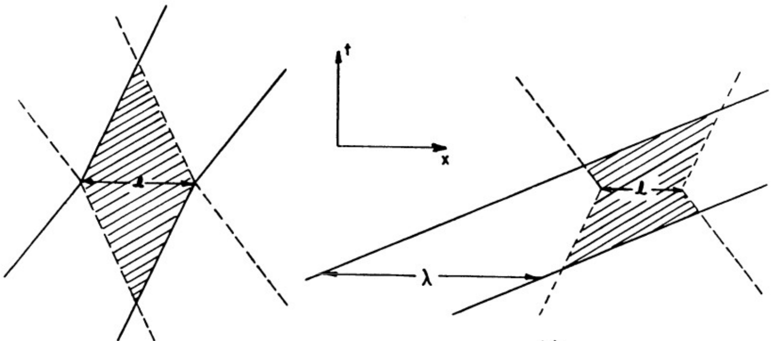
t
x
l
l
λ

(a) (b)

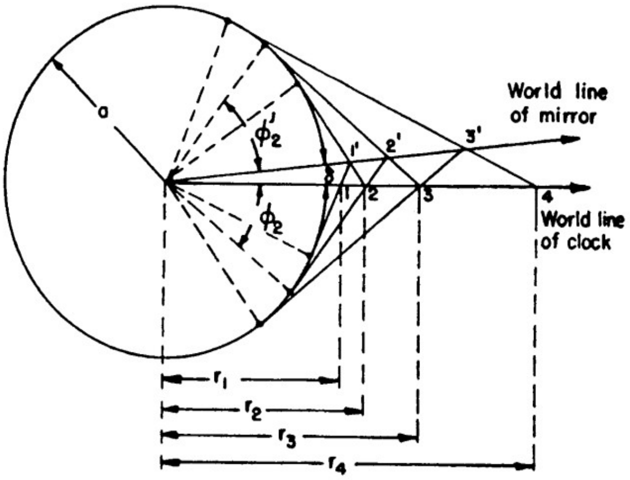
World line
of mirror
World line
of clock
a
ϕ_2
ϕ_2
1'
2'
3'
1
2
3
4
r_1
r_2
r_3
r_4

:

2) Special Relativity and Quantum theory

Source: 20190831_021 A Primer in Tensor Analysis and Relativity / Special Relativity and Quantum theory...pdf

Pág. 10 (23)

Fig. 5 Direct measurement of the curvature by means of a clock and mirror...

Curved space $\frac{t_2 - 2t_2 + t_3}{ct_2^2} = \frac{11}{a} = \sqrt{V_2 R}$

Obs: Interessante !!!

9(:69 9 92 9 , ,,0 A ; ! ? ; (
& ! > ! 3 : * ; .; E . ! 3!
' 0 P(<A: >AE2: A> &RL2C L& (A <& A

CHAPTER 2

DEFORMED MINKOWSKI SPACE-TIME

As is well known, the Minkowski metric[a]

$$g = \mathrm{diag}(1, -1, -1, -1) \tag{2.1}$$

is a generalization of the Euclidean metric $\epsilon = \mathrm{diag}(1, 1, 1)$. On the basis of the discussion of Chap. 1, we assume that the metric describes, in an effective way, the interaction, and that there exist interactions more general than the electromagnetic ones (which, as well known, are long-range and derivable from a potential).

The simplest generalization of the space-time metric which accounts for such more general properties of interactions is provided by a *deformation*, η, of the Minkowski metric (2.1), defined as[3–5]

$$\eta = \mathrm{diag}(b_0^2, -b_1^2, -b_2^2, -b_3^2). \tag{2.2}$$

Of course, from a formal point of view metric (2.2) is not new at all. Deformed Minkowski metrics of the same type have already been proposed in the past[6–9] in various physical frameworks, starting from Finsler's generalization of Riemannian geometry[6] to Bogoslowsky's anisotropic space-time[7] to Santilli's isotopic Minkowski space.[8] A phenomenological deformation of the type (2.2) was also obtained by Nielsen and Picek[9] in the context of the electroweak theory. Moreover, although for quite different purposes, "quantum" deformed Minkowski spaces have been also considered in the context of quantum groups.[10] Leaving to later considerations the true specification of the exact meaning of the deformed metric (2.2) in our framework, let us right now stress two basic points.

---

[a] Throughout this book, lower Latin indices take the values $\{1, 2, 3\}$ and label spatial dimensions, whereas lower Greek indices vary in the range $\{0, 1, 2, 3\}$, with 0 referring to the time dimension. Ordinary 3-vectors are denoted in boldface. Moreover, we adopt the signature $(+, -, -, -)$, and employ the notation "*ESCon*" ("*ESCoff*") to mean that the Einstein sum convention on repeated indices is (is not) used.

1. Firstly, metric (2.2) is supposed to hold at a *local* (and not global) scale, *i.e.* to be valid not everywhere, but only in a suitable (local) space-time region (characteristic of both the system and the interaction considered). We shall therefore refer often to it as a "*topical*" deformed metric.

   In the present case, the term "local" must be understood in the sense that a deformed metric of the kind (2.2) describes the geometry of a 4-dimensional variety attached at a point x of the standard Minkowski space-time, in the same way as a local Lorentz frame is associated (as a tangent space) to each point of the (globally Riemannian) space of Einstein's GR. Another example, on some respects more similar to the present formalism, is provided by a space-time endowed with a vector fibre-bundle structure, where a (maximally symmetric) Riemann space with constant curvature is attached at each point x.[11]
2. Secondly, metric (2.2) is regarded to play a *dynamical role*. So, in order to comply with the solidarity principle, we assume that the parameters b_μ ($\mu = 0, 1, 2, 3$) are, in general, real and positive functions of a given set of observables $\{\mathcal{O}\}$ characterizing the system (in particular, of its total energy, as specified later):

   $$\{b_\mu\} = \{b_\mu(\{\mathcal{O}\})\} \in R_0^+, \quad \forall \{\mathcal{O}\}. \tag{2.3}$$

   The set $\{\mathcal{O}\}$ represents therefore, in general, a set of non-metric variables $(\{x_{n.m.}\})$.

Equation (2.2) therefore becomes:

$$\begin{aligned}\eta_{\mu\nu} &= \eta_{\mu\nu}(\{\mathcal{O}\})\\ &= \mathrm{diag}(b_0^2(\{\mathcal{O}\}), -b_1^2(\{\mathcal{O}\}), -b_2^2(\{\mathcal{O}\}), -b_3^2(\{\mathcal{O}\})) \overset{\text{ESC off}}{=}\\ &= \delta_{\mu\nu}(\delta_{\mu 0} b_0^2(\{\mathcal{O}\}) - \delta_{\mu 1} b_1^2(\{\mathcal{O}\}) - \delta_{\mu 2} b_2^2(\{\mathcal{O}\}) - \delta_{\mu 3} b_3^2(\{\mathcal{O}\})).\end{aligned} \tag{2.4}$$

However, for the moment the deformation of the Minkowski space will be discussed only from a formal point of view, by disregarding the problem of the observables on which the coefficients b_μ actually depend (it will be faced in Chap. 3).

Notice that the first point, *i.e.* the assumed local validity of (2.2), differentiates this approach from those based on Finsler's geometry or from the Bogoslowski's one (which, at least in their standard meaning, do consider deformed metrics at a *global* scale), and makes it similar, on some aspects, to the philosophy and methods of the isotopic generalizations of Minkowski spaces. However, it is well known that Lie-isotopic theories rely in an essential way, from the mathematical standpoint, on (and are strictly characterized by) the very existence of the so-called isotopic unit.[8] In the following, such a formal device will not be exploited (because unessential on all respects), so that the present formalism is not an isotopic one. Moreover, from a physical point of view, the isotopic formalism is expected to apply only to strong interactions.[8] On the contrary, it will be assumed that the (effective) representation of interactions through the deformed metric (2.2) does hold for *all* kinds of interactions (at least for their nonlocal component). In spite of such basic differences this formalism shares some common formal results — as we shall see in the following — with isotopic relativity (like the mathematical expression of the generalized Lorentz transformations).

It is now possible to define a generalized ("*deformed*") Minkowski space $\tilde{M}(x, \eta(\{\mathcal{O}\}))$ with the same local coordinates x of M (the four-vectors of the usual Minkowski space), but with metric given by the metric tensor η (2.4). The generalized interval in $\tilde{M}$ is therefore given by ($x^\mu = (x^0, x^1, x^2, x^3) = (ct, x, y, z)$, with c being the usual light speed in vacuum) (ESC on)[3–5]:

$$\begin{aligned} &ds^{\tilde{2}}(\{\mathcal{O}\}) \\ &\quad\equiv b_0^2(\{\mathcal{O}\})c^2dt^2 - b_1^2(\{\mathcal{O}\})(dx^1)^2 - b_2^2(\{\mathcal{O}\})(dx^2)^2 - b_3^2(\{\mathcal{O}\})(dx^3)^2 \\ &\quad= \eta_{\mu\nu}(\{\mathcal{O}\})dx^\mu dx^\nu = dx * dx. \end{aligned} \tag{2.5}$$

The last step in (2.5) defines the scalar product $*$ in the deformed Minkowski space $\tilde{M}$. Moreover, according to (2.5), we shall use the following notation for the deformed square norm of a four-vector:

$$|x|_*^2 \equiv x * x = \eta_{\mu\nu}(\{\mathcal{O}\})x^\mu x^\nu = x^{\tilde{2}}. \tag{2.5$'$}$$

In order to evidence some preliminary implications of metric (2.4), let us consider (for simplicity sake and without loss of generality) an isotropic 3-dimensional space, *i.e.*

$$b_1^2(\{\mathcal{O}\}) = b_2^2(\{\mathcal{O}\}) = b_3^2(\{\mathcal{O}\}) \equiv b^2(\{\mathcal{O}\}) \tag{2.6}$$

so that the corresponding deformed metric reads

$$\begin{aligned} \eta_{\mu\nu_{\mathrm{ISO}}}(\{\mathcal{O}\}) &= \mathrm{diag}(b_0^2(\{\mathcal{O}\}), -b^2(\{\mathcal{O}\}), -b^2(\{\mathcal{O}\}), -b^2(\{\mathcal{O}\})) \stackrel{\text{ESC off}}{=} \\ &= \delta_{\mu\nu}[\delta_{\mu 0}b^2(\{\mathcal{O}\}) - (\delta_{\mu 1} + \delta_{\mu 1} + \delta_{\mu 1})b^2(\{\mathcal{O}\})]. \end{aligned} \tag{2.7}$$

One gets, for null separation $ds^{\tilde{2}} = 0$:

$$\begin{aligned} ds^{\tilde{2}}(\{\mathcal{O}\}) = 0 &\Leftrightarrow b_0^2(\{\mathcal{O}\})c^2dt^2 - b^2(\{\mathcal{O}\})[(dx^1)^2 + (dx^2)^2 + (dx^3)^2] = 0 \\ &\Leftrightarrow \frac{(dx^2 + dy^2 + dz^2)}{dt^2} = c^2\frac{b_0^2(\{\mathcal{O}\})}{b^2(\{\mathcal{O}\})}. \end{aligned} \tag{2.8}$$

From Eq. (2.8) it is easily seen that the *maximal causal speed* in the generalized Minkowski space is given by:

$$u = \frac{b_0(\{\mathcal{O}\})}{b(\{\mathcal{O}\})}c. \tag{2.9}$$

It is worth noticing that a similar result (namely, a "maximum attainable speed", *a priori* different for different physical processes) was also obtained by Coleman and Glashow,[12] in the framework of a discussion of possible effects breaking Lorentz invariance (essentially on a local scale).

Let us remark that u depends explicitly on the metric parameters b_μ, which are *a priori* different for every physical system. However, since the deformation of the metric represents, on average, the effects of the nonlocal interactions involved, it is expected that physical systems with the same kind of interactions (besides the electromagnetic ones) are described by metric parameters of the same order of magnitude (or, at least, this holds true for the ratio b_0/b). In this sense it is possible to refer to u as a "speed of interaction", rather than "speed of the physical system" considered (of course, at the same energy scale).

In Eq. (2.9), the value of u is parametrized in terms of c, and depends on the physical system (and its interactions). Moreover, it is

$$u \gtreqless c \Leftrightarrow \frac{b_0(\{\mathcal{O}\})}{b(\{\mathcal{O}\})} \gtreqless 1. \tag{2.10}$$

In other words, there may be maximal causal speeds either *subluminal* or *superluminal*, depending on the interaction considered.

It is worth to recall that the deformation of the metric, resulting in the interval (2.5), represents a geometrization of a suitable space-time region (corresponding to the physical system considered) that describes, in the average, the effect of nonlocal interactions on a test particle. It is clear that there exist infinitely many deformations of the Minkowski space (precisely, ∞^4), corresponding to the different possible choices of the parameters b_μ, *a priori* different for each physical system.

Moreover, since the usual, "flat" Minkowski metric g (2.1) is related in an essential way to the electromagnetic interaction, we shall always mean in the following — unless otherwise specified — that electromagnetic interactions imply the presence of a fully Minkowskian metric. Actually, as it will be seen, a deformed metric of the type (2.7) is required if one wants to account for possible nonlocal electromagnetic effects (like those occurring in the superluminal wave propagation in waveguides: see Chap. 7).

CHAPTER 4

GENERALIZED PRINCIPLE OF RELATIVITY AND LORENTZ TRANSFORMATIONS

4.1. Deformed Special Relativity

In order to develop the relativity theory in a deformed Minkowski space-time, one has to suitably generalize and clarify the basic concepts which are at the very foundation of SR.

Let us first of all define a “topical inertial frame”:

(i) a *topical “inertial” frame* (TIF) is a reference frame in which space-time is homogeneous, but space is not necessarily isotropic.

Then, a “*generalized principle of relativity*” (or “*principle of metric invariance*”) can be stated as follows:

(ii) all physical measurements within every topical “inertial” frame must be carried out via the *same* metric.

We shall call “*Deformed Special Relativity*” (DSR) the generalization of SR based on the above two postulates, and whose space-time structure is given by the deformed Minkowski space $\tilde{M}$ introduced in Chap. 2.

4.2. Deriving the Deformed Lorentz Transformations

It follows from the above points (i), (ii) that the transformation equations connecting topical “inertial” frames, *i.e.* the generalized Lorentz transformations, are those which leave invariant the deformed metric when passing from a topical “inertial” frame K, to another frame K', moving with constant velocity with respect to K. Then, physical laws are to be covariant with respect to such generalized transformations.

In other words, the generalized Lorentz transformations are the isometries of the deformed Minkowski space $\tilde{M}$. We shall refer to them in the following as *deformed Lorentz transformations* (DLT). Their explicit form can be derived by the same procedure followed in order to find the Lorentz transformations in the usual Minkowski space.

Consider two TIF, K and K'; by definition, the DLT's leave invariant the deformed interval (2.5), *i.e.*

$$b_0^2c^2t^2 - b_1^2x^2 - b_2^2y^2 - b_3^2z^2 = b_0^2c^2t'^2 - b_1^2x'^2 - b_2^2y'^2 - b_3^2z'^2. \tag{4.1}$$

Moreover, it can be assumed, without loss of generality, that the frames K and K' are in standard configuration (*i.e.* their spatial frames coincide at $t = t' = 0$). By choosing the boost direction along $\hat{x}^1 = \hat{x}$, we have therefore $y' = y$, $z' = z$ and Eq. (4.1) reduces to

$$b_0^2c^2t^2 - b_1^2x^2 = b_0^2c^2t'^2 - b_1^2x'^2. \tag{4.2}$$

From space-time homogeneity it follows that the functional relations between the two sets of coordinates $\{x, y, z, t\}$ and $\{x', y', z', t'\}$ must be linear. Then, in general, the deformed coordinate transformations are to be searched in the form

$$\begin{cases} x' = A_{11}x + A_{14}t \\ y' = y \\ z' = z \\ t' = A_{41}x + A_{44}t \end{cases} \tag{4.3}$$

where the coefficients $A_{11}, A_{14}, A_{41}, A_{44}$ depend *a priori* in general on $\mathbf{v}$ and $\hat{x}$ (and, parametrically, on the energy).

Notice that the origin O' of TIF K' must move in K with velocity $\mathbf{v} = v^1\hat{x}$, and therefore:

$$x' = 0, \quad x = vt \Leftrightarrow A_{14} = -vA_{11} \Leftrightarrow x' = A_{11}(x - vt). \tag{4.4}$$

Replacing (4.3), (4.4) in (4.2) yields

$$b_0^2c^2t^2 - b_1^2x^2 = b_0^2c^2(A_{41}x + A_{44}t)^2 - A_{11}^2b_1^2x^2(x - vt)^2 \tag{4.5}$$

which implies the following 3×3 quadratic system:

$$\begin{cases} c^2 = c^2A_{44}^2 - \left(\dfrac{b_1}{b_0}\right)^2 A_{11}^2v^2 \\ -1 = c^2\left(\dfrac{b_0}{b_1}\right)^2 A_{41}^2 - A_{11}^2 \\ 0 = c^2\left(\dfrac{b_0}{b_1}\right)^2 A_{41}A_{44} + A_{11}^2v \end{cases} \tag{4.6}$$

with general solution

$$A_{11} = A_{44} = \pm \left(1 - \left(\frac{vb_1}{cb_0}\right)^2\right)^{-1/2} \tag{4.7}$$

$$A_{41} = \mp \left(\frac{vb_1^2}{c^2b_0^2}\right)\left(1 - \left(\frac{vb_1}{cb_0}\right)^2\right)^{-1/2} = -\left(\frac{vb_1^2}{c^2b_0^2}\right)A_{11}. \tag{4.8}$$

The final result is

$$\begin{cases} x' = \tilde{\gamma}(x - vt) = \tilde{\gamma}\left(x - \tilde{\beta}\frac{b_0}{b_1}ct\right) \\ y' = y \\ z' = z \\ t' = \tilde{\gamma}\left(t - \frac{vb_1^2}{c^2b_0^2}x\right) = \tilde{\gamma}\left(t - \frac{\tilde{\beta}^2}{v}x\right) \end{cases} \tag{4.9}$$

where v is the relative speed of the reference frames, and

$$\tilde{\beta} = \frac{v}{u}; \tag{4.10}$$

$$\tilde{\gamma} = (1 - \tilde{\beta}^2)^{-1/2}. \tag{4.11}$$

Transformations (4.9) do formally coincide with the isotopic Lorentz transformations. However, in the present context their physical meaning is different, as it is easily seen e.g. by the identification of the maximal causal speed u with the speed characteristic of the quanta of a given interaction (see Sec. 3). In particular, the parametrization (4.10) of the deformed velocity parameter $\tilde{\beta}$ in terms of u immediately shows that is always $\tilde{\beta} < 1$, so that $\tilde{\gamma}$ never takes imaginary values (contrarily to the isotopic case). Moreover, no reference at all is made, in this framework, to the existence of an underlying "medium".

It must be carefully noted that, like the metric, also the generalized Lorentz transformations depend on the energy. This means that one gets different transformation laws for different values of E, but still with the same functional dependence on the energy, so that the invariance of the deformed interval (2.5) is always ensured (provided that the process considered does always occur via the same interaction).

Indeed, the energy E can be considered fixed also because, from a quantum point of view, energy can be transferred only by finite amounts.

Differentiating Eqs. (4.9), we get therefore

$$\begin{cases} udt' + t'du = \tilde{\gamma}(udt - \tilde{\beta}dx) + [d\tilde{\gamma}(ut - \tilde{\beta}x) + \tilde{\gamma}(tdu - xd\tilde{\beta}]; \\ dx' = \tilde{\gamma}(dx - \tilde{\beta}udt) + [d\tilde{\gamma}(x - \tilde{\beta}ut) - \tilde{\gamma}(t\tilde{\beta}du + td\tilde{\beta})], \end{cases} \tag{4.12}$$

where, by the above argument, $dE = 0$ and therefore $d\tilde{\gamma} = d\tilde{\beta} = du = 0$. Squaring (4.12) and subtracting, we find

$$dx'^2 - u^2dt'^2 = \tilde{\gamma}^2[(dx - \tilde{\beta}dt)^2 - (udt - \tilde{\beta}dx)^2] = dx^2 - u^2dt^2 \tag{4.13}$$

where in the last step use has been made of Eq. (4.2). Exploiting the explicit expression of u, Eq. (2.9), one has finally

$$ds'^2 = ds^2, \tag{4.14}$$

i.e. the deformed Lorentz transformations (4.9) are actually the isometries of the deformed Minkowski space $\tilde{M}$, in spite of their dependence on the energy.

As already stressed in Chap. 2 (see Eq. (2.10)), it is possible to have $u \geq c$ and therefore $c \leq v \leq u$, *i.e.* superluminal motions are allowed. Let us remark that the possibility of tachyonic speeds is accomplished, within this framework, without any recourse neither to imaginary quantities nor to singularities in the transformation laws (unlike the standard case), because it is always $v \leq u$ (even if $v \geq c$), so that the relativistic factor $\tilde{\gamma}$ (Eq. (4.11)) takes only real values, as already noted above.

4.3. Maximal Causal Speed Revisited

In $\tilde{M}$, it is possible *a priori* to consider two scalar products between 3-vectors $\mathbf{v}_1, \mathbf{v}_2$: the standard, Euclidean one $\cdot$, defined by means of the metric tensor $g_{ik} = \delta_{ik}$, and the deformed one, induced by the deformed scalar product $*$ in $\tilde{M}$, and defined by means of the metric tensor $-\eta_{ik}(E) \overset{\text{ESC off}}{=} b_i^2(E)\delta_{ik}$ (where the sign $-$ is obviously introduced in order to get a positive 3-vector norm) as follows (cf. Eq. (2.5)):

$$\begin{aligned} \mathbf{v}_1 * \mathbf{v}_2 &\equiv -\sum_{i=1}^{3} \eta_{ij}(E)(v_1)^i(v_2)^j \\ &= \sum_{i=1}^{3} b_i^2(E)\delta_{ij}(v_1)^i(v_2)^j \\ &= b_1^2(E)(v_1)^1(v_2)^1 + b_2^2(E)(v_1)^2(v_2)^2 + b_3^2(E)(v_1)^3(v_2)^3. \end{aligned} \tag{4.15}$$

CHAPTER 11

METRICS OF INTERACTIONS

The metric representation derived in the previous Chapters are, obviously, not ultimate. They constitute a mere phenomenological attempt at applying the formalism of the DSR (developed in Parts I and II), by showing explicitly how to pass from a deformed Minkowski metric to a metric representation of a given interaction.

The main aim of the present Chapter is just to provide a review of the phenomenological metrics derived for all four fundamental interactions, by stressing their general structure and their features as a premise for further developments.

11.1. Review of Phenomenological Metrics

11.1.1. *Energy-Dependent Phenomenological Metrics for the Four Interactions*

In Chaps. 7–10 it was shown that a local breakdown of Lorentz invariance may be envisaged for all four fundamental interactions (electromagnetic, weak, strong and gravitational) whereby one gets evidence for a departure of the space-time metric from the Minkowskian one (at least in the energy range examined). Let us review the basic features of the energy-dependent phenomenological metrics derived above:

(1) Both the electromagnetic and the weak metric show the same functional behavior, namely

$$\eta(E) = \mathrm{diag}(1, -b^2(E), -b^2(E), -b^2(E)); \tag{11.1}$$

$$b^2(E) = \begin{cases} (E/E_0)^{1/3}, & 0 < E \le E_0 \\ 1, & E_0 < E \end{cases} \tag{11.2}$$

$$= 1 + \theta(E_0 - E)\left[\left(\frac{E}{E_0}\right)^{1/3} - 1\right], \quad E > 0 \tag{11.3}$$

(where $\theta(x)$ is the Heaviside theta function) with the only difference between them being the threshold energy E_0, *i.e.* the energy value at which the metric parameters are constant (namely the metric becomes Minkowskian: $\eta_{\mu\nu}(E \geq E_0) \equiv g_{\mu\nu} = \mathrm{diag}(1,-1,-1,-1)$); the fits to the experimental data yield

$$E_{0,\mathrm{e.m}} = (4.5 \pm 0.2)\ \mu\mathrm{eV}; \quad \mathrm{E}_{0,\mathrm{weak}} = (80.4 \pm 0.2)\ \mathrm{GeV}. \tag{11.4}$$

Notice that for either interaction the metric is isochronous, spatially isotropic and subminkowskian, *i.e.* it approaches the Minkowskian limit from below (for $E < E_0$). Both metrics are therefore Minkowskian for $E > E_{0w} \simeq 80\,\mathrm{GeV}$, and then our formalism is fully consistent with electroweak unification, which occurs at an energy scale $\sim$100 GeV.

(2) The strong metric reads

$$\begin{aligned}\eta_{\mathrm{strong}}(E) = \mathrm{diag}(&b^2_{0,\mathrm{strong}}(E), -b^2_{1,\mathrm{strong}}(E),\\ &- b^2_{2,\mathrm{strong}}(E), -b^2_{3,\mathrm{strong}}(E));\end{aligned} \tag{11.5}$$

$$b^2_{1,\mathrm{strong}}(E) = \left(\frac{\sqrt{2}}{5}\right)^2; \quad b^2_{2,\mathrm{strong}}(E) = \left(\frac{2}{5}\right)^2, \quad \forall E > 0 \tag{11.6}$$

$$\begin{aligned}b^2_{0,\mathrm{strong}}(E) = b^2_{3,\mathrm{strong}}(E) &= \begin{cases}1, & 0 < E \leq E_{0,\mathrm{strong}}\\ (E/E_{0,\mathrm{strong}})^2, & E_{0,\mathrm{strong}} < E\end{cases}\\ &= 1 + \theta(E - E_{0,\mathrm{strong}})\left[\left(\frac{E}{E_{0,\mathrm{strong}}}\right)^2 - 1\right], \quad E > 0\end{aligned} \tag{11.7}$$

with

$$E_{0,\mathrm{strong}} = (367.5 \pm 0.4)\,\mathrm{GeV}. \tag{11.8}$$

In this case, contrarily to the electromagnetic and the weak ones, a deformation of the time coordinate occurs; moreover, the three-space is anisotropic, with two spatial parameters constant (but different in value) and the third one variable with energy like the time one.

(3) The explicit form of the gravitational energy-dependent metric is:

$$\eta_{\mathrm{grav}}(E) = \mathrm{diag}(b^2_{0,\mathrm{grav}}(E), -b^2_{1,\mathrm{grav}}(E), -b^2_{2,\mathrm{grav}}(E), -b^2_{3,\mathrm{grav}}(E)); \tag{11.9}$$

$$
\begin{aligned}
b^2_{0,\mathrm{grav}}(E) = b^2_{3,\mathrm{grav}}(E) &= \begin{cases} 1, & 0 < E \leq E_{0,\mathrm{grav}} \\ \dfrac{1}{4}(1 + E/E_{0,\mathrm{grav}})^2, & E_{0,\mathrm{grav}} < E \end{cases} \\
&= 1 + \theta(E - E_{0,\mathrm{grav}}) \left[\frac{1}{4} \left(1 + \frac{E}{E_{0,\mathrm{grav}}} \right)^2 - 1 \right], \quad E > 0
\end{aligned}
\tag{11.10}
$$

(the coefficients $b^2_{1,\mathrm{grav}}(E)$ and $b^2_{2,\mathrm{grav}}(E)$ are presently undetermined at phenomenological level), with

$$E_{0,\mathrm{grav}} = (20.2 \pm 0.1)\,\mu\mathrm{eV}. \tag{11.11}$$

It is worth stressing the analogy between the strong and the gravitational metrics. In both cases, a deformation of the time coordinate occurs. Moreover, one of the spatial parameters (we conventionally assumed as the third one) varies with energy as the time one in an overminkowskian way, namely, they approach the Minkowskian limit from above ($E_0 < E$). The other two spatial parameters are constant, but different in value for the strong case (*i.e.* the three-space is anisotropic for strong interactions), whereas nothing can be said for the gravitational case on an experimental basis.

The general pattern of the four phenomenological metrics is shown in Fig. 11.1.

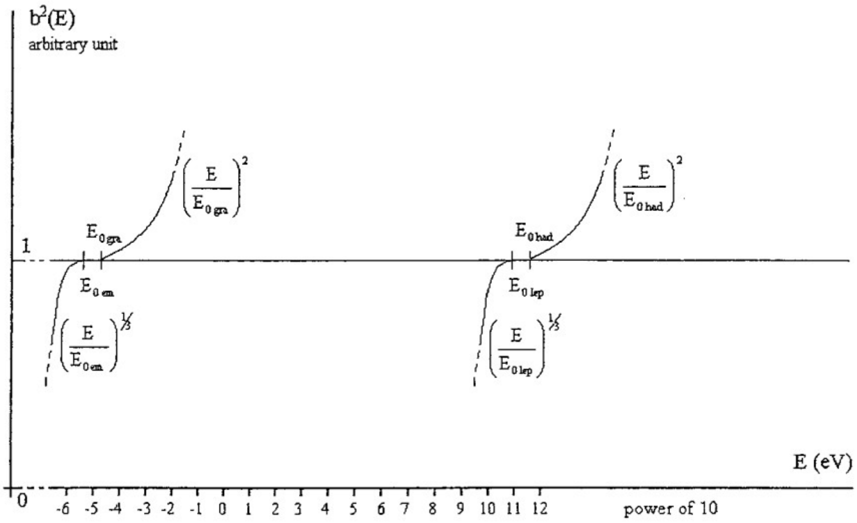

Fig. 11.1. General pattern of the phenomenological metrics for the four fundamental interactions.

in the anisotropic case, one exploits the values of the maximal causal speeds derived from the analysis of the strong interaction based on BE correlation (the only interaction at present described by a spatially anisotropic metric) (cf. Chap. 10).

13.5. Appendix — Einstein and the Aether

When Einstein started his scientific life as a young physicist, the aether was conceived as a material medium allowing propagation of the electromagnetic signals and filling the interplanetary and interstellar spaces. This was for instance the idea of Maxwell, as formulated by him in the article he wrote in the *Encyclopedia Britannica.* According to Drude and Cohn, the aether might also be thought as the vacuum, endowed with suitable physical properties.

Einstein was so convinced about the reality of the aether, that he devised as early as 1895 an experiment aimed to detect the drift of the aether. He pursued such an idea when a student at Zurich ETH in 1897. However, he was not allowed to carry out such an experiment by his professor H. F. Weber. Later, he sent to his uncle C. Koch a manuscript (remained unpublished) entitled *On the Examination of the State of the Aether in a Magnetic field.* In the experiment therein proposed, he wanted to investigate the reciprocal effect between the magnetic field generated by a steady current and the aether, regarded as an elastic medium.

Another experimental proposal was exposed by Einstein in a letter to M. Grossmann from Winterthur in 1901. Subsequently, Einstein abandoned the idea of the aether in the old sense, and rather identified it, within General Relativity, as a substratum without mechanical and kinematical properties, but able to codetermine mechanical and electromagnetic events. In a modern view, we can state that such a new concept of aether is nothing but a static gravitational field (locally the Minkowski space-time).[82]

The possible existence of the aether is indeed nothing, in today language, than the possibility of LLI breakdown. We can therefore conclude that Einstein asked the right question, but looking for the wrong answer. His experiment based on the feedback between a steady current and the aether has strong analogies with Bartocci's one we discussed above. In particular, it belongs to the same class of LLI breakdown tests, namely those based on electromagnetic effects.

CHAPTER 14

THE GRAVITATIONAL MASS OF ELECTRON FROM GEOMETRY

In this Chapter, it will be shown that the DSR formalism yields an expression of the electron mass m_e in terms of the parameter δ of local Lorentz invariance (LLI) breakdown and of the threshold energy for the gravitational metric, $E_{0,\text{grav}}$ (*i.e.* as by now familiar, the energy value under which the metric becomes Minkowskian). This allows one to evaluate m_e from the (experimental) knowledge of such parameters.[83]

14.1. LLI Breaking Factor and Relativistic Energy in DSR

The breakdown of standard local Lorentz invariance (LLI) is expressed by the LLI breaking factor parameter δ discussed in Sec. 13.4. In particular, the isotropic parameter is expressed by Eq. (13.14). To the present aims, one can also define δ as follows:

$$\delta_{\text{int}} \equiv \frac{m_{\text{in,int}} - m_{\text{in,grav}}}{m_{\text{in,int}}} = 1 - \frac{m_{\text{in,grav}}}{m_{\text{in,int}}} \tag{14.1}$$

where $m_{\text{in,int}}$ is the inertial mass of the particle considered with respect to the given interaction.[a] In other words, we assume that the *local* deformation of space-time corresponding to the interaction considered, and described by the metric (3.3), gives rise to a *local violation* of the Principle of Equivalence for interactions different from the gravitational one. Such a departure, just expressed by the parameter δ_{int}, does constitute also a measure of the amount of LLI breakdown. In the framework of DSR, δ_{int} embodies the geometrical contribution to the inertial mass, thus discriminating between two different metric structures of space-time.

---

[a] In the following, "int" denotes a physically detectable fundamental interaction, which can be operationally defined by means of a phenomenological energy-dependent metric of deformed minkowskian type.

Of course, if the interaction considered is the gravitational one, the Principle of Equivalence strictly holds, *i.e.*

$$m_{\text{in,grav}} = m_g \tag{14.2}$$

where m_g is the gravitational mass of the physical object considered, *i.e.* it is its "gravitational charge" (namely its coupling constant to the gravitational field).

Then, (14.1) can be rewritten as:

$$\delta_{\text{int}} \equiv \frac{m_{\text{in,int}} - m_g}{m_{\text{in,int}}} = 1 - \frac{m_g}{m_{\text{in,int}}} \tag{14.3}$$

and therefore, when the particle is subjected only to gravitational interaction, it is

$$\delta_{\text{grav}} = 0. \tag{14.4}$$

Let us recall the expression (5.10) of the deformed relativistic energy, for a particle subjected to a given interaction and moving along $\hat{x}^i$:

$$\begin{aligned} E_{\text{int}} &= m_{\text{in,int}} u_{i,\text{int}}^2(E) \tilde{\gamma}_{\text{int}}(E) \\ &= m_{\text{in,int}} c^2 \frac{b_{0,\text{int}}^2(E)}{b_{i,\text{int}}^2(E)} \left[1 - \left(\frac{v_i b_{i,\text{int}}(E)}{c b_{0,\text{int}}(E)} \right)^2 \right]^{-1/2} \end{aligned} \tag{14.5}$$

(with $\mathbf{u}_{\text{int}}(E)$ the maximal causal velocity for the interaction considered), which, in the non-relativistic (NR) limit of DSR (*i.e.* $v_i \ll u_{i,\text{int}}(E)$), reduces to (cf. Eq. (5.12))

$$E_{\text{int,NR}} = m_{\text{in,int}} u_{i,\text{int}}^2(E) = m_{\text{in,int}} c^2 \frac{b_{0,\text{int}}^2(E)}{b_{i,\text{int}}^2(E)}. \tag{14.6}$$

In the case of the gravitational metric (see Eq. (8.28)), we have

$$\frac{b_{0,\text{grav}}^2(E)}{b_{3,\text{grav}}^2(E)} = 1, \quad \forall E \in R_0^+. \tag{14.7}$$

Therefore, for $i = 3$, Eqs. (14.5) and (14.6) become, respectively ($v_3 = v$):

$$E_{\text{grav}} = m_g c^2 \left[1 - \left(\frac{v}{c} \right)^2 \right]^{-1/2} = m_g c^2 \gamma \tag{14.8}$$

$$E_{\text{grav,NR}} = m_g c^2 \tag{14.9}$$

namely, the gravitational energy takes its standard, special-relativistic values.

This means that the special characterization (corresponding to the choice $i = 3$) of Eqs. (14.5) and (14.6) within the framework of DSR relates the gravitational interaction with SR, which is — as well known — based on the electromagnetic interaction in its Minkowskian form.

14.2. The Electron as a Fundamental Particle and its "Geometrical" Mass

Let us consider for E the threshold energy of the gravitational interaction:

$$E = E_{0,\text{grav}}. \tag{14.10}$$

By definition of $E_{0,\text{grav}}$, one has therefore:

$$\begin{aligned}\eta_{\mu\nu,\text{grav}}(E) &= \operatorname{diag}(1, -b^2_{1,\text{grav}}(E), -b^2_{2,\text{grav}}(E), -1)\\ &\overset{\text{ESC off}}{=} \delta_{\mu\nu}\left[\delta_{\mu 0} - \delta_{\mu 1} b^2_{1,\text{grav}}(E) - \delta_{\mu 2} b^2_{2,\text{grav}}(E) - \delta_{\mu 3}\right],\\ &\quad \forall E \in (0, E_{0,\text{grav}}].\end{aligned}$$

Notice that at the energy $E = E_{0,\text{grav}}$ the electromagnetic metric (7.8) is Minkowskian, too (because $E_{0,\text{grav}} > E_{0,\text{em}}$).

On the basis of the previous considerations, it seems reasonable to assume that the physical object (particle) p with a rest energy (*i.e.* gravitational mass) just equal to the threshold energy $E_{0,\text{grav}}$, namely

$$E_{0,\text{grav}} = m_{g.,p}c^2, \tag{14.11}$$

must play a fundamental role for either e.m. and gravitational interaction. We can e.g. hypothesize that p corresponds to the lightest mass eigenstate which experiences both force fields (*i.e.* from a quantum viewpoint, coupling to the respective interaction carriers, the photon and the graviton). As a consequence, p must be intrinsically stable, due to the impossibility of its decay in lighter mass eigenstates, even in the case such a particle is subject to weak interaction, too (*i.e.* it couples to all gauge bosons of the Glashow–Weinberg–Salam group $SU(2)\times U(1)$, not only to its electromagnetic charge sector).

Since, as above seen, for $E = E_{0,\text{grav}}$ the electromagnetic metric is Minkowskian, too, it is natural to assume, for p:

$$m_{\text{in},p,\text{e.m.}} = m_{\text{in},p} \tag{14.12}$$

namely its inertial mass is that measured with respect to the electromagnetic metric.

Then, due to the Equivalence Principle (see Eq. (14.2)), the mass of p is characterized by

$$p:\begin{cases} m_{\mathrm{in},p,\mathrm{grav}} = m_{g,p} \\ m_{\mathrm{in},p,\mathrm{e.m.}} = m_{\mathrm{in},p}. \end{cases} \tag{14.13}$$

Therefore, for such a fundamental particle the LLI breaking factor (14.1) of the e.m. interaction becomes:

$$\delta_{\mathrm{e.m.}} = \frac{m_{\mathrm{in,p}} - m_{\mathrm{g,p}}}{m_{\mathrm{in,p}}} = 1 - \frac{m_{\mathrm{g,p}}}{m_{\mathrm{in,p}}} \Leftrightarrow m_{\mathrm{g,p}} = m_{\mathrm{in,p}}\left(1 - \delta_{\mathrm{e.m.}}\right). \tag{14.14}$$

Replacing (14.11) in (14.14) yields:

$$E_{0,\mathrm{grav}} = m_{in,\mathrm{p}}\left(1 - \delta_{\mathrm{e.m.}}\right)c^2 \Leftrightarrow m_{in,\mathrm{p}} = \frac{E_{0,\mathrm{grav}}}{c^2}\frac{1}{1 - \delta_{\mathrm{e.m.}}}. \tag{14.15}$$

Equation (14.15) allows one to evaluate the inertial mass of p from the knowledge of the electromagnetic LLI breaking parameter $\delta_{\mathrm{e.m.}}$ and of the threshold energy $E_{0,\mathrm{grav}}$ of the gravitational metric.

The lowest limit to the LLI breaking factor of electromagnetic interaction has been determined by the experiment (based on the detection of a DC voltage across a conductor induced by the steady magnetic field of a coil), discussed in Chap. 13. On account of the definition (13.14) of the isotropic LLI breaking parameter, and the definition (14.1), the value found corresponds to

$$1 - \delta_{\mathrm{e.m.}} \cong 4 \cdot 10^{-11}. \tag{14.16}$$

Then, inserting the value (8.24) for $E_{0,\mathrm{grav}}$ and (14.16) in (14.15), one gets

$$m_{\mathrm{in},p} = \frac{E_{0,\mathrm{grav}}}{c^2}\frac{1}{1 - \delta_{\mathrm{e.m.}}} \geq \frac{2 \cdot 10^{-5}}{4 \cdot 10^{-11}}\frac{\mathrm{eV}}{c^2} = 0.5\frac{\mathrm{MeV}}{c^2} = m_{\mathrm{in},e} \tag{14.17}$$

(with $m_{\mathrm{in,e}}$ being the electron mass) where the $\geq$ is due to the fact that in general the LLI breaking factor constitutes an *upper limit* (*i.e.* it sets the scale *under which* a violation of LLI is expected). If the coil experiment does indeed provide evidence for a LLI breakdown (as it seems the case, although further confirmation is needed), Eq. (14.17) yields $m_{\mathrm{in},p} = m_{\mathrm{in,e}}$. We find therefore the amazing result that *the fundamental particle p is nothing but the electron e^- (or its antiparticle e^+).*[b] The electron is indeed the lightest massive lepton (pointlike, non-composite particle) with electric charge, and therefore subjected to gravitational, electromagnetic and weak interactions, but unable to weakly decay due to its small mass. Consequently, e^- (e^+) shares all the properties we required for the particle p, whereby it plays a fundamental role for gravitational and electromagnetic interactions.

---

[b] Of course, this last statement does strictly holds only if the CPT theorem mantains its validity in the DSR framework, too. Although this problem has not yet been addressed in general on a formal basis, it can be stated that it holds true in the case considered, since the energy value $E = E_{0,\mathrm{grav}}$ corresponds to the minkowskian form of both electromagnetic and gravitational metric.

It was therefore shown that within DSR it is possible — on the basis of simple and plausible assumptions — to evaluate the inertial mass of the electron e^- (and therefore of its antiparticle, the positron e^+) by exploiting the expression of the relativistic energy in the deformed Minkowski space $\tilde{M}_4(E)_{E \in R_0^+}$, the explicit form of the phenomenological metric describing the gravitational interaction (in particular its threshold energy), and the LLI breaking parameter for the electromagnetic interaction $\delta_{\text{e.m.}}$. Let us stress that no use has been made of the electric charge e of the electron.

Therefore, the inertial properties of one of the fundamental constituents of matter and of Universe do find a "geometrical" interpretation in the context of DSR, by admitting for local violations of standard Lorentz invariance.

Leonia

...de año en año la ciudad se expande y los basurales deben retroceder más lejos; la importancia de los desperdicios aumenta..., se despliegan en un perímetro cada vez más vasto.

Las ciudades continuas. 1

En Leonia los desperdicios rodean la ciudad. El visitante se encuentra en el centro de una circunferencia de inmundicia que se desarrolla de año en año según la ciudad se expande. Y lo hace en un perímetro cada vez más vasto. Esa expansión y ese perímetro cada vez mayores, plantean una cuestión fundamental. Si la expansión significa un aumento del área ocupada por la ciudad, ¿el aumento de ése área implica un aumento de su perímetro?

Supongamos que Leonia es plana y su basural circular. El área A de un círculo puede escribirse como función de su perímetro P: $A=P/(4\pi)$. Y entonces, $A'(P)=P/(2\pi)=r$. Por tanto, el radio mide el cambio de área con respecto al perímetro. Y si el radio r se aumenta en r_0, su perímetro y área aumentan en $2\pi r_0$ y $\pi r_0^2+2\pi r r_0$ respectivamente. Es decir, que el aumento del perímetro equivale al perímetro del círculo correspondiente al aumento del radio ($2\pi r_0$), mientras que el aumento de área es mayor que el área del círculo correspondiente a ése aumento. Esto se observa en la figura al margen. El perímetro del círculo mayor, de radio $R=r+r_0$, es la suma de los perímetros de los dos círculos interiores, mientras que su área sobrepasa en mucho la de ambos.

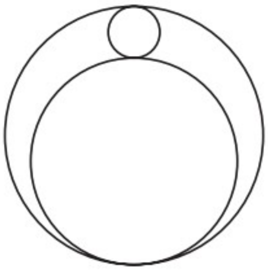

El mundo sobre el que se extiende Leonia no es plano, sino esférico. Si nada limita su expansión, el perímetro y área circulares de Leonia crecerán de la mano, pero sólo hasta que el perímetro iguale el del ecuador terrestre. A partir de ahí el perímetro decrecerá mientras la expansión de la ciudad continúa. Un fenómeno ya señalado por Frabetti (1999) en la *ciudad incontenible*, cuyas murallas crecían hasta un punto en el que con las piedras de la muralla precedente podía construirse un muro que encerraba un área todavía mayor. De ahí en adelante, el perímetro de la ciudad encierra su basural en el polo opuesto al centro de aquella.

Los desperdicios de Leonia ... invadirían el mundo si no estuvieran presionando, más allá de la última cresta, basurales de otras ciudades que también rechazan lejos de sí montañas de deshechos.

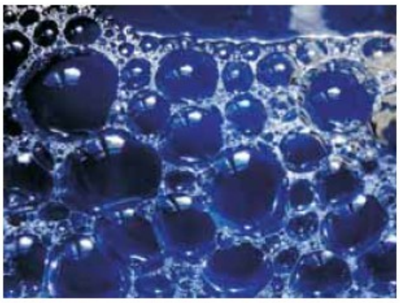

Leonia : *la ciudad que crece como la espuma*

Pero los basurales de Leonia están limitados por la presión de los de otras ciudades como las pompas achatadas de la espuma. ■

suma+ 58
Junio 2008, pp.101-105

Gauss y el polígono regular de 17 lados

Cuando se finaliza un noble edificio no deben quedar visibles los andamios

C.F. Gauss

El sorprendente encuentro entre la Aritmética, el Álgebra y la Geometría.

El 18 de febrero de 1792, antes de cumplir los 15 años el joven Carl Friedrich Gauss hace su inscripción en el Collegium Carolinum de Brunswick. En este colegio da clases de matemáticas y ciencias naturales E. A. W. Von Zimmermann (1743-1815) su valedor ante el duque de Brunswick.

Gauss permanecerá en él hasta 1795, estudiando lenguas clásicas, literatura, filosofía y, por supuesto, matemáticas superiores, siendo un alumno brillante en todas ellas. Entre sus lecturas de matemáticas de esta época están los *Principia Mathematica* de Newton, el *Ars Conjectandi* de Jackob Bernoulli y algunas de las memorias de Euler. En el Collegium Carolinum Gauss iniciará alguna de sus futuras investigaciones matemáticas, según sus propias confesiones posteriores, como la distribución de los números primos o los fundamentos de la geometría.

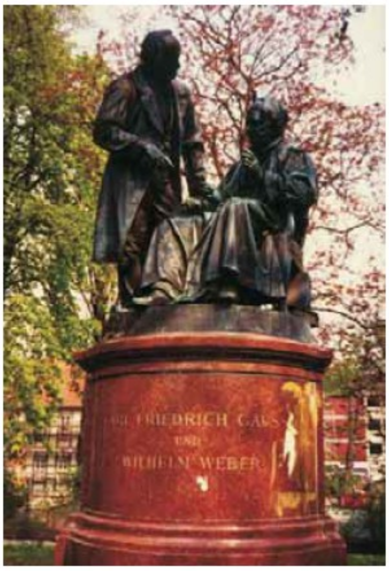

Antonio Pérez Sanz
IES Salvador Dalí (Madrid)
decabeza@revistasuma.es

Cuando en el otoño de 1795 se traslada a la Universidad Georgia Augusta de Göttingen, con una beca del Duque, Gauss aún no ha decidido su futuro académico dudando entre los estudios de Filología clásica y las Matemáticas. Las lecciones de matemáticas, no muy buenas según la opinión de Gauss, las impartía el anciano profesor Gotthelf Abraham Kästner que tenía entonces 76 años. En esta época conoce a Wolfgang (Farkas) Bolyai, que se incorporó a la universidad un año después que él. Gauss, unos años más tarde llegó a afirmar: *Bolyai fue el único que supo interpretar mis criterios metafísicos sobre las Matemáticas.* Y también que Bolyai fue el *espíritu más complicado que jamás conocí*.

Bolyai es más explícito al hablar de su amistad:

> Nos unía la pasión por las Matemáticas y nuestra conciencia moral, y así paseábamos durante largas horas en silencio, cada uno ocupado en sus propios pensamientos.

> *Justo un mes antes de cumplir los 19 años, Gauss se decantará definitivamente por las matemáticas y hará su primera anotación en su diario de notas, un pequeño cuaderno de 19 páginas, que acompañará a Gauss hasta 1814, el diario científico más importante de la historia de las matemáticas*

Construcción con regla y compás del polígono regular de 17 lados

Desde su llegada a Göttingen el joven Gauss siguió desarrollando de forma autónoma sus investigaciones sobre números que había iniciado en el Collegium. Sin duda, más fruto de estas investigaciones que de las enseñanzas de Kästner, cuando Gauss estaba en su casa de Brunswick, se va a producir un descubrimiento que será clave, no sólo en la carrera de Gauss, sino en el futuro de las matemáticas: **el heptadecágono, el polígono regular de 17 lados se puede construir con regla y compás.**

Él mismo, muchos años más tarde, recordará el momento en una carta que dirige a Gerling fechada el 6 de enero de 1819:

> Fue el día 29 de marzo de 1796, durante unas vacaciones en Brunswick, y la casualidad no tuvo la menor participación en ello ya que fue fruto de esforzadas meditaciones; en la mañana del citado día, antes de levantarme de la cama, tuve la suerte de ver con la mayor claridad toda esta correlación, de forma que en el mismo sitio e inmediatamente apliqué al heptadecágono la correspondiente confirmación numérica.

El día siguiente, el 30 de marzo, justo un mes antes de cumplir los 19 años, Gauss se decantará definitivamente por las matemáticas y hará su primera anotación en su diario de notas, un

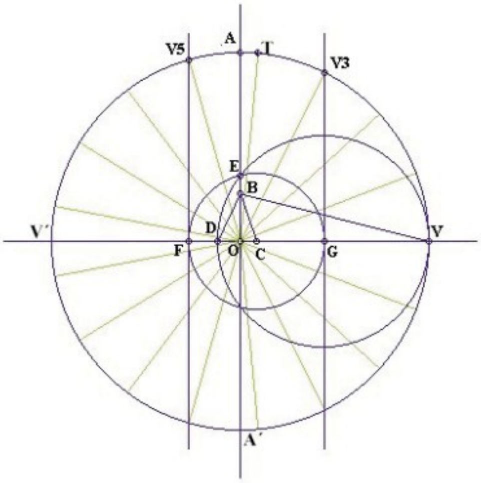

Construcción del polígono regular de 17 lados

Método de Gauss(1796), simplificada por H.W. Richmond (1893)

1. Se construye la circunferencia con centro en O. Se dibujan los diámetros perpendiculares AA´ y VV´
2. Se obtiene un punto B, sobre el radio OA, tal que el segmento OB es la cuarta parte de OA
3. Se obtiene el punto C, sobre OV, tal que el ángulo OBC es la cuarta parte del ángulo OBV (hay que bisectar dos veces un ángulo)
4. Se obtiene un punto D, sobre el diámetro VV´, tal que el ángulo DBV sea de 45° (se puede hacer bisectando un ángulo recto)
5. Se obtiene G, mitad del segmento DV, se dibuja la circunferencia con centro G y radio GV. Esta circunferencia corta al radio OA en el punto E.
6. Se dibuja la circunferencia con centro C y radio CE, dicha circunferencia corta a VV´ en dos puntos: F y G
7. Se levantan perpendiculares a VV´, pasando por F y G , que cortan a la circunferencia en V3 y V5.
8. La mitad del arco V3V5, nos da un punto T. El segmento V 3T es el lado del polígono regular de 17 lados.

pequeño cuaderno de 19 páginas, que acompañará a Gauss hasta 1814, el diario científico más importante de la historia de las matemáticas, en el que irá anotando, a veces de forma críptica, los resultados matemáticos que le vienen a la cabeza, en total 144 anotaciones. Por este diario desfilará un alto porcentaje de los descubrimientos matemáticos del siglo XIX. En este libro no fueron recogidos todos los descubrimientos de Gauss en el período prolífico de 1796 a 1814. Pero muchos de los anotados bastarían para establecer la prioridad de Gauss en campos, donde algunos de sus contemporáneos se niegan a creer que Gauss les precediera.

Muchos hallazgos que quedaron enterrados durante décadas en este diario habrían encumbrado a media docena de grandes matemáticos de haber sido publicados. Algunos jamás se hicieron públicos durante la vida de Gauss, y nunca pretendió la prioridad cuando otros autores se le anticiparon. Sus anotaciones constituían descubrimientos esenciales de la Matemática del siglo XIX. Un documento que por desgracia para la ciencia no verá la luz hasta casi 50 años después de la muerte de Gauss.

> Principia quibus innititur sectio circuli, ac divisibilitas eiusdem geometrica in septemdecim partes, etc. Mart. 30 Brunsv.

Con tan sólo 18 años, el joven Gauss había hecho un descubrimiento que por sí solo le habría hecho pasar a la historia de las matemáticas. Un descubrimiento que constituía sólo la punta del iceberg de una teoría mucho más amplia que dará origen tres años más tarde a las *Disquisitiones Arithmeticae,* obra que Gauss va madurando durante su estancia en la universidad de Gottingën.

En esa misma carta a Gerling, Gauss relata en primera persona cómo llegó a este descubrimiento:

> Por una reflexión muy profunda sobre la relación existente entre el conjunto de raíces de la ecuación
>
> $$\frac{x^p-1}{x-1}=0$$
>
> por razones aritméticas, he llegado a vislumbrar muy claramente la naturaleza de esta relación, tras unos días de vacaciones en Brunswick, la mañana de este día, justo antes de levantarme, de manera que he podido hacer sobre el terreno la aplicación particular al polígono regular de 17 lados y ejecutar los cálculos numéricos correspondientes.

Gauss tiene ya en su cabeza una visión clara del modelo geométrico de representación de los números complejos y de su potencial revolucionario para atacar ecuaciones ciclotómicas. Y justo en un momento de la historia en que la colectividad matemática tiene serios reparos a la utilización alegre de los números imaginarios que Euler ha realizado cincuenta años antes al proponer su revolucionaria fórmula (Prop. 138), que relaciona las cantidades imaginarias con la cantidades trascendentes nacidas del círculo, que así se llama el capítulo correspondiente de la *Introductio in Analisin infinitorum* (1748)

$$e^{+v\sqrt{-1}}=\cos v+\sqrt{-1}\,sen\,v$$

Que tras la introducción de la letra i como notación de la unidad imaginaria, será una de las fórmulas más populares de las matemáticas: $e^{iv}=\cos v+i\,sen\,v$, sobre todo para el valor particular de $v=\pi$

$$e^{i\pi}=\cos\pi+i\,sen\,\pi=-1$$

Porque en la cabeza del joven Gauss las raíces de la ecuación, de la forma

$$e^{i\frac{2k\pi}{p}}$$

se perfilan claramente como los vértices de un polígono regular de p lados.

La Aritmética modular viene en ayuda del Álgebra

Gauss se ocupa del caso especial p = 17.

Si una de las raíces es $r=e^{i\frac{2\pi}{17}}$, las otras raíces son las potencias sucesivas de r, es decir r^m.

Gauss crea una tabla con las 17 raíces ordenadas según las potencias de 3 expresadas en módulo 17.

k	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
m=3k(mod 17)	1	3	9	10	13	5	15	11	16	14	8	7	4	12	2	6	1

Los valores de m forman una permutación de los números enteros del 1 al 16.

¿Por qué ordena Gauss las raíces de esta forma un tanto arbitraria? Sencillamente, esto le permite comprobar las funciones de las raíces que permanecen invariantes para un cierto número de permutaciones.

Los antecedentes algebraicos

La idea de relacionar las funciones de las raíces que permanecen invariantes o que toman un número pequeño de valores al permutar las raíces, con las soluciones de la ecuación, la desarrolla ampliamente Lagrange en una memoria de 1771.

La idea parte del hecho de un resultado explicitado por Vandermonde: todos los radicales que intervienen en las fórmulas de resolución de las ecuaciones de grado menor que 4 se pueden expresar como función de las raíces.

En la ecuación de segundo grado $ax^2 + bx + c = 0$, el discriminante

$$\Delta = \sqrt{b^2 - 4ac} = a\sqrt{(x_1 - x_2)^2}$$

Es una función simétrica de las raíces.

En la ecuación de tercer grado $ax^3 + b\,x^2 + c\,x + d = 0$, la función $T = (x_1 + jx_2 + j^2x_3)^3$ sólo toma dos valores distintos al hacer:

$$j = e^{i\frac{2\pi}{3}}$$

$$T_1 = (x_1 + jx_2 + j^2x_3)^3 \text{ y } T_2 = (x_1 + j^2x_2 + jx_3)^3$$

Por lo tanto, las funciones $T_1 + T_2$, y $T_1 \bullet T_2$ son simétricas respecto de las raíces de la ecuación y por ello calculables utilizando sólo los coeficientes de la ecuación.

Las tres raíces se pueden calcular resolviendo el sistema

$$\begin{cases} x_1 + x_2 + x_3 = -b \\ x_1 + jx_2 + j^2x_3 = \sqrt[3]{T_1} \\ x_1 + j^2x_2 + jx_3 = \sqrt[3]{T_2} \end{cases}$$

La existencia de funciones T_i de las raíces, que toman menos valores distintos que el grado de la ecuación, permiten encontrar las soluciones mediante ecuaciones auxiliares construidas a partir de funciones simétricas de las T_i . A estas ecuaciones se las conoce como *ecuaciones resolventes*.

La genialidad de Gauss

La tabla de Gauss le permite construir ecuaciones resolventes agrupando las raíces de 8 en 8, de 4 en 4 o de 2 en 2.

Gauss comienza con dos periodos de 8 raíces

$$p_1 = \omega_1 + \omega_9 + \omega_{13} + \omega_{15} + \omega_{16} + \omega_8 + \omega_4 + \omega_2$$

$$p_2 = \omega_3 + \omega_{10} + \omega_5 + \omega_{11} + \omega_{14} + \omega_7 + \omega_{12} + \omega_6$$

Tanto p_1 como p_2 contienen al mismo tiempo a ω_n y a su inversa ω_{17-n} .

La suma de $p_1 + p_2 = -1$, ya que la suma de las 17 raíces de $x^{17} - 1 = 0$, las 16 de p_1 y p_2 más la raíz 1, ha de ser 0

Y como $\omega_a \cdot \omega_b = \omega_c \quad \textit{donde} \quad c \equiv a + b \pmod{17}$

El producto de $p_1 \bullet p_2 = -4$

De modo que p_1 y p_2 son las raíces de la ecuación:

$$x^2 + x - 4 = 0$$

A continuación Gauss agrupa las raíces de p_1 y p_2 en cuatro periodos de 4 raíces

$$\begin{cases} q_1 = \omega_1 + \omega_{13} + \omega_{16} + \omega_4 \\ q_2 = \omega_9 + \omega_{15} + \omega_8 + \omega_2 \end{cases} \text{ y } \begin{cases} q_3 = \omega_3 + \omega_5 + \omega_{14} + \omega_{12} \\ q_4 = \omega_{10} + \omega_{11} + \omega_7 + \omega_6 \end{cases}$$

Podemos comprobar que
$q_1 + q_2 = p_1$; $\qquad q_1 \bullet q_2 = p_1 + p_2 = -1$

Y por tanto q_1 y q_2 son las raíces de la ecuación:

$$x^2 - p_1\,x - 1 = 0$$

Análogamente

$q_3 + q_4 = p_2$; $\qquad q_3 \bullet q_4 = p_1 + p_2 = -1$

Y por tanto q_3 y q_4 son las raíces de la ecuación:

$$x^2 - p_2\,x - 1 = 0$$

Es fácil comprobar que $\omega_a + \omega_{17-a} = 2\cos\frac{2a\pi}{17}$

Así que $q_1 = 2\cos\frac{2\pi}{17} + 2\cos\frac{8\pi}{17}$ y $q_2 = 2\cos\frac{4\pi}{17} + 2\cos\frac{16\pi}{17}$

Con la ecuación $x^2 - p_2\,x - 1 = 0$

Obtenemos que:

$$q_3 = 2\cos\frac{6\pi}{17} + 2\cos\frac{10\pi}{17} \text{ y } q_4 = 2\cos\frac{12\pi}{17} + 2\cos\frac{14\pi}{17}$$

Por fin, formando periodos de 2 en 2 tendremos

$$r_1 = \omega_1 + \omega_{16} \quad ; \quad r_2 = \omega_{13} + \omega_4$$

Y por tanto:

$$\begin{cases} r_1 + r_2 = q_1 \\ r_1 \cdot r_2 = q_3 \end{cases}$$

$$\text{Así: } \quad r_1 = 2\cos\frac{2\pi}{17} \quad \text{y} \quad r_2 = 2\cos\frac{8\pi}{17}$$

Y reemplazando en los valores de p_i tendremos... es sólo un problema de paciencia...

$$p_1 = \frac{\sqrt{17} - 1}{2} \quad ; \quad p_2 = -\frac{\sqrt{17} + 1}{2}$$

Y q_i:

$$q_1=\frac{p_1+\sqrt{p_1^2+4}}{2}=\frac{\sqrt{17}-1+\sqrt{34-2\sqrt{17}}}{4}\quad;$$
$$q_2=\frac{p_1-\sqrt{p_1^2+4}}{2}=\frac{\sqrt{17}-1-\sqrt{34-2\sqrt{17}}}{4}$$

$$q_3=\frac{p_2+\sqrt{p_2^2+4}}{2}=\frac{-\sqrt{17}-1+\sqrt{34+2\sqrt{17}}}{4}\quad;$$
$$q_4=\frac{p_2-\sqrt{p_2^2+4}}{2}=\frac{-\sqrt{17}-1-\sqrt{34+2\sqrt{17}}}{4}$$

y por fin r_i:

$$\frac{r_1}{2}=\cos\frac{2\grave{A}}{17}=-\frac{1}{16}+\frac{1}{16}\sqrt{17}+\frac{1}{16}\sqrt{34-2\sqrt{17}}+$$
$$+\frac{1}{8}\sqrt{17+3\sqrt{17}-\sqrt{34-2\sqrt{17}}-2\sqrt{34+2\sqrt{17}}}$$

Sustituyendo en las sucesivas ecuaciones podemos calcular las 16 raíces de la ecuación.

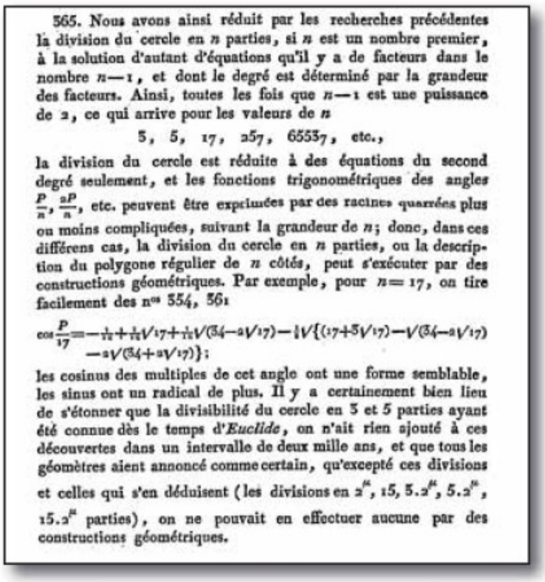

365. Nous avons ainsi réduit par les recherches précédentes la division du cercle en n parties, si n est un nombre premier, à la solution d'autant d'équations qu'il y a de facteurs dans le nombre $n-1$, et dont le degré est déterminé par la grandeur des facteurs. Ainsi, toutes les fois que $n-1$ est une puissance de 2, ce qui arrive pour les valeurs de n

3, 5, 17, 257, 65537, etc.,

la division du cercle est réduite à des équations du second degré seulement, et les fonctions trigonométriques des angles $\frac{P}{n}$, $\frac{2P}{n}$, etc. peuvent être exprimées par des racines quarrées plus ou moins compliquées, suivant la grandeur de n; donc, dans ces différens cas, la division du cercle en n parties, ou la description du polygone régulier de n côtés, peut s'exécuter par des constructions géométriques. Par exemple, pour $n=17$, on tire facilement des n^os 354, 361

$$\cos\frac{P}{17}=-\tfrac{1}{16}+\tfrac{1}{16}\sqrt{17}+\tfrac{1}{16}\sqrt{(34-2\sqrt{17})}-\tfrac{1}{8}\sqrt{\{(17+3\sqrt{17})-\sqrt{(34-2\sqrt{17})}-2\sqrt{(34+2\sqrt{17})}\}};$$

les cosinus des multiples de cet angle ont une forme semblable, les sinus ont un radical de plus. Il y a certainement bien lieu de s'étonner que la divisibilité du cercle en 3 et 5 parties ayant été connue dès le temps d'*Euclide*, on n'ait rien ajouté à ces découvertes dans un intervalle de deux mille ans, et que tous les géomètres aient annoncé comme certain, qu'excepté ces divisions et celles qui s'en déduisent (les divisions en 2^μ, 15, 5.2^μ, 5.2^μ, 15.2^μ parties), on ne pouvait en effectuer aucune par des constructions géométriques.

Disquisitiones Arithmeticae. Ed. Francesa de 1809

Del Álgebra a la Geometría

Al final sólo hemos tenido que resolver de forma sucesiva ecuaciones de 2º grado, es decir todo el proceso se puede realizar geométricamente con regla y compás. Esto ha sido así ya que para n = 17 tenemos n – 1 = 16 = 2 • 2 • 2 • 2

La alegría del joven Gauss, ese 30 de marzo de 1796 aún tiene 18 años, está más que justificada. Acaba de resolver un problema de más 2 milenios de antigüedad.

Este hecho, por sí sólo, le haría pasar a la Historia de las Matemáticas. Definitivamente las lenguas clásicas habían perdido un genio para siempre. Las matemáticas y todas las ciencias habían ganado la batalla en la cabeza de Gauss.

Gauss terminará sus *Disquisitiones Arithmeticae*, con las proposiciones 365 y 366, tantas como los días del año, rebosante de alegría, facilitando la lista de los polígonos de menos de 300 lados que se pueden construir con regla y compás. El joven genio ha tocado la gloria y lo sabe.

> Para que la división geométrica del círculo en N partes sea posible, N debe ser 2, o una potencia de 2, o bien un número primo de la forma 2m +1 o bien le producto de una potencia de 2 por uno o varios números primos diferentes de esta forma

Y todo ello antes de cumplir los 19 años... Cuesta creer que todo lo hiciera DE CABEZA...

DE CABEZA ■

REFERENCIAS BIBLIOGÁFICAS

BÜHLER, W. K. (1981): *Gauss A biographical Study,* Springer-Verlag, New York.

DUNNINGTON, G. W., GAUSS, C. F. (1955): *Titan of Science,* New York.

FRIEDELMEYER, J-P. (1994): "Recherche inconnue désespérément", en *Histoires de problèmes. Histoire des mathématiques,* Ellipses, I.R.E.M. París.

GAUSS, C. F., (1973): *Werke.* Georg Olms, Hildesheim.

GAUSS, C. F., (1996): *Disquisicions aritmètiques.* Traducción de la profesora Pascual Xufrí G., Sociedad Catalana de Matemáticas. Barcelona.

PARDO REGO, V. (2003): *Lagrange. La elegancia matemática,* Nivola, Madrid.

RASSIAS G. M., (1991): *The mathematical heritage of C F Gauss,* Singapore.

REICH, K. (1977): *Gauss. 1777/1977.* Inter-Nationes. Bonn-Bad Gedessberg.

Vídeos

PÉREZ SANZ, A. (2000): *Gauss. De lo real a lo imaginario.* Serie Universo Matemático. RTVE.

Internet

http://www.geocities.com/RainForest/Vines/2977/gauss/formulae/heptadecagon.html
http://divulgamat.ehu.es/weborriak/Historia/MateOspetsuak/Gauss.asp

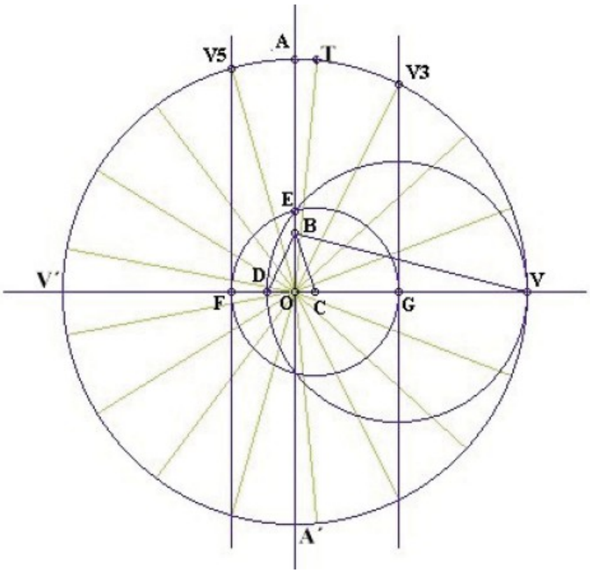

B 9(:69 9 92 9 , < ; : * ; 3

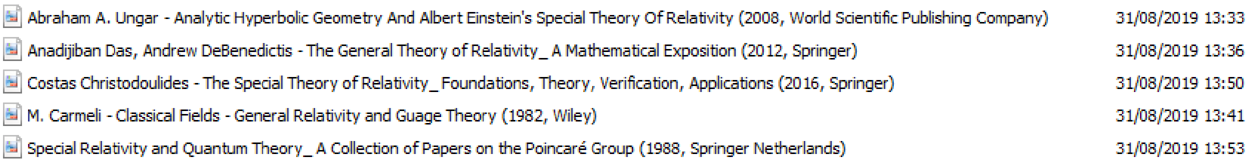

Abraham A. Ungar - Analytic Hyperbolic Geometry And Albert Einstein's Special Theory Of Relativity (2008, World Scientific Publishing Company) 31/08/2019 13:33
Anadijiban Das, Andrew DeBenedictis - The General Theory of Relativity_ A Mathematical Exposition (2012, Springer) 31/08/2019 13:36
Costas Christodoulides - The Special Theory of Relativity_ Foundations, Theory, Verification, Applications (2016, Springer) 31/08/2019 13:50
M. Carmeli - Classical Fields - General Relativity and Guage Theory (1982, Wiley) 31/08/2019 13:41
Special Relativity and Quantum Theory_ A Collection of Papers on the Poincaré Group (1988, Springer Netherlands) 31/08/2019 13:53

B 9(:69 9 92 9 , (. (N (;
P;3 . ? ; (! (. A | 3 < ; 2 : * ; $@ C !
. 3 ; 3!

B 9(:69 9 92 9 , (! 4 . > @ (!
> 8 ! < ? < ; : * ;- (AH3 @ 3 3!

B" 9(:69 9 92 9 , ! ! < 3
< ; : * ;- E ! @ < ;@ @ (33 %@ 3 3!

B/ 9(:69 9 92 9 , E !
? : * ; ! ? < ; ,$ @ C ; 3!

B 9(:69 9 92 9 , 3 : * ; ! 6
< ;- (3 v ? 3 ,$$@ 3 R ! 3!

v 2 * 777

9(:69 9 92 9 , * ! 3 3 3!

9(:69 9 92 9 , * ! 3 3 3!

" 9(:69 9 92 9 , * ! 3 "3 3!

/ 9(:69 9 92 9 , * ! 3 /3 3!

9(:69 9 92 9 , * ! 3 3 3!

% 9(:69 9 92 9 , * ! 3 %3 3!

0 9(:69 9 92 9 , * ! 3 03 3!

$ 9(:69 9 92 9 , * ! 3 $3 3!

, 9(:69 9 92 9 , * ! 3 ,3 3!

9(:69 9 92 9 , * ! 3 3 3!

<v /# # "% ' 0 ! !

& , # # 0 ,

N3!

. 9(:69 9 92 9 , P ! !| ; (
! | 3 (;@ A! > * 3 3!

' /, % AH $

AH $ & 3 3
3 @ ! !
3 3

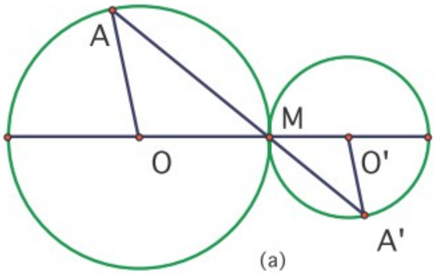

(a)

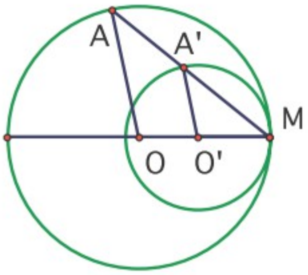

Solution. If line AA' passes through the point M of tangency of the circles O and O' (Figures t58a and 58b), then $\widehat{OAM} = \widehat{OMA} = \widehat{O'MA'} = \widehat{O'A'M}$, so that OA is parallel to $O'A'$.

Notes. This is another good problem for students to cut their teeth on. If they have solved Exercise 57, they may be able to recognize the need to consider both internally and externally tangent circles.

. H • ' / ! * ! @ r ! * }
H 3 3 ! .v 33 | < @ . •! ! . H

9(:69 9 92 9 , (; 4 * ? ; /@
A 3 ; 3!

. 9(:69 9 92 9 , ? 3 -
! @ ! 3 * ;@ 3 , $@ ? P 3!

9(:69 9 92 9 , F * @ !
4 * - ((. ? ; 3!

! 9(:69 9 92 9 , 3 * 4 * ? ;-
(? ! ! < < : ! 3 H ? ; @ 3 8
P ! . 3! 777777777777

2. A * v 3 ' *') . 3) !
! @ 3 3) ! * . !

9(:69 9 92 9 , 4 * ? ; &
0@ !4*

9(:69 9 92 9 , 4 * ? ; &&
0@ 8 B !4*

: ,

9(:69 9 92 9 , * ! "3 3!

' $/ , ! , "

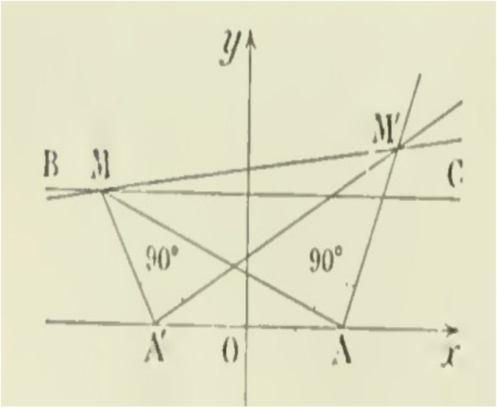

ou, en remarquant que $xdx + ydy + zdz = rdr$,

$$Xdx + Ydy + Zdz = -mKdr - mgdz.$$

Il en résulte qu'il existe une fonction des forces U, définie par

$$U = -m(Kr + gz) = -m(K\sqrt{x^2+y^2+z^2} + gz),$$

et par suite l'équation générale des surfaces de niveau est

$$K\sqrt{x^2+y^2+z^2} + gz = h,$$

ou

$$x^2+y^2+z^2 = \frac{1}{K^2}(gz-h)^2,$$

ou

$$x^2+y^2+z^2 = \frac{g^2}{K^2}\left(z - \frac{h}{g}\right)^2.$$

Cette équation représente le lieu des points dont le rapport des distances à l'origine et au plan (P), $z - \frac{h}{g} = 0$, est constant et égal à $\frac{g}{K}$.

Ce lieu est une quadrique de révolution autour de Oz, dont la méridienne dans le plan des zx a pour foyer l'origine, pour directrice la droite (D), $z - \frac{h}{g} = 0$, et pour excentricité $\frac{g}{K}$. Cette conique a pour axe focal Oz.

Donc si $g < K$, la conique est une ellipse, et les surfaces de niveau sont des ellipsoïdes de révolution allongés ;

si $g > K$, ce sont des hyperboloïdes de révolution à deux nappes ;

si $g = K$, ce sont des paraboloïdes de révolution.

Cas particulier. — Si $h = 0$, la conique se réduit à deux droites, et la surface à un cône de révolution. Ce cône n'est réel que si $g > K$.

André Courtois, à Avesnes-sur-Helpe (Nord).

Bonnes solutions par MM. J. Dagnaux, à Dijon ; G. Lach, à Douai ; A. Rousseau, à Fournes-en Weppes ; Marcel Ullmann, 47e d'artillerie, à Héricourt.

◆

CONCOURS DE 1913

ÉCOLE CENTRALE

Géométrie analytique.

2177. — Soient deux axes de coordonnées rectangulaires Ox et Oy ; sur Ox, un point A d'abscisse positive donnée a et un point A' d'abscisse négative $-a$. On considère la droite BC dont l'équation est $y - a = 0$, et un point M d'abscisse λ sur cette droite. On mène la droite MA et, en A, la perpendiculaire AM' à AM ; puis la droite MA' et, en A', la perpendiculaire A'M' à MA'. On demande de résoudre, dans l'ordre que l'on voudra, les deux questions suivantes :

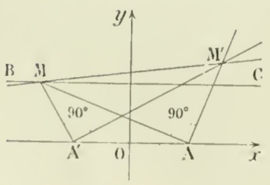

1° Lorsque λ varie, trouver et construire le lieu du point M' ;

2° Exprimer en fonction de λ les coordonnées du point de contact de la droite MM' avec son enveloppe, et construire cette enveloppe en faisant varier λ.

Mécanique.

2178. — Soient trois axes de coordonnées rectangulaires Ox, Oy, Oz, dont l'un Oz est vertical et dirigé positivement en sens contraire de la pesanteur. On donne, sur Ox, le point A d'abscisse positive a ; sur Oy,

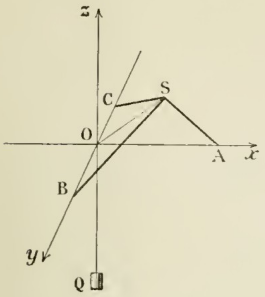

le point B d'ordonnée positive $\frac{a}{2}$ et le point C d'ordonnée négative $-\frac{a}{3}$. Soit, en outre, un point S de coordonnées $x = \frac{a}{2}$, $y = 0$, $z = \frac{a}{2}$. On considère trois barres SA, SB, SC, rectilignes, homogènes, de poids respectifs connus P, P′, P″ et dont on néglige les épaisseurs. Elles sont soudées en S et forment un trièdre de figure invariable qui s'appuie sur le plan xOy, supposé non-déformable, aux points A, B, C.— Au point S est attaché un cordon de poids négligeable qui passe en O, sans frottement, sur une petite poulie fixe que l'on supposera réduite à ce point O. Ce fil est tendu par un poids Q. Il n'y a pas de frottement en B et en C, mais il y a en A un frottement de coefficient $f = 1$.

Trouver en fonction de P, P′, P″, la plus grande valeur de Q pour laquelle le trièdre peut rester en équilibre dans la position donnée et, pour cette valeur de Q, calculer les réactions du plan xOy en B et en C.

Effectuer les calculs, à un gramme près, sachant que $a = 60^{cm}$ et qu'un centimètre de longueur de chaque barre pèse 10^g.

(11 juin, de 7 h. à 11 h.)

Trigonométrie.

2179. — On considère un triangle ABC rectangle en A dont l'hypoténuse BC est égale à une longueur donnée a, et on prend sur la bissectrice intérieure de l'angle ABC une longueur BD égale à une longueur donnée l; on désigne par x la moitié de l'angle ABC.

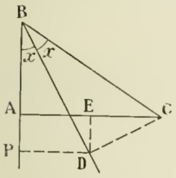

1° Étudier la variation de la distance DE du point D au côté AC quand on fait varier x de 0 à $\frac{\pi}{4}$.

2° Déterminer l'angle x de façon que le rapport $\frac{DC}{DP}$ des distances du point D au point C et au côté AB soit égal à un nombre donné K. Discuter.

Nota. — On pourra traiter ces deux parties dans l'ordre que l'on voudra.

(13 juin, de 3 h. à 5 h.)

Calcul.

2180. — On donne

$$a = 223,$$
$$b = 283,$$
$$c = 324,$$
$$m = 830.$$

Calculer au moyen de la règle à calcul :

1° Les nombres r, u, v, w définis par

$$r = \sqrt{\frac{abc}{m}}, \qquad u = \sqrt{\frac{mbc}{a}}, \qquad v = \sqrt{\frac{mca}{b}}, \qquad w = \sqrt{\frac{mab}{c}};$$

les cubes r^3, u^3, v^3, w^3 et les inverses $\frac{1}{r}$, $\frac{1}{u}$, $\frac{1}{v}$, $\frac{1}{w}$, de ces quatre nombres ;

2° Les plus petits angles positifs x, y, z tels que $\operatorname{tg} x = \frac{r}{a}$, $\operatorname{tg} y = \frac{r}{b}$, $\operatorname{tg} z = \frac{r}{c}$;

3° Les sinus et cosinus de ces trois angles x, y, z.

Avis : La composition de calcul est faite avec une règle à calcul de 26^{cm} environ. Défense d'apporter en même temps une table de logarithmes.

(14 juin, de 3 h. à 4 h.)

Épure.

2181. — Cadre 27^{cm} sur 45^{cm}. Ligne de terre xy (en trait noir fin) petit axe ; ω centre. Les plans de projection existent et sont opaques. Titre extérieur : Ombres. — Titre intérieur : Ellipsoïde et Paraboloïde.

' ,0 u) ! € ! € 00 ! ,

23e Année. Septembre 1913. N° 12.

REVUE DE MATHÉMATIQUES SPÉCIALES

PREMIÈRE PARTIE

AGRÉGATION DES SCIENCES MATHÉMATIQUES

Concours de 1912.

Mathématiques spéciales.

2077. — 1° *Si, dans une équation du troisième degré*

$$(1) \qquad z^3 - p_1 z^2 + p_2 z - p_3 = 0,$$

le coefficient p_1 *est nul, la fonction* $z_1^2 - z_2 z_3 = z_2^2 + z_2 z_3 + z_3^2$, *non symétrique par rapport aux lettres, est équivalente à une fonction symétrique des racines de l'équation. Le cas où* $p_1 = 0$ *est-il le seul où il existe un polynome homogène du second degré à trois variables qui jouisse de la même propriété?*

Le coefficient p_1 *n'étant pas nul, former l'équation* (2) *qui a pour racines les valeurs que prend la fonction* $x^2 + xy + y^2$ *quand on y remplace* x *et* y *par deux racines quelconques de l'équation* (1). *L'équation* (2) *peut-elle être équivalente à l'équation* (1) *dont on l'a déduite?*

2° *Les médianes d'un triangle* ABC *se coupant en un point* O, *on fait tourner ce triangle de* 60° *autour de* O *dans un sens et dans l'autre; on obtient ainsi deux nouvelles positions* A'B'C', A''B''C'' *du triangle* (A' *et* A'' *étant les nouvelles positions du point* A). *Soient* A'_1, A''_1 *les milieux de* BA', BA''; B'_1, B''_1 *les milieux de* CB', CB''; C'_1, C''_1, *les milieux de* AC', AC''; A'_2, A''_2, *les milieux de* CA', CA''; B'_2, B''_2, *les milieux de* AB', AB''; C'_2, C''_2, *les milieux de* BC', BC''. *Montrer que les angles* $A'_1OA''_1$, $A'_2OA''_2$, $B'_1OB''_1$, $B'_2OB''_2$, $C'_1OC''_1$, $C'_2OC''_2$ *ont même bissectrice et qu'il existe sur cette bissectrice deux points* ω', ω'' *symétriques par rapport à* O, *tels que les quadrilatères* $\omega'\omega''A'_1A''_1$, $\omega'\omega''A'_2A''_2$, $\omega'\omega''B'_1B''_1$, ... *soient inscriptibles.*

Démontrer qu'il existe un triangle $A_1B_1C_1$ *homothétique du triangle* ABC *par rapport à* O *et tel que, si l'on prend les inverses* A_2, B_2, C_2 *des sommets* A_1, B_1, C_1, *le pôle d'inversion étant soit* ω', *soit* ω'', *le centre des moyennes distances des points* A_2, B_2, C_2 *est soit* ω', *soit* ω''.

(*On pourra représenter les différents points par leurs affixes, rapportées à un système d'axes rectangulaires ayant* O *pour origine, et utiliser la propriété indiquée au n° 1.*)

3° *On considère les triangles* ABC *variables, tels que les points* ω', ω'' *qui leur correspondent soient fixes et pour lesquels le produit des longueurs des médianes est donné. Trouver le lieu des sommets de ces triangles.* (*On pourra se borner à construire ce lieu dans le cas où les données sont telles que l'un des triangles* ABC *ait deux sommets confondus.*)

Étant donné un point A *du lieu, construire le triangle* ABC *qui lui correspond.*

1. Si p_1 est nul, il est évident que la fonction des racines $u = z_1^2 - z_2 z_3$ est égale à $z_2^2 + z_3^2 + z_2 z_3$; pour le voir, il n'y a qu'à remplacer z_1 par $-(z_2 + z_3)$ dans l'expression de u. On a donc alors

$$2u = z_1^2 - z_2 z_3 + z_2^2 + z_3^2 + z_2 z_3 = \Sigma z_1^2,$$

ce qui montre bien que u est une fonction symétrique des racines; dans le cas actuel, $u = -p_2$.

Considérons maintenant un polynome homogène et du second degré en z_1, z_2, z_3

$$u = a z_1^2 + b z_2^2 + c z_3^2 + 2a' z_2 z_3 + 2b' z_3 z_1 + 2c' z_1 z_2;$$

s'il est équivalent à une fonction symétrique des racines de l'équation du troisième degré, il ne doit pas changer par l'échange de z_1 avec z_2 et nous aurons aussi

$$u = az_2^2 + bz_1^2 + cz_3^2 + 2a'z_1z_3 + 2b'z_2z_3 + 2c'z_1z_2,$$

par suite, $$a(z_1^2 - z_2^2) - b(z_1^2 - z_2^2) - 2a'z_3(z_1 - z_2) + 2b'z_3(z_1 - z_2) = 0.$$

Supposons maintenant que l'équation ne soit pas une équation particulière ayant une racine double; nous pourrons diviser par $z_1 - z_2$ et nous aurons

$$(a - b)(z_1 + z_2) + 2(b' - a')z_3 = 0, \qquad \text{ou} \qquad (a - b)p_1 + [2(b' - a') + b - a]z_3 = 0.$$

Or ici z_3 désigne évidemment l'une quelconque des racines de l'équation du troisième degré; nous devrons donc avoir

$$(a - b)p_1 = 0, \qquad 2(b' - a') + b - a = 0.$$

Si p_1 n'est pas nul, nous aurons $b = a$ et $b' = a'$; de même, $c = a$, $c' = a'$; le polynome u est alors un polynome symétrique.

Si $p_1 = 0$, nous avons aussi

$$2(b' - a') + b - a = 0;$$

nous aurons de même

$$2(c' - b') + c - b = 0 \qquad \text{et} \qquad 2(a' - c') + a - c = 0.$$

Il est facile de voir que ces trois relations se réduisent à deux et que deux des six coefficients se déterminent à l'aide des quatre autres. En tenant compte de l'homogénéité du polynome par rapport aux six coefficients, nous voyons qu'il y a une triple infinité de tels polynomes, non symétriques par rapport aux lettres, mais équivalents à des fonctions symétriques des racines de l'équation du 3^e^ degré. Ceci n'a lieu d'ailleurs que dans le cas où p_1 est nul.

Les polynomes indiqués au début de l'énoncé : $z_1^2 - z_2z_3$, $z_2^2 - z_1z_3$, $z_3^2 - z_1z_2$, $z_1^2 + z_2^2 + z_1z_2$, etc., sont des cas particuliers de ceux que nous venons d'obtenir. Il est facile alors, en faisant $a = 1$ et remplaçant b par $1 + 2(a' - b')$, c par $1 + 2(a' - c')$, de vérifier que tous ces polynomes sont équivalents à des fonctions symétriques.

Appelons maintenant x, y, z les trois racines de l'équation et cherchons celle qui admet pour racines les valeurs de

$$t = x^2 + xy + y^2 = (x + y)^2 - xy.$$

Nous avons $$x + y = p_1 - z, \qquad xy = \frac{p_3}{z},$$

et $$t = (p_1 - z)^2 - \frac{p_3}{z} = \frac{z^3 - 2p_1z^2 + p_1^2z - p_3}{z};$$

puis, comme $z^3 - p_1z^2 + p_2z - p_3 = 0$,

$$t = \frac{-p_1z^2 + p_1^2z - p_2z}{z} = -p_1z + p_1^2 - p_2.$$

Il suffit donc de faire la transformation

$$z = \frac{p_1^2 - p_2 - t}{p_1},$$

ce qui donne l'équation du troisième degré en t

$$(p_1^2 - p_2 - t)^3 - p_1^2(p_1^2 - p_2 - t)^2 + p_1^2p_2(p_1^2 - p_2 - t) - p_1^3p_3 = 0.$$

Cette équation, développée, s'écrit

$$t^3 - (2p_1^2 - 3p_2)t^2 + (p_1^4 - 3p_1^2p_2 + 3p_2^2)t - (p_1^2p_2^2 - p_2^3 - p_1^3p_3) = 0,$$

et il n'y a plus qu'à l'identifier avec l'équation donnée pour obtenir les relations qui doivent exister entre p_1, p_2, p_3 :

$$2p_1^2 - 3p_2 = p_1, \qquad p_1^4 - 3p_1^2p_2 + 3p_2^2 = p_2, \qquad p_1^2p_2^2 - p_2^3 - p_1^3p_3 = p_3.$$

Ces relations nous donnent

$$p_2 = \frac{2p_1^2 - p_1}{3}, \qquad p_3 = \frac{p_1^2p_2^2 - p_2^3}{1 + p_1^3},$$

et enfin $$p_1^4 - p_1^2(2p_1^2 - p_1) + \frac{(2p_1^2 - p_1)^2}{3} - \frac{2p_1^2 - p_1}{3} = 0;$$

cette dernière donne p_1 et les autres donnent ensuite p_2 et p_3.

Développons l'équation en p_1 ; elle s'écrit

$$p_1^4 - p_1^3 - p_1^2 + p_1 = 0,$$

puis

$$p_1(p_1 - 1)^2(p_1 + 1) = 0.$$

Elle donne successivement : $p_1 = 0$, $p_1 = 1$, $p_1 = -1$; mais puisque $p_1 = 0$, il n'y a que deux valeurs acceptables, les valeurs $p_1 = 1$ et $p_1 = -1$.

Pour $p_1 = 1$, nous avons $p_2 = \frac{1}{3}$, $p_3 = \frac{1}{27}$ et l'équation en z est

$$z^3 - z^2 + \frac{z}{3} - \frac{1}{27} = 0 ;$$

la transformation est alors, $z = \frac{2}{3} - t$; deux des racines, z et t, ont pour somme $\frac{2}{3}$, l'autre est égale à $\frac{1}{3}$, le produit zt à $\frac{1}{9}$, les trois racines sont égales à $\frac{1}{3}$. Ceci se vérifie immédiatement en remarquant que le polynome $z^3 - z^2 + \frac{z}{3} - \frac{1}{27}$ est le cube de $z - \frac{1}{3}$.

Pour $p_1 = -1$, nous voyons que p_2 est égal à 1 et que p_3 est indéterminé. L'équation est alors

$$z^3 + z^2 + z - p_3 = 0 ;$$

la transformation est la transformation identique $z = t$, et ceci est facile à expliquer, car nous avons évidemment

$$z_1^3 + z_1^2 + z_1 = z_2^3 + z_2^2 + z_2, \qquad \text{ou} \qquad (z_1 - z_2)(z_1^2 + z_1 z_2 + z_2^2 + z_1 + z_2 + 1) = 0.$$

Ceci nous montre que

$$z_1^2 + z_1 z_2 + z_2^2 + z_1 + z_2 + 1 = 0, \qquad \text{ou} \qquad z_1^2 + z_1 z_2 + z_2^2 = z_3 ;$$

les trois valeurs de la fonction $x^2 + xy + y^2$ sont donc égales aux trois racines de l'équation proposée.

2. Les médianes des triangles ABC se coupant au point O, la somme des affixes de ces trois points, $z_1 + z_2 + z_3$, est nulle et ces trois nombres sont racines d'une équation du troisième degré de la forme $z^3 + p_2 z - p_3 = 0$. Nous savons alors que la quantité $z_1^2 + z_1 z_2 + z_2^2$ est équivalente à une fonction symétrique des racines ; c'est donc un invariant relatif aux sommets du triangle. D'autre part, nous avons

$$I = z_2^2 + z_1 z_2 + z_1^2 = (z_2 - \alpha z_1)(z_2 - \alpha^2 z_1),$$

α et α^2 étant les deux racines cubiques imaginaires de l'unité, par exemple $\alpha = \cos\frac{2\pi}{3} + i \sin\frac{2\pi}{3}$.

Appelons alors z_1' et z_1'' les affixes des points A' et A'' : nous aurons

$$z_1' = z_1\left(\cos\frac{\pi}{3} + i \sin\frac{\pi}{3}\right), \qquad \alpha z_1' = - z_1 ;$$

nous aurons de même $\alpha^2 z_1'' = - z_1$: par suite, $z_1 = - \alpha^2 z_1$, $z_1'' = - \alpha z_1$.

L'invariant I est donc égal à $(z_2 + z_1')(z_2 + z_1'')$.

Soient alors z' et z'' les affixes des milieux de BA′ et BA″ ; nous voyons que

$$I = 4z'z'', \qquad z'z'' = \frac{I}{4},$$

et qu'en posant $z^2 = \frac{I}{4} = z'z''$, nous définirons un couple de points ω' et ω'', qui sera le même pour tous les couples de milieux, tels que A_1' et A_1''. Ces deux points sont symétriques par rapport au point O, sur la bissectrice de l'angle (OA_1', OA_1'') et à des distances de O égales à $\sqrt{|z'|\,|z''|} = \frac{1}{2}\sqrt{|I|}$.

Figurons alors les cinq points O, A_1', A_1'', ω' et ω''. Nous savons que

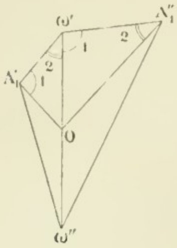

Fig 1.

$$\overline{A'_1 O\omega'} = \overline{\omega' O A''_1} \qquad \text{et} \qquad \overline{O\omega'}^2 = OA'_1 . OA''_1 ;$$

par conséquent $\frac{O\omega'}{OA'_1} = \frac{OA''_1}{O\omega'}$ et les deux triangles $OA'_1\omega'$, $O\omega'A''_1$ sont semblables, les angles 1 et 1, ainsi que 2 et 2, sont égaux ; l'angle ω' est donc le supplément de l'angle $A'_1O\omega'$; de même l'angle ω'' est le supplément de l'angle $A'_1O\omega''$, et le quadrilatère $\omega' A''_1 \omega'' A'_1$ est inscriptible. Il en est de même de tous ceux qu'indique l'énoncé (*fig.* 1 et 2).

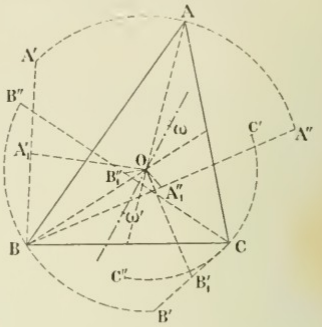

Fig. 2.

Soit λ le rapport d'homothétie du triangle $A_1B_1C_1$ au triangle ABC, les affixes des sommets de ce nouveau triangle sont λz_1, λz_2, λz_3. D'autre part, appelons maintenant z'_1 l'imaginaire conjugué de z_1, z_0 l'affixe du point ω' et z'_0 son imaginaire conjugué, enfin k la puissance d'inversion ; nous aurons, pour transformer un point quelconque du plan, $Z - z_0 = \frac{k}{z' - z'_0}$, $Z = X + iY$ étant l'affixe du point transformé de $z = x + iy$.

Appliquons aux trois points A_1, B_1, C_1, qui se transforment en A_2, B_2, C_2, et appelons Z_1, Z_2, Z_3 les affixes de ces nouveaux points ; nous aurons

$$Z_1 - z_0 = \frac{k}{\lambda z'_1 - z'_0}, \qquad Z_2 - z_0 = \frac{k}{\lambda z'_2 - z'_0}, \qquad Z_3 - z_0 = \frac{k}{\lambda z'_3 - z'_0};$$

puis,

$$Z_1 + Z_2 + Z_3 - 3z_0 = \frac{k}{\lambda z'_1 - z'_0} + \frac{k}{\lambda z'_2 - z'_0} + \frac{k}{\lambda z'_3 - z'_0}.$$

Si nous voulons que le point ω' soit le centre des moyennes distances des trois points A_2, B_2, C_2, il faudra que $Z_1 + Z_2 + Z_3 - 3z_0 = 0$, ou

$$\frac{1}{\lambda z'_1 - z'_0} + \frac{1}{\lambda z'_2 - z'_0} + \frac{1}{\lambda z'_3 - z'_0} = 0 ;$$

ce qui donne

$$\lambda^2 \Sigma z'_1 z'_2 - 2\lambda z'_0 \Sigma z'_1 + 3 z_0'^2 = 0.$$

Or $\Sigma z'_1 = 0$, $\Sigma z'_1 z'_2 = z'_1 z'_2 + z'_3(z'_1 + z'_2) = z'_1 z'_2 - (z'_1 + z'_2)^2$, $\Sigma z'_1 z'_2 = - z_1'^2 - z'_1 z'_2 - z_2'^2 = - I'$; d'autre part, $4z_0^2 = I$, $4z_0'^2 = I'$; donc l'équation qui donne λ devient

$$4 z_0'^2 \lambda^2 - 3 z_0'^2 = 0, \qquad \lambda^2 = \frac{3}{4}.$$

Il suffit donc de prendre $\lambda = \pm \frac{\sqrt{3}}{2}$.

3. Nous prendrons ici pour axe des x la droite $\omega'\omega''$ et pour origine des coordonnées le milieu de ce segment. Nous savons que z_0^2, qui se réduit ici à $\overline{O\omega'}^2$, est égal à $\frac{1}{4}$; d'autre part, nous avons vu au début que I ou u est égal à $\frac{z_1^2 + z_2^2 + z_3^2}{2}$, c'est-à-dire ici à $-p_2$, puisque $z_1 + z_2 + z_3$ est nul ; donc p_2 est constant et réel, et même négatif. D'ailleurs le produit des médianes étant constant, cela veut dire que le produit des longueurs OA, OB, OC est constant ; donc le module de p_3 est constant et égal à a^3.

Posons $x = \rho \cos \omega$, $y = \rho \sin \omega$, $z = \rho(\cos \omega + i \sin \omega)$; l'équation $z^3 + p_2 z - p_3 = 0$ deviendra

$$\rho^3(\cos 3\omega + i \sin 3\omega) + p_2\rho(\cos \omega + i \sin \omega) = p_3 = a^3(\cos \lambda + i \sin \lambda) ;$$

elle se décompose en deux :

$$\rho^3 \cos 3\omega + p_2\rho \cos \omega = a^3 \cos \lambda,$$
$$\rho^3 \sin 3\omega + p_2\rho \sin \omega = a^3 \sin \lambda,$$

et l'élimination de λ entre ces deux équations est immédiate ; elle donne l'équation du lieu des sommets

$$\rho^6 + 2p_2\rho^4 \cos 2\omega + p_2^2\rho^2 = a^6.$$

Quand deux des sommets B et C sont confondus, le triangle s'aplatit sur une droite AB, le point O est sur cette droite aux $\frac{2}{3}$ à partir de A, et la droite $\omega'\omega''$, c'est-à-dire Ox, est sur cette droite AB, qui correspond à ce triangle particulier. Nous avons alors $4p_2^3 + 27p_3^2 = 0$, le p_3 de ce triangle est réel et se réduit à a^3. Changeons de notation et posons alors $p_3^2 = 4l^6$; nous aurons $p_2^3 = -27l^6$, $p_2 = -3l^2$, et l'équation du lieu deviendra, dans ce cas particulier,

$$\rho^6 - 6l^2\rho^4 \cos 2\omega + 9l^4\rho^2 - 4l^6 = 0.$$

Pour construire cette courbe, au moins sommairement, remarquons que l'équation en ρ a des coefficients périodiques, et que leur période est π ; il suffit donc de faire varier ω dans un intervalle d'étendue 2π. Or le changement de ω en $-\omega$ laisse l'équation invariable ; donc la courbe est symétrique par rapport à Ox, et en tenant compte de cette symétrie, il suffit de faire varier ω de 0 à π ; le changement de ω en $\pi - \omega$ laisse encore l'équation invariable ; la courbe est symétrique par rapport à Oy, et, en tenant compte aussi de cette nouvelle symétrie, il suffit de faire varier ω de 0 à $\frac{\pi}{2}$.

Pour $\omega = 0$, l'équation devient

$$\rho^6 - 6l^2\rho^4 + 9l^4\rho^2 - 4l^6 = 0,$$

ou

$$\rho^2(\rho^2 - 3l^2)^2 - 4l^6 = 0 ;$$

elle se décompose en

$$\rho^3 - 3l^2\rho + 2l^3 = 0.$$
$$\rho^3 - 3l^2\rho - 2l^3 = 0 ;$$

la première admet la racine positive double $\rho = l$, et la seconde, la racine positive simple $\rho = 2l$; le premier point de Ox, A, est un point double, le second, B, est un point simple.

Quand ω augmente, $\cos 2\omega$ diminue, la somme des valeurs de ρ^2 diminue ; pour $\omega = \frac{\pi}{4}$, l'équation n'a déjà plus qu'une seule racine positive et cela se poursuit jusqu'à $\omega = \frac{\pi}{2}$. A ce moment, l'équation devient

$$\rho^6 + 6l^2\rho^4 + 9l^4\rho^2 - 4l^6 = 0, \qquad \text{ou} \qquad \rho^2(\rho^2 + 3l^2)^2 - 4l^6 = 0 ;$$

elle se décompose en

$$\rho^3 + 3l^2\rho - 2l^3 = 0, \qquad \text{et} \qquad \rho^3 + 3l^2\rho + 2l^3 = 0,$$

qui n'admettent qu'une racine positive, comprise entre 0 et l.

Pour savoir à quel moment l'équation cesse d'avoir trois racines positives, il suffit d'annuler le discriminant de l'équation cubique en ρ^2. Ceci donne, en posant $\cos 2\omega = u$, la relation

$$32u^3 - 27u^2 - 36u + 31 = 0,$$

qui se décompose en $u - 1 = 0$ et $32u^2 + 5u - 31 = 0$.

L'unique valeur positive de u est donc

$$u = \cos 2\omega = \frac{-5 + \sqrt{3993}}{64}, \qquad \cos 2\omega = \frac{63,1 - 5}{64} = \frac{58,1}{64} = 0,90 ;$$

cette valeur de 2ω est voisine de 65°, et ω a pour valeur correspondante un nombre α voisin de 37 ou 38°.

Alors de 0 à α, il y a trois valeurs positives de ρ ; pour $\omega = \alpha$, deux d'entre elles sont confondues ; la droite $\omega = \alpha$ est tangente à la courbe ; enfin, au delà, de $\omega = \alpha$ à $\omega = \frac{\pi}{2}$, il n'y a plus qu'un seul ρ qui soit positif, les deux autres ont cessé d'être réels.

Vérifions enfin que le point A est un point double. Pour cela, formons tg V en ce point : nous avons

$$\operatorname{tg} V = \rho \frac{d\omega}{d\rho}; \qquad \text{d'autre part} \qquad \cos 2\omega = \frac{\rho^6 + 9l^4\rho^2 - 4l^6}{6l^2\rho^4},$$

$$\cos 2\omega = \frac{\rho^2}{6l^2} + \frac{3}{2} l^2 \frac{1}{\rho^2} - \frac{2}{3} \frac{l^4}{\rho^4}, \qquad -2\sin 2\omega \frac{d\omega}{d\rho} = \frac{\rho}{3l^2} - \frac{3l^2}{\rho^3} + \frac{8}{3} \frac{l^4}{\rho^5} = \frac{\rho^6 - 9l^4\rho^2 + 8l^6}{3l^2\rho^5},$$

$$\frac{d\omega}{d\rho} = - \frac{(\rho^2 - l^2)(\rho^4 + l^2\rho^2 - 8l^4)}{6l^2\rho^5 \sin 2\omega}.$$

Dans le second nombre, l'un des facteurs, $-\frac{\rho^4 + l^2\rho^2 - 8l^4}{6l^2\rho^5}$, a pour limite $\frac{1}{l^3}$. quand ω tend vers 0 et ρ vers l; l'autre facteur, $\frac{\rho^2 - l^2}{\sin 2\omega}$, prend une forme indéterminée, et en lui appliquant la règle de L'Hospital, et appelant λ la limite de $\frac{d\omega}{d\rho}$, on a $\lambda = \frac{1}{l^3}\frac{l}{\lambda}$, $\lambda^2 = \frac{1}{l^2}$, $\lambda = \pm \frac{1}{l}$.

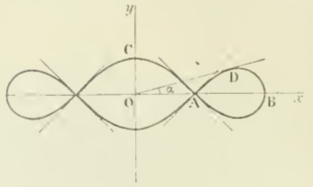

Fig. 3.

Les deux valeurs de $\operatorname{tg} V$ sont donc $\operatorname{tg} V = \pm 1$. La courbe a la forme ci-contre (*fig.* 3).

Lorsqu'on se donne le point A, on se donne l'affixe, z_1, de ce point, et il suffit de trouver les affixes des deux autres points, z_2 et z_3.

Or on a $z_2 + z_3 = - z_1$, $\quad I = - p_2 = 3l^2 = z_2^2 + z_3^2 + z_2 z_3$, $\quad I = 3l^2 = z_1^2 - z_2 z_3$;
ceci donne

$$z_2 z_3 = z_1^2 - 3l^2,$$

et nous sommes ramenés à construire deux affixes dont la somme et le produit sont donnés.

◆

GÉOMÉTRIE ANALYTIQUE

2106 — *On considère un faisceau ponctuel de coniques et une droite* Δ *: le pôle de* Δ *décrit une conique* Γ *harmoniquement circonscrite au faisceau. Les tangentes à* Γ *aux sommets du triangle conjugué rencontrent les côtés opposés en trois points situés sur une même droite* Δ' *: les deux droites* Δ *et* Δ' *sont réciproques. Soit* Γ' *la conique qui correspond à* Δ'.

Démontrer que si l'une des droites Δ, Δ' *tourne autour d'un point fixe, l'autre enveloppe une conique harmoniquement inscrite au faisceau; que le point d'intersection de* Δ *et de* Δ' *décrit une cubique, et que le quatrième point de rencontre des coniques* Γ *et* Γ' *décrit une sextique ayant trois points triples aux sommets du triangle conjugué.*

Prenons pour triangle de référence le triangle conjugué commun à toutes les coniques du faisceau donné : si nous désignons par

$$x^2 + y^2 + z^2 = 0, \qquad Ax^2 + By^2 + Cz^2 = 0$$

les équations de deux de ces coniques, l'équation générale des coniques de ce faisceau sera

$$(A + \lambda)x^2 + (B + \lambda)y^2 + (C + \lambda)z^2 = 0.$$

Le pôle de la droite Δ, $ux + vy + wz = 0$, est défini par les équations

$$\frac{(A + \lambda)x}{u} = \frac{(B + \lambda)y}{v} = \frac{(C + \lambda)z}{w};$$

on aura le lieu de ce point en éliminant λ entre ces trois équations. Pour faire cette élimination, nous égalons ces trois rapports à un paramètre $-\rho$, nous obtenons trois équations linéaires en λ et ρ et

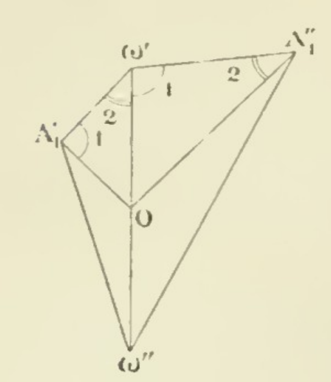
ω'
A''_1
2
1
2
A'_1
1
O
ω''

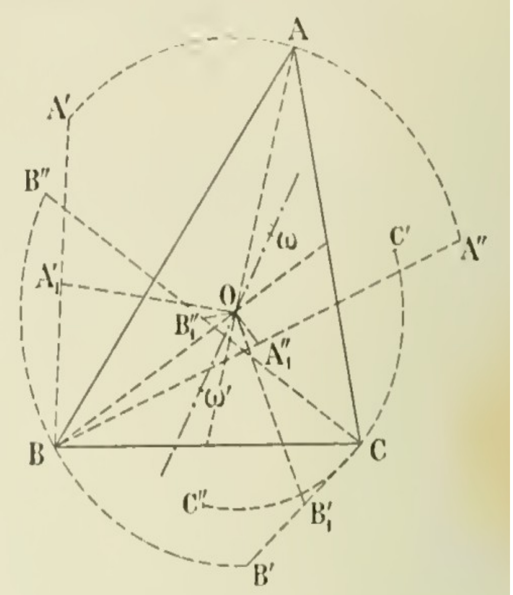
A
A'
B''
ω
C'
A''
A'_1
O
B''_1
A''_1
ω'
B
C
C''
B'_1
B'

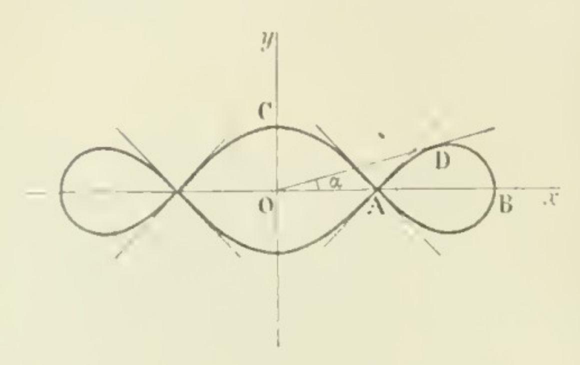
y
C
D
α
O
A
B
x

' "/ 0

Enfin, pour avoir le lieu du quatrième point de rencontre des coniques Γ et Γ′, nous éliminons u, v, w entre les équations

$$ux_0 + vy_0 + wz_0 = 0,$$
$$auyz + bvzx + cwxy = 0,$$
$$\frac{yz}{u} + \frac{zx}{v} + \frac{xy}{w} = 0.$$

Des deux premières nous tirons

$$\frac{u}{x(cyy_0 - bzz_0)} = \frac{v}{y(azz_0 - cxx_0)} = \frac{w}{z(bxx_0 - ayy_0)},$$

et en portant dans la troisième nous avons

$$\frac{yz}{x(cyy_0 - bzz_0)} + \frac{zx}{y(azz_0 - cxx_0)} + \frac{xy}{z(bxx_0 - ayy_0)} = 0,$$

ou

$$\frac{y^2z^2}{cyy_0 - bzz_0} + \frac{z^2x^2}{azz_0 - cxx_0} + \frac{x^2y^2}{bxx_0 - ayy_0} = 0.$$

Sous cette forme on reconnait l'équation d'une sextique ayant trois points triples aux sommets du triangle conjugué.

A. ROUSSEAU, à Fournes-en-Weppes.

Bonnes solutions par MM. Lach, à Douai; L. Naucelle, à La Châtre.

2107. — *Soit* M *le point le plus haut d'une cycloïde engendrée par un cercle* O *qui roule sur une droite* Δ *et qui actuellement touche la cycloïde au point* M. *Soit* AB *une parallèle à* Δ *qui coupe la cycloïde en* A *et le cercle en* B. *Montrer que* AB = arc BM, *et en déduire la construction du centre instantané de rotation qui correspond au point* A *de la cycloïde.*

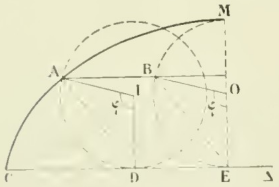

Le cercle O touche Δ en E. Soit I le centre du cercle générateur au moment où il passe en A et D le point de contact sur Δ. Si C est le point de départ sur Δ du mobile, et si on pose

$\widehat{DIA} = \widehat{EOB} = \varphi$, on a CD = arc DA = arc EB = $R\varphi$;

d'où AB = DE = $R(\pi - \varphi)$.

C'est précisément la valeur de l'arc BM. Le centre instantané de rotation, D, qui correspond au point A de la cycloïde, s'obtient donc en menant AD parallèle à BE; d'ailleurs le point A se déduit du point B en menant, en sens contraire du mouvement, BA parallèle à Δ et égal à arc MB.

A. ROUSSEAU, à Fournes-en-Weppes.

Bonnes solutions : MM. R. Manen, à Albi; D. Bellocq, à Bordeaux; M. Ullmann, à Héricourt; G. Lach, à Douai; L. Simon, à Fourmies; R. Dufour; G. Bèle.

◆

CONCOURS DE 1913 (*Suite.*)

ÉCOLE DES PONTS ET CHAUSSÉES

(Cours spéciaux et places réservées aux sous-ingénieurs et conducteurs.)

Algèbre et Analyse.

I. — **2191.** Un tore circulaire circonscrit à une sphère de rayon R satisfait aux conditions suivantes : le plan du parallèle inférieur passe par le centre de la sphère; le diamètre y du méridien est inférieur ou égal à R; le

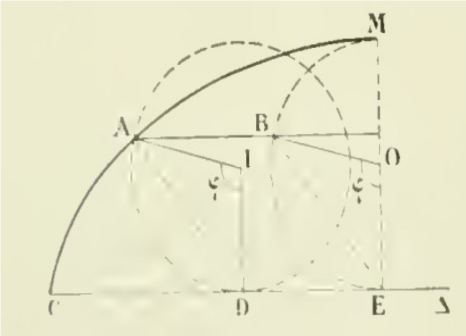

. 9(:69 9 92 9 , * ! 3 3!
' ,/ " :) v v)

En revenant aux coordonnées X, Y, ($x = aX$, $y = bY$), cette équation devient

$$abXY - \alpha bY - \beta aX = 0.$$

Posons, pour la symétrie des notations, $\alpha = aA$, $\beta = bB$, en remarquant que, par suite, on a la condition $A + B = 1$; l'hyperbole lieu de M devient

$$XY - AX - BY = 0.$$

Pour qu'on puisse trouver sur ce lieu deux points M et M′, il faut que cette hyperbole puisse appartenir au faisceau des deux hyperboles

$$XY - X + R = 0, \qquad XY - Y + S = 0.$$

En multipliant la première équation par S et en retranchant la seconde multipliée par R, on en déduit par identification :

$$A = \frac{S}{S - R}, \qquad B = \frac{R}{R - S},$$

qui vérifie bien

$$A + B = 1.$$

A et B étant connus, on voit que R et S sont soumis à une seule condition ; on peut la mettre sous la forme symétrique

$$AR + BS = 0.$$

ou, avec les anciennes notations, $\frac{\alpha r}{a^2} + \frac{\beta s}{b^2} = 0$, ce qui exprime que le lieu de N est le diamètre conjugué par rapport à l'ellipse de centre O, de sommets A et B, de la droite joignant l'origine au point de AB qui a pour coordonnées α, β.

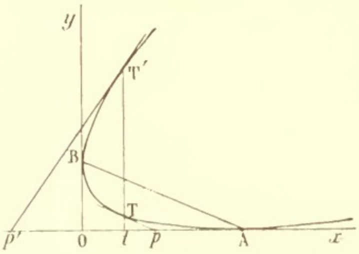

Remarques géométriques. — 1° Il est bien évident, d'après la figure même, qu'à un point p ne correspond qu'un point t ; mais qu'à un point t correspondent deux points T et T′, et par suite deux points p, p', qui forment une involution. Considérons les quatre tangentes fixes à (P) : Tp, T′p', Oy et la droite de l'infini. Le rapport anharmonique qu'elles interceptent sur une tangente quelconque est constant. En prenant la tangente infiniment voisine de Oy et remarquant que TT′ est parallèle à Oy, on en déduit que ce rapport est -1. Donc si nous coupons par Ox tangente à (P) on voit que O est le milieu de pp', ce qui montre que l'involution de ces deux points est une symétrie.

2° La figure montre immédiatement qu'à un point M correspond un seul point N et à un point N deux points M. Si l'on remarque que toute tangente à P, pq', est telle que p et q' soient les projections sur les axes d'un point C de AB, on voit que MCM′D est un rectangle dont AB est une diagonale ; on en déduit immédiatement la construction de M′ connaissant M, telle qu'elle a été donnée ci-dessus. On sait de plus que, dans un rectangle tel que MCM′D, les trois diagonales de MCM′D, pCq'O et qOp'D sont concourantes ; donc N est sur MM′.

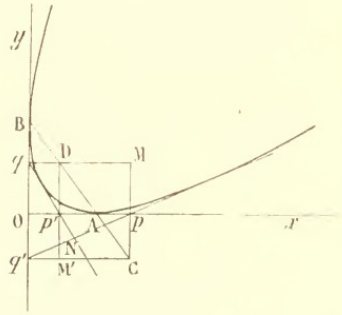

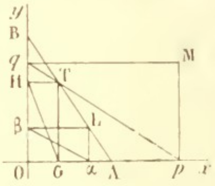

Si M et M′ sont confondus, ils sont sur AB, et p, p' sont confondus ainsi que q, q'. Les deux tangentes pq', $p'q$ étant confondues, leur point de rencontre est leur point de contact ; donc N décrit (P).

Si M et N sont confondus, pM et pN étant confondus, pN est une tangente verticale à (P). On trouverait de même la tangente horizontale à (P), résultats déjà obtenus par le calcul.

3° On sait que si l'on projette un point quelconque L de AB sur les axes en α et β, la droite $\alpha\beta$ est tangente à la parabole bitangente aux axes en A et B. De même pour tout point de la droite pq. Si donc on veut une tangente commune autre que les axes, ou la droite de l'infini, aux deux paraboles de l'énoncé, il suffira de prendre le point de rencontre de AB et pq, **T**, et de le projeter sur les axes, en G, H. GH est Δ_M.

Si cette droite est fixe, T' l'est aussi, et par suite pq passe par un point fixe T. On en déduit qu'il y a correspondance homographique entre p et q, et que par suite M décrit une hyperbole équilatère de directions asymptotiques Ox, Oy.

A. S. L.

Bonnes solutions : MM. J. Ottenheimer, lycée de Nancy ; G. Foucry, à Reims ; P. Villain ; Doligez ; Ch. Guillerme ; R. Dontot, élève de l'École normale supérieure ; G. Lach à Rodez ; A. Courtois ; Jambert ; J. Soubaigné.

◆

MÉCANIQUE

1918. — 1° *On considère un point* M *mobile dont la trajectoire est la cubique gauche* Γ, $x = at$, $y = bt^2$, $z = ct^3$, *a, b, c étant les coordonnées d'un point* P *fixe et t désignant le temps. On suppose a, b, c positifs et non nuls. On suppose dans la suite que le point* P *reste sur une surface* (S) *telle que pour les valeurs ainsi définies de a, b, c, le vecteur-vitesse en chaque point de* Γ *fasse un angle constant, que l'on calculera, avec une droite* Δ *convenablement choisie.*

P *étant fixe, étudier en chaque point de* Γ *la vitesse. Hodographe. (On déterminera en vraie grandeur les éléments de l'hodographe.)*

2° *Calculer en chaque point de* Γ *l'accélération totale et l'accélération tangentielle du mobile. Déduire de leur comparaison le rayon de courbure de* Γ. *Variations de ce rayon de courbure.*

3° *A chaque point* P *de* S *correspond un certain mouvement.* P *étant supposé arbitraire sur* S, *y a-t-il un nombre fini ou infini de mouvements faisant passer le point* M *par un point* Q *donné de l'espace (deux mouvements distincts pouvant d'ailleurs correspondre à une même trajectoire). Discuter suivant la position de* Q.

4° *Étant donné un point* Q *on considère le mouvement qui fait passer* M *en* Q *au temps t', le temps* 0 *correspondant au passage du mobile à l'origine. On suppose que ce mobile quitte sa trajectoire en* Q *suivant la tangente avec une vitesse uniforme égale précisément à sa vitesse au passage par* Q. *Lieu des points* Q' *occupés dans cette hypothèse par le mobile au temps* $t' + 1$ *quand* Q *occupe toutes les positions possibles (On suppose que t' a une valeur fixe donnée).*

1° Soient α, β, γ les cosinus directeurs de la direction cherchée ; l'angle V de la tangente à Γ avec cette direction est donné par

$$\cos^2 V = \frac{(a\alpha + 2b\beta t + 3ct^2\gamma)^2}{a^2 + 4b^2t^2 + 9c^2t^4}.$$

Ce rapport doit être indépendant de t ; en particulier les termes en t et t^3 doivent disparaître au numérateur, ce qui donne, a, b, c ne pouvant pas être nuls, $\alpha = \gamma = 0$, ou $\beta = 0$. La première hypothèse ne convient pas. La seconde donne, en écrivant que les divers coefficients sont proportionnels au numérateur et au dénominateur,

$$\alpha^2 = \frac{6ac\alpha\gamma}{4b^2} = \gamma^2.$$

On a $\alpha = \gamma$, ou $\alpha = -\gamma$. Mais a, b, c devant être positifs, on en déduit $\alpha = \gamma$, ce qui fixe la direction demandée et l'angle V (45°) ; on trouve en outre la condition

$$3ac = 2b^2,$$

équation d'un cône qui est le lieu cherché de P. P étant sur ce cône (S), toute la cubique Γ y est également, comme on le vérifie aisément.

La vitesse en un point de Γ est donnée par $X = a$, $Y = 2bt$, $Z = 3ct^2$. L'hodographe est la parabole du plan $X = a$, d'équation $Y^2 = \frac{4b^2}{3c} Z$ ou $Y^2 = 2aZ$.

&

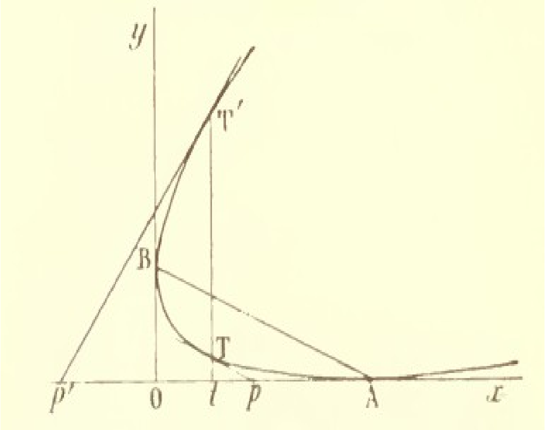
y
T′
B
T
p′
0
t
p
A
x

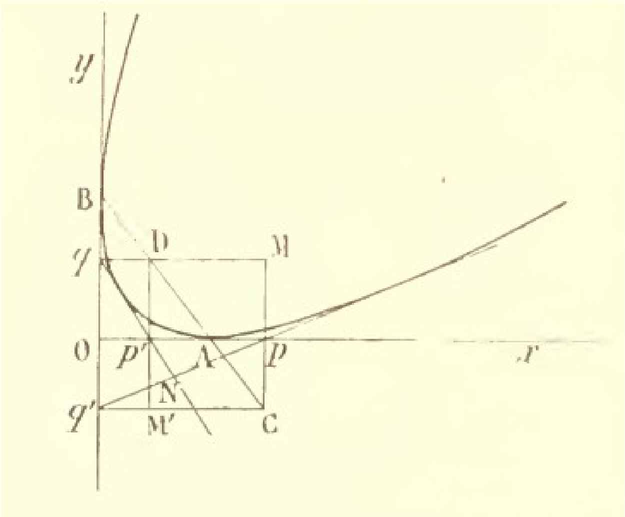
y
B
q
D
M
0
p′
A
p
x
N
q′
M′
C

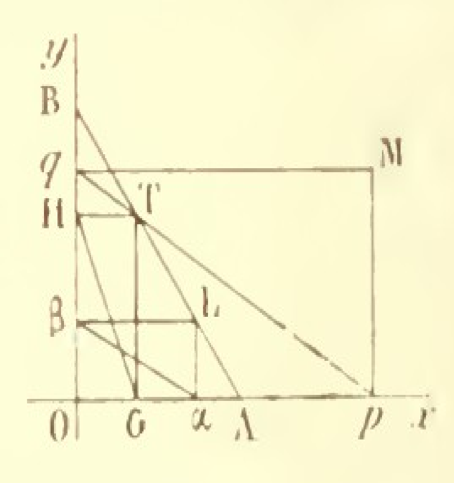
y
B
q
M
H
T
β
L
O
G
α
A
p
x

' ,0 " A 3

CONCOURS DE 1911 *(Suite.)*

AGRÉGATION DES SCIENCES MATHÉMATIQUES

Mathématiques élémentaires [1].

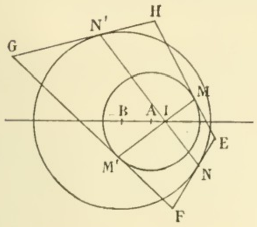

1° Si l'on mène à une ellipse (C) de centre C deux cercles bitangents, l'un (A) ayant son centre A sur l'axe focal, l'autre (B) ayant son centre B sur l'axe non focal, le point d'intersection I des cordes de contact MM' et NN' est un point limite pour le système des deux cercles. (On demande aux candidats une démonstration indépendante de la théorie des pôles et polaires dans les coniques ; s'ils emploient l'équation de la droite AB, rapportée aux axes de l'ellipse, pour montrer d'abord que le point I est sur cette droite, ils établiront géométriquement les formules qui expriment $\overline{\mathrm{CA}}$ et $\overline{\mathrm{CB}}$ au moyen des coordonnées $\overline{\mathrm{CP}}$ et $\overline{\mathrm{CQ}}$ du point I. Ils pourront donner *subsidiairement* une démonstration fondée sur la théorie des pôles et polaires dans les coniques).

2° On donne deux cercles (A) et (B), de centres A et B, dont l'un (A) est intérieur à l'autre ; soit I celui des deux points limites du faisceau (A, B) qui est intérieur aux deux cercles. (Ce choix du point I a pour but d'écarter une discussion que les candidats n'auraient pas le temps de faire convenablement).

Si l'on mène par I deux cordes rectangulaires MM' et NN' appartenant respectivement aux deux cercles, il existe une ellipse (C) doublement tangente aux deux cercles, les cordes de contact étant MM' et NN' ; le démontrer. Les cordes variant, les pôles R et S de ces cordes sont sur une droite fixe Δ perpendiculaire à la droite AB en un point J ; les circonférences qui ont pour diamètres les segments RS sont orthogonales à la circonférence de diamètre AB qui est le lieu des centres des ellipses (C) ; les carrés a^2 et b^2 des demi-axes des ellipses (C) sont, les uns et les autres, proportionnels aux distances CD des centres C à la droite Δ : les ellipses (C) sont, par suite, semblables entre elles.

3° Les données restant les mêmes, les tangentes en M et M' sont rencontrées par les tangentes en N et N' aux points E, F, G, H. Désignant par λ l'angle que fait IM avec le prolongement de BA, on établira les *six* relations qui déterminent en fonction des constantes de la figure $\left(r,\ \mathrm{R},\ \frac{\mathrm{IA}}{r} = \mathrm{A},\ \frac{\mathrm{IB}}{\mathrm{R}} = \mathrm{B}\right)$ et du paramètre λ les éléments essentiels du quadrilatère IMEN, à savoir les angles $\mathrm{IME} = \alpha$, $\mathrm{INE} = \beta$, les côtés IM et IN, les côtés EM et EN ; les deux dernières de ces relations sont

$$\mathrm{EM} \sin \alpha + \mathrm{EN} \cos \beta = \frac{\mathrm{R} \sin (\beta - \lambda)}{\cos \lambda},$$

$$\mathrm{EM} \cos \alpha + \mathrm{EN} \sin \beta = \frac{- r \cos (\lambda + \alpha)}{\sin \lambda}.$$

Vérifier au moyen de ces relations l'égalité $\frac{\mathrm{EM}}{\mathrm{EN}} = \frac{\mathrm{B}}{\mathrm{A}}$, et en déduire que le quadrilatère EFGH reste inscrit à un cercle fixe (O) du faisceau (A, B) lorsque les cordes MM' et NN' varient. On determinera la position du centre O de ce cercle en fonction des longueurs IA, IB, JA, JB, JI, en calculant successivement $\frac{\mathrm{OA}}{\mathrm{OB}}$, OA, OB ; on calculera aussi la distance JO, on interprètera géometriquement la formule obtenue, et l'on en déduira une conséquence pour les circonférences décrites sur les segments RS comme diamètres.

Comparant deux expressions différentes de la valeur du rapport $\frac{\mathrm{ME}}{\mathrm{MH}}$ en fonction des quantités IE, IH, $\widehat{\mathrm{MIE}} = \gamma$, $\widehat{\mathrm{MIH}} = \delta$, on obtiendra $\gamma = \delta$, et on interpretera ce resultat en considerant les diagonales EG et FH. De quelle nature est la correspondance entre les droites EG et FH ?

Mathématiques spéciales.

1968. — On considère dans un plan la parabole (P) et la droite (D) dont les equations sont, en coordonnées rectangulaires,

$$y^2 - 2px = 0, \qquad y - a = 0.$$

---

[1] Des solutions étudiées de la question de mathématiques élémentaires sont publiées dans le *Journal de Mathématiques élémentaires*.

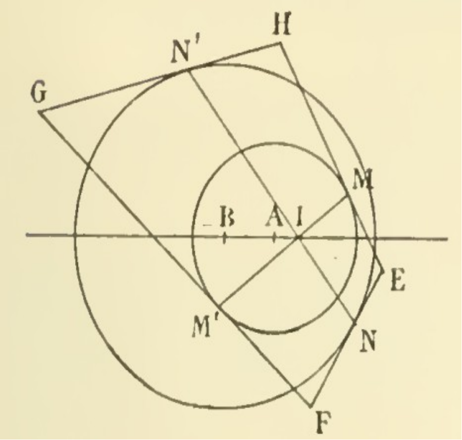
H
N'
G
M
B
A
I
E
M'
N
F

& 3

N } @ !)

?R|P@ |& ERA 4 3 •

. ?R|P@ |& ERA } ! ! ! v

2. @ } E |?@ R&R| A P@ !
A ! 3) " 4 3

9(:69 9 92 9 , * ! 3 3!

'

REVUE
DE
MATHÉMATIQUES SPÉCIALES

11e Année. — N° 1. Octobre 1900.

PREMIÈRE PARTIE

DÉMONSTRATION DU THÉORÈME DE PONCELET SUR LES POLYGONES INSCRITS DANS UNE CONIQUE ET CIRCONSCRITS A UNE AUTRE

par M. **E. Cahen**, professeur de Mathématiques spéciales au Collège Rollin.

Dans le n° 9 de l'année 1900 de ce journal, M. Fontené considère deux points d'une conique V, dont les paramètres t et t' sont liés par une relation *doublement quadratique* et *symétrique*

$$A t^2 t'^2 + B t t'(t + t') + C(t^2 + t'^2) + D t t' + E(t + t') + F = 0. \tag{1}$$

La droite qui joint ces deux points enveloppe une conique U *et, réciproquement.*

Si la droite qui joint deux points d'une conique V *est tangente à une conique* U, *les paramètres de ces deux points sont liés par une relation de la forme précédente.*

Bien entendu, les énoncés corrélatifs sont vrais.

Comme application, M. Fontené indique le *lieu du troisième sommet d'un triangle circonscrit à une conique* U *et dont les deux autres sommets décrivent une conique* S et l'enveloppe corrélative.

Ceci peut se généraliser :

Soient deux coniques V *et* U. *Un polygone de* p *côtés est inscrit dans* V *et* $p - 1$ *de ses côtés sont tangents à* U. *Quelle est l'enveloppe du* p^e *côté ?* (Sur la figure $p = 5$).

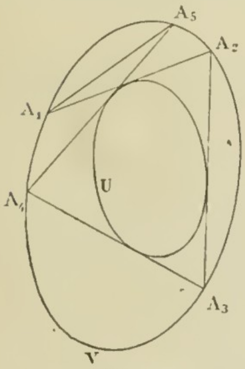

Fig. 1.

Or, il y a entre les paramètres des points A_1 et A_p sur la conique V, une relation qui est évidemment *algébrique* et *symétrique*. D'autre part, cette relation est du second degré. En effet, si l'on se donne A_1 il y a deux positions possibles pour A_2; mais si l'on choisit l'une de ces positions il y a une position déterminée pour A_3 puis une pour A_4 et ainsi de suite. Donc si l'on se donne A_1, il y a deux positions possibles pour A_p. La relation *algébrique*, *symétrique* entre les paramètres de A_1 et de A_p est donc, de plus, du second degré. Donc A_1A_p *enveloppe une conique*.

Corrélativement : *Un polygone de* p *côtés étant circonscrit à une conique et* $p - 1$ *de ses sommets étant sur une autre, le lieu du* p^e *sommet est une conique.*

Pour arriver au théorème de Poncelet, cherchons s'il existe un polygone de p côtés inscrit dans V et circonscrit à U. Or, si d'un point A_1 de U, on déduit, au moyen de p tangentes consécutives, un point A_{p+1}, il y a entre les paramètres de A_1 et de A_{p+1}, une relation

$$A^{(p+1)} t^2 t'^2 + B^{(p+1)} t t'(t + t') + C^{(p+1)}(t^2 + t'^2) + D^{(p+1)} t t' + E^{(p+1)}(t + t') + F^{(p+1)} = 0,$$

$A^{(p+1)}$, $B^{(p+1)}$, ... $F^{(p+1)}$ étant des coefficients qui dépendent de p.

Pour que le point A_1 soit sommet d'un polygone cherché, il faut et il suffit qu'il coïncide avec A_{p+1}.

On obtient donc son t, en faisant $t' = t$ dans l'équation précédente, ce qui donne

$$(2) \qquad A^{(p+1)}t^4 + 2B^{(p+1)}t^3 + [2C^{(p+1)} + D^{(p+1)}]t^2 + 2E^{(p+1)}t + F^{(p+1)} = 0.$$

Cette équation est du quatrième degré ; on trouve donc quatre positions pour le point A_1.

Mais il faut remarquer qu'il y a, quel que soit p, quatre positions du point A_1 dont les paramètres doivent satisfaire à l'équation (2), sans que ces points donnent de vrais polygones répondant à la question. Ce sont les *quatre points de contact avec* V, *des tangentes communes à* U *et* V.

Si on considère un tel point A_1 (voir la figure 2 où les tangentes communes à U et V sont réelles), il est évident qu'il coïncide avec les points A_2, A_3, ... A_{p+1}, ...

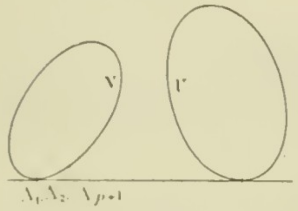

Fig. 2.

Donc l'équation (2) doit donner ces quatre solutions singulières, et comme elle est du quatrième degré, elle n'en donne, en général, pas d'autres.

Donc, en général, *il n'y a pas de vrai polygone de p côtés inscrit dans* V *et circonscrit à* U.

Mais on voit que s'il y en a un, l'équation (2) ayant plus de quatre racines est une identité. *Donc s'il y a un de ces polygones il y en a une infinité.* C'est le théorème de Poncelet.

On peut choisir arbitrairement un sommet du polygone sur V.

Corrélativement, on peut choisir arbitrairement un côté du polygone tangent à U (*).

Voici quelques remarques que suggère la méthode précédente.

Considérons un polygone de p côtés inscrit dans V et circonscrit à U ; si partant d'un sommet, on joint les sommets de q en q, jusqu'à ce qu'on revienne au point de départ, on forme un polygone de $\frac{p}{d}$ côtés (d étant le plus grand commun diviseur de p et q).

Or, entre les paramètres de deux sommets dont les rangs diffèrent de q, existe une relation doublement quadratique et symétrique, qui est la même, quelle que soit la position de ces sommets. Donc le polygone obtenu est circonscrit à une conique $U^{(q)}$, et cette conique reste la même si le polygone primitif varie en restant inscrit dans V et circonscrit à U. Au système de deux coniques U et V, correspondent ainsi p coniques

$$U^{(1)}, \qquad U^{(2)}, \ldots \qquad U^{(p)},$$

(dont la première est V, et la dernière U).

Sur ces p coniques, il y en a $\varphi(p)$ qui correspondent à des polygones de p côtés ; et plus généralement $\varphi\left(\frac{p}{d}\right)$ qui correspondent à des polygones de $\frac{p}{d}$ côtés.

(*) La démonstration précédente ne prouve pas qu'il existe effectivement des polygones de Poncelet. Il est clair qu'il en existe pour $p \leq 5$, car si l'on circonscrit un pentagone, par exemple, à une conique U, on peut faire passer une conique V par les cinq sommets de ce pentagone. Mais pour $p > 5$, ce raisonnement ne s'applique pas. On peut se borner à un exemple. Soit la parabole V,
$$x = t, \qquad y = t^2$$
et la parabole U dont l'équation est
$$x^2(1 + \cos\varphi) - 2y = 0.$$

On trouve facilement que la relation qui doit exister entre les paramètres t et t' de deux points A_1, A_2 de V pour que la droite A_1A_2 soit tangente à U est
$$t^2 + t'^2 - 2tt'\cos\varphi = 0.$$

Quant à la relation entre les paramètres t et $t^{(p)}$ de A_1 et de A_{p+1}, on voit (de proche en proche) que c'est
$$t^2 + t^{(p)2} - 2tt^{(p)}\cos p\varphi = 0.$$

L'équation (2) se réduit donc ici à
$$2(1 - \cos p\varphi)t^2 = 0.$$

Elle est une identité lorsque
$$\cos p\varphi = 1, \qquad \text{ou} \qquad \varphi = \frac{2k\pi}{p}.$$

Pour ces valeurs de φ, il existe des polygones de p côtés inscrits dans V et circonscrits à U.

En particulier, si p est pair, en faisant $q = \frac{p}{2}$, on est amené à joindre deux sommets opposés du polygone. Pour ces deux sommets la relation quadratique se réduit à une relation du premier degré symétrique, c'est-à-dire à une involution (car à un sommet du polygone ne correspond qu'un sommet opposé). Donc

Quand un polygone d'un nombre pair de côtés est inscrit dans une conique et circonscrit à une autre, les diagonales qui joignent les sommets opposés se coupent en un même point qui est fixe quand le polygone varie.

Dans les cas de l'hexagone, ce point est le point de Brianchon.

Tous ces théorèmes ont d'ailleurs leurs corrélatifs.

Par exemple, dans les polygones précédents, les points de rencontre des côtés opposés sont tous sur une même droite fixe. Dans le cas de l'hexagone, c'est la droite de Pascal.

◆

ALGÈBRE

904. — *Étant donnée une équation du quatrième degré* $f(x) = ax^4 + 4bx^3 + 6cx^2 + 4b'x + a' = 0$, *trouver les conditions qui doivent être remplies pour que l'équation* $f(y + h) = 0$ *soit réciproque pour trois valeurs de* h.

Dans ce cas, trouver les valeurs de h *et les racines de l'équation* $f(x) = 0$.

Si dans l'équation proposée, on effectue la substitution indiquée $x = y + h$, cette équation devient

$$ay^4 + 4(ah + b)y^3 + 6(ah^2 + 2bh + c)y^2 + 4(ah^3 + 3bh^2 + 3ch + b')y + ah^4 + 4bh^3 + 6ch^2 + 4b'h + a' = 0.$$

En écrivant que cette équation est une équation réciproque ordinaire, nous avons les deux relations

$$ah^4 + 4bh^3 + 6ch^2 + 4b'h + a' - a = 0,$$
$$ah^3 + 3bh^2 + (3c - a)h + b' - b = 0$$

qui doivent être vérifiées simultanément pour trois valeurs de h. Or, en multipliant la seconde par h et retranchant le résultat ainsi obtenu de la première, celle-ci s'abaisse au troisième degré et devient

$$bh^3 + (3c + a)h^2 + (3b' + b)h + a' - a = 0,$$

et nous avons alors à exprimer que les deux équations du troisième degré en h ont trois racines communes, par suite, qu'elles ont leurs coefficients proportionnels. Cela nous donne les relations

$$\frac{b}{a} = \frac{3c + a}{3b} = \frac{3b' + b}{3c - a} = \frac{a' - a}{b' - b}.$$

Posons alors $\frac{b}{a} = k$; nous aurons $\frac{3c}{a} = 3k^2 - 1$, $\frac{3b' + b}{a} = k(3k^2 - 2)$, d'où $\frac{b'}{a} = k(k^2 - 1)$;

et enfin $$\frac{a' - a}{a} = k^2(k^2 - 2) \quad \text{et} \quad \frac{a'}{a} = (k^2 - 1)^2.$$

L'équation du quatrième degré est donc

$$x^4 + 4kx^3 + 2(3k^2 - 1)x^2 + 4k(k^2 - 1)x + (k^2 - 1)^2 = 0;$$

elle peut s'écrire $(x + k)^4 - 2(x + k)^2 + 1 = 0$ ou $[(x + k)^2 - 1]^2 = 0$.

Elle a deux racines doubles qui sont $1 - k$ et $-1 - k$.

11e Année. N° 2. Novembre 1900.

REVUE DE MATHÉMATIQUES SPÉCIALES

PREMIÈRE PARTIE

PARAMÈTRE TANGENTIEL D'UN CONE DU SECOND ORDRE

par M. **G. Fontené**.

1. Je me propose principalement d'établir le fait suivant :

Soit une conique fixe S, *et soit un cercle* C *dont le plan est parallèle au plan de la conique et dont le centre est sur la perpendiculaire au plan de la conique menée par le centre de cette courbe ; si l'on considère les cônes du second ordre qui passent par la conique et qui ont leurs sommets sur le cercle* C, *si on les suppose indéformables, et si on les fait passer par une conique quelconque* S' *de manière que leurs sommets soient encore dans un plan parallèle au plan de la conique, ces sommets seront encore sur un cercle* C' *ayant son centre sur la perpendiculaire au plan de la conique* S' *menée par le centre de cette courbe.* Si S ou S' est une parabole, le cercle correspondant est remplacé par une parallèle à la directrice.

2. La forme d'un cône du second ordre dépend de deux paramètres, dont l'un est particulièrement intéressant. Les axes étant rectangulaires, l'équation du cône étant $f = 0$ avec le sommet à l'origine, l'équation du cône isotrope de même sommet étant $\varphi \equiv x^2 + y^2 + z^2 = 0$ *sans facteur constant parasite*, soient Δ, Θ, Θ' les coefficients de λ^3, λ^2, λ dans le discriminant de la forme $\lambda f + \varphi$; ces invariants étant des degrés 3, 2, 1 par rapport aux coefficients de f, *les rapports des quantités*

$$\sqrt[3]{\Delta}, \qquad \sqrt[2]{\Theta}, \qquad \Theta'$$

sont indépendants du facteur arbitraire qui entre dans f et sont, par suite, complètement déterminés par la forme du cône.

Nous appellerons *paramètre tangentiel* d'un cône du second ordre la quantité

$$I = \frac{\Theta}{\sqrt[3]{\Delta^2}} ;$$

dans l'hypothèse $I = 0$, ou $\Theta = 0$, le cône est harmoniquement inscrit au cône isotrope de même sommet, c'est-à-dire qu'il admet des trièdres trirectangles circonscrits.

3. Voici une expression géométrique de l'invariant I.

Soit un cône du second ordre sur lequel on trace une conique. S'il s'agit d'une conique à centre, en désignant par a^2 et b^2 les carrés des demi-axes de la conique, par γ la distance du sommet du cône au plan *de la conique, par* P *la puissance du sommet du cône par rapport à la sphère orthoptique de la conique, on a*

$$(1) \qquad I = \frac{P}{\sqrt[3]{a^2 b^2 \gamma^2}}.$$

S'il s'agit d'une parabole, en désignant par p le paramètre de la parabole, par γ la distance du sommet du cône au plan de la parabole, par d la distance du sommet du cône au plan orthoptique de la parabole, distance comptée positivement lorsque ce sommet est séparé de la courbe par le plan orthoptique, on a

$$(1') \qquad I = \frac{2d}{\sqrt[3]{p\gamma^2}}.$$

Soit d'abord une conique à centre. L'origine des coordonnées étant au sommet du cône, les axes Ox et Oy étant pour plus de simplicité parallèles aux axes de la conique, les équations de la conique sont

$$z = \gamma, \qquad A(x-\alpha)^2 + B(y-\beta)^2 + C = 0.$$

et l'on a les deux cônes

$$\left\{\begin{aligned} f &\equiv A(\gamma x - \alpha z)^2 + B(\gamma y - \beta z)^2 + Cz^2 = 0,\\ \varphi &\equiv x^2 + y^2 + z^2 = 0.\end{aligned}\right.$$

On peut écrire $f \equiv ax^2 + by^2 + cz^2 + 2fyz + 2gzx,$

avec $a = A\gamma^2, \qquad b = B\gamma^2, \qquad c = A\alpha^2 + B\beta^2 + C, \qquad f = -B\beta\gamma, \qquad g = -A\alpha\gamma,$

et l'on a $\Delta = abc - af^2 - bg^2 = AB\gamma^4(A\alpha^2 + B\beta^2 + C) - AB^2\beta^2\gamma^4 - BA^2\alpha^2\gamma^4,$

ou $\Delta = ABC\gamma^4,$

$$\Theta = (bc - f^2) + (ac - g^2) + ab = B\gamma^2(A\alpha^2 + B\beta^2 + C) - B^2\beta^2\gamma^2 + A\gamma^2(A\alpha^2 + B\beta^2 + C) - A^2\alpha^2\gamma^2 + AB\gamma^4,$$

ou encore $\Theta = AB\gamma^2(\alpha^2 + \beta^2 + \gamma^2) + C(A + B)\gamma^2 = AB\gamma^2\left(\alpha^2 + \beta^2 + \gamma^2 + \dfrac{C}{A} + \dfrac{C}{B}\right);$

à cause de $A = \dfrac{-C}{a^2}, \qquad B = \dfrac{-C}{b^2},$ on a donc

$$\Delta = \frac{C^3\gamma^4}{a^2b^2}, \qquad \Theta = \frac{C^2\gamma^2}{a^2b^2}\,.\,P.$$

L'élimination de C donne

$$I = \frac{\Theta}{\sqrt[3]{\Delta^2}} = \frac{P\gamma^2}{a^2b^2\sqrt[3]{\dfrac{\gamma^8}{a^4b^4}}} = \frac{P}{\sqrt[3]{a^2b^2\gamma^2}}.$$

Pour passer au cas de la parabole, nous partirons par exemple du cas de l'ellipse. Si R est le rayon de la sphère orthoptique, on peut écrire

$$I = \frac{\dfrac{P}{R}}{\sqrt[3]{\dfrac{a^2b^2}{R^3}\cdot\gamma^2}};$$

en posant $\dfrac{b^2}{a} = p \ (p > 0),$ ce qui donne $R = \sqrt{a(a+p)},$ on a d'ailleurs

$$\frac{a^2b^2}{R^3} = \frac{a^3p}{\sqrt{a^3(a+p)^3}};$$

si l'on fait croître a indéfiniment, d'une part, la quantité $\dfrac{P}{R}$ a pour limite le double de la distance d du sommet du cône au plan orthoptique de la parabole, cette distance étant positive lorsque le sommet du cône est séparé de la parabole par le plan orthoptique, d'autre part, la constante ci-dessus tend vers p ; on a donc la formule (1').

On a cette conséquence :

4. *Le lieu des sommets des cônes de même paramètre tangentiel circonscrits à une conique à centre donnée est une surface de révolution dont l'axe est la perpendiculaire au plan de la conique menée par le centre de cette conique.* La puissance P d'un point de cette surface par rapport à la sphère orthoptique de la conique est proportionnelle à la racine cubique du carré de la distance γ de ce point au plan de la conique ; cette puissance a d'ailleurs le signe de I ou le signe contraire selon qu'il s'agit d'une ellipse ou d'une hyperbole.

Pour une parabole, le lieu en question est un cylindre dont les génératrices sont parallèles à la directrice de la courbe. On a d'ailleurs $d = \lambda.\gamma^{\frac{2}{3}},$ et λ a le signe de I.

5. Le fait énoncé au début de cette note résulte également de la formule (1), les cônes considérés ayant même paramètre tangentiel.

GÉOMÉTRIE ANALYTIQUE

911. — *On donne un angle droit xOy et sur Ox deux points A, A′ symétriques l'un de l'autre par rapport au point O. Trouver l'équation générale des coniques tangentes à Oy, traçant sur Ox une involution dont A et A′ sont les points doubles et ayant pour centre commun le point de la première bissectrice qui se projette sur Ox en A.*

Cela étant, on demande :

1° *De séparer les points du plan par lesquels passent deux coniques réelles du faisceau de ceux par lesquels il passe deux coniques imaginaires ;*

2° *De trouver le lieu des points de rencontre de la parallèle à Ox menée par le point de contact de l'une de ces coniques avec Oy et des deux droites qui joignent le centre aux points de rencontre de cette conique avec Ox ; en outre, de construire ce lieu ;*

3° *De trouver le lieu du point de rencontre ordinaire de chacune de ces coniques avec le cercle osculateur au point situé sur Oy, et de construire ce lieu.*

Si nous désignons OA par a, l'équation demandée sera évidemment de la forme

$$(x-a)^2+2\lambda(x-a)(y-a)+\mu(y-a)^2+\nu=0.$$

Nous aurons successivement les relations qui existent entre les paramètres λ, μ, ν en écrivant d'abord que les deux points de rencontre avec Ox divisent harmoniquement les points A et A′, c'est-à-dire que le produit des racines de l'équation en x obtenue en annulant y dans l'équation de la conique est égal à a^2 ; puis, que l'axe des y rencontre la conique en deux points confondus, c'est-à-dire que l'équation en $y-a$ obtenue en faisant $x=0$ dans l'équation de la conique a une racine double. Le premier calcul donne $\nu=-2\lambda a^2-\mu a^2$; la seconde donne, en tenant compte de cette valeur de ν, $(\lambda+\mu)^2=\mu$. Posons alors $\mu=\rho^2$, nous aurons $\lambda+\rho^2=\pm\rho$, et il suffit de prendre un signe dans cette relation, car, prendre ensuite le second signe revient à changer ρ en $-\rho$ et ceci n'altère pas les valeurs que prend le paramètre, puisque c'est une variable indépendante. Nous avons donc finalement $\mu=\rho^2$ et $\lambda=\rho-\rho^2$; de là nous déduisons sans peine, en développant et simplifiant, l'équation générale des coniques envisagées

$$x^2+2(\rho-\rho^2)xy+\rho^2y^2-2a(1+\rho-\rho^2)x-2a\rho y+a^2=0. \tag{1}$$

La signification géométrique de ρ est facile à apercevoir : en faisant $x=0$ dans l'équation, nous trouvons $(\rho y-a)^2=0$; $\dfrac{a}{\rho}$ est donc l'ordonnée du point de contact avec Oy.

1. Écrivons que la conique (1) passe par un point (x, y) donné dans le plan, nous aurons une équation du second degré en ρ qui nous donnera les paramètres des deux coniques du système qui passent au point choisi. Si ces paramètres sont imaginaires, les coniques (1) seront imaginaires ; sinon elles seront réelles, car chacune d'elles aura alors une équation réelle, un centre et un point réels distincts l'un de l'autre. Tout revient donc à étudier la réalité des racines de l'équation du second degré en ρ,

$$(y^2-2xy+2ax)\rho^2+2(xy-ax-ay)\rho+(x-a)^2=0 ;$$

la conique en deux points D, E, il y aura deux directions asymptotiques AD′ et AE′ isogonales de AD et AE.

L'ellipse n'ayant pas de points à l'infini, la quartique ne coupera point le cercle circonscrit au triangle.

Dans le cas d'une hyperbole, la courbe sera analogue, mais elle coupera le cercle circonscrit en deux points qu'on obtiendra en prenant les isogonales des directions asymptotiques de l'hyperbole. Le cercle ne pourra couper l'hyperbole qu'en deux points qui donneront naissance à deux directions asymptotiques.

Enfin, quand la conique sera une parabole, la quartique aura une forme analogue avec deux directions asymptotiques; mais comme la parabole n'a qu'une direction asymptotique la quartique sera tangente au cercle circonscrit au triangle ABC.

4. Pour avoir ce lieu, il faut éliminer λ, μ, ν entre les équations suivantes :

équation de AM $\mu y = \nu z$, équation de la conique $\sqrt{\lambda x} + \sqrt{\mu y} + \sqrt{\nu z} = 0$,

équation de condition $\sqrt{\lambda x_0} + \sqrt{\mu y_0} + \sqrt{\nu z_0} = 0$.

Des deux premières on tire $\dfrac{\lambda}{4yz} = \dfrac{\mu}{xz} = \dfrac{\nu}{xy}$ ou $\dfrac{\lambda}{\frac{4}{x}} = \dfrac{\mu}{\frac{1}{y}} = \dfrac{\nu}{\frac{1}{z}}$.

En portant les valeurs proportionnelles à λ, μ, ν, dans la troisième équation, on a pour le lieu

$$\sqrt{\frac{4x_0}{x}} + \sqrt{\frac{y_0}{y}} + \sqrt{\frac{z_0}{z}} = 0.$$

La courbe isogonale sera $\sqrt{4x_0X} + \sqrt{y_0Y} + \sqrt{z_0Z} = 0$, conique inscrite dans le triangle ABC.

La quartique sera analogue à la précédente, et tout ce qui a été dit s'applique de nouveau en remplaçant le point P par le point $2x_0$, y_0, z_0 qui est le point de concours des tangentes à la quartique aux sommets du triangle dans le problème précédent.

CANS.

Remarque. — Par une transformation homographique qui change les points B et C en les deux points cycliques, les coniques de l'énoncé deviennent des paraboles ayant un foyer commun, le point A (dans sa nouvelle position) et passant par un point donné P.

Nous laissons au lecteur le soin de développer cette méthode et d'édifier la solution correspondante.

Bonnes solutions : MM. J.-D. Dufaut, école primaire supérieure d'Aiguillon; C. Grolleau, répétiteur général au lycée de Marseille; Pelvoisin.

---

917. — *Etant données deux quadriques* C *et* C′ *qui se coupent suivant deux courbes planes.* P *et* Q, *il est possible de leur circonscrire une infinité de quadriques* S. *Deux de ces quadriques* S_1 *et* S_2 *se coupent suivant deux courbes planes.*

1° *Les plans de ces deux courbes se coupent sur l'intersection de* P, Q, *et forment avec ces dernières un faisceau harmonique.*

2° *Par ces deux courbes passe une quadrique contenant la conique* P *et une contenant la conique* Q.

3° *Par ces deux courbes passent deux cônes du second degré. Ces cônes ont leurs sommets sur une droite fixe et restent circonscrits à une quadrique fixe, tangente aux plans* P, Q *et inscrite dans les deux cônes circonscrits à la fois aux deux quadriques données* C *et* C′.

Faisons une transformation homographique de la figure de manière que la conique Q se transforme en le cercle de l'infini ; les deux quadriques C et C' se transforment alors en deux sphères Γ et Γ', et la conique P devient le cercle Π situé dans le plan radical des deux sphères.

Il suffit alors de résoudre la question dans ce cas particulier, puisque les propriétés à démontrer se conservent dans une transformation homographique quelconque.

Toute quadrique Σ circonscrite à la fois aux deux sphères est de révolution autour de la droite Δ qui joint les centres ω et ω' des deux sphères ; deux de ces quadriques Σ_1 et Σ_2 se coupent suivant deux cercles C_1 et C_2 dont les centres sont sur la droite Δ et dont les plans sont perpendiculaires à Δ.

Coupons la figure par un plan passant par Δ ; les sphères Γ, Γ' sont coupées suivant deux grands cercles γ, γ' ; le plan radical Π suivant une droite π, axe radical des cercles γ, γ', qui rencontre Δ au point i ; les quadriques Σ_1 et Σ_2 sont coupées suivant des coniques σ_1, σ_2 bitangentes aux cercles γ, γ', les cordes de contact étant D_1, D'_1 pour σ_1, D_2, D'_2 pour σ_2. Les coniques σ_1, σ_2 se coupent en quatre points a, b, a', b' symétriques deux à deux par rapport à Δ. Les droites ab, $a'b'$ sont des diamètres des cercles C_1 et C_2.

La conique σ_1 étant bitangente aux cercles γ et γ', les cordes de contact D_1 et D'_1 forment un faisceau harmonique avec l'axe radical π et la droite de l'infini, donc D_1 et D'_1 sont équidistantes de l'axe radical, il en est de même pour D_2 et D'_2.

1. Le cercle γ étant bitangent aux coniques σ_1 et σ_2, les cordes de contact D_1 et D_2 sont conjuguées harmoniques par rapport à ab et $a'b'$: de même ab et $a'b'$ forment un faisceau harmonique avec D_1 et D_2. Comme D_1 et D'_1 d'une part, D_2 et D'_2 d'autre part sont équidistantes du point i, il en est de même de ab et $a'b'$. Par suite les plans des cercles C_1 et C_2 forment un faisceau harmonique avec le plan radical Π et le plan de l'infini.

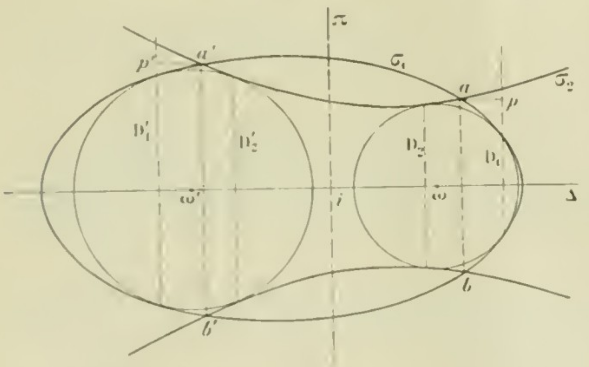

2. Par les quatre points a, b, a', b' on peut faire passer un cercle ayant son centre sur Δ : ce cercle en tournant autour de Δ engendre une sphère contenant les deux cercles C_1 et C_2. Cette sphère est la transformée homographique de la quadrique qui contient les coniques communes aux quadriques S_1 et S_2 et qui passe par la conique Q.

Soient c et d les points où l'axe radical πi rencontre les cercles γ et γ'. Par les six points a, b, a', b', c, d on peut faire passer une seule conique ayant pour axe Δ. En tournant autour de Δ, cette conique engendre une quadrique contenant les cercles C_1, C_2 et le cercle commun aux sphères Γ et Γ'.

3. Par les cercles C_1 et C_2 passent visiblement deux cônes de révolution ayant pour axe Δ et dont les sommets sont les points s_1 et s_2 où aa' et ab' rencontrent Δ.

Pour démontrer que ces cônes sont circonscrits à une quadrique fixe, il suffit d'établir que les droites aa' et ab' sont tangentes à une conique fixe ayant pour axe Δ.

Soient at, $a't'$ les tangentes menées des points a et a' respectivement aux cercles γ et γ', et ap, $a'p'$ les distances des mêmes points aux droites D_1 et D'_1. D'après une propriété bien connue des cercles focaux d'une conique, les rapports $\frac{at}{ap}$ et $\frac{a't'}{a'p'}$ sont égaux à l'excentricité de la conique. Comme $ap = a'p'$, on en conclut $at = a't'$. Donc les puissances des points a et a' par rapport aux cercles γ et γ' respectivement sont égales. Il en résulte immédiatement que les cercles γ_1 et γ_2 qui ont pour centres ω et ω' et qui passent par les points a et a' ont même axe radical que les cercles γ et γ'.

La droite aa' rencontre les cercles γ_1 et γ'_1 et l'axe radical πi aux points e, e', g. Comme on a $ga = ga'$ et $ga.ge = ga'.ge'$, on en conclut $ge = ge'$ et par suite $ae = a'e'$. Si on abaisse ωf et $\omega' f'$ perpendiculaires sur aa', on a $gf = g'f'$ et par suite la perpendiculaire menée par le point g à la droite aa' passe par le point o milieu de $\omega\omega'$.

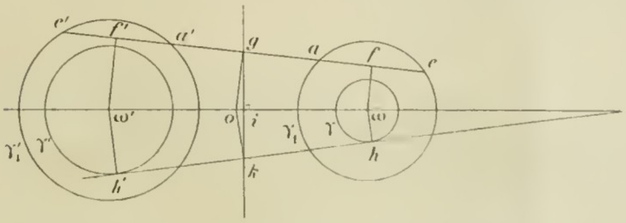

Il en résulte que aa' est tangente à une parabole ayant pour foyer le point o et pour tangente au sommet ig. On verrait de même que ab' touche la même parabole. On en conclut que les deux cônes qui contiennent les cercles C et C' sont circonscrits à un paraboloïde de révolution ayant pour foyer le point o et pour plan tangent au sommet le plan II.

Il nous reste à démontrer que ce paraboloïde est également inscrit dans les cônes circonscrits aux deux sphères Γ et Γ'. Soit hh' une tangente commune aux deux cercles γ et γ', rencontrant ig au point k; comme ωh et $\omega' h'$ sont perpendiculaires à hh', ok l'est aussi, et par suite hh' touche la parabole de foyer o et de tangente au sommet ig.

Observons en terminant que si l'on revient à la première figure, le paraboloïde est la transformée d'une quadrique qui touche les plans P et Q.

On rapprochera avec intérêt cette solution de la solution géométrique de la question **898** (N° de juin 1900, p. 530).

---

918. — *Si un triangle varie en restant inscrit dans une conique* C *et circonscrit à une autre conique* C', *la polaire de tout point de* C *par rapport au système des trois côtés du triangle passe par un point fixe.*

Prenons pour triangle de référence l'un des triangles indiqués et représentons par

$$yz + zx + xy = 0, \qquad avw + bwu + cuv = 0,$$

les équations ponctuelle et tangentielle des deux coniques C et C'. La seconde équation peut s'écrire

$$\frac{a}{u} + \frac{b}{v} + \frac{c}{w} = 0,$$

et donne immédiatement, en fonction d'un paramètre λ,

$$\frac{a}{u} = \lambda, \qquad \frac{b}{v} = 1 - \lambda, \qquad \frac{c}{w} = -1;$$

par conséquent les coordonnées d'une tangente quelconque à la seconde conique s'expriment rationnellement en λ par les formules très simples

$$u = \frac{a}{\lambda}, \qquad v = \frac{b}{1-\lambda}, \qquad w = -c.$$

Nous pouvons démontrer en passant que le pôle de cette droite par rapport aux trois sommets du triangle choisi décrit une droite fixe liée à ce triangle ; en effet, appelons u_0, v_0, w_0 les trois nombres précédents ; le pôle de la droite correspondante par rapport à la courbe impropre de troisième classe $uvw = 0$ a pour équation

$$uv_0w_0 + vw_0u_0 + wu_0v_0 = 0, \qquad \text{ou} \qquad \frac{u}{u_0} + \frac{v}{v_0} + \frac{w}{w_0} = 0,$$

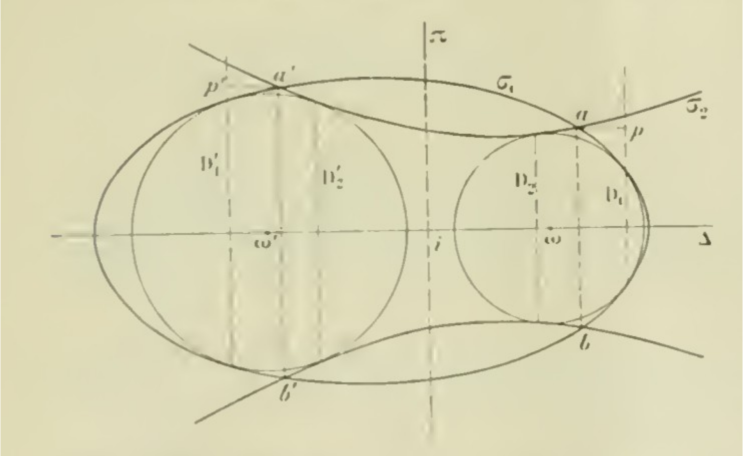
π
σ1
σ2
p'
a'
a
p
D'1
D'2
D2
D1
ω'
i
ω
Δ
b
b'

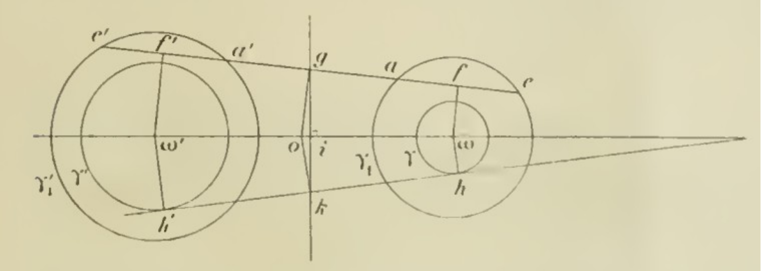
e'
f'
a'
g
a
f
e
ω'
o
i
γ1
γ
ω
γ'1
γ'
h
h'
k

& ,"

9(:69 9 92 9 ," & ! ; A
! 3!

' 0 E / "

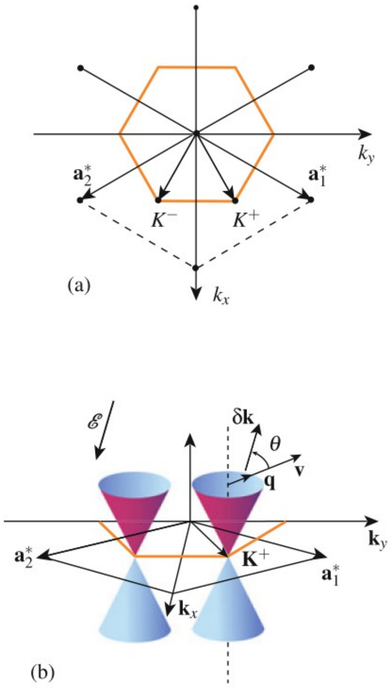
k_y
$\mathbf{a}_2^*$
$\mathbf{a}_1^*$
K^-
K^+
k_x
$\mathscr{E}$
$\delta\mathbf{k}$
θ
$\mathbf{v}$
$\mathbf{q}$
$\mathbf{k}_y$
$\mathbf{a}_2^*$
$\mathbf{K}^+$
$\mathbf{a}_1^*$
$\mathbf{k}_x$

(a)

(b)

Substituting P atoms for Si atoms therefore creates a semiconducting compound with majority electron doping, which is said to be *n*-doped. Note that the electron and hole densities are always related by (3.54).

In the same way, some Si atoms could be replaced by atoms of a trivalent element like B, which is an electron acceptor, introducing an acceptor level E_a above the top of the valence band. One can thus obtain a *p*-doped compound, i.e., doped with holes, with the electronic structure illustrated in Fig. 4.12b.

Note that the presence of positive charges can be detected directly by electron transport experiments with a magnetic field $B = \mu_0 H_\mathrm{a}$ perpendicular to the current. In the plane perpendicular to the field, the Laplace force deflects the carrier trajectories in the same direction whether they be positively or negatively charged. It follows that charges accumulate in the direction perpendicular to the current, and in the steady-state regime, this produces an electric field transverse to the direction of the current. The associated potential difference has the sign of the majority carriers in the material. This effect, called the *Hall effect*, can be used to determine the density and sign of the carriers in a semiconductor.

The industrial applications of semiconductors stem from this ability to control the density and *n* or *p* type of the carriers by doping. This in turn led to the development of microelectronics, where diodes and transistors were constructed by making junctions between *n*- and *p*-doped semiconductors. Another important step was to make transistors by applying an electric field to a semiconducting layer in order to control the level of doping, as we shall see for the specific example of graphene.

4.6 Electrons, Holes, and Transport in Graphene

As we saw in Chap. 3, graphene has the band structure of a zero gap semiconductor. We shall see here how the particular shape of the dispersion curves $E(\mathbf{k})$ shown in Fig. 3.19 thus leads to rather novel properties.

a. Consequences of the Graphene Band Structure

To begin with, consider the dispersion relations (3.60), taking E_0 as the zero energy. The experimental value $t \simeq 3$ eV is high, and we may restrict to eigenstates of energy $|E| \ll t$ if we are interested in low energy properties at experimentally accessible temperatures. Indeed, for neutral graphene, the conduction electrons and holes appear for energies close to $E_\mathrm{F} = 0$. The corresponding momenta lie near the six vertices of the Brillouin zone, to be reduced to two points K^+ and K^- with wave vectors $\mathbf{K}^\varepsilon$ with $\varepsilon = \pm$, called Dirac points (see box below). The four others are obtained from K^+ and K^- by translations along a vector of the reciprocal lattice $(\mathbf{a}_1^*, \mathbf{a}_2^*)$, as shown in Fig. 4.13a. Setting

$$\mathbf{k} = \mathbf{K}^\varepsilon + \mathbf{q} \, ,$$

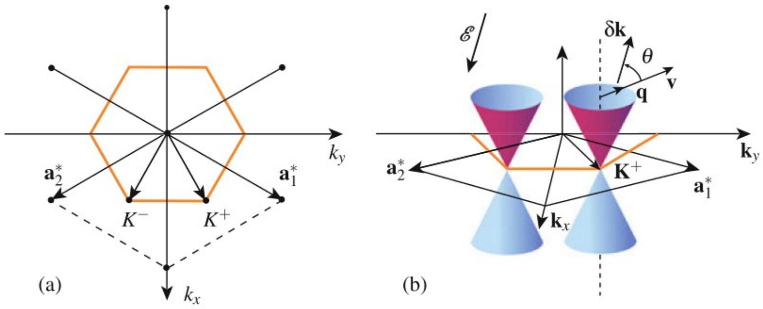

Fig. 4.13 (**a**) Hexagonal reciprocal lattice of graphene and its first Brillouin zone. The primitive cell $(\mathbf{a}_1^*, \mathbf{a}_2^*)$ is shown, as are the two Dirac points K^+ and K^- of this cell. (**b**) The conical dispersion relation $E(\mathbf{k})$ near the Dirac points is shown inside the same primitive cell. This is a zoom on Fig. 3.19 in the vicinity of the Dirac points. The real space velocity $\mathbf{v}$ of an electron with wave vector $\mathbf{K}^+ + \mathbf{q}$ is indicated

we may consider an expansion of (3.60) to first order in $\mathbf{q}$:

$$E_{\mathbf{k}} = \pm\frac{3}{2}at\,|\mathbf{q}| \ . \tag{4.74}$$

Figure 4.13b shows these dispersion relations, which are conical close to the Dirac points. The velocity of an electron in a state $E_{\mathbf{k}}$ is given by

$$\langle \mathbf{v}_{\mathbf{k}} \rangle = \frac{1}{\hbar}\nabla_{\mathbf{k}} E_{\mathbf{k}} = \pm\frac{3at}{2\hbar}\mathbf{n} = \pm v_{\mathrm{F}}\mathbf{n} \, , \tag{4.75}$$

where $\mathbf{n}$ is the unit vector normal to the constant energy circle at $\mathbf{K}^{\varepsilon} + \mathbf{q}$, which thus lies along $\mathbf{q}$ (see Fig. 4.13b). The electron velocity is aligned with its quasi-momentum $\mathbf{q}$ and has magnitude $v_{\mathrm{F}} = 3at/2\hbar$, which is independent of the electron state. Note that the zero curvature of the constant energy surface implies that *the effective mass vanishes*. Moreover, the dispersion relation (4.74) can be rewritten in the form

$$E_{\mathbf{K}^{\varepsilon}+\mathbf{q}} = \pm\hbar v_{\mathrm{F}}\,|\mathbf{q}| \ . \tag{4.76}$$

This dispersion relation provides a simple way of calculating the density of energy states $D_{\mathrm{S}}(E)$ per unit area and spin direction, when $|E| \ll t$. The number $D_{\mathrm{S}}(E)\mathrm{d}E$ of states with energies lying between E and $E + \mathrm{d}E$ is the number of electron states contained in the area $2\pi q\mathrm{d}q$ of reciprocal space, viz.,

$$D_{\mathrm{S}}(E)\mathrm{d}E = 2\pi q\mathrm{d}q \times (2\pi)^{-2} \times 2 \, ,$$

taking into account the existence of the two branches at points K^+ and K^-. We obtain

$$D_{\mathrm{S}}(E) = \frac{|E|}{\pi\hbar^2 v_{\mathrm{F}}^2} \ . \tag{4.77}$$

This density of states, which vanishes at $E = 0$, increases linearly with energy, a distinctly unusual behaviour compared with standard metals or semiconductors. A detailed calculation using the general form of the dispersion curves (3.60) gives Van Hove singularities for $D_S(E)$ when $E = \pm t$, energies at which the dispersion curves have saddle points, and a constant limiting value for the extrema of the energy $E = \pm 3t$, where a parabolic dispersion relation and finite effective mass are obtained (see Fig. 4.14).

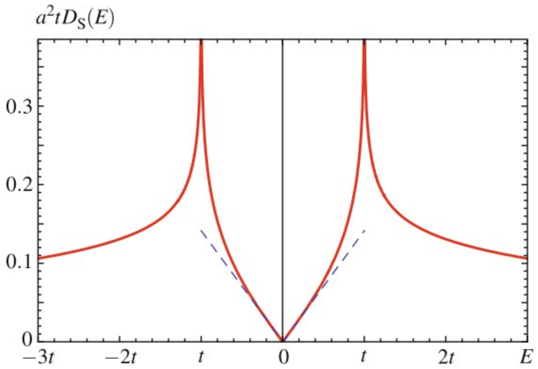

Fig. 4.14 Density of states per unit area for graphene, obtained numerically from the dispersion curves given in (3.60). For $|E| < t$, we recover the linear relation (4.77) close to the Dirac points. Figure courtesy of Montambaux, G.: Laboratoire de Physique des Solides, Orsay, France

Dirac Points and Relativistic Physics

The similarity between (4.76) and the dispersion relation for photons, together with the vanishing effective mass, has inspired an analogy between the equations of motion of the electrons in graphene around the points K^+ and K^- and those of massless quantum particles. The speed v_F given in (4.75), which is independent of the electron energy, plays the role of the speed of light. For this reason, the points K^+ and K^- are commonly referred to as Dirac points today. For the same reason, the surfaces representing the energy dispersion relation are called Dirac cones. The states in the conduction band then represent particles, while those of the valence band are identified with antiparticles.

Note that the degeneracy between the two conical valleys introduces a complication which we have hidden in (4.76), where the sign ± for the energy is obviously not related to the one used to characterise the Dirac points $\mathbf{K}^{\varepsilon}$.

This analogy with 2D relativistic quantum physics can be taken a long way. As in this last case, where the particle spin is represented by 2×2 Pauli matrices, in the case of graphene, these matrices take into account the sublattices A and B. In this representation, the Hamiltonian yielding the conical dispersion relations can then be written

$$\hat{H} = \hbar v_F \, |\mathbf{q}| \, \hat{\boldsymbol{\xi}} \ .$$

The operator $\hat{\boldsymbol{\xi}}$, with eigenvalues ± 1, describes the projection of the sublattice 'spin' in the direction of propagation of the electron in the graphene. It is analogous to helicity in particle physics. Helicity is the projection of the real spin of the particle onto its direction of propagation. This analogy leads us to consider massless charged particles, something unknown in particle physics, since the only known massless particles, photons or neutrinos, are electrically neutral.

b. Electron Transport

At $T = 0$, there are in principle no carriers, and one therefore expects zero conductivity for graphene. However, for $T \neq 0$, electrons with surface density n_{e} are excited in the conduction band, with a resulting hole surface density n_{h} in the valence band, and graphene becomes a conductor. Given the electron–hole symmetry of the band structure, charge conservation implies that $n_{\mathrm{e}} = n_{\mathrm{h}}$ and that the chemical potential remains fixed at $\mu = E_{\mathrm{F}} = 0$.

> **Question 4.5.** Show that the number of carriers thermally excited varies as T^2 and that the same is true of the electronic specific heat.

Insofar as the dispersion relation is not quadratic, and since the effective mass is zero, the Drude relation is not applicable, in contrast to the case of standard semiconductors. We must therefore consider a more general expression for the electrical conductivity, taking into account the total current of the occupied electron states, in the form

$$\mathbf{j} = -2\sum_{\mathbf{q}} e\mathbf{v}(\mathbf{q}) f\big(E(\mathbf{q})\big) , \tag{4.78}$$

where $f(E)$ is the Fermi function and the factor of 2 is due to spin degeneracy. This current is clearly zero when there is no electric field. If we assume, as discussed in this chapter, that the electrons are subjected to viscous friction, in the presence of a field $\mathscr{E}$, the stationary solution implies a shift $\boldsymbol{\delta}\mathbf{k}$ of the electronic states given by

$$\hbar\boldsymbol{\delta}\mathbf{k} = -\tau_{\mathrm{e}} e\mathscr{E} = \hbar\boldsymbol{\delta}\mathbf{q} , \tag{4.79}$$

where τ_{e} is the collision time, assumed to be the same for both electrons and holes, since the problem is symmetric. Taking into account the displacement $\boldsymbol{\delta}\mathbf{q}$ of the Fermi surface in (4.78), we obtain

$$\begin{aligned}\mathbf{j} &= -2e\sum_{\mathbf{q}} \mathbf{v}(\mathbf{q}+\boldsymbol{\delta}\mathbf{q}) f\big[E(\mathbf{q})\big] = -2e\sum_{\mathbf{q}} \mathbf{v}(\mathbf{q}) f\big[E(\mathbf{q}-\boldsymbol{\delta}\mathbf{q})\big] \\ &= 2e\sum_{\mathbf{q}} \mathbf{v}(\mathbf{q})\frac{\mathrm{d}f\big[E(\mathbf{q})\big]}{\mathrm{d}E}(\boldsymbol{\nabla}_{\mathbf{q}}E)\boldsymbol{\cdot}\boldsymbol{\delta}\mathbf{q} = 2e\hbar\boldsymbol{\delta}\mathbf{q}\sum_{\mathbf{q}} v_x^2(\mathbf{q})\frac{\mathrm{d}f\big[E(\mathbf{q})\big]}{\mathrm{d}E} .\end{aligned}$$

Equation (4.79) was used in the last equality, taking $\mathbf{x}$ to be a unit vector in the direction $\mathbf{j} \parallel \mathscr{E} \parallel \boldsymbol{\delta}\mathbf{q}$. For temperatures such that $k_{\mathrm{B}}T \ll t$, we can restrict the dispersion relations to their conical part given by (4.76). Setting $\theta = (\mathbf{q}, \boldsymbol{\delta}\mathbf{q})$ as shown in Fig. 4.13b, this leads to

$$\mathbf{j} = -2e^2\tau_{\mathrm{e}} v_{\mathrm{F}}^2 \mathscr{E} \sum_{\mathbf{q}} \cos^2\theta \,\frac{\mathrm{d}f\big[E(\mathbf{q})\big]}{\mathrm{d}E} . \tag{4.80}$$

The sum can be replaced by an integral over energy, after integrating over the angular orientation of $\mathbf{q}$ at given energy E, viz., $\int_{-\pi}^{\pi}\cos^2\theta(\mathrm{d}\theta/2\pi) = 1/2$. We thus obtain

$$\mathbf{j} = e^2\tau_e v_F^2 \mathscr{E}\int_{-\infty}^{\infty} -\frac{\mathrm{d}f}{\mathrm{d}E} D_S(E)\mathrm{d}E\,. \tag{4.81}$$

The 2D conductivity can then be written in the form

$$\sigma = \frac{2}{\pi\hbar^2}e^2\tau_e\int_0^{\infty} -\frac{\mathrm{d}f}{\mathrm{d}E}E\mathrm{d}E = \frac{2}{\pi\hbar^2}e^2\tau_e k_B T\int_0^{\infty} -\frac{\mathrm{d}f}{\mathrm{d}x}x\mathrm{d}x\,, \tag{4.82}$$

where $x = E/k_BT$. The integral has the form

$$\int_0^{\infty}\frac{x\mathrm{d}x}{\cosh^2 x} = \ln 2\,,$$

so

$$\boxed{\sigma = 2\frac{\ln 2}{\pi}\frac{e^2}{\hbar^2}\tau_e k_B T}\,. \tag{4.83}$$

We thus expect zero conductivity at $T = 0$, and *growing linearly with temperature*. Note that, in two dimensions, the resistivity $\rho = 1/\sigma$ corresponds to a current density $\mathbf{j}$ per unit width of material (not per unit area), so ρ has the dimensions of a resistance (and the symbol $\Omega/\square$ is sometimes used to express the units of ρ).

c. Doping by Field Effect

It is certainly an interesting fact that graphene is metallic at room temperature because carriers are thermally excited. However, for applications in electronics, it is essential to be able to construct active components producing some specific response to an applied voltage. One of the most widely used approaches with semiconductors is to make a metal oxide semiconductor capacitive structure, called a MOSFET (MOS field effect transistor), to control the doping of the semiconductor through an electric field obtained by applying a voltage to the metal gate.

This effect can be illustrated by the equivalent structure used for graphene, shown schematically in Fig. 4.15a. In this case, the gate is a doped Si substrate, oxidised at the surface to produce a film of SiO_2 with a thickness d of a few hundred nanometers. The graphene layer constitutes the second plate of the capacitor. When a voltage V_G is applied across the gate and the graphene, the charges appearing on the capacitor plates dope the graphene. If ε is the dielectric constant of the oxide layer, the capacitance $C = \varepsilon\varepsilon_0/d$ per unit area of the junction induces a charge

$$\boxed{n = \frac{\varepsilon\varepsilon_0}{ed}V_G}\,. \tag{4.84}$$

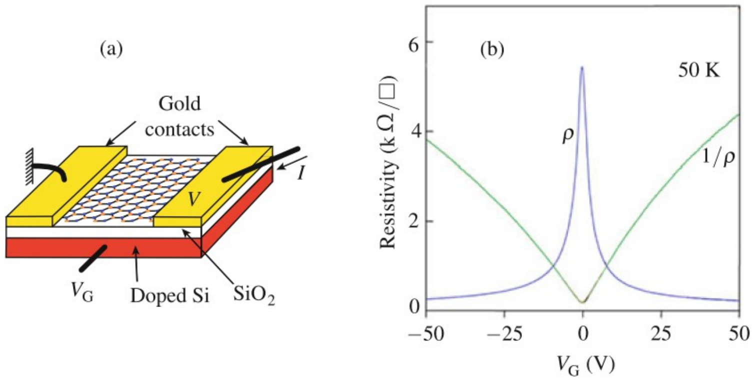

Fig. 4.15 (**a**) Semiconductor–oxide–graphene structure used to dope the graphene sheet by varying the voltage V_G applied to the semiconducting gate. (**b**) Dependence of the resistance and conductivity of graphene on V_G. Figure courtesy of A.K. Geim, from results published by Morozov, S.V., *et al*., Phys. Rev. Lett. **100**, 016602 (2008)

The surface density n of the carriers induced in graphene by an electric field thus varies linearly with the gate voltage V_G. We therefore observe, e.g., in Fig. 4.15b, that the resistivity of graphene falls off rapidly as a function of the gate voltage.

Question 4.6. 1. Which experimental observations in Fig. 4.15b can be explained by the approach discussed in Sects. 4.6b and c? Which cannot be explained? Compare the density of carriers induced experimentally by the field effect with the one obtained in Question 4.5 at thermodynamic equilibrium at 50 K ($\varepsilon = 4$, $d = 300$ nm).
2. What becomes of the chemical potential μ of the carriers in this experiment?

Note that the observations made with the atomic layer of graphene are similar to those carried out on silicon MOSFET transistors, which are made by depositing a metal film on a much thicker semiconducting layer whose carrier density is controlled by the voltage across the metal gate. In the case of graphene, the change in the resistance induced by a voltage $V_G = 10$ V is of the order of a factor of 10, which is a significant transistor effect. This type of structure has been used over the past few years to demonstrate many novel physical properties in graphene, suggesting applications for graphene or graphene multilayers in nanotechnology in the near future.

Question 4.7. What experiment could determine the sign of the injected carriers? How could one check experimentally whether the graphene energy spectrum does indeed correspond to that of the Dirac cones?

& ,/ 9(:69 9 92 9, K C.
;< ; !? * F 8 BS 8 3! ' / 0

Light Deflection

One of the earliest calculations of the deflection of light from a distant source by a foreground gravitational field is due to Henry Cavendish around 1784. Unfortunately, the only evidence that he did such a calculation is on an "isolated scrap" of paper. Will [Wil] offers some details of this story and attempts to reproduce Cavendish's calculation of the bending angle of light passing near the surface of an attracting mass.

The next closest calculations seems to have been in an 1801 paper by Johann von Soldner: "Ueber die Ablenkung eines Lichtstrals von seiner geradlinigen Bewegung, durch die Attraktion eines Weltkörpers, an welchem er nahe vorbeigeht" ["On the Deviation of a Light Ray from its Motion along a Straight Line through the attraction of a Celestial Body which it Passes Closeby"]. This article appeared in "Berliner Astronomisches Jahrbuch für das Jahr 1804" [Sol]; see Jaki's English translation [Jak]. We note that initially von Soldner's paper calculates — using Newtonian mechanics — the bending angle formula for light emitted tangentially from the surface of a gravitating body and travels off to infinity. When applying his formula, however, he reverses the light ray's path, so it comes in from infinity and "grazes" the attracting body [Jak, pp. 945-947] (cf. [Wil, p. 414, Sec. III]). Von Soldner, who later became director of the Munich observatory and a foreign member of the Royal Astronomical Society, figured out: "Thus when a light ray passes by a celestial body, it will, instead of going on in a straight direction, be forced by its attraction to describe a hyperbola whose concave side is directed against the attracting body." He showed that a light ray grazing the sun would be deflected by 0.84 seconds of arc and concluded: "If one could observe the fixed stars very close to the sun, then one would have to take this very much into account." Von Soldner realized that he was far ahead of his time when he apologized at the end: "Hopefully, no one would find it objectionable that I treat a light ray as a heavy body." As a consequence, "Soldner's work on the deflection of light had no influence either on practice or on theory" [Jak].

Even earlier, John Michell in 1783 [Miche] and Pierre-Simon Laplace in 1796 [Lap] independently investigated light being subjected to gravity, but in the context of a body acting on the very light it emits. They both basically arrived at a notion resembling a "black hole," that is, a sufficiently massive star can capture its own light, leaving the star invisible. Laplace stated his result initially without proof, but provided the proof in 1798; see Hawking and Ellis [Haw-E, p. 365] for an English translation. It seems, however, that Michell and Laplace did not write about the situation of light coming from a distant source and undergoing deflection as it passes near a gravitating body. We note, though, that it was Michell and Laplace who inspired the works of Cavendish and von Soldner, respectively (e.g., [Wil, p. 413] and [Jak, pp. 935–936]).

In hindsight, it is slightly surprising that Newton himself did not seem to have studied light deflection by gravity. All the ingredients were there with his gravitational dynamics and the corpuscular theory of light, and the mathematics involved was straightforward compared to other problems he solved successfully. His famous Query 1 of the *Opticks* (1704):

"Do not Bodies act upon Light at a distance, and by their action bend its Rays; and is not this action strongest at the least distance?" [New] is phrased in such a general way that today we tend to assume that it includes gravitational bending as well. This is not the case, though, according to Jaki, who comments [Jak]: "He [Newton], of course, had reflection, refraction and diffraction in mind."

Maybe it is equally surprising that it seems to have taken more than 100 years after von Soldner until somebody else dealt again quantitatively with the problem of light deflection. Even more so, when one learns that a famous contemporary of von Soldner, the physicist-philosopher Georg Christoph Lichtenberg wrote in the year 1796, "Wenn das Licht trotz seiner Geschwindigkeit noch eine Schwere hat, so würde doch so etwas wie Refraktion am Horizonte erscheinen müssen, weil es von der Erde stark gezogen wird" ["If light despite its velocity also had gravity, then something like refraction would have to occur at the horizon, because it would be attracted heavily by the earth"] [Lic].

We know today that the value of the deflection angle that Cavendish and von Soldner calculated is too small by a factor of two.[1] About a century after von Soldner's work, Albert Einstein arrived at a similar numerical value in his first attempt to determine the deflection angle — based on the equivalence principle. Einstein directly addressed the influence of gravity on light in a letter entitled: "Über den Einfluß der Schwerkraft auf die Ausbreitung des Lichtes" ["On the Influence of Gravitation on the Propagation of Light"] [Ein11]; see [Lor-EMW, p. 99] for an English translation of [Ein11]. He predicted that a light ray passing close to a "celestial body" might suffer a deflection towards the side of decreasing gravitational potential. Einstein found a numerical value of 0.83 seconds of arc for the deflection angle of light grazing the sun. In this paper, Einstein also stated that he wished astronomers to investigate this question: "Es wäre dringend zu wünschen, daß sich Astronomen der hier aufgerollten Frage annähmen, auch wenn die im vorigen gegebenen Überlegungen ungenügend fundiert oder gar abenteuerlich erscheinen sollten" ["It would be highly desirable that astronomers take up the question raised here, even if the considerations should seem to be insufficiently founded or entirely speculative"].

By 1915 Einstein succeeded in deriving the correct expression for the light deflection angle with the presentation of the General Theory of Relativity [Ein16, Ein22]. For a light ray passing at a distance R from an object of mass M he found the following deflection angle[2] — see Figure 1.1:

$$\hat{\alpha} = \frac{4GM}{c^2 R}, \tag{1.1}$$

where G is the constant of gravity and c is the speed of light. With the solar values for

---

[1] Will [Wil] compares Cavendish and von Soldner's determinations of the light deflection angle, and includes Einstein's correction to twice the value.

[2] Notice that the bending angle $\hat{\alpha}$ is independent of the light's wavelength and depends on the impact distance and deflector's mass. See Section 3.1.5.

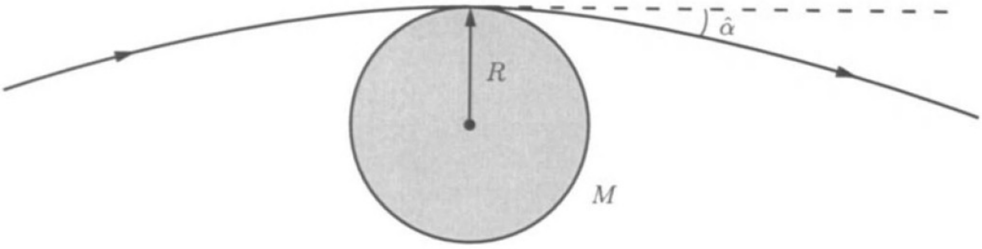

Figure 1.1. Light from a distant source experiences a deflection through angle $\hat{\alpha}$ by a mass M. The angle $\hat{\alpha}$ was calculated incorrectly by Cavendish (1784) and von Soldner (1801) since they used Newton's theory of gravity (see [Wil] for details). It was not until the formulation of General Relativity in 1915 that the correct formula for the bending angle was established by Einstein — see equation (1.1).

radius and mass he obtained

$$\hat{\alpha} = \frac{4GM_\odot}{c^2 R_\odot} \simeq 1.''74.$$

Note that since the light ray grazes the sun, we assume $R \simeq R_\odot$. In Einstein's own words: "Ein an der Sonne vorbeigehender Lichtstrahl erfährt demnach eine Biegung von 1.7 Bogensekunden" ["A light ray grazing the sun therefore experiences a deflection of 1.7 arcseconds"]. We refer readers to [Ray-BB, pp. 59–60] for a derivation of (1.1).

In the following years there were several papers dealing theoretically with light deflection, among them a short letter by Oliver Lodge about the "Einstein deflection" in which he made the point that "it is not permissible to say that the solar gravitational field acts as a lens, for it has no focal length" [Lod]. He pointed out that one would get a focal line starting at "a place seventeen times the distance of Neptune."

Multiple Imaging

While investigating the birth of general relativity, Renn, Sauer and Stachel [Ren-SS] discovered that in notes dated to the spring of 1912 Einstein explored gravitational lensing by a star. He showed that a distant star lying close to the line of sight to a foreground star will be imaged twice, with a faint secondary image increasing the total measurable intensity of the background star [Ren-SS, pp. 185, 186].[3] Einstein did not publish these results until 1936 [Ein36]. The possibility of double images was also mentioned in 1923 by Sir Arthur Eddington in his book *Space, Time and Gravitation* [Edd, p. 133–134]. This effect was also quantitatively studied by Chwolson[Chw] in 1924. Chwolson's article, "Über eine mögliche Form fiktiver Doppelsterne" ["On a possible form of fictitious double stars"], even mentions the "reversed mirror image" (i.e., the negative parity of the second image) and cautioned at the end: "Ob der hier angegebene Fall eines fiktiven Doppelsterns auch wirklich vorkommt, kann ich nicht beurteilen" ["Whether the here mentioned case of a fictitious double star really exists I cannot judge"].

---

[3] In the same notes, Einstein also determined the 'lens equation' for the system, that is, the equation whose solutions give the image positions [Ren-SS, p. 185].

Einstein Rings

In 1936 a Czech engineer named Rudi Mandl approached various scientists (including Einstein) with his idea that a foreground star may act as a "gravitational lens" for light coming from a background star. After Mandl's urging, Einstein wrote a short article for *Science* on the matter, which was published in 1936 [Ein36]. This article was clearly based on Einstein's 1912 notes on the subject [Ren-SS, p. 186]. The paper included the new result that at perfect alignment a background star undergoing lensing by a foreground one will appear as a very bright ring (hence the name "Einstein ring"). However, Einstein concluded that the size of the ring would be far too small to be resolvable, and that the probability of such a configuration was extremely small. He was pessimistic about a possible detection: "Of course, there is not much hope of observing this phenomenon directly." According to [Ren-SS, p. 186], Einstein was so skeptical that in his letter to the editor of *Science* he wrote: "Let me also thank you for your cooperation with the little publication, which Mister Mandl squeezed out of me. It is of little value, but it makes the poor guy happy." Fortunately, history would favor Mandl.

Giant Luminous Arcs, Microlensing, and Blending of Light

Einstein's article inspired (prior to its publication) his Princeton colleague Henry Norris Russell to write the first popular article on gravitational lenses: In *Scientific American*, he presents lensing as a perfect test of general relativity. Russell invented a hypothetical observer on a planet around the White Dwarf companion of the bright star Sirius to illustrate the relativistic effect of bending of light and to make it "conspicuous to the immediate gaze" [Rus]. In this scenario, Russell seems to be the first to study the lensing effect of extended sources, and his figures even show the typical and now well-known images of giant arcs — magnified and tangentially extended images of a background source that look like crescents.

Russell also noticed that star-star lensing (today we call this galactic microlensing) would produce a change of the observable brightness of the background star due to the relative motion between source, lens and observer, and that such an event would typically last for about a day. He mentioned the problem of blending of light due to the brightness of the lensing star[4] as well as the desirability of an early warning system[5] — all basic issues that today's observers have to face (see Section 5.5 on galactic microlensing). In the following year, Tikhov [Tik] determined the intensities of the two images of a star lensed by another star (cf. [Ein36]).

---

[4] "Blending" here relates to the fact that we do not measure the light of the lensed star alone; but that there is additional light, in this case from the lensing star itself. As a consequence, the "baseline" of the light curve is higher than in the case in which we see the background star alone; thus some properties in a "blended" system may be different from those in an isolated one, e.g., the magnification as a function of time is changed (the amplitude is smaller), and the light curve can be chromatic if the stars involved have different colors.

[5] The idea is to observe a star very often in order to be able to notice a possible change due to gravita-

Galaxies as Lenses

Fritz Zwicky — indirectly influenced by Mandl as well — point out in 1937 that "extragalactic *nebulae* offer a much better chance than *stars* for the observation of gravitational lens effects" [Zwi37a]. So he was the first to recognize the potential of galaxies as gravitational lenses. Zwicky reiterated and emphasized these points about lensing in his 1957 book *Morphological Astronomy* [Zwi57]. Aside from this book, there seems to be a 25-year silence in the field of gravitational lensing following the short, productive period in the mid-1930s.

Lens Renaissance in the 1960s

It was not until the early 1960s when there was another sudden outburst of papers that considered the theoretical aspects of gravitational lensing. In particular, Liebes and Refsdal derived in detail some of the basic equations of gravitational lensing with suggestions for applications.

Liebes. In 1964, Liebes [Lie] gave many applications, including discussions of galactic star-star lensing, lensing of stars in the Andromeda galaxy, lensing by "unobservable bodies," and lensing of galaxies. Liebes critically discussed the probability of detecting these effects, and he even showed figures of "Giant Luminous Arcs."

Refsdal's Method: Time Delay and Hubble's Constant. Refsdal derived in 1964 the relations among the distances between observer, lens, and source, and the angles involved [Ref64a]. He showed how one can determine the mass of the lensing object, and he pointed out that "from theory it is reasonable to believe that the deflection of light is independent of frequency." He cautions though that this "... has not been tested by experiment with any accuracy."

In Refsdal's second lensing paper [Ref64b], he analyzed the lens effect by a galaxy on a light source lying far behind. He mainly had supernovae in mind as distant point-like sources. But he even notes that "star-like objects with intense emission ... have recently been discovered" — today we call them quasars. Refsdal realized the potential usefulness of quasars immediately, and he noted that with these objects "observations at greater distances than with supernova will be possible."

Refsdal found that for a double image there is a difference in arrival time between the two paths connecting source and observer. And since this time delay is inversely propor-

---

tional lens magnification early enough so that one is able to notify other observers and make more frequent observations possible.

tional to the Hubble constant, he was the first to note that measuring time delay in a lensed system is a way to determine the Hubble constant.

In two more papers, Refsdal suggested the use of gravitational lensing as a test for cosmological theories [Ref66]. A few years later with one of the first "numerical" projects on lensing [Ref70], he employed the ray-tracing technique to study the propagation of light in universes with inhomogeneous mass distributions. He investigated the influence of lensing on "standard candles" and emphasized the importance of magnification bias.[6]

Lensing, Cosmology, and Quasars

Two consequences of lensing with possible cosmological interest were investigated in the late sixties to early seventies: statistical effects of a lens population on background sources, and the possibility that the newly discovered quasars were just active galaxies, strongly magnified by lenses.

Statistical lensing due to local matter inhomogeneities can change the position, shape, and brightness of background sources, compared to a corresponding homogeneous universe [Gun67a, Gun67b, Noo]. If the universe is filled with a population of compact objects, identifying and counting "double images" on different angular scales could help us detect these objects. Press and Gunn showed that this method can work on mass scales from 10^4 to about 10^{17} solar masses [Pre-G].

The once popular hypothesis that quasars are highly magnified Seyfert galaxies was looked at by different authors. The basic conclusion was that although moderate magnification could be involved, lensing alone could not explain the whole quasar phenomenon. The Barnothys proposed a variety of lensing ideas and scenarios. Although some of their ideas are quite bizarre, they are often not given credit for pursuing the idea of gravitational lensing during a period in which the majority of astrophysicists were rather skeptical about this phenomenon [deS, San, Barn, Barn-B].

Specific Lens Models

The uniform transparent lens was explored in 1972 by Clark [Cla]. He gave an elementary discussion of the gravitational deflection of light due to radially and cylindrically symmetric masses.

Bourassa and colleagues developed a complex formalism to describe lensing situations and investigated a spheroidal gravitational lens, calculating bending angles and determining formulae for intensification, distortion and orientation of the images [Bou-KN]. Using this formalism, the theory of transparent gravitational lenses was fully developed. General expressions for the bending angle of a light ray around a bounded elliptical mass distribution were given, and the locations, magnifications, distortions and orientations of images could be calculated [Bou-K]. Cooke and Kantowski calculated time delays for multiple-imaged quasars. These delays are due to the differences in geometrical path lengths and gravitational potential along the light rays [Cook-K].

---

[6] Magnification bias is the effect whereby a fraction of all light sources are highly magnified due to gravitational lensing. Although the absolute fraction of them is small, there is usually a large reservoir of faint sources which can be magnified. So, in any sample of sources with a certain cut-off in apparent brightness, the fraction of sources that are affected by lensing is higher than in a "volume-limited" sample.

1.2 Detecting Gravitational Lens Effects

Here we list a number of first observations on various aspects of light deflection. Discoveries as well as failures in the search for certain lensing signatures are described, starting in the early years of the twentieth century. Observational phenomena are described in more detail in Chapter 5.

Measuring Light Deflection at Solar Limb

There were various plans to test Einstein's (wrong!) 1911 prediction for the deflection angle at the solar limb. In 1912 an Argentinian expedition was to observe the solar eclipse in Brazil, but was rained out. Another attempt was made in 1914, when Erwin Freundlich led a group to measure the positions of stars near the sun during the solar eclipse on the Russian Crimea. However, when the observers were in Russia, World War I broke out and they were captured by the Russian soldiers and released only after the eclipse (see, e.g., [Bre, Pai]). So, fortunately for Einstein, the measurement of the light deflection by the sun was not made before the correct value was published.

With the full theory of general relativity [Ein16], the correct value for the light deflection was published. In 1917 the then Astronomer Royal Frank Dyson suggested using the solar eclipse in May 1919 for the observational confirmation. This eclipse was particularly well suited, since it occurred in the Hyades constellation with many bright stars in the field. In two famous expeditions to Sobral (Brazil) and Principe Island (Gulf of Guinea, Africa), Eddington measured the displacement of stars close to the sun during the solar eclipse of May 29, 1919 to within 30% of the value predicted by Einstein (the deflection angle has been confirmed very recently with an accuracy of better than 0.02% [Leb] !) .

This experiment became the first and most famous test of general relativity. The confirmation of the result of general relativity — announced by Dyson and Eddington at a joint meeting of the Royal Society and the Royal Astronomical Society in London on November 6, 1919 — was the basis of Einstein's huge popularity starting in the 1920s.

Vain Attempts for other Lensing Effects

As early as 1923 there were attempts to measure the effect of star-star lensing. According to a footnote in [Zwi37b], E.B. Frost, then director of Yerkes Observatory, outlined a program to look for the lensing effect. The lack of any subsequent paper on this project leads us to

2. F * ! ! ' ,/ ,/0 A } 2P=== * ! ! y=

<H ' / 0
< H * . $ 3 3 .; K *
! ,N . ! (. B F * ! 8 @ !
! (B C B 3 @ * . , O,2 > *
F : ; F 8 !;
.;,Q < 33 ! ,8 (K .
! K $ /, O Q• K B | A OK BQ C ; *
! | 3 3 R . !
! ; * . !; ! * ;
C 33 ; @ * @ * ;| 3 @
; ! , } , . !; OK B@ 33 ,/ ,/0Q OC @ 3 / /@ &&&Q
! @ . ! . * ; ! .
: ; (;@ ! ,< ; 3 .; . !;@
@ ! ! @ . ! .; ! . ;
3 . * ! ! ! . !; , P !
; } ! . ! ! .; $/ ! ! ! ! ,& !
. * H ! * ; @ ! * B * ;
, ! } ! ! 3 } !
! ,P 3 ;@ ! ! .4 . & ; *;
. !; , ( ) @ , ! | B ! !
3 ;, OK BQ

!

3 ##3. ! 3 #! . #3 . #" ,0# ! # %$0/y I3

3 ## ! ! . . B ! #

,/

9(:69 9 92 9 ,/
%$0/-(-K . - ! -A3 ! - -! -K $ / } 3

3 ##3. ! 3 #! . #3 . #" ,0# ! # %$0/y I3

. 9(:69 9 92 9 ,/ ! A $ / 3!

< H * . $ 3 3 .; K *
! ,N . ! (. B F * ! 8 @ !
! (B C B 3 @ * . , O,2 > *
F : ; F 8 !;
.;,Q < 33 ! ,8 (K .
! K $ /, O Q• K B | A OK BQ C ; *
! | 3 3 R . !
! ; * . !; ! * ;
C 33 ; @ * @ * ;| 3 @
; ! , } , . !; OK B@ 33 ,/ ,/0Q OC @ 3 / /@ &&&Q
! @ . ! . * ; ! .
: ; (;@ ! ,< ; 3 .; . !;@
@ ! ! @ . ! .; ! . ;

3 . * ! ! ! . !; , P !
; } ! . ! ! .; $/ ! ! ! ! ,& !
. * H ! * ; @ ! * B * ;
, ! } ! ! 3 } !
! ,P 3 ;@ ! ! .4 . & ; *;
. !; , ( ) @ , ! | B ! !
3 ;, OK BQ

9(:69 9 92 9 ,/ . .- .- R/"((((((K 3!

' % %$

! 9(:69 9 92 9 ,/
< -2 - -> - - -F -: ;- - -: - 3!

2 ! ; @ .; . !;
; 3 .;
8; K !
8 @ $

9(:69 9 92 9 ,/ 3!

! !

data 08.09.2019

94) Light Deflection

Source: 20190907/ Singularity Theory and Gravitational Lensing by Arlie O. Petters/ Joachim .... pdf

Pág. 04 (27) - Light Deflection

a) Henry Cavendish (1784) - First (isolated scrap)
b) Johann von Soldner (1801)

"Ueber die Ablenkung eines Lichtstrals von seiner geradlinigen Bewegung, durch die Attraktion eines Weltkörpers, an welchem er nache (nahe) vorbeigeht".
("On the Deviation of a Light Ray from its Motion along a Straight Line through the attraction of a Celestial Body which it Passes Closeby")

This article appeared in
"Berliner Astronomisches Jahrbuch für das Jahr 1804"
Pág. 945 - 947

Einstein utilizou isto em sua teoria (OH!!!) novidade?!

data 08.09.2019
S T Q Q S S [✓ D]

1905 - 1801 = 104 anos
(novíssima descoberta!!!)

Na realidade Einstein utilizou os estudos precedentes do físico neerlandês Hendrik Lorentz e as descobertas de Cavendish e Soldner em sua teoria da relatividade, além de outros cientistas.

Vejamos:

a) Système de coordonnées Soldner

b) La Constante Ramanujan-Soldner

Sources: (others)
https://zenodo.org/record/1424375/files/article.pdf

Obs: Este arquivo não está mais disponível neste website, mas eu tenho uma

data 08.09.2019
S T Q Q S S [✓ D]

cópia armazenada na pasta
20190907/Important/Johann
von Soldner/
article.pdf

Author: P. Lenard
Pág. 593 – 604
20 Juli 1921.

obs (off). Johann Georg von Soldner

https://en.wikisource.org/?
curid=755966.

other (source)
https://archive.org/details/
bub_gb_PN43AAAAMAAJ/page/
n167
Pág 161-172 (169-180)
Copy: 20190907/Important/Johann
von Soldner/

& ,/

> F ! 3 .; * *
2 C B }

H * 3 # / "% *

9(:69 9 92 9 ,/ / "% 3!

F >

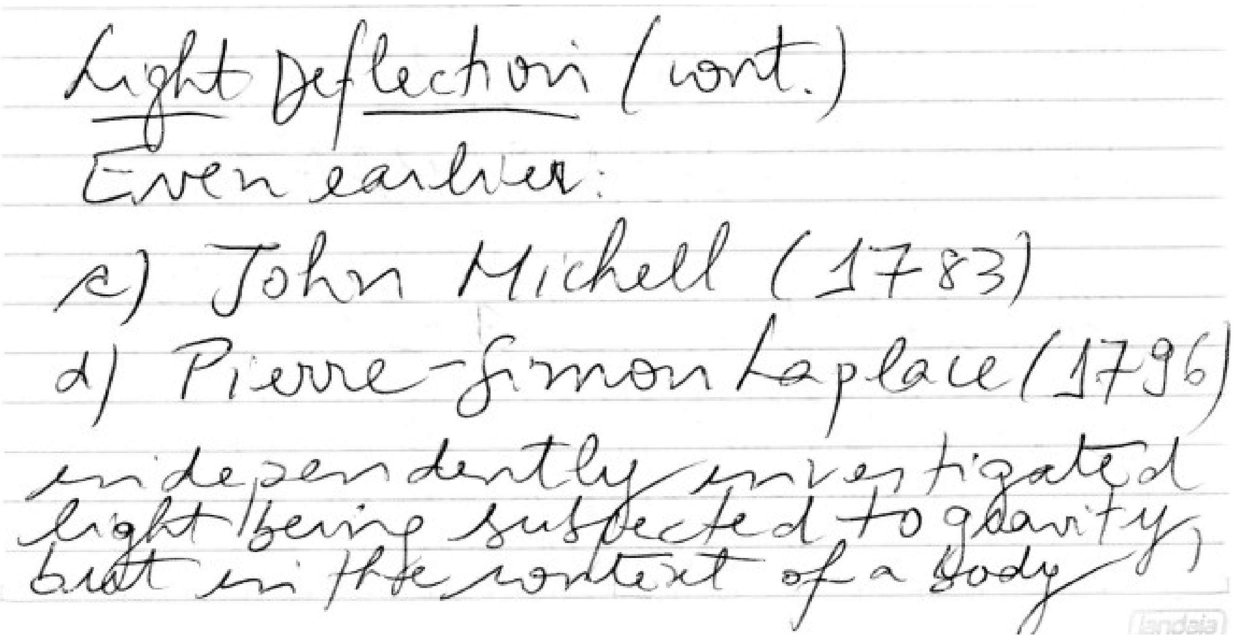

Light Deflection (cont.)

Even earlier:

c) John Michell (1783)

d) Pierre-Simon Laplace (1796)

independently investigated light being subjected to gravity, but in the context of a body

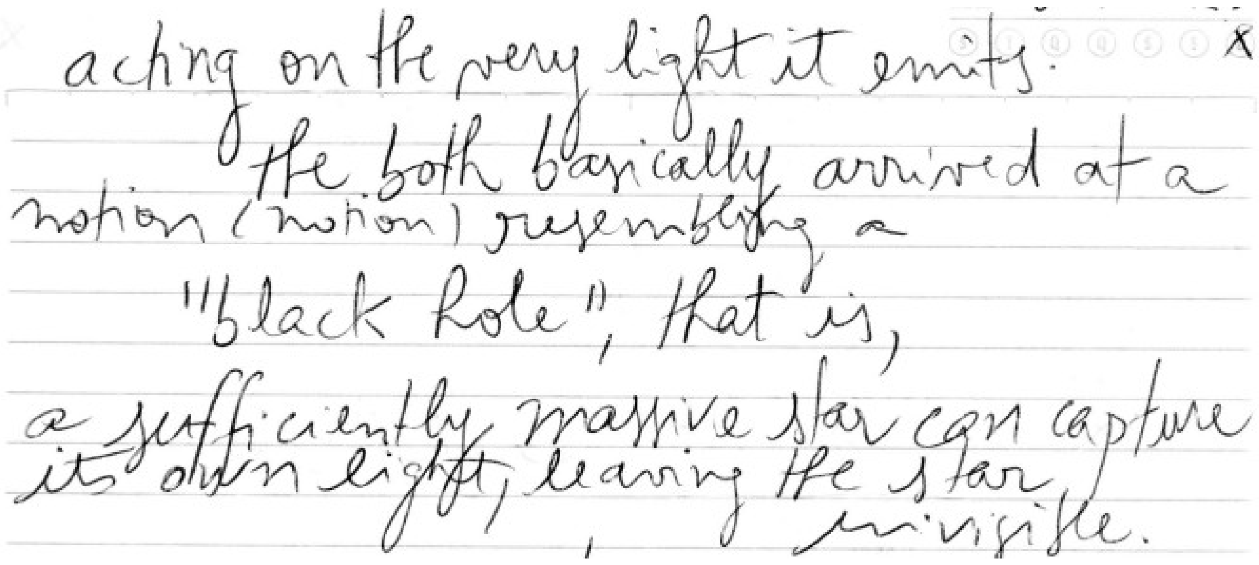

acting on the very light it emits.

He both basically arrived at a notion (notion) resembling a "black hole", that is, a sufficiently massive star can capture its own light, leaving the star invisible.

Michell and Laplace did not write about the situation of light coming from a distant source and undergoing deflection as it passes near a gravitating body.

They are who inspired the works of Cavendish and von Soldner

Soldner:

"Thus when a light ray passes by a celestial body, it will, instead of a going on in a straight direction, be forced by its attraction to describe a hyperbola whose concave side is directed against the attracting body."

He showed that a light grazing the sun would be deflected by 0.84 seconds of arc.

jandaia

F >
2 ! ! .;
! * ! ! P ; * ! ! 0$/ N ;@ ;
*! !! , ! 3, 3 3 C OC Q
! ; ! 3 3 ! * ! | . !
3
< H * . $ 3 3 .; K *
! ,N . ! (. B F * ! 8 @ !
! (B C B 3 @ * . , O,2 > *
F : ; F 8 !;
.;,Q < 33 ! ,8 (K .
! K $ /, O Q• K B | A OK BQ C ; *
! | 3 3 R . !
! ; * . !; ! * ;
C 33 ; @ * @ * ;| 3 @
; ! , } , . !; OK B@ 33 ,/ ,/0Q OC @ 3 / /@ &&&Q
! @ . ! . * ; ! .
: ; (;@ ! ,< ; 3 .; . !;@

@ ! ! @ . ! .; ! . ;
3 . * ! ! ! . !; , P !
; } ! . ! ! .; $/ ! ! ! ! ,& !
. * H ! * ; @ ! * B * ;
, ! } ! ! 3 } !
! ,P 3 ;@ ! ! .4 . & ; *;
. !; , ( ) @ , ! | B ! !
3 ;, OK BQ

A* @ K 0$" O Q ! F 3 0,% OF 3Q
! 3 ! ; * ! . .4 ! * ;@ . H . !;
* ; < ; . . ; * ! .
,. B @, @ ; * 3 @ *
* . F 3 ! ; 3 @ . 3 * ! ! 3 0,$•
P B ! A OP A@ 3 "% Q A & @ * @
! F 3 ! ! . !
! ! ! 3 * . !; C @ @
! F 3 3 ! B * ! ! * ! @ 3 * ;
@ OC @ 3 / "Q ! OK B@ 33 ," ,"%Q
& ! @ ; 3 R ! ! * ! !
! .; * ; (! * !;
! 3 ; @ ! * * ! !
3 ! 3 . * ! ; P 6 ; 23 B 0 /

,> 8 ! 3 F ! @ ! .; . ! : ; • !
! y, OR Q 3 ! ;
! ; ! ! * . ! < @
@ ! K B @ OK BQ ,P OR Q@ @ ! @
! ! ! ,
;.) ; 3 * B ; *
! . !; !) * ; 3 . !
A* @ 3 ; * ! @ 3 ;
3 3 ? 3 F . ; 0,%@ ,C ! F }
? ! B @ ! ! : B
P } @ & * ! A ! B } !, O,& ! 3
* ; ! * ;@ B ! * } @
. ! . ! * ; .; ,Q OF Q
C B ! ; * ! * ! ! * !
! .; = (. ; * ! | B@
(. A * ! * 3 !
! . !) * 3 3 A ! ; !! !
* ; ! ,N. ! A 8 ! B !
(. ! F , O,2 & ? * 3 F ,Q
OA Q• OF A C@ 3 ,,Q A OA P 3 ! !
; 3 , . !;, ! ! !
! * 3 A ! * $" !
! } & 3 3 @ A !
! *) ,A ! ! } @ ! 8
(! E @ ! * .
N. ! ! ! . N , O,& ! .
; ! . B 3) ! @ * !
! . ; ! ! ; 3 * ,

8; , A ! ! ! * H3 !
3 ? < ; : * ; OA &%@ A Q E
; 3 ! : .4 ! !
E

$$\hat{\alpha} = \frac{4GM}{c^2R},$$

? * ; ! 3 ! C *

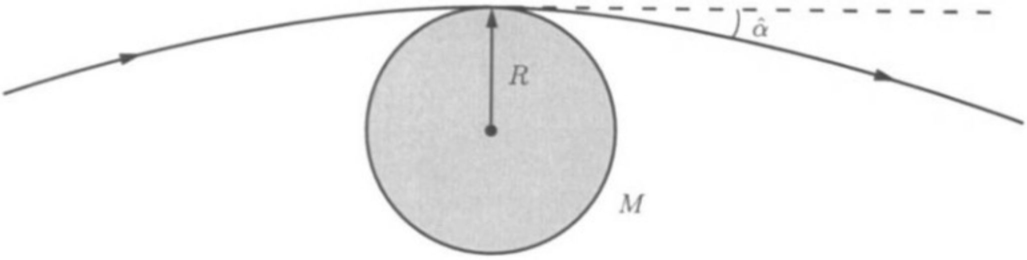

E F ! H3 ! .;
< M ! ; .; * ! 0$/ ! * ! $
; ! R | ; * ; OC ! &
? : * ; , . ! . ! .;
A)

! ! . !

$$\hat{\alpha} = \frac{4GM_\odot}{c^2R_\odot} \simeq 1.''74.$$

Note that since the light ray grazes the sun, we assume $R \simeq R_\odot$. In Einstein's own words: "Ein an der Sonne vorbeigehender Lichtstrahl erfährt demnach eine Biegung von 1.7 Bogensekunden" ["A light ray grazing the sun therefore experiences a deflection of 1.7 arc-seconds"]. We refer readers to [Ray-BB, pp. 59–60] for a derivation of (1.1).

& ; * 3 3 ! ; !
@ .; 2 * F ! . ,A ! ,
! 3 , 3 . ; * !
@ , OF !Q P 3 ! !
, 3 * ! R 3 ,

Light Deflection

One of the earliest calculations of the deflection of light from a distant source by a foreground gravitational field is due to Henry Cavendish around 1784. Unfortunately, the only evidence that he did such a calculation is on an "isolated scrap" of paper. Will [Wil] offers some details of this story and attempts to reproduce Cavendish's calculation of the bending angle of light passing near the surface of an attracting mass.

The next closest calculations seems to have been in an 1801 paper by Johann von Soldner: "Ueber die Ablenkung eines Lichtstrals von seiner geradlinigen Bewegung, durch die Attraktion eines Weltkörpers, an welchem er nahe vorbeigeht" ["On the Deviation of a Light Ray from its Motion along a Straight Line through the attraction of a Celestial Body which it Passes Closeby"]. This article appeared in "Berliner Astronomisches Jahrbuch für das Jahr 1804" [Sol]; see Jaki's English translation [Jak]. We note that initially von Soldner's paper calculates — using Newtonian mechanics — the bending angle formula for light emitted tangentially from the surface of a gravitating body and travels off to infinity. When applying his formula, however, he reverses the light ray's path, so it comes in from infinity and "grazes" the attracting body [Jak, pp. 945-947] (cf. [Wil, p. 414, Sec. III]). Von Soldner, who later became director of the Munich observatory and a foreign member of the Royal Astronomical Society, figured out: "Thus when a light ray passes by a celestial body, it will, instead of going on in a straight direction, be forced by its attraction to describe a hyperbola whose concave side is directed against the attracting body." He showed that a light ray grazing the sun would be deflected by 0.84 seconds of arc and concluded: "If one could observe the fixed stars very close to the sun, then one would have to take this very much into account." Von Soldner realized that he was far ahead of his time when he apologized at the end: "Hopefully, no one would find it objectionable that I treat a light ray as a heavy body." As a consequence, "Soldner's work on the deflection of light had no influence either on practice or on theory" [Jak].

Even earlier, John Michell in 1783 [Miche] and Pierre-Simon Laplace in 1796 [Lap] independently investigated light being subjected to gravity, but in the context of a body acting on the very light it emits. They both basically arrived at a notion resembling a "black hole," that is, a sufficiently massive star can capture its own light, leaving the star invisible. Laplace stated his result initially without proof, but provided the proof in 1798; see Hawking and Ellis [Haw-E, p. 365] for an English translation. It seems, however, that Michell and Laplace did not write about the situation of light coming from a distant source and undergoing deflection as it passes near a gravitating body. We note, though, that it was Michell and Laplace who inspired the works of Cavendish and von Soldner, respectively (e.g., [Wil, p. 413] and [Jak, pp. 935–936]).

In hindsight, it is slightly surprising that Newton himself did not seem to have studied light deflection by gravity. All the ingredients were there with his gravitational dynamics and the corpuscular theory of light, and the mathematics involved was straightforward compared to other problems he solved successfully. His famous Query 1 of the *Opticks* (1704):

"Do not Bodies act upon Light at a distance, and by their action bend its Rays; and is not this action strongest at the least distance?" [New] is phrased in such a general way that today we tend to assume that it includes gravitational bending as well. This is not the case, though, according to Jaki, who comments [Jak]: "He [Newton], of course, had reflection, refraction and diffraction in mind."

Maybe it is equally surprising that it seems to have taken more than 100 years after von Soldner until somebody else dealt again quantitatively with the problem of light deflection. Even more so, when one learns that a famous contemporary of von Soldner, the physicist-philosopher Georg Christoph Lichtenberg wrote in the year 1796, "Wenn das Licht trotz seiner Geschwindigkeit noch eine Schwere hat, so würde doch so etwas wie Refraktion am Horizonte erscheinen müssen, weil es von der Erde stark gezogen wird" ["If light despite its velocity also had gravity, then something like refraction would have to occur at the horizon, because it would be attracted heavily by the earth"] [Lic].

We know today that the value of the deflection angle that Cavendish and von Soldner calculated is too small by a factor of two.[1] About a century after von Soldner's work, Albert Einstein arrived at a similar numerical value in his first attempt to determine the deflection angle — based on the equivalence principle. Einstein directly addressed the influence of gravity on light in a letter entitled: "Über den Einfluß der Schwerkraft auf die Ausbreitung des Lichtes" ["On the Influence of Gravitation on the Propagation of Light"] [Ein11]; see [Lor-EMW, p. 99] for an English translation of [Ein11]. He predicted that a light ray passing close to a "celestial body" might suffer a deflection towards the side of decreasing gravitational potential. Einstein found a numerical value of 0.83 seconds of arc for the deflection angle of light grazing the sun. In this paper, Einstein also stated that he wished astronomers to investigate this question: "Es wäre dringend zu wünschen, daß sich Astronomen der hier aufgerollten Frage annähmen, auch wenn die im vorigen gegebenen Überlegungen ungenügend fundiert oder gar abenteuerlich erscheinen sollten" ["It would be highly desirable that astronomers take up the question raised here, even if the considerations should seem to be insufficiently founded or entirely speculative"].

By 1915 Einstein succeeded in deriving the correct expression for the light deflection angle with the presentation of the General Theory of Relativity [Ein16, Ein22]. For a light ray passing at a distance R from an object of mass M he found the following deflection angle[2] — see Figure 1.1:

$$\hat{\alpha} = \frac{4GM}{c^2 R}, \tag{1.1}$$

where G is the constant of gravity and c is the speed of light. With the solar values for

---

[1] Will [Wil] compares Cavendish and von Soldner's determinations of the light deflection angle, and includes Einstein's correction to twice the value.

[2] Notice that the bending angle $\hat{\alpha}$ is independent of the light's wavelength and depends on the impact distance and deflector's mass. See Section 3.1.5.

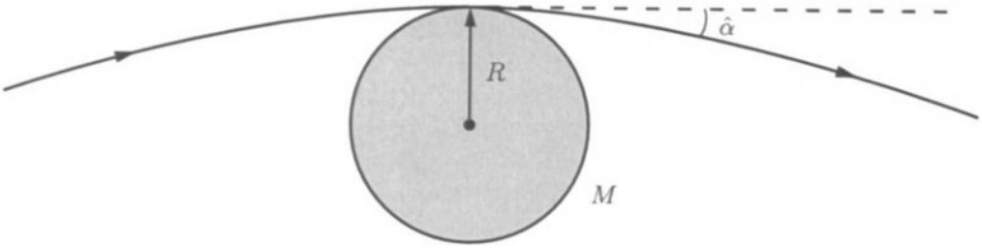

Figure 1.1. Light from a distant source experiences a deflection through angle $\hat{\alpha}$ by a mass M. The angle $\hat{\alpha}$ was calculated incorrectly by Cavendish (1784) and von Soldner (1801) since they used Newton's theory of gravity (see [Wil] for details). It was not until the formulation of General Relativity in 1915 that the correct formula for the bending angle was established by Einstein — see equation (1.1).

radius and mass he obtained

$$\hat{\alpha} = \frac{4GM_\odot}{c^2 R_\odot} \simeq 1.''74.$$

Note that since the light ray grazes the sun, we assume $R \simeq R_\odot$. In Einstein's own words: "Ein an der Sonne vorbeigehender Lichtstrahl erfährt demnach eine Biegung von 1.7 Bogensekunden" ["A light ray grazing the sun therefore experiences a deflection of 1.7 arcseconds"]. We refer readers to [Ray-BB, pp. 59–60] for a derivation of (1.1).

In the following years there were several papers dealing theoretically with light deflection, among them a short letter by Oliver Lodge about the "Einstein deflection" in which he made the point that "it is not permissible to say that the solar gravitational field acts as a lens, for it has no focal length" [Lod]. He pointed out that one would get a focal line starting at "a place seventeen times the distance of Neptune."

Multiple Imaging

While investigating the birth of general relativity, Renn, Sauer and Stachel [Ren-SS] discovered that in notes dated to the spring of 1912 Einstein explored gravitational lensing by a star. He showed that a distant star lying close to the line of sight to a foreground star will be imaged twice, with a faint secondary image increasing the total measurable intensity of the background star [Ren-SS, pp. 185, 186].[3] Einstein did not publish these results until 1936 [Ein36]. The possibility of double images was also mentioned in 1923 by Sir Arthur Eddington in his book *Space, Time and Gravitation* [Edd, p. 133–134]. This effect was also quantitatively studied by Chwolson[Chw] in 1924. Chwolson's article, "Über eine mögliche Form fiktiver Doppelsterne" ["On a possible form of fictitious double stars"], even mentions the "reversed mirror image" (i.e., the negative parity of the second image)

---

[3] In the same notes, Einstein also determined the 'lens equation' for the system, that is, the equation whose solutions give the image positions [Ren-SS, p. 185].

data 08.09.2019

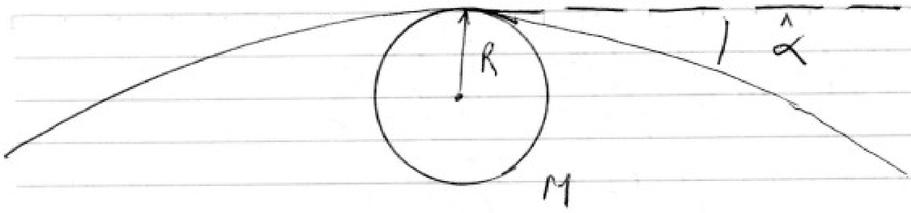

$$\hat{\alpha} = \frac{4GM_\odot}{c^2 R_\odot} \simeq 1.''74$$

Source: Pág. 06(29)
20190907/Singularity Theory and Gravitational Lensing by Arlie O. Petters/.

off: Constante Gravitational.

$$G = 6{,}674184 \times 10^{-11}\, m^3\, Kg^{-1}\, s^{-2}$$

obs:

$$F = G\,\frac{m_1 m_2}{r^2} \qquad (G \neq g)$$

$$g\,(\text{aceleração da gravidade}) = \frac{GM}{r^2}$$

data 08.09.2019

Veja também:

Teoria Gravitacional de Le Sage

$$G = \frac{\varepsilon_c \sigma^2}{4\pi M_n^2}$$ (Sergey Fedosin)

Obs.: S. G. Fedosin

I) The graviton field as the source of mass and gravitational force in the modernized Le Sage's model.

Source: Physical Science International Journal
ISSN: 2348-0130
Vol. 8, Issue 4, pp. 1-18 (2015)

http://dx.doi.org/10.9734/PSIJ/2015/22197

↳ www.sciencedomain.org/index.php?/abstract/12140

off: http://sciencedomain.org/review-history/12140

data 08.09.2019
S T Q Q S S X

Original Manuscript
www.sdiarticle2.org/prh/
PSIJ_33/2015/Ms_PSIJ_22197.pdf
or
www.sciencedomain.org/download/
MTIxNDBAQHBm.pdf

II) Fedosin S. G.
The Force Vacuum Field as
an Alternative to the Ether and
Quantum Vacuum.
Source: WSEAS Transactions
on Applied and Theoretical
Mechanics
ISSN: 1991-8747/2224-3429
Volume 10, Art. #3,
pp. 31-38 (2015)
Source: wseas.org/wseas/cms.action?
id=10178

data 08.09.2019

Download article:

www.wseas.org/multimedia/
journals/mechanics/2015/
a065711-221.pdf

III) Fedosin S. G.

The charged component of the
vacuum field as the source of
eletric force in the modernized
Le Sage's model.

Source: Journal of Fundamental
and Applied Sciences,
Vol. 8, No. 3, pp. 971-1020 (2016)

https://dx.doi.org/10.5281/
zenodo.845357

↳ https://zenodo.org/record/
845357

September 1, 2016

08.09.2019

Notice de la Vie et des
Écrits de George-Louis
LE SAGE de Genève.

Source: Pierre Prevost
https://gdz.sub.uni-goethingen.de/id/PPN526449152/

Download:
https://gdz.sub.uni-goethingen.de/download/pdf/PPN526449152/PPN526449152.pdf

---

Photon-Graviton Recycling
as Cause of Gravitation
Apeiron, Vol. 14, No. 3, July 2007.
Matthew R. Edwards

Source: http://redshift.vif.com/JournalFiles/V14NO3PDF/V14N3EDW.pdf

& ,

9(:69 9 92 9 , K 33 3
< !; 3!

' 0 $ / E ! :

E /

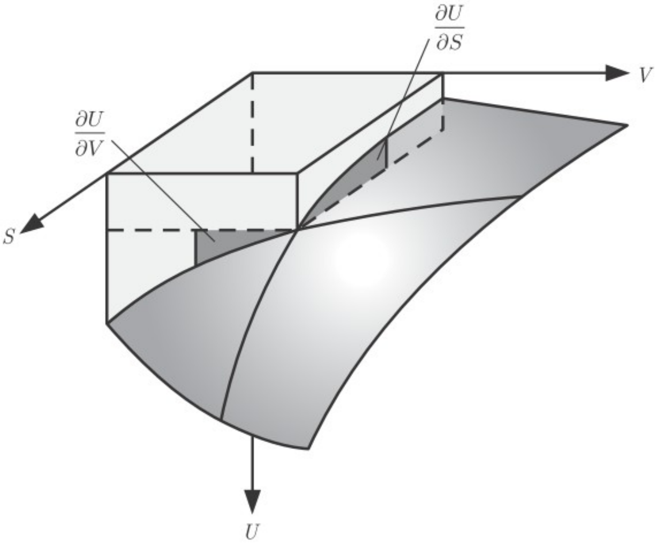
$\frac{\partial U}{\partial S}$
V
$\frac{\partial U}{\partial V}$
S
U

* H 3 ! ; N @ 3; ! *
3 @ N@ < * 3 H ! H ! ! * N @
< 3 * 3 ! 3 ! * *

4.2 Fundamental Relations

In this chapter, we will consider a simple thermodynamic system described by the following state variables: entropy S, volume V and number of moles $N_1, \ldots, N_r$ of the r constituents. From now on, we shall denote by $\{N_A\}$ the set of numbers of moles of the r chemical constituents ($A = 1, \ldots, r$).

4.2.1 Gibbs Relation

According to the definitions (2.9), (2.10) and (2.11) of temperature T, pressure p and chemical potentials μ_A, the differential of $U(S, V, \{N_A\})$ is obtained by applying relation (1.7),

4 Thermodynamic Potentials

Hermann Ludwig Ferdinand von Helmholtz, 1821–1894

H. L. F. von Helmoltz was professor of physiology, first in Königsberg, then in Bonn, Heidelberg and Berlin. He saw friction and inelastic collisions as processes of energy redistribution, which he explained in his first book on thermodynamics : 'Über die Erhaltung der Kraft'. In addition to the vast research that made him famous, he was interested in the sensory perception of colours and music.

4.1 Historical Introduction

The nineteenth century witnessed many developments in thermodynamics, most notably the introduction of the ***thermodynamic potentials*** that will be presented in this chapter.

The study of chemical equilibria raised the question of knowing whether it was possible to define a potential that would play in thermodynamics an analogous role to that of the potential in mechanics. Then, equilibrium would be obtained by minimising the potential [32].

2. P F ! E ! ! * P }@ $ $,/) ! ?
P F E * P } 3 3 ; ;@ D L*f* . @ 8 @
P ! . ! 8 P ! 3 ; ! . @
H3 ! D . B !; „.... ! A ! L G
& !! * ! @ ! ;
3 3 !

In 1869, François-Jacques-Dominique Massieu, general inspector of "Ecole des Mines" and professor at the Science Faculty of the University of Rennes, presented to the "Académie des sciences" his work on what he called the *characteristic functions of fluids*. He considered entropy as a function of internal energy, volume and number of moles of each constituent. He then defined transformations of this function according to a method developed by the French mathematician Adrien-Marie Legendre [33].

In the 1870s, the American physico-chemist and mathematician Josiah Williard Gibbs published a series of articles on the equilibrium of heterogeneous fluids. While his analysis was based on the condition of entropy maximisation, he used the functions denoted nowadays by F, G and H in order to clarify the equilibrium conditions he had found. He acknowledged the contribution of Massieu in his work. Gibbs considered how much energy changed when the mass of one of the constituents changed. Thus, he introduced the notion of chemical potential, but did not use this term. He insisted on avoiding the adjective 'chemical', as he wished to emphasise that his equilibrium conditions pertained not only to chemical substances, but also to phase transitions, electrical effects or elastic media [34].

In 1882, Hermann von Helmholtz determined the potential that defined the equilibrium state of a system when its temperature is kept constant [35]. Thus, he defined the ***free energy***, denoted by the letter F, that is still used nowadays. Another potential, designated today by the letter H, was named ***enthalpy*** by Kammerling Onnes [36].

In 1884, the Frenchman Pierre Duhem presented a thesis on the application of thermodynamic potentials in chemistry. Purportedly, he had to withdraw it under the pressure of Marcelin Berthelot. Later, he published his works in the form of three books, in which he described what he called *chemical mechanics*, which in effect was a description of chemical reactions based on thermodynamics.

In this chapter, we introduce the free energy also called the "Helmholtz free energy", denoted F, and the "Gibbs free energy" denoted G. The international organisation IUPAC recommends using the expressions "***Helmholtz free energy***" A and "***Gibbs free energy***" G. We will discuss the meaning of the adjective "free" in these designations. Although the IUPAC convention denotes Helmholtz free energy by the letter A, the physics community perseveres in the usage of F and we will follow this convention.

4.2 Fundamental Relations

In this chapter, we will consider a simple thermodynamic system described by the following state variables: entropy S, volume V and number of moles $N_1, \ldots, N_r$ of the r constituents. From now on, we shall denote by $\{N_A\}$ the set of numbers of moles of the r chemical constituents ($A = 1, \ldots, r$).

4.2.1 Gibbs Relation

According to the definitions (2.9), (2.10) and (2.11) of temperature T, pressure p and chemical potentials μ_A, the differential of $U(S, V, \{N_A\})$ is obtained by applying relation (1.7),

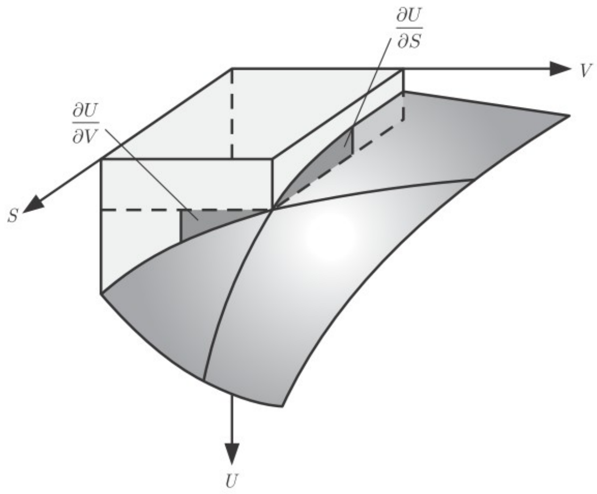

$\frac{\partial U}{\partial S}$
V
$\frac{\partial U}{\partial V}$
S
U

Figure 4.1 Convex surface corresponding to the internal energy $U(S, V)$ as a function of entropy S and volume V in state space (S, U, V). The vertical planes at fixed S or fixed V define curves on the surface $U(S, V)$. The slopes of these curves correspond to the partial derivatives.

$$dU = T\,dS - p\,dV + \sum_{A=1}^{r} \mu_A\,dN_A \tag{4.1}$$

Equation (4.1) is called the ***Gibbs relation***. This equation can also be obtained by multiplying (2.12) by the infinitesimal time interval dt. The Gibbs relation (4.1) expresses the fact that the internal energy U is a state function, i.e. $U = U(S, V, \{N_A\})$, and that the intensive quantities T, p and μ_A are also state functions, i.e.

$$\begin{aligned} T &= T(S, V, \{N_A\}) \\ p &= p(S, V, \{N_A\}) \\ \mu_A &= \mu_A(S, V, \{N_A\}) \end{aligned} \tag{4.2}$$

The relations that appear in (4.2) are called ***equations of state***, because they constitute relationships between the state variables S, V and $\{N_A\}$.

As an ilustration, the internal energy $U(S, V)$ is graphically represented as a state function of the variables entropy S and volume V (Fig. 4.1). This allows a geometric interpretation of the temperature $T(S, V)$ and pressure $p(S, V)$. In general, the internal energy $U(S, V)$ is represented as a surface in the abstract 3-dimensional space (S, U, V), called ***state space***. The temperature $T(S, V)$ and pressure $p(S, V)$ are state functions evaluated at a point on the surface $U(S, V)$. The temperature, defined by $T = \partial U(S, V)/\partial S$, is the derivative with respect to S of the curve located at the intersection of the surface $U(S, V)$ and the plane $V = \text{const}$. The pressure $p = -\,\partial U(S, V)/\partial V$ is the opposite of the derivative with respect to V of the curve located at the intersection of the surface $U(S, V)$ and the plane $S = \text{const}$. The curvature of the surface $U(S, V)$ is such that $T > 0$ and

$p > 0$. Notice that the equations of state $U(S, V)$ and $S(U, V)$ are represented by the same surface in state space.

4.2.2 Euler Relation

Let us consider now a system consisting of λ identical subsystems, each characterised by its entropy S, volume V and number of moles N_A for substances $A = 1, ..., r$. Since the state variables are extensive, the system has a total entropy λS, a total volume λV and λN_A moles of substance A. Since the internal energy of the system is an extensive state function, it is the sum of the internal energies of the subsystems, i.e.

$$U(\lambda S, \lambda V, \{\lambda N_A\}) = \lambda U(S, V, \{N_A\}) \tag{4.3}$$

The derivative of the relation (4.3) with respect to λ can be written as,

$$\frac{\partial U}{\partial(\lambda S)} \frac{d(\lambda S)}{d\lambda} + \frac{\partial U}{\partial(\lambda V)} \frac{d(\lambda V)}{d\lambda} + \sum_{A=1}^{r} \frac{\partial U}{\partial(\lambda N_A)} \frac{d(\lambda N_A)}{d\lambda} = U \tag{4.4}$$

Since S, V and N_A are independent of λ, equation (4.4) reduces to,

$$\frac{\partial U}{\partial(\lambda S)} S + \frac{\partial U}{\partial(\lambda V)} V + \sum_{A=1}^{r} \frac{\partial U}{\partial(\lambda N_A)} N_A = U \tag{4.5}$$

Equation (4.5) has to be satisfied for every value of λ. Thus, choosing $\lambda = 1$,

$$U = \frac{\partial U}{\partial S} S + \frac{\partial U}{\partial V} V + \sum_{A=1}^{r} \frac{\partial U}{\partial N_A} N_A \tag{4.6}$$

Using the definitions (2.9)–(2.11), relation (4.6) yields,

$$U = TS - pV + \sum_{A=1}^{r} \mu_A N_A \tag{4.7}$$

which is called the ***Euler relation***.

4.2.3 Gibbs–Duhem Relation

The infinitesimal variation of internal energy dU is obtained by differentiating the Euler equation (4.7),

$$dU = T\,dS + S\,dT - p\,dV - V\,dp + \sum_{A=1}^{r} \left(\mu_A\,dN_A + N_A\,d\mu_A\right) \tag{4.8}$$

Then, according to the Gibbs relation (4.1), the expression (4.8) requires that,

$$S\,dT - V\,dp + \sum_{A=1}^{r} N_A\,d\mu_A = 0 \tag{4.9}$$

which is called the ***Gibbs–Duhem relation***.

4.3 Legendre Transformation

So far, we have considered the state function internal energy, which is a function of extensive state variables. The conjugated quantities obtained from internal energy by partial derivation are intensive quantities. Often in practical situations, it is these intensive quantities that are measured or controlled. The most obvious case is entropy and its conjugated variable temperature. In a measurement, temperature is more likely to be controlled than entropy.

This raises the question whether it is possible to reformulate our mathematical description in order to use intensive quantities as state variables. The extensive variables will appear then as conjugated quantities. We will now see that such a mathematical operation is possible [37]. This approach will lead us to introduce the thermodynamic potentials called free energy, enthalpy and Gibbs free energy. The simplest case is a state function $F(X_0, X_1, \ldots, X_n)$. Our goal is to use one or more partial derivatives $Y_i(X_0, X_1, \ldots, X_n) = \partial F/\partial X_i$ as independent state variables without loosing the information contained in the function $F(X_0, X_1, \ldots, X_n)$.

Let us consider first the state function $F(X)$ of a single extensive variable X. The slope of the tangent to the state function $F(X)$ at X is given by,

$$Y(X) = \frac{dF(X)}{dX} \tag{4.10}$$

The state function $F(X)$ is a curve determined by a set of points (X, F) corresponding to values of the variable X and the associated function F. The tangent of slope $Y(X)$ to this curve at a fixed point (X, F) intersects the vertical axis at point $(0, G)$ (Fig. 4.2).

Thus, the tangent is determined by a couple of values (Y, G) that correspond to the slope Y and to the coordinate of the intersect on the vertical axis G. Graphically (Fig. 4.2), the slope Y of the tangent is given by

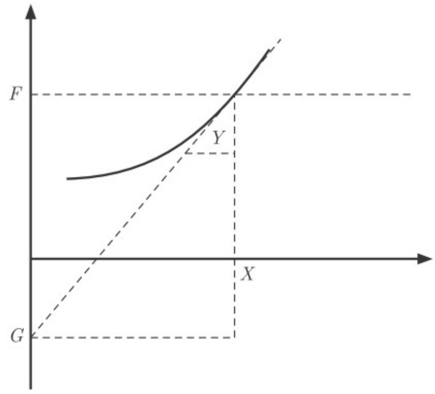

Figure 4.2 $Y(X)$ is the slope of the tangent to the state function $F(X)$ at point X. The tangent intersects the vertical axis at point G.

$$Y = \frac{F - G}{X - 0} \tag{4.11}$$

When inverting the equation of state $Y = Y(X)$, we obtain $X = X(Y)$. Thus, the state function F can be expressed as $F(X(Y))$ and equation (4.11) becomes

$$G(Y) = F(X(Y)) - YX(Y) \tag{4.12}$$

The expression (4.12) is called the ***Legendre transform*** of the state function F with respect to the state variable X. The Legendre transform $G(Y)$ is a state function of Y. The Legendre transformation of $G(Y)$ with respect Y yields again the function $F(X)$. Thus, the state functions $F(X)$ and $G(Y)$ contain the same information.

The Legendre transformation can be generalised from one to several variables. The partial derivative of the state function $F(X_0, X_1, \dots, X_n)$ with respect to X_i, where $i = 0, 1, \dots, n$, evaluated in the state $(X_0, X_1, \dots, X_n)$ is given by,

$$Y_i(X_0, X_1, \dots, X_n) = \frac{\partial F(X_0, X_1, \dots, X_n)}{\partial X_i} \tag{4.13}$$

By analogy with the Legendre transformation with respect to a single state variable, we can write,

$$Y_i = \frac{F - G}{X_i - 0} \tag{4.14}$$

When inverting the equation of state $Y_i = Y_i(X_0, X_1, \dots, X_n)$, we obtain $X_i = X_i(X_0, X_1, \dots, Y_i, \dots, X_n)$. Thus, the state function F can be expressed as $F\Big(X_0, X_1, \dots, X_i(X_0, X_1, \dots, Y_i, \dots, X_n), \dots, X_n\Big)$ and equation (4.11) becomes,

$$\begin{aligned} G(X_0, X_1, \dots, Y_i, \dots, X_n) = {} & F\Big(X_0, X_1, \dots, X_i(X_0, X_1, \dots, Y_i, \dots, X_n), \dots, X_n\Big) \\ & - Y_i X_i(X_0, X_1, \dots, Y_i, \dots, X_n) \end{aligned} \tag{4.15}$$

According to the definition (4.13), the partial derivative of equation (4.15) with respect to Y_i yields,

$$\frac{\partial G}{\partial Y_i} = \frac{\partial F}{\partial X_i}\frac{\partial X_i}{\partial Y_i} - X_i - Y_i\frac{\partial X_i}{\partial Y_i} = -X_i \tag{4.16}$$

There is a simple relationship between the curvatures of the functions $G(Y)$ and $F(X)$ that we can demonstrate as follows. We obtain the second order partial derivative of F with respect to X_i and of G with respect to Y_i, by differentiating relations (4.13) and (4.16),

$$\frac{\partial^2 F}{\partial X_i^2} = \frac{\partial Y_i}{\partial X_i} \quad \text{and} \quad \frac{\partial^2 G}{\partial Y_i^2} = -\frac{\partial X_i}{\partial Y_i} \tag{4.17}$$

This implies that,

$$\frac{\partial^2 G}{\partial Y_i^2} = -\left(\frac{\partial^2 F}{\partial X_i^2}\right)^{-1} \tag{4.18}$$

Hence, we find that the sign of the curvature of the state function G with respect to the variable Y_i is the opposite of the sign of the curvature of the state function F with respect to the variable X_i.

Equation (4.12) defines the Legendre transform of the state function F with respect to the state variable X_i. To simplify the notation, we simply write,

$$G = F - \; Y_i \, X_i \tag{4.19}$$

We obtain the state function $H\,(Y_0, Y_1, \ldots, Y_n)$ by performing successive Legendre transformations on the state function $F\,(X_0, X_1, \ldots, X_n)$, i.e.

$$H = F - \sum_{i=0}^{n} Y_i \, X_i \tag{4.20}$$

It is also possible to perform Legendre transformations on a few variables only, which is often the case in thermodynamics. In the following section, we will obtain the thermodynamic potentials by performing several Legendre transformations of the internal energy with respect to various extensive state variables.

4.3.1 Link between Thermodynamics and Mechanics

In analytical mechanics, a Legendre transformation defined with the opposite sign relates the Lagrangian $L\,(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ of a system containing n material points to the Hamiltonian $H\,(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n, \boldsymbol{P}_1, \ldots, \boldsymbol{P}_n)$ of this system [9], i.e.

$$H = \sum_{i=1}^{n} \boldsymbol{P}_i \cdot \boldsymbol{v}_i - \; L \tag{4.21}$$

Since the Legendre transform of a state function is a state function, its opposite is also a state function and contains the same information.

4.4 Thermodynamic Potentials

The extensive state function that plays a central role in thermodynamics is the internal energy $U\,(S, V, \{N_A\})$. We call ***thermodynamic potential*** every state function obtained by a Legendre transformation of $U\,(S, V, \{N_A\})$. By extension, the internal energy itself can also be called a thermodynamic potential since it can be obtained by two successive Legendre transformations starting from itself. The three other thermodynamic potentials that we will define in this section are the free energy, the enthalpy and the Gibbs free energy.

4.4.1 Free Energy

The ***free energy*** $F\,(T, V, \{N_A\})$, also called ***Helmholtz free energy***, is defined as the Legendre transform (4.19) of the internal energy $U\,(S, V, \{N_A\})$ with respect to the entropy S, i.e.

$$F = U - \; T S \tag{4.22}$$

According to the Euler relation (4.7), the free energy can be written as,

$$F = -\,p\,V + \sum_{A=1}^{r} \mu_A\,N_A \tag{4.23}$$

The differentiation of expression (4.22) is written,

$$dF = dU - \,T\,dS - \,S\,dT \tag{4.24}$$

Substituting for dU its value given by the Gibbs relation (4.1), expression (4.24) becomes,

$$dF = -\,S\,dT - \,p\,dV + \sum_{A=1}^{r} \mu_A\,dN_A \tag{4.25}$$

In conformity with the mathematical description of Legendre transformations, the differential (4.25) shows that the free energy is a state function of the variables T, V and N_A. Moreover, the variables that are conjugates of these state variables are defined as,

$$S\,(T,V,\{N_A\}) = -\,\frac{\partial F\,(T,V,\{N_A\})}{\partial T} \tag{4.26}$$

$$p\,(T,V,\{N_A\}) = -\,\frac{\partial F\,(T,V,\{N_A\})}{\partial V} \tag{4.27}$$

$$\mu_A\,(T,V,\{N_A\}) = \frac{\partial F\,(T,V,\{N_A\})}{\partial N_A} \tag{4.28}$$

4.4.2 Enthalpy

The ***enthalpy*** $H\,(S,p,\{N_A\})$ is defined as the Legendre transform (4.19) of the internal energy $U\,(S,V,\{N_A\})$ with respect to the volume V, i.e.

$$H = U + p\,V \tag{4.29}$$

According to the Euler relation (4.7), the enthalpy H can be written as,

$$H = T\,S + \sum_{A=1}^{r} \mu_A\,N_A \tag{4.30}$$

The differentiation of expression (4.29) yields,

$$dH = dU + p\,dV + V\,dp \tag{4.31}$$

Substituting for dU its value given by the relation (4.1), expression (4.31) becomes,

$$dH = T\,dS + V\,dp + \sum_{A=1}^{r} \mu_A\,dN_A \tag{4.32}$$

The differential (4.32) shows that the enthalpy is a state function of the variables S, p and N_A. Moreover, the conjugates of these state variables are,

$$T(S,p,\{N_A\}) = \frac{\partial H(S,p,\{N_A\})}{\partial S} \tag{4.33}$$

$$V(S,p,\{N_A\}) = \frac{\partial H(S,p,\{N_A\})}{\partial p} \tag{4.34}$$

$$\mu_A(S,p,\{N_A\}) = \frac{\partial H(S,p,\{N_A\})}{\partial N_A} \tag{4.35}$$

4.4.3 Gibbs Free Energy

The ***Gibbs free energy*** $G(T,p,\{N_A\})$ also called ***free enthalpy*** is defined as the Legendre transform (4.19) of the internal energy $U(S,V,\{N_A\})$ with respect to entropy S and volume V or, equivalently, as the Legendre transform of the enthalpy $H(S,p,\{N_A\})$ with respect to entropy S, i.e.

$$G = U + pV - TS = H - TS \tag{4.36}$$

Taking into account the Euler relation (4.7), the Gibbs free energy reads,

$$G = \sum_{A=1}^{r} \mu_A N_A \tag{4.37}$$

The differentiation of (4.36) yields,

$$dG = dU - T\,dS - S\,dT + p\,dV + V\,dp \tag{4.38}$$

Substituting for dU its value given by the Gibbs relation (4.1), the expression (4.38) for dG becomes,

$$dG = -S\,dT + V\,dp + \sum_{A=1}^{r} \mu_A\,dN_A \tag{4.39}$$

The differential (4.39) shows that the Gibbs free energy is a state function of the variables T, p and N_A. Moreover, the conjugates of these state variables are defined as,

$$S(T,p,\{N_A\}) = -\frac{\partial G(T,p,\{N_A\})}{\partial T} \tag{4.40}$$

$$V(T,p,\{N_A\}) = \frac{\partial G(T,p,\{N_A\})}{\partial p} \tag{4.41}$$

$$\mu_A(T,p,\{N_A\}) = \frac{\partial G(T,p,\{N_A\})}{\partial N_A} \tag{4.42}$$

In the following sections of this chapter, we shall see that these thermodynamic potentials enable the characterisation of equilibrium states of systems subjected to constraints imposed by their interaction with an environment that we shall define as reservoirs or baths.

(95)

09.09.2019

95) 20190909

Principles of thermodynamics

Source: 20190909_02/Principles of thermodynamics/[Jean-Philippe....]...pdf

Pág. 72(85)

4.2 Fundamental Relations

4.2.1 Gibbs Relation

→ Fig. 4.1 →

Convex surface corresponding to the internal energy U(S,V) as a function of entropy S and volume V in state space (S,U,V).
The vertical planes at fixed S or fixed V define curves on the surface U(S,V).
The slopes of these curves correspond to the partial derivatives.

$$dU = T\,dS - p\,dV + \sum_{A=1}^{r} \mu_A\, dN_A \quad (4.1)$$

Figure → convex surface.

Important:

data 09.09.2019

[✓ D] S T Q Q S S

Pág. 71(84) ⇒

The international organization IUPAC recommends using the expressions

"Helmholtz free energy" A

and

"Gibbs free energy" G.

—

Important:

Fazer um estudo com este gráfico (convex surface) com o item 94 ($\hat{2}$) – Soldner (1."74)

⇒ Light Deflection

—

96) Redefining the Kg (20190924)

Source:

https://www.youtube.com/watch?

V=Oo0jm1PPRuo
(zero)

Obs: World's Roundest object!

https://www.youtube.com/watch?

& ,% L

data 09.09.2019

Pág. 71(84) ⇒

the international organisation IUPAC recommends using the Expressions
"Helmholtz free energy" A
and
"Gibbs free energy" G.

—

Important:

Fazer um estudo com este gráfico (convex surface) com o item 94 ($\hat{\alpha}$) - Soldner (1."74)
↳ Light Deflection

—

96) Redefining the Kg (20190924)

Source:

https://www.youtube.com/watch?
V=Oo0jm1PPRuo
(zero)

Obs: World's Roundest object!
https://www.youtube.com/watch?

24.09.2019

1Kg = $2,15 \times 10^{25}$ atoms
Si 28 (silício)

Obs.:

Recalibrar a constante de Planck usando a Balança de Watt
(Watt Balance)
↳ chamada agora
Balança Kibble

Important,

Utilizar esta nova redefinição do Kg para ajudar a calcular o desvio da luz
(Cavendish).

—

Kg →

2 formas:

1/ esfera de silício para determinar e definir o número de Avogadro
(Constante de Avogrado)

1 m = distance light travels in $\frac{1}{299.792.458}$ s

light speed = 299,792,458 m/s

: +

3 ## ; . # y*I2 4 :

9(:69 9 92 9 ,% P C : ! B 3/

. 3 ## ; . # y*I 4A ? ?$

9(:69 9 92 9 ,% K : ! !
L 3/

3 ## ; . # y*I - &< -

9(:69 9 92 9 ,% < B ! ! * B 3/

! 3 ## ; . # y*I† 8;&/ > ~

9(:69 9 92 9 ,% C ! : ! 2.4 3/

3 ## ; . # y*I AF*:} 77777 77777

9(:69 9 92 9 ,% < > * ; < < ! 3/

24.09.2019

2) Balança de Watt para definir a constante de Planck.
(Kibble).

2011 → definiu-se Kg baseado na constante de Planck

Planck's Constant (h)

Avogadro's Number (N_A)

$$mg = BLI$$

(campo magnético) e comprimento do fio da bobina.

O movimento induz um tensão na bobina que é igual ao campo magnético.

X o comprimento do fio da bobina,

X sua velocidade.

Equações:

$V = BLv$ $IBL = mg$

→

24.09.2019

$$\frac{V}{v} = BL \qquad BL = \frac{mg}{I}$$

$$\frac{V}{v} = BL = \frac{mg}{I} \Rightarrow \frac{V}{v} = \frac{mg}{I} \Rightarrow$$

$$VI = mgv$$

↳ tensão x intensidade da corrente = massa x aceleração da gravidade x velocidade

Lado esquerdo: potência elétrica — Lado ~~esq~~ direito: potência mecânica

equilíbrio

Mudando disso para a constante de Planck, o número que relaciona a frequência de um fóton com sua energia?

A tensão pode ser medida acuradamente usando um efeito quântico macroscópico que envolve junções Josephson (Efeito Josephson).

24.08.2019

Junção Josephson

consiste em 2 supercondutores separados por um pedaço fino de isolante.

Você aplica uma radiação de microondas a essa junção criando uma tensão em todo o dispositivo e seu valor é precisamente conhecido por ser

$$V = \frac{hf}{2e}$$

h = constante de Planck

f = freqüência da radiação

e = carga em um elétron

Empilhando $\underline{\underline{n}}$ ⇒ number of junctions (em série)

$$V = n\,\frac{hf}{2e}$$

Essas junções são colocadas no circuito com a bobina, como é movido pelo campo e assim, você equilibra exatamente a tensão induzida na bobina usando essas junções Josephson

⇒

24.09.2019

$n\frac{hf}{2e}$ $\qquad$ $n\frac{hf}{2e} \leftarrow (2e)$

$$VI = mgv \Rightarrow$$

$$V\frac{V}{R} = mgv$$

A corrente é passada através de um resistor e nós medimos essa tensão novamente usando junções Josphson

Para medir a resistência, usamos outro efeito quântico macroscópico chamado de efeito Hall Quântico.

A medida da resistência será uma fração inteira

$$\boxed{\frac{1}{p}\frac{h}{e^2}}$$

↳ carga do elétron

Substituindo tudo, teremos h (constant Planck)

$$h = \frac{4}{pn^2}\frac{gv}{f^2}m$$

↳ frequência

Acurance 10 ppb (partes por bilhão).

Vejamos os termos.

24.09.2019

V = velocidade da bobina à medida que esta se move através do campo magnético

Para medir esta velocidade usamos um interferômetro à laser
(interferometer)

À medida que a distância à bobina muda as franjas de interferência passam por um detector.

Quantas franjas passaram em um certo período de tempo podemos determinar a velocidade da bobina.

g = usamos gravímetro

96) 20190924 (write 20190925)

Complementa o item anterior.

Ⓐ Spherical Coordinates

Source: 20190924/Quantum_Physics/
epdf.pub_quantum-physics-3rd-edition.pdf

3 51 ! ,% 3 !
9(:69 9 92 9 ,% 3! 3 .-) 3 ; " !
! 3! ' %0 E 0

Chapter 7

Angular Momentum

In our discussion of the Schrödinger equation in three dimensions we will need to deal with the kinetic energy in three dimensions, which has the form

$$K = \frac{\mathbf{p}^2}{2\mu} \tag{7-1}$$

for problems in which the effective mass of the particle is μ. Just as in classical mechanics, there is a close connection between $\mathbf{p}^2$ and the square of the angular momentum $\mathbf{L} = \mathbf{r} \times \mathbf{p}$. We shall take over this expression for the angular momentum into quantum mechanics, with the recognition that $\mathbf{r}$ and $\mathbf{p}$ are to be treated as operators. We will also see (Supplement 7-A **[www.wiley.com/college/gasiorowicz]**) that when the potential in a three-dimensional Schrödinger equation

$$H\psi(\mathbf{r}) = \left(\frac{\mathbf{p}^2}{2\mu} + V(\mathbf{r})\right)\psi(\mathbf{r}) = E\psi(\mathbf{r}) \tag{7-2}$$

is central, so that

$$V(\mathbf{r}) = V(r) \tag{7-3}$$

then angular momentum is conserved.

The conservation of angular momentum implies the operator equation

$$\frac{d\mathbf{L}}{dt} = 0 \tag{7-4}$$

This, as seen in Chapter 6, is equivalent to

$$[H, \mathbf{L}] = 0 \tag{7-5}$$

Our discussion of the details of the eigenvalues and eigenfunctions of the angular momentum operators is presented in two ways. We may proceed from $\mathbf{p}^2$ and obtain the angular momentum eigenvalue equation directly as a differential equation, which can be solved in a standard way. The other way proceeds from the definition of the angular momentum operator, and deals with the angular momentum commutation relations. This approach is patterned after the way we dealt with the harmonic oscillator in Chapter 6, and is more directly "quantum mechanical." We therefore relegate the direct approach to Supplement 7-B **[www.wiley.com/college/gasiorowicz]**, and turn directly to the algebraic, operator approach.

7-1 THE ANGULAR MOMENTUM COMMUTATION RELATIONS

We might be tempted to look for simultaneous eigenfunctions of H and all three components of $\mathbf{L}$. This, as we have seen in Chapter 5, is only possible if all four operators commute with each other. To proceed we must check whether *all* of the operators L_x, L_y, and L_z commute with each other, as well as with H. In fact, different components of the angular momentum *do not* commute with each other. For example, paying particular attention to the ordering of the operators, we get, using $\mathbf{L} = \mathbf{r} \times \mathbf{p}$,

$$\begin{aligned}[L_x, L_y] &= [yp_z - zp_y, zp_x - xp_z] \\ &= y[p_z, z]p_x + x[z, p_z]p_y \\ &= \frac{\hbar}{i}(yp_x - xp_y) = i\hbar L_z\end{aligned} \tag{7-6a}$$

Similarly we can show that

$$[L_y, L_z] = i\hbar L_x \tag{7-6b}$$

and

$$[L_z, L_x] = i\hbar L_y \tag{7-6c}$$

It is true that each of the components of angular momentum commutes with $\mathbf{L}^2$. For example,

$$\begin{aligned}[L_z, L_x^2 + L_y^2 + L_z^2] &= L_y[L_z, L_y] + [L_z, L_y]L_y + L_x[L_z, L_x] + [L_z, L_x]L_x \\ &= -i\hbar L_yL_x - i\hbar L_xL_y + i\hbar L_xL_y + i\hbar L_yL_x = 0\end{aligned} \tag{7-7}$$

We can see that as a consequence of these commutation relations, only one component of $\mathbf{L}$ may be chosen with H and $\mathbf{L}^2$ to form a simultaneously commuting set. To show this let us assume that we have a set of eigenfunctions that are simultaneous eigenfunctions of all three components of $\mathbf{L}$. Let us assume that

$$L_x|u\rangle = l_1|u\rangle$$

and

$$L_y|u\rangle = l_2|u\rangle$$

which implies that $L_xL_y|u\rangle = l_1l_2|u\rangle$ and $L_yL_x|u\rangle = l_2l_1|u\rangle$. As a consequence of (7-6a) this means that $L_z|u\rangle = 0$. This, however implies that

$$l_2|u\rangle = L_y|u\rangle = \frac{1}{i\hbar}[L_z, L_x]|u\rangle = \frac{1}{i\hbar}L_zl_1|u\rangle = 0$$

Similarly, we can show that $l_1|u\rangle = 0$. This means that only for $\mathbf{L} = 0$ can we have simultaneous eigenfunction for all three components of the angular momentum.

There is nothing to keep us from picking just one component of $\mathbf{L}$ as part of the commuting set. Conventionally the choice is L_z, but there is nothing special about this choice. We thus will deal with simultaneous eigenfunctions of $\mathbf{L}^2$ and L_z. We will denote the eigenkets by $|l, m\rangle$. Our starting point is thus the set of equations

$$\begin{aligned}\mathbf{L}^2|l, m\rangle &= \hbar^2 l(l+1)|l, m\rangle \\ L_z|l, m\rangle &= \hbar m|l, m\rangle\end{aligned} \tag{7-8}$$

At this point we make no statements about what the values of l and m are. They must be real since the operators $\mathbf{L}$ are hermitian. The factors of $\hbar$ are there purely for dimensional reasons, since the product $\mathbf{r} \times \mathbf{p}$ has the dimensions of $\hbar$.

7-2 RAISING AND LOWERING OPERATORS FOR ANGULAR MOMENTUM

Our starting point is (7-8), together with the angular momentum commutation relations (7-6) and the orthonormality relation

$$\langle l', m' | l, m \rangle = \delta_{ll'} \delta_{mm'} \tag{7-9}$$

It will prove convenient to introduce the operators

$$L_\pm = L_x \pm iL_y \tag{7-10}$$

These obey the commutation relations

$$\begin{aligned} [L_+, L_-] &= [L_x + iL_y, L_x - iL_v] = (-2i)[L_x, L_v] \\ &= 2\hbar L_z \end{aligned} \tag{7-11}$$

and

$$\begin{aligned} [L_z, L_\pm] &= [L_z, L_x \pm iL_y] = i\hbar L_y \mp i(i\hbar L_x) = \pm\hbar(L_x \pm iL_y) \\ &= \pm\hbar L_\pm \end{aligned} \tag{7-12}$$

It is also obvious that

$$[\mathbf{L}^2, L_\pm] = 0 \tag{7-13}$$

Furthermore, we have

$$\begin{aligned} L_+L_- &= (L_x + iL_y)(L_x - iL_y) = L_x^2 + L_y^2 - i[L_x, L_y] \\ &= \mathbf{L}^2 - L_z^2 + \hbar L_z \end{aligned} \tag{7-14}$$

and similarly,

$$L_-L_+ = \mathbf{L}^2 - L_z^2 - \hbar L_z \tag{7-15}$$

Thus

$$L_+L_- + L_z^2 - \hbar L_z = L_-L_+ + L_z^2 + \hbar L_z = \mathbf{L}^2 \tag{7-16}$$

We now note that $\langle l, m | L_x^2 | l, m \rangle = \langle L_x(l, m) | L_x(l, m) \rangle \geq 0$ and by extension $\langle l, m | \mathbf{L}^2 | l, m \rangle \geq 0$. This implies that $l(l + 1) \geq 0$. From this it follows that $l \geq 0$. (The alternative that $l \leq -1$ we reject, since we would then call $l + 1 = -l'$ and get $l' \geq 0$.) First, we note that

$$\mathbf{L}^2 L_\pm | l, m \rangle = L_\pm \mathbf{L}^2 | l, m \rangle = \hbar^2 l(l + 1) L_\pm | l, m \rangle \tag{7-17}$$

This means that $L_\pm | l, m \rangle$ is an eigenstate of $\mathbf{L}^2$ with the eigenvalue characterized by l. On the other hand,

$$L_z L_+ | l, m \rangle = (L_+ L_z + \hbar L_+) | l, m \rangle = \hbar(m + 1) L_+ | l, m \rangle \tag{7-18}$$

and similarly

$$L_z L_- | l, m \rangle = \hbar(m - 1) L_- | l, m \rangle \tag{7-19}$$

These equations imply that $L_+ | l, m \rangle$ is an eigenstate of L_z with the m value raised by unity, and $L_- | l, m \rangle$ is an eigenstate of L_z with m value lowered by unity. We therefore call $L_\pm$ *raising* and *lowering* operators, respectively. We may write

$$\begin{aligned} L_+ | l, m \rangle &= C_+(l, m) | l, m + 1 \rangle \\ L_- | l, m \rangle &= C_-(l, m) | l, m - 1 \rangle \end{aligned} \tag{7-20}$$

The conjugate relation of the first of the above is

$$\langle l, m|L_- = \langle l, m+1|C_+^*(l, m) \tag{7-21}$$

Multiplying this with the first of (7-20) yields

$$\begin{aligned} |C_+(l, m)|^2\langle l, m+1|l, m+1\rangle = \langle l, m|L_-L_+|l, m\rangle &= \langle l, m|\mathbf{L}^2 - L_z^2 - \hbar L_z|l, m\rangle \\ &= \hbar^2[l(l+1) - m^2 - m] \\ &= \hbar^2[(l-m)(l+m+1)] \end{aligned} \tag{7-22}$$

Thus

$$C_+(l, m) = \hbar\sqrt{(l-m)(l+m+1)} \tag{7-23}$$

and, similarly

$$C_-(l, m) = \hbar\sqrt{l(l+1) - m(m-1)} = \hbar\sqrt{(l+m)(l-m+1)} \tag{7-24}$$

It follows from

$$\langle L_\pm(l, m)|L_\pm(l, m)\rangle \geq 0$$

that

$$\begin{aligned} \langle L_\pm(l, m)|L_\pm(l, m)\rangle &= \langle l, m|L_\mp L_\pm|l, m\rangle \\ &= \langle l, m|\mathbf{L}^2 - L_z^2 \pm \hbar L_z|l, m\rangle \\ &= \hbar^2[l(l+1) - m(m \mp 1)] \geq 0 \end{aligned} \tag{7-25}$$

This implies that both

$$\begin{aligned} l(l+1) &\geq m(m+1) \\ l(l+1) &\geq m(m-1) \end{aligned} \tag{7-26}$$

are true. Since $l \geq 0$, it follows from the above that

$$-l \leq m \leq l \tag{7-27}$$

Let us assume that the minimum value of m is $m_{\min}$. This means that we cannot lower the m-value any further, and thus

$$L_-|l, m_{\min}\rangle = 0 \tag{7-28}$$

We can see, in a number of ways (by looking at $C_-(l, m)$ for example), that

$$m_{\min} = -l \tag{7-29}$$

Similarly, the maximum value of m, denoted by $m_{\max}$ is such that

$$L_+|l, m_{\max}\rangle = 0 \tag{7-30}$$

and

$$m_{\max} = l \tag{7-31}$$

Since the maximum value is to be reached from the minimum value by unit steps (repeated application of L_+), we find, as seen in Fig. 7-1, that there are $(2l + 1)$ steps. This implies that $(2l + 1)$ is an integer, and m can take on the values

$$m = -l, -l+1, -l+2, \ldots, l-1, l \tag{7-32}$$

The possibility that l is half-integral, $l = 1/2, 3/2, 5/2, \ldots$, will be discussed in Chapter 10, when we discuss *spin*. Until then, we restrict ourselves to *integer values of l*.

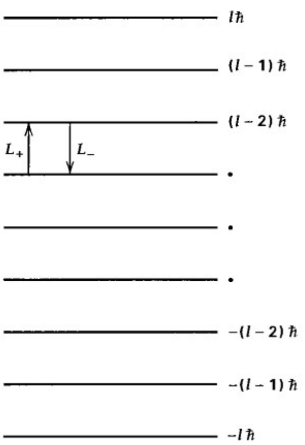

Figure 7-1 Spectrum of the operator L_z for a given value of l.

7-3 REPRESENTATION OF $|l, m\rangle$ STATES IN SPHERICAL COORDINATES

Our next task is to work out the general form of $Y_{lm}(\theta, \varphi)$, defined by

$$Y_{lm}(\theta, \varphi) = \langle \theta, \varphi | l, m \rangle \tag{7-33}$$

The angles (θ, φ) are the coordinates on the surface of a unit sphere (Fig. 7-2). The analog of

$$\int_{-\infty}^{\infty} dx |x\rangle\langle x| = \mathbf{1}$$

is

$$\int d\Omega |\theta, \varphi\rangle\langle \theta, \varphi| \equiv \int_0^{2\pi} d\varphi \int_0^{\pi} \sin\,\theta\, d\theta |\theta, \varphi\rangle\langle \theta, \varphi| = \mathbf{1} \tag{7-34}$$

It follows that the application of $\langle \theta', \varphi'|$ to (7-34) leads to

$$\langle \theta', \varphi' | \theta, \varphi \rangle = \frac{1}{\sin\,\theta} \delta(\theta - \theta')\, \delta(\varphi - \varphi') \tag{7-35}$$

To find an explicit form of the spherical harmonics as defined in (7-33) we make use of the polar coordinate representations

$$\begin{aligned} L_z &= \frac{\hbar}{i} \frac{\partial}{\partial \varphi} \\ L_\pm &= \hbar e^{\pm i\varphi}\left(\pm \frac{\partial}{\partial \theta} + i\, \cot\,\theta \frac{\partial}{\partial \varphi} \right) \end{aligned} \tag{7-36}$$

which are easily obtained from the material in Supplement 7-B.

Consider now $\langle \theta, \varphi | L_z | l, m \rangle$. This can be written in two ways. In one of them we have

$$\langle \theta, \varphi | L_z | l, m \rangle = \hbar m \langle \theta, \varphi | l, m \rangle \tag{7-37}$$

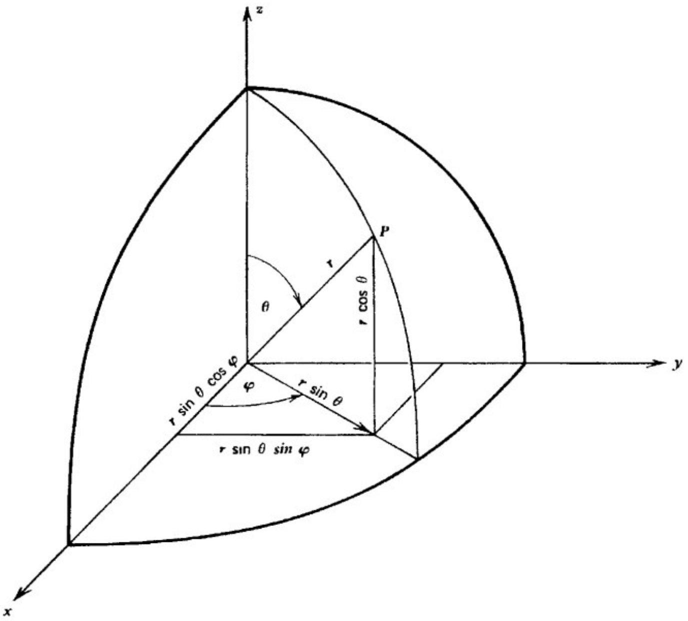

Figure 7-2 The definition of the spherical coordinates used in the text and the relation between the cartesian coordinates (x, y, z) and the spherical coordinates (r, θ, φ).

In the other we use (7-36) to write

$$\begin{aligned}\frac{i}{\hbar}\langle\theta, \varphi|L_z|l, m\rangle &= \lim_{\delta\varphi\to 0}[\langle\theta, \varphi + \delta\varphi|l, m\rangle - \langle\theta, \varphi|l, m\rangle]/\delta\varphi \\ &= \frac{\partial}{\partial\varphi}\langle\theta, \varphi|l, m\rangle\end{aligned} \tag{7-38}$$

Combining (7-37) and (7-38) we get the differential equation

$$\frac{\partial}{\partial\varphi}\langle\theta, \varphi|l, m\rangle = im\langle\theta, \varphi|l, m\rangle \tag{7-39}$$

whose solution is

$$\langle\theta, \varphi|l, m\rangle = F(\theta)e^{im\varphi} \tag{7-40}$$

We next consider the maximum m-state $|l, l\rangle$ for which $L_+|l, l\rangle = 0$. This implies that

$$\begin{aligned}0 = \langle\theta, \varphi|L_+|l, l\rangle &= \hbar e^{i\varphi}\left(\frac{\partial}{\partial\theta} + i\cot\theta\frac{\partial}{\partial\varphi}\right)\langle\theta, \varphi|l, l\rangle \\ = \hbar e^{i\varphi}\left(\frac{\partial}{\partial\theta} - l\cot\theta\right)\langle\theta, \varphi|l, l\rangle &= \hbar e^{i(l+1)\varphi}\left(\frac{\partial}{\partial\theta} - l\cot\theta\right)F(\theta)\end{aligned} \tag{7-41}$$

or

$$\left(\frac{\partial}{\partial\theta} - l \cot \theta\right)F(\theta) = 0$$

The solution of this equation is easily seen to be

$$F(\theta) = (\sin \theta)^l \tag{7-42}$$

States with values of m lower than l can be obtained by repeated application of the *lowering operator* L_- to the state $|l, l\rangle$.[1] This amounts to

$$\begin{aligned} Y_{lm}(\theta, \varphi) &= C(L_-)^{l-m}(\sin \theta)^l e^{il\varphi} \\ &= C\left(-\frac{\partial}{\partial\theta} + i \cot \theta \frac{\partial}{\partial\varphi}\right)^{l-m}[(\sin \theta)^l e^{il\varphi}] \end{aligned} \tag{7-43}$$

The coefficient C is there so that the eigenstates are normalized according to

$$\int d\Omega\, |Y_{lm}(\theta, \varphi)|^2 = 1 \tag{7-44}$$

The details of working out the Y_{lm} can be found in Supplement 7-B. **[www.wiley.com/college/gasiorowicz]** Here we restrict ourselves to writing down the appropriately normalized Y_{lm} with the phases that are conventionally used. We have

$$Y_{lm}(\theta, \varphi) = (-1)^m\left[\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}\right]^{1/2} P_l^m(\cos \theta)e^{im\varphi} \tag{7-45}$$

for $m \geq 0$, with

$$Y_{l,-m} = (-1)^m Y_{lm}^* \tag{7-46}$$

The *associated Legendre polynomials* are given by

$$P_l^m(u) = (-1)^{l+m}\frac{(l+m)!}{(l-m)!}\frac{(1-u^2)^{-m/2}}{2^l l!}\left(\frac{d}{du}\right)^{l-m}(1-u^2)^l \tag{7-47}$$

for $m \geq 0$. The values for negative m are obtained from

$$P_l^{-m}(u) = (-1)^m\frac{(l-m)!}{(l-m)!}P_l^m(u) \tag{7-48}$$

We list some of the spherical harmonics in the following table.

---

[1] We may relate our results here to our discussion of *degeneracy* in Chapter 5. There are $(2l + 1)$ eigenfunctions of $\mathbf{L}^2$. In other words, the eigenfunctions are $(2l + 1)$-fold degenerate We found another operator L_z that commutes with $\mathbf{L}^2$, and we were able to arrange things so that the eigenfunctions are *simultaneous* eigenfunctions of the *mutually commuting operators* $\mathbf{L}^2$ and L_z. They are *not* eigenfunctions of L_x or L_y.

Table of Spherical Harmonics

$$Y_{0,0} = \frac{1}{\sqrt{4\pi}}$$

$$Y_{1,1} = -\sqrt{\frac{3}{8\pi}}\, e^{i\varphi} \sin\theta$$

$$Y_{1,0} = \sqrt{\frac{3}{4\pi}} \cos\theta$$

$$Y_{2,2} = \sqrt{\frac{15}{32\pi}}\, e^{2i\varphi} \sin^2\theta$$

$$Y_{2,1} = -\sqrt{\frac{15}{8\pi}}\, e^{i\varphi} \sin\theta \cos\theta$$

$$Y_{2,0} = \sqrt{\frac{5}{16\pi}} (3\cos^2\theta - 1)$$

$$Y_{3,3} = -\sqrt{\frac{35}{64\pi}}\, e^{3i\varphi} \sin^3\theta$$

$$Y_{3,2} = \sqrt{\frac{105}{64\pi}}\, e^{2i\varphi} \sin^2\theta \cos\theta$$

$$Y_{3,1} = -\sqrt{\frac{21}{64\pi}}\, e^{i\varphi} \sin\theta (5\cos^2\theta - 1)$$

$$Y_{3,0} = \sqrt{\frac{7}{16\pi}} (5\cos^3\theta - 3\cos\theta)$$

EXAMPLE 7-1

Use the raising/lowering operators to obtain the (unnormalized) eigenfunctions $Y_{lm}(\theta, \varphi)$ for $l = 2$ for $m = 2, 1, 0$.

SOLUTION We start with

$$Y_{22}(\theta, \varphi) = A \sin^2\theta e^{2i\varphi}$$

To get $Y_{21}(\theta, \varphi)$ we apply the L_- operator once. From (7-36) we find that

$$Y_{21}(\theta, \varphi) \propto L_- Y_{22}(\theta, \varphi) = \hbar e^{-i\varphi}\left(-\frac{\partial}{\partial\theta} + i \cot\theta \frac{\partial}{\partial\varphi}\right)(\sin^2\theta e^{2i\varphi})$$

A little algebra yields

$$Y_{21}(\theta, \varphi) = B e^{i\varphi} \sin\theta \cos\theta$$

Repeating the application of L_-, we get

$$\begin{aligned} Y_{20}(\theta, \varphi) \propto L_- Y_{21}(\theta, \varphi) &= \hbar e^{-i\varphi}\left(-\frac{\partial}{\partial\theta} + i \cot\theta \frac{\partial}{\partial\varphi}\right) B e^{i\varphi} \sin\theta \cos\theta \\ &= B\hbar(-\cos^2\theta + \sin^2\theta - \cos^2\theta) = C(3\cos^2\theta - 1) \end{aligned}$$

so that

$$Y_{20}(\theta, \varphi) = C(3 \cos^2\theta - 1)$$

We conclude with a comment on the connection with classical angular momentum. We expect that for large l the correspondence principle will lead us to the classical limit—that is, we will find something that we can associate with an angular momentum *vector*. We note that when $m = l$ (or $m = -l$) then

$$|Y_{l,\pm l}(\theta, \varphi)|^2 \propto (\sin\theta)^{2l} \tag{7-49}$$

This implies that for large l the probability distribution is strongly confined to the equatorial region around $\theta = \pi/2$. This is as expected: For $m = l$, L_z has its largest value so that $\mathbf{L}^2$ is as close as it can get to L_z^2. In the limit of large l, the classical limit,

$$\frac{\mathbf{L}^2 - L_z^2}{\mathbf{L}^2} = \frac{l}{l(l+1)} = \frac{1}{l+1} \to 0 \tag{7-50}$$

This means that it is possible to line up the angular momentum in a particular direction, here the z-axis. Such an alignment would require

$$\langle L_x^2 \rangle = \langle L_y^2 \rangle = 0$$

which is not possible when quantum mechanical effects become important, and commutation relations cannot be ignored.

7-4 COMMENTS ON THE EXPANSION THEOREM

The general statement of the expansion theorem allows us to expand a state vector in terms of a complete set of states. Consider

$$|\psi\rangle = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} C_{lm} |l, m\rangle \tag{7-51}$$

With the help of the orthonormality relation

$$\langle l', m' | l, m \rangle = \delta_{ll'} \delta_{mm'} \tag{7-52}$$

we get $C_{lm} = \langle l, m | \psi \rangle$, so that

$$|\psi\rangle = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} |l, m\rangle\langle l, m|\psi\rangle \tag{7-53}$$

Since this is true for every $|\psi\rangle$, we get the expected completeness relation

$$\sum_{l=0}^{\infty} \sum_{m=-l}^{l} |l, m\rangle\langle l, m| = \mathbf{1} \tag{7-54}$$

where $\mathbf{1}$ is the unit operator.

If we multiply this on the left by $\langle \theta', \varphi' |$ and on the right by $|\theta, \varphi\rangle$, we get

$$\sum_{l=0}^{\infty} \sum_{m=-l}^{l} \langle \theta', \varphi' | l, m\rangle\langle l, m|\theta, \varphi\rangle = \langle \theta', \varphi' | \theta, \varphi\rangle$$

which can be written as

$$\sum_{l=0}^{\infty} \sum_{m=-l}^{l} Y_{lm}^*(\theta', \varphi') Y_{lm}(\theta, \varphi) = \frac{1}{\sin\theta} \delta(\theta - \theta')\, \delta(\varphi - \varphi') \tag{7-55}$$

We can use the expansion theorem to answer the question often raised by students. What is so special about the z-direction? Cannot we align the angular momentum (as far as possible) with the x-axis? The answer is that this is indeed possible. Such a state, which would be confined near the equatorial plane about the x-axis (in the vicinity of $\varphi = \pi/2$) will be a particular linear combination of the $Y_{lm}(\theta, \varphi)$. Its physical properties will be exactly the same as the Y_{ll} state.

PROBLEMS

1. Consider the molecule CN, which may be described by a dumbbell consisting of two masses M_1 and M_2 attached by a rigid rod of length a. The dumbbell rotates *in a plane* about an axis going through the center of mass and perpendicular to it.

   (a) Write down a Hamiltonian that describes the motion.

   (b) What is the energy spectrum?

   (c) Write down an expression for the difference in energy between the ground state and the first excited state in terms of the masses and a.

2. Express the spherical harmonics for $l = 0, 1, 2$ in terms of x, y, z.

3. Calculate $\langle l, m_1|L_x|l, m_2\rangle$ and $\langle l, m_1|L_y|l, m_2\rangle$.

4. Calculate $\langle l, m_1|L_x^2|l, m_2\rangle$ and $\langle l, m_1|L_y^2|l, m_2\rangle$. (*Hint:* Things will be easier with the use of $L_\pm$, $L_\pm^2$, and so on.)

5. The Hamiltonian for an axially symmetric rotator is

   $$H = \frac{L_x^2 + L_y^2}{2I_1} + \frac{L_z^2}{2I_3}$$

   (a) What are the eigenvalues of H?

   (b) Sketch the spectrum, assuming that $I_1 > I_3$.

   (c) What is the spectrum in the limit that I_1 is much larger than I_3?

6. Use lowering operators to calculate the angular dependence (without worrying about normalization) of $Y_{4m}(\theta. \varphi)$ for $m = 3, 2, 1, 0$. You are given $Y_{44}(\theta, \varphi) = Ae^{4i\varphi} \sin^4 \theta$.

7. A system is described by the Hamiltonian

   $$H = \frac{\mathbf{L}^2}{2I} + \alpha L_z$$

   What is the energy spectrum of the system?

8. Calculate the commutators $[x, L_x]$, $[y, L_x]$, $[z, L_x]$, $[x, L_y]$, $[y, L_y]$, $[z, L_y]$. Do you detect a pattern that will allow you to state the commutators of x, y, z with L_z?

9. Repeat the calculation with $x, \ldots$ replaced by $p_x, \ldots$. in problem 8.

10. Consider a state of total angular momentum $l = 2$. What are the eigenvalues of the operators **(a)** L_z, **(b)** $\frac{3}{5}L_x - \frac{4}{5}L_y$, and **(c)** $2L_x - 6L_y + 3L_z$?

11. A particle in a spherically symmetric potential is in a state described by the wave packet

    $$\psi(x, y, z) = C(xy + yz + zx)e^{-\alpha r^2}$$

    What is the probability that a measurement of the square of the angular momentum yields 0? What is the probability that it yields $6\hbar^2$? If the value of l is found to be 2, what are the relative probabilities for $m = 2, 1, 0, -1, -2$?

12. Consider the following model of a perfectly smooth cylinder. It is a ring of equally spaced, identical particles, with mass M/N so that the mass of the ring is M and its moment of inertia is MR^2, with R

. 9(:69 9 92 9 ,% 3! 3 .- ! 3 ; " !
! 3!
' $ " F
' "0 3 3 E /

and the other in the spaceship. The observer at rest on Earth (and also assumed to be at rest with respect to the two stars) measures the distance between the stars to be L_p, where L_p is the proper length. According to this observer, the time it takes the spaceship to complete the voyage is $\Delta t = L_p/v$. What does an observer in the moving spaceship measure for the distance between the stars? Because of time dilation, the space traveler measures a smaller time of travel: $\Delta t' = \Delta t/\gamma$. The space traveler claims to be at rest and sees the destination star as moving toward the spaceship with speed v. Because the space traveler reaches the star in the shorter time $\Delta t'$, he or she concludes that the distance, L, between the stars is shorter than L_p. This distance measured by the space traveler is given by

$$L = v\Delta t' = v\frac{\Delta t}{\gamma}$$

Because $L_p = v\Delta t$, we see that $L = L_p/\gamma$ or

$$L = L_p\left(1 - \frac{v^2}{c^2}\right)^{1/2} \tag{1.10}$$

Length contraction

where $(1 - v^2/c^2)^{1/2}$ is a factor less than 1. This result may be interpreted as follows:

> If an object has a proper length L_p when it is measured by an observer at rest with respect to the object, when it moves with speed v in a direction parallel to its length, its length L is measured to be shorter according to $L = L_p\left(1 - \frac{v^2}{c^2}\right)^{1/2}$.

Note that the length contraction takes place only along the direction of motion. For example, suppose a stick moves past a stationary Earth observer with a speed v, as in Figure 1.13b. The length of the stick as measured by an observer in the frame attached to it is the proper length L_p, as illustrated in Figure 1.13a. The length of the stick, L, as measured by the Earth observer is shorter than L_p by the factor $(1 - v^2/c^2)^{1/2}$. Note that length contraction is a symmetric effect: If the stick were at rest on Earth, an observer in a frame moving past the earth at speed v would also measure its length to be shorter by the same factor $(1 - v^2/c^2)^{1/2}$.

As we mentioned earlier, one of the basic tenets of relativity is that all inertial frames are equivalent for analyzing an experiment. Let us return to the example of the decaying muons moving at speeds close to the speed of light to see an example of this. An observer in the muon's reference frame would measure the proper lifetime, whereas an Earth-based observer measures the proper height of the mountain in Figure 1.11. In the muon's reference frame, there is no time dilation, but the distance of travel is observed to be shorter when measured from this frame. Likewise, in the Earth observer's reference frame, there is time dilation, but the distance of travel is measured to be the proper height of the mountain. Thus, when calculations on the muon are performed in both frames, one sees the effect of "offsetting penalties," and the outcome of the experiment is the same!

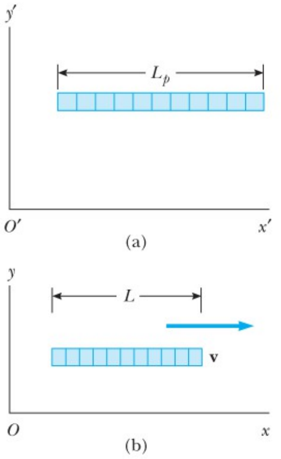

Figure 1.13 A stick moves to the right with a speed v. (a) The stick as viewed in a frame attached to it. (b) The stick as seen by an observer who sees it move past her at v. Any inertial observer finds that the length of a meter stick moving past her with speed v is less than the length of a stationary stick by a factor of $(1 - v^2/c^2)^{1/2}$.

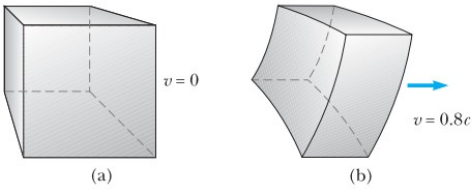

Figure 1.14 Computer-simulated photographs of a box (a) at rest relative to the camera and (b) moving at a speed $v = 0.8c$ relative to the camera.

Note that proper length and proper time are measured in *different* reference frames.

If an object in the shape of a box passing by could be photographed, its image would show length contraction, but its shape would also be distorted. This is illustrated in the computer-simulated drawings shown in Figure 1.14 for a box moving past an observer with a speed $v = 0.8c$. When the shutter of the camera is opened, it records the shape of the object at a given instant of time. Because light from different parts of the object must arrive at the shutter at the same time (when the photograph is taken), light from more distant parts of the object must start its journey earlier than light from closer parts. Hence, the photograph records different parts of the object at different times. This results in a highly distorted image, which shows horizontal length contraction, vertical curvature, and image rotation.

EXAMPLE 1.3 The Contraction of a Spaceship

A spaceship is measured to be 100 m long while it is at rest with respect to an observer. If this spaceship now flies by the observer with a speed of $0.99c$, what length will the observer find for the spaceship?

Solution The proper length of the ship is 100 m. From Equation 1.10, the length measured as the spaceship flies by is

$$L = L_p\sqrt{1 - \frac{v^2}{c^2}} = (100\text{ m})\sqrt{1 - \frac{(0.99c)^2}{c^2}} = 14\text{ m}$$

Exercise 4 If the ship moves past the observer at $0.01000c$, what length will the observer measure?

Answer 99.99 m.

EXAMPLE 1.4 How High Is the Spaceship?

An observer on Earth sees a spaceship at an altitude of 435 m moving downward toward the Earth at $0.970c$. What is the altitude of the spaceship as measured by an observer in the spaceship?

Solution The proper length here is the Earth–ship separation as seen by the Earth-based observer, or 435 m. The moving observer in the ship finds this separation (the altitude) to be

$$L = L_p\sqrt{1 - \frac{v^2}{c^2}} = (435\text{ m})\sqrt{1 - \frac{(0.970c)^2}{c^2}}$$
$$= 106\text{ m}$$

EXAMPLE 1.5 The Triangular Spaceship

A spaceship in the form of a triangle flies by an observer at $0.950c$. When the ship is measured by an observer at rest with respect to the ship (Fig. 1.15a), the distances x and y are found to be 50.0 m and 25.0 m, respectively. What is the shape of the ship as seen by an observer who sees the ship in motion along the direction shown in Figure 1.15b?

Solution The observer sees the horizontal length of the ship to be contracted to a length of

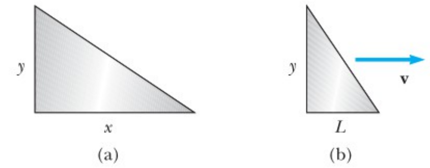

Figure 1.15 (Example 1.5) (a) When the spaceship is at rest, its shape is as shown. (b) The spaceship appears to look like this when it moves to the right with a speed v. Note that only its x dimension is contracted in this case.

$$L = L_p\sqrt{1-\frac{v^2}{c^2}}$$

$$= (50.0\text{ m})\sqrt{1-\frac{(0.950c)^2}{c^2}} = 15.6\text{ m}$$

The 25-m vertical height is unchanged because it is perpendicular to the direction of relative motion between the observer and the spaceship. Figure 1.15b represents the shape of the spaceship as seen by the observer who sees the ship in motion.

THE TWINS PARADOX

OPTIONAL

If we placed a living organism in a box . . . one could arrange that the organism, after an arbitrary lengthy flight, could be returned to its original spot in a scarcely altered condition, while corresponding organisms which had remained in their original positions had long since given way to new generations. **(Einstein's original statement of the twins paradox in 1911)**

An intriguing consequence of time dilation is the so-called clock or twins paradox. Consider an experiment involving a set of identical 20-year-old twins named Speedo and Goslo. The twins carry with them identical clocks that have been synchronized. Speedo, the more adventuresome of the two, sets out on an epic journey to planet X, 10 lightyears from Earth. (Note that 1 lightyear (ly) is the distance light travels through free space in 1 year.) Furthermore, his spaceship is capable of a speed of $0.500c$ relative to the inertial frame of his twin brother. After reaching planet X, Speedo becomes homesick and impetuously sets out on a return trip to Earth at the same high speed of the outbound journey. On his return, Speedo is shocked to discover that many things have changed during his absence. To Speedo, the most significant change is that his twin brother Goslo *has aged more than he* and is now 60 years of age. Speedo, on the other hand, has aged by only 34.6 years.

At this point, it is fair to raise the following question—Which twin is the traveler and which twin would really be the younger of the two? If motion is relative, the twins are in a symmetric situation and either's point of view is equally valid. From Speedo's perspective, it is he who is at rest while Goslo is on a high-speed space journey. To Speedo, it is Goslo and the Earth that have raced away on a 17.3-year journey and then headed back for another 17.3 years. This leads to the paradox: Which twin will have developed the signs of excess aging?

To resolve this apparent paradox, recall that special relativity deals with inertial frames of reference moving with respect to one another at *uniform speed*. However, the trip situation is not symmetric. Speedo, the space traveler, must experience acceleration during his journey. As a result, his state of motion is not always uniform, and consequently Speedo is not in an inertial frame. He cannot regard himself to always be at rest and Goslo to be in uniform motion. Hence Speedo cannot apply simple time dilation to Goslo's motion, because to do so would be an incorrect application of special relativity. Therefore there is no paradox and Speedo will really be the younger twin at the end of the trip.

The conclusion that Speedo is not in a *single* inertial frame is inescapable. We may diminish the length of time needed to accelerate and decelerate Speedo's spaceship to an insignificant interval by using very large and expensive rockets and claim that he spends all but a negligible amount of time coasting to planet X at $0.500c$ in an inertial frame. However, to return to Earth, Speedo must slow down, reverse his motion, and return in a different inertial frame, one which is moving uniformly toward the Earth. At the very best, Speedo is in *two* different inertial frames. The important point is that even when we idealize Speedo's trip, it consists of motion in two different inertial frames and a very real lurch as he hops from the outbound ship to the returning Earth shuttle. Only Goslo remains in a single inertial frame, and so only he can correctly apply the simple time dilation formula of special relativity to Speedo's trip. Thus, Goslo finds that instead of aging 40 years (20 ly/$0.500c$), Speedo actually ages only $(\sqrt{1 - v^2/c^2})(40 \text{ yr})$, or 34.6 yr. Clearly, Speedo spends 17.3 years going to planet X and 17.3 years returning in agreement with our earlier statement.

The result that Speedo ages 34.6 yr while Goslo ages 40 yr can be confirmed in a very direct experimental way from Speedo's frame if we use the special theory of relativity but take into account the fact that *Speedo's idealized trip takes place in two different inertial frames.* In yet another flight of fancy, suppose that Goslo celebrates his birthday each year in a flashy way, sending a powerful laser pulse to inform his twin that Goslo is another year older and wiser. Because Speedo is in an inertial frame on the outbound trip in which the Earth appears to be receding at $0.500c$, the flashes occur at a rate of one every

$$\frac{1}{\sqrt{1 - (v^2/c^2)}}\text{ yr} = \frac{1}{\sqrt{1 - [(0.500c)^2/c^2]}}\text{ yr} = 1.15 \text{ yr}$$

This occurs because moving clocks run slower. Also, because the Earth is receding, each successive flash must travel an additional distance of $(0.500c)(1.15 \text{ yr})$ between flashes. Consequently, Speedo observes flashes to arrive with a total time between flashes of $1.15 \text{ yr} + (0.500c)(1.15 \text{ yr})/c = 1.73$ yr. The total number of flashes seen by Speedo on his outbound voyage is therefore $(1 \text{ flash}/1.73 \text{ yr})(17.3 \text{ yr}) =$ 10 flashes. This means that Speedo views the Earth clocks to run more slowly than his own on the outbound trip because he observes 17.3 years to have passed for him while only 10 years have passed on Earth.

On the return voyage, because the Earth is racing toward Speedo with speed $0.500c$, successive flashes have less distance to travel, and the total time Speedo sees between the arrival of flashes is drastically shortened: $1.15 \text{ yr} - (0.500)(1.15 \text{ yr}) = 0.577$ yr/flash. Thus, during the return trip, Speedo sees $(1 \text{ flash}/0.577 \text{ yr})(17.3 \text{ yr}) = 30$ flashes in total. In sum, during his 34.6 years of travel, Speedo receives $(10 + 30)$ flashes, indicating that his twin has aged 40 years. Notice that there has been no failure of special relativity for Speedo as long as we take his *two* inertial frames into account and assume negligible acceleration and deceleration times. On both the outbound and inbound trips Speedo correctly judges the Earth clocks to run slower than his own, but on the return trip his rapid movement toward the light flashes more than compensates for the slower rate of flashing.

Figure 1.16 "I love hearing that lonesome wail of the train whistle as the frequency of the wave changes due to the Doppler effect."

The Relativistic Doppler Shift

Another important consequence of time dilation is the shift in frequency found for light emitted by atoms in motion as opposed to light emitted by atoms at rest. A similar phenomenon, the mournful drop in pitch of the sound of a passing train's whistle, known as the Doppler effect, is quite familiar to most cowboys (Fig. 1.16). The Doppler shift for sound is usually

9(:69 9 92 9 ,% 3! 3 .- ! 3 "!
! 3!

1 THE MICHELSON INTERFEROMETER

The Michelson interferometer, first introduced by Albert Michelson in 1881, has played a vital role in the development of modern physics. This simple and versatile instrument was used, for example, to establish experimental evidence for the validity of the special theory of relativity, to detect and measure hyperfine structure in line spectra, to measure the tidal effect of the moon on the earth, and to provide a substitute standard for the meter in terms of wavelengths of light. Michelson himself pioneered much of this work.

A schematic of the Michelson interferometer is shown in Figure 1a. From an extended source of light *S*, beam 1 of light is split by a beam splitter (BS) by means of a thin, semitransparent front surface metallic or dielectric film, deposited on glass. The interferometer is therefore of the amplitude-splitting type. Reflected beam 2 and transmitted beam 3, of roughly equal amplitudes, continue to fully-reflecting mirrors *M*2 and *M*1, respectively, where their directions are reversed. On returning to the beam splitter, beam 2 is now transmitted and beam 3 is reflected by the semitransparent film so that they come together again and leave the interferometer as beam 4. The useful aperture of this double-beam interferometer is such that all rays striking *M*1 and *M*2 will be normal, or nearly so. Thus, beam 4 includes rays that have traveled different optical paths and will demonstrate interference. At least one of the mirrors is equipped with tilting adjustment screws that allow the surface of *M*1 to be made perpendicular to that of *M*2. One of the mirrors is also movable along the direction of the beam by means of an accurate track and micrometer screw. In this way, the difference between the optical paths of

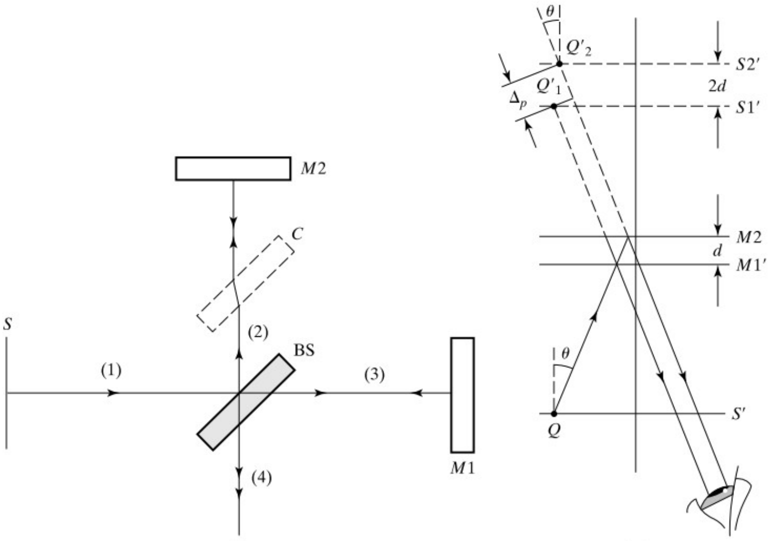

Figure 1 (a) The Michelson interferometer. (b) Equivalent optics for the Michelson interferometer.

beams 2 and 3 can be gradually varied. Notice that beam 3 traverses the beam splitter three times, whereas beam 2 traverses it only once. In some applications, where white light is used, it is essential that the optical paths of the two beams be made precisely equal. Although this can be accomplished at one wavelength by appropriately increasing the distance of $M2$ from BS, the correction would not suffice at another wavelength because of the dispersion of the glass. To compensate for all wavelengths at once, a compensator plate C made of the same material and dimensions as BS is inserted parallel to BS in the path of beam 2. Any small, remaining inequalities in optical paths can be removed by allowing the compensator to rotate, thus varying the optical path through the thickness of its glass plate.

The actual interferometer in Figure 1a possesses two optical axes at right angles to one another. A simpler but equivalent optical system, having a single optical axis, can be drawn by working with virtual images of source S and mirror $M1$ via reflection in the BS mirror. These positions are most simply found by regarding the assembly including S, $M1$, and beams 1 and 3 of Figure 1a as rotated counterclockwise by 90° about the point of intersection of the beams with the BS mirror. The resulting geometry is shown in Figure 1b. The new position of the source plane is S', and the new position of the mirror $M1$ is $M1'$. Light from a point Q on the source plane S' then effectively reflects from both mirrors $M2$ and $M1'$, shown parallel and with an optical path difference of d. The two reflected beams appear to come from the two virtual images, Q_1' and Q_2', of object point Q. Since the images $S1'$ and $S2'$ of the source plane in the mirrors must be separated by twice the mirror separation, the distance between Q_1' and Q_2' is $2d$, and the optical-path difference Δ_p between the two beams emerging from the interferometer is

$$\Delta_p = 2d \cos\theta \tag{1}$$

where the angle θ measures the inclination of the beams relative to the optical axis. For a normal beam, $\theta = 0$ and $\Delta_p = 2d$. We expect this result, since, if one mirror is farther from BS than the other by a distance d, the extra distance traversed by the beam taking the longer route includes distance d twice, once before and once after reflection. If, in addition, $\Delta = m\lambda$, so that the two beams interfere constructively, it follows that they will do so repeatedly for every $\lambda/2$ translation of one of the mirrors so long as the separation Δ_p does not exceed the so-called *coherence length*, l_t, of the source. The coherence length is the length along a wave train over which the phase of the wave remains correlated. The coherence length of the light emitted by a particular source is the ratio of the speed of light c to the spread of frequencies $\Delta\nu$ present in the source. The coherence length of typical laser sources ranges from tens of centimeters to tens of kilometers. Throughout this chapter, we will assume that all effective path-length differences between interfering beams that originate from the same source are much less than the coherence length of the source.

The optical system of Figure 1b is now equivalent to the case of interference due to a plane, parallel air film, illuminated by an extended source. Virtual fringes of equal inclination may be seen by looking into the beam splitter along ray 4, with the eye or a telescope focused at infinity. Assuming that the two interfering beams are of equal amplitude, the irradiance of the fringe system of circles concentric with the optical axis is given by

$$I = 4I_0 \cos^2\left(\frac{\delta}{2}\right) \tag{2}$$

where the phase difference is

$$\delta = k\Delta = \left(\frac{2\pi}{\lambda}\right)\Delta \tag{3}$$

The net optical path difference is $\Delta = \Delta_p + \Delta_r$, as usual. A relative π phase shift between the two beams occurs because the reflection coefficients from opposite sides of a beam splitter differ by $-1 = e^{i\pi}$.[1] For dark fringes, then,

$$\Delta_p + \Delta_r = 2d\cos\theta + \frac{\lambda}{2} = \left(m + \frac{1}{2}\right)\lambda$$

or, more simply,

$$2d\cos\theta = m\lambda \qquad m = 0, 1, 2, \ldots \text{ dark fringes} \tag{4}$$

If d is of such magnitude that the normal rays forming the center of the fringe system satisfy Eq. (4), that is, the center fringe is dark, then its order, given by

$$m_{\max} = \frac{2d}{\lambda} \tag{5}$$

is a large integer. Neighboring dark fringes decrease in order outwards from the center of the pattern, as $\cos\theta$ decreases from its maximum value of 1. This ordering of fringes may be inverted for convenience by associating another integer p with each fringe of order m, where

$$p = m_{\max} - m = \frac{2d}{\lambda} - m \tag{6}$$

Using Eq. (6) to replace m in Eq. (4), we arrive at

$$p\lambda = 2d(1 - \cos\theta) \qquad p = 0, 1, 2, \ldots \text{ dark fringes} \tag{7}$$

where now the central fringe is of order zero and the neighboring fringes increase in order, outward from the center. Figure 2 illustrates the relationship between orders m and p for the arbitrary case where $m_{\max} = 100$. Equation (4) or (7) indicates that, as d is varied, a particular point in the fringe pattern (θ = constant) will correspond to gradually changing values of order m or p.

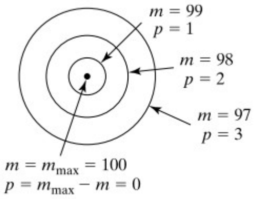

Figure 2 Alternate orderings of fringes.

Integral values occur whenever the point coincides with a dark fringe. Equivalently, this means that as d is varied, fringes of the pattern appear to move inward toward the center, where they disappear, or else move outward from the center, where they seem to originate, depending on whether the optical-path difference is decreasing or increasing. The motion of the fringe pattern thus reverses as one of the mirrors is moved continually through the point of zero path difference. Viewed in another way, Eq. (4) requires an increase in the angular separation $\Delta\theta$ of a given small fringe interval Δm as the mirror spacing d becomes smaller, since taking the differential of Eq. (4) leads to

$$|\Delta\theta| = \frac{\lambda\Delta m}{2d\sin\theta}$$

---

[1] This conclusion assumes real beam-splitter transmission coefficients. The assumption is made to allow discussion of the fringe pattern in a concrete (and common) situation. It does not affect the validity of results like Eq. (8), for example, because measurements depend on the net *motion* of the fringe pattern, not on precisely where it is dark and where it is bright.

This means that the fringes are more widely separated when optical-path differences are small. In fact, if $d = \lambda/2$, then from Eq. (4), $m = \cos\theta$, and the entire field of view encompasses no more than one fringe! For a mirror translation Δd, the number Δm of fringes passing a point at or near the center of the pattern is, according to Eq. (4),

$$\Delta m = \frac{2\Delta d}{\lambda} \tag{8}$$

Equation (8) suggests an experimental way of either measuring λ when Δd is known or calibrating the micrometer translation screw when λ is known.

Example 1

Fringes are observed due to monochromatic light in a Michelson interferometer. When the movable mirror is translated by 0.073 mm, a shift of 300 fringes is observed. What is the wavelength of the light? What displacement of the fringe system takes place when a flake of glass of index 1.51 and 0.005 mm thickness is placed in one arm of the interferometer? (Assume that the light beam is normal to the glass surface.)

Solution

Using Eq. (8),

$$\lambda = \frac{2\Delta d}{\Delta m} = \frac{(2)(0.073)}{300} = 4.87 \times 10^{-4}\text{ mm} = 487\text{ nm}$$

With the glass inserted, one arm is effectively lengthened by a path difference of $\Delta d = n_g t - n_{\text{air}} t$, so that

$$\Delta m = \frac{2\Delta d}{\lambda} = \frac{2(n_g - 1)t}{\lambda} = \frac{2(0.51)(0.005 \times 10^{-3})}{487 \times 10^{-9}} = 10.5\text{ fringes}$$

2 APPLICATIONS OF THE MICHELSON INTERFEROMETER

The Michelson interferometer is easily adaptable to the measurement of thin films, a technique essentially the same as that described in the preceding chapter. It can also be adapted to measure the index of refraction of a gas. An evacuable cell with plane, parallel windows is interposed in the path of beam 3 (Figure 1a) and is filled with a gas at a pressure and temperature for which its index of refraction is desired. The fringe system established under these conditions is monitored as the gas is gradually pumped out of the cell. A count Δm of the net fringe shift is related to the change in optical path during the decrease of the gas pressure. If the actual length of the cell is accurately known to be L, the change in optical path is given by

$$\Delta d = nL - L = L(n - 1) \tag{9}$$

and using Eq. (8), it follows that the index can be determined from

$$n - 1 = \left(\frac{\lambda}{2L}\right)\Delta m \tag{10}$$

Consider another direct application of the Michelson interferometer, the determination of wavelength difference between two closely spaced components of a spectral "line," λ and λ'. Each wavelength forms its own system of circular fringes according to Eq. (4). Suppose we view the circular systems near their center, so that $\cos\theta \cong 1$. Then for a given path difference d of the interferometer, the product $m\lambda$ is fixed, that is, $m\lambda = m'\lambda'$. When the fringe systems coincide, the pattern appears sharp, whereas when the fringes of one system in the region of observation lie midway between the fringes of the second system, the pattern appears rather uniform in brightness, or "washed out." The mirror movement Δd required between consecutive coincidences is related to the wavelength difference $\Delta\lambda$ as follows. At one coincidence, when fringes are "in step," the orders of the two systems corresponding to λ and λ' must be related by

$$m = m' + N$$

where N is an integer. If the optical-path difference at this time is d_1, then from Eq. (4),

$$\frac{2d_1}{\lambda} = \frac{2d_1}{\lambda'} + N \tag{11}$$

Let the optical-path difference be increased to d_2, when the next coincidence is found. Then,

$$m = m' + (N + 1)$$

or

$$\frac{2d_2}{\lambda} = \frac{2d_2}{\lambda'} + N + 1 \tag{12}$$

By subtracting Eq. (11) from Eq. (12) and by writing the mirror movement $\Delta d = d_2 - d_1$, we find

$$\lambda' - \lambda = \frac{\lambda\lambda'}{2\Delta d} \tag{13}$$

Now since λ and λ' are very close, the wavelength difference of the two unresolved components can be approximated by

$$\Delta\lambda = \frac{\lambda^2}{2\Delta d} \tag{14}$$

This technique is often employed in an optics laboratory course to measure the wavelength difference of 6 Å between the two components of the yellow "line" of sodium.

All the preceding discussion of the fringes from a Michelson interferometer has been in terms of virtual fringes of equal inclination. We have assumed that mirrors $M1$ and $M2$ are precisely perpendicular as shown in Figure 1a, or, what amounts to the same thing, precisely parallel in the equivalent optical system of Figure 1b. If the alignment is such that the air space between $M1'$ and $M2$ in Figure 1b is a wedge, fringes of equal thickness may be seen localized at the mirrors. These fringes will be straight, oriented parallel to the line that represents the intersection of $M1'$ and $M2$. If the wedge is of large angle, they will be curved in a way that can be shown to be hyperbolic arcs. Again, if the source is small, then real, nonlocalized fringes appear in the light emerging from the interferometer, as if formed by the two virtual images of the source in $M1'$ and $M2$. These fringes appear without effort when the intense, coherent light of a laser is used. These possibilities have already been discussed in the previous chapter, where we treated the various interference fringes that can be formed by illumination of a film. Figure 3 is a photograph showing the distortion of fringes of equal thickness produced by a candle flame when situated in one arm of a Michelson interferometer. Variations in temperature produce variations in optical-path length by changing the refractive index of the air.

! 9(:69 9 92 9 ,% 3! 3 .- ! 3 ;
! ; " ! 3!
' "/

> Fritz Haber is known for developing the means for synthesizing ammonia and developing fertilizers. He won the Nobel Prize in chemistry in 1918. He is also known for the Born–Haber cycle for finding he lattice energy of ionic solids. However, he was prominent as the father of chemical warfare for developing and directing the use of chorine and other poison gases in war. His first wife committed suicide. Some say that was because the involvement of Haber with the use of poison gases, others say it was because of his alleged infidelity.

1.1.3 Metals and Wigner–Seitz Theory (B)

Examples of metals are sodium (Na) and copper (Cu). A metal such as Na is viewed as being composed of positive ion cores (Na^+) immersed in a "sea" of free conduction electrons that come from the removal of the 3s electron from atomic Na. Metallic binding can be partly understood within the context of the Wigner–Seitz theory. In a full treatment, it would be necessary to confront the problem of electrons in a periodic lattice. (A discussion of the Wigner–Seitz theory will be deferred until Chap. 3.) One reason for the binding is the lowering of the kinetic energy of the "free" electrons relative to their kinetic energy in the atomic 3s state [41]. In a metallic crystal, the valence electrons are free (within the constraints of the Pauli principle) to wander throughout the crystal, causing them to have a smoother wave function and hence less $\nabla^2\psi$. Generally speaking this spreading of the electrons wave function also allows the electrons to make better use of the attractive potential. Lowering of the kinetic and/or potential energy implies binding. However, the electron–electron Coulomb repulsions cannot be neglected (see, e.g., Sect. 3.1.4), and the whole subject of binding in metals is not on so good a quantitative basis as it is in crystals involving the interactions of atoms or molecules which do not have free electrons. One reason why the metallic crystal is prevented from collapsing is the kinetic energy of the electrons. Compressing the solid causes the wave functions of the electrons to "wiggle" more and hence raises their kinetic energy.

A very simple picture[6] suffices to give part of the idea of metallic binding. The ground-state energy of an electron of mass M in a box of volume V is [19]

$$E = \frac{\hbar^2\pi^2}{2M}V^{-2/3}.$$

---

[6] A much more sophisticated approach to the binding of metals is contained in the pedagogical article by Tran and Perdew [1.26]. This article shows how exchange and correlation effects are important and discusses modern density functional methods (see Chap. 3).

Thus the energy of N electrons in N separate boxes is

$$E_A = N\frac{\hbar^2\pi^2}{2M}V^{-2/3}. \tag{1.21}$$

The energy of N electrons in a box of volume NV is (neglecting electron–electron interaction that would tend to increase the energy)

$$E_M = N\frac{\hbar^2\pi^2}{2M}V^{-2/3}N^{-2/3}. \tag{1.22}$$

Therefore $E_M/E_A = N^{-2/3} \ll 1$ for large N and hence the total energy is lowered considerably by letting the electrons spread out. This model of binding is, of course, not adequate for a real metal, since the model neglects not only electron–electron interactions but also the potential energy of interaction between electrons and ions and between ions and other ions. It also ignores the fact that electrons fill up states by satisfying the Pauli principle. That is, they fill up in increasing energy. But it does clearly show how the energy can be lowered by allowing the electronic wave functions to spread out.

In modern times, considerable progress has been made in understanding the cohesion of metals by the density functional method, see Chap. 3. We mention in particular, Daw [1.6].

Due to the important role of the free electrons in binding, metals are good electrical and thermal conductors. They have moderate to fairly strong binding. We do not think of the binding forces in metals as being two-body, central, or short-range.

1.1.4 Valence Crystals and Heitler–London Theory (B)

An example of a valence crystal is carbon in diamond form. One can think of the whole valence crystal as being a huge chemically saturated molecule. As in the case of metals, it is not possible to understand completely the binding of valence crystals without considerable quantum-mechanical calculations, and even then the results are likely to be only qualitative. The quantum-mechanical considerations (Heitler–London theory) will be deferred until Chap. 3.

Some insight into covalent bonds (also called homopolar bonds) of valence crystals can be gained by considering them as being caused by sharing electrons between atoms with unfilled shells. Sharing of electrons can lower the energy because the electrons can get into lower energy states without violating the Pauli principle. In carbon, each atom has four electrons that participate in the valence

9(:69 9 92 9 ,% 3! 3 .- 3!
' 0 0/ 3

Chapter 2

VECTOR SPACES AND COORDINATE SYSTEMS

In this chapter, the concept of an inner product vector space will be developed. Coordinate maps will be introduced, and common orthogonal curvilinear coordinate systems will be defined. Curves and surfaces will be treated as manifolds embedded in Euclidean space. This chapter provides a basic overview of the elements of vector calculus. Detail properties of the rotational group will be discussed later in Chapter 7.

2.1 POINTS AND VECTORS

Space and time define a continuous differentiable arena or manifold in which particles and fields interact. In nonrelativistic mechanics, space is considered to form a three-dimensional continuum, while time forms a separate one-dimensional continuum. In relativistic mechanics, *spacetime* forms a single four-dimensional continuum of indefinite metric signature. The nonrelativistic picture will be used for now.

A physical particle is a point-like object with a well-defined position and velocity, to which additional physical attributes, such as mass and charge, can be assigned. A point is an entity with a location but no extent. The motion of a particle is given by the trajectory of its point. *Points* are fundamental to the language of mathematics. In mathematics, any form of space can be constructed from elemental points.

Newton's fiction of absolute space, existing independent of matter, is perhaps the easiest way to introduce the concept of a geometric vector space.

Consider a point **P** in an absolute three-dimensional Euclidean space, parameterized by three, independent, continuous Cartesian coordinates $(x^1, x^2, x^3) = (x, y, z)$ with respect to a constant, orthogonal Euclidean basis $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$. The origin $\mathbf{O} = (0, 0, 0)$ is embedded in the space. The location of a point defined on E^3 with respect to a Cartesian basis is given by[1]

$$\mathbf{P}(x, y, z) = x^i \mathbf{e}_i = x\,\mathbf{e}_1 + y\,\mathbf{e}_2 + z\,\mathbf{e}_3. \tag{2.1}$$

2.1.1 Displacement

The *displacement* of a point from the origin is given by

$$\mathbf{r} = \mathbf{P} - \mathbf{O}. \tag{2.2}$$

Infinitesimal displacements between points are vectors. With respect to a fixed origin, one gets

$$d\mathbf{r} = dx^i \mathbf{e}_i. \tag{2.3}$$

The Euclidean basis of absolute space is constant in time and position independent; therefore, all of their partial derivatives vanish

$$\partial \mathbf{e}_i / \partial x^j = 0 \quad \partial \mathbf{e}_i / \partial t = 0. \tag{2.4}$$

Euclidean space is translationally and rotationally invariant. Euclidean space is an example of a flat metric space. It has no intrinsic curvature. The length of a relative displacement between two points in this space is given by the *Pythagorean theorem*:

$$(d\mathbf{r})^2 = \delta_{ij} dx^i dx^j = (dx)^2 + (dy)^2 + (dz)^2. \tag{2.5}$$

Infinitesimal displacements $d\mathbf{r}$ are the prototypes for all vectors. A point on a general manifold is not a vector, since a point has a fixed origin of reference, while a vector quantity, like the velocity, only has a magnitude and direction. This distinction becomes important when discussing curved spaces. A common example of a vector is a force—force has a magnitude and an orientation. Multiple forces acting on an object point are summed according to the parallelogram law of vector addition.

The chain rule for partial derivates can be used to map infinitesimal displacements into multiple coordinate systems

$$\begin{gathered} d\mathbf{r}(\xi(x, y, z)) = d\xi^i \frac{\partial \mathbf{r}}{\partial \xi^i}, \\ d\mathbf{r} = dx^i \mathbf{e}_i = d\xi^i \boldsymbol{\lambda}_i. \end{gathered} \tag{2.6}$$

---

[1] Please note that the *Einstein summation convention* will be used throughout this book: the sum (contraction) over repeated raised and lowered tensor indices is implied in an equation unless otherwise noted by use of the acronym (nsc) denoting "No Summation Convention."

A coordinate map or *patch* is allowable in any region of space where it is differentiable and invertible (non-singular). The chain rule yields a *form invariant* expression for generalized *contravariant* elements of displacement $d\xi^i$ with respect to a *covariant* vector basis $\boldsymbol{\lambda}_i$ defined by

$$\boldsymbol{\lambda}_i = \frac{\partial \mathbf{r}}{\partial \xi^i}. \tag{2.7}$$

A space point can be located by specifying its generalized coordinates ξ^i. The motion of the point is a parameterized curve called its *trajectory*

$$\mathbf{r}(t) = \mathbf{r}(\xi^i(t)), \tag{2.8}$$

where the function $\mathbf{r}(t)$ is twice differentiable with respect to the time. Since a particle has a well-defined location at any instance of time, the equation of a particle trajectory must be a *single-valued function* of the time.

2.1.2 Velocity

The velocity of a particle is the vector tangent to the trajectory obtained by taking the total time derivative of its position

$$\mathbf{v} = \frac{d\mathbf{r}}{dt} = \dot{\xi}^i \boldsymbol{\lambda}_i. \tag{2.9}$$

A *vector* is a special case of a *tensor*, or a form-invariant object. The components of velocity $\dot{\xi}^i$ with respect to a general tensor basis $\boldsymbol{\lambda}_i$ are its *generalized velocities*. The tensor itself is an invariant object, but its components transform by covariance under a coordinate transformation. Vectors are rank-1 tensors and can be used to create tensors of greater rank. Vectors form the foundation for vector fields and vector analysis. Tensors of arbitrary rank can be expressed as multidimensional arrays with respect to some tensor basis, although as objects, in and of themselves, they are invariants, independent of any choice of coordinates. They conform to Einstein's principle of relativity, which asserts that the inherent properties of physical objects should not depend on choice of reference frame used to represent them.

2.2 INNER PRODUCT SPACES

Geometric vector spaces are inner product spaces with a metric, which provide a measure of separation. They are defined over the field of real numbers. If the space has a mixed metric signature, such as Minkowski Space, the

E 3 # # " , /0 rH & ,0

data 25.09.2019

Ⓓ Metals and Wigner-Seitz theory^x (B)
(1.1.3)

(Source: 20190924/Quantum_Physics/
epdf.pub-quantum-physics-3rd)
(não)

Source: 20190924/Solid-State Physics
Introduction to the Theory/
epdf.pub_solid-state....pdf

Pág. 11 (34)

The ground-state energy of an electron of mass M in a box of volume V is [19]

$$E = \frac{\hbar^2 \pi^2}{2M} V^{-2/3}.$$

Ⓔ Vector Spaces and Coordinate Systems

Source: 20190924/Newton/
epdf.pub_classical_
mechanics.pdf

Pág. 57 (74) — Chapter 2.

(97) 20190924 — (20190925)

Ⓘ https://einsteinpapers.press.
princeton.edu/vol6-trans/42

25.09.2019

The Formal Foundation of the
General theory of Relativity.

(II) www.mpiwg-berlin.mpg.de/sites/default/files/Preprints/P118.pdf

Source: 20190925/the Einstein theory of Relativity/P118.pdf

Hilbert's Foundation of Physics: from a theory of everything to a constituent of general relativity

Stachel.

(III) www.bu.edu/cphs/files/2015/04/2007-Renn-Stachel.pdf

Source: 20190925/the Einstein theory of Relativity/2007-Renn-Stachel.pdf

(98) The Physical and Mathematical Foundations of the theory of Relativity A Critical Analysis by Antonio Romano.
2019 - 498 pag.
ISBN 3030242362

→

%% *+,-* * *. * &') /"

0%123,45 6.789+4%.8 .6 0: % 6,. + 403.,: .6 3 3,:40%8; 4. +
.8 4%47384 .6 ;383,+1 ,31+4% %4:
<=,;38 ,388 +89 <.08 4+ 031
39%439 2: 436+8 0+<97>

%%% *+,-* * *. * &' '?, /"

0%123,45 6.789+4%.8 .6 0: % 6,. + 403.,: .6 3 3,:40%8; 4. +
.8 4%47384 .6 ;383,+1 ,31+4% %4:

% &)

+,- * *. * &) ! ! ' !@ /"

.A 1 B " CCC

A *+,-* * *. * &) 4 / 4 D , B D?
6 " E 4 DE E +// @E / % A /"

% &&

+,- * *. * && &) !") /"

%

+,- * *. * A " A B
/ F B / &E , /"

A *+,-* * *. * (" " E ;
0 /"

( ))

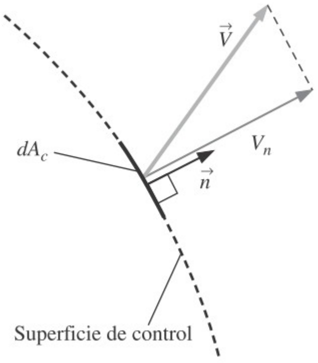
$\vec{V}$
dA_c
V_n
$\vec{n}$
Superficie de control

sideración sólo a las formas mecánicas de la energía. La conservación de la energía se trata en la sección 5-6

5-2 ▪ CONSERVACIÓN DE LA MASA

El principio de conservación de la masa es uno de los principios más fundamentales de la naturaleza. Todos estamos familiarizados con este principio y es fácil entenderlo. Una persona no tiene que ser un científico para imaginarse cuánto aderezo de vinagre y aceite se obtendrá cuando se mezclan 100 g de aceite con 25 g de vinagre. Inclusive las ecuaciones químicas se balancean con base en el principio de conservación de la masa. Cuando 16 kg de oxígeno reaccionan con 2 kg de hidrógeno, se forman 18 kg de agua (Fig. 5-2). En un proceso electrolítico, el agua se separará de vuelta a 2 kg de hidrógeno y 16 kg de oxígeno.

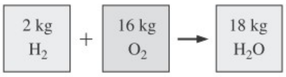

FIGURA 5-2
La masa se conserva, inclusive, durante las reacciones químicas.

Realmente, no es exacto confirmar que la masa se conserva. Ocurre que la masa m y la energía E pueden convertirse una en otra, según la conocida fórmula propuesta por Albert Einstein (1879-1955):

$$E = mc^2 \tag{5-3}$$

donde c es la velocidad de la luz en el vacío, la cual es $c = 2.9979 \times 10^8$ m/s. Esta ecuación sugiere que hay equivalencia entre masa y energía. Todos los sistemas físicos y químicos muestran interacciones de energía con su alrededor; pero la cantidad de energía que participa es equivalente a una masa extremadamente pequeña en comparación con la masa total del sistema. Por ejemplo, cuando se forma 1 kg de agua a partir de oxígeno e hidrógeno, a las condiciones atmosféricas comunes la cantidad de energía liberada es de 15.8 MJ, lo cual corresponde a una masa de 1.76×10^{-10} kg. Sin embargo, en las reacciones nucleares, la equivalencia en masa de la cantidad de energía que interactúa es una fracción importante de la masa total que participa. Por lo tanto, en la mayoría de los análisis de ingeniería, consideramos tanto la masa como la energía como cantidades conservadas.

Para los *sistemas cerrados*, el principio de conservación de la masa se usa de manera implícita cuando se establece que la masa del sistema permanezca constante durante el proceso. Sin embargo, para los *volúmenes de control*, la masa puede cruzar las fronteras y, por consiguiente, se debe considerar la razón de la masa que entra y que sale del volumen de control.

Gastos de masa y de volumen

La cantidad de masa que fluye a través de una sección transversal por unidad de tiempo se llama **razón de flujo de masa**, o simplemente flujo o gasto másico, se denota por $\dot{m}$. Se pone un punto sobre el símbolo para indicar *razón de cambio respecto al tiempo*.

Un fluido fluye hacia dentro o hacia fuera de un volumen de control normalmente por tubos o ductos. El gasto diferencial de masa de fluido que fluye a través de un pequeño elemento de área, dA_c (el subíndice corresponde a la primera letra de la palabra inglesa *cross-section*), en una sección transversal de tubo, es proporcional al propio dA_c, a la densidad ρ del fluido y a la componente de la velocidad del flujo normal a dA_c, la cual se denota como V_n, (Fig. 5-3) y se expresa de la siguiente manera:

$$\delta\dot{m} = \rho V_n \, dA_c \tag{5-4}$$

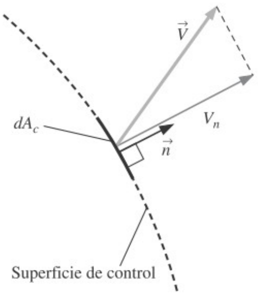

FIGURA 5-3
La velocidad normal V_n para una superficie es la componente de la velocidad perpendicular a esa superficie.

Note que se usan tanto δ como d para indicar las cantidades diferenciales, pero, por lo general, δ se usa para cantidades (como calor, trabajo y transferencia de masa) que son *funciones de trayectoria* y tienen *diferenciales inexactas*, en tanto que d se usa para cantidades (como las propiedades) que son *funciones de punto* y tienen *diferenciales exactas*. Por ejemplo, para el flujo en el tubo exterior de un ducto que consta de dos tubos concéntricos de radio interior r_1 y radio exterior r_2, se tiene $\int_1^2 dA_c = A_{c2} - A_{c1} = \pi(r_2^2 - r_1^2)$, pero $\int_1^2 \delta\dot{m} = \dot{m}_{\text{total}}$ (razón total de flujo de masa total en el tubo exterior), no $\dot{m}_2 - \dot{m}_1$. Para valores especificados de r_1 y r_2, el valor de la integral de dA_c es fijo (de allí provienen los nombres de función de punto y diferencial exacta), pero éste no es el caso para la integral de $\delta\dot{m}$ (de ahí provienen los nombres de función de trayectoria y diferencial inexacta).

La razón de flujo de masa que cruza toda el área de la sección transversal de un tubo o de un ducto se obtiene por integración:

$$\dot{m} = \int_{A_c} \delta\dot{m} = \int_{A_c} \rho V_n \, dA_c \qquad \text{(kg/s)} \tag{5-5}$$

No obstante que la ecuación 5-5 siempre es válida (de hecho, es *exacta*), no siempre es práctica porque implica integración. En lugar de ello, resultaría conveniente expresar el flujo másico en términos de valores promediados sobre la sección transversal del tubo. En un flujo compresible, ρ y V_n varían a lo largo de la sección transversal. Sin embargo, en muchas aplicaciones prácticas, la densidad es esencialmente uniforme sobre la sección transversal del tubo y se puede extraer ρ afuera de la integral de la ecuación 5-5. Pero la velocidad *nunca* es uniforme sobre una sección transversal de un tubo debido a la condición de no deslizamiento en las paredes. Más bien, la velocidad varía desde cero en las paredes hasta algún valor máximo en la línea central del tubo o cerca de éste. Se define la **velocidad promedio** V_{prom} como el valor de V_n promediado a lo largo de toda la sección transversal del tubo (Fig. 5-4):

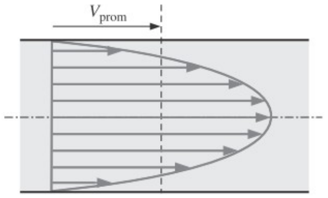

FIGURA 5-4
La velocidad promedio V_{prom} se define como la magnitud de la velocidad promediada a lo largo de la sección transversal.

Velocidad promedio:

$$V_{\text{prom}} = \frac{1}{A_c}\int_{A_c} V_n \, dA_c \tag{5-6}$$

donde A_c es el área de la sección transversal normal a la dirección del flujo. Note que si la magnitud de la velocidad fuera V_{prom} en toda la sección transversal, el gasto de masa sería idéntico al que se obtiene cuando se integra el perfil real de velocidad. De donde, para el flujo incompresible o inclusive para el flujo compresible para el cual ρ sea uniforme a lo largo de A_c, la ecuación 5-5 queda:

$$\dot{m} = \rho V_{\text{prom}} A_c \qquad \text{(kg/s)} \tag{5-7}$$

Para el flujo compresible se puede concebir ρ como la densidad promedio sobre la sección transversal y entonces, no obstante, se usa la ecuación 5-7 como una aproximación razonable. Por sencillez, se elimina el subíndice de la velocidad promedio. A menos que se indique lo contrario, V denota la velocidad promedio en la dirección del flujo. Asimismo, A_c denota el área de la sección transversal normal a la dirección del flujo.

El volumen del fluido que fluye a través de una sección transversal por unidad de tiempo se llama **razón de flujo volumétrico**, o gasto volumétrico, o simplemente flujo volumétrico, $\dot{V}$ (Fig. 5-5), y se da por:

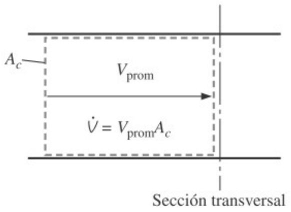

FIGURA 5-5
El gasto volumétrico es el volumen de fluido que fluye a través de una sección transversal por unidad de tiempo.

$$\dot{V} = \int_{A_c} V_n \, dA_c = V_{\text{prom}} A_c = V A_c \qquad (\text{m}^3/\text{s}) \tag{5-8}$$

En 1628, el monje italiano Benedetto Castelli (1577-1644) publicó una primera forma de la ecuación 5-8. Note que en muchos textos de mecánica de fluidos se usa Q en lugar de $\dot{V}$ para el gasto volumétrico. Aquí se usa $\dot{V}$ para evitar confusión con la transferencia de calor.

Las razones de flujo de masa y de volumen están relacionadas por:

$$\dot{m} = \rho \dot{V} = \frac{\dot{V}}{v} \tag{5-9}$$

donde v es el volumen específico. Esta relación es análoga a $m = m = \rho V = V/v$, la cual es la relación entre la masa y el volumen de un fluido en un recipiente.

Principio de conservación de la masa

Este **principio para un volumen de control** se puede expresar como: *la transferencia neta de masa hacia dentro un volumen de control, o hacia fuera de éste, durante un intervalo* Δt*, es igual al cambio neto (aumento o disminución) en la masa total que está dentro de ese volumen en el transcurso de* Δt; es decir:

$$\begin{pmatrix}\text{Masa total que entra}\\ \text{al VC durante } \Delta t\end{pmatrix} - \begin{pmatrix}\text{Masa total que sale}\\ \text{del VC durante } \Delta t\end{pmatrix} = \begin{pmatrix}\text{Cambio neto durante } \Delta t \text{ en la}\\ \text{masa que está dentro del VC}\end{pmatrix}$$

o

$$m_{\text{ent}} - m_{\text{sal}} = \Delta m_{\text{VC}} \qquad \text{(kg)} \tag{5-10}$$

donde $\Delta m_{\text{VC}} = m_{\text{final}} - m_{\text{inicial}}$ es el cambio en la masa del volumen de control durante el proceso (Fig. 5-6). Esto también se puede expresar en la *forma de razones*:

$$\dot{m}_{\text{ent}} - \dot{m}_{\text{sal}} = dm_{\text{VC}}/dt \qquad \text{(kg/s)} \tag{5-11}$$

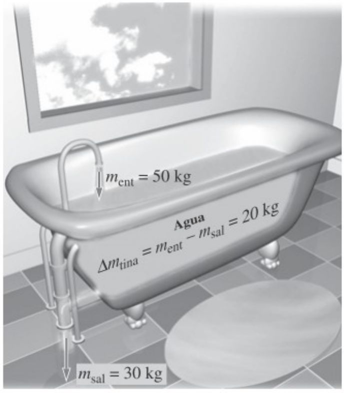
$m_{\text{ent}} = 50$ kg
Agua
$\Delta m_{\text{tina}} = m_{\text{ent}} - m_{\text{sal}} = 20$ kg
$m_{\text{sal}} = 30$ kg

FIGURA 5-6
Principio de conservación de la masa para una tina común de baño.

donde $\dot{m}_{\text{ent}}$ y $\dot{m}_{\text{sal}}$ son las razones totales de flujo de masa hacia dentro y hacia fuera del volumen de control, y dm_{CV}/dt es la razón de cambio de la masa que está dentro de las fronteras de ese volumen. Con frecuencia, se hace mención de las ecuaciones 5-10 y 5-11 como el **balance de masa** y son aplicables a cualquier volumen de control que pase por alguna clase de proceso.

Considere un volumen de control de forma arbitraria, como se muestra en la figura 5-7. La masa de un volumen diferencial dV que esté dentro del volumen de control es $dm = \rho\, dV$. Por integración se determina que la masa total dentro del volumen de control en cualquier instante t es:

Masa total dentro del VC:

$$m_{\text{VC}} = \int_{\text{VC}} \rho\, dV \tag{5-12}$$

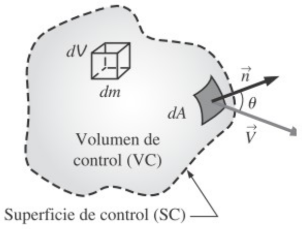

FIGURA 5-7
Volumen diferencial de control, dV, y superficie diferencial de control, dA, usados en la deducción de la relación de conservación de la masa.

Entonces la razón de cambio de la cantidad de masa dentro del volumen de control se puede expresar de la siguiente maneraa:

Razón de cambio de la masa dentro del VC:

$$\frac{dm_{\text{VC}}}{dt} = \frac{d}{dt}\int_{\text{VC}} \rho\, dV \tag{5-13}$$

Para el caso especial en el que nada de masa cruza la superficie de control (es decir, el volumen de control es un sistema cerrado), el principio de conservación de la masa se reduce al de un sistema que se puede expresar como $dm_{\text{VC}}/dt = 0$. Esta relación es válida si el volumen de control está fijo, en movimiento o deformándose.

Considérese ahora el flujo de masa hacia fuera o hacia dentro del volumen de control a través de un área diferencial dA sobre la superficie de control de un volumen fijo. Sea $\vec{n}$ el vector unitario, normal a la suprficie dA y dirigido hacia fuera de volumen de control, normal a ésta y $\vec{V}$ la velocidad del flujo dA en relación con un sistema fijo de coordenadas, como se muestra en la figura 5-7. En general, la velocidad cruza la superficie dA y forma un ángulo θ con la normal exterior de ésta, y la razón de flujo másico es proporcional a la componente normal de la velocidad $\vec{V}_n = \vec{V}\cos\theta$ cambiándose desde un valor de flujo hacia fuera cuando $\theta = 0$ (el flujo es normal a dA), pasando por cero, cuando $\theta = 90°$ (el flujo es tangente a dA), hasta llegar a un *flujo* máximo *hacia dentro* cuando $\theta = 180°$ (el flujo es normal a dA, pero en dirección opuesta). Al aplicar el concepto del producto punto de dos vectores, la magnitud de la componente normal de la velocidad se puede expresar de la siguiente manera:

Componente normal de la velocidad: $$V_n = V\cos\theta = \vec{V}\cdot\vec{n} \qquad \textbf{(5-14)}$$

El flujo másico a través de dA es proporcional a la densidad del fluido ρ, a la velocidad normal V_n y al área de flujo dA, y se puede expresar como:

Razón diferencial de flujo de masa:

$$\delta\dot{m} = \rho V_n\,dA = \rho(V\cos\theta)\,dA = \rho(\vec{V}\cdot\vec{n})\,dA \qquad \textbf{(5-15)}$$

El flujo másico neto hacia adentro o hacia afuera del volumen de control a través de toda la superficie de control se obtiene cuando se integra $\delta\dot{m}$ sobre esa superficie:

La razón neta de flujo de masa: $$\dot{m}_{\text{neto}} = \int_{\text{SC}} \delta\dot{m} = \int_{\text{CS}} \rho V_n\,dA = \int_{\text{SC}} \rho(\vec{V}\cdot\vec{n})\,dA \qquad \textbf{(5-16)}$$

Note que $\vec{V}\cdot\vec{n} = V\cos\theta$ es positiva para $\theta < 90°$ (flujo hacia fuera) y negativa para $\theta > 90°$ (flujo hacia dentro). Por lo tanto, se toma en cuenta de manera automática la dirección del flujo y la integral de la ecuación 5-16 da directamente la razón *neta* de flujo de masa. Un valor positivo para $\dot{m}_{\text{neto}}$ indica flujo neto hacia fuera y uno negativo indica flujo de masa neto hacia dentro.

Si se reordena la ecuación 5-11 como $dm_{\text{CV}}/dt + \dot{m}_{\text{sal}} - \dot{m}_{\text{ent}} = 0$, entonces se puede expresar la relación de conservación de la masa para un volumen fijo de control como:

Conservación de la masa, caso general: $$\frac{d}{dt}\int_{\text{VC}} \rho\,dV + \int_{\text{SC}} \rho(\vec{V}\cdot\vec{n})\,dA = 0 \qquad \textbf{(5-17)}$$

Ésta expresa que *la razón de cambio respecto al tiempo de la masa que está dentro del volumen de control más la razón neta de flujo de masa a través de la superficie de control es igual a cero.*

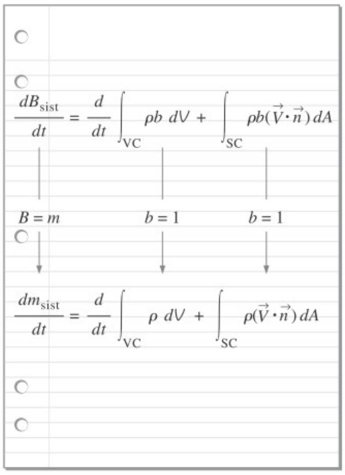
$\frac{dB_{\text{sist}}}{dt} = \frac{d}{dt}\int_{\text{VC}} \rho b\,dV + \int_{\text{SC}} \rho b(\vec{V}\cdot\vec{n})\,dA$
$B = m$
$b = 1$
$b = 1$
$\frac{dm_{\text{sist}}}{dt} = \frac{d}{dt}\int_{\text{VC}} \rho\,dV + \int_{\text{SC}} \rho(\vec{V}\cdot\vec{n})\,dA$

FIGURA 5-8
La ecuación de conservación de masa se obtiene cuando se reemplaza B en el teorema del transporte de Reynolds por la masa m y b por 1 (m por unidad de masa $= m/m = 1$).

También se puede deducir la relación general de conservación de la masa para un volumen de control con la aplicación del teorema del transporte de Reynolds (RTT, por sus siglas en inglés) cuando se toma la propiedad B como la masa m (capítulo 4). Entonces se tiene $b = 1$, ya que cuando se divide la masa, entre la masa, para obtener la propiedad por unidad de masa, da la unidad. Asimismo, la masa de un sistema es constante y su derivada respecto al tiempo es cero. Es decir $dm_{\text{sist}}/dt = 0$. Entonces el teorema del transporte de Reynolds se reduce de inmediato a la ecuación 5-17, como se muestra en la figura 5-8, y de este modo se ilustra que el teorema del transporte de Reynolds en verdad es un recurso muy poderoso.

Cuando se divide la integral de superficie de la ecuación 5-17 en dos partes —una para las corrientes salientes (término positivo) y otra para las entrantes (término negativo)— la relación general de conservación de la masa también se puede expresar como:

$$\frac{d}{dt}\int_{\text{CV}} \rho\,dV + \sum_{\text{sal}} \rho|V_n|A - \sum_{\text{ent}} \rho|V_n|A = 0 \qquad \textbf{(5-18)}$$

donde A representa el área para una entrada o una salida, y se usa el signo de sumatoria para subrayar que deben considerarse *todas* las entradas y salidas. Si se usa la definición de flujo másico, la ecuación 5-18 también se puede expresar como:

$$\frac{d}{dt}\int_{\text{VC}} \rho\,dV = \sum_{\text{ent}} \dot{m} - \sum_{\text{sal}} \dot{m} \quad \text{o} \quad \frac{dm_{\text{VC}}}{dt} = \sum_{\text{ent}} \dot{m} - \sum_{\text{sal}} \dot{m} \qquad \textbf{(5-19)}$$

Se tiene una flexibilidad considerable en la selección de un volumen de control cuando se resuelve un problema. Varias elecciones de este volumen son disponibles, pero algunas son más convenientes que otras. La selección de un volumen de control no debe llevar a las complicaciones innecesarias. La elección apropiada de un volumen de control puede hacer que la resolución de un problema aparentemente complicado sea más bien fácil. Una regla sencilla al elegir un volumen de control es hacer que la superficie de control sea *normal al flujo* en todos los lugares donde se se crucen la superficie y el flujo, siempre que sea posible. De esta manera, el producto punto $\vec{V} \cdot \vec{n}$ se convierte, simplemente, en la magnitud de la velocidad, y la integral $\int_A \rho(\vec{V} \cdot \vec{n})\, dA$ se vuelve ρVA (Fig. 5-9).

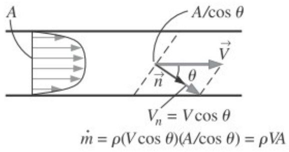

a) Superficie de control *formando un ángulo* con el flujo

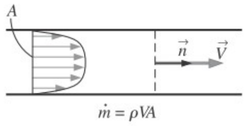

b) Superficie de control *normal* al flujo

FIGURA 5-9
Siempre debe seleccionarse una superficie de control *normal al flujo* en todos los lugares donde se cruce con ese flujo del fluido, para evitar complicaciones, aun cuando el resultado sea el mismo.

Volúmenes de control en movimiento o en deformación

Las ecuaciones 5-17 y 5-19 también son válidas para volúmenes de control en movimiento y en deformación, siempre que se reemplace la *velocidad absoluta* $\vec{V}$ por la *velocidad relativa* $\vec{V}_r$, la cual es la velocidad del fluido con relación a la superficie de control (capítulo 4). En el caso de un volumen de control que está moviéndose pero no esté deformando, la velocidad relativa es la velocidad del fluido que observa una persona en movimiento con el volumen de control y se expresa como $\vec{V}_r = \vec{V} - \vec{V}_{\text{SC}}$, en donde $\vec{V}$ es la velocidad del fluido y $\vec{V}_{\text{SC}}$ es la velocidad de la superficie de control, ambas en relación con un punto fijo en el exterior. Note que ésta es una sustracción *vectorial.*

En algunos problemas prácticos (como la inyección de un medicamento a través de la aguja de una jeringa por el movimiento forzado del émbolo) intervienen volúmenes de control en deformación. Todavía se pueden usar las relaciones de conservación de la masa desarrolladas para esos volúmenes siempre que la velocidad del fluido que cruza una parte en deformación de la superficie de control se exprese en relación con ésta (es decir, la velocidad del fluido se debe expresar en relación con un marco de referencia sujeto a la parte en deformación de la superficie de control). En este caso, una vez más, la velocidad relativa en cualquier punto sobre la superficie de control se expresa como $\vec{V}_r = \vec{V} - \vec{V}_{\text{SC}}$, en donde $\vec{V}_{\text{SC}}$ es la velocidad local de esa superficie de control en ese punto en relación con un punto fijo en el exterior del volumen de control.

Balance de masa para procesos de flujo estacionario

En el transcurso de un proceso de flujo estacionario, la cantidad total de masa contenida dentro de un volumen de control no cambia con el tiempo (m_{VC} = constante). Entonces el principio de conservación de la masa exige que la cantidad total de masa que entra en un volumen de control sea igual a la cantidad total de masa que sale de él. Por ejemplo, para la boquilla de una manguera de jardín en operación estacionaria, la cantidad de agua que entra a la boquilla por unidad de tiempo es igual a la cantidad de agua que sale de ella por unidad de tiempo.

Cuando se trata con procesos de flujo estacionario no se tiene interés en la cantidad de masa que fluye hacia fuera o hacia dentro de un dispositivo en un transcurso de tiempo; en lugar de ello, se tiene interés en la cantidad de masa que fluye por unidad de tiempo; es decir, *la razón de flujo de masa* $\dot{m}$. *El principio de conservación de la masa* para un sistema general de flujo estacionario con entradas y salidas múltiples se puede expresar en la forma de razón como (Fig. 5-10):

Flujo estacionario:
$$\sum_{\text{ent}} \dot{m} = \sum_{\text{sal}} \dot{m} \qquad \text{(kg/s)} \tag{5-20}$$

Ésta expresa que *la razón total de masa que entra en un volumen de control es igual a la razón total de masa que sale de él.*

Numerosos dispositivos de ingeniería, como toberas, difusores, turbinas, compresores y bombas, forman una sola corriente (sólo una entrada y una salida).

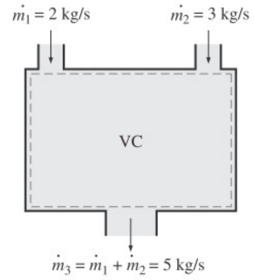

FIGURA 5-10
Principio de conservación de la masa para un sistema de flujo estacionario con dos entradas y una salida.

Para estos casos se denota el estado de entrada por el subíndice 1 y el de salida por el subíndice 2, y se eliminan los signos de suma. Entonces la ecuación 5-20 se reduce, para *sistemas de flujo estacionario con una sola corriente*, a:

$$\text{Flujo estacionario (una sola corriente):} \quad \dot{m}_1 = \dot{m}_2 \quad \rightarrow \quad \rho_1 V_1 A_1 = \rho_2 V_2 A_2 \qquad \textbf{(5-21)}$$

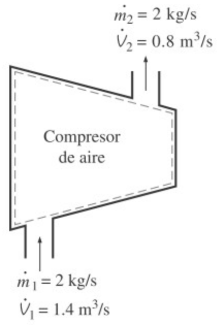

FIGURA 5-11
Cuando transcurre un proceso de flujo estacionario, los gastos volumétricos no se conservan necesariamente, aun cuando sí se conserven los flujos de masa.

Caso especial: flujo incompresible

Las relaciones de conservación de la masa se pueden simplificar todavía más cuando el fluido es incompresible, el cual suele ser el caso para los líquidos. Cuando se cancela la densidad en ambos miembros de la relación general del flujo estacionario da:

$$\text{Flujo estacionario e incompresible:} \quad \sum_{\text{ent}} \dot{V} = \sum_{\text{sal}} \dot{V} \qquad (\text{m}^3/\text{s}) \qquad \textbf{(5-22)}$$

Para sistemas de una sola corriente de flujo uniforme, la ecuación 5-22 se convierte en

$$\text{Flujo estacionario e incompresible (una sola corriente):} \quad \dot{V}_1 = \dot{V}_2 \rightarrow V_1 A_1 = V_2 A_2 \qquad \textbf{(5-23)}$$

Siempre se debe tener presente que no existen cosas como principio de "conservación del volumen". Por lo tanto, los gastos volumétricos hacia dentro y hacia fuera de un aparato pueden ser diferentes. El gasto volumétrico a la salida de un compresor de aire es mucho menor que el que se tiene en la admisión, aun cuando la razón de flujo de masa de aire a través del compresor es constante (Fig. 5-11). Esto se debe a la densidad más alta del aire a la salida del compresor. Sin embargo, para el flujo estacionario de líquidos, los gastos volumétricos, así como los de masa, permanecen constantes, ya que los líquidos son esencialmente sustancias incompresibles (de densidad constante). El flujo de agua por la boquilla de una manguera de jardín es un ejemplo de este último caso.

El principio de conservación de la masa exige tomar en cuenta toda pequeña porción de masa que participe en un proceso. Si el lector puede verificar el saldo de su chequera (mantener al día los depósitos y disposiciones o, sencillamente, si observa el principio de "conservación del dinero"), no debe tener dificultades en aplicar el principio de conservación de la masa a los sistemas de ingeniería.

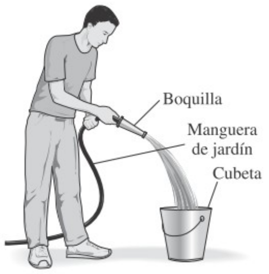

FIGURA 5-12
Esquema para el ejemplo 5-1.

EJEMPLO 5-1 Flujo de agua por la boquilla de una manguera de jardín

Se usa una manguera de jardín que tiene una boquilla de riego para llenar una cubeta de 10 gal. El diámetro de la manguera es de 2 cm y se reduce hasta 0.8 cm en la salida de la boquilla (Fig. 5-12). Si transcurren 50 s para llenar la cubeta con agua, determine *a*) las razones de flujo volumétrico y de masa del agua que pasa por la manguera y *b*) la velocidad promedio del agua a la salida de la boquilla.

SOLUCIÓN Se usa una manguera de jardín para llenar una cubeta con agua. Se deben determinar las razones de flujo volumétrico y de masa y la velocidad a la salida.

Suposiciones **1** El agua es una sustancia incompresible. **2** El flujo por la manguera es estacionario. **3** No hay desperdicio de agua por salpicadura.

Propiedades Se toma la densidad del agua como 1 000 kg/m³ = 1 kg/L.

Análisis *a*) Note que se descargan 10 gal de agua en 50 s, las razones de flujo volumétrico y de masa son:

$$\dot{V} = \frac{V}{\Delta t} = \frac{10 \text{ gal}}{50 \text{ s}} \left(\frac{3.7854 \text{ L}}{1 \text{ gal}} \right) = \mathbf{0.757\ L/s}$$

$$\dot{m} = \rho \dot{V} = (1 \text{ kg/L})(0.757 \text{ L/s}) = \mathbf{0.757\ kg/s}$$

+,- * *. * 9 + / D + +"B "
/ G + D A 2 H? 4 / / E 6 + D E " 0 A
/ + D 6 E 2 H I /"

.A 1 B " CCC

" *+,-* * *. * 6 DE < + % "
9 6 " " &E / % A /"

6 ! $

4 / " / /
(!&

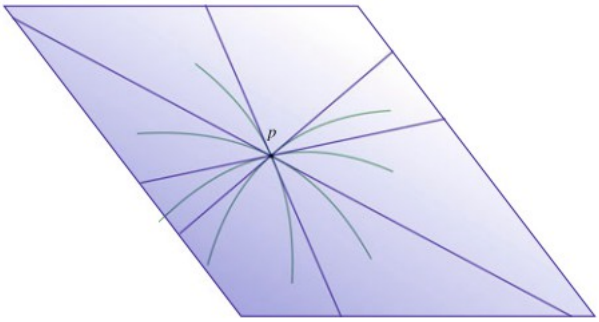

6 ! 4 / " / /
/ B /

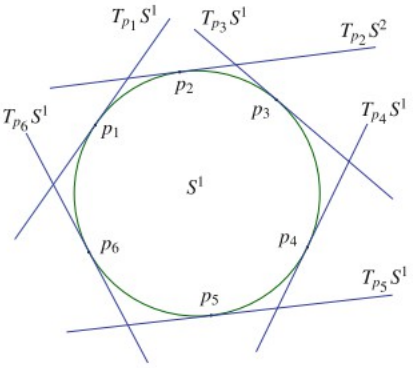
$T_{p_1}S^1$
$T_{p_3}S^1$
$T_{p_2}S^2$
$T_{p_6}S^1$
$T_{p_4}S^1$
p_2
p_3
p_1
S^1
p_6
p_4
$T_{p_5}S^1$
p_5

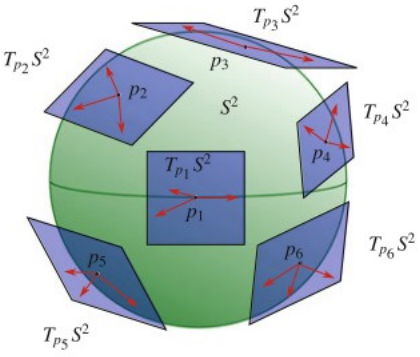
$T_{p_3}S^2$
$T_{p_2}S^2$
p_3
p_2
S^2
$T_{p_4}S^2$
p_4
$T_{p_1}S^2$
p_1
p_5
p_6
$T_{p_6}S^2$
$T_{p_5}S^2$

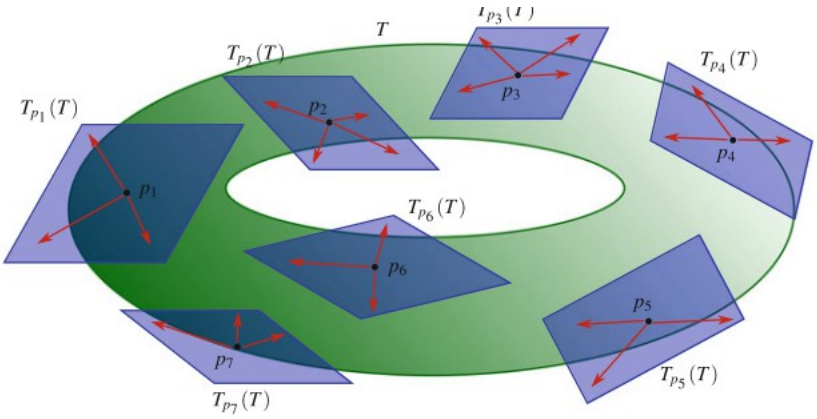
$T_{p_3}(T)$
T
$T_{p_2}(T)$
$T_{p_4}(T)$
$T_{p_1}(T)$
p_3
p_2
p_4
p_1
$T_{p_6}(T)$
p_6
p_5
p_7
$T_{p_7}(T)$
$T_{p_5}(T)$

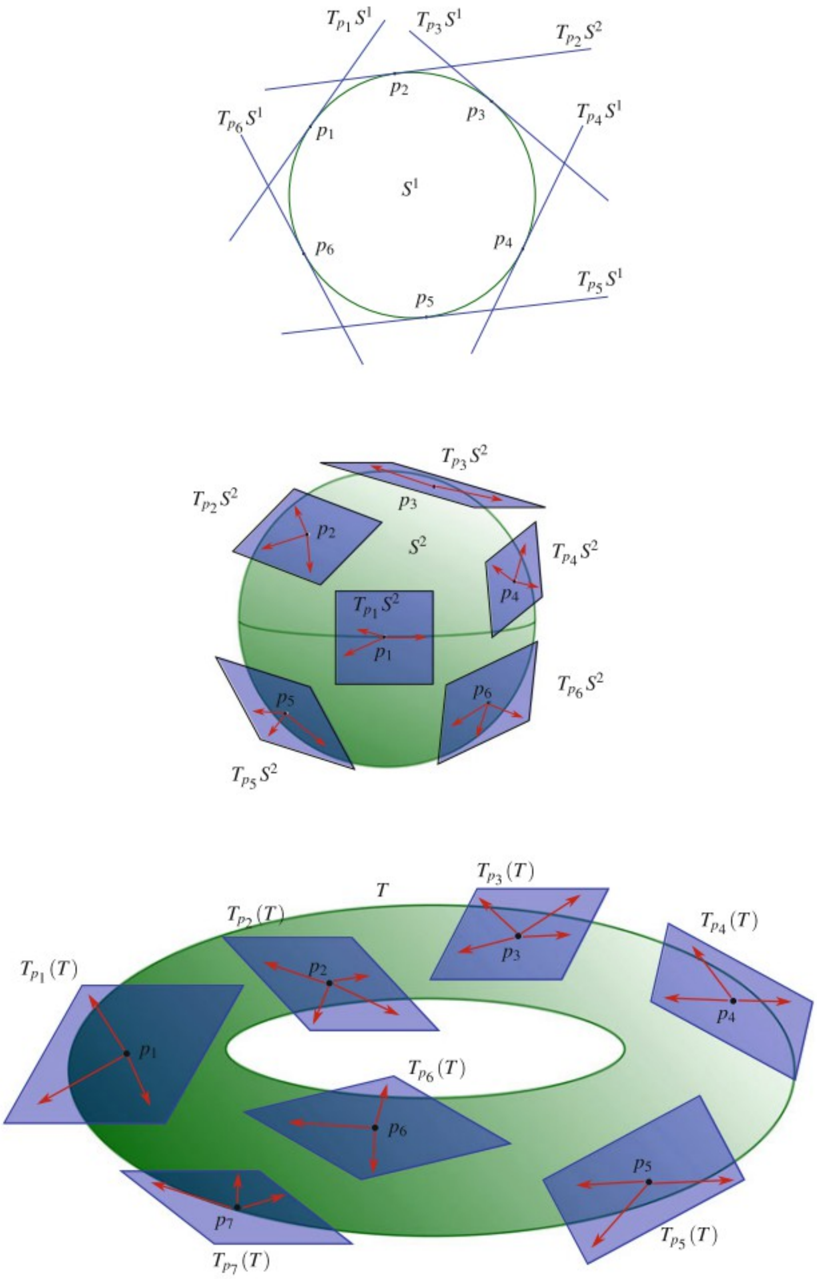
$T_{p_1}S^1$
$T_{p_3}S^1$
$T_{p_2}S^2$
p_2
$T_{p_6}S^1$
p_1
p_3
$T_{p_4}S^1$
S^1
p_6
p_4
$T_{p_5}S^1$
p_5
$T_{p_3}S^2$
$T_{p_2}S^2$
p_3
p_2
S^2
$T_{p_4}S^2$
p_4
$T_{p_1}S^2$
p_1
$T_{p_6}S^2$
p_5
p_6
$T_{p_5}S^2$
$T_{p_3}(T)$
T
$T_{p_2}(T)$
$T_{p_4}(T)$
$T_{p_1}(T)$
p_3
p_2
p_4
p_1
$T_{p_6}(T)$
p_6
p_5
p_7
$T_{p_5}(T)$
$T_{p_7}(T)$

Fig. 2.14 The manifold S^1 along with a number of tangent lines depicted (top left). The manifold S^2 along with a number of tangent planes depicted (top right). The manifold T along with a number of tangent planes depicted (bottom). For both manifolds S^2 and T several vectors in each tangent plane are also depicted

(! $

Chapter 2
An Introduction to Differential Forms

In this chapter we introduce one of the fundamental ideas of this book, the differential one-form. Later chapters will discuss differential k-forms for $k > 1$. We slowly and systematically build up to the concept of a one-forms by first considering a variety of necessary ingredients. We begin in section one by introducing the Cartesian coordinate functions, which play a surprisingly important role in understanding exactly what a one-form is, as well as playing a role in the notation used for one-forms.

In section two we discuss manifolds, tangent spaces, and vector fields in enough detail to give you a good intuitive idea of what they are, leaving a more mathematically rigorous treatment for Chap. 10. In section three we return to the concept of directional derivatives that you are familiar with from vector calculus. The way directional derivatives are usually defined in vector calculus needs to be tweaked just a little for our purposes. Once this is done a rather surprising identification between the standard Euclidian unit vectors and differential operators can be made. This equivalence plays a central role going forward.

Finally in section four differential one-forms are defined. Thoroughly understanding differential one-forms requires pulling together a variety of concepts; vector spaces and dual spaces, the equivalence between Euclidian vectors and differential operators, directional derivatives and the Cartesian coordinate functions. The convergence of all of these mathematical ideas in the definition of the one-form is the goal of the chapter.

2.1 Coordinate Functions

Now that we have reviewed the necessary background we are ready to start getting into the real meat of the course, differential forms. For the first part of this course, until we understand the basics, we will deal with differential forms on $\mathbb{R}^n$, with $\mathbb{R}^2$ and $\mathbb{R}^3$ being our primary examples. Only after we have developed some intuitive feeling for differential forms will we address differential forms on more general manifolds. You can very generally think of a manifold as a space which is locally Euclidian - that means that if you look closely enough at one small part of a manifold then it basically looks like $\mathbb{R}^n$ for some n.

For the time being we will just work with our familiar old Cartesian coordinate system. We will discuss other coordinate systems later on. We are used to seeing $\mathbb{R}^2$ pictured as a plane with an x and a y axis as in Fig. 2.1. Here we have plotted the points $(2, 1)$, $(-4, 3)$, $(-2, -2)$, and $(1, -4)$ on the xy-plane. Similarly, we are used to seeing $\mathbb{R}^3$ pictured with an x, y, and z axis as show in Fig. 2.2. Here we have plotted the points $(2, 3, 4)$, $(-3, -4, 2)$, and $(4, 1, -3)$. Also, notice that we follow the "right-hand rule" convention in this book when drawing $\mathbb{R}^3$.

Notice the difference in the notations between points in $\mathbb{R}^n$ and vectors in $\mathbb{R}^n$. For example, points in $\mathbb{R}^2$ or $\mathbb{R}^3$ are denoted by $p = (x, y)$ or $p = (x, y, z)$, where $x, y, z \in \mathbb{R}$, whereas vectors in $\mathbb{R}^2$ or $\mathbb{R}^3$ are denoted by

$$v = \begin{bmatrix} x \\ y \end{bmatrix} \text{ or } v = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

© Springer Nature Switzerland AG 2018

J. P. Fortney, *A Visual Introduction to Differential Forms and Calculus on Manifolds*,
https://doi.org/10.1007/978-3-319-96992-3_2

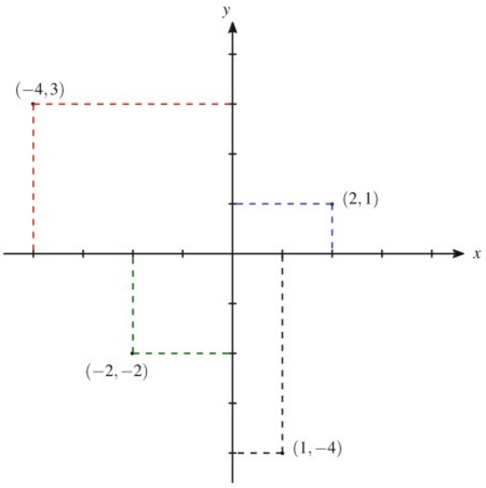
y
x
(−4, 3)
(2, 1)
(−2, −2)
(1, −4)

Fig. 2.1 The Cartesian coordinate system for $\mathbb{R}^2$ with the points (2, 1), (−4, 3), (−2, −2), and (1, −4) shown

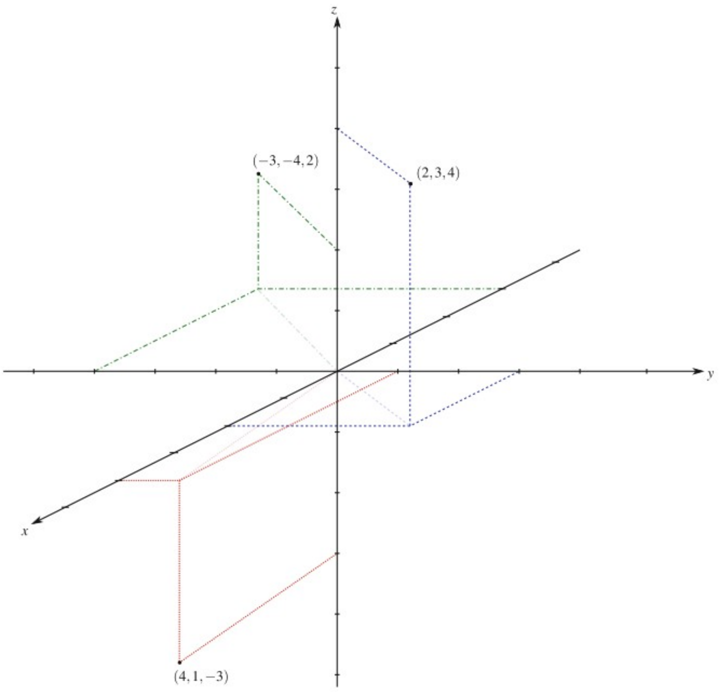
z
y
x
(−3, −4, 2)
(2, 3, 4)
(4, 1, −3)

Fig. 2.2 The Cartesian coordinate system for $\mathbb{R}^3$ with the points (2, 3, 4), (−3, −4, 2), and (4, 1, −3) shown

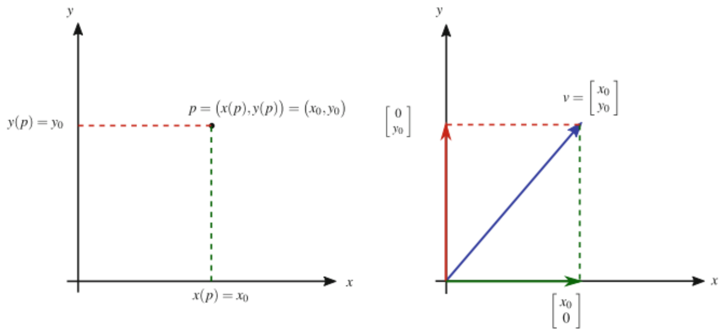
y
$y(p) = y_0$
$p = (x(p), y(p)) = (x_0, y_0)$
x
$x(p) = x_0$
y
$v = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}$
$\begin{bmatrix} 0 \\ y_0 \end{bmatrix}$
x
$\begin{bmatrix} x_0 \\ 0 \end{bmatrix}$

Fig. 2.3 The manifold $\mathbb{R}^2$ (left) and the vector space $\mathbb{R}^2$ (right)

where $x, y, z \in \mathbb{R}$. Points and vectors in $\mathbb{R}^n$ for arbitrary n are denoted similarly. Here comes a subtle distinction that is almost always glossed over in calculus classes, especially in multivariable or vector calculus, yet it is a distinction that is very important.

- We will call the collection of points $p = (x_1, x_2, \dots, x_n)$, with $x_1, x_2, \dots, x_n \in \mathbb{R}$, the *manifold* $\mathbb{R}^n$.
- We will call the collection of vectors $v = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$, where $x_1, x_2, \dots, x_n \in \mathbb{R}$, the *vector space* $\mathbb{R}^n$.

One of the reasons this distinction is glossed over in calculus classes is that most calculus students have not yet taken a course in linear algebra and do not yet know what a vector space is. Also, it is certainly possible to "do" vector calculus without knowing the technical definition of a vector space by being a little vague and imprecise mathematically speaking. The manifold $\mathbb{R}^2$ and the vector space $\mathbb{R}^2$ look exactly alike yet are totally different spaces. They have been pictured in Fig. 2.3. The manifold $\mathbb{R}^2$ contains points and the vector space $\mathbb{R}^2$ contains vectors. Of course, in the case of $\mathbb{R}^2$, $\mathbb{R}^3$, or $\mathbb{R}^n$ for that matter, these spaces can be naturally identified. That is, they are really the same. So why is making the distinction so important? It is important to make the distinction for a couple of reasons. First, keeping clear track of our spaces will help us understand the theory of differential forms better, and second, when we eventually get to more general manifolds, most manifolds will not have a vector space structure so we do not want to get into the bad habit of thinking of our manifolds as vector spaces too.

Now let us spend a moment reviewing good old-fashioned functions on $\mathbb{R}^2$ and $\mathbb{R}^3$. We will consider real-valued functions, which simply means that the range (or codomain) is the reals $\mathbb{R}$ or some subset of the reals. Real-valued functions from $\mathbb{R}^2$ are the easiest to visualize since they can be graphed in three dimensional space. Consider the real-valued functions $f : \mathbb{R}^2 \to \mathbb{R}$ defined by the following $f(x, y) = x^3 + y^2 - 3xy$. For example, $f(2, 3)$ is given by $(2)^3 + (3)^2 - 3(2)(3) = -1$, $f(5, -2)$ is given by $(5)^3 + (-2)^2 - 3(5)(-2) = 159$ and $f(-3, -1)$ is given by $(-3)^3 + (-1)^2 - 3(-3)(-1) = -35$. A function from $\mathbb{R}^2$ to $\mathbb{R}$ is graphed in $\mathbb{R}^3$ by the set of points $(x, y, f(x, y))$. The function $f(x, y) = x^3 + y^2 - 3xy$ is shown in Fig. 2.4. Sometimes we will write $f(p)$ where $p = (x, y)$. By doing this we are emphasizing that the input (x, y) is a point in the domain.

As a further illustration, the graphs of the functions $f(x, y) = xy$ and $f(x, y) = x^2 + y^2$ are shown in Fig. 2.5. As before, the set of points that are actually graphed are the points $(x, y, f(x, y))$ in $\mathbb{R}^3$, thus we can see that it requires three dimensions to draw the graph of a function $f : \mathbb{R}^2 \to \mathbb{R}$. That means that we can not accurately draw the graph of any function $f : \mathbb{R}^n \to \mathbb{R}$ for $n \geq 3$. However, we often continue to draw inaccurate cartoons similar to Figs. 2.4 and 2.5 when we want to try to visualize or depict functions $f : \mathbb{R}^n \to \mathbb{R}$ where $n \geq 3$.

We will use different name for "functions" on the manifold $\mathbb{R}^n$ and "functions" on the vector space $\mathbb{R}^n$. Functions on the manifold $\mathbb{R}^n$ will simply be called **functions**, or sometimes, if we want to emphasize that range is the set of real numbers, **real-valued functions**. However, our "functions" on the vector space $\mathbb{R}^n$ are called **functionals** when the range is the set of real values and **transformations** when the range is another vector space. We have already spent some time considering a

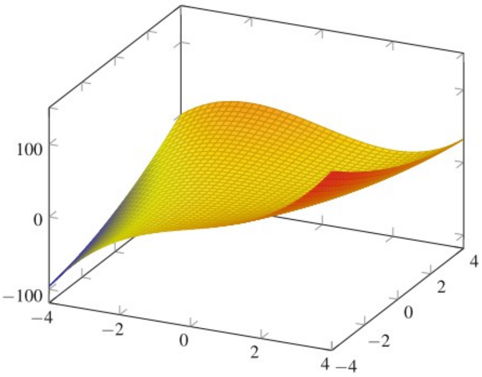

Fig. 2.4 The graph of function $f(x, y) = x^3 + y^2 - 3xy$ is displayed

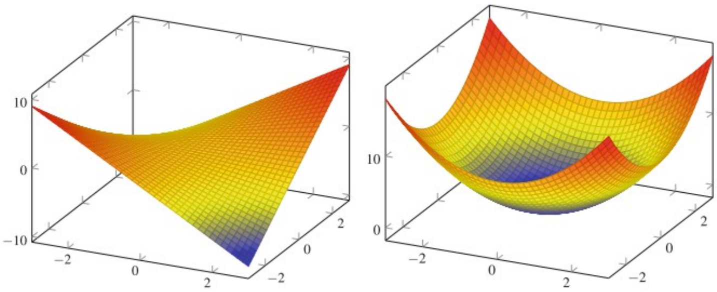

Fig. 2.5 The graphs of the function $f(x, y) = xy$ (left) and $f(x, y) = x^2 + y^2$ (right) as an illustration of the graphs of functions in two variables depicted in $\mathbb{R}^3$

certain class of functionals, the linear functionals that satisfy

$$\begin{aligned} f(v + w) &= f(v) + f(w) \\ f(cv) &= cf(v) \end{aligned}$$

where v and w are vectors and c is a real number. So, the difference between manifolds $\mathbb{R}^n$ and vector spaces $\mathbb{R}^n$ is implicit in our language:

$$\begin{aligned} \text{function } f &: \text{(manifold) } \mathbb{R}^n \longrightarrow \mathbb{R} \\ \text{functional } f &: \text{(vector space) } \mathbb{R}^n \longrightarrow \mathbb{R}. \end{aligned}$$

Now, we will consider a very special function from $\mathbb{R}^n$ to $\mathbb{R}$. We will start by using $\mathbb{R}^2$ as an example. The special functions we want to look at are called the **coordinate functions**. For the moment we will look at the **Cartesian coordinate functions**.

Consider a point p in the manifold $\mathbb{R}^2$ given by the Cartesian coordinates $(3, 5)$. Then the coordinate function

$$\begin{aligned} x : \mathbb{R}^2 &\longrightarrow \mathbb{R} \\ p &\longmapsto x(p) \end{aligned}$$

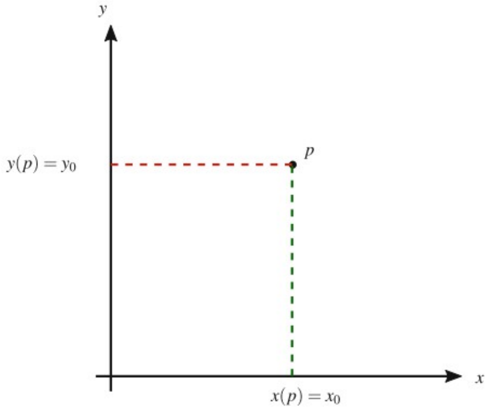

Fig. 2.6 The coordinate functions $x : \mathbb{R}^2 \to \mathbb{R}$ and $y : \mathbb{R}^2 \to \mathbb{R}$ picking off the coordinate values x_0 and y_0, respectively, of the point $p \in \mathbb{R}^2$

picks off the first coordinate, that is, $x(p) = x\big((3, 5)\big) = 3$, and the coordinate function

$$\begin{aligned} y : \mathbb{R}^2 &\longrightarrow \mathbb{R} \\ p &\longmapsto y(p) \end{aligned}$$

picks off the second coordinate, $y(p) = y\big((3, 5)\big) = 5$. This is illustrated in Fig. 2.6.

The fact of the matter is, even though you are used to thinking of (3, 5) as a point, a point is actually more abstract than that. A point exists independent of its coordinates. For example, you will sometimes see something like this,

$$p = \big(x(p), y(p)\big)$$

where, as an example, for the point we were given above we have $x(p) = 3$ and $y(p) = 5$.

There is some ambiguity that we need to get used to here. Often x and y are used to represent the outputs of the coordinate functions x and y. So, we could end up with an ambiguous statement like this

$$\begin{aligned} x(p) &= x \\ y(p) &= y \end{aligned}$$

where the x and the y on the left hand sides of the equations are coordinate functions from $\mathbb{R}^2$ to $\mathbb{R}$, and the x and y on the right hand side are real numbers, that is, elements of $\mathbb{R}$. Learning to recognize when x and y are coordinate functions and when they are real values requires a little practice, you have to look carefully at how they are being used. This distinction may sometimes be pointed out, but it often is glossed over.

Even though points are abstract and independent of coordinates, we can not actually talk about a particular point without giving it an "address." This is actually what coordinate functions do, they are used to give a point an "address."

Coordinate functions on $\mathbb{R}^3$ (and $\mathbb{R}^n$ for that matter) are completely analogous. For example, on $\mathbb{R}^3$ we have the three coordinate functions

$$\begin{aligned} x : \mathbb{R}^3 &\longrightarrow \mathbb{R} \\ p &\longmapsto x(p) \end{aligned} \qquad \begin{aligned} y : \mathbb{R}^3 &\longrightarrow \mathbb{R} \\ p &\longmapsto y(p) \end{aligned} \qquad \begin{aligned} z : \mathbb{R}^3 &\longrightarrow \mathbb{R} \\ p &\longmapsto z(p) \end{aligned}$$

which pick off the x, the y and the z coordinates of the point p as illustrated in Fig. 2.7.

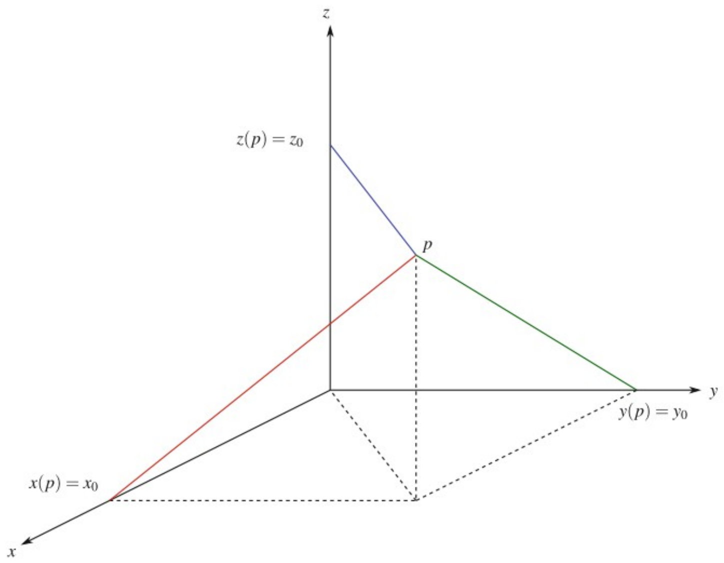

Fig. 2.7 The coordinate functions $x : \mathbb{R}^3 \to \mathbb{R}$, $y : \mathbb{R}^3 \to \mathbb{R}$, and $z : \mathbb{R}^3 \to \mathbb{R}$ picking off the coordinate values x_0, y_0, and z_0 respectively, of the point $p \in \mathbb{R}^3$

Of course Cartesian coordinate functions are not the only coordinate functions there are. Polar coordinate, spherical coordinate, and cylindrical coordinate functions are some other examples you have likely encountered before. We will talk about these coordinate systems in more detail later on but for the moment we consider one example. Consider the point $p = p\big(x(p), y(p)\big) = (3, 5)$. Polar coordinates r and θ are often given by $x = r\cos(\theta)$ and $y = r\sin(\theta)$. Suppose we want to find the "address" of point p in terms of polar coordinates. That is, we want to find $p = \big(r(p), \theta(p)\big)$.

Since we know p in terms of x and y coordinates, we can write $r = \sqrt{x^2 + y^2}$ and $\theta = \arctan\left(\frac{y}{x}\right)$. Then we have

$$\begin{aligned} p = \Big(r(p), \theta(p)\Big) &= \Big(r\Big(x(p), y(p)\Big), \theta\Big(x(p), y(p)\Big)\Big) \\ &= \left(\sqrt{x(p)^2 + y(p)^2}, \arctan\left(\frac{y(p)}{x(p)}\right)\right) \\ &= \left(\sqrt{3^2 + 5^2}, \arctan\left(\frac{5}{3}\right)\right) \\ &= \left(\underbrace{\sqrt{34}}_{r}, \underbrace{\arctan\left(\frac{5}{3}\right)}_{\theta}\right) \\ &\approx \big(5.831, 59.036\big). \end{aligned}$$

Again, the same ambiguity applies, r and θ can either be coordinate functions or the values given by the coordinate functions.

2.2 Tangent Spaces and Vector Fields

Let us recap what we have done so far. We have reviewed vector spaces and talked about $\mathbb{R}^n$ as a vector space, that is, as a collection of vectors of the from

$$v = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

where $x_1, \ldots, x_n \in \mathbb{R}$. Then we talked about $\mathbb{R}^n$ as a manifold, that is, as a collection of points $(x_1, \ldots, x_n)$ where $x_1, \ldots, x_n \in \mathbb{R}$. We have also reviewed linear functionals on the vector space $\mathbb{R}^n$ and functions on the manifold $\mathbb{R}^n$ and introduced coordinate functions. We will now see how the manifold $\mathbb{R}^n$ is related to vector spaces $\mathbb{R}^n$. Our motivation for this is our desire to do calculus. Well, for the moment to take directional derivatives anyway.

However, to help us picture what is going on a little better we are going to start by considering three manifolds, S^1, S^2, and T. The manifold S^1 is simply a circle, the manifold S^2 is the sphere, and the manifold T is the torus, which looks like a donut. For the moment we are going to be quite imprecise. We want you to walk away with a general feeling for what a manifold is and not overwhelm you with technical details. For the moment we will simply say a **manifold** is a space that is **locally Euclidian**. What we mean by that is if we look at a very small portion of the manifold that portion looks like $\mathbb{R}^n$ for some natural number n.

We will illustrate this idea for the three manifolds we have chosen to look at. The circle S^1 is locally like $\mathbb{R}^1$, which explains the superscript 1 on the S. In Fig. 2.8 we zoom in on a small portion of the one-sphere S^1 a couple of times to see that locally the manifold S^1 looks like $\mathbb{R}^1$. Similarly, in Fig. 2.9 we zoom in on a small portion of the two-sphere S^2 to see that it looks like $\mathbb{R}^2$ locally and in Fig. 2.10 we zoom in on a small portion of the torus T see that it also looks locally like $\mathbb{R}^2$.

The point with these examples is that even though these example manifolds *locally* look like $\mathbb{R}^1$ and $\mathbb{R}^2$, *globally* they do not. That is, their global behavior is more complex; these spaces somehow twist around and reconnect with themselves in a way that Euclidian space does not.

Now we will introduce the idea of a **tangent space**. Again, we want to give you an general feeling for what a tangent space is and not overwhelm you with the technical details, which will be presented in 10.2. From calculus you should recall what the tangent line to a curve at some point of the curve is. Each point of the curve has its own tangent line, as is shown in Fig. 2.11. We can see that the one dimensional curve has tangent spaces that are lines, that is, that are also one dimensional.

Similarly, a two dimensional surface has a tangent plane at each point of the surface, as is pictured in Fig. 2.12. The tangent space to a manifold at the point p is basically the set of all lines that are tangent to smooth curves that go through

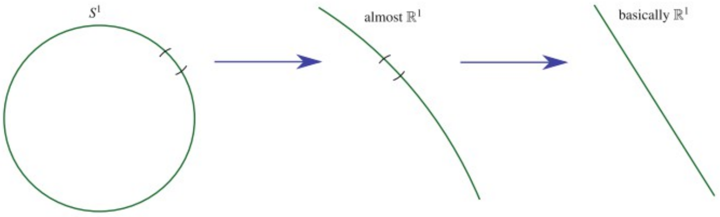

Fig. 2.8 Here a small portion of a one-sphere, or circle, S^1 is zoomed in on to show that locally the manifold S^1 looks like $\mathbb{R}^1$

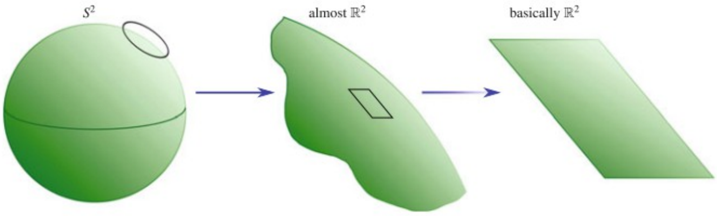

Fig. 2.9 Here a small portion of the two-sphere S^2 is zoomed in on to show that locally the manifold S^2 looks like $\mathbb{R}^2$

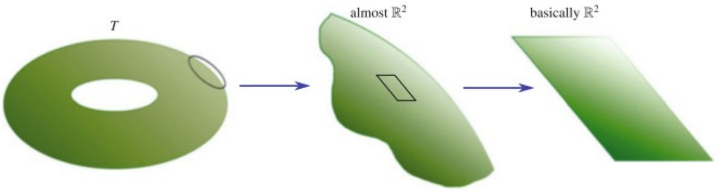

Fig. 2.10 Here a small portion of the torus T is zoomed in on to show that locally the manifold T looks like $\mathbb{R}^2$

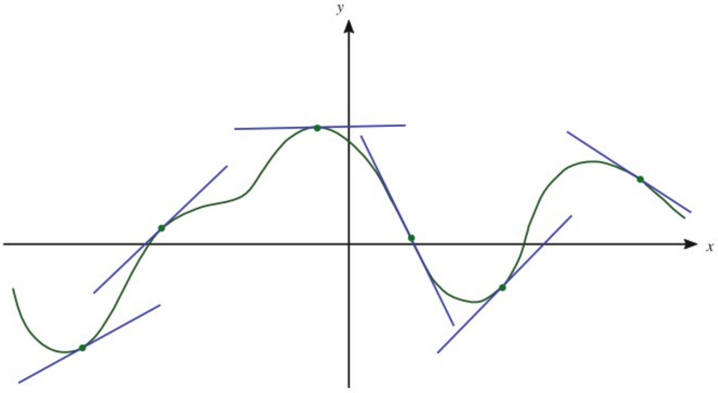

Fig. 2.11 Here a curve is pictured with tangent lines drown at several points

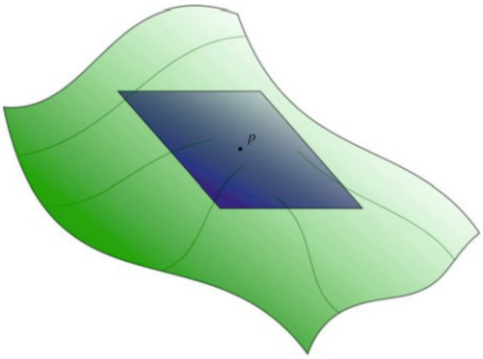

Fig. 2.12 Here a two-dimensional surface is pictured along with a tangent plane drown at one point

the manifold at the point p. This is illustrated in Fig. 2.13 where we have drawn a point p with five smooth curves going through it along with the five tangent lines to the curves at the point p. In both of these pictures the tangent space is a plane.

If you imagine a point in a three dimensional manifold and all the possible curves that can go through that point, and then imagine all the tangent lines to those curves at that point, you will see that the tangent space at that point is three dimensional. It turns out that the tangent space at a "nice" point is the same dimension as the manifold at that point. In this book we will not discuss manifolds with points that are not "nice" in this sense. In fact, the tangent spaces turn out to be nothing more than the vector spaces $\mathbb{R}^n$ for the correct value of n. Because the tangent spaces are vector spaces we think of the tangent space at a point p to be the set of all tangent vectors to the point p. We use special notation for the tangent space. Either T_pM or $T_p(M)$ denotes the tangent space to the manifold M at the point p.

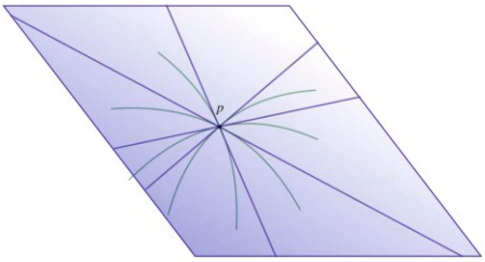

Fig. 2.13 The tangent plane to a two-dimensional surface at the point p is the set of all tangent lines at p to curves in that surface that go through p

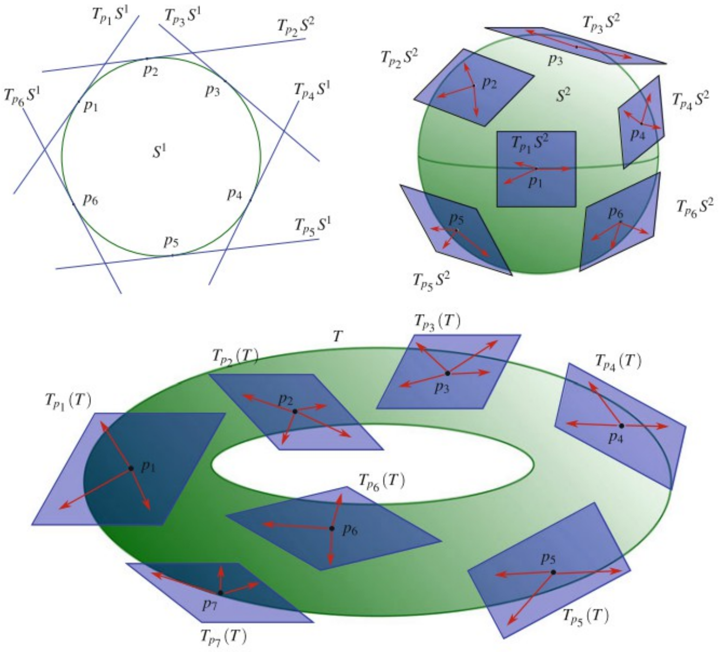

Fig. 2.14 The manifold S^1 along with a number of tangent lines depicted (top left). The manifold S^2 along with a number of tangent planes depicted (top right). The manifold T along with a number of tangent planes depicted (bottom). For both manifolds S^2 and T several vectors in each tangent plane are also depicted

To get a better idea of what the set of tangent spaces to a manifold look like, we have drawn a few tangent spaces to the three simple manifolds we had looked at earlier in Fig. 2.14, the circle S^1, the sphere S^2, and the torus T. For the two dimensional tangent spaces we have also drawn some elements of the tangent spaces, that is, vectors emanating from the point p. The tangent spaces of the circle are denoted by T_pS^1, the tangent spaces of the sphere are denoted by T_pS^2, and the tangent spaces of the torus are denoted by $T_p(T)$. We use the parenthesis here simply because T_pT looks a little odd.

Even though we are quite familiar with the manifolds $\mathbb{R}^2$ and $\mathbb{R}^3$, imagining the tangent spaces to these manifolds is a little strange, though is it something you have been implicitly doing since vector calculus. The tangent space of the manifold $\mathbb{R}^2$ at a point p is the set of all vectors based at the point p. In vector calculus we dealt with vectors with different base points a lot, but we always simply thought of these vectors as being *in* the manifold $\mathbb{R}^2$ or $\mathbb{R}^3$. Now we think of these vectors as belonging to a separate copy of the vector space $\mathbb{R}^2$ or $\mathbb{R}^3$ attached at the point and called the tangent space.

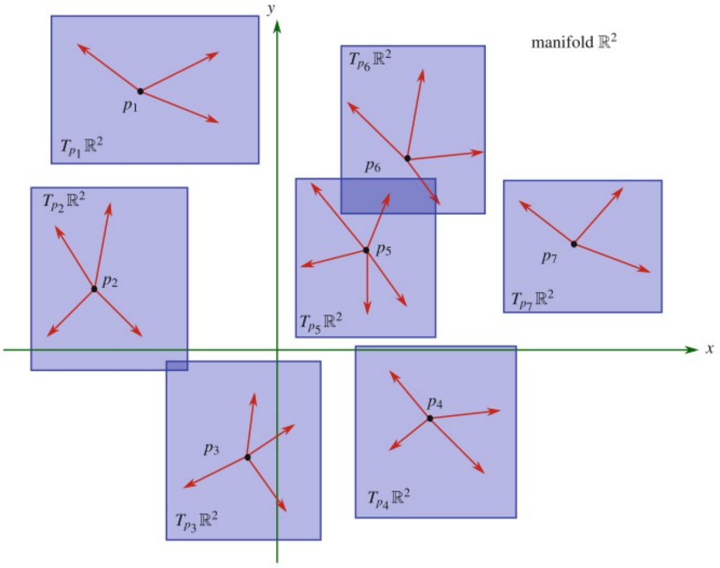
y
manifold $\mathbb{R}^2$
$T_{p_6}\mathbb{R}^2$
p_1
$T_{p_1}\mathbb{R}^2$
p_6
$T_{p_2}\mathbb{R}^2$
p_5
p_2
p_7
$T_{p_7}\mathbb{R}^2$
$T_{p_5}\mathbb{R}^2$
x
p_4
p_3
$T_{p_4}\mathbb{R}^2$
$T_{p_3}\mathbb{R}^2$

Fig. 2.15 The manifold $\mathbb{R}^2$ along with the tangent spaces $T_{p_i}\mathbb{R}^2$ at seven different points. A few vectors in each tangent space are shown

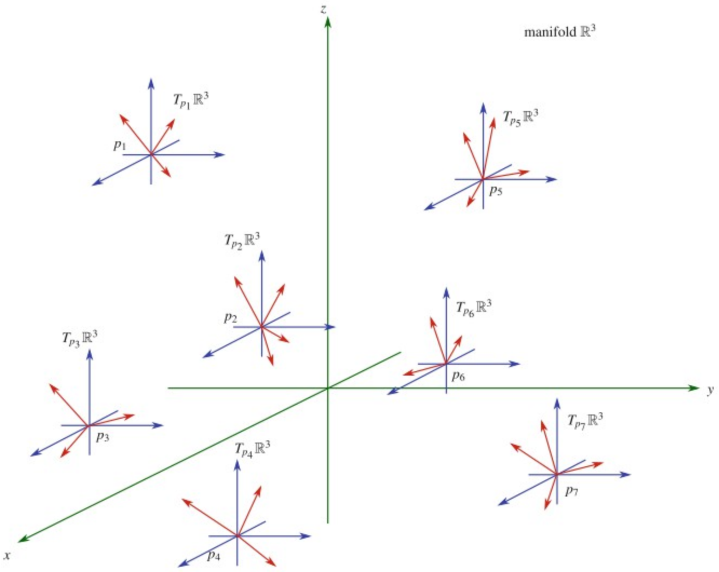
z
manifold $\mathbb{R}^3$
$T_{p_1}\mathbb{R}^3$
p_1
$T_{p_5}\mathbb{R}^3$
p_5
$T_{p_2}\mathbb{R}^3$
p_2
$T_{p_6}\mathbb{R}^3$
p_6
$T_{p_3}\mathbb{R}^3$
p_3
y
$T_{p_7}\mathbb{R}^3$
$T_{p_4}\mathbb{R}^3$
p_7
p_4
x

Fig. 2.16 The manifold $\mathbb{R}^3$ along with the tangent spaces $T_{p_i}\mathbb{R}^3$ at seven different points. A few vectors in each tangent space are shown

Figure 2.15 tries to help you imagine what is going on. The manifold $\mathbb{R}^2$ is pictured along with the tangent spaces $T_{p_i}\mathbb{R}^2$ at seven different points $p_1, \dots, p_7$. Each of the tangent spaces $T_p\mathbb{R}^2$ is **isomorphic** to ("the same as") $\mathbb{R}^2$. Though we have drawn the tangent spaces at only seven points, there is in fact a separate tangent space $T_p\mathbb{R}^2$ for every single point $p \in \mathbb{R}^2$.

Figure 2.16 tries to help you imagine the tangent spaces to $\mathbb{R}^3$. Again, the manifold $\mathbb{R}^3$ is pictured along with the tangent spaces $T_{p_i}\mathbb{R}^3$ at seven different points $p_1, \dots, p_7$. Each of the tangent spaces $T_p\mathbb{R}^3$ is naturally isomorphic to $\mathbb{R}^3$. Again, though we have drawn the tangent spaces at only seven points, there is in fact a separate tangent space $T_p\mathbb{R}^3$ for every single point $p \in \mathbb{R}^3$.

From now on we will often include the base point of a vector in our notation. For example, the vector v_p means vector v at the base point p of the manifold M. Clearly, the vector v_p belongs to, or is an element of, the tangent space T_pM. So, for example, we have

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix}_{(-2,-3)} \in T_{(-2,-3)}\mathbb{R}^2, \qquad \begin{bmatrix} 1 \\ 1 \end{bmatrix}_{(-1,4)} \in T_{(-1,4)}\mathbb{R}^2, \qquad \begin{bmatrix} 1 \\ 1 \end{bmatrix}_{(2,4)} \in T_{(2,4)}\mathbb{R}^2,$$

and

$$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}_{(-2,-3,-5)} \in T_{(-2,-3,-5)}\mathbb{R}^3, \qquad \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}_{(-1,4,3)} \in T_{(-1,4,3)}\mathbb{R}^3, \qquad \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}_{(2,4,-1)} \in T_{(2,4,-1)}\mathbb{R}^3.$$

The manifold M, along with a copy of the vector space T_pM attached at every point of the manifold is called the **tangent bundle** of the manifold M which is denoted as TM or $T(M)$. Often a cartoon version of the tangent bundle of a manifold M is drawn as in Fig. 2.17 using a one dimensional representation for both the manifold and the tangent spaces.

If M is an n dimensional manifold then TM is $2n$ dimensional. That makes intuitive sense since to specify exactly any element (vector) in TM you need to specify the vector part, which requires n numbers, and the base point, which also requires n numbers. Thus, $T\mathbb{R}^2$ has four dimensions and $T\mathbb{R}^3$ has six dimensions. In fact, the tangent bundle TM of a manifold M is itself also a manifold. Thus, if M is an n dimensional manifold then TM is a $2n$ dimensional manifold.

Now we will very briefly introduce the concept of a **vector field**. Doubtless you have been exposed to the idea of a vector field in vector calculus where it is usually defined on $\mathbb{R}^2$ or $\mathbb{R}^3$ as a vector-valued function that assigns to each point $(x, y) \in \mathbb{R}^2$ or $(x, y, z) \in \mathbb{R}^3$ a vector

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}_{(x,y)} \text{ or } \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}_{(x,y,z)}.$$

An example of a vector field on $\mathbb{R}^2$ is shown in Fig. 2.18.

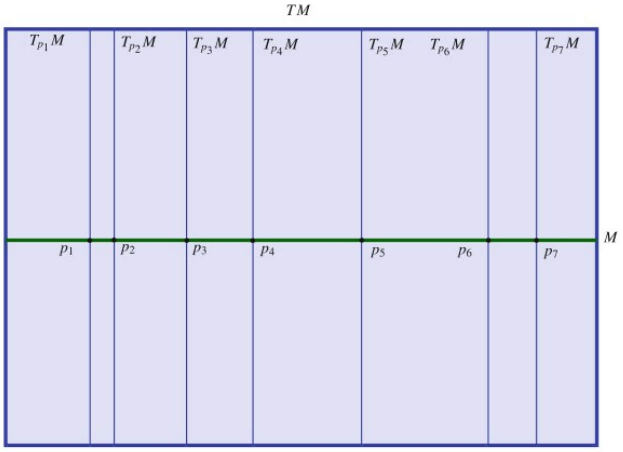

Fig. 2.17 A cartoon of a tangent bundle TM over a manifold M. Seven tangent spaces $T_{p_i}M$ are shown

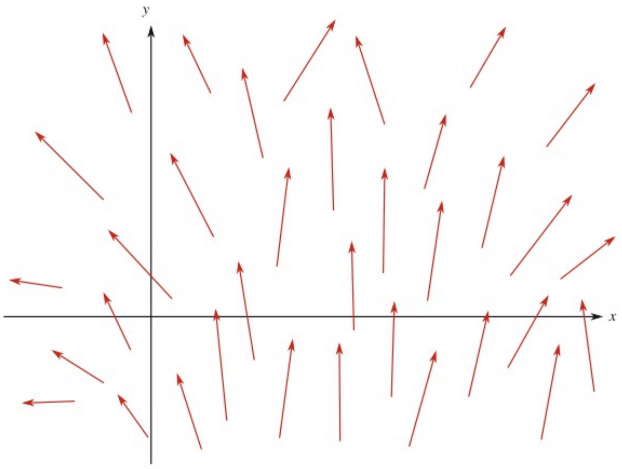

Fig. 2.18 A vector field drawn on the manifold $\mathbb{R}^2$. Recall, each of the vectors show are an element of a tangent space and not actually in the manifold $\mathbb{R}^2$

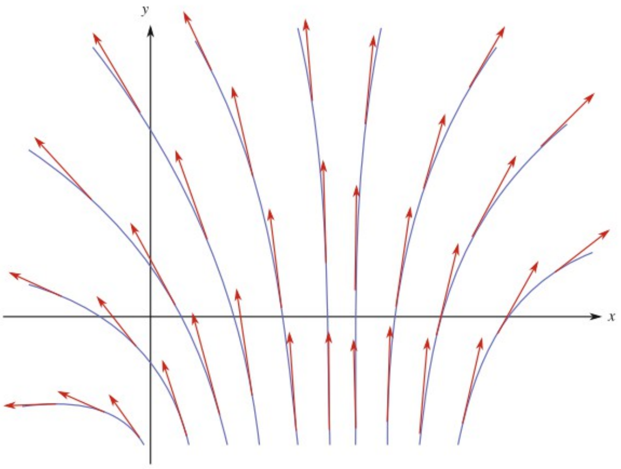

Fig. 2.19 A smooth vector field drawn on the manifold $\mathbb{R}^2$. Again, recall, each of the vectors shown are an element of a tangent space and not actually in the manifold $\mathbb{R}^2$

A vector field is called **smooth** if it is possible to find curves on the manifold such that all the vectors in the vector field are tangent to the curves. These curves are called the **integral curves** of the vector field, and finding them is essentially what differential equations are all about. Figure 2.19 shows several integral curves of a smooth vector field in $\mathbb{R}^3$.

While we certainly are not going to get into differential equations here, what we want to do is introduce a new way of thinking about vector fields. Notice, a vector field on a manifold such as $\mathbb{R}^2$ or $\mathbb{R}^3$ gives us a vector at each point of the manifold. But a vector v_p at a point p of a manifold M is an element of the tangent space at that point, T_pM. So being given a vector field is the same thing as being given an element of each tangent space of M. This is often called a **section** of the tangent bundle TM. Graphically, we often visualize sections of tangent bundles as in Fig. 2.20.

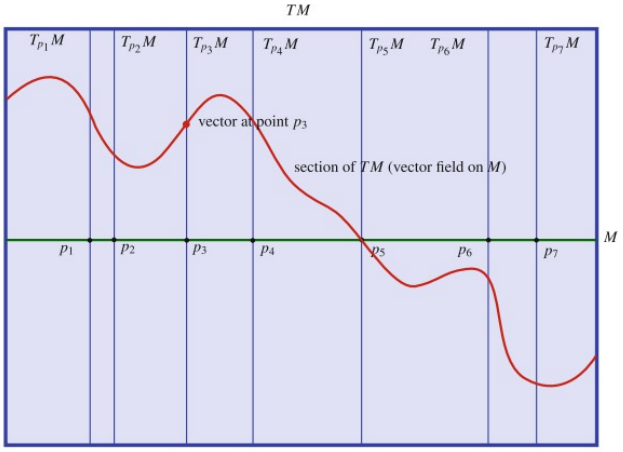

Fig. 2.20 A cartoon of a section of a tangent bundle over the manifold M. The section of TM is simply a vector field on M. The vector in the vector field on M at point p_3 is depicted as the point in $T_{p_3}M$ that is in the section. Notice, in this cartoon representation the vector at the point p_5 would be the zero vector

Question 2.1 Before we get into the next section let's take just a moment to review what derivatives were all about. Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = \frac{x^2}{2} + 2$.

(a) Sketch the graph of the function $f(x) = \frac{x^2}{2} + 2$.
(b) Find the derivative of $f(x)$.
(c) What is $f(2)$? What is $f(3)$? Draw a line between the points $\big(2, f(2)\big)$ and $\big(3, f(3)\big)$.
(d) What is the slope of this secant line?
(e) Write down the limit of the difference quotient definition of derivative.
(f) Use this definition to find the derivative of f at $x = 2$. Does it agree with what you found above?
(g) The derivative of f at $x = 2$ is a number. Explain what this number represents.

From the above question you should now remember that derivatives are basically slopes of lines tangent to the graph of the function at a certain point.

2.3 Directional Derivatives

Now the time comes to explore why we might actually want to use all of this. Suppose we were given a real-valued function f on the manifold $\mathbb{R}^n$. That is, $f : \mathbb{R}^n \to \mathbb{R}$. How might we use a vector v_p at some given point p on the manifold? One natural thing we could do is take the **directional derivative** of f at the point p in the direction of v_p. Consider Fig. 2.21 where a real-valued function $f : \mathbb{R} \to \mathbb{R}$ is shown. Given a point p in the domain there are two directions that it is possible to move in, left or right. Similarly, if you had a vector based at that point it could either point left, as v_p does, or right, as w_p does. We could ask how the function varies in either of these directions. In Fig. 2.22 a real-valued function $f : \mathbb{R}^2 \to \mathbb{R}$ is depicted. Given a point $p = (x, y)$ in the domain one can go any direction in two dimensions, and a vector based at that point could point in any direction in two dimensions as well. And similarly, we could ask how the function varies in any of these directions. In Fig. 2.23 we choose one of those directions, the direction given by the vector

$$v_p = \begin{bmatrix} a \\ b \end{bmatrix}_{(x,y)}.$$

We are able to use this vector to find the directional derivative of f in the direction v_p at the point p. In other words, we are able to find how f is changing in the direction v_p at the point p.

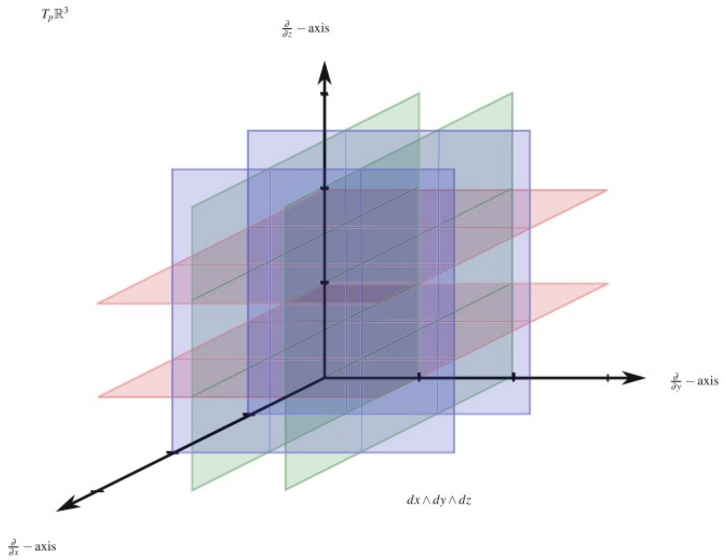
$T_p\mathbb{R}^3$
$\frac{\partial}{\partial z}$ − axis
$\frac{\partial}{\partial y}$ − axis
$\frac{\partial}{\partial x}$ − axis
$dx \wedge dy \wedge dz$

Fig. 5.30 The three-form $dx \wedge dy \wedge dz$ is made up of "cubes" whose sides are the planes perpendicular to the x-axis, the y-axis, and the z-axis

5.4 Three-Forms in $\mathbb{R}^3$

This is probably the shortest section in the entire book. And for a reason. Drawing three-forms in $T_p\mathbb{R}^3$ is in fact exceedingly simple. The space of three-forms on $\mathbb{R}$, denoted $\bigwedge^3(\mathbb{R})$, is actually of dimension one, which means that every three-form on $\mathbb{R}^3$ is of the form $c dx \wedge dy \wedge dz$ for some $c \in \mathbb{R}$. The one-form dx was given by the planes $x = n$, the one-form dy was given by the planes $y = n$, and the one-form dz was given by the planes $z = n$, where we had $n = 0, \pm 1, \pm 2, \ldots$. The three-form is simply all of these planes being drawn simultaneously, thereby filling space with unit cubes. See Fig. 5.30. For three-forms $a dx \wedge dy \wedge dz$, where $a \in \mathbb{R}$, the usual ambiguity arises.

Question 5.20 Based on Fig. 5.30 sketch three different pictures for the three-form $2dx \wedge dy \wedge dz$.

Question 5.21 Let

$$u = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, v = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, w = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Evaluate $dx \wedge dy \wedge dz$ on different permutations of u, v, w and use your answers to determine what kind of orientation is $dx \wedge dy \wedge dz$ has. (Note, somehow all three directions have to be taken into account.)

Question 5.22 How might you depict the positive orientation for $dx \wedge dy \wedge dz$ in Fig. 5.30? What about the negative orientation?

5.5 Pictures of Forms on Manifolds

What we have done so far is to attempt to draw graphical representations of one-forms and two-forms in $\mathbb{R}^2$ and one-forms, two-forms, and three-forms in $\mathbb{R}^3$. In each case we looked at some form α_p at the point p and found a representation of α_p in the tangent space at the point p. For a one-form on $\mathbb{R}^2$ our picture was of lines in $T_p\mathbb{R}^2$ (Figs. 5.1–5.7), while for a two-form on $\mathbb{R}^2$ our picture was of boxes in $T_p\mathbb{R}^2$ (Figs. 5.8–5.10). For a one-form in $\mathbb{R}^3$ our picture was of planes in $T_p\mathbb{R}^3$ (Figs. 5.11–5.18), for a two-form in $\mathbb{R}^3$ our picture was of tubes in $T_p\mathbb{R}^3$ (Figs. 5.19–5.29), and for a three-form in $\mathbb{R}^3$ our pictures was of cubes in $T_p\mathbb{R}^3$ (Fig. 5.30).

Sometimes attempts are made to utilize these graphical representations of one-forms and two-forms on $\mathbb{R}^2$ and one-forms, two-forms, and three-forms on $\mathbb{R}^3$ to "draw" forms on the whole manifold, at least for manifolds that are two or three dimensional. There are two ways that this is generally done. The first method at least makes sense mathematically, though it can still end up being incomplete in some sense and can potentially misrepresent the differential form. The second method, unfortunately, has some serious mathematical shortcomings, in addition to potentially misrepresenting the differential form. But when these methods do work they can be an extremely helpful and illuminating way to think about differential forms on two and three dimensional manifolds. It is best to keep in the back of your mind that the pictures these methods give are merely cartoons at best.

First, we will present the first method of visualizing forms on manifolds. Essentially what is done here is that a lattice of points $p_{ij} = (x_i, y_j)$ on the manifold is chosen and the line-stacks that represent the one-form in $T_{p_{ij}}M$ are drawn on the manifold at each of these lattice points. Figure 5.31 shows how this is done for a series of simple one-forms on $\mathbb{R}^2$. The arrow points indicate positive orientation. Of course, we can choose lattices of points on the manifold with differing densities. Figure 5.32 shows the same one-form $xdx + ydy$ on $\mathbb{R}^2$ using two different lattices of points. A two-form would be similar, only with a small grid of boxes drawn at each point. The drawback of this method is the spacing of the lattice of points; it is entirely possible that important details of the differential form can be missed if one chooses an inappropriate lattice. Be that as it may, one generally still obtains a reasonable cartoon that can at least help one imagine the differential form.

For three dimensional manifolds the pictures are analogous. First a three-dimensional lattice of points $p_{ijk} = (x_i, y_j, z_k)$ is chosen on the manifold. A one-form would then be represented as small stacks of sheets at each point p_{ijk}, a two-form would be represented with small bundles of tubes at each point p_{ijk}, and a three-form would be represented with small sets of cubes at each point p_{ijk}.

The second method does something similar, but tries to take it one step further. Consider Fig. 5.32 again. In the second image the lattice of points was chosen close enough so that, particularly in the middle, they start to merge into each other and form concentric circles. In the second method the little pictures at the lattice points are connected up to each other to provide one continuous image on the manifold. When this is actually possible the images are very appealing and simple to think about. First we will look at how method two works when it is appropriate and then we will briefly discuss how method two fails in general.

As an example, a generic one-form α on $\mathbb{R}^3$ is shown in Fig. 5.33. Here the line-stacks at the lattice points (not shown) are connected, as well as is possible, to generate curves that cover the manifold. The value $\alpha(v)$ is (approximately) given by the number of curves the vector v pierces. The case of two-forms on the manifold $\mathbb{R}^2$ is handled identically, only instead of small stacks of lines at each lattice point one would have small patches of boxes. Again, it may be possible to connect up the boxes to give a rough image of the two-form on the manifold. Consider Fig. 5.34. Here a two-form on manifold $\mathbb{R}^2$ is depicted on the manifold $\mathbb{R}^2$. In order to find the value of the two-form α at any given point p two vectors at that point need to be given, v_1 and v_2. The value $\alpha_p(v_1, v_2)$ is then given by the number of boxes that are inside the parallelepiped spanned by v_1 and v_2.

Attempts to picture one-forms, two-forms, and three-forms on the manifold $\mathbb{R}^3$ are very similar. Again, a lattice of points $p_{ijk} = (x_i, y_j, z_k)$ is chosen and at each of these points the picture of the form in $T_{p_{ijk}}\mathbb{R}^3$ is found. For a one-form α at each point p_{ijk} one would have a stack of two-dimensional sheets, similar to Fig. 5.17. Small copies of each of these pictures would then be superimposed on the manifold $\mathbb{R}^3$ at each lattice point. One would then try to connect up these sheets, as best as one could, to arrive at a picture similar to that of Fig. 5.35. How closely the sheets are packed (in other words, how far apart the sheets are from each other) is determined by the particular one-form. To find the (approximate) value of $\alpha(v)$ one would then count the number of sheets the vector v pierces. So, the sheets of the one-form 10α are ten times as close to each other as the sheets of the one-form α.

For a two-form α at each point p_{ijk} one would have a bundle of "tubes" similar to Fig. 5.23. At each lattice point the bundle of tubes would have varying densities and be in different directions. One would then try to connect up these sheets, as best as one could, to arrive at a picture similar to that of Fig. 5.36, where tubes of varying sizes twist through $\mathbb{R}^3$. The density at which the tubes are packed depends on the exact nature of the two-form α. In order to (approximately) find $\alpha(v_p, w_p)$ one finds the parallelepiped spanned by v_p and w_p and counts the number of tubes that go through this parallelepiped.

For a three-form α at each point p_{ijk} one would have a little set of cubes, similar to Fig. 5.30. One could imagine Rubik's cubes of varying sizes and orientations at each lattice point. Then these cubes are all connected up as well as possible to fill $\mathbb{R}^3$ with little cubes. We will not attempted to show that here. Finding the (approximate) value of $\alpha(u_p, v_p, w_p)$ amounts to counting the number of cubes that are inside the parallelepiped spanned by u_p, v_p, w_p.

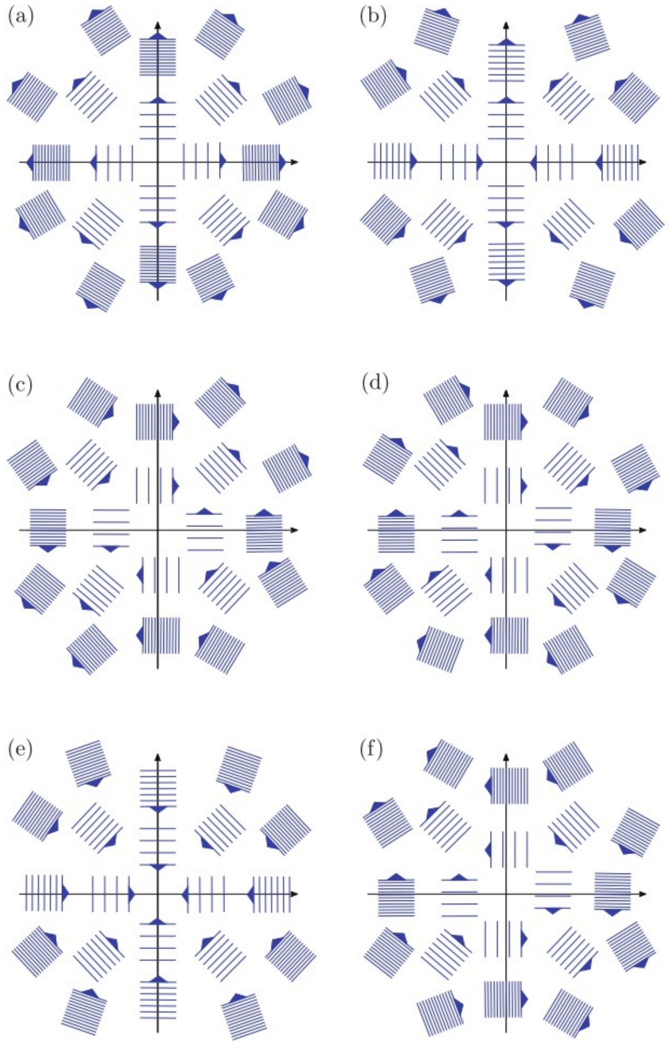

Fig. 5.31 One-forms on the manifold $\mathbb{R}^2$ represented as a grid of line-stacks on the manifold $\mathbb{R}^2$. The arrow points indicate positive direction (Images generated with Vector Field Analyzer II, Version 2, Kawski, 2001) . (**a**) $xdx + ydy$. (**b**) $-xdx + ydy$. (**c**) $ydx + xdy$. (**d**) $ydx - xdy$. (**e**) $-xdx - ydy$. (**f**) $-ydx - xdy$

There is something very appealing about this way of looking at differential forms, and in certain instances it helps provide an additional way of thinking about differential forms on two-dimensional and three-dimensional manifolds. Unfortunately, in general method two fails. In fact, one does not have to look very far to find examples of forms where it is actually impossible to visualize them using method two. One of the real issues is the "connecting" bit. Mathematically this is a complete slight-of-hand with no reasonable justification. In fact, each of the line-stacks for a one-form are in a different tangent space so "connecting" the line-stacks simply doesn't make sense. And even when we draw the pictures on the manifold as we have done, there is no reason to suppose that the line-stacks at points close to each other would be oriented in a way that would allow them to be connected, as was the case in Fig. 5.32. For a general one-form they need not be. And of course this applies to the plane-stacks and tube-bundles for three-dimensional manifolds as well.

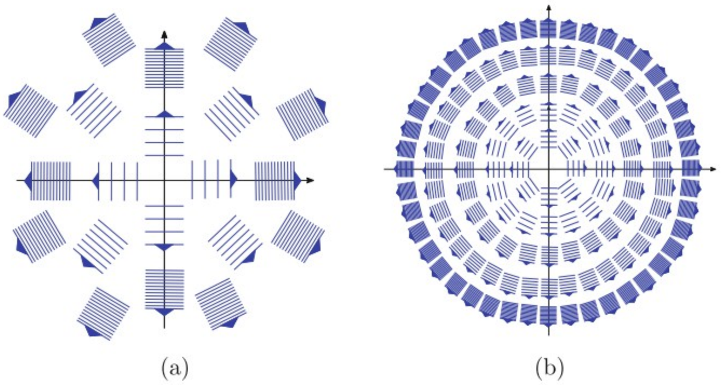

Fig. 5.32 The one-form $xdx + ydy$ on the manifold $\mathbb{R}^2$ represented as a grid of line-stacks using two different arrays of points. The arrow points indicate positive direction. Notice how in the image on the right the line-stacks in the middle start to look like they are merging into concentric circles (Images generated with Vector Field Analyzer II, Version 2, Kawski, 2001)

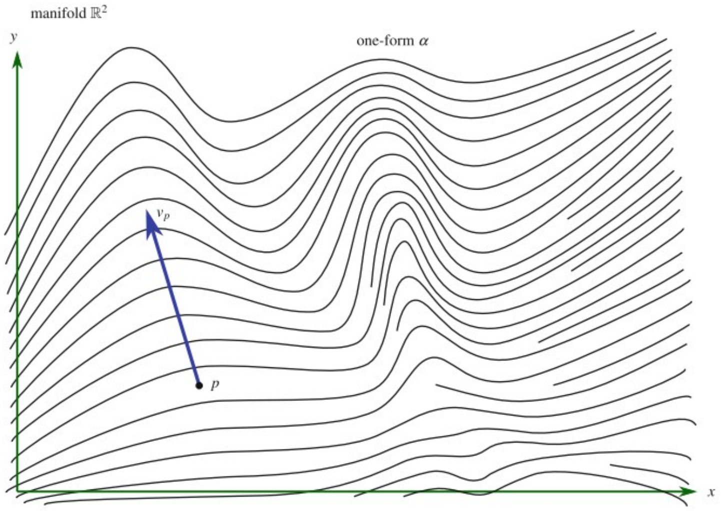

Fig. 5.33 A one-form α is shown on the manifold $\mathbb{R}^2$ with all the "line stacks" connected up smoothly as curves. The one-form applied to a vector is (approximately) equal to the number of curves "pierced" by the vector. Here, $\alpha(v_p) \approx 6$

A necessary, but not sufficient, requirement is that the distribution given by the kernel of the differential form be integrable. This would give what is called a foliation of the manifold which would allow the line-stacks, plane-stacks, or tube-bundles to line up. Exploring and explaining this is beyond the scope of this book. It is simply sufficient to realize that even a one-form on $\mathbb{R}^3$ as simple as $xdy + dz$ can not be visualized using method two.

However, this way of visualizing differential forms does appear in physics sometimes. Physicists are not so concerned about a completely general way of visualizing forms, they are interested in a way of visualizing forms that works for the physical problems and situations they are dealing with. Thus this method does have some utility in certain situations. When it is appropriate it does in fact provides a very nice way of thinking about forms. In fact, it gives a nice picture to go along with Stokes' theorem that is very appealing and will be presented in Sect. 11.6.

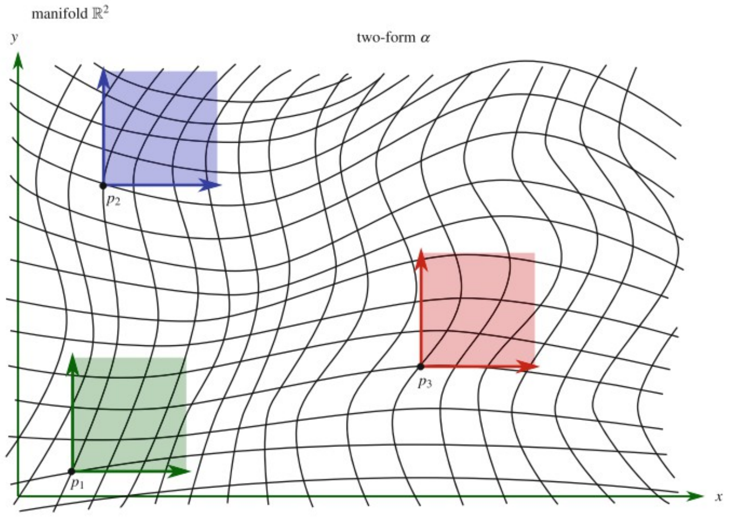

Fig. 5.34 A two-form α is shown on the manifold $\mathbb{R}^2$ with all boxes connected up smoothly. The unit vectors e_1 and e_2 are depicted at points p_1 (green), p_2 (blue), and p_3 (red). One (approximately) finds $\alpha_{p_i}(e_1, e_2)$ by counting the number of boxes that fall within the parallelepiped spanned by e_1 and e_2 at each of these points. Thus $\alpha_{p_1}(e_1, e_2) \approx 12$, $\alpha_{p_2}(e_1, e_2) \approx 16$, and $\alpha_{p_3}(e_1, e_2) \approx 9$. It is clear that while this is a nice picture the answers it gives are only very approximate

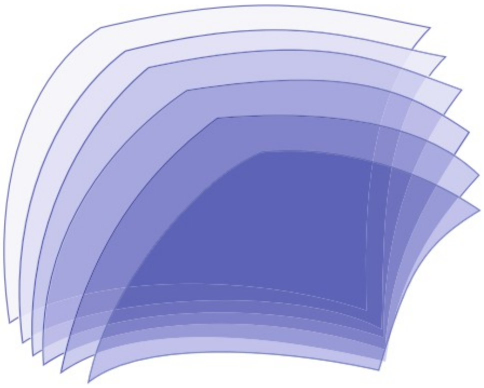

Fig. 5.35 A generic one-form in $\mathbb{R}^3$ is comprised of sheets filling $\mathbb{R}^3$

5.6 A Visual Introduction to the Hodge Star Operator

Introducing the Hodge star operator in a mathematically rigorous way at this point is a little tricky since much of the necessary mathematics has not yet been introduced, yet the geometric pictures of the Hodge star operator that go along with this chapter are really quite nice and easy to understand, so now is a natural time to do it. Thus we will try to strike a balance, providing some of the mathematics behind the pictures but not enough to overwhelm you with theoretical details. For the moment we will stick to the three dimensional case and give plausibility arguments instead of rigorous proofs.

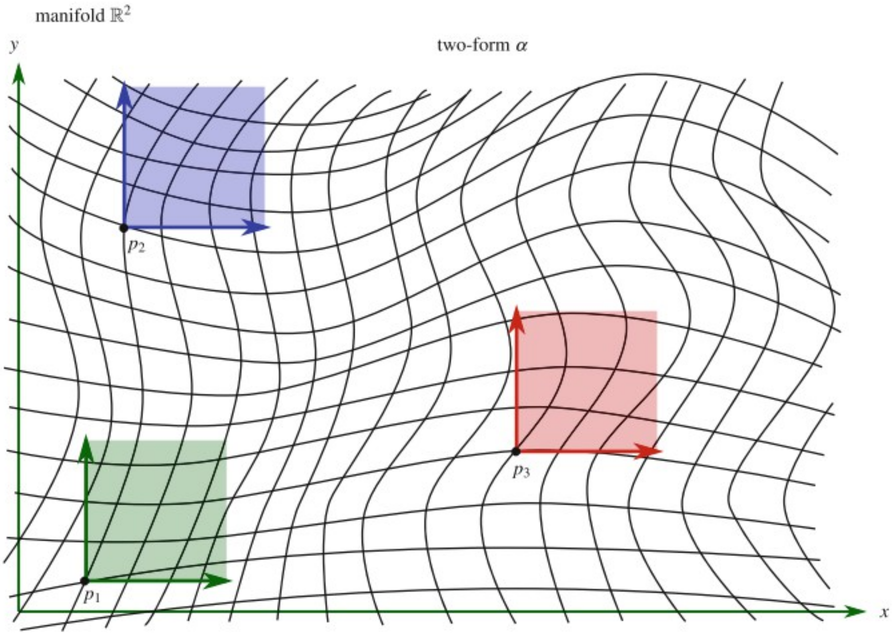

Fig. 5.34 A two-form α is shown on the manifold $\mathbb{R}^2$ with all boxes connected up smoothly. The unit vectors e_1 and e_2 are depicted at points p_1 (green), p_2 (blue), and p_3 (red). One (approximately) finds $\alpha_{p_i}(e_1, e_2)$ by counting the number of boxes that fall within the parallelepiped spanned by e_1 and e_2 at each of these points. Thus $\alpha_{p_1}(e_1, e_2) \approx 12$, $\alpha_{p_2}(e_1, e_2) \approx 16$, and $\alpha_{p_3}(e_1, e_2) \approx 9$. It is clear that while this is a nice picture the answers it gives are only very approximate

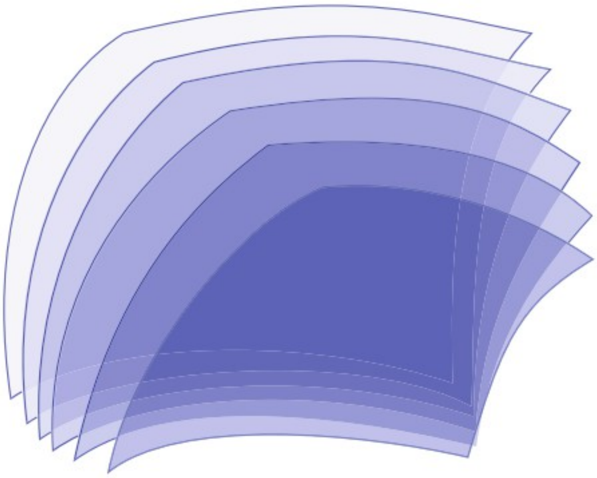

Fig. 5.35 A generic one-form in $\mathbb{R}^3$ is comprised of sheets filling $\mathbb{R}^3$

+,- * *. * + % + A 3 5
/ 4 D , B D? 3 & " 3 D % / & & &&'E
/ /"

(

4 / / / D " D
H " 4 D " / " / 3/ D
A / D / D J
H A J / B " "" "
+ 3 &$@

4 / / 3 5 +
/ G / D " H " "
" B D D & E / A "
/ / D " B " / /
B D / / 4 E E /
D / " %
3 5 B / & E " G / D
A " A H D / B
" / D 2D & / D / " B /
" E " F 3 D : 3
/ " " " " / D " " /
4 A " A H / " / /
" D E " "
E " / "
% & 3 5 / D " / / D "
/ " A / / A
/ / D D % E / E 3 5
B > B " / E / E 0 K "
3 5 B / / " AD H " E
" A B " / / /
/ 2 K " B 0 K " 3 "
/ D / " B H

( )& @
1 " 4 + , B -

5

Length and Time Are Relative Quantities

> [In order to measure the Lorentz] contraction we should have to operate with two perpendicular rods, and with two mutually interfering pencils of light, allowing the one to travel to and fro along the first rod, and the other along the second rod. But in this way we should come back once more to the Michelson experiment . . .
>
> *H. A. Lorentz (1895)*

> The special relativity theory resulted in appreciable advances [among them was that] it enforced the need for clarification of the fundamental concepts in epistemological terms.
>
> *A. Einstein (1923)*

A noncircular and unambiguous method for measuring length using light rays depended on Einstein's extending to every inertial reference system the axiom concerning the propagation of light relative to ether-fixed reference systems. The relativity of time and simultaneity was a startling consequence of Einstein's having combined axiomatics with an epistemological analysis of the measurement of time and length. Consequently, Einstein may well have believed it necessary to prepare the reader of 1905 for the consequences of an axiomatic approach to electrodynamics via kinematics, by presenting in §2 an illustrative example of how an observer in one inertial reference system notes how observers in another inertial system check the synchronization of their clocks.

Einstein began §2 by restating the two postulates that he had proposed in the introductory section [§2, 1–12]:

> §2. *On the Relativity of Lengths and Times*

> The following considerations are based on the principle of relativity and on the principle of the constancy of the velocity of light. We define these two principles thus—
>
> 1. The laws by which the states of physical systems undergo changes are independent of whether these changes of state are referred to one or the other of two coordinate systems moving relatively to each other in uniform translational motion.
>
> 2. Any ray of light moves in the "resting" coordinate system with the definite velocity c, which is independent of whether the ray was emitted by a resting or by a moving body. Consequently,
>
> $$\text{velocity} = \frac{\text{light path}}{\text{time interval}}$$
>
> where time interval is to be understood in the sense of the definition in §1.
>
> [In 1905 Einstein used the terms postulate and principle interchangeably.]

Having in §1 broken the vicious circle in clock synchronization, Einstein could define the velocity of light as:

$$\text{“velocity} = \frac{\text{light path}}{\text{time interval}}\text{”} \tag{5.1}$$

where the light path need not be closed.

In fact, Einstein (1907e) rewrote Equation (5.1) as

$$c = r/(t_B - t_A) \tag{5.2}$$

where r is the distance between A and B and he prefaced Eq. (5.2) with the statement that c was a "universal constant." Einstein continued in (1907e): "It is by no means natural to expect that the hypothesis stated above, which we call the principle of the constancy of the velocity of light, should be actually satisfied in nature, yet—at least for a coordinate system of a certain state of motion—it is made likely by the confirmations which Lorentz's theory, that is based on the assumption of an absolutely resting ether, have obtained by experiment." Einstein's two footnotes to this passage stated that by "Lorentz's theory" he meant the one in the (1895), and in the second footnote he wrote: "In particular the fact that this theory [Lorentz's (1895)] of the dragging coefficient (Fizeau's experiment) is in agreement with the available data." Therefore, for Einstein, Lorentz's (1895) explanation of Fizeau's 1851 experiment furnished sufficient support for the "principle of the constancy of the velocity of light."

The universality of the second principle is perhaps clouded by Einstein's use in 1905 of the term "resting system": yet any inertial system can be the "resting system." Throughout the 1905 paper, Einstein referred to a particular inertial reference system K as the resting system in order to emphasize how, in a certain sense, the relativity theory was a generalization of Lorentz's electromagnetic theory in which K was the ether-fixed reference system. As Einstein (1907e) wrote, the special relativity theory "originally arose from the coalescence of H. A. Lorentz's theory and the principle of relativity [from mechanics]."

Einstein continued in the relativity paper [§2, 13–58] with an epistemological analysis, based upon in principle operational definitions, of the concepts of simultaneity and length as these quantities referred to observers in different inertial reference systems.

Using Einstein's notation from his §3, consider two inertial reference systems k and K, equipped with measuring rods and clocks that are synchronized according to the procedure in §1. Suppose that initially k and K are relatively at rest and are superposed, i.e., their x-, y-, and z-axes are aligned. Observers in k and K measure the length of a rod lying along their common x-axes by comparing it with identical measuring rods at relative rest, i.e., by a congruence measurement. Let this length be l. Then k is accelerated to a velocity v relative to K, and Einstein assumes that the rod's characteristics are unaffected by the sudden acceleration.

Einstein next presented two in-principle operational procedures, or *Gedanken* experiments, in which we could "imagine [the moving rod's] length to be determined." (1) When an observer resting in *k* determined the rod's length by a congruence measurement, "according to the principle of relativity," the result had to be l. The magnitude of any discrepancy would be a measure of violation of this principle. Einstein defined the length l as "the length of the rod in the moving system."[1] (2) An observer at rest in K also measured the rod's length. "At a definite time," i.e., instantaneously, the observer in K marked off on K's x-axis the positions of the moving rod's end points A and B. Einstein referred to this length as r_{AB}—"the length of the (moving) rod in the resting system." According to the "two principles" of Einstein's theory the lengths l and r_{AB} differ, although, as Einstein emphasized, "current kinematics tacitly assumes [them to be] precisely equal." The crux of the matter was that the high velocity of light led scientists to assume the equivalence of "these two operations" and so the moving rod in "geometrical respects could be perfectly represented by *the same* body *at rest* in a definite position." Here Einstein carefully, albeit implicity, differentiated between two different notions of a body's shape—the geometrical and what he referred to (1907e) as the "kinematical." The observer in k measures the geometrical shape, and the K-observer the kinematical shape; "current kinematics tacitly" assumed their equality.

"We imagine further," wrote Einstein in the relativity paper, that the clocks at A and B in k are always synchronized with those in K. Einstein then offered the following *Gedanken* experiment. "Imagine" that two observers in k at A and B check their clocks' synchronization, using the two-way, two-clock procedure from §1, and assume also that an observer in K monitors this operation with clocks that are synchronized with those at A and B in k. In consequence, the observer in K notes three events E_1, E_2 and E_3 on K's clocks: E_1—a ray of light leaves A when A coincides with the position of a clock in K at time t_A; E_2—after traveling the distance r_{AB} (relative to K) the ray is received at B when B coincides with a clock in K at time t_B; E_3—the ray returns to A after traveling the distance r_{AB} when A

coincides with the clock in K at time t'_A. (Einstein's footnote emphasized that the times discussed were always those in K.) Einstein's mathematical analysis relates E_1 with E_2, and E_3 with E_4 as follows:

$$c(t_B - t_A) = r_{AB} + v(t_B - t_A), \tag{5.3}$$

$$c(t'_A - t_B) = r_{AB} - v(t'_A - t_B), \tag{5.4}$$

and he used the second postulate for calculating the left-hand sides of Eqs. (5.3) and (5.4). From Eqs. (5.3) and (5.4) follow Einstein's results:

$$t_B - t_A = r_{AB}/(c - v), \tag{5.5}$$

$$t'_A - t_B = r_{AB}/(c + v). \tag{5.6}$$

Thus, if the clocks at A and B on the moving rod remain in synchrony with those in K, they are not synchronous to observers in k. Consequently, wrote Einstein, "we can attribute no ***absolute*** meaning to the concept of simultaneity"—simultaneity is a relative concept. Einstein continued by emphasizing a consequence of the relativity of simultaneity: two spatially separated events that are simultaneous in K are not simultaneous in k, which "is in motion relatively to" K.

If, for the purpose of defining time, terms of order v/c are neglected, i.e., c is considered to be much larger than every other velocity as the ether theorists assumed, then Eqs. (5.5) and (5.6) become equal and the clocks at A and B remain in synchrony.

As far as I know Einstein's demonstration in §2 of the relativity of simultaneity is rarely analyzed. Most likely the reason is that the relativistic transformation equations for the space and time coordinates provide the quickest means for deducing the properties of time and simultaneity in moving reference systems. But Einstein's purpose in §2 was essentially epistemological and he did not require excess mathematical baggage to make his point—that is, by referring all measured quantities to the "resting system" K, whose analogue in the ether-based theories satisfied a principle of relativity and a light axiom, Einstein analyzed the meaning of the time of two moving clocks A and B by exploring how observers in another inertial reference system measure the time for a to-and-fro light exchange between A and B.

But what about the relativity of lengths which Einstein signaled in the title to this §2? Perhaps Einstein wished the reader to deduce a rough version of this result, thereby also allowing the reader to be convinced that it followed from the relativity of simultaneity. For example, at this point in the relativity paper, the reader could have deduced the relativity of length as follows. The observer in k at A registers the time interval for light to traverse to-and-fro the rod of length l in k, $\Delta t' = t'_A - t_A$ as

$$\Delta t' = 2l/c. \tag{5.7}$$

From Eqs. (5.5) and (5.6) the observer in K registers the time interval as

$$\Delta t = 2r_{AB}/c(1 - v^2/c^2). \tag{5.8}$$

Comparing these two time intervals yields

$$\frac{\Delta t'}{\Delta t} = \frac{l(1 - v^2/c^2)}{r_{AB}}. \tag{5.9}$$

We cannot progress any further toward an exact result, however, without knowing how r_{AB} and l are related—that is, we need the transformation equations relating the space and time coordinates in k and K. Perhaps Einstein wanted the reader to realize from Eq. (5.9) that in a geometry where lengths were measured using light rays and where clocks were synchronized by exchanging light signals—and thus simultaneity was a relative notion (i.e., $\Delta t' \neq \Delta t$)—then r_{AB} could not equal l. Or perhaps Einstein was inviting the reader to substitute into Eq. (5.9) a mathematically equivalent result from Lorentz's electrodynamics—that is, $r_{AB} = l\sqrt{1 - v^2/c^2}$, the Lorentz contraction, thereby obtaining the result from Einstein's §4 for the relativity of time.

The notion of measuring lengths using light rays had been proposed before Einstein by A. A. Michelson (1881) and (1887), Lorentz [(1892b), (1895)] and Poincaré [(1905a), (1906)]. Lorentz and Poincaré had proposed, and then rejected, two methods for measuring the Lorentz contraction: (1) a congruence measurement, but the measuring rod would likewise undergo a contraction; (2) a light geometry using two transverse rods of equal length, but this was just another Michelson-Morley experiment. Einstein realized that unambiguous length measurement of moving objects using light rays had to be taken in combination with properly synchronized clocks. To this we add Einstein's insistence that the problem of determining changes of length should for the present be treated as a kinematical problem.

In summary, the far-reaching results of §2 were that there were no such notions as the true time or the true length of an object; rather these were relative concepts: for example, the length of the rod was either l or r_{AB}, depending upon the rod's motion relative to an inertial observer.

Footnote

1. Einstein (1921b) elaborated on the relativistic notion of congruence as follows: We make two "tracts," i.e., marks, on a practically rigid body. If the tracts on B were congruent with those on another practically rigid body A, when A and B were at relative rest, they would be congruent whenever A and B were brought together.

(& $@

7

The Relativity of Length and Time

> *A priori* it is quite clear that we must learn something about the physical behaviour of measuring-rods and clocks from the equations of transformation, for the magnitudes x, y, z, t are nothing more nor less than the results of measurements obtainable by means of measuring-rods and clocks. If we had based our considerations on the Galilei transformation we should not have obtained a contraction of the rod as a consequence of its motion.
>
> *A. Einstein (1917a)*

Straightaway [§4, 1–2] Einstein emphasized the "physical meaning" of the results to follow: "Physical Meaning of the Equations Obtained Concerning Moving Rigid Bodies and Moving Clocks."

7.1 Moving Bodies: Einstein's Length Contraction

> [If] one should define as 'time' what Lorentz introduced as an auxiliary quantity, and called the local time [then] the hypothesis of H. A. Lorentz and Fitzgerald appears as a secondary consequence of the theory.
>
> *A. Einstein (1907e)*

Einstein quickly demonstrated that a body having a spherical shape relative to an observer in its rest system k, Eq. [§4.1], appeared to be an ellipsoid of revolution to an observer in K, Eq. [§4.2] [§4, 3–13]. Although the sphere's dimensions normal to its direction of motion remain unchanged, its dimension along the common direction of the x-axes of k and K "appears shortened in the ratio 1: $\sqrt{1 - v^2/c^2}$,

i.e., the greater the value of v, the greater the shortening"; consequently, length is a relative quantity.[1] The terseness of Einstein's derivation of the contraction of moving bodies was perhaps meant to demonstrate not only the power of his axiomatic method vis-à-vis Lorentz's in 1895 which required interlocking hypotheses and fictitious reference systems, but also to emphasize that the Lorentz contraction is a "secondary consequence of the theory." In addition, Einstein's derivation set the stage for §8 where he discusses the transformation properties of the energy of a light pulse.

As v approaches c the contraction in the direction of motion becomes more pronounced (from Eq. [§4.3] the semi-minor axis is $R\sqrt{1 - v^2/c^2}$). The ellipsoid gets flatter until when $v = c$ it appears to an observer in K as a plane perpendicular to the ξ-axis; nevertheless, to an observer in k the object remains a sphere as required by the principle of relativity. For $v > c$ the ellipsoid's semi-minor axis becomes imaginary, and so, as Einstein wrote, "our deliberations become meaningless."[2] (This was also Wien's (1904b) reason for c being the ultimate velocity, and from his readings in the *Annalen* Einstein could have been aware of speculations on hyperlight velocities, particularly the controversy between Abraham (1904b) and Wien (1904b).) Perhaps Einstein did not mention here that unphysical results ensue also for $v = c$ because this was obvious from the relativistic transformations of §3. This result may have provided further evidence for the impossibility of catching up with a point on a light wave, thereby viewing it as a standing wave. In (1917a) Einstein again demonstrated from the length contraction that the velocity of light in vacuum is the ultimate velocity, and he added (1917a)—

> Of course this feature of the velocity c as a limiting velocity also follows clearly from the equations of the Lorentz transformation, for these become meaningless if we choose values of v greater than c.

Einstein's statement supports my assertion above that certain results concerning $v \geq c$ were contained already in §3 of the 1905 relativity paper.[3] Einstein concluded his analysis of the shape of a moving body by asserting the reciprocity of length contraction.

Einstein's footnote to the passage under discussion ("That is, a body possessing spherical form when examined at rest.") refers to what in (1907e) he defined as the "geometrical shape"; he went on to define the shape measured by the observer in K as the "kinematical shape." Einstein's definitions served to emphasize that his length contraction differed fundamentally from Lorentz's which asserted that something happened to the rod in S_r, but compensating effects prevented the inertial observer from ascertaining changes in shape. Einstein's (1907e) derivation of the relativistic contraction exposed more clearly than the one in 1905, the relation between the relativities of length and time: In k there is a rod of length $\xi_2 - \xi_1 = l_0$. Two observers in K simultaneously, i.e., $t_1 = t_2$, mark on K's axis the rod's length as $x_2 - x_1 = l$. From Eq. (6.53) Einstein obtained $l = l_0\sqrt{1 - v^2/c^2}$.[4]

7.2 Moving Clocks: The Relativity of Time

Consider [§4, 20–48] that K's x-axis is lined with synchronized clocks, and at k's origin there is one clock. If the origins of k and K coincided at $t = \tau = 0$, then at a later time t the origin of k has moved a distance $x = vt$ relative to K's origin, and [see Eq. (6.52)] the clock at k's origin registers a time observed from K of

$$\tau = t\sqrt{1 - v^2/c^2} \tag{7.1}$$

where the quantity τ is the elapsed time on the clock at k's origin, and t is the time on a clock located at $x = vt$ on K's x-axis; consequently, the slowing down of a moving clock, i.e., time dilation, is measured by comparing one clock with at least two others. In order to emphasize the physical content of time dilation, Einstein rewrote the result as

$$\tau = t - (1 - \sqrt{1 - v^2/c^2})t, \tag{7.2}$$

so the time interval in k is thus slower than the one in K by an amount

$$(1 - \sqrt{1 - v^2/c^2}). \tag{7.3}$$

Expanding the square root in Eq. (7.3), and retaining terms up to second order in v/c, gives $\frac{1}{2}(v^2/c^2)t$. Thus, like length time too was a relative quantity which depended on only the magnitude of the relative velocity between two inertial reference systems. The relativity of time was a new result of Einstein's theory, and Einstein proceeded to investigate it further.

Suppose that separated clocks at A and B on K's x-axis are properly synchronized. Let the clock at A be instantaneously accelerated to the velocity v and transported over to B where it is instantaneously decelerated to rest.[5] A "peculiar consequence" of the prediction of the relativity of time is that the clocked moved from A to B is not synchronous with the one that remained at rest at position B: it is slower than the one at B by an amount of time $(1 - \sqrt{1 - v^2/c^2})$. This difference disposed of Poincaré's 1898 suggestion for synchronizing distant clocks by transporting a clock synchronous with a master clock. Einstein then took this "peculiar consequence" one step further by investigating the case in which the points A and B coincided, and their synchronized clocks C_1 and C_2 were side by side in K. Clock C_1 remains at rest but C_2 undergoes a round trip journey along the sides of a polygon. We may assume that C_2 moves in a succession of inertial reference systems $\mathrm{k}_1, \ldots, \mathrm{k}_n$, and may thus neglect the abrupt accelerations at each corner of the polygonal path when C_2 is transferred from one inertial system to another. Since each inertial system moves with velocity v relative to K, then Einstein claimed that when C_2 returned to K it lagged behind C_1 by the amount $(1 - \sqrt{1 - v^2/c^2})$, and he assumed the validity of this result when the polygon was replaced by a continuous curve. As an example, Einstein predicted that a clock placed at the equator ran slower than a clock placed at one of the poles. His extension to the case where C_2's trajectory was a continuous curve was unwarranted in 1905, but perhaps he considered that this case could always be treated as the limiting case of a many-sided polygon.[6] (Sommerfeld's footnote to this passage asserts that a pendulum-clock had to be excluded from the considerations of §4 because its oscillations are determined by gravity and hence cannot be discussed kinematically. See Appendix.)

Einstein's (1907e) discussion of time dilation did not contain any "peculiar consequences" because he proposed a real experiment, and not a *Gedanken* experiment, to verify the time dilation phenomenon. In 1907 Einstein went far beyond his intent in 1905 of basing relativity theory upon clocks and measuring rods taken as irreducible structures; now he defined a clock as any periodic process—for example, an atomic oscillator emitting a frequency ν_0 as measured in k. From Eq. (7.1) an observer in K measures the frequency

$$\nu = \nu_0\sqrt{1 - v^2/c^2}, \tag{7.4}$$

and Einstein wrote that "the equation [7.4] admits of a very interesting application" since Stark (1906) had observed a Doppler shift in the radiation emitted by canal rays. The "interesting application" of Eq. (7.4) was that it predicted a Doppler effect when the observer's line of sight was normal to the emitter's direction of motion, i.e., a transverse Doppler effect.[7] Einstein asserted that observation of this effect vindicated also time dilation because "we can consider the ion to be a clock." Lorentz's theory did not predict a transverse Doppler effect, although Einstein made no mention of this point (see Chapter 10).

Einstein's prediction was tested successfully in 1937 by two American physicists H. E. Ives and G. R. Stilwell, although Ives maintained that their research verified the "Larmor-Lorentz" theory (1938).[8]

7.3 On the Reciprocity of Relativistic Phenomena

> The behaviour of measuring rods and clocks in translational motion, when viewed superficially, gives rise to a *remarkable paradox* which on closer examination, however, vanishes (italics in original).
>
> *H. A. Lorentz (1910–1912)*

In his lectures at Leiden University (1910–1912) Lorentz analyzed two paradoxes rooted in inferring that the reciprocity of the relativistic contraction and time dilation were achieved merely by interchanging Greek and Roman letters, and replacing v with $-v$. After all, according to the relativity theory all effects were relative. Lorentz first discussed the length contraction.

Consider that the reference systems K and k are equipped with identical meter sticks of length l and are positioned on the systems' x-axes; Lorentz referred to these rods in k and K as 1 and 2, respectively. Rod 1, wrote Lorentz, "appears" to the K-observer to be smaller than the rod 2; rod 2 appears to the k-observer as smaller than rod 1. "But how," continued Lorentz, "can the rod 1 be shorter than and, at the same time, longer than rod 2?" There is no paradox here, rather, Lorentz

25.09.2019

are relative quantities

Obs: Michelson experiment...
H. A. Lorentz (1895)

Pág. 189 (226)

Pág. 209 (246)

7.1. Moving Bodies:
Einstein's Length
Contraction

... what Lorentz introduced as an auxiliary quantity, and called the "local time" [then] the hypothesis of H. A. Lorentz and Fitzgerald appears as a secondary consequence of the theory".
A. Einstein (1907e)

Important: Einstein quickly demonstrated that

25.09.2019

a body having a spherical shape relative to an observer in its rest system K, Eq. [§4.1]...

f) Hyperbolic Triangle centers
Ungar.

Source: 20190926/ the Special Theory of Relativity Einstein's World in New Axiomatic/ Ungar, Abraham.. Hyperbolic..pdf

Pág. 64(79)
3.4 The Resultant Relativistically Invariant Mass Theorem
Lemma 3.1.

(101) 20190927

$\pi = \frac{355}{113}$

Source: 20171111/Aurelio Baldor - Geometria.pdf
Pág. 196(200) → 258. El numero π.
Pág. 198(202) → Metodo Grafico para

+,- * *. * 7 E +A + 0D/ A
? 4 / B // E / /"
(@$ '& 1 !

Accordingly, we wish to determine the relativistically invariant mass m_0 of the system S, and the velocity $\mathbf{v}_0$ relative to $\Sigma_{\mathbf{0}}$ of a fictitious inertial frame, called the center of momentum frame, relative to which the three-momentum of S vanishes.

Assuming that the four-momentum is additive, the sum of the four-momenta of the N particles of the system S gives the four-momentum $(m_0\gamma_{\mathbf{v}_0}, m_0\gamma_{\mathbf{v}_0}\mathbf{v}_0)^t$ of S. Accordingly,

$$\sum_{k=1}^{N} m_k \begin{pmatrix} \gamma_{\mathbf{v}_k} \\ \gamma_{\mathbf{v}_k}\mathbf{v}_k \end{pmatrix} = m_0 \begin{pmatrix} \gamma_{\mathbf{v}_0} \\ \gamma_{\mathbf{v}_0}\mathbf{v}_0 \end{pmatrix}, \tag{3.14}$$

where

(i) The invariant masses $m_k > 0$ and the velocities $\mathbf{v}_k \in \mathbb{R}_c^n$, $k = 1, \ldots, N$, relative to $\Sigma_{\mathbf{0}}$ of the constituent particles of S are given, while

(ii) The invariant mass m_0 of S and the velocity $\mathbf{v}_0$ of the center of momentum frame of S relative to $\Sigma_{\mathbf{0}}$ are to be determined uniquely by the *Resultant Relativistically Invariant Mass Theorem*, which is Theorem 3.2 in Sect. 3.4

If $m_0 > 0$ and $\mathbf{v}_0 \in \mathbb{R}_c^n$ that satisfy (3.14) exist then, as anticipated, the three-momentum of the system S relative to its center of momentum frame vanishes since, by (3.8) and (3.14), the four-momentum of S relative to its center of momentum frame is given by

$$\begin{aligned} L(\ominus\mathbf{v}_0)\sum_{k=1}^{N} m_k \begin{pmatrix} \gamma_{\mathbf{v}_k} \\ \gamma_{\mathbf{v}_k}\mathbf{v}_k \end{pmatrix} &= L(\ominus\mathbf{v}_0) m_0 \begin{pmatrix} \gamma_{\mathbf{v}_0} \\ \gamma_{\mathbf{v}_0}\mathbf{v}_0 \end{pmatrix} \\ &= m_0 \begin{pmatrix} \gamma_{\ominus\mathbf{v}_0\oplus\mathbf{v}_0} \\ \gamma_{\ominus\mathbf{v}_0\oplus\mathbf{v}_0}(\ominus\mathbf{v}_0\oplus\mathbf{v}_0) \end{pmatrix} = m_0 \begin{pmatrix} 1 \\ \mathbf{0} \end{pmatrix}, \end{aligned} \tag{3.15}$$

noting that $\gamma_{\ominus\mathbf{v}_0\oplus\mathbf{v}_0} = \gamma_{\mathbf{0}} = 1$.

3.4 The Resultant Relativistically Invariant Mass Theorem

Lemma 3.1 below presents an identity that we need for the proof of the *Resultant Relativistically Invariant Mass Theorem* 3.2.

Lemma 3.1 *Let N be any positive integer, and let $m_k \in \mathbb{R}$ and $\mathbf{v}_k \in \mathbb{R}_c^n$, $k = 1, \ldots, N$, be N real numbers and N points of an Einstein gyrogroup $\mathbb{R}_c^n = (\mathbb{R}_c^n, \oplus)$. Then*

$$\left(\sum_{k=1}^{N} m_k\gamma_{\mathbf{v}_k}\frac{\mathbf{v}_k}{c}\right)^2 = \left(\sum_{k=1}^{N} m_k\gamma_{\mathbf{v}_k}\right)^2 - \left\{\left(\sum_{k=1}^{N} m_k\right)^2 + 2\sum_{\substack{j,k=1\\ j<k}}^{h} m_j m_k(\gamma_{\ominus\mathbf{v}_j\oplus\mathbf{v}_k} - 1)\right\}. \tag{3.16}$$

Proof The proof is given by the following chain of equations, which are numbered for subsequent explanation:

$$
\begin{aligned}
&\left(\sum_{k=1}^{N} m_k \gamma_{\mathbf{v}_k} \frac{\mathbf{v}_k}{c}\right)^2 \\
&\overbrace{=\!=}^{(1)} \sum_{k=1}^{N} m_k^2 \gamma_{\mathbf{v}_k}^2 \frac{\mathbf{v}_k^2}{c^2} + 2 \sum_{\substack{j,k=1\\ j<k}}^{h} m_j m_k \gamma_{\mathbf{v}_j} \gamma_{\mathbf{v}_k} \frac{\mathbf{v}_j \cdot \mathbf{v}_k}{c^2} \\
&\overbrace{=\!=}^{(2)} \sum_{k=1}^{N} m_k^2 \big(\gamma_{\mathbf{v}_k}^2 - 1\big) + 2 \sum_{\substack{j,k=1\\ j<k}}^{h} m_j m_k (\gamma_{\mathbf{v}_j} \gamma_{\mathbf{v}_k} - \gamma_{\ominus \mathbf{v}_j \oplus \mathbf{v}_k}) \\
&\overbrace{=\!=}^{(3)} \sum_{k=1}^{N} m_k^2 \gamma_{\mathbf{v}_k}^2 - \sum_{k=1}^{N} m_k^2 + 2 \sum_{\substack{j,k=1\\ j<k}}^{h} m_j m_k \gamma_{\mathbf{v}_j} \gamma_{\mathbf{v}_k} \\
&\qquad - 2 \sum_{\substack{j,k=1\\ j<k}}^{h} m_j m_k \gamma_{\ominus \mathbf{v}_j \oplus \mathbf{v}_k} \\
&\overbrace{=\!=}^{(4)} \left(\sum_{k=1}^{N} m_k \gamma_{\mathbf{v}_k}\right)^2 - \left\{\sum_{k=1}^{N} m_k^2 + 2 \sum_{\substack{j,k=1\\ j<k}}^{h} m_j m_k \gamma_{\ominus \mathbf{v}_j \oplus \mathbf{v}_k}\right\} \\
&\overbrace{=\!=}^{(5)} \left(\sum_{k=1}^{N} m_k \gamma_{\mathbf{v}_k}\right)^2 - \left\{\left(\sum_{k=1}^{N} m_k\right)^2 + 2 \sum_{\substack{j,k=1\\ j<k}}^{h} m_j m_k (\gamma_{\ominus \mathbf{v}_j \oplus \mathbf{v}_k} - 1)\right\}. \qquad (3.17)
\end{aligned}
$$

The assumption $\mathbf{v}_k \in \mathbb{R}_c^n$ implies, by (1.3), p. 4, that all gamma factors in (3.16)–(3.17) are real and greater than 1. Derivation of the numbered equalities in (3.17) follows:

1. This equation is obtained by an expansion of the square of a sum of vectors in $\mathbb{R}^n$.
2. Follows from Item 1 by (1.9)–(1.10), p. 5.
3. Follows from Item 2 by an obvious expansion.
4. Follows from Item 3 by an expansion of the square of a sum of real numbers.
5. Follows from Item 4 by an expansion of another square of a sum of real numbers. □

Einstein velocity addition law (1.2), p. 4, admits the following theorem:

Theorem 3.2 (Resultant Relativistically Invariant Mass Theorem) *Let* $(\mathbb{R}_c^n, \oplus)$ *be an Einstein gyrogroup, and let* $m_k \in \mathbb{R}$ *and* $\mathbf{v}_k \in \mathbb{R}_c^n$, $k = 1, 2, \ldots, N$, *be* N *real* *numbers and N elements of* $\mathbb{R}_c^n$ *satisfying*

$$\sum_{k=1}^{N} m_k \gamma_{\mathbf{v}_k} \neq 0. \tag{3.18}$$

Furthermore, let

$$\sum_{k=1}^{N} m_k \begin{pmatrix} \gamma_{\mathbf{v}_k} \\ \gamma_{\mathbf{v}_k} \mathbf{v}_k \end{pmatrix} = m_0 \begin{pmatrix} \gamma_{\mathbf{v}_0} \\ \gamma_{\mathbf{v}_0} \mathbf{v}_0 \end{pmatrix} \tag{3.19}$$

be an $(n+1)$*-vector equation for the two unknowns* $m_0 \in \mathbb{R}$ *and* $\mathbf{v}_0 \in \mathbb{R}^n$.

Then (3.19) *possesses a unique solution* $(m_0, \mathbf{v}_0)$, $m_0 \neq 0$, $\mathbf{v}_0 \in \mathbb{R}_c^n$, *satisfying the following three identities for all* $\mathbf{w} \in \mathbb{R}_c^n$ *(including, in particular, the interesting special case of* $\mathbf{w} = \mathbf{0}$*):*

$$\mathbf{w} \oplus \mathbf{v}_0 = \frac{\sum_{k=1}^{N} m_k \gamma_{\mathbf{w} \oplus \mathbf{v}_k} (\mathbf{w} \oplus \mathbf{v}_k)}{\sum_{k=1}^{N} m_k \gamma_{\mathbf{w} \oplus \mathbf{v}_k}}, \tag{3.20}$$

$$\gamma_{\mathbf{w} \oplus \mathbf{v}_0} = \frac{\sum_{k=1}^{N} m_k \gamma_{\mathbf{w} \oplus \mathbf{v}_k}}{m_0}, \tag{3.21}$$

$$\gamma_{\mathbf{w} \oplus \mathbf{v}_0} (\mathbf{w} \oplus \mathbf{v}_0) = \frac{\sum_{k=1}^{N} m_k \gamma_{\mathbf{w} \oplus \mathbf{v}_k} (\mathbf{w} \oplus \mathbf{v}_0)}{m_0}, \tag{3.22}$$

where

$$m_0 \ = \ \sqrt{\left(\sum_{k=1}^{N} m_k \right)^2 + 2 \sum_{\substack{j,k=1 \\ j<k}}^{N} m_j m_k (\gamma_{\ominus(\mathbf{w} \oplus \mathbf{v}_j) \oplus (\mathbf{w} \oplus \mathbf{v}_k)} - 1)}. \tag{3.23}$$

Proof Following (3.18), we assume, without loss of generality, that

$$\sum_{k=1}^{N} m_k \gamma_{\mathbf{v}_k} > 0 \tag{3.24}$$

(otherwise, we replace each m_k by $-m_k$, resulting in the replacement of m_0 by $-m_0$). Let us consider the following four equations, (3.25)–(3.28), which are specialized from (3.20)–(3.23) by taking $\mathbf{w} = \mathbf{0}$:

$$\mathbf{v}_0 = \frac{\sum_{k=1}^{N} m_k \gamma_{\mathbf{v}_k} \mathbf{v}_k}{\sum_{k=1}^{N} m_k \gamma_{\mathbf{v}_k}}, \tag{3.25}$$

$$\gamma_{\mathbf{v}_0} = \frac{\sum_{k=1}^{N} m_k \gamma_{\mathbf{v}_k}}{m_0}, \tag{3.26}$$

$$\gamma_{\mathbf{v}_0}\mathbf{v}_0 = \frac{\sum_{k=1}^{N} m_k \gamma_{\mathbf{v}_k}\mathbf{v}_0}{m_0}, \tag{3.27}$$

where

$$m_0 \;=\; \sqrt{\left(\sum_{k=1}^{N} m_k\right)^2 + 2\sum_{\substack{j,k=1\\ j<k}}^{N} m_j m_k (\gamma_{\ominus\mathbf{v}_j\oplus\mathbf{v}_k} - 1)}. \tag{3.28}$$

The proof of this theorem consists of two parts. In the first part of the proof, we show that if (3.19) for the unknowns $\mathbf{v}_0 \in \mathbb{R}^n$ and $m_0 \in \mathbb{R}$ possesses a solution, then the solution must be given uniquely by $\mathbf{v}_0$ of (3.25) and m_0 of (3.28), with $\mathbf{v}_0 \in \mathbb{R}^n_c$ and $m_0 > 0$, satisfying (3.26)–(3.27).

In the second part of the proof, we show that $\mathbf{v}_0$ of (3.25) and m_0 of (3.28), indeed, form a solution of (3.19) for the unknowns $\mathbf{v}_0 \in \mathbb{R}^n_c$ and $m_0 > 0$, and that the solution satisfies (3.20)–(3.23).

Part I In this part of the proof, we assume that there exist $m_0 \in \mathbb{R}$ and $\mathbf{v}_0 \in \mathbb{R}^n$ that satisfy (3.19). Then, the norms of the two sides of (3.19) are equal while, by (3.11), the norm of the right-hand side of (3.19) is m_0. Hence, the norm of the left-hand side of (3.19) equals m_0 as well, obtaining the following chain of equations, which are numbered for subsequent explanation:

$$\begin{aligned}
m_0^2 &\overbrace{=\!=}^{(1)} \left\| \sum_{k=1}^{N} m_k \begin{pmatrix} \gamma_{\mathbf{v}_k} \\ \gamma_{\mathbf{v}_k}\mathbf{v}_k \end{pmatrix} \right\|^2 \\
&\overbrace{=\!=}^{(2)} \left\| \begin{pmatrix} \sum_{k=1}^{N} m_k \gamma_{\mathbf{v}_k} \\ \sum_{k=1}^{N} m_k \gamma_{\mathbf{v}_k}\mathbf{v}_k \end{pmatrix} \right\|^2 \\
&\overbrace{=\!=}^{(3)} \left(\sum_{k=1}^{N} m_k \gamma_{\mathbf{v}_k}\right)^2 - \left(\sum_{k=1}^{N} m_k \gamma_{\mathbf{v}_k}\frac{\mathbf{v}_k}{c}\right)^2 \\
&\overbrace{=\!=}^{(4)} \left(\sum_{k=1}^{N} m_k\right)^2 + 2\sum_{\substack{j,k=1\\ j<k}}^{h} m_j m_k \gamma_{\ominus\mathbf{v}_j\oplus\mathbf{v}_k} - 1.
\end{aligned} \tag{3.29}$$

Derivation of the numbered equalities in (3.29) follows:

1. This equation follows from the result that the norm of the left-hand side of (3.19) equals the norm of the right-hand side of (3.19), the latter being m_0 by (3.11).
2. Follows from Item 1 by the common "four-vector" addition of $(n+1)$-vectors (where $n = 3$ in physical applications).
3. Follows from Item 2 by (3.10).
4. Follows from Item 3 by Identity (3.16) of Lemma 3.1.

It follows from the upper entry of (3.19), along with Assumption (3.24) that

$$m_0 > 0. \tag{3.30}$$

We thus obtained in (3.29) the desired equation, (3.28), for m_0.

Hence, if $m_0 \in \mathbb{R}$ and $\mathbf{v}_0 \in \mathbb{R}^n$ that satisfy (3.19) exist, m_0 is positive and must be given by (3.28). Clearly, $\mathbf{v}_0 \in \mathbb{R}_c^n$ since $\gamma_{\mathbf{v}_0}$ is real, as we see from (3.19).

By assumption, $\mathbf{v}_0$ satisfies (3.19). Equation (3.19) is equivalent to two equations, formed by the upper entry and by the lower entry of (3.19). Dividing the lower entry of (3.19) by its upper entry, noting that $m_0 \neq 0$ by (3.30), we obtain (3.25).

Similarly, dividing the upper entry of (3.19) by $m_0 > 0$ we obtain (3.26), and dividing the lower entry of (3.19) by $m_0 > 0$ we obtain (3.27).

Hence, if $m_0 \in \mathbb{R}$ and $\mathbf{v}_0 \in \mathbb{R}^n$ that satisfy (3.19) exist, then $m_0 > 0$, $\mathbf{v}_0 \in \mathbb{R}_c^n$, and they must be given by (3.28) and (3.25), and satisfy (3.26)–(3.27).

Part II In Part I, we have shown that if (3.19) possesses a solution for the unknowns $\mathbf{v}_0 \in \mathbb{R}^n$ and $m_0 \in \mathbb{R}$, then $\mathbf{v}_0 \in \mathbb{R}_c^n$ is given uniquely by (3.25) and $m_0 > 0$ is given uniquely by (3.28), satisfying (3.26)–(3.27). We will now show that, indeed, $\mathbf{v}_0 \in \mathbb{R}_c^n$, given by (3.25), and $m_0 > 0$, given by (3.28), form a solution of (3.19), and that the solution satisfies (3.20)–(3.23). Accordingly, in this second part of the proof we assume that $\mathbf{v}_0 \in \mathbb{R}_c^n$ and $m_0 > 0$ are given by (3.25) and (3.28), and we will prove that the pair $(m_0, \mathbf{v}_0)$ forms a solution of (3.19).

It follows from Identity (3.16) of Lemma 3.1, along with m_0 of (3.28) that

$$\left(\sum_{k=1}^{N} m_k \gamma_{\mathbf{v}_k} \frac{\mathbf{v}_k}{c}\right)^2 = \left(\sum_{k=1}^{N} m_k \gamma_{\mathbf{v}_k}\right)^2 - m_0^2. \tag{3.31}$$

Hence, by (3.25) and (3.31), we have the following chain of equations, which are numbered for subsequent explanation:

$$\begin{aligned}\frac{\mathbf{v}_0^2}{c^2} &\overbrace{=\!=}^{(1)} \frac{(\sum_{k=1}^{N} m_k \gamma_{\mathbf{v}_k} \frac{\mathbf{v}_k}{c})^2}{(\sum_{k=1}^{N} m_k \gamma_{\mathbf{v}_k})^2} \\ &\overbrace{=\!=}^{(2)} \frac{(\sum_{k=1}^{N} m_k \gamma_{\mathbf{v}_k})^2 - m_0^2}{(\sum_{k=1}^{N} m_k \gamma_{\mathbf{v}_k})^2} \\ &=\!= 1 - \frac{m_0^2}{(\sum_{k=1}^{N} m_k \gamma_{\mathbf{v}_k})^2}.\end{aligned} \tag{3.32}$$

Derivation of the numbered equalities in (3.32) follows:

1. This equation is given by Assumption (3.25).
2. Follows from Item 1 by (3.31).

It follows from (3.32) that

$$\gamma_{\mathbf{v}_0} = \frac{1}{\sqrt{1-\frac{\mathbf{v}_0^2}{c^2}}} = \frac{\sum_{k=1}^N m_k \gamma_{\mathbf{v}_k}}{m_0}, \tag{3.33}$$

thus verifying (3.26).

Following (3.26) and (3.25), we have

$$\gamma_{\mathbf{v}_0}\mathbf{v}_0 = \frac{\sum_{k=1}^N m_k \gamma_{\mathbf{v}_k}\mathbf{v}_k}{m_0}, \tag{3.34}$$

thus verifying (3.27).

Finally, (3.33) implies that m_0 and $\mathbf{v}_0$ satisfy the upper entry of (3.19) and, similarly, (3.34) implies that m_0 and $\mathbf{v}_0$ satisfy the lower entry of (3.19). Hence, the pair consisting of m_0 and $\mathbf{v}_0$ forms a solution of (3.19). We have thus shown that $\mathbf{v}_0 \in \mathbb{R}_c^n$ and $m_0 > 0$ given by (3.25) and (3.28) form a solution of (3.19), and that this solution satisfies (3.26)–(3.28).

To complete the proof it remains to show that the pair $(m_0, \mathbf{v}_0)$ satisfies (3.20)–(3.23) as well.

Let us first show that m_0 given by (3.28) is given by (3.23) as well. Indeed, following (1.3), p. 4, (1.33), p. 10, and (1.65), p. 18, we have

$$\gamma_{\ominus(\mathbf{w}\oplus\mathbf{v}_j)\oplus(\mathbf{w}\oplus\mathbf{v}_k)} = \gamma_{\ominus\mathbf{v}_j\oplus\mathbf{v}_k}, \tag{3.35}$$

implying that the right-hand sides of (3.23) and (3.28) are equal, so that m_0 is independent of $\mathbf{w} \in \mathbb{R}_c^n$, as desired. As such, m_0 is given by each of (3.23) and (3.28).

We have thus shown that the unique solution of (3.19) is formed by $\mathbf{v}_0 \in \mathbb{R}_c^n$ and $m_0 > 0$ that are given by (3.25) and (3.28), and that the solution satisfies (3.26)–(3.27). It, therefore, remains to show that the solution satisfies (3.20)–(3.22) as well.

Applying the Lorentz boost $L(\mathbf{w})$, $\mathbf{w} \in \mathbb{R}_c^n$, to each side of (3.19), we have

$$L(\mathbf{w})\sum_{k=1}^N m_k \begin{pmatrix} \gamma_{\mathbf{v}_k} \\ \gamma_{\mathbf{v}_k}\mathbf{v}_k \end{pmatrix} = L(\mathbf{w}) m_0 \begin{pmatrix} \gamma_{\mathbf{v}_0} \\ \gamma_{\mathbf{v}_0}\mathbf{v}_0 \end{pmatrix}. \tag{3.36}$$

Following the linearity of the Lorentz boost, illustrated in (3.8) and (3.9), (3.36) can be written as

$$\sum_{k=1}^N m_k \begin{pmatrix} \gamma_{\mathbf{w}\oplus\mathbf{v}_k} \\ \gamma_{\mathbf{w}\oplus\mathbf{v}_k}(\mathbf{w}\oplus\mathbf{v}_k) \end{pmatrix} = m_0 \begin{pmatrix} \gamma_{\mathbf{w}\oplus\mathbf{v}_0} \\ \gamma_{\mathbf{w}\oplus\mathbf{v}_0}(\mathbf{w}\oplus\mathbf{v}_0\ ) \end{pmatrix}. \tag{3.37}$$

Equation (3.37) is identical with (3.19) in which $\mathbf{v}_k \in \mathbb{R}_c^n$ is replaced by $\mathbf{w}\oplus\mathbf{v}_k \in \mathbb{R}_c^n$, $k = 0, 1, \ldots, N$.

But, the unique solution of (3.19) is the pair $(m_0 > 0, \mathbf{v}_0 \in \mathbb{R}_c^n)$ that satisfies (3.25)–(3.28). Hence, the unique solution of (3.37) is the pair $(m_0 > 0, \mathbf{w}\oplus\mathbf{v}_0 \in \mathbb{R}_c^n)$

that satisfies (3.20)–(3.23). Hence, the unique solution ($m_0 > 0$, $\mathbf{v}_0 \in \mathbb{R}^n_c$) of (3.19) satisfies not only (3.25)–(3.28) but, more generally, (3.20)–(3.23), and the proof is complete. □

In physical applications to particle systems, the real numbers m_k in Theorem 3.2 represent particle masses. As such, they are positive so that Assumption (3.24) is satisfied. However, anticipating applications of Theorem 3.2 to barycentric coordinates in hyperbolic geometry, in Chap. 4, we need the validity of Theorem 3.2 for real numbers m_k that need not be positive as well.

We have thus established in Theorem 3.2 the following four results concerning an isolated system S, (3.12),

$$S = S(m_k, \mathbf{v}_k, \Sigma_{\mathbf{0}}, k = 1, \ldots, N) \tag{3.38}$$

of N noninteracting material particles the kth particle of which has invariant mass $m_k > 0$ and velocity $\mathbf{v}_k \in \mathbb{R}^n_c$ relative to an inertial frame $\Sigma_{\mathbf{0}}$, $k = 1, \ldots, N$:

1. The relativistically invariant (or, rest) mass m_0 of the system S is given by

$$m_0 = \sqrt{\left(\sum_{k=1}^{N} m_k\right)^2 + 2\sum_{\substack{j,k=1\\ j<k}}^{N} m_j m_k (\gamma_{\ominus\mathbf{v}_j \oplus \mathbf{v}_k} - 1)} \tag{3.39}$$

   according to (3.23) with $\mathbf{w} = \mathbf{0}$.
2. The relativistic mass of the system S is

$$m_0 \gamma_{\mathbf{v}_0} \tag{3.40}$$

   relative to the rest frame $\Sigma_{\mathbf{0}}$, where $\mathbf{v}_0$ is the velocity of the center of momentum frame of S relative to $\Sigma_{\mathbf{0}}$, given by

$$\mathbf{v}_0 = \frac{\sum_{k=1}^{N} m_k \gamma_{\mathbf{v}_k} \mathbf{v}_k}{\sum_{k=1}^{N} m_k \gamma_{\mathbf{v}_k}} \tag{3.41}$$

   according to (3.20) with $\mathbf{w} = \mathbf{0}$.
3. Like energy and momentum, the relativistic mass is additive, that is, in particular for the system S relative to the rest frame $\Sigma_{\mathbf{0}}$, by (3.21) with $\mathbf{w} = \mathbf{0}$,

$$m_0 \gamma_{\mathbf{v}_0} = \sum_{k=1}^{N} m_k \gamma_{\mathbf{v}_k} . \tag{3.42}$$

4. The relativistic mass $m_0 \gamma_{\mathbf{v}_0}$ of a system meshes extraordinarily well with the Minkowskian four-vector formalism of special relativity. In particular, for the system S relative to the rest frame $\Sigma_{\mathbf{0}}$, we have, by (3.19),

$$\sum_{k=1}^{N} \begin{pmatrix} m_k \gamma_{\mathbf{v}_k} \\ m_k \gamma_{\mathbf{v}_k} \mathbf{v}_k \end{pmatrix} = \begin{pmatrix} m_0 \gamma_{\mathbf{v}_0} \\ m_0 \gamma_{\mathbf{v}_0} \mathbf{v}_0 \end{pmatrix}, \tag{3.43}$$

where m_0 and $\mathbf{v}_0$ are given uniquely by (3.39) and (3.41).

Thus, the relativistically invariant mass m_0 of a particle system S in (3.39) gives rise to its associated relativistic mass $m_0\gamma_{\mathbf{v}_0}$ relative to the rest frame $\Sigma_\mathbf{0}$. The latter, in turn, brings in (3.43) the concept of the relativistic mass into conformity with the Minkowskian four-vector formalism of special relativity. Moreover, we will see in Sects. 3.7 and 3.8 that the relativistically invariant mass m_0 of a particle system S provides a natural interpretation of observations in astrophysics and in particle physics.

To appreciate the power and elegance of Theorem 3.2 in relativistic mechanics in terms of novel analogies that it shares with familiar results in classical mechanics, we present below the classical counterpart, Theorem 3.3, of Theorem 3.2. The latter is obtained from the former by approaching the Newtonian limit when c tends to infinity. The resulting Theorem 3.3 is immediate, and its importance in classical mechanics is well-known.

Theorem 3.3 (Resultant Newtonian Invariant Mass Theorem) *Let* $(\mathbb{R}^n,+)$ *be a Euclidean* n*-space, and let* $m_k \in \mathbb{R}$ *and* $\mathbf{v}_k \in \mathbb{R}^n$, $k=1,2,\ldots,N$, *be* N *real numbers and* N *elements of* $\mathbb{R}^n$ *satisfying*

$$\sum_{k=1}^{N} m_k \neq 0. \tag{3.44}$$

Furthermore, let

$$\sum_{k=1}^{N} m_k \begin{pmatrix} 1 \\ \mathbf{v}_k \end{pmatrix} = m_0 \begin{pmatrix} 1 \\ \mathbf{v}_0 \end{pmatrix} \tag{3.45}$$

be an $(n+1)$*-vector equation for the two unknowns* $m_0 \in \mathbb{R}$ *and* $\mathbf{v}_0 \in \mathbb{R}^n$.

Then (3.45) *possesses a unique solution* $(m_0, \mathbf{v}_0)$, $m_0 \neq 0$, *satisfying the following equations for all* $\mathbf{w} \in \mathbb{R}^n$ *(including, in particular, the interesting special case of* $\mathbf{w}=\mathbf{0}$*):*

$$\mathbf{w}+\mathbf{v}_0 = \frac{\sum_{k=1}^{N} m_k(\mathbf{w}+\mathbf{v}_k)}{\sum_{k=1}^{N} m_k} \tag{3.46}$$

and

$$m_0 = \sum_{k=1}^{N} m_k. \tag{3.47}$$

Proof While the proof of Theorem 3.3 is trivial, our point is to present a proof that emphasizes how Theorem 3.3 is derived from Theorem 3.2. Indeed, in the limit as $c \to \infty$, the results of Theorem 3.2 tend to corresponding results of Theorem 3.3, noting that in this limit gamma factors tend to 1. Accordingly, Theorem 3.3 is a special case of Theorem 3.2 corresponding to $c=\infty$. □

In physical applications to particle systems, the real numbers m_k in Theorem 3.3 represent particle masses and, hence, they are positive. However, anticipating applications of Theorem 3.3 to barycentric coordinates in Euclidean geometry, in Chap. 4, we need the validity of Theorem 3.3 for real numbers m_k that need not be positive as well.

Identity (3.46) of Theorem 3.3 is immediate. Yet, it is geometrically important. The geometric importance of the validity of (3.46) for all $\mathbf{w} \in \mathbb{R}^n$ lies on its implication that the velocity $\mathbf{v}_0$ of the center of momentum frame of a particle system relative to a given inertial rest frame in classical mechanics is independent of the choice of the origin of the classical velocity space $\mathbb{R}^n$ with its underlying standard model of Euclidean geometry.

Unlike Identity (3.46) of Theorem 3.3, which is immediate, its counterpart in Theorem 3.2, Identity (3.20), is not immediate. Yet, in full analogy with Theorem 3.3, the validity of Identity (3.20) in Theorem 3.2 for all $\mathbf{w} \in \mathbb{R}_c^n$ is geometrically important. This geometric importance of Identity (3.20) lies on its implication that the velocity $\mathbf{v}_0$ of the center of momentum frame of a particle system relative to a given inertial rest frame in relativistic mechanics is independent of the choice of the origin of the relativistic velocity space $\mathbb{R}_c^n$ with its underlying Cartesian–Beltrami–Klein ball model of hyperbolic geometry.

3.5 Mass and Velocity of Particle Systems

In this section, we emphasize the analogies that the classical mass and velocity of a particle system share with their relativistic counterparts. Let

$$\begin{aligned} S_{newton} &= S_{newton}\big(m_k, \mathbf{v}_k \in \mathbb{R}^n, \Sigma_{\mathbf{0}}, k = 1, \ldots, N\big), \\ S_{einstein} &= S_{einstein}\big(m_k, \mathbf{v}_k \in \mathbb{R}_c^n, \Sigma_{\mathbf{0}}, k = 1, \ldots, N\big) \end{aligned} \tag{3.48}$$

be a Newtonian particle system and its corresponding Einsteinian particle system. Each of these is an isolated system of N noninteracting material particles the kth particle of which has invariant mass $m_k > 0$ and velocity $\mathbf{v}_k$ relative to a rest frame $\Sigma_{\mathbf{0}}$. These velocities are Newtonian, $\mathbf{v}_k \in \mathbb{R}^n$, for the Newtonian system S_{newton} and Einsteinian, $\mathbf{v}_k \in \mathbb{R}_c^n$, for the Einsteinian system $S_{einstein}$.

The mass m_0 in Identity (3.45) of Theorem 3.3 is the Newtonian mass of the Newtonian particle system S_{newton}. It is given by, (3.47),

$$m_0 = \sum_{k=1}^{N} m_k. \tag{3.49}$$

The velocity $\mathbf{v}_0 \in \mathbb{R}^n$ in Identity (3.45) is the Newtonian velocity of the center of momentum frame of the system S_{newton} relative to $\Sigma_{\mathbf{0}}$. It is given by (3.46) with $\mathbf{w} = \mathbf{0}$,

$$\mathbf{v}_0 = \frac{\sum_{k=1}^{N} m_k \mathbf{v}_k}{\sum_{k=1}^{N} m_k} \tag{3.50}$$

%

25.09.2019

a body having a spherical shape relative to an observer in its rest system K, Eq. [§4.1]...

f) Hyperbolic Triangle centers
Ungar.

Source: 20190926/ the Special Theory of Relativity Einstein's World in New Axiomatics/ Ungar, Abraham.. Hyperbolic.pdf

Pág. 64(79)
3.4 the Resultant Relativistically Invariant Mass Theorem
Lemma 3.1.

---

(101) 20190927

$$\pi = \frac{355}{113}$$

Source: 20171111 / Aurelio Baldor - Geometria.pdf
Pág. 196(200) → 258. El numero π.
Pág. 198(202) → Metodo Grafico para

data 27.09.2019

rectificar aproximadamente una circunferencia. (important !!!)

Fig. 221.

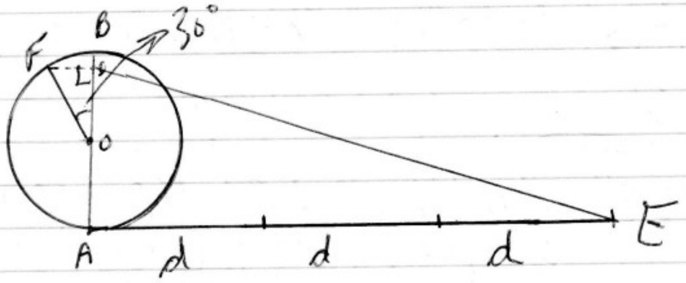

Vejamos:

1) Tracemos un diámetro $\overline{AB}$ (fig. 221);

2) Por A, tracemos la tangente;

3) Sobre dicha tangente y a partir de A tomemos tres diámetros. Se obtiene el punto E;

4) Tracemos ∠ BOF = 30°;

5) Bajemos $\overline{FD} \perp \overline{OB}$ e

6) Unamos D com E. $\overline{DE}$ es la solución del problema.

Obs: Use o Teorema de Pitágoras para calcular a hipotenusa deste triângulo.

Obs: Veja a JUSTIFICACION desta construcción.

+,- * *. * + 2 " ; /"

(&@) 3 /

(&) " ; / / G " / CC

6

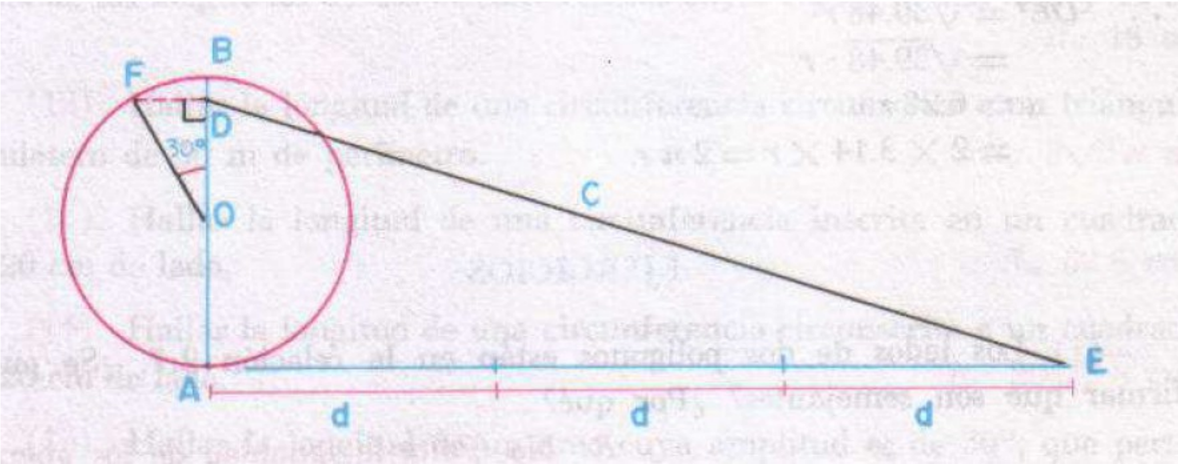
B
F
D
30°
O
C
A
E
d
d
d

" /

(&! &' $ 1 " "

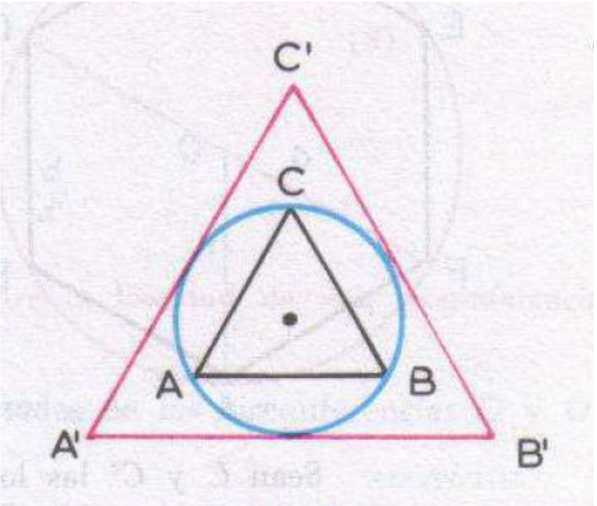

6 '

253. TEOREMA 67. "El perímetro de un polígono regular circunscrito de 2 n lados, es menor que el perímetro del polígono regular de n lados, circunscrito a la misma circunferencia.

HIPÓTESIS: $ABCD \cdots$ (Fig. 216) es un polígono regular de n lados circunscrito a la circunferencia O.

TESIS: $\overline{MN} + \overline{NO} + \overline{OP} + \overline{PQ} + \cdots \overline{AB} + \overline{BC} + \overline{CD} + \cdots.$

NOTA. **Las demostraciones de los teoremas 66 y 67, las dejamos al alumno. Recuerde que: "La menor distancia entre dos puntos, es el segmento que los une".**

254. LONGITUD DE LA CIRCUNFERENCIA.

Observemos que al duplicar el número de lados, el perímetro de un polígono regular inscrito en una circunferencia aumenta y el perímetro disminuye cuando el polígono es circunscrito.

Ejemplo. Sea el triángulo inscrito $\triangle ABC$ (Fig. 217) y el circunscrito $\triangle A'B'C'$ a la misma circunferencia.

El perímetro del triángulo inscrito P_3 es menor que el perímetro del triángulo circunscrito P'_3, es decir: $P_3 < P'_3$.
Si duplicamos el número de lados de ambos polígonos tendremos un hexágono inscrito y otro circunscrito y en ellos resultará:

$$P_6 < P'_6.$$

Nótese que P ha aumentado $(P_6 > P_3)$ y que P' ha disminuído $(P'_6 < P'_3)$. Esto, expresado matemáticamente y restando ordenadamente, sería:

$$\begin{array}{l} P'_3 > P'_6 \\ \underline{P_3 < P_6} \\ \therefore\ P'_3 = P_3 > P'_6 - P_6 \end{array}$$

Análogamente resultaría:

$$P'_6 - P_6 > P'_{12} - P_{12}$$
$$P'_{12} - P_{12} > P'_{24} - P_{24}$$
$$P'_{24} - P_{24} > P'_{48} - P_{48}, \text{ etc.}$$

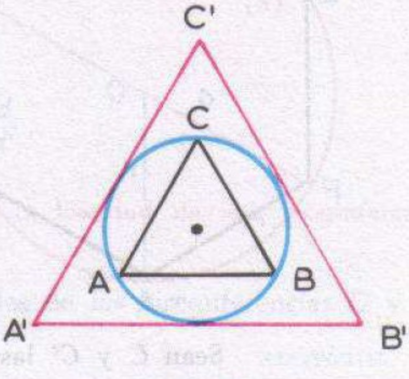

Fig 217

Es decir, a medida que se va duplicando el número de lados de los polígonos inscritos y circunscritos, la diferencia entre sus perímetros se va haciendo cada vez más pequeña, llegando a ser tan pequeña como se quiera.

Si el número de lados de estos polígonos continúa duplicándose indefinidamente, la diferencia entre ambos perímetros tiende a cero. En matemáticas se dice que ambas sucesiones P_3, P_6, P_{12} $\cdots$ y P'_3, P'_6, P'_{12} $\cdots$ de perímetros tienen un límite común el cual se llama *longitud de la circunferencia*.

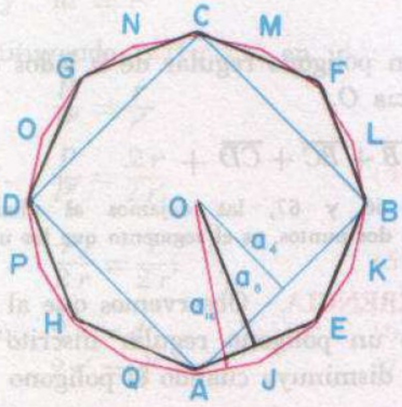

Fig. 218

255. RELACION ENTRE LA APOTEMA Y EL RADIO Observemos (Fig. 218) que a medida que se duplica el número de lados de un polígono inscrito, la apotema se va haciendo cada vez mayor y acercándose indefinidamente al valor del radio. El radio del polígono no varía y siempre es igual al radio de la circunferencia circunscrita.

256. TEOREMA 68. PROPORCIONALIDAD ENTRE LAS LONGITUDES DE CIRCUNFERENCIAS Y SUS RADIOS O DIÁMETROS. **"La razón de las longitudes de dos circunferencias cualesquiera es igual a la razón de sus radios y de sus diámetros"**

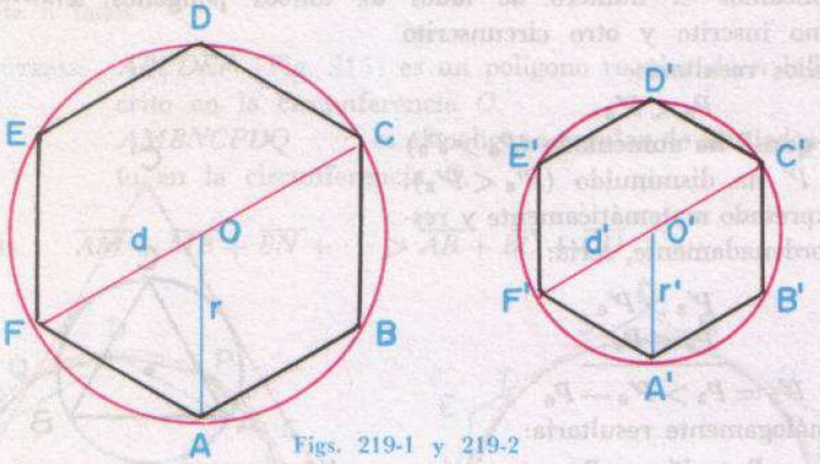

Figs. 219-1 y 219-2

HIPÓTESIS: Sean C y C' las longitudes de las circunferencias O y O' (Fig. 219) de radios r y r' y diámetros d y d'.

TESIS: $\dfrac{C}{C'} = \dfrac{r}{r'} = \dfrac{d}{d'}$.

Construcción auxiliar. Inscribamos en cada una de esas dos circunferencias un polígono regular, ambos del mismo número de lados. Sean $ABCDEF$ y $A'B'C'D'E'F'$ esos polígonos y sean P y P' sus perímetros.

DEMOSTRACIÓN:

$\frac{P}{P'} = \frac{r}{r'}$ Porque la razón de los perímetros de dos polígonos regulares del mismo número de lados es igual a la razón de sus radios (por ser los polígonos semejantes).

$\frac{\text{lím } P}{\text{lím } P'} = \frac{r}{r'}$ (1) Porque la proporcionalidad anterior se cumple cualquiera que sea el número de lados del polígono.

Pero: $\text{lím } P = C$ (2)
y $\text{lím } P' = C'$ (3) Por definición de longitud de circunferencia.

Sustituyendo (2) y (3), en (1), tenemos:

$$\frac{C}{C'} = \frac{r}{r'} \qquad (4)$$

Multiplicando por 2, los dos términos de la 2ª razón:

$$\frac{C}{C'} = \frac{2r}{2r'} \qquad (5)$$

Pero: $2r = d$ (6)
y $2r' = d'$ (7)

Sustituyendo (6) y (7) en (5), tenemos:

$$\frac{C}{C'} = \frac{d}{d'} \qquad (8)$$

Comparando (4) y (8), tenemos:

$$\frac{C}{C'} = \frac{r}{r'} = \frac{d}{d'}.$$

257. COROLARIO. "La razón entre la longitud de una circunferencia y su diámetro, es una cantidad constante"

HIPÓTESIS: Sean C y C' las longitudes de las circunferencias O y O' de diámetros d y d'.

TESIS: $\frac{C}{d} = \frac{C'}{d'}$.

DEMOSTRACIÓN:

$\frac{C}{C'} = \frac{d}{d'}$ Teorema anterior;

$\therefore \frac{C}{d} = \frac{C'}{d'}$ Intercambiando los medios

258. EL NUMERO π. El valor constante de la razón de la longitud de una circunferencia a su diámetro, se representa por la letra griega π (pí). Es decir:

$$\frac{C}{d} = \pi.$$

Este número π, es un número irracional, es decir, no se puede expresar por ningún número entero o fraccionario. Se ha calculado con muchas cifras decimales y unos cuantos valores aproximados son los siguientes:

$$\pi = \frac{22}{7}$$
$$\pi = 3.14$$
$$\pi = 3.1416$$
$$\pi = \frac{355}{113}$$
$$\pi = 3{,}1415926535 \cdots$$

Generalmente se usa el valor 3.14 o 3.1416.

259. COROLARIO. "La longitud de una circunferencia es igual al duplo de π, multiplicado por el radio".

DEMOSTRACIÓN:

$\frac{C}{d} = \pi$ Por definición;

$\therefore \quad C = \pi d \qquad (1)$ Despejando C;

y como: $d = 2r \qquad (2)$ Definición;

sustituyendo (2) en (1), tenemos:

$$C = \pi(2r) \quad \therefore \quad C = 2\pi r.$$

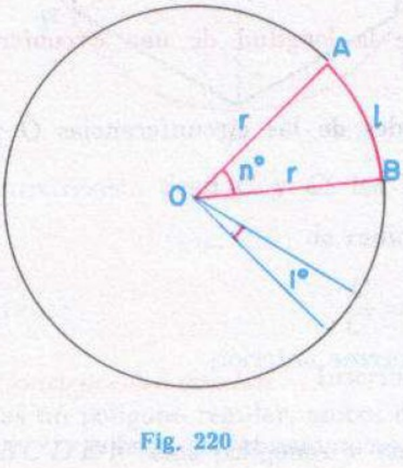

Fig. 220

260. CALCULO DE LA LONGITUD DE UNA CIRCUNFERENCIA.

Hallar la longitud de la circunferencia cuyo radio vale 6 cm.

Fórmula: $C = 2\pi r$

$$C = 2 \times 3.14 \times 6$$
$$C = 12 \times 3.14$$
$$C = 37.68 \text{ cm.}$$

261. LONGITUD DE UN ARCO DE CIRCUNFERENCIA DE $n°$. Si $C = 2\pi r$ es la longitud de la circunferencia (360°) la longitud del arco de 1° será $\frac{2\pi r}{360}$ porque 1° es $\frac{1}{360}$ de una circunferencia. (Fig. 220)

Y la longitud, l, de un arco de $n°$ será:

$$l = \frac{2\pi r n°}{360°}$$

$$\therefore \quad l = \frac{\pi r n°}{180°} \qquad \text{Simplificando.}$$

262. CALCULO DE VALORES APROXIMADOS DE π. Tomando $r = 1$, calculemos los perímetros de los hexágonos regulares inscrito y circunscrito.

INSCRITO. $r = 1,\ l_6 = r$ $\quad P_6 = 6\, l_6,$

$l_6 = 1 \quad \therefore\ P_6 = 6 \times 1 = 6.$

CIRCUNSCRITOS. $L_n = \frac{2\, r\, l_n}{\sqrt{4\, r^2 - l_n^2}}$

$$\therefore \quad L_6 = \frac{2 \times 1 \times 1}{\sqrt{4 \times 1^2 - 1^2}} = \frac{2}{\sqrt{3}} = \frac{2\sqrt{3}}{3} = 1.1547$$

$$\therefore \quad P' = 6\, L_6 = 6 \times 1.1547 = 6.9282$$

Valores aproximados de π:

$$\frac{P}{d} = \frac{6}{2} = 3 \quad \text{y} \quad \frac{P'}{d} = \frac{6.9282}{2} = 3.4641$$

Duplicando el número de lados de dichos polígonos, obtendríamos para el dodecágono:

Fórmula: $l_{2n} = \sqrt{2\, r^2 - r\sqrt{4\, r^2 - l_n^2}}$

$$l_{12} = \sqrt{2 \times 1^2 - 1\sqrt{4 \times 1^2 - 1^2}} = \sqrt{2 - \sqrt{4 - 1}}$$

$$= \sqrt{2 - \sqrt{3}} = \sqrt{2 - 1.7321} = \sqrt{0.2679}$$

$$= 0.5176.$$

$$\therefore \quad P_{12} = 12 \times 0.5176 = 6.2116.$$

Análogamente para el circunscrito, tendríamos:

$$P'_{12} = 12 \times L_{12} = 12 \times 0.5358 = 6.4307.$$

Valores aproximados de π:

$$\frac{P}{d} = \frac{6.2116}{2} = 3.1058 \quad \text{y} \quad \frac{P'}{d} = \frac{6.4307}{2} = 3.2153.$$

Continuando este proceso podríamos formar el siguiente cuadro:

Número de lados	POLIGONO INSCRITO			POLIGONO CIRCUNSCRITO		
	l_n	P	$\frac{P}{d}$	L_n	P'	$\frac{P'}{d}$
6	1	6	3	1.1547	6.9282	3.464
12	0.5176	6.2116	3.1058	0.5358	6.4307	3.2153
24	0.2610	6.2652	3.1326	0.2633	6.3193	3.1596
48	0.1308	6.2787	3.1393	0.1310	6.2921	3.1460
96	0.0654	6.2820	3.1410	0.0654	6.2854	3.1427
192	0.0327	6.2829	3.1414	0.0327	6.2837	3.1418
384	0.0163	6.2831	3.1415	0.0163	6.2833	3.1416
768	0.0081	6.2831	3.1415	0.0081	6.2832	3.1416

Las razones $\frac{P}{d}$ y $\frac{P'}{d}$ tienden al valor de π.

Comparando las tres primeras y tomando las cifras comunes, tenemos el valor $\pi = 3$.

Comparando las tres siguientes tenemos $\pi = 3.1$.

Comparando las razones correspondientes a los polígonos de 96 lados, obtenemos $\pi = 3.14$.

Continuando este proceso, podemos obtener el valor de π con la aproximación que se desee. Este método se conoce con el nombre de "*método de los perímetros*".

Como puede observarse, el método es muy laborioso.

263. METODO GRAFICO PARA RECTIFICAR APROXIMADAMENTE UNA CIRCUNFERENCIA. Rectificar una circunferencia es obtener un segmento rectilíneo cuya longitud sea igual a la de la curva.

1) Tracemos un diámetro $\overline{AB}$ (Fig. 221).

2) Por A, tracemos la tangente.

3) Sobre dicha tangente y a partir de A tomemos tres diámetros. Se obtiene el punto E.

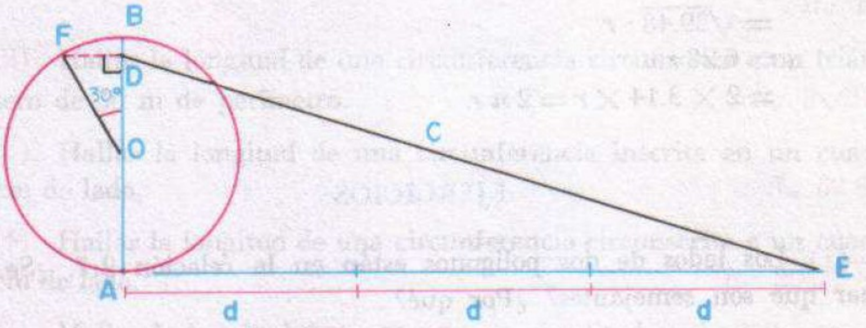

Fig 221

4) Tracemos $\angle BOF = 30°$.

5) Bajemos $\overline{FD} \perp \overline{OB}$.

6) Unamos D con E. $\overline{DE}$ es la solución del problema.

264 JUSTIFICACION DE LA CONSTRUCCION ANTERIOR.

En el $\triangle AED$:

$$(\overline{DE})^2 = (\overline{EA})^2 + (\overline{AD})^2 \qquad (1) \quad \text{Teorema de Pitágoras.}$$

Pero: $\overline{AD} = \overline{AO} + \overline{OD}$ (2) Suma de segmentos.

Sustituyendo (2) en (1), tenemos:

$$(\overline{DE})^2 = (\overline{EA})^2 + (\overline{AO} + \overline{OD})^2 \qquad (3)$$

Pero: $\overline{EA} = 3\,d = 6\,r$ (4) Construcción;

$\overline{AO} = r$ (5) Construcción.

$\overline{OD} = a_6 = \dfrac{r\sqrt{3}}{2}$ (6) Apotema de un hexágono inscrito; $\angle DOF = 30°$.

Sustituyendo (4), (5) y (6), en (3), tenemos:

$$\begin{aligned}(\overline{DE})^2 &= (6\,r)^2 + \left[r + \frac{r\sqrt{3}}{2}\right]^2 \\ &= 36\,r^2 + r^2 + r^2\sqrt{3} + \frac{3\,r^2}{4} \\ &= r^2\,(36 + 1)\,\sqrt{3} + \frac{3}{4} \qquad \text{Sacando } r^2 \text{ en factor común;} \\ &= r^2\,(36 + 1 + 1.73 + 0.75) \\ &= r^2\,(39.48) \\ &= 39.48\,r^2\end{aligned}$$

$$\begin{aligned}\therefore\quad \overline{DE}^2 &= \sqrt{39.48\,r^2}\\ &= \sqrt{39.48}\cdot r\\ &= 6.28\,r\\ &= 2\times 3.14\times r = 2\pi r.\end{aligned}$$

EJERCICIOS

(1) Los lados de dos polígonos están en la relación 2:7. ¿Se puede afirmar que son semejantes? ¿Por qué?

R.: No. Falta la igualdad de los ángulos.

(2) Dos eptágonos son equiángulos. ¿Se puede afirmar que son semejantes? ¿Por qué? *R.:* No. Falta la proporcionalidad de los lados.

(3) Dos rectángulos son semejantes. Los anchos respectivos son 16 y 24 metros y el primero tiene 30 m de largo. ¿Cuál es el largo del segundo?

R.: 45 m.

(4) Los lados de dos decágonos regulares miden 3 y 5 m. Hallar las razones de: *a*) sus lados; *b*) sus perímetros; *c*) sus radios; *d*) sus diámetros; *e*) sus apotemas. *R.:* $\frac{l}{l'}=\frac{P}{P'}=\frac{r}{r'}=\frac{d}{d'}=\frac{a}{a'}=\frac{3}{5}$.

(5) En una circunferencia de 10 m de diámetro, el lado del polígono regular de 48 lados, inscrito en la misma, mide 1.3 m. Calcular el lado de otro polígono regular del mismo número de lados, inscrito en una circunferencia de 12.5 m de radio. *R.:* 3.25 m.

(6) El perímetro de un polígono regular de 96 lados mide 31.2 m y su radio 10 m. Calcular el radio de otro polígono regular del mismo número de lados, si uno de estos lados mide 4.5 m. *R.:* 138.2 m.

(7) Hallar la longitud de una circunferencia cuyo radio mide 9 cm.

R.: 56.54 cm.

(8) Hallar la longitud de una circunferencia cuyo diámetro mide 15 cm.

R.: 47.12 cm.

(9) Hallar el radio de una circunferencia cuya longitud es 628 cm.

R.: 1 m.

(10) Hallar el diámetro de una circunferencia cuya longitud es 424 m.

R.: 134.96 m.

(11) Hallar el radio de una circunferencia cuya longitud es igual a la

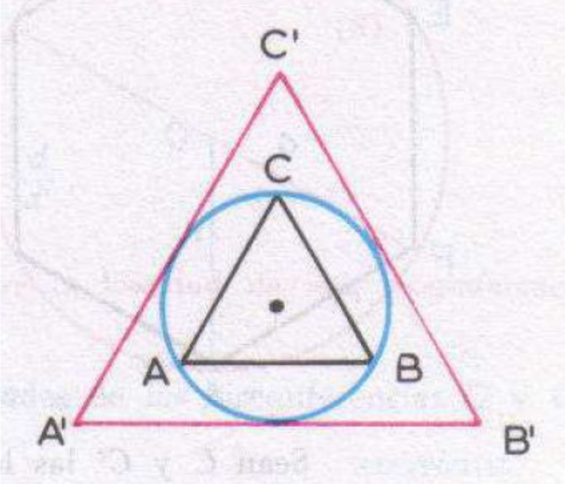

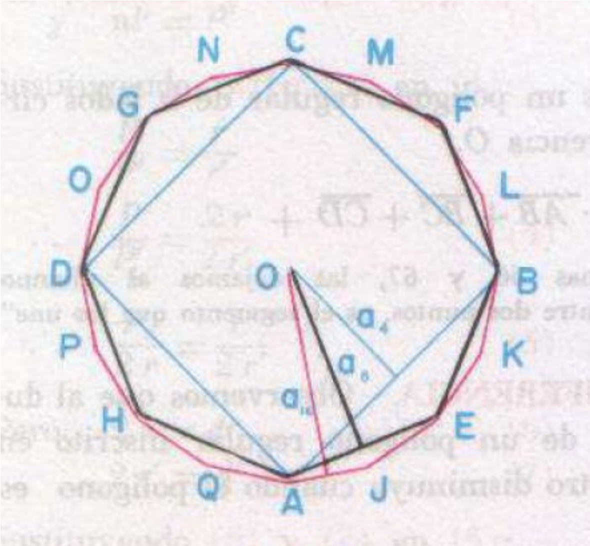
N
C
M
G
F
O
L
D
O
B
a_4
P
K
a_8
a_{16}
H
E
Q
A
J

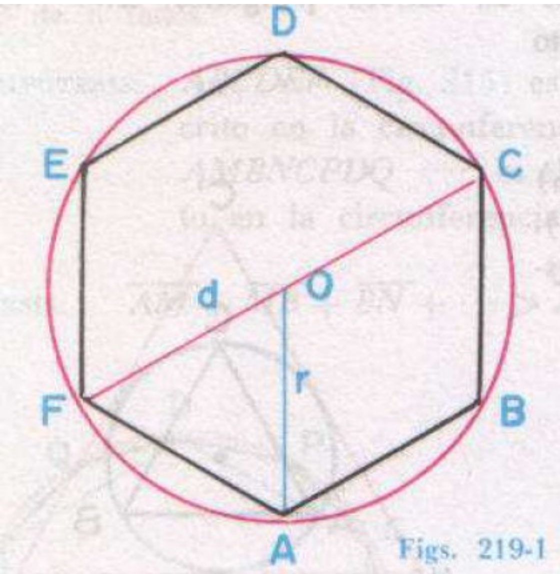

Figs. 219-1

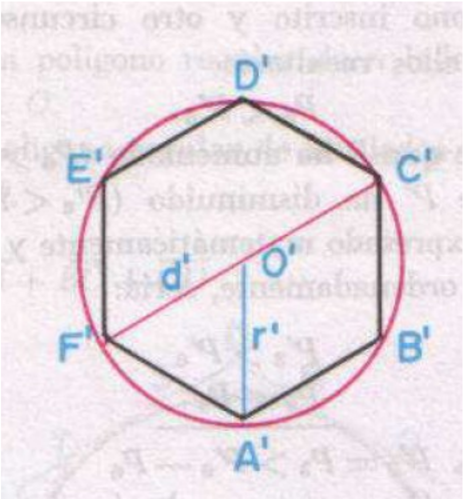

6 &

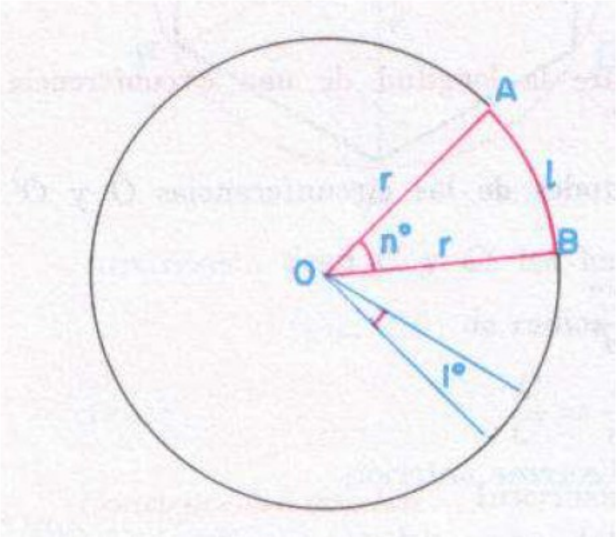

6

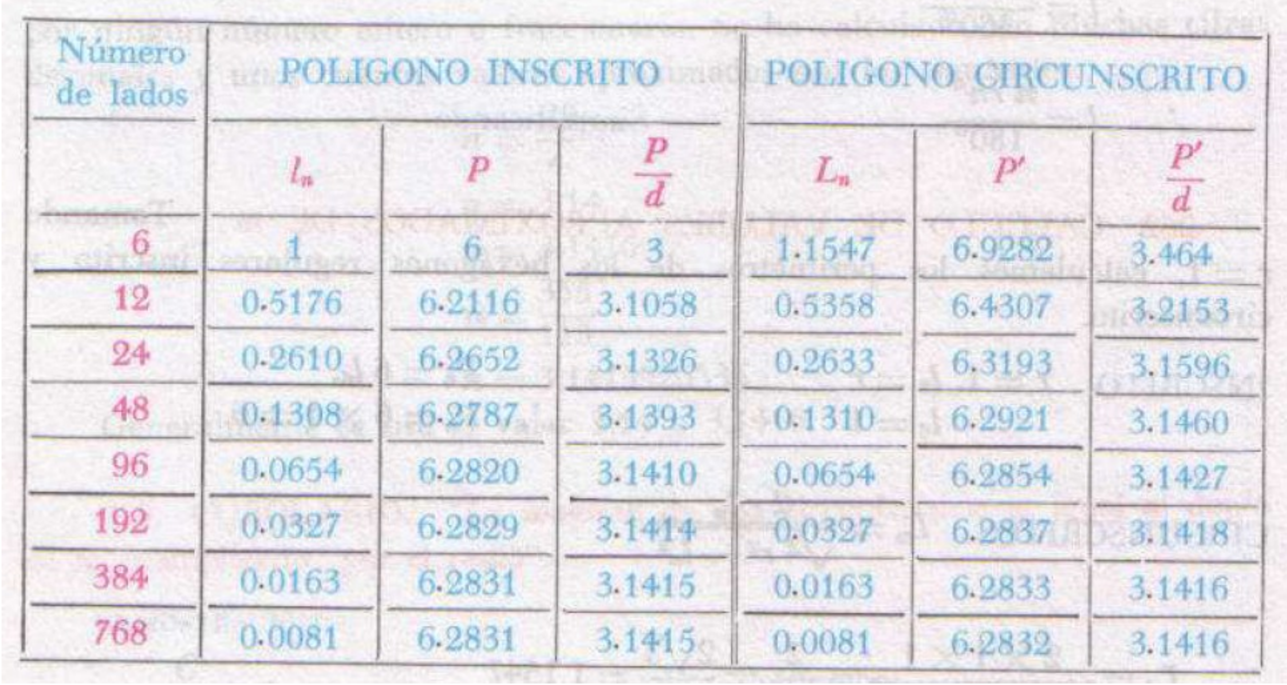

Número de lados	POLIGONO INSCRITO			POLIGONO CIRCUNSCRITO		
	l_n	P	$\frac{P}{d}$	L_n	P'	$\frac{P'}{d}$
6	1	6	3	1.1547	6.9282	3.464
12	0.5176	6.2116	3.1058	0.5358	6.4307	3.2153
24	0.2610	6.2652	3.1326	0.2633	6.3193	3.1596
48	0.1308	6.2787	3.1393	0.1310	6.2921	3.1460
96	0.0654	6.2820	3.1410	0.0654	6.2854	3.1427
192	0.0327	6.2829	3.1414	0.0327	6.2837	3.1418
384	0.0163	6.2831	3.1415	0.0163	6.2833	3.1416
768	0.0081	6.2831	3.1415	0.0081	6.2832	3.1416

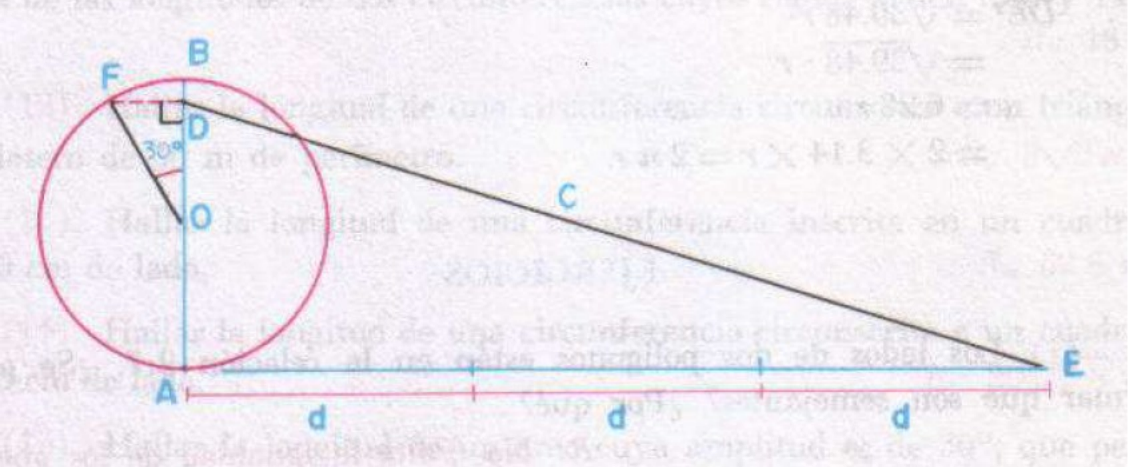

4) Tracemos $\angle BOF = 30°$.

5) Bajemos $\overline{FD} \perp \overline{OB}$.

6) Unamos D con E. $\overline{DE}$ es la solución del problema.

264 JUSTIFICACION DE LA CONSTRUCCION ANTERIOR.

En el $\triangle AED$:

$$(\overline{DE})^2 = (\overline{EA})^2 + (\overline{AD})^2 \qquad (1) \quad \text{Teorema de Pitágoras.}$$

Pero: $\overline{AD} = \overline{AO} + \overline{OD}$ (2) Suma de segmentos.

Sustituyendo (2) en (1), tenemos:

$$(\overline{DE})^2 = (\overline{EA})^2 + (\overline{AO} + \overline{OD})^2 \qquad (3)$$

Pero: $\overline{EA} = 3\,d = 6\,r$ (4) Construcción;

$\overline{AO} = r$ (5) Construcción.

$\overline{OD} = a_6 = \dfrac{r\sqrt{3}}{2}$ (6) Apotema de un hexágono inscrito; $\angle DOF = 30°$.

Sustituyendo (4), (5) y (6), en (3), tenemos:

$$\begin{aligned}
(\overline{DE})^2 &= (6\,r)^2 + \left[r + \frac{r\sqrt{3}}{2}\right]^2 \\
&= 36\,r^2 + r^2 + r^2\sqrt{3} + \frac{3\,r^2}{4} \\
&= r^2\,(36+1)\sqrt{3} + \frac{3}{4} && \text{Sacando } r^2 \text{ en factor común;} \\
&= r^2\,(36 + 1 + 1.73 + 0.75) \\
&= r^2\,(39.48) \\
&= 39.48\,r^2
\end{aligned}$$

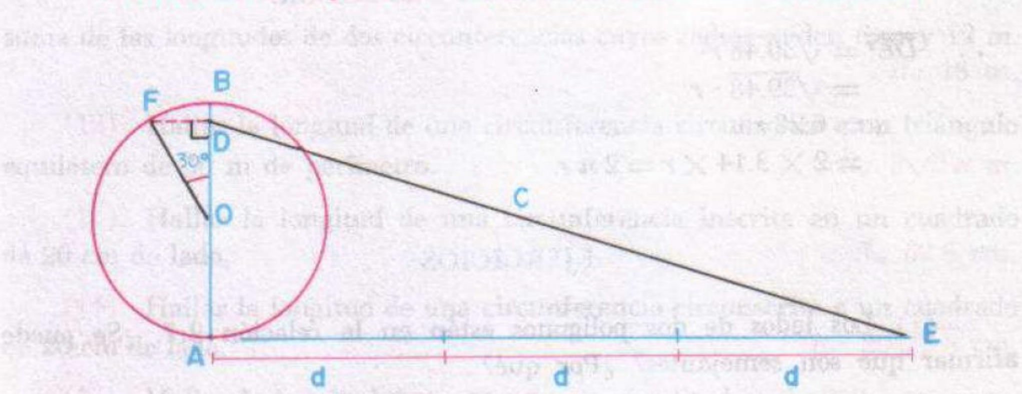

Fig. 221

4) Tracemos $\angle BOF = 30°$.

5) Bajemos $\overline{FD} \perp \overline{OB}$.

6) Unamos D con E. $\overline{DE}$ es la solución del problema.

264 JUSTIFICACION DE LA CONSTRUCCION ANTERIOR.

En el $\triangle AED$:

$(\overline{DE})^2 = (\overline{EA})^2 + (\overline{AD})^2$ (1) Teorema de Pitágoras.

Pero: $\overline{AD} = \overline{AO} + \overline{OD}$ (2) Suma de segmentos.

Sustituyendo (2) en (1), tenemos:

$(\overline{DE})^2 = (\overline{EA})^2 + (\overline{AO} + \overline{OD})^2$ (3)

Pero: $\overline{EA} = 3d = 6r$ (4) Construcción;

$\overline{AO} = r$ (5) Construcción.

$\overline{OD} = a_6 = \frac{r\sqrt{3}}{2}$ (6) Apotema de un hexágono inscrito; $\angle DOF = 30°$.

Sustituyendo (4), (5) y (6), en (3), tenemos:

$$
\begin{aligned}
(\overline{DE})^2 &= (6r)^2 + \left[r + \frac{r\sqrt{3}}{2} \right]^2 \\
&= 36r^2 + r^2 + r^2\sqrt{3} + \frac{3r^2}{4} \\
&= r^2 (36 + 1)\sqrt{3} + \frac{3}{4} \\
&= r^2 (36 + 1 + 1.73 + 0.75) \\
&= r^2 (39.48) \\
&= 39.48\, r^2
\end{aligned}
$$

Sacando r^2 en factor común;

$$\therefore \quad \overline{DE^2} = \sqrt{39.48\,r^2}$$
$$= \sqrt{39.48} \cdot r$$
$$= 6.28\,r$$
$$= 2 \times 3.14 \times r = 2\pi r.$$

EJERCICIOS

(1) Los lados de dos polígonos están en la relación 2:7. ¿Se puede afirmar que son semejantes? ¿Por qué?

R.: No. Falta la igualdad de los ángulos.

(2) Dos eptágonos son equiángulos. ¿Se puede afirmar que son semejantes? ¿Por qué? *R.:* No. Falta la proporcionalidad de los lados.

(3) Dos rectángulos son semejantes. Los anchos respectivos son 16 y 24 metros y el primero tiene 30 m de largo. ¿Cuál es el largo del segundo?

R.: 45 m.

(4) Los lados de dos decágonos regulares miden 3 y 5 m. Hallar las razones de: *a*) sus lados; *b*) sus perímetros; *c*) sus radios; *d*) sus diámetros; *e*) sus apotemas. *R.:* $\frac{l}{l'} = \frac{P}{P'} = \frac{r}{r'} = \frac{d}{d'} = \frac{a}{a'} = \frac{3}{5}$.

(5) En una circunferencia de 10 m de diámetro, el lado del polígono regular de 48 lados, inscrito en la misma, mide 1.3 m. Calcular el lado de otro polígono regular del mismo número de lados, inscrito en una circunferencia de 12.5 m de radio. *R.:* 3 25 m.

(6) El perímetro de un polígono regular de 96 lados mide 31.2 m y su radio 10 m. Calcular el radio de otro polígono regular del mismo número de lados, si uno de estos lados mide 4.5 m. *R.:* 138 2 m.

(7) Hallar la longitud de una circunferencia cuyo radio mide 9 cm.

R.: 56 54 cm.

(8) Hallar la longitud de una circunferencia cuyo diámetro mide 15 cm.

R.: 47 12 cm.

(9) Hallar el radio de una circunferencia cuya longitud es 628 cm.

R.: 1 m.

(10) Hallar el diámetro de una circunferencia cuya longitud es 424 m.

R.: 134 96 m.

(11) Hallar el radio de una circunferencia cuya longitud es igual a la

suma de las longitudes de dos circunferencias cuyos radios miden 6 m y 12 m. *R.:* 18 m.

(12) Hallar la longitud de una circunferencia circunscrita a un triángulo equilátero de 36 m de perímetro. *R.:* $8\sqrt{3}\,\pi$ m.

(13) Hallar la longitud de una circunferencia inscrita en un cuadrado de 20 cm de lado. *R.:* 62.8 cm.

(14) Hallar la longitud de una circunferencia circunscrita a un cuadrado de 20 cm de lado. *R.:* $20\sqrt{2}\,\pi$ cm.

(15) Hallar la longitud de un arco cuya amplitud es de 30°, que pertenece a una circunferencia de 10 cm de diámetro. *R.:* $\frac{5}{6}\pi$ cm.

(16) ¿Cuál es la amplitud del arco cuya longitud es 5.23 cm si pertenece a una circunferencia de 20 cm de radio? *R.:* 15°.

(17) Calcular el radio de un arco cuya amplitud es de 20°, si su longitud es de 2.79 cm. *R.:* 8 cm.

(18) Hallar la longitud de un arco de $3^\circ\,20'$ que pertenece a una circunferencia de 10 m de radio. *R.:* 0.58 m.

(19) Hallar la longitud de un arco de $5^\circ\,2'\,8''$ que pertenece a una circunferencia de 2 m de radio. *R.:* 17.5 cm.

(20) Hallar el perímetro del segmento circular limitado por el lado del triángulo equilátero inscrito en una circunferencia de 4 cm de radio. *R.:* 15.3 cm.

(21) Hallar el perímetro del segmento circular limitado por el lado del cuadrado inscrito en una circunferencia de 3 cm de radio. *R.:* 8.94 cm.

(22) Hallar el perímetro del segmento circular limitado por el lado del hexágono regular inscrito en una circunferencia de 5 cm de radio. *R.:* 10.23 cm.

(23) La longitud de un arco que pertenece a una circunferencia de 4 m de radio, es igual a la longitud de un arco que pertenece a una circunferencia de 10 m de radio. Si el primer arco es de 36°, ¿cuántos grados tiene el segundo arco? *R.:* $14^\circ\,30'$.

(24) El arco $\frown BC$ se ha trazado haciendo centro en A. El arco $\frown CD$ se ha trazado haciendo centro en B.

Si $\overline{AB} = 5$ cm, calcular la longitud de la curva BCD. *R.:* $15\,\pi$ cm.

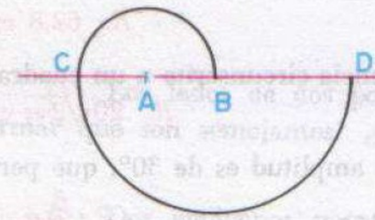

Ejer. 24

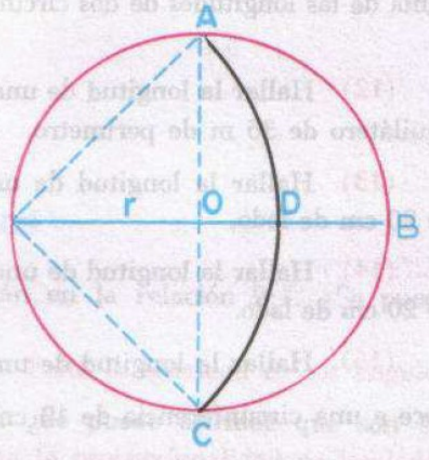

Ejer. 25

(25) Si el radio de la circunferencia *O* es *r*, ¿cuál es el perímetro de la "lúnula" *ABCD*?

R.: $P = \dfrac{2+\sqrt{2}}{2}\pi r$.

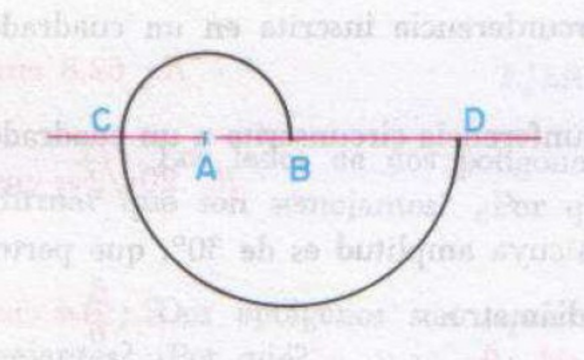

F $

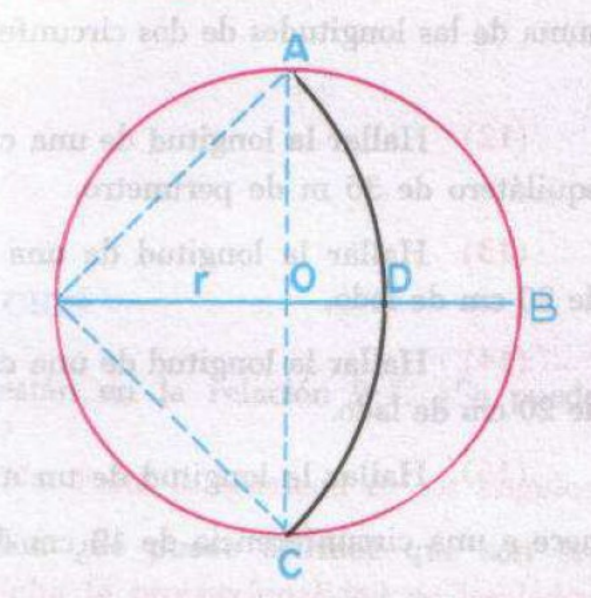

F

, " L B

+,- * *. * + 2 " ; /"

(' $

/ " A

+ M ! #' N 2

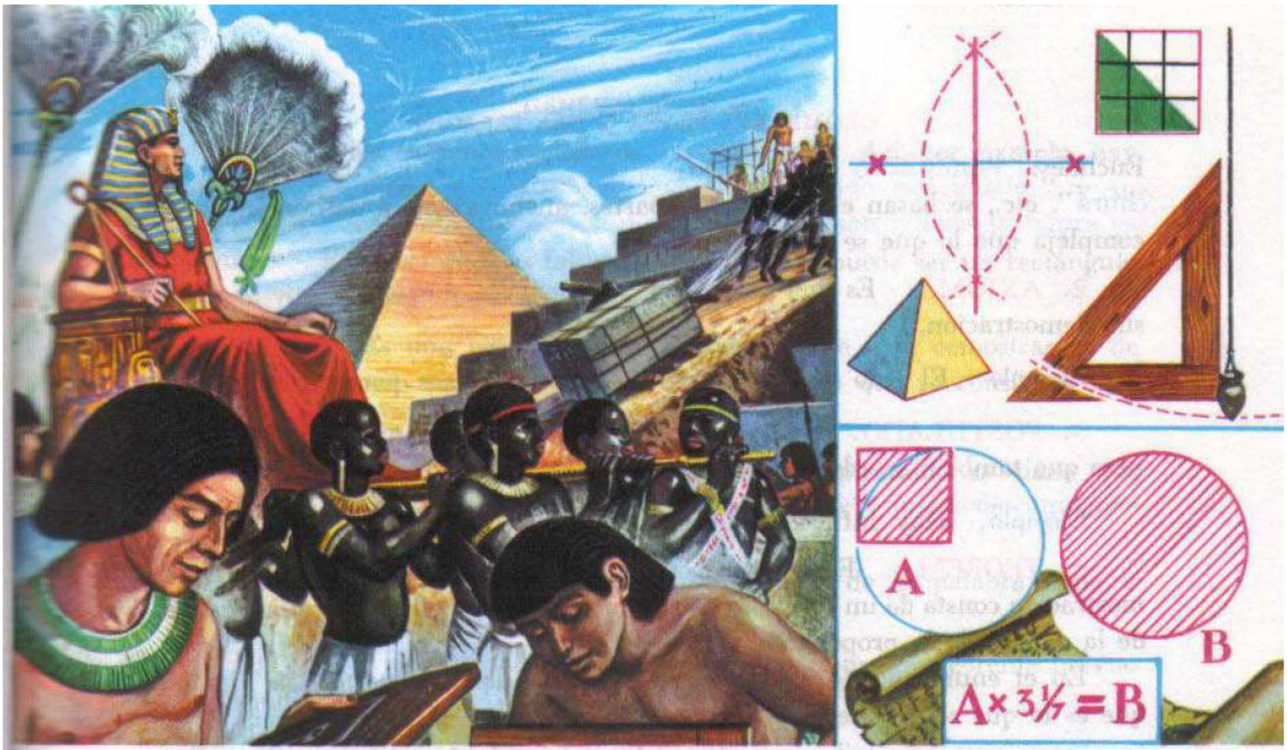

PAPIRO DE RHIND. Centuria (18-16 a. d. C.) El documento egipcio más importante que se conoce es el Papiro de Rhind, atribuido a Ahmes, quien dejó sentado que el área del círculo (B) era casi 3 1/7 veces el área del cuadrado (A) que se trazara con su radio. Al llevar a la realidad las magistrales obras de las pirámides, es evidente que conocían los egipcios cómo trazar una perpendicular a una línea. Asimismo sabían hallar el área del cuadrado y el triángulo y el uso de la plomada.

1

Generalidades

1. EL METODO DEDUCTIVO. Es el usado en la ciencia y, principalmente, en la Geometría. Este método consiste en encadenar conocimientos que se suponen verdaderos de manera tal, que se obtienen nuevos conocimientos. Es decir, obtener nuevas proposiciones como consecuencia lógica de otras anteriores.

No todas las propiedades son consecuencia de otras. Hay algunas que se aceptan como ciertas por sí mismas: son los *axiomas y postulados*.

Hay además las definiciones que son proposiciones que exponen con claridad y precisión los caracteres de una cosa. Una característica de la Geometría moderna consiste en evitar la definición de conceptos primarios que tenían poco o ningún sentido. Así, por ejemplo, las definiciones tan conocidas de

Euclides: "Punto es lo que no tiene partes" "Línea es una longitud sin anchura", etc., se basan en conceptos (partes, anchura) cuya definición es más compleja que lo que se trata de definir.

2. AXIOMA. Es una proposición tan sencilla y evidente que se admite sin demostración.

Ejemplo. El todo es mayor que cualquiera de sus partes.

3. POSTULADO. Es una proposición no tan evidente como un axioma pero que también se admite sin demostración.

Ejemplo. Hay infinitos puntos.

4. TEOREMA. Es una proposición que puede ser demostrada. La demostración consta de un conjunto de razonamientos que conducen a la evidencia de la verdad de la proposición.

En el enunciado de todo teorema se distinguen dos partes: la hipótesis, que es lo que se supone, y la tesis que es lo que se quiere demostrar.

Ejemplo. La suma de los ángulos interiores de un triángulo vale dos rectos.

HIPÓTESIS. A, B y C son los ángulos interiores de un triángulo.

TESIS. La suma de los ángulos A, B y C vale dos rectos.

En la demostración se utilizan los conocimientos adquiridos hasta aquel momento, enlazados de una manera lógica.

5. COROLARIO. Es una proposición que se deduce de un teorema como consecuencia del mismo.

Ejemplo. Del teorema: La suma de los ángulos interiores de un triángulo es igual a dos rectos, se deduce el siguiente corolario: La suma de los ángulos agudos de un triángulo rectángulo vale un recto.

6. TEOREMA RECIPROCO. Todo teorema tiene su recíproco. La hipótesis y la tesis del recíproco son, respectivamente, la tesis y la hipótesis del otro teorema que, en este caso, se llama teorema directo.

Ejemplo. El recíproco del teorema: La suma de los ángulos interiores de un triángulo vale dos rectos dice: Si la suma de los ángulos interiores de un polígono vale dos rectos el polígono es un triángulo.

La hipótesis y la tesis del recíproco son:

HIPÓTESIS. Tenemos un polígono cuyos ángulos interiores suman dos rectos.

TESIS. El polígono es un triángulo.

No siempre los teoremas recíprocos son verdaderos. Así, por ejemplo, hay un teorema que dice: "Las diagonales de un cuadrado son iguales" y su recíproco dice: "Si las diagonales de un paralelogramo son iguales la figura es un cuadrado". Este recíproco es falso porque la figura puede ser un rectángulo que también tiene sus diagonales iguales.

7. LEMA. Es una proposición que sirve de base a la demostración de un teorema. Es como un "teorema preliminar" a otro que se considera más importante.

Ejemplo. Para demostrar el volumen de una pirámide se tiene que demostrar antes el lema que dice: "Un prisma triangular se puede descomponer en tres tetraedros equivalentes".

Actualmente se ha prescindido bastante del uso de la palabra lema y se le suele llamar teorema o bien teorema preliminar.

8. ESCOLIO. Es una observación que se hace sobre un teorema previamente demostrado.

Ejemplo. Después de demostrar el teorema que dice: **En una misma circunferencia o en circunferencias iguales a mayor arco corresponde mayor cuerda** (considerando arcos menores que una semicircunferencia), se podría añadir, como escolio: "**Si no se consideran arcos menores que una semicircunferencia, a mayor arco corresponde menor cuerda**".

Actualmente apenas se usa la palabra escolio para sustituto de observación.

9. PROBLEMA. Es una proposición en la que se pide construir una figura que reuna ciertas condiciones (los problemas gráficos) o bien calcular el valor de alguna magnitud geométrica (los problemas numéricos).

Ejemplo de problema gráfico. Construir la circunferencia que pasa por tres puntos dados.

Ejemplo de problema numérico. Calcular la altura de un triángulo equilátero cuyo lado mide 6 cm.

10. EL PUNTO. Ya hemos dicho que el punto no se define. La idea de punto está sugerida por la huella que deja en el papel un lápiz bien afilado.

Un punto geométrico es imaginado tan pequeño que carece de dimensión

Admitimos el siguiente postulado:

Hay infinitos puntos

Notación: Los puntos se suelen designar por letras mayúsculas y representar por un trazo, un circulito o una cruz. Así decimos el punto A; el punto B; etc. (Fig. 1).

Fig. 1

11. LA LINEA. Son tipos especiales de conjuntos de puntos. Entre los más notables están:

La *línea recta*. **Una imagen de este conjunto de puntos es un rayo luminoso, el borde de una regla, etc.**

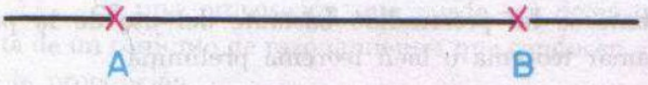

Fig. 2

Una recta geométrica se extiende sin límite en dos sentidos. No comienza ni termina. Admitimos los siguientes postulados:

Por dos puntos pasa una recta y solamente una.

Dos rectas no pueden tener más que un solo punto común.

La recta se suele designar por dos de sus puntos con el símbolo $\leftrightarrow$ encima. Así, la recta AB (Fig. 2) se representa $\overleftrightarrow{AB}$.

La línea curva. **Una imagen de línea curva es la circunferencia. Actualmente se considera que las líneas curvas pueden tener trazos rectos o no tenerlos (Fig 3).**

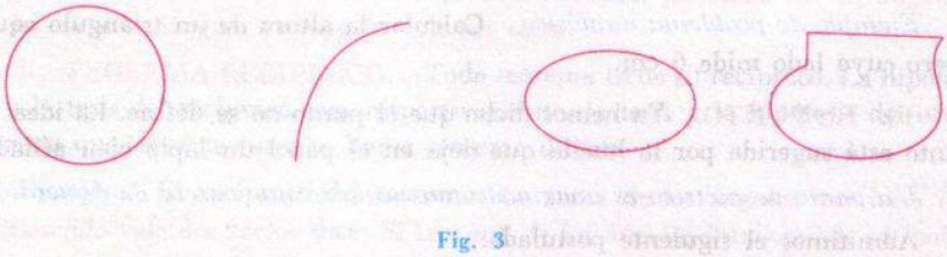

Fig. 3

Un tipo especial de curva es la línea quebrada, formada por trazos rectos. Al principio extraña llamar curva a una línea formada por trazos rectos, pero conviene acostumbrarse a esta nueva nomenclatura por la utilidad que tiene en estudios más avanzados.

Otro tipo especial de curva es la curva simple cerrada. Es la que se puede trazar de tal manera que empieza y termina en el mismo punto y éste es el único que se toca dos veces. Este tipo de curva tiene un interior y un exterior y admitimos el siguiente postulado:

Al unir un punto interior A con uno exterior B **(Fig. 4)** *de una curva simple cerrada se corta dicha curva.*

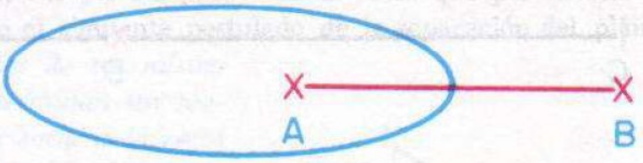

Fig. 4

Una línea tiene una sola dimensión: longitud.

12. CUERPOS FISICOS Y CUERPOS GEOMETRICOS. Son cuerpos físicos todas las cosas que nos rodean: libros, lápices, mesas, etc. Tienen forma, color, están hechos de una sustancia determinada y ocupan un lugar en el espacio.

Hay esquemas ideales de ciertos cuerpos físicos de los cuales la Geometría considera solamente su forma y su tamaño. Son los cuerpos geométricos o sólidos (Fig. 5)**. Son los prismas, conos, esferas, etc.**

Los sólidos tienen tres dimensiones: largo, ancho y alto.

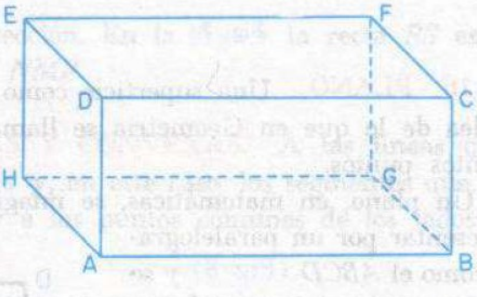

Fig. 5

13. SUPERFICIES. Son los límites que separan a los cuerpos del espacio que los rodea.

Las superficies tienen dos dimensiones: largo y ancho.

Ejemplo. ***ABCD* es una parte, llamada cara, de la superficie de la figura 5.**

14. SEMIRRECTA. Si sobre una recta señalamos un punto *A*, se llama semirrecta al conjunto de puntos formado por el *A* y todos los que le siguen

o todos los que le preceden. El punto A es el origen de la semirrecta.

Una semirrecta se suele representar por el origen y otro punto de ella, con el símbolo — encima.

Así, la semirrecta de origen C y otro punto D (Fig. 6) se representa por $\overrightarrow{CD}$.

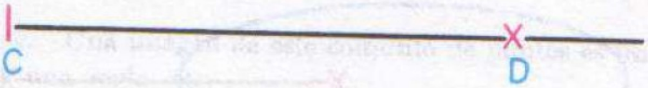

Fig. 6

15. SEGMENTO. Si sobre una recta señalamos dos puntos A y B, se llama segmento al conjunto de puntos comprendidos entre A y B más estos dos puntos que se llaman extremos del segmento. Generalmente al que se nombre en primer lugar se le llama origen y al otro, extremo.

Se admite el siguiente postulado:

La distancia más corta entre dos puntos es el segmento que los une.

Un segmento se designa por las letras de sus extremos y con un trazo encima.

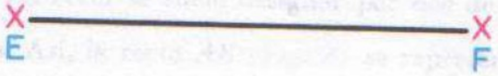

Fig. 7

Ejemplo. El segmento EF (Fig. 7) se representa así: $\overline{EF}$.

16. PLANO. Una superficie como una pared, el piso, etc., nos sugiere la idea de lo que en Geometría se llama plano. Son conjuntos parciales de infinitos puntos.

Un plano, en matemáticas, se imagina de extensión ilimitada. Se suele representar por un paralelogramo como el $ABCD$ (Fig 8) y se nombra por tres de sus puntos no alineados o por una letra griega. Así el plano de la Fig. 8 se nombra: plano ABC o bien plano α.

Dos propiedades características de los planos son las dadas por los siguientes postulados:

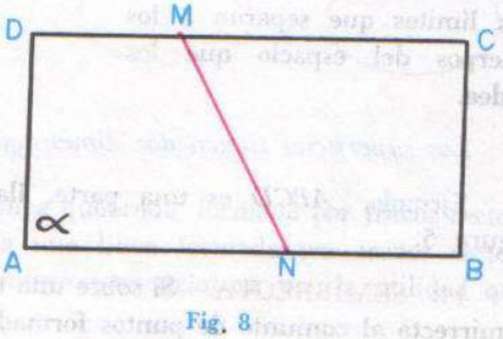

Fig. 8

Por tres puntos no alineados pasa un plano y solamente uno.

Si una recta tiene dos puntos comunes con un plano, toda la recta está contenida en el plano.

17. SEMIPLANO. Toda recta $\overleftrightarrow{MN}$ (Fig. 8) de un plano lo divide en dos regiones llamada semiplanos. Cada punto del plano pertenece a uno de los semiplanos, excepto los puntos de la recta que pertenecen a los dos.

Se admite el siguiente postulado de la separación del plano:

Dos puntos de un mismo semiplano determinan un segmento que no corta a la recta que da origen a los dos semiplanos; y dos puntos de distinto semiplano determinan un segmento que corta a la recta.

18. INTERSECCION DE PLANOS. Se admite el siguiente postulado:

Si dos planos tienen un punto común tienen una recta común.

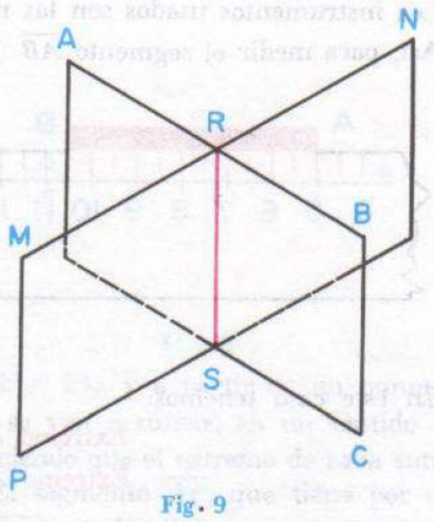

Fig. 9

En este caso se dice que los dos planos se cortan y a la recta común se le llama recta de intersección. En la Fig. 9 la recta *RS* es la intersección de los dos planos *ABC* y *NMP*.

19. POLIGONALES CONCAVAS Y CONVEXAS. A las líneas quebradas se les llama también *poligonales* y, en este caso, los segmentos que las forman reciben el nombre de *lados* y a los puntos comunes de los lados se les llama *vértices*.

Una poligonal es convexa(Fig. 10-A) si al prolongar en los dos sentidos uno cualquiera de sus lados, toda la poligonal queda en un mismo semiplano.

Se llama poligonal cóncava(Fig. 10-B) si al prolongar en los dos sentidos alguno de sus lados, parte de la poligonal queda en un semiplano y parte en el otro.

20. MEDIDA DE SEGMENTOS. Medir un segmento es compararlo con otro elegido como unidad. Para este fin se usan las unidades de longitud del sistema métrico decimal, del sistema inglés o de cualquier otro sistema.

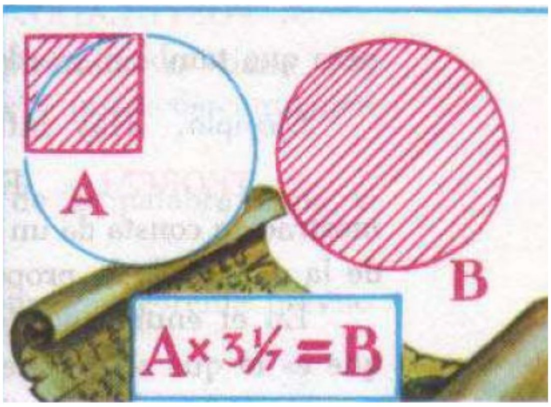

% / B

○ ○ ○ ○ ☒ ○

1021) Space Curve

Source: 20190727-03/ Volume-2.pdf

Source 2: 036_N_O_V_O_S_A_R_Q_U_I_V_O_S_201
volume-2.pdf

Introduction to Calculus
Volume II by
J.H. Heinbockel

Pág. 113(105)

Fig 7-10

Space curve revolved about
line to form surface of revolution,

.

+,- * *. * - $3 ! /"

O - 3 '@ /

A *+,-* * *. * - $3 /"

O - 3

+,- * *. * D ; P /"
.A 1 B " CCC ($ @ $ @

11.6 La razón entre la circunferencia y el diámetro de un círculo

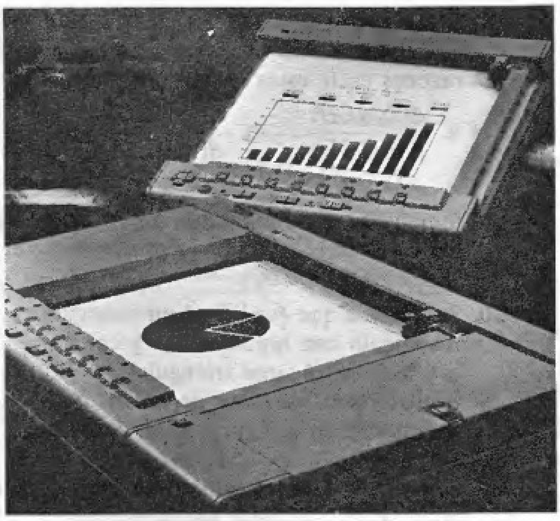

En las gráficas por computador, la impresora dibuja algo que parecen ser líneas curvas. En realidad, la pluma del graficador dibuja segmentos sucesivos de recta tan cortos que el efecto final es una línea curva. Este mismo concepto se emplea para encontrar la circunferencia de un círculo.

Estas figuras muestran una secuencia de polígonos regulares que poco a poco se aproximan a un círculo.

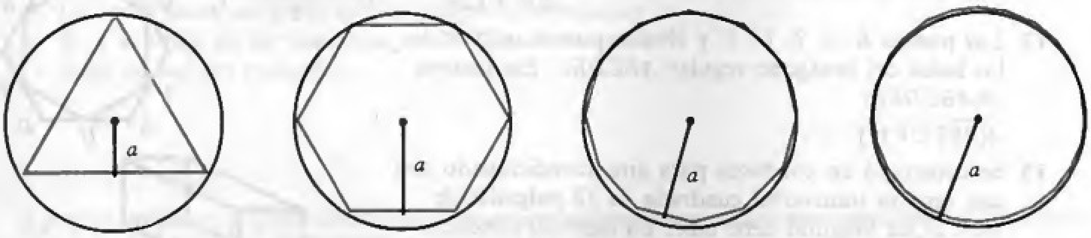

Al incrementarse el número de lados de un polígono regular, poco a poco su forma se aproxima a la de un círculo circunscrito. Además, su perímetro se aproxima a un número fijo que recibe el nombre de *circunferencia* del círculo, y la apotema se acerca al radio del círculo circunscrito.

El siguiente teorema es básico para la teoría de los círculos.

Definición 11.6

La **circunferencia de un círculo** es el número al que se aproximan los perímetros de los polígonos regulares inscritos conforme se incrementa el número de lados de los polígonos regulares.

Teorema 11.8 La razón entre la circunferencia y el diámetro es la misma para todos los círculos.

EXPLICACION DE LA PRUEBA

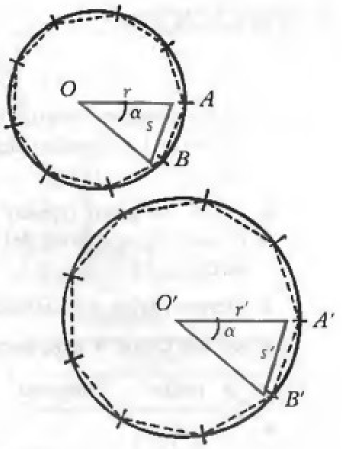

1. Selecciónense dos círculos cualesquiera e inscríbase en ellos un polígono regular de n lados.
2. Un par de triángulos isósceles semejantes $\triangle AOB$ y $\triangle A'O'B'$ están determinados por un lado de cada polígono.
3. Por tanto, las razones $\frac{s}{r}$ y $\frac{s'}{r'}$ son iguales y $\frac{ns}{r} = \frac{n's'}{r'}$.
4. Los números ns y $n's'$ igualan los perímetros p y p' de los dos polígonos regulares. Por tanto,
$$\frac{p}{r} = \frac{p'}{r'}.$$
5. Al incrementarse el número de lados n, los perímetros p y p' se aproximan a las circunferencias C y C'. Por tanto,
$$\frac{C}{r} = \frac{C'}{r'} \text{ y } \frac{C}{2r} = \frac{C'}{2r'}.$$

La razón $\frac{C}{d}$ es un número irracional, lo que significa que no puede escribirse exactamente como un decimal. Algunas aproximaciones de este número son 3.14, $3\frac{1}{7}$ y 3.14159.

Definición 11.7

La **razón** $\frac{C}{d}$, que es el mismo número real para cualquier círculo, se representa por π (la letra griega pi).

Teorema 11.9 Dado un círculo de radio r y diámetro d, la circunferencia C está dada por la fórmula $C = \pi d = 2\pi r$.

APLICACION

El animador de un espectáculo quiere diseñar un modelo para un megáfono. Este diseño es una porción de un círculo limitado por un ángulo central y su arco interceptado. Si los lados del modelo tienen 15 pulgadas y el ángulo central mide 120°, encuéntrese la longitud del arco interceptado.

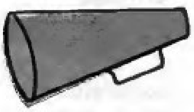

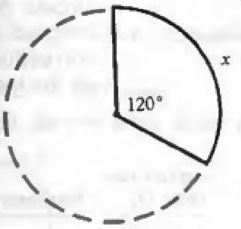

Solución. Puede establecerse una proporción entre la longitud del arco, la circunferencia del círculo, la medida del ángulo central y la medida en grados del círculo.

$$\frac{120}{360} = \frac{x}{\text{circunferencia}} = \frac{x}{2\pi 15}$$

$$\frac{1}{3} = \frac{x}{30\pi \text{ pulgadas}} \quad \text{o} \quad x = 10\pi \doteq 31.4 \text{ pulgadas}$$

($ $!

28. Si $\triangle ABC$ es un triángulo equilátero, ¿qué fracción del triángulo está sombreada?

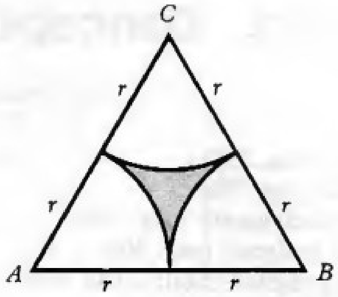

29. En la figura siguiente, el círculo pequeño es tangente a cuatro arcos circulares. ¿Qué fracción del círculo grande está sombreada?

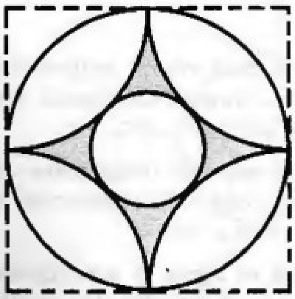

30. Los conductos de cables telefónicos están construidos para contener tres cables (todos circulares y tangentes al conducto y entre sí) de 1 cm de radio. ¿Qué fracción del conducto está ocupada por los cables?

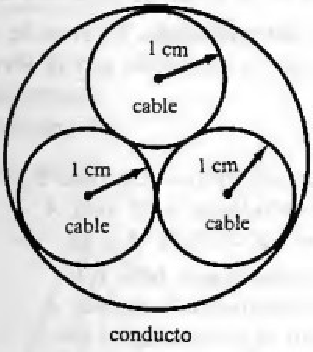

SOLUCION DE PROBLEMAS

Los puntos A', B' y C' son puntos de trisección de los lados de $\triangle ABC$. El área de la región triangular sombreada es $\underline{?}$ del área de $\triangle ABC$.
Empléese una cuadrícula pequeña o alguna técnica de medición para determinar si la región blanca debe llenarse con la fracción $\frac{1}{5}$, $\frac{1}{6}$, $\frac{1}{7}$ ó $\frac{1}{8}$.

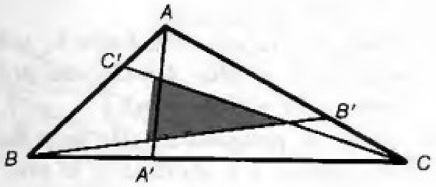

9. $12 \cdot \frac{5\sqrt{3}}{3} = 20\sqrt{3}$. **11.** $\frac{8}{3}$
13. Cuadruplicado. **15.** Igual. **17.** 13.
Solución de problemas **1.** Sí. **2.** No.

páginas 404 a 407

1. 429. **3.** 264. **5.** 425.
7. 198. **9.** $512\frac{1}{2}$
11. 8.75. **13.** 5 pies. **15.** 12.5. **17.** 30.
19. Sea h = distancia entre $\overline{AB}$ y $\overline{CD}$;
$A(AECD) = \frac{1}{2}h(AE + DC)$;
$A(EBCD) = \frac{1}{2}h(EB + DC)$; $AE = EB$.
21. $\triangle ABD$ y $\triangle BDC$ tienen la misma altura y bases iguales, $\overline{AD}$ y $\overline{CD}$.
23. $A(ABCD) = (h_1 + h_2)AB$; $A(\triangle CDX) =$
$= \frac{1}{2}h_1 AB$; $A(\triangle ABX) = \frac{1}{2}h_2 AB$;
$A(\triangle CDX + \triangle ABX) = \frac{1}{2}h_1 AB + \frac{1}{2}h_2 AB =$
$= \frac{1}{2}AB(h_1 + h_2)$.
25. $A(\triangle BAA') = A(\triangle CAA')$, tienen la misma altura y bases iguales; sea H el centroide de $\triangle ABC$; $A(\triangle BHA') = A(\triangle CHA')$;
$A(\triangle AHB) = A(\triangle AHC)$, resta;
$A(\triangle BHC') = A(\triangle AHC') = A(\triangle AHB') =$
$= A(\triangle CHB')$;
repítase el procedimiento con
$A(\triangle BAB') = A(\triangle CBB')$ como primer paso.
27. $A(ABCD) = \frac{1}{2}h(AB + DC)$; $A(\triangle ADE) =$
$= A(ABCD) + (A(\triangle DCE) + A(\triangle ABE))$;
trácense las perpendiculares de E a $\overleftrightarrow{DC}$ y de E a $\overleftrightarrow{AB}$; longitudes de estas alturas =
$= \frac{1}{2}h$; $A(\triangle ADE) = \frac{1}{2}h(AB + DC) -$
$- (\frac{1}{2} \cdot \frac{1}{2}hDC + \frac{1}{2} \cdot \frac{1}{2}hAB) = \frac{1}{4}h(AB + DC)$;
$\frac{1}{2}A(ABCD)$.
29. Usese el ejercicio 25; suma de áreas.
31. $x\sqrt{2} = 3 - 2x$; $x = \dfrac{3}{2 + \sqrt{2}}$
33. En el triángulo 30-60-90, $CG = \dfrac{\sqrt{3}}{3}$;

$$A(\triangle CFG) = \frac{1}{2} \cdot \frac{\sqrt{3}}{3} \cdot 1 = \frac{\sqrt{3}}{6}.$$

Solución de problemas
1. $\sqrt{2}$. **2.** $\sqrt{3}$. **3.** $\dfrac{\sqrt{3}}{2}$. **4.** $\sqrt{2}$. **5.** $\dfrac{\sqrt{2}}{4}$.

páginas 410 y 411

1. 102. **3.** 104. **5.** 72. **7.** $\dfrac{5\sqrt{3}}{3}, 25\sqrt{3}$.
9. 11, 484. **11.** $20k$.
13. $20\sqrt{3}$ m, $50\sqrt{3}$ m^2.
15. $20\sqrt{3}$ pies, 5 pies.
17. 2, 96; 5, 225; 10, 400; 15, 525; 20, 600; 25, 625; encerrar un cuadrado.
Solución de problemas $10 + 2\sqrt{5} + \sqrt{2}$;
$4 + 3\sqrt{5} + \sqrt{17}$;
$3\sqrt{2} + 2\sqrt{10} + 2\sqrt{5} + 2\sqrt{13}$.

páginas 414 y 415

1. 1:2. **3.** 13:20. **5.** 3:4.
7. 200, 72; $25x + 9x = 272$.
9. $\dfrac{YZ}{BC} = \dfrac{1}{2}; \left(\dfrac{1}{2}\right)^2 = \dfrac{1}{4}$.
11. 2:1. **13.** $\sqrt{288} \doteq 17$.
Solución de problemas 1:16.

páginas 418 y 419

1. Circunferencia.
3. $3\frac{1}{7}$, 3.14, 3.1416. **5.** 3, 6π.
7. 4, 8. **8.** Diámetro.
11. 36π pulgadas^2. **13.** $(32 + 4\pi)$ pies.
15. $(10 + 4\pi)$ m. **17.** 6 pies.
Solución de problemas $AD^2 + OD^2 = AO^2$;
$OD^2 = 1 - \dfrac{s^2}{4}$; $OD^2 = \dfrac{4}{4} - \dfrac{s^2}{4}$, $OD = \dfrac{\sqrt{4 - s^2}}{2}$;
$CD = 1 - \dfrac{\sqrt{4 - s^2}}{2}$; $x^2 = CD^2 + AD^2$;
$x^2 = \left(1 - \dfrac{\sqrt{4 - s^2}}{2}\right)^2 + \left(\dfrac{s}{2}\right)^2$;
$x^2 = 2 - \sqrt{4 - s^2}$.

páginas 422 a 425

1. $4\pi, \frac{121}{4}\pi, \pi^3, 3\pi$.
3. $\pi, 9\pi, \frac{3}{2}\pi, \dfrac{25}{\pi}$. **5.** 18π cm^2.

7. $\frac{25\pi}{24}$ cm^2. **9a.** 3.93 cm^2. **9b.** 5.89 m^2.
9c. 13.08 pulgadas^2. **11.** 16:25.
13. $2\sqrt{2}:\sqrt{5}$. **15.** $9 - \frac{9}{4}\pi$.
17. $9 - \frac{9}{4}\pi$. **19.** $12\pi - 9\sqrt{3}$.
21. 21.5 %; $\frac{(9 - \frac{9}{4}\pi)}{9}$.
23. $\frac{1}{2}\pi(\frac{3}{2}AB)^2 - \frac{1}{2}\pi AB^2 + \frac{1}{2}(\pi(\frac{1}{2}AB)^2) = \frac{3}{4}\pi AB^2$ sombreada; la fracción del círculo es $\frac{1}{3}$
25. Teorema de Pitágoras, $\frac{1}{2}\pi\left(\frac{a}{2}\right)^2 + \frac{1}{2}\pi\left(\frac{b}{2}\right)^2 = \frac{\pi}{8}(a^2 + b^2) = \frac{\pi}{8}(c^2) = \frac{1}{2}\pi\left(\frac{c}{2}\right)^2$.
27. $A_1 = \frac{\pi}{2}\left(\frac{AB}{2}\right)^2 - \frac{\pi}{2}\left(\frac{AC}{2}\right)^2 - \frac{\pi}{2}\left(\frac{BC}{2}\right)^2$.

$$A_1 = \frac{\pi((AC + BC)^2 - AC^2 - BC^2)}{8},$$

$$A_1 = \frac{\pi AC \cdot BC}{4};\ A_2 = \pi\left(\frac{CD}{2}\right)^2,$$

$$CD^2 = AC \cdot BC,\ A_2 = \frac{\pi AC \cdot BC}{4}.$$

29. Area del cuadrado —área de cuatro cuartos de círculo— área del círculo pequeño = área sombreada; sea $2x$ = longitud de un lado del cuadrado; dibújese un triángulo 30-60-90 con vértice en el centro y longitudes x, x, $x\sqrt{2}$; radio del círculo pequeño = $x\sqrt{2} - x$; $(2x)^2 - \pi x^2 - \pi(x\sqrt{2} - x)^2 = 4x^2 - 4\pi x^2 + 2\pi x^2\sqrt{2}$; razón del área sombreada con el área del círculo = $\frac{4}{\pi} - 4 + 2\sqrt{2}$ ó 10 %.

Solución de problemas $\frac{1}{7}$

página 427 Capítulo 11 Resumen

1a. $5\sqrt{2}$. **1b.** 40. **1c.** $5\sqrt{119}$. **1d.** 50.
2a. 3. **2b.** $\frac{27}{4}$. **3.** 2:1. **4.** $\frac{27\sqrt{3}}{2}$.
5. $\overline{AD}$.

6. Triángulo 30-60-90, $OB = 2$, $OA = 1$, $BA = \sqrt{3}$; $\pi \cdot 2^2 - \frac{1}{2} \cdot 3 \cdot 2\sqrt{3} = 4\pi - 3\sqrt{3}$.
7. 1:4. **8.** 4π.

página 429 Repaso de álgebra

1. 24 cm^2. **3.** 87.92 cm.
5. 3658.7π cm^3 ó 11 488.2 cm^3.
7. $h = 3.6$ cm.
9. $A = 58.5\pi$ pulgadas^2 o 183.69 pulgadas^2.
11. $-2, 3$. **13.** -2. **15.** $-3, -2$.
17. $10, -9$. **19.** $\frac{1}{5}, \frac{1}{3}$. **21.** 2, 3.
23. 7900 yardas^2. **25.** 93.75 cm^2.
27. 486 mm^2. **29.** 2 m.

CAPITULO 12

páginas 436 a 439

1. b. **3.** $\overline{AB}, \overline{BC}, \overline{CD}, \overline{DE}, \overline{AE}$.
5. $\triangle FBC, \triangle FAB, \triangle FAE, \triangle FDE, \triangle FDC$.
7. $\overline{EF}, \overline{HG}, \overline{IJ}, \overline{LK}, \overline{AB}, \overline{DC}$.
9. Misma.
11. $EFGH, \overline{AE}, \overline{BF}, \overline{CG}, \overline{DH}$.
13. $ADHE$. **15.** 5. **17.** 6.
19. $\overline{BH}, \overline{AG}, \overline{CE}, \overline{DF}$.
23. $15^2 + 8^2 = 289 = 17^2$. **25.** $2\sqrt{33}$.
27. $5\sqrt{3}$. **29.** 6.
31. $\overline{AF}$, $\overline{FC}$ y $\overline{AC}$ son diagonales congruentes y forman la base de un triángulo equilátero; $\overline{FB}$, $\overline{CB}$ y $\overline{AB}$ son aristas congruentes y forman aristas laterales congruentes.
33. 8; $BDEA$, $ACFB$, $BDGC$, $ACHD$, $AFHE$, $BEGF$, $CFHG$, $DEGH$.
35. Las diagonales del paralelogramo $AHGB$ se bisecan entre sí.
37. En el paralelogramo $ABGH$, O es el punto medio de $\overline{AG}$ y $\overline{BH}$ (Ejercicio 35); en el paralelogramo $ACGE$, las diagonales $\overline{AG}$ y $\overline{CE}$ se intersecan en el punto medio; el punto medio de $\overline{AG}$ es O.
39. Sea E la esquina interior; $CABE$ y $ABDE$ son paralelogramos, definición; $CE = ED = 9\sqrt{2}$, lados opuestos congruentes; sumar.

" *+,-* * *. * & F B ; D %
'E /"

+,- * *. * & F B ; D
%% 'E 2 H /"

+,- * *. * & F B ; D %
'E "FB

+,- * *. * & F B ; D
%% 'E 2 H "FB

+/ B " / ; B " " ; B D

+,- * *. * ; // 6 Q ; B DE
; ? ? 9 B / 4 D " 2 D +//
E / /"

+,- * *. * ; // 6 Q ; B DE
; ? ? 2 H 0 E D " % " / B D E
/ /"

% !

+,- * *. * ! 9 D > " H R
D E " 3" 4 D 4, /"

(!&! $ $! 9 ; D

CHAPTER 17

3-D GEOMETRY

In This Chapter

- Overview
- 3-D Vectors
- Rendering
- Casting a Ray

Overview

Thus far, most of the discussion in this book has involved 2-D problems. While adding a third dimension is mostly just a matter of adding another number, doing so has many implications. For one thing, since your computer screen has two dimensions, not three, when you enter the third dimension, you must find some way to flatten the extra dimension back to a 2-D image. Further, objects in 3-D have sides that are visible and that occlude other sides. To understand how to work with such problems, in this chapter you'll start by looking at how 3-D space can be represented.

3-D Vectors

By now you are used to representing a point in 2-D space in terms of a pair of Cartesian coordinates measured along two dimensions (*x* and *y*) from a fixed origin. Adding the third dimension (*z*, for example) involves little more than adding a third number to a list.

Adding the Third Dimension

As with 2-D geometry, with 3-D geometry you start by creating a space. Creating a 3-D space involves defining three orthogonal (mutually perpendicular) axes and an origin. How you orient the axes is arbitrary. You can point your axes in any direction you like. It is conventional, however, to consider the x-axis as "left to right," the y-axis as "down to up" and the z-axis as "front to back." If you move down two units, right one unit and forward three units, your actions correspond to the vector $(1 - 23)^T$.

Note

One of the most common orientations of a 3-D space involves a *left-hand axis*. With the left-handed axes, imagine grasping the *z*-axis with your left hand, with the thumb pointing in the positive *z* direction. Your fingers curl around from the positive *x* to the positive *y* direction.

The three-dimensional vector is represented in the same way as the two-dimensional vector. You just add one more component. Most other aspects of vector geometry stay the same. In fact, there is only one addition, which will be discussed shortly. As an example of how 3-D vectors work, consider applying the Pythagorean Theorem. You represent the magnitude of a 3-D vector $(x \quad y \quad z)^T$ as $\sqrt{x^2 + y^2 + z^2}$. The dot product is given by multiplying three instead of two components pairwise:

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} \cdot \begin{pmatrix} d \\ e \\ f \end{pmatrix} = ad + be + cf$$

Just as in 2-D space, in 3-D space you can define a *line* by using a point with position vector **p** and a direction vector **v** so that every point on the line has position **p** + t**v** for some t. You can also define a *plane* in 3-D space in much the same way. As shown in figure 17.1, you might employ a point **p** and two non-collinear vectors **v** and **w**, so that each point on the plane is **p** + t**v** + s**w** for some s and t. Another approach is to choose a point **p** and a normal vector **n**. The vector **n** is perpendicular to all the vectors in the plane. While this second method is more efficient, it is less convenient for some calculations. For this reason, in some contexts in this and subsequent chapters, the first approach will be used.

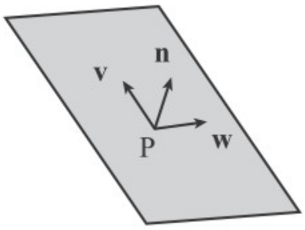

Figure 17.1
Defining a plane.

It's worth noting that the normal gives you a useful equation relating points on a plane. If the normal is $(a \quad b \quad c)^T$, then the points on the plane all conform to the equation $ax + by + cz = d$, where d is the perpendicular distance of the plane from the origin. This is the 3-D equivalent to the line equation $ay + bx = c$. In addition, note that if **n** is a unit vector and **p** is on the plane, then $d = \mathbf{p} \cdot \mathbf{n}$.

The Vector (Cross) Product

Dot products involving 3-D values work much the same way as 2-D dot products. However, when working in three dimensions, you have a new way to combine two vectors. This is known as the *vector product*, or more commonly the *cross product*. The vector or cross product is usually designated with a multiplication sign. For example, you represent the vector or cross product of the vectors **x** and **w** as $\mathbf{x} \times \mathbf{w}$. Unlike the dot product, which returns a scalar value, the cross product returns a vector. In other words given two vectors, the vector product returns a third vector that is perpendicular to the two vectors you have used in the product. The result is essentially the three-dimensional equivalent to the `normalVector()` function introduced in Chapter 5.

Calculating the vector product is a little more awkward than calculating the dot product. The formula for accomplishing this is as follows:

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} \times \begin{pmatrix} d \\ e \\ f \end{pmatrix} = \begin{pmatrix} bf - cd \\ cd - fa \\ ae - bd \end{pmatrix}$$

To remember the equation, one approach is to think of it as a determinant represented by a 3 × 3 matrix:

$$\begin{pmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a & b & c \\ d & e & f \end{pmatrix}$$

In this matrix, since **i**, **j**, **k** are the basis vectors $(1 \quad 0 \quad 0)^T$, $(0 \quad 1 \quad 0)^T$, and $(0 \quad 0 \quad 1)^T$, each element of the determinant corresponds to one component of the cross product vector. Given this understanding, you can then develop the `crossProduct()` function fairly readily:

```
function crossProduct(v1, v2)
   set x to v1[2]*v2[3]-v2[2]*v1[3]
   set y to v1[3]*v2[1]-v1[1]*v2[3]
   set z to v1[1]*v2[2]-v1[2]*v2[1]
   return vector(x,y,z)
end function
```

Here is a list of some of the most prominent features or properties of the cross product:

- The cross product is not commutative. In fact, $\mathbf{v} \times \mathbf{w} = -\mathbf{w} \times \mathbf{v}$. Nor is it associative. In general, $\mathbf{u} \times (\mathbf{v} \times \mathbf{w}) \neq (\mathbf{v} \times \mathbf{w}) \times \mathbf{w}$.
- The cross product is distributive over addition: $\mathbf{u} \times (\mathbf{v} + \mathbf{w}) = \mathbf{u} \times \mathbf{v} + \mathbf{u} \times \mathbf{w}$.
- The cross product of a vector with itself is the zero vector.
- If you take the scalar product with either of the two original vectors, if the two vectors are perpendicular, you get zero.
- If **v** and **w** are orthogonal unit vectors, the cross product is also a unit vector.
- Generally, if the angle between **v** and **w** is θ, then $|\mathbf{v} \times \mathbf{w}| = |\mathbf{v}|\,|\mathbf{w}| \sin \theta$.
- The magnitude of the cross product of two vectors is the area of the parallelogram whose sides are defined by the vectors. In other words, the area of the shape ABCD, where $\overrightarrow{AB} = \overrightarrow{DC} = \mathbf{v}$ and $\overrightarrow{BC} = \overrightarrow{AD} = \mathbf{w}$, equals $|\mathbf{v} \times \mathbf{w}|$.
- The direction of the cross product always follows the same handedness as the axes. Stated differently, if you rotate the space so that the input vectors are aligned as closely as possible to the *x*- and *y*-axes, the output vector is aligned in the positive *z*-direction. As long as your basis maintains the same handedness, the cross-product is independent of the basis.

Using the Cross Product

The cross product, like the dot product, is a useful way to *regularize* a situation. Regularization involves removing unnecessary elements. For example, suppose you have a plane for which you know the normal **n** but want to describe it instead by using two vectors on the plane. You can employ the cross product to do this. First, choose an arbitrary vector **w** that is not collinear with **n**. Next, take the cross product **w** × **n**. This gives a new vector **v** that is perpendicular to **n** (and also to **w**) and that therefore lies in the plane you are interested in. Finally, take the cross product again to get **u** = **v** × **n**. Taking the cross product gives a second vector perpendicular to **n**, which is also perpendicular to **v**.

Note

> When you work with 3-D calculations, it becomes difficult to remember the distinction between *position* and *direction* vectors. A position vector goes from the origin to a given point. Generally, a direction vector goes from a point established by a position vector to another point established by a position vector.

Another common use for the cross product is to find an *axis of rotation*. You'll look at this further later, but for now consider Figure 17.2. An arrow at **p** is pointing along the vector **v**, and you want it instead to point along the vector **w**. By finding the cross product **v** × **w**, you have a vector perpendicular to both. This serves as an axis of rotation. You can rotate the arrow around this vector to turn it in the right direction. What's more, by finding the dot product **v** × **w**, you can also calculate the angle of rotation.

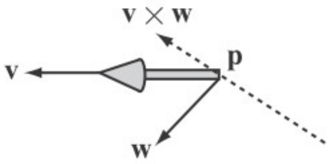

Figure 17.2

Using the vector and scalar products to calculate a rotation.

As a further exploration of the dot products, it's worth noting one more important calculation. This calculation involves the point of intersection of a line and a plane. Suppose you have a line defined by the point P and a vector **u**, and a plane defined by a point Q and a normal **n**. As illustrated by Figure 17.3, your objective is to know a value *t*, such that **p** + *t***u** lies on the plane.

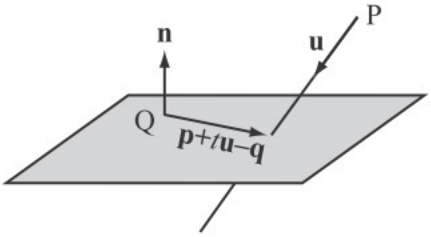

Figure 17.3

Finding the point of intersection of a line and a plane.

There are several ways to discover the value of t. One in particular keeps things simple. Notice that for any point **a** on the plane, **a** – **q** is perpendicular to **n**. In particular,

$$(\mathbf{p} + t\mathbf{u} - \mathbf{q}) \cdot \mathbf{n} = 0$$

Because the scalar product is associative, you can reason that $(\mathbf{p} - \mathbf{q}) \cdot \mathbf{n} = t\mathbf{u} \cdot \mathbf{n}$. As a result, can also reason that

$$t = \frac{(\mathbf{p} - \mathbf{q}) \cdot \mathbf{n}}{\mathbf{u} \cdot \mathbf{n}}$$

As you might expect, however, since there is no intersection between the line and plane, this approach fails when **u** and **n** are perpendicular. Otherwise, it is quick and reliable.

You can use a similar technique to find the line of intersection of two planes. Accordingly, if two planes are represented by the points **p** and **q** and the normals **n** and **m**, respectively, then as shown in Figure 17.4, your objective is to find the line that lies on both planes.

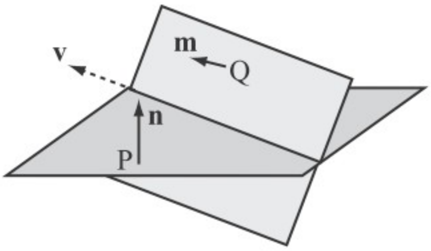

Figure 17.4

The line of intersection of two planes.

To find the line, first notice that since this line lies on both planes, it must be perpendicular to both **n** and **m**. Given that it is perpendicular, you can find its direction vector **v** as **n** × **m**. Having gone this far, you now need only to find a single point on the line. To find this point, you apply the previous result. You choose an arbitrary vector in the first plane (a good choice is **n** × **v**) and see where the line through **p** along this vector intersects the second plane. The `linePlaneIntersection()` and `planePlaneIntersection()` functions encapsulate these activities. Interactions of lines and planes are attended to by the first function:

```
function linePlaneIntersection(linePt, lineVect, planePt, planeNormal)
   set d to dotProduct(lineVect,planeNormal)
   if d=0 then return "no intersection"
   set v to linePt-planePt
   return dotProduct(v,planeNormal)/d
end function
```

The second function attends to intersections of planes:

```
function planePlaneIntersection (pt1, normal1, pt2, normal2)
   set v to crossProduct(normal1,normal2)
   set u to crossProduct(normal1, v)
   set p to linePlaneIntersection(pt1, u, pt2, normal2)
   if p="no intersection" then return p
   return array(p,v)
end function
```

Homogeneous Coordinates

Although you need only three coordinates to represent 3-D space, in practice you often use four. There are a number of reasons for this, and discussion of these reasons will be presented in Chapter 18. However, for now, consider that in addition to the *x*, *y*, and *z* coordinates, you add one more, usually designated as *w*. The result is a four-dimensional vector. This vector follows the rules that have been reviewed for two- and three-dimensional vectors.

While adding a fourth coordinate value might seem arbitrary and confusing, the w-coordinate comes in very handy in a number of instances. The w-coordinate is referred to as a *homogeneous coordinate*. The term *homogeneous* means something like "having a similar dimension." For a position vector, w is set to 1. For a direction vector, w it is 0. To see how this works out, consider the following equation:

$$\begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} - \begin{pmatrix} a \\ b \\ c \\ 1 \end{pmatrix} = \begin{pmatrix} x-a \\ y-b \\ z-c \\ 0 \end{pmatrix}$$

The vector from one position vector to another should have a zero w-component. However, at this point, it is not important to emphasize this point too strongly. As becomes clear upon further study, vector addition with homogeneous coordinates is not quite the same as vector addition using only normal coordinates.

Still, to understand the purpose of homogeneous coordinates in 3-D calculations, it's helpful to briefly draw from a discussion of two-dimensional activities. As shown by Figure 17.5, a two-dimensional plane is defined using axes x' and y'. The plane is placed in a three-dimensional space with axes x, y, and w, and it coincides with the $w = 1$ plane. As a result, any point (x', y') in the plane coincides with a point $(x, y, 1)$ in the 3-D space.

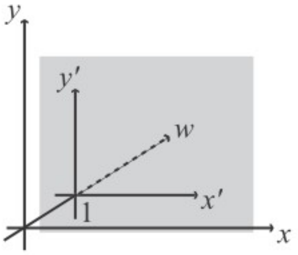

Figure 17.5

Homogeneous coordinates in 2-D.

However, it's not just points on the plane that you can map to the 2-D space. You can also compress the whole of the 3-D space onto the plane by a process of projection. As shown by Figure 17.6, for any point P, you draw a line from the 3-D origin through P and find its point of intersection with the plane.

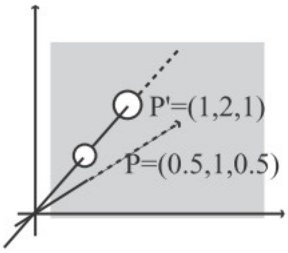

Figure 17.6

Projecting 3-D homogeneous space to the plane.

You can even do this with homogeneous points that have $w = 0$. Although the line OP is parallel to the plane, you still say that it intersects the plane "at infinity." In other words, it intersects the plane infinitely far along the line on the plane parallel to OP. In all other cases, the point on the plane that corresponds to (x, y, w) is $\left(\frac{x}{w}, \frac{y}{w}\right)$. In the strictest sense, in fact, there is no difference between the homogeneous points (x, y, w) and $\left(\frac{x}{w}, \frac{y}{w}, 1\right)$. Mathematically, they are considered to be equal. The outcome is that homogeneous coordinates are scale-invariant. If you multiply them by a constant factor, they remain the same.

As an introduction to using homogeneous coordinates, look again at the problem of finding the point of intersection of two lines in 2-D. The line through (a, b) with a vector $\begin{pmatrix} p \\ q \end{pmatrix}$ can be represented by the equations

$$a + tp = x$$

$$b + tq = y$$

which give

$$\frac{x-a}{p} = \frac{y-b}{q}$$

$$qx - py - aq + bp = 0$$

In general, a line in 2-D can be represented as $ax + by + c = 0$. Using homogeneous coordinates, this is expressed as $ax + by + cw = 0$. Comparing these two forms of expression, you can see that a line in 2-D corresponds to a plane in homogeneous coordinate space. If you represent two lines in this way, you can find their point of intersection using this approach:

$$ax + by + cw = 0$$

$$px + qy + rw = 0$$

$$\frac{x}{w} = \frac{br - cq}{aq - bp}$$

$$\frac{y}{w} = \frac{cp - ar}{aq - bp}$$

From these operations, you end up with the homogeneous coordinates ($br - cq$, $cp - ar$, $aq - bp$). Notice that if $aq - bp = 0$, then the two lines are parallel. One resulting observation is that while in standard coordinates there is no solution to the equation, in homogeneous coordinates the two lines meet at a point at infinity with $w = 0$. Further, the position vector of this point is given by the cross product

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} \times \begin{pmatrix} p \\ q \\ r \end{pmatrix}$$

The cross product makes sense when you consider that since the two vectors give the normals of the planes in 3-D space, their cross product gives the intersection of the planes. The intersection of the planes corresponds to the intersection of the 2-D lines.

Homogeneous coordinates will be addressed again in the next chapter, where attention is given to how they can be applied in 3-D. Still, it is beneficial to keep them in mind as you read the next section, where I'll discuss the projection plane for 3-D rendering. The two are closely related.

Rendering

The process of translating a 3-D scene into a 2-D picture is called *rendering*. While this is a complex process that you'll return to at a later stage, at this point, it is worthwhile to consider a few preliminaries, some of which involve the physical media through which rendering is accomplished.

The Projection Plane

With respect to projection planes, it is helpful first to consider how your eyes receive light reflected from the objects around them. Since light rays travel in a straight line, if you draw a line from your eye out into the world, what you see is the first thing in line with your eye, and your field of vision encompasses anything taken in by the lines coming from you within a particular angle from your eye.

To consider this in light of what happens in the construction of a 3-D world, imagine an observer standing still and looking straight ahead. The world visible to the observer, from the area straight ahead to the peripheral regions on the sides and top and bottom, constitutes the view *frustum*, as shown in Figure 17.7. The frustum is cut or intersected by different *viewing* or *projection planes*. Figure 17.7 shows two planes, the near plane and the far plane.

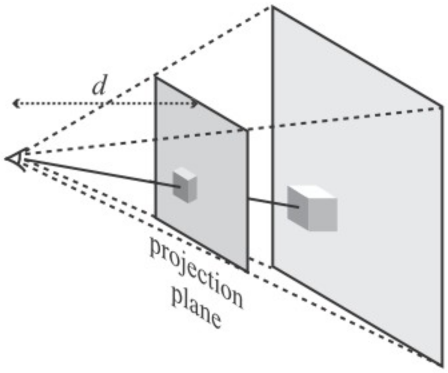

Figure 17.7
The field of view and the view frustum.

Note

> One of the difficulties in describing 3-D geometry is drawing clear diagrams on a 2-D page. Figure 17.7 shows a 3-D cube transformed into a 2-D picture on the projection plane.

The planes in the frustum are the primary devices used to translate a 3-D into a 2-D image. Referring to Figure 17.7, imagine that an observer is viewing the 3-D world as shown by the projection plane at some distance *d*. To work out what is to be found at each point of the plane, you can draw a line from the observer and see what it hits. This is called *raytracing*. For real-time animation, you can calculate where in the space each object is, draw a line to the observer, and see where this line intersects the plane. The precise situation is complicated by lights and texturing, but this is the essential principle.

As illustrated by Figure 17.8, to find the point on the projection plane that corresponds to a particular point in space, assume that the observer is facing in the direction **n**. The point you are interested in is at point **p**. Likewise, assume that the projection plane is at a perpendicular distance *d* from the observer, who is at the point **o**.

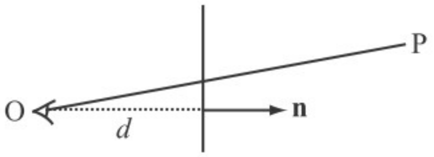

Figure 17.8

Projecting a point to the projection plane.

Figure 17.8 represents a cross-section, and in it a line is drawn from the point at P to the observer. This line passes through the projection plane at some unknown point. The situation depicted, then, is nearly the same as what is represented by Figure 17.7.

In mathematical terms, the plane normal of Figure 17.8 is the same as the direction the observer is facing. You can find a reference point on the plane drawn at the center of the screen as $\mathbf{o} + d\mathbf{n}$, assuming $\mathbf{n}$ is normalized. So the point on the projection plane corresponding to P is given by the intersection of the line starting at P in the direction $\mathbf{o} - \mathbf{p}$ with this plane, which is the point $\mathbf{p} + t(\mathbf{o} - \mathbf{p})$, where

$$t = \frac{(\mathbf{p} - (\mathbf{o} + d\mathbf{n})) \cdot \mathbf{n}}{(\mathbf{o} - \mathbf{p}) \cdot \mathbf{n}} = -d\frac{\mathbf{n} \cdot \mathbf{n}}{(\mathbf{o} - \mathbf{p}) \cdot \mathbf{n}} - 1 = \frac{d}{(\mathbf{p} - \mathbf{o}) \cdot \mathbf{n}} - 1$$

Note that since $\mathbf{n}$ is normalized, $\mathbf{n} \cdot \mathbf{n} = 1$.

Having calculated what the projection plane looks like, you can now draw its contents at the correct size on the computer screen, called a *viewport*. There are a number of different ways you can do this. One way involves specifying the scale. For example, you might determine that one unit of the projection plane corresponds to one pixel on the screen. As shown in Figure 17.9, more commonly you specify the viewport by giving one of the maximum angles in the field of view of the observer, such as the angle θ. In some instances, the whole field of view angle, given by 2θ, is specified.

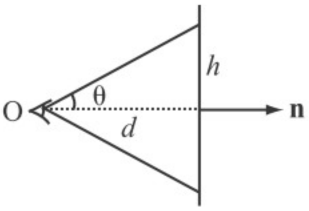

Figure 17.9

Specifying the field of view.

Since you know that $\tan\theta = \frac{h}{d}$, you also know the height of the topmost point on the projection plane. This then becomes a scale for drawing to the screen.

This approach to calculation, however, can become highly overcomplicated, since you don't need the value of d. All you actually need to know is the size of the picture on the screen and the angle of the field of view. To make the figures come out correctly, you can then set your projection screen at the appropriate place. Then you simply need to use similar triangles.

You can calculate the distance of the point from the observer along the camera vector to be $(\mathbf{p} - \mathbf{o}) \cdot \mathbf{n}$ and its vertical displacement as $(\mathbf{p} - \mathbf{o}) \cdot \mathbf{u}$, where $\mathbf{u}$ is the "up" vector of the camera. If the projection plane is to have a height of h, then you must also have that $d = \frac{h}{\tan\theta}$, so the height of the point as projected onto the screen is given by $\frac{h_{\max} \cdot (\mathbf{p} - \mathbf{o}) \cdot \mathbf{u}}{\tan\theta (\mathbf{p} - \mathbf{o}) \cdot \mathbf{n}}$. Here's a function that draws together some of the previous observations, translating a point in space to a point on the screen. As arguments, the `pos3DToScreenPos()` function takes the observer position and normal, the vertical field of view angle, the height of the screen, and the up-vector of the observer.

```
function pos3DToScreenPos(pt, observerPos, observerVect,
                          observerUp, fov, h)
   set observerRight to crossProduct(observerUp, observerVect)
   set v to pt-observerPos
   set z to dotProduct(v,observerVect)
   set d to h * tan(fov)
   set x to d*dotProduct(v,observerRight)/z
   set y to d*dotProduct(v,observerUp)/z
   return vector(x,-y)
end function
```

The `pos3DToScreenPos()` function returns the point's position on the screen relative to the center of the 3-D viewport. The position of y is measured downward. Since in practice you precalculate many such values, the function could be made more efficient. In addition, there are also some complications about points behind the observer. On the other hand, in practice you can let your 3-D card handle this part, as will become apparent momentarily.

To explore further how an image is displayed, as mentioned previously, the set of points that is visible to a particular observer is called the view frustum. As illustrated by Figure 17.7 (shown previously), the view frustum is defined in part by the base of a truncated pyramid. A pyramid is a six-sided shape. Its sides are at an angle determined by the field of view. Its front and back faces are determined arbitrarily and are sometimes referred to as the *hither* and *yon* of the camera. While you can theoretically see infinitely far and infinitely near, when rendering it is convenient to *clip* the scene at a certain distance, often using fogging to obscure the distant objects before they disappear from view.

In the equation for screen projection, notice that although changing $\mathbf{o}$ and θ both affect the image on screen, they do so in different ways. Moving the observer while leaving θ constant moves the view frustum. It changes the scale of the image linearly. By changing the field of view, however, the shape of the view frustum can be changed. This is like altering the angle of a camera lens. Choosing the correct field of view for your scene can take a little fiddling with the parameters to make something that looks right, just as a director or cinematographer may spend a long time choosing the right lens. None is more correct, but they have different visual effects. The most natural-feeling will be a viewing angle that is the same as the angle the computer user has to the computer screen—but this depends on how near they are to the screen and the resolution they are using.

Perspective

Perspective is a technique for representing objects on a flat surface. The technique in art of using perspective was mathematically defined during the Renaissance, but the basic principles of perspective had been used prior to that. As defined mathematically, perspective is the approach used to render a 3-D scene on paper. While useful for drawing, however, it must be technically translated if it is to be an effective approach to rendering images on a monitor.

To draw a scene in perspective, you start by creating a horizon. The horizon is supposed to be drawn at the height of the observer's eyes. Each point on the horizon is a vanishing point. The vanishing point represents an object infinitely far away along a line drawn from the observer's vertical axis. As shown in Figure 17.10, the classic example of this is a road stretching into the distance.

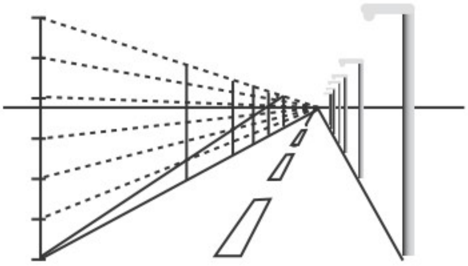

Figure 17.10

A perspective drawing.

In Figure 17.10, in addition to the road, a number of lampposts are drawn. If you were to travel along the road, the lampposts would be spaced along the road, as would be the lines on the road. The dotted lines on the left of the figure illustrate the basic technique. Two types of lines are used. Those that begin with the first post and converge at a point in the horizon are called orthogonal lines. Those that represent the lampposts or are transverse lines.

To develop the perspective, you divide the transverse line of the front post into several equal parts, joining each one to the vanishing point using an orthogonal line. You then draw in a line from the bottom of the first post to the top of the post that represents the visible background. Your remaining lampposts are placed so that they coincide with the points where the this line meets the orthogonal lines.

Mathematically, you can calculate that the image size of an object decays exponentially with distance:

$$\text{Height in image} = \text{height} \times e^{-k(d - d_0)}$$

where k is some constant and d_0 is the distance of an object whose height in the image is equal to its height in reality. There's nothing special about using e in the equation. Any base will do if you scale k accordingly.

The reason that perspective is not particularly useful in computer 3-D engines is that it requires many recalculations if your observer moves. However, if you're working with simple scenes and a fixed observer, it can be an easy way to create a quick 3-D effect. Exploring the difference between drawing and rendering involves seeing how the values k and d_0 in perspective drawing relate to d and θ in standard 3-D rendering.

Orthographic Projections

Although the standard perspective projection method described in the previous section is the most common way to draw a 3-D scene, other methods prove more useful in some circumstances, particularly in architectural and technical drawings.

Standard perspective projection is sometimes called *central projection*. With central projection, something distinctive occurs as you move farther away from the projection plane. As shown in Figure 17.11, as d increases and θ decreases, the objects on the screen always appear the same size, and the view frustum becomes a simple box.

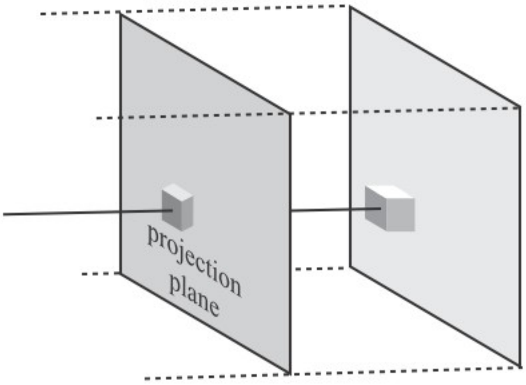

Figure 17.11

Orthographic projection.

The image shown in Figure 17.11 is called *orthographic*. When the object you are looking at is at a different angle, it can also be called *oblique*, *axonometric*, or *isometric*, but all these are equivalent to the orthographic view. Only the orientation of the object has changed. In an orthographic view, the distance of an object from the viewer or any other object has no effect on its size. It does, however, affect the position of the object on the screen and the drawing order. Otherwise, all you need to know is the size of the viewport.

In gaming, the most common of these orthographic views is the *isometric* view. This is where you look down on an object in an orthographic projection at such an angle that each side of the object is at the same angle to the projection plane, so they are all viewed as the same length. (You'll look at isometric games briefly in Part V.)

One more projection technique is worth noting. Instead of a projection plane, you use a *projection sphere* or *cylinder* around the observer. In some ways, you might imagine this to be more realistic. After all, you aren't actually looking at the world through a window but through two movable eyes. However, using a projection sphere creates some rather strange effects. One example is the Mercator projection of the earth, which is the most commonly seen map of the planet. In a Mercator projection, you imagine placing the earth inside a cylinder of paper and shining a light from the center. The shadow cast on the cylinder forms a map, which you can then unroll.

This is a good way to deal with the problem of mapping a spherical object like the earth onto a planar map, but it has some problems. Among other things, measuring distance is extremely difficult. The farther away you move from the equator, the more spread out the map is, until the North Pole is actually infinitely spread out, and infinitely high up the cylinder. On a map, this is why Greenland is so enormously out of proportion to its actual size. Navigating through a world using this kind of projection would be very weird and confusing.

Casting a Ray

As became evident in the discussion of 2-D images, it can often be useful to get a list of all objects along a *ray*. A ray is a line with one point of termination that extends infinitely. Chapter 10 discussed how to calculate the intersection of a ray with a plane. In the next couple of chapters, you'll calculate ray collisions for both simple shapes and polygonal meshes. In the current passage, it is worthwhile to look at a few applications of this method, assuming that you can employ a 3-D engine to calculate the ray for you.

As anticipated by the discussion in Chapter 10, *raycasting* is useful in many circumstances, but two are the most common. The first is in user interaction. For example, consider a situation that arises when someone clicks on the 3-D scene and must calculate what object is underneath it. The second is in collision detection. Consider a situation in which one or more rays cast in front of a shape can be used like the beam of a headlamp to determine whether there are any obstacles ahead.

Using a 3-D Engine to Find Objects Along a Path

The basic principle of raycasting is that you send a query to your real-time engine, passing it a starting point and a direction vector. The query then returns a list of models with which the ray intersects. Depending on your particular 3-D engine, the query might contain some optimization methods. In some cases, for example, you can specify a maximum number of models to return, a list of models to check against, and a maximum length for the ray. More detailed results might involve collision points and collision normals or the texture coordinates clicked on.

A good example of how this can be useful is in *terrain following*. Terrain following involves an object that is moving along a ground of varying height. Suppose your ground is a continuous triangular mesh, and you are creating a 4 × 4 vehicle that is driving along the terrain. How can you make the vehicle move realistically across the various bumps? To do so, you must know how high each wheel needs to be, orienting the vehicle accordingly.

One way to assess the heights of bumps is to store the information about the ground as a 2-D height map. This approach provides rapid information, but at the same time, it also requires high amounts of storage space. Another method—which might be quicker, depending on the speed of your 3-D engine and card and the complexity of the geometry of the terrain—is to cast a ray downward from each wheel and orient according to the collision points. This approach has the additional advantage of working equally well with a terrain that is changing over time, such as water waves.

A similar simple example is gunfire. In a first-person shooter (FPS), each player or enemy has a weapon that generally fires in a straight line. At the moment the gun is fired, a quick raycasting collision check determines where it will hit.

The technique is less useful when dealing with objects three or four times bigger than the mesh against which they are colliding. It is also not very useful with objects that are colliding with other objects that may be smaller. In such cases, you must use a large number of rays to determine whether the path is definitely clear. Certain workarounds are available, however. For example, if the objects are significantly different in size, it becomes possible to approximate the collision by using a proxy shape, such as a sphere or bounding box.

Picking, Dragging, and Dropping

Using interaction requires raycasting. The simplest example is when the player of your game clicks on the screen to select an object in 3-D space. How do you determine which object they have clicked on?

By choosing a point on the screen, your user has actually selected an angle for a ray. In fact, you're running the process of drawing a 3-D point to the screen in reverse. The simplest way to determine this point is to use the projection screen method to find a particular 3-D point under the mouse position. You then use this point to create your direction vector. To do this, as shown in previous calculations in this chapter, you need the field of view angle θ and the height h of the viewport. Then the distance to the projection plane can be found as $d = \frac{h}{\tan\theta}$.

Having found this distance, and knowing the camera's forward and up-vectors, you can quickly calculate the point under the mouse. The `screenPosTo3DPos()` function performs this calculation:

```
function screenPosTo3DPos(viewportPos, observerPos,
                          observerVect, observerUp, fov, h)
   set observerRight to crossProduct(observerUp, observerVect)
   set d to h / tan(fov)
   return observerVect*d - observerUp*viewportPos[2] +
                           observerRight*viewportPos[1]
end function
```

You now have all you need to cast a ray. The start point is the camera position, and the direction is the vector to the point you found. This tells you the model clicked on.

Suppose you need to drag the model to a new position. This question gives you a technical problem. While there are three dimensions to move it in, there are only two dimensions for the mouse. Somehow, the user's movements need to be translated into 3-D space.

How the movements are translated depends on the particular circumstances. Consider what happens if the model is part of a plane. It might be an object on the ground, a sliding tile, or a picture on the wall. If this is the case, then your two directions of motion are sufficient. You need only to find a way to constrain the motion in the particular direction.

If the model is free to move in space but not rotate, then the best solution is likely to combine the mouse movement with a key. You might program your game so that the user presses the Shift key while moving the mouse. This might cause the object to move in the camera's *xz*-plane rather than the *xy*-plane. Alternatively, while making the drag occur only in the current *xy*-plane, you can allow the user to move freely around the object, changing the plane of view or providing multiple views of the same object. Finally, if the model is supposed to be rotating in place, then mouse movements can correspond to rotations in the direction of two axes.

To accomplish such tasks, you can start with constraining to the plane. This is an extension of what you've already seen. As the mouse moves, you cast a ray from its current position to the plane you're interested in. You can use this as the new position for the dragged object, offsetting it by its height or radius as appropriate. As a note, the position does not have to be a on a plane. It can be rough terrain or even other models. Think about the pool game example, which involved a "ghost" ball demonstrating where a collision would occur.

For a more sophisticated variant, you can "fake" the ray position so that the users feel they're dragging the middle of the object instead of its base. Figure 17.12 illustrates how this might work. First, you calculate the current offset of the base of the object from its midpoint as they appear on the screen coordinates. After that, you apply this offset to the current mouse position to get the ground position you're interested in. Now you drag the object to sit correctly at that position. The user isn't precisely dragging the midpoint of the object, but it feels that way.

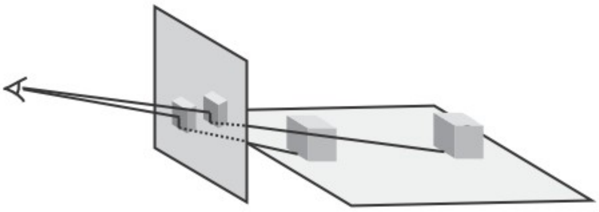

Figure 17.12
Dragging an object along the plane.

Dragging the object freely in space is an extension of this. You'll always have to constrain to a plane. However, you can give a freer choice as to what plane that is. Generally, while the most natural approach involves using a plane normal to the camera, you can also use one normal to the up-vector, or any other vector that seems convenient.

A much more difficult problem comes when you want to allow the user to rotate objects with the mouse. You can start with the simplest of these situations, which is when you want players to be able to rotate themselves, as in a mouse-driven FPS. In that case, a mouse movement corresponds to a camera rotation.

This problem is reasonably easy to solve. This is because there is a straightforward mapping from the mouse movement to the desired movement. At each time-step, you rotate the camera to point along whatever ray is currently underneath the mouse. (You'll see in the next chapter how to rotate an object to point in a particular direction.) To accomplish this, you must decide whether to keep rotating if the mouse is held still but off-center. You must also decide whether to rotate only when the mouse is moved. In the case of mouse movement, you need to reset the cursor position to the center of the screen each time you move. (With this approach, it is best to hide the cursor.)

When using the mouse to rotate an object in the camera view, things get a little trickier. There is no hard-and-fast translation from one to the other. You can make a start, however. The most natural assumption for someone rotating an object is that the point of the object clicked on will remain under the mouse as it rotates. You can do this as long as you know the intersection point of the ray. The difficulty is knowing which way should be up. As will be discussed in Chapter 18, you can specify an up-vector for a point-at operation. In this case, there's no obvious winner.

As an example of how to apply such observations, consider a case in which the object to be rotated has position vector $\mathbf{p}$, and the point on its clicked surface has position vector $\mathbf{q}$. Now the mouse is moved so that it defines a ray $\mathbf{o} + t\mathbf{v}$ from the observer. You want to know which direction to point the object so that it points toward the mouse position. In other words, you want to point the object toward the ray.

To find the correct direction, you need to find the point of intersection of the ray with the sphere centered on $\mathbf{p}$ whose radius is the magnitude of $\mathbf{q} - \mathbf{p}$. If there is such an intersection, then this tells you the new position toward which the object should point. If there is no intersection, then the mouse is outside the range of the object, in which case you can either continue to rotate or point it as closely as possible toward the mouse. Another way to put it is to say that you point it along a vector perpendicular to the vector $\mathbf{v}$ through the plane containing $\mathbf{v}$ and $\mathbf{p} - \mathbf{o}$.

But what about the up-vector? One option is to use the camera's up-vector. This works reasonably well, but it leads to problems when you want to rotate the object to align with this vector. Another approach is to use the axis of rotation, but this causes the object to rotate rather erratically. A good compromise is to use the cross-product of the camera's direction vector and the movement vector of the point.

Exercise

EXERCISE 17.1

Create a function that will draw a three-dimensional cube, culling the back faces. While you can determine which faces of a cube are visible by using the normal of the plane, this is a task that is better left for later. For the moment, concentrate on the projection. See if you can work out how to use the cross product to determine if a particular face of the cube is visible. Use the function to make a cube that spins around its center. Look at the effect of using different projection techniques and fields of view.

Summary

In this chapter, you've had a brief introduction to three-dimensional space and how it works. As you've seen, the third dimension can be understood using the concepts used to discuss 2-D objects. What happens, however, is that calculations become more complex due to the extra freedom of movement and because you have to work with an observer who can't see the whole space.

In the next chapter, you'll look further into how the space works and extend the matrix math you met in Chapter 5 to deal with the much more complex motions and transformations available to you.

You Should Now Know

- How to extend vectors into three dimensions
- The meaning and use of the *cross* (*vector*) product
- How to find points and lines of intersection between *lines* and *planes*
- The meaning of the term *homogeneous coordinates* and how to use them to calculate intersections
- How to use different kinds of *projection* to create your own 3-D engine
- How to use *raycasting* to create user interaction in 3-D space

CHAPTER 18

TRANSFORMS

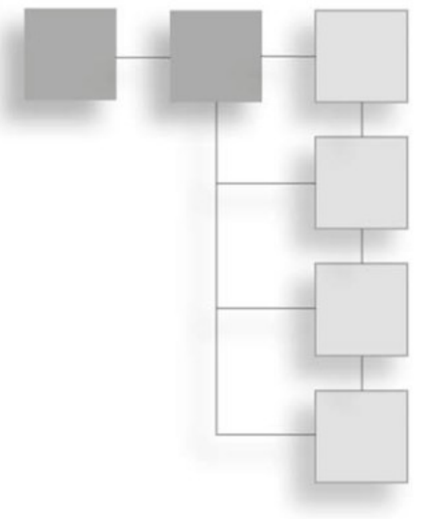

In This Chapter

- Overview
- Describing Locations in Space
- Applying Transforms

Overview

You've already seen how a 2-D matrix can be used to describe a sequence of transformations in 2-D space. In this chapter, you'll look at the 3-D equivalent of this, the *transform*. You'll also see how transforms can be used to create objects with relationships to each other.

Describing Locations in Space

The 3-D world is populated by objects called *models*. Models are sets of vertices joined together in triangles to make a *mesh*. In order to work out where these models are, the 3-D engine treats them as a single object called a *node*. The node is associated with a particular location and orientation. The 3-D engine works out the positions of the vertices relative to the node. The same method is also used to describe the location of lights, cameras, and other members of 3-D space.

Position, Rotation, and Scale

To describe a node, you must know three things: its position, its rotation, and its scale. Rotation and scale can be handled by a transformation matrix. Position is a simple vector. The simplest way to visualize this combination of transformations is as a single process acting on the "raw" model. The first action is scaling. The second is rotating about the origin. The final is translating to the correct point.

All of these transformations are *affine*, meaning that if two lines are parallel before transformation, then they remain parallel afterward. Shears and reflections are also affine transformations. In fact, any transformation that can be performed by a matrix is affine. Rotations, translations, and reflections are slightly stricter. They are rigid-body transformations because, as well as affinity, they preserve angles between lines in space. In other words, while the relative positions of vertices may change, the angles between them remain the same. As illustrated by Figure 18.1, this is not true of scales in general.

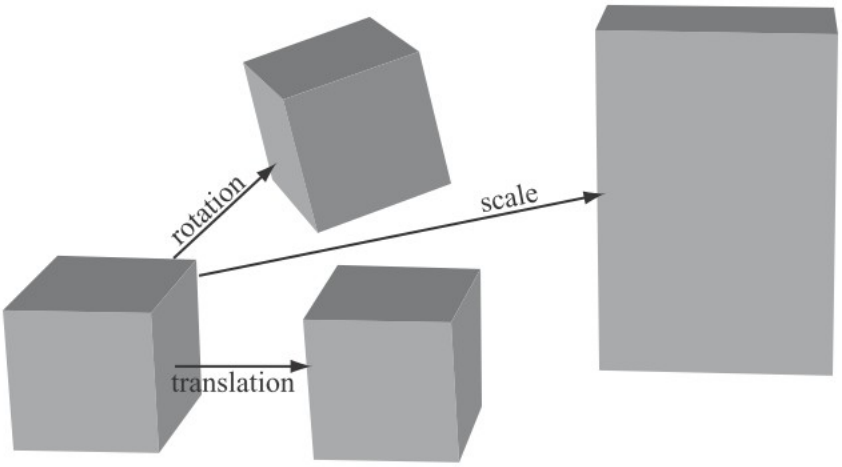

Figure 18.1
The three basic transformations.

Scale is the transformation that is easiest to visualize. Scaling an object by a constant factor involves multiplying each position vector by the scale factor. In terms of matrix mathematics, the matrix is multiplied the matrix $n\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. More generally, you can scale by different amounts in each direction. To accomplish this, you multiply the vectors by the following matrix:

$$\begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix}$$

Any scale transformation can be calculated in this way. For example, a scale factor of 2 in the direction (1 1 0) is equivalent to the scale

$$\begin{pmatrix} \sqrt{2} & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Rotation is more complicated than scaling. In two dimensions, rotation about the origin is relatively straightforward. It is described by a single angle. When a third dimension is added, however, you must specify an axis of rotation. Any vector in space can act as an axis, but as with the pool balls you looked at in previous chapters, due to topspin and sidespin, any rotation can be decomposed into simpler rotations about the basis vectors.

A rotation about one of the basis vectors can be described by a matrix like this:

$$\mathbf{R}_x = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix}$$

Combining all three such matrices gives you a generic rotation:

$$\mathbf{R} \equiv \mathbf{R}_z\mathbf{R}_y\mathbf{R}_x = \begin{pmatrix} \cos\xi & \sin\xi & 0 \\ -\sin\xi & \cos\xi & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} \cos\phi & 0 & \sin\phi \\ 0 & 1 & 0 \\ -\sin\phi & 0 & \cos\phi \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix}$$

Note

> The triple bar symbol (≡) here means "is defined as" or "is identical to." It's like a stronger kind of equals sign.

Multiplying out all these matrices is an extremely involved process, but it's not as bad as it might seem. Suppose you transform the vectors **i**, **j**, and **k** under **R**. You end up with the column vectors **u**, **v**, and **w** of the matrix. One feature of a rotation is that the vector between a point and its image must always be perpendicular to the axis of rotation. As a result, if you now take the cross product (**u** – **i**) × (**v** – **j**), you find a new vector **a** perpendicular to both. This vector **a** is the overall axis of rotation.

What's more, you can calculate the angle of rotation α by finding the components of **i** and **v** perpendicular to **a**. In other words, you find the result of projecting vectors **i** and **v** to a plane normal to **a**. Here is how this can be accomplished:

$$\begin{aligned}\mathbf{i}_N &= \mathbf{i} - (\mathbf{a} \cdot \mathbf{i})\,\mathbf{a} \\ \mathbf{v}_N &= \mathbf{v} - (\mathbf{a} \cdot \mathbf{v})\,\mathbf{a} \\ \mathbf{i}_N \cdot \mathbf{v}_\mathrm{N} &= (\mathbf{i} - (\mathbf{a} \cdot \mathbf{i})\,\mathbf{a}) \cdot (\mathbf{v} - (\mathbf{a} \cdot \mathbf{v})\,\mathbf{a}) \\ &= \mathbf{i} \cdot \mathbf{v} - (\mathbf{a} \cdot \mathbf{i})\,(\mathbf{a} \cdot \mathbf{v})\,(2 - \mathbf{a} \cdot \mathbf{a})\end{aligned}$$

Note

This only works if **i** (and therefore **v**) is not parallel to **a**. If it is parallel to **a**, you must use a different basis vector for comparison.

Assuming **a** is a unit vector (if it is not, you can normalize it), and given the definition of rotation, noting that the dot product with **i** and **v** must be equal, you arrive at this equation:

$$\cos \alpha = \mathbf{i} \cdot \mathbf{v} - (\mathbf{a} \cdot \mathbf{i})^2$$

This process can always be used to combine two or more transformations into one. Conversely, since you can decompose a rotation **R** into primitive rotations about the basis vectors, a rotation is often described by means of a rotation vector giving the angle of rotation about the three axes. You'll be coming back to this issue several times over the course of this section.

While the notions proposed in this section work fine, you still eventually hit a snag. How can you represent translations? It would be good to be able to perform a complete transformation in one step, including translations. Otherwise, among other things, any combination of transformations would require first translating back to the origin, performing rotations and scales, re-adding the position, and then translating again. Since position information is additive, however, and not multiplicative, such activity becomes seemingly impossible.

The Transformation Matrix

The solution to this problem is to use the homogeneous coordinates introduced in the previous chapter. With the aid of the fourth dimension, you can perform translations using a 4 × 4 matrix. Here is how this is done:

$$\begin{pmatrix} 1 & 0 & 0 & a \\ 0 & 1 & 0 & b \\ 0 & 0 & 1 & c \\ 0 & 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} = \begin{pmatrix} x+a \\ y+b \\ z+c \\ 1 \end{pmatrix}$$

Notice that the last row of the calculation is almost irrelevant. It serves largely as a piece of bookkeeping. Also, notice that if you perform a translation on a direction vector instead of a position vector, so that the w-coordinate is zero, the vector is unaffected by the transformation.

If you combine the resultant matrix with the scaling and rotation matrices mentioned previously, you end up with a generic transform

$$\mathbf{T} = \mathbf{P}\mathbf{R}_z\mathbf{R}_y\mathbf{R}_x\mathbf{S}$$

where **P** is a translation, and **S** is a scale that appears like this

$$\begin{pmatrix} a & 0 & 0 & 0 \\ 0 & b & 0 & 0 \\ 0 & 0 & c & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

and

$$\mathbf{R}_z = \begin{pmatrix} \cos\theta & \sin\theta & 0 & 0 \\ -\sin\theta & \cos\theta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

A similar relationship applies for $\mathbf{R}_y$ and $\mathbf{R}_x$. The non-translation matrices are created simply by replacing the 3 × 3 matrix at the top-left of the 4 × 4 matrix with the appropriate 3-D matrix.

Because its bottom row is always $(0 \quad 0 \quad 0 \quad 1)$, **T** has 12 unknown entries. However, not all such matrices are valid transforms. Those with a determinant of zero are invalid. Of the rest, exactly half would perform a reflection and most others would create some form of shear. It's easy to see that there are nine degrees of freedom in creating a valid transform. Three are rotation angles, three are scale factors, and three are translation values.

Each of the matrices that make up a transform can be inverted fairly easily. For a translation, replace each value with its negation. For a rotation, negate the rotation angle. For a scale, replace the scale factors by their reciprocals. In fact, as it happens, inverting translations and rotations is even easier than that. Their inverse is actually their transpose. In such situations, you say the matrices are *orthogonal*, which is equivalent to the rigid-body nature of the transformation. This means that you can invert the whole transform. When you do this, it is important to remember that you must apply the inverses in the opposite order:

$$\mathbf{T}^{-1} = \mathbf{S}^{-1}\mathbf{R}_x^{\mathrm{T}}\mathbf{R}_y^{\mathrm{T}}\mathbf{R}_z^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}$$

This process is straightforward providing that you keep track of the primitive transformations making up the transform. Otherwise, you have to either decompose the general transform into its component parts or invert it as it stands.

To decompose a transform into position, rotation, and scale information, you first strip off the translation element from the last column of the matrix. To calculate the scale and rotation, notice first that when you multiply **S** by the basis vectors, you multiply the basis vectors by the scale factors a, b, and c. For example, $\mathbf{RSi} = \mathbf{R}(a\mathbf{i}) = a(\mathbf{Ri})$. Also, since **R** is a rigid rotation, **Ri** is a unit vector.

A further point is that **Ri** is also the first column of **R**. Given this information, a must be the magnitude of the first column of **RS**, and the first column of **R** is the normalized version of this vector. Since the case is similar for b and c, you can split the transform into the three underlying transformations.

Further decomposing **R** into rotations about the three axes is also possible, although not always necessary. You're looking for θ, ϕ, and ξ such that

$$\begin{pmatrix} \cos\xi & \sin\xi & 0 \\ -\sin\xi & \cos\xi & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} \cos\phi & 0 & \sin\phi \\ 0 & 1 & 0 \\ -\sin\phi & 0 & \cos\phi \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{pmatrix}$$

(Since it is not involved in this transformation, you ignore the w-coordinate.)

Since the z rotation doesn't affect the z-coordinate and the x rotation doesn't affect the x-coordinate, you can see that $z_1 = -\sin\phi$. This tells you the value of ϕ. Having done that, you can see that $z_2 = -\cos\phi\ \sin\theta$, which gives you θ and $x_1 = \cos\phi\ \cos\xi$, which gives you ξ. The only problem is that there is more than one set of solutions to this system, so different sets of rotations will combine to give the right answer. This might seem trivial, but it does cause problems, particularly when interpolating transforms.

One advantage to working with the raw transformations underlying a transform is that they provide you with a validation system against rounding errors. Because of the redundancy in the transform matrix, small errors can accumulate to give a transform that is not valid if a small sheer factor is introduced. You can counteract this by checking a transform against its underlying transformations and ensuring that the two are true to one another. To some extent, this decreases the value of combining transforms into a single matrix. In this respect, it is important to remember that the main value occurs when applying a transform to multiple objects in a scene. In such cases, a validation isn't necessary. You need only to validate the transform when modifying it in some way.

Applying Transforms

Transforms simplify 3-D calculations enormously. This section provides a few examples of how this can be accomplished.

Creating Motion with Transforms

In practice, it's easiest to work with transforms by ignoring the full transform matrix and concentrating on the underlying raw transformations. In general, you'll have some node with a transform **T**, and you'll want to move it somewhere else. The only time you need to use the full transform is when calculating the actual position of the vertices of a model. For example, the basic unit cube has eight vertices at the eight permutations of the homogeneous vector $(\pm 0.5 \ \ \pm 0.5 \ \ \pm 0.5 \ \ 1)^T$. You can use a transform to move this cube to any other position or size in space. Once you have calculated the appropriate transform, you multiply it by each of the vertices to determine the new location of the cube.

The same process also works for direction vectors. As a result, the vector along one side of the cube will be unaffected. However, you do need to be careful about normal vectors. Because the scale transform does not preserve angles between lines, a direction vector that is normal to a line or plane before transforming is unlikely to still be normal afterward if the transform includes a scaling element.

If you know that **n** and **v** are perpendicular vectors, to find a vector normal to the transformed vector **Tv**, you must find some new transform $\mathbf{T}_n$, such that $\mathbf{T}_n\mathbf{n}\mathbf{T}\mathbf{v} = 0$. Recall from the definition of the dot product that since $\mathbf{n} \cdot \mathbf{v} = \mathbf{u}^T\mathbf{v}$, $(\mathbf{T}_n\mathbf{n})^T (\mathbf{T}\mathbf{v}) = 0$. Likewise, $\mathbf{n}^T\mathbf{T}_n^T\mathbf{T}\mathbf{v} = 0$. Since **n** and **v** are perpendicular by definition, if $\mathbf{T}_n^T\mathbf{T} = \mathbf{I}$, then the equation is solved. You can conclude, then, that $\mathbf{T}_n = (\mathbf{T}^{-1})^T$, the inverse transpose of **T**.

Yet again, it should be stressed that since such calculations are mostly handled behind the scenes by a decent real-time 3-D program, you don't need to worry about them. Still, it is worth keeping them in mind. They become more relevant as you work more deeply into the details of collision detection.

To consider examples of how you can use transforms to perform a common task, consider the action of aligning an arrow to point at a particular position vector. Suppose the arrow is a model whose root position (its appearance under the identity transform) is as shown in Figure 18.2. Notice that it is pointing along the positive *z*-axis.

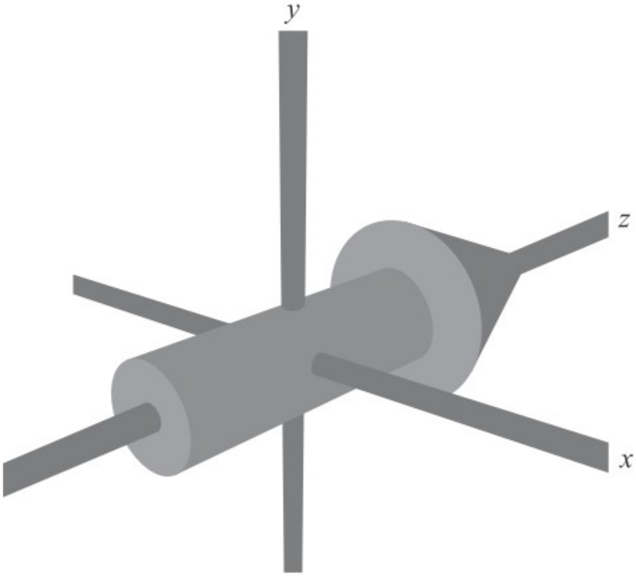

Figure 18.2
An arrow model in its untransformed state.

Imagine that the arrow in Figure 18.2 is located at position **p**, pointing along the vector **v**. To calculate **v**, you start with the transform of the arrow. If you know the transform of the arrow, then **v** is the normalized third column of the transform matrix. You want the arrow to point at a point **q** instead along the vector **v**, so you rotate it appropriately. To do this, you can first calculate the vector $\mathbf{q} - \mathbf{p}$. This is the new direction for the transform's *z*-axis. Normalized, you call this **u**. Taking the cross product $\mathbf{u} \times \mathbf{v}$, you can find the axis around which you would like the arrow to rotate. By taking the dot product $\mathbf{u} \cdot \mathbf{v}$, you can find the cosine of the rotation angle.

Now all you need to do is to perform the rotation. The general matrix for rotating around an arbitrary axis is a little ugly. It appears as follows:

$$\mathbf{R} = \begin{pmatrix} c_+ + a_1^2 c_- & a_1 a_2 c_- - a_3 s & a_1 a_3 c_- + a_2 s \\ a_1 a_2 c_- + a_3 s & c_+ + a_2^2 c_- & a_2 a_3 c_- - a_1 s \\ a_1 a_3 c_- - a_2 s & a_2 a_3 c_- + a_1 s & c_+ + a_3^2 c_- \end{pmatrix}$$

where the axis is given by the vector $(a_1 \quad a_2 \quad a_3)^T$. If the angle is θ, then s, c_+, and c_- are defined as $s = \sin\theta$, $c_+ = \cos\theta$, and $c_- = 1 - c_+$.

If you were working with non-homogeneous coordinates, you'd have to combine this with a translation to the origin, the rotation, and then the translation back. In this case you don't need to worry, however. Since translation is applied last, the translation and rotation parts of the homogeneous matrix are essentially independent. As a result, you can create a transform **S** that has **R** in the top-left portion and an identity position element.

Given this progress, all you need to do is to combine the matrix **S** with the original transform **T** to apply the rotation. While in most instances this might be done readily, however, there is one potential problem. You must decide whether you should right- or left-multiply the transforms. To solve this problem, consider that since **S** is to be applied after **T**, you need to combine it on the left as **ST**, not as **TS**.

Note

> In contrast, since it's generally more useful to apply the scale to the object in its unrotated form before applying rotations, consider that a scale transform is more likely to be applied before **T**, and therefore right-multiplied with it.

Everything seems to be in order at this point, but still there is another complication. What happens if the node you are interested in is not symmetrical? For example, if the node is not an arrow model but a camera, then it's not enough to point at a particular location. You must also specify which way is up. Otherwise, your camera will spin in strange directions.

To resolve this problem, you perform not just one rotation, but two. First, you rotate the object, as before, to point in a particular direction. This might be along its local *z*-axis. You then rotate again around this direction vector to keep the preferred up-vector (usually the local *y*-axis) pointing as close to upward as possible. Figure 18.3 shows this process.

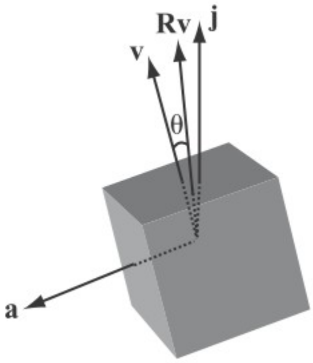

Figure 18.3

Keeping a preferred up-vector.

Note

> There's nothing special about the up direction. You could just as easily align it to keep a preferred left side. In a gravitational world, however, it's often sensible to single out the vertical direction.

Determining the angle of rotation requires calculus. You are seeking an angle that maximizes **Rv** · **j**, where **R** is a rotation matrix around the (normalized) pointing direction **a** = **q** – **p**, and **v** is the current direction of the up-vector. Plugging this into the matrix for **R**, you are trying to find θ that maximizes

$$(a_1a_2v_1 + a_2a_3v_3 + a_2^2v_2)(1 - \cos\theta) + v_2\cos\theta - (a_3v_1 + a_1v_3)\sin\theta$$

Differentiating, you see that θ must solve as

$$\tan\theta = \frac{(a_3v_1 + a_1v_3)}{(a_1a_2v_1 + a_2a_3v_3 + (a_2^2 - 1)v_2)}$$

This yields two possible values for θ. One gives the angle to point down. The other gives an angle to point up. Also, if the numerator and denominator are both zero, you are in a situation in which the pointing vector **a** is parallel to **j**, so all directions are equally far from pointing upward.

Interpolation

One of the most powerful features of transforms is the ability to create smooth motion by *interpolating* between two matrices. For example, you might want to create a camera that automatically follows a race car. If you place the camera at a constant distance relative to the car, the result is that the camera provides a view that makes it appear that it is tied on the back of the car. Instead of this, you want it to appear as though the camera is mounted on a helicopter smoothly following along behind the car.

To do this, you can calculate two transforms. The first is the current transform **C** of the camera. This is something that you are likely to already know. The second is the target transform **T**, which is at a set position relative to the moving car. At this point, instead of moving the camera to the new transform, you move it part way along, to some transform $t\mathbf{C} + (1 - t)\mathbf{T}$.

The effect of this change is to make the camera ease into position. This strategy works particularly well when following an object that is moving continually, like a car. A similar process can be used for following a third-person perspective character in a game like *Super Mario World* or *Tomb Raider*. In such cases, however, calculating the target transform is more complicated, since you have to deal with the possibility that other objects or landscape features are in the way. Generally, camera controls are one of the most difficult aspects of game design and can often make or break the game in terms of playability.

Quaternions

As mentioned previously, there are simple ways to handle rotations. The key concept is the *quaternion*. The quaternion is a special kind of 4-D vector. A quaternion $\mathbf{q}$ can be written as $(w \;\; x \;\; y \;\; z)^T$. Alternatively, it can be written as $w + xi + yj + zk$, where the numbers i, j, and k are related to the imaginary number i. The imaginary number i is defined to be the square root of -1.

In this case, i, j, and k are *orthogonal imaginary numbers*, in the sense that squaring any of them gives -1, and together they satisfy the following equations: $ij = -ji = k$; $jk = -kj = i$; $ki = -ik = j$. No real numbers satisfy these equations. They are not numbers in that sense. Instead, they are mathematical abstractions that do what you want them to do. What they generate is a kind of imaginary vector.

The quaternion $\mathbf{q}$ may also be written in the form $w + \mathbf{v}$, where $\mathbf{v}$ is the vector $(x \;\; y \;\; z)^{\mathrm{T}}$. You also define the conjugate of $\mathbf{q}$ to be the quaternion $w - \mathbf{v}$. Having defined your strange imaginary basis, you can calculate the product of two quaternions:

$$\begin{aligned}
&(w_1 + x_1 i + y_1 j + z_1 k)\,(w_2 + x_2 i + y_2 j + z_2 k) = w_1\,(w_2 + x_2 i + y_2 j + z_2 k)\\
&\quad + x_1 i\,(w_2 + x_2 i + y_2 j + z_2 k) + y_1 j\,(w_2 + x_2 i + y_2 j + z_2 k) + z_1 k\,(w_2 + x_2 i + y_2 j + z_2 k)\\
&\quad = (w_1 w_2 + w_1 x_2 i + w_1 y_2 j + w_1 z_2 k) + (x_1 w_2 i - x_1 x_2 + x_1 y_2 k + x_1 z_2 j)\\
&\quad + (y_1 w_2 j - y_1 x_2 k - y_1 y_2 + y_1 z_2 i) + (z_1 w_2 k - z_1 x_2 j - z_1 y_2 i - z_1 z_2)\\
&\quad = (w_1 w_2 - x_1 x_2 - y_1 y_2 - z_1 z_2) + (w_1 x_2 + x_1 w_2 + y_1 z_2 - z_1 y_2)\, i\\
&\quad + (w_1 y_2 + y_1 w_2 - z_1 x_2 + x_1 z_2)\, j + (w_1 z_2 + z_1 w_2 + x_1 y_2 - y_1 x_2)\, k
\end{aligned}$$

As unpleasant as this product might appear, it can be simplified if you use the vector form. If you say $\mathbf{q}_1 = w_1 + \mathbf{v}_1$ and $\mathbf{q}_2 = w_2 + \mathbf{v}_2$, then the product is given by

$$\mathbf{q}_1\mathbf{q}_2 = w_1 w_2 + \mathbf{v}_1 \cdot \mathbf{v}_2 + w_1\mathbf{v}_2 + w_2\mathbf{v}_1 + \mathbf{v}_1 \times \mathbf{v}_2$$

Notice that because of the cross-product component, the product of quaternions is not commutative.

Armed with this result, you can take the product of a quaternion with its conjugate:

$$\mathbf{q}\overline{\mathbf{q}} = w^2 - \mathrm{v}\cdot\mathrm{v} - \mathrm{v}\times\mathrm{v}$$

Since for any vector, $\mathbf{v} \times \mathbf{v} = 0$, you can reason that

$$\mathbf{q}\overline{\mathbf{q}} = w^2 - \mathrm{v}\cdot\mathrm{v}$$

And because $i^2 = j^2 = k^2 = -1$, you find that $\mathbf{q}\overline{\mathbf{q}} = \mathbf{q}\cdot\mathbf{q} = |\mathbf{q}|^2$. The inverse (both left and right) of a quaternion can be given by $\mathbf{q}^{-1} = \dfrac{\overline{\mathbf{q}}}{|\mathbf{q}|^2}$, or in particular for a unit quaternion, $\mathbf{q}^{-1} = \overline{\mathbf{q}}$. The product of a quaternion with its inverse is the quaternion $(1 \;\; 0 \;\; 0 \;\; 0)^{\mathrm{T}}$, which is equal to the scalar 1.

What does all this mathematical work have to do with rotations? It turns out that if you consider the vector $\mathbf{u}$ to be a quaternion with a zero scalar part, and if you calculate the product $\mathbf{quq}^{-1}$ for some quaternion $\mathbf{q}$, you arrive at the rotation of $\mathbf{v}$ about a particular axis.

Notice first that since the product $\mathbf{quq}^{-1}$ is the same for any scalar multiple of **q**, you can consider **q** to be a unit quaternion **s** + **v**, whose inverse is $\overline{\mathbf{q}} = \mathbf{s} - \mathbf{v}$. This gives the product as

$$\begin{aligned}\mathbf{quq}^{-1} &= (-\mathbf{v}\cdot\mathbf{u} + s\mathbf{u} + \mathbf{v}\times\mathbf{u})\overline{\mathbf{q}} \\ &= -s\mathbf{v}\cdot\mathbf{u} + (s\mathbf{u} + \mathbf{v}\times\mathbf{u})\cdot\mathbf{v} + s(s\mathbf{u} + \mathbf{v}\times\mathbf{u}) + (\mathbf{v}\cdot\mathbf{u})\mathbf{v} - (s\mathbf{u} + \mathbf{v}\times\mathbf{u})\times\mathbf{v} \\ &= \mathbf{v}\cdot(\mathbf{v}\times\mathbf{u}) + s^2\mathbf{u} + 2s(\mathbf{v}\times\mathbf{u}) + (\mathbf{v}\cdot\mathbf{u})\mathbf{v} - (\mathbf{v}\times\mathbf{u})\times\mathbf{v}\end{aligned}$$

By the identities of the dot and cross products, the first term is zero and the last term is $|\mathbf{v}|^2 - \mathbf{u}(\mathbf{v}\cdot\mathbf{u})\mathbf{v}$, so you have

$$\mathbf{quq}^{-1} = (s^2 - |\mathbf{v}|^2)\,\mathbf{u} + 2s(\mathbf{v}\times\mathbf{u}) + 2(\mathbf{v}\cdot\mathbf{u})\,\mathbf{v}$$

If you say $t = |\mathbf{v}|$ and set **a** to the normalized vector $\frac{\mathbf{v}}{t}$, then you can rewrite this as

$$\mathbf{quq}^{-1} = (s^2 - t^2)\,\mathbf{u} + 2st(\mathbf{a}\times\mathbf{u}) + 2t^2(\mathbf{a}\cdot\mathbf{u})\,\mathbf{a}$$

This represents a rotation about the axis **a**, with an angle θ given by $\cos\theta - s^2 = t^2$, $\sin\theta = 2st$, and $1 - \cos\theta = 2t^2$. Of these, you need only the last, since the first two together confirm that $1 - \cos\theta = 2t^2$, which you know because **q** is of unit length. The last is equivalent to saying that $s^2 + t^2 = 1$, which implies that $s = \cos\left(\frac{\theta}{2}\right)$. You can sum all this up by saying that a rotation around an axis **a** with angle θ can be calculated as follows using the quaternion:

$$\mathbf{q} = \cos\left(\frac{\theta}{2}\right) + \sin\left(\frac{\theta}{2}\right)\mathbf{a}$$

This is a fairly significant improvement on the matrix form. One reason is that far fewer calculations are involved in the transformation. Another reason is that it's much easier to see the relationship between the rotation and the quaternion than it is with the matrix. Also, just as with matrices, you can interpolate between two quaternions by taking a linear combination of them, although the rate at which the interpolation occurs is not the same as the rate at which the angle varies.

Parents and Children

In addition to combining transforms to move objects from one place to another, you can use them to define relationships between different nodes. This allows you to move whole groups of nodes together by varying a single transform, just as varying a node's transform can affect all the vertices of its model mesh.

You arrange all the nodes of the 3-D world in the form of a tree. In the tree, each node has exactly one parent and any number of children. The topmost node is the world itself, and each of the world node's children has its own transform relative to the world. But the transform of a child node is applied relative to its parent. In other words, if the child's relative transform is **C** and the parent's transform is **P**, then the child's transform relative to the world's coordinate system is **CP**.

For example, suppose one model is sitting in the world with a world transform consisting of a rotation about the *x*-axis and a uniform scale. Any child of this model will be rotated and scaled before its own transform is applied. As a result, the model and its children act as a group. If the parent model is translated, rotated or scaled, its children, grandchildren, and so on will move along with it.

Note

In this section, what applies to translation in most cases can be applied to rotation and scaling.

Such movement means that there is a potential ambiguity when transforming an object. Say that you want to move an object 5 units along the *z*-axis. Which *z*-axis does that mean? It could be the node's "local" *z*-axis, or its *z*-axis relative to its parent. It could also be the actual *z*-axis of the world. Figure 18.4 illustrates this situation.

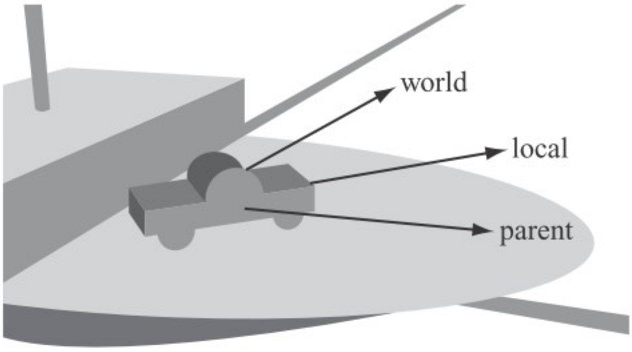

Figure 18.4
Relative motion in 3-D space.

Figure 18.4 represents a model of a car. The parent of the car is a ship on which the car is sitting. The car was modeled with its *z*-axis pointing along the body of the car. If you want to translate the car by 5 units along a *z*-axis, this could mean any of the directions indicated. The first is the car's own *z*-axis, which is pointing along the world vector $(1\ \ 0\ \ 1)^T$. The second is the ship's *z*-axis, which is pointing along the world's *x*-axis. The third is the world's *z*-axis. Ultimately, you might want to use the *z*-axis of any other node in the scene.

Fortunately, each of these is easy to calculate. The *z*-axis of the node is given by the last column of its complete world transform matrix. In general, any direction vector can be found simply by transforming it with the node's world transform. Similarly, while with a little work you can determine the direction of the parent's *z*-axis or any other model's, working out the world's *z*-axis is even easier.

Commonly, you want to go the other way around. You need to know how to change the node's personal transform in order to make it move along a particular vector. In this case, the simplest situation involves transforming the object in its local reference frame. Transforming locally is a matter of altering the transform of the object.

To alter the transform relative to the object's parent's transform is a little harder. To understand how this is so, suppose you want to move the object 5 pixels relative to the parent's *z*-axis. You must first calculate the vector in the child's reference frame. Expressed differently, you need a vector $\mathbf{v}$ such that $\mathbf{Cv} = \mathbf{k}$, which is given by $\mathbf{v} = \mathbf{C}^{-1}\mathbf{k}$. In English, this means that the vector $\mathbf{v}$, when "untransformed" into parent-space, is equal to the *z*-vector $\mathbf{k}$. Notice that this is the same independent of the parent's own transform. Similarly, to translate relative to a world-vector, you must invert the complete world transform of the child node.

Note

The vector **v** does not need to be of unit length, since any of the transforms involved may include a scaling element. Actually, this is convenient, for it means you can perform your translation of 5 units without worrying about scale. To do so, you multiply by the already scaled vector **v**. But for rotations you need to remember to normalize the axis vector before performing the rotation.

Exercise

Exercise 18.1

Create functions that will apply a translation, scale, or rotation for a particular node relative to the world, the node's parent, or the node itself. You should think about which method of storing transform information is most useful: as a single matrix or as three transformation vectors? Using quaternions for rotations?

Summary

In this chapter you have seen how homogeneous coordinates and mathematical gizmos like quaternions can be used to create a simple representation of a node's disposition in 3-D space. In the process, you've also learned how to create smooth motion incorporating simultaneous rotation, translation, and scale transformations by interpolating between transforms.

In the next chapter, you'll take one final look at collision detection by extending your previous work into the third dimension.

You Should Now Know

- The meaning of the term *transform* and how it is represented in various forms: as a single matrix, as a combination of separate matrices, and as three vectors
- How *quaternions* can be used to create a simpler representation of rotations
- How transforms can be combined to move objects from place to place, and to create groups of objects with a fixed position relative to one another

A *+,-* * *. * ! ; K "
+// " 6 , $ /"
(4 " 8 / 4

of the string; deflections with a small (large) slope are (Subsection 2.1.3) linear (nonlinear). The elastic string is the one-dimensional analogue of an elastic membrane (Subsection 2.1.4) in two dimensions (Sections II.6.1 and II.6.2).

2.1.1 Tangential and Normal Components of the Tension

In the absence of transverse loading, the string is straight, and the x-axis can be taken along it; in the presence of transverse loads (Figure 2.1a), the string is deflected (2.1a), and the arc length is given by (2.1b)

$$y = \zeta(x): \quad (\mathrm{d}s)^2 = (\mathrm{d}y)^2 + (\mathrm{d}x)^2 = \left(1 + \zeta'^2\right)(\mathrm{d}x)^2, \quad \zeta' \equiv \frac{\mathrm{d}\zeta}{\mathrm{d}x}, \tag{2.1a–c}$$

where prime denotes the derivative with regard to x; for example, (2.1c) is the slope of the shape (2.1a) of the string. Since the tension T is tangent to the string, its horizontal (T_x) and vertical (T_y) components are given (Figure 2.1a) by

$$T_x = T\frac{\mathrm{d}x}{\mathrm{d}s} = T\left|1 + \zeta'^2\right|^{-1/2}, \quad T_y = T\frac{\mathrm{d}y}{\mathrm{d}s} = T\frac{\mathrm{d}y}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}s} = T_x\zeta' = T\zeta'\left|1 + \zeta'^2\right|^{-1/2}, \tag{2.2a,b}$$

where (2.1a through c) is used. The exact horizontal (2.2a) and vertical (2.2b) tensions are a nonlinear function of the slope, which can be linearized for small slopes (2.3a):

$$\zeta'^2 \ll 1 \Leftrightarrow \mathrm{d}s = \mathrm{d}x \Rightarrow \left(y_{\max}\right)^2 \ll L^2, \tag{2.3a–c}$$

and (1) this is equivalent to the statement (2.3b) that the length of the string is equal to its horizontal projection (Figure 2.1b); (2) it implies that (2.3c) the maximum transverse deflection $y_{\max}$ is small relative to the length L of the string (Figure 2.2a); and (3) the latter condition is necessary but not sufficient, that is, the relative total deflection may be small, yet the slope could be locally large if the string has steep ripples (Figure 2.2b), and in this case the deformation is nonlinear. For a small slope (2.4a) everywhere,

$$\zeta'^2 \ll 1: \quad T_x = T, \quad T_y = T\zeta', \tag{2.4a–c}$$

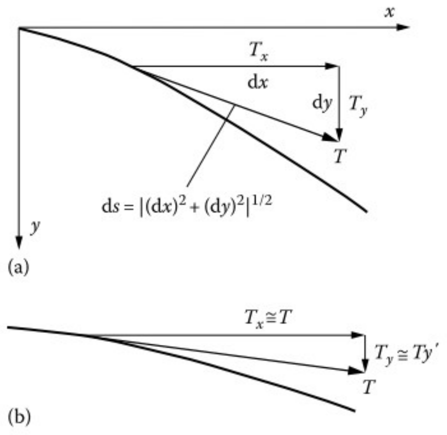

FIGURE 2.1 The tension along the tangent to an elastic string has vertical and horizontal components (a) that simplify in the linear case of small slope (b).

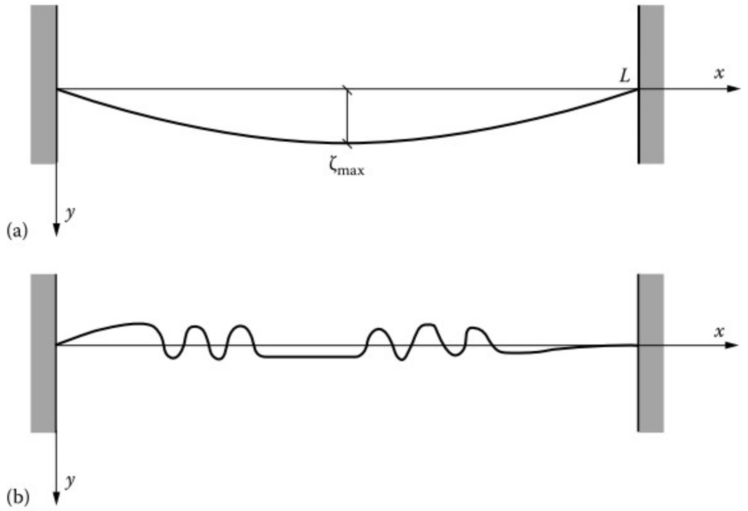

FIGURE 2.2 A linear deflection is defined by a small slope, and implies that the maximum deflection is small compared with the distance between the supports (a). The converse may not be true, for example, if the maximum deflection is small but the slope is large due to the presence of steep "ripples" (b) the deflection of the string is nonlinear.

the horizontal tension (2.4b) coincides with the tension T along the string, and the vertical tension (2.4c) is small but nonnegligible by comparison. Hence, *for an elastic string of shape (2.1a), under uniform or nonuniform tension T, the horizontal and vertical components are given exactly in the nonlinear case by (2.2a and b) and are approximated by (2.4b and c) in the linear case of a small slope everywhere by (2.4a)* ≡ *(2.3a through c).*

2.1.2 Transverse Force and Shear Stress

If a transverse force $F(x)$ is applied to the string, then the latter will deflect to an equilibrium shape such that, along each segment dx, the variation of the **transverse force** dF is balanced by the variation of the vertical component of the tension (2.5a):

$$\mathrm{d}F + \mathrm{d}T_y = 0, \quad F(x) = -T_y(x), \tag{2.5a,b}$$

where the constant of integration in (2.5b) is zero because the two terms must balance at the supported ends of the string, $x = 0, L$. Substituting (2.2b) into (2.5b) leads to the differential equation of first order (2.6) specifying the shape of the string:

$$-F(x) = T\zeta'\left|1+\zeta'^2\right|^{-1/2}, \tag{2.6}$$

under an arbitrary transverse force $F(x)$. The **shear stress** $f(x)$ is defined as the transverse force per unit length (2.7a):

$$f(x) \equiv \frac{\mathrm{d}F}{\mathrm{d}x}, \quad F(x) = F(0) + \int_0^x f(\xi)\,\mathrm{d}\xi, \tag{2.7a,b}$$

(& !$@ $! $ $ 9 ./ " ; D
.A 4" /P

+,- * *. * ! / G + D +//
6 " 6 " , /"
($ @ 6 !

3

Addition, Product, and Inverses

An operation upon complex numbers is a relation between two ordered pairs of real numbers and a third ordered pair, and it can be represented graphically on the complex plane. The sum (Sections 3.1 and 3.2) is most conveniently performed in the Cartesian representation, and the product (Sections 3.3 and 3.4) in the polar representation. The operations sum and product, and their inverses subtraction and division, can be used (Sections 3.5 and 3.6) together with the conjugate operator to determine the real and imaginary parts, and the modulus and argument, of complex expressions. The algebra of complex numbers is related to the geometry of the Cartesian plane (Sections 3.8 and 3.9), and implies a number of trigonometrical relations (Section 3.7). The algebra of complex numbers extends to functions of one or more variables, and thus to complex spaces whose real dimension is twice the number of complex variables.

3.1 Complex Addition and Rule of the Parallelogram

Starting with two complex numbers in the Cartesian representation:

$$u = a + ib, \quad v = c + id, \tag{3.1a,b}$$

the operator addition is defined by the sum of each element of the pair,

$$u + v \equiv (a + c) + i\,(b + d), \tag{3.2}$$

and corresponds (Figure 3.1) to the addition of u and v as vectors, according to the **rule of the parallelogram**. Hence, the real (imaginary) part is the sum of the real (imaginary) parts:

$$\operatorname{Re}(u + v) = \operatorname{Re}(u) + \operatorname{Re}(v), \quad \operatorname{Im}(u + v) = \operatorname{Im}(u) + \operatorname{Im}(v). \tag{3.3a,b}$$

The addition of complex numbers has the following properties: (i) it is closed because the sum of complex numbers is a complex number, that is, (1.1) holds for the operation + in $|C$; (ii) it is commutative (3.4a) and associative (3.4b):

$$u + v = v + u, \quad (u + v) + w = u + (v + w); \tag{3.4a,b}$$

(iii) it has a neutral element, namely, zero $0 + 0i$; (iv) each complex number $u = a + ib$ has a symmetric $-u = -a - ib$, given by (1.18). Hence, the ordered pair $(|C, +)$, where $|C$ is the set of complex numbers and + the operation addition defined by (3.1a,b; 3.2), is a **commutative group**.

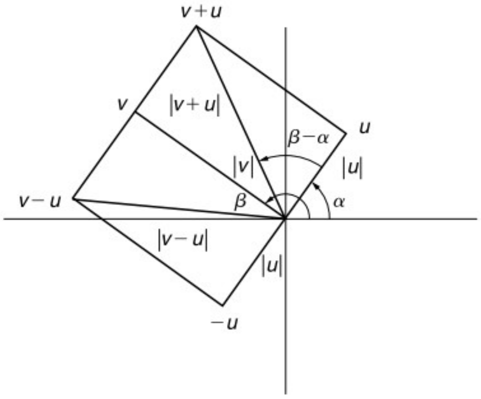

FIGURE 3.1
Taking two complex numbers u, v as the sides of a rectangle their sum is the diagonal. Their difference $u - v$ is the diagonal of the rectangle obtained by reflection on the v-direction.

3.2 Modulus, Argument, and Triangular Equalities (Pythagoras, VI b.c.)

The modulus (1.11) of the sum (3.2) is given by

$$\begin{aligned} |u+v|^2 &= (a+c)^2 + (b+d)^2 = \left(a^2+d^2\right) + 2\,(ac+bd) \\ &= |u|^2 + |v|^2 + 2\,|u|\,|v|\,(\cos\alpha\cos\beta + \sin\alpha\sin\beta)\,, \end{aligned} \tag{3.5}$$

where α, β are the arguments of u,v. The expression (3.5) simplifies

$$\alpha \equiv \arg(u)\,, \beta \equiv \arg(v): \quad |u+v| = \left| |u|^2 + |v|^2 + 2\,|u|\,|v|\cos(\alpha-\beta)\right|^{1/2}, \tag{3.6}$$

which *specifies the length* $|u+v|$ *of one side of a triangle, from the lengths* $|u|, |v|$ *of the other two sides, and the angle* $\alpha - \beta$ *between them* (Figure 3.2a). In the case $\alpha - \beta = \pm\pi/2$, the two perpendicular sides lead to **Pythagoras theorem (VI b.c.)** *for the rectangular triangle (Figure 3.2b):*

$$\arg(u) - \arg(v) = \pm\frac{\pi}{2}: \quad |u+v|^2 = |u|^2 + |v|^2\,. \tag{3.7}$$

From (1.12a) and (3.2) follows the expression:

$$\begin{aligned} \arg(u+v) &= \text{arc }\tan\left(\frac{b+d}{a+c}\right) \\ &= \text{arc}\tan\left(\frac{|u|\sin\alpha + |v|\sin\beta}{|u|\cos\alpha + |v|\cos\beta}\right), \end{aligned} \tag{3.8}$$

for *the argument of the sum of complex numbers (3.8);* other alternative expressions are given in Example 10.6.

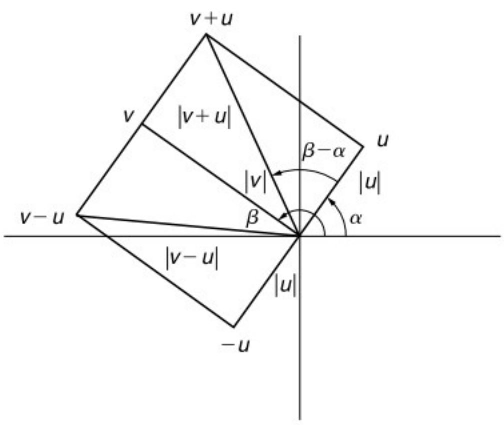

6 !

3.2 Modulus, Argument, and Triangular Equalities (Pythagoras, VI b.c.)

The modulus (1.11) of the sum (3.2) is given by

$$\begin{aligned}|u+v|^2 &= (a+c)^2 + (b+d)^2 = \left(a^2+d^2\right) + 2\,(ac+bd)\\ &= |u|^2 + |v|^2 + 2\,|u|\,|v|\,(\cos\alpha\cos\beta + \sin\alpha\sin\beta)\,,\end{aligned} \tag{3.5}$$

where α, β are the arguments of u,v. The expression (3.5) simplifies

$$\alpha \equiv \arg(u)\,, \beta \equiv \arg(v): \quad |u+v| = \left|\,|u|^2 + |v|^2 + 2\,|u|\,|v|\cos(\alpha-\beta)\right|^{1/2}, \tag{3.6}$$

which *specifies the length* $|u+v|$ *of one side of a triangle, from the lengths* $|u|, |v|$ *of the other two sides, and the angle* $\alpha-\beta$ *between them* (Figure 3.2a). In the case $\alpha-\beta = \pm\pi/2$, the two perpendicular sides lead to **Pythagoras theorem (VI b.c.)** *for the rectangular triangle (Figure 3.2b):*

$$\arg(u) - \arg(v) = \pm\frac{\pi}{2}: \quad |u+v|^2 = |u|^2 + |v|^2\,. \tag{3.7}$$

From (1.12a) and (3.2) follows the expression:

$$\begin{aligned}\arg(u+v) &= \text{arc }\tan\left(\frac{b+d}{a+c}\right)\\ &= \text{arc }\tan\left(\frac{|u|\sin\alpha + |v|\sin\beta}{|u|\cos\alpha + |v|\cos\beta}\right),\end{aligned} \tag{3.8}$$

for *the argument of the sum of complex numbers (3.8);* other alternative expressions are given in Example 10.6.

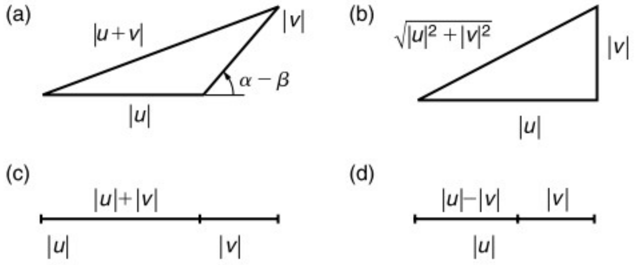

FIGURE 3.2
The sum of complex numbers by the rule of the diagonal of the rectangle (Figure 3.1) is equivalent to summing complex numbers as the sides of a triangle (a). A particular case is a rectangular triangle (b) with orthogonal sides for which holds the Pythagoras theorem. The degenerate cases are collapsed triangles with parallel (c) or antiparallel (d) sides.

" *+,-* * *. * ! + +/ " /"

(

1.8 — ESTABILIDADE DE PÊNDULO E DE NAVIO

1.8.1 — Estabilidade de pêndulo no campo gravítico
1.8.2 — Movimento do pêndulo e tensão na barra
1.8.3 — Distância metacêntrica e momento resultante
1.8.4 — Comprimento do pêndulo equivalente ao casco
1.8.5 — Energias cinética, potencial e gravítica
1.8.6 — Movimento na vizinhança da posição de equilíbrio
1.8.7 — Cortes e domínio angular do movimento
1.8.8 — Oscilação em torno da posição de equilíbrio estável
1.8.9 — Afastamento da posição de equilíbrio instável

1.9 — PROJECÇÃO ESTEREOGRÁFICA DA ESFERA

1.9.1 — A esfera unidade e o plano complexo
1.9.2 — Correspondência bijectiva e "ponto no infinito"
1.9.3 — Projecção estereográfica directa e inversa
1.9.4 — Projecção de circunferência sobre recta
1.9.5 — Projecção de circunferência sobre circunferência
1.9.6 — Isoformismo da circunferência com a recta real.

Exercícios 1-20.

1 — DOMÍNIO COMPLEXO

Os primeiros números usados pelo homem foram os naturais (entiers positifs; positive integers) que permitiam fazer somas sem restrições; a diferença de dois naturais $n-m$ só é um natural se $n>m$. Para poder realizar a subtracção sem restrições introduziram-se os inteiros (entiers; integers); o produto de inteiros não tem restrições, mas a divisão de dois inteiros n/m só é um inteiro se n é múltiplo de m. Para poder realizar a divisão noutros casos introduziram-se os racionais (rationels; rationals); estes não são compactos, visto que entre dois racionais existem outros números (por exemplo π, e) que são os irracionais (irrationels; irrationals). O conjunto dos reais (réels; reals), que consiste dos racionais e irracionais é compacto; permite a operação potência mas a raíz $\sqrt[n]{a}$ nem sempre tem n soluções reais. Introduzem-se então os complexos (complexes, complex), para os quais tanto as operações directas (soma, produto e potência), como as operações inversas (subtracção, divisão e radiação) não tem restrições.

os racionais são densos mas não são compactos

TABELA

Operações em relação às quais os conjuntos de números são fechados (fermées; closed), isto é, a operação aplicada sobre elementos do conjunto dá outro elemento do conjunto.

CONJUNTO	Naturais	Inteiros	Racionais	Reais	Complexos
Operação:	N	I	P	R	S
Soma	x	x	x	x	x
Subtracção	—	x	x	x	x
Produto	x	x	x	x	x
Divisão	—	—	x	x	x
Potência	x	x	x	x	x
Radiciação	—	—	—	—	x

1.1 — REPRESENTAÇÃO CARTESIANA E POLAR

Um número complexo é um par ordenado de números reais e o conjunto dos complexos está em correspondência com os pontos do plano, o que permite a sua representação gráfica.

1.1.1 — **Parte real e parte imaginária**

Um número *complexo z é um par ordenado (x,y) de números reais*, para o qual se usa vulgarmente a notação

$$z = x + iy \tag{1}$$

em que x e y se designam respectivamente de parte real e imaginária (partie réelle et imaginaire; real and imaginary part):

$$x \equiv \mathrm{Re}(z), \qquad y \equiv \mathrm{Im}(z). \tag{2a,b}$$

As partes real e imaginária podem ser tomadas como coordenadas x,y num referencial cartesiano (Descartes 1637), sendo **z** representado por um ponto no plano de Argand (1806; aliás introduzido por Gauss em 1799). Logo existe uma correspondência biunívoca entre o conjunto S dos complexos **z**, o conjunto R^2 dos pares ordenados (x,y) reais e os pontos P do plano complexo (plan complexe; complex plane).

Caspar Wessel

1.1.2 — Coordenadas polares: módulo e argumento

A distância do ponto P (Figura 1) representativo do complexo **z** à origem define o módulo (module, modulus):

$$\rho \equiv |z| = |x^2 + y^2|^{1/2} \tag{3}$$

em que $|\dots|^{1/2}$ significa que se toma a raíz quadrada com sinal positivo. O ângulo do segmento OP com o eixo Ox, medido a partir do eixo Ox, no sentido positivo, que é o contrário ao dos ponteiros do relógio (counter-clockwise), define o argumento (argument):

$$\theta \equiv \arg(z) = \begin{cases} \operatorname{arc\,tan}(y/x) & x > 0, \\ \operatorname{arc\,tan}(y/x) + \pi & x < 0, \end{cases} \tag{4}$$

em que $-\pi/2 < \theta < \pi/2$ para $x > 0$ (semi-plano direito) e $\pi/2 < \theta < 3\pi/2$ para $x < 0$ (semi-plano esquerdo). Um *número complexo é representado univocamente pelo seu módulo (3)* $\rho \geqslant 0$ *e argumento (4)* $0 \leqslant \theta < 2\pi$; o argumento é indeterminado na origem $\rho = 0$, já que (4) depende da forma como $x,y \to 0$.

1.1.3 — Fórmula de Moivre (1730; Euler 1749)

Entre as partes real e imaginária, e o módulo e argumento de um complexo existem as seguintes relações (vide Figura 1):

$$x = \rho\cos\theta, \qquad y = \rho\sin\theta. \tag{5a,b}$$

Substituindo em (l):

$$z = \rho(\cos\theta + i\sin\theta) = \rho\, e^{i\theta} \tag{6}$$

em que usou a fórmula

$$\boxed{e^{i\theta} \equiv \cos\theta + i\sin\theta} \tag{7}$$

que será justificada ulteriormente (§ 5.5.1). *As expressões (1) e (6) são respectivamente a representação cartesiana* (cartesienne; cartesian) *e polar* (polaire; polar) *do complexo.* A passagem de representação cartesiana à polar é feita por (3,4), e a transformação inversa é (5a,b). O Jacobiano respectivo

$$\begin{vmatrix} \partial x/\partial\rho & \partial y/\partial\rho \\ \partial x/\partial\theta & \partial y/\partial\theta \end{vmatrix} = \begin{vmatrix} \cos\theta & \sin\theta \\ -\rho\sin\theta & \rho\cos\theta \end{vmatrix} = \rho \neq (0, \infty) \tag{8}$$

mostra que a transformação é *invertível excepto na origem (§ 1.1.2) e no infinito (§ 1.9), onde* θ *é indeterminado.*

1.1.4 — Conjugado e simetria na origem e eixos

O ponto simétrico de P em relação ao eixo real (axe réel; real axis) define o complexo conjugado:

$$z^* = x - iy = \rho\, e^{-i\theta}\,. \tag{9}$$

Note-se que a dupla aplicação do conjugado é idempotente:

$$(z^*)^* = z\,. \tag{10}$$

O simétrico de P em relação à origem (origine; origin) é o simétrico algébrico:

$$-z = -x - iy = \rho\, e^{i(\pi+\theta)}. \tag{11}$$

O simétrico em relação ao eixo imaginário (axe imaginaire; imaginary axis) é o conjugado do simétrico:

$$-z^* = -x + iy = \rho\, e^{i(\pi-\theta)}. \tag{12}$$

Convencionando que o primeiro quadrante se segue ao quarto conclui-se que (Figura 1) *os complexos z* (1,6), $-z^*$ (12), $-z$ (11) e z^* (9) *estão em quadrantes sucessivos.*

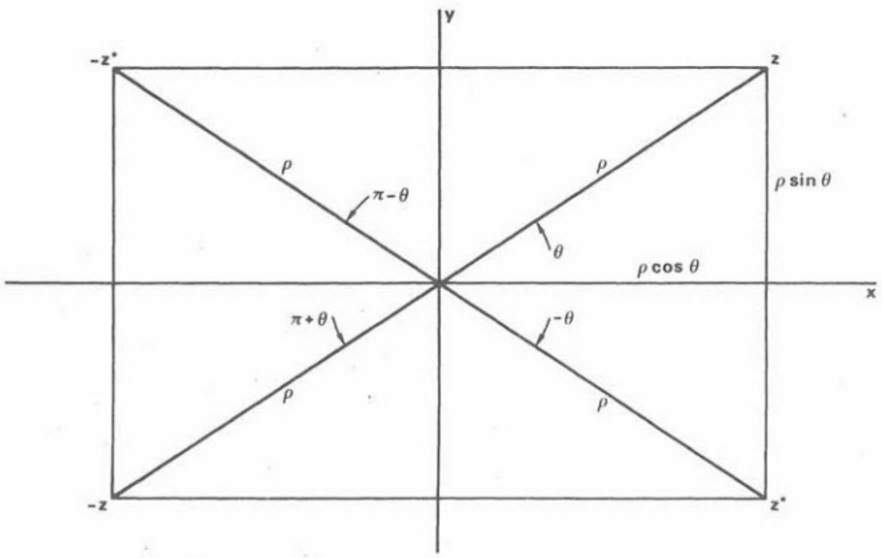

Fig. 1 — Representação de complexos no plano.

.A + / / "

1.2 — ESTABILIDADE DE UMA POSIÇÃO DE EQUILÍBRIO

Quando um corpo (ou sistema) se encontra numa posição (ou estado) de equilíbrio, e é perturbado podem verificar-se três casos: (i) o corpo permanece na posição perturbada, e o equilíbrio diz-se indiferente (indifferent); (ii) o corpo regressa à posição de equilíbrio que se diz estável (stable); (iii) o corpo afasta-se mais da posição de equilíbrio, que se diz instável (unstable). O afastamento ou aproximação da posição de equilíbrio, após a perturbação, pode ser monótono (monotone; monotonic) ou oscilatório (oscillatoire; oscilatory). Estes vários casos de movimento estão incluídos numa só expressão complexa.

1.2.1 — **Trajectória após afastamento da posição de equilíbrio**

Um corpo encontra-se na posição de equilíbrio caracterizada pela coordenada $x=0$. No instante $t=0$ é afastado para a posição x_o, e o seu movimento ulterior é dado pela lei:

$$\zeta = \omega + i\lambda : \; x(t) = \mathrm{Re}(x_o \, e^{i\zeta t}), \tag{1}$$

em que ζ é em geral complexo. Visto que **x** tem que ser real, tomou-se a parte real da expressão entre parêntesis; (1) é um exemplo de representação complexa de uma grandeza real. A representação complexa também pode ser usada para calcular a velocidade **v** e a aceleração **a**,

$$v(t) \equiv dx/dt = \mathrm{Re}(i\zeta x_o \, e^{i\zeta t}), \tag{2a}$$

$$a(t) \equiv dv/dt = d^2x/dt^2 = \mathrm{Re}(-\zeta^2 x_o \, e^{i\zeta t}). \tag{2b}$$

1.2.2 — **Corpo em repouso e equilíbrio indiferente**

Pretende-se agora analisar o movimento para todos os valores possíveis de ζ, e concluir da estabilidade da posição de equilíbrio $x=0$. O caso trivial (I) é $\zeta=0$, em que:

$$\zeta = 0; \quad x(t) = x_o, \quad v(t) = 0 = a(t), \tag{3a,b}$$

isto é *se $\zeta=0$, a posição de equilíbrio $x=0$ é indiferente, visto que o corpo uma vez afastado para x_o mantém-se em repouso na nova posição* (Figura 2a).

( )

1.3 — OPERAÇÕES ADIÇÃO E PRODUTO

As operações sobre números complexos são relações de dois pares ordenados de reais com um terceiro par ordenado de reais e são representáveis graficamente no plano complexo. Para a adição é mais conveniente a representação cartesiana, ao passo que a representação polar simplifica o cálculo do produto. As operações adição e produto, e as suas inversas subtracção e divisão podem ser usadas com o conjugado para determinar a parte real, imaginária, módulo ou argumento de expressões complexas.

1.3.1 — **Adição de complexos e regra do paralelogramo**

Dados dois números complexos na representação cartesiana:

$$u = a + ib, \qquad v = c + id \tag{1a,b}$$

define-se o resultado da operação adição como o complexo:

$$u + v = (a + c) + i\,(b + d) \tag{2}$$

obtido (Figura 3a) adicionando **u** e **v** como vectores, segundo a regra do paralelogramo. Logo, a parte real/imaginária da soma é a soma das respectivas partes real/imaginária:

$$\operatorname{Re}(u + v) = \operatorname{Re}(u) + \operatorname{Re}(v), \qquad \operatorname{Im}(u + v) = \operatorname{Im}(u) + \operatorname{Im}(v). \tag{3}$$

A operação adição de complexos tem as seguintes propriedades: (i) é fechada visto a soma de dois complexos ser um complexo; (ii) é comutativa e associativa:

$$u + v = v + u, \qquad (u + v) + w = u + (v + w); \tag{4}$$

(iii) tem elemento neutro, que é $0 + 0i$; (iv) todo o complexo $u = a + ib$, tem simétrico $-u = -a - ib$. Logo o *par ordenado (S, +) em que S é o conjunto dos complexos, e + a operação adição definida por (1a, b; 2) é um grupo abeliano.*

1.3.2 — **Módulo da soma e teorema de Pitágoras**

Calculando o módulo de $u + v$ (2), usando (1.1.3):

$$\begin{aligned}|u + v|^2 &= (a + c)^2 + (b + d)^2 = (a^2 + b^2) + (c^2 + d^2) + 2\,(ac + bd) = \\ &= |u|^2 + |v|^2 + 2\,|u|\,|v|\,(\cos\alpha\,\cos\beta + \sin\alpha\,\sin\beta),\end{aligned} \tag{5}$$

em que $\alpha = \arg(u)$, $\beta = \arg(v)$. Logo *o módulo da soma é dado pela fórmula:*

$$|u+v| = \{|u|^2 + |v|^2 + 2|u|\,|v| \cos(\alpha - \beta)\}^{1/2}, \tag{6}$$

que exprime (Figura 3c) o comprimento de um lado do triângulo em termos do comprimento dos outros dois lados, e do ângulo $\alpha - \beta$ entre eles. No caso de dois lados serem perpendiculares $\alpha - \beta = \pi/2$, e obtém-se:

$$\alpha - \beta = \pi/2: \quad |u+v|^2 = |u|^2 + |v|^2, \tag{7}$$

que é o *Teorema de Pitágoras para o triângulo rectângulo.*

1.3.3 — Produto de complexos, homotetia e rotação

Dados dois complexos na forma polar:

$$u = |u|\, e^{i\alpha}, \quad v = |v|\, e^{i\beta} \tag{8a,b}$$

define-se o produto como o complexo:

$$uv = |u|\,|v|\, e^{i(\alpha+\beta)}. \tag{9}$$

Logo, multiplicar o complexo **u** pelo complexo **v**, corresponde (Figura 3b) a: (i) multiplicar o módulo de **u** pelo de **v**, o que corresponde a uma extensão se $|v| > 1$, ou contracção se $|v| < 1$; (ii) adicionar ao argumento de **u** ao de **v**, o que corresponde a uma rotação no sentido positivo se $\beta > 0$, e negativo se $\beta < 0$:

$$|uv| = |u|\,|v|, \quad \arg(uv) = \arg(u) + \arg(v), \tag{10a,b}$$

Atendendo a que $|uv|/|v| = |u|/1$, e $\arg(u) = \arg(uv) - \arg(v)$, os triângulos de vértices (0, 1, u) e (0, v, uv) são semelhantes, isto é, têm ângulos internos iguais dois a dois e lados todos na mesma proporção. *O par ordenado (S, x), em que S é o conjunto dos complexos e x a operação produto definida por (8a,b; 9) é um grupo abeliano.* O elemento neutro é 1, e o inverso de u é 1/u (§ 1.3.5).

1.3.4 — Significado do símbolo imaginário

As partes real e imaginária (9) são dadas respectivamente por:

$$\begin{aligned} \mathrm{Re}(uv) = |u|\,|v| \cos(\alpha+\beta) &= |u| \cos\alpha\, |v| \cos\beta - \\ &- |u| \sin\alpha\, |v| \sin\beta = ac - bd, \end{aligned} \tag{11a}$$

$$\operatorname{Im}(uv) = |u|\,|v|\,\sin(\alpha+\beta) = |u|\,\sin\alpha\,|v|\,\cos\beta + {} + |u|\,\cos\alpha\,|v|\,\sin\beta = bc - ad; \tag{11b}$$

Logo o produto de complexos na forma cartesiana é:

$$(a+ib)\,(c+id) = (ac-bd) + i\,(ad+bc), \tag{12}$$

que corresponde à regra vulgar do produto desde que:

$$i^2 = -1, \quad \sqrt{-1} = \pm i\,. \tag{13a,b}$$

A condição (13a,b) a que satisfaz o símbolo imaginário i pode ser deduzida de $i \equiv e^{i\pi/2}$ logo $i^2 = e^{i\pi} = -1$.

1.3.5 — **Conjugado da soma, produto e inversão**

A partir de (2) e (12) verifica-se que o conjugado da soma/produto é a soma/produto dos conjugados:

$$(u+v)^* = u^* + v^*, \quad (uv)^* = u^*\,v^*. \tag{14a,b}$$

De (1.1.1,9) deduzem-se as seguintes fórmulas para a parte real e imaginária dum complexo:

$$x \equiv \operatorname{Re}(z) = (z+z^*)/2, \quad y \equiv \operatorname{Im}(z) = (z-z^*)/2i\,. \tag{15a,b}$$

De (1.1.6,9) deduzem-se fórmulas para o módulo e argumento:

$$\rho \equiv |z| \equiv (z\,z^*)^{1/2}, \quad \theta \equiv \arg(z) = \frac{1}{2}\arg(z/z^*)\,. \tag{16a,b}$$

O conjugado permite obter uma fórmula para o inverso de um complexo:

$$1/z = \rho^{-1}\,e^{-i\theta} = z^*/zz^* = z^*\,|z|^{-2}, \tag{17}$$

cujas partes real e imaginária são respectivamente:

$$\operatorname{Re}(1/z) = \rho^{-1}\cos\theta = x\,(x^2+y^2)^{-1/2}, \tag{18a}$$

$$\operatorname{Im}(1/z) = -\,\rho^{-1}\sin\theta = -\,y\,(x^2+y^2)^{-1/2}. \tag{18b}$$

1.3.6 — Representação complexa de grandezas reais

As fórmulas (15a,b) são o exemplo da representação de reais por meio de complexos. As representações complexas adicionam-se, por exemplo $x + iy = z$, mas o produto simplifica-se:

$$\frac{1}{2}(u + u^*)\,\frac{1}{2}(v + v^*) = \frac{1}{4}\{(uv + uv^*) + (uv + uv^*)^*\},$$

que corresponde a:

$$\mathrm{Re}(u)\,\mathrm{Re}(v) = \frac{1}{2}\,\mathrm{Re}(uv + uv^*). \tag{19}$$

Analogamente se demonstra que:

$$\mathrm{Im}(u)\,\mathrm{Im}(v) = \frac{1}{2}\,\mathrm{Re}(uv^* - uv), \tag{20}$$

$$\mathrm{Re}(u)\,\mathrm{Im}(v) = \frac{1}{2}\,\mathrm{Im}(uv - uv^*). \tag{21}$$

1.3.7 — Desigualdades incluindo a triangular de Schwartz

Elevando (15a,b) ao quadrado deduzem-se as desigualdades:

$$(z + z^*)^2 \geqslant 0 \geqslant (z - z^*)^2, \tag{22}$$

em que a primeira igualdade só se verifica **para z imaginário** puro e a **segunda para z** real. Atendendo a que $-1 \leqslant \cos(\alpha - \beta) \leqslant 1$, deduz-se de (6):

$$|u|^2 + |v|^2 - 2|u|\,|v| \leqslant |u + v|^2 \leqslant |u|^2 + |v|^2 + 2|u||v|,$$

que é a *desigualdade de Schwartz:*

$$|u| - |v| \leqslant |u + v| \leqslant |u| + |v|, \tag{23}$$

em que a *segunda igualdade se verifica se* $\arg(u) = \arg(v)$*, e a primeira se* $\arg(u) = \arg(v) + \pi$*. A interpretação geométrica (Figura 3c) é que o comprimento de um lado de um triângulo não excede a soma dos comprimentos dos outros dois, nem é inferior à sua diferença; a igualdade corresponde ao triângulo degenerado em segmento.*

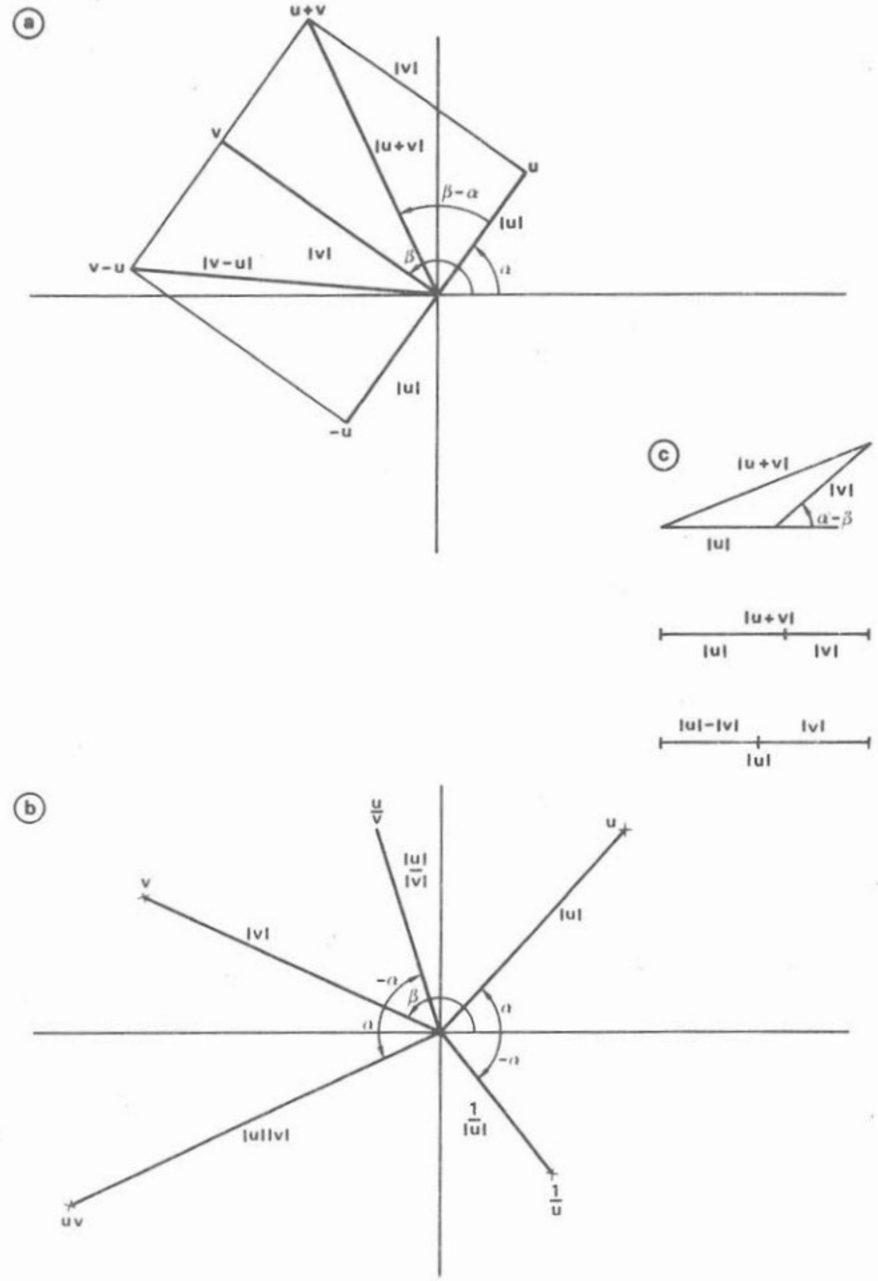
a
u+v
|v|
v
|u+v|
u
β−α
|u|
β
α
v−u
|v−u|
|v|
|u|
−u
c
|u+v|
|v|
α−β
|u|
|u+v|
|u|
|v|
|u|−|v|
|v|
|u|
b
u/v
|u|/|v|
u
v
|v|
|u|
−α
β
α
α
−α
1/|u|
|u||v|
uv
1/u

Fig. 3 — Operações sobre complexos: (a) adição e subtracção; (b) multiplicação e divisão. Ilustração gráfica da desigualdade de Schwartz (c).

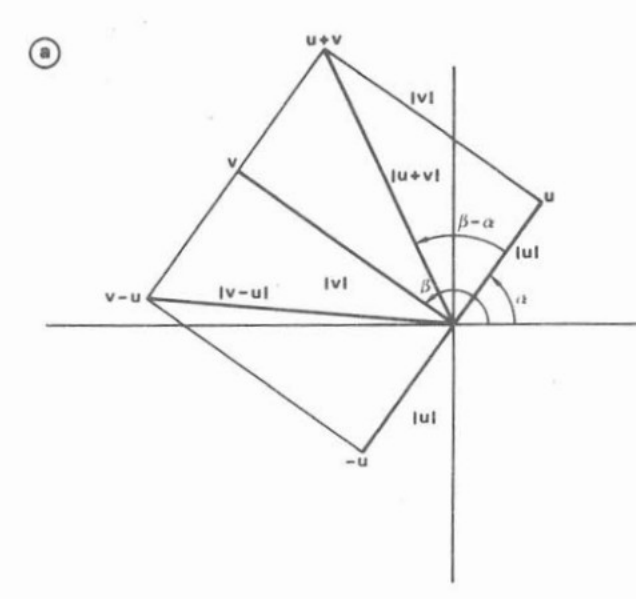
ⓐ
u+v
|v|
v
|u+v|
u
β−α
|u|
v−u
|v−u|
|v|
β
α
|u|
−u

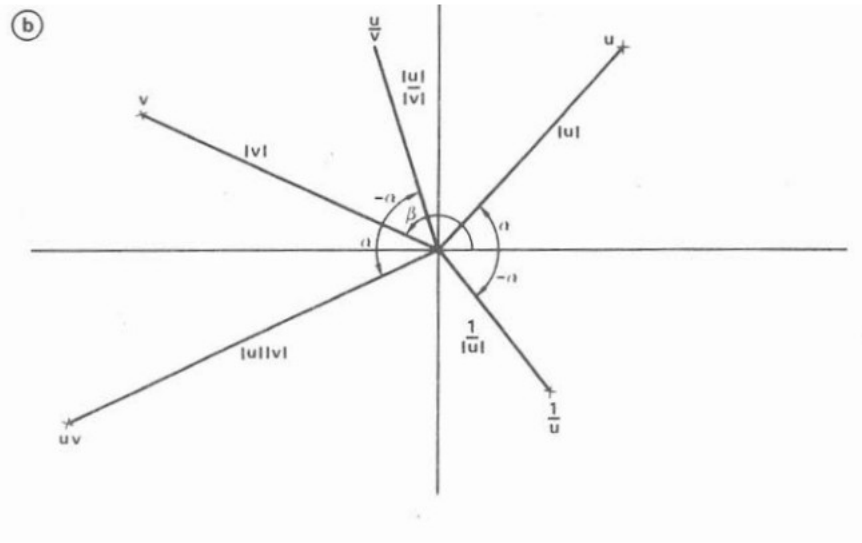
ⓑ
$\frac{u}{v}$
$\frac{|u|}{|v|}$
u
v
|v|
|u|
−α
β
α
α
−α
$\frac{1}{|u|}$
|u||v|
$\frac{1}{u}$
uv

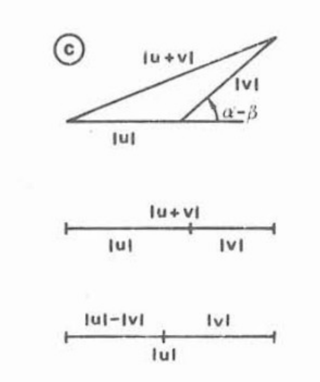
ⓒ
|u+v|
|v|
α−β
|u|
|u+v|
|u|
|v|
|u|−|v|
|v|
|u|

7/"

+,- * *. * ! 4 " /
// " " " , /"

(! ' !' !& $$!

" 0D/ A 6

MMM % /

+,- * *. * ! '!) /"

" D

(!$!

/ "

This famous differential equation is called Bessel's equation. It has two linearly independent solutions, which we will study in some detail later:

$$R = G\,J_m(\rho) + H\,N_m(\rho) = G\,J_m(\alpha r) + H\,N_m(\alpha r). \qquad (2.4.10)$$

The special case $\alpha = 0$ deserves special mention. This arises in the case of z independent solutions of Laplace's equation, for which $\alpha = k = 0$. The equation for $R(r)$ reduces to

$$\frac{d^2R}{dr^2} + \frac{1}{r}\frac{dR}{dr} - \frac{m^2}{r^2} R = 0. \qquad (2.4.11)$$

One can easily find the elementary solutions of this equation:

$$R(r) = \begin{cases} Gr^m + Hr^{-m}, & m \neq 0, \\ G + H \ln r, & m = 0. \end{cases} \qquad (2.4.12)$$

2.5 SPHERICAL COORDINATES (r, θ, φ)

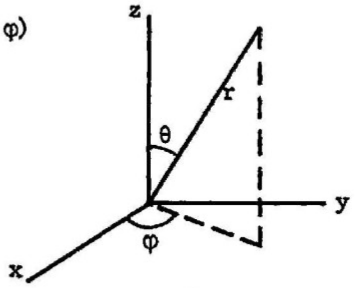

In spherical coordinates Helmholtz' equation assumes the form

$$\frac{1}{r}\frac{\partial^2}{\partial r^2}(ru) + \frac{1}{r^2 \sin\theta}\left[\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial u}{\partial\theta}\right) + \frac{1}{\sin\theta}\frac{\partial^2 u}{\partial\varphi^2}\right] + k^2 u = 0. \qquad (2.5.1)$$

We assume a separated solution of the form

$$u = R(r)\ Y(\theta,\varphi), \tag{2.5.2}$$

substitute in (2.5.1), divide by $u = RY$, and find

$$\frac{1}{R}\frac{1}{r}\frac{d^2}{dr^2}(rR) + \frac{1}{r^2}\frac{1}{Y\sin\theta}\left[\frac{\partial}{\partial\theta}(\sin\theta\frac{\partial Y}{\partial\theta}) + \frac{1}{\sin\theta}\frac{\partial^2 Y}{\partial\varphi^2}\right]$$
$$+ k^2 = 0. \tag{2.5.3}$$

The form of this equation implies

$$\frac{1}{Y}\frac{1}{\sin\theta}\left[\frac{\partial}{\partial\theta}(\sin\theta\frac{\partial Y}{\partial\theta}) + \frac{1}{\sin\theta}\frac{\partial^2 Y}{\partial\varphi^2}\right] = -\lambda \tag{2.5.4}$$

and

$$\frac{1}{R}\frac{1}{r}\frac{d^2}{dr^2}(rR) + k^2 - \frac{\lambda}{r^2} = 0, \tag{2.5.5}$$

where λ is a constant.

Consider first the radial equation (2.5.5). If $k^2 \neq 0$, let $r = \rho/k$ and rewrite the equation in the form

$$\frac{d^2R}{d\rho^2} + \frac{2}{\rho}\frac{dR}{d\rho} + (1 - \frac{\lambda}{\rho^2})R = 0. \tag{2.5.6}$$

This is almost Bessel's equation (2.4.9). To reduce it to the form of Bessel's equation, make the substitution

$$R = \frac{1}{\sqrt{\rho}}S. \tag{2.5.7}$$

After a short calculation one finds Bessel's equation for S:

$$\frac{d^2S}{d\rho^2} + \frac{1}{\rho}\frac{dS}{d\rho} + \left(1 - \frac{\lambda+\frac{1}{4}}{\rho^2}\right)S = 0. \tag{2.5.8}$$

Comparing with (2.4.9), (2.4.10) we see that this has the solution

$$S = A\,J_\beta(\rho) + B\,N_\beta(\rho) \quad , \quad \beta = \sqrt{\lambda+\tfrac{1}{4}}\,, \tag{2.5.9}$$

or equivalently

$$R = A\frac{1}{\sqrt{kr}}J_\beta(kr) + B\frac{1}{\sqrt{kr}}N_\beta(kr)\,. \tag{2.5.10}$$

For Laplace's equation, with $k^2 = 0$, the radial equation (2.5.5) is much simpler:

$$\frac{1}{r}\frac{d^2}{dr^2}(rR) - \frac{\lambda}{r^2}R = 0\,. \tag{2.5.11}$$

Assuming a solution of the form

$$R = r^\alpha, \tag{2.5.12}$$

one finds a quadratic equation for α with roots

$$\alpha = \frac{1}{2}\left(-1 \pm \sqrt{1+4\lambda}\right). \tag{2.5.13}$$

The general solution of (2.5.11) is then

$$R = Ar^{\alpha_1} + Br^{\alpha_2}, \tag{2.5.14}$$

where α_1, α_2 are the two roots (2.5.13).

Returning now to the angular equation (2.5.4), we separate again,

$$Y = P(\theta)\ \Phi(\varphi), \tag{2.5.15}$$

and obtain

$$\frac{1}{P}\frac{1}{\sin\theta}\frac{d}{d\theta}(\sin\theta\frac{dP}{d\theta}) + \frac{1}{\sin^2\theta}\frac{1}{\Phi}\frac{d^2\Phi}{d\varphi^2} + \lambda = 0, \tag{2.5.16}$$

which implies

$$\frac{1}{\Phi}\frac{d^2\Phi}{d\varphi^2} = -m^2 \tag{2.5.17}$$

and

$$\frac{1}{\sin\theta}\frac{d}{d\theta}(\sin\theta\frac{dP}{d\theta}) + (\lambda - \frac{m^2}{\sin^2\theta})P = 0. \tag{2.5.18}$$

The φ equation (2.5.17) has a simple solution

$$\Phi = Ae^{im\varphi} + Be^{-im\varphi}. \tag{2.5.19}$$

The θ equation (2.5.18) is Legendre's equation. To eliminate the trigonometric functions let $x = \cos\theta$. Equation (2.5.18) then assumes the form

$$\frac{d}{dx}\left[(1-x^2)\frac{dP}{dx}\right] + \left[\lambda - \frac{m^2}{1-x^2}\right]P = 0 \tag{2.5.20}$$

or

$$\frac{d^2P}{dx^2} - \frac{2x}{1-x^2}\frac{dP}{dx} + \frac{1}{1-x^2}\left[\lambda - \frac{m^2}{1-x^2}\right]P = 0, \tag{2.5.21}$$

which has two linearly independent solutions to be studied in detail later:

$$P = CP_\ell^m(x) + DQ_\ell^m(x) \quad , \quad \ell(\ell+1) = \lambda . \tag{2.5.22}$$

We have considered here only the three most common coordinate systems. Evidently the method can be employed in the case of the more exotic coordinate systems such as elliptic coordinates.

2.6 SERIES SOLUTIONS OF ORDINARY DIFFERENTIAL EQUATIONS: PRELIMINARIES

We have seen that the separation of variables procedure reduces a partial differential equation to ordinary differential equations. In some cases the ordinary differential equations were trivial. In sections 2.2 - 2.5 we came across two nontrivial cases: Bessel's equation (2.4.9) and Legendre's equation (2.5.21). For these and other cases we need a more general method of obtaining a solution.

We shall study the general class of second order linear differential equations of the form

$$\frac{d^2u}{dx^2} + p(x)\frac{du}{dx} + q(x)\,u = 0\,. \tag{2.6.1}$$

We restrict ourselves to the case of a real variable x and assume that in some region in x of interest p(x) and q(x) are analytic functions except for isolated poles. When we say that a function f(x) is analytic at a point x_o we shall mean that it can be expanded in a Taylor series

$$f(x) = \sum_{n=0}^{\infty} \frac{1}{n!} f^{(n)}(x)(x - x_o)^n \tag{2.6.2}$$

+,- * *. * ! 6 E + " "
/ & /"

( )!

' 4 3 3L "

7.2 The Euler Equation in Streamline Coordinates

Consider a two-dimensional ideal-fluid flow on the (y, z)-plane shown in Fig. 7.1a, in which the solid lines with arrows represent streamlines. At a specific point of a stream line, the direction $\boldsymbol{s}$ is defined as the tangential direction of the streamline at that point. The direction $\boldsymbol{n}$ is perpendicular to $\boldsymbol{s}$ and points outward. The coordinate system spanned by $\{\boldsymbol{s}, \boldsymbol{n}\}$ is termed the *streamline coordinate system*. Taking inner product of the Euler equation with two infinitesimal vectors $\mathrm{d}\boldsymbol{s}$ and $\mathrm{d}\boldsymbol{n}$ yields

$$\begin{aligned} \mathrm{d}\boldsymbol{s} \cdot \left(\frac{\partial \boldsymbol{u}}{\partial t} + (\boldsymbol{u} \cdot \nabla)\boldsymbol{u} = -\frac{1}{\rho}\nabla p + \rho \boldsymbol{b} \right), &\quad \longrightarrow \quad u_s \frac{\partial u_s}{\partial s} = -\frac{1}{\rho}\frac{\partial p}{\partial s} - g\frac{\partial z}{\partial s}, \\ \mathrm{d}\boldsymbol{n} \cdot \left(\frac{\partial \boldsymbol{u}}{\partial t} + (\boldsymbol{u} \cdot \nabla)\boldsymbol{u} = -\frac{1}{\rho}\nabla p + \rho \boldsymbol{b} \right), &\quad \longrightarrow \quad \frac{u_s^2}{R} = \frac{1}{\rho}\frac{\partial p}{\partial n} + g\frac{\partial z}{\partial n}, \end{aligned} \tag{7.2.1}$$

if the gravitational acceleration g is the only body force, and the steady-flow assumption is used, where u_s and u_n are respectively the velocity components in the s- and n-directions, and R denotes the radius of curvature at the evaluation point on the streamline. Equation $(7.2.1)_2$ is obtained by the fact that the acceleration a_n that is experienced by a fluid element at the point is the inverse of centrifugal acceleration given by $a_n = -u_s^2/R$.

---

[2] Daniel Bernoulli, 1700–1782, a Swiss mathematician and physicist, who was one of the many prominent mathematicians in the *Bernoulli family*, with his main contributions in mathematics, mechanics, fluid mechanics, probability, and statistics.

[3] Although the term $\nabla\phi \cdot \nabla\phi$ in Eq. (7.1.4) is nonlinear, it places no difficulty in the analysis.

equally be determined in a simpler manner. The *Bernoulli equation*,[2] to be discussed in Sect. 7.3, is an integration form of the Euler equation give by

$$\frac{\partial \phi}{\partial t} + \frac{p}{\rho} + \frac{1}{2}\nabla\phi \cdot \nabla\phi - G = F(t), \tag{7.1.4}$$

which is shown here to demonstrate the solving procedure, where G is the potential function of the conservative body forces and $F(t)$ represents the *unsteady Bernoulli constant*. The pressure field can be determined by using this equation if ϕ is determined. Consequently, for ideal-fluid flows, instead of solving the equations of hydrodynamics directly, the pressure and velocity fields can be obtained in a simpler manner by using Eqs. (7.1.3) and (7.1.4). The study of ideal-fluid flows by using this simpler solving procedure is termed the *potential-flow theory*. When compared to the hydrodynamic equations, the features of potential-flow theory are twofold: First, Eq. (7.1.4) is linear, whereas Eq. $(7.1.1)_2$ is nonlinear.[3] Second, the *principle of superposition* can be used for linear equations to superimpose simple solutions to obtain solutions to complex circumstances. This latter feature will be used extensively in the analyses of two- and three-dimensional potential flows.

7.2 The Euler Equation in Streamline Coordinates

Consider a two-dimensional ideal-fluid flow on the (y, z)-plane shown in Fig. 7.1a, in which the solid lines with arrows represent streamlines. At a specific point of a stream line, the direction $\boldsymbol{s}$ is defined as the tangential direction of the streamline at that point. The direction $\boldsymbol{n}$ is perpendicular to $\boldsymbol{s}$ and points outward. The coordinate system spanned by $\{\boldsymbol{s}, \boldsymbol{n}\}$ is termed the *streamline coordinate system*. Taking inner product of the Euler equation with two infinitesimal vectors $\mathrm{d}\boldsymbol{s}$ and $\mathrm{d}\boldsymbol{n}$ yields

$$\begin{aligned} &\mathrm{d}\boldsymbol{s} \cdot \left(\frac{\partial \boldsymbol{u}}{\partial t} + (\boldsymbol{u}\cdot\nabla)\boldsymbol{u} = -\frac{1}{\rho}\nabla p + \rho\boldsymbol{b}\right), \quad \longrightarrow \quad u_s\frac{\partial u_s}{\partial s} = -\frac{1}{\rho}\frac{\partial p}{\partial s} - g\frac{\partial z}{\partial s}, \\ &\mathrm{d}\boldsymbol{n} \cdot \left(\frac{\partial \boldsymbol{u}}{\partial t} + (\boldsymbol{u}\cdot\nabla)\boldsymbol{u} = -\frac{1}{\rho}\nabla p + \rho\boldsymbol{b}\right), \quad \longrightarrow \quad \frac{u_s^2}{R} = \frac{1}{\rho}\frac{\partial p}{\partial n} + g\frac{\partial z}{\partial n}, \end{aligned} \tag{7.2.1}$$

if the gravitational acceleration g is the only body force, and the steady-flow assumption is used, where u_s and u_n are respectively the velocity components in the s- and n-directions, and R denotes the radius of curvature at the evaluation point on the streamline. Equation $(7.2.1)_2$ is obtained by the fact that the acceleration a_n that is experienced by a fluid element at the point is the inverse of centrifugal acceleration given by $a_n = -u_s^2/R$.

---

[2] Daniel Bernoulli, 1700–1782, a Swiss mathematician and physicist, who was one of the many prominent mathematicians in the *Bernoulli family*, with his main contributions in mathematics, mechanics, fluid mechanics, probability, and statistics.

[3] Although the term $\nabla\phi \cdot \nabla\phi$ in Eq. (7.1.4) is nonlinear, it places no difficulty in the analysis.

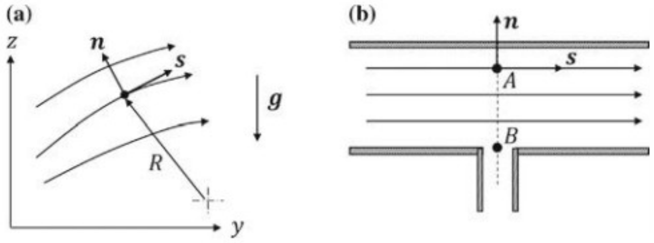

Fig. 7.1 Euler equation in a streamline coordinate system. **a** Illustration of the streamline coordinates. **b** Equivalence between the pressures at point A and point B in view of the Euler equation

The implications of Eq. (7.2.1) are straightforward. For example, consider a steady ideal-fluid flow along a horizontal streamline, for which the Euler equation in the s-direction reduces to

$$u_s \frac{\partial u_s}{\partial s} = -\frac{1}{\rho} \frac{\partial p}{\partial s}. \tag{7.2.2}$$

This equation indicates that a negative pressure gradient in the s-direction is required to have a positive velocity increase and vice versa. On the other hand, if the pressure is maintained as a constant along a streamline which has an elevation difference (i.e., z is not a constant), Eq. $(7.2.1)_1$ reduces to

$$u_s \frac{\partial u_s}{\partial s} = -\frac{\partial z}{\partial s}. \tag{7.2.3}$$

Similarly, in order to have a positive velocity increase, a negative elevation gradient in the s-direction is required. Without solving the Euler equation directly, Eqs. (7.2.2) and (7.2.3) deliver the important physical features of ideal-fluid flows which are summarized as follows:

- A fluid tends to flow locally from a high-pressure region to a low-pressure region.
- A fluid tends to flow locally from a high-elevation region to a low-elevation region.

For circumstances in which both elevation and pressure gradients present, the flow direction is determined by the relative significance between the pressure and gravitation forces. Although the above conclusions are obtained for ideal-fluid flows, they can equally be extended *qualitatively* for viscous fluid flows, except that the influence of viscous force needs to be taken into account.

For a steady horizontal streamline, the Euler equation in the n-direction reduces to

$$\frac{\partial p}{\partial n} = -g \frac{\partial z}{\partial n}, \tag{7.2.4}$$

which corresponds exactly to the hydrostatic equation given in Sect. 3.1. If the gravitational acceleration is further assumed to point in the x-direction (i.e., the direction perpendicular to the page), this equation is simplified to

$$\frac{\partial p}{\partial n} = 0, \tag{7.2.5}$$

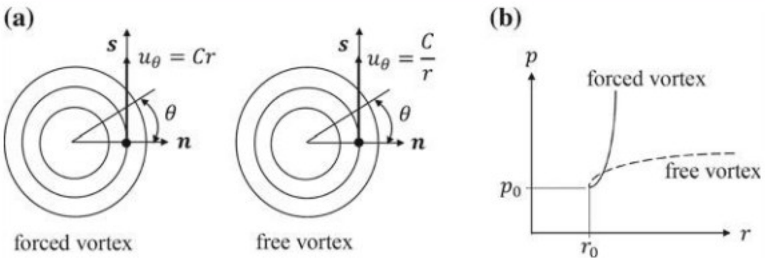

Fig. 7.2 Applications of the Euler equation in a streamline coordinate system. **a** Two-dimensional forced and free vortices. **b** The radial pressure distributions of two vortices

indicating that there exists no pressure variation in the n-direction. This equation provides the theoretical foundation of pressure measurement. For example, consider a two-dimensional flow shown in Fig. 7.1b, in which the gravitational acceleration points perpendicularly to the page. The pressure at point A is exactly the same as that at point B, which can be measured by using e.g. a manometer connected to the wall. However, caution must be made to ensure that the tube of manometer should strictly be perpendicular to the wall with completely flat connecting surfaces. If it is not the case, then at the connection region there exists a non-vanishing R, giving rise to a non-vanishing value of u_s^2/R on the left-hand-side of Eq. $(7.2.1)_2$. In this case, the pressure at point A is no longer the same as the pressure at point B.

Applications of the Euler equation in a streamline coordinate system are demonstrated in Fig. 7.2a by studying e.g. the pressure distributions of a two-dimensional forced and a free vortices, in which the gravitational acceleration points perpendicularly to the page. The pressure variations in the forced and free vortices are described by using the Euler equation in the radial direction given respectively by

$$\frac{\partial p}{\partial r} = \rho C^2 r, \qquad \frac{\partial p}{\partial r} = \frac{\rho C^2}{r^3}, \tag{7.2.6}$$

showing that the pressure variation in a forced vortex is proportional to r, while that in a free vortex is proportional to r^{-3}. Integrating these equations yields

$$p - p_0 = \frac{1}{2}\rho C^2 (r - r_0^2), \qquad p - p_0 = \frac{1}{2}\rho C^2 \left(\frac{1}{r_0^2} - \frac{1}{r^2} \right), \tag{7.2.7}$$

where the reference point is taken at $r = r_0$ with $p = p_0$, i.e., the atmospheric pressure. These results are illustrated graphically in Fig. 7.2b. Applying Eq. (7.2.7) to the free surfaces of two vortices gives rise respectively to

$$\frac{1}{2}\rho C^2 (r - r_0^2) = 0, \qquad \frac{1}{2}\rho C^2 \left(\frac{1}{r_0^2} - \frac{1}{r^2} \right) = 0, \tag{7.2.8}$$

which are the equations of free surfaces. The first equation corresponds to the free surface of a fluid in rigid rotational motion described in Sect. 3.5. Although these two equations have singular points, they serve as the simplest models to demonstrate qualitatively the features of forced and free vortices.[4]

---

[4] The singular point of a forced vortex occurs at $r \to \infty$, while that of a free vortex occurs at $r \to 0$.

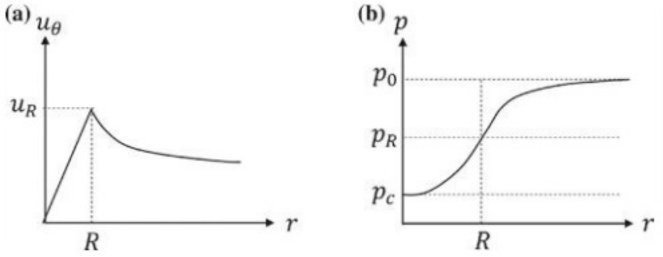

Fig. 7.3 Characteristics of a Rankine vortex in the radial direction. **a** Distribution of the tangential velocity. **b** Distribution of the thermodynamic pressure

A *Rankine vortex* is a combination of a forced vortex in the inner part and a free vortex in the outer part,[5] with the distributions of tangential velocity and pressure shown in Fig. 7.3, where p_c is the pressure at the vortex center, u_R represents the maximum tangential velocity with pressure p_R, and p_0 denotes the surrounding (atmospheric) pressure. The Rankine vortex is the simplest model to describe the features of a typhoon and can be used as a first engineering approximation to estimate the wind loads on structures due to the occurrence of a typhoon. In this circumstance, u_R is used to estimate the typhoon intensity, and R corresponds nearly to the edge of typhoon eye. It follows from the Euler equation that moist air flows from the surrounding toward the center of a typhoon, with the flow direction deflected to the right by the Coriolis force in the Northern Hemisphere, causing a typhoon to rotate counterclockwise. On the contrary, the flow direction is deflected to the left in the Southern Hemisphere, and a typhoon rotates clockwise. During the inward motion, the moist air experiences equally an upward motion to reach the top of a typhoon. Since the atmospheric temperature in high elevation is lower than that near the sea surface, it is likely possible that condensation of water vapor contained in the moist air takes place. The latent heat released by the condensation process provides an energy supply to maintain or even enhance the rotational motion of a typhoon, until all supplied energies are dissipated by the viscous and other effects.

7.3 The Bernoulli Equation

7.3.1 General Formulation

For the Newtonian fluids with constant density and dynamic viscosity, the Navier-Stokes equation with vanishing viscous force reads

$$\rho\frac{\partial \boldsymbol{u}}{\partial t} + \rho(\boldsymbol{u}\cdot\nabla)\boldsymbol{u} = -\nabla p + \rho\nabla G, \tag{7.3.1}$$

---

[5] William John Macquorn Rankine, 1820–1872, a Scottish mechanical engineer, who made contributions to various fields, and together with Rudolf Clausius and William Thomson, founded first law of thermodynamics.

where the body force per unit mass is considered a *conservative field* with its corresponding scalar potential given by G, i.e., $\boldsymbol{b} = \nabla G$. For example, the gravitational acceleration is a conservative force field, which can be determined by the gradient of gravitational potential energy per unit mass. Since

$$(\boldsymbol{u} \cdot \nabla)\boldsymbol{u} = \nabla\left(\frac{1}{2}\boldsymbol{u} \cdot \boldsymbol{u}\right) - \boldsymbol{u} \times \boldsymbol{\omega}, \tag{7.3.2}$$

where $\boldsymbol{\omega} = \nabla \times \boldsymbol{u}$, substituting this expression into Eq. (7.3.1) yields

$$\frac{\partial \boldsymbol{u}}{\partial t} + \nabla\left(\frac{1}{2}\boldsymbol{u} \cdot \boldsymbol{u}\right) - \boldsymbol{u} \times \boldsymbol{\omega} = -\frac{1}{\rho}\nabla p + \nabla G. \tag{7.3.3}$$

Taking inner product of this equation with a line element of a space curve, $\mathrm{d}\boldsymbol{\ell}$, gives

$$\frac{\partial \boldsymbol{u}}{\partial t} \cdot \mathrm{d}\boldsymbol{\ell} + \mathrm{d}\left(\frac{1}{2}\boldsymbol{u} \cdot \boldsymbol{u}\right) + \frac{\mathrm{d}p}{\rho} - \mathrm{d}G = (\boldsymbol{u} \times \boldsymbol{\omega}) \cdot \mathrm{d}\boldsymbol{\ell}, \tag{7.3.4}$$

along the tangential direction at a specific point on the space curve ℓ, where $\mathrm{d}\alpha$ denotes the total derivative of any quantity α. This equation is referred to as the *differential Bernoulli equation* or the *differential Bernoulli integral* for frictionless flow. Although Eq. (7.3.4) originates from the Euler equation, it is in fact a scalar equation denoting an energy balance, for each term represents a kind of energy. Integrating Eq. (7.3.4) results in

$$\int \frac{\partial \boldsymbol{u}}{\partial t} \cdot \mathrm{d}\boldsymbol{\ell} + \frac{1}{2}\boldsymbol{u} \cdot \boldsymbol{u} + \int \frac{\mathrm{d}p}{\rho} - G = \int (\boldsymbol{u} \times \boldsymbol{\omega}) \cdot \mathrm{d}\boldsymbol{\ell}, \tag{7.3.5}$$

which is termed the *Bernoulli equation* or the *Bernoulli integral* for frictionless flows along an arbitrary curve in space.

Several important simplifications to Eq. (7.3.5) can be made, which are explored in the following.

- For steady and incompressible flows along a streamline, Eq. (7.3.5) reduces to

  $$\frac{u^2}{2} + \frac{p}{\rho} - G = \mathcal{C}, \qquad u = \|\boldsymbol{u}\|, \tag{7.3.6}$$

  for $\boldsymbol{u}$ is in parallel with $\mathrm{d}\boldsymbol{\ell}$, giving rise to a vanishing value of $(\boldsymbol{u} \times \boldsymbol{\omega}) \cdot \mathrm{d}\boldsymbol{\ell}$. The term $\mathcal{C}$ is an integration constant, called the *Bernoulli constant*, which is different in different streamlines. Obviously, the right-hand-side of Eq. (7.3.5) vanishes equally if the flow is irrotational. For such a circumstance, $\mathrm{d}\boldsymbol{\ell}$ represents then a line element of any curve in space, which is not necessary a streamline. If the flow experiences both the gravitational acceleration $\boldsymbol{g} = \|g\|$ and a concentric acceleration $r\omega^2$ with r the radius and ω the angular speed, the scalar potential G is then given by

  $$G = -gz + \frac{r^2\omega^2}{2}, \tag{7.3.7}$$

  with which Eq. (7.3.6) becomes

  $$\frac{u^2 - r^2\omega^2}{2} + \frac{p}{\rho} + gz = \mathcal{C}, \tag{7.3.8}$$

where z is the elevation. If the fluid experiences no rotational motion, the above equation is simplified to

$$\frac{u^2}{2} + \frac{p}{\rho} + gz = \mathcal{C}, \tag{7.3.9}$$

which is the most common form of the Bernoulli equation, being valid for every point on a streamline. This equation indicates that the total mechanical energy, including the kinetic energy $u^2/2$, pressure energy p/ρ, and potential energy gz per unit mass, should be a constant through the entire streamline. From this perspective, Eq. (7.3.9) represents the simplest conservation of energy of fluid flows and provides a transformation rule for mechanical energy of steady, ideal-fluid flows along a streamline. The equivalent expressions of Eq. (7.3.9) are given by

$$\frac{u^2}{2g} + \frac{p}{\gamma} + z = \mathcal{C}, \qquad \frac{\rho u^2}{2} + p + \rho g z = \mathcal{C}. \tag{7.3.10}$$

- For steady and compressible flows along a streamline, Eq. (7.3.5) reduces to

$$\frac{u^2 - r^2\omega^2}{2} + \int \frac{\mathrm{d}p}{\rho} + gz = \mathcal{C}. \tag{7.3.11}$$

Further simplifications to this equation become possible if the relations between ρ and p are prescribed. For example, consider an ideal gas characterized by $p = C\rho^\gamma$, where C is a constant and γ denotes the specific-heat ratio. Substituting this expression into the above equation yields

$$\frac{u^2 - r^2\omega^2}{2} + \frac{\gamma}{\gamma - 1}\frac{p}{\rho} + gz = \mathcal{C}, \tag{7.3.12}$$

which is different from Eq. (7.3.8) for incompressible flows due to the amount of energy stored in pressure form. The difference becomes more obvious if Eqs. (7.3.8) and (7.3.12) are expressed in terms of the Mach number $\mathcal{M}_a$ given respectively by[6]

$$\frac{p_2 - p_1}{p_1} = \frac{\gamma(\mathcal{M}_a)_1^2}{2}; \quad \frac{p_2 - p_1}{p_1} = \frac{\gamma(\mathcal{M}_a)_1^2}{2}\left(1 + \frac{1}{4}(\mathcal{M}_a)_1^2 + \frac{2-\gamma}{24}(\mathcal{M}_a)_1^4 + \cdots\right), \tag{7.3.13}$$

between any two points 1 and 2 on a streamline without the contributions of concentric acceleration, where $(\mathcal{M}_a)_1$ represents the Mach number at point 1. For air, $\gamma = 1.4$, and it is ready to verify that the pressure energy stored in a compressible flow is larger than that in an incompressible flow. However, the difference is nearly 2% when $(\mathcal{M}_a)_1 = 0.3$. For larger values of $(\mathcal{M}_a)_1$, the disagreement between Eqs. $(7.3.13)_1$ and $(7.3.13)_2$ becomes obvious. These results imply that compressible flows can be approximated by using the theory of incompressible flows if the Mach number does not exceed 0.3.

---

[6] A more detailed discussion will be provided in Sect. 9.3.4.

Table 7.1 Different forms of the Bernoulli equation with the corresponding restrictions under conservative body forces

Form	Restrictions
$\int \frac{\partial \boldsymbol{u}}{\partial t} \cdot \mathrm{d}\boldsymbol{\ell} + \frac{1}{2}\boldsymbol{u} \cdot \boldsymbol{u} + \int \frac{\mathrm{d}p}{\rho} - G = \int (\boldsymbol{u} \times \boldsymbol{\omega}) \cdot \mathrm{d}\boldsymbol{\ell}$	F
$\frac{u^2 - r^2\omega^2}{2} + \frac{p}{\rho} + gz = \mathcal{C}$	F + I + S + L
$\frac{u^2}{2} + \frac{p}{\rho} + gz = \mathcal{C}$	F + I + S + L (gravitational field)
$\frac{u^2 - r^2\omega^2}{2} + \int \frac{\mathrm{d}p}{\rho} + gz = \mathcal{C}$	F + S + L
$\int \frac{\partial u}{\partial t}\mathrm{d}\ell + \frac{u^2 - r^2\omega^2}{2} + \frac{p}{\rho} + gz = \mathcal{C}$	F + I + L
$\frac{\partial \phi}{\partial t} + \int \frac{\mathrm{d}p}{\rho} + \frac{1}{2}\nabla\phi \cdot \nabla\phi - G = F(t)$	F + IR

F: frictionless flows, I: incompressible flows, IR: irrotational flows, S: steady flows, L: along a streamline

- For unsteady and incompressible flows along a stream line, Eq. (7.3.5) reduces to

$$\begin{aligned} &\int \frac{\partial u}{\partial t}\mathrm{d}\ell + \frac{u^2 - r^2\omega^2}{2} + \frac{p}{\rho} + gz = \mathcal{C}, \\ &\int_1^2 \frac{\partial u}{\partial t}\mathrm{d}\ell + \frac{(u_2^2 - r_2^2\omega^2) - (u_1^2 - r_1^2\omega^2)}{2} + \frac{p_2 - p_1}{\rho} + g(z_2 - z_1) = 0, \end{aligned} \tag{7.3.14}$$

  which should be evaluated between any two points 1 and 2 on a streamline.
- For unsteady and irrotational flows, the right-hand-side of Eq. (7.3.5) vanishes identically, and the equation reduces to

$$\frac{\partial \phi}{\partial t} + \int \frac{\mathrm{d}p}{\rho} + \frac{1}{2}\nabla\phi \cdot \nabla\phi - G = F(t), \tag{7.3.15}$$

  where $\boldsymbol{u} = \nabla\phi$ is used for irrotational flows with ϕ the velocity potential function, and $F(t)$ is termed the unsteady Bernoulli constant, even though it is not strictly a constant.

Table 7.1 summarizes different forms of the Bernoulli equation with the corresponding restrictions. The restriction “along a streamline” can be removed if the flow is irrotational. In that circumstance, the Bernoulli equation can be used between any two points on any curve in space.

The Bernoulli equation is frequently used to estimate the energy loss between any two points of a flow. The difference in total mechanical energies between any two points is defined as the energy loss in-between. For example, consider water flowing through a valve. Let point 1 be located before the valve and point 2 be located after the valve. The energy loss of water passing the valve per unit mass, ΔE, is then obtained as

$$\Delta E = \left(\frac{p_1}{\rho} + \frac{u_1^2}{2} + gz_1\right) - \left(\frac{p_2}{\rho} + \frac{u_2^2}{2} + gz_2\right), \tag{7.3.16}$$

for a steady, incompressible flow along a streamline with the gravitational field. It is noted that this equation only provides the general concept of energy loss. For viscous flows, it should be revised to take into account the influence of non-uniform velocity distributions in laminar and turbulent flows on the estimations on the kinetic energy. The topic will be discussed in Sect. 8.6.8.

7.3.2 Static, Dynamic, and Stagnation Pressures

The pressure in the Bernoulli equation is termed the thermodynamic pressure or static pressure. Instead of expressing the pressure as a gage one in determining the hydrostatic force on a submerged surface described in Sect. 3.2, in applying the Bernoulli equation the pressure needs to be expressed as an absolute value, for it represents a kind of energy stored in the fluid. The term $\rho u^2/2$ is called the *dynamic pressure*, which is an equivalent pressure due to the presence of fluid velocity. The sum of static and dynamic pressures is called the *stagnation pressure* p_s given by

$$p_s = p + \frac{1}{2}\rho u^2. \tag{7.3.17}$$

The stagnation pressure is interpreted as the pressure that a fluid element, initially associated with a velocity in an ideal fluid, experiences if it is brought to rest *isentropically*, in which the entropy of fluid remains unchanged. In such a process, the kinetic energy of a fluid element is converted completely to a form of pressure. For viscous-fluid flows, a moving fluid element experiences equally a pressure which is larger than the static pressure when it is brought to rest. However, this larger pressure is not the stagnation pressure, for the process is not isentropic, and only a part of the kinetic energy is converted to a form of pressure, while the other part is converted to heat due to the dissipative viscous effects.

The static and stagnation pressures of a fluid can be measured by e.g. a *Pitot tube* shown in Fig. 7.4a, which is a combination of two concentric circular tubes with the inner tube open in the front and outer tube having small open holes on the sides. The static pressure is measured at point B, while the stagnation pressure is measured at point A. It is possible to connect these two points with a regular manometer, with which the fluid velocity can be determined by

+,- * *. * ! & & $ 8

% 3" /"

($ '

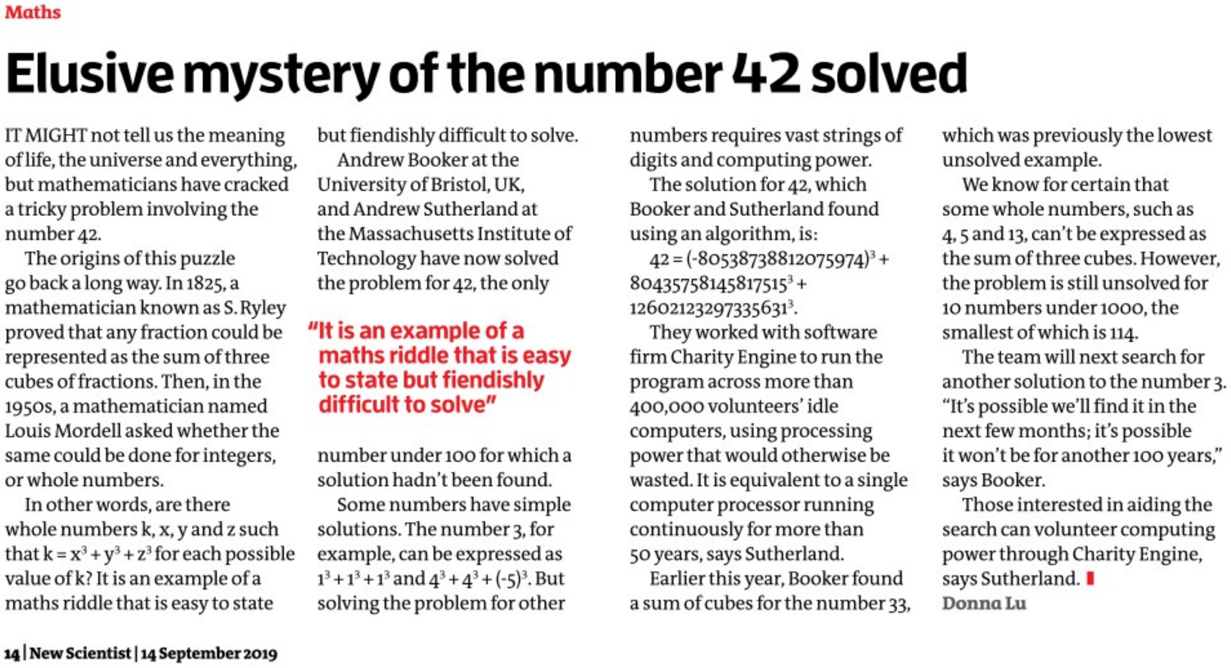

Maths

Elusive mystery of the number 42 solved

IT MIGHT not tell us the meaning of life, the universe and everything, but mathematicians have cracked a tricky problem involving the number 42.

The origins of this puzzle go back a long way. In 1825, a mathematician known as S. Ryley proved that any fraction could be represented as the sum of three cubes of fractions. Then, in the 1950s, a mathematician named Louis Mordell asked whether the same could be done for integers, or whole numbers.

In other words, are there whole numbers k, x, y and z such that $k = x^3 + y^3 + z^3$ for each possible value of k? It is an example of a maths riddle that is easy to state but fiendishly difficult to solve.

Andrew Booker at the University of Bristol, UK, and Andrew Sutherland at the Massachusetts Institute of Technology have now solved the problem for 42, the only number under 100 for which a solution hadn't been found.

"It is an example of a maths riddle that is easy to state but fiendishly difficult to solve"

Some numbers have simple solutions. The number 3, for example, can be expressed as $1^3 + 1^3 + 1^3$ and $4^3 + 4^3 + (-5)^3$. But solving the problem for other numbers requires vast strings of digits and computing power.

The solution for 42, which Booker and Sutherland found using an algorithm, is:

$42 = (-80538738812075974)^3 + 80435758145817515^3 + 12602123297335631^3$.

They worked with software firm Charity Engine to run the program across more than 400,000 volunteers' idle computers, using processing power that would otherwise be wasted. It is equivalent to a single computer processor running continuously for more than 50 years, says Sutherland.

Earlier this year, Booker found a sum of cubes for the number 33, which was previously the lowest unsolved example.

We know for certain that some whole numbers, such as 4, 5 and 13, can't be expressed as the sum of three cubes. However, the problem is still unsolved for 10 numbers under 1000, the smallest of which is 114.

The team will next search for another solution to the number 3. "It's possible we'll find it in the next few months; it's possible it won't be for another 100 years," says Booker.

Those interested in aiding the search can volunteer computing power through Charity Engine, says Sutherland. ▮

Donna Lu

14 | New Scientist | 14 September 2019

(!$!'

/ 4 7 B

B " " " K

Woven from weirdness

The true origins of space-time, the backdrop to reality, are hidden in the quantum realm, writes physicist **Sean Carroll**

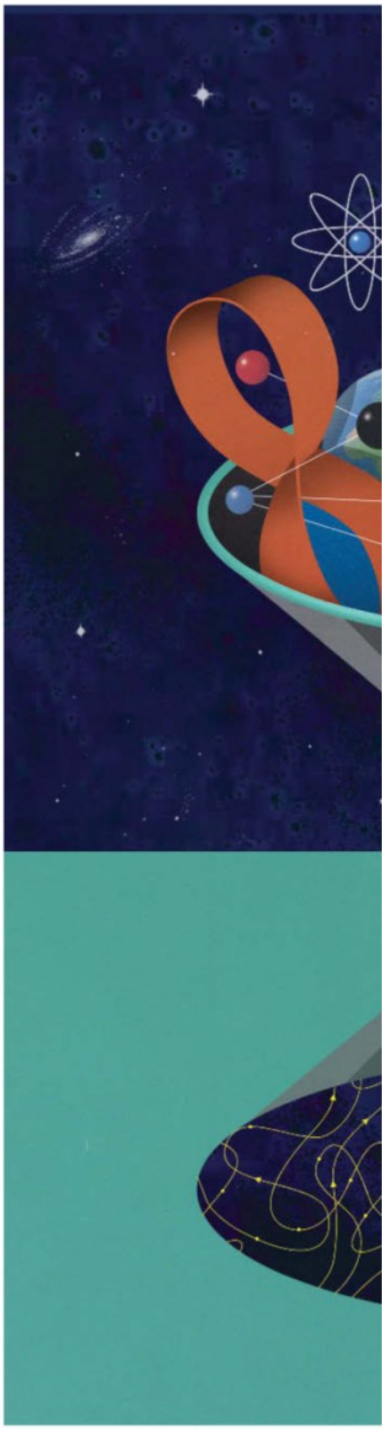
BRETT RYDER

LET'S say you want to meet a friend for coffee. You have to tell them where you are going to be – your location in space – but you also need to let them know when. Both bits of information are necessary because we live in a four-dimensional continuum: three-dimensional space and everything within it, from steaming coffee machines to stars exploding in faraway galaxies, all happening at different moments of one-dimensional time.

"Space-time" is simply the physical universe inside which we and everything else exists. And yet, even after millennia living in it, we still don't know what space-time actually is. Physicists have strived to work it out for more than a century. In recent years, many of us have been trying to figure out what might be the threads from which the fabric of reality is woven. We have ideas, each with its own selling points and shortcomings. But for my money, the most exciting one is the most surprising.

It is the idea that space-time emerges from a weird property of the quantum world that means particles and fields, those fundamental constituents of nature, can be connected even if they are at opposite ends of the universe. If that is correct, we might finally have found a bridge between the two irreconcilable totems of physics, placing us on the threshold of a theory of quantum gravity. We would also have the most startling demonstration yet that the world we see isn't the world as it is – that there is always "something deeply hidden", as Albert Einstein put it – and that the only way to understand the fundamental nature of reality is by confronting quantum mechanics head-on.

Space-time is a relatively new notion. Isaac Newton had no need for it. For him, space and time were individually real and absolute. Only when Einstein formulated his special theory of relativity in 1905 did the two start to come together. He showed that different observers will generally divide space-time into "space" and "time" in different, incompatible ways; what is "space" and what is "time" are relative to how an observer is moving.

Various thinkers had previously speculated that the two should be rolled together. In Edgar Allan Poe's 1848 prose poem *Eureka*, for instance, he wrote that "space and duration are one." But it wasn't until 1908 that mathematician Hermann Minkowski unified them in a scientific way. He dramatically proclaimed: "Henceforth, space for itself, and time for itself, shall completely reduce to a mere shadow, and only some sort of union of the two shall preserve independence."

Einstein was unimpressed, grumbling about "superfluous learnedness". But he eventually came round to the idea, putting the geometry

Hear the latest incredible ideas in Big Physics on the Cosmology stage at New Scientist Live, from 10 to 13 October
newscientistlive.com

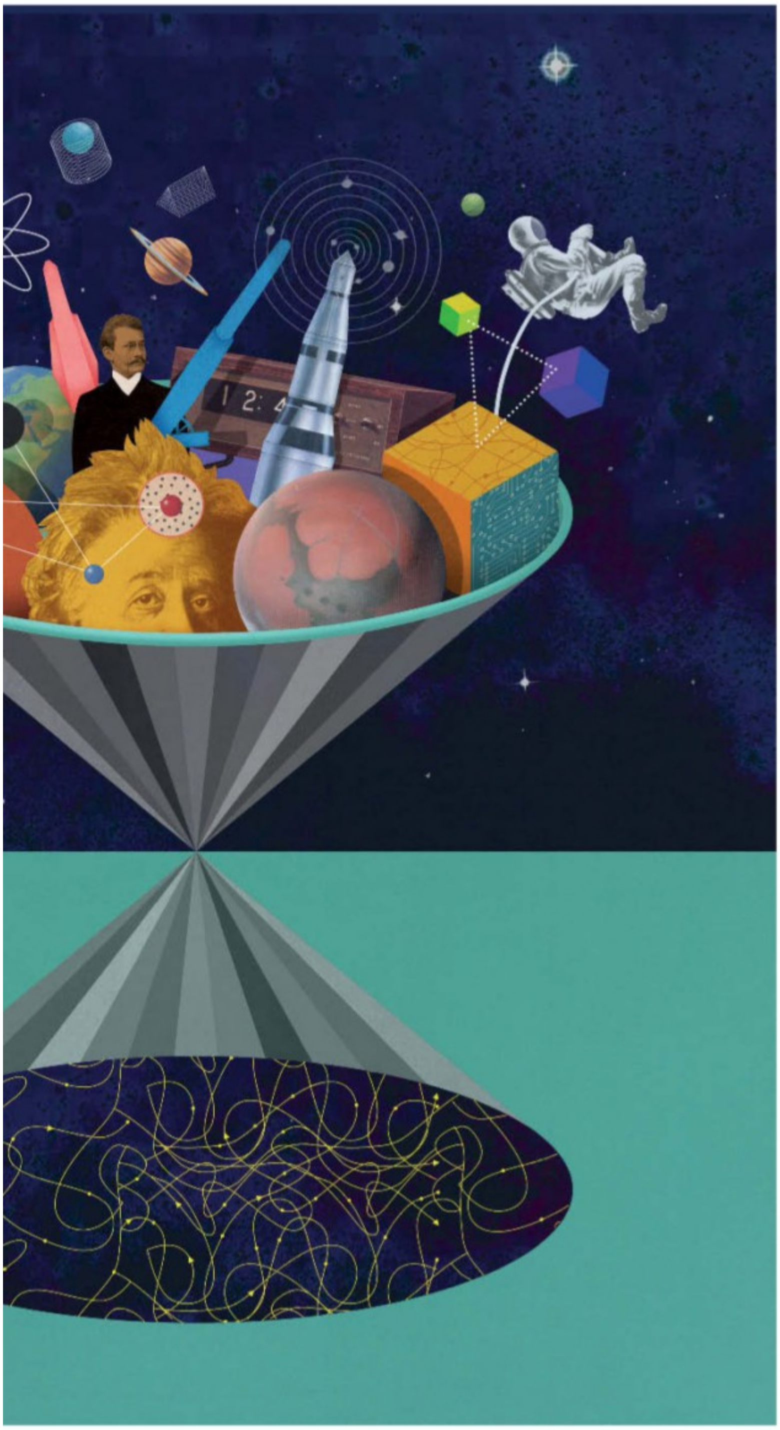

Sean Carroll is a physicist at the California Institute of Technology. His most recent book is *Something Deeply Hidden: Quantum Worlds and the Emergence of Spacetime*

"How in the world can space-time exist in a superposition of various possibilities?"

of space-time firmly at the heart of his general relativity. It said that space-time isn't merely a static background in which things happen. It is a dynamic entity, warping and stretching under the influence of mass and energy. The curvature of space-time manifests itself to us as the force of gravity.

Still, it would seem weird to ask what space-time is "made of" in classical physics. In general relativity, space-time changes over time in response to other stuff. But it is still a background, and a fundamental constituent of nature. It isn't made of anything; it just "is".

The problems with that view started with the discovery of quantum mechanics, the rules that govern the behaviour of subatomic particles and fields. Scientists haven't been able to construct a quantum-mechanical theory of gravity as they have for the three other fundamental forces of nature. Part of the issue is technical: when we try to make classical general relativity into a quantum-mechanical theory using standard techniques, our equations blow up and we get nonsensical answers. But part of it is conceptual.

Quantum mechanics tells us that systems exist in superpositions of different measurable quantities like position and velocity. There is no such thing as "the position" of a quantum particle; there are many possible positions, ›

ENTANGLED TIME

In the quest to figure out what lies behind the backdrop to reality we call space-time, we have begun to grasp how the space part can emerge from quantum entanglement (see main story). Time is a different story. But there is one way to derive the fourth dimension from the same phenomenon.

It was suggested back in 1983 by Don Page, now at the University of Alberta in Canada, and William Wootters at Williams College in Williamstown, Massachusetts. In quantum mechanics, if a system can be in various different states, we can add those states together in any combination to create new states, superpositions of the originals. An electron, for example, can be spinning clockwise or counterclockwise, but it can also be in a superposition of both.

With that in mind, consider a quantum system consisting of two subsystems: one is a clock and the other is everything else. Let the system as a whole evolve through time, so that the clock reads differently at each moment. Now take a series of such moments, say one per second, and add together all the specific quantum states at all the moments—all the ways the world actually was at each reading of the clock.

This gives you a new super-state, a superposition of individual states with specific clock readings and specific configurations of everything else. It doesn't evolve with time. But because this is a quantum system, the clock is entangled with the rest of the world. And if we were to measure the clock to see what it read, the rest of the system would instantly snap into whatever quantum state the original system had at that corresponding time.

In this way, time can appear to emerge even in an unchanging quantum state. The key is entanglement; all we need is a clock subsystem that is entangled with the rest of the universe in the right way. Time is just what your clock reads.

which take on definite values only when we observe them. How in the world can space-time exist in a superposition of different possibilities? That would make it impossible to say for sure that a certain event happened at a definite location in space and time.

Physicists of different persuasions have taken different approaches to constructing a solution in the form of a theory of quantum gravity. The most popular is string theory, which replaces particles with loops and segments of vibrating string. String theory successfully produces a quantum version of gravity, but not one that connects with our world in an obvious way. Nor does it resolve those fundamental conceptual problems. String theory's leading rival, loop quantum gravity, is an attempt to directly quantise general relativity. Loop proponents typically take the conceptual challenges of quantum gravity more seriously than their stringy colleagues, but the challenges remain.

This has led some physicists to take a step back and ask the question in a different way. The standard approach to developing a quantum description of some phenomenon, like the electromagnetic field or a collection of atoms, is to start with a classical description and then "quantise" it. That approach has failed again and again when it comes to gravity and space-time. It also isn't how nature works. The real world doesn't start classically and then somehow quantise. It is quantum from the start, and the classical world emerges as an approximation.

So maybe we shouldn't be trying to quantise gravity at all. Perhaps we should instead formulate a quantum theory from the start, and then show how classical space-time emerges from that. It is a new approach that has dramatic consequences for how we think about what space-time itself is made of.

Spooky action

To make progress in this direction, it is helpful to start with our current best physical theory, which is quantum field theory. According to this theory, the fundamental ingredients of the world are fields, such as the electric and magnetic fields. Even particles like electrons and quarks are simply vibrations

"Maybe it was a mistake to quantise gravity, and space-time was lurking in quantum mechanics all along"

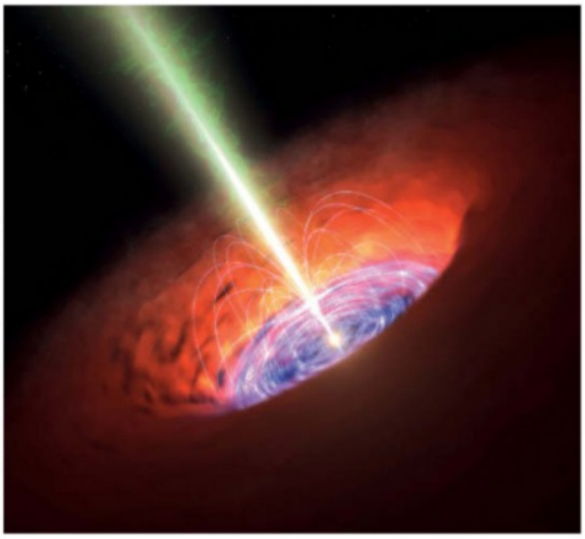

L. CALCADA/EUROPEAN SOUTHERN OBSERVATORY/SCIENCE PHOTO LIBRARY

Black holes give us reason to think space can emerge like a hologram

in fields that stretch through space.

Classically, we can specify the value of a field in an approximate fashion by dividing space into tiny regions, then listing the value of the field in each region. Once we graduate to quantum field theory, an extra feature comes into the game: the values of the field in different regions can be entangled with each other. Due to quantum uncertainty, we don't know exactly what answer we will get if we measure the field at some location, but entanglement means the answer we get at one point will affect what we would measure at any other point.

In the vacuum state of an ordinary quantum field theory – empty space, no particles flying around – the entanglement between fields in different regions is directly tied to the distance between them, and therefore to the geometry of space-time. Nearby regions are highly entangled with each other, while faraway regions share little entanglement.

This suggests an intriguing way to reverse our normal way of thinking and so find space-time within quantum theory. Let us imagine starting with just a quantum state, no pre-existing notion of space-time. Now we can try to work backwards, to extract space-time from entanglement.

If in ordinary physics the entanglement between two regions goes down as the regions get further apart, let us imagine defining the distance as (inversely) related to the entanglement. In that case, having a quantum state automatically gives us the "distance" between any two parts of it, and therefore defines a geometry on this emergent space.

So far so good. But a quantum state exists at each moment of time, so at best it can define the geometry of space at that moment. We want to extend this to four-dimensional space-time.

Thankfully, here we can borrow a trick from physicist Ted Jacobson at the University of Maryland, who, in 1995, showed how we could derive Einstein's equation for general relativity from simple assumptions about the relationship between entropy and geometry.

Entropy, a measure of disorder, is directly related to entanglement: the more entangled a region is with the rest of the world, the more entropy it contains. Einstein said that it is adding matter or energy to a region that causes space-time to curve. Jacobson showed that increasing the entanglement of a region can have the same effect, if we insist that the amount of entropy must be proportional to the area bounding that region. That is automatically true in empty space, but Jacobson suggested that it remains true even when space isn't empty. You can try to add more entanglement, but space-time will bend to compensate, so that entropy always remains proportional to area.

Quantum gum

Space-time might be woven from quantum entanglement

In a model universe, the equations describing gravity in a volume of space are equivalent to those describing the surface of that volume, which don't include gravity. This suggests space on the inside somehow emerges from the properties of the outside, namely entanglement

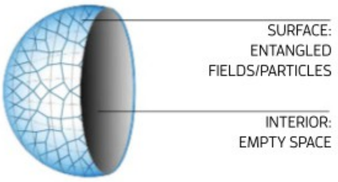

Sure enough, when you reduce the entanglement connecting two regions of the outside surface, the space inside pulls apart as if pulling at two ends of a piece of chewing gum

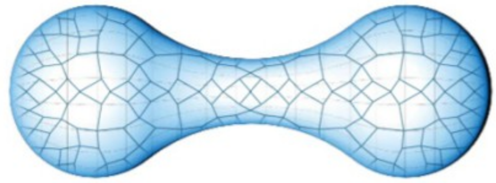

If you eliminate all entanglement, the space inside splits in two, suggesting that entanglement is the thread that binds space-time

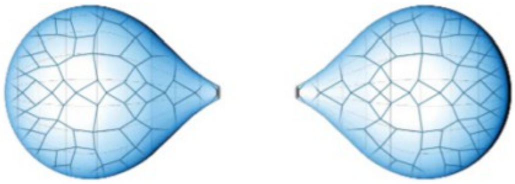

Essence of reality

So Einstein says that energy causes curvature, while Jacobson says entanglement does. But Jacobson also argued that it is really the same thing: whenever you add entanglement, energy necessarily follows. From this logic, he was able to derive that the curvature of space-time in his approach obeyed the same equation that Einstein first wrote down for general relativity. Gravity, it appears, can arise from entanglement, rather than directly from mass and energy. This remarkable result was the beginning of what is now called "thermodynamic" or "entropic" gravity.

But it doesn't quite get us to where we need to be. In deriving his alternative picture of where gravity comes from, Jacobson assumed a classical space-time and imagined that there were quantum fields living within it. Ideally, we would like to keep everything quantum from the start and derive the existence of space-time itself. This is something I recently attempted with my collaborators, ChunJun (Charles) Cao and Spiros Michalakis at the California Institute of Technology. Rather than starting with vibrating quantum fields living in space-time, we started with completely abstract quantum "degrees of freedom".

This is just some quantity that can take on different values, independently of other quantities. In Newtonian mechanics, the degrees of freedom are positions and velocities of particles; in field theory, they are the values and rates of change of the fields. In our approach, the degrees of freedom don't have any direct physical interpretation. They are the basic stuff of reality, the essence out of which everything else is made – a kind of "quantumness" that pre-exists everything. Most importantly, these quantum degrees ›

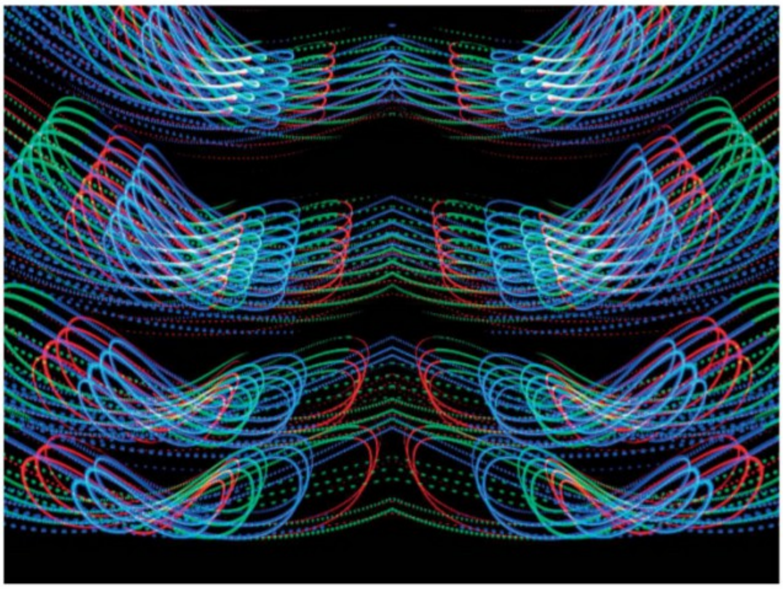
JOSE A. BERNAT BACETE/GETTY

Ethereal connections between distant quantum elements could hold the cosmos together

> **"Only with massive amounts of entanglement do we see large-scale distortions in the fabric of space"**

of freedom are entangled with each other.

With that in mind, we flip around Jacobson's idea. Now we can define the area surrounding a region as the entanglement of its degrees of freedom with the outside world. And sure enough, the corresponding geometry obeys Einstein's equation of general relativity. Gravity, in other words, can emerge directly from the quantum essence of reality, without quantising any assumed classical stuff.

That might sound like a conclusion, but it is more like a promising beginning. Many assumptions went into our derivation, and whether these assumptions hold true in nature remains to be seen. Most importantly, our derivation of Einstein's equation from entanglement only works when gravity is weak and spacetime is nearly flat. Once gravity becomes strong and space-time is curved, as in the Big Bang or near a black hole, radically new phenomena become important.

The most dramatic of these is the "holographic principle," the idea that the degrees of freedom describing a black hole can be thought of as living on its edge, the event horizon, rather than the interior. Juan Maldacena of the Institute for Advanced Study in Princeton used the holographic principle to show an equivalence between two very different theories: quantum field theory without gravity in four-dimensional space-time, and quantum gravity with a negative vacuum energy in five dimensions.

The wormhole connection

Subsequent work by Mark van Raamsdonk at the University of British Columbia in Canada and others has shown that the space-time geometry on the quantum-gravity side of this correspondence is directly tied to quantum entanglement on the field-theory side. As we decrease entanglement in the field theory, space-time on the quantum-gravity side grows apart (see "Quantum gum", page 37).

Maldacena and Leonard Susskind at Stanford University in California have taken this connection to extremes with a bold idea they dubbed "ER=EPR." ER stands for Albert Einstein and Nathan Rosen, who wrote a paper in 1935 proposing the existence of wormholes, or shortcuts through space-time. EPR, meanwhile, stands for Einstein, Boris Podolsky and Rosen, who collaborated on another paper emphasising the role of entanglement in quantum theory. The ER=EPR conjecture therefore posits that whenever you have two entangled particles, there is a tiny wormhole connecting them.

Don't take this too literally. The wormholes that purportedly connect pairs of particles would be microscopically small and impossible for anything to pass through. It is only when massive amounts of entanglement become involved that we begin to see a macroscopic distortion in the fabric of space.

Moreover, our universe has a positive vacuum energy, not a negative one, so the implications of the equivalence revealed in Maldacena's negative-vacuum-energy thought experiment don't translate directly to an actionable strategy for dealing with quantum gravity in the real world. They do, however, serve as another strong hint that quantum entanglement is at the heart of it all.

All of these ideas are, at present, somewhere between promising conjectures and optimistic dreams. We don't know the best way to think about these supposed fundamental degrees of freedom that entangle together to make space-time, nor do we know how they interact with each other in any detailed way. We can't yet derive the emergence of quantum fields living within space-time, obeying the rules of relativity. And we certainly can't answer important questions like why the energy of empty space is so small, or why space has four macroscopic dimensions.

Even so, imagining that space-time emerges from quantum entanglement is a promising way to think about the basic nature of reality. It may be that it was a mistake to start with general relativity and try to quantise it; maybe space-time was lurking within quantum mechanics all along.

And even if formulating a complete theory of quantum gravity isn't your thing, thinking about space-time this way should at least put a new slant on the familiar four-dimensional continuum in which we live, rushing around in space to be on time for coffee. ∎

% $ @# # & !)
+,- * *. * $ " 3 S 4 D
; B D 1 9 E / % A /"
('

have real propagation speeds. A complication that is always present in alternative theories of gravity is that the extra degrees of freedom are always non-minimally coupled to gravity (else there would be matter fields by definition). So, when constructing an action for a theory with a given field content it is nontrivial to judge whether it will satisfy the stability criteria mentioned above. As a result, one of the first calculations one does in every alternative theory of gravity is to check if all excitations satisfy these criteria around flat space (or some maximally symmetric space—the vacuum solution of the theory).

In Sect. 1.2.2 we mentioned that theories that lead to higher-order equations are generically plagued by instabilities. These instabilities are essentially due to the presence of ghosts. It has been shown by Ostrogradski in 1850 that non-degenerate Lagrangians with higher-order derivatives generically lead to Hamiltonians that are linear in at least one of the momenta [11]. Such Hamiltonians are not bound from below. A detailed discussion can be found in [12]. Obviously, Ostragradski's instabilities make higher-order theories particularly unappealing. However, higher-order theories which can be explicitly re-written as second-order theories with more fields evade such instabilities. We will see an example of such a theory below.

Once stability issues have been addressed, and the behaviour of the new degrees of freedom has been tamed, the next step is to find a mechanism that hides them in regimes where general relativity is well tested and no extra degrees of freedom have been seen, but still allows them to be present and lead to different phenomenology in other regimes. How challenging a task this is and how inventive we have been in order to circumvent the difficulties will be demonstrated by the examples from scalar-tensor gravity presented in the next section.

It should be mentioned that a road less taken is to consider alternative theories with non-dynamical extra degrees of freedom. In fact, one could circumvent Lovelock's theorem by considering a gravity theory where fields other than the metric are present, but they are auxiliary fields, so that they do not satisfy dynamical equation but can be instead algebraically eliminated. This way ones has the same degrees of freedom as in general relativity and does not have to worry about instabilities associated with new dynamical fields. However, such an approach is not without serious shortcomings, see [13] for a discussion and references therein. For the rest of these notes we will focus one theories with dynamical new degrees of freedom, as most popular alternative theories of gravity fall under this category.

1.3 Scalar-Tensor Gravity

1.3.1 The Prototype: Brans–Dicke Theory

The action for Brans–Dicke theory is

$$S_{\mathrm{BD}} = \frac{1}{16\pi G}\int d^4x\sqrt{-g}\left(\varphi R - \frac{\omega_0}{\varphi}\nabla^\mu\varphi\nabla_\mu\varphi - V(\varphi)\right) + S_m(g_{\mu\nu},\psi), \qquad (1.8)$$

where φ is a scalar field and ω_0 is known as the Brans–Dicke parameter. After some manipulations, the corresponding field equations can take the form

$$\begin{aligned} R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = {} & \frac{8\pi G}{\varphi} T_{\mu\nu} + \frac{\omega_0}{\varphi^2} \left(\nabla_\mu \varphi \nabla_\nu \varphi - \frac{1}{2} g_{\mu\nu} \nabla^\lambda \varphi \nabla_\lambda \varphi \right) \\ & + \frac{1}{\varphi} \left(\nabla_\mu \nabla_\nu \varphi - g_{\mu\nu} \Box \varphi \right) - \frac{V(\varphi)}{2\varphi} g_{\mu\nu} \,, \qquad (1.9) \end{aligned}$$

$$(2\omega_0 + 3) \Box \varphi = \varphi \, V' - 2V + 8\pi G \, T \,, \qquad (1.10)$$

where $\Box = \nabla^\lambda \nabla_\lambda$ and a prime denotes differentiation with respect to the argument. In its original formulation Brans–Dicke theory did not have a potential.

It is straightforward to see that in vacuo, where $T_{\mu\nu} = 0$, the theory admits solutions where with $\varphi = \varphi_0 =$ constant, provided that $\varphi_0 \, V'(\varphi_0) - 2V(\varphi_0) = 0$. For such solutions the metric actually satisfies Einstein's equations with an effective cosmological constant $V(\phi_0)$. So, one could be misled to think that, as long as $V(\varphi_0)$ has the right value, the predictions of the theory could be the same as those of general relativity. For instant, the space-time around the Sun could be described by such a solution, and then solar system constraints would be automatically satisfied. What invalidates this logic is that the $\varphi = \varphi_0$ solutions are not unique. φ could actually have a nontrivial configuration, which would also force the metric to deviate for the corresponding solution of general relativity.

This is indeed the case for spherically symmetric solution that describe the exterior of stars, and in particular the Sun. Consider for concreteness the case where $V = m^2(\varphi - \varphi_0)^2$. Performing a newtonian expansion one can calculate the newtonian limit of the metric. The perturbations of the metric are

$$h_{00} = \frac{GM_s}{\varphi_0 r} \left(1 - \frac{1}{2\omega_0 + 3} exp \left[-\sqrt{\frac{2\varphi_0}{2\omega_0 + 3}} mr \right] \right) \,, \qquad (1.11)$$

$$h_{ij} = \frac{GM_s}{\varphi_0 r} \delta_{ij} \left(1 + \frac{1}{2\omega_0 + 3} exp \left[-\sqrt{\frac{2\varphi_0}{2\omega_0 + 3}} mr \right] \right) \,, \qquad (1.12)$$

where M_s is the mass of the Sun. There is a Yukawa-like correction to the standard $1/r$ potential, with effective mass $m_{\text{eff}} = \sqrt{\frac{2\varphi_0}{2\omega_0+3}} m$ and range m_{eff}^{-1}. The ratio of the perturbations of the time-time component h_{00} over any space-space diagonal component $h_{ij}|_{i=j}$, which is also known as the γ (Eddington) parameter is then given by [14]

$$\gamma \equiv \frac{h_{ij}|_{i=j}}{h_{00}} = \frac{2\omega_0 + 3 - exp\left[-\sqrt{\frac{2\varphi_0}{2\omega_0+3}} mr \right]}{2\omega_0 + 3 + exp\left[-\sqrt{\frac{2\varphi_0}{2\omega_0+3}} mr \right]} \,. \qquad (1.13)$$

It is clear that in order for γ to be close to 1, which is the value it has in general relativity, either ω_0 or $m_{\rm eff}$ should be very large. Indeed, in the limit where $\omega_0 \to \infty$ or $m \to \infty$ the equation imply that $\varphi \to \varphi_0$ and the constant φ solutions with $g_{\mu\nu}$ satisfying Einstein's equation become unique. Current constraints on γ require that $\gamma - 1 = (2.1 \pm 2.3) \times 10^{-5}$ [15]. For $m = 0$, this constraint would require ω_0 to be larger than 40,000, which would make the theory indistinguishable from general relativity at all scales. For $\omega_0 = O(1)$, the range of the Yukawa correction would have to be below the smaller scale we have currently tested the inverse square law, i.e. a few microns. But if this is indeed the case, then this correction will never play a role at large scales.

The main message here is that weak gravity constraints are very powerful. It seems very hard to satisfy them and still have a theory whose phenomenology differs from that of general relativity at scales where we currently test gravity. One would have to circumvent this problem in order to construct a theory which is phenomenologically interesting.

1.3.2 *Scalar-Tensor Theories*

Scalar-tensor theories are straightforward generalisations of Brans–Dicke theory in which ω_0 is promoted to a general function of φ. Their action is

$$S_{\rm st} = \frac{1}{16\pi G}\int d^4x\sqrt{-g}\Big(\varphi R - \frac{\omega(\varphi)}{\varphi}\nabla^{\mu}\varphi\nabla_{\mu}\varphi - V(\varphi)\Big) + S_m(g_{\mu\nu},\psi)\,. \tag{1.14}$$

This is the most general action one can write for a scalar field non-minimally coupled to gravity which is second order in derivatives of the scalar. It can, therefore, be thought of as an effective field theory which captures, at some appropriate limit, the phenomenology of a more fundamental theory that contains a scalar field. The corresponding field equations are, after some manipulations

$$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} = \frac{8\pi G}{\varphi}T_{\mu\nu} + \frac{\omega(\varphi)}{\varphi^2}\left(\nabla_{\mu}\varphi\nabla_{\nu}\varphi - \frac{1}{2}g_{\mu\nu}\,\nabla^{\lambda}\varphi\nabla_{\lambda}\varphi\right)$$

$$+\frac{1}{\varphi}\left(\nabla_{\mu}\nabla_{\nu}\varphi - g_{\mu\nu}\Box\varphi\right) - \frac{V(\varphi)}{2\varphi}g_{\mu\nu}\,, \tag{1.15}$$

$$[2\omega(\varphi) + 3]\,\Box\varphi = -\omega'(\varphi)\,\nabla^{\lambda}\varphi\nabla_{\lambda}\varphi + \varphi\,V' - 2V + 8\pi G\,T\,. \tag{1.16}$$

Scalar-tensor theories have been extensively studied and we will not review them here. See [16, 17] for detailed reviews. The behaviour of the theories in the weak field limit will be no different than that of Brans–Dicke theory, though allowing ω

A *+,-* * *. * $ 6 < 3
0 "A H? ?" ? D "FB

+,- * *. * $ 6 < 3
0 "A H? ?" ? D /"

($ $!

0. Introduction

A pseudo-Riemannian metric g on a (connected) manifold M is called an *Einstein metric* if its Ricci tensor Ric and scalar curvature s satisfy the relations

$$\mathrm{Ric} = \frac{\mathrm{s}}{n} g, \quad d\mathrm{s} = 0, \quad n = \dim M. \tag{0.1}$$

0. Introduction

A pseudo-Riemannian metric g on a (connected) manifold M is called an *Einstein metric* if its Ricci tensor Ric and scalar curvature s satisfy the relations

$$\mathrm{Ric} = \frac{\mathrm{s}}{n} g, \quad d\mathrm{s} = 0, \quad n = \dim M. \tag{0.1}$$

One then also refers to the pair (M, g) as an *Einstein manifold*. Note that, since M is assumed connected, condition $d\mathrm{s} = 0$ means here that s is constant.

This chapter is an exposition of selected results concerning Einstein metrics on four-manifolds. The focus of our discussion on dimension four, with other dimensions mentioned only in passing, follows what has consistently, with relatively few exceptions, seemed to be the main areas of interest for Einstein metrics both in differential geometry and mathematical physics.

We address two main topics concerning four-dimensional Einstein metrics, often treated as completely separate. One is the question of global properties of compact Einstein 4-manifolds which are *Riemannian* in the sense that their metrics are positive definite. The other topic pertains to local properties of Einstein metrics on 4-dimensional manifolds in Riemannian as well as *Lorentzian* and *neutral* cases; the italicized terms refer to the sign patterns $-+++$ and $--++$, of which the former characterizes spacetime metrics in general relativity.

As a result, the material presented here can naturally be divided into *local* and *global* topics. The former are dealt with in most of Part I and Part IV. As for the latter, they are discussed, usually in the context of compact Riemannian manifolds, in Parts II and III.

Here is a brief outline of the contents of this article. Part I covers a variety of elementary topics, ranging from preliminary material on curvature, flatness, submanifolds, and conformal changes of metrics, through local classification theorems for curvature-homogeneous Riemannian Einstein metrics of dimension four (along with the special cases of metrics which are locally homogeneous, or locally symmetric), and also including topics such as Einstein hypersurfaces in pseudo-Euclidean spaces, Einstein metrics conformal to Kähler metrics or to Riemannian-product metrics, mobility for Einstein metrics, and potentials for Kähler metrics. Parts II and III deal with topological conditions that guarantee nonexistence (or, in some rare cases, existence) of a Riemannian Einstein metric on a given compact four-manifold. Finally, Part IV is devoted to local properties of indefinite Einstein metrics in dimension four and contains Petrov's classification of curvature types and a local classification of locally symmetric pseudo-Riemannian Einstein 4-manifolds.

For more details, see the detailed summaries at the beginning of each of the four parts, and the table of contents.

Besides the local-global distinction, another division in the material reflects four different levels involved in the presentation. One, which might be termed *basic and self-contained*, appears in Part I, and is both easily accessible to nonexperts and accompanied by full proofs. A 'nonexpert' means here someone whose background includes, but does not reach far beyond, the contents of a typical introductory graduate course in modern differential geometry offered at large US universities. Accordingly, the reader is assumed to have some familiarity with objects such as manifolds, differentiable mappings, tensors, pseudo-Riemannian metrics, and vector bundles (along with differentiable sections of, connections in, and fibrewise algebraic operations on the latter); the notions just listed will not be defined. Other concepts, although expected to be known to the reader, are briefly defined here; these include the Levi-Civita connection, the curvature and Ricci tensors, the Lie bracket of vector fields, etc.

As for Part II, it contains material which is still *basic*, but this time relies on several nonelementary results that are quoted without proofs (but, once taken for granted, make the rest of the discussion in Part II quite easy). In Part IV, the presentation level is *intermediate*, which here means that it is elementary and self-contained, but involves lengthy technical details that a nonexpert might not be willing to go into. Finally, Part III contains material on an *advanced* level in the form of brief summaries of results relying on repeated uses of techniques the details of which cannot be presented in a text of this size.

As another consequence of its limited size, the present chapter does not and cannot address all facts concerning four-dimensional Einstein metrics that are deemed important by large groups of experts. The topics left out include:

- techniques of global complex analysis used to prove existence and uniqueness results of the Calabi–Yau–Aubin type;
- moduli spaces of Einstein metrics on compact manifolds;
- the whole subject of cohomogeneity-one Einstein 4-manifolds, with the example of a U(2)-invariant Einstein metric on a compact 4-manifold discovered by Page [53], and Bérard Bergery's classification theorem [4] stating that no further examples of this type are possible;
- further topological obstructions to the existence of specific types of Einstein metrics on compact 4-manifolds, such as a theorem due to Hitchin [36] for nonpositively or nonnegatively curved Riemannian metrics, or results of Law [44] for indefinite metrics of the *neutral* sign pattern $--++$;
- a result of DeTurck and Kazdan [26] which says that every Riemannian Einstein metric is analytic in suitable local coordinates;
- an approach to Einstein metrics using normal coordinates and power-series solutions to (0.1);
- a result of Gasqui [30] stating that any "Einsteinian" algebraic curvature tensor at a point x in a manifold M is realized as the curvature at x of an Einstein metric g defined on a neighborhood of x in M, with the prescribed value $g(x)$ at x;

and many more.

This paper benefited from correspondence and conversations with many people. Some of those discussions go back more than a decade. I am particularly obliged to Andrea Sambusetti for numerous comments and helpful criticism of Section 27. I also greatly appreciate conversations I had in the mid-1980s with Richard Palais and Chuu-Lian Terng, concerning warped-product Einstein metrics, a topic mentioned in Sections 16 and 19 of this text.

PART I: BASICS

Throughout the following twenty-three sections we will usually assume that the metrics in question are both *Riemannian* (positive definite) and *four-dimensional*, unless a generhave real propagation speeds. A complication that is always present in alternative theories of gravity is that the extra degrees of freedom are always non-minimally coupled to gravity (else there would be matter fields by definition). So, when constructing an action for a theory with a given field content it is nontrivial to judge whether it will satisfy the stability criteria mentioned above. As a result, one of the first calculations one does in every alternative theory of gravity is to check if all excitations satisfy these criteria around flat space (or some maximally symmetric space—the vacuum solution of the theory).

In Sect. 1.2.2 we mentioned that theories that lead to higher-order equations are generically plagued by instabilities. These instabilities are essentially due to the presence of ghosts. It has been shown by Ostrogradski in 1850 that non-degenerate Lagrangians with higher-order derivatives generically lead to Hamiltonians that are linear in at least one of the momenta [11]. Such Hamiltonians are not bound from below. A detailed discussion can be found in [12]. Obviously, Ostragradski's instabilities make higher-order theories particularly unappealing. However, higher-order theories which can be explicitly re-written as second-order theories with more fields evade such instabilities. We will see an example of such a theory below.

Once stability issues have been addressed, and the behaviour of the new degrees of freedom has been tamed, the next step is to find a mechanism that hides them in regimes where general relativity is well tested and no extra degrees of freedom have been seen, but still allows them to be present and lead to different phenomenology in other regimes. How challenging a task this is and how inventive we have been in order to circumvent the difficulties will be demonstrated by the examples from scalar-tensor gravity presented in the next section.

It should be mentioned that a road less taken is to consider alternative theories with non-dynamical extra degrees of freedom. In fact, one could circumvent Lovelock's theorem by considering a gravity theory where fields other than the metric are present, but they are auxiliary fields, so that they do not satisfy dynamical equation but can be instead algebraically eliminated. This way ones has the same degrees of freedom as in general relativity and does not have to worry about instabilities associated with new dynamical fields. However, such an approach is not without serious shortcomings, see [13] for a discussion and references therein. For the rest of these notes we will focus one theories with dynamical new degrees of freedom, as most popular alternative theories of gravity fall under this category.

1.3 Scalar-Tensor Gravity

1.3.1 The Prototype: Brans–Dicke Theory

The action for Brans–Dicke theory is

$$S_{\mathrm{BD}} = \frac{1}{16\pi G}\int d^4x\sqrt{-g}\left(\varphi R - \frac{\omega_0}{\varphi}\nabla^{\mu}\varphi\nabla_{\mu}\varphi - V(\varphi)\right) + S_m(g_{\mu\nu},\psi), \quad (1.8)$$

where φ is a scalar field and ω_0 is known as the Brans–Dicke parameter. After some manipulations, the corresponding field equations can take the form

$$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} = \frac{8\pi G}{\varphi}T_{\mu\nu} + \frac{\omega_0}{\varphi^2}\left(\nabla_\mu\varphi\nabla_\nu\varphi - \frac{1}{2}g_{\mu\nu}\,\nabla^\lambda\varphi\nabla_\lambda\varphi\right)$$
$$+\frac{1}{\varphi}\left(\nabla_\mu\nabla_\nu\varphi - g_{\mu\nu}\Box\varphi\right) - \frac{V(\varphi)}{2\varphi}g_{\mu\nu}\,, \tag{1.9}$$
$$(2\omega_0 + 3)\Box\varphi = \varphi\, V' - 2V + 8\pi G\, T\,, \tag{1.10}$$

where $\Box = \nabla^\lambda\nabla_\lambda$ and a prime denotes differentiation with respect to the argument. In its original formulation Brans–Dicke theory did not have a potential.

It is straightforward to see that in vacuo, where $T_{\mu\nu} = 0$, the theory admits solutions where with $\varphi = \varphi_0 =$ constant, provided that $\varphi_0\, V'(\varphi_0) - 2V(\varphi_0) = 0$. For such solutions the metric actually satisfies Einstein's equations with an effective cosmological constant $V(\phi_0)$. So, one could be misled to think that, as long as $V(\varphi_0)$ has the right value, the predictions of the theory could be the same as those of general relativity. For instant, the space-time around the Sun could be described by such a solution, and then solar system constraints would be automatically satisfied. What invalidates this logic is that the $\varphi = \varphi_0$ solutions are not unique. φ could actually have a nontrivial configuration, which would also force the metric to deviate for the corresponding solution of general relativity.

This is indeed the case for spherically symmetric solution that describe the exterior of stars, and in particular the Sun. Consider for concreteness the case where $V = m^2(\varphi - \varphi_0)^2$. Performing a newtonian expansion one can calculate the newtonian limit of the metric. The perturbations of the metric are

$$h_{00} = \frac{GM_s}{\varphi_0 r}\left(1 - \frac{1}{2\omega_0 + 3}exp\left[-\sqrt{\frac{2\varphi_0}{2\omega_0 + 3}}mr\right]\right)\,, \tag{1.11}$$

$$h_{ij} = \frac{GM_s}{\varphi_0 r}\delta_{ij}\left(1 + \frac{1}{2\omega_0 + 3}exp\left[-\sqrt{\frac{2\varphi_0}{2\omega_0 + 3}}mr\right]\right)\,, \tag{1.12}$$

where M_s is the mass of the Sun. There is a Yukawa-like correction to the standard $1/r$ potential, with effective mass $m_{\text{eff}} = \sqrt{\frac{2\varphi_0}{2\omega_0+3}}m$ and range m_{eff}^{-1}. The ratio of the perturbations of the time-time component h_{00} over any space-space diagonal component $h_{ij}|_{i=j}$, which is also known as the γ (Eddington) parameter is then given by [14]

$$\gamma \equiv \frac{h_{ij}|_{i=j}}{h_{00}} = \frac{2\omega_0 + 3 - exp\left[-\sqrt{\frac{2\varphi_0}{2\omega_0+3}}mr\right]}{2\omega_0 + 3 + exp\left[-\sqrt{\frac{2\varphi_0}{2\omega_0+3}}mr\right]}\,. \tag{1.13}$$

It is clear that in order for γ to be close to 1, which is the value it has in general relativity, either ω_0 or m_{eff} should be very large. Indeed, in the limit where $\omega_0 \to \infty$ or $m \to \infty$ the equation imply that $\varphi \to \varphi_0$ and the constant φ solutions with $g_{\mu\nu}$ satisfying Einstein's equation become unique. Current constraints on γ require that $\gamma - 1 = (2.1 \pm 2.3) \times 10^{-5}$ [15]. For $m = 0$, this constraint would require ω_0 to be larger than 40,000, which would make the theory indistinguishable from general relativity at all scales. For $\omega_0 = O(1)$, the range of the Yukawa correction would have to be below the smaller scale we have currently tested the inverse square law, i.e. a few microns. But if this is indeed the case, then this correction will never play a role at large scales.

The main message here is that weak gravity constraints are very powerful. It seems very hard to satisfy them and still have a theory whose phenomenology differs from that of general relativity at scales where we currently test gravity. One would have to circumvent this problem in order to construct a theory which is phenomenologically interesting.

1.3.2 Scalar-Tensor Theories

Scalar-tensor theories are straightforward generalisations of Brans–Dicke theory in which ω_0 is promoted to a general function of φ. Their action is

$$S_{\text{st}} = \frac{1}{16\pi G} \int d^4x \sqrt{-g}\Big(\varphi R - \frac{\omega(\varphi)}{\varphi} \nabla^\mu \varphi \nabla_\mu \varphi - V(\varphi)\Big) + S_m(g_{\mu\nu}, \psi)\,. \tag{1.14}$$

This is the most general action one can write for a scalar field non-minimally coupled to gravity which is second order in derivatives of the scalar. It can, therefore, be thought of as an effective field theory which captures, at some appropriate limit, the phenomenology of a more fundamental theory that contains a scalar field. The corresponding field equations are, after some manipulations

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = \frac{8\pi G}{\varphi} T_{\mu\nu} + \frac{\omega(\varphi)}{\varphi^2} \left(\nabla_\mu \varphi \nabla_\nu \varphi - \frac{1}{2} g_{\mu\nu}\, \nabla^\lambda \varphi \nabla_\lambda \varphi \right)$$

$$+ \frac{1}{\varphi} \left(\nabla_\mu \nabla_\nu \varphi - g_{\mu\nu} \Box \varphi \right) - \frac{V(\varphi)}{2\varphi} g_{\mu\nu}\,, \tag{1.15}$$

$$[2\omega(\varphi) + 3]\, \Box \varphi = -\omega'(\varphi)\, \nabla^\lambda \varphi \nabla_\lambda \varphi + \varphi\, V' - 2V + 8\pi G\, T\,. \tag{1.16}$$

Scalar-tensor theories have been extensively studied and we will not review them here. See [16, 17] for detailed reviews. The behaviour of the theories in the weak field limit will be no different than that of Brans–Dicke theory, though allowing ω

+,- * *. * $ < KD A H + % "
; , B D " D @E A " 7 B D /"
(@ $)

known from very precise experiments that the gravitational mass is for all material bodies proportional to the inertial mass, and the factor of proportionality is a universal constant (taken to be just 1). Hence, all bodies will experience the same acceleration in a given gravitational field. Consequently, the frame of reference connected with a body falling freely in a homogeneous gravitational field is inertial.

Homogeneous gravitational fields do not exist in Nature. However, along the trajectory of a body falling freely in any real gravitational field, the gravitational force and the inertial force cancel each other at every point. Hence, if the gravitational field is continuous (which is the case in all practical instances), then, in a *sufficiently small neighbourhood* of the falling body, the inertial forces will be arbitrarily small. Given the precision ε attainable in measurements, a falling body B will, at every point of its trajectory, determine a sphere of radius δ inside which the inertial force will be smaller than ε, i.e. unmeasurable. Inside that sphere the reference system defined by the body B will thus be 'practically' inertial. It is called the **local inertial frame** of the body B. (In fact, local inertial frames are defined not by individual bodies, but by their trajectories that do not depend on mass.) It is called 'local' because it differs from the universal inertial frame postulated in Newton's theory: there is an infinity of local inertial frames. It is easy to see that two different local inertial frames will in general move with acceleration relative to each other. For example, consider two local inertial frames connected with bodies falling freely towards the Earth from opposite directions. Thus, at a large distance from a freely falling body its local inertial frame ceases to be inertial.

12.3 Trajectories of free motion in Einstein's theory

Let us recall one more conclusion of Chapter 1: since no standard for a straight line exists in Nature, it will be simpler to assume that the geometry of our space is non-Euclidean, and in that geometry the trajectories of material bodies are free-motion trajectories. What was called 'gravitational field' in Newton's theory will be a consequence of the non-Euclidean geometry in which the free motions take place. By the argument of Section 12.1, the Riemann geometries should be appropriate. They do contain the Euclidean geometry as a special case and, because of the coordinate-independent formalism they use, are consistent with the postulate of equivalence of all reference systems.

The generalisation of the notion of a straight line that exists in a Riemann space is the geodesic line. If in a Riemann space the curvature goes to zero, the geometry goes over into the Euclidean geometry and the geodesic lines go over into straight lines. Hence, the geodesic line is a natural candidate for the trajectory of a free motion. In addition, the geodesic lines have the following property:

Theorem 12.1 *For a timelike geodesic* G, *the coordinates in a Riemann space can be chosen so that the Christoffel symbols vanish along* G.

This is an additional argument that the geodesics should be taken as the trajectories of free fall. The vanishing of the Christoffel symbols means that along G the local gravitational field will be approximately zero. (It is not exactly zero because the derivatives of the Christoffel symbols will not vanish, and so the curvature will be nonzero.)

Proof: (Latin indices will label vectors; the lower case ones will run through the values 0, 1, 2, 3, whereas the upper case ones will run through the values 1, 2, 3. Tensor indices running through the values 1, 2, 3 will also be denoted by latin letters. Wherever confusion might arise, vector indices will have a hat over them.)

At a point $p_0 \in G$, we choose such a basis in the tangent space that $e_{\hat{0}}{}^{\alpha}(p_0)$ is tangent to G, and the other vectors, $e_A{}^{\alpha}(p_0)$, $A = 1, 2, 3$ are orthogonal to $e_{\hat{0}}{}^{\alpha}(p_0)$, i.e. $g_{\alpha\beta} e_{\hat{0}}{}^{\alpha} e_A{}^{\alpha}\big|_{p_0} = 0$. Then we define the bases $e_i{}^{\alpha}(p)$ in a neighbourhood of G as follows:

1. For $p \in G$, we transport $e_i{}^{\alpha}(p_0)$ parallely from p_0 to p along G.
2. For $p \notin G$ we draw through p a geodesic G'_p that intersects G orthogonally (see Fig. 12.1). Let $p' \in G$ be the point of intersection of G'_p and G. Then we transport the basis $e_i{}^{\alpha}(p')$ (already defined in point 1) to p along G'_p.

This procedure works provided that there are no singular points on G'_p. It gives a unique result provided that p is not too distant from G, otherwise the geodesics orthogonal to G might intersect each other.

The following equations hold on G:

$$e_i{}^{\alpha}{}_{;\beta}\big|_G = 0, \qquad e^i{}_{\alpha;\beta}\big|_G = 0. \tag{12.1}$$

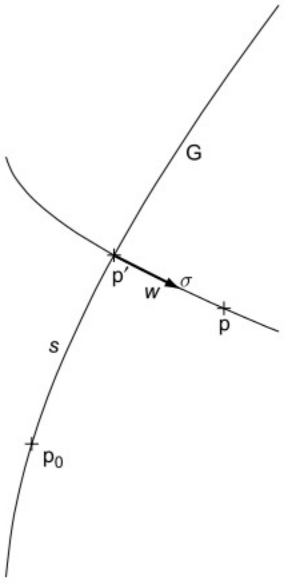

Fig. 12.1. Construction of the Fermi coordinates in which the Christoffel symbols vanish along a given timelike geodesic G. More explanation in the text.

The first equation follows because (i) $e_i{}^{\alpha}{}_{;\mu}\, e_{\widehat{0}}{}^{\mu}\big|_{\mathrm{G}} = 0$ in consequence of the basis $e_i{}^{\alpha}$ being transported parallely along G and $e_{\widehat{0}}{}^{\mu}$ being tangent to G and (ii) $e_i{}^{\alpha}{}_{;\mu} e_S{}^{\mu}\big|_{\mathrm{G}} = 0 (S = 1, 2, 3)$ in consequence of $e_S{}^{\mu}$ at G being tangent to one of the geodesics used to transplant the bases parallely to other points. The second of (12.1) follows from $e_i{}^{\alpha} e^i{}_{\beta} = \delta^{\alpha}{}_{\beta}$.

Now, using (12.1) in (4.19), we obtain

$$\left.\begin{Bmatrix} \alpha \\ \beta\gamma \end{Bmatrix}\right|_{\mathrm{G}} = e_s{}^{\alpha} e^s{}_{\beta,\gamma}\big|_{\mathrm{G}}\,. \tag{12.2}$$

Since the Christoffel symbols are symmetric, this implies at once that

$$e^i{}_{\beta,\gamma}\big|_{\mathrm{G}} = e^i{}_{\gamma,\beta}\big|_{\mathrm{G}}\,. \tag{12.3}$$

We have defined the basis (up to rotations of $e_A{}^{\alpha}(\mathrm{p}_0)$, $A = 1, 2, 3$), but we have not so far defined the coordinates. Let w^{α} be the tangent vector to G'_{p} of unit length (see Fig. 12.1), and let σ be the length of the segment of G'_{p} between p and G ($\sigma = 0$ for points on G). Let s be the length of the segment of G between p_0 and p'. For the point p we define the time coordinate $x^0 = s$ and the space coordinates

$$x^A = \sigma w^{\rho}(\mathrm{p}') e^{\widehat{A}}{}_{\rho}(\mathrm{p}'), \qquad A = 1, 2, 3. \tag{12.4}$$

In the coordinates (s, x^A) thus defined, called **Fermi coordinates**,

$$e_{\widehat{0}}{}^{\alpha}(\mathrm{p}') = \left.\frac{\partial x^{\alpha}}{\partial s}\right|_{\mathrm{G}} \underset{*}{=} \delta^{\alpha}{}_0, \tag{12.5}$$

since $e_{\widehat{0}}{}^{\alpha}$ is tangent to G, and

$$e_A{}^0(\mathrm{p}') \underset{*}{=} 0, \qquad A = 1, 2, 3, \qquad w^0 \underset{*}{=} 0 \tag{12.6}$$

because $e_A{}^{\alpha}(\mathrm{p}')$ and w^{α} lie in the subspace of the tangent space in which $x^0 = \text{constant}$. Equations (12.5) and (12.6) imply that the matrix $e_i{}^{\alpha}(\mathrm{p}')$ has a block form. Consequently, the inverse matrix $e^i{}_{\alpha}(\mathrm{p}')$ must have the same block form:

$$e^{\widehat{0}}{}_{\alpha}(\mathrm{p}') \underset{*}{=} \delta^0{}_{\alpha}, \qquad e^{\widehat{A}}_{0}(\mathrm{p}') \underset{*}{=} 0. \tag{12.7}$$

Now let us differentiate (12.4) by σ. Since $w^{\rho}(\mathrm{p}')$ and $e^{\widehat{A}}{}_{\rho}(\mathrm{p}')$ do not depend on σ, we have

$$w^A(\sigma) \overset{\mathrm{def}}{=} \frac{\mathrm{d}x^A}{\mathrm{d}\sigma} = w^{\rho}(\mathrm{p}') e^{\widehat{A}}{}_{\rho}(\mathrm{p}') \underset{*}{=} w^A(\mathrm{p}'), \tag{12.8}$$

which means that w^A does not depend on σ. In consequence of (12.6) and (12.7) this implies that $w^A(\mathrm{p}') \underset{*}{=} w^K(\mathrm{p}') e^{\widehat{A}}{}_K(\mathrm{p}')$, $K = 1, 2, 3$, so

$$e^{\widehat{A}}{}_K(\mathrm{p}') \underset{*}{=} \delta^{\widehat{A}}{}_K, \tag{12.9}$$

since $w^K(\mathrm{p}')$ is an arbitrary 3-vector at p'. From (12.7) and (12.9) we then obtain

$$e^i{}_{\alpha}(\mathrm{p}') \underset{*}{=} \delta^i{}_{\alpha}, \qquad e_i{}^{\alpha}(\mathrm{p}') \underset{*}{=} \delta^{\alpha}{}_i. \tag{12.10}$$

This implies that $e_i{}^{\alpha}(\mathrm{p}')$ is constant along G, so

$$e^i{}_{\alpha,\beta} e_{\widehat{0}}{}^{\beta}\big|_{\mathrm{G}} = \frac{\partial}{\partial s} e^i{}_{\alpha}(\mathrm{p}') \underset{*}{=} e^i{}_{\alpha,0}\big|_{\mathrm{G}} = 0, \qquad e^i{}_{0,\alpha}\big|_{\mathrm{G}} = 0. \tag{12.11}$$

(The last equation follows from the previous one by virtue of (12.3).)

Now note that

$$\frac{\mathrm{d}}{\mathrm{d}\sigma}\left(w^{\alpha} e^i{}_{\alpha}\right) \equiv \frac{\mathrm{D}}{\mathrm{d}\sigma}\left(w^{\alpha} e^i{}_{\alpha}\right) = 0, \tag{12.12}$$

because w^{α} is tangent to a geodesic on which σ is the affine parameter, while $e^i{}_{\alpha}$ is by definition parallely transported along that geodesic. Moreover, (12.8) implies that $\mathrm{d}w^{\alpha}/\mathrm{d}\sigma = 0$, hence

$$0 = \frac{\mathrm{d}}{\mathrm{d}\sigma}\left(w^{\alpha} e^i{}_{\alpha}\right) = w^{\alpha} \frac{\mathrm{d}e^i{}_{\alpha}}{\mathrm{d}\sigma} = w^{\alpha} w^{\beta} e^i{}_{\alpha,\beta} = w^A w^B e^i{}_{A,B}, \tag{12.13}$$

which means that $e^i{}_{(A,B)}\big|_{\mathrm{G}} = 0$, because along G w^A is an arbitrary vector orthogonal to G. Then, from (12.3),

$$e^i{}_{A,B}\big|_{\mathrm{G}} = 0. \tag{12.14}$$

Equations (12.11) and (12.14) imply that $e^i{}_{\alpha,\beta}\big|_{\mathrm{G}} = 0$, so, from (12.2),

$$\left.\begin{Bmatrix} \alpha \\ \beta\gamma \end{Bmatrix}\right|_{\mathrm{G}} \underset{*}{=} 0. \ \square \tag{12.15}$$

The physical meaning of (12.15) is that, in a sufficiently small neighbourhood of each $\mathrm{p}' \in \mathrm{G}$, other geodesics emanating from p′ are, up to first-order terms, approximated by straight lines. This neighbourhood is thus a local inertial frame. This is one more suggestion that geodesics should be the trajectories of free motion in relativity.

12.4 Special relativity versus gravitation theory

We have so far been dealing with the postulate that general relativity should reduce to the Newtonian kinematics of free motion in the limit of vanishing gravitational field. However, in between these extremes there is special relativity that describes the kinematics of free particles in the absence of gravitation, but takes into account velocities comparable to the velocity of light. In the previous sections we have thus been discussing a two-stage limiting transition: with the gravitational field to zero and with the velocity of light to infinity. When we switch off the gravitational field, but put no limits on the velocities, the geometric theory of gravitation should reproduce special relativity.

The spacetime of special relativity is a 4-dimensional flat Riemann space of signature $(+---)$. The Riemann space of general relativity should thus be of the same signature. This is because a change of signature means either a discontinuity in some metric components or at least one component of the metric passing through zero value, while we expect that the 'switching off' of the gravitational field can be done in a continuous way and does not lead through any singularities.

12.5 The Newtonian limit of relativity

We assumed that with the gravitational field switched off, general relativity should reproduce special relativity and, in addition, with the velocity of light becoming infinite it should reproduce the Newtonian kinematics of free motion. A logical consequence of these two postulates is the requirement that when the velocity of light is made infinite while the gravitational field is still there, general relativity should reproduce the Newtonian theory of gravitation. Hence, the field equations of relativity should, in the limit $c \to \infty$, reproduce the Poisson equation

$$\Delta\phi = 4\pi G\rho, \tag{12.16}$$

where ϕ is the gravitational potential and ρ is the density distribution of the matter generating the gravitational field.

12.6 Sources of the gravitational field

In the theory of gravitation that we are now constructing, the gravitational field should manifest its presence as non-flatness of the metric. Consequently, the metric tensor should be the device to describe gravitation. The metric tensor in four dimensions has in the most general case 10 components. Hence, we will need 10 equations to determine it.

The description of the sources in the equations should thus be correspondingly elaborate. The source in the equations generalising (12.16) should be a quantity that is a generalisation of mass-density. According to special relativity, the mass of a body depends on its energy. Hence, the energy of motion of a continuous medium should contribute to the gravitational field and so should the internal energy, e.g. pressure in a fluid and stresses in a solid.

In special relativity, the physical state of a continuous medium is described by the energy-momentum tensor $T_{\alpha\beta}$. In any chosen coordinate system, its component T_{00} is equal to the energy-density ρc^2, the components T_{0I}, $I = 1, 2, 3$, form the 3-dimensional vector of energy stream through a unit area of surface orthogonal to the direction of flow, and the components T_{ij} form the stress tensor. The tensor $T_{\alpha\beta}$ is symmetric, and so has in general 10 independent components. Consequently, it is a natural candidate for the source in the field equations of gravitation.

The description of motion of matter, i.e. the distribution of velocities inside matter, is part of the energy-momentum tensor. Consequently, the equations of motion are differential relations between various components of the energy-momentum tensor. In the Cartesian coordinates, the equations of motion are $T^{\alpha\beta}{}_{,\beta} \underset{*}{=} 0$. In any other coordinates then, these equations take the form

$$T^{\alpha\beta}{}_{;\beta} = 0. \tag{12.17}$$

The field equations determining the metric tensor in which the energy-momentum tensor should be the source must be consistent with (12.17).

The following equations hold on G:

$$e_i{}^{\alpha}{}_{;\beta}\big|_{\mathrm{G}} = 0, \qquad e^i{}_{\alpha;\beta}\big|_{\mathrm{G}} = 0. \tag{12.1}$$

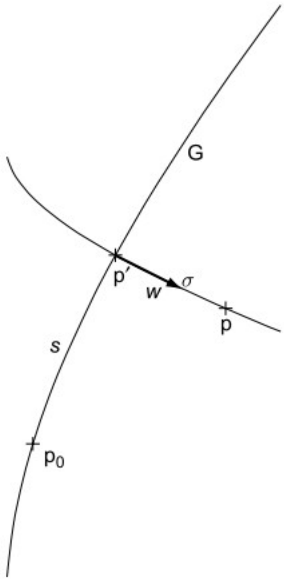

6 6 " D A B
B H " ; G/ G

" *+,-* * *. * $ 4 + D " 3 D
9 ; D D " 3 'E / 2 0 " A /"

($ @

! . D " "

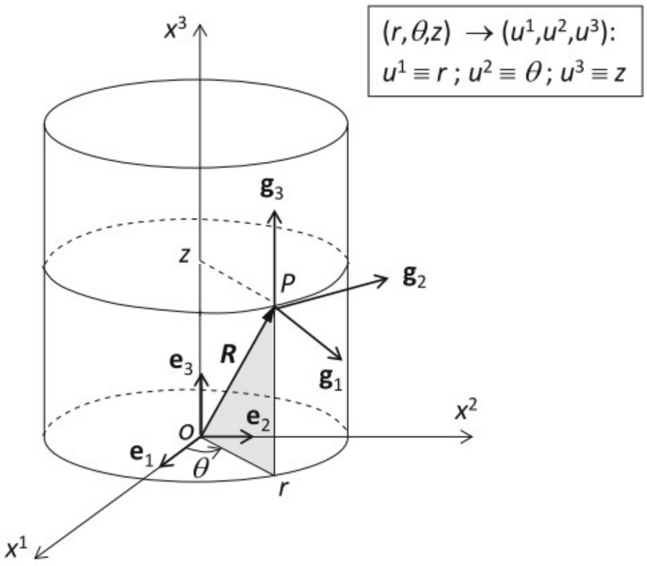

$(r,\theta,z) \rightarrow (u^1,u^2,u^3)$:
$u^1 \equiv r$; $u^2 \equiv \theta$; $u^3 \equiv z$
x^3
x^2
x^1
$\mathbf{g}_3$
$\mathbf{g}_2$
$\mathbf{g}_1$
z
P
$\boldsymbol{R}$
$\mathbf{e}_3$
$\mathbf{e}_2$
$\mathbf{e}_1$
O
θ
r

$$\mathbf{g}^k \cdot \mathbf{g}_i = \frac{\left(\mathbf{g}_i \times \mathbf{g}_j\right) \cdot \mathbf{g}_i}{\left[\mathbf{g}_i, \mathbf{g}_j, \mathbf{g}_k\right]} = \delta_i^k \tag{2.20}$$

As a result, the relation between the contravariant and covariant bases is given in the general curvilinear coordinate system $(u^1,\ldots,u^N)$.

$$\begin{cases} \mathbf{g}_i \cdot \mathbf{g}^k = \mathbf{g}^k \cdot \mathbf{g}_i = \delta_i^k \text{ for } i,k = 1,\ 2,\ \ldots,N \\ \mathbf{g}_i \cdot \mathbf{g}_k = \mathbf{g}_k \cdot \mathbf{g}_i \neq \delta_i^k \text{ for } i,k = 1,\ 2,\ \ldots,N \end{cases} \tag{2.21}$$

The basis $\mathbf{g}_i$ is called dual to the basis $\mathbf{g}^j$ [5] if

$$\mathbf{g}_i \cdot \mathbf{g}^j = \delta_i^j \text{ for } i,j = 1,\ 2,\ \ldots,N \tag{2.22}$$

where δ_i^j is the Kronecker delta.

Let $\{\mathbf{g}_1, \mathbf{g}_2,\ldots,\mathbf{g}_N\}$ be a covariant basis of the curvilinear coordinates $\{u^i\}$. The contravariant basis $\{\mathbf{g}^1, \mathbf{g}^2,\ldots,\mathbf{g}^N\}$, the dual basis to the covariant basis, can be written in the matrix formulation as

$$\mathbf{G} = \left[\mathbf{g}_1 \quad \mathbf{g}_2 \quad \cdot \quad \mathbf{g}_i \quad \mathbf{g}_N\right]; \ \mathbf{G}^{-1} = \begin{bmatrix} \mathbf{g}^1 \\ \mathbf{g}^2 \\ \cdot \\ \mathbf{g}^j \\ \mathbf{g}^N \end{bmatrix} \quad \Rightarrow \quad \mathbf{G}^{-1}\mathbf{G} = \mathbf{I} \tag{2.23}$$

where $\mathbf{g}^j$ is the j row vector of $\mathbf{G}^{-1}$; $\mathbf{g}_i$ is the i column vector of $\mathbf{G}$.

The covariant and contravariant bases (dual bases) of the orthogonal cylindrical and spherical coordinates are computed in the following section.

2.3.1.1 Orthogonal Cylindrical Coordinates

The cylindrical coordinates (r,θ,z) are orthogonal curvilinear coordinates in which the bases are mutually perpendicular but not unitary. Figure 2.3 shows a point P in the cylindrical coordinates (r,θ,z) embedded in the orthonormal Cartesian coordinates (x^1,x^2,x^3). However, the cylindrical coordinates change as the point P varies.

The vector **OP** can be written in Cartesian coordinates (x^1,x^2,x^3):

$$\begin{aligned} \mathbf{R} &= (r\cos\theta)\,\mathbf{e}_1 + (r\sin\theta)\,\mathbf{e}_2 + z\,\mathbf{e}_3 \\ &\equiv x^1\mathbf{e}_1 + x^2\mathbf{e}_2 + x^3\mathbf{e}_3 \end{aligned} \tag{2.24}$$

where

$\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$ are the orthonormal bases of Cartesian coordinates;

θ is the polar angle.

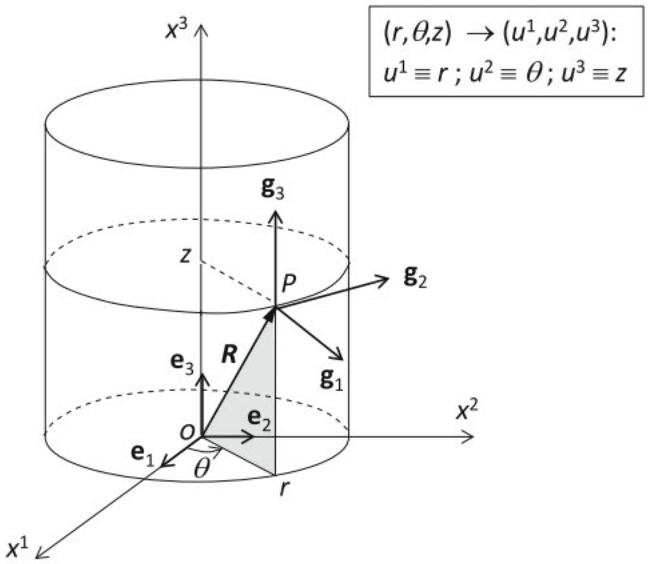
$(r,\theta,z) \rightarrow (u^1,u^2,u^3)$:
$u^1 \equiv r\,;\ u^2 \equiv \theta\,;\ u^3 \equiv z$
x^3
$\mathbf{g}_3$
z
P
$\mathbf{g}_2$
$\mathbf{e}_3$
$\boldsymbol{R}$
$\mathbf{g}_1$
x^2
O
$\mathbf{e}_2$
$\mathbf{e}_1$
θ
r
x^1

Fig. 2.3 Covariant bases of orthogonal cylindrical coordinates

To simplify the formulation with Einstein symbol, the coordinates of u^1, u^2, and u^3 are used for r, θ, and z, respectively. Therefore, the coordinates of $P(u^1,u^2,u^3)$ are given in Cartesian coordinates:

$$P\left(u^1,u^2,u^3\right)=\begin{Bmatrix} x^1 = \ r\cos\theta \equiv u^1\cos u^2 \\ x^2 = \ r\sin\theta \equiv u^1\sin u^2 \\ x^3 = z \equiv u^3 \end{Bmatrix} \tag{2.25}$$

The covariant bases of the curvilinear coordinates are computed from

$$\mathbf{g}_i = \frac{\partial \mathbf{R}}{\partial u^i} = \frac{\partial \mathbf{R}}{\partial x^j}\cdot\frac{\partial x^j}{\partial u^i} = \mathbf{e}_j\frac{\partial x^j}{\partial u^i}\ \text{for } j = 1,2,3 \tag{2.26}$$

The covariant basis matrix $\mathbf{G}$ yields from Eq. (2.26):

$$\begin{aligned}\mathbf{G} &= [\mathbf{g}_1 \quad \mathbf{g}_2 \quad \mathbf{g}_3] \\ &= \begin{pmatrix} \frac{\partial x^1}{\partial u^1} & \frac{\partial x^1}{\partial u^2} & \frac{\partial x^1}{\partial u^3} \\ \frac{\partial x^2}{\partial u^1} & \frac{\partial x^2}{\partial u^2} & \frac{\partial x^2}{\partial u^3} \\ \frac{\partial x^3}{\partial u^1} & \frac{\partial x^3}{\partial u^2} & \frac{\partial x^3}{\partial u^3} \end{pmatrix} = \begin{pmatrix} \cos\theta & -r\sin\theta & 0 \\ \sin\theta & r\cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}\end{aligned} \tag{2.27}$$

The determinant of $\mathbf{G}$ is called the Jacobian J.

$$|\mathbf{G}| \equiv J = \begin{vmatrix} \dfrac{\partial x^1}{\partial u^1} & \dfrac{\partial x^1}{\partial u^2} & \dfrac{\partial x^1}{\partial u^3} \\ \dfrac{\partial x^2}{\partial u^1} & \dfrac{\partial x^2}{\partial u^2} & \dfrac{\partial x^2}{\partial u^3} \\ \dfrac{\partial x^3}{\partial u^1} & \dfrac{\partial x^3}{\partial u^2} & \dfrac{\partial x^3}{\partial u^3} \end{vmatrix} = \begin{vmatrix} \cos\theta & -r\sin\theta & 0 \\ \sin\theta & r\cos\theta & 0 \\ 0 & 0 & 1 \end{vmatrix} = r \tag{2.28}$$

The relation between the covariant and contravariant bases yields from Eq. (2.22):

$$\mathbf{g}^i \cdot \mathbf{g}_j = \delta^i_j \quad \text{(Kronecker delta)} \tag{2.29}$$

Thus, the contravariant basis matrix $\mathbf{G}^{-1}$ results from the inversion of the covariant basis matrix $\mathbf{G}$, as given in Eq. (2.27).

$$\mathbf{G}^{-1} = \begin{bmatrix} \mathbf{g}^1 \\ \mathbf{g}^2 \\ \mathbf{g}^3 \end{bmatrix} = \frac{1}{r} \begin{pmatrix} r\cos\theta & r\sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & r \end{pmatrix} \tag{2.30}$$

The calculation of the determinant and inversion matrix of $\mathbf{G}$ will be discussed in the following section.

According to Eq. (2.27), the covariant bases can be denoted as

$$\begin{cases} \mathbf{g}_1 = (\cos\theta)\,\mathbf{e}_1 + (\sin\theta)\,\mathbf{e}_2 + 0\cdot\,\mathbf{e}_3 \Rightarrow |\mathbf{g}_1| = 1 \\ \mathbf{g}_2 = -(r\sin\theta)\,\mathbf{e}_1 + (r\cos\theta)\,\mathbf{e}_2 + 0\cdot\,\mathbf{e}_3 \Rightarrow |\mathbf{g}_2| = r \\ \mathbf{g}_3 = 0\cdot\,\mathbf{e}_1 + 0\cdot\mathbf{e}_2 + 1\cdot\mathbf{e}_3 \Rightarrow |\mathbf{g}_3| = 1 \end{cases} \tag{2.31}$$

The contravariant bases result from Eq. (2.30).

$$\begin{cases} \mathbf{g}^1 = (\cos\theta)\,\mathbf{e}_1 + (\sin\theta)\,\mathbf{e}_2 + 0\cdot\,\mathbf{e}_3 \Rightarrow \left|\mathbf{g}^1\right| = 1 \\ \mathbf{g}^2 = -\left(\dfrac{\sin\theta}{r}\right)\mathbf{e}_1 + \left(\dfrac{\cos\theta}{r}\right)\mathbf{e}_2 + 0\cdot\,\mathbf{e}_3 \Rightarrow \left|\mathbf{g}^2\right| = \dfrac{1}{r} \\ \mathbf{g}^3 = 0\cdot\mathbf{e}_1 + 0\cdot\,\mathbf{e}_2 + 1\cdot\,\mathbf{e}_3 \Rightarrow \left|\mathbf{g}^3\right| = 1 \end{cases} \tag{2.32}$$

Not only the covariant bases but also the contravariant bases of the cylindrical coordinates are orthogonal due to

$$\begin{aligned} &\mathbf{g}_i \cdot \mathbf{g}^j = \mathbf{g}^j \cdot \mathbf{g}_i = \delta^j_i \\ &\mathbf{g}_i \cdot\, \mathbf{g}_j = 0 \text{ for } i \neq j; \\ &\mathbf{g}^i \cdot\, \mathbf{g}^j = 0 \text{ for } i \neq j. \end{aligned}$$

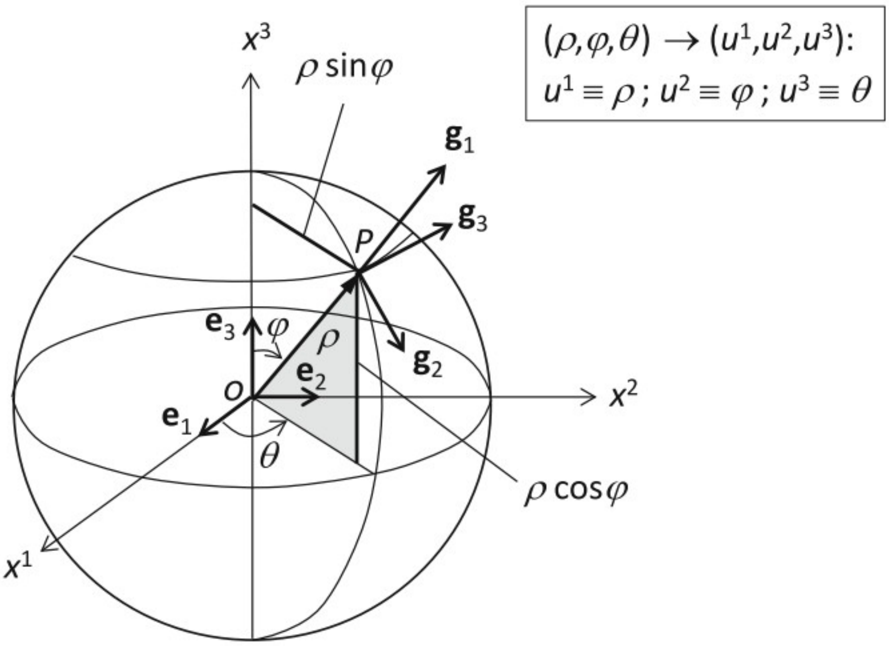
$(\rho,\varphi,\theta) \to (u^1,u^2,u^3)$:
$u^1 \equiv \rho\,;\ u^2 \equiv \varphi\,;\ u^3 \equiv \theta$
x^3
$\rho \sin\varphi$
$\mathbf{g}_1$
$\mathbf{g}_3$
P
$\mathbf{e}_3$
φ
ρ
$\mathbf{g}_2$
$\mathbf{e}_2$
O
x^2
$\mathbf{e}_1$
θ
$\rho \cos\varphi$
x^1

Fig. 2.4 Covariant bases of orthogonal spherical coordinates

2.3.1.2 Orthogonal Spherical Coordinates

The spherical coordinates (ρ,φ,θ) are orthogonal curvilinear coordinates in which the bases are mutually perpendicular but not unitary.

The spherical coordinates (ρ,φ,θ) are orthogonal curvilinear coordinates in which the bases are mutually perpendicular but not unitary. Figure 2.4 shows a point P in the spherical coordinates (r,θ,z) embedded in the orthonormal Cartesian coordinates (x^1,x^2,x^3). However, the spherical coordinates change as the point P varies.

The vector **OP** can be written in Cartesian coordinates (x^1,x^2,x^3):

$$\begin{aligned}\mathbf{R} = &\ (\rho\sin\varphi\cos\theta)\,\mathbf{e}_1 + (\rho\sin\varphi\sin\theta)\,\mathbf{e}_2 + \rho\cos\varphi\,\mathbf{e}_3 \\ \equiv &\ x^1\mathbf{e}_1 + x^2\mathbf{e}_2 + x^3\mathbf{e}_3\end{aligned} \tag{2.33}$$

where

$\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$ are the orthonormal bases of Cartesian coordinates;

φ is the equatorial angle;

θ is the polar angle.

To simplify the formulation with Einstein symbol, the coordinates of u^1, u^2, and u^3 are used for ρ, φ, and θ, respectively. Therefore, the coordinates of $P(u^1,u^2,u^3)$ are given in Cartesian coordinates:

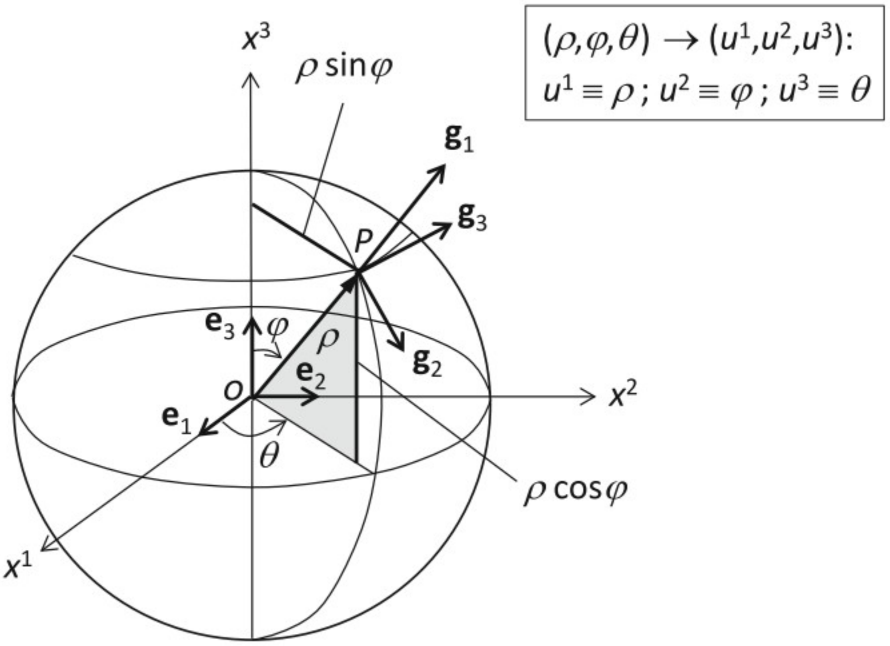
$(\rho,\varphi,\theta) \to (u^1,u^2,u^3)$:
$u^1 \equiv \rho$; $u^2 \equiv \varphi$; $u^3 \equiv \theta$
x^3
$\rho\sin\varphi$
$\mathbf{g}_1$
$\mathbf{g}_3$
P
$\mathbf{e}_3$
φ
ρ
$\mathbf{g}_2$
$\mathbf{e}_2$
O
x^2
$\mathbf{e}_1$
θ
$\rho\cos\varphi$
x^1

6 $ B A / "

B

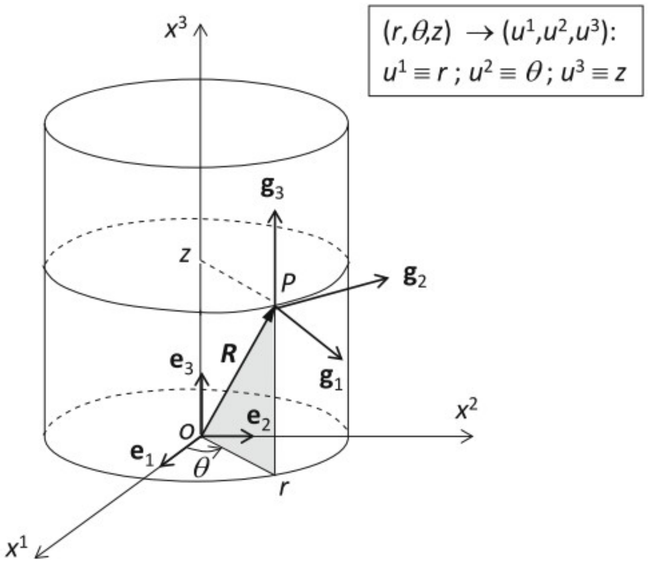
$(r,\theta,z) \to (u^1,u^2,u^3)$:
$u^1 \equiv r$; $u^2 \equiv \theta$; $u^3 \equiv z$
x^3
$\mathbf{g}_3$
z
P
$\mathbf{g}_2$
$\mathbf{e}_3$
$\boldsymbol{R}$
$\mathbf{g}_1$
x^2
O
$\mathbf{e}_2$
$\mathbf{e}_1$
θ
r
x^1

6 ! B A D " "

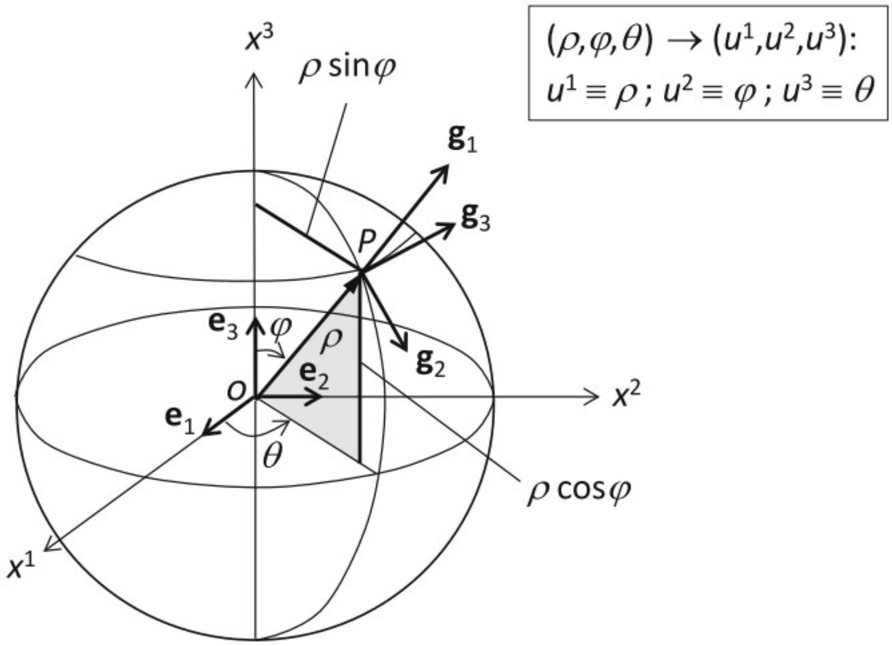
$(\rho,\varphi,\theta) \rightarrow (u^1,u^2,u^3)$:
$u^1 \equiv \rho$; $u^2 \equiv \varphi$; $u^3 \equiv \theta$
x^3
$\rho \sin\varphi$
$\mathbf{g}_1$
$\mathbf{g}_3$
P
$\mathbf{e}_3$
φ
ρ
$\mathbf{g}_2$
O
$\mathbf{e}_2$
x^2
$\mathbf{e}_1$
θ
$\rho \cos\varphi$
x^1

Fig. 2.4 Covariant bases of orthogonal spherical coordinates

2.3.1.2 Orthogonal Spherical Coordinates

The spherical coordinates (ρ,φ,θ) are orthogonal curvilinear coordinates in which the bases are mutually perpendicular but not unitary.

The spherical coordinates (ρ,φ,θ) are orthogonal curvilinear coordinates in which the bases are mutually perpendicular but not unitary. Figure 2.4 shows a point P in the spherical coordinates (r,θ,z) embedded in the orthonormal Cartesian coordinates (x^1,x^2,x^3). However, the spherical coordinates change as the point P varies.

The vector **OP** can be written in Cartesian coordinates (x^1,x^2,x^3):

$$\begin{aligned}\mathbf{R} &= (\rho \sin\varphi \cos\theta)\, \mathbf{e}_1 + (\rho \sin\varphi \sin\theta)\, \mathbf{e}_2 + \rho \cos\varphi\, \mathbf{e}_3 \\ &\equiv x^1\mathbf{e}_1 + x^2\mathbf{e}_2 + x^3\mathbf{e}_3\end{aligned} \tag{2.33}$$

where

$\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$ are the orthonormal bases of Cartesian coordinates;

φ is the equatorial angle;

θ is the polar angle.

To simplify the formulation with Einstein symbol, the coordinates of u^1, u^2, and u^3 are used for ρ, φ, and θ, respectively. Therefore, the coordinates of $P(u^1,u^2,u^3)$ are given in Cartesian coordinates:

$$P\left(u^1,u^2,u^3\right)=\left\{\begin{array}{l} x^1 = \ \rho\sin\varphi\cos\theta \equiv u^1\sin u^2\cos u^3 \\ x^2 = \ \rho\sin\varphi\sin\theta \equiv u^1\sin u^2\cos u^3 \\ x^3 = \rho\cos\varphi \equiv u^1\cos u^2 \end{array}\right\} \tag{2.34}$$

The covariant bases of the curvilinear coordinates are computed from

$$\mathbf{g}_i=\frac{\partial\mathbf{R}}{\partial u^i}=\frac{\partial\mathbf{R}}{\partial x^j}\cdot\frac{\partial x^j}{\partial u^i}=\mathbf{e}_j\frac{\partial x^j}{\partial u^i}\text{ for } j=1,2,3 \tag{2.35}$$

Thus, the covariant basis matrix **G** can be calculated from Eq. (2.35).

$$\begin{aligned}\mathbf{G}=[\,\mathbf{g}_1\quad\mathbf{g}_2\quad\mathbf{g}_3\,]&=\begin{pmatrix}\dfrac{\partial x^1}{\partial u^1} & \dfrac{\partial x^1}{\partial u^2} & \dfrac{\partial x^1}{\partial u^3}\\ \dfrac{\partial x^2}{\partial u^1} & \dfrac{\partial x^2}{\partial u^2} & \dfrac{\partial x^2}{\partial u^3}\\ \dfrac{\partial x^3}{\partial u^1} & \dfrac{\partial x^3}{\partial u^2} & \dfrac{\partial x^3}{\partial u^3}\end{pmatrix}\\ &=\begin{pmatrix}\sin\varphi\cos\theta & \rho\cos\varphi\cos\theta & -\rho\sin\varphi\sin\theta\\ \sin\varphi\sin\theta & \rho\cos\varphi\sin\theta & \rho\sin\varphi\cos\theta\\ \cos\varphi & -\rho\sin\varphi & 0\end{pmatrix}\end{aligned} \tag{2.36}$$

The determinant of the covariant basis matrix **G** is called the Jacobian J.

$$\begin{aligned}|\mathbf{G}|\equiv J&=\begin{vmatrix}\sin\varphi\cos\theta & \rho\cos\varphi\cos\theta & -\rho\sin\varphi\sin\theta\\ \sin\varphi\sin\theta & \rho\cos\varphi\sin\theta & \rho\sin\varphi\cos\theta\\ \cos\varphi & -\rho\sin\varphi & 0\end{vmatrix}\\ &=\rho^2\sin\varphi\end{aligned} \tag{2.37}$$

Similarly, the contravariant basis matrix $\mathbf{G}^{-1}$ is the inversion of the covariant basis matrix **G**.

$$\mathbf{G}^{-1}=\begin{bmatrix}\mathbf{g}^1\\ \mathbf{g}^2\\ \mathbf{g}^3\end{bmatrix}=\frac{1}{\rho}\begin{pmatrix}\rho\sin\varphi\cos\theta & \rho\sin\varphi\sin\theta & \rho\cos\varphi\\ \cos\varphi\cos\theta & \cos\varphi\sin\theta & -\sin\varphi\\ -\left(\dfrac{\sin\theta}{\sin\varphi}\right) & \left(\dfrac{\cos\theta}{\sin\varphi}\right) & 0\end{pmatrix} \tag{2.38}$$

The matrix product $\mathbf{G}\mathbf{G}^{-1}$ must be an identity matrix according to Eq. (2.23).

$$\begin{aligned}\mathbf{G}^{-1}\mathbf{G} &= \frac{1}{\rho}\begin{pmatrix} \rho\sin\varphi\cos\theta & \rho\sin\varphi\sin\theta & \rho\cos\varphi \\ \cos\varphi\cos\theta & \cos\varphi\sin\theta & -\sin\varphi \\ -\left(\dfrac{\sin\theta}{\sin\varphi}\right) & \left(\dfrac{\cos\theta}{\sin\varphi}\right) & 0 \end{pmatrix} \\ &\cdot\begin{pmatrix} \sin\varphi\cos\theta & \rho\cos\varphi\cos\theta & -\rho\sin\varphi\sin\theta \\ \sin\varphi\sin\theta & \rho\cos\varphi\sin\theta & \rho\sin\varphi\cos\theta \\ \cos\varphi & -\rho\sin\varphi & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \equiv \mathbf{I}\end{aligned} \tag{2.39}$$

According to Eq. (2.36), the covariant bases can be written as

$$\begin{aligned}\mathbf{g}_1 &= (\sin\varphi\cos\theta)\,\mathbf{e}_1 + (\sin\varphi\sin\theta)\,\mathbf{e}_2 + \cos\varphi\,\mathbf{e}_3 \Rightarrow |\mathbf{g}_1| = 1 \\ \mathbf{g}_2 &= (\rho\cos\varphi\cos\theta)\,\mathbf{e}_1 + (\rho\cos\varphi\sin\theta)\,\mathbf{e}_2 - (\rho\sin\varphi)\,\mathbf{e}_3 \Rightarrow |\mathbf{g}_2| = \rho \\ \mathbf{g}_3 &= (-\rho\sin\varphi\sin\theta)\,\mathbf{e}_1 + (\rho\sin\varphi\cos\theta)\,\mathbf{e}_2 + 0\cdot\mathbf{e}_3 \Rightarrow |\mathbf{g}_3| = \rho\sin\varphi\end{aligned} \tag{2.40}$$

The contravariant bases result from Eq. (2.38).

$$\begin{aligned}\mathbf{g}^1 &= (\sin\varphi\cos\theta)\,\mathbf{e}_1 + (\sin\varphi\sin\theta)\,\mathbf{e}_2 + \cos\varphi\,\mathbf{e}_3 \Rightarrow \left|\mathbf{g}^1\right| = 1 \\ \mathbf{g}^2 &= \left(\frac{1}{\rho}\cos\varphi\cos\theta\right)\mathbf{e}_1 + \left(\frac{1}{\rho}\cos\varphi\sin\theta\right)\mathbf{e}_2 - \left(\frac{1}{\rho}\sin\varphi\right)\mathbf{e}_3 \Rightarrow \left|\mathbf{g}^2\right| = \frac{1}{\rho} \\ \mathbf{g}^3 &= -\left(\frac{1}{\rho}\frac{\sin\theta}{\sin\varphi}\right)\mathbf{e}_1 + \left(\frac{1}{\rho}\frac{\cos\theta}{\sin\varphi}\right)\mathbf{e}_2 + 0\cdot\mathbf{e}_3 \Rightarrow \left|\mathbf{g}^3\right| = \frac{1}{\rho\sin\varphi}\end{aligned} \tag{2.41}$$

Not only the covariant bases but also the contravariant bases of the spherical coordinates are orthogonal due to

$$\begin{aligned}&\mathbf{g}_i\cdot\mathbf{g}^j = \mathbf{g}^j\cdot\mathbf{g}_i = \delta_i^j; \\ &\mathbf{g}_i\cdot\mathbf{g}_j = 0 \quad \text{for } i\neq j; \\ &\mathbf{g}^i\cdot\,\mathbf{g}^j = 0 \quad \text{for } i\neq j.\end{aligned}$$

2.3.2 Metric Coefficients in General Curvilinear Coordinates

The covariant basis vectors $\mathbf{g}_1$, $\mathbf{g}_2$, and $\mathbf{g}_3$ to the general curvilinear coordinates (u^1, u^2,u^3) at the point P can be calculated from the orthonormal bases ($\mathbf{e}_1$,$\mathbf{e}_2$,$\mathbf{e}_3$) in Cartesian coordinates $x^j = x^j(u^i)$, as shown in Fig. 2.2.

$$\begin{aligned}\mathbf{g}_i &\equiv \frac{\partial \mathbf{r}}{\partial u^i} = \frac{\partial \mathbf{r}}{\partial x^k}\frac{\partial x^k}{\partial u^i} \\ &= \mathbf{e}_k \frac{\partial x^k}{\partial u^i} \text{ for } k = 1, 2, 3\end{aligned} \tag{2.42}$$

The covariant metric coefficients g_{ij} are defined as

$$\begin{aligned}g_{ij} &\equiv \mathbf{g}_i \cdot \mathbf{g}_j = \frac{\partial \mathbf{r}}{\partial u^i}\frac{\partial \mathbf{r}}{\partial u^j} = \mathbf{g}_j \cdot \mathbf{g}_i \equiv g_{ji} \\ &= \frac{\partial x^k}{\partial u^i}\frac{\partial x^l}{\partial u^j}\mathbf{e}_k \mathbf{e}_l = \frac{\partial x^k}{\partial u^i}\frac{\partial x^l}{\partial u^j}\delta_{kl} \\ &= \frac{\partial x^k}{\partial u^i}\frac{\partial x^k}{\partial u^j}\end{aligned} \tag{2.43}$$

Similarly, the contravariant metric coefficients g^{ij} can be denoted as

$$g^{ij} \equiv \mathbf{g}^i \cdot \mathbf{g}^j = \mathbf{g}^j \cdot \mathbf{g}^i = g^{ji} \tag{2.44}$$

Furthermore, the contravariant basis can be rewritten as a linear combination of the covariant bases.

$$\mathbf{g}^i = A^{ij}\mathbf{g}_j \tag{2.45}$$

According to Eq. (2.44) and using Eqs. (2.21) and (2.45), the contravariant metric coefficients can be expressed as

$$\begin{aligned}g^{ik} &\equiv \mathbf{g}^i \cdot \mathbf{g}^k = A^{ij}\mathbf{g}_j \cdot \mathbf{g}^k \\ &= A^{ij}\delta_j^k = A^{ik}\end{aligned} \tag{2.46}$$

Thus,

$$\mathbf{g}^i = g^{ij}\mathbf{g}_j \text{ for } j = 1, 2, 3 \tag{2.47}$$

Analogously, one obtains the covariant basis

$$\mathbf{g}_k = g_{kl}\mathbf{g}^l \text{ for } l = 1, 2, 3 \tag{2.48}$$

The mixed metric coefficients can be defined by

$$\begin{aligned} g_k^i \equiv \mathbf{g}^i \cdot \mathbf{g}_k &= \left(g^{ij}\mathbf{g}_j\right) \cdot \mathbf{g}_k \\ &= \left(g^{ij}\mathbf{g}_j\right) \cdot \left(g_{kl}\mathbf{g}^l\right) = g^{ij} g_{kl} \left(\mathbf{g}_j \cdot \mathbf{g}^l\right) \\ &= g^{ij} g_{kl} \delta_j^l = g^{ij} g_{kj} \\ &= \delta_k^i \end{aligned} \tag{2.49a}$$

Thus,

$$g^{ij} g_{kj} = g_{kj} g^{ij} = \delta_k^i \tag{2.49b}$$

Therefore, the contravariant metric tensor is the inverse of the covariant metric tensor.

$$g^{ij} g_{kj} = g_{kj} g^{ij} = \delta_k^i \ \Leftrightarrow \mathbf{M}^{-1}\mathbf{M} = \mathbf{M}\mathbf{M}^{-1} = \mathbf{I} \tag{2.50}$$

where $\mathbf{M}^{-1}$ and $\mathbf{M}$ are the contravariant and covariant metric tensors.

Thus,

$$\begin{aligned} \mathbf{M}^{-1}\mathbf{M} &= \begin{bmatrix} g^{11} & g^{12} & . & g^{1N} \\ g^{21} & g^{22} & . & g^{2N} \\ . & . & g^{ij} & . \\ g^{N1} & g^{N2} & . & g^{NN} \end{bmatrix} \cdot \begin{bmatrix} g_{11} & g_{12} & . & g_{1N} \\ g_{21} & g_{22} & . & g_{2N} \\ . & . & g_{ij} & . \\ g_{N1} & g_{N2} & . & g_{NN} \end{bmatrix} \\ &= \left(g^{ij} g_{kj}\right) = \left(\delta_k^i\right) = \begin{bmatrix} 1 & 0 & . & 0 \\ 0 & 1 & . & 0 \\ 0 & . & 1 & 0 \\ 0 & 0 & . & 1 \end{bmatrix} = \mathbf{I} \end{aligned} \tag{2.51}$$

According to Eqs. (2.42) and (2.49a), the contravariant bases of the curvilinear coordinates can be derived as

$$\begin{aligned} &\delta_i^k \equiv \mathbf{g}^k \cdot \mathbf{g}_i = \mathbf{g}^k \cdot \frac{\partial x^j}{\partial u^i} \mathbf{e}_j \Rightarrow \\ &\delta_i^k \cdot \mathbf{e}_j = \left(\mathbf{g}^k \cdot \frac{\partial x^j}{\partial u^i} \mathbf{e}_j\right) \cdot \mathbf{e}_j = \mathbf{g}^k \cdot \frac{\partial x^j}{\partial u^i} \Rightarrow \\ &\mathbf{g}^k = \delta_i^k \frac{\partial u^i}{\partial x^j} \mathbf{e}_j = \frac{\partial u^k}{\partial x^j} \mathbf{e}_j \end{aligned} \tag{2.52}$$

Thus,

$$\mathbf{g}^j = \frac{\partial u^j}{\partial x^i} \mathbf{e}_i \text{ for } i = 1, 2, 3 \tag{2.53}$$

Generally, the covariant and contravariant metric coefficients of the general curvilinear coordinates have the following properties:

$$\begin{cases} g_{ji} = g_{ij} = \mathbf{g}_i \cdot \mathbf{g}_j \neq \delta_i^j \text{ (cov. metric coefficient)} \\ g^{ji} = g^{ij} = \mathbf{g}^i \cdot \mathbf{g}^j \neq \delta_i^j \text{ (contrav. metric coefficient)} \\ g_i^j = \mathbf{g}_i \cdot \mathbf{g}^j = \delta_i^j \text{ (mixed metric coefficient)} \end{cases} \tag{2.54}$$

Using Eqs. (2.42) and (2.53), one obtains the Kronecker delta

$$\begin{aligned} \delta_i^j = \mathbf{g}_i \cdot \mathbf{g}^j &= \frac{\partial x^k}{\partial u^i}\frac{\partial u^j}{\partial x^l}\mathbf{e}_k \cdot \mathbf{e}_l = \frac{\partial x^k}{\partial u^i}\frac{\partial u^j}{\partial x^l}\delta_k^l \\ &= \frac{\partial x^k}{\partial u^i}\frac{\partial u^j}{\partial x^k} = \frac{\partial u^j}{\partial u^i} \end{aligned} \tag{2.55a}$$

The Kronecker delta is defined by

$$\delta_i^j \equiv \begin{cases} 0 \text{ for } i \neq j \\ 1 \text{ for } i = j \end{cases} \tag{2.55b}$$

As an example, the covariant metric tensor **M** in the cylindrical coordinates results from Eq. (2.31).

$$\mathbf{M} = \begin{bmatrix} g_{11} & g_{12} & g_{13} \\ g_{21} & g_{22} & g_{23} \\ g_{31} & g_{32} & g_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.56a}$$

The contravariant metric coefficients in the contravariant metric tensor $\mathbf{M}^{-1}$ are calculated from inverting the covariant metric tensor **M**.

$$\mathbf{M}^{-1} = \begin{bmatrix} g^{11} & g^{12} & g^{13} \\ g^{21} & g^{22} & g^{23} \\ g^{31} & g^{32} & g^{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & r^{-2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.56b}$$

Analogously, the covariant metric tensor **M** in the spherical coordinates results from Eq. (2.40).

$$\mathbf{M} = \begin{bmatrix} g_{11} & g_{12} & g_{13} \\ g_{21} & g_{22} & g_{23} \\ g_{31} & g_{32} & g_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \rho^2 & 0 \\ 0 & 0 & (\rho \sin \varphi)^2 \end{bmatrix} \tag{2.57a}$$

The contravariant metric coefficients in the contravariant metric tensor $\mathbf{M}^{-1}$ are calculated from inverting the covariant metric tensor **M**.

$$\mathbf{M}^{-1} = \begin{bmatrix} g^{11} & g^{12} & g^{13} \\ g^{21} & g^{22} & g^{23} \\ g^{31} & g^{32} & g^{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \rho^{-2} & 0 \\ 0 & 0 & (\rho \sin \varphi)^{-2} \end{bmatrix} \tag{2.57b}$$

2.3.3 Tensors of Second Order and Higher Orders

Mapping an arbitrary vector $\mathbf{x} \in \mathbf{R}^N$ by a linear functional $\mathbf{T}$, one obtains its image vector $\mathbf{y} = \mathbf{T}\mathbf{x}$ [2, 3].

$$\begin{aligned} &\mathbf{T} : \mathbf{R}^N \rightarrow \quad \mathbf{R}^N \\ &\mathbf{T} : \mathbf{x} \rightarrow \mathbf{y} = \mathbf{T}\mathbf{x} \equiv \mathbf{T} \cdot \left(x^j \mathbf{g}_{\mathrm{j}} \right) = x^j \mathbf{T} \cdot \mathbf{g}_j \\ &\qquad\qquad\qquad\quad = x^j T_{ik} \mathbf{g}^i \left(\mathbf{g}^k \cdot \mathbf{g}_j \right) = x^j T_{ik} \delta_j^k \mathbf{g}^i = T_{ij} x^j \mathbf{g}^i \end{aligned} \tag{2.58}$$

where $\mathbf{T}$ is a second-order tensor $\in \mathbf{R}^N \times \mathbf{R}^N$.

It is obvious that the image vector $\mathbf{y} = \mathbf{T}\mathbf{x}$ is a tensor of one lower order compared to the tensor $\mathbf{T}$ that can be considered as a linear operator.

The second-order tensor $\mathbf{T}$ can be generated from the tensor product (dyadic product) of two vectors $\mathbf{u}$ and $\mathbf{v}$, as denoted in Eq. (2.60).

Let $\mathbf{u}$ and $\mathbf{v}$ be two arbitrary vectors. They can be written in the covariant and contravariant bases as

$$\begin{cases} \mathbf{u} = u^i \mathbf{g}_i = u_i \mathbf{g}^i \in \mathbf{R}^N; \\ \mathbf{v} = v^j \mathbf{g}_j = v_j \mathbf{g}^j \in \mathbf{R}^N \end{cases} \tag{2.59}$$

The tensor product of two vectors $\mathbf{u}$ and $\mathbf{v}$ results in the second-order tensor $\mathbf{T}$.

$$\begin{aligned} &\mathbf{T} : \mathbf{u}, \mathbf{v} \in \mathbf{R}^N \rightarrow \mathbf{T} \equiv \mathbf{u} \otimes \mathbf{v} \in \mathbf{R}^N \times \mathbf{R}^N \\ &\mathbf{T} = u^i v^j \mathbf{g}_i \otimes \mathbf{g}_j \equiv u^i v^j \mathbf{g}_i \mathbf{g}_j \equiv T^{ij} \mathbf{g}_i \mathbf{g}_j \\ &\mathbf{T} = u_i v_j \mathbf{g}^i \otimes \mathbf{g}^j \equiv u_i v_j \mathbf{g}^i \mathbf{g}^j \equiv T_{ij} \mathbf{g}^i \mathbf{g}^j \end{aligned} \tag{2.60}$$

Note that the terms $\mathbf{g}_i\mathbf{g}_j$ and $\mathbf{g}^i\mathbf{g}^j$ are called the covariant and contravariant basis tensors. Hence, they are not the same notations as the covariant and contravariant metric coefficients g_{ij} and g^{ij}, respectively.

$$\begin{aligned} &\mathbf{g}_i \mathbf{g}_j \neq \mathbf{g}_i \cdot \mathbf{g}_j \equiv g_{ij} \\ &\mathbf{g}^i \mathbf{g}^j \neq \mathbf{g}^i \cdot \mathbf{g}^j \equiv g^{ij} \end{aligned}$$

Similarly, one obtains the properties of the mixed basis tensors:

6 A ;D /
+,- * *. * $ +A 7 + D B /
// D/ A D " D/ A & /"
(!

1.1 FROM MÖBIUS TO GYROGROUPS

Möbius transformations of the complex open unit disc $\mathbb{D}$,

$$\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\} \tag{1.1}$$

of the complex plane $\mathbb{C}$ are studied in most books on function theory of one complex variable. According to [31, p. 176], these are important for at least two reasons: (i) they play a central role in non-Euclidean geometry [51], and (ii) they are the only automorphisms of the disc.

Ahlfors' book [1], *Conformal Invariants: Topics in Geometric Function Theory*, begins with a presentation of the *polar decomposition* of the most general Möbius self-transformation of the complex open unit disc $\mathbb{D}$,

$$z \mapsto e^{i\theta} \frac{a + z}{1 + \bar{a}z} = e^{i\theta}(a \oplus z) , \tag{1.2}$$

where $a, z \in \mathbb{D}$, $\overline{a}$ being the complex conjugate of a, and where $\theta \in \mathbb{R}$ is a real number [12, p. 211], [17, p. 185], [23, pp. 6 – 7], [31, pp. 177 – 178]. In (1.2) we present the Möbius polar decomposition in a form that suggests the *Möbius addition*, $\oplus$, which we define by the equation

$$a \oplus z = \frac{a+z}{1+\overline{a}z} \, . \tag{1.3}$$

Naturally, Möbius subtraction, $\ominus$, is given by $a \ominus z = a \oplus (-z)$, so that $z \ominus z = 0$ and $\ominus z = 0 \ominus z = 0 \oplus (-z) = -z$.

Interestingly, Möbius addition possesses the *gyroautomorphic inverse property*

$$\ominus(a \oplus b) = \ominus a \ominus b \tag{1.4}$$

and the *left cancellation law*

$$\ominus a \oplus (a \oplus z) = z \tag{1.5}$$

for all $a, b, z \in \mathbb{D}$. Contrasting the left cancellation law (1.5) of Möbius addition, care must be taken with its right counterpart since, in general,

$$(a \oplus z) \ominus z \neq a \, . \tag{1.6}$$

Indeed, in (1.15) below we will introduce a secondary binary operation, $\boxplus$, which is a binary *cooperation* called a Möbius *coaddition*, and which gives rise to the *right cancellation law*, (1.18), of Möbius addition.

Möbius addition is neither commutative nor associative. Fortunately, the failure of associativity is a source of rich theory, the theory of gyrogroups and gyrovector spaces, as we will see in this book.

To measure the extent of which Möbius addition deviates from commutativity we define the *gyrations*

$$\mathrm{gyr}[a, b] = \frac{a \oplus b}{b \oplus a} = \frac{1 + a\overline{b}}{1 + \overline{a}b} \tag{1.7}$$

for $a, b \in \mathbb{D}$. Being unimodular complex numbers, $|\mathrm{gyr}[a, b]| = 1$, gyrations represent rotations of the disc $\mathbb{D}$ about its center. In the special case when $a = 0$ or $b = 0$ or $a = \lambda b$ for some real number λ, the gyration $\mathrm{gyr}[a, b]$ is trivial, that, is, $\mathrm{gyr}[a, b] = 1$. The number -1 cannot be represented by a gyration since the equation $\mathrm{gyr}[a, b] = -1$ has no solution for a and b in the disc $\mathbb{D}$.

Gyrations respect Möbius addition,

$$\mathrm{gyr}[a, b](c \oplus d) = \mathrm{gyr}[a, b]c \oplus \mathrm{gyr}[a, b]d \tag{1.8}$$

for all $a, b, c, d \in \mathbb{D}$, and the inverse $\mathrm{gyr}^{-1}[a, b]$ of a gyration $\mathrm{gyr}[a, b]$ is the gyration $\mathrm{gyr}[b, a]$,

$$\mathrm{gyr}^{-1}[a, b] = \mathrm{gyr}[b, a] \, , \tag{1.9}$$

as we see from (1.7). Hence, gyrations are special automorphisms of the Möbius groupoid $(\mathbb{D}, \oplus)$. We recall that a groupoid $(G, \oplus)$ is a nonempty set, G, with a binary operation, $\oplus$, and that an automorphism of a groupoid $(G, \oplus)$ is a bijective self map f of G that respects its binary operation $\oplus$, that is, $f(a \oplus b) = f(a) \oplus f(b)$. The set of all automorphisms of a groupoid $(G, \oplus)$ forms a group, denoted $Aut(G, \oplus)$. The formal definition of groups will be presented in Def. 1.3, p. 5.

To emphasize that gyrations are automorphisms, they are also called *gyroautomorphisms*. The automorphism $\phi(z) = -z$ of the disc is not a gyroautomorphism of the disc since the equation $(1 + a\bar{b})/(1 + \bar{a}b) = -1$ for $a, b \in \mathbb{D}$ has no solution in the disc. Consequently, the gyroautomorphisms of the disc do not form a group under gyroautomorphism composition.

It follows from the gyration definition in (1.7) that Möbius addition obeys the *gyrocommutative law*:

$$a \oplus b = \text{gyr}[a, b](b \oplus a) \tag{1.10}$$

for all $a, b \in \mathbb{D}$. The Möbius gyrocommutative law (1.10) is not terribly surprising since it is generated by definition. But, we are not finished!

Coincidentally, the gyration gyr$[a, b]$, which "repairs" in (1.10) the breakdown of commutativity in Möbius addition, "repairs" the breakdown of associativity in Möbius addition as well. Indeed, it gives rise to the following gyroassociative laws (left and right),

$$\begin{aligned} a \oplus (b \oplus z) &= (a \oplus b) \oplus \text{gyr}[a, b]z\,, \\ (a \oplus b) \oplus z &= a \oplus (b \oplus \text{gyr}[b, a]z)\,, \end{aligned} \tag{1.11}$$

for all $a, b, z \in \mathbb{D}$. Hence, gyrations measure simultaneously the extent to which Möbius addition deviates from both associativity and commutativity.

Möbius gyrations are expressible in terms of Möbius addition. Indeed, solving the first identity in (1.11) for gyr$[a, b]z$ by means of the Möbius left cancellation law (1.5) we have

$$\text{gyr}[a, b]z = \ominus(a \oplus b) \oplus \{a \oplus (b \oplus z)\}\,. \tag{1.12}$$

Möbius gyrations gyr$[a, b]$, $a, b \in \mathbb{D}$, thus endow the Möbius disc groupoid $(\mathbb{D}, \oplus)$ with a group-like structure, naturally called a *gyrogroup*. In fact, Möbius gyrations do more than that. They contribute their own rich structure to the gyrogroup. Thus, for instance, they possess the *loop properties* (left and right)

$$\begin{aligned} \text{gyr}[a \oplus b, b] &= \text{gyr}[a, b]\,, \\ \text{gyr}[a, b \oplus a] &= \text{gyr}[a, b]\,, \end{aligned} \tag{1.13}$$

and the nested gyration identity

$$\text{gyr}[b, \ominus\text{gyr}[b, a]a] = \text{gyr}[a, b]\,, \tag{1.14}$$

as one can straightforwardly check.

Möbius addition $\oplus$, as a primary addition in the disc, comes with *Möbius coaddition*, $\boxplus$, as a secondary addition in the disc. Möbius coaddition is defined in terms of Möbius addition and Möbius gyrations by the equation

$$a \boxplus b = a \oplus \text{gyr}[a, \ominus b]b \tag{1.15}$$

for all $a, b \in \mathbb{D}$. It turns out that Möbius coaddition is commutative (but not associative), given by the equation

$$a \boxplus b = \frac{(1 - |a|^2)b + (1 - |b|^2)a}{1 - |a|^2|b|^2} \tag{1.16}$$

for all $a, b \in \mathbb{D}$, as we see from (1.15) and (1.7).

Replacing b by $\ominus b$ in (1.15), and noting that $\mathrm{gyr}[a, \ominus b]$ is an automorphism, it follows from (1.15) that

$$a \boxminus b = a \oplus \mathrm{gyr}[a, b](\ominus b) = a \ominus \mathrm{gyr}[a, b]b\,. \tag{1.17}$$

for all $a, b \in \mathbb{D}$, where we use the obvious notation, $a \boxminus b = a \boxplus (\ominus b)$.

In terms of Möbius coaddition, the *right cancellation law* of Möbius addition takes the form

$$(a \oplus b) \boxminus b = a\,. \tag{1.18}$$

The right cancellation law (1.18) is associated with the second right cancellation law

$$(a \boxplus b) \ominus b = a\,, \tag{1.19}$$

as one can straightforwardly check. Identities (1.18) and (1.19) present one of the duality symmetries that Möbius addition and coaddition share.

Guided by the group-like structure of Möbius addition, we define a *gyrogroup* to be a group in which the associative law is replaced by the gyroassociative laws (left and right), and a *gyrocommutative gyrogroup* to be a commutative group in which the commutative and associative laws are replaced by the gyrocommutative and gyroassociative laws. Like Möbius gyrations, the gyrations in the gyrocommutative and gyroassociative laws of an abstract gyrogroup must be automorphisms of the gyrogroup.

Thus, the disc $\mathbb{D}$ with its Möbius addition forms a *gyrocommutative gyrogroup* $(\mathbb{D}, \oplus)$ where, strikingly, the concepts of gyrocommutativity and gyroassociativity of the gyrogroup preserve the flavor of their classical counterparts [50, 51]. Mathematical coincidences are not accidental! Hence, in particular, the gyrations that "coincidentally" repair in (1.10) – (1.11) the breakdown of commutativity and associativity need not be peculiar to Möbius addition, as shown in [13, 14], and as we will see in this book.

Indeed, the generalization of groups into gyrogroups that Möbius addition suggests bears an intriguing resemblance to the generalization of the rational numbers to the real ones. The beginner is initially surprised to discover an irrational number, like $\sqrt{2}$, but soon later he or she is likely to realize that there are more irrational numbers than rational ones. Similarly, the gyrogroup structure of Möbius addition initially comes as a surprise. But, interested explorers may soon realize that there are indications that, in some sense, there are more non-group gyrogroups than groups [13].

+,- * *. * $ 4GAH.4 +"
9 ; D E 1 /"
3 / $! ($! $

or

$$\begin{Bmatrix}\alpha\\ \mu\ \ \nu\end{Bmatrix} A_c \begin{Bmatrix}c\\ \alpha\ \ \sigma\end{Bmatrix} + A_\alpha \frac{\partial}{\partial x^\sigma}\begin{Bmatrix}\alpha\\ \mu\ \ \nu\end{Bmatrix} - \begin{Bmatrix}\alpha\\ \mu\ \ \sigma\end{Bmatrix} A_c \begin{Bmatrix}c\\ \alpha\ \ \nu\end{Bmatrix} - A_c \frac{\partial}{\partial x^\nu}\begin{Bmatrix}c\\ \mu\ \ \sigma\end{Bmatrix} = 0 \text{ by Eq. (4.49)}$$

or

$$\begin{Bmatrix}\alpha\\ \mu\ \ \nu\end{Bmatrix} A_c \begin{Bmatrix}c\\ \alpha\ \ \sigma\end{Bmatrix} + A_c \frac{\partial}{\partial x^\sigma}\begin{Bmatrix}c\\ \mu\ \ \nu\end{Bmatrix} - \begin{Bmatrix}\alpha\\ \mu\ \ \sigma\end{Bmatrix}\begin{Bmatrix}c\\ \alpha\ \ \nu\end{Bmatrix} A_c - A_c \frac{\partial}{\partial x^\nu}\begin{Bmatrix}c\\ \mu\ \ \sigma\end{Bmatrix} = 0$$

or

$$\left[\begin{Bmatrix}\alpha\\ \mu\ \ \nu\end{Bmatrix}\begin{Bmatrix}c\\ \alpha\ \ \sigma\end{Bmatrix} + \frac{\partial}{\partial x^\sigma}\begin{Bmatrix}c\\ \mu\ \ \nu\end{Bmatrix} - \begin{Bmatrix}\alpha\\ \mu\ \ \sigma\end{Bmatrix}\begin{Bmatrix}c\\ \alpha\ \ \nu\end{Bmatrix} - \frac{\partial}{\partial x^\nu}\begin{Bmatrix}c\\ \mu\ \ \sigma\end{Bmatrix}\right] A_c = 0$$

or

$$\left[-\frac{\partial}{\partial x^\nu}\begin{Bmatrix}c\\ \mu\ \ \sigma\end{Bmatrix} + \frac{\partial}{\partial x^\sigma}\begin{Bmatrix}c\\ \mu\ \ \nu\end{Bmatrix} - \begin{Bmatrix}\alpha\\ \mu\ \ \sigma\end{Bmatrix}\begin{Bmatrix}c\\ \alpha\ \ \nu\end{Bmatrix} + \begin{Bmatrix}\alpha\\ \mu\ \ \nu\end{Bmatrix}\begin{Bmatrix}c\\ \alpha\ \ \sigma\end{Bmatrix}\right] A_c = 0$$

or

$$R^c_{\mu\sigma\nu} A_c = 0, \quad \text{i.e. } R^c_{\mu\sigma\nu} = 0;\ \text{as } A_c \text{ is arbitrary.}$$

Thus, we see that the RHS of Eq. (4.50) is integrable only when the curvature tensor vanishes. Consequently, Eq. (4.49) is integrable only when $R^c_{\mu\sigma\nu} = 0$.

Now, we can carry the vector A_μ to any point by parallel displacement. This gives a unique result independent of the path of transfer. If a vector is displaced in this way we obtain a uniform vector field. Thus, the construction of uniform vector field is only possible when the curvature tensor vanishes.

4.3 Einstein Space

Einstein space is defined as a space which is homogeneous with regard to the Ricci tensor R_{ij}. Thus, a space for which

$$R_{ij} = \lambda g_{ij};\quad \lambda = \text{invariant at all points} \tag{4.54}$$

at every point of the space, then that space is called *Einstein space.* Inner multiplication of g^{ij} shows that

$$g^{ij} R_{ij} = \lambda g^{ij} g_{ij} \Rightarrow R = \lambda N. \tag{4.55}$$

Thus for an Einstein space,

$$R_{ij} = \lambda g_{ij} = \frac{1}{N} R g_{ij}. \tag{4.56}$$

Hence, for an Einstein space Eq. (4.56) holds at every point of the space.

Theorem 4.3.1 *A space of constant curvature is an Einstein space.*

Proof: Let the Riemannian curvature κ at P of V_N for the orientation determined by p^i and q^i, is given in Eq. (4.34) as

$$\kappa = \frac{p^h q^i p^j q^k R_{hijk}}{p^h q^i p^j q^k \left[g_{hj} g_{ik} - g_{ij} g_{kh}\right]}.$$

Since κ is constant and independent of the orientation so,

$$\kappa = \frac{R_{hijk}}{g_{hj} g_{ik} - g_{ij} g_{kh}}$$

or

$$R_{hijk} = \kappa \left[g_{hj} g_{ik} - g_{ij} g_{kh}\right].$$

Multiplying by g^{hk}, we get

$$\kappa g^{hk} \left[g_{hj} g_{ik} - g_{ij} g_{kh}\right] = g^{hk} R_{hijk}$$

or

$$\kappa \left(\delta_i^h g_{hj} - N g_{ij}\right) = R_{ij}; \quad \text{since, } g^{hk} R_{hijk} = R_{ij}$$

or

$$\kappa \left(g_{ij} - N g_{ij}\right) = R_{ij}$$

or

$$\kappa \left(1 - N\right) g_{ij} g^{ij} = R_{ij} g^{ij}; \text{ multiplying by } g^{ij}$$

or

$$\kappa N (1 - N) = R; \quad \text{as } R_{ij} g^{ij} = R.$$

Therefore,

$$R_{ij} = (1 - N) g_{ij} \frac{R}{N(1 - N)} = \frac{R}{N} g_{ij}.$$

This is the necessary and sufficient condition for the space V_N to be Einstein space.

EXAMPLE 4.3.1 *If $g^{ik} R_{kj} = R^i_j$ and $g^{ij} R_{ij} = R$, show that $R^i_{j,i} = \frac{1}{2} \frac{\partial R}{\partial x^j}$. Hence, show that for an Einstein space, the scalar curvature R is constant.*

Solution: Here, we use the Bianchi identity, Eq. (4.20), $R_{pkjt,i} + R_{pkti,j} + R_{pkij,t} = 0$. The covariant differentiation of the relation $g^{ik} R_{kj} = R^i_j$ with respect to x^l is given by

$$R^i_{j,l} = g^{ik}_{,l} R_{kj} + g^{ik} R_{kj,l} = g^{ik} R_{kj,l}; \quad \text{as } g^{ik}_{,l} = 0.$$

Thus, the divergence of R^i_j is given by

$$\begin{aligned}
R^i_{j,i} &= g^{ik}R_{kj,i} = g^{ik}R^t_{kjt,i};\ \text{as } R^t_{kit} = R_{kj} \\
&= g^{ik}\left(g^{pt}R_{pkjt}\right)_{,i};\quad \text{as } g^{pt}R_{pkjt} = R^t_{kjt} \\
&= g^{ik}g^{pt}R_{pkjt,i};\quad \text{as } g^{pt}_{,i} = 0 \\
&= -g^{ik}g^{pt}\left[R_{pkti,j} + R_{pkij,t}\right] = -g^{pt}\left[g^{ik}R_{pkti,j} + g^{ik}R_{pkij,t}\right] \\
&= -g^{pt}\left[g^{ki}R_{kpti,j} + g^{ik}R_{kpji,t}\right] \\
&\quad\ \text{as } R_{kpti} = -R_{pkti} \text{ and } R_{kpji} = R_{pkij} \\
&= -g^{pt}\left[-R^i_{pti,j} + R^i_{pji,t}\right] = -g^{pt}\left[-R_{pt,j} + R_{pj,t}\right] \\
&= \left(g^{pt}R_{pt}\right)_{,j} - g^{pt}R_{pj,t} = R_{,j} - R^t_{j,t} = R_{,j} - R^i_{j,i} \\
&\quad\ \text{replacing the dummy index } t \text{ by } i
\end{aligned}$$

or

$$2R^i_{j,i} = R_{,j} = \frac{\partial R}{\partial x^j} \Rightarrow R^i_{j,i} = \frac{1}{2}\frac{\partial R}{\partial x^j}.$$

For an Einstein space, $R_{ij} = \frac{R}{N}g_{ij}$, therefore,

$$g^{ki}R_{ij} = \frac{R}{N}g_{ij}g^{ki} \Rightarrow R^k_j = \frac{R}{N}\delta^k_j$$

or

$$R^k_{j,k} = \frac{1}{N}R_{,k}\delta^k_j = \frac{1}{N}R_{,j}$$

or

$$\frac{1}{2}R_{,j} = \frac{1}{N}R_{,j} \Rightarrow R_{,j} = 0;\quad \text{for } N > 2.$$

Therefore, for an Einstein space, the scalar curvature R is constant.

EXAMPLE 4.3.2 *For a V_2 referred to an orthogonal system of parametric curves $(g_{12} = 0)$, show that*

$$\begin{aligned}
R_{12} &= 0,\quad R_{11}g_{22} = R_{22}g_{11} = R_{1221} \\
R &= g^{ij}R_{ij} = \frac{2R_{1221}}{g_{11}g_{22}},\quad \textit{consequently, } R_{ij} = \frac{R}{2}g_{ij}.
\end{aligned}$$

Hence, show that V_2 is an Einstein space.

Solution: Given that, in a V_2, $g_{12} = 0$ so that $g^{12} = 0$. Also,

$$g^{ij} = \frac{1}{g_{ij}} \Rightarrow g^{11} = \frac{1}{g_{11}}, \quad g^{22} = \frac{1}{g_{22}}.$$

The metric of V_2 is given by

$$\begin{aligned} ds^2 &= g_{ij}dx^i dx^j; \quad i, j = 1, 2 \\ &= g_{11}(dx^1)^2 + g_{22}(dx^2)^2; \quad \text{as } g_{12} = 0. \end{aligned}$$

Also, we know that $g^{hk}R_{hijk} = R_{ij}$ and

$$g = \begin{vmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{vmatrix} = \begin{vmatrix} g_{11} & 0 \\ 0 & g_{22} \end{vmatrix} = g_{11}g_{22}.$$

(a) Here, we have to show that $R_{12} = 0$. For this,

$$\begin{aligned} R_{12} &= g^{hk}R_{h12k} = g^{h1}R_{h121} + g^{h2}R_{h122} = g^{h1}R_{h121}; \quad \text{as } R_{h122} = 0 \\ &= g^{11}R_{1121} + g^{21}R_{2121} = 0; \quad \text{as } g^{12} = 0 \text{ and } R_{1121} = 0. \end{aligned}$$

(b) Using the relation $g^{hk}R_{hijk} = R_{ij}$ we get

$$\begin{aligned} R_{11} &= g^{hk}R_{h11k} = g^{2k}R_{211k} = g^{22}R_{2112} = \frac{R_{2112}}{g_{22}}. \\ R_{22} &= g^{hk}R_{h22k} = g^{1k}R_{122k} = g^{11}R_{1221} = \frac{R_{1221}}{g_{11}}. \end{aligned}$$

Since $R_{2112} = R_{1221}$, it follows that, $R_{11}g_{22} = R_{1221} = R_{22}g_{11}$. Therefore, in a two-dimensional Riemannian space the components of a Ricci tensor are proportional to the components of a metric tensor.

(c) From the definition of R, we get

$$\begin{aligned} R &= g^{ij}R_{ij} = g^{i1}R_{i1} + g^{i2}R_{i2} \\ &= g^{11}R_{11} + g^{22}R_{22}; \quad \text{as } g^{12} = 0 \\ &= \frac{R_{11}}{g_{11}} + \frac{R_{22}}{g_{22}} = \frac{R_{1221}}{g_{22}g_{11}} + \frac{R_{1221}}{g_{11}g_{22}} \\ &= \frac{2R_{1221}}{g_{11}g_{22}}; \quad \text{as } R_{2112} = R_{1221}. \end{aligned}$$

(d) From (c), we see that

$$R = \frac{2R_{1221}}{g_{11}g_{22}} = \frac{2R_{1221}}{g}; \quad \text{as } g = g_{11}g_{22}$$
$$\Rightarrow R_{1221} = \frac{1}{2}Rg.$$

Therefore, the relation in (b) can be written in the form

$$R_{11}g_{22} = \frac{1}{2}Rg = R_{22}g_{11}$$

from which it follows that

$$\begin{aligned} R_{11} &= \frac{Rg}{2g_{22}} = \frac{Rg_{11}g_{22}}{2g_{22}} = \frac{1}{2}Rg_{11} \\ R_{22} &= \frac{Rg}{2g_{11}} = \frac{Rg_{11}g_{22}}{2g_{11}} = \frac{1}{2}Rg_{22} \\ R_{12} &= \frac{1}{2}Rg_{12} = 0; \quad \text{as } g_{12} = 0. \end{aligned}$$

Consequently, we get $R_{ij} = \frac{R}{2}g_{ij}$. This shows that V_2 is an Einstein space.

4.4 Mean Curvature

Let $\lambda^i_{h|}$ be the components of unit vector in a direction at a point P of a V_N. Let $\lambda^i_{k|}$ be the components of unit vector forming an orthogonal ennuple. One vector out of N vectors $\lambda^i_{k|}$ is $\lambda^i_{h|}$ and hence, we can say that $\lambda^i_{k|}$ are components of $N-1$ vectors where $h \neq k$ and $k = 1, 2, \ldots, N$.

Let the Riemannian curvature at P of V_N for the orientation determined by $\lambda^i_{h|}$ and $\lambda^i_{k|}$ $(k \neq h)$ be denoted by K_{hk} and is given by

$$\begin{aligned} K_{hk} &= \frac{\lambda^p_{h|}\lambda^q_{k|}\lambda^r_{h|}\lambda^s_{k|}R_{pqrs}}{\lambda^p_{h|}\lambda^q_{k|}\lambda^r_{h|}\lambda^s_{k|}\left(g_{pr}g_{qs} - g_{ps}g_{qr}\right)} \\ &= \frac{\lambda^p_{h|}\lambda^q_{k|}\lambda^r_{h|}\lambda^s_{k|}R_{pqrs}}{\left(\lambda^p_{h|}\lambda^r_{h|}g_{pr}\right)\left(\lambda^q_{k|}\lambda^s_{k|}g_{qs}\right) - \left(\lambda^p_{h|}\lambda^s_{k|}g_{ps}\right)\left(\lambda^q_{k|}\lambda^r_{h|}g_{qr}\right)}. \end{aligned} \tag{4.57}$$

Since unit vectors $\lambda^i_{h|}$ and $\lambda^i_{k|}$ are orthogonal, so,

$$\lambda^p_{h|}\lambda^r_{h|}g_{pr} = 1 \text{ and } \lambda^p_{h|}\lambda^s_{k|}g_{ps} = 0.$$

Therefore, from Eq. (4.57), we get

$$K_{hk} = \frac{\lambda^p_{h|}\lambda^q_{k|}\lambda^r_{h|}\lambda^s_{k|}R_{pqrs}}{1.1 - 0.0} = \lambda^p_{h|}\lambda^q_{k|}\lambda^r_{k|}\lambda^s_{k|}R_{pqrs}. \tag{4.58}$$

Since the right-hand side of Eq. (4.58) vanishes for $h = k$, and so we say K_{hh} is equal to zero. Let $M_h = \sum_{k=1}^{N} K_{hk}$. Hence from Eq. (4.58) we have,

$$\begin{aligned}
\sum_{k=1}^{N} K_{hk} = M_h &= \sum_{k=1}^{N} \lambda^p_{h|}\lambda^q_{k|}\lambda^r_{h|}\lambda^s_{k|}R_{pqrs} \\
&= \lambda^p_{h|}\lambda^q_{h|} \sum_{k=1}^{N} \lambda^q_{k|}\lambda^s_{k|}R_{pqrs} \\
&= \lambda^p_{h|}\lambda^r_{h|}g^{qs}R_{pqrs}; \text{ as } \sum_{k=1}^{N} \lambda^q_{k|}\lambda^s_{k|} = g^{qs} \\
&= -\lambda^p_{h|}\lambda^r_{h|}g^{qs}R_{qprs} = -\lambda^p_{h|}\lambda^r_{h|}R^s_{pqs} = -\lambda^p_{h|}\lambda^r_{h|}R_{pr}.
\end{aligned} \tag{4.59}$$

This shows that M_h is independent of $(N-1)$ orthogonal direction choosen to complete an orthogonal ennuple. Here, M_h is defined as mean curvature or Riccian curvature of V_N for the direction $\lambda^p_{h|}$. Summing Eq. (4.58) over the N mutually orthogonal directions, we get

$$\sum_{h=1}^{N} M_h = -\sum_{h=1}^{N} \lambda^p_{h|}\lambda^r_{h|}R_{pr} = -g^{pr}R_{pr} = -R,$$

where R is the scalar curvature. This proves that the sum of mean curvatures for N mutual orthogonal directions is independent of the directions chosen to complete an orthogonal ennuple and has the value $-R$.

4.4.1 Ricci's Principal Directions

Let us suppose that $\lambda^i_{h|}$ is not a unit vector and therefore, the mean curvature M_h in this case is given by

$$M_h = -\frac{R_{ij}\lambda^i_{h|}\lambda^j_{h|}}{g_{ij}\lambda^i_{h|}\lambda^j_{h|}} \Rightarrow \left(R_{ij} + M_h g_{ij}\right)\lambda^i_{h|}\lambda^j_{h|} = 0.$$

Differentiating this equation with respect to $\lambda^i_{h|}$, we get

$$\frac{\partial M_h}{\partial \lambda^i_{h|}} g_{jk}\lambda^j_{h|}\lambda^k_{h|} + 2\left(R_{ij} + M_h g_{ij}\right)\lambda^j_{h|} = 0.$$

For extreme value of M_h, we must have $\frac{\partial M_h}{\partial \lambda^i_{h|}} = 0$, and so

$$(R_{ij} + M_h g_{ij})\,\lambda^j_{h|} = 0,$$

which are called *Ricci's principal direction* of the space and they are principal directions of Ricci tensor R_{ij}.

EXAMPLE 4.4.1 *For a two-dimensional manifold, prove that* $K_{pq} = -\frac{R}{2}$.

Solution: In V_2, we know R_{1212} is the only non-vanishing Riemann–Christoffel tensor given by

$$R_{1212} = \frac{1}{2}Rg; R = \text{ total curvature.}$$

Let the Riemannian curvature at P of V_N for the orientation determined by $\lambda^i_{p|}$ and $\lambda^i_{q|}$ $(q \neq p)$ be denoted by K_{pq} and from Eq. (4.57) we get,

$$\begin{aligned} K_{pq} &= \frac{\lambda^h_{p|}\lambda^i_{q|}\lambda^j_{p|}\lambda^k_{q|}R_{hijk}}{\lambda^h_{p|}\lambda^i_{q|}\lambda^j_{p|}\lambda^k_{q|}\left(g_{hj}g_{ik} - g_{hk}g_{ij}\right)} \\ &= \frac{\lambda^1_{p|}\lambda^2_{q|}\lambda^1_{p|}\lambda^2_{q|}R_{1212}}{\lambda^1_{p|}\lambda^2_{q|}\lambda^1_{p|}\lambda^2_{q|}\left(g_{11}g_{22} - g_{12}g_{21}\right)} = \frac{R_{1212}}{g_{12}^2 - g_{11}g_{22}} = \frac{R_{1212}}{-g} = -\frac{1}{2}R. \end{aligned}$$

4.5 Geodesics in a V_N

Consider a curve $\Gamma : x^i = x^i(t)$ in a V_N, an N-dimensional Riemannian space, on it and let $x^i(t)$ be the co-ordinates of a general point P on it. Let the points A and B of the path correspond, respectively, to the value t_1 and t_2 of the parameter.

Consider two points on the path close to each other with co-ordinates x^i and $x^i + dx^i$. Let the parameter take values t and $t + dt$ corresponding to these two points, respectively. The interval ds between the two points is given by

$$ds^2 = g_{ij}dx^i dx^j = g_{ij}\frac{dx^i}{dt}\frac{dx^j}{dt}dt^2. \tag{4.60}$$

The total interval between A and B along the path therefore,

$$\begin{aligned} S &= \int_{t_1}^{t_2} \sqrt{g_{ij}\frac{dx^i}{dt}\frac{dx^j}{dt}}\,dt \\ &= \int_A^B F(x^i, \dot{x}^i)dt, \text{ say; where } F(x^i, \dot{x}^i) = \sqrt{g_{ij}\dot{x}^i\dot{x}^j}. \end{aligned} \tag{4.61}$$

In order Γ is a geodesic, integral Eq. (4.61) should be stationary. In order that integral Eq. (4.61) is stationary, we have the Euler-Lagrange condition,

$$\frac{\partial F}{\partial \dot{x}^l} - \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{x}^l}\right) = 0$$

or

$$\frac{1}{2\dot{s}}\frac{\partial g_{ij}}{\partial x^l}\dot{x}^i\dot{x}^j - \left[-\frac{\ddot{s}}{\dot{s}}g_{li}\dot{x}^i + \frac{1}{\dot{s}}\frac{\partial g_{li}}{\partial \dot{x}^j}\dot{x}^j\dot{x}^i + \frac{1}{\dot{s}}g_{li}\ddot{x}^i\right] = 0$$

or

$$g_{li}\ddot{x}^i - \frac{\ddot{s}}{\dot{s}}g_{li}\dot{x}^i + [li, j]\dot{x}^i\dot{x}^j = 0. \tag{4.62}$$

Multiplying this relation by g^{lm} and perform summation over l, we get

$$\ddot{x}^m - \frac{\ddot{s}}{\dot{s}}\dot{x}^m + \left\{\begin{matrix} m \\ i \quad j \end{matrix}\right\}\dot{x}^i\dot{x}^j = 0. \tag{4.63}$$

Equation (4.63) of geodesic can be further simplified by taking $s = t$, so that $\dot{s} = 1$, $\ddot{s} = 0$, and therefore, Eq. (4.63) becomes

$$\frac{d^2x^i}{ds^2} + \left\{\begin{matrix} i \\ j \quad k \end{matrix}\right\}\frac{dx^j}{ds}\frac{dx^k}{ds} = 0, \tag{4.64}$$

where s itself is chosen as the parameter. This, of course, represents a set of N coupled differential equations. Equation (4.64) is known as geodesic equation in V_N.

EXAMPLE 4.5.1 *Show that on the surface of a sphere all great circles are geodesics while no other circle is a geodesic.*

Solution: The components of the metric tensor on the surface of a sphere of radius a are $g_{11} = a^2, g_{22} = a^2\sin^2\theta, g_{12} = g_{21} = 0$ and $g = a^4\sin^2\theta$ so that the metric is $ds^2 = a^2(d\theta)^2 + a^2\sin^2\theta(d\phi)^2$. The non-vanishing Christoffel symbols are

$$\left\{\begin{matrix} 2 \\ 1 \quad 2 \end{matrix}\right\} = \cot\theta = \left\{\begin{matrix} 2 \\ 2 \quad 1 \end{matrix}\right\};\ \left\{\begin{matrix} 1 \\ 2 \quad 2 \end{matrix}\right\} = -\sin\theta\cos\theta.$$

Thus, the geodesic equation reduces to (with $\theta = x^1, \phi = x^2$)

$$\frac{d^2\theta}{ds^2} - \sin\theta\cos\theta\left(\frac{d\phi}{ds}\right)^2 = 0 \tag{i}$$

$$\frac{d^2\phi}{ds^2} + 2\cot\theta\frac{d\theta}{ds}\frac{d\phi}{ds} = 0. \tag{ii}$$

(a) Consider a great circle on the surface of the sphere and choose the z axis $(\theta = 0)$ to be normal to the plane of the circle, so that the great circle is the equator. The parametric equation of the great circle is

$$\theta = \frac{\pi}{2}, \quad \phi = cs + d; \quad c \neq 0, \tag{iii}$$

where c and d are independent of s, θ and ϕ. It is evident that Eq. (iii) satisfy Eqs. (i) and (ii), so that the great circle is a geodesic. Since the choice of the polar axis $\theta = 0$ is arbitrary, it follows that any great circle is a geodesic.

(b) Consider a circle on the sphere such that the plane of the circle does not pass through the centre of the sphere. Again, choose the direction $\theta = 0$ to be normal to the plane of the circle so that the parametric equation of the circle is

$$\theta = \theta_0, \ \phi = ps + q; \ \theta_0 \in (0, \pi) - \left\{\frac{\pi}{2}\right\}; \ p \neq 0, \tag{iv}$$

where p and q are independent of s, θ and ϕ. On substitution, it is seen that Eq. (ii) is satisfied but Eq. (i) is reduces to

$$p^2 \sin\theta_0 \cos\theta_0 = 0.$$

The conditions on θ_0 and p of Eq. (iv) show that this equation cannot be satisfied, proving that the circle is not a geodesic.

Now, we find the lines of shortest distance of the sphere. From Eq. (4.61), we have,

$$S = a \int_A^B F(x^i, \dot{x}^i) dt,$$

where

$$F(x^i, \dot{x}^i) = \sqrt{g_{ij}\dot{x}^i\dot{x}^j} = \sqrt{1 + \dot{\phi}^2 \sin^2\theta}.$$

In order Γ is a geodesic, then Euler–Lagrange equation reduces to,

$$F\dot{\phi} = \text{constant} \ \Rightarrow \frac{\dot{\phi}\sin^2\theta}{\sqrt{1 + \dot{\phi}^2 \sin^2\theta}} = c$$

$$\Rightarrow \dot{\phi}^2 = \frac{c^2}{\sin^4\theta - c^2\sin^2\theta} = \frac{c^2 \operatorname{cosec}^4\theta}{(1 - c^2) - c^2\cot^2\theta}$$

$$\Rightarrow \dot{\phi} = \frac{k \ \operatorname{cosec}^2\theta}{\sqrt{1 - k^2\cot^2\theta}}; k^2 = \frac{c^2}{1 - c^2}$$

$$\Rightarrow \phi = \alpha - \sin^{-1}(k\cot\theta); \ \text{ i.e. } \sin(\alpha - \phi) = k\cot\theta.$$

Let us verify these curves are great circles. Multiplied last equation by $\alpha \sin\theta$, we get

$$(\alpha \sin\theta \cos\phi)\sin\alpha - (\alpha \sin\theta \sin\phi)\cos\alpha = k\alpha\cos\theta$$

or

$$x \sin\alpha - y\cos\alpha = kz.$$

This represents a plane through the centre of the sphere, and the section of the sphere by this plane is a great circle.

4.5.1 Parallel Vector Fields

In an Euclidean space, a vector field $A_i(x^j)$ is said to be a field of parallel vectors if the components A_i are constants, i.e. $A_{i,j} = 0$, i.e. or, suppose that a vector A_i defined at a point x^j undergoes a displacement dx^j to the point $x^j + dx^j$, where, the vector becomes $A_i + \partial A_i$. We shall say that A_i undergoes a parallel displacement if $\frac{\partial A_i}{\partial x^j} = 0$. For example, a uniform electric field is a field of parallel vectors; the gravitational field of the earth in a sufficiently small region of space is a field of parallel vectors.

Here, our object is to generalise the concept of parallel displacement of surface vectors and tensors in Riemannian space, due to Levi Civita. Let us consider a space curve Γ, total of points, whose co-ordinates satisfy equation of the form

$$\Gamma: \; x^i = x^i(t), \tag{4.65}$$

where x^i (belongs to the class C^1) are functions of a single parameter t. Let a vector A be localised at some point P of Γ whose components A^i are functions of t. If we construct at every point of Γ, a vector equal to A in magnitude and parallel to it in direction, we obtain what is known as a parallel field of vectors along the curve Γ. We have shown in Chapter 3 that if A_i is a vector, its co-ordinate derivative $\frac{\partial A_i}{\partial x^j}$ is not a tensor, but the intrinsic derivative $\frac{\delta A_i}{\delta t}$ along the curve Eq. (4.65) is a tensor. To retain the tensor character, therefore, we must define the parallel displacement of a tensor in a Riemannian space differently. If A^i is a solution of the system of differential equation

$$\frac{\delta A^i}{\delta t} = \frac{dA^i}{dt} + \begin{Bmatrix} i \\ j \;\; k \end{Bmatrix} A^j \frac{dx^k}{ds} \tag{4.66}$$

then the vectors A^i are said to form a parallel vector field along the curve Γ given by Eq. (4.65). We say that if the vector A is a parallel field along Γ, then the vectors A do not change along the curve. Thus, the vector A suffers a parallel displacement along Γ if the intrinsic derivative of the tensor along the curve vanishes, (Figure 4.1), i.e.

$$A^i_{,j}\frac{dx^i}{dt} = \frac{\delta A^i}{\delta t} = 0. \tag{4.67}$$

% / 3 " CCC

Conversely, we can show that every solution of system Eq. (4.67) yields a parallel vector field along Γ. Indeed, from the theory of differential equations it is known that this system of first order differential equations has a unique solution when the values of the components A^i are specified at a point of Γ. But the vector field formed by constructing a family of vectors of fixed lengths, parallel to a given vector, satisfies the system. Hence, every solution of Eq. (4.67) satisfies the initial conditions must form a parallel field along Γ.

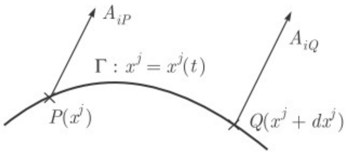

Figure 4.1: Parallel vectors A_{iP} and A_{iQ}.

In other words, we say that A^i is parallely displaced with respect to Riemannian space along the curve Γ, if Eq. (4.67) holds at all points of Γ. Thus, we have a field of parallel vectors, if,

$$A^i_{,j}\frac{dx^i}{dt} = 0 \Rightarrow g_{ik}\left(A^i_{,j}\frac{dx^j}{dt}\right) = 0 \Rightarrow \left(g_{ik}A^i\right)_{,j}\frac{dx^j}{dt} = 0$$

$$\Rightarrow A_{k,j}\frac{dx^j}{dt} = 0; \text{i.e.}\quad A_{i,j}\frac{dx^j}{dt} = 0$$

or

$$\left[\frac{dA^i}{dx^j} + \left\{\begin{matrix} i \\ \alpha \ \ j \end{matrix}\right\} A^\alpha\right]\frac{dx^j}{dt} = 0$$

or

$$\frac{dA^i}{dt} + \left\{\begin{matrix} i \\ \alpha \ \ j \end{matrix}\right\} A^\alpha \frac{dx^j}{dt} = 0$$

or

$$dA^i = -\left\{\begin{matrix} i \\ j \ \ k \end{matrix}\right\} A^j dx^k \tag{4.68}$$

along the curve. The converse follows, as previously, for the existence and uniqueness of solution of such systems of differential equations. Similarly, the condition for a parallel displacement of a covariant vector A_i is

$$dA_i = \left\{\begin{matrix} j \\ i \ \ k \end{matrix}\right\} A_j dx^k. \tag{4.69}$$

Also, it follows from the observation that $A_{i,k} = g_{i\alpha} A^{\alpha}_{,k}$ whenever $A_i = g_{ij} A^j$. Thus, the increments in A^i and A_i due to a displacement dx^j along the curve are given by Eqs. (4.68) and (4.69), respectively.

Result 4.5.1 Now, consider a point $P(x^i)$ and two displacements dx^i and δx^i from P giving rise to the points $Q(x^i + dx^i)$ and $R(x^i + \delta x^i)$, respectively, as seen in Figure 4.2.

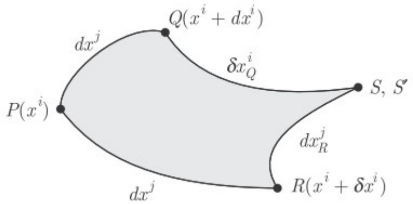

Figure 4.2: Generalised infinitesimal parallelogram.

Imagine the segment $PR = \delta x^i$ being parallely displaced such that as the end point P goes to Q, the other end R goes to S. The displacement QS will be

$$QS \equiv \delta x^i_Q = \delta x^i + d(\delta x^i), \tag{4.70}$$

which makes the co-ordinates of the points S,

$$S \equiv \left[x^i + dx^i + \delta x^i_Q\right] = \left[x^i + dx^i + \delta x^i + d(\delta x^i)\right], \tag{4.71}$$

where by Eq. (4.68), we have

$$d(\delta x^i) = -\left\{ \begin{matrix} i \\ j \quad k \end{matrix} \right\}_P \delta x_j dx^k. \tag{4.72}$$

The subscript P on the Christoffel symbol means is to be evaluated at P. Similarly, imagine the segment $PQ = dx^i$ being parallely displaced such that as the point P goes to R, the other end Q goes to S'. The displacement RS' will be given by

$$RS = dx^i_R = \delta x^i + d(\delta x^i), \tag{4.73}$$

which makes the co-ordinates of the point S'

$$S' \equiv \left[x^i + dx^i + \delta x^i_R\right] = \left[x^i + \delta x^i + dx^i + \delta(dx^i)\right], \tag{4.74}$$

$$\text{where,}\quad \delta(dx^i) = -\left\{ \begin{matrix} i \\ j \quad k \end{matrix} \right\}_P dx_j \delta x^k = -\left\{ \begin{matrix} i \\ j \quad k \end{matrix} \right\}_P dx_k \delta x^j, \tag{4.75}$$

which follows by interchanging j and k and using the symmetric property of the Christoffel symbol. Equations (4.72) and (4.75) show that $d(\delta x^i) = \delta(dx^i)$ indicating that the point S' coincides with S. The closed figure $PQRS$ can be regarded as the generalised infinitesimal parallelogram with opposite sides parallel to each other in the Riemannian space.

Result 4.5.2 (**Curvature of Riemannian space**). Here we have introduced the curvature tensor by considering the parallel displacement of a vector along two different paths. Consider an infinitesimal parallelogram PQRS (Figure 4.3) as described with adjacent sides $PQ = dx^i$ and $PR = \delta x^i$. Imagine a constravariant vector A^i_p defined at the point P being parally displaced is the following two ways:

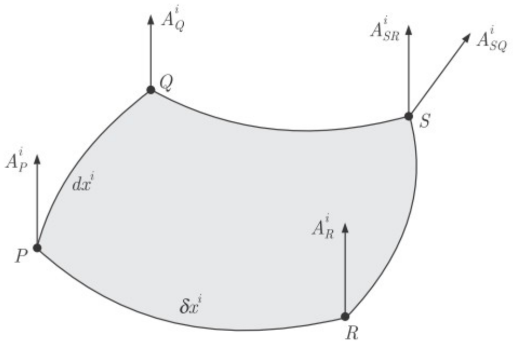

Figure 4.3: Curvature of Riemannian space.

(i) Displace A^i_p parallely from P to Q resulting A^i_Q and then displace A^i_Q parallely from Q to S giving the vector A^i_{SQ}.

(ii) Displace A^i_p parallely from P to R giving A^i_R parallely from R to S resulting the vector A^i_{SR}.

The vector A^i_Q obtained by a parallel displacement of A^i_P is given by

$$A^i_Q = A^i_P + dA^i_P; \quad \text{where } dA^i_P = -\left\{ \begin{matrix} i \\ j \ \ k \end{matrix} \right\}_P A^i_P dx^k. \tag{i}$$

This vector is parallely displaced from Q to S resulting the vector A^i_{SQ} which will be given by

$$A^i_{SQ} = A^i_Q + \delta A^i_Q = A^i_Q - \left\{ \begin{matrix} i \\ l \ \ h \end{matrix} \right\}_Q A^l_Q \, \delta x^h. \tag{ii}$$

The Christoffel symbol depends on the metric tensor which in turn is a function of co-ordinates. For a small displacement, we can therefore write

$$\left\{ \begin{matrix} i \\ l \ \ h \end{matrix} \right\}_Q = \left\{ \begin{matrix} i \\ l \ \ h \end{matrix} \right\}_P + \left\{ \begin{matrix} i \\ l \ \ h \end{matrix} \right\}_{P,m} dx^m \tag{iii}$$

where $\begin{Bmatrix} i \\ l \quad h \end{Bmatrix}_{P,m} = \left[\frac{\partial}{\partial x^m} \begin{Bmatrix} i \\ l \quad h \end{Bmatrix}\right]_{at\ P}$. Substituting Eqs. (iii) and (i) in Eq. (ii), we have

$$A^i_{SQ} = A^i_P - \begin{Bmatrix} i \\ j \quad k \end{Bmatrix}_P A^j_P \, dx^k - \left[\begin{Bmatrix} i \\ l \quad h \end{Bmatrix}_P + \begin{Bmatrix} i \\ l \quad h \end{Bmatrix}_{P,m} dx^m\right]$$

$$\left[A^l_P - \begin{Bmatrix} l \\ j \quad k \end{Bmatrix}_P A^j_P \, dx^k\right] \delta x^h$$

$$= A^i - \begin{Bmatrix} i \\ j \quad k \end{Bmatrix} A^j \, dx^k - \begin{Bmatrix} i \\ l \quad h \end{Bmatrix} A^l \delta x^h - \begin{Bmatrix} i \\ l \quad h \end{Bmatrix}_{,m} A^l \, dx^m \, \delta x^h$$

$$+ \begin{Bmatrix} i \\ l \quad h \end{Bmatrix} \begin{Bmatrix} l \\ j \quad k \end{Bmatrix} A^j \, dx^k \, dx^h + O\left[(dx^k)^3\right], \qquad (iv)$$

where we have dropped the subscript P from the right hand side of (iv) because all the quantities are to be evaluated at P. The vector A^i_{SR} obtained by a parallel displacement along the path PRS will be simply given by an interchange of dx^i and dx^i in the expression for A^i_{SQ}:

$$A^i_{SR} = A^i - \begin{Bmatrix} i \\ j \quad k \end{Bmatrix} A^j \, \delta x^k - \begin{Bmatrix} i \\ l \quad h \end{Bmatrix} A^l dx^h - \begin{Bmatrix} i \\ l \quad h \end{Bmatrix}_{,m} A^l \, \delta x^m \, dx^h$$

$$+ \begin{Bmatrix} l \\ l \quad h \end{Bmatrix} \begin{Bmatrix} l \\ j \quad k \end{Bmatrix} A^j \delta x^k \, dx^h + O\left[(dx^k)^3\right]. \qquad (v)$$

Subtracting (iv) from (v) we get

$$A^i_{SR} - A^i_{SR} = \begin{Bmatrix} i \\ l \quad h \end{Bmatrix}_{,m} A^l \, dx^m \, \delta x^h - \begin{Bmatrix} i \\ l \quad h \end{Bmatrix}_{,m} A^l \, \delta x^m \, dx^h$$

$$+ \begin{Bmatrix} i \\ l \quad h \end{Bmatrix} \begin{Bmatrix} l \\ j \quad k \end{Bmatrix} A^j \, \delta x^k \, dx^h - \begin{Bmatrix} i \\ l \quad h \end{Bmatrix} \begin{Bmatrix} l \\ j \quad k \end{Bmatrix} A^j \, dx^k \, \delta x^h. \qquad (vi)$$

Replacing l by j and m by k in the first two terms on the right hand side of (vi) and interchanging k and h in the second and third terms, the Eq. (vi) can be written as

$$A^i_{SR} - A^i_{SQ} = R^i_{jhk} \, A^j \, dx^k \, \delta x^h, \qquad \text{(vii)}$$

where

$$R^i_{jkh} = \begin{Bmatrix} i \\ j \quad h \end{Bmatrix}_{,k} - \begin{Bmatrix} i \\ j \quad k \end{Bmatrix}_{,h} + \begin{Bmatrix} i \\ l \quad k \end{Bmatrix} \begin{Bmatrix} i \\ j \quad h \end{Bmatrix} - \begin{Bmatrix} i \\ l \quad h \end{Bmatrix} \begin{Bmatrix} l \\ j \quad k \end{Bmatrix} \qquad (viii)$$

Since A^j, dx^k and δx^h are arbitrary vectors, if follows from quotient law that R^i_{jkh} is a tensor of rank four, known as the Riemann-Christoffel curvature tensor. It is independent of the vector A^i and depends only on the metric tensor and its first and second derivatives.

It follows that if a tensor is parallely displaced along a closed curve until we come back to the starting point, the resulting vector will not necessarily be the same as the original vector in a non-Euclidean space.

EXAMPLE 4.5.2 *If A^i and B^i are two vectors of constant magnitudes and undergo parallel displacements along a curve then show that they are inclined at a constant angle.*

Solution: Let the two vectors A^i and B^i be of constant magnitudes, so that their lengths do not change as we move along the curve Γ. Let these vectors suffer parallel displacement along a curve Γ, so by hypothesis

$$\frac{\delta A^i}{\delta t} = 0 = \frac{\delta B^i}{\delta t},$$

at each point of Γ. The angle θ between the vectors A^i and B^i, is given by

$$\cos\theta = g_{ij}A^iB^j.$$

The angle θ between the vectors A^i and B^i remains fixed as the parameter t is allowed to change. Here, $g_{ij}A^iB^j$ is an invariant. Differentiating intrinsically with respect to t, we get

$$\begin{aligned}\frac{\delta}{\delta t}(\cos\theta) &= \frac{\delta}{\delta t}\left(g_{ij}A^iB^j\right) = \frac{d}{dt}\left(g_{ij}A^iB^j\right)\\ &= g_{ij}\frac{\delta A^i}{\delta t}B^j + g_{ij}A^i\frac{\delta B^j}{\delta t} = g_{ij}\cdot 0\cdot B^j + g_{ij}A^i\cdot 0 = 0\end{aligned}$$

or

$$\frac{d}{dt}(\cos\theta) = 0 \Rightarrow \cos\theta = \text{constant, i.e. } \theta = \text{constant.}$$

Thus, we conclude that $g_{ij}A^iB^j$ is constant along Γ. It follows directly from the result that, if $A^i = B^i$, then $g_{ij}A^iA^j = A^2$ is constant along Γ and this implies $\theta = $ constant.

EXAMPLE 4.5.3 *Prove that geodesics are autoparallel curves.*

Solution: To prove the result, it is enough to prove that the unit tangent vector, say λ^i, suffers a parallel displacement along a geodesic curve Γ. Let $\Gamma : x^i = x^i(t)$ be a geodesic in a Riemannian space. Then the co-ordinates of the points on the geodesic satisfy the geodesic equation,

$$\frac{d^2x^i}{dt^2} + \left\{ \begin{matrix} i \\ j \ \ k \end{matrix} \right\} \frac{dx^j}{dt}\frac{dx^k}{dt} = 0$$

or

$$\frac{d\lambda^i}{dt} + \left\{ \begin{matrix} i \\ j \ \ k \end{matrix} \right\} \lambda^j \frac{dx^k}{dt} = 0$$

or

$$\left[\frac{\partial\lambda^i}{\partial x^k} + \left\{ \begin{matrix} i \\ j \ \ k \end{matrix} \right\} \lambda^j\right] \frac{dx^k}{dt} = 0$$

or

$$\lambda^i_{,k}\frac{dx^k}{dt} = 0; \text{ i.e. } \lambda^i_{,j}\frac{dx^j}{dt} = 0,$$

at each point on Γ. This result shows that the unit vectors form a system of parallel vectors along a geodesic. Hence, it follows from the definition of parallel displacement that the field of tangent vectors to a geodesic is a field of parallel vectors.

EXAMPLE 4.5.4 *Any vector of constant magnitude which undergoes a parallel displacement along a geodesic is inclined at a constant angle to the curve.*

Solution: Let a vector A^i of constant magnitude undergo a parallel displacement along a geodesic Γ, so that,

$$\frac{\delta A^i}{\delta t} = A^i_{,j}\frac{dx^j}{dt} = 0$$

at each point of Γ. Let λ^i be a unit tangent vector to the curve Γ, so that,

$$\frac{\delta\lambda^i}{\delta t} = \lambda^i_{,j}\frac{dx^j}{dt} = 0$$

at each point of Γ, since geodesics are autoparallel curves. Let θ be the angle between the vector λ^i and a curve Γ at any point of Γ, i.e. θ is the angle between A^i and λ^i, then,

$$A^i\lambda_i = A \cdot 1 \cdot \cos\theta$$

or

$$\frac{d}{dt}(A\cos\theta) = \frac{d}{dt}(A^i\lambda_i) = (A^i\lambda_i)_{,j}\frac{dx^i}{dt}$$

or

$$\begin{aligned} -A\sin\theta\frac{d\theta}{dt} &= A^i_{,j}\frac{dx^j}{dt}\lambda_i + A^i\lambda_{i,j}\frac{dx^j}{dt} \\ &= 0t_1 + A^i 0 = 0. \end{aligned}$$

or

$$\sin\theta\frac{d\theta}{dt} = 0; \quad \text{as } A \neq 0,$$

or

$$\sin\theta = 0 \quad \text{or } \frac{d\theta}{dt} = 0 \Rightarrow \theta = 0 \quad \text{or } \theta = \text{ constant.}$$

EXAMPLE 4.5.5 *If the intrinsic derivative of a vector vanishes, then show that the magnitude of the vector is constant.*

Solution: Let the intrinsic derivative of a vector A^i be zero so that

$$\frac{\delta A^i}{\delta t} = A^i_{,j}\frac{dx^j}{dt} = 0.$$

We have to show that $A^2 = A^i A_i$ is constant. Differentiating both sides with respect to a parameter t and noting that the intrinsic derivative of a scalar is ordinary derivative, we have

$$\begin{aligned}\frac{dA^2}{dt} &= \frac{d}{dt}\left(A^i A_i\right) = \left(A^i A_i\right)_{,j}\frac{dx^j}{ds}\\ &= A^i_{,j}\frac{dx^j}{dt}A_i + A^i A_{i,j}\frac{dx^j}{ds} = 0A_i + A^i 0 = 0\end{aligned}$$

or

$$A^2 = \text{constant } \Rightarrow A = \text{constant.}$$

Thus, if the derived vector of a given vector A in the direction of a given curve Γ vanishes at all points of a curve, then the magnitude of the vector is constant. If A^i forms a field of parallel vectors along the curve $x^i = x(t)$, then we have

$$\frac{\delta A^i}{\delta t} = A^i_{,j}\frac{dx^j}{dt} = 0.$$

Then similarly, as above we can show that $A =$ constant. Thus, the magnitude of a vector remain invariant under parallel displacement.

EXAMPLE 4.5.6 *Show that the parallel displacement of a vector taken all around a circle on the surface of a sphere does not lead back to the same vector except when the circle is a great circle, that is, a geodesic.*

Solution: Choose spherical polar co-ordinates $x^1 = \theta$ and $x^2 = \phi$ on the surface of the sphere. The non-vanishing Christoffel symbols are,

$$\left\{\begin{matrix} 2 \\ 1\ 2 \end{matrix}\right\} = \cot\theta = \left\{\begin{matrix} 2 \\ 2\ 1 \end{matrix}\right\}; \quad \left\{\begin{matrix} 1 \\ 2\ 2 \end{matrix}\right\} = -\sin\theta\cos\theta.$$

Choose a small circle normal to the polar axis so that its parametric equation is

$$\theta = \alpha, \phi = ps + q; \quad p \neq 0 \text{ with } \frac{d\theta}{dt} = 0.$$

Consider a vector $A^i = (a^1, A^2)$ at a point on the circle which may be taken to be $\phi = 0$. When the vector is displaced parallely along the circle, the change in its components is given, from Eq. (4.69), by

$$\frac{dA^1}{dt} = -\left\{ \begin{matrix} 1 \\ j \quad k \end{matrix} \right\} A^j \frac{dx^k}{dt} = A^2 \sin\alpha \cos\alpha \frac{d\phi}{dt} \tag{i}$$

$$\frac{dA^2}{dt} = -\left\{ \begin{matrix} 2 \\ j \quad k \end{matrix} \right\} A^j \frac{dx^k}{dt} = -A \cot\alpha \frac{d\phi}{dt}. \tag{ii}$$

Now, for any function f of θ and ϕ through t, we have

$$\frac{df}{dt} = \frac{\partial f}{\partial \theta}\frac{d\theta}{dt} + \frac{\partial f}{\partial \phi}\frac{d\phi}{dt} = \frac{\partial f}{\partial \phi}\frac{d\phi}{dt},$$

which reduces Eqs. (i) and (ii) to

$$\frac{\partial A^1}{\partial \phi} = A^2 \sin\alpha\cos\alpha; \quad \frac{\partial A^2}{\partial \phi} = -A^1 \cot\alpha. \tag{iii}$$

Differentiating once again with respect to ϕ and substituting one in the other, Eq. (iii) give rise to the decoupled equations

$$\frac{\partial^2 A^1}{\partial \phi^2} = -A^1 \cos^2\alpha; \quad \frac{\partial^2 A^2}{\partial \phi^2} = -A^2 \cos^2\alpha, \tag{iv}$$

whose solution, consistent with Eq. (iii), is

$$\begin{aligned} A^1 &= \sin\alpha\,[c\cos(\phi\cos\alpha) + d\sin(\phi\cos\alpha)] \\ A^2 &= d\cos(\phi\cos\alpha) - c\sin(\phi\cos\alpha). \end{aligned}$$

At $\phi = 0$, the components of the vector are given by

$$A^1(\phi = 0) = c\sin\alpha, \quad A^2(\phi = 0) = d, \tag{v}$$

whereas, after parallel displacement once around the circle, the components become,

$$\begin{aligned} A^1(\phi = 2\pi) &= \sin\alpha\,[c\cos(2\pi\cos\alpha) + d\sin(2\pi\cos\alpha)] \\ A^2(\phi = 2\pi) &= d\cos(2\pi\cos\alpha) - c\sin(2\pi\cos\alpha), \end{aligned}$$

which are not the same as those given by Eq. (v) except when $\alpha = \frac{\pi}{2}$, i.e. when the displacement is along the geodesic.

CHAPTER 1

Gyrogroups

Gyrogroups are generalized groups, which are best motivated by the algebra of Möbius transformations of the complex open unit disc. Groups are classified into commutative and noncommutative groups and, in full analogy, gyrogroups are classified into gyrocommutative and nongyrocommutative gyrogroups. Some commutative groups admit scalar multiplication, giving rise to vector spaces. In full analogy, some gyrocommutative gyrogroups admit scalar multiplication, giving rise to gyrovector spaces. Furthermore, vector spaces form the algebraic setting for the standard model of Euclidean geometry and, in full analogy, gyrovector spaces form the algebraic setting for various models of the hyperbolic geometry of Bolyai and Lobachevsky.

The evolution from Möbius to gyrogroups began soon after the discovery in 1988 [48] that Einstein velocity addition law of relativistically admissible velocities, which is seemingly structureless, is in fact rich of structure. Later, the rich structure that Einstein velocity addition law encodes turned out to be a gyrocommutative gyrogroup and a gyrovector space. The resulting notions of gyrogroups and gyrovector spaces preserve the flavor of their classical counterparts, groups and vector spaces. They are useful and fascinating enough to be made part of the lore learned by all undergraduate and graduate mathematics and physics students. Being a natural generalization of groups and vector spaces, gyrogroups and gyrovector spaces lay a fruitful bridge between nonassociative algebra and hyperbolic geometry, just as groups and vector spaces lay a fruitful bridge between associative algebra and Euclidean geometry. More than 150 years have passed since the German mathematician August Ferdinand Möbius (1790 – 1868) first studied the transformations that now bear his name [30, p. 71]. Yet, the rich structure he thereby exposed is still far from being exhausted, as the evolution from Möbius to gyrogroups demonstrates in Sections 1.1 – 1.3 of this chapter.

1.1 FROM MÖBIUS TO GYROGROUPS

Möbius transformations of the complex open unit disc $\mathbb{D}$,

$$\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\} \tag{1.1}$$

of the complex plane $\mathbb{C}$ are studied in most books on function theory of one complex variable. According to [31, p. 176], these are important for at least two reasons: (i) they play a central role in non-Euclidean geometry [51], and (ii) they are the only automorphisms of the disc.

Ahlfors' book [1], *Conformal Invariants: Topics in Geometric Function Theory*, begins with a presentation of the *polar decomposition* of the most general Möbius self-transformation of the complex open unit disc $\mathbb{D}$,

$$z \mapsto e^{i\theta} \frac{a+z}{1+\bar{a}z} = e^{i\theta}(a \oplus z) , \tag{1.2}$$

where $a, z \in \mathbb{D}$, $\overline{a}$ being the complex conjugate of a, and where $\theta \in \mathbb{R}$ is a real number [12, p. 211], [17, p. 185], [23, pp. 6 – 7], [31, pp. 177 – 178]. In (1.2) we present the Möbius polar decomposition in a form that suggests the *Möbius addition*, $\oplus$, which we define by the equation

$$a \oplus z = \frac{a + z}{1 + \overline{a}z} \,. \tag{1.3}$$

Naturally, Möbius subtraction, $\ominus$, is given by $a \ominus z = a \oplus (-z)$, so that $z \ominus z = 0$ and $\ominus z = 0 \ominus z = 0 \oplus (-z) = -z$.

Interestingly, Möbius addition possesses the *gyroautomorphic inverse property*

$$\ominus(a \oplus b) = \ominus a \ominus b \tag{1.4}$$

and the *left cancellation law*

$$\ominus a \oplus (a \oplus z) = z \tag{1.5}$$

for all $a, b, z \in \mathbb{D}$. Contrasting the left cancellation law (1.5) of Möbius addition, care must be taken with its right counterpart since, in general,

$$(a \oplus z) \ominus z \neq a \,. \tag{1.6}$$

Indeed, in (1.15) below we will introduce a secondary binary operation, $\boxplus$, which is a binary *cooperation* called a Möbius *coaddition*, and which gives rise to the *right cancellation law*, (1.18), of Möbius addition.

Möbius addition is neither commutative nor associative. Fortunately, the failure of associativity is a source of rich theory, the theory of gyrogroups and gyrovector spaces, as we will see in this book.

To measure the extent of which Möbius addition deviates from commutativity we define the *gyrations*

$$\mathrm{gyr}[a, b] = \frac{a \oplus b}{b \oplus a} = \frac{1 + a\overline{b}}{1 + \overline{a}b} \tag{1.7}$$

for $a, b \in \mathbb{D}$. Being unimodular complex numbers, $|\mathrm{gyr}[a, b]| = 1$, gyrations represent rotations of the disc $\mathbb{D}$ about its center. In the special case when $a = 0$ or $b = 0$ or $a = \lambda b$ for some real number λ, the gyration $\mathrm{gyr}[a, b]$ is trivial, that, is, $\mathrm{gyr}[a, b] = 1$. The number -1 cannot be represented by a gyration since the equation $\mathrm{gyr}[a, b] = -1$ has no solution for a and b in the disc $\mathbb{D}$.

Gyrations respect Möbius addition,

$$\mathrm{gyr}[a, b](c \oplus d) = \mathrm{gyr}[a, b]c \oplus \mathrm{gyr}[a, b]d \tag{1.8}$$

for all $a, b, c, d \in \mathbb{D}$, and the inverse $\mathrm{gyr}^{-1}[a, b]$ of a gyration $\mathrm{gyr}[a, b]$ is the gyration $\mathrm{gyr}[b, a]$,

$$\mathrm{gyr}^{-1}[a, b] = \mathrm{gyr}[b, a] \,, \tag{1.9}$$

as we see from (1.7). Hence, gyrations are special automorphisms of the Möbius groupoid $(\mathbb{D}, \oplus)$. We recall that a groupoid $(G, \oplus)$ is a nonempty set, G, with a binary operation, $\oplus$, and that an automorphism of a groupoid $(G, \oplus)$ is a bijective self map f of G that respects its binary operation $\oplus$, that is, $f(a\oplus b) = f(a)\oplus f(b)$. The set of all automorphisms of a groupoid $(G, \oplus)$ forms a group, denoted $Aut(G, \oplus)$. The formal definition of groups will be presented in Def. 1.3, p. 5.

To emphasize that gyrations are automorphisms, they are also called *gyroautomorphisms*. The automorphism $\phi(z) = -z$ of the disc is not a gyroautomorphism of the disc since the equation $(1 + a\bar{b})/(1 + \bar{a}b) = -1$ for $a, b \in \mathbb{D}$ has no solution in the disc. Consequently, the gyroautomorphisms of the disc do not form a group under gyroautomorphism composition.

It follows from the gyration definition in (1.7) that Möbius addition obeys the *gyrocommutative law*:

$$a\oplus b = \text{gyr}[a, b](b\oplus a) \tag{1.10}$$

for all $a, b \in \mathbb{D}$. The Möbius gyrocommutative law (1.10) is not terribly surprising since it is generated by definition. But, we are not finished!

Coincidentally, the gyration gyr$[a, b]$, which "repairs" in (1.10) the breakdown of commutativity in Möbius addition, "repairs" the breakdown of associativity in Möbius addition as well. Indeed, it gives rise to the following gyroassociative laws (left and right),

$$\begin{aligned} a\oplus(b\oplus z) &= (a\oplus b)\oplus\text{gyr}[a, b]z\,, \\ (a\oplus b)\oplus z &= a\oplus(b\oplus\text{gyr}[b, a]z)\,, \end{aligned} \tag{1.11}$$

for all $a, b, z \in \mathbb{D}$. Hence, gyrations measure simultaneously the extent to which Möbius addition deviates from both associativity and commutativity.

Möbius gyrations are expressible in terms of Möbius addition. Indeed, solving the first identity in (1.11) for gyr$[a, b]z$ by means of the Möbius left cancellation law (1.5) we have

$$\text{gyr}[a, b]z = \ominus(a\oplus b)\oplus\{a\oplus(b\oplus z)\}\,. \tag{1.12}$$

Möbius gyrations gyr$[a, b]$, $a, b \in \mathbb{D}$, thus endow the Möbius disc groupoid $(\mathbb{D}, \oplus)$ with a group-like structure, naturally called a *gyrogroup*. In fact, Möbius gyrations do more than that. They contribute their own rich structure to the gyrogroup. Thus, for instance, they possess the *loop properties* (left and right)

$$\begin{aligned} \text{gyr}[a\oplus b, b] &= \text{gyr}[a, b]\,, \\ \text{gyr}[a, b\oplus a] &= \text{gyr}[a, b]\,, \end{aligned} \tag{1.13}$$

and the nested gyration identity

$$\text{gyr}[b, \ominus\text{gyr}[b, a]a] = \text{gyr}[a, b]\,, \tag{1.14}$$

as one can straightforwardly check.

Möbius addition $\oplus$, as a primary addition in the disc, comes with *Möbius coaddition*, $\boxplus$, as a secondary addition in the disc. Möbius coaddition is defined in terms of Möbius addition and Möbius gyrations by the equation

$$a \boxplus b = a\oplus\text{gyr}[a, \ominus b]b \tag{1.15}$$

for all $a, b \in \mathbb{D}$. It turns out that Möbius coaddition is commutative (but not associative), given by the equation

$$a \boxplus b = \frac{(1-|a|^2)b + (1-|b|^2)a}{1-|a|^2|b|^2} \tag{1.16}$$

for all $a, b \in \mathbb{D}$, as we see from (1.15) and (1.7).

Replacing b by $\ominus b$ in (1.15), and noting that $\mathrm{gyr}[a, \ominus b]$ is an automorphism, it follows from (1.15) that

$$a \boxminus b = a \oplus \mathrm{gyr}[a, b](\ominus b) = a \ominus \mathrm{gyr}[a, b]b\,. \tag{1.17}$$

for all $a, b \in \mathbb{D}$, where we use the obvious notation, $a \boxminus b = a \boxplus (\ominus b)$.

In terms of Möbius coaddition, the *right cancellation law* of Möbius addition takes the form

$$(a \oplus b) \boxminus b = a\,. \tag{1.18}$$

The right cancellation law (1.18) is associated with the second right cancellation law

$$(a \boxplus b) \ominus b = a\,, \tag{1.19}$$

as one can straightforwardly check. Identities (1.18) and (1.19) present one of the duality symmetries that Möbius addition and coaddition share.

Guided by the group-like structure of Möbius addition, we define a *gyrogroup* to be a group in which the associative law is replaced by the gyroassociative laws (left and right), and a *gyrocommutative gyrogroup* to be a commutative group in which the commutative and associative laws are replaced by the gyrocommutative and gyroassociative laws. Like Möbius gyrations, the gyrations in the gyrocommutative and gyroassociative laws of an abstract gyrogroup must be automorphisms of the gyrogroup.

Thus, the disc $\mathbb{D}$ with its Möbius addition forms a *gyrocommutative gyrogroup* $(\mathbb{D}, \oplus)$ where, strikingly, the concepts of gyrocommutativity and gyroassociativity of the gyrogroup preserve the flavor of their classical counterparts [50, 51]. Mathematical coincidences are not accidental! Hence, in particular, the gyrations that "coincidentally" repair in (1.10) – (1.11) the breakdown of commutativity and associativity need not be peculiar to Möbius addition, as shown in [13, 14], and as we will see in this book.

Indeed, the generalization of groups into gyrogroups that Möbius addition suggests bears an intriguing resemblance to the generalization of the rational numbers to the real ones. The beginner is initially surprised to discover an irrational number, like $\sqrt{2}$, but soon later he or she is likely to realize that there are more irrational numbers than rational ones. Similarly, the gyrogroup structure of Möbius addition initially comes as a surprise. But, interested explorers may soon realize that there are indications that, in some sense, there are more non-group gyrogroups than groups [13].

+,- * *. * $ +A + 7 + D
0D/ A ; D + " + A 3 S / 4 D . , B D)E T "
A / D /"
(! 4 @ & 6 @ ' 4 3 ;D "/

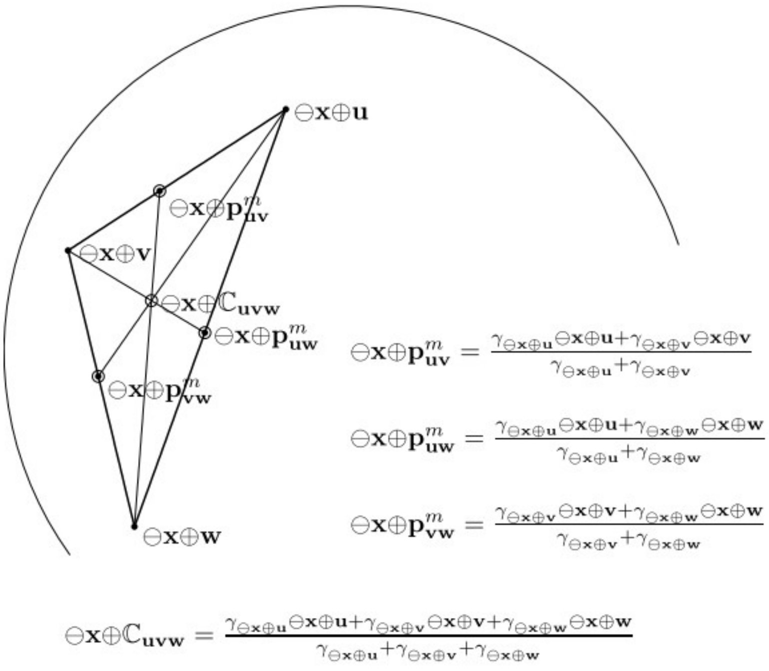

$\ominus\mathbf{x}\oplus\mathbf{u}$
$\ominus\mathbf{x}\oplus\mathbf{p}^m_{\mathbf{uv}}$
$\ominus\mathbf{x}\oplus\mathbf{v}$
$\ominus\mathbf{x}\oplus\mathbb{C}_{\mathbf{uvw}}$
$\ominus\mathbf{x}\oplus\mathbf{p}^m_{\mathbf{uw}}$
$\ominus\mathbf{x}\oplus\mathbf{p}^m_{\mathbf{vw}}$
$\ominus\mathbf{x}\oplus\mathbf{w}$

$$\ominus\mathbf{x}\oplus\mathbf{p}^m_{\mathbf{uv}} = \frac{\gamma_{\ominus\mathbf{x}\oplus\mathbf{u}}\ominus\mathbf{x}\oplus\mathbf{u}+\gamma_{\ominus\mathbf{x}\oplus\mathbf{v}}\ominus\mathbf{x}\oplus\mathbf{v}}{\gamma_{\ominus\mathbf{x}\oplus\mathbf{u}}+\gamma_{\ominus\mathbf{x}\oplus\mathbf{v}}}$$

$$\ominus\mathbf{x}\oplus\mathbf{p}^m_{\mathbf{uw}} = \frac{\gamma_{\ominus\mathbf{x}\oplus\mathbf{u}}\ominus\mathbf{x}\oplus\mathbf{u}+\gamma_{\ominus\mathbf{x}\oplus\mathbf{w}}\ominus\mathbf{x}\oplus\mathbf{w}}{\gamma_{\ominus\mathbf{x}\oplus\mathbf{u}}+\gamma_{\ominus\mathbf{x}\oplus\mathbf{w}}}$$

$$\ominus\mathbf{x}\oplus\mathbf{p}^m_{\mathbf{vw}} = \frac{\gamma_{\ominus\mathbf{x}\oplus\mathbf{v}}\ominus\mathbf{x}\oplus\mathbf{v}+\gamma_{\ominus\mathbf{x}\oplus\mathbf{w}}\ominus\mathbf{x}\oplus\mathbf{w}}{\gamma_{\ominus\mathbf{x}\oplus\mathbf{v}}+\gamma_{\ominus\mathbf{x}\oplus\mathbf{w}}}$$

$$\ominus\mathbf{x}\oplus\mathbb{C}_{\mathbf{uvw}} = \frac{\gamma_{\ominus\mathbf{x}\oplus\mathbf{u}}\ominus\mathbf{x}\oplus\mathbf{u}+\gamma_{\ominus\mathbf{x}\oplus\mathbf{v}}\ominus\mathbf{x}\oplus\mathbf{v}+\gamma_{\ominus\mathbf{x}\oplus\mathbf{w}}\ominus\mathbf{x}\oplus\mathbf{w}}{\gamma_{\ominus\mathbf{x}\oplus\mathbf{u}}+\gamma_{\ominus\mathbf{x}\oplus\mathbf{v}}+\gamma_{\ominus\mathbf{x}\oplus\mathbf{w}}}$$

Fig. 6.17 A left gyrotranslation by $\mathbf{x}\in(\mathbb{R}^2_s, \oplus, \otimes)$ of the gyrotriangle **uvw** of Fig. 6.16 in an Einstein plane is shown. Gyromidpoints are gyrocovariant under left gyrotranslations, Theorem 6.37. Hence, a left gyrotranslation of a gyrotriangle does not distort its side-gyromidpoints and gyrocentroid. Accordingly, related formulas are left gyrotranslated gyrocovariantly.

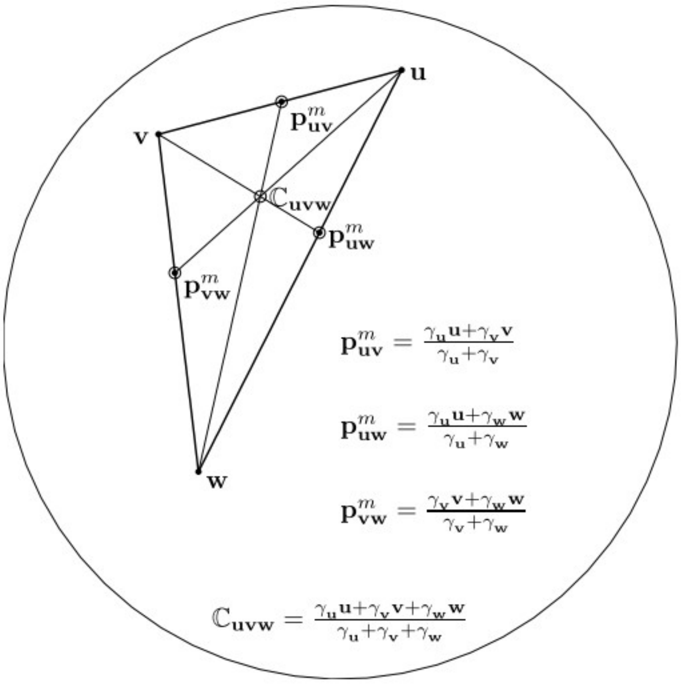
$\mathbf{u}$
$\mathbf{v}$
$\mathbf{w}$
$\mathbf{p}_{\mathbf{uv}}^m$
$\mathbb{C}_{\mathbf{uvw}}$
$\mathbf{p}_{\mathbf{uw}}^m$
$\mathbf{p}_{\mathbf{vw}}^m$
$\mathbf{p}_{\mathbf{uv}}^m = \frac{\gamma_{\mathbf{u}}\mathbf{u}+\gamma_{\mathbf{v}}\mathbf{v}}{\gamma_{\mathbf{u}}+\gamma_{\mathbf{v}}}$
$\mathbf{p}_{\mathbf{uw}}^m = \frac{\gamma_{\mathbf{u}}\mathbf{u}+\gamma_{\mathbf{w}}\mathbf{w}}{\gamma_{\mathbf{u}}+\gamma_{\mathbf{w}}}$
$\mathbf{p}_{\mathbf{vw}}^m = \frac{\gamma_{\mathbf{v}}\mathbf{v}+\gamma_{\mathbf{w}}\mathbf{w}}{\gamma_{\mathbf{v}}+\gamma_{\mathbf{w}}}$
$\mathbb{C}_{\mathbf{uvw}} = \frac{\gamma_{\mathbf{u}}\mathbf{u}+\gamma_{\mathbf{v}}\mathbf{v}+\gamma_{\mathbf{w}}\mathbf{w}}{\gamma_{\mathbf{u}}+\gamma_{\mathbf{v}}+\gamma_{\mathbf{w}}}$

Fig. 6.16 The gyromidpoints of the three sides of a gyrotriangle **uvw** in the Einstein gyrovector plane $(\mathbb{R}_s^2, \oplus_{\text{E}}, \otimes_{\text{E}})$ are shown along with its gyromedians and gyrocentroid. Interestingly, Einsteinian gyromidpoints and gyrocentroids have interpretation in relativistic mechanics, fully analogous to the interpretation of Euclidean midpoints and centroids in classical mechanics that one discovers in the vector space approach to Euclidean geometry [Hausner (1998)]; see Fig. 10.3.

We will now study the gyrotriangle gyromedians and gyrocentroid in Einstein, Fig. 6.16, Möbius,Fig. 6.20, and PV gyrovector spaces, Fig. 6.21, [Ungar (2004a)].

6.22.1 *In Einstein Gyrovector Spaces*

Einstein gyrovector spaces $(\mathbb{V}_s, \oplus_{\text{E}}, \otimes_{\text{E}})$ are particularly suitable for the study of gyromedians and gyrocentroids since gyrolines are Euclidean straight lines so that the calculation of points of intersection of gyrolines can be performed by methods of linear algebra. The resulting determination of gyrocentroids in Einstein gyrovector spaces, Fig. 6.16, can readily be translated into other isomorphic gyrovector spaces.

It follows from (6.330), by (3.195), p. 92, and the scalar associative law of gyrovector spaces, that the gyromidpoint $\mathbf{p}_{\mathbf{uv}}^{m}$ in Einstein gyrovector spaces is given by the equation

$$\begin{aligned}\mathbf{p}_{\mathbf{uv}}^{m} &= \tfrac{1}{2}\otimes_{\rm E}(\mathbf{u}\boxplus_{\rm E}\mathbf{v})\\ &= \tfrac{1}{2}\otimes_{\rm E}\{2\otimes_{\rm E}\tfrac{\gamma_{\mathbf{u}}\mathbf{u}+\gamma_{\mathbf{v}}\mathbf{v}}{\gamma_{\mathbf{u}}+\gamma_{\mathbf{v}}}\}\\ &= \frac{\gamma_{\mathbf{u}}\mathbf{u}+\gamma_{\mathbf{v}}\mathbf{v}}{\gamma_{\mathbf{u}}+\gamma_{\mathbf{v}}}\end{aligned} \tag{6.331}$$

We have thus obtained the following

Theorem 6.92 (The Einstein Gyromidpoint). *Let* $\mathbf{u},\mathbf{v}\in\mathbb{V}_s$ *be any two points of an Einstein gyrovector space* $(\mathbb{V}_s,\oplus_{\rm E},\otimes_{\rm E})$. *The gyromidpoint* $\mathbf{p}_{\mathbf{uv}}^{m}$ *of the gyrosegment* $\mathbf{uv}$ *is given by the equation*

$$\mathbf{p}_{\mathbf{uv}}^{m} = \frac{\gamma_{\mathbf{u}}\mathbf{u}+\gamma_{\mathbf{v}}\mathbf{v}}{\gamma_{\mathbf{u}}+\gamma_{\mathbf{v}}} \tag{6.332}$$

The gyromidpoints of the three sides of a gyrotriangle $\mathbf{uvw}$ in the Einstein gyrovector plane $(\mathbb{R}_s^2,\oplus_{\rm E},\otimes_{\rm E})$, and its gyrocentroid are shown in Fig. 6.16. Gyromidpoints are gyrocovariant under left gyrotranslations, Theorem 6.37. Therefore a left gyrotranslation by $\mathbf{x}\in\mathbb{R}_s^2$ of the gyrosegment $\mathbf{uv}$ in Fig. 6.16 into the gyrosegment $(\ominus\mathbf{x}\oplus\mathbf{u})(\ominus\mathbf{x}\oplus\mathbf{v})$ in Fig. 6.17 does not distort the gyrosegment gyromidpoint. Hence, the gyromidpoint of the gyrotranslated gyrosegment $(\ominus\mathbf{x}\oplus\mathbf{u})(\ominus\mathbf{x}\oplus\mathbf{v})$, Fig. 6.17, is given by the equation

$$\ominus\mathbf{x}\oplus\mathbf{p}_{\mathbf{uv}}^{m} = \mathbf{p}_{(\ominus\mathbf{x}\oplus\mathbf{u})(\ominus\mathbf{x}\oplus\mathbf{v})}^{m} \tag{6.333}$$

thus uncovering the interesting identity

$$\ominus\mathbf{x}\oplus\frac{\gamma_{\mathbf{u}}\mathbf{u}+\gamma_{\mathbf{v}}\mathbf{v}}{\gamma_{\mathbf{u}}+\gamma_{\mathbf{v}}} = \frac{\gamma_{\ominus\mathbf{x}\oplus\mathbf{u}}\ominus\mathbf{x}\oplus\mathbf{u}+\gamma_{\ominus\mathbf{x}\oplus\mathbf{v}}\ominus\mathbf{x}\oplus\mathbf{v}}{\gamma_{\ominus\mathbf{x}\oplus\mathbf{u}}+\gamma_{\ominus\mathbf{x}\oplus\mathbf{v}}} \tag{6.334}$$

in any Einstein gyrovector space.

Gyrolines in the Beltrami-Klein ball model of hyperbolic geometry, that is, in Einstein gyrovector spaces, are Euclidean straight lines. Hence, by elementary techniques of linear algebra, one can verify that the three gyromedians in Fig. 6.16 are concurrent, and that the point of concurrency, that is, the gyrocentroid $C_{\mathbf{uvw}}$ of gyrotriangle $\mathbf{uvw}$, is given by the elegant equation (6.335) that we place in the following theorem.

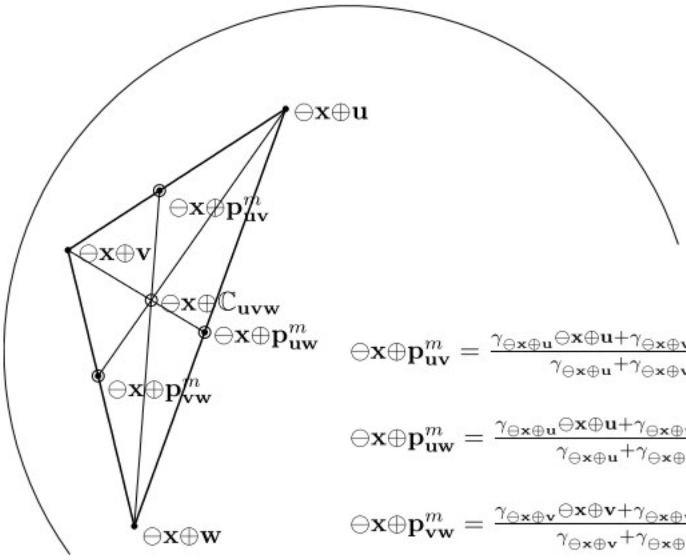
$\ominus\mathbf{x}\oplus\mathbf{u}$
$\ominus\mathbf{x}\oplus\mathbf{p}_{\mathbf{uv}}^{m}$
$\ominus\mathbf{x}\oplus\mathbf{v}$
$\ominus\mathbf{x}\oplus\mathbb{C}_{\mathbf{uvw}}$
$\ominus\mathbf{x}\oplus\mathbf{p}_{\mathbf{uw}}^{m}$
$\ominus\mathbf{x}\oplus\mathbf{p}_{\mathbf{vw}}^{m}$
$\ominus\mathbf{x}\oplus\mathbf{w}$

$$\ominus\mathbf{x}\oplus\mathbf{p}_{\mathbf{uv}}^{m} = \frac{\gamma_{\ominus\mathbf{x}\oplus\mathbf{u}}\ominus\mathbf{x}\oplus\mathbf{u}+\gamma_{\ominus\mathbf{x}\oplus\mathbf{v}}\ominus\mathbf{x}\oplus\mathbf{v}}{\gamma_{\ominus\mathbf{x}\oplus\mathbf{u}}+\gamma_{\ominus\mathbf{x}\oplus\mathbf{v}}}$$

$$\ominus\mathbf{x}\oplus\mathbf{p}_{\mathbf{uw}}^{m} = \frac{\gamma_{\ominus\mathbf{x}\oplus\mathbf{u}}\ominus\mathbf{x}\oplus\mathbf{u}+\gamma_{\ominus\mathbf{x}\oplus\mathbf{w}}\ominus\mathbf{x}\oplus\mathbf{w}}{\gamma_{\ominus\mathbf{x}\oplus\mathbf{u}}+\gamma_{\ominus\mathbf{x}\oplus\mathbf{w}}}$$

$$\ominus\mathbf{x}\oplus\mathbf{p}_{\mathbf{vw}}^{m} = \frac{\gamma_{\ominus\mathbf{x}\oplus\mathbf{v}}\ominus\mathbf{x}\oplus\mathbf{v}+\gamma_{\ominus\mathbf{x}\oplus\mathbf{w}}\ominus\mathbf{x}\oplus\mathbf{w}}{\gamma_{\ominus\mathbf{x}\oplus\mathbf{v}}+\gamma_{\ominus\mathbf{x}\oplus\mathbf{w}}}$$

$$\ominus\mathbf{x}\oplus\mathbb{C}_{\mathbf{uvw}} = \frac{\gamma_{\ominus\mathbf{x}\oplus\mathbf{u}}\ominus\mathbf{x}\oplus\mathbf{u}+\gamma_{\ominus\mathbf{x}\oplus\mathbf{v}}\ominus\mathbf{x}\oplus\mathbf{v}+\gamma_{\ominus\mathbf{x}\oplus\mathbf{w}}\ominus\mathbf{x}\oplus\mathbf{w}}{\gamma_{\ominus\mathbf{x}\oplus\mathbf{u}}+\gamma_{\ominus\mathbf{x}\oplus\mathbf{v}}+\gamma_{\ominus\mathbf{x}\oplus\mathbf{w}}}$$

Fig. 6.17 A left gyrotranslation by $\mathbf{x}\in(\mathbb{R}_s^2, \oplus, \otimes)$ of the gyrotriangle **uvw** of Fig. 6.16 in an Einstein plane is shown. Gyromidpoints are gyrocovariant under left gyrotranslations, Theorem 6.37. Hence, a left gyrotranslation of a gyrotriangle does not distort its side-gyromidpoints and gyrocentroid. Accordingly, related formulas are left gyrotranslated gyrocovariantly.

Theorem 6.93 (The Einstein Gyrotriangle Gyrocentroid). *Let* $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{V}_s$ *be any three nongyrocollinear points of an Einstein gyrovector space* $(\mathbb{V}_s, \oplus_E, \otimes_E)$. *The gyrocentroid* $\mathbb{C}_{\mathbf{uvw}}^{m}$ *of the gyrotriangle* **uvw**, *Fig. 6.16, is given by the equation*

$$\mathbb{C}_{\mathbf{uvw}} = \frac{\gamma_{\mathbf{u}}\mathbf{u} + \gamma_{\mathbf{v}}\mathbf{v} + \gamma_{\mathbf{w}}\mathbf{w}}{\gamma_{\mathbf{u}} + \gamma_{\mathbf{v}} + \gamma_{\mathbf{w}}} \tag{6.335}$$

Gyromidpoints are gyrocovariant, Theorem 6.37. Therefore a left gyrotranslation by $\mathbf{x}\in\mathbb{R}_s^2$ of the gyrotriangle **uvw** in Fig. 6.16 into the gyrotriangle $(\ominus\mathbf{x}\oplus\mathbf{u})(\ominus\mathbf{x}\oplus\mathbf{v})(\ominus\mathbf{x}\oplus\mathbf{w})$ in Fig. 6.17 does not distort the gyrotriangle side gyromidpoints and gyrocentroid. Hence, the gyrocentroid of the gyrotranslated gyrotriangle $(\ominus\mathbf{x}\oplus\mathbf{u})(\ominus\mathbf{x}\oplus\mathbf{v})(\ominus\mathbf{x}\oplus\mathbf{w})$, Fig. 6.17, is given by the

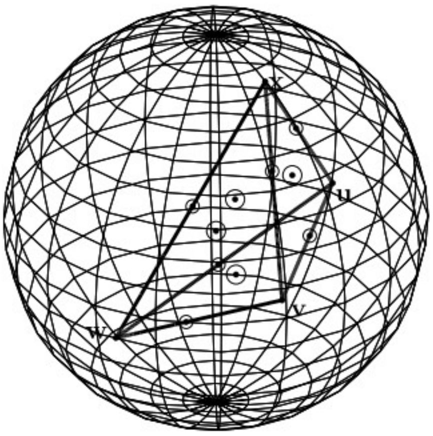

Fig. 6.18 The hyperbolic tetrahedron $\mathbf{uvwx}$, that is, a gyrotetrahedron, is shown in the Einstein gyrovector space
$$\mathbb{R}^3_s = (\mathbb{R}^3_s, \oplus_E, \otimes_E)$$
underlying the Beltrami ball model of hyperbolic geometry. The gyrotetrahedron $\mathbf{uvwx}$ is shown inside the s-ball $\mathbb{R}^3_s$ of the Euclidean 3-space $\mathbb{R}^3$ where it lives. The faces of the gyrotetrahedron are gyrotriangles.

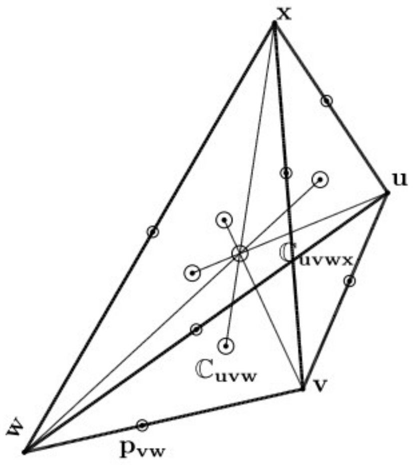

Fig. 6.19 Shown are also the gyromidpoints of the 6 sides and the gyrocentroids of the 4 faces. The gyroline joining a vertex of a gyrotetrahedron and the hyperbolic centroid, gyrocentroid, of the opposite face is called a gyrotetrahedron gyromedian. The four gyromedians of the gyrotetrahedron $\mathbf{uvwx}$ are concurrent. The point of concurrency is the gyrotetrahedron gyrocentroid $\mathbb{C}_{\mathbf{uvwx}}$.

equation

$$\begin{aligned} \ominus\mathbf{x}\oplus\mathbb{C}_{\mathbf{uvw}} &= \frac{\gamma_{\ominus\mathbf{x}\oplus\mathbf{u}}\ominus\mathbf{x}\oplus\mathbf{u} + \gamma_{\ominus\mathbf{x}\oplus\mathbf{v}}\ominus\mathbf{x}\oplus\mathbf{v} + \gamma_{\ominus\mathbf{x}\oplus\mathbf{w}}\ominus\mathbf{x}\oplus\mathbf{w}}{\gamma_{\ominus\mathbf{x}\oplus\mathbf{u}} + \gamma_{\ominus\mathbf{x}\oplus\mathbf{v}} + \gamma_{\ominus\mathbf{x}\oplus\mathbf{w}}} \\ &= \mathbb{C}_{\ominus\mathbf{x}\oplus\mathbf{u}\ominus\mathbf{x}\oplus\mathbf{v}\ominus\mathbf{x}\oplus\mathbf{w}} \end{aligned} \tag{6.336}$$

thus uncovering the interesting identity

$$\begin{aligned} &\ominus\mathbf{x}\oplus\frac{\gamma_{\mathbf{u}}\mathbf{u} + \gamma_{\mathbf{v}}\mathbf{v} + \gamma_{\mathbf{w}}\mathbf{w}}{\gamma_{\mathbf{u}} + \gamma_{\mathbf{v}} + \gamma_{\mathbf{w}}} \\ &\quad = \frac{\gamma_{\ominus\mathbf{x}\oplus\mathbf{u}}\ominus\mathbf{x}\oplus\mathbf{u} + \gamma_{\ominus\mathbf{x}\oplus\mathbf{v}}\ominus\mathbf{x}\oplus\mathbf{v} + \gamma_{\ominus\mathbf{x}\oplus\mathbf{w}}\ominus\mathbf{x}\oplus\mathbf{w}}{\gamma_{\ominus\mathbf{x}\oplus\mathbf{u}} + \gamma_{\ominus\mathbf{x}\oplus\mathbf{v}} + \gamma_{\ominus\mathbf{x}\oplus\mathbf{w}}} \end{aligned} \tag{6.337}$$

in any Einstein gyrovector space.

Euclidean triangle centroids have classical mechanical interpretation [Hausner (1998)]. Interestingly, we will see in Chaps. 10-11 on Einstein's special theory of relativity that gyrocentroids in Einstein two and three-dimensional gyrovector spaces have analogous relativistic mechanical interpretation in which the gamma factor plays the role of the relativistic mass correction.

Further extension to n-dimensional gyrotetrahedrons in n-dimensional Einstein gyrovector spaces, $n \geq 3$, shown in Figs. 6.18 and 6.19 for $n = 3$, is obvious. Thus, for instance, the gyrocentroid $\mathbb{C}_{\mathbf{uvwx}}$ of the gyrotetrahedron **uvwx** in Fig. 6.19 is given by the equation

$$\mathbb{C}_{\mathbf{uvwx}} = \frac{\gamma_{\mathbf{u}}\mathbf{u} + \gamma_{\mathbf{v}}\mathbf{v} + \gamma_{\mathbf{w}}\mathbf{w} + \gamma_{\mathbf{x}}\mathbf{x}}{\gamma_{\mathbf{u}} + \gamma_{\mathbf{v}} + \gamma_{\mathbf{w}} + \gamma_{\mathbf{x}}} \tag{6.338}$$

6.22.2 *In Möbius Gyrovector Spaces*

Let

$$\begin{aligned} G_e &= (\mathbb{V}_s, \oplus_{\mathrm{E}}, \otimes_{\mathrm{E}}) \\ G_m &= (\mathbb{V}_s, \oplus_{\mathrm{M}}, \otimes_{\mathrm{M}}) \end{aligned} \tag{6.339}$$

be, respectively, the Einstein and the Möbius gyrovector spaces of the ball $\mathbb{V}_s$ of a real inner product space $\mathbb{V}$. They are gyrovector space isomorphic, with the isomorphism ϕ_{EM} and its inverse isomorphism ϕ_{ME} from G_m into G_e shown in Table 6.1, p. 226. Accordingly, the correspondence between elements of G_m and G_e is given by the equations

$$\begin{aligned} \mathbf{v}_e &= 2\otimes_{\mathrm{M}}\mathbf{v}_m \\ \mathbf{v}_m &= \tfrac{1}{2}\otimes_{\mathrm{E}}\mathbf{v}_e \end{aligned} \tag{6.340}$$

$\mathbf{v}_e \in G_e$, $\mathbf{v}_m \in G_m$.

Hence, by (6.340) and the first identity in (6.304) we have

$$\gamma_{\mathbf{v}_e} = \gamma_{2\otimes_{\mathrm{M}}\mathbf{v}_m} = 2\gamma_{\mathbf{v}_m}^2 - 1 \tag{6.341}$$

Similarly, following (6.340) and (6.301) we have

$$\gamma_{\mathbf{v}_e}\mathbf{v}_e = \gamma_{2\otimes_{\mathrm{M}}\mathbf{v}_m} 2\otimes_{\mathrm{M}}\mathbf{v}_m = 2\gamma_{\mathbf{v}_m}^2\mathbf{v}_m \tag{6.342}$$

05.10.2019

Pág. 243(254).

g) The Einstein Gyromidpoint theorem 6.92

Source: 20191003_03/Abraham_Ungar_Books/Abraham A. Ungar - Analytic Hyperbolic Geometry and Albert Einstein's Special Theory of Relativity.pdf

Pág. 230(251) Theorem 6.92

Fig. 6.17

A left gyrotranslation by $x \in \{R^2, \oplus, \otimes\}$ of the gyrotriangle uvw of Fig. 6.16 in an Einstein plane is shown.

Pág. 232(253) - Fig. 6.18

The hyperbolic tetrahedron uvwx, that is, a gyrotetrahedron, is shown in the Einstein gyrovector space

$R^3_s = \{R^3_s, \oplus_E, \otimes_E\}$

05.10.2019

underlying the Beltrami ball model of hyperbolic geometry.

The gyrotetrahedron uvwx is shown inside the 1-ball R^3_1 of the Euclidean 3-space R^3 where it lives.

The faces of the gyrotetrahedron are gyrotriangle.

h) A.2 Surfaces in 3-D Euclidean space

A.2.1 – Representations of surfaces; metrics

$$x = x^i(u,v)\, a_i,$$

where a_i is a fixed basis, i, j, k. Equivalently, we have

$$\left.\begin{aligned} x &= x(u,v) \\ y &= x(u,v), \\ z &= z(u,v). \end{aligned}\right| \text{ for a given domain of the parameters.}$$

These actually are coordinates on the surface. We may also give it as $f(x,y,z) = 0$. —

data 05 10 2019

S T Q Q S [✓ S] D

Pág. 248 (267)

Source: 20191005/Important/
Differential geometry for physicists and mathematicians - Moving frames ...pdf.

i) 1.1.1. Three-dimensional vectors.

Source: 20191005/Important/
Ibragimov - Tensors and Riemannian Geometry (2015, De Gruyter).pdf

Pág. 03 (12)

Vector Algebra

$a \in \mathbb{R}^3$ be a three-dimensional (in particular, two-dimensional) vector.

Graphically, a is a directed line segment. Its magnitude is denoted by $|a|$.

The scalar product $a.b$ of vector a and b is a scalar quantity (i.e., a real number) defined by

$a.b = |a||b| \cos\theta$, where $\theta \, (0 \leq \theta \leq \pi)$.

7.3 The Gyroline Element of Möbius Gyrovector Spaces

In this section we uncover the Riemannian line element to which the gyrometric of the Möbius gyrovector space $(\mathbb{R}^n_c, \oplus_{\rm M}, \otimes_{\rm M})$ gives rise.

Let us consider the gyrodifferential (7.12),

$$\begin{aligned}\Delta\mathbf{s}_{\rm M} &= (\mathbf{v} + \Delta\mathbf{v})\ominus_{\rm M}\mathbf{v}\\ &= \begin{pmatrix} x_1 + \Delta x_1 \\ x_2 + \Delta x_2 \end{pmatrix} \ominus_{\rm M} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\end{aligned} \tag{7.22}$$

in the Möbius gyrovector plane $(\mathbb{R}^2_c, \oplus_{\rm M}, \otimes_{\rm M})$, where ambiguously, $+$ is the Euclidean addition in $\mathbb{R}^2$ and in $\mathbb{R}$. To calculate $\mathbf{X}_1$ and $\mathbf{X}_2$ we have

$$\begin{aligned} d\mathbf{s}_{\rm M} &= \left[\frac{\partial\Delta\mathbf{s}_{\rm M}}{\partial\Delta x_1}\right]_{\left\{\begin{array}{l}\Delta x_1 = 0\\ \Delta x_2 = 0\end{array}\right\}} dx_1 \quad + \quad \left[\frac{\partial\Delta\mathbf{s}_{\rm M}}{\partial\Delta x_2}\right]_{\left\{\begin{array}{l}\Delta x_1 = 0\\ \Delta x_2 = 0\end{array}\right\}} dx_2 \\ &= \mathbf{X}_1(x_1, x_2)dx_1 + \mathbf{X}_2(x_1, x_2)dx_2 \end{aligned} \tag{7.23}$$

where $\mathbf{X}_1, \mathbf{X}_2 : \mathbb{R}^2_c \to \mathbb{R}^2$, obtaining

$$\begin{aligned}\mathbf{X}_1(x_1, x_2) &= \frac{c^2}{c^2 - r^2}(1,\, 0) \in \mathbb{R}^2\\ \mathbf{X}_2(x_1, x_2) &= \frac{c^2}{c^2 - r^2}(0,\, 1) \in \mathbb{R}^2\end{aligned} \tag{7.24}$$

where $r^2 = x_1^2 + x_2^2$.

The gyrometric coefficients of the gyrometric of the Möbius gyrovector plane in the Cartesian x_1x_2-coordinates are therefore

$$\begin{aligned} E &= \mathbf{X}_1\cdot\mathbf{X}_1 = \frac{c^4}{(c^2 - r^2)^2}\\ F &= \mathbf{X}_1\cdot\mathbf{X}_2 = 0\\ G &= \mathbf{X}_2\cdot\mathbf{X}_2 = \frac{c^4}{(c^2 - r^2)^2}\end{aligned} \tag{7.25}$$

Hence, the gyroline element of the Möbius gyrovector plane $(\mathbb{R}^2_c, \oplus_{\rm M}, \otimes_{\rm M})$

is the Riemannian line element

$$\begin{aligned} ds_{\mathrm{M}}^2 &= \|d\mathbf{s}_{\mathrm{M}}\|^2 \\ &= Edx_1^2 + 2Fdx_1dx_2 + Gdx_2^2 \\ &= \frac{c^4}{(c^2-r^2)^2}(dx_1^2+dx_2^2) \end{aligned} \tag{7.26}$$

known as the Poincaré metric. Thus, for instance, the Riemannian line element of the Poincaré disc or, equivalently, the Möbius gyrovector plane $(\mathbb{R}^2_{c=2}, \oplus_{\mathrm{M}}, \otimes_{\mathrm{M}})$, is [McCleary (1994), p. 226]

$$ds_{\mathrm{M}}^2 = \frac{dx_1^2 + dx_2^2}{(1-\frac{r^2}{c^2})^2} \tag{7.27}$$

An interesting elementary study of the Riemannian structure (7.26) in the context of the hyperbolic plane is presented in the introductory chapter of [Helgason (1984)]. The Riemannian line element ds_{M}^2 is identified in [Farkas and Kra (1992), p. 216] as a Riemannian metric on the Riemann surface $\mathbb{D}_{c=1}$, where $\mathbb{D}_{c=1}$ is the Poincaré complex open unit disc.

Following Riemann [Stahl (1993), p. 73] we note that E, G and $EG - F^2 = EG$ are all positive in the open disc $\mathbb{R}^2_c$, so that the quadratic form (7.26) is *positive definite* [Kreyszig (1991), p. 84].

The gyrocurvature of the Möbius gyrovector plane is the Gaussian curvature K of the surface with the line element (7.26). It is a negative constant,

$$K = -\frac{4}{c^2} \tag{7.28}$$

as one can calculate from (7.21).

Extending (7.26) from $n = 2$ to $n \geq 2$ we have

$$\begin{aligned} ds_{\mathrm{M}}^2 &= \frac{c^4}{(c^2-\mathbf{r}^2)^2}d\mathbf{r}^2 \\ &= \frac{d\mathbf{r}^2}{(1+\frac{1}{4}K\mathbf{r}^2)^2} \end{aligned} \tag{7.29}$$

The Riemannian line element ds_{M}^2 reduces to its Euclidean counterpart in the limit of large c,

$$\lim_{c\to\infty} ds_{\mathrm{M}}^2 = d\mathbf{r}^2 \tag{7.30}$$

as expected.

The study of the line element (7.29) by Riemann is described in [Coxeter (1998), p. 12]. In two dimensions it is known as the Poincaré metric of the Poincaré disc [McCleary (1994)]. We have thus established the following theorem.

Theorem 7.4 *The gyroline element of a Möbius gyrovector space* $(\mathbb{R}^n_c, \oplus_M, \otimes_M)$ *is given by the equation*

$$ds^2_{\rm M} = \frac{c^4}{(c^2 - \mathbf{r}^2)^2} d\mathbf{r}^2 \tag{7.31}$$

and its gyrocurvature is given by the equation

$$K = -\frac{4}{c^2} \tag{7.32}$$

In particular, for $n = 2$ *and* $c = 1$ *the gyroline element* (7.31) *coincides with the Riemannian line element of the Poincaré disc model of hyperbolic geometry.*

A Riemannian metric g in $\mathbb{R}^n$ has the form

$$g = \sum_{i,j=1}^{n} g_{ij} dx_i dx_j \tag{7.33}$$

Two Riemannian metrics g_1 and g_2 are said to be *conformal* to each other if there is a positive smooth (that is, infinitely differentiable) function $f : \mathbb{R}^n \to \mathbb{R}^{>0}$ such that $g_1 = f g_2$ [Lee (1997)], $\mathbb{R}^{>0} = \{r \in \mathbb{R} : r > 0\}$ being the positive ray of the real line $\mathbb{R}$.

The Riemannian metric of Möbius gyrovector spaces $(\mathbb{R}^n_c, \oplus_{\rm M}, \otimes_{\rm M})$, given by (7.29), is conformal to the Riemannian metric

$$d\mathbf{r}^2 = \sum_{i=1}^{n} dx_i^2 \tag{7.34}$$

of the Euclidean n-space $\mathbb{R}^n$. Hence, hyperbolic spheres in Möbius gyrovector spaces are also Euclidean spheres. However, Euclidean and hyperbolic sphere centers need not coincide, as shown in Fig. 7.1. The reason in terms of analogies that Euclidean and hyperbolic geometry share is clear. For any two tangent Euclidean circles with centers $\mathbf{a}, \mathbf{c}$ and concurrent point $\mathbf{b}$, the points $\mathbf{a}, \mathbf{b}, \mathbf{c}$ are collinear. In full analogy, for any two tangent Möbius circles with centers $\mathbf{a}, \mathbf{c}$ and concurrent point $\mathbf{b}$, the points $\mathbf{a}, \mathbf{b}, \mathbf{c}$ are gyrocollinear, as shown in Fig. 7.1.

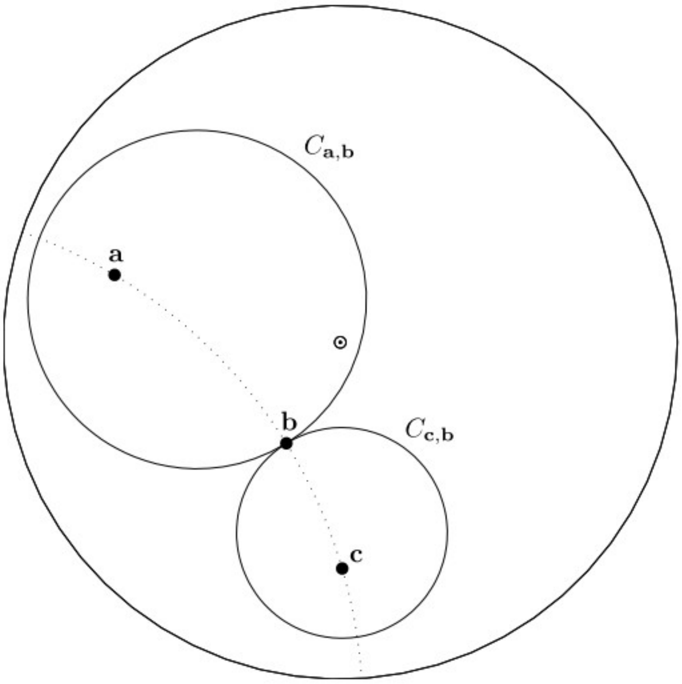

Fig. 7.1 The Möbius gyrocircle $C_{\mathbf{a},\mathbf{b}}$, given by the equation

$$\mathbf{a}\oplus_{\text{M}} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} (\ominus_{\text{M}}\mathbf{a}\oplus_{\text{M}}\mathbf{b}), \qquad 0 \le \theta < 2\pi,$$

is the hyperbolic circle in the Möbius gyrovector plane $(\mathbb{R}^2_c, \oplus_{\text{M}}, \otimes_{\text{M}})$. It is a Euclidean circle with hyperbolic radius $\|\ominus_{\text{M}}\mathbf{a}\oplus_{\text{M}}\mathbf{b}\|$, hyperbolically centered at **a**.
A second gyrocircle, $C_{\mathbf{c},\mathbf{b}}$, that intersects $C_{\mathbf{a},\mathbf{b}}$ at a single point, **b**, is shown. In full analogy with Euclidean geometry, the two gyrocenters **a** and **c** and the concurrent point **b** are gyrocollinear.

7.4 The Cogyroline Element of Möbius Gyrovector Spaces

In this section we uncover the Riemannian line element to which the cogyrometric of the Möbius gyrovector space $(\mathbb{R}^n_c, \oplus_{\text{M}}, \otimes_{\text{M}})$ gives rise.

Let us consider the cogyrodifferential (7.12),

$$\begin{aligned} \Delta\mathbf{s}_{\text{CM}} &= (\mathbf{v} + \Delta\mathbf{v}) \boxminus_{\text{M}} \mathbf{v} \\ &= \begin{pmatrix} x_1 + \Delta x_1 \\ x_2 + \Delta x_2 \end{pmatrix} \boxminus_{\text{M}} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \end{aligned} \tag{7.35}$$

in the Möbius gyrovector plane $(\mathbb{R}^2_c, \oplus_{\text{M}}, \otimes_{\text{M}})$, where $+$ is the Euclidean

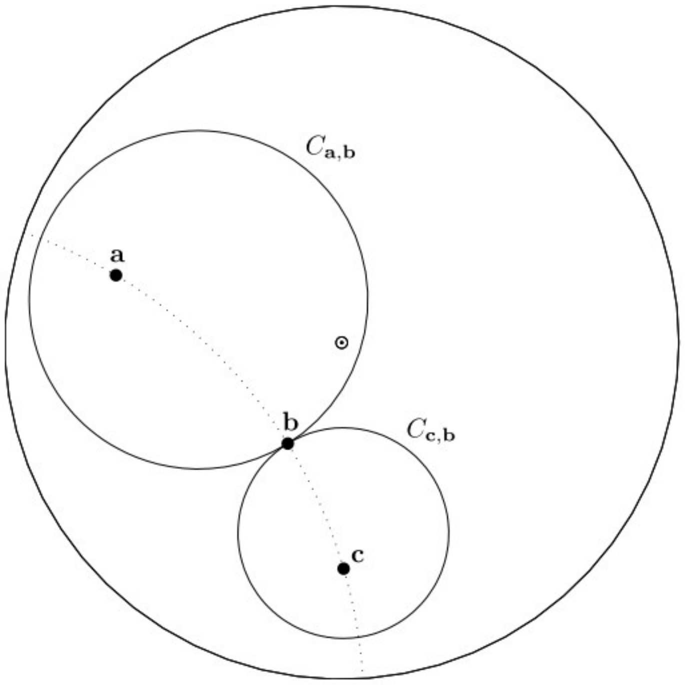
$C_{\mathbf{a},\mathbf{b}}$
$\mathbf{a}$
$\mathbf{b}$
$C_{\mathbf{c},\mathbf{b}}$
$\mathbf{c}$

Fig. 7.1 The Möbius gyrocircle $C_{\mathbf{a},\mathbf{b}}$, given by the equation

$$\mathbf{a}\oplus_{\mathrm{M}}\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}(\ominus_{\mathrm{M}}\mathbf{a}\oplus_{\mathrm{M}}\mathbf{b}), \qquad 0\le\theta<2\pi,$$

+,- * *. * $ 9 D / D
" ? B " " ? 3 " / , $E
T " /"

($) @'

+ ! 9 3 " /
+ , / V

that trivial. In general, the nature of the issues considered involves frame fields adapted to the surface, not constant frame fields. As for the standard concept of (Levi-Civita) connection on a surface, it is here a matter of neglecting the normal component in the differentiation of tangent vectors to the surface.

As has been the case through the text, we shall continue to use the moving frame method. One has not yet taken full advantage of it. We do better than appears to be the case in the literature by the use of surface-adapted 3-D frame fields that are canonically determined by the diagonalization of the second fundamental form of surfaces.

We shall see that the Frenet frame field is canonically determined by curves and that the "geodesic frame field" is canonically determined by the pair of curve and surface. There is the coordinate frame field on surfaces, canonically determined by the pair of surface and coordinate system on it. There are also the canonical frame fields of 3-D Euclidean space, i.e. the constant orthonormal frame fields denoted as $(\mathbf{i}, \mathbf{j}, \mathbf{k})$. Missing is the canonical frame field of an embedded surface.

For comparison purposes, there will be some paragraphs which are not proper part of the theory of curves and surfaces with the theory of moving frames. They illustrate the cumbersomeness of the by now almost abandoned traditional approach. If we removed those comparisons, our treatment would be more than two pages shorter.

A.2 Surfaces in 3-D Euclidean space

A.2.1 Representations of surfaces; metrics

One way of giving a surface is in the form

$$\mathbf{x} = x^i(u, v)\mathbf{a}_i, \tag{2.1a}$$

where $\mathbf{a}_i$ is a fixed basis, $\mathbf{i}, \mathbf{j}, \mathbf{k}$. Equivalently, we have

$$x = x(u, v), \quad y = x(u, v), \quad z = z(u, v), \tag{2.1b}$$

for a given domain of the parameters. These actually are coordinates on the surface. We may also give it as

$$f(x, y, z) = 0. \tag{2.2}$$

The connection between (2.1) and (2.2) can be established as follows. We solve for u and v in two of the equations (2.1b) and substitute in the third of those equations, then taking all terms to the left hand side. On the other hand, we go back from (2.2) to (2.1) by writing, for instance,

$$x = u, \qquad y = v. \tag{2.3}$$

Substitution of (2.3) in (2.2) allows us to view z as an implicit function of x and y. Of course, one may then use other parametrizations through an invertible system of equations

$$u' = u'(u, v), \qquad v' = v'(u, v). \tag{2.4}$$

The metric of 3-D Euclidean space in terms of Cartesian systems is

$$ds^2 = \sum_i (dx^i)^2, \qquad i = 1, 2, 3. \tag{2.5}$$

where

$$dx^i = x^i{}_{,\alpha}\, du^\alpha, \qquad \alpha = 1, 2 \tag{2.6}$$

and where $x^i = (x, y, z)$ and $u^1 = u,\ u^2 = v$. A subscript after a comma means partial differentiation. The variable with respect to which the differentiation takes place is apparent in any such expressions. Once these operations have been performed, any trace of x, y, z, dx, dy and dz has disappeared and we are left with

$$ds^2 = g_{\alpha\beta}(u^\gamma) du^\alpha du^\beta, \tag{2.7}$$

which is also written as

$$ds^2 = Edu^2 + 2Fdudv + Gdv^2. \tag{2.8}$$

It is the called the first fundamental form. Clearly

$$E = g_{11}, \quad F = g_{12}, \quad G = g_{22}. \tag{2.9}$$

The positive definiteness of the metric implies

$$EG - F^2 > 0, \tag{2.10}$$

as we are about to show.

(2.8) is the way in which one expresses distance on curves, which is actually computed as

$$\sqrt{E\left(\frac{du}{d\lambda}\right)^2 + 2F\left(\frac{du}{d\lambda}\right)\left(\frac{dv}{d\lambda}\right) + G\left(\frac{dv}{d\lambda}\right)^2} d\lambda, \tag{2.11}$$

where λ is a parameter on the curve. In general, we can take the parameter to be v itself. (2.11) then becomes

$$\int \sqrt{E\mu^2 + 2F\mu + G}\ dv, \tag{2.12}$$

where μ is du/dv. The equation

$$y = Ex^2 + 2Fx + G \tag{2.13}$$

is a parabola. If the metric is positive definite, this is an upright parabola with vertex above the x axis. It does not cut the line $y = 0$. Thus the roots of $Ex^2 + 2Fx + G = 0$ must be imaginary, which implies

$$EG - F^2 > 0. \tag{2.14}$$

A.2.2 Normal to a surface, orthonormal frames, area

Vectors normal to a surface are not tangent vectors. Hence, they cannot be given as a linear combination of tangent vectors, but rather in terms of, for example, the basis $\mathbf{a}_i = \mathbf{i}, \mathbf{j}, \mathbf{k}$. Given the surface $f(x, y, z) = 0$, its tangent plane at (x_0, y_0, z_0) is obtained by substituting $x^i - x_0^i$ for dx^i in $f_{,i}\, dx^i = 0$, i.e.

$$f_{,i} \cdot (x^i - x_0^i) = 0, \tag{2.15}$$

where the dot is used for product, thus preempting taking $x^i - x_0^i$ as argument of $f_{,i}$. Let $g(x, y, z)$ be defined as $-f(x, y, z)$. The equation $g(x, y, z) = 0$ represents the same surface as $f(x, y, z) = 0$. But the unit vectors

$$\frac{\sum f_{,i}\, \mathbf{a}_i}{\sqrt{\sum (f_{,i})^2}}, \qquad \frac{\sum g_{,i}\, \mathbf{a}_i}{\sqrt{\sum (g_{,i})^2}}$$

are opposite of each other. Whether we take one or the other as the unit normal to the plane is not significant at this point.

Differentiating (2.1), we get

$$d\mathbf{x} = dx^i(u, v)\mathbf{a}_i = \mathbf{x}_{,u} du + \mathbf{x}_{,v} dv. \tag{2.16}$$

$\mathbf{x}_{,u}$ and $\mathbf{x}_{,v}$ are defined by these relations. They constitute a tangent coordinate vector basis field. It is not, therefore, orthonormal in general. The equation of the plane at the point of coordinates (u_0, v_0) is given by

$$\mathbf{x} = \mathbf{x}_0 \ + \ \mathbf{x}_{,u} \cdot (u - u_0) \ + \ \mathbf{x}_{,v} \cdot (v - v_0), \tag{2.17}$$

where $\mathbf{x}_0$, $\mathbf{x}_{,u}$ and $\mathbf{x}_{,v}$ are evaluated at (u_0, v_0). The dots at mid height have been put for the same reason as in (2.15). In moving frame notation, they are the equivalent of

$$d\mathbf{P} = \omega^i \mathbf{e}_i, \qquad \mathbf{P} = \mathbf{P}_0 \ + \ X^i \mathbf{e}_i. \tag{2.18}$$

The following is a necessary step to launch the moving frame method once the equation of a surface has been given. Let $(\hat{\omega}^1, \hat{\omega}^2)$ be any pair of differential 1-forms that satisfies

$$ds^2 = g_{\alpha\beta}(u, v) du^\alpha du^\beta = (\hat{\omega}^1)^2 + (\hat{\omega}^2)^2. \tag{2.19}$$

A first (retrospectively very important) consequence of the normalized form of the metric is that $\hat{\omega}^3$ is zero on the surface. Hence, the pull-back to the surface of the torsion of Euclidean space (torsion which, of course, is zero) reads

$$0 = d\hat{\omega}^3 = \hat{\omega}^\alpha \wedge \hat{\omega}^3_\alpha = \Gamma^3_{\alpha\beta} \hat{\omega}^\alpha \wedge \hat{\omega}^\beta, \tag{2.20}$$

where, for the last step, we have used that the pull-back of $\hat{\omega}^3_\alpha$ is a linear function of only $\hat{\omega}^1$ and $\hat{\omega}^2$. We immediately get $\Gamma^3_{12} = \Gamma^3_{21}$, which we prefer to write as

$$\Gamma^1_{32} = \Gamma^2_{31} \tag{2.21}$$

(on the surface).

We now proceed to diagonalize the expression

$$g_{\alpha\beta}du^{\alpha}du^{\beta} \equiv g_{11}du^2 + 2g_{12}dudv + g_{22}dv^2, \tag{2.22}$$

by first obtaining

$$g_{\alpha\beta}du^{\alpha}du^{\beta} - \frac{1}{g_{11}}(g_{11}du + g_{12}dv)^2 = dv^2\frac{g_{22}g_{11} - g_{12}^2}{g_{11}}, \tag{2.23}$$

as per Eq. (4.51) of chapter 3. We thus have

$$g_{\alpha\beta}du^{\alpha}du^{\beta} = \left(\frac{g_{11}du + g_{12}dv}{\sqrt{g_{11}}}\right)^2 + dv^2\frac{g_{22}g_{11} - g_{12}^2}{g_{11}}, \tag{2.24}$$

and

$$\hat{\omega}^1 = \frac{g_{11}du + g_{12}dv}{\sqrt{g_{11}}}, \qquad \hat{\omega}^2 = dv\frac{(g_{22}g_{11} - g_{12}^2)^{1/2}}{\sqrt{g_{11}}} \tag{2.25}$$

and, further,

$$\hat{\omega}^1 \wedge \hat{\omega}^2 = (g_{22}g_{11} - g_{12}^2)^{1/2}du \wedge dv = \sqrt{EG - F^2}\, du \wedge dv. \tag{2.26}$$

Clearly, there is an infinite number of pairs $(\hat{\omega}^1, \hat{\omega}^2)$ that orthogonalize the metric of the surface. (2.25) is just one of them. We shall use this freedom to determine *frame fields adapted to surfaces*.

A.2.3 The equations of Gauss and Weingarten

The connection equations of Euclidean space are simply $d\mathbf{a}_i = 0$ in terms of a Cartesian frame field. The equations of Gauss and Weingarten are the connection equations of 3-D Euclidean space when the frame field is constituted by $\mathbf{x}_{,u}$, $\mathbf{x}_{,v}$ and the unit vector $\mathbf{N}$ perpendicular to the plane that they determine.

At a time when the concept of connection equations did not exist, Gauss wrote the differentiation of the tangent frame field $(\mathbf{x}_{,u}, \mathbf{x}_{,v})$ as

$$\mathbf{x}_{,u,u} = \Gamma^1_{11}\mathbf{x}_{,u} + \Gamma^2_{11}\mathbf{x}_{,v} + e\mathbf{N}, \tag{2.27a}$$

$$\mathbf{x}_{,u,v} = \Gamma^1_{12}\mathbf{x}_{,u} + \Gamma^2_{12}\mathbf{x}_{,v} + f\mathbf{N}, \tag{2.27b}$$

$$\mathbf{x}_{,v,v} = \Gamma^1_{22}\mathbf{x}_{,u} + \Gamma^2_{22}\mathbf{x}_{,v} + g\mathbf{N}, \tag{2.27c}$$

which reflects the state of the art of the time. $d(\mathbf{x}_{,u})$ is a differential 2-form on the surface, valued in 3-D Euclidean vector space. It can be written as the following bilinear combination

$$d(\mathbf{x}_{,u}) = \Gamma^1_{1\alpha}du^{\alpha}\mathbf{x}_{,u} + \Gamma^2_{1\alpha}du^{\alpha}\mathbf{x}_{,v} + \Gamma^3_{1\alpha}du^{\alpha}\mathbf{N} \tag{2.28}$$

(and similarly for $d\mathbf{x}_{,v}$). Here we can read that

$$\mathbf{x}_{,u,u} = \Gamma^1_{11}\mathbf{x}_{,u} + \Gamma^2_{11}\mathbf{x}_{,v} + \Gamma^3_{11}\mathbf{N}, \tag{2.29}$$

(and similarly for $\mathbf{x}_{,u,v}$ and $\mathbf{x}_{,v,v}$).

The two Weingarten equations are

$$\mathbf{N}_{,u} = \frac{fF - eG}{EG - F^2}\mathbf{x}_{,u} + \frac{eF - fE}{EG - F^2}\mathbf{x}_{,v}, \tag{2.30a}$$

$$\mathbf{N}_{,v} = \frac{gF - fG}{EG - F^2}\mathbf{x}_{,u} + \frac{fF - gE}{EG - F^2}\mathbf{x}_{,v}. \tag{2.30b}$$

In terms of orthonormal frames, we would replace

(a) $\mathbf{x}_{,u}$ and $\mathbf{x}_{,v}$ with $\hat{\mathbf{e}}_\alpha$,

(b) (e, f, g) with Γ's as stated above. We shall later see how they were introduced originally.

(c) (E, G, F) with $(1, 1, 0)$.

Notice the great simplification that then results. Further simplification follows in section 4 when we adopt the canonical frame field of surfaces.

A.3 Curves in 3-D Euclidean space

A.3.1 Frenet's frame field and formulas

Curves in E^3 are 1-dimensional manifolds, if we eliminate the points where the curve intersects itself. They can be given as the intersection of two surfaces

$$f(x, y, z) = 0, \qquad g(x, y, z) = 0, \tag{3.1}$$

or as

$$x = x(u), \quad y = y(u), \quad z = z(u), \tag{3.2}$$

which is the developed form of

$$\mathbf{x} = x^i(u)\mathbf{a}_i. \tag{3.3}$$

The relation between (3.1) and (3.2) is obtained by solving for two of the Cartesian coordinates in (3.1), and then changing parameter if one so wishes. In the opposite direction, we eliminate u in (3.2)

The distance on the curve between two points on the same is obtained by integration of

$$ds = \sqrt{\left(\frac{dx}{du}\right)^2 + \left(\frac{dy}{du}\right)^2 + \left(\frac{dz}{du}\right)^2}\, du. \tag{3.4}$$

Frenet's frame is the 3-D frame field adapted to a curve that, if we require it to be orthonormal, is canonically determined by the curve.

The first element of the frame is the unit tangent vector $\mathbf{t}$,

$$\mathbf{t} \equiv \frac{d\mathbf{x}}{ds} = \frac{d\mathbf{x}}{du}\frac{1}{ds/du}, \tag{3.5}$$

with ds/du obtained from (3.4). $\mathbf{t}$ is obviously a unit vector

$$\mathbf{t} \cdot \mathbf{t} = 1, \tag{3.6}$$

since $d\mathbf{x} \cdot d\mathbf{x}$ is ds^2. The tangent line to a curve at its point $\mathbf{x}_0$ is

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{t}s, \tag{3.7}$$

s being assigned the value zero at $\mathbf{x}_0$.

From (3.6), we define a curvature vector $\boldsymbol{\kappa} \equiv \frac{d\mathbf{t}}{ds}$. We then have:

$$\mathbf{t} \cdot \boldsymbol{\kappa} = \frac{\mathbf{dt}}{\mathbf{ds}} \cdot \boldsymbol{\kappa} = \mathbf{0} \tag{3.8}$$

which means that the equations

$$\boldsymbol{\kappa} = \kappa \mathbf{n} = \frac{1}{\rho} \mathbf{n} \tag{3.9}$$

define $\mathbf{n}$, κ and ρ if we demand that $\mathbf{n}$ be a unit vector. This vector is the second element of a right handed orthonormal frame. One refers to κ and ρ respectively as curvature and radius of curvature. Because of the sought orthonormality of the frame, the coefficient of $\mathbf{n}'$ relative to $\mathbf{b}$ will be the negative of the coefficient of $\mathbf{b}'$ relative to $\mathbf{n}$. Thus:

$$\mathbf{n}' = -\kappa \mathbf{t} + \tau \mathbf{b}, \qquad \mathbf{b}' = -\tau \mathbf{n}. \tag{3.10}$$

τ is called torsion of the curve.

Equations (3.9) and (3.10) are called the Frenet equations. $\mathbf{b}$ is perpendicular to the plane of plane curves. It is then of constant direction and actually a constant vector, since it is of unit size. τ is, therefore, zero. Plane curves are torsionless. On the other hand, a helix has torsion since the plane of $\mathbf{t}$ and $\mathbf{n}$ keeps changing. This plane is called the osculating plane.

A.3.2 Geodesic frame fields and formulas

We are going to study surfaces through curves on them. It is then useful to consider 3-D right handed orthonormal frames ($\hat{\mathbf{u}}_i = \mathbf{t}, \mathbf{u}, \mathbf{N}$). $\mathbf{u}$ is a vector in the tangent plane and $\mathbf{N}$ is usually taken on the opposite side (of the tangent plane) to $\boldsymbol{\kappa}$ (see example in next paragraph). We shall refer to this frame field as the geodesic frame field of the curve on the given surface.

One defines the *normal and geodesic curvatures* $\boldsymbol{\kappa}_n$ and $\boldsymbol{\kappa}_g$ of the curve at a point by means of the decomposition

$$\mathbf{t}' \equiv \frac{d\mathbf{t}}{ds} = \boldsymbol{\kappa} \equiv \boldsymbol{\kappa}_n + \boldsymbol{\kappa}_g = -\kappa_n \mathbf{N} + \kappa_g \mathbf{u}. \tag{3.11}$$

We follow Struik and use the subscript "$_n$" even though the projection is on $\mathbf{N}$, not on $\mathbf{n}$ (see (3.9)). We have introduced the minus sign in front of $\kappa_n \mathbf{N}$ so that

κ_n be positive under the most usual convention when choosing the direction of the normal to the sphere. Thus the normal $\mathbf{n}$ to a parallel is inwards, whereas the normal at the same point to the sphere itself is usually taken to be outwards. This example is also useful in connection with how the radii of curvature of the parallel and of the sphere are related. Recall that the magnitudes of the curvatures and their associated radii are the inverses of each other.

Orthonormality determines the $\mathbf{t}$ component of $\mathbf{N}'$. We thus have,

$$\mathbf{N}' \equiv \frac{d\mathbf{N}}{ds} = \kappa_n \mathbf{t} + \tau_g \mathbf{u}, \tag{3.12}$$

which defines the coefficient τ_g, called geodesic torsion of the curve. Using again orthonormality, we obtain

$$\mathbf{u}' \equiv \frac{d\mathbf{u}}{ds} = -\kappa_g \mathbf{t} - \tau_g \mathbf{N}. \tag{3.13}$$

A.4 Curves on surfaces in 3-D Euclidean space

A.4.1 Canonical frame field of a surface

We clearly have

$$\kappa_n = -\boldsymbol{\kappa} \cdot \mathbf{N} = -\frac{d\mathbf{t}}{ds} \cdot \mathbf{N} = \mathbf{t} \cdot \frac{d\mathbf{N}}{ds} = \frac{d\mathbf{x}}{ds} \cdot \frac{d\mathbf{N}}{ds}. \tag{4.1}$$

$d\mathbf{x} \cdot d\mathbf{N}$ is a quadratic symmetric differential form on du and dv:

$$d\mathbf{x} \cdot d\mathbf{N} = e du^2 + 2f du dv + g dv^2. \tag{4.2}$$

It is called the *second fundamental form.* We do not have any use for this form, as there is a much better way to study normal curvature through the introduction of the concept of canonical frame field of a surface.

Let $\hat{\mathbf{u}}_\alpha$ denote any orthonormal, counterclockwise 2-D frame tangent to the surface and let $\hat{\mathbf{u}}_3$ be $\mathbf{N}$. Since only zero forms and 1-forms live on a curve, the last ones are multiples of each other. Formal division by a differential 1-form is justified and we get

$$\kappa_n = \frac{\hat{\omega}'^{\alpha} \hat{\mathbf{u}}_\alpha}{ds} \cdot \frac{\hat{\omega}_3'^{i} \hat{\mathbf{u}}_i}{ds} = \frac{\hat{\Gamma}'^{1}_{31} (\hat{\omega}'^{1})^2 + (\hat{\Gamma}'^{1}_{32} + \hat{\Gamma}'^{2}_{31}) \hat{\omega}'^{1} \hat{\omega}'^{2} + \hat{\Gamma}'^{2}_{32} (\hat{\omega}'^{2})^2}{ds^2}. \tag{4.3}$$

Equation (4.3) applies to a one-dimensional manifold of frames related at each point by rotations around the normal to the surface. We reserve the use of unprimed quantities for the orthonormal frame $(\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3)$ that satisfies

$$\hat{\Gamma}^{1}_{32} = -\hat{\Gamma}^{2}_{31}, \tag{4.4}$$

and, because of (4.4), also

$$\hat{\Gamma}^{1}_{32} = \hat{\Gamma}^{2}_{31} = 0. \tag{4.5}$$

+,- * *. * $ %A B 4 " ,
; D E 9 ; D /"
(!

data 05 10 2019

Pág. 248 (267)

Source: 20191005/Important/Differential geometry for physicists and mathematicians-Moving frames ....pdf.

i) 1.1.1. Three-dimensional vectors.

Source: 20191005/Important/Ibragimov - Tensors and Riemannian Geometry (2015, De Gruyter).pdf

Pág. 03 (12)

Vector Algebra

$a \in \mathbb{R}^3$ be a three-dimensional (in particular, two-dimensional) vector.

Graphically, a is a directed line segment. Its magnitude is denoted by $|a|$.

The scalar product $a.b$ of vectors a and b is a scalar quantity (i.e., a real number) defined by

$a.b = |a||b| \cos\theta$, where θ $(0 \leq \theta \leq \pi)$.

05.10.2019

is the angle between the vector a and b.
The scalar product is denoted in the literature by (a, b).

The vectors $a, b \in \mathbb{R}^3$ can be represented in the rectangular Cartesian coordinates in the form

$a=(a^1, a^2, a^3)$, $b=(b^1, b^2, b^3)$,

which means that

$a = a^1 i + a^2 j + a^3 k.$

$(a^1 i)$

$b = b^1 i + b^2 j + b^3 k.$

where i, j, and k are the unit vectors along the first, second and third coordinate axes in $\mathbb{R}^3$, respectively.

Then the scalar product is given by

$$a \cdot b = \sum_{i=1}^{3} a^i b^i.$$

1(175) j) Brans-Dicke theory

Source: https://en.wikipedia.org/wiki/Brans-Dicke_theory

05 10 2019

In theoretical physics the Brans-Dicke theory of gravitation (Jordan-Brans-Dicke theory) is a theoretical framework to explain gravitation.

It is a competitor of Einstein's theory of general relativity.

It is an example of a scalar-tensor theory, a gravitational theory in which the gravitational interaction is mediated by a scalar field as well as the tensor field of general relativity.

other:

https://pt.wikipedia.org/wiki/
Geometria_hiperbólica/

105) Advanced Trigonometric Relations through Nbic Functions.

Source: 037_novosarquivos_2019/
Trigonometric/
Sarket.cd_8122430236.pdf

other: Trigonometric Functions and Complex Numbers

Chapter 1

Preliminaries

1.1 Vectors in linear spaces

This section contains basic notion and useful formulae from vector analysis in n-dimensional linear spaces $\mathbb{R}^n$. Following the convenient traditional notation, vectors of dimensions two ($n = 2$) and three ($n = 3$) are denoted by letters in bold faced type.

1.1.1 Three-dimensional vectors

1.1.1.1 Vector algebra

Let $\boldsymbol{a} \in \mathbb{R}^3$ be a three-dimensional (in particular, two-dimensional) vector. Graphically, $\boldsymbol{a}$ is a directed line segment. Its magnitude is denoted by $|\boldsymbol{a}|$.

The *scalar product* $\boldsymbol{a} \cdot \boldsymbol{b}$ of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is a scalar quantity (i.e. a real number) defined by

$$\boldsymbol{a} \cdot \boldsymbol{b} = |\boldsymbol{a}||\boldsymbol{b}| \cos \theta, \tag{1.1.1}$$

where θ ($0 \leq \theta \leq \pi$) is the angle between the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. The scalar product is denoted in the literature by $(\boldsymbol{a}, \boldsymbol{b})$.

The vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^3$ can be represented in the rectangular Cartesian coordinates in the form

$$\boldsymbol{a} = (a^1, a^2, a^3), \quad \boldsymbol{b} = (b^1, b^2, b^3), \tag{1.1.2}$$

which means that

$$\begin{aligned} \boldsymbol{a} &= a^1 \boldsymbol{i} + a^2 \boldsymbol{j} + a^3 \boldsymbol{k}, \\ \boldsymbol{b} &= b^1 \boldsymbol{i} + b^2 \boldsymbol{j} + b^3 \boldsymbol{k}, \end{aligned} \tag{1.1.3}$$

where $\boldsymbol{i}$, $\boldsymbol{j}$ and $\boldsymbol{k}$ are the unit vectors along the first, second and third coordinate axes in $\mathbb{R}^3$, respectively. Then the scalar product (1.1.1) is given by

$$\boldsymbol{a} \cdot \boldsymbol{b} = \sum_{i=1}^{3} a^i b^i. \tag{1.1.4}$$

Brought to you by | Stockholms Universitet
Authenticated
Download Date | 10/31/15 1:40 AM

The *vector product* of the vectors (1.1.3) is a vector given by

$$\boldsymbol{a}\times\boldsymbol{b}=\begin{vmatrix}\boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k}\\ a^1 & a^2 & a^3\\ b^1 & b^2 & b^3\end{vmatrix} \tag{1.1.5}$$

or (see also Eq. (8.2.22))

$$\boldsymbol{a}\times\boldsymbol{b}=(a^2b^3-a^3b^2)\boldsymbol{i}+(a^3b^1-a^1b^3)\boldsymbol{j}+(a^1b^2-a^2b^1)\boldsymbol{k}. \tag{1.1.6}$$

The vector product is also denoted by $[\boldsymbol{a},\boldsymbol{b}]$ or $[\boldsymbol{a}\,\boldsymbol{b}]$.

Proposition 1.1.1. The vector product defined by Eq. (1.1.5) has the following properties:
(i) the vector product is anticommutative,

$$\boldsymbol{a}\times\boldsymbol{b}=-\boldsymbol{b}\times\boldsymbol{a};$$

(ii) the magnitude of the vector $\boldsymbol{a}\times\boldsymbol{b}$ is equal to the area of the parallelogram with the sides $\boldsymbol{a}$ and $\boldsymbol{b}$, i.e.

$$|\boldsymbol{a}\times\boldsymbol{b}|=|\boldsymbol{a}||\boldsymbol{b}|\sin\theta,$$

where $\theta(0\le\theta\le\pi)$ is the angle between the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$;
(iii) the vector $\boldsymbol{a}\times\boldsymbol{b}$ is orthogonal to the plane spanned by $\boldsymbol{a}$ and $\boldsymbol{b}$ and is such that the triplet $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{a}\times\boldsymbol{b}$ forms a right-handed system.

Given three vectors $\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}\in\mathbb{R}^3$, one can produce the *mixed product*

$$\boldsymbol{a}\cdot(\boldsymbol{b}\times\boldsymbol{c})$$

called also the *scalar triple product*. Geometrically, the mixed product is equal to the volume of the parallelepiped having $\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}$ as its edges, provided that the triplet $\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}$ forms a right-handed system. For the vectors written in the rectangular Cartesian coordinates in the form (1.1.2), the mixed product is given by

$$\boldsymbol{a}\cdot(\boldsymbol{b}\times\boldsymbol{c})=\begin{vmatrix}a^1 & a^2 & a^3\\ b^1 & b^2 & b^3\\ c^1 & c^2 & c^3\end{vmatrix}. \tag{1.1.7}$$

One can consider also the *vector triple product* $\boldsymbol{a}\times(\boldsymbol{b}\times\boldsymbol{c})$. It can be computed by the equation

$$\boldsymbol{a}\times(\boldsymbol{b}\times\boldsymbol{c})=(\boldsymbol{a}\cdot\boldsymbol{c})\boldsymbol{b}-(\boldsymbol{a}\cdot\boldsymbol{b})\boldsymbol{c}. \tag{1.1.8}$$

Brought to you by | Stockholms Universitet
Authenticated
Download Date | 10/31/15 1:40 AM

1.1.1.2 Differential calculus of vectors

Here we consider vector fields. The calculations are given in the rectangular Cartesian reference frame. In particular, the independent variables are the coordinates x, y, z of the position vector $\boldsymbol{x} = (x, y, z)$.

A *vector field* $\boldsymbol{a} = (a^1, a^2, a^3)$ is a vector function

$$\boldsymbol{a} = \boldsymbol{a}(x, y, z)$$

depending upon the position vector $\boldsymbol{x} = (x, y, z)$.

Hamilton's operator ∇ is a vector given in the rectangular Cartesian coordinates (x, y, z) by

$$\nabla = \boldsymbol{i}\,\frac{\partial}{\partial x} + \boldsymbol{j}\,\frac{\partial}{\partial y} + \boldsymbol{k}\,\frac{\partial}{\partial z}, \tag{1.1.9}$$

where $\boldsymbol{i}$, $\boldsymbol{j}$ and $\boldsymbol{k}$ are unit vectors along the x, y and z axes, respectively. In other words, Hamilton's operator is the vector $\nabla = (\nabla_x, \nabla_y, \nabla_z)$ with the components

$$\nabla_x = \frac{\partial}{\partial x}, \quad \nabla_y = \frac{\partial}{\partial y}, \quad \nabla_z = \frac{\partial}{\partial z}.$$

The familiar operations of gradient, divergence and curl are written via Hamilton's operator as follows.

The *gradient* of a scalar field $\phi = \phi(x, y, z)$ is the vector field

$$\operatorname{grad}\phi = \nabla\phi \equiv \frac{\partial\phi}{\partial x}\,\boldsymbol{i} + \frac{\partial\phi}{\partial y}\,\boldsymbol{j} + \frac{\partial\phi}{\partial z}\,\boldsymbol{k}. \tag{1.1.10}$$

The *divergence* of a vector field $\boldsymbol{a}$ is the scalar product of the vectors ∇ and $\boldsymbol{a} = (a^1, a^2, a^3)$:

$$\operatorname{div}\boldsymbol{a} \overset{\text{def}}{=} \nabla\cdot\boldsymbol{a} = \nabla_x a^1 + \nabla_y a^2 + \nabla_z a^3 \equiv \frac{\partial a^1}{\partial x} + \frac{\partial a^2}{\partial y} + \frac{\partial a^3}{\partial z}. \tag{1.1.11}$$

The *curl* or *rotation* of a vector field $\boldsymbol{a} = \boldsymbol{a}(x, y, z)$ is the vector product of the vectors ∇ and $\boldsymbol{a}$:

$$\operatorname{curl}\boldsymbol{a} \equiv \operatorname{rot}\boldsymbol{a} \overset{\text{def}}{=} \nabla\times\boldsymbol{a} = \begin{vmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ a^1 & a^2 & a^3 \end{vmatrix} \tag{1.1.12}$$

or (see also Eq. (8.2.23))

$$\nabla\times\boldsymbol{a} = \left(\frac{\partial a^3}{\partial y} - \frac{\partial a^2}{\partial z}\right)\boldsymbol{i} + \left(\frac{\partial a^1}{\partial z} - \frac{\partial a^3}{\partial x}\right)\boldsymbol{j} + \left(\frac{\partial a^2}{\partial x} - \frac{\partial a^1}{\partial y}\right)\boldsymbol{k}. \tag{1.1.13}$$

Brought to you by | Stockholms Universitet
Authenticated
Download Date | 10/31/15 1:40 AM

Let $\boldsymbol{a}$, $\boldsymbol{b}$ and ϕ, ψ be vector and scalar fields, respectively, and let α, β be arbitrary constants. The operator ∇ together with formulae of *vector algebra* enables one to relate scalar and vector fields through differentiation and to obtain the following equations:

$$\begin{aligned}
&1.\quad \nabla(\alpha\phi + \beta\psi) = \alpha\nabla\phi + \beta\nabla\psi,\\
&2.\quad \nabla\cdot(\alpha\boldsymbol{a} + \beta\boldsymbol{b}) = \alpha\nabla\cdot\boldsymbol{a} + \beta\nabla\cdot\boldsymbol{b},\\
&3.\quad \nabla\times(\alpha\boldsymbol{a} + \beta\boldsymbol{b}) = \alpha\nabla\times\boldsymbol{a} + \beta\nabla\times\boldsymbol{b},\\
&4.\quad \nabla(\phi\psi) = \psi\nabla\phi + \phi\nabla\psi,\\
&5.\quad \nabla\cdot(\phi\boldsymbol{a}) = (\nabla\phi)\cdot\boldsymbol{a} + \phi(\nabla\cdot\boldsymbol{a}),\\
&6.\quad \nabla\times(\phi\boldsymbol{a}) = (\nabla\phi)\times\boldsymbol{a} + \phi\nabla\times\boldsymbol{a},\\
&7.\quad \nabla\cdot(\boldsymbol{a}\times\boldsymbol{b}) = \boldsymbol{b}\cdot(\nabla\times\boldsymbol{a}) - \boldsymbol{a}\cdot(\nabla\times\boldsymbol{b}),\\
&8.\quad \nabla\cdot(\nabla\phi) \equiv \nabla^2\phi \equiv \Delta\phi = \frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} + \frac{\partial^2\phi}{\partial z^2},\\
&9.\quad \nabla\cdot(\nabla\times\boldsymbol{a}) = 0,\\
&10.\ \nabla\times(\nabla\phi) = 0,\\
&11.\ \nabla\times(\nabla\times\boldsymbol{a}) = \nabla(\nabla\cdot\boldsymbol{a}) - \nabla^2\boldsymbol{a}.
\end{aligned} \tag{1.1.14}$$

Note that the 11th equation of (1.1.14) follows from Eq. (1.1.8).

Equations (1.1.14) are often written in the notation (1.1.10)–(1.1.11). For example, the equations 5, 9, 10 and 11 are written:

$$\begin{aligned}
&5'.\quad \operatorname{div}(\phi\,\boldsymbol{a}) = \phi\operatorname{div}\boldsymbol{a} + \boldsymbol{a}\cdot\operatorname{grad}\phi,\\
&9'.\quad \operatorname{div}\operatorname{rot}\boldsymbol{a} = 0,\\
&10'.\ \operatorname{rot}\operatorname{grad}\phi = 0,\\
&11'.\ \operatorname{rot}\operatorname{rot}\boldsymbol{a} = \operatorname{grad}(\operatorname{div}\boldsymbol{a}) - \Delta\,\boldsymbol{a},
\end{aligned}$$

where

$$\Delta \equiv \nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

is the *Laplacian.*

1.1.1.3 Integral calculus of vectors

Green's theorem: Let V be an arbitrary region in the (x, y) plane with the boundary ∂V. Then

$$\int_{\partial V} Pdx + Qdy = \int_V \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dxdy \tag{1.1.15}$$

Brought to you by | Stockholms Universitet
Authenticated
Download Date | 10/31/15 1:40 AM

for any (differentiable) functions $P(x,y)$ and $Q(x,y)$.

Stokes' theorem: Let V be an (orientable) surface in the space (x,y,z) with the boundary ∂V, and let $P(x,y,z), Q(x,y,z)$ and $R(x,y,z)$ be any (differentiable) functions. Then

$$\begin{aligned}\int_{\partial V} Pdx + Qdy + Rdz = \int_V \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dx\,dy + \left(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}\right) dy\,dz \\ + \left(\frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}\right) dz\,dx.\end{aligned} \tag{1.1.16}$$

Equation (1.1.16) is written in the vector notation as follows:

$$\int_{\partial V} \boldsymbol{A} \cdot d\boldsymbol{x} = \int_V \operatorname{curl} \boldsymbol{A} \cdot \boldsymbol{dS}.$$

Here $\boldsymbol{A} = (P,Q,R)$, $d\boldsymbol{x} = (dx,dy,dz)$ and $\boldsymbol{dS} = \boldsymbol{\nu} dS$, where $\boldsymbol{\nu}$ is the unit outward normal to the surface V.

The divergence theorem (the Gauss-Ostrogradsky theorem): Let V be a volume in the space (x,y,z) with the closed boundary ∂V and $\boldsymbol{A}$ be any vector field. Then

$$\int_{\partial V} (\boldsymbol{A} \cdot \boldsymbol{\nu})\,dS = \int_V (\nabla \cdot \boldsymbol{A})\,dx\,dy\,dz \equiv \int_V \operatorname{div} \boldsymbol{A}\,dx\,dy\,dz, \tag{1.1.17}$$

where $\boldsymbol{\nu}$ is the unit outward normal to the boundary ∂V of V.

1.1.2 General case

In what follows, we denote by $\mathbb{R}^n$ a real n-dimensional vector (linear) space. Its elements are n-dimensional vectors $x = (x^1, \ldots, x^n)$, where x^i denote the Cartesian coordinates of a point $x = (x^1, \ldots, x^n) \in \mathbb{R}^n$ referred to rectangular axes. Thus, the *scalar product* of vectors $x = (x^1, \ldots, x^n)$ and $y = (y^1, \ldots, y^n) \in \mathbb{R}^n$ are defined by the formula similar to (1.1.4):

$$x \cdot y = \sum_{i=1}^{n} x^i y^i. \tag{1.1.18}$$

The length of a vector $x = (x^1, \ldots, x^n)$ is

$$|x| = \sqrt{x \cdot x} \equiv \sqrt{\sum_{i=1}^{n} (x^i)^2}\,. \tag{1.1.19}$$

Brought to you by | Stockholms Universitet
Authenticated
Download Date | 10/31/15 1:40 AM

05.10.2019

is the angle between the vector a and b. The scalar product is denoted in the literature by (a, b).

The vectors $a, b \in \mathbb{R}^3$ can be represented in the rectangular Cartesian coordinates in the form

$a=(a^1, a^2, a^3), \quad b=(b^1, b^2, b^3),$

which means that

$a = a^1 i + a^2 j + a^3 k.$

$(a^1 i)$

$b = b^1 i + b^2 j + b^3 k.$

where i, j and k are the unit vectors along the first, second and third coordinate axes in $\mathbb{R}^3$, respectively.

Then the scalar product is given by

$$a.b = \sum_{i=1}^{3} a^i b^i.$$

~~1(95)~~ j) Brans-Dicke theory

Source: https://en.wikipedia.org/wiki/Brans-Dicke_theory

05.10.2019 X

In theoretical physics the Brans-Dicke theory of gravitation (Jordan-Brans-Dicke theory) is a theoretical framework to explain gravitation.

It is a competitor of Einstein's theory of general relativity.

It is an example of a scalar-tensor theory, a gravitational theory in which the gravitational interaction is mediated by a scalar field as well as the tensor field of general relativity.

other:

https://pt.wikipedia.org/wiki/Geometria_hiperbólica/

105) a) Advanced Trigonometric Relations through Nbic Functions.

Source: 037_novosarquivos_2019/
Trigonometric/
Sardet.cd_8122430236.pdf

other: Trigonometric Functions and Complex Numbers

! "

" # # $ $!! % & ' " " "

()*+(((, (- !$ $. /"

0 ()*+(((, (1 2) "
/ 3 4 0 /"

*

()*+(((, (1 2) " / 3
4 0 5

&

$$
\begin{aligned}
360^\circ &= 2\pi \text{ radians} \\
180^\circ &= \pi \text{ radians} \\
1^\circ &= \frac{\pi}{180} \text{ radians} \approx 0.017 \text{ radians} \\
1 \text{ radian} &= \left(\frac{180}{\pi}\right)^\circ \approx 57^\circ 17' 44.8'' \approx 57.3^\circ .
\end{aligned}
$$

Chapter 1

Trigonometric Identity

1.1. Trigonometric Ratios for Any Angle

Any angle and its measures

1. Positive angle, negative angle and zero angle

If a ray rotating around the endpoint counterclockwise formed as the positive angle, its measurement is positive, and if a ray rotating around the endpoint clockwise formed as the positive angle, its measurement is negative (see Fig. 1.1):

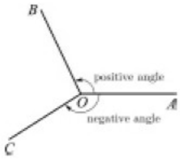

Fig. 1.1.

2. Coterminal angle

In the process of angle forming, we can find there are numerous angles whose initial and terminal sides coincide with a certain angle α. Their range is 360° integer times that of α in size. We put the set of the angles (including α itself) with the same edge as α:

$$\{\beta | \beta = k \cdot 360^\circ + \alpha,\ k \in Z\}.$$

Namely, any angle with the same terminal side as angle α can be expressed as a total of angle α and integer number of perigons.

3. Quadrant angle

If the terminal ray of an angle in standard position lies in the first quadrant, then the angle is said to be a first-quadrant angle. Similarly, second-, third-, and fourth-quadrant angles are defined as follows:

First-quadrant angles: $\{\alpha | k \cdot 360^\circ < \alpha < k \cdot 360^\circ + 90^\circ,\ \mathrm{k} \in Z\}$;

Second-quadrant angles:
$\{\alpha | k \cdot 360° + 90° < \alpha < k \cdot 360° + 180°, k \in Z\}$;
Third-quadrant angles:
$\{\alpha | k \cdot 360° + 180° < \alpha < k \cdot 360° + 270°, k \in Z\}$;
Fourth-quadrant angles:
$\{\alpha | k \cdot 360° + 270° < \alpha < k \cdot 360° + 360°, k \in Z\}$.

4. Quadrantal angle

If the terminal ray of an angle in standard position lies along an axis, the angle is called a quadrantal angle.

- If the terminal ray of the angle lies along positive y-axis, then $\theta = 360°k + 90°$, $k \in Z$.
- If the terminal ray of the angle lies along negative y-axis, then $\theta = 360°k - 90°$, $k \in Z$.
- If the terminal ray of the angle lies along y-axis, then $\{\theta | \theta = 180°k + 90°,\ k \in Z\}$.
- If the terminal ray of the angle lies along positive x-axis, then $\theta = 2 \times 180°k + 0°$, $k \in Z$.
- If the terminal ray of the angle lies along negative x-axis, then $\theta = 2 \times 180°\, k + 180°$, $k \in Z$.
- If the terminal ray of the angle lies along x-axis, then $\{\theta | \theta = 180°k,\ k \in Z\}$.

Exercise

1. Find angles between 0° with 360°, that are coterminal with each given angle, and judging in which quadrantal angle.

(1) $-120°$; (2) $640°$;(3) $-950°12'$.

Solution: (1) Since $-120° = -360° + 240°$, it follows that 240° is coterminal with $-120°$, which is the third-quadrant angle.

(2) Since $640° = 360° + 280°$, it follows that 280° is coterminal with 640°, which is the fourth-quadrant angle.

(3) Since $-950°\ 12' = -3 \times 360° + 129°48'$, it follows that $129°48'$ is coterminal with $-950°\ 12'$, which is the third-quadrant angle.

2. Give a set S for the measure of all angles that are coterminal with each given angle, find angles between $-360°$ and $720°$ in S.

(1) $60°$;(2) $-21°$; (3) $363°14'$.

Solution: (1) $S = \{\beta | \beta = 60° + k \cdot 360°, k \in Z\}$.

The angles between $-360°$ and $720°$ in S are as follows:

$$-1 \times 360^\circ + 60^\circ = -300^\circ;$$
$$0 \times 360^\circ + 60^\circ = 60^\circ;$$
$$1 \times 360^\circ + 60^\circ = 420^\circ.$$

(2) $S = \{\beta | \beta = -21^\circ + k \cdot 360^\circ, k \in Z\}$.

The angles between -360° and 720° in S are as follows:

$$0 \times 360^\circ - 21^\circ = -21^\circ;$$
$$1 \times 360^\circ - 21^\circ = 339^\circ;$$
$$2 \times 360^\circ - 21^\circ = 699^\circ.$$

(3) $S = \{\beta | \beta = 363^\circ\, 14' + k \cdot 360^\circ, k \in Z\}$.

The angles between -360° and 720° in S are as follows:

$$-2 \times 360^\circ + 363^\circ 14' = -356^\circ 46';$$
$$-1 \times 360^\circ + 363^\circ 14' = 3^\circ 14';$$
$$0 \times 360^\circ + 363^\circ 14' = 363^\circ 14'.$$

3. Find the set for the angle, which displays the terminal side of the angle in the set of angles in the shadow area (exclusive of boundaries). (See Fig. 1.2.)

Solution:.

(A) $\{\alpha | 60^\circ + k \cdot 360^\circ < a < 255^\circ + k \cdot 360^\circ, k \in Z\}$.

(B) $\{\alpha | - 120^\circ + k \cdot 360^\circ < \alpha < 45^\circ + k \bullet 360^\circ, k \in Z\}$.

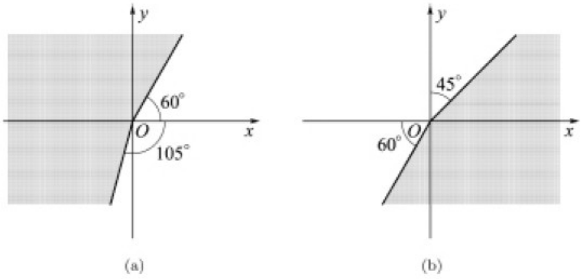

(a) (b)

Fig. 1.2.

4. If α is a second-quadrant angle, how about $\frac{\alpha}{2}$? How about 2α?

Solution: If α is a second-quadrant angle, then
$k \cdot 360^\circ + 90^\circ < \alpha < k \cdot 360^\circ + 180^\circ, k \in Z.$

Hence $k \cdot 180^\circ + 45^\circ < \frac{\alpha}{2} < k \cdot 180^\circ + 90^\circ$, because $k \in Z$. Therefore $k = 2n$ or $k = 2n + 1$.

When $k = 2n$, $n \cdot 360^\circ + 45^\circ < \frac{\alpha}{2} < n \cdot 360^\circ + 90^\circ$, therefore $\frac{\alpha}{2}$ is a first-quadrant angle.

When $k = 2n+1$, $n \cdot 360^\circ + 225^\circ < \frac{\alpha}{2} < n \cdot 360^\circ + 270^\circ$, therefore $\frac{\alpha}{2}$ is a third-quadrant angle.

Hence when α is a second-quadrant angle, therefore $\frac{\alpha}{2}$ is a first-quadrant angle or third-quadrant angle. Similarly, $180°+k\cdot720° < 2a < 360°+k \cdot 720°$, $k \in Z$. Hence, when a is the second-quadrant angle, $2a$ is a third-quadrant angle or the fourth-quadrant angle.

Radian

(1) Radian: When an arc of a circle has the same length as the radius of the circle, as shown in Fig. 1.3, the measure of the central angle $\angle AOB$ is by definition 1 radian.

(2) Radian measure: The radian to measure angle.

(3) Equation for the radian measure of the central angle *α*: $|\alpha| = \frac{l}{r}$, where α is the subtended angle in radians, l the arc length, and r the radius.

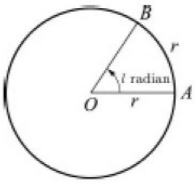

Fig. 1.3.

(4) Conversion formulas between degree measure and radian measure: One revolution measured in radians is 2π and measured in degrees is 360°. We therefore have the conversion formulas:

$$\begin{aligned} 360^{\circ} &= 2\pi \text{ radians} \\ 180^{\circ} &= \pi \text{ radians} \\ 1^{\circ} &= \frac{\pi}{180} \text{ radians} \approx 0.017 \text{ radians} \\ 1 \text{ radian} &= \left(\frac{180}{\pi}\right)^{\circ} \approx 57^{\circ}17'44.8'' \approx 57.3^{\circ}. \end{aligned}$$

Because the radian number is the ratio of the arc length to radius, so it is a real number. Therefore, in expressing the range of angle in radian measure, usually omit the word "radian". Due to different needs, the angle is not only measured in degree, but also in radian. For example, if we want to learn the trigonometric function, we need to draw its image, using the degree measure, which will cause inconvenience. Since the introduction of radian, the set of angles corresponds with the set of real numbers as shown in Fig. 1.4:

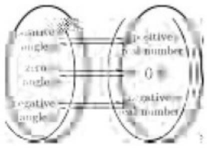

Fig. 1.4.

Degree	0°	15°	30°	45°	60°	75°	90°	105°	120°	135°	150°	180°	210°
Radian	0	$\frac{\pi}{12}$	$\frac{\pi}{6}$	$\frac{\pi}{4}$	$\frac{\pi}{3}$	$\frac{5\pi}{12}$	$\frac{\pi}{2}$	$\frac{7\pi}{12}$	$\frac{2\pi}{3}$	$\frac{3\pi}{4}$	$\frac{5\pi}{6}$	π	$\frac{7\pi}{6}$
Degree	225°	240°	270°	300°	315°	330°	360°						
Radian	$\frac{5\pi}{4}$	$\frac{4\pi}{3}$	$\frac{3\pi}{2}$	$\frac{5\pi}{3}$	$\frac{7\pi}{4}$	$\frac{11\pi}{6}$	2π						

(5) **Formulas for the arc length and area of a sector of a circle**: The following formulas are arc length and area of a sector of a circle, where α $(0 < \alpha < 2\pi)$ is the central angle of a sector, r the radius, l is the length, and S the area (see Fig. 1.5):

$$(1)\ l = \alpha r;$$

$$(2)\ S = \frac{1}{2}\alpha r^2;$$

$$(3)\ S = \frac{1}{2}lr.$$

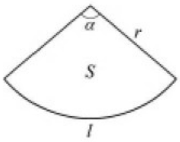

Fig. 1.5.

Exercise

5. The rope is twisted around the rim of 40-cm radius. The lower end B of the rope is hanging an object W (Fig. 1.6). If the wheel rotates counterclockwise by six laps every min, how many seconds does it take to lift objects upward 100 cm from position W?

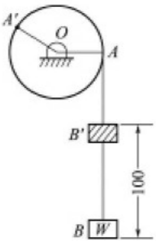

Fig. 1.6.

Solution: When the wheel is rotated in a clockwise direction, the arc length AA $'$ from A to A' is equal to the distance from B to B'. When $BB' = 100$ cm, then $AA' = 100$ cm time arc, and the center angle $\angle AOA'$ corresponding to arc AA' is given by $\frac{100}{40}$ radians.

Since the wheel rotates six times per minute at a uniform speed, the radian $\frac{6\times 2\pi}{60}$ per second is given by $\frac{\pi}{5}$ When T seconds have passed, the radian that passed is

0 ()*+(((, (67 " " 1 3 8
*0 9 : ;<= " " > ; " ? / /"
& ! @

23.3 Let ABC be any triangle. Let D, E, F divide the sides in thirds. Draw AE, BD, CF. Show that the small triangle formed in the inside has area equal to $\frac{1}{7}$ the area of the whole triangle. *Hint*: Review the proof of Proposition 5.4 and use similar triangles.

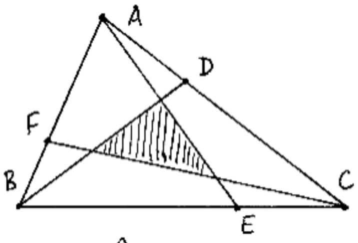

23.4 Let ABC be any triangle, let DE be a line parallel to the base, and let F be any point on DE. Show that the area of the union of the two triangles DBF and ECF is less than or equal to one-fourth the area of the whole triangle, with equality if and only if D and E are the midpoints of AB and AC.

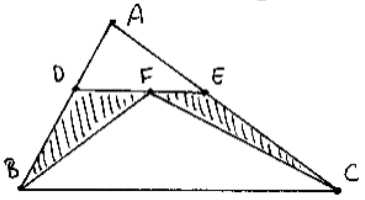

23.5 To get a feeling for ordered abelian groups, try this one. Let $\mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} \mid a, b \in \mathbb{Q}\}$, and similarly $\mathbb{Q}(\sqrt{3})$.

(a) Show that $\mathbb{Q}(\sqrt{2})$ and $\mathbb{Q}(\sqrt{3})$ are isomorphic as abelian groups (under addition).

(b) Show that $\mathbb{Q}(\sqrt{2})$ and $\mathbb{Q}(\sqrt{3})$ are not isomorphic as fields.

(c) Show that $\mathbb{Q}(\sqrt{2})$ and $\mathbb{Q}(\sqrt{3})$ are isomorphic as ordered sets, with the orderings induced by the natural ordering on $\mathbb{R}$. (Here an *ordered set* is a set S, together with a relation $a < b$, having the two properties (i) $a < b$ and $b < c$ implies $a < c$; (ii) if $a, b \in S$, then one and only one of the following holds: $a < b$, $a = b$, $b < a$.)

(d) $\mathbb{Q}(\sqrt{2})$ and $\mathbb{Q}(\sqrt{3})$ are not isomorphic as ordered abelian groups (again taking addition as the group operation).

23.6 Prove Ptolemy's theorem: The rectangle formed of the two diagonals of a cyclic quadrilateral is equal in content to the sum of the two rectangles made by opposite sides of the quadrilateral.

23.7 In a Euclidean plane, show that the results (VI.1) and (VI.14)–(VI.31) hold using equal content for equality of figures.

24 Dissection

As we have seen in the previous sections, Euclid bases his theory of area on adding and subtracting congruent figures. This notion is formalized by Hilbert's definition of equal content (Section 22), which can also be interpreted using a measure of area function (Section 23). While the notion of content gives a good general theory, one could complain that it is not practical, in the sense that if two figures have equal content, one does not know how much must be added in order to make them equidecomposable. In Example 22.1.3, the two figures being compared both have area 1, but to make them equidecomposable, one must add a triangle whose area is infinite in the non-Archimedean field *F*.

In this section we will investigate the stricter notion of when two figures are equidecomposable (Section 22). This leads to the practical problem of *dissection*: given two figures, to find, if possible, an efficient decomposition of the first as a nonoverlapping union of smaller figures, not necessarily triangles, that can be reassembled into the second. A dissection exists if and only if the two figures are equidecomposable. In this case we will also say that one figure can be dissected into the other, or that they are *equivalent by dissection*.

We work in a Hilbert plane with (P), and in some cases use also (A) or (E). We will see that certain of the results, such as the Pythagorean theorem (I.47), which Euclid proved for content, are also true in the stronger sense of dissection. We will prove the theorem, due to Bolyai and Gerwien, that in an Archimedean plane, any two figures of equal area (Section 23) are equivalent by dissection.

The practical problem of finding dissections of one figure into another has received a certain amount of attention among recreational mathematicians, but at this point it seems to be more of an art than a science. Amateurs have discovered a number of clever dissections, but proofs that certain dissections are minimal, or effective bounds on the number of pieces required, seem to be lacking.

We begin this section with some general results on existence of dissections.

Proposition 24.1
In a Hilbert plane with (P), *any triangle can be dissected into a parallelogram.*

Proof (cf. 22.6) Let *ABC* be the triangle. Let *D* be the midpoint of *AC*. Draw lines through *D* parallel to *BC* and through *C* parallel to *AB*, meeting at *F*. Then $\Delta ADE \cong \Delta CDF$ by (ASA). So ΔABC can be dissected into the parallelogram *BCEF*.

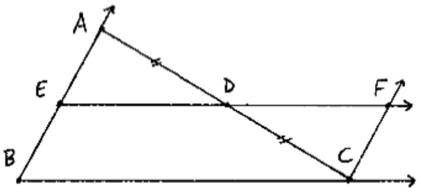

Lemma 24.2
Let ABC be a triangle, and suppose that the foot D of the altitude from A to side BC lands outside the interval BC. Then one of the angles of the triangle, at B or C, is obtuse.

Proof Suppose *B* is between *D* and *C*. Then the angle $\angle ABC$ is an exterior angle of the right triangle *ADB*, so $\angle ABC$ is greater than a right angle, by (I.16).

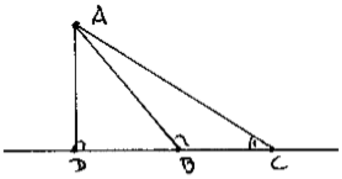

Proposition 24.3

Any parallelogram can be dissected into a rectangle.

Proof Let $ABCD$ be the given parallelogram. Drop the altitudes from C and D to the side AB, with feet E, F. Suppose, for example, that E is inside the interval AB. Then the triangle ΔACE is congruent to the triangle BDF. So the parallelogram is dissected into the rectangle $CDEF$.

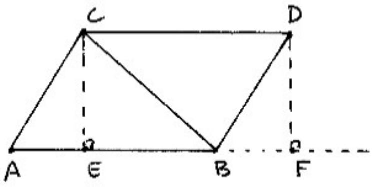

Now let us see that we can always apply this construction. In any parallelogram, the opposite angles are equal (I.34) and the sum of the angles is four right angles, so two of the opposite angles are acute. (If not, all four are right angles and there is noting to prove.) So we may assume that the angle at A is acute.

Now, if the altitude from C lands outside AB, then by (24.2) the angle ABC must be obtuse. This forces the angle ACB to be acute (I.32), and in that case, exchanging the roles of B and C, the altitude from B to AC will land inside the segment AC (24.2), so we can apply the construction above.

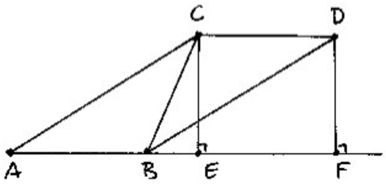

Proposition 24.4

Given a rectangle $ABCD$, and given a segment AE such that $AB < AE < 2AB$, the rectangle $ABCD$ can be dissected into a rectangle having one side equal to AE.

Proof Suppose we are given the rectangle $ABCD$ and the point E. Let C and E be joined, meeting BD at F. Let G be chosen on AC such that $CG \cong BF$. Construct the rectangle $AEGH$, and let K, L be as shown. Because of parallel lines, the angles at C and F are equal. And CG was equal to BF by construction, so by (ASA), $\Delta CGK \cong \Delta FBE$. It follows that $GK \cong BE$, and so by subtraction, $CD \cong AB \cong KH$. Now by (ASA) again, $\Delta CDF \cong \Delta KHE$. This gives a dissection of the rectangle $ABCD$ into the rectangle $AEGH$, as required.

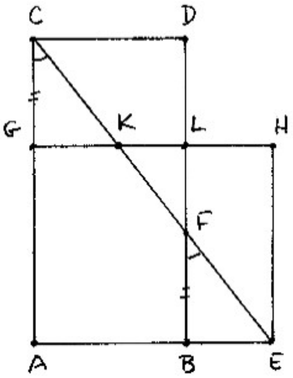

Note that in order for this dissection to work, we need to know that F is below the midpoint of BD, so that G is above the midpoint of AC, and so F is below L. This follows from the hypothesis $AB < AE < 2AB$, because the line from C to the midpoint of BD would meet the line AB at a point M with $AM = 2AB$.

Proposition 24.5
Assume Archimedes' axiom (A). *Given any rectangle ABCD and given any segment EF, there is a rectangle EFGH equivalent by dissection to ABCD.*

Proof Given any rectangle, by cutting it in half and reassembling the two halves along the other sides, we can dissect the original rectangle into a new rectangle with one-half the height and twice the base of the original one. Now by Archimedes' axiom, after doubling or halving AB a finite number of times, we may assume $AB \leq EF < 2AB$. Then we apply (24.4) to get a rectangle $EFGH$ as required.

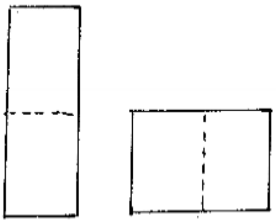

Corollary 24.6
Assuming (A) *and given a segment EF, any rectilineal figure* (*Section* 22) *can be dissected into a rectangle with one side EF.*

Proof Divide the figure P into a finite number of triangles $T_1, T_2, \ldots, T_n$. For each i first dissect the triangle T_i into a parallelogram (24.1), then to a rectangle (24.3), then to a rectangle R_i with base equal to EF (24.5). Now stacking the rectangles $R_1, R_2, \ldots, R_n$, all of which have base equal to EF, on top of each other, we obtain one big rectangle with base EF, as required.

Theorem 24.7 (Bolyai, Gerwien)
In a Hilbert plane with (P) *and* (A), *let* α *be the area function of* (23.2). *Then two figures P and Q are equivalent by dissection if and only if they have the same area.*

Proof One direction follows from (23.1). The new part, which requires the use of Archimedes' axiom, is that figures of equal area are equivalent by dissection. The proof parallels (23.7), where the analogous result was proved for content.

So suppose figures P and Q have equal area: $\alpha(P) = \alpha(Q)$. By (24.6), they can each be dissected into rectangles with sides $1, a$ and $1, b$. Then $\alpha(P) = a$, $\alpha(Q) = b$, so $a = b$, and the rectangles are equal. Thus by transitivity of equidecomposability (22.2), P can be dissected into Q.

Remark 24.7.1
This result is false without (A), as is shown by Example 22.1.3.

Remark 24.7.2
This proof is effective, since given figures P and Q of equal area, and given a decomposition of each into triangles, the method of proof of the results of this section will lead to a dissection of one to the other. Of course, the dissection found in this way may not be efficient in terms of the number of pieces required.

Remark 24.7.3
Combining this result with (23.7) we see that in a Hilbert plane with (P) and (A), two figures have equal content if and only if they are equidecomposable (cf. 22.1.4).

Proposition 24.8
Assume (E). *Let* $ABCD$ *be a rectangle with sides* $a = AB$ *and* $b = AC$ *satisfying* $a \leq b \leq 4a$. *Then there is a segment* $c = AE$ *such that the rectangle* $ABCD$ *can be dissected into a square with side* AE.

Proof Form a segment of length $a + b$, let O be its midpoint, draw a circle with center O passing through the segment's endpoints, and let c be the segment cut off on a perpendicular dividing the segment into $a + b$, from the segment to where it meets the circle. (Here we use (E) for the existence of the intersection point.)

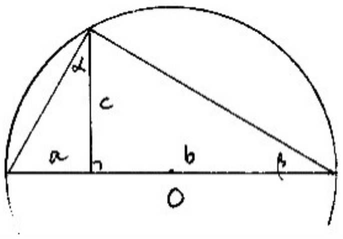

Now let $ABCD$ be the given rectangle, let AE be congruent to the segment c, and let $AEFG$ be the square on side AE. Since the angle in a semicircle is a right angle (III.31), the angles α, β in the circle diagram are equal. These are the same as the angles α, β of the second diagram, by congruent triangles (SAS). Hence the two diagonal lines FB and CE are parallel (I.28). It follows that $CF \cong MB$ and $FK \cong BE$ (I.33). From here the same argument as in the proof of (24.4) shows that $ABCD$ can be dissected into the square $AEFG$.

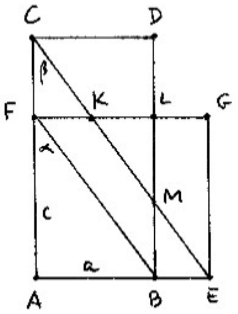

In order for this construction to work, we need to know that $AB \leq AE < 2AB$. We assumed that $a \leq b$, so the angle $\alpha = \beta$ is less than one-half a right angle, and so $a \leq c \leq b$. On the other hand, if $c \geq 2a$, then by similar triangles it follows that $b \geq 2c$. This implies $b \geq 4a$, contrary to our hypothesis.

Corollary 24.9
In a Euclidean plane with (A), *any rectilineal figure can be dissected into a square.*

Proof Choose a segment AB. By (24.6) the original figure can be dissected into a rectangle $ABCD$ with side AB. Cutting the rectangle in half and reassembling a finite number of times (again using (A)), as in the proof of (24.5), we may assume that the sides $a = AB$ and $b = AC$ satisfy $a \leq b \leq 4a$. Then by (24.8) this rectangle can be dissected into a square.

Next, let us look at the Pythagorean theorem (I.47). Euclid's proof shows that the union of the squares on the legs of the right triangle has equal *content* to the square on the hypotenuse. Using Archimedes' axiom (A) one can improve Euclid's results about parallelograms (I.35) and triangles (I.37) to hold also for dissection. Then Euclid's proof will show that (I.47) holds for dissection assuming (A). However, even better than this are several direct proofs of (I.47) by dissection that do not need Archimedes' axiom.

Let us first consider the proof attributed to Thabit b. Qurra (826–901 A.D.).

Proposition 24.10 (I.47)
In a Hilbert plane with (P), *the union of the squares on the legs of a right triangle can be dissected into the square on the hypotenuse.*

Proof Let ABC be the original triangle; let $ABDE$ be the square on AB; let $ACFG$ be the square on AC; fill in the square $GHEK$ and the line LM so that $DM \cong BC$. Then the three triangles ADF, FHG, CKG are congruent to the original triangle ABC, and $GHEK$ is congruent to the square on BC.

Now the square on the hypotenuse is dissected into the five pieces $1, 2, 3, 4, 5$. The first three are congruent to $1', 2', 3'$ by construction. The square on AB is dissected into $1', 3', 5$, while the square on BC is dissected into $2', 4$. So the result is proved.

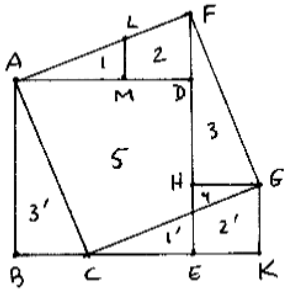

Remark 24.10.1
Here is another proof of (I.47) by dissection. Let ABC be the original right triangle. Draw the squares on the three sides as shown. Let O be the center of the square $ACFG$, and draw lines through O parallel to the sides of the square $BCRS$.

Now let M, N, P, Q be the midpoints of the sides of the square $BCRS$, and draw lines through them parallel to AB and AC, to form the figure shown.

We claim that $1', 2', 3', 4', 5'$ are respectively congruent to $1, 2, 3, 4, 5$, which will complete the proof by dissection.

Since O is the center of the square $ACFG$, it follows that O is also the midpoint of the segments KL and HJ. These two segments are also equal to each other, and from the parallelogram $BCKL$, they are equal to BC. Thus OK, OL, OH, OJ are all equal to one-half of BC; hence are congruent to the segments BM, MC, DP, PS, etc. Since the lines drawn in the big square are parallel to the sides of the square $ACFG$, it follows that $1, 2, 3, 4$ are congruent to $1', 2', 3', 4'$.

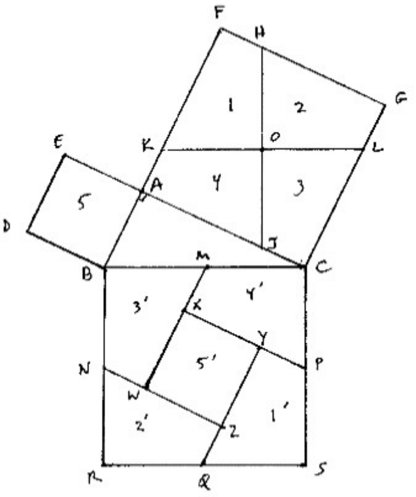

It remains to show that $5 \cong 5'$. Note that $MW \cong LC$, since they are corresponding sides of 3 and $3'$. From the parallelogram $BCKL$ we see that $LC \cong BK$. But also $BK \cong MX$. So by subtraction $XW \cong AB$. The figure is symmetrical, so one can also show that $AJ \cong LC$ and $JC \cong KA$; then a similar argument shows that $XY \cong AB$. Thus $5'$ is a square congruent to 5.

Note that a proof by dissection requires

(1) a diagram to show the pieces and how they correspond,
(2) a careful construction stating how and in what order the diagram was drawn, and
(3) a proof, based on the construction (2), showing that corresponding pieces are congruent.

There may be many ways of making the construction, and it is worth contemplating which will make the proof easier.

Exercises

The following exercises take place in a Hilbert plane with (P). Assume that it is a Euclidean plane, or the Cartesian plane over a field, as needed.

24.1 Using Archimedes' axiom (A) give a direct proof of (I.37) by dissection.

24.2 Dissect a square into three equal smaller squares.

24.3 Use the accompanying diagram to provide another proof of the Pythagorean theorem (I.47) by dissection. You must supply a construction of the diagram, and then a proof based on your construction, that corresponding pieces are congruent.

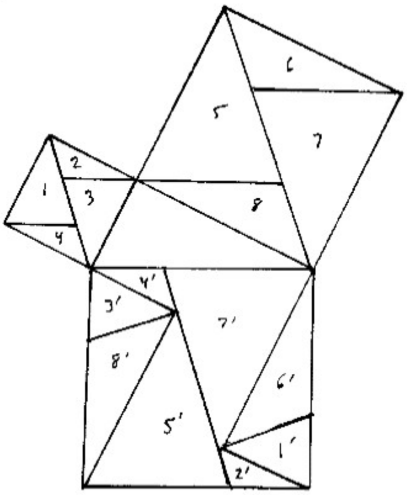

24.4 Find a dissection of an equilateral triangle into a square. After writing your construction and proof, make a model with construction paper and numbered pieces to illustrate your dissection.

24.5 Prove or disprove the existence of a dissection of a square of side 21 into a rectangle of sides 13, 34 according to the diagram below.

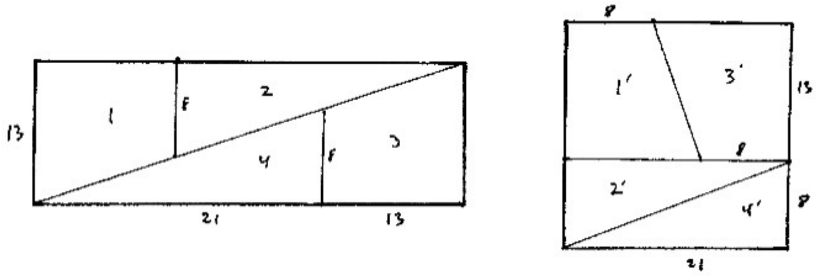

24.6 Let ABC be any triangle. Let D be a point on AC. Prove that there is a dissection of ABC into a trapezoid $PQRS$ as shown. What condition is necessary on the point D for this to work?

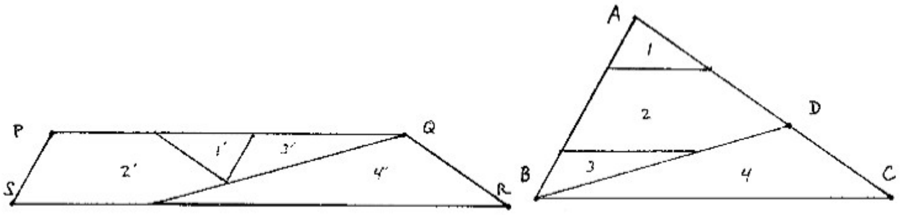

24.7 Use the previous exercise to find a dissection of two equilateral triangles of side 1 into a single equilateral triangle of side $\sqrt{2}$.

24.8 Prove the existence of Dudeney's (1929) dissection of an equilateral triangle into a square using four pieces, as suggested in the accompanying diagram.

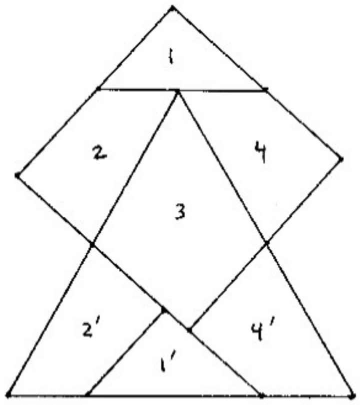

24.9 Dissect a regular pentagon into a square.

24.10 Prove or disprove: For any integer $n \geq 1$, the minimum number of triangles required to dissect a square of side 1 into a rectangle of sides n and $1/n$ is $2n$.

24.11 Dissect a square into eight strictly acute triangles (all angles less than 90°).

24.12 If a square is dissected into n triangles of equal area, then n must be an even number.

24.13 Consider a restricted form of dissection problem, where you can cut the figure in pieces and move the pieces around in the plane, *using translations only* (i.e., no rotations, and no turning pieces over). Show that you can dissect the unit square, with translations only, into the unit square with any other given orientation.

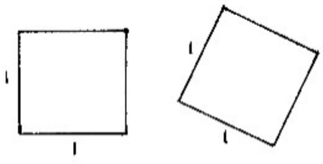

24.14 Divide the unit square by its diagonal into two (congruent) triangles T_1, T_2. Show that it is impossible to dissect T_1 into T_2 using translations only.

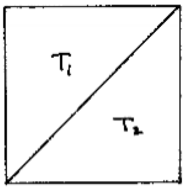

24.15 If P and P' are any two figures with the same area, then assuming Archimedes' axiom, show that it is possible to dissect P into P' using translations and 180° rotations only.

24.16 Let T be a triangle whose smallest angle is greater than or equal to 45°. Then T can be dissected into a square using 5 or fewer pieces.

24.17 Let T be a triangle having an angle α with $\tan\alpha < 1/a$ for some $a \in F$, $a > 1$. Then any dissection of T into a square requires $\geq \frac{1}{2}\sqrt{a}$ pieces. In particular, there exist triangles requiring arbitrarily many pieces to dissect into a square.

24.18 Let T be a triangle whose smallest angle α satisfies $\tan\alpha > 1/a$, for some $a \in F$, $a \geq 1$. Then T can be dissected into a square using no more than $3\sqrt{a} + 4$ pieces.

25 Quadratura Circuli

No discussion of area would be complete without mention of the classical problem of squaring the circle. Through the ages this problem has had an influence way beyond the confines of mathematics as an archetype of the insoluble problem. In one of his religious meditations, John Donne (early seventeenth century) mentions squaring the circle as something only God can achieve. Perhaps the expression "to put a round peg in a square hole" is a modern vestige of this old problem. And yet what a natural problem it is! The circle is the perfect round form, while the square is complete with its four equal sides and equal angles. What could be more natural than to transform the one into the other? But the solution has been as elusive as the alchemist's search for a way to transform lead into gold.

The problem is to construct a square with area equal to a given circle. What does this mean, exactly? By "construction," of course, is meant a geometrical construction with ruler and compass. But what is the area of a circle? In the theory of area presented so far, which corresponds to Euclid's usage in Books I–VI of the *Elements*, we have discussed the area of rectilineal figures only. So there is another problem implicit in the problem of squaring the circle, namely to give a good definition of the area of a circle. In the language of Section 23, we might ask for a measure of area function defined on all plane regions that can be bounded by straight line segments and arcs of circles. A modern approach to this (in the real Cartesian plane) is to use a definite integral to define the area.

Euclid never defines the area of a circle, and wisely avoids the problem of squaring the circle, but in Book XII, he does have a proposition saying that the ratio of the area of one circle to another is equal to the ratio of the squares of their radii. His proof makes use of the "method of exhaustion," which we would call a limiting process. The key result, which is remarkably similar to the treatment of the Riemann integral in modern calculus classes, is this.

Proposition 25.1
In a Euclidean plane with (A), *given a circle* Γ, *one can find an inscribed polygon* P *and a circumscribed polygon* P' *such that the difference in area* $\alpha(P') - \alpha(P)$ *is less than any preassigned quantity* (*as an element of the field* F *of segment arithmetic*).

Proof Start with any inscribed polygon P and any circumscribed polygon P', where the vertices of P are the points of tangency of P'. If we double the number of sides of P and P' by bisecting all the central angles of P, then we get new polygons P_1 and P_1'. We compare the difference of P_1 and P_1' to the difference of P and P'. Let us look at the triangle ABC, which is one piece of $P' - P$. The new difference $P_1' - P_1$ intersected with ABC is represented by the shaded area.

Euclid shows that the shaded area is less than one-half of the area of the triangle ABC. In fact, its area is less than or equal to $\frac{1}{4}$ of the area of that triangle (Exercise 23.4). So every time we double the number of sides, the new difference in areas of outside and inside polygons is less than half of the previous difference. Hence, using Archimedes' axiom (A), it can be made less than any preassigned quantity.

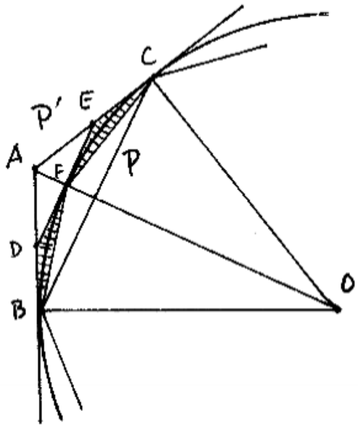

Approximating the circle by its inscribed and circumscribed polygons, Euclid could show that the ratio of the area of a circle to the square of its radius was a constant independent of the size of the circle. This constant, of course, is what we now call π. Archimedes, who lived shortly after Euclid, used the method of approximating by polygons to obtain the estimate $3\frac{10}{71} < \pi < 3\frac{1}{7}$, or in fractions, $\frac{223}{71} < \pi < \frac{22}{7}$. Thus began one facet of the work inspired by the problem of squaring the circle, namely to find increasingly accurate numerical approximations for π. Energetic human calculators worked hard on this problem, so that by

1600, the value of π was known to 15 decimal places, by 1700 to 71 places, and by 1873 to more than 500 places. Even today, the urge to compute is still strong, and with the help of electronic calculators, the value of π has been found to more than one billion decimal places.

Coming back to the original problem of squaring the circle, it has always been regarded as very difficult, if not impossible. Since Lindemann proved in the 1880s that π is transcendental we have known that the problem as originally stated is mathematically impossible (cf. 13.2). Nevertheless, the problem has exerted a great fascination on the inquiring mind, and over the years, many people have come forward with solutions. The study of these solutions forms a colorful chapter in the history of mathematics.

The solutions can be divided into three general categories.

I. Honest ruler and compass constructions that, however, give only an approximate solution to the problem. One of the best of these depends on the coincidental proximity of the number $\frac{1}{5}(3\sqrt{5}+9) = 3.14164\ldots$ to π.

II. Constructions using additional tools that give an exact solution to the problem. One such known from antiquity uses an auxiliary curve, the *quadratrix*.

Take a quadrant AOB of the unit circle. Let OA_1 bisect the right angle, and let B_1 bisect the interval OB. Let the radius OA_1 meet the horizontal line through B_1 at C_1. Bisect the two 45° angles at O with OA_2 and OA_3, and bisect the intervals OB_1 and B_1B with B_2, B_3. Let the radii meet the corresponding horizontal lines in C_2 and C_3. Continuing in this manner, bisecting angles and segments, one can construct as many points as one likes $C_1, C_2, C_3, \ldots$. The curve passing through these points is called the quadratrix.

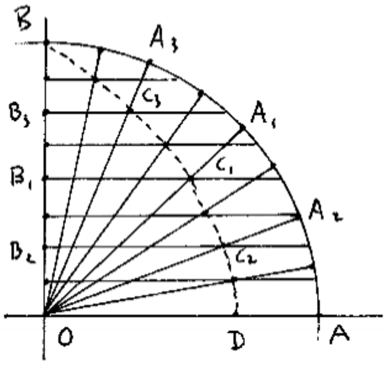

With a little help from modern trigonometry we can express the curve parametrically as

$$\begin{cases} x = \dfrac{2\theta}{\pi \tan \theta}, \\ y = \dfrac{2\theta}{\pi}, \end{cases}$$

where θ is the angle of the radius to the x-axis. As θ tends to 0 we obtain the point D where the curve meets the x-axis. Its x-coordinate is $2/\pi$ (Exercise 25.1). So clearly, if we are given a copy of the quadratrix curve, we can obtain the quantity π, and hence can construct a square with area equal to a given circle. Note, however, that this is not an honest ruler and compass construction:

Though we can obtain as many of the C_i as we like by ruler and compass, the point D is obtained only as a limit of the C_i as they approach the x-axis.

III. Solutions by people I would like to call fringe mathematicians, though they have been called by worse names, who have little mathematical training but are attracted to famous problems like the moth to the flame. The hint that a problem may be insoluble inspires them to work even harder, as they have a conviction that they will succeed where others have failed. Their solutions are a jumble of facts and fallacies lacking in logical coherence. Pointing out their errors and quoting the results of established mathematicians only confirms their belief that they have discovered the real truth that all the scholars were too blind to see. Sometimes their solutions come from divine inspiration, which leaves little room for discussion, since one cannot argue with God.

A case in point is Joseph Ignati Carl von Leistner, a knight in the service of Francis I (1708–1765), duke of Lorraine, and later Holy Roman emperor. Herr Leistner discovered that the exact value of π was 3844/1225. When his discovery did not occasion the acclaim he expected, he used his influence with the duke to appoint a imperial commission to report on his work. This commission naturally pointed out his errors, upon which he published a lengthy rebuttal attacking his critics and defending his work. Apparently, Herr Leistner believed that the true value of π would be a ratio of integers, but that no one had yet been clever enough to find the right ones, as he did. One of his arguments goes like this: "Archimedes proposed the values 22/7 and 223/71, but even he acknowledged that they were not correct, while mine is the correct value." Herr Leistner failed to notice that his value fell outside of the bounds proved by Archimedes. On the occasion of his lord's marriage to Maria Theresa of Austria in 1736, Herr Leistner published a treatise on the wonderful coincidence of the numbers 3844 and 1225 with various dates and other numbers from the life of the duke and his bride.

Exercises

25.1 (a) Derive the parametric equations for the quadratrix given in the text.

(b) Use a little first-year calculus to show that the point D has x-coordinate $2/\pi$.

25.2 (a) Let Γ be the circle of radius 1. Find the area of the inscribed polygons of 6, 12, 24, 48, 96, 192 sides, and express them in standard form (Exercise 13.2) as a rational number times an expression involving integers and square roots. Then verify that the area of the inscribed 192-gon is greater than 223/71, thus confirming Archimedes' estimate $223/71 < \pi$. (Check: To four decimals, area of inscribed 48-gon is 3.1326, 96-gon is 3.1394, 192-gon is 3.1410.)

(b) Show that the area of the circumscribed 96-gon is less than 22/7, confirming the inequality $\pi < 22/7$.

()*+(((, (7 A B : C- ?
> D B > /"

" ()*+(((, (: ; * A " C.'? 1
)) /"

& !' C C ;E

1 C

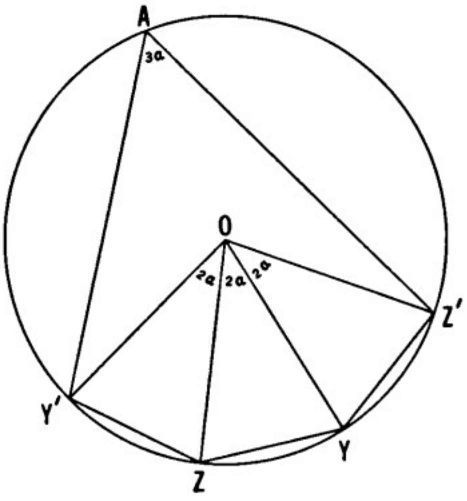
A
3α
O
2α
2α
2α
Z′
Y′
Y
Z

2 C)

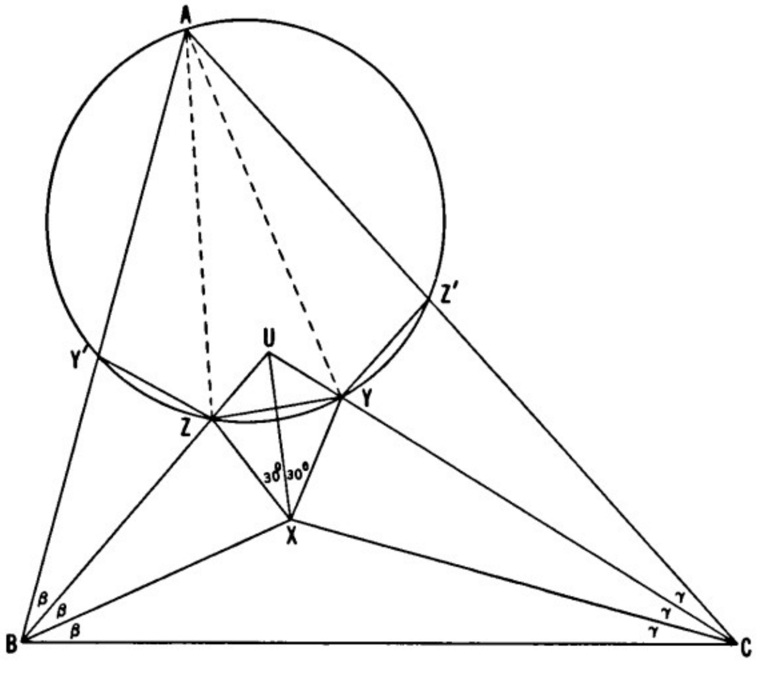
A
Z′
U
Y′
Y
Z
30°
30°
X
β
β
β
γ
γ
γ
B
C

2 C>

3. Let the incircle (with center I) of $\triangle ABC$ touch the side BC at X, and let A' be the midpoint of this side. Then the line $A'I$ (extended) bisects AX.

2.9 Morley's theorem

One of the most surprising theorems in elementary geometry was discovered about 1904 by Frank Morley (the father of Christopher Morley, whose novel, *Thunder on the Left*, has a kink in its time sequence that appeals particularly to geometers). He mentioned it to friends in Cambridge, England, and published it twenty years later in Japan. Meanwhile it was rediscovered and presented as a problem in the Educational Times. Two solutions were sent in, one of which, by M. T. Naraniengar,† is as neat as any of the dozens that have been devised since then. The theorem states:

THEOREM 2.91. *The points of intersection of the adjacent trisectors of the angles of any triangle are the vertices of an equilateral triangle.*

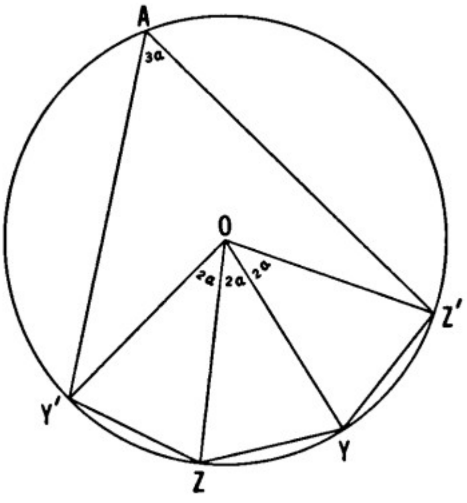

Figure 2.9A

Naraniengar's proof requires a preparatory theorem or lemma (illustrated in Figure 2.9A):

LEMMA. *If four points Y', Z, Y, Z' satisfy the conditions*

---

† Mathematical Questions and Their Solutions from the Educational Times (New Series), **15** (1909), p. 47.

$$Y'Z = ZY = YZ'$$

and

$$\angle YZY' = \angle Z'YZ = 180° - 2\alpha > 60°$$

then they lie on a circle. Moreover, if a point A, *on the side of the line* $Y'Z'$ *away from* Y, *is so situated that* $\angle Y'AZ' = 3\alpha$, *then this fifth point* A *also lies on the same circle.*

To prove the lemma, let the internal bisectors of the equal angles YZY' and $Z'YZ$ meet at O. Then $OY'Z$, OZY, OYZ' are three congruent isosceles triangles having base angles $90° - \alpha$. Their equal sides OY', OZ, OY, OZ' are radii of a circle with center O, and their angles at this common vertex are 2α. In other words, each of the equal chords $Y'Z$, ZY, YZ' subtends an angle 2α at the center O and consequently subtends an angle α at any point on the arc $Y'Z'$ not containing Y. This arc may be described as the locus of points (on the side of the line $Y'Z'$ away from Y) from which the chord $Y'Z'$ subtends an angle 3α. One such point is A; therefore A lies on the circle.

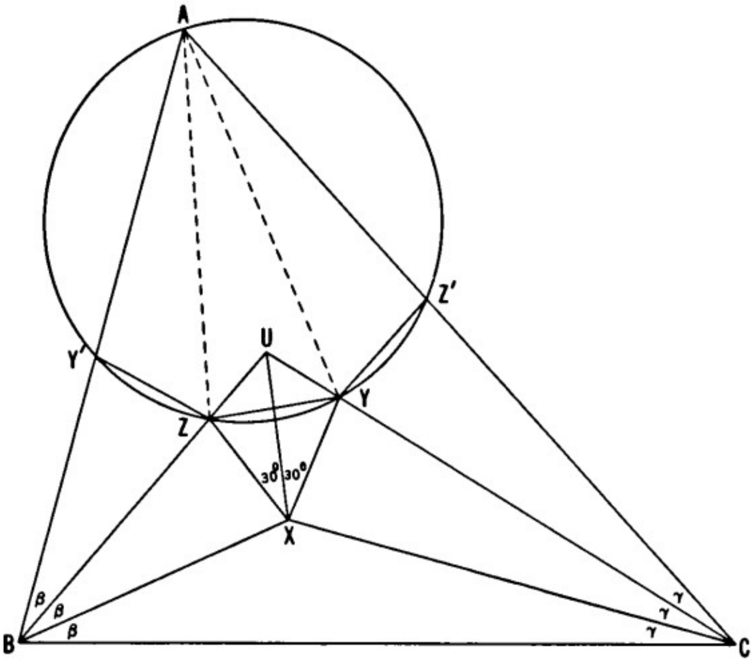

Figure 2.9B

Now we are ready to attack Theorem 2.91 itself. In Figure 2.9B, the trisectors of the angles $B = 3\beta$ and $C = 3\gamma$ meet as shown at the points U and X. In $\triangle BCU$, the angles at B and C are bisected by BX and CX; hence X is the incenter, and the angle at U is bisected by UX. If we construct points Y and Z on the lines CU and BU so that XY and XZ make equal angles 30° with XU on opposite sides, then $\triangle UXY \cong \triangle UXZ$, $XY = XZ$, and since the angle at X is 60° it follows that $\triangle XYZ$ is equilateral.

Also $\triangle UZY$ is isosceles. Its angle at U is the same as that of $\triangle UBC$, whose other angles are 2β and 2γ; therefore the equal angles of $\triangle UYZ$ at Y and Z are each $\beta + \gamma$.

Writing $\alpha = A/3$, we deduce from $A + B + C = 180°$ that

$$\alpha + \beta + \gamma = 60^\circ, \qquad \text{whence} \qquad \beta + \gamma = 60^\circ - \alpha.$$

Thus

$$\angle YZU = 60^\circ - \alpha \qquad \text{and} \qquad \angle XZU = 120^\circ - \alpha.$$

Our next step is to mark off $BY' = BX$ on BA, and $CZ' = CX$ on CA. We now have

$$\triangle BZX \cong \triangle BZY' \qquad \text{and} \qquad \triangle CYX \cong \triangle CYZ',$$

so that

$$Y'Z = ZX = ZY = YX = YZ'.$$

Before we can apply the lemma, we still have to evaluate $\angle YZY'$ and $\angle Z'YZ$. However, this is a simple matter. Since the equal angles BZY' and BZX have equal supplements,

$$\angle UZY' = \angle XZU = 120^\circ - \alpha$$

and

$$\angle YZY' = \angle YZU + \angle UZY' = (60^\circ - \alpha) + (120^\circ - \alpha)$$
$$= 180^\circ - 2\alpha.$$

Similarly, $\angle Z'YZ = 180^\circ - 2\alpha$; and of course $\alpha = \frac{1}{3}A < 60^\circ$.

Applying the lemma, we deduce that the five points Y', Z, Y, Z', A all lie on a circle. Since the equal chords $Y'Z$, ZY, YZ' subtend equal angles α at A, the lines AZ and AY trisect the angle A of $\triangle ABC$. In other words, the points X, Y, Z, which were artificially constructed so as to form an equilateral triangle, are in fact the points described in Morley's theorem. The proof is now complete.

EXERCISES

1. Let the angle trisectors AZ and CX (extended) meet at V, BX and AY at W. Then the three lines UX, VY, WZ are concurrent. (That is, in the language of projective geometry, UVW and XYZ are *perspective* triangles. In general, UVW is *not* equilateral.)

2. For what kind of triangle ABC will the pentagon $AY'ZYZ'$ be regular?

3. When $\triangle ABC$ is equilateral, the four points Y', Z, Y, Z' occur among the vertices of a regular enneagon (9-gon) in which A is the vertex opposite to the side ZY.

4. For a triangle with angles 3α, 3β, 3γ and circumradius R, Morley's triangle has sides $8R \sin \alpha \sin \beta \sin \gamma$.

5. If $Z'Y = YZ = ZY'$ on the side $Z'Y'$ of a rectangle $BCZ'Y'$ whose center X forms an equilateral triangle with Y and Z, then BX and BZ trisect the right angle at B.

()*+(((, ($$ C '!C! /"

1 D 1 F
1 ; 0 ; "

7/" "

% ()*+(((, (!' C. /"

1)"A " : ; A " //

CHAPTER III

THE STRAIGHT LINE

1. *Collinearity of three points, concurrency of three straight lines*

We have already used the formula $z = 1 + ju$ for the straight line, passing through the point 1 in the direction of the vector j. If we wish to represent the line passing through the point z_1 in the direction of the vector z_2, we can generalize the above formula to $z = z_1 + z_2 u$.

The equation of the line passing through two given points z_1 and z_2, is however, more often required. Point P lying on this line in such a way that the distances from P to z_1 and z_2 are in the proportion u to 1 is given by: (fig. 28):

$$z = \frac{z_1 + uz_2}{1 + u}.$$ [illegible]

For $u = 0$, P coincides with z_1, for $u = \infty$, P coincides with z_2. Positive values of u occur between z_1 and z_2, negative values on both sides outside the stretch $z_1 z_2$; the point at infinity of the line is characterized by the parameter value $u = -1$.

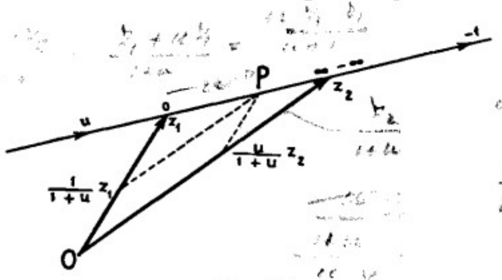

Fig. 28

Calculating u from the equation gives:

$$-u = \frac{z - z_1}{z - z_2}$$

and from this equation we see at once that the three points lie in one line as the reality of either side means that the vectors $z - z_1$ and $z - z_2$ have the same direction.

We can bring the criterion for collinearity of three points z_1, z_2 and z_3 in a more symmetrical form. The argument factor of

$$\frac{z_3 - z_1}{z_3 - z_2}$$

must equal ± 1; in formula:

$$\frac{z_3 - z_1}{z_3 - z_2} \cdot \frac{z_3^* - z_2^*}{z_3^* - z_1^*} = 1$$

and this can be written in the symmetrical form:

$$(z_1 z_2^* - z_1^* z_2) + (z_2 z_3^* - z_2^* z_3) + (z_3 z_1^* - z_3^* z_1) = 0$$

or, in the form of a determinant, the criterion of collinearity becomes:

$$\begin{vmatrix} 1 & 1 & 1 \\ z_1 & z_2 & z_3 \\ z_1^* & z_2^* & z_3^* \end{vmatrix} = 0.$$

We now proceed to calculate the point of intersection of the two lines passing through z_1 and z_2 and through z_3 and z_4 respectively. The two lines are:

$$z = \frac{z_1 + u z_2}{1 + u} \quad \text{and} \quad z = \frac{z_3 + v z_4}{1 + v}.$$

For the point of intersection these two expressions are equal, hence

$$(1 + v)\ (z_1 + u z_2) = (1 + u)\ (z_3 + v z_4)$$

For the conjugate values a similar equation must be valid:

$$(1 + v)\,(z_1^* + u z_2^*) = (1 + u)\,(z_3^* + v z_4^*)$$

and from these two equations u and v of the point of intersection can be calculated. Substituting these values into the line equations yields for the point of intersection:

$$z = \frac{(z_1 z_2^* - z_1^* z_2)\,(z_3 - z_4) - (z_3 z_4^* - z_3^* z_4)\,(z_1 - z_2)}{(z_1 - z_2)\,(z_3^* - z_4^*) - (z_1^* - z_2^*)\,(z_3 - z_4)}.$$

If the vectors $z_1 - z_2$ and $z_3 - z_4$ are parallel, the denominator equals zero by the criterion of parallellity and then $z=\infty$, as was to be expected.

If we cut the line $z_3 z_4$ by a third line $z_5 z_6$, we create a second point of intersection, the place of which is given by the last equation, if only we add 2 to each index. If the three lines have to pass through one point (in a technical term, if the lines are *concurrent*), the two expressions for the points of intersection have to be put equal to each other and the condition resulting from this equality can be brought into the symmetrical form of a determinant:

3 lignes [illegible]

$$\begin{vmatrix} z_1-z_2 & z_3-z_4 & z_5-z_6 \\ \dot{z}_1-\dot{z}_2 & \dot{z}_3-\dot{z}_4 & \dot{z}_5-\dot{z}_6 \\ z_1\dot{z}_2-\dot{z}_1 z_2 & z_3\dot{z}_4-\dot{z}_3 z_4 & z_5\dot{z}_6-\dot{z}_5 z_6 \end{vmatrix}=0.$$

We can check this formula by laying the origin in the common point of intersection. Each element of the third row is proportional to the area of the parallelograms on the vectors z_1z_2, z_3z_4 and z_5z_6 respectively and all three must vanish if the origin is on the lines z_1z_2, z_3z_4 and z_5z_6, but then the determinant is also zero.

2. *The theorems of Ceva, Menelaos and Desargues*

As applications of the general theory let us prove the theorems of Menelaos and Ceva.

Menelaos (58 A. D.): the three points P, Q, R, each lying on a side of the triangle ABC (fig. 29) in such a way that

$$\frac{AP.BQ.CR}{PB.QC.RA}=-1 \text{ are collinear.}$$

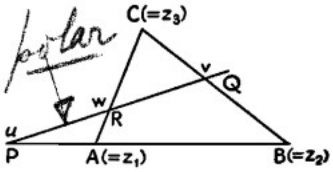

Fig. 29

Let the three points be:

$$z_P=\frac{z_1+uz_2}{1+u};\quad z_Q=\frac{z_2+vz_3}{1+v};\quad z_R=\frac{z_3+wz_1}{1+w}.$$

The condition is then equivalent to $u\cdot v\cdot w=-1$. The determinant occurring in the criterion for collinearity is

$$\begin{vmatrix} 1+u & 1+v & 1+w \\ z_1+uz_2 & z_2+vz_3 & z_3+wz_1 \\ \dot{z}_1+u\dot{z}_2 & \dot{z}_2+v\dot{z}_3 & \dot{z}_3+w\dot{z}_1 \end{vmatrix}.$$

[illegible]

[illegible]

Add the first row and $-u$ times the second to $-1/w=uv$ times the third row. The members of this row then become zero from which it follows that the determinant is zero and hence that the points PQR are collinear, q.e.d.

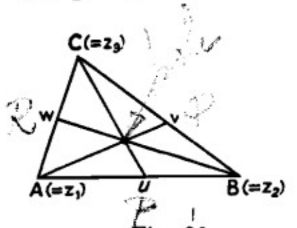

Fig. 30

CEVA (1700): the three lines joining the vertices of a triangle with the points PQR, each lying on one side so that

$$\frac{AP.BQ.CR}{PB.QC.RA}=1 \text{ are concurrent.}$$

Let the three points be:

$$z_P=\frac{z_1+uz_2}{1+u};\quad z_Q=\frac{z_2+vz_3}{1+v};\quad z_R=\frac{z_3+wz_2}{1+w},$$

then the condition leads to $u\cdot v\cdot w=1$. The determinant of the criterion of concurrency is:

$$\begin{vmatrix} (1+u)z_3-(z_1+uz_2) & (1+v)z_1-(z_2+vz_3) & (1+w)z_2-(z_3+wz_1) \\ (1+u)z_3^{\bullet}-(z_1^{\bullet}+uz_2^{\bullet}) & (1+v)z_1^{\bullet}-(z_2^{\bullet}+vz_3^{\bullet}) & (1+w)z_2^{\bullet}-(z_2^{\bullet}+wz_1^{\bullet}) \\ (1+u)\{z_3(z_1^{\bullet}+uz_2^{\bullet})-z_3^{\bullet}(z_1+uz_2)\} & \text{cycl.} & \text{cycl.} \end{vmatrix}.$$

The sum of v times the first column, the second column and $1/w=uv$ times the third is zero; therefore the determinant is zero, which proves the concurrency of the three lines.

Each Ceva's point characterized by three definite values of u, v and w is correlated to one Meneloas' line, characterized by $-u$, $-v$, $-w$. They are taken together as *trilinear pole* and *trilinear polar*. The polar of the centre of gravity of the triangle $(u=v=w=1)$ is the straight line at infinity.

Another important theorem is Desargues' *theorem*, stating that corresponding sides of two perspective triangles intersect in three collinear points, ($\pm$1650), and we may see as follows, that it is a consequence of the theorems of MENELAOS and CEVA. Draw

[illegible] $\triangle ABC$ + [illegible] → $\triangle$ [illegible]

(fig. 31) the triangle ABC and mark the points u, v and w on its sides. By Ceva's theorem the triangles ABC and vwu are perspective if $u \cdot v \cdot w = 1$. Mark the points $-u$, $-v$ and $-w$ and draw the straight line m through these points, which is possible because of Menelaos' theorem. Now as $uvw = 1$, the points w, u, $-v$ must be collinear and also the points v, u, $-w$.... and also the points w, v, $-u$. We see thus that the corresponding sides of the triangles ABC and vwu intersect in three points of the line m, q.e.d. ([illegible] $\triangle uvw$ [illegible])

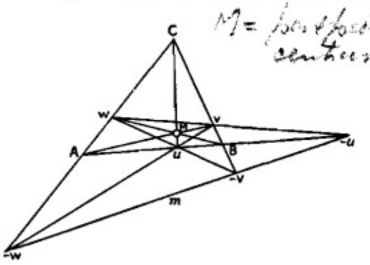

Fig. 31

M = [illegible] centrum.

3. *Line coordinates*

If in the criterion of collinearity of three points we substitute z for z_3, it becomes:

$$\begin{vmatrix} 1 & 1 & 1 \\ z_1 & z_2 & z \\ z_1^* & z_2^* & z^* \end{vmatrix} = 0$$

and this can be considered as the equation of the line passing through z_1 and z_2; written in one line the equation is:

$$(z_1^* - z_2^*)\, z + (z_2 - z_1)\, z^* + (z_1 z_2^* - z_2 z_1^*) = 0.$$

and this can be brought into the simple form:

$$p^* z + p z^* - 2 = 0 \qquad (\text{or } \operatorname{Re}(p^* z) = 1)$$

by introducing:

$$p = 2\,\frac{z_2 - z_1}{z_1^* z_2 - z_1 z_2^*}; \quad p^* = 2\,\frac{z_2^* - z_1^*}{z_1 z_2^* - z_1^* z_2}.$$

The straight line appears to be completely determined by the choice of p. p is called the vector of the line and its components are called the line coordinates. A point is determined by one vector or by two coordinates and we see that the same applies to a straight line.

The use of the line vector appears at once, if we introduce it

incident met

[illegible] →

[illegible] 3 [illegible]

[illegible]

into the formula for the point of intersection of two lines, deduced in a former section. Let the line z_1z_2 be characterized by p_1 and the line z_3z_4 by p_2, where

$$p_1 = 2\,\frac{z_2 - z_1}{\overset{\bullet}{z}_1 z_2 - z_1 \overset{\bullet}{z}_2}\,; \quad p_2 = 2\,\frac{z_4 - z_3}{\overset{\bullet}{z}_3 z_4 - z_3 \overset{\bullet}{z}_4}\,.$$

After introducing this in the formula for the point of intersection z it becomes:

$$z = 2\,\frac{p_2 - p_1}{\overset{\bullet}{p}_1 p_2 - p_1 \overset{\bullet}{p}_2}$$

and there exists a remarkable reciprocity between the point of intersection z of the two lines determined by the vectors p_1 and p_2 and the line p connecting the two points determined by the vectors z_1 and z_2.

Let us also introduce the notion of the line vector into the condition for the concurrency of three lines. This takes then the form:

$$\begin{vmatrix} 1 & 1 & 1 \\ p_1 & p_2 & p_3 \\ \overset{\bullet}{p}_1 & \overset{\bullet}{p}_2 & \overset{\bullet}{p}_3 \end{vmatrix} = 0$$

which is formally identical with the condition for the collinearity of three points. Indeed we may say that three lines are concurrent, if their p's are collinear.

We shall now investigate the geometrical nature of the vector p, defining the point P (fig. 32). As in the definition of p the denominator is purely imaginary, we conclude that the vector from the origin to P is normal to the vector $z_2 - z_1$, that means normal to the

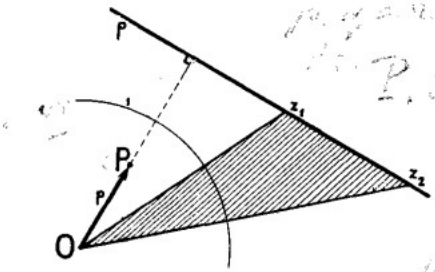

Fig. 32

[illegible]

dualität.

line z_1z_2. As, furthermore, the absolute value of the denominator equals twice the area of the triangle $O z_1 z_2$, of which z_1z_2 is the base, we easily conclude that the modulus of p is the inverse of the length of the perpendicular let down from O on to the line. Point P is the inversion of the base of this perpendicular.

From now on we shall take point P and the corresponding line together as *pole* and *polar* with respect to the unit circle; both the pole and the line will be characterized by the letter p. As we have seen above, by the transformation of pole and polar a pencil of rays passing through one point is transformed into a point assemblage on a straight line. We can prove that this line is the polar of the point which carries the pencil of rays. Let this point be Q with coordinate q. The condition that Q shall lie on an arbitrary member p of the pencil is:

$$p^* q + p q^* - 2 = 0$$

but owing to the symmetry between p and q this is also the condition that P shall fall on the polar q of Q, q.e.d.

The polar of the point of intersection R of the lines p and q (fig. 33) must be the line PQ. PQR is a so-called *polar triangle.*

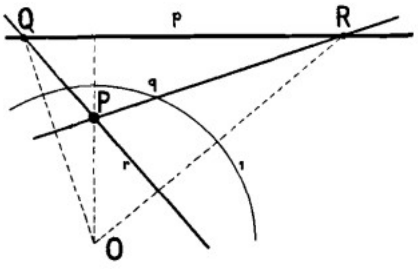

Fig. 33

By the transformation of pole and polar it is transformed into itself.

4. *Polar transformation, dual conceptions*

By the transformation of pole and polar a curve considered as a collection of points is transformed into a collection of straight lines. These lines will envelop a second curve which, therefore, may be considered as the transform of the original curve. A secant

[illegible]

of the original curve, intersecting it in n (= the order) points will be transformed into a point carrying n lines from the collection that envelops the second curve. As these lines are tangents to this curve, the transform will be of the class n. Conversely the class of the original curve will equal the order of the transform. As conics are both of second order and second class, the transformation of pole and polar will transform conics into conics. We know that the inversion transforms circles into circles, but this is not the case with the transformation of pole and polar. By way of example we shall transform the circle, but first we have to indicate the general way of finding the formula of the polar transform.

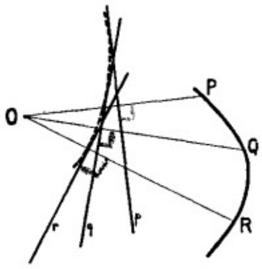

Fig. 34

The point of intersection of the lines p and q which are the transforms of the points P with coordinate p and Q with coordinate q is:

$$z = 2\,\frac{q-p}{p^{\cdot}q - pq^{\cdot}}.$$

Now let Q approach to P, then the line PQ will approach to a limiting position which is a tangent of the original curve. In the transform the point of intersection of p and q will approach a limiting position which by definition will be considered as a point of the transformed curve. Now if $q-p=dp$,

$$p^{\cdot}q - pq^{\cdot} = p^{\cdot}(p+dp) - p\,(p^{\cdot}+dp^{\cdot}) = p^{\cdot}dp - p\,dp^{\cdot}$$

and the point of the transformed curve is:

$$z = 2\,\frac{dp}{p^{\cdot}dp - p\,dp^{\cdot}}.$$

Performing the polar transformation for a second time that is, finding $2\,\dfrac{dz}{z^{\cdot}dz - z\,dz^{\cdot}}$ gives of course again the original curve p, as may be verified by actually performing the calculation.

We shall now apply this general result to the circle (fig. 35)

$$p = a + \exp(ju)$$
$$dp = j \exp(ju)\, du$$
$$dp^{*} = -j \exp(-ju)\, du$$

The transform is therefore:

$$z = \frac{\exp ju}{a \cos u + 1}$$

hyperbool

and this is not a circle, but a hyperbola (fig. 35).

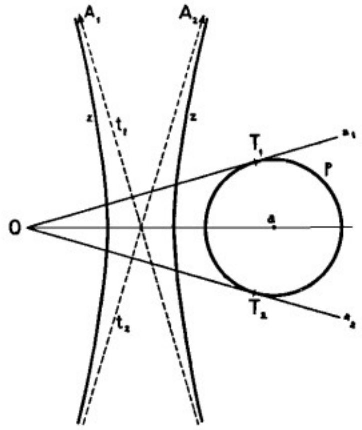

Fig. 35

Any property of the original figure will be found again in the transform, but in the so-called dual form (Poncelet). Dual expressions are point and line; collinearity of points and concurrency of lines; order and class; point on a curve and tangent. As examples of dual theorems we may consider the theorems C of Menelaos and Ceva. Another example is supplied by Pappos' theorems (300 A.D., fig. 36):

Take three points u_1, u_2, u_3 on one line and three points v_1, v_2, v_3 on a second line. The lines u_1v_2 and u_2v_1 intersect in C, the lines u_2v_3 and u_3v_2 in A, the lines u_3v_1 and u_1v_3 in B. The theorem says that A, B and C are collinear. It can be proved with the methods of the preceding sections.

The dual theorem is (fig. 37):

Take three lines u_1, u_2, u_3 through one point U and three lines v_1, v_2, v_3 through a second point V.

The points u_1v_2 and u_2v_1 determine the straight line c, the points u_2v_3 and u_3v_2 determine the line a, the points u_3v_1 and u_1v_3 determine the line b. According to the theorem the lines a, b and c are concurrent.

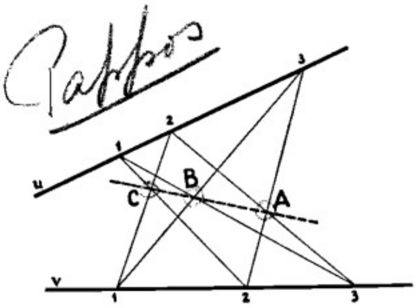

Fig. 36

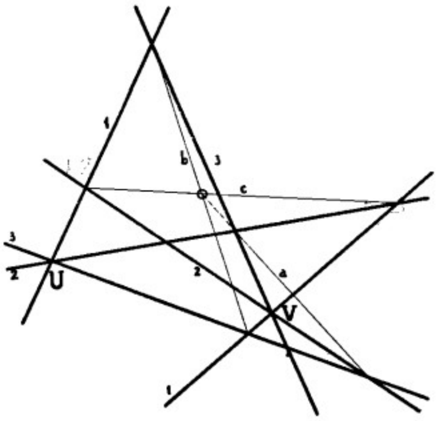

Fig. 37

Although this theorem too can be proved by means of the methods of the previous sections, it is not necessary to do so, once the dual theorem has been proved. More than that, the analytical proof of one of the theorems is identical with that of the second. We have merely to interpret the symbols in the dual way. Finally we may remind that Desargues' theorem is dual with itself.

5. *Projective point assemblages and ray pencils*

The point assemblage on the straight line:

$$z = \frac{z_1 + z_2u}{1+u} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

may be compared with that on the arbitrary other straight line:

$$z' = \frac{z_3 + z_4u}{m+nu} \qquad (m \text{ and } n \text{ are real}). \quad . \quad . \quad . \quad . \quad (2)$$

The latter equation is of a more general character than the former. We remark here, that z_3 and z_4 are no longer points of the straight line; it passes through the points z_3/m $(u=0)$ and z_4/n $(n=\infty)$. We shall show that the transition from (1) to (2) can be accomplished by a combination of parallel translations, rotations and perspective projections.

By way of introduction we may remark that two lines:

$$z_A=\frac{z_1+z_2u}{m+nu} \quad \text{and} \quad z_B=\frac{z_1+z_2u}{p+qu}$$

are perspective with respect to the origin, by which we mean, that the points on the two lines with equal values of u lie on one straight line passing through O. Indeed, the quotient

$$\frac{z_A}{z_B}=\frac{p+qu}{m+nu}$$

is real and that means that z_A and z_B have the same direction.

The construction of the scale of (2)

$$z'=\frac{z_3+z_4u}{m+nu}$$

is shown in fig. 38. We first construct the line z_3+z_4u which is perspective with (2) and has a uniform scale (unit=modulus of z_4). Fig. 39 shows in the same way the construction of (1):

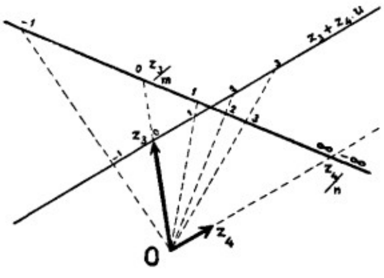

Fig. 38

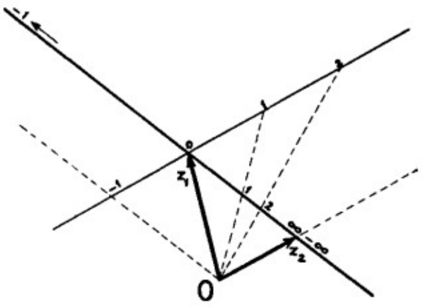

Fig. 39

Now move one of the here mentioned uniform scales until both are parallel and at the same time perspective and the chain of operations leading from (1) to (2) is closed. In general two point assemblages of the type (2) will not be perspective, but we call them *projective*.

Perspective projection leaves certain properties intact; these are called *projective properties*, other properties may be lost in the projection; as opposed to the projective properties they are called *metric properties*.

Examples of projective properties are; incidence of point and line, collinearity of points, concurrency of lines, order and class of a curve, character of tangent, double point, inflection point, cusp. We shall now add one important example more to this collection: the cross ratio of four points *on a straight line* is invariant. To prove this, take four points $u_1 u_2 u_3 u_4$ on the line (2) and calculate the cross ratio. We find:

$$D = \frac{z'(u_1) - z'(u_3)}{z'(u_1) - z'(u_4)} \div \frac{z'(u_2) - z'(u_3)}{z'(u_2) - z'(u_4)} = \frac{u_1 - u_3}{u_1 - u_4} \div \frac{u_2 - u_3}{u_2 - u_4}$$

and this is entirely independent of z_3, z_4, m and n.

The projective character of the cross ratio was already known to Pappos (300 A.D.).

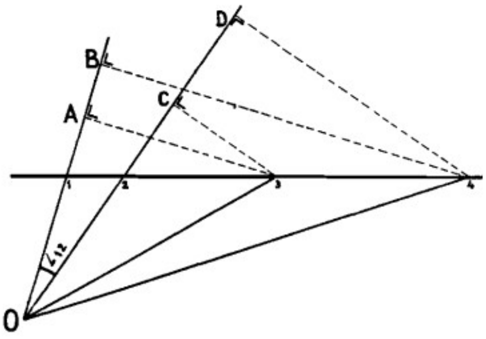

Fig 40

The cross ratio is, so to say, picked up by a projecting pencil of rays and transferred to a second line. We expect that it can be expressed in the angles of the pencil without the help of a secant. Indeed (fig. 40)

$$\frac{z_1-z_3}{z_1-z_4}=\frac{\mathrm{A}\,3}{\mathrm{B}\,4}=\frac{\mathrm{O}\,3\sin\angle 13}{\mathrm{O}\,4\sin\angle 14}$$

In the same way:

$$\frac{z_2-z_3}{z_2-z_4}=\frac{\mathrm{C}\,3}{\mathrm{D}4}=\frac{\mathrm{O}\,3\sin\angle 23}{\mathrm{O}\,4\sin\angle 24}$$

and the cross ratio:

$$\boxed{D=\frac{\sin\angle 13}{\sin\angle 14}\div\frac{\sin\angle 23}{\sin\angle 24},}$$ 2/

where $\angle$ 12 is the angle between the two rays from O to the points z_1 and z_2 etc.; as the choice of O however is entirely arbitrary we have left O out of the expression for D.

A ray pencil projecting a linear point assemblage will be called projective with it. Two pencils projecting projective point assemblages will be called mutually projective; if, more in particular, corresponding rays of these two pencils intersect all on one straight line, we shall call them *perspective*. „Perspective pencils of lines" is the dual conception of „perspective point assemblages".

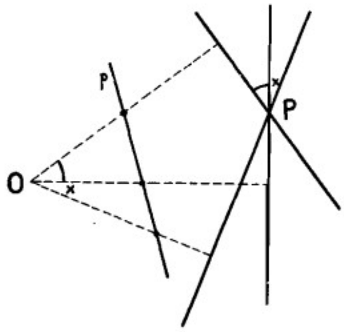

Fig. 41

The poles of a ray pencil lie on the polar of the carrier point of the pencil and as the pole of each member is situated on the perpendicular let down from O on the line concerned, the arguments of the poles differ by the same amounts as the slopes of the lines of the original pencil (fig. 41). We come therefore, to the conclusion that the cross ratio is also invariant with respect to the transformation of pole and polar.

6. *Projective geometry*

As a rule projective properties of figures are much more easily proved by direct projective methods than by analytical calculation.

*crossratio van een [illegible] [illegible] [illegible] in (3) [illegible]

We shall give a number of examples of theorems proved by the former methods.

a. Harmonic point quadruples in the complete quadrangle.

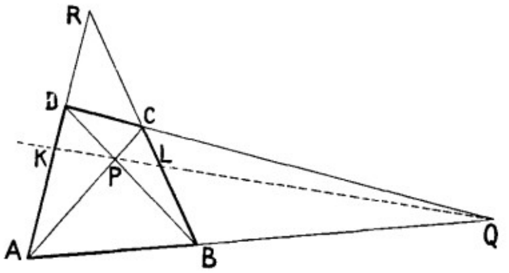

Fig. 42

The complete quadrangle ABCD has six sides and three diagonal points, P, Q and R (fig. 42). Connect P and Q by a line which intersects the two remaining sides in K and L, then the points P, Q, K and L are lying *harmonically* ($D = -1$). The proof is as follows:

$$D(\text{PQKL}) = (\text{projected from A on RB})) = D(\text{CBRL}).$$

D

$$= (\text{projected from D on KQ}) = D(\text{QPKL}).$$

Now P and Q have changed places, but from the definition of the cross ratio it follows that it is inverted by changing two elements. So D must be either $+1$ or -1, and from the figure we see, that in this case it is -1.

b. Desargues' theorem was proved analytically in a previous section. The projective proof is as follows: (fig. 31). The point quadruples $-w\text{A}w\text{C}$ and $-v\text{B}v\text{C}$ are both harmonic ($D = -1$) and as they have one point in common (C) they must be perspective, so that w $-v$ and $-w$ v must intersect on the side AB (in the point u); for the same reason w v and $-w$ v must meet on AB.; q.e.d.

c. Pappos' theorem I follows from the fact, that in fig. 43, where the lines u_2C and u_3B are parallel and so on, defining three points v_1, v_2 and v_3 on the straight line at infinity, A, B and C are collinear for elementary reasons. By a collineation this figure is transformed into fig. 36 of page 42, but the three points A, B and C will remain collinear, and this is what the theorem states.

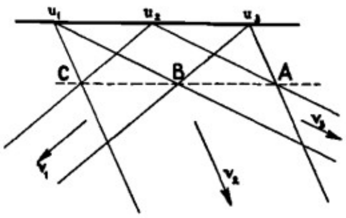

Fig. 43

7. *Involution*

We shall now return to the study of the point assemblage:

$$z = \frac{z_1 + z_2 u}{1 + u}$$

lying on the straight line through the two points z_1 and z_2, and try to find the point which lies harmonically with the point u with respect to z_1 and z_2. This point u' has to be determined from

$$D(z_1 z_2 z(u) z(u')) = -1.$$

or, as $u = 0$ for z_1, $u = \infty$ for z_2, from:

$$\frac{u-0}{u'-0} \div \frac{u-\infty}{u'-\infty} = -1; \text{ hence } u' = -u.$$

The transformation transforming each point into the fourth harmonic with respect to two fixed points, is called the *involution*, its analytical expression is simply:

$$u' = -u.$$

Evidently the relation is reciprocal, the points u and $-u$ form a *pair of the involution*. There are two points which coincide with their partners, namely $u=0$ and $u=\infty$; these are precisely the points z_1 and z_2. Because of this property they are called the *double points* of the involution. The middle of the section $z_1 z_2$ $(u=1)$ is the partner of the infinite point of the line $(u=-1)$.

The harmonic mean of the distances of two partners of the involution to one of the double points is constant;

$$z(u)-z_1=(z_2-z_1)\frac{u}{1+u}; \qquad z(-u)-z_1=(z_1-z_2)\frac{u}{1-u},$$

hence

$$\frac{1}{z(u)-z_1}+\frac{1}{z(-u)-z_1}=\frac{2}{z_2-z_1}=\text{constant}=\frac{1}{f}$$

and as all vectors involved are in the same direction, the same relation holds for the absolute values. With respect to the second double point (z_2) we come to the same result with opposite sign:

$$\frac{1}{z(u)-z_2}+\frac{1}{(z-u)-z_2}=\frac{2}{z_1-z_2}=\text{constant}.$$

Another theorem states that the product of the distances of the partners of a pair to the „*centre*" of the involution, that is the point $\frac{1}{2}(z_1+z_2)$, is constant. Indeed:

$$\left(z(u)-\frac{z_1+z_2}{2}\right)\left(z(-u)-\frac{z_1+z_2}{2}\right)=$$

$$=\left(\frac{z_1+z_2u}{1+u}-\frac{z_1+z_2}{2}\right)\left(\frac{z_1-z_2u}{1-u}-\frac{z_1+z_2}{2}\right)=\left(\frac{z_1-z_2}{2}\right)^2=\text{constant}$$

and as all vectors involved are in the same direction the same holds for the absolute values.

This theorem provides us with a second definition of the involution: The involution transforms a point into its partner in such a way that the product of the distances of the two partners to a fixed centre is constant.

This definition is more general than the original one. If the two partners are situated on the same side of the centre, as was the case up to now, the product is positive, if, however, we choose the points on opposite sides of the center, the product is negative and there exist no real double points of the involution. We distinguish *hyperbolic* (product positive, real double points) from *elliptic* (product negative, imaginary double points) and *parabolic* (product zero, two coinciding double points) involutions. In the case of the parabolic involution one of the points of each pair coincides always with the double point.

Now the point $1/u$ lies symmetrical with the point u with respect to the centre $\frac{1}{2}(z_1 + z_2)$:

$$z(u) - \frac{z_1 + z_2}{2} = \frac{u-1}{u+1} \cdot \frac{z_2 - z_1}{2}; \; z\left(\frac{1}{u}\right) - \frac{z_1 + z_2}{2} = \frac{1-u}{1+u} \cdot \frac{z_2 - z_1}{2};$$

and, as the pairs $(u, -u)$ form a hyperbolic involution, the pairs of points $(u, -1/u)$, form an elliptic involution. The basic points $z_1 (u = +0)$ and $z_2 (u = -\infty)$ are no longer double points, but form a pair. The double points are imaginary $(u = \pm i)$ and with respect to these imaginary double points the harmonic configuration of the pairs is maintained.

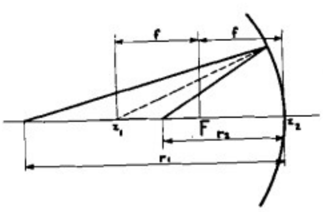

Fig. 44

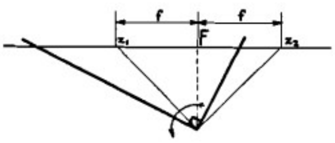

Fig 45

An example of a hyperbolic involution is provided by the pairs of objects and images in the case of reflection of light by a spherical mirror, for which the equation $1/r_1 + 1/r_2 = 1/f$ holds. This case is depicted in fig. 44; the mirror is placed in the point z_2, the centre of the spherical surface is placed in z_1.

An example of an elliptic involution is provided by the pairs of points of insersection of the sides of a right angle with a straight line, if the right angle is rotated round its vertex, fig. 45.

A point involution can be picked up by a projecting ray pencil and projected on a second line. If in particular, as is the case in fig. 45, the rays of each pair of the ray involution are at right angles we speak of an *orthogonal involution.* In projecting an involution on an other line carrier, the double points are perspective, but the centre is not projected as the centre of the new involution; the centre is not a projective but a metric characteristic of the involution.

CHAPTER TWO

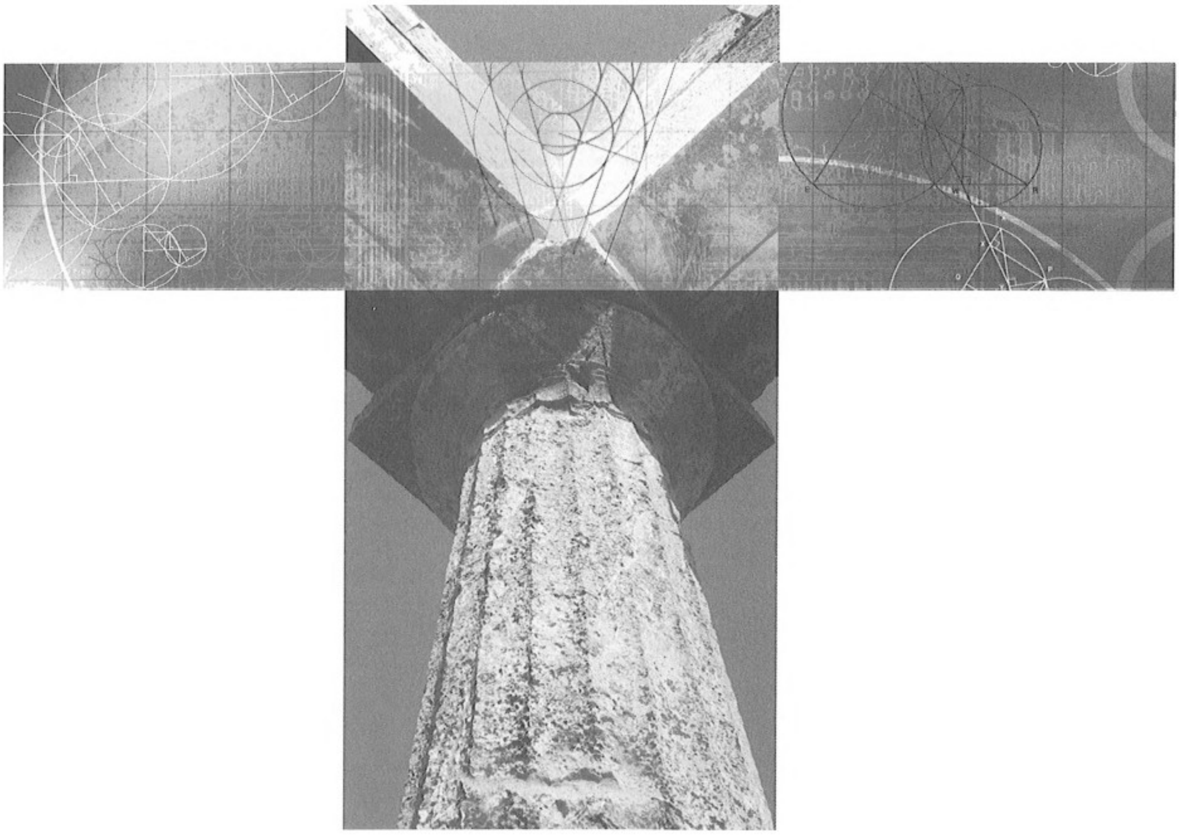

CONCURRENCY OF LINES IN A TRIANGLE

INTRODUCTION

In spite of its importance, the concept of concurrency of lines (i.e., three or more lines containing a common point) usually gets a light treatment in an elementary geometry course because of higher priorities. Acquiring a truly good facility with the concept would require that more theorems be explored than time permits in the first geometry course. Familiar concurrencies such as the medians, angle bisectors, and altitudes of a triangle are mentioned but not often established by proof. Introducing a few new theorems makes the topic of concurrency quite simple and presents a new vista in Euclidean geometry. This chapter begins by demonstrating the importance of establishing concurrency. With the help of an important theorem, first published by Giovanni Ceva in 1678, we present a variety of interesting relationships and theorems. You will soon see how simply some previously difficult theorems can be proved.

Because we want to show the importance of concurrency, let us consider the following problem:

> Two wires are placed in straight lines meeting in an inaccessible region (Figure 2-1). How would you locate the proper placement for a wire that is to *bisect* the angle formed by the two wires without touching the inaccessible region?

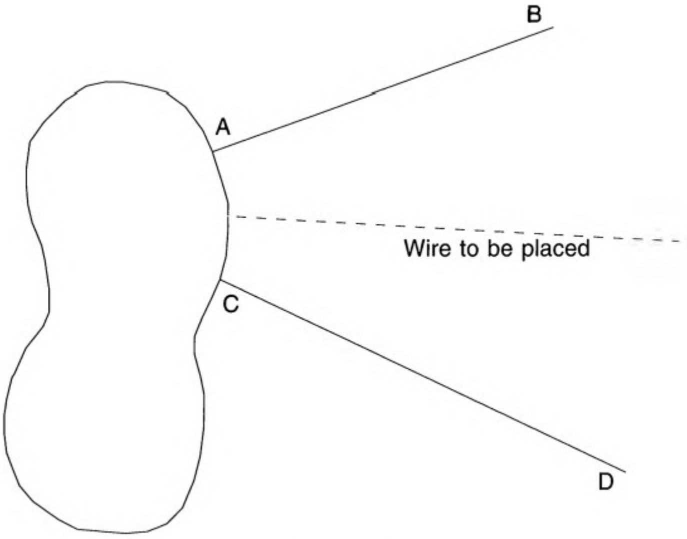

FIGURE 2-1

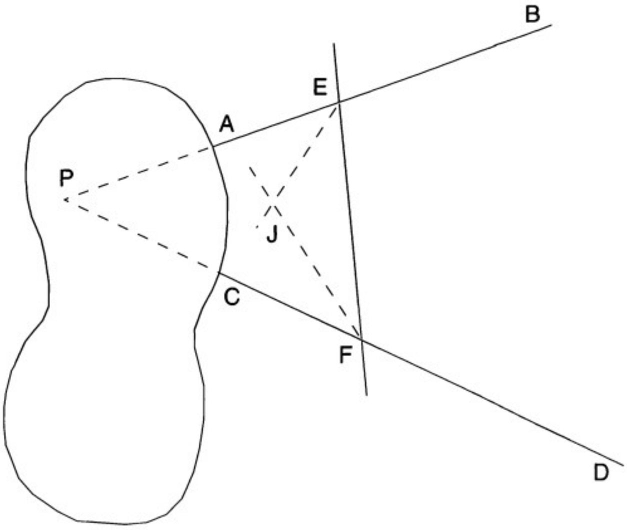

FIGURE 2-2

Although there are many possible methods of solution to this problem, we chose the following solution for a reason that will soon become clear.

Solution Draw any line through $\overleftrightarrow{AB}$ and $\overleftrightarrow{CD}$, intersecting them at points E and F, respectively (see Figure 2-2). Construct the bisectors of $\angle AEF$ and $\angle CFE$, which meet at point J. Suppose $\triangle PEF$ were complete. The bisector of $\angle P$ would then have to contain point J because the angle bisectors of a triangle are concurrent.

Repeat this process for any other line $\overleftrightarrow{GH}$ that meets $\overleftrightarrow{AB}$ and $\overleftrightarrow{CD}$ at points G and H, respectively (Figure 2-3). This time, bisect $\angle AGH$ and $\angle CHG$. These bisectors meet at point K. Once again, we notice that the required angle bisector (that of $\angle P$) must contain point K. Because this required angle bisector *must* contain both J and K, these two points determine our desired line, which is the location of the wire to be installed. ●

This solution relies heavily on the notion that the angle bisectors of a triangle are concurrent. As we have said, the topic of concurrency in a triangle deserves more attention than it usually gets in the elementary geometry course. In a very simple way, we will prove that the angle bisectors of a triangle are concurrent. First, we must establish an extremely useful relationship.

Recall from elementary geometry that there are many "centers" of a triangle. Some examples are:

- *centroid*—the center of gravity of the triangle, determined by the intersection of the medians;
- *orthocenter*—the point of intersection of the altitudes of the triangle;
- *incenter*—the center of the inscribed circle of the triangle, determined by the intersection of the angle bisectors of the triangle;

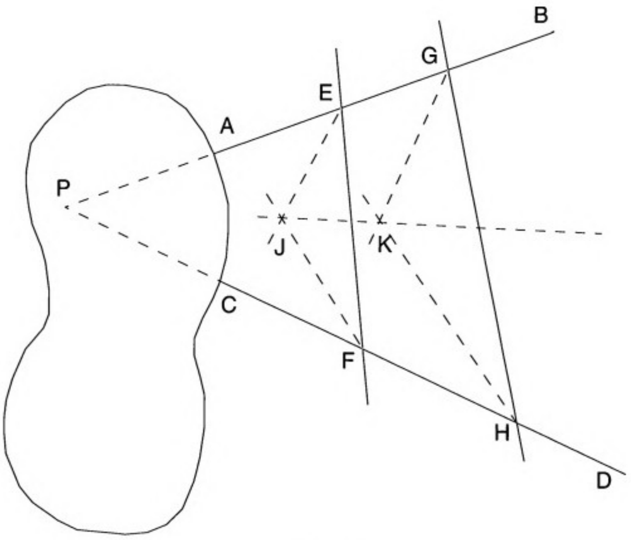

FIGURE 2-3

- *circumcenter*—the center of the circumscribed circle (or circumcircle), determined by the intersection of the perpendicular bisectors of the sides of the triangle.

Numerous applications of these triangle centers are considered in the elementary geometry course. On occasion, students will consider "practical" applications that rely on the concurrency property of these points. We offered one such application at the beginning of this chapter. Yet because the traditional proofs of these concurrency relationships are somewhat cumbersome, the relationships are frequently accepted without proof. With the help of the famous theorem first published* by the Italian mathematician Giovanni Ceva (1647–1734), which bears his name, we will produce simple proofs of the concurrency relationships named previously, as well as many others.

CEVA'S THEOREM

THEOREM 2.1 **(Ceva's Theorem)** The three lines containing the vertices A, B, and C of $\triangle ABC$ and intersecting the opposite sides at points L, M, and N, respectively, are concurrent if and only if $\frac{AN}{NB} \cdot \frac{BL}{LC} \cdot \frac{CM}{MA} = 1$.

---

* *De lineis se invicem secantibus statica constructio* (Milan, 1678).

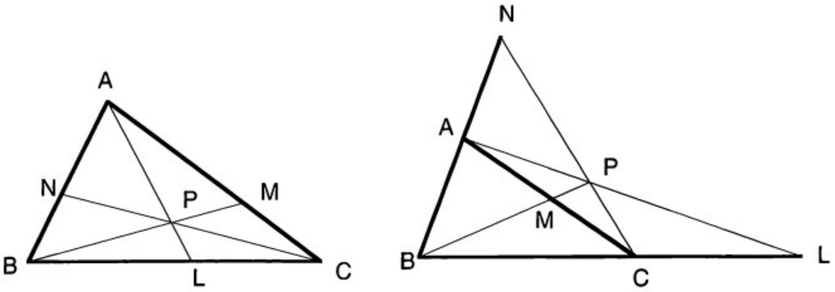

FIGURE 2-4

To prove this theorem, we must note that there are two possible situations in which the three lines drawn from the vertices may intersect the sides and still be concurrent. These situations are pictured in Figure 2-4. It is perhaps easier to follow the proof with the diagram on the left and verify the validity of the statements with the diagram on the right. In any case, the statements made in the proof hold for *both* diagrams.

Ceva's theorem is an equivalence (or biconditional) and therefore requires two proofs (one the converse of the other). We will first prove that *if the three lines containing the vertices of* $\triangle ABC$ *and intersecting the opposite sides at points L, M, and N, respectively, are concurrent, then* $\frac{AN}{NB} \cdot \frac{BL}{LC} \cdot \frac{CM}{MA} = 1$. We offer three proofs. The first (though not the simplest) requires no auxiliary lines.

Proof I In Figure 2-4, $\overline{AL}$, $\overline{BM}$, and $\overline{CN}$ meet at point *P*. Because $\triangle ABL$ and $\triangle ACL$ share the same altitude (i.e., from point *A*):

$$\frac{\text{area } \triangle ABL}{\text{area } \triangle ACL} = \frac{BL}{LC} \tag{I}$$

Similarly:

$$\frac{\text{area } \triangle PBL}{\text{area } \triangle PCL} = \frac{BL}{LC} \tag{II}$$

From (I) and (II):

$$\frac{\text{area } \triangle ABL}{\text{area } \triangle ACL} = \frac{\text{area } \triangle PBL}{\text{area } \triangle PCL}$$

A basic property of proportions $\left(\frac{w}{x} = \frac{y}{z} = \frac{w-y}{x-z}\right)$ provides that:

$$\frac{BL}{LC} = \frac{\text{area } \triangle ABL - \text{area } \triangle PBL}{\text{area } \triangle ACL - \text{area } \triangle PCL} = \frac{\text{area } \triangle ABP}{\text{area } \triangle ACP} \tag{III}$$

We now repeat the process, using $\overline{BM}$ instead of $\overline{AL}$:

$$\frac{CM}{MA} = \frac{\text{area } \triangle BMC}{\text{area } \triangle BMA} = \frac{\text{area } \triangle PMC}{\text{area } \triangle PMA}$$

It follows that:

$$\frac{CM}{MA} = \frac{\text{area } \triangle BMC - \text{area } \triangle PMC}{\text{area } \triangle BMA - \text{area } \triangle PMA} = \frac{\text{area } \triangle BCP}{\text{area } \triangle BAP} \qquad \text{(IV)}$$

Once again we repeat the process, this time using $\overline{CN}$ instead of $\overline{AL}$:

$$\frac{AN}{NB} = \frac{\text{area } \triangle ACN}{\text{area } \triangle BCN} = \frac{\text{area } \triangle APN}{\text{area } \triangle BPN}$$

This gives us:

$$\frac{AN}{NB} = \frac{\text{area } \triangle ACN - \text{area } \triangle APN}{\text{area } \triangle BCN - \text{area } \triangle BPN} = \frac{\text{area } \triangle ACP}{\text{area } \triangle BCP} \qquad \text{(V)}$$

We now simply multiply (III), (IV), and (V) to get the desired result:

$$\frac{BL}{LC} \cdot \frac{CM}{MA} \cdot \frac{AN}{NB} = \frac{\text{area } \triangle ABP}{\text{area } \triangle ACP} \cdot \frac{\text{area } \triangle BCP}{\text{area } \triangle BAP} \cdot \frac{\text{area } \triangle ACP}{\text{area } \triangle BCP} = 1 \; \bullet$$

By introducing an auxiliary line, we can produce a simpler proof.

Proof II Consider Figure 2-4, but add a line containing point A and parallel to $\overline{BC}$ that intersects $\overleftrightarrow{CP}$ at point S and $\overleftrightarrow{BP}$ at point R (see Figure 2-5). The parallel lines enable us to establish the following pairs of similar triangles:

$$\triangle AMR \sim \triangle CMB \Rightarrow \frac{AM}{MC} = \frac{AR}{CB} \qquad \text{(I)}$$

$$\triangle BNC \sim \triangle ANS \Rightarrow \frac{BN}{NA} = \frac{CB}{SA} \qquad \text{(II)}$$

$$\triangle CLP \sim \triangle SAP \Rightarrow \frac{CL}{SA} = \frac{LP}{AP} \qquad \text{(III)}$$

$$\triangle BLP \sim \triangle RAP \Rightarrow \frac{BL}{RA} = \frac{LP}{AP} \qquad \text{(IV)}$$

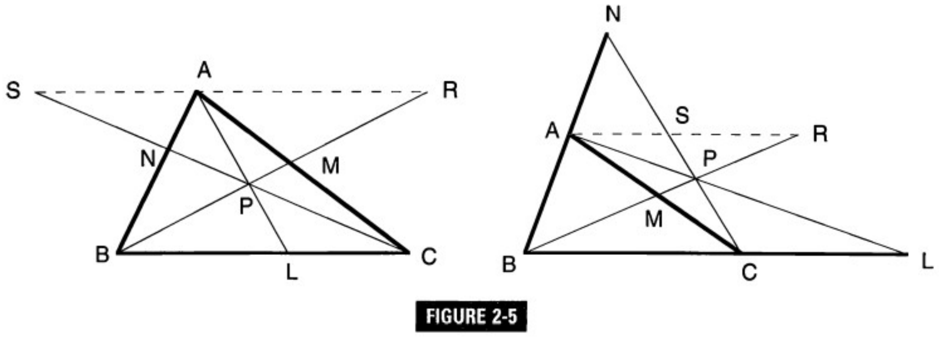

FIGURE 2-5

From (III) and (IV) we get:

$$\frac{CL}{SA} = \frac{BL}{RA}$$

This can be rewritten as:

$$\frac{CL}{BL} = \frac{SA}{RA} \qquad \text{(V)}$$

By multiplying (I), (II), and (V), we obtain our desired result:

$$\frac{AM}{MC} \cdot \frac{BN}{NA} \cdot \frac{CL}{BL} = \frac{AR}{CB} \cdot \frac{CB}{SA} \cdot \frac{SA}{RA} = 1$$

We rearrange the terms and invert the fractions to get:

$$\frac{AN}{NB} \cdot \frac{BL}{LC} \cdot \frac{CM}{MA} = 1 \text{ (the same as the conclusion of Proof I)} \; \bullet$$

By adding two auxiliary lines to the diagrams in Figure 2-4, we are able to produce another proof, again using the properties of similar triangles.

Proof III We begin with the diagrams shown in Figure 2-4 but add two lines to each diagram. We draw a line through point A and a line through point C each parallel to $\overline{BP}$ and intersecting $\overleftrightarrow{CP}$ and $\overleftrightarrow{AP}$ at points S and R, respectively (see Figure 2-6).

$$\triangle ASN \sim \triangle BPN \Rightarrow \frac{AN}{NB} = \frac{AS}{BP} \qquad \text{(I)}$$

$$\triangle BPL \sim \triangle CRL \Rightarrow \frac{BL}{LC} = \frac{BP}{CR} \qquad \text{(II)}$$

$$\triangle PAM \sim \triangle RAC \Rightarrow \frac{CA}{MA} = \frac{RC}{PM}$$

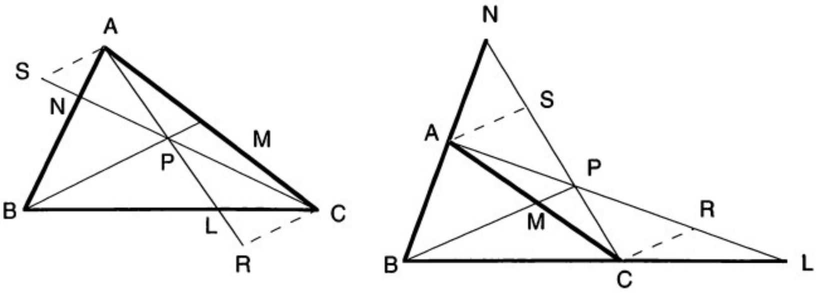

FIGURE 2-6

This can be written as:

$$CA = \frac{(RC)(MA)}{PM} \tag{III}$$

$$\triangle PCM \sim \triangle SCA \Rightarrow \frac{CM}{CA} = \frac{PM}{AS}$$

This can be written as:

$$CA = \frac{(AS)(CM)}{PM} \tag{IV}$$

From (III) and (IV):

$$\frac{(RC)(MA)}{PM} = \frac{(AS)(CM)}{PM}$$

This can be written as:

$$\frac{CM}{MA} = \frac{RC}{AS} \tag{V}$$

To obtain our desired result, we multiply (I), (II), and (V):

$$\frac{AN}{NB} \cdot \frac{BL}{LC} \cdot \frac{CM}{MA} = \frac{AS}{BP} \cdot \frac{BP}{CR} \cdot \frac{RC}{AS} = 1$$ ●

To complete the proof of Ceva's theorem, we must now prove the converse of the implication proved above; that is, we will now prove that *if the lines containing the vertices of $\triangle ABC$ intersect the opposite sides in points L, M, and N, respectively, so that* $\frac{AN}{NB} \cdot \frac{BL}{LC} \cdot \frac{CM}{MA} = 1$, *then these lines,* $\overleftrightarrow{AL}$, $\overleftrightarrow{BM}$, *and* $\overleftrightarrow{CN}$, *are concurrent.*

Proof Suppose $\overleftrightarrow{BM}$ and $\overleftrightarrow{AL}$ intersect at point P. Draw $\overleftrightarrow{PC}$ and call its intersection with $\overleftrightarrow{AB}$ N'. Now that $\overleftrightarrow{AL}$, $\overleftrightarrow{BM}$, and $\overleftrightarrow{CN'}$ are concurrent, we can use the part of Ceva's theorem proved earlier to state the following:

$$\frac{AN'}{N'B} \cdot \frac{BL}{LC} \cdot \frac{CM}{MA} = 1$$

Our hypothesis stated that:

$$\frac{AN}{NB} \cdot \frac{BL}{LC} \cdot \frac{CM}{MA} = 1$$

Therefore $\frac{AN'}{N'B} = \frac{AN}{NB}$, so N and N' must coincide, proving concurrency. ●

APPLICATIONS OF CEVA'S THEOREM

One of the best ways to show the usefulness of Ceva's theorem is to apply it to the proof of the concurrency of the various line segments encountered in elementary geometry. The simplest application is to prove the concurrency of the medians of a triangle. To best appreciate the "power" of Ceva's theorem, you should first recall the conventional method of proving the medians of a triangle concurrent. Suffice it to say that it is quite long and complex. Compared to this rather cumbersome proof, the method we use here should provoke some excitement about Ceva's theorem.

Application 1 Prove that the medians of a triangle are concurrent. ●

Proof In $\triangle ABC$, $\overline{AL}$, $\overline{BM}$, and $\overline{CN}$ are medians (see Figure 2-7). Therefore $AN = NB$, $BL = LC$, and $CM = MA$. Multiplying these equalities gives us:

$$(AN)(BL)(CM) = (NB)(LC)(MA) \quad \text{or} \quad \frac{AN}{NB} \cdot \frac{BL}{LC} \cdot \frac{CM}{MA} = 1$$

INTERACTIVE 2-7

Drag vertices *A*, *B*, and *C* to change the shape of the triangle and see that the medians always meet at one point.

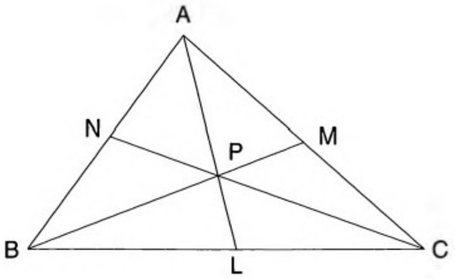

FIGURE 2-7

Thus by Ceva's theorem, $\overline{AL}$, $\overline{BM}$, and $\overline{CN}$ are concurrent. ●

Again, it would be advisable to compare the conventional proof (that presented in the context of elementary geometry) for the concurrency of the altitudes of a triangle to the following proof, which uses Ceva's theorem.

Application 2 Prove that the altitudes of a triangle are concurrent. ●

Proof In $\triangle ABC$, $\overline{AL}$, $\overline{BM}$, and $\overline{CN}$ are altitudes. You can follow this proof for both diagrams of Figure 2-8 because the same proof holds true for both an acute and an obtuse triangle.

$$\triangle ANC \sim \triangle AMB \Rightarrow \frac{AN}{MA} = \frac{AC}{AB} \qquad \text{(I)}$$

$$\triangle BLA \sim \triangle BNC \Rightarrow \frac{BL}{NB} = \frac{AB}{BC} \qquad \text{(II)}$$

$$\triangle CMB \sim \triangle CLA \Rightarrow \frac{CM}{LC} = \frac{BC}{AC} \qquad \text{(III)}$$

%/ AE 1 G // E 1

INTERACTIVE 2-8

Drag vertices *A*, *B*, and *C* to change the shape of the triangle and see that the altitudes always meet at one point.

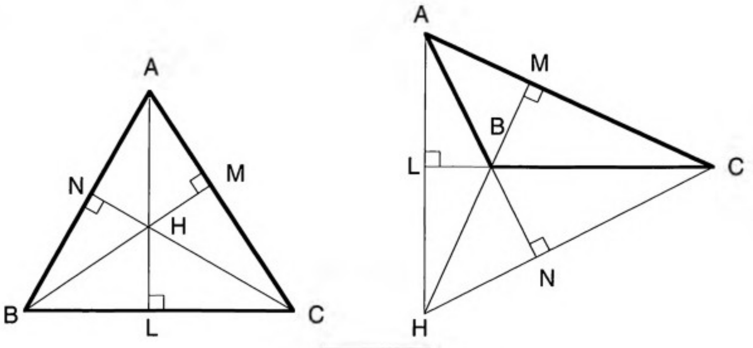

FIGURE 2-8

Multiplying (I), (II), and (III) gives us:

$$\frac{AN}{MA}\cdot\frac{BL}{NB}\cdot\frac{CM}{LC}=\frac{AC}{AB}\cdot\frac{AB}{BC}\cdot\frac{BC}{AC}=1$$

Therefore the altitudes are concurrent (by Ceva's theorem). ●

The proof that the three angle bisectors of a triangle are concurrent is left as an exercise. The following proof should be helpful in working that exercise.

Application 3 Prove that the bisector of any interior angle of a nonisosceles triangle and the bisectors of the two exterior angles at the other vertices are concurrent. ●

Proof In $\triangle ABC$, $\overleftrightarrow{AL}$ bisects $\angle BAC$ and meets $\overleftrightarrow{BC}$ at point L, $\overleftrightarrow{BM}$ bisects exterior $\angle ABE$ and meets $\overleftrightarrow{AC}$ at point M, and $\overleftrightarrow{CN}$ bisects exterior $\angle ACF$ and meets $\overleftrightarrow{AB}$ at point N (see Figure 2-9).

Because the bisector ($\overleftrightarrow{AL}$) of an interior angle of a triangle partitions the opposite side proportionally to the remaining two sides of the triangle:

$$\frac{BL}{LC}=\frac{AB}{AC} \tag{I}$$

An exterior angle bisector partitions the side that it intersects proportionally to the remaining sides of the triangle. This property produces the following proportions:

$$\text{For } \overleftrightarrow{BM}:\quad \frac{CM}{MA}=\frac{BC}{AB} \tag{II}$$

$$\text{For } \overleftrightarrow{CN}:\quad \frac{AN}{NB}=\frac{AC}{BC} \tag{III}$$

By multiplying (I), (II), and (III), we get:

$$\frac{BL}{LC}\cdot\frac{CM}{MA}\cdot\frac{AN}{NB}=\frac{AB}{AC}\cdot\frac{BC}{AB}\cdot\frac{AC}{BC}=1$$

INTERACTIVE 2-9

Drag vertices *A*, *B*, and *C* to change the shape of the triangle and see that the indicated bisectors of the angles (interior and exterior) always meet at one point.

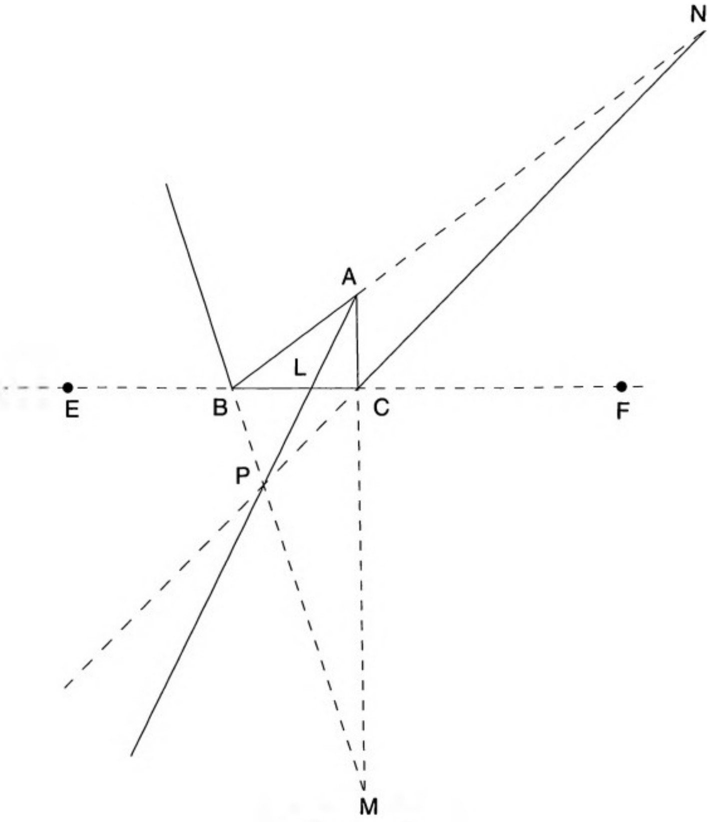

FIGURE 2-9

By Ceva's theorem, we may conclude that $\overleftrightarrow{AL}$, $\overleftrightarrow{BM}$, and $\overleftrightarrow{CN}$ are concurrent. ●

Sometimes the question of concurrency is a bit disguised, as in the following application.

Application 4 In $\triangle ABC$, $\overline{PQ} \parallel \overline{BC}$ and intersects $\overline{AB}$ and $\overline{AC}$ at points *P* and *Q*, respectively (see Figure 2-10). Prove that $\overline{PC}$ and $\overline{QB}$ intersect at a point on median $\overline{AM}$. ●

INTERACTIVE 2-10

Drag vertices *A*, *B*, and *C* to change the shape of the triangle; drag point *P* on $\overline{AB}$ and see that $\overline{PC}$ and $\overline{QB}$ always meet at a point on $\overline{AM}$.

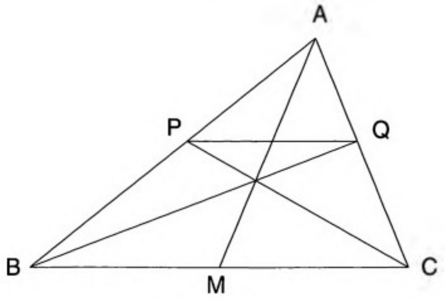

FIGURE 2-10

Proof Because $\overline{PQ} \parallel \overline{BC}$:

$$\frac{AP}{PB} = \frac{AQ}{QC} \quad \text{or} \quad \frac{AP}{PB} \cdot \frac{QC}{AQ} = 1 \qquad \text{(I)}$$

Because $\overline{AM}$ is a median, $BM = MC$. Therefore:

$$\frac{BM}{MC} = 1 \qquad \text{(II)}$$

By multiplying (I) and (II), we get:

$$\frac{AP}{PB} \cdot \frac{QC}{AQ} \cdot \frac{BM}{MC} = 1$$

Thus, by Ceva's theorem, $\overline{AM}$, $\overline{QB}$, and $\overline{PC}$ are concurrent, or $\overline{QB}$ and $\overline{PC}$ intersect at a point on $\overline{AM}$. ●

Up to this point, all our applications have been used to prove concurrency. The following application demonstrates a somewhat different use of Ceva's theorem.

Application 5 In $\triangle ABC$, where $\overline{CD}$ is the altitude to $\overline{AB}$ and P is any point on $\overline{DC}$, $\overline{AP}$ intersects $\overline{CB}$ at point Q and $\overline{BP}$ intersects $\overline{CA}$ at point R (see Figure 2-11). Prove that $\angle RDC \cong \angle QDC$. ●

Proof Let $\overrightarrow{DR}$ and $\overrightarrow{DQ}$ intersect the line containing C and parallel to $\overline{AB}$, at points G and H, respectively.

$$\triangle CGR \sim \triangle ADR \Rightarrow \frac{CR}{RA} = \frac{GC}{AD} \qquad \text{(I)}$$

$$\triangle BDQ \sim \triangle CHQ \Rightarrow \frac{BQ}{QC} = \frac{DB}{CH} \qquad \text{(II)}$$

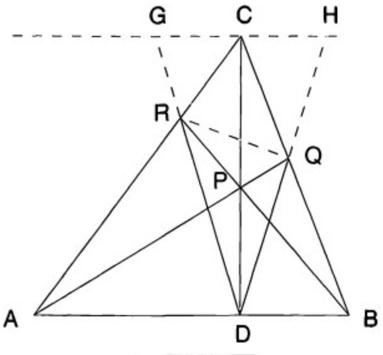

FIGURE 2-11

We now apply Ceva's theorem to $\triangle ABC$ to get:

$$\frac{CR}{RA}\cdot\frac{AD}{DB}\cdot\frac{BQ}{QC}=1 \tag{III}$$

Substituting (I) and (II) into (III) gives us:

$$\frac{GC}{AD}\cdot\frac{AD}{DB}\cdot\frac{DB}{CH}=1 \quad\text{or}\quad \frac{GC}{CH}=1$$

This implies that $GC = CH$. Thus $\overline{CD}$ is the perpendicular bisector of $\overline{GH}$. Hence $\triangle GCD \cong \triangle HCD$, and therefore $\angle RDC \cong \angle QDC$. ●

From the preceding applications, we have seen how Ceva's theorem easily enables us to prove theorems whose proofs would otherwise be quite complex. Ceva's theorem again demonstrates its usefulness in assisting us in proving an interesting point of concurrency in a triangle known as the *Gergonne point.*

THE GERGONNE POINT

A fascinating point of concurrency in a triangle was first established by French mathematician Joseph-Diaz Gergonne (1771–1859). Gergonne reserved a distinct place in the history of mathematics as the initiator (in 1810) of the first purely mathematical journal, *Annales des mathématiques pures et appliqués.* The journal appeared monthly until 1832 and was known as *Annales del Gergonne.* During the time of its publication, Gergonne published about two hundred papers, mostly on geometry. Gergonne's *Annales* played an important role in the establishment of projective and algebraic geometry by giving some of the greatest minds of the times an opportunity to share information. Here we consider a rather simple theorem established by Gergonne that exhibits concurrency and is easily proved using Ceva's theorem.

THEOREM 2.2 The lines containing a vertex of a triangle and the point of tangency of the opposite side with the inscribed circle are concurrent. (This point of concurrency is known as the **Gergonne point** of the triangle.)

Proof Circle O is tangent to sides $\overline{AB}$, $\overline{AC}$, and $\overline{BC}$ of $\triangle ABC$ at points N, M, and L, respectively (see Figure 2-12). It follows that $AN = AM$, $BL = BN$, and $CM = CL$. Each of these equalities can be written as:

$$\frac{AN}{AM}=1 \quad \frac{BL}{BN}=1 \quad \frac{CM}{CL}=1$$

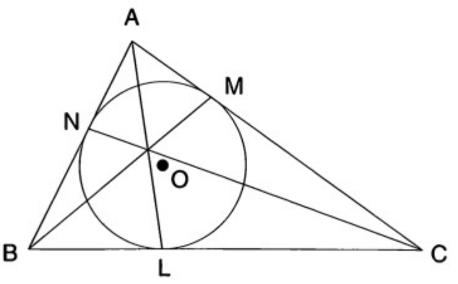

FIGURE 2-12

By multiplying these three fractions, we get:

$$\frac{AN}{AM} \cdot \frac{BL}{BN} \cdot \frac{CM}{CL} = 1$$

Therefore:

$$\frac{AN}{BN} \cdot \frac{BL}{CL} \cdot \frac{CM}{AM} = 1$$

By Ceva's theorem, this equality implies that $\overline{AL}$, $\overline{BM}$, and $\overline{CN}$ are concurrent. The point of concurrency is the Gergonne point of $\triangle ABC$. ●

EXERCISES

1. Prove that the angle bisectors of a triangle are concurrent.

2. If point P is situated on $\overline{BC}$ so that $AB + BP = AC + CP$, point Q is situated on $\overline{AC}$ so that $BC + CQ = AB + AQ$, and point R is situated on $\overline{AB}$ so that $BC + BR = AC + AR$, prove that $\overline{AP}$, $\overline{BQ}$, and $\overline{CR}$ are concurrent (see Figure 2-13). (This point of concurrency is known as the *Nagel point** of $\triangle ABC$.)

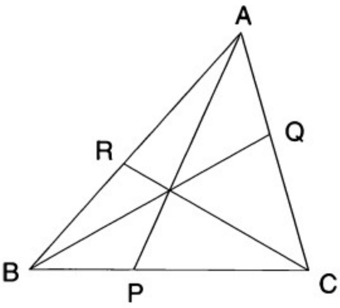

FIGURE 2-13

---

* Discovered by C. H. Nagel (1803–1882), the point can also be described as the intersection of the lines from the vertices of a triangle to the points of tangency of the opposite escribed circles.

3. $\triangle ABC$ cuts a circle at points E, E', D, D', F, and F', as in Figure 2-14. Prove that if $\overline{AD}$, $\overline{BF}$, and $\overline{CE}$ are concurrent, then $\overline{AD'}$, $\overline{BF'}$, and $\overline{CE'}$ are also concurrent.

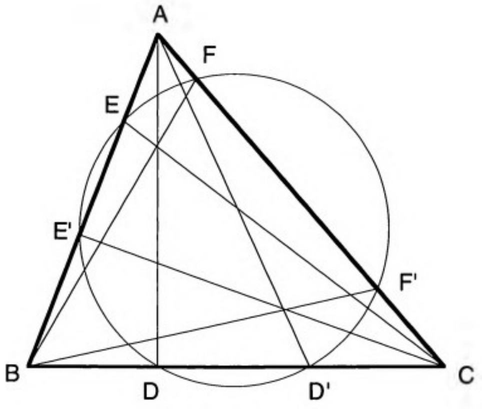

FIGURE 2-14

4. In $\triangle ABC$ (Figure 2-15), $\overline{AL}$, $\overline{BM}$, and $\overline{CN}$ are concurrent at point P. Points R, S, and T are chosen on $\overline{BC}$, $\overline{AC}$, and $\overline{AB}$, respectively, so that $\overline{NR} \parallel \overline{AC}$, $\overline{LS} \parallel \overline{AB}$, and $\overline{MT} \parallel \overline{BC}$. Prove that $\overline{AR}$, $\overline{BS}$, and $\overline{CT}$ are concurrent (at point Q).

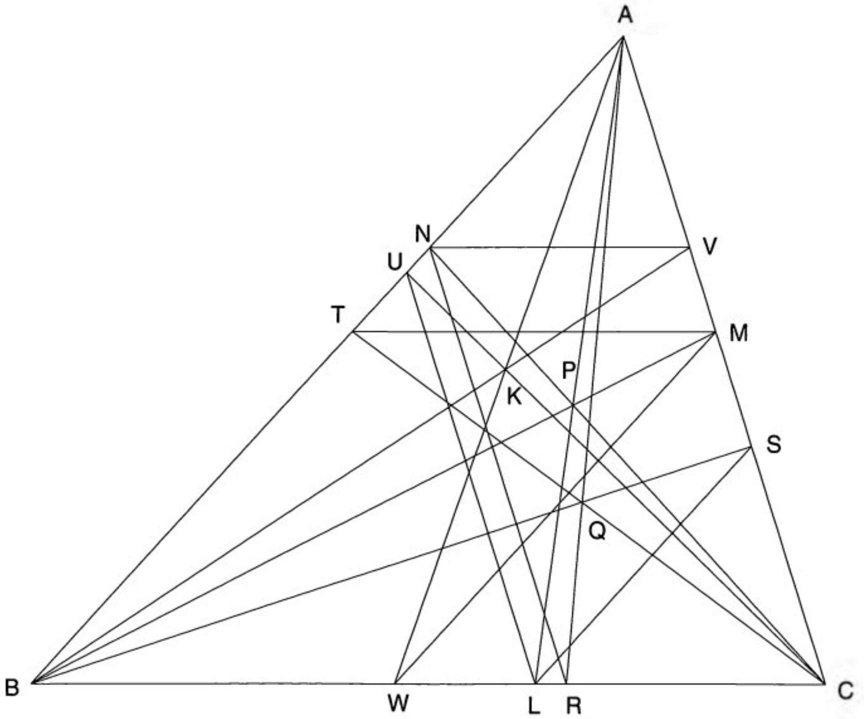

FIGURE 2-15

5. In $\triangle ABC$ (Figure 2-15), $\overline{AL}$, $\overline{BM}$, and $\overline{CN}$ are concurrent at point P. Points U, V, and W are chosen on $\overline{AB}$, $\overline{AC}$, and $\overline{BC}$, respectively, so that $\overline{LU} \parallel \overline{AC}$, $\overline{NV} \parallel \overline{BC}$, and $\overline{MW} \parallel \overline{AB}$. Prove that $\overline{AW}$, $\overline{BV}$, and $\overline{CU}$ are concurrent (at point K).

6. In $\triangle ABC$ (Figure 2-16), $\overline{AL}$, $\overline{BM}$, and $\overline{CN}$ are concurrent at point K and L, M, and N are points on $\overline{BC}$, $\overline{AC}$, and $\overline{AB}$, respectively. Points P, R, and Q are respective midpoints of $\overline{AL}$, $\overline{CN}$, and $\overline{BM}$. Prove that $\overline{DP}$, $\overline{EQ}$, and $\overline{FR}$ are concurrent if D, E, and F are respective midpoints of $\overline{BC}$, $\overline{AC}$, and $\overline{AB}$.

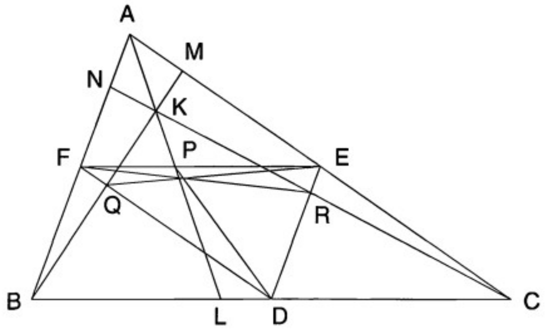

FIGURE 2-16

7. In $\triangle ABC$ (Figure 2-17), $\overline{AL}$, $\overline{BM}$, and $\overline{CN}$ are concurrent at point P. Points S, Q, and R are midpoints of $\overline{MN}$, $\overline{ML}$, and $\overline{NL}$, respectively. Prove that $\overrightarrow{AS}$, $\overrightarrow{BR}$, and $\overrightarrow{CQ}$ are concurrent.

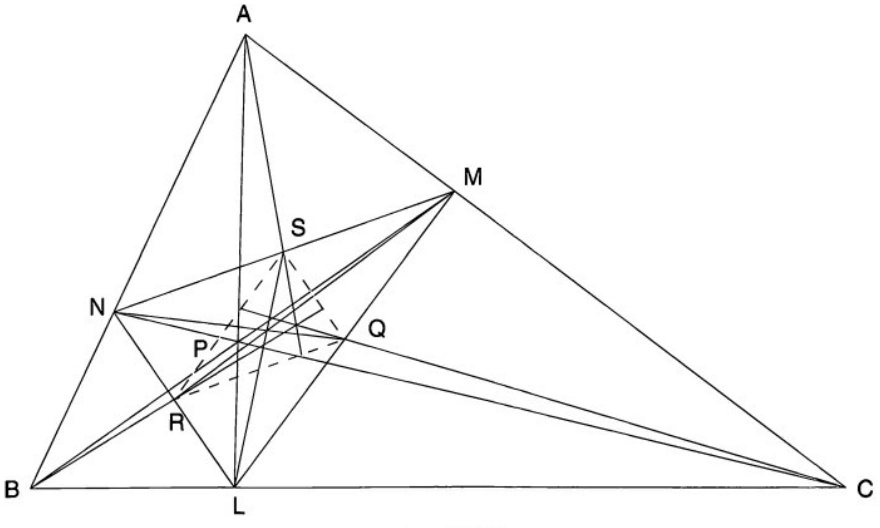

FIGURE 2-17

8. In $\triangle ABC$ (Figure 2-17), $\overline{AL}$, $\overline{BM}$, and $\overline{CN}$ are concurrent at point P. Points S, Q, and R are points on $\overline{MN}$, $\overline{ML}$, and $\overline{NL}$, respectively. If $\overline{LS}$, $\overline{MR}$, and $\overline{NQ}$ are concurrent, prove that $\overrightarrow{AS}$, $\overrightarrow{BR}$, and $\overrightarrow{CQ}$ are also concurrent.
9. Circles P, O, and Q are escribed circles of $\triangle ABC$, with the points of tangency indicated in Figure 2-18. Prove that $\overline{AD}$, $\overline{BE}$, and $\overline{CF}$ are concurrent.

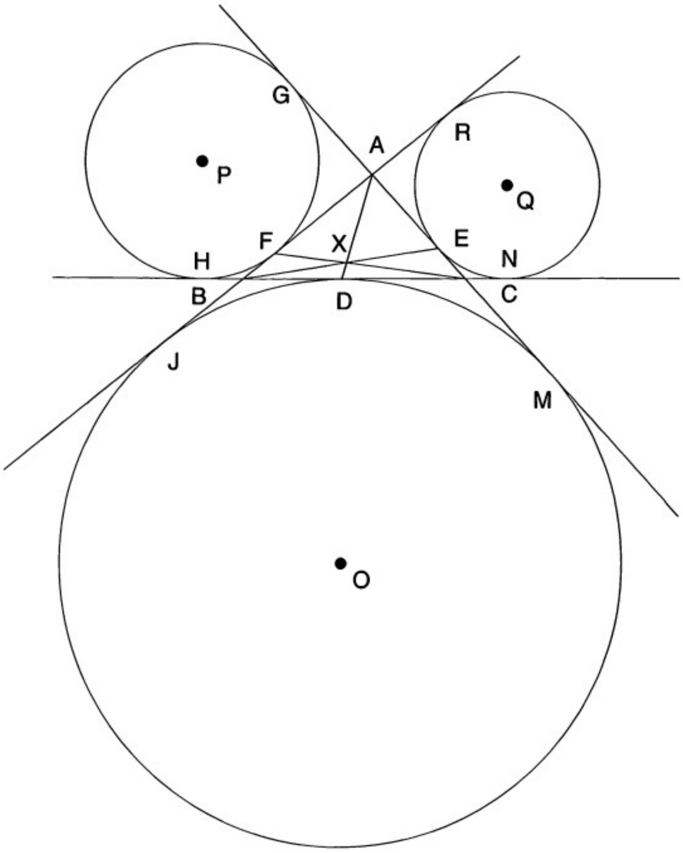

FIGURE 2-18

10. Given three circles, nonintersecting, mutually external, and with distinct radii, connect the intersection of internal common tangents of each pair of circles with the center of the other circle, as in Figure 2-19. Prove that the three resulting line segments are concurrent.

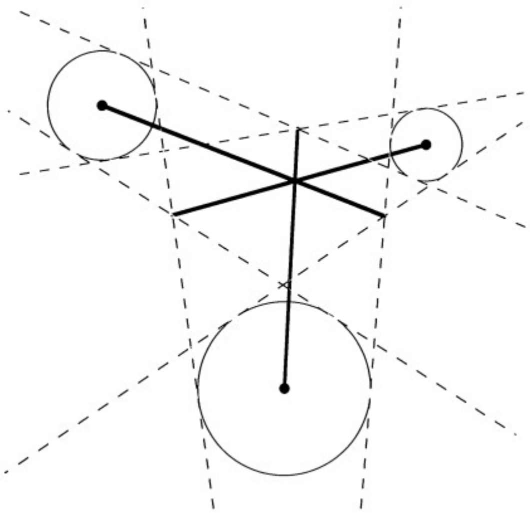

FIGURE 2-19

CHAPTER
THREE

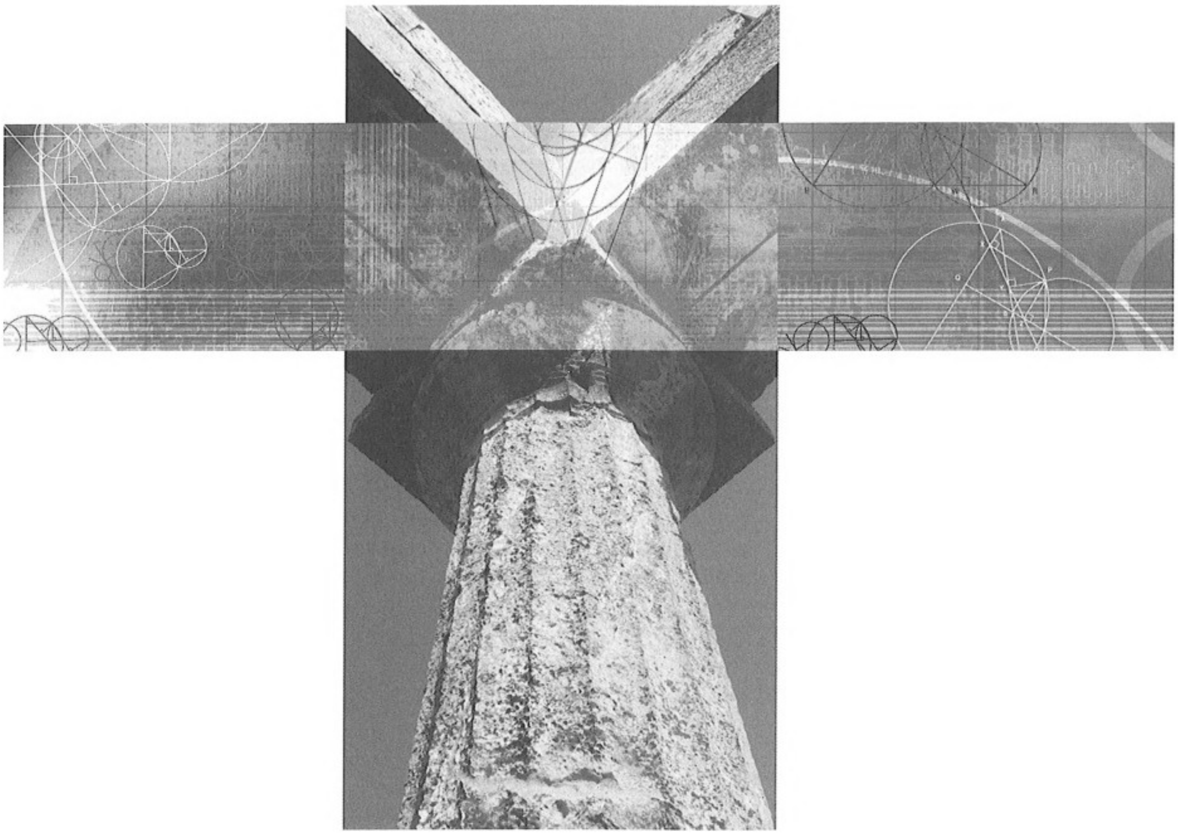

COLLINEARITY OF POINTS

DUALITY

Many statements in geometry involve relationships between points and lines. In a statement concerning points and lines in a plane, when the word *point* is replaced by the word *line* and the word *line* is replaced by the word *point* each time these words are used in the statement, the newly formed statement is said to be the *dual* of the original statement. Occasionally, other modifications may need to be made in order to preserve proper sentence structure. This principle of duality was discovered by Charles Julien Brianchon (1785–1864) while using this relationship on a theorem by Blaise Pascal. We will visit these theorems later in this chapter. However, the transition from Chapter 2 to Chapter 3 follows the principle of duality because concurrency of lines is the dual of collinearity of points. The primary focus of this chapter is collinearity.

Let us first familiarize ourselves with the principle of duality. Consider the following examples of dual statements:

Statement	**Dual Statement**
1. Two distinct points determine a unique line.	1. Two distinct lines determine a unique point.
2. Any point contains an infinite number of lines.	2. Any line contains an infinite number of points.
3. Only one triangle is determined by three noncollinear points.	3. Only one trilateral is determined by three nonconcurrent lines.

This last example of duality demonstrates that related words also need to be changed when forming the dual of a statement. Specifically, note that *collinear* and *concurrent* are dual words, as are *triangle* and *trilateral.*

Recall Ceva's theorem (see Figure 3-1):

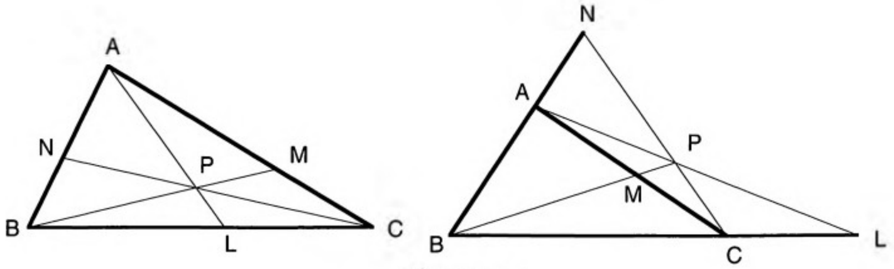

FIGURE 3-1

The three lines containing the vertices A, B, and C of $\triangle ABC$ and intersecting the opposite sides at points L, M, and N, respectively, are concurrent if and only if $\frac{AN}{NB}\cdot\frac{BL}{LC}\cdot\frac{CM}{MA}=1$.

For the most part, the dual of a postulate is also a postulate, and the dual of a definition is itself a definition. Thus, if a statement is a theorem, its dual is likely to be a theorem as well.* In any case, we would at least have a statement that would be a good candidate to be a theorem. A valid proof would be needed to establish the statement as a theorem.

This is precisely what we will now investigate. With our knowledge of duality, we will form the dual statement of Ceva's theorem. Actually, it was the rediscovery of Menelaus of Alexandria's famous but forgotten theorem,† which we will discuss in the next section, that led Giovanni Ceva in the first book of his *De lineis rectis se invicem secantibus statica constructio* (Milan, 1678) to produce his theorem by the principle of duality. Note the duality relationship between the two theorems.

The three points P, Q, and R on the sides $\overleftrightarrow{AC}$, $\overleftrightarrow{AB}$, and $\overleftrightarrow{BC}$, respectively, of $\triangle ABC$ (see Figure 3-2) are collinear if and only if $\frac{AQ}{QB}\cdot\frac{BR}{RC}\cdot\frac{CP}{PA}=-1$.‡ This statement is in fact a theorem, known as *Menelaus's theorem,* that is the subject of the next section.

INTERACTIVE 3-2

Drag vertices *A*, *B*, and *C* to change the shape of the triangle; drag points *P* and *Q* and see that Menelaus's theorem is always true.

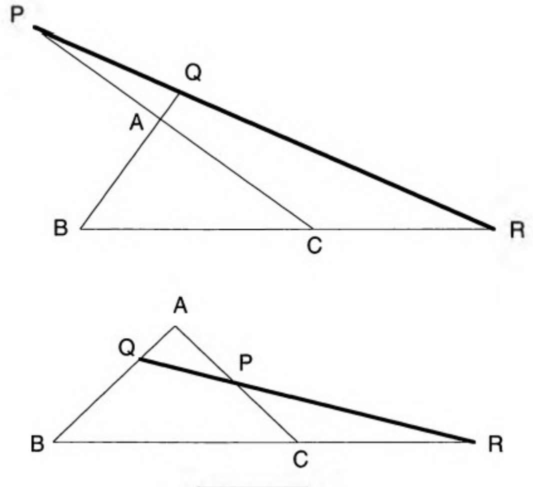

FIGURE 3-2

---

* In a geometric system exclusively based on postulates and definitions whose respective duals are all true, the dual of every theorem is also a theorem. This claim is easily justified by realizing that the proof of the dual of a theorem can be produced by simply replacing each statement of the proof of the original theorem by its dual.

† During the Dark Ages, much of classical Greek mathematics was lost and forgotten.

‡ The reason for the negative sign is explained in the proof of this theorem.

MENELAUS'S THEOREM

Menelaus of Alexandria, in about A.D. 100 in a work titled *Sphaerica,* produced the well-known plane version of the theorem that we will present here. He used the plane version to develop the spherical analogue,* which was the purpose of his treatise. As we have mentioned, this theorem, which today bears Menelaus's name, did not become popular until it was rediscovered by Giovanni Ceva as a part of his work in 1678.

THEOREM 3.1 **(Menelaus's theorem)** The three points P, Q, and R on the sides $\overleftrightarrow{AC}$, $\overleftrightarrow{AB}$, and $\overleftrightarrow{BC}$, respectively, of $\triangle ABC$ are collinear if and only if $\frac{AQ}{QB} \cdot \frac{BR}{RC} \cdot \frac{CP}{PA} = -1$.

Like Ceva's theorem, Menelaus's theorem is an equivalence and therefore requires proofs for each of the two statements (converses of each other) that comprise the entire theorem. We will first prove that *if the three points P, Q, and R on the sides $\overleftrightarrow{AC}$, $\overleftrightarrow{AB}$, and $\overleftrightarrow{BC}$,, respectively, of $\triangle ABC$ are collinear, then* $\frac{AQ}{QB} \cdot \frac{BR}{RC} \cdot \frac{CP}{PA} = -1$. We offer two proofs of this part of Menelaus's theorem.

Proof I Draw a line containing C, parallel to $\overleftrightarrow{AB}$ and intersecting $\overline{PQR}$ (or $\overline{QPR}$) at point D (see Figure 3-3).

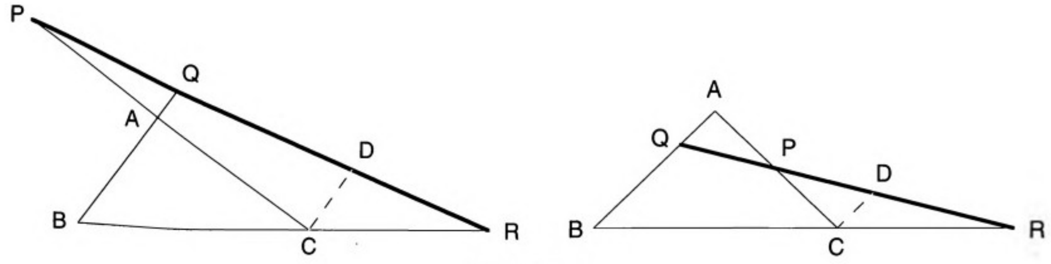

FIGURE 3-3

$$\triangle DCR \sim \triangle QBR \Rightarrow \frac{DC}{QB} = \frac{RC}{BR} \quad or \quad DC = \frac{(QB)(RC)}{BR} \qquad \text{(I)}$$

$$\triangle PDC \sim \triangle PQA \Rightarrow \frac{DC}{AQ} = \frac{CP}{PA} \quad or \quad DC = \frac{(AQ)(CP)}{PA} \qquad \text{(II)}$$

---

* The spherical analogue to Theorem 3.1 for spherical triangle ABC is $\frac{\sin \widehat{AQ}}{\sin \widehat{QB}} \cdot \frac{\sin \widehat{BR}}{\sin \widehat{RC}} \cdot \frac{\sin \widehat{CP}}{\sin \widehat{PA}} = -1$.

From (I) and (II), we get:

$$\frac{(QB)(RC)}{BR} = \frac{(AQ)(CP)}{PA} \quad \text{or} \quad (QB)(RC)(PA) = (AQ)(CP)(BR)$$

This gives us:

$$\frac{AQ}{QB} \cdot \frac{BR}{RC} \cdot \frac{CP}{PA} = 1$$

By taking direction into account in the left-hand diagram of Figure 3-3, we see that $\frac{AQ}{QB}$, $\frac{BR}{RC}$, and $\frac{CP}{PA}$ are each negative ratios; in the right-hand diagram of Figure 3-3, $\frac{BR}{RC}$ is a negative ratio, whereas $\frac{AQ}{QB}$ and $\frac{CP}{PA}$ are positive ratios. Because in each case there is an odd number of negative ratios:

$$\frac{AQ}{QB} \cdot \frac{BR}{RC} \cdot \frac{CP}{PA} = -1 \bullet$$

Proof II Once again we begin by assuming collinearity of P, Q, and R. Draw $\overline{BM} \perp \overleftrightarrow{PR}$, $\overline{AN} \perp \overleftrightarrow{PR}$, and $\overline{CL} \perp \overleftrightarrow{PR}$ (see Figure 3-4).

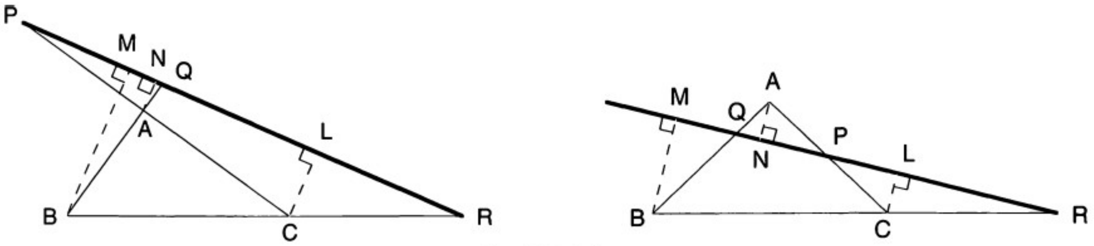

FIGURE 3-4

$$\triangle BMQ \sim \triangle ANQ \Rightarrow \frac{AQ}{QB} = \frac{AN}{BM} \qquad \text{(I)}$$

$$\triangle LCP \sim \triangle NAP \Rightarrow \frac{CP}{PA} = \frac{LC}{AN} \qquad \text{(II)}$$

$$\triangle MRB \sim \triangle LRC \Rightarrow \frac{BR}{RC} = \frac{BM}{LC} \qquad \text{(III)}$$

By multiplying (I), (II), and (III), we get, numerically:

$$\frac{AQ}{QB} \cdot \frac{CP}{PA} \cdot \frac{BR}{RC} = \frac{AN}{BM} \cdot \frac{LC}{AN} \cdot \frac{BM}{LC} = 1$$

From (I) and (II), we get:

$$\frac{(QB)(RC)}{BR} = \frac{(AQ)(CP)}{PA} \quad \text{or} \quad (QB)(RC)(PA) = (AQ)(CP)(BR)$$

This gives us:

$$\frac{AQ}{QB} \cdot \frac{BR}{RC} \cdot \frac{CP}{PA} = 1$$

By taking direction into account in the left-hand diagram of Figure 3-3, we see that $\frac{AQ}{QB}$, $\frac{BR}{RC}$, and $\frac{CP}{PA}$ are each negative ratios; in the right-hand diagram of Figure 3-3, $\frac{BR}{RC}$ is a negative ratio, whereas $\frac{AQ}{QB}$ and $\frac{CP}{PA}$ are positive ratios. Because in each case there is an odd number of negative ratios:

$$\frac{AQ}{QB} \cdot \frac{BR}{RC} \cdot \frac{CP}{PA} = -1 \; \bullet$$

Ⓟroof II Once again we begin by assuming collinearity of P, Q, and R. Draw $\overline{BM} \perp \overleftrightarrow{PR}$, $\overline{AN} \perp \overleftrightarrow{PR}$, and $\overline{CL} \perp \overleftrightarrow{PR}$ (see Figure 3-4).

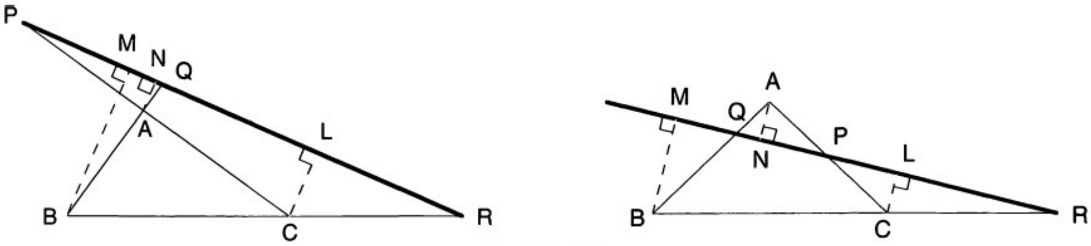

FIGURE 3-4

$$\triangle BMQ \sim \triangle ANQ \Rightarrow \frac{AQ}{QB} = \frac{AN}{BM} \tag{I}$$

$$\triangle LCP \sim \triangle NAP \Rightarrow \frac{CP}{PA} = \frac{LC}{AN} \tag{II}$$

$$\triangle MRB \sim \triangle LRC \Rightarrow \frac{BR}{RC} = \frac{BM}{LC} \tag{III}$$

By multiplying (I), (II), and (III), we get, numerically:

$$\frac{AQ}{QB} \cdot \frac{CP}{PA} \cdot \frac{BR}{RC} = \frac{AN}{BM} \cdot \frac{LC}{AN} \cdot \frac{BM}{LC} = 1$$

APPLICATIONS OF MENELAUS'S THEOREM

Before investigating other famous theorems that can be proved using Menelaus's theorem, we will consider a few applications of Menelaus's theorem. Each of these unnamed theorems presents some very interesting results that are very easily proved using Menelaus's theorem.

Application 1 Prove that the interior angle bisectors of two angles of a nonisosceles triangle and the exterior angle bisector of the third angle meet the opposite sides in three collinear points. ●

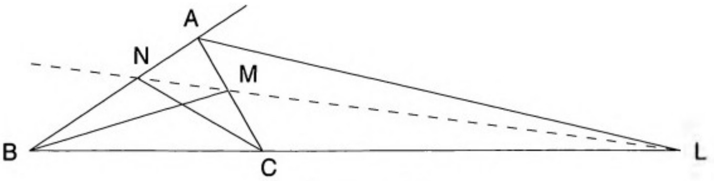

FIGURE 3-5

Proof In $\triangle ABC$, $\overline{BM}$ and $\overline{CN}$ are the interior angle bisectors, while $\overline{AL}$ bisects the exterior angle at point A (see Figure 3-5). Because the bisector of an angle (interior or exterior) of a triangle partitions the opposite side proportionally to the two remaining sides, we have:

$$\frac{AM}{MC}=\frac{AB}{BC}\quad\frac{BN}{NA}=\frac{BC}{AC}\quad\text{and}\quad\frac{CL}{BL}=\frac{AC}{AB}$$

Therefore, by multiplication:

$$\frac{AM}{MC}\cdot\frac{BN}{NA}\cdot\frac{CL}{BL}=\frac{AB}{BC}\cdot\frac{BC}{AC}\cdot\frac{AC}{AB}=1$$

However, because $\frac{CL}{BL}=\frac{-CL}{LB}$:

$$\frac{AM}{MC}\cdot\frac{BN}{NA}\cdot\frac{CL}{LB}=-1$$

Thus, by Menelaus's theorem, points N, M, and L must be collinear. ●

Application 2 Prove that the exterior angle bisectors of any nonisosceles triangle meet the opposite sides in three collinear points. ●

Proof In $\triangle ABC$, the bisectors of the exterior angles at A, B, and C meet the opposite sides (extended) at points N, L, and M, respectively (see Figure 3-6). Because the bisector of an angle (interior or exterior) of a triangle partitions the opposite side proportionally to the two remaining sides, we have:

$$\frac{CL}{AL} = \frac{BC}{AB} \quad \frac{AM}{BM} = \frac{AC}{BC} \quad \text{and} \quad \frac{BN}{CN} = \frac{AB}{AC}$$

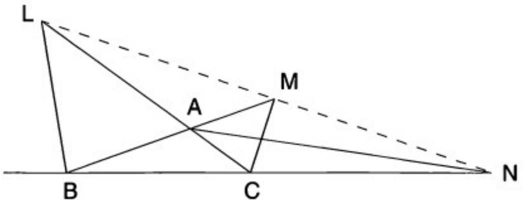

FIGURE 3-6

Therefore:

$$\frac{CL}{AL} \cdot \frac{AM}{BM} \cdot \frac{BN}{CN} = \frac{BC}{AB} \cdot \frac{AC}{BC} \cdot \frac{AB}{AC} = -1 \text{ (because all three ratios are negative)}$$

Thus, by Menelaus's theorem, points L, M, and N are collinear. ●

Application 3 A circle through vertices B and C of $\triangle ABC$ meets $\overline{AB}$ at point P and $\overline{AC}$ at point R. If $\overleftrightarrow{PR}$ meets $\overleftrightarrow{BC}$ at point Q, prove that $\frac{QC}{QB} = \frac{(RC)(AC)}{(PB)(AB)}$. ●

Proof Consider $\triangle ABC$ with transversal $\overline{QPR}$ (Figure 3-7). By Menelaus's theorem:

$$\frac{RC}{AR} \cdot \frac{AP}{PB} \cdot \frac{QB}{CQ} = -1$$

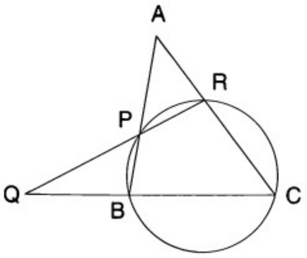

FIGURE 3-7

Then, considering absolute values, we have:

$$\frac{QC}{QB} = \frac{RC}{AR} \cdot \frac{AP}{PB} \tag{I}$$

However, $(AP)(AB) = (AR)(AC)$. (If two secant segments intersect outside the circle, the product of the lengths of one secant segment and its external segment equals the product of the lengths of the other secant segment and its external segment.) Therefore:

$$\frac{AP}{AR} = \frac{AC}{AB} \tag{II}$$

By substituting (II) into (I), we get our desired result:

$$\frac{QC}{QB} = \frac{(RC)(AC)}{(PB)(AB)} \bullet$$

Often, both Menelaus's theorem and its dual, Ceva's theorem, are needed to solve a problem or prove a theorem. The next applications demonstrate this.

Application 4 In right triangle ABC, points P and Q are on $\overline{BC}$ and $\overline{AC}$, respectively, such that $CP = CQ = 2$. Through R, the point of intersection of $\overline{AP}$ and $\overline{BQ}$, a line is drawn also passing through point C and meeting $\overline{AB}$ at point S. $\overleftrightarrow{PQ}$ meets $\overleftrightarrow{AB}$ at point T. If hypotenuse $AB = 10$ and $AC = 8$, find TS (Figure 3-8). ●

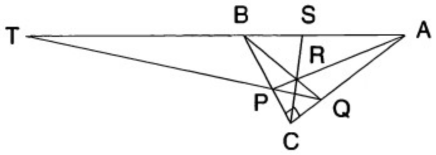

FIGURE 3-8

Solution In right triangle ABC, hypotenuse $AB = 10$ and $AC = 8$, so $BC = 6$ (by the Pythagorean theorem). In $\triangle ABC$, because $\overline{AP}$, $\overline{BQ}$, and $\overline{CS}$ are concurrent, by Ceva's theorem:

$$\frac{AQ}{QC} \cdot \frac{CP}{PB} \cdot \frac{BS}{SA} = 1$$

Substituting, we get:

$$\frac{6}{2} \cdot \frac{2}{4} \cdot \frac{BS}{10 - BS} = 1 \quad \text{or} \quad BS = 4$$

Now consider $\triangle ABC$ with transversal $\overline{QPT}$. By Menelaus's theorem:

$$\frac{AQ}{QC} \cdot \frac{CP}{PB} \cdot \frac{BT}{TA} = -1$$

Because we are not dealing with directed line segments, this may be restated as:

$$(AQ)(CP)(BT) = (QC)(PB)(AT)$$

Substituting, we get:

$$(6)(2)(BT) = (2)(4)(BT + 10)$$

Solving for BT gives us $BT = 20$, and thus $TS = 24$. ●

Application 5 In quadrilateral $ABCD$, $\overleftrightarrow{AB}$ and $\overleftrightarrow{CD}$ meet at point P, while $\overleftrightarrow{AD}$ and $\overleftrightarrow{BC}$ meet at point Q. Diagonals $\overleftrightarrow{AC}$ and $\overleftrightarrow{BD}$ meet $\overleftrightarrow{PQ}$ at points X and Y, respectively. Prove that $\frac{PX}{XQ} = -\frac{PY}{YQ}$ (see Figure 3-9). ●

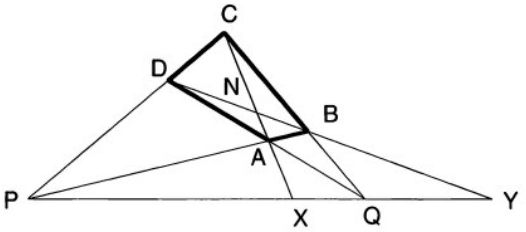

FIGURE 3-9

Proof Consider $\triangle PQC$ with $\overline{PB}$, $\overline{QD}$, and $\overline{CX}$ concurrent. By Ceva's theorem:

$$\frac{PX}{XQ} \cdot \frac{QB}{BC} \cdot \frac{CD}{DP} = 1 \qquad \text{(I)}$$

Now consider $\triangle PQC$ with $\overline{DBY}$ as a transversal. By Menelaus's theorem:

$$\frac{PY}{YQ} \cdot \frac{QB}{BC} \cdot \frac{CD}{DP} = -1 \qquad \text{(II)}$$

Therefore, from (I) and (II):

$$\frac{PX}{XQ} = -\frac{PY}{YQ}$$ ●

We will now consider some rather famous theorems that can be proved using Menelaus's theorem.

DESARGUES'S THEOREM

During his lifetime, Gérard Desargues (1591–1661) did not enjoy the important stature as a mathematician that he has attained in later years. This lack of popularity was in part due to the then recent development of analytic geometry by René Descartes (1596–1650) and to Desargues's introduction of many new and largely unfamiliar terms. (Incidentally, we make every effort in this book not to introduce any new terms in order to make it more reader-friendly—we want to learn from Desargues's misjudgment.)

In 1648, his pupil, Abraham Bosse, a master engraver, published a book titled *Manière universelle de M. Desargues, pour pratiquer la perspective,* which was not popularized until about two centuries later. This book contained a theorem that in the nineteenth century became one of the fundamental propositions of projective geometry. It is this theorem that is of interest to us here. It involves placing any two triangles in a position that will enable the three lines joining corresponding vertices to be concurrent. Remarkably, when this is achieved the pairs of corresponding sides meet in three collinear points. We will prove Desargues's theorem by using Menelaus's theorem.

THEOREM 3.2 **(Desargues's theorem)** If $\triangle A_1B_1C_1$ and $\triangle A_2B_2C_2$ are situated so that the lines joining the corresponding vertices, $\overleftrightarrow{A_1A_2}$, $\overleftrightarrow{B_1B_2}$, and $\overleftrightarrow{C_1C_2}$, are concurrent, then the pairs of corresponding sides intersect in three collinear points.

INTERACTIVE 3-10

Drag vertices A_1, B_1, C_1 and A_2, B_2, C_2 to change the shape of the triangles and point P and see that Desargues's theorem is true.

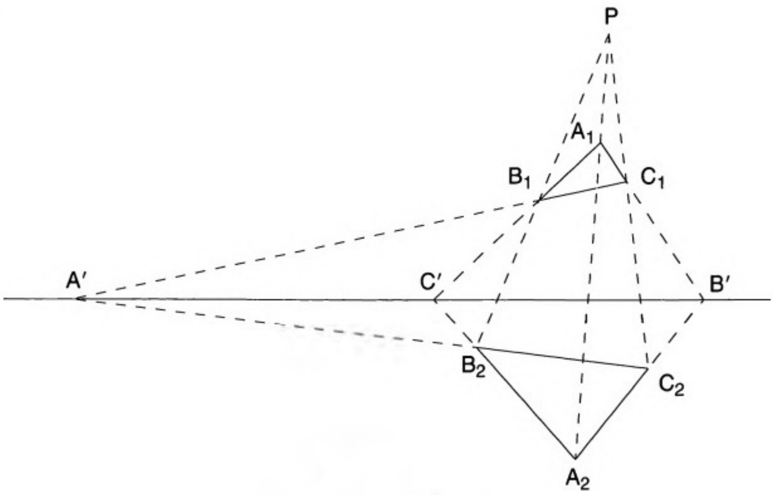

FIGURE 3-10

Proof In Figure 3-10, $\overleftrightarrow{A_1A_2}$, $\overleftrightarrow{B_1B_2}$, and $\overleftrightarrow{C_1C_2}$ all meet at point P, by hypothesis. $\overleftrightarrow{B_2C_2}$ and $\overleftrightarrow{B_1C_1}$ meet at A′; $\overleftrightarrow{A_2C_2}$ and $\overleftrightarrow{A_1C_1}$ meet at B′; $\overleftrightarrow{B_2A_2}$ and $\overleftrightarrow{B_1A_1}$ meet at C′.

Consider $\overline{A'B_1C_1}$ to be a transversal of $\triangle PB_2C_2$. Therefore, by Menelaus's theorem:

$$\frac{PB_1}{B_1B_2} \cdot \frac{B_2A'}{A'C_2} \cdot \frac{C_2C_1}{C_1P} = -1 \qquad \text{(I)}$$

Similarly, considering $\overline{C'B_1A_1}$ as a transversal of $\triangle PB_2A_2$:

$$\frac{PA_1}{A_1A_2} \cdot \frac{A_2C'}{C'B_2} \cdot \frac{B_2B_1}{B_1P} = -1 \text{ (Menelaus's theorem)} \qquad \text{(II)}$$

Now, taking $\overline{B'A_1C_1}$ as a transversal of $\triangle PA_2C_2$:

$$\frac{PC_1}{C_1C_2} \cdot \frac{C_2B'}{B'A_2} \cdot \frac{A_2A_1}{A_1P} = -1 \text{ (Menelaus's theorem)} \qquad \text{(III)}$$

By multiplying (I), (II), and (III), we get:

$$\frac{B_2A'}{A'C_2} \cdot \frac{A_2C'}{C'B_2} \cdot \frac{C_2B'}{B'A_2} = -1$$

Thus, by Menelaus's theorem, applied to $\triangle A_2B_2C_2$, we have points A', B', and C' collinear. ●

It should be noted that the converse of Desargues's theorem is also true. It is the dual of the original theorem. We leave the proof as an exercise.

To appreciate the value of Desargues's theorem, we will examine some applications. Although each may be proved in other ways, we will use the method employing Desargues's theorem.

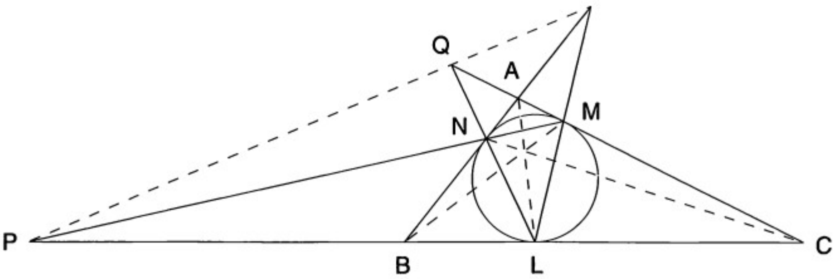

FIGURE 3-11

Application 6 A circle inscribed in $\triangle ABC$ is tangent to sides $\overline{BC}$, $\overline{CA}$, and $\overline{AB}$ at points L, M, and N, respectively. $\overleftrightarrow{MN}$ intersects $\overleftrightarrow{BC}$ at point P, $\overleftrightarrow{NL}$ intersects $\overleftrightarrow{AC}$ at point Q, and $\overleftrightarrow{ML}$ intersects $\overleftrightarrow{AB}$ at point R. Prove that points P, Q, and R are collinear (see Figure 3-11). ●

Proof Because the tangent segments from an external point to a circle are congruent:

$$AN = AM \quad NB = BL \quad MC = LC$$

Therefore:

$$\frac{AN}{NB} \cdot \frac{BL}{LC} \cdot \frac{MC}{AM} = 1$$

By Ceva's theorem, $\overleftrightarrow{AL}$, $\overleftrightarrow{BM}$, and $\overleftrightarrow{CN}$ are concurrent. Because these are the lines joining the corresponding vertices of $\triangle ABC$ and $\triangle LMN$, by Desargues's theorem the intersections of the corresponding sides are collinear; therefore points P, Q, and R are collinear. ●

Application 7 In $\triangle ABC$, points F, E, and D are the feet of the altitudes drawn from the vertices A, B, and C, respectively. The sides of pedal* triangle FED, $\overleftrightarrow{EF}$, $\overleftrightarrow{DF}$, and $\overleftrightarrow{DE}$, intersect the sides of $\triangle ABC$, $\overleftrightarrow{AB}$, $\overleftrightarrow{AC}$, and $\overleftrightarrow{BC}$, at points M, N, and L, respectively. Prove that points M, N, and L are collinear (see Figure 3-12). ●

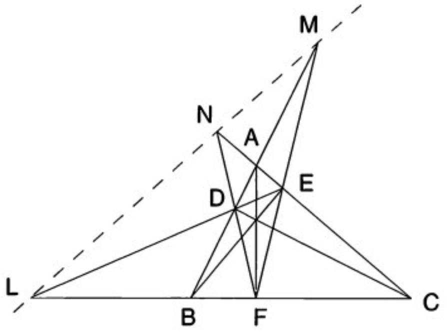

FIGURE 3-12

Proof Let A, B, C and F, E, D be corresponding vertices of $\triangle ABC$ and $\triangle FED$. Because $\overline{AF}$, $\overline{CD}$, and $\overline{BE}$ are concurrent (they are the altitudes of $\triangle ABC$), the intersections of the corresponding sides $\overleftrightarrow{DE}$ and $\overleftrightarrow{BC}$, $\overleftrightarrow{FE}$ and $\overleftrightarrow{BA}$, and $\overleftrightarrow{FD}$ and $\overleftrightarrow{CA}$ are collinear by Desargues's theorem. ●

---

* A pedal triangle (of a given triangle) is formed by joining the feet of the perpendiculars from any given point to the sides of the given triangle.

PASCAL'S THEOREM

Blaise Pascal (1623–1662), a contemporary of Desargues, is regarded today as one of the true geniuses in the history of mathematics. Although eccentricities kept him from achieving his true potential, he is considered one of the originators of the formalized study of probability (an outgrowth of his correspondences with Fermat), and he made many important contributions to other branches of mathematics. We concern ourselves here with one of his contributions to geometry.

In 1640, at the age of sixteen, Pascal published a one-page paper titled *Essay pour les coniques.* It contained a theorem that Pascal referred to as *mysterium hexagrammicum.* The work highly impressed Descartes, who couldn't believe it was the work of a boy. This theorem states that the intersections of the opposite sides of a hexagon inscribed in a conic section are collinear. For our purposes, we will consider only the case in which the conic section is a circle and the hexagon has no pair of opposite sides parallel.

THEOREM 3.3 **(Pascal's theorem)** If a hexagon with no pair of opposite sides parallel is inscribed in a circle, then the intersections of the opposite sides are collinear.

Proof Hexagon $ABCDEF$ is inscribed in a circle (see Figure 3-13). The pairs of opposite sides $\overleftrightarrow{AB}$ and $\overleftrightarrow{DE}$ meet at point L, $\overleftrightarrow{CB}$ and $\overleftrightarrow{EF}$ meet at point M, and $\overleftrightarrow{CD}$ and $\overleftrightarrow{AF}$ meet at point N. Also, $\overleftrightarrow{AB}$ meets $\overleftrightarrow{CN}$ at point X, $\overleftrightarrow{EF}$ meets $\overleftrightarrow{CN}$ at point Y, and $\overleftrightarrow{EF}$ meets $\overleftrightarrow{AB}$ at point Z.

Consider $\overleftrightarrow{BC}$ to be a transversal of $\triangle XYZ$. Then, by Menelaus's theorem:

$$\frac{ZB}{BX} \cdot \frac{XC}{CY} \cdot \frac{YM}{MZ} = -1 \tag{I}$$

Taking $\overleftrightarrow{AF}$ to be a transversal of $\triangle XYZ$:

$$\frac{ZA}{AX} \cdot \frac{YF}{FZ} \cdot \frac{XN}{NY} = -1 \quad \text{(Menelaus's theorem)} \tag{II}$$

Also, because $\overleftrightarrow{DE}$ is a transversal of $\triangle XYZ$:

$$\frac{XD}{DY} \cdot \frac{YE}{EZ} \cdot \frac{ZL}{LX} = -1 \quad \text{(Menelaus's theorem)} \tag{III}$$

By multiplying (I), (II), and (III), we get:

$$\frac{YM}{MZ} \cdot \frac{XN}{NY} \cdot \frac{ZL}{LX} \cdot \frac{(ZB)(ZA)}{(EZ)(FZ)} \cdot \frac{(XD)(XC)}{(AX)(BX)} \cdot \frac{(YE)(YF)}{(DY)(CY)} = -1 \tag{IV}$$

When two secant segments are drawn to a circle from an external point, the product of the lengths of one secant and its external segment equals the product of the lengths of the other secant and its external segment.

INTERACTIVE 3-13

Drag points *A, B, C, D, E,* and *F* on the circle to see that Pascal's theorem is true.

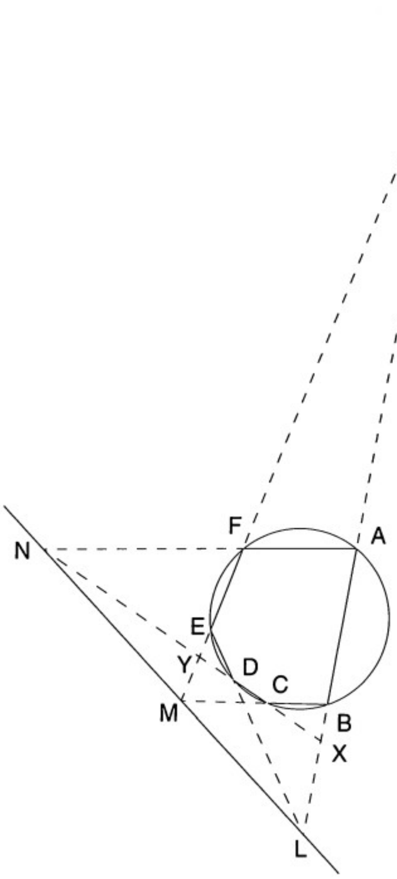

FIGURE 3-13

Therefore:

$$\frac{(ZB)(ZA)}{(EZ)(FZ)} = 1 \qquad \text{(V)}$$

$$\frac{(XD)(XC)}{(AX)(BX)} = 1 \qquad \text{(VI)}$$

$$\frac{(YE)(YF)}{(DY)(CY)} = 1 \qquad \text{(VII)}$$

By substituting (V), (VI), and (VII) into (IV), we get:

$$\frac{YM}{MZ} \cdot \frac{XN}{NY} \cdot \frac{ZL}{LX} = -1$$

Thus, by Menelaus's theorem, points *M, N,* and *L* must be collinear. ●

It is interesting to note that Pascal's theorem can be extended in the following manner.

INTERACTIVE 3-13

Drag points *A, B, C, D, E,* and *F* on the circle to see that Pascal's theorem is true.

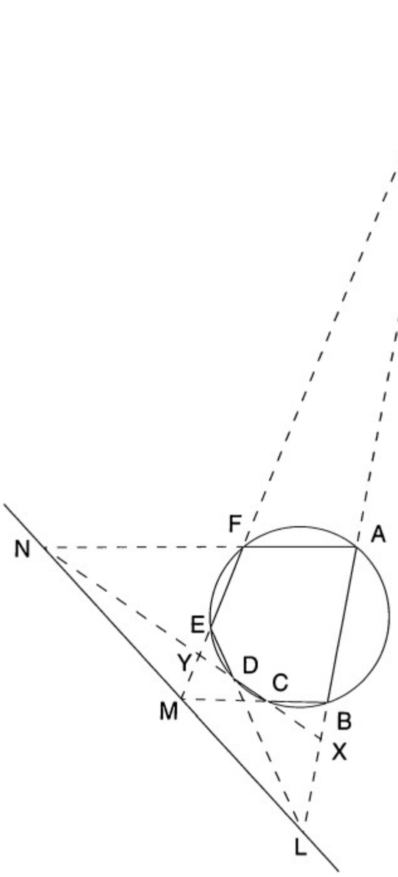

FIGURE 3-13

Therefore:

$$\frac{(ZB)(ZA)}{(EZ)(FZ)} = 1 \qquad \text{(V)}$$

$$\frac{(XD)(XC)}{(AX)(BX)} = 1 \qquad \text{(VI)}$$

$$\frac{(YE)(YF)}{(DY)(CY)} = 1 \qquad \text{(VII)}$$

By substituting (V), (VI), and (VII) into (IV), we get:

$$\frac{YM}{MZ} \cdot \frac{XN}{NY} \cdot \frac{ZL}{LX} = -1$$

Thus, by Menelaus's theorem, points *M, N,* and *L* must be collinear. ●

It is interesting to note that Pascal's theorem can be extended in the following manner.

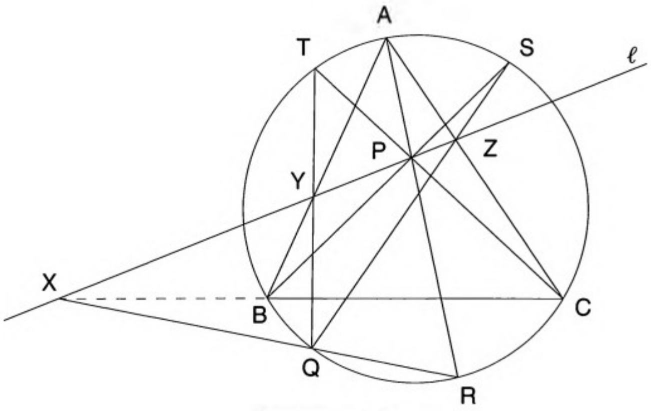

FIGURE 3-16

Application 9 Select any point P not on $\triangle ABC$ and a line ℓ containing P and intersecting sides $\overleftrightarrow{BC}$, $\overleftrightarrow{AB}$, and $\overleftrightarrow{AC}$ at points X, Y, and Z, respectively. Let $\overleftrightarrow{AP}$, $\overleftrightarrow{BP}$, and $\overleftrightarrow{CP}$ intersect the circumcircle of $\triangle ABC$ at points R, S, and T, respectively (see Figure 3-16). Prove that $\overleftrightarrow{RX}$, $\overleftrightarrow{SZ}$, and $\overleftrightarrow{TY}$ are concurrent. ●

Proof Let $\overleftrightarrow{RX}$ intersect the circumcircle at point Q. Consider hexagon $ARQTCB$ and apply Pascal's theorem to it. We notice that because $\overleftrightarrow{AR} \cap \overleftrightarrow{AB}$ at point P and $\overleftrightarrow{RQ} \cap \overleftrightarrow{CB}$ at point X, $\overleftrightarrow{TQ} \cap \overleftrightarrow{AB}$ at a point on ℓ, which must be Y (because $\overleftrightarrow{AB} \cap \ell$ at point Y).

Now consider hexagon $ARQSBC$. Because $\overleftrightarrow{AR} \cap \overleftrightarrow{SB}$ at P and $\overleftrightarrow{RQ} \cap \overleftrightarrow{CB}$ at X, $\overleftrightarrow{SQ} \cap \overleftrightarrow{AC}$ at a point on ℓ, which must be Z. Thus $\overleftrightarrow{RX}$, $\overleftrightarrow{SZ}$, and $\overleftrightarrow{TY}$ are and concurrent. ●

BRIANCHON'S THEOREM

In 1806, at the age of twenty-one, a student at the École Polytechnique, Charles Julien Brianchon (1785–1864), published an article in the *Journal de L'École Polytechnique* that was to become one of the fundamental contributions to the study of conic sections in projective geometry. His development led to a restatement of the somewhat forgotten theorem of Pascal and its extension, after which Brianchon stated a new theorem that now bears his name. Brianchon's theorem, which states "In any hexagon circumscribed about a conic section, the three diagonals cross each other in the same point,"* bears a curious resemblance to Pascal's theorem. They are, in fact, duals of each other. This can be easily seen by comparing the following versions of each theorem:

---

* *Source Book in Mathematics,* edited by D. E. Smith (New York: McGraw-Hill Book Co., 1929), p. 336.

Pascal's Theorem	**Brianchon's Theorem**
The points of intersection of the opposite sides of a hexagon inscribed in a conic section are collinear.	The lines joining the opposite vertices of a hexagon circumscribed about a conic section are concurrent.

Notice that the two statements above are alike except for the underlined words, which are duals of one another. As with Pascal's theorem, we will consider only the conic section that is a circle.

THEOREM 3.5 **(Brianchon's theorem)** If a hexagon is circumscribed about a circle (see Figure 3-17), the lines containing opposite vertices are concurrent.

INTERACTIVE 3-17

Drag points *A, B, C, D, E,* and *F* on the circle and see that Brianchon's theorem is true.

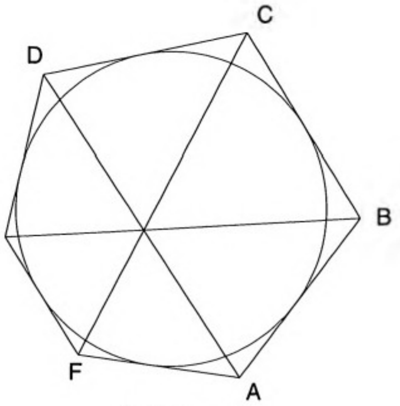

FIGURE 3-17

The simplest proofs of this theorem require knowledge of concepts from projective geometry. Although we are prepared at this point to prove this theorem by Euclidean methods, our proof will be more concise if we wait until we study radical axes later in this chapter on page 71.

Brianchon suggested the following application immediately after the statement of his new theorem.

Application 10 Pentagon *ABCDE* is circumscribed about a circle, with points of tangency at *F*, *M*, *N*, *R*, and *S*. If diagonals $\overline{AD}$ and $\overline{BE}$ intersect at point *P*, prove that points *C*, *P*, and *F* are collinear (see Figure 3-18). ●

Proof Consider the hexagon circumscribed about a circle (Figure 3-17) having its sides $\overline{AF}$ and $\overline{EF}$ merge into one line segment. Thus $\overline{AFE}$ is now a side of a circumscribed pentagon with *F* as one point of tangency (see Figure 3-18). Thus we can view the pentagon in Figure 3-18 as a degenerate hexagon. We then simply apply Brianchon's theorem to this degenerate hexagon to obtain our desired conclusion. That is, $\overline{AD}$, $\overline{BE}$, and $\overline{CF}$ are concurrent at point *P*, or points *C*, *P*, and *F* are collinear. ●

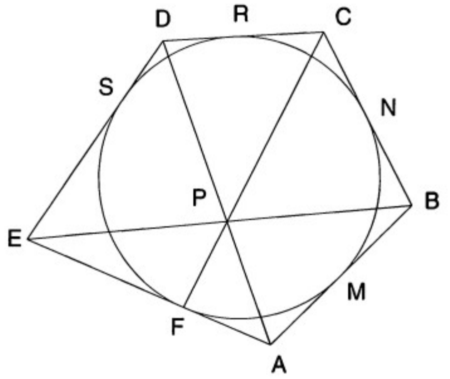

FIGURE 3-18

PAPPUS'S THEOREM

Consider the vertices of a hexagon $AB'CA'BC'$ (Figure 3-19) located alternately on two lines (see Figure 3-20). Suppose we now draw the lines that were the opposite sides of the hexagon to locate their point of intersection. We find that the three points of intersection of these pairs of "opposite sides" are collinear. This conclusion was first published by Pappus of Alexandria in his *Mathematical Collection* circa A.D. 300.

For the purpose of providing a proof, we will restate Pappus's theorem. You will notice once again that the proof uses Menelaus's theorem repeatedly.

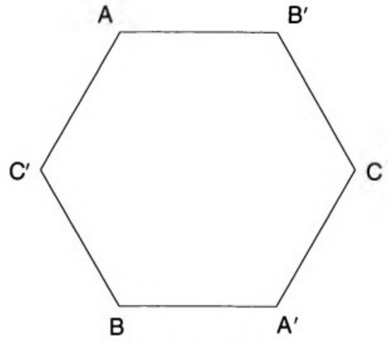

FIGURE 3-19

INTERACTIVE 3-20

Drag points *A′, B′, C′, A, B,* and *C* and see that Pappus's theorem is true.

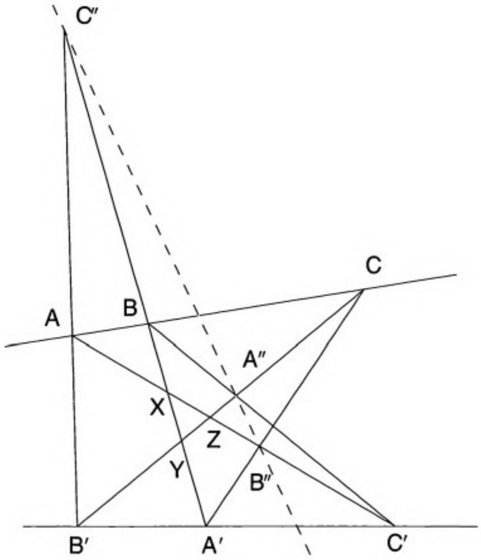

FIGURE 3-20

THEOREM 3.6 **(Pappus's theorem)** Points A, B, and C are on one line, and points A', B', and C' are on another line (in any order). If $\overleftrightarrow{AB'}$ and $\overleftrightarrow{A'B}$ meet at C'', $\overleftrightarrow{AC'}$ and $\overleftrightarrow{A'C}$ meet at B'', and $\overleftrightarrow{BC'}$ and $\overleftrightarrow{B'C}$ meet at A'', then points A'', B'', and C'' are collinear.

Proof In Figure 3-20, $\overline{B'C}$ meets $\overline{A'B}$ at point Y, $\overline{AC'}$ meets $\overline{A'B}$ at point X, and $\overline{B'C}$ meets $\overline{AC'}$ at point Z. Consider $\overline{C''AB'}$ as a transversal of $\triangle XYZ$. By Menelaus's theorem:

$$\frac{ZB'}{YB'} \cdot \frac{XA}{ZA} \cdot \frac{YC''}{XC''} = -1 \qquad \text{(I)}$$

Taking $\overline{A'B''C}$ as a transversal of $\triangle XYZ$:

$$\frac{YA'}{XA'} \cdot \frac{XB''}{ZB''} \cdot \frac{ZC}{YC} = -1 \text{ (Menelaus's theorem)} \qquad \text{(II)}$$

PAPPUS'S THEOREM

Consider the vertices of a hexagon $AB'CA'BC'$ (Figure 3-19) located alternately on two lines (see Figure 3-20). Suppose we now draw the lines that were the opposite sides of the hexagon to locate their point of intersection. We find that the three points of intersection of these pairs of "opposite sides" are collinear. This conclusion was first published by Pappus of Alexandria in his *Mathematical Collection* circa A.D. 300.

For the purpose of providing a proof, we will restate Pappus's theorem. You will notice once again that the proof uses Menelaus's theorem repeatedly.

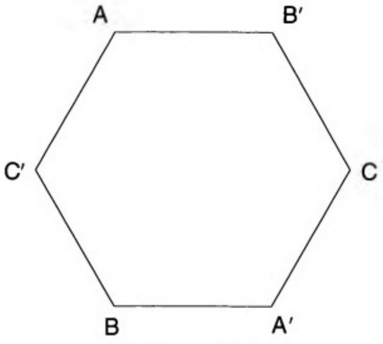

FIGURE 3-19

INTERACTIVE 3-20

Drag points A', B', C', A, B, and C and see that Pappus's theorem is true.

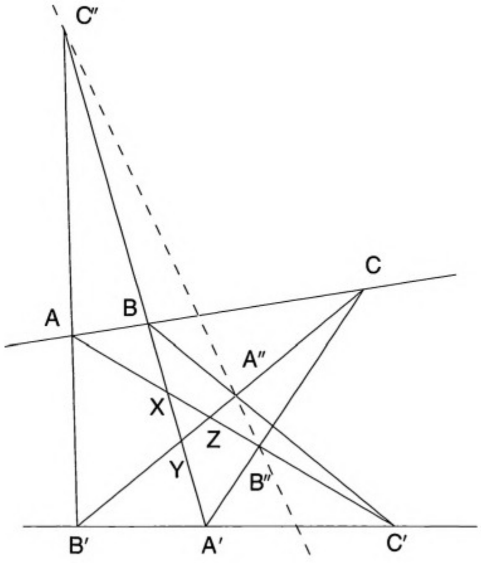

FIGURE 3-20

THEOREM 3.6 **(Pappus's theorem)** Points A, B, and C are on one line, and points A', B', and C' are on another line (in any order). If $\overleftrightarrow{AB'}$ and $\overleftrightarrow{A'B}$ meet at C'', $\overleftrightarrow{AC'}$ and $\overleftrightarrow{A'C}$ meet at B'', and $\overleftrightarrow{BC'}$ and $\overleftrightarrow{B'C}$ meet at A'', then points A'', B'', and C'' are collinear.

Proof In Figure 3-20, $\overline{B'C}$ meets $\overline{A'B}$ at point Y, $\overline{AC'}$ meets $\overline{A'B}$ at point X, and $\overline{B'C}$ meets $\overline{AC'}$ at point Z. Consider $\overline{C''AB'}$ as a transversal of $\triangle XYZ$. By Menelaus's theorem:

$$\frac{ZB'}{YB'} \cdot \frac{XA}{ZA} \cdot \frac{YC''}{XC''} = -1 \qquad \text{(I)}$$

Taking $\overline{A'B''C}$ as a transversal of $\triangle XYZ$:

$$\frac{YA'}{XA'} \cdot \frac{XB''}{ZB''} \cdot \frac{ZC}{YC} = -1 \text{ (Menelaus's theorem)} \qquad \text{(II)}$$

INTERACTIVE 3-21

Drag vertices *A*, *B*, and *C* to change the shape of the triangle; drag point *P* on the circle and watch the Simson line.

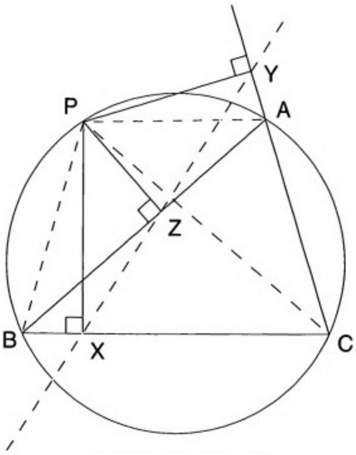

FIGURE 3-21

Proof I (See Figure 3-21.) Draw $\overline{PA}$, $\overline{PB}$, and $\overline{PC}$.

$$m\angle PBA = \frac{1}{2} m\,\widehat{AP}$$

$$m\angle PCA = \frac{1}{2} m\,\widehat{AP}$$

Therefore $m\angle PBA = m\angle PCA = a$. Thus:

$$\frac{BZ}{PZ} = \cot a = \frac{CY}{PY} \text{ (in } \triangle PZB \text{ and } \triangle PYC) \quad \text{or} \quad \frac{BZ}{PZ} = \frac{CY}{PY}$$

This implies:

$$\frac{BZ}{CY} = \frac{PZ}{PY} \tag{I}$$

Similarly, $m\angle PAB = m\angle PCB = b$ (both are $\frac{1}{2}m\widehat{PB}$). Therefore:

$$\frac{AZ}{PZ} = \cot b = \frac{CX}{PX} \text{ (in } \triangle PAZ \text{ and } \triangle PCX) \quad \text{or} \quad \frac{AZ}{PZ} = \frac{CX}{PX}$$

This implies:

$$\frac{CX}{AZ} = \frac{PX}{PZ} \tag{II}$$

Because $\angle PBC$ and $\angle PAC$ are opposite angles of an inscribed (cyclic) quadrilateral, they are supplementary. However, $\angle PAY$ is also supplementary to $\angle PAC$. Therefore:

$$m\angle PBC = m\angle PAY = c$$

Thus:

$$\frac{BX}{PX} = \cot c = \frac{AY}{PY}\,(\text{in } \triangle PBX \text{ and } \triangle PAY) \quad \text{or} \quad \frac{BX}{PX} = \frac{AY}{PY}$$

This implies:

$$\frac{AY}{BX} = \frac{PY}{PX} \tag{III}$$

By multiplying (I), (II), and (III), we obtain:

$$\frac{BZ}{CY} \cdot \frac{CX}{AZ} \cdot \frac{AY}{BX} = \frac{PZ}{PY} \cdot \frac{PX}{PZ} \cdot \frac{PY}{PX} = 1 \text{ (or } -1\text{, had we considered direction)}$$

Thus, by Menelaus's theorem, points X, Y, and Z are collinear. These three points determine the Simson line of $\triangle ABC$ with respect to point P. ●

Proof II (See Figure 3-21.) Because $\angle PYA$ is supplementary to $\angle PZA$, quadrilateral $PZAY$ is cyclic. Draw $\overline{PA}$, $\overline{PB}$, and $\overline{PC}$. Therefore:

$$m\angle PYZ = m\angle PAZ \tag{I}$$

Similarly, because $\angle PYC$ is supplementary to $\angle PXC$, quadrilateral $PXCY$ is cyclic, and therefore:

$$m\angle PYX = m\angle PCB \tag{II}$$

However, quadrilateral $PACB$ is also cyclic because it is inscribed in the given circumcircle, and therefore:

$$m\angle PAZ (m\angle PAB) = m\angle PCB \tag{III}$$

From (I), (II), and (III), $m\angle PYZ = m\angle PYX$, and thus points X, Y, and Z are collinear. ●

For other proofs of Simson's theorem, see *Challenging Problems in Geometry* by A. S. Posamentier and C. T. Salkind (New York: Dover, 1996), pages 43–45.

There are many simple applications of the Simson line. We will consider a few of them.

Application 11 Sides $\overleftrightarrow{AB}$, $\overleftrightarrow{BC}$, and $\overleftrightarrow{CA}$ of $\triangle ABC$ are cut by a transversal at points Q, R, and S, respectively. The circumcircles of $\triangle ABC$ and $\triangle SCR$ intersect at point P. Prove that quadrilateral $APSQ$ is cyclic. ●

Proof Draw perpendiculars $\overline{PX}$, $\overline{PY}$, $\overline{PZ}$, and $\overline{PW}$ to $\overleftrightarrow{AB}$, $\overleftrightarrow{AC}$, $\overleftrightarrow{QR}$, and $\overleftrightarrow{BC}$, respectively, as in Figure 3-22.

Because point P is on the circumcircle of $\triangle ABC$, points X, Y, and W are collinear (Simson's theorem). Similarly, because point P is on the circumcircle of $\triangle SCR$, points Y, Z, and W are collinear. It then follows that points X, Y, and Z are collinear.

Thus point P must lie on the circumcircle of $\triangle AQS$ (the converse of Simson's theorem, whose proof we leave as an exercise). Therefore quadrilateral $APSQ$ is cyclic. ●

INTERACTIVE 3-22

Drag vertices *A*, *B*, and *C* to change the shape of the triangle and note that *APSQ* is always cyclic.

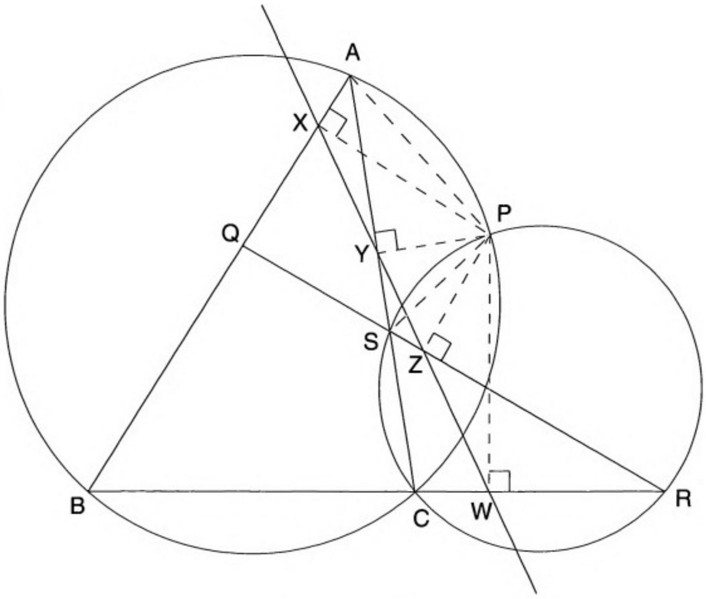

FIGURE 3-22

Application 12 $\overline{AB}$, $\overline{BC}$, $\overline{EC}$, and $\overline{ED}$ form triangles *ABC*, *FBD*, *EFA*, and *EDC*. Prove that the four circumcircles of these triangles meet at a common point. ●

Proof Consider the circumcircles of △*ABC* and △*FBD*, which meet at points *B* and *P*. From point *P* draw perpendiculars $\overline{PX}$, $\overline{PY}$, $\overline{PZ}$, and $\overline{PW}$ to $\overline{BC}$, $\overline{AB}$, $\overline{ED}$, and $\overline{EC}$, respectively (see Figure 3-23). Because point *P* is on the circumcircle of △*ABC*, points *X*, *Y*, and *W* are collinear. Therefore points *X*, *Y*, *Z*, and *W* are collinear.

Because points *Y*, *Z*, and *W* are collinear, point *P* must lie on the circumcircle of △*EFA* (the converse of Simson's theorem). By the same reasoning, because points *X*, *Z*, and *W* are collinear, point *P* lies on the circumcircle of △*EDC*. Thus all four circles pass through point *P*. ●

The Simson line has many interesting properties; we present a few here.

THEOREM 3.8 **(Simson line property I)** If the altitude $\overline{AD}$ of △*ABC* meets the circumcircle at point *P*, then the Simson line of *P* with respect to △*ABC* is parallel to the line tangent to the circle at point *A*.

INTERACTIVE 3-23

Drag vertices *A, B, C, D,* and *E* to change the shape of the triangles and see that the circles always meet at a point.

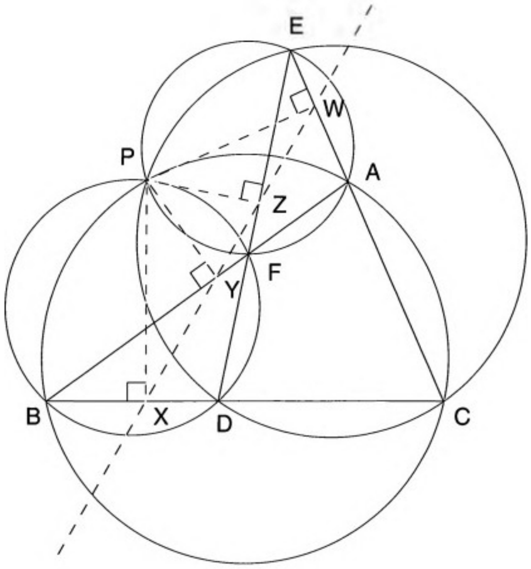

FIGURE 3-23

Proof Because $\overline{PX}$ and $\overline{PZ}$ are perpendicular, respectively, to sides $\overleftrightarrow{AC}$ and $\overleftrightarrow{AB}$ of $\triangle ABC$, points X, D, and Z determine the Simson line of point P with respect to $\triangle ABC$.

Draw $\overline{PB}$ (see Figure 3-24). Consider quadrilateral $PDBZ$, where $m\angle PDB = m\angle PZB = 90°$, thus making quadrilateral $PDBZ$ a cyclic quadrilateral.* In quadrilateral $PDBZ$:

$$m\angle DZB = m\angle DPB \qquad \text{(I)}$$

INTERACTIVE 3-24

Drag vertices *A, B,* and *C* to change the shape of the triangle and see that the required lines are parallel.

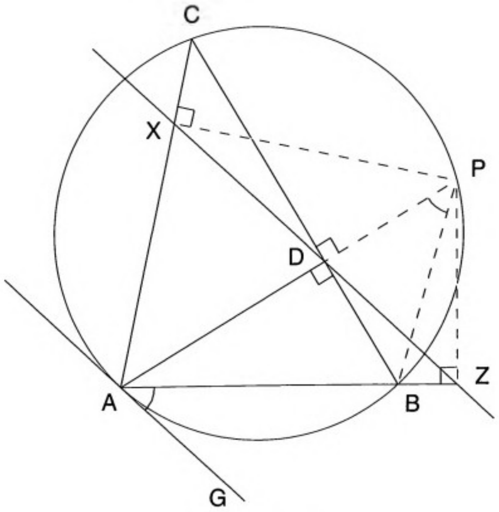

FIGURE 3-24

---

* Because opposite angles are supplementary.

However, in the circumcircle of $\triangle ABC$, $m\angle GAB = \frac{1}{2}(m\overset{\frown}{AB})$, and $m\angle DPB(m\angle APB) = \frac{1}{2}(m\overset{\frown}{AB})$. Therefore:

$$m\angle GAB = m\angle DPB \qquad \text{(II)}$$

From (I) and (II), by transitivity, $m\angle DZB = m\angle GAB$, and thus Simson line $\overleftrightarrow{XDZ}$ is parallel to tangent $\overleftrightarrow{GA}$. ●

THEOREM 3.9 **(Simson line property II)** From point P on the circumcircle of $\triangle ABC$, if perpendiculars $\overline{PX}$, $\overline{PY}$, and $\overline{PZ}$ are drawn to sides $\overleftrightarrow{AC}$, $\overleftrightarrow{AB}$, and $\overleftrightarrow{BC}$, respectively, then $(PA)(PZ) = (PB)(PX)$ (see Figure 3-25).

Proof Because $m\angle PYB = m\angle PZB = 90°$, quadrilateral $PYZB$ is cyclic.* Therefore:

$$m\angle PBY = m\angle PZY \qquad \text{(I)}$$

Likewise, because $m\angle PXA = m\angle PYA = 90°$, quadrilateral $PXAY$ is cyclic, and:

$$m\angle PXY = m\angle PAY \qquad \text{(II)}$$

Points X, Y, and Z are collinear (the Simson line). Therefore, from (I) and (II):

$$\triangle PAB \sim \triangle PXZ \Rightarrow \frac{PA}{PX} = \frac{PB}{PZ} \quad \text{or} \quad (PA)(PZ) = (PB)(PX)$$ ●

INTERACTIVE 3-25

Drag vertices *A*, *B*, and *C* to change the shape of the triangle; drag point *P* on the circle and see that the products are equal.

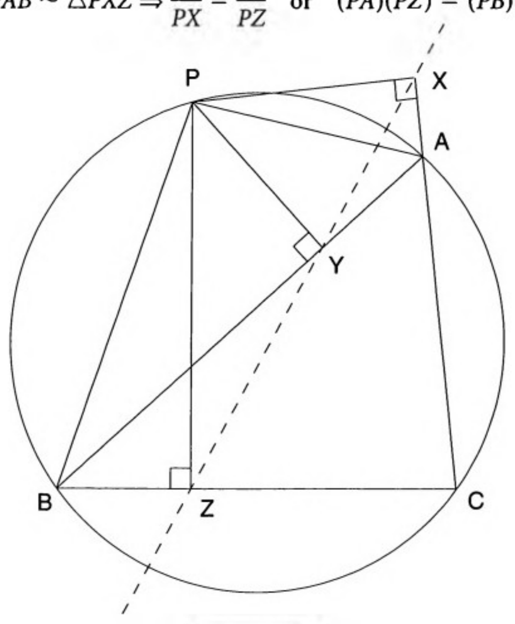

FIGURE 3-25

---

* A quadrilateral is cyclic (i.e., may be inscribed in a circle) if one side subtends congruent angles at the two opposite vertices.

THEOREM 3.10 **(Simson line property III)** The measure of the angle determined by the Simson line of two given points on the circumcircle of a given triangle is equal to one-half the measure of the arc determined by the two points.

Proof In Figure 3-26, $\overleftrightarrow{XYZ}$ is the Simson line for point P and $\overleftrightarrow{UVW}$ is the Simson line for point Q. Extend $\overline{PX}$ and $\overline{QW}$ to meet the circle at points M and N, respectively. Then draw $\overline{AM}$ and $\overline{AN}$. Because $m\angle PZB = m\angle PXB = 90°$, quadrilateral $PZXB$ is cyclic, and:

$$m\angle ZXP = m\angle ZBP \qquad \text{(I)}$$

$$m\angle ABP = m\angle AMP \quad \text{or} \quad m\angle ZBP = m\angle AMP \qquad \text{(II)}$$

From (I) and (II), $m\angle ZXP = m\angle AMP$. Therefore:

$$\overleftrightarrow{XYZ} \parallel \overline{AM}$$

In a similar fashion, it may be shown that $\overleftrightarrow{UVW} \parallel \overline{AN}$.

Hence, if T is the point of intersection of the two Simson lines, then $m\angle XTW = m\angle MAN$ because their corresponding sides are parallel. Now, $m\angle MAN = \frac{1}{2}(m\overset{\frown}{MN})$, but because $\overline{PM} \parallel \overline{QN}$, we have $m\overset{\frown}{MN} = m\overset{\frown}{PQ}$ and therefore $m\angle MAN = \frac{1}{2}(m\overset{\frown}{PQ})$. Thus $m\angle XTW = \frac{1}{2}(m\overset{\frown}{PQ})$. ●

INTERACTIVE 3-26

Drag vertices *A*, *B*, and *C* to change the shape of the triangle; drag points *P* and *Q* on the circle, and the angle is one-half the measure of the arc.

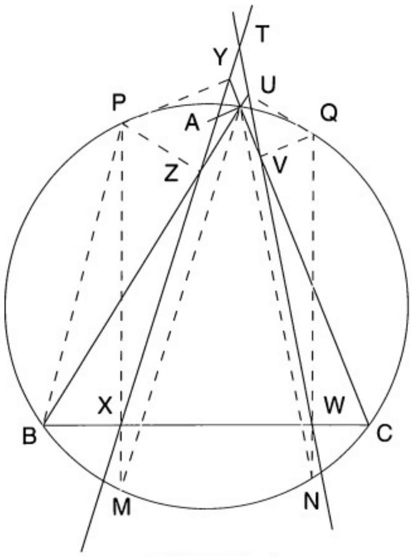

FIGURE 3-26

Here is an interesting application of Simson's theorem to an earlier dilemma. Recall the proof of the fallacy that all scalene triangles are isosceles in Chapter 1. Perpendiculars were drawn from point G to $\overleftrightarrow{AC}$, $\overleftrightarrow{BC}$, and $\overleftrightarrow{AB}$, meeting these lines at points D, F, and E, respectively. Because point G is on the circumcircle of $\triangle ABC$, Simson's theorem establishes that points D, E, and

F are collinear. The famous postulate by Moritz Pasch (1843–1930) states that a straight line intersecting one side of a triangle (internally) must intersect exactly one of the other two sides (internally) except if the line contains a vertex of the triangle. Euclid had quietly assumed this idea. Yet with this postulate available to us, we are assured that the two critical perpendiculars can neither both fall inside nor both fall outside the triangle, which enables us to avoid the fallacious proof previously offered.

RADICAL AXES

Earlier in this chapter, we stated Brianchon's theorem as the dual of Pascal's theorem. At that juncture we deferred the proof because we needed some more knowledge about a radical axis. We will now establish some important properties of radical axes and then use them to prove Brianchon's theorem.

Consider two circles R and Q (see Figure 3-27) intersecting at points A and B. P is any point on $\overleftrightarrow{AB}$ not between points A and B. $\overline{PT}$ and $\overline{PS}$ are tangents to circles R and Q at points T and S, respectively.

INTERACTIVE 3-27

Drag points Q, R, and P and note the radical axis.

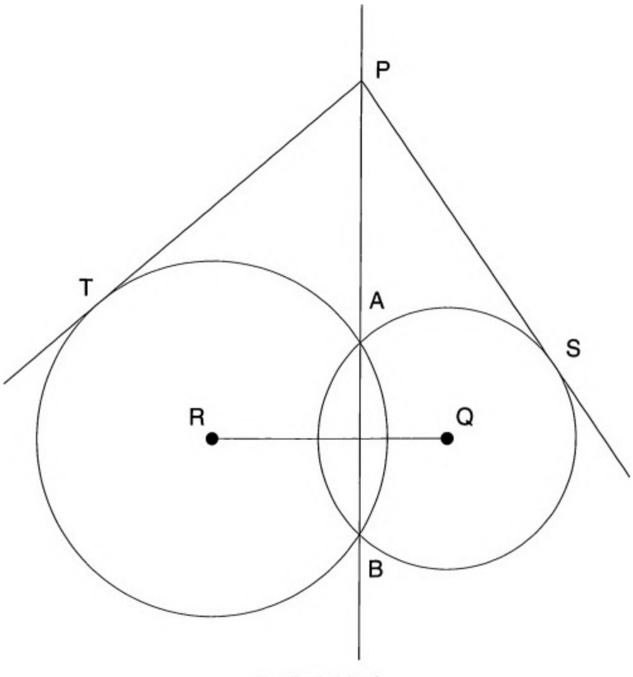

FIGURE 3-27

From our study of elementary geometry, we know that PT is the mean proportional between PB and PA. Therefore $(PT)^2 = (PB)(PA)$. Similarly, for circle Q, $(PS)^2 = (PB)(PA)$. It then follows that $PT = PS$.

Because point P was selected as *any* external point on $\overleftrightarrow{AB}$, we can conclude that from any external point on $\overleftrightarrow{AB}$, tangent segments to circles R and Q are congruent. Before we can state this as a locus theorem, we must prove that any point P that generates congruent tangents to circles R and Q must lie on $\overleftrightarrow{AB}$.

Suppose P is any point where tangent segments $\overline{PT}$ and $\overline{PS}$ are congruent. Let $\overrightarrow{PA}$ intersect circle R at point B and circle Q at point B'. As before, $(PB)(PA) = (PT)^2$ and $(PB')(PA) = (PS)^2$. Because $PT = PS$, $PB = PB'$. Therefore B and B' must coincide, and P lies on the common secant, $\overleftrightarrow{PA}$, of the two circles. We call the line consisting of points that are common endpoints of congruent tangent segments to two circles the *radical axis* of the two circles.

We now state this result as our next theorem.

THEOREM 3.11 The radical axis of two intersecting circles is their common secant.

It follows immediately that the radical axis of two tangent circles is their common tangent. Before we can investigate the radical axis of two nonintersecting circles, we need to consider the following theorem.

THEOREM 3.12 The locus of a point the difference of whose distances squared from two fixed points is a constant is a line perpendicular to the segment determined by the two fixed points.

Proof Let R and Q be the fixed points and let P be a point on the locus (see Figure 3-28). Draw $\overline{PR}$ and $\overline{PQ}$. Construct $\overline{PN} \perp \overline{RQ}$. We use the Pythagorean theorem to get:

$$(PR)^2 - (RN)^2 = (PN)^2 \quad \text{and} \quad (PQ)^2 - (QN)^2 = (PN)^2$$

Therefore:

$$(PR)^2 - (RN)^2 = (PQ)^2 - (QN)^2 \quad \text{or} \quad (PR)^2 - (PQ)^2 = (RN)^2 - (QN)^2 = k$$

Let $RQ = d$. Then, by factoring the last equality, we have:

$$(RN + QN)(RN - QN) = k$$

$$d(RN - QN) = k$$

$$RN - QN = \frac{k}{d} \qquad \text{(I)}$$

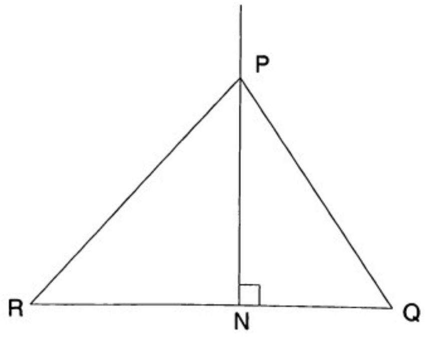

FIGURE 3-28

Remember that:

$$RN + QN = d \tag{II}$$

Solving equations (I) and (II) simultaneously, we get:

$$RN = \frac{d^2 + k}{2d} \quad \text{and} \quad QN = \frac{d^2 - k}{2d}$$

This fixes the position of point N.

Because d and k are constant for any given situation, point P must lie on the line perpendicular to $\overline{RQ}$ at point N, which divides $\overline{RQ}$ in the ratio:

$$\frac{RN}{QN} = \frac{d^2 + k}{d^2 - k}$$

We can conclude this locus proof by showing that any point on $\overleftrightarrow{PN}$ satisfies the given conditions. This is left to the reader. ●

Theorem 3.12 enables us to continue our study of radical axes. We must now determine the radical axis of two nonintersecting circles. Our intuition would probably predict the next theorem.

THEOREM 3.13 The radical axis of two nonintersecting circles is a line perpendicular to their line of centers.

Proof Begin by letting r and q be the radii of circles R and Q, respectively. Let P be a point on the required locus, that is, so that tangent segments $\overline{PT}$ and $\overline{PS}$ are congruent (see Figure 3-29).

By applying the Pythagorean theorem to $\triangle PTR$ and $\triangle PSQ$, we get:

$$(PR)^2 - r^2 = (PT)^2 \quad \text{and} \quad (PQ)^2 - q^2 = (PS)^2$$

But $PT = PS$; therefore:

$$(PR)^2 - r^2 = (PQ)^2 - q^2 \quad \text{or} \quad (PR)^2 - (PQ)^2 = r^2 - q^2$$

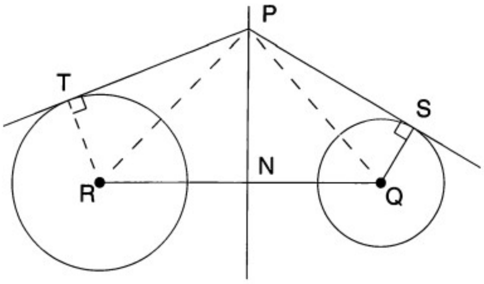

FIGURE 3-29

Because the right-hand side of this equality is a constant, we can conclude (by Theorem 3.12) that the locus of P is the line containing point P, which is perpendicular to the line of centers $\overleftrightarrow{RQ}$. ●

In a manner similar to that used in the previous proof, we can determine the location of point N in terms of the radii and the distance between the centers. As a direct consequence of Theorem 3.13, we have the following theorem.

THEOREM 3.14 The radical axes of three given circles whose centers are not collinear are concurrent.

Proof Let us consider circles R, Q, and U whose radical axes are $\overleftrightarrow{AB}$, $\overleftrightarrow{CD}$, and $\overleftrightarrow{EF}$ (see Figure 3-30).

Let point P be the intersection of $\overleftrightarrow{AB}$ and $\overleftrightarrow{CD}$. Using radical axis $\overleftrightarrow{AB}$ of circles R and Q, we have $PT = PS$. Using radical axis $\overleftrightarrow{CD}$ of circles Q and U, we have $PV = PS$. (Note: $\overleftrightarrow{PT}$, $\overleftrightarrow{PS}$, and $\overleftrightarrow{PV}$ are tangents to the given circles.)

Thus $PT = PV$, which indicates that point P must lie on the radical axis, $\overleftrightarrow{EF}$, of circles R and U. This proves that the radical axes are concurrent at point P. ●

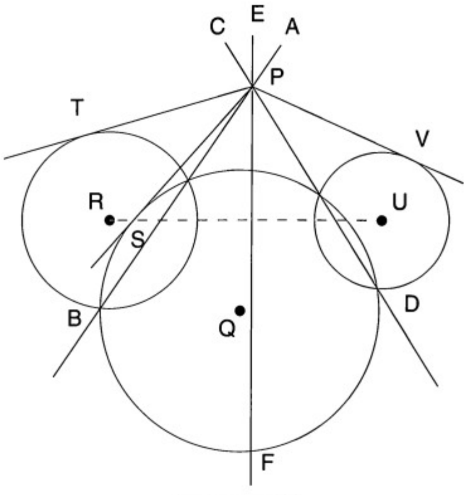

FIGURE 3-30

We are now ready to prove Brianchon's theorem, which we discussed earlier in the chapter (on page 58). The proof we will use is by A. S. Smogorzhevskii (*The Ruler in Geometrical Constructions,* New York: Blaisdell Publishing Company, 1961, pp. 33–35).

THEOREM 3.5 **(Brianchon's theorem)** If a hexagon is circumscribed about a circle, the lines containing opposite vertices are concurrent.

Proof As seen in Figure 3-31, the sides of hexagon $ABCDEF$ are tangent to a circle at points T, N, L, S, M, and K. Points K', L', N', M', S', and T' are chosen on $\overrightarrow{FA}$, $\overrightarrow{DC}$, $\overrightarrow{BC}$, $\overrightarrow{FE}$, $\overrightarrow{DE}$, and $\overrightarrow{BA}$, respectively, so that:

$$KK' = LL' = NN' = MM' = SS' = TT'$$

Now construct circle P tangent to $\overrightarrow{BA}$ and $\overrightarrow{DE}$ at points T' and S', respectively (the existence of this circle is easily justified). Similarly, construct circle Q tangent to $\overrightarrow{FA}$ and $\overrightarrow{DC}$ at points K' and L', respectively. Then construct circle R tangent to $\overrightarrow{FE}$ and $\overrightarrow{BC}$ at points M' and N', respectively.

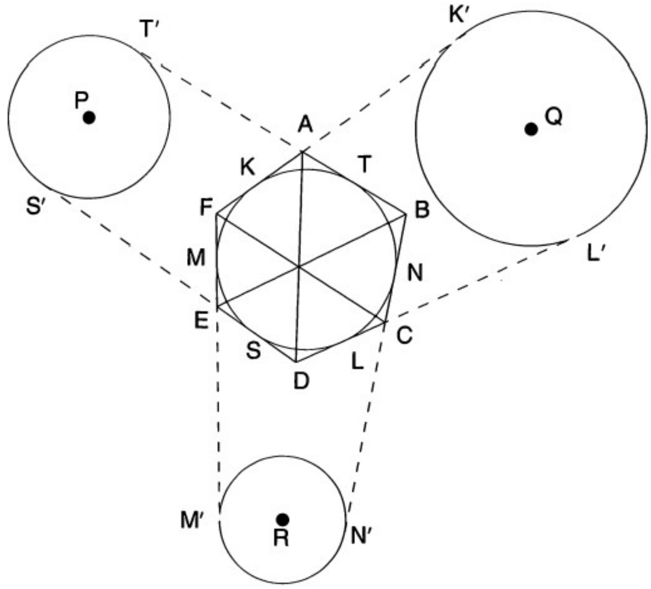

FIGURE 3-31

Because two tangent segments to a circle from an external point have the same length, $FM = FK$. We already know that $MM' = KK'$. Therefore, by addition:

$$FM' = FK'$$

Similarly:

$$CL = CN \quad \text{and} \quad LL' = NN'$$

By subtraction:

$$CL' = CN'$$

We now notice that points F and C are each endpoints of a pair of congruent tangent segments to circles R and Q. Thus these points determine the radical axis, $\overleftrightarrow{CF}$, of circles R and Q. Using the same technique, we can easily show that $\overleftrightarrow{AD}$ is the radical axis of circles P and Q and that $\overleftrightarrow{BE}$ is the radical axis of circles P and R.

We proved that the radical axes of three circles with noncollinear centers (taken in pairs) are concurrent (Theorem 3.14). Therefore $\overleftrightarrow{CF}$, $\overleftrightarrow{AD}$, and $\overleftrightarrow{BE}$ are concurrent.

We should note that the only way in which these circles would have had collinear centers is if the diagonals were to have coincided, which is impossible.

EXERCISES

1. Sides $\overleftrightarrow{AB}$, $\overleftrightarrow{BC}$, $\overleftrightarrow{CD}$, and $\overleftrightarrow{DA}$ of quadrilateral $ABCD$ are intersected by a straight line at points K, L, M, and N, respectively. Prove that $\frac{BL}{LC} \cdot \frac{AK}{KB} \cdot \frac{DN}{NA} \cdot \frac{CM}{MD} = 1$.

2. Side $\overline{AB}$ of square $ABCD$ is extended to point P so that $BP = 2(AB)$. With M the midpoint of $\overline{DC}$, $\overline{BM}$ intersects $\overline{AC}$ at point Q. Also, $\overline{PQ}$ intersects $\overline{BC}$ at point R. Use Menelaus's theorem to find the numerical value of $\frac{CR}{RB}$.

3. Points P and R are on sides $\overline{AB}$ and $\overline{AC}$, respectively, of $\triangle ABC$ so that $\overline{AP} \cong \overline{AR}$. Prove that median $\overline{AM}$ partitions $\overline{PR}$ into segments proportional to $\overline{AB}$ and $\overline{AC}$.

4. Prove that the tangents to the circumcircle of a triangle at its vertices intersect the opposite sides in three collinear points.

5. Prove that if a line contains the centroid, G, of $\triangle ABC$ and intersects sides $\overleftrightarrow{AB}$ and $\overleftrightarrow{AC}$ at points M and N, respectively, then $(AM)(NC) + (AN)(MB) = (AM)(AN)$.

6. A circle is tangent to side $\overline{BC}$ of $\triangle ABC$ at M, its midpoint, and intersects $\overleftrightarrow{AB}$ and $\overleftrightarrow{AC}$ at points R, R' and S, S', respectively. If $\overline{RS}$ and $\overline{R'S'}$ are each extended to meet $\overleftrightarrow{BC}$ at points P and P', respectively, prove that $(BP)(BP') = (CP)(CP')$ (see Figure 3-32).

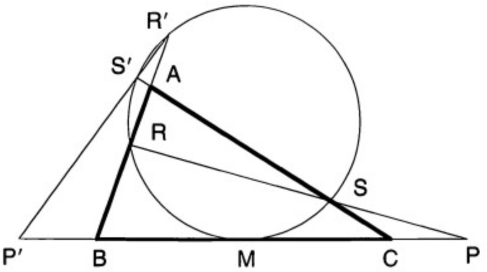

FIGURE 3-32

7. In $\triangle ABC$, P, Q, and R are the midpoints of sides $\overline{AB}$, $\overline{BC}$, and, $\overline{AC}$ respectively. $\overleftrightarrow{AN}$, $\overleftrightarrow{BL}$, and $\overleftrightarrow{CM}$ are concurrent, meeting the opposite sides at points N, L, and M, respectively. If $\overleftrightarrow{PL}$ meets $\overleftrightarrow{BC}$ at point J, $\overleftrightarrow{MQ}$ meets $\overleftrightarrow{AC}$ at point I, and $\overleftrightarrow{RN}$ meets $\overleftrightarrow{AB}$ at point H, prove that points H, I, and J are collinear (see Figure 3-33).

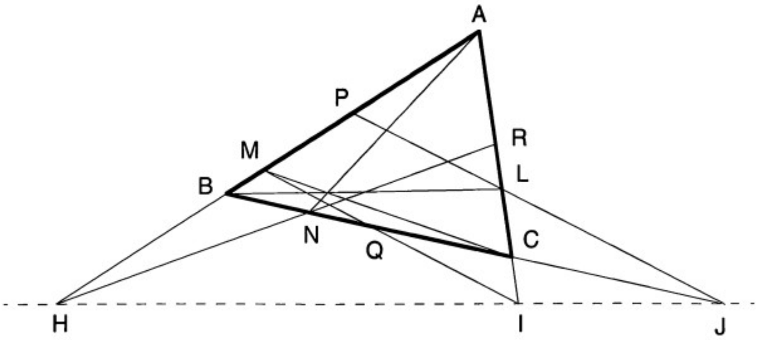

FIGURE 3-33

8. Prove that the three pairs of common external tangents to three circles (no two of which are equal or concentric), taken two at a time, intersect in three collinear points.

9. Prove that the perpendicular bisectors of the interior angle bisectors of any triangle meet the sides opposite the angles being bisected in three collinear points.

10. Provide a proof for Application 6, using Menelaus's theorem.

11. How can Brianchon's theorem be used to prove the existence of the Gergonne point of a triangle?

12. Compare Pappus's theorem to Pascal's theorem.

13. State and prove the converse of Desargues's theorem.

14. State and prove the converse of Simson's theorem.

15. In Figure 3-34, $\triangle ABC$, with right angle at A, is inscribed in circle O. The Simson line of point P with respect to $\triangle ABC$ meets $\overline{PA}$ at point M. Prove that $\overline{MO}$ is perpendicular to $\overline{PA}$ at point M.

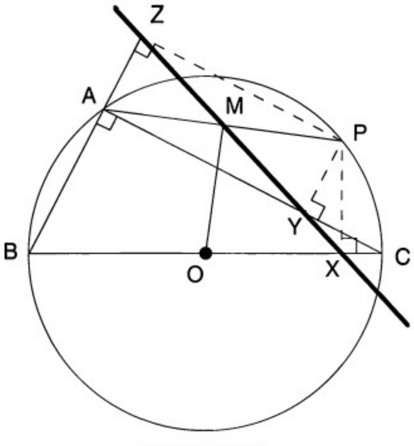

FIGURE 3-34

16. From a point P on the circumference of circle O, three chords are drawn meeting the circle at points A, B, and C. Prove that the three points of intersection of the three circles with $\overline{PA}$, $\overline{PB}$, and $\overline{PC}$ as diameters are collinear.

17. If two triangles are inscribed in the same circle, a single point on the circumcircle determines a Simson line for each triangle. Prove that the angle formed by these two Simson lines is constant, regardless of the position of the point.

18. Prove that the common tangent segments (if they exist) of two given circles are bisected by their radical axis.

19. Prove that the radical axis of the two circles whose diameters are the diagonals of a trapezoid contains the point of intersection of the nonparallel sides of the trapezoid.

20. Prove that the four points determined by the intersections of two secants drawn from a point on the radical axis of two circles with the two circles lie on a third circle.

CHAPTER 5

HISTORY OF THE PARALLEL POSTULATE

> Like the goblin "Puck," [the feat of proving Euclid V] has led me "up and down, up and down," through many a wakeful night: but always, just as I thought I had it, some unforeseen fallacy was sure to trip me up, and the tricksy sprite would "leap out, laughing ho, ho, ho!"
>
> C. L. DODGSON (LEWIS CARROLL)

Let us summarize what we have done so far. We have discovered certain gaps in Euclid's definitions and postulates for plane geometry. We filled in these gaps and firmed up the foundations for this geometry by presenting (a modified version of) Hilbert's definitions and axioms. We then built a structure of theorems on these foundations. However, the structure thus far erected does not rest on the parallel postulate, and we called that structure "neutral geometry." One reason we postponed building on the parallel postulate is that we have less confidence in it than in the other axioms.

You may feel that to deny the Euclidean parallel postulate would go against common sense. Albert Einstein once said that "common sense is, as a matter of fact, nothing more than layers of preconceived notions stored in our memories and emotions for the most part before age eighteen."

That Euclid himself did not quite trust this postulate is shown by the fact that he postponed using it in a proof for as long as possible — until his 29th proposition. In this chapter, we will examine a few illuminating attempts to prove Euclid's parallel postulate (many other attempts are presented in Bonola, 1955; Gray, 1989; and Rosenfeld, 1988). It should be emphasized that most of these attempts were made by outstanding mathematicians, not incompetents. And even though each attempt was flawed, the effort was usually not wasted; for, assuming that all but one step can be justified, when we detect the flawed step, we find another statement which to our surprise is equivalent[1] to the parallel postulate. You have the opportunity to do this enjoyable detective work in Exercises 8 through 13.

PROCLUS

Proclus (A.D. 410–485), whose commentary is one of the main sources of information on Greek geometry, criticized the parallel postulate as follows: "This ought even to be struck out of the Postulates altogether; for it is a theorem involving many difficulties, which Ptolemy, in a certain book, set himself to solve. . . . The statement that since [the two lines] converge more and more as they are produced, they will sometime meet is plausible but not necessary." Proclus offers the example of a hyperbola that approaches its asymptotes as closely as you like without ever meeting them (see Figure 5.1). This example shows that the opposite of Euclid's conclusions can at least be imagined.[2] Proclus says: "It is then clear from this that we must seek a proof of the present theorem, and that it is alien to the special character of postulates."

For over two thousand years some of the best mathematicians tried to prove Euclid's fifth postulate. What does it mean, according to our terminology, to have a proof? It should not be necessary to *assume* the parallel postulate as an axiom; we should be able to prove it *from the other axioms.* If we were able to prove Euclid V in this way, it would become a theorem in neutral geometry and neutral geometry would encompass all of Euclidean geometry.

The first known attempted proof was by Ptolemy. Without going through the details of his argument (see Heath, 1956, pp. 204–206), we might say that he assumed Hilbert's parallel postulate without

---

[1] Actually, the flawed argument only proves that the unjustified statement *implies* the parallel postulate; the converse requires further argument. I do not present any attempts that are uninformative.

[2] Students always object to Figure 5.1 on the grounds that the hyperbola is not "straight." We agreed not to use this word because we don't have a precise definition. A precise definition can be given in differential geometry. See Appendix A.

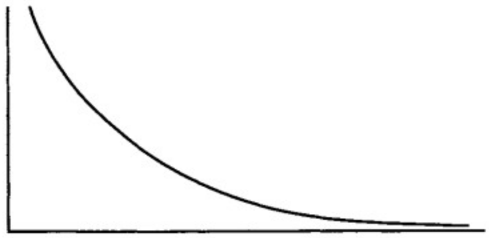

FIGURE 5.1

realizing it. We have seen in Chapter 4 that Hilbert's parallel postulate is logically equivalent to Euclid V (Theorem 4.5), so that Ptolemy was assuming what he was trying to prove; i.e., his reasoning was essentially circular.

Proclus attempted to prove the parallel postulate as follows (see Figure 5.2): Given two parallel lines *l* and *m*. Suppose line *n* cuts *m* at P. We wish to show *n* intersects *l* also (see Proposition 4.7). Let Q be the foot of the perpendicular from P to *l* (Corollary 1 to Theorem 4.1). If *n* coincides with $\overleftrightarrow{PQ}$, then it intersects *l* at Q. Otherwise, one ray $\overrightarrow{PY}$ of *n* lies between $\overrightarrow{PQ}$ and a ray $\overrightarrow{PX}$ of *m*. Take X to be the foot of the perpendicular from Y to *m*.

Now, as the point Y recedes endlessly from P on *n*, the segment XY increases indefinitely in size, and so eventually becomes greater than segment PQ. Therefore, Y must cross over to the other side of *l*, so that *n* must meet *l*.

The preceding paragraph is the heart of Proclus' argument; it is a rather sophisticated argument, involving motion and continuity. Moreover, every step in the argument can be shown to be correct—except that the conclusion doesn't follow! (In Exercise 6 you are asked to prove Aristotle's principle that XY increases indefinitely, where "indefinitely" means "without bound." For example, the sequence of numbers $\frac{1}{2}, \frac{3}{4}, \frac{7}{8}, \frac{15}{16}, \frac{31}{32}, \ldots$ increases but not "indefinitely" in the sense of "without bound," because 1 is a bound for these numbers.)

How could one justify the last step? Let us drop a perpendicular YZ from Y to *l*. You might then say that (1) X, Y, and Z are collinear, and (2) XZ $\cong$ PQ. Thus, when XY becomes greater than PQ, XY must also be greater than XZ, so that Y must be on the other side of *l*. Here the conclusion does indeed follow from statements 1 and 2. The trouble is that there is no justification for these statements!

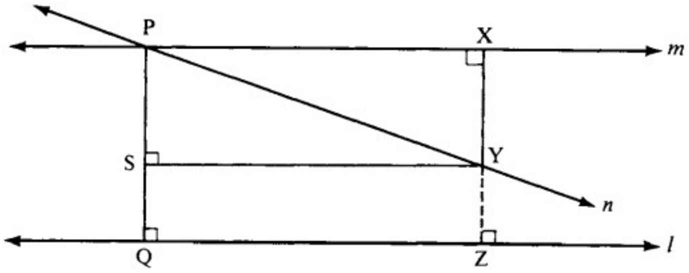

FIGURE 5.2

If this boggles your mind, it may be because Figure 5.2 makes statements 1 and 2 *seem* correct. You recall, however, that we are not allowed to use a diagram to justify a step in a proof. Each step must be proved from stated axioms or previously proven theorems. (We will show later that it is not possible in neutral geometry to prove statements 1 and 2. They can be proved only in Euclidean geometry and only by using the parallel postulate; this makes Proclus' argument circular.)

This analysis of Proclus' faulty argument illustrates how careful you must be in the way you think about parallel lines from now on. You probably visualize parallel lines as railroad tracks, everywhere equidistant from each other, and the ties of the tracks perpendicular to both parallels. This imagery is valid only in Euclidean geometry. Without the parallel postulate, the only thing we can say about two lines that are "parallel" is that, by definition of "parallel," they have no point in common. You can't assume they are equidistant; you can't even assume they have one common perpendicular segment. As Humpty Dumpty remarked: "When I use a word it means what I wish it to mean, neither more nor less."

WALLIS

The next important attempt to prove the parallel postulate was made by the Persian astronomer and mathematician Nasir Eddin al-Tusi (1201 – 1274). But since his attempted proof had several unjustified

John Wallis

assumptions, let us move ahead to John Wallis (1616–1703).[3] Wallis gave up trying to prove the parallel postulate in neutral geometry. Instead, he proposed a new axiom, which he felt was more plausible than the parallel postulate, and then proved the parallel postulate from his new axiom and the other axioms of neutral geometry.

WALLIS' POSTULATE. Given any triangle $\triangle ABC$ and given any segment DE. There exists a triangle $\triangle DEF$ (having DE as one of its sides) that is similar to $\triangle ABC$ (denoted $\triangle DEF \sim \triangle ABC$). (See Figure 5.3.)

---

[3] Wallis was the leading English mathematician before Isaac Newton. In his treatise *Arithmetica infinitorum* (which Newton studied), Wallis introduced the symbol ∞ for "infinity," developed formulas for certain integrals, and presented his famous infinite product formula

$$\frac{\pi}{2}=\frac{2\cdot 2\cdot 4\cdot 4\cdot 6\cdot 6\cdot 8\cdots}{1\cdot 3\cdot 3\cdot 5\cdot 5\cdot 7\cdot 7\cdots}$$

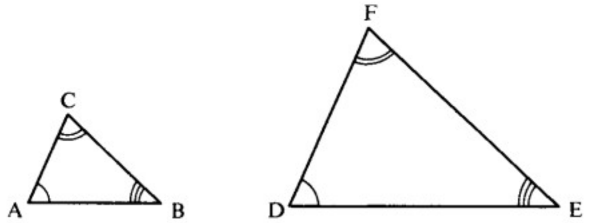

FIGURE 5.3

Similar triangles are triangles whose vertices can be put in one-to-one correspondence so that corresponding angles are congruent. In Euclidean geometry it is proved that corresponding sides of similar triangles are proportional (see Exercise 18); for example, each side of $\triangle$DEF might be twice as long as the corresponding side of $\triangle$ABC. Thus, the intuitive meaning of Wallis' postulate is that you can either magnify or shrink a triangle as much as you like, without distortion.

Using Wallis' postulate, the parallel postulate can be proved as follows (see Figure 5.4):

Proof:

Given a point P not on line *l*, construct one parallel *m* to *l* through P as before—by dropping a perpendicular $\overleftrightarrow{PQ}$ to *l* and erecting *m* perpendicular to $\overleftrightarrow{PQ}$. Let *n* be any other line through P. We must show that *n* meets *l*. As before, we consider a ray of *n* emanating from P that is between a ray of *m* and $\overrightarrow{PQ}$; for any point R on this ray, we drop $\overleftrightarrow{RS}$ perpendicular to $\overleftrightarrow{PQ}$ (see Proposition 3.16 and Corollary 1 to Theorem 4.1 for the existence and uniqueness of all our perpendiculars).

We now apply Wallis' postulate to $\triangle$PSR and segment PQ. It tells us that there is a point T such that $\triangle$PSR is similar to $\triangle$PQT.

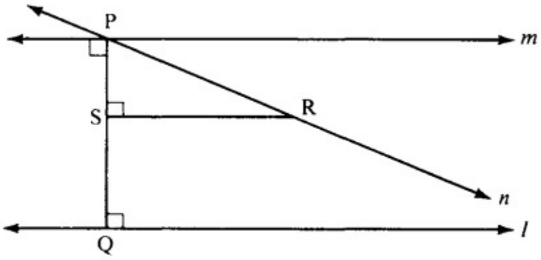

FIGURE 5.4

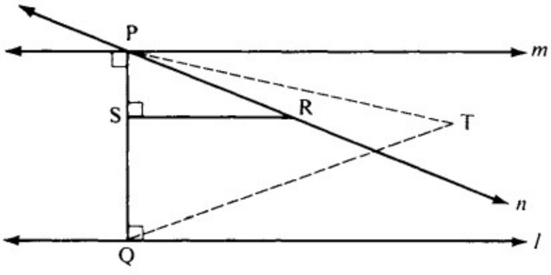

FIGURE 5.5

Assume T lies on the same side of $\overleftrightarrow{PQ}$ as R (Figure 5.5) — if not, reflect across $\overleftrightarrow{PQ}$.

By definition of similar triangles, $\sphericalangle TPQ \cong \sphericalangle RPS$. But since these angles have the ray $\overrightarrow{PQ} = \overrightarrow{PS}$ as a common side and since T lies on the same side of $\overleftrightarrow{PQ}$ as R, the only way they can be congruent is to be equal (Axiom C-4). Thus, $\overrightarrow{PR} = \overrightarrow{PT}$, so that T lies on n. Similarly, $\sphericalangle PQT \cong \sphericalangle PSR$, a right angle; hence, T lies on l. Thus, n and l meet at T; m is therefore the only line through P parallel to l. ■

There is no longer reason to consider Wallis' postulate any more plausible than Euclid V, because it turns out to be logically equivalent to Euclid V (see Exercise 7(a)).

SACCHERI

We next consider the remarkable work of the logician and Jesuit priest Girolamo Saccheri (1667 – 1733). Just before he died he published a little book entitled *Euclides ab omni naevo vindicatus (Euclid Freed of Every Flaw)*, which was not really noticed until a century and a half later, when Eugenio Beltrami rediscovered it.

Saccheri's idea was to use a reductio ad absurdum argument. He assumed the negation of the parallel postulate and tried to deduce a contradiction. Specifically, he studied certain quadrilaterals (Figure 5.6) whose base angles are right angles and whose base-adjacent sides

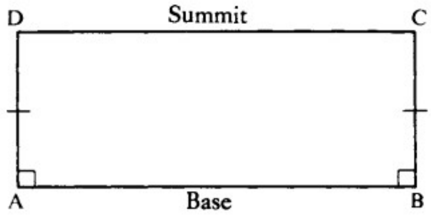

FIGURE 5.6 Saccheri quadrilateral.

are congruent to each other. These quadrilaterals have subsequently become known as *Saccheri quadrilaterals* (though they were studied many centuries earlier by the poet Omar Khayyám and by Nasir Eddin al-Tusi). It is easy to prove in neutral geometry that the summit angles are congruent (Exercise 1), i.e., $\sphericalangle C \cong \sphericalangle D$.

There are three possible cases:

Case 1: The summit angles are right angles.
Case 2: The summit angles are obtuse.
Case 3: The summit angles are acute.

Wanting to prove the first case, which is the case in Euclidean geometry, Saccheri tried to show that the other two cases led to contradictions. He succeeded in showing that case 2 leads to a contradiction: if the summit angles were obtuse, the angle sum of the quadrilateral would be more than 360°, contradicting Corollary 2 to the Saccheri-Legendre theorem (to verify the hypothesis of Corollary 2, see Exercise 17).

However hard he tried, he could not squeeze a contradiction out of case 3, "the inimical acute angle hypothesis," as he called it. He was able to deduce many strange results,[4] but not a contradiction. Finally, he exclaimed in frustration: "The hypothesis of the acute angle is absolutely false, because [it is] repugnant to the nature of the straight line!" It is as if a man had discovered a rare diamond, but, unable to believe what he saw, announced it was glass. Although he did not recognize it, Saccheri had discovered non-Euclidean geometry.

---

[4] See the translation of Saccheri's treatise by G. B. Halsted (Saccheri, 1970).

CLAIRAUT

Alexis Claude Clairaut (1713–1765) was a leading French geometer. Like Wallis, he did not try to prove the parallel postulate in neutral geometry but replaced it in his 1741 text *Éléments de géometrie* with another axiom.

CLAIRAUT'S AXIOM. Rectangles exist.

One can argue that Euclid V is not obvious because one might have to travel very far indeed to verify that the "physical lines" predicted to meet by that postulate actually do meet. According to Saccheri, it suffices to show the existence of one rectangle, which could be quite "small." Clairaut made that his axiom, arguing that "we observe rectangles all around us in houses, gardens, rooms, walls." So why didn't that settle the matter? Perhaps because the game of trying to prove Euclid V had been going on for so many centuries that it became a challenging obsession for mathematicians. Or did mathematicians finally recognize that geometry was not about "physical space"? After all, if you believe that a rectangle can be drawn on the ground, then you cannot also believe that the earth is spherical, because rectangles do not exist on a sphere. If you think you have drawn a "physical rectangle," you could be mistaken because exact measurements are physically impossible. Or did it finally dawn on mathematicians that any postulate proposed to replace Euclid V—no matter how intuitively appealing—was logically equivalent to Euclid V and therefore nothing was gained *logically* by the replacement?

Let us prove that *Clairaut's axiom is logically equivalent in neutral geometry to the parallel postulate.*

Proof:
If we assume the latter, then the existence of rectangles follows easily from Proposition 4.11 and Theorem 4.7. Conversely, assume Clairaut's axiom. Then by Theorem 4.7, all triangles have angle sum 180°, and by introducing a diagonal, all convex quadrilaterals have angle sum 360°. Return to Proclus' argument as illustrated in Figure 5.2. Let S be the foot of the perpendicular from Y to $\overleftrightarrow{PQ}$. S is on the same side of m as Y and Q because $\overleftrightarrow{SY}$ is parallel to m

(Corollary 1 to Theorem 4.1). Moreover, □PXYS, which has three right angles, is now known to be a rectangle. You can easily prove (Exercise 4) that opposite sides of a rectangle are congruent, so $PS \cong XY$. By Aristotle's axiom (Chapter 3), Y can be chosen on the given ray of n so that $XY > PQ$. Then $PS > PQ$ and $P * Q * S$. As above, Y is on the same side of l as S, hence on the opposite side of l from P. Therefore l meets n at some point between P and Y. ■

LEGENDRE

Legendre did not know of Saccheri's work and rediscovered Saccheri's theorems in neutral geometry that are our Theorems 4.4 and 4.7. Legendre certainly knew of Clairaut's text and rejected Clairaut's axiom. We already discussed, in Chapter 1, one of Legendre's attempts to prove the parallel postulate, whose flaw we ask you to detect in Exercise 8. Legendre published a collection of his many attempts as late as 1833, the year he died. Here is his attempt to prove that the angle sum of every triangle is 180°. (Using our modification of Proclus' argument above, we could then prove Hilbert's parallel postulate.)

Proof (see Figure 5.7):

Suppose, on the contrary, there exists a triangle $\triangle ABC$ having defect $d \neq 0$. By the Saccheri-Legendre Theorem 4.4, $d > 0$. One of the angles of the triangle—say $\sphericalangle A$—must then be acute (in fact, less than 60°). On the opposite side of $\overleftrightarrow{BC}$ from A, let D be the unique point such that $\sphericalangle DBC \cong \sphericalangle ACB$ and $BD \cong AC$ (Axioms

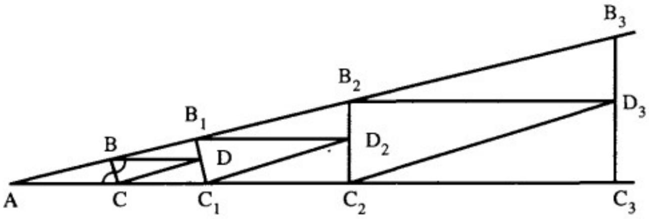

FIGURE 5.7

C-1 and C-4). Then $\triangle ACB \cong \triangle DBC$ (SAS). Also $\overleftrightarrow{BD} \parallel \overleftrightarrow{AC}$ and $\overleftrightarrow{BA} \parallel \overleftrightarrow{DC}$ (by the alternate interior angle theorem, Theorem 4.1), so that D lies in the interior of the acute $\sphericalangle A$. Hence there is a line l through D such that l intersects side $\overrightarrow{AB}$ in a point $B_1 \neq A$ and side $\overrightarrow{AC}$ in a point $C_1 \neq A$. Because of the parallel lines, we know that $B_1 \neq B$ and $C_1 \neq C$.

Suppose B_1 was on segment AB. Then A and B_1 would be on the same side of $\overleftrightarrow{BD}$. Since $\overleftrightarrow{BD} \parallel \overleftrightarrow{AC}$, A and C_1 are on the same side of $\overleftrightarrow{BD}$. Thus B_1 and C_1 are on the same side of $\overleftrightarrow{BD}$ (Axiom B-4). But since D lies in the interior of $\sphericalangle A$, $B_1 * D * C_1$ (Proposition 3.7). This contradiction shows that $A * B * B_1$. Similarly, we have $A * C * C_1$. Since $\triangle ACB \cong \triangle DBC$, the defect of $\triangle DBC$ is also d. Therefore, by the additivity of the defect applied to the four triangles into which $\triangle AB_1C_1$ has been decomposed, the defect of $\triangle AB_1C_1$ is greater than or equal to $2d$.

Repeating this construction for $\triangle AB_1C_1$, we obtain $\triangle AB_2C_2$ with defect greater than or equal to $4d$. Iterating the construction n times, we obtain a triangle with defect greater than or equal to $2^n d$, which can be made as large as we like by taking n sufficiently large. But the defect of a triangle cannot be more than 180°! This contradiction shows that every triangle $\triangle ABC$ has defect 0. ■

Can you see the flaw? It is easy, because we have justified every step but one, the sentence beginning with "Hence." That is the assumption you were warned in Chapter 3 not to make. Legendre made the same error as was made many centuries earlier by Simplicius (Byzantine, sixth century), al-Jawhari (Persian, ninth century), Nasir Eddin al-Tusi, and others. He has failed to prove in neutral geometry that the defect of every triangle is zero. Nevertheless, Legendre has succeeded in proving the following theorem in neutral geometry.

THEOREM 5.1. Hypothesis: For any acute $\sphericalangle A$ and any point D in the interior of $\sphericalangle A$, there exists a line through D and not through A which intersects both sides of $\sphericalangle A$. Conclusion: The angle sum of every triangle is 180°.

You will easily see from the Klein model in Chapter 7 that the hypothesis of Theorem 5.1 fails in non-Euclidean geometry (Figure 7.5). Let us show that *the hypothesis can be proved in Euclidean geometry* (hence this hypothesis is another statement equivalent to Euclid V). Drop a perpendicular from interior point D to one of the sides of ∢A, and let B be the foot of that perpendicular. Since ∢A is acute, $(\sphericalangle A)^\circ + (\sphericalangle DBA)^\circ = (\sphericalangle A)^\circ + 90^\circ < 180^\circ$. So $\overrightarrow{BD}$ meets the other side of ∢A, by Euclid V. ■

LAMBERT AND TAURINUS

Regarding Euclid V, Johann Heinrich Lambert (1728–1777) wrote:

> Undoubtedly, this basic assertion is far less clear and obvious than the others. Not only does it naturally give the impression that it should be proved, but to some extent it makes the reader feel that he is capable

Johann Heinrich Lambert

> of giving a proof, or that he should give it. However, to the extent to which I understand this matter, that is just a *first* impression. He who reads Euclid further is bound to be amazed not only at the rigor of his proofs but also at the delightful simplicity of his exposition. This being so, he will marvel all the more at the position of the fifth postulate when he finds out that Euclid proved propositions that could far more easily be left unproved.

Lambert studied quadrilaterals having at least three right angles, which are now named after him (though they were studied seven centuries earlier by the Egyptian scientist ibn-al-Haytham). You showed in Exercise 19, Chapter 4, that Lambert quadrilaterals exist. A Lambert quadrilateral can be "doubled" (by reflecting it across an included side of two right angles) to obtain a Saccheri quadrilateral. Like Saccheri, Lambert disproved the obtuse angle hypothesis and studied the implications of the "inimical" acute angle hypothesis. He observed that it implied that similar triangles must then be congruent, which in turn implied the existence of an absolute unit of length (see Theorem 6.2, Chapter 6). He called this consequence "exquisite" but did not want it to be true, worrying that the absence of similar, proportional figures "would result in countless inconveniences," especially for astronomers (he did not realize that an elegant non-Euclidean trigonometry could be developed).

He also noticed that the defect of a triangle was proportional to its area (see Chapter 10). He recalled that on a sphere in Euclidean space, the angle sum of a triangle formed by arcs of great circles was greater than 180°, and the excess over 180° of the angle sum of the triangle was proportional to the area of the triangle, the constant of proportionality being the square r^2 of the radius of the sphere (see Rosenfeld, 1988, Chapter 1). If r is replaced by ir ($i = \sqrt{-1}$), squaring introduces a minus sign that converts the excess into the defect in that proportionality. Lambert therefore speculated that the acute angle hypothesis described geometry on a "sphere of imaginary radius."[5]

---

[5] In fact, this idea can be explained in terms of a natural embedding of the non-Euclidean plane in *relativistic* three-space (see Chapter 7). Lambert is known for proving the irrationality of π and of e^x and tan x when x is rational, as well as for important laws he discovered in optics and astronomy. The quote is from B. A. Rosenfeld (1988), p. 100.

Fifty years passed before this brilliant idea was further elaborated in a booklet dated 1826 by F. A. Taurinus, who transformed the formulas of spherical trigonometry into formulas for what he called "log-spherical geometry" by substituting ir for r (his formulas are proved by a different method in Theorem 10.4, Chapter 10). Taurinus vacillated over whether such a geometry actually "existed." He sent a copy of his booklet to C. F. Gauss (see Chapter 6) and later burned the remaining copies in despair when Gauss did not respond.

Lambert cautiously did not submit his *Theory of Parallels* for publication (it was published posthumously in 1786). It contained an erroneous attempt to disprove the acute angle hypothesis. Given line l and distance d, let us call the locus m of all points on a given side of l at perpendicular distance d from l an *equidistant curve*. The flaw in many early attempts to prove the parallel postulate was the tacit assumption that m was a line. Lambert tried to prove this assumption, but he only succeeded in proving that an arc of m could not be a circular arc. Saccheri also erred using differential calculus in his attempt to prove that m was a line.[6]

FARKAS BOLYAI

There were so many attempts to prove Euclid V that by 1763 G. S. Klügel was able to submit a doctoral thesis finding the flaws in 28 different supposed proofs of the parallel postulate, expressing doubt that it could be proved. The French encyclopedist and mathematician J. L. R. d'Alembert called this "the scandal of geometry." Mathematicians were becoming discouraged. The Hungarian Farkas Bolyai wrote to his son János:

> You must not attempt this approach to parallels. I know this way to its very end. I have traversed this bottomless night, which extinguished all light and joy of my life. I entreat you, leave the science of parallels alone. . . . I thought I would sacrifice myself for the sake of the truth. I was ready to become a martyr who would remove the flaw from

---

[6] As an example in elliptic geometry: If l is the equator of the sphere, the equidistant curves are the other circles of latitude.

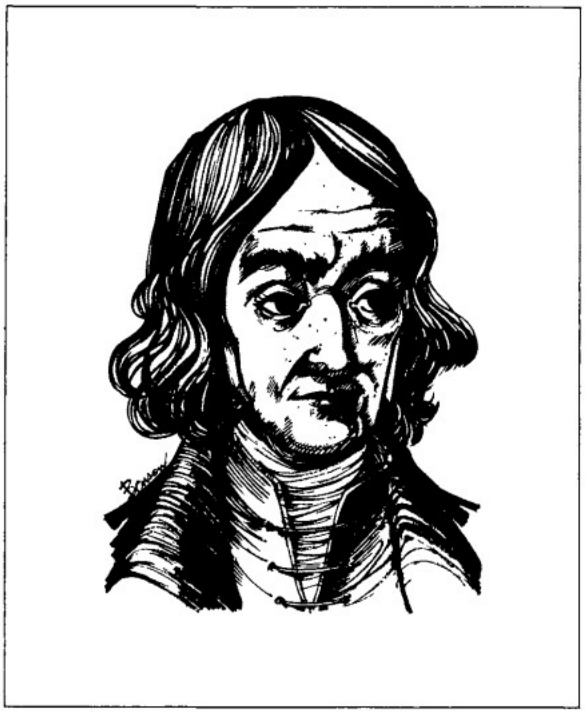

Farkas Bolyai

> geometry and return it purified to mankind. I accomplished monstrous, enormous labors; my creations are far better than those of others and yet I have not achieved complete satisfaction. . . . I turned back when I saw that no man can reach the bottom of the night. I turned back unconsoled, pitying myself and all mankind.
>
> I admit that I expect little from the deviation of your lines. It seems to me that I have been in these regions; that I have traveled past all reefs of this infernal Dead Sea and have always come back with broken mast and torn sail. The ruin of my disposition and my fall date back to this time. I thoughtlessly risked my life and happiness — *aut Caesar aut nihil.*[7]

But the young Bolyai was not deterred by his father's warnings, for he had a completely new idea. He assumed that the negation of

---

[7] The correspondence between Farkas and János Bolyai is from Meschkowski (1964).

Euclid's parallel postulate was not absurd, and in 1823 was able to write to his father:

> It is now my definite plan to publish a work on parallels as soon as I can complete and arrange the material and an opportunity presents itself; at the moment I still do not clearly see my way through, but the path which I have followed gives positive evidence that the goal will be reached, if it is at all possible; I have not quite reached it, but I have discovered such wonderful things that I was amazed, and it would be an everlasting piece of bad fortune if they were lost. When you, my dear Father, see them, you will understand; at present I can say nothing except this: that *out of nothing I have created a strange new universe.* All that I have sent you previously is like a house of cards in comparison with a tower. I am no less convinced that these discoveries will bring me honor than I would be if they were completed.

We will explore this "strange new universe" in the following chapters. A century after János Bolyai wrote this letter, the English physicist J. J. Thomson remarked, somewhat facetiously:

> We have Einstein's space, de Sitter's space, expanding universes, contracting universes, vibrating universes, mysterious universes. In fact, the pure mathematician may create universes just by writing down an equation, and indeed if he is an individualist he can have a universe of his own.

In fact, in 1949 the renowned logician Kurt Gödel found a model of the universe that satisfies Einstein's gravitational equations, one in which it is theoretically possible to travel backward in time![8]

REVIEW EXERCISE

Which of the following statements are correct?

(1) Wallis' postulate implies that there exist two triangles that are similar but not congruent.

---

[8] To date, attempts to refute Gödel's model on either mathematical or philosophical grounds have failed. See "On the paradoxical time-structures of Gödel," by Howard Stein, *Journal of the Philosophy of Science*, v. 37, December 1970, p. 589.

(2) A "Saccheri quadrilateral" is a quadrilateral □ABDC such that ∢CAB and ∢DBA are right angles and AC ≅ BD.

(3) A "Lambert quadrilateral" is a quadrilateral having at least three right angles.

(4) A quadrilateral that is both a Saccheri and a Lambert quadrilateral must be a rectangle.

(5) A hyperbola comes arbitrarily close to its asymptotes without ever intersecting them.

(6) János Bolyai warned his son Farkas not to work on the parallel problem.

(7) Saccheri succeeded in disproving the "inimical" acute angle hypothesis.

(8) In trying to prove Euclid's fifth postulate, Ptolemy tacitly assumed what we have been calling Hilbert's parallel postulate.

(9) It is a theorem in neutral geometry that if $l \parallel m$ and $m \parallel n$, then $l \parallel n$.

(10) It is a theorem in neutral geometry that every segment has a unique midpoint.

(11) It is a theorem in neutral geometry that if a rectangle exists, then the angle sum of any triangle is 180°.

(12) It is a theorem in neutral geometry that if l and m are parallel lines, then alternate interior angles cut out by any transversal to l and m are congruent to each other.

(13) Legendre proved in neutral geometry that for any acute ∢A and any point D in the interior of ∢A, there exists a line through D and not through A which intersects both sides of ∢A.

(14) Clairaut showed that Euclid's fifth postulate could be replaced in the logical presentation of Euclidean geometry by the "more obvious" postulate that rectangles exist, yet mathematicians were not appeased by Clairaut's replacement and they continued to try to prove Euclid V.

EXERCISES

Again, in proofs in Exercises 1 – 17 you are allowed to use only our previous results from neutral geometry.

1. Let □ABDC be a Saccheri quadrilateral, so that ∢B and ∢A are right angles and CA ≅ DB (Figure 5.8). Prove that ∢C ≅ ∢D. (Hint: Prove △CAB ≅ △DBA, then △CDB ≅ △DCA.) Also prove that Saccheri quadrilaterals exist.

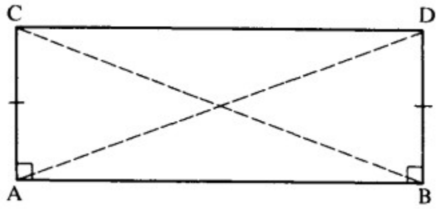

FIGURE 5.8

2. Let □ABDC be a quadrilateral whose base angles ∢A and ∢B are right angles. Prove that if AC < BD, then (∢D)° < (∢C)° (Figure 5.9). (Hint: If AC ≅ BE, with B * E * D, use Exercise 23, Chapter 4, to show that E is interior to ∢ACD, then apply Exercise 1 and the exterior angle theorem.)
3. With the same hypothesis as in Exercise 2, prove the converse, that if (∢D)° < (∢C)°, then AC < BD. (Hint: Assume the contrary, which involves the two cases AC ≅ BD and AC > BD. In each case, derive a contradiction.)
4. The Swiss-German mathematician Lambert considered quadrilaterals with at least three right angles, which are now named after him (Figure 5.10). Prove the following:
   (a) The fourth angle ∢D of a Lambert quadrilateral is never obtuse.
   (b) If ∢D is a right angle, then the opposite sides of □ABCD are congruent (use Exercise 2 and an RAA argument).

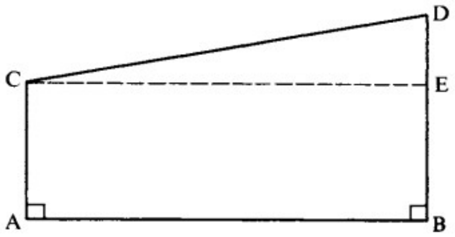

FIGURE 5.9

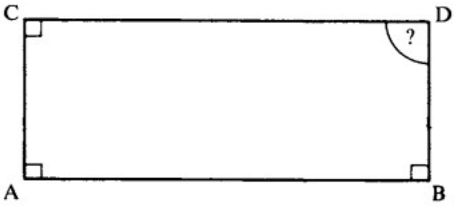

FIGURE 5.10 Lambert quadrilateral.

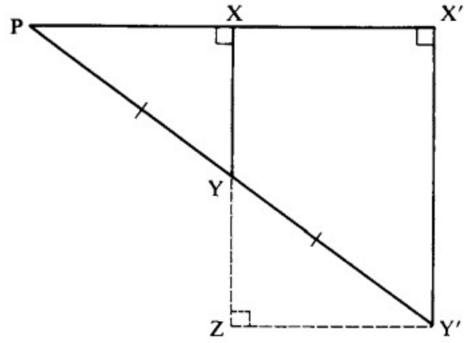

FIGURE 5.11

(c) If ∢D is acute, then each side adjacent to ∢D is greater than its opposite side, i.e., DB > CA and CD > AB (use Exercise 3).

(d) A quadrilateral is both a Lambert and a Saccheri quadrilateral if and only if it is a rectangle.

We can combine statements (a), (b), and (c) of this exercise into the following statement: a side adjacent to the fourth angle of a Lambert quadrilateral is greater than or congruent to its opposite side. As you know, case (b) always holds if the geometry is Euclidean; in the next chapter we will show that case (c) always holds if the geometry is hyperbolic. In elliptic geometry the fourth angle of a Lambert quadrilateral is always obtuse, and a side adjacent to the fourth angle is always smaller than its opposite side.

5. Given a right triangle △PXY with right angle at X, form a new right triangle △PX′Y′ that has acute angle ∢P in common with the given triangle but double the hypotenuse (prove that this can be done); see Figure 5.11. Prove that the side opposite the acute angle is *at least* doubled, whereas the side adjacent to the acute angle is *at most* doubled. (Hint: Extend side XY far enough to drop a perpendicular Y′Z to $\overleftrightarrow{XY}$. Prove that △PXY ≅ △Y′ZY, and apply Exercise 4 to the Lambert quadrilateral □XZY′X′.)
6. Use Exercise 5 to prove Aristotle's axiom (used in Proclus' argument) that as Y recedes endlessly from P, segment XY increases indefinitely (see p. 96). (Hint: Use Archimedes' axiom and the fact that $2^n \to \infty$ as $n \to \infty$.[9]) Does segment PX also increase indefinitely?
7. (a) Prove that Euclid's fifth postulate implies Wallis' postulate (see Figure 5.12). (Hint: Use Axiom C-4 and the fact that in Euclidean

---

[9] Euclid had a version of Archimedes' axiom in Book V, but he presented it as a "definition": If $a < b$, then there is a number n such that $2^n a > b$.

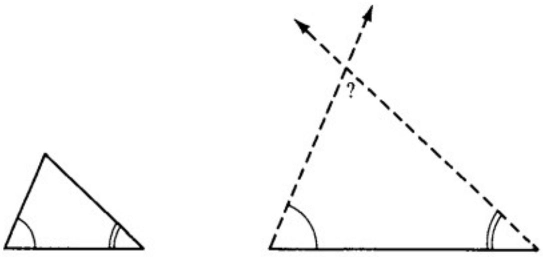

FIGURE 5.12

geometry the angle sum of a triangle is 180° — Proposition 4.11.)

(b) Suppose that in the statement of Wallis' postulate we add the assumption $AB \cong DE$ and replace the word "similar" by "congruent." Prove this new statement in neutral geometry.

8. Reread Legendre's attempted proof of the parallel postulate in Chapter 1. Find the flaw, and justify all the steps that are correct. Prove the flawed statement in Euclidean geometry.
9. Find the unjustified assumption in the following "proof" of the parallel postulate by Farkas Bolyai (see Figure 5.13). Given P not on line l, $\overleftrightarrow{PQ}$ perpendicular to l at Q, and line m perpendicular to $\overleftrightarrow{PQ}$ at P. Let n be any line through P distinct from m and $\overleftrightarrow{PQ}$. We must show that n meets l. Let A be any point between P and Q. Let B be the unique point such that

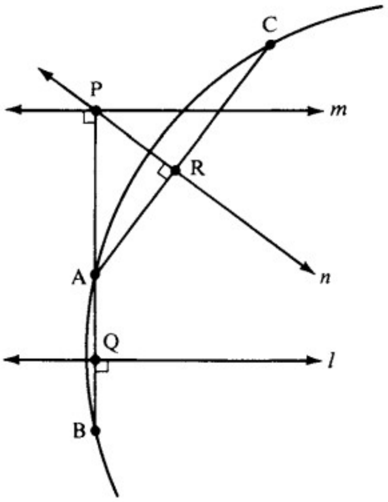

FIGURE 5.13

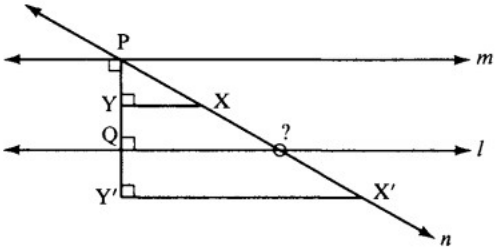

FIGURE 5.14

$A * Q * B$ and $AQ \cong QB$. Let R be the foot of the perpendicular from A to n. Let C be the unique point such that $A * R * C$ and $AR \cong RC$. Then A, B, and C are not collinear (else $R = P$); hence there is a unique circle γ passing through them. Since l is the perpendicular bisector of chord AB of γ and n is the perpendicular bisector of chord AC of γ, l and n meet at the center of γ (Exercise 17(b), Chapter 4).

10. The following attempted proof of the parallel postulate is similar to Proclus' but the flaw is different; detect the flaw with the help of Exercise 5. (See Figure 5.14.) Given P not on line l, $\overleftrightarrow{PQ}$ perpendicular to l at Q, and line m perpendicular to $\overleftrightarrow{PQ}$ at P. Let n be any line through P distinct from m and $\overleftrightarrow{PQ}$. We must show that n meets l. Let $\overrightarrow{PX}$ be a ray of n between $\overleftrightarrow{PQ}$ and a ray of m, and let Y be the foot of the perpendicular from X to $\overleftrightarrow{PQ}$. As X recedes endlessly from P, PY increases indefinitely. Hence, Y eventually reaches a position Y′ on $\overrightarrow{PQ}$ such that $PY' > PQ$. Let X′ be the corresponding position reached by X on line n. Now X′ and Y′ are on the same side of l because $\overleftrightarrow{X'Y'}$ is parallel to l. But Y′ and P are on opposite sides of l. Hence, X′ and P are on opposite sides of l, so that segment PX′ (which is part of n) meets l.
11. Find the flaw in the following attempted proof of the parallel postulate given by J. D. Gergonne (see Figure 5.15). Given P not on line l, $\overleftrightarrow{PQ}$ perpendicular to l at Q, line m perpendicular to $\overleftrightarrow{PQ}$ at P, and point $A \neq P$ on m. Let $\overrightarrow{PB}$ be the last ray between $\overrightarrow{PA}$ and $\overrightarrow{PQ}$ that intersects l, B being the point of intersection. There exists a point C on l such that $Q * B * C$ (Axioms B-1 and B-2). It follows that $\overrightarrow{PB}$ is not the last ray between $\overrightarrow{PA}$ and $\overrightarrow{PQ}$ that intersects l, and hence all rays between $\overrightarrow{PA}$ and $\overrightarrow{PQ}$ meet l. Thus m is the only parallel to l through P.
12. Legendre made another attempt to prove that the defect of every triangle is zero, as follows. In any triangle $\triangle ABC$, if we are given the numbers α, β that measure the angles at A and B, respectively, and the number x that measures the length of the included side AB, then by ASA the number γ that measures the third angle is uniquely determined, so we can write $\gamma = f(\alpha, \beta, x)$. Now if right angles are measured by the unit 1,

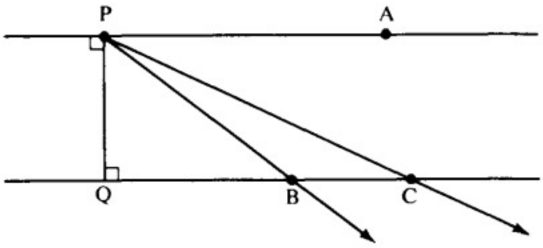

FIGURE 5.15

γ is some number between 0 and 2. But x is not a dimensionless number, since it depends on the arbitrary unit of length (e.g., inch, foot, or meter). Thus the formula for γ cannot actually contain x, and so $\gamma = f(\alpha, \beta)$. (If we knew that the geometry was Euclidean, we would have $f(\alpha, \beta) = 2 - \alpha - \beta$.) Now let D be the midpoint of AB. By Axiom C-4, there is a unique ray $\overrightarrow{DE}$ with E on the same side of $\overleftrightarrow{AB}$ as C such that $\sphericalangle ADE \cong \sphericalangle B$. Then $\overleftrightarrow{DE} \parallel \overleftrightarrow{BC}$, by corresponding angles (Exercise 32, Chapter 4), so, by Pasch's theorem, we can choose label E for a point on AC. Since $\gamma = f(\alpha, \beta)$ for any triangle, $\sphericalangle AED \cong \sphericalangle C$ (i.e., $\triangle ADE$ is similar to $\triangle ABC$). Hence $\square$CEDB has angle sum 360°, so $\triangle CEB$ and $\triangle BED$ have angle sum 180°. By Theorem 4.7, all triangles have defect zero. What is the flaw? (Hint: See the remarks following Theorem 6.2.)

13. It was stated at the beginning of this chapter that if all steps but one of an attempt to prove the parallel postulate are correct, then the flawed step yields another statement equivalent to Hilbert's parallel postulate. Show that for Proclus' attempt, that statement is: Given parallel lines l, m having a common perpendicular and a point Y not lying on l or m, if X (resp. Z) is the foot of the perpendicular from Y to l (resp. to m), then X, Y, and Z are collinear. (Hint: Use Exercise 4b.)
14. Prove that if the defect of every triangle is 0 then Hilbert's parallel postulate holds (this is the converse to Proposition 4.11). (Hint: See the discussion of Clairaut's axiom.)
15. You will show in Exercise 24 that the following statement can be proved in Euclidean geometry: If points P, Q, R lie on a circle with center O, and if $\sphericalangle PQR$ is acute, then $(\sphericalangle PQR)^\circ = \frac{1}{2}(\sphericalangle POR)^\circ$. In neutral geometry, show that this statement implies the existence of a triangle of defect zero (hence, by Prop. 4.7 and Clairaut's axiom, the geometry is Euclidean).
16. *(Difficult)* Here is Legendre's desperate attempt in the 12th and final edition of his geometry text to prove that the defect d of any triangle $\triangle ABC$ is zero. Label so that $\overline{AC} \geq \overline{AB}$. Then the method of Exercise 15,

Chapter 4 (invented by Legendre to prove Theorem 4.4), gives us $\triangle AEC$ having the same defect d (where $AE = 2AD$, D being the midpoint of BC). From $\overline{AC} \geq \overline{AB}$ it is not difficult to show that $\sphericalangle EAC \leq \sphericalangle EAB$, so that $\sphericalangle EAC \leq \frac{1}{2}\sphericalangle BAC$. If $\overline{AC} \geq \overline{AE}$, set $B_1 = E$ and $C_1 = C$; otherwise, reflect $\triangle AEC$ across the angle bisector of $\sphericalangle EAC$ to obtain $\triangle AB_1C_1$ with defect d, $\sphericalangle B_1AC_1 \leq \frac{1}{2}\sphericalangle BAC$, C_1 on ray $\overrightarrow{AC}$, and $\overline{AC_1} \geq \overline{AB_1}$. Iterate this construction n times to obtain $\triangle AB_nC_n$ with defect d, $\sphericalangle B_nAC_n \leq 2^{-n}\sphericalangle BAC$, and C_n on ray $\overrightarrow{AC}$. Let $n \to \infty$. Since $\sphericalangle B_nAC_n \to 0$, the triangles $\triangle AB_nC_n$ (which all have defect d) converge in the limit to a degenerate triangle on ray $\overrightarrow{AC}$ having angles 0°, 0°, and 180°. Hence $d = 0$. Criticize this argument (Hint: Show that B_n, $C_n \to \infty$.)

17. Prove that a Saccheri quadrilateral is convex. Prove that a Lambert quadrilateral is a parallelogram and that every parallelogram is convex.

The remaining exercises in this chapter are exercises in Euclidean geometry, which means you are allowed to use the parallel postulate and its consequences already established. We will refer to these results in Chapter 7. You are also allowed to use the following result, a proof of which is indicated in the Major Exercises:

PARALLEL PROJECTION THEOREM. Given three parallel lines l, m, and n. Let t and t' be transversals to these parallels, cutting them in points A, B, and C and in points A′, B′, and C′, respectively. Then $\overline{AB}/\overline{BC} = \overline{A'B'}/\overline{B'C'}$. (Figure 5.16.)

18. *Fundamental theorem on similar triangles.* Given $\triangle ABC \sim \triangle A'B'C'$, i.e., given $\sphericalangle A \cong \sphericalangle A'$, $\sphericalangle B \cong \sphericalangle B'$, and $\sphericalangle C \cong \sphericalangle C'$. Then corresponding sides are proportional, i.e., $\overline{AB}/\overline{A'B'} = \overline{AC}/\overline{A'C'} = \overline{BC}/\overline{B'C'}$. (See Figure 5.17.) Prove the theorem. (Hint: Let B″ be the point on $\overrightarrow{AB}$ such that $AB'' \cong A'B'$, and let C″ be the point on $\overrightarrow{AC}$ such that $AC'' \cong A'C'$. Use

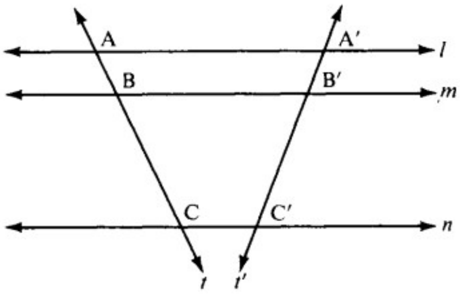

FIGURE 5.16

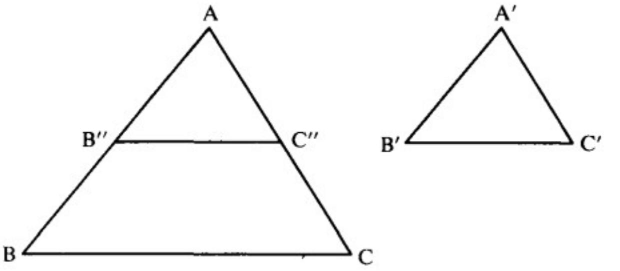

FIGURE 5.17

the hypothesis to show that $\triangle AB''C'' \cong \triangle A'B'C'$ and deduce from corresponding angles that $\overleftrightarrow{B''C''}$ is parallel to $\overleftrightarrow{BC}$. Now apply the parallel projection theorem.)

19. Prove the converse to the fundamental theorem on similar triangles. (Hint: Choose B″ as before. Use Pasch's theorem to show that the parallel to $\overleftrightarrow{BC}$ through B″ cuts AC at a point C″. Then use the hypothesis, Exercise 18, and the SSS criterion to show that $\triangle ABC \sim \triangle AB''C'' \cong \triangle A'B'C'$.)
20. *SAS similarity criterion.* If $\sphericalangle A \cong \sphericalangle A'$ and $\overline{AB}/\overline{A'B'} = \overline{AC}/\overline{A'C'}$, prove that $\triangle ABC \sim \triangle A'B'C'$. (Hint: Same method as in Exercise 19, but using SAS instead of SSS.)
21. Prove the Pythagorean theorem. (Hint: Let CD be the altitude to the hypotenuse; see Figure 5.18. Use the fact that the angle sum of a triangle equals 180° (Exercise 8, Chapter 4) to show that $\triangle ACD \sim \triangle ABC \sim \triangle CBD$. Apply Exercise 18 and a little algebra based on $\overline{AB} = \overline{AD} + \overline{DB}$ to get the result.)
22. The fundamental theorem on similar triangles (Exercise 18) allows the trigonometric functions such as sine and cosine to be defined. Namely, given an acute angle $\sphericalangle A$, make it part of a right triangle $\triangle BAC$ with right angle at C, and set

FIGURE 5.18

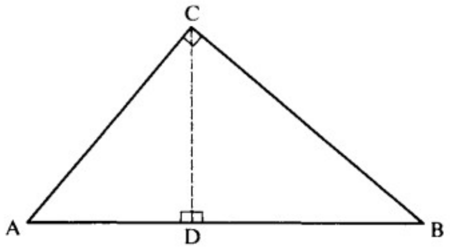

$$\sin \sphericalangle A = (\overline{BC})/(\overline{AB})$$
$$\cos \sphericalangle A = (\overline{AC})/(\overline{AB}).$$

These definitions are then independent of the choice of the right triangle used. If $\sphericalangle A$ is obtuse and $\sphericalangle A'$ is its supplement, set

$$\sin \sphericalangle A = +\sin \sphericalangle A'$$
$$\cos \sphericalangle A = -\cos \sphericalangle A'.$$

If $\sphericalangle A$ is a right angle, set

$$\sin \sphericalangle A = 1$$
$$\cos \sphericalangle A = 0.$$

Now, given any triangle $\triangle ABC$, if a and b are the lengths of the sides opposite A and B, respectively, prove the law of sines,

$$\frac{a}{b} = \frac{\sin \sphericalangle A}{\sin \sphericalangle B}.$$

(Hint: Drop altitude CD and use the two right triangles $\triangle ADC$ and $\triangle BDC$ to show that $b \sin \sphericalangle A = \overline{CD} = a \sin \sphericalangle B$; see Figure 5.19.) Similarly, prove the law of cosines,

$$c^2 = a^2 + b^2 - 2ab \cos \sphericalangle C,$$

and deduce the converse to the Pythagorean theorem.

23. Given A * B * C and point D not collinear with A, B, and C (Figure 5.20). Prove that

$$\frac{\overline{AB}}{\overline{BC}} = \frac{\overline{AD} \sin \sphericalangle ADB}{\overline{CD} \sin \sphericalangle CDB}$$
$$\frac{\overline{AC}}{\overline{BC}} = \frac{\overline{AD} \sin \sphericalangle ADC}{\overline{BD} \sin \sphericalangle BDC}$$

(Hint: Use the law of sines to compute $\overline{AB}/\overline{AD}$, $\overline{CD}/\overline{BC}$, and $\overline{BD}/\overline{BC}$, and remember that $\sin \sphericalangle ABD = \sin \sphericalangle CBD$.)

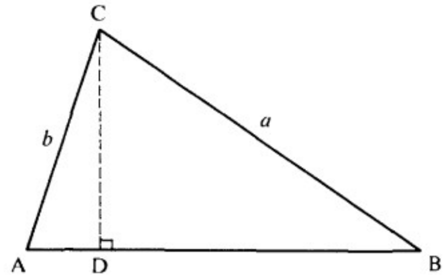

FIGURE 5.19

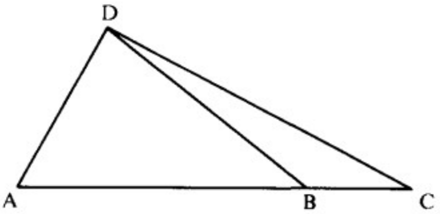

FIGURE 5.20

24. Let γ be a circle with center O, and let P, Q, and R be three points on γ. Prove that if P and R are diametrically opposite, then $\measuredangle$PQR is a right angle, and if O and Q are on the same side of $\overleftrightarrow{PR}$, then $(\measuredangle PQR)^\circ = \frac{1}{2}(\measuredangle POR)^\circ$. (Hint: Again use the fact that the triangular angle sum is 180°. There are four cases to consider, as in Figure 5.21.) State and prove the analogous result when O and Q are on opposite sides of $\overleftrightarrow{PR}$.
25. Prove that if two angles inscribed in a circle subtend the same arc, then they are congruent; see Figure 5.22. (Hint: Apply the previous exercise after carefully defining "subtend the same arc.")
26. Prove that if $\measuredangle$PQR is a right angle, then Q lies on the circle γ having PR as diameter. (Hint: Use uniqueness of perpendiculars and Exercise 24.)

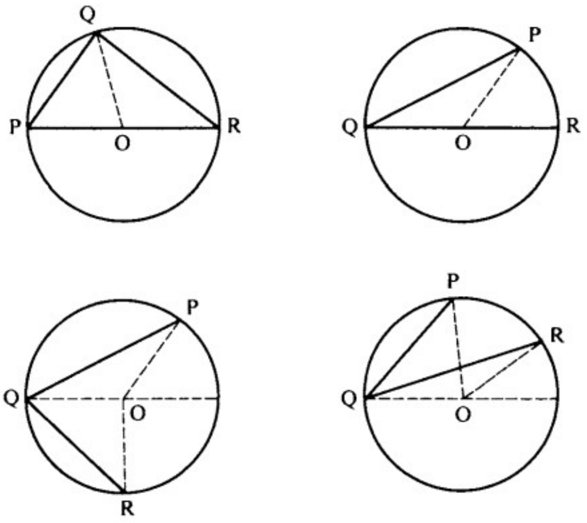

FIGURE 5.21

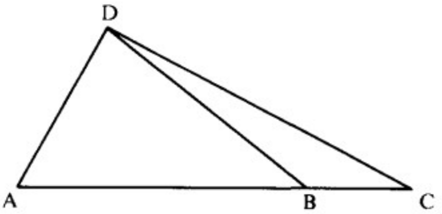

FIGURE 5.20

24. Let γ be a circle with center O, and let P, Q, and R be three points on γ. Prove that if P and R are diametrically opposite, then ∢PQR is a right angle, and if O and Q are on the same side of $\overleftrightarrow{PR}$, then $(∢PQR)^\circ = \frac{1}{2}(∢POR)^\circ$. (Hint: Again use the fact that the triangular angle sum is 180°. There are four cases to consider, as in Figure 5.21.) State and prove the analogous result when O and Q are on opposite sides of $\overleftrightarrow{PR}$.
25. Prove that if two angles inscribed in a circle subtend the same arc, then they are congruent; see Figure 5.22. (Hint: Apply the previous exercise after carefully defining "subtend the same arc.")
26. Prove that if ∢PQR is a right angle, then Q lies on the circle γ having PR as diameter. (Hint: Use uniqueness of perpendiculars and Exercise 24.)

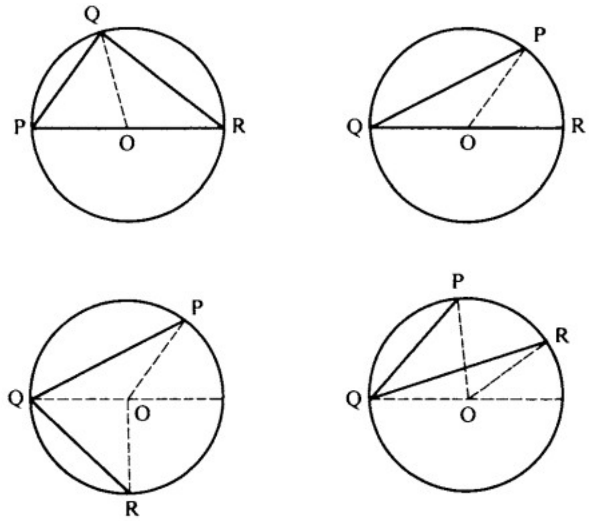

FIGURE 5.21

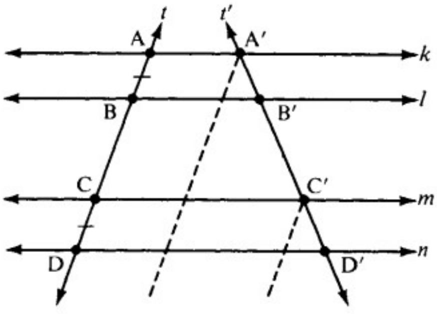

FIGURE 5.23

$x = x'$, since a real number is the least upper bound of all the rational numbers less than it (see any good text on real analysis). To show this, lay off on $\overrightarrow{BA}$ a segment BD of length $p\overline{CB}/q$, and let D′ be the parallel projection of D onto t'. From $p/q < x$, deduce B * D * A. Now apply Major Exercises 3 and 4 to show that $p/q < x'$.[10]

6. Given a segment AB of length a with respect to some unit segment OI (see Theorem 4.3). Using straightedge and compass only, show how to construct a segment of length $\sqrt{a}$. (Hint: Extend AB to a segment AC of length $a + 1$; erect a perpendicular through B and let D be one of its intersections with the circle having AC as diameter; apply the theory of similar triangles to show that $\overline{BD} = \sqrt{a}$. Review the construction in Major Exercise 1, Chapter 1.)
7. Prove that given any line l, two points A and B not on l are on the same side S of l if and only if they lie on a circle contained in S. (Hint: If A and B are on opposite sides of l, apply Major Exercise 1, Chapter 4. If they are on the same side S, let M be the midpoint and m the perpendicular bisector of AB. Any circle through A and B has its center on m. If $\overleftrightarrow{AB} \parallel l$, take any point P between M and the point where m meets l, and use the circle through A, B and P (see Exercise 12, Chapter 6). Otherwise, if A is closer to l than B, let the perpendicular from A to l meet m at O. Show that the circle centered at O with radius OA ≅ OB lies in S. Be sure to indicate where the hypothesis that *the geometry is Euclidean* is used; see Exercise P-20, Chapter 7.)

---

[10] This clever method of proof was essentially discovered by the ancient Greek mathematician Eudoxus—see E. C. Zeeman, "Research, Ancient and Modern," *Bulletin of the Institute of Mathematics and Its Applications*, 10 (1974): 272–281, Warwick University, England.

PROJECTS

1. Since there is no algebra in Euclid's *Elements*, it was quite a feat for Eudoxus to have discovered a purely geometric treatment of proportions and similar triangles. Report on this, using Heath (1956) and Moise (1990, Chapter 20) as references.
2. Eudoxus was also the founder of theoretical astronomy in antiquity (his work later refined by Ptolemy). In his model, the universe was bounded by "the celestial sphere," so that the physical interpretation of Euclid's second and third postulates was false! Even Kepler and Galileo believed in an outer limit to the world. It was René Descartes (1596–1650) who promoted the idea that we live in infinite, unbounded Euclidean space. Report on these issues, using Torretti (1978) as one reference.
3. Report on several other attempts to prove the Euclidean parallel postulate in neutral geometry, using Rosenfeld (1988) as a reference.
4. We remarked after Theorem 4.3 that the definition of π as the ratio of the circumference of any circle to its diameter could only be justified in Euclidean geometry, not neutral geometry. Report on the justification in Moise (1990), section 21.2, which uses the theory of similar triangles.

% ()*+(((, (/ " :) ; ?
@ ? " ,/ G)"A @ " /"

()*+(((, (/ " : /"

% ()*+(((, (: ; % ") % "
% " = " " 9;/ 0 : ; /"

& ' 1)3

The goal of this book is to provide an introduction to both Euclidean and non-Euclidean (hyperbolic) geometry. The two geometries share many features, but they have fundamental and radical differences. Neutral geometry is the part of the path they have in common and that is what we have been studying so far, but we have now come to the fork in the path. That fork comes when you try to answer this question:

> Given a line ℓ and a point P that is not on ℓ, how many lines pass through P and are parallel to ℓ?

Using just the axioms of neutral geometry, we can prove that there is always at least one such parallel. We can also prove that if there is more than one parallel, then there must be infinitely many. But that is the extent of what the neutral axioms can say. The neutral axioms are not enough to determine whether there is one parallel or many. A single axiom is what separates Euclidean and non-Euclidean geometry: the final axiom of Euclidean geometry calls for exactly one parallel, the final axiom of non-Euclidean geometry calls for more than one parallel.

The next several chapters are devoted to Euclidean geometry. Euclidean geometry is very old, and so there is a lot of it. We only cover the fundamentals in this book, but there are many excellent books that do much more, such as *Geometry Revisited* [CG67] by Coxeter and Greitzer.

Part II

Euclidean Geometry

The goal of this book is to provide an introduction to both Euclidean and non-Euclidean (hyperbolic) geometry. The two geometries share many features, but they have fundamental and radical differences. Neutral geometry is the part of the path they have in common and that is what we have been studying so far, but we have now come to the fork in the path. That fork comes when you try to answer this question:

> Given a line ℓ and a point P that is not on ℓ, how many lines pass through P and are parallel to ℓ?

Using just the axioms of neutral geometry, we can prove that there is always at least one such parallel. We can also prove that if there is more than one parallel, then there must be infinitely many. But that is the extent of what the neutral axioms can say. The neutral axioms are not enough to determine whether there is one parallel or many. A single axiom is what separates Euclidean and non-Euclidean geometry: the final axiom of Euclidean geometry calls for exactly one parallel, the final axiom of non-Euclidean geometry calls for more than one parallel.

The next several chapters are devoted to Euclidean geometry. Euclidean geometry is very old, and so there is a lot of it. We only cover the fundamentals in this book, but there are many excellent books that do much more, such as *Geometry Revisited* [CG67] by Coxeter and Greitzer.

11

The Axiom on Parallels

Euclidean geometry is distinguished from neutral geometry by one axiom, one that describes the nature of parallel lines. There are many formulations of the parallel axiom for Euclidean geometry, but the one that gets right to the heart of the matter is *Playfair's axiom*, named after the Scottish mathematician John Playfair.

> *(Playfair's axiom) Let ℓ be a line, and let P be a point that is not on ℓ. Then there is exactly one line through P that is parallel to ℓ.*

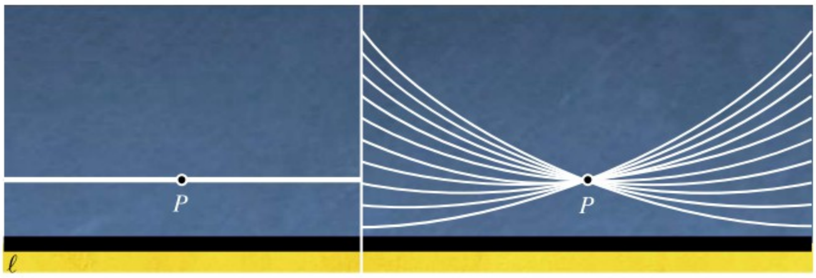

Figure 1. On the left, Euclidean geometry: a single line through P parallel to ℓ. On the right, more than one line through P parallel to ℓ.

In this chapter, we look at a small collection of theorems that are almost immediate consequences of Playfair's axiom, and as such are at the very core of Euclidean geometry. The first of these is Euclid's fifth postulate. This is the controversial postulate in the *Elements*, but also the one that guarantees the same parallel behavior that Playfair's axiom provides. Euclid's fifth is a little unwieldy, particularly when compared to Playfair's axiom, but it is the historical impetus for much of what followed. So let's use Playfair's axiom to prove Euclid's fifth postulate.

Theorem 11.1 *(Euclid's fifth postulate) If lines ℓ_1 and ℓ_2 are crossed by a transversal t, and the sum of adjacent interior angles on one side of t measure less than 180°, then ℓ_1 and ℓ_2 intersect on that side of t.*

Proof First some labels.

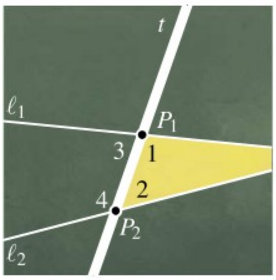

Figure 2. The labels.

Start with lines ℓ_1 and ℓ_2 crossed by transversal t. Label P_1 and P_2, the points of intersection of t with ℓ_1 and ℓ_2 respectively. On one side of t, the two adjacent interior angles should add up to less than 180°. Label the one at P_1 as $\angle 1$ and the one at P_2 at $\angle 2$. Label the supplement of $\angle 1$ as $\angle 3$ and label the supplement of $\angle 2$ as $\angle 4$.

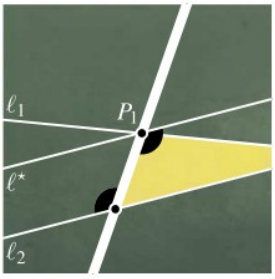

Figure 3. Constructing the unique parallel.

This theorem asserts that the lines intersect, and says where the intersection occurs. We will tackle those two statements in that order. Note that $\angle 1$ and $\angle 4$ are alternate interior angles, but they are not congruent: if they were, their supplements $\angle 2$ and $\angle 3$ would be too, and then

$$(\angle 1) + (\angle 2) = (\angle 1) + (\angle 3) = 180^\circ.$$

There is, however, another line $\ell^\star$ through P_1 that forms an angle congruent to $\angle 4$ (because of the angle construction axiom), and by the alternate interior angle theorem, $\ell^\star$ must be parallel to ℓ_2. Because of Playfair's axiom, $\ell^\star$ is the only parallel to ℓ_2 through P_1. Therefore ℓ_1 intersects ℓ_2.

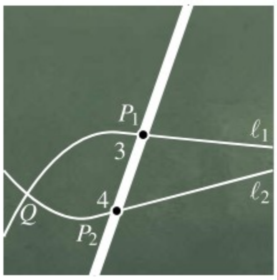

Figure 4. An impossible triangle on the wrong side of t.

The second part of the proof is to figure out on which side of t that ℓ_1 and ℓ_2 cross. Let's see what would happen if they intersected at a point Q on the wrong side of t, the side with $\angle 3$ and $\angle 4$. Then $\triangle P_1 P_2 Q$ would have two interior angles, $\angle 3$ and $\angle 4$, whose measures add up to more than 180°. This would violate the Saccheri-Legendre theorem. So ℓ_1 and ℓ_2 cannot intersect on the side of t with $\angle 3$ and $\angle 4$ and so they must intersect on the side with $\angle 1$ and $\angle 2$.

☐

One of the truly useful theorems of neutral geometry is the alternate interior angle theorem. In fact, we just used it in the last proof. You may recall from high school geometry that the converse of the theorem is often even more useful. The problem is that the converse of the alternate interior angle theorem can't be proved using just the axioms of neutral geometry. It depends on the Euclidean behavior of parallel lines.

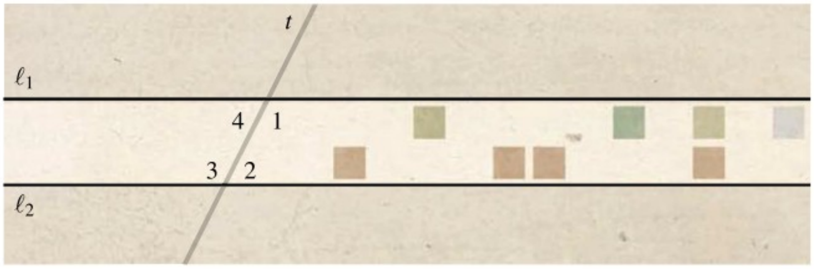

Figure 5. The converse of the alternate interior angle theorem: if $\ell_1 \parallel \ell_2$, then $\angle 1 \simeq \angle 3$ and $\angle 2 \simeq \angle 4$.

Theorem 11.2 *(Converse of the alternate interior angle theorem) If ℓ_1 and ℓ_2 are parallel, then the pairs of alternate interior angles formed by a transversal t are congruent.*

Proof Consider two parallel lines crossed by a transversal. Label adjacent interior angles: $\angle 1$ and $\angle 2$, and $\angle 3$ and $\angle 4$, so that $\angle 1$ and $\angle 4$ are supplementary and $\angle 2$ and $\angle 3$ are supplementary. That means that the pairs of alternate interior angles are $\angle 1$ and $\angle 3$, and $\angle 2$ and $\angle 4$. Now, we just have to do a little arithmetic. From the two pairs of supplementary angles:

(i): $(\angle 1) + (\angle 4) = 180^\circ$
(ii): $(\angle 2) + (\angle 3) = 180^\circ$.

Adding all four angles together,

$$(\angle 1) + (\angle 2) + (\angle 3) + (\angle 4) = 360^\circ.$$

Now we turn to Euclid's fifth postulate, or rather its contrapositive. The lines ℓ_1 and ℓ_2 are parallel, which means that they do not intersect on either side of t. Therefore Euclid's fifth says that on neither side of t may the sum of adjacent interior angles be less than 180°:

- $(\angle 1) + (\angle 2) \geq 180^\circ$
- $(\angle 3) + (\angle 4) \geq 180^\circ$.

If either sum was greater than 180°, though, the sum of all four angles would have to be more than 360°. We saw above that is not the case, so the inequalities are actually equalities:

(iii): $(\angle 1) + (\angle 2) = 180^\circ$
(iv): $(\angle 3) + (\angle 4) = 180^\circ$.

Now we have two systems of equations with four unknowns, and it is basic algebra from here. Subtract *(iv)* from *(i)* to get

$$(\angle 1) - (\angle 3) = 0 \implies (\angle 1) = (\angle 3).$$

Subtract *(i)* from *(iii)* to get

$$(\angle 2) - (\angle 4) = 0 \implies (\angle 2) = (\angle 4).$$

Therefore the alternate interior angles are congruent. □

One of the key theorems we proved in the neutral geometry section was the Saccheri-Legendre theorem: that the angle sum of a triangle is at most 180°. That is all you can say

with the axioms of neutral geometry, but in a world with Playfair's axiom and the converse of the alternate interior angle theorem, we can make a more bold statement about angle sums of triangles.

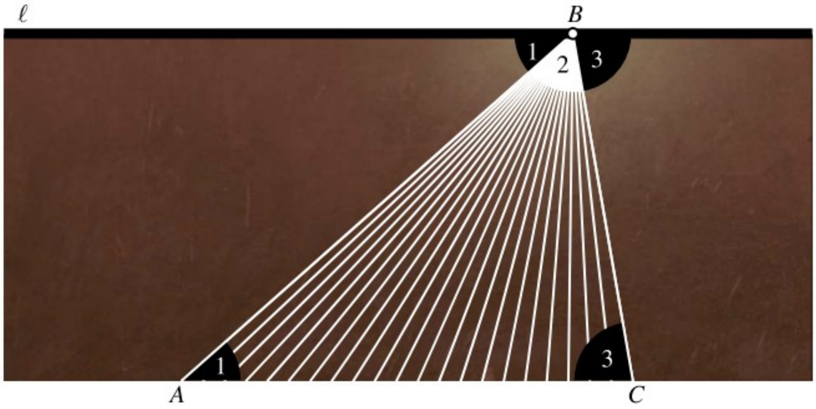

Figure 6. Use the converse of the alternate interior angle theorem to find the angle sum of a triangle.

Theorem 11.3 *The angle sum of a triangle is* 180°.

Proof Given $\triangle ABC$, by Playfair's axiom, there is a unique line ℓ through B that is parallel to $\overline{AC}$. That line and the rays $\overrightarrow{BA}$ and $\overrightarrow{BC}$ form three angles, $\angle 1$, $\angle 2$, and $\angle 3$ as shown in figure 6. By the converse of the alternate interior angle theorem, two pairs of alternate interior angles are congruent: $\angle 1 \simeq \angle A$ and $\angle 3 \simeq \angle C$. Therefore, the angle sum of $\triangle ABC$ is

$$s(\triangle ABC) = (\angle A) + (\angle B) + (\angle C) = (\angle 1) + (\angle 2) + (\angle 3) = 180^\circ. \qquad \square$$

So, the three angles in an equilateral triangle all measure 60° and the measures of the two acute angles in a right triangle must add up to 90°.

In the last chapter on quadrilaterals we talked a little bit about the uncertain status of rectangles in neutral geometry, saying that it is easy to make a convex quadrilateral with three right angles, but that once you have done that, there is no guarantee that the fourth angle will be a right angle. Here it is now in the Euclidean context:

Theorem 11.4 *Let* $\angle ABC$ *be a right angle. Let* r_A *and* r_C *be rays so that* r_A *has endpoint* A*, is on the same side of* $\overline{AB}$ *as* C*, and is perpendicular to* $\overline{AB}$*;* r_C *has endpoint* C*, is on the same side of* $\overline{BC}$ *as* A*, and is perpendicular to* $\overline{BC}$*. Then* r_A *and* r_C *intersect at a point* D*, and the angle formed at this intersection,* $\angle ADC$*, is a right angle. Therefore* $\square ABCD$ *is a rectangle.*

Proof First we need to make sure that r_A and r_C intersect. Let ℓ_A and ℓ_C be the lines containing r_A and r_C respectively. By the alternate interior angle theorem, the right angles at A and B imply that ℓ_A and $\overline{BC}$ are parallel. So $\overline{BC}$ is the one line parallel to ℓ_A through C, and that implies that ℓ_C cannot be parallel to ℓ_A: it has to intersect ℓ_A. Let's call that point of intersection D. This

theorem claims that it is the rays, not the lines, that intersect. Therefore we need to rule out the possibility that the intersection of ℓ_A and ℓ_C might happen on one (or both) of the opposite rays. Since ℓ_A is parallel to $\overline{BC}$, all the points of ℓ_A are on the same side of $\overline{BC}$ as A. None of the points of r_C^{op} is on that side of BC, so D cannot be on r_C^{op}. Likewise, all the points of ℓ_C are on the same side of $\overline{AB}$ as C. None of the points of r_A^{op} is on that side of AB, so D cannot be on r_A^{op}.

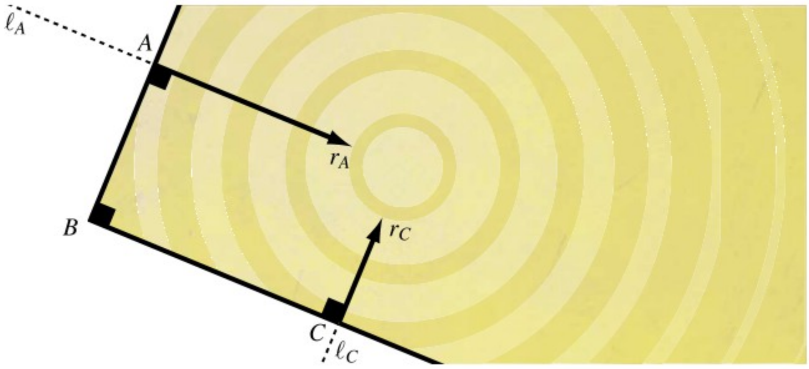

Figure 7. Constructing a rectangle.

So now $\square ABCD$ has three right angles, $\angle A$, $\angle B$, and $\angle C$. It is actually a convex quadrilateral too (verifying this is left as an exercise), so the diagonal AC divides $\square ABCD$ into two triangles $\triangle ABC$ and $\triangle ADC$ (see figure 8). Then, since the angle sum of a triangle is 180°,

$$\begin{gathered} s(\triangle ABC) + s(\triangle ADC) = 180^\circ + 180^\circ \\ (\angle CAB) + (\angle B) + (\angle ACB) + (\angle CAD) + (\angle D) + (\angle ACD) = 360^\circ \\ (\angle A) + (\angle B) + (\angle C) + (\angle D) = 360^\circ \\ 90^\circ + 90^\circ + 90^\circ + (\angle D) = 360^\circ \\ (\angle D) = 90^\circ. \end{gathered}$$

Therefore $\angle D$ is a right angle. $\square$

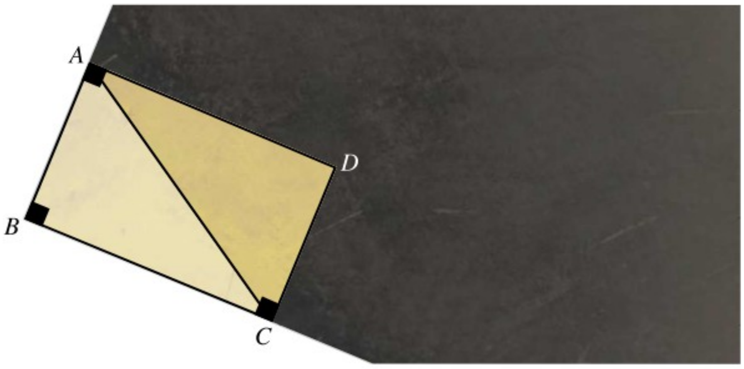

Figure 8. Verifying that the fourth angle is right.

Exercise 11.1 Fill in the gap in the proof by explaining why $\square ABCD$ must be convex.

We see then that, yes, rectangles exist in Euclidean geometry. Two essential properties of rectangles are provided in the following (easy) exercises.

Exercise 11.2 Let $\square ABCD$ be a rectangle. Prove that $\overline{AB}$ is parallel to $\overline{CD}$ and $\overline{AD}$ is parallel to $\overline{BC}$.

Exercise 11.3 Let $\square ABCD$ be a rectangle. Prove that $AB \simeq CD$ and $AD \simeq BC$ and $AC \simeq BD$.

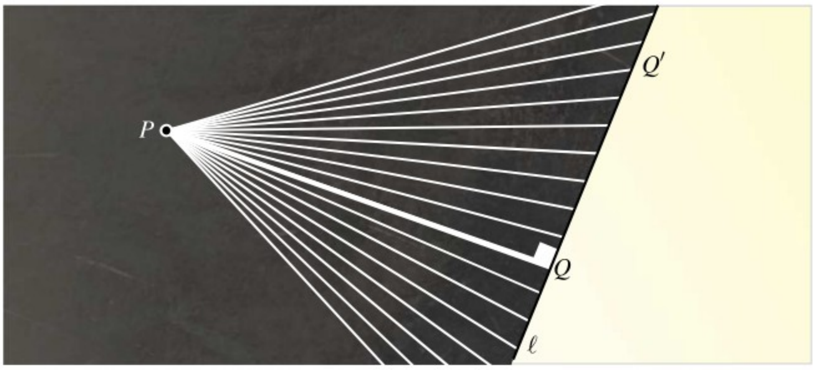

Figure 9. The distance from P to ℓ is measured along the segment from P to ℓ that is perpendicular to ℓ.

For the last result of this section, we return to parallel lines. One of the things that we will see when we study non-Euclidean geometry is that parallel lines tend to diverge from each other. That doesn't happen in Euclidean geometry. This is one of the key differences between the two geometries. Let's make this more precise. Suppose that P is a point that is not on a line ℓ. Define the distance from P to ℓ to be the minimum distance from P to a point on ℓ:

$$d(P,\ell) = \min\left\{|PQ| \,\middle|\, Q \text{ is on } \ell\right\}.$$

Exercise 11.4 Prove that the minimum actually occurs when Q is the foot of the perpendicular to ℓ through P (this can be proved using solely the axioms of neutral geometry).

For a pair of parallel lines, that distance as measured along perpendiculars does not change.

Theorem 11.5 *If ℓ and ℓ' are parallel lines, then the distance from a point on ℓ to ℓ' is constant. In other words, if P and Q are points on ℓ, then*

$$d(P,\ell') = d(Q,\ell').$$

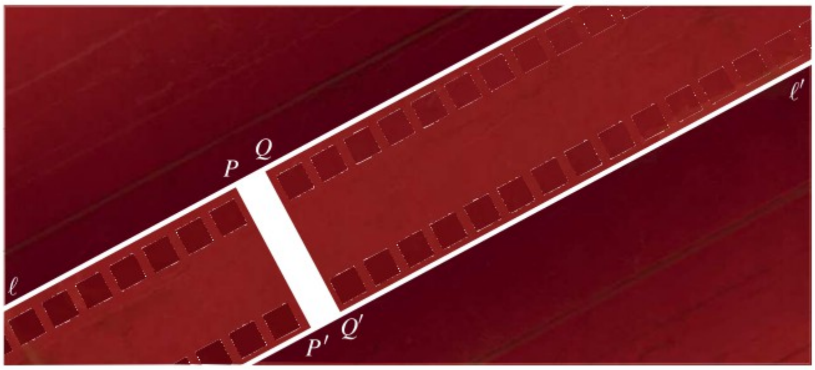

Figure 10. Perpendicular segments PP' and QQ' between parallel lines have the same length. That is, parallel lines are equidistant.

Proof Let P' and Q' be the feet of the perpendiculars on ℓ' from P and Q respectively. That way, by the previous exercise,

$$d(P, \ell') = |PP'| \text{ and } d(Q, \ell') = |QQ'|.$$

Then $\angle PP'Q'$ and $\angle QQ'P'$ are right angles. By the converse of the alternate interior angle theorem, $\angle P$ and $\angle Q$ are right angles too, so $\square PQQ'P'$ is a rectangle. According to exercise 11.2, PP' and QQ', which are opposite sides of a rectangle, are congruent. □

11.1 Exercises

11.5. Let ℓ be a line and P a point not on ℓ. Using only the results of neutral geometry, prove that there is at least one line through P that is parallel to ℓ.

11.6. Suppose that ℓ_1, ℓ_2, and ℓ_3 are three distinct lines such that ℓ_1 and ℓ_2 are parallel, and ℓ_2 and ℓ_3 are parallel. Prove that ℓ_1 and ℓ_3 are parallel.

11.7. Suppose that ℓ_1 and ℓ_2 are parallel lines, and that t is a line that intersects (but is distinct from) ℓ_1. Prove that t intersects ℓ_2.

11.8. Given three distinct parallel lines, prove that there is a labeling of them as ℓ_1, ℓ_2, and ℓ_3 so that

- no point of ℓ_1 lies between a point of ℓ_2 and a point of ℓ_3
- every point of ℓ_2 lies between a point of ℓ_1 and of ℓ_3
- no point of ℓ_3 lies between a point of ℓ_1 and a point of ℓ_2.

In this case, we can say that ℓ_2 is *between* ℓ_1 and ℓ_3.

11.9. This exercise is an extension of the previous one. Given n distinct parallel lines, prove that there is an ordering of them as $\ell_1, \ell_2, \ldots, \ell_n$ so that if $i < j < k$ then ℓ_j is between ℓ_i and ℓ_k.

11.10. Find the angle sum of a convex n-gon as a function of n.

11.11. What is the measure of an interior angle of a regular n-gon (as a function of n)?

11.12. Consider a convex quadrilateral $\square ABCD$. Prove that the two diagonals of $\square ABCD$ bisect each other if and only if $\square ABCD$ is a parallelogram.

11.13. Show that a parallelogram $\square ABCD$ is a rectangle if and only if $AC \simeq BD$.

11.14. Suppose that the diagonals of a convex quadrilateral $\square ABCD$ intersect one another at a point P and that

$$AP \simeq BP \simeq CP \simeq DP.$$

Prove that $\square ABCD$ is a rectangle.

11.15. Suppose that the diagonals of a convex quadrilateral bisect one another at right angles. Prove that the quadrilateral must be a rhombus.

11.16. Let $\triangle ABC$ be isosceles with $AB \simeq AC$, and let a be the midpoint of BC. Let b and c be points on the line through A that is parallel to BC so that $b * A * c$ and so that $bA \simeq cA \simeq Ba$. Prove that $\triangle ABC \simeq \triangle abc$.

11.17. Verify that the Cartesian model (as developed through the exercises in chapters 1–3) satisfies Playfair's axiom.

12

Parallel Projection

In the course of this chapter, we will need to use a few properties of parallelograms. It is rather common to confuse the properties of a parallelogram with its definition, so first, recall the definition of a parallelogram.

Definition 12.1 A *parallelogram* is a simple quadrilateral whose opposite sides are parallel.

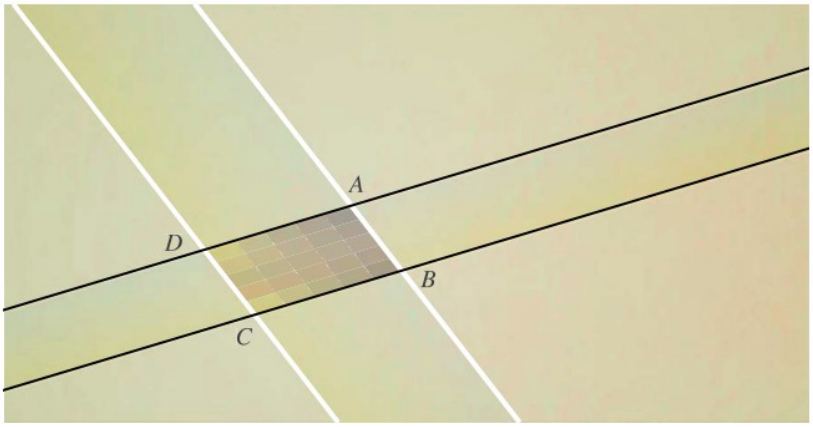

Figure 1. A parallelogram.

Now on to the properties of parallelograms that we will need for this chapter.

Exercise 12.1 Prove that in a parallelogram, the two pairs of opposite sides are congruent and the two pairs of opposite angles are congruent.

Exercise 12.2 Prove that if a convex quadrilateral has one pair of opposite sides that are both parallel and congruent, then it is a parallelogram.

Exercise 12.3 Let $\square ABB'A'$ be a simple quadrilateral. Verify that if AA' and BB' are parallel, but AB and $A'B'$ are not, then AA' and BB' cannot be congruent.

12.1 Parallel projection

The purpose of this chapter is to introduce a mechanism called *parallel projection*, a particular kind of mapping from the points on one line to the points on another. Parallel projection is the piece of machinery that you have to have in place to understand similarity, which is in turn essential for much of what we will be doing in the next chapters. The primary goal at this point is to understand how distances between points may be distorted by the parallel projection mapping. Once that is figured out, we will be able to turn our attention to the geometry of similarity.

Definition 12.2 A *parallel projection* from one line ℓ to another ℓ' is a map $\Phi : \ell \to \ell'$ that assigns to each point P on ℓ a point $\Phi(P)$ on ℓ' so that all the lines connecting a point and its image are parallel to one another.

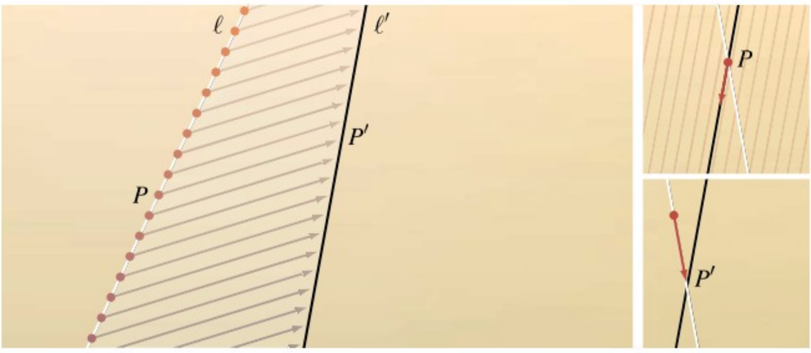

Figure 2. The path from a point P on ℓ to a point P' on ℓ' defines a parallel projection as long as neither P nor P' is the intersection of ℓ and ℓ' (as shown at right).

It is easy to construct parallel projections. Any point P on ℓ and its image $\Phi(P)$ on ℓ' completely determines the projection: for any other point Q on ℓ there is a unique line that passes through Q and is parallel to $\overleftrightarrow{P\,\Phi(P)}$. Wherever the line intersects ℓ' will have to be $\Phi(Q)$. There are only two cases where this construction will not work: if P is the intersection of ℓ and ℓ', then the lines of projection run parallel to ℓ' and so fail to provide a point of intersection; and if $\Phi(P)$ is the intersection of ℓ and ℓ', then the lines of projection actually coincide rather than being parallel.

Theorem 12.3 *A parallel projection is both one-to-one and onto.*

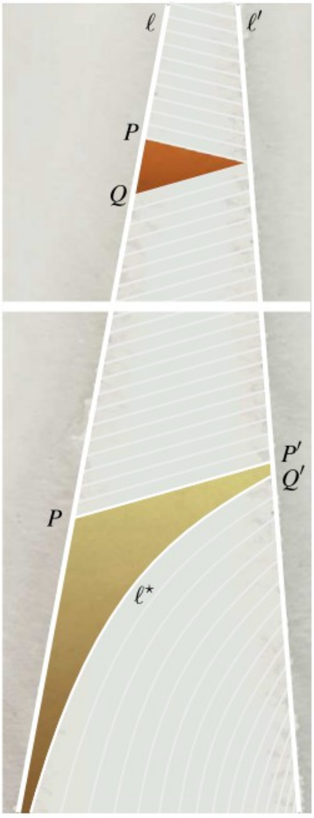

Figure 3. Using proofs by contradiction, we can show that parallel projection is both one-to-one (top) and onto (bottom).

Proof Let $\Phi : \ell \to \ell'$ be a parallel projection. First let's see why Φ is one-to-one. Suppose that it is not. That is, suppose that P and Q are two distinct points on ℓ but that $\Phi(P) = \Phi(Q)$. Then the two projecting lines $\overline{P\,\Phi(P)}$ and $\overline{Q\,\Phi(Q)}$, which ought to be parallel, actually share a point. This can't happen.

Now let's see why Φ is onto. Suppose that Q' is a point on ℓ' that is not in the image of Φ. To get a sense of how Φ is casting points from ℓ to ℓ', consider a point P on ℓ and its image $P' = \Phi(P)$ on ℓ'. By Playfair's axiom, there is a unique line $\ell^\star$ that passes through Q' and is parallel to $\overline{PP'}$. If $\ell^\star$ intersects ℓ at a point Q, then $\Phi(Q) = Q'$, contradicting the initial supposition that Q' is not in the image of Φ. If $\ell^\star$ does not intersect ℓ, then both ℓ and $\overline{PP'}$ are lines that pass through P and are parallel to $\ell^\star$, in violation of Playfair's axiom. In either case, we arrive at a contradiction. Therefore, an arbitrary point Q' of ℓ must be in the image of Φ, and so Φ is onto. $\square$

Now that we have established that parallel projection is a bijection, it helps to use a naming convention to makes things more readable. I will use a prime mark $'$ to indicate the parallel projection of a point, so $\Phi(P) = P'$, $\Phi(Q) = Q'$, and so on.

12.2 Parallel projection, order, and congruence

So far we have seen that parallel projection establishes a correspondence between the points of one line and the points of another. What about the order of the points? Can points get shuffled up in the process of a parallel projection? No.

Theorem 12.4 *Let $\Phi : \ell \to \ell'$ be a parallel projection. If A, B, and C are points on ℓ and $A * B * C$, then $A' * B' * C'$.*

Proof Because B is between A and C, A and C must be on opposite sides of $\overline{BB'}$. But, $\overline{AA'}$ does not intersect $\overline{BB'}$, so A' has to be on the same side of $\overline{BB'}$ as A; and $\overline{CC'}$ does not

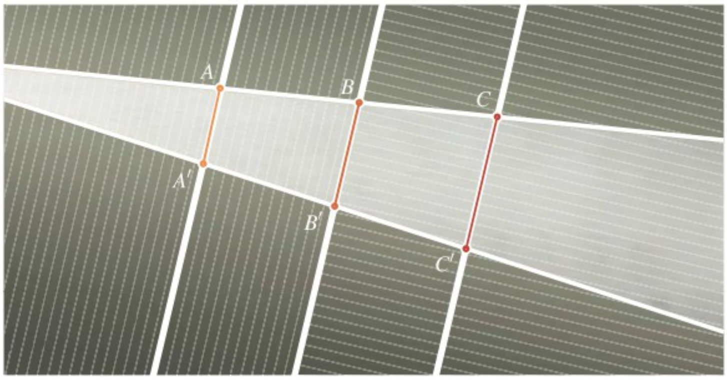

Figure 4. Parallel projection keeps points in order.

intersect $\overline{BB'}$, so C' has to be on the same side of $\overline{BB'}$ as C. Therefore A' and C' have to be on opposite sides of $\overline{BB'}$, and so the intersection of $\overline{BB'}$ and $A'C'$, which is B', must be between A' and C'. □

Now what about congruence?

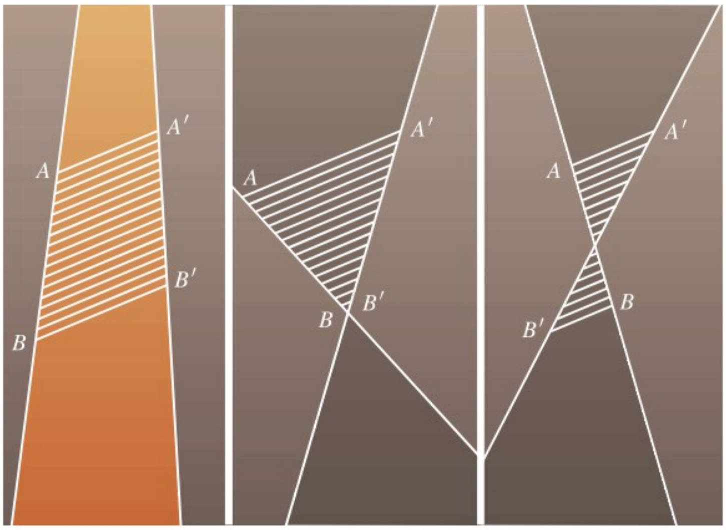

Figure 5. There are three positions for A and B relative to the image line: both on the same side, one on the image line, or one on each side. Likewise, there are three positions for a and b. There are nine cases in all.

Theorem 12.5 *Let* $\Phi : \ell \to \ell'$ *be a parallel projection. If* a, b, A, *and* B *are points on* ℓ *and if* $ab \simeq AB$, *then* $a'b' \simeq A'B'$.

Proof There are several cases, depending on the positions of ab and AB relative to ℓ'. They could lie on the same side of ℓ', they could lie on opposite sides of ℓ', one or both could straddle

ℓ', or one or both could have an endpoint on ℓ'. Each case must be handled differently. We will look at the case where both segments are on the same side of ℓ'. The others are left as exercises. There are still two cases to consider, depending on whether or not ℓ and ℓ' are parallel.

Case 1: when ℓ and ℓ' are parallel (see figure 6). This is the simpler case, but it helps illuminate the more general case. Because aa' is parallel to bb' and ab is parallel to $a'b'$, $\square aa'b'b$ is a parallelogram; AA' is parallel to BB' and AB is parallel to $A'B'$ so $\square AA'B'B$ is also a parallelogram. Because the opposite sides of a parallelogram are congruent (exercise 12.1), $a'b' \simeq ab$ and $AB \simeq A'B'$. Since $ab \simeq AB$, $a'b' \simeq A'B'$.

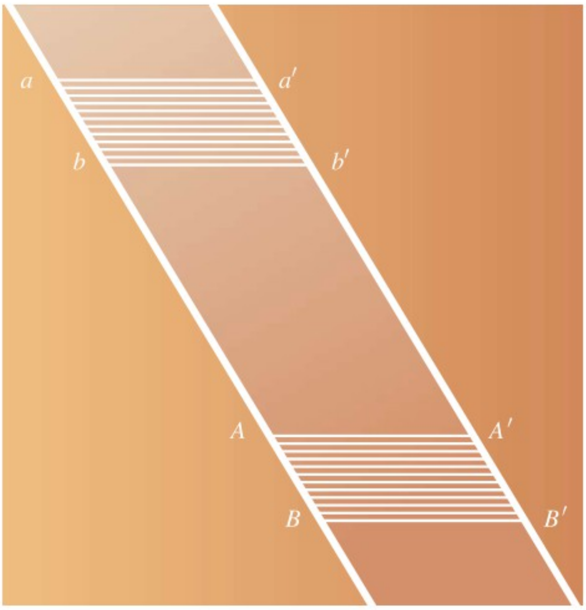

Figure 6. Case 1, when the two lines are parallel.

Case 2: when ℓ and ℓ' are not parallel (see figure 7). This is the more likely case, in which $\square aa'b'b$ and $\square AA'B'B$ will not be parallelograms. We can takes a similar approach to case 1, though, by building some parallelograms into the problem. Because ℓ and ℓ' are not parallel, the segments aa' and bb' cannot have the same length (exercise 12.3), and the segments AA' and BB' cannot have the same length. Let's assume that aa' is shorter than bb' and that AA' is shorter than BB'. If this is not the case, then it is just a matter of switching some labels to make it so. Then

- there is a point c between b and b' so that $bc \simeq aa'$, and
- there is a point C between B and B' so that $BC \simeq AA'$.

This creates four shapes of interest: the quadrilaterals $\square a'abc$ and $\square A'ABC$ that are parallelograms (exercise 12.2), and the triangles $\triangle a'b'c$ and $\triangle A'B'C$. The key here is to prove that $\triangle a'b'c \simeq \triangle A'B'C$.

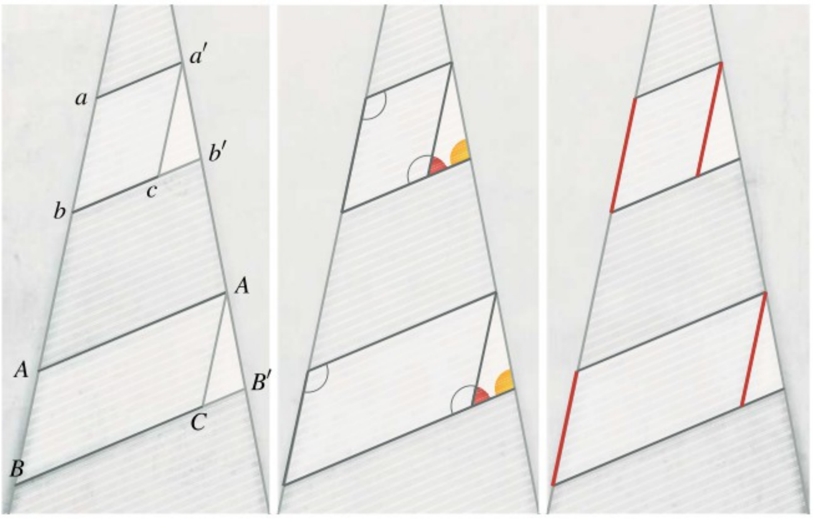

Figure 7. The non-parallel case. (l) First construct parallelograms. (c) The two angles and (r) the sides needed for AAS to show that $\triangle a'b'c' \simeq \triangle A'B'C'$.

- **A:** $\angle b' \simeq \angle B'$. The lines $\overline{cb'}$ and $\overline{CB'}$ are parallel (they are two of the projecting lines) and they are crossed by the transversal ℓ'. By the converse of the alternate interior angle theorem, $\angle a'b'c$ and $\angle A'B'C$ are congruent.
- **A:** $\angle c \simeq \angle C$. The opposite angles of the two parallelograms are congruent. Therefore $\angle a'cb \simeq \angle a'ab$ and $\angle A'AB \simeq \angle A'CB$. But $\overline{aa'}$ and $\overline{AA'}$ are parallel lines cut by the transversal ℓ, so $\angle a'ab \simeq \angle A'AB$, so $\angle a'cb \simeq \angle A'CB$, and so their supplements $\angle a'cb'$ and $\angle A'CB'$ are also congruent.
- **S:** $a'c \simeq A'C$. The opposite sides of the two parallelograms are congruent too. Therefore $a'c \simeq ab$ and $AB \simeq A'C$, and since $ab \simeq AB$, we get $a'c \simeq A'C$.

By AAS, then, $\triangle a'b'c \simeq \triangle A'B'C$. The corresponding sides $a'b'$ and $A'B'$ have to be congruent.
□

Exercise 12.4 Fill in some gaps:

(1) Prove the theorem in the case when both a and A lie on ℓ' (at the intersection of ℓ and ℓ').
(2) Prove the theorem in the case when a and A lie on opposite sides of ℓ' from b and B.

12.3 Parallel projection and distance

That brings us to the question at the very heart of parallel projection. If Φ is a parallel projection and A and B are two points on ℓ, how do the lengths $|AB|$ and $|A'B'|$ compare? In case 1 of the last proof, AB and $A'B'$ ended up being congruent, but that was because ℓ and ℓ' were parallel. In general, AB and $A'B'$ do not have to be congruent. But it turns out that in the process of parallel projecting from one line to another, all distances are scaled by a constant multiple.

Theorem 12.6 *If $\Phi : \ell \to \ell'$ is a parallel projection, then there is a constant k such that*

$$|A'B'| = k|AB|$$

for all points A and B on ℓ.

It is helpful to get a little perspective on this before the proof. First note that the previous theorem gives us a way to narrow the scope of the problem. Fix a point O on ℓ and let r be one of the two rays along ℓ with O as its endpoint. The segment construction axiom says that every segment AB on ℓ is congruent to a segment OP where P is some point on r. We have just seen that parallel projection maps congruent segments to congruent segments. So if Φ scales all segments of the form OP by a factor of k, then it must scale all the segments of ℓ by the same factor.

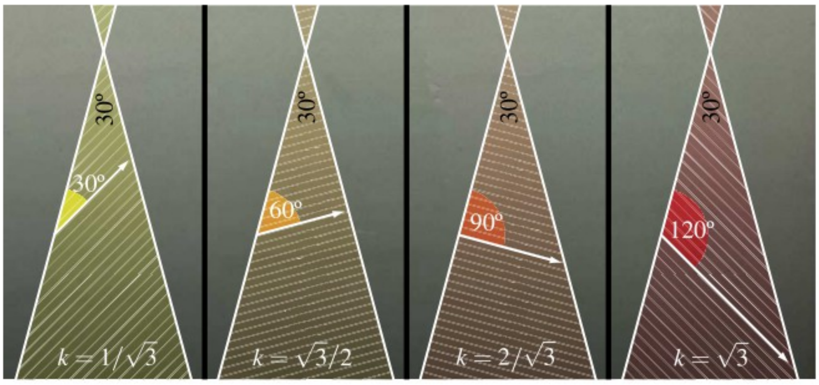

Figure 8. Some parallel projections and their scaling constants.

Second, we can extend that by considering what happens when we parallel project end-to-end congruent copies of a segment. For this, let me introduce another convenient notation convention: for the rest of this argument, let P_d denote the point on r that is a distance d from O. Now, pick a positive real value x, and let

$$k = |O'P'_x|/|OP_x|,$$

so that Φ scales the segment OP_x by a factor of k. Eventually we will have to show that Φ scales all segments by that factor, but for now let's restrict our attention to the segments OP_{nx}, where n is a positive integer. Between O and P_{nx} are P_x, P_{2x}, … $P_{(n-1)x}$ in order:

$$O * P_x * P_{2x} * \cdots * P_{(n-1)x} * P_{nx}.$$

We have seen that parallel projection preserves the order of points, so

$$O' * P'_x * P'_{2x} * \cdots * P'_{(n-1)x} * P'_{nx}.$$

Each segment $P_{ix}P_{(i+1)x}$ is congruent to OP_x and consequently each parallel projection $P'_{ix}P'_{(i+1)x}$ is congruent to $O'P'_x$. Add their lengths:

$$\begin{aligned}|O'P'_{nx}| &= |O'P'_x| + |P'_xP'_{2x}| + |P'_{2x}P'_{3x}| + \cdots + |P'_{(n-1)x}P'_{nx}| \\ &= kx + kx + kx + \cdots + kx \quad (n \text{ times}) \\ &= k \cdot nx\end{aligned}$$

and so Φ scales OP_{nx} by a factor of k.

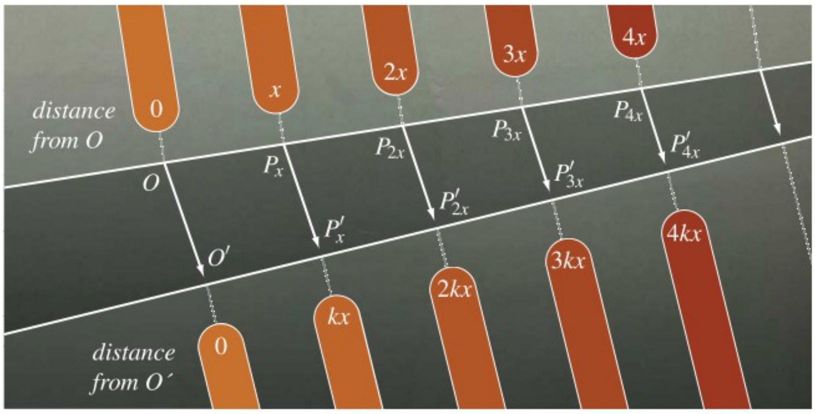

Figure 9. Scaling of segments of length kx, where k is an integer.

Sadly, no matter what x is, the points P_{nx} account for an inconsequential portion of the set of all points of r. However, if OP_x and OP_y were to have two different scaling factors we could use this end-to-end copying to magnify the difference between them. Third, then, consider an example that demonstrates how this works, and which shows how this ultimately prevents there being two different scaling factors. In this example, let's suppose that $|O'P'_1| = 2$, so all integer length segments on ℓ are scaled by a factor of 2. Let's take a look at what this means for $P_{3.45}$. Let k be the scaling factor for $OP_{3.45}$. Figure 10 shows how the first few end-to-end copies of $OP_{3.45}$ begin to limit the possible values of k.

Fourth, note that the floor function, $f(x) = \lfloor x \rfloor$, assigns to each real number x the largest integer that is less than or equal to it. The ceiling function, $f(x) = \lceil x \rceil$, assigns to each real number x the smallest integer that is greater than or equal to it.

It is finally time to prove that parallel projection scales distance.

Proof Let $k = |O'P'_1|$ so that k is the scaling factor for the segment of length 1 (and consequently for all integer length segments). Now take a point P_x on ℓ and let k' be the scaling factor for the segment OP_x. We want to show that $k' = k$ and we can do that by following the same strategy as in the example above, by capturing k' in an increasingly narrow band around k by looking at the parallel projection of P_{nx} as n increases. Start by working back and forth between

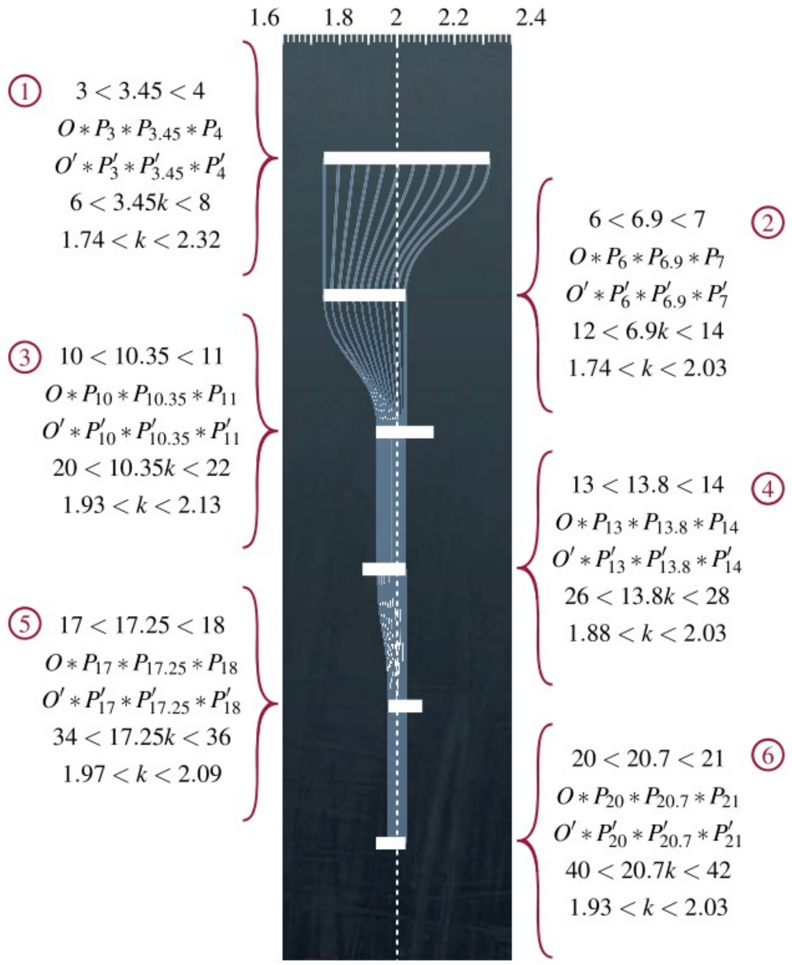

① $3 < 3.45 < 4$
$O * P_3 * P_{3.45} * P_4$
$O' * P'_3 * P'_{3.45} * P'_4$
$6 < 3.45k < 8$
$1.74 < k < 2.32$

② $6 < 6.9 < 7$
$O * P_6 * P_{6.9} * P_7$
$O' * P'_6 * P'_{6.9} * P'_7$
$12 < 6.9k < 14$
$1.74 < k < 2.03$

③ $10 < 10.35 < 11$
$O * P_{10} * P_{10.35} * P_{11}$
$O' * P'_{10} * P'_{10.35} * P'_{11}$
$20 < 10.35k < 22$
$1.93 < k < 2.13$

④ $13 < 13.8 < 14$
$O * P_{13} * P_{13.8} * P_{14}$
$O' * P'_{13} * P'_{13.8} * P'_{14}$
$26 < 13.8k < 28$
$1.88 < k < 2.03$

⑤ $17 < 17.25 < 18$
$O * P_{17} * P_{17.25} * P_{18}$
$O' * P'_{17} * P'_{17.25} * P'_{18}$
$34 < 17.25k < 36$
$1.97 < k < 2.09$

⑥ $20 < 20.7 < 21$
$O * P_{20} * P_{20.7} * P_{21}$
$O' * P'_{20} * P'_{20.7} * P'_{21}$
$40 < 20.7k < 42$
$1.93 < k < 2.03$

Figure 10. An example. Projecting end-to-end congruent copies of a segment of length 3.45 by a parallel projection that scales all integer length segments by a factor of two.

inequalities of real numbers and orders of points:

$$\begin{gathered}\lfloor nx \rfloor \leq nx \leq \lceil nx \rceil \\ O * P_{\lfloor nx \rfloor} * P_{nx} * P_{\lceil nx \rceil} \\ O' * P'_{\lfloor nx \rfloor} * P'_{nx} * P'_{\lceil nx \rceil} \\ k\lfloor nx \rfloor \leq k'nx \leq k\lceil nx \rceil.\end{gathered}$$

(The notation here is a bit casual—if nx is an integer, then $\lfloor nx \rfloor = nx$, so $P_{\lfloor nx \rfloor}$ is not actually between O and P_{nx}. This does not cause any problems for the proof.) Because $(nx - 1) \leq \lfloor nx \rfloor$

and $\lceil nx \rceil \leq (nx+1)$ we can move to a less precise but more manageable set of inequalities:

$$k(nx-1) \leq k'nx \leq k(nx+1).$$

Divide by nx to get

$$k\left(\frac{nx-1}{nx}\right) \leq k' \leq k\left(\frac{nx+1}{nx}\right).$$

As n increases, the ratios $(nx-1)/(nx)$ and $(nx+1)/(nx)$ both approach 1. In the limit as n goes to infinity, they are 1. Since the inequalities have to be true for all n, the only possible value for k' , then, is k. □

12.4 Exercises

12.5. Suppose that Φ is a parallel projection from ℓ to ℓ'. If ℓ and ℓ' intersect at P, prove that $\Phi(P) = P$.

12.6. Prove that if ℓ and ℓ' are parallel, then the scaling factor of a parallel projection between them must be one, but that if ℓ and ℓ' are not parallel, then there is a parallel projection with every possible scaling factor k where $0 < k < \infty$.

12.7. Let $\Phi : \ell \to \ell'$ be a parallel projection. Show that the inverse map $\Phi^{-1} : \ell' \to \ell$ is also a parallel projection.

12.8. Let $\Phi_1 : \ell \to \ell'$ and $\Phi_2 : \ell' \to \ell''$ be parallel projections. Is the composition $\Phi_2 \circ \Phi_1 : \ell \to \ell''$ necessarily a parallel projection?

12.9. In chapter 6, we constructed a distance function, and one of the keys to the construction was locating the points on a ray that were a distance of $m/2^n$ from its endpoint. In Euclidean geometry, there is a construction that locates all the points on a ray that are any rational distance m/n from its endpoint. Take two non-opposite rays r and r' with a common endpoint O. Along r, lay out m congruent copies of a segment of length 1, ending at the point P_m. Along r', lay out n congruent copies of a segment of length 1, ending at the point Q_n. Mark the point Q_1 on r' that is a distance 1 from O. Verify that the line that passes through Q_1 and is parallel to P_mQ_n intersects r a distance of m/n from O.

% .

()*+(((, (. <) > / ; H "E
; 4 0 > D ! /"

& 'C -

Chapter 3

Calculating the Value of π?

Up to now we have described π and mentioned ways in which attempts have been made to calculate its value. They varied from highly intelligent (say, ingenious) guesses by mathematicians, to attempts at performing calculations that were later proved impossible (i.e., squaring the circle), to carefully planned constructions that would yield the value of π if carried out far enough and carefully enough. Some methods of calculating the value of π, strangely enough, relied on probability, or in one case on mysterious insights. Here we will provide you with a variety of methods for calculating the value of π. We chose those that should be easily understood by the general reader. Where a concept is used that may be a bit off the beaten path, or simply unfamiliar to some, we provide some background information. We will be presenting the classical attempts, rather than those used in the more recent computer-driven methods. We begin with one of the most famous classical methods, by one of the most gifted mathematicians in the history of mathematics, Archimedes.

Archimedes' Method for Finding the Value of π

Perhaps the easiest way to begin to calculate the value of π was developed by Archimedes. It is a method that can appeal to one's intuition. He noticed that as the number of sides of a regular polygon increases, while keeping the radius or the apothem[1] constant, the limiting value of the perimeter is the circumference of a circle. That is, suppose we take the first few regular polygons (an equilateral triangle, a square, a regular pentagon, and a regular hexagon) and inscribe them in the same-size circle. As the number of sides of the regular polygon increases, the perimeter of the polygon gets closer and closer to the circumference (i.e., perimeter) of the circle. Remember that the circumscribed circle must contain each of the vertices of the polygon. Here is what it can look like.

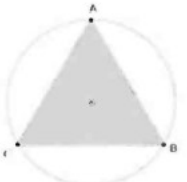

Fig. 3-1a

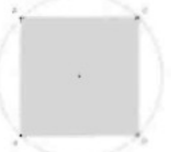

Fig. 3-1b

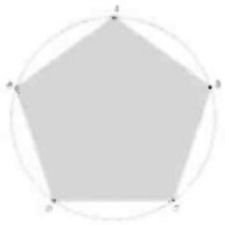

Fig. 3-1c

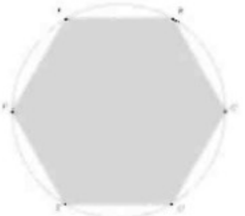

Fig. 3-1d

---

[1] The apothem is the segment from the center of a regular polygon to the midpoint of one of its sides. It is perpendicular to the side.

This may be easier to see when the regular polygon's sides increase further so that it becomes a dodecagon (which has twelve sides). We can actually calculate the increasing perimeters and see them gradually approach the circumference of the circle.

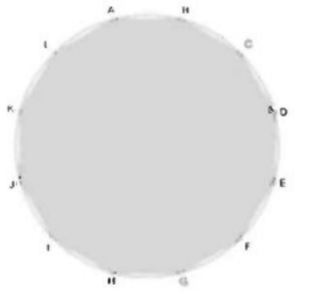

Fig. 3-2

Let's take the hexagon as our example of a "general polygon." From this we will then generalize to polygons of many more (or fewer) sides. We begin with a regular hexagon inscribed in a circle of radius $\frac{1}{2}$. The measure of $\angle AOB$ is one-sixth of a complete revolution of 360°, or 60°. Since $\overline{OK} \perp \overline{AB}$ at K, $BK = AK = a$.

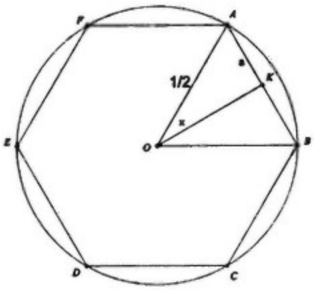

Fig. 3-3

We seek to find the perimeter of the hexagon, when we know the length of the radius $\left(\frac{1}{2}\right)$ and the measure of

$$\angle AOK = \frac{1}{2}(60°) = 30°$$

Using the trigonometric function sine,[2] we get

$$\angle AOK = \sin 30° = \frac{a}{\frac{1}{2}} = 2a$$

Since $\sin 30° = \frac{1}{2}$, then $2a = \frac{1}{2}$, and $a = \frac{1}{4}$. The perimeter of the hexagon is then 12 times a, which equals 3.

Let's generalize this for any regular polygon of n sides.

$$\angle x = \frac{1}{2} \bullet \frac{360°}{n} = \frac{180°}{n}$$

Therefore, for the general regular polygon of n sides

$$\sin \frac{180°}{n} = 2a$$

The perimeter of the n-sided regular polygon is then n times $2a$, which makes this perimeter equal to

$$n \sin \frac{180°}{n}$$

We can then take various values of n and compute the perimeter of the regular polygon whose circumscribed circle has a radius of $\frac{1}{2}$.

---

[2] The sine function is defined for a right triangle as the ratio of the side opposite the angle in question and the hypotenuse (the side opposite the right angle).

We shall work out the first few examples here and then provide the results of others in a table.

When $n = 3$:

$$3\sin\frac{180^\circ}{3} = 3\sin 60^\circ \approx 3(0.86602540378443864676372317075294)$$
$$= 2.5980762113533159402911695122588$$

When $n = 4$:

$$4\sin\frac{180^\circ}{4} = 4\sin 45^\circ \approx 4(0.70710678118654752440084436210485)$$
$$= 2.8284271247461900976033774484194$$

When $n = 5$:

$$5\sin\frac{180^\circ}{5} = 5\sin 36^\circ \approx 5(0.58778525229247312916870595463907)$$
$$= 2.9389262614623656458435297731954$$

When $n = 6$:

$$6\sin\frac{180^\circ}{6} = 6\sin 30^\circ = 6(0.50000000000000000000000000000000)$$
$$= 3.00000000000000000000000000000000$$

We just calculated the first four entries in the table below and here provide you with the remaining ones.

n	**Perimeter of inscribed polygon of *n* sides**
3	2.5980762113533159402911695122588…
4	2.8284271247461900976033774484194…
5	2.9389262614623656458435297731954…
6	3.0000000000000000000000000000000…
7	3.0371861738229068433303783299385…
8	3.0614674589207181738276798722432…
9	3.0781812899310185973968965321403…
10	3.0901699437494742410229341718282…
11	3.0990581252557266748255970688128…
12	3.1058285412302491481867860514886…
13	3.111036357382509729337984413828…
14	3.1152930753884016600446359029551…
15	3.1186753622663900565261342660769…
24	3.1326286132812381971617494694917…
36	3.137606738915694248090313750149…
54	3.1398207611656947410923929097419…
72	3.140595890304191984286221559116…
90	3.14095470322508744813956634628…
120	3.1412337969447783132734022664935…
180	3.1414331587110323074954161329369…
250	3.1415099708381519785686472871987…
500	3.1415719827794756248676550789799…
1,000	3.1415874858795633519332270354959…
10,000	3.141592601912665692979346479289…

Now compare this 10,000-sided regular polygon (our last entry) to the value of π that we already know. Remember it is inscribed in the circle with radius $\frac{1}{2}$. This 10,000-sided polygon is optically quite indistinguishable from the circle (obviously, without magnification enhancements). The circumference of the circumscribed circle of radius $\frac{1}{2}$ is $2\pi r = 2\pi\left(\frac{1}{2}\right) = \pi$.

Look at the known value of π for comparison.

π ≈ 3.1415926535 8979323846 2643383279 5028841971 6939937510 5820974944…

Up to the seventh decimal place, the approximation with a 10,000-sided regular polygon perimeter is correct. If we were to calculate the perimeter of a regular polygon of 100,000 sides, we would get an even closer approximation. The perimeter of a regular polygon of 100,000 sides is 3.14159265307302196048314802075 31…, which approximates π correct to nine decimal places.

Archimedes (obviously) did not have the luxury of using electronic (or even mechanical) calculating devices to assist him in his calculations.[3] He also did not have the facility brought about by the place value system (such as our decimal system), nor did he have the use of trigonometry available to him. Yet he still used a 96-sided regular polygon. He saw the circle as the limiting figure of an inscribed circle as well as the circumcircle we just used above. By taking the average of the perimeters of each pair of circles of n-sided regular polygons, he would "sandwich in" the perimeter of the circle, which in the case of a circle with radius of $\frac{1}{2}$ is π.

---

[3] The mechanical calculator was invented by four mathematicians over a rather wide stretch of time. Wilhelm Schickardt (1592–1635), a German mathematician, built the first digital calculator in 1623. Blaise Pascaal (1623–1662) built the first mechanical calculating machine in 1642 for his father, who was a tax collector. The machine, called Pascaline, was commercially sold after 1645. Gottfried Wilhelm Leibniz (1646–1716) developed a mechanical calculator in 1673 that failed during a demonstration in London but nonetheless, because of the spectacular concept involved, was accepted in the Royal Society. The English mathematician Charles Babbage (1792–1871), despite devoting a greater part of his professional life to its development, never reached a completed product.

Let us now repeat the above exercise with the polygon circumscribed about the circle, or, put another way, where the circle of radius $\frac{1}{2}$ is inscribed in the polygon (i.e., the circle must be tangent to each side of the polygon). As before, we will consider regular polygons with successively greater numbers of sides, each with our given circle inscribed.

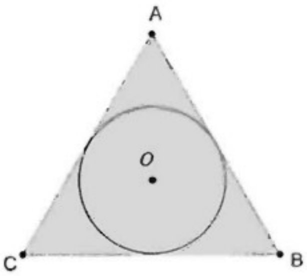

Fig. 3-4a

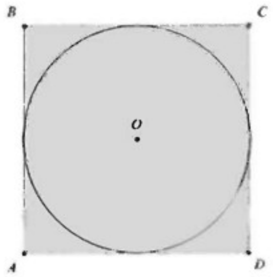

Fig. 3-4b

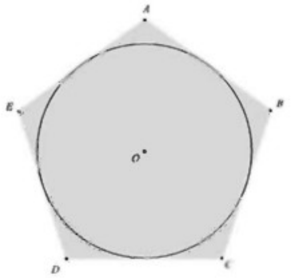

Fig. 3-5

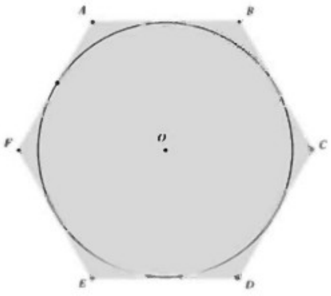

Fig. 3-6

Notice how gradually the perimeter of the polygon appears to get closer and closer to the circumference of the circle.

He began his work in 1812 and worked for decades on the project. In the end, the lack of precision tools prevented him from achieving his "analytical engine" in 1833. His work was first realized in the form of a working machine in 1944 when the IBM Corporation and Harvard University collaborated to produce the Automatic Sequence Controlled Calculator.

This time we will consider a regular pentagon circumscribed about our circle of radius $\frac{1}{2}$ as our first polygon to study. Then we will generalize our procedure and extend it to many others.

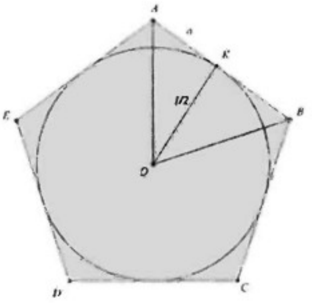

Fig. 3-7

Our objective is to find the perimeter of the pentagon with a side of $2a$. We know that

$$\tan \angle AOK = \frac{a}{OK}{}^{4} \text{ and } m\,\angle AOB = 72^\circ$$

so that

$$m\angle AOK = 36^\circ \text{, while } OK = \frac{1}{2}$$

Therefore,

$$a = \frac{1}{2}\tan 36^\circ \approx \frac{1}{2}(0.72654252800536088589546675748062)$$
$$= 0.36327126400268044294773337874031$$

---

[4] The tangent function is defined for a right triangle as the ratio of the side opposite the angle being considered to the side adjacent to this angle.

Thus the perimeter of the pentagon is 10 times a, or about 3.6327126400268044294773337874031 (this is found by taking 5 times $2a$, or about 3.6327), not yet a very close approximation of π. The circumference of the circle is $2\pi r = 2\pi\left(\frac{1}{2}\right) = \pi$.

In the general case of a regular polygon of n sides

$$m\angle AOK = \frac{1}{2}\bullet\frac{360^\circ}{n} = \frac{180^\circ}{n}$$

From the example of the pentagon, $\tan\angle AOK = \frac{a}{OK}$. It follows that

$$a = OK\tan\angle AOK = \frac{1}{2}\ \bullet \tan\ \frac{180^\circ}{n}$$

The perimeter of the polygon is then

$$n\bullet 2a = n\bullet 2\bullet\frac{1}{2}\bullet\tan\frac{180^\circ}{n} = n\tan\frac{180^\circ}{n}$$

As before, we will calculate the perimeters of the various polygons, this time, though, circumscribed about our circle with radius $\frac{1}{2}$.

We already have the calculated perimeter for the pentagon, so we will do the calculation for the hexagon now.

When $n = 6$:

$$n\bullet\tan\frac{180^\circ}{n} = 6\bullet\tan\frac{180^\circ}{6}\bullet\tan 30^\circ = 6\bullet\frac{\sqrt{3}}{3} \approx 3.4641$$

For more than four decimal places we get the following:

n	**Perimeter of circumscribed polygon of *n* sides**
3	5.1961524227066318805823390245176…
4	4.00000000000000000000000000000000…
5	3.6327126400268044294773337874031…
6	3.4641016151377545870548926830117…
7	3.3710223316527005103251364713988…
8	3.3137084989847603904135097936776…
9	3.2757321083958212521594309449915…
10	3.2491969623290632615587141221513…
11	3.2298914223220338542066829685944…
12	3.2153903091734724776706439019295…
13	3.2042122194157076473003149216291…
14	3.1954086414620991330865590688542…
15	3.1883484250503318788938749085512…
24	3.1596599420975004833166349778332…
36	3.1495918869332641879926720996586…
54	3.1451418433791039391493421086004…
72	3.1435878894128684595626030399174…
90	3.1428692542572957450362363196353…
96	3.1427145996453682981688590937721…
120	3.1423105883024314667236592753428…
180	3.141911687079165437723201139551…
250	3.1417580308448944353707690613384…
500	3.1416339959448860645952957694732…
1,000	3.1416029890561561260413432901054…
10,000	3.1415927569440529197246707719118…

Again you will notice how the more sides the polygon has, the closer its perimeter gets to the circumference of the circle—which we now know is π.

Archimedes, as we said before, saw the inscribed and the circumscribed polygons "sandwiching in" the circle, as seen below by the inscribed and circumscribed dodecagons ($n = 12$).

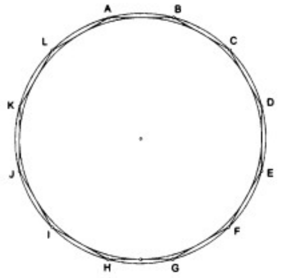

Fig. 3-8

He essentially suggested taking the average of the two perimeters for each type of polygon to get a better approximation of π.

n	Perimeter of inscribed polygon of *n* sides	Perimeter of circumscribed polygon of *n* sides	Average of the perimeters of the inscribed and circumscribed polygons of *n* sides
3	2 5980762113533159402911695122588	5 1961524227066318805823390245176	3 8971143170299739104367542683875
4	2 8284271247461900976033774484194	4 0000000000000000000000000000000	3 4142135623730950488016887242095
5	2 9389262614623656458435297731954	3 6327126400268044294773337874031	3 285819450744585037660431780299
6	3 0000000000000000000000000000000	3 4641016151377545870548926830117	3 2320508075688772935274463415055
7	3 0371861738229068433303783299385	3 3710223316527005103251364713988	3 204104252737803676827757400668
8	3 0614674589207181738276798722432	3 3137084989847603904135097936776	3 18758797895273928212059483296
9	3 0781812899310185973968965321403	3 2757321083958212521594309449915	3 1769566991634199247781637385655
10	3 0901699437494742410229341718282	3 2491969623290632615587141221513	3 1696834530392687512908241469895
11	3 0990581252557266748255970688128	3 2298914223220338542066829685944	3 164474773788880264516140018703
12	3 1058285412302491481867860514886	3 2153903091734724776706439019295	3 1606094252018608129287149767085
13	3 111036357382509729337984413828	3 2042122194157076473003149216291	3 1576579275769793101170566815055
14	3 1152930753884016600446359029551	3 1954086414620991330865590688542	3 1553508584252503965655974859045
15	3 1186753622663900565261342660769	3 1883484250503318788938749085512	3 1535118936583609677100045873135
24	3 1326286132812381971617494694917	3 1596599420975004833166349778332	3 146144277689369340239192223662
36	3 137606738915694248090313750149	3 1495918869332641879926720996586	3 1435993129244792180414929249035
54	3 1398207611656947410923929097419	3 1451418433791039391493421086004	3 1424813022723993401208675091705
72	3 140595890304191984286221559116	3 1435878894128684595626030399174	3 1420918898585302219244122995165
90	3 14095470322508744813956634628	3 1428692542572957450362363196353	3 1419119787411915965879013329575
96	3 1410319508905096381113529264597	3 1427145996453682981688590937721	3 1418732752679389681401060101155
120	3 1412337969447783132734022664935	3 1423105883024314667236592753428	3 1417721926236048899985307709175
180	3 1414331587110323074954161329369	3 141911687079165437723201139551	3 1416724228950988726093086362435
250	3 1415099708381519785686472871987	3 1417580308448944353707690613384	3 141634000841523206969708174268
500	3 1415719827794756248676550789799	3 1416339959448860645952957694732	3 141602989362180844731475424226
1 000	3 1415874858795633519332270354959	3 1416029890561561260413432901054	3 1415952374678597389872851628
10 000	3 141592601912665692979346479289	3 1415927569440529197246707719118	3 1415926794283593063520086256

The average of the two perimeters (the right-hand column) is closest to the value of π for each type of polygon. When Archimedes did these calculations, he didn't take as many examples as we did here. He began with two regular hexagons, then doubled the number of sides, using two dodecagons (12-sided polygons), then used two 24-gons,[5] then two 48-gons, and then two 96-gons.

Although his calculations were probably not as accurate as ours are, and we do not have a record of how he did his calculations, he did conclude from the 96-gon that the ratio of the circumference of a circle to its diameter—which is π—is greater than $3\frac{10}{71}$ and less than $3\frac{1}{7}$. We can write this symbolically as

$$3\frac{10}{71} < \pi < 3\frac{1}{7}$$

For comparison to the above, this is

$$3.1408450704225352112676056338028\ldots < \pi < 3.1428571428571428571428571428571\ldots$$

We have come a long way since Archimedes' ingenious methods. As we noted earlier, we can now calculate π to many more places than ever thought possible; however, this "primitive" method gives much intuitive insight into what this ratio that π represents really is.

A Reverse Method to Archimedes by Cusanus

Archimedes had used inscribed and circumscribed regular polygons within and about a given circle, each time increasing the number of sides. The argument was that as the number of sides of the polygons increased, the circumference of the circle, "sandwiched" between the two polygons, was the limiting value of the polygon.

---

[5] 24-gon is a short way of referring to a 24-sided polygon.

An analogous method developed by Nicholas of Cusa (1401–1464) has us "sandwiching" in regular polygons with increasing numbers of sides by inscribed and circumscribed circles. Nicholas of Cusa[6] took his name from his hometown of Cues (today Kues) on the Mosel River in Germany. By today's assessments, he is considered one of the pioneering German thinkers in the transition from the Middle Ages to modern times, yet he was not too well known as a mathematician. He was better known for his substantial career in the church. He became a cardinal in 1488 and was bishop of Brixen (northern Italy) and a governor (or vicar general) of Rome. As a mathematician, he made ill-fated attempts to square the circle[7] and to trisect the general angle,[8] both of which we now know are impossible. As with many mathematicians fascinated with one of the three "famous problems of antiquity," namely, squaring the circle, Cusanus's attempts led him to a fine approximation of π. Let's take a look at what Cusanus achieved in these attempts. We will demonstrate this here, but in more modern terms.

In 1450 Cusanus nested a given regular polygon with the fixed perimeter 2 with inscribed and circumscribed circles. He used a sequence of the regular n-gons ($n = 4, 8, 16, 32, \ldots$).

---

[6] Sometimes referred to by his Latin name, Cusanus.

[7] One of the three famous problems of antiquity—no longer a problem today—is how to "square a circle." That means how to construct (with only an unmarked straightedge and a pair of compasses) the side length of a square equal in area to a given circle. Today, we know this to be impossible.

[8] Another of the three famous problems of antiquity—no longer a problem today—is how to "trisect an angle." That means how to construct (with only an unmarked straightedge and a pair of compasses) angle trisectors of a general angle—not any specific number of degrees, for it would be possible for some special angles, such as a right angle Today, we know this general angle trisection to be impossible

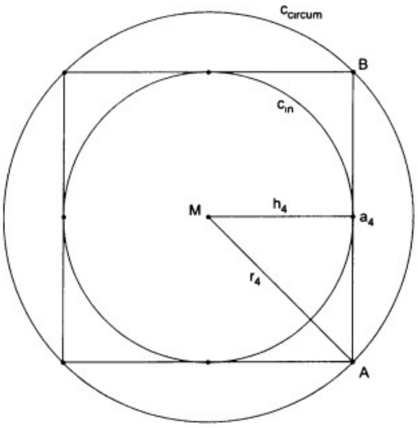

Fig. 3-9

Let's begin as Cusanus did by starting with a square (which may, of course, be referred to as a regular 4-gon). We will call the perimeter of the square p_4. Since each side of the square is $a_4 = \frac{1}{2}$, then $p_4 = 4 \bullet a_4 = 2$.

Consider the inscribed circle with circumference C_{in} and the circumscribed circle with circumference C_{circum} of the square pictured above. The radius of the inscribed circle is

$$h_4 = \frac{a_4}{2} = \frac{1}{4} = 0.25$$

The radius of the circumscribed circle is

$$r_4 = \sqrt{h_4^2 + \left(\frac{a_4}{2}\right)^2} = \frac{\sqrt{2}}{4} \approx 0.3535533905$$

We can plainly see that the perimeter of the square is somewhere between the circumferences of the two circles, so we can compare the perimeter, p_4, and the two circumferences, C_{in} and C_{circum}, to get

$$C_{\text{in}} < p_4 < C_{\text{circum}} \quad \text{or} \quad 2\pi\, h_4 < 2 < 2\pi\, r_4$$

Dividing all terms by 2 gives us

$$\pi\, h_4 < 1 < \pi\, r_4$$

Then dividing all terms by π we get

$$h_4 < \frac{1}{\pi} < r_4$$

Taking the reciprocal values of each of the terms reverses the inequality, so we get

$$\frac{1}{r_4} < \pi < \frac{1}{h_4}$$

Since $r_4 = \frac{\sqrt{2}}{4}$, it follows that $\frac{1}{r_4} = \frac{4}{\sqrt{2}} \approx 2.82842713$ and $\frac{1}{h_4} = 4$. Therefore $2.82842713 < \pi < 4$, a rather rough approximation for the value of π. But wait, as we increase the number of sides of the regular polygon, the estimates should get better.

The next approximation was done by doubling the number of sides of the regular polygon we just used to get a regular octagon. Cusanus considered a regular 8-gon with h_8 as the radius of the inscribed circle with circumference C_{in} and r_8 as the radius of the circumscribed circle with circumference C_{circum}.

Since each side is $a_8 = \frac{1}{4}$, the perimeter, p_8, of the regular 8-gon is $p_8 = 8 \cdot a_8 = 2$.

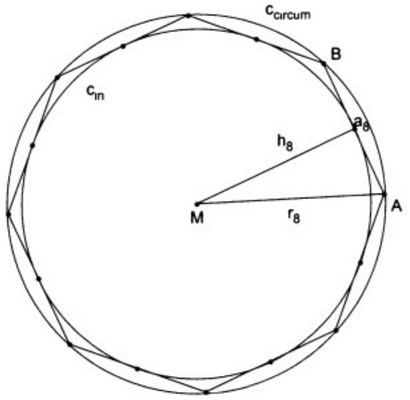

Fig. 3-10

$m\angle AMB = 45°$, therefore, $m\angle AMC = 22.5°$

$$\tan \angle AMC = \tan 22.5° = \sqrt{2} - 1 = \frac{\frac{a_8}{2}}{h_8} = \frac{a_8}{2h_8}$$ [9]

This can then be transformed into the following equation:

$$h_8 = \frac{a_8}{2\tan \angle AMC} = \frac{\frac{1}{4}}{2\bullet\left(\sqrt{2}-1\right)} = \frac{1}{8\left(\sqrt{2}-1\right)} = \frac{\sqrt{2}+1}{8} \approx 0.3017766952$$ [10]

With the Pythagorean theorem, $r_8^2 = h_8^2 + \left(\frac{a_8}{2}\right)^2$, we get

$$r_8 = \sqrt{\frac{\sqrt{2}}{32} + \frac{1}{16}} \approx 0.3266407412$$

For the perimeter, p_8, and circumferences, C_{in} and C_{circum}, we get

$$C_{\text{in}} < p_8 < C_{\text{circum}} \quad \text{or} \qquad 2\pi\, h_8 < 2 < 2\pi\, r_8$$

Dividing both sides by 2, it follows that $\pi h_8 < 1 < \pi r_8$, which, when dividing each term by π, yields $h_8 < \frac{1}{\pi} < r_8$.

---

[9] Tan $22.5° = \sqrt{2} - 1$ can be obtained by applying a theorem from the high school course to an isosceles right triangle with one of its base angles bisected to get 22.5°

The theorem states that the angle bisector of a triangle divides the side to which it is drawn proportionally to the two adjacent sides Hence,

$$\frac{\sqrt{2}}{1} = \frac{1-x}{x}\text{, then } x\sqrt{2} = 1 - x\text{, and } x = \frac{1}{\sqrt{2}+1}$$

Rationalizing the denominator gives us

$$x = \frac{1}{\sqrt{2}+1} \bullet \frac{\sqrt{2}-1}{\sqrt{2}-1} = \sqrt{2} - 1$$

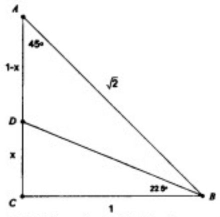

[10] This was obtained by rationalizing the denominator, that is, by multiplying by 1 in the form of $\frac{\sqrt{2}+1}{\sqrt{2}+1}$

Again, taking the reciprocal of each term reverses the inequality to give us

$$\frac{1}{r_8} < \pi < \frac{1}{h_8}$$

That means for the reciprocal values

$$\frac{1}{r_8} = \sqrt{32 - 16\sqrt{2}}$$

$$= 4\sqrt{2 - \sqrt{2}} \approx 3.061467458$$

and $\frac{1}{h_8} = 8\,(\sqrt{2} - 1) \approx 3.313708498$

We finally have a more accurate value range for π:

$$3.061467458 < \pi < 3.313708498$$

We will now take a giant leap to the general case, where we will try to sandwich in the value of π. This may be a bit complicated for the reader no longer familiar with some of the intricacies of high school mathematics, yet it is the conclusion of this generalization that is of greater importance than the process.

For the general case, h_n is the radius of the inscribed circle with circumference C_{in}, and r_n is the radius of the circumscribed circle with circumference C_{circum} of a regular n-gon ($n = 4, 8, 16, 32, \ldots$). We have established above that

$$\frac{1}{r_n} < \pi < \frac{1}{h_n}$$

In this way one can generalize nested intervals for π, if one iteratively[11] determines the radii of the inscribed and circumscribed circles with increasing numbers of sides of the regular n-

---

[11] Iteration is a computational procedure in which the desired result is approached through a repeated cycle of operations, each of which more closely approximates the desired result

gon (with the perimeter 2).

How did Cusanus get his iteration method? To explain this, we look again at a regular n-gon ($n = 4, 8, 16, 32, \ldots$):

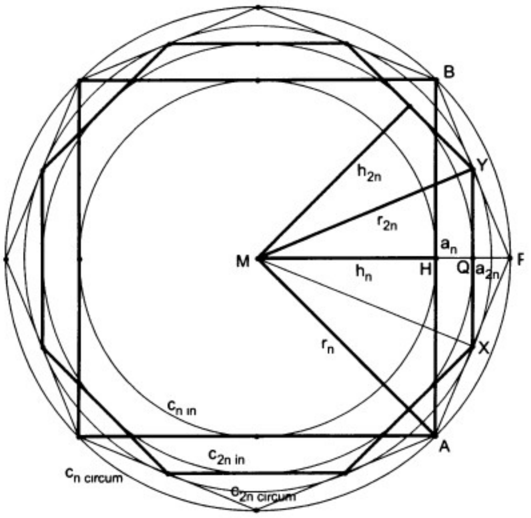

Fig. 3-11

We assume $AB = a_n$, $MA = MB = r_n$, and $MH = h_n$. After doubling the number of sides of the polygon, we get the regular $2n$-gon. Here P is the midpoint of the arc $\widehat{AB}$, and X and Y are the midpoints of the sides $\overline{AP}$ and $\overline{BP}$ in the triangle ΔABP. Therefore, $XY = \frac{AB}{2}$. $\overline{XY}$ is the side of the regular $2n$-gon with the perimeter 2 and the center M. It follows that $MP = MA = r_n$, $MX = MY = r_{2n}$, and $MQ = h_{2n}$ (compare the cases where $n = 4$ and $2n = 8$ in the above figure).

Because Q is the midpoint of the segment $\overline{PH}$, we have $h_{2n} = \frac{r_n + h_n}{2}$.

In the right triangle ΔMPX it follows that $MX^2 = MQ \bullet MP$.[12] This may be written as $r_{2n}^2 = r_n \bullet h_{2n}$, which then leads to

$$r_{2n} = \sqrt{r_n \bullet h_{2n}}$$

To generate the values for the rest of the n-gons (where n = 16, 32, 64, etc.), we can use the general case. We use the following general terms:

$$h_4 = \frac{1}{4}; r_4 = \frac{\sqrt{2}}{4} \qquad \text{(start values)}$$

$$h_{2n} = \frac{r_n + h_n}{2}; r_{2n} = \sqrt{r_n \bullet h_{2n}} \qquad \text{(iteration terms)}$$

These yield the following table of values:

n	h_n	r_n	$\frac{1}{r_n}$	$\frac{1}{h_n}$
4	0 25	0 3535533905	2 828427124	4
8	0 3017766952	0 3266407412	3 061467458	3 313708498
16	0 3142087182	0 3203644309	3 121445152	3 182597878
32	0 3172865746	0 3188217886	3 136548490	3 151724907
64	0 3180541816	0 3184377538	3 140331156	3 144118385
128	0 3182459677	0 3183418463	3 141277250	3 142223629
256	0 3182939070	0 3183178758	3 141513801	3 141750369
512	0 3183058914	0 3183118835	3 141572940	3 141632080
1024	0 3183088874	0 3183103855	3 141587725	3 141602510
2048	0.3183096365	0 3183100110	3 141591421	3 141595117
4096	0 3183098237	0 3183099173	3 141592345	3 141593269
8192	0 3183098705	0 3183098939	3 141592576	3 141592807
16384	0 3183098822	0 3183098881	3 141592634	3 141592692
32768	0 3183098852	0 3183098866	3 141592648	3 141592663

We have achieved seven decimal places accuracy for the value of π: 3.1415926. Here for puposes of comparison is the value of π correct to thirty-one decimal places:

π = 3.1415926535897932384626433832795...

---

[12] This comes from similar triangles MXP and MQX, or by applying the familiar mean proportional theorems.

While the method of Archimedes relies on the trigonometric functions sine and tangent, the method of Cusanus depends only on elementary theorems like the Pythagorean theorem, similarity, and the basic definition of the trigonometric functions. Furthermore, the arithmetic and geometric means are used for the iteration:

$$A(x,y)=\frac{x+y}{2}=\frac{r_n+h_n}{2}=h_{2n}$$
$$G(x,y)=\sqrt{x\bullet y}=\sqrt{r_n\bullet h_{2n}}=r_{2n}$$

Calculation of the Value of π by Counting Squares

It is always challenging to determine the actual value of π. There isn't any arithmetically comfortable method for calculating the value of π. On the one hand, we only need an elementary knowledge of mathematics for the following methods, but, on the other hand, these approximations of the value of π are not as exact as the calculations done by Archimedes or Cusanus. We offer here a few relatively simple methods for calculating the value of π.

To determine the area of a circle, we can cover it with a lattice of squares (each with a side of length 1), and we will count the number of squares (*a*) in the interior of the circle. We then will count the number of squares (*b*) that are intersected by the circle's circumference.

We will *assume*[13] that one-half of the area of these intersected squares lies in the interior of the circle, and the other half of the area of these intersected squares lies outside the circle.

---

[13] It is this assumption that will limit our accuracy of the approximation of π.

So we have $a + \frac{b}{2}$ as an approximate value for the area of the circle. Let's consider this with the following example.

Example 1:
Circle with the radius $r = 8$

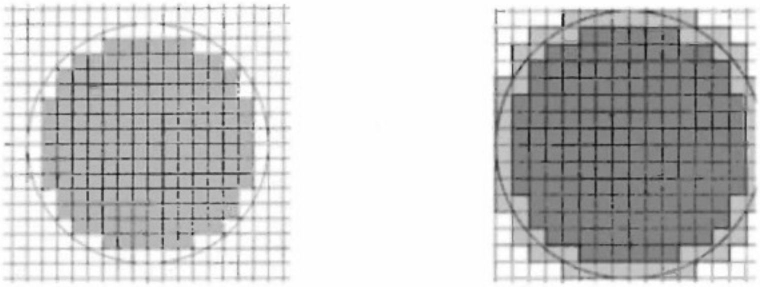

$a = 14 \cdot 6 + 12 \cdot 4 + 10 \cdot 2 + 6 \cdot 2 = 164$ $b = 60$

Fig. 3-12

By the known method (using the formula for the area of a circle) we get

$Area_{\text{circle}} = \pi \cdot r^2 \approx 3.14 \cdot 8^2 = 200.96$

Now using the counting method, we get the following result, which compares favorably with the traditional method above.

Approximate value: $Area_{\text{circle}} \approx a + \frac{b}{2} = 164 + 30 = 194$, which is close to the "actual" 201 found by the formula.

This approximation leads to π as follows: The approximate area of the circle is 194. This should then be equal to $8^2 \pi = 64\pi$, so that $64\pi \approx 194$. This gives a value for $\pi = \frac{194}{64} \approx 3.03125$.

The approximation of π becomes better when a larger number of squares is used, say, when $r = 10$:

Example 2:
Circle with the radius $r = 10$

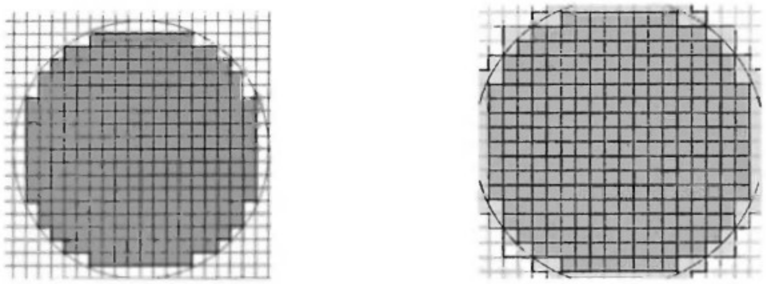

$a = 18 \cdot 8 + 16 \cdot 4 + 14 \cdot 2 + 12 \cdot 2 + 8 \cdot 2 = 276$ $b = 68$

Fig. 3-13

$Area_{\text{circle}} = \pi \cdot r^2 \approx 3.14 \cdot 10^2 = 314$

Approximate value: $Area_{\text{circle}} \approx a + \frac{b}{2} = 276 + 34 = 310$, which is now closer to the "actual" value, 314, as determined by the formula above. Again, $Area_{\text{circle}} = 310 = 10^2\pi = 100\pi$. Thus, $\pi \approx \frac{310}{100} = 3.1$, which compares favorably to the previous approximation.

Instead of the whole circle, it suffices to look at a quadrant, count the respective squares, and then quadruple it.

Calculating the Value of π by Counting Lattice Points

The method used by the German mathematician Carl Friedrich Gauss (1777–1855) is relatively simpler. Instead of counting the number of squares, he determined the area of a circle by counting the number of lattice points of the square lattice in the interior of the circle. Lattice points are points with integer coordinates. We can locate all the lattice points of the circle with radius r with $x^2 + y^2 \leq r^2$ (the Pythagorean theorem).

If $f(r)$ represents the number of lattice points that lie in the circular area with the radius r, then we get (with the help of Gauss's idea) an approximation for π:

$$\pi \approx \frac{f(r)}{r^2}$$

This should bring back thoughts (especially if you were to multiply both side of this "almost-equation" by r^2) of the now-famous formula for the area of a circle, $Area = \pi r^2$. Here we have $\pi r^2 \approx f(r)$.

There is a formula for finding $f(r)$, but this formula is complicated. Instead, we will give some examples. Consider the following.

Example:

$$f(r) = 317,\ r^2 = 100, \text{ therefore, } \frac{f(r)}{r^2} = 3.17 \approx \pi$$

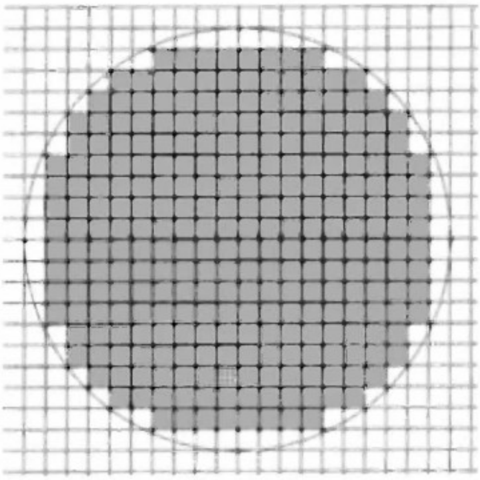

Fig. 3-14

Instead of the whole circle, it suffices to look at a quadrant and to count the respective lattice points again. (Be careful to count the origin only once.)

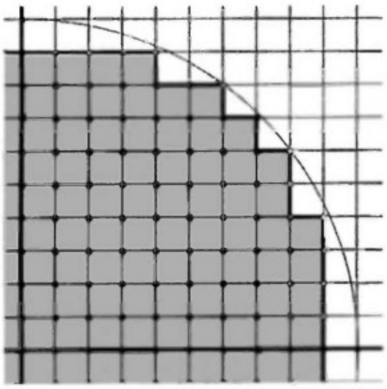

Fig. 3-15

Further values:

r	5	10	20	30	100	200	300	... 100,000
$f(r)$	81	317	1,257	2,821	31,417	125,629	282,697	31,415,925,457
$\frac{f(r)}{r^2}$	3.24	3.17	3.1425	3.134	3.1417	3.140725	3.14107	3.1415925457

It appears that this sequence heads for the actual value of π, 3.1415926....

For $r = 20$, we already get the correct second place after the decimal point. Strangely enough, for $r = 30$ the value of π gets less accurate, but then eventually gets closer to the true value of π.

Using Physical Properties to Calculate the Value of π

A physicist might determine the value of π by using what may be considered a simpler method than was involved with the tedious task of counting squares or lattice points.

He would weigh (as exactly as possible) one circle with a radius of 10 cm, which he would cut out of an evenly thick piece of cardboard. Then he would compare its weight (or its mass) with that of a square (of 10 cm length side) cut out of the same material.

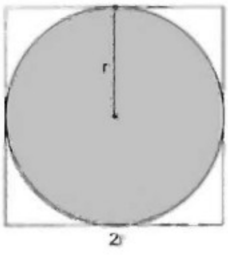

Fig. 3-16

We now compare the area of the circle, radius r, to the area of the square, side $2r$, by considering their weights.[14]

$\frac{m_{circle}}{m_{square}} = \frac{(2r)^2}{\pi r^2} = \frac{4r^2}{\pi r^2} = \frac{4}{\pi} \approx 1.273239$.[15] Therefore, $\pi \approx 3.141594$.

In the eighteenth century, the French agronomist Franzose Olivier de Serres "proved," by using a scale, that a circle weighs as much as a square whose side has the length of that of an equilateral triangle inscribed in the circle—this assumes that both figures are cut out of the same material. When you follow the discussion below, you will see that this implies that $\pi = 3$!

---

[14] Technically speaking, we actually have a cylinder and a rectangular solid, if one were to consider the width of the cardboard as the height of these objects.

[15] Actually a constant value

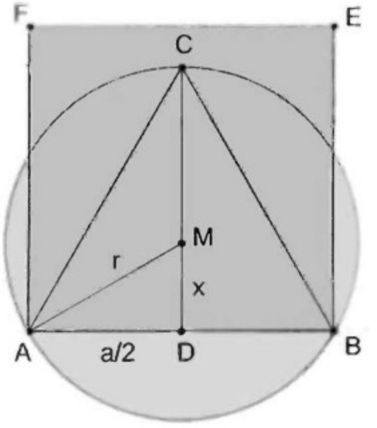

in ΔADM: $m\angle DAM = \frac{1}{2}(60°) = 30°$

$\sin\angle DAM = \frac{x}{r} = \frac{1}{2}$; therefore, $x = \frac{r}{2}$

$\left(\frac{a}{2}\right)^2 = r^2 - x^2 = \frac{3}{4}r^2$; therefore, $\frac{a}{2} = \frac{\sqrt{3}}{2}$

Area of the square $= a^2 = \left(r\sqrt{3}\right)^2 = \mathbf{3}\, r^2$

Area of the circle $= \boldsymbol{\pi}\, r^2$

We can then conclude that $\pi = 3$.

Fig. 3-17

The Monte-Carlo Method to Determine the Value of π

The Monte-Carlo method[16] is a procedure that makes use of probability, calculus, and statistics to form a summary to establish facts with a large number of tests for a random experiment. The Buffon's needle problem (see chapter 1, pages 38–39) is considered one of the Monte-Carlo methods.

Another such procedure can be simulated by means of raindrops that fall on a predefined square, or similarly by using a number of random dart throws. This "dartboard algorithm," which can be used in a school setting, shall serve as an example here.

To do this, a chance rain is simulated and the hits counted within and outside the inscribed circle (with radius of length 1) on a square with side length 2. Using dart throws, instead of raindrops, may be a better procedure. With the following considerations, one then reaches a possibility to determine the value of π.

---

[16] The name is taken from the gambling paradise

The relationship between the hits in the circle and the complete number of throws yields an approximation for π:

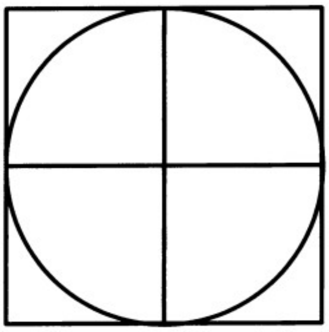

square
side: $2r = 2$
area: $A_{SQ} = (2r)^2 = 4$

circle
radius: $r = 1$
area: $A_C = \pi \bullet r^2 = \pi$

Fig. 3-18

$$\frac{A_C}{A_{SQ}} = \frac{\pi \bullet r^2}{4r^2} = \frac{\pi}{4}$$

$$\frac{\pi}{4} = \text{probability (circle hits)} = \frac{\text{number of circle hits}}{\text{number of throws}}$$

The method yields a good approximation for π only after a very large number of throws. The randomizer (the dart thrower or water dropper) must produce really coincidental numbers and may not be subject to any regularities, that is, the person may not influence where the darts or water droplets fall.

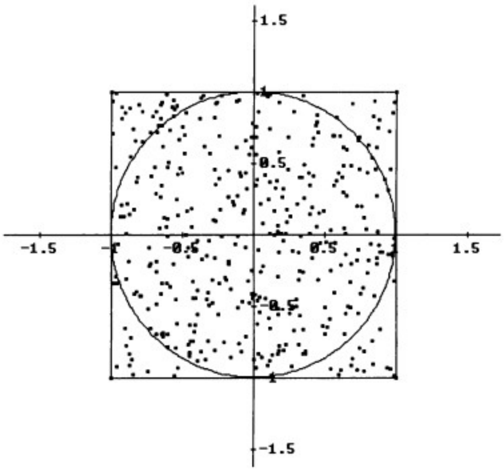

Fig. 3-19

For the first quadrant we get, for example, in the case of ten throws, that the first and the fourth throws don't satisfy the condition $x^2 + y^2 \leq 1$. Therefore, since two points are outside the target region, only eight points are drawn:

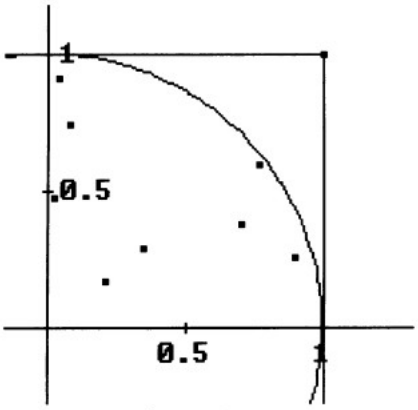

Fig. 3-20

For the calculation of the ratios of the areas, a Monte-Carlo integration is now used. One may proceed as follows:

- With a random numbers generator, an x and a y value between 0 and $2r$ are "thrown."
- The Pythagorean theorem is used to check if the thrown point $P(x, y)$ lies within or outside the circle.
- The hits in the circle are counted.
- The procedure is repeated—the more often repeated the more accurate the expected value of π will be.

Calculating π from a Series of Numbers

Earlier we considered the following formula for π, which was developed by the famous German mathematician Gottfried Wilhelm Leibniz (1646–1716), who, together with Isaac Newton, is credited with developing the modern calculus:

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \frac{1}{11} + \cdots$$

Leibniz, who is also considered one of the great philosophers of the Western world, commented on the unusual connection between the number π and the pattern of alternately adding and subtracting odd unit fractions with the words "Numero deus impare gaudet" (God is happy with the odd number).

We mentioned then that it approaches the value of π rather slowly, since it will take one hundred thousand terms to get to a five-place accuracy of π, and for six-place accuracy we will need to carry this series out for one million terms.

Let's take a look at how this series "behaves." We multiply both sides by 4:

$$\pi = 4\bullet\left(1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\frac{1}{9}-\frac{1}{11}+\frac{1}{13}-\frac{1}{15}\pm\ldots\right) = 4\bullet\sum_{i=1}^{\infty}\frac{(-1)^{i+1}}{2i-1}$$ [17]

n	**Partial Sum**	**Exact Value**	**Approximate Value**
1	$4\bullet 1$	4	4
2	$4\bullet\left(1-\frac{1}{3}\right)$	$\frac{8}{3}$	2.666666666
3	$4\bullet\left(1-\frac{1}{3}+\frac{1}{5}\right)$	$\frac{52}{15}$	**3**.466666666
4	$4\bullet\left(1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}\right)$	$\frac{304}{105}$	2.895238095
5	$4\bullet\left(1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\frac{1}{9}\right)$	$\frac{1{,}052}{315}$	**3**.339682539
6	$4\bullet\left(1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\frac{1}{9}-\frac{1}{11}\right)$	$\frac{10{,}312}{3{,}465}$	2.976046176
7	$4\bullet\left(1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\frac{1}{9}-\frac{1}{11}+\frac{1}{13}\right)$	$\frac{147{,}916}{45{,}045}$	**3**.283738483
8	$4\bullet\left(1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\frac{1}{9}-\frac{1}{11}+\frac{1}{13}-\frac{1}{15}\right)$	$\frac{135{,}904}{45{,}045}$	**3**.017071817
9	$4\bullet\left(1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\frac{1}{9}-\frac{1}{11}+\frac{1}{13}-\frac{1}{15}+\frac{1}{17}\right)$	$\frac{2{,}490{,}548}{765{,}765}$	**3**.252365934
10	$4\bullet\left(1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\frac{1}{9}-\frac{1}{11}+\frac{1}{13}-\frac{1}{15}+\frac{1}{17}-\frac{1}{19}\right)$	$\frac{44{,}257{,}352}{14{,}549{,}535}$	**3**.041839618
100	$4\bullet\left(1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}\pm\quad-\frac{1}{199}\right)$	numerator and denominator have 88 places (see below)	**3.1**31592903
1,000	$4\bullet\left(1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}\pm\quad-\frac{1}{1{,}999}\right)$	–	**3.1**31592902
10,000	$4\bullet\left(1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}\pm\quad-\frac{1}{19{,}999}\right)$	–	**3.141**492653
100,000	$4\bullet\left(1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}\pm\quad-\frac{1}{199{,}999}\right)$	–	**3.1415**82653

From the chart above, when $n = 100$ (exact value):

$$\frac{8252079759413970386664454687621174435983101115012912631997769614579677862845786070667088}{2635106162757236442495826303084698495556581115509040892412867358728390766099042109898375}$$

---

[17] This expression is merely a mathematical shorthand way of writing the series.

The approximation value "jumps" back and forth around the actual value of π, since the terms are adding or subtracting alternately; it is, therefore, alternately bigger or smaller than π.

We compare the approximate value for $n = 100{,}000$ (the last entry in the chart) with the correct value of π = **3.1415**9265358979323 84..., and we have indeed (only) four correct places. Then as we increase the value of n, the approximation for π becomes increasingly more accurate (i.e., closer to the true value of π).

$n = 1{,}000{,}000$: approximation value = **3.141591**653
$n = 10{,}000{,}000$: approximation value = **3.141592**553
$n = 100{,}000{,}000$: approximation value = **3.1415926**43

A Better Series Calculation for the Value of π

There are other series that converge[18] to the value of π faster than the Leibniz series.

Earlier we mentioned the following formula for deriving the value of π, which was discovered by the famous Swiss mathematician Leonhard Euler:

$$\frac{\pi^2}{6} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \cdots$$

or (after multiplication by a 6 and square root extraction):

$$\pi = \sqrt{6 \bullet \left(1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \cdots\right)} = \sqrt{6 \bullet \sum_{i=1}^{\infty} \frac{1}{i^2}}$$

This formula is sensationally fast in comparison to that of Leibniz. In the case of the Leibniz series, the computer needed more than two and a half hours for $n = 10^8$; the Euler series[19] delivers the accompanying approximation value for $n = 10^8$ in virtually zero seconds of computer time!

---

[18] To "converge," in this sense, means to approach a particular value as a limit

The approximate values of π are about the same quality as before. For $n = 10^8$ we get eight-place accuracy: π ≈ **3.1415926**44. Let's take a look at some partial sums:

n	Partial Sum	Exact Value	Approximate Value
1	$\sqrt{6\cdot\frac{1}{1^2}}$	$\sqrt{6}$	2.449489742
2	$\sqrt{6\cdot\left(1+\frac{1}{2^2}\right)}$	$\frac{\sqrt{30}}{2}$	2.738612787
3	$\sqrt{6\cdot\left(1+\frac{1}{2^2}+\frac{1}{3^2}\right)}$	$\frac{7\sqrt{6}}{6}$	2.857738033
4	$\sqrt{6\cdot\left(1+\frac{1}{2^2}+\frac{1}{3^2}+\frac{1}{4^2}\right)}$	$\frac{\sqrt{1,230}}{12}$	2.922612986
5	$\sqrt{6\cdot\left(1+\frac{1}{2^2}+\frac{1}{3^2}+\frac{1}{4^2}+\frac{1}{5^2}\right)}$	$\frac{\sqrt{32,214}}{60}$	2.963387701
10	$\sqrt{6\cdot\left(1+\frac{1}{2^2}+\frac{1}{3^2}+\quad+\frac{1}{9^2}+\frac{1}{10^2}\right)}$	$\frac{\sqrt{59,049,870}}{2,520}$	**3**.049361635
100	$\sqrt{6\cdot\left(1+\frac{1}{2^2}+\frac{1}{3^2}+\quad+\frac{1}{99^2}+\frac{1}{100^2}\right)}$	**See Below***	**3.1**32076531
1,000	$\sqrt{6\cdot\left(1+\frac{1}{2^2}+\frac{1}{3^2}+\quad+\frac{1}{999^2}+\frac{1}{1,000^2}\right)}$		**3.14**0638056
10,000	$\sqrt{6\cdot\left(1+\frac{1}{2^2}+\frac{1}{3^2}+\quad+\frac{1}{9,999^2}+\frac{1}{10,000^2}\right)}$		**3.141**497163
100,000	$\sqrt{6\cdot\left(1+\frac{1}{2^2}+\frac{1}{3^2}+\quad+\frac{1}{99,999^2}+\frac{1}{100,000^2}\right)}$		**3.1415**83104
1,000,000	$\sqrt{6\cdot\left(1+\frac{1}{2^2}+\frac{1}{3^2}+\quad+\frac{1}{(10^6)^2}\right)}$		**3.14159**1698
10,000,000	$\sqrt{6\cdot\left(1+\frac{1}{2^2}+\frac{1}{3^2}+\quad+\frac{1}{(10^7)^2}\right)}$		**3.141592**558
100,000,000	$\sqrt{6\cdot\left(1+\frac{1}{2^2}+\frac{1}{3^2}+\quad+\frac{1}{(10^8)^2}\right)}$		**3.1415926**44
1,000,000,000	$\sqrt{6\cdot\left(1+\frac{1}{2^2}+\frac{1}{3^2}+\quad+\frac{1}{(10^9)^2}\right)}$		**3.14159265**2

* $n = 100$ (exact value):

$$\frac{\sqrt{47685260823991136193368937855526615791071504956311295796682985044114543932711097030}}{69720375229712477164533808935312303556800}$$

---

[19] Euler also found the sums of the reciprocals of the fourth and sixth powers (by the way, to date no one has been able to do this for third powers!).

For purposes of comparison, here is the value of π:

$\pi = 3.14159265358979323 84...$

The Genius's Method for Finding the Value of π

The extraordinarily brilliant Indian mathematician Srinivasa Ramanujan (1887–1920) made contributions to the generation of the value of π but left little evidence on how he arrived at his results. Born in 1887 in the small south Indian town of Erode, Ramanujan spent his youth fascinated with mathematics to the detriment of other subjects. As the fledgling Indian Mathematical Society was being founded, it provided a forum for Ramanujan to exhibit his mathematical prowess. For example, in 1911 he posed problems based on his earlier work and found no solvers among the readership. One such example was to evaluate

$$\sqrt{1+2\sqrt{1+3\sqrt{1+4\sqrt{1+\cdots}}}},$$

which appeared harmlessly simple, but yet found no successful solvers. The trick was found in his notebook of theorems that he established. Here, he simply applied the following theorem, which said that if you could represent a number as $(x + n + a)$, the above expression could be represented as

$$x+n+a=\sqrt{ax+(n+a)^2+x\sqrt{a(x+n)+(n+a)^2+(x+n)\sqrt{\cdots}}}$$

So if $3 = x + n + a$, where, say, $x = 2$, $n = 1$, and $a = 0$, then the value of this nest of radicals is simply equal to 3. This is nearly impossible to do without knowledge of Ramanujan's theorem.

With this new exposure, he wrote to three of the top mathematicians in England, E. W. Hobson, H. F. Baker, and G. H. Hardy.[20] Of these three Cambridge professors, only Godfrey Harold Hardy (1877–1947) responded and ultimately invited Ramanujan to England. Hardy thought that the statements contained in the letter had to be correct. For, if they were wrong, nobody would have had such a wild imagination to make them up.

Despite a clash of cultures, the two got along very well and mutually assisted each other. This was the beginning of Ramanujan's popularity outside India. It should be remembered that, even though wearing shoes and using eating utensils were new to this Indian, he came from a long heritage of mathematical culture. The Indians were using our numeration system (including the zero) for over a thousand years before it was introduced in Europe with the publication of Fibonacci's book, *Liber abaci,* in 1202.

---

[20] The letter to Hardy, dated "Madras, 16th January 1913" and which enticed Hardy to respond, was the following

Dear Sir,

I beg to introduce myself to you as a clerk in the Accounts Department of the Port Trust Office at Madras on a salary of only £ 20 per annum. I am now about 23 years of age. I have no University education but I have undergone the ordinary school course. After leaving school I have been employing the spare time at my disposal to work at mathematics. I have not trodden through the conventional regular course which is followed in a University course, but I am striking out a new path for myself. I have made a special investigation of divergent series in general and the results I get are termed by the local mathematicians as "startling.". . .

I would request you to go through the enclosed papers. Being poor, if you are convinced that there is anything of value I would like to have my theorems published. I have not given the actual investigations nor the expressions that I get but I have indicated the lines on which I proceed. Being inexperienced I would very highly value any advice you give me. Requesting to be excused for the trouble I give you.

I remain,
Dear Sir,
Yours truly,
S. Ramanujan.

(As reprnted in Robert Kanigel, *The Man Who Knew Infinity: A Life of the Genius Ramanujan* [New York: Charles Scribner's Sons, 1991], pp. 159–60.)

In our context, Ramanujan came up with some amazing results in the determination of the value of π. He empirically (his word) obtained the approximate value of π with the following expression:

$$\left(9^2+\frac{19^2}{22}\right)^{\frac{1}{4}}=\left(81+\frac{361}{22}\right)^{\frac{1}{4}}=\left(\frac{2{,}143}{22}\right)^{\frac{1}{4}}$$

$$=\left(97.4\overline{09}\right)^{\frac{1}{4}}$$

$$\approx 3.14159265258264612520603717964402237155 7\ldots$$

He further stated that the value he used for π for purposes of calculation was

$$\frac{355}{113}\left(1-\frac{.0003}{3{,}533}\right)=3.1415926535897943\ldots,$$

which he went on to say "is greater than π by about 10^{-15}" and "is obtained by simply taking the reciprocal of

$$1-\left(\frac{113\pi}{355}\right)=11{,}776{,}666.6185424743744652803554 3.\ldots\text{"}^{21}$$

The more ambitious reader can find the justification for this work in appendix B.

Srinivasa Ramanujan gave us other uncanny approximations of π. Even today we are mystified by how he arrived at the various results. Although we are becoming more able to understand his derivations, we still cannot fully appreciate the complexity of the way his unique mind functioned. The following are some of his findings on the value of π.

---

[21] Srinivasa Ramanujan, "Modular Equations and Approximations to π," *Quarterly Journal of Mathematics* 45 (1914): 350–72. Reprinted in *S. Ramanujan: Collected Papers*, ed. G. H. Hardy, P. V Seshuaigar, and B. M. Wilson (New York: Chelsea, 1962), pp. 22–39.

In chapter 2 we have already mentioned the following formula:

$$\frac{1}{\pi}=\frac{\sqrt{8}}{9{,}801}\sum_{n=0}^{\infty}\frac{(4n)!(1{,}103+26{,}390n)}{(n!)^4\,396^{4n}}$$

Another formula is

$$\frac{1}{\pi}=\sum_{n=0}^{\infty}\binom{2n}{n}^3\frac{42n+5}{2^{12n+4}}$$

The following are some approximations of π that are due to Ramanujan:

$\frac{355}{113}\approx 3.141592920$ was originally discovered by Adriaen Métius (1571-1635),[22] and later Ramanujan gave a geometric construction for this term.

$\frac{9}{5}+\sqrt{\frac{9}{5}}\approx 3.141640786$

$\frac{19\sqrt{7}}{16}\approx 3.141829681$

Some series discovered by Ramanujan follow. However, the important point is that evaluating such series to huge numbers of digits requires developing specific algorithms.

$$\frac{1}{\pi}=\frac{2\sqrt{2}}{9{,}801}\sum_{k=0}^{\infty}\frac{(4k)!}{(k!)^4\,4^{4k}}\frac{(1{,}103+26{,}390k)}{99^{4k}} \qquad [\text{Ramanujan}]$$

$$\frac{1}{\pi}=12\sum_{k=0}^{\infty}(-1)^k\frac{(6k)!}{(3k)!(k!)^3}\frac{(13{,}591{,}409+545{,}140{,}134k)}{640{,}320^{3k+3/2}} \qquad [\text{Chudnovsky}]$$

---

[22] He and his father, Adriaen Anthoniszoon (c. 1600), took the approximation $3\frac{15}{106}<\pi<3\frac{17}{120}$, added the numerators (15 + 17 = 32) and the denominators (106 + 120 = 226), took the means (16 and 113), and gave $\pi=3\frac{16}{113}=\frac{355}{113}=3.1415929$, a very close approximation. (D. E. Smith, *History of Mathematics*, vol. 2 [New York: Dover, 1958])

Here are some more of the wondrous discoveries by the genius Ramanujan that you may wish to ponder.

$$\frac{2}{\pi}=1-5\left(\frac{1}{2}\right)^3+9\left(\frac{1\bullet 3}{2\bullet 4}\right)^3-13\left(\frac{1\bullet 3\bullet 5}{2\bullet 4\bullet 6}\right)^3+\ldots$$

$$\frac{4}{\pi}=1+\left(\frac{1}{2}\right)^2+\left(\frac{1}{2\bullet 4}\right)^2+\left(\frac{1\bullet 3}{2\bullet 4\bullet 6}\right)^2+\left(\frac{1\bullet 3\bullet 5}{2\bullet 4\bullet 6\bullet 8}\right)^2+\ldots \qquad [\text{Forsyth}]$$

$$\frac{1}{\pi}=\frac{1}{72}\sum_{k=0}^{\infty}(-1)^k\frac{(4k)!}{(k!)^4\,4^{4k}}\frac{(23+260k)}{18^{2k}}$$

$$\frac{1}{\pi}=\frac{1}{3{,}528}\sum_{k=0}^{\infty}(-1)^k\frac{(4k)!}{(k!)^4\,4^{4k}}\frac{(1{,}123+21{,}460k)}{882^{2k}}$$

$$\frac{1}{\pi}=12\sum_{k=0}^{\infty}(-1)^k\frac{(6k)!}{(3k)!(k!)^3}\frac{(A+Bk)}{C^{3k+3/2}} \qquad [\text{Borwein}]$$

In the last formula

$$A = 1{,}657{,}145{,}277{,}365 + 212{,}175{,}710{,}912\ \sqrt{61}$$

$$B = 107{,}578{,}229{,}802{,}750 = 13{,}773{,}980{,}892{,}672\ \sqrt{61}$$

$$C = 5{,}280\ (236{,}674 + 30{,}303)\ \sqrt{61}$$

and each additional term in the series adds about thirty-one digits.

We have seen numerous ways that the value of π was calculated. Some were primitive, while others were quite sophisticated. Most remarkably are those that would have appeared to have evolved from spectacular guesses. Today's methods all involve the computer, and how accurate the future calculations of the value of π will be is going to be merely limited by man's creativity and the computer's ability.

0 ()*+(((, (.)0 * / " 1
; ")// / $ /"

()*+(((, (.)"A " <) % "

& - . . 1 0 " 1 ;

In modern physics the color of light is determined by the frequency of oscillation of an electromagnetic wave, and the pitch of a musical note is determined by the frequency of oscillation of a vibrating string, membrane, or air column in a musical instrument. But the timbre of a musical instrument is determined by combining numerous oscillations of different frequencies, each with its own characteristic amplitude. Thus the recognizable differences of *concert tone A* on a flute, a piano, or a violin reflects in each case the summation of many different frequencies of relatively small amplitude added to the lowest frequency, which determines the perceived pitch.

In all the examples listed above, complex oscillations are analyzed as sums of many simple oscillations. This subject is known as *harmonic analysis*; it is known also as *Fourier analysis* after Joseph Fourier, who successively applied it to analyze the flow of heat through a metal rod in 1822. There are many distinguished books about Fourier series, though most presume knowledge of the Lebesgue integral. The author recommends especially the book by Dym and McKean [6]. It provides a useful summary of the needed properties of the Lebesgue integral in the beginning.

6.1 THE VIBRATING STRING AND TRIGONOMETRIC SERIES

We will begin with an important example from physics,[15] which explains why one might wish to express all functions from a broad class as the sums of infinite series of sine waves and cosine waves.

■ EXAMPLE 6.1

Suppose a piece of music wire (such as a violin, guitar, or piano string) is stretched taut and fastened down at the endpoints of the interval $[0,1]$ on the x-axis. Now either a finger or an implement is used to stretch the string away from the axis, into some initial plane curve described by a displacement function $f(x)$. At time $t = 0$ the string is released, and, depending on the manner in which the plucking or hammering of the string was done, the string has an initial velocity at each point $x \in [0,1]$ given by $g(x)$. This leads to the following partial differential equation governing the function $u(x,t)$, which is the displacement of the string at coordinate x and at time t.

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2}. \tag{6.1}$$

Here a is a constant reflecting the tension to which the string is tightened and the density of the steel in the string. Since our purpose here is to motivate a mathematical problem and not to keep track of the effect of physical constants on the solution, *we will take* $a = 1$ *henceforth*.

The boundary conditions on the string and the initial conditions on the displacement and velocity at time $t = 0$ are given as follows.

---

[15] Neither first-year physics nor differential equations (ordinary or partial) is required for this course. The Example is presented for physical motivation and historical perspective.

$$\begin{cases} u(0,t) &= 0, \\ u(1,t) &= 0, \\ u(x,0) &= f(x), \\ \frac{\partial u}{\partial t}(x,0) &= g(x). \end{cases} \tag{6.2}$$

What is needed is a technique that may be enable us to construct a function $u(x, t)$ which satisfies Equation (6.1) and conditions (6.2). A classic method of solution is to seek solutions having the special simplified form

$$u(x,t) = F(x)G(t),$$

where F and G must be determined. (This method is called *separation of variables*.) Substitution into Equation (6.1) yields

$$\frac{F''(x)}{F(x)} = \frac{G''(t)}{G(t)},$$

which proves that both sides must be equal to the same constant since the left side is independent of t and the right side is independent of x. If that constant were a *nonnegative* number λ^2, then we would have

$$F''(x) = \lambda^2 F(x), \tag{6.3}$$

which has the general solution

$$F(x) = c_1 e^{\lambda x} + c_2 e^{-\lambda x}.$$

But then there would be no solution for $F(x)$ other than the identically zero solution given the boundary conditions (6.2). (See Exercise 6.5.)

Thus we take the constant to be a *negative* number $-\lambda^2$, which yields the general solution

$$F_n(x) = c \sin 2n\pi x \tag{6.4}$$

for some integer $n \in \mathbb{Z}$ and an arbitrary constant c. Similar work for the function G yields a sequence of corresponding solutions

$$u_n(t) = (A_n \cos 2n\pi t + B_n \sin 2n\pi t) \sin 2\pi n x.$$

We observe that Equation (6.1) will be satisfied by any function of the form

$$u(x,t) = \sum_{n=0}^{\infty} (A_n \cos nt + B_n \sin nt) \sin nx, \tag{6.5}$$

provided that all questions of convergence of the series itself and its twice derived series can be managed.

The goal of this example is to satisfy both the *boundary* conditions and the *initial* conditions in (6.2). In order to achieve this, we will need to be able to express both f and g as sums of infinite *trigonometric series*:

$$f(x) = \sum_{n=1}^{\infty} A_n \sin 2n\pi x$$

and

$$g(x) = \sum_{n=1}^{\infty} 2n\pi B_n \cos 2n\pi x.$$

Trigonometric Series

Thus the problem of describing the motion of a vibrating string leads us to the importance of determining which functions f and g can be expressed as *convergent* infinite series of sine and cosine functions of various periods. And we will need a method of computation of the necessary coefficients of such expansions.

As a problem of pure mathematics, we ask for which functions f defined on $[0, 1]$ it is possible to expand

$$f(x) = \sum_{n=0}^{\infty} a_n \cos 2n\pi x + b_n \sin 2n\pi x \tag{6.6}$$

in a suitably convergent manner. (We have included the index $n = 0$ in the summation in order to allow for nonzero constant functions.) Also, we need to ask whether given such a function f we can find the coefficients a_n and b_n.

Theorem 6.1.1 If *the series in Equation (6.6) converges* uniformly *on* $[0, 1]$ *to some function* f, *then* $f \in \mathcal{C}[0, 1]$,

$$\begin{cases} a_n = & 2\int_0^1 f(x) \cos 2n\pi x \, dx, \\ b_n = & 2\int_0^1 f(x) \sin 2n\pi x \, dx \end{cases} \tag{6.7}$$

for each $n > 0$, *and* $a_0 = \int_0^1 f(x)\, dx$. *Moreover,* if *the series*

$$\sum_{n=0}^{\infty} (-2n\pi a_n \sin 2n\pi x + 2n\pi b_n \cos 2n\pi x)$$

converges uniformly *on* $[0, 1]$, *then* f *is differentiable and*

$$f'(x) = \sum_{n=0}^{\infty} -2n\pi a_n \sin 2n\pi x + 2n\pi b_n \cos 2n\pi x. \tag{6.8}$$

Proof: Since each summand on the right side of Equation (6.6) is a continuous function on $[0, 1]$, the uniform convergence of this series guarantees the continuity of

CHAPTER 8

EUCLIDEAN SPACE

8.1 EUCLIDEAN SPACE AS A COMPLETE NORMED VECTOR SPACE

Throughout pure and applied mathematics, it is necessary to consider functions of more than one variable. If a function depends on n real variables, $x_1, x_2, \ldots, x_n$, it is possible to combine these n real variables into one *vector variable*

$$\mathbf{x} = (x_1, x_2, \ldots, x_n).$$

In this notation, $\mathbf{x}$ is not a real number, but rather an n-tuple of real numbers. In addition, it is often necessary to consider functions that have vector values instead of real values. Up to this point, our course has focused on real-valued functions of one real variable. In this chapter we will begin the rigorous study of vector-valued functions of vector variables.

In the study of real-valued functions of a single real variable, we saw the advantages of considering normed vector spaces of more than one dimension. For example, the reader has seen that the vector space $C[a, b]$ has infinitely many dimensions (Exercise 2.57). In the present chapter, we will focus on finite-dimensional vector spaces equipped with what is called the *Euclidean* norm. However, we begin with a more general context.

Advanced Calculus: An Introduction to Linear Analysis. By Leonard F. Richardson

Copyright © 2008 John Wiley & Sons, Inc.

Definition 8.1.1 *In any (real) vector space V (Tabel 2.1, p. 59), we call a function*

$$\langle\cdot,\cdot\rangle : V \times V \to \mathbb{R}$$

a scalar product *if and only if it has the following three properties.*

i. $\langle a\mathbf{x} + \mathbf{y}, \mathbf{z}\rangle = a\langle \mathbf{x}, \mathbf{z}\rangle + \langle \mathbf{y}, \mathbf{z}\rangle$ *for all* $a \in \mathbb{R}$ *and for all* $\mathbf{x}$ *and* $\mathbf{y}$ *in* V.

ii. $\langle \mathbf{x}, \mathbf{y}\rangle = \langle \mathbf{y}, \mathbf{x}\rangle$ *for all* $\mathbf{x}$ *and* $\mathbf{y}$ *in* V.

iii. $\langle \mathbf{x}, \mathbf{x}\rangle \geq 0$ *for all* $\mathbf{x} \in V$ *and* $\langle \mathbf{x}, \mathbf{x}\rangle = 0 \Leftrightarrow \mathbf{x} = \mathbf{0} \in V$.

■ EXAMPLE 8.1

In the vector space $\mathbb{R}^n$, define the *Euclidean* scalar product by

$$\langle \mathbf{x}, \mathbf{y}\rangle = \sum_{i=1}^{n} x_i y_i. \tag{8.1}$$

The reader should verify that this product satisfies all three conditions to be called a scalar product.

Theorem 8.1.1 *In any (real) vector space V equipped with a scalar product as defined in Definition 8.1.1, we define*

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle} \tag{8.2}$$

for all $\mathbf{x} \in V$. *The function* $\|\cdot\|$ *as defined in Equation (8.2)* is *a* norm, *as in Definition 2.4.4, and the* Cauchy–Schwarz *Inequality is satisfied:*

$$|\langle \mathbf{x}, \mathbf{y}\rangle| \leq \|\mathbf{x}\|\|\mathbf{y}\|.$$

Proof: To prove the Cauchy–Schwarz inequality, we fix $\mathbf{x}$ and $\mathbf{y}$, and we proceed as follows. For all $t \in \mathbb{R}$, define a polynomial

$$p(t) = \langle t\mathbf{x} + \mathbf{y}, t\mathbf{x} + \mathbf{y}\rangle.$$

Observe that $p(t) \geq 0$ for all t. By linearity of $\langle\cdot,\cdot\rangle$ in each variable we see that

$$p(t) = \|\mathbf{x}\|^2 t^2 + 2\langle \mathbf{x}, \mathbf{y}\rangle t + \|\mathbf{y}\|^2 = at^2 + bt + c,$$

where $a = \|\mathbf{x}\|^2$, $b = 2\langle \mathbf{x}, \mathbf{y}\rangle$, and $c = \|\mathbf{y}\|^2$. But the quadratic polynomial $p(t) \geq 0$ for all $t \in \mathbb{R}$ if and only if

$$b^2 - 4ac \leq 0,$$

which is equivalent to $b^2 \leq 4ac$. Hence

$$|\langle \mathbf{x}, \mathbf{y}\rangle|^2 \leq \|\mathbf{x}\|^2\|\mathbf{y}\|^2.$$

The first two conditions of Definition 2.4.4 are easily verified for $\|\cdot\|$. The third condition, the triangle inequality, is left for Exercise 8.1. ∎

Definition 8.1.2 *If a vector space V is equipped with a norm $\|\cdot\|$, we define the* open ball *of radius $r \geq 0$ about $\mathbf{p} \in V$ by*

$$B_r(\mathbf{p}) = \left\{\mathbf{v} \in V \,\middle|\, \|\mathbf{v} - \mathbf{p}\| < r\right\}.$$

Similarly, we define the closed ball *of radius $r \geq 0$ about $\mathbf{p} \in V$ by*

$$\bar{B}_r(\mathbf{p}) = \left\{\mathbf{v} \in V \,\middle|\, \|\mathbf{v} - \mathbf{p}\| \leq r\right\}.$$

The *bar* above the symbol B_r indicates that the set $\bar{B}_r$ is the *closed* ball, meaning that it includes the spherical boundary surface.

■ EXAMPLE 8.2

In the familiar Cartesian plane of Euclidean geometry, with the norm of a vector being its geometrical length, we have

$$\langle \mathbf{x}, \mathbf{y} \rangle = x_1 y_1 + x_2 y_2 = \|\mathbf{x}\|\|\mathbf{y}\| \cos\theta,$$

where θ is the angle between $\mathbf{x}$ and $\mathbf{y}$. Thus the Cauchy–Schwarz inequality in the plane follows from the fact that $|\cos\theta| \leq 1$ for all θ. (The Cauchy–Schwarz inequality follows alternatively from the argument given in Theorem 8.1.1.) In the plane of Euclidean geometry, $B_1(\mathbf{0})$ is the region strictly inside the circle of radius 1 centered at the origin. The reader should check that $B_0(\mathbf{p}) = \emptyset$, the empty set, for all points $\mathbf{p}$.

Definition 8.1.3 *If $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ are in $\mathbb{R}^n$, the set of all n-tuples of real numbers, we define the* Euclidean scalar product *of $\mathbf{x}$ and $\mathbf{y}$ as in Equation (8.1) and the* Euclidean norm *of $\mathbf{x}$ as in Equation (8.2). The* Euclidean space $\mathbb{E}^n$ *of n dimensions is defined to be the vector space $\mathbb{R}^n$ equipped with the Euclidean scalar product and the Euclidean norm.*

The reader should have checked that the Euclidean scalar product as defined above does satisfy the three properties required to be a scalar product, and that the Euclidean norm is in fact a norm.

We remark that it is very common when dealing with $\mathbb{E}^n$ to call it $\mathbb{R}^n$ informally. This means that we should suppose that the Euclidean inner product and norm are in use unless stated explicitly to the contrary. (See Exercise 8.4.)

Theorem 8.1.2 *The Euclidean normed vector space $\mathbb{E}^n$ has the following two properties.*

i. The Euclidean space $\mathbb{E}^n$ is complete in the sense of Definition 2.5.3.

ii. For a sequence of vectors $\mathbf{x}^{(j)}$ in $\mathbb{E}^n$, the sequence

$$\mathbf{x}^{(j)} \to \mathbf{x} \in \mathbb{E}^n \Leftrightarrow x_k^{(j)} \to x_k$$

as $j \to \infty$, *for each* $k = 1, 2, \ldots, n$.

Remark 8.1.1 The second part of the theorem tells us that convergence of a sequence in $\mathbb{E}^n$ is equivalent to convergence in each separate coordinate sequence.

Proof: The reader will recall that the vectors $\mathbf{x} \in \mathbb{R}^n$, which are the vectors of $\mathbb{E}^n$, comprise a vector space using the familiar operations under which

$$c\mathbf{x} = (cx_1, \ldots, cx_n)$$

and

$$\mathbf{x} + \mathbf{y} = (x_1 + y_1, \ldots, x_n + y_n).$$

We begin by proving claim (i). To see that every convergent sequence in any normed vector space is Cauchy in the sense of Definition 2.5.3, see Exercise 8.8. We will prove that every Cauchy sequence $\mathbf{x}^{(k)}$ in V converges. By hypothesis, For each $\epsilon > 0$ there exists K such that j and $k \geq K$ implies $\left\|\mathbf{x}^{(j)} - \mathbf{x}^{(k)}\right\| < \epsilon$. However, for each $l = 1, \ldots, n$, we have

$$\left|x_l^{(j)} - x_l^{(k)}\right| \leq \sqrt{\sum_{m=1}^{n} \left(x_m^{(j)} - x_m^{(k)}\right)^2} = \left\|\mathbf{x}^{(j)} - \mathbf{x}^{(k)}\right\|.$$

Thus the sequence $x_l^{(j)}$ is Cauchy in $\mathbb{R}$, and so it has a limit: $x_l^{(j)} \to x_l$ as $j \to \infty$, for each $l = 1, \ldots, n$. Now we denote $\mathbf{x} = (x_1, \ldots, x_n)$. Then we conclude that

$$\left\|\mathbf{x}^{(j)} - \mathbf{x}\right\| = \sqrt{\sum_{l=1}^{n} \left(x_l^{(j)} - x_l\right)^2} \to 0$$

as $j \to \infty$, by the theorems governing limits of sequences. Hence the original Cauchy sequence does converge in the sense of the norm of $\mathbb{E}^n$. Note that we have also proven that if each sequence of coordinates converges, then the sequence of vectors in $\mathbb{E}^n$ converges as well. We leave the second claim to Exercise 8.9.

■ EXAMPLE 8.3

Here are two examples of convergence and *divergence*, which means *failure of convergence*, for sequences of vectors.

i. Let $\mathbf{x}^{(j)} = \left(\frac{j^2+1}{2j^2-3}, j \sin\left(\frac{1}{j}\right)\right)$. Then $x^{(j)} \to \left(\frac{1}{2}, 1\right) \in \mathbb{E}^2$ as $j \to \infty$, because of the convergence of the sequence in each coordinate separately.

ii. In $\mathbb{E}^2$, the sequence $x^{(j)} = \left(\frac{1}{j}, j\right)$ diverges as $j \to \infty$.

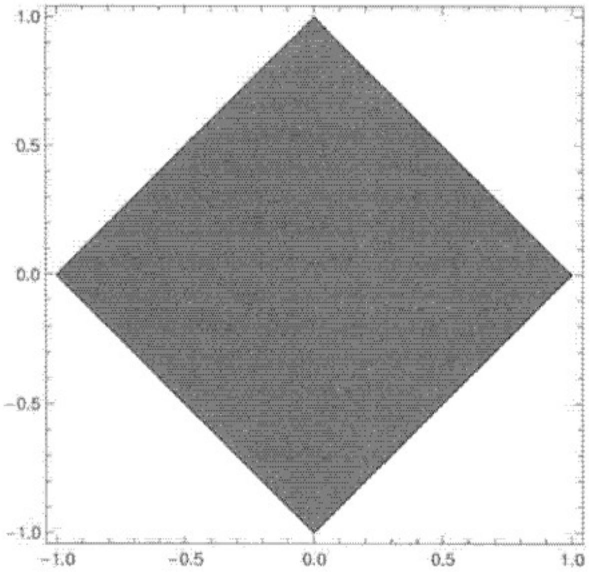

Figure 8.1 The unit ball $B_1(\mathbf{0})$ in the taxicab norm.

■ **EXAMPLE 8.4**

Here is an example of a *non-Euclidean* norm that can be placed on the vector space $\mathbb{R}^2$. Define the *taxicab* norm on $\mathbb{R}^2$ by

$$\|\mathbf{x}\|_t = |x_1| + |x_2|.$$

In Exercise 8.4 the reader will show that the taxicab norm satisfies the requirements of Definition 2.4.4 to be a norm. To understand the name *taxicab norm* consider a city in which all roads run parallel to the x-axis or the y-axis. The length of an honest taxicab ride would be determined using the taxicab norm. Fig. 8.1 shows the unit ball around the origin in the taxicab norm.

EXERCISES

8.1 † Complete the proof of Theorem 8.1.1 by proving the triangle inequality:

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$$

for all $\mathbf{x}$ and $\mathbf{y}$ in V. That is, prove that the norm defined in terms of a given scalar product does satisfy the triangle inequality. (Hint: Apply the Cauchy–Schwarz Inequality.)

8.2 In $\mathbb{E}^n$, $\|\mathbf{x} - \mathbf{y}\|$ is interpreted geometrically as being the distance between $\mathbf{x}$ and $\mathbf{y}$. Prove the *geometrical version* of the triangle inequality in $\mathbb{E}^n$:

$$\|\mathbf{x} - \mathbf{z}\| \leq \|\mathbf{x} - \mathbf{y}\| + \|\mathbf{y} - \mathbf{z}\|$$

for all $\mathbf{x}, \mathbf{y}$, and $\mathbf{z}$ in $\mathbb{E}^n$. Interpret this inequality in terms of a geometrical triangle in $\mathbb{E}^n$.

8.3 Two vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{E}^n$ are called *orthogonal* if and only if

$$\langle \mathbf{x}, \mathbf{y} \rangle = 0.$$

We will prove the Pythagorean Theorem in $\mathbb{E}^n$.

a) Prove: The vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal if and only if

$$\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 = \|\mathbf{x} + \mathbf{y}\|^2.$$

b) Interpret part (a) as a theorem about triangles in $\mathbb{E}^n$ with a vertex at the origin.

c) Let $\mathbf{x}, \mathbf{y}$, and $\mathbf{z}$ be in $\mathbb{E}^n$. Prove that the triangle with vertices at $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$ is a right triangle with right angle at $\mathbf{x}$ if and only if

$$\|\mathbf{x} - \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{z}\|^2 = \|\mathbf{y} - \mathbf{z}\|^2.$$

8.4 † Define the *taxicab norm* as in Example 8.4. Show that $\| \cdot \|_t$ satisfies the requirements of Definition 2.4.4 to be a norm.

8.5 Find necessary and sufficient conditions on the real numbers a_1 and a_2 to assure that

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{a}} = a_1 x_1 y_1 + a_2 x_2 y_2$$

satisfies Definition 8.1.1 and is thus a bona fide scalar product on the two-dimensional vector space $\mathbb{R}^2$. Then sketch the unit ball around the origin in $\mathbb{R}^2$ using the norm $\| \cdot \|_{\mathbf{a}}$ that is determined by this scalar product.

8.6 In $\mathbb{E}^1$, what is the open ball $B_1(\mathbf{0})$? Answer the same question in $\mathbb{E}^3$.

8.7 Find the $\lim_{j \to \infty} \mathbf{x}^{(j)}$ in $\mathbb{E}^2$, or state that it does not exist.

a) $\mathbf{x}^{(j)} = \left(\frac{1}{j}, \left(1 + \frac{1}{j}\right)^j \right)$.

b) $\mathbf{x}^{(j)} = \left(\frac{1}{j}, \frac{e^j}{j^{100}} \right)$.

8.8 † Prove that every convergent sequence $\mathbf{x}^{(k)}$ in a normed vector space V must be Cauchy, as claimed in Theorem 8.1.2.

8.9 † Show that $\mathbf{x}^{(j)} \to \mathbf{x}$ in the sense of the norm of $\mathbb{E}^n$ if and only if the sequence $x_l^{(j)} \to x_l$ for each $l = 1, \ldots, n$, completing the proof of Theorem 8.1.2.

8.10 Suppose $t_k \to t \in \mathbb{R}$ and $\mathbf{x}^{(k)} \to \mathbf{x} \in \mathbb{E}^n$ as $k \to \infty$. Prove: $t_k \mathbf{x}^{(k)} \to t\mathbf{x}$ as $k \to \infty$.

8.11 A sequence $\mathbf{x}^{(j)}$ in $\mathbb{E}^n$ is called *bounded* if there exists a number $M > 0$ such that $\|\mathbf{x}^{(j)}\| \leq M$ for all $j \in \mathbb{N}$.

a) Prove that every convergent sequence in $\mathbb{E}^n$ is bounded.

b) Give an example of a bounded sequence in $\mathbb{E}^2$ that is not convergent.

c) † *(The Bolzano–Weierstrass Theorem)* Prove that every bounded sequence in $\mathbb{E}^n$ has a convergent subsequence. (Hint: One approach is to give a proof by induction on the dimension n. Or one can work informally as follows, applying the Bolzano–Weierstrass Theorem for $\mathbb{R}$ a total of n times. Here is a notational suggestion: If one wishes to denote a subsequence of a subsequence, use two *strictly increasing* functions ϕ_1 and ϕ_2 mapping $\mathbb{N}$ into itself, and $\mathbf{x}^{(\phi_1 \circ \phi_2(j))}$ can be used as in Definition 1.5.1 to denote the jth term of a subsequence of a subsequence of a sequence $\mathbf{x}^j$.)

8.12 Prove: A sequence $\mathbf{x}_n \to \mathbf{L} \in \mathbb{E}^k$ if and only if *every* subsequence $\mathbf{x}_{n_i}$ possesses a sub-subsequence $\mathbf{x}_{n_{i_j}}$ that converges to $\mathbf{L}$ as $j \to \infty$. (Hint: To prove the *only if* part, suppose false and write out the logical negation of convergence of $\mathbf{x}_n$ to $\mathbf{L}$.)

8.13 *Prove or give a counterexample*: A sequence $\mathbf{x}_n \in \mathbb{E}^k$ converges if and only if *every* subsequence $\mathbf{x}_{n_i}$ possesses a sub-subsequence $\mathbf{x}_{n_{i_j}}$ that converges as $j \to \infty$.

8.14

a) Suppose a vector space V is equipped with an inner product $\langle \cdot, \cdot \rangle$, and suppose we define a corresponding norm by $\|\mathbf{x}\|^2 = \langle \mathbf{x}, \mathbf{x} \rangle$. Prove the *Parallelogram Law:*

$$\|\mathbf{x}+\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{y}\|^2 = 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2.$$

b) Prove that the taxicab norm defined in Exercise 8.4 does not correspond as in part (a) above to any inner product on $\mathbb{R}^2$.

c) Under the hypotheses of part (a) above, prove the identity

$$\langle \mathbf{x}, \mathbf{y} \rangle = \frac{1}{4}\left(\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2\right).$$

d) *(For this part, you might be able to negotiate extra credit from your teacher.)* Suppose only that V has a norm. Define what is *hoped* to be a scalar product on V by the formula in the preceding part. Prove that this defines a legitimate scalar product on V provided the norm satisfies the Parallelogram Law of part. (Hints: Positivity and symmetry are easy to verify. To show additivity of the product in the first variable, express $4\langle \mathbf{x}+\mathbf{y}, \mathbf{z} \rangle$ in terms of the identity given in the previous part. Then add and subtract $\|\mathbf{x}-\mathbf{y}+\mathbf{z}\|^2$ to show that

$$4\langle \mathbf{x}+\mathbf{y}, \mathbf{z} \rangle = 2(\|\mathbf{x}+\mathbf{z}\|^2 + \|\mathbf{y}\|^2 - \|\mathbf{x}\|^2 - \|\mathbf{y}-\mathbf{z}\|^2).$$

Then let $\mathbf{x}$ and $\mathbf{y}$ change roles and add the results. When this is done, show that $\langle \alpha\mathbf{x}, \mathbf{y} \rangle = \alpha\langle \mathbf{x}, \mathbf{y} \rangle$ for all $\alpha \in \mathbb{N}$. Extend this to $\alpha \in \mathbb{Q}$. Then justify that the inner product is a continuous function of $\mathbf{x}$ and extend to all $\alpha \in \mathbb{R}$.)

8.15 Prove that Euclidean space is *self-dual.* That is, prove that $T : \mathbb{E}^n \to \mathbb{R}$ is a bounded linear functional if and only if there exists $\mathbf{y} \in \mathbb{E}^n$ such that $T(\mathbf{x}) = \langle \mathbf{x}, \mathbf{y} \rangle$ for each $\mathbf{x} \in \mathbb{E}^n$.

8.2 OPEN SETS AND CLOSED SETS

In the calculus of one variable, it is possible to accomplish much with functions defined on an interval. Intervals are quite simple sets to consider as domains on which functions may be defined. Many useful theorems concerning the continuity, differentiability, or integrability of functions can be proven based on whether a domain is a closed or an open interval, or a finite or an infinite interval. In two or more variables, however, the domain of definition of a function can be much more complicated than an interval. For example, the real-valued function

$$f(x, y) = \frac{1}{x\sqrt{1 - 2x^2 - 3y^2}}$$

is defined on the domain that is the region lying *inside* but *not on* an ellipse, *excluding* those points inside the ellipse that are on the y-axis. It is easy to construct more intricate examples than this. Hence a rigorous study of the calculus of functions defined on a Euclidean space $\mathbb{E}^n$ necessitates a careful study of those *topological* properties of sets in $\mathbb{E}^n$ that are needed to prove important theorems regarding continuity, differentiability, and integrability of such functions. Most of the definitions and theorems work for all vector spaces V equipped with a norm, so we state them in this context.

Definition 8.2.1 *In a normed vector space V a subset $S \subseteq V$ is called an* open set *if and only if for each $\mathbf{x} \in S$ there exists a number $r > 0$ corresponding to $\mathbf{x}$ such that the open ball $B_r(\mathbf{x}) \subseteq S$.*

Expressed informally, a set S is said to be open if each point $\mathbf{p}$ of S has at least some small open ball, or *buffer zone* around itself that remains entirely inside S. The radius of the buffer zone may need to be smaller for some points of S than for others.

■ EXAMPLE 8.5

The first example is the open ball $B_r(\mathbf{x})$, which we claim is an open set. (The language *open ball* would be a poor choice if the claim just made were false, but a name alone does not establish that a definition is satisfied.)

Proof: To show that the definition of openness is satisfied, we must show that if $\mathbf{y} \in B_r(\mathbf{x})$ then there exists $d > 0$ such that $B_d(\mathbf{y}) \subseteq B_r(\mathbf{x})$. We claim that $d = r - \|\mathbf{y} - \mathbf{x}\| > 0$ works. In fact, if $\mathbf{z} \in B_d(\mathbf{y})$, then

$$\begin{aligned} \|\mathbf{z} - \mathbf{x}\| &\le \|\mathbf{z} - \mathbf{y}\| + \|\mathbf{y} - \mathbf{x}\| \\ &< d + \|\mathbf{y} - \mathbf{x}\| = r, \end{aligned}$$

proving that $\mathbf{z} \in B_r(\mathbf{x})$ as claimed. ■

Theorem 8.2.1 *Suppose A is an arbitrary set. Suppose for each $a \in A$ there is an* open *set $O_a \subseteq V$* indexed *by a. Then $\mathcal{O} = \bigcup_{a\in A} O_a$ is an open set.*

Remark 8.2.1 In this theorem, the *index set* A may be finite or infinite, countable or uncountable.

Proof: In words, this theorem says that the union of a family of open sets must always be open. Let $\mathbf{x} \in \mathcal{O}$. We must show there exists $r > 0$ such that $B_r(\mathbf{x}) \subseteq \mathcal{O}$. But there exists $a \in A$ such that $\mathbf{x} \in O_a$. Since O_a is open, there exists $r > 0$ such that $B_r(\mathbf{x}) \subseteq O_a \subseteq \mathcal{O}$. ■

It is natural at this point to ask whether the intersection of an arbitrary family of open sets must be open. This is false, as shown in Exercise 8.21.

Theorem 8.2.2 *If $O_1, \ldots, O_k$ is a family of finitely many open sets, then $\mathcal{O} = \bigcap_{j=1}^{k} O_j$ is open.*

Proof: In words, this says the intersection of finitely many open sets must be open. For each $\mathbf{x} \in \mathcal{O}$, $\mathbf{x} \in O_j$ for each j. Hence there exists for each j a number $r_j > 0$ such that $B_{r_j}(\mathbf{x}) \subseteq O_j$. Let $r = \inf\{r_1, \ldots, r_k\}$, so that $r > 0$. Then $B_r(\mathbf{x}) \subseteq \mathcal{O}$. ■

Definition 8.2.2 *In a normed vector space V, a set $S \subseteq V$ is called a* closed set *if and only if its complement $S^c = V \setminus S$ is open.*

We will use the following definition to prove an alternative and equivalent form of the concept of *closed* set.

Definition 8.2.3 *A point* $\mathbf{p}$ *is called a* cluster point *of a subset S of a normed vector space V if and only if there exists a sequence of vectors $\mathbf{x}^{(j)} \in S \setminus \{\mathbf{p}\}$ such that $\mathbf{x}^{(j)} \to \mathbf{p}$ as $j \to \infty$. If $\mathbf{p} \in S$ but $\mathbf{p}$ is* not *a cluster point of S, then $\mathbf{p}$ is called an* isolated point *of S.*

Note that a cluster point of S may or may not be an element of the set S. An isolated point of S must, however, be an element of S.

Theorem 8.2.3 *A subset S of a normed vector space V is closed if and only if every cluster point $\mathbf{p}$ of S is an element of S.*

Proof: Suppose first that S is closed. Let $\mathbf{p}$ be any cluster point of S. We need to prove that $\mathbf{p} \in S$. By hypothesis, there is a sequence $\mathbf{x}^{(j)} \in S \setminus \{\mathbf{p}\}$ such that $\mathbf{x}^{(j)} \to \mathbf{p}$ as $j \to \infty$. Thus, for every open ball $B_r(\mathbf{p})$ there exists $J \in \mathbb{N}$ such that $j \geq J \implies \mathbf{x}^{(j)} \in B_r(\mathbf{p})$. Thus $\mathbf{p} \notin S^c = V \setminus S$ since the latter set is open. Hence $\mathbf{p} \in S$.

Next, we suppose every cluster point of S is in S. We need to prove that the complement of S is open. Let $\mathbf{x} \in S^c = V \setminus S$. We claim there exists $r > 0$ such that $B_r(\mathbf{x}) \subseteq S^c = V \setminus S$. Suppose this claim were false. Then for each $j \in \mathbb{N}$

there exists $\mathbf{x}^{(j)} \in S \setminus \{\mathbf{x}\}$ such that $\mathbf{x}^{(j)} \in B_{\frac{1}{j}}(\mathbf{x})$. But then $\mathbf{x}^{(j)} \to \mathbf{x}$. Thus $\mathbf{x}$ is a cluster point of S, which forces $\mathbf{x} \in S$. This is a contradiction, which proves the claim. ∎

■ EXAMPLE 8.6

Every closed ball $\bar{B}_r(\mathbf{x})$ is a closed set.

Proof: We will show that the complement of the closed ball is open. If $\mathbf{y} \in V \setminus \bar{B}_r(\mathbf{x})$, let $\delta = \|\mathbf{y} - \mathbf{x}\| - r$. We claim that $B_\delta(\mathbf{y})$ is contained in the complement of the closed ball. So let $\mathbf{z} \in B_\delta(\mathbf{y})$. Then

$$\|\mathbf{z} - \mathbf{x}\| + \|\mathbf{z} - \mathbf{y}\| \geq \|\mathbf{x} - \mathbf{y}\|$$

by the triangle inequality, so

$$\begin{aligned} \|\mathbf{z} - \mathbf{x}\| &\geq \|\mathbf{x} - \mathbf{y}\| - \|\mathbf{z} - \mathbf{y}\| \\ &> \|\mathbf{x} - \mathbf{y}\| - \delta = r. \end{aligned}$$

Thus $\mathbf{z}$ lies outside the closed ball, as required. ∎

■ EXAMPLE 8.7

It follows from the preceding example that for each $\mathbf{p} \in \mathbb{E}^n$, the singleton set $\{\mathbf{p}\} = \bar{B}_0(\mathbf{p})$ is a closed set. The reader should show that this singleton set $\{\mathbf{p}\}$ is not an open set. Note that this does *not* follow from some linguistic relationship between the words *open* and *closed* in every day parlance. Instead, it is necessary to show that if $r > 0$ then $B_r(\mathbf{p})$ is not a subset of $\{\mathbf{p}\}$.

Theorem 8.2.4 *Suppose A is an arbitrary set. (A may be finite or infinite, countable or uncountable.) Suppose for each $a \in A$ there is a* closed *set $F_a \subseteq \mathbb{E}^n$ indexed by a. Then $\mathcal{F} = \bigcap_{a \in A} F_a$ is a closed set.*

Proof: It suffices to observe the

$$\mathcal{F} = \left(\bigcup_{a \in A} F_a^c \right)^c .$$

Since each set F_a^c is open being the complement of the closed set F_a, the union is open and its complement, $\mathcal{F}$, is closed. ∎

EXERCISES

8.16 Prove that Theorem 8.2.2 implies that the *empty set* $\emptyset$ is open in any normed vector space V, and that the whole space V is closed.

8.17 For each of the following sets, state whether it is open, closed, both, or neither.

- **a)** $A = \{\mathbf{x} \in \mathbb{E}^2 \,|\, x_1 \neq 0\}$.
- **b)** $B = \{\mathbf{x} \in \mathbb{E}^2 \,|\, x_1 = 0\}$.
- **c)** $C = \{\mathbf{x} \in \mathbb{E}^2 \,|\, x_1 \geq 0 \text{ and } x_2 > 0\}$.
- **d)** $E = \{\mathbf{x} \in \mathbb{E}^2 \,|\, x_1 + x_2 > 0\}$.
- **e)** $F = \{\mathbf{x} \in \mathbb{E}^2 \,|\, x_1 + x_2 \leq 0\}$.
- **f)** $D = A \cup B$.
- **g)** $G = \{\mathbf{x} \in \mathbb{E}^2 \,|\, 1 < x_1^2 + x_2^2 < 4\}$.

8.18 Prove that a point $\mathbf{p}$ is a cluster point of a set $D \subseteq \mathbb{E}^n$ if and only if for each $\delta > 0$ there exists $\mathbf{x} \in D$ such that $0 < \|\mathbf{x} - \mathbf{p}\| < \delta$.

8.19 Prove that if $\mathbf{p}$ is an isolated point of D, once $\delta > 0$ is sufficiently small the set $\{\mathbf{x} \in D \mid 0 < \|\mathbf{x} - \mathbf{p}\| < \delta\} = \emptyset$, the empty set.

8.20 Let $\mathcal{O}$ denote an *arbitrary open* subset of $\mathbb{E}^n$. Prove that $\mathcal{O}$ is the union of a (possibly infinite) family of open *balls*.

8.21 † Let $\mathbf{p} \in \mathbb{E}^n$. Show that the singleton set $\{\mathbf{p}\}$ can be written as the intersection of an infinite sequence of open balls, thereby showing that intersections of open sets need not be open. (Hint: Define a sequence $r_n > 0$ for which $\inf\{r_n \mid n \in \mathbb{N}\} = 0$.)

8.22 Prove or give a counterexample: No set $S \subseteq \mathbb{E}^n$ can be *both* open and closed.

8.23 Prove or give a counterexample: Every set $S \subseteq \mathbb{E}^n$ must be either open or closed.

8.24 Prove that if $F_1, \ldots, F_k$ is a family of finitely many closed sets, then $\mathcal{F} = \bigcup_{j=1}^k F_j$ is closed.

8.25 Use Example 8.7 to prove that every set $S \subseteq \mathbb{E}^n$ is the union of a family of closed sets.

8.26 † If $S \subset \mathbb{E}^n$, define the *interior* of S, denoted by S^o, to be

$$S^o = \{\mathbf{x} \in S \mid \exists\, r > 0 \text{ such that } B_r(\mathbf{x}) \subseteq S\}.$$

Prove that $S \subseteq \mathbb{E}^n$ is an open set if and only if $S = S^o$.

8.27 Prove that $\left(\bar{B}_r(\mathbf{x})\right)^o = B_r(\mathbf{x})$ in the space $\mathbb{E}^n$.

8.28 Define the *closure* $\bar{S}$ of $S \subseteq \mathbb{E}^n$ to be the intersection of all closed sets that contain S. Prove that $\bar{S}$ is a closed set and that $\bar{S} = S \cup C$, where C is the set of all cluster points of S.

8.29 Prove: The closed ball $\bar{B}_r(\mathbf{x})$, defined in Definition 8.1.2, is actually the closure of the open ball $B_r(\mathbf{x})$, provided $r > 0$.

8.30 Define the *boundary* of $S \subseteq \mathbb{E}^n$, denoted by ∂S, to be $\partial S = \bar{S} \setminus S^o$. Find the boundary of $B_r(\mathbf{x})$ in $\mathbb{E}^n$.

8.31 Define a set $S \subseteq \mathbb{E}^n$ to be dense in $\mathbb{E}^n$ if and only if $\bar{S} = \mathbb{E}^n$. Denote by $\mathbb{Q}^n$ the subset of $\mathbb{E}^n$ consisting of vectors with exclusively rational coordinates. Prove that $\mathbb{Q}^n$ is dense in $\mathbb{E}^n$.

8.32 Prove or give a counterexample: Every open set in $\mathbb{E}^n$ can be expressed as the union of either a finite or a countable family of open balls.

8.33 Find $(\mathbb{Q}^n)^o$ and $\partial\mathbb{Q}^n$. Justify your answers. (Hint: It will help to consider the density of the set of irrational numbers in the set $\mathbb{R}$.)

8.3 COMPACT SETS

In Chapters 1 through 3, we saw that closed, finite intervals $[a, b]$ have very useful properties as domains of continuous functions. Such intervals were very important for establishing an Extreme Value Theorem (Theorem 2.4.2) and a uniform continuity theorem (Theorem 2.3.2), for example. In this section we seek to describe the class of subsets of $\mathbb{E}^n$ that have similar properties as domains for functions of vector variables. We take our cue from the Heine–Borel Theorem (Theorem 1.7.2) for functions defined on intervals $[a, b]$.

Definition 8.3.1 *If $E \subseteq \mathbb{E}^n$, we call a family*

$$\mathcal{O} = \{O_a \mid a \in A\}$$

of open sets O_a an open cover *of E if and only if*

$$E \subseteq \bigcup_{a \in A} O_a.$$

If there exists a finite *subset F of the* index *set A such that*

$$E \subseteq \bigcup_{a \in F} O_a,$$

then we call $\mathcal{O}_F = \{O_a \mid a \in F\}$ a finite subcover *of E. We call the set E* compact *if and only if* every *open cover of E has a finite subcover. A set E is called* bounded *if and only if* $\sup \left\{ \|\mathbf{x}\| \,\middle|\, \mathbf{x} \in E \right\} < \infty$.

The reader should take care not to confuse the concept of an open cover with the *union* of its elements.

■ **EXAMPLE 8.8**

Let

$$S = \{(m, n) \mid m \in \mathbb{Z},\ n \in \mathbb{Z}\} = \mathbb{Z}^2 = \mathbb{Z} \times \mathbb{Z}.$$

We claim that the subset S of $\mathbb{E}^2$ is not compact. For a proof, we observe that

$$S \subset \bigcup_{(m,n) \in S} B_1((m, n)).$$

However, the open cover $\{B_1((m,n)) \mid m \in \mathbb{Z},\ n \in \mathbb{Z}\}$ of S has no finite subcover. Do you see why not?

■ EXAMPLE 8.9

The set $B_1(\mathbf{0})$ is not a compact subset of $\mathbb{E}^2$. For a proof, show that

$$\mathcal{O} = \{B_r(\mathbf{0}) \mid 0 < r < 1\}$$

is an open cover of $B_1(\mathbf{0})$, yet there is no finite subcover. The reader should show why this is true.

The following theorem provides a powerful generalization of the phenomena exhibited by the two preceding examples.

Theorem 8.3.1 (Heine–Borel Theorem for $\mathbb{E}^n$) *A set $E \subseteq \mathbb{E}^n$ is compact if and only if E is* both *closed and bounded.*

Proof: We leave the proof that every compact set must be both bounded and closed to Exercises 8.38 and 8.39, respectively. Here we suppose that E is both closed and bounded, and we will prove that E is necessarily compact. We suppose that this were *false*, and deduce a contradiction.

Thus we assume there exists an open cover $\mathcal{O} = \{O_a \mid a \in A\}$ of E for which there exists *no* finite subcover. Since E is bounded, there exists $R > 0$ such that $E \subseteq B_R(\mathbf{0})$, which is, in turn, contained in the *hyper*cube $[-R, R]^{\times n}$, an n-fold Cartesian product of the interval $[-R, R]$. It will be convenient in this proof to denote the hypercube by a pair of diagonally opposite corner-vectors $\mathbf{a}_1 = (-R, -R, \dots, -R)$ and $\mathbf{b}_1 = (R, R, \dots, R)$. Thus we will denote the cube

$$[-R, R]^{\times n} = [\mathbf{a}_1, \mathbf{b}_1].$$

Next we subdivide the cube $[\mathbf{a}_1, \mathbf{b}_1]$ into 2^n congruent subcubes as follows. Simply bisect the interval $[-R, R]$ on each of the n coordinate axes, and form all 2^n possible Cartesian products of a half-interval from each of the axes. Each such subcube can be denoted $[\mathbf{a}, \mathbf{b}]$, where the coordinates of $\mathbf{a} = (a_1, a_2, \dots, a_n)$ denote the left-hand endpoints of the n chosen subintervals on the axes, and with a similar convention is employed for $\mathbf{b}$ using right-hand endpoints. Observe that for each such subcube $[\mathbf{a}, \mathbf{b}] \subset [\mathbf{a}_1, \mathbf{b}_1]$, the set $E \cap [\mathbf{a}, \mathbf{b}]$ is covered by $\mathcal{O}$. Among the 2^n subcubes formed, there must exist at least one subcube $[\mathbf{a}_2, \mathbf{b}_2]$ having the property that $E \cap [\mathbf{a}_2, \mathbf{b}_2]$ has no finite subcover from $\mathcal{O}$. (Otherwise, there would exist a finite subcover for E itself.) Now we repeat the process by subdividing $[\mathbf{a}_2, \mathbf{b}_2]$ into 2^n subcubes and we select a subcube $[\mathbf{a}_3, \mathbf{b}_3]$ in the same manner as before. In this way, we obtain a decreasing nest

$$[\mathbf{a}_1, \mathbf{b}_1] \supset [\mathbf{a}_2, \mathbf{b}_2] \supset \dots \supset [\mathbf{a}_k, \mathbf{b}_k] \supset \dots$$

of subcubes having the property that for each $k \in \mathbb{N}$ the set $E_k = E \cap [\mathbf{a}_k, \mathbf{b}_k]$ is covered by $\mathcal{O}$ but has no finite subcover.

Select a point $\mathbf{p}_k \in E_k$ for each $k \in \mathbb{N}$. If j and $k \geq N$ then $\mathbf{p}_j$ and $\mathbf{p}_k \in E_N$. Hence

$$\|\mathbf{p}_j - \mathbf{p}_k\| \leq \frac{R\sqrt{n}}{2^{N-2}} \to 0$$

as $N \to \infty$. Thus the sequence $\mathbf{p}_k$ is a Cauchy sequence, and $\mathbf{p}_k \to \mathbf{p} \in E_N$ for each $N \in \mathbb{N}$. (Here we use the fact that E_N is a closed set.) Since $\mathbf{p} \in E$, there exists $a \in A$ such that $\mathbf{p} \in O_a$. Since O_a is open, there exists $r > 0$ such that $B_r(\mathbf{p}) \subseteq O_a$. Now select $N \in \mathbb{N}$ such that $k \geq N$ implies

$$\frac{R\sqrt{n}}{2^{k-2}} < r,$$

so that $\mathbf{p} \in E_N \subset B_r(\mathbf{p}) \subseteq O_a$. Thus we have covered E_N with a *single* set O_a from $\mathcal{O}$, contradicting the claim that E_N has no finite subcover from $\mathcal{O}$. ■

EXERCISES

8.34 If E is a compact subset of $\mathbb{E}^n$, prove (*without using the Heine–Borel Theorem*) that every closed subset of E is compact.

8.35 Use *only* the definition of compactness to show that every finite subset of $\mathbb{E}^n$ is compact.

8.36

- **a)** Let $S = \{\mathbf{x}^{(j)} \mid j \in \mathbb{N}\} \subset \mathbb{E}^n$ be any *convergent* sequence. Show that S is a compact set if and only if $\lim_{j\to\infty} \mathbf{x}^{(j)} \in S$.
- **b)** Let $S = \{\mathbf{x}_n = (1 + \frac{n}{2^n}, \frac{n}{2^n}) \mid n = 0, 1, 2, 3, \ldots\} \subset \mathbb{E}^2$. True or False: S is compact.

8.37 Let $S = \{\mathbf{x}^{(j)} \mid j \in \mathbb{N}\} \subset \mathbb{E}^n$ be *any* sequence. Prove or give a counterexample: The set S is a compact set if and only if the sequence $\mathbf{x}^{(j)}$ is convergent to an element of S.

8.38 † Prove part of the Heine–Borel theorem by showing that if $E \subseteq \mathbb{E}^n$ is *compact*, then E is *bounded*. (Hint: Show how to cover E with suitable open balls centered at $\mathbf{0}$.)

8.39 † Prove part of the Heine–Borel theorem by showing that if $E \subseteq \mathbb{E}^n$ is *compact*, then E is *closed*. (Hint: Suppose that $\mathbf{p}$ is any point not in E. Show how to cover E with the complements of suitable closed balls centered at $\mathbf{p}$. Conclude that E^c is *open*.)

8.40 † Let f be a real-valued function defined on a closed finite interval $[a, b]$ in $\mathbb{E}^1$. Define the *graph* $G_f = \{(x_1, x_2) \in \mathbb{E}^2 \mid x_2 = f(x_1), x_1 \in [a, b]\}$.

- **a)** Prove: If $f \in \mathcal{C}[a, b]$, then G_f is a compact subset of $\mathbb{E}^2$. (Hint: Use the Heine–Borel Theorem.)

b) Let

$$g(x_1) = \begin{cases} \sin \frac{\pi}{x_1} & \text{if } x_1 \in (0,1], \\ 0 & \text{if } x_1 = 0. \end{cases}$$

Is the graph G_g a compact subset of $\mathbb{E}^2$? Prove your conclusion.

8.41 Let $E_1 \supseteq E_2 \supseteq \dots \supseteq E_k \supseteq \dots$ be a decreasing nest of *nonempty closed* subsets of $\mathbb{E}^n$.

a) Give an example to show it is possible for $\bigcap_{k=1}^{\infty} E_k$ to be empty.

b) If E_1 is compact, show that $\bigcap_{k=1}^{\infty} E_k \neq \emptyset$. (Hint: Select a point $\mathbf{x}_k \in E_k$ for each $k \in \mathbb{N}$. Apply Exercise 8.11.c.)

8.42 For each of the following subsets of $\mathbb{E}^n$, determine whether or not it is compact and justify your conclusion.

a) $B_r(\mathbf{x})$, with $r > 0$.

b) $\bar{B}_r(\mathbf{x})$, with $r > 0$.

c) $S^{n-1} = \bar{B}_r(\mathbf{x}) \setminus B_r(\mathbf{x})$, with $r > 0$.

8.43 Let $K \subset \mathbb{E}^n$ be compact. Suppose f and g are in $\mathcal{C}(K)$ and suppose that $f(\mathbf{x}) = 0$ for each $\mathbf{x} \in K$ such that $g(\mathbf{x}) = 0$. Prove: If $\epsilon > 0$, there exists $M \geq 0$ such that

$$|f(\mathbf{x})| < M|g(\mathbf{x})| + \epsilon$$

for all $\mathbf{x} \in K$. (Hint: Use the Heine–Borel Theorem. For interesting applications of this exercise in approximation theory, see [13].)

8.4 CONNECTED SETS

In order to generalize the Intermediate Value Theorem (Theorem 2.3.1) to functions defined on Euclidean space, it will be necessary to identify a class of subsets of $\mathbb{E}^n$ with properties sufficiently similar to those of intervals. Let us begin by writing a formal definition of the concept of *interval*.

Definition 8.4.1 *An* interval *is any subset I of $\mathbb{R}$ such that for all a and b in I with $a \leq b$ the set $\{x \in \mathbb{R} \mid a \leq x \leq b\} \subseteq I$.*

This concept includes all open, closed, half-open and half-closed, finite or infinite intervals. (See Exercise 8.44.) The difficulty in generalizing this concept is that there is no natural *linear ordering*, or notion of *inequality* in $\mathbb{E}^n$ when $n > 1$. The concept we seek is that of being a *connected* set. In order to define this concept it is convenient to begin with its opposite: a set is called *disconnected*, in the sense that it can be *separated* into two parts, or components.

Definition 8.4.2 *We say that a set $E \subseteq V$, a normed vector space, is* disconnected *if there exists a pair of* disjoint open *subsets A and B of V for which $E \subseteq A \cup B$ and such that both*

$$E_1 = E \cap A \neq \emptyset \text{ and } E_2 = E \cap B \neq \emptyset.$$

In this case, we say that A and B separate *E. If E is* not *disconnected, then E is called* connected.

■ EXAMPLE 8.10

Let $E = \mathbb{Q}$, the set of all rational numbers. We claim that E is a *disconnected* subset of $\mathbb{E}^1$. In fact, let

$$A = \left\{x \in \mathbb{E}^1 \,\middle|\, x < \sqrt{2}\right\} \text{ and } B = \left\{x \in \mathbb{E}^1 \,\middle|\, x > \sqrt{2}\right\}.$$

Then A and B are disjoint open sets that separate E. The key to this construction is that there exists $\sqrt{2} \in \mathbb{E}^1 \setminus \mathbb{Q}$.

In Exercise 8.45 the reader will prove that a subset of $\mathbb{E}^1$ is connected if and only if it is an interval. The set $\mathbb{Q}$, which is shown above to be disconnected, is not an interval. We show below that the concept of a set E being disconnected can be expressed without reference to open sets A and B as in the definition. This will give us a more visually intuitive concept of the meaning of connectivity.

Theorem 8.4.1 *A set $E \subseteq V$, a normed vector space, is disconnected if and only if we can decompose E as $E = E_1 \cup E_2$, both sets nonempty, where $E_1 \cap E_2 = \emptyset$ and where neither E_1 nor E_2 contains any cluster points of the other set.*

Proof: First we suppose that E is separated. Thus A and B exist as in the definition. If $\mathbf{e} \in E_1$, then there exists $r > 0$ such that $B_r(\mathbf{e}) \subseteq A$. Since E_2 is disjoint from A, $\mathbf{e}$ is not a cluster point of E_2. Similarly, E_2 has no cluster point of E_1.

Next we suppose that $E = E_1 \cup E_2$, a disjoint union, where neither E_1 nor E_2 has any cluster point of the other, and each set is nonempty. Let $\mathbf{e} \in E$. Since $\mathbf{e}$ is in one of the two sets E_1 or E_2 without being a cluster point of the other, there exists $r_{\mathbf{e}} > 0$ such that $B_{r_{\mathbf{e}}}(\mathbf{e})$ is disjoint from the set to which $\mathbf{e}$ does not belong. Let

$$A = \bigcup_{\mathbf{e} \in E_1} B_{\frac{r_{\mathbf{e}}}{2}}(\mathbf{e}).$$

Similarly, we let

$$B = \bigcup_{\mathbf{e} \in E_2} B_{\frac{r_{\mathbf{e}}}{2}}(\mathbf{e}).$$

It follows that A and B are open sets and that $A \cap E = E_1$ and $B \cap E = E_2$. We need prove only that $A \cap B = \emptyset$. Suppose this claim were false. Then there exists $\mathbf{p} \in A \cap B$. Hence there exist $\mathbf{e}_1 \in E_1$ and $\mathbf{e}_2 \in E_2$ such that $\|\mathbf{e}_1 - \mathbf{p}\| < \frac{r_{\mathbf{e}1}}{2}$ and $\|\mathbf{e}_2 - \mathbf{p}\| < \frac{r_{\mathbf{e}2}}{2}$. By the triangle inequality for norms, it follows that

$$\|\mathbf{e}_1 - \mathbf{e}_2\| < \frac{r_{\mathbf{e}_1} + r_{\mathbf{e}_2}}{2} \leq \max\{r_{\mathbf{e}_1}, r_{\mathbf{e}_2}\}.$$

This is a contradiction. ■

Theorem 8.4.2 *Let f be any continuous real-valued function defined on an interval I. Let G_f denote the graph of f as defined in Exercise 8.40. We claim that G_f is a connected subset of $\mathbb{E}^2$.*

Proof: We suppose the claim were false, and we will deduce a contradiction. If G_f were disconnected, then $G_f = E_1 \cup E_2$, a disjoint union of two nonempty sets, neither one containing any cluster point of the other. So there exist $(x_i, f(x_i)) \in E_i$, for $i = 1, 2$. Suppose without loss of generality that $x_1 < x_2$. (We know $x_1 \neq x_2$, since G_f is the graph of a function.) Define

$$c = \sup \left\{ t \in [x_1, x_2] \,\middle|\, (q, f(q)) \in E_1, \forall q \in [x_1, t] \right\}.$$

Thus $x_1 \leq c \leq x_2$. If $(c, f(c)) \in E_1$ then $c < x_2$. Hence for $n \in \mathbb{N}$ there exists

$$c_n \in \left(c, c + \frac{1}{n} \right) \cap [x_1, x_2]$$

for which $(c_n, f(c_n)) \in E_2$. This would imply that $(c_n, f(c_n)) \to (c, f(c))$, by continuity of f. Hence $(c, f(c))$ is a cluster point of E_2, which is a contradiction. On the other hand, if $(c, f(c)) \in E_2$, then $c > x_1$, $(c, f(c))$ is a cluster point of E_1. This yields a similar contradiction. ■

Corollary 8.4.1 *The Euclidean space $\mathbb{E}^n$ is connected, for each $n \in \mathbb{N}$.*

Proof: Suppose the corollary were false. Then $\mathbb{E}^n = E_1 \cup E_2$, a disjoint union of two sets, neither of which contains any of the other's cluster points. Let $\mathbf{p} \in E_1$ and $\mathbf{q} \in E_2$. Let

$$\phi(t) = \mathbf{p} + t(\mathbf{q} - \mathbf{p})$$

for each $t \in [0, 1]$. Then $\phi(0) \in E_1$ and $\phi(1) \in E_2$. Define

$$c = \sup \left\{ t \in [0, 1] \,\middle|\, (q, \phi(q)) \in E_1, \forall q \in [0, t] \right\}.$$

Now complete the proof just as in Theorem 8.4.2. ■

Theorem 8.4.3 *Suppose E and F are* nondisjoint *connected subsets of a normed vector space V. Then $G = E \cup F$ is connected.*

Proof: Suppose false. Then there exist nonempty disjoint open sets A and B such that $A \cap G \neq \emptyset \neq B \cap G$. Since E cannot be separated, E is contained entirely in one open set or the other. Without loss of generality, suppose $E \subseteq A$. Similarly, F is contained entirely in one of the open sets. Since $B \cap G \neq \emptyset$, $F \subseteq B$. But this is impossible since $E \cap F \neq \emptyset$ and $A \cap B = \emptyset$. ■

EXERCISES

8.44 † Prove that if $I \subseteq \mathbb{R}$ is an interval, as defined in Definition 8.4.1, then I must have one of the following forms: $[a, b], (a, b), [a, b)$ or $(a, b]$. In this notation $-\infty \leq a \leq b \leq \infty$, but closed endpoints must be finite. (Hint: Consider $\sup(I)$ and $\inf(I)$.)

8.45 †

a) Prove that every connected subset $S \subseteq \mathbb{E}^1$ is an interval.

b) Prove that every interval I is a connected subset of $\mathbb{E}^1$. In particular, this includes the claim that the real line $\mathbb{E}^1$ is itself a connected set. (Hint: Consider Theorem 8.4.2 and the function $f(x) \equiv 0$. Explain how connectedness in E^2 implies connectedness in E^1.)

8.46 Prove that the union of the x_1-axis and the x_2-axis in $\mathbb{E}^2$ is a connected set.

8.47 Prove that $S^1 = \{\mathbf{x} \in \mathbb{E}^2 \mid \|\mathbf{x}\| = 1\}$ is a connected set.

8.48 Suppose $E \subseteq \mathbb{E}^2$ such that for all $\mathbf{x} \in E$, $x_2 = 0$. Suppose E is a connected subset of $\mathbb{E}^1$, which we identify with the x_1-axis of $\mathbb{E}^2$. Prove that E is also connected as a subset of $\mathbb{E}^2$.

8.49 Let $E = \{\mathbf{x} \in \mathbb{E}^2 \mid x_1x_2 > 0\}$. Is E connected? Prove your conclusion.

8.50 Let $E = \{\mathbf{x} \in \mathbb{E}^2 \mid x_1x_2 \geq 0\}$. Is E connected? Prove your conclusion.

8.51 Let $E = \{\mathbf{x} \in \mathbb{E}^2 \mid x_1^2 = 1\}$. Is E connected? Prove your conclusion.

8.52 In E^2, let $E = \{\mathbf{x} \mid x_1 = 0\} \cup \{\mathbf{x} \mid x_1x_2 = 1, x_1 > 0\}$. Is E connected? Prove your conclusion.

8.53 Let

$$g(x) = \begin{cases} x \sin \frac{\pi}{x} & \text{if } x > 0, \\ 0 & \text{if } x = 0. \end{cases}$$

Prove that the graph of g is a connected subset of $\mathbb{E}^2$.

8.54 Let

$$f(x) = \begin{cases} \sin \frac{\pi}{x} & \text{if } x > 0, \\ 0 & \text{if } x = 0. \end{cases}$$

Prove that the graph of f is a connected subset of $\mathbb{E}^2$. (Hint: Try writing $G_f \subseteq A \cup B$ with A and B open and disjoint. Prove that G_f must be contained entirely within one of the two open sets.)

8.55 Let $E \subseteq \mathbb{E}^n$ be any connected set and let $\mathbf{p}$ be a cluster point of E. *Prove or disprove*: $E \cup \{\mathbf{p}\}$ is connected.

8.56 Let $E \subseteq \mathbb{E}^n$ be any connected set and let $\bar{E}$ be the *closure* of E. *Prove or disprove*: $\bar{E}$ is connected.

8.57 Let $f(x) = e^{-x^2}$ for all $x \in \mathbb{R}$. Denote the graph of f by G_f and the x-axis in $\mathbb{E}^2$ by G_0. *Prove or disprove*: $S = G_f \cup G_0$ is a connected subset of $\mathbb{E}^2$. (Hint: Apply Theorem 8.4.1.)

8.58 Prove or give a counterexample: Every open set $\mathcal{O} \subseteq \mathbb{E}^n$ can be expressed as the union of a finite or countable family of *disjoint* open balls. (Hint: See Exercise 8.32.)

8.5 TEST YOURSELF

EXERCISES

8.59 Let $\mathbf{x} = (5, -2, 3) \in \mathbb{E}^3$. Give an example of a non-$\mathbf{0}$ vector $\mathbf{y} \in \mathbb{E}^3$ for which $\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 = \|\mathbf{x} + \mathbf{y}\|^2$.

8.60 Let $S \subseteq \mathbb{E}^n$ be a set of vectors for which each sequence

$$\left\{\mathbf{x}^{(k)} \mid k \in \mathbb{N}\right\} \subseteq S$$

has a convergent *sub*sequence. True or False: S must be bounded.

8.61 The *taxicab* norm in $\mathbb{R}^2$ is defined by $\|\mathbf{x}\|_t = |x_1| + |x_2|$. Draw a sketch of the unit ball $B_1(\mathbf{0})$ in $\mathbb{R}^2$ using the taxicab norm. Label your axes.

8.62 Give an example of a set $D \subseteq \mathbb{E}^2$ and a point $\mathbf{p}$ which is an *isolated* point of D.

8.63 Give an example of a set $S \subseteq \mathbb{E}^2$ for which S is dense $\mathbb{E}^2$ but has empty interior.

8.64 True or False: The interior of the closure of an *open* set $S \subseteq \mathbb{E}^n$ need *not* be S itself.

8.65 Let $f(x) = \sin \frac{\pi}{x}$ for all $x \neq 0$. True or False: The graph G_f is compact.

8.66 Let $E_1 \supseteq E_2 \supseteq \ldots \supseteq E_k \supseteq \ldots$ be a decreasing nest of closed subsets of $\mathbb{E}^n$. True or False: $\bigcap_{k=1}^{\infty} E_k$ must be non-$\emptyset$.

8.67 Let $S = \left\{\mathbf{x}_n = \left(1 + \frac{n}{2^n}, \frac{n}{2^n}\right) \mid n = 0, 1, 2, 3, \ldots\right\} \subset \mathbb{E}^2$. True or False: S is compact.

8.68 Let

$$g(x) = \begin{cases} x^2 \cos \frac{1}{x} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

True or False: The graph G_g is connected.

8.69 For each one of the following two sets, decide whether it is *connected* or *not connected*.

- **a)** $S = \left\{\mathbf{x} \in \mathbb{E}^2 \mid |x_2| = |x_1|\right\}$.
- **b)** $T = \left\{\mathbf{x} \in \mathbb{E}^2 \mid |x_2| < |x_1|\right\}$.

8.70 True or False: If $S \neq \emptyset$ is open in $\mathbb{E}^n$, then $S = \cup_{a \in A} B_a$, a union of nonempty *disjoint* open balls of finite radius.

()*+(((, (. !' $ --C /"
& $$ $!! * % = " /

" ()*+(((, (. 1 < ; < <1 /"
& '!

The Five "Reflecting Red" Types

The all-red signatures for sphere patterns must have the form `*AB...N` since we can no longer afford two $*$'s. The ones for which *ch* is positive are in perfect correspondence with the true blue types, since $ch(*\texttt{AB...N})$ is exactly half of $ch(\mathbf{AB...N})$, as we see from the following:

$$ch(*\texttt{AB...N}) = {}^{\$}2 - 1 - \left(\frac{A-1}{2A} + \cdots + \frac{N-1}{2N}\right),$$

$$ch(\mathbf{AB...N}) = {}^{\$}2 - \left(\frac{A-1}{A} + \cdots + \frac{N-1}{N}\right).$$

But remember the proviso: `*MN` exists only if $M = N$.

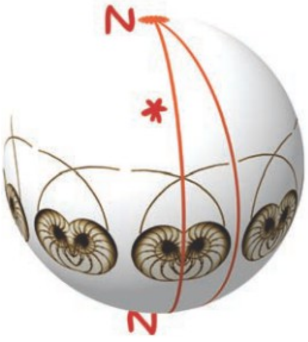

Signature *NN

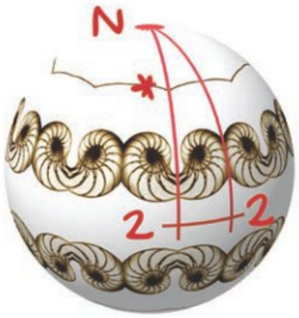

Signature *22N

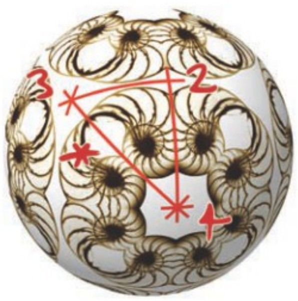

Signature *432

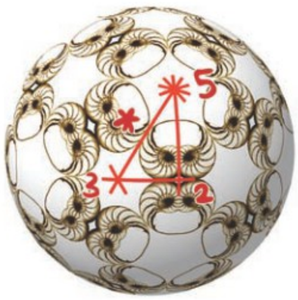

Signature *532

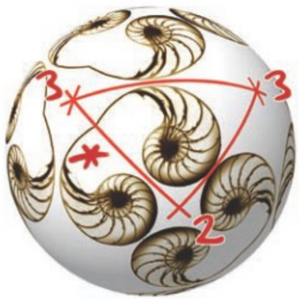

Signature *332

The Four Hybrid Types

As in the plane case, these must all be obtainable by promotion from the red reflective cases. Here are all the possibilities:

$$\begin{tikzcd} {*532} & {*432} & {*332} \arrow[d] & {*22N} \arrow[d] & {*NN} \arrow[d] \\ & & 3{*}2 & 2{*}N & N{*} \arrow[d] \\ & & & & N\times \end{tikzcd}$$

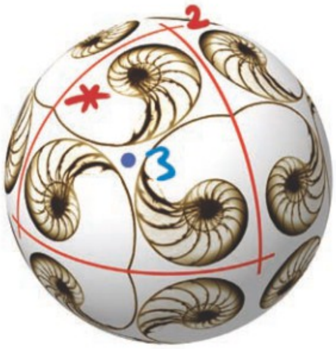

Signature **3**∗2

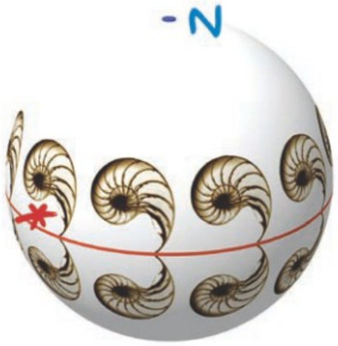

Signature **N**∗

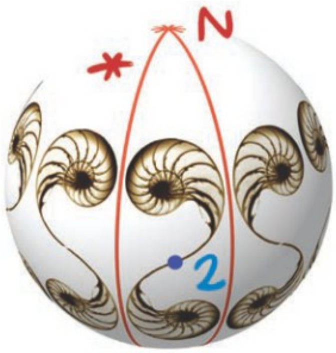

Signature **2**∗N

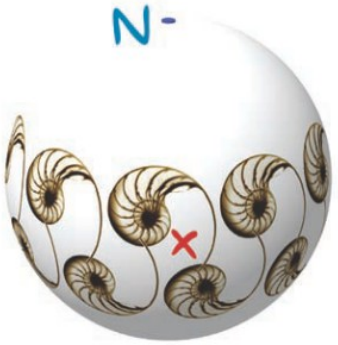

Signature **N**×

The Existence Problem: Proving the Proviso

All 17 possibilities that we enumerated for plane patterns actually arose. In the spherical case, the corresponding statement is not quite true; the types **MN** and *MN only exist if $M = N$. The other cases cause no problem.

For example, *442 was generated by reflections in a triangle of angles $\frac{\pi}{4}$, $\frac{\pi}{4}$, $\frac{\pi}{2}$, and a plane pattern with this type of symmetry exists because such a triangle exists in the Euclidean plane. Similarly, *532 is generated by reflections in a triangle of angles $\frac{\pi}{5}$, $\frac{\pi}{3}$, $\frac{\pi}{2}$, and a spherical pattern with this symmetry exists because there is a spherical triangle with these angles.[1]

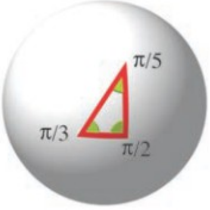

A triangle on the sphere with angles $\frac{\pi}{5}$, $\frac{\pi}{3}$, and $\frac{\pi}{2}$.

Now for the proviso! The type *MN, when it exists, is generated by the reflections in the sides of a two-sided polygon with angles $\frac{\pi}{M}$ and $\frac{\pi}{N}$. This *does* exist when $M = N$; it's the lune bounded by two great semicircles at angle $\frac{\pi}{N}$ (at right), but does not when $M \neq N$. (For the same reason *M, which equals *M1, fails to exist for $M > 1$.)

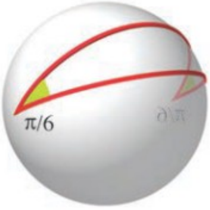

A two-sided spherical polygon with two angles of $\frac{\pi}{6}$ radians.

A hypothetical pattern of type **MN** with $M \neq N$ would contain just two types of gyration point. But then, by superposing it with its image under a reflection fixing a gyration point of each type, we should obtain one of type *MN, which is impossible. Therefore, **MN** also fails to exist if $M \neq N$, and **M** fails to exist if $M \neq 1$.

Group Theory and All the Spherical Symmetry Types

Group theory is not discussed in detail here, at least not until much later in this book; there are many texts available that teach group theory better than we are able to in the space available here. In brief, the symmetries of a pattern form a group; suppose A and B are symmetries of a pattern, described by some motion of the pattern that takes a fundamental region to a copy of itself. Then, the "product" symmetry AB is what you get when the motion associated with A is followed by the motion associated with B. You may wish

---

[1] In the plane, the sum of the angles of a triangle is always π radians $= 180^\circ$. On a sphere, the sum can be larger. For example, a triangle on the globe with one vertex on the North Pole and two vertices on the equator has two angles of $\frac{\pi}{2}$ radians at the equator. The sum of the angles of that triangle will be π plus the angle at the pole.

Order	NN	N×	N*	2*N	22N	*NN	*22N			Number of Groups
1	11									1
2	22	×	*		22	*				3(1)
3	33									1
4	44	2×	2*	2*	222	*22	*22			5(2)
5	55									1
6	66	3×	3*		223	*33				5(2)
7	77									1
8	88	4×	4*	2*2	224	*44	*222			7(4)
9	99									1
10	10 10	5×	5*		225	*55				5(2)
11	11 11									1
12	12 12	6×	6*	2*3	226	*66	*223	332		8(4)
13	13 13									1
14	14 14	7×	7*		227	*77				5(2)
15	15 15									1
16	16 16	8×	8*	2*4	228	*88	*224			7(4)
17	17 17									1
18	18 18	9×	9*		229	*99				5(2)
19	19 19									1
20	20 20	10×	10*	2*5	22 10	*10 10	*225			7(3)
21	21 21									1
22	22 22	11×	11*		22 11	*11 11				5(2)
23	23 23									1
24	24 24	12×	12*	2*6	22 12	*12 12	*226	*332 432	3*2	10(6)
27	27 27									1
30	30 30	15×	15*		22 15	*15 15				5(2)
33	33 33									1
36	36 36	18×	18*	2*9	22 18	*18 18	*229			7(3)
39	39 39									1
42	42 42	21×	21*		22 21	*21 21				5(2)
45	45 45									1
48	48 48	24×	24*	2*12	22 24	*24 24	*22 12	*432		8(5)
51	51 51									1
54	54 54	27×	27*		22 27	*27 27				5(2)
57	57 57									1
60	60 60	30×	30*	2*15	22 30	*30 30	*22 15		532	8(4)
63	63 63									1
66	66 66	33×	33*		22 33	*33 33				5(2)
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮			⋮
120	120 120	60×	60*	2*30	22 60	*60 60	*22 30	*532		8(5)
	C		2 × *C*	*D*			2 × *D*	*P*	2 × *P*	

Table 4.1. Types of spherical repeating patterns and their equivalences as groups.

		for $n = 3, 4, 5$
C	cyclic	$\langle a \mid a^n = 1\rangle$
D	dihedral	$\langle a, b \mid a^n = b^2 = (ab)^2 = 1\rangle$
P	polyhedral	$\langle a, b, c \mid a^n = b^3 = c^2 = abc = 1\rangle$
$2 \times G$	direct product of G with a group of order 2.	

Table 4.2. Legend for Table 4.1.

to confirm for yourself that in the examples we have seen so far this composition is associative (i.e., that (AB)C = A(BC) for all choices of A, B, and C) and that in the relation AB = C, any two of A, B, and C uniquely determine the third.

The size or *order* of a group is the number of elements in it—here it is the number of symmetries. Two different geometrical groups have the same abstract structure if their elements multiply in the same way.

All the Spherical Types

Table 4.1 lists the types of spherical repeating patterns and their equivalences as groups. The table is complete for patterns with up to 24 symmetries, after which we restrict to multiples of 3. The last column gives the number of distinct geometrical groups, followed in parentheses by the number of different abstract structures (as separated by the lines). Each row of Table 4.1 lists all the groups of a given size; groups with different abstract structures are separated by the curved and straight lines. Codes for these structures are given below the table and are explained in Table 4.2: the polyhedral groups are isomorphic to the alternating and symmetric groups A_4, S_4, or A_5 according as n is 3, 4, or 5.

Where Are We?

We have shown in this chapter that the spherical form of the Magic Theorem implies that the spherical symmetry types fall into seven infinite families plus seven individual types. The groups of symmetries of these patterns are listed by increasing number of symmetries in Table 4.1. As before, the proof of the Magic Theorem that leads to this conclusion is postponed to Chapter 6.

data 21.10.2019

c) Advanced Calculus - An Introduction to Linear Analysis...pdf

Source: 20191021/ Advanced Calculus An Introduction to Linear Analysis/.

Pag. 180 (206)

6.1 The Vibrating String and Trigonometric Series.

d) The Symmetries of Things

Source: 20191021/ The Symmetries of Things/ Conwt.M - The_Symmetries_of_Things.pdf

Pag. 57 (76)

The Existence Problem: Proving the Proviso.

Figures (2)

1) Triangle on the sphere with angles $\frac{\pi}{5}$, $\frac{\pi}{3}$, and $\frac{\pi}{2}$

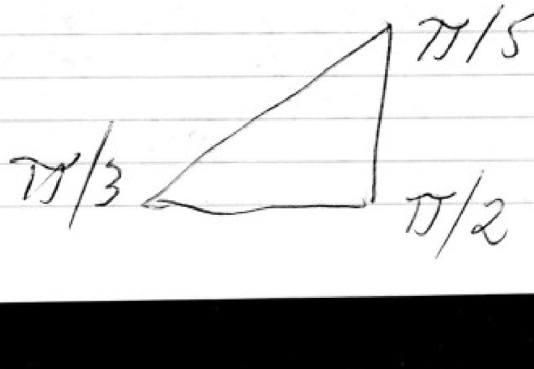

21.10.2019

2) A two-sided spherical polygon with two angles of $\frac{\pi}{6}$ radians

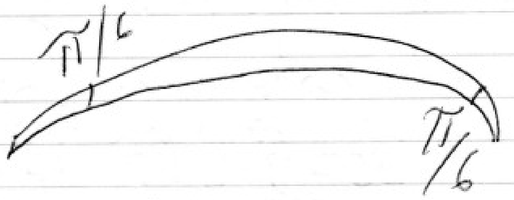

e) 20191022 / Cosmic DNA at the Origin.... /

Chris H. Hardy - Cosmic DNA at the Origin.... epub.

Pág. 60/339

Dirac Sea: the vacuum as a sea of negative energy

The British physicist Paul Dirac brought a new understanding of the vacuum as an infinite energy, as we conceive it now.

He predicted the existence of anti-matter back in 1928, with his famous Dirac Equation, that was the first bridge between quantum physics and relativity theory.

—

107) 20200125

()*+(((, (. 9 9 "; @4)
,) 9;/ " 0 > > < % / 1 ; /
% " / " 0 / 0

&

@ A A ;
1 > / ; @ 0 " " A ;?
A 9 / " " 3 0 D C -? @
=I ? 0 " 0 I / ; " A ; ; H ?
/ " " / ? / A ; "? 0 "
/ 3/ ; A " " 0;) " C$ > @
/ ? ; / ; ?
/ // ; H D / " @ < J I
/ / A ; = ? / " ; = K

e) 20191022 / Cosmic DNA at the Origin..../

Chris H. Hardy – Cosmic DNA at the Origin....epub.

Pág. 60/339

Dirac Sea: the vacuum as a sea of negative energy

The British physicist Paul Dirac brought a new understanding of the vacuum as an infinite energy, as we conceive it now.

He predicted the existence of anti-matter back in 1928, with his famous Dirac Equation, that was the first bridge between quantum physics and relativity theory.

25.01.2020

DATA_NOVOVIVO_20_2020/images

I) Screenshot... objeto giratório mais rápido do mundo.png

Use image: o rotor gira impulsionado pela força da luz
/1579795322979_v2-450x450.jpeg

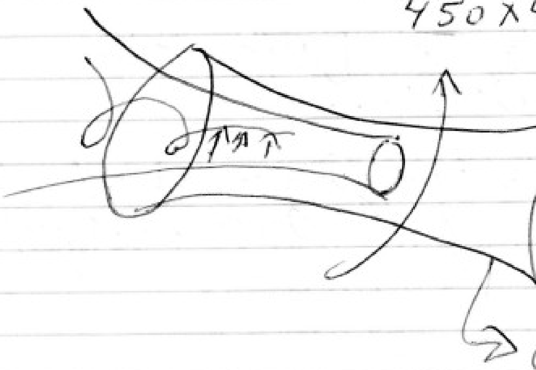

→ curvature (space)
Research!!!

Copy files:

II) NOVOVIVO_20_2020/Important/
Patricia M. Schwarz...
Special Relativity - From Einstein to String 2004.pdf

Pág. 13 (26)

A very small rotation with
$\theta \sim 0$ can be written:

(1.30)

25 01 2020

$$\tilde{x} \simeq x + \theta y + O(\theta^2)$$
$$\tilde{y} \simeq y - \theta x + O(\theta^2)$$

In this infinitesimal limit, the rotation matrix $R(\theta')$ can be written in terms of a matrix r

$$R(\theta) \simeq I + \theta r + \dots,$$
$$r = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \qquad (1.31)$$

Obs: Using of rotations preserve the Euclidean metric with Gyrocommutative Gyrogroups,
I will be able to develop a new glimpse.

III) PC_Oculto/ –

Electromagnetic Fields (IEEE Press Series on Electromagnetic Wave theory)_images/

Electromagnetic Fields..._0028.jpg

1.4 Green's Functions

25.01.2020

Green's functions by considering
the displacement of a beam under
a distributed force $p(x)$.

... Except at x' there is no force.

sobre uma superfície.

Veja o item I,

uso da luz da laser sobre uma
superfície.

Important:

A luz exerce uma pressão
constante, conhecida como:

"pressão de radiação"!

108) 20200127

1) www2.math.uu.se/~thomase/
GeometryOverFields.pdf
From Geometry to Number
The Arithmetic Field
implicit in Geometry
Johan Alm.
49 pages.

Jandaia

25.01.2020

the Green's functions by considering
the displacement of a beam under
a distributed force $p(x)$.

... Except at x' there is no force.

sobre uma superfície.

Veja o item I,

uso da luz da laser sobre uma
superfície.

Important:

A luz exerce uma pressão
constante, conhecida como
"pressão de radiação"!

108) 20200127

1) www2.math.uu.se/~thomase/
GeometryoverFields.pdf
From Geometry to Number
The Arithmetic Field
implicit in Geometry
Johan Alm.
49 pages.

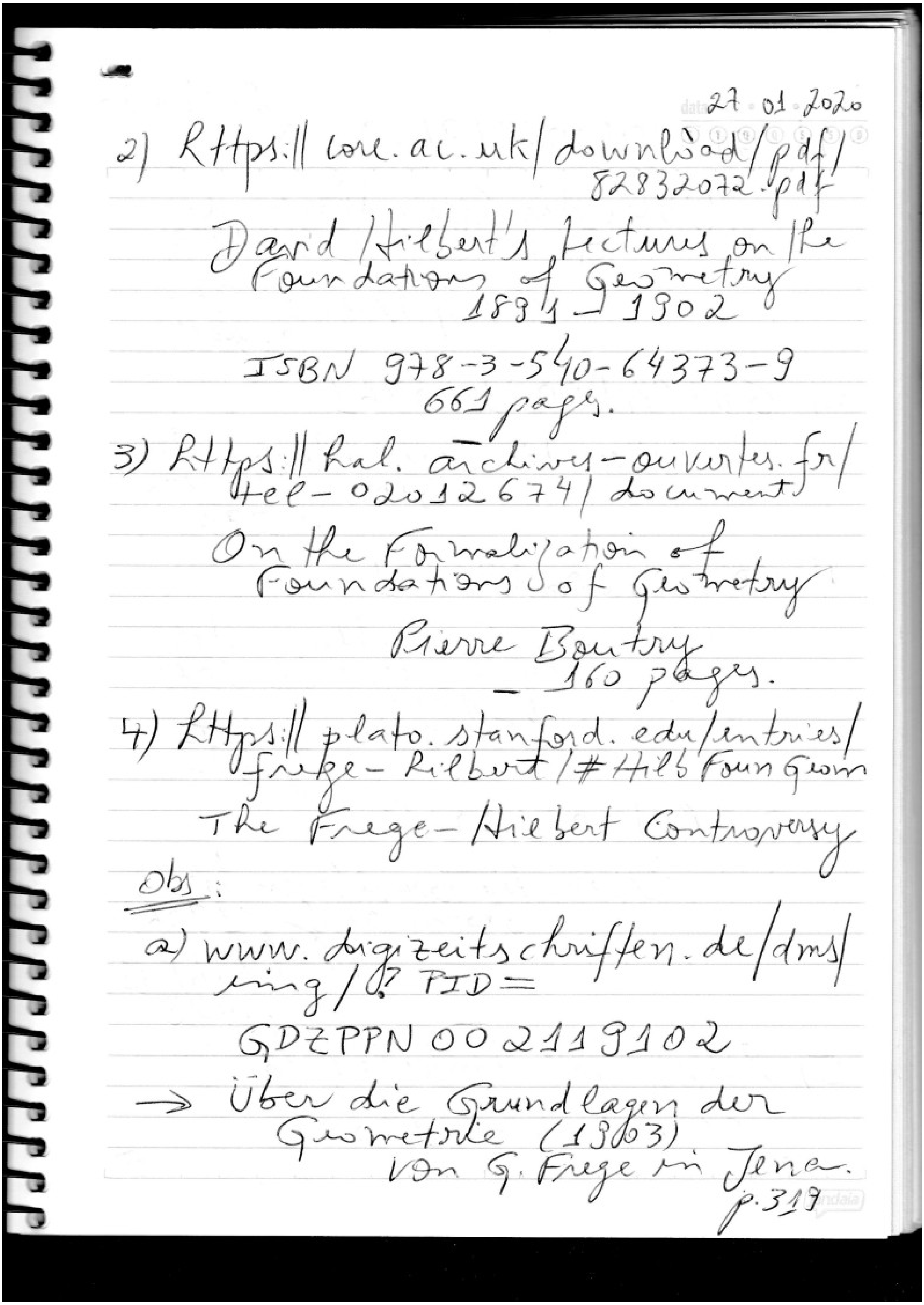

27.01.2020

2) https://core.ac.uk/download/pdf/82832072.pdf

David Hilbert's Lectures on the Foundations of Geometry
1891-1902

ISBN 978-3-540-64373-9
661 pages.

3) https://hal.archives-ouvertes.fr/tel-02012674/document

On the Formalization of Foundations of Geometry

Pierre Boutry
– 160 pages.

4) https://plato.stanford.edu/entries/frege-hilbert/#HilbFounGeom

The Frege-Hilbert Controversy

Obs:

a) www.digizeitschriften.de/dms/img/?PID=
GDZPPN002119102

→ Über die Grundlagen der Geometrie (1903)
von G. Frege in Jena.
p.319

% 2 ## ! : ! #! # #; '6<46= > $

./01. . .+ . -
>&((- (?* *** >&((- (?* *** - 2!

@9 ! 4 ! ! 4

9 2 ## ! : ! #! # #; '6<46= > $ $

./01. . .+ . -
>&((- (?* *** >&((- (?* *** $ 2!

@9 ! 4 ! ! 4 "

2 ## ! : ! #! # #; '6<46= > %"

./01. . .+ . -
>&((- (?* *** >&((- (?* *** (" 2!

@9 ! 4 ! ! 4
3 A 4

! 2 ## ! : ! #! # #; '6<46= > "&

./01. . .+ . -
>&((- (?* *** >&((- (?* *** $& 2!

@9 ! 4 ! ! 4 "
3 A 4

2 ## ! : ! #! # #; '6<46= > """

./01. . .+ . -
>&((- (?* *** >&((- (?* *** - 2!

@9 ! 4 ! ! 4 "'
3 A 4

27.01.2020

2) https://core.ac.uk/download/pdf/82832072.pdf

David Hilbert's Lectures on the Foundations of Geometry
1891-1902

ISBN 978-3-540-64373-9
661 pages.

3) https://hal.archives-ouvertes.fr/tel-02012674/document

On the Formalization of Foundations of Geometry

Pierre Boutry
- 160 pages.

4) https://plato.stanford.edu/entries/frege-hilbert/#HilbFounGeom

The Frege-Hilbert Controversy

Obs:

a) www.digizeitschriften.de/dms/img/?PID=

GDZPPN002119102

→ Über die Grundlagen der Geometrie (1903)
von G. Frege in Jena.
p.319

27.01.2020

b) GDZPPN002119129

Über die Grundlagen der Geometrie II (1903)
Von G. Frege in Jena
p. 368

c) GDZPPN002120461

Über die Grundlangen der Geometrie-I (1906)
p. 293

d) GDZPPN002120631

Über die Grundlagen der Geometrie (1906)
Frege
(Fortsetzung) — II
p. 377

e) GDZPPN002120666

Über die Grundlagen der Geometry (1906)
Frege
(Schluß) III
p. 423

—

5) https://math.berkeley.edu/~wodzicki/160/

⇒

27.01.2020

MATH 160 – Introduction to History of Mathematics.

Links to: Papyrus Rhind
1 – 79.(pdf)

Papyrus Moscow → 1 – 25

Lahun Papyrus ... etc.

6) The Plimpton 322 Tablet

https://arxiv.org/pdf/1004.0025.pdf

obs:

The Plimpton 322 Tablet and the Babylonian Method of Generating Pythagorean Triples.
34 pages.

7) ftp.math.utah.edu/pub/text/toc/histmath.html

Volume 1 → 49 (2019).

8) https://geometryexpressions.com

Obs.: Pode-se usar na internet.
Download → pago.

27.01.2020

9) CORE. AC. UK

10) https://www.nbarchive.dk/publications/bcw/
Niels Bohr Collected Works.
12 Volumes

obs.:
Harald August Bohr (1887-1951)

Niels Bohr (Niels Henrick David Bohr)
(1885 - 1962)

Harald August Bohr →
Collected Mathematical Works
3 volumes.
www.archive.org/details/
collectedmathema0000bohr
→ Volume I
collectedmathema0003bohr
→ Volume III

obs: Falta o Volume II.

27.01.2020

9) CORE. AC. UK

10) https://www.nbarchive.dk/publications/bcw/
Niels Bohr Collected Works.
12 Volumes

obs.:
Harald August Bohr (1887-1951)

Niels Bohr (Niels Henrick David Bohr)
(1885 – 1962)

Harald August Bohr →
Collected Mathematical Works
3 volumes.

www.archive.org/details/
collectedmathema0000bohr
→ Volume I
collectedmathema0003bohr
→ Volume III

obs: Falta o Volume II.

27.01.2020

11) www.plouffe.fr/simon

10) cont.

Harald Bohr Papers

web.math.ku.dk/arkivet/hbohr/hbpapers.htm

12) 20200202

The Mathematical Papers of Georg Friedrich Bernhard Riemann (1826-1866)

Source:

https://www.emis.de/classics/Riemann/

13/ Latex

linorg.usp.br/CTAN/systems/win32/miktex/tm/packages/

109) 20200202

Inkscape program

Extensions -> Render -> Spirograph
-> Wireframe Sphere
-> Gear

27-01-2020

11) www.plouffe.fr/simon

10) cont.

Harald Bohr Papers

web.math.ku.dk/arkivet/hbohr/
hbpapers.htm

12) 20200202

The Mathematical Papers of
Georg Friedrich Bernhard
Riemann (1826-1866)

Source:

https://www.emis.de/classics/
Riemann/

13/ Latex

linorg.usp.br/CTAN/systems/win32/
miktex/tm/packages/

109) 20200202

Inkscape program

Extensions → Render → Spirograph
→ Wireframe
→ Gear Sphere

02 02 2020

Source: Arq-Important_03/Inkscape_files/

a) drawing-20200202.svg

I) Gear ⇒ 24 teeth (360÷24=15°)

Configuration:

Nº teeth: 24

Circular pitch: 20 ⇒ 15

Pressure angle: 20

Diameter: 20

Units: px ⇒ mm (altere)

II) Spirograph

R – Ring Radius (px): 100

r – Gear Radius (px): 60

d – Pen Radius (px): 50

Gear placement { Inside (Hypotrochoid)
Outside (Epitrochoid)

Rotation – 0

Jandaia

data 02.02.2020

Quality (Default) — 16

b) Spirograph

spirograph_200.60.50.e.200.100.50.svg

I) Spirograph

200. 60. 50. Inside. 24. 16
(R) (r) (d) (GP) (R) (Q)

→ spirograph 10 points

II) Spirograph

200. 100. 50. Inside. 24. 16

ellipse.

Source: spirograph_200.60.50. e. 200.100.50.svg

Position:

Spiro 1 → x 55,430
y 100,207
W 100,806
H 95,919 mm

Spiro 2 → x = 65
y = 134,805
W = 81,540
H = 26,723 mm

Obs: Use as setas para movimentá-los até a posição.

data 02 02 2020

c) spirograph_100.200.50.e.100.60.100.svg

Spiro 1 : 100 . 200 . 50
(R) (r) (d)

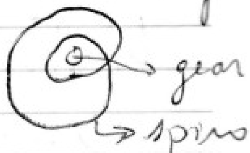

Spiro 2 : 100 . 60 . 100
(R) (r) (d) } 5 pontas (pentagrama)

d) spirograph_100.60.50.e.300.60.50.svg

R - 300
r - 60
d - 50
G - Inside
R - 0
Q - 16

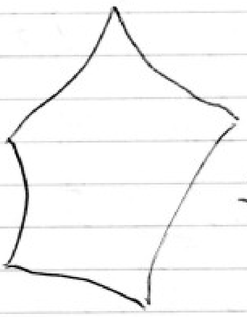

5 pontas.

e) ~~spirograph~~ gear_25.20.15.20.mm.e.spiro.300.60.50.inside.0.16.svg

A engrenagem e o spirograph ficam alinhados

25 dentes divididos por 5 pontas

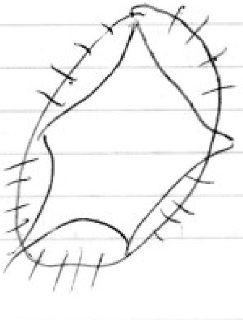

data 02.02.2020

Outras simulações:

f) spirograph_200.60.50.e.200.100.50_ajustado.svg

g) gear_25.20.15.20.mm.e.spiro_300.60.50.inside.24.16.svg

110) 20200202 Riemann

geometrias não-euclidianas

Lobachevsky
Bolyai
Gauss
Riemann

—

Bernhard Riemann
On the hypothesis which lie at the bases of geometry

—

postulado paralelo (euclides)

Não há geodésicas paralelas.

(Gauss foi o primeiro a testar a curvatura do espaço físico.

179° 59' 59.32"

02.02.2020

Outras simulações:

f) spirograph_200.60.50.e.200.100.50_ajustado.svg

g) gear_25.20.15.20.mm.e.spiro_300.60.50.inside.24.16.svg

110) 20200202 Riemann

geometrias não-euclidianas
Lobachevsky
Bolyai
Gauss
Riemann

—

Bernhard Riemann
On the hypothesis which lie at the bases of geometry

—

postulado paralelo (euclides)

Não há geodésicas paralelas.

(Gauss foi o primeiro a testar a curvatura do espaço físico.

179°59'59.32"

data 02.02.2020

usou os montanhas da Alemanha

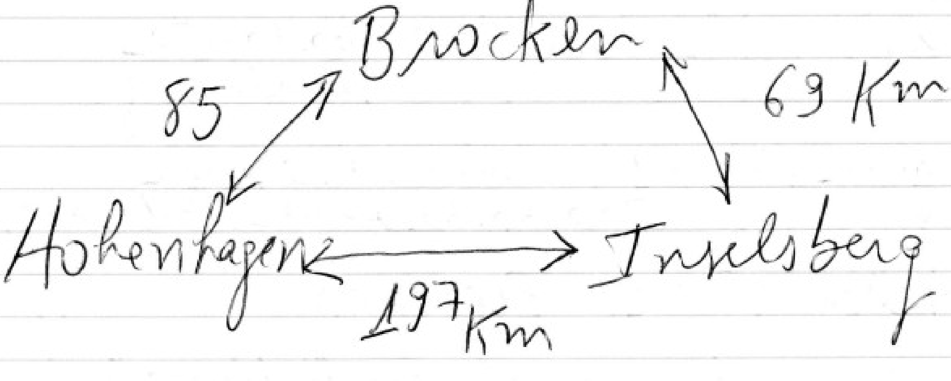

Beltrami (Eugenio)
"pseudo-esfera"

descrevendo um espaço hiperbólico de constante curvatura negativa por meio de uma equação análoga à da hiperesfera, evidencia a natureza comum das hipóteses de Bolyai e Lobachevsky e de Riemann.

Book: Uma história não-euclidiana.
Rosenfeld.

Source: Books_2020_06/Book1_27_2020/
Important
Boris A. Rosenfeld _ A_history_of_non-Euclidean ... djvu

02.02.2020
8

Source: Books_2020_06/OTHERS/
Math24.net/
Screenshot_2020-02-03_
Curvature of Plane Curves.png

Osculating → Osculate

Verb:

(of a curve or surface) touch (another curve or surface) so as to have a common tangent at the point of contact.

Should the students be asked to sketch the osculating circle to a given curve at a given point, and then notice that as the point of contact changes, the radius of the circle is inversely proportional to the measure of curvature?

If seven points on a twisted cubic be joined, two and two, by twenty-one lines, then any seven planes that contain these 21 lines will osculate a second cubic curve.

Book: A history of non-Euclidean geometry
B. A. Rosenfeld

) - $

Chapter 8

The Curvature of Space

Curvature and Intrinsic Geometry of a Surface in the Works of Euler

By the *curvature* of a curve at a point we mean the limit of the ratio of the angle $\Delta\alpha$ between the tangents at the endpoints of an arc to the length Δs of that arc as the latter contracts to the point; that is, the limit

$$k = \frac{d\alpha}{ds} = \lim_{\Delta s \to 0} \frac{\Delta\alpha}{\Delta s}. \tag{8.1}$$

k is also the reciprocal of the *radius of curvature* at the point in question, that is, the radius of the osculating circle at that point. (The osculating circle at a point P of a curve is defined as the limit of circles determined by three points on the curve as they tend to P.) The concepts of the curvature of a curve and of the osculating (literally "kissing," from the Latin *osculans*) circle were already known to Leibniz.[1] Leibniz also suggested the possibility of characterizing the curvature of a surface by means of an osculating sphere.[2]

Euler characterized the curvature of a surface in his *Investigations of the curvature of surfaces* (Recherches sur la courbure des surfaces. Berlin, 1767) *[176, vol. 28, pp. 1–22]*. Euler called a normal section of a surface $z = f(x, y)$ perpendicular to the xOy plane a *principal section*. By varying the angle φ between a normal section and the principal section Euler found that at each point of the surface there is a maximal radius of curvature f and a minimal radius of curvature g, that their planes are perpendicular to one another, and

---

[1] Leibniz first considered osculation of curves—in particular, osculation of curves and circles—in his *New reflections on the nature of the angles of tangency and osculation and their use in mathematical practice to replace complex figures with simpler ones* (Meditatio nova de natura Anguli contactus et osculi, horumque usu in practica Mathesi ad figuras faciliores succedaneas difficilioribus substituendas. Leipzig 1686) *[312, v. 7, pp. 326–329]*.

[2] In a letter to Johann Bernoulli, dated July 29, 1698.

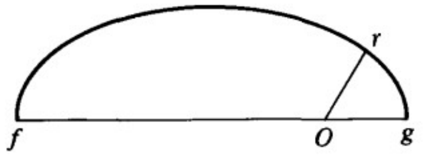

Figure 101

that the connection between an arbitrary radius of curvature $r = r(\varphi)$ and the radii of curvature f and g is given by

$$r = \frac{2fg}{f + g - (f - g)\cos 2\varphi}.$$

Euler writes:

> For a simple derivation of this formula we join the maximal and minimal osculating radii by putting $Of = f$ and $Og = g$ and describe on fg a semi-ellipse with one focus at O (Figure 101). Then for the section MN one must take the angle fOr, doubled with respect to the angle EZM, and the line Oz will be equal to the osculating radius for the section MN. Thus drawing conclusions about the curvature of surfaces, at first glance a matter of great complexity, reduces for each element to the determination of two osculating radii one of which is maximal and the other minimal for that element; these two things entirely determine the nature of the curvature, and we obtain the curvature of all possible sections perpendicular to the given element *[176, vol. 28, p. 21]*.

In 1837 Charles Dupin (1784 1873) *[157, p. 109]* modified Euler's formula to read

$$\frac{1}{r} = \frac{1}{f}\cos^2 \varphi + \frac{1}{g}\sin^2 \varphi.$$

In this paper *On solids whose surfaces can be developed onto a plane* (De solidis quarum superficiem in planum explicare licet. Petersburg, 1772) *[176, vol. 28, pp. 161–186]* Euler introduced the concept of a *developable surface*, that is, a surface that can be applied to a plane without folding or tearing, and proved the fundamental theorem that such a surface is a cylinder, or a cone, or a surface formed by the tangents to a space curve. His point of departure was that on such a surface an infinitesimal triangle must be congruent to the corresponding triangle in the plane to which the surface is applied. Further, he introduced on the surface curvilinear coordinates t, u equal to the rectangular coordinates of the corresponding points of the plane and, denoting the partial derivatives $\partial x/\partial t$, $\partial y/\partial t$, $\partial z/\partial t$ by l, m, n, and $\partial x/\partial u$, $\partial y/\partial u$, $\partial z/\partial u$ by λ, μ, ν, set down the developability conditions in the form $l^2 + m^2 + n^2 = 1, \lambda^2 + \mu^2 + \nu^2 = 1, l\lambda + m\mu + n\nu = 0$, that is, as the requirements of unicity and perpendicularity of the partial derivatives with respect to t and u of the radius vector of a point of the surface. An equivalent form of this condition is that the square of an arc element of a curve on a developable surface equals the sum of the squares of the differentials of the coordinates,

$$ds^2 = dt^2 + du^2,$$

that is, the arc element of a developable surface coincides with the arc element of a plane.

In a note published only in 1862 *[176, vol. 29, pp. 437–440]* Euler also established the general conditions of applicability of a surface to other surfaces. These results of Euler pertain to the so-called *intrinsic geometry of a surface*, that is, the study of properties of a surface that are unchanged when the surface is bent.

Euler's work on *geodesics* also pertains to the intrinsic geometry of surfaces. A *geodesic* is the shortest of all curves on a surface joining two of its points (in the plane the geodesics are straight lines and on a sphere they are great circles).[3]

The problem of finding geodesics on a surface was first posed in 1697 by Johann Bernoulli (1667–1748). Shortly after that, Johann Bernoulli wrote a letter to l'Hospital to the effect that he had found the general differential equation of geodesics. This result was published only in 1742 *[50, p. 364]*. Jacob Bernoulli (1654–1705) showed in 1698 that when cylindrical and conical surfaces are applied to planes, their geodesics go over into straight lines.

Euler was the first to publish the differential equation of geodesics. He did this in the paper *On the shortest curve on an arbitrary surface joining two arbitrary points* (De linea brevissima in superficie quacumque duo quaelibet puncta jungente. Petersburg, 1732) *[176, v. 25, pp. 1–12]*. Euler's equation was

$$\frac{Qddx + Pddy}{Qdx + Pdx} = \frac{dxddx + dyddy}{dt^2 + dx^2 + dy^2},$$

where t, x, y are rectangular coordinates in three-dimensional space, and the functions P and Q are determined from the differential equation of the surface $Pdx = Qdy + Rdt$. Euler returned to the problem of geodesics many times. In the second volume of his *Mechanics* (Mechanica. Petersburg) *[177, v. 2, p. 426]* he showed that, "in the absence of forces," the path of a point on a surface is a geodesic (an assumption used earlier by Bernoulli). In this connection, Euler showed that the *principal normal of a geodesic* (a straight line in the osculating plane of that curve passing through the point of contact and perpendicular to the tangent) always coincides with the normal to the surface.

---

[3] Many of Euler's papers deal with the determination of geodesics. The first was published in 1728 *[176, vol. 25, pp. 1–12]*, and the last was published posthumously *[176, vol. 25, pp. 269–279]*. Euler called geodesics *shortest curves*. The term *geodesic* first appeared in Laplace's *Celestial Mechanics* (Paris, 1799). Later it was applied to all quadrics. After the appearance of Liouville's paper (1814) it was applied to all surfaces.

We note that in the cartography paper *On the representation of the surface of a sphere in the plane* Euler compared the arc length elements of the plane and the sphere and proved the impossibility of an isometric mapping of a sphere to a plane. Also, he posed the question about the three mappings of the sphere to the plane that are of greatest importance in cartography, namely, the mapping in which the meridians and parallels are mapped onto an orthogonal system of lines, the mapping that preserves angles between curves (conformal mapping), and the mapping that preserves areas (equiareal mapping).

Independently of Euler, Monge considered developable surfaces in his *Memoir on evolutes, radii of curvature and various kinds of inflection of curves with twofold curvature* (Mémoire sur les développées, des rayons de courbure et les differents genres d'inflexion des courbes à double courbure. Paris, 1785) *[374]*. We note that Monge's book *Applications of analysis to geometry* (Applications de l'analyse à la géométrie. Paris, 1807) *[373]* contained the first systematic exposition of the theory of surfaces.

Intrinsic Geometry of Surfaces in the Work of Gauss

The general theory of the intrinsic geometry of a surface was formulated by Carl Friedrich Gauss in his *General investigations of curved surfaces* (Disquisitiones generales circa superficies curvas. Göttingen, 1828) *[196, vol. 4, pp. 217–258]*. Gauss's paper was the result of the assignment, given to him in 1820, to produce a cartographic survey of the kingdom of Hanover. This assignment made him reflect on problems of geodesy. In turn, this brought him to the theory of surfaces.

Gauss's theory unified the surface theories of Euler and Monge and the theory of quadratic forms developed by him in his *Arithmetical investigations* (Disquisitiones arithmeticae. Göttingen, 1801) *[196, vol. 1]*. In *Investigations of curved surfaces* Gauss introduced curvilinear coordinates p, q of points of the surface and defined the functions

$$a = \frac{dx}{dp}, \qquad b = \frac{dy}{dp}, \qquad c = \frac{dz}{dp}, \qquad a' = \frac{dx}{dq}, \qquad b' = \frac{dy}{dq},$$

$$c' = \frac{dz}{dq}; \qquad bc' - cb' = A, \qquad ca' - ac' = B', \qquad ab' - ba' = C;$$

$$\frac{ddx}{dp^2} = \alpha, \qquad \frac{ddx}{dpdq} = \beta, \qquad \frac{ddx}{dq^2} = \gamma, \qquad \frac{ddy}{dp^2} = \alpha',$$

$$\frac{ddy}{dpdq} = \beta', \qquad \frac{ddy}{dq^2} = \gamma', \qquad \frac{ddz}{dp^2} = \alpha'', \qquad \frac{ddz}{dpdq} = \beta'', \qquad \frac{ddz}{dq^2} = \gamma''.$$

In modern terms, the vectors $\{a, b, c\}$ and $\{a', b', c'\}$ are the partial derivatives of the radius vector of a point on the surface with respect to p and q

(Gauss denoted ordinary and partial derivatives the same way); the vector $\{A, B, C\}$ is the cross product of $\{a, b, c\}$ and $\{a', b', c'\}$ and its direction is that of the normal to the surface; and $\{\alpha, \beta, \gamma\}$, $\{\alpha', \beta', \gamma'\}$, and $\{\alpha'', \beta'', \gamma''\}$ are the second partial derivatives of the radius vector of a point on the surface with respect to p and q.

Further, Gauss introduced the quadratic forms

$$ds^2 = Edp^2 + 2Fdpdq + Gdq^2$$

and

$$Ddp^2 + D'dpdq + D''dq^2,$$

where

$$aa + bb + cc = E, \quad aa' + bb' + cc' = F, \quad a'a' + b'b' + c'c' = G,$$

$$A\alpha + B\beta + C\gamma = D, \quad A\alpha' + B\beta' + C\gamma' = D', \quad A\alpha'' + B\beta'' + C\gamma'' = D''.$$

The first of these forms, the so-called first fundamental form, gives the square of an arc length element of the surface, that is, of an element of length of an arc of a curve on the surface, and the second differs from the modern second fundamental form only by the multiplier $AA + BB + CC = EG - FF$.

Gauss introduced the all-important concept of "measure of curvature" now known as the *Gaussian curvature* of a surface. Making use of the spherical image of the surface, that is, the mapping that associates to a point P on the surface, a point P' on the unit sphere, such that the radius OP' is parallel to the normal to the surface at P (in modern terms, the end of the unit normal to the surface with beginning at a fixed point), Gauss defined the concepts of total curvature and "measure of curvature" as follows:

> We shall say that a part of a curved surface bounded by a given contour has *total curvature* given by the area of the corresponding figure on the surface of the sphere. One must draw a clear distinction between the total curvature and a kind of specific curvature that we shall call the *measure of curvature*. The latter pertains to a point on the surface and is the fraction obtained by dividing the total curvature of an element of the surface adjacent to the point by the area of that element, and thus gives the ratio of corresponding infinitesimal areas on the sphere and on the curved surface. We hope that our subsequent exposition will fully explain the utility of these new concepts" *[196, vol. 4, p. 226]*.

Gauss goes on to show that

> the "measure of curvature" at a point of the surface is equal to a fraction whose numerator is 1 and whose denominator is the product of the two principal curvatures of the normal sections.
>
> It is clear that the measure of curvature is positive for convex-convex or concave-concave surfaces (this is a trivial distinction) and negative for convex-concave ones. If the surface is made up of parts of both

kinds, then on their boundary the measure of curvature must vanish *[196, vol. 4, pp. 231]*.

Then Gauss obtains further expressions for the "measure of curvature" of which the most important are the expression of the "measure of curvature" k in terms of the coefficients of his two forms,

$$k = \frac{DD'' - D'D'}{EG - FF},$$

and the expression in terms of the coefficients of the first form alone and their derivatives with respect to p and q:

$$\begin{aligned} 4(EG - FF)^2 k = {} & E\left(\frac{dE}{dq}\frac{dG}{dq} - 2\frac{dF}{dp}\frac{dG}{dq} + \left(\frac{dG}{dp}\right)^2\right) + \\ & + F\left(\frac{dE}{dp}\frac{dG}{dq} - \frac{dE}{dq}\frac{dG}{dp} - 2\frac{dE}{dq}\frac{dF}{dq} + \frac{dF}{dp}\frac{dF}{dq} - \right. \\ & \left. - 2\frac{dF}{dp}\frac{dG}{dp}\right) + G\left(\frac{dE}{dp}\frac{dG}{dp} - \frac{dE}{dp}\frac{dF}{dq} + \left(\frac{dE}{dq}\right)^2\right) - \\ & - 2(EG - FF)\left(\frac{ddE}{dq^2} - 2\frac{ddF}{dpdq} + \frac{ddG}{dp^2}\right). \end{aligned}$$

In connection with the last formula Gauss makes the following comment:

> The formula in the last section leads, of itself, to the following remarkable
>
> *Theorem*. If a curved surface is applied to any other surface then the measure of curvature at each of its points remains unchanged.
>
> Also, it is clear that every finite part of a curved surface will, after application to another surface, retain its total curvature.
>
> The special case to which geometers have until now limited their investigations is the case of surfaces applicable to a plane. Our theory readily shows that the measure of curvature of such surfaces at any point is zero. . . .
>
> What we have set forth in the previous section involves a special approach to the investigation of surfaces worthy of the close attention of geometers. Namely, when a surface is regarded not as the boundary of a solid but as a solid with one vanishing dimension, flexible but not stretchable, then the properties of the surface are partly dependent on the form to which it has been reduced and in which it is being studied, and partly independent [of it in the sense that] they remain invariant regardless of its form under bending. The latter properties, whose investigation opens a new and fruitful area of geometry, include the measure of curvature and total curvature as defined by us. They also include the study of geodesic curves, as well as many other matters to be discussed in the sequel. When this approach is adopted, then the plane, and surfaces applicable to a plane, such as cylindrical and conical surfaces,

> are regarded as essentially the same, and the general method of characterizing a surface studied in this manner is based on the formula $\sqrt{Edp^2 + 2Fdpdq + Gdq^2}$ that expresses the connection of the line element of the surface with two variables p and q *[196, vol. 4, pp. 237–238]*.

Gauss is defining the *intrinsic geometry* of a surface invariant under bending and application to another surface. Some of the properties studied in intrinsic geometry are the lengths of curves on the surface, the angles between curves, and what Gauss calls *geodesic curves*, that is, geodesics. Gauss's theorem, called by him a *remarkable theorem* (Theorema Egregium), asserts that the *measure of curvature* is a property that belongs to intrinsic geometry.

Further, Gauss considers the total curvature of a geodesic triangle (that is, a triangle whose sides are arcs of geodesics). The total curvature of such a triangle is given by $\int k d\sigma$, where $d\sigma$ is an element of the surface of the triangle. He finds that

> the total curvature of a triangle is equal to the part of the spherical surface that corresponds to the triangle taken with a plus or minus sign according as the surface on which the triangle lies is concave-concave or concave-convex; as the unit of area we take a square with side one (the radius of the sphere); then the area of the sphere is 4π. It follows that the [surface of the] part of the sphere corresponding to the triangle is to the surface of the sphere as $\pm(A + B + C - \pi)$ to 4π. This theorem, that undoubtedly belongs with the most beautiful theorems of the theory of surfaces, can be stated as follows:
>
> *The excess above* 180° *of the sum of the angles of a triangle formed by geodesic curves on a curved concave-concave surface and the defect below* 180° *of the sum of the angles of a triangle formed by geodesic curves on a concave-convex surface are each measured by the area of the part of the spherical surface that corresponds to the given triangle via normal directions, provided that the whole surface* (*of the sphere*) *is taken as* 720 *degrees [196, vol. 4, pp. 245–246]*.

This theorem of Gauss implies that the Gaussian curvature at a point is equal to the limit of the ratio of the angular excess $A + B + C - \pi$ of a geodesic triangle ABC on the surface to its area as the triangle shrinks to the given point.

Minding's Theory of Surfaces of Constant Curvature

Gauss's work on the intrinsic geometry of surfaces was continued by Ferdinand Minding (1806–1885), who worked in Dorpat (now Tartu). In his *Remark on the development of curved lines on surfaces* (Bemerkung über die Abwickelung krummer Linien auf Flächen. Berlin, 1830) *[366]*, Minding introduced the fundamental concept of *geodesic curvature* of a curve on a surface, defined as the limit of the ratio of the angle between tangent geodesics at the endpoints of an arc of the curve to the arc length as the arc shrinks to a point, and showed that this limit is invariant under bending. In the paper *How to decide whether two surfaces are mutually applicable; including remarks on surfaces of constant measure of curvature* (Wie sich entscheiden lässt ob zwei gegebene krumme Flächen auf einander abwickelbar sind, nebst Bemerkungen über die Flächen von unveränderlicher Krümmungsweise. Berlin, 1839) *[367]* Minding expressed the line element of a surface with constant "measure of curvature" as

$$ds^2 = dp^2 + \left(\frac{1}{\sqrt{k}} \sin p\sqrt{k}\right)^2 dq^2$$

and concluded that

> two surfaces of equal constant measure of curvature can be applied to each other in infinitely many ways, since any two points of one can be associated to two arbitrary points of the other provided that the lengths of the shortest curves between the points of each pair are equal. This implies the following consequences:
>
> Every surface whose measure of curvature is zero [results from] the bending of a plane—this much is well known.
>
> Every surface whose measure of curvature (k) is constant and positive can be applied to a sphere of radius $1/\sqrt{k}$ *[367, pp. 375–376]*.

Further, Minding discovered helical surfaces and, in particular, surfaces of constant negative curvature. One such surface is given by the parametric equations

$$r = \sqrt{x^2 + y^2} = \frac{1}{\cosh\varphi}, \qquad z = \varphi - \tanh\varphi.$$

Minding comments that

> This surface is the result of rotating a curve like ... *CBE* ... (Figure 102) about the zz-axis which it approaches asymptotically *[367, p. 380]*.

The curve *CBE*, which is characterized by the constancy of the length of the segment of the tangent between the point of tangency and the zz-axis, is called a *tractrix*, and the surface of revolution referred to by Minding is called a *pseudosphere*.

In *Contributions to the theory of shortest curves on curved surfaces* (Beiträge zur Theorie der kürzesten Linien auf krummen Flächen. Berlin, 1840) *[368]* Minding established the trigonometric relations in geodesic triangles on such surfaces and noted that these formulas could be obtained from the corresponding formulas of spherical geometry on a sphere of radius r by multiplying r by the complex number i.

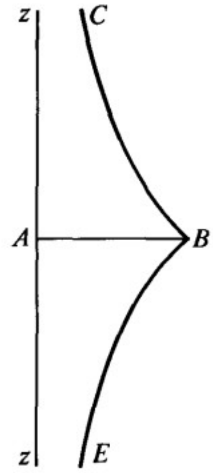

Figure 102

Interpretation of the Lobačevskian Plane on a Pseudosphere

In spite of the fact that Minding's paper appeared in the same journal (*Crelle's*) as Lobačevskiĭ's *Imaginary geometry*, neither of them noticed that the trigonometric formulas of the hyperbolic plane coincide with the trigonometric formulas of a surface of constant negative curvature.[4]

The person who *did* notice the sameness of these formulas was E. Beltrami, who considered the matter in *An attempt at an interpretation of non-Euclidean geometry [42, pp. 374–405]*. Beltrami constructed a model of the hyperbolic plane in a circle and associated to the points of the hyperbolic plane of curvature $-1/R^2$ the coordinates u, v of the corresponding points of the circle. He found that in this coordinate system the line element is given by

$$ds^2 = R^2\frac{(a^2-u^2)du^2+2uv\,du\,dv+(a^2-v^2)dv^2}{a^2-u^2-v^2}.$$

Beltrami computed the total curvature of a surface with this line element and noticed that the Gaussian curvature of the hyperbolic plane is everywhere equal to the same number $-1/R^2$; that is, the hyperbolic plane can be viewed as a surface of constant negative curvature.

Thus, speaking of surfaces of constant negative curvature, Beltrami concluded that

---

[4] Lobačevskiĭ's paper appeared in vol. 17 (1837) of *Crelle's Journal* and Minding's paper in vol. 20 (1840) of that journal. B. L. Laptev *[265, p. 20]* explains that Lobačevskiĭ did not borrow the latter volume from the library of Kazan University.

> the theorems of non-Euclidean planimetry apply to them. Further, most of these theorems can be thought of concretely only if we refer them to such surfaces and not to the plane *[42, p. 381]*.

That is why Beltrami suggested that all surfaces of constant negative curvature be called *pseudospherical* surfaces.

Beltrami also showed that each of the sheets of Minding's pseudosphere is isometric to the part of the hyperbolic plane enclosed between two parallel lines and a horocycle perpendicular to them. He also found parts of the hyperbolic plane isometric to other surfaces of revolution of constant negative curvature determined by Minding.

Riemannian Geometry

In the paper *On the hypotheses which lie at the foundations of geometry [122, pp. 55–71]* Riemann introduced a notion of an n-*ply extended manifold* broader than either of the similar notions introduced by Grassmann and Schläfli. After defining this manifold Riemann posed the question of the "measure-relations of which such a manifold is capable" and of the possibility "to express geometrically the calculated results." After noting that the foundations of these two parts of the question are established in Gauss's celebrated memoir *Disquisitiones generales circa superficies curvas [122, pp. 59–60]*, Riemann writes:

> Position-fixing being reduced to quantity-fixings, and the position of a point in the n-ply extended manifold being consequently expressed by means of n variables $x_1, x_2, x_3, \ldots x_n$, the determination of a line comes to the giving of these quantities as functions of one variable. The problem consists then in establishing a mathematical expression for the length of a line, and to this end we must consider the quantities x as expressible in terms of certain units. I shall treat this problem only under certain restrictions, and I shall confine myself in the first place to lines in which the ratios of the increments dx of the respective variables vary continuously. We may then conceive these lines broken up into elements, within which the ratios of the quantities dx may be regarded as constant; and the problem is then reduced to establishing for each point a general expression for the linear element ds starting from that point, an expression which will thus contain the quantities x and the quantities dx. I shall suppose, secondly, that the length of the linear element, to the first order, is unaltered when all the points of this element undergo the same infinitesimal displacement, which implies at the same time that if all the quantities dx are increased in the same ratio, the linear element will vary also in the same ratio. On these suppositions, the linear element may be any homogeneous function of the first degree of the quantities

> dx, which is unchanged when we change the signs of all the dx, and in which the arbitrary constants are continuous functions of the quantities x. To find the simplest cases, I shall seek first an expression for manifoldness of $n - 1$ dimensions which are everywhere equidistant from the origin of the linear element; that is, I shall seek a continuous function of position whose values distinguish them from one another. In going outwords from the origin, this must either increase in all directions or decrease in all directions; I assume that it increases in all directions, and therefore has a minimum at that point. If, then, the first and second differential coefficients of this function are finite, its first differential must vanish, and the second differential cannot become negative; I assume that it is always positive. This differential expression, then, of the second order remains constant when ds remains constant, and increases in the duplicate ratio when the dx, and therefore also ds, increase in the same ratio; it must therefore be ds^2 multiplied by a constant, and consequently ds is the square root of an always positive integral homogeneous function of the second order of the quantities dx, in which the coefficients are continuous functions of the quantities x. For Space, when the position of points is expressed by rectilinear co-ordinates, $ds = \sqrt{\sum (dx)^2}$ *[122, pp. 60–61]*.

(In Clifford's translation we changed his term n-*dimensional manifoldedness* to Riemann's n-ply extended manifold). By *Space* Riemann means three-dimensional Euclidean space. After considering various possible cases of the dependence of ds on the differentials dx, Riemann restricts himself to "manifolds for which the line element is given by the square root of a differential expression of the second degree," that is,

$$ds^2 = \sum_i \sum_j g_{ij} dx_i dx_j, \tag{8.2}$$

where $g_{ij} = g_{ji}$ are functions of the variables x_i.

Riemann's requirement guarantees that his n-dimensional space is locally Euclidean. The quadratic form (8.2) is assumed to be positive definite, that is, for all dx_i not simultaneously zero $ds^2 > 0$. Hence each pair of points with infinitesimally different coordinates x_i and $x_i + dx_i$ are a definite distance ds apart. By integrating this distance along different lines we determine their length. Among the different lines we can find the shortest (geodesic) lines $x_i = x_i(t)$ that are solutions of the differential equations

$$\frac{d^2 x_i}{ds^2} + \sum_j \sum_k \Gamma^i_{jk} \frac{dx_j}{ds} \frac{dx_k}{ds} = 0, \tag{8.3}$$

where the Γ^i_{jk} are point functions expressed in terms of the coefficients g_{ij} by means of the relations

$$\sum_i g_{il} \Gamma^l_{jk} = \frac{1}{2}\left(\frac{\partial g_{lk}}{\partial x_j} + \frac{\partial g_{il}}{\partial x_k} - \frac{\partial g_{jk}}{\partial x_l}\right). \tag{8.4}$$

The spaces defined in this way are called *Riemannian spaces*. Surfaces in ordinary space are particular Riemannian spaces for $n = 2$; in this case, an instance of a geometry of the kind defined by Riemann is the so-called intrinsic geometry of a surface, elaborated by Gauss in the work *General investigations of curved surfaces* alluded to by Riemann. Clearly, Euclidean space is a special case of a Riemannian space whose geodesics are Euclidean straight lines.

A more general example of a Riemannian space is a hypersurface in $(n + 1)$-dimensional Euclidean space and, in particular, a hypersphere in this space. The geodesics of a hypersphere are its great circles—sections by two-dimensional planes passing through its center. Elliptic space, obtained from a sphere by identifying antipodal points, and Lobačevskian space are also examples of Riemannian spaces.

One of the most important concepts of Riemannian geometry is the *curvature of a space*. The curvature is defined at every point in every two-dimensional direction passing through this point. To determine the curvature of a Riemannian space at a point one considers a geodesic triangle, that is, a curvilinear triangle bounded by arcs of three geodesics, one of whose vertices is at the given point and two of whose sides issuing from this vertex are tangent to the given two-dimensional direction. For this triangle ABC one computes the angular excess $A + B + C - \pi$ (as defined, the angular excess can be positive, zero, or negative), as well as the area, and forms the ratio of the angular excess over the area of the triangle. Now one lets the triangle shrink to the given vertex but insists that the sides issuing from it continue to be tangent to the given two-dimensional direction. It is obvious that in this limiting process the area of the geodesic triangle tends to zero. Since the geometry of the two-dimensional surface in which the passage to the limit is taking place is locally Euclidean and the angular excess of a Euclidean plane is zero, it follows that the angular excess of the geodesic triangle also tends to zero. But the ratio of the angular excess over the area of the triangle tends to a definite limit called the *curvature of the Riemannian space* at the given point in the given two-dimensional direction.

The curvature of n-dimensional Euclidean space, regarded as a special case of a Riemannian space, is equal to zero at all of its points; that is why a Euclidean space is called Riemannian space of *zero curvature* and Riemannian spaces different from Euclidean space are called *curved spaces*.

Riemann defined the curvature of his space in the following manner; in a neighborhood of the point at which he wanted to define curvature he introduced coordinates x_i equal to zero at the point and such that distance s of points in the neighborhood of this point was given in terms of the coordinates by the relation $s^2 = \sum_i x_i^2$. Riemann writes:

> When we introduce these quantities, the square of the line-element is Σdx^2 for infinitesimal values of the x, but the term of next order in it is equal to a homogeneous function of the second order of the $\frac{1}{2}n(n - 1)$ quantities $(x_1 dx_2 - x_2 dx_1)$, $(x_1 dx_3 - x_3 dx_1)$ ... an infinitesimal, there-

> fore, of the fourth order; so that we obtain a finite quantity on dividing this by the square of the infinitesimal triangle, whose vertices are $(0,0,0,\dots)$, $(x_1,x_2,x_3,\dots)$, $(dx_1,dx_2,dx_3,\dots)$. This quantity retains the same value so long as the x and the dx are included in the same binary linear form, or so long as the two geodesics from 0 to x and from 0 to dx remain in the same surface-element; it depends therefore only on place and direction. It is obviously zero when the manifold represented is flat, i.e., when the squared line-element is reducible to $\Sigma\, dr^2$, and may therefore be regarded as the measure of the deviation of the manifold from flatness at the given point in the given surface-direction. Multiplied by $-\frac{3}{4}$ it becomes equal to the quantity which Privy Councillor Gauss has called the total curvature of a surface *[122, pp. 62–63]*.

In view of Riemann's choice of coordinates, the expansion of the square of the line element ds^2 in the neighborhood of the point he considered is given by

$$ds^2 = \sum_i dx_i^2 + \sum_i \sum_j \sum_k c_{ij,k} x_k dx_i dx_j + \sum_i \sum_j \sum_k \sum_l c_{ij,kl} x_k x_l dx_i dx_j + \dots,$$

where $c_{ij,k}$ and $2c_{ij,kl}$ are the values of the derivatives $\dfrac{\partial g_{ij}}{\partial x_k}$ and $\dfrac{\partial^2 g_{ij}}{\partial x_k \partial x_l}$ at the point under consideration. In the first place, Riemann notes that there is no linear term in this expansion, that is, that $c_{ij,k} = 0$. The reason for this is that his coordinate lines are geodesics satisfying equation (8.3), where the coefficients Γ^i_{jk} are expressed in terms of g_{ij} by formulas (8.4). Then Riemann claims that the second-order terms in the expansion of ds^2 form a quadratic form in the $x_i dx_j - x_j dx_i$. If for the sake of uniformity we denote the infinitesimals x_i by δx_i then the latter magnitudes can be written as $\Delta x_{ij} = \delta x_i dx_j - \delta x_j dx_i$. These quantities determine a parallelogram on the vectors $\{dx_i\}$ and $\{\delta x_i\}$, so that the quadratic terms in the expansion can be written as

$$\Delta\sigma^2 = \sum_i \sum_j \sum_k \sum_l R_{ij,kl} \Delta x_{ij} \Delta x_{kl}.$$

The quantity obtained by Riemann, which he calls the fraction from the division of $\Delta\sigma^2$ by the area of the triangle on the vectors $\{dx_i\}$ and $\{\delta x_i\}$, is actually the limit of this fraction as the triangle in question shrinks to the point. Nowadays, a quantity proportional to that defined by Riemann (with a proportionality factor such that in the case of a surface it coincides with the Gaussian curvature) is called the *Riemannian curvature* of the space at a given point in a given two-dimensional direction. Riemann goes on to give our earlier geometric interpretation of the curvature he defined. He notes that on surfaces in ordinary space "the difference between the angle sum of an infinitesimal triangle and two right angles is proportional to the area of the triangle" and writes:

> To give an intelligible meaning to the curvature of an *n*-fold manifold at a given point and in a given surface-direction through it, we must start from the fact that a geodesic proceeding from a point is entirely determined when its initial direction is given. According to this we obtain a determinate surface if we prolong all the geodesics proceeding from the given point and lying initially in the given surface-direction; this surface has at the given point a definite curvature, which is also the curvature of the *n*-fold continuum at the given point in the given surface-direction *[122, p. 64]*.

Riemannian geometry was applied by Riemann to the theory of differential equations in *A mathematical work containing an attempt to answer the question proposed by the most illustrious Paris Academy: To determine the heat state of a homogeneous solid so that a system of isothermal curves given at a certain moment in time remains a system of isothermal curves at an arbitrary moment in time and so that the temperature at a point is expressed as a function of time and two more independent variables* (Commentatio mathematica qua respondere tentatur questioni ab Ill$^{\text{ma}}$ Academia Parisensi proposita: Trouver quel doit être l'état calorifique d'un corps solide homogène pour qu'une système de courbes isothermes, à un instant donné, restent isothermes après un temps quelconque, de telle sorte que la température d'un point puisse s'exprimer en fonction du temps et de deux autres variables indépendentes), written in 1861 and published in his collected works (Leipzig, 1876) *[454, pp. 391–404]*. In this paper Riemann solved the problem of reducing the differential equation of heat conduction

$$\sum_i \frac{\partial}{\partial s_i}\left(\sum_j b_{ij}\frac{\partial u}{\partial s_j}\right) = h\frac{\partial u}{\partial t}$$

to simplest form— a problem equivalent to that of transforming the quadratic form $\sum_i \sum_j b_{ij} ds_i ds_j$ to a sum of squares. Riemann found that this can be done if and only if the expression

$$K = \frac{1}{2}\frac{\sum\limits_i \sum\limits_j \sum\limits_k \sum\limits_l (ij, kl)(ds_i\delta s_j - ds_j\delta s_i)(ds_k\delta s_l - ds_l\delta s_k)}{\left(\sum\limits_i \sum\limits_j b_{ij} ds_i ds_j\right)\left(\sum\limits_i \sum\limits_j b_{ij}\delta s_i\delta s_j\right) - \left(\sum\limits_i \sum\limits_j b_{ij} ds_i\delta s_j\right)^2},$$

that is invariant under a change of variables, vanishes. Of this expression, he denotes by (111), Riemann says:

> The expression $\sqrt{\sum_i \sum_j b_{ij} ds_i ds_j}$ can be regarded as a line element of an *n*-tuply extended space that is outside the bounds of our intuition. If we lead from point $(s_1, s_2, \ldots, s_n)$ in this space all possible geodesics whose initial directions are characterized by the ratios $\alpha ds_1 + \beta\delta s_1 : \alpha ds_2 + \beta\delta s_2 : \ldots : \alpha ds_n + \beta\delta s_n$ (where α and β are arbitrary quantities) then these lines form a certain surface which we can think of as located in the

> ordinary space of our intuition. Then expression (111) is the curvature of this surface at the point $(s_1, s_2, \ldots, s_n)$ *[454, p. 403]*.

The quantities (ij, kl), the so-called four-index Riemann symbols, are the same quantities that we denoted earlier by $R_{ij,kl}$. We note that the quantities $R_{ij,kl}$ can be expressed in terms of the coefficients g_{ij} and their derivatives, using the quantities Γ^i_{jk} (expressed in terms of the g_{ij} and their derivatives by formulas (8.4)), by means of the formula

$$R_{ij,kl} = \left(\frac{\partial \Gamma^h_{jk}}{\partial x_i} - \frac{\partial \Gamma^h_{ik}}{\partial x_j} + \Gamma^h_{ir}\Gamma^r_{jk} - \Gamma^h_{jr}\Gamma^r_{ik}\right) g_{hl}.$$

In the paper mentioned previously Riemann also used the quantities Γ^i_{jk} but denoted them as p_{ijk}.

Riemann's investigations were continued by Elwin Bruno Christoffel (1829–1900) in the paper *On the transformation of homogeneous differential expression of second degree* (Über die Transformation der homogenen Differentialausdrücke zweiten Grades, Berlin, 1869) *[114, vol. 1, pp. 368–377]*, in which he posed the question of the conditions under which the geometry determined by a form $\sum_i \sum_j g_{ij} dx_i dx_j$ coincides with the geometry determined by a form $\sum_i \sum_j h_{ij} dy_i dy_j$. Just as Riemann's intention was to develop Gauss's theory of surfaces, so Christoffel's intention was to generalize the problem of superposition of surfaces. Christoffel's necessary condition for the coincidence of geometries turned out to be the coincidence of the differential forms $\sum_i \sum_j \sum_k \sum_l R_{ij,kl} dx_i dx_k \delta x_j \delta x_l$, computed for the two given forms. Christoffel denoted the quantities Γ^i_{jk} by $\left\{ {jk \atop i} \right\}$ and the quantities $\sum_l g_{il} \Gamma^l_{jk}$ by $\left[{jk \atop i} \right]$. That is why these quantities are often called *Christoffel symbols of the first and second kind*, respectively.

Riemannian Spaces of Constant Curvature

After defining Riemannian spaces of variable curvature and noting that Euclidean spaces are spaces of zero curvature Riemann dwelled at length on spaces of constant nonzero curvature:

> Manifolds whose curvature is constantly zero may be treated as a special case of those whose curvature is constant. The common character of these continua whose curvature is constant may be also expressed thus, that figures may be moved in them without stretching. For clearly figures could not be arbitrarily shifted and turned round in them if the curvature at each point were not the same in all directions. On the other hand, however, the measure-relations of the manifold are entirely determined by the curvature; they are therefore exactly the same in all directions at one point as at another, and consequently the same constructions can be made from it: whence it follows that in aggregates with

> constant curvature figures may have any arbitrary position given them. The measure-relations of these manifolds depend only on the value of the curvature, and in relation to the analytic expression it may be remarked that if this value is denoted by α, the expression for the line-element may be written
>
> $$\frac{1}{1+\frac{1}{4}\alpha\sum x^2}\sqrt{\sum dx^2}$$
>
> *[122, p. 65].*

The latter expression generalizes the previously mentioned expression, obtained by Minding, of the line element of a surface of constant curvature in terms of its Gaussian curvature.

The simplest example of a Riemannian space of constant positive curvature is a sphere in $(n + 1)$-dimensional Euclidean space; the Riemannian curvature of a sphere of radius r, of any dimension, in all two-dimensional directions is equal to $1/r^2$, for the area of any spherical triangle is the product of r^2 and the angular excess of the triangle.

An example of an n-dimensional Riemannian space of constant negative curvature is n-dimensional Lobačevskian space, first defined by Beltrami in the paper *Fundamental theory of spaces of constant curvature* (Teoria fondamentale degli spazî di curvatura costante. Milan, 1869) *[42, pp. 406–429]* published in the same year as his previously discussed paper. This space can also be viewed as one of the sheets of a sphere of radius qi in $(n + 1)$-dimensional pseudo-Euclidean space. The Riemannian curvature of such a sphere in all two-dimensional directions is $-1/q^2$ since the area of an arbitrary spherical triangle on this sphere is the product of its angular defect by $-q^2$.

We wish to mention here a remarkable theorem of F. Schur established in the paper *On the connection between spaces of constant curvature and projective spaces* (Über dem Zusammenhang der Räume constanten Krümmungmasses mit den projectiven Räumen. Leipzig, 1886) *[517]* which asserts that, on a Riemannian manifold, the constancy of the Riemannian curvature in all two-dimensional directions at each point implies the constancy of the Riemannian curvature at all points. Schur's proof is purely geometric and is based on a projective mapping of the bundle of linear elements at one point onto a similar bundle at another point. An analytic proof of this result was given by Luigi Bianchi (1856–1928) in the paper *On four-index symbols and on Riemannian curvature* (Sui simboli a quatro indici e sulla curvatura di Riemann. Rome, 1902) *[56]*.

Elliptic Geometry

Another form of n-dimensional Riemannian space of constant positive curvature is a sphere in $(n + 1)$-dimensional Euclidean space with identified antipodal points called n-dimensional *elliptic space* or *non-Euclidean space of Riemann.* When $(n + 1)$-dimensional Euclidean space is completed to $(n + 1)$-dimensional projective space, pairs of antipodal points of a sphere in $(n + 1)$-dimensional Euclidean space are projected from its center onto the points of its plane at infinity. Since this plane is an n-dimensional projective space, elliptic space can be viewed as a metrized projective space. The details are as follows.

An arbitrary sphere of $(n + 1)$-dimensional Euclidean space is given by an equation of the form

$$A \sum_i x_i^2 + 2 \sum_i b_i x_i + c = 0,$$

which can be written in homogeneous coordinates as

$$A \sum_i x_i^2 + 2 \sum_i b_i x_i x_0 + c(x_0)^2 = 0.$$

The latter equation shows that our sphere intersects the plane at infinity, $x_0 = 0$, in the imaginary quadric $\sum_i x_i^2 = 0$. Hence an elliptic space can be described as a projective space with a given imaginary conic. The distance ω between two points X and Y of an elliptic space of curvature $1/r^2$ is connected with the angle φ between the corresponding diameters of the sphere by the relation $\omega = \varphi r$. On the other hand, Edmond Laguerre (1834–1886) showed in the paper *On the theory of foci* (Sur la theorie des foyers. Paris, 1853) *[298, vol. 2, pp. 6–15]* that the angle φ between two straight lines in Euclidean space can be expressed in terms of the cross ratio (ij, xy) of the straight lines x, y and two *isotropic straight lines* i, j, that is, imaginary lines of zero length that lie in the plane of x and y and pass through the point of intersection of x and y, by means of the relation

$$\varphi = \frac{1}{2i} \ln(ij, xy). \tag{8.5}$$

Since the imaginary lines i, j that join the center of the sphere to the points I, J of the imaginary quadric $\sum_i x_i^2 = 0$ in the plane at infinity are, in view of the equation of the quadric, isotropic, Laguerre's formula (8.5) yields an expression for the distance ω between points X, Y of elliptic space in terms of the cross ratio (IJ, XY) of these points and the two points of intersection of the straight line XY and the imaginary quadric:

$$\omega = \frac{r}{2i} \ln(IJ, XY). \tag{8.6}$$

An elliptic metric in the projective plane was defined in 1859 by Cayley in the *Sixth memoir upon quantics [103, vol. 2, pp. 561–592]*. An elliptic metric in space, and the term *elliptic geometry*, were introduced by Klein in 1871 in the paper *On the so-called non-Euclidean geometry [282, vol. 1, pp. 254–305]*, in which he wrote that

> The basis of a general projective metric in space is provided by an arbitrary *fundamental surface of the second order.*
>
> To define the distance between two points one joins them by a straight line. It intersects the fundamental surface in two new points that are in a definite cross ratio with the two given points. *The logarithm of this cross ratio multiplied by an arbitrary constant c yields what one should call the distance between the two given points [282, vol. 1, p. 300].*

Klein gave a similar definition of the angle between two planes (see p. 238). Further Klein wrote:

> By a *motion* is meant the totality of linear transformations that leave the fundamental surface invariant.
>
> By *spheres* one means quadric surfaces that meet the fundamental surface along a plane curve. The center of the sphere is the pole of the plane that contains the osculating curve....
>
> If the fundamental surface is *imaginary* then all straight lines have finite length and each pencil of planes has finite angle sum. *Elliptic* geometry comes under this case (provided that the constant c' in the definition of angles is taken as $\sqrt{-1}/2$, so that the sum of the angles in a pencil of planes is π).
>
> We do not investigate the case when the fundamental surface is *real* and *ruled* (like a hyperboloid of one sheet) for this case is no way connected with the three geometries (elliptic, hyperbolic, parabolic) considered here.
>
> Finally, if the fundamental surface is *real* and *nonruled* then for the interior points of the surface we obtain a metric that includes the metric of hyperbolic geometry provided that we again take the constant c' to be $\sqrt{-1}/2$.
>
> *Parabolic* geometry is included as a special case of the general metric; this case arises if the fundamental surface specializes (degenerates) into an imaginary conic section. The fundamental conic section of parabolic geometry is the so-called imaginary circle at infinity *[282, vol. 1, p. 301].*

Klein's *fundamental surface* is now called by Cayley's term *the absolute*. We note that Klein's *linear transformations* are collineations, his *hyperbolic geometry* is Lobačevskian geometry, and his *parabolic geometry* is Euclidean geometry; the *imaginary circle at infinity* is the imaginary circle $\sum_i x_i^2 = 0$ in the plane at infinity of Euclidean space that plays the role of the absolute of this space. Klein did not investigate the geometry of the exterior of an oval (nonruled) quadric and the geometry of a ruled quadric. The reason for this is that, unlike elliptic geometry and Lobačevskian geometry, the geometry of these spaces is locally not Euclidean but pseudo-Euclidean.

We also note that the only difference between formula (8.6) and formula (6.27) for Lobačevskian space is that in one case we have the pure imaginary constant r/i and in the other the real constant q.

Clifford Parallels and Surfaces

The geometry of elliptic space was significantly developed by William Kingdon Clifford (1845–1879) in his *Preliminary sketch of biquaternions* (London, 1879) *[122, pp. 181–200]*.

After defining elliptic, hyperbolic, and parabolic geometries whose absolutes are, respectively, an imaginary quadric, a real (oval) quadric, and an imaginary conic in a real plane, Clifford investigates *elliptic geometry*. After defining poles and polar planes and mutually polar straight lines with respect to an absolute and pointing out that two points that are polar conjugates with respect to the absolute "are a *quadrant* apart," that is, are $(\pi/2)r$ apart, or, if $r = 1$, are $\pi/2$ apart, Clifford notes that

> Through an arbitrary point can in general be drawn *one* line perpendicular to a given plane; namely, the line joining the point to the pole of the plane. If, however, the point *is* the pole of the plane, every line through it is perpendicular to the plane. Similarly, from a point not on the polar of a given line can be drawn one and only one perpendicular to the line; namely, the line through the point which meets the given line and its polar *[122, p. 192]*.

and further:

> *In general, two lines can be drawn so that each meets two given lines at right angles, and these are polars of one another*. One line may therefore be converted into another by rotation about two polar axes. These axes are determined as the lines which meet the two given lines and their polars. If we travel continuously along one of these lines and draw perpendiculars on the other, one of these axes determines the shortest distance between the lines, and the other the longest. If then these two are equal, the lines are equidistant along their whole length. Thus *there is a case of exception in which two lines and their polars belong to the same set of generators of a hyperboloid; the lines are then equidistant along their whole length, and meet the same two generators of one system of the absolute*. I shall use the word *parallel* to denote two lines so situated; and they shall be called *right* parallel or *left* parallel according as one is converted into the other by a right-handed or left-handed twist. Through an arbitrary point can be drawn one right parallel and one left parallel to a given line; the angle between them is twice the distance of the point from the line. There are many points of analogy between the *parallels* here defined and those of parabolic geometry. Thus, if a line meets two parallel lines, it makes equal angles with them; and a series of parallel lines meeting a given line constitute a ruled surface of zero curvature. The geometry of this surface is the same as that of a finite parallelogram whose opposite sides are regarded as identical *[122, pp. 192–193]*.

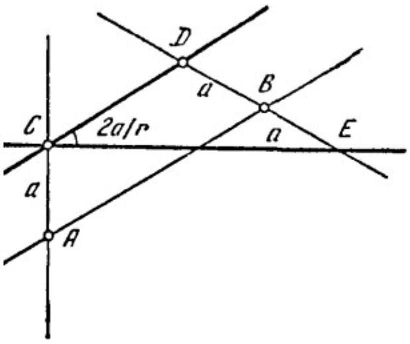

Figure 103

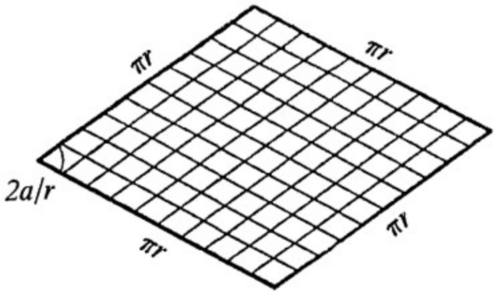

Figure 104

The parallels defined by Clifford are now called *paratactic straight lines*;[5] one must bear in mind that, unlike parallels in Euclidean (*parabolic*) geometry, Clifford parallels are skew straight lines. Figure 103 illustrates the construction of a right, CD, and left, CE, Clifford parallel through a point C to a straight line AB. From the point C one drops the perpendicular CA to the straight line AB and constructs its polar DBE. On the latter one lays off on either side of the point B segments BD and BE equal to the segment $CA = a$. The required parallels are the straight lines CD and CE; the angle DCE is measured in terms of the segment $DE = 2a$ and is equal to $2a/r$. The surface constructed by Clifford is now known as a *Clifford surface*. It is a ruled quadric obtained by rotating one of two paratactic straight lines about the other; its rectilinear generators in both families are paratactic to the axes; at the same time it is also a surface of revolution about the polar of the first axis. The geometry of a Clifford surface is Euclidean. Its area is finite and equal to $\pi^2 r^2 \sin 2a/r$, for it is isometric to a rhombus with sides πr and acute angle $2a/r$ (Figure 104) with sides in each pair of opposite sides glued together. To prove that the geometry of this surface is Euclidean it suffices to write down its parametric equations in homogeneous coordinates whose basis points are located on the axes of the surface,

$$x_0 = \cos\frac{a}{r}\cos u, \qquad x_1 = \cos\frac{a}{r}\sin u,$$

$$x_2 = \sin\frac{a}{r}\cos v, \qquad x_3 = \sin\frac{a}{r}\sin v,$$

and to compute the line element

$$ds^2 = \sum_i dx_i^2 = r^2\left(\cos^2\frac{a}{r}\,du^2 + \sin^2\frac{a}{r}\,dv^2\right).$$

---

[5] This term was introduced by E. Study in the paper *[568]*.

If in the latter we put

$$U = r\cos\frac{a}{r}\cdot u, \qquad V = r\sin\frac{a}{r}\cdot v,$$

then we have $ds^2 = dU^2 + dV^2$.

If one eliminates from the parametric equations of the Clifford surface the parameters u and v then one obtains its equation in the form

$$\sin^2\frac{a}{r}(x_0^2 + x_1^2) - \cos^2\frac{a}{r}(x_2^2 + x_3^2) = 0.$$

A Clifford surface is the simplest solution of the Clifford-Klein problem of finding spaces with a Euclidean metric that are not isometric to Euclidean space in the large.

Clifford's Idea of the Geometrization of Physics

W. K. Clifford was also interested in questions of the philosophy of space. In his philosophical work *Philosophy of pure science* (London, 1873) there are his famous words on Lobačevskian geometry!

> What Vesalius was to Galen, what Copernicus was to Ptolemy, that was Lobatchewsky to Euclid. There is, indeed, a somewhat instructive parallel between the last two cases. Copernicus and Lobatchewsky were both of Slavic origin. Each of them has brought about a revolution in scientific ideas so great that it can only be compared with that wrought by the other. And the reason of the transcendent importance of these two changes is that they are changes in the conception of the Cosmos *[123, vol. 1, p. 356]*.

What Clifford is saying here is that Lobačevskiĭ's discovery that Euclidean geometry is not the only conceivable geometry is just as revolutionary as Vesalius' shattering of the myth of the exclusive position of man in the animal kingdom and Copernicus' discovery that the Earth is just one of the planets.

Clifford expressed his interesting thoughts on space in his posthumously published *The Common sense of the exact sciences* (London, 1885) *[121]*. Clifford begins the chapter on space with the words

> Geometry is a physical science *[121, p. 43]*,

and at the end of the chapter on position he writes:

> We may conceive our space to have everywhere a nearly uniform curvature, but that slight variations of the curvature may occur from point to point, and themselves vary with the time. These variations of the curvature with the time may produce effects which we not unnaturally attribute to physical causes independent of the geometry of our space.

> We might even go so far as to assign to this variation of the curvature of space "what really happens" in that phenomenon which we term the motion of matter *[121, pp. 202–203]*.

Clifford developed these ideas in greater detail in his report *On the space theory of matter* (Cambridge, 1876). First he paraphrases Riemann:

> Riemann has shewn that as there are different kinds of lines and surfaces, so there are different kinds of space of three dimensions; and that we can only find out by experience to which of these kinds the space in which we live belongs. In particular, the axioms of plane geometry are true within the limits of experiment on the surface of a sheet of paper, and yet we know that the sheet is really covered with a number of small ridges and furrows, upon which (the total curvature not being zero) these axioms are not true. Similarly, he says although the axioms of solid geometry are true within the limits of experiment for finite portions of our space, yet we have no reason to conclude that they are true for very small portions; and if any help can be got thereby for the explanation of physical phenomena, we may have reason to conclude that they are not true for very small portions of space *[122, p. 21]*.

Then he states the following four principles:

> (1) That small portions of space *are* in fact of a nature analogous to little hills on a surface which is on the average flat; namely, that the ordinary laws of geometry are not valid in them.
> (2) That this property of being curved or distorted is continually being passed on from one portion of space to another after the manner of a wave.
> (3) That this variation of the curvature of space is what really happens in that phenomenon which we call the *motion of matter*, whether ponderable or etherial.
> (4) That in the physical world nothing else takes place but this variation, subject (possibly) to the law of continuity *[122, pp. 21–22]*

These principles form a program of geometrization of physics.

Riemann's Topology

In addition to the theory of Riemannian spaces Riemann also founded topology—one of the most important disciplines that significantly broadened our notions of space.

We have already pointed out that Euler interpreted Leibniz's term *geometry of position* in a topological sense, whereas Carnot and Grassmann gave it narrower interpretations. Euler's interpretation of the term was developed by the German physicist Johann Benedict Listing (1808–1862), who suggested the now generally accepted term *topology* in his *Preliminary studies on topology* (Vorstudien zur Topologie. Göttingen, 1847) *[329]*. Listing's term comes from replacing Leibniz's Latin *situs* (place) with its Greek counterpart *topos*. Listing studied

> *linear complections, that is, lines or curves or sets of such, located on a surface, such as a plane or a sphere, or arbitrarily disposed in space [329, p. 867].*

He investigated knots, chains, plaits, and other forms of mutual disposition of linear complections - later he called them *linear complexes*—with numerous examples drawn from biology and technology. Listing's work *The census of spatial complexes or a generalization of Euler's therorem on polyhedra* Der Census räumlicher Komplexe oder Verallgemeinerung des Euler'schen Satzes von den Polyedern. Göttingen, 1862) *[330]* dealt with more general *spatial complexes*.

Gauss interpreted Leibniz's ideas in the same topological sense. Gauss devoted to topological investigations, as to non-Euclidean geometry, a number of rough notes and letters (see *[196, vol. 8, pp. 407–410]* and the study of Jean Claude Pont *[435a, pp. 31–38]*). When he founded the topology of two-dimensional manifolds and laid the foundations of multidimensional manifolds Riemann also interpreted Leibniz's ideas in the same way. Riemann laid the foundations of two-dimensional topology in his *Theory of Abelian functions* (1857) *[454, pp. 82–142]*, mentioned earlier. In this work he wrote, with reference to Leibniz, that

> For the study of functions which arise as integrals of exact differentials, some theorems belonging to *analysis situs* are nearly indispensable *[454, p. 91]*.

Riemann associates to algebraic functions $f(x, y)$ in the complex variables x and y multisheeted surfaces now called *Riemann surfaces*. He divides these surfaces into

> simply connected ones, in which every closed curve bounds a region of the surface - as, for example, a disk—and multiply connected ones, for which this does not happen—as, for example, an annulus bounded by two concentric circles *[454, p. 22]*.

and points out that by means of a system of cuts it is possible to make a multiply connected surface into a simply connected one. Riemann introduces the characteristic p of a plane algebraic curve (later called by Clebsch *[119]* the genus of the curve) which he defines as half the number of cuts needed to make the corresponding multiply connected Riemann surface into a simply connected one. The number p is now called the *genus of the surface*. We note that, as S. L'Huillier showed *[323]*, for polyhedra the genus p is connected with the Euler characteristic $\chi = N_0 - N_1 + N_2$ of a polyhedron by means of the relation $\chi = 2 - 2p$. For a sphere $p = 0$, for a torus $p = 1$, and for a sphere with p handles it is equal to p. Riemann showed that for a Riemann surface the genus p, the number n of its sheets, and the number w of its branch points are connected by the relation $w - 2n = 2p - 2$.

In a survey that is a supplement to the article *On the hypotheses that lie at the foundations of geometry* Riemann adds a remark in connection with the first chapter of the article *Notion of an n-ply extended magnitude*:

> Chapter 1 is at once an introduction to investigations on *analysis situs* *[454, p. 286]*.

Fragments dealing with *analysis situs* were published in 1876 in an edition of Riemann's collected works *[454, pp. 479–482]*. In these fragments Riemann generalized the topological properties of a two-dimensional surface to an n-dimensional manifold, here called an *n-stretch* (*n-Streck*). Here Riemann defines what are now called *homologous* n-stretches:

> an n-stretch A is said to be transformable into an n-stretch B if A and parts of B together form the complete boundary of an interior $(n + 1)$-stretch *[454, p. 479]*.

Then he gives the extremely important definition:

> If in the interior of a continuously extended manifold it is possible to make every unbounded n-stretch bounding by means of m definite parts of nonbounding n-stretches then this manifold has $(m + 1)$-tuple connection of the n-th dimension.
>
> A continuously extended manifold is called *simply connected* if the connection of every dimension is simple *[454, p. 479]*.

It is easy to verify that for a two-dimensional manifold of genus p Riemann's *connection* is $(2p + 1)$-ple. Then Riemann explains the dependence of the connection of the boundary of a manifold on the connection of the manifold itself.

These ideas of Riemann were set forth by his friend Enrico Betti in the previously mentioned paper *On spaces of an arbitrary number of dimensions [55, vol. 2, pp. 273–290]*. By "spaces" Betti meant manifolds in multidimensional Euclidean spaces. He wrote:

> If in an n-dimensional space R, bounded by one or more $(n - 1)$-dimensional spaces, every closed m-dimensional space, $m < n$, is the boundary of a part of a linear connected $(m + 1)$-dimensional space, entirely contained in R, then we shall have a connection in $m + 1$ dimensions and we shall say that R has a *simple* connection of the m-th species. If a space R has only simple connections, then we shall say that it is simply connected. If, however, one can imagine in R a number p_m of closed m-dimensional spaces which cannot form the boundary of a linear connected part of an $(m + 1)$-dimensional space, entirely contained in R, and such that every other closed m-dimensional space forms

> by itself or with some or all of these spaces the boundary of a linear connected part of an $(m + 1)$-dimensional space, entirely contained in R, then we shall say that R has a connection of $(p_m + 1)$-th order of the m-th species *[55, vol. 2, p. 278]*.

These ideas were further developed by Poincaré. Already in the memoir *On curves defined by differential equations* (Sur les courbes defines par les équations differentielles. Paris, 1881–1885) *[431, vol. 1, pp. 90–161]* he made extensive use of topological properties of curves to give qualitative descriptions of solutions of differential equations. Poincaré devoted to the topology of multidimensional manifolds the large memoir *Analysis situs* (the title is taken over from Riemann) (Paris, 1895) *[431, vol. 6, pp. 193–288]* and five supplements to it (Palermo-London-Paris, 1899 1904).

Poincaré determined $(n - p)$-dimensional manifold in n-dimensional space by means of p equations and q inequalities between the n coordinates. Then he defined the boundary of a manifold, homomorphism between manifolds, and homology—the fundamental concept of combinatorial topology. Using this concept he defined the *Betti numbers* of a manifold, that is, Betti's *orders of connection* (in the twenties of this century Solomon Lefschetz (1884–1972) and James Alexander (1888–1971) proposed to call *Betti numbers* numbers b_i one less than Poincaré's *Betti numbers*; cf. *[308]*). Poincaré extended Euler's theorem to polyhedra with arbitrary Betti numbers by showing that

$$N_0 - N_1 + N_2 - \cdots + (-1)^n N_{n-1} = b_0 - b_1 + b_2 - \cdots - (-1)^n b_{n-1}.$$

The polynomial $\sum b_i t^i$ is called the *Poincaré polynomial.* In the same paper Poincaré also laid the foundation for the homotopic theory of manifolds in which a leading role is played by the noncommutative *Poincaré group* introduced by him.

Topological Spaces

At the beginning of the 20th century, in connection with the spread of the group-theoretic viewpoint, mathematicians began to study *abstract spaces* together with the study of topological invariants of manifolds that are submanifolds of Euclidean spaces or are obtained from such manifolds by identification of points (as in the case of obtaining the projective plane from a sphere) or by splitting them into points of various sheets (as in the case of constructing the Riemann surface of an algebraic function out of its domain of definition in the complex plane). One of the first works in this direction was the paper *On certain points of the functional calculus* (Sur quelques points du calcul fonctionelles. Palermo, 1906) *[184]* of the French mathematician Maurice Fréchet (1878–1973). In this paper Fréchet defines for the first time an *abstract metric space* exemplified by a space of functions or curves. To construct an abstract space Fréchet found it necessary

> to generalize, first of all, the theory of linear sets that has brought about so much progress in the theory of functions of a single variable,

by which theory he meant the, by that time, extensively developed theory of sets of points of the real line. Fréchet goes on to say:

> If we assume this preliminary investigation of sets then there arises a difficulty. The first generalization that seems natural is that of the notion of a continuous function. But if one wishes to consider operations where the variable is an element of arbitrary nature, then we must first know what is to be meant by neighboring elements or by the limit of a sequence of elements. This seems impossible: usually one gives a special definition of limit for each category of elements under study— points, curves, and so on. I circumvented this difficulty by a method similar to that which allows one to reason in the theory of abstract groups about a composition of explicitly indeterminate form.
>
> Now I note that almost all (but not all) classical definitions of limit can be formulated as follows: in the given category of elements one can associate to each pair of elements a number $\rho(A, B)$ whose properties are very close to the properties of distance of two points, namely that A coincides with B if $\rho(A, B) = 0$ and A tends to B if $\rho(A, B)$ tends to zero. If we accept this hypothesis, less general but nevertheless very broad, then we obtain numerous more specific results.
>
> The approach just outlined leads to the generalizing of almost all theorems about linear sets and about continuous functions (at least those that can be formulated independently of the nature of the investigated sets) *[184, pp. 1–2]*.

When mentioning abstract groups Fréchet emphasizes that he is applying the same methods of creating abstract mathematical concepts that have hitherto been used only in algebra, where, together with the abstract theory of groups, founded in the seventies of the 19th century, there also appeared Dedekind's abstract ring theory.

Developing his program, Fréchet defines a general metric space, which he calls *class* (V), as

> a set of elements of arbitrary nature such that we know how to determine whether two given elements are identical or not and, in addition, such that to any two of them, A, B, we can associate a number $(A, B) = (B, A)$ with the following two properties:
>
> 1′ The necessary and sufficient condition that (A, B) is zero is the identity of A and B.
>
> 2′ There exists a completely determined positive function $f(\varepsilon)$ such that the inequalities $(A, B) \leq \varepsilon$ and $(B, C) \leq \varepsilon$ imply $(A, C) \leq f(\varepsilon)$ for all elements A, B, C. In other words, for (A, C) to be small it suffices that (A, B) and (B, C) are small *[184, p. 18]*.

At present, metric spaces are defined by means of three axioms: the *identity axiom* that coincides with Frechet's axiom 1′; the *symmetry axiom* $(A, B) = (B, A)$ that Frechet regards as part of the definition of distance; and the *triangle inequality axiom* $(A, B) + (B, C) \geq (A, C)$ that replaces Fréchet's axiom 2′.

Fréchet goes on to consider sets in metric spaces, limits of sequences, and limit points of sets. All limit points of a set form its *derived set*. Sets that contain all their limit points are called *closed*, and complements of closed sets with respect to the whole space are called *open* sets. The intersection of all closed sets containing a given set M is called the *closure* $\bar{M}$ of the set M. Then, after defining the Cauchy criterion for a sequence of elements $A_1, A_2, \ldots$ of the metric space as the possibility of associating to every $\varepsilon > 0$ an integer n with the property that the inequality $(A_n, A_{n+p}) > \varepsilon$ is satisfied for every p, Fréchet defines a *complete* metric space as a metric space in which every sequence satisfying the Cauchy criterion has a, necessarily unique, limit.

Fréchet restricts himself to spaces that can be regarded in at least one way as the derived sets of countable sets of their elements. He considers metric spaces of functions with different definitions of distance.

The definition of limit points and closed sets made it possible to introduce continuous mappings and homeomorphisms of the metric spaces considered by Fréchet. Fréchet's paper became the starting point for the development of a general theory of abstract spaces and for the development of the basic concepts of functional analysis.

Since distance is not needed for the study of topological properties of abstract spaces, just as it is not needed for the study of topological properties of lines and surfaces, soon after the appearance of abstract spaces Felix Hausdorff (1868–1942) in his *Foundations of set theory* (Grundzüge der Mengenlehre. Leipzig, 1914) *[217; 218]* defined an *abstract topological space*. Hausdorff defined a topological space as any collection of elements, called *points*, with a distinguished collection of subsets $\{U\}$, such that to each point x there are associated some, and, at the very least, one of the sets of the system $\{U\}$, called *neighborhoods* $U(x)$ of x; the intersection of any two neighborhoods $U_1(x)$ and $U_2(x)$ contains a neighborhood $U_3(x)$ of this point, and if y is a point of a neighborhood $U(x)$ then there exists a neighborhood $U(y)$ of y contained in $U(x)$. Prescribing neighborhoods makes it possible to define limit points of any set M as points x such that every neighborhood of x contains at least one point of M other than x. Just as in Fréchet spaces so too in a topological (Hausdorff) space one defines derived, closed, and open sets. It is easy to see that all neighborhoods are open sets.

Every Fréchet metric space is a topological space. In fact, one can take as neighborhoods the sets of points whose distances from the points of a (fixed) countable set, whose derived set is the whole space, are less than some rational numbers. It is easy to see that the number of such neighborhoods is countable, or, as one says, the space has a *countable basis*.

A further development of the Hausdorff definition is the definition of a topological space proposed by the Polish mathematician Kazimierz Kuratowski (1896–1980) in the paper *The operation $\bar{A}$ of analysis situs* (L'operation $\bar{A}$ de l'analysis situs. Warsaw, 1922) *[296]*. By *operation $\bar{A}$* Kuratowski means the transition from a set A to its closure $\bar{A}$. According to Kuratowski, a topological space is a collection of elements of arbitrary nature, called *points*, on whose subsets A there is defined an operation of *closure* satisfying the following axioms:

1′ for the union $A + B$ of two sets A and B we have $\overline{A + B} = \bar{A} + \bar{B}$;
2′ a set A is contained in its closure, $A \subset \bar{A}$;
3′ the closure of the empty set $\varnothing$ coincides with it, $\overline{\varnothing} = \varnothing$;
4′ the closure of the closure coincides with the closure, $\bar{\bar{A}} = \bar{A}$.

If for every set there is defined its closure $\bar{A}$ then the points of $\bar{A}$ not in A are called the *limit points of A*; the sets A that coincide with their closures A are called the *closed sets*, and their complements relative to the whole space are called the *open sets*.

The Soviet mathematician Pavel Sergeevič Aleksandrov [Alexandroff] (1896–1983) proposed in the paper *On the foundation of* n*-dimensional topology* (Zur Begründung der *n*-dimensionalen Topologie. Leipzig, 1925) *[14]* a more symmetric form of Kuratowski's definition:

1. A topological space is a set of elements of arbitrary nature, called *points*, in which certain subsets, called *open sets*, have been singled out such that:
   1′ the whole space is an open set;
   2′ the null set is an open set;
   3′ the intersection of finitely many open sets is an open set;
   4′ every union of open sets is an open set.
2. A topological space is a set of elements of arbitrary nature, called points, in which certain subsets, called *closed sets*, have been singled out such that:
   1′ the whole space is a closed set;
   2′ the null set is a closed set;
   3′ every union of closed sets is a closed set;
   4′ the intersection of finitely many closed sets is a closed set.

In the first case a closed set is defined as the complement of an open set and in the second case an open set is defined as the complement of a closed set. Neighborhoods can be defined as subsystems of open sets such that an arbitrary open set can be represented as the union of sets of this subsystem (see *[218, p. 258]*).

To different systems of closed or open subsets in the same set of points there correspond different topologies on this set. For example, our axiom systems are satisfied if

(A) the only closed (and open) sets are the whole space and the null set;
(B) the closed (and open) sets are all sets of points of the space.

In case (A) the closures of all nonempty sets, including single points, coincide with the whole space. In case (B) the closure of each set coincides with this set, and a minimal system of neighborhoods is a system in which the neighborhood of each point is that point alone. Then the space has no limit points and is called *discrete*; the space of case (A) is called *trivial*.

Hausdorff ruled out case (A) by means of the following axiom: for any two points of the space there are disjoint neighborhoods *[218, p. 260]*. Sometimes Hausdorff's axiom is replaced with a weaker one due to Frigyes Riesz, that the closures of all single points of the space are the points themselves *[456, vol. 1, pp. 155 169]*—or by an even weaker axiom introduced by the Soviet mathematician Andreĭ Nikolaevič Kolmogorov (1903 1987), that any two points of the space have different closures *[15, p. 58]*.

A very important class of topological spaces is the class of *compact spaces* such that every covering of the space by means of open sets contains a finite subcovering. The Soviet mathematicians P. S. Aleksandrov and Pavel Samuilovič Uryson [Urysohn] (1898–1924) studied such spaces in the case when the minimal cardinality of a system of neighborhoods is greater than the cardinality of a countable set and called them *bicompact [16]*. For these spaces P. S. Aleksandrov developed a homology theory analogous to the homology theory of manifolds in Euclidean spaces. By now mathematicians have found a great many topological invariants of the most varied topological spaces.

One of the most obvious topological invariants is the dimension of a topological space. Whereas it is possible to have a one-to-one correspondence between manifolds of different dimensions, a one-to-one bicontinuous correspondence can only be established between manifolds of the same dimension. If the points of a topological space have neighborhoods homeomorphic to n-dimensional Euclidean space then it is natural to say that the dimension of the space is n. In more complicated cases dimension was defined by the Dutch mathematician Luitzen Egbertus Jan Brouwer (1882–1966) in the paper *Proof of the invariance of the dimension number* (Beweis der Invarianz der Dimensionzahl. Leipzig, 1911) *[80]*. The ideas of Brouwer were developed by Urysohn in *A memoir on Cantor manifolds* (Memuar o kantorovyh mnogoobraziyah. 1928) *[599]* and by the German mathematician Karl Menger (b. 1902) in *Dimension theory* (Dimensionstheorie. Leipzig, 1932) *[362]*. They defined dimension by induction on the dimension number, beginning with the dimension -1 which they assigned to the empty set.

Urysohn and Menger also proposed another definition of dimension, equivalent to the preceding for the most important topological spaces: a nonempty space is n-dimensional if every finite covering of the space has a finite subdivision of order $\leq n$ and there exists a finite covering of the space without finite subdivisions of order $< n$. A nonempty space is infinite-dimensional if it is not n-dimensional for any nonnegative n. The dimension of the null set is again taken to be -1.

In the previously mentioned paper Brouwer proved his famous *fixed-point* *theorem*: every continuous mapping of an n-dimensional simplex into itself has at least one fixed point. This theorem was first proved by the Latvian mathematician Piers Bohl (1865–1921), who worked in Riga, in the paper *On the motion of a mechanical system near an equilibrium position* (Über die Bewegung eines mechanischen Systems in der Nähe einer Gleichgewichtlage. Berlin, 1904) *[64, pp. 79–125]*. The Bohl-Brouwer theorem was extended to infinite-dimensional analogues of an n-dimensional simplex by the Soviet mathematician Andrei Nikolaevič Tihonov [Tychonoff] (b. 1906) *[597]*. Bohl came to the fixed-point theorem in connection with the problem of proving the existence of a solution of a system of differential equations connected with the notion of certain mechanical systems he investigated. The Bohl-Brouwer theorem was later used to prove the existence of solutions of finite systems of ordinary differential equations, and the Tihonov theorem was used to prove the existence of solutions of infinite systems of differential equations.

The Influence of the General Theory of Relativity

We have already pointed out the importance of Einstein's discovery of the special theory of relativity for the elaboration of the concept of pseudo-Euclidean geometry. For example, the appearance of Weyl's book *Space, time, matter [626]* in which, among other matters, he set forth his axiomatization of n-dimensional Euclidean space, was connected with the general theory of relativity. But the importance of the general theory of relativity for the development of geometry is far broader and deeper. Whereas in the special theory of relativity spacetime was viewed as a pseudo-Euclidean space, in the general theory it is viewed as an analogue of a Riemannian space that stands in the same relation to such a space as pseudo-Euclidean space to Euclidean space. Such a space is now called a *pseudo-Riemannian* or *general Riemannian space*. Just as in a Riemannian space, so too at each point of this space there is given the square of a line element

$$ds^2 = \sum_i \sum_j g_{ij} dx^i dx^j,$$

where the g_{ij} are point functions and, although the quadratic form is no longer positive definite, it can be reduced at each point to the form

$$ds^2 = -(dx^0)^2 + (dx^1)^2 + (dx^2)^2 + (dx^3)^2.$$

The first sketch of the general theory of relativity was set forth by Einstein together with the German mathematician Marcel Grossmann (1878–1936) in *Outline of a generalized theory of relativity and theory of gravitation* (Entwurf einer verallgemeinerten Relativitätstheorie und Theorie der Gravitation. Leipzig, 1913), and the final theory was presented in *Foundations of the general theory of relativity* (Grundlagen der allgemeinen Relativitätstheorie. Leipzig, 1916) *[340, pp. 109–164]*.

Already in the 1913 paper Einstein and Grossmann used the *tensor calculus* created by the Italian geometer Gregorio Ricci-Curbastro (1853–1925) in *Principles of a theory of differential quadratic forms* (Principii di una theoria delle forme differenziale quadratiche. Milan, 1884) *[453, vol. 1, pp. 138–171]* and subsequently developed together with Tullio Levi-Civita (1873–1941) in *Methods of the absolute differential calculus and their applications* (Méthodes du calcul différentiel absolu et leurs applications. Leipzig, 1901) *[453, vol. 2, pp. 185–271]*. Ricci gave the name *covariant system of the first order* to the functions a_i of the coordinates x^i that change as a result of a coordinate transformation $x^{i'} = x^{i'}(x^1, \ldots, x^n)$ in accordance with the rule

$$a_{i'} = \sum_i a_i \frac{\partial x^i}{\partial x^{i'}}, \tag{8.7}$$

exemplified by the law of transformation of partial derivatives $\frac{\partial \varphi}{\partial x^i}$, and the name *contravariant system of the first order* to functions a^i of the coordinates x^i that transform in accordance with the opposite rule

$$a^{i'} = \sum_i a^i \frac{\partial x^{i'}}{\partial x^i}, \tag{8.8}$$

exemplified by the law of transformation of the differentials dx^i of the coordinates. Ricci considered scalars as systems of zero order.

Ricci called the functions a_{ij}, a^{ij}, and a^i_j that transform in accordance with the respective rules

$$a_{i'j'} = \sum_i \sum_j a_{ij} \frac{\partial x^i}{\partial x^{i'}} \frac{\partial x^j}{\partial x^{j'}}, \quad a^{i'j'} = \sum_i \sum_j a^{ij} \frac{\partial x^{i'}}{\partial x^i} \frac{\partial x^{j'}}{\partial x^j}, \quad a^{i'}_{j'} = \sum_i \sum_j a^i_j \frac{\partial x^{i'}}{\partial x^i} \frac{\partial x^j}{\partial x^{j'}}, \tag{8.9}$$

twice covariant, *twice contravariant*, and *mixed* systems of the second order.

Ricci defined in a similar way systems of higher order. Contravariant systems of the first order at a point can be regarded as coordinates of vectors; covariant systems of the first order at the same point can be regarded as coefficients of linear forms defined on these vectors; twice covariant systems of the second order can be regarded as coefficients of bilinear forms defined on these vectors, and mixed systems of the second order can be regarded as matrices of linear transformations defined on these vectors. Einstein and Grossmann proposed that Ricci's "systems" be called *tensors* (and thus extended to them the term *elastic tensor* and that the "orders" of the systems be called *ranks*. The Dutch geometer Jan Arnoldus Schouten (1883–1971) in his *Ricci calculus* (Der Ricci-Kalkül. Berlin, 1924) *[510]* called Ricci's systems *affinors* and thereby extended to them a term used by F. Jung for linear operators; he called "orders" of the systems *valences*, borrowing this term from chemistry.

Most mathematicians now use the term *tensor* in the sense of Einstein and *valence* in the sense of Schouten. Following Einstein's suggestion, in the tensor calculus one omits the summation sign when summing with respect to covariant and contravariant indices. We will follow this rule in the following.

The coefficients of the quadratic form that defines the metric of a Riemannian space form a twice covariant tensor g_{ij}; the quadratic form is written in tensor notation as

$$ds^2 = g_{ij}dx^i dx^j, \tag{8.10}$$

and the equation of the geodesics is written as

$$\frac{d^2x^i}{dt^2} + \Gamma^i_{jk}\frac{dx^j}{dt}\frac{dx^k}{dt} = 0. \tag{8.11}$$

The Christoffel symbols Γ^i_{jk} do not form a tensor and transform under coordinate transformations in accordance with a more involved law. But the four-index Riemann symbols, now written as $R_{ij,k}^{\cdot\cdot\,\cdot l}$, do form a tensor of valence four called the *curvature tensor*.

Covariant differentiation as defined by Ricci associates to every tensor field in a Riemannian space—that is, to every function that defines a tensor at every point of a region in a Riemannian space—the field of a new tensor that has one additional covariant valence. The covariant derivative $\nabla_i\varphi$ of a scalar is its partial derivative $\frac{\partial\varphi}{\partial x^i}$ with respect to the coordinate x^i. The covariant derivatives $\nabla_j a^i$ and $\nabla_j a_i$ of vectors a^i and a_i are given by

$$\nabla_j a^i = \frac{\partial a^i}{\partial x^j} + \Gamma^i_{jk}a^k, \qquad \nabla_j a_i = \frac{\partial a_i}{\partial x^j} - \Gamma^k_{ji}a_k. \tag{8.12}$$

The covariant derivatives $\nabla_k a_{ij}$ and $\nabla_k a^i_j$ of the tensors a_{ij} and a^i_j have the form

$$\nabla_k a_{ij} = \frac{\partial a_{ij}}{\partial x^k} - \Gamma^l_{ki}a_{lj} - \Gamma^l_{kj}a_{il}, \qquad \nabla_k a^i_j = \frac{\partial a^i_j}{\partial x^k} + \Gamma^i_{kl}a^l_j - \Gamma^l_{kj}a^i_l. \tag{8.13}$$

The Geometry of the General Theory of Relativity

Einstein formulated the principle of general relativity in *The foundations of the general theory of relativity* as follows:

> The general laws of nature are to be expressed by equations which hold good for all systems of co-ordinates, that is, are covariant with respect to any substitutions whatever (generally covariant) *[340, p. 117]*.

The most important element of the general relativity theory is the geometric interpretation of gravity. In this theory gravitation is linked to the curvature of space, whose definition in a pseudo-Riemannian space is the same as that in a Riemannian space; the greater the density of matter in a certain region, and thus the intensity of the gravitational field, the greater the curvature of pseudo-Riemannian spacetime. A mass point subject to gravitational forces alone moves along a geodesic which satisfies the same equations as in Riemannian space.

Light rays, too, propagate along geodesics of pseudo-Riemannian spacetime. The deflection of the trajectories of material particles and of light rays from straight lines represents the attraction of particles of matter and of light by the heavy masses that gives rise to the gravitational field. In this connection we quote from Einstein's paper *On the ether* (Über den Aether. Zürich, 1924) *[166]*:

> General relativity theory sets aside one other shortcoming of classical dynamics: in the latter, inertia and weight appear as phenomena that are completely independent from one another in spite of the fact that both are conditioned by the same material constant -- mass. Relativity theory surmounts this shortcoming by establishing for the dynamical behavior of an electrically neutral particle the law of geodesics in which the interaction of inertia and gravitation is already inseparable. This interaction imparts to the ether a metric that varies from point to point and properties that determine the dynamical behavior of material points. In turn, these are determined by physical factors, namely the distribution of mass or energy. Thus the ether of general relativity differs from the ether of classical mechanics in that it is no longer "absolute" but is determined in the sense of its properties, that vary in space, by the distribution of weighted matter. This definition is complete provided that the world is spatially finite and closed *[166, p. 89]*.

Here the word *ether* is used in the sense of curved space whose geometric properties are determined by matter; in other papers Einstein replaces this term by the term *space-time continuum*. Einstein's reference to the finite and closed character of the universe reflects his original view that the curvature of space implies that space is finite and closed. (Cf. the paper *[649]* of Yakov Borisovič Zel'dovič (1914–1988).)

General relativity theory has given a physical interpretation to the curvature of pseudo-Riemannian spacetime. Light rays propagate along the isotropic geodesics of this spacetime. The deviation of the trajectories of material points and light rays from geodesics represents the attraction of material points and light by the heavy masses that give rise to the gravitational field.

The primary connection between non-Euclidean geometry and general relativity is that the discovery of non-Euclidean geometry extended the range of conceivable spaces and thus prepared the ground for the discovery of spaces of variable curvature, both Riemannian and pseudo-Riemannian; the latter represents the mathematical apparatus of general relativity. However, there is also a direct link between general relativity and non-Euclidean

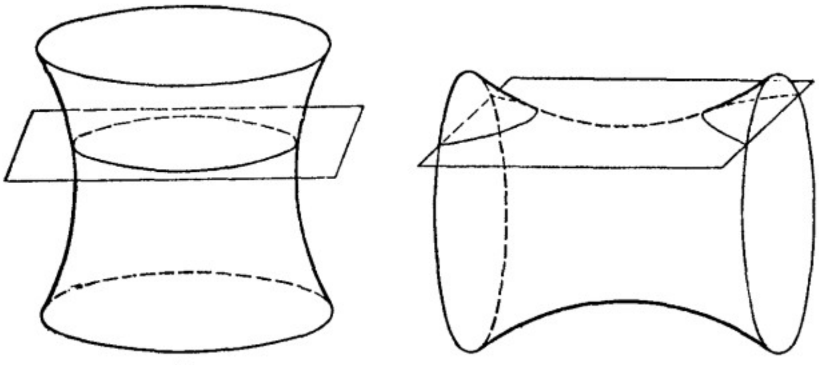

Figure 105

geometry, but what is involved is a four-dimensional rather than a three-dimensional space as in the case of special relativity. The geometry in question describes the spacetime of general relativity "in the mean," under the assumption of the uniform distribution of matter. This view was presented by Einstein already in the paper *Cosmological considerations linked to the general relativity theory* (Kosmologische Betrachtungen zur allgemeinen Relativitätstheorie. Berlin, 1917) *[340, pp. 175–189]*. Here Einstein considered a uniform distribution of matter in space alone, regarding space itself as a sphere in four-dimensional Euclidean space and spacetime as a cylinder in five-dimensional space. However, if we proceed from the assumption that we are dealing with a uniform distribution of matter throughout the spacetime continuum then spacetime is a sphere of real or imaginary radius in five-dimensional pseudo-Euclidean spaces of indices 4 and 3, respectively. If we assume, for simplicity, that the hypersurfaces $t =$ const. are parallel hyperplanes, then, in time, the "space section" of the world decreases or increases, depending on the position of the cutting hyperplane. In the first case the curvature of the "space section" is constant and positive, whereas in the second case it is constant and negative (Figures 105a and 105b), in full agreement with astronomical observations. This confirmation shows that real spacetime, that is, a pseudo-Riemannian space of variable curvature, corresponds in the mean to the picture just outlined; the theory of an expanding universe of constant negative curvature was proposed by A. A. Friedmann in the paper *On the possibility of a world of constant negative curvature* (Über die Möglichkeit einer Welt mit konstanter negativer Krümmung des Raumes. Leipzig, 1924) *[188]*.

The view of spacetime as a four-dimensional pseudo-Riemannian space of index 3 with positive curvature was proposed by the Belgian astronomer Willem de Sitter (1872 1934) in the paper *On Einstein's theory of gravitation and its astronomical consequences* (Brussels, 1917) *[541]*. In the paper *Geometrical note on de Sitter's world* (London, 1924) *[158]* Patrick du Val showed that "de Sitter's world" is a Lobačevskian space of positive curvature, that is, a sphere of real radius in five-dimensional pseudo-Euclidean space of index 4 with identified antipodal points (see also *[129a]*).

Einstein's Philosophy of Space

The discovery of the general theory of relativity was a brilliant concretization of the connection between matter and space and time, which are an existence form of matter— an issue often addressed by the classics of Marxism.

Einstein dealt with the philosophical questions of space in many of his papers. One relevant passage is the preceding quotation from his paper *On the ether*. The following quotation is from the paper *Geometry and experience* (Geometrie und Erfahrung. Berlin, 1921) *[165]*:

> Of all sciences mathematics is held in especially high esteem, for its theorems are absolutely true and indisputable while the areas of other sciences are to some extent debatable and there is always the danger of their results being overturned by new discoveries. But it is not fitting for a researcher working in some other area of science to envy mathematics, for the propositions of mathematics rest not on real objects but exclusively on objects of our imagination. In fact, it is little wonder that one attains logical agreement of deductions if one earlier reached agreement concerning the principal propositions (axioms) as well as the methods to be employed for deriving from the principal propositions other propositions. At the same time, the deep esteem for mathematics has another basis, which is that mathematics is that which gives to the exact sciences a measure of confidence which they could not otherwise attain.
>
> In this connection there arises a riddle that has worried researchers of all times. Whence the remarkable correspondence between mathematics and real things if mathematics is just a product of human thought unrelated to any experience? Can the human mind understand the properties of real things without any experience, just by way of reflection?
>
> To my mind, a concise answer to this question is this: to the extent to which the theorems of mathematics can be applied to reflect the real world they are not exact; they are exact to the extent to which they do not refer to reality. It seems to me that complete clarity in this matter can be achieved only by following the kind of mathematics known as axiomatics. The progress achieved through axiomatics consists in the fact that it sharply separated its formal-logical content from its objective and intuitive content. By point, straight line, and so on, in axiomatic geometry one should mean only contentless ideas. That which gives them content lies outside mathematics.
>
> On the other hand, however, it is true that mathematics in general and geometry in particular owe to their origins the necessity of learning

> a little about the behavior of objects existing in reality. This is shown by the very word "geometry" which means "measurement of the earth" *[165, pp. 3–5]*.

In this paper Einstein arrives

> at the viewpoint held by as original and deep a thinker as Henri Poincaré: Euclidean geometry differs from all conceivably possible axiomatic geometries by its simplicity. And since axiomatic geometry contains no statements about actual reality and can make such statements only in conjunction with physical laws, it seems possible and reasonable to adhere to Euclidean geometry regardless of the properties of reality. On the other hand, in case of an observed disagreement between theory and experience, it is easier to agree to change physical laws than axiomatic Euclidean geometry. If one forgets all about the connection between a practically rigid body and geometry, then it is not easy to refuse to admit that Euclidean geometry must be viewed as simplest.
>
> Why did Poincaré and other researchers reject the equivalence that thrusts itself upon one between a practically rigid body of real experience and a geometric body? Simply because, when one considers them more closely, real rigid bodies in nature turn out not to be rigid at all, since their geometric behavior, that is, their possible mutual disposition, depends on temperature, external forces, and so on. Thus the initial direct connection between geometry and physical reality is destroyed and we are forced to shift to the following more general position characteristic of Poincaré's viewpoint. Geometry (G) tells us nothing about the behavior of real objects; this behavior is described only by geometry together with the totality of physical laws (Ph). In symbolic terms we can say that only the sum (G) + (Ph) lends itself to experimental verification. Thus one can choose arbitrarily (G) as well as different parts of (Ph); all of these laws represent agreements. To avoid contradictions it is necessary to choose the rest of (Ph) so that (G) and all (Ph) are experimentally verifiable *[165, pp. 7–8]*.

We see that Einstein refers here to the philosophical works of Poincaré mentioned earlier but, unlike Poincaré, who treated "objective reality" and "laws of nature" in the spirit of Mach, Einstein proceeds from "physical reality" and the possibility of "choosing arbitrarily" (G), that is, geometric axioms which, together with (Ph), that is, physical laws, must confront experience. Thus Einstein interprets Poincaré's position in materialistic terms and, like Lobačevskiĭ, believes that the question of which of the conceivable geometric systems corresponds to the geometry of the real world must be decided experimentally.

It should be pointed out that Einstein frequently refers to Mach, whom he regards as one of his predecessors. In the obituary *Ernst Mach* (Ernst Mach. Leipzig, 1916) Einstein considered Mach's critique of Newton's mechanics and wrote:

> The quoted lines show that Mach clearly understood the weak points of classical mechanics and came close to discovering the general relativity theory. And that half a century before its creation! *[167]*.

Einstein regarded Mach's idea that the inertia of a body is explainable by its interaction with other bodies—what Einstein called *Mach's principle*—as one of the factors in the preparation of the general relativity theory. Mach's idea that "space and time cannot be completely separated from one another" *[347, p. 350]* is very close to the ideas of special relativity theory.

Undoubtedly, Einstein's famous *thought experiments* were influenced by Mach. At the same time Einstein did not agree with Mach's doctrine of sensations as elements of the world. He not only interpreted the terms *sensations* and *experience* in a materialist spirit but also ascribed his viewpoint to Mach when he said (apropos Mach's use of the term *sensations*) that

> frequently, as a result of insufficient familiarity with his, that is, Mach's, works, some persons tend to confuse them with the terminology of philosophical idealism and solipsism *[167]*.

Here Einstein's opinion was based on the fact that he took Mach's materialist terminology literally and throught it inconceivable that a serious physicist could be a subjective idealist. Later Einstein changed his opinion about Mach's philosophy and during a meeting of the French philosophical society, while answering a question from Emile Meyerson, he referred to Mach as a "deplorable philosopher" (*un déplorable philosophe*) and accused him of a "myopic view of science that led him to reject the existence of atoms" *[364, p. 62]*.

Parallel Displacement

The heightened interest in Riemannian geometry brought about by general relativity theory resulted in an extremely important discovery in this geometry, a concept known as *parallel displacement of vectors*. The discovery was made by Levi-Civita in the paper *The notion of parallelism in an arbitrary manifold and the geometric characterization of Riemannian curvature implied by it* (Nozione di parallelismo in una varieta qualunque e consequente specificazione geometrica della curvatura Riemannina. Palermo, 1917) *[320]*.

If a vector a^i is defined at a point A of a Riemannian space then the result of its displacement to a point B of the line $x^i(t)$ is a solution of the differential equation

$$\frac{da^i}{dt} + \Gamma^i_{jk} a^j \frac{dx^k}{dt} = 0, \tag{8.14}$$

that can be written with the help of the covariant derivative as

$$\nabla_k a^i \frac{dx^k}{dt} = 0. \tag{8.15}$$

It follows that if one defines along a coordinate line a vector field of vectors obtained by parallel displacement of a given vector then at all points of this line the covariant derivative with respect to the corresponding coordinate is equal to zero. Parallel displacement of vectors admits of a visual geometric definition. Parallel displacement of vectors from a point A of the space to an infinitely near point B is generated by mapping a neighborhood of the point A onto a neighborhood of the point B that is the result of applying first a reflection (along geodesics) in the point A in the neighborhood of A, and then a reflection in C, the midpoint of the geodesic arc from A to B, in a neighborhood of C that contains the neighborhood of A. Under this neighborhood mapping the increments of the coordinates of the point A map to the increments of the coordinates of the point B and, consequently, the differentials of the coordinates of A, that is, the vectors at A, map to the differentials of the coordinates of B, that is, the vectors at B. The name *parallel displacement* is due to the fact that the product of two reflections in two points in ordinary space is a translation of this space. Levi-Civita showed that the curvature of a Riemannian space, at a given point, in a two-dimensional direction is the limit of the ratio of the angle by which a vector is rotated as a result of parallel displacement along a closed contour passing through the given point to the area of the surface bounded by this contour when the contour shrinks to the given point (this is Levi-Civita's *geometric characterization of Riemannian curvature*).

The Problem of a Unified Field Theory

In general relativity theory the geometric properties of space are determined by the gravitational field alone, without any effect of the electromagnetic field on the geometry of space. After creating the general theory of relativity Einstein posed the problem of constructing a geometry of four-dimensional spacetime determined by the electromagnetic as well as the gravitational field. The first attempt of construction of a unified field theory was made by Hermann Weyl in his paper *Gravitation and electricity* (Gravitation und Elektrizität. Leipzig, 1918) *[340, pp. 200–216]*. Besides the metric quadratic form

$$ds^2 = \sum_{\mu,\nu} g_{\mu\nu} dx^\mu dx_\nu \tag{2}$$

Weyl considers also the linear form

$$d\phi = \sum_{\mu} \phi_\mu dx_\mu \tag{7}$$

and writes that

> *The internal metrical connexion of space thus depends on a linear form* (7) *besides the quadratic form* (2)—which is determined except as to an arbitrary factor of proportionality *[340, p. 206]*.

The vector ϕ_μ is the *vector-potential* of the electromagnetic field $F_{\mu\nu}$ given by

$$F_{\mu\nu} = \partial\phi_\mu/\partial x_\nu - \partial\phi_\nu/\partial x_\mu.$$

The geometry constructed by Weyl is called the *Weyl connection*. Weyl expounded this theory also in his *Space, Time and Matter* (1918) *[626]*.

Einstein himself devoted the last 35 years of his life to the search for a unified field theory. Of other attempts at constructing a unified field theory we note, first of all, the paper of Theodor Kaluza (1885–1954) *On the unicity problem of physics* (Zum Unitätsproblem der Physik. Berlin, 1921) *[259]*. To increase the number of relevant parameters Kaluza went in the direction of increasing the number of dimensions of his space and assumed that in addition to the four dimensions of physical spacetime there is a fifth dimension without a direct physical sense. Proceeding from Kaluza's idea the mathematician Veblen and the physicist Pauli proposed in 1930–1933 a unified field theory in which five coordinates are viewed as the projective coordinates of particles in a four-dimensional space.

A detailed history of unified field theories in the first third of the 20th century is given by Vladimir Pavlovič Vizgin in the book *Unified field theories in the first third of the 20th century* (Edinye teorii polya v pervoĭ treti XX veka. Moscow, 1985) *[615]*.

We note that when Einstein posed the question of a unified field theory he assumed that the theories of interactions of the gravitational and electromagnetic fields exhaust all forms of physical interactions. The discovery, in the last few decades, of "strong" and "weak" interactions that do not coincide with these interactions deprived the problem posed by Einstein of the importance he attached to it. In his "Remarks to the Einsteinian sketch of a unified field theory" the famous German physicist Werner Heisenberg (1901–1976) wrote:

> This attempt, grandiose in its conception, failed from the very beginning. At the time when Einstein concerned himself with the unified field theory new elementary particles were continuously discovered and with them their associated new fields. This meant that there existed no hard, solid, empirical basis for the implementation of the Einsteinian program and Einstein's attempts produced no conclusive results whatever. However, the failure that dogged the Einsteinian program also had deeper foundations than the uncertainty of empirical facts; these foundations have to do with the relations of Einstein's field-theoretic view of quantum theory. Einstein proceeded from the classical nonlinear field theory of matter for the metric tensor of the field that defined his geometry.

> Einstein hoped that in the end it will be possible to think of atoms and elementary particles in such a theory as singular solutions of the nonlinear equations of the field. In reality, however, the element of discreteness that expresses the existence of elementary particles has a far wider character. It becomes dominant as soon as we go over to the domain of atoms and elementary particles; which is why their description by means of a field theory of the classical type is out of the question. Conversely, we now know that there operate here quantum-mechanical laws whose structure was understood in the twenties. Since Einstein could not come to terms with such a structure, he did not make an attempt to approach the unified field theory using quantum laws.
>
> Nevertheless, experiments with elementary particles carried out in connection with quantum theory contain very many arguments in favor of Einstein's program.
>
> In the last ten years many elementary particles have been discovered and thus also many new fields. Also, it became clear that elementary particles can change into one another. If two elementary particles with very high kinetic energy collide in some way then there can arise new particles, and the laws that govern the coming into existence and the disappearance of elementary particles can, apparently, be formulated by means of relatively simple selection rules and the corresponding quantum numbers. It follows that all elementary particles "consist," so to say, of one and the same substance that can be simply called "energy" or "matter"; their structure and their ability to go over into one another should follow from a simple law for matter.
>
> Thus, by the nature of things, a satisfactory theory of elementary particles must, at the same time, be a unified field theory of matter *[221a, pp. 120 - 125]*.

An attempt to create such a unified theory of elementary particles was made by Heisenberg himself.

Spaces with Affine Connection

Although the problem of a unified field theory has not yielded any substantial results for physics, it has been remarkably fruitful for geometry. Weyl's theory was further developed by Schouten, who introduced the general notion of a space with an affine connection in the paper *On different forms of connection that can be laid at the foundation of a differential geometry* (Über die verschiedenen Arten der Übertragung die einer Differentialgeometrie zugrunde gelegt werden können. Berlin, 1922) *[511]*. A detailed exposition of Schouten's theory is found in his *Ricci calculus*. A space with an affine connection is defined as a manifold of points with coordinates x^i at each of whose points there is given a point function $\Gamma^i_{jk} = \Gamma^i_{kj}$ that is not expressible in the general case in terms of the point functions g_{ij}. Because of this last condition it is not possible, in the general case, to define in a space with an affine connection the length of lines, but it *is* possible to define parallel displacement of vectors by formula (8.14). In a space with an affine connection geodesics are defined. They are solutions of the equations (8.11). For each arc of a geodesic there is defined its midpoint. In such a space one can define reflections in points along geodesics. Just as in a Riemannian space, the geometric sense of parallel displacement of vectors is that it is determined by a mapping of a neighborhood of a point of the space onto a neighborhood of an infinitely near point that is the result of two reflections in two points along geodesics. The name *affine connection* is due to the fact that the mapping of vectors at a point of space onto vectors at an infinitely near point of this space is an affine transformation of vector spaces.

It is clear that Riemannian and pseudo-Riemannian spaces are special cases of spaces with an affine connection. In such spaces the mapping of vectors at one point onto vectors at an infinitely close point under parallel displacement is an isometry of Euclidean and pseudo-Euclidean spaces, respectively.

The space introduced by Weyl in his unified field theory is also a space with an affine connection. In this space the mapping of vectors at one point onto vectors at an infinitely close point under parallel displacement is a similarity transformation.

Even more general spaces were introduced by the French mathematician Élie Cartan (1869–1951). In the papers *On manifolds with an affine connection and the generalized relativity theory* (Sur les variétés à connexion affine et la théorie de la relativité généralisée. Paris, 1923) *[96, part 3, pp. 659–746, 798–825, 921–992]*, *Spaces with a conformal connection* (Les espaces à connexion conforme. Paris, 1923) *[96, part 3, pp. 747–797]* and *On manifolds with a projective connection* (Sur les variétés à connexion projective. Paris, 1924) *[96, part 3, pp. 825–861]* Cartan introduced the notion of a *space with a homogeneous connection*. Whereas in the case of a space with an affine connection there is associated to every point a vector space that can be regarded as an affine space, and under parallel displacement of vectors the space at one point is mapped by an affine transformation onto the space at an infinitely near point, in the case of a space with a homogeneous connection there corresponds to every point of the space a homogeneous space with a definite group of transformations. Also, for every pair of infinitely near points there is specified a transformation of the spaces associated with these points that preserves their geometry. In addition to spaces with an affine connection and special cases of such spaces noted above, the most important spaces with a homogeneous connection are *spaces with a projective* and *conformal connection*, respectively. To each point of these spaces there is associated, respectively, a projective or conformal space, and for each pair of infinitely near points of a space with a connection there is specified, respectively, a projective or conformal mapping of one of the spaces associated with these points onto the other.

Differentiable Manifolds

The spaces introduced by Riemann and the groups introduced by Lie, bearing the names of their respective inventors, as well as pseudo-Riemannian spaces, Weyl spaces, and the more general spaces with an affine or other homogeneous connection were defined only locally—in the domain of a certain coordinate system. The introduction of topological manifolds made possible the study of differential-geometric spaces in the large. In every n-dimensional topological manifold one can define a homeomorphic mapping of a neighborhood U of every point onto some domain of n-dimensional Euclidean space, and the coordinate system in the Euclidean space provides coordinates for the neighborhood U. This system of coordinates in the neighborhood U is now called a *system of local coordinates* or a *local map* of the manifold, and the totality of local maps is called the *atlas* of the manifold. A manifold is called *differentiable of class* v if (1) the domains of the maps of the atlas cover the whole manifold; and (2) if U and U' are two domains of maps of the atlas with nonempty intersection and a point x in the intersection has coordinates x^i and $x^{i'}$ in the two maps then the functions $x^{i'} = f(x^1, \ldots, x^n)$ have continuous partial derivatives up to and including order v, and the jacobian $\det\left(\frac{\partial x^{i'}}{\partial x^j}\right)$ is different from zero throughout the intersection of the domains. The manifold is called *analytic* if the functions in (2) are analytic.

The definitions of a differentiable manifold and of an analytic manifold were formulated by Oswald Veblen (1880–1960) and by John Henry Constantine Whitehead (1904–1960) in *The foundations of differential geometry* (Cambridge, 1932) *[612]*.

With every point of a differentiable manifold there is associated a vector space with vectors $\{dx^i\}$ whose coordinates are the differentials of the coordinates x^i; these spaces are called the *tangent spaces* of the differentiable manifold. Under coordinate transformations $x^i \to x^{i'}$ the coordinates of the vectors and tensors of the tangent spaces transform according to the rules (8.7)–(8.9), and so on. If in every tangent space of a differentiable manifold there is given a metric tensor g_{ij} that determines an inner product in the Euclidean or pseudo-Euclidean space then the differentiable manifold is called, respectively, a *Riemannian* or *pseudo-Riemannian space in the large*. If in each of the tangent spaces of a differentiable manifold there is given an object Γ^i_{jk} that transforms under coordinate transformations according to the same rule as the coefficients Γ^i_{jk} of Riemannian spaces then the differentiable space is called a *space with an affine connection in the large*.

Fibrations

Spaces with connections are special cases of fiber spaces. Unlike spaces with Schouten-Cartan connections, fiber spaces are considered in the large. The initial form of the theory of fiber spaces in the large was the theory of "sphere-spaces," that is, fiber spaces whose "fibers" are spheres. This theory arose in connection with the solution of the problem of continuous mapping of an n-dimensional sphere onto an m-dimensional one for, if $n > m$, then to every point of the m-dimensional sphere there corresponds an $(n - m)$-dimensional manifold of the n-dimensional sphere, so that the latter "stratifies" or "fibrates" into these manifolds; the resulting fibres are also called *spheres*. The theory of such spaces was constructed by the American mathematician Hassler Whitney (b. 1907) in the paper *Sphere-spaces* (Boston, 1935) *[632]*. In the paper *On the theory of sphere bundles* (Boston, 1940) *[633]* Whitney calls these spaces *sphere-bundles*. The most complete account of Whitney's theory is contained in his lectures *On the topology of differentiable manifolds* (Ann Arbor, 1941) *[634]*. Together with *sphere-bundles* Whitney also considered *plane-bundles*. Whitney associated to every differentiable manifold a *tangent plane-bundle* consisting of the tangent spaces of the manifold and *tangent sphere-bundle* consisting of spheres in these tangent spaces with centers at the points of tangency, with one sphere in each tangent space. In the first case, the points of the *fibers* characterize the vectors of the differentiable manifold; in the second, they characterize the directions issuing from a point of the manifold.

Soon Whitney's *sphere-bundles* were replaced by arbitrary manifolds. Then there arose the more general concept of a *fiber bundle* or a *fiber space*. The general theory of such spaces was set forth by the American mathematician Norman Steenrod (1910–1971) in the monograph *The topology of fiber bundles* (Princeton, 1951) *[560]*, which completed the development of this theory.

The simplest fiber space is the *direct product* of two topological spaces S and T, that is, the totality of pairs (s, t) of elements s in S and t in T where the closed sets of the topological product are sets of pairs (s, t) in which both elements s and t vary over closed sets in the spaces S and T, respectively. In this case one may view the set of pairs (s, t_0) with fixed t_0 as a *fiber* and the space T of all possible elements t_0 as a basis of the fiber space. One can also view the set of pairs (s_0, t) with fixed s_0 as a *fiber* and the space S as a basis of the same fiber space. In this case the fibration is called *trivial*.

In more complicated cases, a fibre space is not simply the direct product of a *fiber* by a *basis* but also consists of homeomorphic fibers, and, if the latter are viewed as points of some new space, then this space can be regarded as a basis of the fiber space. The simplest fiber space of this kind is a *Möbius strip* obtained from the lateral surface of a right circular cylinder by identification of pairs of points that are symmetric with respect to its center, or from a rectangle (which may be thought of as the application of half of the preceding cylinder to a plane) by gluing the points of a pair of its opposite sides that are symmetric with respect to its center. (Hence the name *skew products* for general fibrations of a space used in the Russian translation of N. Steenrod's book *[560]*). One usually considers also the group of topological mappings of a fiber onto itself that preserves a certain structure of these fibers and mappings of fibers, one onto another, that preserve this structure.

The most important fiber spaces are the *tangent plane-bundles* defined by Whitney for differentiable manifolds. Now they are called simply *tangent bundles* of differentiable manifolds. The bases of these fiber spaces are the differentiable manifolds, and the fibres are their tangent spaces. To the class of fiber spaces there belong Riemannian spaces, pseudo-Riemannian spaces, and spaces with an affine connection in the large together with their tangent spaces that are, respectively, Euclidean, pseudo-Euclidean, and affine, as well as spaces with a projective, conformal, and arbitrary homogeneous connection together with the corresponding homogeneous spaces that also form *fibers*. The groups of transformations of the *fibers* of these spaces are, respectively, the groups of Euclidean and pseudo-Euclidean motions, of affine, projective, conformal, and other transformations.

Exterior Forms, Curvature and Betti Numbers

The field of a skew-symmetric covariant tensor $a_{i_1 i_2 \dots i_k}$ in a domain of a differentiable manifold defines at every point of this domain an exterior differential form

$$\omega_k = \sum_{i_1} \dots \sum_{i_k} a_{i_1 i_2 \dots i_k} dx^{i_1} \wedge dx^{i_2} \wedge \cdots \wedge dx^{i_k},$$

where $\wedge$ denotes exterior multiplication. The form is invariant under even permutations of the dx^{i_a} and changes sign under odd permutations of the dx^{i_a} (in virtue of the skew symmetry of $a_{i_1 \dots i_k}$ the ordinary differential form $\sum_{i_1} \dots \sum_{i_k} a_{i_1 i_2 \dots i_k} dx^{i_1} dx^{i_2} \dots dx^{i_k}$ is equal to zero). For exterior forms we define the operations of exterior multiplication

$$\omega_{k+l} = \omega_k \wedge \omega_l$$

and exterior differentiation

$$\omega_{k+1} = \mathscr{D}\omega_k = \sum_{i_1} \dots \sum_{i_k} da_{i_1 \dots i_k} \wedge dx^{i_1} \wedge \cdots \wedge dx^{i_k}.$$

The differential forms in multiple integrals defining areas and volumes are exterior forms, and the Ostrogradskiĭ formula in the beginning of chapter 7 can be written as

$$\int_{\partial M} \omega = \int_M \mathscr{D}\omega,$$

where M is a bounded domain and ∂M is its boundary.

The theory of exterior forms was founded by É. Cartan in the paper *On certain differential expressions and the Pfaff problem* (Sur certaines expressions différentielles et le problème de Pfaff. Paris, 1899) *[96, part 2, pp. 303–396]*. It was extensively developed by É. Cartan (see his book *Exterior differential systems and their geometrical applications* (Les systèmes différentiels extérieurs et leurs applications géométriques. Paris, 1945) *[97]*, by the Soviet geometer Sergeĭ Pavlovič Finikov (1883–1964), and by their students.

If a moving orthonormal frame $\{\mathbf{e}_i\}$ is associated with every point of a curve or surface in the space R_n or with every figure of a family of figures in this space, then the differentials of the radius vector $\mathbf{x}$ and of the vectors $\mathbf{e}_i$ of the frame can be written as

$$d\mathbf{x} = \mathbf{e}_i\omega^i, \qquad d\mathbf{e}_i = \mathbf{e}_j\omega_i^j, \tag{8.16}$$

where

$$\omega_j^i = -\omega_i^j, \tag{8.17}$$

and the exterior differentials of the forms ω^i and ω_j^i satisfy the *equations of structure*

$$\mathscr{D}\omega^i = \sum_k \omega^k \wedge \omega_k^i, \qquad \mathscr{D}\omega_j^i = \sum_k \omega_j^k \wedge \omega_k^i. \tag{8.18}$$

A curve, surface or family of figures is defined by a system of Pfaff equations $\theta^1 = \theta^2 = \cdots = \theta^r = 0$, where the forms θ^A are linear combinations of the forms ω^i, ω_j^i. For a curve with arc length s and tangent vector $\mathbf{e}_1$ these equations are $\omega^2 = \cdots = \omega^n = 0$ and for a hypersurface with normal vector $\mathbf{e}_n$ the equation is $\omega^n = 0$. If for the curve we put $d\mathbf{e}_1/ds = k_1\mathbf{e}_2$ and choose vectors $\mathbf{e}_3, \ldots, \mathbf{e}_n$ so that $d\mathbf{e}_i/ds$ is a linear combination of $\mathbf{e}_1, \ldots, \mathbf{e}_{i+1}$, then $\omega_i^{i+1}/ds = -\omega_{i+1}^i/ds = k_i$ are curvatures of the curve and the remaining ω_j^i are zeros. In this way we obtain the *Frenet formulas*

$$\begin{aligned} &d\mathbf{e}_1/ds = k_1\mathbf{e}_2,\\ &d\mathbf{e}_2/ds = -k_1\mathbf{e}_1 + k_2\mathbf{e}_3, \ldots,\\ &\quad d\mathbf{e}_i/ds = -k_{i-1}\mathbf{e}_{i-1} + k_i\mathbf{e}_{i+1}, \ldots, d\mathbf{e}_n/ds = -k_{n-1}\mathbf{e}_{n-1} \end{aligned}$$

discovered for R_3 in 1847 by Frederic Frenet (1816–1900) and for R_n in 1874 by C. Jordan. The curvatures k_i define the curve up to a motion of R_n.

For a hypersurface we direct the vectors $\mathbf{e}_1, \ldots, \mathbf{e}_{n-1}$ along curvature lines of the hypersurface and the vector $\mathbf{e}_n$ along its normal. Exterior differentiation of the Pfaff equation $\omega^n = 0$ gives, in view of (8.18), $\mathscr{D}\omega^n = \sum_k \omega^k \wedge \omega_k^n = 0$. Hence $\omega_\alpha^n = \sum_\beta b_{\alpha\beta}\omega^\beta$ $(\alpha, \beta = 1, 2, \ldots, n-1)$. We note that the analogues of Gauss's quadratic forms for surfaces in R_3 and for hypersurfaces in R_n are

$$d\mathbf{x}^2 = \sum_\alpha (\omega^\alpha)^2,$$

$$\mathbf{e}_n d^2\mathbf{x} = -d\mathbf{e}_n d\mathbf{x} = -\left(-\sum_\alpha \omega_\alpha^n \mathbf{e}_\alpha\right)\left(\sum_\beta \omega^\beta \mathbf{e}_\beta\right) = \sum_\alpha \sum_\beta b_{\alpha\beta}\omega^\alpha\omega^\beta.$$

In our basis $b_{\alpha\beta} = b_\alpha\delta_{\alpha\beta}$, $-b_\alpha = k_\alpha$ are main curvatures of the hypersurface; they are differential invariants of the hypersurface. Exterior differentiation of other forms yields a complete system of differential invariants of the hypersurface which defines it up to a motion of the space R_n.

Formulas (8.16) and (8.18) also hold in the affine space E_n and in the pseudo-Euclidean space ${}^lR_n$. In the case of E_n the forms ω_j^i are not connected by any condition. In the case of ${}^lR_n$ with scalar square of a vector $\mathbf{x}^2 = \sum_i \varepsilon_i (x^i)^2$ the conditions (8.17) become

$$\varepsilon_i \omega^i_j = -\varepsilon_j \omega^j_i. \tag{8.19}$$

In the spaces P_n, S_n and ${}^lS_n$ the points are represented by vectors of the spaces E_{n+1}, R_{n+1} and ${}^lR_{n+1}$, respectively. In these spaces the role of the radius vector $\mathbf{x}$ is played by e_0. In these spaces hold the second formulas (8.16) and (8.18). In P_n the forms ω^i_j are connected by the condition $\sum_i \omega^i_i = 0$, in S_n by the condition (8.17), and in ${}^lR_n$ by the condition (8.19).

Analogous methods can be applied in all homogeneous spaces.

These methods can also be applied in spaces with affine connection and in Riemannian and pseudo-Riemannian spaces. In these spaces the vectors $\mathbf{e}_i$ are vectors of the tangent spaces of E_n, R_n and ${}^lR_n$ respectively. In these spaces hold formulas (8.16) and in Riemannian and pseudo-Riemannian spaces formulas (8.17) and (8.19), respectively. The equations of structure there have the form

$$\mathscr{D}\omega^i = \sum_k \omega^k \wedge \omega^i_k + \sum_j \sum_k S^i_{jk} \omega^j \wedge \omega^k,$$
$$\mathscr{D}\omega^i_j = \sum \omega^k_j \wedge \omega^i_k + \sum_k \sum_h R_{jkh}{}^{\cdots i} \omega^k \wedge \omega^h$$

where S^i_{jk} is the *torsion tensor* (equal to zero in spaces without torsion, in particular, in Riemannian and pseudo-Riemannian spaces), $R_{jkh}{}^{\cdots i}$ is the *curvature tensor* (in Riemannian spaces $R_{ijkl} = g_{hl} R^h{}_{ijk}$ is the Riemann tensor).

An exterior differential form ω is called *closed* if $\mathscr{D}\omega = 0$ and *exact* if $\omega = \mathscr{D}\omega'$. Every exact form is closed since $\mathscr{D}(\mathscr{D}\omega) = 0$. Exterior differentiation is similar to the taking of a boundary, except that in the latter case the dimension of a manifold decreases and in the former case the "order" of a form increases. This analogy was made use of in the paper *On analysis situs* [i.e. topology] *of n-dimensional manifolds* (Sur l'analysis situs des variétés à n-dimensions. Paris, 1931) *[452]* by the Belgian mathematician Georges de Rham (1903–1969). In this paper de Rham defined the *cohomology groups*, in a certain sense dual to Betti's *homology groups*, and applied exterior differential forms for the calculation of Betti numbers. The integral of a closed form is said to be homologous to zero, and h integrals of closed forms on a p-dimensional submanifold which are independent in the usual topological sense are said to be homologically independent if no linear combination of these integrals with constant coefficients not all zero is homologous to zero. De Rham has shown that the Betti number b_p is equal to the number of homologically independent integrals of exact differential forms of order p. This method of calculation of Betti numbers was applied by Cartan in his *Topology of compact homogeneous spaces* (La topologie des espaces homogènes clos. Moscow, 1937) *[96, part 1, pp. 1331–1338]* and by his student Charles Ehresmann in the paper *On the topology of certain homogeneous spaces* (Sur la topologie de certains espaces homogènes. Princeton, 1934) *[164]*.

The Betti numbers are connected with the Gaussian curvature of a surface: if the Gaussian integral curvature $\int K d\sigma$ is taken over a closed compact surface in R_3, then this integral is equal to $4\pi\left(1-\frac{b_1}{2}\right)$, where b_1 is the first Betti number of the surface. This number is equal to double the number of handles of the surface ($b_1 = 0$ for a sphere and $= 2$ for a torus). About this problem see the book of Kentaro Yano and Salomon Bochner *Curvature and Betti numbers* (Princeton, 1953) *[645]*.

data 02.02.2020

Pág. 280 (292) - Chapter 8

The Curvature of Space

Curvature and Intrinsic Geometry of a Surface in the Works of Euler

$$K = \frac{d\alpha}{ds} = \lim_{\Delta s \to 0} \frac{\Delta \alpha}{\Delta s}$$

K is also the reciprocal of the radius of curvature at the point in question, that is, the radius of the OSCULATING circle at that point.

(The osculating circle at a point P of a curve is defined as the limit of circles determined by three points on the curve as they tend to P.)

The concepts of the curvature of a curve and of the osculating circle were already know to Leibniz.

↳ obra: New reflections on the nature of the angles of tangency and osculation and their use in mathematical

02.02.2020

practice to replace complex figures with simpler ones.

Leipzig 1686 [312, v.7, pp. 326-329]

Leibniz also suggested the possibility of characterizing the curvature of a surface by means of an osculating sphere

↳ In a letter to Johann Bernoulli, dated July 29, 1698

111) ARQ-Important_05/Géométrie/Libre-Géométrie-Projective/Geometrie-Projective/

a) livre-exercices-de-geometrie-fgm-bnf.pdf

Pág. 78 (110)

177a. Théorème de Desargues

b) livre-geometrie-projective-daniel-perrin-~~1~~328pages-pdf

Pág. 65 (100)

Théorème de Newton

Figure 3.7 Le théorème de Newton projectif

3e Cas. *On emploie un volume auxiliaire, afin d'étudier les propriétés d'une figure plane, que l'on peut considérer comme étant une section du solide.*

Exercice.

174. Théorème. *Sur une sécante quelconque, l'hyperbole et ses asymptotes interceptent des segments égaux.*

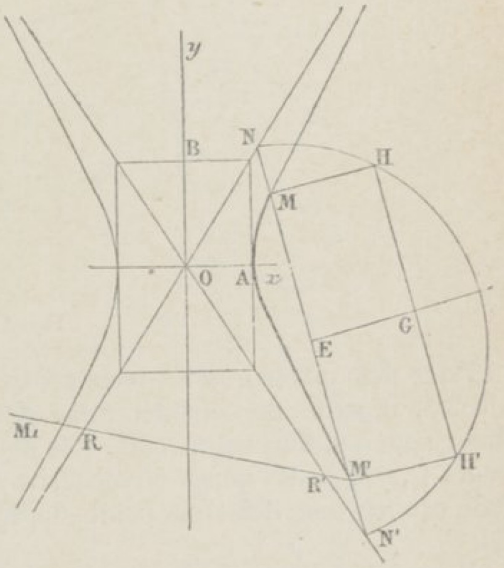

Fig. 110.

Considérons le cône formé par la rotation de ON autour de Ox. Un plan sécant perpendiculaire au méridien principal, et dont la trace serait NN', couperait le cône suivant une ellipse, puisque toutes les génératrices de la même nappe seraient rencontrées. (G., n° 844.)

Soit NHN' le rabattement de la moitié de l'ellipse : le plan qui donne l'hyperbole est éloigné de l'axe du cône de la longueur OB ; donc sa trace sur l'ellipse est une corde HH' parallèle à NN' et telle que $EG = OB$. Mais, NN' étant le grand axe de l'ellipse, la perpendiculaire élevée au milieu de NN' divise toute corde parallèle en deux parties égales : ainsi $GH = GH'$; donc $MN = M'N'$.

Si la sécante coupait les deux branches, la section serait une hyperbole dont RR' serait l'axe transverse ; $M'M_1$ serait la projection d'une corde parallèle ; donc encore, $M_1R = M'R'$.

Application. A l'aide de la propriété démontrée, on construit très facilement une hyperbole lorsqu'on connaît les asymptotes et un point de la courbe.

175. Corollaire. *Toute tangente limitée aux asymptotes est divisée en deux parties égales par le point de contact*[*].

Exercice.

176. Théorème de d'Alembert[*]. *Trois circonférences, considérées deux à deux, ont six centres de similitude ; les trois centres extérieurs sont*

---

[*] On trouve facilement la plupart des propriétés de l'ellipse lorsqu'on la considère comme étant obtenue par la section oblique d'un cône de révolution. MAC-LAURIN, dès 1742, en donne de nombreux exemples dans son *Traité des fluxions*, tome II, chap. XIV, page 96.

[*] D'ALEMBERT, né à Paris en 1717, mort en 1783. On lui doit un *Traité de dynamique*, le *Traité de l'équilibre et du mouvement des fluides*, et un grand nombre d'autres écrits.

en ligne droite ; il en est de même de deux centres intérieurs et d'un centre extérieur.

Démonstration de Monge *. Considérons des sphères ayant pour grands cercles les cercles donnés A, B, C. Les cônes, circonscrits à ces sphères prises deux à deux, ont respectivement pour sommets les centres de similitude.

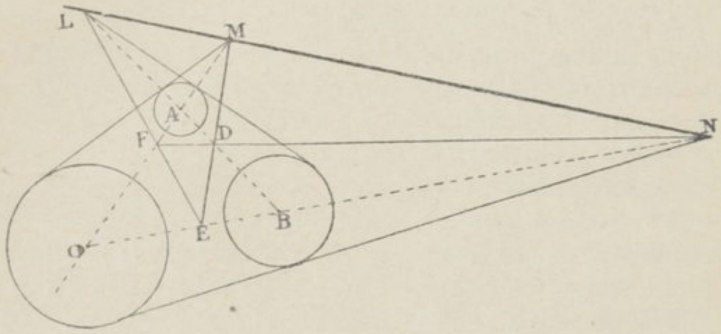

Fig. 111.

Pour démontrer que les trois centres extérieurs L, M, N sont en ligne droite, il suffit de considérer les deux plans tangents qui laissent les trois sphères d'un même côté. Ces deux plans contiennent les trois sommets L, M, N des cônes circonscrits ; or deux plans se coupent suivant une droite ; donc les trois points L, M, N sont en ligne droite.

Remarque. Pour F, D, N, on considère les deux plans tangents qui laissent les sphères B et C d'un même côté, tandis que la sphère A est de l'autre côté, etc.

Exercice.

177 *a.* **Théorème de Desargues** **. *Lorsque les côtés de deux triangles* ABC, abc *se coupent deux à deux en trois points situés en ligne droite, les droites* Aa, Bb, Cc, *qui joignent les sommets correspondants, se coupent au même point.*

Admettons que *abc* soit la base d'un prisme triangulaire, dont A'B'C' serait la section par un plan mené par la droite LMN.

Les droites AB, A'B' concourent au point L, car LA'B' est l'intersection du plan sécant et du plan conduit par L*ab*.

Il est évident que les droites AA', BB', CC' concourent en un même point S ; car si par les lignes concourantes LAB, LA'B' on fait passer un premier plan, puis un second par MAC, MA′C′, un troisième par NBC, NB′C′, les trois plans se coupent en un même point S.

---

* Monge, né à Beaune en 1746, mort en 1818, élève, puis répétiteur à l'école militaire de Mézières, est le principal créateur de la *géométrie descriptive*. Après avoir accompagné Bonaparte en Égypte, il eut à son retour la direction de l'École polytechnique.

** Desargues, né à Lyon en 1593, mort en 1662, s'occupa surtout de la partie pratique des mathématiques. Pascal, Descartes, Fermat, La Hire, ont profité des idées de cet auteur. On doit à Desargues le théorème relatif à deux triangles dont les sommets sont deux à deux sur trois droites concourantes, théorème que Poncelet a pris pour base de sa théorie des figures homologiques.

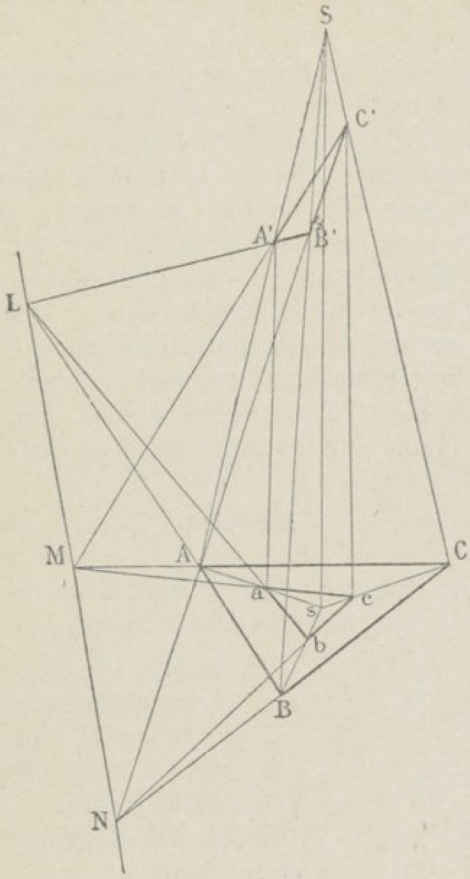

Fig. 112.

Donc les trois droites A*a*, B*b*, C*c* concourent en un même point *s*, projection du sommet S de la pyramide sur la base ABC.

Exercice.

177 *b*. Théorème réciproque. *Lorsque les sommets de deux triangles* ABC, abc *sont deux à deux sur trois droites qui concourent en un même point* s, *les côtés des triangles se coupent deux à deux, en trois points* L, M, N *situés en ligne droite.*

Considérons une pyramide dont *sa*A, *sb*B, *sc*C seraient les projections des arêtes latérales. Les projetantes qui correspondent aux sommets *a*, *b*, *c* donneraient A′, B′, C′, sur les arêtes correspondantes.

Or le plan de la section A′B′C′ coupe celui de la base suivant une certaine droite, et les côtés correspondants AB, A′B′ se coupent sur cette droite, en L, par exemple; donc *ab* passe aussi par ce point, car *ab* est la projection de A′B′.

Remarque. Les deux théorèmes de *Desargues* sont fondamentaux dans la théorie de l'*homologie* *.

---

* L'***homologie***, comme corps de doctrine et procédé général de transformation des figures,

§ VI. — Projections ou Sections.

178. La méthode des projections ou des sections est, en quelque sorte, la contre-partie de la méthode qui emploie des surfaces ou des volumes auxiliaires pour étudier des questions de géométrie plane. En effet, par la méthode des projections, on se propose d'obtenir une figure plus simple que la figure proposée, ou bien on ramène une question de géométrie dans l'espace à un exercice plan, se bornant à étudier la section obtenue en coupant le solide par un plan convenablement choisi.

Nous ne considérons ici que la *projection cylindrique,* c'est-à-dire la projection obtenue par des droites parallèles entre elles, mais dans une direction d'ailleurs quelconque par rapport au plan de la section. Ainsi, étudier la projection plane d'une figure donnée revient à considérer la section du cylindre formé par les projetantes de cette figure donnée.

Exercice.

179. Théorème. *La bissectrice de l'angle d'un triangle divise le côté opposé en segments proportionnels aux côtés adjacents.*

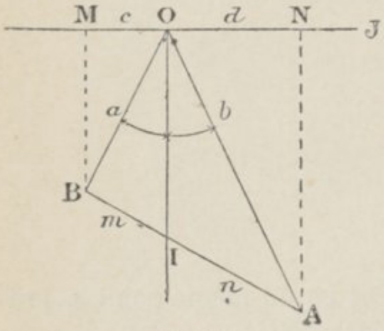

Fig. 113.

Projetons les sommets A et B sur la bissectrice extérieure OJ.

Les côtés a et b étant également inclinés sur OI sont proportionnels à leurs projections c et d.

Il en est de même des segments m et n; donc

$$\frac{a}{b}=\frac{c}{d}=\frac{m}{n} \quad C.\ Q.\ F.\ D.$$

Remarque. Pour les segments déterminés par la bissectrice extérieure, on projette ces segments et les côtés adjacents sur la bissectrice intérieure. La démonstration est tout aussi simple que la précédente.

Exercice.

180. Théorème de Ménélaüs. *Lorsqu'une transversale coupe les trois côtés d'un triangle, le produit de trois segments, n'ayant pas d'extrémité commune, égale le produit des trois autres segments* (n° 166).

---

est due à Poncelet. Pour se rendre compte de la fécondité de cette méthode et de l'esprit investigateur du créateur de l'homologie, il faut lire son *Traité des propriétés projectives des figures*, et les *Applications d'analyse et de géométrie*, du même auteur.

Traité des propriétés projectives des figures, 2 vol. in-4°. La première édition est de 1822, et la seconde de 1865. Les premières recherches datent de 1813, pendant la captivité de l'auteur en Russie; elles furent communiquées dès 1814 à MM. François et Servois, professeurs aux écoles d'artillerie et du génie à Metz.

Quelques fragments de ces recherches ont été publiés en 1817-1818 dans le tome VIII des *Annales de Gergonne.*

Sur une droite quelconque, projetons les trois sommets du triangle par des lignes Aa, Bb, Cc parallèles à la transversale. Le point o est la projection des trois points L, M, N.

On sait que les parallèles divisent les sécantes en parties proportionnelles; on peut donc remplacer le rapport $\frac{AL}{BL}$ par $\frac{ao}{bo}$, etc.

Mais $AL.BM.CN = BL.CM.AN$ ou $\frac{AL}{BL} \cdot \frac{BM}{CM} \cdot \frac{CN}{AN} = 1$ peut être remplacé par

$$\frac{ao}{bo} \cdot \frac{bo}{co} \cdot \frac{co}{ao} = 1$$

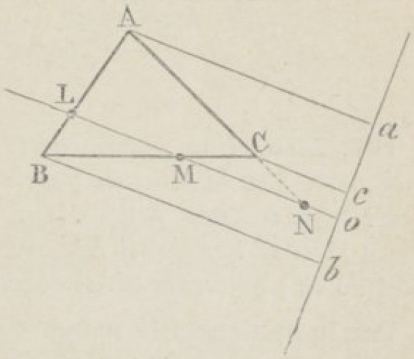

Fig. 114.

Or cette dernière égalité est évidente; la relation demandée est par suite démontrée.

Exercice.

181. Théorème de Carnot *. *Lorsqu'une transversale coupe les côtés d'un polygone plan, chaque côté est divisé en deux segments; le produit de tous les segments, n'ayant pas d'extrémité commune, égale le produit de tous les autres segments.*

Soit, par exemple, un pentagone ABCDE dont les côtés successifs AB, BC,... sont coupés par une transversale en des points H, K, L, M, N.

Projetons la figure sur une droite quelconque xy située dans son plan, par des droites parallèles à la transversale, et soit O le point où cette ligne rencontre xy.

Il faut prouver qu'on a :

$$\frac{AH}{BH} \cdot \frac{BK}{CK} \cdot \frac{CL}{DL} \cdot \frac{DM}{EM} \cdot \frac{EN}{AN} = 1$$

ou

$$\frac{ao}{bo} \cdot \frac{bo}{co} \cdot \frac{co}{do} \cdot \frac{do}{eo} \cdot \frac{eo}{ao} = 1$$

Or cette dernière relation est évidente. La première est donc démontrée.

Remarque. On démontre aussi d'une manière fort simple la généralisation suivante :

Lorsqu'un plan coupe les côtés d'un polygogne gauche, chaque côté est divisé en deux segments; le produit de tous les segments; n'ayant pas d'extrémité commune, égale le produit de tous les autres segments.

---

* CARNOT, né à Nolay (Côte-d'Or) en 1753, mort à Magdebourg en 1823, élève de MONGE à l'école de Mézières, publia un *Essai sur les transversales, De la corrélation dans les figures de géométrie*, la *Géométrie de position*.

On ne cite généralement que la *Géométrie de position*, publiée en 1803; mais le théorème ci-dessus, ainsi que son extension à un polygone gauche, se trouve déjà dans l'ouvrage publié en 1801 : *De la corrélation des figures de géométrie*, n^{os} 220 et 221, page 162.

Il suffit de projeter la figure sur un plan non parallèle au plan sécant, en recourant à des droites parallèles à ce même plan sécant; car on retombe sur le *théorème de Carnot.*

Exercice.

182. Lieu. *Une pyramide triangulaire* SABC *est coupée par un plan qui rencontre le plan de base suivant* LMN, *et détermine dans la pyramide une section* A'B'C'. *On fait tourner la section* A'B'C' *autour de l'axe* MN, *et l'on joint* AA', BB', CC'; *quel est le lieu décrit par le sommet de la pyramide ainsi obtenue* * ?

Par la hauteur SH menons un plan SHOR perpendiculaire à l'axe de rotation; ce plan détermine deux droites DE, D'E' dont il suffit d'étudier la position respective, car elles sont invariablement liées à la base et à la section. Le problème revient donc à une question connue de géométrie plane. *On demande le lieu décrit par le point de concours* S *des droites* DD', EE' (nº 84).

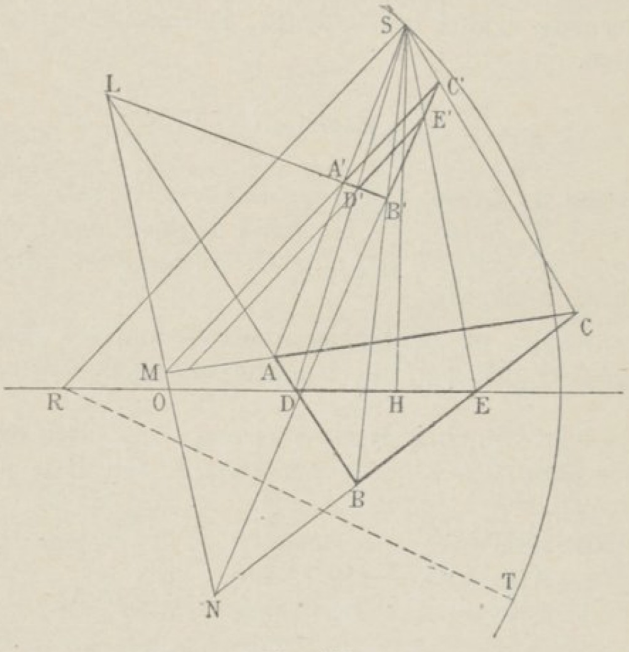

Fig. 115.

Le sommet S décrit une circonférence dont le plan est perpendiculaire à MN; R en est le centre et RS le rayon.

---

* Ce théorème a été indiqué par Poncelet dans l'étude de l'homologie; mais il est attribué ordinairement à Steiner, qui l'a formulé explicitement dans le *Journal de Crelle.*

La démonstration que nous donnons est très simple; néanmoins on lira avec fruit celle de A. Amiot, *Leçons nouvelles de géométrie élémentaire*, 2ᵉ édition, page 570.

A. Amiot, ancien professeur de mathématiques spéciales au lycée Saint-Louis, est surtout connu par les nombreux élèves qu'il a préparés pour l'École normale supérieure et pour l'École polytechnique. On lui doit divers ouvrages classiques, entre autres des *Eléments de géométrie* et les *Leçons nouvelles de géométrie descriptive.*

9 ./01. . .+ . 7 2 H 7 !
2 & -2 2!) "

3.6.3 Application : le théorème de Newton

3.6.5 Théorème. *Soit* a, b, c, d *un repère et soient* e *et* f *les intersections de* $(ab), (cd)$ *et* $(ad), (bc)$ *respectivement (voir fig. 3.7). Soit* D *une droite et* p, q, r *les intersections de* D *avec* (ca)*,* (bd) *et* (ef)*. On appelle* i, j, k *les conjugués harmoniques de* p, q, r *par rapport à* c, a *;* b, d *et* e, f*. Alors* i, j, k *sont alignés.*

Démonstration. On note u, v, w les intersections de $(bd), (ef)$; $(ef), (ac)$ et $(ac), (bd)$ respectivement et on applique la partie réciproque de 3.6.3 au triangle uvw. Il s'agit donc de montrer qu'on a :

$$\llbracket v, w, p, i \rrbracket \, \llbracket w, u, q, j \rrbracket \, \llbracket u, v, r, k \rrbracket = 1,$$

ou encore, en permutant les termes :

$$\llbracket i, p, w, v \rrbracket \, \llbracket j, q, u, w \rrbracket \, \llbracket k, r, v, u \rrbracket = 1.$$

Mais, on a le lemme suivant :

3.6.6 Lemme. *On a* $\llbracket i, p, w, v \rrbracket = \llbracket p, a, v, w \rrbracket^2 = \llbracket p, c, v, w \rrbracket^2$ *et les égalités analogues sur les droites* (bd) *et* (ef)*.*

Ce lemme implique le théorème car on a $\llbracket p, a, v, w \rrbracket \, \llbracket q, d, w, u \rrbracket \, \llbracket r, f, u, v \rrbracket = 1$ par application du sens direct de 3.6.3 au triangle uvw et aux points alignés p, q, r d'une part et a, d, f d'autre part.

Démonstration. (du lemme) On a $\llbracket a, c, p, i \rrbracket = -1$ par définition de i. Par ailleurs, la polaire de w par rapport aux droites (ad) et (bc) n'est autre que (fe) par construction de la polaire et on en déduit $\llbracket a, c, v, w \rrbracket = -1$. L'unique involution φ de (ac) admettant a et c pour points fixes (voir 3.5.13) échange donc i, p (resp. v, w) (c'est d'ailleurs l'involution de Desargues de (ac) associée au quadrilatère $jkqr$). L'application de φ donne l'égalité $\llbracket i, p, w, v \rrbracket = \llbracket p, i, v, w \rrbracket$. On en déduit, avec le lemme des trois birapports 3.6.1, $\llbracket i, p, w, v \rrbracket = \llbracket p, a, v, w \rrbracket \, \llbracket a, i, v, w \rrbracket = \llbracket p, a, v, w \rrbracket \, \llbracket a, p, w, v \rrbracket$ par application de φ et le lemme en résulte par permutation.

Le théorème originel de Newton correspond au cas où D est la droite de l'infini (voir fig. 3.8) :

3.6.7 Corollaire. (Newton) *Soit* a, b, c, d *quatre points du plan affine non trois à trois alignés. Soient* e *et* f *les intersections de* $(ab), (cd)$ *et* $(ad), (bc)$ *respectivement. Alors les milieux* i, j, k *de* c, a *;* b, d *et* e, f *respectivement, sont alignés.*

Nous reviendrons sur ce théorème dans les Parties II et III. Le lecteur qui voudrait en voir d'autres preuves pourra se reporter au livre de Jean-Louis Ayme (voir [Aym03]) où il en trouvera une trentaine !

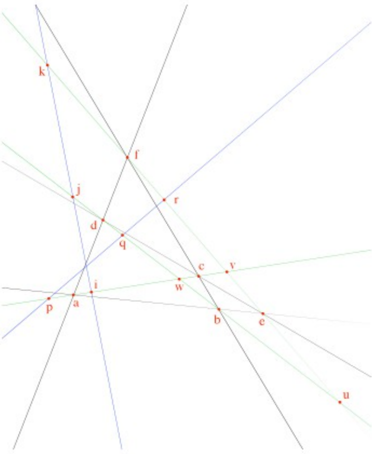

FIGURE 3.7 – Le théorème de Newton projectif

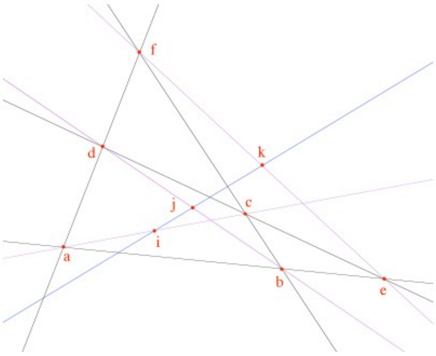

FIGURE 3.8 – Le théorème de Newton affine

3.6.4 Involutions et birapports

3.6.8 Lemme. *Soit D une droite projective et soient α, β, γ, α', β', γ' six points distincts de D. Soit $f : D \to D$ l'homographie définie par $f(\alpha) = \alpha'$, $f(\beta) = \beta'$, $f(\gamma) = \gamma'$. Alors f est une involution si et seulement si on a la relation suivante entre birapports :*

$$(*) \qquad [\![\alpha, \beta, \gamma, \gamma']\!]\, [\![\beta, \gamma, \alpha, \alpha']\!]\, [\![\gamma, \alpha, \beta, \beta']\!] = -1.$$

Démonstration. On vérifie que la relation $(*)$ équivaut à

$$\frac{\overline{\gamma'\beta}}{\overline{\gamma'\alpha}} \times \frac{\overline{\alpha'\gamma}}{\overline{\alpha'\beta}} \times \frac{\overline{\beta'\alpha}}{\overline{\beta'\gamma}} = 1,$$

ce qui, en multipliant haut et bas par $\overline{\gamma'\gamma}$, conduit à :

$$[\![\gamma', \alpha', \beta, \gamma]\!] = [\![\gamma', \beta', \gamma, \alpha]\!]^{-1} = [\![\gamma', \beta', \alpha, \gamma]\!] = [\![\gamma, \alpha, \beta', \gamma']\!].$$

Il en résulte que l'homographie qui envoie γ' sur γ, α' sur α, β sur β' envoie aussi γ sur γ'. C'est donc une involution. Mais alors elle envoie α sur α', donc cette involution n'est autre que f.

3.6.9 *Remarque.* On peut affaiblir l'hypothèse que les six points sont distincts. Il faut bien entendu supposer α, β, γ distincts, ainsi que α', β', γ' et supposer aussi que l'un des couples homologues par f est non trivial (par exemple $\gamma \neq \gamma'$), sinon on a $f = \mathrm{Id}$. Sous ces hypothèses, le résultat reste vrai avec la convention que si l'un des birapports est infini (par exemple si $\gamma' = \alpha$) alors l'un des autres est nul (et ici on a nécessairement $\alpha' = \gamma$).

3.6.5 Involutions et concours

3.6.10 Théorème. *Soient a, b, c trois points non alignés de $\mathbf{P}(E)$ et soient a', b', c' des points de (bc), (ca) et (ab) respectivement, distincts de a, b, c. Soit D une droite ne passant ni par les points a, b, c, a', b', c', ni par les points d'intersections de (aa'), (bb'), (cc'). La droite D coupe respectivement (bc), (ca), (ab), (aa'), (bb'), (cc') en des points $\alpha, \beta, \gamma, \alpha', \beta', \gamma'$.*
Alors les droites (aa'), (bb') et (cc') sont concourantes si et seulement si l'homographie de D qui envoie α sur α', β sur β', γ sur γ' est une involution.

Démonstration. Supposons (aa'), (bb'), (cc') concourantes en d. Alors le résultat n'est rien d'autre que le théorème de Desargues, cf. 3.5.6, appliqué au repère a, b, c, d.

Pour la réciproque on appelle d le point d'intersection de (aa') et (bb') et γ'' celui de D et de (cd). On applique encore Desargues au repère $abcd$. On en déduit une involution de D qui échange α et α', β et β' et γ et γ''. C'est nécessairement l'homographie de l'énoncé et on a $\gamma' = \gamma''$, ce qui signifie que d est sur (cc').

3.6.11 *Remarque.* C'est cet argument qui a été utilisé en 3.5.9 pour prouver le concours des hauteurs d'un triangle.

3.6.12 Corollaire. *Avec les notations ci-dessus, les droites (aa'), (bb') et (cc') sont concourantes si et seulement si on a la relation*

$$(*) \qquad \llbracket \alpha, \beta, \gamma, \gamma' \rrbracket \, \llbracket \beta, \gamma, \alpha, \alpha' \rrbracket \, \llbracket \gamma, \alpha, \beta, \beta' \rrbracket = -1,$$

ou encore la relation $(**)$: $\llbracket b, a, \gamma, c' \rrbracket \, \llbracket c, b, \alpha, a' \rrbracket \, \llbracket a, c, \beta, b' \rrbracket = -1$.

Démonstration. Le premier point vient de 3.6.8 (car les six points $\alpha, \beta, \gamma, \alpha', \beta', \gamma'$ sont distincts). Pour le second on montre successivement les égalités $\llbracket \alpha, \beta, \gamma, \gamma' \rrbracket = \llbracket b, a, \gamma, c' \rrbracket$, $\llbracket \beta, \gamma, \alpha, \alpha' \rrbracket = \llbracket c, b, \alpha, a' \rrbracket$ et $\llbracket \gamma, \alpha, \beta, \beta' \rrbracket = \llbracket a, c, \beta, b' \rrbracket$ en utilisant les perspectives de centres c, a, b.

3.6.6 Céva

Il s'agit de la variante usuelle de la relation $(**)$:

./01. . .+ . I 7 4J J H 7 &$(
2 (2! 2

data 02 02 2020

practice to replace complex figures with simpler ones.

Leipzig 1686 [312, v.7, pp. 326-329]

Leibniz also suggested the possibility of characterizing the curvature of a surface by means of an osculating sphere

↳ In a letter to Johann Bernoulli, dated July 29, 1698

111) ARQ-Important_05/Géométrie/Libre-Géométrie-Projective/Geometrie-Projective/

a) livre-exercices-de-geometrie-fgm-bnf.pdf

Pág. 78 (110)

177a. Théorème de Desargues

b) livre-geometrie-projective-daniel-perrin-328pages.pdf

Pág. 65 (100)

Théorème de Newton

Figure 3.7 Le théorème de Newton projectif

data 02 02 2020
1

Figure 3.8 Le Théorème de Newton affine.

Affine Adjetivo: allowing for or preserving parallel relationship

"still to be implemented are geometric primitives and affine transformation of image".

The simplest way to see why this assertion is true is to observe that any triangle in the plane can be mapped by an invertible affine transformation to an equilateral triangle.

Using parallelograms ~~from the previous~~ to define affine transformations may be a mixed blessing.

It can be obtained from the previous description by applying affine transformation to keep the polygon bounded as it degenerates.

Our interest is in the space of affine equivalence classes of equal-area polygons.

Jandaia

03.02.2020

112) Trigonometria Cubo*
Curvilínea Vectorial
Projetiva Musichiana.
↓ (TM)
Projetiva Diferencial Ortogonal
sobre a curva ciclóide.

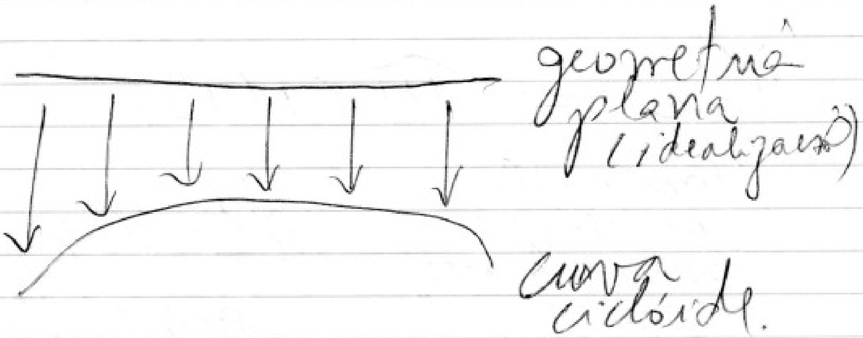
geometria plana (idealização)
curva ciclóide.

A velocidade maior entre 2 pontos ocorre no percurso de uma ciclóide e não em linha reta.
(espaço curvo → gravidade)
A luz segue o mesmo princípio.

A gravidade distorce o plano e a luz acompanha.

Veja: Grigori Perelman → Conjectura Poincaré
Einstein baseou-se em Poincaré e Lorentz.

data 03.02.2020

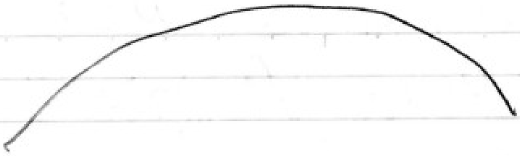

curvature

$\pi\!\!\!\pi > \pi$

(Mus) ✓ fator de conversão

$\pi \times c = \pi\!\!\!\pi\,(MUS)$

(constante curvilínea munichiana)

obs: Vejá Constante cosmológica
↓
expansão do universo (acelerando).

Veja:
GAUSS (item 110) e 111 (osculate)
179° 59' 59.32''

113) Mathematics: 20200208
Michael Atiyah →

' &

data 03.02.2020

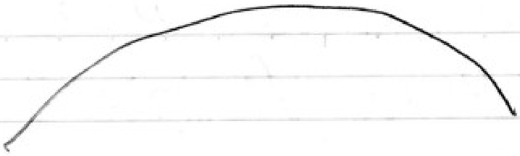
curvature

$\widetilde{\Pi} > \tilde{\pi}$

(Mus)

✓ fator de conversão

$\tilde{\pi} \times c = \widetilde{\Pi}$ (MUS)

(constante curvilínea munichiana)

obs: Veja Constante cosmológica
↓
expansão do universo (acelerando).

Veja:
GAUSS (item 110) e 111 (osculate)
$\underline{1}79° 59' 59.32''$

113) Mathematics: 20200208
Michael Atiyah →

data 08.02.2020

22 April 1929 > 89 anos.
11 January 2019 (+3 meses para 90 anos)

H. F. Baker

03 July 1866
17 March 1956 > 89 anos (+3 meses para 90 anos)

Meu nono também 14 Jan 1997 (faleceu)
(4 meses para 90 anos) (14)
Maio 1907. ↳ correto.

NONA - 20 Nov 1991.

Michael Atiyah →

A twisted cubic curve → subjects of Atiyah's first paper.

Michael Atiyah and Friedrich Hirzebruch, the creators of topological K-theory.

September 2018, at the 2018 Heidelberg Laureate Forum, he claimed a simple proof of the Riemann hypothesis, one of the most important and challenging problems in mathematics.

His claims have been met with skepticism by the mathematical community.

data 08.02.2020

Twisted cubic is a smooth, rational curve C of degree three in projective 3-space. P^3.

It is a fundamental example of a skew curve.

It is essentially unique, up to projective transformation.

Obs: Atiyah, Michael F (1988a)
Collected works
vol. 1 Early papers: general papers.

ISBN 978-0-19-853275-0

Source: (Book)

Data_Novouivo_30_2020/
Books_30_2020002[illegible]/
The Riemann Hypothesis
Michael Atiyah - Collected Works vol. 1. Early Papers. 1988.djvu
↳ A note on the tangents

08.02.2020

of a twisted cubic

Pag. 5(27)

obs: usou como referências:

a) Baker, H. F. Principles of geometry

(Cambridge, 1940), vol. 4, p. 52

Important:

Principles of Geometry Baker
6 volumes

Source: Data_Novonio_30_2020/
Books_30_2020/
The Riemann Hypothesis/
Principles of Geometry
Baker_6_Volumes//

Pág. 52(72)

the chords of a cubic curve in the original threefold space.

Veronese's surface

até

Pág. 55(75) Compare the theorem
(55) of Kronecker
and Castelnuovo
quoted in vol. III, p. 223.

data 08.02.2020

Volume III - Chapter IV
the Steiner quartic surface
Pág. 221 (245)

As the equation ~~which~~ ~~is dual~~ of the ~~tangent~~ plane of the cubic surface is

$$x_0^{-2}x + y_0^{-2}y + \dots = 0,$$

the tangential equation of this cubic surface is

$$l^{\frac{1}{2}} + m^{\frac{1}{2}} + n^{\frac{1}{2}} + p^{\frac{1}{2}} = 0$$

and the equation of the quartic surface, dual to the cubic surface, is

$$x^{\frac{1}{2}} + y^{\frac{1}{2}} + z^{\frac{1}{2}} + t^{\frac{1}{2}} = 0$$

this meets each of the planes

$$x = 0, \ y = 0 \text{ etc},$$

in a conic at every point of which it touches the plane, any two of these four conics having a point of contact, on

(osculate)

08.02.2020

the line of intersection of their planes.

...

Pág. 222 (246)

Ex. 3 The inflexional curves of the Steiner surface are rational quartic curves, having the three double lines of the surface as chords...

Pág. 223 (247)

Ex. 5 ... known as Veronese's surface.

Ex. 4 ... the inflexional curves of the surface represented by the equation

$$x^m + y^m + z^m + t^m = 0$$

satisfy the equation

$$(ax^m)^{\frac{1}{2}} + (by^m)^{\frac{1}{2}} + (cz^m)^{\frac{1}{2}} = 0$$

where a, b, c are any constants whose sum is zero.

Pág. 247 (223)

VERONESE'S SURFACE,

data 08.02.2020

see Cayley, Papers, $\underline{VI}$, p. 198*
(1868), and Bertini,
Geometria proiettiva d.
iperspazi, 1907, Chapter $\underline{XIV}$, $\underline{XV}$.)

See also Castelnuovo,
Atti d. r. Acc. d. Lincei, $\underline{III}$,
1894, p. 22,
who proves, after Kronecker,
that an irreducible algebraic
surface which is met by
∞^2 planes in reducible
curves, if not a
ruled surface,
is a Steiner quartic
surface.

Twisted Cubic

is most easily given
parametrically as the
image of the map
$\nu: P^1 \to P^3$
which assigns to the homogeneous
coordinates [S : T] the value
$\nu: [S:T] \to [S^3 : S^2T : ST^2 : T^3]$.

08.02.2020

In one coordinate patch of projective space, the map is simply the moment curve

$\nu : x \rightarrow (x, x^2, x^3)$

that is, it is closure by a single point at infinity of the affine curve (x, x^2, x^3).

Homogeneous coordinates
or
projective coordinates

Möbius in 1827 work

Der barycentrische Calcül.

→ points at infinity can be represented using ~~finite~~ coordinates.
(finite)

Formulas involving homogeneous coordinates are often simpler and more symmetric than their Cartesian counterparts.

in general, projective transformations to be easily represented by a matrix.

08 02 2020
1

Rational Bézier curve –

polynomial curve, defined in homogeneous coordinates (blue) and its projection on plane – rational curve (red).

Obs: Uso do LATEX.

Book (Source)

ARQ-Important/latex-books/
Books I important/
Guide to Latex.pdf
Helmut Kopka

Pág. 299 (312)

Curved lines

\qbezier[num]$(x_1, y_1)(x_2, y_2)(x_3, y_3)$

Ex:

```
\begin{picture}(40,20)
\qbezier(0,0)(20,20)(40,10)
\end{picture}
```

08.02.2020

real numbers → real projective spaces

complex numbers → complex projective space.

obs: the complex projective line uses two homogeneous complex coordinates and is known as the Riemann Sphere.

114) Research:

a) On Michael Atiyah and the Riemann Hypothesis

https://blogs.ams.org/blogonmathblogs/2018/10/01/on-michael-atiyah-and-the-riemann-hypothesis/

b) Michael Atiyah claims proof of Riemann Hypothesis

https://aperiodical.com/2018/09/michael-atiyah-claims-proof-of-riemann-hypothesis/

' % -# # $ %

./01. . .+ . % + / 5 ! 0
852 2

9 ./01. . . % / 5 2 0
852 C /2 ! 2

! <

'

6 *> 7 7 7 *&(* #D B *&(* #' 2 L #> 7 * K #

M ! N)! ! K 7 D B * * -*O2! !* (
+9
6 ! A N
D B * * (D B *& * J &&

D B * * - D B *&%* J %

./01. . .+ . 7 -2 2!

) 3 3 ! ! IP/ 5 2 Q K

NOUVELLES ANNALES
DE
MATHÉMATIQUES.

FORMULES FONDAMENTALES DE L'ANALYSE SPHÉRIQUE

(voir t. XVII, p. 307);

PAR M. VANNSON.

Propriétés des tangentes.

L'équation de la tangente sera

$$\frac{yy'}{b^2}+\frac{xx'}{a^2}=1.$$

Étant donnée une circonférence de grand cercle $y=\mathrm{A}x+\mathrm{B}$, si on demande la condition pour qu'elle soit tangente à l'ellipse, on écrira que les deux équations sont identiques, ce qui conduira à la condition

$$\mathrm{B}=\pm\sqrt{\mathrm{A}^2a^2+b^2}.$$

L'équation d'une tangente pourra donc se représenter par la formule

$$y=\mathrm{A}x\pm\sqrt{\mathrm{A}^2a^2+\mathrm{B}^2}.$$

Si on cherche la rencontre des deux lignes ainsi représentées, on trouvera qu'elle a lieu à 90 degrés de l'origine ou du centre en un point dont la latitude a pour tangente

A (*voyez* la tangente au cercle). Le signe + se rapporte à la tangente LC située au-dessus du diamètre LO, et le signe — à la tangente LD au-dessous de LO; LO a pour équation

$$y = mx.$$

Si on mène un rayon au point de contact C, et qu'on désigne par A′ le coefficient angulaire de ce rayon, on aura

$$A' = \frac{y'}{x'};$$

donc entre ce coefficient et celui A de la tangente on a la relation

$$AA' = -\frac{b^2}{a^2}.$$

Fig. 1.

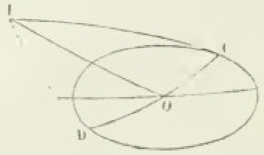

Soit OL le conjugué de OC et A″ son coefficient, on a aussi

$$A'A'' = -\frac{b^2}{a^2},$$

donc

$$A = A'';$$

d'où il résulte que la tangente et le diamètre OL se rencontrent à 90 degrés du centre en un point dont la latitude a pour tangente

$$A = -\frac{b^2 x'}{a^2 y'}.$$

On pourrait donc mener une tangente en C en menant le diamètre OC, déterminant son conjugué (nous avons vu le moyen plus haut), prenant sur ce conjugué OL un arc égal à 90 degrés, et joignant le point obtenu au point C.

On peut aussi se servir de la relation

$$X x' = a^2,$$

X étant la tangente du segment déterminé par la rencontre de la tangente avec l'axe.

Théorème. *La tangente divise en parties égales l'angle formé par un rayon vecteur mené du foyer au point de contact et le prolongement de l'autre rayon vecteur.*

Fig. 2.

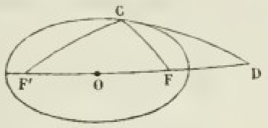

Il suffit pour cela de démontrer qu'on a la relation

$$\frac{\sin F'C}{\sin FC} = \frac{\sin F'D}{\sin FD}.$$

Or, nous avons trouvé précédemment

$$\operatorname{tang}\left(\frac{F'C + FC}{2}\right) = a \quad \text{et} \quad \operatorname{tang}\left(\frac{F'C - FC}{2}\right) = \frac{Cx'}{a}.$$

Ajoutant et soustrayant membre à membre, puis divisant, on trouve

$$\frac{\sin F'C}{\sin FC} = \frac{a^2 + Cx'}{a^2 - Cx'};$$

d'autre part nous avons trouvé

$$\operatorname{tang} OD = \frac{a^2}{x'};$$

on a d'ailleurs

$$\text{tang OF} = C;$$

opérant sur ces équations comme sur les précédentes, on a

$$\frac{\sin F'D}{\sin FD} = \frac{a^2 + Cx'}{a^2 - Cx'};$$

donc

$$\frac{\sin F'C}{\sin FC} = \frac{\sin F'D}{\sin FD}. \qquad \text{C. Q. F. D.}$$

Cela donne le moyen de mener par un point quelconque une tangente à l'ellipse, comme sur un plan.

Quand la courbe est rapportée à deux diamètres conjugués, l'équation de la tangente ne change pas de forme, c'est toujours

$$\frac{yy'}{b_{,}^2} + \frac{xx'}{a_{,}^2} = 1.$$

Fig. 3.

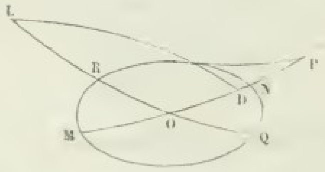

Si on cherche l'intersection de la tangente avec l'axe des x, on trouve

$$xx' = a_{,}^2;$$

d'où l'on voit que si l'on prolonge une tangente jusqu'à la rencontre d'un diamètre quelconque en P, et qu'on joigne le point du conjugué à 90 degrés de l'origine au point de contact, qu'on prolonge l'arc obtenu jusqu'au premier diamètre en D, on aura

$$\text{tang ON tang OP} = \text{tang}^2\,\text{ON}.$$

on a d'ailleurs

$$\text{tang}\,\text{OF} = \text{C};$$

opérant sur ces équations comme sur les précédentes, on a

$$\frac{\sin \text{F}'\text{D}}{\sin \text{FD}} = \frac{a^2 + \text{C}x'}{a^2 - \text{C}x'};$$

donc

$$\frac{\sin \text{F}'\text{C}}{\sin \text{FC}} = \frac{\sin \text{F}'\text{D}}{\sin \text{FD}}. \qquad \text{C. Q. F. D.}$$

Cela donne le moyen de mener par un point quelconque une tangente à l'ellipse, comme sur un plan.

Quand la courbe est rapportée à deux diamètres conjugués, l'équation de la tangente ne change pas de forme, c'est toujours

$$\frac{yy'}{b_{\prime}^2} + \frac{xx'}{a_{\prime}^2} = 1.$$

Fig. 3.

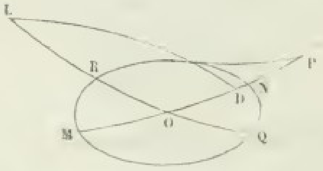

Si on cherche l'intersection de la tangente avec l'axe des x, on trouve

$$xx' = a_{\prime}^2;$$

d'où l'on voit que si l'on prolonge une tangente jusqu'à la rencontre d'un diamètre quelconque en P, et qu'on joigne le point du conjugué à 90 degrés de l'origine au point de contact, qu'on prolonge l'arc obtenu jusqu'au premier diamètre en D, on aura

$$\text{tang}\,\text{ON}\;\text{tang}\,\text{OP} = \text{tang}^2\,\text{ON}.$$

Prenons OCM $= 90°$. Joignons MA par un arc qui coupe en D l'arc OP. Nous aurons par le théorème précédent

$$\operatorname{tang} \mathrm{OD} \operatorname{tang} \mathrm{OP} = b_{,}^{2}.$$

Soit mené l'arc DS perperdiculaire sur OP, nous aurons

$$\frac{\operatorname{tang} \mathrm{OD}}{\operatorname{tang} \mathrm{OS}} = \frac{\operatorname{tang} \mathrm{OA}}{\operatorname{tang} \mathrm{OP}};$$

d'où

$$\operatorname{tang} \mathrm{OS} \operatorname{tang} \mathrm{OA} = \operatorname{tang} \mathrm{OD} \operatorname{tang} \mathrm{OP} = b_{,}^{2}.$$

On a donc

$$\operatorname{tang} \mathrm{OS} = \frac{b_{,}^{2}}{a},$$

ce qui détermine le point S.

Mais le lieu des sommets des triangles rectangles ayant pour hypoténuse OS est une ellipse dont nous avons déjà déterminé les axes. OS est l'un des axes et l'autre β se trouve par la formule

$$\sin \beta = \operatorname{tang} \frac{\mathrm{OS}}{2}:$$

la question revient donc à trouver la rencontre d'un grand cercle MA avec une ellipse dont on connaît les axes. (Problème déjà résolu.)

Discussion. Pour que le problème soit possible, il faut que le cercle MAD coupe l'ellipse. Or celle-ci a pour équation (résultat déjà trouvé)

$$x^2 + y^2 - \mathrm{A}x = 0,$$

A étant la tangente de OS$= \frac{b_{,}^{2}}{a}$. Le cercle MA passant par un point M de latitude α et par le point A a pour équation

$$y = \alpha'(x - a), \quad (\alpha' = \operatorname{tang} \alpha);$$

combinant ces deux équations, on trouve par l'élimination de y,

$$(\alpha_,^2 + 1)\, x^2 - (2a\alpha_,^2 + \mathrm{A})\, x + a^2\alpha_,^2 = 0.$$

Pour que les racines soient réelles, il faut avoir

$$\mathrm{A} > 2a \operatorname{tang}\left(45^\circ - \frac{\alpha}{2}\right);$$

d'où

$$b_,^2 > 2a^2 \operatorname{tang}\left(45^\circ - \frac{\alpha}{2}\right).$$

Ce qui donne le minimum de b'. Si on suppose pour b' sa valeur minimum, les racines sont égales et on trouve

$$x = a \sin\alpha';$$

d'où

$$y = -\alpha' a (1 - \sin\alpha').$$

Si on divise y par x, on aura la tangente de l'angle AOZ compris entre la direction du diamètre b' et l'axe de la courbe a. On trouve que cet angle $\mathrm{AOZ} = \frac{\alpha' + 90^\circ}{2} + 90^\circ$. Ce qui démontre que la direction du diamètre b', dans le cas du minimum, est bissectrice de l'angle formé par l'axe des y et OM′ prolongement de OM. De là résulte ce théorème :

Parmi toutes les ellipses qu'on peut construire sur un axe donné, il y en a une pour laquelle le conjugué d'un diamètre OM *de direction connue est minimum, et dans cette ellipse, la direction de ce conjugué divise en parties égales l'angle formé par le second axe et le prolongement du diamètre donné* OM.

Le même théorème a lieu sur un plan.

Théorème a démontrer. *Si aux extrémités* AB *d'un* *axe, on mène deux tangentes et qu'on les coupe par une troisième aux points* M *et* N, *qu'on projette* AM *et* BN *sur l'autre axe, le produit des tangentes des deux projections est égal à la tangente carrée de la moitié de cet axe, et si on multiplie* tang AM *par* tang BN, *on trouve la tangente carrée de ce même demi-axe multipliée par le cosinus carré du premier.*

La première partie de l'énoncé est encore vraie, quand on remplace les axes par un système de diamètres conjugués.

Ce théorème peut servir à résoudre le problème suivant :

On donne un des axes d'une ellipse et une tangente, trouver l'autre axe.

On peut aussi en tirer le corollaire :

Étant donné un triangle bi-rectangle ABC, si on coupe par un arc les côtés en M et P, de manière que le produit des tangentes des segments à partir de la base soit constante $= \tang^2 \varphi$, la courbe enveloppe de l'arc sécant est une ellipse sphérique, dont un axe est AB et dont l'autre s'obtient en prenant $AP = BM = \varphi$, et joignant PM, l'intersection de cet arc avec la perpendiculaire Oy donne l'extrémité du second axe.

Fig. 6.

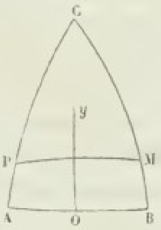

Théorème à démontrer. *Si on mène en un point* A *d'une ellipse une tangente, qu'on la prolonge jusqu'à la rencontre en M, N de deux diamètres conjugués pris pour axes, qu'on construise le demi-diamètre conjugué de celui qui passe au point A, la tangente carrée de la projection de ce demi-diamètre sur un des deux premiers égale le produit des tangentes des projections des deux segments de la tangente (il s'agit ici de projections obliques).*

Si au lieu de deux diamètres on prend les axes, et qu'on appelle b' la tangente du diamètre conjugué de celui qui passe au point de contact, le produit des tangentes trigonométriques des deux segments de la tangente égalera

$$b'^2 \cos^2 a \cos^2 b + \sin^2 a \sin^2 b.$$

Cette relation sert à résoudre le problème suivant :

Étant donnés les axes d'une ellipse non construite, lui mener une tangente telle, que le produit des tangentes des deux segments interceptés entre le point de contact et les axes soit égal à une quantité donnée P.

La relation donnera b'^2. Connaissant a, b et b', on pourra par une construction déjà faite trouver la direction du diamètre b', ainsi que la direction et la grandeur OA de son conjugué; il suffira alors de joindre le point A à un point pris sur b' à 90 degrés de l'origine.

Discussion. Pour que le problème soit possible, il ne suffit pas qu'on trouve pour b'^2 une valeur positive, il faut qu'on la trouve comprise entre a^2 et b^2. Ce qui conduit aux conditions

$$P < \sin^2 \alpha, \quad P > \sin^2 \beta,$$

α et β désignant les demi-axes. Dans le cas du maximum, la tangente sera menée par l'extrémité du petit axe et réciproquement.

Théorème. *Si des deux foyers d'une ellipse on abaisse des arcs perpendiculaires sur une tangente, le produit des sinus de ces arcs est constant et égal à* $\sin^2 \beta \cos^2 \gamma$, *$\gamma$ étant la distance du centre au foyer, et β le demi-axe perpendiculaire à l'axe des foyers.*

On sait qu'une tangente quelconque peut se représenter par l'équation

$$y = mx + \sqrt{a^2 m^2 + b^2}.$$

La distance δ du foyer à cette tangente est donnée d'après une formule ci-dessus démontrée par l'équation

$$\sin \delta = \frac{\sqrt{a^2 m^2 + b^2} + mc}{\sqrt{1 + b^2 + m^2 (1 + a^2)} \sqrt{1 + c^2}},$$

l'autre distance est donnée par l'équation

$$\sin \delta' = \frac{\sqrt{a^2 m^2 + b^2} - mc}{\sqrt{1 + b^2 + m^2 (1 + a^2)} \sqrt{1 + c^2}};$$

multipliant membre à membre et observant que

$$\cos a' = \cos b \cos \gamma,$$

on trouve

$$\sin \delta . \sin \delta' = \sin^2 b \cos^2 c.$$

Ce théorème peut servir à trouver l'axe β, connaissant les foyers et une tangente.

Application. On donne un foyer et deux tangentes,

Fig. 7.

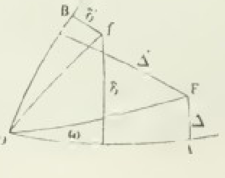

trouver le lieu du second foyer. Soient F le foyer donné, Δ et Δ' ses sinus distances aux deux tangentes OA, OB, f l'autre foyer, δ et δ' les sinus des deux distances.

On aura, par le théorème précédent,

$$\frac{\delta}{\delta'} = \frac{\Delta'}{\Delta};$$

mais

$$\frac{\delta}{\delta'} = \frac{\sin f\mathrm{OA}}{\sin(\theta - f\mathrm{OA})};$$

soit

$$f\mathrm{OA} = \alpha,$$

on aura

$$\frac{\delta}{\delta'} = \frac{\sin\alpha}{\sin(\theta - \alpha)};$$

soit β l'angle FOB, on aura de même

$$\frac{\Delta'}{\Delta} = \frac{\sin\beta}{\sin(\theta - \beta)}.$$

D'où on conclut aisément $\alpha = \beta$; donc le lieu du point f est une circonférence de grand cercle passant à l'origine et par un point symétrique de F relativement à la bissectrice.

Corollaire. Si d'un point O on mène deux tangentes à

Fig. 8.

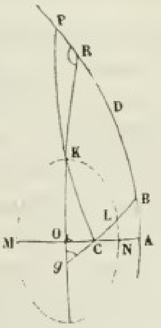

une ellipse et qu'on joigne le point O aux deux foyers par deux arcs de grands cercles, ces arcs feront avec les deux tangentes, des angles égaux. Si on demandait le lieu du centre qui sur un plan est une ligne droite parallèle au lieu du foyer, je dis qu'on trouverait une ellipse sphérique. Cela revient à faire voir que le lieu du milieu de l'arc variable CB terminé d'une part à un point fixe C, et d'autre part à sa rencontre B avec une circonférence AD, est une ellipse.

Posons

$$\mathrm{CB} = 2\rho, \quad \operatorname{tang} \mathrm{CA} = p \quad \text{et} \quad \mathrm{BCA} = \omega,$$

nous avons dans le triangle ABC,

$$p = \operatorname{tang} 2\rho \cos\omega \quad \text{ou} \quad 2 \operatorname{tang}\rho \cos\omega = p(1 - \operatorname{tang}\rho),$$

ou passant aux coordonnées rectangles,

$$1 - \frac{2x}{p} = y^2 + x^2.$$

C'est une ellipse dont on pourrait trouver les axes par son équation, mais on les obtient plus aisément par la géométrie. Soit N le point milieu de CA, N sera l'extrémité d'un axe. Prenons $\mathrm{NM} = \frac{\pi}{2}$, M sera l'autre extrémité, OK perpendiculaire sur MN sera la direction du second axe. Pour en avoir l'extrémité, construisons le le triangle OCg, de manière que l'angle $g = \mathrm{R}$. Soit K le milieu de gR, joignez CK, prolongez l'arc CK jusqu'en P, K sera, par suite d'une égalité de triangles, le milieu de CKP, donc K est un point du lieu, c'est un sommet. La courbe est alors facile à construire.

On voit donc que *quand on donne deux tangentes et un foyer, le lieu du centre est une ellipse.*

Problème. *Trouver le lieu des projections du foyer sur les tangentes.*

On peut prendre l'équation générale des tangentes

$$y = mx + \sqrt{a^2 m^2 + b^2},$$

et celle d'une perpendiculaire abaissée du foyer sur cette tangente, puis éliminer m entre ces deux équations ; mais le calcul est long et compliqué : on arrive beaucoup plus simplement au résultat par le moyen suivant : Soit MK une tangente en M ; joignons le point M aux foyers par les

Fig. 9.

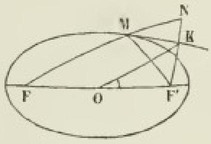

arcs MF, MF', prenons sur FM prolongé MN = MF', joignons NF' par un arc qui coupe la tangente en K. K est un point du lieu ; j'appelle ρ la distance OK, ω l'angle KOF', et p la distance F'K ou KN dans le triangle KOF', nous aurons

$$(1) \qquad \cos p = \cos c \cos \rho + \sin c \sin \rho \cos \omega,$$

et dans le triangle NFF',

$$(2) \qquad \cos 2\alpha = \cos 2c \cos 2\rho + \sin 2c \sin 2p \cos F' ;$$

on a encore dans ORF'

$$\cos F' = \frac{\cos \rho - \cos p \cos c}{\sin p \sin c}.$$

Je remplace $\cos F'$ par cette expression dans l'équation (2), puis j'élimine $\cos p$ par l'équation (1), on trouve ainsi l'équation

$$\sin^2 \rho (\cos^2 c + \sin^2 c \cos^2 \omega) = \sin^2 \alpha,$$

et passant aux coordonnées rectangles

$$\frac{y^2(\cos^2 c - \sin^2\alpha)}{\sin^2\alpha} + \frac{x^2}{a^2} = 1.$$

Le lieu demandé est donc une ellipse concentrique à la première ayant un axe commun 2α ; l'autre axe 2β est donné par la formule

$$\operatorname{tang}^2\beta = \frac{\sin^2\alpha}{\cos^2 e - \sin^2\alpha}.$$

Si le rayon de la sphère est infini, on trouve un cercle, car $\beta = \alpha$.

Fig. 10.

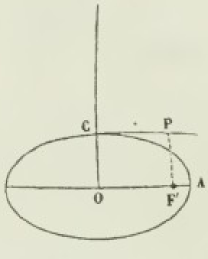

Quand on a reconnu que la courbe est du second degré, on peut trouver les axes sur la figure. Il est d'abord évident que OA est un axe de la courbe. Pour avoir l'autre, au sommet C menons une tangente CP et projetons le foyer F' sur cette tangente ; l'arc F'P projetant ira couper l'axe des y en un point D à $\frac{\pi}{2} + \beta$ de l'origine ($\beta =$ OC).

Donc on a pour le point P, $y' = -\frac{1}{b}$, l'équation de l'arc F'P sera donc

$$\frac{y}{-\frac{1}{b}} + \frac{x}{c} = 1$$

(c étant la tangente de OF'). Si dans cette équation nous faisons $y = b$, nous aurons les coordonnées du point D. Mais ces coordonnées doivent vérifier l'équation du lieu

$$\frac{y^2}{\beta'^2} + \frac{x^2}{a^2} = 1,$$

il est donc facile d'avoir β ou $\text{tang}\,\beta$; on trouve comme par l'autre méthode

$$\text{tang}^2\beta = \frac{\sin^2\alpha}{\cos^2 c - \sin^2\alpha}.$$

Si d'un point A, pris hors d'une ellipse sphérique, on mène deux tangentes et qu'on divise en parties égales les angles qu'elles font entre elles, les deux bissectrices viennent couper l'axe en deux points tels, que le produit des tangentes de leurs distances au centre est une constante c^2 égale au carré de la demi-distance des foyers. Les tangentes ont pour équations

$$y = mx + \sqrt{a^2 m^2 + b^2}, \quad y = m'x + \sqrt{a^2 m_1^2 + \beta^2}.$$

Les bissectrices ont pour équations

$$\frac{y - mx - \sqrt{a^2 m^2 + b^2}}{\sqrt{1 + b^2 + m^2(1 + a^2)}} = \frac{y - m'x - \sqrt{a^2 m_1^2 + b^2}}{\pm\sqrt{1 + b^2 + m_1^2(1 + a^2)}},$$

ou

$$r(y - mx - s) = r'(y - m'x - s'),$$

en appelant r et r' les inverses des dénominateurs et s, s', les deux radicaux du numérateur. Faisant $y = 0$, on trouve

$$x = \frac{rs - r's'}{m'r' - mr};$$

pour l'autre bissectrice, il faut changer le signe de r', ce qui donne

$$X = -\frac{rs + r's'}{m'r' + mr};$$

d'où on tire

$$
\begin{aligned}
\mathrm{X}x &= \frac{r_{,}^{2}s_{,}^{2} - r^{2}s^{2}}{m_{,}^{2}r_{,}^{2} - m^{2}r^{2}} = \frac{r_{,}^{2}(a^{2}m_{,}^{2} + b^{2}) - r^{2}(a^{2}m_{,}^{2} + b^{2})}{m_{,}^{2}r_{,}^{2} - m^{2}r^{2}} \\
&= a^{2} + \frac{b^{2}(r_{,}^{2} - r^{2})}{m_{,}^{2}r_{,}^{2} - m^{2}r^{2}} = a^{2} + \frac{b^{2}(1 + a^{2})(m^{2} - m_{,}^{2})}{m_{,}^{2}(1 + b^{2}) - m^{2}(1 + b^{2})} \\
&= a^{2} - \frac{\sin^{2}\beta}{\cos^{2}\alpha} = \frac{\sin^{2}\alpha - \sin^{2}\beta}{\cos^{2}\beta} = \mathrm{C}^{2},
\end{aligned}
$$

C désignant la tangente de la demi-distance des foyers.

FIG. 11.

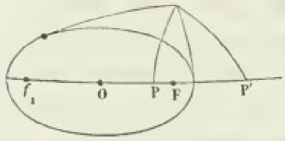

Il suit de là que si avec 2 C comme arc diamétral on traçait un cercle, les deux points P et P' seraient tels, que la polaire de l'un relative à ce cercle passerait par l'autre, donc les points de rencontre des bissectrices avec l'axe et les deux foyers sont quatre points harmoniques. Ce théorème peut servir à construire une ellipse connaissant le centre, la direction de l'axe et deux tangentes. On peut encore en conclure que si on donne deux tangentes et un foyer, le lieu de l'autre foyer est une circonférence de grand cercle.

PROBLÈME. *Trouver le lieu des sommets d'angles droits dont les côtés sont tangents à une ellipse.*

Soient $x'y'$, $x''y''$ les tangentes des coordonnées des points de contact, les deux tangentes en ces points auront pour équations

$$a^{2}yy' + b^{2}xx' = a^{2}b^{2},$$
$$a^{2}yy'' + b^{2}xx'' = a^{2}b^{2},$$

on a aussi

$$a^2 y_{,}^2 + b^2 x_{,}^2 = a^2 b^2,$$
$$a^2 y_{,,}^2 + b^2 x_{,,}^2 = a^2 b^2,$$

et pour que les deux tangentes soient rectangulaires, on a la condition connue

$$a'' y' y'' + b'' x' x'' + a'' b'' = 0,$$

il faut éliminer x', x'', y', y'' entre ces quatre équations. L'élimination se fait comme pour l'ellipse plane, et on trouve pour équation du lieu

$$(1 - a^2) y^2 + (1 - b^2) x^2 = a^2 + b^2.$$

Premier cas.

$$a^2 < 1, \quad b^2 < 1,$$

on a uneellipse sphérique. Si par les points C et B, sommets de la courbe donnée, on mène deux tangentes se coupant en D, le cercle concentrique à l'ellipse et passant en D aura pour équation

Fig. 12.

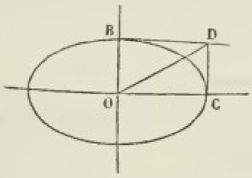

$$x^2 + y^2 = a^2 + b^2,$$

et comme dans la courbe trouvée on a

$$x^2 + y^2 > a^2 + b^2,$$

tous les points de cette courbe sont extérieurs au cercle (OD), par suite ils sont extérieurs à l'ellipse; on pourra donc par chacun d'eux mener deux tangentes, qui se couperont à angle droit.

Deuxième cas. Si on a

$$a^2 > 1 \quad \text{et} \quad b^2 > 1,$$

l'équation trouvée ne sera satisfaite par aucune valeur réelle de y et de x, en sorte que le problème de mener à l'ellipse deux tangentes rectangulaires sera dans ce cas impossible.

Troisième cas.

$$a^2 > 1, \quad b^2 < 1.$$

L'équation trouvée pourra s'écrire ainsi :

$$(a^2 - 1)y^2 - (1 - b^2)x^2 = -(a^2 + b^2),$$

on pourra l'appeler une *hyperbole sphérique*, et tous ses points conviendront encore à la question.

Quatrième cas.

$$a = 1, \quad b < 1.$$

La courbe donnée est une *parabole sphérique*, le lieu trouvé a pour équation

$$x = \pm\sqrt{\frac{1 + b^2}{1 - b^2}} = \pm\sqrt{\frac{1}{\cos 2\hat{c}}},$$

$2\hat{c}$ représentant le petit axe de la courbe donnée; cette équation représente deux circonférences perpendiculaires à l'axe de 90 degrés : il est aisé de reconnaître que ces circonférences sont les deux directrices de la parabole. Ainsi les directrices d'une parabole sphérique ont, comme sur le plan, la propriété de contenir tous les sommets des angles droits circonscrits à la courbe.

Si on construit une courbe dont les points soient les pôles des tangentes de la courbe donnée, on trouve une

FIG. 13.

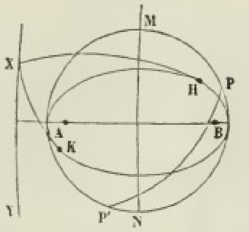

ellipse qu'on nomme ellipse *supplémentaire* de la première; elle a avec elle l'axe commun AB; le second axe MN est plus grand que AB. Cela posé, si par un point X de la directrice on mène les tangentes XH, XK, leurs pôles seront deux points P et P_1 de la courbe supplémentaire, l'arc PP' passera donc par le pôle de la directrice XY qui est le foyer de la parabole donnée, et il sera supplémentaire de l'angle X, c'est-à-dire de 90 degrés. On voit donc que si un ellipse a son petit axe de 90 degrés, tous les arcs de même longueur $\frac{\pi}{2}$, inscrits à la courbe, passeront par un point fixe qui sera le foyer de la parabole supplémentaire. (BORGNET.)

On résout, d'une manière analogue, le problème suivant : Trouver le lieu des intersections des tangentes menées par les extrémités de deux diamètres conjugués. L'équation est la même que sur le plan, savoir :

$$\frac{y^2}{(b\sqrt{2})^2}+\frac{x^2}{(a\sqrt{2})^2}=1.$$

Si par le centre O on mène une sécante coupant les deux courbes aux points A' et A, on aura

$$\operatorname{tang} OA' = \operatorname{tang} OA \,.\, \sqrt{2}.$$

De là résulte un moyen simple d'obtenir les axes. Menons

Fig. 14.

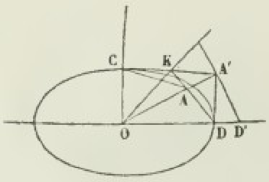

aux sommets D et C les tangentes DA', CA'; A' sera un point du lieu; joignons AD, et abaissons OK perpendiculaire sur AD, puis de A' abaissons aussi une perpendiculaire sur OK, cette perpendiculaire viendra couper OD en un point D' qui sera l'extrémité d'un des axes; l'autre sommet C' se trouve de même.

Propriétés des cordes supplémentaires.

Leur définition et la relation entre leurs coefficients angulaires sont comme pour l'ellipse plane. Ainsi en appelant ces coefficients α, α', on aura

$$\alpha\alpha' = -\frac{b^2}{a^2} \quad \text{ou} \quad -\frac{b_{\prime}^2}{a_{\prime}^2},$$

selon que la courbe sera rapportée à ses axes ou à un système de diamètres conjugués. Or on a la même relation entre les coefficients angulaires β, β' d'une tangente et du

Fig. 15.

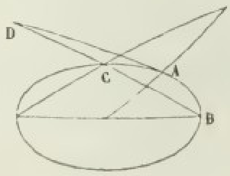

rayon mené au point de contact; si donc $\alpha = \beta$, on aura

$$\alpha' = \beta',$$

ce qui démontre que quand une tangente rencontre une corde CB à 90 degrés du centre, la corde supplémentaire de CB rencontre aussi à 90 degrés du centre le rayon mené au point de contact.

Cela donne un moyen facile de mener une tangente par un point donné sur la courbe ou par un point situé à 90 degrés du centre.

Si on appelle γ le coefficient angulaire d'un grand cercle et γ' le coefficient du rayon qui joint son pôle au centre, on a aussi

$$\gamma\gamma' = -\frac{b^2}{a^2}.$$

Si donc $\alpha = \gamma$, on aura

$$\alpha' = \gamma',$$

c'est-à-dire que si une corde CB rencontre à 90 degrés du centre un grand cercle, la corde supplémentaire de CB rencontrera aussi à 90 degrés du centre l'arc allant du centre au pôle de ce cercle. Si donc on veut construire la polaire d'un point A, on joindra OA, on prendra sur OA prolongée $OM = 90^\circ$, on joindra M à l'extrémité L de l'axe ou d'un autre diamètre par un arc coupant l'ellipse en B. On mènera CB corde supplémentaire de BL, on prendra sur CB un point N à 90 degrés du centre, et le point N

Fig. 16.

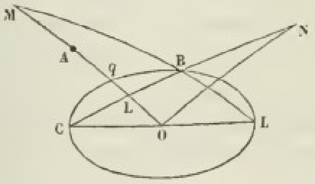

sera un point de la polaire de A. Il sera facile d'en trouver un second par la relation connue

$$\text{tang}^2\, \text{O}q = \text{tang OA}\,.\,\text{tang OL}.$$

On démontrera aussi facilement que si un diamètre et une corde se coupent à 90 degrés du centre, le conjugué et la corde supplémentaire se coupent à la même distance du centre, ce qui servira à construire deux diamètres conjugués dont l'un passe par un point donné.

SUR L'INTERPOLATION (*);

PAR M. EUGÈNE ROUCHÉ,
Professeur au lycée Charlemagne, Docteur ès Sciences mathematiques.

I.

1. Dans l'un de ses intéressants articles sur divers points du Cours de Mathématiques spéciales, M. Gerono a indiqué un moyen de déduire la formule d'interpolation de Newton de celle de Lagrange.

Je me propose de donner de ce même problème une solution assez facile et assez courte pour être introduite dans les Éléments. Toutefois, je ne serais peut-être pas revenu sur cette question, sans une circonstance particulière, qui se présente d'ailleurs assez fréquemment en mathématiques.

Lorsqu'on aborde le problème dont je viens de parler, il suffit d'un coup d'œil pour voir que la formule de Lagrange, dans sa forme habituelle, ne se prête pas aisément à la transformation demandée. Une préparation préalable est nécessaire pour éviter de longs calculs. En cherchant à grouper convenablement les termes, j'ai été

(*) D'après une thèse récemment soutenue dont nous rendrons compte.

9 ./01. . .+ . & 7 2 2!
) "

NOUVELLES ANNALES
DE
MATHÉMATIQUES.

MÉTHODE ÉLÉMENTAIRE POUR CALCULER LE RAPPORT DE LA CIRCONFÉRENCE AU DIAMÈTRE;

PAR M. L. MALEYX.

En étudiant la méthode inventée par Archimède pour déterminer une valeur approchée du rapport de la circonférence au diamètre, j'ai été conduit à faire deux remarques principales.

La première consiste en ce que Archimède, en doublant le nombre des côtés d'un polygone régulier, fait varier à la fois le périmètre du polygone et les diamètres des cercles inscrit et circonscrit; ce fait distingue sa méthode de celle dite *des périmètres*, où l'on suppose que, le périmètre du polygone variant, le rayon du cercle circonscrit reste fixe, et de celle dite *des polygones isopérimètres*, où l'on suppose que le périmètre reste fixe. S'il existe une constante dans sa méthode, cette constante est le côté du polygone, et encore ne l'est-elle pas d'une manière absolue.

La deuxième remarque consiste en ce que l'application de ses principes peut être rendue plus simple que celle des principes des deux autres méthodes que je viens de citer.

J'ai fondé sur ces remarques une étude que je me propose de publier ultérieurement. et dont j'extrais ce qui suit.

I. *Un côté de l'angle droit d'un triangle rectangle est représenté par* 1, *son hypoténuse par* D *et le second côté par* d, *l'angle opposé au premier côté l'étant par* α; *on se propose de calculer le second côté de l'angle droit* d_1 *et l'hypoténuse* D_1 *d'un second triangle rectangle dont le premier côté est aussi représenté par* 1, *et opposé à un angle égal à* $\frac{\alpha}{2}$.

Soit (*fig.* 1) ABC le premier triangle rectangle dans lequel on a

$$A = 1^{\text{dt}}, \quad ABC = \alpha, \quad AC = 1, \quad BC = D, \quad AB = d.$$

Pour construire le second, il suffira de prolonger AB de la longueur $BB_1 = BC = D$, et d'unir par une ligne droite le point B_1 au point C; car, le triangle CBB_1 étant isoscèle par construction, on a

Fig. 1.

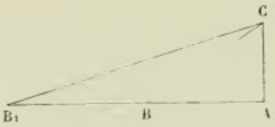

$$BB_1C = BCB_1 = \tfrac{1}{2}ABC.$$

On a du reste évidemment

$$AB_1 = d_1 = D + d$$

et, d'après le théorème de Pythagore,

$$B_1C = D_1 = \sqrt{d_1^2 + 1}.$$

II. *Extraction de la racine carrée d'un nombre donné sous la forme* $a^2 + R$, *d'après les données* a *et* R, *et sans effectuer le carré* a^2.

Prenons a pour valeur approchée par défaut de la racine cherchée, et soit $\frac{b}{10^\alpha}$ la valeur aussi approchée par défaut, et à moins de $\frac{1}{10^\alpha}$ de la partie complémentaire de la racine, on a

$$\left(a+\frac{b}{10^\alpha}\right)^2 \leq (a^2+\mathrm{R}):$$

développant le premier membre et réduisant,

$$2a\times\frac{b}{10^\alpha}+\frac{b^2}{10^{2\alpha}}\leq\mathrm{R},$$

d'où

$$\frac{b}{10^\alpha}\leq\frac{\mathrm{R}}{2a}.$$

Nous pouvons développer le quotient $\frac{\mathrm{R}}{2a}$ en fraction décimale, à moins de $\frac{1}{10^\alpha}$ par défaut, et, si $\frac{q}{10^\alpha}$ représente ce développement, R_1 étant le reste de la division, nous aurons

$$\mathrm{R}=2a\times\frac{q}{10^\alpha}+\mathrm{R}_1$$

et, en conséquence,

$$\frac{b}{10^\alpha}\leq\frac{q}{10^\alpha}+\frac{\mathrm{R}_1}{2a},\qquad\frac{\mathrm{R}_1}{2a}<\frac{1}{10^\alpha}.$$

$\frac{b}{10^\alpha}$, étant un multiple de $\frac{1}{10^\alpha}$ contenu dans le second membre de l'inégalité précédente, ne peut surpasser $\frac{q}{10^\alpha}$, qui est le plus grand multiple de $\frac{1}{10^\alpha}$ contenu dans le même nombre ; $\frac{q}{10^\alpha}$ ne peut donc être que supérieur ou égal à $\frac{b}{10^\alpha}$, et, pour qu'il ne lui soit pas supérieur, il suffit que du nombre donné, $a^2+\mathrm{R}$, on puisse retrancher le carré de $a+\frac{q}{10^\alpha}$, c'est-à-dire qu'on ait l'inégalité

$$\left(a+\frac{q}{10^\alpha}\right)^2\leq a^2+\mathrm{R}$$

ou

$$2a \times \frac{q}{10^{\alpha}} + \frac{q^2}{10^{2\alpha}} \leq R$$

ou encore

$$\frac{q^2}{10^{2\alpha}} \leq R_1.$$

Si donc on développe le quotient de la division de R par $2a$ en décimales, tant que le carré de la partie trouvée au quotient ne surpassera pas le reste de la division, cette partie sera égale à la partie complémentaire de la racine, à moins d'une unité du dernier ordre, et par défaut.

Supposons que nous nous arrêtions dans la division de R par $2a$ à la partie $\frac{q}{10^{\alpha}}$ du quotient, $\frac{q^2}{10^{2\alpha}}$ ne surpassant pas le reste R_1 de la division, et désignons par R_2 l'excès de R_1 sur $\frac{q^2}{10^{2\alpha}}$; nous aurons successivement

$$\frac{q^2}{10^{2\alpha}} + R_2 = R_1,$$

$$\frac{q^2}{10^{2\alpha}} + R_2 = R - 2a \times \frac{q}{10^{\alpha}};$$

d'où, ajoutant aux deux membres $a^2 + 2a \times \frac{q}{10^{\alpha}}$,

$$\left(a + \frac{q}{10^{\alpha}}\right)^2 + R_2 = a^2 + R.$$

On pourra alors reprendre le calcul précédent en traitant $a + \frac{q}{10^{\alpha}}$ et R_2 de la même manière que a et R, et continuer ainsi jusqu'à ce que l'on ait obtenu une valeur suffisamment approchée de la racine cherchée.

Observations. — On pourra encore abréger le calcul en ne conservant que la partie utile des produits qu'on devra successivement retrancher dans les divisions qu'on aura à faire, à condition de tenir compte à la fin du calcul des erreurs commises sur ces produits.

III. Pour ne rien laisser à désirer, donnons un exemple de ce calcul : soit à extraire la racine carrée de la somme $(3,732050807568 + \alpha)^2 + 1$, α étant un nombre positif inconnu, mais moindre que $\frac{1}{10^{12}}$. Agissons comme si α était nul, nous en tiendrons compte ultérieurement; commençons la division de 1 par le double de la partie entre parenthèses diminuée de α, ne conservant dans chaque produit que la partie représentant des unités de l'ordre $\frac{1}{10^{13}}$:

```
1,0000000000000000 | 7,464101615136
   7464101815136    |---------------
 ----------------   | 0,13
   2535898384864o
   2239230484540 8
   --------------
    29666790032 32
```

Retranchons du second reste le carré de la partie trouvée au quotient $(0,13)^2 = 0,0169$, puis divisons le résultat de cette soustraction par la somme du diviseur précédent et du double de la partie trouvée au quotient, 0,26 :

```
0,0296667900323 2
0,0169
---------
0,01276679003232o | 7,724101615136
     7724101615136 |---------------
    -------------- | 0,001652
     5042688417184
     4634460969078
     -------------
      408227448106
      386205080755
      ------------
       22022367351
       15448203230
       -----------
        6574164121
```

Opérons sur le nouveau quotient comme sur le précédent, modifions de même le diviseur et poursuivons :

$$
\begin{array}{r|l}
0{,}000006574164121 & \\
(0{,}001652)^2 = 0{,}000002729104 & \\
\hline
3845060121 & 7{,}727405615136 \\
\cline{2-2}
3090962244 & 0{,}000000497587 \\
\hline
754097877 & \\
695466504 & \\
\hline
58631373 & \\
54091835 & \\
\hline
4539538 & \\
3863700 & \\
\hline
675838 & \\
618192 & \\
\hline
57646 & \\
54089 & \\
\hline
3557 &
\end{array}
$$

Dans les opérations que nous venons d'exécuter, les produits que nous avons successivement retranchés du dividende initial sont trop faibles : il en résulte que le reste final est trop grand de la somme des parties négligées dans chaque produit ; il est facile de former une limite supérieure de la somme de ces parties négligées.

Nous avons d'abord négligé le produit de 2α (partie complémentaire du diviseur) par le nombre formé des trois premiers chiffres des quotients obtenus, soit $0{,}131$; cette première partie négligée a une valeur moindre que $0{,}131 \times \frac{2}{10^{12}} = \frac{131 \times 2}{10^{15}}$.

En second lieu, dans chacun des produits suivants, la partie négligée n'atteint pas un nombre d'unités de l'ordre $\frac{1}{10^{15}}$ marqué par le chiffre employé au multiplicateur ; donc la somme des parties négligées n'atteint pas le produit

$$\frac{1}{10^{15}} \times (131 \times 2 + 6 + 5 + 2 + 4 + 9 + 7 + 5 + 8 + 7),$$

et, si nous désignons cette somme par E, nous aurons

$$E < \frac{315}{10^{15}}.$$

De là résulte que, si nous désignons par R le reste final exact, nous avons les inégalités

$$\frac{3557 - 315}{10^{15}} = \frac{3242}{10^{15}} < R < \frac{3557}{10^{15}}.$$

Le dernier quotient trouvé ne pourrait être que trop grand, puisque nous avons opéré sur des dividendes trop grands, et il ne l'est pas, en tant que quotient, puisque son produit par le diviseur retranché du dividende fournit le reste R qui est positif; il n'est pas trop grand non plus, en tant que partie complémentaire de la racine, puisque son carré est inférieur à $\frac{25}{10^{14}}$ qui est lui-même certainement inférieur à R.

Donc la partie complémentaire de la racine carrée cherchée est, à moins de $\frac{1}{10^{12}}$ par défaut,

$$0,131652497587,$$

et la racine carrée est elle-même :

$$3,863703305155,$$

à moins de $\frac{2}{10^{12}}$ par défaut.

IV. L'exemple sur lequel nous venons d'appliquer le procédé d'extraction de la racine carrée exposé au n° II n'a pas été pris au hasard; si dans le triangle rectangle ABC, considéré au n° I, nous supposons

$$BC = D = 2,$$

nous aurons

$$AB = d = \sqrt{3} \quad \text{et} \quad \alpha = 30^\circ.$$

Or $\sqrt{3}$ est égale à 1,732050807568, à moins de $\frac{1}{10^{12}}$ par défaut; dès lors, et par application des formules du n° I, nous aurons, en partant des données,

$$\begin{array}{llll}
D = 2, & & & \\
d = \sqrt{3} = 1,732050807568 & \text{à moins de} & \frac{1}{10^{12}} & \text{par défaut}, \\
d_1 = \quad 3,732050807568 & \text{»} & \frac{1}{10^{12}} & \text{»} \\
D_1 = \quad 3,863703305155 & \text{»} & \frac{2}{10^{12}} & \text{»}
\end{array}$$

l'angle compris entre D_1 et d_1 étant de 15°.

On voit du reste immédiatement, en unissant le point A au point milieu de BC, ou à celui de B_1C, que D et d sont les diamètres des cercles circonscrit ou inscrit à l'hexagone régulier dont le côté est 1, et que D_1 et d_1 sont les diamètres des cercles circonscrit et inscrit au dodécagone régulier ayant aussi pour côté 1.

Reprenant deux fois le même calcul à partir de D_1 et d_1, nous nous procurerons de même les diamètres D_2 et d_2, D_3 et d_3, des cercles circonscrit et inscrit aux polygones réguliers de vingt-quatre et quarante-huit côtés, respectivement, et ayant chacun pour côté 1.

V. Les détails dans lesquels nous sommes entrés à l'occasion de ces calculs, dans le n° III, doivent nous dispenser de les rapporter *in extenso :* nous n'en donnerons donc que les résultats.

Toutefois, avant de les inscrire, nous ferons la remarque suivante, qui nous conduira à une notation utile : on voit que, dans chaque calcul, tel que celui du n° III, on ne fait que déterminer la différence qui existe entre les diamètres des cercles circonscrit et inscrit à un même polygone régulier dont le côté est 1; nous désignerons donc, d'une façon générale, par D_k et d_k les diamètres des cercles circonscrit et inscrit au polygone régulier dont le côté est 1, et dont le nombre des côtés est 6×2^k, et par Δ_k la différence $D_k - d_k$.

Cela posé, les résultats trouvés sont caractérisés par les égalités suivantes :

$$d_2 = D_1 + d_1 = 7{,}595754112723 \quad \text{à moins de} \quad \frac{3}{10^{12}} \quad \text{par défaut.}$$

$$\Delta_2 = 0{,}065543462815 \quad \text{»} \quad \frac{1}{10^{12}} \quad \text{»}$$

$$D_2 = d_2 + \Delta_2 = 7{,}661297575538 \quad \text{»} \quad \frac{1}{10^{12}} \quad \text{»}$$

$$d_3 = D_2 + d_2 = 15{,}257051688261 \quad \text{»} \quad \frac{7}{10^{12}} \quad \text{»}$$

$$\Delta_3 = 0{,}032736610412 \quad \text{»} \quad \frac{1}{10^{12}} \quad \text{»}$$

$$D_3 = d_3 + \Delta_3 = 15{,}289788298673 \quad \text{»} \quad \frac{8}{10^{12}} \quad \text{»}$$

VI. La partie la plus pénible de ces calculs est évidemment la détermination des nombres Δ : aussi est-on conduit à rechercher s'il n'existe pas quelque propriété géométrique de la figure qui nous permette d'atteindre plus facilement ce résultat.

Reprenons la figure du n° I et ajoutons y la circonférence décrite de B comme centre avec $BC = D$ pour rayon, elle coupe AB en B_1 et en E, joignons CE par une ligne droite; conservant les notations des numéros précédents, nous aurons

$$BE - BA = D - d = AE = \Delta.$$

Dans le triangle ACE (*fig.* 2), l'angle ACE est égal à l'angle

$$BB_1C = \frac{\alpha}{2},$$

car ils ont leurs côtés respectivement perpendiculaires et sont tous les deux aigus ; donc, pour avoir la différence $AE = \Delta$, il suffira de construire CE faisant avec CA l'angle $ACE = \frac{1}{2}ABC$; de même, pour avoir la différence

Fig. 2.

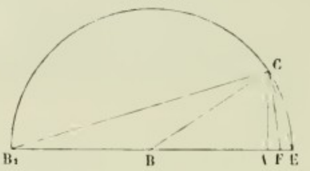

$$AF = B_1C - B_1A = D_1 - d_1 = \Delta_1,$$

il suffira de construire la droite CF faisant avec CA l'angle $ACF = \frac{1}{2}AB_1C$. On conclut de cette construction que CF est bissectrice de l'angle ACE et de la propriété connue de cette droite, l'égalité de rapports

$$\frac{\Delta - \Delta_1}{\Delta_1} = \frac{CE}{CA} = \sqrt{\Delta^2 + 1}.$$

Élevons les deux membres au carré et ramenons à la forme entière

$$\Delta^2 - 2\Delta\Delta_1 + \Delta_1^2 = \Delta_1^2(\Delta^2 + 1).$$

Réduisant et transposant,

$$\Delta\Delta_1^2 + 2\Delta_1 - \Delta = 0.$$

Cette équation du second degré définit Δ_1 au moyen de Δ par sa racine positive ; mais, comme Δ a généralement une petite valeur, on peut calculer Δ_1 au moyen d'une suite d'inégalités que nous allons déduire de l'équation même, et sans extraction de racine carrée.

Nous pouvons donner à l'équation précédente la forme

$$\Delta_1 = \tfrac{1}{2}\Delta - \tfrac{1}{2}\Delta\Delta_1^2 : \tag{1}$$

d'où l'inégalité évidente

$$\Delta_1 < \tfrac{1}{2}\Delta.$$

Remplaçons dans l'équation (1), et au second membre seulement, Δ_1 par le nombre plus grand $\frac{1}{2}\Delta$; nous aurons

$$\Delta_1 > \tfrac{1}{2}\Delta - \tfrac{1}{8}\Delta^3.$$

Ainsi, en prenant pour valeur approchée par excès de Δ_1 la valeur $\frac{1}{2}\Delta$, l'erreur commise a pour limite supérieure $\frac{1}{8}\Delta^3$.

Si nous remplaçons maintenant au second membre de l'équation (1) Δ_1 par le nombre plus petit $\frac{1}{2}\Delta - \frac{1}{8}\Delta^3$, nous aurons, en négligeant un terme soustractif, ce qui ne fait qu'augmenter l'inégalité,

$$\Delta_1 < \tfrac{1}{2}\Delta - \tfrac{1}{8}\Delta^3 + \tfrac{1}{16}\Delta^5.$$

Dès lors, en prenant pour valeur approchée par défaut de Δ_1 la différence $\frac{1}{2}\Delta - \frac{1}{8}\Delta^3$, l'erreur commise sera inférieure à $\frac{1}{16}\Delta^5$.

Enfin, en substituant à Δ_1 dans le second membre de l'équation (1) la valeur trop grande $\frac{1}{2}\Delta - \frac{1}{8}\Delta^3 + \frac{1}{16}\Delta^5$, nous aurons, en négligeant des termes dont la somme est certainement positive,

$$\Delta_1 > \tfrac{1}{2}\Delta - \tfrac{1}{8}\Delta^3 + \tfrac{1}{16}\Delta^5 - \tfrac{5}{16}\Delta^7 ;$$

d'où l'on conclut que si, dans la suite

$$\tfrac{1}{2}\Delta - \tfrac{1}{8}\Delta^3 + \tfrac{1}{16}\Delta^5 - \tfrac{5}{16}\Delta^7.$$

on prend un, deux ou trois termes comme valeur approchée de Δ_1, cette valeur est approchée par excès ou par défaut suivant qu'on s'arrête à un terme affecté du signe + ou du signe —, et que, dans tous les cas, la valeur absolue du premier terme négligé est limite supérieure de l'erreur commise.

Ces inégalités permettent de calculer un des nombres Δ au moyen du précédent avec une assez grande approximation : on pourrait même l'obtenir avec une approximation illimitée par l'emploi de l'égalité (1) et par la méthode des approximations successives ; mais leur emploi est encore pénible, eu égard aux multiplications à faire : aussi ne ferons-nous usage dans ce qui suit que des inégalités

$$\tfrac{1}{2}\Delta - \tfrac{1}{8}\Delta^3 < \Delta_1 < \tfrac{1}{2}\Delta,$$

et nous allons nous occuper d'établir des formules se déduisant de l'équation (1) et permettant d'arriver au même résultat au moyen d'opérations plus simples.

VII. Reprenons l'égalité initiale réduite du n° VI

$$\Delta - 2\Delta_1 - \Delta\Delta_1^2 = 0, \tag{1}$$

d'où

$$\Delta = \frac{2\Delta_1}{1 - \Delta_1^2}. \tag{1'}$$

Appliquons-la aux différences d'ordre supérieur de une unité

$$\Delta_1 - 2\Delta_2 - \Delta_1\Delta_2^2 = 0, \tag{2}$$

d'où

$$\Delta_1 = \frac{2\Delta_2}{1 - \Delta_2^2}. \tag{2'}$$

Δ étant peu différent de $2\Delta_1$, et Δ_1 l'étant peu de $2\Delta_2$, la valeur numérique du dernier terme du premier membre de l'égalité (1) est sensiblement égale à huit fois celle du terme correspondant de l'égalité (2); retranchant les égalités (1) et (2) membre à membre après avoir multiplié la seconde par 8, on a

$$\Delta - 10\Delta_1 + 16\Delta_2 - \Delta_1(\Delta\Delta_1 - 8\Delta_2^2) = 0. \tag{3}$$

Transformons le dernier terme du premier membre de l'égalité (3); si nous multiplions membre à membre l'égalité (1)' et l'égalité obtenue en élevant au carré les deux membres de l'égalité (2)', nous aurons

$$\Delta\Delta_1 = \frac{8\Delta_2^2}{(1-\Delta_1^2)(1-\Delta_2^2)^2}.$$

Remplaçons au dénominateur Δ_1 par sa valeur tirée de l'égalité (2)'

$$\Delta\Delta_1 = \frac{8\Delta_2^2}{(1-\Delta_2^2)^2 - 4\Delta_2^2}.$$

De là on déduit

$$\begin{aligned}\Delta\Delta_1 - 8\Delta_2^2 &= 8\Delta_2^2\left[\frac{1}{(1-\Delta_2^2)^2 - 4\Delta_2^2} - 1\right]\\ &= 8\Delta_2^2 \times \frac{6\Delta_2^2 - \Delta_2^4}{(1-\Delta_2^2)^2 - 4\Delta_2^2}.\end{aligned}$$

En substituant au dénominateur $\frac{2\Delta_2}{\Delta_1}$ à $1 - \Delta_2^2$ qui lui est égal, d'après l'égalité (2)',

$$\Delta\Delta_1 - 8\Delta_2^2 = 8\Delta_2^2 \times \frac{6\Delta_2^2 - \Delta_2^4}{\frac{4\Delta_2^2}{\Delta_1^2} - 4\Delta_2^2} = 8\Delta_2^2\Delta_1^2 \times \frac{6-\Delta_2^2}{4(1-\Delta_1^2)}.$$

Remplaçons enfin au dénominateur de la dernière fraction $1 - \Delta_1^2$ par sa valeur $\frac{2\Delta_1}{\Delta}$ déduite de l'égalité (1)', on a, après réduction,

$$\Delta\Delta_1 - 8\Delta_2^2 = \Delta\Delta_1\Delta_2^2(6 - \Delta_2^2):$$

d'où, reportant dans l'égalité (3), on a

$$\Delta - 10\Delta_1 + 16\Delta_2 - \Delta\Delta_1^2\Delta_2^2(6 - \Delta_2^2) = 0. \tag{α}$$

De l'égalité (*a*) on peut déduire

$$\Delta_2 = \frac{1}{16}(10\Delta_1 - \Delta) + \frac{(6-\Delta_2^2)\Delta\Delta_1^2\Delta_2^2}{16}.$$

Le premier terme du second membre fournit une valeur approchée par défaut de Δ_2, et le second terme qui représente l'erreur est moindre que

$$\frac{6\Delta\left(\frac{\Delta}{2}\right)^2\left(\frac{\Delta}{4}\right)^2}{16} = \frac{3}{512}\Delta^5 < \frac{\Delta^5}{170}.$$

VIII. Reprenons encore l'égalité (*a*) du numéro précédent, et appliquons-la aux différences d'indice supérieur d'une unité :

$$(a) \qquad \Delta - 10\Delta_1 + 16\Delta_2 - \Delta\Delta_1^2\Delta_2^2(6-\Delta_1^2) = 0.$$

$$(b) \qquad \Delta_1 - 10\Delta_2 + 16\Delta_3 - \Delta_1\Delta_2^2\Delta_3^2(6-\Delta_3^2) = 0.$$

Retranchons membre à membre de l'égalité (*a*) l'égalité (*b*) dont les deux membres ont été précédemment multipliés par 32; on en déduit

$$(c) \qquad \left\{\begin{aligned} &\Delta - 42\Delta_1 + 336\Delta_2 - 512\Delta_3 \\ &\quad - \Delta_1\Delta_2^2[\Delta\Delta_1(6-\Delta_2^2) - 32\Delta_3^2(6-\Delta_3^2)] = 0. \end{aligned}\right.$$

Nous allons réduire la parenthèse du dernier terme de l'égalité (*c*) en l'exprimant d'abord au moyen de Δ_3; nous avons trouvé dans le numéro précédent l'expression du produit $\Delta\Delta_1$ au moyen de Δ_2; on en déduit, en tenant compte de l'égalité $\Delta_2 = \frac{2\Delta_3}{1-\Delta_3^2}$,

$$\begin{aligned} \Delta\Delta_1(6-\Delta_2^2) &= \frac{8\Delta_2^2(6-\Delta_2^2)}{(1-\Delta_2^2)^2 - 4\Delta_2^2} \\ &= \frac{\dfrac{32\Delta_3^2}{(1-\Delta_3^2)^2}\left[6 - \dfrac{4\Delta_3^2}{(1-\Delta_3^2)^2}\right]}{\left[1 - \dfrac{4\Delta_3^2}{(1-\Delta_3^2)^2}\right]^2 - \dfrac{16\Delta_3^2}{(1-\Delta_3^2)^2}} \\ &= \frac{32\Delta_3^2[6(1-\Delta_3^2)^2 - 4\Delta_3^2]}{[(1-\Delta_3^2)^2 - 4\Delta_3^2]^2 - 16\Delta_3^2(1-\Delta_3^2)^2}. \end{aligned}$$

La parenthèse du dernier terme de l'égalité (c) peut alors, par substitution de la valeur précédente, s'écrire sous la forme

$$\frac{32\Delta_3^2}{[(1-\Delta_3^2)^2-4\Delta_3^2]^2-16\Delta_3^2(1-\Delta_3^2)^2} \times \{6(1-\Delta_3^2)^2-4\Delta_3^2-(6-\Delta_3^2)\{[(1-\Delta_3^2)^2-4\Delta_3^2]^2-16\Delta_3^2(1-\Delta_3^2)^2\}\}$$

Effectuons les calculs indiqués dans la parenthèse; l'expression précédente devient

$$\frac{32\Delta_3^2}{[(1-\Delta_3^2)^2-4\Delta_3^2]^2-16\Delta_3^2(1-\Delta_3^2)^2} \times \Delta_3^2(153-442\Delta_3^2+238\Delta_3^4-34\Delta_3^6+\Delta_3^8).$$

Transformons actuellement le facteur fractionnaire en remplaçant dans son dénominateur $(1-\Delta_3^2)$ par la quantité égale $\frac{2\Delta_3}{\Delta_2}$, on trouve

$$\begin{aligned}\frac{32\Delta_3^2}{[(1-\Delta_3^2)^2-4\Delta_3^2]^2-16\Delta_3^2(1-\Delta_3^2)^2} &= \frac{32\Delta_3^2}{\left(\frac{4\Delta_3^2}{\Delta_2^2}-4\Delta_3^2\right)^2-16\Delta_3^2\times\frac{4\Delta_3^2}{\Delta_2^2}} \\ &= \frac{2\Delta_2^4}{\Delta_3^2[(1-\Delta_2^2)^2-4\Delta_2^2]}.\end{aligned}$$

Remplaçons de nouveau $(1-\Delta_2^2)$ par la fraction égale $\frac{2\Delta_2}{\Delta_1}$; nous trouvons, pour valeur du facteur fractionnaire,

$$\frac{2\Delta_2^4}{\Delta_3^2[(1-\Delta_2^2)^2-4\Delta_2^2]} = \frac{2\Delta_2^4}{\Delta_3^2\left(\frac{4\Delta_2^2}{\Delta_1^2}-4\Delta_2^2\right)} = \frac{\Delta_2^2\times\Delta_1^2}{2\Delta_3^2(1-\Delta_1^2)}.$$

Substituons enfin à $1-\Delta_1^2$ la valeur égale $\frac{2\Delta_1}{\Delta}$

$$\frac{\Delta_2^2\times\Delta_1^2}{2\Delta_3^2(1-\Delta_1^2)} = \frac{\Delta\Delta_1\Delta_2^2}{4\Delta_3^2}.$$

Il en résulte que la parenthèse du dernier terme de l'égalité (c) peut s'écrire

$$\frac{\Delta\Delta_1\Delta_2^2}{4}(153 - 442\Delta_3^2 + 238\Delta_3^4 - 34\Delta_3^6 + \Delta_3^8),$$

et qu'enfin l'égalité (c) peut prendre la forme

$$(d)\quad \left\{ \begin{aligned} &\Delta - 42\Delta_1 + 336\Delta_2 - 512\Delta_3 \\ &\quad - \frac{\Delta\Delta_1^2\Delta_2^4}{4}(153 - 442\Delta_3^2 + 238\Delta_3^4 - 34\Delta_3^6 + \Delta_3^8) = 0; \end{aligned} \right.$$

d'où l'on déduit

$$\begin{aligned} \Delta_3 = {} & \frac{1}{512}[(8\Delta_2 - \Delta_1) \times 42 + \Delta] \\ & - \frac{\Delta\Delta_1^2\Delta_2^4}{2048}(153 - 442\Delta_3^2 + 238\Delta_3^4 - 34\Delta_3^6 + \Delta_3^8). \end{aligned}$$

Comme la parenthèse du second terme du second membre est évidemment positive pour les valeurs de Δ_3 qui ne surpassent pas $\sqrt{\frac{153}{442}}$ et à plus forte raison $\frac{1}{2} < \sqrt{\frac{153}{442}}$; que de plus, pour les mêmes valeurs de Δ_3, cette parenthèse a une valeur inférieure à 153, il en résulte que le premier terme du second membre peut être pris comme valeur approchée par excès de Δ_3, et que l'erreur commise est moindre que la fraction $\frac{153\Delta\Delta_1^2\Delta_2^4}{2048}$, qui admet elle-même, comme limite supérieure, $\frac{153\Delta^7}{2048 \times 2^{10}} = \frac{153\Delta^7}{2^{21}}$.

IX. Les formules établies aux n^os VI, VII et VIII, prises sous les formes

$$\begin{aligned} \Delta_1 &= \tfrac{1}{2}\Delta - \tfrac{1}{2}\Delta\Delta_1^2, \\ \Delta_2 &= \frac{1}{16}(10\Delta_1 - \Delta) + \frac{(6 - \Delta_2^2)\Delta.\Delta_1^2.\Delta_2^2}{16}, \\ \Delta_3 &= \frac{1}{512}[(8\Delta_2 - \Delta_1) \times 42 + \Delta] \\ &\quad - \frac{\Delta\Delta_1^2\Delta_2^4}{2048}(153 - 442\Delta_3^2 + 238\Delta_3^4 - 34\Delta_3^6 + \Delta_3^8), \end{aligned}$$

permettent de calculer une des valeurs Δ au moyen de la précédente, ou des deux ou trois précédentes, avec une approximation indéfinie, à l'aide de la méthode des approximations successives, et cela indépendamment des valeurs de D et d.

On pourrait encore les appliquer en suivant des méthodes diverses pour obtenir des valeurs approchées de l'inconnue, mais le procédé qui nous paraît le plus avantageux, et celui que nous allons appliquer, consiste à prendre pour valeur approchée du premier membre le premier terme du second ; nous éviterons ainsi l'emploi des logarithmes, et nous obtiendrons une approximation suffisante, pourvu que Δ soit assez petit.

Nous avons déjà trouvé directement, aux n^{os} III et V, les valeurs

$$\Delta_1 = 0,131652497587\ldots,$$
$$\Delta_2 = 0,065543462815\ldots,$$
$$\Delta_3 = 0,032736610412\ldots,$$

chacune de ces valeurs étant approchée à moins de $\frac{1}{10^{12}}$ par défaut.

En acceptant comme valeur approchée par excès de Δ_4, $\frac{1}{512}[(8\Delta_3 - \Delta_2) \times 42 + \Delta_1]$, l'erreur sera, d'après la fin du n° VIII, inférieure à $\frac{153 \times 132^7}{2^{21} \times 10^{21}}$.

Mais on a évidemment les égalités ou inégalités suivantes :

$$\begin{aligned}\frac{153 \times 132^7}{2^{21} \times 10^{21}} &= \frac{153 \times 33^7}{2^7 \times 10^{21}}\\ &= 153 \times \left(\frac{33}{100}\right)^7 \times \frac{1}{2^7 \times 10^7} < 153 \times \frac{1}{3^7} \times \frac{1}{2^7 \times 10^7}\\ &= \frac{17}{3^5 \times 2^7 \times 10^7} < \frac{55}{10^{12}}.\end{aligned}$$

Si nous faisons le calcul de

$$\frac{1}{512}[(8\Delta_3 - \Delta_2) \times 42 + \Delta_1].$$

en prenant les valeurs approchées que nous avons trouvées pour Δ_1, Δ_2, Δ_3, la première et la dernière par excès, la seconde par défaut, nous aurons pour valeur approchée par excès $0,01636392218 5$, quotient décimal approché à moins de $\frac{1}{10^{12}}$ par excès; du reste, l'erreur commise en remplaçant Δ_1, Δ_2, Δ_3 par leurs valeurs approchées n'atteint pas

$$\frac{1}{10^{12}} \times \frac{1}{512} \times (336 + 42 + 1) < \frac{1}{10^{12}},$$

d'où résulte qu'on peut prendre le quotient

$$0,016363922185$$

comme valeur approchée par excès de Δ_4 à moins de $\frac{57}{10^{12}}$, et le nombre $0,016363922128$ comme valeur approchée par défaut du même nombre, avec la même limite d'erreur.

Si actuellement nous répétons des calculs analogues pour trouver une valeur approchée de Δ_5 au moyen de Δ_2, Δ_3, Δ_4, en observant que la limite supérieure du terme négligé, à savoir $\frac{153\Delta_2\Delta_3^2\Delta_4^4}{2048}$, est moindre que le $\frac{1}{128}$ de celle du terme négligé dans le calcul précédent, puisque Δ_2, Δ_3, Δ_4 sont respectivement moindres que les moitiés de Δ_1, Δ_2, Δ_3, et qu'en conséquence ce terme négligé est moindre que $\frac{55}{10^{14}}$; que, de plus, l'erreur commise en substituant à Δ_2, Δ_3, Δ_4 leurs valeurs approchées, et en développant le quotient en fraction décimale à moins de $\frac{1}{10^{12}}$, par excès, est moindre que

$$\frac{1}{10^{12}} \times \frac{1}{512} \times (336 \times 57 + 42 + 1) + \frac{1}{10^{12}} < \frac{39}{10^{12}},$$

nous en conclurons qu'on peut prendre comme valeur approchée de Δ_5 par excès et à moins de $\frac{40}{10^{12}}$ le nombre $0,0081813437$, et comme valeur approchée de Δ_5 par défaut et avec la même approximation le nombre

$$0,0081814133397;$$

d'où

$$0,016363922128 < \Delta_4 < 0,016363922185,$$
$$0,0081814133397 < \Delta_5 < 0,0081814133437.$$

X. Les valeurs de Δ auxquelles nous sommes actuellement parvenus sont suffisamment petites pour que nous puissions faire usage de la formule établie au n° VII pour obtenir les valeurs consécutives de Δ; nous accepterons donc, comme valeur approchée par défaut de Δ_6, $\frac{1}{16}(10\,\Delta_5 - \Delta_4)$, et nous avons vu à la fin du n° VII que la limite de l'erreur commise était $\frac{\Delta_4^5}{170} < \frac{8}{10^{12}}$. Calculons $\frac{1}{16}(10\Delta_5 - \Delta_4)$, par défaut, en prenant Δ_5 par défaut et Δ_4 par excès; la limite de l'erreur commise sera $\frac{1}{16}(10 \times 40 + 57) \times \frac{1}{10^{12}}$, moindre que $\frac{29}{10^{12}}$. Il résulte de là que la valeur approchée par défaut et à moins de $\frac{1}{10^{12}}$ de

$$\tfrac{1}{16}(0,081814133970 - 0,016363922185)$$

est aussi approchée par défaut de Δ_6 à moins de $\frac{38}{10^{12}}$; d'où

$$0,004090638236 < \Delta_6 < 0,004090638274.$$

On trouve de même, par l'emploi réitéré de la même formule et en prenant les mêmes précautions relatives à l'approximation,

$$0,002045310557 < \Delta_7 < 0,002045310586,$$
$$0,001022654206 < \Delta_8 < 0,001022654229.$$

Nous avons ainsi trouvé des valeurs approchées de Δ_6 à moins de $\frac{38}{10^{12}}$, de Δ_7 à moins de $\frac{29}{10^{12}}$ et de Δ_8 à moins de $\frac{23}{10^{12}}$.

Δ_8 est assez petit, comme nous allons le voir, pour que nous puissions tirer tout le parti possible des nombres que nous avons calculés, en tenant compte des erreurs qui les affectent.

XI. Nous avons posé au n° V $\Delta_k = D_k - d_k$, nous avons du reste vu, d'après le n° I, que

$$d_{k+1} = D_k + d_k;$$

on en déduit

$$d_{k+1} = 2\,d_k + \Delta_k.$$

Faisant varier k, on trouve

$$\begin{aligned}
d_4 &= 2\,d_3 + \Delta_3,\\
d_5 &= 2\,d_4 + \Delta_4,\\
&\dots\dots\dots\dots\\
d_p &= 2\,d_{p-1} + \Delta_{p-1}.
\end{aligned}$$

Multipliant ces égalités par les puissances décroissantes de 2 à partir de 2^{p-4}, on trouve, en les ajoutant membre à membre,

$$(1) \qquad d_p = 2^{p-3}\,d_3 + 2^{p-4}\,\Delta_3 + 2^{p-5}\,\Delta_4 + \ldots + 2\,\Delta_{p-2} + \Delta_{p-1}.$$

Si nous posons $p = 8$, nous aurons

$$(2) \qquad d_8 = 2^5\,d_3 + 2^4\,\Delta_3 + 2^3\,\Delta_4 + 2^2\,\Delta_5 + 2\,\Delta_6 + \Delta_7.$$

Remplaçons dans cette égalité (2) d_3, Δ_3, Δ_4, …, Δ_7, par les valeurs approchées par défaut que nous en avons trouvées, nous aurons une valeur approchée par défaut de d_8, et la limite supérieure de l'erreur commise sur d_8 s'obtiendra en remplaçant au second membre de l'égalité (2) d_3, Δ_3, ... Δ_7, chacun par la limite supérieure de l'erreur commise sur ce nombre.

Nous rapportons ci-dessous le Tableau de ce calcul, en observant que nous n'écrivons que le numérateur de la limite supérieure de chaque erreur, tous ces nombres ayant pour dénominateur commun 10^{12}.

Sur une ligne horizontale nous écrivons séparément la valeur approchée par défaut de d_3 et le numérateur de la limite de l'erreur commise sur ce nombre. Nous multiplions ces deux nombres par 2; les produits sont placés sur la seconde ligne horizontale. A ces produits nous ajoutons respectivement la valeur approchée par défaut de Δ_3, et le numérateur de la limite de l'erreur commise sur Δ_3; nous multiplions les sommes par 2, et nous continuons de même jusqu'à ce que nous ayons fait l'addition de la valeur approchée par défaut de Δ_7 et du numérateur de la limite de l'erreur correspondante.

$$\begin{array}{rr}
15,257051688261 & 7 \\
30,514103376522 & 14 \\
32736610412 & 1 \\
\hline
30,546839986934 & 15 \\
61,093679973868 & 30 \\
16363922128 & 57 \\
\hline
61,110043895996 & 87 \\
122,220087791992 & 174 \\
8181413397 & 40 \\
\hline
122,228269205389 & 214 \\
244,456538410778 & 428 \\
4090638236 & 38 \\
\hline
244,460629049014 & 466 \\
488,921258098028 & 932 \\
2045310557 & 29 \\
\hline
488,923303408585 & 961
\end{array}$$

De là nous pouvons conclure que d_8 admet pour valeur approchée par défaut 488,923303408585, et que la limite supérieure de l'erreur est $\frac{961}{10^{12}}$.

XII. Par le procédé employé au n° XI, on trouve l'égalité

$$(1)\qquad \left\{\begin{aligned} d_{p+n} = 2^n d_p + 2^{n-1}\Delta_p + 2^{n-2}\Delta_{p+1} + \ldots \\ + 2\Delta_{p+n-2} + \Delta_{p+n-1}. \end{aligned}\right.$$

Si k est le nombre des côtés du polygone admettant pour diamètre du cercle inscrit d_p, d_{p+n} est le diamètre du cercle inscrit dans le polygone dont le nombre des côtés est $k \times 2^n$, et dont le périmètre est aussi représenté par $k \times 2^n$; nous pouvons en conclure, d'après l'égalité des rapports des périmètres de deux polygones réguliers d'un même nombre de côtés à celui des diamètres des cercles qui leur sont respectivement inscrits, que $\frac{d_{p+n}}{2^n}$ est le diamètre du cercle inscrit, dans le polygone régulier dont le nombre des côtés est $k \times 2^n$ et dont le périmètre est k.

Divisant les deux membres de la dernière égalité (1) par 2^n, on en déduit

$$\frac{d_{p+n}}{2^n} = d_p + \frac{1}{2}\Delta_p + \frac{1}{2^2}\Delta_{p+1} + \ldots + \frac{1}{2^n}\Delta_{p+n-1}.$$

On trouve de même, pour le diamètre du cercle circonscrit au même polygone,

$$\frac{\mathrm{D}_{p+n}}{2^n} = d_p + \frac{1}{2}\Delta_p + \frac{1}{2^2}\Delta_{p+1} + \ldots + \frac{1}{2^n}\Delta_{p+n-1} + \frac{1}{2^n}\Delta_{p+n}.$$

On voit facilement, d'après ces égalités, que $\frac{d_{p+n}}{2^n}$ croît avec n, tandis que $\frac{\mathrm{D}_{p+n}}{2^n}$ décroît dans les mêmes circonstances, tout en restant supérieur à $\frac{d_{p+n}}{2^n}$ d'une quantité qui a pour limite zéro quand n augmente indéfiniment; du reste, $\frac{d_{p+n}}{2^n}$ et $\frac{D_{p+n}}{2^n}$ comprennent entre eux le diamètre de la circonférence dont la longueur est k; on peut donc considérer le diamètre de cette circonférence dont la longueur est k comme la limite vers laquelle tend $\frac{d_{p+n}}{2^n}$ quand n augmente au delà de toute limite, et nous poserons

$$\Theta = \lim \frac{d_{p+n}}{2^n}.$$

XIII. D'après les inégalités établies au n° VI, nous pouvons écrire

$$\begin{array}{llll}
 & \Delta_p & = \Delta_p & = \Delta_p, \\
\frac{1}{2}\Delta_p & -\frac{1}{8}\Delta_p^3 & < \Delta_{p+1} & < \frac{1}{2}\Delta_p. \\
\frac{1}{2}\Delta_{p+1} & -\frac{1}{8}\Delta_{p+1}^3 & < \Delta_{p+2} & < \frac{1}{2}\Delta_{p+1}, \\
\frac{1}{2}\Delta_{p+2} & -\frac{1}{8}\Delta_{p+2}^3 & < \Delta_{p+3} & < \frac{1}{2}\Delta_{p+2}, \\
\multicolumn{4}{l}{\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots,} \\
\frac{1}{2}\Delta_{p+n-2} & -\frac{1}{8}\Delta_{p+n-2}^3 & < \Delta_{p+n-1} & < \frac{1}{2}\Delta_{p+n-2}, \\
\frac{1}{2}\Delta_{p+n-1} & -\frac{1}{2}\Delta_{p+n-1} = & 0 & = \frac{1}{2}\Delta_{p+n-1} - \frac{1}{2}\Delta_{p+n-1}.
\end{array}$$

Multiplions les nombres écrits dans chaque ligne horizontale par une puissance de $\frac{1}{2}$ dont l'exposant soit égal au rang de la ligne correspondante, et ajoutons membre à membre ces deux séries d'inégalités; on trouve, en posant, pour abréger,

$$\Sigma = \frac{1}{2}\Delta_p + \frac{1}{2^2}\Delta_{p+1} + \dots + \frac{1}{2^n}\Delta_{p+n-1},$$

$$\Sigma' = \frac{1}{8}\left(\frac{1}{2^2}\Delta_p^3 + \frac{1}{2^3}\Delta_{p+1}^3 + \dots + \frac{1}{2^n}\Delta_{p+n-2}^3\right),$$

$$\frac{1}{2}\Delta_p + \frac{1}{4}\Sigma - \frac{1}{2^{n+2}}\Delta_{p+n-1} - \Sigma' < \Sigma < \frac{1}{2}\Delta + \frac{1}{4}\Sigma - \frac{1}{2^{n+2}}\Delta_{p+n-1}.$$

Observons actuellement que les termes contenus dans la parenthèse qui donne la valeur de Σ' sont, toujours d'après les inégalités du n° VI, respectivement moindres que les termes de même rang de la progression géométrique décroissante

$$\frac{1}{2^2}\Delta_p^3,\quad \frac{1}{2^3}\times\frac{1}{2^3}\Delta_p^3,\quad \frac{1}{2^4}\times\frac{1}{(2^2)^3}\Delta_p^3,$$
$$\frac{1}{2^5}\times\frac{1}{(2^3)^3}\Delta_p^3,\quad \ldots,\quad \frac{1}{2^n}\times\frac{1}{(2^{n-2})^3}\Delta_p^3;$$

il en résultera que Σ' est inférieur au $\frac{1}{8}$ de la limite de la somme des termes de cette progression, soit à

$$\frac{1}{8}\Delta_p^3\times\frac{\frac{1}{2^2}}{1-\frac{1}{2^4}}=\frac{\Delta_p^3}{2(2^4-1)}.$$

Dès lors, remplaçant Σ' par une quantité plus grande, nous aurons

$$\frac{1}{2}\Delta_p+\frac{1}{4}\Sigma-\frac{1}{2^{n+2}}\Delta_{p+n-1}-\frac{\Delta_p^3}{2(2^4-1)}$$
$$<\Sigma<\frac{1}{2}\Delta_p+\frac{1}{4}\Sigma-\frac{1}{2^{n+2}}\Delta_{p+n-1},$$

ou

$$\frac{1}{2}\Delta_p-\frac{1}{2^{n+2}}\Delta_{p+n-1}-\frac{\Delta_p^3}{2(2^4-1)}<\frac{3}{4}\Sigma<\frac{1}{2}\Delta_p-\frac{1}{2^{n+2}}\Delta_{p+n-1}.$$

Ces inégalités, ayant lieu quel que soit n, auront lieu si on le fait croître au de là detoute limite, et comme, dans ces conditions,

$$\lim\Sigma=\Theta-d_p,\qquad \lim\Delta_{p+n-1}=0,$$

nous trouverons par cette hypothèse

$$\frac{1}{2}\Delta_p-\frac{\Delta_p^3}{2\times(2^4-1)}<\frac{3}{4}(\Theta-d_p)<\frac{1}{2}\Delta_p,$$

d'où

$$d_p+\frac{2}{3}\Delta_p-\frac{2}{3}\frac{\Delta_p^3}{15}<\Theta<d_p+\frac{2}{3}\Delta_p.$$

On peut donc accepter comme valeur approchée par excès du diamètre de la circonférence dont la longueur est k

$$d_p + \tfrac{2}{3}\Delta_p,$$

et comme limite de l'erreur commise $\dfrac{2}{3}\,\dfrac{\Delta_p^3}{15}$.

XIV. Nous reportant au n° XI, le nombre des côtés du polygone régulier, dont le cercle inscrit admet pour diamètre d_8, est

$$6 \times 2^8 = 1536,$$

nous pouvons donc déduire du numéro précédent que

$$d_8 + \tfrac{2}{3}\Delta_8$$

représente par excès le diamètre de la circonférence dont la longueur est représentée par 1536.

D'après les valeurs numériques trouvées pour d_8 et Δ_8, et tenant compte des erreurs dont elles sont affectées, on trouve dans ce cas particulier

$$\Theta < 488,923985179032;$$

la limite de l'erreur commise sur Θ est

$$\frac{1037}{10^{12}}.$$

Le quotient à moins de $\frac{1}{10^{12}}$ par défaut de 1536 par

$$488,923985179032$$

est

$$3,141592653585.$$

On peut prendre pour limite de l'erreur commise sur le quotient $\dfrac{3,2 \times 1037}{488,9 \times 10^{12}} < \dfrac{7}{10^{12}}$, d'où

$$3,141592653585 < \pi < 3,141592653593.$$

$3,141592653593$ est donc approché de π à moins de $\frac{5}{10^{12}}$ sans qu'on connaisse le sens de l'approximation.

On trouverait $\frac{1}{\pi}$ en divisant Θ par 1536.

NOTE SUR LA DÉCOMPOSITION D'UNE FORME QUADRATIQUE À m VARIABLES EN UNE SOMME DE $m-n$ CARRÉS;

PAR M. BENOIT.

On sait que, pour exprimer qu'une forme quadratique à m variables peut se décomposer en une somme de $m-n$ carrés, il faut écrire que tous les déterminants mineurs de l'ordre $n-1$ de son discriminant sont nuls; mais les équations ainsi obtenues ne sont pas distinctes : je vais chercher quelles sont celles d'entre elles dont les autres sont des conséquences.

Soit d'abord le déterminant

$$\Delta = \begin{vmatrix} a_1^1 & a_1^2 & \dots & a_1^m \\ a_2^1 & \cdot\cdot & \dots & a_2^m \\ \cdot\cdot & \cdot\cdot & \dots & \cdot\cdot \\ a_m^1 & \cdot\cdot & \dots & a_m^m \end{vmatrix}.$$

Considérons l'un des mineurs de l'ordre n différent de zéro, soit

$$\delta = \begin{vmatrix} a_1^1 & a_1^2 & \dots & a_1^{m-n} \\ a_2^1 & \cdot\cdot & \dots & a_2^{m-n} \\ \cdot\cdot & \cdot\cdot & \dots & \cdot\cdot\cdot\cdot\cdot \\ a_{m-n}^1 & \cdot\cdot & \dots & a_{m-n}^{m-n} \end{vmatrix} \neq 0.$$

./01. . .+ . % 7 &2 2!
) $

NOUVELLES ANNALES
DE
MATHÉMATIQUES.

[O¹5n]

SURFACES ENGENDRÉES PAR LE DÉPLACEMENT D'UNE COURBE PLANE AVEC CONE CIRCONSCRIT LE LONG DE LA COURBE;

Par M. E. KERAVAL,
Professeur au lycée Louis-le-Grand.

Pour abréger, j'appelle *surface* Σ toute surface engendrée par le déplacement d'une courbe plane (C), quand il existe un cône circonscrit à la surface tout le long de chaque courbe (C). Je me servirai de coordonnées homogènes x_1, x_2, x_3, x_4 que je supposerai fonctions de deux paramètres u, v. Le paramètre u restera constant sur chaque courbe (C); je supposerai que les courbes obtenues en faisant v constant soient conjuguées des premières.

Je sais que dans ces conditions x_1, x_2, x_3, x_4 sont des solutions d'une même équation de la forme

$$\frac{\partial^2 \theta}{\partial u\, \partial v} + p\frac{\partial \theta}{\partial u} + q\frac{\partial \theta}{\partial v} + r\theta = 0, \tag{1}$$

où p, q, r sont des fonctions de u et de v. Si je désigne par (S) la courbe lieu du sommet des cônes circonscrits à Σ le long des courbes (C), les tangentes aux courbes v

rencontrent la courbe S, donc si l'on désigne par h et k les invariants de (1) savoir

$$h = \frac{\partial p}{\partial u} + pq - r,$$

$$k = \frac{\partial q}{\partial v} + pq - r,$$

l'invariant k est ici nul; nous dirons que l'équation (1) est de rang zéro à gauche. D'autre part, les courbes u étant planes, il existe une relation de la forme

$$(2) \qquad U_1 x_1 + U_2 x_2 + U_3 x_3 + U_4 x_4 \equiv 0,$$

où les coefficients de x_1, x_2, x_3, x_4 ne dépendent que de u. On sait que, dans ce cas, l'équation (1) est au plus de rang deux à droite, c'est-à-dire qu'après avoir appliqué deux fois au plus la transformation de Laplace on aura une nouvelle équation où le nouvel invariant h sera nul. Alors les équations de la surface Σ doivent être de la forme

$$(3) \qquad x_k = \begin{vmatrix} B_k & A_k & A'_k & A''_k \\ \beta_1 & \alpha_1 & \alpha'_1 & \alpha''_1 \\ \beta_2 & \alpha_2 & \alpha'_2 & \alpha''_2 \\ \beta_3 & \alpha_3 & \alpha'_3 & \alpha''_3 \end{vmatrix} \qquad (k = 1, 2, 3, 4).$$

où les fonctions B_k, β_1, β_2, β_3 dépendent de v et A_k, α_1, α_2, α_3 de u. Mais il peut arriver que le plan des courbes u roule sur un cône; nous prendrons le sommet du cône à l'origine

$$x_1 = 0, \qquad x_2 = 0, \qquad x_3 = 0.$$

Alors la relation (2) ne contiendra plus que trois coordonnées et l'équation (1) sera de rang un à droite, donc

$$(4)\qquad x_k = \begin{vmatrix} B_k & A_k & A'_k \\ \beta_1 & \alpha_1 & \alpha'_1 \\ \beta_2 & \alpha_2 & \alpha'_2 \end{vmatrix}.$$

Enfin le plan des courbe u peut passer par une droite fixe. Dans ce cas, ci cette droite a pour équation

$$x_1 = 0, \qquad x_2 = 0$$

la relation (2) ne contient plus que les deux variables x_1, x_2, l'équation (1) est de rang zéro à droite et les équations de la surface prennent la forme

$$(5)\qquad x_k = B_k + A_k.$$

Nous sommes ainsi conduits à ranger les surfaces Σ en trois catégories, nous aurons ainsi les surfaces Σ de troisième, seconde et première espèce. Dans les équations précédentes on suppose toujours qu'il n'existe pas de relation linéaire homogène à coefficients constants soit entre les quantités α, soit entre les β.

Étude des surfaces Σ de troisième espèce.

Les équations de la surface sont de la forme (3), mais il reste à exprimer que les courbes u sont planes. En portant les valeurs de x_1, x_2, x_3, x_4 dans la relation (2), on trouve une relation linéaire en B_1, B_2, B_3, B_4, β_1, β_2, β_3 et homogène. En divisant par un déterminant différent de zéro, cette relation prend la forme

$$U_1 B_1 + U_2 B_2 + U_3 B_3 + U_4 B_4 + \varphi(\beta_1, \beta_2, \beta_3) \equiv 0,$$

φ étant linéaire et homogène en β_1, β_2, β_3. Dans cette relation donnons à u quatre valeurs telles que le déterminant des coefficients de U_1, U_2, U_3, U_4 soit différent de zéro, ce qui est possible, puisque le plan des courbes u ne roule pas sur un cône, nous pourrons tirer pour B_1, B_2, B_3, B_4 des valeurs de la forme

$$B_k = a_k \beta_1 + b_k \beta_2 + c_k \beta_3,$$

les coefficients a_k, b_k, c_k étant des constantes. Alors en changeant la signification de A_k nous pourrons prendre, pour équations de la surface Σ,

$$(A) \qquad x_k = \begin{vmatrix} 0 & A_k & A'_k & A''_k \\ \beta_1 & \alpha_1 & \alpha'_1 & \alpha''_1 \\ \beta_2 & \alpha_2 & \alpha'_2 & \alpha''_2 \\ \beta_3 & \alpha_3 & \alpha'_3 & \alpha''_3 \end{vmatrix}$$

et cette fois les courbes u sont bien planes. La forme de ces équations nous permet d'énoncer le théorème suivant :

THÉORÈME. — *Les courbes planes* $u = \text{const.}$ *qui engendrent la surface sont homographiques, et la correspondance homographique est établie par les courbes conjuguées* $v = \text{const.}$

On peut imaginer par exemple dans un plan P un point de coordonnées homogènes β_1, β_2, β_3 décrivant une courbe (β). Pour chaque valeur de u on a dans un plan P_u une transformation homographique de cette courbe (β). Nous verrons plus loin si cette homographie peut devenir *singulière* pour certaines valeurs de u.

Pour le moment nous allons chercher les plans P_u et la courbe lieu du sommet des cônes circonscrits. On pourrait se servir des méthodes indiquées par M. Darboux pour des questions analogues, je préfère indiquer une méthode ne dépassant pas le cadre de la classe de Mathématiques spéciales.

Équations de la courbe (S).

Les tangentes aux courbes *v* rencontrent cette courbe. L'une des courbes *v* a pour équations

$$y_k = \begin{vmatrix} A_k & A'_k & A''_k \\ \alpha_1 & \alpha'_1 & \alpha''_1 \\ \alpha_2 & \alpha'_2 & \alpha''_2 \end{vmatrix}.$$

Or la théorie des déterminants nous donne l'identité

$$(6)\qquad \begin{vmatrix} A_k & A'_k & A''_k & A'''_k \\ \alpha_1 & \alpha'_1 & \alpha''_1 & \alpha'''_1 \\ \alpha_2 & \alpha'_2 & \alpha''_2 & \alpha'''_2 \\ \alpha_3 & \alpha'_3 & \alpha''_3 & \alpha'''_3 \end{vmatrix} \times (\alpha_1\alpha'_2 - \alpha_2\alpha'_1)$$
$$= \begin{vmatrix} A_k & A'_k & A''_k \\ \alpha_1 & \alpha'_1 & \alpha''_1 \\ \alpha_2 & \alpha'_2 & \alpha''_2 \end{vmatrix} \times \begin{vmatrix} \alpha_1 & \alpha'_1 & \alpha'''_1 \\ \alpha_2 & \alpha'_2 & \alpha'''_2 \\ \alpha_3 & \alpha'_3 & \alpha'''_3 \end{vmatrix}$$
$$- \begin{vmatrix} A_k & A'_k & A'''_k \\ \alpha_1 & \alpha'_1 & \alpha'''_1 \\ \alpha_2 & \alpha'_2 & \alpha'''_2 \end{vmatrix} \times \begin{vmatrix} \alpha_1 & \alpha'_1 & \alpha''_1 \\ \alpha_2 & \alpha'_2 & \alpha''_2 \\ \alpha_3 & \alpha'_3 & \alpha''_3 \end{vmatrix}.$$

Si donc on pose

$$(B)\qquad X_k = \begin{vmatrix} A_k & A'_k & A''_k & A'''_k \\ \alpha_1 & \alpha'_1 & \alpha''_1 & \alpha'''_1 \\ \alpha_2 & \alpha'_2 & \alpha''_2 & \alpha'''_2 \\ \alpha_3 & \alpha'_3 & \alpha''_3 & \alpha'''_3 \end{vmatrix},$$

cette identité prend la forme

$$y_k \times \begin{vmatrix} \alpha_1 & \alpha'_1 & \alpha'''_1 \\ \alpha_2 & \alpha'_2 & \alpha'''_2 \\ \alpha_3 & \alpha'_3 & \alpha'''_3 \end{vmatrix} - y'_k \times \begin{vmatrix} \alpha_1 & \alpha'_1 & \alpha''_1 \\ \alpha_2 & \alpha'_2 & \alpha''_2 \\ \alpha_3 & \alpha'_3 & \alpha''_3 \end{vmatrix} = X_k(\alpha_1\alpha'_2 - \alpha_2\alpha'_1),$$

qui prouve que le point X_1, X_2, X_3, X_4 se trouve bien sur la tangente à la courbe décrite par le point y_1, y_2, y_3, y_4. Il en est de même pour deux autres courbes *v* dont les coordonnées du sommet du cône sont données par les formules (B).

Plan des courbes u.

Le plan P_u d'une courbe u c'est le plan osculateur de la courbe (λ),

$$\lambda_k = A_k.$$

c'est-à-dire de la courbe décrite par le point de coordonnées A_1, A_2, A_3, A_4. Ceci résulte immédiatement de la forme des équations de la surface Σ, le plan osculateur contenant les trois points de coordonnées A_k, A'_k, A''_k.

Enveloppes des courbes u.

Un point d'une courbe u décrivant une enveloppe ne peut être que sur la caractéristique du plan P_u. Pour ces points, u et v sont liés par la relation

$$\begin{vmatrix} \beta_1 & \alpha_1 & \alpha'_1 \\ \beta_2 & \alpha_2 & \alpha'_2 \\ \beta_3 & \alpha_3 & \alpha'_3 \end{vmatrix} = 0,$$

que j'écrirai pour abréger

$$| \beta \alpha \alpha' | = 0.$$

Pour chaque valeur de u cette relation détermine un certain nombre de points situés sur la caractéristique du plan P_u. Soit M l'un de ces points, ses coordonnées sont fonctions de u et v qui sont eux-mêmes liés par la relation précédente. On peut supposer u et v fonctions d'un paramètre t, on aura

$$\frac{dx_k}{dt} = \frac{\partial x_k}{\partial u}\frac{du}{dt} + \frac{\partial x_k}{\partial v}\frac{dv}{dt}.$$

Or pour ce point M, on a

$$x_k = - A_k | \beta \alpha' \alpha'' | + A'_k | \beta \alpha \alpha'' |,$$

$$\frac{\partial x_k}{\partial u} = - A_k | \beta \alpha' \alpha''' | + A'_k | \beta \alpha \alpha''' | + A''_k | \beta \alpha \alpha'' |,$$

$$\frac{\partial x_k}{\partial v} = - A_k | \beta' \alpha' \alpha'' | + A'_k | \beta' \alpha \alpha'' |,$$

donc $\frac{dx_k}{dt}$ est une fonction linéaire et homogène de A_k, A'_k, A''_k. Le point de coordonnées $\frac{dx_1}{dt}$ $\frac{dx_2}{dt}$ $\frac{dx_3}{dt}$ $\frac{dx_4}{dt}$ est donc dans le plan P_u, il en est de même de la tangente à la courbe décrite par le point M, et comme elle est dans le plan tangent à la surface elle coïncide avec la tangente à la courbe u. Donc tous les points M que nous avons définis décrivent des courbes tangentes aux courbes u. Si par exemple, β_1, β_2, β_3 sont des polynomes de degré m en v, c'est-à-dire si les courbes u sont unicursales et de degré m, on aura m enveloppes. Pour $m = 2$, on retrouve les résultats de M. Blutel. (*Annales de l'École Normale*, mai-juin 1890. *Voir* aussi la thèse de M. Lelieuvre 1894 qui s'occupe du cas d'une courbe plane unicursale se déplaçant de telle sorte qu'elle soit divisée homographiquement par ses conjuguées.)

Une transformation par polaires réciproques change une surface Σ en une surface Σ et nous montre que le cône circonscrit à Σ roule sur des développables décrites par les génératrices de contact du cône avec les plans tangents qu'on peut lui mener par la tangente à la courbe que décrit le sommet du cône.

Définition d'une congruence Σ.

Dans les équations (A) traitons β_1, β_2, β_3 comme des constantes et désignons-les par a, b, c ; au lieu de

α_1, α_2, α_3 mettons α, β, γ, nous aurons les équations

$$(C) \qquad x_k = \begin{vmatrix} 0 & A_k & A'_k & A''_k \\ a & \alpha & \alpha' & \alpha'' \\ b & \beta & \beta' & \beta'' \\ c & \gamma & \gamma' & \gamma'' \end{vmatrix} \qquad (k = 1, 2, 3, 4).$$

Dans ces formules, a, b, c sont donc trois constantes arbitraires et A_1, A_2, A_3, A_4, α, β, γ des fonctions de u. Pour chaque valeur de u, les points x_k qui correspondent aux différentes valeurs a, b, c décrivent un plan P_u et les tangentes aux courbes décrites par ces points concourent en un point S

$$(D) \qquad X_k = \begin{vmatrix} A_k & A'_k & A''_k & A'''_k \\ \alpha & \alpha' & \alpha'' & \alpha''' \\ \beta & \beta' & \beta'' & \beta''' \\ \gamma & \gamma' & \gamma'' & \gamma''' \end{vmatrix}.$$

Si a, b, c restent fixes et que u varie, les formules (C) définissent une courbe v, on a donc une congruence de courbes v et les surfaces de cette congruence sont les surfaces Σ. Il est naturel de chercher quelle peut être la surface focale de la congruence Σ. Il est évident que si pour une valeur particulière de u le point S se trouve dans le plan P_u toutes les courbes de la congruence seront tangentes à ce plan qui formera une des nappes de la surface focale. La valeur de X_k est de la forme

$$X_k = \lambda_1 A_k + \lambda_2 A'_k + \lambda_3 A''_k - \Delta A'''_k,$$

où

$$\Delta = \begin{vmatrix} \alpha & \alpha' & \alpha'' \\ \beta & \beta' & \beta'' \\ \gamma & \gamma' & \gamma'' \end{vmatrix}.$$

J'écris que ce point est dans le plan P_u, c'est-à-dire

$$\begin{vmatrix} x_1 & x_2 & x_3 & x_4 \\ A_1 & A_2 & A_3 & A_4 \\ A'_1 & A'_2 & A'_3 & A'_4 \\ A''_1 & A''_2 & A''_3 & A''_4 \end{vmatrix} = 0.$$

Les coefficients de λ_1, λ_2, λ_3 sont nuls et il reste

$$\Delta \times \begin{vmatrix} A_1 & A_2 & A_3 & A_4 \\ A'_1 & A'_2 & A'_3 & A'_4 \\ A''_1 & A''_2 & A''_3 & A''_4 \\ A'''_1 & A'''_2 & A'''_3 & A'''_4 \end{vmatrix} = 0,$$

ce qui donne : 1° les plans P_u stationnaires c'est-à-dire osculateurs stationnaires de la courbe A_k; 2° les plans P_u pour lesquels $\Delta = 0$ qui correspondent aux valeurs de u pour lesquels un point décrivant la courbe plane $x = \alpha$, $y = \beta$, $z = \gamma$ serait un point d'inflexion. Je vais faire voir que pour ces valeurs de u qui rendent Δ nul, toutes les courbes v sont concourantes et concourent au point S correspondant. Il suffit de se reporter à l'identité (6) où l'on tiendra compte du changement de notations. Ainsi la surface focale se compose des points qui correspondent à Δ nul et des plans P_u stationnaires. D'ailleurs il n'existe pas d'autres points focaux, car si F est un point focal et si S n'est pas dans le plan P_u correspondant, le plan tangent en F à la surface de la congruence dépendra évidemment du mode d'assemblage des courbes v.

Cas d'homographie singulière.

Supposons qu'on donne à u une valeur fixe, les formules (C) font correspondre point par point le plan P_u et un plan Q. Dans le premier on a le point x_1, x_2,

x_3, x_4; dans le deuxième, le point de coordonnées a, b, c homogènes. Il faut s'assurer que ces formules (C) permettent de tirer a, b, c en fonction de trois des variables x_k. Désignons par

$$\begin{vmatrix} \alpha_1 & \alpha_2 & \alpha_3 \\ \beta_1 & \beta_2 & \beta_3 \\ \gamma_1 & \gamma_2 & \gamma_3 \end{vmatrix} \text{ l'adjoint de } \begin{vmatrix} \alpha & \alpha' & \alpha'' \\ \beta & \beta' & \beta'' \\ \gamma & \gamma' & \gamma'' \end{vmatrix},$$

les équations (C) s'écriront

$$\begin{aligned} x_k = &- a(A_k\alpha_1 + A'_k\alpha_2 + A''_k\alpha_3) \\ &- b(A_k\beta_1 + A'_k\beta_2 + A''_k\beta_3) \\ &- c(A_k\gamma_1 + A'_k\gamma_2 + A''_k\gamma_3), \end{aligned}$$

ce qui fait quatre équations. Pour qu'on puisse tirer a, b, c des trois premières, par exemple, il faut que

$$\begin{vmatrix} A_1\alpha_1 + A'_1\alpha_2 + A''_1\alpha_3 & A_1\beta_1 + A'_1\beta_2 + A''_1\beta_3 & A_1\gamma_1 + A'_1\gamma_2 + A''_1\gamma_3 \\ A_2\alpha_1 + A'_2\alpha_2 + A''_2\alpha_3 & A_2\beta_1 + A'_2\beta_2 + A''_2\beta_3 & A_2\gamma_1 + A'_2\gamma_2 + A''_2\gamma_3 \\ A_3\alpha_1 + A'_3\alpha_2 + A''_3\alpha_3 & A_3\beta_1 + A'_3\beta_2 + A''_3\beta_3 & A_3\gamma_1 + A'_3\gamma_2 + A''_3\gamma_3 \end{vmatrix} \neq 0$$

ou

$$\begin{vmatrix} A_1 & A'_1 & A''_1 \\ A_2 & A'_2 & A''_2 \\ A_3 & A'_3 & A''_3 \end{vmatrix} \times \begin{vmatrix} \alpha_1 & \alpha_2 & \alpha_3 \\ \beta_1 & \beta_2 & \beta_3 \\ \gamma_1 & \gamma_2 & \gamma_3 \end{vmatrix} = 0.$$

Si le second de ces déterminants est nul, son adjoint Δ l'est aussi et l'homographie est singulière. Pour cette valeur de u nous avons vu que les courbes v concouraient au point S correspondant. Si $\Delta \neq 0$, l'homographie ne sera singulière que si les quatre déterminants à trois lignes du Tableau

$$T : \begin{vmatrix} A_1 & A_2 & A_3 & A_4 \\ A'_1 & A'_2 & A'_3 & A'_4 \\ A''_1 & A''_2 & A''_3 & A''_4 \end{vmatrix}$$

sont nuls à la fois. Or ceci ne peut arriver que pour certaines valeurs particulières de u, car nous supposons évidemment que le point A_1, A_2, A_3, A_4 décrive une courbe gauche.

Conclusion. — Ayant choisi les fonctions A_k telles que le point A_k décrive une courbe gauche, et les fonctions α, β, γ telles qu'il n'existe pas entre elles une relation linéaire et homogène, les équations d'une surface Σ seront données par les formules (C), où a, b, c sont des fonctions arbitraires d'un paramètre v. Cela revient à choisir arbitrairement la courbe décrite par le point a, b, c dans le plan ou encore l'une des courbes u dans le plan P_u correspondant.

DEUXIÈME SOLUTION.

On arrive aux mêmes conclusions en modifiant très légèrement les formules données par M. Blutel pour le cas des coniques. Ces formules sont les suivantes :

$$x = \frac{P_1 v^2 + Q_1 v + R_1}{P\ v^2 + Q\ v + R},$$
$$y = \frac{P_2 v^2 + Q_2 v + R_2}{P\ v^2 + Q\ v + R},$$
$$z = \frac{P_3 v^2 + Q_3 v + R_3}{P\ v^2 + Q\ v + R}.$$

Les fonctions P, Q, R, P_1, ..., R_3, de la variable u étant déterminées par les équations

$$\frac{P'_1}{P'} = \frac{Q'_1}{Q'} = \frac{R'_1}{R'} = X,$$
$$\frac{P'_2}{P'} = \frac{Q'_2}{Q'} = \frac{R'_2}{R'} = Y,$$
$$\frac{P'_3}{P'} = \frac{Q'_3}{Q'} = \frac{R'_3}{R'} = Z.$$

De telle sorte qu'on peut se donner X, Y, Z sommet du cône circonscrit et P, Q, R, on aura alors P_1, ..., R_3 par neuf quadratures. Si je considère les trois courbes P, Q, R

$$\left\{\begin{aligned} x &= \frac{P_1}{P} \\ y &= \frac{P_2}{P} \\ z &= \frac{P_3}{P} \end{aligned}\right. \qquad \left\{\begin{aligned} x &= \frac{Q_1}{Q} \\ y &= \frac{Q_2}{Q} \\ z &= \frac{Q_3}{Q} \end{aligned}\right. \qquad \left\{\begin{aligned} x &= \frac{R_1}{R} \\ y &= \frac{R_2}{R} \\ z &= \frac{R_3}{R} \end{aligned}\right.$$

les tangentes aux points de ces trois courbes, qui correspondent à une même valeur de u, concourent au point X, Y, Z. La congruence Σ peut alors s'écrire

$$(C') \qquad \left\{\begin{aligned} x &= \frac{aP_1 + bQ_1 + cR_1}{aP + bQ + cR}, \\ y &= \frac{aP_2 + bQ_2 + cR_2}{aP + bQ + cR}, \\ z &= \frac{aP_3 + bQ_3 + cR_3}{aP + bQ + cR}. \end{aligned}\right.$$

Si dans ces formules je donne à u une valeur particulière, la tangente au point x, y, z va passer par X, Y, Z quelles que soient les valeurs des constantes a, b, c; en effet en désignant par D l'expression $aP + bQ + cR$, on a

$$\begin{aligned} D^2 \frac{dx}{du} &= D(aP' + bQ' + cR')X \\ &\quad - (aP_1 + bQ_1 + cR_1)(aP' + bQ' + cR'), \\ D^2 \frac{dx}{du} &= (aP' + bQ' + cR')(DX - Dx), \end{aligned}$$

d'où

$$\frac{X - x}{\dfrac{dx}{du}} = \frac{Y - y}{\dfrac{dy}{du}} = \frac{Z - z}{\dfrac{dz}{du}},$$

qui exprime la propriété indiquée. Les formules (C') sont équivalentes aux formules (C). On peut d'ailleurs les identifier. Posons par exemple

$$P_1 = \frac{\begin{vmatrix} A_1 & A'_1 & A''_1 \\ \beta & \beta' & \beta'' \\ \gamma & \gamma' & \gamma'' \end{vmatrix}}{\begin{vmatrix} \alpha & \alpha' & \alpha'' \\ \beta & \beta' & \beta'' \\ \gamma & \gamma' & \gamma'' \end{vmatrix}}, \qquad P = \frac{\begin{vmatrix} A_4 & A'_4 & A''_4 \\ \beta & \beta' & \beta'' \\ \gamma & \gamma' & \gamma'' \end{vmatrix}}{\begin{vmatrix} \alpha & \alpha' & \alpha'' \\ \beta & \beta' & \beta'' \\ \gamma & \gamma' & \gamma'' \end{vmatrix}}.$$

Nous aurons, en vertu d'une identité analogue à l'identité (6),

$$P'_1 = \begin{vmatrix} A_1 & A'_1 & A''_1 & A'''_1 \\ \alpha & \alpha' & \alpha'' & \alpha''' \\ \beta & \beta' & \beta'' & \beta''' \\ \gamma & \gamma' & \gamma'' & \gamma''' \end{vmatrix} \times (\gamma\beta' - \beta\gamma')$$

et

$$P' = \begin{vmatrix} A_4 & A'_4 & A''_4 & A'''_4 \\ \alpha & \alpha' & \alpha'' & \alpha''' \\ \beta & \beta' & \beta'' & \beta''' \\ \gamma & \gamma' & \gamma'' & \gamma''' \end{vmatrix} \times (\gamma\beta' - \beta\gamma').$$

On voit de suite quelles valeurs il faudrait prendre pour P_2, ... et l'on a bien

$$\frac{P'_1}{P'} = X, \qquad \dots \qquad \text{avec} \qquad X = \frac{X_1}{X_4},$$

X_1 et X_4 étant les valeurs de la première solution.

Détermination d'une congruence Σ quand on donne la courbe lieu des sommets S et la développable enveloppe des plans P_u avec la correspondance entre les sommets et les plans.

Voici la solution avec mes formules. Je me donne les fonctions A_k et, en outre, les trois fonctions F_1, F_2,

F_3 de la variables u telles que

$$\frac{X_1}{X_4} = F_1, \qquad \frac{X_2}{X_4} = F_2, \qquad \frac{X_3}{X_4} = F_3.$$

On trouve alors, pour déterminer la fonction α, l'équation différentielle du troisième ordre

$$\begin{vmatrix} \alpha & \alpha' & \alpha'' & \alpha''' \\ A_1 - F_1 A_4 & A'_1 - F_1 A'_4 & A''_1 - F_1 A''_4 & A'''_1 - F_1 A'''_4 \\ A_2 - F_2 A_4 & A'_2 - F_2 A'_4 & A''_2 - F_2 A''_4 & A'''_2 - F_2 A'''_4 \\ A_2 - F_3 A_4 & A'_3 - F_3 A'_4 & A''_3 - F_3 A''_4 & A'''_3 - F_3 A'''_4 \end{vmatrix} = 0;$$

β et γ vérifient la même équation. Je laisse au lecteur le soin de développer cette méthode; voici celle qu'on peut déduire des calculs de M. Blutel. Soit

$$U_x + V_y + W_z + 1 = 0$$

le plan P_u. On se donne U, V, W, X, Y, Z comme des fonctions de u. Il s'agit de déterminer P, Q, R, $P_1, \ldots, R_3$.

Le plan P_u contient le point $\frac{P_1}{P}, \frac{P_2}{P}, \frac{P_3}{P}$; donc

$$P + UP_1 + VP_2 + WP_3 = 0.$$

Dérivons trois fois en u et posons pour abréger

$$\begin{aligned} H &= U\ X + V\ Y + W\ Z, \\ K &= U'X + V'Y + W'Z, \\ L &= U''X + V''Y + W''Z. \end{aligned}$$

Nous aurons

$$\begin{gathered} UP_1 + VP_2 + WP_3 + P = 0, \\ U'P_1 + V'P_2 + W'P_3 + P' + P'H = 0, \\ U''P_1 + V''P_2 + W''P_3 + P'K + P'' + P''H + P'H' = 0, \\ U'''P_1 + V'''P_3 + W'''P_3 + P'L \\ \qquad P''' + P''K + P''' + H + 2P''H' + P'H'' = 0 \end{gathered}$$

L'élimination de P_1, P_2, P_3 entre ces quatre équations donne, pour déterminer P, une équation différentielle du troisième ordre; P une fois connu, les équations précédentes donnent P_1, P_2, P_3. On a des calculs identiques pour Q_1, Q_2, Q_3, R_1, R_2, R_3. L'équation différentielle qui est linéaire et homogène par rapport à la fonction inconnue et à ses trois dérivées est d'ailleurs la même dans les trois cas. Soient donc P, Q, R trois intégrales de cette équation, les valeurs les plus générales de P, Q, R seront

$$\begin{gathered} a_1 P + b_1 Q + c_1 R, \\ a_2 P + b_2 Q + c_2 R, \\ a_3 P + b_3 Q + c_3 R; \end{gathered}$$

d'où, pour P_1,

$$a_1 P_1 + b_1 Q_1 + c_1 R_1; \qquad \ldots,$$

d'où, en remplaçant dans les formules (C'),

$$\begin{aligned} x &= \frac{a(a_1P_1+b_1Q_1+c_1R_1)+b(a_2P_1+b_2Q_1+c_2R_1)+c(a_3P_1+b_3Q_1+c_3R_1)}{a(a_1P\ +b_1Q_1+c_1R\ )+b(a_2P\ +b_2Q\ +c_2R\ )+c(a_3P\ +b_3Q\ +c_3Q\ )}, \\ y &= \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \\ z &= \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \end{aligned}$$

ou, avec un changement de notations,

$$\begin{aligned} x &= \frac{AP_1 + BQ_1 + CR_1}{AP\ + BQ\ + CR}, \\ y &= \frac{AP_2 + BQ_2 + CR_2}{AP\ + BQ\ + CR}, \\ z &= \frac{AP_3 + BQ_3 + CR_3}{AP\ + BQ\ + CR}, \end{aligned}$$

avec trois constantes A, B, C ou plutôt deux constantes, $\frac{A}{C}$, $\frac{B}{C}$ par exemple. On a donc une seule congruence Σ.

Conclusion. — Il résulte de là que le problème que nous nous sommes proposé conduit exactement aux mêmes calculs que le problème analogue dans le cas des coniques. Toute la difficulté consistera à résoudre l'équation différentielle du troisième ordre.

Remarque. — Dans un article précédent j'avais considéré les surfaces Σ dans le cas où la courbe plane était indéformable, il fallait donc que cette courbe admette un groupe de transformations homographiques; on sait qu'il en est bien ainsi pour les courbes triangulaires que j'ai trouvées.

Surfaces de deuxième espèce.

Je supposerai que le plan P_u roule sur un cône ayant son sommet à l'origine.

On trouve facilement, pour la congruence Σ, les équations

$$(\mathrm{E}) \qquad x_k = \frac{\begin{vmatrix} 0 & A_k & A'_k \\ a & \alpha & \alpha' \\ b & \beta & \beta' \end{vmatrix}}{\begin{vmatrix} 1 & A_4 & A'_4 \\ a & \alpha & \alpha' \\ b & \beta & \beta' \end{vmatrix}} \qquad (k = 1, 2, 3).$$

Le sommet du cône a pour coordonnées

$$X_k = \frac{\begin{vmatrix} A_k & A'_k & A''_k \\ \alpha & \alpha' & \alpha'' \\ \beta & \beta' & \beta'' \end{vmatrix}}{\begin{vmatrix} A_4 & A'_4 & A''_4 \\ \alpha & \alpha' & \alpha'' \\ \beta & \beta' & \beta'' \end{vmatrix}}.$$

Les plans P_u roulent sur un cône ayant pour sommet l'origine des coordonnées et contenant la courbe

$$x_k = A_k \qquad (k = 1, 2, 3).$$

Avec la méthode de M. Blutel, j'adopterai les formules

$$(\mathrm{E}') \qquad \left\{ \begin{aligned} x &= \frac{aP_1 + bQ_1}{aP + bQ + 1}, \\ y &= \frac{aP_2 + bQ_2}{aP + bQ + 1}, \\ z &= \frac{aP_3 + bQ_3}{aP + bQ + 1}, \end{aligned} \right.$$

avec

$$\frac{P'_1}{P'} = \frac{Q'_1}{Q'} = X,$$

$$\frac{P'_2}{P'} = \frac{Q'_2}{Q'} = Y,$$

$$\frac{P'_3}{P'} = \frac{Q'_3}{Q'} = Z.$$

On se donne alors X, Y, Z sommet du cône puis P, Q; on a alors P_1, P_2, P_3, Q_1, Q_2, Q_3 par six quadratures. Le plan P_u passe par l'origine et les deux points de coordonnées P_1, P_2, P_3 et Q_1, Q_2, Q_3.

Les deux points $\frac{P_1}{P}$, $\frac{P_2}{P}$, $\frac{P_3}{P}$ et $\frac{Q_1}{Q}$, $\frac{Q_2}{Q}$, $\frac{O_3}{Q}$ décrivent deux courbes dont les tangentes se coupent au point X, Y, Z.

Avec ces notations on peut reprendre les calculs faits plus haut pour trouver la congruence Σ quand on se donne la courbe lieu des sommets des cônes et la développable enveloppe des plans P_u avec la correspondance entre les sommets et les plans, c'est-à-dire X, Y, Z, U, V, W; le plan P_u ayant pour équation

$$U_x + V_y + W_z = 0.$$

On trouve alors, pour calculer P et Q, une équation du second ordre linéaire et homogène par rapport à la fonction inconnue, et ses deux premières dérivées se ramenant par suite à une équation de Riccati.

Exemple. — On trouve un exemple intéressant en partant des surfaces Σ de révolution. Si le plan P_u roule sur un cône de révolution, S étant lié au plan, on est ramené à une équation différentielle du type suivant (Oz est l'axe du cône) qu'on est certain de pouvoir intégrer

$$y'' - \frac{2\alpha(x-a)}{1+x^2}y' + \frac{2\beta(x^2-2ax+b)}{(1+x^2)^2}y = 0,$$

où α, β, a sont des constantes et y la fonction inconnue de x. On suppose toutefois α, β liées par la relation

$$\alpha^2+\alpha = 2\beta$$

ou

$$\beta = \frac{\alpha(\alpha+1)}{2}.$$

On peut ramener cette équation à la forme

$$\frac{y''}{y} = \frac{K}{(1+x^2)^2},$$

qui s'intègre facilement.

Surfaces de première espèce.

Généralisation des surfaces de translation.

Ces surfaces ont des équations de la forme

$$x_k = A_k + B_k \qquad (k = 1, 2, 3, 4),$$

où les A_k sont des fonctions de u et les B_k de v. Le long des courbes u (ou v) il existe un cône circonscrit à la surface, il resterait à exprimer que les courbes u sont planes. Pour le moment je garderai ces surfaces dans toute leur généralité. On a là une généralisation des surfaces de translation. Les surfaces transformées par polaires réciproques ont été étudiées par M. Darboux au Tome I de la *Théorie des Surfaces*. Ce sont les surfaces pour lesquelles il existe deux familles conjuguées formées exclusivement de courbes planes. Ici les surfaces que j'étudie sont caractérisées par ce fait qu'il existe un cône circonscrit le long de toutes les courbes du système conjugué u, v. Appelons Σ_1 ces nouvelles surfaces, nous allons en donner une définition géométrique.

Définition des surfaces Σ_1. — Donnons-nous deux courbes, l'une A décrite par le point M′

$$x' = A_1, \qquad y' = A_2, \qquad z' = A_3,$$

A_1, A_2, A_3 étant des fonctions quelconques de u. L'autre courbe B sera décrite par le point M″

$$x'' = B_1, \qquad y'' = B_2, \qquad z'' = B_3, \qquad \text{fonctions de U.}$$

Donnons-nous encore deux autres fonctions A_4 et B_4 l'une de u l'autre de v et sur la droite M′M″ marquons le point M déterminé par

$$\frac{\overline{\mathrm{MM'}}}{\overline{\mathrm{MM''}}} = -\frac{A_4}{B_4}.$$

Les coordonnées du point M seront

$$(\Sigma_1) \qquad \left\{ \begin{aligned} x &= \frac{A_1 B_4 + A_4 B_1}{A_4 + B_4}, \\ y &= \frac{A_2 B_4 + A_4 B_2}{A_4 + B_4}, \\ z &= \frac{A_3 B_4 + A_4 B_3}{A_4 + B_4}. \end{aligned} \right.$$

Comme on peut écrire

$$x = \frac{\frac{A_1}{A_4} + \frac{B_1}{B_4}}{\frac{1}{A_4} + \frac{1}{B_4}},$$

on voit qu'on a bien les mêmes surfaces que celles définies plus haut.

Les courbes u se correspondent point par point et cette correspondance est établie au moyen des courbes conjuguées v. Dans les surfaces de translation les droites joignant les points correspondants de deux courbes u sont égales, parallèles et de même sens; retenons cette propriété que les droites joignant les points correspondants des courbes u (ou v) forment un cylindre. Comme généralisation je vais démontrer le théorème suivant :

Théorème. — *Dans les surfaces* Σ_1 *les droites qui joignent les points correspondants de deux courbes* u (*ou* v) *quelconques forment un cône.*

Donnons à u deux valeurs fixes et soient a et a_1 les valeurs correspondantes de la fonction A_4. Soient M' et M'_1 les positions fixes correspondantes du point M'; enfin M'' correspond à une valeur de v qu'on fera varier. On a ainsi deux positions de M : M et M_1 déterminées par

$$\frac{MM'}{MM''} = -\frac{a}{B_4}, \qquad \frac{M_1M'_1}{M_1M''} = -\frac{a_1}{B_4}.$$

Or

$$\frac{MM'}{MM''}\,\frac{M_1M''}{M_1M'_1}\,\frac{PM'_1}{PM'} = +1.$$

Donc

$$\frac{PM'}{PM'_1} = \frac{a}{a_1},$$

qui prouve que le point P est fixe.

Comme on peut écrire

$$x = \frac{\dfrac{A_1}{A_4} + \dfrac{B_1}{B_4}}{\dfrac{1}{A_4} + \dfrac{1}{B_4}},$$

on voit qu'on a bien les mêmes surfaces que celles définies plus haut.

Les courbes u se correspondent point par point et cette correspondance est établie au moyen des courbes conjuguées v. Dans les surfaces de translation les droites joignant les points correspondants de deux courbes u sont égales, parallèles et de même sens; retenons cette propriété que les droites joignant les points correspondants des courbes u (ou v) forment un cylindre. Comme généralisation je vais démontrer le théorème suivant :

Théorème. — *Dans les surfaces Σ_1 les droites qui joignent les points correspondants de deux courbes u (ou v) quelconques forment un cône.*

Donnons à u deux valeurs fixes et soient a et a_1 les valeurs correspondantes de la fonction A_4. Soient M' et M'_1 les positions fixes correspondantes du point M'; enfin M'' correspond à une valeur de v qu'on fera varier. On a ainsi deux positions de M : M et M_1 déterminées par

$$\frac{MM'}{MM''} = -\frac{a}{B_4}, \qquad \frac{M_1M'_1}{M_1M''} = -\frac{a_1}{B_4}.$$

Or

$$\frac{MM'}{MM''}\,\frac{M_1M''}{M_1M'_1}\,\frac{PM'_1}{PM'} = +1.$$

Donc

$$\frac{PM'}{PM'_1} = \frac{a}{a_1},$$

qui prouve que le point P est fixe.

D'autre part

$$\frac{MM'}{a} = \frac{MM''}{-B_4} = \frac{M''M'}{a + B_4},$$

$$\frac{M_1M'_1}{a_1} = \frac{M_1M''}{-B_4} = \frac{M''M'_1}{a_1 + B_4},$$

$$\frac{PM}{PM_1}\,\frac{M'_1M_1}{M_1M''}\,\frac{M'M''}{M'M} = +1.$$

Donc

$$\frac{PM}{PM_1} = \frac{a(a_1 + B_4)}{a_1(a + B_4)}$$

montre que le rapport $\frac{PM}{PM_1}$ varie avec v et peut être une fonction arbitraire de cette variable.

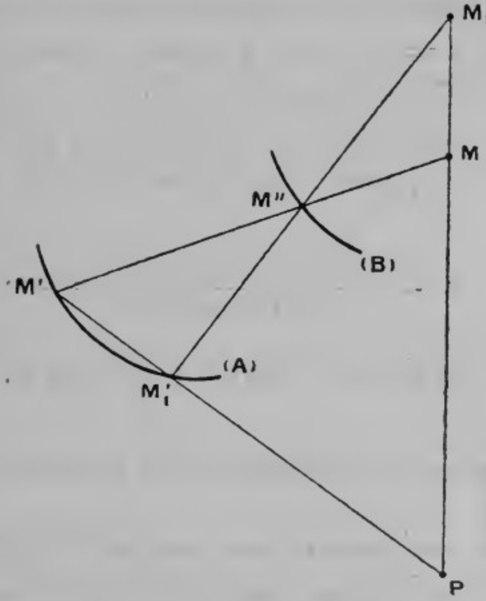

Corollaire. — Il en résulte facilement que :

1° Trois courbes u sont situées deux à deux sur trois cônes et que les sommets des cônes sont en ligne droite ;

2° Quatre courbes u sont deux à deux sur six cônes et les six sommets forment les six sommets d'un quadrilatère complet.

Surface lieu du sommet des cônes.

Sur la corde $M'M'_1$ la position du point P est définie par l'égalité

$$\frac{PM'}{PM'_1} = \frac{A_4(u)}{A_4(u_1)}$$

si M' et M'_1 correspondent respectivement aux valeurs u et u_1. Cette surface (P) est donc définie d'une façon tout à fait analogue à la surface Σ_1, avec cette différence qu'on se sert d'une seule courbe au lieu de deux et d'une seule fonction A_4. Les courbes u et u_1 tracent sur cette surface (P) un système conjugué symétrique et la courbe $u = u_1$ est le lieu des sommets des cônes circonscrits à Σ_1 le long des courbes u de cette surface. En tous les points de cette courbe $u = u_1$, les courbes u et u_1 sont tangentes. Les équations de la surface (P) sont

$$\text{(P)} \qquad x = \frac{A_1(u)A_4(u_1) - A_4(u)A_1(u_1)}{A_4(u_1) - A_4(u)};$$

pour y et z on change l'indice 1 en 2 ou 3.

Cas particulier des surfaces de première espèce.

Je reviens maintenant aux surfaces Σ de première espèce, c'est-à-dire pour lesquelles le plan des courbes u passe par une droite fixe Oz, par exemple :

$$x = \frac{A_1 + B_1}{A_4 + B_4}, \qquad y = \frac{A_2 + B_2}{A_4 + B_4}, \qquad z = \frac{A_3 + B_3}{A_4 + B_4}.$$

Mais pour que $\frac{y}{x}$ dépende de u et pas de v, on montre facilement qu'il faut que B_1 et B_2 soient constants et alors on peut les supposer nuls. Un simple changement de notations conduit alors aux formules

$$x = \frac{A_1 B_4}{A_4 + B_4},$$

$$y = \frac{A_2 B_4}{A_4 + B_4}.$$

$$z = \frac{A_3 B_4 + A_4 B_3}{A_4 + B_4}$$

qui sont les formules (Σ_1) où l'on fait

$$B_1 = 0, \qquad B_2 = 0.$$

Il suffit donc de supposer que la courbe (B) décrite par le point M″ soit une droite pour avoir les nouvelles surfaces. Donc on peut énoncer le théorème suivant :

Théorème. — *Des surfaces* Σ *de première espèce sont telles que deux courbes planes u se correspondent par homologie. Il en résulte que, si l'une de ces courbes coupe l'axe* O*z en un certain nombre de points, les autres courbes u passent par les mêmes points.*

Remarque. — Pour obtenir les surfaces Σ_1 au lieu de poser

$$\frac{MM'}{MM''} = -\frac{A_4}{B_4},$$

on pourrait prendre

$$\frac{MM'}{MM''} = \frac{A_4 + K}{B_4 + K},$$

k étant une constante. Chaque valeur de *k* donne une surface, on a ainsi des surfaces qui se correspondent point par point avec plans tangents parallèles. On trouve ainsi un cas particulier intéressant du théorème suivant, énoncé par M. Darboux au Tome II de la *Théorie des Surfaces :*

Si l'on mène à deux surfaces S *et* S_1 *des plans tangents parallèles, la droite qui joint les points de contact engendre une congruence dont les développables interceptent sur* S *et* S_1 *un réseau conjugué.*

[K¹14f]

SUR UN HEXAÈDRE PARTICULIER;

Par M. R. BRICARD.

1. Il existe dans le plan deux espèces de quadrilatères ayant leurs côtés opposés égaux deux à deux : le *parallélogramme* et le *contre-parallélogramme* (figure formée par les côtés non parallèles et les diagonales d'un trapèze isoscèle).

Dans l'espace, il existe un hexaèdre, le *parallélépipède*, dont toutes les faces sont des parallélogrammes. On peut rechercher, par analogie, s'il existe un hexaèdre dont toutes les faces sont des contre-parallélogrammes. Nous allons voir que la réponse est affirmative. Donnons, dès maintenant, le nom de *contre-parallélépipède* à l'hexaèdre satisfaisant.

2. Soit $abcda'b'c'd'$ le schéma d'un contre-parallélépipède (*fig.* 1). Les lettres ont été tellement disposées que deux quelconques des sommets a, b, c, d soient les sommets opposés d'une même face de l'hexaèdre; a', b', c', d' sont les sommets de l'hexaèdre respectivement opposés aux premiers.

Par hypothèse, le quadrilatère $a'cb'd$ par exemple est un contre-parallélogramme; ses diagonales $a'b'$ et cd sont donc parallèles. De même, $a'c'$ est parallèle à db, et $a'd'$ est parallèle à bc. Autrement dit, $a'b'$, $a'c'$

' 2
>) 2) !) 2 RS
> ! 2) A

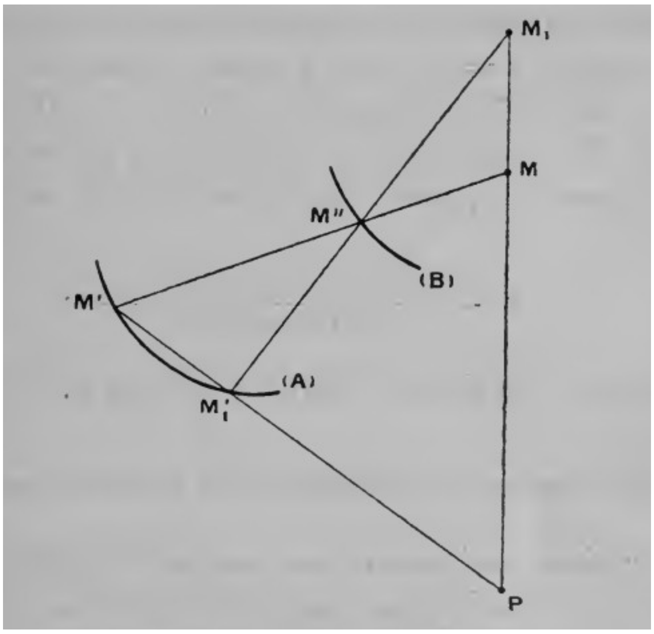

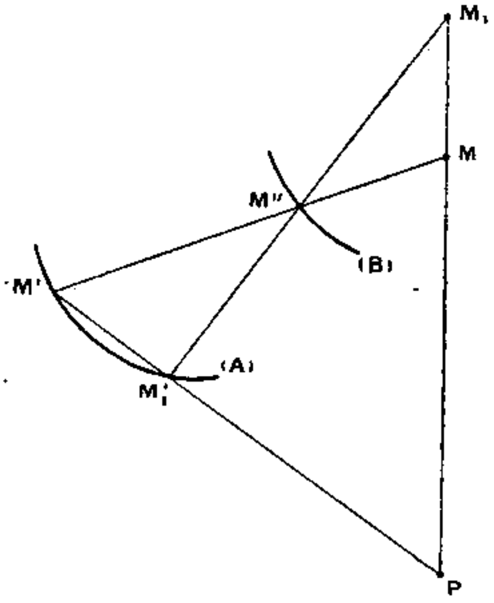

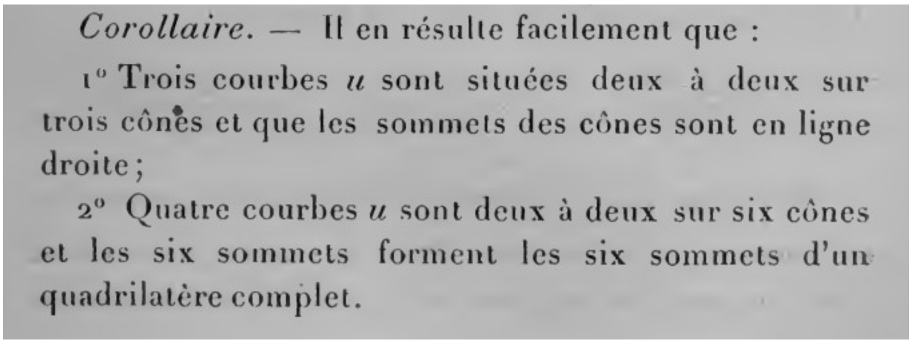

Corollaire. — Il en résulte facilement que :

1° Trois courbes *u* sont situées deux à deux sur trois cônes et que les sommets des cônes sont en ligne droite ;

2° Quatre courbes *u* sont deux à deux sur six cônes et les six sommets forment les six sommets d'un quadrilatère complet.

+9 ! 2 L
) % & 8 G Q! D !

Si l'on mène à deux surfaces S *et* S_1 *des plans tangents parallèles, la droite qui joint les points de contact engendre une congruence dont les développables interceptent sur* S *et* S_1 *un réseau conjugué.*

[$K^1$14f]

SUR UN HEXAÈDRE PARTICULIER;

Par M. R. BRICARD.

1. Il existe dans le plan deux espèces de quadrilatéres ayant leurs côtés opposés égaux deux à deux : le *parallélogramme* et le *contre-parallélogramme* (figure formée par les côtés non parallèles et les diagonales d'un trapèze isoscèle).

Dans l'espace, il existe un hexaèdre, le *parallélépipède*, dont toutes les faces sont des parallélogrammes. On peut rechercher, par analogie, s'il existe un hexaèdre dont toutes les faces sont des contre-parallélogrammes. Nous allons voir que la réponse est affirmative. Donnons, dès maintenant, le nom de *contre-parallélépipède* à l'hexaèdre satisfaisant.

2. Soit $abcda'b'c'd'$ le schéma d'un contre-parallélépipède (*fig.* 1). Les lettres ont été tellement disposées que deux quelconques des sommets a, b, c, d soient les sommets opposés d'une même face de l'hexaèdre; a', b', c', d' sont les sommets de l'hexaèdre respectivement opposés aux premiers.

Par hypothèse, le quadrilatère $a'cb'd$ par exemple est un contre-parallélogramme; ses diagonales $a'b'$ et cd sont donc parallèles. De même, $a'c'$ est parallèle à db, et $a'd'$ est parallèle à bc. Autrement dit, $a'b'$, $a'c'$

et $a'd'$ sont respectivement parallèles aux côtés du triangle bcd; ces trois droites sont donc dans un même plan, d'ailleurs parallèle au plan du triangle. On verra, par un raisonnement semblable, que ce dernier plan contient aussi le point a.

Ainsi, les deux quadrangles $abcd$ et $a'b'c'd'$ appartiennent respectivement à deux plans parallèles P et P'.

Fig. 1.

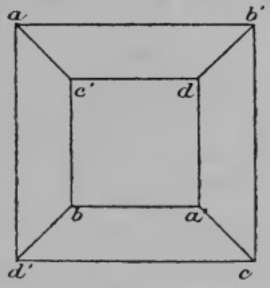

En outre, les côtés du premier quadrangle sont parallèles à ceux du second, deux côtés parallèles étant désignés par des lettres différentes (par exemple ab et $c'd'$). On voit qu'à trois côtés concourants de l'un des quadrangles correspondent dans l'autre trois côtés formant un triangle. On reconnaît là les *quadrangles réciproques* de la statique graphique.

Pour aller plus loin, faisons une projection orthogonale de l'hexaèdre sur le plan P. Le contre-parallélogramme $a'db'c$ restera visiblement tel en projection : il en résulte que les médiatrices des segments $a'b'$ et cd (c'est-à-dire les perpendiculaires élevées aux milieux de ces segments) sont confondues. De même les médiatrices des segments $a'c'$ et db, et celles des segments $a'd'$ et bc. Mais les médiatrices des segments cd, db, bc concourent en un point O, centre du cercle circonscrit au triangle bcd. Il résulte immédiatement de là que les quatre points a', b', c', d' appartiennent à un cercle ayant le même centre O. On verra aussi que le point a appartient au premier cercle contenant les points b, c, d. En résumé les deux quadrangles $abcd$ et $a'b'c'd'$ sont, en projection sur le plan P, inscrits à deux cercles concentriques.

Il est facile d'achever. Soient α, β, ..., γ', δ' les angles que font les rayons Oa, Ob, ..., Oc', Od' avec un axe quelconque Ox. Pour exprimer que le quadrilatère $a'cb'd$ (*fig.* 2) est un contre-parallélogramme, il suffit évidemment d'écrire que les deux angles $\widehat{a'Ob'}$ et $\widehat{cOd}$ ont les mêmes bissectrices, ce qui donne

Fig. 2.

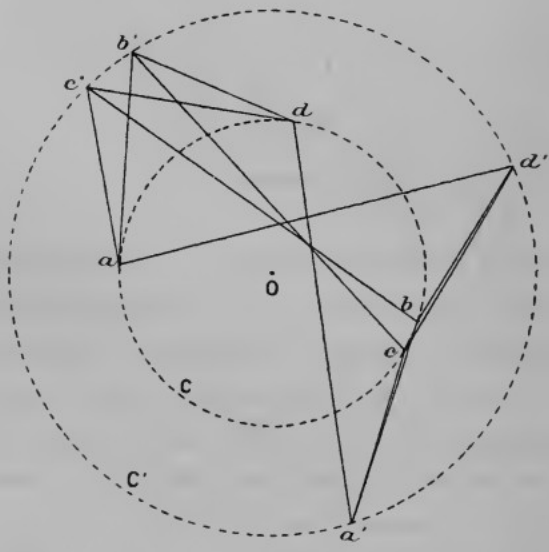

$$\alpha' + \beta' = \gamma + \delta.$$

On aura de même

$$\alpha' + \gamma' = \beta + \delta,$$
$$\alpha' + \delta' = \beta + \gamma,$$
$$\beta' + \gamma' = \alpha + \delta,$$
$$\beta' + \delta' = \alpha + \gamma,$$
$$\gamma' + \delta' = \alpha + \beta.$$

On reconnaît tout de suite que ces six équations sont compatibles, *et cela établit définitivement l'existence du contre-parallélépipède.* En résolvant le système par rapport à α', β', γ', δ', il vient

$$\alpha' = \frac{-\alpha+\beta+\gamma+\delta}{2},$$

$$\beta' = \frac{\alpha-\beta+\gamma+\delta}{2},$$

$$\gamma' = \frac{\alpha+\beta-\gamma+\delta}{2},$$

$$\delta' = \frac{\alpha+\beta+\gamma-\delta}{2}.$$

On peut donc se donner arbitrairement les points a, b, c, d sur le cercle C. Les positions des points a', b', c', d' sur le cercle C′ en résultent. Pratiquement, il vaut mieux se donner trois points sur le cercle C et un point sur le cercle C′, par exemple les points a, b, c et d'. On obtient les autres points en utilisant la réciprocité des deux quadrangles. On aboutit ainsi à la construction suivante :

Tracer deux cercles concentriques C *et* C′ ; *marquer arbitrairement les points* a, b, c *sur le premier et* d' *sur le second. Construire successivement les points* c', a', d *et* b', *appartenant au cercle* C′ *ou au cercle* C *suivant qu'ils sont désignés par des lettres accentuées ou non, en menant* $d'c'$, *parallèle à* ba ; $d'a'$, *parallèle à* bc ; bd, *parallèle à* $a'c'$; $c'b'$, *parallèle à* ad.

Les points a, b, c, d *et ceux qu'on obtient en donnant au quadrangle* $a'b'c'd'$ *une translation quelconque perpendiculaire au plan de la figure sont les sommets d'un contre-parallélépipède.*

On reconnaît tout de suite que ces six équations sont compatibles, *et cela établit définitivement l'existence du contre-parallélépipède*. En résolvant le système par rapport à α', β', γ', δ', il vient

$$\alpha' = \frac{-\alpha+\beta+\gamma+\delta}{2},$$
$$\beta' = \frac{\alpha-\beta+\gamma+\delta}{2},$$
$$\gamma' = \frac{\alpha+\beta-\gamma+\delta}{2},$$
$$\delta' = \frac{\alpha+\beta+\gamma-\delta}{2}.$$

On peut donc se donner arbitrairement les points a, b, c, d sur le cercle C. Les positions des points a', b', c', d' sur le cercle C' en résultent. Pratiquement, il vaut mieux se donner trois points sur le cercle C et un point sur le cercle C', par exemple les points a, b, c et d'. On obtient les autres points en utilisant la réciprocité des deux quadrangles. On aboutit ainsi à la construction suivante :

Tracer deux cercles concentriques C *et* C' ; *marquer arbitrairement les points* a, b, c *sur le premier et* d' *sur le second. Construire successivement les points* c', a', d *et* b', *appartenant au cercle* C' *ou au cercle* C *suivant qu'ils sont désignés par des lettres accentuées ou non, en menant* $d'c'$, *parallèle à* ba ; $d'a'$, *parallèle à* bc ; bd, *parallèle à* $a'c'$; $c'b'$, *parallèle à* ad.

Les points a, b, c, d *et ceux qu'on obtient en donnant au quadrangle* $a'b'c'd'$ *une translation quelconque perpendiculaire au plan de la figure sont les sommets d'un contre-parallélépipède.*

3. Le contre-parallélépipède jouit de propriétés qui résultent immédiatement de sa construction. Tout d'abord, il est évident qu'il est inscrit à une sphère. On voit tout aussi facilement que les douze arêtes du polyèdre sont réparties quatre par quatre sur trois hyperboloïdes de révolution autour de l'axe commun des cercles C et C', savoir :

H_1, contenant les arêtes ab', ba', cd', dc' ;
H_2, contenant les arêtes ac', ca', bd', db' ;
H_3, contenant les arêtes ad', da', bc', cb'.

Il existe d'autres hyperboloïdes attachés au contre-parallélépipède. Considérons en effet les deux systèmes de quatre droites

$$ab, \quad a'b', \quad cd', \quad dc'$$

et

$$ab', \quad ba', \quad cd, \quad c'd';$$

on se rend compte que chaque droite de l'un des systèmes rencontre chaque droite de l'autre système (à distance finie ou infinie). Ces huit droites sont donc à

Fig. 3.

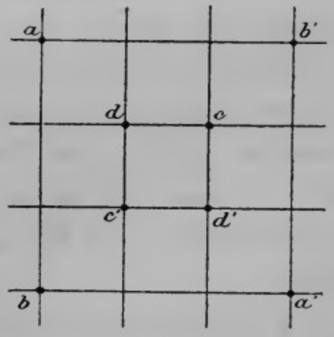

un hyperboloïde H'_1 où elles se répartissent en deux groupes de génératrices de l'un et de l'autre système, comme le montre le schéma de la figure 3. La configuration des huit sommets forme donc ce que M. Fontené appelle un *octuple gauche complet*.

L'hyperboloïde H'_1 touche les deux plans P et P′, puisqu'il les coupe respectivement suivant les génératrices ab et cd, $a'b'$ et $c'd'$. Ces deux plans étant parallèles, le centre de l'hyperboloïde se trouve sur le plan parallèle et équidistant P″.

Il existe deux hyperboloïdes analogues :

$$H'_2, \text{ contenant les droites } \left\{ \begin{array}{llll} ac, & a'c', & bd', & db'. \\ ac', & ca', & bd, & b'd', \end{array} \right.$$

et

$$H'_3, \text{ contenant les droites } \left\{ \begin{array}{llll} ad, & a'd', & bc', & cb', \\ ad', & da', & bc, & b'c', \end{array} \right.$$

Ils ont aussi leur centre sur le plan P″.

Enfin une propriété intéressante du contre-parallélépipède est relative à la possibilité de sa déformation avec conservation des longueurs des arêtes. En effet, le polyèdre le plus général de cette nature dépend de six paramètres de grandeur, à savoir les rayons des cercles C et C′, les longueurs des segments ab, ac, ad', et la valeur de la translation finale qui sert à le définir. D'autre part, ses douze arêtes sont égales quatre par quatre, et la connaissance de leurs longueurs se traduit par trois relations seulement entre les six paramètres. On voit donc *qu'il existe une triple infinité de contre-parallélépipèdes ayant des arêtes de longueurs données*.

Il est curieux de constater que le contre-parallélépipède se comporte, au point de la déformabilité, exactement comme le parallélépipède.

+9 O ! 2 2L2 ! ! 2 H
) ! '

' "

./01. . .+ . " $- %% 2!

) "(-% " / !

C 52 T ! 2 5 !
2 2 A ! 9 A ! !
2 2 !

./01. . .+ . " I B / U
I A $A A D 7 I 5 2!

& ./01. . .+ . " I B / (A
D B 2 * / ' 2 8 7 ! O 7 5 2!

% ./01. . .+ . " I B A
D B 2 2!

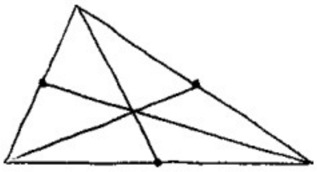

9 C 2 0 7 5 5 :

./01. . .+ . " &"(&%$&% 2!

)

date 20.02.2020
S T Q [✓ Q] S S D

Calculer le rapport de la circonférence au diamètre

Pág. 6(11)

c) s4nouvelleannal13pari.pdf

Surfaces engendrées par le déplacement d'une courbe plane avec cone circonscrit le long de la courbe

Pág. 1(9)

116) 20200223 Source: Data_novoVIVO_38_2020/Books-38_2020/Measurement/Lockhart/sanet_st_9811210144.pdf

a) An Eventful Journey to Unification of All the Fundamental Forces
Delbourgo

Pág. 67(84)-

6.1. Anticommuting coordinates.

There are two types of fields in physics: those connected with integer spin particles, called bosons, and those connected with half-integer spin particles called fermions.

b) The Mathematical Principles of Scale Relativity Physics
Mazilu.

23.07.2020

Chapter 7. The Physical Point of View in the Theory of Surfaces.

Pag. 110 (115).

We start from the idea that a surface should play a double role in physics.

matter from space of the space from matter

for there are altogether two different situations. The statement cannot be made understandable without

=> Second role:

that of locus of an ensemble of positions in the geometrical acceptance.

Source: Data_Novoviva_38_2020/Books_38_2020/the Mathematical Principles of Scale Relativity Physics.../sanet_st_0367349345.pdf.

a) Solid Mensuration with proofs
Kern

Chapter 7. The Physical Point of View in the Theory of Surfaces

We start from the idea that a surface should play a double role in physics. First, it plays the obvious role of separator, specifically for separating the matter from space or of the space from matter, for these are altogether two different situations. This statement cannot be made understandable without the second role of the surface which is closer to the regular mathematical role: that of *locus of an ensemble of positions* in the geometrical acceptance. It pays, however, to notice that it is in this instance that the surface allows the mathematical description and the consequent physical understanding of that property of the material points of 'penetrating without merging', to use the phrasing of Erwin Madelung. In order to get a better grip of this idea, some introductory observations seem to be in order here.

A Few Mathematical Prerequisites

It seems to us obvious that a physical theory of surfaces cannot be developed but by the idea of embedding. As a matter of fact, the embedding is an interpretation and, forcing a little the idea, the essential point of the Darwin's definition of interpretation is the existence of an ensemble of positions giving the possibility of interpretation of the surface as a geometrical locus. Here a position can be taken geometrically as a point and physically as a material particle in the sense of Hertz. In this last instance, the material particle is to us a little more than it was to Hertz himself, in the sense that we added to it the feature of *indicator of a point in space*, by a process of application of a field to a position, being it in space or in matter in a sense which will be explained here shortly. For instance, in keeping with the classical physics, one can imagine the point of application of a force or of a resultant of many forces, according to the third principle of classical dynamics, but this is quite a particular example.

The mathematical method itself, for carying out the task of embedding a surface, and therefore the task of interpretation in general, is based on an almost trivial statement that emerged apparently largely unnoticed, or if noticed, it has not been properly used for physical purposes. In order to make it obvious, we reproduce here two of the Élie Cartan's algebraical theorems which form the ground of his approach of differential geometry by *moving frames* [for a clear description of the idea from the point of view we adopt here, see (Spivak 1999), Volume II, Chapter 7]. These theorems are borrowed directly from one of Cartan's courses, published via the Russian geometrical school of S. P. Finikov [(Cartan, 2001); pp. 16 17, Theorems 7 & 9]. We just appropriate them here under the name of *Cartan Lemmas 1 and 2*, only in order to be suitably used in making our point as explicit as possible:

Lemma 1. Suppose that $s^1, s^2, ..., s^p$ is a set of linearly independent 1-forms. Then we have

$$s^\alpha \wedge \phi_\alpha = 0 \quad \leftrightarrow \quad \phi_\alpha = a_{\alpha\beta} s^\beta \quad , \quad a_{\alpha\beta} = a_{\beta\alpha}$$

where $\phi_1, \phi_2, ...\phi_p$ is another set of linearly independent 1-forms, and summation over repeated indices is understood.

Lemma 2. Suppose the basis differential elements $du^1, du^2, ..., du^n$ are connected by a system of equations

$$\omega^1 = 0, \quad \omega^2 = 0, ..., \ \omega^p = 0$$

where ω^α, $\alpha = 1, 2, ..., p$ are linearly independent 1-forms. Then the 2-form 'f' *vanishes as a consequence of this system* if, and only if, 'f' can be written as

$$f = \omega^\alpha \wedge \phi_\alpha$$

where, again, summation over α is understood, and ϕ_α are p *conveniently chosen* 1-forms.

The first one of these theorems is, by and large, known as *Cartan's Lemma* proper in the specialty literature, while the second one carries no specific name. What seems to be essential in these lemmas, and is always stressed in many old treatises, but apparently forgotten lately perhaps due to exclusive mathematical applications − is the fact that the symmetric matrix **a** from Lemma 1, as well as the 1-forms ϕ_α from Lemma 2, are *things external to the geometrical problem*, things through which the physics can therefore be introduced. In concentrating on the local description at a point of a surface, without being interested of the global aspects of that surface, as it is almost always the case in physics, this observation is essential. In this respect another convention referring to our use of indices is again in order: insofar as the space contemplated as environment in the embedding problem necessary to physical interpretation is here the usual, possibly Euclidean, three-dimensional space of our intuition, we reserve the Latin indices exclusively for this case. The Greek indices are used for any other dimension, as in the case above, but especially for dimension reproduce here two of the Élie Cartan's algebraical theorems which form the ground of his approach of differential geometry by *moving frames* [for a clear description of the idea from the point of view we adopt here, see (Spivak 1999), Volume II, Chapter 7]. These theorems are borrowed directly from one of Cartan's courses, published via the Russian geometrical school of S. P. Finikov [(Cartan, 2001); pp. 16 17, Theorems 7 & 9]. We just appropriate them here under the name of *Cartan Lemmas 1 and 2*, only in order to be suitably used in making our point as explicit as possible:

Lemma 1. Suppose that $s^1, s^2, ..., s^p$ is a set of linearly independent 1-forms. Then we have

$$s^\alpha \wedge \phi_\alpha = 0 \quad \leftrightarrow \quad \phi_\alpha = a_{\alpha\beta} s^\beta \quad , \quad a_{\alpha\beta} = a_{\beta\alpha}$$

where $\phi_1, \phi_2, ...\phi_p$ is another set of linearly independent 1-forms, and summation over repeated indices is understood.

Lemma 2. Suppose the basis differential elements $du^1, du^2, ..., du^n$ are connected by a system of equations

$$\omega^1 = 0, \quad \omega^2 = 0, ..., \ \omega^p = 0$$

where ω^α, $\alpha = 1, 2, ..., p$ are linearly independent 1-forms. Then the 2-form 'f' *vanishes as a consequence of this system* if, and only if, 'f' can be written as

$$f = \omega^\alpha \wedge \phi_\alpha$$

where, again, summation over α is understood, and ϕ_α are p *conveniently chosen* 1-forms.

The first one of these theorems is, by and large, known as *Cartan's Lemma* proper in the specialty literature, while the second one carries no specific name. What seems to be essential in these lemmas, and is always stressed in many old treatises, but apparently forgotten lately perhaps due to exclusive mathematical applications − is the fact that the symmetric matrix **a** from Lemma 1, as well as the 1-forms ϕ_α from Lemma 2, are *things external to the geometrical problem*, things through which the physics can therefore be introduced. In concentrating on the local description at a point of a surface, without being interested of the global aspects of that surface, as it is almost always the case in physics, this observation is essential. In this respect another convention referring to our use of indices is again in order: insofar as the space contemplated as environment in the embedding problem necessary to physical interpretation is here the usual, possibly Euclidean, three-dimensional space of our intuition, we reserve the Latin indices exclusively for this case. The Greek indices are used for any other dimension, as in the case above, but especially for dimension i.e. properties representative to matter as well as properties representative to space. In other words, here, on this surface, the matter meets the space, and physics needs to describe this 'collision of the two worlds', so to speak. In an electromagnetic structure, the space is simply represented by a 2-form, 'b' say, while the matter is represented by a 2-form, 'd' say. These symbols suggest electromagnetism in the sense of David van Dantzig, as representative for the purely electromagnetic structure of matter (van Dantzig, 1934). To wit, these exterior differential 2-forms can indeed be appropriated here as two forms denoted above ω^ν for a general formulation of theory:

$$\omega^1 \equiv b = b_{ij} dx^i \wedge dx^j; \quad \omega^2 \equiv d = d_{ij} dx^i \wedge dx^j.$$

They correspond to *magnetic induction* and respectively *electric induction*, from the classical theory of electricity and associated magnetism. An explanation is in order: if the surface delimiting matter can be globally defined, as in the case of Dirac's electron for instance, these two exterior forms satisfy the integral equations:

$$\oiint_{Surface} b_{ij} dx^i \wedge dx^j = 0;$$

$$\oiint_{Surface} d_{ij} dx^i \wedge dx^j = \iiint_{Volume} \rho_{ijk}\, dV^{ijk}$$

which justify their names, insofar as the magnetic field is source-free while the electric field has the electirc charge as its source. Here ρ is the third order tensor of the *electric density of matter*, while $d\mathbf{V}$ is the oriented volume 3-form in the space occupied by matter. These relations can be used to justify our choices for describing the matter and space in their common limit.

Indeed, in spite of the global aspect of the problem, as the previous integral equations seem to suggest, the physical definition of the two fields involved here *is always local*, at least experimentally. To wit, we have the infinitesimal *voids* in the magnetic matter for defining 'b', or the *dipoles* in the electric matter for defining 'd'. This means that the tensor **b** should be always realizable *in matter* as a vector, say **h**, defining the 1-form of the magnetic field. Likewise, the tensor **d** is representable *in space* by a vector, **e** say, defining the 1-form of electric field. It is here the point where a theorem which we termed as *Agaoka's theorem*, makes its mark of necessity, because we need a *constitutive law* relating the fields **b** and **d** to the fields **e** and **h**. Thus, between the corresponding differential forms we always need a general relationship of the form:

$$b \wedge h + d \wedge e = 0$$

for only in this case, *according to Agaoka's theorem*, there should be a matrix of exterior differential 1-forms, γ say, which realizes a linear correlation between the 2-forms and 1-forms:

$$b = \gamma^{11} \wedge h + \gamma^{12} \wedge e; \quad d = \gamma^{12} \wedge h + \gamma^{22} \wedge e.$$

In terms of the field vectors themselves, these equations can be turned into linear constitutive equations of the kind used in the classical theory of electromagnetism. The exhaustive characterization of electromagnetism from this point of view is due to Reginald Aubrey Fessenden, at the beginning of the last century (Fessenden, 1900). But the situation is quite general, so that now, after this introduction of the mathematical basis and the reasons for such an approach to a physics of surfaces, we need to present in detail the essence of the geometric theory necessary to achieve this physics.

The Differential Theory of Surfaces

Élie Cartan's exterior differential approach to the theory of surfaces, allows us to say that the local differential theory of surfaces is simply a consequence of the fact that the elementary (differential) displacement of a point of surface is an intrinsic vector of that surface (Guggenheimer, 1977). Everything in the differential theory of surfaces, and in the embedding theory in general, follows from this simple fact through the rules of handling of the differential forms. To wit, if $\mathbf{x}$ is the position vector of a point of surface, then $\mathrm{d}\mathbf{x}$ should be an intrinsic vector of the surface, at least as long as no phenomenon occurs forcing us to consider that the position leaves the surface. But even in such a case, important from a physical point of view, the elementary variation of position *in space* can be described by a process of deformation through a continuous family of surfaces containing the position during that variation. At first, the deformation appears as a purely mathematical device, but as soon as we get into details it becomes gradually obvious that it needs further consideration, especially from physical point of view.

Let us now lay down the proper mathematical form of this portrayal of surfaces. Referring the local geometry at a position $\mathbf{x}$ in space to an orthonormal frame of reference: $(\hat{\mathbf{e}}_|, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3) \equiv \langle \hat{\mathbf{e}}|$, we write

$$d\mathbf{x} = s^1\hat{\mathbf{e}}_1 + s^2\hat{\mathbf{e}}_2 + s^3\hat{\mathbf{e}}_3 \quad \leftrightarrow \quad d\mathbf{x} = \langle s|\hat{\mathbf{e}}\rangle. \tag{7.1}$$

These relations call for a further elucidation regarding the notations. A caret over a bold symbol means a unit vector. Further on, a 'Dirac representation' with 'kets' and 'bras' is used for any sequence of mathematically meaningful objects. For instance, $|s\rangle$ here represents the matrix 3×1 having the components of $\mathbf{x}$ as entries. It is thus understood that the vector notation $\mathbf{x}$ involves also the reference frame $|\hat{\mathbf{e}}\rangle$ in its definition, which, according to our notation

! 2 V

23.02.2020

Source: ARQ-IMPORTANT/Cinderella.de/
Pappus/
1) 2015.205959.Solid-
Mensuration.pdf

2) Sommerville D - An
introduction ... pdf

✡ 3) 10-piscoran-final.pdf

Pappus's harmonic theorem
in the Einstein relativistic
velocity model of
hyperbolic geometry.

Veja:

Cinderella.de/

{ Pappus_20200223.pdf
Pappus-~~202002~~ Spherical_20200223.pdf

Source:

Cinderella.de/ the Cinderella.2.Manual.
pdf

Pag. 43(58)-49(64)

4) Analytic Hyperbolic Geometry
in N Dimensions.
Ungar

23.02.2020

Pag. 590(611)

[87] Pappu's harmonic theorem in the Einstein relativistic.

---

3 cont.) Pappu's harmonic theorem in the Einstein relativistic

www.cs.ubbcluj.ro/~studia-m/
2011-1/10-piscoran-final.pdf

mathworld.wolfran.com/
PappusCentroidTheorem.html

theoremoftheday.org/Geometry
andTrigonometry/Guldin/
TotDPappusGuldin.pdf

---

117) 20200226

a) Sullivan Conjecture

https://ocw.mit.edu/faculty/
haynes-miller/

b) $2\pi = \tau$ (Tau)
the radio of the great pyramids perimeter to its height approximates 2π or what is know as TAU.

/01 ' +0C/>C# ! ! # 22 #

./01. . .+ . " $ $! 2!

./01. . .+ . " 2 2!

22 U M 7 7 5 ! 52 9 5

& ./01. . .+ . " 7 6 / !
5 > ! $ $A 2!

./01. . .+ . " 7 6 / !
5 > ! $ $A !H7

+

! ! #

22 * & 2!

22 * 2 * & 2!

! ! #C ! 2!

) %& - %$ "%

! / 5 852 9 4 5 > 6 O

) $ "

W-(X 22 P M 7

./01. . .+ . "
(*/9 */ 9 *O **/ 5 *852 9 *4 5* *>*6 *
/ *' ! * % 2!

) $ "

22 U M 7 7 5 ! 52 9 5

22 P M 7
99 H #Y ! # # 2 2!
./01. . .+ . " 2 2!

data 23.02.2020

Pág. 590(611)

[87] Pappu's harmonic theorem in the Einstein relativistic.

---

3 cont.) Pappu's harmonic theorem in the Einstein relativistic

www.cs.ubbcluj.ro/~studia-m/2011-1/10-piscoran-final.pdf

mathworld.wolfram.com/PappussCentroidTheorem.html

theoremoftheday.org/GeometryandTrigonometry/Guldin/TotDPappusGuldin.pdf

---

117) 20200226

a) Sullivan Conjecture

https://ocw.mit.edu/faculty/haynes-miller/

b) $2\pi = \tau$ (Tau)

the radio of the great pyramids perimeter to its height approximates 2π or what is know as TAU.

data 26.02.2020

c) Data_Novovivo_40_2020/Books_40_2020/
Atoms, Molecules and optical Physics_1/

c1) Atoms, Molecules and optical
Physics 1.pdf

Pág. 90(122) –

2.2 ~~Schrödinge~~

Schrödinger Equation

2.2.3 Time Dependent
Schrödinger Equation

... which is a 2nd order
partial differential equation (PDE)
in space and time derived
from the
MAXWELL equations.

c2) Atoms, Molecules and optical
Physics 2.pdf

Pág. 27(58)

Snellius diffraction

Pág. 28(59)

ABCD matrices.

Pág. 31(62) – Fig 1.16 ⇒

26.02.2020

De-focussing of a Gaussian beam.

Look:

Fig 1.15 - Pág. 30 (61)

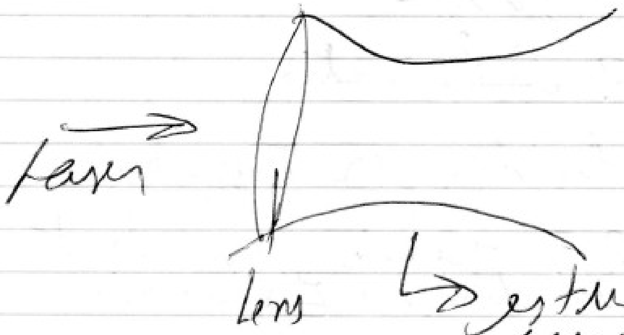

↳ estude esta curvatura com a curva que distorce a gravidade.

Obs:

Fig 1.16

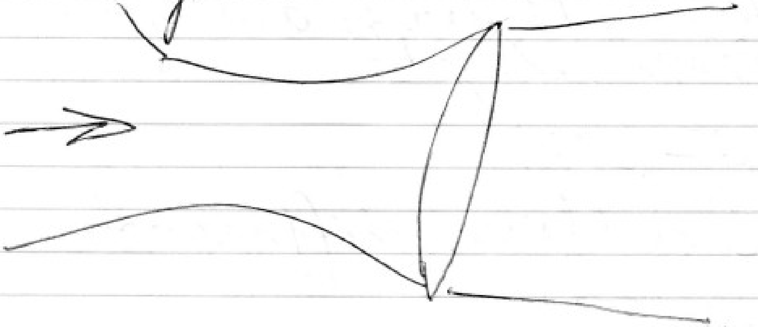

c3) Books_40_2020/ Ordinary Differential Equations.../

saret.st-10167.pdf

26.02.2020

Pág. 61(71) Theorem 1.7

Jordan curve J

U will thus contain (from the Poincaré-Bendixson theorem) the limite cycle $\Gamma(\varepsilon)$.

Look: Fig II.1.7.

c4) Book_40_2020/ Handbook_of_Homotopy_Theory/

Handbook of Homotopy theory.pdf

Pág. 315(324)

8.7.4 Persistence images

A standard choice is that of a spherical symmetric Gaussian with mean μ and a fixed variance σ.

Pág. 316(325)

Pág. 320(329)

Example 8.8.6-

For examples, give a RIEMANNIAN manifold, one can

26.02.2020

compute Gaussian curvature

↳ comparar com a curvatura do espaço (gravidade).

at each point and filter by that quantity.

Pág. 326(335) - Gaussian random

Pág. 327(336) - Gaussian field.

Pág. 328(337) →

[9] Bobrowski and Borman

Euler integration of Gaussian random fields and persistent homology.

Pág. 970(979) →

[69] Levine

Motivic Gauss-Bonnet formulas

e4) Books_40_2020/
Geometry_the Line and the Circle

26.02.2020

compute Gaussian curvature

↳ comparar com a curvatura do espaço (gravidade).

at each point and filter by that quantity.

Pág. 326 (335) - Gaussian random

Pág. 327 (336) - Gaussian field.

Pág. 328 (337) →

[9] Bobrowski and Borman

Euler integration of Gaussian random fields and persistent homology.

Pág. 970 (979) →

[69] Levine

Motivic Gauss-Bonnet formulas

a4) Books_40_2020/Geometry_the Line and the Circle

26.02.2020

an algebraic curve whose x
coefficients and variables
are complex numbers.

Newton → "imaginary
points"

In the 1820s,
Jean Poncelet and
Michel Chasles
argued for using imaginary
points systematically in
Synthetic projective geometry.

Plücker (1830) →
homogeneous coordinates,
he introduced made it possible
to consider points with
complex coordinates.

118) 20200228

Número de Perrin

pt.wikipedia.org/wiki/Número_de_perrin

./01. . .+ . (/ A ! +2
5 2!) $

also characterized by an amplitude $\psi(\boldsymbol{r})$, determining the probability for finding a particle at position $\boldsymbol{r}$ in a volume element $\mathrm{d}^3\boldsymbol{r}$:

$$\mathrm{d}w(\boldsymbol{r}) = w(\boldsymbol{r})\mathrm{d}^3\boldsymbol{r} = \left|\psi(\boldsymbol{r})\right|^2\mathrm{d}^3\boldsymbol{r}. \tag{2.5}$$

And as in optics, diffraction and interference occurs, e.g. at a double slit for which (2.4) describes the experimentally observable signal of particles.

However, in contrast to electromagnetic radiation, where ψ was identified with the electric (or magnetic) field, for matter waves $\psi(\boldsymbol{r})$ cannot be attributed to any directly measurable physical quantity: the observable physics is described by the probability $w(\boldsymbol{r})$. So we call $\psi(\boldsymbol{r})$ simply the *probability amplitude* for finding a particle at position $\boldsymbol{r}$.

Otherwise, analogue considerations are valid for photons and particles of matter. If e.g. we try to localize them on their pathways, we loose the interference patterns! One *important general rule* holds: *interference phenomena are observed if different but indistinguishable pathways exist* on which the particles may proceed toward the detector. In contrast, no interference is observed if the two pathways may in principle be distinguished – even when no distinction is made in the actual experiment.

Section summary

- In classical mechanics well defined trajectories in position $\boldsymbol{r}(t)$ and momentum $\boldsymbol{p}(t)$ space describe the motion of particles. In contrast, quantum mechanics describes probability amplitudes $\psi(\boldsymbol{r})$ and probabilities $\mathrm{d}w(\boldsymbol{r}) = |\psi(\boldsymbol{r})|^2\mathrm{d}^3\boldsymbol{r}$ for finding a particle at a certain position $\boldsymbol{r}$ in a volume element $\mathrm{d}^3\boldsymbol{r}$.
- Diffraction and interference of matter waves can be described by concepts in analogy to those used in wave optics. The big difference is that for photons the “probability amplitude” is also a directly measurable quantity (electric or magnetic field) while for matter waves $\psi(\boldsymbol{r})$ cannot be measured directly – only probabilities $|\psi(\boldsymbol{r})|^2$ refer to the real world.

2.2 SCHRÖDINGER Equation

Wave equations for matter waves were derived independently by Erwin SCHRÖDINGER and PAUL DIRAC who in 1933 jointly received the NOBEL prize in physics for their ground breaking work. The nonrelativistic SCHRÖDINGER equation is most commonly used in AMO physics and will briefly be introduced here. As we shall see in Chap. 6, the relativistically correct DIRAC equation, valid for fermions, is significantly more complex and describes inherently also the electron spin. Its key results may, however, also be introduced as perturbations into the SCHRÖDINGER equation which then leads to sufficiently accurate results in the non relativistic energy regime.

2.2.1 Stationary SCHRÖDINGER Equation

In contrast to photons, particles of matter may be exposed to external forces which change their momentum and hence their wavelength. One may start from the wave equation (2.2) and try to guess, how to modify it for matter waves. Considering a particle of mass m with a total energy W moving in a conservative potential $V(\boldsymbol{r})$, we simply use the (nonrelativistic) energy conservation law of classical mechanics $W = W_{\text{kin}} + V$ to determine its momentum from the kinetic energy W_{kin}:

$$p^2 = 2mW_{\text{kin}} = 2m\big(W - V(\boldsymbol{r})\big).$$

With W being a constant of motion we 'derive' from this and (1.118) the absolute value of the wave vector $k = p/\hbar$ and insert it into (2.2):

$$\Delta\psi(\boldsymbol{r}) + \frac{p^2}{\hbar^2}\psi(\boldsymbol{r}) = \Delta\psi(\boldsymbol{r}) + \frac{2m(W - V(\boldsymbol{r}))}{\hbar^2}\psi(\boldsymbol{r}) = 0.$$

Rewritten, this is already the *stationary* SCHRÖDINGER *equation*

$$-\frac{\hbar^2}{2m}\Delta\psi(\boldsymbol{r}) + V(\boldsymbol{r})\psi(\boldsymbol{r}) = W\psi(\boldsymbol{r}), \tag{2.6}$$

or somewhat more compact

$$\widehat{H}\psi(\boldsymbol{r}) = W\psi(\boldsymbol{r}) \tag{2.7}$$

with the *eigenfunction* $\psi(\boldsymbol{r})$ and the *eigenenergy* W. The HAMILTON *operator* $\widehat{H}$ represents the total energy (briefly just *Hamiltonian*):

$$\widehat{H} = -\frac{\hbar^2}{2m}\Delta + V(\boldsymbol{r}). \tag{2.8}$$

In the often encountered one-dimensional case the SCHRÖDINGER equation (2.6) is further simplified:

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\psi(x)}{\mathrm{d}x^2} + V(x)\psi(x) = W\psi(x). \tag{2.9}$$

2.2.2 HAMILTON and Momentum Operators

We may write the HAMILTON operator (2.8) even more suggestive:

$$\widehat{H} = -\frac{\hbar^2}{2m}\nabla^2 + V(\boldsymbol{r}) = \frac{\widehat{\boldsymbol{p}}^2}{2m} + V(\boldsymbol{r}). \tag{2.10}$$

Here we have introduced the *momentum operator* (a *vector operator*)

$$\widehat{\boldsymbol{p}} = -\mathrm{i}\hbar\nabla = -\mathrm{i}\hbar \begin{pmatrix} \partial/\partial x \\ \partial/\partial y \\ \partial/\partial z \end{pmatrix} \tag{2.11}$$

so that

$$\widehat{\boldsymbol{p}}^2 = \widehat{\boldsymbol{p}} \cdot \widehat{\boldsymbol{p}} = -\hbar^2\nabla^2 = -\hbar^2\Delta = -\hbar^2\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right). \tag{2.12}$$

Thus, (2.10) is the operator form of the *classical energy conservation law*

$$W = W_{\mathrm{kin}} + V = \frac{p^2}{2m} + V(\boldsymbol{r}).$$

2.2.3 Time Dependent SCHRÖDINGER Equation

So far we have discussed probability amplitudes only as a function of $\boldsymbol{r}$, the particle's position in space. For many problems in AMO physics such a description of quantum systems by stationary states is fully sufficient. However, often time dependence is equally important. In respect of photons we know, that electromagnetic waves follow a time dependent wave equation, which is a 2nd order partial differential equation (PDE) in space *and* time derived from the MAXWELL equations. In contrast, for matter waves the *time dependent* SCHRÖDINGER *equation* holds:

$$\widehat{H}\Psi(\boldsymbol{r},t) = \mathrm{i}\hbar\frac{\partial\Psi(\boldsymbol{r},t)}{\partial t} \tag{2.13}$$

or more explicitly: $$-\frac{\hbar^2}{2m}\Delta\Psi(\boldsymbol{r},t) + V(\boldsymbol{r})\Psi(\boldsymbol{r},t) = \mathrm{i}\hbar\frac{\partial\Psi(\boldsymbol{r},t)}{\partial t}.$$

It cannot be derived, and we just communicate it here as it was 'found' by Erwin SCHRÖDINGER in the beginning of 1926 – by the way: during a winter ski holiday in the Swiss Alps. *We point out some key aspects:*

- The SCHRÖDINGER equation is a linear PDE of 2nd order in space and 1st order in time! As a consequence, time dependence of stationary states is truly complex.
- The linear superposition principle may be applied to the solutions.
- The statistical interpretation of quantum mechanics understands solutions $\Psi(\boldsymbol{r},t)$ of this PDE as probability amplitudes for finding a particle at a position $\boldsymbol{r}$ in space at time t according to (2.1).
- Although the SCHRÖDINGER equation cannot be 'derived', it is able to describe a wealth of atomic phenomena and to predict observables quantitatively (under nonrelativistic conditions) with excellent accuracy: we emphasize that is this very fact which defines the "validity" of a physical theory.
- Equivalently, formal (algebraic) quantum mechanics can be deduced from a consistent set of axioms, which are, however, also heuristically assumed.

- Consistent alternatives of wave equations for matter waves are the DIRAC equation for fermions, a multi component spinor equation and the KLEIN-GORDON equation, a single component PDE of 2nd order in time, which turns out to be valid for bosons – both are relativistically correct.

If the *Hamiltonian itself does not depend on time*, $\widehat{H}(\boldsymbol{r},t) = \widehat{H}(\boldsymbol{r})$, the time dependence of the wave function may be *factored by a product ansatz*:

$$\Psi(\boldsymbol{r},t) = \psi(\boldsymbol{r})\varphi(t) \tag{2.14}$$

$$\widehat{H}\Psi(\boldsymbol{r},t) = \mathrm{i}\hbar\frac{\partial\Psi(\boldsymbol{r},t)}{\partial t} \quad \Rightarrow \quad \widehat{H}\psi(\boldsymbol{r})\varphi(t) = \mathrm{i}\hbar\frac{\partial\psi(\boldsymbol{r})\varphi(t)}{\partial t}$$

$$\frac{\widehat{H}\psi(\boldsymbol{r})}{\psi(\boldsymbol{r})} = \frac{\mathrm{i}\hbar}{\varphi(t)}\frac{\partial\varphi(t)}{\partial t} \equiv W.$$

The last identity (with the constant W which has to be determined) must hold so that the former equality can be valid for all values of $\boldsymbol{r}$ and t. One thus has to solve $\mathrm{i}\hbar\mathrm{d}\varphi(t)/\mathrm{d}t = W\varphi(t)$ and $\widehat{H}\psi(\boldsymbol{r}) = W\psi(\boldsymbol{r})$. In this case, the time dependent equation leads to the trivial solution

$$\varphi(t) \propto \exp\left(-\mathrm{i}\frac{W}{\hbar}t\right). \tag{2.15}$$

The position dependent part is nothing but *the stationary* SCHRÖDINGER *equation* (2.6) and the parameter W introduced *is the total energy of the system*. The overall wave function is given by

$$\Psi(\boldsymbol{r},t) = \psi(\boldsymbol{r})\exp\left(-\mathrm{i}\frac{W}{\hbar}t\right). \tag{2.16}$$

Note: The time dependence is truly complex and the imaginary unit i in the prefactor is needed for the solution! In the present case, however, with $\widehat{H} \neq \widehat{H}(t)$ the time dependence is trivial in the sense that only

$$w(\boldsymbol{r},t) = \left|\Psi(\boldsymbol{r},t)\right|^2 = \left|\psi(\boldsymbol{r})\right|^2 \tag{2.17}$$

can be measured, which contains only information about the stationary state.

In order to find these stationary atomic states for a given potential $V(\boldsymbol{r})$ – which are the equivalent to BOHR's stationary orbits – one has to solve the stationary SCHRÖDINGER equation (2.6) under suitably chosen boundary conditions. For bound states this typically leads to a whole *series of discrete total energies* $W_\gamma < 0$ with a set of *quantum numbers* γ characterizing the states. The corresponding *series of wave functions* – for individual electrons called *orbitals* – describe the probability of finding the particles at a given position in space.

./01. . .+ . (/ A ! +2 5
2!) (- !

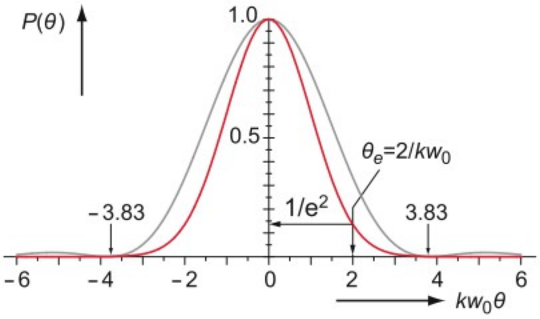

Fig. 1.13 Comparison of the angular dependence of a Gaussian beam (*red*), with a beam waist radius w_0, and the far field of a plane wave (*grey*), diffraction limited by an aperture of radius w_0. The two profiles are normalized to each other at $\theta = 0$

$\hbar k_\perp = \hbar k \theta_e$ one finds $\hbar k_\perp w_0 \geq 2\hbar$. Generally, one defines a so called *beam parameter product* BPP $\equiv w_0\theta_e$ as a measure for the quality of a light beam (specifically for lasers). For a *Gaussian beam* we have

$$\mathrm{BPP} = w_0\theta_e = 2/k = \lambda/\pi, \tag{1.69}$$

while for the *aperture limited plane wave* according to (1.68) the divergence angle

$$w_0\theta_e \geq 3.83/k = 1.22\frac{\lambda}{2} \quad \text{and} \quad s_{\mathrm{Airy}} = \theta_e z = 1.22\frac{z\lambda}{2w_0} \tag{1.70}$$

holds. The characteristic quantity here is the radius s_{Airy} of the central diffraction pattern (the so called AIRY *disc*) on the detector screen.[10] We shall come back to these expressions again in Sect. 2.1.7 in the context of *lateral coherence of natural light* and laser radiation.

1.2.3 Ray Transfer Matrices

Before we can efficiently treat the manipulation of laser beams (focussing, defocussing, deflecting etc.) we have to introduce an important tool of geometrical optics, applicable also to Gaussian beams: the so called *ray transfer matrices* or *ABCD matrices*. They are used for *ray tracing* of paraxial light rays (with $\sin\theta \cong \theta$), i.e. they allow to describe in a simple manner the propagation of light rays through an optical arrangement. These ABCD matrices are efficient tools e.g. for designing laser systems and for computing the properties of laser resonators – typically describing the propagation of well collimated Gaussian beams through a number of optical elements such as mirrors, lenses, and prisms.

---

[10] Note that this formula is the basis of the well known RAYLEIGH criterium for the diffraction limited resolving power of optical instruments. It states that s_{Airy} is the smallest distance between two objects that can be resolved by an instrument with an effective limiting aperture of diameter $D = 2w_0$ and focal length $f = z$.

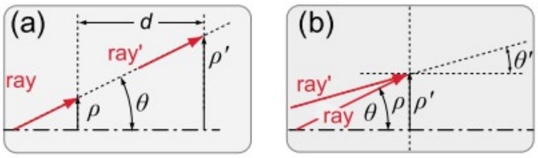

Fig. 1.14 ABCD matrices operating on a light ray. **(a)** Ray translation through a distance d. **(b)** Refraction at an interface between two materials

A paraxial light ray that propagates at a (small) angle θ in a distance ρ from the optical axis, is described by a *ray vector*:

$\begin{pmatrix}\rho\\ \theta\end{pmatrix}$ corresponding to the scheme 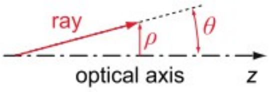

In Fig. 1.14 the change of a *ray* into *ray*′ is illustrated for two important cases: *translation* and *refraction*. Such a change upon transition of the ray through the optical setup corresponds to a linear transformation of the ray vector,[11] described by an ABCD matrix:

$$\begin{pmatrix}\rho'\\ \theta'\end{pmatrix}=\begin{pmatrix}A & B\\ C & D\end{pmatrix}\begin{pmatrix}\rho\\ \theta\end{pmatrix}. \tag{1.71}$$

Specifically, for a Gaussian beam the complex beam parameter (1.41) changes under such transformation from q to q' (here without proof; see e.g. KOGELNIK and LI 1966):

$$q'=\frac{Aq+B}{Cq+D} \quad \text{or} \quad \frac{1}{q'}=\frac{C+D/q}{A+B/q}. \tag{1.72}$$

The most important simple cases are collected in Table 1.2. One verifies

$$\text{that case 1} \quad \begin{pmatrix}\rho'\\ \theta'\end{pmatrix}=\begin{pmatrix}1 & d\\ 0 & 1\end{pmatrix}\begin{pmatrix}\rho\\ \theta\end{pmatrix}=\begin{pmatrix}\rho+\theta d\\ \theta\end{pmatrix}$$

describes a translation,

$$\text{while case 2} \quad \begin{pmatrix}\rho'\\ \theta'\end{pmatrix}=\begin{pmatrix}1 & 0\\ 0 & \frac{n}{n'}\end{pmatrix}\begin{pmatrix}\rho\\ \theta\end{pmatrix}=\begin{pmatrix}\rho\\ \frac{n\theta}{n'}\end{pmatrix}$$

represents the SNELLIUS diffraction law for small angles, $n'\theta'=n\theta$ (paraxial rays). Note that in the latter case $\rho'=\rho$ does not change.

The propagation of a ray through several optical devices is described by the product of the respective ABCD matrices. For example, the ray vector after translation *and* refraction by a thin lens is derived by multiplying matrix 1 with matrix 3 from

---

[11] In a mathematical sense the ray vector is simply the position vector expressed by ρ, the distance from the optical axis (z-axis), and $\theta=\rho/z$, the polar angle. The pictographs used in the following are not always correct in that sense, but quite instructive.

Table 1.2 The most important, simple ABCD matrices

Case	Description	ABCD	Scheme
1	translation through distance d in medium with index of refraction n	$\begin{pmatrix}1 & \frac{d}{n}\\ 0 & 1\end{pmatrix}$	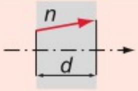
2	refraction by a plane surface with index of refraction left n, right n'	$\begin{pmatrix}1 & 0\\ 0 & \frac{n}{n'}\end{pmatrix}$	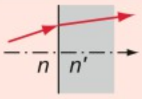
3	thin lens of focal length f	$\begin{pmatrix}1 & 0\\ -\frac{1}{f} & 1\end{pmatrix}$	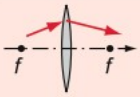
4	reflection from a concave mirror	$\begin{pmatrix}1 & 0\\ -\frac{2}{R} & 1\end{pmatrix}$	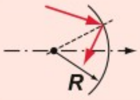
5	refraction at a spherical surface	$\begin{pmatrix}1 & 0\\ -\frac{n'-n}{n'R'} & \frac{n'}{n}\end{pmatrix}$	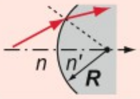

Table 1.2:

$$\begin{pmatrix}1 & 0\\ -\frac{1}{f} & 1\end{pmatrix}\begin{pmatrix}1 & d\\ 0 & 1\end{pmatrix}\begin{pmatrix}\rho\\ \theta\end{pmatrix}=\begin{pmatrix}1 & d\\ -\frac{1}{f} & 1-\frac{d}{f}\end{pmatrix}\begin{pmatrix}\rho\\ \theta\end{pmatrix}=\begin{pmatrix}\rho+\theta d\\ \theta-\frac{\rho+\theta d}{f}\end{pmatrix}. \qquad (1.73)$$

One verifies e.g. for any ray coming out of the focal point ($d=f$ and $\rho=0$), that after passing the lens the exit angle is always $\theta'=0$, i.e. all such rays leave the lens parallel – as well known from geometrical optics. Conversely, a ray entering the lens parallel to the optical axis ($\theta=0$) always leaves the lens at $\rho'=\rho$ with $\theta'=-\rho/f$, independently of d, i.e. the ray is refracted into the focal point.

We may also describe ray propagation behind the lens to a position d' if we know the ray vector in front of the lens:

$$\begin{pmatrix}1 & d'\\ 0 & 1\end{pmatrix}\begin{pmatrix}1 & 0\\ -\frac{1}{f} & 1\end{pmatrix}\begin{pmatrix}\rho\\ \theta\end{pmatrix}=\begin{pmatrix}1-\frac{d'}{f} & d'\\ -\frac{1}{f} & 1\end{pmatrix}\begin{pmatrix}\rho\\ \theta\end{pmatrix}=\begin{pmatrix}\rho-\frac{\rho d'}{f}+d'\theta\\ -\frac{\rho}{f}+\theta\end{pmatrix}. \qquad (1.74)$$

One verifies e.g. that rays which are parallel to the axis before entering the lens ($\theta=0$) cross the axis behind the lens at $d'=f$.

Finally, the product of matrix 1 with matrix 3 and again matrix 1 allows one to describe imaging by a lens. Further multiplication with the matrices for additional optical pathways and other elements allows one to easily describe even complex optical setups very efficiently. We shall now apply this technique to the manipulation of a Gaussian beam with lenses.

1.2.4 Focussing a Gaussian Beam

Focussing and de-focussing, expanding and concentrating Gaussian beams is very important in experimental practice. In principle, very wide beams with large radii may be generated as well as tightly focused ones with extremely high intensity at the focal point. In any case, however, the *product of beam radius w and divergence angle θ_e remains constant*, and is according to (1.69) for a Gaussian beam $\theta_e w = \lambda/\pi$.

To familiarize ourselves with the tools we let a Gaussian beam propagate freely, say, emerging from a laser system. Let us assume that we start with an ideally parallel beam ($R = \infty$) of radius w_L. With the RAYLEIGH length $z_R = \pi w_L^2/\lambda$ the complex beam parameter according to (1.41) is

$$\frac{1}{q} = -\mathrm{i}\frac{\lambda}{\pi w_L^2} = -\frac{\mathrm{i}}{z_R}. \tag{1.75}$$

It changes with distance travelled according to (1.72). With the propagation matrix 1 in Table 1.2 we find at a distance $z' = d$

$$\frac{1}{q'} = \frac{C + D/q}{A + B/q} = \frac{1/q}{1 + z'/q},$$

which can be rewritten in the form (1.41) as

$$\frac{1}{q'} = \frac{1}{z'(1 + z_R^2/z'^2)} - \frac{\mathrm{i}}{z_R(1 + z'^2/z_R^2)} = \frac{1}{R'} - \mathrm{i}\frac{\lambda}{\pi w'^2}.$$

Thus we find $R' = z'(1 + z_R^2/z'^2)$ and $w' = w_L\sqrt{1 + z'^2/z_R^2}$, expressions which are completely equivalent to (1.42) and (1.43). We see that the matrix method reproduces indeed the behaviour of a Gaussian beam as it propagates along the optical axis – including the correct divergence angle $\theta_e = \lambda/\pi w_L$ corresponding to (1.52).

We now want to *focus this laser beam by a collecting lens* of focal length f. Let the lens be positioned at $z = 0$, and we state now more precisely that the assumed initial parallelism implies $R_L \gg f$. In geometrical optics such a "plane wave" is assumed to be focused in the focal *point* of the lens, whereby the initial plane wave front ($R_L = \infty$) is converted into a spherical wave with a radius of curvature $R' = (z - f)$. In reality, however, due to diffraction the focal *point* becomes a circular disc of radius w_0, as sketched in Fig. 1.15.

The complex beam parameter in front of the lens is given by (1.75) as in the previous example. According to (1.72) we use this time the product matrix (1.74) to obtain the new complex beam parameter at a distance z' behind the lens:

$$\frac{1}{q'} = \frac{C + D/q}{A + B/q} = \frac{-1/f - \mathrm{i}\lambda/(\pi w_L^2)}{(1 - \frac{z'}{f}) - z'\mathrm{i}\lambda/(\pi w_L^2)}. \tag{1.76}$$

+9 4 9

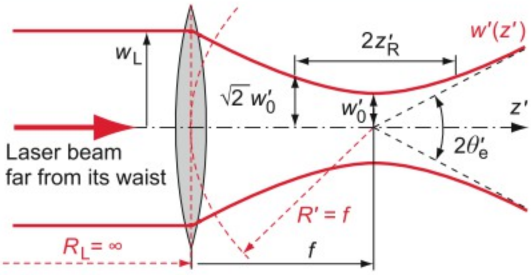

Fig. 1.15 Focussing a parallel Gaussian beam

Without going into further details we concentrate onto the focal point of the lens, $z' = f$, in order to determine the waist $w'_0 = w'(f)$.[12] The result is simply

$$\frac{1}{q'} = \frac{1}{f} - \mathrm{i}\frac{\pi w_L^2}{\lambda f^2} = \frac{1}{R'} - \mathrm{i}\frac{\lambda}{\pi w_0'^2}$$

with the latter equality again according to (1.41). Comparing real and imaginary parts we obtain the *waist radius w'_0 of the focused beam*

$$w'_0 = \frac{f\lambda}{\pi w_L}, \tag{1.77}$$

while the radius of curvature of the wave front becomes $R'(f) = f$. For practical purposes we also note the relation with the $1/\mathrm{e}$ intensity GAUSS radius $a = w/\sqrt{2}$

$$a'_0 = \frac{f\lambda}{2\pi a_L}, \tag{1.78}$$

and recall that the maximum intensity (1.60) in the beam centre which may be obtained by this kind of focussing is $I_0 = P_{\mathrm{tot}}/\pi a_0'^2$.

Inserting (1.77) into (1.45) gives for the focused beam the RAYLEIGH length z'_R and the *confocal parameter* (the distance around the focus for which the intensity on axis is above 50 % of the maximum):

$$b' = 2z'_R = \frac{2\pi w_0'^2}{\lambda} = \frac{2\lambda f^2}{\pi w_L^2}.$$

Table 1.3 presents numerical examples for three different focal lengths f and two different initial beam diameters w_L, all at the 800 nm wavelength of the Titanium-sapphire laser – the workhorse of ultrafast spectroscopy.

---

[12] A closer look into the algebra shows, however, that the smallest cross section is found at $z = fz_0^2/(z_R^2 + f^2)$, i.e. slightly in front of the focal point of the lens. There the radius of curvature even approaches $R \to \infty$. Since typically we start with a relatively large beam radius w_L the RAYLEIGH length is very large, $z_R = \pi w_L^2/\lambda \gg f$, so that the small difference does not play a practical role.

Table 1.3 Beam radius w_0' (at $1/\mathrm{e}^2$ intensity) and confocal parameter $b' = z_R'$ in the focus of a lens with focal length f. An initially parallel Gaussian beam is assumed, at $\lambda = 800\,\text{nm}$ with radius w_L

f/mm	w_L/mm	w_L/f	w_0'/μm	b'
100	0.4	0.004	64	31 mm
100	2	0.02	13	1.3 mm
50	0.4	0.008	31	8 mm
50	2	0.04	6.4	318 μm
10	0.4	0.04	6.4	318 μm
10	2	0.2	1.2	12 μm

Conversely, one may *parallelize a diverging laser beam* by putting a lens of focal lengths f at a distance f behind its waist. As illustrated in Fig. 1.16 the lens converts the curved wave front into an essentially plane wave. Starting with a beam waist w_0 at $z = 0$ and a radius of curvature $R_L = -f$ the complex beam parameter at the waist is

$$\frac{1}{q} = -\frac{1}{f} - \mathrm{i}\frac{\lambda}{\pi w_0^2}.$$

We use again (1.72), this time with the product matrix (1.73). After a brief calculation we obtain the new complex beam parameter directly behind the lens:

$$\frac{1}{q'} = \frac{C + D/q}{A + B/q} = -\mathrm{i}\frac{\pi}{f^2\lambda} w_0^2.$$

Beam radius and radius of curvature immediately behind the lens are thus

$$w_0' = \frac{f\lambda}{\pi w_0} \quad \text{and} \quad R' = \infty, \quad \text{respectively.}$$

As expected, this is just the inverse of the focussing process according to (1.77). We have to be aware that the 'new' laser beam thus generated has now its waist just behind the lens. From there it will propagate in a slightly divergent manner. According to (1.55) the new divergence angle of the 'parallelized' Gaussian beam

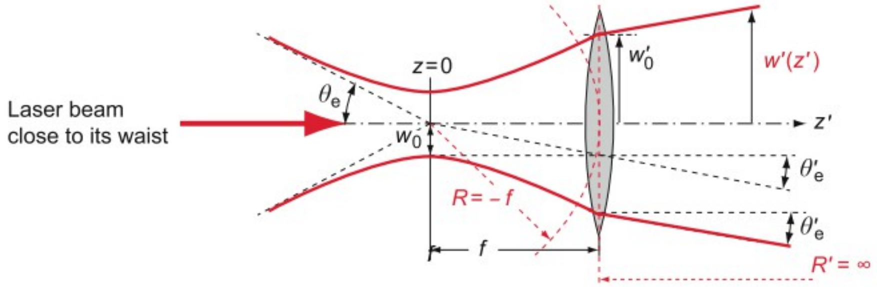

Fig. 1.16 De-focussing of a Gaussian beam

Table 1.4 Beam radii of Gaussian beams as a function from the beam waist for different initial waist radii w_0 according to (1.43). The RAYLEIGH length z_R and the far field divergence angle θ_e are related to w_0 by (1.45) and (1.55), respectively. The numbers are given for the Titanium-sapphire laser wavelength $\lambda = 800\,\text{nm}$

Nr.	Beam parameter			w/mm Beam radius at $1/\text{e}^2$ Intensity at z (distance to waist)				
	w_0/mm	z_R/mm	θ_e/mrad	1 m	2 m	20 m	100 m	1000 m
1	10	400×10^3	2.5×10^{-2}	10	10	10	10.3	27
2	3	35×10^3	3×10^{-2}	3	3	3.4	9	85
3	1	3.9×10^3	0.25	1	1.1	5.2	25	255
4	0.3	353	0.84	0.9	1.7	16	85	850
5	0.1	39	2.5	2.5	5	51	254	2546
6	0.03	3.5	8.5	8.5	16	170	850	8500
7	0.01	0.4	25	25	50	250	2550	25 465

is

$$\theta'_e = \frac{\lambda}{\pi w'_0} = \frac{\lambda \pi w_0}{\pi f \lambda} = \frac{w_0}{f}. \tag{1.79}$$

This divergence angle may even be interpreted geometrically as indicated in Fig. 1.16: it corresponds to the divergence one would expect from classical ray optics imaging an extended light source with a radius w_0 to infinity.

To give a feeling for the divergence of Gaussian beams, in Table 1.4 numerical examples are collected for different initial beam radii w_0 (equivalently RAYLEIGH lengths z_R or divergence angles θ_e). The beam radii w at different distances z from the waist illustrate clearly how the beams diverge – one may often observe this by laser illuminations in the nightly skies of big cities (thanks to RAYLEIGH scattering in dry weather, or even better in humid air due to MIE scattering as discussed in Sect. 8.4.5, Vol. 1).

These numbers are astonishing indeed, at least at first glance. Compare e.g. beam 1, 3, and 5 in Table 1.4. Beam 1 may initially appear as a rather broad 'brush' of light. Note, however, that it widens only by a factor of 3 over the distance of 1km, and it would thus be well suited e.g. in telecommunication or metrology. In contrast, beams 3 or 5 look on the laser table nicely thin and 'laser-like'. They expand, however, very rapidly and get a rather large radius in some distance. Beam 5 will hardly be useable in a neighbouring lab at 20 m distance, while beam 6, finally, is a highly focussed beam generating very high intensity in its focal point, but with a RAYLEIGH length of 0.4 mm, its on axis intensity decays rapidly to less than 50 % outside the confocal length of 0.8 mm around the focus.

For transporting beams over wide distances one first has to expand them. To this end one uses telescope systems as sketched in Fig. 1.17. In the destination where the beam is to be used, one may place a second telescope system – more or less identical to the first one – in order to re-compress the beam. Even if one wants to

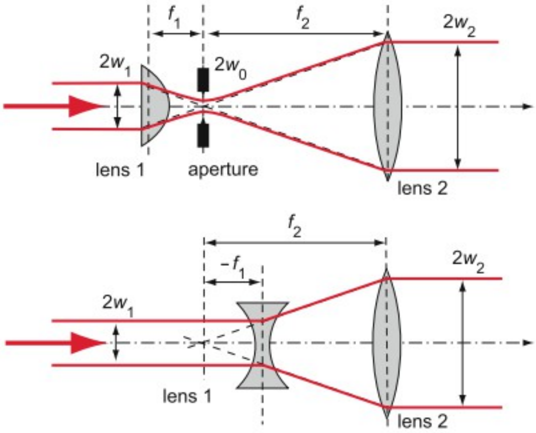

Fig. 1.17 Telescope systems for expanding a Gaussian beam. The *top setup* according to Kepler (two collecting lenses) contains an aperture, positioned in the focus, for beam improvement (spatial filter). The *lower setup* according to Galilei (dispersing lens and collecting lens) can be used even for very high intensities

generate a particularly tight focus, it is recommendable to first expand the beam: according to (1.77) the minimal radius w_0 achievable in the focus of a lens is inversely proportional to the radius w_L of the laser beam in front of the focusing lens.

As illustrated in Fig. 1.17 one may construct telescope systems essentially in two ways: either according to Kepler with two collecting lenses, arranged con-focally or according to Galilei with one dispersing and one collecting lens, the focus of the latter matching the backward focus of the dispersing lens. From the geometry shown in Fig. 1.17 one reads the beam expansion from radius w_1 to w_2 in both cases to be

$$w_2 = w_1 |f_2/f_1|. \tag{1.80}$$

The divergence angle of the expanded beam is again estimated from (1.55):

$$\theta_2 = \frac{\lambda}{\pi w_2}. \tag{1.81}$$

Strictly, this only holds if the radius of the phase surfaces at lens 2 exactly matches f_2 – which may always be achieved by fine adjustment of the lens positions.

Both types of telescope systems have their specific merits. Between two collecting lenses a real focal-point exists, into which one may e.g. place a small, circular aperture (diameter typically 3 to 5 × w_0) as indicated in Fig. 1.17. With such a *spatial filter* one often may improve the lateral profile of not completely perfect Gaussian beams considerably. The (lateral and longitudinal) adjustment of such an aperture requires some experimental skills.

On the other hand, for very high laser intensities – as often encountered when working with femtosecond laser pulses – one must avoid a real focus where ionization and electric breakdown in air may occur, and destroy the laser pulse. In such cases only the Galilei type of telescopes is applicable.

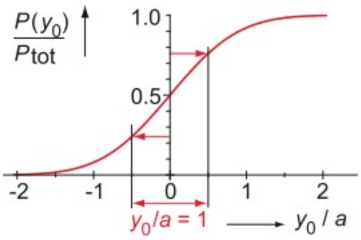

Fig. 1.18 Signal from a Gaussian beam measured with the knife edge method. One determines a from the y_0 position for a signal ratio of 0.24 and 0.76 – as indicated

1.2.5 Measuring Beam Profiles with a Razor Blade

If one already knows that the radial beam profile is Gaussian, or is least described well enough by (1.58), one may determine the radii a or w according to (1.59) by the rather simple, practical "knife-edge" method. One mounts a razor blade onto a precision optical table, which can be moved perpendicularly into the beam, say in y-direction. One registers the power $P(y_0)$ which reaches a detector positioned behind it as a function of y. If the beam is covered from $y = y_0$ to $y = \infty$ this amounts to (with $\rho = \sqrt{x^2 + y^2}$)

$$P(y_0) = \frac{P_{\mathrm{tot}}}{\pi a^2} \int_{-\infty}^{y_0} \mathrm{d}y \left(\int_{-\infty}^{+\infty} \mathrm{d}x \exp\left(-\frac{x^2 + y^2}{a^2} \right) \right). \tag{1.82}$$

This expression may be integrated in closed form and gives

$$\frac{P(y_0)}{P_{\mathrm{tot}}} = \frac{1}{a\sqrt{\pi}} \int_{-\infty}^{y_0} \exp\left(-\frac{y^2}{a^2} \right) \mathrm{d}y = \frac{1}{2} \left(\mathrm{erf}\left(\frac{y_0}{a} \right) + 1 \right) \tag{1.83}$$

with the error function $\mathrm{erf}(y_0/a)$. $P(y_0/a)/P_{\mathrm{tot}}$ is shown in Fig. 1.18, being 0.2398 and 0.7602 for $y_0/a = \mp 0.5$, respectively. At these two positions one reads the GAUSS radius $a = w/\sqrt{2}$ as marked in Fig. 1.18.

1.2.6 The M^2 Factor

In practice, of course, laser beams are never completely perfect. More or less pronounced deviations from the ideal Gaussian beam profile are the rule. This is very important when describing the quality of a laser system. Most concisely this is expressed by the beam parameter product (BPP) according to (1.69): the larger the BPP, the poorer the beam quality. The international norm ISO 11146 defines as a quantitative characteristic of a laser beam with a divergence angle θ_e and a beam radius w_0 the so called M^2 factor by

$$\theta_e = M^2 \frac{\lambda}{\pi w_0}. \tag{1.84}$$

2.2.1 Stationary SCHRÖDINGER Equation

In contrast to photons, particles of matter may be exposed to external forces which change their momentum and hence their wavelength. One may start from the wave equation (2.2) and try to guess, how to modify it for matter waves. Considering a particle of mass m with a total energy W moving in a conservative potential $V(\boldsymbol{r})$, we simply use the (nonrelativistic) energy conservation law of classical mechanics $W = W_{\text{kin}} + V$ to determine its momentum from the kinetic energy W_{kin}:

$$p^2 = 2m W_{\text{kin}} = 2m\big(W - V(\boldsymbol{r})\big).$$

With W being a constant of motion we 'derive' from this and (1.118) the absolute value of the wave vector $k = p/\hbar$ and insert it into (2.2):

$$\Delta\psi(\boldsymbol{r}) + \frac{p^2}{\hbar^2}\psi(\boldsymbol{r}) = \Delta\psi(\boldsymbol{r}) + \frac{2m(W - V(\boldsymbol{r}))}{\hbar^2}\psi(\boldsymbol{r}) = 0.$$

Rewritten, this is already the *stationary* SCHRÖDINGER *equation*

$$-\frac{\hbar^2}{2m}\Delta\psi(\boldsymbol{r}) + V(\boldsymbol{r})\psi(\boldsymbol{r}) = W\psi(\boldsymbol{r}), \tag{2.6}$$

or somewhat more compact

$$\widehat{H}\psi(\boldsymbol{r}) = W\psi(\boldsymbol{r}) \tag{2.7}$$

with the *eigenfunction* $\psi(\boldsymbol{r})$ and the *eigenenergy* W. The HAMILTON *operator* $\widehat{H}$ represents the total energy (briefly just *Hamiltonian*):

$$\widehat{H} = -\frac{\hbar^2}{2m}\Delta + V(\boldsymbol{r}). \tag{2.8}$$

In the often encountered one-dimensional case the SCHRÖDINGER equation (2.6) is further simplified:

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\psi(x)}{\mathrm{d}x^2} + V(x)\psi(x) = W\psi(x). \tag{2.9}$$

2.2.2 HAMILTON and Momentum Operators

We may write the HAMILTON operator (2.8) even more suggestive:

$$\widehat{H} = -\frac{\hbar^2}{2m}\nabla^2 + V(\boldsymbol{r}) = \frac{\widehat{\boldsymbol{p}}^2}{2m} + V(\boldsymbol{r}). \tag{2.10}$$

Here we have introduced the *momentum operator* (a *vector operator*)

data 26.02.2020

c) Data_Novovivo_40_2020/Books_40_2020/Atoms, Molecules and optical physics../

c1) Atoms, Molecules and optical physics 1.pdf

Pág. 90(122) -

2.2 ~~Schrödinge~~

Schrödinger Equation

2.2.3 Time Dependent Schrödinger Equation

... which is a 2nd order partial differential equation (PDE) in space and time derived from the MAXWELL equations.

c2) Atoms, Molecules and optical Physics 2.pdf

Pág. 27(58)

Snellius diffraction

Pág. 28(59)

ABCD matrices.

Pág. 31(62) - Fig 1.16 ⇒

26.02.2020

De-focussing of a Gaussian beam.

Look:

Fig 1.15 - Pág. 30(61)

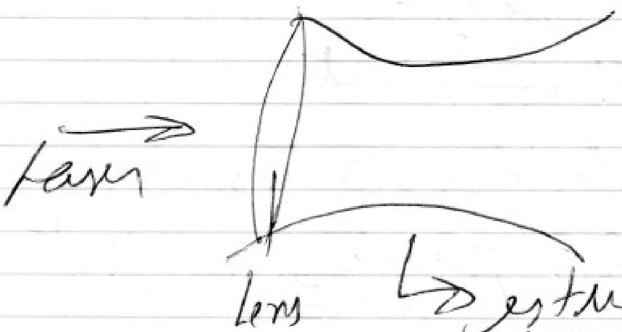

→ estude esta curvatura com a curva que distorce a gravidade.

Obs:

Fig 1.16

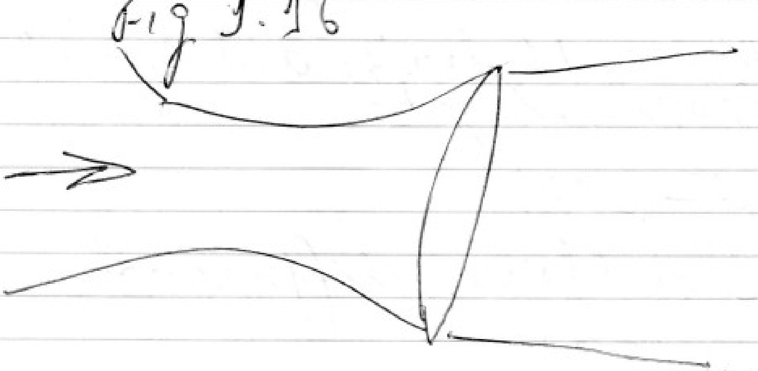

a3) Books_40_2020/
Ordinary Differential
Equations.../
saret.st-10167.pdf

26.02.2020

De-focussing of a Gaussian beam.

Look:

Fig 1.15 - Pág. 30(61)

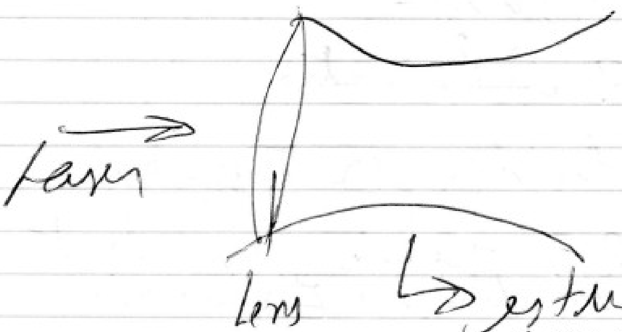

↳ estude esta curvatura com a curva que distorce a gravidade.

Obs:

Fig 1.16

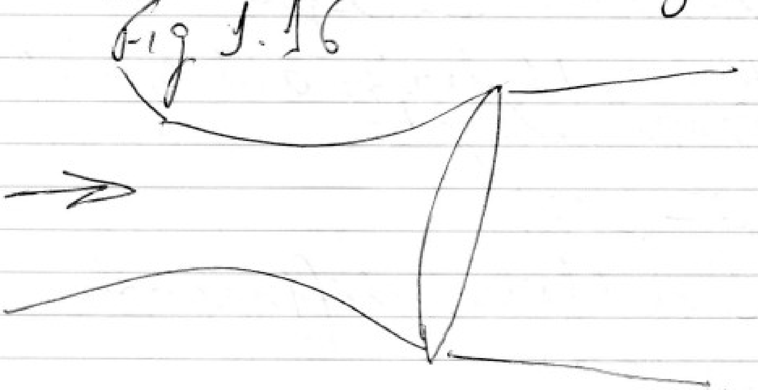

a3) Books_40_2020/ Ordinary Differential Equations.../

saret.st-10167.pdf

& ./01. . .+ . ("(2!

) " (C (

In the above crude analysis, we obtained very little information concerning the location of the unique limit cycle given in Theorem 1.6. When a differential equation contains a parameter, one can sometimes discuss the precise limiting behavior as the parameter tends to some value. This is illustrated with van der Pol's equation (1.7). Suppose k is very large; more specifically, suppose $k = \varepsilon^{-1}$ and let us determine the behavior of the periodic solution as $\varepsilon \to 0^+$. Oscillations of this type are called *relaxation oscillations*. System (1.7) is equivalent to

$$\varepsilon\dot{u} = v - G(u), \qquad \dot{v} = -\varepsilon u, \tag{1.8}$$

where $G(u) = u^3/3 - u$. From Theorem 1.6, equation (1.8) has a unique asymptotically stable limit cycle $\Gamma(\varepsilon)$ for every $\varepsilon > 0$. From (1.8), if ε is small and the orbit is away from the curve $v = G(u)$ in Fig. 1.6, then the u coordinate has a large velocity and the v coordinate is moving slowly. Therefore, the orbit has a tendency to jump in horizontal directions except when it is very close to the curve $v = G(u)$. These intuitive remarks are made precise in

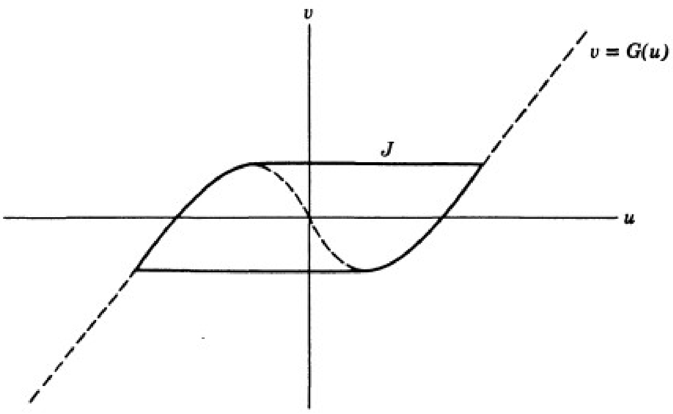

Figure II.1.6

Theorem 1.7. As $\varepsilon \to 0$, the limit cycle of (1.8) approaches the Jordan curve J shown in Fig. 1.6 consisting of arcs of the curve $v = G(u)$ and horizontal line segments.

To prove this, we construct a closed annular region U containing J such that dist(U, J) is any preassigned constant and yet for ε sufficiently small, all paths cross the boundary of U inward. U will thus contain (from the Poincaré-Bendixson theorem) the limit cycle $\Gamma(\varepsilon)$. The construction of

U is shown in Fig. 1.7 where h is a positive constant. The straight lines 81 and 45 are tangent to $v = G(u) + h$, $v = G(u) - h$ respectively and the lines 56, 12, 9–10, 13–14, are horizontal while 23, 67, 11–12, 15–16 are vertical. The remainder of the construction should be clear. The inner and outer boundaries are chosen to be symmetrical about the origin. Also marked on the figure are arrows designating the direction segments of the boundaries crossed. These are obtained directly from the differential equation and are independent of $\varepsilon > 0$. It is necessary to show that the other segments of the boundary are also crossed inward by orbits if ε is small. By symmetry, it is only necessary to discuss 34, 45 and 10–11.

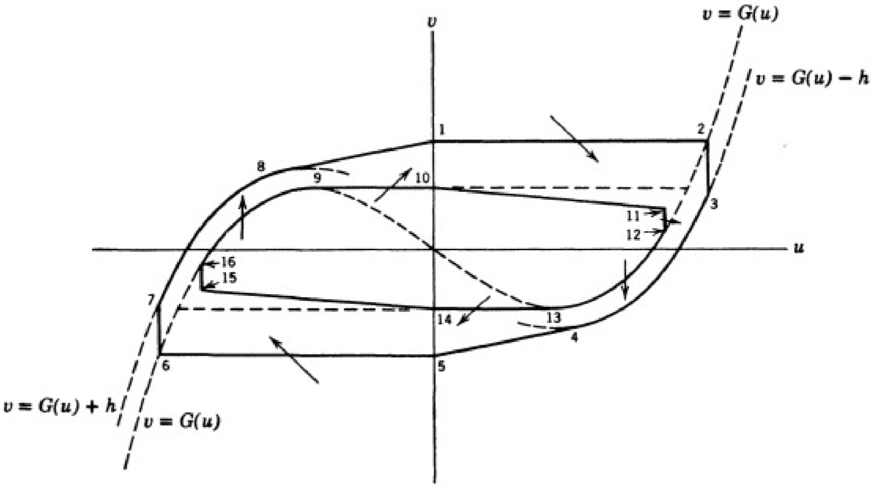

Figure II.1.7

At any point $(u, G(u) - h)$ on 34, along the orbits of (1.8), we have

$$\frac{dv}{du} = \frac{-\varepsilon^2 u}{v - G(u)} = \frac{\varepsilon^2 u}{h} < \frac{\varepsilon^2 u(3)}{h}$$

where $u(3)$ is the value of u at point 3. Hence for ε small enough, this is less than $g(4) < g(u)$ which is the slope of the curve $G(u) - h$. Thus, $\dot{v} < 0$ on this arc implies the orbits enter the region along this arc.

Along the arc 45, we have $|v - G(u)| > h$ and, hence, the absolute value of the slope of the path $|dv/du| = |-\varepsilon^2 u/[v - G(u)]| < \varepsilon^2 u(4)/h$ approaches zero as $\varepsilon \to 0$. For ε small enough this can be made $< g(4)$ which is the slope of the line 45. Thus, $\dot{v} < 0$ on this arc implies the orbits enter into U if ε is small enough.

Let K be the length of the arc 11–12. For K small enough, $|v - G(u)| > K$ along the arc 10–11. Hence, $|dv/du|$ along orbits of (1.8) is less than $\varepsilon^2 u/K < \varepsilon^2 u(11)/K$, which approaches zero as $\varepsilon \to 0$. Thus, for ε small, the orbits enter U since $\dot{u} > 0$ on this arc.

This shows that given a region U of the above type, one can always choose ε small enough to ensure that the orbits cross the boundary of U inward. This proves the desired result since it is clear that U can be made to approximate J as well as desired by appropriately choosing the parameters used in the construction.

Exercise 1.1. Prove the following *Theorem*. Any open disk in R^2 which contains a bounded semiorbit of (1.1) must contain an equilibrium point. Hint: Use the Poincaré-Bendixson Theorem and Theorem I.8.2.

Exercise 1.2. Give a generalization of Exercise 1.1 which remains valid in R^3? Give an example.

Exercise 1.3. Prove the following *Theorem*. If div f has a fixed sign (excluding zero) in a closed two cell Ω, then Ω has no periodic orbits. Hint: Prove by contradiction using Green's theorem over the region bounded by a periodic orbit in Ω.

Exercise 1.4. Consider the two dimensional system $\dot{x} = f(t, x)$, $f(t+1, x) = f(t, x)$, where f has continuous first derivatives with respect to x. Suppose Ω is a subset of R^2 which is homeomorphic to the closed unit disk. Also, for any solution $x(t, x_0)$, $x(0, x_0) = x_0$, suppose there is a $T(x_0)$ such that $x(t, x_0)$ is in Ω for all $t \geqq T(x_0)$. Prove by Brouwer's fixed point theorem that there is an integer m such that the equation has a periodic solution of period m. Does there exist a periodic solution of period 1?

Exercise 1.5. Suppose f as in exercise (1.4) and there is a $\lambda > 0$ such that $x'f(t, x) \leqq -\lambda|x|^2$ for all t, x. If $g(t) = g(t+1)$ is a continuous function, prove the equation $\dot{x} = f(t, x) + g(t)$ has a periodic solution of period 1.

Exercise 1.6. Suppose γ_0 is a periodic orbit of a two dimensional system and let G_e and G_i be the sets defined in Theorem 1.4. Is it possible for an equation to have $\alpha(\gamma(p)) = \gamma_0$ for all noncritical points p in G_i and $\omega(\gamma(q)) = \gamma_0$ for all q in G_e? Explain.

Exercise 1.7. For Lienard's equation, must there always be a periodic orbit which is stable from the outside? Must there be one stable from the inside? Explain.

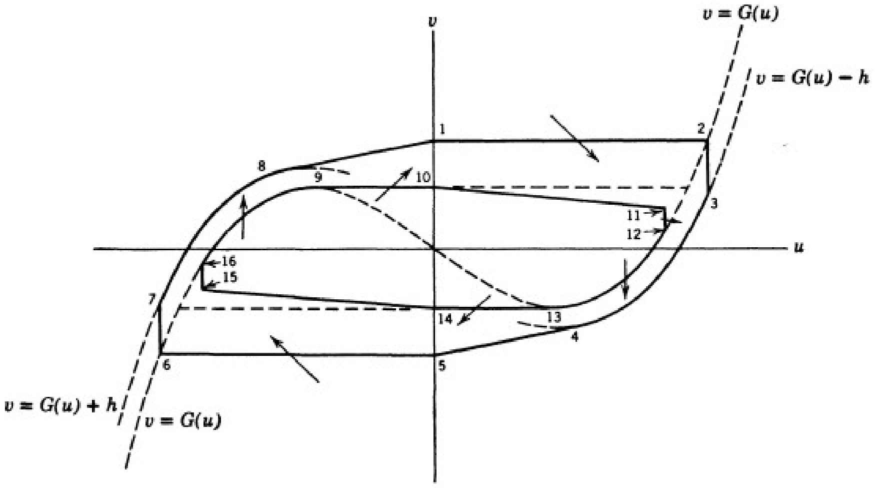

3 " (

% ./01. . .+ . (8 !9 B 8 25 C 5 2!

) & & % - (%

8.7.4 Persistence images

There is another approach that proceeds by treating a barcode, recoded as a persistence diagram, as a collection of point masses and then smoothing the corresponding measure to produce and image, which is finally discretized by selecting pixels and assigning each pixel the average value of the function on a box surrounding that pixel. It is described in [2]. The detailed description is given in several steps. The input is a persistence diagram (it is more natural to use the persistence diagram view in this case), a collection of points $B = \{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathbb{R}^2$. We assume we are given a function $\phi : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ so that (a) $\phi_u = \phi(u, -)$ is a probability distribution on $\mathbb{R}^2$ for each $u \in \mathbb{R}^2$ and (b) the mean of ϕ_u is u. A standard choice is that of a spherically symmetric Gaussian with mean u and a fixed variance σ. We also assume we are given a continuous and piecewise differentiable nonnegative weighting function $f : \mathbb{R}^2 \to \mathbb{R}$ that is zero along the x-axis.

that

$$\lambda_k(B)(t) \geq \lambda_{k+1}(B)(t)$$

and that

$$\lambda_k(B)(l) = 0 \text{ for } k > n.$$

In [12] it is also proved that each function $\lambda_k(B)$ is 1-Lipschitz, i.e. that

$$|\lambda_k(B)(t) - \lambda_k B)(t')| \leq |t - t'|.$$

To summarize, the persistence landscape lies in the vector space $\mathfrak{F}$ of real valued functions on $\mathbb{N} \times \mathbb{R}$, and it follows directly from Remark 8.7.2 that the definition gives us a function $PL : \mathfrak{B} \to \mathfrak{F}$.

One extremely useful fact about the persistence landscape is that it is compatible with the various distances assigned to barcode spaces. We let $\mathfrak{F}_p \subset \mathfrak{F}$ denote the space of functions with finite L_p-norm $|| \; ||_p$. It is clear that the function PL takes values in $\mathfrak{F}_p$ for all p. Recall the definition of the p-Wasserstein distance W_p between barcodes from Section 8.5.1. Bubenik now proves the following in [12].

Theorem 8.7.3. *The function* $d_p(B, B') = ||PL(B) - PL(B')||_p$ *is a metric on* $\mathfrak{B}$. *The two metrics* W_{p+1} *and* d_p *generate the same topology on* $\mathfrak{B}$. *It follows that the map* PL *is continuous when* $\mathfrak{B}$ *is equipped with the metric* W_{p+1} *and* $\mathfrak{F}_p$ *is equipped with the metric associated with the* L_p *norm.*

Remark 8.7.4. Bubenik also provides explicit inequalities involving the two metrics in [12]

Bubenik also proves the following continuity theorem.

Theorem 8.7.5. *The map* PL *is 1-Lipschitz from* $\mathfrak{B}$ *equipped with the bottleneck distance to the space of persistent landscapes equipped with the sup norm distance. This is equivalent to the algebraic statement*

$$|\lambda_k(B)(t) - \lambda_k(B')(t)| \leq W_\infty(B, B').$$

Remark 8.7.6. The map PL separates points.

8.7.4 Persistence images

There is another approach that proceeds by treating a barcode, recoded as a persistence diagram, as a collection of point masses and then smoothing the corresponding measure to produce and image, which is finally discretized by selecting pixels and assigning each pixel the average value of the function on a box surrounding that pixel. It is described in [2]. The detailed description is given in several steps. The input is a persistence diagram (it is more natural to use the persistence diagram view in this case), a collection of points $B = \{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathbb{R}^2$. We assume we are given a function $\phi : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ so that (a) $\phi_u = \phi(u, -)$ is a probability distribution on $\mathbb{R}^2$ for each $u \in \mathbb{R}^2$ and (b) the mean of ϕ_u is u. A standard choice is that of a spherically symmetric Gaussian with mean u and a fixed variance σ. We also assume we are given a continuous and piecewise differentiable nonnegative weighting function $f : \mathbb{R}^2 \to \mathbb{R}$ that is zero along the x-axis.

- Apply the coordinate change $(x,y) \to (x, y-x) = (\xi, \eta)$ to $\mathbb{R}^2$, to obtain the new set of points $\mathcal{B} = \{(\xi_1, \eta_1) \ldots, (\xi_n, \eta_n)\}$ of the same cardinality in $\mathbb{R}^2$. The points that correspond to short intervals are now all located near the ξ-axis. The ξ-axis itself corresponds to intervals of length zero.

- Construct a new function $\rho_{\mathcal{B}}(z) : \mathbb{R}^2 \to \mathbb{R}$, called the *persistence surface of* $\mathcal{B}$, defined by

$$\rho(z) = \sum_{i=1}^{n} f(\xi_i, \eta_i) \phi((\xi_i, \eta_i), z).$$

  Notice that ρ vanishes on the x-axis.

- To construct a finite dimensional representation, we first assume that the persistence diagrams we will be dealing with will always lie in a bounded region in $\mathbb{R}^2$, and divide a box containing that region into a square grid. Construct the vector with coordinates in one to one correspondence with the squares of the grid, and assign the entry corresponding to a square to be the integral over that square of $\rho_{\mathcal{B}}$.

The following is proved in [2].

Theorem 8.7.7. *The map* $PI : \mathfrak{B} \to \mathbb{R}^N$ *that assigns to a persistence diagram a vector using the above procedure is continuous when the metric on* $\mathfrak{B}$ *is the* 1*-Wasserstein distance.*

Remark 8.7.8. In [2], estimates proving this result are given both in the case of a general choice of ϕ and the special case where ϕ is given by Gaussian distributions with fixed variance. Of course, the estimates in the latter case are stronger.

8.8 Generalized persistence

8.8.1 Introduction

Persistent homology operates on functors F from the category $\mathbb{R}_+$ to simplicial complexes, by composing them with the homology functor to obtain a persistence vector space. It is useful to consider other parameter categories for the diagrams, which can also help clarify the structure of data sets. There are at least two such constructions that have been discussed. The first is *zig-zag* persistence, introduced in [19], and the second is *multidimensional persistence*, discussed for example in [22]. The first is designed to study the relationship between homology of constructions that are not nested within each other, such as distinct samples from a given space, and the second provides invariants of situations where it is natural to study filtrations of spaces involving more than one parameter, such as filtering by both the scale parameter R and a measure of density. We describe both extensions of the standard persistent homology methods.

8.8.2 Zig-zag persistence

Consider the triangulation of $\mathbb{R}$ whose vertices are the integers. The set of vertices of the barycentric subdivision of this simplicial complex is equipped with a partial ordering, by recognizing that its elements are in one-to-one correspondence with the simplices in the original triangulation, and that that set is equipped with a partial ordering by treating it as a subset of the power set of $\mathbb{R}$. In concrete terms, we may view it as in one to one correspondence with the set all integers and half integers, with every integer n being less than or equal to the elements $n \pm \frac{1}{2}$. We'll refer to this partially ordered set as $\mathfrak{Z}$. A partially ordered set (and its corresponding category) is said to be connected if any two objects can be connected by a zig-zag sequence of morphisms. Connected partially ordered subsets of $\mathfrak{Z}$ are always determined by a pair (x, y), where x and y are both integers or half integers, via the rule that assigns to the pair (x, y) the collection of objects z for which $x \leq z \leq y$ (in the total ordering on $\mathbb{R}$).

Definition 8.8.1. *For any category $\underline{C}$, a* zig-zag persistence object *in $\underline{C}$ is a functor from a connected subcategory of $\mathfrak{Z}$ to $\underline{C}$. Suppose further that $\underline{C}$ is equipped with an object c_0 that is both initial and terminal. Then for any connected subcategory $\mathfrak{Z}_0 \subseteq \mathfrak{Z}$ and object $c \in \underline{C}$, we define the* interval object *for $\mathfrak{Z}_0$ and c to be the functor $F = F_{\mathfrak{Z}_0, c} : \mathfrak{Z} \to \underline{C}$ defined on objects by $F(x) = c$ for all $x \in \mathfrak{Z}_0$ and $F(x) = c_0$ for any $x \notin \mathfrak{Z}_0$, and on morphisms by $F(x \leq y) = id_c$ whenever $x, y \in \mathfrak{Z}_0$. The behavior on morphisms into or out of $\mathfrak{Z}_0$ is uniquely determined by the fact that c_0 is both initial and terminal. When $\underline{C}$ is the category of vector spaces over a field K, then it is understood that c will be chosen to be a one-dimensional vector space over K.*

Given a zig-zag persistence object in the category of simplicial complexes, we may apply a homology functor $H_i(-; A)$ for an abelian group A to obtain a zig-zag persistence object in the category of abelian groups. The category of abelian groups has the zero object as an object that is both initial and terminal, and so the notion of interval objects makes sense. Of course if $A = K$, where K is a field, then we obtain a zig-zag persistence vector space. It turns out that they can be classified up to isomorphism.

Theorem 8.8.2. (P. Gabriel, [37]) *Let $\mathfrak{Z}' \subset \mathfrak{Z}$ be any finite connected subcategory, and let F denote any zig-zag persistence object in the category of finite dimensional vector spaces over a field K defined on $\mathfrak{Z}'$. Then there is a finite direct sum decomposition*

$$F \cong \bigoplus_i F_i$$

where each F_i is an interval object for $\mathfrak{Z}_0$ and K, where $\mathfrak{Z}_0 \subseteq \mathfrak{Z}'$ is a connected subcategory of $\mathfrak{Z}'$. Moreover, the sum is unique up to isomorphism and reordering of the sum.

Corollary 8.8.3. *The classification of zig-zag persistence vector spaces based on a connected subcategory $\mathfrak{Z}$ is given by barcodes where the intervals have endpoints integers or half integers. We'll refer to these barcodes as the zig-zag persistence barcodes of the zig-zag persistence vector space.*

This theory with applications is discussed in [19] and [20]. Here are some particular situations in which this construction can be used.

1. **Samples:** Given a finite metric space, one can ask to what extent the persistent homology is captured on smaller samples from the data set. For example, suppose that we have taken a very large uniform sample X from the circle, and equip them with a metric by restricting the intrinsic metric (say) on the circle to these points. We will then with high probability find that the one-dimensional persistence barcode for the Vietoris-Rips construction on X will contain one long bar and many much shorter ones. Supposing that we do not actually know that the sample is coming from a circle, but simply observe that we obtain a one-dimensional barcode with one long bar and many shorter ones. A hypothesis suggested by this observation is that the data is obtained by sampling from a space with the homotopy type of the circle, but we may wonder if instead it has somehow appeared "by accident". One way to provide confirmation of our hypothesis would be to observe that we obtain the same result for various subsamples of our space, and that they are compatible in an appropriate sense. Zig-zag persistence provides a way for carrying this out. We suppose that we have chosen samples $X_1, \ldots, X_n \subseteq X$, and create a persistence object in the category of finite metric spaces and distance non-increasing maps

$$\begin{tikzcd}
& X_1 \cap X_2 \arrow[dl] \arrow[dr] & & \cdots \arrow[dl] \arrow[dr] & & X_n \cap X_{n-1} \arrow[dl] \arrow[dr] & \\
X_1 & & X_2 & & X_{n-1} & & X_n.
\end{tikzcd}$$

   If we apply $\mathcal{V}(-, r)$ for a fixed choice of r, guided by the beginning and endpoints of the observed long bar in the barcode for X, we obtain a persistence object in the category of simplicial complexes. If we further apply $H_1(-, K)$ for a field K, we obtain a zig-zag persistence K-vector space. By the classification Theorem 8.8.2 above, we obtain a decomposition of the resulting zig-zag persistence K-vector space. The interpretation of the informal idea that each sample has a one-dimensional homology class and that they are consistent is the presence of an interval K-vector space for a relatively long interval within the set $\{1, \frac{3}{2}, 2, \ldots n - \frac{1}{2}, n\}$, or equivalently a relatively long bar in the zig-zag persistence barcode.

2. **Functions on spaces:** Suppose that we have a topological space X equipped with a continuous map $f : X \to \mathbb{R}_+$. Then we have the ordinary persistence K-vector spaces $\{H_i(f^{-1}([0, r], K)\}_r$, which encode information about the evolution of the homology of the sublevel sets of f as r increases. However, one might be interested in gaining information about approximations to *level sets* instead. They can provide more useful invariants in a number of cases, and are approachable through zig-zag persistent homology. We construct a zig-zag diagram of topological spaces as follows.

$$\begin{tikzcd}
& f^{-1}(1) \arrow[dl] \arrow[dr] & & \cdots \arrow[dl] \arrow[dr] & & f^{-1}(N-1) \arrow[dl] \arrow[dr] & \\
f^{-1}I_0 & & f^{-1}I_1 & & f^{-1}I_{N-2} & & f^{-1}I_{N-1}
\end{tikzcd}$$

   where $I_k = [k, k+1]$ for all k. Again we can apply $H_i(-, K)$ to this diagram, an obtain a zig-zag persistence barcode. It contains information about how the spaces $f^{-1}[k, k+1]$ change as k changes, and how they assemble together. This situation is studied in [20].

3. **Witness complexes:** One of the problems with the witness complex is that we have very little theory about the extent to which it reflects accurately the persistent homology of the underlying metric space. A related problem is that there is no direct relationship between the construction for two different landmark sets. Even if $\mathcal{L}_1 \subseteq \mathcal{L}_2$, there are no maps directly relating $W(X, \mathcal{L}_1, R)$ and $W(X, \mathcal{L}_2, R)$ for two different landmark sets. One approach is to attempt to assess in some manner how consistent the results of the constructions based on $\mathcal{L}_1$ and $\mathcal{L}_2$ are. It turns out that given two landmark sets $\mathcal{L}_1$ and $\mathcal{L}_2$, it is possible to construct an intermediate bivariant construction $W(X, \{\mathcal{L}_1, \mathcal{L}_2\}, R)$ for which there is an evident diagram

$$W(X, \mathcal{L}_1, R) \longleftarrow W(X, \{\mathcal{L}_1, \mathcal{L}_2\}, R) \longrightarrow W(X, \mathcal{L}_2, R).$$

This is itself a short zig-zag diagram of length three, but if we have landmark sets $\{\mathcal{L}_i\}_{i=0}^{N}$ we can clearly construct a longer diagram that includes the bivariant constructions $W(X, \{\mathcal{L}_i, \mathcal{L}_{i+1}\}, R)$ for $i = 0, \ldots, N-1$. The construction is quite simple. Its vertex set is $\mathcal{L}_1 \times \mathcal{L}_2$, and a subset

$$\{l_1^1 \times l_2^1, l_1^2 \times l_2^2, \ldots, l_1^k \times l_2^k\} \subseteq \mathcal{L}_1 \times \mathcal{L}_2$$

spans a simplex in $W(X, \{\mathcal{L}_i, \mathcal{L}_{i+1}\}, R)$ if and only if there is a point $x \in X$ so that x is a witness for $\{l_1^1, l_1^2, \ldots, l_1^k\}$ and $\{l_2^1, l_2^2, \ldots, l_2^k\}$ in the complexes $W(X, \mathcal{L}_1, R)$ and $W(X, \mathcal{L}_1, R)$, respectively. The projections to $W(X, \mathcal{L}_1, R)$ and $W(X, \mathcal{L}_1, R)$ are defined in the evident way.

8.8.3 Multidimensional persistence

There are many situations where it can be useful to introduce families of spaces varying with more than one real parameter. For example, in any kind of topological analysis of data sets, it usually is the case that if one considers the persistent homology of the entire data set, the presence of outliers means that we do not typically obtain the"right homology". For example, if we have data sampled from the unit circle, but a small number of points sprinkled throughout the unit disc, then persistent homology will end up reflecting the homology of the disc rather than that of the circle. This is often circumvented by selecting only the points of sufficient density, as measured by a density estimator, since outliers will typically have very low density. The question then becomes, though, how to choose the threshold for density. Also, it turns out that in general, there will be variation through different topologies as one changes the threshold. The solution to this problem is to attempt to study all thresholds at once, just as we do when considering the scale parameter in the Vietoris-Rips construction. This leads us to the following definition.

Definition 8.8.4. *Let $\underline{C}$ be a category. Then by an n-dimensional persistence object in $\underline{C}$ we mean a functor $\mathbb{R}_+^n \to \underline{C}$, where $\mathbb{R}_+^n$ denotes the n-fold product of copies of the category $\mathbb{R}_+$.*

Example 8.8.5. Let X be any metric space, and suppose that X is equipped with a function $f : X \to \mathbb{R}$. It might be a density estimator, but it might also be a measure of centrality. Then if $X[s]$ denotes the subset $\{x \in X | f(x) \leq s\}$, we obtain family of spaces $\mathcal{V}(X[s], r)$, and by applying homology with coefficients in a field K, we obtain a 2-dimensional persistence K-vector space parametrized by the pair (r, s). While density is used as described above to remove outliers or noise, the case of a centrality measure allows one to capture the presence of the analogues of ends in a finite metric space.

Example 8.8.6. Multidimensional persistent homology can also be used to capture geometric information that is not strictly topological (see [23]). For example, given a Riemannian manifold, one can compute Gaussian curvature at each point and filter by that quantity. Considering the entire manifold, one can obtain a one-dimensional persistence vector space by applying homology over a field K. For computational purposes, though, we would need to deal with a sample and use a second parameter, namely the scale parameter in a Vietoris-Rips complex. This kind of analysis can for example be used to distinguish between various ellipsoids.

The equivalence of categories in Proposition 8.3.1 has the following straightforward analogue.

Proposition 8.8.7. *The category of n-dimensional persistence vector spaces over K is equivalent to the category of $\mathbb{R}^n_+$-graded modules over the $\mathbb{R}^n_+$-graded ring $K[\mathbb{R}^n_+]$. The analogous result with $\mathbb{R}^n_+$ replaced by $\mathbb{N}^n$ also holds.*

Although useful, this result does not give us a classification of multidimensional persistence K-vector spaces analogous to the barcode classification that works in the one-dimensional case. The reason can be understood as analogous to the commutative algebraic situation, where finitely generated $K[x]$-modules can be classified because $K[x]$ is a principal ideal domain, but $K[x_1, \ldots, x_n]$-modules cannot be classified for $n \geq 2$. In a sense it is provable that there is no classification strictly analogous to the $n = 1$ case, because the classification in the case $n \geq 2$ depends on the structure of the field K, as is demonstrated in [22]. This is not the case when $n = 1$, because the classification is always by barcodes, and that is independent of the structure of K.

Although there is no complete classification of multidimensional persistence vector spaces, there exist interesting invariants. For any k-dimensional persistence K-vector space $\{V_x\}_{x \in \mathbb{R}^k}$, and pair of points $x, y \in \mathbb{R}^k$, with $x \leq y$ in the natural partial ordering on $\mathbb{R}^k$, we can define $r(x, y)$ to be the rank of the linear transformation $V_x \to V_y$. We extend the definition to all pairs $x, y \in \mathbb{R}^k$ by setting $r(x, y) = 0$ when x is not less than or equal to y. The function r is therefore defined on $\mathbb{R}^k \times \mathbb{R}^k$, and we refer to it as the *rank invariant*. In the case $k = 1$, the rank invariant is complete, in that it differentiates between distinct barcodes. There is an analogue for multidimensional persistence K-vector spaces of the results described in Section 8.7.2. To state it, we first define an analogue of barcodes. By a *cube* in $\mathbb{R}^k$, we mean a set of the form $I_1 \times I_2 \times \cdots \times I_k$, where each I_s is an interval $[a_s, b_s]$, and write $C(a_1, \ldots, a_k, b_1, \ldots, b_k)$ for this cube. We write $\mathcal{C}_k$ for the set of all k-dimensional cubes. There is a straightforward analogue to the space $\mathfrak{B}$ defined in Section 8.7.1. We first define $B_n = Sp^n(\mathcal{C}_k)$, and define $\mathfrak{B}(k)_+ = \coprod Sp^n(\mathcal{C}_k)$. Next, we say a cube $C(a_1, \ldots, a_k, b_1, \ldots, b_k)$

is *negligible* if $a_i = b_i$ for some i, and define an equivalence relation $\sim$ on $\mathfrak{B}(k)_+$ to be the equivalence relation generated by the relation

$$\{C_1, C_2, \ldots C_n\} \sim \{C_1, \ldots, C_{n-1}\} \text{ for all negligible } C_n.$$

We define $\mathfrak{B}(k)$ to be $\mathfrak{B}(k)_+/\sim$. Every equivalence class under $\sim$ has a unique minimal representative consisting entirely of non-negligible cubes. Let $\mathcal{M}(k)$ denote the set of isomorphism classes of k-dimensional persistence K-vector spaces. For each cube $C = C(a_1, \ldots, a_k, b_1, \ldots, b_k)$ we let $\mu(C)$ denote the isomorphism class of the the k-dimensional persistence K-vector space $\{V_{\vec{x}}\}_{\vec{x} \in \mathbb{R}^k_+}$ for which $V_{\vec{x}} = K$ whenever $\vec{x} \in C$, $V_{\vec{x}} = \{0\}$ whenever $\vec{x} \notin C$, and for which all induced morphisms $V_{\vec{x}} \to V_{\vec{y}}$ for $\vec{x} \leq \vec{y}$ and $\vec{x}, \vec{y} \in C$ are equal to the identity. There is an obvious map $\theta : \mathfrak{B}(k) \to \mathcal{M}(k)$ which assigns to each minimal representative $\{C_1, \ldots, C_n\}$ the direct sum $\oplus_i \mu(C_i)$.

Theorem 8.8.8. [64] *The constructions above satisfy the following properties.*

1. *The set $\mathfrak{B}(k)$ is a subset of the set of real points of an affine scheme Spec(A).*
2. *The ring A is complicated, but there is a $\mathbb{G}_m$-action on Spec(A) that allows us to define a more manageable subring $A^{fin} \subseteq A$.*
3. *A^{fin} is isomorphic to the polynomial ring $\mathbb{R}[x_{\vec{a},\vec{b}}]$ where $\vec{a}$ and $\vec{b}$ are k-vectors of integers for which $a_i \geq 1$ and $b_i \geq 0$ for all i.*
4. *The ring A^{fin} separates points in $\mathfrak{B}(k)$, and maps injectively to the ring of all real valued functions on $\mathfrak{B}(k)$.*
5. *There is a natural lift of the ring homomorphism $A^{fin} \to F(\mathfrak{B}(k), \mathbb{R})$ along θ to a ring homomorphism $j : A^{fin} \to F(\mathcal{M}(k), \mathbb{R})$. $F(X, \mathbb{R})$ denotes the ring of real-valued functions on a set X.*
6. *For any $\alpha \in A^{fin}$, the function $j(\alpha) : \mathcal{M}(k) \to \mathbb{R}$ factors through the rank invariant. For any two elements $X, Y \in \mathcal{M}(k)$, if for all $\alpha \in A^{fin}$, $j(\alpha)(X) = j(\alpha)(Y)$, then the rank invariants of X and Y are equal.*

This result gives one approach to the study of multidimensional persistence. There is a great deal of other work on this topic. The paper [24] deals with computing persistent homology using commutative algebra techniques, via the equivalence of categories from Proposition 8.8.7. They demonstrate that the multigrading yields significant simplification. In [62] and [24], an algebraic framework is constructed for dealing with the fact that there is alway noise in the applications to data analysis. In [14], multidimensional persistence is studied by examining the family of all one dimensional persistence modules obtained by considering lines with varying angles in the persistence domain. M. Lesnick in [49] has defined metric properties of the set of isomorphism classes of multidimensional persistence vector spaces, and proven uniqueness results for them. Finally, in [50], software is developed for visualization and interrogation of two-dimensional persistence vector spaces.

One generalization that has not been studied yet is to multidimensional persistence where some of the persistence directions might be "zig-zag" directions rather than ordinary persistence directions. Formally, this would mean functors from the categories of the form $\mathbb{R}_+^m \times \mathfrak{Z}^n$. This would be very useful in a number of situations. For example, in the zig-zag constructions discussed in Section 8.8 for samples and for witness complexes, we were forced to choose a threshold for the scale parameter. If we had a way of representing functors from $\mathbb{R}_+ \times \mathfrak{Z}$ to vector spaces, we would not be forced to make this selection.

8.9 Coverage and evasion problems

There is an interesting set of technologies used for sensing of various kinds called *sensor nets*. A sensor net consists of a collection of sensors distributed throughout a domain. The sensors are very primitive in the sense that they are only capable of sensing the presence of an intruder or of another sensor within a fixed detection radius R. We also assume each sensor is given an identifying label or number, and that other sensors can sense that identifier when they are within the radius R of each other. One problem of interest is whether or not the balls of radius R cover the domain, and it does not have an immediate solution due to the fact that the positions of the sensors are not available. V. de Silva, and R. Ghrist have developed a very interesting method for addressing this problem based on persistent homology (see [31] and [32]). The rough idea is as follows. Suppose that one has a domain D in the plane, with a connected and compact curve boundary ∂D, that one has a collection of points $\{v_i\}_{i\in I}$ in the region, one for each sensor, and that one knows in some way that ∂D is covered by the open balls $B_R(v_i)$. Let $\mathcal{U}$ denote the family of open subsets $\{B_R(v_i)\cap D\}_{i\in I}$ of D, and let $\mathcal{U}^\partial$ denote the covering $\{B_R(v_i)\cap \partial D\}_{i\in I}$ of ∂D. Suppose further that one knows that spaces

$$B_R(v_{i_1})\cap\cdots\cap B_R(v_{i_s})\cap D \quad\text{and}\quad B_R(v_{i_1})\cap\cdots\cap B_R(v_{i_s})\cap \partial D$$

are all either contractible or empty, for all choices of subsets $\{i_1,\ldots,i_s\}\subseteq I$. The conditions ensure that D is homotopy equivalent to the nerve of $\mathcal{U}$, and that ∂D is homotopy equivalent to the nerve of the covering $\mathcal{U}^\partial$, as a consequence of the nerve theorem. It further ensures that the pair $(N.\mathcal{U}, N.\mathcal{U}^\partial)$ is equivalent to the pair $(D,\partial D)$. For any field, the relative group $H_2(D,\partial D;K)\cong K$, since D is a connected orientable manifold with boundary ∂D, and it follows that the relative group $H_2(N.\mathcal{U}, N.\mathcal{U}^\partial;K)\cong K$. On the other hand, suppose that the sets $B_R(v_i)$ do not cover all of D, and let $D_0\subseteq D$ denote the union

$$\bigcup_i B_r(v_i)\cap D.$$

The space D_0 is a non-compact manifold with boundary, and consequently the relative group $H_2(D_0,\partial D;K)\cong 0$. As above, it will follows that

$$H_2(N.\mathcal{U}, N.\mathcal{U}^\partial;K)\cong 0.$$

date 26.02.2020

S T [✓ Q] Q S S D

Pág. 61(71) Theorem 17

Jordan curve J

U will thus contain (from the Poincaré-Bendixson theorem) the limite cycle Γ (E).

Look: Fig II.1.7.

c4) Book_40_2020/ Handbook_of_Homotopy_Theory/

Handbook of Homotopy theory.pdf

Pág. 315(324)

8.7.4 Persistence images

A standard choice is that of a spherical symmetric Gaussian with mean μ and a fixed variance σ.

Pág. 316(325)

Pág. 320(329)

Example 8.8.6-

For examples, give a RIEMANNIAN manifold, one can

26.02.2020

compute Gaussian curvature

↳ comparar com a curvatura do espaço (gravidade).

at each point and filter by that quantity.

Pág. 326(335) - Gaussian random

Pág. 327(336) - Gaussian field.

Pág. 328(337) →

[9] Bobrowski and Borman

Euler integration of Gaussian random fields and persistent homology.

Pág. 970(979) →

[69] Levine

Motivic Gauss-Bonnet formulas

e4) Books_40_2020/
Geometry_the Line and the Circle

Theorem 8.10.6. [8] *For fixed n and k, and $\mathcal{P}$, $\Delta_{\mathcal{P}}(n,k,-)$ is uniformly robust with robustness coefficient r and estimate bound $nr/(1+r)$ for any r.*

It is also possible to obtain a somewhat simpler result, which does not require calculation of the full distribution $\mu(n,k,X)$. Instead of fixing a reference distribution $\mathcal{P}$, we choose a reference barcode $B \in \overline{\mathfrak{B}}$, and define $\Delta_B^{med}(n,k,X)$ to the median of the distribution of $d_B(B,-)$ applied to samples of k-dimensional barcodes attached to samples of size n taken from the metric measure space X.

Theorem 8.10.7. *For fixed n, k, and B, the function $\Delta_B^{med}(n,k,-)$ from finite metric spaces (with uniform probability measure) is robust with robustness coefficient greater than* $\ln 2/n$.

8.10.3 Random fields

Suppose that we have a function f on a manifold. We have seen in Section 8.5 that we can associate to f the persistence vector spaces $\{H_i(f^{-1}((-\infty,r],K)\}_r$ for i a nonnegative integer and K a field, and further that there are stability results that show that small changes in the function f lead to small changes, as measured by the bottleneck distance, in the corresponding sublevel set barcode. One can also ask, though, what the expected behavior of various of the features attached to barcode space is for a class of functions chosen at random. An initial question is what one means by a function chosen at random. The notion of a *random field* is defined in [4] as follows.

Definition 8.10.8. *By a* real-valued random field *on a topological space T we mean a measurable mapping*

$$F : \Omega \to \mathbb{R}^T$$

*where $\mathbb{R}^T$ denotes the set of all real-valued functions on T and $(\Omega, \mathcal{F}, \mathbb{P})$ is a complete probability space. The set $\mathbb{R}^T$ is equipped with the product σ-algebra over the set T. F creates a probability measure on $\mathbb{R}^T$, from which one can sample. Similarly, one can define an s-*dimensional vector-valued random field *as a measurable mapping*

$$F : \Omega \to (\mathbb{R}^T)^s.$$

Remark 8.10.9. Note that the definition does not use the topology of T. However, in much of the work done in this area, one studies additional hypotheses on the random fields in which the topology of T plays a role.

The idea here is that rather than being a function, a random field is an assignment to each $t \in T$ a distribution on $\mathbb{R}$, rather than a fixed value. Each of the restrictions $\mathbb{R}^T \to \mathbb{R}^{\{t\}} \cong \mathbb{R}$ produces a random variable, and therefore the corresponding distribution, which we denote F_t. In fact, for any finite set of points $t_1, \ldots, t_n \in T$, we obtain a distribution on $\mathbb{R}^n$ which we denote $F_{t_1,\ldots,t_n}$. There is a particular class of random fields called the *Gaussian random fields* that is particularly amenable to analysis.

Definition 8.10.10. *A real-valued random field F is a* Gaussian random field *if for all n, the n-dimensional distributions $F_{t_1,\ldots,t_n}$ are multivariate Gaussian distributions on $\mathbb{R}^n$.*

An s-dimensional random field is Gaussian, if all the distributions $F_{t_1,\ldots,t_n}$ are multivariate Gaussian distributions on $\mathbb{R}^{n\times s}$. Note that in Gaussian fields, the behavior of the random function is completely determined by the expectation function $m(t) = E(F_{t_1,\ldots,t_n})$, and the covariance function C.

The first example of this construction comes out of work of Wiener (see [67] and [6]), using analysis of Brownian motion. Wiener studied the case $T = \mathbb{R}_+$, and produced a Gaussian random field W, where the expected value of W_t is always $= 0$ and the variance is given by $C(s,t) = \min(s,t)$. He also showed that when one samples from the associated measure on $\mathbb{R}^{\mathbb{R}_+}$, one obtains continuous functions with probability 1, and so one calls W a *continuous Gaussian field.* Given any analytic property of functions on manifolds, such as k-th order differentiability, smoothness, or the property of being a Morse function, one can create and study Gaussian random fields whose samples have the given property with probability 1. Further, there are frequently a priori conditions on the covariance function of the random field that can be readily verified and guarantee the satisfaction of such properties.

The paper [9] proves a result concerning the persistent homology of the sublevel sets of functions sampled from Gaussian random fields. We consider the real-valued function σ on barcodes given by

$$\sigma\{[a_1,b_1],\ldots,[a_n,b_n]\} = \sum_i (b_i - a_i).$$

For any fixed $x \in \mathbb{R}$ and barcode $\beta = \{[a_1,b_1],\ldots,[a_n,b_n]\}$, we define the *$x$-truncation of β*, $\beta[x]$, to be the barcode

$$\{[a_1,\min(b_1,x)],\ldots,[a_n,\min(b_n,x)]\}$$

where it is understood that for any i such that $x \leq a_i$, the interval $[a_i,b_i]$ is simply deleted. Finally, we define

$$\chi^{pers}(M,f,x) = \sum_{i=0}^{\infty}(-1)^i\sigma(\{H_i(f^{-1}((-\infty,r],K)\}_r[x]).$$

In [9], the following result is proved concerning the distribution of $\chi^{pers}(M,f,x)$ for Gaussian random fields on Riemannian manifolds which produce Morse functions with probability one.

Theorem 8.10.11. *Let M be a closed d-dimensional Riemannian manifold, and let F be a smooth real valued Gaussian random field so that F is Morse with probability 1, and so that the mean is identically zero and the variance is identically equal to 1. Then for any $x \in \mathbb{R}$, we have*

$$\mathbb{E}\{\chi^{pers}(M,f,x)\} = \chi(M)(\varphi(x) + x\Phi(x)) + \varphi(x)\sum_{j=1}^{d}(2\pi)^{-j/2}\mathcal{L}_j(M)H_{j-2}(-x)$$

where

1. *H_n denotes the n-th Hermite polynomial.*
2. *$\varphi(x) = (2\pi)^{-1/2}e^{-x^2/2}$ is the density function for the standard Gaussian distribution.*

3. $\Phi(x) = \int_{-\infty}^{x} \varphi(u)du$.

4. $\mathcal{L}_j(M)$ *denotes the* j*-th Lipschitz-Killing curvature of* M *(defined for example in [4], Section 7.6) with respect to a metric constructed from the covariance metric attached to* F.

Remark 8.10.12. The point of this result is that it gives a theoretical estimate for the persistent Euler characteristic of sublevel sets in terms of classical invariants of the manifols. Also, the result in [9] is actually proved in a much more general context, that of regular stratified spaces and stratified Morse theory, which in particular permits the study of manifolds with boundary. It also includes the study of random fields that are of the form $G \circ (F_1, \ldots, F_k)$, where G is a deterministic function from $\mathbb{R}^k$ to $\mathbb{R}$, and $(F_1, \ldots, F_k)$ is a vector-valued Gaussian random field.

Bibliography

[1] H. Adams and G. Carlsson, *Evasion paths in mobile sensor networks*, The International Journal of Robotics Research, 34 (1), 2015, 90–104.

[2] H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta, and L. Ziegelmeier, *Persistence images: a stable vector representation of persistent homology*, J. Mach. Learn. Res., 18, 2017, 1–35.

[3] A. Adcock, E. Carlsson, and G. Carlsson, *The ring of algebraic functions on persistence barcodes*, Homology, Homotopy, and Appl., 18, 2016, 381–402.

[4] R. Adler and J. Taylor, *Random fields and geometry*, Springer, 2009.

[5] N. Akkiraju, H. Edelsbrunner, M. Facello, P. Fu, E. Mucke, and C. Varela, *Alpha shapes: defintion and software*, in Proc. Internat. Comput. Geom. Software Workshop 1995.

[6] P. Baldi, *Stochastic calculus, an introduction through theory and exercises*, Springer Universitext, 2017.

[7] S. Barannikov, *The framed Morse complex and its invariants* Adv. Soviet Math., 21, 1994, 93–115.

[8] A. Blumberg, I. Gal, M. Mandell, and M. Pancia, *Robust statistics, hypothesis testing, and confidence intervals for persistent homology on metric measure spaces*, Found. Comput. Math., 14, 2014, 745–789.

[9] O. Bobrowski and M. Borman, *Euler integration of Gaussian random fields and persistent homology*, J. Topol. and Anal., 4 (1), 2012, 49–70.

[10] O. Bobrowski, M. Kahle, and P. Skraba, *Maximally persistent cycles in random geometric complexes*, Ann. Appl. Probab., 27 (4), 2017, 2032–2060.

[11] B. Bollobás, *Random graphs*, second edition, Cambridge University Press, 2011.

[12] P. Bubenik, *Statistical topological data analysis using persistence landscapes*, The Journal of Machine Learning Research 16 (1), 2015, 77–102.

[13] L.Buhovsky, V. Humilière, S. Seyfaddini. *The action spectrum and* C^0 *symplectic topology*, arXiv:1808.09790, 2018.

[14] F. Cagliari, B. Di Fabio, and M. Ferri, *One-dimensional reduction of multidimensional persistent homology*, Proc. Amer. Mat. Soc., 138 (8), 2010, 3003–3017.

[15] Z. X. Cang, Lin Mu and G. Wei, *Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening*, PLOS Computational Biology, 14 (1), 2018, e100592.

[16] Z. X. Cang and G. Wei, *TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions*, PLOS Computational Biology, 13 (7), 2017, e1005690.

[17] G. Carlsson, *Topology and data*, Bull. Amer. Math. Soc., 46 (2), 2009, 255–308.

[18] G. Carlsson, *Topological pattern recognition for point cloud data*, Acta Numer., 23, 2014, 289-368

[19] G. Carlsson and V. de Silva, *Zigzag persistence*, Found. Comput. Math., 10 (4), 2010, 367–405

M 4 ! ! !2 5
) $($($ W"$X I7 7 4 D

[43] Jean-Marc Fontaine and Barry Mazur. Geometric Galois representations. In *Elliptic curves, modular forms, & Fermat's last theorem (Hong Kong, 1993)*, Ser. Number Theory, I, pages 41–78. Int. Press, Cambridge, MA, 1995.

[44] Eberhard Freitag and Reinhardt Kiehl. *Étale cohomology and the Weil conjecture, Ergeb. Math. Grenzgeb. (3)*, 13. Springer-Verlag, 1988.

[45] E. M. Friedlander. *Etale homotopy of simplicial schemes.* Ann. of Math. Stud., 104, Princeton University Press, 1982.

[46] William Fulton. *Intersection theory, Ergeb. Math. Grenzgeb. (3).* 2, Springer Verlag, 1984.

[47] Alexander Grothendieck. Éléments de géométrie algébrique: IV. Étude locale des schémas et des morphismes de schémas, Quatrième partie. *Publ. Math. Inst. Hautes Études Sci.*, 32:5–361, 1967.

[48] A. Grothendieck and M. Raynaud. *Revêtements étales et groupe fondamental.* Lect. Notes in Math., Vol. 224, Springer-Verlag, 1971.

[49] Yonatan Harpaz and Tomer M. Schlank. Homotopy obstructions to rational points. In *Torsors, étale homotopy and applications to rational points, London Math. Soc. Lecture Note Ser.*, 405:280–413. Cambridge University Press, 2013.

[50] Robin Hartshorne. *Algebraic geometry, Grad. Texts in Math.* 52, Springer-Verlag, 1977.

[51] I. N. Herstein. *Noncommutative rings, Carus Math. Monogr.* 15 Math. Assoc. of America, 1994.

[52] Philip S. Hirschhorn. *Model categories and their localizations. Math. Surveys Monogr.* 99 Amer. Math. Soc., 2003.

[53] Asaf Horev and Jake P. Solomon. The open Gromov–Witten–Welschinger theory of blowups of the projective plane. arXiv:1210.4034, 2012.

[54] Marc Hoyois. A quadratic refinement of the Grothendieck-Lefschetz-Verdier trace formula. *Algebr. Geom. Topol.*, 14(6):3603–3658, 2014.

[55] Marc Hoyois. The six operations in equivariant motivic homotopy theory. *Adv. Math.*, 305:197–279, 2017.

[56] Daniel C. Isaksen. Etale realization on the $\mathbb{A}^1$-homotopy theory of schemes. *Adv. Math.*, 184(1):37–63, 2004.

[57] Daniel C. Isaksen. Flasque model structures for simplicial presheaves. *K-Theory*, 36(3-4):371–395, 2005.

[58] J. F. Jardine. Simplicial presheaves. *J. Pure Appl. Alg.*, 47(1):35–87, 1987.

[59] John F. Jardine, *Local homotopy theory.* Springer Monogr. Math., 2015.

[60] André Joyal. Letter to Grothendieck, 1983, Available at https://webusers.imj-prg.fr/~georges.maltsiniotis/ps/lettreJoyal.pdf.

[61] Jesse Kass and Kirsten Wickelgren. The class of Eisenbud–Khimshiashvili–Levine is the local $\mathbb{A}^1$-brouwer degree. Duke Math. J. 168(3):429–469, 2016.

[62] Jesse Kass and Kirsten Wickelgren. An arithmetic count of the lines on a smooth cubic surface. arXiv:1708.01175, 2017.

[63] Jesse Leo Kass and Kirsten Wickelgren. An étale realization which does NOT exist. In *New Directions in Homotopy Theory, Contemp. Math.* 707, pages 11–30. Amer. Math. Soc., 2018

[64] N. Mohan Kuma and M. Pavaman Murthy. Algebraic cycles and vector bundles over affine three-folds. *Ann. of Math.*, 116(3):579–591, 1982.

[65] T. Y. Lam. *Introduction to quadratic forms over fields, Grad. Stud. Math.* 67, Amer. Math. Soc., 2005.

[66] Marc Levine. Toward a new enumerative geometry. arXiv:1703.03049, 2017.

[67] Marc Levine. Motivic Euler characteristics and Witt-valued characteristic classes. arXiv:1806.10108, 2018.

[68] Marc Levine. Toward an algebraic theory of Welschinger invariants. arXiv:1808.02238, 2018.

[69] Marc Levine and Arpon Raksit. Motivic Gauss–Bonnet formulas. arxiv:1808.08385, 2018.

[70] Jacob Lurie. *Higher topos theory, Ann of Math. Stud.*, 170. Princeton University Press, 2009.

[71] Saunders Mac Lane and Ieke Moerdijk. *Sheaves in Geometry and Logic: A First Introduction to Topos Theory.* Universitext. Springer-Verlag, 1992.

[72] Satya Mandal and Raja Sridharan. Euler classes and complete intersections. *J. Math. Kyoto Univ.*, 36(3):453–470, 1996.

[73] J. Peter May. Classifying spaces and fibrations. *Mem. Amer. Math. Soc.*, 1(1,155), 1975.

%A J ./01. . .+ . (0 9 D G !
9 ! 9 7 "A 2 2!
) & &$

and the first half of the 1700s. The first examples involved arc lengths of ellipses and led to the name "elliptic integrals." In the mid-1700s, Leonhard Euler revolutionized the study of elliptic integrals by establishing the identity

$$\int_a^{x_1} g(t)^{-1/2}\,dt + \int_a^{x_2} g(t)^{-1/2}\,dt = \int_a^{x_3} g(t)^{-1/2}\,dt \tag{8}$$

for any polynomial $g(t)$ of degree 3 of 4, where x_3 is a rational function of x_1, x_2, a, $g(x_1)^{1/2}$, $g(x_2)^{1/2}$, $g(a)^{1/2}$.

Certain cubics are now called "elliptic curves" because of their connection with elliptic integrals. This connection was discovered by Gauss, Niels Abel, and Carl Jacobi in the 1820s. Their results were clarified and extended by Riemann in the 1850s, Weierstrass in 1863, and Henri Poincaré in 1901. We summarize a small part of this work below.

Let

$$g(t) = 4t^3 + ct + d$$

be a polynomial of degree 3 without repeated roots. The Weierstrass P-function

$$x = P(u) \tag{9}$$

parametrizes the nonsingular, irreducible complex cubic

$$y^2 = g(x) \tag{10}$$

in the following sense: Equation (9) and the equation

$$y = P'(u) \tag{11}$$

match up the complex numbers u on and inside a parallelogram in the complex plane with the points (x, y) of the complex cubic (10), except that any two complex numbers u in corresponding positions on opposite sides of the parallelogram map to the same point (x, y). The function $P(u)$ can be written in the form

$$P(u) = \frac{1}{u^2} + a_2u^2 + a_4u^4 + \cdots$$

for complex numbers a_2, a_4, ...

Equations (9)–(11) imply that

$$\frac{dx}{du} = P'(u) = y = g(x)^{1/2},$$

and taking reciprocals gives

$$\frac{du}{dx} = g(x)^{-1/2}.$$

C G

9 ! ZZ 2 7 UU 9
2 C ! 7 ! 95 4 A
> /9 A ! [9 - C T ! !
G ! ! 95 0 - A E -"&A ! 8
\ $ E : 2 B 9

) & - &&" 4 P I

know this result for polynomials in two variables over the complex numbers, since Theorem 15.7 applies to the homogenizations.

Theorem 16.11 is a version of Gauss's Lemma, a standard result in abstract algebra. We continue to work over the real numbers except where we state otherwise.

Theorem 16.11
Let $f(u, w, x)$ *and* $g(u, w, x)$ *be polynomials in indeterminates* u, w, *and* x, *and let* $b(u, w)$ *be an irreducible polynomial in* u *and* w. *If neither* $f(u, w, x)$ *nor* $g(u, w, x)$ *has* $b(u, w)$ *as a factor, then neither does* $f(u, w, x)g(u, w, x)$.

know this result for polynomials in two variables over the complex numbers, since Theorem 15.7 applies to the homogenizations.

Theorem 16.11 is a version of Gauss's Lemma, a standard result in abstract algebra. We continue to work over the real numbers except where we state otherwise.

Theorem 16.11
Let $f(u,w,x)$ and $g(u,w,x)$ be polynomials in indeterminates u, w, and x, and let $b(u,w)$ be an irreducible polynomial in u and w. If neither $f(u,w,x)$ nor $g(u,w,x)$ has $b(u,w)$ as a factor, then neither does $f(u,w,x)g(u,w,x)$.

Proof
Collecting terms by powers of x gives

$$f(u,w,x)=\sum c_i(u,w)x^i \qquad \text{and} \qquad g(u,w,x)=\sum d_i(u,w)x^i$$

for polynomials c_i and d_i in u and w. Because neither f nor g has $b(u,w)$ as a factor, we can let k and l be the smallest integers such that $c_k(u,w)$ and $d_l(u,w)$ do not have $b(u,w)$ as a factor over the real numbers. Over the complex numbers, neither $c_k(u,w)$ nor $d_l(u,w)$ has an irreducible factor in common with $b(u,w)$ (by Theorem 11.9), and so neither does $c_k(u,w)d_l(u,w)$ (by Theorem 15.7 applied to the homogenizations). Thus, $b(u,w)$ is not a factor of $c_k(u,w)d_l(u,w)$.

We write

$$f(u,w,x)g(u,w,x)=\sum e_i(u,w)x^i \tag{88}$$

for polynomials e_i in u and w. The coefficient $e_{k+l}(u,w)$ of x^{k+l} is

$$\begin{aligned}&c_kd_l+c_{k-1}d_{l+1}+c_{k-2}d_{l+2}+\cdots\\&\qquad+c_{k+1}d_{l-1}+c_{k+2}d_{l-2}+\cdots.\end{aligned}$$

Because b is a factor of $c_{k-1},\ldots,c_0$ and $d_{l-1},\ldots,d_0$, it is not a factor of e_{k+l}: if it were, then it would be a factor of

$$\begin{aligned}&e_{k+l}-c_{k-1}d_{l+1}-c_{k-2}d_{l+2}-\cdots\\&\qquad-c_{k+1}d_{l-1}-c_{k+2}d_{l-2}-\cdots=c_kd_l,\end{aligned}$$

contradicting the previous paragraph. It follows from (88) that $b(u,w)$ is not a factor of fg. □

Applying the previous theorem repeatedly gives another special case not of the result that polynomials factor uniquely.

Theorem 16.12
Let $p(u,w,x)$ and $f(u,w,x)$ be polynomials in indeterminates u, w, and x, and let $b(u,w)$ be a nonzero polynomial that does not involve x. If f is irreducible and a factor of bp, then it is a factor of b or p.

) & &- !P/ 9 P I 6 ! B
C ! 2 G 7

peatedly can produce any number of points of C with rational coordinates from just one. This addresses a central problem of number theory, finding the rational solutions of equations. An ad hoc algebraic version of the tangent–secant construction was introduced by Diophantus, who lived in Alexandria during the third century A.D. Fermat systematized the construction algebraically, and Newton interpreted it geometrically in terms of tangents and secants.

Complex numbers were introduced in the 1500s to solve cubic equations in one variable. They were reintroduced in the 1700s to facilitate integration by partial fractions. Mathematicians gradually developed proficiency in working with complex numbers, and their confidence increased when Carl Friedrich Gauss gave four proofs of the Fundamental Theorem of Algebra in the first decade of the 1800s. Jean d'Alembert had given an incomplete proof of the Fundamental Theorem in 1746. The main gap in his proof was filled in 1806 when Jean Argand proved a result generally called "d'Alembert's Lemma" (Claim 5 of Section 10). In fact, no truly complete proof of the Fundamental Theorem could be given until the 1870s, when Georg Cantor and Richard Dedekind developed the real numbers formally and Karl Weierstrass derived the basic properties of continuous functions. The proof in Section 10 of the Fundamental Theorem is based on the paper of Charles Fefferman cited in the References, which modernizes and simplifies the work of d'Alembert and Argand.

The idea of a complex curve—an algebraic curve whose coefficients and variables are complex numbers—emerged over centuries. Analytic geometers from Newton onward considered "imaginary points" on curves without clearly specfying the nature of these points. In the 1820s, Jean Poncelet and Michel Chasles argued for using imaginary points systematically in synthetic projective geometry. In 1830, Plücker clarified the nature of imaginary points when the homogeneous coordinates he introduced made it possible to consider points with complex coordinates. Nevertheless, complex curves were not generally considered natural objects of study until Georg Riemann proposed in 1851 a way to consider them topologically: the "Riemann surface" of a polynomial equation $f(w, z) = 0$ consists of sheets that lie over the complex z-plane and correspond to the values of w determined by the equation. In the 1860s, Alfred Clebsch and Paul Gordan recast Riemann's ideas from complex analytic to geometric form, and the modern view of complex curves was established.

"Elliptic integrals" are, speaking roughly, integrals that involve the square root of a polynomial of degree 3 or 4. Unlike integrals that involve the square root of a polynomial of degree 2, elliptic integrals cannot generally be evaluated in closed form. Examples of elliptic integrals arose from scientific and geometric considerations in the last half of the 1600s

8 5 9

III

CHAPTER

Cubics

Introduction and History

Introduction

This chapter is devoted to classifying irreducible cubics. These are curves of degree 3 given by polynomials that do not have factors of degree 1 or 2. We prove that every irreducible cubic can be transformed into the form

$$y^2 = x^3 + fx^2 + gx + h \tag{1}$$

for real numbers f, g, h.

The proof has two main steps. First, we prove in Section 8 that an irreducible cubic C can be transformed into (1) if it has a flex (i.e., a generalized inflection point) or a singular point. Second, we prove in Section 12 that there is a flex on every irreducible cubic that is nonsingular (i.e., has no singular points), and so the previous sentence applies to all irreducible cubics.

In Section 9, we interrupt our work on the classification of cubics to discuss one of their most important properties. We use collinearity of points to define addition on a nonsingular, irreducible cubic C that has a flex O. This definition makes C an abelian group, which means that the sum of two points of C is again a point of C, addition is commutative and associative, O is an identity element, and every point of C has an additive inverse. A central problem in number theory is to determine the set C^* of points of C that have rational coordinates, when C is given by (1) for rational numbers f, g, h. The key to this problem is to observe that C^* is itself a group whose structure can be analyzed.

Sections 10 and 11 lay the groundwork for us to complete the classification of cubics in Section 12. We introduce the complex numbers in Section 10 and prove the Fundamental Theorem of Algebra, which states that every polynomial in one variable factors completely over the complex numbers. We introduce points with complex coordinates in Section 11. The Fundamental Theorem of Algebra ensures that curves have "as many intersections as possible" over the complex numbers. This yields Bezout's Theorem, which states that the number of times that two curves without a common factor intersect in the complex projective plane is the product of their degrees.

We complete the classification of irreducible cubics in Section 12 by proving that every nonsingular, irreducible cubic C has a flex. The Hessian H of C is a cubic formed from the second partial derivatives of C. The points of intersection of C and H are the flexes of C. We use Bezout's Theorem from Section 11 to prove that C and H intersect exactly nine times, counting multiplicities, over the complex numbers. Because nine is odd, and because the intersections of C and H over the complex numbers are interchanged in pairs by conjugating their coordinates, it follows that C and H intersect at least once over the real numbers. Thus, every nonsingular, irreducible cubic C has a flex over the real numbers, as desired.

We end the chapter by asking how many points determine a cubic. We seek an analogue of Theorem 5.10, which says that a conic is uniquely determined by five points, no three collinear. Because the equation of the general cubic

$$\begin{aligned} &ax^3 + bx^2y + cxy^2 + dy^3 + ex^2 + fxy + gy^2 \\ &\quad + hx + iy + j = 0 \end{aligned}$$

has ten coefficients but can be multiplied by a nonzero constant, cubics have nine "degrees of freedom." This suggests that nine points generally lie on a unique cubic, and we see in Section 13 when they do.

History

Newton's classification of cubics in the late 1600s was the first great success of analytic geometry apart from its role in calculus. Newton claimed that the equation of every cubic in the Euclidean plane could be simplified to one of the forms

$$xy^2 + ey = ax^3 + bx^2 + cx + d, \tag{2}$$

$$xy = ax^3 + bx^2 + cx + d, \tag{3}$$

$$y^2 = ax^3 + bx^2 + cx + d, \tag{4}$$

$$y = ax^3 + bx^2 + cx + d, \tag{5}$$

by an appropriate choice of the coordinate axes, which were not required to be perpendicular in Newton's time. James Stirling published a proof of Newton's claim in 1717, possibly in collaboration with Newton. The key to Stirling's proof is to consider the family of chords of the cubic parallel to an asymptote, find the locus of the midpoints of the chords, and choose coordinate axes to simplify the equation of the locus.

Newton multiplied (2) by x and completed the square of the left-hand side to obtain the equation

$$(xy + \tfrac{1}{2}e)^2 = ax^4 + bx^3 + cx^2 + dx + \tfrac{1}{4}e^2. \tag{6}$$

By considering the roots of the right-hand sides of (2)–(6), he divided cubics into 72 species. Stirling identified four more species, and Jean-Paul de Gua de Malves found another two in 1740, giving a total of 78 species of cubics.

Newton also made the remarkable assertion that all cubics could be obtained from those in (4) by projecting between planes. This chapter centers around proving Newton's assertion for irreducible cubics, with the change that projections are replaced by their algebraic equivalent, transformations.

The first proofs of Newton's assertion appeared in 1731, due independently to Alexis Clairaut and Francois Nicole. Clairaut considered the graph of the equation

$$zy^2 = ax^3 + bx^2z + cxz^2 + dz^3 \tag{7}$$

in three-dimensional-Euclidean space. Equation (7) is homogeneous and yields (4) when we set $z = 1$. It follows that (7) describes a cubical cone having the origin as vertex; that is, the graph consists of the lines joining the origin to the cubic given by (4) in the plane $z = 1$. Clairaut showed that every cubic in (2)–(5) is the intersection of a plane and a cubical cone given by (7), which proves Newton's assertion.

Among the important attributes of cubics are flexes, which are generalizations of inflection points. Clairaut asserted in 1731 that an irreducible cubic has from one through three inflection points over the real numbers (Exercises 12.8 and 12.18). For irreducible cubics having three inflection points, de Gua proved in 1740 that the inflection points are collinear (Exercises 8.6 and 9.2(c)). Plücker argued in 1834 that a nonsingular cubic has nine flexes over the complex numbers that lie by threes on twelve lines (Exercise 12.24). His argument was completed in 1844 by Ludwig Hesse, who characterized flexes with a determinant of second partial derivatives that is now called a Hessian (Theorem 12.4).

Suppose that the cubic C in (1) has rational coefficients. If we take the tangent line through a point of C with rational coordinates, or if we take the secant line through two points of C with rational coordinates, the line intersects C at another point that has rational coordinates, as we discuss in Section 9. Applying this tangent–secant construction repeatedly can produce any number of points of C with rational coordinates from just one. This addresses a central problem of number theory, finding the rational solutions of equations. An ad hoc algebraic version of the tangent–secant construction was introduced by Diophantus, who lived in Alexandria during the third century A.D. Fermat systematized the construction algebraically, and Newton interpreted it geometrically in terms of tangents and secants.

Complex numbers were introduced in the 1500s to solve cubic equations in one variable. They were reintroduced in the 1700s to facilitate integration by partial fractions. Mathematicians gradually developed proficiency in working with complex numbers, and their confidence increased when Carl Friedrich Gauss gave four proofs of the Fundamental Theorem of Algebra in the first decade of the 1800s. Jean d'Alembert had given an incomplete proof of the Fundamental Theorem in 1746. The main gap in his proof was filled in 1806 when Jean Argand proved a result generally called "d'Alembert's Lemma" (Claim 5 of Section 10). In fact, no truly complete proof of the Fundamental Theorem could be given until the 1870s, when Georg Cantor and Richard Dedekind developed the real numbers formally and Karl Weierstrass derived the basic properties of continuous functions. The proof in Section 10 of the Fundamental Theorem is based on the paper of Charles Fefferman cited in the References, which modernizes and simplifies the work of d'Alembert and Argand.

The idea of a complex curve—an algebraic curve whose coefficients and variables are complex numbers—emerged over centuries. Analytic geometers from Newton onward considered "imaginary points" on curves without clearly specfying the nature of these points. In the 1820s, Jean Poncelet and Michel Chasles argued for using imaginary points systematically in synthetic projective geometry. In 1830, Plücker clarified the nature of imaginary points when the homogeneous coordinates he introduced made it possible to consider points with complex coordinates. Nevertheless, complex curves were not generally considered natural objects of study until Georg Riemann proposed in 1851 a way to consider them topologically: the "Riemann surface" of a polynomial equation $f(w, z) = 0$ consists of sheets that lie over the complex z-plane and correspond to the values of w determined by the equation. In the 1860s, Alfred Clebsch and Paul Gordan recast Riemann's ideas from complex analytic to geometric form, and the modern view of complex curves was established.

"Elliptic integrals" are, speaking roughly, integrals that involve the square root of a polynomial of degree 3 or 4. Unlike integrals that involve the square root of a polynomial of degree 2, elliptic integrals cannot generally be evaluated in closed form. Examples of elliptic integrals arose from scientific and geometric considerations in the last half of the 1600s and the first half of the 1700s. The first examples involved arc lengths of ellipses and led to the name "elliptic integrals." In the mid-1700s, Leonhard Euler revolutionized the study of elliptic integrals by establishing the identity

$$\int_a^{x_1} g(t)^{-1/2}\,dt + \int_a^{x_2} g(t)^{-1/2}\,dt = \int_a^{x_3} g(t)^{-1/2}\,dt \tag{8}$$

for any polynomial $g(t)$ of degree 3 of 4, where x_3 is a rational function of x_1, x_2, a, $g(x_1)^{1/2}$, $g(x_2)^{1/2}$, $g(a)^{1/2}$.

Certain cubics are now called "elliptic curves" because of their connection with elliptic integrals. This connection was discovered by Gauss, Niels Abel, and Carl Jacobi in the 1820s. Their results were clarified and extended by Riemann in the 1850s, Weierstrass in 1863, and Henri Poincaré in 1901. We summarize a small part of this work below.

Let

$$g(t) = 4t^3 + ct + d$$

be a polynomial of degree 3 without repeated roots. The Weierstrass P-function

$$x = P(u) \tag{9}$$

parametrizes the nonsingular, irreducible complex cubic

$$y^2 = g(x) \tag{10}$$

in the following sense: Equation (9) and the equation

$$y = P'(u) \tag{11}$$

match up the complex numbers u on and inside a parallelogram in the complex plane with the points (x, y) of the complex cubic (10), except that any two complex numbers u in corresponding positions on opposite sides of the parallelogram map to the same point (x, y). The function $P(u)$ can be written in the form

$$P(u) = \frac{1}{u^2} + a_2u^2 + a_4u^4 + \cdots$$

for complex numbers a_2, a_4, ...

Equations (9)–(11) imply that

$$\frac{dx}{du} = P'(u) = y = g(x)^{1/2},$$

and taking reciprocals gives

$$\frac{du}{dx} = g(x)^{-1/2}.$$

This implies that

$$u = \int g(x)^{-1/2}\,dx, \tag{12}$$

which means that u is a multivalued indefinite integral of the two-valued function $g(x)^{-1/2}$ of the complex variable x. In short, we obtain the Weierstrass P-function in (9) by inverting the elliptic integral in (12) and considering x as a function of u.

The idea of parametrizing the complex cubic in (10) by inverting the elliptic integral in (12) arose by drawing analogies with the following familiar facts: the unit circle

$$y^2 = 1 - x^2 \tag{13}$$

is parametrized by setting

$$x = \sin(u) \tag{14}$$

and

$$y = \sin'(u) = \cos(u), \tag{15}$$

where the relation given by (14) is the inverse of the relation

$$u = \arcsin(x) = \int_0^x (1 - t^2)^{-1/2}\,dt. \tag{16}$$

Drawing parallels between the cubic $g(t)$ and the quadratic $1 - t^2$ and between the Weierstrass P-function $x = P(u)$ and the sine function $x = \sin(u)$ creates analogies between (9) and (14), (10) and (13), (11) and (15), and (12) and (16).

In Section 9, we use secants and tangents to define addition of points on a nonsingular cubic given by (10). This method of adding points on the cubic corresponds via the Weierstrass P-function to addition of complex numbers. Specifically, for any complex numbers u_1 and u_2, the point of the complex cubic (10) that corresponds via (9) and (11) to the complex number $u_1 + u_2$ is the sum of the points on the complex cubic that correspond to u_1 and u_2. This is the geometric form of Euler's relation (8). It corresponds to the angle-addition formula for sines via the analogies in the previous paragraph.

The discussion accompanying (9)–(11) shows that the points of a nonsingular complex cubic correspond to the points of a parallelogram whose two pairs of opposite sides are glued together. Gluing together one pair of opposite sides of a parallelogram gives a cylinder. Gluing together the opposite ends of the cylinder gives a torus—the surface of a doughnut. Thus, a nonsingular complex cubic is topologically equivalent to a torus; that is, it can be continuously bent into the surface of a doughnut.

A nonsingular curve $f(x, y) = 0$ over the real numbers can be divided into pieces that are each parametrized by x or y. For example, the unit circle $x^2 + y^2 = 1$ can be divided into the upper and lower half-circles

$$y = (1 - x^2)^{1/2} \qquad \text{and} \qquad y = -(1 - x^2)^{1/2},$$

which are each parametrized by x. The analogous result holds over the complex numbers. In this sense, we can think of complex curves as "one-dimensional over the complex numbers." On the other hand, the complex numbers are themselves two-dimensional over the reals. That explains why a nonsingular complex cubic is topologically equivalent to a two-dimensional surface, the torus. In this text, we always think of complex curves as "one-dimensional over the complex numbers." We work with complex curves algebraically just like real curves.

Colin Maclaurin raised the following issue in 1720. On the one hand, requiring a curve to contain a particular point imposes a linear condition on the coefficients of the curve. A general curve of degree n has $\binom{n+2}{2}$ coefficients. Because we can multiply the coefficients by a nonzero number without changing the curve, a curve of degree n has

$$\binom{n+2}{2} - 1 = \frac{(n+2)(n+1)}{2} - 1 = \frac{n(n+3)}{2}$$

"degrees of freedom." Accordingly, we expect that a curve of degree n is uniquely determined by $n(n+3)/2$ of its points. On the other hand, Bezout's Theorem shows that two complex curves of degree n without multiple intersections intersect at n^2 points. The last two sentences appear to conflict because

$$n^2 \geq n(n+3)/2$$

for $n \geq 3$.

This apparent difficulty was explored by Leonhard Euler in 1748 and by Gabriel Cramer in 1750, and it is now known as "Cramer's paradox." Euler and Cramer suggested that there may be redundancies among the conditions that points impose on curves. We examine this idea for cubics in Section 13. Taking $n = 3$ in the first part of the previous paragraph shows that a cubic is uniquely determined by $n(n+3)/2 = 9$ conditions, provided that the conditions are not redundant. In fact, the $n^2 = 9$ points where two cubics intersect impose redundant conditions—otherwise, the points would not lie on two cubics. If two cubics intersect in nine points, any cubic through eight of the points necessarily contains the ninth, as Exercise 13.4 shows. If nine points are to determine a unique cubic, we show in Section 13 that we can choose eight of the points quite generally, but we must ensure that the ninth point does not lie on every cubic through the first eight.

3 -# # - " NG ' -) % ! ! NG K 7
" !

26.02.2020

an algebraic curve whose *
coefficients and variables
are complex numbers.

Newton → "imaginary points"

In the 1820s,
Jean Poncelet and
Michel Chasles
argued for using imaginary
points systematically in
Synthetic projective geometry.

Plücker (1830) →
homogeneous coordinates,
he introduced made it possible
to consider points with
complex coordinates.

118) 20200228
Números de Perrin
pt.wikipedia.org/wiki/Números_de_Perrin

data 28.02.2020
S T Q Q [✓ S] S D

Número Plástico ψ

é uma constante matemática.
É a solução real inequívoca da equação cúbica
(cúbica)

$$\underbrace{x^3 - x - 1 = 0}$$

valor aproximado
1,324718

função geratriz

$$G(P(n); x) = \frac{3 - x^2}{1 - x^2 - x^3}$$

Source:

wikipedia.org
mathworld.wolfram.com/PerrinSequence.html
www.universite-paris-saclay.fr/~perrin
Daniel.Perrin@math.u-psud.fr

date 28 02 2020

www.imo.universite-paris-saclay.fr/
~perrin/Livre_de_geometrie_projective.html

Géométrie Projective

Pappus

www.math2.fr/geometrie-projective/

math2.fr/2019/10/06/geometrie-projective/

/2019/12/18/geometrie-projective-exemple-de-calculs2/

others:

Data_NOVOVIVO_41_2020/
Books_41_2020/
Pappus/
Enseignement(hors CAPES)/

a) polysite.pdf
↳ Introduction générale et rappels

b) Pappus/TER/constructions.pdf
Constructions à la règle et au compas.

28.02.2020

c) Pappus/SEVRES/polyedres_pdf
Les polyèdres - Perrin

119) 20200302

Books:

a) Data_NOVOVIVO_44_2020/Books_44_2020/
Math_Books/Multivariable_Calculus/
Multivariable_Calculus_(formatted_
for_printing).pdf

Pág. 38(50)

2.6 The geometry of space
curves: Frenet vectors

b) .../Books_44_2020/Math_Books/
Zakon-analysisII-a4-one.pdf

Pág. 97(109) → Chapter 7

Volume and Measure

c) .../Math_Books/
Nicholson-OpenLAWA-2019A.pdf

Pág. 415(437)

8. Orthogonality

8.1 Orthogonal Complements and
Projections.

/+01/ / /2 / (. , 34 4 5 ,
&$* 6 $ 6"

6 ## , 6 7 #86 #. , 9" 9 96 5 ,

6 ## # 6 5 , #

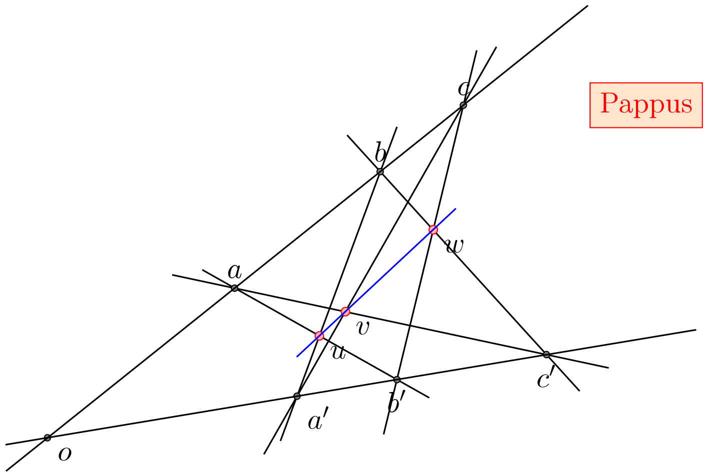
Pappus
c
b
w
a
v
u
c′
b′
a′
o

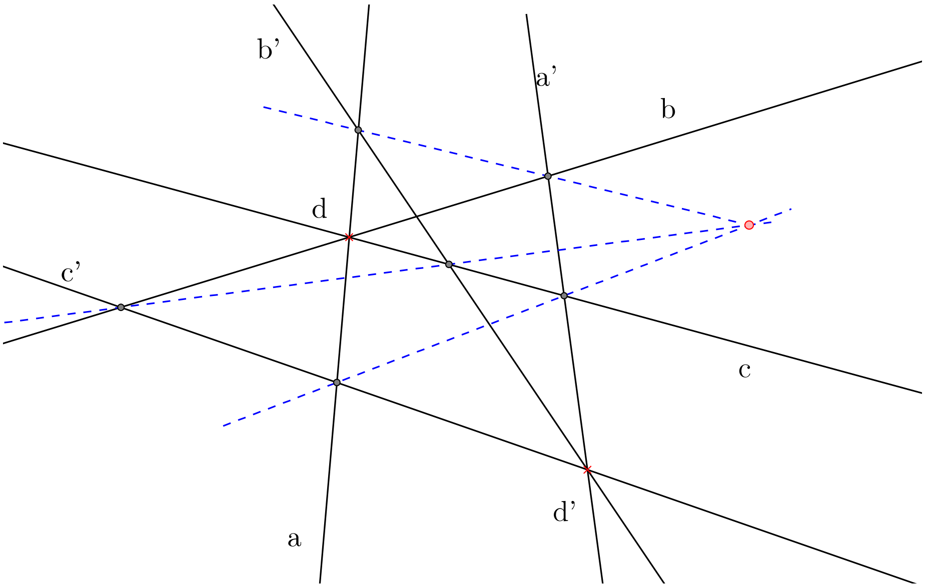

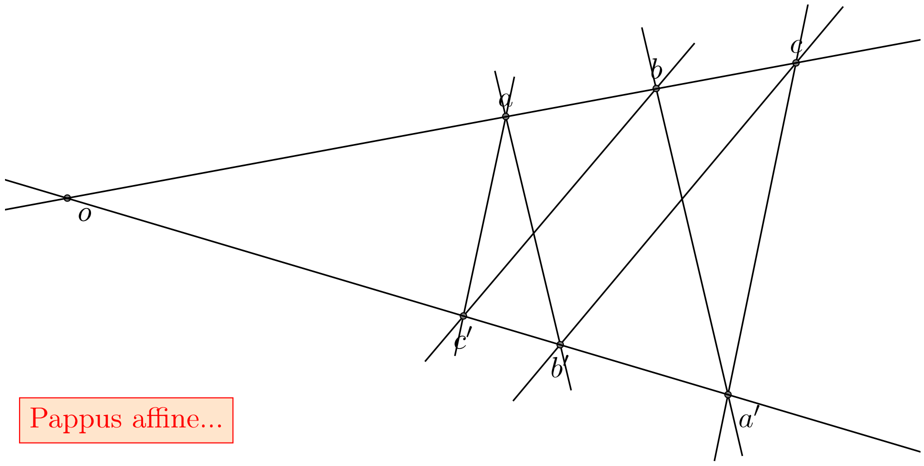
c
b
a
o
c′
b′
a′
Pappus affine...

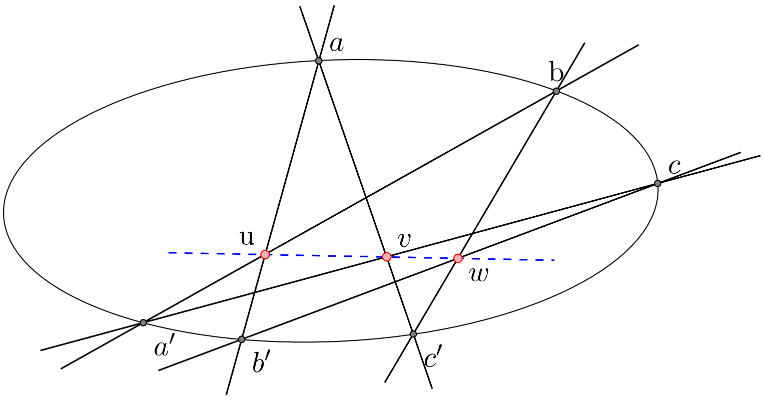

...Et la version coniques

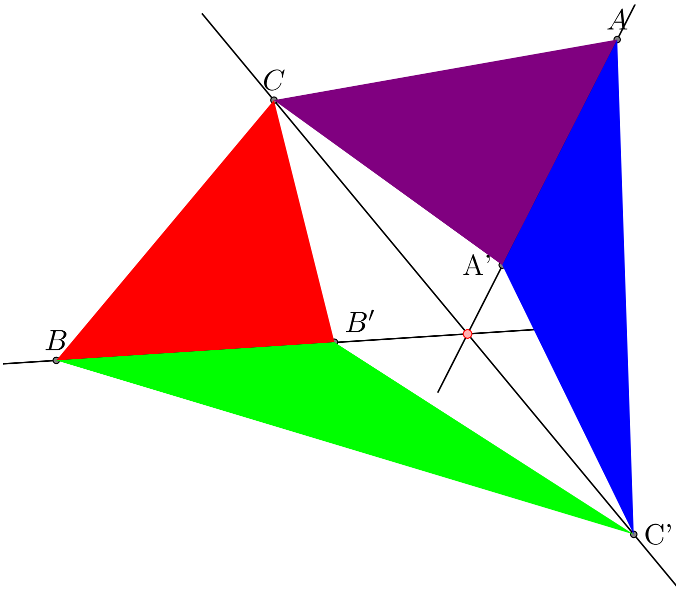

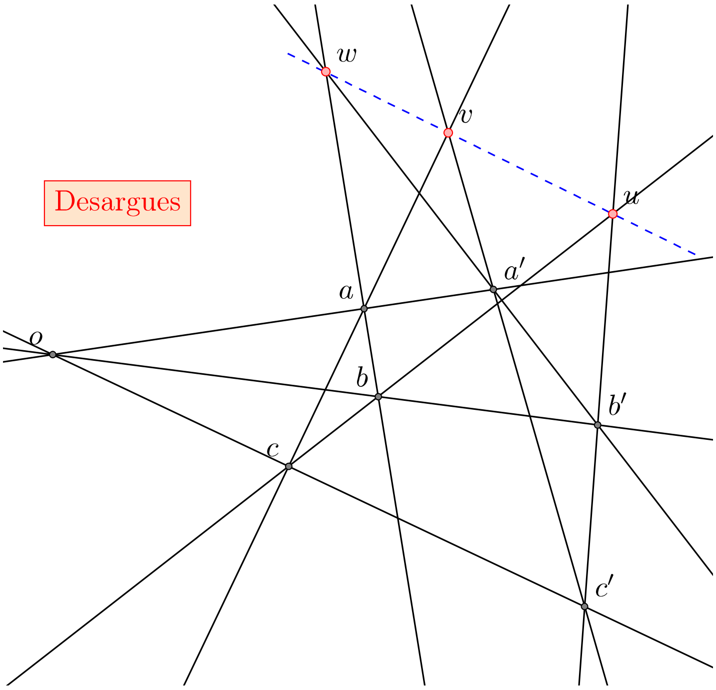
Desargues
w
v
u
a
a′
o
b
b′
c
c′

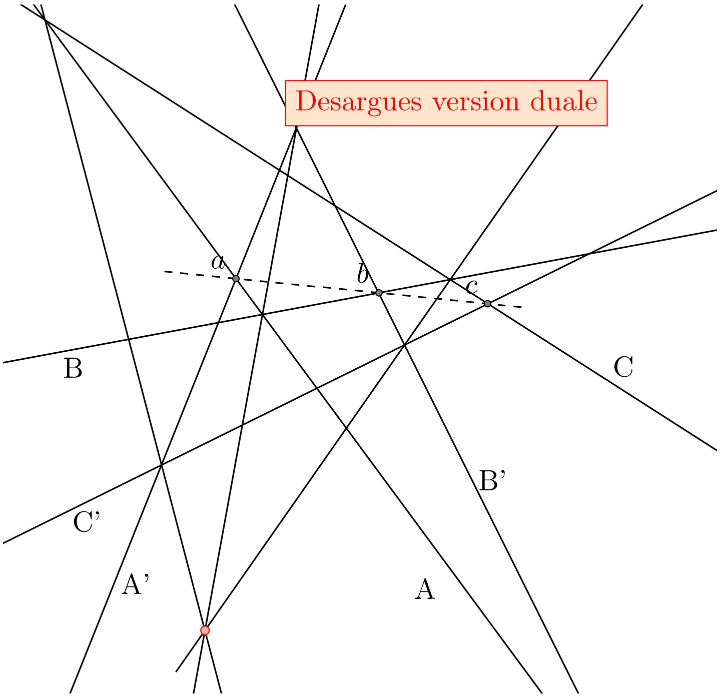
Desargues version duale
a
b
c
B
C
B’
C’
A’
A

Desargues sur les pinceaux de coniques

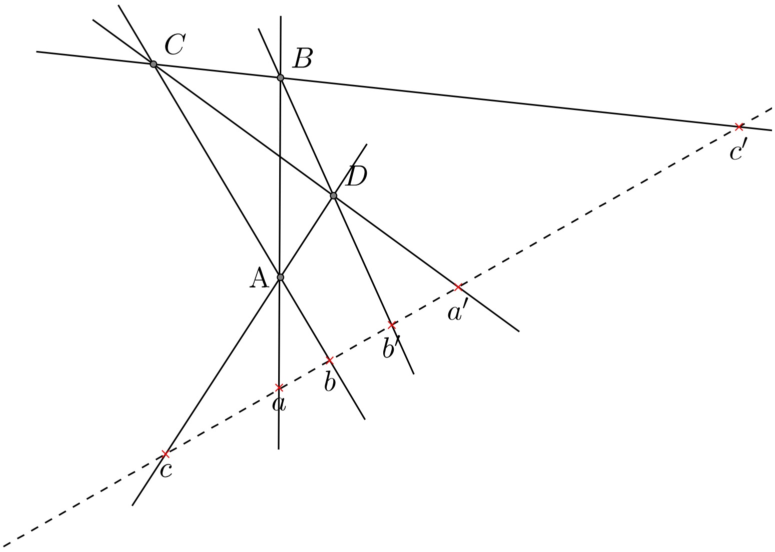
C
B
D
A
c′
a′
b′
b
a
c

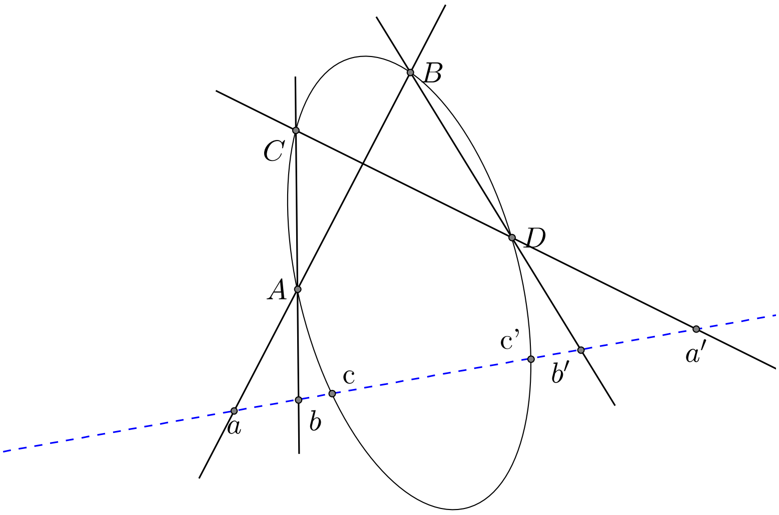
B
C
D
A
c'
b′
a′
c
b
a

polaire de a par rapport aux droites D et D'

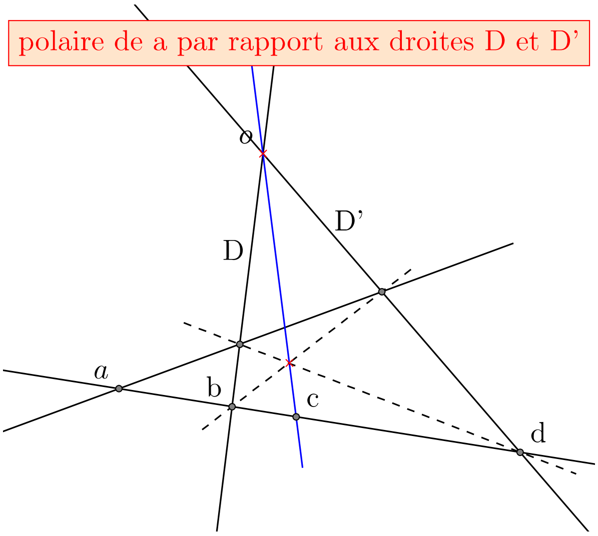

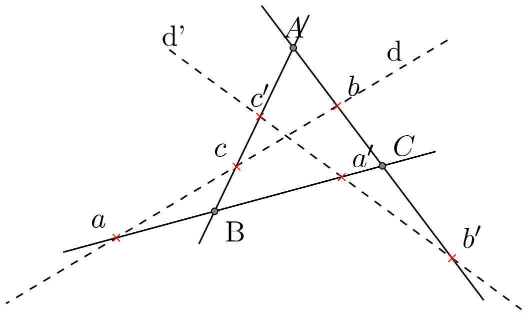
d’
A
d
c′
b
c
C
a′
a
B
b′

Lemme des trois birapports

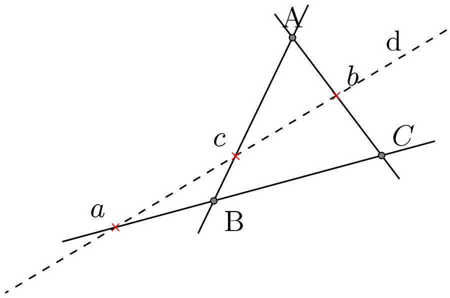

Ménélaüs

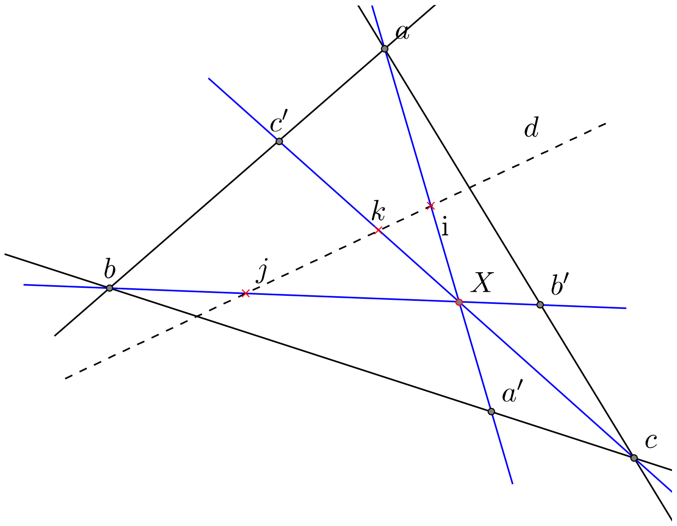
a
c′
d
k
i
b
j
X
b′
a′
c

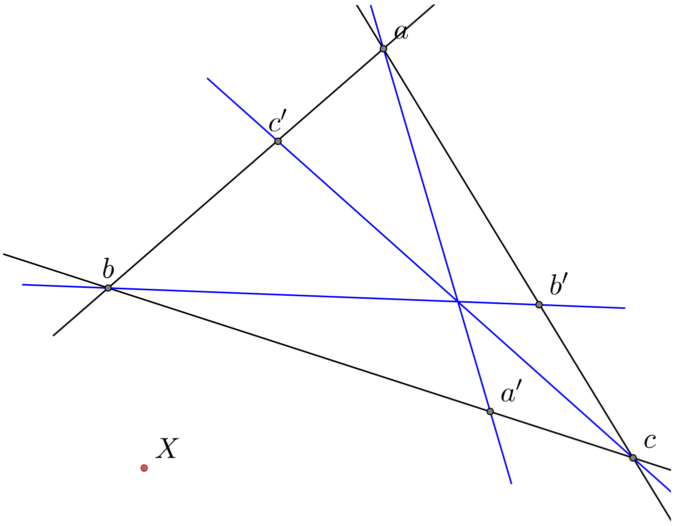
a
c′
b
b′
a′
c
X

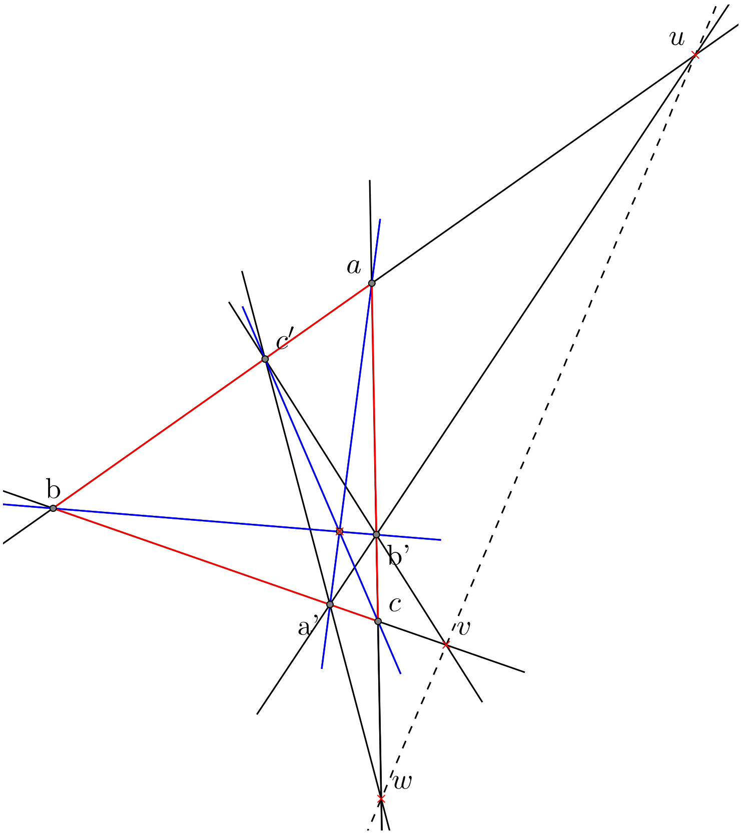
u
a
c′
b
b’
a’
c
v
w

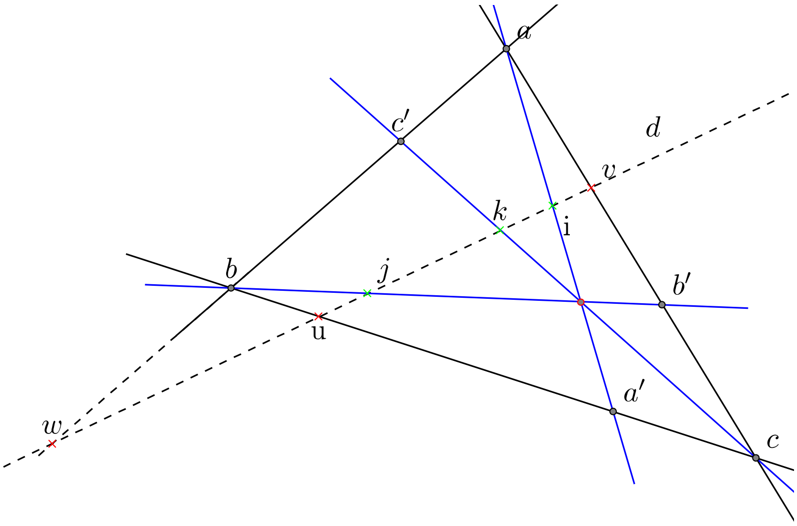

a
c′
d
v
k
i
j
b
b′
u
a′
w
c

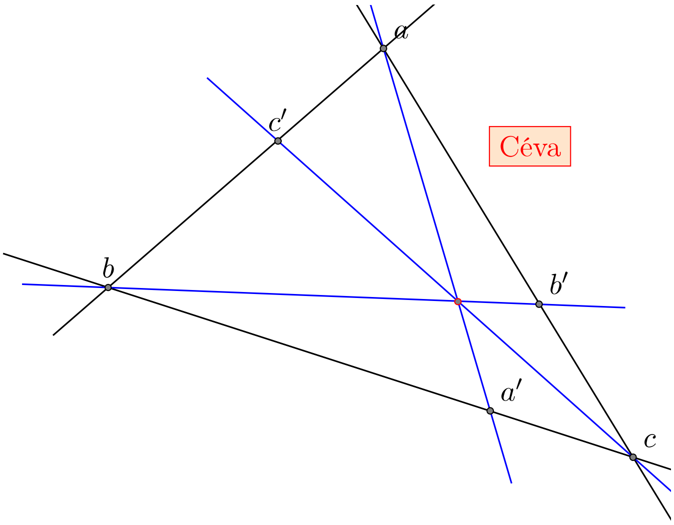

a
c′
Céva
b
b′
a′
c

Le théorème de Newton projectif

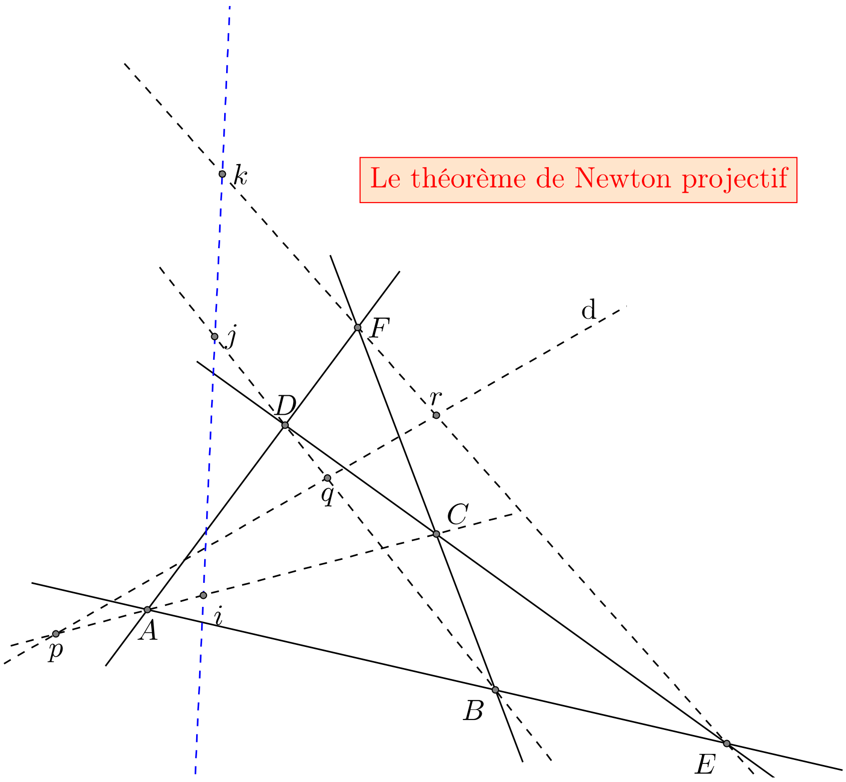
k
j
F
d
D
r
q
C
i
A
p
B
E

Le théorème de Newton affine

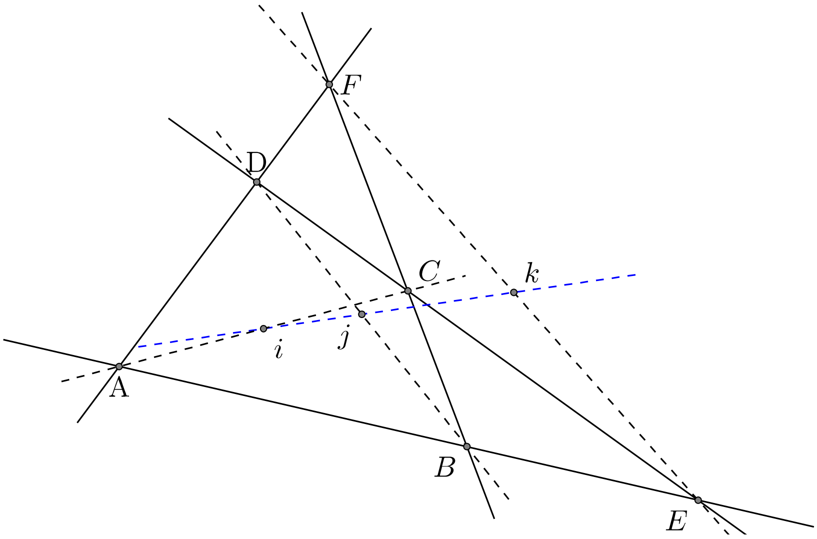

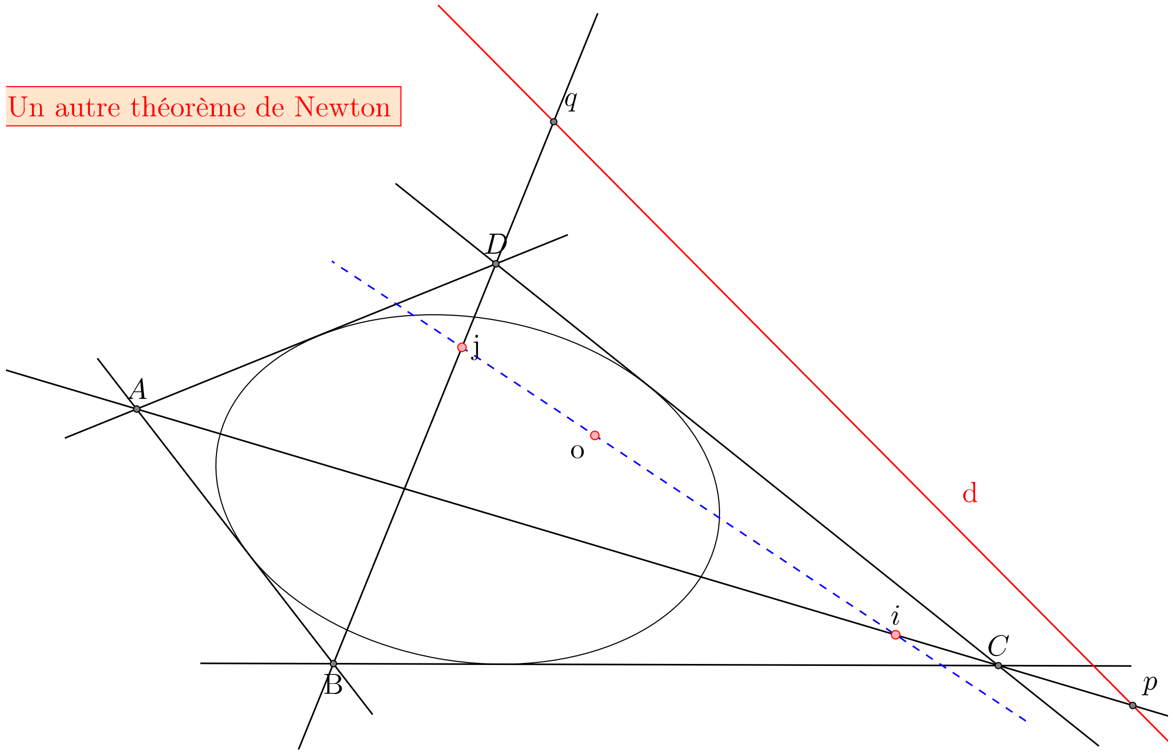
Un autre théorème de Newton
q
D
j
A
o
d
i
C
p
B

/+01/ / /2 / (34 4 6 5 ,
+:;< +:'1=< 6

> 6 ## # $# # (# 6 5 , ? 6 " #

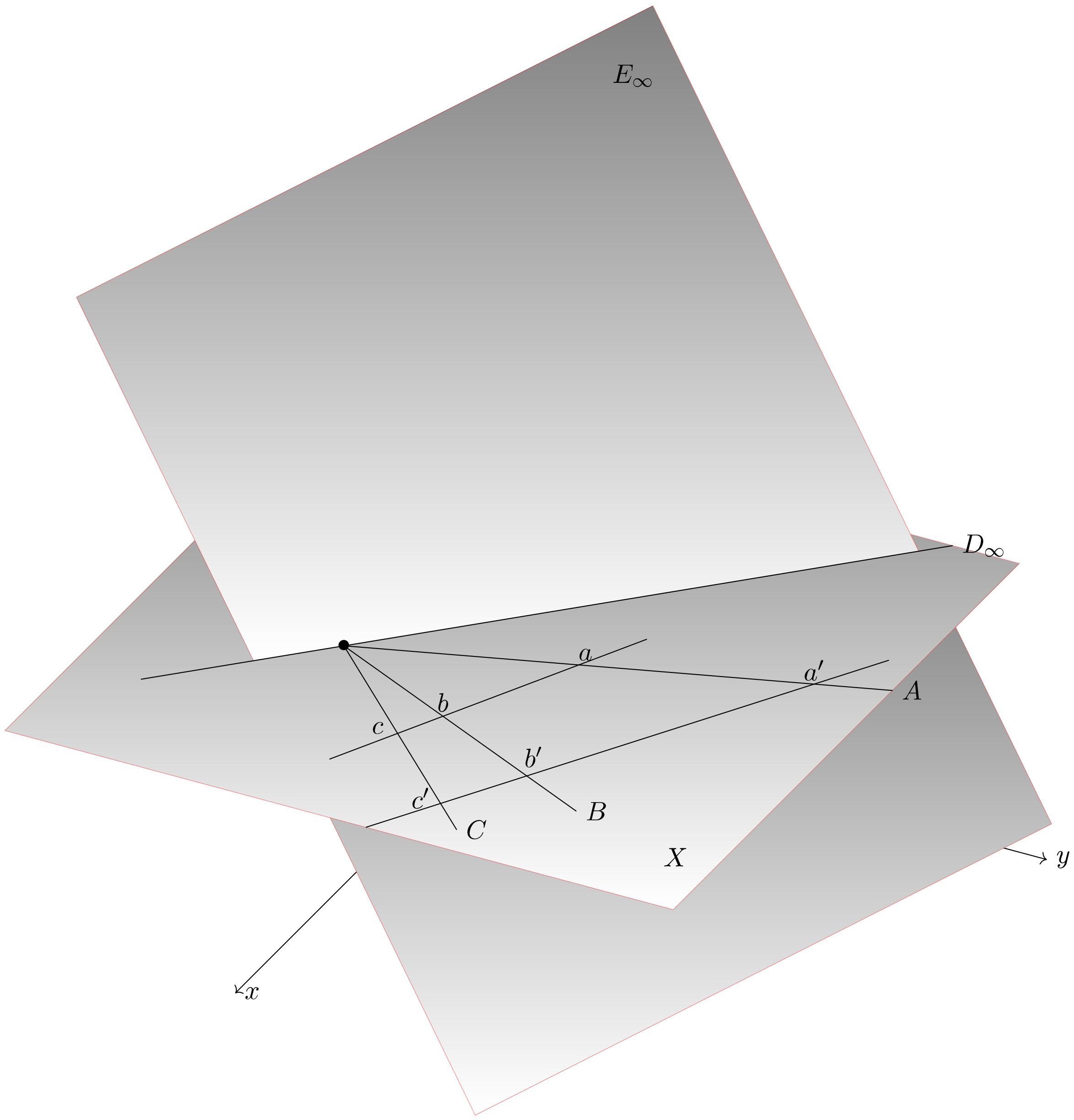
E_∞
D_∞
a
a'
A
b
c
b'
c'
B
C
X
y
x

/+01/ / /2 / (4 4 6 5 , @ ? 6 "
+:;< +:'1=< 6

2

/+01/ / /2 / (> 6 " 6 "
+:;< +:'1=< 6

6 ## # $# # !# 6 5 , #

/+01/ / /2 / (34 4 6 5 ,
+:;< +:'1=< 6

2

/+01/ / /2 / (6 7 6"

> /+01/ / /2 / (6"

/+01/ / /2 / (6 7 " 6"

' $

/+01/ / /2 / $
, > 9 9 "9 96 6"

) &(

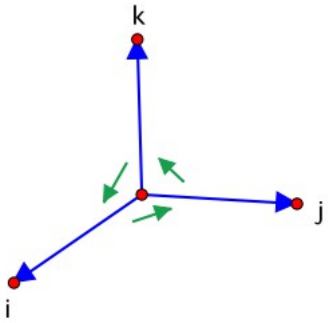

Figure 2.15: The circle of life: $\mathbf{i} \times \mathbf{j} = \mathbf{k}, \quad \mathbf{j} \times \mathbf{k} = \mathbf{i}, \quad \mathbf{k} \times \mathbf{i} = \mathbf{j}$

2.6 The geometry of space curves: Frenet vectors

We take up the following question: *How can one tell when two curves in $\mathbb{R}^3$ are congruent, that is, either curve can be translated, rotated and/or reflected so that it fits exactly on top of the other?* As mentioned earlier, the main idea is to construct a basis of orthogonal unit vectors that moves along the curve and then to study how the basis vectors change during the motion.

So let C be a curve in $\mathbb{R}^3$, parametrized by a differentiable path $\alpha\colon I \to \mathbb{R}^3$, where I is an interval in $\mathbb{R}$. In Section 2.4, we defined two of the vectors in our moving basis:

- the unit tangent vector $\mathbf{T}(t) = \frac{1}{\|\alpha'(t)\|}\alpha'(t) = \frac{1}{v(t)}\mathbf{v}(t)$ and
- the principal normal vector $\mathbf{N}(t) = \frac{1}{\|\mathbf{T}'(t)\|}\mathbf{T}'(t)$.

These definitions make sense only if $\alpha'(t) \neq \mathbf{0}$ and $\mathbf{T}'(t) \neq \mathbf{0}$, so we assume that this is the case from now on. Both $\mathbf{T}(t)$ and $\mathbf{N}(t)$ are constructed to be unit vectors, and we showed before that $\mathbf{T}(t) \cdot \mathbf{N}(t) = 0$. At this point, the third basis vector is easily defined.

Definition. The cross product $\mathbf{B}(t) = \mathbf{T}(t) \times \mathbf{N}(t)$ is called the **binormal vector**.

fingers of your right hand from $\mathbf{i}$ to $\mathbf{j}$, your thumb points in the direction of $\mathbf{k}$. This is not so much a fact as it is an agreement: whenever we draw $\mathbb{R}^3$, we agree to orient the positive x, y, and z-axes so that the right-hand rule for $(\mathbf{i}, \mathbf{j}, \mathbf{k})$ is satisfied, as in Figure 2.15. As we continue with further material, figures in $\mathbb{R}^3$ become increasingly prominent, so it is probably good to state this convention explicitly.

More generally, given a basis $(\mathbf{v}, \mathbf{w}, \mathbf{z})$ of $\mathbb{R}^3$, the plane $\mathcal{P}$ spanned by $\mathbf{v}$ and $\mathbf{w}$ separates $\mathbb{R}^3$ into two half-spaces, or "sides." Once we adopt the right-hand rule for $(\mathbf{i}, \mathbf{j}, \mathbf{k})$, it turns out that $(\mathbf{v}, \mathbf{w}, \mathbf{z})$ is right-handed in the sense of the definition if and only if, when you rotate the fingers of your right hand from $\mathbf{v}$ to $\mathbf{w}$, your thumb lies on the same side of $\mathcal{P}$ as $\mathbf{z}$. In particular, by Proposition 2.19, your thumb lies on the same side as $\mathbf{v} \times \mathbf{w}$. In this way, your right thumb gives the direction of $\mathbf{v} \times \mathbf{w}$.

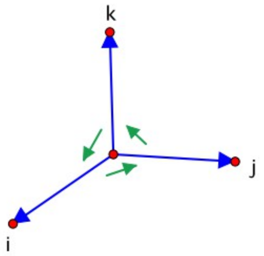

Figure 2.15: The circle of life: $\mathbf{i} \times \mathbf{j} = \mathbf{k}$, $\mathbf{j} \times \mathbf{k} = \mathbf{i}$, $\mathbf{k} \times \mathbf{i} = \mathbf{j}$

Proving the connection between the sign of a determinant and the right-hand rule would be too great a digression. One approach uses the continuity of the determinant and the geometry of rotations to show that, if the rule works for $(\mathbf{i}, \mathbf{j}, \mathbf{k})$, then it works in general. In any case, the right-hand rule is sometimes a convenient way to find cross products geometrically. For instance, in addition to $\mathbf{i} \times \mathbf{j} = \mathbf{k}$, it allows us to see that $\mathbf{j} \times \mathbf{k} = \mathbf{i}$, not $-\mathbf{i}$, and that $\mathbf{k} \times \mathbf{i} = \mathbf{j}$, as pictured in Figure 2.15.

2.6 The geometry of space curves: Frenet vectors

We take up the following question: *How can one tell when two curves in $\mathbb{R}^3$ are congruent, that is, either curve can be translated, rotated and/or reflected so that it fits exactly on top of the other?* As mentioned earlier, the main idea is to construct a basis of orthogonal unit vectors that moves along the curve and then to study how the basis vectors change during the motion.

So let C be a curve in $\mathbb{R}^3$, parametrized by a differentiable path $\alpha \colon I \to \mathbb{R}^3$, where I is an interval in $\mathbb{R}$. In Section 2.4, we defined two of the vectors in our moving basis:

- the unit tangent vector $\mathbf{T}(t) = \frac{1}{\|\alpha'(t)\|}\alpha'(t) = \frac{1}{v(t)}\mathbf{v}(t)$ and
- the principal normal vector $\mathbf{N}(t) = \frac{1}{\|\mathbf{T}'(t)\|}\mathbf{T}'(t)$.

These definitions make sense only if $\alpha'(t) \neq \mathbf{0}$ and $\mathbf{T}'(t) \neq \mathbf{0}$, so we assume that this is the case from now on. Both $\mathbf{T}(t)$ and $\mathbf{N}(t)$ are constructed to be unit vectors, and we showed before that $\mathbf{T}(t) \cdot \mathbf{N}(t) = 0$. At this point, the third basis vector is easily defined.

Definition. The cross product $\mathbf{B}(t) = \mathbf{T}(t) \times \mathbf{N}(t)$ is called the **binormal vector**.

As a cross product, $\mathbf{B}(t)$ is orthogonal to $\mathbf{T}(t)$ and $\mathbf{N}(t)$. Moreover, its length is $\|\mathbf{B}(t)\| = \|\mathbf{T}(t)\|\,\|\mathbf{N}(t)\|\,\sin\theta = 1\cdot 1\cdot \sin\frac{\pi}{2} = 1$. Thus $(\mathbf{T}(t), \mathbf{N}(t), \mathbf{B}(t))$ is a set of orthogonal unit vectors. The vectors are called the **Frenet vectors** of α and are illustrated in Figure 2.16.

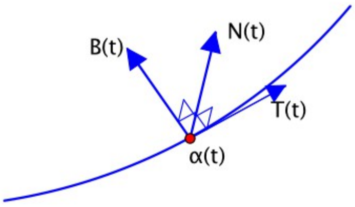

Figure 2.16: The Frenet vectors $\mathbf{T}(t)$, $\mathbf{N}(t)$, $\mathbf{B}(t)$

Example 2.20 (The Frenet vectors of a helix)**.** We shall use the helix

$$\alpha(t) = (a\cos t, a\sin t, bt)$$

as a running example to illustrate new concepts as they arise. Here, a and b are constant and $a > 0$. Recall that the helix is called right-handed if $b > 0$ and left-handed if $b < 0$. If $b = 0$, the curve collapses to a circle of radius a in the xy-plane.

To find the Frenet vectors $\mathbf{T}(t)$, $\mathbf{N}(t)$, and $\mathbf{B}(t)$ of the helix, we simply and carefully follow the definitions.
Unit tangent: $\alpha'(t) = (-a\sin t, a\cos t, b)$, so $\|\alpha'(t)\| = \sqrt{a^2\sin^2 t + a^2\cos^2 t + b^2} = \sqrt{a^2+b^2}$. Thus:

$$\mathbf{T}(t) = \frac{1}{\|\alpha'(t)\|}\alpha'(t) = \frac{1}{\sqrt{a^2+b^2}}(-a\sin t, a\cos t, b).$$

Principal normal: From the last calculation, $\mathbf{T}'(t) = \frac{1}{\sqrt{a^2+b^2}}(-a\cos t, -a\sin t, 0)$, so $\|\mathbf{T}'(t)\| = \frac{1}{\sqrt{a^2+b^2}}\sqrt{a^2\cos^2 t + a^2\sin^2 t + 0^2} = \frac{a}{\sqrt{a^2+b^2}}$. Hence:

$$\begin{aligned}\mathbf{N}(t) = \frac{1}{\|\mathbf{T}'(t)\|}\mathbf{T}'(t) &= \frac{1}{\frac{a}{\sqrt{a^2+b^2}}}\cdot\frac{1}{\sqrt{a^2+b^2}}(-a\cos t, -a\sin t, 0)\\ &= \frac{1}{a}(-a\cos t, -a\sin t, 0)\\ &= (-\cos t, -\sin t, 0).\end{aligned}$$

Binormal: Lastly:

$$\begin{aligned}\mathbf{B}(t) = \mathbf{T}(t)\times\mathbf{N}(t) &= \det\begin{bmatrix}\mathbf{i} & \mathbf{j} & \mathbf{k}\\ -\frac{a\sin t}{\sqrt{a^2+b^2}} & \frac{a\cos t}{\sqrt{a^2+b^2}} & \frac{b}{\sqrt{a^2+b^2}}\\ -\cos t & -\sin t & 0\end{bmatrix}\\ &= \mathbf{i}\Big(\frac{b\sin t}{\sqrt{a^2+b^2}}\Big) - \mathbf{j}\Big(\frac{b\cos t}{\sqrt{a^2+b^2}}\Big) + \mathbf{k}\Big(\frac{a\sin^2 t}{\sqrt{a^2+b^2}} + \frac{a\cos^2 t}{\sqrt{a^2+b^2}}\Big)\\ &= \frac{1}{\sqrt{a^2+b^2}}\big(b\sin t, -b\cos t, a\big).\end{aligned}$$

□

> /+01/ / /2 / $ A B 7 ' % 6"

/+01/ / /2 / $ A B 7 " % 6"

) $* $ 6 * "

/+01/ / /2 / $ C 26 .+D+ $+ 6"

) % %&*

8. Orthogonality

In Section 5.3 we introduced the dot product in $\mathbb{R}^n$ and extended the basic geometric notions of length and distance. A set $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m\}$ of nonzero vectors in $\mathbb{R}^n$ was called an **orthogonal set** if $\mathbf{f}_i \cdot \mathbf{f}_j = 0$ for all $i \neq j$, and it was proved that every orthogonal set is independent. In particular, it was observed that the expansion of a vector as a linear combination of orthogonal basis vectors is easy to obtain because formulas exist for the coefficients. Hence the orthogonal bases are the "nice" bases, and much of this chapter is devoted to extending results about bases to orthogonal bases. This leads to some very powerful methods and theorems. Our first task is to show that every subspace of $\mathbb{R}^n$ *has* an orthogonal basis.

8. Orthogonality

In Section 5.3 we introduced the dot product in $\mathbb{R}^n$ and extended the basic geometric notions of length and distance. A set $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m\}$ of nonzero vectors in $\mathbb{R}^n$ was called an **orthogonal set** if $\mathbf{f}_i \cdot \mathbf{f}_j = 0$ for all $i \neq j$, and it was proved that every orthogonal set is independent. In particular, it was observed that the expansion of a vector as a linear combination of orthogonal basis vectors is easy to obtain because formulas exist for the coefficients. Hence the orthogonal bases are the "nice" bases, and much of this chapter is devoted to extending results about bases to orthogonal bases. This leads to some very powerful methods and theorems. Our first task is to show that every subspace of $\mathbb{R}^n$ *has* an orthogonal basis.

8.1 Orthogonal Complements and Projections

If $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$ is linearly independent in a general vector space, and if $\mathbf{v}_{m+1}$ is not in $\text{span}\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$, then $\{\mathbf{v}_1, \dots, \mathbf{v}_m, \mathbf{v}_{m+1}\}$ is independent (Lemma 6.4.1). Here is the analog for *orthogonal* sets in $\mathbb{R}^n$.

Lemma 8.1.1: Orthogonal Lemma

Let $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m\}$ be an orthogonal set in $\mathbb{R}^n$. Given $\mathbf{x}$ in $\mathbb{R}^n$, write

$$\mathbf{f}_{m+1} = \mathbf{x} - \frac{\mathbf{x}\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 - \frac{\mathbf{x}\cdot\mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 - \cdots - \frac{\mathbf{x}\cdot\mathbf{f}_m}{\|\mathbf{f}_m\|^2}\mathbf{f}_m$$

Then:

1. $\mathbf{f}_{m+1} \cdot \mathbf{f}_k = 0$ *for* $k = 1, 2, \dots, m$.
2. *If* $\mathbf{x}$ *is not in* $\text{span}\{\mathbf{f}_1, \dots, \mathbf{f}_m\}$, *then* $\mathbf{f}_{m+1} \neq \mathbf{0}$ *and* $\{\mathbf{f}_1, \dots, \mathbf{f}_m, \mathbf{f}_{m+1}\}$ *is an orthogonal set.*

Proof. For convenience, write $t_i = (\mathbf{x}\cdot\mathbf{f}_i)/\|\mathbf{f}_i\|^2$ for each i. Given $1 \leq k \leq m$:

$$\begin{aligned}
\mathbf{f}_{m+1}\cdot\mathbf{f}_k &= (\mathbf{x} - t_1\mathbf{f}_1 - \cdots - t_k\mathbf{f}_k - \cdots - t_m\mathbf{f}_m)\cdot\mathbf{f}_k \\
&= \mathbf{x}\cdot\mathbf{f}_k - t_1(\mathbf{f}_1\cdot\mathbf{f}_k) - \cdots - t_k(\mathbf{f}_k\cdot\mathbf{f}_k) - \cdots - t_m(\mathbf{f}_m\cdot\mathbf{f}_k) \\
&= \mathbf{x}\cdot\mathbf{f}_k - t_k\|\mathbf{f}_k\|^2 \\
&= 0
\end{aligned}$$

This proves (1), and (2) follows because $\mathbf{f}_{m+1} \neq \mathbf{0}$ if $\mathbf{x}$ is not in $\text{span}\{\mathbf{f}_1, \dots, \mathbf{f}_m\}$. □

The orthogonal lemma has three important consequences for $\mathbb{R}^n$. The first is an extension for orthogonal sets of the fundamental fact that any independent set is part of a basis (Theorem 6.4.1).

Theorem 8.1.1

Let U be a subspace of $\mathbb{R}^n$.

1. *Every orthogonal subset $\{\mathbf{f}_1, \dots, \mathbf{f}_m\}$ in U is a subset of an orthogonal basis of U.*
2. *U has an orthogonal basis.*

Proof.

1. If span $\{\mathbf{f}_1, \dots, \mathbf{f}_m\} = U$, it is *already* a basis. Otherwise, there exists $\mathbf{x}$ in U outside span $\{\mathbf{f}_1, \dots, \mathbf{f}_m\}$. If $\mathbf{f}_{m+1}$ is as given in the orthogonal lemma, then $\mathbf{f}_{m+1}$ is in U and $\{\mathbf{f}_1, \dots, \mathbf{f}_m, \mathbf{f}_{m+1}\}$ is orthogonal. If span $\{\mathbf{f}_1, \dots, \mathbf{f}_m, \mathbf{f}_{m+1}\} = U$, we are done. Otherwise, the process continues to create larger and larger orthogonal subsets of U. They are all independent by Theorem 5.3.5, so we have a basis when we reach a subset containing dim U vectors.
2. If $U = \{\mathbf{0}\}$, the empty basis is orthogonal. Otherwise, if $\mathbf{f} \neq \mathbf{0}$ is in U, then $\{\mathbf{f}\}$ is orthogonal, so (2) follows from (1). □

We can improve upon (2) of Theorem 8.1.1. In fact, the second consequence of the orthogonal lemma is a procedure by which *any* basis $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ of a subspace U of $\mathbb{R}^n$ can be systematically modified to yield an orthogonal basis $\{\mathbf{f}_1, \dots, \mathbf{f}_m\}$ of U. The $\mathbf{f}_i$ are constructed one at a time from the $\mathbf{x}_i$.

To start the process, take $\mathbf{f}_1 = \mathbf{x}_1$. Then $\mathbf{x}_2$ is not in span $\{\mathbf{f}_1\}$ because $\{\mathbf{x}_1, \mathbf{x}_2\}$ is independent, so take

$$\mathbf{f}_2 = \mathbf{x}_2 - \frac{\mathbf{x}_2 \cdot \mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1$$

Thus $\{\mathbf{f}_1, \mathbf{f}_2\}$ is orthogonal by Lemma 8.1.1. Moreover, span $\{\mathbf{f}_1, \mathbf{f}_2\}$ = span $\{\mathbf{x}_1, \mathbf{x}_2\}$ (verify), so $\mathbf{x}_3$ is not in span $\{\mathbf{f}_1, \mathbf{f}_2\}$. Hence $\{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3\}$ is orthogonal where

$$\mathbf{f}_3 = \mathbf{x}_3 - \frac{\mathbf{x}_3 \cdot \mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 - \frac{\mathbf{x}_3 \cdot \mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2$$

Again, span $\{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3\}$ = span $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, so $\mathbf{x}_4$ is not in span $\{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3\}$ and the process continues. At the mth iteration we construct an orthogonal set $\{\mathbf{f}_1, \dots, \mathbf{f}_m\}$ such that

$$\text{span}\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m\} = \text{span}\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\} = U$$

Hence $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m\}$ is the desired orthogonal basis of U. The procedure can be summarized as follows.

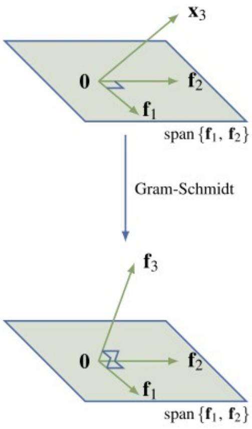

Theorem 8.1.2: Gram-Schmidt Orthogonalization Algorithm[1]

If $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ *is any basis of a subspace* U *of* $\mathbb{R}^n$*, construct* $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m$ *in* U *successively as follows:*

$$\begin{aligned}
\mathbf{f}_1 &= \mathbf{x}_1 \\
\mathbf{f}_2 &= \mathbf{x}_2 - \frac{\mathbf{x}_2 \cdot \mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 \\
\mathbf{f}_3 &= \mathbf{x}_3 - \frac{\mathbf{x}_3 \cdot \mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 - \frac{\mathbf{x}_3 \cdot \mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 \\
&\vdots \\
\mathbf{f}_k &= \mathbf{x}_k - \frac{\mathbf{x}_k \cdot \mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 - \frac{\mathbf{x}_k \cdot \mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 - \cdots - \frac{\mathbf{x}_k \cdot \mathbf{f}_{k-1}}{\|\mathbf{f}_{k-1}\|^2}\mathbf{f}_{k-1}
\end{aligned}$$

for each $k = 2, 3, \dots, m$*. Then*

1. $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m\}$ *is an orthogonal basis of* U*.*
2. $\operatorname{span}\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_k\} = \operatorname{span}\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ *for each* $k = 1, 2, \dots, m$*.*

The process (for $k = 3$) is depicted in the diagrams. Of course, the algorithm converts any basis of $\mathbb{R}^n$ itself into an orthogonal basis.

Example 8.1.1

Find an orthogonal basis of the row space of $A = \left[\begin{array}{rrrr} 1 & 1 & -1 & -1 \\ 3 & 2 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{array}\right]$.

Solution. Let $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ denote the rows of A and observe that $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ is linearly independent. Take $\mathbf{f}_1 = \mathbf{x}_1$. The algorithm gives

$$\begin{aligned}
\mathbf{f}_2 &= \mathbf{x}_2 - \frac{\mathbf{x}_2 \cdot \mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 = (3, 2, 0, 1) - \frac{4}{4}(1, 1, -1, -1) = (2, 1, 1, 2) \\
\mathbf{f}_3 &= \mathbf{x}_3 - \frac{\mathbf{x}_3 \cdot \mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 - \frac{\mathbf{x}_3 \cdot \mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 = \mathbf{x}_3 - \frac{0}{4}\mathbf{f}_1 - \frac{3}{10}\mathbf{f}_2 = \frac{1}{10}(4, -3, 7, -6)
\end{aligned}$$

Hence $\{(1, 1, -1, -1), (2, 1, 1, 2), \frac{1}{10}(4, -3, 7, -6)\}$ is the orthogonal basis provided by the algorithm. In hand calculations it may be convenient to eliminate fractions (see the Remark below), so $\{(1, 1, -1, -1), (2, 1, 1, 2), (4, -3, 7, -6)\}$ is also an orthogonal basis for row A.

---

[1] Erhardt Schmidt (1876–1959) was a German mathematician who studied under the great David Hilbert and later developed the theory of Hilbert spaces. He first described the present algorithm in 1907. Jörgen Pederson Gram (1850–1916) was a Danish actuary.

Remark

Observe that the vector $\frac{\mathbf{x}\cdot\mathbf{f}_i}{\|\mathbf{f}_i\|^2}\mathbf{f}_i$ is unchanged if a nonzero scalar multiple of $\mathbf{f}_i$ is used in place of $\mathbf{f}_i$. Hence, if a newly constructed $\mathbf{f}_i$ is multiplied by a nonzero scalar at some stage of the Gram-Schmidt algorithm, the subsequent **f**s will be unchanged. This is useful in actual calculations.

Projections

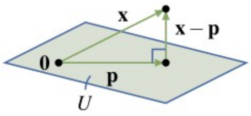

Suppose a point $\mathbf{x}$ and a plane U through the origin in $\mathbb{R}^3$ are given, and we want to find the point $\mathbf{p}$ in the plane that is closest to $\mathbf{x}$. Our geometric intuition assures us that such a point $\mathbf{p}$ exists. In fact (see the diagram), $\mathbf{p}$ must be chosen in such a way that $\mathbf{x}-\mathbf{p}$ is *perpendicular* to the plane.

Now we make two observations: first, the plane U is a *subspace of* $\mathbb{R}^3$ (because U contains the origin); and second, that the condition that $\mathbf{x}-\mathbf{p}$ is perpendicular to the plane U means that $\mathbf{x}-\mathbf{p}$ is *orthogonal* to every vector in U. In these terms the whole discussion makes sense in $\mathbb{R}^n$. Furthermore, the orthogonal lemma provides exactly what is needed to find $\mathbf{p}$ in this more general setting.

Definition 8.1 Orthogonal Complement of a Subspace of $\mathbb{R}^n$

If U is a subspace of $\mathbb{R}^n$, define the ***orthogonal complement*** *$U^\perp$ of U (pronounced "U-perp") by*

$$U^\perp = \{\mathbf{x} \text{ in } \mathbb{R}^n \mid \mathbf{x}\cdot\mathbf{y} = 0 \text{ for all } \mathbf{y} \text{ in } U\}$$

The following lemma collects some useful properties of the orthogonal complement; the proof of (1) and (2) is left as Exercise 8.1.6.

Lemma 8.1.2

Let U be a subspace of $\mathbb{R}^n$.

1. *$U^\perp$ is a subspace of $\mathbb{R}^n$.*
2. *$\{\mathbf{0}\}^\perp = \mathbb{R}^n$ and $(\mathbb{R}^n)^\perp = \{\mathbf{0}\}$.*
3. *If $U = \text{span}\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$, then $U^\perp = \{\mathbf{x} \text{ in } \mathbb{R}^n \mid \mathbf{x}\cdot\mathbf{x}_i = 0 \text{ for } i = 1, 2, \dots, k\}$.*

Proof.

3. Let $U = \text{span}\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$; we must show that $U^\perp = \{\mathbf{x} \mid \mathbf{x}\cdot\mathbf{x}_i = 0 \text{ for each } i\}$. If $\mathbf{x}$ is in $U^\perp$ then $\mathbf{x}\cdot\mathbf{x}_i = 0$ for all i because each $\mathbf{x}_i$ is in U. Conversely, suppose that $\mathbf{x}\cdot\mathbf{x}_i = 0$ for all i; we must show that $\mathbf{x}$ is in $U^\perp$, that is, $\mathbf{x}\cdot\mathbf{y} = 0$ for each $\mathbf{y}$ in U. Write $\mathbf{y} = r_1\mathbf{x}_1 + r_2\mathbf{x}_2 + \cdots + r_k\mathbf{x}_k$, where each r_i is in $\mathbb{R}$. Then, using Theorem 5.3.1,

$$\mathbf{x}\cdot\mathbf{y} = r_1(\mathbf{x}\cdot\mathbf{x}_1) + r_2(\mathbf{x}\cdot\mathbf{x}_2) + \cdots + r_k(\mathbf{x}\cdot\mathbf{x}_k) = r_1 0 + r_2 0 + \cdots + r_k 0 = 0$$

as required. □

Example 8.1.2

Find $U^\perp$ if $U = \text{span}\,\{(1,\ -1,\ 2,\ 0),\ (1,\ 0,\ -2,\ 3)\}$ in $\mathbb{R}^4$.

Solution. By Lemma 8.1.2, $\mathbf{x} = (x,\ y,\ z,\ w)$ is in $U^\perp$ if and only if it is orthogonal to both $(1,\ -1,\ 2,\ 0)$ and $(1,\ 0,\ -2,\ 3)$; that is,

$$\begin{array}{rl} x - y + 2z \phantom{{}+3w} &= 0 \\ x \phantom{{}-y} - 2z + 3w &= 0 \end{array}$$

Gaussian elimination gives $U^\perp = \text{span}\,\{(2,\ 4,\ 1,\ 0),\ (3,\ 3,\ 0,\ -1)\}$.

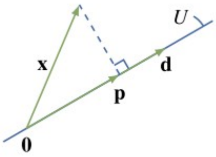

Now consider vectors $\mathbf{x}$ and $\mathbf{d} \neq \mathbf{0}$ in $\mathbb{R}^3$. The projection $\mathbf{p} = \text{proj}_{\mathbf{d}}\,\mathbf{x}$ of $\mathbf{x}$ on $\mathbf{d}$ was defined in Section 4.2 as in the diagram.

The following formula for $\mathbf{p}$ was derived in Theorem 4.2.4

$$\mathbf{p} = \text{proj}_{\mathbf{d}}\,\mathbf{x} = \left(\frac{\mathbf{x}\cdot\mathbf{d}}{\|\mathbf{d}\|^2}\right)\mathbf{d}$$

where it is shown that $\mathbf{x} - \mathbf{p}$ is orthogonal to $\mathbf{d}$. Now observe that the line $U = \mathbb{R}\mathbf{d} = \{t\mathbf{d} \mid t \in \mathbb{R}\}$ is a subspace of $\mathbb{R}^3$, that $\{\mathbf{d}\}$ is an orthogonal basis of U, and that $\mathbf{p} \in U$ and $\mathbf{x} - \mathbf{p} \in U^\perp$ (by Theorem 4.2.4).

In this form, this makes sense for any vector $\mathbf{x}$ in $\mathbb{R}^n$ and any subspace U of $\mathbb{R}^n$, so we generalize it as follows. If $\{\mathbf{f}_1,\ \mathbf{f}_2,\ \dots,\ \mathbf{f}_m\}$ is an orthogonal basis of U, we define the projection $\mathbf{p}$ of $\mathbf{x}$ on U by the formula

$$\mathbf{p} = \left(\frac{\mathbf{x}\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\right)\mathbf{f}_1 + \left(\frac{\mathbf{x}\cdot\mathbf{f}_2}{\|\mathbf{f}_2\|^2}\right)\mathbf{f}_2 + \cdots + \left(\frac{\mathbf{x}\cdot\mathbf{f}_m}{\|\mathbf{f}_m\|^2}\right)\mathbf{f}_m \tag{8.1}$$

Then $\mathbf{p} \in U$ and (by the orthogonal lemma) $\mathbf{x} - \mathbf{p} \in U^\perp$, so it looks like we have a generalization of Theorem 4.2.4.

However there is a potential problem: the formula (8.1) for $\mathbf{p}$ must be shown to be independent of the choice of the orthogonal basis $\{\mathbf{f}_1,\ \mathbf{f}_2,\ \dots,\ \mathbf{f}_m\}$. To verify this, suppose that $\{\mathbf{f}'_1,\ \mathbf{f}'_2,\ \dots,\ \mathbf{f}'_m\}$ is another orthogonal basis of U, and write

$$\mathbf{p}' = \left(\frac{\mathbf{x}\cdot\mathbf{f}'_1}{\|\mathbf{f}'_1\|^2}\right)\mathbf{f}'_1 + \left(\frac{\mathbf{x}\cdot\mathbf{f}'_2}{\|\mathbf{f}'_2\|^2}\right)\mathbf{f}'_2 + \cdots + \left(\frac{\mathbf{x}\cdot\mathbf{f}'_m}{\|\mathbf{f}'_m\|^2}\right)\mathbf{f}'_m$$

As before, $\mathbf{p}' \in U$ and $\mathbf{x} - \mathbf{p}' \in U^\perp$, and we must show that $\mathbf{p}' = \mathbf{p}$. To see this, write the vector $\mathbf{p} - \mathbf{p}'$ as follows:

$$\mathbf{p} - \mathbf{p}' = (\mathbf{x} - \mathbf{p}') - (\mathbf{x} - \mathbf{p})$$

This vector is in U (because $\mathbf{p}$ and $\mathbf{p}'$ are in U) and it is in $U^\perp$ (because $\mathbf{x} - \mathbf{p}'$ and $\mathbf{x} - \mathbf{p}$ are in $U^\perp$), and so it must be zero (it is orthogonal to itself!). This means $\mathbf{p}' = \mathbf{p}$ as desired.

Hence, the vector $\mathbf{p}$ in equation (8.1) depends only on $\mathbf{x}$ and the subspace U, and *not* on the choice of orthogonal basis $\{\mathbf{f}_1,\ \dots,\ \mathbf{f}_m\}$ of U used to compute it. Thus, we are entitled to make the following definition:

Definition 8.2 Projection onto a Subspace of $\mathbb{R}^n$

Let U be a subspace of $\mathbb{R}^n$ with orthogonal basis $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_m\}$. If $\mathbf{x}$ is in $\mathbb{R}^n$, the vector

$$\operatorname{proj}_U \mathbf{x} = \frac{\mathbf{x}\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 + \frac{\mathbf{x}\cdot\mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 + \cdots + \frac{\mathbf{x}\cdot\mathbf{f}_m}{\|\mathbf{f}_m\|^2}\mathbf{f}_m$$

*is called the **orthogonal projection** of $\mathbf{x}$ on U. For the zero subspace $U = \{\mathbf{0}\}$, we define*

$$\operatorname{proj}_{\{\mathbf{0}\}} \mathbf{x} = \mathbf{0}$$

The preceding discussion proves (1) of the following theorem.

Theorem 8.1.3: Projection Theorem

If U is a subspace of $\mathbb{R}^n$ and $\mathbf{x}$ is in $\mathbb{R}^n$, write $\mathbf{p} = \operatorname{proj}_U \mathbf{x}$. Then:

1. *$\mathbf{p}$ is in U and $\mathbf{x} - \mathbf{p}$ is in $U^\perp$.*
2. *$\mathbf{p}$ is the vector in U closest to $\mathbf{x}$ in the sense that*

$$\|\mathbf{x} - \mathbf{p}\| < \|\mathbf{x} - \mathbf{y}\| \quad \text{for all } \mathbf{y} \in U,\ \mathbf{y} \neq \mathbf{p}$$

Proof.

1. This is proved in the preceding discussion (it is clear if $U = \{\mathbf{0}\}$).
2. Write $\mathbf{x} - \mathbf{y} = (\mathbf{x} - \mathbf{p}) + (\mathbf{p} - \mathbf{y})$. Then $\mathbf{p} - \mathbf{y}$ is in U and so is orthogonal to $\mathbf{x} - \mathbf{p}$ by (1). Hence, the Pythagorean theorem gives

$$\|\mathbf{x} - \mathbf{y}\|^2 = \|\mathbf{x} - \mathbf{p}\|^2 + \|\mathbf{p} - \mathbf{y}\|^2 > \|\mathbf{x} - \mathbf{p}\|^2$$

because $\mathbf{p} - \mathbf{y} \neq \mathbf{0}$. This gives (2). □

Example 8.1.3

Let $U = \operatorname{span}\{\mathbf{x}_1, \mathbf{x}_2\}$ in $\mathbb{R}^4$ where $\mathbf{x}_1 = (1, 1, 0, 1)$ and $\mathbf{x}_2 = (0, 1, 1, 2)$. If $\mathbf{x} = (3, -1, 0, 2)$, find the vector in U closest to $\mathbf{x}$ and express $\mathbf{x}$ as the sum of a vector in U and a vector orthogonal to U.

Solution. $\{\mathbf{x}_1, \mathbf{x}_2\}$ is independent but not orthogonal. The Gram-Schmidt process gives an orthogonal basis $\{\mathbf{f}_1, \mathbf{f}_2\}$ of U where $\mathbf{f}_1 = \mathbf{x}_1 = (1, 1, 0, 1)$ and

$$\mathbf{f}_2 = \mathbf{x}_2 - \frac{\mathbf{x}_2\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 = \mathbf{x}_2 - \frac{3}{3}\mathbf{f}_1 = (-1, 0, 1, 1)$$

Hence, we can compute the projection using $\{\mathbf{f}_1, \mathbf{f}_2\}$:

$$\mathbf{p} = \operatorname{proj}_U \mathbf{x} = \frac{\mathbf{x}\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1 + \frac{\mathbf{x}\cdot\mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2 = \frac{4}{3}\mathbf{f}_1 + \frac{-1}{3}\mathbf{f}_2 = \frac{1}{3}\left[\begin{array}{cccc} 5 & 4 & -1 & 3 \end{array}\right]$$

Thus, $\mathbf{p}$ is the vector in U closest to $\mathbf{x}$, and $\mathbf{x}-\mathbf{p}=\frac{1}{3}(4,\ -7,\ 1,\ 3)$ is orthogonal to every vector in U. (This can be verified by checking that it is orthogonal to the generators $\mathbf{x}_1$ and $\mathbf{x}_2$ of U.) The required decomposition of $\mathbf{x}$ is thus

$$\mathbf{x}=\mathbf{p}+(\mathbf{x}-\mathbf{p})=\tfrac{1}{3}(5,\ 4,\ -1,\ 3)+\tfrac{1}{3}(4,\ -7,\ 1,\ 3)$$

Example 8.1.4

Find the point in the plane with equation $2x+y-z=0$ that is closest to the point $(2,\ -1,\ -3)$.

Solution. We write $\mathbb{R}^3$ as rows. The plane is the subspace U whose points $(x,\ y,\ z)$ satisfy $z=2x+y$. Hence

$$U=\{(s,\ t,\ 2s+t)\mid s,\ t \text{ in } \mathbb{R}\}=\operatorname{span}\{(0,\ 1,\ 1),\ (1,\ 0,\ 2)\}$$

The Gram-Schmidt process produces an orthogonal basis $\{\mathbf{f}_1,\ \mathbf{f}_2\}$ of U where $\mathbf{f}_1=(0,\ 1,\ 1)$ and $\mathbf{f}_2=(1,\ -1,\ 1)$. Hence, the vector in U closest to $\mathbf{x}=(2,\ -1,\ -3)$ is

$$\operatorname{proj}_U \mathbf{x}=\tfrac{\mathbf{x}\cdot\mathbf{f}_1}{\|\mathbf{f}_1\|^2}\mathbf{f}_1+\tfrac{\mathbf{x}\cdot\mathbf{f}_2}{\|\mathbf{f}_2\|^2}\mathbf{f}_2=-2\mathbf{f}_1+0\mathbf{f}_2=(0,\ -2,\ -2)$$

Thus, the point in U closest to $(2,\ -1,\ -3)$ is $(0,\ -2,\ -2)$.

The next theorem shows that projection on a subspace of $\mathbb{R}^n$ is actually a linear operator $\mathbb{R}^n\to\mathbb{R}^n$.

Theorem 8.1.4

Let U be a fixed subspace of $\mathbb{R}^n$. If we define $T:\mathbb{R}^n\to\mathbb{R}^n$ by

$$T(\mathbf{x})=\operatorname{proj}_U \mathbf{x} \quad \text{for all } \mathbf{x} \text{ in } \mathbb{R}^n$$

1. *T is a linear operator.*
2. *$\operatorname{im} T=U$ and $\ker T=U^\perp$.*
3. *$\dim U+\dim U^\perp=n$.*

Proof. If $U=\{\mathbf{0}\}$, then $U^\perp=\mathbb{R}^n$, and so $T(\mathbf{x})=\operatorname{proj}_{\{\mathbf{0}\}}\mathbf{x}=\mathbf{0}$ for all $\mathbf{x}$. Thus $T=0$ is the zero (linear) operator, so (1), (2), and (3) hold. Hence assume that $U\neq\{\mathbf{0}\}$.

1. If $\{\mathbf{f}_1,\ \mathbf{f}_2,\ \dots,\ \mathbf{f}_m\}$ is an orthonormal basis of U, then

$$T(\mathbf{x})=(\mathbf{x}\cdot\mathbf{f}_1)\mathbf{f}_1+(\mathbf{x}\cdot\mathbf{f}_2)\mathbf{f}_2+\cdots+(\mathbf{x}\cdot\mathbf{f}_m)\mathbf{f}_m \quad \text{for all } \mathbf{x} \text{ in } \mathbb{R}^n \tag{8.2}$$

by the definition of the projection. Thus T is linear because

$$(\mathbf{x}+\mathbf{y})\cdot\mathbf{f}_i=\mathbf{x}\cdot\mathbf{f}_i+\mathbf{y}\cdot\mathbf{f}_i \quad \text{and} \quad (r\mathbf{x})\cdot\mathbf{f}_i=r(\mathbf{x}\cdot\mathbf{f}_i) \quad \text{for each } i$$

2. We have $\text{im } T \subseteq U$ by (8.2) because each $\mathbf{f}_i$ is in U. But if $\mathbf{x}$ is in U, then $\mathbf{x} = T(\mathbf{x})$ by (8.2) and the expansion theorem applied to the space U. This shows that $U \subseteq \text{im } T$, so $\text{im } T = U$.

   Now suppose that $\mathbf{x}$ is in $U^\perp$. Then $\mathbf{x} \cdot \mathbf{f}_i = 0$ for each i (again because each $\mathbf{f}_i$ is in U) so $\mathbf{x}$ is in $\ker T$ by (8.2). Hence $U^\perp \subseteq \ker T$. On the other hand, Theorem 8.1.3 shows that $\mathbf{x} - T(\mathbf{x})$ is in $U^\perp$ for all $\mathbf{x}$ in $\mathbb{R}^n$, and it follows that $\ker T \subseteq U^\perp$. Hence $\ker T = U^\perp$, proving (2).

3. This follows from (1), (2), and the dimension theorem (Theorem 7.2.4). □

Exercises for 8.1

Exercise 8.1.1 In each case, use the Gram-Schmidt algorithm to convert the given basis B of V into an orthogonal basis.

a. $V = \mathbb{R}^2$, $B = \{(1, -1), (2, 1)\}$

b. $V = \mathbb{R}^2$, $B = \{(2, 1), (1, 2)\}$

c. $V = \mathbb{R}^3$, $B = \{(1, -1, 1), (1, 0, 1), (1, 1, 2)\}$

d. $V = \mathbb{R}^3$, $B = \{(0, 1, 1), (1, 1, 1), (1, -2, 2)\}$

Exercise 8.1.2 In each case, write $\mathbf{x}$ as the sum of a vector in U and a vector in $U^\perp$.

a. $\mathbf{x} = (1, 5, 7)$, $U = \text{span}\{(1, -2, 3), (-1, 1, 1)\}$

b. $\mathbf{x} = (2, 1, 6)$, $U = \text{span}\{(3, -1, 2), (2, 0, -3)\}$

c. $\mathbf{x} = (3, 1, 5, 9)$,
$U = \text{span}\{(1, 0, 1, 1), (0, 1, -1, 1), (-2, 0, 1, 1)\}$

d. $\mathbf{x} = (2, 0, 1, 6)$,
$U = \text{span}\{(1, 1, 1, 1), (1, 1, -1, -1), (1, -1, 1, -1)\}$

e. $\mathbf{x} = (a, b, c, d)$,
$U = \text{span}\{(1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0)\}$

f. $\mathbf{x} = (a, b, c, d)$,
$U = \text{span}\{(1, -1, 2, 0), (-1, 1, 1, 1)\}$

Exercise 8.1.3 Let $\mathbf{x} = (1, -2, 1, 6)$ in $\mathbb{R}^4$, and let $U = \text{span}\{(2, 1, 3, -4), (1, 2, 0, 1)\}$.

a. Compute $\text{proj}_U \mathbf{x}$.

b. Show that $\{(1, 0, 2, -3), (4, 7, 1, 2)\}$ is another orthogonal basis of U.

c. Use the basis in part (b) to compute $\text{proj}_U \mathbf{x}$.

Exercise 8.1.4 In each case, use the Gram-Schmidt algorithm to find an orthogonal basis of the subspace U, and find the vector in U closest to $\mathbf{x}$.

a. $U = \text{span}\{(1, 1, 1), (0, 1, 1)\}$, $\mathbf{x} = (-1, 2, 1)$

b. $U = \text{span}\{(1, -1, 0), (-1, 0, 1)\}$, $\mathbf{x} = (2, 1, 0)$

c. $U = \text{span}\{(1, 0, 1, 0), (1, 1, 1, 0), (1, 1, 0, 0)\}$,
$\mathbf{x} = (2, 0, -1, 3)$

d. $U = \text{span}\{(1, -1, 0, 1), (1, 1, 0, 0), (1, 1, 0, 1)\}$,
$\mathbf{x} = (2, 0, 3, 1)$

Exercise 8.1.5 Let $U = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$, $\mathbf{v}_i$ in $\mathbb{R}^n$, and let A be the $k \times n$ matrix with the $\mathbf{v}_i$ as rows.

a. Show that $U^\perp = \{\mathbf{x} \mid \mathbf{x} \text{ in } \mathbb{R}^n, A\mathbf{x}^T = \mathbf{0}\}$.

b. Use part (a) to find $U^\perp$ if
$U = \text{span}\{(1, -1, 2, 1), (1, 0, -1, 1)\}$.

Exercise 8.1.6

a. Prove part 1 of Lemma 8.1.2.

b. Prove part 2 of Lemma 8.1.2.

Exercise 8.1.7 Let U be a subspace of $\mathbb{R}^n$. If $\mathbf{x}$ in $\mathbb{R}^n$ can be written in any way at all as $\mathbf{x} = \mathbf{p} + \mathbf{q}$ with $\mathbf{p}$ in U and $\mathbf{q}$ in $U^\perp$, show that necessarily $\mathbf{p} = \operatorname{proj}_U \mathbf{x}$.

Exercise 8.1.8 Let U be a subspace of $\mathbb{R}^n$ and let $\mathbf{x}$ be a vector in $\mathbb{R}^n$. Using Exercise 8.1.7, or otherwise, show that $\mathbf{x}$ is in U if and only if $\mathbf{x} = \operatorname{proj}_U \mathbf{x}$.

Exercise 8.1.9 Let U be a subspace of $\mathbb{R}^n$.

a. Show that $U^\perp = \mathbb{R}^n$ if and only if $U = \{\mathbf{0}\}$.

b. Show that $U^\perp = \{\mathbf{0}\}$ if and only if $U = \mathbb{R}^n$.

Exercise 8.1.10 If U is a subspace of $\mathbb{R}^n$, show that $\operatorname{proj}_U \mathbf{x} = \mathbf{x}$ for all $\mathbf{x}$ in U.

Exercise 8.1.11 If U is a subspace of $\mathbb{R}^n$, show that $\mathbf{x} = \operatorname{proj}_U \mathbf{x} + \operatorname{proj}_{U^\perp} \mathbf{x}$ for all $\mathbf{x}$ in $\mathbb{R}^n$.

Exercise 8.1.12 If $\{\mathbf{f}_1, \dots, \mathbf{f}_n\}$ is an orthogonal basis of $\mathbb{R}^n$ and $U = \operatorname{span}\{\mathbf{f}_1, \dots, \mathbf{f}_m\}$, show that $U^\perp = \operatorname{span}\{\mathbf{f}_{m+1}, \dots, \mathbf{f}_n\}$.

Exercise 8.1.13 If U is a subspace of $\mathbb{R}^n$, show that $U^{\perp\perp} = U$. [*Hint*: Show that $U \subseteq U^{\perp\perp}$, then use Theorem 8.1.4 (3) twice.]

Exercise 8.1.14 If U is a subspace of $\mathbb{R}^n$, show how to find an $n \times n$ matrix A such that $U = \{\mathbf{x} \mid A\mathbf{x} = \mathbf{0}\}$. [*Hint*: Exercise 8.1.13.]

Exercise 8.1.15 Write $\mathbb{R}^n$ as rows. If A is an $n \times n$ matrix, write its null space as $\operatorname{null} A = \{\mathbf{x} \text{ in } \mathbb{R}^n \mid A\mathbf{x}^T = \mathbf{0}\}$. Show that:

a. $\operatorname{null} A = (\operatorname{row} A)^\perp$;

b. $\operatorname{null} A^T = (\operatorname{col} A)^\perp$.

Exercise 8.1.16 If U and W are subspaces, show that $(U+W)^\perp = U^\perp \cap W^\perp$. [See Exercise 5.1.22.]

Exercise 8.1.17 Think of $\mathbb{R}^n$ as consisting of rows.

a. Let E be an $n \times n$ matrix, and let $U = \{\mathbf{x}E \mid \mathbf{x} \text{ in } \mathbb{R}^n\}$. Show that the following are equivalent.

   i. $E^2 = E = E^T$ (E is a **projection matrix**).

   ii. $(\mathbf{x} - \mathbf{x}E) \cdot (\mathbf{y}E) = 0$ for all $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$.

   iii. $\operatorname{proj}_U \mathbf{x} = \mathbf{x}E$ for all $\mathbf{x}$ in $\mathbb{R}^n$.

   [*Hint*: For (ii) implies (iii): Write $\mathbf{x} = \mathbf{x}E + (\mathbf{x} - \mathbf{x}E)$ and use the uniqueness argument preceding the definition of $\operatorname{proj}_U \mathbf{x}$. For (iii) implies (ii): $\mathbf{x} - \mathbf{x}E$ is in $U^\perp$ for all $\mathbf{x}$ in $\mathbb{R}^n$.]

b. If E is a projection matrix, show that $I - E$ is also a projection matrix.

c. If $EF = 0 = FE$ and E and F are projection matrices, show that $E + F$ is also a projection matrix.

d. If A is $m \times n$ and AA^T is invertible, show that $E = A^T(AA^T)^{-1}A$ is a projection matrix.

Exercise 8.1.18 Let A be an $n \times n$ matrix of rank r. Show that there is an invertible $n \times n$ matrix U such that UA is a row-echelon matrix with the property that the first r rows are orthogonal. [*Hint*: Let R be the row-echelon form of A, and use the Gram-Schmidt process on the nonzero rows of R from the bottom up. Use Lemma 2.4.1.]

Exercise 8.1.19 Let A be an $(n-1) \times n$ matrix with rows $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n-1}$ and let A_i denote the $(n-1) \times (n-1)$ matrix obtained from A by deleting column i. Define the vector $\mathbf{y}$ in $\mathbb{R}^n$ by

$$\mathbf{y} = \left[\begin{array}{ccccc} \det A_1 & -\det A_2 & \det A_3 & \cdots & (-1)^{n+1} \det A_n \end{array}\right]$$

Show that:

a. $\mathbf{x}_i \cdot \mathbf{y} = 0$ for all $i = 1, 2, \dots, n-1$. [*Hint*: Write $B_i = \left[\begin{array}{c} x_i \\ A \end{array}\right]$ and show that $\det B_i = 0$.]

b. $\mathbf{y} \neq \mathbf{0}$ if and only if $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n-1}\}$ is linearly independent. [*Hint*: If some $\det A_i \neq 0$, the rows of A_i are linearly independent. Conversely, if the $\mathbf{x}_i$ are independent, consider $A = UR$ where R is in reduced row-echelon form.]

c. If $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n-1}\}$ is linearly independent, use Theorem 8.1.3(3) to show that all solutions to the system of $n-1$ homogeneous equations

$$A\mathbf{x}^T = \mathbf{0}$$

are given by $t\mathbf{y}$, t a parameter.

8.2 Orthogonal Diagonalization

Recall (Theorem 5.5.3) that an $n \times n$ matrix A is diagonalizable if and only if it has n linearly independent eigenvectors. Moreover, the matrix P with these eigenvectors as columns is a diagonalizing matrix for A, that is

$$P^{-1}AP \text{ is diagonal.}$$

As we have seen, the really nice bases of $\mathbb{R}^n$ are the orthogonal ones, so a natural question is: which $n \times n$ matrices have an *orthogonal* basis of eigenvectors? These turn out to be precisely the symmetric matrices, and this is the main result of this section.

Before proceeding, recall that an orthogonal set of vectors is called *orthonormal* if $\|\mathbf{v}\| = 1$ for each vector $\mathbf{v}$ in the set, and that any orthogonal set $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ can be "*normalized*", that is converted into an orthonormal set $\{\frac{1}{\|\mathbf{v}_1\|}\mathbf{v}_1, \frac{1}{\|\mathbf{v}_2\|}\mathbf{v}_2, \dots, \frac{1}{\|\mathbf{v}_k\|}\mathbf{v}_k\}$. In particular, if a matrix A has n orthogonal eigenvectors, they can (by normalizing) be taken to be orthonormal. The corresponding diagonalizing matrix P has orthonormal columns, and such matrices are very easy to invert.

Theorem 8.2.1

The following conditions are equivalent for an $n \times n$ matrix P.

1. *P is invertible and $P^{-1} = P^T$.*
2. *The rows of P are orthonormal.*
3. *The columns of P are orthonormal.*

Proof. First recall that condition (1) is equivalent to $PP^T = I$ by Corollary 2.4.1 of Theorem 2.4.5. Let $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ denote the rows of P. Then $\mathbf{x}_j^T$ is the jth column of P^T, so the (i, j)-entry of PP^T is $\mathbf{x}_i \cdot \mathbf{x}_j$. Thus $PP^T = I$ means that $\mathbf{x}_i \cdot \mathbf{x}_j = 0$ if $i \neq j$ and $\mathbf{x}_i \cdot \mathbf{x}_j = 1$ if $i = j$. Hence condition (1) is equivalent to (2). The proof of the equivalence of (1) and (3) is similar. □

Definition 8.3 Orthogonal Matrices

An $n \times n$ matrix P is called an **orthogonal matrix**[2] *if it satisfies one (and hence all) of the conditions in Theorem 8.2.1.*

Example 8.2.1

The rotation matrix $\left[\begin{array}{rr} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{array} \right]$ is orthogonal for any angle θ.

These orthogonal matrices have the virtue that they are easy to invert—simply take the transpose. But they have many other important properties as well. If $T : \mathbb{R}^n \to \mathbb{R}^n$ is a linear operator, we will prove

---

[2] In view of (2) and (3) of Theorem 8.2.1, *orthonormal matrix* might be a better name. But *orthogonal matrix* is standard.

(Theorem 10.4.3) that T is distance preserving if and only if its matrix is orthogonal. In particular, the matrices of rotations and reflections about the origin in $\mathbb{R}^2$ and $\mathbb{R}^3$ are all orthogonal (see Example 8.2.1).

It is not enough that the rows of a matrix A are merely orthogonal for A to be an orthogonal matrix. Here is an example.

Example 8.2.2

The matrix $\left[\begin{array}{rrr} 2 & 1 & 1 \\ -1 & 1 & 1 \\ 0 & -1 & 1 \end{array}\right]$ has orthogonal rows but the columns are not orthogonal. However, if the rows are normalized, the resulting matrix $\left[\begin{array}{ccc} \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{-1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{array}\right]$ is orthogonal (so the columns are now orthonormal as the reader can verify).

Example 8.2.3

If P and Q are orthogonal matrices, then PQ is also orthogonal, as is $P^{-1} = P^T$.

Solution. P and Q are invertible, so PQ is also invertible and

$$(PQ)^{-1} = Q^{-1}P^{-1} = Q^TP^T = (PQ)^T$$

Hence PQ is orthogonal. Similarly,

$$(P^{-1})^{-1} = P = (P^T)^T = (P^{-1})^T$$

shows that P^{-1} is orthogonal.

Definition 8.4 Orthogonally Diagonalizable Matrices

An $n \times n$ matrix A is said to be ***orthogonally diagonalizable*** *when an orthogonal matrix P can be found such that $P^{-1}AP = P^TAP$ is diagonal.*

This condition turns out to characterize the symmetric matrices.

Theorem 8.2.2: Principal Axes Theorem

The following conditions are equivalent for an $n \times n$ matrix A.

1. *A has an orthonormal set of n eigenvectors.*
2. *A is orthogonally diagonalizable.*
3. *A is symmetric.*

Proof. (1) $\Leftrightarrow$ (2). Given (1), let $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ be orthonormal eigenvectors of A. Then $P = \left[\begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_n \end{array}\right]$ is orthogonal, and $P^{-1}AP$ is diagonal by Theorem 3.3.4. This proves (2). Conversely, given (2) let $P^{-1}AP$ be diagonal where P is orthogonal. If $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ are the columns of P then $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ is an orthonormal basis of $\mathbb{R}^n$ that consists of eigenvectors of A by Theorem 3.3.4. This proves (1).

(2) $\Rightarrow$ (3). If $P^TAP = D$ is diagonal, where $P^{-1} = P^T$, then $A = PDP^T$. But $D^T = D$, so this gives $A^T = P^{TT}D^TP^T = PDP^T = A$.

(3) $\Rightarrow$ (2). If A is an $n \times n$ symmetric matrix, we proceed by induction on n. If $n = 1$, A is already diagonal. If $n > 1$, assume that (3) $\Rightarrow$ (2) for $(n-1)\times(n-1)$ symmetric matrices. By Theorem 5.5.7 let λ_1 be a (real) eigenvalue of A, and let $A\mathbf{x}_1 = \lambda_1\mathbf{x}_1$, where $\|\mathbf{x}_1\| = 1$. Use the Gram-Schmidt algorithm to find an orthonormal basis $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ for $\mathbb{R}^n$. Let $P_1 = \left[\begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_n \end{array}\right]$, so P_1 is an orthogonal matrix and $P_1^TAP_1 = \left[\begin{array}{cc} \lambda_1 & B \\ 0 & A_1 \end{array}\right]$ in block form by Lemma 5.5.2. But $P_1^TAP_1$ is symmetric (A is), so it follows that $B = 0$ and A_1 is symmetric. Then, by induction, there exists an $(n-1)\times(n-1)$ orthogonal matrix Q such that $Q^TA_1Q = D_1$ is diagonal. Observe that $P_2 = \left[\begin{array}{cc} 1 & 0 \\ 0 & Q \end{array}\right]$ is orthogonal, and compute:

$$\begin{aligned}(P_1P_2)^TA(P_1P_2) &= P_2^T(P_1^TAP_1)P_2 \\ &= \left[\begin{array}{cc} 1 & 0 \\ 0 & Q^T \end{array}\right]\left[\begin{array}{cc} \lambda_1 & 0 \\ 0 & A_1 \end{array}\right]\left[\begin{array}{cc} 1 & 0 \\ 0 & Q \end{array}\right] \\ &= \left[\begin{array}{cc} \lambda_1 & 0 \\ 0 & D_1 \end{array}\right]\end{aligned}$$

is diagonal. Because P_1P_2 is orthogonal, this proves (2). ☐

A set of orthonormal eigenvectors of a symmetric matrix A is called a set of **principal axes** for A. The name comes from geometry, and this is discussed in Section 8.9. Because the eigenvalues of a (real) symmetric matrix are real, Theorem 8.2.2 is also called the **real spectral theorem**, and the set of distinct eigenvalues is called the **spectrum** of the matrix. In full generality, the spectral theorem is a similar result for matrices with complex entries (Theorem 8.7.8).

Example 8.2.4

Find an orthogonal matrix P such that $P^{-1}AP$ is diagonal, where $A = \left[\begin{array}{rrr} 1 & 0 & -1 \\ 0 & 1 & 2 \\ -1 & 2 & 5 \end{array}\right]$.

Solution. The characteristic polynomial of A is (adding twice row 1 to row 2):

$$c_A(x) = \det \left[\begin{array}{ccc} x-1 & 0 & 1 \\ 0 & x-1 & -2 \\ 1 & -2 & x-5 \end{array}\right] = x(x-1)(x-6)$$

Thus the eigenvalues are $\lambda = 0$, 1, and 6, and corresponding eigenvectors are

$$\mathbf{x}_1 = \left[\begin{array}{r} 1 \\ -2 \\ 1 \end{array}\right] \quad \mathbf{x}_2 = \left[\begin{array}{r} 2 \\ 1 \\ 0 \end{array}\right] \quad \mathbf{x}_3 = \left[\begin{array}{r} -1 \\ 2 \\ 5 \end{array}\right]$$

respectively. Moreover, by what appears to be remarkably good luck, these eigenvectors are *orthogonal*. We have $\|\mathbf{x}_1\|^2 = 6$, $\|\mathbf{x}_2\|^2 = 5$, and $\|\mathbf{x}_3\|^2 = 30$, so

$$P = \left[\begin{array}{ccc} \frac{1}{\sqrt{6}}\mathbf{x}_1 & \frac{1}{\sqrt{5}}\mathbf{x}_2 & \frac{1}{\sqrt{30}}\mathbf{x}_3 \end{array} \right] = \frac{1}{\sqrt{30}} \left[\begin{array}{ccc} \sqrt{5} & 2\sqrt{6} & -1 \\ -2\sqrt{5} & \sqrt{6} & 2 \\ \sqrt{5} & 0 & 5 \end{array} \right]$$

is an orthogonal matrix. Thus $P^{-1} = P^T$ and

$$P^TAP = \left[\begin{array}{ccc} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 6 \end{array} \right]$$

by the diagonalization algorithm.

Actually, the fact that the eigenvectors in Example 8.2.4 are orthogonal is no coincidence. Theorem 5.5.4 guarantees they are linearly independent (they correspond to distinct eigenvalues); the fact that the matrix is *symmetric* implies that they are orthogonal. To prove this we need the following useful fact about symmetric matrices.

Theorem 8.2.3

If A is an $n \times n$ symmetric matrix, then

$$(A\mathbf{x}) \cdot \mathbf{y} = \mathbf{x} \cdot (A\mathbf{y})$$

for all columns $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$.[3]

Proof. Recall that $\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T\mathbf{y}$ for all columns $\mathbf{x}$ and $\mathbf{y}$. Because $A^T = A$, we get

$$(A\mathbf{x}) \cdot \mathbf{y} = (A\mathbf{x})^T\mathbf{y} = \mathbf{x}^TA^T\mathbf{y} = \mathbf{x}^TA\mathbf{y} = \mathbf{x} \cdot (A\mathbf{y})$$

☐

Theorem 8.2.4

If A is a symmetric matrix, then eigenvectors of A corresponding to distinct eigenvalues are orthogonal.

Proof. Let $A\mathbf{x} = \lambda\mathbf{x}$ and $A\mathbf{y} = \mu\mathbf{y}$, where $\lambda \neq \mu$. Using Theorem 8.2.3, we compute

$$\lambda(\mathbf{x} \cdot \mathbf{y}) = (\lambda\mathbf{x}) \cdot \mathbf{y} = (A\mathbf{x}) \cdot \mathbf{y} = \mathbf{x} \cdot (A\mathbf{y}) = \mathbf{x} \cdot (\mu\mathbf{y}) = \mu(\mathbf{x} \cdot \mathbf{y})$$

Hence $(\lambda - \mu)(\mathbf{x} \cdot \mathbf{y}) = 0$, and so $\mathbf{x} \cdot \mathbf{y} = 0$ because $\lambda \neq \mu$. ☐

---

[3] The converse also holds (Exercise 8.2.15).

Now the procedure for diagonalizing a symmetric $n \times n$ matrix is clear. Find the distinct eigenvalues (all real by Theorem 5.5.7) and find orthonormal bases for each eigenspace (the Gram-Schmidt algorithm may be needed). Then the set of all these basis vectors is orthonormal (by Theorem 8.2.4) and contains n vectors. Here is an example.

Example 8.2.5

Orthogonally diagonalize the symmetric matrix $A = \left[\begin{array}{rrr} 8 & -2 & 2 \\ -2 & 5 & 4 \\ 2 & 4 & 5 \end{array}\right]$.

Solution. The characteristic polynomial is

$$c_A(x) = \det \left[\begin{array}{ccc} x-8 & 2 & -2 \\ 2 & x-5 & -4 \\ -2 & -4 & x-5 \end{array}\right] = x(x-9)^2$$

Hence the distinct eigenvalues are 0 and 9 of multiplicities 1 and 2, respectively, so $\dim(E_0) = 1$ and $\dim(E_9) = 2$ by Theorem 5.5.6 (A is diagonalizable, being symmetric). Gaussian elimination gives

$$E_0(A) = \text{span}\,\{\mathbf{x}_1\}, \quad \mathbf{x}_1 = \left[\begin{array}{r} 1 \\ 2 \\ -2 \end{array}\right], \quad \text{and} \quad E_9(A) = \text{span}\left\{ \left[\begin{array}{r} -2 \\ 1 \\ 0 \end{array}\right], \left[\begin{array}{r} 2 \\ 0 \\ 1 \end{array}\right] \right\}$$

The eigenvectors in E_9 are both orthogonal to $\mathbf{x}_1$ as Theorem 8.2.4 guarantees, but not to each other. However, the Gram-Schmidt process yields an orthogonal basis

$$\{\mathbf{x}_2,\ \mathbf{x}_3\} \text{ of } E_9(A) \quad \text{where} \quad \mathbf{x}_2 = \left[\begin{array}{r} -2 \\ 1 \\ 0 \end{array}\right] \text{ and } \mathbf{x}_3 = \left[\begin{array}{r} 2 \\ 4 \\ 5 \end{array}\right]$$

Normalizing gives orthonormal vectors $\{\frac{1}{3}\mathbf{x}_1,\ \frac{1}{\sqrt{5}}\mathbf{x}_2,\ \frac{1}{3\sqrt{5}}\mathbf{x}_3\}$, so

$$P = \left[\begin{array}{ccc} \frac{1}{3}\mathbf{x}_1 & \frac{1}{\sqrt{5}}\mathbf{x}_2 & \frac{1}{3\sqrt{5}}\mathbf{x}_3 \end{array}\right] = \frac{1}{3\sqrt{5}} \left[\begin{array}{rrr} \sqrt{5} & -6 & 2 \\ 2\sqrt{5} & 3 & 4 \\ -2\sqrt{5} & 0 & 5 \end{array}\right]$$

is an orthogonal matrix such that $P^{-1}AP$ is diagonal.

It is worth noting that other, more convenient, diagonalizing matrices P exist. For example, $\mathbf{y}_2 = \left[\begin{array}{r} 2 \\ 1 \\ 2 \end{array}\right]$ and $\mathbf{y}_3 = \left[\begin{array}{r} -2 \\ 2 \\ 1 \end{array}\right]$ lie in $E_9(A)$ and they are orthogonal. Moreover, they both have norm 3 (as does $\mathbf{x}_1$), so

$$Q = \left[\begin{array}{ccc} \frac{1}{3}\mathbf{x}_1 & \frac{1}{3}\mathbf{y}_2 & \frac{1}{3}\mathbf{y}_3 \end{array}\right] = \frac{1}{3} \left[\begin{array}{rrr} 1 & 2 & -2 \\ 2 & 1 & 2 \\ -2 & 2 & 1 \end{array}\right]$$

is a nicer orthogonal matrix with the property that $Q^{-1}AQ$ is diagonal.

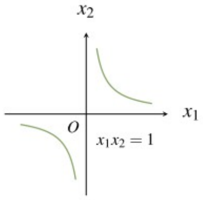

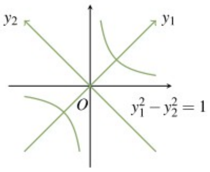

If A is symmetric and a set of orthogonal eigenvectors of A is given, the eigenvectors are called principal axes of A. The name comes from geometry. An expression $q = ax_1^2 + bx_1x_2 + cx_2^2$ is called a **quadratic form** in the variables x_1 and x_2, and the graph of the equation $q = 1$ is called a **conic** in these variables. For example, if $q = x_1x_2$, the graph of $q = 1$ is given in the first diagram.

But if we introduce new variables y_1 and y_2 by setting $x_1 = y_1 + y_2$ and $x_2 = y_1 - y_2$, then q becomes $q = y_1^2 - y_2^2$, a diagonal form with no cross term y_1y_2 (see the second diagram). Because of this, the y_1 and y_2 axes are called the principal axes for the conic (hence the name). Orthogonal diagonalization provides a systematic method for finding principal axes. Here is an illustration.

Example 8.2.6

Find principal axes for the quadratic form $q = x_1^2 - 4x_1x_2 + x_2^2$.

Solution. In order to utilize diagonalization, we first express q in matrix form. Observe that

$$q = \left[\begin{array}{cc} x_1 & x_2 \end{array}\right] \left[\begin{array}{rr} 1 & -4 \\ 0 & 1 \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \end{array}\right]$$

The matrix here is not symmetric, but we can remedy that by writing

$$q = x_1^2 - 2x_1x_2 - 2x_2x_1 + x_2^2$$

Then we have

$$q = \left[\begin{array}{cc} x_1 & x_2 \end{array}\right] \left[\begin{array}{rr} 1 & -2 \\ -2 & 1 \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \end{array}\right] = \mathbf{x}^T A\mathbf{x}$$

where $\mathbf{x} = \left[\begin{array}{c} x_1 \\ x_2 \end{array}\right]$ and $A = \left[\begin{array}{rr} 1 & -2 \\ -2 & 1 \end{array}\right]$ is symmetric. The eigenvalues of A are $\lambda_1 = 3$ and $\lambda_2 = -1$, with corresponding (orthogonal) eigenvectors $\mathbf{x}_1 = \left[\begin{array}{r} 1 \\ -1 \end{array}\right]$ and $\mathbf{x}_2 = \left[\begin{array}{c} 1 \\ 1 \end{array}\right]$. Since $\|\mathbf{x}_1\| = \|\mathbf{x}_2\| = \sqrt{2}$, so

$$P = \tfrac{1}{\sqrt{2}} \left[\begin{array}{rr} 1 & 1 \\ -1 & 1 \end{array}\right] \text{ is orthogonal and } P^TAP = D = \left[\begin{array}{rr} 3 & 0 \\ 0 & -1 \end{array}\right]$$

Now define new variables $\left[\begin{array}{c} y_1 \\ y_2 \end{array}\right] = \mathbf{y}$ by $\mathbf{y} = P^T\mathbf{x}$, equivalently $\mathbf{x} = P\mathbf{y}$ (since $P^{-1} = P^T$). Hence

$$y_1 = \tfrac{1}{\sqrt{2}}(x_1 - x_2) \quad \text{and} \quad y_2 = \tfrac{1}{\sqrt{2}}(x_1 + x_2)$$

In terms of y_1 and y_2, q takes the form

$$q = \mathbf{x}^T A\mathbf{x} = (P\mathbf{y})^T A(P\mathbf{y}) = \mathbf{y}^T(P^TAP)\mathbf{y} = \mathbf{y}^T D\mathbf{y} = 3y_1^2 - y_2^2$$

Note that $\mathbf{y} = P^T\mathbf{x}$ is obtained from $\mathbf{x}$ by a counterclockwise rotation of $\frac{\pi}{4}$ (see Theorem 2.4.6).

Observe that the quadratic form q in Example 8.2.6 can be diagonalized in other ways. For example

$$q = x_1^2 - 4x_1x_2 + x_2^2 = z_1^2 - \tfrac{1}{3}z_2^2$$

where $z_1 = x_1 - 2x_2$ and $z_2 = 3x_2$. We examine this more carefully in Section 8.9.

If we are willing to replace "diagonal" by "upper triangular" in the principal axes theorem, we can weaken the requirement that A is symmetric to insisting only that A has real eigenvalues.

Theorem 8.2.5: Triangulation Theorem

If A is an $n \times n$ matrix with n real eigenvalues, an orthogonal matrix P exists such that P^TAP is upper triangular.[4]

Proof. We modify the proof of Theorem 8.2.2. If $A\mathbf{x}_1 = \lambda_1\mathbf{x}_1$ where $\|\mathbf{x}_1\| = 1$, let $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ be an orthonormal basis of $\mathbb{R}^n$, and let $P_1 = \left[\begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{array}\right]$. Then P_1 is orthogonal and $P_1^TAP_1 = \left[\begin{array}{cc} \lambda_1 & B \\ 0 & A_1 \end{array}\right]$ in block form. By induction, let $Q^TA_1Q = T_1$ be upper triangular where Q is of size $(n-1)\times(n-1)$ and orthogonal. Then $P_2 = \left[\begin{array}{cc} 1 & 0 \\ 0 & Q \end{array}\right]$ is orthogonal, so $P = P_1P_2$ is also orthogonal and $P^TAP = \left[\begin{array}{cc} \lambda_1 & BQ \\ 0 & T_1 \end{array}\right]$ is upper triangular. □

The proof of Theorem 8.2.5 gives no way to construct the matrix P. However, an algorithm will be given in Section 11.1 where an improved version of Theorem 8.2.5 is presented. In a different direction, a version of Theorem 8.2.5 holds for an arbitrary matrix with complex entries (Schur's theorem in Section 8.7).

As for a diagonal matrix, the eigenvalues of an upper triangular matrix are displayed along the main diagonal. Because A and P^TAP have the same determinant and trace whenever P is orthogonal, Theorem 8.2.5 gives:

Corollary 8.2.1

If A is an $n \times n$ matrix with real eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ (possibly not all distinct), then $\det A = \lambda_1\lambda_2\dots\lambda_n$ and $\operatorname{tr} A = \lambda_1 + \lambda_2 + \dots + \lambda_n$.

This corollary remains true even if the eigenvalues are not real (using Schur's theorem).

Exercises for 8.2

Exercise 8.2.1 Normalize the rows to make each of the following matrices orthogonal.

a. $A = \left[\begin{array}{rr} 1 & 1 \\ -1 & 1 \end{array}\right]$

b. $A = \left[\begin{array}{rr} 3 & -4 \\ 4 & 3 \end{array}\right]$

c. $A = \left[\begin{array}{rr} 1 & 2 \\ -4 & 2 \end{array}\right]$

d. $A = \left[\begin{array}{rr} a & b \\ -b & a \end{array}\right]$, $(a, b) \neq (0, 0)$

---

[4] There is also a lower triangular version.

"

02 03 2020

d) PG_Oculto/Math_Sphere_Equations/

d1) geometriaa5.pdf
Pág. 231(244) - Chapter 6
Círculos e Esferas

d2) geometriaanaliticaevetorial-SGD.pdf
Pág. 149(157) - Chapter 6
Círculos e Esferas.

e) ... Math_Books/
Spherical Geometry and Its Applications/
sanet.st_0367180464.pdf
Pág. 252(267)
6.1 Riemannian Metrics

120) 20200304
A) Lecture Notes on General Relativity
Matthias Blau
http://www.blau.itp.unibe.ch/GRLecturenotes.html

"   92 # 9 6 9<E   #

"   6"

/+01/   /   /2   / $ 6"

) & %%   6 ! F   <

6 CÍRCULOS E ESFERAS

6.1 EQUAÇÕES CANÔNICAS DE CÍRCULOS E ESFERAS

Um círculo é o conjunto de pontos no plano que estão a uma certa distância r de um ponto dado (a, b).

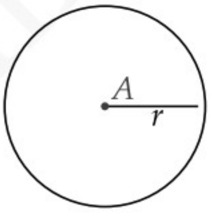

Figura 6.1: Círculo de centro A e raio r.

Desta forma temos que um ponto (x, y) pertence ao círculo de centro (a, b) e raio r se e somente se satisfaz a equação:

$$\sqrt{(x-a)^2 + (y-b)^2} = r$$

ou equivalentemente:

$$(x-a)^2 + (y-b)^2 = r^2$$

De modo análogo, a equação reduzida de uma esfera de centro (a, b, c) e raio r é

$$(x-a)^2 + (y-b)^2 + (z-c)^2 = r^2$$

Versão Preliminar

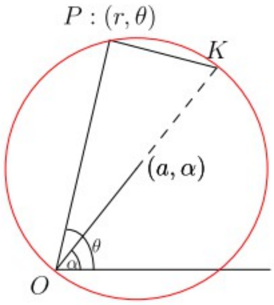

Figura 6.2: Esfera de Centro C e raio r.

Versão Preliminar

Exemplo 6.1 Determine a equação do círculo de centro $(-3,1)$ que

é tangente a reta $3x-4y-2=0$

Solução: Já conhecemos o centro e precisamos determinar o raio. Nesse caso o raio é a distância entre a reta e o ponto, já que a tangente a um círculo é perpendicular ao raio que liga o centro ao ponto de tangência. Logo:

$$r=\frac{|3(-3)-4\cdot 1-2|}{\sqrt{3^2+4^2}}=3$$

e assim a equação do círculo é:

$$(x+3)^2+(y-1)^2=9 \text{ ou } x^2+y^2+6x-2y+1=0$$

□

Exemplo 6.2 Determine a equação da esfera cujo diâmetro é o segmento que liga $(3,-1,2)$ a $(5,3,4)$.

Solução: Nesse caso aparentemente não conhecemos nem o centro nem o raio. Mas temos que o centro é o ponto médio do segmento e que o raio é metade do diâmetro. Logo:

$$r = \frac{1}{2}\sqrt{(5-3)^2 + (3+1)^2 + (4-2)^2} = \sqrt{6}$$

O ponto médio é $(4, 1, 3)$ e logo a equação da esfera é:

$$(x-4)^2 + (y-1)^2 + (z-3)^2 = 6$$

□

Exemplo 6.3 Identifique a curva cuja equação é:

$$x^2 + y^2 - 6x - 4y - 12 = 0$$

Solução: Identificaremos a curva completando quadrados. O termo $x^2 - 6x$ pode ser convertido num quadrado, se somarmos 9 e $y^2 - 4y$ pode ser convertido num quadrado somando 4. Desta forma, somaremos $4 + 9$ em cada lado da equação $x^2 + y^2 - 6x - 4y - 12 = 0$. Logo temos:

$$x^2 + y^2 - 6x - 4y - 12 = 0 \tag{6.1}$$

$$\Rightarrow (x^2 - 6x + 9) + (y^2 - 4y + 4) = 12 + 4 + 9 \tag{6.2}$$

$$\Rightarrow (x-3)^2 + (y-2)^2 = 5^2 \tag{6.3}$$

Logo a curva é um círculo de raio 5 e centro $(3, 2)$. □

Versão Preliminar

Versão Preliminar

Podemos generalizar o exemplo anterior:

Exemplo 6.4 Identifique a curva cuja equação é:

$$x^2 + y^2 + Ax + By + C = 0$$

Solução: Como no exemplo anterior, identificaremos a curva completando quadrados. O termo $x^2 + Ax$ pode ser convertido num quadrado, se somarmos $\dfrac{A^2}{4}$ e $y^2 + By$ pode ser convertido num quadrado somando $\dfrac{B^2}{4}$. Desta forma, somaremos $\dfrac{A^2}{4} + \dfrac{B^2}{4}$ em cada lado da equação:

$$x^2 + y^2 + Ax + By + C = 0 \tag{6.4}$$

$$\Rightarrow \left(x^2 + Ax + \frac{A^2}{4}\right) + \left(y^2 + By + \frac{B^2}{4}\right) = \frac{A^2}{4} + \frac{B^2}{4} - C \tag{6.5}$$

$$\Rightarrow \left(x + \frac{A}{2}\right)^2 + \left(y + \frac{B}{2}\right)^2 = \frac{A^2}{4} + \frac{B^2}{4} - C \tag{6.6}$$

Observamos que para a equação anterior ser a equação de um circulo, $r^2 = \dfrac{A^2}{4} + \dfrac{B^2}{4} - C$, e assim temos que ter $\dfrac{A^2}{4} + \dfrac{B^2}{4} - C > 0$.

No caso em que $\dfrac{A^2}{4} + \dfrac{B^2}{4} - C < 0$, o lugar geométrico descrito pela equação 6.6 é vazio, pois a equação não pode ser satisfeita pois a soma de quadrados é necessariamente negativa.

No caso em que $\dfrac{A^2}{4} + \dfrac{B^2}{4} - C = 0$, o lugar geométrico descrito pela equação 6.6 é o ponto $\left(-\dfrac{A}{2}, -\dfrac{B}{2}\right)$, pois se a soma de quadrados perfeitos é 0 cada termo da soma é zero. ☐

De modo análogo, podemos demonstrar que a equação

$$x^2 + y^2 + z^2 + Ax + By + Cz + D = 0$$

descreve uma esfera se $\frac{A^2}{4} + \frac{B^2}{4} + \frac{C^2}{4} - D > 0$, um ponto se $\frac{A^2}{4} + \frac{B^2}{4} + \frac{C^2}{4} - D = 0$ e o conjunto vazio se $\frac{A^2}{4} + \frac{B^2}{4} + \frac{C^2}{4} - D < 0$.

Exemplo 6.5 A superfície cuja equação é:

$$12 - 2x + x^2 + 4y + y^2 + 8z + z^2 = 0$$

é uma esfera. Encontre seu centro e raio.

Solução: Completando os quadrados temos

$$(x^2 - 2x + 1) + (y^2 + 4y + 4) + (z^2 + 8z + 16) - 1 - 4 - 16 + 12 = 0$$

Daí segue que:

$$(x - 1)^2 + (y + 2)^2 + (z + 4)^2 = 9$$

E logo o centro dessa esfera é $(1, -2, -4)$ e o raio é 3. ☐

6.1.1 Círculo por três pontos

É conhecido que três pontos não colineares determinam um único círculo. Assim sendo, fixados P_1, P_2 e P_3 não colineares podemos

Versão Preliminar

facilmente encontrar a equação do círculo que passa por tais pontos. Tal equação pode ser encontrada observando que a equação geral de um círculo é da forma

$$x^2 + y^2 + Ax + By + C = 0$$

e que um ponto pertence ao círculo se e somente se suas coordenadas satisfazem tal equação. A substituição de cada ponto resulta assim numa equação linear nas variáveis A, B, C e assim o fato dos três pontos pertencerem ao círculo nos fornecem um sistema linear em três equações e três variáveis A, B, C. Resolvendo tal sistema encontramos, então, a equação do círculo.

Exemplo 6.6 Determine a equação do círculo que passa pelos pontos $(-1, 2)$, $(0, 1)$ e $(-3, 2)$.

Solução: Substituindo os pontos na equação

temos o sistema:

$$\begin{cases} 5 - A + 2B + C = 0 \\ 1 + B + C = 0 \\ 13 - 3A + 2B + C \end{cases}$$

cujas solução é $A = 4, B = 0, C = -1$. E logo a equação é

$$x^2 + y^2 + 4x - 1 = 0.$$

Completando quadrado obtemos, então:

$$(x^2 + 4x + 4) + y^2 - 4 - 1 = 0.$$

Versão Preliminar

Donde segue:

$$(x+2)^2 + y^2 = 5.$$

Desse modo vemos que o círculo que passa por tais pontos tem centro $(-2,0)$ e raio $\sqrt{5}$. $\square$

É possível encontrar a equação de um círculo por três pontos não colineares de uma outra maneira. Para esse fim consideramos o triângulo determinado pelos pontos P_1, P_2, P_3 e esse circunscrito na circunferência. Assim o seu centro é o circuncentro desse triângulo, isto é, o encontro das mediatrizes.

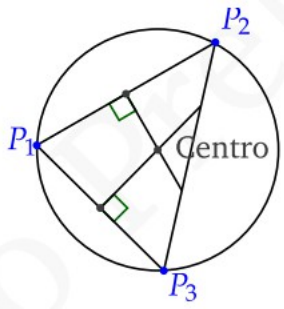

Exemplo 6.7 Determine a equação do círculo que passa pelos pontos $(-1,2)$, $(0,1)$ e $(-3,2)$.

Solução: A equação da reta passando pelos pontos $(-1,2)$, $(0,1)$ é $y-1=-x$, e como o ponto médio desses pontos é: $(-\frac{1}{2}, \frac{3}{2})$ temos que a mediatriz relativa a esse lado é: $y - \frac{3}{2} = x + \frac{1}{2}$ (lembrando

Versão Preliminar

que como a mediatriz é perpendicular ao lado seu coeficiente angular é igual a menos o inverso do coeficiente da reta).

De modo análogo a equação da reta passando pelos pontos $(0,1)$ e $(-3,2)$ é $y = -\frac{x}{3} + 1$ e a equação da mediatriz é: $3x = -6 + y$

temos o sistema:

$$\begin{cases} 3x = -6 + y \\ y - \frac{3}{2} = x + \frac{1}{2} \end{cases}$$

cujas solução é $x = -2, y = 0$, ou seja o centro da circunferência é $(-2,0)$. O raio pode ser calculado observando que este será a distância do centro $(-2,0)$ a um dos vértices do triângulo, por exemplo $(0,1)$. Assim $r^2 = 5$, e logo a equação é:

$$(x+2)^2 + y^2 = 5.$$

□

Versão Preliminar

Exemplo 6.8 Obtenha a equação da esfera que passa pelos pontos

$(0,0,1), (2,0,0), (1,1,1), (0,1,0)$ **Solução:** Impondo que os pontos pertençam a esfera temos o seguinte sistema linear:

$$\begin{aligned} &1 + C + D = 0 \\ &4 + 2A + D = 0 \\ &3 + A + B + C + D = 0 \\ &1 + B + D = 0 \end{aligned}$$

cuja solução é $A = -\frac{5}{3}, B = -\frac{1}{3}, C = -\frac{1}{3}, D = -\frac{2}{3}$ e assim a equação da esfera é:

$$x^2 + y^2 + z^2 - \frac{5x}{3} - \frac{y}{3} - \frac{z}{3} - \frac{2}{3} = 0$$

Completando quadrado obtemos:

$$\left(x^2 - \frac{5x}{3} + \left(\frac{5}{6}\right)^2\right) + \left(y^2 - \frac{y}{3} + \left(\frac{1}{6}\right)^2\right) +$$
$$+ \left(z^2 - \frac{z}{3} + \left(\frac{1}{6}\right)^2\right) - \left(\frac{5}{6}\right)^2 - \left(\frac{1}{6}\right)^2 - \left(\frac{1}{6}\right)^2 - \frac{24}{36} = 0.$$

Donde segue:

$$\left(x^2 - \frac{5}{6}\right)^2 + \left(y^2 - \frac{1}{6}\right)^2 + \left(z^2 - \frac{1}{6}\right)^2 = \frac{51}{36}.$$

□

Exercícios

Ex. 1.1 — Determine a equação dos seguintes círculos:

a) Centro $(-2, 5)$ e raio $r = 3$.

b) Centro $(1, 3)$ e raio $r = 2$

c) Centro a origem e raio $r = a$

d) Centro $(5, 2)$ e passando pelo ponto $(2, 3)$

e) Tangente ao eixo y na origem e raio a

f) Diâmetro $(5, 2)$ a $(-2, 10)$

Versão Preliminar

g) Centro $(3,-2)$ tangente a $2x-y=0$

h) Tangente a $2x-5y+1=0$ no ponto $(2,1)$ e raio 3 (duas respostas)

Ex. 1.2 — Identifique, dando o centro e o raio.

a) $x^2+y^2-4x+6y=12$

b) $x^2+y^2-2x-4y+5$

c) $x^2+y^2=2ax$

d) $4x^2-4x=5y-4y^2$

e) $x^2+y^2+z^2=2az$

Ex. 1.3 — Encontre a equação do círculo que passa pelos pontos $(4,0),(0,3)$ e a origem.

Ex. 1.4 — Encontre a equação dos seguintes círculos

a) Tangente aos eixos coordenados coordenados no segundo quadrante e com raio $r=4$.

b) Tangente ao eixo x, ao eixo y e a linha que intercepta o eixo x e o eixo y em 3 e 2 respectivamente.

Ex. 1.5 — Verifique que as equações abaixo descrevem esferas, em caso afirmativo identifique o centro e o raio:

a) $x^2+y^2+z^2-2x-4y+10=0$

Versão Preliminar

b) $x^2 - 6x + y^2 - 4y + z^2 + 14z + 58$

c) $x^2 + y^2 - 6y + z^2 + 4z + 16$

d) $x^2 + 2x + y^2 + 4y - z^2 + 6z - 29$

Ex. 1.6 — Dados $P_1 = (x_1, y_1, z_1)$ e $P_2 = (x_2, y_2, z_2)$ então a equação da esfera que tem P_1P_2 como diâmetro é

$$(x - x_1)(x - x_2) + (y - y_1)(y - y_2) + (z - z_1)(z - z_2) = 0$$

6.2 RETAS TANGENTES E PLANOS TANGENTES

Versão Preliminar

Uma reta é dita tangente a um círculo se a intersecção entre essa reta e o círculo for somente um ponto. Para uma reta tangente o seu vetor diretor é perpendicular ao vetor ligando o raio ao ponto de intersecção. Além disso a distância do centro do círculo a reta tangente é igual ao raio do círculo.

De modo análogo, dizemos que um plano é tangente a uma esfera se esse plano interceptar a esfera num único ponto. Nesse caso o vetor normal ao plano é paralelo ao vetor radial ligando o centro da esfera ao ponto onde o plano intercepta a esfera. E a distância do plano tangente ao centro da esfera é igual ao raio da mesma.

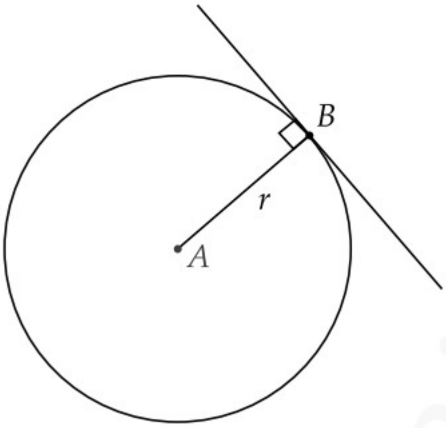

Figura 6.3: Reta tangente a um círculo

Versão Preliminar

Exemplo 6.9 Encontre a reta tangente ao círculo de equação $x^2 + y^2 - 2y - 4x = 0$ no ponto $(3,3)$

Solução: Completando quadrados podemos colocar a equação $x^2 + y^2 - 2y - 4x = 0$ na forma reduzida:

$$(x-2)^2 + (y-1)^2 = 0$$

Logo o centro do círculo tem coordenadas $(2,1)$. Logo, o vetor ligando o centro do círculo ao ponto $(3,3)$ é $\mathbf{i} + 2\mathbf{k}$ e assim o coeficiente angular da reta passando por estes pontos é igual a 2. Logo, o coeficiente da reta tangente é $-\frac{1}{2}$ (Por quê? Tente escrever a equação da reta tangente na forma padrão obtendo antes equações paramétricas para a mesma.). E assim a equação da reta tangente é:

$$y - 3 = -\frac{1}{2}(x-3)$$

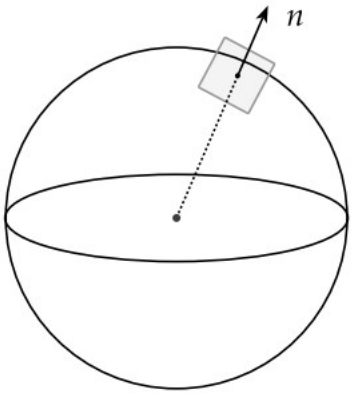

Figura 6.4: Plano tangente a uma esfera

ou

$$x + 2y = 9.$$

□

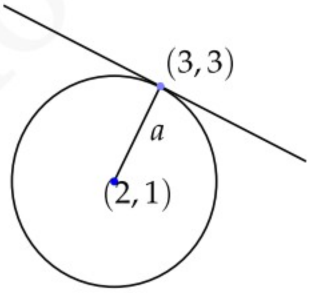

Versão Preliminar

Podemos generalizar o exemplo anterior. Dado um círculo de equação

$$(x-a)^2 + (y-b)^2 = r^2$$

Vamos calcular a equação da reta tangente no ponto (x_1, y_1).

Para tanto, consideraremos o vetor ligando o centro do círculo ao ponto de tangencia: $(x_1 - a)\mathbf{i} + (y_1 - b)\mathbf{j}$. Consequentemente a inclinação da reta passando por esses pontos é: $\dfrac{y_1 - b}{x_1 - a}$ Logo o coeficiente angular da reta tangente é $-\dfrac{x_1 - a}{y_1 - b}$. E assim a equação da reta tangente é da forma

$$(y - y_1) = -\frac{x_1 - a}{y_1 - b}(x + x_1)$$

e logo

$$(y - y_1)(y_1 - b) = -(x_1 - a)(x - x_1)$$

e assim expandindo:

$$(x_1 - a)x + (y_1 - b)y = k$$

para alguma constante k. Somando $(x_1 - a)(-a) + (y_1 - b)(-b)$ em ambos os lados da equação obtemos:

$$(x_1 - a)(x - a) + (y_1 - b)(y - b) = k_2$$

para alguma constante k_2, que determinaremos agora. Se substituirmos $x = x_1$ e $y = y_1$ teremos que

$$k_2 = (x_1 - a)^2 + (y_1 - b)^2 = r^2$$

Versão Preliminar

e assim a equação da reta tangente no ponto (x_1, y_1) é

$$(x_1 - a)(x - a) + (y_1 - b)(y - b) = r^2.$$

Exemplo 6.10 Obtenha as equações dos planos tangentes a esfera

$-3 - 2x + x^2 + 4y + y^2 + 2z + z^2 = 0$ que são paralelos ao plano $x - 2y + 2z = 3$.

Solução: Completando quadrados temos que a equação da esfera pode ser escrita como:

$$(x-1)^2 + (y+2)^2 + (z+1)^2 = 9$$

Logo o centro dessa esfera é $(1, -2, -1)$ e o raio é 3.

A equação geral de um plano paralelo a $x - 2y + 2z = 3$ tem equação da forma: $x - 2y + 2z = d$

Como esse plano é tangente a esfera a distância do centro dessas esferas ao plano é igual ao raio dessa esfera. E assim:

$$d(C, \pi) = \frac{|1 - 2(-2) + 2(-1) - d|}{9} = 3$$

e logo $d = -6$ ou $d = 12$ e assim as equações dos planos são $x - 2y + 2z = -6$ e $x - 2y + 2z = 12$.

□

Versão Preliminar

Exercícios

Ex. 2.1 — Encontre a equação a reta tangente no ponto indicado:

a) $x^2 + y^2 = 25$, $(-3, 4)$

b) $x^2 + y^2 = 2x - 4y$, origem.

c) Encontre as retas tangentes ao circulo $x^2 + y^2 = 4x$ que passam pelo ponto $(3, 2)$.

d) Uma corda da circunferência $x^2 + y^2 = 25$ se encontra sobre a reta cuja equação é $x - 7y + 25 = 0$. Qual o comprimento dessa corda?

Ex. 2.2 — Para um triângulo qualquer encontrar:

a) a equação da circunferência circunscrita ao triângulo

b) a equação da circunferência inscrita ao triângulo

c) a equação da circunferência que passa pelos pontos médios dos lados do triângulo.

Ex. 2.3 — As equações dos lados de um triângulo são $9x + 2y + 13 = 0, 3x + 8y - 47 = 0$ e $x - y - 1 = 0$. Encontrar a equação da circunferência circunscrita.

Ex. 2.4 — Mostrar que as tangentes de inclinação m à circunferência $x^2 + y^2 = r^2$ são $y = mx \pm r\sqrt{1 + m^2}$.

Ex. 2.5 — Qual a equação da circûnferencia que passa pelos pontos $(1, 2), (3, 4)$ e que tem centro sobre o eixo y?

Versão Preliminar

Ex. 2.6 — Fixado a, quais devem ser os dois valores de b para que a reta $y = ax + b$ seja tangente ao círculo de centro na origem e raio r?

Ex. 2.7 — Uma circunferência de raio 5 é tangente a reta $3x - 4y - 1 = 0$ no ponto $(3, 2)$. Determinar sua equação (duas soluções).

Ex. 2.8 — Mostrar analiticamente que qualquer reta que passa pelo ponto $(-1, 5)$ não pode ser tangente a circunferência $x^2 + y^2 + 4x - 6y + 6 = 0$. Interprete o resultado geometricamente.

Ex. 2.9 — Encontre a equação dos círculos que passam pelos seguintes conjuntos de pontos. Diga qual o centro, o raio e desenhe.

a) $(3, 4), (-1, 2), (-2, 4)$
b) $(4, 2), (-2, 3), (-1, 6)$
c) $(a, 0), (b, 0), (0, c)$

Ex. 2.10 — Mostrar que o plano tangente à esfera $x^2 + y^2 + z^2 = r^2$ no ponto (a, b, c) tem equação $ax + by + cz = r^2$

Ex. 2.11 — Encontre a equação da esfera que passa pelos pontos $(0, 0, 1)$,$(1, 0, 0)$, $(0, 1, 0)$ e cujo centro esta no plano $x + y - z = 0$

Versão Preliminar

Ex. 2.12 — Encontre a esfera que tem centro na reta

$$r : \begin{cases} x = 2z - 3 \\ y = z - 1 \end{cases}$$

e passa pelos pontos $(6, -1, 3)$ e $(0, 7, 5)$

Ex. 2.13 — Calcule a distância do ponto $(2, 3, 4)$ à esfera $x^2 + 4x + y^2 - 2y + z^2 + 4$.

Ex. 2.14 — Determine a equação da esfera cujo centro é $(3, 2, -2)$ é que é tangente ao plano

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} -3 \\ 1 \\ 0 \end{pmatrix} t + \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix} s$$

Ex. 2.15 — Determine a equação da esfera cujo centro se encontra sobre o eixo X e que passa pelos pontos $(3, -4, 2)$ e $(6, 2, -1)$.

Ex. 2.16 — A equação de uma esfera é $x^2 + y^2 + z^2 + 6y - 4z + 9 = 0$. Determinar a equação da esfera concêntrica que é tangente ao plano:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ -1 \end{pmatrix} + \begin{pmatrix} \frac{1}{2} \\ 1 \\ 1 \end{pmatrix} s + \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} t$$

Versão Preliminar

Ex. 2.17 — Encontre os planos tangentes a esfera $x^2 + y^2 + (z-1)^2 = 1$ que são paralelos ao plano $4x - y + 3z = 2$

Ex. 2.18 — Encontre a equação dos planos que contem a reta r e são tangentes a esfera S:

$$r : \frac{x+6}{2} = y + 3 = z + 1$$

e $S : x^2 + y^2 + z^2 - 4x + 2y - 4z + 4 = 0$.

Versão Preliminar

6.3 CIRCUNFERÊNCIA EM COORDENADAS POLARES

CENTRADA NA ORIGEM O caso mais simples ocorre quando a circunferência está centrada na origem nesse caso a circunferência é o conjunto de pontos que distam uma constante a da origem ou seja a equação em coordenadas polares é

$$r = a.$$

É fácil de ver que essa equação coincide com a em equação em coordenadas cartesianas. Observe que, em coordenadas cartesianas,

$P = (x, y)$ pertence a tal círculo se e somente se: $x = a\cos\theta$ e $y = a \operatorname{sen}\theta$. Daí segue que:

$$x^2 + y^2 = a^2(\cos^2\theta + \operatorname{sen}^2\theta) = a^2.$$

PASSANDO PELA ORIGEM Dada uma circunferência de raio a e passando pela origem. As coordenadas polares do centro dessa circunferência são (a, α).

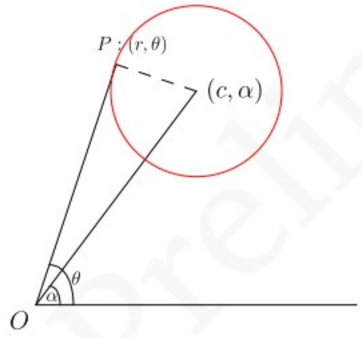

Considere o triângulo ΔOKP. Como $\overline{OK}$ é diâmetro da circunferência circunscrita ao triângulo vemos que ΔOKP é retângulo em P. Da definição de cosseno segue então:

$$r = 2a\cos\left(\theta - \alpha\right).$$

FORMA GERAL Dado uma circunferência de centro (c, α) e raio a, usando a lei dos cossenos temos que:

$$a^2 = r^2 + c^2 - 2rc\cos\left(\theta - \alpha\right)$$

Versão Preliminar

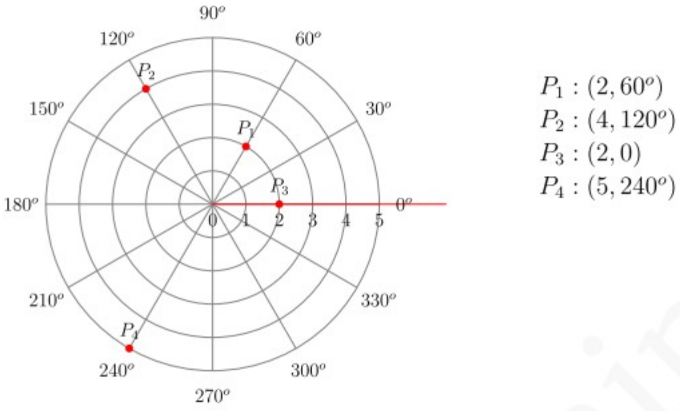
90°
120°
60°
150°
30°
180°
0°
210°
330°
240°
300°
270°
0
1
2
3
4
5
P_1
P_2
P_3
P_4
$P_1 : (2, 60^o)$
$P_2 : (4, 120^o)$
$P_3 : (2, 0)$
$P_4 : (5, 240^o)$

que é a equação da circunferência na forma geral.

Exercícios

Ex. 3.1 — Mostre que o centro do círculo de equação $r = A\cos\theta + B\operatorname{sen}\theta$ é

$$\left(\frac{\sqrt{A^2 + B^2}}{2}, \operatorname{arctg}\frac{B}{A}\right)$$

Ex. 3.2 — Mostre que a reta $r\operatorname{sen}\theta = 4$ é tangente ao círculo $r = 8\cos\theta$

Ex. 3.3 — Mostre que a equação da tangente ao círculo

$$r = 2a\cos\theta$$

Versão Preliminar

no ponto (r_1, θ_1) é:

$$r\cos(\theta - 2\theta_1) = 2a\cos^2\theta_1$$

Ex. 3.4 — Mostre que para todos os valores de a a reta

$$r\cos(\theta - \alpha) = a + r_1\cos\alpha$$

é tangente ao círculo

$$r^2 - 2rr_1\cos\theta + r_1^2 - a^2 = 0$$

Versão Preliminar

G /+01/ / /2 / $,
3G 6"

) %$ * 6 ! F <

/+01/ / /2 / $. , G 7
" 0 6"

) !* ! 0

vature and of geodesy for Riemann spaces. However, the concept of a differential manifold as presented in Chapter 3 did not appear until 1913 in the work of H. Weyl [59, I.§4]. According to Steenrod [53, p. v], general definitions for fiber bundles and vector bundles, which we introduced in part in Chapter 5, did not appear until the work of Whitney in 1935–1940.

Turning to physics, general relativity, one of the landmark achievements in science of the early 20th century, stands as the most visible application of Riemann manifolds to science. Starting from the principle that the speed of light in a vacuum is constant regardless of reference frame [20, p. 42], Einstein developed the theory of special relativity, defined in the absence of gravity. The "interpretation" of the law that "the gravitational mass of a body is equal to its inertial mass" [20, p. 65] and the intention to preserve the principle of the constancy of the speed of light led Einstein to understand spacetime as a curved space where "the geometrical properties of space are not independent, but [...] determined by matter" [20, p. 113]. Riemannian metrics, curvature, and the associated theorems for geodesics gave Einstein precisely the mathematical tools he needed to express his conception of a curvilinear spacetime.

The reader should be aware that other applications of manifolds to science do not (and should not) always require a metric structure. Applications of manifolds to either geometry or physics may require a different structure from or additional structure to a Riemann metric. For example, in its properly generalized context, Hamiltonian mechanics require the structure of what is called a symplectic manifold.

6.1 Riemannian Metrics

6.1.1 Definitions and Examples

Definition 6.1.1. Let M be a smooth manifold. A *Riemannian metric* on M is a tensor field g in $\mathrm{Sym}^2 TM^*$ that is positive definite. In more detail, at each point $p \in M$, a Riemannian metric determines an inner product on T_pM. A smooth manifold M together with a Riemannian metric g is called a *Riemannian manifold* and is denoted by the pair (M, g).

Over a coordinate patch of M with coordinate system $(x^1, \dots, x^n)$, as a section of $TM^{*\otimes 2}$, one writes the metric g as

$$g_{ij}\, dx^i \otimes dx^j,$$

where g_{ij} are smooth functions on M. (In this chapter, we regularly use Einstein's summation convention.) Since at each point, g is a symmetric tensor, $g_{ij} = g_{ji}$ identically. Furthermore, using the notation from Section 4.6, since g is a section of $\mathrm{Sym}^2 TM^*$, we write

$$g_{ij}\, dx^i\, dx^j. \tag{6.1}$$

The square root of the expression in Equation (6.1) is called the *line element* ds

associated to this metric. Many texts, in particular, physics texts, give the metric in reference to the line element by writing $ds^2 = g_{ij}\,dx^i\,dx^j$.

For vectors $X, Y \in T_pM$, we sometimes use the same notation as the first fundamental form ([6, Section 6.1]) and write $\langle X, Y\rangle_p$ for $g_p(X, Y)$, and it is also common to drop the subscript p whenever the point p is implied by context. By analogy with the dot product, the Riemannian metric allows one to define many common notions in geometry.

Definition 6.1.2. Let (M, g) be a Riemannian manifold. Suppose that X, Y are vectors in T_pM.

1. The *length* of X, denoted $\|X\|$, is defined by $\|X\| = \sqrt{g(X, X)}$.
2. The *angle* θ between X and Y is defined by
$$\cos\theta = \frac{g(X, Y)}{\|X\|\,\|Y\|}.$$
3. X and Y are called orthogonal if $g(X, Y) = 0$.

Whenever one introduces a new mathematical structure, one must discuss functions between any two instances of them and when two structures are considered equivalent. In the context of Riemannian manifolds, one still studies any smooth functions between two manifolds. However, two Riemannian manifolds are considered the same if they have the same metric. The following definition makes this precise.

Definition 6.1.3. Let M and N be two Riemannian manifolds. A diffeomorphism $f : M \to N$ is called an *isometry* if for all $p \in M$,

$$\langle X, Y\rangle_p = \langle df_p(X), df_p(Y)\rangle_{f(p)} \qquad \text{for all } X, Y \in T_pM.$$

Two Riemannian manifolds are called *isometric* if there exists an isometry between them.

From an intuitive perspective, an isometry is a transformation that bends (which also includes rigid motions) one manifold into another without stretching or cutting. Problem 6.1.6 asks the reader to show that the catenoid and the helicoid are isometric. Figure 6.1 shows intermediate stages of bending the catenoid into the helicoid. Though one might think this transformation incorporates some stretching because the longitudinal lines straighten out, the twist created in the helicoid strip "balances out" the flattening of the lines in just the right way so that one only needs to bend the surface.

Many examples of Riemannian metrics arise naturally as submanifolds of Riemannian manifolds.

Definition 6.1.4. Let $(N, \tilde{g})$ be a Riemannian manifold and M any smooth manifold. Let $f : M \to N$ be an immersion of M into N, i.e., f is differentiable and df_p

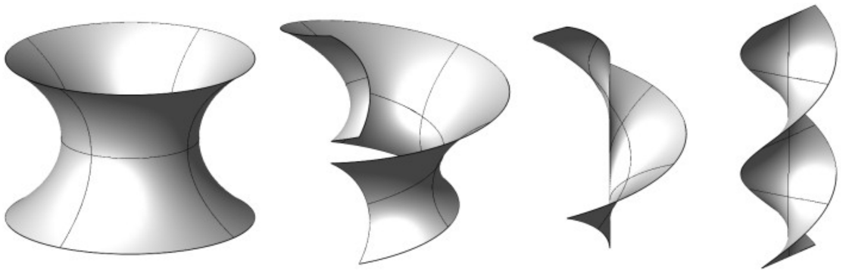

Figure 6.1: Bending the catenoid into the helicoid.

is injective for all p. The metric g on M *induced* by f (or "from N") is defined as the pull-back $g = f^*\tilde{g}$. In other words,

$$\langle X, Y\rangle_p = \langle df_p(X), df_p(Y)\rangle_{f(p)} \qquad \text{for all } p \in M \text{ and } X, Y \in T_pM.$$

The property that df_p is injective ensures that $\langle\, ,\, \rangle_p$ remains positive definite when induced on M.

Example 6.1.5 (Euclidean Spaces)**.** Consider the manifold $M = \mathbb{R}^n$, where $T_p(\mathbb{R}^n) = \mathbb{R}^n$ is naturally equipped with a Riemannian metric: the usual dot product. In particular,

$$g(\partial_i, \partial_j) = \delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$$

This metric is called the *Euclidean metric*.

Example 6.1.6 (First Fundamental Form)**.** A regular surface S is a 2-manifold embedded in $\mathbb{R}^3$, where the embedding map is simply the injection $i : S \to \mathbb{R}^3$. The first fundamental form (see Example 5.2.3) is precisely the metric on S induced by i from the Euclidean metric on $\mathbb{R}^3$. This connection gives us immediately a whole host of examples of Riemannian 2-manifolds that we take from the local theory of regular surfaces.

Proposition 6.1.7. *Let M be an m-dimensional manifold embedded in $\mathbb{R}^n$. If $\vec{F}(u_1, \dots, u_m)$ is a parametrization of a coordinate patch of M, then over this coordinate patch, the coefficients of the metric g on M induced from $\mathbb{R}^n$ are*

$$g_{ij} = \frac{\partial \vec{F}}{\partial u^i} \cdot \frac{\partial \vec{F}}{\partial u^j}.$$

Proof. Let $(x^1, \dots, x^n)$ be the coordinates on $\mathbb{R}^n$. Suppose that a coordinate patch (U, ϕ) of M has coordinates $(u_1, \dots, u_m)$ and that a parametrization of this coordinate patch is $\vec{F}(u_1, \dots, u_m) = \phi^{-1}(u_1, \dots, u_m)$. By Equation (3.14) the matrix of

$d\vec{F}$ in the given coordinate systems is

$$\left(\frac{\partial F^i}{\partial u^j}\right),$$

where $\vec{F} = (F^1, \dots, F^n)$.

Set $\langle\, , \rangle$ as the usual dot product in $\mathbb{R}^n$. Then at each point in U, the coefficients $g_{k\ell}$ of the metric g satisfy

$$\begin{aligned} g_{k\ell} &= g\left(\frac{\partial}{\partial u^k}, \frac{\partial}{\partial u^\ell}\right) = \left\langle d\vec{F}\left(\frac{\partial}{\partial u^k}\right), d\vec{F}\left(\frac{\partial}{\partial u^\ell}\right)\right\rangle = \frac{\partial F^i}{\partial u^k}\frac{\partial F^j}{\partial u^\ell}\left\langle \frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right\rangle \\ &= \frac{\partial F^i}{\partial u^k}\frac{\partial F^j}{\partial u^\ell}\delta_{ij} = \frac{\partial \vec{F}}{\partial u^k}\cdot\frac{\partial \vec{F}}{\partial u^\ell}. \end{aligned}$$ □

We should emphasize at this point that a given manifold can be equipped with nonisometric Riemannian metrics. Problem 6.1.2 presents two different metrics on the 3-torus, each depending on a different embedding into some Euclidean space. In both cases, the 3-torus can be equipped with the same atlas, and so in both situations, the 3-torus is the same as a smooth manifold.

As another example, already in his seminal dissertation [47], Riemann introduced the following metric on the open unit ball in $\mathbb{R}^n$:

$$g_{ii} = \frac{4}{(1 - \|x\|^2)^2} \qquad \text{and} \qquad g_{ij} = 0 \text{ if } i \neq j. \tag{6.2}$$

As we will see, this is not isometric with the open unit ball equipped with the Euclidean metric.

Example 6.1.5 could be misleading in its simplicity. The reader might consider the possibility of defining a metric on any smooth manifold M by taking $\langle\, , \rangle_p$ as the usual dot product in each T_pM with respect to the coordinate basis associated to a particular coordinate system. The problem with this idea is that it does not define a smooth section in $\mathrm{Sym}^2\, TM^*$ over the whole manifold. Nonetheless, as the proof of the following proposition shows, we can use a partition of unity and stitch these bilinear forms together.

Proposition 6.1.8. *Every smooth manifold M has a Riemannian metric.*

Proof. Let M be a smooth manifold with atlas $\mathcal{A} = \{(U_\alpha, \phi_\alpha)\}_{\alpha\in I}$. For each $\alpha \in I$, label $\langle\, , \rangle^\alpha$ as the usual dot product with respect to the coordinate basis over U_α. Let $\{\psi_\alpha\}$ be a partition of unity that is subordinate to $\mathcal{A}$. For each $p \in M$, define the bilinear form $\langle\, , \rangle_p$ on T_pM by

$$\langle X, Y\rangle_p \overset{\text{def}}{=} \sum_{\alpha\in I} \psi_\alpha(p)\langle X, Y\rangle_p^\alpha \tag{6.3}$$

for any $X, Y \in T_pM$.

Since for each $p \in M$, only a finite number of $\alpha \in I$ have $\psi_\alpha(p) \neq 0$, then the sum in Equation (6.3) is finite. It is obvious by construction that $\langle X, Y \rangle_p$ is symmetric. To prove that $\langle\, , \rangle_p$ is positive definite, note that each $\langle X, Y \rangle_p^\alpha$ is. Let I' be the set of all indices $\alpha \in I$ such that $\psi_\alpha(p) \neq 0$. By definition of a partition of unity, $0 < \psi_\alpha(p) \leq 1$. Thus, for all $X \in T_pM$, clearly $\langle X, X \rangle_p \geq 0$. Furthermore, if $\langle X, X \rangle_p = 0$, then at least one summand in

$$\sum_{\alpha \in I'} \psi_\alpha(p) \langle X, X \rangle_p^\alpha$$

is 0. (In fact, all summands are 0.) Thus, there exists an $\alpha \in I$ with $\langle X, X \rangle_p^\alpha = 0$. Since $\langle\, , \rangle_p^\alpha$ is positive definite, then $X = 0$. Hence, $\langle X, X \rangle_p$ itself is positive definite. □

Though Equation (6.3) presents a Riemannian metric on any smooth manifold M, this is not in general easy to work with for specific calculations since it uses a partition of unity, which involves functions that are usually complicated. At any given point $p \in M$, Equation (6.3) does not involve one coordinate system around p but all of the atlas's coordinate neighborhoods of p.

More importantly, the Riemannian metric constructed in the above proof might not have any natural meaning.

Example 6.1.9 (Projective Space)**.** There is a natural metric on projective space $\mathbb{RP}^n$ that is induced from the Euclidean metric on $\mathbb{R}^{n+1}$. First, let g be the metric on $\mathbb{S}^n$ induced from Euclidean $\mathbb{R}^n$ as an embedded submanifold. Recall that the projection map $\pi : \mathbb{S}^n \to \mathbb{RP}^n$ as presented for $n = 2$ in Example 3.2.3 is a smooth function between manifolds. Define the metric $\tilde{g}$ on $\mathbb{RP}^n$ by

$$\tilde{g}_{\pi(p)}(v, w) = g_p\left((d\pi_p)^{-1}(v), (d\pi_p)^{-1}(w)\right), \tag{6.4}$$

for all $v, w \in T_{\pi(p)}\mathbb{RP}^n$. Note first that for all $p \in \mathbb{S}^n$, the linear transformation $d\pi_p$ is surjective between spaces of the same dimension, so is invertible. More importantly, $\tilde{g}$ is well-defined: If $A : \mathbb{S}^n \to \mathbb{S}^n$ is the antipodal map $A(p) = -p$, then $\pi \circ A = \pi$. Hence, $d\pi_p = d(\pi \circ A)_p = d\pi_{-p} \circ dA_p$. Hence

$$\begin{aligned} g_p &\left((d\pi_p)^{-1}(v), (d\pi_p)^{-1}(w)\right) \\ &= g_p\left((dA_p)^{-1}((d\pi_{-p})^{-1}(v)), (dA_p)^{-1}((d\pi_{-p})^{-1}(w))\right) \\ &= g_{-p}((d\pi_{-p})^{-1}(v), (d\pi_{-p})^{-1}(w)), \end{aligned}$$

where the second equality follows because $A : \mathbb{S}^n \to \mathbb{S}^n$ is an isometry. Consequently, in (6.4), the choice of p or $-p$ for the pre-image of $\pi(p)$ in $\mathbb{RP}^n$ is irrelevant.

6.1.2 Arclength and Volume

Using integration, the Riemannian metric allows for formulas that measure nonlocal properties, such as length of a curve and volume of a region on a manifold.

For example, consider a C^1 curve $\gamma : [a,b] \to M$ on a Riemannian manifold (M,g). At each point $\gamma(t)$ in M, the vector $\gamma'(t) = \frac{d\gamma}{dt}$ is a tangent vector, called the *velocity vector*. The Riemannian metric $g = \langle\, , \rangle$ allows one to calculate the length $\|\gamma'(t)\|$, which we call the *speed*. This motivates the following definition.

Definition 6.1.10. Let $\gamma : [a,b] \to M$ be a curve on a Riemannian manifold M of class C^1. The arclength of the curve γ is

$$\ell(\gamma) = \int_a^b \sqrt{\left\langle \frac{d\gamma}{dt}, \frac{d\gamma}{dt} \right\rangle_{\gamma(t)}}\, dt.$$

Proposition 6.1.11. *Let (M,g) be an oriented Riemannian manifold of dimension n. There exists a unique n-form, denoted dV, such that at all $p \in M$, it satisfies $dV_p(e_1, \dots, e_n) = 1$ for all bases $(e_1, \dots, e_n)$ in T_pM that are orthonormal with respect to $g_p(\, ,)$. Furthermore, over any coordinate patch U with coordinates $x = (x^1, \dots, x^n)$,*

$$dV = \sqrt{\det(g_{ij})}\, dx^1 \wedge \cdots \wedge dx^n, \tag{6.5}$$

where $g_{ij} = g(\partial_i, \partial_j) = \langle \partial/\partial x^i, \partial/\partial x^j \rangle$.

Proof. The content of this proposition is primarily linear algebra. By Proposition C.2.1, on each coordinate patch U_α, the form $dV|_{U_\alpha} = \omega_\alpha$ exists on each T_pM and is given by Equation (6.5). The existence of this n-form ω with the desired property explicitly requires that g_p be an inner product on T_pM. In order to define the form dV on the whole manifold, we refer to a partition of unity $\{\psi_\alpha\}$ subordinate to the atlas on M and define

$$dV = \sum_\alpha \psi_\alpha \omega_\alpha. \qquad \square$$

Definition 6.1.12. The form dV described in Proposition 6.1.11 is called the *volume form* of (M,g) and is denoted dV_M if there is a chance of confusion about the manifold. If M^m is a compact manifold, then the m-volume of M is the integral

$$\operatorname{Vol}(M) = \int_M dV. \tag{6.6}$$

If $i : M^m \to N^n$ is an embedded submanifold of a Riemannian manifold $(N, \tilde{g})$ then we can also calculate the m-volume of the submanifold M by equipping M with the metric $g = i^*\tilde{g}$, i.e., the metric induced from N. The reader should be aware that the volume form on M is not necessarily $i^*(dV_N)$. In particular, if $m < n$, then $i^*(dV_N)$ would be an n-form on M, but there are no n-forms on M. We illustrate this with a few examples.

Example 6.1.13 (Arclength)**.** Definition 6.1.10 should actually be a corollary of Proposition 6.1.11 and Definition 6.1.12. Let $(M, \tilde{g})$ be a Riemannian manifold and let $\gamma : [a, b] \to M$ be a curve of class C^1 on M. The induced metric g on γ is defined by

$$g_t(v, w) = \tilde{g}_{\gamma(t)} \left(\frac{d\gamma}{dt} v, \frac{d\gamma}{dt} w \right).$$

However, since a curve is one dimensional, both tangent vectors v and w are scalars. So

$$g_t(v, w) = \tilde{g}_{\gamma(t)}((d\gamma/dt), (d\gamma/dt))vw.$$

In the coordinate t on $[a, b]$, the domain of γ, the metric tensor can be represented by a 1×1 matrix $(\tilde{g}_{\gamma(t)}((d\gamma/dt), (d\gamma/dt)))$. Thus, the volume form on γ from the metric induced from M, is

$$dV = \sqrt{\tilde{g}_{\gamma(t)}((d\gamma/dt), (d\gamma/dt))}\, dt.$$

Definition 6.1.10 follows as a corollary.

Example 6.1.14 (Volume form on $\mathbb{S}^n$)**.** Consider the unit n-sphere $\mathbb{S}^n$ as a Riemannian manifold, equipped with the metric induced from its usual embedding in $\mathbb{R}^{n+1}$.

Consider the usual longitude-latitude parametrization of the sphere $\mathbb{S}^2$:

$$\vec{X}(u, v) = (\cos u \sin v, \sin u \sin v, \cos v) \quad \text{for} \quad (u, v) \in [0, 2\pi] \times [0, \pi].$$

Note that if we restrict the domain to $(0, 2\pi) \times (0, \pi)$, we obtain a dense open subset of $\mathbb{S}^2$. By Proposition 6.1.7, with respect to this coordinate system, the coefficients of the metric tensor are

$$g_{ij} = \begin{pmatrix} \sin^2 v & 0 \\ 0 & 1 \end{pmatrix}.$$

Since $\sin v \geq 0$ for $v \in [0, \pi]$, the volume form on $\mathbb{S}^2$ with respect to this coordinate system is $dV = \sin v\, du \wedge dv$. By Proposition 5.8.1, the volume of the sphere is calculated by

$$V = \int_{\mathbb{S}^2} dV = \int_0^{\pi} \int_0^{2\pi} \sin v\, du\, dv = 2\pi \big[-\cos v \big]_0^{\pi} = 4\pi.$$

We now calculate the volume form on $\mathbb{S}^n$ using an alternate approach. By Example 4.6.24, we see that the volume form on $\mathbb{R}^{n+1}$ is

$$e^{*1} \wedge \cdots \wedge e^{*(n+1)},$$

where $\{e_1, \ldots, e_{n+1}\}$ is the standard basis on $\mathbb{R}^{n+1}$. Furthermore, recall that as an alternating function,

$$e^{*1} \wedge \cdots \wedge e^{*(n+1)}(v_1, \ldots, v_{n+1}) = \det \begin{pmatrix} v_1 & \cdots & v_{n+1} \end{pmatrix}$$

for any $(n+1)$-tuple of vectors $(v_1, \ldots, v_{n+1})$.

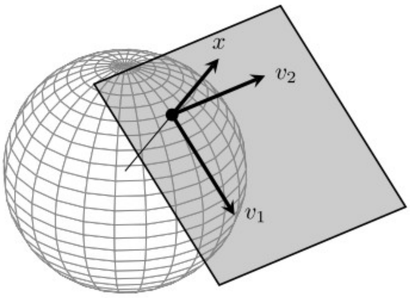

Figure 6.2: Volume form on the sphere.

Define a form $\omega \in \Omega^n(\mathbb{R}^{n+1})$, where for each $x \in \mathbb{R}^{n+1}$ and for any vectors u_i

$$\omega_x(u_1, \dots, u_n) = \det \begin{pmatrix} x & u_1 & \cdots & u_n \end{pmatrix}.$$

This is in fact, the same construction as the inner product on forms $i_x\omega$, but where we are taking advantage of the identification of $\mathbb{R}^{n+1}$ with its own tangent space. By the properties of the determinant, for each x, ω_x is an alternating n-multilinear function on $\mathbb{R}^{n+1}$, so ω is indeed an n-form. Using the Laplace expansion of the determinant, it is easy to show that

$$\omega = \sum_{i=1}^{n+1} (-1)^{i-1} x^i \, dx^1 \wedge \cdots \wedge \widehat{dx^i} \wedge \cdots \wedge dx^{n+1}, \tag{6.7}$$

where the $\widehat{\ }$ notation means to exclude the bracketed term. Using the forms η^j introduced in Equation (5.24), we can write $\omega = \sum_j x^j \eta^j$.

Now if $x \in \mathbb{S}^n$, then from the geometry of the sphere, x is perpendicular to $T_x\mathbb{S}^n$ as a subspace of $\mathbb{R}^{n+1}$ (equipped with the Euclidean metric). Thus, if $\{v_1, \dots, v_n\}$ forms a basis of $T_x\mathbb{S}^n$, then $\{x, v_1, \dots, v_n\}$ forms a basis of $\mathbb{R}^{n+1}$. See Figure 6.2. Furthermore, if the n-tuple $(v_1, \dots, v_n)$ is an orthonormal, positively-oriented basis of $T_x\mathbb{S}^n$, then $(x, v_1, \dots, v_n)$ is an orthonormal, positively-oriented basis of $\mathbb{R}^{n+1}$. But then the restriction of ω to $\mathbb{S}^n$ has the properties described in Proposition 6.1.11. Hence, if $f : \mathbb{S}^n \to \mathbb{R}^{n+1}$ is the usual embedding of the sphere in Euclidean space, then we obtain the volume form of $\mathbb{S}^n$ as

$$dV_{\mathbb{S}^n} = \omega\big|_{\mathbb{S}^n} = f^*(\omega).$$

In Section 5.7, we attempted to generalize with the single technique of integration on manifolds all the types of integration introduced in a standard calculus sequence. However, there were two types of integrals in the list at the beginning of the section that did not fit in the formalism we had developed for the integration of n-forms on n-dimensional smooth manifolds, namely:

- Line integrals of functions in $\mathbb{R}^n$.
- Surface integrals of a real-valued function defined over a closed and bounded region of a regular surface in $\mathbb{R}^3$.

Both of these types of integrals fit into the theory of integration volume forms on manifolds in the following ways.

For the line integral of functions in $\mathbb{R}^n$ over a piecewise-smooth curve C, let $\gamma : [a, b] \to \mathbb{R}^n$ be a parametrization of C. Let $\langle\, ,\, \rangle$ be the Euclidean form on $\mathbb{R}^n$ (i.e., the dot product). Then each smooth piece of γ is a one-dimensional submanifold of $\mathbb{R}^n$, equipped with the metric induced from $\mathbb{R}^n$. The volume form on γ is dV_γ so that for any smooth function f defined on a neighborhood of C,

$$\int_\gamma f\, dV_\gamma = \int_a^b f(t)\sqrt{\left\langle \frac{d\gamma}{dt}, \frac{d\gamma}{dt} \right\rangle_{\gamma(t)}}\, dt = \int_C f\, ds. \tag{6.8}$$

For surface integrals of a function f on a compact regular surface $\mathcal{S} \subset \mathbb{R}^3$, it is not hard to show (see Problem 6.1.1) that $dS = dV_{\mathcal{S}}$, where $dV_{\mathcal{S}}$ is the volume form on $\mathcal{S}$ equipped with the metric induced from the Euclidean metric. Thus, connecting the classical notation with the notation introduced in this section,

$$\int_{\mathcal{S}} f\, dS = \int_{\mathcal{S}} f\, dV_{\mathcal{S}}. \tag{6.9}$$

6.1.3 Raising and Lowering Indices

One interesting property of metrics on manifolds is that they give us a natural way to go back and forth between vectors and covectors.

Recall from Section 4.1 that for any real vector space V, the dual vector space V^* consists of all linear transformations $V \to \mathbb{R}$. If V is a vector space equipped with any bilinear form $\langle\, ,\, \rangle$, then this form defines a linear transformation into the dual V^* by

$$\begin{aligned} i : V &\longrightarrow V^*, \\ v &\longmapsto i_v \quad = (w \mapsto \langle v, w\rangle). \end{aligned}$$

If V is finite-dimensional and the bilinear form $\langle\, ,\, \rangle$ is nondegenerate, then the mapping $(v \mapsto i_v)$ is an isomorphism. We will assume from now on that $\langle\, ,\, \rangle$ is an inner product. The positive-definite implies the bilinear form is nondegenerate so the above mapping is an isomorphism. This isomorphism allows us to define a natural bilinear form $\langle\, ,\, \rangle^*$ on V^* by

$$\langle \eta, \tau\rangle^* = \langle i^{-1}(\eta), i^{-1}(\tau)\rangle.$$

Then $\langle\, ,\, \rangle^*$ is an element of $V \otimes V$, so is a tensor of type of $(2, 0)$. Furthermore, if the components of $\langle\, ,\, \rangle$ are c_{jk}, then the components of $\langle\, ,\, \rangle^*$ with respect to the associated basis are denoted by c^{jk}, where these are the entries of the inverse matrix C^{-1}, where $C = (c_{jk})$. (See Exercises 4.2.4 and 4.2.5 for where we prove these results.)

In coordinates, let $\mathcal{B} = \{u_1, \dots, u_n\}$ be a basis of V and let $\mathcal{B}^* = \{u^{*1}, \dots, u^{*n}\}$ be the associated cobasis. Let $C = (c_{jk})$ be the matrix of $\langle\, , \,\rangle$ with respect to $\mathcal{B}$, viz., $c_{jk} = \langle u_j, u_k\rangle$. Then

$$\langle v, w\rangle = [v]_{\mathcal{B}}^T C[w]_{\mathcal{B}} = c_{jk} v^j w^k. \tag{6.10}$$

This gives the coordinates of i_v with respect to $\mathcal{B}^*$ as $c_{jk}v^j$. Note that the indices for the components of i_v arise naturally as subscripts, consistent with our notation. Similarly, if $\lambda \in V^*$ is a functional on V, then $\lambda = \lambda_i u^{*i}$. If $\lambda = i_v$ for some $v \in V$, then

$$c^{k\ell}\lambda_k = c^{k\ell} a_{jk} v^j = c_{jk} a^{k\ell} v^j = \delta^\ell_j v^j = v^\ell. \tag{6.11}$$

We say that the process of mapping v to i_v "lowers the indices," while mapping λ to $i^{-1}(\lambda)$ "raises the indices."

Now consider a Riemannian manifold (M, g). If $X \in \mathfrak{X}(M)$ is a vector field on M, we define the covector field $X^\flat$ by

$$X^\flat(Y) \stackrel{\text{def}}{=} g(X, Y)$$

for all vector fields $Y \in \mathfrak{X}(M)$. On a coordinate patch, X has coordinates X^i. By the process described in the previous paragraphs, $X^\flat$ has components $g_{ij}X^i$. Mimicking musical notation, the function $\Gamma(TM) \to \Gamma(TM^*)$ that sends $X \mapsto X^\flat$ is call the *flat*, since it lowers the indices of the vector field X.

As we saw, a metric g^* on M also induces an inner product in $\operatorname{Sym}^2 T_pM$ defined by $g^*(\eta, \tau) = g(i^{-1}(\eta), i^{-1}(\tau))$ for any $\eta, \tau \in T_pM^*$. So if $\omega \in \Omega^1(M)$ is a covector field on M, we define the vector field $\omega^\sharp$ by

$$\eta(\omega^\sharp) \stackrel{\text{def}}{=} g^*(\eta, \omega)$$

for all covector fields $\eta \in \Omega^1(M)$. On a given coordinate patch, ω has component functions ω_i and the components of $\omega^\sharp$ are $g^{ij}\omega_i$. Keeping the musical analogy, we denote call this vector field the *sharp* of ω since the process raises the indices.

More generally, if T is any tensor field of type (p, q) on M, then

$$g_{i_\alpha k} T^{i_1\cdots i_p}_{j_1\cdots j_q} \qquad \text{and} \qquad g^{j_\beta l} T^{i_1\cdots i_p}_{j_1\cdots j_q}$$

define tensors fields of type $(p-1, q+1)$ and type $(p+1, q-1)$, respectively. It is common to still use the $\flat$ and $\sharp$ notation, here $T^\flat$ and $T^\sharp$, but one must indicate upon which index one performs the lowering or raising operations.

Recall that the trace of a matrix A is defined as the sum of the diagonal elements. If A has components A^i_j, then the trace is just A^i_i, using the Einstein summation convention of summing along i. Now A corresponds to a linear transformation

$T(\vec{v}) = A\vec{v}$ on a vector space V. Since the trace $\operatorname{Tr} A$ is also the sum of the eigenvalues, the trace remains unchanged under a change in basis in V.

Now, if A is a symmetric $(0,2)$-tensor, then $A^\sharp$ is a $(1,1)$-tensor, and the trace $\operatorname{Tr} A$ is defined in its usual linear algebraic sense. This process is common enough that we define the *trace with respect to* g of A to be

$$\operatorname{Tr}_g A \overset{\text{def}}{=} \operatorname{Tr} A^\sharp. \tag{6.12}$$

In coordinates, $\operatorname{Tr}_g A = g^{ij} A_{ij}$.

PROBLEMS

6.1.1. Recall the following formula from calculus. Let K be a compact set on a regular surface S in $\mathbb{R}^3$. Suppose that S is parametrized by $\vec{r}(u,v)$ and that under this parametrization, $K = \vec{R}(D)$ for some compact region D. Then the surface area of K on S is

$$\iint_K dS = \iint_D \|\vec{r}_u \times \vec{r}_v\| \, du \, dv.$$

Consider the regular surface S as a 2-manifold embedded in the Riemannian manifold of $\mathbb{R}^3$, equipped with the dot product as its usual metric. The parametrization $\vec{r}(u,v)$ describes the embedding of S in $\mathbb{R}^3$ in reference to the coordinates in a chart of S. Prove that the volume form on S of metric on S induced from the dot product in $\mathbb{R}^3$ is

$$dV_S = \|\vec{r}_u \times \vec{r}_v\| \, du \wedge dv.$$

This establishes the familiar surface area integral from Definition 6.1.12.

6.1.2. Consider the 3-torus $\mathbb{T}^3 = \mathbb{S}^1 \times \mathbb{S}^1 \times \mathbb{S}^1$. Calculate the induced metrics for the following two embeddings:

(a) Into $\mathbb{R}^6$ as the image of the parametrization

$$\vec{F}(u^1,u^2,u^3) = \left(\cos u^1, \sin u^1, \cos u^2, \sin u^2, \cos u^3, \sin u^3\right).$$

(b) Into $\mathbb{R}^4$ as the image of the parametrization $\vec{F}(u^1,u^2,u^3)$ given by

$$\Big((c + (b + a\cos u^1)\cos u^2)\cos u^3, (c + (b + a\cos u^1)\cos u^2)\sin u^3,$$
$$(b + a\cos u^1)\sin u^2, a\sin u^1\Big),$$

where $b > a > 0$ and $c > a + b$.

(c) Prove that these two Riemannian manifolds are not isometric.

[Hint: This gives two different metrics on the 3-torus that can be equipped with the same atlas in each case.]

6.1.3. We consider an embedding of $\mathbb{S}^1 \times \mathbb{S}^2$ in $\mathbb{R}^4$ by analogy with the embedding of the torus $\mathbb{S}^1 \times \mathbb{S}^1$ in $\mathbb{R}^3$. Place the sphere $\mathbb{S}^2$ with radius a at $(0,0,b,0)$ (where $b > a$) as a subset of the $x^1x^2x^3$-subspace, and rotate this sphere about the origin with a motion parallel to the x^3x^4-axis. Call this submanifold M and equip it with the metric induced from $\mathbb{R}^4$.

(a) Show that the described manifold M is an embedding as claimed.

(b) Find a parametrization $F : D \to \mathbb{R}^4$ where $D \subset \mathbb{R}^3$ such that, as sets of points $F(D) = M$, and $F(D^\circ)$ is an open subset of M that is homeomorphic to D°. (D° is the interior of the set D.)

(c) Calculate the coefficients g_{ij} of the metric on M in the coordinate patch defined by the above parametrization.

6.1.4. Let (M_1, g_1) and (M_2, g_2) be two Riemannian manifolds, and consider the product manifold $M_1 \times M_2$ with a $(0, 2)$-tensor field defined by

$$g(X_1 + X_2, Y_1 + Y_2) = g_1(X_1, Y_1) + g_2(X_2, Y_2).$$

(a) Show that g defines a metric on $M_1 \times M_2$.

(b) Let $(x^1, \ldots, x^n)$ be local coordinates on M_1 and $(x^{n+1}, \ldots, x^{n+m})$ be local coordinates on M_2 so that $(x^1, \ldots, x^{m+n})$ are local coordinates on $M_1 \times M_2$. Determine the components of the metric g on $M_1 \times M_2$ in terms of g_1 and g_2.

6.1.5. Repeat Problem 6.1.3 with $\mathbb{S}^1 \times S$, where S is a regular surface in $\mathbb{R}^3$ that does not intersect the plane $z = x^3 = 0$.

6.1.6. Consider the following two regular surfaces in $\mathbb{R}^3$. The catenoid parametrized by

$$F(\bar{u}^1, \bar{u}^2) = (\bar{u}^2 \cos \bar{u}^1, \bar{u}^2 \sin \bar{u}^1, \cosh^{-1} \bar{u}^2) \text{ for } (\bar{u}^1, \bar{u}^2) \in [0, 2\pi) \times [1, +\infty)$$

and the helicoid parametrized by

$$F(u^1, u^2) = (u^2 \cos u^1, u^2 \sin u^1, u^1) \text{ for } (u^1, u^2) \in [0, 2\pi) \times \mathbb{R}.$$

Prove that the helicoid and catenoid are isometric, and find an isometry between them.

6.1.7. Let M be a hypersurface of $\mathbb{R}^n$ (submanifold of dimension $n-1$), and equip M with Riemannian structure with the metric induced from $\mathbb{R}^n$. Suppose that an open set U of M is a graph of an $(n-1)$-variable function f, i.e., the parametrization of U is

$$x^1 = u^1, \ldots, x^{n-1} = u^{n-1}, x^n = f(u^1, \ldots, u^{n-1}), .$$

for $(u^1, \ldots, u^{n-1}) \in D$.

(a) Find the coefficients of the metric tensor g on M, and conclude that a formula for the $(n-1)$-volume of U is

$$\int_D \sqrt{1 + \|\text{grad } f\|^2} \, dV.$$

(b) Use this result to calculate the 3-volume of the surface in $\mathbb{R}^4$ given by $w = x^2 + y^2 + z^2$ for $x^2 + y^2 + z^2 \leq 4$.

6.1.8. Let M, N, and S be Riemannian manifolds, and let $f : M \to N$ and $h : N \to S$ be isometries.

(a) Show that f^{-1} is an isometry.

(b) Show that $h \circ f$ is an isometry.

(c) If you have seen some group theory, show that the set of isometries on a Riemannian manifold M forms a group.

6.1.9. Two metrics g and $\tilde{g}$ on a smooth manifold M are called *conformal* if there exists a smooth function $f \in C^{\infty}(M, \mathbb{R})$ such that $\tilde{g} = fg$. Prove that for all $p \in M$ and for all $X, Y \in T_pM$, the angle between X and Y with respect to $\tilde{g}$ is the same as the angle with respect to g.

6.1.10. Let (M, g) and $(N, \tilde{g})$ be Reimannian manifolds. A diffeomorphism $f : M \to N$ is called a conformal mapping if $f^*\tilde{g}$ is conformal to g. Repeat Problem 6.1.8 but replacing "isometry" with "conformal mapping." (See Problem 6.1.9.)

6.1.11. Let γ be a curve on a Riemannian manifold (M, g). Show precisely how the induced metric on γ generalizes Definition 6.1.10.

6.1.12. *Poincaré ball.* The Poincaré Ball is the open ball $\mathbb{B}^n_R$ in n dimensions of radius R equipped with the metric

$$\frac{4R^4}{(R^2 - \|x\|^2)^2} \left((dx^1)^2 + \cdots + (dx^n)^2 \right).$$

Note that this metric is conformal with the metric (see Problem 6.1.9) induced from the Euclidean metric in $\mathbb{R}^{n+1}$.

(a) Set $n = 2$ and $R = 1$. (This choice of parameters is called the unit Poincaré disk.) Calculate the area of the region $\mathcal{R}$ defined by $\|x\| \leq \frac{1}{2}$ and $0 \leq \theta \leq \pi/2$.

(b) Set $n = 3$ and $R = 2$. Calculate the length of the curve $\gamma(t) = (\cos t, \sin t, t/10)$ in the Poincaré ball for $0 \leq t \leq 4\pi$.

6.1.13. *Divergence Theorem.* Let (M, g) be an oriented, compact, Riemannian manifold with boundary. Given any vector field $X \in \mathfrak{X}(M)$ and any tensor field T of type (p, q), with $q \geq 1$, we define the contraction of X with T, denoted $i_X T$, as the tensor field of type $(p, q-1)$ that over any coordinate chart has components

$$X^l T^{i_1 \cdots i_p}_{l j_2 \cdots j_q}.$$

We define the *divergence operator* $\operatorname{div} : \mathfrak{X}(M) \to C^{\infty}(M)$ implicitly by

$$d(i_X \, dV) = (\operatorname{div} X) \, dV.$$

Prove the Divergence Theorem, which states that for any $X \in \mathfrak{X}(M)$

$$\int_M \operatorname{div} X \, dV = \int_{\partial M} g(X, N) \, d\tilde{V},$$

where N is the outward unit normal to ∂M and $d\tilde{V}$ is the volume form associated to the metric on ∂M induced from M.

6.1.14. We consider the sphere $\mathbb{S}^3$ of radius R as a submanifold in $\mathbb{R}^4$ with the induced Euclidean metric.

(a) Show that

$$F(u^1, u^2, u^3) = (R\cos u^1 \sin u^2 \sin u^3, R\sin u^1 \sin u^2 \sin u^3, R\cos u^2 \sin u^3, R\cos u^3),$$

where $(u^1, u^2, u^3) \in [0, 2\pi] \times [0, \pi]^2$ gives a parametrization for $\mathbb{S}^3$ that is homeomorphic to its image when restricted to the open set $V = (0, 2\pi) \times (0, \pi)^2$.

(b) Calculate the components of the metric tensor on the coordinate patch $F(V) = U$.

(c) Use part (b) to calculate the volume of a 3-sphere of radius R.

(d) Leaving R unspecified, consider the function $f(x^1, x^2, x^3, x^4) = (x^1)^2 + (x^2)^2 + (x^3)^2$ and calculate the volume integral $\int_{\mathbb{S}^3} f\, dV$. (Note that this integral would give the radius of gyration of the spherical shell of radius R about a principal axis – if such a thing existed in $\mathbb{R}^4$!)

6.1.15. Calculate the 5-volume of the 5-sphere $\mathbb{S}^5$ of radius R as a submanifold of $\mathbb{R}^6$.

6.1.16. (**ODE**) A loxodrome on the unit sphere $\mathbb{S}^2$ is a curve that makes a constant angle with all meridian lines. We propose to study analogues of loxodromes on $\mathbb{S}^3$. Consider the unit 3-sphere $\mathbb{S}^3$ with the parametrization from Problem 6.1.14. Set $R = 1$. We will call a loxodrome on $\mathbb{S}^3$ any curve γ such that γ' makes a constant angle of α_2 with $\frac{\partial}{\partial u^2}$ and a constant angle of α_3 with $\frac{\partial}{\partial u^3}$.

(a) Find equations that the components of γ must satisfy.

(b) Solve the differential equations we get in part (a). [Hint: Obtain u^1 and u^2 as functions of u^3. You might only be able to obtain one of these functions implicitly.]

6.1.17. Consider the function $r : \mathbb{R}^{n+1} - \{0\} \to \mathbb{S}^n$ given by $r(x) = x/\|x\|$.

(a) Using Example 6.1.14, prove that

$$\tilde{\omega} = r^*(dV_{\mathbb{S}^n}) = \frac{1}{\|x\|^{n+1}} \sum_{j=1}^{n+1} x^j \eta^j.$$

(b) Show that $\tilde{\omega}$ is closed but not exact in $\mathbb{R}^{n+1} - \{0\}$.

(c) Use $dV_{\mathbb{S}^n}$ to show that

$$\mathrm{Vol}(\mathbb{S}^n) = (n+1)\mathrm{Vol}(\mathbb{B}^{n+1}),$$

where $\mathbb{B}^{n+1}$ is the unit ball in $\mathbb{R}^{n+1}$.

6.1.18. Let M be a Riemannian manifold, and let $f : M \to M$ be an isometry on M. Prove that $f^*(dV_M) = \pm dV_M$. (The isometry f is called *orientation-preserving* if $f^*(dV_M) = dV_M$ and *orientation-reversing* if $f^*(dV_M) = -dV_M$.)

6.1.19. Suppose that J and K are disjoint, compact, oriented, connected, smooth submanifolds of $\mathbb{R}^{n+1}$ whose dimensions are greater than 0 and such that $\dim J + \dim K = n$. Define the function Ψ by

$$\begin{aligned} \Psi : J \times K &\longrightarrow \mathbb{S}^n \\ (x, y) &\longmapsto \frac{y - x}{\|y - x\|}. \end{aligned}$$

The *linking number* between J and K is defined as

$$\text{link}(J, K) = \frac{1}{\text{Vol}(\mathbb{S}^n)} \int_{J \times K} \Psi^*(dV_{\mathbb{S}^n}).$$

Prove Gauss's Linking Formula for the linking number of two closed space curves:

$$\text{link}(C_1, C_2) = \frac{1}{4\pi} \int_I \int_J \frac{\det(\vec{\alpha}(u) - \vec{\beta}(v), \vec{\alpha}'(u), \vec{\beta}'(v))}{\|\vec{\alpha}(u) - \vec{\beta}(v)\|^3} \, du \, dv. \tag{6.13}$$

[Note that a closed curve is homeomorphic to a circle $\mathbb{S}^1$ so $J \times K$ is homeomorphic to a torus. Hence, we can view Ψ as a function $\mathbb{T}^2 = \mathbb{S}^1 \times \mathbb{S}^1 \to \mathbb{S}^2$.]

The following exercises involve the Hodge star operation, which is introduced in Appendix C.3.

6.1.20. Let (M, g) be an oriented Riemannian manifold. Section C.3 defines the Hodge star operator on inner product spaces. Given a form $\eta \in \Omega^k(M)$, at each $p \in M$, the Hodge star operator defines an isomorphism $\star : \bigwedge^k T_pM^* \to \bigwedge^{n-k} T_pM^*$.

(a) Show that the Hodge star operator $\star$ is a vector bundle map $\bigwedge^k TM^* \to \bigwedge^{n-k} TM^*$ that leaves every base point fixed and that varies smoothly.

(b) Show that for all functions $f : C^\infty(M)$, the Hodge star operator is given by $\star f = f dV_g$.

6.1.21. Consider $\mathbb{R}^n$ as a manifold with the standard Euclidean metric.

(a) Calculate $\star dx^i$ for any $i = 1, \dots, n$.

(b) Set $n = 4$, and calculate $\star(dx^i \wedge dx^j)$.

6.1.22. Let (M, g) be an oriented Riemannian manifold. Prove the following identities for any vector field $X \in \mathfrak{X}(M)$.

(a) $\text{div}\, X = \star d \star X^\flat$, where div is the divergence operator defined in Problem 6.1.13.

(b) $i_X dV_g = \star X^\flat$.

6.1.23. Let (M, g) be a Riemannian manifold. Consider the operation that consists of $\star d \star d$.

(a) Show that $\star d \star d$ is an $\mathbb{R}$-linear operator $\Omega^k(M) \to \Omega^k(M)$ for $k < \dim M$.

(b) Let $M\mathbb{R}^n$ be a standard Euclidean space. Recalling that $\Omega^0(M) = C^\infty(M)$, show that for any smooth function f,

$$\star d \star df = \nabla^2 f,$$

where ∇^2 is the usual Laplacian $\nabla^2 = \dfrac{\partial^2}{\partial(x^1)^2} + \cdots + \dfrac{\partial^2}{\partial(x^n)^2}$.

(c) Find an expression in coordinates of $\star d \star df$ for smooth functions $f : \mathbb{S}^2 \to \mathbb{R}$, express in (u, v) longitude-latitude coordinates as used in Example 6.1.14.

6.2 Connections and Covariant Differentiation

6.2.1 Motivation

Despite all the "heavy machinery" we have developed in order to create a theory of analysis on manifolds, we are still unable to calculate or even define certain things that are simple in $\mathbb{R}^n$.

For example, if $\gamma : [a, b] \to M$ is a smooth curve on a smooth manifold, we have no way at present to talk about the acceleration of γ. Let $p \in M$, with $\gamma(t_0) = p$. In Definition 3.3.1, we presented the tangent vector $\gamma'(t_0)$ at p as the operator $D_\gamma : C^1(M) \to \mathbb{R}$ that evaluates

$$D_\gamma(f) = \frac{d}{dt} f(\gamma(t))\Big|_{t_0}.$$

In Section 3.3, we developed the linear algebra of expressions D_γ for curves γ through p. The vector space of such operators is what we called the tangent space T_pM.

Mimicking what one does in standard calculus, one could try to define the acceleration vector $\gamma''(t_0)$ at p as a limiting ratio as $t \to t_0$ of $\gamma'(t) - \gamma'(t_0)$ with $t - t_0$. However, what we just wrote does not make sense in the context of manifolds because $\gamma'(t)$ and $\gamma'(t_0)$ are not even in the same vector spaces and so their difference is not defined.

Another attempt to define the acceleration might follow Definition 3.3.1 and try to define $\gamma''(t_0)$ at p as the operator $D^{(2)}_\gamma : C^2(M) \to \mathbb{R}$, where

$$D^{(2)}_\gamma(f) = \frac{d^2}{dt^2} f(\gamma(t))\big|_{t_0}.$$

This operator is well defined and linear. However, $D^{(2)}_\gamma$ does not satisfy Leibniz's rule, and therefore, there does not exist another curve $\tilde{\gamma}$, with $\tilde{\gamma}(t_0) = p$ such that $D^{(2)}_\gamma = D_{\tilde{\gamma}}$. Hence, $D^{(2)}_\gamma \notin T_pM$. We could study properties for operators of the form $D^{(2)}_\gamma$ but, since the operators do not exist in any $TM^{\otimes p} \otimes TM^{*\otimes q}$, this is not the direction the theory of manifolds developed.

Another lack in our current theory is the ability to take partial derivatives or, more generally, directional derivatives of a vector field. Over $\mathbb{R}^n$, it is easy to define $\partial\vec{F}/\partial x^j$, where $\vec{F}$ is a vector field, and, under suitable differentiability conditions, $\partial\vec{F}/\partial x^j$ is again another vector field. In contrast, if X is a vector field over a smooth manifold M and U is a coordinate neighborhood of $p \in M$, one encounters the same problem with defining a vector $\partial_i X_p$ as one does in defining the acceleration of a curve.

) &%(&!& * H ' " 3 0 , 7

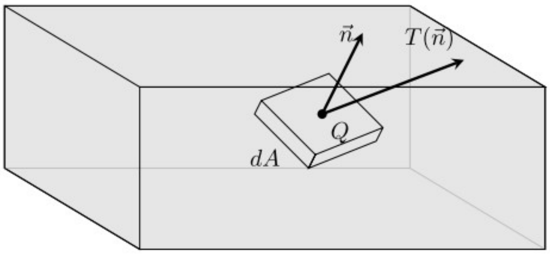

Figure 7.10: Stress tensor: an area element in a continuous medium.

7.5 Brief Introduction to General Relativity

As with the previous sections, the reader might consider it outlandish that we only allow one section to discuss general relativity. General relativity is a vast subject with contributions from a host of scientists and mathematicians, and it stands alongside quantum mechanics as one of the most revolutionary ideas in physics of the 20th century.

On the other hand, most textbooks on general relativity take a considerable amount of time to develop the techniques of analysis on manifolds, in particular, pseudo-Riemannian manifolds. However, these are precisely the mathematical methods we have developed in the previous chapters, so we are in a position to introduce some differential geometric concepts in general relativity as applications.

From the perspective of mathematical structures, general relativity builds on special relativity. The postulates of special relativity brought us to the notion of spacetime, which is a Minkowski space $\mathbb{R}^{3,1}$. In general relativity, we will want to consider our space as *locally Minkowski*, meaning that each tangent space is a Minkowski space. Hence, we model the universe as a pseudo-Riemannian manifold with signature $(3, 1)$. More importantly, the Einstein field equations propose a relationship between the presence of energy and the curvature of this spacetime manifold.

7.5.1 Stress-Energy Tensor

In the mechanics of elastic media, one encounters the concept of a stress tensor, which is a tensor-valued function defined at each point within the body or medium. Suppose the body is in equilibrium but subject to external forces and/or body forces (i.e., forces that act through the whole body). Then there must exist internal forces. Let Q be a point, $\vec{n}$ a vector based at Q, and consider the area element ΔA that is in the plane perpendicular to $\vec{n}$ and has area equal to $\|\vec{n}\|$ (see Figure 7.10). Let $\Delta\vec{F}$ be the overall internal forces distributed over the area element ΔA. The stress vector through the area element ΔA is the vector

$$\vec{T}(\vec{n}) = \lim_{\Delta A \to 0} \frac{\Delta \vec{F}}{\Delta A}. \tag{7.63}$$

It is not hard to show that the function $\vec{T}(\vec{n})$ is a linear function.[56, Section 10.6] Thus the stress tensor with respect to an orthogonal basis $\mathcal{B}$ based at the point Q is the matrix σ such that

$$\vec{T}(\vec{n}) = \sigma \big[\vec{n}\big]_{\mathcal{B}}.$$

Consider now a small rectangular parallelepiped with sides parallel to the coordinate planes. The stress acts on each face as depicted in Figure 7.11. Then the columns of σ are given by

$$\sigma \vec{e}_i = T(\vec{e}_i) = \begin{pmatrix} \sigma_{i1} \\ \sigma_{i2} \\ \sigma_{i3} \end{pmatrix}.$$

Minimal assumptions that are logical for physics (angular momentum in a medium cannot grow to be infinite at a point) imply that the stress tensor σ is symmetric.

As a simple example, in an ideal fluid, the stress on any small area element is composed only of pressure, and there is no shearing force. Consequently, the stress tensor is $\sigma = P\mathbf{I}$, where P is the pressure and $\mathbf{I}$ is the 3×3 identity matrix. (This restates the claim given in calculus texts on the applications of integration to hydrostatics when one says that "at any point in a liquid the pressure is the same in all directions."[55, p. 576]) The stress tensor arises also in the dynamics of viscous fluids where it is no longer necessarily diagonal. The stress tensor at a point "may be a function of the density and temperature, of the relative positions and velocities of elements near [the point], and perhaps also the previous history of the medium."[56, p. 434] This characterization describes the stress tensor as a function of many ambient quantities, but the reference to "relative positions" indicates that the stress tensor need not be diagonal.

Einstein's equation in general relativity involves the so-called *stress-energy tensor*. This tensor is different from the stress tensor but is based on the same concept. We assume that we are in a Minkowski space with metric $\mathbf{g}$ of signature $(3, 1)$.

As in special relativity, the four-velocity of a particle on a world line $\mathcal{P}$ parametrized by $\vec{x}$ is the tangent vector along $\mathcal{P}$ given by

$$\vec{U} = \frac{d\vec{x}}{d\tau}, \tag{7.64}$$

where, the proper time τ of an object along its world line is given in (7.20). From the theory of special relativity, by (7.22), the four-velocity is

$$\vec{U} = (U^0, U^1, U^2, U^3) = (\gamma c, \gamma v_x, \gamma v_y, \gamma v_z), \tag{7.65}$$

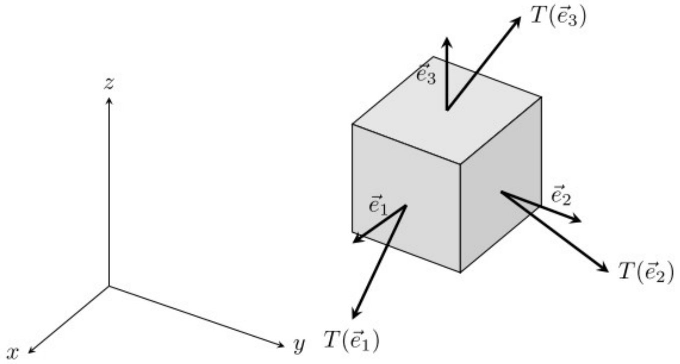

Figure 7.11: Action of stress on an infinitesimal coordinate cube.

where $\gamma = (1 - v^2/c^2)^{-1/2}$ and $\vec{v} = (dx/dt, dy/dt, dz/dt)$. In (7.23), we defined the momentum 4-vector of a particle of rest mass m by $\vec{p} = m\vec{u}$. Equally useful is the four-momentum covector given by

$$\mathbf{p} = \vec{p}^{\flat}. \tag{7.66}$$

In special relativity where the metric $\mathbf{g} = \eta$ is the standard Minkowski metric, the components of the momentum covector are

$$(p_0, p_1, p_2, p_3) = \left(-\frac{E}{c}, p_x, p_y, p_z \right), \tag{7.67}$$

where $(p_x, p_y, p_z) = m_0\gamma(v^1, v^2, v^3)$ is the relativistic 3-vector momentum.

Underlying the assumptions that define the stress-energy tensor, we assume that "spacetime contains a flowing river of 4-momentum"[41, p.130]. Any mass that is moving or anything with energy contributes to the 4-momentum. We could think of an individual particle, in which case the 4-momentum would only be defined on the particle's world line, or we could consider a system of many particles carrying this 4-momentum. In the latter case, we should think of the 4-momentum as a covector field on the spacetime manifold M, that is, as a 1-form.

Let $\mathbf{n}$ be any 1-form on M. Then at each point $Q \in M$, $\mathbf{n}_Q$ is perpendicular (using $\mathbf{g} = \langle \, , \rangle$ at Q) to a three-dimensional subspace of $T_Q M$. This subspace can be spanned by vectors A_Q, B_Q, and C_Q such that

$$\mathbf{n}_Q(u) = -\mathrm{Vol}_{\mathbf{g}}(u, A_Q, B_Q, C_Q), \tag{7.68}$$

where on the right-hand side we mean the 4-volume (with respect to $\mathbf{g}$) of the 4-parallelepiped spanned by $u, A_Q, B_Q, C_Q \in T_Q M$. Then the 3-volume of the 3-

parallelepiped A_Q, B_Q, C_Q is the length $\|\mathbf{n}\|$. In this way, every 1-form represents a volume element in $\mathbb{R}^4$.

We define the *stress-energy tensor* $\mathbf{T}$ of type $(2, 0)$ by $\mathbf{T}(\mathbf{p}, \mathbf{n})$ being the flux of the momentum covector $\mathbf{p}$ across a volume element represented by $\mathbf{n}$. At each point $Q \in M$, we have

$$\mathbf{T}(\mathbf{p}, \mathbf{n}) = \langle \mathbf{p}, \mathbf{n} \rangle_{\mathbf{g}}. \tag{7.69}$$

Over any coordinate chart (x^i), the components of $\mathbf{T}$ are

$$T^{\alpha\beta} = \mathbf{T}(dx^\alpha, dx^\beta)$$

to represent that flux of unit four-momentum in direction dx^α across a volume-element of constant β.

The stress-energy tensor $\mathbf{T}$ is also called the *energy-momentum tensor* because it contains information pertaining to the momentum flowing through space and the presence of static or moving energy in space. The name "stress-energy tensor" is commonly used since it is modeled off the stress tensor in mechanics of elastic media.

The following gives a summary of the information included in the stress-energy tensor. Assuming $j, k > 0$,

$$T^{00} = \text{density of energy (including mass)}, \tag{7.70}$$

$$T^{j0} = j\text{th component of the momentum density}, \tag{7.71}$$

$$T^{0k} = k\text{th component of the energy flux}, \tag{7.72}$$

$$\begin{aligned} T^{jk} &= (j,k)\text{th component of the momentum stress} \\ &= k\text{th component of the flux of the } j\text{th component of momentum.} \end{aligned} \tag{7.73}$$

The notion of *flux* in this context refers to a similar limit as in Equation (7.63) but in the situation where one is concerned with the movement of something (energy, fluid momentum, heat,...) through the infinitessimal area element $d\vec{A}$. In fact, with this particular concept of flux, one can define the stress-energy tensor in short by saying that T^{kj} is the kth component of the flux of the jth component of the 4-momentum.

We now state two facts about the stress-energy tensor that we do not fully justify here.

Proposition 7.5.1. *The (contravariant) stress-energy tensor* $\mathbf{T}$ *is symmetric.*

The symmetry in the components $1 \leq i, j \leq 3$ follows from the same physical reasoning for why the stress tensor in fluid dynamics is symmetric.

Proposition 7.5.2 (Einstein's Conservation Law). *The conservation of energy is equivalent to the identity*

$$\operatorname{div} \mathbf{T} = T^{\alpha\beta}{}_{;\beta} = 0.$$

Proof Sketch. Suppose that energy is conserved in a certain region of M. In other words, though energy and mass may move around, no energy or mass is created or annihilated in M. Then given any four-dimensional submanifold $\mathcal{V}$ with boundary $\partial\mathcal{V}$, the total flux of 4-momentum passing through $\partial\mathcal{V}$ must be 0. We can restate this as

$$\iiint_{\partial\mathcal{V}} \mathbf{T} \cdot d\mathbf{V}^{(3)} = 0,$$

where $d\mathbf{V}^{(3)}$ is the 3-volume element with direction along the outward-pointing normal vector to $\partial\mathcal{V}$. (We can view this as a volume 1-form.) By the product $\cdot$ we mean the contraction of $\mathbf{T}$ with the volume 1-form element $d\mathbf{V}^{(3)}$. Stokes' Theorem applied to pseudo-Riemannian manifolds gives

$$\iiiint_{\mathcal{V}} \operatorname{div} \mathbf{T} dV^{(4)} = \iiint_{\partial\mathcal{V}} \mathbf{T} \cdot d\mathbf{V}^{(3)} = 0.$$

Since this is true for all $\mathcal{V}$ as described above, using a limiting argument similar to that used to show Gauss' Law that $\operatorname{div} \vec{E} = 0$ for an electric field, a limiting argument establishes $\operatorname{div} \mathbf{T} = 0$ everywhere. ☐

Example 7.5.3 (Perfect Fluid Stress-Energy Tensor)**.** A perfect fluid is a fluid in which the pressure p is the same in any direction. The fluid must be free of heat conduction and viscosity and any process that can cause internal sheers. Using the interpretation of $\mathbf{T}$ from Equations (7.70)–(7.73), we see that $T^{jk} = 0$ if $j \neq k$ and $0 < j, k$. Furthermore, since the pressure is the same in all directions, $T^{jj} = p$ for $j = 1, 2, 3$. For components involving $j = 0$ or $k = 0$, we first have $T^{00} = \rho$, the energy density. This quantity includes mass density but also other types of energy such as compression energy. For the remaining off diagonal terms $T^{0j} = T^{j0}$, these are 0 because of the assumption that there is no heat conduction in the perfect fluid. Thus, the stress-energy tensor has components

$$T^{\alpha\beta} = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & p & 0 & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & p \end{pmatrix}. \tag{7.74}$$

If we suppose that an observer is in the Lorentz frame that is at rest with respect to the movement of the fluid, then the velocity has components $u^\alpha = (1, 0, 0, 0)$. With respect to the Minkowski metric η, we can write (7.74) as

$$T^{\alpha\beta} = (\rho + p)u^\alpha u^\beta + p\eta^{\alpha\beta}.$$

We can rewrite this in a coordinate-free way in any metric as

$$\mathbf{T} = p\mathbf{g}^{-1} + (p + \rho)\mathbf{u} \otimes \mathbf{u}, \tag{7.75}$$

where we have written $\mathbf{g}^{-1}$ for the contravariant tensor of type $(2, 0)$ associated to the metric tensor $\mathbf{g}$.

Example 7.5.4 (Electromagnetic Stress-Energy Tensor)**.** Directly using the interpretation of $\mathbf{T}$ given in Equations (7.70)–(7.73) and results from electromagnetism, which we do not recreate here, one can determine the components of the stress-energy tensor for the electromagnetic field in free space. If $F^{\mu\nu}$ are the components of the electromagnetic field tensor, then

$$\begin{aligned} T^{\alpha\beta} &= \frac{1}{\mu_0}\left(F^{\alpha\mu}g_{\mu\nu}F^{\beta\nu} - \frac{1}{4}g^{\alpha\beta}F^{\mu\nu}F_{\mu\nu}\right) \qquad \text{in SI units} \qquad (7.76)\\ &= \frac{1}{4\pi}\left(F^{\alpha\mu}g_{\mu\nu}F^{\beta\nu} - \frac{1}{4}g^{\alpha\beta}F^{\mu\nu}F_{\mu\nu}\right) \qquad \text{in CGS units} \qquad (7.77)\end{aligned}$$

where $\mu_0 = 4\pi \times 10^{-7}\,\mathrm{N/A^{-2}}$ is a constant sometimes called the vacuum permeability.

As the context requires, we may need the stress-energy tensor to be of type $(2, 0)$ as defined, of type $(1, 1)$ or of type $(0, 2)$. To pass between any of these, we raise or lower the indices as needed using the metric.

7.5.2 Einstein Field Equations

The Einstein field equations (EFE) are the heart of general relativity. They stem from the juxtaposition of the two following principles:

1. Every aspect of gravity is a description of the spacetime geometry.
2. Mass (energy) is the source of gravity.

The metric tensor $\mathbf{g}$ encapsulates all the information about the geometry of the spacetime. The metric $\mathbf{g}$ has the associated Levi-Civita connection ∇, the Riemann curvature tensor $\mathbf{R}$ of type $(1, 3)$, the Ricci curvature tensor $\mathbf{Rc}$, and the scalar curvature function R, defined in Chapter 6. (Note: In math texts on Riemannian geometry, one often denotes by S the scalar curvature while texts on general relativity invariably denote it by R. Using the bold font $\mathbf{R}$ to indicate the curvature tensor alleviates any confusion between the scalar and tensor curvature.)

On the other hand, the stress-energy tensor describes the spacetime content of mass-energy. In fact, any observer with 4-velocity $\vec{U}$ measures the density of mass-energy as

$$\rho = \mathbf{u} \cdot \mathbf{T} \cdot \mathbf{u} = T_{\alpha\beta}u^{\alpha}u^{\beta}.$$

In order to put together the two above principles, we should be able to write the tensor $\mathbf{T}$ exclusively in terms of the components of the metric tensor $\mathbf{g}$. The conservation of energy states that $\operatorname{div}\mathbf{T} = 0$. Also, if $\mathbf{T}$ is to serve as a measure of the curvature of spacetime, we propose that it should explicitly involve only components of $\mathbf{R}$ and of $\mathbf{g}$ (no derivatives of any of these terms) and it should be linear in the components of $\mathbf{R}$. It turns out that under these restrictions, there are only a few options for a geometric description of $\mathbf{T}$. In Problem 7.5.2, we show that for purely mathematical reasons, these constraints impose that

$$T_{\alpha\beta} = C\left(R_{\alpha\beta} - \frac{1}{2}Rg_{\alpha\beta} + \Lambda g_{\alpha\beta}\right), \tag{7.78}$$

where $R_{\alpha\beta}$ are the components of the Ricci curvature tensor, R is the scalar curvature, and Λ and C are real constants. This leads to Einstein's field equations.

Let $\mathbf{G}$ be the Einstein curvature tensor described in Definition 6.5.4. General relativity is summarized in this following equation. The presence of mass-energy deforms spacetime according to

$$\mathbf{G} + \Lambda\mathbf{g} = \frac{8\pi G}{c^4}\mathbf{T} \qquad \text{in SI units,} \tag{7.79}$$

where $G = 6.67 \times 10^{-11}\,\mathrm{m}^3\mathrm{s}^{-2}\mathrm{kg}^{-1}$ is the gravity constant and Λ is the *cosmological constant*. If we assume that empty (devoid of energy) spacetime is flat, then

$$\mathbf{G} = \frac{8\pi G}{c^4}\mathbf{T}. \tag{7.80}$$

Equation (7.80) is called collectively the *Einstein field equations* (EFE), and the formulas in (7.79) are the Einstein field equations with cosmological constant. These equations are as important in astrophysics as Newton's second law of motion is in classic mechanics. Though we do not give the calculation here, the constant $8\pi G/c^4$, called *Einstein's constant* is chosen so that (1) when the gravitational field is weak and (2) velocities are small compare to the speed of light, the theory reduces to the Newtonian theory of gravitation in approximation.

In Equation (2) of Einstein's original paper on general relativity [19], Einstein made the assumption that $\mathbf{G}$ vanishes when spacetime is empty of mass-energy. This corresponds to the mathematical assumption that $\Lambda = 0$. However, Equation (7.80) predicts a dynamic universe. This result did not appeal to Einstein and, at the time, there existed no astronomical evidence to support this. In 1917, he introduced the constant Λ because it allows for a static universe. Physically, $\Lambda \neq 0$ would imply the presence of an otherwise unexplained force that counteracts gravity or a sort of negative pressure.

When Hubble discovered that the universe is expanding, the cosmological constant no longer appeared to be necessary and many physicists did away with it. In fact, in his autobiography, George Gamow relays that Einstein told Gamow that he considered the introduction of the cosmological constant as "the biggest blunder of my life." [26] However, the possibility of a small nonzero Λ has resurfaced and regularly enters into the conversations around the current most vexing problems in physics, namely, the nature of dark energy and the effort to unify gravity and quantum mechanics.

We should note the Einstein field equations (EFE) are very complicated. Finding a solution to the EFE means finding the metric tensor $\mathbf{g}$ that satisfies (7.80), which consists of 10 second-order, nonlinear, partial differential equations of 10 functions $g_{ij}(x^0, x^1, x^2, x^3)$, with $0 \leq i \leq j \leq 3$. Then, determining the trajectory of a particle or of radiation amounts to determining the geodesics in this metric. Surprisingly, under some circumstances, especially scenarios that involve a high level of symmetry, it is possible to provide an exact solution.

Whole books have been written about consequences of solutions to (7.80) or (7.79) that deviate from Newtonian mechanics. After the introduction of general relativity, some scientists balked at such mathematical complexity. Yet experiment has repeatedly confirmed predictions in favor of general relativity over Newtonian mechanics. The approach taken here to justify the Einstein field equations cited mathematical esthetics, the condition of being divergence-free, and linear in the components of the curvature tensor **R**. During the 20th century, scientists arrived at (7.79) from other, more physical principles. Of note, in [36], Lovelock showed that the Einstein field equations arise as the unique second-order equations that can follow from the Euler-Lagrange equations of a Lagrange density involving g_{ij} and its derivatives up to second order. Furthermore, there exist natural generalizations to Einstein's theory of general relativity. However, any theory that can unify gravity and quantum mechanics should be able to derive (7.79) as an approximation.

7.5.3 Schwarzschild Metric

We finish this section with one of the earliest proven consequences of general relativity, i.e., the Schwarzschild metric which is an exact solution to Einstein's field equations. Instead of simply showing that the Schwarzschild metric satisfies EFE, we show how the metric was discovered.

One of the main contexts in which one can expect to see the effects of general relativity against Newtonian mechanics is in the context of astronomy. The simplest dynamical problem in astronomy involves calculating the orbit of a single planet around the sun. One can hope that the EFE for the effect of the sun on the space around it will become simple under the following two assumptions (approximations):

1. The sun is a spherically symmetric distribution of mass-energy density.
2. Outside of the sun, the stress-energy tensor should vanish.

The spherical symmetry implies that the components of the metric tensor should be given as functions of x^0 and r alone, where $r^2 = (x^1)^2 + (x^2)^2 + (x^3)^2$. Since we are looking only for solutions outside the sun, we are looking for solutions in a vacuum. Thus $\mathbf{T} = 0$, from which we deduce that $\mathbf{G} = 0$. Thus, $\operatorname{Tr}_{\mathbf{g}} \mathbf{G} = 0$. However,

$$\operatorname{Tr}_{\mathbf{g}} \mathbf{G} = \operatorname{Tr}_{\mathbf{g}} \left(\mathbf{Rc} - \frac{1}{2} R \mathbf{g} \right) = R - \frac{1}{2} R \cdot 4 = -R,$$

where this follows from (6.56) and the fact that $\operatorname{Tr}_{\mathbf{g}} \mathbf{g} = \dim M = 4$. Thus, $R = 0$ and the fact that $\mathbf{G} = 0$ implies that we are looking for spherically symmetric solutions to the equation

$$R_{\alpha\beta} = 0. \tag{7.81}$$

Since we are looking for solutions in a vacuum, it seems as though we have lost information, but, as we shall see, that is not the case.

The following derivation follows the treatment in [54]. We leave some of the details as exercises for the reader.

A judicious choice of coordinates and a few coordinate transformations will simplify the problem. We first start with the coordinates

$$(\bar{x}^0, x^1, x^2, x^3) = (\bar{x}^0, \bar{r}, \theta, \varphi),$$

where $\bar{x}^0 = ct$ for some timelike variable t, $\bar{r}^2 = (x^1)^2 + (x^2)^2 + (x^3)^2$ and where θ and φ are given in the physics style of defining spherical coordinates, i.e., so that φ is the longitudinal angle and θ is the latitude angle measured down from a "positive" vertical direction. We know that the standard line element in spherical coordinates is

$$ds^2 = d\bar{r}^2 + \bar{r}^2 d\theta^2 + \bar{r}^2 \sin^2\theta d\varphi^2.$$

Though we are not working with the Euclidean metric, spherical symmetry does imply that the metric tensor in the space coordinates is orthogonal and that no perpendicular direction to the radial direction is singled out. Thus, the metric tensor has the form

$$g_{\alpha\beta} = \begin{pmatrix} g_{00}(\bar{x}^0, \bar{r}) & g_{01}(\bar{x}^0, \bar{r}) & g_{02}(\bar{x}^0, \bar{r}) & g_{03}(\bar{x}^0, \bar{r}) \\ g_{10}(\bar{x}^0, \bar{r}) & g_{11}(\bar{x}^0, \bar{r}) & 0 & 0 \\ g_{20}(\bar{x}^0, \bar{r}) & 0 & f(\bar{x}^0, \bar{r})^2 & 0 \\ g_{30}(\bar{x}^0, \bar{r}) & 0 & 0 & f(\bar{x}^0, \bar{r})^2 \sin^2\theta \end{pmatrix}, \tag{7.82}$$

where f is any smooth function. We actually have some choice on θ and φ because they are usually given in reference to some preferred x-axis and z-axis. We choose θ and φ (which may change over time with respect to some fixed Cartesian frame) so that $g_{20} = g_{30} = 0$, and then the metric looks like

$$g_{\alpha\beta} = \begin{pmatrix} g_{00}(\bar{x}^0, \bar{r}) & g_{01}(\bar{x}^0, \bar{r}) & 0 & 0 \\ g_{10}(\bar{x}^0, \bar{r}) & g_{11}(\bar{x}^0, \bar{r}) & 0 & 0 \\ 0 & 0 & f(\bar{x}^0, \bar{r})^2 & 0 \\ 0 & 0 & 0 & f(\bar{x}^0, \bar{r})^2 \sin^2\theta \end{pmatrix}. \tag{7.83}$$

We make the coordinate transformation $r = f(\bar{x}^0, \bar{r})$ and all the other coordinates remain the same. In this coordinate system, the metric looks like

$$g_{\alpha\beta} = \begin{pmatrix} g_{00}(\bar{x}^0, r) & g_{01}(\bar{x}^0, r) & 0 & 0 \\ g_{10}(\bar{x}^0, r) & g_{11}(\bar{x}^0, r) & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & r^2 \sin^2\theta \end{pmatrix}. \tag{7.84}$$

Finally, we can orthogonalize the metric tensor by a suitable coordinate transformation of $x^0 = ct = h(\bar{x}^0, r)$, with r staying fixed (see Problem 7.5.4). Since we know that the metric has signature $(3,1)$, we can write the metric in the coordinate system $(x^0, r, \theta, \varphi)$ as

$$g_{\alpha\beta} = \begin{pmatrix} -e^{\nu(x^0,r)} & 0 & 0 & 0 \\ 0 & e^{\lambda(x^0,r)} & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & r^2 \sin^2\theta \end{pmatrix}, \tag{7.85}$$

where λ and ν are smooth functions. The metric in Equation (7.85) is an orthogonal metric that is spherically symmetric in the space variables.

Using the notation $\dot{u} = \partial u/\partial x^0$ and $u' = \partial u/\partial r$, we can show (see Problem 7.5.5) that the independent nonzero Christoffel symbols for the Levi-Civita connection are

$$\Gamma^0_{00} = \frac{1}{2}\dot{\nu}, \qquad \Gamma^0_{01} = \frac{1}{2}\nu', \qquad \Gamma^0_{11} = \frac{1}{2}\dot{\lambda}e^{\lambda-\nu}, \quad \Gamma^1_{00} = \frac{1}{2}\nu' e^{\nu-\lambda}, \tag{7.86}$$

$$\Gamma^1_{01} = \frac{1}{2}\dot{\lambda}, \qquad \Gamma^1_{11} = \frac{1}{2}\lambda', \qquad \Gamma^1_{22} = -re^{-\lambda}, \quad \Gamma^1_{33} = -r\sin^2\theta e^{-\lambda}, \tag{7.87}$$

$$\Gamma^2_{12} = \frac{1}{r}, \qquad \Gamma^2_{33} = -\sin\theta\cos\theta, \quad \Gamma^3_{13} = \frac{1}{r}, \qquad \Gamma^3_{23} = \cot\theta. \tag{7.88}$$

Though it is a little long to calculate (see Problem 7.5.6), we then determine that the only nonzero components of the Ricci tensor are

$$\begin{aligned} R_{00} &= e^{\nu-\lambda}\Big(\frac{\nu''}{2} + \frac{(\nu')^2}{4} - \frac{\nu'\lambda'}{4} + \frac{\nu'}{r}\Big) - \frac{\ddot{\lambda}}{2} - \frac{\dot{\lambda}^2}{4} + \frac{\dot{\lambda}\dot{\nu}}{4}, \\ R_{01} &= R_{10} = \frac{\dot{\lambda}}{r}, \\ R_{11} &= -\frac{\nu''}{2} - \frac{(\nu')^2}{4} + \frac{\nu'\lambda'}{4} + \frac{\lambda'}{r} + e^{\lambda-\nu}\Big(\frac{\ddot{\lambda}}{2} + \frac{\dot{\lambda}^2}{4} - \frac{\dot{\lambda}\dot{\nu}}{4}\Big), \\ R_{22} &= -e^{-\lambda}\Big(1 + \frac{r}{2}(\nu' - \lambda')\Big) + 1, \\ R_{33} &= \sin^2\theta\, R_{22}. \end{aligned} \tag{7.89}$$

Since we are trying to solve $R_{\alpha\beta} = 0$, we obtain conditions on the functions λ and ν. Since $R_{01} = 0$, we deduce immediately that $\dot{\lambda} = 0$, which means that λ is a function of r alone. Also, since $\partial R_{22}/\partial t = 0$, we find that $\partial \nu'/\partial t = 0$. Therefore, we can write the function ν as

$$\nu = \nu(r) + f(t)$$

for some function $f(t)$.

We now make one final coordinate change. In the metric line element, t appears only in the summand $e^{\nu} d(x^0)^2 = e^{\nu(r)} e^{f(t)} d(ct)^2$. So by choosing the variable $\bar{t}$ in such a way that

$$\frac{d\bar{t}}{dt} = e^{f(t)/2}$$

and then renaming $\bar{t}$ to just t, we obtain a metric which is independent of any timelike variable. (The variable $\bar{t}$, relabeled as t, is not necessarily time anymore so we cannot necessarily call a solution with $t = 0$ as a solution static.)

We can now assume there is no t dependence. Simplifying the expression $R_{00} + e^{\nu-\lambda}R_{11}$ leads to

$$\frac{1}{r}(\lambda' + \nu') = 0.$$

This implies that $\lambda(r) = -\nu(r) + C$ for some constant C. Without loss of generality, we can assume that $C = 0$ since we have not specified λ or ν. Thus, we set $\lambda(r) = -\nu(r)$. Then $R_{22} = 0$ in (7.89) implies that

$$e^{-\lambda}(1 - r\lambda') = 1.$$

Now setting $h(r) = e^{-\lambda(r)}$, this last equation becomes

$$h' + \frac{h}{r} = \frac{1}{r}.$$

This is a linear, first-order, ordinary, differential equation whose general solution is

$$h(r) = e^{-\lambda(r)} = 1 - \frac{2M}{r},$$

where M is a constant of integration. One can verify directly that $R_{11} = R_{00} = 0$ in Equation (7.89) are satisfied by this solution and therefore give no additional conditions. Therefore, the spherically symmetric vacuum solution to the EFE gives a metric with line element

$$ds^2 = -\left(1 - \frac{2M}{r}\right)c^2 dt^2 + \left(1 - \frac{2M}{r}\right)^{-1} dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2. \qquad (7.90)$$

This is called the Schwarzschild metric. This metric provided the first exact solution to the Einstein field equations. Though it is still complicated, this metric can be compared in fundamental importance to the solution in mechanics to the differential equations $d^2\vec{x}/dt^2 = -m\vec{g}$. Many of the verifiable predictions of general relativity arise from this metric.

In order to understand (7.90), we need to have some sense of the meaning of the constant M. Obviously, if $M = 0$, then the Schwarzschild metric is simply the flat Minkowski metric for spacetime.

To derive an interpretation for $M \neq 0$, we study some consequences of (7.90) for small velocities. If the velocity v is much smaller than the speed of light, i.e., $v \ll c$, then special relativity tells us that proper time is approximately coordinate time $\tau \cong t = x^0/c$. Furthermore, from (7.65) we can approximate the velocity of any particle as $\vec{U} = (c, 0, 0, 0)$. Plugging these into the geodesic equation

$$\frac{d^2x^i}{d\tau^2} = -\Gamma^i_{jk}\frac{dx^j}{d\tau}\frac{dx^k}{d\tau},$$

we obtain the approximate relationship

$$\frac{d^2x^i}{dt^2} = -\Gamma^i_{00}c^2 = \frac{c^2}{2}g^{il}\frac{\partial g_{00}}{\partial x^l}, \tag{7.91}$$

where the second equality follows from the formula for Christoffel symbols and the fact that the functions g_{ij} are not x^0 dependent. However, since $\mathbf{g}$ is diagonal and g_{00} depends only on r, we find that the only nonzero derivative is

$$\frac{d^2r}{dt^2} = -\frac{c^2}{2}\left(1 - \frac{2M}{r}\right)\frac{\partial}{\partial r}\left(1 - \frac{2M}{r}\right) = -\frac{Mc^2}{r^2}\left(1 - \frac{2M}{r}\right).$$

We must compare this to the formula for gravitational attraction in Newtonian mechanics, namely,

$$\frac{d^2r}{dt^2} = -\frac{GM_S}{r^2}$$

where M_S is the mass of the attracting body (and G is the gravitational constant). Thus, we find as a first approximation that the constant of integration M is

$$M \cong \frac{GM_S}{c^2}. \tag{7.92}$$

Hence, M is a constant multiple of the mass of the attracting body. The constant $2M$ has the dimensions of length, and one calls $r_G = 2M$ the *Schwarzschild radius*. The formula for it is

$$r_G = \frac{2G}{c^2} \cdot M_S = (1.48 \times 10^{-27}\ \text{m/kg})M_S.$$

The Schwarzschild radius is 2.95 km for the sun and 8.8 mm for the Earth. Evidently, for spherically symmetric objects that one encounters in common experience, the Schwarzschild radius is much smaller than the object's actual radius. In fact, if a spherically symmetric object has radius R_S and mass M_S, then

$$r_G < R_S \Longleftrightarrow \frac{2G}{c^2}M_S < R_S.$$

A sphere with $r_G > R_S$ would need to have an enormous density. Furthermore, this situation would seem to be physically impossible for the following reason. It is understood that the Schwarzschild metric holds only in the vacuum outside of the body (planet or star). However, if $r_G > R_S$, then the Schwarzschild radius would correspond to a sphere outside of the spherical body where the Schwarzschild metric has a singularity $g_{11} = 1/0$. For this reason, some physicists initially claimed this to be a result of the successive approximations or simply a physically impossible situation.

The history of science has occasionally shown that singularities in the equations do not immediately imply that the scenario is impossible. The possibility of traveling at the speed of sound was thought to be impossible because of the consequences for the Doppler effect equation. Now, military jets regularly fly faster than the speed of sound. Similarly, in recent decades, physicists regularly study objects considered to be so dense that $2GM_S/c^2 > R_S$. Such objects are called *black holes*. For a time, the existence of black holes remained in the realm of hypothesis, but now astronomers are convinced they have observed many such objects, and astrophysicists have worked out many of their dynamic properties.

PROBLEMS

7.5.1. Show that we can rephrase the explanation for Equation (7.68) by saying that for any three vectors A, B, and C in $T_P M$,

$$(\star \mathbf{n}_P)(A, B, C) = \mathrm{Vol}_{\mathbf{g}}(\mathbf{n}_P^\sharp, A, B, C),$$

where $\mathrm{Vol}_{\mathbf{g}}$ is the volume form with respect to the metric $\mathbf{g}$.

7.5.2. Let $\mathbf{L}$ be a symmetric tensor of type $(0, 2)$ consisting of components that are constructible from those of $\mathbf{R}$ and Tg and are linear in the components of the Riemann curvature tensor $\mathbf{R}$.

(a) Show that $\mathbf{L}$ can only have the form

$$L_{\alpha\beta} = aR_{\alpha\beta} + bRg_{\alpha\beta} + \lambda g_{\alpha\beta},$$

where $R_{\alpha\beta}$ are the components of the Ricci curvature tensor, R is the scalar curvature, and a, b, and λ are real constants. [Hint: Consider Bianchi identities.]

(b) Show that $\operatorname{div} \mathbf{L} = 0$ if and only if $b = -\frac{1}{2}a$.

(c) If $\mathbf{g} = \eta$, the standard Minkowski metric, show that $\mathbf{L} = 0$ if and only if $\lambda = 0$.

7.5.3. Calculate the curvature tensor and the Ricci curvature tensor for the Schwarzschild metric.

7.5.4. Find the "suitable" coordinate transformation h that allows one to pass from Equation (7.84) to Equation (7.85).

7.5.5. Prove that Equation (7.87) is correct.

7.5.6. Prove Equation (7.89).

7.5.7. *Light Propagation in the Schwarzschild Metric.* In the Schwarzschild metric, light travels along the null-geodesics, i.e., where $ds^2 = 0$.

(a) Explain why setting $\theta = 0$ does not lose any generality to finding the null-geodesics.

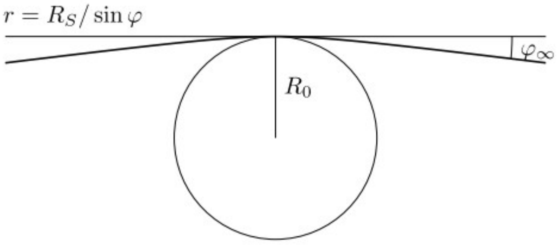

Figure 7.12: Deviation of light near a massive body.

(b) Prove that if one sets $u = 1/r$, then $ds^2 = 0$ implies that

$$\frac{d^2u}{d\varphi^2} + u = 3Mu^2. \tag{7.93}$$

(c) Deduce that in the vicinity of a black hole, light travels in a circle precisely at the radius $r = \frac{3}{2}r_G$.

(d) Solve Equation (7.93) for $M = 0$. This corresponds to empty space (no mass present). Call this solution $u_0(\varphi)$.

(e) Now look for general solutions u to Equation (7.93) by setting $u = u_1 + u_0$. Then $u_1(\varphi)$ must satisfy

$$\frac{d^2u_1}{d\varphi^2} + u_1 = \frac{3M}{R_0}\sin^2(\varphi - \varphi_0).$$

Solve this differential equation explicitly, and find the complete solution to Equation (7.93).

(f) In the complete solution, show that $M = 0$ (empty space) corresponds to traveling along a straight line $u = \frac{1}{R_0}\sin(\varphi - \varphi_0)$, where R_0 is the distance from the line to the origin.

(g) Show that the general solution to Equation (7.93) is asymptotically a line.

(h) We now consider Eddington's famous experiment to measure the deviation of light by the sun. Consider a geodesic $\mathcal{G}$ in the Schwarzschild metric that passes right alongside the sun, i.e., passes through the point $r = R_S$ and $\varphi = 0$. Define φ_∞ as the limiting angle of deviation between the line $r = R_S/\sin\varphi$ and the geodesic $\mathcal{G}$ (see Figure 7.12). The sun bends the light away from the straight line by a total of $2\varphi_\infty$. Using (at a judicious point) the approximation that $\sin\varphi \cong \varphi$, prove that the total deviation of light is

$$2\varphi_\infty = \frac{4GM_S}{R_S c^2},$$

where M_S is the mass of the sun and R_S is the sun's radius.

2> '6 G , , >"7

7.5.8. Consider the metric with line element $ds^2 = -e^{2ax}dt^2 + dx^2 + dy^2 + dz^2$.

(a) From the geodesic equation associated to this metric, show that at every point in this spacetime, a free particle experiences an acceleration of $d^2x/d\tau^2 = -a$.

(b) Show that the only nonzero Christoffel symbols for this metric are $\Gamma^0_{10} = \Gamma^0_{01} = a$ and $\Gamma^1_{00} = ae^{2ax}$.

(c) Show that the only nonzero components of the $(0,4)$ curvature tensor are R_{0101} and its permutations. Calculate R_{0101}.

(d) Find the Ricci curvature tensor and notice that it is diagonal.

2> ' 6 G , , > "7

7.5.8. Consider the metric with line element $ds^2 = -e^{2ax}dt^2 + dx^2 + dy^2 + dz^2$.

(a) From the geodesic equation associated to this metric, show that at every point in this spacetime, a free particle experiences an acceleration of $d^2x/d\tau^2 = -a$.

(b) Show that the only nonzero Christoffel symbols for this metric are $\Gamma^0_{10} = \Gamma^0_{01} = a$ and $\Gamma^1_{00} = ae^{2ax}$.

(c) Show that the only nonzero components of the $(0,4)$ curvature tensor are R_{0101} and its permutations. Calculate R_{0101}.

(d) Find the Ricci curvature tensor and notice that it is diagonal.

) &!& &*(

APPENDIX A

Point Set Topology

Though mathematicians, when developing a new area of mathematics, may define and study any object as they choose, the "natural" notion of a surface in $\mathbb{R}^3$ requires a rather intricate definition (Definition 3.1.1). Though at first somewhat unwieldy, this definition and also the definition for a differentiable manifold are necessary to appropriately generalize calculus and geometry to non-Euclidean spaces.

On the other hand, numerous concepts from geometry and calculus can be generalized not by formulating more constrained definitions but by expanding the context in which we define these concepts. The first wider context presented in this appendix is that of a metric space, a set equipped with some notion of distance. Many concepts from Euclidean geometry, including continuity, have natural generalizations to metric spaces. As it turns out, many useful concepts for analysis, like continuity, limit of sequences, or connectedness, arise in the yet more general context of topological spaces, where instead of a distance function, we have a looser notion of "nearness."

Though topology is a vast branch of mathematics, this appendix presents just the basic notions that support this book's presentation of differential geometry. A reader might encounter many of these concepts in a typical analysis course. We refer the reader to [27] for a gentle but thorough introduction to point set topology and to [43] and [2] for an introduction to topology that includes homology, the fundamental group, algebraic topology, and the classification of surfaces.

A.1 Metric Spaces

A.1.1 Metric Spaces: Definition

A metric space is a set that comes with the notion of "distance" between two points, where this distance function involves a few numerical conditions that mimic geometry in Euclidean spaces.

Definition A.1.1. Let X be any set. A *metric* on X is a function $D : X \times X \to \mathbb{R}^{\geq 0}$ such that

1. equality: $D(x, y) = 0$ if and only if $x = y$;
2. symmetry: $D(x, y) = D(y, x)$ for all $x, y \in X$;
3. triangle inequality: $D(x, y) + D(y, z) \geq D(x, z)$ for all $x, y, z \in X$.

A pair (X, D) where X is a set with a metric D is called a *metric space*.

Example A.1.2 (Euclidean Spaces). The Euclidean space $\mathbb{R}^n$ is a metric space where D is the usual Euclidean distance formula between two points, namely, if $P = (p_1, p_2, \ldots, p_n)$ and $Q = (q_1, q_2, \ldots, q_n)$, then

$$D(P, Q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Many notions in usual geometry (circles, parallelism, midpoint...) depend vitally on this particular distance formula. Furthermore, if $n = 1$, this formula simplifies to the usual distance formula on the real line $\mathbb{R}$, namely,

$$d(x, y) = |y - x|.$$

To prove that $(\mathbb{R}^n, D)$ is indeed a metric space, we must verify the three axioms in Definition A.1.1. The first holds because

$$D(P, Q) = 0 \Longleftrightarrow \sum_{i=1}^{n} (q_i - p_i)^2 = 0,$$

which is equivalent to $(q_i - p_i)^2 = 0$ for all $1 \leq i \leq n$, and hence $q_i = p_i$ for all $1 \leq i \leq n$. The second obviously holds, and we prove the third axiom as follows. The Cauchy-Schwarz inequality on the vectors $\overrightarrow{PQ}$ and $\overrightarrow{QR}$ gives

$$\overrightarrow{PQ} \cdot \overrightarrow{QR} \leq \overrightarrow{PQ} \cdot \overrightarrow{QR}| \leq \|\overrightarrow{PQ}\| \, \|\overrightarrow{QR}\|$$

so

$$2 \sum_{i=1}^{n} (q_i - p_i)(r_i - q_i) \leq 2 \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \sqrt{\sum_{i=1}^{n} (r_i - q_i)^2}.$$

Using the property that $2ab = (a + b)^2 - a^2 - b^2$ with $a = q_i - p_i$ and $b = r_i - q_i$, we get

$$\sum_{i=1}^{n} (r_i - p_i)^2 \leq 2 \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \sqrt{\sum_{i=1}^{n} (r_i - q_i)^2} + \sum_{i=1}^{n} (q_i - p_i)^2 + \sum_{i=1}^{n} (r_i - q_i)^2.$$

Hence,

$$\sum_{i=1}^{n} (r_i - p_i)^2 \leq \left(\sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} + \sqrt{\sum_{i=1}^{n} (r_i - q_i)^2} \right)^2$$

from which it follows that $D(P,Q)+D(Q,R)\geq D(P,R)$.

Of course, the triangle inequality is used in Definition A.1.1 precisely because it is one of the fundamental properties of the Euclidean distance function. However, we needed to verify the triangle inequality based on the formula given for the Euclidean metric, and the example illustrates what is required in order to establish the three axioms.

Example A.1.3. There exists a variety of other metrics on Euclidean space, and we illustrate a few of these alternate metrics for $\mathbb{R}^2$. Let $P=(x_1,y_1)$ and $Q=(x_2,y_2)$. We leave to the reader the proofs that the following functions are metrics on $\mathbb{R}^2$:

$$\begin{aligned} D_1(P,Q) &= |x_2-x_1|+|y_2-y_1|, \\ D_3(P,Q) &= \sqrt[3]{|x_2-x_1|^3+|y_2-y_1|^3}, \\ D_\infty(P,Q) &= \max\big\{|x_2-x_1|,|y_2-y_1|\big\}. \end{aligned}$$

Example A.1.4 (Six Degrees of Kevin Bacon)**.** A humorous example of a metric space is the set of syndicated actors A equipped with the function D defined as follows. Consider the graph whose set of vertices is A and has an edge between two actors a_1 and a_2 if they acted in a movie together. Define $D(a_1,a_2)$ as 0 if $a_1=a_2$ and, otherwise, as the minimum number of edges it takes to create a path connecting a_1 and a_2. The pair (A,D) is a metric space.

The party game called "Six Degrees of Kevin Bacon" asks players to find $D(a_1,a_2)$ given any pair (a_1,a_2).

Having a notion of distance in a set, we may want to consider the subset of all points that are within a certain distance of a fixed point.

Definition A.1.5. Let (X,D) be a metric space, and let $p\in X$ be a point. We define the *open ball* of radius r around p as the set

$$B_r(p)=\{y\in X \mid D(p,y)<r\}.$$

The reader who is new to topology should note that the terminology "open ball" might be initially misleading since the set $B_r(p)$ only takes the shape of an actual ball (disk, sphere, etc.) in the case of the Euclidean metric on $\mathbb{R}^n$.

Example A.1.6. Consider the metric D_1 from Example A.1.3 above, and let $O=(0,0)$. The ball of radius 1 around the origin O using the metric D_1 is the set

$$B_1(O)=\{(x,y)\in\mathbb{R}^2 \mid |x|+|y|<1\}.$$

Notice that the equation $|x|+|y|=1$ has a locus that is symmetric about the x-axis and about the y-axis. So to determine its locus we only need to see what happens in the first quadrant. In the first quadrant, the equation $|x|+|y|=1$ becomes $x+y=1$, which is a line segment from $(1,0)$ to $(0,1)$. Thus, the open ball $B_1(O)$ is the open square with corners at $\{(1,0),(0,1),(-1,0),(0,-1)\}$.

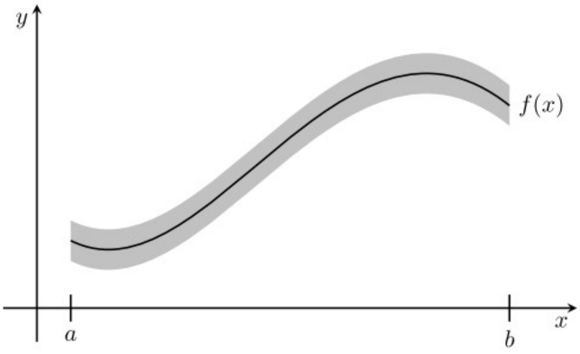

Figure A.1: Example A.1.7: a collar around $f(x)$.

Example A.1.7. Metric spaces can encompass a much wider range than the above examples have illustrated so far. Let $X = C^0([a,b])$ be the set of continuous real functions defined on the closed interval $[a,b]$, or let $X = F_{\text{bounded}}([a,b])$ be the set of all bounded functions on interval $[a,b]$. (A theorem of calculus tells us that any function f continuous over $[a,b]$ is bounded so $C^0([a,b]) \subset F_{\text{bounded}}([a,b])$.) Define the function $D : X \times X \to \mathbb{R}^{\geq 0}$ as

$$D(f,g) = \text{lub}\{|g(x) - f(x)| \, : \, x \in [a,b]\},$$

where lub refers to the least upper bound of a subset of reals.

The open ball of radius r around a function f is the set of all the functions $g \in X$ such that $|f(x) - g(x)| < r$ for all $x \in [a,b]$, or in other words,

$$f(x) - r < g(x) < f(x) + r \qquad \text{for all } x \in [a,b].$$

In this context, we call the region $f(x) - r < y < f(x) + r$ with $a \leq x \leq b$ the *r-collar* of $f(x)$. See Figure A.1.

Above, we introduced the notion of an open ball in any metric space. Though this appendix introduces notions that primarily support an overview of point-set topology, the notion of distance between two points allows us to generalize concepts from geometry to any metric space (X, D). We list here below a few of these concepts, which exist in any metric space but do not generalize further to topological spaces.

- A point $C \in X$ is said to be *between* two points A and B if $D(A,B) = D(A,C) + D(C,B)$. When D is the Euclidean metric on $\mathbb{R}^n$, this equality occurs for a degenerate triangle and only in the case when C lies on the segment $\overline{AB}$. It is in this sense that this definition directly generalizes the notion of betweenness from Euclidean geometry.

- The *bisector* of two points A and B is the set of points

$$\{P \in X \mid D(P, A) = D(P, B)\}.$$

  This is the usual definition for the segment bisector in Euclidean geometry but in other metric spaces this set may look quite different.

- If $A, B \in X$ and $c \in \mathbb{R}$ with $2c > D(A, B)$, then the set of points

$$\{M \in X \mid D(A, M) + D(B, M) = 2c\}$$

  is the ellipse with foci A and B and with half axis c.

- Let $S \subset X$ be a subset. We define the *diameter* of S to be

$$\operatorname{diam} S = \operatorname{lub}\{D(x, y) \mid x, y \in S\}.$$

- A subset S of X is called *bounded* if $\operatorname{diam} S < \infty$.

- Let S_1 and S_2 be two subsets of the metric space X. Then the distance between S_1 and S_2 is

$$D(S_1, S_2) = \operatorname{glb}\{D(x, y) \mid x \in S_1,\ y \in S_2\},$$

  where glb is the greatest lower bound. The distance between a point $x \in X$ and a subset $A \subset X$ is $D(\{x\}, A)$. We observe that this definition of distance between subsets does not establish a metric on $\mathcal{P}(X)$, the set of subsets of X. Indeed, for any two subsets S_1 and S_2 in X such that $S_1 \neq S_2$ and $S_1 \cap S_2 \neq 0$, the distance between them is $D(S_1, S_2) = 0$, and hence, even the first axiom for metric spaces fails. However, in geometry, the notion of distance between sets, especially disjoint sets, is quite useful.

A.1.2 Open and Closed Sets

In the study of real functions, we often use the notions of open intervals and closed intervals. In this context, we simply say that a bounded interval is open if it does not include its endpoints and closed if it includes both of them; a similar definition is given for an unbounded interval. Then a subset of $\mathbb{R}$ is called open if it is a disjoint union of open intervals. In contrast, in $\mathbb{R}^n$ or in a metric space, given the wide range of possibilities for the shape of sets, we cannot legitimately talk about endpoints, though we could attempt to make sense of the concept of "including its boundary points." Regardless, a different definition for openness and closedness is required.

Definition A.1.8. Let (X, D) be a metric space. A subset $U \subseteq X$ is called *open* if for all $p \in U$ there exists $r > 0$ such that the open ball $B_r(p) \subset U$. A subset $F \subseteq X$ is called *closed* if the complement $F^c \overset{\text{def}}{=} X - F$ is open.

Intuitively, this definition states that a subset U of a metric space is called open if around every point there is an open ball, perhaps with a small radius, that is completely contained in U. Note that we may wish to consider more than one metric at the same time on the same set X. In this case, we will refer to a D-open set.

Proposition A.1.9. *Let (X, D) be a metric space. Then*

1. *X and $\emptyset$ are both open;*
2. *the intersection of any two open sets is open;*
3. *the union of any collection of open sets is open.*

Proof. For part 1, if $p \in X$, then any open ball satisfies $B_r(p) \subset X$. Also, since $\emptyset$ is empty, the criteria for openness holds trivially for $\emptyset$.

To prove part 2, let U_1 and U_2 be two open sets and let $p \in U_1 \cap U_2$. Since U_1 and U_2 are open, there exist r_1 and r_2 such that $B_{r_1}(p) \subset U_1$ and $B_{r_2}(p) \subset U_2$. Take $r = \min(r_1, r_2)$. Then $B_r(p) \subseteq B_{r_1}(p) \subset U_1$ and $B_r(p) \subseteq B_{r_2}(p) \subset U_2$ so $B_r(p) \subset U_1 \cap U_2$. Thus, $U_1 \cap U_2$ is open.

Finally, consider a collection of open sets U_α where α is an index taken from some indexing set I, which is not necessarily finite. Define

$$U = \bigcup_{\alpha \in I} U_\alpha \,.$$

For any $p \in U$, there exists some $\alpha_0 \in I$ such that $p \in U_{\alpha_0}$. Since U_{α_0} is open, there exists r such that $B_r(p) \subset U_{\alpha_0}$, and thus, $B_r(p) \subset U$. Consequently, U is open. ☐

Using Proposition A.1.9(2), it is easy to show that any intersection of a finite number of open sets is again open. In contrast, part 3 states that the union of any collection of open subsets of X is again open, regardless of whether this collection is finite or not. This difference between unions and intersections of open sets is not an insufficiency of this proposition but rather a fundamental aspect of open sets in a metric space. In fact, as the following simple example shows, the infinite intersection of open sets need not be open.

Example A.1.10. For each integer $n \geq 1$, consider the open intervals $I_n = \left(0, 1 + \frac{1}{n}\right)$, and define

$$S = \bigcap_{n=1}^{\infty} I_n \,.$$

Obviously, $I_{n+1} \subsetneq I_n$, and so the intervals form a decreasing, nested chain. Since $\lim_{n\to\infty} \frac{1}{n} = 0$, we expect S to contain $(0, 1)$, but we must determine whether it

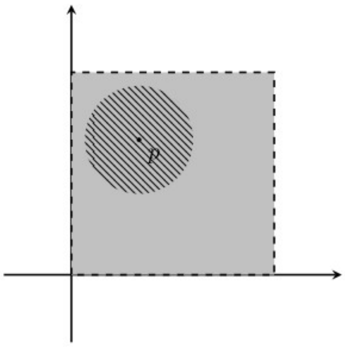

Figure A.2: Example A.1.11.

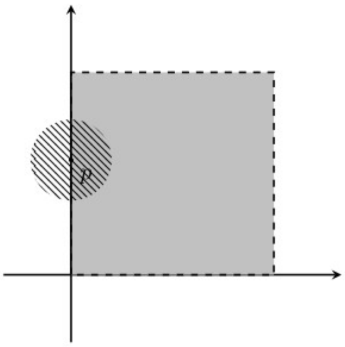

Figure A.3: Example A.1.12.

contains anything more. If $r > 1$, then if n is large enough so that $\frac{1}{n} < r - 1$, we have $r \notin I_n$. On the other hand, for all $n \in \mathbb{Z}^{\geq 1}$, $\frac{1}{n} > 0$, so $1 < 1 + \frac{1}{n}$. Hence, $1 \in I_n$ for all $n \in \mathbb{Z}^{\geq 1}$, and thus, $1 \in S$. Thus, we conclude that $S = (0, 1]$. This shows that the infinite intersection of open sets need not be open.

Example A.1.11. As a more down-to-earth example, we wish to show that according to this definition, the set $S = \{(x, y) \in \mathbb{R}^2 \,|\, 0 < x < 1 \text{ and } 0 < y < 1\}$ is open in $\mathbb{R}^2$ equipped with the Euclidean metric. Let $p = (x_0, y_0)$ be a point in S. Since $p \in S$, we see that $x_0 > 0$, $1 - x_0 > 0$, $y_0 > 0$, and $1 - y_0 > 0$. Since the closed distance from a point p to any line L is along a perpendicular to L, then the closest distance between p and any of the lines $x = 0$, $x = 1$, $y = 0$, and $y = 1$ is $\min\{x_0, 1 - x_0, y_0, 1 - y_0\}$. Consequently, if r is any positive real number such that

$$r \leq \min\{x_0, 1 - x_0, y_0, 1 - y_0\},$$

then $B_r(p) \subset S$. (See Figure A.2.)

Example A.1.12. In contrast to the previous example, consider the set

$$T = \{(x, y) \in \mathbb{R}^2 \,|\, 0 \leq x < 1 \text{ and } 0 < y < 1\},$$

where again we assume $\mathbb{R}^2$ is equipped with the Euclidean metric. The work in Example A.1.11 shows that for any point $p = (x_0, y_0)$, with $0 < x_0 < 1$ and $0 < y_0 < 1$, there exists a positive radius r such that $B_r(p) \subset S \subset T$. Thus, consider now points $p \in T$ with coordinates $(0, y_0)$. For all positive r, the open ball $B_r(p)$ contains the point $(-r/2, y_0)$, which is not in T. Hence, no open ball centered around points $(0, y_0)$ is contained in T, and hence T, is not open. (See Figure A.3.)

It is very common in proofs and definitions that rely on topology to refer to an open set that contains a particular point. Here is the common terminology.

Definition A.1.13. Let p be a point in a metric space (X, D). An *open neighborhood* (or simply *neighborhood*) of p is any open set of X that contains p.

Closed sets satisfy properties quite similar to those described in Proposition A.1.9, with a slight but crucial difference.

Proposition A.1.14. *Let (X, D) be a metric space. Then*

1. *X and $\emptyset$ are both closed;*
2. *the union of any two closed sets is closed;*
3. *the intersection of any collection of closed sets is closed.*

Because a set is defined as closed if its complement is open and because of DeMorgan laws for sets, this proposition is actually a simple corollary of Proposition A.1.9. Therefore, we leave the details of the proof to the reader.

Note that in any metric space, the whole set X and the empty set $\emptyset$ are both open and closed. Depending on the particular metric space, these are not necessarily the only subsets of X that are both open and closed.

Proposition A.1.15. *Let (X, D) be a metric space, and let $x \in X$. The singleton set $\{x\}$ is a closed subset of X.*

Proof. To prove that $\{x\}$ is closed, we must prove that $X - \{x\}$ is open. Let y be a point in $X - \{x\}$. Since $x \neq y$, by the axioms of a metric space, $D(x, y) > 0$. Let $r = \frac{1}{2}D(x, y)$. The real number r is positive, and we consider the open ball $B_r(y)$. Since $D(x, y) > r$, then $x \notin B_r(y)$, and hence, $B_r(y) \subset X - \{x\}$. Hence, we have shown that $X - \{x\}$ is open and thus that $\{x\}$ is closed. ☐

The notion of distance between sets provides an alternate characterization of closed sets in metric spaces. Recall that for any subset $A \subset X$, $x \in A$ implies that $D(x, A) = 0$. The following proposition shows that the converse holds precisely for closed sets.

Proposition A.1.16. *Let (X, D) be a metric space. A subset F is closed if and only if $D(x, F) = 0$ implies $x \in F$.*

Proof. Suppose first that F is closed. If $x \notin F$, then $x \in X - F$, which is open, so there exists an open ball $B_r(x)$ around x contained entirely in $X - F$. Hence, the distance between any point $a \in F$ and x is greater than the radius $r > 0$, thus, $D(x, F) > 0$, and in particular, $D(x, F) \neq 0$. Thus, $D(x, F) = 0$ implies that $x \in F$.

We now prove the converse. Suppose that F is a subset of X such that $D(x, F) = 0$ implies that $x \in F$. Then for all $x \in X - F$, we have $D(x, F) > 0$. Take the positive number $r = \frac{1}{2}D(x, F)$, and consider the open ball $B_r(x)$. Let p be any point in F and a any point in $B_r(x)$. Form the triangle inequality

$$D(p, x) \leq D(p, a) + D(a, x) \iff D(p, a) \geq D(p, x) - D(a, x).$$

The least possible value for $D(p,a)$ occurs when $D(p,x)$ is the least possible and when $D(a,x)$ is the greatest possible, that is when $D(p,x)=D(x,F)$ and $D(a,x)=r$. Thus, we find that $D(p,a)\geq r>0$. Hence, for all $a\in B_r(x)$, we have $D(F,a)>0$ and thus $B_r(x)\cap F=\emptyset$. Therefore, $B_r(x)\subset X-F$ so $X-F$ is open and F is closed. □

Proposition A.1.16 indicates that given any subset A of a metric space X, one can obtain a closed subset of X by adjoining all the points with 0 distance from A. This motivates the following definition.

Definition A.1.17. Let (X,D) be a metric space, and let $A\subset X$ be any subset. Define the *closure* of A as

$$\operatorname{Cl} A=\{x\in X \mid D(x,A)=0\}.$$

Proposition A.1.18. *Let (X,D) be a metric space and A any subset of X.* $\operatorname{Cl} A$ *is the smallest closed set containing A. In other words,*

$$\operatorname{Cl} A=\bigcap_{A\subset F,\ F\ \mathit{closed}} F\ .$$

Proof. (Left as an exercise for the reader. See Problem A.1.20.) □

A.1.3 Sequences

In standard calculus courses, one is introduced to the notion of a sequence of real numbers along with issues of convergence and limits. The definition given in such courses for when we say a sequence converges to a certain limit formalizes the idea of all terms in the sequence ultimately coming arbitrarily close to the limit point. Consequently, since limits formalize a concept about closeness and distance, the natural and most general context for convergence and limits is in a metric space.

Definition A.1.19. Let (X,D) be a metric space and let $\{x_n\}_{n\in\mathbb{N}}$ be a sequence in X. The sequence $\{x_n\}$ is said to converge to the limit $\ell\in X$ if for all $\varepsilon\in\mathbb{R}^{>0}$ there exists $N\in\mathbb{N}$ such that if $n>N$, then $D(x_n,\ell)<\varepsilon$ (i.e., $x_n\in B_\varepsilon(\ell)$). If $\{x_n\}$ converges to ℓ, then we write

$$\lim_{n\to\infty} x_n=\ell.$$

Note that we can restate Definition A.1.19 to say that $\{x_n\}$ converges to ℓ if for all positive $\varepsilon\in\mathbb{R}^{>0}$, only finitely many elements of the sequence $\{x_n\}$ are not in the open ball $B_\varepsilon(\ell)$.

Example A.1.20. Consider the sequence $\{x_n\}_{n\geq 1}$ in $\mathbb{R}^3$ given by $x_n=(3,\frac{1}{n+2},\frac{2n}{n+1})$. We prove that $\{x_n\}$ converges to $(3,0,2)$. We know that as sequences of real numbers,

$$\lim_{n\to\infty}\frac{1}{n+2}=0 \qquad \text{and} \qquad \lim_{n\to\infty}\frac{2n}{n+1}=2.$$

Pick any positive ε. Choose N_1 such that $n > N_1$ implies that $\frac{1}{n+2} < \frac{\varepsilon}{\sqrt{2}}$, and choose N_2 such that $n > N_2$ implies that $\left|\frac{2n}{n+1} - 2\right| < \frac{\varepsilon}{\sqrt{2}}$. Using the Euclidean distance

$$D(x_n, (3,0,2)) = \sqrt{(3-3)^2 + \left(\frac{1}{n+2} - 0\right)^2 + \left(\frac{2n}{n+1} - 2\right)^2},$$

one sees that if $n > N = \max(N_1, N_2)$, then

$$D(x_n, (3,0,2)) < \sqrt{\frac{\varepsilon}{2} + \frac{\varepsilon}{2}} = \varepsilon.$$

This proves that $\lim x_n = (3,0,2)$.

Note that we could have proved directly that $\lim x_n = (3,0,2)$ by considering the limit of $D(x_n, (3,0,2))$ as a sequence of real numbers and proving that this converges to 0.

Example A.1.21. Consider the set X of bounded, real-valued functions defined over the interval $[0,1]$ equipped with the metric defined in Example A.1.7. For $n \geq 1$, consider the sequence of functions given by

$$f_n(x) = \begin{cases} 1 - nx, & \text{for } 0 \leq x \leq \frac{1}{n}, \\ 0, & \text{for } \frac{1}{n} \leq x \leq 1. \end{cases}$$

Figure A.4 shows the functions for $n = 1, 2, 3$. One might suspect that the limit of this sequence $f_n(x)$ would be the function

$$f(x) = \begin{cases} 1, & \text{if } x = 0, \\ 0, & \text{for } x > 0, \end{cases}$$

but this is not the case. Let $r = \frac{1}{4}$, and consider the r-collar around $f(x)$. There is no n such that $f_n(x)$ lies within the $\frac{1}{4}$-collar around $f(x)$. (See Figure A.5.) Consequently, $f_n(x)$ does not converge to $f(x)$ in the metric space (X, D). Note, however, that for all $x \in [0,1]$, as sequences of real numbers $\lim_{n\to\infty} f_n(x) = f(x)$. We say that $f_n(x)$ converges *pointwise*.

Proposition A.1.22. *Let (X, D) be a metric space. Any sequence $\{x_n\}$ can converge to at most one limit point.*

Proof. Suppose that

$$\lim_{n\to\infty} x_n = \ell \qquad \text{and} \qquad \lim_{n\to\infty} x_n = \ell'.$$

Let ε be any positive real number. There exists N_1 such that $n > N_1$ implies that $D(x_n, \ell) < \frac{\varepsilon}{2}$, and there exists N_2 such that $n > N_s$ implies that $D(x_n, \ell') < \frac{\varepsilon}{2}$. Thus, taking some $n > \max(N_1, N_2)$, we deduce from the triangle inequality that

$$D(\ell, \ell') \leq D(x_n, \ell) + D(x_n, \ell') \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

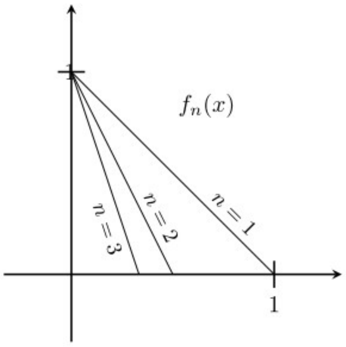

Figure A.4: Sequence of functions.

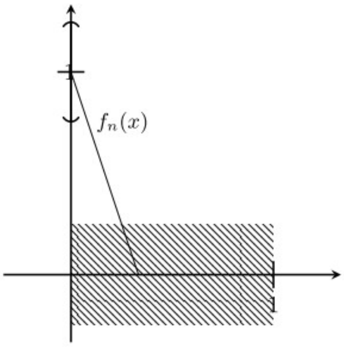

Figure A.5: $\frac{1}{4}$-collar around $f(x)$.

Thus, since $D(\ell, \ell')$ is less than any positive real number, we deduce that $D(\ell, \ell') = 0$ and hence that $\ell = \ell'$. ☐

In any metric space, there are plenty of sequences that do not converge to any limit. For example, the sequence $\{a_n\}_{n\geq 1}$ of real numbers given by $a_n = (-1)^n + \frac{1}{n}$ does not converge toward anything but, in the long term, alternates between being very close to 1 and very close to -1. Referring to the restatement of Definition A.1.19, one can loosen the definition of limit to incorporate the behavior of such sequences as the one just mentioned.

Definition A.1.23. Let (X, D) be a metric space, and let $\{x_n\}$ be a sequence in X. A point $p \in X$ is called an *accumulation point* of $\{x_n\}$ if for all real $\varepsilon > 0$, an infinite number of elements x_n are in $B_\varepsilon(p)$. The accumulation set of $\{x_n\}$ is the set of all accumulation points.

Example A.1.24. Consider again the real sequence $a_n = (-1)^n + \frac{1}{n}$. Let ε be any positive real number. If $n > \frac{1}{\varepsilon}$ and n is even, then $a_n \in B_\varepsilon(1)$. If $n > \frac{1}{\varepsilon}$ and n is odd, then $a_n \in B_\varepsilon(-1)$. Hence, 1 and -1 are accumulation points. However, for any r different than 1 or -1, suppose we choose a ε such that $\varepsilon < \min(|r-1|, |r+1|)$. If n is large enough, then

$$\frac{1}{n} < \big|\min(|r-1|, |r+1|) - \varepsilon\big|,$$

and for such n, we have $a_n \notin B_\varepsilon(r)$. Thus, 1 and -1 are the only accumulation points of $\{a_n\}$. In the terminology of Definition, A.1.23, the accumulation set is $\{-1, 1\}$.

A.1.4 Continuity

For the same reason as for the convergence of sequences, the notion of continuity, first introduced in the context of real functions over an interval, generalizes naturally to the category of metric spaces. Here is the definition.

Definition A.1.25. Let (X, D) and (Y, D') be two metric spaces. A function $f : X \to Y$ is called *continuous at* $a \in X$ if for all $\varepsilon \in \mathbb{R}^{>0}$, there exists $\delta \in \mathbb{R}^{>0}$ such that $D(x, a) < \delta$ implies that $D(f(x), f(a)) < \varepsilon$. The function f is called *continuous* if it is continuous at all points $a \in X$.

Example A.1.26. As a first example of Definition A.1.25, consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by $f(x, y) = x + y$, where we assume $\mathbb{R}^2$ and $\mathbb{R}$ are equipped with the usual Euclidean metrics. Consider some point $(a_1, a_2) \in \mathbb{R}^2$. Let $\varepsilon > 0$ be any positive real number. Choosing $\delta = \frac{\varepsilon}{2}$ will suffice, as we now show. First note that

$$D((x, y), (a_1, a_2)) = \sqrt{(x - a_1)^2 + (y - a_2)^2} < \frac{\varepsilon}{2}$$

implies that

$$|x - a_1| < \frac{\varepsilon}{2} \qquad \text{and} \qquad |y - a_2| < \frac{\varepsilon}{2}.$$

But if this is so, then

$$|f(x, y) - f(a_1, a_2)| = |x + y - (a_1 + a_2)| < |x - a_1| + |y - a_2| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Thus, f is continuous.

Example A.1.27. Definition A.1.25 allows one to study the continuity of functions in much more general contexts, as we show with this example. Let X be a proper subset of $\mathbb{R}^n$, and let $\vec{p}$ be a point in $\mathbb{R}^n - X$. We view X as a metric space by restricting the Euclidean metric to it. Let $\mathbb{S}^{n-1}$ be the unit sphere in $\mathbb{R}^n$ also with its metric coming from the Euclidean one in $\mathbb{R}^n$. Define a function $f : X \to \mathbb{S}^{n-1}$ by

$$f(\vec{x}) = \frac{\vec{x} - \vec{p}}{\|\vec{x} - \vec{p}\|}.$$

We will show that f is continuous.

Let $\vec{a} \in X$. The Euclidean metric is $D(\vec{x}, \vec{a}) = \|\vec{x} - \vec{a}\|$. Hence

$$\begin{aligned} D(f(\vec{a}), f(\vec{x})) &= \left\| \frac{\vec{a} - \vec{p}}{\|\vec{a} - \vec{p}\|} - \frac{\vec{x} - \vec{p}}{\|\vec{x} - \vec{p}\|} \right\| \\ &= \sqrt{\left(\frac{\vec{a} - \vec{p}}{\|\vec{a} - \vec{p}\|} - \frac{\vec{x} - \vec{p}}{\|\vec{x} - \vec{p}\|} \right) \cdot \left(\frac{\vec{a} - \vec{p}}{\|\vec{a} - \vec{p}\|} - \frac{\vec{x} - \vec{p}}{\|\vec{x} - \vec{p}\|} \right)} \\ &= \sqrt{2 - 2\frac{(\vec{a} - \vec{p}) \cdot (\vec{x} - \vec{p})}{\|\vec{a} - \vec{p}\| \, \|\vec{x} - \vec{p}\|}} \\ &= \sqrt{2 - 2\cos\alpha}, \end{aligned}$$

where α is the angle between the vectors $(\vec{a}-\vec{p})$ and $(\vec{x}-\vec{p})$. However, from the trigonometric identity $\sin^2\theta = (1-\cos 2\theta)/2$, we deduce that

$$D(f(\vec{a}), f(\vec{x})) = 2\sin\left(\frac{\alpha}{2}\right).$$

If $\vec{x}$ and $\vec{a}$ are close enough, then $(\vec{a}-\vec{p})$ and $(\vec{x}-\vec{p})$ form an acute angle, and hence, if d is the height from $\vec{a}$ to the segment between $\vec{p}$ and $\vec{x}$, we have

$$D(f(\vec{x}), f(\vec{a})) = 2\sin\left(\frac{\alpha}{2}\right) \leq 2\sin\alpha = 2\frac{d}{\|\vec{p}-\vec{a}\|} \leq 2\frac{\|\vec{x}-\vec{a}\|}{\|\vec{p}-\vec{a}\|}.$$

Therefore, choosing δ small enough so that the angle between $(\vec{a}-\vec{p})$ and $(\vec{x}-\vec{p})$ is acute and $\delta < \frac{1}{2}\|\vec{p}-\vec{a}\|\,\varepsilon$, we conclude that

$$\|\vec{x}-\vec{a}\| < \delta \Longrightarrow D(f(\vec{x}), f(\vec{a})) < \varepsilon,$$

proving that f is continuous at all points $\vec{a} \in X$.

Proposition A.1.28. *Let (X, D), (Y, D'), and (Z, D'') be metric spaces. Let $f : X \to Y$ and $g : Y \to Z$ be functions such that f is continuous at a point $a \in X$ and g is continuous at $f(a) \in Y$. Then the composite function $g \circ f : X \to Z$ is continuous at a.*

Proof. Since f is continuous at a, for all $\varepsilon_1 \in \mathbb{R}^{>0}$, there exists $\delta_1 \in \mathbb{R}^{>0}$ such that $D(x, a) < \delta_1$ implies that $D(f(x), f(a)) < \varepsilon_1$. Since g is continuous at $f(a)$, for all $\varepsilon_2 \in \mathbb{R}^{>0}$, there exists $\delta_2 \in \mathbb{R}^{>0}$ such that $D(y, f(a)) < \delta_2$ implies that $D\big(g(y), g(f(a))\big) < \varepsilon_2$. Therefore, given any $\varepsilon > 0$, set $\varepsilon_2 = \varepsilon$ and choose ε_1 so that $\varepsilon_1 < \delta_2$. Then

$$D(x, a) < \delta_1 \Rightarrow D(f(x), f(a)) < \varepsilon_1 < \delta_2 \Rightarrow D\big(g(f(x)), g(f(a))\big) < \varepsilon,$$

showing that $g \circ f$ is continuous at a. □

Using the concepts of open sets, we can give alternate formulations for when a function between metric spaces is continuous.

Proposition A.1.29. *Let (X, D) and (Y, D') be two metric spaces, and let $f : X \to Y$ be a function. The function f is continuous if and only if for all open subsets $U \subset Y$, the set*

$$f^{-1}(U) = \{x \in X \mid f(x) \in U\}$$

is an open subset of X.

Proof. First suppose that f is continuous. Let U be an open subset of Y, and let x be some point in $f^{-1}(U)$. Of course $f(x) \in U$. Since U is open, there exists a real

$\varepsilon > 0$ such that $B_\varepsilon(f(x)) \subset U$. Since f is continuous, there exists a $\delta > 0$ such that $y \in B_\delta(x)$ implies that $f(y) \in B_\varepsilon(f(x))$. Hence,

$$f(B_\delta(x)) \subset B_\varepsilon(f(x)) \subset U,$$

and thus, $B_\delta(x) \subset f^{-1}(U)$.

Conversely, suppose that $f^{-1}(U)$ is an open set in X for every open set U in Y. Let $f(x)$ be a point in U, and let ε be a positive real number. Then

$$f^{-1}\big(B_\varepsilon(f(x))\big)$$

is an open set in X. Since $x \in f^{-1}\big(B_\varepsilon(f(x))\big)$ is open, there exists some δ such that $B_\delta(x) \subset f^{-1}\big(B_\varepsilon(f(x))\big)$. Thus, $f(B_\delta(x)) \subset B_\varepsilon(f(x))$, and therefore, f is continuous. ☐

The following proposition is an equivalent formulation to Proposition A.1.29 but often more convenient for proofs.

Proposition A.1.30. *Let (X, D) and (Y, D') be two metric spaces and let $f : X \to Y$ be a function. The function f is continuous if and only if for all open balls $B_r(p)$ in Y, the set $f^{-1}(B_r(p))$ is an open subset of X.*

Proof. (Left as an exercise for the reader. See Problem A.1.27.) ☐

Much more could be included in an introduction to metric spaces. However, many properties of metric spaces and continuous functions between them hold simply because of the properties of open sets (Proposition A.1.9) and the characterization of continuous functions in terms of open sets (Proposition A.1.29). This fact motivates the definition of topological spaces.

PROBLEMS

A.1.1. Prove that D_1, D_3, and D_∞ from Example A.1.3 are in fact metrics on $\mathbb{R}^2$.

A.1.2. In the following functions on $\mathbb{R}^2 \times \mathbb{R}^2$, which axioms fail to make the function into a metric?

(a) $D_1((x_1, y_1), (x_2, y_2)) = |x_1| + |x_2| + |y_1| + |y_2|$.

(b) $D_2((x_1, y_1), (x_2, y_2)) = -\big((x_2 - x_1)^2 + (y_2 - y_1)^2\big)$.

(c) $D_3((x_1, y_1), (x_2, y_2)) = |x_2 - x_1| \cdot |y_2 - y_1|$.

(d) $D_4((x_1, y_1), (x_2, y_2)) = |x_2^2 - x_1^2| + |y_2 - y_1|$.

A.1.3. Let (X_1, D_1) and (X_2, D_2) be metric spaces. Consider the Cartesian product $X = X_1 \times X_2$. Prove that the following function is a metric on X:

$$D((p_1, p_2), (q_1, q_2)) = D(p_1, q_1) + D(p_2, q_2).$$

A.1.4. Let $X = \mathcal{P}_{\text{fin}}(\mathbb{Z})$ be the set of all finite subsets of the integers. Recall that the symmetric difference between two sets A and B is $A \triangle B = (A - B) \cup (B - A)$. Define the function $D : X \times X \to \mathbb{R}^{\geq 0}$ by

$$D(A, B) = |A \triangle B|,$$

the cardinality of $A \triangle B$. Prove that D is a metric on X.

A.1.5. In Euclidean geometry, the median line between two points p_1 and p_2 in $\mathbb{R}^2$ is defined as the set of points that are of equal distance from p_1 and p_2, i.e.,

$$M = \{q \in \mathbb{R}^2 \mid D(q, p_1) = D(q, p_2)\}.$$

What is the shape of the median lines in $\mathbb{R}^2$ for D_1, D_2, and D_∞ from Example A.1.3?

A.1.6. Prove that if (X, D) is any metric space, then $D(x, y)^n$ where n is any positive integer, is also metric on X.

A.1.7. Let (X, D) be a metric space, and let S be any subset of X. Prove that (S, D) is also a metric space. (The metric space (S, D) is referred to as the *restriction of D to S*.)

A.1.8. Let $\mathbb{S}^2$ be the unit sphere in $\mathbb{R}^3$, i.e.,

$$\mathbb{S}^2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}.$$

Sketch the open balls on $\mathbb{S}^2$ obtained by the restriction of the Euclidean metric to $\mathbb{S}^2$ (see Problem A.1.7). Setting the radius $r < 2$, for some point $p \in \mathbb{S}^2$, describe $B_r(p)$ algebraically by the equation $x^2 + y^2 + z^2 = 1$ and some linear inequality in x, y, and z.

A.1.9. Prove that Example A.1.7 is in fact a metric space.

A.1.10. Consider the metric space $(\mathbb{R}^2, D_1)$, where D_1 is defined in Example A.1.3.

(a) Let A and B be two points in $\mathbb{R}^2$. Prove that the set of points between A and B is the rectangle with vertical or horizontal edges with A and B as opposite corners.

(b) Determine the bisector of the points A and B in this metric.

A.1.11. Find the distance between the following pairs of sets in $\mathbb{R}^2$:

(a) $A = \{(x, y) \mid x^2 + y^2 < 1\}$ and $B = \{(x, y) \mid (x - 3)^2 + (y - 2)^2 < 1\}$.

(b) $A = \{(x, y) \mid xy = 1\}$ and $B = \{(x, y) \mid xy = 0\}$.

(c) $A = \{(x, y) \mid xy = 2\}$ and $B = \{(x, y) \mid x^2 + y^2 < 1\}$.

A.1.12. Prove that a subset A of a metric space (X, D) is bounded if and only if $A \subset B_r(p)$ for some $r \in \mathbb{R}^{\geq 0}$ and $p \in X$.

A.1.13. *Infinite Intersections and Unions.* Let $A_n = [n, +\infty)$ and let $B_n = [\frac{1}{n}, \sin n]$. Find

$$\text{(a)}\ \bigcup_{n=0}^{\infty} A_n, \qquad \text{(b)}\ \bigcap_{n=0}^{\infty} A_n, \qquad \text{(c)}\ \bigcup_{n=2}^{\infty} B_n. \tag{A.1}$$

A.1.14. Define the metric D on $\mathbb{R}^2$ as follows. For any $p = (p_x, p_y)$ and $q = (q_x, q_y)$, let $D(p,q) = \sqrt{4(p_x - q_x)^2 + (p_y - q_y)^2}$. Prove that this is in fact a metric. What is the shape of a unit ball $B_r\big((x,y)\big)$? What is the shape of the "median" between two points? [Hint: see Problem A.1.5.]

A.1.15. Are the following subsets of the plane (using the usual Euclidean metric) open, closed, or neither:

(a) $\{(x,y) \in \mathbf{R}^2 : x^2 + y^2 < 1\}$.

(b) $\{(x,y) \in \mathbf{R}^2 : x^2 + y^2 \geq 1\}$.

(c) $\{(x,y) \in \mathbf{R}^2 : x + y = 0\}$.

(d) $\{(x,y) \in \mathbf{R}^2 : x + y \neq 0\}$.

(e) $\{(x,y) \in \mathbf{R}^2 : x^2 + y^2 \leq 1 \text{ or } x = 0\}$.

(f) $\{(x,y) \in \mathbf{R}^2 : x^2 + y^2 < 1 \text{ or } x = 0\}$.

(g) The complement A^c where $A = \{(x,y) \in \mathbf{R}^2 \mid x = 0 \text{ and } -1 \leq y \leq 1\}$.

A.1.16. Let L be a line in the plane $\mathbb{R}^2$. Prove that $\mathbb{R}^2 - L$ is open in the Euclidean metric and in the three metrics presented in Example A.1.3.

A.1.17. Let $X = \mathbb{R}^2$, and define D_2 as the Euclidean metric and D_1 as $D_1((x_1,y_1),(x_2,y_2)) = |x_2 - x_1| + |y_2 - y_1|$. Prove that any D_2-open ball contains a D_1-open ball and is also contained in a D_1-open ball. Conclude that a subset of $\mathbb{R}^2$ is D_2-open if and only if it is D_1-open.

A.1.18. Prove that the set $\{\frac{1}{n} \mid n \in \mathbb{Z}^{>0}\}$ is not closed in $\mathbb{R}$ whereas the set $\{\frac{1}{n} \mid n \in \mathbb{Z}^{>0}\} \cup \{0\}$ is.

A.1.19. Consider a metric space (X, D), and let x and y be two distinct points of X. Prove that there exists a neighborhood U of x and a neighborhood V of y such that $U \cap V = \emptyset$. (In general topology, this property is called the *Hausdorff property*, and this exercise shows that all metric spaces are Hausdorff.)

A.1.20. Prove Proposition A.1.18.

A.1.21. Let A be a subset of a metric space (X, D). Suppose that every sequence $\{x_n\}$ in A that converges in X converges to an element of A. Prove that A is closed.

A.1.22. Let $\{x_n\}_{n\in\mathbb{N}}$ be a sequence in a metric space (X, D). Prove that the closure of the set of elements $\{x_n\}$ is $\{x_n \mid n \in \mathbb{N}\}$ together with the accumulation set of the sequence $\{x_n\}$.

A.1.23. Using Definition A.1.25, prove that the real function $f(x) = 2x - 5$ is continuous over $\mathbb{R}$.

A.1.24. Using Definition A.1.25, prove that

(a) the real function $f(x) = x^2$ is continuous over $\mathbb{R}$;

(b) the real function of two variables $f(x,y) = 1/(x^2 + y^2 + 1)$ is continuous over all $\mathbb{R}^2$ (using the usual Euclidean metric).

A.1.25. Let (X, D) and (Y, D') be metric spaces, and let $f : X \to Y$ be a continuous function. Prove that if a sequence $\{x_n\}$ in X converges to a limit point ℓ, then the sequence $\{f(x_n)\}$ in Y converges to $f(\ell)$.

A.1.26. Let $f : X \to Y$ be a function from the metric space (X, D) to the metric space (Y, D'). Suppose that f is such that there exists a positive real number λ, with

$$\frac{D'(f(x_1), f(x_2))}{D(x_1, x_2)} \leq \lambda \qquad \text{for all } x_1 \neq x_2,$$

i.e., that the stretching ratio for f is bounded. Show that f is continuous.

A.1.27. Prove Proposition A.1.30.

A.1.28. Let $X = \mathbb{R}^m$ and $Y = \mathbb{R}^n$ equipped with the usual Euclidean metric. Prove that any linear transformation from X to Y is continuous.

A.2 Topological Spaces

A.2.1 Definitions and Examples

Definition A.2.1. A *topological space* is a pair (X, τ) where X is a set and where τ is a set of subsets of X satisfying the following:

1. X and $\emptyset$ are in τ.
2. For all U and V in τ, $U \cap V \in \tau$.
3. For any collection $\{U_\alpha\}_{\alpha \in I}$ of sets in τ, the union $\bigcup_{\alpha \in I} U_\alpha$ is in τ.

The elements in τ are called open subsets of X and a subset $F \subset X$ is called closed if $X - F$ is open, i.e., if $X - F \in \tau$.

As an alternate terminology we talk about τ satisfying the above three properties as a *topology* on X. In the introduction to this appendix, we promised that topology attempts to provide a mental model that generalizes the notion of nearness. The following concept is key to this way of thinking.

Definition A.2.2. Let (X, τ) be a topological space and let $x \in X$. Any $U \in \tau$ such that $x \in U$ is called a *neighborhood* of x.

If we work with more than one topology on the same underlying set X, we refer to τ-open and τ-closed subsets to avoid ambiguity.

As with the properties of open sets in metric spaces, in criterion (3) of Definition A.2.1, the indexing set I need not be countable, and hence, we should not assume that the collection $\{U_\alpha\}_{\alpha \in I}$ can be presented as a sequence of subsets.

Example A.2.3. According to Proposition A.1.9, if (X, D) is a metric space, it is also a topological space, where we use the topology τ to be the open sets as defined by Definition A.1.8. Historically, it was precisely Proposition A.1.9 along with the discovery by mathematicians that collections of sets with the properties described in this proposition arise naturally in numerous other contexts that led to the given definition of a topological space.

Example A.2.4 (Euclidean topology). Consider the metric space $(\mathbb{R}^2, D)$, where D is the Euclidean metric. The topology induced on $\mathbb{R}^n$ according to Example A.2.3 is called the Euclidean topology.

Example A.2.5 (Discrete topology). Let X be any set. Setting $\tau = \mathcal{P}(X)$ to be the set of all subsets of X is a topology on X called the *discrete topology* on X. In the discrete topology, all subsets of X are both closed and open.

Example A.2.6 (Trivial topology). On the opposite end of the spectrum from, setting $\tau = \{X, \emptyset\}$ also satisfies the axioms of a topology, and this is called the *trivial topology* on X. These two examples represent the largest and the smallest possible examples of topologies on a set X.

Example A.2.7. Let $X = \{a, b, c\}$ be a set with three elements. Consider the set of subsets $\tau = \{\emptyset, \{a\}, \{a, b\}, X\}$. A simple check shows that τ is a topology on X, namely that τ satisfies all the axioms for a topology. Notice that $\{a\}$ is open, $\{c\}$ is closed (since $\{a, b\}$ is open) and that $\{b\}$ is neither open nor closed. By Proposition A.1.15, there is no metric D on X such that the D-open sets of X are the open sets in the topology of τ. We say that (X, τ) is not metrizable.

It is not always easy to specify a subset of $\mathcal{P}(X)$ that satisfy the axioms for a topology on X. The concept of a basis makes this possible and Proposition A.2.9 gives a practical characterization of a basis.

Definition A.2.8. Let (X, τ) be a topological space. A collection of open sets $\mathcal{B} \subset \tau$ is called a *basis* of the topology if every open set is a union of elements in $\mathcal{B} \subset \tau$.

Proposition A.2.9. *Let (X, τ) be a topological space, and suppose that $\mathcal{B}$ is a basis. Then:*

1. *the elements of $\mathcal{B}$ cover X;*
2. *if $B_1, B_2 \in \mathcal{B}$, then for all $x \in B_1 \cap B_2$ there exists $B_3 \in \mathcal{B}$ such that $x \in B_3 \subset B_1 \cap B_2$.*

Conversely, if any collection $\mathcal{B}$ of open sets satisfies the above two properties, then there exists a unique topology on X for which $\mathcal{B}$ is a basis. (This topology is said to be generated *by $\mathcal{B}$.)*

Proof. (Left as an exercise for the reader.) ☐

This characterization allows one to easily describe topologies by presenting a basis of open sets.

Example A.2.10. By Definition A.1.8, in a metric space, the topology associated to a metric has the set of open balls as a basis.

Example A.2.11. Let $X = \mathbb{R}^2$, and consider the collection $\mathcal{B}$ of sets of the form

$$U = \{(x, y) \in \mathbb{R}^2 \mid x > x_0 \text{ and } y > y_0\},$$

where x_0 and y_0 are constants. This collection $\mathcal{B}$ satisfies both of the criteria in Proposition A.2.9, hence there exists a unique topology τ on $\mathbb{R}^2$ with $\mathcal{B}$ as a basis. It is easy to see that τ is different from the usual Euclidean topology. Note that for any open set $U \in \tau$, if $(a, b) \in U$, then the half-infinite ray

$$\{(a + t, b + t) \mid t \geq 0\}$$

is a subset of U. This is not a property of the Euclidean topology on $\mathbb{R}^2$, so this gives a topology different from the Euclidean topology.

Proposition A.2.12. *Let (X, τ) be a topological space. Then the following are true about the τ-closed sets of X:*

1. *X and $\emptyset$ are closed.*
2. *The union of any two closed sets is closed.*
3. *The intersection of any collection of closed sets is closed.*

Proof. Part 1 is obviously true since both X and $\emptyset$ are open.

For part 2, let F_1 and F_2 be any two closed subsets of X. Then ${F_1}^c$ and ${F_2}^c$ are open sets. Thus ${F_1}^c \cap {F_2}^c$ is open. However, by the DeMorgan laws,

$${F_1}^c \cap {F_2}^c = (F_1 \cup F_2)^c.$$

Thus, since $F_1 \cup {F_2}^c$ is open, $F_1 \cup F_2$ is closed.

For part 3, let $\{F_\alpha\}$, where α is in some indexing set I, be a collection of closed subsets. The collection $\{{F_\alpha}^c\}$ is a collection of open sets. Therefore, $\bigcup_\alpha {F_\alpha}^c$ is open. Thus,

$$\bigcap_{\alpha \in I} F_\alpha = \left(\bigcup_{\alpha \in I} {F_\alpha}^c \right)^c$$

is closed. □

A converse to this proposition turns out to be useful for defining certain classes of topologies on sets.

Proposition A.2.13. *Let X be a set. Suppose that a collection $\mathcal{C}$ of subsets of X satisfies the following properties:*

1. *X and $\emptyset$ are in $\mathcal{C}$.*
2. *The union of any two sets in $\mathcal{C}$ is again in $\mathcal{C}$.*
3. *The intersection of any collection of sets in $\mathcal{C}$ is again in $\mathcal{C}$.*

Then the set of all complements of sets in $\mathcal{C}$ form a topology on X.

Proof. (Left as an exercise for the reader.) ☐

Example A.2.14. Let X be any set. Consider the collection $\mathcal{C}$ of subsets of X that include X, $\emptyset$, and all finite subsets of X. The collection $\mathcal{C}$ satisfies all three criteria in Proposition A.2.13, so X, $\emptyset$, and the complements of finite subsets of X form a topology on X. This is often called the *finite complement* topology.

We now present a topology on $\mathbb{R}^n$ that has more open sets than the finite complement topology but fewer than the usual Euclidean topology.

Example A.2.15. Let $X = \mathbb{R}^n$. Let $\mathcal{C}$ be the collection of all finite unions of affine subspaces of $\mathbb{R}^n$, where by affine subspace we mean any set of points that is the solution set to a set of linear equations in $x_1, x_2, \ldots, x_n$ (i.e., points, lines, planes, etc.). Taking the empty set of linear equations or an inconsistent set of linear equations, we obtain X and $\emptyset$ as elements of $\mathcal{C}$. Since the union operation of sets is associative, a finite union of finite unions of affine spaces is again just a union of affine spaces.

To establish that the third criterion in A.2.13 holds for $\mathcal{C}$, we must prove that the intersection of any collection of finite unions of affine spaces is a finite union of affine spaces. Note first that if the intersection of two affine subspaces A_1 and A_2 is a strict subspace of both A_1 and A_2, then $\dim A_1 \cap A_2$ is strictly less than $\dim A_1$ and $\dim A_2$. Let $\{F_\alpha\}_{\alpha \in I}$ be a collection of sets in $\mathcal{C}$, and let $\{\alpha_i\}_{i \in \mathbb{N}}$ be a sequence of indices. Given the sequence $\{\alpha_i\}$, create a sequence of "intersection" trees according to the following recursive definition. The tree T_0 is the tree with a base node, and an edge for each $A_{\alpha_0, j}$ in

$$F_{\alpha_0} = \bigcup_j A_{\alpha_0, j} \ ,$$

with a corresponding leaf for each $A_{\alpha_0, j}$. For each tree T_i, construct T_{i+1} from T_i as follows. Writing

$$F_{\alpha_i} = \bigcup_j A_{\alpha_i, j} \ ,$$

for each leaf F of T_i and for each j such that $F \cap A_{\alpha_i, j} \neq F$, adjoin an edge labeled by $A_{\alpha_i, j}$ to F and label the resulting new leaf by $F \cap A_{\alpha_i, j}$.

As constructed, for each k, the leaves of the tree T_k are labeled by intersections of affine spaces so that

$$\bigcap_{i=0}^{k} F_{\alpha_i}$$

is the union of the leaves of T_k. Since only a finite number of edges gets added to every leaf, for all $i \geq 0$, there can be only a finite number of vertices at a fixed distance from the base node. Furthermore, since any nontrivial intersection $A \cap B$ of affine spaces has dimension

$$\dim(A \cap B) \leq \min(\dim A, \dim B) - 1$$

and since the ambient space is $\mathbb{R}^n$, each branch (descending path) in any tree T_i can have at most $n+1$ edges. In conclusion, for all T_i, there can only be a finite number of leaves. Thus,

$$\bigcap_{i=0}^{\infty} F_{\alpha_i}$$

is a finite union of intersections of affine spaces. Since this holds for all sequences $\{\alpha_i\}$ in I, this then proves that

$$\bigcap_{\alpha \in I} F_{\alpha}$$

is a finite union of intersections of affine spaces.

This rather lengthy proof shows that the collection $\mathcal{C}$ of finite unions of affine spaces in $\mathbb{R}^n$ satisfies the criteria of Proposition A.2.13. Consequently, the set τ of subsets of $\mathbb{R}^n$ that are complements of a finite union of affine spaces forms a topology on $\mathbb{R}^n$.

Example A.2.15, along with the result of Problem A.2.3, shows that given a set X, it is possible to have two topologies τ and τ' on X such that $\tau \subsetneq \tau'$, or in other words, that every τ-open set in X is τ'-open but not vice versa. This leads to a useful notion.

Definition A.2.16. Let X be a set and let τ and τ' be two topologies on X. If $\tau \subseteq \tau'$, then we say that τ' is *finer* than τ and that τ is *coarser* than τ'. If in addition $\tau \neq \tau'$, we say that τ' is *strictly finer* than τ and that τ is *strictly coarser* than τ'.

As a few examples, consider $X = \mathbb{R}^n$. Then the finite complement topology (Example A.2.14) is coarser than the topology defined in Example A.2.15, which is in turn coarser than the Euclidean topology.

In Section A.1.2, Proposition A.1.18 proved that the closure of a subset A (as defined by Definition A.1.17) of a metric space (X, D) is the intersection of all closed subsets of X containing A. This formulation of the closure of a set does not rely explicitly on the the metric D and carries over without changes to topological spaces. The closure of a subset $A \subset X$ is an example of a *topological operator*, three of which we mention here.

Definition A.2.17. Let (X, τ) be a topological space and A a subset of X. We define

1. the *closure* of A, written $\operatorname{Cl} A$, as the intersection of all closed sets in X containing A;
2. the *interior* of A, written A°, as the union of all open sets of X contained in A;
3. the *frontier* of A, written $\operatorname{Fr} A$, as the set of all $x \in X$ such that for every neighborhood U of x intersects A and A^c nontrivially, i.e., $U \cap A \neq \emptyset$ and $U \cap A^c \neq \emptyset$.

We leave some of the basic properties of the above topological operators to the exercises but present one common characterization of closed sets.

Proposition A.2.18. *Let* (X, τ) *be a topological space, and let* A *be any subset of* X*. The set* A *is closed if and only if* $A = \operatorname{Cl} A$*.*

Proof. If A is closed, then A is among the closed sets $F \subset X$ that contain A, and there is no smaller closed subset containing A. Hence $A = \operatorname{Cl} A$. Conversely, if $A = \operatorname{Cl} A$, then since the intersection of any collection of closed sets is closed, $\operatorname{Cl} A$ is closed, so A is closed. □

Another useful characterization of closed sets relies not on a topological operator but on properties of its limit points.

Definition A.2.19. Let A be any subset of a topological space (X, τ). A *limit point* of A is any point $p \in X$ such that $U \cap (A - \{p\})\ ne\emptyset$ for every open neighborhood U of p.

In the vocabulary of sequences in a metric space, if the subset A is a sequence $\{x_n\}$, then in Definition A.1.23, we would call the limit points of A the accumulation points of A. For this reason, some authors use the alternate terminology of accumulation point for limit points of a subset A in a topological space. The discrepancy in terminology is unfortunate, but in topology, the majority of authors use the vocabulary of Definition A.2.19.

Proposition A.2.20. *Let* A *be a subset of a topological space* X*. The set* A *is closed if and only if it contains all of its limit points.*

Proof. Assume first that A is closed. The complement $X - A$ is open and hence is a neighborhood of each of its points. Therefore, there is no point in $X - A$ that is a limit point of A. Hence, A contains all its limit points.

Assume now that A contains all of its limit points. Let $p \in X - A$. Since p is not a limit point of A, there exists an open neighborhood U of p such that $U \cap A = \emptyset$. Thus, $X - A$ is a neighborhood of each of its points, and hence, it is open. Thus, A is closed. □

Finally, we mention one last term related to closures.

Definition A.2.21. Let (X, τ) be a topological space. A subset A of X is called *dense* in X if $\operatorname{Cl} A = X$.

By Proposition A.2.20 a set A is dense in X if every point of X is a limit point of A. We give a few common examples of dense subsets.

Example A.2.22. Let $I = [a, b]$ be a closed interval in $\mathbb{R}$, and equip I with the topology induced from the Euclidean metric on $\mathbb{R}$. The open subsets in this topology on I are of the form $U \cap I$, where U is an open subset of $\mathbb{R}$. Furthermore, the open interval (a, b) is dense in I.

Proposition A.2.23. *The set $\mathbb{Q}$ of rational numbers is dense in $\mathbb{R}$.*

Proof. A precise proof of this statement must rely on a definition of real numbers, as constructed from the rationals. One may find this definition in any introductory analysis text. However, using a high school understanding of real numbers as numbers with an infinite decimal expansion that is not periodic, we can supply a simple proof of this fact.

Let $x_0 \in \mathbb{R}$ be any real number, and let U be an open neighborhood of x_0. By definition, there exists a positive real $\varepsilon > 0$ such that $(x_0 - \varepsilon, x_0 + \varepsilon) \in U$. Let $N = 1 - \lfloor \log_{10} \varepsilon \rfloor$. Consider q the fraction that represents the decimal approximation of x_0 that stops at N digits after the decimal period. Then, $q \in \mathbb{Q}$ and $q \in U$. Hence, x_0 is a limit point of $\mathbb{Q}$. □

Knowing that $\mathbb{Q}$ is a countable and dense subset of $\mathbb{R}$, we can introduce a property of the Euclidean topology that is a key part of the definition of a topological manifold.

Definition A.2.24. A topological space (X, τ) is called *second countable* if there exists a basis of τ that is countable.

Example A.2.25. The Euclidean space $\mathbb{R}^n$ is second countable. Consider the collection $\mathcal{B}$ of open balls whose centers have rational coordinates and of rational radius. This collection is in bijection with $\mathbb{Q}^n \times \mathbb{Q}^{>0}$. Since the Cartesian product of countable sets is countable, $\mathbb{B}$ is countable. It is not hard to see that this collection satisfies the conditions of Proposition A.2.9, so this $\mathcal{B}$ is a countable basis of $\mathbb{R}^n$.

A.2.2 Continuity

When working with topological spaces (X, τ) and (Y, τ'), we are usually not interested in studying the properties of just any function between $f : X \to Y$ because a function without any special properties will not necessarily relate the topology on X to that on Y or vice versa. In Section A.1.4, Proposition A.1.29 provided a characterization of continuous functions between metric spaces only in terms of the open sets in the metric space topology. This motivates the definition of continuity for functions between topological spaces.

Definition A.2.26. Let (X, τ) and (Y, τ') be two topological spaces, and let $f : X \to Y$ be a function. We call f continuous (with respect to τ and τ') if for every open set $U \subset Y$, the set $f^{-1}(U)$ is open in X.

Example A.2.27. Proposition A.1.29 shows that any function called continuous between two metric spaces (X, D) and (Y, D') is continuous with respect to the topologies induced by the metrics on X and Y.

In previous contexts (e.g., functions from $\mathbb{R}^n$ to $\mathbb{R}^m$, and functions between metric spaces), we first defined the notion of continuity at a point and then expanded to continuous over a domain. In this case, as in all texts on topology, we first defined continuity over a whole domain. However, there is a natural definition for continuity at a point.

Definition A.2.28. Let $f : X \to Y$ be as in the previous definition. We call f continuous at $x \in X$, if for every neighborhood V of $f(x)$, there is a neighborhood U of x with $f(U) \subset V$.

The following proposition shows why this definition of continuity at a point is natural.

Proposition A.2.29. *Let (X, τ) and (Y, τ') be topological spaces, and let $f : X \to Y$ be a function. Then f is continuous if and only if f is continuous at all $x \in X$.*

Proof. First, suppose that f is continuous. Then for all $f(x) \in Y$ and for all open neighborhoods V of $f(x)$, $f^{-1}(V)$ is open in X. Furthermore, $x \in f^{-1}(V)$ so $f^{-1}(V)$ is an open neighborhood of x. Also, since $f(f^{-1}(V)) = V$ for any set, we see that setting $U = f^{-1}(V)$ proves one direction.

Second, assume that f is continuous at all $x \in X$. Let V be any open set in Y. If V contains no image $f(x)$, then $f^{-1}(V) = \emptyset$, which is open in X. Therefore, assume that V contains some image $f(x)$. Let $W = f^{-1}(V)$. According to the assumption, for all $x \in W$ there exists open neighborhoods U_x of x such that $f(U_x) \subset V$. Since $f(U_x) \subset V$, then $U_x \subset f^{-1}(V) = W$. But then

$$W \subset \bigcup_{x \in W} U_x \subset W^\circ,$$

and since $W^\circ \subset W$ always, we conclude that $W = W^\circ$, which implies (see Problem A.2.7) that W is open. This shows that f is continuous. ☐

Proposition A.2.30. *Let (X, τ), (Y, τ'), and (Z, τ'') be three topological spaces. If $f : X \to Y$ and $g : Y \to Z$ are continuous functions, then $g \circ f : X \to Z$ is also a continuous function.*

Proof. (Left as an exercise for the reader.) ☐

For metric spaces, a continuous function is one that preserves “nearness” of points. Though topological spaces are not necessarily metric spaces, a continuous function $f : X \to Y$ between topological spaces preserves nearness in the sense that if two images $f(x_1)$ and $f(x_2)$ are in the same open set V, then there is an open set U that contains both x_1 and x_2, with $f(U) \subset V$.

In set theory, we view two sets as “the same” if there exists a bijection between them: they are identical except for how we label the specific elements. For topological spaces to be considered “the same,” not only do the underlying sets need to be in bijection, but this bijection must preserve the topology. This is the concept of a homeomorphism.

Definition A.2.31. Let (X, τ) and (Y, τ') be two topological spaces, and let $f : X \to Y$ be a function. The function f is called a *homeomorphism* if

1. f is a bijection;
2. $f : X \to Y$ is continuous;
3. $f^{-1} : Y \to X$ is continuous.

If there exists a homeomorphism between two topological spaces, we call them *homeomorphic*.

Example A.2.32. Any two squares S_1 and S_2, as subsets of $\mathbb{R}^2$ with the Euclidean metric, are homeomorphic. For $i = 1, 2$, let t_i be the translation that brings the center of S_i to the origin, R_i a rotation that makes the edges of $t_i(S_i)$ parallel to the x- and y-axes, and let h_i be a scaling (homothetie) that changes $R_i \circ t_i(S_i)$ into the square with vertices $\{(-1,-1), (-1,1), (1,1), (1,-1)\}$. It is easy to see that translations are homeomorphisms of $\mathbb{R}^2$ to itself. Furthermore, by Problem A.1.28, we see that R_i and h_i are continuous, and since they are invertible linear transformations, their inverses are continuous as well. Thus, R_i and h_i are homeomorphisms. Thus, the function

$$f = t_2^{-1} \circ R_2^{-1} \circ h_2^{-1} \circ h_1 \circ R_1 \circ t_1$$

is a homeomorphism of $\mathbb{R}^2$ into itself that sends S_1 to S_2. Thus, S_1 and S_2 are homeomorphic.

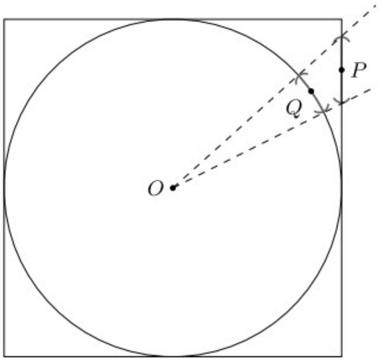

Figure A.6: A homeomorphism between a circle and a square from Example A.2.33.

Example A.2.33. Any circle and any square, as subsets of $\mathbb{R}^2$, are homeomorphic. We must present a homeomorphism between a circle and a square. By Example A.2.32 we see that any two squares are homeomorphic. By a similar argument, any two circles are homeomorphic. So without loss of generality, consider the consider the unit circle S and the unit square T, both with center $(0, 0)$. (See Figure A.6.)

Let $f : T \to S$ be the function defined by $f(x) = x/\|x\|$, where x is viewed as a point in $\mathbb{R}^2$. It is clear that this is a continuous function on its domain, $\mathbb{R}^2 - \{(0,0)\}$ so it is continuous on the subset T. For $\theta \in \mathbb{R}$, consider the function $g(\theta) = \max(|\sin\theta|, |\cos\theta|)$ and then the curve $\gamma : [0, 2\pi] \to \mathbb{R}^2$ parametrized by $\gamma(t) = (\cos(t)/g(t), \sin(t)/g(t))$. It is not hard to verify that this curve traces out the unit square T. Identifying a point on the circle with its angle, we can think of γ as a function $S \to T$. It is not hard to see that g is continuous and that γ is as well. Furthermore, $(f \circ \gamma)(p) = p$ for all $p \in S$ and also $(\gamma \circ f)(q) = q$ for all $q \in T$. Hence, f is a homeomorphism between the unit square and the unit circle.

Figure A.6 shows corresponding points P and Q as well as corresponding neighborhoods of these points.

Example A.2.34. The above two examples only begin to illustrate how different homeomorphic spaces may look. In this example, we prove that any closed, simple, parametrized curve γ in $\mathbb{R}^n$ is homeomorphic to the unit circle $\mathbb{S}^1$ (in $\mathbb{R}^2$). Let $\vec{x} : [0, l] \to \mathbb{R}^n$ be a parametrization by arclength for the curve γ such that $\vec{x}(t_1) = \vec{x}(t_2)$ implies that $t_1 = 0$ and $t_2 = l$, assuming $t_1 < t_2$. The function $f : \gamma \to \mathbb{S}^1$ defined by

$$f(P) = \Big(\cos\Big(\frac{2\pi}{l}\vec{x}^{-1}(P)\Big), \sin\Big(\frac{2\pi}{l}\vec{x}^{-1}(P)\Big)\Big) \tag{A.2}$$

produces the appropriate homeomorphism. Note that $\vec{x}^{-1}$ is not well defined only at the point $\vec{x}(0)$ because $\vec{x}^{-1}(\vec{x}(0)) = \{0, l\}$. However, using either 0 or l in Equation (A.2) is irrelevant.

Example A.2.35. In contrast to the previous example, consider the closed, regular curve γ parametrized by $\vec{x} : [0, 2\pi] \to \mathbb{R}^2$ with

$$\vec{x}(t) = (\cos t, \sin 2t).$$

The curve γ traces out a figure eight of sorts and is not simple because $\vec{x}(\pi/2) = (0,0) = \vec{x}(3\pi/2)$ (and $\vec{x}'(\pi/2) \neq \vec{x}'(3\pi/2)$). We show that it is not homeomorphic to a circle $\mathbb{S}^1$. Call $P = (0,0)$.

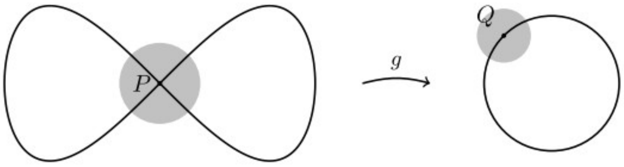

Figure A.7: Figure eight not homeomorphic to a circle.

There does exist a surjective continuous map f of γ onto the circle $\mathbb{S}^1$, namely, using the parametrization $\vec{x}(t)$,

$$f(\vec{x}(t)) = (\cos 2t, \sin 2t)$$

which amounts to folding the figure eight back onto itself so that the circle is covered twice. However, there exists no continuous bijection $g : \gamma \to \mathbb{S}^1$. To see this, call $Q = g(P)$, and let U be a small open neighborhood of Q. If g is a bijection, it has exactly one preimage for every element $x \in U$. Thus, $g^{-1}(U)$ is the image of a nonintersecting parametrized curve. However, every open neighborhood of P includes two segments of curves. Thus, the circle and γ are not homeomorphic.

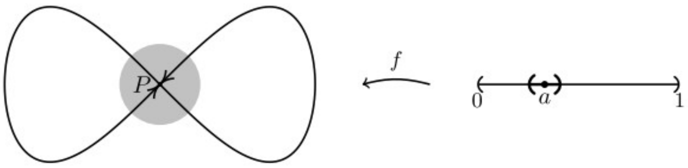

Figure A.8: Figure eight not homeomorphic to a segment.

Example A.2.36. Consider the same regular, closed parametrized curve as in the previous example. The function $f : (0,1) \to \gamma_2$ defined by $f(t) = \vec{x}\big(2\pi t - \frac{\pi}{2}\big)$ is a bijection. Furthermore, f is continuous since it is continuous as a function $(0,1) \to \mathbb{R}^2$. However, the inverse function is not continuous and in fact no continuous bijection $g : (0,1) \to \gamma_2$ can have a continuous inverse.

Let $a \in (0,1)$ be the real number such that $f(a) = P$, the point of self-intersection on γ_2. Take an open segment U around a. The image $f(U)$ is a portion of γ_2 through P (see the heavy lines in Figure A.8). However, this portion $f(U)$ is not an open subset of γ_2 since every open neighborhood of P includes $f((0,\varepsilon_1))$ and $f((1-\varepsilon_2,1))$. Thus there exists no homeomorphism between the open interval and the figure eight.

We conclude this section on continuity by mentioning one particular result, the proof of which exceeds the scope of this appendix.

Theorem A.2.37. *The Euclidean spaces* $\mathbb{R}^n$ *and* $\mathbb{R}^m$ *are homeomorphic if and only if* $m = n$.

This theorem states that Euclidean spaces can only be homeomorphic if they are of the same dimension. This might seem obvious to the casual reader but this fact hides a number of subtleties. First of all, the notion of dimension of a set in topology is not at all a simple one. Secondly, we must be careful to consider space-filling curves, such as the Peano curve, which is a continuous surjection of the closed interval $[0,1]$ onto the closed unit square $[0,1] \times [0,1]$. (See [28], Section 3-3, for a construction.) The construction for space-filling curves can be generalized to find continuous surjections of $\mathbb{R}^n$ onto $\mathbb{R}^m$ even if $n < m$. However, Theorem A.2.37 implies that no space-filling curve is bijective and has a continuous inverse. That Euclidean spaces of different dimensions are not homeomorphic means that they are distinct from the perspective of topology.

A.2.3 Derived Topological Spaces

Given any topological space (X, τ), there are a number of ways to create a new topological space. We present two common ways – subset topology and quotient spaces – which are used throughout this book.

Definition A.2.38. Let (X, τ) be a topological space, and let S be any subset of X. We define the *subset topology* τ' on S by calling a subset $A \subset S$ open if and only if there exists an open subset U of X such that $A = S \cap U$.

The subset topology is sometimes called the topology induced on S from X. There is an alternate way to characterize it.

Proposition A.2.39. *Let (X, τ) be a topological space, and let $S \subset X$. Let $i : S \to X$ be the inclusion function. The subset topology on S is the coarsest topology such that i is a continuous function.*

Proof. Given any subset A of X, we have $i^{-1}(A) = A \cap S$. If i is continuous, then for all open subsets $U \subset X$, the set $i^{-1}(U) = U \cap S$ is open. However, according to Definition A.2.38, the subset topology on S has no other open subsets and therefore is coarser than any other topology on S, making i continuous. ☐

Example A.2.40. Consider $\mathbb{R}$ equipped with the usual topology. Let $S = [a, b]$ be a closed interval. If $a < c < b$, in the subset topology on S, the interval $[a, c)$ is open. To see this, take any real $d < a$. Then

$$[a, c) = (d, c) \cap [a, b],$$

and (d, c) is open in $\mathbb{R}$.

A second and often rather useful way to create new topological spaces is to induce a topology on a quotient set.

Definition A.2.41. Let X be a set, and let R be an equivalence relation on X. The set of equivalence classes of R is denoted by X/R and is called the *quotient set* of X by R.

The concept of a quotient set arises in many areas of mathematics (congruence classes in number theory, quotient groups in group theory, quotient rings in ring theory, etc.) but also serves as a convenient way to define interesting objects in topology and geometry. We provide a few such examples before discussing topologies on quotient sets.

Example A.2.42 (Real Projective Space)**.** The typical construction of the real projective space is given in Example 3.1.6. We present two other equivalent constructions.

Let X be the set of all lines in $\mathbb{R}^{n+1}$, and let R be the equivalence relation of parallelism on X. We can therefore discuss the set of equivalence classes X/R. Each line in X is uniquely parallel to a line that passes through the origin, and hence, X/R may be equated with the set of lines passing through the origin. This set is called the *real projective space* of dimension n and is usually denoted by $\mathbb{RP}^n$.

As an alternate characterization for $\mathbb{RP}^n$, consider the unit n-sphere $\mathbb{S}^n$ in $\mathbb{R}^{n+1}$ and centered at the origin. Each line through the origin intersects $\mathbb{S}^n$ at two antipodal points. Therefore, $\mathbb{RP}^n$ is the quotient set of $\mathbb{S}^n$ with respect to the equivalence relation, in which two points are called equivalent if they are antipodal (form the ends of a diameter through $\mathbb{S}^n$).

Example A.2.43 (Grassmannian)**.** Let X_r be the set of all r dimensional vector subspaces in $\mathbb{R}^n$, and let R be the equivalence relation of parallelism between hyperplanes. The set of equivalence classes is called a Grassmannian and is denoted $\mathbb{G}(r, n)$. Again, since each r-dimensional hyperplane is uniquely parallel to one hyperplane through the origin, $\mathbb{G}(r, n)$ is the set of r-hyperplanes through the origin.

Of particular interest to geometry is the question of how to give a topology to X/R if X is equipped with a topology. Proposition A.2.39 illustrates how to make a reasonable definition.

Definition A.2.44. Let (X, τ) be a topological space and let R be an equivalence relation on X. Define $f : X \to X/R$ as the function that sends an element in X to its equivalence class; f is called the *quotient map* (or sometimes *identification map*). We call *quotient topology* (or *identification topology*) on X/R the finest topology that makes f continuous.

The above definition for the quotient topology does not make it too clear what the open sets of X/R should be. The following proposition provides a different characterization.

Proposition A.2.45. *Let (X, τ) be a topological space, let R be an equivalence relation on X and let f be the quotient map. The quotient topology on X/R is*

$$\tau' = \{U \in \mathcal{P}(X/R) \mid f^{-1}(U) \in \tau\}.$$

Proof. Let τ' be a topology on X/R such that $f : X \to X/R$ is continuous. Then for all open sets $U \in \tau'$, we must have $f^{-1}(U)$ be open in X. Note first that $f^{-1}(\emptyset) = \emptyset$ and $f^{-1}(X/R) = X$, which are both open in X.

From basic set theory, for any function $F : A \to B$ and any collection $\mathcal{C}$ of subsets of B, we have

$$F^{-1}\Big(\bigcup_{S\in\mathcal{C}} S\Big) = \bigcup_{S\in\mathcal{C}} F^{-1}(S) \qquad \text{and} \qquad F^{-1}\Big(\bigcap_{S\in\mathcal{C}} S\Big) = \bigcap_{S\in\mathcal{C}} F^{-1}(S).$$

(See Section 1.3 in [48].) Therefore, if U_1 and U_2 are such that $f^{-1}(U_1) \cap f^{-1}(U_2)$ is open, then $f^{-1}(U_1 \cap U_2)$ is open. Also, for any collection of sets $\{U_\alpha\}_{\alpha\in I}$ in X/R,

$$f^{-1}\Big(\bigcup_{\alpha\in I} U_\alpha\Big) = \bigcup_{\alpha\in I} f^{-1}(U_\alpha)$$

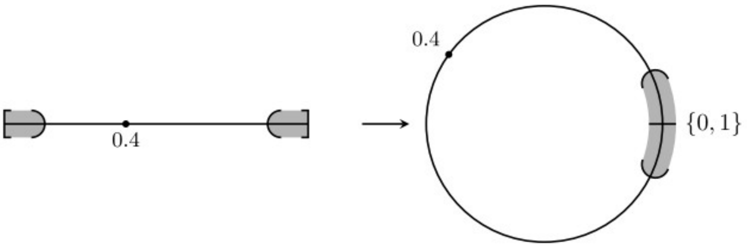

Figure A.9: Open set around $\{0,1\}$.

so if the right-hand side is open, then so is the left-hand side. Consequently, the collection of subsets $\mathcal{B}$ of X/R such that $U \in \mathcal{B}$ if and only if $f^{-1}(U) \in \tau$ is a topology on X/R. However, any finer topology on X/R would include some subset $S \subset X/R$ such that $f^{-1}(S)$ would not be open in X and, hence, would make f not continuous. □

Example A.2.46 (Circles)**.** Let I be the interval $[0,1]$ equipped with the topology induced from $\mathbb{R}$. Consider the equivalence relation that identifies 0 with 1, and everything else inequivalent. The identification space I/R is homeomorphic to a circle. We may use the function $f : I/R \to \mathbb{S}^1$ defined by

$$f(t) = (\cos 2\pi t, \sin 2\pi t),$$

which is well-defined since $f(0) = f(1)$, so whether we take 0 or 1 for the equivalence class $\{0,1\}$, we obtain the same image. This function f is clearly bijective.

To prove that f is continuous, let $P \in \mathbb{S}^1$, and let $x \in I/R$. If $P \neq (1,0)$, then any open neighborhood U of P contains an open interval of angles $\theta_1 < \theta < \theta_2$, where $0 < \theta_1 < \theta_2 < 2\pi$ and the angle θ_0 corresponding to P satisfies $\theta_1 < \theta_0 < \theta_2$. Then

$$f\Big(\frac{\theta_1}{2\pi}, \frac{\theta_2}{2\pi}\Big)$$

contains P and is a subset of U. On the other hand, if $P = (1,0)$, any open neighborhood U' of P in $\mathbb{S}^1$ contains an open arc of angles $\theta_1 < \theta < \theta_2$, with $\theta_1 < 0 < \theta_2$. Then if $g : I \to I/R$ is the quotient map,

$$f \circ g \left(\left[0, \frac{\theta_2}{2\pi}\right) \cup \left(1 - \frac{\theta_1}{2\pi}, 1\right] \right)$$

contains P and is contained in U'. Furthermore, by definition, $g([0,\theta_2/2\pi) \cup (1 - \theta_1/2\pi, 1])$ is open in I/R since $[0,\theta_2/2\pi) \cup (1 - \theta_1/2\pi, 1]$ is open in the subset topology of $[0,1]$. By Proposition A.2.29, f is continuous.

Using a similar argument, it is not hard to show that f^{-1} is continuous, concluding that I/R is homeomorphic to $\mathbb{S}^1$.

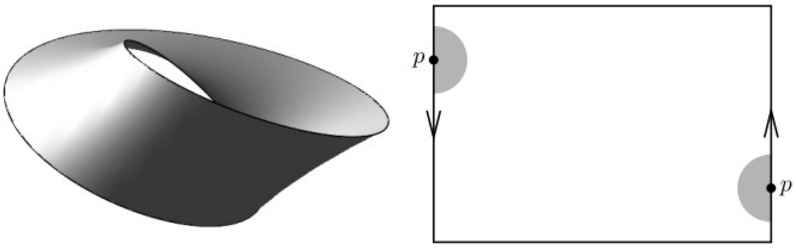

Figure A.10: Möbius strip.

Example A.2.47. Consider the real projective space $\mathbb{RP}^n$ given as the quotient space $\mathbb{S}^n/\sim$ where $\sim$ is the equivalence relation on $\mathbb{S}^n$ where $p_1 \sim p_2$ if and only if they are antipodal to each other, i.e., form a diameter of the sphere. The unit sphere $\mathbb{S}^n$ naturally inherits the subspace topology from $\mathbb{R}^{n+1}$. Definition A.2.44 provides the induced topology for $\mathbb{RP}^n$.

Example A.2.48 (Möbius Strip)**.** Let $I = [0,1] \times [0,1]$ be the unit square with the topology induced from $\mathbb{R}^n$. Define the identification (equivalence relation $\sim$) between points by $(0,y) \sim (1,1-y)$, for all $y \in [0,1]$ and no other points are equivalent to any others. The topological space obtained is called the *Möbius strip*. In $\mathbb{R}^3$, the Möbius strip can be viewed as a strip of paper twisted once and with ends glued together. Figure A.10 shows an embedding of the Möbius strip in $\mathbb{R}^3$, as well as a diagrammatic representation of the Möbius strip. In the diagrammatic representation, the arrows indicate that the opposite edges are identified but in inverse direction. The shaded area shows a disk around a point p on the identified edge.

A.2.4 Compactness

In any calculus course, we encounter the Extreme Value Theorem, a result in analysis that forms an essential ingredient to Rolle's Theorem, hence the Mean Value Theorem, and therefore the Fundamental Theorem of Calculus.

Theorem A.2.49 (Extreme Value Theorem)**.** *Let $f : [a,b] \to \mathbb{R}$ be a continuous real-valued function. Then f attains a maximum and a minimum over the interval $[a,b]$.*

As topology developed, a variety of attempts were made to generalize the idea contained in this fact. In the context of metric spaces, the key properties that would allow for a generalization are that the interval $[a,b]$ is closed and bounded. Though closed is a concept in topology, the notion of being bounded does not have an equivalent notion. It turns out that in topological spaces, the notion of *compactness* provides this desired generalization.

2> > 6

Definition A.2.50. Let (X, τ) be a topological space. Let $\mathcal{U} = \{U_i\}_{i \in I}$ be a collection of open sets in X. We call $\mathcal{U}$ an *open cover* of X if

$$X = \bigcup_{i \in I} U_i.$$

If $J \subset I$, the collection $\mathcal{V} = \{V_j\}_{j \in J}$ is called a *subcover* of $\mathcal{U}$ if $\mathcal{V}$ is itself an open cover of X.

Definition A.2.51. A topological space (X, τ) is called *compact* if every open cover of X has a finite subcover.

Remark A.2.52. A subset A of X is called compact if it is compact when equipped with the subspace topology induced from (X, τ). This is equivalent to the property that whenever there exists a collection of $\mathcal{U} = \{U_i\}_{i \in I}$ of open sets with $A \subset \bigcup_{i \in I} U_i$, there exists a finite subset $J \subset I$ with $A \subset \bigcup_{i \in J} U_j$.

Some examples of compact spaces are obvious, such as that any finite subset of a topological space is compact. However, the following theorem justifies, at least in part, the given definition of compactness.

Theorem A.2.53 (Heine-Borel)**.** *A closed and bounded interval* $[a, b]$ *of* $\mathbb{R}$ *is compact.*

Proof. (A variety of proofs exist for the Heine-Borel Theorem. The following proof is called the "creeping along" proof.)

Let $\mathcal{U} = \{U_\alpha\}_{\alpha \in I}$ be an open cover of $[a, b]$. Define the subset of $[a, b]$ by

$$E = \{x \in [a, b] \mid [a, x] \text{ is contained in a finite subfamily of } \mathcal{U}\}.$$

Obviously, E is an interval, but a priori we do not know whether it is open or closed or even nonempty. We will show that $b \in E$ to establish the theorem.

Now $a \in E$ because there exists some U_{α_0} such that $a \in U_{\alpha_0}$. Let $c = \text{lub } E$. Clearly $a \leq c \leq b$. Suppose that $c < b$. Since $\mathcal{U}$ covers $[a, b]$, there exists some index β such that $c \in U_\beta$, and therefore there exists ε such that $(c - \varepsilon, c + \varepsilon) \subset U_\beta$. Since c is the least upper bound of E, any $x \in (c - \varepsilon, c)$ is in E. Therefore, there exists a finite set $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ such that

$$[a, x] \subset \bigcup_{i=1}^{n} U_{\alpha_i}.$$

But then the finite union of open sets

$$(c - \varepsilon, c + \varepsilon) \cup \bigcup_{i=1}^{n} U_{\alpha_i}$$

contains the point $c + \varepsilon/2$, contradicting the assumption that $c < b$ and $c = \text{lub } E$. Thus $c = b$, and $b \in E$. $\square$

Compactness is an important property of topological spaces, and we refer the reader to Chapter 3 in [2], Section 6.6 in [15], or Chapter 7 in [27] for complete treatments. For the purposes of this book however, we are primarily interested in two results, namely Theorem A.2.57 and Theorem A.2.58, and a few necessary propositions to establish them. We will not prove Theorem A.2.57 completely but again refer the reader to the above sources.

Proposition A.2.54. *Let (X, τ) be a topological space, and let K be a compact subset of X. Every closed subset of K is compact.*

Proof. Let $F \subset K$ be a closed set. Then $X - F$ is open. If $\mathcal{U}$ is an open cover of F, then $\mathcal{U} \cup \{X - F\}$ is an open cover of K. Since K is compact, $\mathcal{U} \cup \{X - F\}$ must admit a finite subcover of K. This finite subcover of K is of the form $\mathcal{U}' \cup \{X - F\}$, where $\mathcal{U}'$ is a finite subcover of $\mathcal{U}$ of F. Thus, F is compact. □

Problem A.1.19 established the fact that two distinct points p_1 and p_2 in a metric space possess, respectively, open neighborhoods U_1 and U_2 such that $U_1 \cap U_2 = \emptyset$. This type of property is called a *separation* property of a topological space because it gives some qualification for how much we can distinguish points in the topological space. There exists a variety of separation axioms, but we only present the one that is relevant for differential geometry.

Definition A.2.55. A topological space (X, τ) is called *Hausdorff* if given any two points p_1 and p_2 in X, there exist open neighborhoods U_1 of p_1 and U_2 of p_2 such that $U_1 \cap U_2 = \emptyset$.

Proposition A.2.56. *If (X, τ) is a Hausdorff topological space, then every compact subset K of X is closed.*

Proof. Let K be compact. Since X is Hausdorff, for every $x \in X - K$ and $y \in K$, there exist open sets U_{xy} and V_{xy}, with $x \in U_{xy}$ and $y \in V_{xy}$, such that $U_{xy} \cap V_{xy} = \emptyset$. For each $x \in X - K$,

$$\{V_{xy} \mid y \in K\}$$

is an open cover of K, so it must possess a finite subcover that we index with a finite number of points $y_1, y_2, \dots, y_n$. But then for each x,

$$U_x = \bigcap_{i=1}^{n} U_{xy_i}$$

is open since it is a finite intersection of open sets. Since $K \subset \bigcup_{i=1}^{n} V_{xy_i}$, we conclude that $K \cap U_x = \emptyset$ for all $x \in X - K$. Thus, $X - K$ is a neighborhood of all of its points and hence it is open. Thus, K is closed. □

Theorem A.2.57. *Let A be any subset of $\mathbb{R}^n$ (equipped with the Euclidean topology). The set A is compact if and only if it is closed and bounded.*

Proof. (We only prove $\Longrightarrow$.)

Suppose that A is compact. Since by Problem A.1.19 any metric space is Hausdorff, Proposition A.2.56 allows us to conclude that A is closed. Since every open set in a metric space is the union of open balls, any open cover of A can be viewed as an open cover of open balls. If A is compact, it is contained in only a finite number of such open balls $\{B_{r_i}(p_i)\}_{i=1}^{m}$. There exists an open ball $B_r(p)$ that contains all the $B_{r_i}(p_i)$. The radius r will be less than $(m-1)\max\{d(p_i,p_j)\}+2\max\{r_i\}$. Then $K \subset B_r(P)$, and hence, K is bounded.

(The proof of the converse is more difficult and uses other techniques that we do not have the time to develop here.) $\square$

Theorem A.2.57 establishes that we might view closed and bounded subsets of $\mathbb{R}^n$ as the topological analog to $[a,b] \subset \mathbb{R}$ in Theorem A.2.49. We now complete the generalization to topological spaces.

Theorem A.2.58. *Let $f : X \to Y$ be a continuous function between topological spaces X and Y. If X is a compact space, then $f(X)$ is compact in Y.*

Proof. Let $\mathcal{U}$ be an open cover of $f(X)$. Since f is continuous, each $f^{-1}(U)$ is open and the collection $\{f^{-1}(U) \mid U \in \mathcal{U}\}$ is an open cover of X. Since X is compact, there exists a finite set $\{U_1, U_2, \ldots, U_n\} \subset \mathcal{U}$ such that

$$X = \bigcup_{i=1}^{n} f^{-1}(U_i).$$

Since for any functions $f(f^{-1}(A)) \subset A$ always and $f(\bigcup_\lambda A_\lambda) = \bigcup_\lambda f(A_\lambda)$, then

$$f(X) = f\Big(\bigcup_{i=1}^{n} f^{-1}(U_i)\Big) = \bigcup_{i=1}^{n} f(f^{-1}(U_i)) \subset \bigcup_{i=1}^{n} U_i\,.$$

Thus, $f(X)$ is compact. $\square$

Corollary A.2.59. *Let X be a compact topological space, and let $f : X \to \mathbb{R}$ be a real-valued function from X. Then f attains both a maximum and a minimum.*

Proof. By Theorem A.2.58, the image $f(X)$ is compact. By Theorem A.2.57, $f(X)$ is a closed and bounded subset of $\mathbb{R}$. Hence, lub $\{f(x) \mid x \in X\}$ and glb $\{f(x) \mid x \in X\}$ are both elements of $f(X)$ and hence are the maximum and minimum of f over X. $\square$

Corollary A.2.60. *Let (X,τ) and (Y,τ') be topological spaces, and let $f : X \to Y$ be a continuous function onto Y. If X is compact, then so is Y.*

Corollary A.2.61. *Let X be a compact topological space. If a space Y is homeomorphic to X, then Y is compact.*

'
/+01/ / /2 / & % !(6"
3 " > 7 , 7
> /+01/ / /2 / 30 6"
. C 3 0 , 7 $((6 2> . , " III
) * 6 (, ' : 0 , :

8 Curvature I: The Riemann Curvature Tensor

8.1 Curvature: Preliminary Remarks

We now come to one of the most important concepts of General Relativity and Riemannian Geometry, that of curvature and how to describe it in tensorial terms. Among other things, this will finally allow us to decide unambiguously if a given metric is just the (flat) Minkowski metric in disguise or the metric of a genuinely curved space (but a proof of this statement is postponed to section 11). More importantly (for present purposes) it will allow us to construct tensors that depend on the 2nd derivatives of the metric and will thus allow us to construct tensorial (generally covariant) differential equations for the metric. In particular, this will then lead us fairly directly to the Einstein equations (section 19), i.e. to the field equations for the gravitational field.

Recall that the equations that describe the behaviour of particles and fields in a gravitational field involve the metric and the Christoffel symbols determined by the metric. Thus the equations for the gravitational field should be generally covariant (tensorial) differential equations for the metric.

At first, here we seem to face a dilemma. How can we write down covariant differential equations for the metric when the covariant derivative of the metric is identically zero? Having come to this point, Einstein himself reached an impasse and required the help of his mathematician friend Marcel Grossmann ("Grossmann, you have to help me, or else I'll go crazy!") whom he had asked to investigate if there were any tensors that could be built from the second derivatives of the metric.

Grossmann soon found that this problem had indeed been addressed and solved in the mathematics literature, in particular by Riemann (generalising work of Gauss on curved surfaces), Ricci-Curbastro and Levi-Civita. It was shown by them that there are indeed non-trivial tensors that can be constructed from (ordinary) derivatives of the metric. These can then be used to write down covariant differential equations for the metric.[18]

The most important among these are the Riemann curvature tensor and its various contractions. In fact, it is known that these are the only tensors that can be constructed from the metric and its first and second derivatives, and they will therefore play a central role in all that follows.

Technically the most straightforward way of introducing the Riemann curvature tensor is via the commutator of covariant derivatives. In this section we will adopt this pragmatic (and relatively streamlined) approach, as it is sufficient to

---

[18] Of course, the story is not as simple and straightforward as that. For an account of Marcel Grossmann's (often overlooked) contributions to tensor calculus and the development of general relativity, see T. Sauer, *Marcel Grossmann and his contribution to the general theory of relativity*, `arXiv:1312.4068 [physics.hist-ph]`.

- determine the most important algebraic and differential properties of the curvature tensor (symmetries and Bianchi identities)
- assess its physical significance (gravitational tidal forces) via the influence of the curvature tensor on the motion of (families of) freely falling particles
- and to thus provide us with all the information and ingredients we need to then discuss the Einstein equations (section 19) and their formulation in terms of an action principle (section 20).

However, this is not geometrically the most intuitive way to introduce the concept of curvature, and it downplays the extent to which the curvature tensor reflects and encodes the geometric properties of space-time and, more generally, does not do justice to the fundamental differential geometric notion and significance of curvature. Some of these aspects are discussed in Part B of these notes, in particular in sections 11, 12, 13 and 14.

8.2 RIEMANN TENSOR FROM THE COMMUTATOR OF COVARIANT DERIVATIVES

As mentioned before, second covariant derivatives do not commute on (p,q)-tensors unless $p=q=0$. However, the fact that they *do* commute on scalars has the pleasant consequence that e.g. the commutator of covariant derivatives acting on a vector field V^μ does not involve any derivatives of V^μ. In fact, I will first show, without actually calculating the commutator, that

$$[\nabla_\mu, \nabla_\nu](\phi V^\lambda) = \phi[\nabla_\mu, \nabla_\nu]V^\lambda \tag{8.1}$$

for any scalar field ϕ. This implies that $[\nabla_\mu, \nabla_\nu]V^\lambda$ cannot depend on derivatives of V because if it did it would also have to depend on derivatives of ϕ.

Hence, the commutator can be expressed purely algebraically in terms of V. As the dependence on V is clearly linear, the commutator of covariant derivatives must then act like a linear transformation. There must therefore be an object $R^\lambda_{\ \sigma\mu\nu}$ such that

$$[\nabla_\mu, \nabla_\nu]V^\lambda = R^\lambda_{\ \sigma\mu\nu}V^\sigma \quad . \tag{8.2}$$

This can of course also be verified by a direct calculation, and we will come back to this below. For now let us just note that, since the left hand side of this equation is clearly a tensor for any V, the quotient theorem implies that the quantities $R^\lambda_{\ \sigma\mu\nu}$ are the components of a tensor.

Let us first verify (8.1). We have

$$\nabla_\mu\nabla_\nu\phi V^\lambda = (\nabla_\mu\nabla_\nu\phi)V^\lambda + (\nabla_\nu\phi)(\nabla_\mu V^\lambda) + (\nabla_\mu\phi)(\nabla_\nu V^\lambda) + \phi\nabla_\mu\nabla_\nu V^\lambda \quad . \tag{8.3}$$

Thus, upon taking the commutator the 2nd and 3rd terms drop out (because the 3rd is the symmetrisation of the 2nd), and we are left with

$$\begin{aligned}[\nabla_\mu, \nabla_\nu]\phi V^\lambda &= ([\nabla_\mu, \nabla_\nu]\phi)V^\lambda + \phi[\nabla_\mu, \nabla_\nu]V^\lambda \\ &= \phi[\nabla_\mu, \nabla_\nu]V^\lambda \ , \end{aligned} \tag{8.4}$$

where the last line follows from the fact that 2nd covariant derivatives do commute on scalars. Thus we have established (8.1).

By explicitly calculating the commutator, one can confirm the structure displayed in (8.2). This explicit calculation shows that the *Riemann-Christoffel Curvature Tensor* (or Riemann tensor for short) is given by

$$\boxed{R^\lambda_{\ \sigma\mu\nu} = \partial_\mu \Gamma^\lambda_{\ \sigma\nu} - \partial_\nu \Gamma^\lambda_{\ \sigma\mu} + \Gamma^\lambda_{\ \mu\rho}\Gamma^\rho_{\ \nu\sigma} - \Gamma^\lambda_{\ \nu\rho}\Gamma^\rho_{\ \mu\sigma}} \tag{8.5}$$

REMARKS:

1. Note how useful the quotient theorem is in this case. It would be quite unpleasant to have to verify the tensorial nature of this expression by explicitly checking its behaviour under coordinate transformations.
2. Note also that this tensor is clearly zero for the Minkowski metric written in Cartesian coordinates. Hence it is also zero for the Minkowski metric written in any other coordinate system. We will prove the converse, that vanishing of the Riemann curvature tensor implies that the metric is (locally) equivalent to the Minkowski metric, in section 11.2.
3. In the above we have defined the Riemann tensor by the relation (8.2) and then deduced the explicit expression (8.5). While this is, pragmatically speaking, a useful way of proceeding, it may be more logical to initially define the Riemann tensor in a different way, e.g. directly by (8.5) (for instance because by painful calculations one has discovered that this particular combination of non-tensorial objects miraculously happens to transform as a tensor). In that case, (8.2) is a result rather than a definition, known as the *Ricci identity.*

It is straightforward to extend the above to an action of the commutator $[\nabla_\mu, \nabla_\nu]$ on arbitrary tensors. For covectors we have, since we can raise and lower the indices with the metric with impunity,

$$\begin{aligned}[\nabla_\mu, \nabla_\nu]V_\rho &= g_{\rho\lambda}[\nabla_\mu, \nabla_\nu]V^\lambda \\ &= g_{\rho\lambda}R^\lambda_{\ \sigma\mu\nu}V^\sigma \\ &= R_{\rho\sigma\mu\nu}V^\sigma \\ &= R_\rho^{\ \sigma}{}_{\mu\nu}V_\sigma \ . \end{aligned} \tag{8.6}$$

We will see later that the Riemann tensor is anti-symmetric in its first two indices. Hence we can also write

$$[\nabla_\mu, \nabla_\nu] V_\rho = -R^\sigma_{\ \rho\mu\nu} V_\sigma \quad . \tag{8.7}$$

The extension to arbitrary (p, q)-tensors now follows the usual pattern, with one Riemann curvature tensor, contracted as for vectors, appearing for each of the p upper indices, and one Riemann curvature tensor, contracted as for covectors, for each of the q lower indices. Thus, e.g. for a $(2, 0)$-tensor $T^{\alpha\beta}$ one has

$$[\nabla_\mu, \nabla_\nu] T^{\alpha\beta} = R^\alpha_{\ \gamma\mu\nu} T^{\gamma\beta} + R^\beta_{\ \gamma\mu\nu} T^{\alpha\gamma} \tag{8.8}$$

and for a (1,1)-tensor $A^\lambda_{\ \rho}$ one has

$$[\nabla_\mu, \nabla_\nu] A^\lambda_{\ \rho} = R^\lambda_{\ \sigma\mu\nu} A^\sigma_{\ \rho} - R^\sigma_{\ \rho\mu\nu} A^\lambda_{\ \sigma} \quad . \tag{8.9}$$

I will give two other versions of the fundamental formula (8.2) which are occasionally useful and used.

1. Instead of looking at the commutator $[\nabla_\mu, \nabla_\nu]$ of two derivatives in the coordinate directions x^μ and x^ν, we can look at the commutator $[\nabla_X, \nabla_Y]$ of two directional covariant derivatives. Evidently, in calculating this commutator one will pick up new terms involving $\nabla_X Y^\mu - \nabla_Y X^\mu$. Denoting this vector field by $[X, Y[^\mu$ (the rationale for this notation will be explained in section 9.3), we can also write the formula for the curvature tensor as

$$([\nabla_X, \nabla_Y] - \nabla_{[X,Y]}) V^\lambda = R^\lambda_{\ \sigma\mu\nu} X^\mu Y^\nu V^\sigma \quad . \tag{8.10}$$

2. Secondly, one can consider a net of curves $x^\mu(s_1, s_2)$ parametrising, say, a two-dimensional surface, and look at the commutators of the covariant derivatives along the s_1- and s_2-curves. The formula one obtains in this case (it can be obtained from (8.10) by noting that X and Y commute in this case) is

$$(D_{s_1} D_{s_2} - D_{s_2} D_{s_1}) V^\lambda = R^\lambda_{\ \sigma\mu\nu} \frac{dx^\mu}{ds_1} \frac{dx^\nu}{ds_2} V^\sigma \quad , \tag{8.11}$$

where D_{s_k} denotes the covariant derivative along the curve parametrised by s_k, i.e. (section 5.7)

$$D_{s_k} = \frac{\partial x^\mu(s_1, s_2)}{\partial s_k} \nabla_\mu \quad . \tag{8.12}$$

8.3 SYMMETRIES AND ALGEBRAIC PROPERTIES OF THE RIEMANN TENSOR

A priori, the Riemann tensor has $256 = 4^4$ components in 4 dimensions. However, because of a large number of symmetries, the actual number of independent components is much smaller.

In general, to read off all the symmetries from the formula (8.5) is difficult. One way to simplify things is to look at the Riemann curvature tensor at the origin x_0 of a Riemann normal coordinate system (or some other inertial coordinate system). In that case, all the first derivatives of the metric disappear and only the first two terms of (8.5) contribute. One finds

$$\begin{aligned} R_{\alpha\beta\gamma\delta}(x_0) &= g_{\alpha\lambda}(\partial_\gamma \Gamma^\lambda_{\beta\delta} - \partial_\delta \Gamma^\lambda_{\beta\gamma})(x_0) \\ &= (\partial_\gamma \Gamma_{\alpha\beta\delta} - \partial_\delta \Gamma_{\alpha\beta\gamma})(x_0) \\ &= \tfrac{1}{2}(g_{\alpha\delta,\beta\gamma} + g_{\beta\gamma,\alpha\delta} - g_{\alpha\gamma,\beta\delta} - g_{\beta\delta,\alpha\gamma})(x_0) \ . \end{aligned} \tag{8.13}$$

In principle, this expression is sufficiently simple to allow one to read off all the symmetries of the Riemann tensor. However, it is more insightful to derive these symmetries in a different way, one which will also make clear *why* the Riemann tensor has these symmetries.

1. Anti-symmetry in the second pair of indices:

   $$R_{\alpha\beta\gamma\delta} = -R_{\alpha\beta\delta\gamma} \tag{8.14}$$

   This is obviously true from the definition or by construction.

2. Anti-symmetry in the first pair of indices:

   $$R_{\alpha\beta\gamma\delta} = -R_{\beta\alpha\gamma\delta} \tag{8.15}$$

   This is a consequence of the fact that the metric is covariantly constant. In fact, we can calculate

   $$\begin{aligned} 0 &= [\nabla_\gamma, \nabla_\delta] g_{\alpha\beta} \\ &= R_\alpha{}^\lambda{}_{\gamma\delta} g_{\lambda\beta} + R_\beta{}^\lambda{}_{\gamma\delta} g_{\alpha\lambda} \\ &= (R_{\alpha\beta\gamma\delta} + R_{\beta\alpha\gamma\delta}) \ . \end{aligned} \tag{8.16}$$

   As mentioned before, this implies that we can write the commutator of covariant derivatives on a covector as

   $$[\nabla_\mu, \nabla_\nu] V_\rho = R_\rho{}^\sigma{}_{\mu\nu} V_\sigma = -R^\sigma{}_{\rho\mu\nu} V_\sigma \ . \tag{8.17}$$

3. Cyclic permutation symmetry (or first Bianchi identity)

   $$R_{\alpha[\beta\gamma\delta]} = 0 \quad \Leftrightarrow \quad R_{\alpha\beta\gamma\delta} + R_{\alpha\delta\beta\gamma} + R_{\alpha\gamma\delta\beta} = 0 \tag{8.18}$$

   This Bianchi identity is a consequence of the fact that there is no torsion. In fact, applying $[\nabla_\gamma, \nabla_\delta]$ to the covector $\nabla_\beta \phi$, ϕ a scalar, one has

   $$\nabla_{[\gamma} \nabla_\delta \nabla_{\beta]} \phi = 0 \quad \Rightarrow \quad R^\lambda{}_{[\beta\gamma\delta]} \nabla_\lambda \phi = 0 \ . \tag{8.19}$$

As this has to be true for all scalars ϕ, this implies $R_{\alpha[\beta\gamma\delta]} = 0$ (to see this you could e.g. choose the (locally defined) coordinate functions $\phi(x) = x^\mu$ with $\nabla_\lambda \phi = \delta^\mu_{\ \lambda}$).

In turn this identity now implies that for any covector (not necessarily a gradient) one has the identity (also called Bianchi identity)

$$\nabla_{[\gamma}\nabla_\delta V_{\beta]} = -R^\alpha_{\ [\beta\gamma\delta]} V_\alpha = 0 \ . \tag{8.20}$$

4. Symmetry under exchange of the two pairs of indices

$$R_{\alpha\beta\gamma\delta} = R_{\gamma\delta\alpha\beta} \tag{8.21}$$

This identity, stating that the Riemann tensor is symmetric in its two pairs of indices, is not an independent symmetry but can be deduced from the three other symmetries by some not particularly interesting algebraic manipulations. One (quite possibly not optimal or minimal) possibility is

$$\begin{aligned} R_{\gamma\delta\alpha\beta} &\overset{(3)}{=} -R_{\gamma\alpha\beta\delta} - R_{\gamma\beta\delta\alpha} \\ &\overset{(2)}{=} R_{\alpha\gamma\beta\delta} + R_{\beta\gamma\delta\alpha} \\ &\overset{(3)}{=} -R_{\alpha\delta\gamma\beta} - R_{\alpha\beta\delta\gamma} - R_{\beta\alpha\gamma\delta} - R_{\beta\delta\alpha\gamma} \\ &\overset{(1,2)}{=} 2R_{\alpha\beta\gamma\delta} + R_{\delta\alpha\gamma\beta} R_{\delta\beta\alpha\gamma} \\ &\overset{(3)}{=} 2R_{\alpha\beta\gamma\delta} - R_{\delta\gamma\beta\alpha} \\ &\overset{(1,2)}{=} 2R_{\alpha\beta\gamma\delta} - R_{\gamma\delta\alpha\beta} \ , \end{aligned} \tag{8.22}$$

from which the claim follows.

Slightly more elegant (but equally obtuse) is the following argument.[19] Consider the matrix

$$\mathcal{R} = \begin{pmatrix} R_{\alpha\beta\gamma\delta} & R_{\alpha\delta\beta\gamma} & R_{\alpha\gamma\delta\beta} \\ R_{\delta\alpha\beta\gamma} & R_{\delta\gamma\alpha\beta} & R_{\delta\beta\gamma\alpha} \\ \vdots & \dots & \dots \end{pmatrix} \tag{8.23}$$

where the first column consists of the 4 cyclic permutations of all 4 indices of the Riemann tensor, while each row consists of the 3 cyclic permutations of the last 3 indices. Thus the sum σ_k of the entries of the k'th row is zero for all k (by symmetry (3)),

$$\sigma_k = \sum_{l=1}^{3} \mathcal{R}_{kl} \overset{(3)}{=} 0 \ , \tag{8.24}$$

while

$$\sigma_1 + \sigma_2 - \sigma_3 - \sigma_4 \overset{(1,2)}{=} 2R_{\alpha\beta\gamma\delta} - 2R_{\gamma\delta\alpha\beta} \ . \tag{8.25}$$

---

[19] See e.g. D. Bleecker, *Gauge Theory and Variational Principles.*

We can now count how many independent components the Riemann tensor really has. (1) implies that the second pair of indices can only take $N = (4 \times 3)/2 = 6$ independent values. (2) implies the same for the first pair of indices. (4) thus says that the Riemann curvature tensor behaves like a symmetric (6×6) matrix and therefore has $(6 \times 7)/2 = 21$ components. We now come to the remaining condition (3): if two of the indices in (3) are equal, (3) is equivalent to (4) and (4) we have already taken into account. With all indices unequal, (3) then provides one and only one more additional constraint. We conclude that the total number of independent components is 20.

Remarks:

1. Note that this agrees precisely with our previous counting in section 3.6 of how many of the second derivatives of the metric cannot be set to zero by a coordinate transformation: the second derivative of the metric has 100 independent components, to be compared with the $4 \times (4 \times 5 \times 6)/(2 \times 3) = 80$ components of the third derivatives of the coordinates. This also leaves 20 components. We thus see very explicitly that the Riemann curvature tensor contains all the coordinate independent information about the geometry up to second derivatives of the metric. In fact, it can be shown that in a Riemann normal coordinate system one has

$$g_{\mu\nu}(x) = \eta_{\mu\nu} + 0 + \tfrac{1}{3} R_{\mu\lambda\sigma\nu}(x_0)(x - x_0)^\lambda (x - x_0)^\sigma + \mathcal{O}((x - x_0)^3) \ . \qquad (8.26)$$

2. Just for the record, I note here that in general dimension $D = d + 1$ the Riemann tensor has $D^2(D^2 - 1)/12$ independent components. This number arises as

$$\begin{aligned} \frac{D^2(D^2-1)}{12} &= \frac{N(N+1)}{2} - \binom{D}{4} \\ N &= \frac{D(D-1)}{2} \end{aligned} \qquad (8.27)$$

and describes (as above) the number of independent components of a symmetric $(N \times N)$-matrix, now subject to $\binom{D}{4}$ conditions which arise from all the possibilities of choosing 4 out of D possible distinct values for the indices in (3). Just as for $D = 4$, this number of components of the Riemann tensor coincides with the number of second derivatives of the metric minus the number of independent components of the third derivatives of the coordinates determined in (3.176),

$$\frac{D(D+1)}{2} \times \frac{D(D+1)}{2} - D \times \frac{D(D+1)(D+2)}{2 \times 3} = \frac{D^2(D^2-1)}{12} \ . \qquad (8.28)$$

For $D = 2$ this formula predicts one independent component, and this is as it should be. Rather obviously the only independent non-vanishing component of the Riemann tensor in this case is R_{1212}. We will discuss curvature in 2 dimensions in more detail in sections 8.6 and 11.3 below.

Finally, a word of warning: there are a large number of sign conventions involved in the definition of the Riemann tensor (and its contractions we will discuss below), so whenever reading a book or article, in particular when you want to use results or equations presented there, make sure what conventions are being used and either adopt those or translate the results into some other convention. As a check: the conventions used here are such that $R_{\phi\theta\phi\theta}$ as well as the curvature scalar (to be introduced below) are *positive* for the standard metric on the two-sphere.

8.4 TIDAL FORCES: INFLUENCE OF CURVATURE ON PARTICLE TRAJECTORIES

In a certain sense the main effect of curvature (or gravity) is that initially parallel trajectories of freely falling non-interacting particles (dust, pebbles,…) do not remain parallel, i.e. that gravity is an attractive force that has the tendency to focus matter. This statement find its mathematically precise formulation in equations describing the influence of space-time curvature on the behaviour of (families of) geodesics.

Let us, as we will need this later anyway, recall first the situation in the Newtonian theory. One particle moving under the influence of a gravitational field is governed by the equation

$$\tfrac{d^2}{dt^2}x^i = -\partial^i\phi(x) \ , \tag{8.29}$$

where ϕ is the potential. Now consider a family of particles, or just two nearby particles, one at $x^i(t)$ and the other at $x^i(t)+\delta x^i(t)$. The other particle will of course obey the equation

$$\tfrac{d^2}{dt^2}(x^i+\delta x^i) = -\partial^i\phi(x+\delta x) \ . \tag{8.30}$$

From these two equations one can deduce an equation for δx itself, namely

$$\tfrac{d^2}{dt^2}\delta x^i = -\partial^i\partial_j\phi(x)\delta x^j \ . \tag{8.31}$$

It describes the effect of *gravitational tidal forces* (the gradient of the gravitational force) on a family of particles moving in a gravitational field.

In particular, when there is no gravitational force, and the trajectories are straight lines, one has

$$\tfrac{d^2}{dt^2}\delta x^i = 0 \quad \Rightarrow \quad \delta x^i = (\delta x^i)_0 + (\delta v^i)t \ . \tag{8.32}$$

Thus one recovers Euclid's parallel axiom, that two straight lines intersect at most once (for suitable choices of $\delta v^i \neq 0$) and that they never intersect when they are initially parallel ($\delta v^i = 0$). Any departure from this equation or its Minkowskian counterpart

$$\tfrac{d^2}{d\tau^2}\delta\xi^a = 0 \ , \tag{8.33}$$

will therefore indicate a departure from Euclidean geometry!

It is the counterpart of (8.31) that we will be seeking in the context of General Relativity. One derivation of this can be modelled on the Newtonian derivation above. It is elementary but looks non-covariant (and therefore somewhat messy) at intermediate stages of the calculation (see section 12.1 for a manifestly covariant derivation).

The starting point is of course the geodesic equation for x^μ and for its nearby partner $x^\mu + \delta x^\mu$,

$$\tfrac{d^2}{d\tau^2}x^\mu + \Gamma^\mu_{\ \nu\lambda}(x)\tfrac{d}{d\tau}x^\nu\tfrac{d}{d\tau}x^\lambda = 0 \ , \tag{8.34}$$

and

$$\tfrac{d^2}{d\tau^2}(x^\mu + \delta x^\mu) + \Gamma^\mu_{\ \nu\lambda}(x+\delta x)\tfrac{d}{d\tau}(x^\nu + \delta x^\nu)\tfrac{d}{d\tau}(x^\lambda + \delta x^\lambda) = 0 \ . \tag{8.35}$$

As above, from these one can deduce an equation for δx, namely

$$\tfrac{d^2}{d\tau^2}\delta x^\mu + 2\Gamma^\mu_{\ \nu\lambda}(x)\tfrac{d}{d\tau}x^\nu\tfrac{d}{d\tau}\delta x^\lambda + \partial_\rho\Gamma^\mu_{\ \nu\lambda}(x)\delta x^\rho\tfrac{d}{d\tau}x^\nu\tfrac{d}{d\tau}x^\lambda = 0 \ . \tag{8.36}$$

Now this does not look particularly covariant. Thus instead of in terms of $d/d\tau$ we would like to rewrite this in terms of the covariant operator D_τ, with

$$D_\tau \delta x^\mu = \frac{d}{d\tau}\delta x^\mu + \Gamma^\mu_{\ \nu\lambda}\frac{dx^\nu}{d\tau}\delta x^\lambda \ . \tag{8.37}$$

Calculating $(D_\tau)^2\delta x^\mu$, replacing $\ddot{x}^\mu$ appearing in that expression by $-\Gamma^\mu_{\ \nu\lambda}\dot{x}^\nu\dot{x}^\lambda$ (because x^μ satisfies the geodesic equation) and using (8.36), one eventually finds the nice covariant *geodesic deviation equation*

$$(D_\tau)^2\delta x^\mu = R^\mu_{\ \nu\lambda\rho}\dot{x}^\nu\dot{x}^\lambda\delta x^\rho \tag{8.38}$$

REMARKS:

1. This shows very clearly that curvature, as captured by the Riemann curvature tensor, leads to non-Euclidean geometry in which e.g. the parallel axiom is not necessarily satisifed.

2. In general, solutions to the geodesic deviation equation are called *Jacobi fields*. They describe the difference between the given geodesic and a (hypothetical) infinitely close neighbouring geodesic.

3. Clearly the present derivation of this result leaves something to be desired. It is also possible to give a manifestly covariant, and thus perhaps slightly more satisfactory, derivation of the above geodesic deviation equation, and we will return to this in section 12.1.

The Riemann tensor, as we have seen, is a four-index tensor. For many purposes this is not the most useful object, but we can create new tensors by contractions of the Riemann tensor. Due to the symmetries of the Riemann tensor, there is essentially only one possibility, namely the *Ricci tensor*

$$R_{\mu\nu} := R^{\lambda}_{\ \mu\lambda\nu} = g^{\lambda\sigma} R_{\sigma\mu\lambda\nu} \quad . \tag{8.39}$$

It arises naturally from the definition (8.2) of the Riemann tensor in terms of commutators of covariant derivatives, when one considers a contracted commutator,

$$[\nabla_\mu, \nabla_\nu] V^\lambda = R^{\lambda}_{\ \sigma\mu\nu} V^\sigma \quad \Rightarrow \quad [\nabla_\mu, \nabla_\nu] V^\mu = R^{\mu}_{\ \sigma\mu\nu} V^\sigma \equiv R_{\sigma\nu} V^\sigma \quad . \tag{8.40}$$

In particular, this identity explains why the Maxwell equations in the covariant Lorenz gauge (6.45) take the non-minimally coupled form (6.46).

It follows from the symmetries of the Riemann tensor that $R_{\mu\nu}$ is symmetric. Indeed

$$R_{\nu\mu} = g^{\lambda\sigma} R_{\sigma\nu\lambda\mu} = g^{\lambda\sigma} R_{\lambda\mu\sigma\nu} = R^{\sigma}_{\ \mu\sigma\nu} = R_{\mu\nu} \quad . \tag{8.41}$$

Thus, for $D = 4$, the Ricci tensor has 10 independent components, for $D = 3$ it has 6, while for $D = 2$ there is only 1 because there is only one independent component of the Riemann curvature tensor to start off with.

There is one more contraction of the Riemann tensor we can perform, namely on the Ricci tensor itself, to obtain what is called the *Ricci scalar* or *curvature scalar*

$$R := g^{\mu\nu} R_{\mu\nu} \quad . \tag{8.42}$$

REMARKS:

1. One might have thought that at least in four dimensions there is another way of constructing a (pseudo-)scalar, by contracting the Riemann tensor with the Levi-Civita tensor, but
$$\epsilon^{\mu\nu\rho\sigma} R_{\mu\nu\rho\sigma} = 0 \tag{8.43}$$
because of the Bianchi identity (cyclic symmetry of the Riemann tensor).

2. Note that for $D = 2$ the Riemann curvature tensor has as many independent components as the Ricci scalar, namely one, and that for $D = 3$ the Ricci tensor has as many components as the Riemann tensor, namely 6. Thus in $D = 2$ one can express the entire Riemann tensor in terms of the Ricci scalar (and the metric) alone, and one has
$$D = 2: \quad R_{\alpha\beta\gamma\delta} = \tfrac{1}{2}(g_{\alpha\gamma} g_{\beta\delta} - g_{\alpha\delta} g_{\beta\gamma}) R \tag{8.44}$$

(we will establish this relation in section 11.3, see (11.29)), while in $D = 3$ one has

$$D = 3: \quad \begin{aligned} R_{\alpha\beta\gamma\delta} = &(g_{\alpha\gamma}R_{\beta\delta} + R_{\alpha\gamma}g_{\beta\delta} - g_{\alpha\delta}R_{\beta\gamma} - R_{\alpha\delta}g_{\beta\gamma}) \\ &+ \tfrac{1}{2}(g_{\alpha\delta}g_{\beta\gamma} - g_{\alpha\gamma}g_{\beta\delta})R \end{aligned} \tag{8.45}$$

(and we will prove this in section 11.4).

3. It is thus only in four (and more) dimensions that there are strictly less components of the Ricci tensor than of the Riemann tensor. This has profound implications for the dynamics of gravity in these dimensions. In fact, we will see that it is only in dimensions $D > 3$ that gravity becomes truly dynamical, where empty space can be curved, where gravitational waves can exist etc.

4. Contracting (8.8), one consequence of the symmetry of the Ricci tensor is the useful general result

$$[\nabla_\mu, \nabla_\nu]T^{\mu\nu} = R_{\mu\nu}(T^{\mu\nu} - T^{\nu\mu}) = 0 \tag{8.46}$$

for any tensor $T^{\mu\nu}$. If $T^{\mu\nu} = F^{\mu\nu}$ is anti-symmetric, $F^{\mu\nu} = -F^{\nu\mu}$, it is not necessary to take the commutator, so one also has

$$F^{\mu\nu} = -F^{\nu\mu} \quad \Rightarrow \quad \nabla_\mu(\nabla_\nu F^{\mu\nu}) = \tfrac{1}{2}[\nabla_\mu, \nabla_\nu]F^{\mu\nu} = 0 \ . \tag{8.47}$$

Note that this can also be deduced (without knowing anything about curvature in general or the Ricci tensor in particular) from the general expression (5.64) for the divergence of an anti-symmetric tensor,

$$\nabla_\mu F^{\mu\nu} = g^{-1/2}\partial_\mu(g^{1/2}F^{\mu\nu}) \quad \Rightarrow \quad \nabla_\nu\nabla_\mu F^{\mu\nu} = g^{-1/2}\partial_\nu\partial_\mu(g^{1/2}F^{\mu\nu}) = 0 \ . \tag{8.48}$$

This is how we had shown in section 6.6 that the Maxwell equations imply covariant current conservation,

$$\nabla_\mu F^{\mu\nu} = -J^\nu \quad \Rightarrow \quad \nabla_\nu J^\nu = 0 \ . \tag{8.49}$$

Now we see that we can alternatively directly use the identity (8.47) to arrive at this result.

5. There are other scalars that can be built from the curvature tensor, but these are necessarily of higher order in the curvature tensor, such as (trivially) R^2 or (somewhat less trivially) $R_{\mu\nu}R^{\mu\nu}$ or the square of the Riemann tensor, the so-called *Kretschmann scalar*

$$K = R_{\mu\nu\rho\sigma}R^{\mu\nu\rho\sigma} \ . \tag{8.50}$$

Analogously, scalars can be built from higher powers of the Riemann tensor and or from powers of covariant derivatives of the Riemann tensor ($\Box R$ being the simplest example).

6. Such scalars are useful in analysing a given metric because, since they are scalars they are invariant under coordinate transformations. Thus they directly provide coordinate-invariant information about a metric. For instance if K is singular at some point in some coordinate system then it will be singular at that point in all coordinate systems, and thus such a singularity is not an artefact of a bad choice of coordinate system but a property of the space(-time) itself described by that metric. A prominent example is the singularity at the origin $r = 0$ of the Schwarzschild metric, unambiguously unveiled by the singularity of its Kretschmann scalar (27.163).

7. Contracting (8.40) with V^ν, one finds

$$V^\nu \nabla_\mu \nabla_\nu V^\mu - V^\nu \nabla_\nu \nabla_\mu V^\mu = R_{\mu\nu} V^\mu V^\nu \quad . \tag{8.51}$$

   Rewriting the first term as

$$V^\nu \nabla_\mu \nabla_\nu V^\mu = \nabla_\mu (V^\nu \nabla_\nu V^\mu) - (\nabla_\mu V^\nu)(\nabla_\nu V^\mu) \tag{8.52}$$

   this identity can be written as

$$V^\nu \nabla_\nu (\nabla_\mu V^\mu) + (\nabla_\mu V_\nu)(\nabla^\nu V^\mu) - \nabla_\mu (V^\nu \nabla_\nu V^\mu) + R_{\mu\nu} V^\mu V^\nu = 0 \quad . \tag{8.53}$$

   This is a very useful and versatile "master equation" which provides valuable information about the relation between vector fields and curvature when specialised e.g. to geodesic vector fields, $V^\nu \nabla_\nu V^\mu = 0$, or Killing vector fields, $\nabla_\mu V_\mu = -\nabla_\nu V_\mu$ and $\nabla_\mu V^\mu = 0$. Various specialisations of this equation will therefore appear later on in these notes, and even though we will then usually rederive them from scratch in the case at hand, it is good to keep in mind that e.g. (12.22) (our starting point for the discussion of the Raychaudhuri equation in section 12.2) and (13.12) (a useful identity relating Killing vectors and curvature) are special cases of (8.53).

8. As an a(far)side, and as an illustration of what one can do with (8.53), assume that V is such that its curl $\nabla_\mu V_\nu - \nabla_\nu V_\mu = 0$ and its divergence $\nabla_\mu V^\mu = 0$ are zero. Locally, the first condition has the solution $V_\mu = \partial_\mu f$, and then the second condition says that $\Box f = 0$, i.e. that f is *harmonic*. Therefore let us call a V^μ *harmonic* if it satisfies the above two conditions (to the mathematically more sophisticated: yes, I know that this is backwards, but we will specialise to the compact Riemannian case below . . .).

   For V harmonic in this sense, (8.53) reduces to

$$(\nabla_\mu V_\nu)(\nabla^\mu V^\nu) + R_{\mu\nu} V^\mu V^\nu = \nabla_\mu (V^\nu \nabla_\nu V^\mu) \quad . \tag{8.54}$$

   The simplest (albeit perhaps not of most direct relevance for physics) situation where one can deduce something of substance from this equation is when one has a Riemannian (i.e. positive-definite) metric and the space one is dealing with is compact, without boundary. Then (a) the first term is non-negative, and (b) upon integration over the space the total derivative term on the right-hand side gives zero upon use of the Gauss theorem (5.62) (discussed in some more detail in section 16.3).

This implies that for a harmonic V to exist on such a space, the integral of $R_{\mu\nu}V^\mu V^\nu$ must be non-positive. In particular,

- if the Ricci tensor is positive (as a quadratic form), there are no harmonic vector fields at all,
- and if $R_{\mu\nu}V^\mu V^\nu = 0$, then a harmonic vector field is necessarily covariantly constant, $\nabla_\mu V_\nu = 0$.

In more mathematical terms this means that the first Betti number of a compact manifold admitting a metric with positive Ricci curvature is equal to zero. A variant of this kind of argument for Killing vectors will be given in section 13.3.[20]

8.6 Example: Curvature Tensor of the 2-Sphere

To see how calculations of the curvature tensor can be done in practice, let us work out the example of the two-sphere of unit radius, i.e. with line element

$$ds^2 = d\theta^2 + \sin^2\theta d\phi^2 \equiv g_{ab}dx^a dx^b \quad . \tag{8.55}$$

We already know that the non-zero Christoffel symbols necessarily have two ϕ-indices and one θ-index (from $g_{\phi\phi} = \sin^2\theta$), and are given by

$$\Gamma^\phi{}_{\phi\theta} = \cot\theta \quad , \qquad \Gamma^\theta{}_{\phi\phi} = -\sin\theta\cos\theta \quad . \tag{8.56}$$

We also know that the Riemann curvature tensor has only one independent component. Let us therefore work out $R^\theta{}_{\phi\theta\phi}$. From the definition we find

$$R^\theta{}_{\phi\theta\phi} = \partial_\theta \Gamma^\theta{}_{\phi\phi} - \partial_\phi \Gamma^\theta{}_{\theta\phi} + \Gamma^\theta{}_{\theta c}\Gamma^c{}_{\phi\phi} - \Gamma^\theta{}_{\phi c}\Gamma^c{}_{\theta\phi} \quad . \tag{8.57}$$

The second and third terms are manifestly zero, and we are left with

$$R^\theta{}_{\phi\theta\phi} = \partial_\theta(-\sin\theta\cos\theta) + \sin\theta\cos\theta\cot\theta = \sin^2\theta \quad . \tag{8.58}$$

Thus we have

$$\begin{aligned} R^\theta{}_{\phi\theta\phi} &= R_{\theta\phi\theta\phi} = \sin^2\theta \\ R^\phi{}_{\theta\phi\theta} &= 1 \quad . \end{aligned} \tag{8.59}$$

---

[20] Theorems of this kind, and proved with the help of these kinds of techniques, were pioneered by S. Bochner and K. Yano in the 1930s - 1950s. See e.g. S. Bochner, K. Yano, *Curvature and Betti Numbers* or W. Poor, *Differential Geometric Structures* for more details, rigour and applications.

Therefore the Ricci tensor R_{ab} has the components

$$\begin{aligned} R_{\theta\theta} &= 1 \\ R_{\theta\phi} &= 0 \\ R_{\phi\phi} &= \sin^2\theta \ . \end{aligned} \tag{8.60}$$

These equations can succinctly be written as

$$R_{ab} = g_{ab} \ , \tag{8.61}$$

showing that the standard metric on the two-sphere is what we will later call an *Einstein metric.* The Ricci scalar R is

$$R = g^{\theta\theta}R_{\theta\theta} + g^{\phi\phi}R_{\phi\phi} = 1 + \frac{1}{\sin^2\theta}\sin^2\theta = 2 \ . \tag{8.62}$$

In particular, we have here our first concrete example of a space with non-trivial, in fact positive, curvature.

The result for the Riemann tensor can be written succinctly as

$$R^a{}_{bcd} = \delta^a{}_c g_{bd} - \delta^a{}_d g_{bc} \ , \tag{8.63}$$

which also immediately implies (8.61),

$$R_{bd} = R^a{}_{bad} = g_{bd} \ . \tag{8.64}$$

We will see later on, in section 14, that this form of the curvature tensor, or its equivalent,

$$R_{abcd} = g_{ac}g_{bd} - g_{ad}g_{bc} \ , \tag{8.65}$$

is characteristic of the curvature tensor of the sphere in any dimension.

8.7 More Examples: Curvature Tensor and Polar/Spherical Coordinates

We now turn to some variations of the above theme (and some other generalisations are discussed in section 11.3 below).

1. First of all, let us address the question what is the curvature (scalar) of a sphere of radius L, i.e. of the space with line element

   $$ds^2 = L^2(d\theta^2 + \sin^2\theta d\phi^2) \ . \tag{8.66}$$

   There are at least 3 ways to answer this question:

   - The first is to simply and blindly redo the above calculations in this case and to see what one gets.

- Alternatively, and somewhat more insightfully, rather than redoing the calculation in that case one can argue as follows. Let us observe first of all that the Christoffel symbols are invariant under constant rescalings of the metric because they are schematically of the form $g^{-1}\partial g$. Therefore the Riemann curvature tensor, which only involves derivatives and products of Christoffel symbols, is also invariant. Hence the Ricci tensor, which is just a contraction of the Riemann tensor, is also invariant:

  $$g_{ab} \to L^2 g_{ab} \quad \Rightarrow \quad \Gamma^a_{\ bc} \to \Gamma^a_{\ bc} \quad \Rightarrow \quad R^a_{\ bcd} \to R^a_{\ bcd} \quad \Rightarrow \quad R_{bd} \to R_{bd} \ . \tag{8.67}$$

  However, to construct the Ricci scalar, one needs the inverse metric. This introduces an explicit L-dependence and the result is that the curvature scalar of a sphere of radius L is $R = 2/L^2$,

  $$R(L^2 g_{ab}) = L^{-2} R(g_{ab}) = 2/L^2 \ . \tag{8.68}$$

  In particular, the curvature scalar of a large sphere is smaller than that of a small sphere, something which makes intuitve sense, a very large sphere locally "looking flatter" than a small sphere. However, one should use this intuition with care since, as we have seen, e.g. the Ricci tensor is independent of the size of the sphere.

- Finally, this result could also have been obtained on purely dimensional grounds. The curvature scalar is constructed from second derivatives of the metric. Hence it has length-dimension (-2). Therefore for a sphere of radius L, R has to be proportional to $1/L^2$. Comparing with the known result for $L = 1$ determines $R = 2/L^2$, as before.

2. Now let us consider, instead of the unit 2-sphere, the unit hyperboloid H^2 with metric (2.31)

   $$ds^2(H^2) = d\sigma^2 + \sinh^2\sigma \, d\phi^2 \ . \tag{8.69}$$

   It is clear that, apart from a few sign changes here and there, the calculation of the Riemann curvature tensor is identical to that for S^2. These sign changes ultimately lead to the conclusion that the curvature scalar of H^2 is (-2). While the sphere is the prototypical example of a space with positive curvature, the hyperboloid is the prototypical example of a space with negative curvature.

   Instead of just doing the calculation for this specific example, it is slightly more instructive to do it for the class of metrics

   $$ds^2 = dx^2 + f(x)^2 d\phi^2 \ , \tag{8.70}$$

   for some (for the time being unspecified) function $f = f(x)$. Denoting the derivative with respect to x by a prime, $f'(x) = df/dx$, one finds (this is a simple but ionstructive exercise) that the Ricci tensor and Ricci scalar are

$$R_{ab} = -(f''/f)g_{ab} \quad , \quad R(x) = -2f''(x)/f(x) \ . \tag{8.71}$$

In particular, for the Euclidean metric and the standard metrics on the sphere and the hyperboloid one finds

$$f(x) = \begin{cases} x & (R^2) \\ \sin x & (S^2) \\ \sinh x & (H^2) \end{cases} \quad \Rightarrow \quad R = \begin{cases} 0 \\ +2 \\ -2 \end{cases} \tag{8.72}$$

In 2 dimensions, R is related to the *Gauss Curvature* K of a surface by $K = R/2$ so that $K = 0, \pm 1$ in these examples. See section 11.3 for some more information.

3. Now let us promote the constant radius L of S^2 to a new radial coordinate r and ask the question what is the curvature tensor of the 3-dimensional space with coordinates $(r, x^a) = (r, \theta, \phi)$ and line element

$$ds^2 = dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2) \ . \tag{8.73}$$

On the one hand, because one seems to have just added a trivial r-direction to the 2-sphere, one might be tempted to suspect that also this 3-dimensional space has non-trivial curvature. On the other hand, we recognise the above metric as the Euclidean metric on $\mathbb{R}^3$, written in spherical coordinates, and as such we expect its curvature (in fact, all components of the Riemann tensor) to be zero.

The latter expectation is of course borne out, but it is instructive to see explicitly how this cancellation occurs. In fact, it will be even more instructive to consider an apparently harmless and innocuous modification of the above metric which consists in replacing dr^2 by some constant multiple of dr^2,

$$ds^2 = p\, dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2) \ . \tag{8.74}$$

Equivalently, up to a truly harmless overall constant factor, we can think of this as the Euclidean metric, but with the metric on the unit-sphere replaced by that of a sphere of radius $1/\sqrt{p} \neq 1$),

$$ds^2 = p\ \left(dr^2 + (r^2/p)(d\theta^2 + \sin^2\theta d\phi^2)\right) \ . \tag{8.75}$$

Proceeding in a pedestrian way, we thus have a metric $g_{\alpha\beta}$ with components

$$g_{rr} = p \quad , \quad g_{ar} = 0 \quad , \quad g_{ab} = r^2\gamma_{ab} \ , \tag{8.76}$$

with γ_{ab} in this example denoting the components of the metric on the unit sphere (and with $\gamma^a_{\ bc}$ and $r^a_{\ bcd}$ its associated Christoffel symbols and components of the Riemann curvature tensor determined in the previous section). From these we can deduce that for $r > 0$ the non-trivial Christoffel symbols are

$$\Gamma^r_{\ ab} = -p^{-1}r\gamma_{ab} \quad , \quad \Gamma^a_{\ br} = r^{-1}\delta^a_b \quad , \quad \Gamma^a_{\ bc} = \gamma^a_{\ bc} \ . \tag{8.77}$$

From this, in turn, one finds that all the components of the Riemann tensor involving at least one r-index are zero, whereas for the purely angular components one finds

$$R^a{}_{bcd} = r^a{}_{bcd} + \Gamma^a{}_{cr}\Gamma^r{}_{bd} - \Gamma^a{}_{dr}\Gamma^r{}_{bc} \quad . \tag{8.78}$$

Using (8.63) and (8.77), one sees that

$$R^a{}_{bcd} = (1-p^{-1})r^a{}_{bcd} = (1-p^{-1})(\delta^a{}_c\gamma_{bd} - \delta^a{}_d\gamma_{bc}) \quad . \tag{8.79}$$

Therefore precisely for $p=1$ the two contributions to the curvature tensor indeed cancel and the curvature tensor is identically zero, as expected.

Equally interesting is the fact that for $p \neq 1$ the curvature is non-zero even away from $r=0$ (in addition, there is a conical deficit angle singularity at $r=0$, as in the next example below, but this shall not be our concern here). In particular it follows from the above result that the only non-vanishing components of the Ricci tensor of this 3-dimensional space are

$$R^a{}_{bcd} = (1-p^{-1})r^a{}_{bcd} \quad \Rightarrow \quad R_{bd} = (1-p^{-1})r_{bd} = (1-p^{-1})\gamma_{bd} \quad . \tag{8.80}$$

Therefore also its Ricci scalar is non-zero,

$$R = g^{\alpha\beta}R_{\alpha\beta} = g^{ab}R_{ab} = 2(1-p^{-1})r^{-2} \quad . \tag{8.81}$$

We also see from this that this space actually has a curvature singularity as $r \to 0$. Since the Ricci scalar is a scalar (under coordinate transformations), this divergence cannot be an artefact of a bad choice of coordinates, and indicates that there is a genuine geometric singularity for $r \to 0$.

Extended to a four-dimensional space-time metric via

$$ds^2 = -dt^2 + p\,dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2) \quad , \tag{8.82}$$

this describes the gravitational field outside a "monopole".[21]

4. As a final variation of this theme, we consider the above example in one dimension less, i.e. we look at the metric one obtains if one replaces the Euclidean metric on $\mathbb{R}^2$ written in polar coordinates by

   $$dr^2 + r^2 d\phi^2 \to p\,dr^2 + r^2 d\phi^2 \quad , \tag{8.83}$$

   where the angle ϕ has period 2π.

   In this case there is an interesting twist (pun intended) and the situation is somewhat different. Pulling out the factor of p, one sees that (up to this irrelevant overall constant factor) the metric can be written as

   $$ds^2 = dr^2 + r^2 d(\phi/\sqrt{p})^2 \equiv dr^2 + C^2 r^2 d\phi^2 \quad . \tag{8.84}$$

---

[21] M. Barriola, A. Vilenkin, *Gravitational Field of a Global Monopole*, Phys. Rev. Lett. 63 (1989) 341-343.

This would be the standard Euclidean metric on $\mathbb{R}^2$ either for $p = 1$ or if the angle ϕ had periodicity $2\pi\sqrt{p}$, but since ϕ has period 2π, this results in a misidentification of the points in a plane, like when one rolls up a flat piece of paper into a cone. Away from $r = 0$, this space is intrinsically flat (all the components of the Riemann curvature tensor are zero, as one can easily calculate - see section 11.1 for an explanation of this use of the word "intrinsic"). There is, however, a conical singularity at the tip of the cone $r = 0$, which can be thought of as a δ-function contribution to the curvature localised at $r = 0$. Extended to a four-dimensional space-time metric,

$$ds^2 = -dt^2 + dz^2 + dr^2 + C^2r^2d\phi^2 \ , \qquad (8.85)$$

it can be interpreted as the space-time metric of an idealised cosmic string extended in the z-direction.[22]

8.8 BIANCHI IDENTITIES AND THE EINSTEIN TENSOR

So far, we have discussed algebraic properties of the Riemann tensor. The Riemann tensor also satisfies some differential identities which, in particular in their contracted form, will be of fundamental importance in the following.

The first identity is easy to derive. As a (differential) operator the covariant derivative clearly satisfies the Jacobi identity

$$[\nabla_{[\mu}, [\nabla_\nu, \nabla_{\lambda]}]] = 0 \qquad (8.86)$$

(total anti-symmetrisation over all 3 indices). Since the commutator $[\nabla_\nu, \nabla_\lambda]$ is already anti-symmetric in the indices ν, λ, this anti-symmetrisation is equivalent to cyclic permutation of the 3 indices,

$$[\nabla_{[\mu}, [\nabla_\nu, \nabla_{\lambda]}]] = 0 \quad \Leftrightarrow \quad [\nabla_\mu, [\nabla_\nu, \nabla_\lambda]] + \circlearrowleft (\mu, \nu, \lambda) = 0 \ . \qquad (8.87)$$

If you do not believe this identity (valid for any 3 associative linear operators), you can just write out the twelve relevant terms explicitly to see that there is indeed a complete cancellation:

---

[22] It is far from straightforward, however, to find a formalism which allows one to calculate and derive the distributional Riemann tensor of this space-time - see R. Geroch, J. Traschen, *Strings and other distributional sources in general relativity*, Phys. Rev. D36 (1987) 1017-1031 for a general analysis of the problem and issues arising in this and related contexts, C. Clarke, J. Vickers, J. Wilson, *Generalized functions and distributional curvature of cosmic strings*, Class. Quantum Grav. 13 (1996) 2485-2498 for one approach (based on the Colombeau algebra of distributions), and D. Garfinkle, *Metrics with distributional curvature*, `arXiv:gr-qc/9906053` for a different approach. We will (mostly) stay away from distributional curvatures in these notes.

$$\begin{aligned}
[\nabla_{[\mu},[\nabla_\nu,\nabla_{\lambda]}]] &\sim \nabla_\mu\nabla_\nu\nabla_\lambda-\nabla_\mu\nabla_\lambda\nabla_\nu-\nabla_\nu\nabla_\lambda\nabla_\mu+\nabla_\lambda\nabla_\nu\nabla_\mu\\
&+\nabla_\lambda\nabla_\mu\nabla_\nu-\nabla_\lambda\nabla_\nu\nabla_\mu+\nabla_\nu\nabla_\mu\nabla_\lambda-\nabla_\mu\nabla_\nu\nabla_\lambda\\
&+\nabla_\nu\nabla_\lambda\nabla_\mu-\nabla_\nu\nabla_\mu\nabla_\lambda-\nabla_\lambda\nabla_\mu\nabla_\nu+\nabla_\mu\nabla_\lambda\nabla_\nu\\
&= 0\ . \qquad (8.88)
\end{aligned}$$

To determine the implications of this identity for the Riemann tensor, we apply it to a vector field V, say. The first term in (8.87) is

$$\begin{aligned}
[\nabla_\mu,[\nabla_\nu,\nabla_\lambda]]V^\rho &= \nabla_\mu(R^\rho_{\ \sigma\nu\lambda}V^\sigma)-[\nabla_\nu,\nabla_\lambda](\nabla_\mu V^\rho)\\
&=(\nabla_\mu R^\rho_{\ \sigma\nu\lambda})V^\sigma+R^\rho_{\ \sigma\nu\lambda}\nabla_\mu V^\sigma-R^\rho_{\ \sigma\nu\lambda}\nabla_\mu V^\sigma+R^\sigma_{\ \mu\nu\lambda}\nabla_\sigma V^\rho \qquad (8.89)\\
&=(\nabla_\mu R^\rho_{\ \sigma\nu\lambda})V^\sigma+R^\sigma_{\ \mu\nu\lambda}\nabla_\sigma V^\rho\ .
\end{aligned}$$

Upon taking the cyclic permutations, the sum of the 2nd terms vanishes by the cyclic symmetry of the Riemann tensor, and therefore one finds

$$(\nabla_\mu R^\rho_{\ \sigma\nu\lambda})V^\sigma+\circlearrowleft(\mu,\nu,\lambda)=0\ . \qquad (8.90)$$

Since this holds for any V, one deduces the *Bianchi identity*

$$\nabla_\mu R^\rho_{\ \sigma\nu\lambda}+\circlearrowleft(\mu,\nu,\lambda)=0 \quad\Leftrightarrow\quad \nabla_{[\mu}R^\rho_{\ |\sigma|\nu\lambda]}=0 \qquad (8.91)$$

(where $|\sigma|$ indicates that this index is to be excluded from the anti-symmetrisation). Using the symmetry (IV) of the Riemann tensor, this can equivalently be written as

$$\nabla_{[\mu}R_{\nu\lambda]\rho\sigma}=0\ . \qquad (8.92)$$

We will mainly be interested in a (double) contraction of this identity. To that end we write out (8.91) explicitly as

$$\nabla_\lambda R_{\alpha\beta\mu\nu}+\nabla_\nu R_{\alpha\beta\lambda\mu}+\nabla_\mu R_{\alpha\beta\nu\lambda}=0\ . \qquad (8.93)$$

By contracting this with $g^{\alpha\mu}$ we obtain

$$\nabla_\lambda R_{\beta\nu}-\nabla_\nu R_{\beta\lambda}+\nabla_\mu R^\mu_{\ \beta\nu\lambda}=0\ . \qquad (8.94)$$

This is not yet particularly useful. To also turn the last term into a Ricci tensor we contract once more, with $g^{\beta\lambda}$ to obtain the contracted Bianchi identity

$$\nabla_\lambda R^\lambda_\nu-\nabla_\nu R+\nabla_\mu R^\mu_\nu=0\ , \qquad (8.95)$$

or

$$\nabla^\mu(R_{\mu\nu}-\tfrac{1}{2}g_{\mu\nu}R)=0\ . \qquad (8.96)$$

The tensor appearing in this equation is the so-called *Einstein tensor* $G_{\mu\nu}$,

$$G_{\mu\nu}=R_{\mu\nu}-\tfrac{1}{2}g_{\mu\nu}R\ . \qquad (8.97)$$

It is the unique divergence-free tensor that can be built from the metric and its first and second derivatives (apart from $g_{\mu\nu}$ itself, of course),

$$\nabla^\mu G_{\mu\nu} = 0 \ , \tag{8.98}$$

and this is why it will play *the* central role in the Einstein equations for the gravitational field.

A minor *caveat* regarding the above statement about the uniqueness of the Einstein tensor is that, as it stands, it is only true in $D = 4$ space-time dimensions. In $D > 4$, there are other tensors with this property, but they are non-linear in 2nd derivatives of the metric. The uniqueness statement continues to be true for $D > 4$ if one adds the requirement that the tensor is linear in 2nd derivatives of the metric. I will briefly come back to this in the discussion of the action principle for general relativity in section 20.1.

8.9 RIEMANN NORMAL COORDINATES REVISITED

In section 3.6 we had introduced Riemann normal coordinates as a special class of inertial coordinate systems, based on geodesics. The main idea was to introduce new coordinates $x^\alpha \to \xi^a$ in such a way that the coordinate lines of the new coordinates ξ^a are geodesics passing through the point p at which one wants to erect this coordinate system.

In particular,

1. we considered the Taylor expansion (3.167)

   $$x^\alpha(\tau) = x_0^\alpha + \tau u_0^\alpha - \tfrac{1}{2}\tau^2 \Gamma^\alpha_{\beta\gamma}(x_0) u_0^\beta u_0^\gamma + \dots \ , \tag{8.99}$$

   of a solution to the geodesic equation;

2. this led us to consider the coordinate transformation (3.169)

   $$x^\alpha(\xi) = x_0^\alpha + (\xi - \xi_0)^a e_a^\alpha - \tfrac{1}{2}(\xi - \xi_0)^b (\xi - \xi_0)^c \Gamma^\alpha_{\beta\gamma}(x_0) e_b^\beta e_c^\gamma + \dots \tag{8.100}$$

   which has the property that the lines

   $$\xi^a(\tau) = \xi_0^a + \tau\lambda^a \tag{8.101}$$

   are geodesics for any constant λ^a;

3. for these geodesics the geodesic equation reduces to

   $$\ddot{\xi}^a + \Gamma^a_{bc}\dot{\xi}^b\dot{\xi}^c = 0 \quad \Rightarrow \quad \Gamma^a_{bc}(\xi_0^a + \tau\lambda^a)\lambda^b\lambda^c = 0 \ , \tag{8.102}$$

   implying at $\tau = 0$

   $$\Gamma^a_{bc}(\xi_0)\lambda^b\lambda^c = 0 \quad \forall\lambda^a \quad \Rightarrow \Gamma^a_{bc}(\xi_0) = 0 \tag{8.103}$$

   (and we will look at the implications of the next term in the Taylor expansion of (8.102) below).

Therefore the Taylor expansion of the metric around $\xi = \xi_0$ has the form

$$g_{ab}(\xi) = \eta_{ab} + \tfrac{1}{2}(\xi-\xi_0)^c(\xi-\xi_0)^d g_{ab,\,cd}(\xi_0) + \dots \quad , \tag{8.104}$$

and we will now determine the quadratic term in this expansion (and be able to express it in terms of the components of the Riemann tensor $R_{abcd}(\xi_0)$ at the point p in these coordinates). To that end we look at the next term in the Taylor expansion of (8.102). Thus we differentiate (8.102) along the geodesic, i.e. with respect to τ, and evaluate the results at $\tau = 0$ to deduce

$$\partial_d \Gamma^a_{\ bc}(\xi_0)\lambda^d\lambda^b\lambda^c = 0 \quad \forall \ \lambda^b \tag{8.105}$$

or, equivalently

$$\partial_{(d}\Gamma^a_{\ bc)}(\xi_0) = 0 \quad \Leftrightarrow \quad \partial_d\Gamma^a_{\ bc}(\xi_0) + \partial_b\Gamma^a_{\ cd}(\xi_0) + \partial_c\Gamma^a_{\ db}(\xi_0) = 0 \ . \tag{8.106}$$

This condition, and the analogous conditions

$$\partial_{(d\ldots e}\Gamma^a_{\ bc)}(\xi_0) = 0 \tag{8.107}$$

arising from the higher-order terms in the Taylor expansion of (8.102) impose constraints on the Christoffel symbols and their derivatives that are satisfied in Riemann normal coordinates (but not in general inertial coordinate systems).

A useful way of reexpressing the condition (8.106) is the following (a certain amount of hindsight or trial-and-error is required for this): because $\Gamma^a_{\ bc}(\xi_0) = 0$, from the definition of the Riemann tensor we have

$$\begin{aligned} R^a_{\ bcd}(\xi_0) + R^a_{\ cbd}(\xi_0) &= \partial_c\Gamma^a_{\ bd}(\xi_0) - \partial_d\Gamma^a_{\ bc}(\xi_0) + \partial_b\Gamma^a_{\ cd}(\xi_0) - \partial_d\Gamma^a_{\ cb}(\xi_0) \\ &= \partial_c\Gamma^a_{\ bd}(\xi_0) + \partial_b\Gamma^a_{\ cd}(\xi_0) - 2\partial_d\Gamma^a_{\ bc}(\xi_0) \quad , \end{aligned} \tag{8.108}$$

and using (8.106) this can be written as

$$\partial_d\Gamma^a_{\ bc}(\xi_0) = -\tfrac{1}{3}(R^a_{\ bcd}(\xi_0) + R^a_{\ cbd}(\xi_0)) \ . \tag{8.109}$$

On the other hand, from

$$g_{ab,\,c}(\xi) = \Gamma_{abc}(\xi) + \Gamma_{bac}(\xi) \tag{8.110}$$

we have

$$g_{ab,\,cd}(\xi) = \partial_d\Gamma_{abc}(\xi) + \partial_d\Gamma_{bac}(\xi) \tag{8.111}$$

and at ξ_0 we can use (8.109) and the symmetries of the Riemann tensor to deduce

$$\begin{aligned} g_{ab,\,cd}(\xi_0) &= -\tfrac{1}{3}(R_{abcd} + R_{acbd} + R_{bacd} + R_{bcad})(\xi_0) \\ &= -\tfrac{1}{3}(R_{acbd} + R_{bcad})(\xi_0) = -\tfrac{1}{3}(R_{acbd} + R_{adbc})(\xi_0) \ . \end{aligned} \tag{8.112}$$

We have thus found that, to quadratic order in a Taylor expansion of the metric around the origin of a Riemann normal coordinate system, the metric can be written as

$$\begin{aligned} g_{ab}(\xi) &= \eta_{ab} - \tfrac{1}{6}(R_{acbd}(\xi_0) + R_{adbc}(\xi_0))(\xi-\xi_0)^c(\xi-\xi_0)^d + \mathcal{O}(\xi^3) \\ &= \eta_{ab} - \tfrac{1}{3}R_{acbd}(\xi_0)(\xi-\xi_0)^c(\xi-\xi_0)^d + \mathcal{O}(\xi^3) \ . \end{aligned} \tag{8.113}$$

If required, higher order terms can be determined analogously with the help of the higher order terms in the Taylor expansion of (8.102), and (with a steady hand) can be expressed in terms of the covariant derivatives of the Riemann tensor at ξ_0.

8.10 PRINCIPLE OF MINIMAL COUPLING REVISITED

In sections 4.1 and 6.1 on the principles of general covariance and minimal coupling respectively, I mentioned that these do not necessarily fix the equations uniquely. In other words, there could be more than one generally covariant equation which reduces to a given equation in Minkowski space. Having the curvature tensor at our disposal now, we can construct examples of this kind.

Given some tensorial equation, obtained by the minimal coupling prescription, say, one can always contemplate the possiblity to add additional terms to it involving the curvature tensor. Since such terms take the form of higher derivative corrections to the original equation, multiplied by appropriate dimensionful constants, one can usually get away with ignoring such terms when dealing with weak fields and other low-energy phenomena, and under such conditions the minimal coupling rule can usually be trusted. However, such terms are not negligible under extreme conditions involving e.g. very strong or strongly fluctuating gravitational fields.

An example which shows very clearly that the minimal coupling prescription, at least the way we have formulated it, is itself ambiguous is, as already briefly pointed out in section 6.6, provided by Maxwell theory. In that case, we saw that in the covariant Lorenz gauge one has (6.45)

$$\nabla_\mu A^\mu = 0 \quad \Rightarrow \quad \nabla_\mu F^{\mu\nu} = \nabla_\mu(\nabla^\mu A^\nu - \nabla^\nu A^\mu) = \Box A^\nu - [\nabla_\mu, \nabla^\nu]A^\mu \quad , \tag{8.114}$$

where $\Box A^\nu = \nabla^\mu \nabla_\mu A^\nu$. It thus follows from (8.40) that the Maxwell equations in the covariant Lorenz gauge can be written as (6.46)

$$\nabla_\mu A^\mu = 0 \quad \Rightarrow \quad \nabla_\mu F^{\mu\nu} = -J^\nu \quad \rightarrow \quad \Box A^\nu - R^\nu_{\ \mu} A^\mu = -J^\nu \quad . \tag{8.115}$$

What this shows is that "minimal coupling" all by itself is not a unique prescription, as we would have obtained (8.115) without the curvature terms by applying the minimal coupling prescription to the special relativity Maxwell equation in the Lorenz gauge, namely just $\Box A_\nu = -J_\nu$.

In the present situation, (6.46) is superior to the equation without the curvature term because

- it follows from a variational principle (involving the minimally coupled counterpart of the Maxwell action)

- and (related to this) because (8.115) implies that the current is covariantly conserved (as we had verified in section 6.6 in an arbitrary gauge), while for the equation without the curvature term covariant current conservation would then be violated by a curvature term (as can easily be verified).

Thus occasionally some such additional criteria can be used to eliminate (or reduce) the ambiguity in the minimal coupling prescription, but this need not always be the case.

As another example, consider the wave equation for a (massless, say) scalar field Φ. In Minkowski space, this is the Klein-Gordon equation which has the obvious curved space analogue (5.56)

$$\Box\Phi = 0 \tag{8.116}$$

obtained by the minimal coupling description. However, one could equally well postulate the equation

$$(\Box + \xi R)\Phi = 0 \ \ , \tag{8.117}$$

where ξ is a (dimensionless) constant and R is the scalar curvature. This equation is generally covariant, and reduces to the ordinary Klein-Gordon equation in Minkowski space, so this is an acceptable curved-space extension of the wave equation for a scalar field. This equation of motion arises (in D sapce-time dimensions) from the action

$$S_\xi[\phi, g_{\alpha\beta}] = -\tfrac{1}{2}\int \sqrt{g} d^D x \ \left(g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi + \xi R\phi^2\right) \ \ , \tag{8.118}$$

exhibiting the non-minimal (yet generally covariant) coupling of the scalar field to gravity via the term $\xi R\phi^2$ (which acts as an x-dependent mass term for the scalar field).

Moreover, ξ is dimensionless, so one cannot argue that on dimensional grounds this ambiguity is irrelevant for weak fields. Indeed, one frequently postulates a specific non-zero value for ξ which makes the wave equation *conformally invariant* (invariant under position-dependent Weyl rescalings of the metric) for massless fields, and this criterion can be imposed to select a particular non-zero value for ξ (e.g. for a 4-dimensional space-time this turns out to be the value $\xi = 1/6$). This will be discussed and explained in section 22.3.

Thus in general such ambiguities are present and are something one has to live with.

11 CURVATURE II: GEOMETRY AND CURVATURE

In this section, we will first discuss two properties of the Riemann curvature tensor that illustrate its geometric significance and thus, *a posteriori*, justify equating the commutator of covariant derivatives with the intuitive concept of curvature. These properties are

- the *path-dependence of parallel transport* in the presence of curvature,
- the fact that the space-time metric is equivalent to the (in an obvious sense flat) Minkowski metric if and only if the Riemann curvature tensor vanishes.

We then briefly discuss some other general aspects of the relation between geometry and curvature (while the interplay between geodesics and curvature and Killing vectors and curvature will be discussed in sections 12 and 13 respectively).

11.1 INTRINSIC GEOMETRY, CURVATURE AND PARALLEL TRANSPORT

The Riemann curvature tensor and its relatives, introduced above, measure the *intrinsic geometry* and curvature of a space or space-time. This means that they can be calculated by making experiments and measurements in the space itself. Such experiments might involve things like checking if the interior angles of a triangle add up to π or not.

This intrinsic geometry and curvature described above should be contrasted with the *extrinsic geometry* which depends on how the space may be embedded in some larger space. As we have no intention of embedding space-time into something higher dimensional, we will mainly be concerned with intrinsic geometry in the following. However, if you would for example be interested in the properties of spacelike hypersurfaces in space-time, then aspects of both intrinsic and extrinsic geometry of that hypersurface would be relevant. See section 18 for some further comments on this.

Let us return to intrinsic geometry. An even better method, the subject of this section, to determine the curvature is to check the properties of parallel transport. The tell-tale sign (or smoking gun) of the presence of curvature is the fact that parallel transport is path dependent, i.e. that parallel transporting a vector V from a point A to a point B along two different paths will in general produce two different vectors at B. Another way of saying this is that parallel transporting a vector around a closed loop at A will in general produce a new vector at A which differs from the initial vector.

This is easy to see in the case of the two-sphere, for which we also worked out explicitly the parallel transport in section 5.9 (see Figure 8). Since all the great circles on a two-sphere are geodesics, in particular the segments N-C, N-E, and E-C in the figure, we know that in order to parallel transport a vector along such a line we just need to

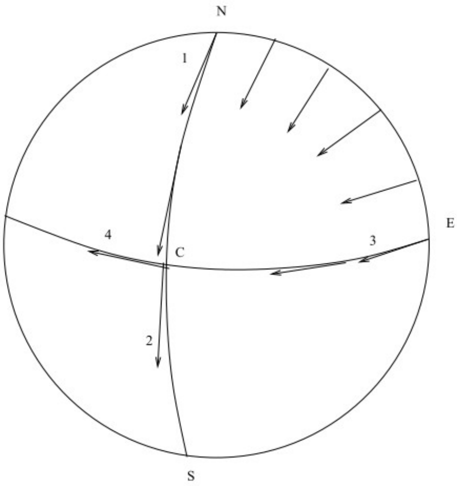

Figure 8: Figure illustrating the path dependence of parallel transport on a curved space: Vector 1 at N can be parallel transported along the geodesic N-S to C, giving rise to Vector 2. Alternatively, it can first be transported along the geodesic N-E (Vector 3) and then along E-C to give the Vector 4. Clearly these two are different. The angle between them reflects the curvature of the two-sphere.

make sure that its length and the angle between the vector and the geodesic line are constant. Thus imagine a vector 1 at the north pole N, pointing downwards along the line N-C-S. First parallel transport this along N-C to the point C. There we will obtain the vector 2, pointing downwards along C-S. Alternatively imagine parallel transporting the vector 1 first to the point E. Since the vector has to remain at a constant (right) angle to the line N-E, at the point E parallel transport will produce the vector 3 pointing westwards along E-C. Now parallel transporting this vector along E-C to C will produce the vector 4 at C. This vector clearly differs from the vector 2 that was obtained by parallel transporting along N-C instead of N-E-C.

To illustrate the claim about closed loops above, imagine parallel transporting vector 1 along the closed loop N-E-C-N from N to N. In order to complete this loop, we still have to parallel transport vector 4 back up to N. Clearly this will give a vector, not indicated in the figure, different from (and pointing roughly at a right angle to) the vector 1 we started off with.

The precise statement regarding the relation between the path dependence of parallel transport and the presence of curvature is the following. If one parallel transports a covector V_μ (I use a covector instead of a vector only to save myself a few minus signs here and there) along a closed infinitesimal loop $x^\mu(\tau)$ with, say, $x(\tau_0) = x(\tau_1) = x_0$, then one has

$$V_\mu(\tau_1) - V_\mu(\tau_0) = \tfrac{1}{2}(\oint x^\rho dx^\nu) R^\sigma{}_{\mu\rho\nu}(x_0) V_\sigma(\tau_0) \quad . \tag{11.1}$$

Thus an arbitrary vector V^μ will not change under parallel transport around an arbitrary small loop at x_0 only if the curvature tensor at x_0 is zero. This can of course be extended to finite loops, but the important point is that in order to detect curvature at a given point one only requires parallel transport along infinitesimal loops.

Before turning to a proof of this result, I just want to note that intuitively it can be understood directly from the definition of the curvature tensor (8.2). Imagine that the infinitesimal loop is actually a tiny parallelogram made up of the coordinate lines x^1 and x^2. Parallel transport along x^1 is governed by the equation $\nabla_1 V^\mu = 0$, that along x^2 by $\nabla_2 V^\mu = 0$. The fact that parallel transporting first along x^1 and then along x^2 can be different from doing it the other way around is precisely the statement that ∇_1 and ∇_2 do not commute, i.e. that some of the components $R_{\mu\nu 12}$ of the curvature tensor are non-zero.

To establish (11.1) we first reformulate the condition of parallel transport,

$$D_\tau V_\mu = 0 \quad \Leftrightarrow \quad \frac{d}{d\tau} V_\mu = \Gamma^\lambda_{\mu\nu} \dot{x}^\nu V_\lambda \tag{11.2}$$

with the initial condition at $\tau = \tau_0$ as the integral equation

$$V_\mu(\tau) = V_\mu(\tau_0) + \int_{\tau_0}^{\tau} d\tau' \, \Gamma^\lambda_{\mu\nu}(x(\tau')) \dot{x}^\nu(\tau') V_\lambda(\tau') \quad . \tag{11.3}$$

As usual, such an equation can be 'solved' by iteration (leading to a time-ordered exponential). Keeping only the first two non-trivial terms in the iteration, one has

$$\begin{aligned} V_\mu(\tau) \; = \; & V_\mu(\tau_0) + \int_{\tau_0}^{\tau} d\tau' \, \Gamma^\lambda_{\mu\nu}(x(\tau')) \dot{x}^\nu(\tau') V_\lambda(\tau_0) \\ & + \int_{\tau_0}^{\tau} d\tau' \int_{\tau_0}^{\tau'} d\tau'' \, \Gamma^\lambda_{\mu\nu}(x(\tau')) \dot{x}^\nu(\tau') \Gamma^\sigma_{\lambda\rho}(x(\tau'')) \dot{x}^\rho(\tau'') V_\sigma(\tau_0) \\ & + \ldots \end{aligned} \tag{11.4}$$

For sufficiently small (infinitesimal) loops, we can expand the Christoffel symbols as

$$\Gamma^\lambda_{\mu\nu}(x(\tau)) = \Gamma^\lambda_{\mu\nu}(x_0) + (x(\tau) - x_0)^\rho (\partial_\rho \Gamma^\lambda_{\mu\nu})(x_0) + \ldots \tag{11.5}$$

The linear term in the expansion of $V_\mu(\tau)$ arises from the zero'th order contribution $\Gamma^\lambda_{\mu\nu}(x_0)$ in the first order (single integral) term in (11.4),

$$[V_\mu(\tau_1) - V_\mu(\tau_0)]^{(1)} = \Gamma^\lambda_{\mu\nu}(x_0) V_\lambda(\tau_0) (\int_{\tau_0}^{\tau_1} d\tau' \, \dot{x}^\nu(\tau')) \quad . \tag{11.6}$$

Now the important observation is that, for a closed loop, the integral in brackets is zero,

$$\int_{\tau_0}^{\tau_1} d\tau' \dot{x}^\nu(\tau') = x^\nu(\tau_1) - x^\nu(\tau_0) = 0 \quad . \tag{11.7}$$

Thus the change in $V_\mu(\tau)$, when transported along a small loop, is at least of second order. Such second order terms arise in two different ways, from the first order term in the expansion of $\Gamma^\lambda_{\mu\nu}(x)$ in the first order term in (11.4), and from the zero'th order terms $\Gamma^\lambda_{\mu\nu}(x_0)$ in the quadratic (double integral) term in (11.4),

$$\begin{aligned}[V_\mu(\tau_1) - V_\mu(\tau_0)]^{(2)} &= (\partial_\rho \Gamma^\lambda_{\mu\nu})(x_0) V_\lambda(\tau_0) (\int_{\tau_0}^{\tau_1} d\tau' \; (x(\tau') - x_0)^\rho \dot{x}^\nu(\tau')) \\ &+ (\Gamma^\lambda_{\mu\nu}\Gamma^\sigma_{\lambda\rho})(x_0) V_\sigma(\tau_0) \int_{\tau_0}^{\tau_1} d\tau' \int_{\tau_0}^{\tau'} d\tau'' \; \dot{x}^\nu(\tau')\dot{x}^\rho(\tau'') \quad (11.8)\end{aligned}$$

The τ''-integral can be performed explicitly,

$$\int_{\tau_0}^{\tau_1} d\tau' \int_{\tau_0}^{\tau'} d\tau'' \; \dot{x}^\nu(\tau')\dot{x}^\rho(\tau'') = \int_{\tau_0}^{\tau_1} d\tau' \; \dot{x}^\nu(\tau')(x(\tau') - x_0)^\rho = \int_{\tau_0}^{\tau_1} d\tau' \; \dot{x}^\nu(\tau') x^\rho(\tau') \tag{11.9}$$

and therefore we find

$$V_\mu(\tau_1) - V_\mu(\tau_0) \approx (\partial_\rho \Gamma^\sigma_{\mu\nu} + \Gamma^\lambda_{\mu\nu}\Gamma^\sigma_{\lambda\rho})(x_0) V_\sigma(\tau_0) (\int_{\tau_0}^{\tau_1} d\tau' \; \dot{x}^\nu(\tau') x^\rho(\tau')) \tag{11.10}$$

The final observation we need is that the remaining integral is anti-symmetric in the indices ν, ρ, which follows immediately from

$$\int_{\tau_0}^{\tau_1} d\tau' \; (\dot{x}^\nu(\tau') x^\rho(\tau') + x^\nu(\tau')\dot{x}^\rho(\tau') = \int_{\tau_0}^{\tau_1} d\tau' \; \frac{d}{d\tau'}(x^\nu(\tau') x^\rho(\tau')) = 0 \;\; . \tag{11.11}$$

It now follows from (11.10) and the definition of the Riemann tensor that

$$V_\mu(\tau_1) - V_\mu(\tau_0) = \tfrac{1}{2}(\oint x^\rho dx^\nu) R^\sigma_{\;\mu\rho\nu}(x_0) V_\sigma(\tau_0) \;\; . \tag{11.12}$$

Simply by raising and lowering of the indices, and using the symmetry properties of the Riemann tensor, we can deduce that the corresponding equation for the parallel tansport of vectors is

$$V^\mu(\tau_1) - V^\mu(\tau_0) = -\tfrac{1}{2}(\oint x^\rho dx^\nu) R^\mu_{\;\sigma\rho\nu}(x_0) V^\sigma(\tau_0) \;\; . \tag{11.13}$$

As an example, recall that in section 8.6 we already determined explicitly the parallel transport of vectors on the 2-sphere along the circles with fixed $\theta = \theta_0$. Choosing θ_0 infinitesimal corresponds to an infinitesimal loop around the north pole. Expanding the result (5.116) for small θ_0, in particular using

$$\sin(2\pi\cos\theta_0) \approx \sin(2\pi(1 - \theta_0^2/2)) \approx -\tfrac{1}{2}(2\pi)\theta_0^2 \;\; , \tag{11.14}$$

one finds complete agreement between (11.13) and the components of the Riemann tensor of the 2-sphere, determined in (8.59),

$$r^\theta_{\;\phi\theta\phi} = \sin^2\theta \quad , \quad r^\phi_{\;\theta\phi\theta} = 1 \;\; , \tag{11.15}$$

evaluated for $\theta_0 \to 0$. In verifying this, some care should be taken with the fact that $\theta = 0$ is a coordinate singularity so that one should never strictly set $\theta_0 = 0$. Alternatively, and to be on the safe side, one can rewrite (11.13) as an equation for orthonormal frame components and use the result (5.121) for the parallel transport of the frame components (which is not sensitive to coordinate singularities).

11.2 VANISHING RIEMANN TENSOR AND EXISTENCE OF FLAT COORDINATES

We are now finally in a position to prove the converse to the statement that the Minkowski metric has vanishing Riemann tensor. Namely, we will see that when the Riemann tensor of a metric vanishes, locally there are coordinates in which the metric is the standard Minkowski metric. Since the opposite of curved is flat, this then allows one to unambiguously refer to the Minkowski metric as *the* flat metric (locally at least), and to Minkowski space as *flat space(-time)*.

So let us assume that we are given a metric with vanishing Riemann tensor. Then, by the above, parallel transport is path independent and we can, in particular, extend a vector $V^\mu(x_0)$ to a vector field everywhere in space-time: to define $V^\mu(x_1)$ we choose any path from x_0 to x_1 and use parallel transport along that path. In particular, the vector field V^μ, defined in this way, will be covariantly constant or *parallel*, $\nabla_\mu V^\nu = 0$. We can also do this for four linearly independent vectors V_a^μ at x_0 and obtain four covariantly constant (parallel) vector fields which are linearly independent at every point.

An alternative way of saying or seeing this is the following: The integrability condition for the equation $\nabla_\mu V^\lambda = 0$ is

$$\nabla_\mu V^\lambda = 0 \;\;\Rightarrow\;\; [\nabla_\mu, \nabla_\nu] V^\lambda = R^\lambda_{\;\sigma\mu\nu} V^\sigma = 0 \;\; . \tag{11.16}$$

This means that the (4×4) matrices $M(\mu, \nu)$ with coefficients $M(\mu,\nu)^\lambda_{\;\sigma} = R^\lambda_{\;\sigma\mu\nu}$ have a zero eigenvalue. If this integrability condition is satisfied, a solution to $\nabla_\mu V^\lambda$ can be found. If one wants four linearly independent parallel vector fields, then the matrices $M(\mu, \nu)$ must have four zero eigenvalues, i.e. they are zero and therefore $R^\lambda_{\;\sigma\mu\nu} = 0$. If this condition is satisfied, all the integrability conditions are satisfied and there will be four linearly independent covariantly constant vector fields - the same conclusion as above.

We will now use this result in the proof, but for covectors instead of vectors. Clearly this makes no difference: if V^μ is a parallel vector field, then $g_{\mu\nu}V^\nu$ is a parallel covector field.

Fix some point x_0. At x_0, there will be an invertible matrix e^a_μ such that

$$g^{\mu\nu}(x_0) e^a_\mu e^b_\nu = \eta^{ab} \;\; . \tag{11.17}$$

Now we solve the equations

$$\nabla_\nu E^a_\mu = 0 \Leftrightarrow \partial_\nu E^a_\mu = \Gamma^\lambda{}_{\mu\nu} E^a_\lambda \tag{11.18}$$

with the initial condition $E^a_\mu(x_0) = e^a_\mu$. This gives rise to four linearly independent parallel covectors E^a_μ.

Now it follows from (11.18) that

$$\partial_\mu E^a_\nu = \partial_\nu E^a_\mu \quad . \tag{11.19}$$

Therefore locally there are four scalars ξ^a such that

$$E^a_\mu = \frac{\partial \xi^a}{\partial x^\mu} \quad . \tag{11.20}$$

These are already the flat coordinates we have been looking for. To see this, consider the expression $g^{\mu\nu} E^a_\mu E^b_\nu$. This is clearly constant because the metric and the E^a_μ are covariantly constant,

$$\partial_\lambda (g^{\mu\nu} E^a_\mu E^b_\nu) = \nabla_\lambda (g^{\mu\nu} E^a_\mu E^b_\nu) = 0 \quad . \tag{11.21}$$

At x_0, this is just the flat metric and thus

$$(g^{\mu\nu} E^a_\mu E^b_\nu)(x) = (g^{\mu\nu} E^a_\mu E^b_\nu)(x_0) = \eta^{ab} \quad . \tag{11.22}$$

Summing this up, we have seen that, starting from the assumption that the Riemann curvature tensor of a metric $g_{\mu\nu}$ is zero, we have proven the existence of coordinates ξ^a in which the metric takes the Minkowski form,

$$g_{\mu\nu} = \frac{\partial \xi^a}{\partial x^\mu} \frac{\partial \xi^b}{\partial x^\nu} \eta_{ab} \quad . \tag{11.23}$$

The argument given above is local in the sense that the existence of these coordinates ξ^a is only guaranteed locally, i.e. in the neighbourhood of some point. Whether or not these coordinates can be used to cover the space-time globally depends on gobal (topological) properties of the space-time which are not captured by the intrinsic local and locally determined Riemann tensor.

For example, imagine starting with Minkowski space $\mathbb{R}^{1,3}$ with inertial coordinates ξ^a, and then making a periodic identification of ξ^1, say,

$$\xi^1 \sim \xi^1 + 2\pi L_1 \quad \Rightarrow \quad \mathbb{R}^{1,3} \to \mathbb{R}^{1,2} \times S^1 \quad . \tag{11.24}$$

Since the Minkowksi metric is translation-invariant, it gives rise to a well-defined metric on the periodically identified space-time, and the metric of this space-time still has zero curvature tensor. Nevertheless, in this case

- in the new space-time the coordinate ξ^1, which is now an angular variable, is not globally well defined,
- and the space-time looks like Minkowski space only locally, not globally.

11.3 CURVATURE OF SURFACES: EULER, GAUSS(-BONNET) AND LIOUVILLE

We can generalise the example of the curvature of the 2-sphere, discussed in section 8.6, somewhat, in this way connecting our considerations with the classical realm of the differential geometry of surfaces, in particular with the *Gauss Curvature*, the *Euler characteristic*, the *Gauss-Bonnet* theorem and the *Liouville Equation*.

For any 2-dimensional metric g_{ab} it is a simple exercise to derive the relation between the one independent component, say R_{1212}, of the Riemann tensor, and the scalar curvature. First of all, the Ricci tensor is

$$R_{ab} = R^c{}_{acb} = R^1{}_{a1b} + R^2{}_{a2b} \tag{11.25}$$

so that the scalar curvature is

$$R(g_{ab}) = g^{ab}R_{ab} = g^{11}R^2{}_{121} + g^{12}R^1{}_{112} + g^{21}R^2{}_{221} + g^{22}R^1{}_{212} \ . \tag{11.26}$$

Using the fact that in 2 dimensions the components of the inverse metric are explicitly given by

$$\left(g^{ab}\right) = \frac{1}{g_{11}g_{22} - g_{12}g_{21}} \begin{pmatrix} g_{22} & -g_{12} \\ -g_{21} & g_{11} \end{pmatrix} \tag{11.27}$$

and the (anti-)symmetry properties (1) and (2) of the Riemann tensor, one finds

$$R(g_{ab}) = \frac{2}{g_{11}g_{22} - g_{12}g_{21}} R_{1212} \ . \tag{11.28}$$

This is precisely the relation (8.44) between the Riemann tensor and Ricci scalar. The factor of 2 in this equation is a consequence of our (and the conventional) definition of the Riemann curvature tensor, and is responsible for the fact that the scalar curvature of the unit 2-sphere is $R = +2$. We can also write this result as

$$R_{abcd} = \tfrac{1}{2}(g_{ac}g_{bd} - g_{ad}g_{bc})R \quad \Leftrightarrow \quad R^a{}_{bcd} = \tfrac{1}{2}(\delta^a{}_c g_{bd} - \delta^a{}_d g_{bc})R \ . \tag{11.29}$$

In two dimensions, it is often convenient and natural to absorb this ubiquitous factor of 2 into the definition of the (scalar) curvature, and what one then gets is the classical Gauss Curvature

$$K := \frac{1}{2}R(g_{ab}) \tag{11.30}$$

of a two-dimensional surface.

It follows from (11.29) that the Ricci tensor is related to the Ricci scalar by

$$R_{ab} = (R/2)g_{ab} = K g_{ab} \ . \tag{11.31}$$

This generalises the result for the standard metric on the 2-sphere found by explicit calculation in section 8.6. It shows that in complete generality the Ricci tensor of a two-dimensional space or space-time, thought of as the linear map

$$R^a_{\ b} = K\delta^a_{\ b} : \quad v^a \mapsto R^a_{\ b}\, v^b \ , \tag{11.32}$$

has only one (double) eigenvalue, namely the Gauss curvature K. It can also be interpreted as saying that in 2 dimensions the Einstein tensor (8.97) is identically zero,

$$G_{ab} = R_{ab} - \tfrac{1}{2} g_{ab} R = 0 \ . \tag{11.33}$$

We will now briefly look at two important and interesting consequences of the above formulae, one related to the Euler characteristic of a surface and its integral representation (the Gauss-Bonnet theorem), and the other to the Liouville equation describing metrics with constant Gauss curvauter $K = k = \pm 1$.

1. Euler Characteristic and the Gauss-Bonnet Theorem

   Let us consider a compact closed surface S, i.e. topologically something like a sphere, or a torus (a sphere with one handle), or a sphere with several handles. A surface S_h with h handles is called a surface of *genus h*.

   Given a metric on S_h, we can associate to S_h its area with respect to the metric,

   $$A(S_h) := \int_{S_h} \sqrt{g} d^2x \ . \tag{11.34}$$

   Clearly, this areas depends on a choice of metric, and under a variation δ_g of the metric it transforms as

   $$\delta_g A(S_h) = \int_{S_h} \delta_g \sqrt{g} d^2x = \tfrac{1}{2} \int_{S_h} \sqrt{g} d^2x \ g^{ab} \delta g_{ab} \ . \tag{11.35}$$

   Given a metric on S_g, we can also naturally associate to it the real number

   $$\chi(S_h) = \frac{1}{2\pi} \int \sqrt{g} d^2x \ K = \frac{1}{4\pi} \int \sqrt{g} d^2x \ R \ . \tag{11.36}$$

   Remarkably, this number turns out to be independent of the metric, in particular *$\chi(S_h)$ is invariant under variations of the metric g_{ab}*,

   $$\delta_g \chi(S_h) = 0 \ . \tag{11.37}$$

   Here are two rather explicit ways of establishing this remarkable result:

   (a) The variation of the integrand is

   $$\delta_g(\sqrt{g} R) = \delta_g(\sqrt{g} g^{ab} R_{ab}) \ . \tag{11.38}$$

   Since $R = g^{ab} R_{ab}$, it is convenient to express variations of the metric in terms of variations of the inverse metric, with (5.85)

   $$\delta \sqrt{g} = \tfrac{1}{2} g^{ab} \delta g_{ab} = -\tfrac{1}{2} g_{ab} \delta g^{ab} \ . \tag{11.39}$$

Then one finds

$$\begin{aligned}\delta_g(\sqrt{g}R) &= (\delta_g\sqrt{g})g^{ab}R_{ab} + \sqrt{g}(\delta g^{ab})R_{ab}) + \sqrt{g}g^{ab}\delta_g R_{ab} \\ &= \sqrt{g}(-\tfrac{1}{2}Rg_{ab} + R_{ab})\delta g^{ab} + \sqrt{g}g^{ab}\delta_g R_{ab} \\ &= \sqrt{g}G_{ab}\delta g^{ab} + \sqrt{g}g^{ab}\delta_g R_{ab} \quad .\end{aligned} \tag{11.40}$$

Now, as shown above, in 2 dimensions the Einstein tensor is identically zero, $G_{ab} = 0$. Moreover, in section 20.2 we will show that in any dimension the 2nd term is a total derivative (20.18),

$$g^{ab}\delta_g R_{ab} = \nabla_a B^a \tag{11.41}$$

for some well-defined B^a built from the covariant derivatives of the variations of the metric, as in (20.19). Taken together, these two facts imply that for a closed surface S_h (without boundary) one has

$$\delta_g\chi(S_h) = \frac{1}{4\pi}\int \sqrt{g}d^2x(G_{ab}\delta g^{ab} + \nabla_a B^a) = 0 \ \ , \tag{11.42}$$

as was to be shown.

(b) Alternatively, somewhat less covariantly but very explicitly, one can show that the integrand $\sqrt{g}K$ or $\sqrt{g}R$ can itself locally be written as a total derivative. Indeed, using (11.29) to write

$$R^1{}_{212} = \tfrac{1}{2}g_{11}R \quad \Leftrightarrow \quad K = \frac{1}{g_{11}}R^2{}_{121} \tag{11.43}$$

and simply writing out explicitly this Riemann curvature tensor component in terms of the Christoffel symbols,

$$R^2{}_{121} = \partial_2\Gamma^2{}_{11} - \partial_1\Gamma^2{}_{12} + \Gamma^2{}_{2a}\Gamma^a{}_{11} - \Gamma^2{}_{1a}\Gamma^a{}_{12} \ \ , \tag{11.44}$$

one finds that $\sqrt{g}K$ can be written as

$$\sqrt{g}K = \in^{ab}\partial_a\beta_b \tag{11.45}$$

with $\in^{12} = -\in^{21} = 1$ the Levi-Civita symbol, and

$$\beta_b = -\frac{\sqrt{g}}{g_{11}}\,\Gamma^2{}_{1b} \ \ . \tag{11.46}$$

The fact that the integrand is locally a total derivative does not mean that the integral is zero (because of the non-tensorial nature of β_a, which will typically exhibit coordinate singularities). It does mean however, that the integral of the metric variation of this expression is zero (because that is tensorial and well-defined on S_h),

$$\sqrt{g}K = \varepsilon^{ab}\partial_a\beta_b \quad \Rightarrow \quad \delta_g\chi(S_h) = \frac{1}{2\pi}\int d^2x\delta_g(\sqrt{g}K) = 0 \ \ . \tag{11.47}$$

Either way we have seen that the real number $\chi(S_h)$ is independent of the metric one uses to calculate it. For example, for $h = 0$ and for the standard metric on the sphere S^2 one finds

$$\chi(S_{h=0}) = \chi(S^2) = \frac{1}{4\pi}\int \sqrt{g}R = \frac{1}{2\pi}\int \sqrt{g} = 2 \ , \tag{11.48}$$

and this will therefore be the result for any metric on S^2. Likewise, for $h = 1$, i.e. a torus, by choosing the flat metric on T^2 (see e.g. the discussion and construction in section 18.1), one finds

$$\chi(S_{h=1}) = \chi(T^2) = 0 \ , \tag{11.49}$$

and this will therefore be the result for any metric on T^2 (and it is instructive to check this explicitly for the non-trivial, non-flat metric on T^2 induced by its embedding into $\mathbb{R}^3$ constructed in section 18.1). I am not aware of an equally elementary calculation to determine $\chi(S_h)$ for $h > 1$ in this way but fact of the matter is that

$$\chi(S_h) = 2 - 2h \tag{11.50}$$

is the *Euler characteristic* of S_h, which can also be defined purely combinatorially as the number

$$\chi(S) = n_F - n_E + n_V \tag{11.51}$$

of faces minus vertices plus edges of any cubist rendition of a surface S (and $\chi(S)$ is independent of such a cubist realisation or *triangulation*). The remarkable fact that this *topological invariant* of a surface S can be calculated in terms of differential geometric quantities, namely as the integral of the curvature scalar, is known as the *Gauss-Bonnet theorem.*

2. Constant Curvature and the Liouville Equation

   When we specialise the above to the class of *conformally flat* metrics with line element

$$ds^2 = \mathrm{e}^{2h(x,y)}(dx^2 + dy^2) \quad \Leftrightarrow \quad g_{ab} = \exp 2h(x,y)\delta_{ab} \tag{11.52}$$

   the calculation of the Riemann tensor is particularly simple and one finds the (easy to memorise) results

$$R^x{}_{yxy} = -\Delta h \tag{11.53}$$

   and

$$K = -\mathrm{e}^{-2h}\Delta h \tag{11.54}$$

   where $\Delta = \partial_x^2 + \partial_y^2$ is the 2-dimensional Laplacian with respect to the flat Euclidean metric $dx^2 + dy^2$. Thus a surface with *constant curvature* $K = k$ is given by a solution to the non-linear differential equation

$$\Delta h + k\mathrm{e}^{2h} = 0 \ . \tag{11.55}$$

This is the (in-)famous *Liouville equation*, which plays a fundamental role in many branches of mathematics (and mathematical physics).

In terms of the intrinsic Laplacian Δ_g associated to the metric g_{ab}, the Gaussian cuvature and the Liouville equation can also simply be written as

$$K = -\Delta_g h \qquad , \qquad \Delta_g h + k = 0 \ , \tag{11.56}$$

since, due to the peculiarities of 2 dimensions, $\sqrt{g}g^{ab}$ in independent of h, i.e. is conformally invariant (as we already observed in a different context in section 7.7, cf. (7.120)),

$$\begin{aligned} &\sqrt{g}g^{ab} = \mathrm{e}^{2h}\mathrm{e}^{-2h}\delta^{ab} = \delta^{ab} \\ \Rightarrow \quad &\Delta_g = \frac{1}{\sqrt{g}}\partial_a(\sqrt{g}g^{ab}\partial_b) = \frac{1}{\sqrt{g}}\partial_a(\delta^{ab}\partial_b) = \mathrm{e}^{-2h}\Delta \ . \end{aligned} \tag{11.57}$$

I will not attempt to say anything about the general (local) solution of this equation (which roughly speaking depends on an arbitrary meromorphic function of the complex coordinate $z = x + iy$), but close this section with some special (and particularly prominent) solutions of this equation.

(a) It is easy to see that

$$\mathrm{e}^{2h(x,y)} = y^{-2} \quad \Leftrightarrow \quad h(x,y) = -\ln y \tag{11.58}$$

solves the Liouville equation with $k = -1$. The corresponding space of constant negative curvature is the *Poincaré upper-half plane* model of the hyperbolic geometry,

$$ds^2 = \frac{dx^2 + dy^2}{y^2} \qquad (\ (x,y) \in \mathbb{R}^2,\ y > 0\ ) \ . \tag{11.59}$$

By the coordinate transformation $y = e^z$ this is mapped to the equivalent metric

$$ds^2 = dz^2 + \mathrm{e}^{-2z}dx^2 \tag{11.60}$$

on the entire (x, z)-plane.

(b) Another solution (for any k) is the rotationally invariant function

$$\mathrm{e}^{2h(x,y)} = 4(1 + k(x^2 + y^2))^{-2} \quad \Leftrightarrow \quad h = -\ln(1 + k(x^2 + y^2)) + \text{const.} \tag{11.61}$$

i. For $k = 0$ one finds the flat (zero curvature) Euclidean metric on $\mathbb{R}^2$.

ii. For $k = +1$, one obtains the metric

$$ds^2 = 4\frac{dx^2 + dy^2}{(1 + x^2 + y^2)^2} \ . \tag{11.62}$$

This is the constant positive curvature metric on the *Riemann sphere* one gets by stereographic projection of the standard metric on the two-sphere S^2 to the (x,y)-plane.

In terms of polar coordinates (r,ϕ) on the Euclidean plane, this metric takes the form

$$ds^2 = 4\frac{dr^2 + r^2 d\phi^2}{(1+r^2)^2} \quad , \tag{11.63}$$

and the further change of variables $r = \tan\theta/2$ shows that this is indeed the standard line element $d\Omega^2$ on the 2-sphere,

$$r = \tan\theta/2 \quad \Rightarrow \quad ds^2 = d\theta^2 + \sin^2\theta d\phi^2 = d\Omega^2 \quad . \tag{11.64}$$

Read backwards, this can also be read as the statement that via the above change of variables the Euclidean metric on $\mathbb{R}^2$ can be written as

$$dr^2 + r^2 d\phi^2 = \frac{(1+r(\theta)^2)^2}{4}(d\theta^2 + \sin^2\theta d\phi^2) = \frac{1}{4\cos^4\theta/2}d\Omega^2 \quad . \tag{11.65}$$

In this process the points "at infinity" in $\mathbb{R}^2$ (where $r \to \infty$) are mapped to the south pole $\theta = \pi$ of the sphere (where the conformal factor in front of the line element of the sphere diverges accordingly). This exhibits S^2 as the *conformal compactification* of $\mathbb{R}^2$.

iii. For $k = -1$, one finds

$$ds^2 = 4\frac{dx^2 + dy^2}{(1-(x^2+y^2))^2} \qquad (\ \{x,y\} \in \mathbb{R}^2, x^2+y^2 < 1\ ) \quad . \tag{11.66}$$

This is the *Poincaré disc* model of the hyperbolic geometry, defined in the interior of the unit disc in $\mathbb{R}^2$. In terms of polar coordinates, it can also be written as

$$ds^2 = 4\frac{dr^2 + r^2 d\phi^2}{(1-r^2)^2} \qquad (\ 0 \leq r < 1\ ) \tag{11.67}$$

The two metrics (11.59) and (11.66) are isometric, i.e. related by a (albeit not completely evident) coordinate transformation.

(c) As our final example, one other solution (given here only for $k = -1$) is

$$\mathrm{e}^{2h(x,y)} = \mathrm{e}^{2h(x)} = 4\frac{\mathrm{e}^{2x}}{(1-\mathrm{e}^{2x})^2} \quad . \tag{11.68}$$

This leads to the form (2.31) of the unit metric on H^2,

$$ds^2 = d\sigma^2 + \sinh^2\sigma\ d\phi^2 \quad . \tag{11.69}$$

Indeed, we can write this in conformally flat form as

$$ds^2 = \sinh^2\sigma(x)(dx^2 + d\phi^2) \tag{11.70}$$

where

$$x = \log\tanh(\sigma/2) \quad \Rightarrow \quad dx = d\sigma/\sinh\sigma \ . \tag{11.71}$$

Performing the exponential / hyperbolic gymnastics required to write the conformal factor $\sinh^2\sigma$ as a function of x, one finds

$$\sinh^2\sigma(x) = 4\frac{\mathrm{e}^{2x}}{(1-\mathrm{e}^{2x})^2} \ , \tag{11.72}$$

giving rise to the solution given in (11.68).

It is worth remarking that the Poincaré upper-half plane model of a space with constant negative curvature readily generalises to arbitrary dimensions and signature. Thus

$$ds^2 = \frac{d\vec{x}^2 + dy^2}{y^2} \quad , \quad d\vec{x}^2 = \delta_{ab}dx^a dx^b \quad \text{or} \quad d\vec{x}^2 = \eta_{ab}dx^a dx^b \tag{11.73}$$

is the metric of a $D = (d+1)$-dimensional space(-time) with constant negative curvature.

The Lorentzian metric will reappear later as a solution to the Einstein equations with a negative cosmological constant, and is in this context known as the *anti-de Sitter metric* (in Poincaré coordinates, which cover only a part of the complete space-time), and we will discuss this solution in some detail in section 39.

11.4 THE WEYL TENSOR AND ITS USES

In section 8 from the Riemann tensor we have extracted its traces, the Ricci tensor and the Ricci scalar, as well as a particular linear combination of them, the Einstein tensor. We can therefore also explicitly decompose the Riemann tensor into these trace parts and the remaining traceless part.

We noted in section 8.5 that for $D = 2$ and $D = 3$ the Riemann tensor would be "pure trace", i.e. could be written entirely in terms of the Ricci tensor and Ricci scalar. For $D = 2$ we have already established this explicitly by proving the relation (11.29),

$$D = 2: \quad R_{\alpha\beta\gamma\delta} = \tfrac{1}{2}(g_{\alpha\gamma}g_{\beta\delta} - g_{\alpha\delta}g_{\beta\gamma})R \tag{11.74}$$

in section 11.3.

We now look at this issue for $D \geq 3$. Simply by linear algebra one finds, for $D \geq 3$, the decomposition

$$\begin{aligned} R_{\mu\nu\rho\sigma} &= C_{\mu\nu\rho\sigma} \\ &+ \frac{1}{D-2}(g_{\mu\rho}R_{\nu\sigma} + R_{\mu\rho}g_{\nu\sigma} - g_{\nu\rho}R_{\mu\sigma} - R_{\nu\rho}g_{\mu\sigma}) \\ &- \frac{1}{(D-1)(D-2)}R(g_{\mu\rho}g_{\nu\sigma} - g_{\nu\rho}g_{\mu\sigma}) \ . \end{aligned} \tag{11.75}$$

This definition is such that $C_{\mu\nu\rho\sigma}$ has all the symmetries of the Riemann tensor (this is manifest) and such that all of its traces are zero, i.e.

$$C^{\mu}_{\ \nu\mu\sigma} = 0 \ , \tag{11.76}$$

as is easily verified. This traceless part $C_{\mu\nu\rho\sigma}$ of the Riemann tensor is called the *Weyl tensor*.

Occasionally it is more convenient and transparent to decompose the Riemann tensor not into the Weyl tensor, the Ricci tensor and the Ricci scalar, but to perform an orthogonal decomposition (with respect to the metric) into the Weyl tensor, the traceless part $S_{\mu\nu}$ of the Ricci tensor,

$$S_{\mu\nu} = R_{\mu\nu} - \frac{1}{D} g_{\mu\nu} R \ , \tag{11.77}$$

and the trace R. Then the decomposition becomes

$$\begin{aligned} R_{\mu\nu\rho\sigma} &= C_{\mu\nu\rho\sigma} \\ &\quad + \frac{1}{D-2}(g_{\mu\rho}S_{\nu\sigma} + S_{\mu\rho}g_{\nu\sigma} - g_{\nu\rho}S_{\mu\sigma} - S_{\nu\rho}g_{\mu\sigma}) \\ &\quad + \frac{1}{D(D-1)} R(g_{\mu\rho}g_{\nu\sigma} - g_{\nu\rho}g_{\mu\sigma}) \ . \end{aligned} \tag{11.78}$$

One other common and convenient decomposition is in terms of a tensor $P_{\mu\nu}$ such that (11.75) takes the form

$$R_{\mu\nu\rho\sigma} = C_{\mu\nu\rho\sigma} + (g_{\mu\rho}P_{\nu\sigma} + P_{\mu\rho}g_{\nu\sigma} - g_{\nu\rho}P_{\mu\sigma} - P_{\nu\rho}g_{\mu\sigma}) \ . \tag{11.79}$$

Comparison with (11.75) shows that this is accomplished by the choice

$$P_{\mu\nu} = \frac{1}{D-2}\left(R_{\mu\nu} - \frac{1}{2(D-1)} g_{\mu\nu} R\right) \ . \tag{11.80}$$

This tensor $P_{\mu\nu}$ is known as the *Schouten Tensor*.

Regardless of how we write the trace part of the Riemann tensor, it turns out that for $D = 3$ the Weyl tensor vanishes identically,

$$D = 3: \quad C_{\alpha\beta\gamma\delta} \equiv 0 \tag{11.81}$$

(I will give an elementary proof of this momentarily). Therefore, for $D = 3$ one has the decomposition

$$\begin{aligned} D = 3: \quad R_{\alpha\beta\gamma\delta} &= (g_{\alpha\gamma}R_{\beta\delta} + R_{\alpha\gamma}g_{\beta\delta} - g_{\alpha\delta}R_{\beta\gamma} - R_{\alpha\delta}g_{\beta\gamma}) \\ &\quad + \tfrac{1}{2}(g_{\alpha\delta}g_{\beta\gamma} - g_{\alpha\gamma}g_{\beta\delta})R \end{aligned} \tag{11.82}$$

This is precisely the result claimed previously in (8.45).

To establish (11.81), in order to trivialise the algebra let us fix a point x_0 and choose coordinates there such that $g_{\alpha\beta}(x_0) = \delta_{\alpha\beta}$ (or $\eta_{\alpha\beta}$, depending on the signature of the metric, but let us assume that we are in the case of Euclidean signature - the same argument works in the Lorentzian case). Now the proof consists of the following elementary steps:

- Since we are in $D = 3$, at least two of the indices in $C_{\alpha\beta\gamma\delta}$ must be equal. Since the Weyl tensor has all the symmetries of the Riemann tensor, if more than two indices are equal, the Weyl tensor component is zero. Thus we only need to consider the components where 2 indices are equal and we can without loss of generality choose these to be $C_{1\beta 1\delta}$, say, with $\beta, \delta \neq 1$.

- Because the Weyl tensor is traceless, one has the relation

  $$C_{1\beta 1\delta} = -C_{2\beta 2\delta} - C_{3\beta 3\delta} \quad . \tag{11.83}$$

  This implies that $C_{1213} = C_{1312} = 0$, so a non-zero component requires $\beta = \delta$.

- For $\beta = \delta$, one derives from this

  $$C_{1212} = -C_{3232} \quad \text{and} \quad C_{1313} = -C_{2323} \tag{11.84}$$

  and likewise for $\alpha = \gamma = 2$ and $\alpha = \gamma = 3$.

- Thus all in all the Weyl tensor can have only 3 independent non-zero components, namely $C_{1212} = C_{2121}, C_{1313} = C_{3131}, C_{2323} = C_{3232}$, and they are all required to be pairwise negatives of each other. This is impossible for non-trivial $C_{\alpha\beta\gamma\delta}$,

  $$C_{1212} = -C_{3232} = +C_{3131} = -C_{2121} = -C_{1212} \quad \Rightarrow \quad C_{1212} = 0 \ , \tag{11.85}$$

  and implies that all of the components of the Weyl tensor are identically zero in $D = 3$.

Thus the Weyl tensor is only non-trivial for $D \geq 4$. Using the Bianchi identies discussed in section 8.8, in particular also (8.94),

$$\nabla^\mu R_{\mu\nu\rho\sigma} = \nabla_\rho R_{\nu\sigma} - \nabla_\sigma R_{\nu\rho} \tag{11.86}$$

from (11.75) one finds a simple expression for the divergence, namely

$$\nabla^\mu C_{\mu\nu\rho\sigma} = (D-3)\left(\nabla_\rho P_{\nu\sigma} - \nabla_\sigma P_{\nu\rho}\right) \quad . \tag{11.87}$$

The tensor appearing on the right-hand side also has its own name. It is called the *Cotton Tensor* $C_{\nu\rho\sigma}$,

$$C_{\nu\rho\sigma} = \nabla_\rho P_{\nu\sigma} - \nabla_\sigma P_{\nu\rho} \quad . \tag{11.88}$$

The content of (11.87) is evidently trivial in $D = 3$, but the Cotton tensor itself is not (and I will briefly come back to this below).

The Weyl tensor plays an important role in many aspects of gravitational physics:

1. For example, the Weyl tensor has traditionally been one of the central objects of interest in the invariant algebraic classification of gravitational fields and in

the characterisation of what are known as algebraically special solutions to the Einstein equations (the so-called *Petrov classification* and related procedures). Originally, this was (of course) developed for $D = 4$, and this case has a number of special features. It is based on the classification of the properties of the eigenvalues λ of the Weyl tensor (at a point x_0), thought of as a map on the space of anti-symmetric $(2,0)$-tensors (*bivectors*),

$$\tfrac{1}{2}C^{\alpha\beta}{}_{\gamma\delta}X^{\gamma\delta} = \lambda X^{\alpha\beta} \tag{11.89}$$

or

$$C^A_B X^B = \lambda X^A \tag{11.90}$$

with $C_{\alpha\beta\gamma\delta} \equiv C_{AB}$ thought of as a symmetric (6×6) matrix.[26]

An equivalent (as it turns out) classification arises from determining the number and multiplicity of linearly independent null vectors ℓ satisfying the condition

$$\ell_{[\gamma}C_{\alpha]\mu\nu[\beta}\ell_{\delta]}\ell^\mu\ell^\nu = 0 \ . \tag{11.91}$$

Such ℓ^α are called the *principal null directions* of the Weyl tensor. More recently, this classification scheme (based on the latter approach) has been (partially) extended to higher dimensions.[27]

2. As we will see in section 19.6, the Einstein equations imply that the Weyl tensor describes the gravitational field in vacuum. Specifically, when (or where) the energy-momentum tensor is zero, the Riemann curvature tensor is equal to the Weyl tensor,

$$T_{\alpha\beta}(x) = 0 \quad \Rightarrow \quad R_{\alpha\beta\gamma\delta}(x) = C_{\alpha\beta\gamma\delta}(x) \ . \tag{11.92}$$

The Weyl tensor thus encodes the information about things like gravitational waves and the asymptotic behaviour of a gravitational field and has been studied extensively from this point of view.

3. In the presence of matter, on the other hand, (11.87), in conjunction with the Einstein equations, becomes an evolution equation for these vacuum components of the gravitational field in terms of the sources - see equations (19.57) and (19.58).

The Weyl tensor also plays an important role in geometry, as it is *conformally invariant*, i.e. $C^\mu_{\nu\rho\sigma}$ is invariant under conformal (Weyl) rescalings of the metric,

$$g_{\mu\nu}(x) \to \mathrm{e}^{2f(x)}g_{\mu\nu}(x) \quad \Rightarrow \quad C^\mu_{\ \nu\rho\sigma} \to C^\mu_{\ \nu\rho\sigma} \ , \tag{11.93}$$

---

[26] See e.g. sections 7.16 and 11 of J. Plebanski, A. Krasinski, *An Introduction to General Relativity and Cosmology*, or section 4 of H. Stephani, D. Kramer, M. MacCallum, C. Hoenslaers, E. Herlt, *Exact Solutions to Einstein's Field Equations* (2nd Edition) for recent expositions of this subject.

[27] See A. Coley, R. Milson, V. Pravda, A. Pravdova, *Classification of the Weyl Tensor in Higher Dimensions*, `arXiv:arXiv:gr-qc/0401008`; A. Coley, *Classification of the Weyl Tensor in Higher Dimensions and Applications*, `arXiv:0710.1598 [gr-qc]`; M. Ortaggio, V. Pravda, A. Pravdova, *Algebraic classification of higher dimensional spacetimes based on null alignment*, `arXiv:1211.7289 [gr-qc]`.

equivalently

$$g_{\mu\nu}(x) \to \mathrm{e}^{2f(x)} g_{\mu\nu}(x) \quad \Rightarrow \quad C_{\mu\nu\rho\sigma} \to \mathrm{e}^{2f(x)} C_{\mu\nu\rho\sigma} \ . \tag{11.94}$$

In particular, the Weyl tensor is zero if the metric is *conformally flat*, i.e. related by a conformal transformation to the flat metric $\eta_{\mu\nu}$ (of any signature),

$$g_{\mu\nu}(x) = \mathrm{e}^{2f(x)} \eta_{\mu\nu}(x) \quad \Rightarrow \quad C^{\mu}_{\ \nu\rho\sigma} = 0 \ . \tag{11.95}$$

This can be established by brute force calculation and is not *per se* particularly enlightning.

Conversely for $D \geq 4$ vanishing of the Weyl tensor is also a sufficient condition for a metric to be (locally) conformal to the flat metric. This is a non-trivial result because at face value one seems to obtain a completely overdetermined system of equations for the single function f, of the form

$$P_{\mu\nu} = \{\text{1st and 2nd derivatives of } f\}_{\mu\nu} \ . \tag{11.96}$$

However, it turns out that the integrability conditions for this system of equations are equivalent to the vanishing of the Weyl tensor, and then a variant of the *Frobenius integrability theorem* (mentioned in a different context in section 15.5) can be used to establish the local existence of a solution f.

For $D = 3$, the situation is slightly (but not fundamentally) different. We see from (11.88) that for any $D \geq 4$ conformal flatness implies vanishing of the Cotton tensor. It turns out that for $D = 3$ the Cotton tensor takes over the role of the Weyl tensor (which, as proven above, is itself trivial for $D = 3$), i.e. one has the statement that for $D = 3$ a metric is (locally) conformally flat if and only if the Cotton tensor vanishes.

11.5 Generalisations: Torsion and Non-Metricity

In section 5.4 we had seen that the Levi-Civita connection (defined by the Christoffel symbols) is characterised by the fact that

1. the metric is covariantly constant, $\nabla_\mu g_{\nu\lambda} = 0$, and
2. the torsion is zero, i.e. the second covariant derivatives of a scalar commute.

It is of course possible to relax either of the conditions (1) or (2), or both of them and, in particular, connections with torsion (relaxation of condition 2) are popular in certain circles and/or arise naturally in certain generalised (gauge) theories of gravity and in string theory.

To discuss this a bit more systematically, we consider a general connection

$$\tilde{\Gamma}^{\mu}_{\ \nu\lambda} = \Gamma^{\mu}_{\ \nu\lambda} + C^{\mu}_{\ \nu\lambda} \tag{11.97}$$

with $\Gamma^{\mu}_{\ \nu\lambda}$ the canonical Levi-Civita connection, and $C^{\mu}_{\ \nu\lambda}$ a $(1,2)$-tensor. We will also use the corresponding $(0,3)$-tensor

$$C_{\mu\nu\lambda} = g_{\mu\rho} C^{\rho}_{\ \nu\lambda} \quad . \tag{11.98}$$

Associated with $\tilde{\Gamma}^{\mu}_{\ \nu\lambda}$ we have the covariant derivative $\tilde{\nabla}_{\mu}$. Since $\tilde{\Gamma}^{\mu}_{\ \nu\lambda}$ will in general not be symmetric in its lower indices, in this section we need to be particularly careful with (and choose a convention for) the ordering of the lower indices in the covariant derivative. We will choose the convention that the last index always refers to the direction along which one is differentiating, i.e.

$$\tilde{\nabla}_{\mu} V^{\nu} = \partial_{\mu} V^{\nu} + \tilde{\Gamma}^{\nu}_{\ \lambda\mu} V^{\lambda} \tag{11.99}$$

etc. The reason for this choice is that one should think of the collection of objects $\Gamma^{\nu}_{\ \lambda\mu}$ (and $\tilde{\Gamma}^{\nu}_{\ \lambda\mu}$) as the coefficients of a matrix-valued 1-form (cf. section 4.6) $\Gamma^{\nu}_{\ \lambda} = \Gamma^{\nu}_{\ \lambda\mu} dx^{\mu}$, the matrices acting by rotation on vectors (and more general tensors), as in (5.21).

We now define the *torsion tensor* $T^{\lambda}_{\mu\nu}$ by

$$[\tilde{\nabla}_{\mu}, \tilde{\nabla}_{\nu}]\phi = T^{\lambda}_{\mu\nu} \partial_{\lambda}\phi \quad , \quad T_{\lambda\mu\nu} = g_{\lambda\rho} T^{\rho}_{\ \mu\nu} \quad , \tag{11.100}$$

and the *non-metricity tensor* $Q_{\nu\lambda\mu}$ by

$$\tilde{\nabla}_{\mu} g_{\nu\lambda} = -Q_{\nu\lambda\mu} \quad . \tag{11.101}$$

In terms of the $C^{\mu}_{\ \nu\lambda}$ these tensors can be written as

$$\begin{aligned} T_{\lambda\mu\nu} &= C_{\lambda\mu\nu} - C_{\lambda\nu\mu} = 2C_{\lambda[\mu\nu]} \\ Q_{\nu\lambda\mu} &= C_{\nu\lambda\mu} + C_{\lambda\nu\mu} = 2C_{(\nu\lambda)\mu} \quad . \end{aligned} \tag{11.102}$$

Thus the torsion is zero iff $C^{\lambda}_{\mu\nu}$ (and hence $\tilde{\Gamma}^{\lambda}_{\mu\nu}$) is symmetric in its lower indices, and the connection is compatible with the metric iff $C_{\nu\lambda\mu}$ is anti-symmetric in its first two indices. In particular, if the torsion is zero and the connection is metric-compatible, one has

$$C_{\lambda\mu\nu} = C_{\lambda\nu\mu} \quad \text{and} \quad C_{\lambda\mu\nu} = -C_{\mu\lambda\nu} \quad \Rightarrow \quad C_{\lambda\mu\nu} = 0 \quad , \tag{11.103}$$

as one can see by the gymnastics

$$C_{\lambda\mu\nu} = C_{\lambda\nu\mu} = -C_{\nu\lambda\mu} = -C_{\nu\mu\lambda} = C_{\mu\nu\lambda} = C_{\mu\lambda\nu} = -C_{\lambda\mu\nu} \quad . \tag{11.104}$$

Conversely, since the absence of torsion and non-metricity characterises the Levi-Civita connection, it should be possible to express the deviation $C_{\lambda\mu\nu}$ from the Levi-Civita connection entirely in terms of torsion and non-metricity. This is indeed the case. By repeating the calculation (5.44) in this more general context, one finds

$$2C_{\lambda(\mu\nu)} = Q_{\nu\lambda\mu} + Q_{\mu\lambda\nu} - Q_{\mu\nu\lambda} - T_{\mu\lambda\nu} - T_{\nu\lambda\mu} \quad . \tag{11.105}$$

Combining this with $2C_{\lambda[\mu\nu]} = T_{\lambda\mu\nu}$, one obtains

$$\begin{aligned} C_{\lambda\mu\nu} &= \tfrac{1}{2}(T_{\lambda\mu\nu} - T_{\mu\lambda\nu} - T_{\nu\lambda\mu}) + \tfrac{1}{2}(Q_{\mu\lambda\nu} + Q_{\nu\lambda\mu} - Q_{\mu\nu\lambda}) \\ &\equiv \tilde{T}_{\lambda\mu\nu} + \tilde{Q}_{\lambda\mu\nu} \;\;, \end{aligned} \tag{11.106}$$

with

$$\tilde{T}_{\lambda\mu\nu} = \tfrac{1}{2}(T_{\lambda\mu\nu} - T_{\mu\lambda\nu} - T_{\nu\lambda\mu}) = -\tilde{T}_{\mu\lambda\nu} \tag{11.107}$$

and

$$\tilde{Q}_{\lambda\mu\nu} = \tfrac{1}{2}(Q_{\mu\lambda\nu} + Q_{\nu\lambda\mu} - Q_{\mu\nu\lambda}) = \tilde{Q}_{\lambda\nu\mu} \;\;. \tag{11.108}$$

Thus we can now split a general connection more informatively into the 3 pieces

$$\tilde{\Gamma}^{\lambda}_{\ \mu\nu} = \Gamma^{\lambda}_{\ \mu\nu} + \tilde{T}^{\lambda}_{\ \mu\nu} + \tilde{Q}^{\lambda}_{\ \mu\nu} \;\;. \tag{11.109}$$

REMARKS:

1. The tensor $\tilde{T}_{\lambda\mu\nu}$ is known as the *contorsion tensor* (frequently (mis-)spelled as "contortion" tensor). I am not aware of a commonly used name for $\tilde{Q}_{\lambda\mu\nu}$, and will not try to invent one. The contorsion tensor is the linear combination of components of the torsion tensor that appear as the connection coefficents of a general metric-compatible connection with torsion,

   $$Q_{\lambda\mu\nu} = 0 \quad \Rightarrow \quad \tilde{\Gamma}^{\lambda}_{\ \mu\nu} = \Gamma^{\lambda}_{\ \mu\nu} + \tilde{T}^{\lambda}_{\ \mu\nu} \;\;. \tag{11.110}$$

   In general it can have both symmetric and anti-symmetric components,

   $$\tilde{T}_{\lambda(\mu\nu)} = -\tfrac{1}{2}(T_{\mu\lambda\nu} + T_{\nu\lambda\mu}) = T_{(\mu\nu)\lambda} \quad , \quad \tilde{T}^{\lambda}_{\ [\mu\nu]} = \tfrac{1}{2}T^{\lambda}_{\ \mu\nu} \;\;, \tag{11.111}$$

   but it cannot be symmetric (if the contorsion were symmetric, the torsion, and hence the contorsion, would be zero). If its symmetric part vanishes, then $\tilde{T}_{\lambda\mu\nu}$ is completely anti-symmetric,

   $$\tilde{T}^{\lambda}_{\ (\mu\nu)} = 0 \quad \Rightarrow \quad \tilde{T}_{\lambda\mu\nu} = \tilde{T}_{[\lambda\mu\nu]} \;\;. \tag{11.112}$$

2. Since $\Gamma_{\lambda\mu\nu}$ contains a part that is antisymmetric in the first 2 indices,

   $$\Gamma_{\lambda\mu\nu} = \tfrac{1}{2}g_{\lambda\mu,\nu} + \tfrac{1}{2}(g_{\nu\lambda,\mu} - g_{\nu\mu,\lambda}) \;\;, \tag{11.113}$$

   one might be tempted to think that that part can be cancelled (or absorbed) by a metric-compatible $C_{\lambda\mu\nu} = C_{[\lambda\mu]\nu}$, so that a very simple metric-compatible connection would be

   $$\tilde{\Gamma}_{\lambda\mu\nu} \overset{?}{=} \tfrac{1}{2}g_{\lambda\mu,\nu} \quad , \quad \tilde{\Gamma}^{\lambda}_{\ \mu\nu} \overset{?}{=} \tfrac{1}{2}g^{\lambda\rho}g_{\rho\mu,\nu} \;\;. \tag{11.114}$$

   However, the term that one has cancelled (or absorbed) is *not* a tensor. Therefore, this candidate "connection" does not transform as (and therefore does not qualify as) a connection and cannot be used to define a covariant derivative.

3. In general, for a connection $\tilde{\nabla}$, the notions of autoparallels (section 5.8),

$$\tilde{\nabla}_\tau X^\mu = 0 \quad \Leftrightarrow \quad \ddot{x}^\mu + \tilde{\Gamma}^\mu_{\ \nu\lambda}\dot{x}^\nu \dot{x}^\lambda = 0 \ , \tag{11.115}$$

(i.e. curves characterised by the fact that their tangent vectors are parallel transported along the curve - this depends on a choice of connection) no longer coincides with the notion of geodesics (which are obtained by extremising proper time or distance, and which always lead to the Levi-Civita connection). However, this difference disappears if $C^\mu_{\ \nu\lambda}$ happens to be anti-symmetric in its lower indices (e.g. for a metric-compatible connection with totally anti-symmetric contorsion tensor), as one then has

$$\ddot{x}^\mu + \tilde{\Gamma}^\mu_{\ \nu\lambda}\dot{x}^\nu \dot{x}^\lambda = \ddot{x}^\mu + \Gamma^\mu_{\ \nu\lambda}\dot{x}^\nu \dot{x}^\lambda \ . \tag{11.116}$$

We have defined the Riemann tensor via the commutator of covariant derivatives (8.2)

$$[\nabla_\mu, \nabla_\nu]V^\lambda = R^\lambda_{\ \sigma\mu\nu}V^\sigma \tag{11.117}$$

associated to the Levi-Civita connection (Christoffel symbols), or, equivalently, by the relation (8.5) (now being careful with the positioning of the lower indices)

$$R^\lambda_{\ \sigma\mu\nu} = \partial_\mu \Gamma^\lambda_{\ \sigma\nu} - \partial_\nu \Gamma^\lambda_{\ \sigma\mu} + \Gamma^\lambda_{\ \rho\mu}\Gamma^\rho_{\ \sigma\nu} - \Gamma^\lambda_{\ \rho\nu}\Gamma^\rho_{\ \sigma\mu} \ . \tag{11.118}$$

In order to show explicitly (rather than by appealing to (11.117)) that this transforms as a tensor, all that one needs is the characteristic non-tensorial transformation behaviour of the Christoffel symbols $\Gamma^\lambda_{\ \mu\nu}$. As discussed in section 5.4 and above, an arbitrary connection $\tilde{\Gamma}^\lambda_{\ \mu\nu}$ that can be used to define a tensorial covariant derivative has the same non-tensorial transformation behaviour. Therefore

$$\tilde{R}^\lambda_{\ \sigma\mu\nu} \equiv R^\lambda_{\ \sigma\mu\nu}(\tilde{\Gamma}) = \partial_\mu \tilde{\Gamma}^\lambda_{\ \sigma\nu} - \partial_\nu \tilde{\Gamma}^\lambda_{\ \sigma\mu} + \tilde{\Gamma}^\lambda_{\ \rho\mu}\tilde{\Gamma}^\rho_{\ \sigma\nu} - \tilde{\Gamma}^\lambda_{\ \rho\mu}\tilde{\Gamma}^\rho_{\ \sigma\mu} \tag{11.119}$$

defines a tensor for any connection, namely the curvature tensor of the connection $\tilde{\Gamma}^\lambda_{\ \mu\nu}$. It is related to the commutator of covariant derivatives by

$$[\tilde{\nabla}_\mu, \tilde{\nabla}_\nu]V^\lambda = \tilde{R}^\lambda_{\ \sigma\mu\nu}V^\sigma + (\tilde{\Gamma}^\rho_{\ \mu\nu} - \tilde{\Gamma}^\rho_{\ \nu\mu})\nabla_\rho V^\lambda = \tilde{R}^\lambda_{\ \sigma\mu\nu}V^\sigma + T^\rho_{\mu\nu}\nabla_\rho V^\lambda \ , \tag{11.120}$$

where $T^\rho_{\mu\nu}$ is the torsion tensor. As before, one can also define the Ricci tensor and Ricci scalar by

$$\tilde{R}_{\mu\nu} \equiv R_{\mu\nu}(\tilde{\Gamma}) = \tilde{R}^\lambda_{\ \mu\lambda\nu} \quad , \quad \tilde{R} \equiv R(\tilde{\Gamma}) = g^{\mu\nu}\tilde{R}_{\mu\nu} \ . \tag{11.121}$$

However, it is crucial to keep in mind that the symmetry properties and Bianchi identities satisfied by these generalised curvature tensors will in general differ from those of the Riemann-Christoffel tensor. This should be clear from the way we derived the symmetries of the Riemann tensor in section 8.3, where we related the symmetries to the properties (metricity, no torsion) that characterise the canonical Levi-Civita connection (Christoffel symbols). For example, in general the Ricci tensor will not be symmetric, the Bianchi identity $R_{\alpha[\beta\gamma\delta]} = 0$ will be replaced by an identity relating $\tilde{R}_{\alpha[\beta\gamma\delta]}$ to the torsion (and its covariant derivative), etc.[28]

---

[28] For more on this and related topics, see e.g. section 1 of T. Ortin, *Gravity and Strings*.

12 CURVATURE III: CURVATURE AND GEODESIC CONGRUENCES

In section 8.4 we had already encountered the so-called *geodesic deviation equation* (8.38),

$$(D_\tau)^2 \delta x^\mu = R^\mu_{\ \nu\lambda\rho} \dot{x}^\nu \dot{x}^\lambda \delta x^\rho \quad , \tag{12.1}$$

describing the evolution of a separation (or deviation) vector along a given geodesic. In this section we will rederive this result in a more satisfactory and covariant manner and also use the same covariant framework to discuss the extension of these results to the so-called *Raychaudhuri equation*, which descibes the focussing properties of congruences of geodesics.

12.1 COVARIANT DERIVATION OF THE GEODESIC DEVIATION EQUATION

The starting point is a geodesic with tangent vector field u^α,

$$u^\beta \nabla_\beta u^\alpha = 0 \quad , \tag{12.2}$$

and a deviation vector field $\delta x^\alpha = \xi^\alpha$ characterised by the condition

$$[u, \xi]^\alpha = u^\beta \nabla_\beta \xi^\alpha - \xi^\beta \nabla_\beta u^\alpha = 0 \quad \Leftrightarrow \quad D_\tau \xi^\alpha = \xi^\beta \nabla_\beta u^\alpha \quad . \tag{12.3}$$

The rationale for this condition is that, if $x^\alpha(\tau, s)$ is a family of geodesics labelled by s, one has the identifications

$$u^\alpha = \frac{\partial}{\partial \tau} x^\alpha(\tau, s) \quad , \quad \xi^\alpha = \frac{\partial}{\partial s} x^\alpha(\tau, s) \quad . \tag{12.4}$$

Since second partial derivatives commute, this implies the relation

$$\frac{\partial}{\partial \tau} \xi^\alpha(\tau, s) = \frac{\partial}{\partial s} u^\alpha(\tau, s) \quad , \tag{12.5}$$

(implicit in the identification $\delta \dot{x} = (d/d\tau)\delta x$ employed in the derivation in section 8.4). Condition (12.3) is nothing other than the covariant way of writing (12.5).

Introducing the tensor

$$B_{\alpha\beta} = \nabla_\beta u_\alpha \tag{12.6}$$

we can write (12.3) as

$$D_\tau \xi^\alpha = B^\alpha_{\ \beta} \xi^\beta \quad , \tag{12.7}$$

so the matrix $B_{\alpha\beta}$ describes the evolution and deformation of the deviation vector ξ^α along the geodesic. Because u^α is affinely geodesic, it satisfies

$$B_{\alpha\beta} u^\beta = u^\beta \nabla_\beta u_\alpha = 0 \tag{12.8}$$

and

$$u^\alpha B_{\alpha\beta} = \tfrac{1}{2} \nabla_\beta (u^\alpha u_\alpha) = 0 \quad , \tag{12.9}$$

and is thus transverse to u^α. This is a crucial property we will come back to in the discussion of the Raychaudhuri equation below. As a consequence one has

$$u_\alpha D_\tau \xi^\alpha = 0 \tag{12.10}$$

(i.e. $D_\tau \xi^\alpha$ is transverse to u^α) and therefore

$$\frac{d}{d\tau}(u_\alpha \xi^\alpha) = D_\tau(u_\alpha \xi^\alpha) = u_\alpha D_\tau \xi^\alpha = 0 \ . \tag{12.11}$$

This means that the u-component of a geodesic deviation vector ξ in the sense of $u_\alpha \xi^\alpha$ is simply constant and contains no interesting information about the geodesic itself.

In the timelike case this means that a vector of the form $\xi^\alpha = \xi u^\alpha$ is a deviation vector only if ξ is constant, and then ξ^α is simply a translation along the geodesic and therefore not a deviation vector of interest (and certainly anyhow not a vector of the kind one has in mind when thinking about a deviation vector, which should point away from the geodesic). In the null case, the interpretation is slightly different (and we will return to this in section 12.4), but the fact that $u_\alpha \xi^\alpha$ is simply constant for a deviation vector remains, and we can without loss of information choose the deviation vector to satisfy the condition $\xi^\alpha u_\alpha = 0$.

Given this set-up, we now want to calculate

$$\begin{aligned}(D_\tau)^2 \xi^\alpha &= (D_\tau B^\alpha_{\ \beta})\xi^\beta + B^\alpha_{\ \beta} D_\tau \xi^\beta \\ &= (D_\tau B^\alpha_{\ \gamma} + B^\alpha_{\ \beta} B^\beta_{\ \gamma})\xi^\gamma \ .\end{aligned} \tag{12.12}$$

Note that, along with $D_\tau \xi^\alpha$, also $D_\tau^2 \xi^\alpha$ is automatically transverse to u^α, $u_\alpha D_\tau^2 \xi^\alpha = 0$, regardless of whether or not one imposes the condition $\xi^\alpha u_\alpha = 0$.

For the term in brackets we find, using the geodesic equation for u^α,

$$\begin{aligned}D_\tau B^\alpha_{\ \gamma} + B^\alpha_{\ \beta} B^\beta_{\ \gamma} &= u^\beta \nabla_\beta \nabla_\gamma u^\alpha + (\nabla_\gamma u^\beta)\nabla_\beta u^\alpha \\ &= u^\beta \nabla_\beta \nabla_\gamma u^\alpha + \nabla_\gamma(u^\beta \nabla_\beta u^\alpha) - u^\beta \nabla_\gamma \nabla_\beta u^\alpha \\ &= u^\beta(\nabla_\beta \nabla_\gamma - \nabla_\gamma \nabla_\beta)u^\alpha = R^\alpha_{\ \delta\beta\gamma} u^\beta u^\delta \ ,\end{aligned} \tag{12.13}$$

and plugging this back into (12.12), we obtain straightaway the covariant version (8.38) of the geodesic deviation equation in the form

$$(D_\tau)^2 \xi^\alpha = R^\alpha_{\ \delta\beta\gamma} u^\delta u^\beta \xi^\gamma \ . \tag{12.14}$$

Note that this result is automatically transverse to u^α,

$$u_\alpha (D_\tau)^2 \xi^\alpha = R_{\alpha\delta\beta\gamma} u^\alpha u^\delta u^\beta \xi^\gamma = 0 \ . \tag{12.15}$$

REMARKS:

1. I hope you agree that this derivation is somewhat more satisfactory than the one given in section 8.4.

2. The object we have called $B_{\alpha\beta}$ in (12.6) and its evolution equation (12.13) will play a central role in our derivation of the Raychaudhuri equation below.

3. When considering a null geodesic, the condition $\xi^\alpha u_\alpha = 0$ does not eliminate the component of ξ^α tangent to the geodesic. In that case it is convenient to introduce an auxiliarly linearly independent null vector field in order to be able to project the deviation vector ξ^α and its derivatives $D_\tau \xi^\alpha$ and $D_\tau^2 \xi^\alpha$ into some spatial codimension 2 plane transverse to these null directions. This *transverse null geodesic equation* will be derived and discussed in section 12.3.

4. If the curve is not a geodesic (but still parametrised by proper time, so that $u^\alpha u_\alpha = -1$), then the above derivation shows that in addition to the force exerted by the space-time curvature the deviation vector feels a force proportional to the change of the acceleration $a^\alpha = u^\beta \nabla_\beta u^\alpha$ along the curve,

   $$(D_\tau)^2 \xi^\alpha = R^\alpha_{\ \delta\beta\gamma} u^\delta u^\beta \xi^\gamma + D_\tau a^\alpha \quad . \tag{12.16}$$

   In flat space, only the last term is present and describes the (tidal) forces arising from the possible non-uniformity of the external force acting on the particle (or, better: on the extended object described by a family of worldlines) to produce the acceleration a^α. Thus, in precise analogy with the Newtonian situation, the gravitational (i.e. here now Riemann curvature tensor) contribution to the geodesic deviation equation should be interpreted as the *gravitational tidal force*.

12.2 RAYCHAUDHURI EQUATION FOR TIMELIKE GEODESIC CONGRUENCES

A *congruence of curves* is a (locally) space-time filling family of curves, i.e. it is such that locally for any space-time point there is a unique curve passing through that point. A (timelike) *geodesic congruence* is then a congruence of (timelike) geodesics.

Manipulations similar to those leading to (12.14) allow one to derive an equation for the rate of change of the divergence $\nabla_\alpha u^\alpha$ of a family of geodesics along the geodesics. This simple result, known as the *Raychaudhuri equation*, has important implications and ramifications in general relativity, in particular in the context of the so-called *singularity theorems* of Penrose, Hawking and others, none of which will, however, be explored here (see footnote 96 of section 29.3 for some references).

Thus u^α now denotes a tangent vector field to an affinely parametrised geodesic congruence, $u^\alpha \nabla_\alpha u^\beta = 0$ (and $u^\alpha u_\alpha = -1$ or $u^\alpha u_\alpha = 0$ everywhere for a timelike or null congruence). As in section 12.1, we introduce the tensor field (12.6)

$$B_{\alpha\beta} = \nabla_\beta u_\alpha \quad . \tag{12.17}$$

Recall from section 12.1 that $B_{\alpha\beta}$ satisfies (12.8), (12.9),

$$B_{\alpha\beta} u^\beta = u^\alpha B_{\alpha\beta} = 0 \tag{12.18}$$

and therefore only has components in the directions transverse to u^α. Its trace

$$\theta = B^\alpha_{\ \alpha} = g^{\alpha\beta} B_{\alpha\beta} = \nabla_\alpha u^\alpha \tag{12.19}$$

is the divergence of u^α and is known as the *expansion* of the (affinely parametrised) geodesic congruence.

The key equation governing the evolution of $B_{\alpha\beta}$ along the integral curves of the geodesic vector field is (12.13)

$$D_\tau B^\alpha_{\ \gamma} + B^\alpha_{\ \beta} B^\beta_{\ \gamma} = R^\alpha_{\ \delta\beta\gamma} u^\beta u^\delta \quad . \tag{12.20}$$

By taking the trace of this equation, we evidently obtain an evolution equation for the expansion θ, namely

$$\frac{d}{d\tau}\theta = -(\nabla_\alpha u_\beta)(\nabla^\beta u^\alpha) - R_{\alpha\beta} u^\alpha u^\beta \quad . \tag{12.21}$$

Note that this equation, written in the form

$$u^\beta \nabla_\beta (\nabla_\alpha u^\alpha) + (\nabla_\alpha u_\beta)(\nabla^\beta u^\alpha) + R_{\alpha\beta} u^\alpha u^\beta = 0 \quad . \tag{12.22}$$

is a special case of the "master equation" (8.53) for $V^\alpha \to u^\alpha$ with $u^\beta \nabla_\beta u^\alpha = 0$.

To gain some more insight into the geometric significance of this equation, we now consider the case that the geodesic congruence u^α is timelike and normalised in the standard way as $u^\alpha u_\alpha = -1$ (so that τ is proper time).

Given this timelike geodesic congruence, we can introduce the tensor

$$h_{\alpha\beta} = g_{\alpha\beta} + u_\alpha u_\beta \quad . \tag{12.23}$$

The properties of this tensor are closely related to those of the (induced metric) tensor $h_{\alpha\beta} = g_{\alpha\beta} - \epsilon N_\alpha N_\beta$ (16.1) studied in section 16.1 in the context of hypersurfaces. The main difference in the present context is that u_α is not necessarily hypersurface-orthogonal (section 15.5) and therefore, in particular, not necessarily a normal vector field to a familiy of spacelike hypersurfaces. Therefore $h_{\alpha\beta}$ does not necesarily have an interpretation as the induced metric on some hypersurface. Nevertheless, pointwise it can be interpreted as a metric on the space of vectors transverse to the geodesic and its purely algebraic properties are identical to those of the induced metric.

In particular,

- $h_{\alpha\beta}$ has the characteristic property that it is orthogonal to u^α,

$$u^\alpha h_{\alpha\beta} = h_{\alpha\beta} u^\beta = 0 \quad . \tag{12.24}$$

- It can therefore be interpreted as the spatial projection of the metric in the directions orthogonal to the timelike vector field u^α. This can be seen more explicitly in terms of the projectors

$$\begin{aligned} h^{\alpha}_{\ \beta} &= \delta^{\alpha}_{\ \beta} + u^{\alpha}u_{\beta} \\ h^{\alpha}_{\ \beta}h^{\beta}_{\ \gamma} &= h^{\alpha}_{\ \gamma} \ . \end{aligned} \tag{12.25}$$

On directions tangential to u^{α} they act as

$$h^{\alpha}_{\ \beta}u^{\beta} = 0 \ , \tag{12.26}$$

whereas on vectors ξ^{α} orthogonal to u^{α}, $u_{\alpha}\xi^{\alpha} = 0$ (spacelike vectors), one has

$$h^{\alpha}_{\ \beta}\xi^{\beta} = \xi^{\alpha} \ . \tag{12.27}$$

- Thus, acting on an arbitrary vector field V^{α}, $v^{\alpha} = h^{\alpha}_{\ \beta}V^{\beta}$ is the projection of this vector into the plane orthogonal to u^{α}. In the same way one can project an arbitrary tensor to a spatial or transverse tensor. E.g.

  $$t_{\alpha\ldots\beta} = T_{\gamma\ldots\delta}\, h^{\gamma}_{\ \alpha}\ldots h^{\delta}_{\ \beta} \tag{12.28}$$

  satisfies

  $$u^{\alpha}t_{\alpha\ldots\beta} = \ldots = u^{\beta}t_{\alpha\ldots\beta} = 0 \ . \tag{12.29}$$

- In particular, the projection of the metric is

  $$g_{\alpha\beta} \to g_{\gamma\delta}h^{\gamma}_{\ \alpha}h^{\delta}_{\ \beta} = g_{\alpha\beta} + u_{\alpha}u_{\beta} = h_{\alpha\beta} \ , \tag{12.30}$$

  as anticipated above. Whereas for the space-time metric one obviously has $g^{\alpha\beta}g_{\alpha\beta} = 4$, the trace of $h_{\alpha\beta}$ is (in the 4-dimensional case)

  $$g^{\alpha\beta}h_{\alpha\beta} = g^{\alpha\beta}g_{\alpha\beta} + g^{\alpha\beta}u_{\alpha}u_{\beta} = 4 - 1 = 3 = h^{\alpha\beta}h_{\alpha\beta} \ . \tag{12.31}$$

Thus for an affinely parametrised congruence the properties (12.8) and (12.9) show that $B_{\alpha\beta}$ is automatically a spatial or transverse tensor in the sense above,

$$b_{\alpha\beta} \equiv h_{\alpha}^{\gamma}h_{\beta}^{\delta}B_{\gamma\delta} = B_{\alpha\beta} \ . \tag{12.32}$$

Note that the affine parametrisation of the timelike geodesic congruence, expressed by the normalisation condition $u^{\alpha}u_{\alpha} = -1$, is crucial for this entire set-up, since the projection operator requires a unit vector field. This is to be contrasted with the situation for null geodesic congruences ℓ^{α}, to be discussed below, where the property $\ell^{\alpha}\ell_{\alpha} = 0$ is independent of the parametrisation and one can (and we will) also consider the case of non-affine parametrisations.

In the spirit of elasticity theory, we now decompose $b_{\alpha\beta}$ into its anti-symmetric, symmetric traceless and trace part,

$$b_{\alpha\beta} = \omega_{\alpha\beta} + \sigma_{\alpha\beta} + \tfrac{1}{3}\theta h_{\alpha\beta} \ , \tag{12.33}$$

with

$$\begin{aligned} \omega_{\alpha\beta} &= \tfrac{1}{2}(b_{\alpha\beta} - b_{\alpha\beta}) \\ \sigma_{\alpha\beta} &= \tfrac{1}{2}(b_{\alpha\beta} + b_{\alpha\beta}) - \tfrac{1}{3}\theta h_{\alpha\beta} \\ \theta &= h^{\alpha\beta} b_{\alpha\beta} = g^{\alpha\beta} B_{\alpha\beta} = \nabla_\alpha u^\alpha \quad . \end{aligned} \tag{12.34}$$

The quantities $\omega_{\alpha\beta}$, $\sigma_{\alpha\beta}$ and θ are known as the *rotation tensor*, *shear tensor*, and *expansion* of the congruence (family) of geodesics defined by u^α.

In terms of these quantities we can write the evolution equation (12.7) for deviation vectors as

$$D_\tau \xi^\alpha = \omega^\alpha_{\ \beta}\xi^\beta + \sigma^\alpha_{\ \beta}\xi^\beta + \tfrac{1}{3}\theta\xi^\alpha \quad , \tag{12.35}$$

and the evolution equation (12.21) for the expansion θ as

$$\frac{d}{d\tau}\theta = -\tfrac{1}{3}\theta^2 - \sigma^{\alpha\beta}\sigma_{\alpha\beta} + \omega^{\alpha\beta}\omega_{\alpha\beta} - R_{\alpha\beta}u^\alpha u^\beta \quad . \tag{12.36}$$

This is the *Raychaudhuri equation* for timelike geodesic congruences.

REMARKS:

1. The expansion θ can be written as

   $$\begin{aligned} \theta &= h^{\alpha\beta} b_{\alpha\beta} = h^{\alpha\beta} B_{\alpha\beta} = h^{\alpha\beta}\nabla_\beta u_\alpha \\ &= \tfrac{1}{2}h^{\alpha\beta}(\nabla_\beta u_\alpha + \nabla_\alpha u_\beta) = \tfrac{1}{2}h^{\alpha\beta} L_u g_{\alpha\beta} \end{aligned} \tag{12.37}$$

   where L_u denotes the Lie derivative along the vector field u. Substituting $g_{\alpha\beta} = h_{\alpha\beta} - u_\alpha u_\beta$, one finds

   $$\theta = \tfrac{1}{2}h^{\alpha\beta} L_u(h_{\alpha\beta} - u_\alpha u_\beta) = \tfrac{1}{2}h^{\alpha\beta} L_u h_{\alpha\beta} \quad . \tag{12.38}$$

   Recalling the formula

   $$\delta\sqrt{g} = \sqrt{g} g^{\alpha\beta}\delta g_{\alpha\beta}/2 \tag{12.39}$$

   for the variation of a volume element induced by a variation of the metric, morally speaking the above equation says that θ measures the change of a transverse (cross-sectional) volume of the congruence with volume element $\sqrt{h}$ as one moves along the geodesics,

   $$\theta = \frac{1}{\sqrt{h}}\frac{d}{d\tau}\sqrt{h} \quad . \tag{12.40}$$

   The statement as such is correct and provides the correct intuition for the meaning of θ, but the definition of the cross-sectional volume and its volume element require a bit of care. When the congruence is hypersurface orthogonal, with the induced metric (16.6)

   $$h_{ab} = E^\alpha_a E^\beta_b h_{\alpha\beta} \quad , \tag{12.41}$$

then (12.40) with $h = \det(h_{ab})$ follows from (12.38), because

$$\begin{aligned} h^{ab} L_u h_{ab} &= h^{ab} L_u (E^\alpha_a E^\beta_b h_{\alpha\beta}) \\ &= h^{ab} E^\alpha_a E^\beta_b L_u h_{\alpha\beta} = h^{\alpha\beta} L_u h_{\alpha\beta} \ . \end{aligned} \tag{12.42}$$

Here we have made use of the fact that u and E_a have vanishing Lie bracket, because (introducing τ and y^a as coordinates, instead of the x^α)

$$u^\alpha = \frac{\partial x^\alpha}{\partial \tau} \quad , \quad E^\alpha_a = \frac{\partial x^\alpha}{\partial y^a} \tag{12.43}$$

and the Lie bracket gives the commutator of the second partical derivatives of x^α.

When the congruence is not hypersurface orthogonal, one can still construct a transverse cross-sectional volume, but one can only choose it to be orthogonal at a given geodesic. Introducing in a neighbourhood of a point on this geodesic coordinates y^a labelling the geodesics, as well as the parameter τ along the geodesic, the above calculation will then still go through.[29]

2. If required and desired, from (12.20) similar (but somewhat less transparent) equations can be derived for the evolution of the shear and rotation tensors along the geodesic congruence, i.e. for $(d/d\tau)\sigma_{\alpha\beta}$ and $(d/d\tau)\omega_{\alpha\beta}$.

3. Since $\omega_{\alpha\beta}$ and $\sigma_{\alpha\beta}$ are purely spatial tensors, their squares are non-negative,

$$\sigma^{\alpha\beta}\sigma_{\alpha\beta} \geq 0 \quad , \quad \omega^{\alpha\beta}\omega_{\alpha\beta} \geq 0 \ , \tag{12.44}$$

with $\sigma^{\alpha\beta}\sigma_{\alpha\beta} = 0$ only for $\sigma_{\alpha\beta} = 0$ (and likewise for the rotation). They thus enter the Raychaudhuri equation with opposite signs.

4. In the presence of both these terms it is difficult to say something general about the evolution of θ. Since the first term $(-\theta^2/3)$ is non-positive, an important special case of the Raychaudhuri equation arises when the rotation is zero, $\omega_{\alpha\beta} = 0$. This happens for example when $u_\alpha = \partial_\alpha S$ is the gradient co-vector of some function S. In this case u_α is orthogonal to the level-surfaces of S. In fact, more generally we have the statement that

$$\omega_{\alpha\beta} = 0 \quad \Leftrightarrow \quad u^\alpha \quad \text{hypersurface orthogonal} \ . \tag{12.45}$$

Indeed, assume that u_α is hypersurface orthogonal, i.e. (15.55)

$$u_{[\alpha}\nabla_\beta u_{\gamma]} = 0 \quad \Leftrightarrow \quad \omega_{\alpha\beta}u_\gamma + \omega_{\beta\gamma}u_\alpha + \omega_{\gamma\alpha}u_\beta = 0 \ . \tag{12.46}$$

Contracting this with u^γ and using $u^\gamma u_\gamma = -1$ and $u^\gamma \omega_{\gamma\beta} = 0$, only the first term survives and one finds on the nose that $\omega_{\alpha\beta} = 0$,

$$u_{[\alpha}\nabla_\beta u_{\gamma]} = 0 \quad \Rightarrow \quad \omega_{\alpha\beta} = 0 \ , \tag{12.47}$$

---

[29] For a more careful proof of this statement see the discussion in section 2.4.8 of E. Poisson, *A Relativist's Toolkit*.

and the Frobenius theorem provides one with the converse statement. Alternatively, $\omega_{\alpha\beta} = 0$ follows from assuming that u_α has the explicit hypersurface-orthogonal form $u_\alpha = f\partial_\alpha S$. Then one has (15.52)

$$\omega_{\beta\alpha} = \nabla_{[\alpha} u_{\beta]} = (\nabla_{[\alpha} \log f) u_{\beta]} \tag{12.48}$$

and by contraction with u^α one deduces

$$\partial_\alpha f = -(u^\beta \partial_\beta f) u_\alpha \sim u_\alpha \quad \Rightarrow \qquad \omega_{\alpha\beta} = 0 \ . \tag{12.49}$$

5. Either way, for a hypersurface orthogonal congruence of timelike geodesics one has

$$\frac{d}{d\tau}\theta = -\tfrac{1}{3}\theta^2 - \sigma^{\alpha\beta}\sigma_{\alpha\beta} - R_{\alpha\beta}u^\alpha u^\beta \ . \tag{12.50}$$

The first two terms on the right hand side are manifestly non-positive (recall that $\sigma_{\alpha\beta}$ is a spatial tensor and hence $\sigma_{\alpha\beta}\sigma^{\alpha\beta} \geq 0$). Thus, if one assumes that the geometry is such that

$$R_{\alpha\beta}u^\alpha u^\beta \geq 0 \tag{12.51}$$

(by the Einstein equations to be discussed in the section 19, this translates into a positivity condition on the energy-momentum tensor known as the *strong energy condition*, cf. section 22.1), one finds

$$\frac{d}{d\tau}\theta = -\tfrac{1}{3}\theta^2 - \sigma^{\alpha\beta}\sigma_{\alpha\beta} - R_{\alpha\beta}u^\alpha u^\beta \leq 0 \ . \tag{12.52}$$

This means that the divergence (convergence) of geodesics will decrease (increase) in time. The interpretation of this result is that gravity is an *attractive* force (for matter satisfying the strong energy condition) whose effect is to focus geodesics.

6. According to (12.52), $d\theta/d\tau$ is not only negative but actually bounded from above by

$$\frac{d}{d\tau}\theta \leq -\tfrac{1}{3}\theta^2 \ . \tag{12.53}$$

Rewriting this equation as

$$\frac{d}{d\tau}\frac{1}{\theta} \geq \frac{1}{3} \ , \tag{12.54}$$

one deduces immediately that

$$\frac{1}{\theta(\tau)} \geq \frac{1}{\theta(0)} + \frac{\tau}{3} \ . \tag{12.55}$$

This has the rather dramatic implication that, if $\theta(0) < 0$ (i.e. the geodesics are initially converging), then $\theta(\tau) \to -\infty$ within finite proper time $\tau \leq 3/|\theta(0)|$,

$$\theta(\tau) \to -\infty \quad \text{for} \quad \tau \leq 3/|\theta(0)| \tag{12.56}$$

provided that the geodesics can be extended that far.

7. If one thinks of the geodesics as trajectories of physical particles, this is obviously a rather catastrophic situation in which these particles will be infinitely squashed. In general, however, the divergence of θ only indicates that the family of geodesics develops what is known as a *caustic* where different geodesics meet.

8. Simple non-catastrophic examples of such caustics are e.g. the poles of a sphere where great circles meet, or even just the origin in Euclidean space $\mathbb{R}^n$ when considering the family of radial geodesics passing through the origin. E.g. in the latter case the tangent vector field is simply ∂_r, and its divergence is

$$\nabla_\alpha(\partial_r)^\alpha = \frac{1}{\sqrt{g}}\partial_\alpha(\sqrt{g}(\partial_r)^\alpha) \sim r^{-1} \quad , \tag{12.57}$$

   which diverges as $r \to 0$. This divergent behaviour is strictly related to the breakdown at the origin $r = 0$ of spherical coordinates adapted to this congruence (cf. also the related discussion in section 43.6).

9. Nevertheless, the above result plays a crucial role in establishing the occurrence of true singularities in general relativity if supplemented e.g. by conditions which ensure that such "harmless" caustics cannot appear, as this means that the geodesic cannot be extended to where one would find $\theta \to -\infty$. This kind of argument (leading to the conclusion of *geodesic incompleteness* of a space-time) is one of the typical ingredients of the singularity theorems of general relativity (see footnote 96 of section 29.3 for some references).

10. The adaptation of this formalism in general and the Raychaudhuri equation in particular to congruences of null geodesics requires some more care (and is ultimately expressed in terms of 2-dimensional rather than 3-dimensional spatial tensors), and we will discuss this in section 12.4 below.

12.3 TRANSVERSE NULL GEODESIC DEVIATION EQUATION

In section 12.4 we will derive the null counterpart of the Raychaudhuri equation for timelike geodesic congruences discussed in section 12.2 above. The set-up we will use is a suitable combination of that for timelike geodesics and the formalism of projectors adapted to null directions. As a preparation for this, and a useful by-product, in this section we will first derive a variant of the geodesic deviation equation for null geodesics, the *transverse null geodesic deviation equation*.

Thus we consider a null geodesic (or congruence of null geodesics), with tangent vector field ℓ^α, and we will initially choose these null geodesics to be affinely parametrised so that one has

$$\ell^\alpha \ell_\alpha = 0 \quad , \quad \ell^\alpha \nabla_\alpha \ell^\beta = 0 \ . \tag{12.58}$$

The affine parameter along the null geodesics of this congruence will (for lack of imagination) be called τ.

Now recall from the discussion of the geodesic deviation equation in section 12.1 that for any geodesic deviation vector ξ^α, i.e. a vector satisfying the condition

$$D_\tau \xi^\alpha = \xi^\beta \nabla_\beta u^\alpha \quad , \tag{12.59}$$

the quantity $\xi^\alpha u_\alpha$ is constant along the geodesic,

$$\frac{d}{d\tau}(\xi^\alpha u_\alpha) = (D_\tau \xi^\alpha) u_\alpha = \xi^\beta u^\alpha \nabla_\beta u^\alpha = 0 \tag{12.60}$$

because $u^\alpha u_\alpha = \epsilon$ is constant, regardless of whether u^α is timelike or null. However, the interpretation of this and its implications depend on whether one is dealing with a timelike geodesic (congruence) or a null geodesic (congruence):

1. In the timelike case the quantity $(-\xi^\alpha u_\alpha)$ is the uninteresting component of ξ along the geodesic u^α, and there was clearly no point in not setting it to zero. The transversality condition $\xi^\alpha u_\alpha = 0$ on the deviation vector could be consistently imposed and was sufficient to remove this component.

2. In the null case, however, the condition $\xi^\alpha \ell_\alpha = 0$ does not accomplish this, i.e. does not remove the component of ξ pointing in the direction of ℓ^α because it imposes no condition precisely on that component. Thus we expect the deviation vector ξ to have *two* uninteresting components in the null case, $\xi^\alpha \ell_\alpha$ and the component of ξ in the direction of ℓ^α:

   - Indeed, as recalled above, we already know in general that $\xi^\alpha \ell_\alpha$ is constant,

     $$\frac{d}{d\tau}(\xi^\alpha \ell_\alpha) = 0 \quad . \tag{12.61}$$

   - To verify this also for a component of ξ along ℓ^α, we impose the condition that ξ be a deviation vector on the vector

     $$\xi^\alpha = \xi \ell^\alpha \tag{12.62}$$

     (note once again that this ξ is completely unrelated to $\xi^\alpha \ell_\alpha$). Since ℓ_α is a geodesic, we have

     $$D_\tau \xi^\alpha = (\frac{d}{d\tau}\xi)\ell^\alpha \overset{!}{=} \xi^\beta \nabla_\beta \ell^\alpha = \xi \ell^\beta \nabla_\beta \ell^\alpha = 0 \quad . \tag{12.63}$$

     Thus $\xi^\alpha = \xi \ell^\alpha$ is a (boring non-) deviation vector iff ξ is constant, and one should impose the condition $\xi = 0$.

Therefore it is natural to project out both these components from ξ_α. In order to construct a suitable projection operator, one can proceed as in section 17.4 and introduce a complementary null vector (field) n^α with

$$n^\alpha n_\alpha = 0 \quad , \quad n^\alpha \ell_\alpha = -1 \ . \tag{12.64}$$

Then

$$\xi^\alpha = \xi \ell^\alpha + \ldots \quad \Rightarrow \quad \xi = -\xi^\alpha n_\alpha \tag{12.65}$$

and we can elininate both boring components by imposing the transversality conditions

$$\xi^\alpha \ell_\alpha = \xi^\alpha n_\alpha = 0 \tag{12.66}$$

on the deviation vector ξ^α, in addition to the deviation vector condition (12.3),

$$\ell^\alpha \nabla_\alpha \xi^\beta = \xi^\alpha \nabla_\alpha \ell^\beta \ . \tag{12.67}$$

As in (12.6) we introduce

$$B_{\alpha\beta} = \nabla_\beta \ell_\alpha \ , \tag{12.68}$$

so that the above condition that ξ^α is a deviation vector can be written as

$$D_\tau \xi^\alpha = B^\alpha_{\ \beta} \xi^\beta \ . \tag{12.69}$$

Because ℓ^α is null and (affinely) geodesic, one has

$$\ell^\alpha B_{\alpha\beta} = B_{\alpha\beta} \ell^\beta = 0 \ , \tag{12.70}$$

but $B_{\alpha\beta}$ is not automatically orthogonal to n^α (and we will come back to and recitify this below). Exactly the same calculation as (12.13) in section 12.1 now shows that

$$D_\tau B_{\alpha\beta} + B^\alpha_{\ \gamma} B^\gamma_{\ \beta} = R^\alpha_{\ \delta\gamma\beta} \ell^\gamma \ell^\delta \tag{12.71}$$

so that one also has the null counterpart of (12.14), namely

$$(D_\tau)^2 \xi^\alpha = R^\alpha_{\ \delta\gamma\beta} \ell^\gamma \ell^\delta \xi^\beta \ . \tag{12.72}$$

Again this equation is automatically transverse to ℓ, in the sense that

$$\ell_\alpha (D_\tau)^2 \xi^\alpha = R_{\alpha\delta\gamma\beta} \ell^\alpha \ell^\gamma \ell^\delta \xi^\beta = 0 \ , \tag{12.73}$$

but not transverse to n^α.

In order to pick up only the transverse components of $D_\tau \xi^\alpha$ in (12.69) (and in subsequent equations), we thus need to project $B_{\alpha\beta}$ onto its transverse components. The construction of this projector is identical to that in section 17.4:

- Associated with a choice of n^α we have a decomposition of the metric into a longitudinal and a transverse spatial part,

  $$g_{\alpha\beta} = s_{\alpha\beta} - (\ell_\alpha n_\beta + \ell_\beta n_\alpha) \ , \tag{12.74}$$

  with the properties

  $$s_{\alpha\beta}\ell^\beta = s_{\alpha\beta}n^\beta = 0 \tag{12.75}$$

  and

  $$g^{\alpha\beta}s_{\alpha\beta} = s^{\alpha\beta}s_{\alpha\beta} = s^\alpha_{\ \alpha} = 2 \ . \tag{12.76}$$

  (we are considering the case $D = 4$).

- We then also have the corresponding transverse (spatial) projectors

  $$\begin{aligned} s^\alpha_{\ \beta} = \delta^\alpha_{\ \beta} + (\ell^\alpha n_\beta + n^\alpha \ell_\beta): \quad & s^\alpha_{\ \gamma}s^\gamma_{\ \beta} = s^\alpha_{\ \beta} \\ & s^\alpha_{\ \beta}\ell^\beta = s^\alpha_{\ \beta}n^\beta = 0 \ , \end{aligned} \tag{12.77}$$

With the aid of these projectors, we can now write the fully projected version of (12.69) as

$$s^\alpha_{\ \beta}D_\tau(s^\beta_{\ \gamma}\xi^\gamma) = b^\alpha_{\ \beta}\xi^\beta \tag{12.78}$$

where $b_{\alpha\beta}$ is the projection of $B_{\alpha\beta}$,

$$b_{\alpha\beta} = s_\alpha^{\ \gamma}s_\beta^{\ \delta}B_{\gamma\delta} \ . \tag{12.79}$$

Likewise the purely transverse (to ℓ and n) variant of the null geodesic equation (12.72) can be written as

$$s^\alpha_{\ \beta}(D_\tau)^2(s^\beta_{\ \gamma}\xi^\gamma) = s^\alpha_{\ \beta}s^\gamma_{\ \delta}R^\beta_{\ \mu\nu\gamma}\ell^\mu\ell^\nu\xi^\delta \ . \tag{12.80}$$

While this is essentially the final result, it is not particularly transparent yet. We will put this equation into a somewhat more attractive form below, in which manifestly only the transverse components of the deviation vector and $R^\beta_{\ \mu\nu\gamma}\ell^\mu\ell^\nu$ appear.

First of all, note that the auxiliary normal vector n^α is not unique. For a fixed choice of ℓ^α, at a point on the geodesic, that is for a given value of τ, it is uniquely determined up to null rotations around ℓ^α,

$$\ell \to \ell \quad , \quad n \to n + \beta^a E_a + \tfrac{1}{2}\beta^2\ell \quad , \quad E_a \to E_a + \beta_a \ell \ , \tag{12.81}$$

where E_a are (necessarily spatial) vectors orthogonal to ℓ and n, and $\beta_a = \beta_a(\tau)$. This ambiguity does not affect any of the results in this section, so in principle one can make any choice of auxiliary n^α. In particular, one can choose n^α to be parallel-transported along ℓ^α,

$$\ell^\alpha\nabla_\alpha n^\beta = 0 \ . \tag{12.82}$$

Then the properties of parallel transport obtained in section 5.8 imply that the conditions (12.64) on n^α hold everywhere along the null geodesic (or congruence of null geodesics) if they are satisified initially. This reduces the ambiguity in (12.81) to τ-independent null rotations.

In fact, one can do even better than that and choose (see also the discussion at the end of section 17.4, in particular around (17.57)) an entire pseudo-orthonormal frame

$$\{E_A\} = \{E_+ = \ell, E_- = n, E_a\} : \quad g_{\alpha\beta} E^\alpha_A E^\beta_B = \eta_{AB} \tag{12.83}$$

where

$$\eta_{++} = \eta_{--} = 0 \quad , \quad \eta_{+-} = -1 \quad , \quad \eta_{a+} = \eta_{a-} = 0 \quad , \quad \eta_{ab} = \delta_{ab} \quad . \tag{12.84}$$

If one selects such a frame at one point along the geodesic and then parallel transports the frame along the geodesic, the orthogonality relations (12.83) will hold everywhere along the geodesic. Thus we can always choose a basis E_A such that

$$D_\tau E^\alpha_A = 0 \quad , \quad g_{\alpha\beta} E^\alpha_A E^\beta_B = \eta_{AB} \quad . \tag{12.85}$$

With this choice, a transverse geodesic deviation vector is simply one which has components only in the E_a-directions,

$$\xi^\alpha \ell_\alpha = \xi^\alpha n_\alpha = 0 \quad \Leftrightarrow \quad \xi^\alpha = \xi^a E^\alpha_a \quad , \tag{12.86}$$

or simply

$$\xi = \xi^a E_a \quad . \tag{12.87}$$

Since $D_\tau E_a$=0, one has

$$D_\tau E_a = 0 \quad \Rightarrow \quad D_\tau \xi^a = \frac{d}{d\tau}\xi^a \quad , \tag{12.88}$$

so that with respect to this basis covariant differentiation along the curve reduces to ordinary differentiation of the components. Then the null geodesic deviation equation (12.80) can simply be written as

$$\frac{d^2}{d\tau^2}\xi^a = R^a{}_{++b}\xi^b = -R^a{}_{+b+}\xi^b \quad , \tag{12.89}$$

where the R_{a+b+} are the frame components of the Riemann tensor,

$$R_{a+b+} = E^\alpha_a E^\mu_+ E^\beta_b E^\nu_+ R_{\alpha\mu\beta\nu} = E^\alpha_a E^\beta_b R_{\alpha\mu\beta\nu} \ell^\mu \ell^\nu \quad . \tag{12.90}$$

Thus the transverse null geodesic deviation equation has the form of a $(D-2)$-dimensional (transverse) harmonic oscillator equation,

$$\frac{d^2}{d\tau^2}\xi^a = -(\Omega^2)^a{}_b \xi^b \quad , \tag{12.91}$$

with the time-dependent symmetric frequency matrix

$$(\Omega^2)(\tau)^a{}_b = R^a{}_{+b+}(\tau) \quad . \tag{12.92}$$

The notation used here is perhaps suggestive but it is not meant to imply that Ω^2 is necessarily positive - the frequencies can be real or imaginary. Using the decomposition (11.75) of the Riemann tensor into its traceless and trace parts, we can (with $D = 4$, $\eta_{a+} = \eta_{++} = 0, \eta_{ab} = \delta_{ab}$) decompose R_{a+b+} as

$$R_{a+b+} = C_{a+b+} + \tfrac{1}{2}\delta_{ab}R_{++} \quad . \tag{12.93}$$

In particular, if the Ricci tensor is zero (as we will see this means that the metric solves the vacuum Einstein equations), the frequency matrix Ω^2 is symmetric traceless and thus necesarily has positive and negative eigenvalues (corresponding to real and imaginary frequencies).

12.4 RAYCHAUDHURI EQUATION FOR AFFINE NULL GEODESIC CONGRUENCES

We now consider a null geodesic congruence, with tangent vector field again denoted by ℓ^α, and we will initially choose these null geodesics to be affinely parametrised so that one has

$$\ell^\alpha \ell_\alpha = 0 \quad , \quad \ell^\alpha \nabla_\alpha \ell^\beta = 0 \quad . \tag{12.94}$$

We use the same framework as in the previous setion, with an auxiliary null vector field n^α with $\ell^\alpha n_\alpha = -1$, the associated projectors etc.

In (12.79) we introduced the transverse projection $b_{\alpha\beta}$ of the tensor $B_{\alpha\beta} = \nabla_\beta \ell_\alpha$,

$$b_{\alpha\beta} = s_\alpha^{\ \gamma} s_\beta^{\ \delta} B_{\gamma\delta} \quad . \tag{12.95}$$

Performing this projection explicitly, one sees that this spatial projection $b_{\alpha\beta}$ is equal to

$$b_{\alpha\beta} = s_\alpha^{\ \gamma} s_\beta^{\ \delta} B_{\gamma\delta} = B_{\alpha\beta} + \ell_\alpha n^\gamma B_{\gamma\beta} + \ell_\beta n^\delta B_{\alpha\delta} + \ell_\alpha \ell_\beta n^\gamma n^\delta B_{\gamma\delta} \quad . \tag{12.96}$$

This has two useful immediate consequences that we will make use of in the following, namely

- that the spatial trace of $b_{\alpha\beta}$ with respect to $s_{\alpha\beta}$ is equal to the space-time trace of $B_{\alpha\beta}$ (with respect to $g_{\alpha\beta}$),

$$g^{\alpha\beta} B_{\alpha\beta} = g^{\alpha\beta} b_{\alpha\beta} = s^{\alpha\beta} b_{\alpha\beta} \quad , \tag{12.97}$$

- and that the square of $b_{\alpha\beta}$ is identical to that of $B_{\alpha\beta}$,

$$B^{\alpha\beta} B_{\beta\alpha} = b^{\alpha\beta} b_{\beta\alpha} \quad . \tag{12.98}$$

We can now, as in the timelike case, decompose $b_{\alpha\beta}$ orthogonally into its irreducible (trace, symmetric traceless, anti-symmetric) parts,

$$\begin{aligned} b_{\alpha\beta} &= \tfrac{1}{2}\theta_\ell s_{\alpha\beta} + \tfrac{1}{2}(b_{\alpha\beta} + b_{\beta\alpha} - \theta_\ell s_{\alpha\beta}) + \tfrac{1}{2}(b_{\alpha\beta} - b_{\beta\alpha}) \\ &= \tfrac{1}{2}\theta_\ell s_{\alpha\beta} + \sigma_{\alpha\beta} + \omega_{\alpha\beta} \quad . \end{aligned} \tag{12.99}$$

Here θ_ℓ is the *expansion*

$$\theta_\ell = s^{\alpha\beta} b_{\alpha\beta} = s^{\alpha\beta}\nabla_\beta \ell_\alpha = g^{\alpha\beta}\nabla_\alpha \ell_\beta = \nabla^\alpha \ell_\alpha \ , \tag{12.100}$$

REMARKS:

1. As in the timelike case (12.38), the expansion θ_ℓ can be written as

   $$\theta_\ell = \tfrac{1}{2} s^{\alpha\beta} L_\ell s_{\alpha\beta} \ . \tag{12.101}$$

   and leads to an analogous interpretation of θ_ℓ as measuring the change in the cross-sectional area element $\sqrt{s}$ of the congruence (12.40),

   $$\theta_\ell = \frac{1}{\sqrt{s}} L_\ell \sqrt{s} \ . \tag{12.102}$$

2. The equivalence between the spatial and space-time traces of $\nabla_\alpha \ell_\beta$ in the above equation is due to the fact that we have chosen ℓ to be affinely parametrised. We will always define θ to be the spatial trace (divergence) of $\nabla_\alpha \ell_\beta$, even when ℓ is not affinely parametrised, but in that case θ and $\nabla_\alpha \ell^\alpha$ are no longer equal (see (12.126)). We will return to this issue below.

3. As regards the other terms, $\sigma_{\alpha\beta}$ and $\omega_{\alpha\beta}$ are again known as the *shear tensor* and *rotation tensor* respectively.

4. As in the timelike case, the rotation is zero if ℓ^α is hypersurface orthogonal. We will establish this result below.

5. Because the above decomposition is orthogonal, we have

   $$B^{\alpha\beta} B_{\beta\alpha} = b^{\alpha\beta} b_{\beta\alpha} = +\tfrac{1}{2}\theta_\ell^2 + \sigma^{\alpha\beta}\sigma_{\alpha\beta} - \omega^{\alpha\beta}\omega_{\alpha\beta} \ . \tag{12.103}$$

6. Because the tensors appearing on the right-hand side of this equation are spatial tensors, their squares are non-negative,

   $$\sigma^{\alpha\beta}\sigma_{\alpha\beta} \geq 0 \quad , \quad \omega^{\alpha\beta}\omega_{\alpha\beta} \geq 0 \ . \tag{12.104}$$

We now want to determine

$$\frac{d}{d\tau}\theta_\ell = \ell^\alpha \nabla_\alpha \theta_\ell \ . \tag{12.105}$$

We can do this either by again deriving an evolution equation for all of $b_{\alpha\beta}$, as in (12.71), or by calculating directly the derivative of θ_ℓ along ℓ. Adopting the latter procedure here, just following one's tensor calculus nose one finds

$$\begin{aligned}\frac{d}{d\tau}\theta_\ell &= \ell^\alpha \nabla_\alpha (\nabla_\beta \ell^\beta) \\ &= \ell^\alpha [\nabla_\alpha, \nabla_\beta] \ell^\beta + \ell^\alpha \nabla_\beta \nabla_\alpha \ell^\beta \\ &= -R_{\alpha\beta}\ell^\alpha \ell^\beta - (\nabla^\beta \ell^\alpha)\nabla_\alpha \ell_\beta + \nabla_\beta(\ell^\alpha \nabla_\alpha \ell^\beta) \ .\end{aligned} \tag{12.106}$$

The 2nd term is just $-B^{\alpha\beta}B_{\beta\alpha}$ and the 3rd term is zero because ℓ is geodesic. Thus one finds the *Raychaudhuri equation for null congruences*

$$\frac{d}{d\tau}\theta_\ell = -R_{\alpha\beta}\ell^\alpha\ell^\beta - \tfrac{1}{2}\theta_\ell^2 - \sigma^{\alpha\beta}\sigma_{\alpha\beta} + \omega^{\alpha\beta}\omega_{\alpha\beta} \ . \tag{12.107}$$

Using (12.102) in the form

$$\frac{d}{d\tau}\sqrt{s} = \theta_\ell\sqrt{s} \ , \tag{12.108}$$

we can also write this as an equation for the change in the expansion rate of the cross-sectional area $\sqrt{s}$ of the congruence. This leads to an additional $+\theta_\ell^2$ in the evolution equation, and thus flips the sign of the 2nd term of (12.107), resulting in

$$\frac{d^2}{d\tau^2}\sqrt{s} = \left(-R_{\alpha\beta}\ell^\alpha\ell^\beta + \tfrac{1}{2}\theta_\ell^2 - \sigma^{\alpha\beta}\sigma_{\alpha\beta} + \omega^{\alpha\beta}\omega_{\alpha\beta}\right)\sqrt{s} \ . \tag{12.109}$$

REMARKS:

1. If the geometry is such that

$$R_{\alpha\beta}\ell^\alpha\ell^\beta \geq 0 \tag{12.110}$$

(by the Einstein equations to be discussed in the section 19, this translates into a positivity condition on the energy-momentum tensor known as the *null energy condition*, cf. section 22.1), this gives a negative contribution to the right-hand side of the Raychaudhuri equation (indicating the focussing effect of gravity on lightrays).

2. If moreover the rotation $\omega_{\alpha\beta}$ is zero, for an affinely parametrised null congruence one has

$$\frac{d}{d\tau}\theta_\ell = -R_{\alpha\beta}\ell^\alpha\ell^\beta - \tfrac{1}{2}\theta_\ell^2 - \sigma^{\alpha\beta}\sigma_{\alpha\beta} \tag{12.111}$$

and therefore, in particular,

$$\frac{d}{d\tau}\theta_\ell \leq -\tfrac{1}{2}\theta_\ell^2 \leq 0 \ . \tag{12.112}$$

In Minkowski space-time (with zero curvature) we would have an equality instead of the first inequality sign, so (12.112) is the (reasonable) statement that the expansion is smaller/slower than in Minkowski space, i.e. that (reasonable) matter has the tendency to focus lightrays.

3. Analogously to the timelike case, (12.112) has the consequence that if one has an initially converging null congruence, $\theta_\ell(\tau_0) < 0$, then because of

$$\frac{d}{d\tau}\theta_\ell \leq -\tfrac{1}{2}\theta_\ell^2 \quad \Rightarrow \quad \frac{1}{\theta_\ell(\tau)} \geq \frac{1}{\theta_\ell(\tau_0)} + \frac{\tau-\tau_0}{2} \tag{12.113}$$

$1/\theta(\tau) \to 0_-$ or $\theta(\tau) \to -\infty$ at the latest at

$$\theta_\ell(\tau) \to -\infty \quad \text{for} \quad \tau \leq \tau_0 + 2/|\theta_\ell(\tau_0)| \tag{12.114}$$

(if the geodesics can be extended that far). As in the timelike case, this usually indicates the formation of a (harmless) *caustic* where these null geodesics cross.

4. E.g. in Minkowski space radially outgoing lightrays $\ell = \partial_v$, $v = t + r$, have expansion

$$\theta_\ell = \nabla_\alpha(\partial_v)^\alpha = \frac{1}{r^2}\partial_\alpha(r^2(\partial_t + \partial_r)^\alpha) = +\frac{2}{r} > 0 \ , \tag{12.115}$$

while radially ingoing lightrays $n = \partial_u$, $u = t - r$, have expansion

$$\theta_n = \nabla_\alpha(\partial_u)^\alpha = \frac{1}{r^2}\partial_\alpha(r^2(\partial_t - \partial_r)^\alpha) = -\frac{2}{r} < 0 \ , \tag{12.116}$$

reflecting the fact that outgoing lightrays expand while ingoing lightrays contract. Both expansions diverge as $r \to 0$, but this evidently does not indicate a pathology of Minkowski space-time, but only of the chosen congruences at the origin (where all the lightrays of the congruence meet).

Note also that e.g. θ_ℓ satisfies the Raychaudhuri equation

$$\frac{d}{d\tau}\theta_\ell = \ell^\alpha\nabla_\alpha\theta_\ell = \partial_r(+2/r) = -2/r^2 = -\tfrac{1}{2}\theta_\ell^2 \ , \tag{12.117}$$

as it should (no curvature, no shear, no rotation). With cross-sectional area $\sqrt{s} \sim r^2$, one also has

$$\frac{1}{\sqrt{s}}\frac{d^2}{d\tau^2}\sqrt{s} = \frac{1}{r^2}(\partial_r)^2 r^2 = \frac{2}{r^2} = +\tfrac{1}{2}\theta_\ell^2 \ , \tag{12.118}$$

illustrating the variant (12.109) of the Raychaudhuri equation: while $\theta_\ell \sim r^{-1} \to 0$ as $r \to \infty$, indicating that the cross-sectional spheres become flatter and flatter for large r, the cross-sectional area grows like $\sqrt{s} \sim r^2$, leading to an acceleration of its growth.

5. An argument similar to that in the timelike case shows that the rotation vanishes if (and locally only if, by Frobenius) ℓ^α is hypersurface orthogonal

$$\omega_{\alpha\beta} = 0 \quad \Leftrightarrow \quad \ell^\alpha \quad \text{hypersurface-orthogonal} \ . \tag{12.119}$$

In order to establish this we begin with the condition

$$\ell_{[\alpha}\nabla_\beta\ell_{\gamma]} = 0 \quad \Leftrightarrow \quad B_{[\alpha\beta]}\ell_\gamma + B_{[\beta\gamma]}\ell_\alpha + B_{[\gamma\alpha]}\ell_\beta = 0 \ . \tag{12.120}$$

Contracting this with ℓ^γ, we would get $0 = 0$, which is true but unenlightning. To get something non-trivial and more useful, we contract with n^γ instead. Then we deduce

$$B_{[\alpha\beta]} - B_{[\beta\gamma]}n^\gamma\ell_\alpha - B_{[\gamma\alpha]}n^\gamma\ell_\beta = 0 \ . \tag{12.121}$$

Disentangling this, one sees that this is precisely the statement that the anti-symmetric part of $b_{\alpha\beta}$ is zero, i.e. that $\omega_{\alpha\beta} = 0$, so we have established the desired result

$$\ell_{[\alpha}\nabla_\beta\ell_{\gamma]} = 0 \quad \Rightarrow \quad \omega_{\alpha\beta} = 0 \ . \tag{12.122}$$

6. The expansion properties of families of null geodesics play a crucial role both in the singularity theorems of general relativity (where for example so-called *trapped surfaces* are characterised by negative expansions for both ingoing and "outgoing" families of lightrays), and in the study of black holes and the laws governing the evolution of their event horizons (where the interest is in the null geodesic congruences generating the horizon). In particular, in the latter case the Raychaudhuri equation is one crucial ingredient in the proof of the statement (Hawking's theorem) that under reasonable conditions the cross-sectional area of the event horizon of a black hole cannot decrease.

12.5 RAYCHAUDHURI EQUATION FOR NON-AFFINELY PARAMETRISED NULL GEODESICS

Let us now look at the case when the null geodesic congruence is not affinely parametrised, i.e. when, instead of (12.94), the starting point is a null vector field ℓ^α satisfying

$$\ell^\alpha \ell_\alpha = 0 \quad , \quad \ell^\alpha \nabla_\alpha \ell^\beta = \kappa_\ell \ell^\beta \quad , \tag{12.123}$$

with κ_ℓ the inaffinity. Then a couple of things change in the derivation, but the end result (12.129) turns out to differ from (12.107) by only one term (and I will give an alternative and much quicker derivation of the result below).

As before, we can choose an auxiliary null vector field n^α, construct the projectors $s^\alpha_{\ \beta}$ etc. Defining again $B_{\alpha\beta} = \nabla_\beta \ell_\alpha$, one still has $\ell^\alpha B_{\alpha\beta} = 0$ (because this is implied by $\ell^\alpha \ell_\alpha = 0$), but instead of $B_{\alpha\beta}\ell^\beta = 0$ one now has

$$B_{\alpha\beta}\ell^\beta = \kappa_\ell \ell_\alpha \quad . \tag{12.124}$$

While the projection (12.96) remains unchanged, i.e. the relation between $b_{\alpha\beta}$ and $B_{\alpha\beta}$ has the same form as in (12.96), the equations (12.97) and (12.98) for the trace and square of $B_{\alpha\beta}$ differ. Instead of (12.97) one has

$$\begin{aligned} s^{\alpha\beta} b_{\alpha\beta} &= (g^{\alpha\beta} + \ell^\alpha n^\beta + \ell^\beta n^\alpha) B_{\alpha\beta} \\ &= \nabla_\alpha \ell^\alpha + \kappa_\ell n^\alpha \ell_\alpha = \nabla_\alpha \ell^\alpha - \kappa_\ell \quad . \end{aligned} \tag{12.125}$$

Thus, if we define the expansion θ_ℓ as the spatial trace of $b_{\alpha\beta}$, instead of (12.100) we find

$$\nabla_\alpha \ell^\alpha = \theta_\ell + \kappa_\ell \quad . \tag{12.126}$$

Analogously, for the square of $B_{\alpha\beta}$ one finds, instead of (12.98),

$$B^{\alpha\beta} B_{\beta\alpha} = b^{\alpha\beta} b_{\beta\alpha} + \kappa_\ell^2 \quad . \tag{12.127}$$

Putting everything together and calculating $(d/d\tau)\nabla_\alpha \ell^\alpha$ as in (12.106), one then finds

$$\begin{aligned}\frac{d}{d\tau}(\theta_\ell + \kappa_\ell) &= -R_{\alpha\beta}\ell^\alpha\ell^\beta - (\nabla^\beta\ell^\alpha)\nabla_\alpha\ell_\beta + \nabla_\beta(\ell^\alpha\nabla_\alpha\ell^\beta)\\ &= -R_{\alpha\beta}\ell^\alpha\ell^\beta - B_{\alpha\beta}B^{\beta\alpha} + \nabla_\beta(\kappa_\ell\ell^\beta)\\ &= -R_{\alpha\beta}\ell^\alpha\ell^\beta - b_{\alpha\beta}b^{\beta\alpha} - \kappa_\ell^2 + \frac{d}{d\tau}\kappa_\ell + \kappa_\ell(\theta_\ell + \kappa_\ell)\\ &= -R_{\alpha\beta}\ell^\alpha\ell^\beta - b_{\alpha\beta}b^{\beta\alpha} + \frac{d}{d\tau}\kappa_\ell + \kappa_\ell\theta_\ell \ .\end{aligned} \tag{12.128}$$

Thus the net effect of dealing with a non-affinely parametrised null congruence is that one just picks up one additional term on the right-hand side of the Raychaudhuri equation,

$$\frac{d}{d\tau}\theta_\ell = \kappa_\ell\theta_\ell - R_{\alpha\beta}\ell^\alpha\ell^\beta - \tfrac{1}{2}\theta_\ell^2 - \sigma^{\alpha\beta}\sigma_{\alpha\beta} + \omega^{\alpha\beta}\omega_{\alpha\beta} \ . \tag{12.129}$$

A quick(er) way to derive (12.129) is from the result (12.107) for affinely parametrised null geodesics, by determining how the quantities appearing in (12.107) change under a reparametrisation

$$\ell^\alpha \to \tilde{\ell}^\alpha = f\ell^\alpha \tag{12.130}$$

(with ℓ^α affinely parametrised, say).

- On the one hand, for the inaffinity one has

$$\tilde{\ell}^\alpha\nabla_\alpha\tilde{\ell}^\beta = \left(\frac{d}{d\tilde{\tau}}\log|f|\right)\tilde{\ell}^\beta \equiv \kappa_{\tilde{\ell}}\,\tilde{\ell}^\beta \ , \tag{12.131}$$

  where $\tilde{\tau}$ is the non-affine parameter along ℓ^α,

$$\frac{d}{d\tilde{\tau}}f = \tilde{\ell}^\alpha\nabla_\alpha f \tag{12.132}$$

- On the other hand for the expansion parameter etc one deduces from

$$\tilde{B}_{\alpha\beta} = \nabla_\beta\tilde{\ell}_\alpha = fB_{\alpha\beta} + \ell_\alpha\nabla_\beta f \tag{12.133}$$

  that the transverse projection is simply

$$\tilde{b}_{\alpha\beta} = fb_{\alpha\beta} \tag{12.134}$$

  which implies

$$\tilde{b}_{\alpha\beta} = fb_{\alpha\beta} \quad \Rightarrow \quad \left\{\begin{array}{ll}\theta_{\tilde{\ell}} & = f\theta_\ell\\ \tilde{\sigma}_{\alpha\beta} & = f\sigma_{\alpha\beta}\\ \tilde{\omega}_{\alpha\beta} & = f\omega_{\alpha\beta}\end{array}\right. \tag{12.135}$$

Plugging these results into (12.107) one obtains on the nose (12.129) (with $\tau \to \tilde{\tau}, \ell \to \tilde{\ell}$ etc.).

12.6 Expansions and Inaffinities of Radial Null Congruences

In this section, we look at some general properties of radial null congruences in a spherically symmetric space-time. All of the results of the previous sections 12.4 and 12.5 are of course also valid in this case, but the spherically symmetric case also has some special and simplifying features.

Thus we consider a spherically symmetric metric. Such a metric could always be written in the form

$$ds^2 = -A(t,r)dt^2 + B(t,r)dr^2 + r^2 d\Omega^2 \tag{12.136}$$

by a suitable choice of coordinates. However, we will not need to commit ourselves to this particular choice of coordinates. By making an arbitrary coordinate transformation

$$(t,r) \to z^a(t,r) \tag{12.137}$$

preserving the manifest spherical symmetry, this metric can be written in the form

$$ds^2 = g_{ab}(z)dz^a dz^b + r(z)^2 d\Omega^2 \tag{12.138}$$

for some 2-dimensional Lorentzian metric $g_{ab}(z)$, and with $r = r(z^a)$ now a function of the new coordinates.

We now consider two linearly independent radial and spherically symmetric null vector fields ℓ^α and n^α, which we choose to be cross normalised such that

$$\ell^\alpha \ell_\alpha = n^\alpha n_\alpha = 0 \quad , \quad \ell^\alpha n_\alpha = -1 \ . \tag{12.139}$$

Remarks:

1. Here "radial" means that it has components only in the ∂_{z^a}-directions transverse to the sphere, and "spherically symmetric" that the coefficients only depend on the z^a and not on the coordinates of the sphere (this can of course also, if desired, be phrased in a more coordinate-independent way, e.g. as the statement that the Lie derivatives of ℓ^α and n^α along the Killing vectors generating the rotational symmetry vanish, but for present purposes not much is gained by this).

2. In concrete applications we will choose n^α to be ingoing (in the sense that future directed null rays tangent to n^α will move towards smaller values of r) and ℓ^α to be (asymptotically) outgoing.

3. The minus sign in the cross normalisation is such that both vector fields are either future or past oriented (and we will of course choose the former).

4. Note that the individual normalisation of the ℓ^α and n^α is not fixed by the above conditions, i.e. one can still perform the boost

$$\ell^\beta \to \mathrm{e}^{+\alpha(x)}\ell^\beta \quad , \quad n^\beta \to \mathrm{e}^{-\alpha(x)}n^\beta \ . \tag{12.140}$$

> This can e.g. be used to select a preferred normalisation for one of them. If ℓ^α has been fixed, then, in spherical symmetry and with the assumption that n^α is also purely radial (longitudinal), n^α is uniquely determined by the 2 conditions $n^\alpha n_\alpha = 0$ and $n^\alpha \ell_\alpha = -1$. This should be contrasted with the situation without spherical symmetry where, as discussed in section 12.3, there is still the additional freedom to perform null rotations on n^α.

Spherical symmetry (and the choice of spherically symmetric null vector fields) also has other implications. For instance, it follows from spherical symmetry that $\ell^\alpha \nabla_\alpha \ell^\beta$ will be some linear combination of ℓ^β and n^β (i.e. no component tangent to the transverse sphere),

$$\ell^\alpha \nabla_\alpha \ell^\beta = A\ell^\beta + Bn^\beta \tag{12.141}$$

(and likewise for n^α). Taking the scalar product with ℓ_α and using

$$(\nabla_\alpha \ell^\beta)\ell_\beta = \tfrac{1}{2}\nabla_\alpha(\ell^\beta \ell_\beta) = 0 \ , \tag{12.142}$$

one finds that $B = 0$, so that automatically $\ell^\alpha \nabla_\alpha \ell^\beta \sim \ell^\beta$. Thus ℓ^α is automatically a geodesic null vector field, but perhaps not affinely parametrised. The proportionality constant A is then the *inaffinity* $A = \kappa_\ell$ (and likewise for n^α). Thus one has

$$\ell^\alpha \nabla_\alpha \ell^\beta = \kappa_\ell \ell^\beta \quad , \quad n^\alpha \nabla_\alpha n^\beta = \kappa_n n^\beta \ . \tag{12.143}$$

The boost freedom can then e.g. be used to choose either ℓ^α or n^α to be affinely parametrised (but usually not both of them simultaneously).

The same argument as above leads to the conclusion that necessarily $n^\alpha \nabla_\alpha \ell^\beta \sim \ell^\beta$ as well, and in this case the constant of proportionality is fixed by taking the scalar product with n_β (and likewise with $\ell \leftrightarrow n$), leading to the conclusions

$$\ell^\alpha \nabla_\alpha n^\beta = -\kappa_\ell n^\beta \quad , \quad n^\alpha \nabla_\alpha \ell^\beta = -\kappa_n \ell^\beta \ . \tag{12.144}$$

In particular, if $\kappa_\ell = 0$, say, this implies that n^α is parallel transported along ℓ, $\ell^\alpha \nabla_\alpha n^\beta = 0$.

It follows from either (12.143) or (12.144) that κ_ℓ and κ_n can be written as

$$\begin{aligned} \kappa_\ell &= -\ell^\alpha n^\beta \nabla_\alpha \ell_\beta = \ell^\alpha \ell^\beta \nabla_\alpha n_\beta \\ \kappa_n &= -n^\alpha \ell^\beta \nabla_\alpha n_\beta = n^\alpha n^\beta \nabla_\alpha \ell_\beta \ , \end{aligned} \tag{12.145}$$

or

$$\begin{aligned} \kappa_\ell &= \tfrac{1}{2}\ell^\alpha \ell^\beta (\nabla_\alpha n_\beta + \nabla_\beta n_\alpha) = \tfrac{1}{2}\ell^\alpha \ell^\beta L_n g_{\alpha\beta} \\ \kappa_n &= \tfrac{1}{2}n^\alpha n^\beta (\nabla_\alpha \ell_\beta + \nabla_\beta \ell_\alpha) = \tfrac{1}{2}n^\alpha n^\beta L_\ell g_{\alpha\beta} \end{aligned} \tag{12.146}$$

Here L_ℓ and L_n are the Lie derivatives. Thus the inaffinities encode the information about the longitudinal projections of the derivatives $\nabla_\alpha \ell_\beta$ and $\nabla_\alpha n_\beta$, or of the Lie derivatives $L_\ell g_{\alpha\beta}$ and $L_n g_{\alpha\beta}$.

Other useful information is contained in the transverse (i.e. parallel to the sphere) projections of these objects. To define them, note that, as in section 12.3, associated with a choice of ℓ^α and n^α we have the decomposition of the metric

$$g_{\alpha\beta} = s_{\alpha\beta} - (\ell_\alpha n_\beta + \ell_\beta n_\alpha) \tag{12.147}$$

with $s_{\alpha\beta}$ the transverse spatial metric (on the sphere),

$$s_{\alpha\beta} dx^\alpha dx^\beta = r(z)^2 d\Omega^2 \ \ , \tag{12.148}$$

but that in the current context this decomposition and the corresponding projectors $s^\alpha_{\ \beta}$ are now unique as the combination $\ell_\alpha n_\beta$ is boost-invariant.

The expansions of ℓ^α and n^α are defined as the transverse spatial projections of the divergence of ℓ^α respectively n^α, i.e.

$$\theta_\ell = s^{\alpha\beta}\nabla_\alpha \ell_\beta \quad , \quad \theta_n = s^{\alpha\beta}\nabla_\alpha n_\beta \ \ . \tag{12.149}$$

As in (12.101) and (12.102) of section 12.4, these can be written as

$$\begin{aligned} \theta_\ell &= \tfrac{1}{2} s^{\alpha\beta} L_\ell s_{\alpha\beta} = \frac{1}{\sqrt{s}} L_\ell \sqrt{s} \\ \theta_n &= \tfrac{1}{2} s^{\alpha\beta} L_n s_{\alpha\beta} = \frac{1}{\sqrt{s}} L_n \sqrt{s} \end{aligned} \tag{12.150}$$

With $\sqrt{s} = r(z)^2 \sin\theta$, one finds more explicitly

$$\theta_\ell = \frac{2}{r}\ell^\alpha \partial_\alpha r \quad , \quad \theta_n = \frac{2}{r} n^\alpha \partial_\alpha r \ \ . \tag{12.151}$$

If one works with r as one of the coordinates, then this can also succinctly be written as

$$\theta_\ell = \frac{2}{r}\ell^r \quad , \quad \theta_n = \frac{2}{r} n^r \ \ . \tag{12.152}$$

As in (12.126) of section 12.5, we also have the relations

$$\nabla^\alpha \ell_\alpha = \theta_\ell + \kappa_\ell \quad , \quad \nabla^\alpha n_\alpha = \theta_n + \kappa_n \ \ . \tag{12.153}$$

Turning now to the Raychaudhuri equation for a spherically symmetric radial null congruence ℓ, the general result (12.129) (for $\kappa_\ell \neq 0$), i.e.

$$\frac{d}{d\tau}\theta_\ell = \kappa_\ell \theta_\ell - R_{\alpha\beta}\ell^\alpha \ell^\beta - \tfrac{1}{2}\theta_\ell^2 - \sigma^{\alpha\beta}\sigma_{\alpha\beta} + \omega^{\alpha\beta}\omega_{\alpha\beta} \tag{12.154}$$

simplifies considerably. Spherical symmetry implies that the spatial shear and rotation tensors are zero (a spatial rotationally invariant 2-tensor is proportional to δ_{ik} which has neither a traceless nor an anti-symmetric part),

$$\sigma_{\alpha\beta} = \omega_{\alpha\beta} = 0 \ \ . \tag{12.155}$$

The vanishing of the rotation can also be deduced from the fact that ℓ is hypersurface orthogonal (specifically orthogonal to the family of null hypersurfaces generated by ℓ). Thus the Rauchaudhuri equation reduces to

$$\frac{d}{d\tau}\theta_\ell = \kappa_\ell \theta_\ell - R_{\alpha\beta}\ell^\alpha \ell^\beta - \tfrac{1}{2}\theta_\ell^2 \ \ . \tag{12.156}$$

13 CURVATURE IV: CURVATURE AND KILLING VECTORS

In (pseudo-)Riemannian geometry the rich interplay between symmetries and geometry is reflected in relations between the curvature tensor and Killing vectors of a metric. Here we will explore some of these relations and their consequences.

13.1 USEFUL IDENTITIES RELATING CURVATURE AND KILLING VECTORS

Using the defining relation of the Riemann curvature tensor,

$$(\nabla_\mu\nabla_\nu - \nabla_\nu\nabla_\mu)V_\lambda = -R^\rho{}_{\lambda\mu\nu}V_\rho \tag{13.1}$$

and its cyclic symmetry (8.18), it is possible to deduce that for a Killing vector K^μ,

$$\nabla_\mu K_\nu + \nabla_\nu K_\mu = 0 \ \ , \tag{13.2}$$

one has the following basic identity relating Killing vectors and the curvature tensor,

$$\nabla_\lambda\nabla_\mu K_\nu = R^\rho{}_{\lambda\mu\nu}\, K_\rho \ \ . \tag{13.3}$$

Indeed, proceeding as in the proof of the cyclic permutation identity (8.18), we deduce that

$$\nabla_{[\mu}\nabla_\nu K_{\lambda]} \sim R^\rho{}_{[\lambda\mu\nu]}K_\rho = 0 \ \ . \tag{13.4}$$

Since $\nabla_\nu K_\lambda$ is anti-symmetric, the total anti-symmetrisation is equivalent to cyclic permutation, and we therefore have

$$\nabla_\mu\nabla_\nu K_\lambda + \nabla_\nu\nabla_\lambda K_\mu + \nabla_\lambda\nabla_\mu K_\nu = 0 \ \ . \tag{13.5}$$

Using the Killing property in the second term, we can write this as

$$\nabla_\lambda\nabla_\mu K_\nu = -[\nabla_\mu, \nabla_\nu]K_\lambda = R^\rho{}_{\lambda\mu\nu}K_\rho \tag{13.6}$$

which is (13.3).

This identity can be interpreted as the statement (and can alternatively be derived from the fact) that the Lie derivative of the Christoffel symbols of a metric along a Killing vector of the metric is zero.

Indeed, first of all it is easy to see that under a general variation of the metric, the induced variation of the Christoffel symbol can be written as (20.14)

$$\delta\Gamma^\mu{}_{\nu\lambda} = \tfrac{1}{2}g^{\mu\rho}(\nabla_\nu\delta g_{\rho\lambda} + \nabla_\lambda\delta g_{\rho\nu} - \nabla_\rho\delta g_{\nu\lambda}) \ \ . \tag{13.7}$$

(this is easy to derive and also easy to remember as it takes exactly the same form as the definition of the Christoffel symbol, only with the metric replaced by the metric variation and the partial derivatives by covariant derivatives - see section 20.2 for a derivation and discussion of this identity). In particular, this exhibits the fact that the metric variation of the Christoffel symbols is a tensor (as could have been anticipated from the fact that the non-tensorial term in the transformation of the Christoffel symbols is independent of the metric), and additionally provides us with an explicit expression for this tensor.

Next, if the variation $\delta g_{\mu\nu} = L_\xi g_{\mu\nu}$ is the Lie derivative, i.e. the variation in the metric induced by an infinitesimal coordinate transformation $\delta x^\mu = \xi^\mu$, one can write this as

$$L_\xi \Gamma^\mu_{\ \nu\lambda} = \tfrac{1}{2} g^{\mu\rho}(\nabla_\nu L_\xi g_{\rho\lambda} + \nabla_\lambda L_\xi g_{\rho\nu} - \nabla_\rho L_\xi g_{\nu\lambda}) \ . \tag{13.8}$$

Note that in general the Lie derivative of a non-tensorial quantity is not well defined (or at least its definition requires a bit more thought). Here, however, it is natural to use the general formula (13.7) for the variation of the Christoffel symbols under metric variations to in particular define their Lie derivative (as the change in the Christoffel symbols induced by the Lie derivative of the metric).

Thus, adopting this definition and using $L_\xi g_{\mu\nu} = \nabla_\mu \xi_\nu + \nabla_\nu \xi_\nu$, the right-hand side can (upon using the definition and cyclic symmetry of the Riemann tensor) be written as

$$\begin{aligned} L_\xi \Gamma^\mu_{\ \nu\lambda} &= \nabla_\lambda \nabla_\nu \xi^\mu - R^\mu_{\ \nu\lambda\rho} \xi^\rho \\ &= \nabla_\nu \nabla_\lambda \xi^\mu + R^\mu_{\ \lambda\rho\nu} \xi^\rho \ . \end{aligned} \tag{13.9}$$

In particular, if $\xi^\mu = K^\mu$ is a Killing vector, one has

$$L_K g_{\mu\nu} = 0 \quad \Rightarrow \quad L_K \Gamma^\mu_{\ \nu\lambda} = 0 \quad \Leftrightarrow \quad \nabla_\lambda \nabla_\nu K_\mu = R_{\mu\nu\lambda\rho} K^\rho \ , \tag{13.10}$$

which is equivalent to (13.3).

Contracting (13.3) over λ and ν, one obtains the next useful and frequently used identity

$$\nabla^\nu \nabla_\mu K_\nu = K^\nu R_{\mu\nu} \ . \tag{13.11}$$

In turn, one immediate consequence of this identity is (contract with K^μ, "integrate by parts" and use the anti-symmetry of $\nabla_\mu K_\nu$)

$$R_{\mu\nu} K^\mu K^\nu = (\nabla^\mu K^\nu)(\nabla_\mu K_\nu) + \nabla_\nu (K^\mu \nabla_\mu K^\nu) \ . \tag{13.12}$$

Note that this can also be deduced directly from (8.53) for $V^\mu \to K^\mu$ a Killing vector.

We will now look at various consequences of the identities (13.3), (13.11) and (13.12) which are useful and interesting in their own right. The implications of these identities for *maximal symmetry* and *maximally symmetric spaces* will be discussed separately in section 14 below.

13.2 Killing Vectors form a Lie algebra

As the first application, we will explicitly prove the assertion (9.45) of section 9.5 that the Lie bracket of two Killing vectors is again a Killing vector. While this follows from the general property (9.34) of the Lie derivative, which itself can (with some work) be deduced from the general definition of the Lie derivative (as the generator of the action of coordinate transformations on tensors), it is instructive and reassuring to verify this by an explicit calculation, also because similar manipulations are required when extending the analysis from Killing vectors to Killing tensors or Killing-Yano tensors briefly mentioned in section 10.5.[30]

Thus consider two Killing vectors A^μ and B^μ, say, i.e. vector fields satisfying

$$\nabla_\mu A_\nu + \nabla_\nu A_\mu = \nabla_\mu B_\nu + \nabla_\nu B_\mu = 0 \tag{13.13}$$

or, equivalently,

$$\nabla_\mu A_\nu = \nabla_{[\mu} A_{\nu]} \quad , \quad \nabla_\mu B_\nu = \nabla_{[\mu} B_{\nu]} \ . \tag{13.14}$$

Explicitly, from (9.18) their Lie bracket is the vector field

$$C^\mu = [A,B]^\mu = A^\nu \nabla_\nu B^\mu - B^\nu \nabla_\nu A^\mu \ , \tag{13.15}$$

and the claim is that C^μ is also Killing, i.e. that $\nabla_\mu C_\nu$ is anti-symmetric,

$$C^\mu = [A,B]^\mu \quad \Rightarrow \quad \nabla_\mu C_\nu = \nabla_{[\mu} C_{\nu]} \ . \tag{13.16}$$

In calculating $\nabla_\mu C_\nu$ one encounters new first derivatives of A^μ and B^μ (which can be manipulated by the Killing equations), as well as second covariant derivatives, which can be reduced to zero-derivative terms by using the fundamental identity (13.3),

$$\begin{aligned}\nabla_\mu C_\nu &= (\nabla_\mu A_\lambda)\nabla^\lambda B_\nu - (\nabla_\mu B_\lambda)\nabla^\lambda A_\nu + A_\lambda \nabla_\mu \nabla^\lambda B_\nu - B_\lambda \nabla_\mu \nabla^\lambda A_\nu \\ &= -(\nabla_\lambda A_\mu)\nabla^\lambda B_\nu + (\nabla^\lambda B_\mu)\nabla_\lambda A_\nu + R_{\rho\mu\lambda\nu}(A^\lambda B^\rho - B^\lambda A^\rho) \ .\end{aligned} \tag{13.17}$$

The first two terms are already manifestly anti-symmetric (the second being the anti-symmetrisation of the first), and by the cyclic identity and other symmetries of the Riemann tensor, so is the last term,

$$R_{\rho\mu\lambda\nu} - R_{\lambda\mu\rho\nu} = R_{\mu\lambda\rho\nu} - R_{\mu\rho\lambda\nu} = R_{\mu\nu\lambda\rho} = -R_{\nu\mu\lambda\rho} \ . \tag{13.18}$$

Thus the Lie bracket of two Killing vectors is indeed again a Killing vector, as claimed.

---

[30] The interesting question if or when Killing-Yano tensors form a Lie algebra, extending and generalising the Lie algebra of the isometry group generated by the Killing vectors, is analysed in D. Kastor, S. Ray, J. Traschen, *Do Killing-Yano tensors form a Lie Algebra?*, `arXiv:0705.0535 [hep-th]`.

13.3 On the Isometry Algebra of a Compact Riemannian Space

In this section we will look at one immediate application of the identity (13.12),

$$R_{\mu\nu}K^\mu K^\nu = (\nabla^\mu K^\nu)(\nabla_\mu K_\nu) + \nabla_\nu(K^\mu \nabla_\mu K^\nu) \;\; , \tag{13.19}$$

namely an analogue of the Bochner-Yano type argument (given in remark 8 of section 8.5) for Killing vectors. Again, in order to be able to say something of substance we assume that the space we are dealing with is compact without boundary, and Riemannian, i.e. equipped with a positive-definite metric. In spite of this, the result we will derive is relevant also for physics, at least as long as one is willing to entertain the possibility that some higher-dimensional generalisations of general relativity (such as Kaluza-Klein theories discussed in section 44) plays a role in some more fundamental description of nature.

With the above assumptions, the first term on the right-hand side of (13.19) is non-negative and the second is a total derivative term that vanishes upon integration. Therefore for a Killing vector to exist on a compact Riemannian space, the integral of $R_{\mu\nu}V^\mu V^\nu$ must be non-negative as well.

This has two immediate implications:

1. If the Ricci tensor (regarded as a quadratic form) of a Riemannian metric on a compact space is negative, that metric can have no continuous isometries whatsoever.
2. If the metric on a compact Riemannian space has vanishing Ricci tensor, $R_{\mu\nu} = 0$, then any Killing vector is covariantly constant, $\nabla_\mu K_\nu = 0$.

Since the Lie bracket of two covariantly constant vector fields is zero,

$$\nabla_\mu V_\nu = \nabla_\mu W_\mu = 0 \quad \Rightarrow \quad [V,W]^\mu = V^\nu \nabla_\nu W^\mu - W^\nu \nabla_\nu V^\mu = 0 \;\; , \tag{13.20}$$

this means that continuous isometries of a space with vanishing Ricci tensor can at most be Abelian. An example is provided by the torus T^n equipped with the flat metric it inherits from regarding T^n as the periodic identification of $\mathbb{R}^n$. This metric has vanishing Ricci tensor (because evidently even the Riemann tensor is zero), but there are n linearly-independent (covariantly) constant translational Killing vectors (inherited from $\mathbb{R}^n$) that generate the Abelian isomtery group $U(1)^n$.

In Kaluza-Klein theory, one of the basic ideas is that gauge symmetries arise from isometries of the "internal" space living in the extra dimensions. This internal space is usually assumed to be compact (so as to be sufficiently small to have escaped our attention). Thus, if one wants to generate non-Abelian gauge theories in this way the above results provide one of the most basic constraints on the internal geometry, namely that the Ricci tensor should not be non-positive (but it does not have to be strictly positive everywhere).

13.4 Invariance of the Curvature along Killing Directions

It should be obvious and obviously true that for any Killing vector of a metric the scalar curvature of the metric inherits the corresponding symmetries of the metric, i.e. that it does not change along the orbits of that Killing vector,

$$\nabla_\mu K_\nu + \nabla_\nu K_\mu = 0 \quad \Rightarrow \quad K^\mu \nabla_\mu R = 0 \ , \tag{13.21}$$

or

$$L_K g_{\mu\nu} = 0 \quad \Rightarrow \quad L_K R = 0 \tag{13.22}$$

(with analogous statements for the Riemann and Ricci tensors).

While true, a covariant proof of this is a bit roundabout:

- One can start with the contracted Bianchi identity $\nabla^\mu G_{\mu\nu} = 0$, and contract it with K^ν to find

$$0 = (\nabla^\mu G_{\mu\nu}) K^\nu = (\nabla^\mu R_{\mu\nu}) K^\nu - \tfrac{1}{2} K^\nu \nabla_\nu R \ . \tag{13.23}$$

- Using the Killing equations, i.e. the anti-symmetry of $\nabla_\mu K_\nu$, and the symmetry of the Ricci tensor, one can write this as

$$K^\nu \nabla_\nu R = 2\nabla^\mu (K^\nu R_{\mu\nu}) \ . \tag{13.24}$$

- Using (13.11), one finally arrives at

$$K^\nu \nabla_\nu R = 2\nabla^\mu (\nabla^\nu \nabla_\mu K_\nu) = [\nabla^\mu, \nabla^\nu] \nabla_\mu K_\nu = 0 \tag{13.25}$$

  by (8.46).

Alternatively (and more quickly but somewhat less covariantly) one could have simply locally introduced an adapted coordinate system (9.59) in which $K = \partial_y$ and $\partial_y g_{\mu\nu} = 0$, to immediately deduce that then necessarily also $\partial_y R = 0$. However, on general grounds, and with an eye towards possible generalisations, it is always useful to have different arguments at one's disposal, in particular among them one which is covariant.

Analogously one can prove, either covariantly or non-covariantly (recommended in this case), that the Lie derivative of the Riemann tensor and the Ricci tensor along a Killing vector are zero,

$$L_K g_{\mu\nu} = 0 \quad \Rightarrow \quad L_K R_{\lambda\sigma\mu\nu} = 0 \quad , \quad L_K R_{\mu\nu} = 0 \ . \tag{13.26}$$

Since $\nabla_\mu K_\nu$ is anti-symmetric, one can write (13.11) more explicitly with the help of the formula (5.64) for the covariant divergence of an anti-symmetric tensor as

$$R^\mu_{\ \nu} K^\nu = \frac{1}{\sqrt{g}} \partial_\nu (\sqrt{g} \nabla^\mu K^\nu) \ . \tag{13.27}$$

This can be a quite efficient way to calculate certain components of the Ricci tensor of a metric, namely those which are of the form $R^\mu_{\ \nu} K^\nu$ for some Killing vector (the components referred to glibly as the "Killing components of the Ricci tensor" in the heading). In spite of this, this shortcut does not appear to be widely known or commonly used.

As an illustration of how this works, consider again the general static spherically symmetric metric (3.22),

$$ds^2 = -A(r)dt^2 + B(r)dr^2 + r^2 d\Omega^2 \ . \tag{13.28}$$

Among the Killing vectors of this metric is the vector field $\xi = \partial_t$ generating time-translations, and thus we can use (13.27) to determine the components $R^\mu_{\ t}$ of the Ricci tensor.

Since the only non-trivial component of ξ_μ is $\xi_t = -A(r)$, one has

$$\nabla_\mu \xi_\nu = \partial_\mu \xi_\nu - \Gamma^\lambda_{\ \mu\nu} \xi_\lambda = \partial_\mu \xi_\nu + A(r) \Gamma^t_{\ \mu\nu} \ , \tag{13.29}$$

and since according to (3.25) the only non-trivial component of $\Gamma^t_{\ \mu\nu}$ is

$$\Gamma^t_{\ rt} = \Gamma^t_{\ tr} = \frac{A'}{2A} \quad , \quad \Gamma^t_{\ \mu\nu} = 0 \quad \text{otherwise} \ , \tag{13.30}$$

the only non-trivial components of $\nabla_\mu \xi_\nu$ are

$$\nabla_t \xi_r = -\nabla_r \xi_t = A'/2 \quad , \quad \nabla_\mu \xi_\nu = 0 \quad \text{otherwise} \ . \tag{13.31}$$

Thus the only non-zero components of $\nabla^\mu \xi^\nu$ are

$$\nabla^r \xi^t = -\nabla^t \xi^r = A'/2AB \ . \tag{13.32}$$

With

$$\sqrt{g} = \sqrt{AB} r^2 \sin\theta \tag{13.33}$$

one then has

$$R^\mu_{\ t} = \frac{1}{\sqrt{AB} r^2} \partial_r (\sqrt{AB} r^2 \nabla^\mu \xi^r) \tag{13.34}$$

so that evidently

$$R^k_{\ t} = 0 \quad \text{with} \quad x^k = (r, \theta, \phi) \tag{13.35}$$

while

$$R^t{}_t = -\frac{1}{2\sqrt{AB}r^2}\partial_r(r^2A'/\sqrt{AB}) \quad . \tag{13.36}$$

Explicitly one can write this as

$$R_{tt} = -AR^t{}_t = \frac{A''}{2B} - \frac{A'}{4B}(\frac{A'}{A} + \frac{B'}{B}) + \frac{A'}{rB} \quad . \tag{13.37}$$

Remarks:

1. As you can check for yourself, this way of determining R_{tt} is much quicker than working it out from the general formula for the Ricci tensor involving the Christoffel symbols squared as well as their derivatives. In fact it is the quickest and slickest way to obtain R_{tt} by a calculation in coordinate components that I am aware of.
2. In the same way, one can also determine the angular components $R^\mu{}_\phi$, say, using the Killing vector $\eta = \partial_\phi$.
3. The only not obviously vanishing component of $R_{\mu\nu}$ (see the discussion in section 24.3) that cannot be obtained in this way is R_{rr}.

13.6 Killing Vectors as Solutions to the Maxwell Equations

A cute application of the identity (13.11) is the following. Recall that in the covariant Lorenz gauge

$$\nabla_\mu A^\mu = 0 \tag{13.38}$$

the vacuum Maxwell equations

$$\nabla^\mu F_{\mu\nu} = \nabla^\mu(\nabla_\mu A_\nu - \nabla_\nu A_\mu) = 0 \tag{13.39}$$

can be written as (8.115)

$$\nabla^\nu\nabla_\nu A^\mu = R^\mu_\nu A^\nu \quad , \tag{13.40}$$

while a Killing vector automatically satisfies

$$\nabla_\mu K_\nu + \nabla_\nu K_\mu = 0 \quad \Rightarrow \quad \nabla_\mu K^\mu = 0 \tag{13.41}$$

and the identity (13.11),

$$\nabla^\nu\nabla_\mu K^\nu = R_{\mu\nu}K^\nu \quad \Leftrightarrow \quad \nabla^\nu\nabla_\nu K^\mu = -R^\mu_\nu K^\nu \quad . \tag{13.42}$$

Thus the sign of the Ricci tensor in (13.40) and (13.42) is different, but evidently this difference disappears for a metric with vanishing Ricci tensor. This does not imply at all that the Riemann tensor is zero. Indeed, we will learn in section 19 that the vacuum Einstein equations (i.e. the gravitational field equations without or outside of matter sources) are simply the "Ricci flatness" conditions $R_{\mu\nu} = 0$ (19.36).

In that case (13.42) reduces to $\nabla^\nu \nabla_\nu K_\mu = 0$ which is of the same form as (13.40). Alternatively, because of the anti-symmetry of $\nabla_\mu K_\nu$, we can equivalently write (13.42) as

$$R_{\mu\nu} = 0 \quad \Rightarrow \quad \nabla^\mu (\nabla_\mu K_\nu - \nabla_\nu K_\mu) = 0 \ , \tag{13.43}$$

which, with the dictionary

$$A_\mu = K_\mu \quad , \quad F_{\mu\nu} = \nabla_\mu K_\mu - \nabla_\nu K_\mu \quad \Rightarrow \quad \nabla^\mu F_{\mu\nu} = 0 \ , \tag{13.44}$$

is identical to (13.39).

This means that any Killing vector of a solution to the vacuum Einstein equations automatically gives rise to a solution of the vacuum Maxwell equations in that gravitational background. Depending on the Killing vector this may or may not be a non-trivial ($F_{\mu\nu} \neq 0$) solution to the Maxwell equations, "rotational" Killing vectors typically giving rise to non-trivial solutions while for "translational" Killing vectors A_μ is pure gauge (see section 14.1 for a more precise characterisation of what is meant by "rotational" and "translational" Killing vectors at a given point).

For example, taking the general Killing vector (9.48) of Minkowski space (which certainly has vanishing Ricci tensor),

$$K^\alpha = \omega^\alpha_{\ \beta}\, x^\beta + \epsilon^\alpha \quad \Rightarrow \quad A_\alpha \equiv K_\alpha = \omega_{\alpha\beta}\, x^\beta + \epsilon_\alpha \ , \tag{13.45}$$

one finds that the associated Maxwell field strength tensor is

$$F_{\alpha\beta} = \partial_\alpha K_\beta - \partial_\beta K_\alpha = -2\omega_{\alpha\beta} \ . \tag{13.46}$$

Thus it vanishes for a purely translational Killing vector while a boost is associated with a constant electric field ($E_k \sim \omega_{0k}$) and a spatial rotation gives rise to a constant magnetic field ($B_k \sim \epsilon_{kij}\omega_{ij}$).

13.7 KILLING VECTORS AND KOMAR CURRENTS

Because the Einstein tensor $G_{\mu\nu}$ (8.97) is symmetric and conserved (the contracted Bianchi identity (8.96)), to any Killing vector one can associate (cf. the discussion in section 10.1) the conserved current

$$J_1^\mu = G^\mu_{\ \nu} K^\nu = R^\mu_{\ \nu} K^\nu - \tfrac{1}{2} R \delta^\mu_{\ \nu} K^\nu = R^\mu_{\ \nu} K^\nu - \tfrac{1}{2} R K^\mu \ . \tag{13.47}$$

However, because K^μ is Killing, one has $\nabla_\mu K^\mu = 0$ identically, as well as $K^\mu \nabla_\mu R = 0$ (as shown above), and hence one has a conserved current

$$J_a^\mu = R^\mu_{\ \nu} K^\nu - \tfrac{1}{2} a\, R K^\mu \quad , \quad \nabla_\mu J_a^\mu = 0 \ , \tag{13.48}$$

for any value of the real parameter a. Among this 1-parameter family of conserved currents, the choice

$$J_{a=0}^\mu \equiv J^\mu(K) = R^{\mu\nu} K_\nu \tag{13.49}$$

(the *Komar current*) is singled out by the fact that, by (13.11), it is not only conserved but can actually be written as the divergence of an anti-symmetric tensor,

$$J^{\mu}(K) = \nabla_{\nu} A^{\mu\nu} \quad , \quad A^{\mu\nu} = -A^{\nu\mu} = \nabla^{\mu} K^{\nu} \tag{13.50}$$

(which also shows directly, by (8.47), that $\nabla_{\mu} J^{\mu}(K) = 0$).

Thus the corresponding conserved charge, written as a hypersurface integral, can actually be written as a *surface integral* of components of $A_{\mu\nu}$. These define the so-called *Komar charges* associated to symmetries of the metric. They will make a brief appearance in section 23.4.

As an aside, note that while in the above we started off with Killing vectors, a similar story is actually true for any vector field. Namely, for any vector field ξ^{μ} define the current

$$J^{\mu}(\xi) = \nabla_{\nu}(\nabla^{[\mu}\xi^{\nu]}) = \tfrac{1}{2}\nabla_{\nu}(\nabla^{\mu}\xi^{\nu} - \nabla^{\nu}\xi^{\mu}) \ . \tag{13.51}$$

Note that this reduces to (13.50) for $\xi^{\mu} = K^{\mu}$ a Killing vector. Moreover, by (8.47) this current is conserved,

$$\nabla_{\mu} J^{\mu}(\xi) = 0 \quad \forall\, \xi^{\mu} \ . \tag{13.52}$$

When $\xi^{\mu} = K^{\mu}$ is a Killing vector, the current can alternatively be written in terms of the Ricci tensor as in (13.49). For a general ξ^{μ} one has, instead,

$$\begin{aligned} J^{\mu}(\xi) &= \tfrac{1}{2}\nabla_{\nu}(\nabla^{\mu}\xi^{\nu} - \nabla^{\nu}\xi^{\mu}) = \nabla_{\nu}\nabla^{\mu}\xi^{\nu} - \tfrac{1}{2}\nabla_{\nu}(\nabla^{\nu}\xi^{\mu} + \nabla^{\mu}\xi^{\nu}) \\ &= [\nabla^{\nu}, \nabla^{\mu}]\xi_{\nu} + \nabla^{\mu}(\nabla_{\nu}\xi^{\nu}) - \tfrac{1}{2}\nabla_{\nu}(\nabla^{\nu}\xi^{\mu} + \nabla^{\mu}\xi^{\nu}) \\ &= R^{\mu}_{\ \nu}\xi^{\nu} + \tfrac{1}{2}(g^{\alpha\beta}g^{\mu\nu} - g^{\mu\alpha}g^{\nu\beta})\nabla_{\nu}(\nabla_{\alpha}\xi_{\beta} + \nabla_{\beta}\xi_{\alpha}) \ , \end{aligned} \tag{13.53}$$

where we made use of (8.40). Note that this indeed reduces to (13.49) for a Killing vector, for which the second term on the right-hand side is absent.

The existence of these identically conserved currents and the corresponding surface charge densities $\nabla^{[\mu}\xi^{\nu]}$ reflects the fact that in general relativity (more generally, in any generally covariant theory) all vector fields can be considered as the generators of symmetries (in the sense of coordinate transformations). Indeed, the currents $J^{\mu}(\xi)$ can be shown to be precisely the corresponding Noether currents arising from the Lagrangian formulation of general relativity to be discussed in section 20. We will establish this result in section 20.6. Nevertheless, the currents and charges associated with Killing vectors turn out to play a privileged role, and we will in particular relate the Komar charge for a timelike Killing vector to the ADM mass of an isolated (asymptotically) static system in section 23.4.

14 CURVATURE V: MAXIMAL SYMMETRY AND CONSTANT CURVATURE

As a preparation for our discussion of cosmology in sections 33 - 38, in this section we will discuss some aspects of what are known as *maximally symmetric spaces*. These are spaces that admit the maximal number of Killing vectors (which turns out to be $n(n+1)/2$ for an n-dimensional space or space-time).

As we will discuss later on, in the context of the *Cosmological Principle*, such spaces, which are simultaneously homogeneous ("the same at every point") and isotropic ("the same in every direction") provide an (admittedly highly idealised) description of space in a cosmological space-time.

If you already know (or are willing to believe) that in any spatial dimension n there are essentially only 3 such spaces, namely the Euclidean space $\mathbb{R}^n$, the sphere S^n, and its negative curvature counterpart, the hyperbolic space H^n (all equipped with their standard metrics), you can skip this section, and may just want to refer to section 14.3 where it is shown that these 3 standard metrics can be written in a unified way as

$$ds^2 = \frac{dr^2}{1-kr^2} + r^2 d\Omega^2_{n-1} \tag{14.1}$$

for $k = 0, \pm 1$ respectively.

The discussion of maximally symmetric spaces and, in particular, space-times will be taken up again and continued in section 39 (which can also be read as a direct sequel to this section, without the intervening sections 33 - 38 on cosmology).

14.1 HOMOGENEOUS, ISOTROPIC AND MAXIMALLY SYMMETRIC SPACES

In order to understand how to define and characterise maximally symmetric spaces, we will need to obtain some more information about how Killing vectors can be classified.

Our starting point is, as in the previous section, the identity (13.3), reproduced here with the explicit x-dependence included for present purposes,

$$\nabla_\lambda \nabla_\mu K_\nu(x) = R^\rho{}_{\lambda\mu\nu}(x) K_\rho(x) \quad . \tag{14.2}$$

In particular, this shows that the second derivatives of the Killing vector at a point x_0 are again expressed in terms of the value of the Killing vector itself at that point. This means (think of Taylor expansions) that, remarkably, a Killing vector field $K^\mu(x)$ is completely and uniquely determined everywhere by the values of $K_\mu(x_0)$ and $\nabla_\mu K_\nu(x_0)$ at a single point x_0.

A set of Killing vectors $\{K^{(i)}_\mu(x)\}$ is said to be *linearly independent* if any linear relation of the form

$$\sum_i c_i K^{(i)}_\mu(x) = 0 \quad , \tag{14.3}$$

with *constant* coefficients c_i implies $c_i = 0$ (the reason for insisting on constant coefficients rather than functions $c_i(x)$ in this definition is of course that if K^μ is a Killing vector, then so is cK^μ iff c is constant).

Since, in an n-dimensional space(-time) there can be at most n linearly independent vectors ($K_\mu(x_0)$) at a point, and at most $n(n-1)/2$ independent anti-symmetric matrices ($\nabla_\mu K_\nu(x_0)$), we reach the conclusion that an n-dimensional space(-time) can have at most

$$n + \frac{n(n-1)}{2} = \frac{n(n+1)}{2} \tag{14.4}$$

independent Killing vectors. A space(-time) with this maximal number of Killing vectors is called *maximally symmetric*.

An example of a metric with the maximal number of Killing vectors is, none too surprisingly, n-dimensional Minkowski space, where $n(n+1)/2$ agrees with the dimension of the Poincaré group, the group of transformations that leave the Minkowski metric invariant.

Other examples of spaces that are maximally symmetric spaces are provided by spheres with their standard metric (e.g. we already know that the 2-sphere has $3 = 2(2+1)/2$ linearly independent Killing vectors, given explicitly in (9.55)). We will show below that spheres and their negative curvature hyperbolic counterparts are the unique non-trivial maximally symmetric spaces (with a corresponding statement for maximally symmetric space-times, which we will study in detail in section 39).

We will now see how the data $K^\mu(x_0)$ and $\nabla_\mu K_\nu(x_0)$ are related to translations and rotations:

- We define a *homogeneous space* to be such that it has infinitesimal isometries that carry any given point x_0 into any other point in its immediate neighbourhood (this could be stated in more fancy terms!). Thus the metric must admit Killing vectors that, at any given point, can take all possible values. Thus we require the existence of Killing vectors for arbitrary $K_\mu(x_0)$. This means that the n-dimensional space admits the maximal number n of *translational Killing vectors.*
- We define a space to be *isotropic at a point* x_0 if it has isometries that leave the given point x_0 fixed and such that they can rotate any vector at x_0 into any other vector at x_0. Therefore the metric must admit Killing vectors such that $K_\mu(x_0) = 0$ but such that $\nabla_\mu K_\nu(x_0)$ is an arbitrary anti-symmetric matrix (for instance to be thought of as an element of the Lie algebra of $SO(n)$). This means that the n-dimensional space admits the maximal number $n(n-1)/2$ of *rotational Killing vectors.*
- Finally, we define a *maximally symmetric space* to be a space with a metric with the maximal number $n(n+1)/2$ of Killing vectors.

For example, as mentioned before, the 2-sphere is maximally symmetric, with 3 linearly independent Killing vectors, given explicitly e.g. in (9.55). The decomposition of these 3 Killing vectors into 1 rotational and 2 translational Killing vectors depends on the point on the 2-sphere, the rotational Killing vector always being associated with the rotations around the axis through that point, and the translational Killing vectors being formed by the remaining 2 linearly independent combinations of Killing vectors. The decomposition given in (9.55) is adapted to rotations around the north (or south) pole, with $V_{(3)} = \partial_\phi$ the corresponding rotational Killing vector. Note that this Killing vector acts as a rotation at / around the poles but that it acts as a translation away from the poles (where some other linear combination of the 3 Killing vectors would be the rotational Killing vector). We will come back to this in slightly more general terms below.

Some simple and fairly obvious consequences of these definitions are the following:

1. A homogeneous and isotropic space is maximally symmetric.
2. A space that is isotropic for all x is also homogeneous.
3. (1) and (2) now imply that a space which is isotropic around every point is maximally symmetric.
4. Finally one also has the converse, namely that a maximally symmetric space is homogeneous and isotropic.

Property (2) is a consequence of the fact that constant linear combinations of Killing vectors are again Killing vectors and that, as mentioned above in the context of the 2-sphere, away from the origin of the rotation a rotation acts just like a translation. Technically, the difference between two rotational Killing vectors at x and $x + dx$ can be shown to be a translational Killing vector. To see this (roughly), consider 2 Killing vectors K and L describing rotations about a point x_0 and a point $x_0 + dx$ respectively, i.e.

$$K^\mu(x_0) = 0 \quad , \quad L^\mu(x_0 + dx) = 0 \ . \tag{14.5}$$

and, in particular,

$$(\nabla_\mu L_\nu)(x_0 + dx) \neq 0 \ . \tag{14.6}$$

Now consider the difference

$$M^\mu(x) = K^\mu(x) - L^\mu(x) \tag{14.7}$$

which is still a Killing vector. At $x = x_0$ one has

$$M^\mu(x_0) = K^\mu(x_0) - L^\mu(x_0) = -L^\mu(x_0 + dx - dx) \ . \tag{14.8}$$

Now expanding $L^\mu(x)$ around the point $x+dx$, one has (in an inertial coordinate system at x_0, say)

$$M_\mu(x_0) = dx^\lambda \nabla_\lambda L_\mu(x_0 + dx) \neq 0 \tag{14.9}$$

while its matrix of covariant derivatives ∇M vanishes there due to the crucial identity $\nabla\nabla L \sim L$ (14.2). Thus M defines a translational Killing vector at x_0.

In practice the characterisation of a maximally symmetric space which is easiest to use is (3) because it requires consideration of only one type of symmetries, namely rotational symmetries.

14.2 CURVATURE TENSOR OF A MAXIMALLY SYMMETRIC SPACE

On the basis of these simple considerations we can already determine the form of the Riemann curvature tensor of a maximally symmetric space. We will see that maximally symmetric spaces are spaces of constant curvature in the sense that the Riemann curvature tensor is simply and purely algebraically related to the metric by

$$R_{ijkl} = k(g_{ik}g_{jl} - g_{il}g_{jk}) \tag{14.10}$$

for some constant k.

This result could be obtained by making systematic use of the higher order integrability conditions for the existence of a maximal number of Killing vectors. The argument given below is less covariant but more elementary.

Assume for starters that the space is isotropic at x_0 and choose a Riemann normal coordinate system centered at x_0. Thus the metric at x_0 is $g_{ij}(x_0) = \eta_{ij}$ where we may just as well be completely general and assume that

$$\eta_{ij} = \text{diag}(\underbrace{-1,\ldots,-1}_{p \text{ times}}\underbrace{+1,\ldots,+1}_{q \text{ times}}) \quad , \tag{14.11}$$

where $p + q = n$ and we only assume $n > 2$.

If the metric is supposed to be isotropic at x_0 then, in particular, the curvature tensor at the origin must be invariant under Lorentz rotations. Now we know (i.e. you should know from your Special Relativity course) that the only invariants of the Lorentz group are the Minkowski metric and products thereof, and the totally anti-symmetric Levi-Civita tensor. Thus the Riemann curvature tensor has to be of the form

$$R_{ijkl}(x_0) = a\eta_{ij}\eta_{kl} + b\eta_{ik}\eta_{jl} + c\eta_{il}\eta_{jk} + d\epsilon_{ijkl} \quad , \tag{14.12}$$

where the last term is only possible for $D = 4$. The symmetries of the Riemann tensor imply that $a = d = b + c = 0$, and hence we are left with

$$R_{ijkl}(x_0) = b(\eta_{ik}\eta_{jl} - \eta_{il}\eta_{jk}) \quad , \tag{14.13}$$

Thus in an arbitrary coordinate system we will have

$$R_{ijkl}(x_0) = b(g_{ik}(x_0)g_{jl}(x_0) - g_{il}(x_0)g_{jk}(x_0)) \ , \tag{14.14}$$

If we now assume that the space is isotropic around every point, then we can deduce that

$$R_{ijkl}(x) = b(x)(g_{ik}(x)g_{jl}(x) - g_{il}(x)g_{jk}(x)) \tag{14.15}$$

for some function $b(x)$. Therefore the Ricci tensor and the Ricci scalar are

$$\begin{aligned} R_{ij}(x) &= (n-1)b(x)g_{ij} \\ R(x) &= n(n-1)b(x) \ . \end{aligned} \tag{14.16}$$

and the Riemann curvature tensor can also be written as

$$R_{ijkl} = \frac{R}{n(n-1)}(g_{ik}g_{jl} - g_{il}g_{jk}) \ , \tag{14.17}$$

while the Einstein tensor is

$$G_{ij} = b[(n-1)(1-n/2)]g_{ij} \ . \tag{14.18}$$

For $n > 2$ the contracted Bianchi identity $\nabla^i G_{ij} = 0$ now implies that $b(x)$ has to be a constant, and we have thus established (14.10). Note that we also have

$$R_{ij} = k(n-1)g_{ij} \ , \tag{14.19}$$

so that a maximally symmetric space(-time) is automatically a solution to the vacuum Einstein equations with a cosmological constant. In the physically relevant case $p = 1$ these are known as de Sitter or anti de Sitter space-times. We will come back to them in detail later on, in section 39.

14.3 MAXIMALLY SYMMETRIC METRICS I: SOLVING THE CONSTANT CURVATURE CONDITIONS

We are interested not just in the curvature tensor of a maximally symmetric space but in the metric itself. I will give you two derivations of the metric of a maximally symmetric space, one by directly solving the differential equation

$$R_{ij} = k(n-1)g_{ij} \tag{14.20}$$

for the metric g_{ij}, the other by a direct geometrical construction of the metric which makes the isometries of the metric manifest.

As a maximally symmetric space is in particular spherically symmetric, we can write its metric in the form

$$ds^2 = B(r)dr^2 + r^2 d\Omega^2_{(n-1)} \ , \tag{14.21}$$

where $d\Omega^2_{(n-1)} = d\theta^2 + \dots$ is the volume-element for the $(n-1)$-dimensional sphere or its counterpart in other signatures. For concreteness, we now fix on $n = 3$, but the argument given below goes through in general.

It is straighforward to calculate the components of the Ricci tensor of this metric. This can be viewed as a special case of the calculations leading to the Schwarzschild metric in section 24, setting the function called $A(r)$ there to zero (of course before having divided by it anywhere ...).

The only independent components are R_{rr} and $R_{\theta\theta}$,

$$\begin{aligned} R_{rr} &= \frac{1}{r}\frac{B'}{B} \\ R_{\theta\theta} &= -\frac{1}{B} + 1 + \frac{rB'}{2B^2} \ . \end{aligned} \tag{14.22}$$

We now want to solve the equations

$$\begin{aligned} R_{rr} &= 2kg_{rr} = 2kB(r) \\ R_{\theta\theta} &= 2kg_{\theta\theta} = 2kr^2 \ . \end{aligned} \tag{14.23}$$

From the first equation we obtain

$$B' = 2krB^2 \ , \tag{14.24}$$

and from the second equation we deduce

$$\begin{aligned} 2kr^2 &= -\frac{1}{B} + 1 + \frac{rB'}{2B^2} \\ &= -\frac{1}{B} + 1 + \frac{2kr^2B^2}{2B^2} \\ &= -\frac{1}{B} + 1 + kr^2 \ . \end{aligned} \tag{14.25}$$

This is an algebraic equation for B solved by

$$B = \frac{1}{1-kr^2} \tag{14.26}$$

(and this also solves the first equation). Therefore we have determined the metric of a a maximally symmetric space to be

$$ds^2 = \frac{dr^2}{1-kr^2} + r^2 d\Omega^2_{(n-1)} \ . \tag{14.27}$$

Clearly, for $k = 0$ this is just the flat metric on $\mathbb{R}^n$. For $k = 1$, this should also look familiar as the standard metric on the sphere. If not, don't worry, we will be more explicit about this below.

We will also rederive these metrics in the next section in a way that makes the isometries of the metric manifest (and which thus also excludes the possibility, not logically ruled out by the arguments given so far, that the metrics we have found here for $k \neq 0$ are spherically symmetric and have constant Ricci curvature but are not actually maximally symmetric).

REMARKS:

1. First of all let us note that for $k \neq 0$ essentially only the sign of k matters as $|k|$ only affects the overall size of the space and nothing else (and can therefore be absorbed in the scale factor $a(t)$ of the metric (33.1) that will be the starting point for our investigations of cosmology). To see this note that a metric of the form (14.27), but with k replaced by k/L^2,
$$ds^2 = \frac{dr^2}{1-kr^2/L^2} + r^2 d\Omega^2 \ , \tag{14.28}$$
can, by introducing $\tilde{r} = r/L$, be put into the form
$$ds^2 = \frac{dr^2}{1-kr^2/L^2} + r^2 d\Omega^2 = L^2 \left(\frac{d\tilde{r}^2}{1-k\tilde{r}^2} + \tilde{r}^2 d\Omega^2 \right) \ . \tag{14.29}$$
We now see explicitly that a rescaling of k by a constant factor is equivalent to an overall rescaling of the metric, and thus we will just need to consider the cases $k = 0, \pm 1$. However, occasionally it will also be convenient to think of k as a continuous parameter, the 3 geometries then being distinguished by $k < 0, k = 0$ and $k > 0$ respectively.

2. For $k = +1$, we have
$$ds^2 = \frac{dr^2}{1-r^2} + r^2 d\Omega^2_{(n-1)} \ . \tag{14.30}$$
Thus, obviously the range of r is restricted to $r \leq 1$ and by the change of variables $r = \sin\psi$, the metric can be put into the standard form of the metric on S^n in polar coordinates,
$$ds^2 \equiv d\Omega^2_n = d\psi^2 + \sin^2\psi d\Omega^2_{n-1} \ . \tag{14.31}$$
This makes it clear that the singularity at $r = 1$ is just a coordinate singularity.

3. For $k = -1$, on the other hand, we have
$$ds^2 = \frac{dr^2}{1+r^2} + r^2 d\Omega^2_{n-1} \ . \tag{14.32}$$
Thus the range of r is $0 \leq r < \infty$, and we can use the change of variables $r = \sinh\psi$ to write the metric as
$$ds^2 \equiv d\tilde{\Omega}^2_n = d\psi^2 + \sinh^2\psi d\Omega^2_{n-1} \ . \tag{14.33}$$
This is the standard metric of a hyperboloid H^n in polar coordinates, and I have introduced the notation $d\tilde{\Omega}^2_n$ for the line-element on the "unit" n-hyperboloid in analogy with the standard notation $d\Omega^2_n$ for the line-element on the unit n-sphere.

4. Thus, collectively we can write the three metrics as

$$ds^2 = \frac{dr^2}{1-kr^2} + r^2 d\Omega^2_{(n-1)} = d\psi^2 + g_k(\psi)^2 d\Omega^2_{n-1} \tag{14.34}$$

where

$$g_k(\psi) = \begin{cases} \psi & k=0 \\ \sin\psi & k=+1 \\ \sinh\psi & k=-1 \end{cases} \tag{14.35}$$

5. Finally, by making the change of variables

$$r = \bar{r}(1+k\bar{r}^2/4)^{-1} \;\; , \tag{14.36}$$

one can put the metric into the isotropic form

$$ds^2 = (1+k\bar{r}^2/4)^{-2}(d\bar{r}^2 + \bar{r}^2 d\Omega^2_{(n-1)}) = (1+k\vec{x}^2/4)^{-2}\, d\vec{x}^2 \;\; . \tag{14.37}$$

Note that this differs by the *conformal factor* $(1+k\bar{r}^2/4)^{-2} > 0$ from the flat metric. One says that such a metric is *conformally flat*. Thus what we have shown is that every maximally symmetric space is conformally flat. Conformally flat, on the other hand, does not by any means imply maximally symmetric (the conformal factor could be any function of the radial and angular variables).

Note also that the metric in this form is just the 3- (or n-) dimensional generalisation of the 2-dimensional constant curvature metric on the 2-sphere in stereographic coordinates (11.62) (for $k=+1$) or of the Poincaré disc metric of H^2 (11.66) (for $k=-1$).

14.4 MAXIMALLY SYMMETRIC METRICS II: EMBEDDINGS

Recall that the standard metric on the n-sphere can be obtained by restricting the flat metric on an ambient $\mathbb{R}^{n+1}$ to the sphere. We will generalise this construction a bit to allow for $k<0$ and other signatures as well.

Consider a flat auxiliary vector space V of dimension $(n+1)$ with metric

$$ds^2 = d\vec{x}^2 + \frac{1}{k}dz^2 \;\; , \tag{14.38}$$

where $\vec{x} = (x^1,\ldots,x^n)$ and $d\vec{x}^2 = \eta_{ij}dx^i dx^j$. Thus the metric on V has signature $(p,q+1)$ for k positive and $(p+1,q)$ for k negative. The group $G = SO(p,q+1)$ or $G = SO(p+1,q)$ has a natural action on V by isometries of the metric. The full isometry group of V is the semi-direct product of this group with the Abelian group of translations (just as in the case of the Euclidean or Poincaré group).

Now consider in V the hypersurface Σ defined by

$$k\vec{x}^2 + z^2 = 1 \;\; . \tag{14.39}$$

This equation breaks all the translational isometries, but by the very definition of the group G it leaves this equation, and therefore the hypersurface Σ, invariant. It follows that G will act by isometries on Σ with its induced metric. Since $\dim G = n(n+1)/2$, the n-dimensional space has $n(n+1)/2$ Killing vectors and is therefore maximally symmetric.

REMARKS:

1. In fact, G acts transitively on Σ (thus Σ is homogeneous) and the stabiliser at a given point is isomorphic to $H = SO(p, q)$ (so Σ is isotropic), and therefore Σ can also be described as the homogeneous space
$$\begin{aligned} \Sigma_{k>0} &= SO(p, q+1)/SO(p, q) \\ \Sigma_{k<0} &= SO(p+1, q)/SO(p, q) \ . \end{aligned} \tag{14.40}$$
2. The Killing vectors of the induced metric are simply the restriction to Σ of the standard generators of G on the vector space V.
3. For Euclidean signature, these spaces are spheres for $k > 0$ and hyperboloids for $k < 0$, and in other signatures they are the corresponding generalisations. In particular, for $(p, q) = (1, n-1)$ we obtain *de Sitter space-time* for $k = 1$ and *anti-de Sitter space-time* for $k = -1$. We will discuss their embeddings, and coordinate systems for them, in much more detail in section 39.

It just remains to determine explicitly this induced metric. For this we start with the defining relation of Σ and differentiate it to find that on Σ one has
$$dz = -\frac{k\vec{x}.d\vec{x}}{z} \ , \tag{14.41}$$
so that
$$dz^2 = \frac{k^2(\vec{x}.d\vec{x})^2}{1 - k\vec{x}^2} \ . \tag{14.42}$$
Thus the metric (14.38) restricted to Σ is
$$\begin{aligned} ds^2|_\Sigma &= d\vec{x}^2 + \frac{1}{k}dz^2|_\Sigma \\ &= d\vec{x}^2 + \frac{k(\vec{x}.d\vec{x})^2}{1 - k\vec{x}^2} \ . \end{aligned} \tag{14.43}$$
Passing from Cartesian coordinates $\vec{x}$ to spherical coordinates (r, θ, ϕ), with
$$r^2 = \vec{x}^2 = \eta_{ij}x^i x^j \quad , \quad r dr = \vec{x}.d\vec{x} \quad , \quad dr^2 = (\vec{x}.d\vec{x})^2/\vec{x}^2 \ , \tag{14.44}$$
this metric can also be written as
$$ds^2 = \frac{dr^2}{1 - kr^2} + r^2 d\Omega^2_{(n-1)} \ . \tag{14.45}$$
This is precisely the metric (14.27) we obtained in the previous section.

data 04.03.2020

General Relativity
Tilman Sauer

https://cds.cern.ch/record/
735319/files/0405066.pdf

Source: Books_46_2020/
Importantissimo_Gauss/
0405066.pdf

others:

I) Einstein's Pathway to General Relativity

Source: https://www.pitt.edu/
~jdnorton/teaching/
HPS_0410/chapters/
general_relativity_pathway/
index.html

II) Einstein's Zurich Notebook

Source: www.pitt.edu/~jdnorton/
Goodies/Zurich_Notebook/

III) Shadow of the Relativity Revolution

/+01/ / /2 / % !! 6"

+ > < K $! 0 , + 3 0 , 7

2

' 6 ## 6 " #85" # #; 9 % # 6 # 9 , 796 7#

/+01/ / /2 / < 7 6

' $ *@ < " 6 " " , > " H " " A
7 > " 6 + $ * @ < L
" " " 6 AA E ' $ @ , @ " A
> , 7 ; A " 6 " , > " " >
" 6 ; 6 6 " , 6 " " 6
7

< B " B : " 7 E "
6 " 6 , 6 7 6 ' $ &@
< + 3 < ; B " > " B
6

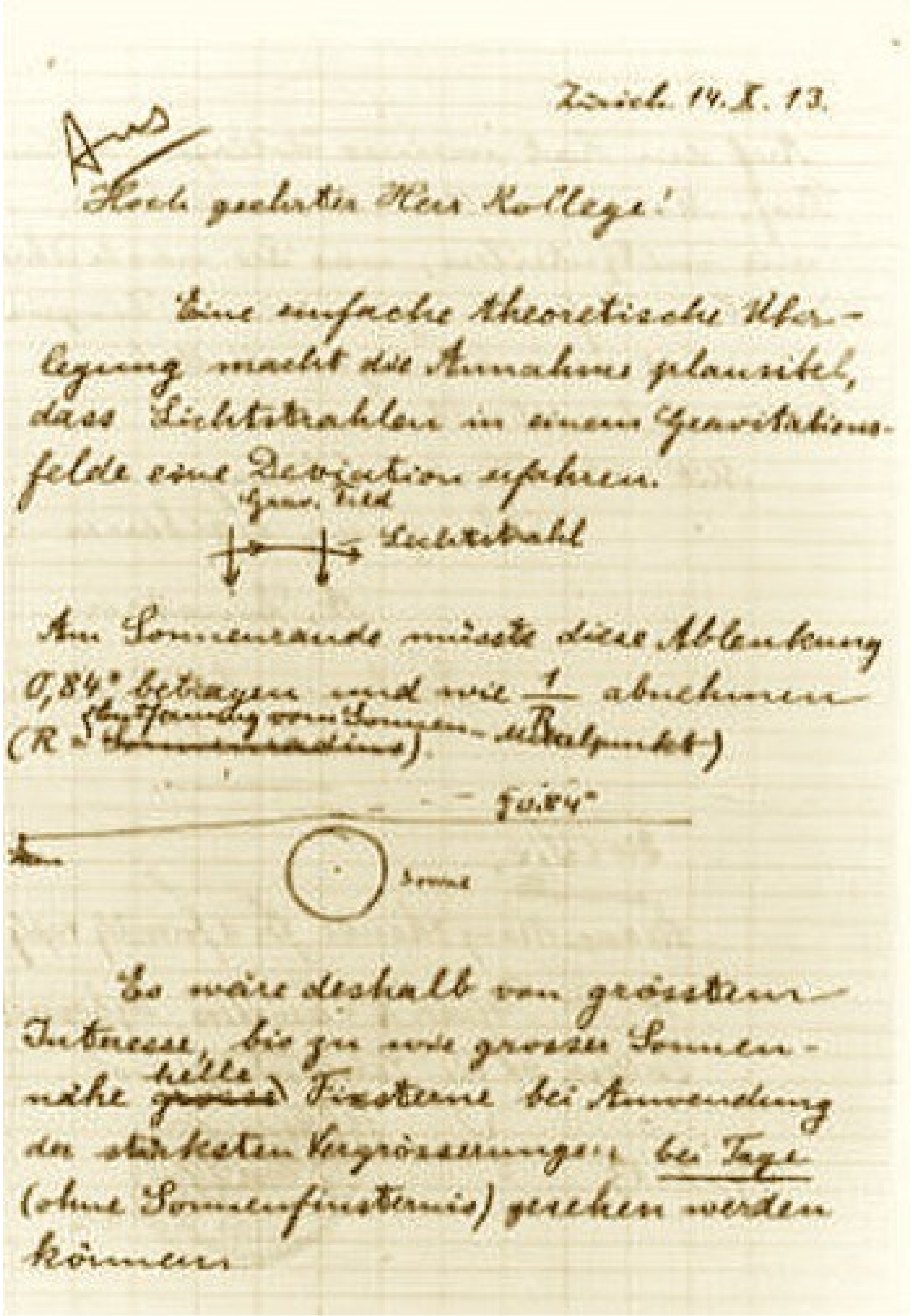

Zürich 14. X. 13.

Ans

Hoch geehrter Herr Kollege!

Eine einfache theoretische Überlegung macht die Annahme plausibel, dass Lichtstrahlen in einem Gravitationsfelde eine Deviation erfahren.

Grav. Feld

Lichtstrahl

Am Sonnenrande müsste diese Ablenkung 0,84" betragen und wie $\frac{1}{R}$ abnehmen (R = ~~Sonnenradius~~ Entfernung vom Sonnen-Mittelpunkt)

0,84"

Sonne

Es wäre deshalb von grösstem Interesse, bis zu wie grosser Sonnennähe ~~grosse~~ helle Fixsterne bei Anwendung der stärksten Vergrösserungen bei Tage (ohne Sonnenfinsternis) gesehen werden können.

6 ## 6 " #85" # #; 9 % # 6 # 9 , 796 7#< 9
9; 56

6 ## " #" # #6 # "# (%!

/+01/ / /2 / 6 9 (%!9 6

/+01/ / /2 / M 6 + >
< 3 < 7 ; @ " " 2 > %@ $ & 6 ; G
. > 7 6

6 ## 7 6 # ? > # # %

6 ## 7 6 # ? > # # N %

/+01/ / /2 / < <? > D " M ' 6

. < ;

6 ## " #" # #6 # "# (%*

/+01/ / /2 / 6 9 (%*9 6

" 6 ## 6 " #85" #3 " #O 9C > B#

/+01/ / /2 / < O C > B 6

2

6 ## > 5 #

6 ## > #, " #0 "#<+0 &%%

6 ## 6 6 6 6 " #, % "
2>

!"#$

"' " 0 , 7 0 ,

6 ## 6 > 6 " # #" # # 6 # * 6"

/+01/ / /2 / * 6"

' " 0 , 7 0 ,

+

30 6"

/+01/ / /2 / 30 6"

) % (% : " M ' , : 5

Finally, a word of warning: there are a large number of sign conventions involved in the definition of the Riemann tensor (and its contractions we will discuss below), so whenever reading a book or article, in particular when you want to use results or equations presented there, make sure what conventions are being used and either adopt those or translate the results into some other convention. As a check: the conventions used here are such that $R_{\phi\theta\phi\theta}$ as well as the curvature scalar (to be introduced below) are *positive* for the standard metric on the two-sphere.

8.4 TIDAL FORCES: INFLUENCE OF CURVATURE ON PARTICLE TRAJECTORIES

In a certain sense the main effect of curvature (or gravity) is that initially parallel trajectories of freely falling non-interacting particles (dust, pebbles,…) do not remain parallel, i.e. that gravity is an attractive force that has the tendency to focus matter. This statement find its mathematically precise formulation in equations describing the influence of space-time curvature on the behaviour of (families of) geodesics.

Let us, as we will need this later anyway, recall first the situation in the Newtonian theory. One particle moving under the influence of a gravitational field is governed by the equation

$$\tfrac{d^2}{dt^2}x^i = -\partial^i\phi(x) \;\;, \tag{8.29}$$

where ϕ is the potential. Now consider a family of particles, or just two nearby particles, one at $x^i(t)$ and the other at $x^i(t) + \delta x^i(t)$. The other particle will of course obey the equation

$$\tfrac{d^2}{dt^2}(x^i + \delta x^i) = -\partial^i\phi(x + \delta x) \;\;. \tag{8.30}$$

From these two equations one can deduce an equation for δx itself, namely

$$\tfrac{d^2}{dt^2}\delta x^i = -\partial^i\partial_j\phi(x)\delta x^j \;\;. \tag{8.31}$$

It describes the effect of *gravitational tidal forces* (the gradient of the gravitational force) on a family of particles moving in a gravitational field.

In particular, when there is no gravitational force, and the trajectories are straight lines, one has

$$\tfrac{d^2}{dt^2}\delta x^i = 0 \quad \Rightarrow \quad \delta x^i = (\delta x^i)_0 + (\delta v^i)t \;\;. \tag{8.32}$$

Thus one recovers Euclid's parallel axiom, that two straight lines intersect at most once (for suitable choices of $\delta v^i \neq 0$) and that they never intersect when they are initially parallel ($\delta v^i = 0$). Any departure from this equation or its Minkowskian counterpart

$$\tfrac{d^2}{d\tau^2}\delta\xi^a = 0 \;\;, \tag{8.33}$$

will therefore indicate a departure from Euclidean geometry!

It is the counterpart of (8.31) that we will be seeking in the context of General Relativity. One derivation of this can be modelled on the Newtonian derivation above. It is elementary but looks non-covariant (and therefore somewhat messy) at intermediate stages of the calculation (see section 12.1 for a manifestly covariant derivation).

The starting point is of course the geodesic equation for x^μ and for its nearby partner $x^\mu + \delta x^\mu$,

$$\tfrac{d^2}{d\tau^2}x^\mu + \Gamma^\mu_{\ \nu\lambda}(x)\tfrac{d}{d\tau}x^\nu \tfrac{d}{d\tau}x^\lambda = 0 \quad , \tag{8.34}$$

and

$$\tfrac{d^2}{d\tau^2}(x^\mu + \delta x^\mu) + \Gamma^\mu_{\ \nu\lambda}(x+\delta x)\tfrac{d}{d\tau}(x^\nu + \delta x^\nu)\tfrac{d}{d\tau}(x^\lambda + \delta x^\lambda) = 0 \quad . \tag{8.35}$$

As above, from these one can deduce an equation for δx, namely

$$\tfrac{d^2}{d\tau^2}\delta x^\mu + 2\Gamma^\mu_{\ \nu\lambda}(x)\tfrac{d}{d\tau}x^\nu \tfrac{d}{d\tau}\delta x^\lambda + \partial_\rho \Gamma^\mu_{\ \nu\lambda}(x)\delta x^\rho \tfrac{d}{d\tau}x^\nu \tfrac{d}{d\tau}x^\lambda = 0 \quad . \tag{8.36}$$

Now this does not look particularly covariant. Thus instead of in terms of $d/d\tau$ we would like to rewrite this in terms of the covariant operator D_τ, with

$$D_\tau \delta x^\mu = \frac{d}{d\tau}\delta x^\mu + \Gamma^\mu_{\ \nu\lambda}\frac{dx^\nu}{d\tau}\delta x^\lambda \quad . \tag{8.37}$$

Calculating $(D_\tau)^2\delta x^\mu$, replacing $\ddot{x}^\mu$ appearing in that expression by $-\Gamma^\mu_{\ \nu\lambda}\dot{x}^\nu\dot{x}^\lambda$ (because x^μ satisfies the geodesic equation) and using (8.36), one eventually finds the nice covariant *geodesic deviation equation*

$$(D_\tau)^2\delta x^\mu = R^\mu_{\ \nu\lambda\rho}\dot{x}^\nu\dot{x}^\lambda\delta x^\rho \tag{8.38}$$

REMARKS:

1. This shows very clearly that curvature, as captured by the Riemann curvature tensor, leads to non-Euclidean geometry in which e.g. the parallel axiom is not necessarily satisifed.

2. In general, solutions to the geodesic deviation equation are called *Jacobi fields.* They describe the difference between the given geodesic and a (hypothetical) infinitely close neighbouring geodesic.

3. Clearly the present derivation of this result leaves something to be desired. It is also possible to give a manifestly covariant, and thus perhaps slightly more satisfactory, derivation of the above geodesic deviation equation, and we will return to this in section 12.1.

data 04.03.2020

Source: https://www.
mpiwg-berlin.mpg.de/
Preprints/P271.pdf

A) Cont.

* newlecturesGR.pdf

Pág. 214

8.4 Tidal Force: Influence of Curvature on Particle Trajectories

Pág. 215

Remarks:

1. this shows very clearly that curvature, as captured by the Riemann curvature tensor, leads to non-Euclidean in which e.g. the parallel axiom is not necessarily satisfied.

other:

Gauss Curvature →

Pág. 260 11.3.

11.3 CURVATURE OF SURFACES: EULER, GAUSS(-BONNET) AND LIOUVILLE

We can generalise the example of the curvature of the 2-sphere, discussed in section 8.6, somewhat, in this way connecting our considerations with the classical realm of the differential geometry of surfaces, in particular with the *Gauss Curvature*, the *Euler characteristic*, the *Gauss-Bonnet* theorem and the *Liouville Equation*.

For any 2-dimensional metric g_{ab} it is a simple exercise to derive the relation between the one independent component, say R_{1212}, of the Riemann tensor, and the scalar curvature. First of all, the Ricci tensor is

$$R_{ab} = R^c{}_{acb} = R^1{}_{a1b} + R^2{}_{a2b} \tag{11.25}$$

so that the scalar curvature is

$$R(g_{ab}) = g^{ab}R_{ab} = g^{11}R^2{}_{121} + g^{12}R^1{}_{112} + g^{21}R^2{}_{221} + g^{22}R^1{}_{212} \ . \tag{11.26}$$

Using the fact that in 2 dimensions the components of the inverse metric are explicitly given by

$$\left(g^{ab}\right) = \frac{1}{g_{11}g_{22} - g_{12}g_{21}} \begin{pmatrix} g_{22} & -g_{12} \\ -g_{21} & g_{11} \end{pmatrix} \tag{11.27}$$

and the (anti-)symmetry properties (1) and (2) of the Riemann tensor, one finds

$$R(g_{ab}) = \frac{2}{g_{11}g_{22} - g_{12}g_{21}} R_{1212} \ . \tag{11.28}$$

This is precisely the relation (8.44) between the Riemann tensor and Ricci scalar. The factor of 2 in this equation is a consequence of our (and the conventional) definition of the Riemann curvature tensor, and is responsible for the fact that the scalar curvature of the unit 2-sphere is $R = +2$. We can also write this result as

$$R_{abcd} = \tfrac{1}{2}(g_{ac}g_{bd} - g_{ad}g_{bc})R \quad \Leftrightarrow \quad R^a{}_{bcd} = \tfrac{1}{2}(\delta^a{}_c g_{bd} - \delta^a{}_d g_{bc})R \ . \tag{11.29}$$

In two dimensions, it is often convenient and natural to absorb this ubiquitous factor of 2 into the definition of the (scalar) curvature, and what one then gets is the classical Gauss Curvature

$$K := \frac{1}{2}R(g_{ab}) \tag{11.30}$$

of a two-dimensional surface.

It follows from (11.29) that the Ricci tensor is related to the Ricci scalar by

$$R_{ab} = (R/2)g_{ab} = K g_{ab} \ . \tag{11.31}$$

This generalises the result for the standard metric on the 2-sphere found by explicit calculation in section 8.6. It shows that in complete generality the Ricci tensor of a two-dimensional space or space-time, thought of as the linear map

$$R^a_b = K\delta^a_b : \quad v^a \mapsto R^a_b v^b \ , \tag{11.32}$$

has only one (double) eigenvalue, namely the Gauss curvature K. It can also be interpreted as saying that in 2 dimensions the Einstein tensor (8.97) is identically zero,

$$G_{ab} = R_{ab} - \tfrac{1}{2} g_{ab} R = 0 \ . \tag{11.33}$$

We will now briefly look at two important and interesting consequences of the above formulae, one related to the Euler characteristic of a surface and its integral representation (the Gauss-Bonnet theorem), and the other to the Liouville equation describing metrics with constant Gauss curvauter $K = k = \pm 1$.

1. Euler Characteristic and the Gauss-Bonnet Theorem

   Let us consider a compact closed surface S, i.e. topologically something like a sphere, or a torus (a sphere with one handle), or a sphere with several handles. A surface S_h with h handles is called a surface of *genus h*.

   Given a metric on S_h, we can associate to S_h its area with respect to the metric,

$$A(S_h) := \int_{S_h} \sqrt{g} d^2x \ . \tag{11.34}$$

   Clearly, this areas depends on a choice of metric, and under a variation δ_g of the metric it transforms as

$$\delta_g A(S_h) = \int_{S_h} \delta_g \sqrt{g} d^2x = \tfrac{1}{2} \int_{S_h} \sqrt{g} d^2x \ g^{ab} \delta g_{ab} \ . \tag{11.35}$$

   Given a metric on S_g, we can also naturally associate to it the real number

$$\chi(S_h) = \frac{1}{2\pi} \int \sqrt{g} d^2x \ K = \frac{1}{4\pi} \int \sqrt{g} d^2x \ R \ . \tag{11.36}$$

   Remarkably, this number turns out to be independent of the metric, in particular *$\chi(S_h)$ is invariant under variations of the metric g_{ab}*,

$$\delta_g \chi(S_h) = 0 \ . \tag{11.37}$$

   Here are two rather explicit ways of establishing this remarkable result:

   (a) The variation of the integrand is

$$\delta_g(\sqrt{g} R) = \delta_g(\sqrt{g} g^{ab} R_{ab}) \ . \tag{11.38}$$

   Since $R = g^{ab} R_{ab}$, it is convenient to express variations of the metric in terms of variations of the inverse metric, with (5.85)

$$\delta \sqrt{g} = \tfrac{1}{2} g^{ab} \delta g_{ab} = -\tfrac{1}{2} g_{ab} \delta g^{ab} \ . \tag{11.39}$$

Then one finds

$$\begin{aligned}\delta_g(\sqrt{g}R) &= (\delta_g\sqrt{g})g^{ab}R_{ab} + \sqrt{g}(\delta g^{ab})R_{ab}) + \sqrt{g}g^{ab}\delta_g R_{ab} \\ &= \sqrt{g}(-\tfrac{1}{2}Rg_{ab} + R_{ab})\delta g^{ab} + \sqrt{g}g^{ab}\delta_g R_{ab} \\ &= \sqrt{g}G_{ab}\delta g^{ab} + \sqrt{g}g^{ab}\delta_g R_{ab} \quad .\end{aligned} \tag{11.40}$$

Now, as shown above, in 2 dimensions the Einstein tensor is identically zero, $G_{ab} = 0$. Moreover, in section 20.2 we will show that in any dimension the 2nd term is a total derivative (20.18),

$$g^{ab}\delta_g R_{ab} = \nabla_a B^a \tag{11.41}$$

for some well-defined B^a built from the covariant derivatives of the variations of the metric, as in (20.19). Taken together, these two facts imply that for a closed surface S_h (without boundary) one has

$$\delta_g\chi(S_h) = \frac{1}{4\pi}\int \sqrt{g}d^2x(G_{ab}\delta g^{ab} + \nabla_a B^a) = 0 \ \ , \tag{11.42}$$

as was to be shown.

(b) Alternatively, somewhat less covariantly but very explicitly, one can show that the integrand $\sqrt{g}K$ or $\sqrt{g}R$ can itself locally be written as a total derivative. Indeed, using (11.29) to write

$$R^1{}_{212} = \tfrac{1}{2}g_{11}R \quad \Leftrightarrow \quad K = \frac{1}{g_{11}}R^2{}_{121} \tag{11.43}$$

and simply writing out explicitly this Riemann curvature tensor component in terms of the Christoffel symbols,

$$R^2{}_{121} = \partial_2\Gamma^2{}_{11} - \partial_1\Gamma^2{}_{12} + \Gamma^2{}_{2a}\Gamma^a{}_{11} - \Gamma^2{}_{1a}\Gamma^a{}_{12} \ \ , \tag{11.44}$$

one finds that $\sqrt{g}K$ can be written as

$$\sqrt{g}K = \in^{ab}\partial_a\beta_b \tag{11.45}$$

with $\in^{12} = -\in^{21} = 1$ the Levi-Civita symbol, and

$$\beta_b = -\frac{\sqrt{g}}{g_{11}}\,\Gamma^2{}_{1b} \quad . \tag{11.46}$$

The fact that the integrand is locally a total derivative does not mean that the integral is zero (because of the non-tensorial nature of β_a, which will typically exhibit coordinate singularities). It does mean however, that the integral of the metric variation of this expression is zero (because that is tensorial and well-defined on S_h),

$$\sqrt{g}K = \varepsilon^{ab}\partial_a\beta_b \quad \Rightarrow \quad \delta_g\chi(S_h) = \frac{1}{2\pi}\int d^2x\delta_g(\sqrt{g}K) = 0 \ \ . \tag{11.47}$$

Either way we have seen that the real number $\chi(S_h)$ is independent of the metric one uses to calculate it. For example, for $h = 0$ and for the standard metric on the sphere S^2 one finds

$$\chi(S_{h=0}) = \chi(S^2) = \frac{1}{4\pi}\int \sqrt{g}R = \frac{1}{2\pi}\int \sqrt{g} = 2 \ , \tag{11.48}$$

and this will therefore be the result for any metric on S^2. Likewise, for $h = 1$, i.e. a torus, by choosing the flat metric on T^2 (see e.g. the discussion and construction in section 18.1), one finds

$$\chi(S_{h=1}) = \chi(T^2) = 0 \ , \tag{11.49}$$

and this will therefore be the result for any metric on T^2 (and it is instructive to check this explicitly for the non-trivial, non-flat metric on T^2 induced by its embedding into $\mathbb{R}^3$ constructed in section 18.1). I am not aware of an equally elementary calculation to determine $\chi(S_h)$ for $h > 1$ in this way but fact of the matter is that

$$\chi(S_h) = 2 - 2h \tag{11.50}$$

is the *Euler characteristic* of S_h, which can also be defined purely combinatorially as the number

$$\chi(S) = n_F - n_E + n_V \tag{11.51}$$

of faces minus vertices plus edges of any cubist rendition of a surface S (and $\chi(S)$ is independent of such a cubist realisation or *triangulation*). The remarkable fact that this *topological invariant* of a surface S can be calculated in terms of differential geometric quantities, namely as the integral of the curvature scalar, is known as the *Gauss-Bonnet theorem.*

2. Constant Curvature and the Liouville Equation

When we specialise the above to the class of *conformally flat* metrics with line element

$$ds^2 = \mathrm{e}^{2h(x,y)}(dx^2 + dy^2) \quad \Leftrightarrow \quad g_{ab} = \exp 2h(x,y)\delta_{ab} \tag{11.52}$$

the calculation of the Riemann tensor is particularly simple and one finds the (easy to memorise) results

$$R^x{}_{yxy} = -\Delta h \tag{11.53}$$

and

$$K = -\mathrm{e}^{-2h}\Delta h \tag{11.54}$$

where $\Delta = \partial_x^2 + \partial_y^2$ is the 2-dimensional Laplacian with respect to the flat Euclidean metric $dx^2 + dy^2$. Thus a surface with *constant curvature* $K = k$ is given by a solution to the non-linear differential equation

$$\Delta h + k\mathrm{e}^{2h} = 0 \ . \tag{11.55}$$

This is the (in-)famous *Liouville equation*, which plays a fundamental role in many branches of mathematics (and mathematical physics).

In terms of the intrinsic Laplacian Δ_g associated to the metric g_{ab}, the Gaussian cuvature and the Liouville equation can also simply be written as

$$K = -\Delta_g h \quad , \quad \Delta_g h + k = 0 \ , \tag{11.56}$$

since, due to the peculiarities of 2 dimensions, $\sqrt{g}g^{ab}$ in independent of h, i.e. is conformally invariant (as we already observed in a different context in section 7.7, cf. (7.120)),

$$\begin{aligned} & \sqrt{g}g^{ab} = \mathrm{e}^{2h}\mathrm{e}^{-2h}\delta^{ab} = \delta^{ab} \\ \Rightarrow \quad & \Delta_g = \frac{1}{\sqrt{g}}\partial_a(\sqrt{g}g^{ab}\partial_b) = \frac{1}{\sqrt{g}}\partial_a(\delta^{ab}\partial_b) = \mathrm{e}^{-2h}\Delta \ . \end{aligned} \tag{11.57}$$

I will not attempt to say anything about the general (local) solution of this equation (which roughly speaking depends on an arbitrary meromorphic function of the complex coordinate $z = x + iy$), but close this section with some special (and particularly prominent) solutions of this equation.

(a) It is easy to see that

$$\mathrm{e}^{2h(x,y)} = y^{-2} \quad \Leftrightarrow \quad h(x,y) = -\ln y \tag{11.58}$$

solves the Liouville equation with $k = -1$. The corresponding space of constant negative curvature is the *Poincaré upper-half plane* model of the hyperbolic geometry,

$$ds^2 = \frac{dx^2 + dy^2}{y^2} \qquad (\ (x,y) \in \mathbb{R}^2,\ y > 0\ )\ . \tag{11.59}$$

By the coordinate transformation $y = e^z$ this is mapped to the equivalent metric

$$ds^2 = dz^2 + \mathrm{e}^{-2z}dx^2 \tag{11.60}$$

on the entire (x, z)-plane.

(b) Another solution (for any k) is the rotationally invariant function

$$\mathrm{e}^{2h(x,y)} = 4(1 + k(x^2 + y^2))^{-2} \quad \Leftrightarrow \quad h = -\ln(1 + k(x^2 + y^2)) + \text{const.} \tag{11.61}$$

  i. For $k = 0$ one finds the flat (zero curvature) Euclidean metric on $\mathbb{R}^2$.

  ii. For $k = +1$, one obtains the metric

$$ds^2 = 4\frac{dx^2 + dy^2}{(1 + x^2 + y^2)^2} \ . \tag{11.62}$$

This is the constant positive curvature metric on the *Riemann sphere* one gets by stereographic projection of the standard metric on the two-sphere S^2 to the (x,y)-plane.

In terms of polar coordinates (r,ϕ) on the Euclidean plane, this metric takes the form

$$ds^2 = 4\frac{dr^2 + r^2 d\phi^2}{(1+r^2)^2} \quad , \tag{11.63}$$

and the further change of variables $r = \tan\theta/2$ shows that this is indeed the standard line element $d\Omega^2$ on the 2-sphere,

$$r = \tan\theta/2 \quad \Rightarrow \quad ds^2 = d\theta^2 + \sin^2\theta d\phi^2 = d\Omega^2 \quad . \tag{11.64}$$

Read backwards, this can also be read as the statement that via the above change of variables the Euclidean metric on $\mathbb{R}^2$ can be written as

$$dr^2 + r^2 d\phi^2 = \frac{(1+r(\theta)^2)^2}{4}(d\theta^2 + \sin^2\theta d\phi^2) = \frac{1}{4\cos^4\theta/2}d\Omega^2 \quad . \tag{11.65}$$

In this process the points “at infinity” in $\mathbb{R}^2$ (where $r \to \infty$) are mapped to the south pole $\theta = \pi$ of the sphere (where the conformal factor in front of the line element of the sphere diverges accordingly). This exhibits S^2 as the *conformal compactification* of $\mathbb{R}^2$.

iii. For $k = -1$, one finds

$$ds^2 = 4\frac{dx^2 + dy^2}{(1-(x^2+y^2))^2} \qquad (\ \{x,y\} \in \mathbb{R}^2, x^2 + y^2 < 1\ ) \quad . \tag{11.66}$$

This is the *Poincaré disc* model of the hyperbolic geometry, defined in the interior of the unit disc in $\mathbb{R}^2$. In terms of polar coordinates, it can also be written as

$$ds^2 = 4\frac{dr^2 + r^2 d\phi^2}{(1-r^2)^2} \qquad (\ 0 \leq r < 1\ ) \tag{11.67}$$

The two metrics (11.59) and (11.66) are isometric, i.e. related by a (albeit not completely evident) coordinate transformation.

(c) As our final example, one other solution (given here only for $k = -1$) is

$$\mathrm{e}^{2h(x,y)} = \mathrm{e}^{2h(x)} = 4\frac{\mathrm{e}^{2x}}{(1-\mathrm{e}^{2x})^2} \quad . \tag{11.68}$$

This leads to the form (2.31) of the unit metric on H^2,

$$ds^2 = d\sigma^2 + \sinh^2\sigma\, d\phi^2 \quad . \tag{11.69}$$

Indeed, we can write this in conformally flat form as

$$ds^2 = \sinh^2\sigma(x)(dx^2 + d\phi^2) \tag{11.70}$$

where

$$x = \log\tanh(\sigma/2) \quad \Rightarrow \quad dx = d\sigma/\sinh\sigma \ . \tag{11.71}$$

Performing the exponential / hyperbolic gymnastics required to write the conformal factor $\sinh^2\sigma$ as a function of x, one finds

$$\sinh^2\sigma(x) = 4\frac{\mathrm{e}^{2x}}{(1-\mathrm{e}^{2x})^2} \ , \tag{11.72}$$

giving rise to the solution given in (11.68).

It is worth remarking that the Poincaré upper-half plane model of a space with constant negative curvature readily generalises to arbitrary dimensions and signature. Thus

$$ds^2 = \frac{d\vec{x}^2 + dy^2}{y^2} \quad , \quad d\vec{x}^2 = \delta_{ab}dx^a dx^b \quad \text{or} \quad d\vec{x}^2 = \eta_{ab}dx^a dx^b \tag{11.73}$$

is the metric of a $D = (d+1)$-dimensional space(-time) with constant negative curvature.

The Lorentzian metric will reappear later as a solution to the Einstein equations with a negative cosmological constant, and is in this context known as the *anti-de Sitter metric* (in Poincaré coordinates, which cover only a part of the complete space-time), and we will discuss this solution in some detail in section 39.

11.4 THE WEYL TENSOR AND ITS USES

In section 8 from the Riemann tensor we have extracted its traces, the Ricci tensor and the Ricci scalar, as well as a particular linear combination of them, the Einstein tensor. We can therefore also explicitly decompose the Riemann tensor into these trace parts and the remaining traceless part.

We noted in section 8.5 that for $D = 2$ and $D = 3$ the Riemann tensor would be "pure trace", i.e. could be written entirely in terms of the Ricci tensor and Ricci scalar. For $D = 2$ we have already established this explicitly by proving the relation (11.29),

$$D = 2: \quad R_{\alpha\beta\gamma\delta} = \tfrac{1}{2}(g_{\alpha\gamma}g_{\beta\delta} - g_{\alpha\delta}g_{\beta\gamma})R \tag{11.74}$$

in section 11.3.

We now look at this issue for $D \geq 3$. Simply by linear algebra one finds, for $D \geq 3$, the decomposition

$$\begin{aligned} R_{\mu\nu\rho\sigma} &= C_{\mu\nu\rho\sigma} \\ &\quad + \frac{1}{D-2}(g_{\mu\rho}R_{\nu\sigma} + R_{\mu\rho}g_{\nu\sigma} - g_{\nu\rho}R_{\mu\sigma} - R_{\nu\rho}g_{\mu\sigma}) \\ &\quad - \frac{1}{(D-1)(D-2)}R(g_{\mu\rho}g_{\nu\sigma} - g_{\nu\rho}g_{\mu\sigma}) \ . \end{aligned} \tag{11.75}$$

This definition is such that $C_{\mu\nu\rho\sigma}$ has all the symmetries of the Riemann tensor (this is manifest) and such that all of its traces are zero, i.e.

$$C^{\mu}_{\ \nu\mu\sigma} = 0 \ , \tag{11.76}$$

as is easily verified. This traceless part $C_{\mu\nu\rho\sigma}$ of the Riemann tensor is called the *Weyl tensor*.

Occasionally it is more convenient and transparent to decompose the Riemann tensor not into the Weyl tensor, the Ricci tensor and the Ricci scalar, but to perform an orthogonal decomposition (with respect to the metric) into the Weyl tensor, the traceless part $S_{\mu\nu}$ of the Ricci tensor,

$$S_{\mu\nu} = R_{\mu\nu} - \frac{1}{D} g_{\mu\nu} R \ , \tag{11.77}$$

and the trace R. Then the decomposition becomes

$$\begin{aligned} R_{\mu\nu\rho\sigma} &= C_{\mu\nu\rho\sigma} \\ &\quad + \frac{1}{D-2}(g_{\mu\rho}S_{\nu\sigma} + S_{\mu\rho}g_{\nu\sigma} - g_{\nu\rho}S_{\mu\sigma} - S_{\nu\rho}g_{\mu\sigma}) \\ &\quad + \frac{1}{D(D-1)} R(g_{\mu\rho}g_{\nu\sigma} - g_{\nu\rho}g_{\mu\sigma}) \ . \end{aligned} \tag{11.78}$$

One other common and convenient decomposition is in terms of a tensor $P_{\mu\nu}$ such that (11.75) takes the form

$$R_{\mu\nu\rho\sigma} = C_{\mu\nu\rho\sigma} + (g_{\mu\rho}P_{\nu\sigma} + P_{\mu\rho}g_{\nu\sigma} - g_{\nu\rho}P_{\mu\sigma} - P_{\nu\rho}g_{\mu\sigma}) \ . \tag{11.79}$$

Comparison with (11.75) shows that this is accomplished by the choice

$$P_{\mu\nu} = \frac{1}{D-2}\left(R_{\mu\nu} - \frac{1}{2(D-1)} g_{\mu\nu} R\right) \ . \tag{11.80}$$

This tensor $P_{\mu\nu}$ is known as the *Schouten Tensor*.

Regardless of how we write the trace part of the Riemann tensor, it turns out that for $D = 3$ the Weyl tensor vanishes identically,

$$D = 3: \quad C_{\alpha\beta\gamma\delta} \equiv 0 \tag{11.81}$$

(I will give an elementary proof of this momentarily). Therefore, for $D = 3$ one has the decomposition

$$\begin{aligned} D = 3: \quad R_{\alpha\beta\gamma\delta} &= (g_{\alpha\gamma}R_{\beta\delta} + R_{\alpha\gamma}g_{\beta\delta} - g_{\alpha\delta}R_{\beta\gamma} - R_{\alpha\delta}g_{\beta\gamma}) \\ &\quad + \tfrac{1}{2}(g_{\alpha\delta}g_{\beta\gamma} - g_{\alpha\gamma}g_{\beta\delta})R \end{aligned} \tag{11.82}$$

This is precisely the result claimed previously in (8.45).

To establish (11.81), in order to trivialise the algebra let us fix a point x_0 and choose coordinates there such that $g_{\alpha\beta}(x_0) = \delta_{\alpha\beta}$ (or $\eta_{\alpha\beta}$, depending on the signature of the metric, but let us assume that we are in the case of Euclidean signature - the same argument works in the Lorentzian case). Now the proof consists of the following elementary steps:

- Since we are in $D = 3$, at least two of the indices in $C_{\alpha\beta\gamma\delta}$ must be equal. Since the Weyl tensor has all the symmetries of the Riemann tensor, if more than two indices are equal, the Weyl tensor component is zero. Thus we only need to consider the components where 2 indices are equal and we can without loss of generality choose these to be $C_{1\beta 1\delta}$, say, with $\beta, \delta \neq 1$.

- Because the Weyl tensor is traceless, one has the relation

  $$C_{1\beta 1\delta} = -C_{2\beta 2\delta} - C_{3\beta 3\delta} \quad . \tag{11.83}$$

  This implies that $C_{1213} = C_{1312} = 0$, so a non-zero component requires $\beta = \delta$.

- For $\beta = \delta$, one derives from this

  $$C_{1212} = -C_{3232} \quad \text{and} \quad C_{1313} = -C_{2323} \tag{11.84}$$

  and likewise for $\alpha = \gamma = 2$ and $\alpha = \gamma = 3$.

- Thus all in all the Weyl tensor can have only 3 independent non-zero components, namely $C_{1212} = C_{2121}, C_{1313} = C_{3131}, C_{2323} = C_{3232}$, and they are all required to be pairwise negatives of each other. This is impossible for non-trivial $C_{\alpha\beta\gamma\delta}$,

  $$C_{1212} = -C_{3232} = +C_{3131} = -C_{2121} = -C_{1212} \quad \Rightarrow \quad C_{1212} = 0 \ , \tag{11.85}$$

  and implies that all of the components of the Weyl tensor are identically zero in $D = 3$.

Thus the Weyl tensor is only non-trivial for $D \geq 4$. Using the Bianchi identies discussed in section 8.8, in particular also (8.94),

$$\nabla^\mu R_{\mu\nu\rho\sigma} = \nabla_\rho R_{\nu\sigma} - \nabla_\sigma R_{\nu\rho} \tag{11.86}$$

from (11.75) one finds a simple expression for the divergence, namely

$$\nabla^\mu C_{\mu\nu\rho\sigma} = (D-3)\left(\nabla_\rho P_{\nu\sigma} - \nabla_\sigma P_{\nu\rho}\right) \quad . \tag{11.87}$$

The tensor appearing on the right-hand side also has its own name. It is called the *Cotton Tensor* $C_{\nu\rho\sigma}$,

$$C_{\nu\rho\sigma} = \nabla_\rho P_{\nu\sigma} - \nabla_\sigma P_{\nu\rho} \quad . \tag{11.88}$$

The content of (11.87) is evidently trivial in $D = 3$, but the Cotton tensor itself is not (and I will briefly come back to this below).

The Weyl tensor plays an important role in many aspects of gravitational physics:

1. For example, the Weyl tensor has traditionally been one of the central objects of interest in the invariant algebraic classification of gravitational fields and in

the characterisation of what are known as algebraically special solutions to the Einstein equations (the so-called *Petrov classification* and related procedures). Originally, this was (of course) developed for $D = 4$, and this case has a number of special features. It is based on the classification of the properties of the eigenvalues λ of the Weyl tensor (at a point x_0), thought of as a map on the space of anti-symmetric $(2,0)$-tensors (*bivectors*),

$$\tfrac{1}{2}C^{\alpha\beta}{}_{\gamma\delta}X^{\gamma\delta} = \lambda X^{\alpha\beta} \tag{11.89}$$

or

$$C^A_B X^B = \lambda X^A \tag{11.90}$$

with $C_{\alpha\beta\gamma\delta} \equiv C_{AB}$ thought of as a symmetric (6×6) matrix.[26]

An equivalent (as it turns out) classification arises from determining the number and multiplicity of linearly independent null vectors ℓ satisfying the condition

$$\ell_{[\gamma}C_{\alpha]\mu\nu[\beta}\ell_{\delta]}\ell^{\mu}\ell^{\nu} = 0 \ . \tag{11.91}$$

Such ℓ^α are called the *principal null directions* of the Weyl tensor. More recently, this classification scheme (based on the latter approach) has been (partially) extended to higher dimensions.[27]

2. As we will see in section 19.6, the Einstein equations imply that the Weyl tensor describes the gravitational field in vacuum. Specifically, when (or where) the energy-momentum tensor is zero, the Riemann curvature tensor is equal to the Weyl tensor,

$$T_{\alpha\beta}(x) = 0 \quad \Rightarrow \quad R_{\alpha\beta\gamma\delta}(x) = C_{\alpha\beta\gamma\delta}(x) \ . \tag{11.92}$$

   The Weyl tensor thus encodes the information about things like gravitational waves and the asymptotic behaviour of a gravitational field and has been studied extensively from this point of view.

3. In the presence of matter, on the other hand, (11.87), in conjunction with the Einstein equations, becomes an evolution equation for these vacuum components of the gravitational field in terms of the sources - see equations (19.57) and (19.58).

The Weyl tensor also plays an important role in geometry, as it is *conformally invariant*, i.e. $C^{\mu}_{\nu\rho\sigma}$ is invariant under conformal (Weyl) rescalings of the metric,

$$g_{\mu\nu}(x) \to \mathrm{e}^{2f(x)}g_{\mu\nu}(x) \quad \Rightarrow \quad C^{\mu}_{\ \nu\rho\sigma} \to C^{\mu}_{\ \nu\rho\sigma} \ , \tag{11.93}$$

---

[26] See e.g. sections 7.16 and 11 of J. Plebanski, A. Krasinski, *An Introduction to General Relativity and Cosmology*, or section 4 of H. Stephani, D. Kramer, M. MacCallum, C. Hoenslaers, E. Herlt, *Exact Solutions to Einstein's Field Equations* (2nd Edition) for recent expositions of this subject.

[27] See A. Coley, R. Milson, V. Pravda, A. Pravdova, *Classification of the Weyl Tensor in Higher Dimensions*, `arXiv:arXiv:gr-qc/0401008`; A. Coley, *Classification of the Weyl Tensor in Higher Dimensions and Applications*, `arXiv:0710.1598 [gr-qc]`; M. Ortaggio, V. Pravda, A. Pravdova, *Algebraic classification of higher dimensional spacetimes based on null alignment*, `arXiv:1211.7289 [gr-qc]`.

equivalently

$$g_{\mu\nu}(x) \to \mathrm{e}^{2f(x)} g_{\mu\nu}(x) \quad \Rightarrow \quad C_{\mu\nu\rho\sigma} \to \mathrm{e}^{2f(x)} C_{\mu\nu\rho\sigma} \ . \tag{11.94}$$

In particular, the Weyl tensor is zero if the metric is *conformally flat*, i.e. related by a conformal transformation to the flat metric $\eta_{\mu\nu}$ (of any signature),

$$g_{\mu\nu}(x) = \mathrm{e}^{2f(x)} \eta_{\mu\nu}(x) \quad \Rightarrow \quad C^{\mu}_{\ \nu\rho\sigma} = 0 \ . \tag{11.95}$$

This can be established by brute force calculation and is not *per se* particularly enlightning.

Conversely for $D \geq 4$ vanishing of the Weyl tensor is also a sufficient condition for a metric to be (locally) conformal to the flat metric. This is a non-trivial result because at face value one seems to obtain a completely overdetermined system of equations for the single function f, of the form

$$P_{\mu\nu} = \{\text{1st and 2nd derivatives of } f\}_{\mu\nu} \ . \tag{11.96}$$

However, it turns out that the integrability conditions for this system of equations are equivalent to the vanishing of the Weyl tensor, and then a variant of the *Frobenius integrability theorem* (mentioned in a different context in section 15.5) can be used to establish the local existence of a solution f.

For $D = 3$, the situation is slightly (but not fundamentally) different. We see from (11.88) that for any $D \geq 4$ conformal flatness implies vanishing of the Cotton tensor. It turns out that for $D = 3$ the Cotton tensor takes over the role of the Weyl tensor (which, as proven above, is itself trivial for $D = 3$), i.e. one has the statement that for $D = 3$ a metric is (locally) conformally flat if and only if the Cotton tensor vanishes.

11.5 GENERALISATIONS: TORSION AND NON-METRICITY

In section 5.4 we had seen that the Levi-Civita connection (defined by the Christoffel symbols) is characterised by the fact that

1. the metric is covariantly constant, $\nabla_\mu g_{\nu\lambda} = 0$, and
2. the torsion is zero, i.e. the second covariant derivatives of a scalar commute.

It is of course possible to relax either of the conditions (1) or (2), or both of them and, in particular, connections with torsion (relaxation of condition 2) are popular in certain circles and/or arise naturally in certain generalised (gauge) theories of gravity and in string theory.

To discuss this a bit more systematically, we consider a general connection

$$\tilde{\Gamma}^{\mu}_{\ \nu\lambda} = \Gamma^{\mu}_{\ \nu\lambda} + C^{\mu}_{\ \nu\lambda} \tag{11.97}$$

with $\Gamma^{\mu}_{\ \nu\lambda}$ the canonical Levi-Civita connection, and $C^{\mu}_{\ \nu\lambda}$ a $(1,2)$-tensor. We will also use the corresponding $(0,3)$-tensor

$$C_{\mu\nu\lambda} = g_{\mu\rho}C^{\rho}_{\ \nu\lambda} \quad . \tag{11.98}$$

Associated with $\tilde{\Gamma}^{\mu}_{\ \nu\lambda}$ we have the covariant derivative $\tilde{\nabla}_{\mu}$. Since $\tilde{\Gamma}^{\mu}_{\ \nu\lambda}$ will in general not be symmetric in its lower indices, in this section we need to be particularly careful with (and choose a convention for) the ordering of the lower indices in the covariant derivative. We will choose the convention that the last index always refers to the direction along which one is differentiating, i.e.

$$\tilde{\nabla}_{\mu}V^{\nu} = \partial_{\mu}V^{\nu} + \tilde{\Gamma}^{\nu}_{\ \lambda\mu}V^{\lambda} \tag{11.99}$$

etc. The reason for this choice is that one should think of the collection of objects $\Gamma^{\nu}_{\ \lambda\mu}$ (and $\tilde{\Gamma}^{\nu}_{\ \lambda\mu}$) as the coefficients of a matrix-valued 1-form (cf. section 4.6) $\Gamma^{\nu}_{\ \lambda} = \Gamma^{\nu}_{\ \lambda\mu}dx^{\mu}$, the matrices acting by rotation on vectors (and more general tensors), as in (5.21).

We now define the *torsion tensor* $T^{\lambda}_{\ \mu\nu}$ by

$$[\tilde{\nabla}_{\mu}, \tilde{\nabla}_{\nu}]\phi = T^{\lambda}_{\ \mu\nu}\partial_{\lambda}\phi \quad , \quad T_{\lambda\mu\nu} = g_{\lambda\rho}T^{\rho}_{\ \mu\nu} \quad , \tag{11.100}$$

and the *non-metricity tensor* $Q_{\nu\lambda\mu}$ by

$$\tilde{\nabla}_{\mu}g_{\nu\lambda} = -Q_{\nu\lambda\mu} \quad . \tag{11.101}$$

In terms of the $C^{\mu}_{\ \nu\lambda}$ these tensors can be written as

$$\begin{aligned} T_{\lambda\mu\nu} &= C_{\lambda\mu\nu} - C_{\lambda\nu\mu} = 2C_{\lambda[\mu\nu]} \\ Q_{\nu\lambda\mu} &= C_{\nu\lambda\mu} + C_{\lambda\nu\mu} = 2C_{(\nu\lambda)\mu} \quad . \end{aligned} \tag{11.102}$$

Thus the torsion is zero iff $C^{\lambda}_{\ \mu\nu}$ (and hence $\tilde{\Gamma}^{\lambda}_{\ \mu\nu}$) is symmetric in its lower indices, and the connection is compatible with the metric iff $C_{\nu\lambda\mu}$ is anti-symmetric in its first two indices. In particular, if the torsion is zero and the connection is metric-compatible, one has

$$C_{\lambda\mu\nu} = C_{\lambda\nu\mu} \quad \text{and} \quad C_{\lambda\mu\nu} = -C_{\mu\lambda\nu} \quad \Rightarrow \quad C_{\lambda\mu\nu} = 0 \quad , \tag{11.103}$$

as one can see by the gymnastics

$$C_{\lambda\mu\nu} = C_{\lambda\nu\mu} = -C_{\nu\lambda\mu} = -C_{\nu\mu\lambda} = C_{\mu\nu\lambda} = C_{\mu\lambda\nu} = -C_{\lambda\mu\nu} \quad . \tag{11.104}$$

Conversely, since the absence of torsion and non-metricity characterises the Levi-Civita connection, it should be possible to express the deviation $C_{\lambda\mu\nu}$ from the Levi-Civita connection entirely in terms of torsion and non-metricity. This is indeed the case. By repeating the calculation (5.44) in this more general context, one finds

$$2C_{\lambda(\mu\nu)} = Q_{\nu\lambda\mu} + Q_{\mu\lambda\nu} - Q_{\mu\nu\lambda} - T_{\mu\lambda\nu} - T_{\nu\lambda\mu} \quad . \tag{11.105}$$

Combining this with $2C_{\lambda[\mu\nu]} = T_{\lambda\mu\nu}$, one obtains

$$\begin{aligned} C_{\lambda\mu\nu} &= \tfrac{1}{2}(T_{\lambda\mu\nu} - T_{\mu\lambda\nu} - T_{\nu\lambda\mu}) + \tfrac{1}{2}(Q_{\mu\lambda\nu} + Q_{\nu\lambda\mu} - Q_{\mu\nu\lambda}) \\ &\equiv \tilde{T}_{\lambda\mu\nu} + \tilde{Q}_{\lambda\mu\nu} \quad , \end{aligned} \tag{11.106}$$

with

$$\tilde{T}_{\lambda\mu\nu} = \tfrac{1}{2}(T_{\lambda\mu\nu} - T_{\mu\lambda\nu} - T_{\nu\lambda\mu}) = -\tilde{T}_{\mu\lambda\nu} \tag{11.107}$$

and

$$\tilde{Q}_{\lambda\mu\nu} = \tfrac{1}{2}(Q_{\mu\lambda\nu} + Q_{\nu\lambda\mu} - Q_{\mu\nu\lambda}) = \tilde{Q}_{\lambda\nu\mu} \quad . \tag{11.108}$$

Thus we can now split a general connection more informatively into the 3 pieces

$$\tilde{\Gamma}^{\lambda}{}_{\mu\nu} = \Gamma^{\lambda}{}_{\mu\nu} + \tilde{T}^{\lambda}{}_{\mu\nu} + \tilde{Q}^{\lambda}{}_{\mu\nu} \quad . \tag{11.109}$$

REMARKS:

1. The tensor $\tilde{T}_{\lambda\mu\nu}$ is known as the *contorsion tensor* (frequently (mis-)spelled as "contortion" tensor). I am not aware of a commonly used name for $\tilde{Q}_{\lambda\mu\nu}$, and will not try to invent one. The contorsion tensor is the linear combination of components of the torsion tensor that appear as the connection coefficents of a general metric-compatible connection with torsion,

$$Q_{\lambda\mu\nu} = 0 \quad \Rightarrow \quad \tilde{\Gamma}^{\lambda}{}_{\mu\nu} = \Gamma^{\lambda}{}_{\mu\nu} + \tilde{T}^{\lambda}{}_{\mu\nu} \quad . \tag{11.110}$$

   In general it can have both symmetric and anti-symmetric components,

$$\tilde{T}_{\lambda(\mu\nu)} = -\tfrac{1}{2}(T_{\mu\lambda\nu} + T_{\nu\lambda\mu}) = T_{(\mu\nu)\lambda} \quad , \quad \tilde{T}^{\lambda}{}_{[\mu\nu]} = \tfrac{1}{2}T^{\lambda}{}_{\mu\nu} \quad , \tag{11.111}$$

   but it cannot be symmetric (if the contorsion were symmetric, the torsion, and hence the contorsion, would be zero). If its symmetric part vanishes, then $\tilde{T}_{\lambda\mu\nu}$ is completely anti-symmetric,

$$\tilde{T}^{\lambda}{}_{(\mu\nu)} = 0 \quad \Rightarrow \quad \tilde{T}_{\lambda\mu\nu} = \tilde{T}_{[\lambda\mu\nu]} \quad . \tag{11.112}$$

2. Since $\Gamma_{\lambda\mu\nu}$ contains a part that is antisymmetric in the first 2 indices,

$$\Gamma_{\lambda\mu\nu} = \tfrac{1}{2}g_{\lambda\mu,\nu} + \tfrac{1}{2}(g_{\nu\lambda,\mu} - g_{\nu\mu,\lambda}) \quad , \tag{11.113}$$

   one might be tempted to think that that part can be cancelled (or absorbed) by a metric-compatible $C_{\lambda\mu\nu} = C_{[\lambda\mu]\nu}$, so that a very simple metric-compatible connection would be

$$\tilde{\Gamma}_{\lambda\mu\nu} \overset{?}{=} \tfrac{1}{2}g_{\lambda\mu,\nu} \quad , \quad \tilde{\Gamma}^{\lambda}{}_{\mu\nu} \overset{?}{=} \tfrac{1}{2}g^{\lambda\rho}g_{\rho\mu,\nu} \quad . \tag{11.114}$$

   However, the term that one has cancelled (or absorbed) is *not* a tensor. Therefore, this candidate "connection" does not transform as (and therefore does not qualify as) a connection and cannot be used to define a covariant derivative.

3. In general, for a connection $\tilde{\nabla}$, the notions of autoparallels (section 5.8),

$$\tilde{\nabla}_\tau X^\mu = 0 \quad \Leftrightarrow \quad \ddot{x}^\mu + \tilde{\Gamma}^\mu_{\ \nu\lambda}\dot{x}^\nu\dot{x}^\lambda = 0 \ , \tag{11.115}$$

(i.e. curves characterised by the fact that their tangent vectors are parallel transported along the curve - this depends on a choice of connection) no longer coincides with the notion of geodesics (which are obtained by extremising proper time or distance, and which always lead to the Levi-Civita connection). However, this difference disappears if $C^\mu_{\ \nu\lambda}$ happens to be anti-symmetric in its lower indices (e.g. for a metric-compatible connection with totally anti-symmetric contorsion tensor), as one then has

$$\ddot{x}^\mu + \tilde{\Gamma}^\mu_{\ \nu\lambda}\dot{x}^\nu\dot{x}^\lambda = \ddot{x}^\mu + \Gamma^\mu_{\ \nu\lambda}\dot{x}^\nu\dot{x}^\lambda \ . \tag{11.116}$$

We have defined the Riemann tensor via the commutator of covariant derivatives (8.2)

$$[\nabla_\mu, \nabla_\nu]V^\lambda = R^\lambda_{\ \sigma\mu\nu}V^\sigma \tag{11.117}$$

associated to the Levi-Civita connection (Christoffel symbols), or, equivalently, by the relation (8.5) (now being careful with the positioning of the lower indices)

$$R^\lambda_{\ \sigma\mu\nu} = \partial_\mu\Gamma^\lambda_{\ \sigma\nu} - \partial_\nu\Gamma^\lambda_{\ \sigma\mu} + \Gamma^\lambda_{\ \rho\mu}\Gamma^\rho_{\ \sigma\nu} - \Gamma^\lambda_{\ \rho\nu}\Gamma^\rho_{\ \sigma\mu} \ . \tag{11.118}$$

In order to show explicitly (rather than by appealing to (11.117)) that this transforms as a tensor, all that one needs is the characteristic non-tensorial transformation behaviour of the Christoffel symbols $\Gamma^\lambda_{\ \mu\nu}$. As discussed in section 5.4 and above, an arbitrary connection $\tilde{\Gamma}^\lambda_{\ \mu\nu}$ that can be used to define a tensorial covariant derivative has the same non-tensorial transformation behaviour. Therefore

$$\tilde{R}^\lambda_{\ \sigma\mu\nu} \equiv R^\lambda_{\ \sigma\mu\nu}(\tilde{\Gamma}) = \partial_\mu\tilde{\Gamma}^\lambda_{\ \sigma\nu} - \partial_\nu\tilde{\Gamma}^\lambda_{\ \sigma\mu} + \tilde{\Gamma}^\lambda_{\ \rho\mu}\tilde{\Gamma}^\rho_{\ \sigma\nu} - \tilde{\Gamma}^\lambda_{\ \rho\mu}\tilde{\Gamma}^\rho_{\ \sigma\mu} \tag{11.119}$$

defines a tensor for any connection, namely the curvature tensor of the connection $\tilde{\Gamma}^\lambda_{\ \mu\nu}$. It is related to the commutator of covariant derivatives by

$$[\tilde{\nabla}_\mu, \tilde{\nabla}_\nu]V^\lambda = \tilde{R}^\lambda_{\ \sigma\mu\nu}V^\sigma + (\tilde{\Gamma}^\rho_{\ \mu\nu} - \tilde{\Gamma}^\rho_{\ \nu\mu})\nabla_\rho V^\lambda = \tilde{R}^\lambda_{\ \sigma\mu\nu}V^\sigma + T^\rho_{\mu\nu}\nabla_\rho V^\lambda \ , \tag{11.120}$$

where $T^\rho_{\mu\nu}$ is the torsion tensor. As before, one can also define the Ricci tensor and Ricci scalar by

$$\tilde{R}_{\mu\nu} \equiv R_{\mu\nu}(\tilde{\Gamma}) = \tilde{R}^\lambda_{\ \mu\lambda\nu} \quad , \quad \tilde{R} \equiv R(\tilde{\Gamma}) = g^{\mu\nu}\tilde{R}_{\mu\nu} \ . \tag{11.121}$$

However, it is crucial to keep in mind that the symmetry properties and Bianchi identities satisfied by these generalised curvature tensors will in general differ from those of the Riemann-Christoffel tensor. This should be clear from the way we derived the symmetries of the Riemann tensor in section 8.3, where we related the symmetries to the properties (metricity, no torsion) that characterise the canonical Levi-Civita connection (Christoffel symbols). For example, in general the Ricci tensor will not be symmetric, the Bianchi identity $R_{\alpha[\beta\gamma\delta]} = 0$ will be replaced by an identity relating $\tilde{R}_{\alpha[\beta\gamma\delta]}$ to the torsion (and its covariant derivative), etc.[28]

---

[28] For more on this and related topics, see e.g. section 1 of T. Ortin, *Gravity and Strings*.

data 04.03.2020

11.3 Curvature of Surfaces: Euler, Gauss(-Bonnet) and Liouville

... We can generalise the example of the curvature of the 2-sphere, discussed in section 8.6, somewhat, in this way connecting our considerations with the classical realm of the differential geometry of surfaces, in particular with the Gauss curvature, the Euler characteristic, the Gauss-Bonnet theorem and the Liouville Equation.

Pág. 261 – Gauss curvature K.

It can also be interpreted as saying that in 2 dimensions the Einstein tensor (8.97)
Pág. 226 ← → book.

B) There's no such things as the Force of Gravity.

04 03 2020

Source: Books_46_2020/
Loop Quantum Gravity for
Everyone/

Quantum Space.pdf

Pág. 27(48).

c) Space-Time, Yang-Mills Gravity,
and Dynamics of Cosmic
Expansion.

Source: Books_46_2020/
Space-Time.../
Space-Time....pdf

Pág. XXXI (32)

" Einstein's approach involving the geometrization of physical fields in curved space-time appears to be successful only in the formulation of classical gravity with the conservation of the energy-momentum pseudo-tensor, rather than the true energy-momentum tensor.

... Finsler geometry,

where the fundamental metric tensors depend on both positions and

>

> /+01/ / /2 /
. 691 93 , 79 9<, 7 6"

> /+01/ / /2 / 1 6 6"

) * %(: L , 7

6 :

/+01/ / /2 / 6 : @ P
3 , 7@ " G7 <?6 9 E 7 , 7 6
6 " $@ D20.G '<C:'M' =H 6"

) QQQ' &

space-time translational gauge symmetry for the motion of quantum particles in the classical limit. In the first-order approximation, Yang–Mills gravity is consistent with the experimentally measured gravitational redshift. In the second-order approximation, it correctly predicts the perihelion shift of Mercury and the gravitational quadrupole radiation of binary pulsars.[13] Furthermore, Yang–Mills gravity allows a quantum theory of gravity to be formulated and Feynman–Dyson rules for the gravitational tensor fields and the associated "ghost particles"[13] to be derived, as we will show in Chapters 10 and 11. Thus, Yang–Mills gravity brings the gravitational interaction back into the experimentally established arena of gauge fields based on flat space-time.

In contrast, Einstein's approach involving the geometrization of physical fields in curved space-time appears to be successful only in the formulation of classical gravity, with the conservation of the energy–momentum pseudo-tensor, rather than the true energy–momentum tensor. If one were to adopt a similar approach to, say, the electromagnetic interaction, which is velocity-dependent, it would be natural to employ the Finsler geometry, where the fundamental metric tensors depend on both positions and velocities (i.e., the differentials of coordinates), rather than the Riemannian geometry.[14] However, all attempts thus far to geometrize classical electrodynamics have not even begun to approach the usefulness of quantum electrodynamics.

The classical theory of Einstein gravity is usually considered to be the most coherent mathematically and satisfying aesthetically.[2] However, there is little chance of making progress simply by going over its successes. Because of the success of Einstein gravity, many have made arguments that attempted to make it unique or necessary.[15,h] There is an old argument due to Feynman (and perhaps others) that the only consistent theory (with no more than two derivatives) of a spin-2 field coupled to itself is, in fact, general relativity. It is not necessary to analyze the detailed arguments one by one, all that is necessary is to present a counterexample.

---

[h] A century ago, Einstein observed "Concepts which have proved useful for ordering things easily assume so great an authority over us, that we forget their terrestrial origin and accept them as unalterable facts. They then become labeled as "conceptual necessities," "*a priori situations*," etc. The road to scientific progress is frequently blocked for long periods by such errors. It is therefore not just an idle game to exercise our ability to analyze familiar concepts, and to demonstrate the conditions on which their justification and usefulness depend, and the way in which these developed, little by little, from the data of experience. In this way they are deprived of their excessive authority."

Yang–Mills gravity, with its translational gauge symmetry in flat space-time and consistency with all known experiments, as well as other desirable features such as the preservation of the conservation law of energy–momentum, operational definitions of space-time coordinates, and quantizability of gravitational fields, is just such a counterexample from outside the curved space-time framework of general relativity.[i]

0.4. Unification of All Interactions Based on the Yang–Mills–Utiyama–Weyl Framework

In Chapters 14–16, we consider a model for unifying all interactions in nature (including gravity) within the taiji symmetry and the generalized Yang–Mills frameworks, following the ideas of Glashow, Salam, Ward, and Weinberg for the unified electroweak interaction.[16] We here present a brief preview of the unification of the gravitational interaction with just the electroweak interaction.

The T_4 gauge covariant derivative (0.5) paves the way for a unification of the gravitational and electroweak ("gravity-electroweak") interactions based on $T_4 \times SU_2 \times U_1$. Consider a unified "gravity-electroweak" model based on a generalized covariant derivative d_μ,[17]

$$d_\mu \equiv J_\mu^\nu \partial_\nu - if W_\mu^a t^a - if' U_\mu t_o, \quad J_\mu^\nu = \delta_\mu^\nu + g\phi_\mu^\nu. \tag{0.8}$$

This is a linear combination of the usual gauge covariant derivative in the electroweak theory and the gravitational T_4 gauge covariant derivative (0.5). The gauge curvatures, $C_{\mu\nu\alpha}$, $W_{\mu\nu}^a$, and $U_{\mu\nu}$, associated with each group in the "gravity-electroweak" covariant derivative d_μ are given by

$$[d_\mu, d_\nu] = C_{\mu\nu\alpha} T^\alpha - if W_{\mu\nu}^a t^a - if' U_{\mu\nu} t_o, \qquad [t^a, t^b] = i\epsilon^{abc} t^c, \tag{0.9}$$

and the $T(4)$ gauge curvature $C_{\mu\nu\alpha}$ in flat space-time is[17]

$$C_{\mu\nu\alpha} = J_{\mu\sigma}\partial^\sigma J_{\nu\alpha} - J_{\nu\sigma}\partial^\sigma J_{\mu\alpha} = \Delta_\mu J_{\nu\alpha} - \Delta_\nu J_{\mu\alpha} = -C_{\nu\mu\alpha}. \tag{0.10}$$

---

[i] For key features of Yang–Mills gravity and Einstein gravity, see a comparison at the end of this overview.

data 04.03.2020

S T [✓ Q] Q S S D

Source: Books_46_2020/
Loop Quantum Gravity for Everyone/
Quantum Space.pdf

Pág. 27(48).

c) Space-Time, Yang-Mills Gravity, and Dynamics of Cosmic Expansion.

Source: Books_46_2020/
Space-Time.../
Space-Time....pdf

Pág. XXXI (32)

" Einstein's approach involving the geometrization of physical fields in curved space-time appears to be successful only in the formulation of classical gravity with the conservation of the energy-momentum pseudo-tensor, rather than the true energy-momentum tensor.

... Finsler geometry,

where the fundamental metric tensors depend on both positions and

04.03.2020

velocity (i.e, the differentials of coordinates),

rather than the Riemannian geometry.

D) Free-surface flow
environmental fluid mechanics ...pdf

Pág. 483(497)
6.6.2 Cauchy-Riemann Equations

Pág. 486(500)
Figure 6.30
Conformal mapping of flow net.

121) 20200403

Books:

A) DATA-NOVOVIW_63_2020/
Books_63_2020/
Approximate Calculation of Integrals/
Approximate Calculation of Integrals ... pdf

" M , "

/O' /H22J 9 9 $9=6" "9 & * *A/H22J 9 9 $/H B 9%!9 /

/O'
/H22J 9 9 $9=6" "9 & * *A/H22J 9 9 $/H B 9%!9 /
M " M /

/+01/ / /2 / M 9 , "
$@ < , @ H ; 6"

%(& %$* ! ! 7 0 <E

6.6.2 Cauchy-Riemann Equations

Functions of a complex variable can be defined by introducing another complex variable, as follows

$$f(z) = \phi(x, y) + \jmath\psi(x, y) \tag{6.162}$$

where $\phi(x, y)$ and $\psi(x, y)$ are real functions of x and y. If the function $f(z)$ is uniquely defined, and its derivatives are also uniquely defined within some region of the $f(z)$ plane, the function is called *analytic* or *holomorphic*. This last condition, however, is not trivial, as there are multiple paths along which a complex derivative may approach its limit as $z \to 0$, i.e.

$$\frac{df}{dz} = \lim_{\Delta z=0} \frac{f(z + \Delta z) - f(z)}{\Delta z}$$

If, for convenience, we select two paths along the Cartesian coordinate directions, x and y, respectively, the corresponding partial derivatives may be written as follows

$$\frac{\partial f}{\partial x} = \lim_{\Delta x=0} \frac{f(z + \Delta x) - f(z)}{\Delta x} \qquad \frac{1}{\jmath}\frac{\partial f}{\partial y} = \lim_{\Delta y=0} \frac{f(z + \Delta y) - f(z)}{\jmath\Delta y}$$

For the derivative of z to be unique, we must have

$$\frac{\partial f}{\partial x} = \frac{1}{\jmath}\frac{\partial f}{\partial y}$$

This can also be written in terms of the real and imaginary components of z, as follows

$$\frac{\partial \phi}{\partial x} + \jmath\frac{\partial \psi}{\partial x} = \frac{1}{\jmath}\left(\frac{\partial \phi}{\partial y} + \jmath\frac{\partial \psi}{\partial y}\right)$$

The *modulus* of the complex variable is given by

$$r = |z| = \sqrt{x^2 + y^2}$$

Therefore, the complex variable can be also expressed in polar coordinates in terms of its *argument*, θ, as follows

$$z = r\left(\cos\theta + \jmath\sin\theta\right) \tag{6.160}$$

Furthermore, using Taylor series expansions, the trigonometric functions can be related to the exponential function. Thus, based on what is known as *DeMoivre's theorem*, a complex variable can be written in exponential form, as follows (Streeter, 1948, p. 87)

$$z = re^{\jmath\theta} \tag{6.161}$$

6.6.2 Cauchy-Riemann Equations

Functions of a complex variable can be defined by introducing another complex variable, as follows

$$f(z) = \phi(x, y) + \jmath\psi(x, y) \tag{6.162}$$

where $\phi(x, y)$ and $\psi(x, y)$ are real functions of x and y. If the function $f(z)$ is uniquely defined, and its derivatives are also uniquely defined within some region of the $f(z)$ plane, the function is called *analytic* or *holomorphic*. This last condition, however, is not trivial, as there are multiple paths along which a complex derivative may approach its limit as $z \to 0$, i.e.

$$\frac{df}{dz} = \lim_{\Delta z = 0} \frac{f\left(z + \Delta z\right) - f\left(z\right)}{\Delta z}$$

If, for convenience, we select two paths along the Cartesian coordinate directions, x and y, respectively, the corresponding partial derivatives may be written as follows

$$\frac{\partial f}{\partial x} = \lim_{\Delta x = 0} \frac{f\left(z + \Delta x\right) - f\left(z\right)}{\Delta x} \qquad \frac{1}{\jmath}\frac{\partial f}{\partial y} = \lim_{\Delta y = 0} \frac{f\left(z + \Delta y\right) - f\left(z\right)}{\jmath\Delta y}$$

For the derivative of z to be unique, we must have

$$\frac{\partial f}{\partial x} = \frac{1}{\jmath}\frac{\partial f}{\partial y}$$

This can also be written in terms of the real and imaginary components of z, as follows

$$\frac{\partial\phi}{\partial x} + \jmath\frac{\partial\psi}{\partial x} = \frac{1}{\jmath}\left(\frac{\partial\phi}{\partial y} + \jmath\frac{\partial\psi}{\partial y}\right)$$

Thus, provided that the partial derivatives of ϕ and ψ are continuous, a necessary, but not sufficient, condition for $f(z)$ to be analytic is

$$\frac{\partial\phi}{\partial x}=\frac{\partial\psi}{\partial y} \qquad \text{and} \qquad \frac{\partial\phi}{\partial y}=-\frac{\partial\psi}{\partial x} \tag{6.163}$$

These are called the *Cauchy-Riemann equations*. Points on the complex plane where these equations are not satisfied are called *singular points*.

6.6.3 Complex Potential

For ideal fluid flow, it is convenient to identify the aforementioned real variables with the velocity potential, ϕ, and the stream function, ψ. For then, differentiation of the first of Eq. (6.163) with respect to x, the second with respect to y, and adding the results, leads to exactly the two-dimensional form of the Laplace equation for the velocity potential, i.e. Eq. (6.15). Similarly, differentiation of the first of Eq. (6.163) with respect to y, the second with respect to x, and subtracting the results, leads to the Laplace equation for the stream function, i.e. Eq. (6.17).

Because the velocity potential, ϕ, and the stream function, ψ, satisfy the Cauchy-Riemann equations, they may also be interpreted as the real and imaginary parts of an analytic function called the *complex potential*, i.e.

$$w(z)=\phi+\jmath\psi \tag{6.164}$$

The rate of change of w with respect to z is called the *complex velocity*. Since w is a function of z only, its derivative can be determined by first computing its total differential, as follows

$$dw=\frac{\partial\phi}{\partial x}dx+\frac{\partial\phi}{\partial x}dy+\jmath\frac{\partial\psi}{\partial x}dx+\jmath\frac{\partial\psi}{\partial y}dy$$

The partial derivatives with respect to y can be substituted using the Cauchy-Riemann equations, i.e. Eq. (6.163), leading to

$$dw=\frac{\partial\phi}{\partial x}\left(dx+\jmath dy\right)+\jmath\frac{\partial\psi}{\partial x}dx\left(dx+\jmath dy\right)$$

Therefore, the derivative of the complex potential is given by

$$\frac{dw}{dz}=\frac{\partial\phi}{\partial x}+\jmath\frac{\partial\psi}{\partial x} \tag{6.165}$$

Finally, recalling the relations between the velocity potential and stream function to the physical velocity components, i.e. Eqs. (6.3) and (5.17), the derivative of w if usually called the *complex velocity*, i.e.

$$\frac{dw}{dz}=u-\jmath v=Ve^{-\jmath\alpha} \tag{6.166}$$

Thus, provided that the partial derivatives of ϕ and ψ are continuous, a necessary, but not sufficient, condition for $f(z)$ to be analytic is

$$\frac{\partial \phi}{\partial x} = \frac{\partial \psi}{\partial y} \qquad \text{and} \qquad \frac{\partial \phi}{\partial y} = -\frac{\partial \psi}{\partial x} \tag{6.163}$$

These are called the *Cauchy-Riemann equations*. Points on the complex plane where these equations are not satisfied are called *singular points*.

6.6.3 Complex Potential

For ideal fluid flow, it is convenient to identify the aforementioned real variables with the velocity potential, ϕ, and the stream function, ψ. For then, differentiation of the first of Eq. (6.163) with respect to x, the second with respect to y, and adding the results, leads to exactly the two-dimensional form of the Laplace equation for the velocity potential, i.e. Eq. (6.15). Similarly, differentiation of the first of Eq. (6.163) with respect to y, the second with respect to x, and subtracting the results, leads to the Laplace equation for the stream function, i.e. Eq. (6.17).

Because the velocity potential, ϕ, and the stream function, ψ, satisfy the Cauchy-Riemann equations, they may also be interpreted as the real and imaginary parts of an analytic function called the *complex potential*, i.e.

$$w(z) = \phi + \jmath\psi \tag{6.164}$$

The rate of change of w with respect to z is called the *complex velocity*. Since w is a function of z only, its derivative can be determined by first computing its total differential, as follows

$$dw = \frac{\partial \phi}{\partial x}dx + \frac{\partial \phi}{\partial x}dy + \jmath\frac{\partial \psi}{\partial x}dx + \jmath\frac{\partial \psi}{\partial y}dy$$

The partial derivatives with respect to y can be substituted using the Cauchy-Riemann equations, i.e. Eq. (6.163), leading to

$$dw = \frac{\partial \phi}{\partial x}(dx + \jmath dy) + \jmath\frac{\partial \psi}{\partial x}dx\,(dx + \jmath dy)$$

Therefore, the derivative of the complex potential is given by

$$\frac{dw}{dz} = \frac{\partial \phi}{\partial x} + \jmath\frac{\partial \psi}{\partial x} \tag{6.165}$$

Finally, recalling the relations between the velocity potential and stream function to the physical velocity components, i.e. Eqs. (6.3) and (5.17), the derivative of w if usually called the *complex velocity*, i.e.

$$\frac{dw}{dz} = u - \jmath v = Ve^{-\jmath\alpha} \tag{6.166}$$

where V is the magnitude of the physical, two-dimensional velocity vector, and α is the angle that it makes with the x axis.

The conditions of irrotationality and incompressibility are automatically satisfied by the complex velocity. This is evident since the closed integral of w' on the complex plane vanishes, as it is comprised of the circulation around, and the flux through the path of integration on the physical plane, i.e.

$$\oint_C \frac{dw}{dz}\, dz = \oint_C \mathbf{V} \cdot \mathbf{t} ds + \oint_C \mathbf{V} \cdot \mathbf{n} ds = 0 \tag{6.167}$$

where $\mathbf{V}$ is the physical velocity vector, and $\mathbf{t}$ and $\mathbf{n}$ are the tangent and normal unit vectors to the path of integration C.

6.6.4 Conformal Transformations

The relations between the complex variables z and w allow a graphical transformation of ideal flow problems from the $x - y$ plane to the $\phi - \psi$ plane. These are commonly called the z and ζ planes, respectively, where ζ is an analytic function of z, i.e.

$$\zeta = f(z) \tag{6.168}$$

Thus, provided that z_0 is some non-singular point on the z plane, the transformation to the ζ plane is called a *conformal mapping* because infinitesimal configurations on the two planes are similar, and obey the relation

$$\delta\zeta = f'(z_0)\, \delta z \tag{6.169}$$

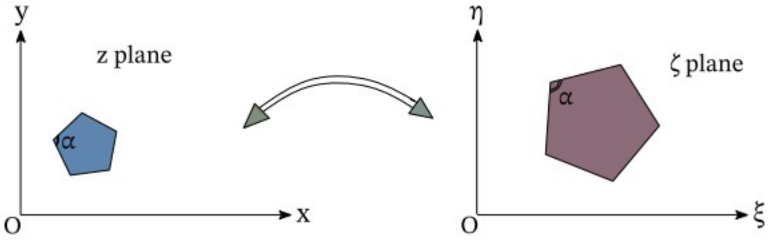

FIGURE 6.29 Schematic of conformal mapping

The goal of conformal mapping is to transform a geometrically complicated region on the z plane to a simpler region on the ζ plane. For example, a wedge shaped boundary may be transformed into the semi-infinite plane, where the flow pattern can be established easier. Thus, as shown in Fig. 6.29, images may be stretched and rotated, as they are transferred to the ζ plane, but always remain geometrically similar. This implies that conformal maps preserve the angles of geometric elements between the z and ζ planes. If an angle of an isolated point is changed in the transformation, that point is singular, and must be excluded from the map. However, singularities on the z plane are mapped to corresponding singularities on the ζ plane (Howe, 2006).

For an ideal fluid, a convenient analytic function for conformal mapping is the complex potential, $w(z)$, and its counterpart on the ζ plane

$$W(\zeta) = \Phi(\xi, \eta) + \jmath \Psi(\xi, \eta) \tag{6.170}$$

Then, the analytic property of Eq. (6.168) implies that

$$w(z) = W\left[f(z)\right] = \phi(x, y) + \jmath \psi(x, y) \tag{6.171}$$

Therefore, we must have

$$\Phi(\xi, \eta) = \phi(x, y) \qquad \text{and} \qquad \Psi(\xi, \eta) = \psi(x, y)$$

It follows that corresponding streamlines and equipotential lines on the z and ζ planes carry the same level set values although their shapes may be different. Furthermore, the flow net on the z plane is always transformed to a rectangular grid on the ζ plane, with its grid lines parallel to the ϕ and ψ axes, as shown in Fig. 6.30.

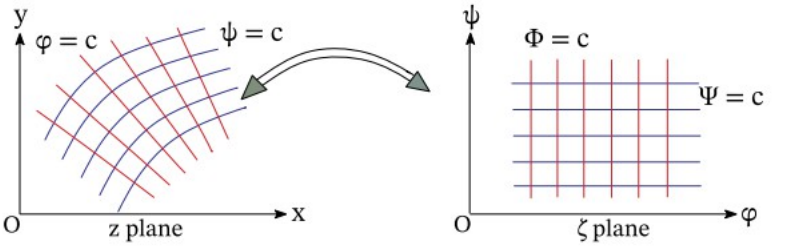

FIGURE 6.30 Conformal mapping of flow net

As a result, the flow on the ζ plane always corresponds to uniform flow directed along the ϕ axis. Since the flow between streamlines is constant, the transformation preserves the flow rate, and therefore the challenge is to find a mapping operator that can transform the uniform flow on the ζ plane to any desired flow on the z plane. The desired transformation operator is none other than the derivative of the complex potential, which according to the Cauchy-Riemann equations is uniquely defined at every point of the z plane where $\frac{dw}{dz} \neq 0$. Then, the general mapping of Eq. (6.169) can now be written as follows

$$\delta\zeta = \frac{dw}{dz}\,\delta z \tag{6.172}$$

In addition to assigning the same level set values to streamlines and equipotential lines, conformal maps also preserve the streamline property of solid boundaries. This is true because on the z plane, $\psi = C$ on the wall, where C is some arbitrary constant. The conformal transformation preserves this constant, regardless of the new orientation of the streamline, thus a solid wall always remains a solid wall.

When solving boundary-value problems on the ζ plane, essential boundary conditions are automatically satisfied because the values of ϕ and Φ are identical. When the values of normal derivatives are prescribed as natural boundary conditions, conformal mapping suggests that (Rouse, 1959)

$$\frac{\partial \phi}{\partial n} = \left| \frac{dw}{dz} \right| \frac{\partial \Phi}{\partial \nu} \tag{6.173}$$

where ν is the normal to the boundary in the ζ plane. A similar transformation is needed to map the complex velocity on the ζ plane, i.e.

$$\frac{dw}{dz} = \frac{dW}{d\zeta}\frac{d\zeta}{dz} \tag{6.174}$$

Finally, it should be mentioned that line sources and vortices map into identical sources and vortices on the ζ plane. The flow features correspond to singularities on both the z and ζ planes; however the strength of both sources and vortices is preserved during the conformal transformation, as it will be shown in the next section.

6.6.5 Power-Law Mapping

Any differential function may be used to accomplish the transform $w = f(z)$. Some simple examples are presented next corresponding to power-law functions of the form $w = Az^n$, where A is in general complex.

6.6.5.1 Uniform Stream; $n = 1$

Consider the uniform flow at an angle α to the x axis, as described in section 6.5.2. Therefore

$$\phi = Ur\cos(\theta - \alpha) \qquad \psi = Ur\sin(\theta - \alpha)$$

The simplest power law transformation employs an exponent of unity for z, thus we may define $z = re^{\dot{\jmath}\theta}$. Then, the complex potential can be written as follows

$$w = Uze^{-\dot{\jmath}\alpha} \tag{6.175}$$

Thus, elements on the z plane are elongated by a factor U and rotated by an angle $-\alpha$ when mapped to the ζ plane. This makes the problem marginally easier to solve on the ζ plane although more advantages can be found with higher exponents.

'

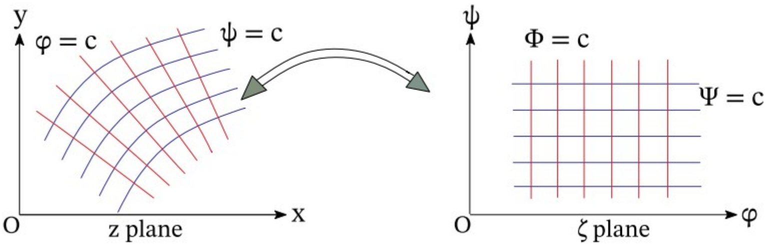
y
φ = c
ψ = c
O
z plane
x
ψ
Φ = c
Ψ = c
O
ζ plane
φ

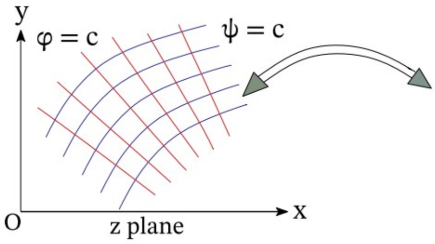
y
φ = c
ψ = c
O
z plane
x

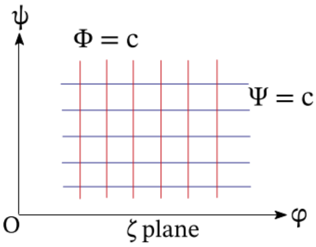
ψ
Φ = c
Ψ = c
O
ζ plane
φ

M $# # % % %

R? ') % " " R? E ,
* "

! "   "   #$ $ %&   '   (   )
"   "   "   #$ $!) #
+,-   *   *.   *   +// 0   '
1 2   3 4   & /"
(%%! %)5

APPENDIX A

GAUSSIAN QUADRATURE FORMULAS FOR CONSTANT WEIGHT FUNCTION

Here we give values of the $A_k^{(n)}$ and $x_k^{(n)}$ which make the approximation

$$\int_{-1}^{1} f(x)\,dx \approx \sum_{k=1}^{n} A_k^{(n)} f(x_k^{(n)})$$

exact whenever $f(x)$ is a polynomial of degree $\leq 2n - 1$. These formulas are discussed in Section 7.2. The $A_k^{(n)}$ and $x_k^{(n)}$ are symmetric with respect to $x = 0$:

$$A_k^{(n)} = A_{n-k+1}^{(n)}, \qquad x_k^{(n)} = -x_{n-k+1}^{(n)}$$

and the tables give only the values corresponding to $0 \leq x_k^{(n)} \leq 1$.

The tabulated values are taken from

H. J. Gawlik, "Zeros of Legendre Polynomials of orders 2-64 and weight coefficients of Gauss quadrature formulae," Armament Research and Development Establishment, Memorandum (B) 77/58. Fort Halstead, Kent, 1958.

where the $A_k^{(n)}$ and $x_k^{(n)}$ are given to 20 decimal places. Values for n = 2, 4, 8, 16, 24, 32, 40, 48, 64, 80, 96 of the same accuracy are given in the two tables

P. Davis and P. Rabinowitz, "Abscissas and weights for Gaussian quadratures of high order," *J. Res. Nat. Bur. Standards*, Vol. 56, 1956, pp. 35–37.

P. Davis and P. Rabinowitz, "Additional abscissas and weights for Gaussian quadrature of high order: values for n = 64, 80 and 96," *J. Res. Nat. Bur. Standards*, Vol. 60, 1958, pp. 613–14.

These three tables are the most extensive values of the $A_k^{(n)}$ and $x_k^{(n)}$ which are known and are believed to be accurate to within a few units in the last significant figures.

$x_k^{(n)}$	$A_k^{(n)}$
$n = 2$	
0.57735 02691 89625 76451	1.00000 00000 00000 00000
$n = 3$	
0.77459 66692 41483 37704	0.55555 55555 55555 55556
0.00000 00000 00000 00000	0.88888 88888 88888 88889
$n = 4$	
0.86113 63115 94052 57522	0.34785 48451 37453 85737
0.33998 10435 84856 26480	0.65214 51548 62546 14263

6 *+,-* * *. * 7 8 9 9 8 :
;/ : 1 2 3 4 <1 2 6 /"
( ) %

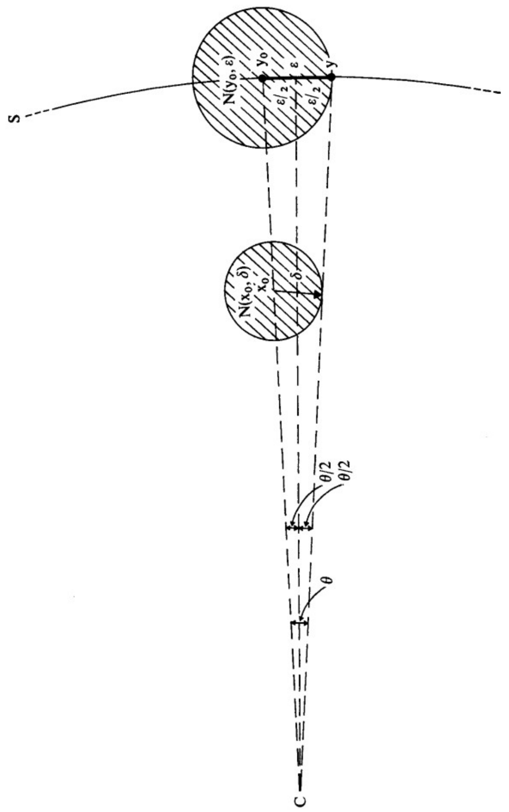

Fig. 16.3

tinuity cannot be established by the argument applying to exterior points.

The continuity of f may be established, however, as follows. Let x_0 be some point of $\mathbf{R}^3$ inside S, and let y_0 be the corresponding image $f(x_0)$ on S. Let a neighbourhood of y_0 be denoted by $N(y_0; \varepsilon)$ and let y be some point on the intersection of the sphere, having centre y_0 and radius ε, with the surface S. Now, let S be the perpendicular distance from x_0 to the radius c–y. The continuity of f at x_0 follows, since it is clear that for any ε the chosen δ will satisfy $f(N(x; \delta)) \subseteq N(y; \varepsilon)$, and since x_0 is an arbitrary point inside S, continuity follows for all points inside S, and hence also for all points belonging to $\mathbf{R}^3 - c$. The functional relationship between δ and ε is given by:

$$\begin{aligned}\delta &= d(x_0, c)\cdot\sin\theta\\ &= 2\cdot d(x_0, c)\cdot\sin\frac{\theta}{2}\cos\frac{\theta}{2}\\ &= 2\cdot d(x_0, c)\cdot\frac{\varepsilon}{2}\sqrt{\left(1-\frac{\varepsilon^2}{4}\right)}\\ &= d(x_0, c)\cdot\sqrt{\left(\varepsilon^2-\frac{\varepsilon^4}{4}\right)}\end{aligned}$$

Since topology is concerned with the study of properties which remain invariant under certain permitted transformations, and since these transformations are defined as *continuous functions having continuous inverses*, the establishment of continuity by the methods discussed in this chapter is an important stage in the investigation of topological invariants. However, these methods are dependent upon the existence of a metric, and so can apply to metric spaces only. A still more general approach is needed, and this will require continuity to be defined without recourse to any metric so that sets representing non-metric spaces may be embraced. It has already been seen that neighbourhoods are a particular kind of *open set*, and it is this concept of *open set* which provides the basis for a definition of continuity not requiring any reference to a metric. Such a general approach will be developed in the next chapter.

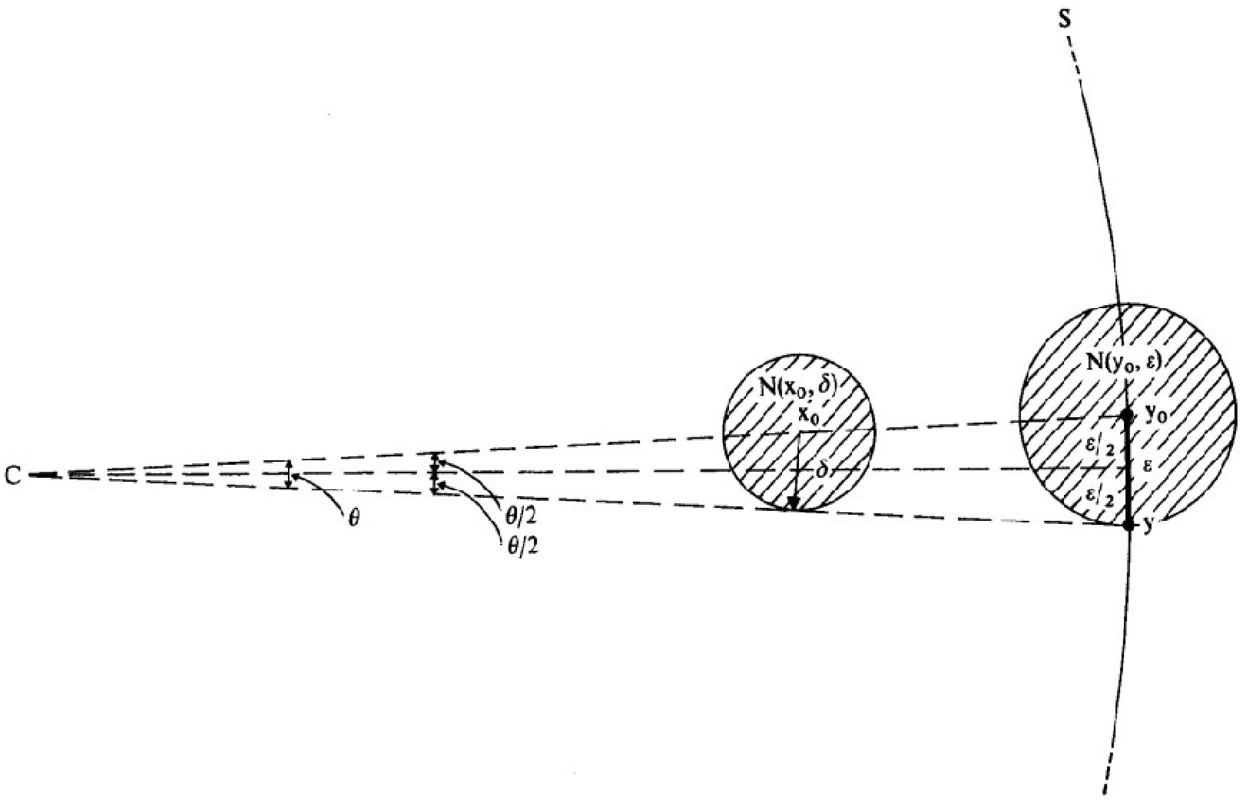
S
N(x_0, δ)
x_0
δ
N(y_0, ε)
y_0
ε/2
ε
ε/2
y
C
θ
θ/2
θ/2

6 *+,-* * *. * 9 ; <
/ /"

9 & %

16

Metric Spaces

Distance in $\mathbf{R}^n$—definition of metric—neighbourhoods—continuity in terms of neighbourhoods—complete system of neighbourhoods—requirement for proof of non-continuity—functional relationships between δ and ε—limitations of metric.

In Chapter 13 it was seen that for three-dimensional Euclidean space, the distance between two points P_0, P_1 satisfies the conditions:

$$d(P_0, P_1) \geqq 0;$$

$$d(P_0, P_1) = 0 \text{ if and only if } P_0 = P_1;$$

$$d(P_0, P_1) = d(P_1, P_0);$$

$$d(P_0, P_1)+d(P_1, P_2) \geqq d(P_0, P_2).$$

Generally, for x and y belonging to $\mathbf{R}^n$, the distance $d(x, y)$ may be defined as

$$d(x, y) = \sqrt{[(x_1-y_1)^2+ \dots +(x_n-y_n)^2]},$$

where the n-tuples $(x_1, \dots, x_n)$, $(y_1, \dots, y_n)$ are the coordinates of x and y respectively. For any points x, y, z of $\mathbf{R}^n$, distance satisfies the conditions:

$$d(x, y) \geqq 0;$$

$$d(x, y) = 0 \text{ if and only if } x = y;$$

$$d(x, y) = d(y, x);$$

$$d(x, y)+d(y, z) \geqq d(x, z).$$

It is not, however, necessary to confine the concept of distance to n-dimensional Euclidean space. If X is any non-empty set for which a non-negative real number $d(x, y)$ is defined for all elements $x, y \in X$ satisfying the four conditions above, then d is termed a *metric* and the set X together with the metric d is termed a *metric space*. The metric d is a function $X \times X \to \mathbf{R}_0^+$. A simple example of a metric space is provided by any non-empty set X together with a metric d defined by

$$d(x, y) = \begin{cases} 0 \text{ if } x = y \\ a > 0 \text{ if } x \neq y \end{cases}$$

for all $x, y \in X$. It can easily be seen that the four conditions for a metric are satisfied by d as defined above.

Another example providing a set of metric spaces is the set of all n-figure binary numbers with the metric d defined as the number of changes of digits required in going from one number to another. Thus if in the set of five-figure binary numbers $x = 01101$ and $y = 10111$, then $d(x, y) = 3$, since there is a change in each of the first, second and fourth digit places (reading from left to right).

In Chapter 13, it was seen that for any point $x_0 \in \mathbf{R}$, the set of all points x of $\mathbf{R}$ satisfying

$$d(x_0, x) < \delta$$

(where d is the usual metric in $\mathbf{R}$) is termed a *neighbourhood* (or *open ball*) of x_0. Clearly, a different neighbourhood of x_0 is defined for each chosen value of δ. The subset satisfying $d(x_0, x) < \rho$ is termed the neighbourhood of x_0 of *radius* ρ and denoted by $N(x_0; \rho)$.

Figure 16.1 depicts a point x_0 belonging to a subset X of $\mathbf{R}^2$, where X is the set of all points within and on the continuous closed curve C. The neighbourhood of x_0 in $\mathbf{R}^2$ of radius ρ is represented by the circle with centre x_0 and radius r, excluding the circumference. The neighbourhood of x_0 in X of radius ρ is represented by that part of the interior of the circle which is shaded, including interior points of the circle belonging to C. Clearly,

$$N(x_0 \in X; \rho) = X \cap N(x_0 \in \mathbf{R}^2; \rho),$$

and a similar relationship will hold for any point $x \in X \subset \mathbf{R}^2$.

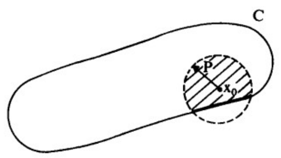

Fig. 16.1

In the case of the non-empty set X and the metric d defined by

$$d(x, y) = \begin{cases} 0 \text{ if } x = Y \\ a > 0 \text{ if } x \neq y, \end{cases}$$

the neighbourhood in X of any point $x \in X$ of radius a consists of the single point x itself, since any point other than x is defined to have distance a from x. In the case of the set of five-figure binary numbers with d defined as the number of changes of digits in going from one number to another, the neighbourhood of 01101 of radius 3 is the subset

$$\{01101,\ 11101,\ 00101,\ 01001,\ 01111,\ 01100,\ 10101,\ 11001,\ 11111,\ 11100,\ 00001,\ 00111,\ 00100,\ 01011,\ 01000,\ 01110\}.$$

This is the subset of all five-figure binary numbers differing by zero, one, or two digits from 01101. The concept of a neighbourhood of given radius may thus be generalised to sets other than subsets of $\mathbf{R}^n$.

The definition of continuity of functions, given in Chapter 13, may now be re-expressed in terms of neighbourhoods where functions are mapping one metric space to another. If X, Y are two metric spaces with metrics d, e respectively, and if $x_0 \in X$, then a function $f: X \to Y$ is said to be *continuous at* x_0 if, for each neighbourhood of $f(x_0)$ in Y, there is some neighbourhood of x_0 in X whose image is in the neighbourhood of $f(x_0)$ in Y. Symbolically, if for every radius ε there is a radius δ such that

$$f(N(x \in X;\delta)) \subseteq N(f(x) \in Y;\varepsilon),$$

then f is continuous at x_0. If this holds for all $x \in X$, then the function is continuous at each and every point of its domain and is said simply to be *continuous*. Thus 'continuous' means 'continuous everywhere'.

The term *open ball* for a neighbourhood of given radius arises from a consideration of neighbourhoods in $\mathbf{R}^3$. Such a neighbourhood in $\mathbf{R}^3$ has a radius defining its boundary and consists of the set of all the points *within* this boundary. Such a set is an open set, and a given point of $\mathbf{R}^3$ together with the given radius defines the interior of a sphere. Hence the term 'open ball' (rather than *open sphere*) is used to emphasise that it is not the surface of the sphere which is being considered. A neighbourhood of a general nature is simply a subset of points containing a given point together with some open ball about that point. The set of all neighbourhoods of a given point $x \in X$ is termed the *complete system of neighbourhoods* of x.

An example of a function which is continuous over a certain subset of $\mathbf{R}^2$ but not continuous everywhere was depicted in Figure 15.16 (page 148). The same function is depicted in Figure 16.2 where $X = \{(x, y): x < 0\}$, and $Y = \{(x, y): x \geqq 0\}$. The images $f(X)$, $f(Y)$ are as depicted in the figure. If $a = (0, a)$ is some point of the subset $\{(0, y)\}$, then $a' = f(a) = (-1, a)$ is a point of the subset $\{(-1, y)\}$. If $0 < \varepsilon < 2$, then the neighbourhood of a' in $\mathbf{R}^2$, $N(a';\varepsilon)$ contains no points of $f(X)$. However, all neighbourhoods of a in $\mathbf{R}^2$ do contain points belonging to X. There is thus no δ corresponding to ε such that

$$f(N(a;\delta)) \subseteq N(a';\varepsilon),$$

and the function is therefore not continuous at a. Since a was taken as an arbitrary point of $\{(0, y)\}$ it follows that f is nowhere continuous on $\{(0, y)\}$. If, however, any point of X or Y is chosen not belonging to $\{(0, y)\}$, there will always be some δ, however small, satisfying the continuity condition. The function is thus continuous everywhere in $\mathbf{R}$ except for the subset $\{(0, y)\}$.

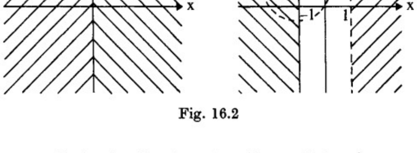

Fig. 16.2

To prove that a function is **not** continuous it is only necessary to find some ε for which no δ exists satisfying the continuity condition. Proving that a function **is** continuous is more difficult because it is necessary to establish that some δ exists for every point of its domain and every possible choice of ε. This may sometimes be done by finding a suitable functional relationship between δ and ε. In certain cases it is possible to take δ equal to ε, the simplest example of which is the identity function $f: \mathbf{R}^n \to \mathbf{R}^n$ for which $f(x) = x$. Other examples are provided by the various rigid transformations. In the case of similarity transformations δ may be taken equal to ε if the transformation is one of contraction. But, if the function increases distances by a factor a greater than unity it is necessary to take $\delta = \varepsilon/a$.

An example of finding an appropriate functional relationship between δ and ε is provided by consideration of the continuity of the radial projection $f: \mathbf{R}^3 - c \to S$, which projects all points of $\mathbf{R}^3$ except c on to the surface S of a sphere of unit radius and centre c. This is depicted in Figure 16.3. The distance between any two points exterior to the sphere is shrunk by the function f, and the distance between any two points on the surface S remains unchanged. Hence, f is continuous if its domain is restricted to elements x of $\mathbf{R}^3$ for which

$$d(x, c) \geqq 1.$$

For pairs of points inside S, however, the function f magnifies distances more and more as the points become closer and closer to c, and con-

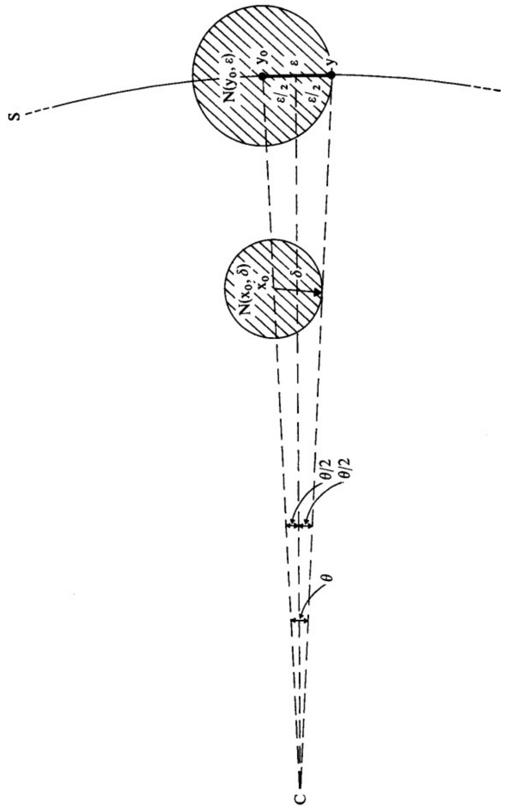

Fig. 16.3

tinuity cannot be established by the argument applying to exterior points.

The continuity of f may be established, however, as follows. Let x_0 be some point of $\mathbf{R}^3$ inside S, and let y_0 be the corresponding image $f(x_0)$ on S. Let a neighbourhood of y_0 be denoted by $N(y_0; \varepsilon)$ and let y be some point on the intersection of the sphere, having centre y_0 and radius ε, with the surface S. Now, let S be the perpendicular distance from x_0 to the radius c–y. The continuity of f at x_0 follows, since it is clear that for any ε the chosen δ will satisfy $f(N(x; \delta)) \subseteq N(y; \varepsilon)$, and since x_0 is an arbitrary point inside S, continuity follows for all points inside S, and hence also for all points belonging to $\mathbf{R}^3 - c$. The functional relationship between δ and ε is given by:

$$\begin{aligned}\delta &= d(x_0, c)\cdot\sin\theta \\ &= 2\cdot d(x_0, c)\cdot\sin\frac{\theta}{2}\cos\frac{\theta}{2} \\ &= 2\cdot d(x_0, c)\cdot\frac{\varepsilon}{2}\sqrt{\left(1-\frac{\varepsilon^2}{4}\right)} \\ &= d(x_0, c)\cdot\sqrt{\left(\varepsilon^2-\frac{\varepsilon^4}{4}\right)}\end{aligned}$$

Since topology is concerned with the study of properties which remain invariant under certain permitted transformations, and since these transformations are defined as *continuous functions having continuous inverses*, the establishment of continuity by the methods discussed in this chapter is an important stage in the investigation of topological invariants. However, these methods are dependent upon the existence of a metric, and so can apply to metric spaces only. A still more general approach is needed, and this will require continuity to be defined without recourse to any metric so that sets representing non-metric spaces may be embraced. It has already been seen that neighbourhoods are a particular kind of *open set*, and it is this concept of *open set* which provides the basis for a definition of continuity not requiring any reference to a metric. Such a general approach will be developed in the next chapter.

6 *+,-* * *. * 9 ; <
/ /"

data 03.04.2020

- β) Books_63_2020/
From Geometry to Topology.../
H. Graham Flegg - From
Geometry to topology... pdf

Pág. 154 (104) - Metric Space

Fig 16.3

radial projection

The distance between any two points exterior to the sphere is shrunk by the function f, and the distance between any two points on the surface S remains unchanged.

c) Books_63_2020/
Laminar Boundary Layers/
sanet.st_12184.pdf

Pág. 457 (490)

20. Parabolic flow over
a flate plate

Pág. 458 (491) →

Fig. VIII. 7. Family of parabolic translates as streamlines of a curved external flow over a flate plate.

data 03.04.2020

- β) Books_63_2020/
From Geometry to Topology.../
H. Graham Flegg - From
Geometry to Topology... pdf

Pág. 154 (104) - Metric Space

Fig 16.3

radial projection...

The distance between any two points exterior to the sphere is shrunk by the function f, and the distance between any two points on the surface S remains unchanged.

C) Books_63_2020/
Laminar Boundary Layers/
sanet.st_12184.pdf

Pág. 457 (490)

20. Parabolic flow over a flate plate

Pág. 458 (491) →

Fig. VIII. 7. Family of parabolic translates as streamlines of a curved external flow over a flate plate.

data 03.04.2020

Equations:

$$y = \lambda + x\left(B - \frac{1}{2} c x\right), \quad (156)$$

$$B = \cot \theta_0, \quad (157)$$

where

and B and c are constant for a given system of translates.

D) Books_63_2020/ Voronoi Diagrams and Delaunay Triangulations/

Jean Gallier... Aspects of Convex Geometry - Polyhedra ....pdf

Pág. 100

Fig 4.11. The polar dual a polygon

Obs: Pág. 98

Fig. 4.10: The polar, $a^{\dagger}$, of a point, a, outside the sphere S^{n-1}.

Obs: Pág. 96 4.3 Polarity and Duality.

+,- * *. * 5) /"
( ) !)# 6 2 /

If there is no secondary flow in the layer, then $v_2 = 0$ everywhere. It follows therefore from equation (18) that

$$\frac{v_1^2}{h_1}\frac{\partial h_1}{\partial x_2} = \frac{1}{\rho}\frac{\partial p}{\partial x_2},$$

and if κ is the curvature of a streamline at any point,

$$\kappa = (h_1 h_2)^{-1}(\partial h_1/\partial x_2),$$

so that the above relation becomes

$$\kappa v_1^2 h_2 = \frac{1}{\rho}\frac{\partial p}{\partial x_2}. \tag{154}$$

Now the right-hand side of this relation is independent of x_3, as is also κ, whereas v_1 is a non-constant function of x_3 by virtue of the boundary conditions. It follows therefore that relation (154) cannot be true unless $\kappa = 0$ everywhere, in which case the streamlines are straight. Hence curvature of the streamlines in planes parallel to the plate implies the existence of secondary flow in the boundary layer.

In the case of steady flow over a curved surface, a special feature is associated with geodesic curves on the surface. L. C. Squire (1955) has shown that for streamlines which are geodesics on the surface the boundary-layer flow parallel to the surface is in the direction of the main stream. Let $x_2 =$ const be such a streamline, so that $\partial h_1/\partial x_2 = 0$ since it is a geodesic. Equations (18) and (21) then give

$$\frac{v_1}{h_1}\frac{\partial v_2}{\partial x_1} + \frac{v_2}{h_2}\frac{\partial v_2}{\partial x_2} + v_3\frac{\partial v_2}{\partial x_3} + \frac{v_1 v_2}{h_1 h_2}\frac{\partial h_2}{\partial x_1} = \nu\frac{\partial^2 v_2}{\partial x_3^2}, \tag{155}$$

since $V_2 = 0$. The boundary conditions on v_2 are $v_2 = 0$ when $x_3 = 0$ and $v_2 \to 0$ as $x_3 \to \infty$, and hence the required solution of equation (155) is $v_2 = 0$. Conversely, if v_2 and V_2 are zero on any curve, that curve must be a geodesic, for in this case equations (18) and (21) give

$$(h_1 h_2)^{-1}(V_1^2 - v_1^2)(\partial h_1/\partial x_2) = 0,$$

which is true for all values of x_3. Thus, since $V_1 \neq v_1$ for all values of x_3, it follows that $\partial h_1/\partial x_2 = 0$; that is, the curve is a geodesic.

20. Parabolic flow over a flat plate

An example of secondary flow in a boundary layer, arising from curvature of streamlines in the main-stream flow, was discussed by Sowerby (1954) and by Loos (1955). The case is that of the boundary layer on a flat plate at zero incidence; the flow in the main stream is steady, and the streamlines lie in planes parallel to the plate and form a system of parabolic translates. The common axis of the system is taken parallel to the leading edge of the plate.

A system of rectangular Cartesian coordinates $Oxyz$ may be chosen, with the plate forming the half-plane $z = 0$, $x \geqslant 0$. Thus, with the introduction of parameters as indicated in Fig. VIII. 7, the family of parabolic translates is represented by the equation

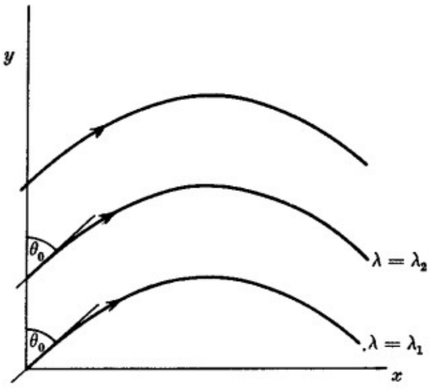

FIG. VIII. 7. Family of parabolic translates as streamlines of a curved external flow over a flat plate.

$$y = \lambda + x(\beta - \tfrac{1}{2}cx), \tag{156}$$

where
$$\beta = \cot\theta_0, \tag{157}$$

and β and c are constant for a given system of translates. Variation of the parameter λ gives different members of the family, while variation of c changes the scale of the parabolas; c may be regarded therefore as a parameter which determines the curvature of the streamlines, and in particular the streamlines are straight when $c = 0$.

The Cartesian components of the main-stream flow are

$$U = \text{constant}, \qquad V = U(\beta - cx), \tag{158}$$

so that the main-stream flow is rotational, and there is a spanwise pressure gradient. The flow everywhere is independent of y, and if u, v, w are the components of velocity in the boundary layer the equations of motion are

$$u\frac{\partial u}{\partial x} + w\frac{\partial u}{\partial z} = \nu\frac{\partial^2 u}{\partial z^2}, \tag{159}$$

$$u\frac{\partial v}{\partial x} + w\frac{\partial v}{\partial z} = -cU^2 + \nu\frac{\partial^2 v}{\partial z^2}, \tag{160}$$

$$\frac{\partial u}{\partial x} + \frac{\partial w}{\partial z} = 0, \tag{161}$$

with boundary conditions

$$u = v = w = 0 \quad \text{when } z = 0 \quad (x > 0),$$
$$u \to U, \quad v \to U(\beta - cx) \quad \text{when } z \to \infty \text{ or } x \to 0. \tag{162}$$

It is clear that the required solution of the above equations provides an extension of the Blasius solution for two-dimensional uniform flow over a flat plate (Section V. 12), since the latter is the case $\beta = 0$, $c = 0$; also the case of uniform flow across a sheared plate is obtained by taking $c = 0$. In the general case the additional term in the equations of motion is the term cU^2 in equation (160), which clearly arises from the curvature of the streamlines.

The solution is obtained by introducing Prandtl's transformation

$$\eta = \tfrac{1}{2}(U/\nu x)^{\frac{1}{2}}z, \tag{163}$$

and writing

$$u = \tfrac{1}{2}Uf'(\eta), \qquad v = U\{\beta g(\eta) - cxh(\eta)\}, \tag{164}$$

where f, g, h are functions of η only. Then all the equations of motion and boundary conditions are satisfied provided

$$w = \tfrac{1}{2}(U\nu/x)^{\frac{1}{2}}(\eta f' - f), \tag{165}$$

and f, g, h satisfy the equations

$$f''' + ff'' = 0, \tag{166}$$

$$g'' + fg' = 0, \tag{167}$$

$$h'' + fh' - 2f'h + 4 = 0, \tag{168}$$

with boundary conditions

$$f = f' = g = h = 0 \quad \text{when } \eta = 0,$$
$$f' \to 2, \quad g \to 1, \quad h \to 1 \text{ when } \eta \to \infty. \tag{169}$$

The required solution of equation (166) was obtained by Blasius, and the appropriate solution of equation (167) is

$$g = \tfrac{1}{2}f'. \tag{170}$$

The remaining equation (168) is linear, and was solved numerically by Sowerby with the use of Howarth's values (1938) for the coefficients f and f'. Loos has shown, however, that the solution is related to a solution of another equation given previously by Mager and Hansen (1952). The asymptotic form of (168) is of some interest, since the equation can be transformed, for large values of η, to a special case of Weber's equation. Numerical values of h and $h - \frac{1}{2}f'$ are given in Table VIII. 4.

The expressions for the components of velocity in the boundary layer are easily obtained. If the component along a streamline is denoted by

TABLE VIII. 4

Functions for the components of velocity in parabolic flow over a flat plate, calculated from equations (166) to (168)

η	h	$h-\frac{1}{2}f'$	η	h	$h-\frac{1}{2}f'$
0	0·0	0·0	2·1	1·0284	0·0615
0·1	0·2637	0·1973	2·2	1·0210	0·0452
0·2	0·4879	0·3552	2·3	1·0153	0·0326
0·3	0·6743	0·4754	2·4	1·0109	0·0231
0·4	0·8252	0·5604	2·5	1·0076	0·0161
0·5	0·9433	0·6135	2·6	1·0053	0·0110
0·6	1·0322	0·6384	2·7	1·0035	0·0074
0·7	1·0956	0·6393	2·8	1·0023	0·0048
0·8	1·1373	0·6205	2·9	1·0015	0·0032
0·9	1·1611	0·5864	3·0	1·0010	0·0020
1·0	1·1708	0·5410	3·1	1·0006	0·0013
1·1	1·1697	0·4884	3·2	1·0004	0·0008
1·2	1·1608	0·4318	3·3	1·0003	0·0005
1·3	1·1469	0·3744	3·4	1·0002	0·0003
1·4	1·1300	0·3185	3·5	1·0001	0·0002
1·5	1·1120	0·2659	3·6	1·0001	0·0002
1·6	1·0941	0·2180	3·7	1·0001	0·0001
1·7	1·0772	0·1755	3·8	1·0001	0·0001
1·8	1·0621	0·1387	3·9	1·0000	0·0001
1·9	1·0488	0·1077	4·0	1·0000	0·0000
2·0	1·0377	0·0821			

v_1 and the component normal to it by v_n, so that v_n is the secondary flow and is measured positive towards the centre of curvature, the results are

$$v_1 = U\{1+(\beta-cx)^2\}^{-\frac{1}{2}}\{\tfrac{1}{2}f'+(\beta-cx)(\tfrac{1}{2}\beta f'-cxh)\}, \tag{171}$$

$$v_n = Ucx\{1+(\beta-cx)^2\}^{-\frac{1}{2}}\{h-\tfrac{1}{2}f'\}. \tag{172}$$

The third component w, normal to the plate, is given by result (165).

It is clear that the secondary flow here is always directed towards the centre of curvature of the streamlines. As an illustration of its magnitude, consider the case where $\beta = 1$ and $c > 0$. Then when $cx = 2$,

$$v_1 = V_1 h \quad \text{and} \quad v_n = V_1(h-\tfrac{1}{2}f'),$$

where V_1 is the main-stream speed for this value of cx. These two profiles are shown in Fig. VIII. 8, and it will be seen that v_1 also departs from a normal profile. It appears that v_1 attains a maximum value (when $\eta = 1{\cdot}0$) at about $1{\cdot}17V_1$, and the secondary flow v_n a maximum value (when $\eta = 0{\cdot}7$) at about $0{\cdot}64V_1$. For the same value of cx the change in direction of flow through the boundary layer is shown in Fig. VIII. 9; the changes are seen to be considerable. Further illustrations may be found in the paper by Loos.

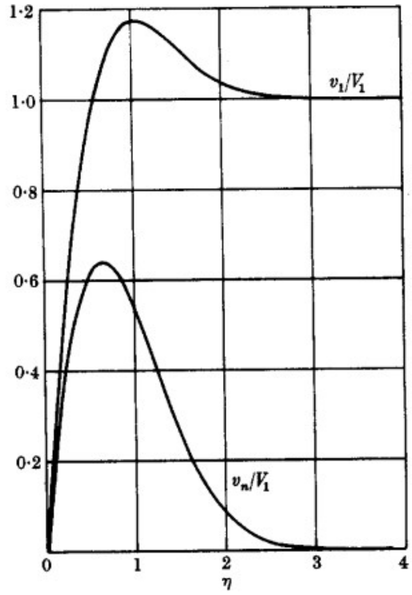

FIG. VIII. 8. Velocity components in the boundary layer on a flat plate with a parabolic main stream, for the values $\beta = 1$, $cx = 2$. The component v_n is the secondary flow in the layer.

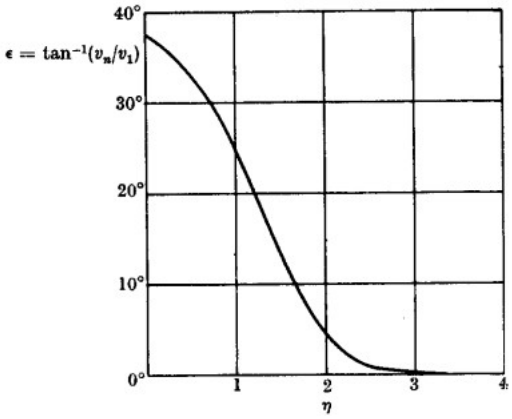

FIG. VIII. 9. Parabolic flow over a flat plate. ϵ is the deviation of the direction of flow in the boundary layer from an external streamline, for the values $\beta = 1$, $cx = 2$.

21. Flow near a stagnation point on a general curved surface

The discussion of the flow in the neighbourhood of a stagnation point on a general surface was given by Howarth (1951*c*). If A is a stagna-

data 03.04.2020

S T Q Q [✓ S] S D

• β) Books_63_2020/
From Geometry to Topology.../
H. Graham Flegg - From
Geometry to Topology... pdf

Pág. 154 (104) - Metric Space

Fig 16.3

radial projection...

The distance between any two points exterior to the sphere is shrunk by the function f, and the distance between any two points on the surface S remains unchanged.

c) Books_63_2020/
Laminar Boundary Layers/
sanet.st_12184.pdf

Pág. 457 (490)

20. Parabolic flow over a flate plate

Pág. 458 (491) →

Fig. VIII. 7. Family of parabolic translates as streamlines of a curved external flow over a flate plate.

data 03.04.2020

Equations:

$$y = \lambda + x\left(B - \frac{1}{2} c x\right), \qquad (156)$$

$$B = \cot \theta_0, \qquad (157)$$

where

and B and c are constant for a given system of translates.

D) Books_63_2020/ Voronoi Diagrams and Delaunay Triangulations/

Jean Gallier... Aspects of Convex Geometry - Polyhedra ....pdf

Pág. 100

Fig 4.11. The polar dual a polygon

Obs: Pág. 98

Fig. 4.10: The polar, $a^{\dagger}$, of a point, a, outside the sphere S^{n-1}.

Obs: Pág. 96 4.3 Polarity and Duality.

"   *+,-*   *   *.   *   =   8   + /   2 0
8   :>   : " < ?   <   <   1   < 1   : ;
@   A /"
(#&) %   : " 1   :

Beware that Proposition 4.19 is false when the dimension of X is infinite and when $\overset{\circ}{A} = \emptyset$.

The proposition below gives a sufficient condition for a closed subset to be convex.

Proposition 4.20. *Let A be a closed subset with nonempty interior. If there is a supporting hyperplane for every point $a \in \partial A$, then A is convex.*

Proof. We leave it as an exercise (see Berger [8], Proposition 11.5.4). ☐

The condition that A has nonempty interior is crucial!

The proposition below characterizes closed convex sets in terms of (closed) half–spaces. It is another intuitive fact whose rigorous proof is nontrivial.

Proposition 4.21. *Let A be a nonempty closed and convex subset. Then, A is the intersection of all the closed half–spaces containing it.*

Proof. Let A' be the intersection of all the closed half–spaces containing A. It is immediately checked that A' is closed and convex and that $A \subseteq A'$. Assume that $A' \neq A$, and pick $a \in A' - A$. Then, we can apply Corollary 4.10 to $\{a\}$ and A and we find a hyperplane, H, strictly separating A and $\{a\}$; this shows that A belongs to one of the two half-spaces determined by H, yet a does not belong to the same half-space, contradicting the definition of A'. ☐

4.3 Polarity and Duality

Let $E = \mathbb{E}^n$ be the Euclidean affine space of dimension n. We will denote the origin $(0, \ldots, 0)$ in $\mathbb{E}^n$ by O. We know that the inner product on $E = \mathbb{E}^n$ induces a duality between E and its dual E^* (for example, see Chapter 6, Section 2 of Gallier [30]), namely, $u \mapsto \varphi_u$, where φ_u is the linear form defined by $\varphi_u(v) = u \cdot v$, for all $v \in E$. For geometric purposes, it is more convenient to recast this duality as a correspondence between points and hyperplanes, using the notion of polarity with respect to the unit sphere, $S^{n-1} = \{a \in \mathbb{E}^n \mid \|\mathbf{Oa}\| = 1\}$.

First, we need the following simple fact: For every hyperplane H not passing through O, there is a *unique* point h, so that

$$H = \{a \in \mathbb{E}^n \mid \mathbf{Oh} \cdot \mathbf{Oa} = 1\}.$$

Indeed, any hyperplane H in $\mathbb{E}^n$ is the null set of some equation of the form

$$\alpha_1 x_1 + \cdots + \alpha_n x_n = \beta,$$

and if $O \notin H$, then $\beta \neq 0$. Thus, any hyperplane H not passing through O is defined by an equation of the form

$$h_1 x_1 + \cdots + h_n x_n = 1,$$

if we set $h_i = \alpha_i/\beta$. So, if we let $h = (h_1, \ldots, h_n)$, we see that

$$H = \{a \in \mathbb{E}^n \mid \mathbf{Oh} \cdot \mathbf{Oa} = 1\},$$

as claimed. Now, assume that

$$H = \{a \in \mathbb{E}^n \mid \mathbf{Oh_1} \cdot \mathbf{Oa} = 1\} = \{a \in \mathbb{E}^n \mid \mathbf{Oh_2} \cdot \mathbf{Oa} = 1\}.$$

The functions $a \mapsto \mathbf{Oh_1} \cdot \mathbf{Oa} - 1$ and $a \mapsto \mathbf{Oh_2} \cdot \mathbf{Oa} - 1$ are two affine forms defining the same hyperplane, so there is a nonzero scalar λ so that

$$\mathbf{Oh_1} \cdot \mathbf{Oa} - 1 = \lambda(\mathbf{Oh_2} \cdot \mathbf{Oa} - 1) \quad \text{for all } a \in \mathbb{E}^n$$

(see Gallier [30], Chapter 2, Section 2.10). In particular, for $a = O$, we find that $\lambda = 1$, and so,

$$\mathbf{Oh_1} \cdot \mathbf{Oa} = \mathbf{Oh_2} \cdot \mathbf{Oa} \quad \text{for all } a,$$

which implies $h_1 = h_2$. This proves the uniqueness of h.

Using the above, we make the following definition:

Definition 4.2. Given any point $a \neq O$, the *polar hyperplane of a (w.r.t. S^{n-1})* or *dual of a* is the hyperplane $a^\dagger$ given by

$$a^\dagger = \{b \in \mathbb{E}^n \mid \mathbf{Oa} \cdot \mathbf{Ob} = 1\}.$$

Given a hyperplane H not containing O, the *pole of H (w.r.t S^{n-1})* or *dual of H* is the (unique) point $H^\dagger$ so that

$$H = \{a \in \mathbb{E}^n \mid \mathbf{OH}^\dagger \cdot \mathbf{Oa} = 1\}.$$

We often abbreviate polar hyperplane to polar. We immediately check that $a^{\dagger\dagger} = a$ and $H^{\dagger\dagger} = H$, so we obtain a bijective correspondence between $\mathbb{E}^n - \{O\}$ and the set of hyperplanes not passing through O.

When a is outside the sphere S^{n-1}, there is a nice geometric interpetation for the polar hyperplane $H = a^\dagger$. Indeed, in this case, since

$$H = a^\dagger = \{b \in \mathbb{E}^n \mid \mathbf{Oa} \cdot \mathbf{Ob} = 1\}$$

and $\|\mathbf{Oa}\| > 1$, the hyperplane H intersects S^{n-1} (along an $(n-2)$-dimensional sphere) and if b is any point on $H \cap S^{n-1}$, we claim that $\mathbf{Ob}$ and $\mathbf{ba}$ are orthogonal. This means that $H \cap S^{n-1}$ is the set of points on S^{n-1} where the lines through a and tangent to S^{n-1} touch S^{n-1} (they form a cone tangent to S^{n-1} with apex a). Indeed, as $\mathbf{Oa} = \mathbf{Ob} + \mathbf{ba}$ and $b \in H \cap S^{n-1}$ i.e., $\mathbf{Oa} \cdot \mathbf{Ob} = 1$ and $\|\mathbf{Ob}\|^2 = 1$, we get

$$1 = \mathbf{Oa} \cdot \mathbf{Ob} = (\mathbf{Ob} + \mathbf{ba}) \cdot \mathbf{Ob} = \|\mathbf{Ob}\|^2 + \mathbf{ba} \cdot \mathbf{Ob} = 1 + \mathbf{ba} \cdot \mathbf{Ob},$$

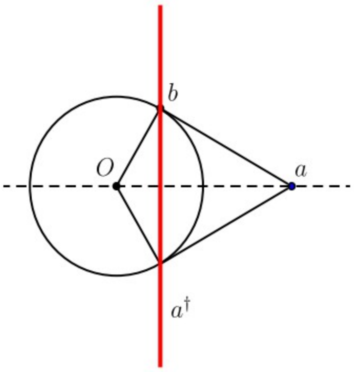

Figure 4.10: The polar, $a^\dagger$, of a point, a, outside the sphere S^{n-1}

which implies $\mathbf{ba} \cdot \mathbf{Ob} = 0$. When $a \in S^{n-1}$, the hyperplane $a^\dagger$ is tangent to S^{n-1} at a.

Also, observe that for any point $a \neq O$, and any hyperplane H not passing through O, if $a \in H$, then, $H^\dagger \in a^\dagger$, i.e, the pole $H^\dagger$ of H belongs to the polar $a^\dagger$ of a. Indeed, $H^\dagger$ is the unique point so that

$$H = \{b \in \mathbb{E}^n \mid \mathbf{OH}^\dagger \cdot \mathbf{Ob} = 1\}$$

and

$$a^\dagger = \{b \in \mathbb{E}^n \mid \mathbf{Oa} \cdot \mathbf{Ob} = 1\};$$

since $a \in H$, we have $\mathbf{OH}^\dagger \cdot \mathbf{Oa} = 1$, which shows that $H^\dagger \in a^\dagger$.

If $a = (a_1, \ldots, a_n)$, the equation of the polar hyperplane $a^\dagger$ is

$$a_1X_1 + \cdots + a_nX_n = 1.$$

Remark: As we noted, polarity in a Euclidean space suffers from the minor defect that the polar of the origin is undefined and, similarly, the pole of a hyperplane through the origin does not make sense. If we embed $\mathbb{E}^n$ into the projective space, $\mathbb{P}^n$, by adding a "hyperplane at infinity" (a copy of $\mathbb{P}^{n-1}$), thereby viewing $\mathbb{P}^n$ as the disjoint union $\mathbb{P}^n = \mathbb{E}^n \cup \mathbb{P}^{n-1}$, then the polarity correspondence can be defined everywhere. Indeed, the polar of the origin is the hyperplane at infinity ($\mathbb{P}^{n-1}$) and since $\mathbb{P}^{n-1}$ can be viewed as the set of hyperplanes through the origin in $\mathbb{E}^n$, the pole of a hyperplane through the origin is the corresponding "point at infinity" in $\mathbb{P}^{n-1}$.

Now, we would like to extend this correspondence to subsets of $\mathbb{E}^n$, in particular, to convex sets. Given a hyperplane, H, not containing O, we denote by H_- the closed half-space containing O.

Definition 4.3. Given any subset A of $\mathbb{E}^n$, the set

$$A^* = \{b \in \mathbb{E}^n \mid \mathbf{Oa} \cdot \mathbf{Ob} \leq 1, \quad \text{for all } a \in A\} = \bigcap_{\substack{a \in A \\ a \neq O}} (a^\dagger)_-,$$

is called the *polar dual* or *reciprocal* of A.

For simplicity of notation, we write $a^\dagger_-$ for $(a^\dagger)_-$. Observe that $\{O\}^* = \mathbb{E}^n$, so it is convenient to set $O^\dagger_- = \mathbb{E}^n$, even though $O^\dagger$ is undefined. By definition, A^* is convex even if A is not. Furthermore, note that

(1) $A \subseteq A^{**}$.

(2) If $A \subseteq B$, then $B^* \subseteq A^*$.

(3) If A is convex and closed, then $A^* = (\partial A)^*$.

It follows immediately from (1) and (2) that $A^{***} = A^*$. Also, if $B^n(r)$ is the (closed) ball of radius $r > 0$ and center O, it is obvious by definition that $B^n(r)^* = B^n(1/r)$.

In Figure 4.11, the polar dual of the polygon $(v_1, v_2, v_3, v_4, v_5)$ is the polygon shown in green. This polygon is cut out by the half-planes determined by the polars of the vertices $(v_1, v_2, v_3, v_4, v_5)$ and containing the center of the circle. These polar lines are all easy to determine by drawing for each vertex, v_i, the tangent lines to the circle and joining the contact points. The construction of the polar of v_3 is shown in detail.

Remark: We chose a different notation for polar hyperplanes and polars ($a^\dagger$ and $H^\dagger$) and polar duals (A^*), to avoid the potential confusion between $H^\dagger$ and H^*, where H is a hyperplane (or $a^\dagger$ and $\{a\}^*$, where a is a point). Indeed, they are completely different! For example, the polar dual of a hyperplane is either a line orthogonal to H through O, if $O \in H$, or a semi-infinite line through O and orthogonal to H whose endpoint is the pole, $H^\dagger$, of H, whereas, $H^\dagger$ is a single point! Ziegler ([67], Chapter 2) use the notation $A^\triangle$ instead of A^* for the polar dual of A.

We would like to investigate the duality induced by the operation $A \mapsto A^*$. Unfortunately, it is not always the case that $A^{**} = A$, but this is true when A is closed and convex, as shown in the following proposition:

Proposition 4.22. *Let A be any subset of $\mathbb{E}^n$ (with origin O).*

(i) If A is bounded, then $O \in \overset{\circ}{A^*}$; if $O \in \overset{\circ}{A}$, then A^* is bounded.

(ii) If A is a closed and convex subset containing O, then $A^{**} = A$.

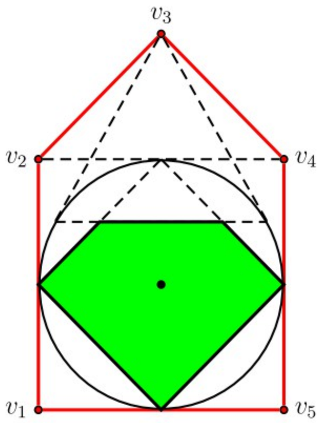

Figure 4.11: The polar dual of a polygon

Proof. (i) If A is bounded, then $A \subseteq B^n(r)$ for some $r > 0$ large enough. Then, $B^n(r)^* = B^n(1/r) \subseteq A^*$, so that $O \in \overset{\circ}{A^*}$. If $O \in \overset{\circ}{A}$, then $B^n(r) \subseteq A$ for some r small enough, so $A^* \subseteq B^n(r)^* = B^r(1/r)$ and A^* is bounded.

(ii) We always have $A \subseteq A^{**}$. We prove that if $b \notin A$, then $b \notin A^{**}$; this shows that $A^{**} \subseteq A$ and thus, $A = A^{**}$. Since A is closed and convex and $\{b\}$ is compact (and convex!), by Corollary 4.10, there is a hyperplane, H, strictly separating A and b and, in particular, $O \notin H$, as $O \in A$. If $h = H^\dagger$ is the pole of H, we have

$$\mathbf{Oh} \cdot \mathbf{Ob} > 1 \quad \text{and} \quad \mathbf{Oh} \cdot \mathbf{Oa} < 1, \quad \text{for all } a \in A$$

since $H_- = \{a \in \mathbb{E}^n \mid \mathbf{Oh} \cdot \mathbf{Oa} \leq 1\}$. This shows that $b \notin A^{**}$, since

$$\begin{aligned} A^{**} &= \{c \in \mathbb{E}^n \mid \mathbf{Od} \cdot \mathbf{Oc} \leq 1 \quad \text{for all } d \in A^*\} \\ &= \{c \in \mathbb{E}^n \mid (\forall d \in \mathbb{E}^n)(\text{if} \quad \mathbf{Od} \cdot \mathbf{Oa} \leq 1 \quad \text{for all } a \in A, \quad \text{then} \quad \mathbf{Od} \cdot \mathbf{Oc} \leq 1)\}, \end{aligned}$$

just let $c = b$ and $d = h$. ☐

Remark: For an arbitrary subset $A \subseteq \mathbb{E}^n$, it can be shown that $A^{**} = \overline{\mathrm{conv}(A \cup \{O\})}$, the topological closure of the convex hull of $A \cup \{O\}$.

Proposition 4.22 will play a key role in studying polytopes, but before doing this, we need one more proposition.

Proposition 4.23. *Let A be any closed convex subset of $\mathbb{E}^n$ such that $O \in \overset{\circ}{A}$. The polar hyperplanes of the points of the boundary of A constitute the set of supporting hyperplanes of A^*. Furthermore, for any $a \in \partial A$, the points of A^* where $H = a^\dagger$ is a supporting hyperplane of A^* are the poles of supporting hyperplanes of A at a.*

Proof. Since $O \in \overset{\circ}{A}$, we have $O \notin \partial A$, and so, for every $a \in \partial A$, the polar hyperplane $a^\dagger$ is well-defined. Pick any $a \in \partial A$ and let $H = a^\dagger$ be its polar hyperplane. By definition, $A^* \subseteq H_-$, the closed half-space determined by H and containing O. If T is any supporting hyperplane to A at a, as $a \in T$, we have $t = T^\dagger \in a^\dagger = H$. Furthermore, it is a simple exercise to prove that $t \in (T_-)^*$ (in fact, $(T_-)^*$ is the interval with endpoints O and t). Since $A \subseteq T_-$ (because T is a supporting hyperplane to A at a), we deduce that $t \in A^*$, and thus, H is a supporting hyperplane to A^* at t. By Proposition 4.22, as A is closed and convex, $A^{**} = A$; it follows that all supporting hyperplanes to A^* are indeed obtained this way. ☐

D

03.04.2020
8

E) Books_63_2020/
Advanced Differential Quadrature Methods/

Advanced Differential Quadrature Methods ... pdf

Pág. 98(122)
3.3 Geometric transformation
Pág. 99
Fig. 3.5: Points for quintic transformation of a curvilinear triangle.
Pág. 103(126)
102(125) -
3.4.4 Symmetry Consideration
All elliptic plates on Pasternak foundation are assumed subjected to uniformly distributed load.
Pág. 103(127)
Fig. 3.6 A quadrant of an elliptic plate.

data 03 04 2020

S T Q Q [✓ S] S D

Pág. 104 (128)

Table 3.2:

Central deflection and central moment of thin circular plates on a Pasternak foundation with clamped or simply supported edges.

Pág. 105 (129)

Fig 3.7: Deflections of the circular plate

Fig 3.8: Radial bending moments of the circular plate.

Pág. 273 (297)

Chapter 8

Mathematical Compendium

8.1. Gauss elimination

Pág. 281 (305)

8.5 Complex Analysis

8.5.1 Complex Variable

D *+,-* * *. * +"2 " 1
- " " / E 7 > , +// " E F /
E 7 /"
+,- * *. * +"2 " " G "
" , # /"
(#5 % % 8 ;

3.3 Geometric transformation

The geometric transformation of a curvilinear triangle into the standard triangle is usually based on the following point transformation expression

$$x = \sum_{j=0}^{N} \sum_{i=0}^{N-j} f_{\overline{pqr}} x_{\overline{pqr}}, \quad y = \sum_{j=0}^{N} \sum_{i=0}^{N-j} f_{\overline{pqr}} y_{\overline{pqr}} \tag{3.36}$$

where N is the order of transformation; $(x_{\overline{pqr}}, y_{\overline{pqr}})$ are the Cartesian coordinates of the transformation point in the curvilinear triangle, $f_{\overline{pqr}}$ are functions given in the similar form as Eq. (3.30). The three digit subscripts $(\overline{p}, \overline{q}, \overline{r}) = (N - j - i, i, j)$ are introduced to identify the nodes in the transformation. Unlike in finite element analysis where piecewise lower order isoparametric transformations suffice to provide accurate modelling of curved domain boundaries, a global transformation of the curvilinear triangle onto the standard triangle is sought herein. In principle, high order transformations in Eq. (3.36) should be able to model the curved boundary adequately. In practice, however, difficulty may arise since the transformation is rather sensitive to the location of inner transformation points as reported in Mitchell and Wait (1977), Zhong (2002). Ideally, the difficulty can be circumvented by employing transformations leaving out inner transformation points. Fortunately, this can be realized by adopting the "serendipity" idea in finite element method. Thus, a triangular serendipity transformation is developed as

$$x = \sum_{j=1}^{15} f_j x_j, \; y = \sum_{j=1}^{15} f_j y_j \tag{3.37}$$

to a space variable at a grid point is approximated by a weighted linear sum of the function values at grid points in a local small triangle containing the grid point in question rather than at all the grid points in the entire triangular domain. Analogous to Eq. (3.28), the expression of a partial derivative at a grid point (α, β, γ) is changed in

$$D_n \{f(\xi, \eta)\}_{\alpha\beta\gamma} = \sum_{j=0}^{m'} \sum_{i=0}^{m'-j} C^{(n)}_{\alpha\beta\gamma,pqr} f_{pqr}$$

$$\alpha + \beta + \gamma = m, \quad (p, q, r) = (m' - i - j, i, j) \tag{3.35}$$

where m'(¡m) is the number of the equal segments of an edge in a local triangle containing the grid point (α, β, γ). The weighting coefficients $C^{(n)}_{\alpha\beta\gamma,pqr}$ are also determined in the local triangle.

3.3 Geometric transformation

The geometric transformation of a curvilinear triangle into the standard triangle is usually based on the following point transformation expression

$$x = \sum_{j=0}^{N} \sum_{i=0}^{N-j} f_{\overline{pqr}} x_{\overline{pqr}}, \quad y = \sum_{j=0}^{N} \sum_{i=0}^{N-j} f_{\overline{pqr}} y_{\overline{pqr}} \tag{3.36}$$

where N is the order of transformation; $(x_{\overline{pqr}}, y_{\overline{pqr}})$ are the Cartesian coordinates of the transformation point in the curvilinear triangle, $f_{\overline{pqr}}$ are functions given in the similar form as Eq. (3.30). The three digit subscripts $(\overline{p}, \overline{q}, \overline{r}) = (N - j - i, i, j)$ are introduced to identify the nodes in the transformation. Unlike in finite element analysis where piecewise lower order isoparametric transformations suffice to provide accurate modelling of curved domain boundaries, a global transformation of the curvilinear triangle onto the standard triangle is sought herein. In principle, high order transformations in Eq. (3.36) should be able to model the curved boundary adequately. In practice, however, difficulty may arise since the transformation is rather sensitive to the location of inner transformation points as reported in Mitchell and Wait (1977), Zhong (2002). Ideally, the difficulty can be circumvented by employing transformations leaving out inner transformation points. Fortunately, this can be realized by adopting the "serendipity" idea in finite element method. Thus, a triangular serendipity transformation is developed as

$$x = \sum_{j=1}^{15} f_j x_j, \ y = \sum_{j=1}^{15} f_j y_j \tag{3.37}$$

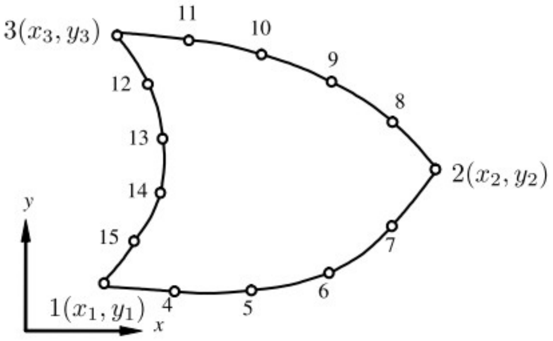

FIGURE 3.5: Points for quintic transformation of a curvilinear triangle.

It is well-known that the quintic order transformation often meets the accuracy requirement of geometric modelling. The shape functions associated with the fifteen nodes on the boundary of a curvilinear triangle (see Fig. 3.5) are given in Appendix A at the end of this chapter. It is worth noting that the high-order serendipity transformation for triangles is developed without the use of inner grid points, different from transformation for quadrilaterals. Thus, the present serendipity transformation enables the TDQM to deal with the problem on complex geometric domains.

The Jacobian matrix of the transformation is

$$[J] = \begin{bmatrix} x_\xi & y_\xi \\ x_\eta & y_\eta \end{bmatrix} \tag{3.38}$$

Its inverse matrix is given as follows:

$$\begin{bmatrix} \xi_x & \eta_x \\ \xi_y & \eta_y \end{bmatrix} = [J]^{-1} = \frac{1}{|J|} \begin{bmatrix} y_\eta & -y_\xi \\ -x_\eta & x_\xi \end{bmatrix}, \; |J| = x_\xi y_\eta - x_\eta y_\xi \tag{3.39}$$

where $|J|$ is the determinant of the Jacobian matrix. With the chain rule of differentiation, the second derivatives of a function in the two coordinate systems are related as

$$\begin{aligned}
\xi_{xx} &= |J|^{-3}\left[|J|\left(y_\eta y_{\xi\eta} - y_\xi y_{\eta\eta}\right) - y_\eta(y_\eta\left.|J|\right|_\xi - y_\xi\left.|J|\right|_\eta)\right] \\
\xi_{yy} &= |J|^{-3}\left[|J|\left(x_\eta x_{\xi\eta} - x_\xi x_{\eta\eta}\right) - x_\eta(x_\eta\left.|J|\right|_\xi - x_\xi\left.|J|\right|_\eta)\right] \\
\eta_{xx} &= |J|^{-3}\left[|J|\left(y_\xi y_{\xi\eta} - y_\eta y_{\xi\xi}\right) - y_\xi(y_\eta\left.|J|\right|_\xi - y_\xi\left.|J|\right|_\eta)\right] \\
\eta_{yy} &= |J|^{-3}\left[|J|\left(x_\xi x_{\xi\eta} - x_\eta x_{\xi\xi}\right) - x_\xi(x_\eta\left.|J|\right|_\xi - x_\xi\left.|J|\right|_\eta)\right] \qquad (3.40) \\
\xi_{xy} &= |J|^{-3}\left[|J|\left(x_\xi y_{\eta\eta} - x_\eta y_{\xi\eta}\right) - y_\eta(x_\eta\left.|J|\right|_\xi - x_\xi\left.|J|\right|_\eta)\right] \\
\eta_{xy} &= |J|^{-3}\left[|J|\left(x_\eta y_{\xi\xi} - x_\xi y_{\xi\eta}\right) - y_\xi(x_\eta\left.|J|\right|_\xi - x_\xi\left.|J|\right|_\eta)\right]
\end{aligned}$$

The above relations will be used later to transform the governing equations and boundary conditions from the $x - y$ domain into the $\xi - \eta$ domain.

3.4 Governing equations of Reissner-Mindlin plates on Pasternak foundation

The TMDQ coupled with geometric transformation is employed to deal with elliptic Reissner-Mindlin plates on Pasternak foundation under uniformly distributed load. Due to the double symmetry of the problem, a quadrant of the plate is considered in the analysis and a triangular serendipity transformation is adopted to map the quadrant into the computational domain−the unit isosceles right triangle.

3.4.1 Governing equations in physical coordinate system

For flexural analysis of Reissner plates on a Pasternak foundation, the differential equilibrium equations are given as follows (Kobayashi and Sonoda, 1989):

$$\frac{D}{2}\left((1-\nu)\nabla^2\psi_x + (1+\nu)\frac{\partial\phi}{\partial x}\right) + \kappa Gh\left(\frac{\partial w}{\partial x} - \psi_x\right) = 0, \qquad (3.41a)$$

$$\frac{D}{2}\left((1-\nu)\nabla^2\psi_y + (1+\nu)\frac{\partial\phi}{\partial y}\right) + \kappa Gh\left(\frac{\partial w}{\partial y} - \psi_y\right) = 0, \qquad (3.41b)$$

$$\kappa Gh\left(\nabla^2 w - \phi\right) + q - p = 0. \qquad (3.41c)$$

where

$$\phi = \frac{\partial\psi_x}{\partial x} + \frac{\partial\psi_y}{\partial y}, \qquad (3.42)$$

$$D = \frac{Eh^3}{12(1-\nu^2)} \tag{3.43}$$

$$p = k_f w - G_f \nabla^2 w \tag{3.44}$$

and $w(x,y)$ is the transverse deflection; $\psi_x(x,y)$ and $\psi_y(x,y)$ the rotations of the normal about $y-$axis and $x-$axis; D, E, G and ν the plate flexural rigidity, Young's modulus, shear modulus and Poisson's ratio, respectively; h is the plate thickness; $q(x,y)$ the surface load intensity; and ∇^2 is the Laplacian operator in Cartesian coordinates. The shear correction factor κ is taken to be $5/6$, p is the subgrade reaction intensity, k_f and G_f are the subgrade reaction coefficient and shear modulus of foundation, respectively. The force resultants are given by:

$$M_x = -D\left(\frac{\partial\psi_x}{\partial x} + \nu\frac{\partial\psi_y}{\partial y}\right) \tag{3.45}$$

$$M_y = -D\left(\nu\frac{\partial\psi_x}{\partial x} + \frac{\partial\psi_y}{\partial y}\right) \tag{3.46}$$

$$M_{xy} = -\frac{1-\nu}{2}D\left(\frac{\partial\psi_x}{\partial y} + \frac{\partial\psi_y}{\partial x}\right) \tag{3.47}$$

$$Q_x = \kappa Gh\left(\frac{\partial w}{\partial x} - \psi_x\right) \tag{3.48}$$

$$Q_y = \kappa Gh\left(\frac{\partial w}{\partial y} - \psi_y\right) \tag{3.49}$$

where M_x, M_y and M_{xy} are the moments and twisting moment; Q_x and Q_y the shear forces.

3.4.2 Governing equations in the standard triangle

By using the geometric coordinate mapping relations, the governing equations in Eqs. (3.41) in the physical domain $x-y$ can be transformed into the computational domain $\xi-\eta$ as follows:

$$[K]_{3\times 18}\begin{bmatrix} [\partial]_{6\times 1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & [\partial]_{6\times 1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & [\partial]_{6\times 1} \end{bmatrix}\begin{Bmatrix} \psi_x \\ \psi_y \\ w \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ q/D \end{Bmatrix} \tag{3.50}$$

where

$$[\partial]_{6\times 1} = \left\{ \tfrac{\partial^2}{\partial\xi^2} \; \tfrac{\partial^2}{\partial\xi\partial\eta} \; \tfrac{\partial^2}{\partial\eta^2} \; \tfrac{\partial}{\partial\xi} \; \tfrac{\partial}{\partial\eta} \; 1 \right\}^T \tag{3.51}$$

is a differential operator vector and the entries in $[K]_{3\times 18}$ are given in Appendix B. With the TDQ approximation, Eq. (3.50) will be recast into a set of algebraic equations at any inner grid point in the standard triangular domain.

3.4.3 Boundary conditions

To impose boundary conditions, the following transformations of displacement components and internal forces are used

$$\psi_n = n_1\psi_x + n_2\psi_y;\ \psi_s = -n_2\psi_x + n_1\psi_y. \tag{3.52}$$

$$\begin{aligned} M_n &= n_1^2 M_x + 2n_1 n_2 M_{xy} + n_2^2 M_y;\\ M_{ns} &= n_1 n_2(M_y - M_x) + (n_1^2 - n_2^2)M_{xy};\\ Q_n &= n_1 Q_x + n_2 Q_y. \end{aligned} \tag{3.53}$$

where (n_1, n_2) are the directional cosines of a unit normal at an arbitrary point on the boundary of the physical domain (see Fig. 3.6). Two typical boundary conditions are considered in this chapter, the clamped and the simply supported boundary conditions, which are mathematically expressed as follows:

(a) *Clamped (C)*

$$w = 0, \quad \psi_n = 0, \quad \psi_s = 0; \tag{3.54}$$

(b) *Simply supported (S):*

$$(i)\ \mathrm{Type\,I(S)} : \mathrm{w} = 0, \mathrm{M_n} = 0, \psi_\mathrm{s} = 0; \tag{3.55}$$

$$(ii)\ \mathrm{Type\,II(S')} : \mathrm{w} = 0, \mathrm{M_n} = 0, \mathrm{M_{ns}} = 0. \tag{3.56}$$

In Eqs. (3.53), (3.55) and (3.56), M_n, M_{ns} are the bending and twisting moments on the edge of the plate respectively; ψ_n and ψ_s are the rotations of the mid-plane in the normal plane, nz, and in the tangent plane, sz, to the boundary; the subscripts n and s represent the normal and tangent directions at a point on the boundary, respectively. Equation (3.55) is designated for plates with simply supported boundary conditions.

3.4.4 Symmetry consideration

In the present case, all elliptic plates on Pasternak foundation are assumed subjected to uniformly distributed load. Consequently, the flexural deformation symmetry of plates can be employed to reduce the computational effort

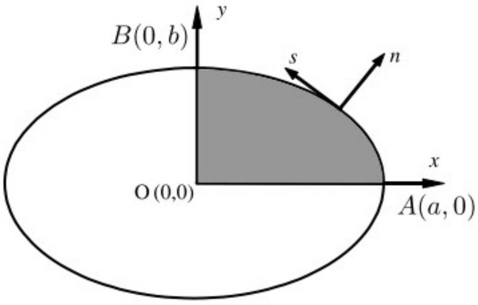

FIGURE 3.6: A quadrant of an elliptic plate.

and increase the computational accuracy. The deformation symmetry conditions of the Reissner/Mindlin plate about a symmetric plane are expressed as

$$\psi_n = 0, \quad \frac{\partial \psi_s}{\partial n} = 0, \quad \frac{\partial w}{\partial n} = 0 \tag{3.57}$$

By combining the boundary conditions and symmetric conditions at the boundary nodes with Eqs. (3.50) at all inner grid points, a set of algebraic equations with at most $3M$ unknowns is established.

Example 3.3 *Plates*

For clarity of presentation, the following normalized quantities are introduced

$$r_0 = \frac{rD}{qb^4}, w_0 = \frac{wD}{qb^4}, M_{r0} = \frac{M_r}{qb^2}, M_{x0} = \frac{M_x}{qb^2}, M_{y0} = \frac{M_y}{qb^2} \tag{3.58}$$

The following parameters are taken in all computations: flexural rigidity $D = 10^6$kNm, Poisson's ratio $\nu = 0.3$, subgrade reaction coefficient $k_f = 2 \times 10^4$kN/m^3. The foundation shear modulus is taken as $G_f = 3 \times 10^4$kN/m unless otherwise stated.

(a) Convergence study

First a circular plate resting on a Pasternak foundation under uniformly distributed load is studied to examine the convergence of the present method. The thickness to width ratio is taken as $h/b = 0.01$. Both simply supported and clamped boundary conditions are considered and the results are compared with the analytical ones (Yu, 1957).

In view of the double symmetry of the plate, a quadrant of the plate is analyzed. The central deflection and the central moment of circular plates are given in Table 3.2. It is found that results with relative error less than

TABLE 3.2: Central deflection and central moment of thin circular plates on a Pasternak foundation with clamped or simply supported edge

m	M	Clamped		Simply supported	
		ω_0	M_{r0}	ω_0	M_{r0}
14	120	0.004651	0.01735	0.005748	0.01088
15	136	0.004637	0.01775	0.005754	0.01022
16	153	0.004639	0.01772	0.005710	0.00971
17	171	0.004628	0.01762	0.005754	0.01059
18	190	0.004638	0.01784	0.005778	0.01079
19	210	0.004648	0.01794	0.005776	0.01058
20	231	0.004650	0.01792	0.005773	0.01046
21	253	0.004649	0.01788	0.005769	0.01047
22	276	0.004648	0.01788	0.005773	0.01051
23	300	0.004648	0.01789	0.005774	0.01048
24	325	0.004649	0.01789	0.005774	0.01047
25	351	0.004649	0.01789	0.005774	0.01049
26	378	0.004649	0.01789	0.005774	0.01050
27	406	0.004649	0.01790	0.005774	0.01049
28	435	0.004649	0.01790	0.005774	0.01049
29	465	0.004649	0.01790	0.005774	0.01049
30	496	0.004649	0.01790	0.005774	0.01049
Analytic[15]		0.004648	0.01791	0.005771	0.01050

1% can be achieved when the global approximation order m is increased to 18. It is seen clearly that the errors are reduced with the increase of the global approximation order. It is worth mentioning that the localized TDQ ($m' = 9$) is implemented in all computations for the sake of consistency and coding simplicity. As reported by Zhong et al., (2003), the loss of efficiency or accuracy is insignificant as long as the local approximation order is higher than 5. On the other hand, the localized implementation leads to remarkable numerical stability.

The radial variation of deflection and the radial variation of bending moment are shown in Figs. 3.7 and 3.8. The shear modulus of foundation are taken as the following different values $G_f = 0.3 \times 10^4$kN/m,2.82843×10^5kN/m and 3×10^6 kN/m in order to investigate the effects of the shear modulus on the behavior of plates. Excellent agreement with the available analytical solution (Yu, 1957) is achieved for both simply supported and clamped plates resting on Pasternak foundation as demonstrated in Figs. 3.7 and 3.8.

(b) Results for clamped and simply supported elliptic plates

To investigate the effect of thickness-to-width ratio, the central deflection and moments for clamped and simply supported elliptic plates are shown in Tables 3.3 and 3.4. Also listed is the least order of global approximation m from which results with four stabilized digits are obtained. It is clearly seen that more grid points are needed to attain convergent results for thin

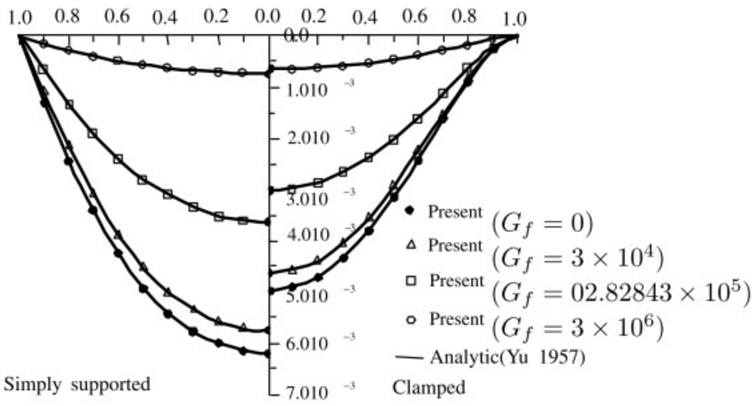

FIGURE 3.7: Deflections of the circular plate.

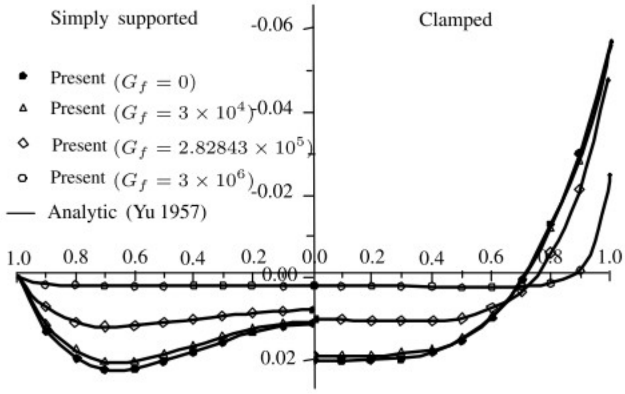

FIGURE 3.8: Radial bending moments of the circular plate.

plates. It is worth mentioning that $m + 1$ merely denotes the number of grid points on an edge of the triangular domain to furnish convergent results with four stabilized digits and a localized TDQ approximation of order $m' = 9$ is adopted in all cases. For clamped elliptic plates, the normalized deflection and the two moments decrease with the increase of thickness-to-width ratio except for circular plates. For simply supported plates, the normalized central deflection decreases with the increase of thickness-to-width ratio while the two moments at the center of elliptic plates show the reverse trend. It should be pointed out that the relative extremes of moments M_x and M_y may not occur at the center of the elliptic plates even for circular plates in the presence of elastic foundation, as shown in Fig. 3.8.

The effect of ellipticity on the deflection and moments is also investigated. Moderately thick elliptic plates ($h/b = 0.1$) with clamped or simply supported edges resting on Pasternak foundations are studied for plates with ellipticity $a/b = 1.5$, 2.0, 2.5, 3.0 and 5.0. The distributions of normalized deflection and moments along the major and the minor axes are plotted for simply supported and clamped boundary conditions (see Figs. 3.9–3.14). It is noted that the convergence rate drops with the increase of ellipticity a/b of plates. In addition, the convergence rate for simply supported elliptic plates is somewhat slower than that for fully clamped elliptic plates and the convergence rate of bending moments is slower than that of the deflection of plates. In the case of elliptic plates with $a/b = 5.0$, four significant digits are stabilized at $m = 20$ and $m = 27$ for the deflection and bending moments at the center of fully clamped elliptic plates, respectively; whereas $m = 22$ and $m = 28$ are required to keep the same level of convergence of the deflection and bending moments at the center of the simply supported elliptic plates.

3.5 Conclusions

The TDQM introduced in this chapter overcomes the difficulties associated with the direct DQ method in dealing with problems on a triangular domain. Transformation of the physical domain is avoided, resulting in simple algebraic equations. Singularity does not appear in the process of the TDQ analysis. In the two examples provided, excellent agreement with the theoretical values is achieved with rather fewer grid points, indicating that the present method is an attractive new numerical tool. In view of the great geometric flexibility of triangles, various geometric domains may be decomposed into a number of triangles. Therefore, the potential applications of the present method lie in the incorporation of domain decomposition technique into the TDQM.

Herein an effort has been made to extend the TDQM to problems with arbitrary geometric domains. A triangular serendipity transformation with

TABLE 3.3: Computed results at the center of clamped elliptic plates resting on a Pasternak foundation

	m	w_0	M_{x0}	M_{y0}
h/b		$a/b = 1.0$		
0.01	27	0.004649	0.01790	0.01789
0.05	18	0.004653	0.01769	0.01769
0.10	21	0.004668	0.01706	0.01706
0.15	14	0.004689	0.01611	0.01611
		$a/b = 2.0$		
0.01	35	0.005153	0.004976	0.01390
0.05	28	0.005149	0.004926	0.01375
0.10	23	0.005141	0.004782	0.01328
0.15	22	0.005129	0.004561	0.01256
		$a/b = 3.0$		
0.01	40	0.005111	0.004170	0.01314
0.05	28	0.005108	0.004126	0.01300
0.10	25	0.005102	0.003993	0.01255
0.15	23	0.005093	0.003790	0.01187

TABLE 3.4: Computed results at the center of simply supported elliptic plates resting on a Pasternak foundation

	m	w_0	M_{x0}	M_{y0}
h/b		$a/b = 1.0$		
0.01	26	0.005774	0.01049	0.01048
0.05	16	0.005765	0.01051	0.01051
0.10	16	0.005745	0.01053	0.01053
0.15	14	0.005714	0.01056	0.01056
		$a/b = 2.0$		
0.01	37	0.005571	0.001425	0.005602
0.05	29	0.005564	0.001449	0.005666
0.10	29	0.005551	0.001524	0.005775
0.15	27	0.005531	0.001641	0.005940
		$a/b = 3.0$		
0.01	39	0.005494	0.001520	0.005233
0.05	30	0.005487	0.001534	0.005299
0.10	29	0.005476	0.001573	0.005398
0.15	28	0.005458	0.001634	0.005549

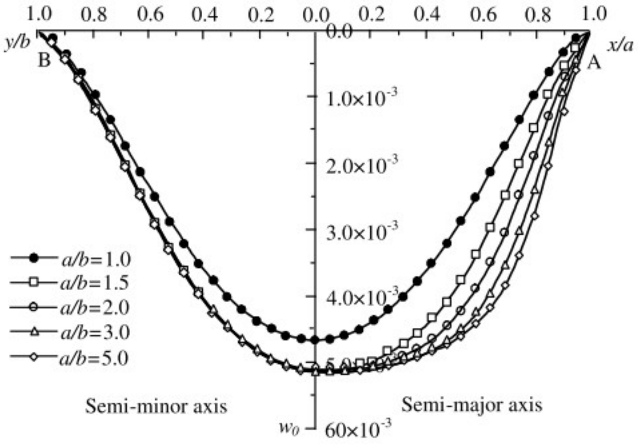

FIGURE 3.9: Deflections of clamped elliptic plates.

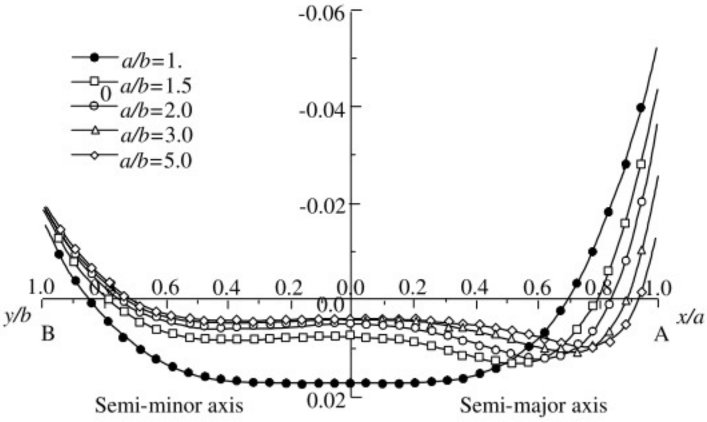

FIGURE 3.10: Bending moment M_{x0} of clamped elliptic plates.

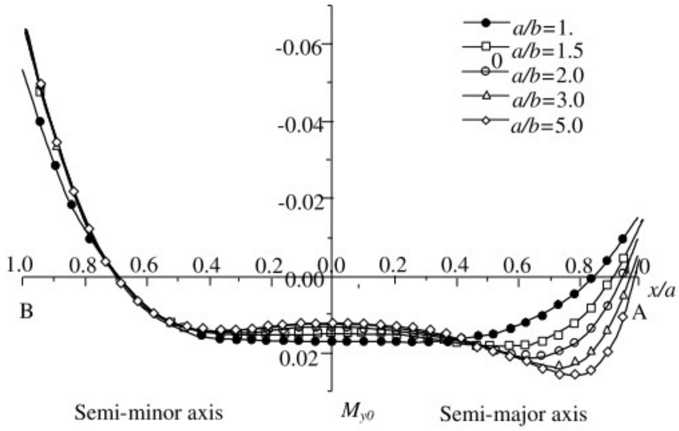

FIGURE 3.11: Bending moment M_{y0} of clamped elliptic platese.

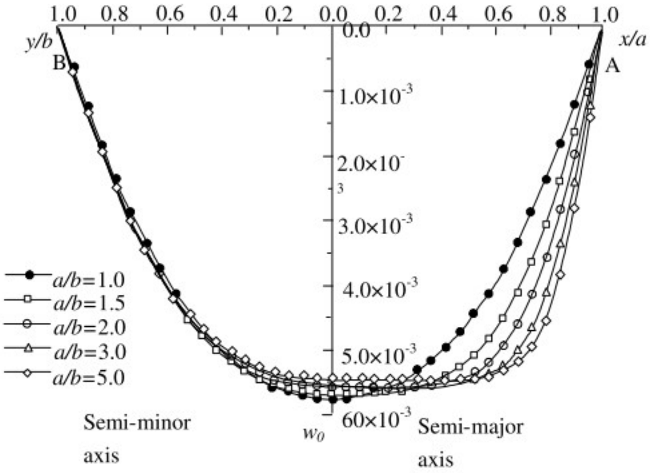

FIGURE 3.12: Deflections of simply supported elliptic plates.

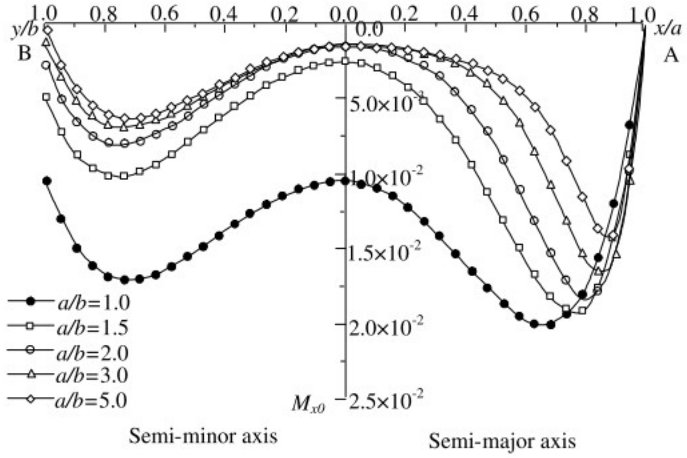

FIGURE 3.13: Bending moment M_{x0} of simply supported elliptic plates.

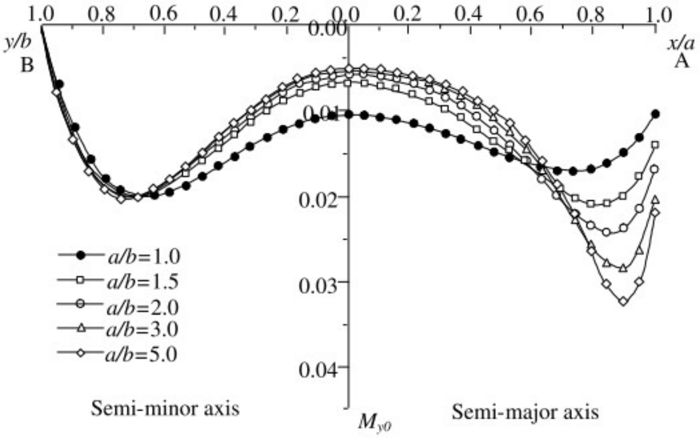

FIGURE 3.14: Bending moment M_{y0} of simply supported elliptic plates.

no inner points is introduced. Consequently, the sensitivity associated with the usual geometric transformations with inner points is removed. Although the triangular serendipity transformation of any order can be developed in the same manner, the quintic order transformation is accurate enough in all cases. Static flexural analysis of elliptic Reissner-Mindlin plates resting on a Pasternak foundation is investigated using the TDQM. Parametric study of elliptic plates with various aspect ratios is performed. The excellent agreement with the available analytical solution for circular plates verifies the effectiveness of the present technique.

Of course, the present strategy may only be applied successfully to the problems with simple and relatively regular geometry. However, it should be pointed out that the geometric transformation is not the only way to make TDQM applicable to problems with curvilinear boundaries. By using the localized TDQM (Zhong et al., 2003), the TDQM can be applied directly to the problems with curvilinear boundaries without the use of geometric transformation. In view of this, the TDQM is more attractive in the analysis of problems with irregular geometric domain.

Appendix A

Although the selection of transformation points on the boundary of a curvilinear domain is a simple matter, uniform distribution is desirable to avoid severe distortion. The functions for the transformation (see Fig. 3.5) are given as

$$f_4 = \frac{3125}{24} L_1 L_2 (L_1 - \frac{1}{5})(L_1 - \frac{2}{5})(L_1 - \frac{3}{5}), \tag{A1}$$

$$f_5 = -\frac{3125}{12} L_1 L_2 (L_1 - \frac{1}{5})(L_1 - \frac{2}{5})(L_1 - \frac{4}{5}) \tag{A2}$$

$$f_6 = \frac{3125}{12} L_1 L_2 (L_1 - \frac{1}{5})(L_1 - \frac{3}{5})(L_1 - \frac{4}{5}) \tag{A3}$$

$$f_7 = -\frac{3125}{24} L_1 L_2 (L_1 - \frac{2}{5})(L_1 - \frac{3}{5})(L_1 - \frac{4}{5}) \tag{A4}$$

$$f_8 = \frac{3125}{24} L_2 L_3 (L_2 - \frac{1}{5})(L_2 - \frac{2}{5})(L_2 - \frac{3}{5}) \tag{A5}$$

$$f_9 = -\frac{3125}{12} L_2 L_3 (L_2 - \frac{1}{5})(L_2 - \frac{2}{5})(L_2 - \frac{4}{5}) \tag{A6}$$

$$f_{10} = \frac{3125}{12} L_2 L_3 (L_2 - \frac{1}{5})(L_2 - \frac{3}{5})(L_2 - \frac{4}{5}) \tag{A7}$$

$$f_{11} = -\frac{3125}{24} L_2 L_3 (L_2 - \frac{2}{5})(L_2 - \frac{3}{5})(L_2 - \frac{4}{5}) \tag{A8}$$

$$f_{12} = \frac{3125}{24} L_3 L_1 (L_3 - \frac{1}{5})(L_3 - \frac{2}{5})(L_3 - \frac{3}{5}) \tag{A9}$$

$$f_{13} = -\frac{3125}{12} L_3 L_1 (L_3 - \frac{1}{5})(L_3 - \frac{2}{5})(L_3 - \frac{4}{5}) \tag{A10}$$

$$f_{14} = \frac{3125}{12} L_3 L_1 (L_3 - \frac{1}{5})(L_3 - \frac{3}{5})(L_3 - \frac{4}{5}) \tag{A11}$$

$$f_{15} = -\frac{3125}{24} L_3 L_1 (L_3 - \frac{2}{5})(L_3 - \frac{3}{5})(L_3 - \frac{4}{5}) \tag{A12}$$

$$f_1 = L_1 - \frac{4}{5}(f_4 + f_{15}) - \frac{3}{5}(f_5 + f_{14}) - \frac{2}{5}(f_6 + f_{13}) - \frac{1}{5}(f_7 + f_{12}) \tag{A13}$$

$$f_2 = L_2 - \frac{4}{5}(f_7 + f_8) - \frac{3}{5}(f_6 + f_9) - \frac{2}{5}(f_5 + f_{10}) - \frac{1}{5}(f_4 + f_{11}) \tag{A14}$$

$$f_3 = L_3 - \frac{4}{5}(f_{11} + f_{12}) - \frac{3}{5}(f_{10} + f_{13}) - \frac{2}{5}(f_9 + f_{14}) - \frac{1}{5}(f_8 + f_{15}) \tag{A15}$$

where $L_1 = 1 - \xi - \eta$, $L_2 = \xi$ and $L_3 = \eta$ are the three area coordinates of an arbitrary point in the standard triangle.

Appendix B

The entries in matrix $[K]$ of Eq. (3.50) are given as

$$\begin{aligned}
&K_{1,1} = \xi_x^2 + c_1\xi_y^2, K_{1,2} = 2\xi_x\eta_x + 2c_1\xi_y\eta_y, K_{1,3} = \eta_x^2 + c_1\eta_y^2,\\
&K_{1,4} = \xi_{xx} + c_1\xi_{yy}, K_{1,5} = \eta_{xx} + c_1\eta_{yy}, K_{1,6} = -c_3,\\
&K_{1,7} = c_2\xi_x\xi_y, K_{1,8} = c_2\left(\xi_x\eta_y + \xi_y\eta_x\right), K_{1,9} = c_2\eta_x\eta_y,\\
&K_{1,10} = c_2\xi_{xy}, K_{1,11} = c_2\eta_{xy}, K_{1,12} = K_{1,13} = K_{1,14} = K_{1,15} = 0,\\
&K_{1,16} = c_3\xi_x, K_{1,17} = c_3\eta_x, K_{1,18} = 0,
\end{aligned} \tag{B1}$$

$$\begin{aligned}
&K_{2,1} = c_2\xi_x\xi_y, K_{2,2} = c_2\left(\xi_x\eta_y + \xi_y\eta_x\right), K_{2,3} = c_2\eta_x\eta_y, K_{2,4} = c_2\xi_{xy},\\
&K_{2,5} = c_2\eta_{xy}, K_{2,6} = 0, K_{2,7} = c_1\xi_x^2 + \xi_y^2, K_{2,8} = 2(c_1\xi_x\eta_x + \xi_y\eta_y),\\
&K_{2,9} = c_1\eta_x^2 + \eta_y^2, K_{2,10} = c_1\xi_{xx} + \xi_{yy}, K_{2,11} = c_1\eta_{xx} + \eta_{yy}, K_{2,12} = -c_3,\\
&K_{2,13} = K_{2,14} = K_{2,15} = 0, K_{2,16} = c_3\xi_y, K_{2,17} = c_3\eta_y, K_{2,18} = 0,
\end{aligned} \tag{B2}$$

$$\begin{aligned}
&K_{3,1} = K_{3,2} = K_{3,3} = K_{3,6} = 0, K_{3,4} = c_3\xi_x, K_{3,5} = c_3\eta_x,\\
&K_{3,7} = K_{3,8} = K_{3,9} = K_{3,12} = 0, K_{3,10} = c_3\xi_y, K_{3,11} = c_3\eta_y,\\
&K_{3,13} = -\left(c_3 + \tfrac{G_f}{D}\right)\left(\xi_x^2 + \xi_y^2\right), K_{3,14} = -2\left(c_3 + \tfrac{G_f}{D}\right)\left(\xi_x\eta_x + \xi_y\eta_y\right),\\
&K_{3,15} = -\left(c_3 + \tfrac{G_f}{D}\right)\left(\eta_x^2 + \eta_y^2\right), K_{3,16} = -\left(c_3 + \tfrac{G_f}{D}\right)\left(\xi_{xx} + \xi_{yy}\right),\\
&K_{3,17} = -\left(c_3 + \tfrac{G_f}{D}\right)\left(\eta_{xx} + \eta_{yy}\right), K_{3,18} = -\left(c_3 + \tfrac{G_f}{D}\right)\tfrac{k_f}{D}
\end{aligned} \tag{B3}$$

with

$$c_1 = \frac{1-\nu}{2}; \quad c_2 = \frac{1+\nu}{2}; \quad c_3 = \frac{6\kappa(1-\nu)}{h^2} \tag{B4}$$

Chapter 8

Mathematical Compendium

The mathematical compendium consists of four independent sections dealing in detail with specific mathematics used in the main text. It is provided as an aid to the readers who may not be familiar with these mathematical applications and thus may find it difficult to collect and study the necessary literature.

8.1 Gauss elimination

In linear algebra, Gauss elimination, named after German mathematician and scientist Carl Friedrich Gauss (1777-1855), is one of the most important numerical methods for solving a system of linear equations. The process of Gauss elimination consists of two steps, i.e., Forward and Backward eliminations. First, Forward elimination is performed to reduce a given system of linear equations to triangular forms, or lead to a degenerate equation with no solution, indicating that the system has no solution. This can be achieved by using elementary row operation. In triangular form, the solutions of the system can than be determined by back-substitution (Backward elimination).

A system of linear equations with n unknowns $x_1, x_2, \cdots, x_n$ is a set of n equations of the form

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n \end{cases} \tag{8.1}$$

In matrix form, we have

$$\mathbf{Ax} = \mathbf{b} \tag{8.2}$$

with

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \tag{8.3}$$

The system is homogeneous if all the b_i are zero, otherwise it is non-homogeneous.

Gauss elimination consists of combining the coefficient matrix **A** with the right-hand side b to the "augmented" $(n, n+1)$ matrix

$$\begin{bmatrix} \mathbf{A}\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & b_n \end{bmatrix} \tag{8.4}$$

A sequence of elementary row operations is then applied to this matrix so as to transform the coefficient part to upper triangular form:

(1) multiply a row by a non-zero real number c,

(2) swap two rows,

(3) add c times one row to another one.

$\begin{bmatrix} \mathbf{A}\ \mathbf{b} \end{bmatrix}$ will then have taken the following form:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ 0 & a'_{22} & \cdots & a'_{2n} & b'_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & a'_{nn} & b'_n \end{bmatrix} \tag{8.5}$$

and the original equation is transformed to $\mathbf{Rx} = \mathbf{c}$ with an upper triangular matrix R, from which the unknowns x can be found by back substitution.

Assume we have transformed the first column, and we want to continue the elimination with the following matrix

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ 0 & a'_{22} & \cdots & a'_{2n} & b'_2 \\ 0 & a'_{32} & \ldots & a'_{3n} & b'_3 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & a'_{n2} & \cdots & a'_{nn} & b'_n \end{bmatrix} \tag{8.6}$$

To zero a'_{32} we want to divide the second row by the "pivot" a'_{22}, multiply it with a'_{32} and subtract it from the third row. If the pivot is zero we have to swap two rows. This procedure frequently breaks down, not only for ill-conditioned matrices. Therefore, most programs perform "partial pivoting", i.e. they swap with the row that has the maximum absolute value of that column.

"Complete pivoting", always putting the absolute biggest element of the whole matrix into the right position, implying reordering of rows and columns, is normally not necessary.

Another variant is Gauss-Jordan elimination, which is closely related to Gaussian elimination. With the same elementary operations it does not only zero the elements below the diagonal but also above. The resulting augmented matrix will then look like:

$$\begin{bmatrix} a_{11} & 0 & \cdots 0 & b_1 \\ 0 & a'_{22} & \cdots 0 & b'_2 \\ & & & \\ 0 & 0 & \cdots a'_{nn} & b'_n \end{bmatrix} \tag{8.7}$$

Therefore, back substitution is not necessary and the values of the unknowns can be computed directly.

Example 8.1 *A system of linear equations with three unknowns*

In the following, a simple system of linear equations with three unknowns is considered to demonstrate the procedure of Gauss elimination. The given system is written as

$$\begin{cases} x_1 & +2x_2 & +2x_3 & = & 2 & L_1 \\ x_1 & +3x_2 & -2x_3 & = & -1 & L_2 \\ 3x_1 & +5x_2 & +8x_3 & = & 8 & L_3 \end{cases} \tag{8.8}$$

with

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 3 & -2 \\ 3 & 5 & 8 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 2 \\ -1 \\ 8 \end{bmatrix} \tag{8.9}$$

First, we eliminate x_2 in the lines L_2 and L_3:

$$\begin{cases} x_1 & + & 2x_2 & + & 2x_3 & = & 2 & L_1 \\ & & x_2 & - & 4x_3 & = & -3 & L_2 \leftarrow L_2 - L_1 \\ & & -x_2 & + & 2x_3 & = & 2 & L_3 \leftarrow L_3 - 3L_1 \end{cases} \tag{8.10}$$

Next, x_1 is eliminated in the line L_3 as

$$\begin{cases} x_1 & + & 2x_2 & + & 2x_3 & = & 2 & L_1 \\ & & x_2 & - & 4x_3 & = & -3 & L_2 \\ & & & - & 2x_3 & = & -1 & L_3 \leftarrow L_3 + L_2 \end{cases} \tag{8.11}$$

Now it is obvious that the coefficient matrix A is reduced in triangular form as

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 2 \\ 0 & 1 & -4 \\ 0 & 0 & -2 \end{bmatrix} \tag{8.12}$$

The first step of Gauss elimination−Forward elimination, is completed. The second step, backward elimination, aims to solve the triangular form of linear equations in reverse order. It can be readily seen from the line L_3 that

$$x_3 = \frac{1}{2} \qquad (L_3) \tag{8.13}$$

Then back-substituting $x_3 = \frac{1}{2}$ into the line L_2, we have

$$x_2 = -1 \qquad (L_2) \tag{8.14}$$

Finally, $x_3 = 1/2$ and $x_2 = -1$ can be substituted into line L_1 for the solution of x_1 as

$$x_1 = 3 \qquad (L_1) \tag{8.15}$$

Thus the solutions of the system are determined.

For a system of linear equations with n unknowns solved by Gauss elimination, approximately $2n^3/3$ operations are required. Therefore, Gauss elimination requires $O(n^3)$ order of computation. The algorithm can be coded by any computer language and applied for systems with thousands of equations and unknowns. However, for millions of equations, the algorithm is computationally expensive. In this case, these large systems are generally solved by interactive methods instead (Kreyszig, 1999; Press et al., 1986).

8.2 Successive over-relaxation (SOR) method

There are two approaches that can be used to determine the solutions of a linear system, i.e., direct methods and iterative methods. Direct methods, such as Gauss elimination method, aim to solve a system of linear equations by performing a finite number of operations to obtain the exact solutions. Direct methods are effective for a relatively small system of equations. However, the intensive computation prohibits the direct methods from solving large systems. In this case, we should resort to iterative methods, a very effective one of which is Successive Over-Relaxation (SOR) iterative method.

Suppose a system of linear equations is of the form

$$\mathbf{Ax} = \mathbf{b} \tag{8.16}$$

The coefficient matrix A can be expressed as the sum of the diagonal, lower triangular, and upper triangular matrices denoted by D, L and U, respectively. The three matrices can be expressed as

$$\mathbf{L} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ l_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & 0 \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_{nn} \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} 0 & u_{12} & \cdots & u_{1n} \\ 0 & 0 & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

(5 % 5 / 0+ :

The equations above (except the last equation) are vector equations. Denoting the i-th component of $\mathbf{k}$ by k_{ij} (j = 1, 2, 3, 4) and denoting the i-th component of $\mathbf{y}_k$ by y_{ik} (k = 1, 2, 3,...), we obtain by equating components

$$\begin{aligned}
&k_{i1} = hf_i(t^n, y_1^n, \cdots, y_m^n) \\
&k_{i2} = hf_i(t^n + \frac{h}{2}, y_1^n + \frac{a_{11}}{2}, \cdots, y_m^n + \frac{a_{m1}}{2}) \\
&k_{i3} = hf_i(t^n + \frac{h}{2}, y_1^n + \frac{a_{12}}{2}, \cdots, y_m^n + \frac{a_{m2}}{2}) \quad i = 1, 2, \cdots, m \\
&k_{i4} = hf_i(t^n + h, y_1^n + a_{13}, \cdots, y_m^n + a_{m3}) \\
&y_i^{n+1} = y_i^n + \frac{1}{6}(k_{i1} + 2k_{i2} + 2k_{i3} + k_{i4}) + O(h^5) \\
&t^{n+1} = t^n + h
\end{aligned} \tag{8.33}$$

8.5 Complex analysis

8.5.1 Complex variable

Complex analysis, traditionally known as the theory of functions of a complex variable, is the branch of mathematics investigating functions of complex numbers. It is useful in many branches of mathematics, including number theory and applied mathematics, and in physics.

Complex analysis is particularly concerned with the analytic functions of complex variables, which are commonly divided into two main classes: the holomorphic functions and the meromorphic functions. Because the separable real and imaginary parts of any analytic function must satisfy Laplace's equation, complex analysis is widely applicable to two-dimensional problems in physics.

A complex function is a function in which the independent variable and the dependent variable are both complex numbers. More precisely, a complex function is a function whose domain Ω is a subset of the complex plane and whose range is also a subset of the complex plane.

For any complex function, both the independent variable and the dependent variable may be separated into real and imaginary parts:

$$z = x + iy \tag{8.34a}$$

$$w = f(z) = u(z) + iv(z) \tag{8.34b}$$

where $x, y \in \mathbf{R}$ and $u(z)$ and $v(z)$ are real-valued functions. In other words, the components of the function $f(z)$

$$u = u(x, y) \tag{8.35a}$$

$$v = v(x, y) \tag{8.35b}$$

can be interpreted as real valued functions of the two real variables, x and y. The basic concepts of complex analysis are often introduced by extending the elementary real functions (e.g., exponentials, logarithms, and trigonometric functions) into the complex domain.

Just as in real analysis, a "smooth" complex function $w = f(z)$ may have a derivative at a particular point in its domain Ω. In fact, the definition of the derivative

$$f'(z) = \lim_{h \to 0} \frac{f(z+h) - f(z)}{h} \tag{8.36}$$

is analogous to the real case but with one very important difference. In real analysis, the limit can only be approached by moving along the one-dimensional number line. In complex analysis, the limit can be approached from any direction in the two-dimensional complex plane. (The claim that "in real analysis, the limit can only be approached by moving along the one dimensional number line" should not be confused with directional derivatives. It may be explained that in directional derivatives, one still moves along the one dimensional x line but it can be in "discrete" units; that is, if one follows $y = x^2$ curve, that does not mean that one is moving on the plane (instead of the one dimensional x line) but means that one is approaching in steps of discrete units.)

If this limit, the derivative, exists for every point$z$in Ω, then$f(z)$ is said to be differentiable on Ω. It can be shown that any differentiable $f(z)$ is analytic. This is a much more powerful result than the analogous theorem that can be proved for real-valued functions of real numbers. In the calculus of real numbers, we can construct a function $f(x)$ that has a first derivative everywhere, but for which the second derivative does not exist at one or more points in the function's domain. But in the complex plane, if a function$f(z)$ is differentiable in a neighborhood it must also be infinitely differentiable in that neighborhood.

By applying the methods of vector calculus to compute the partial derivatives of the two real functions $u(x, y)$ and $v(x, y)$ into which $f(z)$ can be decomposed, and by considering two paths leading to a point z in Ω, it can be shown that the existence of derivative implies

$$f'(z) = \frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x} = \frac{\partial v}{\partial y} - i\frac{\partial u}{\partial y} \tag{8.37}$$

Equating the real and imaginary parts of these two expressions we obtain the traditional formulation of the Cauchy-Riemann Equations:

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \text{and} \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x} \tag{8.38}$$

or, in another common notation

$$u_x = v_y, \text{and} \quad u_{\mathrm{y}} = -v_{\mathrm{x}} \tag{8.39}$$

By differentiating this system of two partial differential equations first with respect to x, and then with respect to y, we can easily show that

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \text{and} \quad \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = 0 \tag{8.40}$$

In other words, the real and imaginary parts of a differentiable function of a complex variable are harmonic functions because they satisfy Laplace's equation.

8.5.2 Holomorphic functions

Holomorphic functions are complex functions defined on an open subset of the complex plane which are differentiable. Complex differentiability has much stronger consequences than usual (real) differentiability. For instance, holomorphic functions are infinitely differentiable, a fact that is far from true for real differentiable functions. Most elementary functions, including the exponential function, the trigonometric functions, and all polynomial functions, are holomorphic.

8.5.3 Meromorphic functions

A meromorphic function on an open subset D of the complex plane is a function that is holomorphic on all D except a set of isolated points, which are poles for the function.

Every meromorphic function on D can be expressed as the ratio between two holomorphic functions (with the denominator not constant 0) defined on D: the poles then occur at the zeroes of the denominator. The Gamma function is meromorphic in the whole complex plane.

Intuitively then, a meromorphic function is a ratio of two nice (holomorphic) functions. Such a function will still be "nice", except at the points where the denominator of the fraction is zero, when the value of the function will be infinite.

From an algebraic point of view, if D is connected, then the set of meromorphic functions is the field of fractions of the integral domain of the set of holomorphic functions. This is analogous to the relationship between **Q**, the rational numbers, and **Z**, the integers.

'

+,- * *. * 9 " 1: ; "
/ +// /"
(% % 9

As this particle moves about, its state properties will be understood to be the same as the local state properties of the continuum.

1.3 INERTIAL COORDINATE FRAMES

It is sufficient to consider Euclidean space to describe the laws of fluid motion. A Euclidean space is a curvature-free space in which a set of rectangular Cartesian coordinates can always be introduced on a global scale. One can introduce any other system of coordinates in this space without altering the basic nature of the space itself. These ideas form a logical environment for understanding coordinate transformation in Newtonian mechanics.

In fluid dynamics, the speeds encountered are far smaller than the speed of light, so that the relativistic effects are negligibly small. Time is considered an absolute entity irrespective of the state of motion, and Newton's laws of motion are assumed to be exact. The concept of an inertial frame of reference is very important. It is precisely defined by Newton's first law of motion. It is a coordinate frame with respect to which bodies, under the absence of external forces, move with zero acceleration. (For transformation between reference frames, refer to ME9.)

In this regard, the idea of a Galilean frame is also important. A coordinate frame that has only uniform translation with respect to a primitive intertial frame embedded in the nearby "fixed stars" is called Galilean:

$$\bar{\mathbf{r}} = \mathbf{r} - \mathbf{V}\mathrm{t} \quad \text{or,} \quad \bar{\mathrm{x}}_{\mathrm{i}} = \mathrm{x}_{\mathrm{i}} - \mathrm{V}_{\mathrm{i}}\mathrm{t}$$

$$\bar{t} = t$$

or

$$\mathbf{r} = \bar{\mathbf{r}} + \mathbf{V}\bar{\mathrm{t}} \quad \text{or,} \quad x_i = \bar{x}_i + V_i\bar{t}$$

where **V** is a constant velocity.

Thus,

$$\bar{\mathbf{u}} = \mathbf{u} - \mathbf{V}$$

$$\mathbf{u} = \bar{\mathbf{u}} + \mathbf{V}$$

$$\bar{\mathbf{a}} = \mathbf{a}$$

The quantities that remain invariant between Galilean frames are called Galilean invariants. Examples are acceleration, stress, vorticity, pressure, temperature, and heat flux. Some Galilean noninvariants are velocity, kinetic energy, and molecular mean free path.

1.4 MOTION OF A CONTINUUM

The continuum hypothesis postulates fluid particles or the material points distributed continuously. To describe any kind of motion, a reference coordinate system is needed. Because fluid motion does not require any relativistic considerations, we take time as an absolute quantity common to

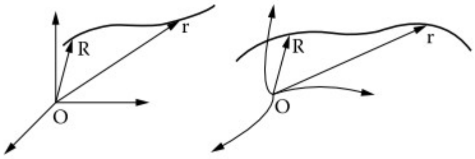

Figure 1.1 Inertial Cartesian or curvilinear reference frame.

any reference frame at rest or in motion. For simplicity, choose a reference frame at rest and refer the motion of all particles with respect to it. This coordinate system can be a curvilinear coordinate system. In what follows, we shall consider an inertial frame for the description of fluid motion, as shown in Figure 1.1.

The material points at the fixed reference time t_0 or at any time t can now be located with respect to this inertial coordinate system by drawing position vectors from point o. Let **R** and **r** denote the position vectors of the fluid particles at time t_0 and t, respectively.* The vectors **R** and **r** can be denoted as an ordered triple of numbers in any of the following forms (refer to ME1 for details):

$$\mathbf{R} = (X, Y, Z) = (X_i) = (X^i),$$

$$\mathbf{r} = (x, y, z) = (x_i) = (x^i)$$

where **R** is the reference configuration at t_0, the reference time.

In this text subscripted variables X_i, x_i denote rectangular Cartesian coordinates, whereas X^i, x^i denote any general nonrectangular coordinates. Thus, **r** referred to a rectangular Cartesian system is

$$\mathbf{r} = x_1\mathbf{i}_1 + x_2\mathbf{i}_2 + x_3\mathbf{i}_3 = x_k\mathbf{i}_k \tag{1.1a}$$

where any two repeated indices imply summation, and $\mathbf{i}_k$ for $k = 1, 2, 3$ are the constant unit vectors along x_1, x_2, and x_3 axes, respectively. For general curvilinear coordinates (x^1, x^2, x^3) the position vector **r** is a function of these coordinates, i.e., $\mathbf{r} = \mathbf{r}(x^i)$, and it is not possible to write **r** linearly in terms of x^i as in Equation 1.1a. It must, however, be noted that the first differential $d\mathbf{r}$ is expressible linearly in dx^i for all coordinates as

$$\begin{aligned} d\mathbf{r} &= \frac{\partial \mathbf{r}}{\partial x^m}\, dx^m \,(\text{sum on } m) \\ &= \mathbf{a}_m\, dx^m \end{aligned} \tag{1.1b}$$

and $\mathbf{a}_m$'s are called the *covariant base vectors*. If x^m's are the rectangular Cartesian coordinates, i.e.,

$$x^1 = x_1, \qquad x^2 = x_2, \qquad x^3 = x_3$$

then from Equation 1.1a we find that

$$\mathbf{a}_m = \mathbf{i}_m, \qquad m = 1, 2, 3 \tag{1.1c}$$

where $\mathbf{i}_m$'s are the constant unit vectors. (Refer to ME1.)

---

* At some places we have also used the symbol **x** in lieu of **r**.

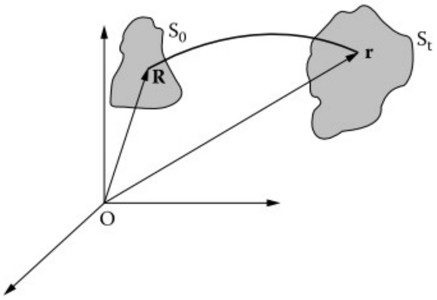

Figure 1.2 Motion of an imaginary volume in a continuum.

In this text, we will use both rectangular Cartesian and general curvilinear coordinates. From this discussion, the reader should have no difficulty in identifying the implied coordinate system. Functions of the coordinates will be written as $F(\mathbf{R})$ or $F(\mathbf{r})$, etc.

Suppose at time $t = t_0$ we isolate a portion of the continuum and denote its bounding surface by S_0, as shown in Figure 1.2. A specific particle in the region enclosed by S_0 can be named $X_1 = a$, $X_2 = b$, $X_3 = c$, where (a, b, c) are the Cartesian coordinates of a particle at $t = t_0$. As this particle moves, its position at some other time $t > t_0$ can be located with respect to the reference system by (x_1, x_2, x_3), where

$$x_i = \phi_i(a,b,c,t), \qquad i = 1, 2, 3 \tag{1.2}$$

The functions ϕ_i are to be determined by the laws of fluid motion. (The determination of ϕ_i is actually the "ultimate problem" of fluid dynamics, and the functions are not known *a priori*.) In the same manner any other particle labeled by its coordinates can be followed to a time $t > t_0$. Collectively, all these particles at time $t > t_0$ will be enclosed by a surface S_t.

This notion of mapping points at $t = t_0$ enclosed by an imaginary surface S_0 onto the points of the region enclosed by S_t at $t > t_0$ is of great help. This mapping should be one-to-one for distinct points to remain distinct. That is, two or more points enclosed by S_0 should not map onto a single point of the region enclosed by S_t.

Collectively, consider all the points at time $t = t_0$ and follow their motion in time $t > t_0$. The law of fluid motion should then be written as

$$x_i = \phi_i(X_1, X_2, X_3, t), \qquad i = 1, 2, 3 \tag{1.3}$$

so that (X_1, X_2, X_3) can be any fluid particle at $t = t_0$. In vector form Equation 1.3 is written as

$$\mathbf{r} = \boldsymbol{\phi}(\mathbf{R}, t) \tag{1.4}$$

where the vector-valued function $\boldsymbol{\phi}$ must be such that at the time $t = t_0$, the position vectors $\mathbf{r}$ and $\mathbf{R}$ coincide. Thus,

$$\mathbf{r}(\mathbf{R}, t_0) = \boldsymbol{\phi}(\mathbf{R}, t_0) = \mathbf{R} \tag{1.5}$$

The coordinates (X^i or X_i) are usually called the material or *Lagrangian coordinates*, and (x^i or x_i) are called the spatial or *Eulerian coordinates*. The following interpretation is in order:

1. If **R** is kept fixed and t varies, then Equation 1.4 represents the path or the trajectory followed by the chosen particle through time.
2. If t is kept fixed, then Equation 1.4 represents the transformation of coordinates occupied by the fluid particles at $t = t_0$ to the coordinates at time t. This transformation is generally nonlinear. Because the space remains Euclidean at $t = t_0$ and $t > t_0$, Equation 1.4 transforms an Euclidean space into itself. In essence, the map **F** defined as $\mathbf{F} : \mathbf{R} \rightarrow \boldsymbol{\phi}\,(\mathbf{R}, t)$ is the fluid flow map.

Besides $\boldsymbol{\phi}$ being a one-to-one mapping of points, the function $\boldsymbol{\phi}$ in Equation 1.4 must be spatially differentiable to some specified order. We will consider $\boldsymbol{\phi}$ to be single valued and at least first-order differentiable in the spatial variables. Let the inverse of Equation 1.4 be written as

$$\mathbf{R} = \psi(\mathbf{r}, t) \tag{1.6}$$

For Equation 1.6 to exist, the Jacobian J of the map **F** defined below must not vanish at any point of the flow field:

$$J(\mathbf{R}, t) = \det\left(\frac{\partial x_i}{\partial X_j}\right) \neq 0 \tag{1.7a}$$

Another way of defining J is to introduce the gradient operator "Grad" in the reference configuration **R** and then a second-order tensor or matrix **G** is obtained as

$$\mathbf{G} = \text{Grad } \mathbf{r}$$

Thus,

$$J = \det(\mathbf{G})$$

In fully expanded form Equation 1.7a is

$$\begin{vmatrix} \dfrac{\partial x_1}{\partial X_1} & \dfrac{\partial x_1}{\partial X_2} & \dfrac{\partial x_1}{\partial X_3} \\ \dfrac{\partial x_2}{\partial X_1} & \dfrac{\partial x_2}{\partial X_2} & \dfrac{\partial x_2}{\partial X_3} \\ \dfrac{\partial x_3}{\partial X_1} & \dfrac{\partial \boldsymbol{x}_3}{\partial X_2} & \dfrac{\partial x_3}{\partial X_3} \end{vmatrix} \neq 0 \tag{1.7b}$$

The condition $J \neq 0$ is known as the *smoothness condition* of fluid motion. Because of the continuity and differentiability of $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ in Equation 1.4 and Equation 1.6, we have

$$\boldsymbol{\phi}[\boldsymbol{\psi}(\mathbf{r}, t), t] = \mathbf{r} \tag{1.8a}$$

$$\boldsymbol{\psi}[\boldsymbol{\phi}(\mathbf{R}, t), t] = \mathbf{R} \tag{1.8b}$$

$$J(\mathbf{R}, t_0) = 1 \tag{1.8c}$$

The preceding can formally be stated using the implicit function theorem. According to the implicit function theorem, if the vector-valued function $\boldsymbol{r} = \boldsymbol{\varphi}(\boldsymbol{R}, \mathrm{t})$ for a fixed time t is continuous and possesses continuous first-order partial derivatives with respect to $\boldsymbol{R}$ in a neighborhood of the point $\boldsymbol{R}$, and if $J \neq 0$ there, then a unique inverse $\boldsymbol{R} = \boldsymbol{\psi}(\boldsymbol{r}, \mathrm{t})$ exists at a point $\boldsymbol{r}$ at time t.

1.5 THE TIME DERIVATIVES

Suppose F to be either a scalar, vector, or tensor function of position and time representing some physical property of the flow. We can write it as $F(\mathbf{r}, t)$ or $F(\mathbf{R}, t)$ because the inverse of the transformation from $\mathbf{r}$ to $\mathbf{R}$ or $\mathbf{R}$ to $\mathbf{r}$ is assumed to exist. Physically, $F(\mathbf{R}, t)$ is the value of F experienced by a fluid particle at time t that at time t_0 was at $\mathbf{R}$, and $F(\mathbf{r}, t)$ is the value experienced by an observer at the fixed position $\mathbf{r}$ at time t. There are two possible time derivatives:

$$(i) \quad \frac{\partial F}{\partial t} = \left.\frac{\partial F(\mathbf{r}, t)}{\partial t}\right|_{\mathbf{r}\ \text{fixed}} \tag{1.9a}$$

$$(i) \quad \frac{dF}{dt} = \left.\frac{\partial F(\mathbf{R}, t)}{\partial t}\right|_{\mathbf{R}\ \text{fixed}} \tag{1.9b}$$

Here, $\partial F/\partial t$ is the rate of change of F measured by an observer stationed at the fixed point $\mathbf{r}$ and is a local time variation of F. On the other hand, dF/dt is the rate of change of F measured by an observer moving with the fluid particle. The time derivative dF/dt is called the *material* or *substantive derivative*, or the *derivative following the motion*. The reason for these names is based on the fact that $F(\mathbf{R}, t)$ is purely a function of time and its derivative is a total derivative.

1.6 VELOCITY AND ACCELERATION

Any description of fluid motion using the Lagrangian coordinates $\mathbf{R} = (X_i)$ and time t is called a description in the *sense of Lagrange*. From an intuitive viewpoint, one should start describing velocity and acceleration in the sense of Lagrange, as is done in classical particle dynamics. It must be noted in this context that the $\mathbf{r}$ defined in Equation 1.4 represents the changing coordinates of a moving fluid particle. The velocity of a fluid particle at time t, which was at the position $\mathbf{R}$ at time $t = t_0$, is obtained by differentiating Equation 1.4, i.e.,

$$\frac{d\mathbf{r}}{dt} = \frac{\partial \boldsymbol{\phi}(\mathbf{R}, t)}{\partial t} = \mathbf{U}(\mathbf{R}, t) \tag{1.10}$$

and is called the *Lagrangian velocity*. Also,

$$d\mathbf{r} = \mathbf{U}(\mathbf{R}, t)dt$$

which on integration and using Equation 1.5 yields the particle trajectory:

$$\mathbf{r} = \mathbf{R} + \int_{t_0}^{t} \mathbf{U}(\mathbf{R}, t)\, dt \tag{1.11}$$

' %
+,- * *. * % 1 2 " = 8 H -H
< 7 /"
(# 6 % %

Problema 3.21 Demonstre que os operadores de projeção são **idempotentes:** $\hat{P}^2 = \hat{P}$. Determine os autovalores de $\hat{P}$ e caracterize seus autovetores.

Problema 3.22 Considere um espaço vetorial tridimensional gerado por uma base ortonormal $|1\rangle$, $|2\rangle$, $|3\rangle$. *Kets* $|\alpha\rangle$ e $|\beta\rangle$ são dados por

$$|\alpha\rangle = i|1\rangle - 2|2\rangle - i|3\rangle, \quad |\beta\rangle = i|1\rangle + 2|3\rangle.$$

(a) Monte $\langle\alpha|$ e $\langle\beta|$ (nos termos da base dual $\langle 1|$, $\langle 2|$, $\langle 3|$).

(b) Calcule $\langle\alpha|\beta\rangle$ e $\langle\beta|\alpha\rangle$ e confirme que $\langle\beta|\alpha\rangle = \langle\alpha|\beta\rangle^*$.

(c) Calcule os nove elementos da matriz do operador $\hat{A} \equiv |\alpha\rangle\langle\beta|$ nos termos dessa base e monte a matriz **A**. Ela é hermitiana?

Problema 3.23 O Hamiltoniano para determinado sistema de dois níveis é

$$\hat{H} = \epsilon(|1\rangle\langle 1| - |2\rangle\langle 2| + |1\rangle\langle 2| + |2\rangle\langle 1|),$$

em que $|1\rangle$, $|2\rangle$ é uma base ortonormal e ϵ é um número com as dimensões da energia. Encontre seus autovalores e autovetores (como combinações lineares de $|1\rangle$ e $|2\rangle$). Qual é a matriz **H** que representa $\hat{H}$ em relação a essa base?

Problema 3.24 Seja $\hat{Q}$ um operador com um conjunto completo de autovetores ortonormais:

$$\hat{Q}|e_n\rangle = q_n|e_n\rangle \quad (n = 1, 2, 3, \ldots).$$

Demonstre que $\hat{Q}$ pode ser escrito em termos de sua **decomposição espectral:**

$$\hat{Q} = \sum_n q_n |e_n\rangle\langle e_n|.$$

Dica: um operador é caracterizado por sua ação em todos os vetores possíveis, portanto, o que você deve mostrar é que

$$\hat{Q}|\alpha\rangle = \left\{\sum_n q_n |e_n\rangle\langle e_n|\right\}|\alpha\rangle,$$

para qualquer vetor $|\alpha\rangle$.

Mais problemas para o Capítulo 3

Problema 3.25 **Polinômios de Legendre.** Use o procedimento de Gram-Schmidt (Problema A.4) para ortonormalizar as funções 1, x, x^2 e x^3 no intervalo $-1 \leq x \leq 1$. Talvez você reconheça os resultados: eles são (à parte da normalização)[30] os **polinômios de Legendre** (Tabela 4.1).

Problema 3.26 Um operador **anti-hermitiano** é igual a *menos* o seu conjugado hermitiano:

$$\hat{Q}^\dagger = -\hat{Q}. \qquad [3.95]$$

(a) Demonstre que o valor esperado de um operador anti-hermitiano é imaginário.

(b) Demonstre que o comutador de dois operadores hermitianos é anti-hermitiano. E o comutador de dois operadores anti-hermitianos, o que é?

Problema 3.27 **Medições sequenciais.** Um operador $\hat{A}$ representando o observável A tem dois autoestados normalizados, Ψ_1 e Ψ_2, com autovalores a_1 e a_2, respectiva-

---

[30] Legendre não sabia qual seria a melhor convenção; ele escolheu o fator global para que todas as suas funções fossem 1 em $x = 1$, e agora estamos presos a essa escolha infeliz.

data 03.04.2020

S T Q Q S S D

122) Books_2020_10/
Books_53_2020/
Theoretical Fluid Dynamics/
Fluid Dynamics Theoretical and Computational Approach...
pdf

Pag. 03(30)

Fig. 1.1

Inertial Cartesian or curvilinear reference frame

Date: 20200318

123) DATA_NOVOVIVO_63_2020/
Books_63_2020/
Introduction to Quantum Mechanics/
David J. Griffiths_
Mecânica Quântica..pdf

Pag. 95(100) -

Problema 3.23

O Hamiltoniano para determinado sistema de dois níveis é

data 03.04.2020

$$\hat{H} = \epsilon(|1\rangle\langle 1| - |2\rangle\langle 2| + |1\rangle\langle 2| + |2\rangle\langle 1|),$$

em que $|1\rangle$, $|2\rangle$
é uma base ortonormal e ϵ é
um número com as dimensões
da energia.

Problema 3.25

Polinômios de Legendre
— use procedimento de
Gram-Schmidt (Problema A.4)
para ortonormalizar as
funções.

Problema 3.26

Um operador anti-hermitiano
é igual a menos o seu conjugado
hermitiano:

$$\hat{Q}^\dagger = -\hat{Q} \qquad [3.95]$$

(124) Books_63_2020/
Introduction to Quantum
Mechanics/
David J. Griffiths - Introduction

Problema 3.21 Demonstre que os operadores de projeção são **idempotentes:** $\hat{P}^2 = \hat{P}$. Determine os autovalores de $\hat{P}$ e caracterize seus autovetores.

Problema 3.22 Considere um espaço vetorial tridimensional gerado por uma base ortonormal $|1\rangle$, $|2\rangle$, $|3\rangle$. *Kets* $|\alpha\rangle$ e $|\beta\rangle$ são dados por

$$|\alpha\rangle = i|1\rangle - 2|2\rangle - i|3\rangle, \quad |\beta\rangle = i|1\rangle + 2|3\rangle.$$

(a) Monte $\langle\alpha|$ e $\langle\beta|$ (nos termos da base dual $\langle 1|$, $\langle 2|$, $\langle 3|$).

(b) Calcule $\langle\alpha|\beta\rangle$ e $\langle\beta|\alpha\rangle$ e confirme que $\langle\beta|\alpha\rangle = \langle\alpha|\beta\rangle^*$.

(c) Calcule os nove elementos da matriz do operador $\hat{A} \equiv |\alpha\rangle\langle\beta|$ nos termos dessa base e monte a matriz **A**. Ela é hermitiana?

Problema 3.23 O Hamiltoniano para determinado sistema de dois níveis é

$$\hat{H} = \epsilon(|1\rangle\langle 1| - |2\rangle\langle 2| + |1\rangle\langle 2| + |2\rangle\langle 1|),$$

em que $|1\rangle$, $|2\rangle$ é uma base ortonormal e ϵ é um número com as dimensões da energia. Encontre seus autovalores e autovetores (como combinações lineares de $|1\rangle$ e $|2\rangle$). Qual é a matriz **H** que representa $\hat{H}$ em relação a essa base?

Problema 3.24 Seja $\hat{Q}$ um operador com um conjunto completo de autovetores ortonormais:

$$\hat{Q}|e_n\rangle = q_n|e_n\rangle \quad (n = 1, 2, 3, ...).$$

Demonstre que $\hat{Q}$ pode ser escrito em termos de sua **decomposição espectral:**

$$\hat{Q} = \sum_n q_n |e_n\rangle\langle e_n|.$$

Dica: um operador é caracterizado por sua ação em todos os vetores possíveis, portanto, o que você deve mostrar é que

$$\hat{Q}|\alpha\rangle = \left\{\sum_n q_n |e_n\rangle\langle e_n|\right\}|\alpha\rangle,$$

para qualquer vetor $|\alpha\rangle$.

Mais problemas para o Capítulo 3

Problema 3.25 **Polinômios de Legendre.** Use o procedimento de Gram-Schmidt (Problema A.4) para ortonormalizar as funções 1, x, x^2 e x^3 no intervalo $-1 \leq x \leq 1$. Talvez você reconheça os resultados: eles são (à parte da normalização)[30] os **polinômios de Legendre** (Tabela 4.1).

Problema 3.26 Um operador **anti-hermitiano** é igual a *menos* o seu conjugado hermitiano:

$$\hat{Q}^\dagger = -\hat{Q}. \qquad [3.95]$$

(a) Demonstre que o valor esperado de um operador anti-hermitiano é imaginário.

(b) Demonstre que o comutador de dois operadores hermitianos é anti-hermitiano. E o comutador de dois operadores anti-hermitianos, o que é?

Problema 3.27 **Medições sequenciais.** Um operador $\hat{A}$ representando o observável A tem dois autoestados normalizados, Ψ_1 e Ψ_2, com autovalores a_1 e a_2, respectivamente. O operador $\hat{B}$, representando o observável B, tem dois autoestados normalizados, ϕ_1 e ϕ_2, com autovalores b_1 e b_2. Os autoestados estão relacionados por

---

[30] Legendre não sabia qual seria a melhor convenção; ele escolheu o fator global para que todas as suas funções fossem 1 em $x = 1$, e agora estamos presos a essa escolha infeliz.

$$\psi_1 = (3\phi_1 = 4\phi_2)/5, \quad \psi_2 = (4\phi_1 - 3\phi_2) \ /5.$$

(a) O observável A é medido e o valor a_1 é obtido. Qual é o estado do sistema (imediatamente) após a medida?

(b) Se a medição de B já foi feita, quais são os resultados possíveis, e quais são as probabilidades para cada um deles?

(c) Logo após a medição de B, A é medido novamente. Qual é a probabilidade de se obter a_1? (Observe que a resposta seria bem diferente se eu tivesse informado o resultado das medições de B.)

****Problema 3.28** Calcule a função de onda para o espaço de momento $\Phi_n(p, t)$ para o n-ésimo estado estacionário do poço quadrado infinito. Esboce $|\Phi_1(p, t)|^2$ e $|\Phi_2(p, t)|^2$, como funções de p (preste atenção aos pontos $p = \pm n\pi\hbar/a$). Use $\Phi_n(p, t)$ para calcular o valor esperado de p^2. Compare sua resposta à do Problema 2.4.

Problema 3.29 Considere a função de onda

$$\Psi(x,0) = \begin{cases} \dfrac{1}{\sqrt{2n\lambda}} e^{i2\pi x/\lambda}, & -n\lambda < x < n\lambda, \\ 0, & \text{caso contrário,} \end{cases}$$

em que n é um número positivo inteiro. Essa função é puramente sinusoidal (com comprimentos de onda λ) no intervalo $-n\lambda < x < n\lambda$, mas ainda carrega uma *faixa* de momentos, pois as oscilações não seguem até o infinito. Descubra a função de onda para o espaço de momento $\Phi(p, 0)$. Faça o gráfico de $|\Psi(x, 0)|^2$ e $|\Phi(p, 0)|^2$ e determine suas larguras, w_x e w_p (a distância entre os zeros de cada lado do ponto principal). Observe o que acontece com cada largura quando $n \Rightarrow \infty$. Usando w_x e w_p como estimativas de Δx e Δp, verifique se o princípio da incerteza é satisfeito. *Atenção:* se você tentar calcular σ_p, terá uma desagradável surpresa. Você consegue diagnosticar o problema?

Problema 3.30 Suponha que

$$\Psi(x,0) = \frac{A}{x^2 + a^2}, \quad (-\infty < x < \infty)$$

para constantes A e a.

(a) Determine A, normalizando $\Psi_1(x, 0)$.

(b) Calcule $\langle x \rangle$, $\langle x^2 \rangle$ e σ_x (em tempo $t = 0$).

(c) Calcule a função de onda para o espaço de momento $\Phi(p, 0)$ e certifique-se de que ela esteja normalizada.

(d) Use $\Phi(p, 0)$ para calcular $\langle p \rangle$, $\langle p \rangle^2$ e σ_p (em tempo $t = 0$).

(e) Verifique o princípio da incerteza de Heisenberg para esse estado.

***Problema 3.31 Teorema do virial.** Use a Equação 3.71 para mostrar que

$$\frac{d}{dt}\langle xp \rangle = 2\langle T \rangle - \left\langle x\frac{dV}{dx} \right\rangle, \qquad [3.96]$$

em que T é a energia cinética ($H = T + V$). Em um estado *estacionário*, o lado esquerdo é zero (por quê?) assim que

$$2\langle T \rangle - \left\langle x\frac{dV}{dx} \right\rangle. \qquad [3.97]$$

Isso é chamado de **teorema do virial**. Use-o para provar que $\langle T \rangle = \langle V \rangle$ para estados estacionários do oscilador harmônico e verifique se isso é compatível com os resultados obtidos nos problemas 2.11 e 2.12.

Problema 3.32 Em uma versão interessante do princípio da incerteza para a energia-tempo[31] $\Delta t = \tau/\pi$, em que τ é o tempo que $\Psi(x, t)$ leva para evoluir a um estado ortogonal $\Psi(x, 0)$. Faça um teste usando uma função de onda que é uma mistura equivalente de dois estados estacionários (ortonormais) de um potencial (arbitrário): $\Psi(x, 0) = (1\sqrt{2})[\Psi_1(x) + \Psi_2(x)]$.

****Problema 3.33** Calcule os elementos da matriz $\langle n|x|n'\rangle$ e $\langle n|p|n'\rangle$ na base (ortonormal) de estados estacionários para o oscilador harmônico (Equação 2.67). Você já calculou os elementos 'diagonais' ($n = n'$) no Problema 2.12; use a mesma técnica para o caso geral. Construa as matrizes (infinitas) correspondentes, **X** e **P**. Mostre que $(1/2m)\mathbf{P}^2 + (m\omega^2/2)\mathbf{X}^2 = \mathbf{H}$ é *diagonal* nessa base. Os elementos diagonais que você encontrou são o que você esperava? *Resposta parcial:*

$$\langle n|x|n'\rangle = \sqrt{\frac{\hbar}{2m\omega}}\left(\sqrt{n'}\delta_{n,n'-1} + \sqrt{n}\delta_{n',n-1}\right). \qquad [3.98]$$

Problema 3.34 O oscilador harmônico está em tal estado que a medida da energia produziria tanto $(1/2)\hbar\omega$ quanto $(3/2)\ \hbar\omega$, com igual probabilidade. Qual é o maior valor possível de $\langle p \rangle$ nesse caso? Se esse valor máximo for aceito no tempo $t = 0$, qual é o $\Psi(x, t)$?

*****Problema 3.35 Estados coerentes para o oscilador harmônico.** Entre os estados estacionários do oscilador harmônico ($|n\rangle = \Psi_n(x)$, Equação 2.67), somente $n = 0$ atinge o limite da incerteza ($\sigma_x\sigma_p = \hbar/2$); em geral, $\sigma_x\sigma_p = (2n + 1)\hbar/2$, conforme você viu no Problema 2.12. Porém, certas *combinações lineares* (conhecidas como **estados coerentes**) também minimizam o produto da incerteza. Acontece que eles são *autofunções do operador de abaixamento*:[32]

$$a_-|\alpha\rangle = \alpha|\alpha\rangle$$

(O autovalor α pode ser qualquer número complexo.)

(a) Calcule $\langle x \rangle$, $\langle x^2 \rangle$, $\langle p \rangle$, $\langle p^2 \rangle$ no estado $|\alpha\rangle$. *Dica:* use a técnica do Exemplo 2.5 e lembre-se de que a_+ é o conjugado hermitiano de a_-. *Não* suponha que α seja real.

---

[31] Veja Lev Vaidman, *Am. J. Phys.* 60, 182 (1992) para uma prova disso.

[32] Não há autofunções normalizáveis do operador de *levantamento*.

(b) Calcule σ_x e σ_p; demonstre que $\sigma_x\sigma_p = \hbar/2$.

(c) Assim como qualquer outra função de onda, um estado coerente pode ser expandido em termos dos autoestados das energias:

$$|\alpha\rangle = \sum_{n=0}^{\infty} c_n\,|n\rangle.$$

Demonstre que os coeficientes de expansão são

$$c_n = \frac{\alpha^n}{\sqrt{n!}}c_0.$$

(d) Determine c_0 por meio da normalização de $|\alpha\rangle$. *Resposta:* $\exp(-|\alpha|^2/2)$.

(e) Agora coloque a dependência do tempo:

$$|n\rangle \to e^{-iE_n t/\hbar}\,|n\rangle,$$

e demonstre que $|\alpha(t)\rangle$ continua sendo um autoestado de a_-, porém, o *autovalor* evolui com o tempo:

$$\alpha(t) = e^{-i\omega t}\alpha.$$

Portanto, um estado coerente *permanece* coerente e continua a minimizar o produto da incerteza.

(f) O estado fundamental ($|n=0\rangle$) é um estado coerente? Se sim, qual é o autovalor?

Problema 3.36 Princípio da incerteza expandido.[33] O princípio da incerteza generalizado (Equação 3.62) afirma que

$$\sigma_A^2\sigma_B^2 \geq \frac{1}{4}\langle C\rangle^2,$$

em que $\hat{C} \equiv -i[\hat{A}, \hat{B}]$.

(a) Demonstre que a equação pode ser reforçada para dizer

$$\sigma_A^2\sigma_B^2 \geq \frac{1}{4}(\langle C\rangle^2 + \langle D\rangle^2), \qquad [3.99]$$

em que $\hat{D} \equiv \hat{A}\hat{B} + \hat{B}\hat{A} - 2\langle A\rangle\langle B\rangle$. *Dica:* mantenha o termo $\mathrm{Re}(z)$ na Equação 3.60.

(b) Verifique a Equação 3.99 para o caso $B = A$ (nesse caso, o princípio da incerteza padrão é trivial, sendo que $\hat{C} = 0$; infelizmente, o princípio da incerteza expandido não ajuda muito).

Problema 3.37 O Hamiltoniano para um determinado sistema de três níveis está representado pela seguinte matriz:

$$\mathbf{H} = \begin{pmatrix} a & 0 & b \\ 0 & c & 0 \\ b & 0 & a \end{pmatrix},$$

em que a, b e c são números reais (suponha $a - c \neq \pm b$).

(a) Se o sistema se inicia no estado

$$|\mathcal{S}(0)\rangle = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix},$$

qual é a $|\mathcal{S}(t)\rangle$?

(b) Se o sistema se inicia no estado

$$|\mathcal{S}(0)\rangle = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

qual é a $|\mathcal{S}(t)\rangle$?

Problema 3.38 O Hamiltoniano para um determinado sistema de três níveis está representado pela seguinte matriz:

$$\mathbf{H} = \hbar\omega\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

Dois outros observáveis, A e B, estão representados pelas matrizes

$$\mathbf{A} = \lambda\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix}, \quad \mathbf{B} = \mu\begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

em que ω, λ e μ são números reais positivos.

(a) Calcule os autovalores e autovetores normalizados de **H**, **A** e **B**.

(b) Suponha que o sistema seja iniciado no estado genérico

$$|\mathcal{S}(0)\rangle = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix},$$

com $|c_1|^2 + |c_2|^2 + |c_3|^2 = 1$. Descubra os valores esperados de H, A e B (em $t = 0$).

(c) Qual é a $|\mathcal{S}(t)\rangle$? Se você medir a energia desse estado (em tempo t), quais valores você obterá, e quais as probabilidades de cada um? Responda a essas mesmas perguntas para A e B.

****Problema 3.39**

(a) Para uma função $f(x)$, que pode ser expandida em uma série de Taylor, demonstre que

$$f(x + x_0) = e^{i\hat{p}x_0/\hbar}f(x)$$

(em que x_0 é qualquer distância constante). Por essa razão, $\hat{p}/\hbar$ é chamado de **gerador de translações no**

---

[33] Para comentários e referências interessantes, veja R. R. Puri, *Phys. Rev. A* **49**, 2178 (1994).

espaço. *Nota:* o exponencial de um *operador* é definido pela expansão em série de potências: $e^{\hat{Q}} \equiv 1 + \hat{Q} + (1/2)\hat{Q}^2 + (1/3!)\hat{Q}^3 + \ldots$.

(b) Se $\Psi(x, t)$ satisfaz a equação de Schrödinger (dependente do tempo), demonstre que

$$\Psi(x, t + t_0) = e^{-i\hat{H}t_0/\hbar}\Psi(x, t)$$

(em que t_0 é qualquer tempo constante); $-\hat{H}/\hbar$ é chamado de **gerador de translações no tempo**.

(c) Demonstre que o valor esperado de uma variável dinâmica $Q(x, p, t)$ em tempo $t + t_0$ pode ser escrito[34]

$$\langle Q \rangle_{t=t_0} = \langle \Psi(x, t) | e^{i\hat{H}t_0/\hbar}\hat{Q}(\hat{x}, \hat{p}, t = t_0)e^{-i\hat{H}t_0/\hbar} | \Psi(x, t) \rangle.$$

Use isso para recuperar a Equação 3.71. *Dica:* faça $t_0 = dt$, e expanda para primeira ordem em dt.

****Problema 3.40**

(a) Escreva a 'equação de Schrödinger' dependente do tempo no espaço de momento para uma partícula livre e obtenha o resultado. *Resposta:* $\exp(-ip^2t/2m\hbar)\Phi(p, 0)$.

(b) Calcule $\Phi(p, 0)$ para o pacote de onda gaussiano móvel (Problema 2.43) e monte $\Phi(p, t)$ para esse caso. Monte também $|\Phi(p, t)|^2$, e note que é independente do tempo.

(c) Calcule $\langle p \rangle$ e $\langle p^2 \rangle$ por meio de avaliação das integrais apropriadas que envolvem Φ e compare suas respostas ao Problema 2.43.

(d) Demonstre que $\langle H \rangle \equiv \langle p \rangle^2/2m + \langle H \rangle_0$ (em que o índice 0 representa o *estacionário* gaussiano) e comente esse resultado.

---

[34] Especialmente se estabelecermos $t = 0$ e descartarmos o índice em t_0,

$$\langle Q(t) \rangle = \langle \Psi(x, t) | \hat{Q} | \Psi(x, t) \rangle = \langle \Psi(x, 0) | \hat{U}^{-1}\hat{Q}\hat{U} | \Psi(x, 0) \rangle,$$

em que $\hat{U} \equiv \exp(-i\hat{H}t/\hbar)$. Isso significa que você pode calcular os valores esperados de Q tanto imprensando $\hat{Q}$ entre $\Psi(x, t)^*$ e $\Psi(x, t)$ quanto como sempre fazemos (deixando as funções de onda carregar a dependência de tempo), ou, ainda, imprensando $\hat{U}^{-1}\hat{Q}\hat{U}$ entre $\Psi(x, 0)^*$ e $\Psi(x, 0)$, deixando assim que o *operador* carregue a dependência do tempo. O primeiro é chamado de **representação de Schrödinger**, e o último é chamado de **representação de Heisenberg**.

')
+,- * *. *) 1 2 " = 8 ' "
D ": %< /"
+,- * *. *) 1 2 " = 8 ' "
D ": !< 6 " I 2 : /"
(! %& %) ? / J DG "

where

$$\text{}\mathscr{r}^2 = z^2 + R^2 - 2zR\cos\theta,$$

so

$$\begin{aligned} V_{\text{ave}} &= \frac{1}{4\pi R^2}\frac{q}{4\pi\epsilon_0}\int [z^2 + R^2 - 2zR\cos\theta]^{-1/2} R^2 \,\sin\theta\, d\theta\, d\phi \\ &= \frac{q}{4\pi\epsilon_0}\frac{1}{2zR}\sqrt{z^2 + R^2 - 2zR\cos\theta}\,\Big|_0^{\pi} \\ &= \frac{q}{4\pi\epsilon_0}\frac{1}{2zR}[(z+R) - (z-R)] = \frac{1}{4\pi\epsilon_0}\frac{q}{z}. \end{aligned}$$

But this is precisely the potential due to q at the *center* of the sphere! By the superposition principle, the same goes for any *collection* of charges outside the sphere: their average potential over the sphere is equal to the net potential they produce at the center. □

Problem 3.1 Find the average potential over a spherical surface of radius R due to a point charge q located *inside* (same as above, in other words, only with $z < R$). (In this case, of course, Laplace's equation does not hold within the sphere.) Show that, in general,

$$V_{\text{ave}} = V_{\text{center}} + \frac{Q_{\text{enc}}}{4\pi\epsilon_0 R},$$

where V_{center} is the potential at the center due to all the *external* charges, and Q_{enc} is the total enclosed charge.

Problem 3.2 In one sentence, justify **Earnshaw's Theorem:** *A charged particle cannot be held in a stable equilibrium by electrostatic forces alone.* As an example, consider the cubical arrangement of fixed charges in Fig. 3.4. It *looks,* off hand, as though a positive charge at the center would be suspended in midair, since it is repelled away from each corner. Where is the leak in this "electrostatic bottle"? [To harness nuclear fusion as a practical energy source it is necessary to heat a plasma (soup of charged particles) to fantastic temperatures—so hot that contact would vaporize any ordinary pot. Earnshaw's theorem says that electrostatic containment is also out of the question. Fortunately, it *is* possible to confine a hot plasma magnetically.]

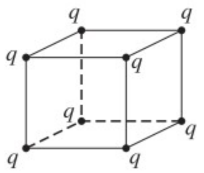

FIGURE 3.4

Problem 3.3 Find the general solution to Laplace's equation in spherical coordinates, for the case where V depends only on r. Do the same for cylindrical coordinates, assuming V depends only on s.

Problem 3.4

(a) Show that the average electric *field* over a spherical surface, due to charges outside the sphere, is the same as the field at the center.

(b) What is the average due to charges *inside* the sphere?

3.1.5 ■ Boundary Conditions and Uniqueness Theorems

Laplace's equation does not by itself determine V; in addition, suitable boundary conditions must be supplied. This raises a delicate question: What are appropriate boundary conditions, sufficient to determine the answer and yet not so strong as to generate inconsistencies? The one-dimensional case is easy, for here the general solution $V = mx + b$ contains two arbitrary constants, and we therefore require two boundary conditions. We might, for instance, specify the value of the function at each end, or we might give the value of the function and its derivative at one end, or the value at one end and the derivative at the other, and so on. But we cannot get away with *just* the value or *just* the derivative at *one* end—this is insufficient information. Nor would it do to specify the derivatives at both ends—this would either be redundant (if the two are equal) or inconsistent (if they are not).

In two or three dimensions we are confronted by a *partial* differential equation, and it is not so obvious what would constitute acceptable boundary conditions. Is the shape of a taut rubber membrane, for instance, uniquely determined by the frame over which it is stretched, or, like a canning jar lid, can it snap from one stable configuration to another? The answer, as I think your intuition would suggest, is that V *is* uniquely determined by its value at the boundary (canning jars evidently do not obey Laplace's equation). However, other boundary conditions can also be used (see Prob. 3.5). The *proof* that a proposed set of boundary conditions will suffice is usually presented in the form of a **uniqueness theorem.** There are many such theorems for electrostatics, all sharing the same basic format—I'll show you the two most useful ones.[4]

> **First uniqueness theorem:** The solution to Laplace's equation in some volume $\mathcal{V}$ is uniquely determined if V is specified on the boundary surface $\mathcal{S}$.

Proof. In Fig. 3.5 I have drawn such a region and its boundary. (There could also be "islands" inside, so long as V is given on all their surfaces; also, the outer

---

[4] I do not intend to prove the *existence* of solutions here—that's a much more difficult job. In context, the existence is generally clear on physical grounds.

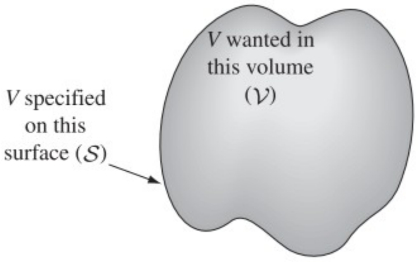

FIGURE 3.5

boundary could be at infinity, where V is ordinarily taken to be zero.) Suppose there were *two* solutions to Laplace's equation:

$$\nabla^2 V_1 = 0 \quad \text{and} \quad \nabla^2 V_2 = 0,$$

both of which assume the specified value on the surface. I want to prove that they must be equal. The trick is look at their *difference:*

$$V_3 \equiv V_1 - V_2.$$

This obeys Laplace's equation,

$$\nabla^2 V_3 = \nabla^2 V_1 - \nabla^2 V_2 = 0,$$

and it takes the value *zero* on all boundaries (since V_1 and V_2 are equal there). But Laplace's equation allows no local maxima or minima—all extrema occur on the boundaries. So the maximum and minimum of V_3 are both zero. Therefore V_3 must be zero everywhere, and hence

$$V_1 = V_2. \qquad \square$$

Example 3.1. Show that the potential is *constant* inside an enclosure completely surrounded by conducting material, provided there is no charge within the enclosure.

Solution

The potential on the cavity wall is some constant, V_0 (that's item (iv), in Sect. 2.5.1), so the potential inside is a function that satisfies Laplace's equation and has the constant value V_0 at the boundary. It doesn't take a genius to think of *one* solution to this problem: $V = V_0$ everywhere. The uniqueness theorem guarantees that this is the *only* solution. (It follows that the *field* inside an empty cavity is zero—the same result we found in Sect. 2.5.2 on rather different grounds.)

The uniqueness theorem is a license to your imagination. It doesn't matter *how* you come by your solution; if (a) it satisfies Laplace's equation and (b) it has the correct value on the boundaries, then it's *right.* You'll see the power of this argument when we come to the method of images.

Incidentally, it is easy to improve on the first uniqueness theorem: I assumed there was no charge inside the region in question, so the potential obeyed Laplace's equation, but we may as well throw in some charge (in which case V obeys Poisson's equation). The argument is the same, only this time

$$\nabla^2 V_1 = -\frac{1}{\epsilon_0}\rho, \qquad \nabla^2 V_2 = -\frac{1}{\epsilon_0}\rho,$$

so

$$\nabla^2 V_3 = \nabla^2 V_1 - \nabla^2 V_2 = -\frac{1}{\epsilon_0}\rho + \frac{1}{\epsilon_0}\rho = 0.$$

Once again the *difference* ($V_3 \equiv V_1 - V_2$) satisfies Laplace's equation and has the value zero on all boundaries, so $V_3 = 0$ and hence $V_1 = V_2$.

Corollary: The potential in a volume $\mathcal{V}$ is uniquely determined if (a) the charge density throughout the region, and (b) the value of V on all boundaries, are specified.

3.1.6 ■ Conductors and the Second Uniqueness Theorem

The *simplest* way to set the boundary conditions for an electrostatic problem is to specify the value of V on all surfaces surrounding the region of interest. And this situation often occurs in practice: In the laboratory, we have conductors connected to batteries, which maintain a given potential, or to **ground**, which is the experimentalist's word for $V = 0$. However, there are other circumstances in which we do not know the *potential* at the boundary, but rather the *charges* on various conducting surfaces. Suppose I put charge Q_a on the first conductor, Q_b on the second, and so on—I'm not telling you how the charge distributes itself over each conducting surface, because as soon as I put it on, it moves around in a way I do not control. And for good measure, let's say there is some specified charge density ρ in the region between the conductors. Is the electric field now uniquely determined? Or are there perhaps a number of different ways the charges could arrange themselves on their respective conductors, each leading to a different field?

Second uniqueness theorem: In a volume $\mathcal{V}$ surrounded by conductors and containing a specified charge density ρ, the electric field is uniquely determined if the *total charge* on each conductor is given (Fig. 3.6). (The region as a whole can be bounded by another conductor, or else unbounded.)

Proof. Suppose there are *two* fields satisfying the conditions of the problem. Both obey Gauss's law in differential form in the space between the conductors:

$$\nabla \cdot \mathbf{E}_1 = \frac{1}{\epsilon_0}\rho, \qquad \nabla \cdot \mathbf{E}_2 = \frac{1}{\epsilon_0}\rho.$$

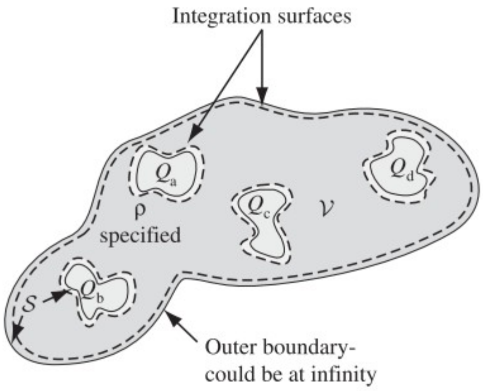

FIGURE 3.6

And both obey Gauss's law in integral form for a Gaussian surface enclosing each conductor:

$$\oint\limits_{\substack{i\text{th conducting}\\ \text{surface}}} \mathbf{E}_1 \cdot d\mathbf{a} = \frac{1}{\epsilon_0} Q_i, \qquad \oint\limits_{\substack{i\text{th conducting}\\ \text{surface}}} \mathbf{E}_2 \cdot d\mathbf{a} = \frac{1}{\epsilon_0} Q_i.$$

Likewise, for the outer boundary (whether this is just inside an enclosing conductor or at infinity),

$$\oint\limits_{\substack{\text{outer}\\ \text{boundary}}} \mathbf{E}_1 \cdot d\mathbf{a} = \frac{1}{\epsilon_0} Q_{\text{tot}}, \qquad \oint\limits_{\substack{\text{outer}\\ \text{boundary}}} \mathbf{E}_2 \cdot d\mathbf{a} = \frac{1}{\epsilon_0} Q_{\text{tot}}.$$

As before, we examine the difference

$$\mathbf{E}_3 \equiv \mathbf{E}_1 - \mathbf{E}_2,$$

which obeys

$$\nabla \cdot \mathbf{E}_3 = 0 \tag{3.7}$$

in the region between the conductors, and

$$\oint \mathbf{E}_3 \cdot d\mathbf{a} = 0 \tag{3.8}$$

over each boundary surface.

Now there is one final piece of information we must exploit: Although we do not know how the charge Q_i distributes itself over the ith conductor, we *do* know that each conductor is an equipotential, and hence V_3 is a *constant* (not necessarily the *same* constant) over each conducting surface. (It need not be *zero,* for the potentials V_1 and V_2 may not be equal—all we know for sure is that *both* are *constant* over any given conductor.) Next comes a trick. Invoking product rule number 5 (inside front cover), we find that

$$\nabla \cdot (V_3 \mathbf{E}_3) = V_3(\nabla \cdot \mathbf{E}_3) + \mathbf{E}_3 \cdot (\nabla V_3) = -(E_3)^2.$$

Here I have used Eq. 3.7, and $\mathbf{E}_3 = -\nabla V_3$. Integrating this over $\mathcal{V}$, and applying the divergence theorem to the left side:

$$\int_{\mathcal{V}} \nabla \cdot (V_3 \mathbf{E}_3)\, d\tau = \oint_{\mathcal{S}} V_3 \mathbf{E}_3 \cdot d\mathbf{a} = -\int_{\mathcal{V}} (E_3)^2\, d\tau.$$

The surface integral covers all boundaries of the region in question—the conductors and outer boundary. Now V_3 is a constant over each surface (if the outer boundary is infinity, $V_3 = 0$ there), so it comes outside each integral, and what remains is zero, according to Eq. 3.8. Therefore,

$$\int_{\mathcal{V}} (E_3)^2\, d\tau = 0.$$

But this integrand is never negative; the only way the integral can vanish is if $E_3 = 0$ everywhere. Consequently, $\mathbf{E}_1 = \mathbf{E}_2$, and the theorem is proved. □

This proof was not easy, and there is a real danger that the theorem itself will seem more plausible to you than the proof. In case you think the second uniqueness theorem is "obvious," consider this example of Purcell's: Figure 3.7 shows a simple electrostatic configuration, consisting of four conductors with charges $\pm Q$, situated so that the plusses are near the minuses. It all looks very comfortable. Now, what happens if we join them in pairs, by tiny wires, as indicated in Fig. 3.8? Since the positive charges are very near negative charges (which is where they *like* to be) you might well guess that *nothing* will happen—the configuration looks stable.

Well, that sounds reasonable, but it's wrong. The configuration in Fig. 3.8 is *impossible.* For there are now effectively *two* conductors, and the total charge on each is *zero*. *One* possible way to distribute zero charge over these conductors is to have no accumulation of charge anywhere, and hence zero field

FIGURE 3.7

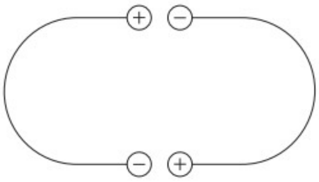

FIGURE 3.8

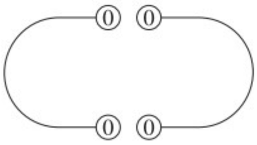

FIGURE 3.9

everywhere (Fig. 3.9). By the second uniqueness theorem, this must be *the* solution: The charge will flow down the tiny wires, canceling itself off.

Problem 3.5 Prove that the field is uniquely determined when the charge density ρ is given and *either* V *or* the normal derivative $\partial V/\partial n$ is specified on each boundary surface. Do not assume the boundaries are conductors, or that V is constant over any given surface.

Problem 3.6 A more elegant proof of the second uniqueness theorem uses Green's identity (Prob. 1.61c), with $T = U = V_3$. Supply the details.

3.2 ■ THE METHOD OF IMAGES

3.2.1 ■ The Classic Image Problem

Suppose a point charge q is held a distance d above an infinite grounded conducting plane (Fig. 3.10). *Question:* What is the potential in the region above the plane? It's not just $(1/4\pi\epsilon_0)q/\mathscr{r}$, for q will induce a certain amount of negative charge on the nearby surface of the conductor; the total potential is due in part to q directly, and in part to this induced charge. But how can we possibly calculate the potential, when we don't know how much charge is induced or how it is distributed?

From a mathematical point of view, our problem is to solve Poisson's equation in the region $z > 0$, with a single point charge q at $(0, 0, d)$, subject to the boundary conditions:

1. $V = 0$ when $z = 0$ (since the conducting plane is grounded), and
2. $V \to 0$ far from the charge $\big($that is, for $x^2 + y^2 + z^2 \gg d^2\big)$.

The first uniqueness theorem (actually, its corollary) guarantees that there is only one function that meets these requirements. If by trick or clever guess we can discover such a function, it's got to be the answer.

Trick: Forget about the actual problem; we're going to study a *completely different* situation. This new configuration consists of *two* point charges, $+q$ at

data 03.04.2020

$$\hat{H} = \epsilon(|1\rangle\langle 1| - |2\rangle\langle 2| + |1\rangle\langle 2| + |2\rangle\langle 1|),$$

em que $|1\rangle, |2\rangle$ é uma base ortonormal e ϵ é um número com as dimensões da energia.

Problema 3.25

Polinômios de Legendre — use procedimento de Gram-Schmidt (Problema A.4) para ortonormalizar as funções.

Problema 3.26

Um operador anti-hermitiano é igual a menos o seu conjugado hermitiano:

$$\hat{Q}^{\dagger} = -\hat{Q} \qquad [3.95]$$

124) Books_63_2020/
Introduction to Quantum Mechanics/
David J. Griffiths - Introduction

03.04.2020

to Electrodynamics(2013, Pearson).
pdf

Pág. 117(122)

1.4 Laplace's Equation in three Dimensions

$$V(r) = \frac{1}{4\pi R^2} \oint_{\text{sphere}} V \, da$$

Figure 3

Obs.: 3 For a proof does not rely on Coulomb's Law (only on Laplace's equation), see Prob. 37

---

125) 20200404

Data_NOVOVIVO_64_2020/Books_64_2020/Mathematics for physics A Guide Tour for Graduate Students/Michael Stone ... Mathematics for physics ... pdf

Pág. 41(57)

Problem 1.6: the catenary revisited

data 04.04.2020

Pág. 42(57)

Figure 1.15 Weighted line

Figure 1.16 the Poincaré disc of Exercise 1.7.

Problem 1.7.

Another model for Lobachevski geometry (see Exercise 1.3) is the Poincaré disc (Figure 1.16)

This space consists of the interior of the unit disc

$$D^2 = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$$

equipped with the Riemann metric

$$ds^2 = \frac{dx^2 + dy^2}{(1 - x^2 - y^2)^2}$$

$(-)$

126) DATA_NOVOVIVO_64_2020/
Books_64_2020/
Geometry/
The Ricci Flow in Riemannian Geometry_A Complete Proof of the Differentiable 1-4 -

Jandaia

data 04 04 2020
S T Q Q S [✓ S] D

Pinching Sphere theorem
Chapter 4
Evolution of the Curvature
Source: Pág. 63 (82)

127) 20200406 —

A) Data-NOVOVIVO_65/2020/
Books_65_2020/
Higher Algebra Classical/
Higher Algebra Classical by Sadhan...pdf
Pág. 4 → 2. Complex Numbers

B) Books_65_2020/
Descriptive Geometry/
Geometric and Engineering Drawing
by K. Morling...pdf

Pág. 5(22)

Diagonal Scales

Pág. 6(23) Figure 1.4

Pág. 7(24) Figure 1.5
Figure 1.6

Pág. 8(25) Figure 1.7

Pág. 11(28) 2 → The Construction of

1.7 Further exercises and problems

Here is a collection of problems relating to the calculus of variations. Some date back to the sixteenth century, others are quite recent in origin.

Exercise 1.1: A smooth path in the xy-plane is given by $\mathbf{r}(t) = (x(t), y(t))$ with $\mathbf{r}(0) = \mathbf{a}$, and $\mathbf{r}(1) = \mathbf{b}$. The length of the path from $\mathbf{a}$ to $\mathbf{b}$ is therefore

$$S[\mathbf{r}] = \int_0^1 \sqrt{\dot{x}^2 + \dot{y}^2}\, dt,$$

where $\dot{x} \equiv dx/dt$, $\dot{y} \equiv dy/dt$. Write down the Euler–Lagrange conditions for $S[\mathbf{r}]$ to be stationary under small variations of the path that keep the endpoints fixed, and hence show that the shortest path between two points is a straight line.

Exercise 1.2: *Fermat's principle*. A medium is characterized optically by its *refractive index n*, such that the speed of light in the medium is c/n. According to Fermat (1657), the path taken by a ray of light between any two points makes the travel time stationary between those points. Assume that the ray propagates in the xy-plane in a layered medium with refractive index $n(x)$. Use Fermat's principle to establish Snell's law in its general form $n(x) \sin\psi = \text{constant}$, by finding the equation giving the stationary paths $y(x)$ for

$$F_1[y] = \int n(x)\sqrt{1 + y'^2}\, dx.$$

(Here the prime denotes differentiation with respect to x.) Repeat this exercise for the case that n depends only on y and find a similar equation for the stationary paths of

$$F_2[y] = \int n(y)\sqrt{1 + y'^2}\, dx.$$

By using suitable definitions of the angle of incidence ψ in each case, show that the two formulations of the problem give physically equivalent answers. In the second formulation you will find it easiest to use the first integral of Euler's equation.

How are we to make sense of this expression? We begin in the next chapter where we explain what it means to differentiate $\delta(x)$, and show that (1.186) does indeed correspond to the differential operator L. In subsequent chapters we explore the manner in which differential operators and matrices are related. We will learn that just as some matrices can be diagonalized so can some differential operators, and that the class of diagonalizable operators includes (1.182).

If all the eigenvalues of L are positive, our stationary point was a minimum. For each negative eigenvalue, there is direction in function space in which $J[y]$ decreases as we move away from the stationary point.

1.7 Further exercises and problems

Here is a collection of problems relating to the calculus of variations. Some date back to the sixteenth century, others are quite recent in origin.

Exercise 1.1: A smooth path in the xy-plane is given by $\mathbf{r}(t) = (x(t), y(t))$ with $\mathbf{r}(0) = \mathbf{a}$, and $\mathbf{r}(1) = \mathbf{b}$. The length of the path from $\mathbf{a}$ to $\mathbf{b}$ is therefore

$$S[\mathbf{r}] = \int_0^1 \sqrt{\dot{x}^2 + \dot{y}^2}\, dt,$$

where $\dot{x} \equiv dx/dt$, $\dot{y} \equiv dy/dt$. Write down the Euler–Lagrange conditions for $S[\mathbf{r}]$ to be stationary under small variations of the path that keep the endpoints fixed, and hence show that the shortest path between two points is a straight line.

Exercise 1.2: *Fermat's principle*. A medium is characterized optically by its *refractive index n*, such that the speed of light in the medium is c/n. According to Fermat (1657), the path taken by a ray of light between any two points makes the travel time stationary between those points. Assume that the ray propagates in the xy-plane in a layered medium with refractive index $n(x)$. Use Fermat's principle to establish Snell's law in its general form $n(x)\sin\psi = \text{constant}$, by finding the equation giving the stationary paths $y(x)$ for

$$F_1[y] = \int n(x)\sqrt{1 + y'^2}\, dx.$$

(Here the prime denotes differentiation with respect to x.) Repeat this exercise for the case that n depends only on y and find a similar equation for the stationary paths of

$$F_2[y] = \int n(y)\sqrt{1 + y'^2}\, dx.$$

By using suitable definitions of the angle of incidence ψ in each case, show that the two formulations of the problem give physically equivalent answers. In the second formulation you will find it easiest to use the first integral of Euler's equation.

.6 9 J /

Problem 1.3: *Hyperbolic geometry*. This problem introduces a version of the Poincaré model for the non-Euclidean geometry of Lobachevski.

(a) Show that the stationary paths for the functional

$$F_3[y] = \int \frac{1}{y}\sqrt{1+y'^2}dx,$$

with $y(x)$ restricted to lying in the upper half-plane, are semicircles of arbitrary radius and with centres on the x-axis. These paths are the *geodesics*, or minimum length paths, in a space with Riemann metric

$$ds^2 = \frac{1}{y^2}(dx^2 + dy^2), \qquad y > 0.$$

(b) Show that if we call these geodesics "lines", then one and only one line can be drawn though two given points.

(c) Two lines are said to be *parallel* if, and only if, they meet at "infinity", i.e. on the x-axis. (Verify that the x-axis is indeed infinitely far from any point with $y > 0$.) Show that given a line q and a point A not lying on that line, there are *two* lines passing through A that are parallel to q, and that between these two lines lies a pencil of lines passing through A that never meet q.

Problem 1.4: *Elastic rods.* The elastic energy per unit length of a bent steel rod is given by $\frac{1}{2}YI/R^2$. Here R is the radius of curvature due to the bending, Y is the Young's modulus of the steel and $I = \iint y^2 dxdy$ is the moment of inertia of the rod's cross-section about an axis through its centroid and perpendicular to the plane in which the rod is bent. If the rod is only slightly bent into the yz-plane and lies close to the z-axis, show that this elastic energy can be approximated as

$$U[y] = \int_0^L \frac{1}{2} YI \left(y''\right)^2 \, dz,$$

where the prime denotes differentiation with respect to z and L is the length of the rod. We will use this approximate energy functional to discuss two practical problems.

(a) *Euler's problem: The buckling of a slender column*. The rod is used as a column which supports a compressive load Mg directed along the z-axis (which is vertical; see Figure (1.14a)). Show that when the rod buckles slightly (i.e. deforms with both ends remaining on the z-axis) the total energy, including the gravitational potential energy of the loading mass M, can be approximated by

$$U[y] = \int_0^L \left\{ \frac{YI}{2} \left(y''\right)^2 - \frac{Mg}{2} \left(y'\right)^2 \right\} dz.$$

.6 7:/ 6 8 : "D J 6

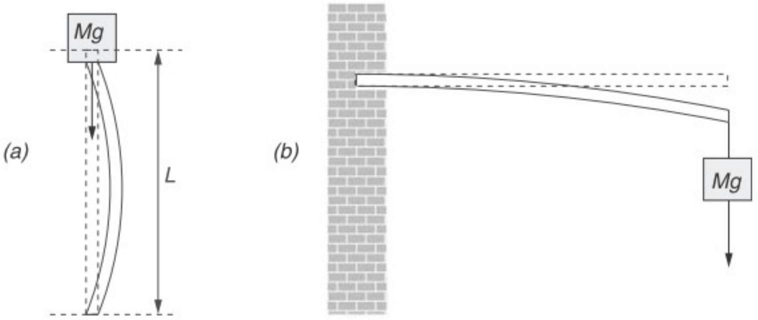

Figure 1.14 A rod used as: (a) a column, (b) a cantilever.

By considering small deformations of the form

$$y(z) = \sum_{n=1}^{\infty} a_n \sin \frac{n\pi z}{L}$$

show that the column is unstable to buckling and collapse if $Mg \geq \pi^2 YI/L^2$.

(b) *Leonardo da Vinci's problem: The light cantilever.* Here we take the z-axis as horizontal and the y-axis as being vertical (Figure 1.14b). The rod is used as a beam or cantilever and is fixed into a wall so that $y(0) = 0 = y'(0)$. A weight Mg is hung from the end $z = L$ and the beam sags in the $(-y)$-direction. We wish to find $y(z)$ for $0 < z < L$. We will ignore the weight of the beam itself.

- Write down the complete expression for the energy, including the gravitational potential energy of the weight.
- Find the differential equation and boundary conditions at $z = 0, L$ that arise from minimizing the total energy. In doing this take care not to throw away any term arising from the integration by parts. You may find the following identity to be of use:

$$\frac{d}{dz}(f'g'' - fg''') = f''g'' - fg''''.$$

- Solve the equation. You should find that the displacement of the end of the beam is $y(L) = -\frac{1}{3}MgL^3/YI$.

Exercise 1.5: Suppose that an elastic body Ω of density ρ is slightly deformed so that the point that was at cartesian coordinate x_i is moved to $x_i + \eta_i(x)$. We define the resulting *strain tensor* e_{ij} by

$$e_{ij} = \frac{1}{2}\left(\frac{\partial \eta_j}{\partial x_i} + \frac{\partial \eta_i}{\partial x_j}\right).$$

It is automatically symmetric in its indices. The Lagrangian for small-amplitude elastic motion of the body is

$$L[\boldsymbol{\eta}] = \int_\Omega \left\{ \frac{1}{2}\rho\dot{\eta}_i^2 - \frac{1}{2}e_{ij}c_{ijkl}e_{kl} \right\} d^3x.$$

Here, c_{ijkl} is the tensor of *elastic constants*, which has the symmetries

$$c_{ijkl} = c_{klij} = c_{jikl} = c_{ijlk}.$$

By varying the η_i, show that the equation of motion for the body is

$$\rho\frac{\partial^2\eta_i}{\partial t^2} - \frac{\partial}{\partial x_j}\sigma_{ji} = 0,$$

where

$$\sigma_{ij} = c_{ijkl}e_{kl}$$

is the *stress tensor*. Show that variations of η_i on the boundary $\partial\Omega$ give as boundary conditions

$$\sigma_{ij}n_j = 0,$$

where n_i are the components of the outward normal on $\partial\Omega$.

Problem 1.6: *The catenary revisited.* We can describe a catenary curve in parametric form as $x(s)$, $y(s)$, where s is the arc-length. The potential energy is then simply $\int_0^L \rho g y(s) ds$ where ρ is the mass per unit length of the hanging chain. The x, y are not independent functions of s, however, because $\dot{x}^2 + \dot{y}^2 = 1$ at every point on the curve. Here a dot denotes a derivative with respect to s.

(a) Introduce infinitely many Lagrange multipliers $\lambda(s)$ to enforce the $\dot{x}^2 + \dot{y}^2$ constraint, one for each point s on the curve. From the resulting functional derive two coupled equations describing the catenary, one for $x(s)$ and one for $y(s)$. By thinking about the forces acting on a small section of the cable, and perhaps by introducing the angle ψ where $\dot{x} = \cos\psi$ and $\dot{y} = \sin\psi$, so that s and ψ are *intrinsic coordinates* for the curve, interpret these equations and show that $\lambda(s)$ is proportional to the position-dependent tension $T(s)$ in the chain.

(b) You are provided with a lightweight line of length $\pi a/2$ and some lead shot of total mass M. By using equations from the previous part (suitably modified to take into account the position dependent $\rho(s)$) or otherwise, determine how the lead should be distributed along the line if the loaded line is to hang in an arc of a circle of radius a (see Figure 1.15) when its ends are attached to two points at the same height.

.6 ? " : 2 "

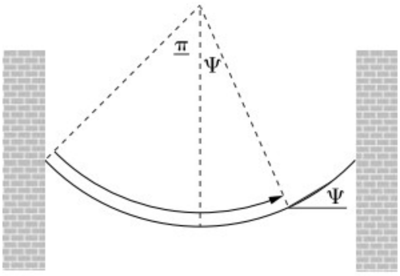

Figure 1.15 Weighted line.

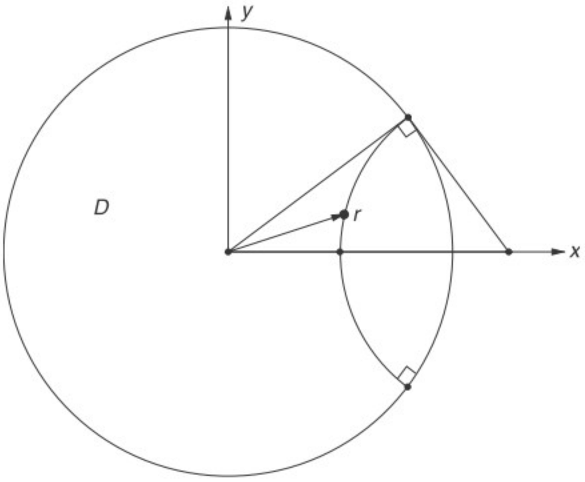

Figure 1.16 The Poincaré disc of Exercise 1.7. The radius OP of the Poincaré disc is unity, while the radius of the geodesic arc PQR is PX = QX = RX = R. The distance between the centres of the disc and arc is OX = x_0. Your task in part (c) is to show that ∠OPX = ∠ORX = 90°.

Problem 1.7: Another model for Lobachevski geometry (see Exercise 1.3) is the *Poincaré disc* (Figure 1.16). This space consists of the interior of the unit disc $D^2 = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$ equipped with the Riemann metric

$$ds^2 = \frac{dx^2 + dy^2}{(1 - x^2 - y^2)^2}.$$

The geodesic paths are found by minimizing the arc-length functional

$$s[\mathbf{r}] \equiv \int ds = \int \left\{ \frac{1}{1 - x^2 - y^2} \sqrt{\dot{x}^2 + \dot{y}^2} \right\} dt,$$

where $\mathbf{r}(t) = (x(t), y(t))$ and a dot indicates a derivative with respect to the parameter t.

(a) Either by manipulating the two Euler–Lagrange equations that give the conditions for $s[\mathbf{r}]$ to be stationary under variations in $\mathbf{r}(t)$, or, more efficiently, by observing that $s[\mathbf{r}]$ is invariant under the infinitesimal rotation

$$\begin{aligned} \delta x &= \varepsilon y \\ \delta y &= -\varepsilon x \end{aligned}$$

and applying Noether's theorem, show that the parametrized geodesics obey

$$\frac{d}{dt}\left(\frac{1}{1-x^2-y^2}\frac{x\dot{y}-y\dot{x}}{\sqrt{\dot{x}^2+\dot{y}^2}}\right)=0.$$

(b) Given a point (a, b) within D^2, and a direction through it, show that the equation you derived in part (a) determines a unique geodesic curve passing through (a, b) in the given direction, but does not determine the parametrization of the curve.

(c) Show that there exists a solution to the equation in part (a) in the form

$$\begin{aligned} x(t) &= R\cos t + x_0 \\ y(t) &= R\sin t. \end{aligned}$$

Find a relation between x_0 and R, and from it deduce that the geodesics are circular arcs that cut the bounding unit circle (which plays the role of the line at infinity in the Lobachevski plane) at right angles.

Exercise 1.8: The Lagrangian for a particle of charge q is

$$L[\mathbf{x}, \dot{\mathbf{x}}] = \frac{1}{2}m\dot{\mathbf{x}}^2 - q\phi(\mathbf{x}) + q\dot{\mathbf{x}}\cdot\mathbf{A}(\mathbf{x}).$$

Show that Lagrange's equation leads to

$$m\ddot{\mathbf{x}} = q(\mathbf{E} + \dot{\mathbf{x}}\times\mathbf{B}),$$

where

$$\mathbf{E} = -\nabla\phi - \frac{\partial\mathbf{A}}{\partial t}, \quad \mathbf{B} = \operatorname{curl}\mathbf{A}.$$

Exercise 1.9: Consider the action functional

$$S[\boldsymbol{\omega}, \mathbf{p}, \mathbf{r}] = \int\left(\frac{1}{2}I_1\omega_1^2 + \frac{1}{2}I_2\omega_2^2 + \frac{1}{2}I_3\omega_3^2 + \mathbf{p}\cdot(\dot{\mathbf{r}} + \boldsymbol{\omega}\times\mathbf{r})\right\}\,dt,$$

where $\mathbf{r}$ and $\mathbf{p}$ are time-dependent 3-vectors, as is $\boldsymbol{\omega} = (\omega_1, \omega_2, \omega_3)$. Apply the action principle to obtain the equations of motion for $\mathbf{r}, \mathbf{p}, \boldsymbol{\omega}$ and show that they lead to Euler's

.6 D ? DG

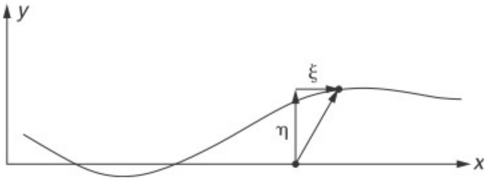

Figure 1.17 Vibrating piano string.

equations

$$\begin{aligned} I_1\dot{\omega}_1 - (I_2 - I_3)\omega_2\omega_3 &= 0, \\ I_2\dot{\omega}_2 - (I_3 - I_1)\omega_3\omega_1 &= 0, \\ I_3\dot{\omega}_3 - (I_1 - I_2)\omega_1\omega_2 &= 0, \end{aligned}$$

governing the angular velocity of a freely rotating rigid body.

Problem 1.10: *Piano string*. An elastic piano string can vibrate both transversely and longitudinally, and the two vibrations influence one another (Figure 1.17). A Lagrangian that takes into account the lowest-order effect of stretching on the local string tension, and can therefore model this coupled motion, is

$$L[\xi, \eta] = \int dx \left\{ \frac{1}{2}\rho_0 \left[\left(\frac{\partial \xi}{\partial t}\right)^2 + \left(\frac{\partial \eta}{\partial t}\right)^2 \right] - \frac{\lambda}{2} \left[\frac{\tau_0}{\lambda} + \frac{\partial \xi}{\partial x} + \frac{1}{2}\left(\frac{\partial \eta}{\partial x}\right)^2 \right]^2 \right\}.$$

Here $\xi(x,t)$ is the longitudinal displacement and $\eta(x,t)$ the transverse displacement of the string. Thus, the point that in the undisturbed string had coordinates $[x, 0]$ is moved to the point with coordinates $[x+\xi(x,t), \eta(x,t)]$. The parameter τ_0 represents the tension in the undisturbed string, λ is the product of Young's modulus and the cross-sectional area of the string and ρ_0 is the mass per unit length.

(a) Use the action principle to derive the two coupled equations of motion, one involving $\dfrac{\partial^2 \xi}{\partial t^2}$ and one involving $\dfrac{\partial^2 \eta}{\partial t^2}$.

(b) Show that when we linearize these two equations of motion, the longitudinal and transverse motions decouple. Find expressions for the longitudinal (c_{L}) and transverse (c_{T}) wave velocities in terms of τ_0, ρ_0 and λ.

(c) Assume that a given transverse pulse $\eta(x,t) = \eta_0(x - c_{\mathrm{T}}t)$ propagates along the string. Show that this induces a concurrent longitudinal pulse of the form $\xi(x - c_{\mathrm{T}}t)$. Show further that the longitudinal Newtonian momentum density in this concurrent pulse is given by

$$\rho_0 \frac{\partial \xi}{\partial t} = \frac{1}{2} \frac{c_{\mathrm{L}}^2}{c_{\mathrm{L}}^2 - c_{\mathrm{T}}^2} T^0{}_1$$

.6

where

$$T^0{}_1 \equiv -\rho_0 \frac{\partial \eta}{\partial x}\frac{\partial \eta}{\partial t}$$

is the associated pseudo-momentum density.

The forces that created the transverse pulse will also have created other longitudinal waves that travel at c_L. Consequently the Newtonian x-momentum moving at c_T is not the only x-momentum on the string, and the total "true" longitudinal momentum density is not simply proportional to the pseudo-momentum density.

Exercise 1.11: Obtain the canonical energy–momentum tensor $T^\nu{}_\mu$ for the barotropic fluid described by (1.119). Show that its conservation leads to both the momentum conservation equation (1.128), and the energy conservation equation

$$\partial_t \mathcal{E} + \partial_i \{v_i(\mathcal{E} + P)\},$$

where the energy density is

$$\mathcal{E} = \frac{1}{2}\rho(\nabla\phi)^2 + u(\rho).$$

Interpret the energy flux as being the sum of the convective transport of energy together with the rate of working by an element of fluid on its neighbours.

Problem 1.12: Consider the action functional[3]

$$S[\mathbf{v}, \rho, \phi, \beta, \gamma] = \int d^4x \left\{ -\frac{1}{2}\rho\mathbf{v}^2 - \phi\left(\frac{\partial \rho}{\partial t} + \mathrm{div}\,(\rho\mathbf{v})\right) + \rho\beta\left(\frac{\partial \gamma}{\partial t} + (\mathbf{v}\cdot\nabla)\gamma\right) + u(\rho)\right\},$$

which is a generalization of (1.177) to include two new scalar fields β and γ. Show that varying $\mathbf{v}$ leads to

$$\mathbf{v} = \nabla\phi + \beta\nabla\gamma.$$

This is the *Clebsch representation* of the velocity field. It allows for flows with non-zero vorticity

$$\boldsymbol{\omega} \equiv \mathrm{curl}\,\mathbf{v} = \nabla\beta \times \nabla\gamma.$$

---

[3] H. Bateman, *Proc. Roy. Soc. Lond.* A, **125** (1929) 598; C. C. Lin, *Liquid Helium* in *Proc. Int. Sch. Phys. "Enrico Fermi", Course XXI* (Academic Press, 1965).

.6 6 / 2 : "

Show that the equations that arise from varying the remaining fields $\rho, \phi, \beta, \gamma$ together imply the mass conservation equation

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \mathbf{v}) = 0,$$

and Bernoulli's equation in the form

$$\frac{\partial \mathbf{v}}{\partial t} + \boldsymbol{\omega} \times \mathbf{v} = -\nabla \left(\frac{1}{2}\mathbf{v}^2 + h \right).$$

(Recall that $h = du/d\rho$.) Show that this form of Bernoulli's equation is equivalent to Euler's equation

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} = -\nabla h.$$

Consequently S provides an action principle for a general inviscid barotropic flow.

Exercise 1.13: *Drums and membranes*. The shape of a distorted drumskin is described by the function $h(x, y)$, which gives the height to which the point (x, y) of the flat undistorted drumskin is displaced.

(a) Show that the area of the distorted drumskin is equal to

$$\text{Area}[h] = \int dx\, dy \sqrt{1 + \left(\frac{\partial h}{\partial x}\right)^2 + \left(\frac{\partial h}{\partial y}\right)^2},$$

where the integral is taken over the area of the flat drumskin.

(b) Show that for small distortions, the area reduces to

$$\mathcal{A}[h] = \text{const.} + \frac{1}{2}\int dx\, dy\, |\nabla h|^2.$$

(c) Show that if h satisfies the two-dimensional Laplace equation then $\mathcal{A}$ is stationary with respect to variations that vanish at the boundary.

(d) Suppose the drumskin has mass ρ_0 per unit area, and surface tension T. Write down the Lagrangian controlling the motion of the drumskin, and derive the equation of motion that follows from it.

Problem 1.14: *The Wulff construction.* The surface-area functional of the previous exercise can be generalized so as to find the equilibrium shape of a crystal. We describe the crystal surface by giving its height $z(x, y)$ above the xy-plane, and introduce the direction-dependent surface tension (the surface free-energy per unit area) $\alpha(p, q)$, where

$$p = \frac{\partial z}{\partial x}, \quad q = \frac{\partial z}{\partial y}. \qquad (\star)$$

.6 ; L

We seek to minimize the total surface free energy

$$F[z] = \int dxdy \left\{ \alpha(p,q)\sqrt{1+p^2+q^2} \right\},$$

subject to the constraint that the volume of the crystal

$$V[z] = \int z\, dxdy$$

remains constant.

(a) Enforce the volume constraint by introducing a Lagrange multiplier $2\lambda^{-1}$, and so obtain the Euler–Lagrange equation

$$\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial p}\right) + \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial q}\right) = 2\lambda^{-1}.$$

Here

$$f(p,q) = \alpha(p.q)\sqrt{1+p^2+q^2}.$$

(b) Show in the isotropic case, where α is constant, that

$$z(x,y) = \sqrt{(\alpha\lambda)^2 - (x-a)^2 - (y-b)^2} + \text{const.}$$

is a solution of the Euler–Lagrange equation. In this case, therefore, the equilibrium shape is a sphere.

An obvious way to satisfy the Euler–Lagrange equation in the general anisotropic case would be to arrange things so that

$$x = \lambda\frac{\partial f}{\partial p}, \quad y = \lambda\frac{\partial f}{\partial q}. \tag{$\star\star$}$$

(c) Show that ($\star\star$) is exactly the relationship we would have if $z(x,y)$ and $\lambda f(p,q)$ were Legendre transforms of each other, i.e. if

$$\lambda f(p,q) = px + qy - z(x,y),$$

where the x and y on the right-hand side are functions of p, q obtained by solving ($\star$). Do this by showing that the inverse relation is

$$z(x,y) = px + qy - \lambda f(p,q)$$

where now the p, q on the right-hand side become functions of x and y, and are obtained by solving ($\star\star$).

.6 D ? G /

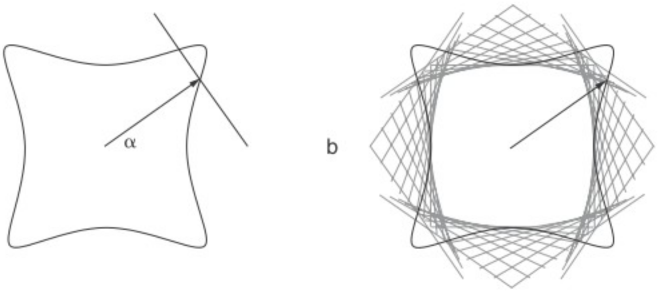

Figure 1.18 Two-dimensional Wulff crystal. (a) Polar plot of surface tension α as a function of the normal **n** to a crystal face, together with a line perpendicular to **n** at distance α from the origin. (b) Wulff's construction of the corresponding crystal surface as the envelope of the family of perpendicular lines. In this case, the minimum-energy crystal has curved faces, but sharp corners. The envelope continues beyond the corners, but these parts are unphysical.

For real crystals, $\alpha(p,q)$ can have the property of being a continuous-but-nowhere-differentiable function, and so the differential calculus used in deriving the Euler–Lagrange equation is inapplicable. The Legendre transformation, however, has a geometric interpretation that is more robust than its calculus-based derivation.

Recall that if we have a two-parameter family of surfaces in $\mathbb{R}^3$ given by $F(x,y,z;p,q) = 0$, then the equation of the *envelope* of the surfaces is found by solving the equations

$$0 = F = \frac{\partial F}{\partial p} = \frac{\partial F}{\partial q}$$

so as to eliminate the parameters p, q.

(d) Show that the equation

$$F(x,y,z;p,q) \equiv px + qy - z - \lambda\alpha(p,q)\sqrt{1+p^2+q^2} = 0$$

describes a family of planes perpendicular to the unit vectors

$$\mathbf{n} = \frac{(p,q,-1)}{\sqrt{1+p^2+q^2}}$$

and at a distance $\lambda\alpha(p,q)$ away from the origin.

(e) Show that the equations to be solved for the envelope of this family of planes are exactly those that determine $z(x,y)$. Deduce that, for smooth $\alpha(p,q)$, the profile $z(x,y)$ is this envelope.

Wulff conjectured[4] that, even for non-smooth $\alpha(p,q)$, the minimum-energy shape is given by an equivalent geometric construction: erect the planes from part (d) and, for each plane, discard the half-space of $\mathbb{R}^3$ that lies on the far side of the plane from the origin. The convex region consisting of the intersection of the retained half-spaces is the crystal. When $\alpha(p, q)$ is smooth this "Wulff body" is bounded by part of the envelope of the planes. (The parts of the envelope not bounding the convex body – the "swallowtails" visible in Figure 1.18 – are unphysical.) When $\alpha(p, q)$ has cusps, these singularities can give rise to flat facets which are often joined by rounded edges. A proof of Wulff's claim had to wait 43 years until 1944, when it was established by use of the Brunn–Minkowski inequality.[5]

---

[4] G. Wulff, *Zeitschrift für Kristallografie,* **34** (1901) 449.

[5] A. Dinghas, *Zeitshrift für Kristallografie,* **105** (1944) 304. For a readable modern account see: R. Gardner, *Bulletin Amer. Math. Soc.* **39** (2002) 355.

03 04 2020

to Electrodynamics(2013, Pearson).
pdf

Pág. 117(122)

1.4 Laplace's Equation in three Dimensions

$$V(r) = \frac{1}{4\pi R^2} \oint_{sphere} V \, da$$

Figure 3

Obs: 3 For a proof does not rely on Coulomb's law (only on Laplace's equation), see Prob. 37

125) 20200404

Data_NOVOVIVO_64_2020/Books_64_2020/Mathematics for physics A Guide Tour for Graduate Students/Michael Stone ... Mathematics for physics ...pdf

Pág. 41(57)

Problem 1.6: the catenary revisited

data 04 · 04 2020

Pág. 42(57)

Figure 1.15 Weighted line

Figure 1.16 The Poincaré disc of Exercise 1.7.

Problem 1.7.

Another model for Lobachevski geometry (see Exercise 1.3) is the Poincaré disc (Figure 1.16)

This space consists of the interior of the unit disc

$$D^2=\{(x,y)\in\mathbb{R}^2 : x^2+y^2\leq 1\}$$

equipped with the Riemann metric

$$ds^2 = \frac{dx^2+dy^2}{(1-x^2 \underset{(-)}{-} y^2)^2}$$

126) DATA_NOVOVIVO_64_2020/
Books_64_2020/
Geometry/
The Ricci Flow in Riemannian Geometry_A Complete Proof of the Differentiable 1-4 -

' &
+,- * *. * & ; , 9 ,
8 :> + / 1 6 >) / ; < /
3 7 " 6 /"
(&% 5 /) D2 2

Chapter 4
Evolution of the Curvature

The Ricci flow is introduced in this chapter as a geometric heat-type equation for the metric. In Sect. 4.4 we derive evolution equations for the curvature, and its various contractions, whenever the metric evolves by Ricci flow. These equations, particularly that of Theorem 4.14, are pivotal to our analysis throughout the coming chapters. In Sect. 4.5.3 we discuss a historical result concerning the convergence theory for the Ricci flow in n-dimensions. This will motivational much of the Böhm and Wilking analysis discussed in Chap. 11.

4.1 Introducing the Ricci Flow

The Ricci flow was first introduced by Richard Hamilton in the early 1980s. Inspired by Eells and Sampson's work on Harmonic map heat flow [ES64] (where they take maps between manifolds and try to 'make them better'), Hamilton speculated that it should be possible to take other geometric objects, for instance the metric g_{ij}, and try to 'improve it' by means of a heat-type equation.

In looking for a suitable parabolic equation, one would like $\frac{\partial g_{ij}}{\partial t}$ to equal a Laplacian-type expression involving second-order derivatives of the metric. Computing derivatives of the metric with the Levi–Civita connection is of no help, as they vanish in normal coordinates. However, computing derivatives with respect to a fixed background connection does give something non-trivial. In particular, the Ricci tensor in Lemma 2.47 has an expression which involves second derivatives the components of the metric, thus it is a natural candidate for such a $(2,0)$-tensor (specifically, the last term in the bracket in Lemma 2.47 is $-\frac{1}{2}$ times a 'Laplacian' of the metric computed using coordinate second derivatives). Taking this factor into account, we are led to the evolution equation

$$\frac{\partial}{\partial t} g_{ij} = -2\,\mathrm{Ric}_{ij} \tag{4.1}$$

B. Andrews and C. Hopper, *The Ricci Flow in Riemannian Geometry*,
Lecture Notes in Mathematics 2011, DOI 10.1007/978-3-642-16286-2_4,
© Springer-Verlag Berlin Heidelberg 2011

known simply as the Ricci flow. Further motivation comes from the expression for the metric in exponential coordinates given in Theorem 2.65, as it implies that the Laplacian of the metric computed at the origin in exponential coordinates is equal to $-\frac{2}{3}\mathrm{Ric}_{ij}$.

In contrast to other natural geometric equations (such as the minimal surface equation and its parabolic analogue the mean curvature flow; the harmonic map equation and the associated flow; and many other examples), the Ricci flow was not initially derived as the gradient flow of a geometric functional. However despite this, Perelman's work [Per02] shows there is a natural energy functional lying behind the Ricci flow. We discuss this in Chap. 10.

In light of (4.1), it is clear that we are no longer interested in just a single manifold (M, g), but rather – assuming a suitable local existence theory – a manifold with a one-parameter family of metrics $t \mapsto g(t)$ parametrised by a 'time' t. We adopt this point of view here, taking the metrics $g(t)$ as sections of the fixed bundle $\mathrm{Sym}^2 T^*M$. The rate of change in time of the metric makes sense since at each point $p \in M$ we are simply differentiating $g(t)$ in the vector space given by the fibre of this bundle at p. Later (in Chap. 6) we will see that there are advantages in instead working with bundles defined over $M \times \mathbb{R}$, where a more geometrically meaningful notion of the time derivative can be used.

4.1.1 Exact Solutions

In order to get a feel for the evolution equation, we present some simple solutions of the Ricci flow.

4.1.1.1 Einstein Metrics

If the initial metric is Ricci flat, that is Ric=0, then clearly the metric remains stationary for all subsequent times. Concrete examples of this are the Euclidean space $\mathbb{R}^n$ and the flat torus $\mathbb{T}^n = S^1 \times \cdots \times S^1$.

If the initial metric is Einstein, that is $\mathrm{Ric}(g_0) = \lambda g_0$ for some $\lambda \in \mathbb{R}$, then a solution $g(t)$ with $g(0) = g_0$ is given by $g(t) = (1 - 2\lambda t)g_0$. The cases $\lambda > 0$, $\lambda = 0$ and $\lambda < 0$ correspond to shrinking, steady and expanding solutions. The simplest shrinking solution is that of the unit sphere (S^n, g_0). Here $\mathrm{Ric}(g_0) = (n-1)g_0$, so $g(t) = (1 - 2(n-1)t)g_0$. Thus the sphere will collapse to a point in finite time $T = 1/2(n-1)$. By contrast, if the initial metric g_0 were hyperbolic, $\mathrm{Ric}(g_0) = -(n-1)g_0$ so the evolution $g(t) = (1 + 2(n-1)t)g_0$ will expand the manifold homothetically for all time. In this case, the solution only goes back to $T = -1/2(n-1)$ upon which the metric explodes out of a single point.

4.1.1.2 Quotient Metrics

If the initial Riemannian manifold $N = M/\mathcal{G}$ is a quotient of a Riemannian manifold M by a discrete group of isometries $\mathcal{G}$, it will remain so under the Ricci flow – as the flow on M preserves the isometry group. For example, the projective space $\mathbb{R}P^n = S^n/\mathbb{Z}_2$ of constant curvature shrinks to a point, as does its cover S^n.

4.1.1.3 Product Metrics

If we take the product metric (cf. Sect. 2.4.2) on a product manifold $M \times N$ initially, the metric will remain a product metric under the Ricci flow. Hence the metric on each factor evolves by the Ricci flow independently of the other.

For example, on $S^2 \times S^1$ the S^2 shrinks to a point while S^1 stays fixed so that the manifold collapses to a circle. Moreover, if we take any Riemannian manifold (M, g) and evolve the metric g by the Ricci flow, then it will also evolve on the product manifold $M \times \mathbb{R}^n$, $n \geq 1$, with product metric $g \oplus (dx_1^2 + \cdots + dx_n^2)$ – since the flow is stationary on the $\mathbb{R}^n$ part.

4.1.2 Diffeomorphism Invariance

The curvature of a manifold M can be thought of as the obstruction to being locally isometric to Euclidean space. Indeed, a Riemannian manifold is flat if and only if its curvature tensor vanishes identically. Moreover, recall that:

Theorem 4.1. *The Riemann curvature tensor R is invariant under local isometries: If $\varphi : (M, g) \to (\widetilde{M}, \widetilde{g})$ is a local isometry, then $\varphi^* \widetilde{R} = R$.*

So if $\varphi : M \to \widetilde{M}$ is a time-*independent* diffeomorphism such that $g(t) = \varphi^* \widetilde{g}(t)$ and $\widetilde{g}$ is a solution of the Ricci flow, we see that

$$\begin{aligned} \frac{\partial}{\partial t} g &= \varphi^* \Big(\frac{\partial}{\partial t} \widetilde{g} \Big) \\ &= \varphi^* \big(-2 \operatorname{Ric}(\widetilde{g}) \big) \\ &= -2 \operatorname{Ric}(\varphi^* \widetilde{g}) \\ &= -2 \operatorname{Ric}(g), \end{aligned}$$

where the second last equality is due to Theorem 4.1. Hence g is also a solution to the Ricci flow. Therefore we conclude that (4.1) is invariant under the full diffeomorphism group.[1]

---

[1] Note that if φ is time-*dependent*, an extra Lie derivative term is introduced into the equation (cf. Sect. 5.4).

4.1.2.1 Preservation of Symmetries

The Ricci flow preserves any symmetries that are present in the initial metric. To see this, note that each symmetry is an isometric diffeomorphism of the initial metric; so the pullback of a solution of the Ricci flow by this diffeomorphism gives a solution of Ricci flow. Since the symmetry is an isometry of the initial metric, these are solutions of Ricci flow with the same initial data, and so are identical (assuming the uniqueness result proved in Chap. 5). Therefore the symmetry is an isometry of the metric at any positive time.

4.1.3 Parabolic Rescaling of the Ricci Flow

Aside from the geometric symmetries of diffeomorphism invariance, the Ricci flow has additional scaling properties that are essential for blow-up analysis of singularities. The time-independent diffeomorphism invariance allows for changes in the spatial coordinates. We can also translate in the time coordinates: If $g(x,t)$ satisfies Ricci flow, then so does $g(x,t-t_0)$ for any $t_0 \in \mathbb{R}$. Further, the Ricci flow has a *scale invariance*: If g is a solution of Ricci flow, and $\lambda > 0$, then g_λ is also a solution, where

$$g_\lambda(x,t) = \lambda^2 g\Big(x, \frac{t}{\lambda^2}\Big).$$

The main use of this rescaling will be to analyse singularities that develop under the Ricci flow. In such a case the curvature tends to infinity, so we perform a rescaling to produce metrics with bounded curvature and try to produce a smooth limit of these. We will discuss the machinery required for this in Chap. 9, and return to the blow-up procedure in Sect. 9.5.

4.2 The Laplacian of Curvature

We devote this section to the derivation of an expression for the Laplacian of $R_{ijk\ell}$. This will be used to derive an evolution equation for the curvature in Sect. 4.4, where the Ricci flow (4.1) implies a heat-type equation for the Riemannian curvature R. In order to do this, we need to introduce quadratic $B_{ijk\ell}$ terms which will help simplify our computation.

4.2.1 Quadratic Curvature Tensor

Various second order derivatives of the curvature tensor are likely to differ by terms quadratic in the curvature tensor. To this end we introduce the $(4,0)$-tensor $B \in \mathscr{T}_0^4(M)$ defined by

$$B(X,Y,W,Z) = \langle R(X,\cdot,Y,\star), R(W,\cdot,Z,\star)\rangle\,,$$

where the inner product $\langle,\rangle$ is given by (2.3). In components this becomes

$$B_{ijk\ell} = g^{pr}g^{qs}R_{piqj}R_{rks\ell} = R^{p\ q}_{\ i\ j}R_{pkq\ell}. \tag{4.2}$$

This tensor has some of the symmetries of the curvature tensor, namely

$$B_{ijk\ell} = B_{ji\ell k} = B_{k\ell ij}.$$

However other symmetries of the curvature tensor may fail to hold for $B_{ijk\ell}$.

4.2.2 Calculating the Connection Laplacian $\Delta R_{ijk\ell}$

Using the definition of the Laplacian explicated in Sect. 2.6, we compute the connection Laplacian acting on the $\mathscr{T}_0^4(M)$ tensor bundle.

Proposition 4.2. *On a Riemannian manifold (M,g), the Laplacian of the curvature tensor R satisfies*

$$\begin{aligned}\Delta R_{ijk\ell} &= \nabla_i\nabla_k R_{j\ell} - \nabla_j\nabla_k R_{i\ell} + \nabla_j\nabla_\ell R_{ik} - \nabla_i\nabla_\ell R_{jk}\\ &\quad - 2\big(B_{ijk\ell} - B_{ij\ell k} - B_{i\ell jk} + B_{ikj\ell}\big) + g^{pq}\big(R_{qjk\ell}R_{pi} + R_{iqk\ell}R_{pj}\big).\end{aligned}$$

Proof. Using the second Bianchi identity – together with the linearity of ∇ over the space of tensor fields – we find that

$$\begin{aligned}0 &= (\nabla_p\nabla_q R)(\partial_i,\partial_j,\partial_k,\partial_\ell)\\ &\quad + (\nabla_p\nabla_i R)(\partial_j,\partial_q,\partial_k,\partial_\ell) + (\nabla_p\nabla_j R)(\partial_q,\partial_i,\partial_k,\partial_\ell).\end{aligned}$$

Combining this identity with (2.10) implies that the Laplacian of R takes the form

$$\begin{aligned}\Delta R_{ijk\ell} &= (\mathrm{tr}_g\nabla^2 R)_{ijk\ell}\\ &= g^{pq}(\nabla_{\partial_p}\nabla_{\partial_q}R)(\partial_i,\partial_j,\partial_k,\partial_\ell)\\ &= g^{pq}\left(-(\nabla_p\nabla_i R)(\partial_j,\partial_q,\partial_k,\partial_\ell) - (\nabla_p\nabla_j R)(\partial_q,\partial_i,\partial_k,\partial_\ell)\right)\\ &= g^{pq}\big(\nabla_p\nabla_i R_{qjk\ell} - \nabla_p\nabla_j R_{qik\ell}\big).\end{aligned}$$

From this it suffices to express $g^{pq}\nabla_p\nabla_i R_{qjk\ell}$ in terms of lower order terms by commuting derivatives and contracting with the metric.

Firstly, Proposition 2.44 implies that

$$\nabla_p\nabla_i R_{qjk\ell} = \nabla_i\nabla_p R_{qjk\ell} + \big(R(\partial_i,\partial_p)R\big)(\partial_q,\partial_j,\partial_k,\partial_\ell). \tag{4.3}$$

From the second Bianchi identity, the first term on the right-hand side of (4.3) becomes

$$\nabla_i\nabla_p R_{qjk\ell} = \nabla_i\nabla_k R_{jq\ell p} - \nabla_i\nabla_\ell R_{jqkp}.$$

Contracting with the metric and invoking Proposition 2.48, (2.15), gives

$$\begin{aligned} g^{pq}\nabla_i\nabla_p R_{qjk\ell} &= g^{pq}\nabla_i\nabla_k R_{jq\ell p} - g^{pq}\nabla_i\nabla_\ell R_{jqkp} \\ &= \nabla_i\nabla_k R_{j\ell} - \nabla_i\nabla_\ell R_{jk}. \end{aligned} \tag{4.4}$$

Turning our attention to the second term on the right-hand side of (4.3), we find by Proposition 2.43 that

$$\begin{aligned} &\big(R(\partial_i,\partial_p)R\big)(\partial_q,\partial_j,\partial_k,\partial_\ell) \\ &\quad= \cancel{R(\partial_i,\partial_p)(R_{qjk\ell})} \\ &\qquad - R(R(\partial_i,\partial_p)\partial_q,\partial_j,\partial_k,\partial_\ell) - \cdots - R(\partial_q,\partial_j,\partial_k,R(\partial_i,\partial_p)\partial_\ell) \\ &\quad= -R_{ipq}{}^{n}R_{njk\ell} - R_{ipq}{}^{n}R_{qnk\ell} - R_{ipq}{}^{n}R_{qjn\ell} - R_{ipq}{}^{n}R_{qjkn} \\ &\quad= g^{mn}\big(R_{piqm}R_{njk\ell} + R_{pijm}R_{qnk\ell} + R_{pikm}R_{qjn\ell} + R_{pi\ell m}R_{qjkn}\big). \end{aligned}$$

Therefore

$$\begin{aligned} &g^{pq}\big(R(\partial_i,\partial_p)R\big)(\partial_q,\partial_j,\partial_k,\partial_\ell) \\ &\quad= g^{pq}g^{mn}\big(R_{piqm}R_{njk\ell} + R_{pijm}R_{qnk\ell} + R_{pikm}R_{qjn\ell} + R_{pi\ell m}R_{qjkn}\big). \end{aligned}$$

We now look to simplify this expression by observing that the first term contracts to

$$g^{pq}g^{mn}R_{piqm}R_{njk\ell} = g^{mn}R_{ipm}{}^{p}R_{njk\ell} = g^{mn}R_{njk\ell}R_{im} = g^{pq}R_{pjk\ell}R_{iq},$$

the second term contracts to

$$\begin{aligned} g^{pq}g^{mn}R_{pijm}R_{qnk\ell} &= g^{pq}g^{mn}R_{pijm}(-R_{nkq\ell} - R_{kqn\ell}) \\ &= g^{pq}g^{mn}(R_{pimj}R_{q\ell nk} - R_{pimj}R_{qkn\ell}) \\ &= R^{q}{}_{i}{}^{n}{}_{j}R_{q\ell nk} - R^{q}{}_{i}{}^{n}{}_{j}R_{qkn\ell} \\ &= B_{ij\ell k} - B_{ijk\ell}, \end{aligned}$$

and the third and forth terms contract to

$$\begin{aligned} g^{pq}g^{mn}(R_{pikm}R_{qjn\ell} + R_{pi\ell m}R_{qjkn}) &= g^{pq}g^{mn}(-R_{pimk}R_{qjn\ell} + R_{pim\ell}R_{qjnk}) \\ &= -R^{q\ n}_{\ i\ k}R_{qjn\ell} + R^{q\ n}_{\ i\ \ell}R_{qjnk} \\ &= -B_{ikj\ell} + B_{i\ell jk}. \end{aligned}$$

Combining these results gives

$$\begin{aligned} &g^{pq}\big(R(\partial_i, \partial_p)R\big)(\partial_q, \partial_j, \partial_k, \partial_\ell) \\ &\qquad = g^{pq}R_{pjk\ell}R_{iq} - (B_{ijk\ell} - B_{ij\ell k} - B_{i\ell jk} + B_{ikj\ell}). \end{aligned} \tag{4.5}$$

Finally, by putting (4.4) and (4.5) together we get

$$\begin{aligned} g^{pq}\nabla_p\nabla_i R_{qjk\ell} &= \nabla_i\nabla_k R_{j\ell} - \nabla_i\nabla_\ell R_{jk} \\ &\quad - \ (B_{ijk\ell} - B_{ij\ell k} - B_{i\ell jk} + B_{ikj\ell}) + g^{pq}R_{pjk\ell}R_{iq}. \end{aligned}$$

Since $\Delta R_{ijk\ell} = g^{pq}\nabla_p\nabla_i R_{qjk\ell} - (i \leftrightarrow j)$ the desired formula now follows. □

4.3 Metric Variation Formulas

In this section we establish how one can formally take the time derivative of a metric and the associated Levi–Civita connection. Thereafter we derive various variational equations for the Levi–Civita connection, curvature tensor and various traces thereof.

4.3.1 *Interpreting the Time Derivative*

Consider a one-parameter family of smooth metrics $g = g(t) \in \Gamma(\mathrm{Sym}^2\, T^*M)$ parametrised by 'time' t. We define the time derivative $\frac{\partial}{\partial t}g : \mathscr{X}(M) \times \mathscr{X}(M) \to C^\infty(M)$ of the metric g by letting

$$\Big(\frac{\partial}{\partial t}g\Big)(X, Y) := \frac{\partial}{\partial t}g(X, Y) \tag{4.6}$$

for any time *independent* vector fields $X, Y \in \mathscr{X}(M)$ (where $\frac{\partial}{\partial t}g(X, Y)$ is the time derivative of the smooth function $g(X, Y) \in C^\infty(M)$ given by the standard difference quotient). The metric in local coordinates can be expressed as $g(t) = g_{ij}(t)dx^i \otimes dx^j$, in which case (4.6) implies that

$$\frac{\partial}{\partial t}g = \dot{g}_{ij}(t)\, dx^i \otimes dx^j.$$

Therefore we regard the time derivative of the metric as the derivative of its the component functions with respect to a fixed basis.

Since the Levi–Civita connection ∇ can be written locally in terms of the metric (2.9), it too will be time dependent. So in a similar fashion, we define the time derivative of $\nabla = \nabla^{(t)}$ by letting

$$\Big(\frac{\partial}{\partial t}\nabla\Big)(X,Y) := \frac{\partial}{\partial t}\nabla_X Y \tag{4.7}$$

for time-independent vector fields X and Y. As ∇ satisfies the product rule by definition, it is not tensorial. However, we observe the following special properties of its time derivative:

Lemma 4.3. *The time derivative* $\frac{\partial}{\partial t}\nabla$ *of the Levi–Civita connection* ∇ *is tensorial.*

Proof. For $f \in C^\infty(M)$ and any time independent vector fields $X, Y \in \mathscr{X}(M)$, we see that

$$\Big(\frac{\partial}{\partial t}\nabla\Big)(X, fY) = \frac{\partial}{\partial t}\Big((Xf)Y + f\nabla_X Y\Big) = f\,\frac{\partial}{\partial t}\nabla_X Y = f\Big(\frac{\partial}{\partial t}\nabla\Big)(X,Y).$$

Thus $\frac{\partial}{\partial t}\nabla$ is a tensor by Proposition 2.19. □

Lemma 4.4. *If X is a time independent vector field and $V = V(t)$ is a time dependent vector field, then*

$$\frac{\partial}{\partial t}\nabla_X V = \Big(\frac{\partial}{\partial t}\nabla\Big)(X,V) + \nabla_X\frac{\partial V}{\partial t}.$$

Proof. Fix a time independent vector field X. As $\nabla_X = \nabla_X^{(t)} : \mathscr{X}(M) \to \mathscr{X}(M)$ we see that

$$\frac{\partial}{\partial t}\nabla_X V \lim_{\delta\to 0}\frac{\nabla^{(t+\delta)}(X, V(t+\delta)) - \nabla^{(t)}(X, V(t))}{\delta}.$$

Since

$$\begin{aligned}
\nabla^{(t+\delta)}&(X, V(t+\delta)) - \nabla^{(t)}(X, V(t)) \\
&= \nabla^{(t+\delta)}(X, V(t+\delta)) - \nabla^{(t+\delta)}(X, V(t)) \\
&\quad + \Big(\nabla^{(t+\delta)} - \nabla^{(t)}\Big)(X, V(t)) \\
&= \nabla^{(t+\delta)}(X, V(t+\delta) - V(t)) + \big(\nabla^{(t+\delta)} - \nabla^{(t)}\big)(X, V(t)),
\end{aligned}$$

we conclude that

$$\begin{aligned}\frac{\partial}{\partial t}\nabla_X V &= \lim_{\delta\to 0}\frac{1}{\delta}\nabla^{(t+\delta)}(X, V(t+\delta)-V(t))\\ &\quad+\lim_{\delta\to 0}\frac{1}{\delta}\Big(\nabla^{(t+\delta)}-\nabla^{(t)}\Big)(X, V(t))\\ &= \nabla^{(t)}_X\Big(\lim_{\delta\to 0}\frac{V(t+\delta)-V(t)}{\delta}\Big)+\Big(\frac{\partial}{\partial t}\nabla\Big)(X,V). \qquad \square\end{aligned}$$

Furthermore, we would like to differentiate the Christoffel symbols $\Gamma^k_{ij} = dx^k(\nabla_{\partial_i}\partial_j)$ of the Levi–Civita connection ∇. To do this, we consider them as a map $\Gamma : \mathscr{X}(M)\times\mathscr{X}(M)\times\mathscr{T}^1_0(M)\to C^\infty(M)$ defined by $\Gamma(X,Y,\omega) := \omega(\nabla_X Y)$. With this we proceed in a similar fashion by defining $\frac{\partial}{\partial t}\Gamma$ to be

$$\Big(\frac{\partial}{\partial t}\Gamma\Big)(X,Y,\omega) := \frac{\partial}{\partial t}\Gamma(X,Y,\omega). \tag{4.8}$$

It is easy so see that $\frac{\partial}{\partial t}\Gamma$ is tensorial since $(\frac{\partial}{\partial t}\Gamma)(X,Y,\omega) = \frac{\partial}{\partial t}\omega(\nabla_X Y) = \omega(\frac{\partial}{\partial t}\nabla_X Y)$ and $\frac{\partial}{\partial t}\nabla$ is a tensor by Lemma 4.3.

4.3.2 Variation Formulas of the Curvature

We now derive evolution equations for geometric quantities under arbitrary metric variations. Note that $\frac{\partial}{\partial t}T = (\delta_{\dot g}T)(g) = (\delta_h T)(g)$ for a (k,ℓ)-tensor T, whenever the direction $h = \frac{\partial}{\partial t}g$ (where the first variation δ_h is taken in the sense of Appendix A).

Lemma 4.5. *Suppose $g(t)$ is a smooth one-parameter family of metrics on a manifold M such that $\frac{\partial}{\partial t}g = h$. Then*

$$\frac{\partial}{\partial t}g^{ij} = -g^{ik}g^{j\ell}h_{k\ell}. \tag{4.9}$$

Proof. As $g^{ik}g_{k\ell} = \delta^i_\ell$ we find that $0 = (\frac{\partial}{\partial t}g^{ik})\,g_{k\ell} + g^{ik}h_{k\ell}$. In which case $\frac{\partial}{\partial t}g^{ij} = -g^{ik}g^{j\ell}h_{k\ell} = -h^{ij}$. $\square$

Proposition 4.6. *Suppose $g(t)$ is a smooth one-parameter family of metrics on a manifold M such that $\frac{\partial}{\partial t}g = h$. Then the Levi–Civita connection Γ^k_{ij} of g evolves by*

$$\frac{\partial}{\partial t}\Gamma^k_{ij} = \frac{1}{2}g^{k\ell}\Big((\nabla_j h)(\partial_i,\partial_\ell)+(\nabla_i h)(\partial_j,\partial_\ell)-(\nabla_\ell h)(\partial_i,\partial_j)\Big).$$

Proof. As $\frac{\partial}{\partial t}\Gamma$ and ∇h are tensorial, we are free to work in any local coordinate chart (x^i). In particular, by choosing normal coordinates about a point p (and evaluating the computation at that point), (2.9) implies that

$$\frac{\partial}{\partial t}\Gamma_{ij}^{k} = \frac{1}{2}\Big(\frac{\partial}{\partial t}g^{k\ell}\Big)\cancel{(\partial_j g_{i\ell} + \partial_i g_{j\ell} - \partial_\ell g_{ij})} + \frac{1}{2}g^{k\ell}\frac{\partial}{\partial t}\Big(\partial_j g_{i\ell} + \partial_i g_{j\ell} - \partial_\ell g_{ij}\Big).$$

Using Proposition 2.29 we observe that

$$\frac{\partial}{\partial t}\partial_j g_{i\ell} = \partial_j\big(h(\partial_i, \partial_\ell)\big) = (\nabla_j h)(\partial_i, \partial_\ell) + h(\cancel{\nabla_j \partial_i}, \partial_\ell) + h(\partial_i, \cancel{\nabla_j \partial_\ell}),$$

from which the result now follows. ☐

Proposition 4.7. *Suppose $g(t)$ is a smooth one-parameter family of metrics on a manifold M such that $\frac{\partial}{\partial t}g = h$. Then the Riemannian curvature tensor evolves by*

$$\begin{aligned}\frac{\partial}{\partial t}R_{ijk}{}^{\ell} = \frac{1}{2}g^{\ell p}\Big(&\nabla_i\nabla_p h_{jk} + \nabla_j\nabla_k h_{ip} - \nabla_i\nabla_k h_{jp} - \nabla_j\nabla_p h_{ik}\\ &- R_{ijk}{}^{q}h_{qp} - R_{ijp}{}^{q}h_{qk}\Big).\end{aligned}$$

Proof. As the expression is tensorial, we are free to work in normal coordinates in a neighbourhood about a point p (and evaluating the expression at that point). Using Lemma 2.46 in these coordinates, we find that

$$\begin{aligned}\frac{\partial}{\partial t}R_{ijk}{}^{\ell} &= \frac{\partial}{\partial t}\big(\partial_j\Gamma_{ik}^{\ell} - \partial_i\Gamma_{jk}^{\ell}\big)\\ &\quad + \Big(\Big(\frac{\partial}{\partial t}\Gamma_{ik}^{m}\Big)\cancel{\Gamma_{jm}^{\ell}} + \cancel{\Gamma_{ik}^{m}}\Big(\frac{\partial}{\partial t}\Gamma_{jm}^{\ell}\Big)\Big) - \cancel{(i \leftrightarrow j)}\\ &= \partial_j\Big(\frac{\partial}{\partial t}\Gamma_{ik}^{\ell}\Big) - \partial_i\Big(\frac{\partial}{\partial t}\Gamma_{jk}^{\ell}\Big).\end{aligned}$$

From this, together with Proposition 4.6, we have that

$$\begin{aligned}\partial_j\Big(\frac{\partial}{\partial t}\Gamma_{ik}^{\ell}\Big) &= \frac{1}{2}\cancel{\partial_j g^{\ell p}}\Big(\nabla_k h_{ip} + \nabla_i h_{kp} - \nabla_p h_{ik}\Big)\\ &\quad + \frac{1}{2}g^{\ell p}\Big(\partial_j\big(\nabla_k h_{ip} + \nabla_i h_{kp} - \nabla_p h_{ik}\big)\Big)\\ &= \frac{1}{2}g^{\ell p}\Big(\nabla_j\nabla_k h_{ip} + \nabla_j\nabla_i h_{kp} - \nabla_j\nabla_p h_{ik}\Big),\end{aligned}$$

where the last equality – using Proposition 2.29 – is due to

$$\begin{aligned}\partial_j\big((\nabla h)(\partial_i, \partial_k, \partial_\ell)\big) &= \big(\nabla_j(\nabla h)\big)(\partial_i, \partial_k, \partial_\ell)\\ &\quad + (\nabla h)(\cancel{\nabla_j\partial_i}, \ldots) + \cdots + (\nabla h)(\ldots, \cancel{\nabla_j\partial_\ell})\\ &= (\nabla_j\nabla_i h)(\partial_k, \partial_\ell).\end{aligned}$$

Therefore,

$$
\begin{aligned}
\frac{\partial}{\partial t} R_{ijk}{}^{\ell} &= \frac{1}{2} g^{\ell p}\big(\nabla_j \nabla_k h_{ip} + \nabla_j \nabla_i h_{kp} - \nabla_j \nabla_p h_{ik}\big) \\
&\quad - \frac{1}{2} g^{\ell p}\big(\nabla_i \nabla_k h_{jp} + \nabla_i \nabla_j h_{kp} - \nabla_i \nabla_p h_{jk}\big) \qquad (4.10) \\
&= \frac{1}{2} g^{\ell p}\Big(\nabla_i \nabla_p h_{jk} + \nabla_j \nabla_k h_{ip} - \nabla_i \nabla_k h_{jp} - \nabla_j \nabla_p h_{ik} \\
&\qquad + \big(R(\partial_i, \partial_j) h\big)(\partial_k, \partial_p)\Big) \qquad (4.11)
\end{aligned}
$$

since $\nabla_j \nabla_i h_{kp} - \nabla_i \nabla_j h_{kp} = (R(\partial_i, \partial_j)h)(\partial_k, \partial_p)$ by Proposition 2.44. The desired equation now follows since we find, by Proposition 2.43, that

$$
\begin{aligned}
\big(R(\partial_i, \partial_j)h\big)(\partial_k, \partial_p) &= \cancel{R(\partial_i, \partial_j) h_{kp}} - h(R(\partial_i, \partial_j)\partial_k, \partial_p) - h(\partial_k, R(\partial_i, \partial_j)\partial_p) \\
&= -R_{ijk}{}^{q} h_{qp} - R_{ijp}{}^{q} h_{qk}. \qquad \square
\end{aligned}
$$

Proposition 4.8. *Suppose $g(t)$ is a smooth family of metrics on a manifold M with $\frac{\partial}{\partial t} g = h$. Then the $(4,0)$-Riemann curvature tensor R evolves by*

$$
\begin{aligned}
\frac{\partial}{\partial t} R_{ijk\ell} &= \frac{1}{2}\big(\nabla^2_{i,\ell} h_{jk} + \nabla^2_{j,k} h_{i\ell} - \nabla^2_{i,k} h_{j\ell} - \nabla^2_{j,\ell} h_{ik}\big) \\
&\quad + \frac{1}{2} g^{pq}\big(R_{ijkp} h_{q\ell} + R_{ijp\ell} h_{qk}\big).
\end{aligned}
$$

Proof. As $R_{ijk\ell} = R_{ijk}{}^{m} g_{m\ell}$, we have

$$
\frac{\partial}{\partial t} R_{ijk\ell} = R_{ijk}{}^{q} \frac{\partial}{\partial t} g_{q\ell} + g_{a\ell} \frac{\partial}{\partial t} R_{ijk}{}^{a}.
$$

Now since

$$
R_{ijk}{}^{q} \frac{\partial}{\partial t} g_{q\ell} = g^{pq} R_{ijkp} h_{q\ell},
$$

and (by Proposition 4.7)

$$
\begin{aligned}
g_{a\ell} \frac{\partial}{\partial t} R_{ijk}{}^{a} &= \frac{1}{2}\big(\nabla^2_{i,\ell} h_{jk} + \nabla^2_{j,k} h_{i\ell} - \nabla^2_{i,k} h_{j\ell} - \nabla^2_{j,\ell} h_{ik}\big) \\
&\quad - \frac{1}{2} g^{pq}\big(R_{ijkp} h_{q\ell} + R_{ij\ell p} h_{qk}\big)
\end{aligned}
$$

the result naturally follows. $\square$

Proposition 4.9. *Suppose $g(t)$ is a smooth one-parameter family of metrics on a manifold M such that $\frac{\partial}{\partial t} g = h$. Then the Ricci tensor* Ric *evolves by*

$$
\frac{\partial}{\partial t} R_{ik} = \frac{1}{2} g^{pq}\big(\nabla^2_{q,k} h_{ip} - \nabla^2_{i,k} h_{qp} + \nabla^2_{q,i} h_{kp} - \nabla^2_{q,p} h_{ik}\big).
$$

Proof. Recall from (4.10) that

$$\begin{aligned}\frac{\partial}{\partial t}R_{ijk}{}^{\ell} = \frac{1}{2}g^{\ell p}\big(&\nabla^2_{j,k}h_{ip} + \nabla^2_{j,i}h_{kp} - \nabla^2_{j,p}h_{ik}\\ &- \nabla^2_{i,k}h_{jp} - \nabla^2_{i,j}h_{kp} + \nabla^2_{i,p}h_{jk}\big).\end{aligned}$$

From which we find that

$$\begin{aligned}\frac{\partial}{\partial t}R_{ik} = \frac{\partial}{\partial t}R_{ijk}{}^{j} &= \frac{1}{2}g^{jp}\big(\nabla^2_{j,k}h_{ip} - \nabla^2_{i,k}h_{jp} + \nabla^2_{j,i}h_{kp} - \nabla^2_{j,p}h_{ik}\big)\\ &\quad + \frac{1}{2}\cancel{g^{qp}\big(-\nabla^2_{i,q}h_{kp} + \nabla^2_{i,p}h_{qk}\big)}. \qquad \square\end{aligned}$$

Proposition 4.10. *Suppose $g(t)$ is a smooth one-parameter family of metrics on a manifold M with $\frac{\partial}{\partial t}g = h$. Then the scalar curvature* Scal *evolves by*

$$\frac{\partial}{\partial t}\mathrm{Scal} = -\Delta \mathrm{tr}_g h + \delta^2 h - \langle h, \mathrm{Ric}\rangle\,,$$

where $\delta^2 h = g^{ij}g^{pq}\nabla^2_{q,j}h_{pi}$ is the 'divergence term'.

Proof. Using (4.9) and Proposition 4.9 we find that

$$\begin{aligned}\frac{\partial}{\partial t}\mathrm{Scal} &= R_{ik}\frac{\partial}{\partial t}g^{ik} + g^{ik}\frac{\partial}{\partial t}R_{ik}\\ &= -g^{ij}g^{k\ell}h_{j\ell}R_{ik} + \frac{1}{2}g^{ik}g^{pq}\Big(\nabla^2_{q,k}h_{ip} - \nabla^2_{i,k}h_{qp} + \nabla^2_{q,i}h_{kp} - \nabla^2_{q,p}h_{ik}\Big)\\ &= -g^{ij}g^{pq}h_{jq}R_{ip} + g^{ij}g^{pq}\nabla^2_{q,j}h_{ip} - g^{ij}g^{pq}\nabla^2_{i,j}h_{pq}\\ &= -h_{jq}R^{jq} + g^{ij}g^{pq}\nabla^2_{q,j}h_{pi} - \Delta\mathrm{tr}_g h. \qquad \square\end{aligned}$$

Proposition 4.11. *Suppose $g(t)$ is a smooth one-parameter family of metrics on a manifold M with $\frac{\partial}{\partial t}g = h$. Then the volume form $d\mu(g(t))$ evolves by*

$$\frac{\partial}{\partial t}d\mu = \frac{1}{2}\mathrm{tr}\,h\,d\mu.$$

To prove this, recall that:

Definition 4.12. The *adjunct* of a square matrix A is defined as the transpose of the cofactor matrix of A, that is

$$\mathrm{adj}\,A = \begin{pmatrix} \det A_{11} & -\det A_{21} & \dots \\ -\det A_{12} & \det A_{22} & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

where A_{ij} is obtained from A by striking out the i-th row and the j-th column.

Also recall that:

Lemma 4.13. *If A is a square matrix then* $A \operatorname{adj} A = \operatorname{adj} A\, A = \det A\, I_{n\times n}$. *Moreover, if* $\det A \neq 0$ *then* $A^{-1} = \frac{1}{\det A} \operatorname{adj} A$.

So if A is a square matrix with (i, j)-th entry a_{ij}, we can expand $\det A$ along the i-th row to get: $\det A = \sum_{k=1}^{n} (-1)^{i+k} \det A_{ik}$. Therefore by Lemma 4.13, the partial derivative of $\det A$ with respect to the (i, j)-th entry are

$$\begin{aligned}\frac{\partial}{\partial a_{ij}} \det A &= (-1)^{i+j} \det A_{ij} \\ &= (\operatorname{adj} A)_{ji} = \det A\, (A^{-1})_{ji}.\end{aligned} \tag{4.12}$$

Proof (Proposition 4.11). In local coordinates (x^i) the volume form can be written as $d\mu = \sqrt{\det g}\, dx^1 \wedge \ldots \wedge x^n$. So by (4.12) and the chain rule,

$$\begin{aligned}\frac{\partial}{\partial t}\sqrt{\det g} &= \frac{1}{2}\frac{1}{\sqrt{\det g}}\frac{\partial}{\partial t}\det g \\ &= \frac{1}{2}\frac{1}{\sqrt{\det g}}\frac{\partial \det g}{\partial g_{ij}}\frac{\partial g_{ij}}{\partial t} \\ &= \frac{1}{2}\sqrt{\det g}\,(g^{-1})_{ji} h_{ij} \\ &= \frac{1}{2} g^{ij} h_{ij}\,\sqrt{\det g} = \frac{1}{2}\operatorname{tr} h\,\sqrt{\det g}.\end{aligned}$$

Thus,

$$\frac{\partial}{\partial t} d\mu = \frac{\partial\sqrt{\det g}}{\partial t} dx^1 \wedge \ldots \wedge x^n = \frac{1}{2}\operatorname{tr} h\, d\mu. \qquad \square$$

4.4 Evolution of the Curvature Under the Ricci Flow

Using the results of the previous sections, it is now a relatively easy task to derive the evolution equations of the curvature, and its various traces, under the Ricci flow.

Theorem 4.14. *Suppose $g(t)$ is a solution of the Ricci flow, the $(4, 0)$-Riemannian tensor R evolves by*

$$\begin{aligned}\frac{\partial}{\partial t} R_{ijk\ell} &= \Delta R_{ijk\ell} + 2(B_{ijk\ell} - B_{ij\ell k} - B_{i\ell jk} + B_{ikj\ell}) \\ &\quad - g^{pq}(R_{pjk\ell} R_{qi} + R_{ipk\ell} R_{qj} + R_{ijkp} R_{q\ell} + R_{ijp\ell} R_{qk}).\end{aligned}$$

Proof. By Proposition 4.8, with $\frac{\partial}{\partial t} g_{ij} = -2R_{ij}$, the time derivative of $R_{ijk\ell}$ satisfies

$$\begin{aligned}&\nabla^2_{i,\ell}R_{jk}+\nabla^2_{j,k}R_{i\ell}-\nabla^2_{i,k}R_{j\ell}-\nabla^2_{j,\ell}R_{ik}\\&\quad=-\frac{\partial}{\partial t}R_{ijk\ell}-g^{pq}\big(R_{ijkp}R_{q\ell}+R_{ijp\ell}R_{qk}\big).\end{aligned}$$

By Proposition 4.2, with indices k and ℓ switched, the Laplacian of R satisfies

$$\begin{aligned}&\nabla^2_{i,\ell}R_{jk}-\nabla^2_{j,\ell}R_{ik}+\nabla^2_{j,k}R_{i\ell}-\nabla^2_{i,k}R_{j\ell}\\&\quad=\Delta R_{ij\ell k}+2(B_{ij\ell k}-B_{ijk\ell}-B_{ikj\ell}+B_{i\ell jk})\\&\qquad-g^{pq}(R_{qj\ell k}R_{pi}+R_{iq\ell k}R_{pj}).\end{aligned}$$

Combining these equations gives

$$\begin{aligned}-\Delta R_{ijk\ell}=\Delta R_{ij\ell k}=&-\frac{\partial}{\partial t}R_{ijk\ell}-2(B_{ij\ell k}-B_{ijk\ell}-B_{ikj\ell}+B_{i\ell jk})\\&+g^{pq}(R_{qj\ell k}R_{pi}+R_{iq\ell k}R_{pj})\\&-g^{pq}(R_{ijkp}R_{q\ell}+R_{ijp\ell}R_{qk}).\end{aligned}$$ □

We can also derive, without too much effort, the evolution of the following quantities under the Ricci flow.

Corollary 4.15. *Under the Ricci flow, the connection coefficients evolve by*

$$\frac{\partial}{\partial t}\Gamma^k_{ij}=-g^{k\ell}\Big((\nabla_j\mathrm{Ric})(\partial_i,\partial_\ell)+(\nabla_i\mathrm{Ric})(\partial_j,\partial_\ell)-(\nabla_\ell\mathrm{Ric})(\partial_i,\partial_j)\Big).$$

Corollary 4.16. *Under the Ricci flow, the volume form of g evolves by*

$$\frac{\partial}{\partial t}d\mu=-\mathrm{Scal}\,d\mu.$$

Corollary 4.17. *Under the Ricci flow,*

$$\begin{aligned}\frac{\partial}{\partial t}R_{ik}&=\Delta R_{ik}+\nabla^2_{ik}\mathrm{Scal}-g^{pq}(\nabla^2_{q,i}R_{kp}+\nabla^2_{q,k}R_{ip}),\\\frac{\partial}{\partial t}\mathrm{Scal}&=2\Delta\mathrm{Scal}-2g^{ij}g^{pq}\nabla^2_{q,j}R_{pi}+2|\mathrm{Ric}|^2.\end{aligned}$$

The proof of these corollaries follow easily by substituting $\frac{\partial}{\partial t}g_{ij}=-2R_{ij}$ into Propositions 4.6, 4.11, 4.9 and 4.10 respectively.

We note that the formulas in the last corollary can be simplified as follows.

Corollary 4.18. *Under the Ricci flow,*

$$\frac{\partial}{\partial t}R_{ik}=\Delta R_{ik}+2g^{pq}g^{rs}R_{pikr}R_{qs}-2g^{pq}R_{ip}R_{qk}.$$

Proof. By (4.9) the time derivative of $R_{ik} = g^{j\ell} R_{ijk\ell}$ is $\frac{\partial}{\partial t} R_{ik} = g^{j\ell} \frac{\partial}{\partial t} R_{ijk\ell} - 2g^{jp} g^{\ell q} R_{pq} R_{ijk\ell}$. Substituting the expression for $\frac{\partial}{\partial t} R_{ijk\ell}$ in Theorem 4.14 (with $g^{j\ell} \Delta R_{ijk\ell} = \Delta R_{ik}$ from (2.15)) results in

$$\begin{aligned}\frac{\partial}{\partial t} R_{ik} &= \Delta R_{ik} + 2g^{j\ell}(B_{ijk\ell} - B_{ij\ell k} - B_{i\ell jk} + B_{ikj\ell}) \\ &\quad - g^{j\ell} g^{pq}(R_{pjk\ell} R_{qi} + R_{ipk\ell} R_{qj} + R_{ijp\ell} R_{qk} + R_{ijkp} R_{q\ell}).\end{aligned}$$

As we find that

$$\begin{aligned}&2g^{j\ell}(B_{ijk\ell} - B_{ij\ell k} - B_{i\ell jk} + B_{ikj\ell}) \\ &\quad = 2g^{j\ell} B_{ijk\ell} - 2g^{j\ell}(B_{i\ell jk} + B_{ij\ell k}) + 2g^{pr} g^{qs} R_{piqk} R_{rs} \\ &\quad = 2g^{j\ell} B_{ijk\ell} - 4g^{j\ell} B_{ij\ell k} + 2g^{pr} g^{qs} R_{piqk} R_{rs} \\ &\quad = 2g^{j\ell}(B_{ijk\ell} - 2B_{ij\ell k}) + 2g^{pr} g^{qs} R_{piqk} R_{rs}\end{aligned}$$

and

$$\begin{aligned}&g^{j\ell} g^{pq}(R_{pjk\ell} R_{qi} + R_{ipk\ell} R_{qj} + R_{ijp\ell} R_{qk} + R_{ijkp} R_{q\ell}) \\ &\quad = 2g^{pq} R_{pi} R_{qk} + g^{j\ell} g^{pq} R_{ipk\ell} R_{qj} + g^{j\ell} g^{pq} R_{ijkp} R_{q\ell} \\ &\quad = 2g^{pq} R_{pi} R_{qk} + 2g^{pr} g^{qs} R_{piqk} R_{rs},\end{aligned}$$

it follows that

$$\frac{\partial}{\partial t} R_{ik} = \Delta R_{ik} + 2g^{j\ell}(B_{ijk\ell} - 2B_{ij\ell k}) + 2g^{pr} g^{qs} R_{piqk} R_{rs} - 2g^{pq} R_{pi} R_{qk}.$$

The desired result now follows from the following claim.

Claim 4.19. For any metric g_{ij}, the tensor $B_{ijk\ell}$ satisfies the identity

$$g^{j\ell}(B_{ijk\ell} - 2B_{ij\ell k}) = 0.$$

Proof of Claim. Using the Bianchi identities,

$$\begin{aligned}g^{j\ell} B_{ijk\ell} &= g^{j\ell} g^{pr} g^{qs} R_{piqj} R_{rks\ell} \\ &= g^{j\ell} g^{pr} g^{qs} R_{pqij} R_{rsk\ell} \\ &= g^{j\ell} g^{pr} g^{qs}(R_{piqj} - R_{pjqi})(R_{rks\ell} - R_{r\ell sk}) \\ &= 2g^{j\ell}(B_{ijk\ell} - B_{ij\ell k}).\end{aligned}$$

□

Corollary 4.20. *Under the Ricci flow,*

$$\frac{\partial}{\partial t} \mathrm{Scal} = \Delta \mathrm{Scal} + 2|\mathrm{Ric}|^2.$$

Proof. From Corollary 4.17 it suffices to show that

$$2g^{ij}g^{pq}\nabla^2_{q,j}R_{pi} = g^{pq}\nabla_q\nabla_p\mathrm{Scal} = \Delta\mathrm{Scal}.$$

To do this we claim:

Claim 4.21. The identity

$$\frac{1}{2}\nabla_q\nabla_p\mathrm{Scal} = g^{ij}\nabla_q\nabla_j R_{pi}$$

holds true.

Proof of Claim. By the contracted second Bianchi identity (2.16),

$$\begin{aligned}\frac{1}{2}(\nabla\mathrm{Scal})(\partial_p) = \frac{1}{2}\nabla_p\mathrm{Scal} = g^{ij}\nabla_j R_{pi} &= (g^{-1}\otimes\nabla\mathrm{Ric})(\partial_i,\partial_j,\partial_j,\partial_p,\partial_i)\\ &= (\mathrm{tr}_{14}\,\mathrm{tr}_{23}\,g^{-1}\otimes\nabla\mathrm{Ric})(\partial_p)\\ &= (\mathrm{tr}\,g^{-1}\otimes\nabla\mathrm{Ric})(\partial_p).\end{aligned}$$

As $\nabla_q(\mathrm{tr}\,g^{-1}\otimes\nabla\mathrm{Ric}) = \mathrm{tr}\,g^{-1}\otimes\nabla_q(\nabla\mathrm{Ric})$ we find that

$$\begin{aligned}\frac{1}{2}\nabla_q\nabla_p\mathrm{Scal} &= \frac{1}{2}(\nabla_q(\nabla\mathrm{Scal}))(\partial_p)\\ &= (\mathrm{tr}\,g^{-1}\otimes\nabla_q(\nabla\mathrm{Ric}))(\partial_p)\\ &= g^{ij}\nabla_q\nabla_j R_{pi}.\end{aligned}$$ □

Remark 4.22. The evolution of the scalar curvature by Corollary 4.20 provides a simple illustration of the fact that the Ricci flow 'prefers' positive curvature. In this case the two components $\Delta\mathrm{Scal}$ and $2|\mathrm{Ric}|^2$ can be interpreted in the following way: The dissipative term $\Delta\mathrm{Scal}$ reflects the fact that a point in M with a higher average curvature than its neighbours will tend to revert to the mean. The nonlinear term $2|\mathrm{Ric}|^2$ reflects the fact that if one is in a positive curvature region (e.g. a region behaving like a sphere), then the metric will contract, thus increasing the curvature to be even more positive. Conversely, if one is in a negative curvature region (such as a region behaving like a saddle), then the metric will expand, thus weakening the negativity of curvature. In both cases the curvature is trending upwards, consistent with the non-negativity of $2|\mathrm{Ric}|^2$.

4.5 A Closer Look at the Curvature Tensor

So far we have managed to derive, in Theorem 4.14, a heat-type evolution equation for the curvature tensor under the Ricci flow. As we seek to deform the metric so it has constant sectional curvature, we need to look at the relationship between the algebraic properties of the curvature tensor and the global topological and geometry of the manifold.

In three dimensions the curvature is relatively simple algebraically, and in four dimensions still somewhat tractable, but in higher dimensions the curvature becomes very complicated and hard to study. To get a glimpse of this we will look closely at the algebraic structure, in particular that of the Weyl curvature tensor. To do this though, we first need to define the following product.

4.5.1 Kulkarni–Nomizu Product

Given two $(2,0)$-tensors we want to build a $(4,0)$-tensor that has the same symmetries as that of the algebraic Riemannian curvature tensor (that is, a $(4,0)$-tensor that satisfies symmetry properties (i)–(iv) in Sect. 2.7.4.1). To do this we need a map, called the Kulkarni–Nomizu product, of the form

$$\owedge : \mathrm{Sym}^2(M) \times \mathrm{Sym}^2(M) \to \mathrm{Curv}(M),$$

where $\mathrm{Sym}^2(M)$ is the bundle of symmetric $(2,0)$-tensors and $\mathrm{Curv}(M)$ is the bundle of curvature tensors. Thus given $\alpha, \beta \in \mathrm{Sym}^2(M)$ we require the product $\alpha \owedge \beta$ to satisfy the following symmetries:

(a) Antisymmetric in the first two arguments: $(\alpha \owedge \beta)_{ijk\ell} = -(\alpha \odot \beta)_{jik\ell}$
(b) Antisymmetric in the last two arguments: $(\alpha \owedge \beta)_{ijk\ell} = -(\alpha \owedge \beta)_{ij\ell k}$
(c) Symmetric paired arguments: $(\alpha \owedge \beta)_{ijk\ell} = (\alpha \owedge \beta)_{k\ell ij}$
(d) Satisfy the Bianchi identity: $(\alpha \owedge \beta)_{ijk\ell} + (\alpha \owedge \beta)_{jki\ell} + (\alpha \owedge \beta)_{kij\ell} = 0$

A natural way to build $\alpha \owedge \beta$ is to use the tensor product. By linearity we expect $\alpha \owedge \beta$ to be a sum of $\alpha \otimes \beta$ with components permuted in such a way that the symmetries matches that of the algebraic Riemann curvature tensor. Hence all that is needed is to find the correct permutations.

Now as α and β are symmetric, terms of the form $\alpha_{ij}\beta_{k\ell}$ are disallowed since this would contradict (a) and (b). So we need to mix i, j with k, ℓ across the tensors α and β. If we naïvely suppose one of the terms is $\alpha_{ik}\beta_{j\ell}$, then (c) implies we will also need $\alpha_{j\ell}\beta_{ik}$ and by (a), (b) we will also need $-\alpha_{jk}\beta_{i\ell}$ and $-\alpha_{i\ell}\beta_{jk}$. Therefore by defining

$$(\alpha \owedge \beta)_{ijk\ell} := \alpha_{ik}\beta_{j\ell} + \alpha_{j\ell}\beta_{ik} - \alpha_{jk}\beta_{i\ell} - \alpha_{i\ell}\beta_{jk} \tag{4.13}$$

or alternatively

$$\begin{aligned}(\alpha \owedge \beta)(v_1, v_2, v_3, v_4) &= (\alpha \otimes \beta)(v_1, v_3, v_2, v_4) + (\alpha \otimes \beta)(v_2, v_4, v_1, v_3) \\ &\quad - (\alpha \otimes \beta)(v_1, v_4, v_2, v_3) - (\alpha \otimes \beta)(v_2, v_3, v_1, v_4),\end{aligned}$$

properties (a)–(c) are immediately satisfied; inspection also shows that the Bianchi identity holds as well. Thus we have constructed a $(4,0)$-tensor that has all of the symmetries of that of the algebraic curvature tensor.

4.5.2 Weyl Curvature Tensor

The Weyl curvature tensor is defined to be the *traceless component* of the Riemann curvature tensor. It can be obtained from the full curvature tensor R by subtracting out various traces. To find an exact expression, we seek to subtract a tensor C which is the sum of the scalar and traceless Ricci[2] parts of R, with the additional conditions that C must have the same traces and algebraic structures as that of R. By using the Kulkarni–Nomizu product, we consider the tensor C as taking the form

$$C = c_1 \mathrm{Scal}\, g \owedge g + c_2 \overset{\circ}{\mathrm{Ric}} \owedge g,$$

where c_1 and c_2 are scalars. As $\operatorname{tr} R = \operatorname{tr} C$, where the trace is taken over *any* pair of indices, we must have

$$\mathrm{Ric}_{ik} = g^{j\ell} R_{ijk\ell} = g^{j\ell} C_{ijk\ell}.$$

Now since

$$\begin{aligned} C_{ijk\ell} &= 2c_1 \mathrm{Scal}(g_{ik} g_{j\ell} - g_{i\ell} g_{jk}) \\ &\quad + c_2(\overset{\circ}{\mathrm{Ric}}_{ik} g_{j\ell} + \overset{\circ}{\mathrm{Ric}}_{j\ell} g_{ik} - \overset{\circ}{\mathrm{Ric}}_{i\ell} g_{jk} - \overset{\circ}{\mathrm{Ric}}_{jk} g_{i\ell}), \end{aligned}$$

it follows that

$$\begin{aligned} \frac{1}{n} \mathrm{Scal}\, g_{ik} + \overset{\circ}{\mathrm{Ric}}_{ik} = \mathrm{Ric}_{ik} &= g^{j\ell} C_{ijk\ell} \\ &= 2c_1(n-1) \mathrm{Scal}\, g_{ik} + c_2(n-2)\, \overset{\circ}{\mathrm{Ric}}_{ik}. \end{aligned}$$

In which case the scalars $c_1 = \frac{1}{2n(n-1)}$ and $c_2 = \frac{1}{n-2}$.

Therefore the Weyl tensor W must be of the form

$$\mathrm{Weyl} = R - \frac{1}{n-2} \overset{\circ}{\mathrm{Ric}} \owedge g - \frac{\mathrm{Scal}}{2n(n-1)} g \owedge g. \tag{4.14}$$

---

[2] Where the traceless Ricci tensor $\overset{\circ}{\mathrm{Ric}} := \mathrm{Ric} - \frac{\mathrm{Scal}}{n} g$, since $g^{ik} \overset{\circ}{\mathrm{Ric}}_{ik} = \mathrm{Scal} - \frac{\mathrm{Scal}}{n} g^{ik} \mathrm{Ric}_{ik} = \mathrm{Scal} - \frac{\mathrm{Scal}}{n} \delta^i_i = 0$.

Alternatively, by using $\overset{\circ}{\mathrm{Ric}} = \mathrm{Ric} - \frac{\mathrm{Scal}}{n} g$, we could also write

$$\mathrm{Weyl} = R - \frac{1}{n-2}\mathrm{Ric} \owedge g + \frac{\mathrm{Scal}}{2(n-1)(n-2)} g \owedge g. \tag{4.15}$$

Since W is defined using the Kulkarni–Nomizu product it has the same symmetries as the curvature tensor, but by definition all of its traces vanish. Indeed, one can check by hand that ${W_{ijk}}^{j} = 0$, or indeed any other trace, as per definition. Also note that

$$\left|R\right|^2 = \left|\frac{\mathrm{Scal}}{2n(n-1)} g \owedge g\right|^2 + \left|\frac{1}{n-2}\overset{\circ}{\mathrm{Ric}} \owedge g\right|^2 + \left|\mathrm{Weyl}\right|^2$$

as the decomposition is orthogonal and that the metric g has constant sectional curvature if and only if $\overset{\circ}{\mathrm{Ric}} = 0$ and $\mathrm{Weyl} = 0$ (cf. Sect. 12.3.1).

4.5.3 Sphere Theorem of Huisken–Margerin–Nishikawa

In three dimensions the Weyl tensor vanishes, and the curvature can be understood solely in terms of the Ricci tensor (see Sect. 7.5.3 for one way of doing this). One of the first insights into the differentiable pinching problem, for dimensions $n \geq 4$, was made independently by Huisken [Hui85], Nishikawa [Nis86] and Margerin [Mar86]. By using the Ricci flow, they were able show that a manifold is diffeomorphic to a spherical space form, provided the norm of the Weyl curvature tensor and the norm of the traceless Ricci tensor are not too large compared to the scalar curvature at each point.

Theorem 4.23. *Let $n \geq 4$. If the curvature tensor of a smooth compact n-dimensional Riemannian manifold M of positive scalar curvature satisfies*

$$\left|\mathrm{Weyl}\right|^2 + \left|\frac{1}{n-2}\overset{\circ}{\mathrm{Ric}} \owedge g\right|^2 \leq \delta_n \frac{2}{n(n-1)}\mathrm{Scal}^2 \tag{4.16}$$

with $\delta_4 = \frac{1}{5}$, $\delta_5 = \frac{1}{10}$, and $\delta_n = \frac{2}{(n-2)(n+1)}$ for $n \geq 6$, then the Ricci flow has a solution $g(t)$ on a maximal finite time interval $[0, T)$, and $\frac{g(t)}{2(n-1)(T-t)}$ converges in C^∞ to a metric of constant curvature 1 as $t \to \infty$. In particular, M is diffeomorphic to a spherical space form.

Remark 4.24. The Riemann curvature tensor can be considered as an element of the vector space of symmetric bilinear forms acting on the space $\bigwedge^2 TM$ (see Sect. 12.2). The inequality (4.16) defines a cone in this vector space around the line of constant curvature tensors, in which the Riemann tensor must lie. In Chaps. 12 and 13 we will return to this idea of constructing cones in the space of curvature tensors.

The technique of the proof follows Hamilton's original paper [Ham82b]. The approach is to show that if the scalar-free part of the curvature is small compared to the scalar curvature initially, then it must remain so for all time. In which case we need to show, under the hypothesis (4.16) with an appropriate δ_n, there exists $\varepsilon > 0$ such that

$$\left|\mathrm{Weyl}\right|^2 + \left|\frac{1}{n-2}\mathring{\mathrm{Ric}} \owedge g\right|^2 \leq \delta_n \, (1-\varepsilon)^2 \, \frac{2}{n(n-1)}\mathrm{Scal}^2 \tag{4.17}$$

remains valid as long as the solution to the Ricci flow exists for times $t \in [0, T)$. This is achieved by working with

$$\mathring{R} := R - \frac{2\,\mathrm{Scal}}{n(n-1)} g \owedge g,$$

which measures the failure of g to have (pointwise) constant positive sectional curvature (cf. Sect. 12.3.1). It is proved using the maximum principle (which we discuss in Chap. 7) that a bound can be obtained on a function of the form

$$F_\sigma = \frac{|\mathring{R}|^2}{\mathrm{Scal}^{2-\sigma}},$$

for $\sigma \geq 0$, yielding the following result.[3]

Proposition 4.25 ([Hui85, Sect. 3])**.** *If the inequality (4.17), that is*

$$|\mathring{R}|^2 \leq \delta_n \, (1-\varepsilon)^2 \, \frac{2}{n(n-1)}\mathrm{Scal}^2,$$

holds at time $t = 0$, then it remains so on $0 \leq t < T$. Moreover, there are constants $C < \infty$ and $\sigma > 0$, depending on n and the initial metric, such that

$$|\mathring{R}|^2 \leq C\,\mathrm{Scal}^{2-\sigma}$$

holds for $0 \leq t < T$.

Remark 4.26. From this analysis we see that the main obstacle to obtaining similar sphere-type theorem, under weaker pinching conditions than that of Theorem 4.23, is the Weyl curvature tensor W. Controlling the behaviour of W under the Ricci flow, for dimensions $n \geq 4$, has proved to be a major technical hurdle, which was finally overcome by the efforts of Böhm and Wilking [BW08]. We discuss their method in Chap. 12.

---

[3] The quantity $|\mathring{R}|^2$ is a higher dimensional analogue of the quantity $|\mathring{\mathrm{Ric}}|^2 = |\mathrm{Ric}|^2 - \frac{1}{3}\mathrm{Scal}^2$ used by Hamilton [Ham82b] in the $n = 3$ case, which measures how far the eigenvalues of the Ricci tensor diverge from each other.

' ! % $ $ 5 !
+,- * *. * ! 7 + 6 6: "
M / /" /(" "
6 *+,-* * *. * ! 8 " D 1
6: M /" (1

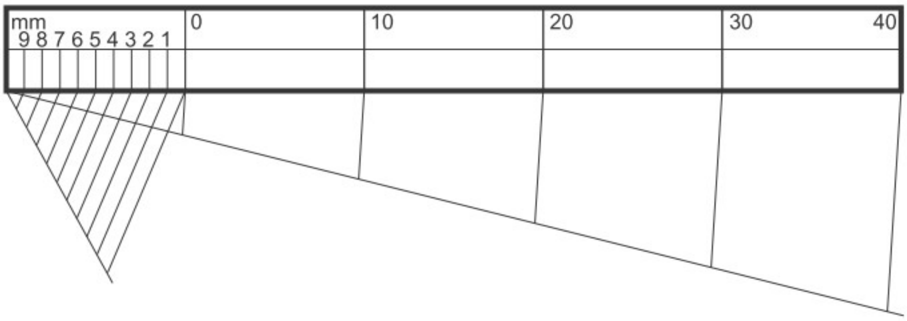

Figure 1.2 Plain scale 30 mm = 10 mm (RF = $\frac{3}{1}$).

Diagonal Scales

There is a limit to the number of divisions that can be constructed on a plain scale. Try to divide 10 mm into 50 parts; you will find that it is almost impossible. The architect, cartographer and surveyor all have the problem of having to subdivide into smaller units than a plain scale allows. A diagonal scale allows you to divide into smaller units.

Before looking at any particular diagonal scale, let us first look at the underlying principle.

Figure 1.3 shows a triangle ABC. Suppose that AB is 10 mm long and BC is divided into 10 equal parts. Lines from these equal parts have been drawn parallel to AB and numbered from 1 to 10.

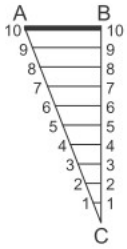

Figure 1.3 Showing how to divide a line AB into 10 equal parts.

It should be obvious that the line 5–5 is half the length of AB. Similarly, the line 1–1 is $\frac{1}{10}$ the length of AB and line 7–7 is $\frac{7}{10}$ the length of AB. (If you wish to prove this mathematically use similar triangles.)

You can see that the lengths of the lines 1–1 to 10–10 increase by 1 mm each time you go up a line. If the length of AB had been 1 mm to begin with, the increases would have been $\frac{1}{10}$ mm each time. In this way small lengths can be divided into very much smaller lengths, and can be easily picked out.

Three examples of diagonal scales follow.

This scale would be used where the drawing is twice the size of the natural object and the draughtsperson has to be able to measure on a scale accurate to 0.1 mm.

The longest natural dimension is 60 mm. This length is first divided into six 10 mm intervals. The first 10 mm is then divided into 10 parts, each 1 mm wide (scaled). Each of these 1 mm intervals is divided with a diagonal into 10 more equal parts (Fig. 1.4).

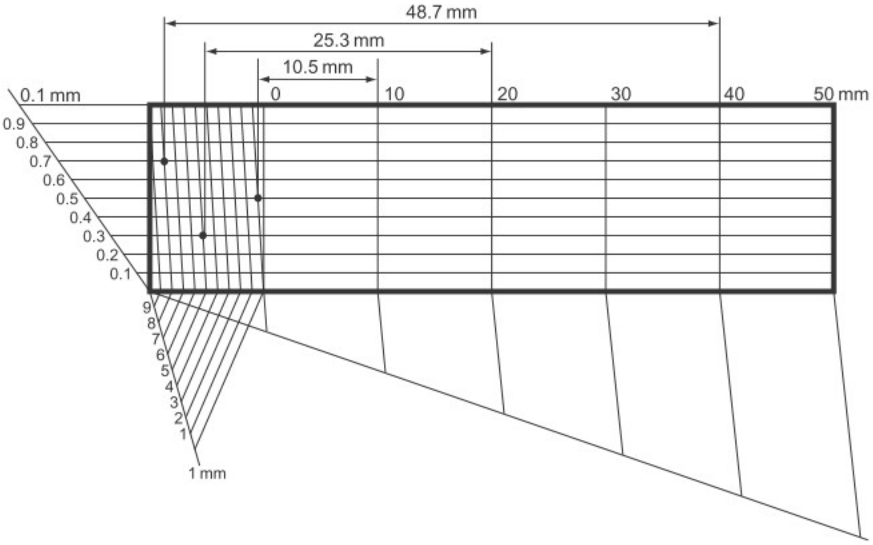

Figure 1.4 Diagonal scale 20 mm = 10 mm to read to 0.1 mm (RF = $\frac{2}{1}$).

To construct a diagonal scale, 30 mm = 1 mm, 4 m long to read to 10 mm (Fig. 1.5).

Length of scale = 4 × 30 mm = 120 mm
First division 4 × 30 mm lengths
Second division 10 × 100 mm lengths
Diagonal division 10 × 10 mm lengths

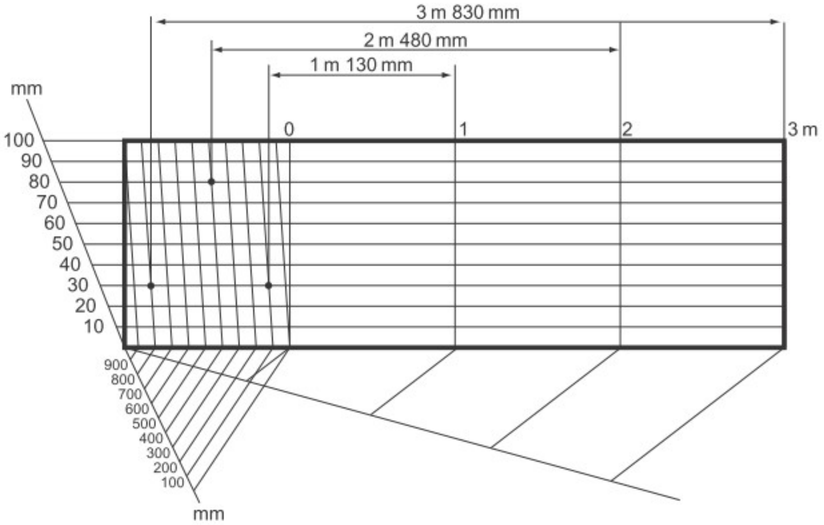
3 m 830 mm
2 m 480 mm
1 m 130 mm
mm
0
1
2
3 m
100
90
80
70
60
50
40
30
20
10
900
800
700
600
500
400
300
200
100
mm

Figure 1.5 A diagonal scale, 3 cm = 1 m, to read m and cm.

To construct a diagonal scale, 50 mm = 1 mm, 3 mm long to read 0.01 mm (Fig. 1.6).

Length of scale = 3 × 50 mm = 150 mm
First division 3 × 50 mm lengths
Second division 10 × 0.1 mm lengths
Diagonal division 10 × 0.01 mm lengths

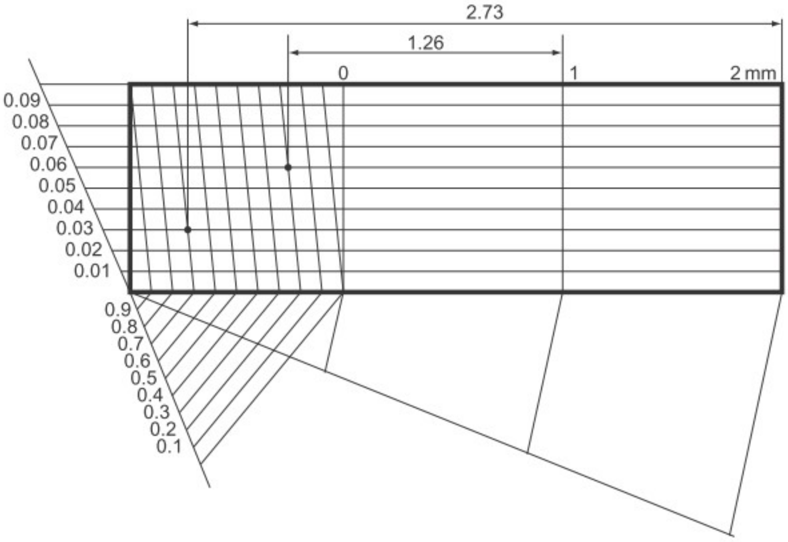
2.73
1.26
0
1
2 mm
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

Figure 1.6 A diagonal scale, 50 mm = 1 mm to read to 0.01 mm (RF = $\frac{1}{50}$).

Proportional Scales

It is possible to construct one plain scale directly from another, so that the new scale is proportional to the original one. An example of this is given in Fig. 1.7. The new scale is a copy of the original one but is $\frac{7}{4}$ times larger. The proportions of the scales can be varied by changing the ratios of the lines AB to BC.

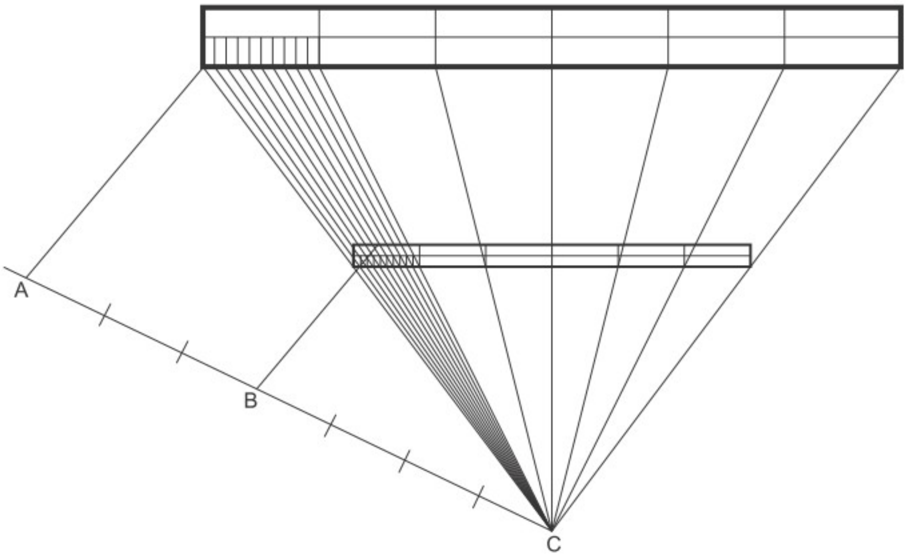

Figure 1.7 To enlarge a scale by a proportion of 7:4.

Exercise 1

1. **(a)** Draw the simple key shown in Fig. 1 full size.

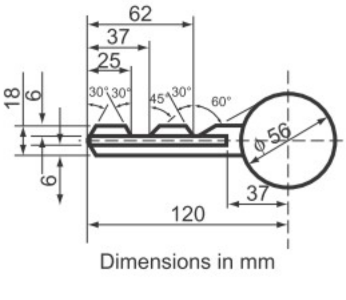
62
37
25
18
6
30°
30°
45°
30°
60°
ϕ 56
6
37
120
Dimensions in mm

Figure 1

(b) Construct a plain scale with an RF of $\frac{5}{4}$ suitable for use in the making of an enlarged drawing of this key. Do *not* draw this key again.

Southern Regional Examinations Board (question originally set in imperial units).

2. Construct a plain scale of 50 mm = 300 mm to read to 10 mm up to 1200 mm. Using this scale, draw to scale a triangle having a perimeter of 1200 mm and having sides in the ratio 3:4:6. Print neatly along each side the length to the nearest 10 mm.

Oxford Local Examinations (question originally set in imperial units; see Chapter 2 for information not in Chapter 1).

3. Construct the plain figure shown in Fig. 2 and then, by means of a proportional scale, draw a similar figure standing on the base AG. All angles must be constructed geometrically in the first figure. Measure and state the length of the side corresponding with CD.

Oxford Local Examinations (question originally set in imperial units; see Chapter 2 for information not in Chapter 1).

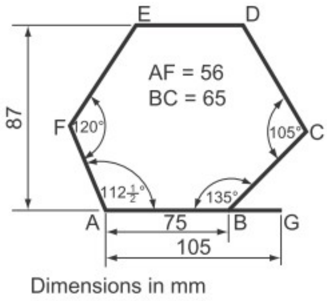

Figure 2

4. Construct a diagonal scale in which 40 mm represents 1 m. The scale is to read down to 10 mm and is to cover a range of 5 m. Mark off a distance of 4 m 780 mm.

5. Construct a diagonal scale of 25 mm to represent 1 m which can be used to measure m and 10 mm up to 8 m. Using this scale construct a quadrilateral ABCD which stands on a base AB of length 4 m 720 mm and having BC = 3 m 530 mm, AD = 4 m 170 mm, $\angle ABC = 120°$ and $\angle ADC = 90°$. Measure and state the lengths of the two diagonals and the perpendicular height, all correct to the nearest 10 mm. Angles must be constructed geometrically.

Oxford Local Examinations (question originally set in imperial units; see Chapter 2 for information not in Chapter 1).

6. Construct a diagonal scale, 10 times full size, to show mm and tenths of a mm and to read to a maximum of 20 mm. Using the scale, construct a triangle ABC with AB 17.4 mm, BC 13.8 mm and AC 11 mm.

Oxford and Cambridge Schools Examinations Board.

2 The Construction of Geometric Figures from Given Data

This chapter is concerned with the construction of plane geometric figures. Plane geometry is the geometry of figures that are two-dimensional, i.e. figures that have only length and breadth. Solid geometry is the geometry of three-dimensional figures.

There are an endless number of plane figures but we will concern ourselves only with the more common ones – the triangle, the quadrilateral and the better known polygons.

Before we look at any particular figure, there are a few constructions that must be revised (Figs. 2.1–2.13).

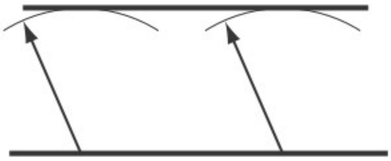

Figure 2.1 To construct a parallel line.

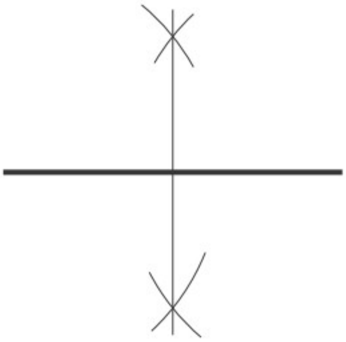

Figure 2.2 To bisect a line.

Geometric and Engineering Drawing. DOI: 10.1016/B978-0-08-096768-4.00002-4

© 2010 Elsevier Ltd. All rights reserved.

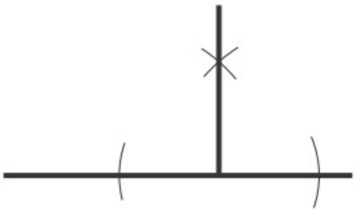

Figure 2.3 To erect a perpendicular from a point on a line.

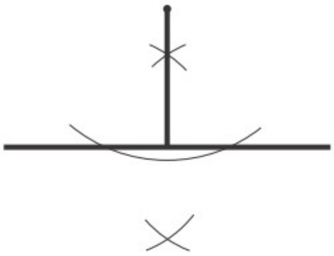

Figure 2.4 To erect a perpendicular from a point to a line.

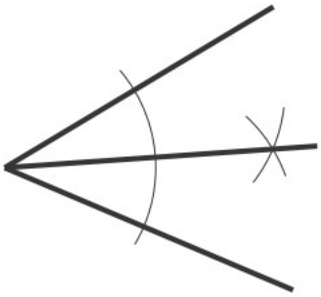

Figure 2.5 To bisect an angle.

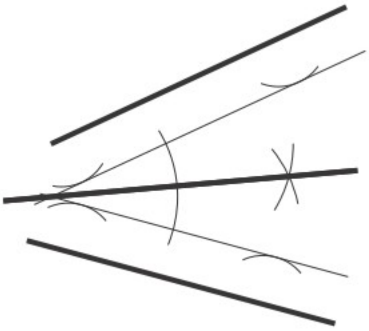

Figure 2.6 To bisect the angle formed by two converging lines.

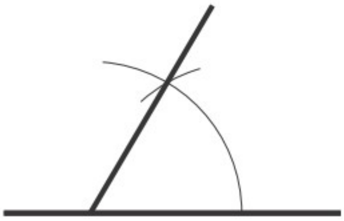

Figure 2.7 To construct 60°.

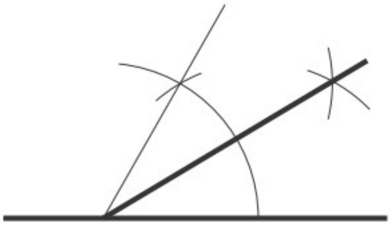

Figure 2.8 To construct 30°.

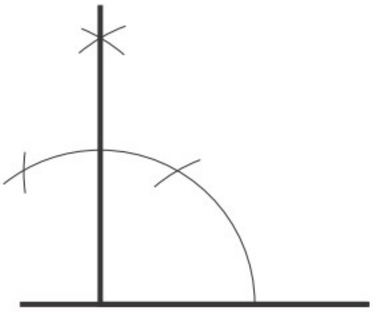

Figure 2.9 To construct 90°.

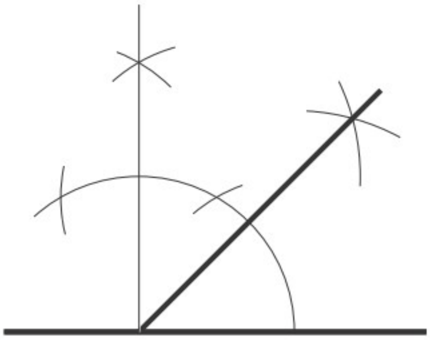

Figure 2.10 To construct 45°.

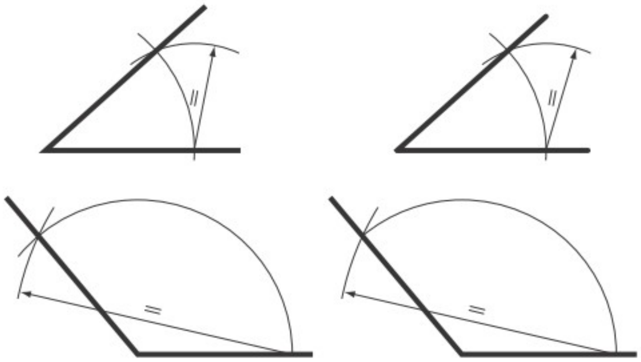

Figure 2.11 To construct an angle similar to another angle.

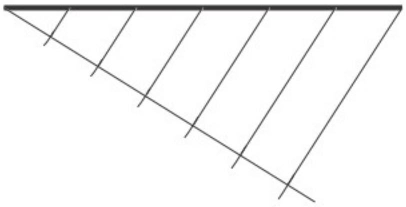

Figure 2.12 To divide a line into a number of equal parts (e.g. six).

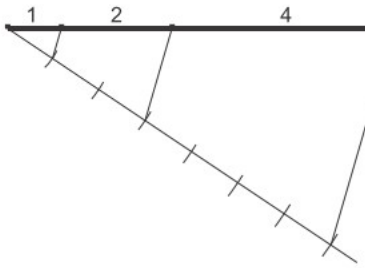

Figure 2.13 To divide a line proportionally (e.g. 1:2:4).

The Triangle

Definitions

The triangle is a plane figure bounded by three straight sides.

A *scalene* triangle is a triangle with three unequal sides and three unequal angles.
An *isosceles* triangle is a triangle with two sides, and hence two angles, equal.
An *equilateral* triangle is a triangle with all the sides, and hence all the angles, equal.
A *right-angled* triangle is a triangle containing one right angle. The side opposite the right angle is called the 'hypotenuse'.

Constructions

To construct an equilateral triangle, given one of the sides (Fig. 2.14).

1. Draw a line AB, equal to the length of the side.
2. With compass point on A and radius AB, draw an arc as shown.
3. With compass point on B, and with the same radius, draw another arc to cut the first arc at C.

Triangle ABC is equilateral.

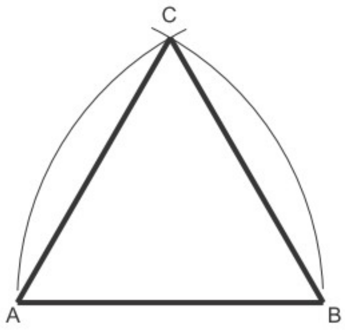

Figure 2.14

To construct an isosceles triangle given the perimeter and the altitude (Fig. 2.15).

1. Draw line AB equal to half the perimeter.
2. From B erect a perpendicular and make BC equal to the altitude.
3. Join AC and bisect it to cut AB in D.
4. Produce DB so that BE = BD. CDE is the required triangle.

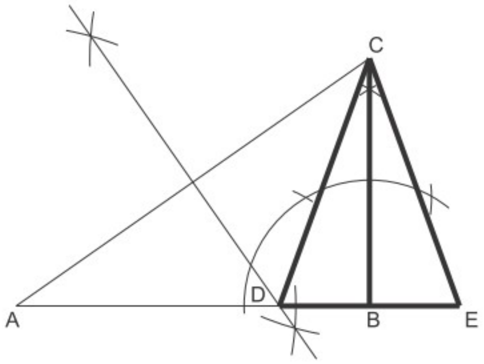

Figure 2.15

To construct a triangle, given the base angles and the altitude (Fig. 2.16).

1. Draw a line AB.
2. Construct CD parallel to AB so that the distance between them is equal to the altitude.
3. From any point E, on CD, draw CÊF and DÊG so that they cut AB in F and G, respectively.

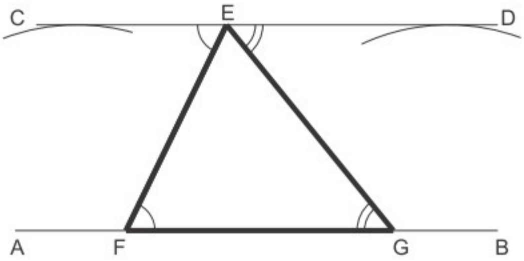

Figure 2.16

Since CÊF = EF̂G and DÊG = EĜF (alternate angles), then EFG is the required triangle.

To construct a triangle given the base, the altitude and the vertical angle (Fig. 2.17).

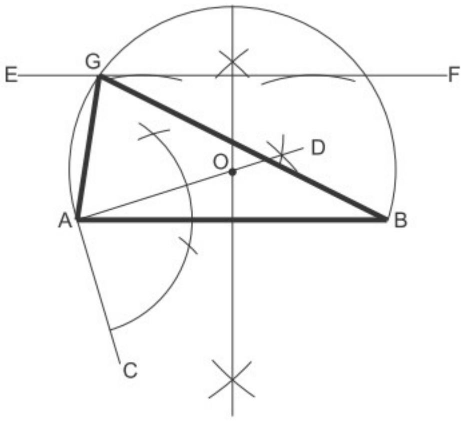

Figure 2.17

1. Draw the base AB.
2. Construct BÂC equal to the vertical angle.
3. Erect AD perpendicular to AC.
4. Bisect AB to meet AD in O.
5. With centre O and radius OA (= OB), draw a circle.
6. Construct EF parallel to AB so that the distance between them is equal to the altitude.

Let EF intersect the circle in G.

ABG is the required triangle.

To construct a triangle given the perimeter and the ratio of the sides (Fig. 2.18).

1. Draw the line AB equal in length to the perimeter.
2. Divide AB into the required ratio (e.g. 4:3:6).
3. With centre C and radius CA, draw an arc.
4. With centre D and radius DB, draw an arc to intersect the first arc in E.

ECD is the required triangle.

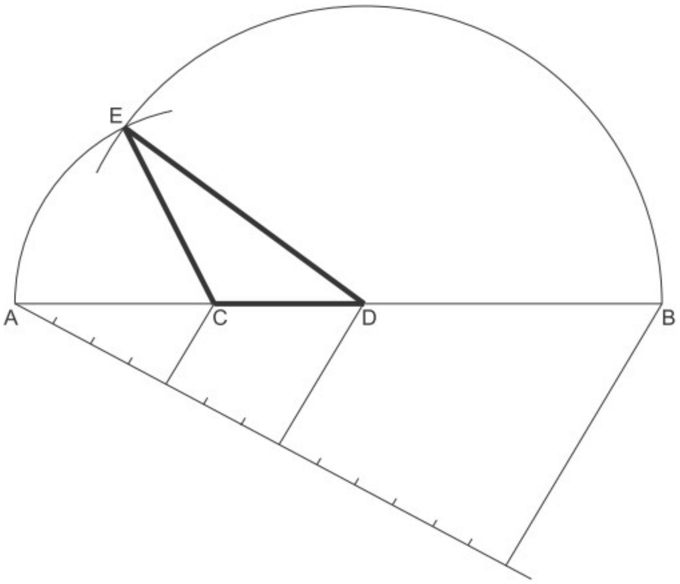

Figure 2.18

To construct a triangle given the perimeter, the altitude and the vertical angle (Fig. 2.19).

1. Draw AB and AC each equal to half the perimeter, so that CAB is the vertical angle.
2. From B and C erect perpendiculars to meet in D.
3. With centre D, draw a circle, radius DB (= DC).
4. With centre A and radius equal to the altitude, draw an arc.
5. Construct the common tangent between the circle and the arc. Let this tangent intersect AC in F and AB in E. (For tangent construction see Chapter 4.)

FEA is the required triangle.

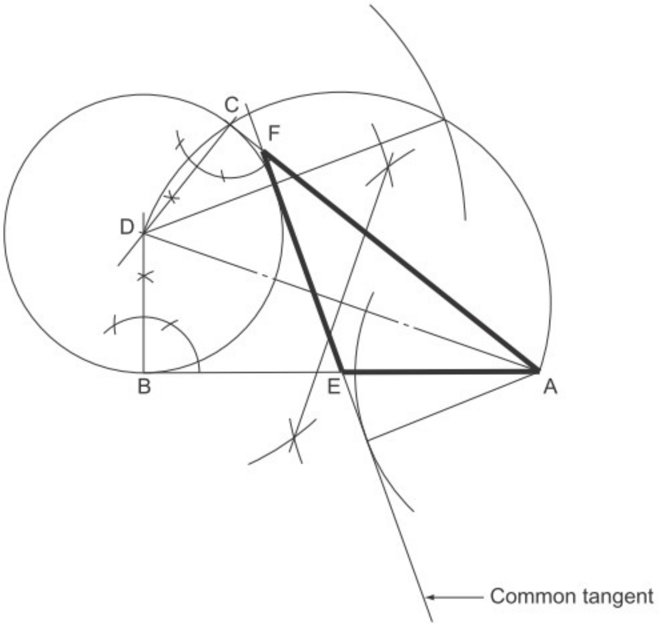
C
F
D
B
E
A
Common tangent

Figure 2.19

To construct a triangle similar to another triangle but with a different perimeter (Fig. 2.20).

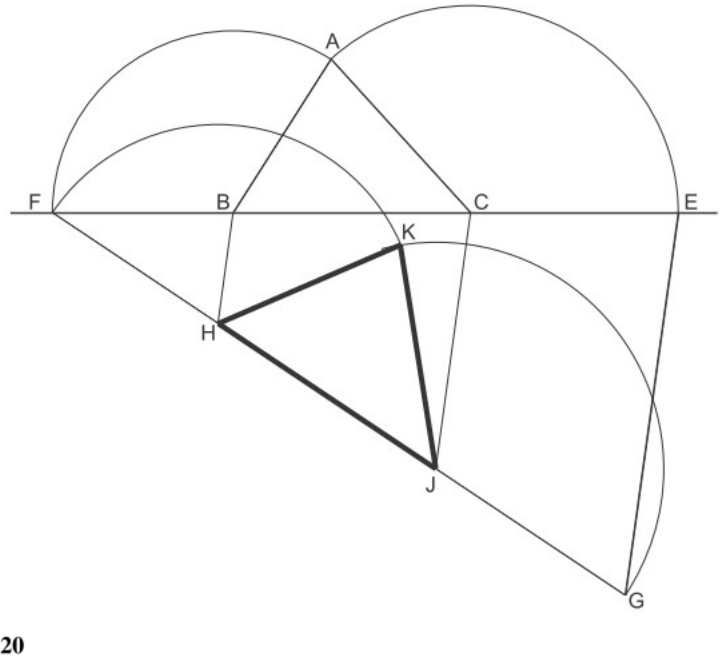

Figure 2.20

1. Draw the given triangle ABC.
2. Produce BC in both directions.
3. With compass point on B and radius BA, draw an arc to cut CB produced in F.
4. With compass point on C and radius CA, draw an arc to cut BC produced in E.
5. Draw a line FG equal in length to the required perimeter.
6. Join EG and draw CJ and BH parallel to it.
7. With centre H and radius HF, draw an arc.
8. With centre J and radius JG, draw another arc to intersect the first arc in K.

HKJ is the required triangle.

The Quadrilateral

Definitions

The quadrilateral is a plane figure bounded by four straight sides.

A *square* is a quadrilateral with all four sides of equal length and one of its angles (and hence the other three) a right angle.
A *rectangle* is a quadrilateral with its opposite sides of equal length and one of its angles (and hence the other three) a right angle.
A *parallelogram* is a quadrilateral with opposite sides equal and therefore parallel.
A *rhombus* is a quadrilateral with all four sides equal.
A *trapezium* is a quadrilateral with one pair of opposite sides parallel.
A *trapezoid* is a quadrilateral with all four sides and angles unequal.

Constructions

To construct a square given the length of the side (Fig. 2.21).

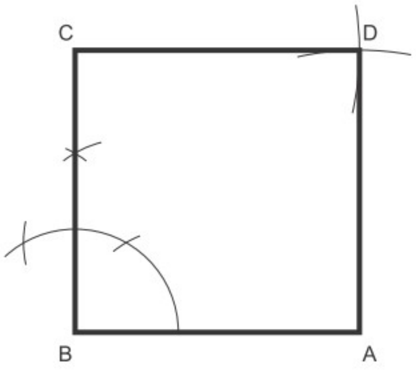

Figure 2.21

1. Draw the side AB.
2. From B erect a perpendicular.
3. Mark off the length of side BC.
4. With centres A and C draw arcs, radius equal to the length of the side of the square, to intersect at D.

ABCD is the required square.

To construct a square given the diagonal (Fig. 2.22).

1. Draw the diagonal AC.
2. Bisect AC.
3. With centre O and radius OA (= OC), draw a circle to cut the bisecting line in B and D.

ABCD is the required square.

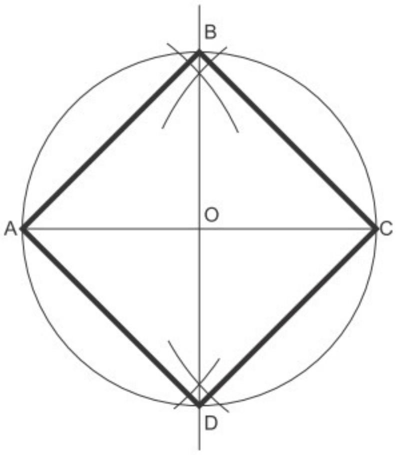

Figure 2.22

To construct a rectangle given the length of the diagonal and one of the sides (Fig. 2.23).

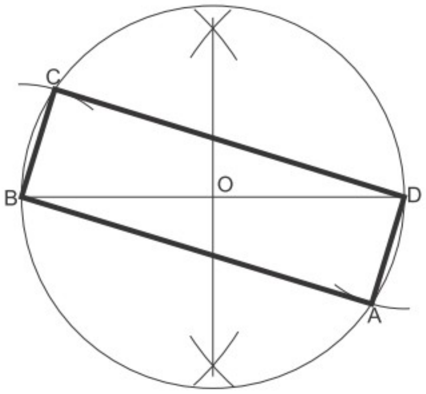

Figure 2.23

1. Draw the diagonal BD.
2. Bisect BD.
3. With centre O and radius OB (= OD), draw a circle.
4. With centre B and radius equal to the length of the known side, draw an arc to cut the circle in C.
5. Repeat step 4 with centre D to cut at A.

ABCD is the required rectangle.

To construct a parallelogram given two sides and an angle (Fig. 2.24).

1. Draw AD equal to the length of one of the sides.
2. From A construct the known angle.
3. Mark off AB equal in length to the other known side.
4. With compass point at B, draw an arc equal in radius to AD.
5. With compass point at D, draw an arc equal in radius to AB.

ABCD is the required parallelogram.

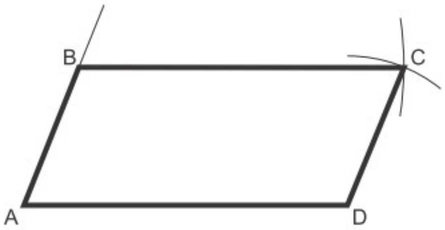

Figure 2.24

To construct a rhombus given the diagonal and the length of the sides (Fig. 2.25).

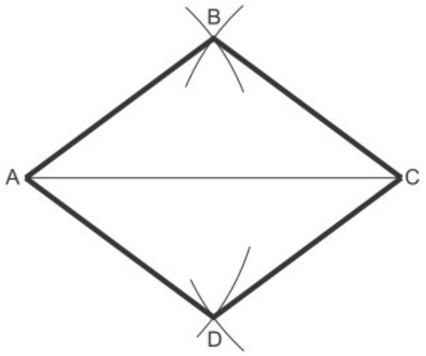

Figure 2.25

1. Draw the diagonal AC.
2. From A and C draw intersecting arcs, equal in length to the sides, to meet at B and D.

ABCD is the required rhombus.

To construct a trapezium given the lengths of the parallel sides, the perpendicular distance between them and one angle (Fig. 2.26).

1. Draw one of the parallels AB.
2. Construct the parallel line.
3. Construct the known angle from B to intersect the parallel line in C.
4. Mark off the known length CD.

ABCD is the required trapezium.

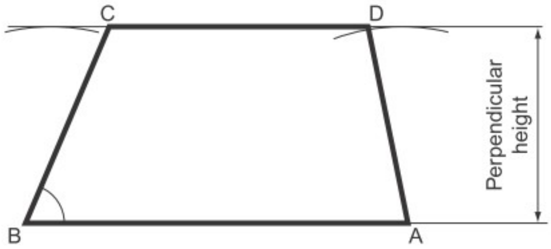

Figure 2.26

Polygons

Definitions

A polygon is a plane figure bounded by more than four straight sides. Polygons that are frequently referred to have particular names. Some of these are listed below.

A *pentagon* is a plane figure bounded by five sides.
A *hexagon* is a plane figure bounded by six sides.
A *heptagon* is a plane figure bounded by seven sides.
An *octagon* is a plane figure bounded by eight sides.
A *nonagon* is a plane figure bounded by nine sides.
A *decagon* is a plane figure bounded by ten sides.

A regular polygon is one that has all its sides equal and therefore all its exterior angles equal and all its interior angles equal.

It is possible to construct a circle within a regular polygon so that all the sides of the polygon are tangential to that circle. The diameter of that circle is called the 'diameter of the polygon'. If the polygon has an even number of sides, the diameter is the distance between two diametrically opposed faces. This dimension is often called the 'across-flats' dimension.

The diagonal of a polygon is the distance from one corner to the corner furthest away from it. If the polygon has an even number of sides, then this distance is the dimension between two diametrically opposed corners.

Constructions

To construct a regular hexagon given the length of the sides (Fig. 2.27).

1. Draw a circle, radius equal to the length of the side.
2. From any point on the circumference, step the radius around the circle six times. If your construction is accurate, you will finish at exactly the same place that you started.
3. Connect the six points to form a regular hexagon.

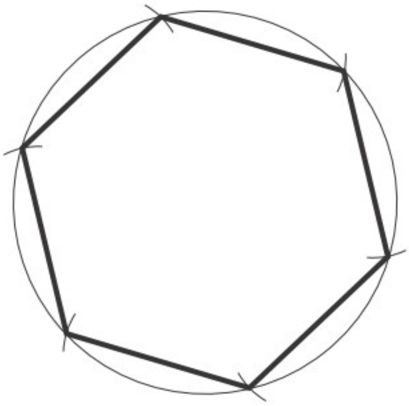

Figure 2.27

To construct a regular hexagon given the diameter (Fig. 2.28).

This construction, using compasses and straight edge only, is quite feasible but is relatively unimportant. What is important is to recognise that a hexagon can be constructed, given the diameter or across-flats dimension, by drawing tangents to the circle with a 60° set square. This is very important when drawing hexagonal-headed nuts and bolts.

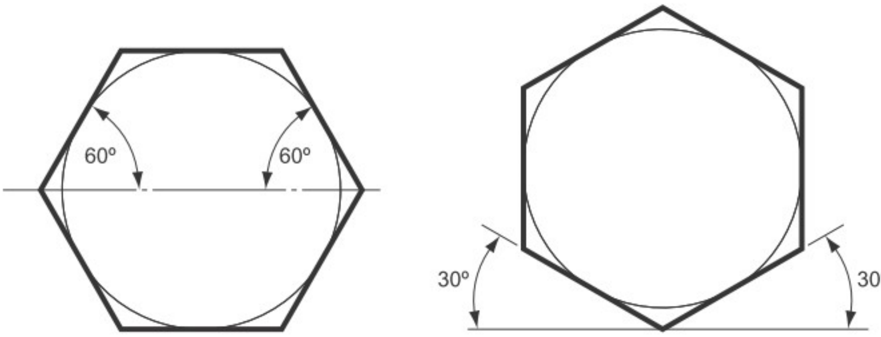

Figure 2.28

To construct a regular octagon given the diagonal, i.e. within a given circle (Fig. 2.29).

1. Draw the circle and insert a diameter AE.
2. Construct another diagonal CG, perpendicular to the first diagonal.
3. Bisect the four quadrants thus produced to cut the circle in B, D, F and H.

ABCDEFGH is the required octagon.

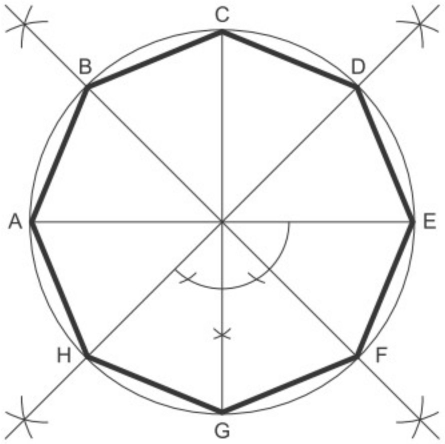

Figure 2.29

To construct a regular octagon given the diameter, i.e. within a given square (Fig. 2.30).

1. Construct a square PQRS, length of side equal to the diameter.
2. Draw the diagonals SQ and PR to intersect in T.
3. With centres P, Q, R and S draw four arcs, radius PT (= QT = RT = ST) to cut the square in A, B, C, D, E, F, G and H.

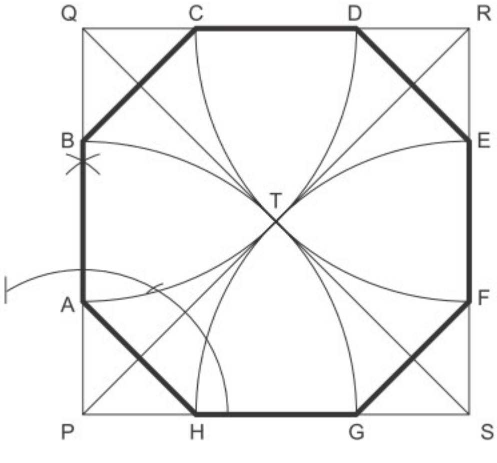

Figure 2.30

ABCDEFGH is the required octagon.

To construct any given polygon, given the length of a side.

There are three fairly simple ways of constructing a regular polygon. Two methods require a simple calculation and the third requires very careful construction if it is to be exact. All three methods are shown. The construction work for any polygon, and a heptagon (seven sides), has been chosen to illustrate them.

Method 1 (Fig. 2.31).

1. Draw a line AB equal in length to one of the sides and produce AB to P.
2. Calculate the exterior angle of the polygon by dividing 360° by the number of sides. In this case the exterior angle is 360°/7 = 51 3°/7.
3. Draw the exterior angle PBC so that BC = AB.
4. Bisect AB and BC to intersect in O.
5. Draw a circle, centre O and radius OA (= OB = OC).
6. Step off the sides of the figure from C to D, D to E, etc.

ABCDEFG is the required heptagon.

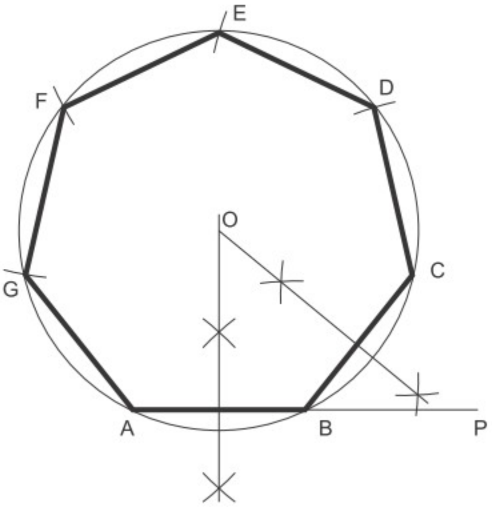

Figure 2.31

Method 2 (Fig. 2.32).

1. Draw a line AB equal in length to one of the sides.
2. From A erect a semi-circle, radius AB to meet BA produced in P.
3. Divide the semi-circle into the same number of equal parts as the proposed polygon has sides. This may be done by trial and error or by calculation (180°/7 = 25 5°/7 for each arc).
4. Draw a line from A to point 2 (for ALL polygons). This forms a second side to the polygon.
5. Bisect AB and A_2 to intersect in O.
6. With centre O draw a circle, radius OB (= OA = O_2).
7. Step off the sides of the figure from B to C, C to D, etc.

ABCDEFG is the required septagon.

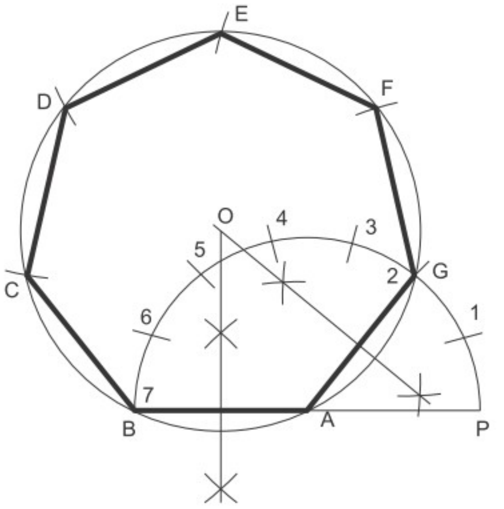

Figure 2.32

Method 3 (Fig. 2.33).

1. Draw a line GA equal in length to one of the sides.
2. Bisect GA.
3. From A construct an angle of 45° to intersect the bisector at point 4.
4. From G construct an angle of 60° to intersect the bisector at point 6.
5. Bisect between points 4 and 6 to give point 5.
   Point 4 is the centre of a circle containing a square. Point 5 is the centre of a circle containing a pentagon. Point 6 is the centre of a circle containing a hexagon. By marking off points at similar distances the centres of circles containing any regular polygon can be obtained.

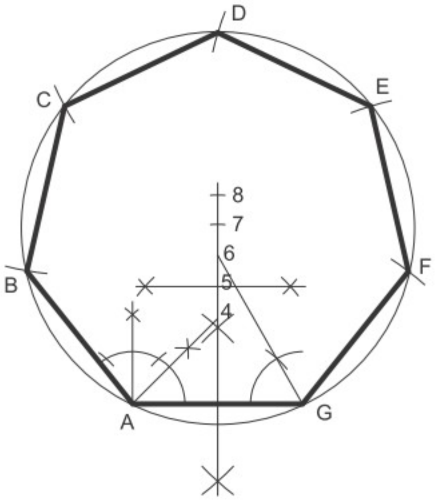

Figure 2.33

6. Mark off point 7 so that 6 to 7 = 5 to 6 (= 4 to 5).
7. With centre at point 7 draw a circle, radius 7 to A (= 7 to G).
8. Step off the sides of the figure from A to B, B to C, etc.

ABCDEFG is the required heptagon.

To construct a regular polygon given a diagonal, i.e. within a given circle (Fig. 2.34).

1. Draw the given circle and insert a diameter AM.
2. Divide the diameter into the same number of divisions as the polygon has sides.
3. With centre M draw an arc, radius MA. With centre A draw another arc of the same radius to intersect the first arc in N.
4. Draw N_2 and produce to intersect the circle in B (for any polygon).
5. AB is the first side of the polygon. Step out the other sides BC, CD, etc.

ABCDE is the required polygon.

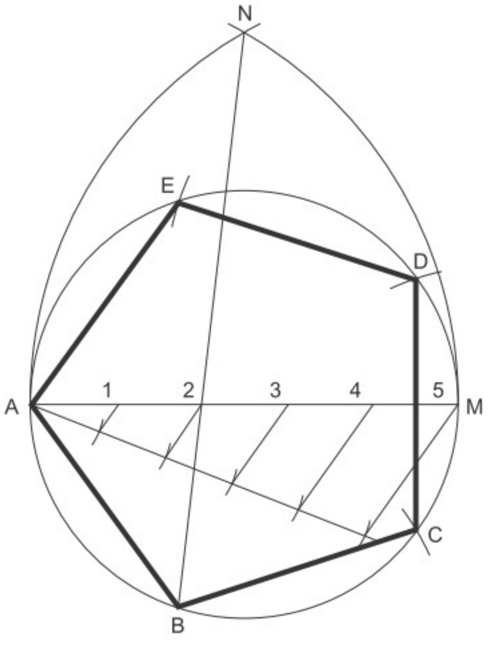

Figure 2.34

To construct a regular polygon given a diameter (Fig. 2.35).

1. Draw a line MN.
2. From some point A on the line draw a semi-circle of any convenient radius.
3. Divide the semi-circle into the same number of equal sectors as the polygon has sides (in this case nine, i.e. 20° intervals).
4. From A draw radial lines through points 1 to 8.

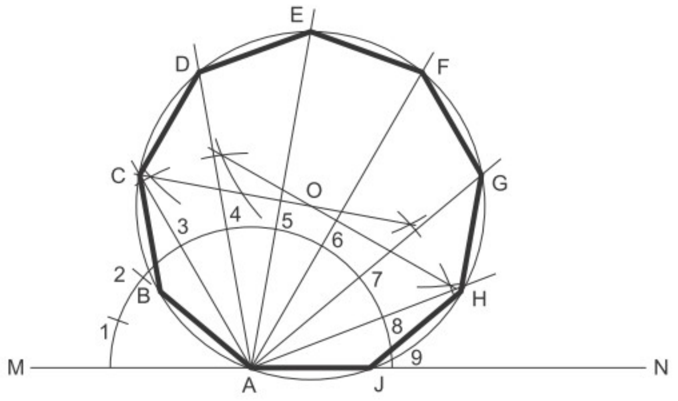

Figure 2.35

5. If the polygon has an even number of sides, there is only one diameter passing through A. In this case, bisect the known diameter to give centre O. If, as in this case, there are two diameters passing through A (there can never be more than two), then bisect both diameters to intersect in O.
6. With centre O and radius OA, draw a circle to intersect the radial lines in C, D, E, F, G and H.
7. From A mark off AB and AJ equal to CD, DE, etc.

ABCDEFGHJ is the required polygon.

The constructions shown above are by no means all the constructions that you may be required to do, but they are representative of the type that you may meet.

If your geometry needs a little extra practice, it is well worthwhile proving these constructions by Euclidean proofs. A knowledge of some geometric theorems is needed when answering many of the questions shown below, and proving the above constructions will make sure that you are familiar with them.

Exercise 2

1. Construct an equilateral triangle with sides 60 mm long.
2. Construct an isosceles triangle that has a perimeter of 135 mm and an altitude of 55 mm.
3. Construct a triangle with base angles 60° and 45° and an altitude of 76 mm.
4. Construct a triangle with a base of 55 mm, an altitude of 62 mm and a vertical angle of $37\frac{1}{2}^\circ$.
5. Construct a triangle with a perimeter measuring 160 mm and sides in the ratio 3:5:6.
6. Construct a triangle with a perimeter of 170 mm and sides in the ratio 7:3:5.
7. Construct a triangle given that the perimeter is 115 mm, the altitude is 40 mm and the vertical angle is 45°.
8. Construct a triangle with a base measuring 62 mm, an altitude of 50 mm and a vertical angle of 60°. Now draw a similar triangle with a perimeter of 250 mm.
9. Construct a triangle with a perimeter of 125 mm, whose sides are in the ratio 2:4:5. Now draw a similar triangle whose perimeter is 170 mm.

10. Construct a square of side 50 mm. Find the mid-point of each side by construction and join up the points with straight lines to produce a second square.

11. Construct a square whose diagonal is 68 mm.

12. Construct a square whose diagonal is 85 mm.

13. Construct a parallelogram given two sides 42 mm and 90 mm long, and the angle between them 67°.

14. Construct a rectangle that has a diagonal 55 mm long and one side 35 mm long.

15. Construct a rhombus if the diagonal is 75 mm long and one side is 44 mm long.

16. Construct a trapezium given that the parallel sides are 50 mm and 80 mm long and are 45 mm apart.

17. Construct a regular hexagon, 45 mm side.

18. Construct a regular hexagon if the diameter is 75 mm.

19. Construct a regular hexagon within an 80 mm diameter circle. The corners of the hexagon must all lie on the circumference of the circle.

20. Construct a square, side 100 mm. Within the square, construct a regular octagon. Four alternate sides of the octagon must lie on the sides of the square.

21. Construct the following regular polygons:
a pentagon, side 65 mm;
a heptagon, side 55 mm;
a nonagon, side 45 mm;
a decagon, side 35 mm.

22. Construct a regular pentagon, diameter 82 mm.

23. Construct a regular heptagon within a circle, radius 60 mm. The corners of the heptagon must lie on the circumference of the circle.

.6 ? 2 " NNN

3 Isometric Projection

Engineering drawings are always drawn in orthographic projection. For the presentation of detailed drawings, this system has been found to be far superior to all others. The system has, however, the disadvantage of being very difficult to understand by people not trained in its usage. It is always essential that an engineer be able to communicate his ideas to anybody, particularly people who are not engineers, and it is therefore an advantage to be able to draw using a system of projection that is more easily understood. There are many systems of projection and this book deals with three: isometric, oblique and orthographic projections. Isometric and oblique projections present the more pictorial view of an object.

True isometric projection is an application of orthographic projection and is dealt with in greater detail later in this chapter. The most common form of isometric projection is called ‘conventional isometric’.

Conventional Isometric Projection (Isometric Drawing)

If you were to make a freehand drawing of a row of houses, the house furthest away from you would be the smallest house on your drawing. This is called the ‘perspective’ of the drawing and, in a perspective drawing, none of the lines are parallel. Isometric drawing ignores perspective altogether. Lines are drawn parallel to each other and drawings can be made using a T-square and a set square. This is much simpler than perspective drawing.

Figure 3.1 shows a shaped block drawn in conventional isometric projection.

You will note that there are three isometric axes. These are inclined at 120° to each other. One axis is vertical and the other two axes are therefore at 30° to the horizontal. Dimensions measured along these axes, or parallel to them, are true lengths.

The faces of the shaped block shown in Fig. 3.1 are all at 90° to each other. The result of this is that all of the lines in the isometric drawing are parallel to the isometric axes. If the lines are not parallel to any of the isometric axes, they are no longer true lengths. An example of this is shown in Fig. 3.2 which shows an isometric drawing of a regular hexagonal prism. The hexagon is first drawn as a plane figure and a simple shape, in this case a rectangle, is drawn around the hexagon. The rectangle is easily drawn in isometric and the positions of the corners of the hexagon can be transferred from the plane figure to the isometric drawing with a pair of dividers.

Geometric and Engineering Drawing. DOI: 10.1016/B978-0-08-096768-4.00003-6
© 2010 Elsevier Ltd. All rights reserved.

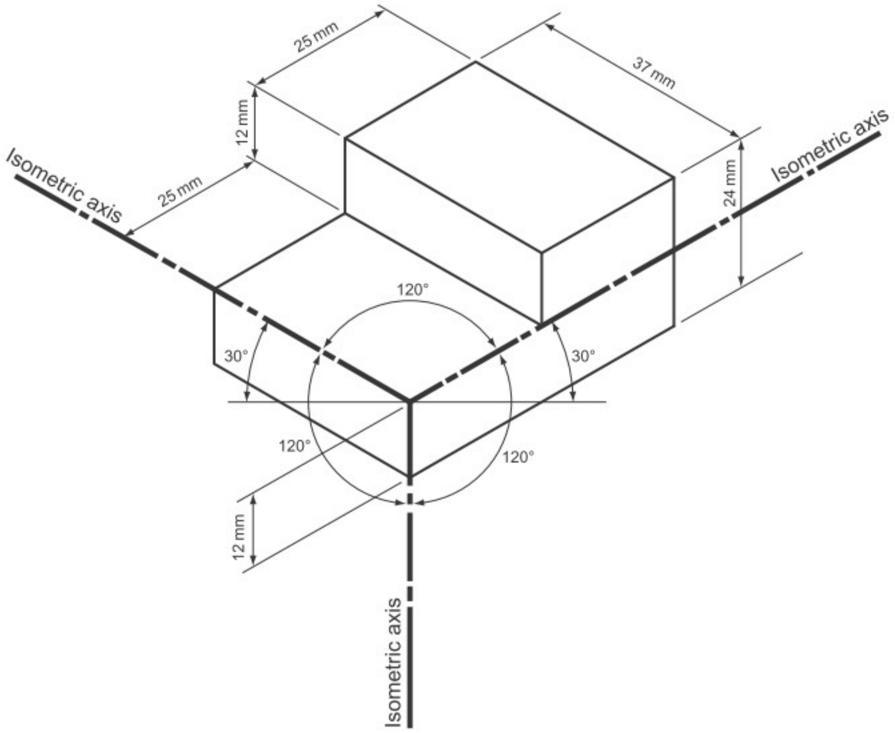
25 mm
37 mm
12 mm
25 mm
24 mm
Isometric axis
Isometric axis
120°
30°
30°
120°
120°
12 mm
Isometric axis

Figure 3.1 Shaped block drawn in conventional isometric projection.

The dimensions of the hexagon should all be 25 mm and you can see from Fig. 3.2 that lines not parallel to the isometric axes do not have true lengths.

Figure 3.3 shows another hexagonal prism. This prism has been cut at an incline and this means that two extra views must be drawn so that sufficient information to draw the prism in isometric can be transferred from the plane views to the isometric drawing.

This figure shows that, when making an isometric drawing, all dimensions must be measured parallel to one of the isometric axes.

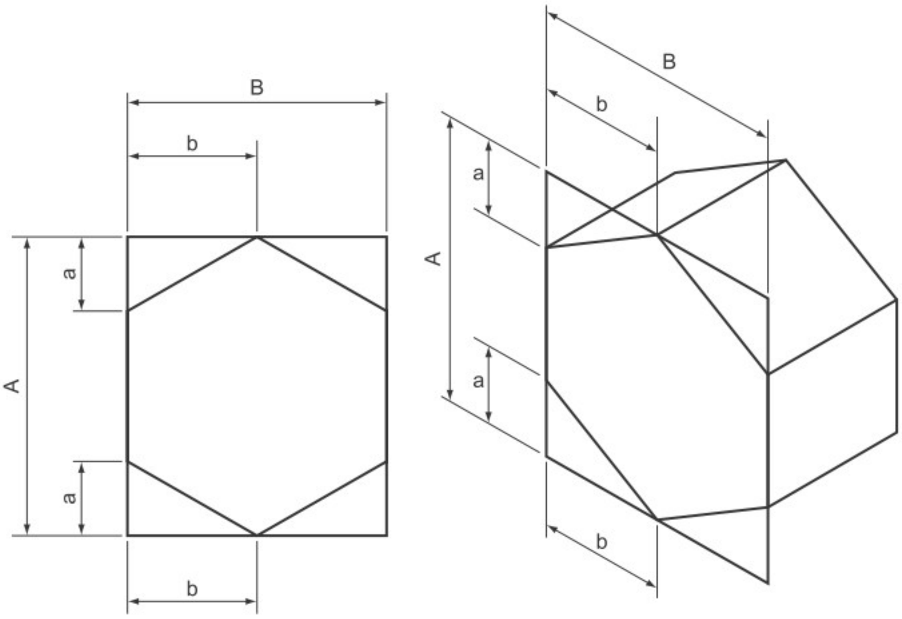
B
b
a
A
a
b
B
b
a
A
a
b

Figure 3.2 An isometric drawing of a regular hexagonal prism.

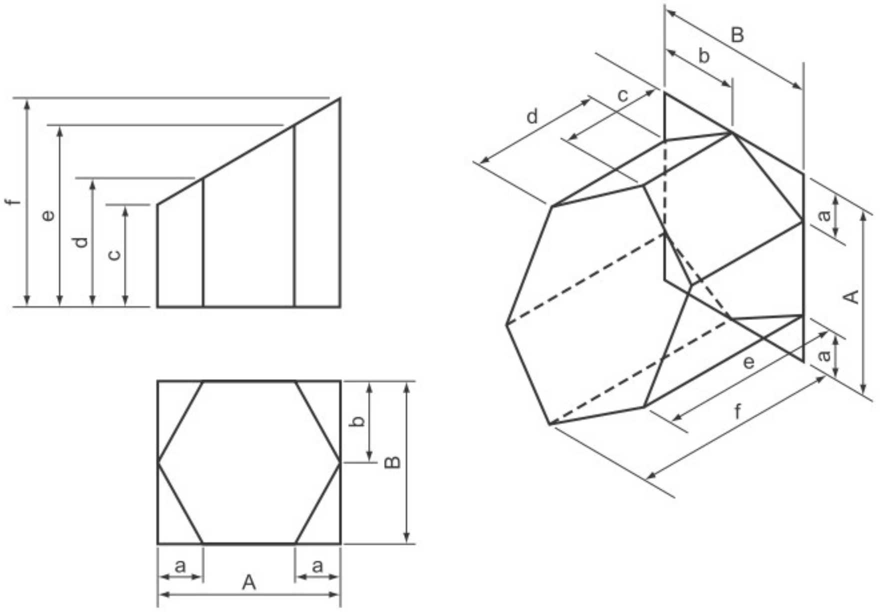
f
e
d
c
b
B
a
A
a
B
b
c
d
a
A
a
e
f

Figure 3.3 A hexagonal prism cut obliquely.

Circles and Curves Drawn in Isometric Projection

All of the faces of a cube are square. If a cube is drawn in isometric projection, each square side becomes a rhombus. If a circle is drawn on the face of a cube, the circle will change shape when the cube is drawn in isometric projection. Figure 3.4 shows how to plot the new shape of the circle.

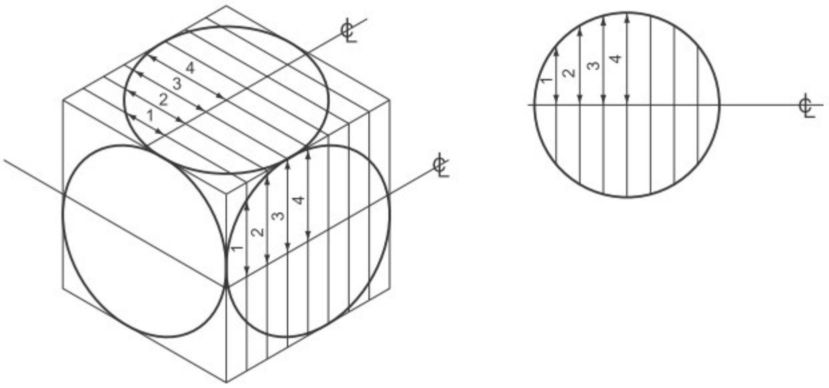

Figure 3.4

The circle is first drawn as a plane figure, and is then divided into an even number of equal strips. The face of the cube is then divided into the same number of equal strips. Centre lines are added and the measurement from the centre line of the circle to the point where strip 1 crosses the circle is transferred from the plane drawing to the isometric drawing with a pair of dividers. This measurement is applied above and below the centre line. This process is repeated for strips 2, 3, etc.

The points that have been plotted should then be carefully joined together with a neat freehand curve.

Figure 3.5 illustrates how this system is used in practice.

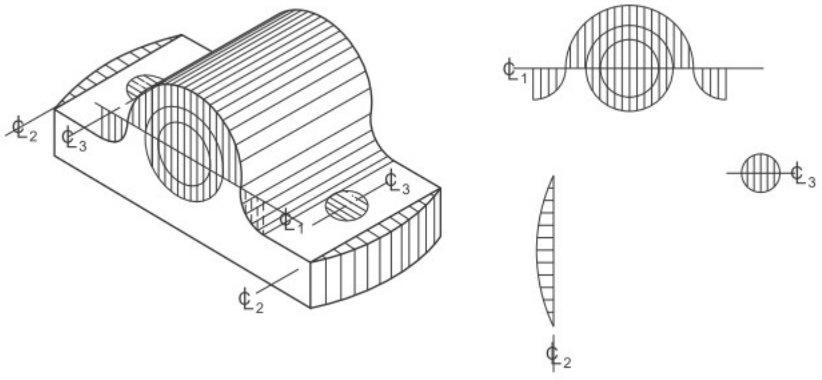

Figure 3.5 How circles and part circles can be drawn isometrically.

Since a circle can be divided into four symmetrical quadrants, it is really necessary to draw only a quarter of a circle instead of a whole plane circle.

The dimensions which are transferred from the plane circle to the isometric view are called 'ordinates' and the system of transferring ordinates from plane figures to isometric views is not confined to circles. It may be used for any regular or irregular shape. Figure 3.6 shows a shaped plate.

There are several points worth noting from Fig. 3.6.

- **(a)** Since the plate is symmetrical about its centre line, only half has been divided into strips on the plane figure.
- **(b)** In proportion to the plate, the holes are small. They have, therefore, ordinates much closer together so that they can be drawn accurately.
- **(c)** The point where the vee cut-out meets the elliptical outline has its own ordinate so that this point can be transferred exactly to the isometric view.
- **(d)** Since the plate has a constant thickness, the top and bottom profiles are the same. A quick way of plotting the bottom profile is to draw several vertical lines down from the top profile and, with dividers set at the required thickness of the plate, follow the top curve with the dividers, marking the thickness of the plate on each vertical line.

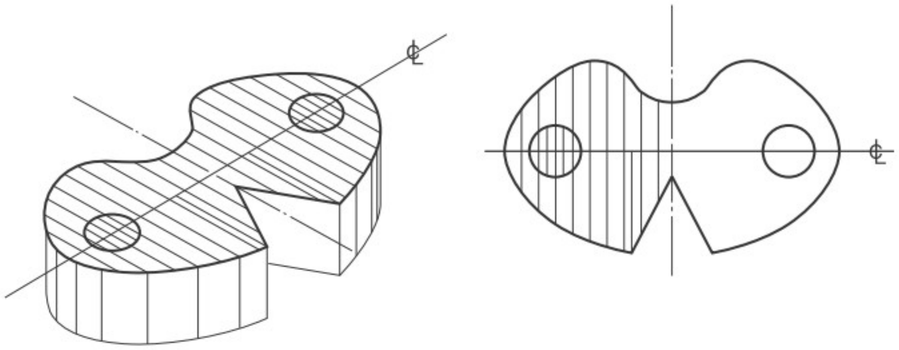

Figure 3.6 A shaped plate drawn in isometric projection.

It is sometimes necessary to draw circles or curves on faces that are not parallel to any of the three isometric axes. Figure 3.7 shows a cylinder cut at 45°. Two views of the cylinder have to be drawn: a plan view and an elevation. The plan view is divided into strips and the positions of these strips are projected onto the elevation.

The base of the cylinder is drawn in isometric projection in the usual way. Points 1 to 20, where the strips cross the circle, are projected vertically upwards and the height of the cylinder, measured from the base with dividers, is transferred for each point in turn from the elevation to the isometric view. These points are then carefully joined together with a neat freehand curve.

True Isometric Projection

Isometric projection is a method of drawing with instruments which gives a pictorial view of an object. It is not often used in industry and, when it is, the vast majority of

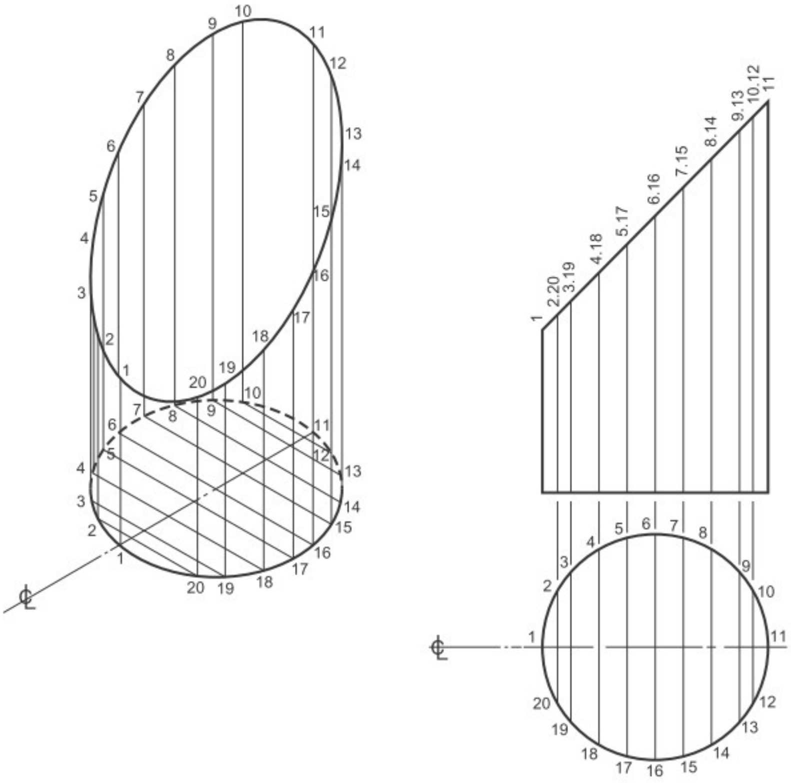
10
9
11
8
12
7
13
6
14
5
15
4
16
3
17
18
2
1
19
20
10
7
8
9
6
11
5
12
4
13
3
14
2
15
1
16
17
20
19
18
℄
1
2.20
3.19
4.18
5.17
6.16
7.15
8.14
9.13
10.12
11
5
6
7
8
4
3
9
2
10
1
11
℄
20
12
19
13
18
14
17
16
15

Figure 3.7 An cylinder cut obliquely.

drawings are made using conventional isometric projection. Conventional isometric projection is a distorted and simplified form of true isometric. True isometric is found by taking a particular view from an orthographic projection of an object. Figure 3.8 shows a cube, about 25 mm side, drawn in orthographic projection with the cube so positioned that the front elevation is a true isometric projection of the cube. The three isometric axes are still at 120° to each other. In conventional isometric, distances measured parallel to these axes are true lengths. In true isometric projection they are no longer true lengths although they are proportional to their true lengths. However, the horizontal distances on a true isometric projection *are* true lengths. The reduction of lengths measured parallel to the isometric axes makes the overall size of the true isometric drawing appear to look more natural, particularly when directly compared with an orthographic or plane view of the same object (compare the relative sizes of the prism in Fig. 3.2).

If the horizontal length and the length parallel to the isometric axes were both to be true lengths, the isometric axes would have to be at 45° (Fig. 3.9). Since the isometric axes are at 30°, the 45° lengths must be reduced.

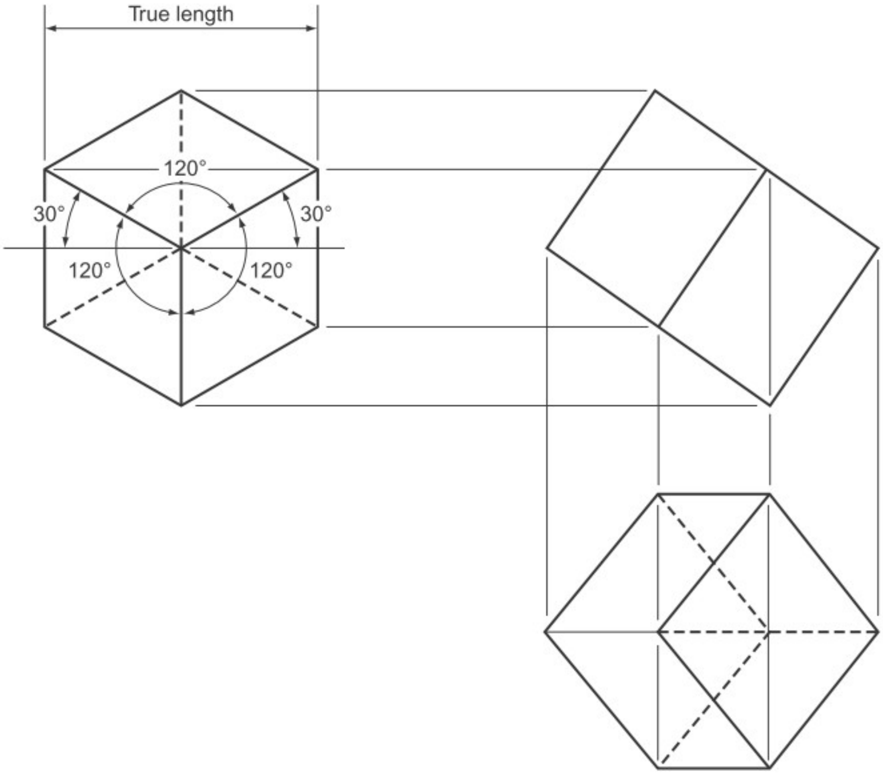
True length
120°
30°
30°
120°
120°

Figure 3.8 True isometric projection (drawn orthographically).

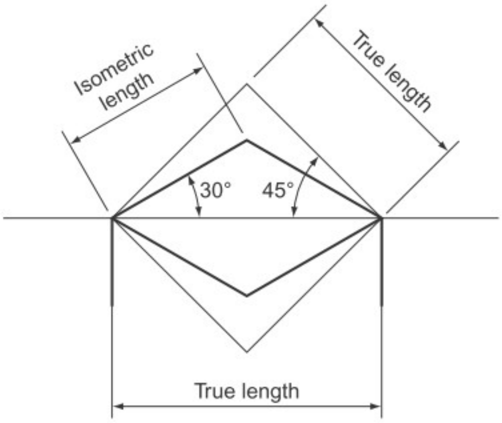
Isometric length
True length
30°
45°
True length

Figure 3.9

This operation is shown in Fig. 3.10.

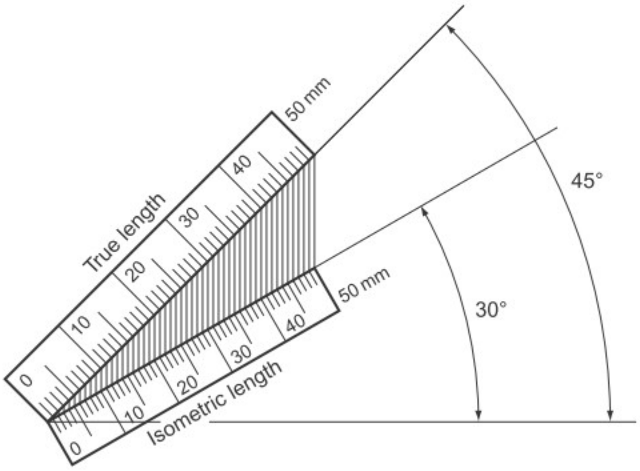
50 mm
0
10
20
30
40
True length
45°
30°
50 mm
0
10
20
30
40
Isometric length

Figure 3.10

The ratio between the true length and the isometric length is isometric length = true length × 0.8165.

This ratio is constant for all lines measured parallel to any of the isometric axes. If you are asked to draw an object using an isometric scale, your scale may be constructed as in Fig. 3.10 or you may construct a conventional plain scale as shown in Fig. 3.11. The initial length of this scale is 100 × 0.8165 = 81.65 mm. The scale is then completed as shown in Chapter 1.

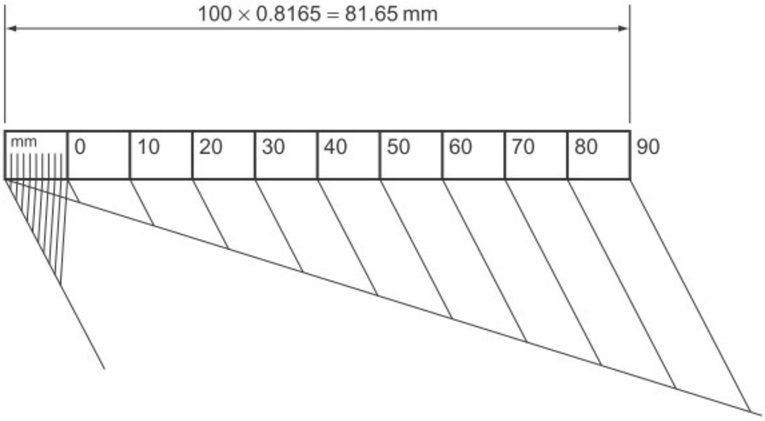
100 × 0.8165 = 81.65 mm
mm
0
10
20
30
40
50
60
70
80
90

Figure 3.11 A plain scale to measure true isometric lengths.

Exercise 3

(All questions originally set in imperial units.)

1. Draw, full size, an isometric projection of the component shown in Fig. 1 looking in the direction of the arrow A. Hidden details are not to be shown.
   Associated Lancashire Schools Examining Board

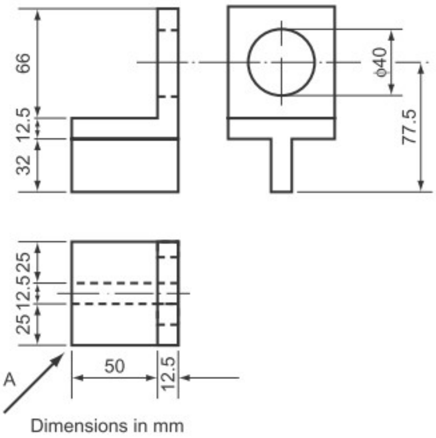
66
12.5
32
ϕ40
77.5
25
12.5
25
A
50
12.5
Dimensions in mm

Figure 1

2. Figure 2 shows the front elevation and plan of an ink bottle stand. Make a full size isometric drawing of the stand with corner A nearest to you. Hidden details should *not* be shown.
   West Midlands Examinations Board

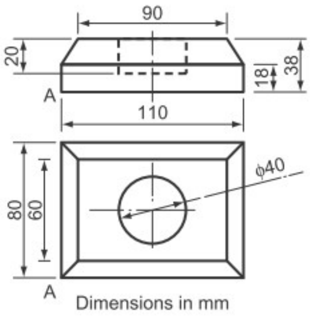
90
20
18
38
A
110
80
60
ϕ40
A
Dimensions in mm

Figure 2

3. Figure 3 shows the development of a hexagonal box. Draw, in isometric projection, the assembled box standing on its base. Ignore the thickness of the material and omit hidden detail.
North Western Secondary School Examinations Board (see Chapter 14 for information not in Chapter 3).

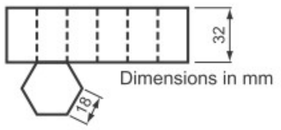
32
18
Dimensions in mm

Figure 3

4. Three views of a bearing are shown in Fig. 4. Make an isometric drawing of the bearing. Corner A should be the lowest point on your drawing. No hidden details are required.
South-East Regional Examinations Board

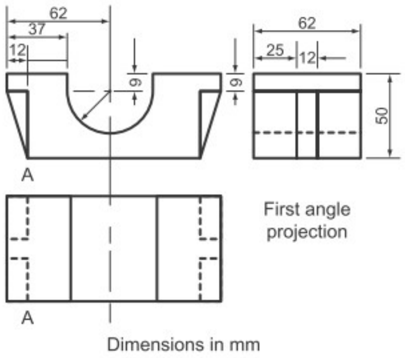
62
37
12
9
9
62
25
12
50
A
A
First angle projection
Dimensions in mm

Figure 4

5. Two views of a plain shaft bearing are shown in Fig. 5. Make a full size isometric drawing of the bearing. Hidden details should *not* be shown.
West Midlands Examinations Board

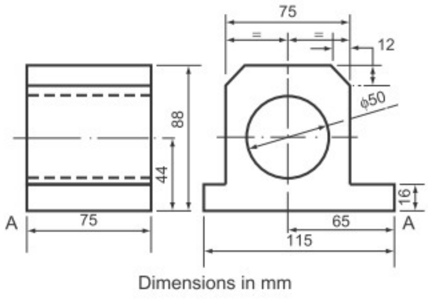
75
=
=
12
ϕ50
88
44
16
A
75
65
A
115
Dimensions in mm

Figure 5

6. Construct an isometric drawing of the casting shown in Fig. 6. Make point X the lowest point of your drawing. Do not use an isometric scale.
University of London School Examinations

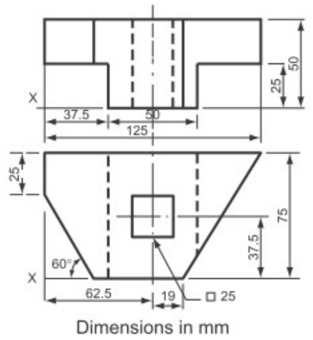
X
37.5
50
125
25
50
25
60°
X
62.5
19
□ 25
37.5
75
Dimensions in mm

Figure 6

7. Construct an isometric scale and use it to make a true isometric view of the casting shown in Fig. 7. Corner X is to be the lowest corner of your drawing.
University of London School Examinations

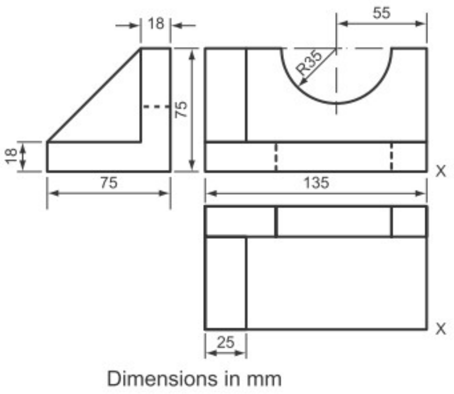
18
55
R35
75
18
75
135
X
X
25
Dimensions in mm

Figure 7

8. A plan and elevation of the base of a candlestick are shown in Fig. 8. Draw (a) another elevation when the base is viewed in the direction of the arrow and (b) an accurate isometric view of that half of the candlestick base indicated by the letters *abcd* on the plan view. The edge *ab* is to be in the foreground of your drawing.
Southern Universities' Joint Board (see Chapter 8 for information not in Chapter 3).

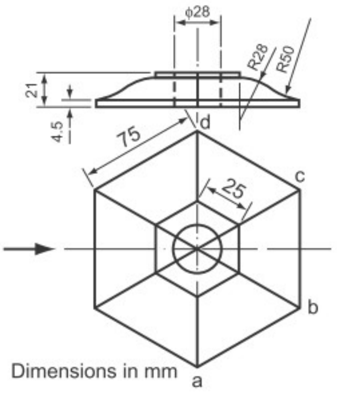
ϕ28
R28
R50
21
4.5
75
d
c
25
b
a
Dimensions in mm

Figure 8

9. Figure 9 shows the plan and elevation of an angle block. Make a full size isometric projection of the block, making the radiused corner the lowest part of your drawing. Do *not* use an isometric scale. Hidden detail should be shown.
Oxford and Cambridge Schools Examinations Board

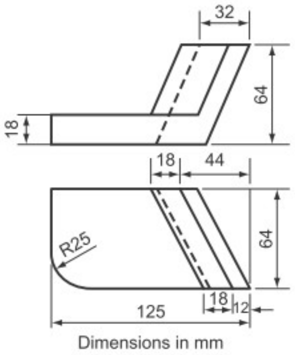
32
64
18
18
44
64
R25
125
18
12
Dimensions in mm

Figure 9

10. Figure 10 shows two views of a cylindrical rod with a circular hole. Make an isometric drawing of the rod in the direction of the arrows R and S. Hidden detail is not required.
Associated Examining Board

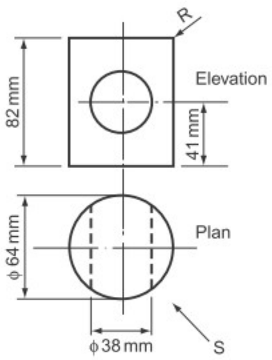
R
Elevation
82 mm
41 mm
Plan
ϕ 64 mm
ϕ 38 mm
S

Figure 10

4 The Construction of Circles to Satisfy Given Conditions

The Greatest Invention of All Time

About 6000 years ago, an unknown Mesopotamian made one of the greatest inventions of all time: the wheel. This was the most important practical application ever made of a shape that fascinated early mathematicians. The shape is, of course, the circle. After the wheel had been invented, the Mesopotamians found many more applications for the circle than just for transport. The potter's wheel was developed and vessels were made much more accurately and quickly. Pulleys were invented and engineers and builders were able to raise heavy weights. Since that time, the circle has been the most important geometric shape in the development of all forms of engineering.

Apart from its practical applications, the circle has an aesthetic value which makes it unique among plane figures. The ancients called it 'the perfect curve' and its symmetry and simplicity has led artists and craftsmen to use the circle as a basis for design for many thousands of years.

Definitions

A circle is the locus of a point which moves so that it is always a fixed distance from another stationary point.

Concentric circles are circles that have the same centre.
Eccentric circles are circles that are not concentric.

Figure 4.1 shows some of the parts of the circle.

Constructions

The length of the circumference of a circle is πD or $2\pi R$, where D is the diameter and R the radius of the circle, π is the ratio of the diameter to the circumference and may be taken as 22/7 or, more accurately, as 3.142.

If you need to draw the circumference of a circle (this is required quite often in subsequent chapters), you should either calculate it, or use the construction shown in Fig. 4.2. This construction is not exact but is accurate enough for most needs. For the

Geometric and Engineering Drawing. DOI: 10.1016/B978-0-08-096768-4.00004-8
© 2010 Elsevier Ltd. All rights reserved.

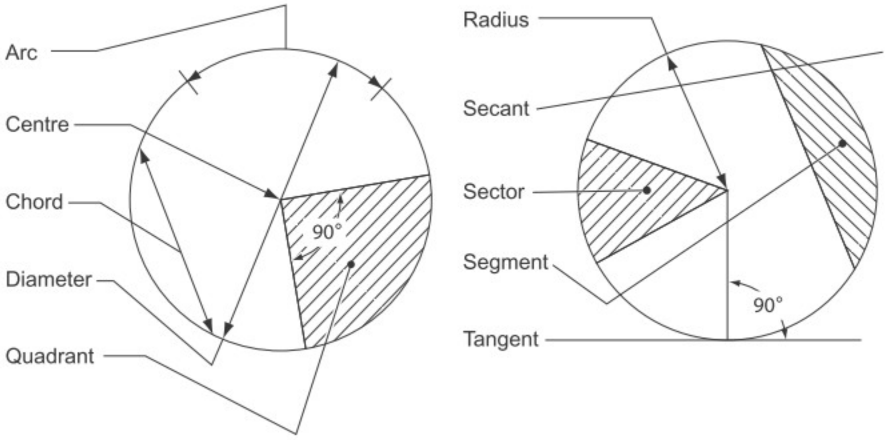
Arc
Centre
Chord
Diameter
Quadrant
90°
Radius
Secant
Sector
Segment
Tangent
90°

Figure 4.1 Parts of the circle.

sake of thoroughness, the corresponding construction, that of finding the diameter from the circumference, is shown in Fig. 4.3.

To construct the circumference of a circle, given the diameter (Fig. 4.2).

1. Draw a semi-circle of the given diameter AB, centre O.
2. From B mark off three times the diameter, BC.
3. From O draw a line at 30° to OA to meet the semi-circle in D.
4. From D draw a line perpendicular to OA to meet OA in E.
5. Join EC.

EC is the required circumference.

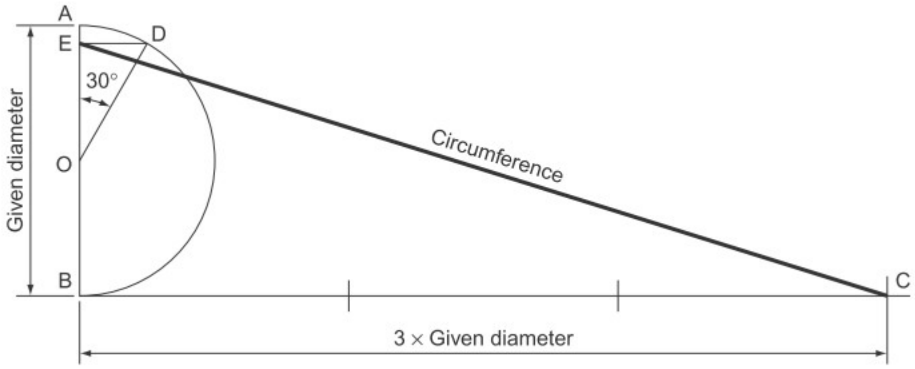
A
E
D
30°
O
B
C
Given diameter
Circumference
3 × Given diameter

Figure 4.2

.6 ' /

To construct the diameter of a circle, given the circumference (Fig. 4.3).

1. Draw the given circumference AB.
2. Bisect AB in C.
3. With centre C, and radius CA, draw a semi-circle.
4. With centre B, and radius BC, draw an arc to cut the semi-circle in D.
5. From D draw a perpendicular to AB, to cut AB in E.
6. With centre E and radius ED, draw an arc to cut AB in F.

AF is the required diameter.

The rest of this chapter shows some of the constructions for finding circles drawn to satisfy certain given conditions.

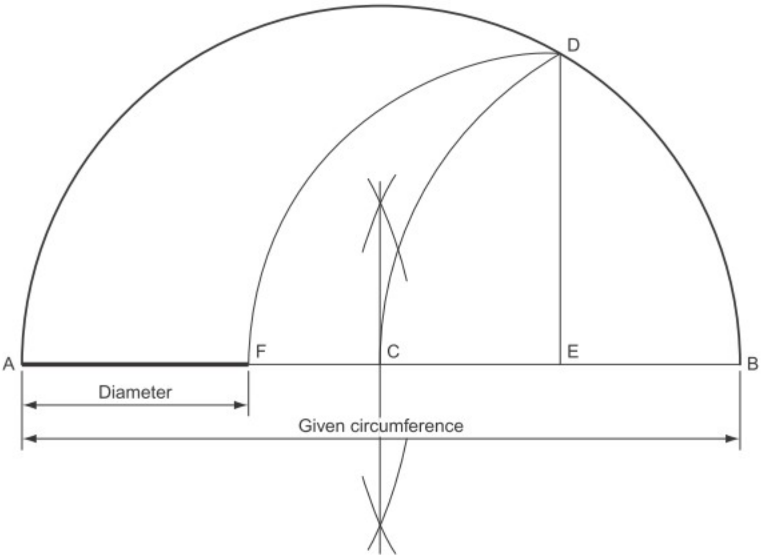

Figure 4.3

To find the centre of any circle (Fig. 4.4).

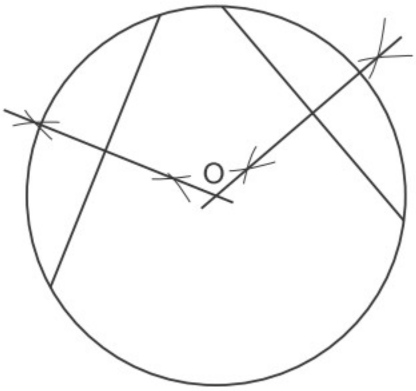

Figure 4.4

1. Draw any two chords.
2. Construct perpendicular bisectors to these chords to intersect in O.

O is the centre of the circle.

To construct a circle to pass through three given points (Fig. 4.5).

1. Draw straight lines connecting the points as shown. These lines are, in fact, chords of the circle.
2. Draw perpendicular bisectors through these lines to intersect in O.

O is the centre of a circle that passes through all three points.

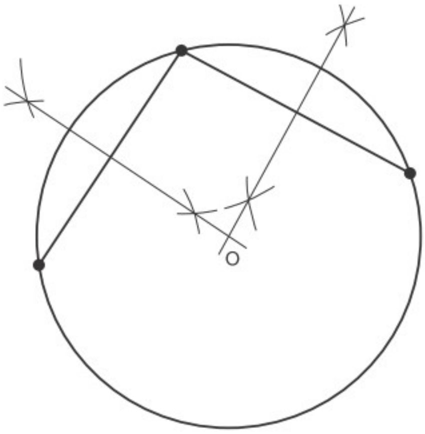

Figure 4.5

To construct the inscribed circle of any regular polygon (in this case, a triangle) (Fig. 4.6).

1. Bisect any two of the interior angles to intersect in O. (If the third angle is bisected it should also pass through O.)

O is the centre of the inscribed circle. This centre is called the 'incentre'.

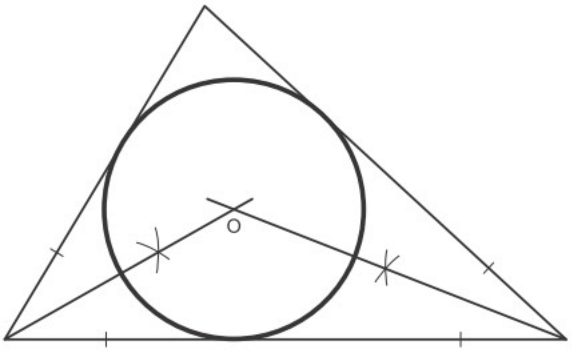

Figure 4.6

To construct the circumscribed circle of any regular polygon (in this case a triangle) (Fig. 4.7).

1. Perpendicularly bisect any two sides to intersect in O. (If the third side is bisected it should also pass through O.)
   O is the centre of the circumscribed circle. This centre is called the 'circumcentre'.

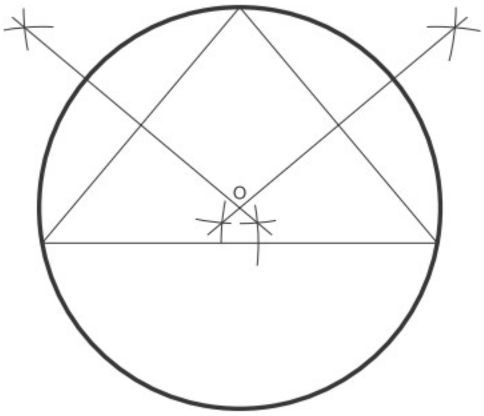

Figure 4.7

To construct the escribed circle to any regular polygon (in this case a triangle) (Fig. 4.8).

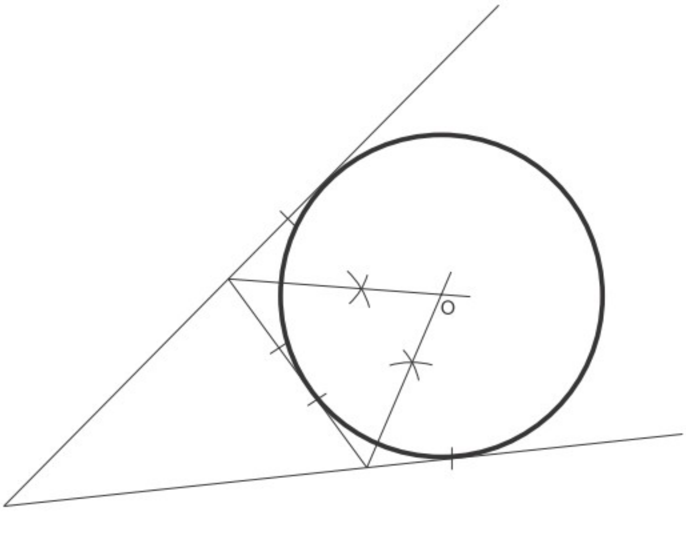

Figure 4.8

1. An escribed circle is a circle that touches a side and the two adjacent sides produced. Thus, the first step is to produce the adjacent sides.
2. Bisect the exterior angles thus formed to intersect in O.

O is the centre of the escribed circle.

To construct a circle that passes through a fixed point A and touches a line at a given point B (Fig. 4.9).

1. Join AB.
2. From B erect a perpendicular BC.
3. From A construct angle $B\hat{A}O$ similar to angle $C\hat{B}A$ to intersect the perpendicular in O.

O is the centre of the required circle.

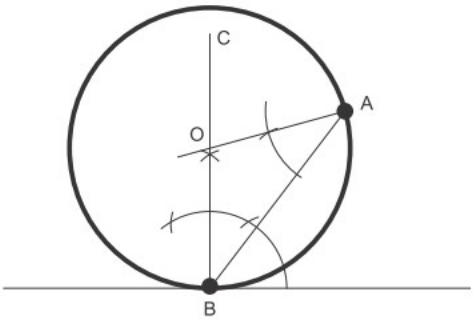

Figure 4.9

To construct a circle that passes through two given points, A and B, and touches a given line (Fig. 4.10).

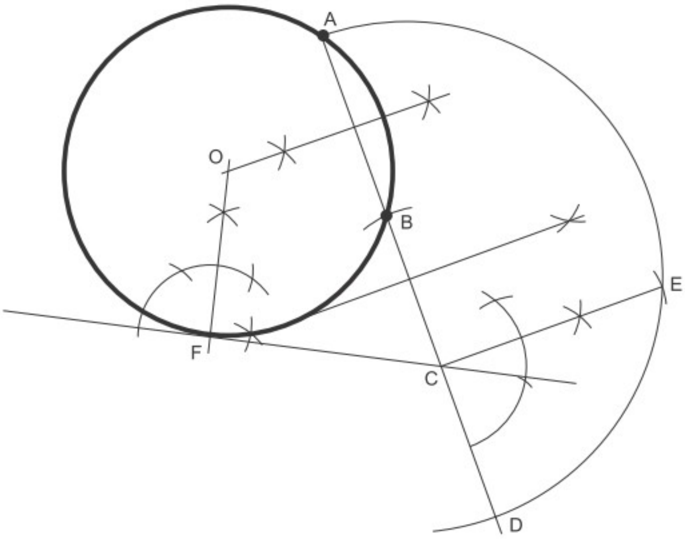

Figure 4.10

1. Join AB and produce this line to D (cutting the given line in C) so that BC = CD.
2. Construct a semi-circle on AD.
3. Erect a perpendicular from C to cut the semi-circle in E.
4. Make CF = CE.
5. From F erect a perpendicular.
6. Perpendicularly bisect AB to meet the perpendicular from F in O.

O is the centre of the required circle.

To construct a circle that touches two given lines and passes through a given point P. There are two circles which satisfy these conditions (Fig. 4.11).

1. If the two lines do not meet, produce them to intersect in A.
2. Bisect the angle thus formed.
3. From any point on the bisector draw a circle, centre B, to touch the two given lines.
4. Join PA to cut the circle in C and D.
5. Draw PO, parallel to CB and PO, parallel to DB.

O_1 and O_2 are the centres of the required circles.

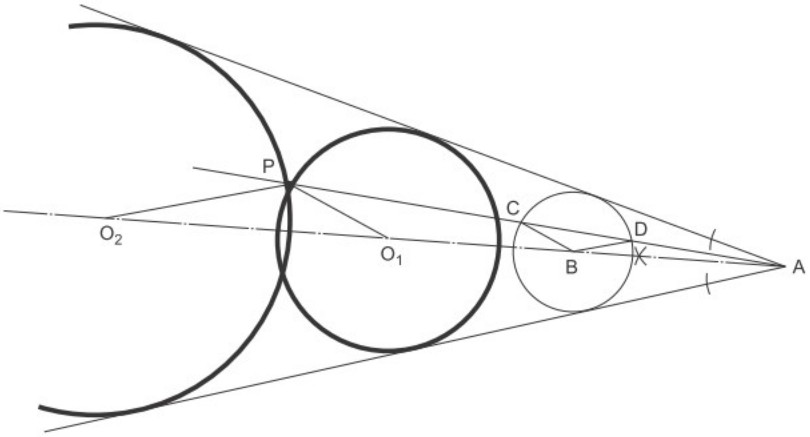

Figure 4.11

To construct a circle, radius *R*, to touch another given circle radius *r*, and a given line (Fig. 4.12).

1. Draw a line parallel to the given line, the distance between the lines equal to R.
2. With compass point at the centre of the given circle and radius set at $R + r$, draw an arc to cut the parallel line in O.

O is the centre of the required circle.

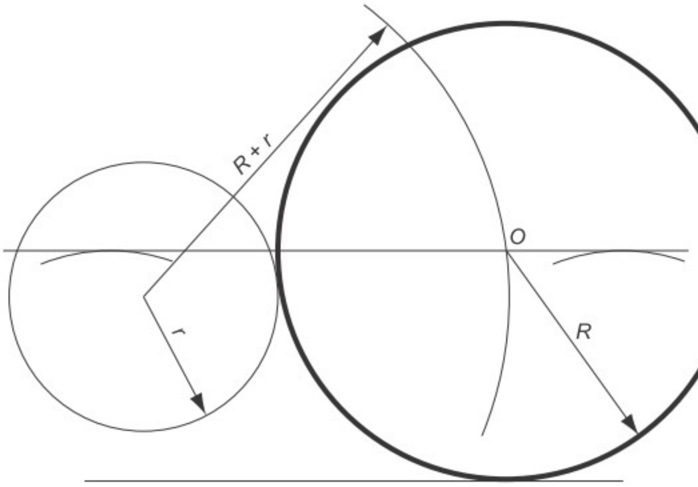

Figure 4.12

To construct a circle that touches another circle and two tangents of that circle (Fig. 4.13).

1. If the tangents do not intersect, produce them to intersect in A.
2. Bisect the angle formed by the tangents.
3. From B, the point of contact of the circle and one of its tangents, construct a perpendicular to cut the bisector in O_1. This is the centre of the given circle.
4. Join BD.
5. Draw EF parallel to DB and FO_2 parallel to BO_1.

O_2 is the centre of the required circle.

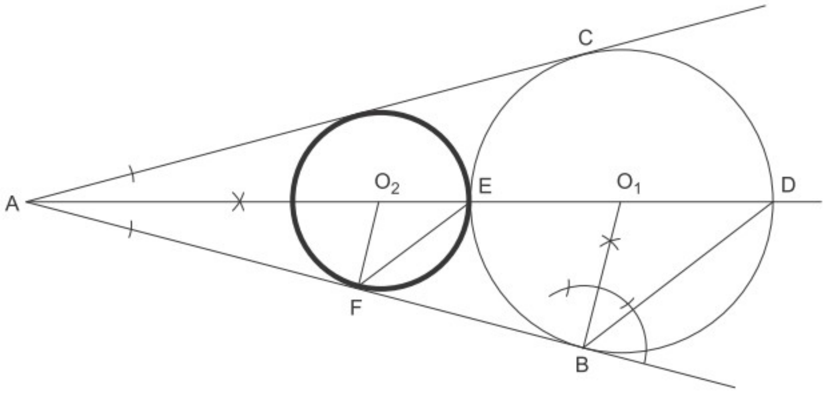

Figure 4.13

To construct a circle that touches another circle and two lines (Fig. 4.14).

1. Draw intersecting lines parallel to the given lines. These lines, AB and AC, must be distance *r*, the radius of the given circle, from the given lines.
2. Repeat construction of Fig. 4.11 and construct a circle that passes through O_1, the centre of the given circle, and touches the two parallel lines.
3. The centre of this circle, O_2, is also the centre of the required circle.

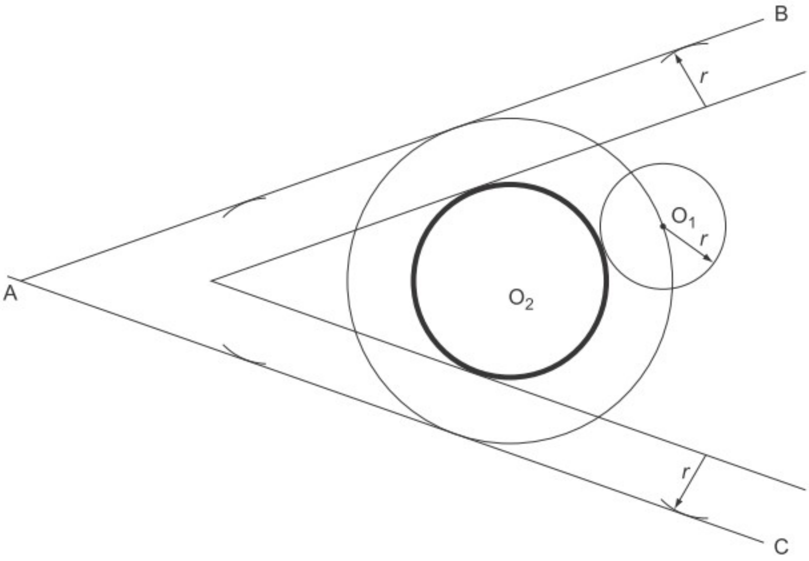

Figure 4.14

To construct a circle that passes through two given points, P and Q, and touches a given circle, centre D (Fig. 4.15).

All of the above constructions are for finding single circles which satisfy given conditions. The rest of the constructions in this chapter are concerned with more than one circle at a time.

1. Join PQ and produce this line.
2. Perpendicularly bisect PQ and, with centre somewhere on this bisector, draw a circle to pass through points P and Q and cut the given circle in A and B.
3. Join AB and produce to cut PQ produced in C.
4. Construct the tangent from C to the given circle. (Join CD, bisect CD in E, compass point at E draw a radius ED to cut the circle in F.)
5. From F erect a perpendicular to cut the bisector of PQ in O.

O is the centre of the required circle.

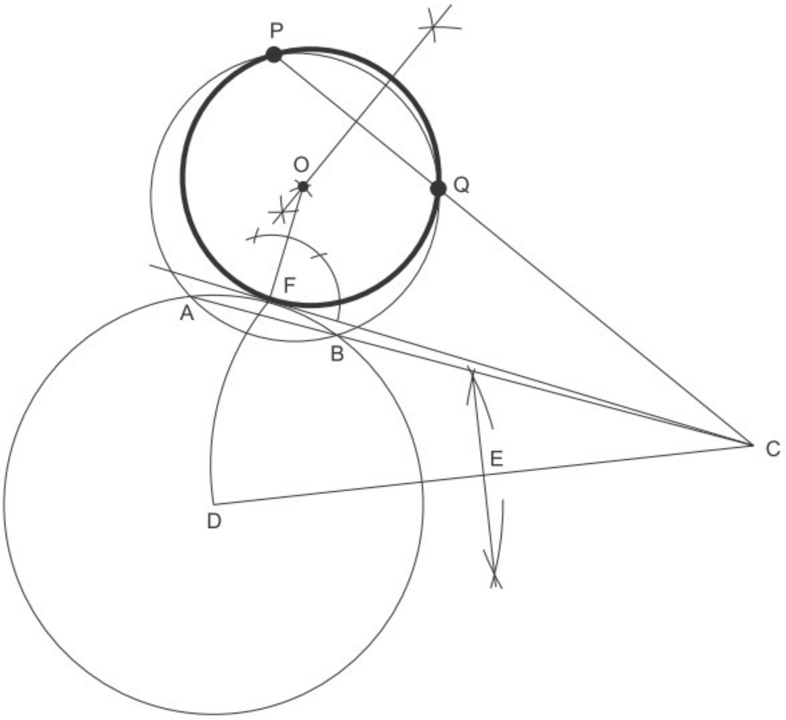

Figure 4.15

To draw three circles which touch each other, given the position of their centres $\mathbf{O_1}$, $\mathbf{O_2}$ and $\mathbf{O_3}$ (Fig. 4.16).

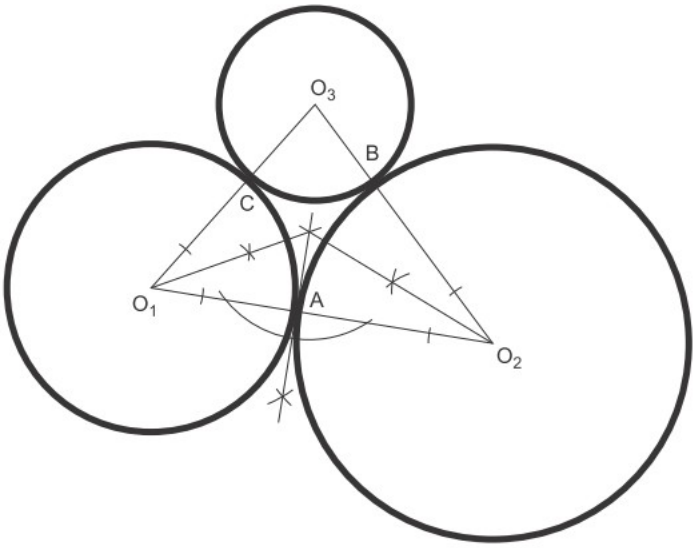

Figure 4.16

1. Draw straight lines connecting the centres.
2. Find the centre of the triangle thus formed by bisecting two of the interior angles.
3. From this centre, drop a perpendicular to cut O_1O_2 in A.
4. With centre O_1 and radius O_1A, draw the first circle.
5. With centre O_2 and radius O_2A, draw the second circle.
6. With centre O_3 and radius O_3C (= O_3B), draw the third circle.

To draw two circles, given both their radii, within a third circle, all three circles to touch each other (Fig. 4.17).

1. Mark off the diameter AB of the largest circle.
2. Mark off AO_1 equal to the radius of one of the other circles and draw this circle, centre O_1, to cut the diameter in C.
3. From C mark off CD equal to the radius of the third circle.
4. Mark off BE equal to the radius of the third circle.
5. With centre O_1 and radius O_1D, draw an arc.
6. With centre O and radius OE, draw an arc to cut the first arc in O_2.

O_2 is the centre of the third circle.

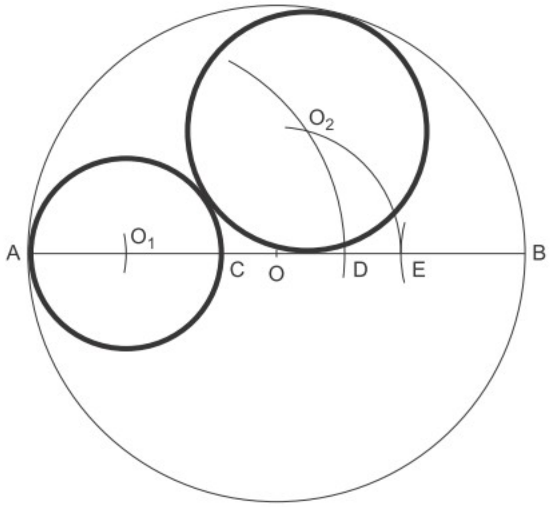

Figure 4.17

To draw any number of equal circles within another circle, the circles all to be in contact (in this case 5) (Fig. 4.18).

1. Divide the circle into the same number of sectors as there are proposed circles.
2. Bisect all the sectors and produce one of the bisectors to cut the circle in D.
3. From D erect a perpendicular to meet OB produced in E.
4. Bisect $D\hat{E}O$ to meet OD in F.
5. F is the centre of the first circle. The other circles have the same radius and have centres on the intersections of the sector bisectors and a circle, centre O and radius OF.

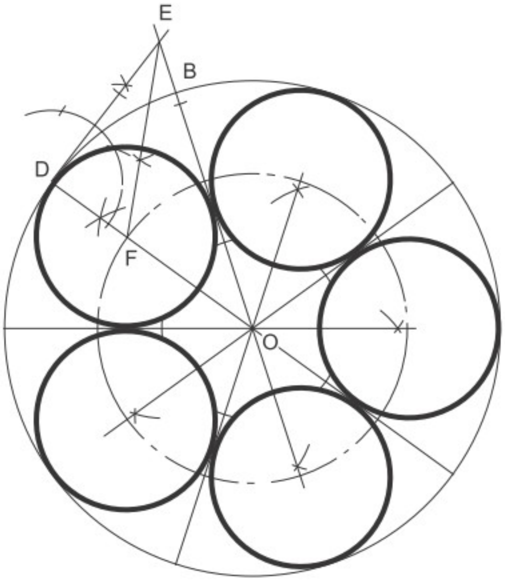

Figure 4.18

To draw several equal circles within a regular polygon to touch each other and one side of the polygon (in this case, a septagon) (Fig. 4.19).

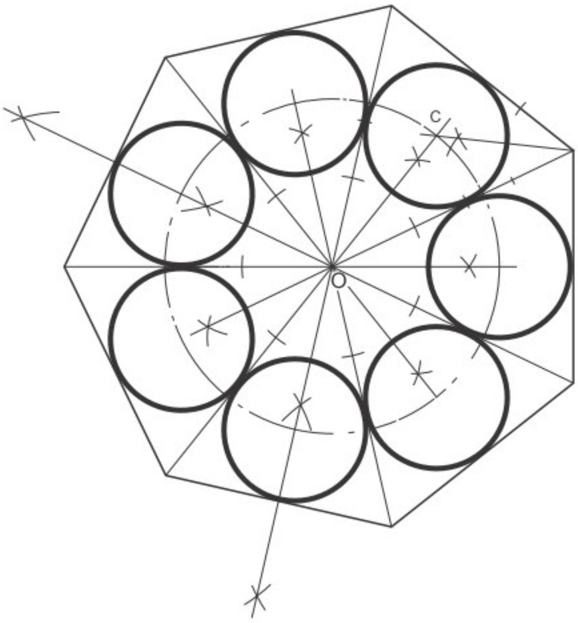

Figure 4.19

1. Find the centre of the polygon by bisecting two of the sides.
2. From this centre, draw lines to all of the corners.
3. This produces several congruent triangles. All we now need to do is to draw the inscribed circle in each of these triangles. This is done by bisecting any two of the interior angles to give the centre C.
4. The circles have equal radii and their centres lie on the intersection of a circle, radius OC and the bisectors of the seven equal angles formed by step 2.

To draw equal circles around a regular polygon to touch each other and one side of the polygon (in this case, a septagon) (Fig. 4.20).

1. Find the centre of the polygon by bisecting two of the sides.
2. From the centre O draw lines through all of the corners and produce them.
3. Bisect angles $C\hat{A}B$ and $D\hat{B}A$ to intersect in E.
4. E is the centre of the first circle. The rest can be obtained by drawing a circle, radius OE, and bisecting the seven angles formed by step 2. The intersections of this circle and these lines give the centres of the other six circles.

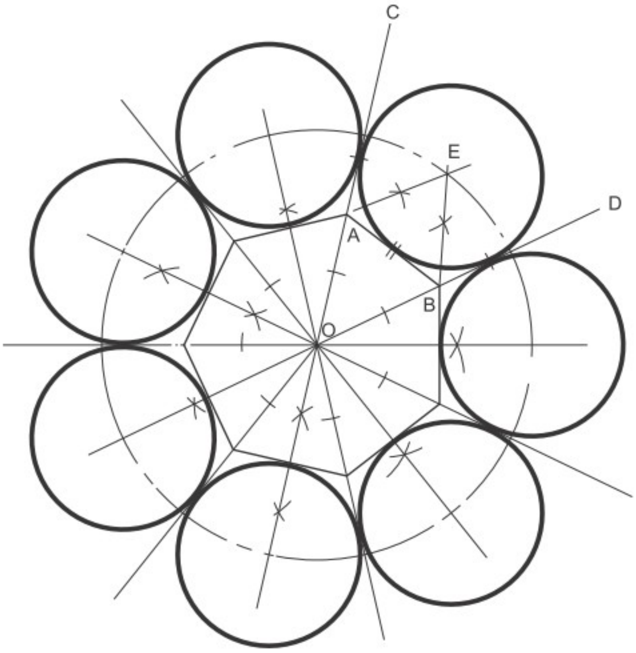

Figure 4.20

Exercise 4

(All questions originally set in imperial units.)

1. Construct a regular octagon on a base line 25 mm long and draw the inscribed circle. Measure and state the diameter of this circle in mm.

North Western Secondary School Examinations Board (see Chapter 2 for information not in Chapter 4).

2. Describe *three* circles, each one touching the other two externally, their radii being 12, 18 and 24 mm, respectively.
   North Western Secondary School Examinations Board
3. No construction has been shown in Fig. 1. You are required to draw the figure full size showing all construction lines necessary to ensure the circles are tangential to their adjacent lines.
   Southern Regional Examinations Board

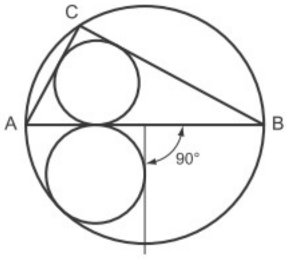

AB = ϕ144 mm
AC = 68 mm

Figure 1

4. Construct the triangle ABC in which the base BC = 108 mm, the vertical angle Â = 70° and the altitude is 65 mm.
   D is a point on AB 34 mm from A. Describe a circle to pass through the points A and D and touch (tangential to) the line BC.
   Southern Universities' Joint Board (see Chapter 2 for information not in Chapter 4).
5. Figure 2 shows two touching circles placed in the corner made by two lines which are perpendicular to one another. Draw the view shown and state the diameter of the smaller circle. Your construction must show clearly the method of obtaining the centre of the smaller circle.
   University of London School Examinations

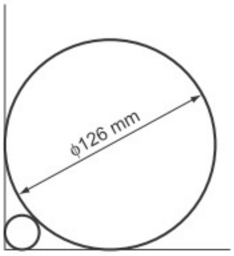

Figure 2

6. Figure 3 shows two intersecting lines AB and BC and the position of a point P. Draw the given figure and find the centre of a circle that will pass through P and touch the lines AB and BC. Draw the circle and state its radius as accurately as possible.
   University of London School Examinations

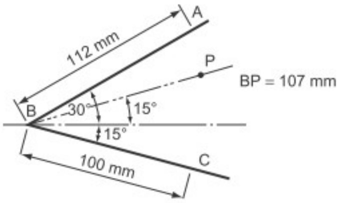

Figure 3

7. A triangle has sides 100, 106 and 60 mm long. Draw the triangle and construct the following: (a) the inscribed circle, (b) the circumscribed circle, (c) the smallest escribed circle.
   University of London School Examinations (see Chapter 2 for information not in Chapter 4).
8. Construct an isosceles triangle ABC where the included angle $A = 67\frac{1}{2}°$, and AB = AC = 104 mm. Draw circles of 43, 37 and 32 mm radius using as centres A, B and C, respectively.
   Construct the smallest circle that touches all three circles.
   Measure and state the diameter of the constructed circle.
   Associated Examining Board (see Chapter 2 for information not in Chapter 4).
9. AB and AC are two straight lines which intersect at an angle of 30°. D is a point between the two lines at perpendicular distances of 37 and 62 mm, respectively from AB and AC. Describe the circle that touches the two converging lines and passes through point D; the centre of this circle is to lie between the points A and D. Now draw two other circles each touching the constructed circle externally and also the converging lines. Measure and state the diameters of the constructed circles.
   Oxford Local Examinations
10. OA and OB are two straight lines meeting at an angle of 30°. Construct a circle of diameter 76 mm to touch these two lines and a smaller circle that will touch the two converging lines and the first circle.
    Also construct a third circle of diameter 64 mm which touches each of the other two circles.
    Oxford Local Examinations
11. Construct a regular octagon of side 75 mm and within this octagon describe eight equal circles each touching one side of the octagon and two adjacent circles. Now draw the smallest circle that will touch all eight circles. Measure and state the diameter of this circle.
    Oxford Local Examinations (see Chapter 2 for information not in Chapter 4).

data 04 04 2020

Pinching Sphere theorem
Chapter 4
Evolution of the Curvature

Source: Pag. 63 (82)

127) 20200406

A) Data-NOVOVIVO_65/2020/
Books_65_2020/
Higher Algebra Classical/
Higher Algebra Classical by Sadhan...pdf
Pag. 4 → 2. Complex Numbers

B) Books_65_2020/
Descriptive Geometry/
Geometric and Engineering Drawing
by K. Morling...pdf

Pág. 5 (22)

Diagonal Scales

Pág. 6 (23) Figure 1.4

Pag. 7 (24) Figure 1.5
Figure 1.6

Pag. 8 (25) Figure 1.7

Pag. 11 (28) 2 → the Construction of

data 06 04 2020

Geometric Figures from given data

Obs: Usar este livro.

Pág. 44 (61)

Figure 4.1 Parts of the circle

Figure 4.2

The construct the circumference of a circle, given the diametre

(important)

Figure 4.3

To construct the diameter of a circle, given the circumference

Figure 4.4

To find the centre of any circle

Figure 4.5

To construct a circle to pass through three given points

Figure 4.6

To construct the inscribed circle

Jandaia

data 06.04.2020

of any regular polygon
(in this case, a triangle)

Figure 4.7

To construct the circumscribed circle of any regular polygon (triangle)

Figure 4.8

To construct the escribed circle to any regular polygon (triangle)

Figure 4.9

To construct a circle that passes through a fixed point A and touches a line at a given point B.

... Figure 4.20

Pag. 55(72)

C) Books_65_2020/
Descriptive Geometry/
A.N. Whitehead - The Axioms of Descriptive Geometry....pdf
Pag. 4(13)

Peano uses the following useful notation which is an extension of his notation for segments and prolongations.

+,- * *. * ! + F L " ; +0
1 / 2 8 : 1 2 0 D" /"
( ) %

CHAPTER I.

FORMULATIONS OF THE AXIOMS.

1. THE general considerations which must govern a mathematical investigation on the foundations of Geometry have been explained in Chapter I of the previous tract of this series, on the Axioms of Projective Geometry*. It is explained there that 'Descriptive Geometry' is here used as a generic term for any Geometry in which two straight lines in a plane do not necessarily intersect. Also it is pointed out that the purely classificatory portions of a Descriptive Geometry are clumsy and uninteresting, and that accordingly the idea of order is introduced from the very beginning.

There are three main ways by which this introduction of order can be conveniently managed. In one way, which is represented by Peano's axioms given below (§§ 3—6), the *class* of points which lie *between* any two points is taken as a fundamental idea. It is then easy to define the class of points collinear with the two points and lying *beyond* one of them. Thus these three classes of points, namely the two classes lying beyond the two points respectively and the class lying between the two points, together with the two points themselves form the straight line defined by the two points. Then a set of axioms of the straight line are required, concerned with the idea of 'between,' and also axioms are required respecting coplanar lines.

Another way, which was pointed out by Vailati† and Russell‡, is to conceive a straight line as essentially a serial relation involving two terms. The whole field of such a relation, namely the terms which are thus ranged in order by it, forms the class of points on the straight line. Thus the Geometry starts with the fundamental conception of a class of *relations*. The axioms of the straight line are the axioms which secure that each of these relations is a serial relation. The points are the entities occurring in the fields of any of these relations. The axioms of the plane are the same as in the previous mode of development.

---

* In the sequel this tract will be referred to as 'Proj. Geom.'

† Cf. *Rivista di Matematica*, vol. IV.

‡ Cf. *Principles of Mathematics*, § 376.

The third way, recently developed by Prof. O. Veblen*, is to consider the science of Descriptive Geometry as the study of the properties of one single three-termed relation of order. The entities forming the field of this relation are the points. When this relation holds between three points A, B, and C, it is said that 'the points A, B, and C, are in the linear order ABC.' This method of conceiving the subject results in a notable simplification, and combines advantages from the two previous methods. Veblen's axioms will be stated in full (cf. § 8).

2. The enunciation of the axioms of Descriptive Geometry, which is given in the sections (§§ 3—6) immediately following, is that due to Peano†. His formulation is based upon that of Pasch‡, to whom is due the first satisfactory systematic exposition of the subject. The undefined fundamental ideas are two in number, namely that of a class of entities called 'points,' and that of the 'class of points lying between any two given points.' It has already been explained§ that this undetermined class of points is in fact any class of entities with interrelations, such that the axioms are satisfied when considered as referring to them.

The symbol AB will represent the class of points lying between the points A and B. This class will be called the segment AB.

The first group of axioms, eleven in number, secure the ordinary properties of a straight line with respect to the order of points on it, and also with respect to the division of a line‖ into three parts by any two points on it, and into two parts by any single point on it. The Dedekind property* is not secured by them, but, compactness† is secured by axiom IV.

---

* Cf. *A System of Axioms for Geometry*, Trans. of the Amer. Math. Soc., vol. v., 1904.

† Cf. *I principii di Geometria*, Turin, 1889. These axioms are repeated by him in an article, *Sui fondamenti della Geometria*, Rivista di Matematica, vol. IV. 1894. In this latter article the minute mathematical deductions are omitted, and their place is taken by valuable observations on the main points to be considered. Also a treatment of congruence is given which does not appear in the earlier tract. This article should be studied carefully by every student of the subject.

‡ Cf. *Vorlesungen über neuere Geometrie*, Leipzig, 1882. This treatise is the classic work on the subject.

§ Cf. Proj. Geom. § 2.

‖ Note that 'line' will be habitually used for 'straight line.'

3. Peano's axioms of the straight line are as follows:

I. There is at least one point.

II. If A is any point, there is a point distinct from A.

III. If A is a point, there is no point lying between A and A.

It follows that the class AA possesses no members.

IV. If A and B are distinct points, there is at least one point lying between A and B.

Thus the class AB is not the null class.

V. If the point C lies between A and B, it also lies between B and A.

It easily follows that the classes AB and BA are identical.

VI. The point A does not lie between the points A and B.

Thus the class, or segment, AB does not include its end-points A and B.

Definition. If A and B are points, the symbol $A'B$ represents the class of points, such as C, with the property that B lies between A and C. Thus $A'B$ is the prolongation of the line beyond B, and $B'A$ is its prolongation beyond A.

VII. If A and B are distinct points, there exists at least one member of $A'B$.

VIII. If A and D are distinct points, and C is a member of AD, and B of AC, then B is a member of AD.

IX. If A and D are distinct points, and B and C are members of AD, then either B is a member of AC, or B is identical with C, or B is a member of CD.

X. If A and B are distinct points, and C and D are members of $A'B$, then either C is identical with D, or C is a member of BD, or D is a member of BC.

XI. If A, B, C, D are points, and B is a member of AC, and C of BD, then C is a member of AD.

Definition. The straight line possessing A and B, symbolized by str (A, B), is composed of the three classes $A'B$, AB, $B'A$, together with the points A and B themselves.

Then by the aid of the previous axioms the usual theorems, excluding the Dedekind property, respecting the order of points on a line can be proved. Also any two points are both contained by one and only one line.

---

* Cf. Proj. Geom. § 19, and § 9 of the present tract.

† Cf. Proj. Geom. § 16.

4. Peano uses the following useful notation which is an extension of his notation for segments and prolongations. If A is a point and u is a class of points, then Au is the class of points lying on the segments between A and points of u, and $A'u$ is the class of points on the prolongations of these segments beyond the points of u.

Then in conformity with this notation the seven regions into which a plane is divided by three lines are as in the figure.

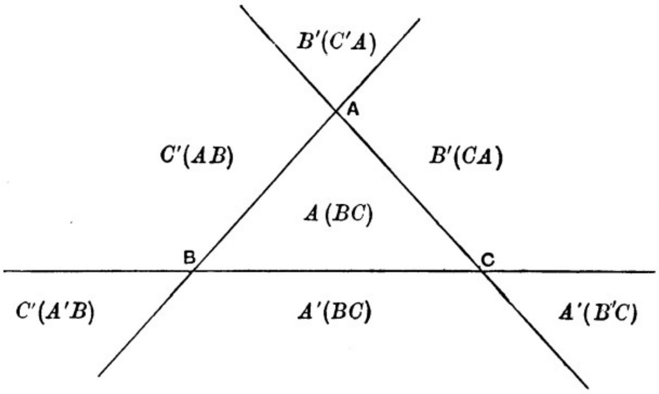

The plane determined by the three non-collinear points A, B, C—written ple (A, B, C)—is defined to be the class of points consisting of the points on the three lines str (BC), str (CA), and str (AB), and of the points in the seven regions $A\,(BC)$, $A'\,(BC)$, $B'\,(CA)$, $C'\,(AB)$, $A'\,(B'C)$, $B'\,(C'A)$, $C'\,(A'B)$.

5. Three axioms are required to establish the Geometry of a plane.

XII. If r is a straight line, there exists a point which does not lie on r.

Note that it would be sufficient to enunciate this axiom for one straight line.

XIII. If A, B, C are three non-collinear points, and D lies on the segment BC, and E on the segment AD, there exists a point F on both the segment AC and the prolongation $B'E$ (cf. fig. i, p. 5).

XIV. If A, B, C are three non-collinear points, and D lies on the segment BC, and F on the segment AC, there exists a point E lying on both the segments AD and BF (cf. fig. ii).

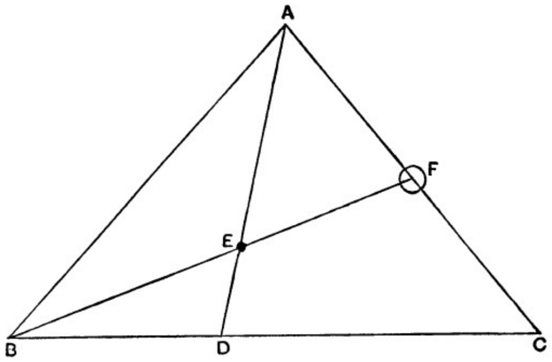

Fig. i.

With these axioms all the usual properties of the division of a plane by a line, and of the inside and outside of a plane closed figure,

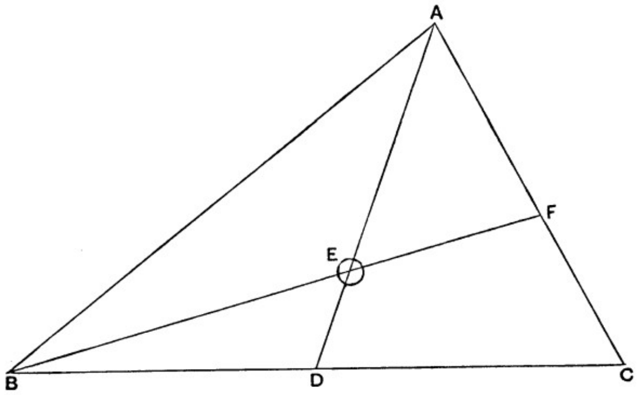

Fig. ii.

can be proved. Thus if ABC form a triangle and a coplanar line intersect the segment BC, it must intersect one and only one of the segments CA and AB.

Also any three non-collinear points lie in one and only one plane; and the line determined by any two points lying in a plane lies entirely in that plane. But, as the case of Euclidean Geometry shews, we cannot prove from these axioms that any two lines in a plane intersect.

6. For three-dimensional Geometry two other axioms are required.

XV. A point can be found external to any plane. The enunciation of this axiom can be restricted to a particular plane.

XVI. Given any plane p, and any point A outside it, and any point Q on it, and any point B on the prolongation $A'Q$, then, if X is any other point, either X lies on the plane p, or AX intersects the plane p, or BX intersects the plane p.

The annexed figure illustrates the axiom, the points X_1, X_2, X_3 being positions of X which illustrate the three alternatives contemplated in the axiom. Thus X_1 lies on the plane p; X_2 lies on the same side of p as B, so that AX_2 must cut p in some point L; X_3 lies on the same side of p as A, so that BX_3 must cut p in some point M.

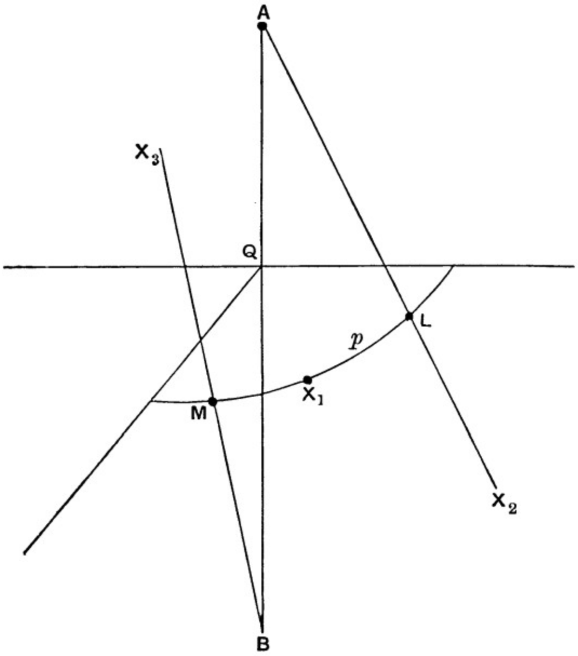

Axiom XVI secures the limitation to three dimensions, and the division of space by a plane. It can also be proved from the axioms that, if two planes intersect in at least one point, they intersect in a straight line.

7. A point will be said to divide a line into two half-rays which emanate from it.

A line will be said to divide a plane into two half-planes which are bounded by it.

A plane will be said to divide space into two half-spaces which are bounded by it.

A sheaf of lines is a complete set of coplanar lines concurrent at one point (the vertex). A sheaf of half-rays is a complete set of coplanar half-rays emanating from one point (the vertex).

A bundle of lines is a complete set of lines concurrent at one point (the vertex). A bundle of half-rays is a complete set of half-rays emanating from one point (the vertex).

If p, q, r are three half-rays belonging to a sheaf of half-rays, then r is said to 'lie between' p and q, if points A and B can be found on p and q respectively, such that the segment AB intersects r.

It can be proved that if r lies between p and q, then p does not lie between r and q.

The complete set of planes through a given line (the axis) is called a sheaf of planes. The axis divides each plane into two half-planes. These half-planes form a sheaf of half-planes.

If p, q, r are three half-planes belonging to a sheaf of half-planes, then r is said to 'lie between' p and q, if points A and B can be found on p and q respectively, such that the segment AB intersects r.

It can be proved that if r lies between p and q, then p does not lie between r and q.

The theorems indicated in this and in the preceding sections, and allied theorems, are not always very easy to prove. But their proofs depend so largely upon the particular mode of formulation of the axioms, that it would be outside the scope of this tract to enter into a consideration of them. In the sequel we shall assume that the whole class of theorems of the types, which have been thus generally indicated, can be proved from the axioms stated.

8. Formulations of the axioms of Descriptive Geometry have also been given by Hilbert*, and by E. H. Moore†, and by B. Russell‡, and by O. Veblen§. Veblen's memoir represents the final outcome of these successive labours, and his formulation will be given now. The axioms are stated in terms of 'points' and of a relation among three points called 'order.' Points and order are not defined.

I. There exist at least two distinct points.

---

* *Grundlagen der Geometrie*, Leipzig, 1899, English Translation by E. J. Townsend, Chicago, 1902.

† *On the Projective Axioms of Geometry*, Trans. of the Amer. Math. Soc., vol. III., 1902.

‡ *The Principles of Mathematics*, Cambridge, 1903, ch. XLVI.

§ *A System of Axioms for Geometry*, Trans. of the Amer. Math. Soc., vol. V., 1904.

II. If the points A, B, C are in the order ABC, they are in the order CBA.

III. If the points A, B, C are in the order ABC, they are not in the order BCA.

IV. If the points A, B, C are in the order ABC, then A is distinct from C.

V. If A and B are any two distinct points, there exists a point C such that A, B, C are in the order ABC.

Definition 1. The line AB $(A \neq B)$ consists of A and B, and of all points X in one of the possible orders ABX, AXB, XAB. The points X in the order AXB constitute the 'segment' AB. A and B are the 'end-points' of the segment, but are not included in it.

VI. If points C and D $(C \neq D)$ lie on the line AB, then A lies on the line CD.

VII. If there exist three distinct points, there exist three points A, B, C not in any of the orders ABC, BCA, or CAB.

Definition 2. Three distinct points not lying on the same line are the 'vertices' of a 'triangle' ABC, whose sides are the segments AB, BC, CA, and whose 'boundary' consists of its vertices and the points of its sides.

VIII. If three distinct points A, B, C do not lie on the same line, and D and E are two points in the orders BCD and CEA, then a point F exists in the order AFB and such that D, E, F lie on the same line.

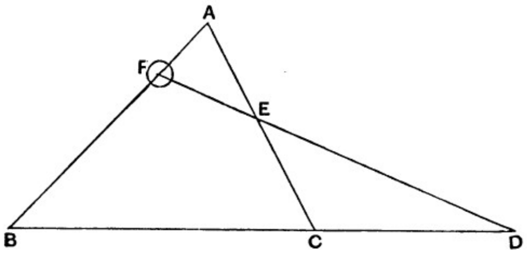

Definition 5. A point O is 'in the interior of' a triangle, if it lies on a segment, the end-points of which are points of different sides of a triangle. The set of such points O is 'the interior' of the triangle.

Definition 6. If A, B, C form a triangle, the 'plane' ABC consists of all points collinear with any two points of the sides of the triangle.

IX. If there exist three points not lying in the same line, there exists a plane *ABC* such that there is a point *D* not lying in the plane *ABC*.

Definition 7. If *A*, *B*, *C*, and *D* are four points not lying in the same plane, they form a 'tetrahedron' *ABCD*, whose 'faces' are the interiors of the triangles *ABC*, *BCD*, *CDA*, *DAB*, whose 'vertices' are the four points *A*, *B*, *C*, and *D*, and whose 'edges' are the segments *AB*, *BC*, *CD*, *DA*, *AC*, *BD*. The points of faces, edges, and vertices constitute the 'surface' of the tetrahedron.

Definition 8. If *A*, *B*, *C*, *D* are the vertices of a tetrahedron, the space *ABCD* consists of all points collinear with any two points of the faces of the tetrahedron.

X. If there exist four points, neither lying in the same line, nor lying in the same plane, there exists a space *ABCD*, such that there is no point *E* not collinear with two points of the space *ABCD*.

The above axioms of Veblen are equivalent to the axioms of Peano which have been previously given. Both Peano and Veblen give an axiom securing the Dedekind property (cf. § 9). Also Veblen gives an axiom securing the 'Euclidean' property (cf. § 10).

9. Dedekind's original formulation* of his famous property applies directly to the case of a descriptive line and is as follows:

"If all points of the straight line fall into two classes such that every point of the first class lies to the left of every point of the second class, then there exists one and only one point which produces this division of all points into two classes, this severing of the straight line into two portions."

It is of course to be understood that the dividing point itself belongs to one of the two classes.

It follows immediately that the boundary of a triangle consists of points in a compact closed order possessing the Dedekind property as already formulated for closed series†.

This definition may be repeated here to exhibit its essential independence of the special definition of projective segments upon which the previous formulation rests.

Let *A*, *B*, *C* be any three points of a closed series. Then by

---

* Cf. his *Continuity and Irrational Numbers*, ch. III.; the quotation here is from Beman's translation, Chicago, 1901.

† Cf. Proj. Geom. § 19 (*a*).

hypothesis the series is such that there are two ways round from A to C, namely, one through B and one not through B. Let segm (ABC) denote the points, excluding A and C, which are traversed from A to C through B, and let segm $(A\hat{B}C)$ denote the remaining points of the series. Again let a class u of the points of the series be called a segment of the series, when (1) there is a point B of the series which does not belong to u, and (2) if P and Q be any two points of u then segm $(P\hat{B}Q)$ belongs entirely to u.

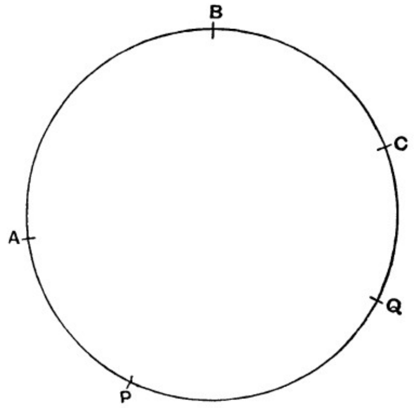

Then the series possesses the Dedekind property if any segment such as u (which excludes more than one point of the series) must possess two boundary points, that is to say, if there must exist points A and C such that segm $(A\hat{B}C)$, with the possible exception of either or both of A and C, is identical with u. Here—as above—B is a point which does not belong to u.

Hence a sheaf of half-rays can also be considered as a closed compact series with the Dedekind property. This is made immediately evident by surrounding the vertex by a triangle in the plane of the sheaf. Then each half-ray of the sheaf intersects the boundary of the triangle in one and only one point. Also the order of the points on the boundary is the order of the corresponding half-rays of the sheaf. But the boundary of the triangle is a closed series with the Dedekind property.

10. By the aid of the Dedekind axiom and of the preceding axioms, it can now be proved that, if l be any line and A be any point, not incident in l, then in the plane Al at least one line can be drawn through A, which does not intersect l.

For take any point B on l, and let p and q be the two supplementary half-rays of l which emanate from B. Consider the sheaf of half-rays, vertex A, in the plane Al. Some of these half-rays intersect p and some intersect q, and these classes are mutually exclusive.

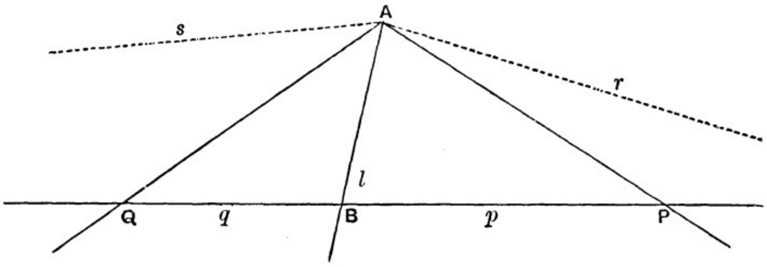

Also, from the Dedekind property, there exist two semi-rays which are limits of the semi-rays intersecting p. AB is one of the semi-rays, let r be the other. Now the semi-ray p has no end-point. Hence r is not among the semi-rays intersecting p. Again by similar reasoning there is a semi-ray s which is the limit of the semi-rays intersecting q, and s does not intersect q.

Now first let r and s be not collinear, and let r' and s' be the half-rays supplementary to r and s respectively. Consider the set (α, say) of lines through A with one set of their half-rays between r and s', and therefore with their supplementary half-rays between r' and s. There are an infinite number of such lines. Now all half-rays emanating from A and lying between the half-ray AB and r intersect p, and no other half-rays from A intersect p. Similarly for the half-rays AB and s and q. Also if s' lie between the half-ray AB and r, then r' lies between the half-ray AB and s; and in this case every line of the set α intersects the line l twice, namely once for each of its pair of supplementary half-rays emanating from A. But this is impossible. Hence neither the supplement of r nor that of s can intersect l. Secondly if r and s are collinear, then the complete line formed by r and s cannot intersect l. For neither r nor s intersects l.

Thus taking any point A and any line l, the sheaf of lines, vertex A and in the plane Al, falls into two parts, namely the lines which intersect l, called the lines 'secant' to l, and the lines which do not intersect l, called the lines 'non-secant' to l. The non-secant lines of the sheaf may reduce to one line. The supposition that this is the case is the 'Euclidean Axiom.'

CHAPTER II.

THE ASSOCIATED PROJECTIVE SPACE.

11. We have now to establish the relation of Descriptive Geometry to an associated Projective Geometry. In a Projective Space let a 'convex region' be defined to be a region which (1) does not include the whole of any line, and (2) includes the whole of one of the two segments between any two points within it. It is easy to prove that such regions exist. For remembering* that we can employ the ordinary theory of homogeneous coordinates, the surface

$$x^2 + y^2 + z^2 - u^2 = 0$$

is well known to enclose such a region. Let a quadric enclosing a convex region be called a 'convex quadric.' Again in two dimensions let A, B, C be any three non-collinear points, and let P be any point not collinear with any two of them. Let AP meet the line BC in L, BP meet CA in M, CP meet AB in N.

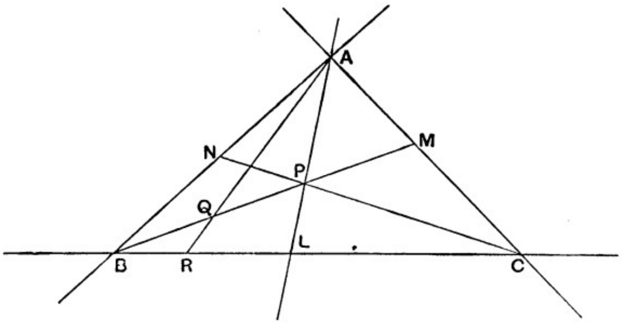

Define the triangular region (ABC/P) to be the set of points formed by the collection of segments such as segm (AQR)†, where Q is any point on segm (BPM), and R is the point where the line

---

* Cf. Proj. Geom. §§ 37 and 42.

† Cf. Proj. Geom. § 13.

AQ intersects BC. The points A, B, C can be interchanged in this definition without altering the region obtained.

Similarly in three dimensions, if A, B, C, D are the vertices of a non-degenerate tetrahedron, and P be any point not on any of the planes, ABC, BCD, etc., the 'tetrahedral region' $(ABCD/P)$ can be similarly defined. From the ordinary theory of homogeneous coordinates, it is well known that a triangular region in two dimensions, and a tetrahedral region in three dimensions, are both convex regions.

Again the triangular region (ABC/P) considered above has as its 'boundary' the segments (BLC), (CMA), (ANB), together with the points A, B, C. Also considering the tetrahedral region $(ABCD/P)$, let AP intersect BCD in L, BP intersect CDA in M, CP intersect DAB in N, DP intersect ABC in O. Then the 'boundary' of the tetrahedral region $(ABCD/P)$ consists of the triangular regions (BCD/L), (CDA/M), (DAB/N), (ABC/O), together with the boundaries of these triangular regions.

It is now a well-known result from the use of coordinates that in two dimensions any line through a point in a triangular region cuts the boundary in two points only; and that in three dimensions any line through a point in a tetrahedral region cuts the boundary in two points only.

12. Now consider a convex region, let it be either the region within a convex quadric, or a tetrahedral region. Call the points within it 'Descriptive points'; and call the portions of lines within it 'Descriptive lines.' The projective order of points on a line becomes an open order for Descriptive points on Descriptive lines. Then by the use of coordinate Geometry it is easy to prove that all the Descriptive axioms of the present tract, either in Peano's form or in

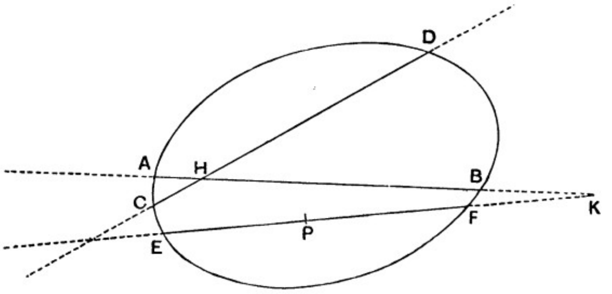

Veblen's form, are satisfied, including the Dedekind axiom, but excluding the Euclidean axiom. Thus in the figure the lines AB and CD

intersect at a point H in Descriptive Space; but the lines AB and EF do not intersect in Descriptive Space, since K lies outside it. Also it is evident that through any point P an infinite number of lines can be drawn, coplanar with AB, and not intersecting it in Descriptive Space.

13. The previous article (§ 12) proves* the existence theorem for Descriptive Space with the negation of the Euclidean axiom; in other words, it proves the independence of the Euclidean axiom.

The existence theorem for Descriptive Space with the Euclidean axiom is immediately proved by considering the region of Projective Space found by excluding all the points on one projective plane. The region is convex according to the above definition; also all the Descriptive axioms, together with the Dedekind axiom and the Euclidean axiom, hold for it†.

14. The independence of the Dedekind axiom of the other axioms, combined with the negation of the Euclidean axiom, is proved by considering, as in § 12, Descriptive Space to be a tetrahedral region in Projective Space, but confining ourselves to the points whose coordinates are algebraic numbers‡, as in the corresponding proof for Projective Geometry.

The independence of the Dedekind axiom of the other axioms, combined with the Euclidean axiom, is similarly proved by considering Descriptive Space to be the region in Projective Space found by excluding a particular plane; and further, as before, we confine our consideration to the points whose coordinates are algebraic numbers.

---

* Cf. Proj. Geom. § 43.

† In the later Greek period, and during the seventeenth and eighteenth centuries, the discussion of the foundations of Geometry was almost entirely confined to attempts to prove the Euclidean axiom. The explicit recognitions of its independence by Lobatschefskij (1828), and by J. Bolyai (1832) laid the foundation of the existing theories of non-Euclidean Geometry. For the literature of the whole question cf. Stäckel and Engel, *Die Theorie der Parallellinien von Euklid bis auf Gauss*, Leipzig, 1895, and also their *Urkunden zur Geschichte der Nichteuklidischen Geometrie, I. Lobatschefskij*, Leipzig, 1898.

‡ Cf. Proj. Geom. § 43 (α). An oversight in this proof may be here corrected. The proof, as printed, proceeds by considering only points with rational coordinates. But then a difficulty arises as to the theory of segments given in Chapter IV. of Proj. Geom. For it is necessary that the real double points of a hyperbolic involution should belong to the points considered. But these double points are given by a quadratic equation. Thus algebraic numbers (*i.e.* numbers which can occur as the roots of equations with integral coefficients) should be substituted for rational numbers. The proof proceeds without other alteration. I am indebted to Mr G. G. Berry of the Bodleian Library for this correction.

15. It has been proved in §§ 12 and 13 that a convex region of a Projective Space is a Descriptive Space. The converse problem has now to be considered in this and in the next chapter; namely, given a Descriptive Space, to construct a Projective Space of which the Descriptive Space is part. This effects a very considerable simplification in the investigation of the properties of Descriptive Space owing to the superior generality of the analogous properties of Projective Space. Thus a Projective Space affords a complete interpretation of all the entities indicated in coordinate geometry. It is in order to gain this simplification that the 'plane at infinity' is introduced into ordinary Euclidean Geometry. We have in effect to seek the logical justification for this procedure by indicating the exact nature of the entities which are vaguely defined as the 'points at infinity'; and the procedure is extended by shewing that it is not necessarily connected with the assumption of the Euclidean axiom. This investigation is the Theory of Ideal Points*, or of the generation of 'Proper and Improper Projective Points' in Descriptive Geometry. The Euclidean axiom will not be assumed except when it is explicitly introduced. The remainder of this chapter will be occupied with the general theorems which are required for the investigation.

16. If A be any point and l be any line not containing A, then the plane Al divides the bundle of half-rays emanating from A into three sets, (1) the half-rays in the plane Al, (2) the half-rays on one side of the plane, (3) the half-rays on the other side of the plane. The sets (2) and (3) are formed of half-rays supplementary one to the other.

Lemma. With the above notation, it is possible to find a plane through the line l and intersecting any finite number of the half-rays either of set (2) or of set (3).

For let $a_1, \dots a_n$ be n half-rays of one of the two sets. Let B_1 be any point on a_1, and B_2 be any point on a_2. Then either the plane

---

* Originally suggested by Klein (extending an earlier suggestion of von Staudt), *Math. Annal.* vols. IV. and VI., 1871 and 1872; first worked out in detail by Pasch, *loc. cit.*, §§ 6—9. In the text I have followed very closely a simplification of the argument given by R. Bonola, *Sulla Introduzione degli Enti Improprii in Geometria Projectiva*, Giornale di Matematiche, vol. XXXVIII., 1900.

B_1l lies between the planes B_2l and Al, or the plane B_2l lies between the planes B_1l and Al, or the planes B_1l and B_2l are identical. But in either of the first two cases the intermediate plane intersects both semi-rays a_1 and a_2. Hence a plane is found through l, intersecting both a_1 and a_2. Call it the plane $B_2'l$. Again take any point B_3 on a_3; and the same argument shews that at least one of $B_2'l$ and B_3l intersects a_1, a_2, and a_3. Proceeding in this way, a plane is finally found which intersects each of the n semi-rays.

17. Desargues' Perspective theorems* can be enunciated in the following modified forms:

(1) If two coplanar triangles ABC and $A'B'C'$ are such that the lines AA', BB', CC' are concurrent in a point O, then the three intersections of BC and $B'C'$, CA and $C'A'$, AB and $A'B'$, if they exist, are collinear.

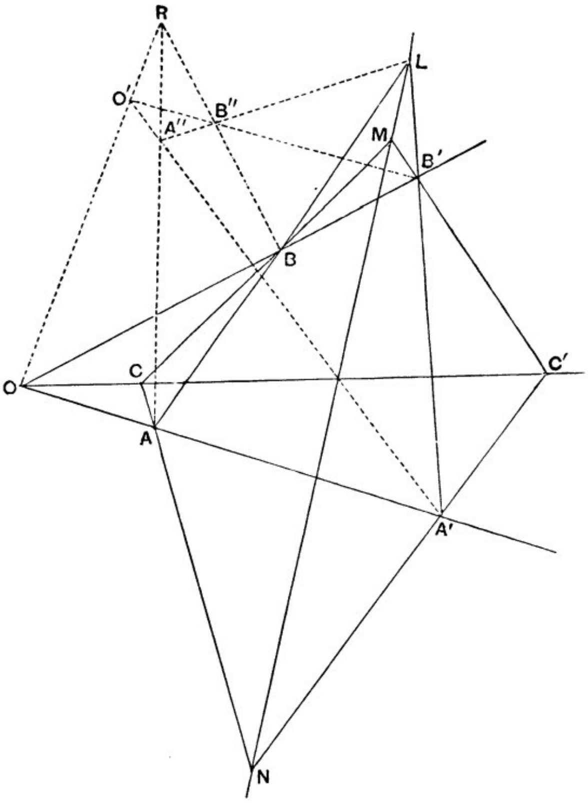

---

* Cf. Proj. Geom. § 7.

(2) If the pairs of homologous sides of the two coplanar triangles ABC and $A'B'C'$, namely, BC and $B'C'$, CA and $C'A'$, AB and $A'B'$, intersect in three collinear points, then the lines AA', BB', CC', if any two intersect, are concurrent in the same point.

Considering the first proposition let AB and $A'B'$ intersect in L, BC and $B'C'$ in M, CA and $C'A'$ in N. Now it is not possible both for A' to lie on the segment OA, and for A to lie on the segment OA'. Assume that A' does not lie on the segment OA. Let R be any point external to the given plane (α, say). Now by the lemma of § 16, it is possible to find a plane through LM, lying between the planes LMR and LMA (*i.e.* the plane α), and intersecting the three lines RA, RB, RC, say, in the points A'', B'', C'' (in the figure C'' is not shewn). Then evidently A'' must lie in the segment RA. Hence $A'A''$, since A' does not lie in the segment OA, must intersect OR in the segment OR. Thus the intersection of the lines $A'A''$ and OR is secured. Let it be the point O'. Again the lines $O'A'$ and $O'B'$ are the projections from R on the plane $A'O'B'$ of the lines OA' and OB'. Now $A''B''$ passes through L. Hence B'' lies on the plane $A'A''B'$, *i.e.* on the plane $OA'B'$. Hence B'' is on the projection of the line OB' on the plane $O'A'B'$, *i.e.* B'' lies on $O'B'$. Thus $B'B''$ passes through O'.

Reasoning in exactly the same way for BC and $B'C'$, it follows that $C'C''$ passes through O'. The same figure has now been constructed as in the proof of the corresponding theorem for Projective Geometry*. Accordingly the theorem follows by the same reasoning.

In order to demonstrate the converse theorem, we proceed exactly as above, except that, L, M, N are now assumed to be collinear, O is the point of intersection of AA' and BB'. Then the same construction is made as before, and it is successively proved by similar reasoning that every pair of the lines $A'A''$, $B'B''$, $C'C''$ intersect. But the lines are not coplanar. Hence they intersect in the same point O'. But O' must lie on RO. Thus CC' passes through O.

Corollary. The enunciation of the first theorem can be modified by removing the assumption that AC and $A'C'$ intersect, but by adding the assumption that AC intersects LM.

18. A trihedron is the figure formed by three lines concurrent in the same point, and not all coplanar. The three lines form the edges of the trihedron; the three planes containing the lines, two by two, form the faces of the trihedron; the point of concurrence of the three edges is the vertex of the trihedron.

---

* Cf. Proj. Geom. § 7.

It follows (cf. § 6) that, if two trihedrons have the same vertex, any two faces, one from each trihedron, must intersect in a line through the vertex; also that any two planes each containing two edges, one edge from each trihedron, must intersect in a line through the vertex.

Desargues' theorems can be applied to two trihedrons with the same vertex; only in this case, as in Projective Geometry, there are no exceptional cases depending on non-intersection.

The enunciations are as follows:

(1) If a, b, c and a', b', c' are the edges of two trihedrons with the same vertex, such that the planes containing a and a', b and b', c and c', are concurrent in a line s (*i.e.* belong to the same sheaf), then the three intersections of the planes bc and $b'c'$, ca and $c'a'$, ab and $a'b'$ are coplanar.

(2) If a, b, c and a', b', c' are the edges of two trihedrons with the same vertex, such that the three intersections of the planes bc and $b'c'$, ca and $c'a'$, ab and $a'b'$, are coplanar, then the three planes containing a and a', b and b', c and c' belong to the same sheaf.

These propositions immediately follow from the case of triangles by noticing that, by the lemma of § 16, the six edges of the trihedrons can be cut by a plane, not through the vertex. Hence by the previous remarks on trihedrons, Desarguesian triangles are obtained without the exceptions due to non-intersection.

19. The two theorems of the present and next articles are the central theorems of the whole theory of Ideal Points.

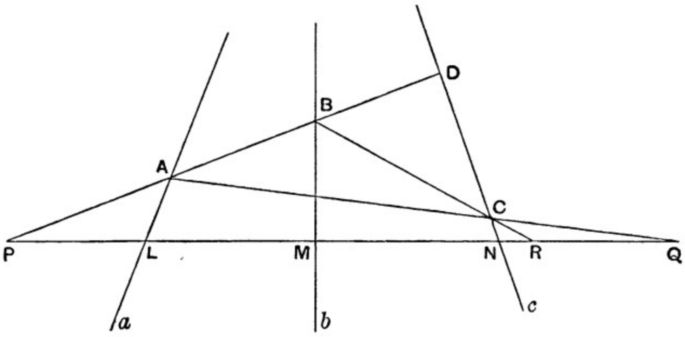

If the lines a, b, c are the intersections of three planes α, β, γ of a sheaf with a plane π, not belonging to the sheaf, and if O be any point not incident in π, then the three planes Oa, Ob, Oc belong to one sheaf.

If the axis (r) of the sheaf intersect the plane π in a point S, then the three lines a, b, c pass through S, and the line OS is evidently contained in Oa, Ob, Oc, and thus forms the axis of the new sheaf.

Consider now the case when the axis (r) of the sheaf does not intersect π. Then a, b, c are not concurrent, and no two of them intersect. Hence one of the three (b, say) must lie between [*i.e.* any segment, joining a point on a and a point on c, intersects b] the other two. Take any two points L and N on a and c respectively, then the segment LN intersects b in a point M. Take two other points P and Q on LN so that we have the order P, L, M, N, Q. Take any point D on c; then the segment PD must intersect a and b in two points A and B respectively; and the segment AQ must intersect the segment DN in a point C. Then the line BC must intersect the segment NQ in a point R. Thus a triangle ABC has been formed, whose vertices lie on a, b, c, and whose sides AB, AC, BC pass through P, Q, R respectively.

By taking another point D' on c, another triangle $A'B'C'$ can be similarly formed, whose vertices lie on a, b, c, and whose sides $A'B'$ and $A'C'$ pass through P and Q respectively.

We have first to shew that $B'C'$ passes through R. For taking any point T on the axis (r) of the sheaf, the lines TA, TB, TC form the edges of one trihedron, and the lines TA', TB', TC' form the edges of another trihedron with the same vertex.

Also the planes TAA', TBB', TCC' belong to the same sheaf. Hence the three intersections of the pairs of planes TAB and $TA'B'$, TAC and $TA'C'$, TBC and $TB'C'$ are coplanar; hence they lie in the plane TPQ. Hence $B'C'$ passes through R.

Now considering the two trihedrons with edges OA, OB, OC, and OA', OB', OC', the intersections of the pairs of faces OAB and $OA'B'$, OAC and $OA'C'$, OBC and $OB'C'$, are respectively OP, OQ, OR; and these are coplanar. Hence by the converse part of Desargues' theorem for trihedrons, the planes OAA', OBB', OCC' belong to the same sheaf. Hence Oa, Ob, Oc belong to the same sheaf (*i.e.* have a common line of intersection).

20. If any two of the lines a, b, c are coplanar, but the three lines are not coplanar, and similarly for the lines a, b, d, then c and d are coplanar.

If a and b intersect, the theorem is evident; for a, b, c are concurrent, and a, b, d are concurrent. Hence c and d are concurrent.

Assume that a and b do not intersect. Then it is easy to prove that no two of the lines intersect. It follows that no one of the lines c, b, d can intersect any of the planes ab, ac, ad in which it does not lie.

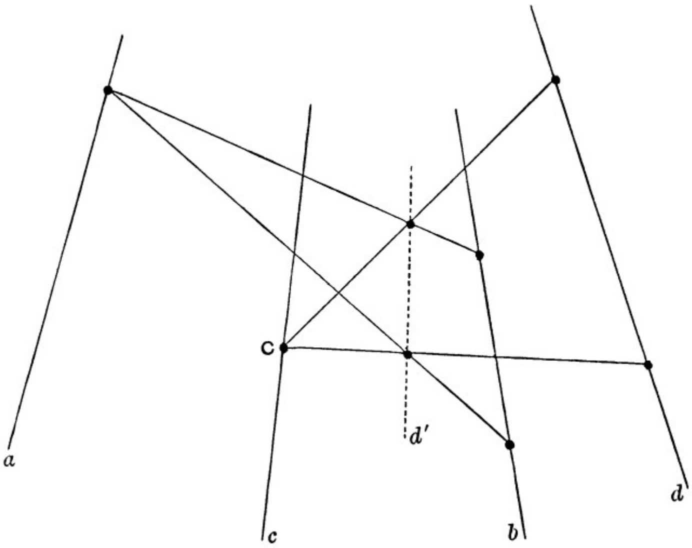

Fig. 1.

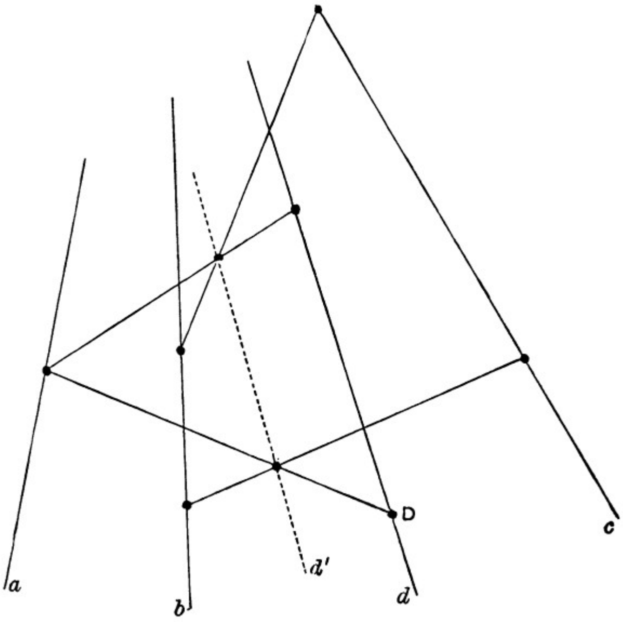

Fig. 2.

Hence it follows that either *c* and *d* lie on opposite sides of the plane *ab*, or *d* and *b* on opposite sides of the plane *ac*, or *b* and *c* on opposite sides of the plane *ad*.

First, let *c* and *d* lie on opposite sides of the plane *ab* (cf. fig. 1). Take any point *C* on *c*. Then the plane *Cd* must intersect the plane *ab* in a line, *d'*, say. Then the lines *a*, *b*, *d'* are the intersections of the three planes *da*, *db*, *dC* with the plane *ab*; and these three planes belong to the same sheaf. Hence (cf. § 19) the three planes through the lines *a*, *b*, *d'* respectively and through any point not on *ab* belong to the same sheaf. But *C* is such a point. Hence the planes *Ca*, *Cb*, *Cd'* belong to the same sheaf. But *c* is the common line of *Ca* and *Cb*. Hence *Cd'* contains the line *c*. Hence *c* and *d* are coplanar.

Secondly, let the plane *ad* lie between *b* and *c*. Then the plane *bc* must intersect the plane *ad* in some line, *d'* say. Thus the three lines *b*, *d'*, *c* are the intersections of the three planes *ab*, *ad*, *ac* with the plane *bc*. These three planes belong to the same sheaf. Hence (cf. § 19), if *D* is any point on *d*, not on *bc*, the planes *Db*, *Dc*, *Dd'* belong to the same sheaf. But *Db* and *Dd'* intersect in the line *d*; hence *Dc* passes through the line *d*. Thus *c* and *d* are coplanar.

Thirdly, let the plane *ac* lie between *b* and *d*. Then the proof is as in the second case, interchanging *c* and *d*.

data 06.04.2020

of any regular polygon
(in this case, a triangle)

Figure 4.7

To construct the circumscribed circle of any regular polygon (triangle)

Figure 4.8

To construct the escribed circle to any regular polygon (triangle)

Figure 4.9

To construct a circle that passes through a fixed point A and touches a line at a given point B.

... Figure 4.20

Pag. 55(72).

C) Books_65_2020/
Descriptive Geometry/
A. N. Whitehead - the Axioms of Descriptive Geometry....pdf

Pag. 4(13)

Peano uses the following useful notation which is an extension of his notation for segments and prolongations.

data 06.04.2020

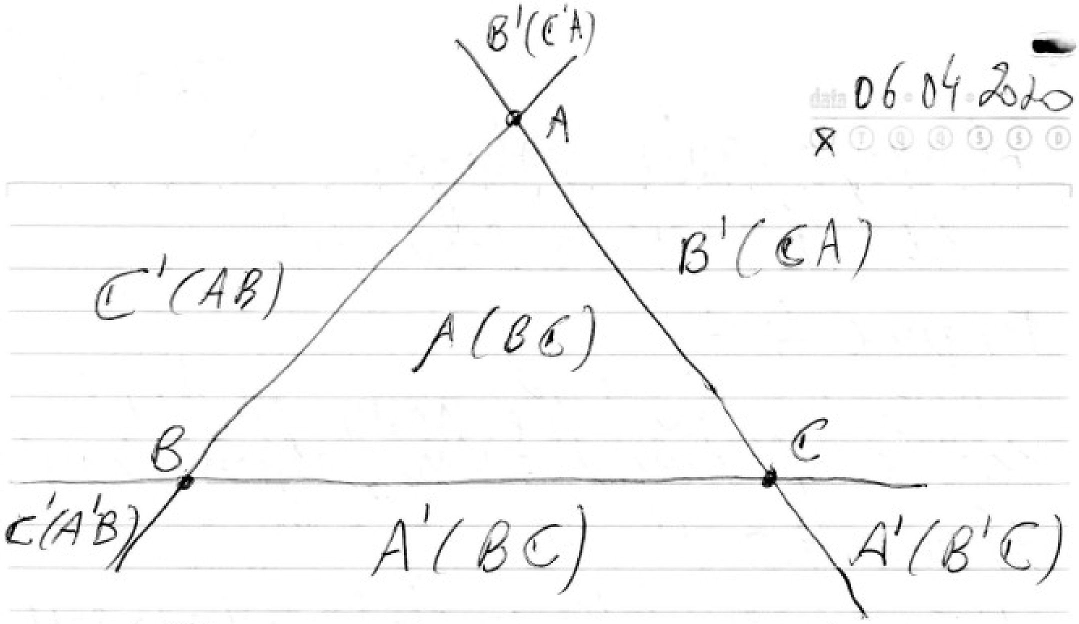

Seven regions into which a plane is divided by three lines are as in the figure

Pag. 18(27) Sheaves of Planes

Pag. 18(27) → 19

The two theorems of the present and next articles are the central theorems of the whole of Ideal Points

Pag. 19 Coplanarity of Lines

128) Graphic Science:

Thomas Ewing French

Source:

https://archive.org/details/graphicscienceen00fren

' 5
+,- * *. * 5 8 / " <
" / 2 :< / /"

' #
+,- * *. * # M 3 : 3 D
" D D ' > 9 L3I; D" /"
(%)) 9) 5

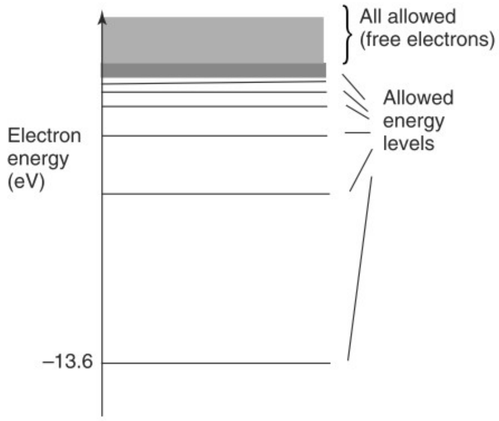
All allowed
(free electrons)
Allowed
energy
levels
Electron
energy
(eV)
–13.6

9) !O D : 2 " :"

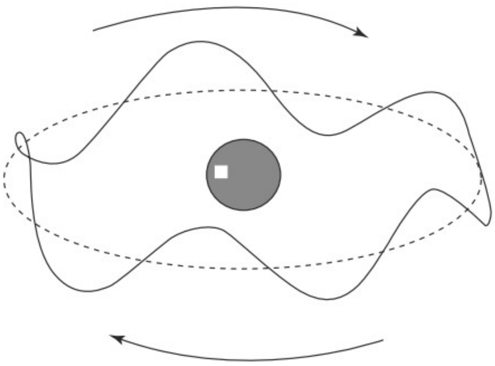

9) 5O : 2
"
6 2
: /

Growth of semiconductor crystals

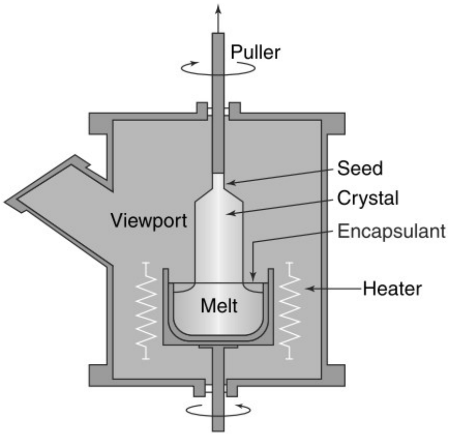

Figure 4-6 Experimental setup of Czochralski growth of semiconductor crystals

The fabrication of semiconductor devices requires highly pure and mono-crystalline semiconductors. The Czochralski method, invented by the Polish scientist Jan Czochralski in 1916, is the preferred method for high-volume production of high quality Si single crystals. The silicon is pulled from a bath of molten silicon to form mono-crystalline silicon using a small seed crystal, as shown in Fig. 4-6. The semiconductor crystals can be obtained by cooling the molten semiconductor material. Incidentally, this procedure yields poly-crystalline material since crystals start growing in different locations with a different orientation. Single-crystal silicon can be grown by dipping one end of a seed crystal in the melt and by controlling the temperature difference between the seed crystal and the molten silicon. The output is large single-crystal silicon, cylindrical in shape. The diameter of the silicon will depend on the rotation of the seed crystal during the stages of its growth and partly on the cylindrical shape of the crucible containing the melt.

The cylindrical silicon is then cut into smaller wafers with a diamond saw and again polished to yield the substrate material for the silicon device fabrication.

4-2-2 Mechanical Properties

In the presence of stress, the tensorial nature of Hooke's law becomes prominent. Only the second- and third-order elastic constants are being affected in the presence of stress. In the *n*-type germanium, the shear elastic constant is being reduced by 8 per cent. In view of large changes of the elastic constants with the carrier density, the experiments on the velocity of sound involving shear mode as a function of carrier concentration exhibit the contribution of the electrons to the mechanical properties.

4-2-3 Energy Band Theory

In the energy band theory, we are only concerned with the conduction band and the valance band separated by a band gap. By using this theory, we can differentiate among insulators, metals and semiconductors. Another attribute of this theory is the formulation of the carrier energy spectrum, which in turn controls all the physical properties of semiconductors and semi metals in general.

Energy levels in atoms

The sum of the kinetic and potential energies of an electron orbiting a nucleus can only assume a set of discrete values; those for the simplest atom (hydrogen) are shown in Fig. 4-7. This happens because the electrons exhibit wave-like properties.

To be stable, the electron wave must repeat itself exactly at the same time when it is on a circular path around the nucleus. This is illustrated schematically in Fig. 4-8. It leads to the conclusion that in such a state an integral number of wavelengths must fit into its path. This state is termed *quantization condition*.

The electron's wavelength (λ) is inversely proportional to its momentum (p) as defined by de Broglie's relation of the wave-particle duality:

$$\lambda = \frac{h}{p} \tag{4-1}$$

The constant h is called Planck constant, and has the value 6.6×10^{-34} Js. The graph of Eq. (4-1) is rectangular hyperbolic in $p - \lambda$ plane. The quantization condition, thus, fixes the electron momentum to one of a set of discrete values, depending on the number of wavelengths fitting the circular path. The number, known as a *quantum number*, determines which set of discrete energies the electron has, since the energy is directly related to the electron's momentum. Four quantum numbers are needed to describe the electron in the hydrogen atom.

It should be noted that an electron freed from its parent nucleus can have any wavelength, and hence any kinetic energy. This explains the continuum of energy levels above the zero of energy, given in Fig. 4-7. When an atom contains many electrons, no more than two are found to exist simultaneously in the same state of nature. One of these electrons spins clockwise, the other anticlockwise. So the above rule can be expressed by stating, "In a closed system, no two electrons can occupy the same state." This is known as the Pauli exclusion principle.

In the simplest cases, when each state has a different energy we find electrons filling energy levels from the lowest up, one to each state. In the more general cases, it is possible for two or more orbits, or states, to have the same energy. Thus, a limited number of electrons are in the highest filled atomic energy levels, and these are the valence electrons, which take part in bonding an atom within a molecule or solid.

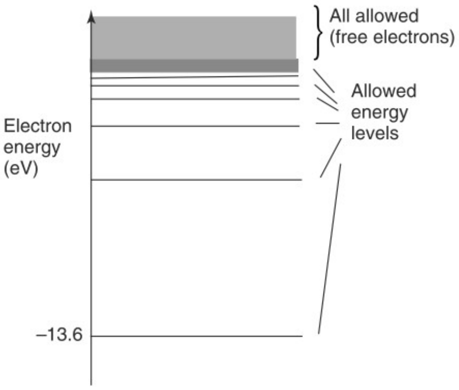

Figure 4-7 Energy level diagram for a hydrogen atom

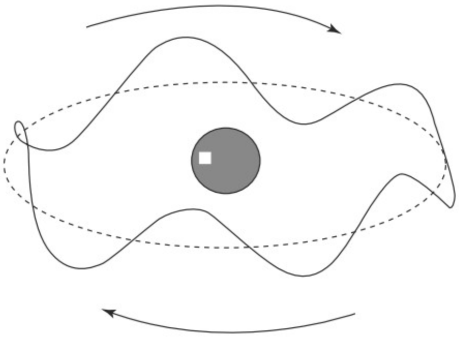

Figure 4-8 Continuity of the electron wave around the nucleus resulting in an integral number of wavelengths fitting into any circular path

Solved Examples

Example 4-1 When a photon impinges upon a semiconductor it can take an electron from the valence band to the conduction band. The momentum is conserved in such transitions. Calculate the value of the wave vector that a 2.0 eV photon carries. The electron in the valence band can go into the conduction band with the momentum change of the photon.

6 *+,-* * *. * # M 3 : < , 6 :<
+ 6 M 3 D " D D > 9 ,8 6B " 3D)
D" /"
(5 5% & 0 J "

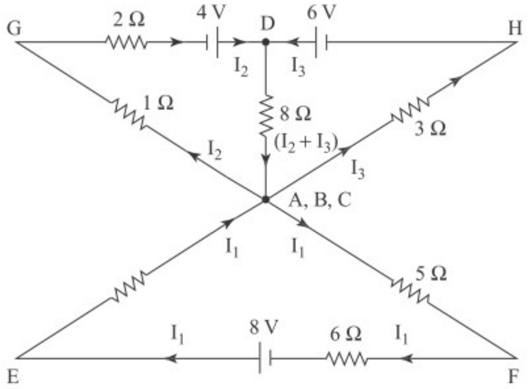

Figure 2.25

In the loop EA FE, current I_1 will flow. No current from this loop will flow to the other two loops. Current flowing from E to A is to be the same as the current flowing from A to F.

The distribution of currents in loop GDAG and HDAH have been shown. By applying KVL in these loops we write:

for loop GDAG

$$-2\,I_2 + 4 - 8(I_2 + I_3) - I_2 = 0$$

or, $$11\,I_2 + 8\,I_3 = 4 \qquad \text{(i)}$$

for loop HDAH

$$6 - 8\,(I_2 + I_3) - 3\,I_3 = 0 \qquad \text{(ii)}$$

or, $$8I_2 + 11\,I_3 = 6$$

Solving eqs. (i) and (ii)

$$I_3 = 0.6 \text{ A and } I_2 = -0.07 \text{ A}$$

Current through the 8 Ω resistor $= I_2 + I_3$

$$= -0.07 + 0.6$$
$$= 0.53 \text{ A}$$

So far, we have assumed branch currents in a network and applying KVL written the voltage equations. From the loop equations, we have calculated the branch currents. Two other methods, namely Maxwell's mesh current method and Node voltage method, are described in the following sections.

2.6 MAXWELL'S MESH CURRENT METHOD

A mesh is a smallest loop in a network. KVL is applied to each mesh in terms of mesh currents instead of branch currents. As a convention, mesh currents are assumed to be flowing in the clockwise direction without branching out at the junctions. Applying KVL, the voltage equations are framed. By knowing the mesh currents, the branch currents can be determined. The procedure followed is explained through an example. Let us calculate the current flowing through the branches in the circuit given in Fig. 2.26.

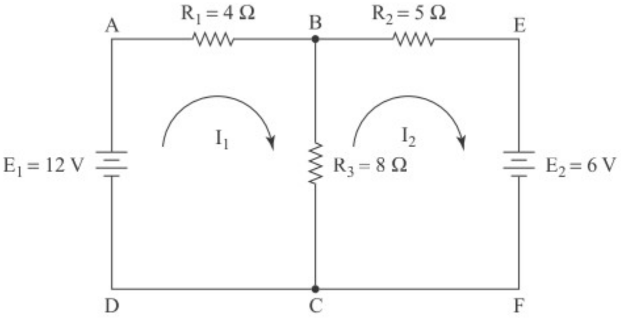

Figure 2.26

We have assumed loop currents I_1 and I_2 flowing in the clockwise direction as shown.

It may be noted that current flowing through the resistor R_3 is the algebraic sum of the two currents I_1 and I_2. Here I_1 is flowing in the downward direction while I_2 is flowing in the upward direction.

We will now write the voltage equations for the two loops applying KVL and then solve the equations. If the value of any mesh currents is calculated as negative, we will take the direction of that mesh current opposite to the assumed clockwise direction.

For loop DABCD, the voltage equation is

$$12 - 4\,I_1 - 8(I_1 - I_2) = 0$$

or,
$$3\,I_1 - 2I_2 = 3 \tag{i}$$

For loop BEFCB, the voltage equation is

$$-5\,I_2 - 6 - (I_2 - I_1)\,8 = 0$$

or,
$$8\,I_1 - 13\,I_2 = 6 \tag{ii}$$

solving eqs. (i) and (ii), we get

$$I_1 = 1.17\text{ A},\ I_2 = 0.26\text{ A}$$

and current flowing through R_3 is $(I_1 - I_2) = 0.91$ A

I_1 is flowing through R_1, I_2 is flowing through R_2 and $(I_1 - I_2)$ is flowing through R_3.

Example 2.8 Using the mesh current method calculate the current flowing through the resistors in the circuit shown in Fig. 2.27.

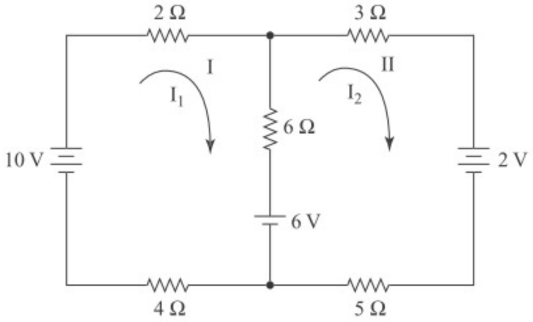

Figure 2.27

Solution:

Applying KVL in mesh I,

$$10 - 2\,I_1 - 6\,(I_1 - I_2) - 6 - 4\,I_1 = 0$$

or,
$$6\,I_1 + 3\,I_2 = 2 \qquad \text{(i)}$$

Applying KVL in mesh II,

$$-3I_2 - 2 - 5I_2 + 6 + 6\,(I_1 - I_2) = 0$$

or,
$$-6I_1 + 14\,I_2 = 4 \qquad \text{(ii)}$$

Adding eqs. (i) and (ii)

$$11\,I_2 = 6$$

or,
$$I_2 = \frac{6}{11}\,\text{A}$$

and
$$I_1 = \frac{2+3I_2}{6} = \frac{2+3\times 6/11}{6} = \frac{20}{33}\,\text{A}$$

Current through the 6 Ω resistor is $(I_1 - I_2)$ which is equal to $\frac{2}{33}$ A.

Example 2.9 A network with three meshes has been shown in Fig. 2.28. Applying Maxwell's mesh current method determine the value of the unknown voltage, V for which the mesh current, I_1 will be zero.

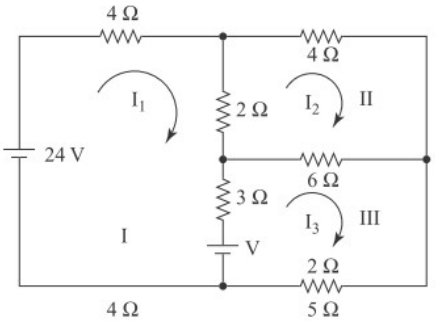

Figure 2.28

Solution:

Applying KVL in mesh I, II, and III respectively, we get

$$-4\,I_1 - 2\left(I_1 - I_2\right) - 3\left(I_1 - I_3\right) - V + 24 = 0$$

or,
$$9\,I_1 - 2\,I_2 - 3\,I_3 = 24 - V \qquad \text{(i)}$$

$$-4\,I_2 - 6\left(I_2 - I_3\right) - 2\left(I_2 - I_1\right) = 0$$

or,
$$I_1 - 6\,I_2 + 3\,I_3 = 0 \qquad \text{(ii)}$$

$$-6\left(I_3 - I_2\right) - 2\,I_3 + V - 3\left(I_3 - I_1\right) = 0$$

or,
$$3I_1 + 6\,I_2 - 11\,I_3 = -V \qquad \text{(iii)}$$

From (i), (ii), and (iii) the determinants Δ and Δ_a or Δ_1 are

$$\Delta = \begin{vmatrix} 9 & -2 & -3 \\ 1 & -6 & 3 \\ 3 & 6 & -11 \end{vmatrix}$$

$$\Delta_1 = \begin{vmatrix} 24-v & -2 & -3 \\ 0 & -6 & 3 \\ -v & 6 & -11 \end{vmatrix}$$

According to Cramer's rule,

$$I_1 = \frac{\Delta_1}{\Delta}, \quad I_2 = \frac{\Delta_2}{\Delta} \quad I_3 = \frac{\Delta_3}{\Delta}$$

Here condition is that I_1 must be zero.

$$I_1 = \frac{\Delta_1}{\Delta} = 0. \text{ So, } \Delta_1 \text{ must be zero}$$

Thus, we equate Δ_1 to zero.

$$\begin{vmatrix} 24-v & -2 & -3 \\ 0 & -6 & 3 \\ -v & 6 & -11 \end{vmatrix} = 0$$

or, $$(24-v)\begin{vmatrix} -6 & 3 \\ 6 & -11 \end{vmatrix} - (-2)\begin{vmatrix} 0 & 3 \\ -v & -11 \end{vmatrix} - 3\begin{vmatrix} 0 & -6 \\ -v & 6 \end{vmatrix} = 0$$

$$(24 - V)\,[66 - 18] + 2\,[0 - (-3V)] - 3[0 - 6V] = 0$$

or, $$(24 - V)\,48 + 6V + 18V = 0$$

or, $$24 \times 48 - 48V + 24V = 0$$

or, $$24V = 24 \times 48$$

or, $$V = 48 \text{ Volts}$$

2.7 NODAL VOLTAGE METHOD (NODAL ANALYSIS)

In the nodal analysis method a reference node in the network is chosen. Then the unknown voltages at the other nodes are determined with respect to the reference node. After determining the node voltages, currents in all branches can be calculated. This method of circuit analysis is suitable where a network has a number of loops, and hence a large number of simultaneous equations are to be solved. The procedure for the node voltage method is explained through an example.

Example 2.10 For the circuit shown in Fig. 2.29 determine the voltages at nodes B and C and calculate the current through the 8 Ω resistor.

+,- * *. * # %5 & !5 /"
( )) /

Chapter 9

The Sphere

In a sequel to *Flatland*, an octagon by the name of Mr. Puncto surveys some rather large triangles and finds that each has angles that add up to slightly more than 180°! This discovery excites Mr. Puncto, and, after double checking his data, he makes the discovery public. Unfortunately neither the civil authorities nor the scientific establishment shares his excitement. They suspect he is merely inventing excuses to explain some errors in his measurements, and they dismiss him from his job. The true explanation is that the Flatlanders are living not in a plane but on a sphere, and on a sphere the angles of a triangle really do add up to more than 180°, as we shall soon see. This incident and much more is described in the book *Sphereland* by Dionys Burger. I heartily recommend *Sphereland* to all readers of the present book (just don't be put off by the rather dull summary of *Flatland* at the beginning).

A triangle drawn on a sphere is called a *spherical triangle* (Figure 9.1). Each side of a spherical triangle is required to be a geodesic; that is, it is required to be intrinsically straight in the sense that a Flatlander on the sphere would perceive it as bending neither to the left nor to the

07:40:41.

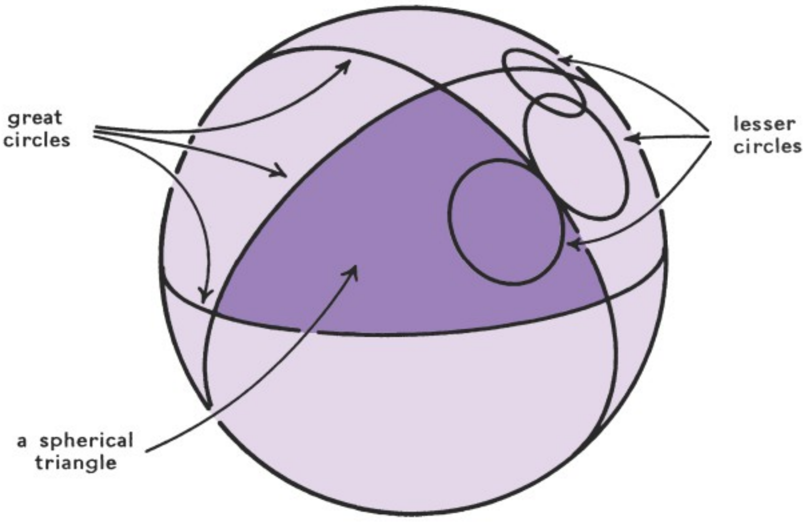

Figure 9.1: On any sphere, the great circles are those circles that are as big as possible. A great circle appears straight to a Flatlander on the sphere. By contrast, any lesser circle appears to bend to one side or the other.

right. A side of a spherical triangle is thus an arc of a so-called *great circle* (see Figure 9.1).

From now on we will *measure all angles in radians*, to facilitate easier comparison of angles and areas (in a minute you'll see how and why we want to do this). Recall that π radians = 180°, $\pi/2$ radians = 90°, etc. Except when specified otherwise, we will henceforth assume that all spheres are unit spheres, i.e. they all have radius one.

Exercise 9.1 For each spherical triangle in Figure 9.2 compute (1) the sum of its angles in radians, and (2) its area. To compute the areas, use the fact that the unit sphere has area 4π. For example, the first triangle shown occupies $1/8$ of the sphere, so its area is $(4\pi)/8 = \pi/2$.

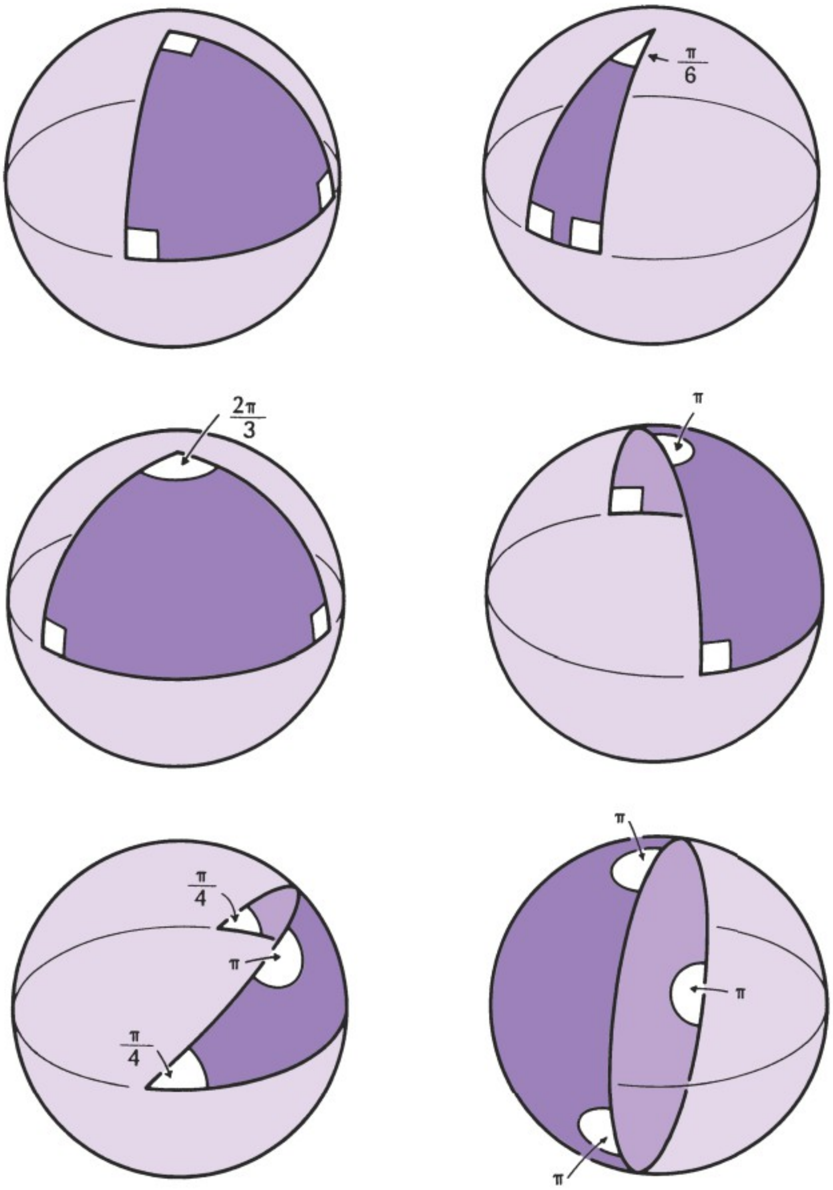

Figure 9.2: Some assorted spherical triangles. Three of the triangles are "degenerate" in the sense that each has one or more angles equal to π*. The last triangle occupies an entire hemisphere, and its three sides all lie on the same great circle.*

Exercise 9.2 Using your answers to Exercise 9.1, find a formula relating a spherical triangle's angle-sum to its area. This formula appeared in 1629 in the section "De la mesure de la superfice des triangles et polygones sphéricques, nouvellement inventée" of the book *Invention nouvelle en l'algebre* by Albert Girard. You should try to find the formula before reading on, because the following paragraphs give it away.

Exercise 9.3 What is the area of a spherical triangle whose angles in radians are $\pi/2$, $\pi/3$, and $\pi/4$? What is the area of a spherical triangle with angles of 61°, 62°, and 63°?

The proof of the formula you just discovered is simple and elegant. (It is not, however, the sort of thing you're likely to stumble onto on your own. I struggled for hours without being able to prove the formula at all.) The easiest proof uses "lunes".

A *lune* is a region on a sphere bounded by two great semi-circles, as shown in Figure 9.3. As you increase a lune's angle θ from 0 to 2π, the area A that it encloses increases proportionally from 0 to the whole area of the sphere, which is 4π. In symbols, $A = 2\theta$.

Now we'll find the formula for the area of a spherical triangle with angles α, β, and γ. First extend the sides of the triangle all the way around the sphere to form three great circles, as shown in Figure 9.4. An "antipodal triangle", identical to the original, is formed on the back side of the sphere. Figure 9.5*(a)* shows three colored lunes. These lunes have respective angles $\pi - \alpha$, $\pi - \beta$, and $\pi - \gamma$, and therefore have areas $2(\pi \quad \alpha)$, $2(\pi - \beta)$, and $2(\pi - \gamma)$.

07:40:41.

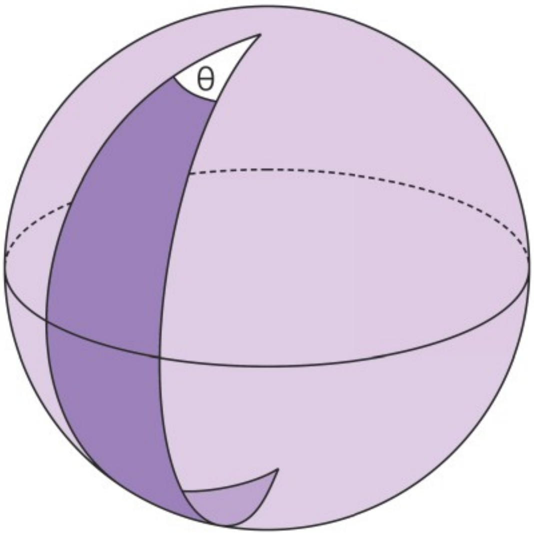

Figure 9.3: A lune with angle θ

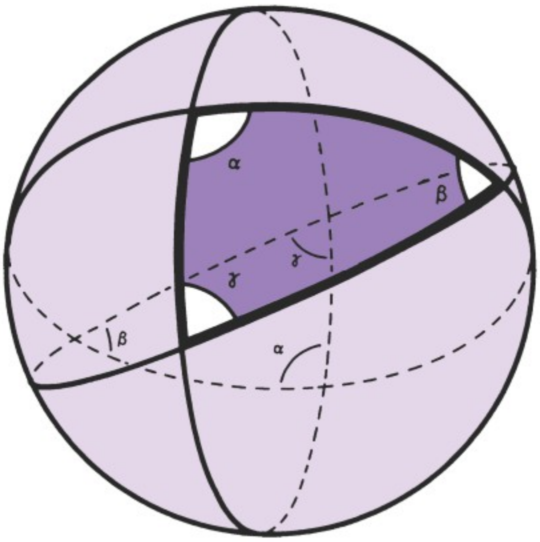

Figure 9.4: Extend the edges of the spherical triangle, and the resulting great circles will form an "antipodal triangle" on the back side of the sphere.

07:40:41.

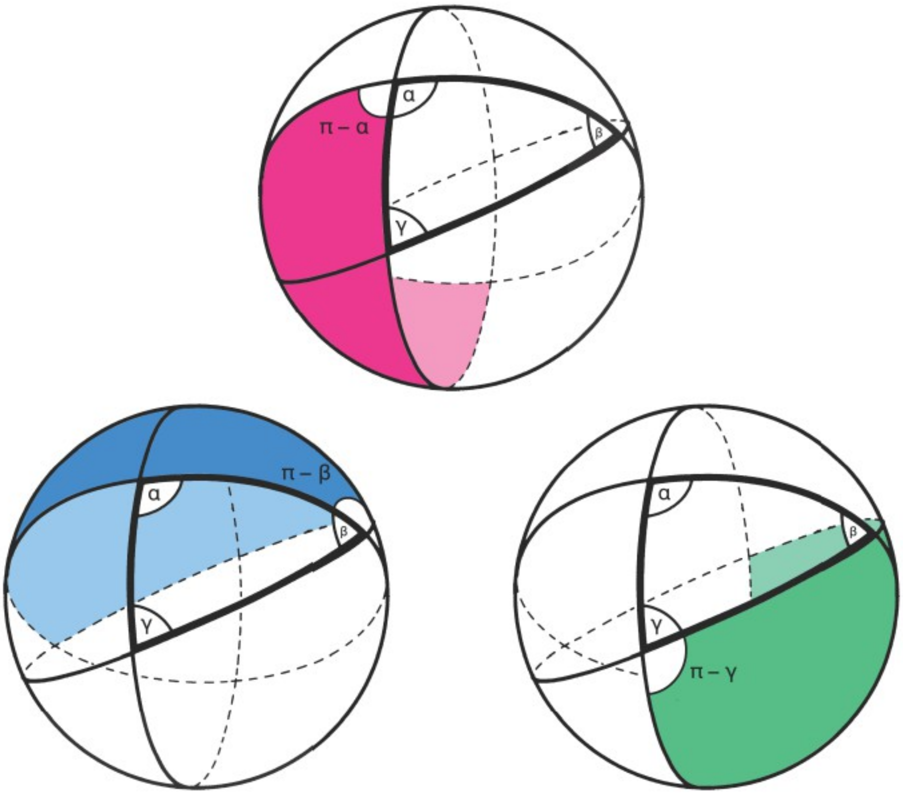
α
π − α
β
γ
α
π − β
β
γ
α
β
γ
π − γ

(a) Three colored lunes

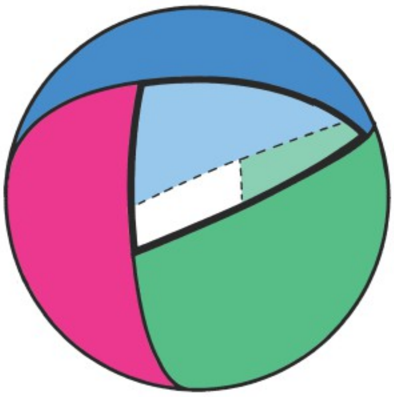

(b) The three colored lunes fill the whole sphere except for the original and antipodal triangles.

Figure 9.5

07:40:41.

Taken together (Figure 9.5*(b)*), the three colored lunes fill almost the whole sphere, except for the original and antipodal triangles. So . . .

$$\begin{array}{ccccccccccc} \begin{pmatrix}\text{area}\\ \text{of red}\\ \text{lune}\end{pmatrix} & + & \begin{pmatrix}\text{area}\\ \text{of blue}\\ \text{lune}\end{pmatrix} & + & \begin{pmatrix}\text{area}\\ \text{of green}\\ \text{lune}\end{pmatrix} & + & \begin{pmatrix}\text{area of}\\ \text{original}\\ \text{triangle}\end{pmatrix} & + & \begin{pmatrix}\text{area of}\\ \text{antipodal}\\ \text{triangle}\end{pmatrix} & = & \begin{pmatrix}\text{area}\\ \text{of entire}\\ \text{sphere}\end{pmatrix} \\ 2(\pi-\alpha) & + & 2(\pi-\beta) & + & 2(\pi-\gamma) & + & A & + & A & = & 4\pi \end{array}$$

$$6\pi - 2(\alpha+\beta+\gamma) + 2A = 4\pi$$

$$A = (\alpha+\beta+\gamma) - \pi$$

which is just what we wanted to prove! In words, this formula says that the sum of the angles of a spherical triangle exceeds π by an amount equal to the triangle's area.

Exercise 9.4 The formula $A = (\alpha+\beta+\gamma) - \pi$ applies only to triangles on a sphere of radius one. How must you modify the formula to apply to triangles on a sphere of radius two? What about radius three? Write down a general formula for triangles on a sphere of radius r.

Exercise 9.5 A society of Flatlanders lives on a sphere whose radius is exactly 1000 meters. A farmer has a triangular field with perfectly straight (i.e. geodesic) sides and angles, which have been carefully measured as 43.624°, 85.123°, and 51.270°. What is the area of the field? Don't forget to convert the angles to radians. (Bonus Question: How accurately do you know the field's area? That is, by plus or minus what percent?)

07:40:41.

Exercise 9.6 A society of Flatlanders lives on a sphere. They carefully survey a triangle and find its angles to be 60.0013°, 60.0007°, and 60.0011°, while its area is 5410.52 m^2. What is the area of the entire sphere?

Exercise 9.7 Estimate the angle-sum of each of the following spherical triangles. *Hint:* First make a rough estimate of the area enclosed by each triangle, and then apply the formula from Exercise 9.4. Because you have to guess the areas of the triangles, you'll get only approximate, not exact, answers. The radius of the Earth is roughly 6400 km.

1. The triangle formed by Providence, Newport, and Westerly, Rhode Island. These cities are roughly 50 km apart.
2. The triangle formed by Houston, El Paso, and Amarillo, Texas. These cities are roughly 1000 km apart.
3. The triangle formed by Madras, India; Tokyo, Japan; and St. Petersburg, Russia. These cities are roughly 7000 km apart.

We've now seen the first major way in which the geometry of a sphere differs from the geometry of a plane. Namely, the sum of the angles of a spherical triangle exceeds π by an amount proportional to the triangle's area, whereas the sum of the angles of a Euclidean (= flat) triangle equals π exactly (study Figure 9.6 for a proof of this last fact).

07:40:41.

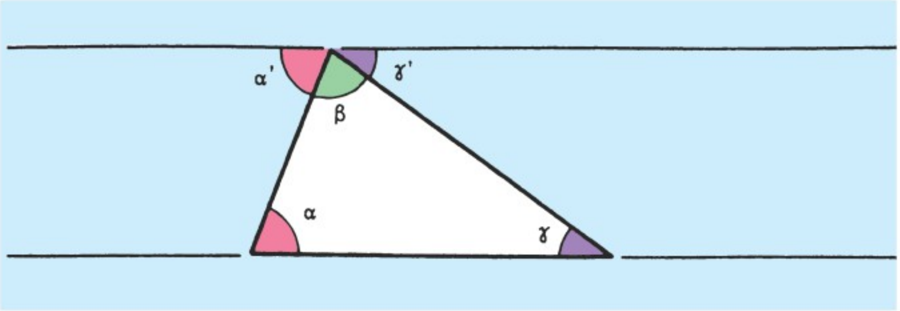

Figure 9.6: A quick proof that the sum of the angles in a Euclidean triangle is π*:*

$$\alpha' + \beta + \gamma' = \pi,$$

$$\alpha = \alpha' \text{ and } \gamma = \gamma',$$

$$\text{therefore } \alpha + \beta + \gamma = \pi.$$

A piece of a sphere rips open when flattened onto a plane (Figure 9.7). This shows that a circle on a sphere has a smaller circumference and encloses less area than a circle of the same radius in a plane. I should stress that the

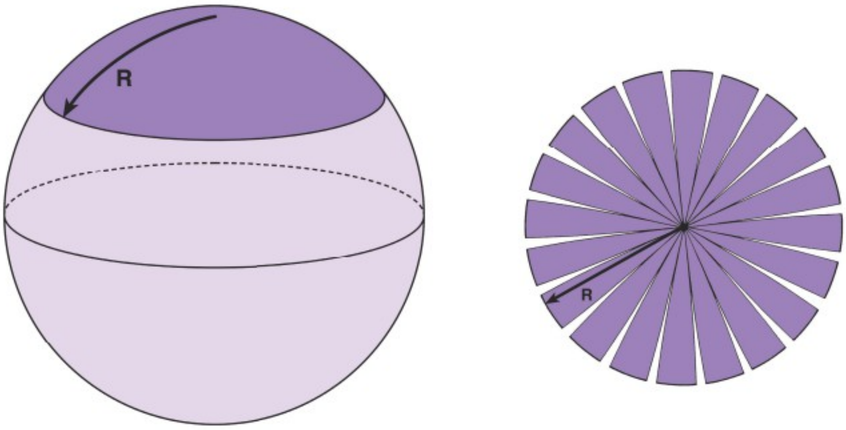

Figure 9.7: A piece of a sphere splits open when flattened.

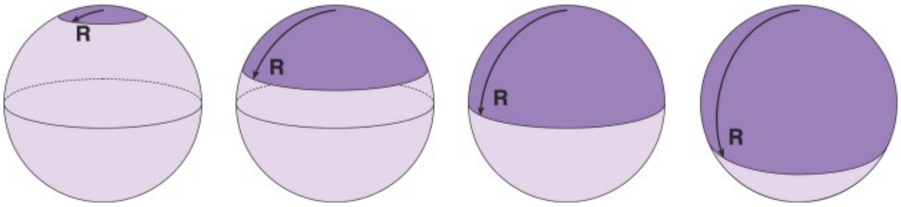

Figure 9.8: The circle's circumference first increases, but then decreases once the circle is past the equator.

radius of a circle on a sphere is measured along the sphere itself—the way a Flatlander would measure it. Figure 9.8 shows that on a sphere a circle's circumference can actually shrink, even though the circle's (intrinsically measured) radius is still increasing.

Exercise 9.8 *(Review exercise)* What other surface has the local geometry of a sphere? Could Flatlanders living on this other surface tell that they weren't on a sphere?

+,- * *. * # : :< / " / +
' " 8 : M : & /" (

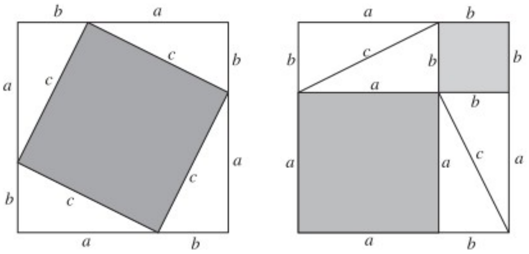

1. The Basics

◆ 1.1. MEASUREMENT

SUPPLIES

tape measure
ruler

These first two sections are intended to remind you of some basic tools and formulae which will be used throughout the text. The word geometry comes from the Greek words for earth and measurement. The first known application of geometry was surveying, most especially in Egypt where annual floods both enriched the soil and eradicated the demarcations between fields. Geometry provided the theoretical tools needed to accurately and efficiently divide up the land into individual holdings. Thus, we begin this book with some quick and easy exercises in measurement.

While the Egyptians developed a great deal of practical geometry, the Greeks developed its theoretical underpinnings, culminating in Euclid's *The Elements*, circa 300 B.C. Euclid's text contains most of the developments in mathematics known at the time, all contained in a superlatively clear and logical structure, wherein each result leads naturally to the next. *The Elements* is one of the most influential texts in the history of western civilization, perhaps only surpassed by the Bible. However, our purpose is not to recapitulate the standard results of Euclidean geometry. We are going to assume that you have already studied these, though we will provide significant prompting at times. We will make note of some of the particular results when we need to use them, such as the properties of isosceles triangles and the various ways of showing congruence.

It is often useful to be able to get a rough idea how big something is, even if you don't happen to have a ruler handy. The exercises that follow are intended to lead you to find readily portable substitutes for a ruler or yardstick.

▷ **Exercise 1.** Find a part of yourself an inch long. I use the first joint on the middle finger of my left hand which happens to be almost exactly one inch long. For some reason the corresponding joint on my right hand is a little longer.

▷ **Exercise 2.** Find a part of yourself a foot long. Use this and your "inch" measure of Exercise 1 to measure the length and width of your desk. Compare this with the actual length and width you find using your ruler. If you are too far off (more than 10%), explain why.

▷ **Exercise 3.** Find a part of yourself a yard long. Use this to measure the length and width of the blackboard. Compare this with the actual dimensions. If you are too far off (more than 10%), explain why.

▷ **Exercise 4.** Measure the length of your stride. Use this to measure the dimensions of your classroom. Compare this with the actual dimensions. If you are too far off (more than 10%), explain why.

The Pythagorean Theorem

One of the most useful tools in geometry is the Pythagorean theorem:

Pythagorean Theorem: If *ABC* is a right triangle with the right angle at vertex *C*, then the side *AB* opposite vertex *C* is called the hypotenuse and its length squared is the sum of the squares of the other two sides, *BC* and *AC*. I.e., $(AB)^2 = (BC)^2 + (AC)^2$, or if we label the sides *a*, *b*, and *c* as in the picture, then $c^2 = a^2 + b^2$.

Another way of stating the Pythagorean theorem in purely geometric terms is that the area of the square built with side *c* is equal to the combined areas of the squares built with sides *a* and *b*.

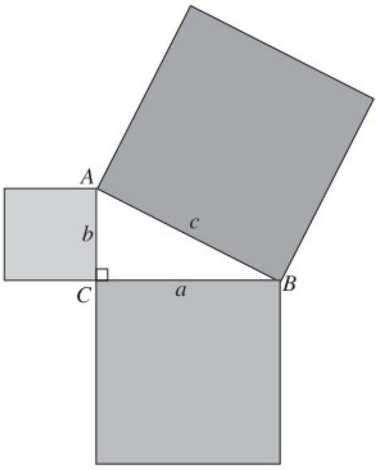

There are many different proofs of this theorem. Elisha Loomis has collected 367 of these in his book *The Pythagorean Proposition*, including one by President Garfield. My favorite of these is the following: Four copies of triangle ΔABC can be arranged as in the picture below on the left to form part of a square with side $(a + b)$. Thus the big square has area equal to the areas of four triangles plus the area of a $c \times c$ square. But the same four triangles can be rearranged inside the same big square as on the right, so that the area of the big square with side $(a + b)$ is also equal to the areas of the four triangles plus the areas of an $a \times a$ square and a $b \times b$ square. Thus, $c^2 = a^2 + b^2$:

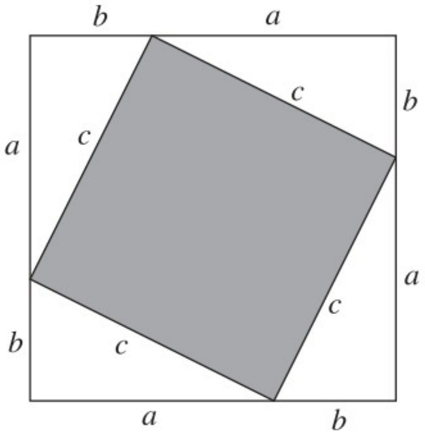

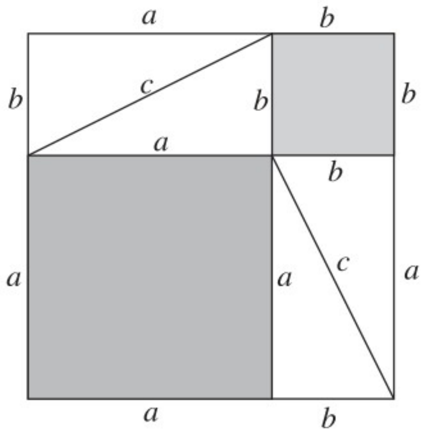

Another proof of the Pythagorean theorem involves cutting up the middle-sized square, which has side a. A copy of the triangle ΔABC can be positioned inside the $a \times a$ square so the edge c goes through the point O at the center of the square as in the figure on the next page. The point O divides length c into two equal pieces, labelled x in the diagram, so that $c = 2x$. The side of the square is also divided into two pieces: a longer one marked y and a shorter one marked z. The diagram makes it clear that $y - z = b$, where b is the shortest side of the triangle. Draw another line perpendicular to the edge c as shown at the bottom left. These two lines divide the $a \times a$ square up into four identical pieces. These four pieces can be rearranged in the larger $c \times c$ square to leave a small square at the center. We must show that this small square is $b \times b$. From the lengths of the sides indicated, the small square has side equal to $y - z$, which we already know is equal to b. Thus the hole is a $b \times b$ square and so the largest $c \times c$ square has area equal to the sum of the areas of the middle-sized $a \times a$ square plus the $b \times b$ square. Therefore $c^2 = a^2 + b^2$:

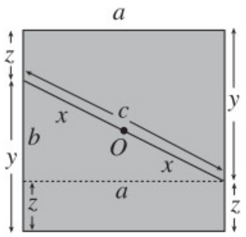

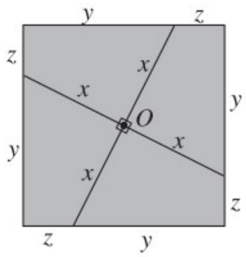

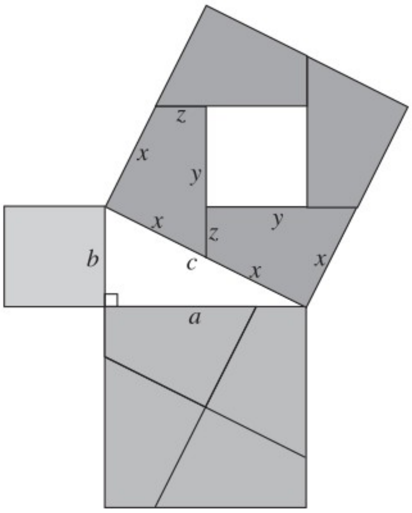

▷ **Exercise 5.** Redraw the figures for the proof above for the case in which *a* and *b* are the same length.

The converse of the Pythagorean theorem is also true:

Theorem: If a triangle has sides *a*, *b*, and *c*, and $c^2 = a^2 + b^2$, then the triangle is a right triangle with the right angle opposite the side *c*.

▷ **Exercise 6.** If a triangle has sides of lengths 3, 4, and 5, is it a right triangle, and, if so, where is the right angle?

► **Exercise 7.** If a right triangle has shorter sides of lengths 5 and 12, how long is the hypotenuse?

▷ **Exercise 8.** Triples of integers, such as those occurring in Exercise 6: 3-4-5 and Exercise 7: 5-12-?, are called Pythagorean triples, since they occur as the side lengths in right triangles and thus satisfy the Pythagorean theorem. Give one more example of an **integer** Pythagorean triple.

Since 1993 there has been a lot of publicity about Fermat's Last Theorem. This all started in 1637 when Pierre de Fermat, a French lawyer and very good amateur mathematician, was reading a book in which there was a proof that there are infinitely many Pythagorean triples. He scribbled in the book that he had a wonderful proof that this statement was false for higher powers, but there was not enough space in the margin for him to write it down. In other words, he was claiming that there are never integers *a*, *b*, and *c* satisfying the equation $a^n + b^n = c^n$ for any integer $n \geq 3$. For example, we showed above that a $c \times c$ square can be made out of an $a \times a$ square and a $b \times b$ square. Fermat's Last

Theorem says, for example, that a $c \times c \times c$ cube cannot be made from the pieces of a $a \times a \times a$ cube and $b \times b \times b$ cube for any integers a, b, and c. For many years, mathematicians tried to prove this theorem. It was finally proven in 1994 by Andrew Wiles. Wiles used extremely sophisticated techniques, so it is unlikely that Fermat's idea for a proof, whatever it was, was correct.

Perimeter

The *perimeter*, or *circumference*, of a geometric figure is the sum of the lengths of all of the sides, or the distance to walk all the way around the figure. For circles, one can use the formula

$$P = 2\pi r$$

to figure out the perimeter, or distance around the circle, where r is the radius of the circle (the distance from the center to the outside). This formula can be used to find the perimeter if you know the radius, or to find the radius if you know the perimeter.

▷ **Exercise 9.** (a) If a half-circle has radius 3, what is its perimeter? (Measure only the curved part of the semicircle.)
(b) If the perimeter of a circle is 8π, what is its radius?

► **Exercise 10.** Find the perimeter of each of these regions. Some of the figures in this exercise are composites of rectangles and portions of circles.

(a)

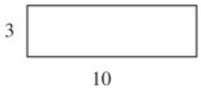

(b)

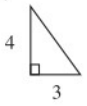

(c)

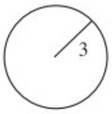

(d)

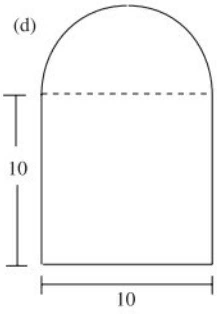

(e)

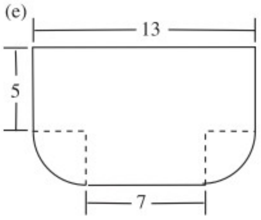

(f)

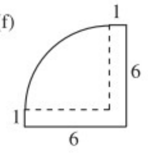

(g)

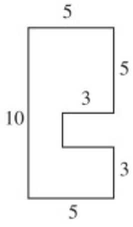

Area

The figure whose area is easiest to measure is the rectangle. For example, below is a 7 by 4 rectangle, which is 7 units wide and 4 units tall, and thus has area 7 times 4, or 28, square units. You can see in the picture that the rectangle contains 28 small squares.

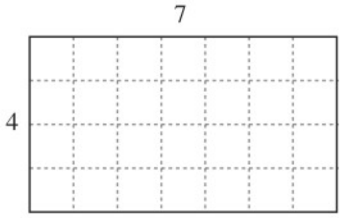

Thus the area of a rectangle is always the width times the height:

$$A(\text{rectangle}) = w \cdot h.$$

Next, we can figure out the area of a parallelogram labeled *ABCD* by dropping a perpendicular line from *D* to the point *E* on line *AB*.

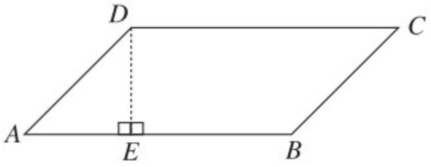

Notice that if the triangle *AED* is cut off and then glued back so that *AD* is glued to *BC*, a rectangle is formed that has the same area as the original parallelogram:

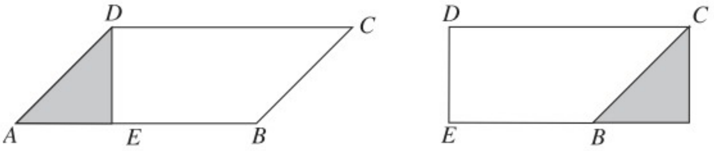

Thus, the area of a parallelogram is also width times height, but the height must be measured by dropping a perpendicular to the side that represents the width.

$$A(\text{parallelogram}) = w \cdot h.$$

▷ **Exercise 11.** You probably remember the formula for the area of a triangle from high school. State the formula and give an explanation of why this is the correct rule. You should model your explanation on the one above for the area of the parallelogram.

▷ **Exercise 12.** Figure out the area of the isosceles trapezoid shown below:

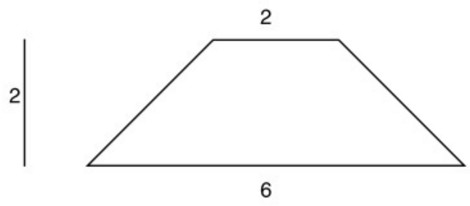

▷ **Exercise 13.** Figure out and state clearly a rule for the area of any trapezoid. Give an explanation of why your formula is correct. Be careful not to assume all trapezoids are as symmetrical as the one in Exercise 12.

The area formulae for other polygons are deduced from the ones above. The exception is the area of a circle.

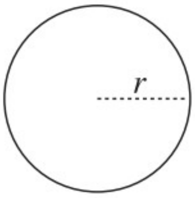

The radius r is measured from the center of the circle to the rim. The area is given by

$$A(\text{circle}) = \pi \cdot r^2.$$

Most other curved shapes require calculus in order to find their areas.

Now use these formulae to figure out the areas of the following figures. Some of the figures in Exercise 14 are composites of rectangles and portions of circles.

▶ **Exercise 14.** Find the area of each of these regions:

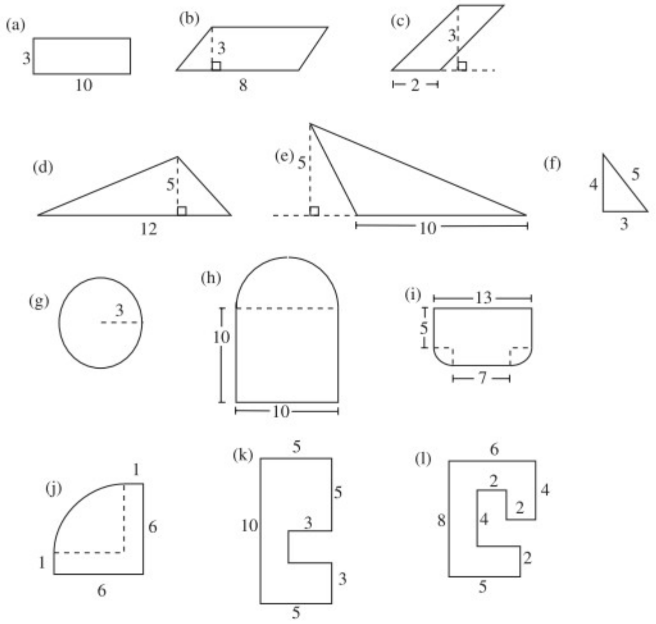
(a)
3
10
(b)
3
8
(c)
3
2
(d)
5
12
(e)
5
10
(f)
4
5
3
(g)
3
(h)
10
10
(i)
13
5
7
(j)
1
6
1
6
(k)
5
5
10
3
3
5
(l)
6
2
4
2
8
4
2
5

Angles

Angles are used to measure steepness or inclination. A 0° angle is considered horizontal, while a 90°, or right angle, is perpendicular to the horizontal. To draw a right angle, an architect or builder uses ruler and compass, or, either a T-square (a sort of T-shaped ruler that has a precut right angle) or a triangle template. We shall assume that you had plenty of practice with angle measure in previous courses.

Some professions use other ways to describe angles or steepness. Builders talk of the pitch of a roof and will describe it, say, as a 4 in 12 pitch, meaning that for every 12 feet of horizontal distance, the roof goes up 4 feet.

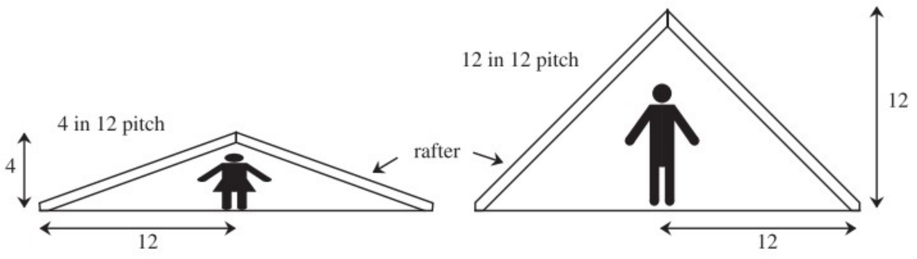
4 in 12 pitch
4
12
rafter
12 in 12 pitch
12
12

Different roof pitches have different advantages and disadvantages. A flat roof or a roof with low pitch (generally, a pitch less than 4 in 12 is considered low) minimizes the area of the roof, requiring less material for construction, but building codes require that leak-prone continuous membrane roofing be used on flat roofs and on roofs with pitch less than 3 in 12. Higher pitch roofs provide better runoff for rain and snow but require more building materials and form awkward interior spaces. Pitches greater than 12 in 12 are very rarely used.

▷ **Exercise 15.** What angle does a 12 in 12 pitch roof make with the horizontal?

► **Exercise 16.** If the building below has a roof with a 12 in 12 pitch and the building is 50 feet wide, how high is the peak of the roof from the attic floor? If the overhang is 1 foot, how long must the rafters be? What is the total area of the roof surface if the building is 75 feet long?

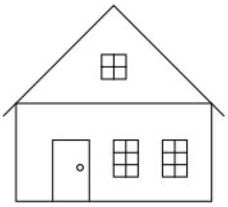

▷ **Exercise 17.** If a building has a roof with a 4 in 12 pitch and the building is 50 feet wide, how high is the peak of the roof from the attic floor? If the overhang is 1 foot, how long must the rafters be? What is the total area of the roof surface if the building is 75 feet long?

Stairs in buildings are commonly specified by riser and tread: the riser is the vertical height of the step, and the tread is the depth of the horizontal part of the step. Thus, an architect may specify a stair with a 7.5 inch riser and an 11 inch tread with a 1 inch nose, as pictured below:

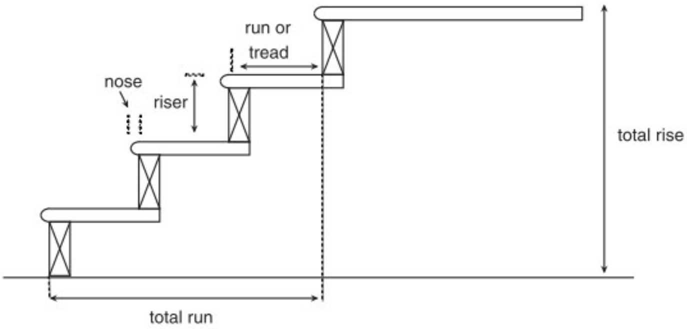

▷ **Exercise 18.** Design two different stairs from the first to the second story of a house if the distance from the floor to the ceiling is 7 feet and the interstitial space (the space between the ceiling and the floor of the story above) is 6 inches. Your stairs must satisfy the National Fire Codes for public buildings, which specify a maximum riser height of 7 inches and a minimum tread depth of 11 inches, and that treads must have uniform depth and risers uniform height. Use different risers and treads for each of your stairs.

Ramps are also specified by rise and run. The maximum steepness allowed for handicapped accessibility is a 1:12 slope, i.e., a handicapped ramp can go up at most 1 foot for every 12 feet of horizontal distance.

▷ **Exercise 19.** Check one handicapped ramp for compliance with the building codes.

In mathematics, slope is also described as the relation between rise and run, but is usually written as a fraction $\frac{rise}{run}$, so that instead of 1 in 12 or 1:12, a mathematician would describe the ramp as having slope $\frac{1}{12}$. A line 45° from the horizontal will have slope 1. Slopes more than 1 give steeper lines, while slopes less than 1 are more shallow.

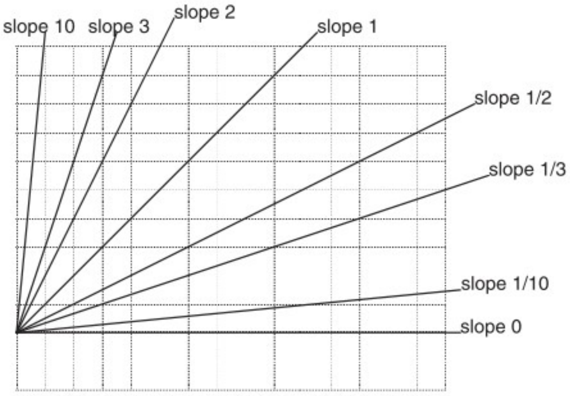

While finding the angle between two lines is a familiar task, sometimes one needs to find the angle between two curves. The angle between two curves is defined to be the angle between their *tangent lines* at the point where they intersect. The tangent line is the line that best approximates the curve at that point. The angle θ is formed between the curves shown:

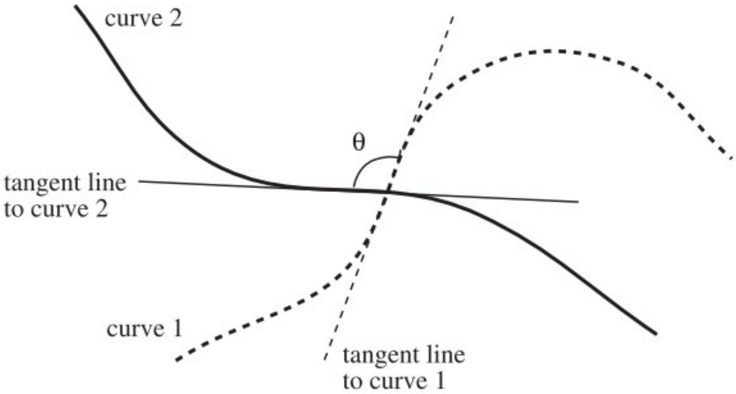
curve 2
θ
tangent line
to curve 2
curve 1
tangent line
to curve 1

In general, one needs calculus in order to find a tangent line. The one easy case is a circle: it happens that the tangent line to a circle at a point is perpendicular to the radial line from the center to that point.

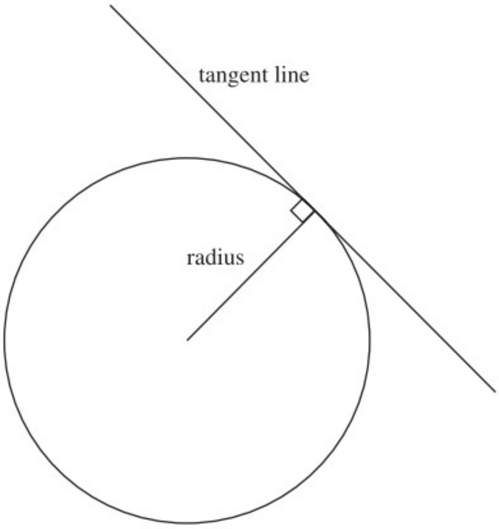
tangent line
radius

When one talks about the angle formed between a line and a circle, one refers to the angle θ formed by the line and this tangent line:

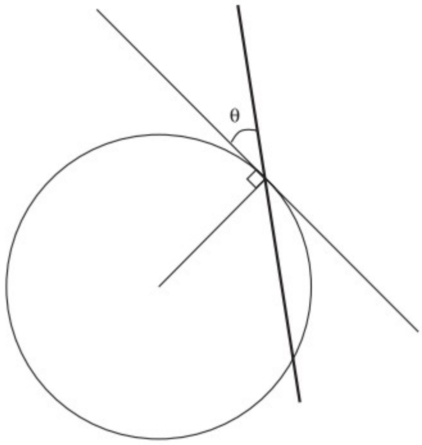

SUGGESTED READING

Elisha S. Loomis, *The Pythagorean Proposition*, NCTM, Washington, DC, 1968.

Hints and Solutions to Selected Problems

1.1. MEASUREMENT

- **7.** If a right triangle has legs of lengths 5 and 12, the hypotenuse is $\sqrt{5^2 + 12^2} = \sqrt{169} = 13$ long.
- **10. a.** 26 **d.** $30 + 5\pi \approx 45.71$ **f.** $14 + \frac{5\pi}{2} \approx 21.85$
- **14. b.** 24 **e.** 25 **h.** $100 + 12.5\pi \approx 139.27$
  j. $11 + 6.25\pi \approx 30.63$
- **16.** The peak is 25 feet from the attic floor. The rafters must be $1 + 25\sqrt{2} \approx 36.36$ feet. The area of the roof is $2 \cdot 75 \cdot (1 + 25\sqrt{2}) \approx 5453.3$ square feet.

1.2. POLYGONS

- **1.** The sizes of the tangram pieces are figured out by using the way they fit together and the Pythagorean theorem: the square is 1 on each side; the smallest triangles have legs 1 and hypotenuse $\sqrt{2}$; the parallelogram has sides of lengths 1 and $\sqrt{2}$; the middle-sized triangle has legs $\sqrt{2}$ and hypotenuse 2; the big triangles have legs 2 and hypotenuse $2\sqrt{2}$.
- **3.** Many of these have more than one solution, but here is one for each of the first four:

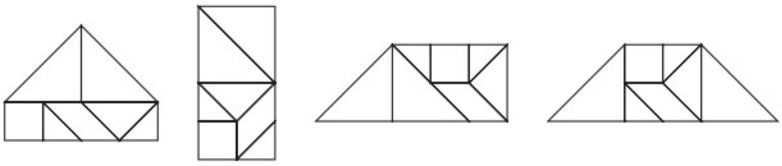

% *+,-* * *. * # ; / / >
: " : 3 3 4 /"

" *+,-* * *. * # < " . '
F 6 /"
.6 ? 2 " NNN

9 % $ $ %)5

P0 ' % (%5 " " P0 G 2
5 "

! " " # $ $ % & ' # (#!
" " "

)*+,)))-) # . / + 0
' ' 1 21 0" 3"

)*+,)))-) # . / + 0
" " 1 21 0" 3"

)*+,)))-) # . / + 0
"' +4 0" 3"

(%5 66 2 78 + " 2

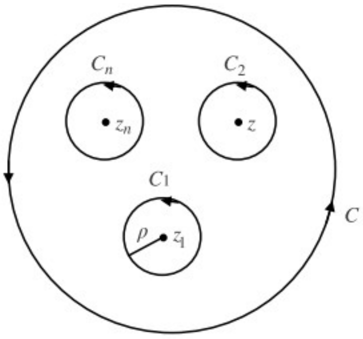
C_n
C_2
z_n
z
C_1
ρ
z_1
C

9 %

Solution. We have

$$\frac{e^z}{z^3} = \frac{1}{z^3}\left(1+z+\frac{z^2}{2!}+\frac{z^3}{3!}+\dots\right)$$

$$= \frac{1}{z^3}+\frac{1}{z^2}+\frac{1}{2!z}+\frac{1}{3!}+\frac{1}{4!}z+\dots$$

Thus, the principal part of the Laurent expansion consists of three terms and so $\frac{e^z}{z^3}$ has a pole of order 3 at $z = 0$.

EXAMPLE 1.76

Show that $z \sin \frac{1}{z}$ has essential singularity at $z = 0$.

Solution. We have

$$z\sin\frac{1}{z} = z\left\{\frac{1}{z}-\frac{1}{3!z^3}+\frac{1}{5!z^5}-\dots\right\}$$

$$= 1-\frac{1}{3!z^2}+\frac{1}{5!z^4}+\dots$$

Since the series *does not terminate*, $z = 0$ is an *essential singularity*.

Definition 1.31. A function $f(z)$ is said to be *meromorphic* if it is analytic in the finite part of the plane except at a finite number of poles.

1.6 RESIDUES AND CAUCHY'S RESIDUE THEOREM

Definition 1.32. Let the Laurent series expansion of a function $f(z)$ at isolated singularity z_0 be

$$f(z) = \sum_{n=1}^{\infty}\frac{b_n}{(z-z_0)^n}+\sum_{n=0}^{\infty}a_n(z-z_0)^n \quad (1.40)$$

The coefficient b_1, in the principal part of the expansion, given by

$$b_1 = \frac{1}{2\pi i}\int_C f(\xi)\,d\xi \quad (1.41)$$

for the contour C: $|z - z_0| = r < R$ is called *residue* of $f(z)$ at z_0 and is denoted by Res(z_0).

The residue of $f(z)$ at $z = \infty$ is defined by

$$-b_1 = \frac{-1}{2\pi i}\int_C f(\xi)\,d\xi.$$

It is the coefficient of $\frac{1}{z}$ with its sign changed in the expansion of $f(z)$ in the neighbourhood of $z = \infty$.

If $f(z)$ has a pole of order m at z_0, then the Laurent expansion of $f(z)$ is

$$f(z) = \frac{b_m}{(z-z_0)^m}+\frac{b_{m-1}}{(z-z_0)^{m-1}}+\dots+\frac{b_1}{(z-z_0)}$$
$$+a_0+a_1(z-z_0)+a_2(z-z_0^2)+\dots$$

Multiplying both sides by $(z - z_0)^m$, we have

$$(z-z_0)^m f(z) = b_m + b_{m-1}(z-z_0)$$
$$+\dots+b_1(z-z_0)^{m-1}$$
$$+a_0(z-z_0)^m$$
$$+a_1(z-z_0)^{m+1}+\dots,$$

which is Taylor's series of the analytic function $(z - z_0)^m f(z)$. Differentiating both sides $m - 1$ times with respect to z, we have

$$\frac{d^{m-1}}{dz^{m-1}}[(z-z_0)^m f(z)] = b_1(m-1)!$$
$$+m(m-1)\dots 2a_0(z-z_0)+\dots$$

Letting $z \to z_0$, we get

$$\lim_{z\to z_0}\frac{d^{m-1}}{dz^{m-1}}[(z-z_0)^m f(z)] = b_1(m-1)!$$

and so

$$b_1 = \text{Res}(z_0)$$
$$= \frac{1}{(m-1)!}\lim_{z\to z_0}\frac{d^{m-1}}{dz^{m-1}}[(z-z_0)^m f(z)]. \quad (1.42)$$

If z_0 is a simple pole, that is, $m = 1$, then

$$\text{Res}(z_0) = b_1 = \lim_{z\to z_0}(z-z_0)f(z). \quad (1.43)$$

If

$$f(z) = \frac{p(z)}{q(z)},$$

where $p(z)$ and $q(z)$ are analytic at $z = z_0$, $p(z_0) \neq 0$ and $q(z)$ has a simple zero at z_0, that is, $f(z)$ has

Solution. We have

$$\frac{e^z}{z^3} = \frac{1}{z^3}\left(1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \ldots\right)$$

$$= \frac{1}{z^3} + \frac{1}{z^2} + \frac{1}{2!\,z} + \frac{1}{3!} + \frac{1}{4!}z + \ldots$$

Thus, the principal part of the Laurent expansion consists of three terms and so $\frac{e^z}{z^3}$ has a pole of order 3 at $z = 0$.

EXAMPLE 1.76

Show that $z \sin \frac{1}{z}$ has essential singularity at $z = 0$.

Solution. We have

$$z \sin\frac{1}{z} = z\left\{\frac{1}{z} - \frac{1}{3!\,z^3} + \frac{1}{5!\,z^5} - \ldots\right\}$$

$$= 1 - \frac{1}{3!\,z^2} + \frac{1}{5!\,z^4} + \ldots$$

Since the series *does not terminate*, $z = 0$ is an *essential singularity*.

Definition 1.31. A function $f(z)$ is said to be *meromorphic* if it is analytic in the finite part of the plane except at a finite number of poles.

1.6 RESIDUES AND CAUCHY'S RESIDUE THEOREM

Definition 1.32. Let the Laurent series expansion of a function $f(z)$ at isolated singularity z_0 be

$$f(z) = \sum_{n=1}^{\infty} \frac{b_n}{(z - z_0)^n} + \sum_{n=0}^{\infty} a_n (z - z_0)^n \quad (1.40)$$

The coefficient b_1, in the principal part of the expansion, given by

$$b_1 = \frac{1}{2\pi i}\int_C f(\xi)\, d\xi \quad (1.41)$$

for the contour C: $|z - z_0| = r < R$ is called *residue* of $f(z)$ at z_0 and is denoted by Res(z_0).

The residue of $f(z)$ at $z = \infty$ is defined by

$$-b_1 = \frac{-1}{2\pi i}\int_C f(\xi)\, d\xi.$$

It is the coefficient of $\frac{1}{z}$ with its sign changed in the expansion of $f(z)$ in the neighbourhood of $z = \infty$.

If $f(z)$ has a pole of order m at z_0, then the Laurent expansion of $f(z)$ is

$$f(z) = \frac{b_m}{(z - z_0)^m} + \frac{b_{m-1}}{(z - z_0)^{m-1}} + \ldots + \frac{b_1}{(z - z_0)}$$
$$+ a_0 + a_1(z - z_0) + a_2(z - z_0^2) + \ldots$$

Multiplying both sides by $(z - z_0)^m$, we have

$$\begin{aligned}(z - z_0)^m f(z) = b_m &+ b_{m-1}(z - z_0) \\ &+ \ldots + b_1(z - z_0)^{m-1} \\ &+ a_0(z - z_0)^m \\ &+ a_1(z - z_0)^{m+1} + \ldots,\end{aligned}$$

which is Taylor's series of the analytic function $(z - z_0)^m f(z)$. Differentiating both sides $m - 1$ times with respect to z, we have

$$\frac{d^{m-1}}{dz^{m-1}}[(z - z_0)^m f(z)] = b_1(m-1)!$$
$$+ m(m-1)\ldots 2a_0(z - z_0) + \ldots$$

Letting $z \to z_0$, we get

$$\lim_{z \to z_0} \frac{d^{m-1}}{dz^{m-1}}[(z - z_0)^m f(z)] = b_1(m-1)!$$

and so

$$\begin{aligned} b_1 &= \text{Res}(z_0) \\ &= \frac{1}{(m-1)!}\lim_{z \to z_0} \frac{d^{m-1}}{dz^{m-1}}[(z - z_0)^m f(z)]. \quad (1.42)\end{aligned}$$

If z_0 is a simple pole, that is, $m = 1$, then

$$\text{Res}(z_0) = b_1 = \lim_{z \to z_0}(z - z_0) f(z). \quad (1.43)$$

If

$$f(z) = \frac{p(z)}{q(z)},$$

where $p(z)$ and $q(z)$ are analytic at $z = z_0$, $p(z_0) \neq 0$ and $q(z)$ has a simple zero at z_0, that is, $f(z)$ has a simple pole at z_0. Then $q(z) = (z - z_0)\, g(z)$, $g(z_0) \neq 0$ and $q(z_0) = 0$. Hence (1.43) reduces to

$$\operatorname{Res}(z_0) = \lim_{z\to z_0} (z - z_0)\, f(z) = \lim_{z\to z_0} (z - z_0)\, \frac{p(z)}{q(z)}$$

$$= \lim_{z\to z_0} \frac{p(z)}{\dfrac{q(z) - q(z_0)}{z - z_0}} = \frac{p(z_0)}{q'(z_0)} \qquad (1.44)$$

Thus, the residues at poles can be calculated using the formulas (1.42), (1.43), and (1.44).

EXAMPLE 1.77

Find residues of

(i) $f(z) = \dfrac{z^2 - 2z}{(z+1)^2\,(z^2+4)}$

(ii) $f(z) = \dfrac{e^{z^2}}{(z-i)^3}$

at all its poles.

Solution. (i) The function $f(z)$ has a pole of order 2 at $z = -1$ and simple poles at $z = \pm 2i$. Therefore,

$$\operatorname{Res}(-1) = \lim_{z\to -1} \frac{d}{dz}\left[(z+1)^2\, \frac{z^2 - 2z}{(z+1)^2 (z^2+4)}\right]$$

$$= \lim_{z\to -1} \frac{(z^2+4)\,(2z-2) - (z^2 - 2z)(2z)}{(z^2+4)^2}$$

$$= -\frac{14}{25}$$

$$\operatorname{Res}(2i) = \lim_{z\to 2i} (z - 2i)\frac{z^2 - 2z}{(z+1)^2\,(z^2+4)}$$

$$= \lim_{z\to 2i} \frac{z^2 - 2z}{(z+1)^2 (z+2i)} = \frac{7+i}{25}$$

$$\operatorname{Res}(-2i) = \lim_{z\to -2i} (z+2i)\frac{z^2 - 2z}{(z+1)^2\,(z-2i)\,(z+2i)}$$

$$= \frac{7-i}{25}$$

(ii) $f(z)$ has a pole of order 3 at $z = i$. Hence

$$\operatorname{Res}(i) = \frac{1}{2!}\lim_{z\to i}\frac{d^2}{dz^2}[(z-i)^3\, f(z)]$$

$$= \frac{1}{2}\lim_{z\to i}\frac{d^2}{dz^2}(e^{z^2})$$

$$= \lim_{z\to i}[2z^2 e^{z^2} + e^{z^2}]$$

$$= -\frac{1}{e}.$$

EXAMPLE 1.78

Find the residue of $f(z) = \cot z$ at its poles.

Solution. We have

$$f(z) = \cot z = \frac{\cos z}{\sin z}.$$

The poles of $f(z)$ are given by $\sin z = 0$. Thus $z = n\pi$, $n = 0, \pm 1, \pm 2, \ldots$ are the simple poles. Using formula (1.44), the residue at $z = n\pi$ is given by

$$\operatorname{Res}(n\pi) = \lim_{z\to n\pi} \frac{\cos z}{\dfrac{d}{dz}(\sin z)} = \lim_{z\to n\pi} \frac{\cos z}{\cos z} = 1.$$

EXAMPLE 1.79

Find the residue at each pole of $f(z) = \dfrac{ze^{iz}}{z^2 + a^2}$.

Solution. We have

$$f(z) = \frac{ze^{iz}}{(z+ai)\,(z-ai)}$$

Therefore, $f(z)$ has simple poles at $z = \pm\, ai$. Now

$$\operatorname{Res}(ai) = \lim_{z\to ai} (z - ai) f(z)$$

$$= \lim_{z\to ai} (z - ai)\frac{ze^{iz}}{(z+ai)\,(z-ai)}$$

$$= \lim_{z\to ai} \frac{ze^{iz}}{z+ai} = \frac{e^{-a}}{2},$$

$$\operatorname{Res}(-ai) = \lim_{z\to -ai} (z+ai)\, f(z)$$

$$= \lim_{z\to -ai} \frac{ze^{iz}}{z - ai} = \frac{e^{a}}{2}.$$

EXAMPLE 1.80

Find the residue of $f(z) = \dfrac{1-e^{2z}}{z^4}$ at its poles.

Solution. The function $f(z)$ has a pole of order 4 at $z = 0$. Therefore,

$$\text{Res}(0) = \frac{1}{3!}\lim_{z\to 0}\frac{d^3}{dz^3}[(z-0)^4\ f(z)]$$

$$= \frac{1}{6}\lim_{z\to 0}\frac{d^3}{dz^3}[1-e^{2z}]$$

$$= \frac{1}{6}\lim_{z\to 0}\frac{d^3}{dz^3}\left[1-\left\{1+2z+4z^2+8z^3+\frac{16z^4}{4!}+\dots\right\}\right]$$

$$= \frac{1}{6}\lim_{z\to 0}\frac{d^3}{dz^3}\left[-2z-2z^2-\frac{8}{6}z^3-\frac{16}{24}z^4-\dots\right]$$

$$= \frac{1}{6}\lim_{z\to 0}\left[-8-\frac{294}{24}z-\dots\right] = -\frac{4}{3}.$$

EXAMPLE 1.81

Find the residues of $f(z) = \dfrac{z^3}{(z-1)(z-2)(z-3)}$. at $z = 1$, 2, and 3 and ∞ and show that their sum is zero.

Solution. The function

$$f(z) = \frac{z^3}{(z-1)(z-2)(z-3)}$$

has simple poles at $z = 1$, 2, and 3. Now

$$\text{Res}(1) = \lim_{z\to 1}(z-1)\ f(z) = \lim_{z\to 1}\frac{z^3}{(z-2)(z-3)} = \frac{1}{2}$$

$$\text{Res}(2) = \lim_{z\to 2}(z-2)\ f(z) = \lim_{z\to 2}\frac{z^3}{(z-1)(z-3)} = -8$$

$$\text{Res}(3) = \lim_{z\to 3}(z-3)\ f(z) = \lim_{z\to 3}\frac{z^3}{(z-1)(z-2)} = \frac{27}{2}.$$

To find residue at ∞, we expand $f(z)$ in the neighbourhood of $z = \infty$ as follows:

$$f(z) = \frac{z^3}{z^3\left(1-\frac{1}{z}\right)\left(1-\frac{2}{z}\right)\left(1-\frac{3}{z}\right)}$$

$$= \left(1-\frac{1}{z}\right)^{-1}\left(1-\frac{2}{z}\right)^{-1}\left(1-\frac{3}{z}\right)^{-1}$$

$$= 1+\frac{6}{z}+ \textit{higher powers} \text{ of } \frac{1}{z}.$$

Now residue at ∞ is coefficient of $\frac{1}{z}$ with sign changed. Thus Res(∞) = –6. Hence, the sum of the residues equals $\frac{1}{2}-8+\frac{27}{2}-6 = 0$.

To compute the values of integrals in our study, we shall require the following theorem.

Theorem 1.21. (Cauchy's Residue Theorem). If $f(z)$ is analytic within and on a closed contour C except at finitely many poles lying in C, then

$$\int_C f(z)\ dz = 2\pi i\ \Sigma\ R,$$

where ΣR denotes the sum of residues of $f(z)$ at the poles within C.

Proof: Let $z_1, z_2, \dots, z_n$ be the n poles lying in C. Let $C_1, C_2, \dots, C_n$ be the circles with centre $z_1, z_2, \dots, z_n$ and radius ρ such that all these circles lie entirely within C and do not overlap (Fig. 1.14)

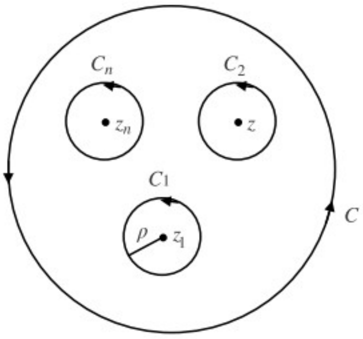

Figure 1.14

Then $f(z)$ is analytic in the region between C and the circles. Hence, by Cauchy-Goursat theorem

$$\int_C f(z)\ dz = \int_{C_1} f(z)\ dz + \int_{C_2} f(z)\ dz + \dots$$
$$+ \int_{C_n} f(z)\ dz. \tag{1.45}$$

If $f(z)$ has a pole of order m_1 at $z = z_1$, then

$$f(z) = \phi_1(z) + \sum_{r=1}^{m_1} \frac{b_r}{(z-z_1)^r},$$

where $\phi_1(z)$ is regular within and on C_1. Therefore,

$$\int_{C_1} f(z)\,dz = \int_{C_1} \phi_1(z)\,dz + \int_{C_1} \frac{b_1}{(z-z_1)}\,dz$$

$$+ \int_{C_1} \frac{b_2}{(z-z_2)^2}\,dz + \ldots + \int_{C_1} \frac{bm_1}{(z-z_1)^r}\,dz. \qquad (1.46)$$

Since $\phi_1(z)$ is analytic within and on C_1, by Cauchy-Goursat theorem

$$\int_{C_1} \phi_1(z)\,dz = 0.$$

Moreover, substituting $z - z_1 = \rho\, e^{i\theta}$, we have

$$\int_{C_1} \frac{bm_1}{(z-z_1)^r}\,dz \;= \int_0^{2\pi} \frac{bm_1 .\rho i e^{i\theta}}{e^{m_1}\; e^{m_1 i\theta}}\,d\theta$$

$$= \frac{ibm_1}{e^{m_1-1}} \int_0^{2\pi} e^{-(m_1-1)i\theta} d\theta = 0$$

for $m_1 \neq 1$

and

$$\int_{C_1} \frac{b_1}{(z-z_1)}\,dz = \int_0^{2\pi} \frac{b_1 \rho i e^{i\theta}\,d\theta}{\rho e^{i\theta}} = 2\pi i\, b_1.$$

Hence (1.46) reduces to

$$\int_{C_1} f(z)\,dz = 2\pi i R_1,$$

where R_1 is the residue of $f(z)$ at $z = z_1$. Similarly,

$$\int_{C_2} f(z)\,dz = 2\pi i\, R_2,$$

....

$$\int_{C_n} f(z)\,dz = 2\pi i\, R_n,$$

where R_i is the residue of $f(z)$ at $z = z_i$. Hence (1.45) becomes

$$\int_C f(z)\,dz = 2\pi i(R_1 + R_2 + \ldots + R_n) = 2\pi i\, \Sigma R,$$

where $\Sigma R = R_1 + R_2 + \ldots + R_n$.

Remark 1.10. In the assumptions of Cauchy's integral formula, $f(z)$ is assumed to be analytic within and on a closed curve. Therefore, $\frac{f(z)}{z-a}$ has a simple pole at $z = a$. Then

$$\text{Res}(a) = \lim_{z\to a} (z-a)\frac{f(z)}{(z-a)} = f(a).$$

Hence, by Cauchy's Residue theorem, we have

$$\int_C \frac{f(z)}{z-a}\,dz = 2\pi i\; f(a),$$

that is,

$$f(a) = \frac{1}{2\pi i}\int_C \frac{f(z)}{z-a}\,dz.$$

It follows, therefore, that Cauchy's integral formula is a particular case of Cauchy's Residue theorem.

EXAMPLE 1.82

Evaluate

$$\int_C \frac{dz}{z^2(z+1)(z-1)}, \quad C: |z| = 3.$$

Solution. The integrand has simple poles at $z = 1$ and $z = -1$ and double poles at $z = 0$ lying in C. Therefore,

$$\text{Res}(1) = \lim_{z\to 1}(z-1) f(z)$$

$$= \lim_{z\to 1} \frac{1}{z^2(z+1)} = \frac{1}{2},$$

$$\text{Res}(-1) = \lim_{z\to -1}(z+1) f(z)$$

$$= \lim_{z\to -1} \frac{1}{z^2(z-1)} = -\frac{1}{2},$$

$$\text{Res}(0) = \lim_{z\to 0} \frac{d}{dz}[(z-0)^2 f(z)]$$

$$= \lim_{z\to 0} \frac{d}{dz}\left[\frac{1}{(z+2)(z-1)}\right]$$

$$= \lim_{z\to 0} \frac{-2z-1}{(z^2+z-2)^2} = \frac{-1}{4}.$$

Hence, by Cauchy-Residue theorem,

$$\int_C \frac{dz}{z^2(z+1)(z-1)} = 2\pi i\left[\frac{1}{2}-\frac{1}{2}-\frac{1}{4}\right] = -\frac{\pi i}{2}.$$

EXAMPLE 1.83

Evaluate the integral

$$\int_{|z|=1} \frac{4z^2-4z+1}{(z-2)(4+z^2)}\,dz.$$

Solution. Let

$$f(z) = \frac{4z^2-4z+1}{(z-2)(4+z^2)}.$$

The poles of $f(z)$ are $z = 2$, $z = \pm 2i$. We note that none of these poles lie in the curve $|z| = 1$. Thus the function is analytic within and on $|z| = 1$. Hence, by Cauchy-Goursat theorem,

$$\int_{|z|=1} \frac{4z^2-4z+1}{(z-2)(4+z^2)}\,dz = 0.$$

EXAMPLE 1.84

Evaluate $\int_C \frac{dz}{(z^2+4)^2}$, where C is the curve $|z-i| = 2$.

Solution. Let

$$f(z) = \frac{1}{(z^2+4)^2} = \frac{1}{(z+2i)^2\,(z-2i)^2}.$$

Thus $f(z)$ has two double poles at $z = 2i$ and $z = -2i$, out of which only $z = 2i$ lies within $|z-i| = 2$. Now

$$\begin{aligned}\text{Res}(2i) &= \lim_{z\to 2i} \frac{d}{dz}[(z-2i)^2.f(z)]\\ &= \lim_{z\to 2i} \frac{d}{dz}\left[\frac{1}{(z+2i)^2}\right]\\ &= \lim_{z\to 2i}\left[\frac{-2z-4i}{(z+2i)^4}\right] = \frac{-i}{32}.\end{aligned}$$

Hence, by Cauchy's Residue theorem, we have

$$\int_C \frac{dz}{(z^2+4)^2} = 2\pi i\left(-\frac{i}{32}\right) = \frac{\pi}{16}.$$

EXAMPLE 1.85

Evaluate

$$\int_C \frac{1-\cos 2(z-3)}{(z-3)^3}\,dz,$$

where C is the curve $|z-3| = 1$.

Solution. Expanding $\cos 2(z-3)$, we have

$$\begin{aligned} f(z) &= \frac{1-\cos 2(z-3)}{(z-3)^3}\\ &= \frac{1}{(z-3)^3}\left[1-1+\frac{4(z-3)^2}{2!}-\frac{16(z-3)^4}{4!}+\ldots\right]\\ &= \frac{2}{z-3}-\frac{16}{4!}(z-3)+\ldots.\end{aligned}$$

Thus $f(z)$ has a simple pole at $z = 3$. The Laurent's series is in the power of $z-3$. The coefficient of $\frac{1}{z-3}$ is 2. Hence, the residue of $f(z)$ at $z = 3$ is 2 and so by Cauchys Residue theorem

$$\int_C f(z)dz = 2\pi i(2) = 4\pi i.$$

EXAMPLE 1.86

Evaluate

$$\int_{|z|=3} \frac{\sin \pi z^2+\cos \pi z^2}{(z-1)^2(z-2)}\,dz.$$

Solution. The integrand has simple pole at $z = 2$ and a pole of order 2 at $z = 1$. But these poles lie within $|z| = 3$. Now

/)*+,)))-) # . / + 0
0" 5 3" & 3
(# :

3 Curvature

Let P be any point on a given curve and Q a neighboring point of P such that the arc PQ is concave towards its chord. Let the normals at P and Q intersects at N.

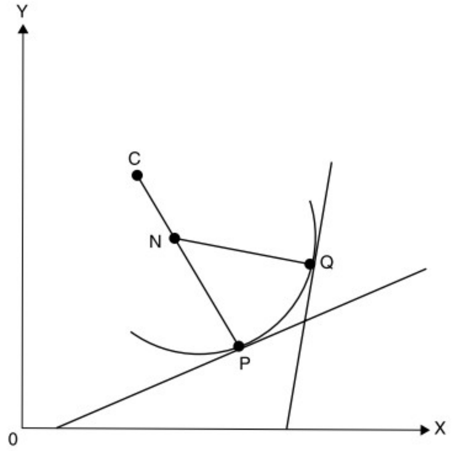

When Q → P, N tends to a definite position C, called the *center of curvature* of the curve at P. The distance CP is called the *radius of curvature* of the curve at the point P and is denoted by ρ. The circle with center at C and the radius ρ, equal to CP, is called the *circle of curvature* of the given curve at the point P. Any chord of the circle of curvature drawn through the point P is called the *chord of curvature*. The reciprocal of the radius of curvature is called the *curvature of the curve* at the point P and is denoted by K.

3.1 RADIUS OF CURVATURE OF INTRINSIC CURVES

Let A be a fixed point on a given curve and P and Q be two neighboring points on the curve. Let arcs AP and PQ be s and δs, respectively. Let the tangents at P and Q make angles ψ and $\psi + \delta\psi$ with the x-axis so that the angle between the tangents at P and Q is $\delta\psi$. Let R be the point of intersection of the tangents at P and Q and let N be the intersection of normals at P and Q. We have

$$\angle PNQ = \angle SRT = \delta\psi.$$

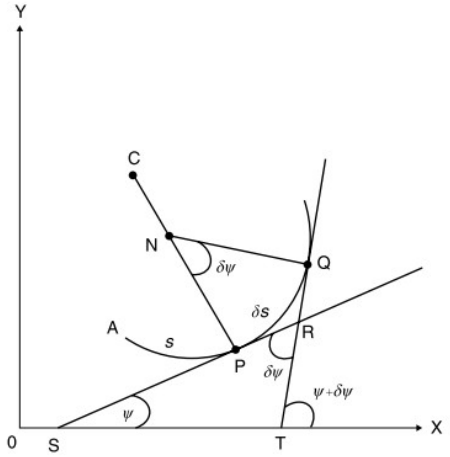

From the triangle PNQ,

$$\frac{PN}{\sin\angle NQP} = \frac{\text{chord } PQ}{\sin\angle PNQ} = \frac{\text{chord } PQ}{\sin\delta\psi}$$

and so,

$$\begin{aligned} PN &= \frac{\text{chord } PQ}{\sin\delta\psi}\cdot\sin\angle NQP \\ &= \frac{\text{chord } PQ}{\delta s}\frac{\delta s}{\delta\psi}\frac{\delta\psi}{\sin\delta\psi}\sin\angle NQP. \end{aligned}$$

If Q→P, $\delta\psi\to 0$, and $\delta s\to 0$, and the chord PQ tends to the tangent at P and QN tends to the normal at P. Consequently, $\angle NQP\to\frac{\pi}{2}$.

Suppose N approaches C as Q approaches P. Then the radius of curvature at P is given by

$$\rho = \lim_{Q\to P} \mathrm{PN}$$

$$= \lim_{Q\to P}\left(\frac{\text{chord PQ}}{\text{arc PQ}}\right)\left(\lim_{\delta\psi\to 0}\frac{\delta s}{\delta\psi}\right).\left(\lim_{\delta\psi\to 0}\frac{\delta\psi}{\sin\delta\psi}\right)$$

$$\times\left(\lim_{Q\to P}\sin\angle \mathrm{NQP}\right)$$

$$= 1.\frac{ds}{d\psi}.1.\sin\frac{\pi}{2} = \frac{ds}{d\psi}.$$

Consequently,

$$\mathrm{K} = \frac{1}{\rho} = \frac{d\psi}{ds}.$$

Since the relation between arc length s and ψ is called the *intrinsic equation of a curve*, the formula $\rho = \frac{ds}{d\psi}$ is known as the *intrinsic formula for the radius of curvature* of the curve. Since the unit of $\delta\psi$ is radian, the unit of curvature is radian per unit length.

3.2 RADIUS OF CURVATURE FOR CARTESIAN CURVES

Let the equation of the curve be $y = f(x)$. Let ψ be the angle which the tangent at any point (x,y) makes with the x-axis. Then

$$\tan\psi = \frac{dy}{dx}.$$

Differentiating with respect to s, we have

$$\sec^2\psi\frac{d\psi}{ds} = \frac{d}{ds}\left(\frac{dy}{dx}\right) = \frac{d}{dx}\left(\frac{dy}{dx}\right).\frac{dx}{ds}$$

or

$$\frac{1}{\rho}\sec^2\psi = \frac{d^2y}{dx^2}.\frac{dx}{ds} = \frac{d^2y}{dx^2}\cos\psi,\ \text{since}\ \frac{dx}{ds} = \cos\psi$$

or

$$\rho = \frac{\sec^3\psi}{\frac{d^2y}{dx^2}} = \frac{(1+\tan^2\psi)^{\frac{3}{2}}}{\frac{d^2y}{dx^2}} = \frac{\left[1+\left(\frac{dy}{dx}\right)^2\right]^{\frac{3}{2}}}{\frac{d^2y}{dx^2}} = \frac{(1+y_1^2)^{\frac{3}{2}}}{y_2}.$$

Remark 3.1.

(i) If the tangent at the point is parallel to the y-axis, then $\frac{dy}{dx}$ is infinite and so the above formula for ρ does not hold good in that case.

(ii) The value ρ is positive or negative in accordance with y_2 being positive or negative. However, we will ignore the negative sign as we shall be interested only in finding the length of the radius of curvature.

(iii) The value of ρ depends on the curve and not on the coordinate axes chosen. Therefore, interchanging x and y in the above formula, we get

$$\rho = \frac{\left[1+\left(\frac{dx}{dy}\right)^2\right]^{\frac{3}{2}}}{\frac{d^2x}{dy^2}},$$

the formula which is used when $\frac{dy}{dx}$ is infinite, that is, when the tangent is parallel to the y-axis.

(iv) The point (x, y) on a curve is called a *point of inflexion* if $\frac{d^2y}{dx^2} = 0$ at that point.

(v) If the equation of the curve is given in an implicit form $f(x,y) = 0$, then since $\frac{dy}{dx} = \frac{-f_x}{f_y}$, $f_y \neq 0$, we have

$$\frac{d^2y}{dx^2} = -\frac{f_{xx}(f_y)^2 - 2f_x f_y f_{xy} + f_{yy}\,(f_x)^2}{(f_y)^3}.$$

Hence, the formula for radius of curvature reduces to

$$\rho = \frac{\left[1+\left(\frac{dy}{dx}\right)^2\right]^{\frac{3}{2}}}{\frac{d^2y}{dx^2}} = \frac{-\left[1+\left(\frac{-f_x}{f_y}\right)^2\right]^{\frac{3}{2}}(f_y)^3}{f_{xx}(f_y)^2 - 2f_xf_yf_{xy} + f_{yy}(f_x)^2}$$

$$= \frac{\left[(f_x)^2+(f_y)^2\right]^{\frac{3}{2}}}{f_{xx}(f_y)^2 - 2f_xf_yf_{xy} + f_{yy}(f_x)^2}\ \text{(in magnitude)}.$$

EXAMPLE 3.1

Find the radius of curvature at the point (s, ψ) on the following curves:

(i) $s = c\log\sec\psi$.

(ii) $s = a\log\tan\left(\frac{\pi}{4}+\frac{\psi}{2}\right)$.

(iii) $s = a\log\cot\left(\frac{\pi}{4}-\frac{\psi}{2}\right) + a\frac{\sin\psi}{\cos^2\psi}$.

(iv) $s^2 = 8ay$.

Solution.

(i) We have $s = c\log\sec\psi$. Therefore,

$$\rho = \frac{ds}{d\psi} = c.\frac{1}{\sec\psi}.\sec\psi\tan\psi = c\tan\psi.$$

(ii) We are given that

$$s = a\log\tan\left(\frac{\pi}{4}+\frac{\psi}{2}\right).$$

Therefore,

$$\begin{aligned}\rho &= \frac{ds}{d\psi}\\ &= a\frac{1}{\tan\left(\frac{\pi}{4}+\frac{\psi}{2}\right)}\cdot \sec^2\left(\frac{\pi}{4}+\frac{\psi}{2}\right)\cdot\frac{1}{2}\\ &= a\frac{\cos\left(\frac{\pi}{4}+\frac{\psi}{2}\right)}{\sin\left(\frac{\pi}{4}+\frac{\psi}{2}\right)}\cdot\frac{1}{\cos^2\left(\frac{\pi}{4}+\frac{\psi}{2}\right)}\cdot\frac{1}{2}\\ &= \frac{a}{2\,\sin\left(\frac{\pi}{4}+\frac{\psi}{2}\right)\cos\left(\frac{\pi}{4}+\frac{\psi}{2}\right)}\\ &= \frac{a}{\sin 2\left(\frac{\pi}{4}+\frac{\psi}{2}\right)}\\ &= \frac{a}{\sin\left(\frac{\pi}{2}+\psi\right)}\\ &= \frac{a}{\cos\psi}\\ &= a\ \sec\psi.\end{aligned}$$

(iii) Since $s = a\ \log\ \cot\left(\frac{\pi}{4}-\frac{\psi}{2}\right) + a\frac{\sin\psi}{\cos^2\psi}$, we have

$$\begin{aligned}\rho=\frac{ds}{d\psi}&=\frac{d}{d\psi}\left[a\log\cot\left(\frac{\pi}{4}-\frac{\psi}{2}\right)+a\sin\psi\sec^2\psi\right]\\ &=a.\frac{1}{\cot\left(\frac{\pi}{4}-\frac{\psi}{2}\right)}\left[-\mathrm{cosec}^2\left(\frac{\pi}{4}-\frac{\psi}{2}\right)\right]\left(-\frac{1}{2}\right)\\ &\quad+a\sin\psi(2\sec\psi\sec\psi\tan\psi)+a\cos\psi\sec^2\psi\\ &=\frac{a}{2\sin\left(\frac{\pi}{4}-\frac{\psi}{2}\right)\cos\left(\frac{\pi}{4}-\frac{\psi}{2}\right)}+\frac{2a\sin^2\psi}{\cos^3\psi}+\frac{a}{\cos\psi}\\ &=\frac{a}{\sin 2\left(\frac{\pi}{4}-\frac{\psi}{2}\right)}+\frac{2a\sin^2\psi}{\cos^3\psi}+\frac{a}{\cos\psi}\\ &=\frac{a}{\sin\left(\frac{\pi}{2}-\psi\right)}+\frac{2a\sin^2\psi}{\cos^3\psi}+\frac{a}{\cos\psi}\\ &=\frac{2a}{\cos\psi}+\frac{2a\sin^2\psi}{\cos^3\psi}\\ &=\frac{2a}{\cos^3\psi}(\cos^2\psi+\sin^2\psi)=2a\sec^3\psi.\end{aligned}$$

(iv) We are given that

$$s^2 = 8\,a\,y.$$

Differentiating with respect to s, we get

$$2s = 8a\frac{dy}{ds} = 8a\ \sin\psi, \text{since}\sin\psi = \frac{dy}{ds}$$

or

$$s = 4a\ \sin\psi.$$

Therefore,

$$\rho = \frac{ds}{d\psi} = 4a\cos\psi = 4a\sqrt{1-\sin^2\psi}$$

$$= 4a\sqrt{1-\frac{s^2}{16a^2}} = 4a\sqrt{1-\frac{8ay}{16a^2}} = 4a\sqrt{1-\frac{y}{2a}}.$$

EXAMPLE 3.2

Show that for any curve the following relation holds:

$$\frac{1}{\rho} = \frac{d}{dx}\left(\frac{dy}{ds}\right).$$

Solution. We know that

$$\frac{dy}{ds} = \sin\psi.$$

Therefore,

$$\begin{aligned}\frac{d}{dx}\left(\frac{dy}{ds}\right) &= \frac{d}{dx}[\sin\psi] = \cos\psi\frac{d\psi}{dx} = \cos\psi\frac{d\psi}{ds}\cdot\frac{ds}{dx}\\ &= \cos\psi.\frac{1}{\rho}.\sec\psi,\ \sec\psi = \frac{ds}{dx}\\ &= \frac{1}{\rho}, \text{which proves the result.}\end{aligned}$$

EXAMPLE 3.3

Find the radius of curvature of the curve, $s = a\ e^{\frac{x}{a}}$.

Solution. The equation of the curve is $s = a\ e^{\frac{x}{a}}$. Differentiating with respect to x, we get

$$\frac{ds}{dx} = a\ e^{\frac{x}{a}}.\frac{1}{a} = e^{\frac{x}{a}} = \frac{s}{a}$$

or

$$s = a\frac{ds}{dx} = a\ \sec\psi,\ \text{since}\ \sec\psi = \frac{ds}{dx}.$$

Therefore,

$$\begin{aligned}\rho &= \frac{ds}{d\psi}\\ &= a\ \sec\psi\tan\psi\\ &= s\tan\psi = s\sqrt{\sec^2\psi - 1}\\ &= s\sqrt{\frac{s^2}{a^2}-1} = \frac{s}{a}\sqrt{s^2-a^2}.\end{aligned}$$

EXAMPLE 3.4

Find the radius of curvature of the following curves:

(i) $x^{\frac{2}{3}} + y^{\frac{2}{3}} = a^{\frac{2}{3}}$ at (x, y).

(ii) $xy = c^2$ at (x, y).

(iii) $y = 4\sin x - \sin 2x$ at $x = \frac{\pi}{2}$.

(iv) $x^3 + y^3 = 3axy$ at the point $\left(\frac{3a}{2}, \frac{3a}{2}\right)$.

Solution.

(i) The equation of the given curve is

$$x^{\frac{2}{3}} + y^{\frac{2}{3}} = a^{\frac{2}{3}}.$$

Any point on this curve may be taken as ($a\cos^3 t$, $a\sin^3 t$) where t is a parameter. Then,

$$\frac{dx}{dt} = -3a\cos^2 t\sin t \text{ and } \frac{dy}{dt} = 3a\sin^2 t\cos t.$$

Therefore,

$$\frac{dy}{dx} = \frac{\frac{dy}{dt}}{\frac{dx}{dt}} = -\frac{3a\sin^2 t\ \cos t}{3a\cos^2 t\ \sin t} = -\tan t$$

and

$$\frac{d^2y}{dx^2} = \frac{d}{dx}(-\tan t) = \frac{d}{dt}(-\tan t)\frac{dt}{dx} = -\sec^2 t.\frac{dt}{dx}$$

$$= -\sec^2 t.\frac{1}{-3a\cos^2 t\ \sin t} = \frac{1}{3a}\sec^4 t\operatorname{cosec} t.$$

Hence ρ at the point ($a\cos^3 t$, $a\sin^3 t$) is given by

$$\rho = \frac{\left[1 + \left(\frac{dy}{dx}\right)^2\right]^{\frac{3}{2}}}{\frac{d^2y}{dx^2}} = \frac{3a\,(1 + \tan^2 t)^{\frac{3}{2}}}{\sec^4 t\operatorname{cosec} t}$$

$$= \frac{3a\sec^3 t}{\sec^4 t\operatorname{cosec} t} = 3a\sin t\cos t.$$

But $x = a\cos^3 t$ and $y = a\sin^3 t$ imply $\cos t = \left(\frac{x}{a}\right)^{\frac{1}{3}}$ and $\sin t = \left(\frac{y}{a}\right)^{\frac{1}{3}}$. Hence,

$$\rho = 3a\left(\frac{x}{a}\right)^{\frac{1}{3}}\left(\frac{y}{a}\right)^{\frac{1}{3}} = 3a^{\frac{1}{3}}x^{\frac{1}{3}}y^{\frac{1}{3}}.$$

(ii) The given curve is

$$xy = c^2 \quad \text{or} \quad y = \frac{c^2}{x}.$$

Therefore,

$$\frac{dy}{dx} = -\frac{c^2}{x^2} \text{ and } \frac{d^2y}{dx^2} = \frac{2c^2}{x^3}.$$

Hence,

$$\rho = \frac{\left[1 + \left(\frac{dy}{dx}\right)^2\right]^{\frac{3}{2}}}{\frac{d^2y}{dx^2}} = \frac{\left(1 + \frac{c^4}{x^4}\right)^{\frac{3}{2}}}{\frac{2c^2}{x^3}}$$

$$= \frac{(x^4 + c^4)^{\frac{3}{2}}}{2c^2x^3} = \frac{(x^4 + x^2y^2)^{\frac{3}{2}}}{2c^2x^3}, \text{ since } xy = c^2$$

$$= \frac{(x^2 + y^2)^{\frac{3}{2}}}{2c^2} = \frac{r^3}{2c^2}, \text{where } r^2 = x^2 + y^2.$$

(iii) The equation of the given curve is

$$y = 4\sin x - \sin 2x.$$

Therefore,

$$\frac{dy}{dx} = 4\cos x - 2\cos 2x \text{ and}$$

$$\frac{d^2y}{dx^2} = -4\sin x + 4\sin 2x$$

and so,

$$\left(\frac{dy}{dx}\right)_{x=\frac{\pi}{2}} = 4\cos\frac{\pi}{2} - 2\cos\pi = 2 \text{ and}$$

$$\left(\frac{d^2y}{dx^2}\right)_{x=\frac{\pi}{2}} = -4\sin\frac{\pi}{2} + 4\sin\pi = -4.$$

Hence,

$$\rho = \frac{\left[1 + \left(\frac{dy}{dx}\right)^2\right]^{\frac{3}{2}}}{\frac{d^2y}{dx^2}} = \frac{(1+4)^{\frac{3}{2}}}{-4}$$

$$= \frac{5\sqrt{5}}{4} \text{ (ignoring sign).}$$

(iv) The given curve is

$$x^3 + y^3 = 3axy.$$

Differentiating both sides with respect to x, we get

$$3x^2 + 3y^2\frac{dy}{dx} = 3a\left(x\frac{dy}{dx} + y\right)$$

or

$$(y^2 - ax)\frac{dy}{dx} = ay - x^2 \tag{1}$$

or

$$\frac{dy}{dx} = \frac{ay - x^2}{y^2 - ax}$$

so that,

$$\left(\frac{dy}{dx}\right)_{\left(\frac{3a}{2},\frac{3a}{2}\right)} = -1. \tag{2}$$

Further, differentiating (1) again, we get

$$(y^2 - ax)\frac{d^2y}{dx^2} + \left(2y\frac{dy}{dx} - a\right)\frac{dy}{dx} = a\frac{dy}{dx} - 2x$$

or

$$(ax - y^2)\frac{d^2y}{dx^2} = 2x + 2y\left(\frac{dy}{dx}\right)^2 - 2a\frac{dy}{dx}.$$

Therefore at the point $\left(\frac{3a}{2}, \frac{3a}{2}\right)$, we have

$$\left(\frac{d^2y}{dx^2}\right)_{\left(\frac{3a}{2},\frac{3a}{2}\right)} = -\frac{32}{3a}, \text{[using(2)]}.$$

Hence, the radius of curvature ρ at $\left(\frac{3a}{2}, \frac{3a}{2}\right)$ is

$$\rho = \frac{\left[1 + \left(\frac{dy}{dx}\right)^2\right]^{\frac{3}{2}}}{\frac{d^2y}{dx^2}} = \frac{(1+1)^{\frac{3}{2}}}{-\frac{32}{3a}}$$

$$= \frac{3a}{8\sqrt{2}} \text{(in magnitude)}.$$

EXAMPLE 3.5

Find the radius of curvature of the following curves:

(i) $y = c\cosh\left(\frac{x}{c}\right)$ at $(0, c)$.

(ii) $\sqrt{x} + \sqrt{y} = \sqrt{a}$ at (x, y).

Solution.

(i) The equation of the curve is

$$y = c\ \cosh\left(\frac{x}{c}\right).$$

Therefore,

$$\frac{dy}{dx} = c\sinh\left(\frac{x}{c}\right)\frac{1}{c} = \sinh\left(\frac{x}{c}\right)$$

and

$$\frac{d^2y}{dx^2} = \frac{1}{c}\cosh\left(\frac{x}{c}\right).$$

Hence, the radius of curvature is

$$\rho = \frac{\left[1 + \left(\frac{dy}{dx}\right)^2\right]^{\frac{3}{2}}}{\frac{d^2y}{dx^2}} = \frac{\left[1 + \sinh^2\left(\frac{x}{c}\right)\right]^{\frac{3}{2}}}{\frac{1}{c}\cosh\left(\frac{x}{c}\right)}$$

$$= \frac{c\left[\cosh^2\left(\frac{x}{c}\right)\right]^{\frac{3}{2}}}{\cosh\left(\frac{x}{c}\right)}$$

$$= c\cosh^2\left(\frac{x}{c}\right)$$

$$= c\left(\frac{y}{c}\right)^2 = \frac{y^2}{c}.$$

Thus, ρ at $(0, c)$ is

$$\rho = \frac{c^2}{c} = c.$$

(ii) The equation of the curve is

$$\sqrt{x} + \sqrt{y} = \sqrt{a}.$$

Differentiating with respect to x, we get

$$\frac{1}{2}x^{-\frac{1}{2}} + \frac{1}{2}y^{-\frac{1}{2}}\frac{dy}{dx} = 0,$$

which yields

$$\frac{dy}{dx} = -\frac{x^{-\frac{1}{2}}}{y^{-\frac{1}{2}}} = -y^{\frac{1}{2}}x^{-\frac{1}{2}}.$$

Differentiating again with respect to x, we have

$$\frac{d^2y}{dx^2} = -\left(\frac{1}{2}y^{-\frac{1}{2}}\frac{dy}{dx}\right)x^{-\frac{1}{2}} - y^{\frac{1}{2}}\left(-\frac{1}{2}x^{-\frac{3}{2}}\right)$$

$$= -\frac{1}{2}y^{-\frac{1}{2}}x^{-\frac{1}{2}}\left(-y^{\frac{1}{2}}x^{-\frac{1}{2}}\right) + \frac{1}{2}y^{\frac{1}{2}}x^{-\frac{3}{2}}$$

$$= \frac{1}{2x} + \frac{y^{\frac{1}{2}}}{2x^{\frac{3}{2}}} = \frac{1}{2x}\left(1 + \frac{y^{\frac{1}{2}}}{x^{\frac{1}{2}}}\right)$$

$$= \frac{1}{2x}\left(\frac{\sqrt{x} + \sqrt{y}}{\sqrt{x}}\right) = \frac{\sqrt{a}}{2x^{\frac{3}{2}}}.$$

Hence,

$$\rho = \frac{\left[1 + \left(\frac{dy}{dx}\right)^2\right]^{\frac{3}{2}}}{\frac{d^2y}{dx^2}} = \frac{\left[1 + \frac{y}{x}\right]^{\frac{3}{2}}}{\frac{\sqrt{a}}{2x^{\frac{3}{2}}}} = \frac{2(x+y)^{\frac{3}{2}}}{\sqrt{a}}.$$

EXAMPLE 3.6

Prove that for the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$, the radius of curvature $\rho = \frac{a^2b^2}{p^3}$, where p is the perpendicular from the center to the tangent at (x,y).

Solution. The equation of the ellipse is

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \text{ which implies } b^2x^2 + a^2y^2 = a^2b^2.$$

Differentiating with respect to x, we get

$$\frac{2x}{a^2} + \frac{2y}{b^2}\frac{dy}{dx} = 0$$

or

$$\frac{dy}{dx} = -\frac{b^2}{a^2}\left(\frac{x}{y}\right).$$

Differentiating again with respect to x, we have

$$\frac{d^2y}{dx^2} = -\frac{b^2}{a^2}\left(\frac{y - x\frac{dy}{dx}}{y^2}\right) = -\frac{b^2}{a^2y^2}\left[y - x\left(-\frac{b^2x}{a^2y}\right)\right]$$

$$= -\frac{b^2}{a^2y^2}\left[\frac{a^2y^2 + b^2x^2}{a^2y}\right] = -\frac{b^2}{a^2y^2}\left[\frac{a^2b^2}{a^2y}\right]$$

$$= -\frac{b^4}{a^2y^3}.$$

Therefore ρ at the point (x,y) is

$$\rho = \frac{\left[1 + \left(-\frac{b^2x}{a^2y}\right)^2\right]^{\frac{3}{2}}}{-\frac{b^4}{a^2y^3}}$$

$$= \frac{[a^4y^2 + b^4x^2]^{\frac{3}{2}}}{a^4b^4} \text{ (in magnitude).}$$

The equation of the tangent to the given ellipse at (x, y) is $\frac{Xx}{a^2} + \frac{Yy}{b^2} = 1$. Therefore, $p =$ length of the perpendicular from the center (0,0) to the tangent

$$= \frac{1}{\sqrt{\left(\frac{x}{a^2}\right)^2+\left(\frac{y}{b^2}\right)^2}} = \frac{a^2b^2}{\sqrt{b^4x^2+a^4y^2}}.$$

Thus,

$$p^3 = \frac{a^6b^6}{(b^4x^2+a^4y^2)^{\frac{3}{2}}}$$

so that,

$$(b^4x^2+a^4y^2)^{\frac{3}{2}} = \frac{a^6b^6}{p^3}.$$

Substituting this value in the expression for ρ, we get

$$\rho = \frac{\frac{a^6b^6}{p^3}}{a^4b^4} = \frac{a^2b^2}{p^3}.$$

3.3 RADIUS OF CURVATURE FOR PARAMETRIC CURVES

Let the parametric equation of the curve be

$$x = f(t) \text{ and } y = \phi(t).$$

Then if $x' = \frac{dx}{dt}$ and $y' = \frac{dy}{dt}$, we have

$$\frac{dy}{dx} = \frac{\frac{dy}{dt}}{\frac{dx}{dt}} = \frac{y'}{x'}$$

and

$$\frac{d^2y}{dx^2} = \frac{d}{dx}\left(\frac{dy}{dx}\right) = \frac{d}{dx}\left(\frac{y'}{x'}\right) = \left[\frac{d}{dt}\left(\frac{y'}{x'}\right)\right]\frac{dt}{dx}$$

$$= \frac{x'y''-y'x''}{(x')^2}\cdot\frac{1}{x'} = \frac{x'y''-y'x''}{(x')^3}.$$

Therefore,

$$\rho = \frac{\left[1+\left(\frac{dy}{dx}\right)^2\right]^{\frac{3}{2}}}{\frac{d^2y}{dx^2}}$$

$$= \frac{\left[1+\left(\frac{y'}{x'}\right)^2\right]^{\frac{3}{2}}}{\frac{x'y''-y'x''}{x'^3}}$$

$$= \frac{(x'^2+y'^2)^{\frac{3}{2}}}{x'y''-y'x''},\ x'y''-y'x'' \neq 0.$$

EXAMPLE 3.7

Solve Example 3.6 using parametric equation of the ellipse.

Solution. Any point on the ellipse $\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$ is $(a\cos\theta, b\sin\theta)$ and the parametric equation of the ellipse is

$$x = a\ \cos\theta \text{ and } y = b\ \sin\theta.$$

Therefore,

$$x' = -a\sin\theta \text{ and } y' = b\cos\theta,$$

$$x'' = -a\cos\theta \text{ and } y'' = -b\sin\theta.$$

Hence, ρ at $(a\cos\theta, b\sin\theta)$ is

$$\rho = \frac{(x'^2+y'^2)^{\frac{3}{2}}}{x'y''-y'x''} = \frac{(a^2\sin^2\theta+b^2\cos^2\theta)^{\frac{3}{2}}}{ab\ \sin^2\theta+ab\ \cos^2\theta}$$

$$= \frac{(a^2\sin^2\theta+b^2\cos^2\theta)^{\frac{3}{2}}}{ab}. \qquad (1)$$

Now the equation of tangent to the ellipse at $(a\cos\theta, b\sin\theta)$ is

$$\frac{xa\cos\theta}{a^2}+\frac{yb\sin\theta}{b^2} = 1 \text{ or } \frac{x}{a}\cos\theta+\frac{y}{b}\sin\theta = 1.$$

Therefore,

p = length of the perpendicular from the center $(0,0)$ on the tangent

$$= \frac{1}{\sqrt{\frac{\cos^2\theta}{a^2}+\frac{\sin^2\theta}{b^2}}} = \frac{ab}{\sqrt{a^2\sin^2\theta+b^2\cos^2\theta}}$$

Thus,

$$p^3 = \frac{a^3b^3}{(a^2\sin^2\theta+b^2\cos^2\theta)^{\frac{3}{2}}}.$$

Putting this value in (1), we get

$$\rho = \frac{a^2b^2}{p^3}.$$

EXAMPLE 3.8

Find the radius of curvature at any point θ of the curve

$$x = a(\theta-\sin\theta) \text{ and } y = a(1-\cos\theta).$$

Solution. The equation of the curve is

$$x = a(\theta-\sin\theta) \text{ and } y = a(1-\cos\theta).$$

Therefore,

$$x' = a(1-\cos\theta) = 2a\sin^2\frac{\theta}{2} \text{ and}$$

$$y' = a\sin\theta = 2a\sin\frac{\theta}{2}\cos\frac{\theta}{2}.$$

Thus,

$$\frac{dy}{dx} = \frac{y'}{x'} = \frac{2a\sin\frac{\theta}{2}\cos\frac{\theta}{2}}{2a\sin^2\frac{\theta}{2}} = \cot\frac{\theta}{2}.$$

Further,

$$\begin{aligned}\frac{d^2y}{dx^2} &= \frac{d}{dx}\left(\frac{dy}{dx}\right) = \frac{d}{dx}\left(\cot\frac{\theta}{2}\right)\\ &= \frac{d}{d\theta}\left(\cot\frac{\theta}{2}\right)\frac{d\theta}{dx}\\ &= -\frac{1}{2}\operatorname{cosec}^2\frac{\theta}{2}\cdot\frac{1}{2a\sin^2\frac{\theta}{2}}\\ &= -\frac{1}{4a}\operatorname{cosec}^4\frac{\theta}{2}.\end{aligned}$$

Hence,

$$\begin{aligned}\rho &= \frac{\left[1+\left(\frac{dy}{dx}\right)^2\right]^{\frac{3}{2}}}{\frac{d^2y}{dx^2}} = \frac{\left(1+\cot^2\frac{\theta}{2}\right)^{\frac{3}{2}}}{-\frac{1}{4}\operatorname{cosec}^4\frac{\theta}{2}}\\ &= -\frac{4a}{\operatorname{cosec}\frac{\theta}{2}} = 4a\sin\frac{\theta}{2} \text{ (in magnitude).}\end{aligned}$$

EXAMPLE 3.9

Show that the radius of curvature at any point of the cycloid

$x = a\,(t + \sin t)$ and $y = a(1 - \cos t)$ is $4a\cos\frac{t}{2}$.

Solution. The equation of the curve is

$$x = a(t + \sin t) \text{ and } y = a(1 - \cos t).$$

Therefore,

$$\begin{aligned}x' &= a(1+\cos t) \text{ and } y' = a\sin t,\\ x'' &= -a\sin t \text{ and } y'' = a\cos t.\end{aligned}$$

Hence,

$$\begin{aligned}\rho &= \frac{(x'^2+y'^2)^{\frac{3}{2}}}{x'y''-y'x''}\\ &= \frac{[a^2(1+\cos t)^2 + a^2\sin^2 t]^{\frac{3}{2}}}{a(1+\cos t)a\cos t - a\sin t(-a\sin t)}\\ &= \frac{a\,[2\,(1+\cos t)]^{\frac{3}{2}}}{1+\cos t} = (2\sqrt{2})a\,(1+\cos t)^{\frac{1}{2}}\\ &= (2\sqrt{2})a\left(2\cos^2\frac{t}{2}\right)^{\frac{1}{2}} = 4a\cos\frac{t}{2}.\end{aligned}$$

EXAMPLE 3.10

If ρ_1 and ρ_2 are the radii of curvatures at the extremities of a focal chord of the parabola $y^2 = 4ax$, prove that

$$\rho_1^{-\frac{2}{3}} + \rho_2^{-\frac{2}{3}} = (2a)^{-\frac{2}{3}}.$$

Solution. The parametric equation of the parabola $y^2 = 4ax$ is

$$x = at^2 \text{ and } y = 2at.$$

Therefore,

$$\begin{aligned}x' &= 2at \text{ and } y' = 2a,\\ x'' &= 2a \text{ and } y'' = 0.\end{aligned}$$

Therefore, ρ at the point $(at^2, 2at)$ is

$$\begin{aligned}\rho &= \frac{[x'^2+y'^2]^{\frac{3}{2}}}{x'\,y''-y'\,x''} = \frac{[4a^2t^2+4a^2]^{\frac{3}{2}}}{-4a^2}\\ &= \frac{8a^3(t^2+1)^{\frac{3}{2}}}{-4a^2} = 2a(1+t^2)^{\frac{3}{2}} \text{ (in magnitude).}\end{aligned}$$

Let the extremities of the focal chord be

$$P(at_1^2,\ 2at_1) \text{ and } Q\,(at_2^2,\ 2at_2).$$

Then, by the property of focal chord, $t_1t_2 = -1$. Further, by the above-obtained expression for ρ, we have

$$\begin{aligned}\rho_1 &= \text{radius of curvature at } \rho\,(at_1^2,\ 2at_1)\\ &= 2a\left(1+t_1^2\right)^{\frac{3}{2}} \text{ and}\\ \rho_2 &= \text{radius of curvature at } Q\left(at_2^2,\ 2at_2\right)\\ &= 2a\left(1+t_2^2\right)^{\frac{3}{2}}.\end{aligned}$$

Therefore, since $t_1^2\,t_2^2 = 1$, we have

$$\begin{aligned}\rho_1^{-\frac{2}{3}} + \rho_2^{-\frac{2}{3}} &= (2a)^{-\frac{2}{3}}\left[\frac{1}{1-t_1^2} + \frac{1}{1-t_2^2}\right]\\ &= (2a)^{-\frac{2}{3}}\left[\frac{t_1^2+t_2^2+2}{t_1^2+t_2^2+2}\right] = (2a)^{-\frac{2}{3}}.\end{aligned}$$

EXAMPLE 3.11

If ρ_1 and ρ_2 are the radii of curvatures at the extremities of two conjugate semi-diameters of an ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$, show that

$$\left(\rho_1^{\frac{2}{3}} + \rho_2^{\frac{2}{3}}\right)(ab)^{\frac{2}{3}} = a^2 + b^2.$$

Solution. The parametric equations of the ellipse are $x = a\cos\theta$ and $y = b\sin\theta$. As in Example 3.7, we have

$$\rho = \frac{(a^2\sin^2\theta + b^2\cos^2\theta)^{\frac{3}{2}}}{ab}$$

and so,

$$\rho ab = (a^2\sin^2\theta + b^2\cos^2\theta)^{\frac{3}{2}}$$

or

$$(ab)^{\frac{2}{3}}\,\rho^{\frac{2}{3}} = a^2\sin^2\theta + b^2\cos^2\theta.$$

If C is the center of the ellipse and CP and CQ be a pair of conjugate semi-diameters of the ellipse, then the eccentric angles of P and Q are θ and $\theta + \frac{\pi}{2}$, respectively. Thus if ρ is the radius of curvature at P $(a\cos\theta, b\sin\theta)$, then the curvature at Q is obtained by replacing θ by $\theta+\frac{\pi}{2}$. Thus, if ρ_1 and ρ_2 are the radii of curvatures at P and Q, then from the above expressions we get

$$\begin{aligned}(ab)^{\frac{2}{3}}\,\rho_1^{\frac{2}{3}} &= a^2\sin^2\theta+b^2\cos^2\theta\\ (ab)^{\frac{2}{3}}\,\rho_2^{\frac{2}{3}} &= a^2\sin^2\left(\theta+\frac{\pi}{2}\right)+b^2\cos^2\left(\theta+\frac{\pi}{2}\right)\\ &= a^2\cos^2\theta+b^2\sin^2\theta.\end{aligned}$$

Adding these two relations, we get

$$\left(\rho_1^{\frac{2}{3}}+\rho_2^{\frac{2}{3}}\right)(ab)^{\frac{2}{3}} = a^2(\sin^2\theta+\cos^2\theta) + b^2(\sin^2\theta+\cos^2\theta) = a^2+b^2.$$

EXAMPLE 3.12

Show that the radius of curvature at a point P on an ellipse $\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$ is given by $\rho=\frac{(CQ)^3}{ab}$, where CQ is the semi-diameter conjugate to CP.

Solution. If $P(a\cos\theta,\ b\sin\theta)$ is the point on the ellipse, then the point Q shall be $Q\left(a\cos\left(\theta+\frac{\pi}{2}\right), b\sin\left(\theta+\frac{\pi}{2}\right)\right)=Q(-a\sin\theta,\ b\cos\theta)$. Therefore,

$$CQ=\sqrt{a^2\sin^2\theta+b^2\cos^2\theta}\quad\text{or}$$

$$(CQ)^3=(a^2\sin^2\theta+b^2\cos^2\theta)^{\frac{3}{2}}.$$

Hence, from Example 3.7, we have

$$\rho=\frac{(a^2\sin^2\theta+b^2\cos^2\theta)^{\frac{3}{2}}}{ab}=\frac{(CQ)^3}{ab}.$$

EXAMPLE 3.13

Show that the radius of curvature at the end of the major axis of the ellipse $\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$ is equal to the semi-latus rectum of the ellipse.

Solution. As in the above examples, we have

$$\rho=\frac{(a^2\sin^2\theta+b^2\cos^2\theta)^{\frac{3}{2}}}{ab}.$$

Now at the ends of the major axis, $\theta=0$ and π. Therefore, putting $\theta=0$, we get

$$\begin{aligned}\rho&=\frac{(b^2)^{\frac{3}{2}}}{ab}\\ &=\frac{b^2}{a}=\text{Semi-latus rectum of the ellipse.}\end{aligned}$$

3.4 RADIUS OF CURVATURE FOR PEDAL CURVES

Let $P(r,\theta)$ be a given point on the curve such that the tangent to the curve at P makes an angle ψ with the x-axis. Let ϕ be the angle between the tangent and the radius vector. Then $\psi=\theta+\phi$.

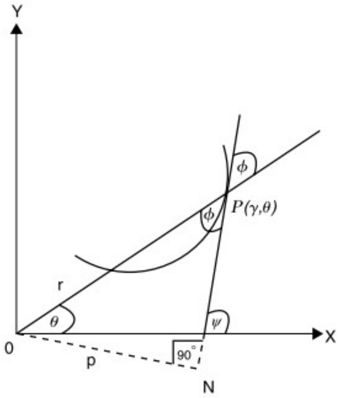

Differentiating with respect to s, we get

$$\frac{d\psi}{ds}=\frac{d\theta}{ds}+\frac{d\phi}{ds}. \tag{1}$$

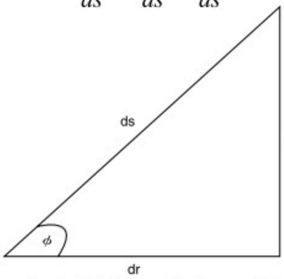

Also $p=r\sin\phi$. Differentiating with respect to r, we get

$$\begin{aligned}\frac{dp}{dr}&=\sin\phi+r\cos\phi\,\frac{d\phi}{dr}=r\frac{d\theta}{ds}+r\frac{dr}{ds}\cdot\frac{d\phi}{dr}\\ &=r\left[\frac{d\theta}{ds}+\frac{d\phi}{ds}\right]=r\frac{d\psi}{ds}=\frac{r}{\rho},\ \rho=\frac{ds}{d\psi}.\end{aligned}$$

Hence,

$$\rho=r\frac{dr}{dp}.$$

3.5 RADIUS OF CURVATURE FOR POLAR CURVES

From the figure of article 3.4, we have $\psi=\theta+\phi$ and so differentiation yields

$$\frac{d\psi}{ds}=\frac{d\theta}{ds}+\frac{d\phi}{ds}$$

or

$$\frac{1}{\rho}=\frac{d\theta}{ds}+\frac{d\phi}{d\theta}\cdot\frac{d\theta}{ds}. \tag{1}$$

But,

$$\tan\phi = r\frac{d\theta}{dr} = \frac{\frac{r}{dr}}{d\theta} = \frac{r}{r_1}, \text{ where } r_1 = \frac{dr}{d\theta}.$$

Differentiating with respect to θ, we get

$$\sec^2\phi\frac{d\phi}{d\theta} = \frac{r_1 r_1 - r.r_2}{{r_1}^2}, \text{ where } r_2 = \frac{d^2r}{d\theta^2}.$$

Therefore,

$$\frac{d\phi}{d\theta} = \frac{{r_1}^2 - rr_2}{{r_1}^2\sec^2\phi} = \frac{{r_1}^2 - rr_2}{{r_1}^2(1+\tan^2\phi)}$$

$$= \frac{{r_1}^2 - rr_2}{{r_1}^2\left(1+\frac{r^2}{{r_1}^2}\right)}, \text{since}$$

$$\tan\phi = \frac{rd\theta}{dr} = \frac{r}{r_1}.$$

$$= \frac{{r_1}^2 - rr_2}{{r_1}^2 + r^2}.$$

Moreover, since $\sin\phi = r\frac{d\theta}{ds}$, we have

$$\frac{d\theta}{ds} = \frac{\sin\phi}{r} = \frac{1}{r}\cdot\frac{r}{\sqrt{r^2+{r_1}^2}} = \frac{1}{\sqrt{r^2+{r_1}^2}}.$$

Hence from (1), we have

$$\frac{1}{\rho} = \frac{1}{\sqrt{r^2+{r_1}^2}}\left[1+\frac{{r_1}^2-rr_2}{{r_1}^2+r^2}\right] = \frac{r^2-rr_2+2{r_1}^2}{(r^2+{r_1}^2)^{\frac{3}{2}}},$$

which yields

$$\rho = \frac{(r^2+{r_1}^2)^{\frac{3}{2}}}{r^2-rr_2+2{r_1}^2}.$$

3.5.1 Second Method

The relation between p, r, and θ is

$$\frac{1}{p^2} = \frac{1}{r^2} + \frac{1}{r^4}{r_1}^2, \; r_1 = \frac{dr}{d\theta}.$$

Differentiating with respect to r, we get

$$-\frac{2}{p^3}\frac{dp}{dr} = -\frac{2}{r^3} - \frac{4}{r^5}{r_1}^2 + \frac{2}{r^4}r_1\frac{d}{dr}r_1$$

$$= -\frac{2}{r^3} - \frac{4}{r^5}{r_1}^2 + \frac{2}{r^4}r_1\left[\frac{d}{d\theta}\left(\frac{dr}{d\theta}\right)\right]\frac{d\theta}{dr}$$

$$= -\frac{2}{r^3} - \frac{4}{r^5}{r_1}^2 + \frac{2}{r^4}r_1\cdot r_2\cdot\frac{1}{r_1}$$

$$= -\frac{2}{r^3} - \frac{4}{r^5}{r_1}^2 + \frac{2}{r^4}r_2.$$

Therefore,

$$\frac{1}{p^3}\frac{dp}{dr} = \frac{1}{r^3} + \frac{2}{r^5}r_1^2 - \frac{1}{r^4}r_2 = \frac{r^2+2{r_1}^2-rr_2}{r^5}. \quad (1)$$

Since,

$$\frac{1}{p^3} = \left[\left(\frac{1}{r^2}+\frac{1}{r^4}{r_1}^2\right)^2\right]^{\frac{3}{2}} \text{ and } \frac{dp}{dr} = \frac{r}{\rho},$$

we have,

$$\rho = \frac{r^6\left[\left(\frac{1}{r^2}+\frac{1}{r^4}{r_1}^2\right)^2\right]^{\frac{3}{2}}}{r^2+2{r_1}^2-rr_2} == \frac{(r^2+{r_1}^2)^{\frac{3}{2}}}{r^2+2{r_1}^2-rr_2}.$$

Corollary: Let the equation of the curve be $\frac{1}{r} = f(\theta)$. Put $u = \frac{1}{r}$ or $r = \frac{1}{u}$. Then,

$$\frac{dr}{d\theta} = -\frac{1}{u^2}\cdot\frac{du}{d\theta} \quad \text{and}$$

$$\frac{d^2r}{d\theta^2} = -\frac{1}{u^2}\cdot\frac{d^2u}{d\theta^2} + \frac{2}{u^3}\left(\frac{du}{d\theta}\right)^2.$$

Therefore,

$$\rho = \frac{\left[r^2+\left(\frac{dr}{d\theta}\right)^2\right]^{\frac{3}{2}}}{r^2+2\left(\frac{dr}{d\theta}\right)^2 - r\frac{d^2r}{d\theta^2}}$$

$$= \frac{\left[\frac{1}{u^2}+\frac{1}{u^4}\left(\frac{du}{d\theta}\right)^2\right]^{\frac{3}{2}}}{\frac{1}{u^2}+2.\frac{1}{u^4}\left(\frac{du}{d\theta}\right)^2 - \frac{1}{u}\left[-\frac{1}{u^2}\frac{d^2u}{d\theta^2}+\frac{2}{u^3}\left(\frac{du}{d\theta}\right)^2\right]}$$

$$= \frac{\frac{1}{u^6}\left[u^2+\left(\frac{du}{d\theta}\right)^2\right]^{\frac{3}{2}}}{\frac{1}{u^2}+\frac{1}{u^3}\frac{d^2u}{d\theta^2}} = \frac{\left[u^2+\left(\frac{du}{d\theta}\right)^2\right]^{\frac{3}{2}}}{u^2\left(u+\frac{d^2u}{d\theta^2}\right)}$$

$$= \frac{[u^2+u_1^2]^{\frac{3}{2}}}{u^3(u+u_2)},$$

where

$$u_1 = \frac{du}{d\theta} \text{ and } u_2 = \frac{d^2u}{d\theta^2}.$$

EXAMPLE 3.14

Find the radius of curvature at the point (r, θ) of the following curves:

(i) $r^2\cos 2\theta = a^2$.

(ii) $r^n = a^n\sin n\theta$.

(iii) $r = a(1-\cos\theta)$.

Solution.

(i) The equation of the curve is

$$r^2\cos 2\theta = a^2. \quad (1)$$

Taking logarithm, we get

$$2\log r + \log\cos 2\theta = \log a^2. \quad (2)$$

Differentiating (2) with respect to θ, we get

$$\frac{2}{r}\frac{dr}{d\theta} + \frac{-2\sin 2\theta}{\cos 2\theta} = 0 \quad \text{or} \quad r_1 = \frac{dr}{d\theta} = r\tan 2\theta.$$

Differentiating again with respect to θ, we get

$$r_2 = \frac{d^2r}{d\theta^2} = 2r\sec^2 2\theta + \tan 2\theta \frac{dr}{d\theta}$$
$$= 2r\sec^2 2\theta + r\tan^2 2\theta.$$

Hence,

$$\rho = \frac{(r^2 + {r_1}^2)^{\frac{3}{2}}}{r^2 + 2{r_1}^2 - rr_2}$$
$$= \frac{(r^2 + r^2\tan^2 2\theta)^{\frac{3}{2}}}{r^2 + 2r^2\tan^2 2\theta - (2r^2\sec^2\theta + r^2\tan^2 2\theta)}$$
$$= -\frac{r^3\sec^3 2\theta}{r^2\sec^2 2\theta} = -r\sec 2\theta = -r\left(\frac{r^2}{a^2}\right), \text{using}(1)$$
$$= \frac{r^3}{a^2} \text{ (in magnitude)}.$$

(ii) The given curve is

$$r^n = a^n \sin\ n\theta \tag{1}$$

Taking log, we get

$$n\log r = n\log a + \log\sin n\theta.$$

Differentiating with respect to θ, we get

$$\frac{n}{r}\frac{dr}{d\theta} = n\cot n\theta.$$

Therefore,

$$r_1 = \frac{dr}{d\theta} = r\cot n\theta.$$

Differentiating once more, we get

$$r_2 = \frac{d^2r}{d\theta^2} = -nr\,\text{cosec}^2 n\theta + \cot n\theta\frac{dr}{d\theta}$$
$$= -nr\,\text{cosec}^2 n\theta + r\cot^2 n\theta.$$

Hence,

$$\rho = \frac{(r^2 + r_1^2)^{\frac{3}{2}}}{r^2 + 2r_1^2 - rr_2}$$
$$= \frac{(r^2 + r^2\cot^2 n\theta)^{\frac{3}{2}}}{r^2 + 2r^2\cot^2 n\theta - r(-nr\ \text{cosec}^2 n\theta + r\cot^2 n\theta)}$$
$$= \frac{r^3\,\text{cosec}^3 n\theta}{r^2 + r^2\cot^2 n\theta + nr^2\,\text{cosec}^2 n\theta}$$
$$= \frac{r^3\text{cosec}^3 n\theta}{r^2(1 + \cot^2 n\theta) + nr^2\,\text{cosec}^2 n\theta}$$
$$= \frac{r^3\,\text{cosec}^3 n\theta}{r^2\text{cosec}^2 n\theta(1 + n)}$$
$$= \frac{r\,\text{cosec}\ n\theta}{n + 1}$$
$$= \frac{a^n}{(n + 1)r^{n-1}}, \text{using}(1).$$

(iii) The equation of the given curve is

$$r = a(1 - \cos\theta). \tag{1}$$

Differentiating with respect to θ, we get

$$r_1 = \frac{dr}{d\theta} = a\sin\theta \quad \text{and} \quad r_2 = \frac{d^2r}{d\theta^2} = a\cos\theta.$$

Therefore,

$$\rho = \frac{(r^2 + r_1^2)^{\frac{3}{2}}}{r^2 + 2r_1^2 - rr_2}$$
$$= \frac{[a^2(1 - \cos\theta)^2 + a^2\sin^2\theta]^{\frac{3}{2}}}{a^2(1 - \cos\theta)^2 + 2a^2\sin^2\theta - a^2\cos\theta(1 - \cos\theta)},$$
using (1)
$$= \frac{a^3(1 + \cos^2\theta + \sin^2\theta - 2\cos\theta)^{\frac{3}{2}}}{a^2[1 + \cos^2\theta - 2\cos\theta + 2\sin^2\theta - \cos\theta + \cos^2\theta]}$$
$$= \frac{2^{\frac{3}{2}}a(1 - \cos\theta)^{\frac{3}{2}}}{3(1 - \cos\theta)} = \frac{2^{\frac{3}{2}}a\sqrt{r}}{3\sqrt{a}} = \frac{2}{3}\sqrt{2ar}.$$

It also follows that the radius of curvature at any point of the cardioid $r = a(1 - \cos\theta)$ varies as $\sqrt{r}$.

EXAMPLE 3.15

Show that the radius of curvature at any point of the cardioid $r = a(1 + \cos\theta)$ is $\frac{2}{3}\sqrt{2ar}$ and that $\frac{\rho^2}{r}$ is constant.

Solution. The equation of the curve is

$$r = a(1 + \cos\theta).$$

Therefore,

$$r_1 = \frac{dr}{d\theta} = -a\sin\theta \text{ and } r_2 = \frac{d^2r}{d\theta^2} = -a\cos\theta.$$

Therefore,

$$\rho = \frac{(r^2 + r_1^2)^{\frac{3}{2}}}{r^2 + 2r_1^2 - rr_2}$$
$$= \frac{[a^2(1 + \cos\theta)^2 + a^2\sin^2\theta]^{\frac{3}{2}}}{a^2(1 + \cos\theta)^2 + 2a^2\sin^2\theta + a^2\cos\theta(1 + \cos\theta)}$$
$$= \frac{a^3(1 + \cos^2\theta + \sin^2\theta + 2\cos\theta)^{\frac{3}{2}}}{a^2[1 + \cos^2\theta + 2\cos\theta + 2\sin^2\theta + \cos\theta + \cos^2\theta]}$$
$$= \frac{2^{\frac{3}{2}}a(1 + \cos\theta)^{\frac{3}{2}}}{3(1 + \cos\theta)} = \frac{2^{\frac{3}{2}}a\sqrt{r}}{3\sqrt{a}}$$
$$= \frac{2}{3}\sqrt{2ar} = \frac{4a}{3}\cos\frac{\theta}{2} \text{(in terms of } \theta\text{)}.$$

It follows that ρ varies as $\sqrt{r}$. Further,

$$\rho^2 = \frac{8ar}{9} \text{ or } \frac{\rho^2}{r} = \frac{8a}{9} \text{ (constant).}$$

EXAMPLE 3.16

Find the radius of curvature of the curve

$$\frac{l}{r} = 1 + e \cos\theta.$$

Solution. The equation of the given curve is

$$\frac{l}{r} = 1 + e \cos\theta.$$

Let $\frac{1}{r} = u$. Then the equation of the curve becomes

$$u = \frac{1}{l}(1 + e \cos\theta).$$

Therefore,

$$u_1 = \frac{du}{d\theta} = \frac{1}{l}(-e \sin\theta),$$

$$u_2 = \frac{d^2u}{d\theta^2} = \frac{1}{l}(-e \cos\theta).$$

Hence, the radius of curvature is

$$\begin{aligned}\rho &= \frac{(u^2 + u_1^2)}{u^3(u + u_2)}\\ &= \frac{\left(\frac{1}{l^3}\right)\left[(1 + e\cos\theta)^2 + e^2\sin^2\theta\right]^{\frac{3}{2}}}{\left(\frac{1}{l^4}\right)(1 + e\cos\theta)^3\left[(1 + e\cos\theta) - e\cos\theta\right]}\\ &= \frac{l\left[1 + e^2 + 2e\cos\theta\right]^{\frac{3}{2}}}{(1 + e\cos\theta)^3}.\end{aligned}$$

EXAMPLE 3.17

If ρ is the radius of curvature and $\tan\phi = r\frac{d\theta}{dr}$, show that

$$\frac{r}{\rho} = \sin\phi\left(1 + \frac{d\phi}{d\theta}\right) \text{ for any curve.}$$

Solution. We know that

$$\psi = \theta + \phi.$$

Differentiating with respect to s, we get

$$\begin{aligned}\frac{d\psi}{ds} = \frac{d\theta}{ds} + \frac{d\phi}{ds} &= \frac{d\theta}{ds} + \frac{d\phi}{d\theta}\cdot\frac{d\theta}{ds}\\ &= \frac{d\theta}{ds}\left(1 + \frac{d\phi}{d\theta}\right), \text{ that is,}\\ \frac{1}{\rho} &= \frac{\sin\phi}{r}\left(1 + \frac{d\phi}{d\theta}\right), \text{ since}\\ \rho &= \frac{ds}{d\psi} \text{ and } \sin\phi = r\frac{d\theta}{ds}.\end{aligned}$$

Hence,

$$\frac{r}{\rho} = \sin\phi\left(1 + \frac{d\phi}{d\theta}\right) \quad \text{so that} \quad \rho = \frac{r\,\mathrm{cosec}\,\phi}{1 + \frac{d\phi}{d\theta}}.$$

EXAMPLE 3.18

If ρ_1 and ρ_2 are the radii of curvatures at the extremities of any chord of the cardioid $r = a(1 + \cos\theta)$ which passes through the pole, show that

$$\rho_1^2 + \rho_2^2 = \frac{16a^2}{9}.$$

Solution. As in Example 3.15, the radius of curvature at any point of the cardioid $r = a(1 + \cos\theta)$ is

$$\rho = \frac{2}{3}\sqrt{2ar} \quad \text{so that} \quad \rho^2 = \frac{8}{9}ar.$$

If PQ is any chord of the cardioid passing through the pole and if P and Q are the points (r_1, θ_1) and (r_2, θ_2), respectively, then $\theta_2 = \pi + \theta_1$. Hence, if ρ_1 and ρ_2 are the radii of curvatures at P and Q, respectively, then

$$\rho_1{}^2 = \frac{8}{9}ar_1 \quad \text{and} \quad \rho_2{}^2 = \frac{8}{9}ar_2.$$

Therefore,

$$\begin{aligned}\rho_1{}^2 + \rho_2{}^2 &= \frac{8a}{9}[r_1 + r_2]\\ &= \frac{8a}{9}[a(1 + \cos\theta_1) + a(1 + \cos\theta_2)]\\ &= \frac{8a^2}{9}[2 + \cos\theta_1 + \cos\theta_2]\\ &= \frac{8a^2}{9}[2 + \cos\theta_1 + \cos(\pi + \theta_1)]\\ &= \frac{8a^2}{9}[2 + \cos\theta_1 - \cos\theta_1] = \frac{16a^2}{9}.\end{aligned}$$

EXAMPLE 3.19

Find the radius of curvature at any point (p, r) of the following curves:

(i) ellipse $\frac{1}{p^2} = \frac{1}{a^2} + \frac{1}{b^2} - \frac{r^2}{a^2b^2}$.

(ii) $pa^2 = r^3$.

(iii) $pa^n = r^{m+1}$

Solution.

(i) The equation of the ellipse is

$$\frac{1}{p^2} = \frac{1}{a^2} + \frac{1}{b^2} - \frac{r^2}{a^2b^2}.$$

Differentiating with respect to p, we get

$$-\frac{2}{p^3} = -\frac{2r}{a^2b^2}\cdot\frac{dr}{dp} \quad \text{or} \quad r\frac{dr}{dp} = \frac{a^2b^2}{p^3}.$$

Hence,

$$\rho = r\frac{dr}{dp} = \frac{a^2b^2}{p^3}.$$

(ii) The equation of the curve is

$$pa^2 = r^3.$$

Differentiating with respect to p, we have

$$a^2 = 3r^2\frac{dr}{dp} \quad \text{or} \quad r\frac{dr}{dp} = \frac{a^2}{3r}.$$

Hence,

$$\rho = r\frac{dr}{dp} = \frac{a^2}{3r}$$

(iii) The given equation is

$$pa^n = r^{m+1}.$$

Therefore, differentiation with respect to p yields

$$a^n = (m+1)r^m\frac{dr}{dp} \quad \text{or} \quad r\frac{dr}{dp} = \frac{a^n}{(m+1)r^{m-1}}.$$

Hence,

$$\rho = r\frac{dr}{dp} = \frac{a^n}{(m+1)r^{m-1}}.$$

Thus the radius of curvature varies inversely as the $(m-1)$th power of the radius vector.

EXAMPLE 3.20

Show that the radius of curvature of the equiangular spiral $r = ae^{\theta \cot\alpha}$ at any point (r, θ) is $\rho = r \operatorname{cosec} \alpha$.

Solution. The given curve is

$$r = ae^{\theta \cot\alpha}.$$

Differentiating with respect to θ, we have

$$r_1 = \frac{dr}{d\theta} = ae^{\theta\cot\alpha}\cdot\cot\alpha = r\cot\alpha \text{ and}$$

$$r_2 = \frac{d^2r}{d\theta^2} = r\cot^2\alpha.$$

Therefore,

$$\rho = \frac{(r+r_1^2)^{\frac{3}{2}}}{r^2+2r_1^2-rr_2} = \frac{(r^2+r^2\cot^2\alpha)^{\frac{3}{2}}}{r^2+2r^2\cot^2\alpha-r^2\cot^2\alpha}$$

$$= \frac{r^3(1+\cot^2\alpha)^{\frac{3}{2}}}{r^2(1+\cot^2\alpha)} = r(\operatorname{cosec}^2\alpha)^{\frac{1}{2}} = r\operatorname{cosec}\alpha.$$

3.6 RADIUS OF CURVATURE AT THE ORIGIN

The radius of curvature at the origin is determined by the following methods:

1. Newton's Method

Suppose that a curve passes through the origin and the axis of x is tangent at the origin. Then,

$$\text{slope of the tangent at the origin} = \left(\frac{dy}{dx}\right)_{(0,0)}$$
$$= y_1(0,0).$$

Since the tangent at the origin is x-axis, the slope of the tangent should be zero. Therefore $y_1(0,0) = 0$. The radius of curvature at the origin is

$$\rho \text{ at } (0,0) = \frac{[1+y_1{}^2(0,0)]^{\frac{3}{2}}}{y_2(0,0)} = \frac{1}{y_2(0,0)}. \qquad (1)$$

On the other hand,

$$\lim_{\substack{x\to 0\\ y\to 0}}\frac{x^2}{2y} = \lim_{\substack{x\to 0\\ y\to 0}}\frac{2x}{2y_1} \quad \left(\frac{0}{0}\text{ form}\right)$$

$$= \lim_{\substack{x\to 0\\ y\to 0}}\frac{1}{y_2} = \frac{1}{y_2(0,0)}. \qquad (2)$$

From (1) and (2), it follows that

$$\rho \; at \; (0,0) = \lim_{\substack{x\to 0\\ y\to 0}}\frac{x^2}{2y}.$$

Similarly, if the curve passes through the origin and y-axis is tangent to it at the origin, then

$$\rho \text{ at } (0,0) = \lim_{\substack{x\to 0\\ y\to 0}}\frac{y^2}{2x}.$$

If the initial line is tangent at the pole (origin), then

$$\rho = \lim_{\substack{x\to 0\\ y\to 0}}\frac{x^2}{2y} = \lim_{\substack{r\to 0\\ \theta\to 0}}\frac{r^2\cos^2\theta}{2r\sin\theta}$$

$$= \lim_{\substack{r\to 0\\ \theta\to 0}}\frac{r}{2\theta}\cdot\frac{\theta}{\sin\theta}\cdot\cos^2\theta = \lim_{\substack{r\to 0\\ \theta\to 0}}\frac{r}{2\theta}\left(\frac{0}{0}\text{form}\right)$$

$$= \lim_{\substack{r\to 0\\ \theta\to 0}}\frac{\frac{dr}{d\theta}}{2} = \frac{1}{2}\cdot\frac{dr}{d\theta}.$$

2. Method of Expansion

This method is used when neither x- nor y-axis is tangent to the curve at the origin. Suppose that the equation of the curve is $y = f(x)$. Since the curve passes through the origin, $f(0) = 0$. By Maclaurin's expansion, we have

$$y = f(0) + xf'(0) + \frac{x^2}{2!}f''(0) + \frac{x^3}{3!}f'''(0) + \ldots$$

$$= xf'(0) + \frac{x^2}{2!}f''(0) + \frac{x^3}{3!}f'''(0) + \ldots$$

$$= px + \frac{1}{2!}qx^2 + \frac{1}{3!}rx^3 + \ldots \qquad (1)$$

where,

$$p = f'(0) = y_1(0),$$
$$q = f''(0) = y_2(0),$$
$$r = f'''(0) = y_3(0), \text{ and so on.}$$

Differentiating (1) with respect to x, we get

$$y_1 = p + qx + \frac{1}{2}rx^2 + \dots \text{ and}$$
$$y_2 = q + rx + \dots .$$

Thus,

$$y_1(0,0) = p,\ y_2(0,0) = q.$$

Hence,

$$\rho \text{ at the origin} = \frac{[1 + y_1{}^2(0,0)]^{\frac{3}{2}}}{y_2(0)} = \frac{(1+p^2)^{\frac{3}{2}}}{q}.$$

EXAMPLE 3.21

Find the radius of curvature at the origin for the curve

$5x^3 + 7y^2x^2 + 4x^2y + xy^2 + 2x^2 + 3xy + y^2 + 4x = 0.$

Solution. The given curve passes through the origin. The lowest-degree term in the given equation is 4x. Therefore, the tangent at the origin is $4x = 0$ or $x = 0$, that is, the y-axis. Therefore, by Newton's Method,

$$\rho \text{ at the origin} = \lim_{\substack{x\to 0\\ y\to 0}} \frac{y^2}{2x}.$$

Dividing the equation of the curve throughout by $2x$, we get

$$\frac{5}{2}x^2 + \frac{7y^2x^2}{2x} + 2xy + \frac{y^2}{2} + x + \frac{3}{2}y + \frac{y^2}{2x} + 2 = 0.$$

Taking limit as $x \to 0$ and $y \to 0$, we get

$$\lim_{\substack{x\to 0\\ y\to 0}} \frac{y^2}{2x} + 2 = 0$$

or

$$\rho = -2 = 2\,(\text{numerically}).$$

EXAMPLE 3.22

Show that the radii of curvatures at the origin on the curve $x^3 + y^3 = 3axy$ is each equal to $\frac{3a}{2}$.

Solution. The equation of the given curve is

$$x^3 + y^3 = 3axy.$$

The curve passes through the origin. Further, the tangent at the origin is given by $3axy = 0$. Thus, $x = 0$ and $y = 0$ are the tangents at the origin. Therefore, the radii of curvatures are given by

$$\rho_1 \text{ at the origin } = \lim_{\substack{x\to 0\\ y\to 0}} \frac{x^2}{2y}.$$

and

$$\rho_2 \text{ at the origin } = \lim_{\substack{x\to 0\\ y\to 0}} \frac{y^2}{2x}.$$

Dividing the given equation throughout by 2xy, we get

$$\frac{x^2}{2y} + \frac{y^2}{2x} = \frac{3a}{2}.$$

Taking limit as $x \to 0$ and $y \to 0$, we have

$$\lim_{\substack{x\to 0\\ y\to 0}} \frac{x^2}{2y} + \lim_{\substack{x\to 0\\ y\to 0}} \frac{y^2}{2x} = \frac{3a}{2}$$

or

$$\lim_{\substack{x\to 0\\ y\to 0}} \frac{x^2}{2y} + \lim_{\substack{x\to 0\\ y\to 0}} \frac{1}{4}xy\frac{2y}{x^2} = \frac{3a}{2}$$

or

$$\rho_1 + 0\frac{1}{\rho_1} = \frac{3a}{2} \quad \text{or} \quad \rho_1 = \frac{3a}{2}.$$

Similarly, we can show that $\rho_2 = \frac{3a}{2}$.

EXAMPLE 3.23

Show that the radius of curvature of the curve

$$y^2 = x^2\frac{a+x}{a-x}$$

is $\pm a\sqrt{2}$ (or $a\sqrt{2}$ in magnitude).

Solution. The curve is

$$y^2(a-x) = x^2(a+x).$$

Obviously, the curve passes through the origin. But equating to zero the lowest-degree terms, we get

$$a(y^2 - x^2) = 0.$$

Therefore $y = \pm x$ is the tangent at the origin. Hence Newton's method cannot be applied to this curve. So let

$$y = px + \frac{1}{2!}qx^2 + \frac{1}{3!}rx^3 + \dots$$

Substituting this value of y in the given equation of the curve, we get

$$\left(px + \frac{1}{2}qx^2 + \frac{1}{3!}rx^3 + \dots\right)^2 (a-x) = x^2(a+x).$$

Equating coefficients of x^2 and x^3 on both sides, we get

$$ap^2 = a \text{ and } -p^2 + apq = 1,$$

which yield

$$p = 1, q = \frac{2}{a} \text{ and } p = -1, q = \frac{-2}{a}.$$

Therefore,

$$\rho_1 = \frac{(1+p^2)^{\frac{3}{2}}}{q} = \frac{2^{\frac{3}{2}}}{\frac{2}{a}} = \sqrt{2}a$$

and

$$\rho_2 = \frac{(1+1)^{\frac{3}{2}}}{-\frac{2}{a}} = -\sqrt{2}a = \sqrt{2}a \text{ (in magnitude).}$$

EXAMPLE 3.24

Find the radius of curvature at the origin of the cycloid

$$x = a(\theta + \sin\theta) \text{ and } y = a(1-\cos\theta).$$

Solution. The given curve is

$$x = a(\theta + \sin\theta) \text{ and } y = a(1-\cos\theta).$$

Therefore, differentiation with respect to θ yields

$$\frac{dx}{d\theta} = a(1+\cos\theta) \text{ and } \frac{dy}{d\theta} = a\sin\theta.$$

Then,

$$\frac{dy}{dx} = \frac{dy/d\theta}{dx/d\theta} = \frac{a\sin\theta}{a(1+\cos\theta)} = \frac{2\sin\frac{\theta}{2}\cos\frac{\theta}{2}}{2\cos^2\frac{\theta}{2}}$$
$$= \tan\frac{\theta}{2}.$$

Since $\theta = 0$ at $(0,0)$, we have $\left(\frac{dy}{dx}\right)_{(0,0)} = 0$. Therefore, x-axis is tangent to the curve at the origin and so, the Newton's method is applicable. Therefore, ρ at the origin is given by

$$\rho(0,0) = \lim_{\substack{x\to 0\\ y\to 0}} \frac{x^2}{2y}$$
$$= \lim_{\theta\to 0}\left[\frac{a^2(\theta+\sin\theta)^2}{2a(1-\cos\theta)}\right] \quad \left(\frac{0}{0}\text{form}\right)$$
$$= \lim_{\theta\to 0}\left[\frac{a\{2(\theta+\sin\theta)(1+\cos\theta)\}}{2\sin\theta}\right] \quad \left(\frac{0}{0}\text{form}\right)$$
$$= \lim_{\theta\to 0}\left[a\frac{(\theta+\sin\theta)(-\sin\theta)+(1+\cos\theta)^2}{\cos\theta}\right]$$
$$= 4a.$$

3.7 CENTRE OF CURVATURE

Let P(x, y) be any point on the curve and let PT be the tangent at P making an angle ψ with the positive direction of x-axis. Let $C(X, Y)$ be the center of curvature corresponding to P(x, y).

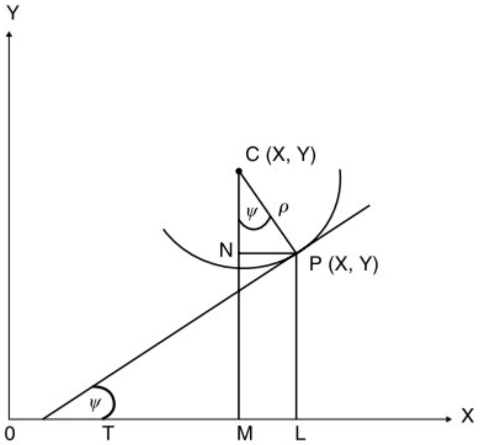

Draw PL and CM perpendicular to the x-axis from P and C, respectively. Draw PN ⊥ CM. Then

$$X = OM = OL - ML = OL - NP = x - \rho\sin\psi \quad (1)$$

and

$$Y = CM = NM + CN = \text{NM} + \rho\cos\psi$$
$$= y + \rho\cos\psi \quad (2)$$

But,

$$\tan\psi = \frac{dy}{dx} = y_1.$$

Therefore,

$$\sec\psi = \sqrt{1+\tan^2\psi}$$
$$= \sqrt{1+y_1^2} \text{ or}$$
$$\cos\psi = \frac{1}{\sqrt{1+y_1^2}}$$

and so,

$$\sin\psi = \sqrt{1-\cos^2\psi} = \sqrt{1-\frac{1}{1+y_1^2}} = \frac{y_1}{\sqrt{1+y_1^2}}.$$

Also,

$$\rho = \frac{(1+y_1^2)^{\frac{3}{2}}}{y_2}.$$

Hence (1) yields

$$X = x - \frac{(1+y_1^2)^{\frac{3}{2}}}{y_2}.\frac{y_1}{\sqrt{1+y_1^2}} = x - \frac{y_1(1+y_1^2)}{y_2}$$

and

$$Y = y + \frac{(1+y_1^2)^{\frac{3}{2}}}{y_2}.\frac{1}{\sqrt{1+y_1^2}} = y + \frac{1+y_1^2}{y_2}.$$

Also (1) and (2) can be expressed as

$$X = x - \rho \sin \psi = x - \frac{ds}{d\psi} \cdot \frac{dy}{ds} = x - \frac{dy}{d\psi} \text{ and}$$

$$Y = y + \rho \cos \psi = y + \frac{ds}{d\psi} \cdot \frac{dx}{ds} = y + \frac{dx}{d\psi}.$$

3.8 EVOLUTES AND INVOLUTES

The locus of the center of curvature for a curve is called the *evolute* of the curve and the curve itself is called the *involute* of its evolute.

3.9 EQUATION OF THE CIRCLE OF CURVATURE

Let (X,Y) be the coordinates of the center of curvature and ρ the radius of curvature at any point (x, y) of the curve. Then the equation of the circle of curvature is

$$(x - X)^2 + (y - Y)^2 = \rho^2.$$

EXAMPLE 3.25

Find the center of curvature of the following curves:

(i) $y = x^3 - 6x^2 + 3x + 1$ at (1,−1).

(ii) the parabola $y^2 = 4ax$ at (x, y). Also find the equation of the evolute of the given parabola.

(iii) the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ at (x, y). Also find the evolute.

Solution.

(i) The given curve is

$$y = x^3 - 6x^2 + 3x + 1.$$

Therefore,

$$y_1 = \frac{dy}{dx} = 3x^2 - 12x + 3 \text{ and so } y_1(1, -1) = -6,$$

$$y_2 = \frac{d^2y}{dx^2} = 6x - 12 \text{ and so } y_2(1, -1) = -6,$$

$$X = x - \frac{y_1(1 + y_1^2)}{y_2}$$

$$= 1 - \frac{(-6)[1 + (-6)^2]}{-6} = -36, \text{ and}$$

$$Y = -1 + \frac{1 + (-6)^2}{-6} = -\frac{43}{6}.$$

Hence, the centre of curvature is

$$(X, Y) = \left(-36, -\frac{43}{6}\right).$$

(ii) The equation of the given parabola is $y^2 = 4ax$. Therefore,

$$y_1 = \frac{2a}{y} = \frac{2a}{\sqrt{4ax}} = \sqrt{\frac{a}{x}} \text{ and } y_2 = -\frac{1}{2}\sqrt{a}\, x^{-\frac{3}{2}}.$$

Therefore,

$$X = x - \frac{\sqrt{\frac{a}{x}}\left(1 + \frac{a}{x}\right)}{-\frac{1}{2}\sqrt{a}\, x^{-\frac{3}{2}}}$$

$$= x + 2(x + a)$$

$$= 3x + 2a \tag{1}$$

and

$$Y = y + \frac{1 + \frac{a}{x}}{-\frac{1}{2}\sqrt{a}\, x^{-\frac{3}{2}}}$$

$$= 2\sqrt{a}\sqrt{x} - \frac{2(x + a)}{\sqrt{a}\, x^{-\frac{1}{2}}}$$

$$= 2\sqrt{a}\sqrt{x}\left(1 - \frac{x + a}{a}\right)$$

$$= -2\frac{x^{\frac{3}{2}}}{\sqrt{a}}. \tag{2}$$

Hence, the center of curvature of the given parabola is

$$(X, Y) = \left(3x + 2a, -\frac{2x^{\frac{3}{2}}}{\sqrt{a}}\right).$$

From (1), we have $x = \frac{X - 2a}{3}$. Putting this value in (2), we get

$$Y = -\frac{2(X - 2a)^{\frac{3}{2}}}{\sqrt{a}} \text{ or } aY^2 = 4\left(\frac{X - 2a}{3}\right)^3$$

or

$$27aY^2 = 4(X - 2a)^3.$$

Therefore, the locus of the center of curvature (X, Y) is

$$27ay^2 = 4(x - 2a)^3,$$

which is the required evolute.

(iii) The equation of the ellipse is

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

Therefore,

$$y_1 = -\frac{b^2x}{a^2y} \text{ and } y_2 = -\frac{b^4}{a^2y^3}.$$

Therefore,

$$X = x - \frac{y_1(1 + y_1^2)}{y_2} = x - \frac{b^4x^2 + a^4y^2}{a^4b^2} \text{ and}$$

$$Y = y + \frac{1 + y_1^2}{y_2} = y + \frac{b^4x^2 + a^4y^2}{a^2b^4}.$$

Second Method

Equations of the ellipse are

$$x = a\cos\theta \text{ and } y = b\sin\theta.$$

Therefore,

$$\frac{dx}{d\theta} = -a\sin\theta \text{ and } \frac{dy}{d\theta} = b\cos\theta$$

and so,

$$y_1 = \frac{dy}{dx} = \frac{\frac{dy}{d\theta}}{\frac{dx}{d\theta}} = -\frac{b}{a}\cot\theta \text{ and}$$

$$y_2 = \frac{d^2y}{dx^2} = \frac{d}{dx}\left(\frac{dy}{dx}\right) = \left[\frac{d}{d\theta}\left(-\frac{b}{a}\cot\theta\right)\right]\frac{d\theta}{dx}$$

$$= \frac{b}{a}\operatorname{cosec}^2\theta\left(\frac{1}{-a\sin\theta}\right) = -\frac{b}{a^2}\operatorname{cosec}^3\theta.$$

Therefore,

$$X = x - \frac{y_1(1+y_1^2)}{y_2}$$

$$= x - \frac{-\frac{b}{a}\cot\theta\left(1+\frac{b^2}{a^2}\cot^2\theta\right)}{-\frac{b}{a^2}\operatorname{cosec}^3\theta}$$

$$= a\cos\theta - \frac{(a^2\sin^2\theta + b^2\cos^2\theta)}{a}\cos\theta$$

$$= \frac{a^2-b^2}{a}\cos^3\theta \qquad (1)$$

and

$$Y = y + \frac{1+y_1^2}{y_2} = y + \frac{1+\frac{b^2}{a^2}\cot^2\theta}{-\frac{b}{a^2}\operatorname{cosec}^3\theta}$$

$$= b\sin\theta - \frac{\sin\theta}{b}(a^2\sin^2\theta + b^2\cos^2\theta)$$

$$= \frac{1}{b}\sin\theta\,[b^2 - a^2\sin^2\theta - b^2\cos^2\theta]$$

$$= \frac{b^2-a^2}{b}\sin^3\theta = -\frac{a^2-b^2}{b}\sin^3\theta \qquad (2)$$

Hence the center of curvature is

$$(X, Y) = \left(\frac{a^2-b^2}{a}\cos^3\theta,\ -\frac{a^2-b^2}{b}\sin^3 t\right).$$

From (1) and (2), we have

$$aX = (a^2-b^2)\cos^3\theta \quad \text{or}$$

$$(aX)^{\frac{2}{3}} = (a^2-b^2)^{\frac{2}{3}}\cos^2\theta$$

and

$$bY = -(a^2-b^2)\sin^3\theta \quad \text{or}$$

$$(bY)^{\frac{2}{3}} = -(a^2-b^2)^{\frac{2}{3}}\sin^3\theta.$$

Adding both we get

$$(aX)^{\frac{2}{3}} + (bY)^{\frac{2}{3}} = (a^2-b^2)^{\frac{2}{3}}.$$

Therefore, the evolute of the ellipse is

$$(ax)^{\frac{2}{3}} + (by)^{\frac{2}{3}} = (a^2-b^2)^{\frac{2}{3}}.$$

EXAMPLE 3.26

Show that the evolute of the cycloid $x = a(\theta - \sin\theta)$ and $y = a(1-\cos\theta)$ lies on an equal cycloid.

Solution. For the given cycloid, we have

$$\frac{dx}{x\theta} = a\,(1-\cos\theta) \text{ and } \frac{dy}{d\theta} = a\sin\theta.$$

Therefore,

$$y_1 = \frac{dy}{dx} = \frac{\frac{dy}{d\theta}}{\frac{dx}{d\theta}} = \frac{a\sin\theta}{a(1-\cos\theta)} = \cot\frac{\theta}{2}$$

and

$$y_2 = \frac{d}{dx}(y_1) = \frac{d}{d\theta}(y_1)\frac{d\theta}{dx}$$

$$= -\frac{1}{2}\operatorname{cosec}^2\frac{\theta}{2}\cdot\frac{1}{a(1-\cos\theta)} = \frac{-1}{4a\sin^4\frac{\theta}{2}}.$$

Therefore,

$$X = x - \frac{y_1(1+y_1^2)}{y_2}$$

$$= a(\theta-\sin\theta) - \frac{\cot\frac{\theta}{2}\left[1+\cot^2\frac{\theta}{2}\right]}{-\frac{1}{4}\operatorname{cosec}^4\frac{\theta}{2}}$$

$$= a(\theta-\sin\theta) + 4a\cos\frac{\theta}{2}\sin\frac{\theta}{2}$$

$$= a\theta - a\sin\theta + 2a\sin\theta = a(\theta+\sin\theta) \qquad (1)$$

and

$$Y = y + \frac{1+y_1^2}{y_2}$$

$$= a(1-\cos\theta) - 4a\sin^2\frac{\theta}{2}$$

$$= a - a\cos\theta - 2a + 2a\cos\theta$$

$$= -a(1-\cos\theta). \qquad (2)$$

Hence, the center of curvature of the given cycloid is

$$(X, Y) = (a(\theta+\sin\theta) \text{ and } -a(1-\cos\theta)).$$

The locus of (*X*, *Y*) is

$$x = a(\theta+\sin\theta) \text{ and } y = -a(1-\cos\theta),$$

which is another equal cycloid.

EXAMPLE 3.27

Show that the evolute of the rectangular hyperbola $xy = c^2$ is the curve

$$(x+y)^{\frac{2}{3}} - (x-y)^{\frac{2}{3}} = (4c)^{\frac{2}{3}}.$$

Solution. We have

$$xy = c^2 \quad \text{or} \quad y = \frac{c^2}{x}.$$

Therefore,

$$y_1 = \frac{dy}{dx} = -\frac{c^2}{x^2} \text{ and } y_2 = \frac{d^2y}{dx^2} = \frac{2c^2}{x^3}.$$

Let (X, Y) be the coordinate of the center of curvature. Then,

$$\begin{aligned} X &= x - \frac{y_1(1+y_1^2)}{y_2} = x - \frac{\frac{-c^2}{x^2}\left(1+\frac{c^4}{x^4}\right)}{\frac{2c^2}{x^3}} \\ &= x + \frac{x^4+c^4}{2x^3} = \frac{2x^4+x^4+c^4}{2x^3} \\ &= \frac{3x^4+x^2y^2}{2x^3}, \text{ since } xy = c^2 \\ &= \frac{3x}{2} + \frac{y^2}{2x} \quad \text{and} \end{aligned}$$

$$\begin{aligned} Y &= y + \frac{1+y_1^2}{y_2} = y + \frac{1+\frac{c^4}{x^4}}{\frac{2c^2}{x^3}} = y + \frac{x^4+c^4}{2c^2x} \\ &= y + \frac{x^4+x^2y^2}{2x^2y}, \text{ since } xy = c^2 \\ &= y + \frac{x^2+y^2}{2y} = \frac{3y^2+x^2}{2y} = \frac{3}{2}y + \frac{x^2}{2y}. \end{aligned}$$

Hence, the center of curvature is

$$(X, Y) = \left(\frac{3x}{2} + \frac{y^2}{2x}, \frac{3y}{2} + \frac{x^2}{2y}\right).$$

Further, we note that

$$X + Y = \frac{1}{2xy}[x^3 + y^3 + 3x^2y + 3xy^2] = \frac{1}{2c^2}(x+y)^3$$

and so,

$$(X+Y)^{\frac{2}{3}} = \frac{1}{(2c^2)^{\frac{2}{3}}}(x+y)^2.$$

Similarly,

$$(X-Y)^{\frac{2}{3}} = -\frac{1}{(2c^2)^{\frac{2}{3}}}(x-y)^2.$$

Therefore,

$$\begin{aligned} (X+Y)^{\frac{2}{3}} - (X-Y)^{\frac{2}{3}} &= \frac{1}{(2c^2)^{\frac{2}{3}}}[(x+y)^2 + (x-y)^2] \\ &= \frac{1}{(2c^2)^{\frac{2}{3}}}(4xy) = \frac{1}{(2c^2)^{\frac{2}{3}}}(4c^2) \\ &= (4c)^{\frac{2}{3}}. \end{aligned}$$

The locus of (X, Y) is therefore,

$$(x+y)^{\frac{2}{3}} - (x-y)^{\frac{2}{3}} = (4c)^{\frac{2}{3}}.$$

Remark 3.2. The question can also be solved by taking a parametric equation of the curve as $x = ct$ and $y = \frac{c}{t}$, and then eliminating t from the expression for X and Y.

EXAMPLE 3.28

Show that the circle of curvature at the origin of the parabola $y = mx + \frac{x^2}{a}$ is $x^2 + y^2 = a(1 + m^2)(y - mx)$.

Solution. The equation of the given parabola is

$$y = mx + \frac{x^2}{a}.$$

Therefore,

$$y_1 = m + \frac{2x}{a} = \frac{am+2x}{a}$$

so that $y_1(0,0) = m$ and

$$y_2 = \frac{2}{a} \text{ so that } y_2(0,0) = \frac{2}{a}.$$

Let (X, Y) be the center of curvature at the origin. Then,

$$X = x - \frac{y_1(1+y_1^2)}{y_2} = 0 - \frac{m(1+m^2)}{\frac{2}{a}} = \frac{-am(1+m^2)}{2}$$

and

$$Y = y + \frac{1+y_1^2}{y_2} = 0 + \frac{1+m^2}{\frac{2}{a}} = \frac{a(1+m^2)}{2}.$$

Also,

$$\rho \text{ at } (0,0) = \frac{(1+y_1^2)^{\frac{3}{2}}}{y_2} = \frac{(1+m^2)^{\frac{3}{2}}}{\frac{2}{a}} = \frac{a(1+m^2)^{\frac{3}{2}}}{2}$$

Therefore, the equation of the circle of curvature is

$$(x-X)^2 + (y-Y)^2 = \rho^2$$

or

$$\left[x + \frac{am(1+m^2)}{2}\right]^2 + \left[y - \frac{a(1+m^2)}{2}\right]^2 = \frac{a^2(1+m^2)^3}{4}$$

or

$$\begin{aligned} 4x^2 + 4y^2 &= a(1+m^2)[a(1+m^2)^2 - am^2(1+m^2) \\ &\quad - a(1+m^2) + 4(y-mx)] \\ &= 4a(1+m^2)(y-mx), \end{aligned}$$

which yields

$$x^2 + y^2 = a(1+m^2)(y-mx).$$

3.10 CHORDS OF CURVATURE PARALLEL TO THE COORDINATE AXES

Let P(X, Y) be a point on the curve and let the tangent at P make an angle ψ with the positive direction of x-axis. Then $CP = \rho$ and $PD = 2\rho$. Let PA and PB be the chords of curvature parallel to the x- and y-axis, respectively.

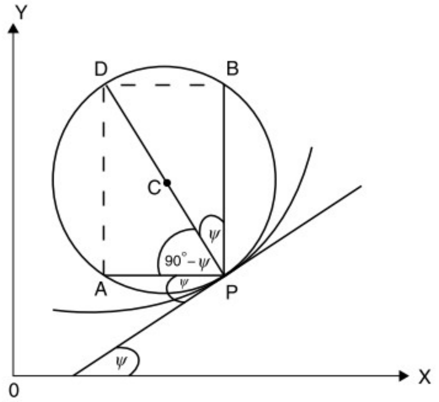

The chord PA makes an angle $\frac{\pi}{2} - \psi$ with CP and the chord PB makes an angle ψ with CP. Let C_x denote the length of PA and C_y denote the length of the chord PB. Then

$$C_x = PD\cos\left(\frac{\pi}{2} - \psi\right) = 2\rho\sin\psi$$

$$= 2\frac{(1+y_1^2)^{\frac{3}{2}}}{y_2}\cdot\frac{y_1}{\sqrt{1+y_1^2}} = \frac{2y_1(1+y_1^2)}{y_2}$$

and

$$C_y = PD\cos\psi = 2\rho\cos\psi$$

$$= 2\frac{(1+y_1^2)^{\frac{3}{2}}}{y_2}\cdot\frac{1}{\sqrt{1+y_1^2}} = \frac{2(1+y_1^2)}{y_2}.$$

3.11 CHORD OF CURVATURE IN POLAR COORDINATES

1. Chord of Curvature Through the Pole (Origin)

Let P(X, Y) be a point on the curve and let the tangent at P make an angle ϕ with the radius vector OP. Then PA, the chord of curvature through the pole O makes an angle $\frac{\pi}{2} - \phi$ with CP, the radius of curvature.

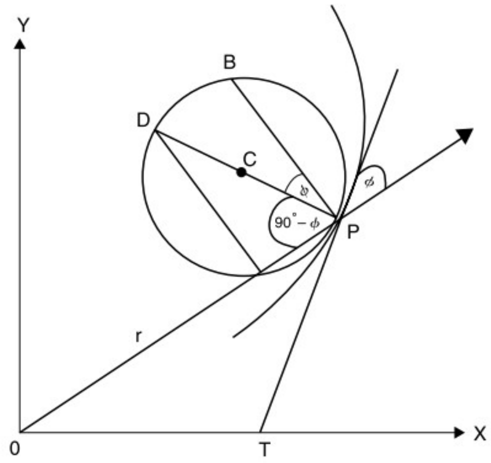

If C_o denotes the length of the chord of curvature through the pole, then

$$C_o = \text{PA} = \text{PD}\cos\left(\frac{\pi}{2} - \phi\right) = 2\rho\sin\phi$$

$$= 2\frac{(r^2+r_1^2)^{\frac{3}{2}}}{r^2+2r_1^2-rr_2}\cdot\frac{r}{\sqrt{r^2+r_1^2}}, \text{ since } \tan\phi = \frac{r}{r_1}$$

$$= \frac{2r(r^2+r_1^2)}{r^2+2r_1^2-rr_2}.$$

2. Chord of Curvature Perpendicular to the Radius Vector

The chord PB of curvature perpendicular to the radius vector makes an angle ϕ with CP. Therefore, the length C_p of this chord is given by

$$C_p = \text{PB} = \text{PD}\cos\varphi = 2\rho\sin\phi$$

$$= 2\frac{(r^2+r_1^2)^{\frac{3}{2}}}{r^2+2r_1^2-rr_2}\cdot\frac{r_1}{\sqrt{r^2+r_1^2}}$$

$$= \frac{2r_1(r^2+r_1^2)}{r^2+2r_1^2-rr_2}.$$

EXAMPLE 3.29

Show that in the curve $y = a\log\sec\frac{x}{a}$, the chord of curvature parallel to the axis of y is of constant length.

Solution. The equation of the curve is

$$y = a\log\sec\frac{x}{a}.$$

Therefore,

$$y_1 = \frac{dy}{dx} = a\frac{1}{\sec\frac{x}{a}}\sec\frac{x}{a}\tan\frac{x}{a}.\frac{1}{a} = \tan\frac{x}{a},$$

$$y_2 = \frac{d^2y}{dx^2} = \frac{1}{a}\sec^2\frac{x}{a}.$$

Therefore, the length of chord of curvature parallel to the y-axis is

$$C_y = \frac{2(1+y_1^2)}{y_2} = \frac{2\left(1+\tan^2\frac{x}{a}\right)}{\frac{1}{a}\sec^2\frac{x}{a}} = 2a \quad \text{(constant)}.$$

EXAMPLE 3.30

Find the length of the chord of curvature through the pole of the cardioid $r = a\,(1 + \cos\,\theta)$.

Solution. The cardioid is $r = a\,(1 + \cos\theta)$. Therefore,

$$r_1 = \frac{dr}{d\theta} = -a\sin\theta \text{ and } r_2 = \frac{d^2r}{d\theta^2} = -a\cos\theta.$$

Hence,

$$\begin{aligned} C_o &= \frac{2r(r^2+r_1^2)}{r^2+2r_1^2-rr_2} \\ &= \frac{2r[a^2(1+\cos\theta)^2+a^2\sin^2\theta]}{a^2(1+\cos\theta)^2+2a^2\sin^2\theta+a^2(1+\cos\theta)\cos\theta} \\ &= \frac{2r(a^2+a^2\cos^2\theta+2a^2\cos\theta+a^2\sin^2\theta)}{3a^2+3a^2\cos\theta} \\ &= \frac{2r[2a^2+2a^2\cos\theta]}{3a^2(1+\cos\theta)} = \frac{4r^2a}{3ar} = \frac{4r}{3}. \end{aligned}$$

EXAMPLE 3.31

If C_x and C_y are the chords of curvature parallel to the axes at any point of the curve $y = ae^{\frac{x}{a}}$, show that

$$\frac{1}{C_x^2} + \frac{1}{C_y^2} = \frac{1}{2aC_x}.$$

Solution. We know that

$$C_x = \frac{2y_1(1+y_1^2)}{y_2} \text{ and } C_y = \frac{2(1+y_1^2)}{y_2}.$$

The given curve is $y = \text{a}e^{\frac{x}{a}}$. Therefore,

$$y_1 = \frac{dy}{dx} = \text{a}e^{\frac{x}{a}}\cdot\frac{1}{a} = e^{\frac{x}{a}} \text{ and } y_2 = \frac{d^2y}{dx^2} = e^{\frac{x}{a}}\cdot\frac{1}{a}.$$

Hence,

$$C_x = \frac{2e^{\frac{x}{a}}\left(1+e^{\frac{2x}{a}}\right)}{\frac{1}{a}e^{\frac{x}{a}}} = 2a\left(1+e^{\frac{2x}{a}}\right)$$

and

$$C_y = \frac{2\left(1+e^{\frac{2x}{a}}\right)}{\frac{1}{a}\;e^{\frac{x}{a}}} = \frac{2a}{e^{\frac{x}{a}}}\left(1+e^{\frac{2x}{a}}\right).$$

Then,

$$\begin{aligned} \frac{1}{C_x^2}+\frac{1}{C_y^2} &= \frac{1}{4a^2\left(1+e^{\frac{2x}{a}}\right)^2}+\frac{1}{\frac{4a^2}{e^{\frac{2x}{a}}}\left(1+e^{\frac{2x}{a}}\right)^2} \\ &= \frac{1}{4a^2\left(1+e^{\frac{2a}{x}}\right)^2}\left(1+e^{\frac{2x}{a}}\right) = \frac{1}{4a^2\left(1+e^{\frac{2x}{a}}\right)} = \frac{1}{2aC_x}. \end{aligned}$$

EXAMPLE 3.32

Find the length of the chord of curvature through the pole of the curve $r^n = a^n\cos n\theta$.

Solution. The equation of the given curve is $r^n = a^n \cos n\theta$. Differentiating with respect to θ, we have

$$nr^{n-1}\frac{dr}{d\theta} = -na^n\sin\;n\theta$$

or

$$\begin{aligned} r_1 &= \frac{dr}{d\theta} = -\frac{a^n\sin\;n\theta}{r^{n-1}} = -r\frac{a^n\sin\;n\theta}{r^n} \\ &= -r\frac{a^n\sin\;n\theta}{a^n\cos n\theta} = -r\tan\;n\theta \quad \text{and} \\ r_2 &= -r_1\tan\;n\theta - rn\sec^2 n\theta \\ &= r\tan^2\;n\theta - rn\sec^2 n\theta. \end{aligned}$$

Therefore,

$$\begin{aligned} C_o &= \frac{2r(r^2+r_1^2)}{r^2-rr_2+2r_1^2} \\ &= \frac{2r(r^2+r^2\tan\;n\theta)}{r^2-r(r\tan^2 n\theta-r\,n\sec^2 n\theta)+2r^2\tan^2\theta} \\ &= \frac{2r^3\sec^2 n\theta}{r^2[1-\tan^2 n\theta+n\sec^2 n\theta+2\tan^2 n\theta]} \\ &= \frac{2r\sec^2 n\theta}{1+\tan^2 n\theta+n\sec^2 n\theta} \\ &= \frac{2r\sec^2 n\theta}{\sec^2 n\theta+n\sec^2 n\theta} = \frac{2r}{n+1}. \end{aligned}$$

EXAMPLE 3.33

If C_o and C_p denote the length of the chords of curvature of the cardioid $r = a(1 + \cos\theta)$, along and perpendicular to the radius vector through any point, show that

$$3(C_o^2 + C_p^2) = 8aC_o.$$

Solution. From Example 3.30, we note that

$$r_1 = \frac{dr}{d\theta} = -a\sin\theta,$$

$$r_2 = \frac{d^2r}{d\theta^2} = -a\cos\theta, \text{ and } C_o = \frac{4r}{3}.$$

Further,

$$C_p = \frac{2r_1(r^2 + r_1^2)}{r^2 - rr_2 + 2r_1^2} = -\frac{2a\sin\theta(2a^2 + 2a^2\cos\theta)}{3a^2(1+\cos\theta)}$$

$$= -\frac{4a^2\sin\theta(1+\cos\theta)}{3a^2(1+\cos\theta)} = -\frac{4a\sin\theta}{3}.$$

Therefore,

$$3(C_o^2 + C_p^2) = 3\left[\frac{16r^2}{9} + \frac{16a^2}{9}\sin^2\theta\right]$$

$$= \frac{16}{3}[r^2 + a^2\sin^2\theta]$$

$$= \frac{16}{3}[a^2(1+\cos^2\theta + 2\cos\theta) + a^2\sin^2\theta]$$

$$= \frac{16}{3}[2a^2(1+\cos\theta)] = \frac{32ra}{3} = 8aC_o.$$

EXAMPLE 3.34

Prove that the points on the curve $r = f(\theta)$, is the circle of curvature at which it passes through the origin (pole) are given by the equation

$$f(\theta) + f''(\theta) = 0.$$

Solution. Since the circle of curvature passes through the pole, the chord of curvature through the pole is r. But the chord of curvature C_o is given by

$$C_o = \frac{2r(r^2 + r_1^2)}{r^2 - rr_2 + 2r_1^2}.$$

Therefore,

$$r = \frac{2r(r^2 + r_1^2)}{r^2 - rr_2 + 2r_1^2}$$

or

$$r^2 - rr_2 + 2r_1^2 = 2(r^2 + r_1^2)$$

or

$$r^2 + rr_2 = 0 \text{ or } r(r + r_2) = 0$$

and so, $r + r_2 = 0$. Hence,

$$f(\theta) + f''(\theta) = 0.$$

EXAMPLE 3.35

Show that in any curve the chord of curvature perpendicular to the radius vector is

$$\frac{2\rho}{r}\sqrt{r^2 - p^2}.$$

Solution. The chord of curvature perpendicular to the radius victor is

$$C_p = 2\rho\cos\phi = 2\rho\sqrt{1 - \sin^2\phi}$$

$$= 2\rho\sqrt{1 - \left(\frac{p}{r}\right)^2}, \text{ since } p = r\sin\phi$$

$$= \frac{2\rho}{r}\sqrt{r^2 - p^2}.$$

EXERCISES

1. Find the radius of curvature at any point (x, y) of the following curves:

   (i) $ay^2 = x^3$ **Ans.** (i) $\frac{\sqrt{x}(4a+9x)^{\frac{3}{2}}}{6a}$.

   (ii) $xy = c^2$ **Ans.** (ii) $\frac{x^2+y^2}{2c^2}$.

2. Find the radius of curvature at a given point of the following curves:

   (i) $y = e^x$ at the point where it cross the y-axis.
   Hint: The point is $x = 0$ and $y = e^0 = 1$.

   (ii) $y = \frac{c}{2}\left(e^{\frac{x}{c}} + e^{-\frac{x}{c}}\right)$ at (x, y).
   Hint: The curve is $y = c\cosh\frac{x}{c}$.

   (iii) $\sqrt{x} + \sqrt{y} = \sqrt{a}$ at $\left(\frac{a}{4}, \frac{a}{4}\right)$.

   (iv) $y^2 = \frac{a^2(a-x)}{x}$ at $(a, 0)$.

   Ans. (i) $2\sqrt{2}$ (ii) $\frac{y^2}{c}$ (iii) $\frac{a}{\sqrt{2}}$ (iv) $\frac{a}{2}$.

3. Show that the radius of curvature at any point of the astroid $x^{\frac{2}{3}} + y^{\frac{2}{3}} = a^{\frac{2}{3}}$ is three times the length of the perpendicular from the origin to the tangent at that point.
   Hint: If parametric equations are $x = a\cos^2 t$ and $y = a\sin^3 t$, then $\rho = 3a\sin t\ \cos t$, tangent is $y - a\sin^3 t = \frac{dy}{dx}(x - a\cos^3 t)$, and length of perpendicular from (0,0) on tangent is $\sin t \cos t$.

4. If ρ is the radius of curvature of the parabola $y^2 = 4ax$ at a point P, show that ρ^2 varies as $(SP)^3$, where S is the focus.
   Hint: ρ at $(at^2, 2at)$ is $2a(1+t^2)^{\frac{3}{2}}$, $S(a, 0)$ is the focus and so,
   $SP = \sqrt{(at^2 - a)^2 + (2at - 0)^2} = a(1 + t^2)$.
   Then $\rho^2 = \frac{4}{a}(SP)^3$.

5. For the curve $y = \frac{ax}{a+x}$, show that $\left(\frac{2\rho}{a}\right)^{\frac{2}{3}} = \left(\frac{y}{x}\right)^2 + \left(\frac{x}{y}\right)^2$, where ρ is the radius of curvature at (x,y).

6. Show that for the curve $x = a\cos\theta(1+\sin\theta)$ and $y = a\sin\theta\,(1+\cos\theta)$, the radius of curvature at $\theta = -\frac{\pi}{4}$ is a.

7. Prove that for the curve $r = a(\sec^3\psi - 1)$ the radius of curvature is $\rho = 3a\tan\psi\sec^3\psi$ and hence, show that $3a\frac{dy}{dx}\cdot\frac{d^2y}{dx^2} = 1$.

8. Prove that for the curve $4x = \theta + \sin\theta$ and $4y = 1 - \cos\theta$, the radius of curvature is $\rho = \cos\frac{\theta}{2}$.

9. Find the radius of curvature at the point θ of the curve $x = a\left(\cos\theta + \log\tan\frac{\theta}{2}\right)$ and $y = a\sin\theta$.
Hint: $\frac{dx}{d\theta} = \frac{a\cos^2\theta}{\sin\theta}$, $\frac{dy}{d\theta} = a\cos\theta$, $\frac{dy}{dx} = \tan\theta$, and $\frac{d^2y}{dx^2} = \frac{\sin\theta}{a\cos^4\theta}$. **Ans.** $a\cos\theta$.

10. Find the radius of curvature of the point (p,r) of the following curves:

    (i) $r^3 = 2ap^2$.

    (ii) $p^2 = \frac{r^4}{r^2+a^2}$. **Ans**. (i) $\frac{2}{3}\sqrt{2ar}$ (ii) $\frac{(r^2+a^2)^{\frac{3}{2}}}{r^2+2a^2}$.

11. Find the radius of curvature at any point of the curve $r^n = a^n\cos n\theta$.
Hint: $r_1 = -r\tan n\theta$ and $r_2 = r\tan^2\theta - nr\sec^2 n\theta$. Putting in $\rho = \frac{(r^2+r_1^2)}{r^2+2r_1^2-rr_2}$, we get $\rho = \frac{a^n}{(n+1)r^{n-1}}$.

12. Find the radius of curvature at (r,θ) of the curve $r = 6\left(1-\sin^2\frac{\theta}{2}\right)$.
Ans. $\rho = 4\cos\frac{\theta}{2}$.

13. Show that the radius of curvature at the point (r,θ) on the curve
$\theta = \frac{1}{a}(r^2-a^2)^{\frac{1}{2}} - \cos^{-1}\left(\frac{a}{r}\right)$ is $\rho = (r^2-a^2)^{\frac{1}{2}}$.
Hint: $\frac{d\theta}{dr} = \frac{(r^2-a^2)^{\frac{1}{2}}}{ar}$ and so, $\frac{dr}{d\theta} = \frac{ar}{\sqrt{r^2-a^2}}$.
Also $\frac{1}{p^2} = \frac{1}{r^2} + \frac{1}{r^4}\left(\frac{dr}{d\theta}\right)^2 = \frac{1}{r^2-a^2}$ and so pedal equation is $p^2 = r^2 - a^2$. Differentiating, we get $2p = 2r\frac{dr}{dp}$. Then $\rho = r\frac{dr}{dp} = p = \sqrt{r^2-a^2}$.

14. Show that the curvatures of the curves $r = a\theta$ and $r\theta = a$ at their common point are in the ratio 3:1.
Hint: Eliminating r between the two equations, we get $\theta^2 = 1$. Thus, at the common point $\theta^2 = 1$, for the first curve $r_1 = a$ and $r_2 = 0$ and so, $\rho_1(at\ \theta^2 = 1) = \frac{2a\sqrt{2}}{3}$. For the second curve $r_1 = -\frac{a}{\theta^2}$ and $r_2 = \frac{2a}{\theta^3}$ and so, $\rho_2(at\ \theta^2 = 1) = 2a\sqrt{2}$.

15. Prove that for any curve $\frac{d^2r}{ds^2} = \frac{\sin^2\phi}{r} - \frac{\sin\phi}{\rho}$
Hint: $\frac{dr}{ds} = \cos\phi$, $\rho = \frac{ds}{d\psi}$, $\phi = \psi - \theta$, and $\sin\phi = r\frac{d\theta}{ds}$.

16. Find the radius of curvature for the curve
$x = c\log\left\{s + \sqrt{s^2+c^2}\right\}$ and $y = \sqrt{s^2+c^2}$.
Hint: $\frac{dx}{ds} = -\frac{c}{\sqrt{s^2+c^2}}$ and $\frac{dy}{ds} = \frac{s}{s^2+c^2}\ \frac{d^2y}{ds^2} = \frac{d}{ds}\left(\frac{dy}{ds}\right) = \frac{c^2}{(s^2+c^2)^{\frac{3}{2}}}$ and so, $\rho = \frac{\frac{dx}{ds}}{\frac{d^2y}{ds^2}} = \frac{y^2}{c}$.

17. Find the radius of curvature at the origin of the following curves:

    (i) $x^3 - y^3 - 2x^2 + 6y = 0$.

    (ii) $x^3 - 2x^2y + 3xy^2 - 4y^3 + 5x^2 - 6xy + 7y^2 - 8y = 0$.

    Ans. (i) $\rho(0,0) = \frac{3}{2}$ (ii) $\rho(0,0) = \frac{4}{5}$.

18. Find the radius of curvature at the pole for the curve $r = a\sin n\theta$.
Ans. $\rho = \frac{1}{2}\frac{dr}{d\theta} = \frac{na}{2}$.

19. Find the radius of curvature at the origin for the curve $a(y^2 - x^2) = x^3$.
Ans. $2\sqrt{2}a$.

20. Find the coordinates of the center of curvature for the curve $a^2y = x^3$.
Ans. $\left(\frac{x}{2}\left(1 - \frac{9x^4}{a^4}\right), \left(\frac{5x^3}{2a^3} + \frac{a^2}{6x}\right)\right)$.

21. Show that the evolute of the tractrix $x = c\cos t + c\log\tan\frac{t}{2}$ and $y = c\sin t$ is the catenary $y = c\cosh\frac{x}{c}$.

22. Find the circle of curvature for the curve $x^3 + y^3 = 3xy$ at the point $\left(\frac{3}{2}, \frac{3}{2}\right)$.
Ans. $8(x^2 + y^2) - 21(x + \mathrm{y}) + 27 = 0$.

23. Find the length of the chord of curvature through the pole for the curves $r = \mathrm{a}e^{m\theta}$.
Ans. $2r$.

24. Find the coordinates of the center of curvature at $(at^2, 2at)$ on the parabola $y^2 = 4ax$.
Ans. $(2a + 3at^2, -2at^3)$.

25. Show that in the curve $y = a\cosh\left(\frac{x}{a}\right)$, the chord of curvature parallel to the axis of x is of length $a\sinh\left(\frac{2x}{a}\right)$.
Hint: $y_1 = \sinh\frac{x}{a}$ and $y_2 = \frac{1}{a}\cosh\frac{x}{a}$. Therefore, $C_x = \frac{2y_1(1+y_1^2)}{y_2} = a\sinh\frac{2x}{a}$.

26. If C_x and C_y are the chords of curvature parallel to the axis of x and y, respectively, at any point of the curve $y = a\cosh\frac{x}{a}$, show that

$$4a^2(C_x^2 + C_y^2) = C_y^4.$$

)*+,)))-) # ! # &!%& # 3"

(% % %!5 3 % 2 ; 4 7

")*+,)))-) # < 4 0
0" # 3"

(5 %% 9 # ;

VOLTAGES, CURRENTS AND POWER IN STAR CONNECTIONS

Figure 27 shows the three-phase three-wire system. The voltages V_{an}, V_{bn} and V_{cn} are called phase voltages, whereas the voltages V_{ab}, V_{bc} and V_{ca} are called line voltages, and the phase current (I_p) in each phase is equal to the line current in each line because each line is in series with its individual phase winding.

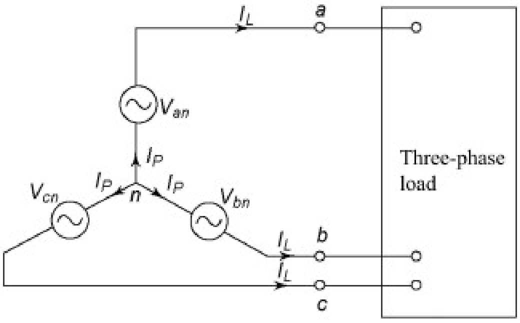

I_L
a
V_{an}
I_P
V_{cn}
I_P
n
I_P
V_{bn}
I_L
b
I_L
c
Three-phase load

9 &

STAR AND DELTA CONNECTIONS

The *three armature conductors* in a three-phase generator are not generally interconnected. They may be connected in the star (*Y*) or delta (Δ) form. The star connection of the three-armature conductors is shown in Figure 25 where all the similar ends a', b' and c' are at the starting point. The terminals a', b' and c' are joined to form the neutral terminal (*N*). In a star connection, similar ends are connected (the '*s*' in star indicates similar ends). The neutral conductor may or may not be present. If it is present, it is called a three-phase four-wire system. Otherwise, it is a three-phase three-wire system. The delta connection of the three-armature conductor is shown in Figure 26. In a delta connection (the *'d'* in delta indicates dissimilar ends), dissimilar ends are interconnected.

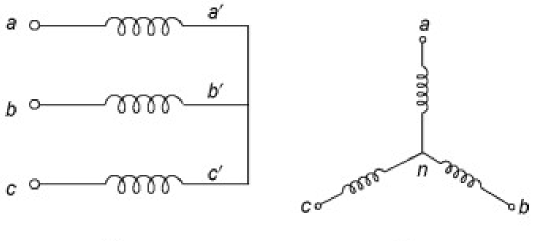

(a) (b)

Figure 25 Star Connection

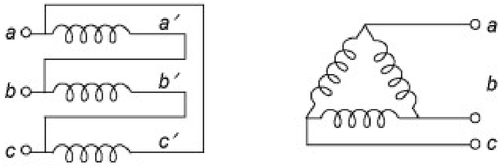

(a) (b)

Figure 26 Delta Connection

VOLTAGES, CURRENTS AND POWER IN STAR CONNECTIONS

Figure 27 shows the three-phase three-wire system. The voltages V_{an}, V_{bn} and V_{cn} are called phase voltages, whereas the voltages V_{ab}, V_{bc} and V_{ca} are called line voltages, and the phase current (I_p) in each phase is equal to the line current in each line because each line is in series with its individual phase winding.

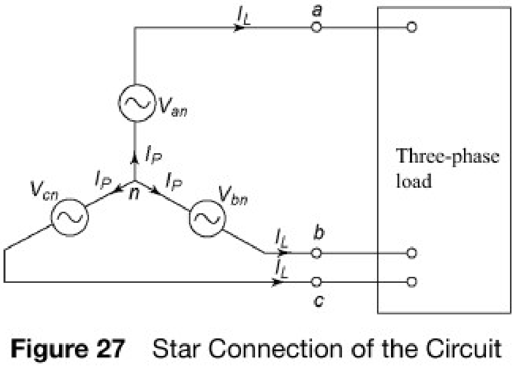

Figure 27 Star Connection of the Circuit

Let
$$\begin{aligned} V_{an} &= V_P \angle 0^\circ \\ V_{bn} &= V_P \angle -120^\circ \\ V_{cn} &= V_P \angle -240^\circ \end{aligned}$$

V_p indicates the magnitude of phase voltage.

Now,
$$\begin{aligned} V_{ab} &= V_{an} + V_{nb} = V_{an} - V_{bn} \\ &= V_P \angle 0^\circ - V_P \angle -120^\circ \\ &= V_P (\cos 0^\circ + j \sin 0^\circ) - V_P [\cos(-120^\circ) + j \sin(-120^\circ)] \\ &= V_P - V_P (\cos 120^\circ - j \sin 120^\circ) \\ &= V_P - V_P \left(-\frac{1}{2} - j\frac{\sqrt{3}}{2} \right) = V_P \left(1 + \frac{1}{2} + j\frac{\sqrt{3}}{2} \right) \\ &= V_P \left(\frac{3}{2} + j\frac{\sqrt{3}}{2} \right) = \sqrt{3}\, V_P \left(\frac{\sqrt{3}}{2} + j\frac{1}{2} \right) \\ &= \sqrt{3}\, V_P (\cos 30^\circ + j \sin 30^\circ) = \sqrt{3}\, V_P \angle 30^\circ \end{aligned}$$

Similarly,
$$\begin{aligned} V_{bc} &= V_{bn} + V_{nc} = V_{bn} - V_{cn} \\ &= V_P \angle -120^\circ - V_P \angle -240^\circ = V_P \angle -90^\circ \end{aligned}$$

and
$$\begin{aligned} V_{ca} &= V_{cn} + V_{na} = V_{cn} - V_{an} = V_P \angle -240^\circ - V_P \angle 0^\circ \\ &= \sqrt{3}\, V_P \angle -210^\circ \end{aligned}$$

Figure 28 shows the phasor diagram.

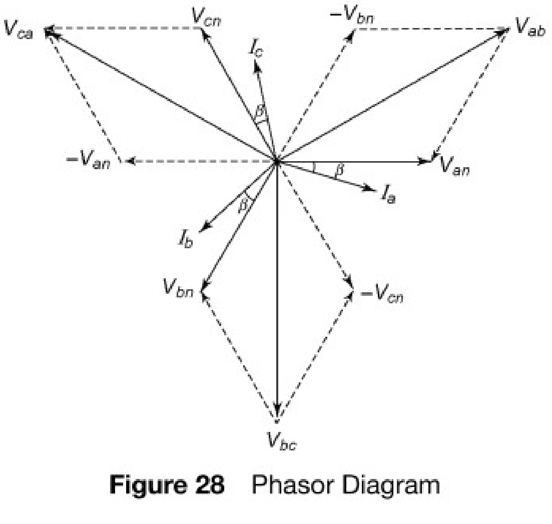

Figure 28 Phasor Diagram

So the magnitude of line voltage is $\sqrt{3}$ times the magnitude of phase voltage. In this case, $I_P = I_L$. The active or true power in each phase of the circuit is given by $V_P I_P \cos\beta$, where β is the angle between each phase voltage and its corresponding phase current. The total three-phase power (P) is expressed as

$$P = 3V_P I_P \cos\beta = 3\left(\frac{V_L}{\sqrt{3}}\right) I_L \cos\beta$$

i.e.,

$$P = \sqrt{3}V_L I_L \cos\beta \tag{29}$$

Similarly, total reactive power (Q) is expressed as

$$Q = 3V_P I_P \sin\beta = 3\left(\frac{V_L}{\sqrt{3}}\right) I_L \sin\beta$$

i.e.,

$$Q = \sqrt{3}V_L I_L \sin\beta \tag{30}$$

Q is positive when the load is of the inductive type, and it is negative when the load is of the capacitive type.

The total apparent power of the three phases (S) = $\sqrt{3}V_L I_L$.

VOLTAGES, CURRENTS AND POWER IN DELTA CONNECTIONS

Figure 29 shows three voltage sources connected in delta. In Figure 29, the phase voltages and line voltages are identical, but the line currents and phase currents are not identical.

Here, taking I_{ab} as reference, we have

$$\begin{aligned}
I_{ab} &= I_P \angle 0^\circ \\
I_{bc} &= I_P \angle -120^\circ \\
I_{ca} &= I_P \angle -240^\circ \\
I_a &= I_{ca} - I_{ab} = I_P \angle -240^\circ - I_P \angle 0^\circ \\
&= I_P(\cos 240^\circ - j\sin 240^\circ) - I_P(\cos 0^\circ + j\sin 0^\circ) \\
&= I_P\left(-\frac{1}{2} + j\frac{\sqrt{3}}{2}\right) - I_P = I_P\left(-\frac{3}{2} + j\frac{\sqrt{3}}{2}\right) \\
&= \sqrt{3}\,I_P\left(-\frac{\sqrt{3}}{2} + j\frac{1}{2}\right) = \sqrt{3}\,I_P(\cos 150^\circ + j\sin 150^\circ) \\
&= \sqrt{3}\,I_P \angle 150^\circ
\end{aligned}$$

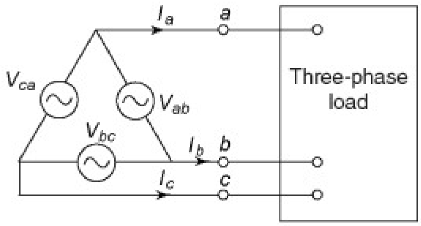

Figure 29 Delta Connection of the Circuit

data 09.04.2020

S T Q [✓ Q] S S D

Pearson(2013).pdf

Pag. 19(44)

Figure 30 Phasor Diagram

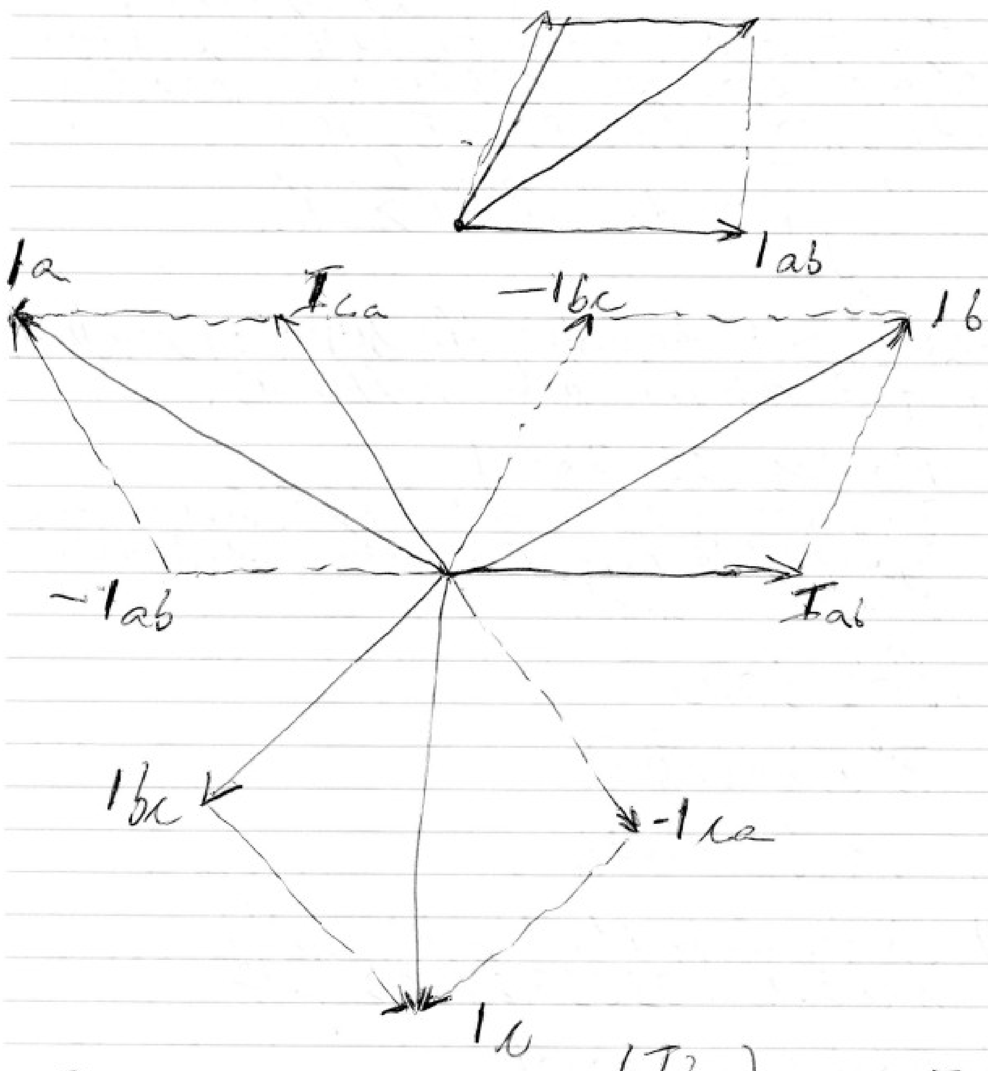
I_{ab}
I_a
I_{ca}
$-I_{bc}$
I_b
$-I_{ab}$
I_{ab}
I_{bc}
$-I_{ca}$
I_c

$$P = 3V_p I_p \cos\phi = 3V_L\left(\frac{I_L}{\sqrt{3}}\right)\cos\beta = \sqrt{3}\, V_L I_L \underset{\cos}{\underset{\downarrow}{\cos}}\beta$$

' #
)*+,)))-) # 6!5% 6 3"
)*+,)))-) # 4 7 " 0
4 3 = * 2 > / ? 0 " - " 5& @ 3 A " 3"
(5 6 #%% #%5 -/ B : " CCC
; 3 : 4 7 - 3 <
(! !6 *33 ' : + 3 :

100. Applications of the Integral Curve of the Reciprocal Curve.

Fig. 2 shows an **acceleration-velocity** curve, the scales being marked. The integral curve of its reciprocal, drawn in the direct manner described above, is shown below OX. The scale and units for ordinates of the integral curve, from the above reasoning, are:

$$1'' = \frac{20 \text{ feet}}{\text{secs.}} \div \text{NO} \, . \, \frac{\cdot 02 \text{ feet}}{(\text{secs.})^2} = \frac{20}{\cdot 75 \times \cdot 02} \, . \, \frac{\text{feet secs.}^2}{\text{secs. feet}} = 1333\tfrac{1}{3} \text{ seconds.}$$

The integral curve is therefore a **velocity-time** curve; a space-time curve may be obtained from it as on page 78.

If the given curve is a **velocity-space** curve (space along OX) the integral of the reciprocal will give a **space-time** curve.

EXAMPLES

(1) Plot the graph $y = \cdot 1x^2 + 3$ from $x = 0$ to 6. Scales: values of x, $1'' = 1$; values of y, $1'' = 2$. Draw the reciprocal curve using a pole distance of $2''$. Deduce the scale and read off the value of $1/y$ when $x = 3$. Check by calculation.

(2) The following table shows the relation between acceleration α and velocity v for a body, starting from rest.

α, ft./sec.^{2}	·6	·55	·45	·34	·246	·19	·16
v, ft./sec.	0	10	20	30	40	50	60

Plot the acceleration-velocity graph and obtain from it a velocity-time curve. What times have elapsed when the body has velocities of (*a*) 35 f.s., (*b*) 60 f.s.? Through what distance has the body moved when the velocity is 60 f.s.? Ans. (*a*) 80 sec., (*b*) 205 sec.; 7655 ft.

Note.—An additional example (No. 3) is given on page 86

Problem 99. To construct the Reciprocal Curve of a given Curve, and the Integral Curve of the Reciprocal.

Let $y = f(x)$, fig. 1, be any given curve. The reciprocal of this curve is one which will give values of $1/y$ for values of x.

Draw a succession of ordinates, produce them below OX, and mark the mid-points on the curve. Consider the mid-point p. Draw pA par¹ to OX to intersect OY in A, and join A to any pole N, on XO produced. From N draw NB perp. to NA to intersect OY in B. Draw Bp_1 par¹ to OX to intersect in p_1 the mid-ordinate from p: p_1 is a point on the required reciprocal curve. Obtain other points in the same way and draw the complete reciprocal curve.

To construct the integral curve of the reciprocal, make use of the radials below OX, from the pole N. Beginning at O draw through each ordinate space a line par¹ to the corresponding radial. The construction is similar to that given on page 76

It will be seen that this integral curve may be drawn *without first constructing the reciprocal*, the first segment of the integral curve, from O, being perp. to the first radial, as indicated by the dotted line; and so on.

Scales. (1) Reciprocal Curve. Let the scales be as marked: to determine n[C] in terms of h[A], m[B], and d.

From the figure, $po = \text{AO}$ and $p_1o = \text{BO}$.

By similar △s, AO : NO :: NO : BO, i.e. $\text{AO} \,.\, \text{BO} = (\text{NO})^2$ . . . (1)

$$\text{i.e. BO} = \frac{1}{\text{AO}} \,.\, (\text{NO})^2 = \frac{1}{\text{AO}} \,.\, \text{constant.}$$

Because ordinates of the reciprocal curve are to represent values of $\frac{1}{y}$, then

$$\text{BO}'' \,.\, n[\text{C}] = \frac{1}{\text{AO}'' \,.\, m[\text{B}]}; \text{ i.e. } n[\text{C}] = \frac{1}{\text{AO}'' \,.\, \text{BO}'' \,.\, m[\text{B}]}.$$

Substituting from (1), $n[\text{C}] = \dfrac{1}{(\text{NO}'')^2 \,.\, m[\text{B}]}$; i.e. $\boldsymbol{n[\text{C}] = \dfrac{1}{d^2 \,.\, m[\text{B}]}}$

(2) Integral Curve. The ordinate scale for this curve, $1'' = o[\text{D}]$, is deduced as on page 76 and is $1'' = n[\text{C}] \,.\, h[\text{A}] \,.\, d = \dfrac{1}{d^2 \,.\, m[\text{B}]} \,.\, h[\text{A}] \,.\, d$.

$$\therefore\ \boldsymbol{o[\text{D}] = \frac{h[\text{A}]}{d \,.\, m[\text{B}]}}.$$

100. Applications of the Integral Curve of the Reciprocal Curve.

Fig. 2 shows an **acceleration-velocity** curve, the scales being marked. The integral curve of its reciprocal, drawn in the direct manner described above, is shown below OX. The scale and units for ordinates of the integral curve, from the above reasoning, are:

$$1'' = \frac{20 \text{ feet}}{\text{secs.}} \div \text{NO} \,.\, \frac{\cdot 02 \text{ feet}}{(\text{secs.})^2} = \frac{20}{\cdot 75 \times \cdot 02} \,.\, \frac{\text{feet secs.}^2}{\text{secs. feet}} = 1333\tfrac{1}{3} \text{ seconds.}$$

The integral curve is therefore a **velocity-time** curve; a space-time curve may be obtained from it as on page 78.

If the given curve is a **velocity-space** curve (space along OX) the integral of the reciprocal will give a **space-time** curve.

EXAMPLES

(1) Plot the graph $y = \cdot 1x^2 + 3$ from $x = 0$ to 6. Scales: values of x, $1'' = 1$; values of y, $1'' = 2$. Draw the reciprocal curve using a pole distance of 2″. Deduce the scale and read off the value of $1/y$ when $x = 3$. Check by calculation.

(2) The following table shows the relation between acceleration α and velocity v for a body, starting from rest.

α, ft./sec.²	·6	·55	·45	·34	·246	·19	·16
v, ft./sec.	0	10	20	30	40	50	60

Plot the acceleration-velocity graph and obtain from it a velocity-time curve. What times have elapsed when the body has velocities of (*a*) 35 f.s., (*b*) 60 f.s.? Through what distance has the body moved when the velocity is 60 f.s.? Ans. (*a*) 80 sec., (*b*) 205 sec.; 7655 ft.

Note.—An additional example (No. 3) is given on page 86

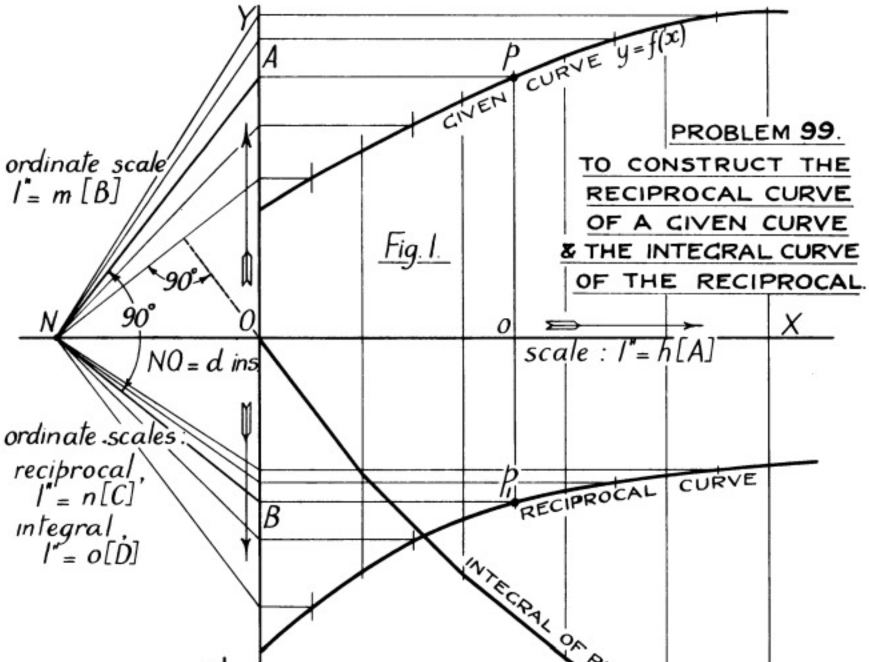
Y
A
P
GIVEN CURVE y = f(x)
ordinate scale
1" = m [B]
90°
90°
N
O
o
X
NO = d ins
scale : 1" = h [A]
ordinate scales:
reciprocal,
1" = n [C]
integral,
1" = o [D]
P₁
B
RECIPROCAL CURVE
INTEGRAL OF RECIPROCAL CURVE

Fig. 1.

PROBLEM 99.
TO CONSTRUCT THE
RECIPROCAL CURVE
OF A GIVEN CURVE
& THE INTEGRAL CURVE
OF THE RECIPROCAL.

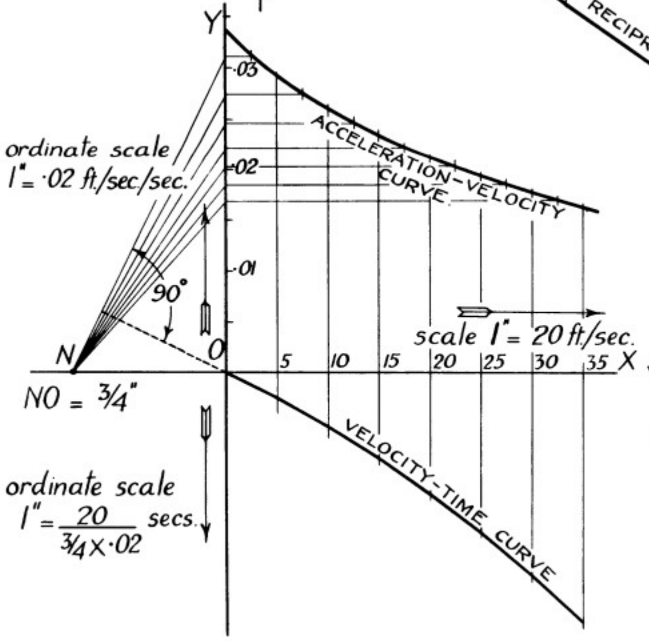
Y
·03
·02
·01
ordinate scale
1" = ·02 ft/sec/sec.
ACCELERATION-VELOCITY CURVE.
90°
N
O
scale 1" = 20 ft/sec.
5
10
15
20
25
30
35
X
NO = 3/4"
VELOCITY-TIME CURVE
ordinate scale
1" = 20/(3/4 × ·02) secs.

Fig. 2.

GIVEN AN
ACCELERATION-
VELOCITY CURVE,
TO CONSTRUCT
A VELOCITY-TIME
CURVE.

' #
)*+,)))-) # D 3 " 6 3"
; * 3 0E : < : "

' ##
)*+,)))-) ## 2 7
> " !@ + 7 7 7 3"
(! 9

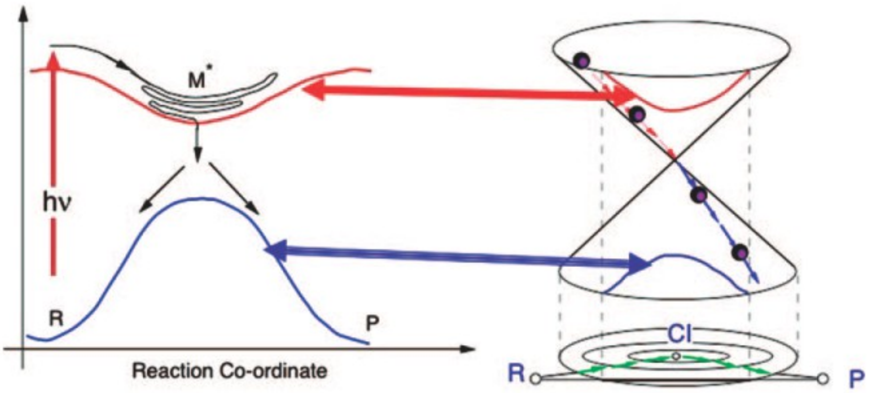

Figure 1.5 (a) Avoided crossing, (b) conical intersection, true crossing. Adapted from M. Robb, M. Garavelli, M. Olivucci and F. Bernardi, A Computational Strategy for Organic Photochemistry, *Rev. Comp. Ch.*,[15] John Wiley and Sons, © 2000 by Wiley-VCH, Inc.

6

)*+,)))-) ## &! #= 3 3"

' - 7

1(33) Books_67_2020/
Theoretical Chemistry for/
Electronic Excited States/
theoretical chemistry for
electronic excited fstates
(2018.~).pdf

Pág. 08(22)

Figure 1.5 a) Avoided crossing
b) Canonical intersec-
tion, true crossing.

10.04.2020

26. http://www.worldscientific.com/doi/suppl/10.1142/7803/suppl_file/7803_chap01.pdf

→ For an introductory article on conical intersections in photochemistry see Robb[22], which is available as a free download.[25]

Pág. 15(29)

Figure 1.10

Pág. 16(30)

Figure 1.11

Pág. 19(33)

Figure 1.13

(34) 20200411

A) Books_67_2020/
Solving problems in geometry/
Projective geometry an introduction by Rey Casse .. pdf

Projective Planes
Pág. 29(42)

' #%
13" "
')*+,)))-) #% %& 56! #% 3"
3 " 3 2 7
")*+,)))-) #% # ! & 3"
3 3 D " 3 7
(&& !& 6 5 0E

6.9 Equilateral triangles

Explicit formulas for the equilateral triangle are developed in this section, thus allowing the qualitative max/min theorems in this chapter for triangles to be rephrased as computable estimates. For example, the maximality of $\mu_2 A$ for the equilateral triangle (as stated in Theorem 6.33) is equivalent to saying

$$\mu_2 A \le \frac{4\pi^2}{3\sqrt{3}},$$

as one sees by evaluating $\mu_2 A$ for the equilateral using the values stated below.

The modes and frequencies of the equilateral triangle were derived roughly 150 years ago by Lamé. For a modern treatment of the Dirichlet case, see Mathews and Walker's text [695, pp. 237–239], and for a comprehensive treatment under all boundary conditions see McCartin's monograph [701] or his earlier papers upon which the monograph is based.

Consider now the equilateral triangle with vertices at $(0,0)$, $(\sqrt{3}/2, 1/2)$ and $(\sqrt{3}/2, -1/2)$. This triangle has sidelength 1, and

$$A = \sqrt{3}/4, \qquad D = 1, \qquad L = 3, \qquad S^2 = 3.$$

We state below the Dirichlet and Neumann spectra for this triangle.

6.9.1 Dirichlet spectrum of equilateral triangle with unit sidelength

The first three eigenvalues are

$$\lambda_1 = \frac{3 \cdot 16\pi^2}{9}, \qquad \lambda_2 = \lambda_3 = \frac{7 \cdot 16\pi^2}{9}$$

with eigenfunctions

$$u_1(x,y) = \sin\left(\frac{4}{\sqrt{3}}\pi x\right) - 2\cos(2\pi y)\sin\left(\frac{2}{\sqrt{3}}\pi x\right) \tag{6.13}$$

$$u_2(x,y) = \sin\left(\frac{6}{\sqrt{3}}\pi x\right)\cos\left(\frac{2}{3}\pi y\right) - \sin\left(\frac{4}{\sqrt{3}}\pi x\right)\cos\left(\frac{8}{3}\pi y\right) - \sin\left(\frac{2}{\sqrt{3}}\pi x\right)\cos\left(\frac{10}{3}\pi y\right) \tag{6.14}$$

$$u_3(x,y) = \sin\left(\frac{6}{\sqrt{3}}\pi x\right)\sin\left(\frac{2}{3}\pi y\right) - \sin\left(\frac{4}{\sqrt{3}}\pi x\right)\sin\left(\frac{8}{3}\pi y\right) + \sin\left(\frac{2}{\sqrt{3}}\pi x\right)\sin\left(\frac{10}{3}\pi y\right) \tag{6.15}$$

Figure 6.11 displays the nodal patterns of these first three Dirichlet eigenfunctions of the equilateral triangle, extended by odd reflection across the sides of the triangle.

The higher eigenfunctions are more complicated, although the corresponding eigenvalues retain a simple form:

$$\frac{16\pi^2}{9}(j_1^2 + j_1 j_2 + j_2^2), \qquad j_1, j_2 \ge 1. \tag{6.16}$$

Choosing $(j_1, j_2) = (1,1)$ gives the first eigenvalue $48\pi^2/9$ stated above. Note that modes $(j_1, j_2) = (5,6)$ and $(j_1, j_2) = (1,9)$ belong to the same eigenvalue (see Figure 6.11 for some linear combinations).

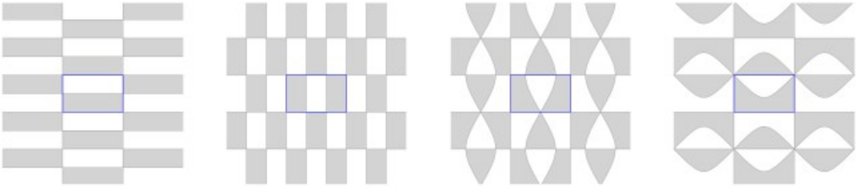

Fig. 6.9. Nodal patterns for the smallest multiple eigenvalue on a rectangle with $a/b = \sqrt{8/3}$.

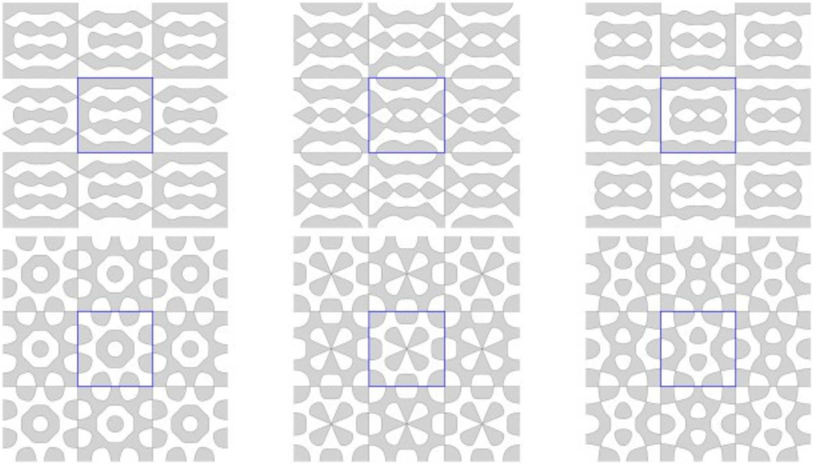

Fig. 6.10. Various nodal patterns belonging to the 31st Dirichlet eigenvalue of a square. This eigenvalue has multiplicity three, with pure modes (1, 7), (7, 1) and (5, 5). Different linear combinations of these modes generate rather diverse nodal domains, including the ones shown here.

multiplicity can arise in interesting ways, such as for $(0, 5)$ and $(3, 4)$ in the Neumann case, or $(1, 7)$ and $(5, 5)$ in the Dirichlet case (which in fact gives triple multiplicity). In Section 6.13 we describe nodal problems that have been solved recently in a few special cases, including squares.

6.9 Equilateral triangles

Explicit formulas for the equilateral triangle are developed in this section, thus allowing the qualitative max/min theorems in this chapter for triangles to be rephrased as computable estimates. For example, the maximality of $\mu_2 A$ for the equilateral triangle (as stated in Theorem 6.33) is equivalent to saying

$$\mu_2 A \leq \frac{4\pi^2}{3\sqrt{3}},$$

as one sees by evaluating $\mu_2 A$ for the equilateral using the values stated below.

The modes and frequencies of the equilateral triangle were derived roughly 150 years ago by Lamé. For a modern treatment of the Dirichlet case, see Mathews and Walker's text [695, pp. 237–239], and for a comprehensive treatment under all boundary conditions see McCartin's monograph [701] or his earlier papers upon which the monograph is based.

Consider now the equilateral triangle with vertices at $(0,0)$, $(\sqrt{3}/2, 1/2)$ and $(\sqrt{3}/2, -1/2)$. This triangle has sidelength 1, and

$$A = \sqrt{3}/4, \qquad D = 1, \qquad L = 3, \qquad S^2 = 3.$$

We state below the Dirichlet and Neumann spectra for this triangle.

6.9.1 Dirichlet spectrum of equilateral triangle with unit sidelength

The first three eigenvalues are

$$\lambda_1 = \frac{3 \cdot 16\pi^2}{9}, \qquad \lambda_2 = \lambda_3 = \frac{7 \cdot 16\pi^2}{9}$$

with eigenfunctions

$$u_1(x,y) = \sin\left(\frac{4}{\sqrt{3}}\pi x\right) - 2\cos(2\pi y)\sin\left(\frac{2}{\sqrt{3}}\pi x\right) \tag{6.13}$$

$$u_2(x,y) = \sin\left(\frac{6}{\sqrt{3}}\pi x\right)\cos\left(\frac{2}{3}\pi y\right) - \sin\left(\frac{4}{\sqrt{3}}\pi x\right)\cos\left(\frac{8}{3}\pi y\right) - \sin\left(\frac{2}{\sqrt{3}}\pi x\right)\cos\left(\frac{10}{3}\pi y\right) \tag{6.14}$$

$$u_3(x,y) = \sin\left(\frac{6}{\sqrt{3}}\pi x\right)\sin\left(\frac{2}{3}\pi y\right) - \sin\left(\frac{4}{\sqrt{3}}\pi x\right)\sin\left(\frac{8}{3}\pi y\right) + \sin\left(\frac{2}{\sqrt{3}}\pi x\right)\sin\left(\frac{10}{3}\pi y\right) \tag{6.15}$$

Figure 6.11 displays the nodal patterns of these first three Dirichlet eigenfunctions of the equilateral triangle, extended by odd reflection across the sides of the triangle.

The higher eigenfunctions are more complicated, although the corresponding eigenvalues retain a simple form:

$$\frac{16\pi^2}{9}(j_1^2 + j_1 j_2 + j_2^2), \qquad j_1, j_2 \geq 1. \tag{6.16}$$

Choosing $(j_1, j_2) = (1,1)$ gives the first eigenvalue $48\pi^2/9$ stated above. Note that modes $(j_1, j_2) = (5,6)$ and $(j_1, j_2) = (1,9)$ belong to the same eigenvalue (see Figure 6.11 for some linear combinations).

6.9.2 Neumann spectrum of equilateral triangle with unit sidelength

We have $\mu_1 = 0$, with eigenfunction $u_0 \equiv 1$, and

$$\mu_2 = \mu_3 = \frac{16\pi^2}{9}$$

)*+,)))-) #% < : 7 "
/7 + 7 3"
(5 %

5. Let ℓ, m, n be three *mutually skew* lines (i.e. no two of the lines intersect) of a projective space S_3 of dimension 3. Show that through each point of ℓ, there exists a unique line r which intersects both m and n.
   Such a line r is called an (ℓ, m, n)**-transversal**. The set $\mathfrak{R}$ of all (ℓ, m, n)-transversals is called a **regulus**, and is sometimes denoted by $\mathfrak{R}(\ell, m, n)$. Prove that no two distinct (ℓ, m, n)-transversals intersect in a point.
6. Show that if p, q are two skew lines in a projective space S_4 of dimension 4, then there exists a unique hyperplane of S_4 containing them. Deduce that *in general* there exists a unique line of S_4 intersecting three given lines which are pairwise non-intersecting.
7. Show that if S_r, S_n are two given projective spaces, then the points of $S_r \oplus S_n$ are those (and only those) on all the lines joining any point of S_r to any point of S_n.

Projective planes

A projective space of dimension 2 is called a projective plane. In that case, the set of axioms reduces to the following:

Definition 3.11 *A* **projective plane** π *is a set* $\mathcal{P}$ *of points and a set* $\mathcal{L}$ *of subsets of* $\mathcal{P}$*, called lines, satisfying the following:*

P1. There exists a unique line joining two distinct points.

P2. There exists a unique point of intersection of two distinct lines.

P3. There exist at least three non-collinear points.

P4. There exist at least three points on each line.

Example 3.12 The EEP is a projective plane since all the axioms are satisfied. However the Euclidean plane is not a projective plane, since axiom P2 is not satisfied by any pair of parallel lines.

Definition 3.13 *In a projective plane, a set of four points* $A, B, C, D,$ no three collinear *is called a* **quadrangle**. *A set of four lines* $a, b, c, d,$ no three concurrent *is called a* **quadrilateral**.

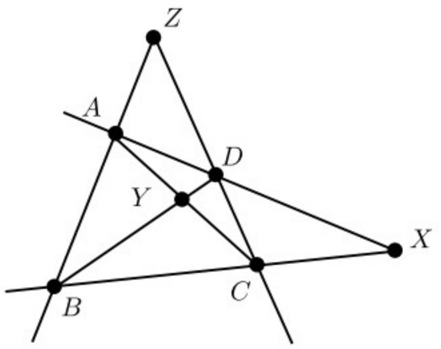

Example 3.14 Let π be a projective plane. Prove that π has a quadrangle

Solution. By Axiom P3, there exist three non-collinear points A, B, and X of π. Axiom P1 gives the lines AB, AX, and BX. By Axiom P4, there exist points $C \in BX$, $D \in AX$, distinct from A, B, and X. By Axioms P1 and P2, the line DC intersects the line AB in a point Z, which is distinct from A,B,C,D, and X. Thus $ABCD$ is a quadrangle. □

Note 6 *In the last example, BD, AC intersect in a point Y, which is distinct from the points A, B, C, D, X, and Z. There may be no other point in π, in which case the points X, Y, Z lie on a line.*

Definition 3.15 *This plane of seven points and seven lines is called the* **Fano** *plane.*

Either of the following two diagrams represents the Fano plane!

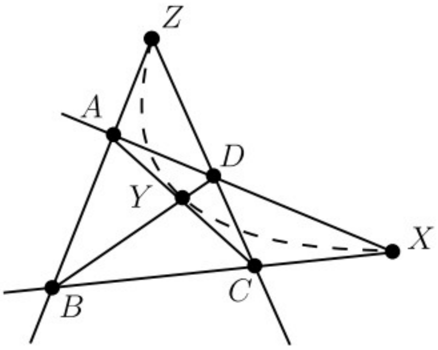

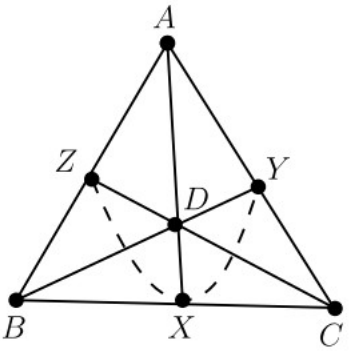

The Fano plane has exactly *three* points on *each* line, and *three* lines through *each* point. We now prove that this is a consequence of the Fano plane being a projective plane with a *finite* number of points.

Theorem 3.16 *Let π be a projective plane which has a finite number N_2 of points. Then*

1. *Every line of π has the same number $N_1 = q + 1$ of points. The number q is called the* **order** *of the plane π.*
2. *The number N_2 of points in π is $q^2 + q + 1 = (q^3 - 1)/(q - 1)$.*

Proof.

1. Let ℓ, m be any two distinct lines of π. Let $P = \ell \cap m$. On ℓ, there exist at least two points A, B, other than P. On m, there exist two points C, D, other than P.
   Let $AD \cap BC = Q$. Then $Q \notin \ell$, and $Q \notin m$. For each $P_i \in \ell$, the line QP_i intersects m in a point. Hence m has at least as many points as ℓ. Similarly, ℓ has at least as many points as m. Hence ℓ and m have the same number N_1 of points. Put $N_1 = q + 1$.

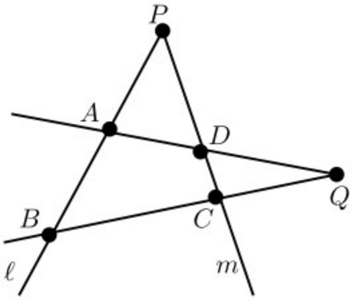

2. Let ℓ be any line of π. Take a point $P \notin \ell$. Then the points of π lie on the lines joining P to the $q+1$ points of ℓ. There are $q+1$ such lines, and each line contains q points, other than P. Hence

$$\begin{aligned} N_2 &= q(q+1) + 1 \\ &= q^2 + q + 1 \\ &= \frac{q^3 - 1}{q - 1}. \end{aligned}$$

□

Exercise 3.2 Let π be a projective plane. Use Definition 3.11 to prove:

P3'. There exist at least three non-concurrent lines in π.

P4'. There exist at least three lines through each point in π.

Deduce that the *Principle of Duality is valid in a projective plane.*

Order of a finite projective space

In Theorem 3.16, we proved that every line of a finite projective plane π has the same number of points. We next show that this is also true of a finite projective space.

Theorem 3.17 *Let $S_r, r \geq 2$, be an r-dimensional projective space which has a finite number N_r of points. Then*

1. *Every line of S_r has the same number $q+1$ of points. The number q is called the* **order** *of the space S_r.*

2. $N_r = q^r + q^{r-1} + \cdots + q + 1 = (q^{r+1} - 1)/(q - 1)$.

Proof.

1. Let ℓ, m be any two lines of S_r. Two cases are to be considered.
   Case 1: Suppose ℓ and m are coplanar. By Theorem 3.16, ℓ and m have the same number of points.
   Case 2: Suppose ℓ and m are not coplanar. Take a point P on ℓ, and a point Q on m. Then the line PQ is coplanar with, but distinct from, both ℓ and m. From Case 1, ℓ has the same number of points as PQ, and PQ has the same number of points as m. Hence ℓ and m have the same number N_1 of points.

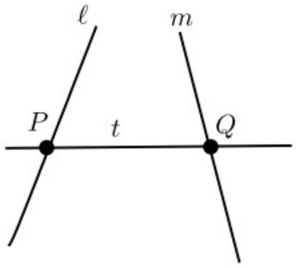

2. Put $N_1 = q + 1$. We use induction to find N_r. Assume that the result is true for S_{r-1}. Take a point $P \notin S_{r-1}$. Then $P \oplus S_{r-1}$ is a projective space S_r of dimension r. (See Example 3.10.) The points of S_r lie on the lines joining P to the points of S_{r-1}. There are N_{r-1} such lines, and each line contains q points, other than P. Hence

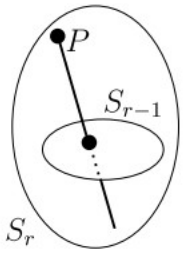

$$\begin{aligned} N_r &= qN_{r-1} + 1 \\ &= q(q^{r-1} + q^{r-2} + \cdots + q + 1) + 1 \\ &= q^r + q^{r-1} + \cdots + q + 1 \\ &= \frac{q^{r+1} - 1}{q - 1}. \end{aligned}$$ □

Example 3.18 The Fano plane is of order 2, since it has $2 + 1$ points on each line. In fact, it has $7 = 2^2 + 2 + 1$ points and $7 = 2^2 + 2 + 1$ lines.

Note 7 1. *By the principle of duality in the plane:*

$$\begin{aligned} \textit{number of points on a line} &= \textit{number of lines through a point} \\ &= q + 1, \\ \textit{number of points in the plane} &= \textit{number of lines in the plane} \\ &= q^2 + q + 1. \end{aligned}$$

2. *A careful investigation reveals that the principle of duality is valid in a projective space S_r of dimension r. In S_r, for all n, with $0 \leq n \leq r$, a subspace S_n of dimension n is dual to a subspace S_{r-n-1} of dimension $r - n - 1$.*

   In the plane ($r = 2$), S_0 and S_1 are dual spaces.

   In the projective space S_3 of dimension 3,

   - *a point (of dimension 0) is dual to a plane (of dimension $2 = 3 - 0 - 1$),*
   - *a line (of dimension 1) is dual to a line (of dimension $1 = 3 - 1 - 1$).*

Example 3.19 Let S_3 be a three-dimensional projective space of order q. By the principle of duality (a point is dual to a plane, and a line is dual to a line), we have:

$$\begin{aligned}\text{Number of points} &= q^3+q^2+q+1\\ &= \text{number of planes.}\end{aligned}$$

$$\begin{aligned}\text{Number of planes containing a line } \ell &= \text{number of points on a line } \ell\\ &= q+1.\end{aligned}$$

$$\begin{aligned}\text{Number of lines through a point} &= \text{number of lines in a plane}\\ &= q^2+q+1.\end{aligned}$$

Now, let N be the number of lines in S_3. We count the number M of pairs (ℓ, P), where P is a point lying on the line ℓ. On each of the N lines of S_3 there are $q+1$ points. Thus,

$$M = N(q+1).$$

Further, through each of the q^3+q^2+q+1 points of S_3, there pass q^2+q+1 lines. Therefore,

$$M = (q^2+q+1)(q^3+q^2+q+1).$$

Hence

$$\begin{aligned}N = \frac{M}{(q+1)} &= (q^2+q+1)\frac{(q^3+q^2+q+1)}{(q+1)}\\ &= (q^2+q+1)(q^2+1)\\ &= q^4+q^3+2q^2+q+1.\end{aligned}$$

□

In the next two examples, we use different techniques to find again the number of lines in a three-dimensional projective space S_3 of order q; the rationale of this repetition is to expose the reader to combinatorial arguments which are typical of what one encounters in Finite Geometry.

Example 3.20 Let π and α be two distinct planes of a three-dimensional projective space S_3 of order q. Find the number N of lines in S_3 by investigating how a line intersects π and α.

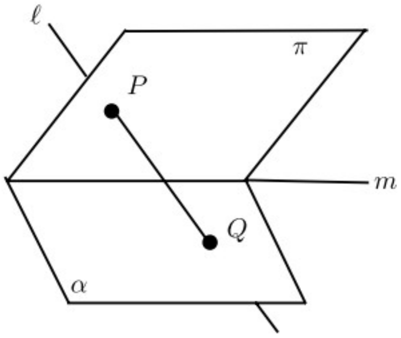

Solution. In S_3, the planes π and α necessarily intersect in a line; let m denote the line $\pi \cap \alpha$. A line ℓ of S_3 is one of three types:

Type I: $\ell = PQ$, with $P \in \pi \backslash m$ and $Q \in \alpha \backslash m$.

Type II: $\ell = m$.

Type III: ℓ intersects m in a point.

There are q^2 choices for P, and q^2 choices for Q; therefore there are $q^2 \times q^2$ lines of Type I. There is one line of Type II, namely m. Through each of the $q+1$ points of m, there pass q^2+q lines of Type III. Therefore, there are $(q^2+q)(q+1)$ lines of Type III. Hence,

$$N = q^4 + 1 + (q^2+q)(q+1) = q^4 + q^3 + 2q^2 + q + 1.$$

□

Example 3.21 Find the number N of lines in a three-dimensional projective space S_3 of order q by counting the number of ways of choosing two distinct points from the points of S_3.

Solution. Since each pair of the q^3+q^2+q+1 points of S_3 gives rise to a line, and any line is specified by a pair of its $q+1$ points, the number of lines of S_3 is

$$\begin{aligned} N &= \frac{\binom{q^3+q^2+q+1}{2}}{\binom{q+1}{2}} \\ &= \frac{(q^3+q^2+q+1)(q^3+q^2+q)}{(q+1)q} \\ &= \frac{(q^2+1)(q+1)q(q^2+q+1)}{(q+1)q} \\ &= (q^2+1)(q^2+q+1) \\ &= q^4+q^3+2q^2+q+1. \end{aligned}$$

□

Exercise 3.3

1. Consider the triple $(\mathcal{P}, \mathcal{L}, \mathcal{I})$ with $\mathcal{P} = \{1,2,3,4\}$, $\mathcal{L} = \{a, b, c, d, e, f\}$, and $\mathcal{I} = \{(1,a), (2,a), (3,b), (4,b), (1,c), (3,c), (2,d), (4,d), (1,e), (4,e), (2,f), (3,f)\}$.
   (a) Draw a diagram of this triple.
   (b) Verify that it satisfies only two of the axioms of a projective plane.

2. Using your knowledge of the Euclidean three-dimensional space $\mathbb{R}^3$, show that the following triple $(\mathcal{P}, \mathcal{L}, \mathcal{I})$ is a projective plane where:
   $\mathcal{P}$ is the set of lines passing through the origin $O = (0,0,0)$;
   $\mathcal{L}$ is the set of planes passing through O;
   $\mathcal{I} = \{(\ell, \alpha) \mid \ell \text{ is a line in the plane } \alpha;\ \ell \in \mathcal{P},\ \alpha \in \mathcal{L}\}$.

3. Prove that the dual of a projective plane is a projective plane.

4. Show that a finite projective plane has an odd number of points. Can there exist a projective plane with 23 points? What about 111 points?

5. Let π be a projective plane. Show that there exists lines ℓ, m in π and a point P not on ℓ or m.

6. Assume that Desargues' Theorem is valid in a given projective plane. Prove its converse, *without appealing to the Principle of Duality.*

   (*Hint:* Draw carefully the configuration of the converse of Desargues' Theorem; find, in the configuration, triangles in perspective from a point and apply Desargues' Theorem.)

7. Recall that in a two-dimensional projective space, (see Exercise 2.1(3)), the dual of Desargues' Theorem is its converse. Thus, by the principle of duality, Desargues' Theorem implies its converse.

   State, for the three-dimensional case: Desargues' Theorem, its dual, and its converse. Comment.

8. Consider the following

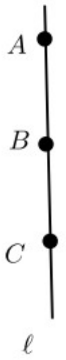

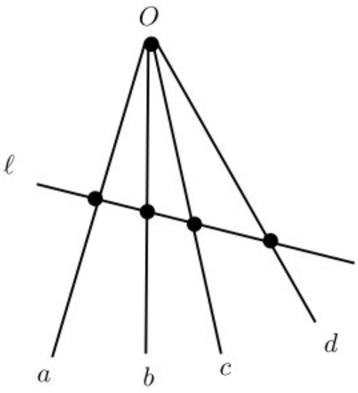

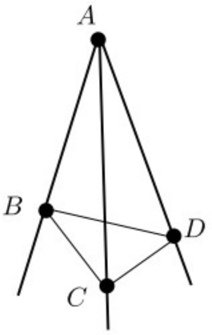

   as configurations, first, in a plane, and second, in a three-dimensional projective space. In each case, draw the dual configuration.

9. Let $\pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6$ be six planes in three-dimensional projective space such that each of the following sets of three planes have a common line of intersection

$$\{\pi_1, \pi_2, \pi_3\}, \{\pi_1, \pi_4, \pi_5\}, \{\pi_3, \pi_5, \pi_6\}, \{\pi_2, \pi_4, \pi_6\}.$$

   Further, no four planes have a common line of intersection.

   Prove that the six planes have a common point.

   (*Hint:* Solve the dual problem.)

10. Count all the incidences in S_4, a four-dimensional projective space of order q. That is, determine the number of
    (a) points and hyperplanes in S_4;
    (b) lines and planes in S_4;
    (c) points in each of a line, a plane and a hyperplane;
    (d) lines in each of a plane and a hyperplane;
    (e) lines through a point;
    (f) planes in a hyperplane;

(g) planes through a point;

(h) planes containing a line;

(i) hyperplanes through each of a point, a line, and a plane.

11. Let π be a projective plane of order q.

    Definition: A **k-arc** K is a set of k points of π, *no three collinear*. A line ℓ of π is a **chord**, a **tangent**, or an **external line** to K depending on whether $|\ell \cap K|$ is $2, 1$, or 0, respectively.

    (a) By considering the lines of π through a point P of a k-arc K, show that the number t of tangents through each point of K is
    $$t = q - k + 2.$$
    Deduce that
    $$k \leq q + 2$$
    and that a $(q+2)$-arc has no tangents.

    (b) Let K be a k-arc of π, where k is an odd number. By considering the lines of π through a point P not on K, show that there pass an odd number of tangents through P. Deduce that if q is odd,
    $$k \leq q + 1.$$

    (c) Let K be a $(q+1)$-arc of π, where q is *even*. Let P, Q be any two distinct points of K. Show that through each point of PQ there passes at least one tangent. Deduce that through each point of PQ there passes exactly one tangent. Deduce also that the $q+1$ tangents of K meet in a point N, and that $K \cup \{N\}$ is a $(q+2)$-arc. The point N is called the **nucleus** of the $(q+1)$-arc K.

12. **Definition**: An **affine plane** $\mathcal{A}$ is a triple $(\mathcal{P}, \mathcal{L}, \mathcal{I})$ with points $\mathcal{P}$, lines $\mathcal{L}$, and incidence $\mathcal{I}$, satisfying the following axioms:

    A1. Every pair of points lies on a unique line.

    A2. Given any line ℓ and any point P not on ℓ, there is a unique line m such that P lies on m, and ℓ and m have no common point.

    A3. There are three non-collinear points.

    We call two lines of $\mathcal{A}$ *parallel* if they coincide or have no common point.

    (a) Let π be a projective plane and let ℓ_∞ be a line of π. Let π^{ℓ_∞} denote the triple obtained from π by deleting the line ℓ_∞ and all the points on ℓ_∞.
        (i) Show that π^{ℓ_∞} is an affine plane.
        (ii) Deduce that the Euclidean plane is an affine plane.

    (b) Show that parallelism in an affine plane $\mathcal{A}$ is an equivalence relation and that each point of $\mathcal{A}$ lies on exactly one line from each class of parallel lines.

    (c) Prove that the dual of an affine plane is not an affine plane.

13. **Definition**: Let $(\mathcal{P}, \mathcal{B}, \mathcal{I})$ be a triple with points $\mathcal{P}$, lines $\mathcal{B}$, and incidence $\mathcal{I}$, satisfying the following axioms:

    A1. Each point is incident with $t+1$ distinct lines.
    Two distinct points are incident with *at most one* line.

A2. Each line is incident with $s+1$ distinct points.
Two distinct lines are incident with *at most one* point.

A3. If ℓ is a line and P a point not incident with ℓ, then there exists a *unique* point Q and a *unique* line m such that $m = PQ$ and $\ell \cap m = Q$.

Then $(\mathcal{P}, \mathcal{B}, \mathcal{I})$ is called a **Generalised Quadrangle** with parameters s and t and is denoted $\mathrm{GQ}(s,t)$.

Note: There is a point–line duality for $\mathrm{GQ}(s,t)$ which interchanges points and lines (and so interchanges s and t) in any definition or theorem.

(a) Let ℓ be a given line of $\mathrm{GQ}(s,t)$. Label the $s+1$ points on ℓ by $P_1, P_2, \ldots, P_{s+1}$. Let the set of t lines through P_i distinct from ℓ be $\ell_{i1}, \ldots, \ell_{it}$ for $i = 1, \ldots, s+1$. Let C^* be the configuration made up of the lines ℓ, ℓ_{ij} ($i = 1, \ldots, s+1$, $j = 1, \ldots, t$), and the points incident with them.

  (i) Show that if $i \neq k$, $(1 \leq i, k \leq s+1)$, then $\ell_{ij} \cap \ell_{km} = \phi$ for any j, m such that $1 \leq j, m \leq t$.

  (ii) Using (i), deduce that C^* has $(1+s)(1+st)$ points.

  (iii) Show that each point of the $\mathrm{GQ}(s,t)$ lies on a line of C^*.

  (iv) Using (ii) and (iii), deduce that the $\mathrm{GQ}(s,t)$ also has $(1+s)(1+st)$ points.

  (v) Using duality, write down the number of lines of the $\mathrm{GQ}(s,t)$.

(b) Let $q = 2^h$, where h is some positive integer. Let $\mathrm{PG}(3,q)$ denote a projective space of dimension 3 and of order q. Let α be a plane of $\mathrm{PG}(3,q)$. Assume that there exists a set $\mathcal{K}$ of $q+2$ points in α such that no three points of $\mathcal{K}$ are collinear. (Note that any line of α meets $\mathcal{K}$ in 0 or 2 points.) Consider the following triple $(\mathcal{P}, \mathcal{B}, \mathcal{I})$:

  - $\mathcal{P}$ is the set of points in $\mathrm{PG}(3,q)\backslash\alpha$.
  - $\mathcal{B}$ is the set of lines of $\mathrm{PG}(3,q)$ which do not lie in α and which contain exactly one point of $\mathcal{K}$.
  - $\mathcal{I}$ is the natural incidence between points and lines.

  (i) Show that $\mathcal{P}$ has q^3 points.

  (ii) Let P be a point of $\mathcal{K}$. Show that there are q^2 distinct lines of $\mathcal{B}$ passing through P and deduce that $\mathcal{B}$ has $q^2(q+2)$ lines.

  (iii) Show that each point in $\mathcal{P}$ lies on $q+2$ distinct lines in $\mathcal{B}$.
  Prove that each line in $\mathcal{B}$ contains q distinct points of $\mathcal{P}$.

  (iv) Let ℓ be a line in $\mathcal{B}$ and let P be a point in $\mathcal{P}$ not incident with ℓ.

    A. *Existence:* Show that there exists a point Q in $\mathcal{P}$ and a line m in $\mathcal{B}$ such that $m = PQ$ and $\ell \cap m = Q$.
    (*Hint:* Consider the plane β containing P and ℓ and investigate $\beta \cap \alpha$. Also, use the fact that any line in α intersects $\mathcal{K}$ in either 0 or 2 points.)

    B. *Uniqueness:* Show that the point–line pair obtained above is unique.
    (*Hint:* Suppose that (Q_1, m_1) and (Q_2, m_2) are two such pairs. Consider the points $\ell \cap \mathcal{K}$, $m_1 \cap \mathcal{K}$, and $m_2 \cap \mathcal{K}$.)

  (v) Deduce that $(\mathcal{P}, \mathcal{B}, \mathcal{I})$ is a generalised quadrangle and determine the parameters s and t.

Desargues' Theorem and non-Desarguesian planes

Desargues' Theorem plays a central role in the study of projective planes. But the theorem is not a consequence of the axioms, as we will soon see.

Definition 3.22 *A projective plane π is* **Desarguesian** *if Desargues' Theorem is valid in π for any pair of triangles which have distinct vertices, and are in perspective from a point.*

Example 3.23 As will be proved later, the EEP is Desarguesian.

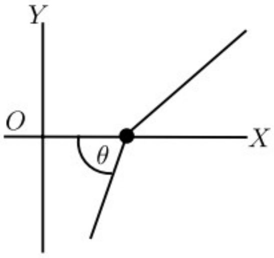

Example 3.24 (**A non-Desarguesian projective plane**) Consider the following triple $\pi = (\mathcal{P}, \mathcal{L}, \mathcal{I})$:

1. The *points* of π are the points of the extended Euclidean plane.
2. The *lines* of π are
   (a) the Euclidean lines perpendicular to the x-axis;
   (b) the Euclidean lines parallel to the x-axis;
   (c) the Euclidean lines with negative slopes;
   (d) the 'refracted' Euclidean lines given by:

$$y = \begin{cases} \frac{1}{2}(x-a)\tan\theta & y \geq 0, \\ (x-a)\tan\theta & y < 0, \end{cases}$$

   where $a \in \mathbb{R}$ and $0 < \theta < 90$;

   (e) the line at infinity of the extended Euclidean plane.
3. *Incidence* $\mathcal{I}$ in π is as follows:
   Incidence of the extended Euclidean plane and
   - parallel Euclidean lines of type (a), (b), or (c) intersect in their point at infinity of the extended Euclidean plane.
   - two lines of type (d) that do not intersect in the Euclidean plane, intersect in the point at infinity of their 'upper halves'.

The proof that π *is a projective plane* is left as an exercise for the reader.

Theorem: *π is non-Desarguesian.*
Sketch of Proof. In the diagram, the corresponding sides of triangles ABC and $A'B'C'$ are parallel, and therefore their intersections lie on the line at infinity, but the join of corresponding vertices are not concurrent. Hence the converse of Desargues' Theorem (and so Desargues' Theorem) is not valid in π. □

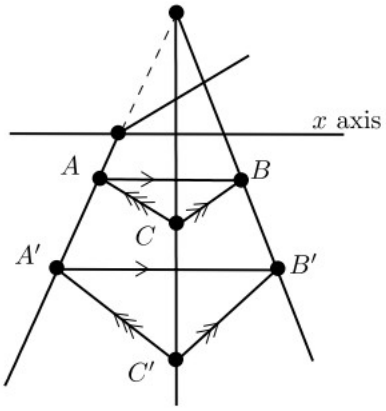

Note 8 *The Desargues' configuration requires* 10 *points, namely one vertex, two triangles, and three points of intersection of corresponding sides. The Fano plane has only seven points, and therefore Desargues' Theorem is meaningless in the Fano plane. In fact, if a quadrangle is chosen in the Fano plane to play the role of a vertex and a triangle, the remaining three points are collinear!*

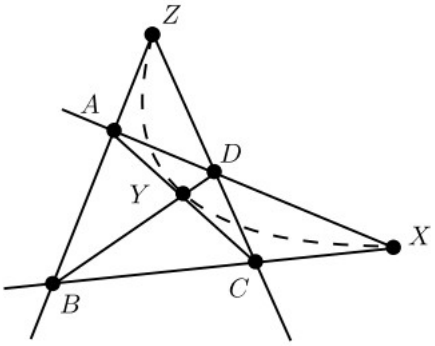

Theorem 3.25 *Let π be a finite projective plane of order $q \geq 4$. Then π contains 'many' Desarguesian configurations.*

Proof. Take any point V, any line ℓ with $V \notin \ell$. Take p, q, r, any three lines through V, and let

$$p \cap \ell = X, \quad q \cap \ell = Y, \quad r \cap \ell = Z.$$

Let N, L be points of ℓ, distinct from X, Y, and Z. This is possible since ℓ has at least five points.

Define a map $\sigma\colon \mathcal{A} \to \mathcal{L}$, where

$$\mathcal{A} = p \backslash \{V, X\}, \mathcal{L} = \ell \backslash \{N, L, Y, X\},$$

as follows:

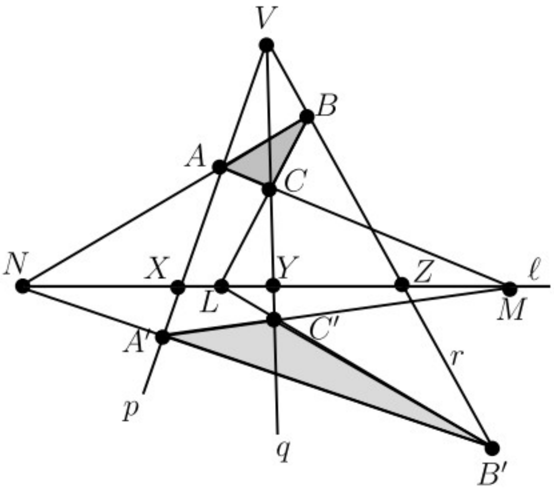

- Let $A \in \mathcal{A}$, and let $NA \cap VZ = B$. Thus $B \neq V, Z$.
- Join BL and let $BL \cap q = C$. Thus $C \neq Y$.
- Let $AC \cap \ell = M$. Thus $M \notin \{N, L, Y, X\}$.
- Define $\sigma \colon \mathcal{A} \to \mathcal{L}$ by $A^\sigma = M$.

Now $\mathcal{A}$ has $q-1$ elements, while $\mathcal{L}$ has $q-3$ elements. Therefore, σ is not one-to-one. Thus, there exist distinct points $A, A' \in \mathcal{A}$, with $\sigma(A) = \sigma(A')$. Let $B' = NA' \cap r$ and $C' = LB' \cap q$. Then, we have: triangles $ABC, A'B'C'$ are in perspective, with vertex V and axis ℓ. □

Note 9 *The word 'many' in the above theorem is justified by the fact that*

1. V *is any point.*
2. ℓ *is any line,* $V \notin \ell$.
3. p, q, r, *are any three lines through* V.
4. N, L *are any two points of* ℓ, *distinct from* X, Y, Z.

Theorem 3.26 *Desargues' Theorem is valid in a projective space of dimension* n, $n \geq 3$, *for any two non-coplanar triangles* $ABC, A'B'C'$, *in perspective from a point* V.

Proof. Let triangles $ABC, A'B'C'$ belong to the distinct planes π, α, respectively. We first note that the two triangles $ABC, A'B'C'$ belong to a space Σ of dimension 3 because $\Sigma = \pi \oplus V = \alpha \oplus V$. Therefore, $\pi \cap \alpha$ is a line. Now, AB and $A'B'$ are lines of the plane VAB, and therefore they intersect in a point. Similarly, BC and $B'C'$ intersect in a point; so also do the lines $AC, A'C'$. It follows that the points $AB \cap A'B'$, $BC \cap B'C'$, $CA \cap C'A'$ belong to the line $\pi \cap \alpha$. ☐

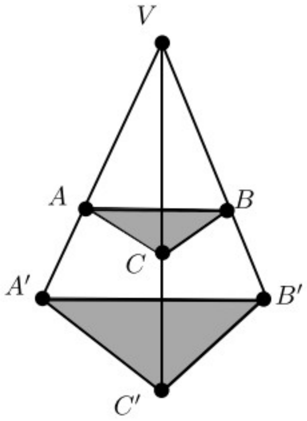

Theorem 3.27 *Let π be a projective plane, of order greater than two, immersed in a projective space S_n, of dimension $n \geq 3$. Then π is Desarguesian.*

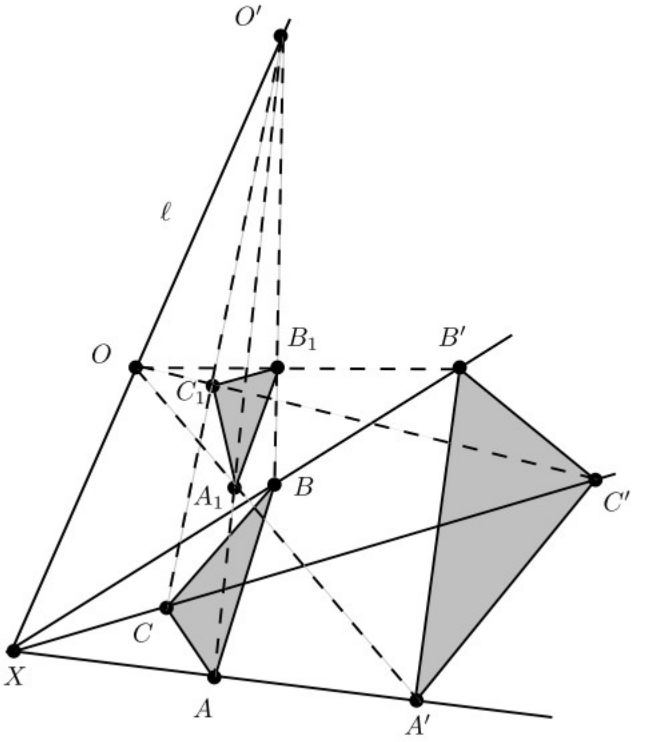

Proof. Let π be a plane (of order greater than two) of a projective space S_n, of dimension $n \geq 3$. Let $ABC, A'B'C'$ be two triangles of π, in perspective from a point X. Take any line $\ell \notin \pi$, through X. There are at least two other points (distinct from X) on ℓ. Let O, O' be two such points. Now $A'O, O'A$ intersect in a point since they are distinct lines of the plane OXA. Let $A'O \cap O'A = A_1$. Similarly, define the points B_1 and C_1. Note that the points A_1, B_1, and $C_1 \notin \pi$, and are not collinear.

Denote the plane $A_1B_1C_1$ by π_1. Thus $\pi \neq \pi_1$. Since triangles $ABC, A_1B_1C_1$ are in perspective from O', from the proof of Theorem 3.26, $A_1C_1 \cap AC \in \pi \cap \pi_1$. Let $A_1C_1 \cap \pi = P$. Similarly, triangles $A'B'C'$ and $A_1B_1C_1$ are in perspective from O; hence $A_1C_1 \cap A'C' \in \pi \cap \pi_1$. Hence $AC \cap A'C' = P$ lies on the line $\pi \cap \pi_1$. Similarly, $BC \cap B'C'$ and $AB \cap A'B'$ lie on the line $\pi \cap \pi_1$. In other words, the line $\pi \cap \pi_1$ is the axis of the perspectivity. ☐

Note 10 1. *A projective plane π need not be Desarguesian, although it could have many Desargues' configurations. Thus, it is not a consequence of the axioms that π is Desarguesian or not.*

2. *In Exercise 3.4(2b), it is proved that if Pappus' Theorem holds for any pair of lines of a projective plane π, (π is said to be* **Pappian***) then π is Desarguesian. The converse of this is not true for infinite planes.*

3. *It will be shown that any projective plane π can be coordinatised by means of a set $\mathcal{S}$ which has (as a consequence of the axioms of a projective plane only) an addition $+$ and a multiplication $\cdot$. But, unless π is assumed to have additional geometrical properties, $\{\mathcal{S}, +, \cdot\}$ has few algebraic properties; in fact, as will be indicated later, if the order of π is greater than two, then π is Desarguesian if and only if it can be coordinatised by a* skew field, *and Pappian if and only if it can be coordinatised by a field.*

   By definition, a **skew field** *has all the properties of a field, except that multiplication is not prescribed to be commutative. There exist infinite skew-fields which are not fields, the* quaternions *for example. However, Wedderburn's Theorem states that a finite skew-field is a field. Therefore,* a finite Desarguesian plane is Pappian.

   Since any plane of a projective space S_r, of dimension $r, r \geq 3$, is Desarguesian (Theorem 3.27), it follows that a non-Desarguesian plane cannot be embedded in a three-dimensional projective space.

Exercise 3.4

1. Consider three triangles in the Euclidean plane which have their corresponding sides parallel. Assume that the EEP is Desarguesian, and prove:
   (a) The three triangles are pairwise in perspective. That is, for each pair of triangles, the joins of corresponding vertices are concurrent.
   (b) If the vertices of the three perspectivities are distinct, then these vertices are collinear.

2. Let π be a given Pappian plane (i.e. a projective plane in which Pappus' Theorem is valid).

(a) Let ABC, DEF be two triangles in π such that lines AD, BE, CF meet in a point O *and* the lines AE, BF, CD meet in a point O'. Prove that AF, BD, CE meet in a point.
(*Hint:* Dualise.)

(b) Prove that π is Desarguesian.
(*Hint:* Let triangles $ABC, A'B'C'$ be in perspective from V. Let

$$\begin{aligned} B'C' \cap AC &= S, & B'A \cap CC' &= T, \\ B'A' \cap VS &= P, & BA \cap VS &= U, \\ BC \cap B'C' &= L, & CA \cap C'A' &= M, \\ AB \cap A'B' &= N. \end{aligned}$$

Use Pappus' Theorem to show that

(i) the points U, T, L are collinear by considering the lines $BB'V$ and SCA,

(ii) the points P, M, T are collinear by considering the lines $AA'V$ and $C'SB'$,

(iii) the points L, M, N are collinear by considering the lines $B'AT$ and UPS.

This question is hard unless you draw the three diagrams suggested by the hint separately (i.e. do not crowd any one diagram).)

3. Consider the following diagram and suppose the points and lines are those of a Pappian plane π. By considering the two triads of points $\{A_2, D, B_1\}$ and $\{B_2, C, A_1\}$, prove that the line QR passes though the point $a \cap b$.

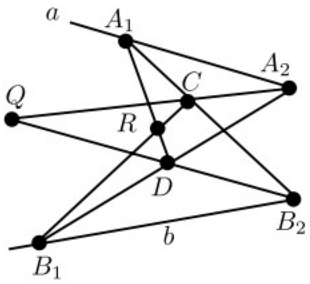

4. Let π be a Desarguesian plane. In π, let A, B, C lie on a line m and A', B', C' lie on a line n such that AA', BB', and CC' are concurrent in a point V. Prove that the points $AB' \cap A'B$, $AC' \cap A'C$, and $BC' \cap B'C$ lie on a line passing through the point of intersection of m and n.

(*Hint:* Identify triangles in perspective from the point V, and apply Desargues' Theorem.)

5. Prove that if the lines joining corresponding vertices of two tetrahedra in a three-dimensional projective space S_3 are concurrent, then the intersection of corresponding faces are coplanar. (A tetrahedron is a set of four non-coplanar points; the points are called vertices of the tetrahedron, the plane spanned by any three vertices is called a face, and the line joining two vertices is called an edge.)

(*Hint:* Desargues' Theorem is valid for non-coplanar triangles. Therefore, corresponding edges intersect in collinear points.)

6. Let π be a projective plane. Let D, C, X be points on a given line m of π. Let ℓ be any line through X, $\ell \neq m$. Let A, B be two points, distinct from X, on ℓ. Let $DA \cap CB = Z$, $BD \cap AC = U$.

   Define $Y = ZU \cap DC$ as a **harmonic conjugate** of X with respect to D, C. *Prove:* If π is Desarguesian, then the construction for a harmonic conjugate is unique.

   (*Hint:* Take another line ℓ' through X. Get Z', A', B', U'. You have to prove: $Z'U'$ passes through Y.
   Apply Desargues' Theorem or its converse to:

   (a) triangles ZAB and $Z'A'B'$;
   (b) triangles UAB and $U'A'B'$;
   (c) triangles AZU and $A'Z'U'$.)

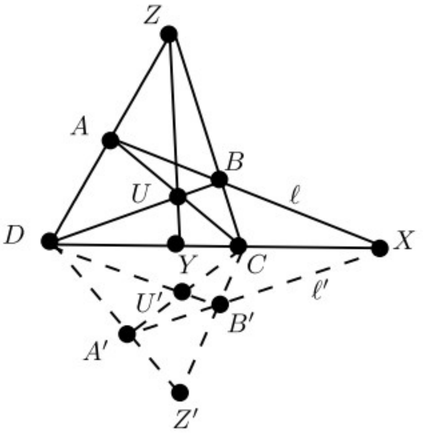

5 Coordinatising a projective plane

Preliminaries

There are many ways to coordinatise a projective plane. The most popular one is to attempt to make the plane look as close to a field plane as possible.

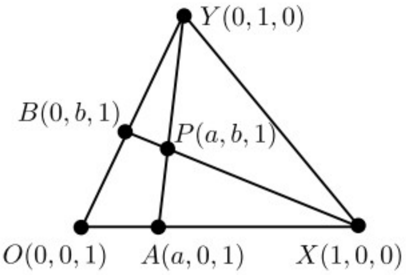

Consider the plane PG$(2,F)$. Let O, X, and Y be any three non-collinear points of PG$(2,F)$. Without loss of generality, we may take the homogeneous coordinates of O, X, and Y to be $(0,0,1)$, $(1,0,0)$, and $(0,1,0)$, respectively. Any point A on OX, with $A \neq X$, has homogeneous coordinates $(a,0,1)$, $a \in F$. Similarly, let $B=(0,b,1)$ be any point of OY, distinct from Y. Then $P = AY \cap XB$ has homogeneous coordinates $(a,b,1)$. Thus, any point P not on XY has non-homogeneous coordinates of type (a,b), with $a,b \in F$.

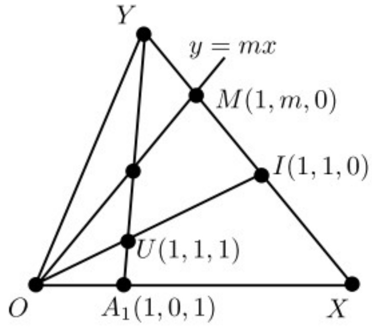

The line $y=mx$ intersects XY in the point $M=(1,m,0)$. We call (m) the non-homogeneous coordinate of M.

If U is the point $(1,1,1)$, then any point on OU, distinct from $I = OU \cap XY$, has non-homogeneous coordinates of type (a,a), $a \in F$.

The *coordinatisation* of a projective plane, which is introduced in the next section, retains the above-mentioned features.

Introducing coordinates

Let π be any projective plane. Let OXY be a triangle in π, and let I be any point on XY, with I distinct from X and Y. Let U be any point of OI, with U

distinct from O and I. Define γ to be an abstract set of elements in one-to-one correspondence with the points of OI distinct from I. Denote by 0 and 1 the elements of γ corresponding to O and U, respectively. Let $c \in \gamma$ correspond to the point C of OI, $C \neq I$. The *coordinates* of C are defined to be (c, c). Thus, O has coordinates $(0, 0)$, and U has coordinates $(1, 1)$.

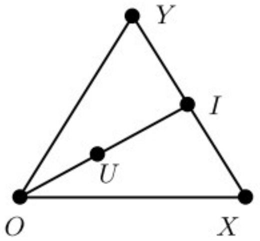

Let P be any point of π, not on XY. Let $A = YP \cap OI$ and $B = XP \cap OI$ have coordinates (a, a) and (b, b), respectively. *Define (a, b) to be the coordinates of P.* Thus, if $P \in OX$, its coordinates are of type $(a, 0)$; if $P \in OY$, its coordinates are of type $(0, b)$.

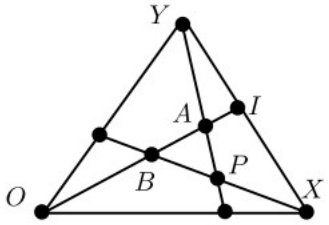

Note 29 *No two points of $\pi \backslash XY$ have the same coordinates.*

Note 30 *Two points have the same 'x' coordinate if and only if they are collinear with Y; two points have the same 'y' coordinates if and only if they are collinear with X.*

Let M be any point of XY, distinct from Y. Let $T = YU \cap OM$ have coordinates $(1, m)$, $m \in \gamma$. *Define (m) to be the coordinates of M.* Lastly, let ∞ be an abstract element. *Define (∞) to be the coordinates of Y.*

The coordinatisation of the projective plane π is now complete.

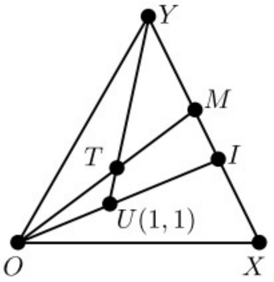

Addition in γ

The aim is to discover any algebraic structure that γ inherits from the fact that π is a projective plane.

Let $A = (x, 0)$ and $B = (0, y)$ be points on OX, OY, respectively. Let $M = YA \cap IB$. Let coordinates of M be (x, m). In γ we define the operation + (to be called **addition**) by

$$x + y = m.$$

Thus, M has coordinates $(x, x + y)$.

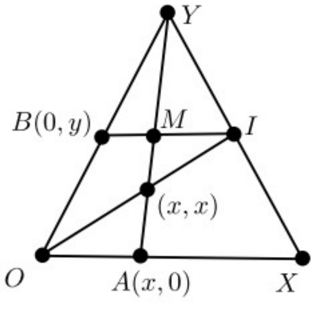

Example 5.1
1. If $M \in OI$, $M \neq I$, we have $m = x$, $y = 0$ and therefore $x + 0 = x$ for all $x \in \gamma$.
2. If $x = 0$, we have $0 + y = y$, for all $y \in \gamma$.
3. If any two of x, c, y are given, the third can be determined by the equation

$$x + c = y.$$

For example, suppose c and y are given. Let C be the point of coordinates $(0, c)$; let B be the point (y, y). Let $T = CI \cap BX$. Then the coordinates of T are $(x, y) = (x, x + c)$.

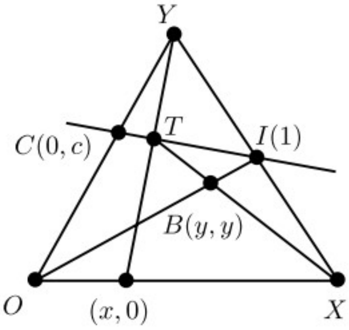

4. Suppose $a, b, u \in \gamma$ and $a \neq b$. Then the points $V = (0, u)$, $A' = (a, a + u)$, $B' = (b, b + u)$, and I are collinear. Suppose $a + u = b + u$. Then, the points $(a, a + u)$, $(b, b + u)$, having the same 'y' coordinates are collinear with X. This is a contradiction. Therefore, $a + u = b + u$ if and only if $a = b$, for all $u \in \gamma$.

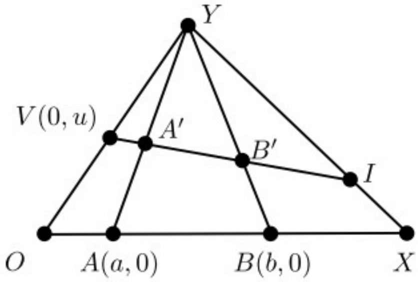

□

Definition 5.2 *A* **loop** $(H, \circ)$ *is a set* H *on which is defined a binary operation* $\circ$*, such that*

1. *For all* $x, y \in H$, $x \circ y \in H$.
2. *Given* $a, b \in H$*, each of the equations* $a \circ x = b$*, and* $y \circ a = b$ *has a unique solution in* H.
3. *There exists an element* $e \in H$*, such that* $e \circ x = x \circ e = x$*, for all* $x \in H$.

Theorem 5.3 $(\gamma, +)$ *is a loop.*

Exercise 5.1

1. Show that loops of orders $1, 2, 3, 4$ are groups.
   (*Hint:* Construct a multiplication table for a loop $(H, \circ)$ of order $2, 3$, or 4. In each case, a multiplication table for a group is obtained.)

2. The following multiplication table is that of a loop of order 5.

	1	a	b	ab	ba
1	1	a	b	ab	ba
a	a	1	ab	ba	b
b	b	ba	1	a	ab
ab	ab	b	ba	1	a
ba	ba	ab	a	b	1

Justify the statement: 'This loop is not a group'.

Multiplication in γ

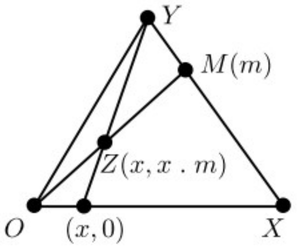

Let M be a point of XY, distinct from Y. Let its coordinate be (m). Let (x, y) be the coordinates of a point Z on OM, distinct from M. In γ, define the operation '$\cdot$' (to be called **multiplication**) by

$$z = x \cdot m$$

Example 5.4 Prove the following:

1. For all $a \in \gamma$, $a \cdot 0 = 0$ and $0 \cdot a = 0$.
2. If $m \neq 0$, then for all $a, b \in \gamma$

$$\begin{aligned} a \cdot m = b \cdot m &\Leftrightarrow a = b \\ m \cdot a = m \cdot b &\Leftrightarrow a = b \end{aligned}$$

   In particular, $a \cdot b = 0 \Leftrightarrow a = 0$ or $b = 0$.

3. If any two of x, m, c are given, the third can be uniquely determined by the equation

$$x \cdot m = c.$$

4. $a \cdot 1 = 1 \cdot a = a$ for all $a \in \gamma$.

Solution.

1. By definition of '$\cdot$', the point $(x, x \cdot 0)$ is the point $(x, 0)$, and the point $(0, 0 \cdot a)$ is the point $(0, 0)$.
2. Suppose $m \neq 0$ and $a \neq b$. Then the points $O = (0, 0)$, $P = (a, a \cdot m)$, $Q = (b, b \cdot m)$, and $M = (m)$ are collinear. Suppose also that $a \cdot m = b \cdot m$. Then the points P, Q, X are collinear; this is a contradiction. Hence $a \cdot m = b \cdot m$ if and only if $a = b$. The other parts are similarly proved.

3. Suppose m and c are given. Let M, C be points with coordinates (m) and (c, c), respectively. Let $P = OM \cap XC$. Then P has coordinates $(x, x \cdot m) = (x, c)$.
4. Let A', A be points of coordinates (a), and (a, a), respectively. Then $T = YU \cap OA'$ has coordinates $(1, a)$ and also $(1, 1 \cdot a)$. Also, A has coordinates $(a, a \cdot 1)$. Thus,

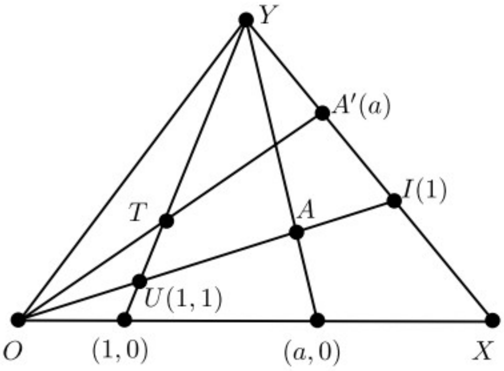

$$1 \cdot a = a \cdot 1 = a$$

for all $a \in \gamma$. □

Theorem 5.5 $(\gamma \backslash \{0\}, \cdot)$ *is a loop.*

Proof. This follows from Example 5.4(iii) and (iv). □

Hall Ternary Ring

Our definitions of '$\cdot$' and of '$+$' do not allow us to say that the point (x, y) of a line satisfies an equation of type $y = x \cdot m + c$.

Let ℓ be a line. Let (m) and $(0, c)$ be the coordinates of $M = \ell \cap XY$ and $C = \ell \cap OY$, respectively. Let $P = (x, y)$ be a point of ℓ. This incidence is expressed with the help of a ternary operation $(\circ, *)$ as follows:

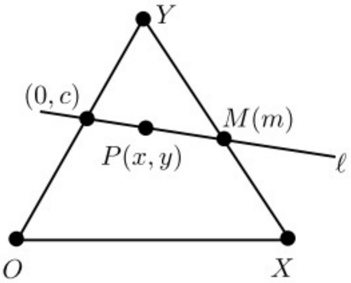

$$x \circ m * c = y.$$

Example 5.6 Show that in $[\gamma, \cdot, +, (\circ, *)]$,

1. $x \circ 1 * c = x + c$.
2. $x \circ m * 0 = x \cdot m$.
3. $0 \circ m * c = x \circ 0 * c = c$.
4. $1 \circ m * 0 = m \circ 1 * 0 = m$.
5. Given x, m, y, then there exists a unique c such that

$$x \circ m * c = y.$$

6. If $m_1 \neq m_2$, $c_1 \neq c_2$ are given, then there exists a unique x such that
$$x \circ m_1 * c_1 = x \circ m_2 * c_2.$$
7. If x_1, x_2, y_1, y_2 are given, then there exists a unique (m, c) such that
$$x_1 \circ m * c = y_1 \quad \text{and} \quad x_2 \circ m * c = y_2.$$

Solution. The reader will find the solutions easier if the corresponding diagrams are drawn.

1. $x \circ 1 * c = y$ if and only if $P = (x, y)$ lies on the line joining $C = (0, c)$ to $I = (1)$, if and only if $y = x + c$ by definition of $+$.
2. Let $P = YA \cap OM$, where $A = (x, 0)$, $M = (m)$. Then $x \circ m * 0 = x \cdot m$, since the coordinates of P are $(x, x \cdot m)$ by definition of $\cdot$, and $(x, x \circ m * 0)$, by definition of $(\circ, *)$.
3. The point P on CM has coordinates $(x, x \circ m * c)$. When P is at C, the coordinates of $C = (0, c) = (0, 0 \circ m * c)$. When $m = 0$, P has coordinates $(x, x \circ m * c)$ and also (x, c). Hence $0 \circ m * c = x \circ 0 * c = c$.
4. We have
$$\begin{aligned} 1 \circ m * 0 &= 1 \cdot m \quad \text{by (2).} \\ &= m \quad \text{by Example 5.4(4).} \end{aligned}$$
5. There exists a unique line joining the points $P = (x, y)$ and $M = (m)$. This line intersects OY in a unique point.
6. The line joining the points (m_1) and $(0, c_1)$ intersects the line joining (m_2) and $(0, c_2)$ in a unique point $(x, y) = (x, x \circ m_1 * c_1) = (x, x \circ m_2 * c_2)$.
7. There exists a unique line joining the points (x_1, y_1) and (x_2, y_2). Hence, there exists a unique (m, c) such that
$$x_1 \circ m * c = y_1 \quad \text{and} \quad x_2 \circ m * c = y_2. \qquad \square$$

Definition 5.7 1. *The system* $[\gamma, \cdot, +, (\circ, *)]$ *is called* **Hall's ternary ring**.

2. *If* $x \circ m * c = x \cdot m + c$, *Hall's Ternary Ring is said to be* **linear**.

The Veblen–Wedderburn Ring

It is a consequence of π being a projective plane that $(\gamma, +)$ and $(\gamma \backslash \{0\}, \cdot)$ are loops. Any further algebraic relation which γ may be made to satisfy causes the closure of some geometrical configuration.

Definition 5.8 *Let π be a projective plane. Let ABC, $A'B'C'$ be two triangles with distinct vertices such that AA', BB', CC' pass through a point V. Let ℓ be a line through V such that ℓ contains two of the points $L = BC \cap B'C'$, $M = AC \cap A'C'$, and $N = AB \cap A'B'$.*

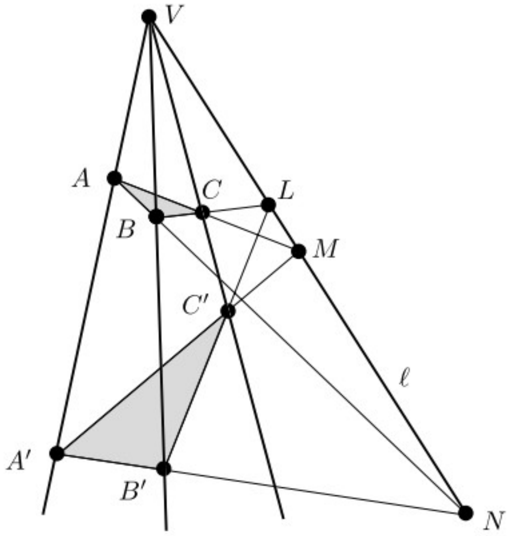

Then π is said to satisfy the Little Desargues Axiom with respect to triangles ABC, $A'B'C'$ in perspective from the vertex V with ℓ as axis*, if all three points L, M, N necessarily lie on ℓ.*

Example 5.9 A field plane $\mathrm{PG}(2, F)$ satisfies the Little Desargues Axiom with respect to any pair of triangles in perspective from any point V, and any axis ℓ, $V \in \ell$.

Theorem 5.10 *Let π be a projective plane which satisfies the Little Desargues Axiom with respect to triangles in perspective from Y and with YX as axis, then*

1. $[\gamma, +]$ *is a group.*
2. $x \circ m * c = x \cdot m + c$ *(i.e. γ is linear).*

Proof.

1. We have to prove that for all $a, b, c \in \gamma$,

$$a + (b + c) = (a + b) + c.$$

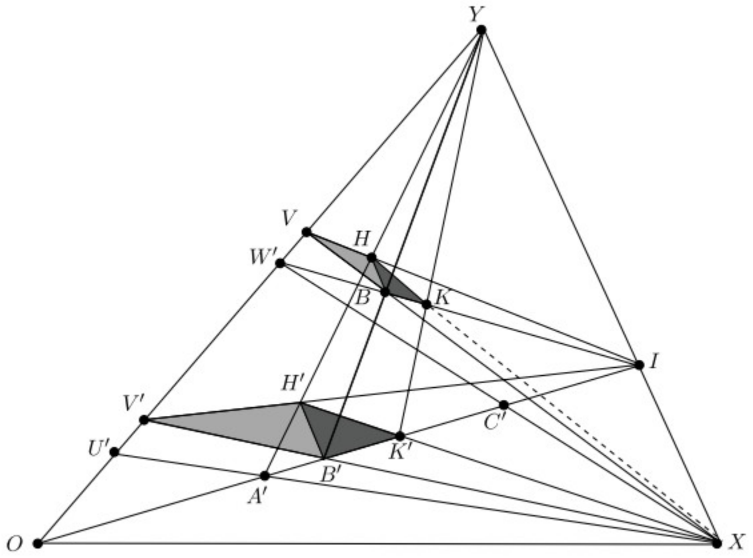

To do this, we construct two points H, K of coordinates $(a, a+(b+c))$ and $(a+b, (a+b)+c)$, respectively, and show that the Little Desargues Axiom implies that H, K, X are collinear. Since points collinear with X have the same 'y' coordinate, the result follows.

Let $A' = (a,a)$, $B' = (b,b)$, $C' = (c,c)$ be three distinct points of the line OI, distinct from I. The following points are successively constructed, and their coordinates are identified:

$$\begin{aligned}
U' &= OY \cap XA' &&= (0,a),\\
V' &= OY \cap XB' &&= (0,b),\\
W' &= OY \cap XC' &&= (0,c),\\
H' &= YA' \cap V'I &&= (a, a+b),\\
B &= YB' \cap W'I &&= (b, b+c),\\
V &= XB \cap OY &&= (0, b+c),\\
H &= YA' \cap VI &&= (a, a+(b+c)),\\
K' &= H'X \cap OI &&= (a+b, a+b),\\
K &= W'I \cap YK' &&= (a+b, (a+b)+c).
\end{aligned}$$

Consider triangles VHB and $V'H'B'$. They are in perspective from Y, and

$$\begin{aligned}
VH \cap V'H' &= I \in XY,\\
VB \cap V'B' &= X \in XY.
\end{aligned}$$

By the Little Desargues Axiom, we have

$$HB \cap H'B' \in XY.$$

Next, consider triangles HBK and $H'B'K'$. They are in perspective from Y and both $HB \cap H'B'$ and $BK \cap B'K' = I$ belong to XY. By the Little Desargues Axiom,

$$HK \cap H'K' \in XY.$$

Since $H'K' \cap XY = X$, it follows that $HK \cap H'K' = X$. In other words, H, K, X are collinear.

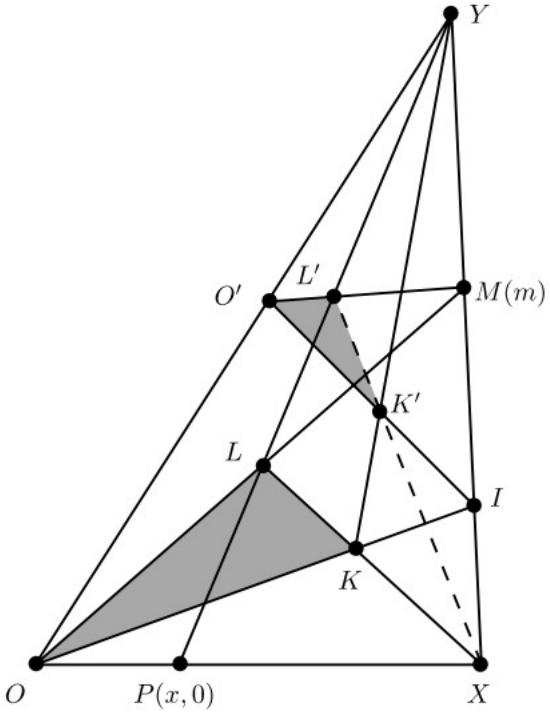

2. Starting with the points $P = (x, 0)$, $O' = (0, c)$, and $M = (m)$, the following points are successively constructed, and their coordinates are identified:

$$\begin{aligned} L &= YP \cap OM = (x, x \cdot m), \\ K &= LX \cap OI = (x \cdot m, x \cdot m), \\ K' &= O'I \cap YK = (x \cdot m, x \cdot m + c), \\ L' &= YP \cap O'M = (x, x \circ m * c). \end{aligned}$$

Now triangles OLK, $O'L'K'$ are in perspective from Y, and

$$\begin{aligned} OL \cap O'L' &= M \in YX, \\ OK \cap O'K' &= I \in YX. \end{aligned}$$

By the Little Desargues Axiom, we have

$$LK \cap L'K' \in YX.$$

Since $LK \cap YX = X$, it follows that $L'K' \cap YX = X$. Hence L', K', being collinear with X, have the same 'y' coordinate; and so

$$x \circ m * c = x \cdot m + c. \qquad \square$$

The following theorem is stated without proof. The rationale is to indicate the improvements in the algebraic properties of γ as a result of assuming certain geometrical properties of the projective plane π.

Theorem 5.11 *If the Little Desargues Axiom holds for triangles in perspective from any point* P *of* XY*, with* XY *as axis, then* $[\gamma, +, \cdot, (\circ, *)]$ *is such that*

1. $[\gamma, +]$ *is an abelian group.*
2. $x \circ m * c = x \cdot m + c$*, for all* $x, m, c \in \gamma$*.*
3. $[\gamma \backslash \{0\}, \cdot]$ *is a loop with 1 as unity.*
4. $(a + b) \cdot m = a \cdot m + b \cdot m$ *for all* $a, b, m \in \gamma$*.*
5. *If* $a \neq b$*, then the equation* $x \cdot a = x \cdot b + c$ *has a unique solution* x*, for all* $a, b, c \in \gamma$*.*

Note 31 *We shall, from now on, write* ab *instead of* $a \cdot b$*.*

Definition 5.12

1. *If the Hall Ternary Ring* $\gamma = (\gamma, +, \cdot)$ *has the properties described in Theorem 5.11, then it is called a* **Veblen–Wedderburn Ring** *or a* **quasifield***.*
2. *A quasifield with an associative multiplication is called a* **nearfield***.*
3. *A quasifield with the left distributive law is called a* **semifield***.*
4. *A* **skew field** *has all the properties of a field, except that multiplication is not prescribed to be commutative. Thus a skew field is both a semifield and a nearfield.*
5. *A semifield* γ *is an* **alternative division ring** *if for all* $a, b \in \gamma$

$$
\begin{aligned}
a(ab) &= (aa)b, \\
(ab)b &= a(bb).
\end{aligned}
$$

We quote, without proof, the following two important theorems:

Theorem 5.13 *(***Wedderburn's Theorem***) A* finite *skew field is a field.*

Theorem 5.14 *(***Artin-Zorn Theorem***) A* finite *alternative division ring is a field.*

For comparison sake, we tabulate some results. [Here $a.b$ is written ab, $\gamma \backslash \{0\}$ is written as γ^*, and Little Desargues' Theorem is denoted by L.D.]

Geometric Properties	Properties of Hall Ternary Ring γ
L.D., vertex Y Axis XY	$x.m * c = xm + c$ $(\gamma, +)$ is a group
L.D., vertex anywhere on axis XY	$(\gamma, +)$ is abelian $(\gamma^*, \cdot)$ is a loop with 1 as unity $(a+b)m = am + bm$ $a \neq b \Rightarrow$ unique solution to $xa = xb + c$
L.D., axis any line through Y, vertex anywhere on this line	$(\gamma^*, \cdot)$ has inverses which are two-sided $a \neq 0 \Rightarrow a^{-1}(ab) = b$ The Moufang Identity $a[c(ab)] = [a(ca)]b$ holds. So does the alternative law $a(ab) = (aa)b$
L.D., any axis, vertex anywhere on such axis	$a \neq 0 \Rightarrow a^{-1}(ab) = b = (ba)a^{-1}$ Moufang Identity is replaced by $a(ab) = (aa)b$ $(ba)a = b(aa)$ The plane is now called a Moufang Plane. [Harmonic construction yields a unique conjugate point (line) provided Ring does not have characteristic 2]
Plane is Desarguesian	γ is a skew field

Note 32 *As we have seen, Pappus' Theorem implies Desargues' Theorem; the converse is not true. (*Hessenberg, 1905.*)*

Exercise 5.2

1. If $[\gamma, +, \cdot]$ is a quasifield, then for all $a, b \in \gamma$
$$-(ab) = (-a)b.$$
2. If $(\gamma, +, \cdot)$ has all the properties of a quasifield, except possibly commutative addition, then γ is a quasifield.
3. Let π be a projective plane, with γ as coordinatising set. Let addition $+$ in γ be defined as in Section 'Addition in γ'. (Recall Theorem 5.10: to be a group, $[\gamma, +]$ requires the validity of the Little Desargues' configuration, vertex Y, axis ℓ_∞.) Display a configuration whose validity in π ensures that
$$(a+a)+b = a+(a+b)$$
for all choices of a, b from γ. Show that the configuration has eleven points and may be regarded as consisting of two quadrilaterals Q, Q' where,
   (a) Q, Q' are in perspective from a point;
   (b) the pairs of opposite sides of each quadrilateral are concurrent.

11.04.2020

A) 5 Coordinatising a projective plane

Pag. 99 (112)

6 Non-Desarguesian Planes

Pag. 110 (123)

Pag. 112 (125)

Example 6.4

Prove that if a collineation σ ~~that~~ of a projective plane π is such

1. some line l is fixed pointwise;
2. two points $P, Q \in l$ are fixed,

then $\sigma = 1$.

B) Books_67_2020/
Mathematics of Shape Description/
Pijush K. Ghosh _ Mathematics of Shape Description ... pdf

Pag. 77 (96)

3. Algebraic Structures for Shape Description

5. Arithmetics of Geometric Shape
Pag. 165 (184)

6 Non-Desarguesian planes

(V,ℓ)-perspectivities

The aim of this chapter is to construct some classes of non-Desarguesian projective planes. The starting point we use here for the construction of such planes is the assumption that the planes have a certain type of collineation group.

We adopt the following notation: If τ and σ are collineations of a projective space Π, and P is any point of Π, then $\tau\sigma(P)$ or $P^{\tau\sigma}$ are used to denote $\tau(\sigma(P))$ (i.e. σ is first applied to P, then τ is applied to $\sigma(P)$.)

Definition 6.1 *Let σ be a collineation of a projective plane π. Then*

1. *A point P is* **fixed** *by σ if and only if $P^\sigma = P$.*
2. *A line ℓ is* **fixed** *by σ if and only if $\ell^\sigma = \ell$.*
3. *A line ℓ is* **fixed pointwise** *by σ if and only if $P^\sigma = P$ for all points $P \in \ell$.*
4. *A point P is* **fixed linewise** *by σ if and only if $\ell^\sigma = \ell$ for all lines ℓ through P.*

Example 6.2 Let σ be the homography of a field plane $\mathrm{PG}(2,F)$ defined by the matrix

$$A = \begin{bmatrix} a & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad a \in F\backslash\{0,1\}.$$

Investigate whether σ has any fixed points or lines.

Solution. A point X is fixed if

$$X \mapsto AX = \lambda X, \text{ for some } \lambda \in F.$$

Thus the fixed points are the eigenvectors of A. Now

$$|\lambda I - A| = \begin{vmatrix} \lambda - a & 0 & 0 \\ 0 & \lambda - 1 & 0 \\ 0 & 0 & \lambda - 1 \end{vmatrix} = (\lambda - a)(\lambda - 1)^2.$$

So A has two eigenvalues $1, a$. Corresponding to the eigenvalue $\lambda = 1$, solving $(I - A)X = 0$ gives the eigenvectors and therefore the fixed points

$$\{(x,y,z) = (0,r,t) \mid r,t \in F, \text{ not both zero}\}.$$

Thus every point on the line ℓ of equation $x = 0$ is fixed. In other words, ℓ is fixed pointwise.

Corresponding to the eigenvalue $\lambda = a$, solving $(aI - A)X = 0$ gives the fixed point $V = (1, 0, 0)$.

By Example 4.30, if $L = [\ell, m, n]$ is a line, then

$$\sigma\colon L \mapsto LA^{-1}.$$

Therefore, a line L is fixed if

$$L \mapsto LA^{-1} = \lambda L, \quad \text{for some } \lambda \in F.$$

Thus L is a fixed line if L^t is an eigenvector of $(A^{-1})^t$. In our example, the matrix A is a diagonal matrix, and therefore $(A^{-1})^t = A^{-1}$. Since A and A^{-1} have the same eigenvectors, the fixed lines of σ are $[1, 0, 0]$, and

$$\{[\ell, m, n] = [0, r, t] \mid r, t \in F, \text{ not both zero}\}.$$

Thus the line $[1, 0, 0]$ is fixed; also, every line through the point $V = (1, 0, 0)$ is fixed, and therefore V is fixed linewise. □

Example 6.3 Recall from Example 4.18 that the collineation $\sigma\colon i \mapsto i + 1$ of the Fano plane has no fixed point. By taking the points $0, 1, 2, 5$ to be the points $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$, $(1, 1, 1)$, respectively, then σ is seen to be the homography with matrix

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

The characteristic polynomial of A is

$$|\lambda I - A| = \begin{vmatrix} \lambda & 0 & 1 \\ 1 & \lambda & 1 \\ 0 & 1 & \lambda \end{vmatrix} = \lambda^3 + \lambda + 1.$$

This is irreducible in GF(2). Thus there are no fixed points in PG(2, 2); the fixed points are in the cubic extension PG$(2, 2^3)$.

Exercise 6.1 Let σ be a collineation of a projective plane π. Prove that

1. If the points P and Q are fixed by σ, so is the line PQ.
2. If the lines ℓ and m are fixed by σ, so is the point $\ell \cap m$.
3. If the lines ℓ and m are fixed pointwise by σ, then σ is necessarily the identity collineation i.

   (*Hint:* Let P be any point of π not on ℓ or m. Take two lines a, b through P, but not through $\ell \cap m$. Prove successively that a, b, and P are fixed.)
4. If the points P and Q are fixed linewise by σ, then σ fixes every point of the plane π; σ is the identity collineation i.

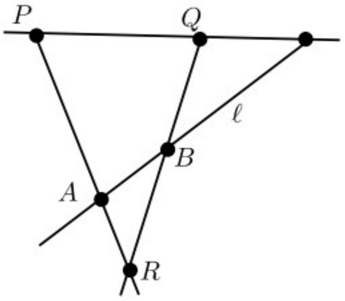

Example 6.4 Prove that if a collineation σ of a projective plane π is such that

1. some line ℓ is fixed pointwise;
2. two points $P, Q \notin \ell$ are fixed,

then $\sigma = i$.

Solution. Let R be a point of $\pi\backslash\{\ell, PQ\}$. Let $\ell \cap RP = A$ and $\ell \cap RQ = B$. Since the points P, Q, A, B are fixed, it follows that $R = PA \cap QB$ is fixed. Similarly, every point $S \in PR\backslash\{A, P\}$ is fixed. So the line PR is fixed pointwise. Thus we have two lines (namely ℓ and PR fixed pointwise by σ). Hence by Exercise 6.1(3), $\sigma = i$. □

Exercise 6.2 Let σ be a collineation of a projective plane π. Let $\mathcal{P}$, $\mathcal{L}$ denote the set of all fixed points and fixed lines, respectively, and let $\mathcal{I}$ be the natural incidence. Assume that $(\mathcal{P}, \mathcal{L}, \mathcal{I})$ contains a quadrangle. Prove that $(\mathcal{P}, \mathcal{L}, \mathcal{I})$ is a subplane of π.

Theorem 6.5 *Let σ be a collineation of a projective plane π, $\sigma \neq i$, fixing a line ℓ pointwise. Then there exists a point V fixed linewise by σ.*

Proof. By Example 6.4, σ can fix at most one point P which is not on ℓ.

Case 1: Suppose σ fixes a point $V \notin \ell$. Let P be any point of ℓ. Then $(VP)^\sigma = V^\sigma P^\sigma = VP$. Thus V is fixed linewise.

Case 2: Suppose no point of $\pi\backslash\{\ell\}$ is fixed by σ. Let P be any point of $\pi\backslash\{\ell\}$. Then $P^\sigma \neq P$. Let $V = PP^\sigma \cap \ell$. Now

$$\begin{aligned}(VP)^\sigma &= V^\sigma P^\sigma \\ &= VP^\sigma \\ &= VP, \quad \text{since } V, P, P^\sigma \text{ are collinear.}\end{aligned}$$

Hence $VP = PP^\sigma$ is fixed. Similarly, if Q is any point of $\pi\backslash\{\ell, VP\}$, then QQ^σ is fixed. Therefore, $PP^\sigma \cap QQ^\sigma$ is a fixed point and so lies on ℓ. Hence $PP^\sigma \cap QQ^\sigma = V$. Thus V is fixed linewise. □

Note 33 *Dually we have: if σ fixes a point linewise, then σ fixes a line pointwise.*

Definition 6.6 *Let σ be a collineation of a projective plane π that fixes a line ℓ pointwise and a point V linewise. Then:*

1. *σ is called a* **(V, ℓ)-perspectivity** *or a* **central collineation**.
2. *If $V \in \ell$, σ is called an* **elation**.

3. *If $V \notin \ell$, σ is called a* **homology**.
4. *V is called the* **vertex** *of σ and ℓ is called the* **axis** *of σ.*

Example 6.7 In the homography σ in Example 6.2, $V = (1, 0, 0)$ is the vertex of σ and the line ℓ of equation $x = 0$ is the axis of σ. Also, σ is a homology as $V \notin \ell$.

Example 6.8 Prove that the homography σ of a field plane $\mathrm{PG}(2, F)$ with matrix

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

is an elation. Find the vertex and axis of σ.

Solution. The characteristic polynomial of the matrix A is

$$|\lambda I - A| = \begin{vmatrix} \lambda - 1 & 0 & -1 \\ 0 & \lambda - 1 & -1 \\ 0 & 0 & \lambda - 1 \end{vmatrix} = (\lambda - 1)^3.$$

The fixed points of σ are the solutions of $(I - A)X = 0$, that is,

$$\begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Therefore, the line ℓ of equation $z = 0$ is fixed pointwise by σ, and no point of $\mathrm{PG}(2, F) \backslash \{\ell\}$ is fixed. To find the vertex V, let P be any point of $\mathrm{PG}(2, F) \backslash \{\ell\}$. Then by Case 2 in the proof of Theorem 6.5, $V = \ell \cap PP^\sigma$. Let $P = (0, 0, 1)$. Then P^σ is given by

$$AP = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

So $V = PP^\sigma \cap \ell = (1, 1, 0)$. □

Exercise 6.3 Prove that, for a given point V and a given line ℓ of a projective plane π, with respect to composition of functions, the set of all (V, ℓ)-perspectivities of π is a group.

Note 34 *The identity collineation i is considered both as a homology and an elation.*

Example 6.9 Let σ be a (V, ℓ)-perspectivity and τ a collineation of a projective plane π. Prove that $\tau\sigma\tau^{-1}$ is a (V^τ, ℓ^τ)-perspectivity of π.

Solution. Let P be a point of the line ℓ^τ. Let $Q = \tau^{-1}(P)$. Therefore $Q \in \ell$. So,

$$\tau\sigma\tau^{-1}(P) = \tau(\sigma(\tau^{-1}(P))) = \tau(\sigma(Q)) = \tau(Q) = P.$$

Hence ℓ^τ is fixed pointwise by $\tau\sigma\tau^{-1}$. Similarly, V^τ is fixed linewise by $\tau\sigma\tau^{-1}$. □

Theorem 6.10 *A (V,ℓ)-perspectivity σ of a projective plane π is uniquely determined by the vertex V, the axis ℓ and the image $Q = P^\sigma$ of* one *point P, $P \notin \ell$, $P \neq V$.*

Proof. Note that as $P \notin \ell$ and $P \neq V$, necessarily $Q \neq V$ and $Q \notin \ell$.

We first prove that such a σ must be unique. Suppose there are two (V,ℓ)-perspectivities σ, τ such that $P^\sigma = P^\tau = Q$. Then $\tau^{-1}\sigma$ fixes P, V and fixes ℓ pointwise.

By Example 6.4, $\tau^{-1}\sigma = i$, and therefore $\tau = \sigma$.

We must now show that we can uniquely find the image R^σ of any point R of $\pi\backslash\{V,\ell\}$. Let R be any point of $\pi\backslash\{V,\ell\}$ not on the line $VPP^\sigma = VPQ$. Let $PR \cap \ell = U$. Then,

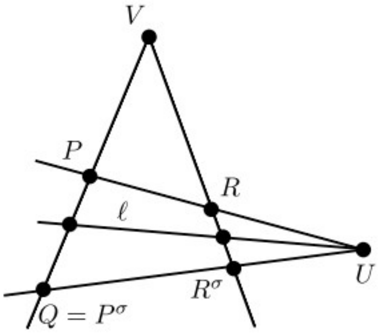

$$\begin{aligned}
R^\sigma &= (PU \cap VR)^\sigma \\
&= (PU)^\sigma \cap (VR)^\sigma \\
&= P^\sigma U^\sigma \cap (VR), \quad \text{as } V \text{ is fixed linewise,} \\
&= QU \cap VR, \quad \text{as } P^\sigma = Q \text{ and } U \in \ell \text{ is fixed.}
\end{aligned}$$

Lastly, if S is a point on the line VPP^σ distinct from V and not in ℓ, then we use R, R^σ in a similar way to find S^σ. □

Example 6.11 Find the equation of the (V,ℓ)-perspectivity σ of a field plane $\mathrm{PG}(2,F)$, where $V = (1,0,0)$, ℓ is the line $x = 0$, $P = (1,1,1)$, $Q = (1,b,b)$, $(b \in F)$, and $P^\sigma = Q$.

Solution. Using the same notation as in the proof of Theorem 6.10, let $R = (x_1, y_1, z_1)$.

Then $U = (0, y_1 - x_1, z_1 - x_1)$. The equation of the line QU is

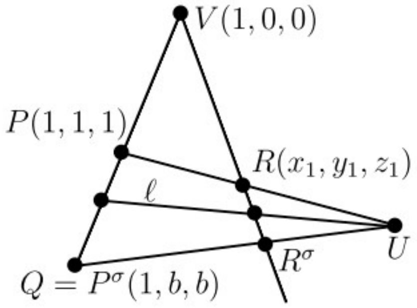

$$0 = \begin{vmatrix} x & y & z \\ 1 & b & b \\ 0 & y_1 - x_1 & z_1 - x_1 \end{vmatrix}$$

$$\begin{aligned}
&= [b(z_1 - x_1) - b(y_1 - x_1)]x - (z_1 - x_1)y + (y_1 - x_1)z \\
&= b(z_1 - y_1)x - (z_1 - x_1)y + (y_1 - x_1)z.
\end{aligned}$$

The equation of line VR is $z_1y - y_1z = 0$.
The point of intersection of QU and VR is

$$\begin{aligned} R^\sigma &= (y_1(z_1 - x_1) - z_1(y_1 - x_1), by_1(z_1 - y_1), bz_1(z_1 - y_1)) \\ &= (x_1(z_1 - y_1), by_1(z_1 - y_1), bz_1(z_1 - y_1)) \\ &= (x_1, by_1, bz_1). \end{aligned}$$

Thus σ is the homology with matrix

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & b \end{bmatrix}. \qquad \square$$

Example 6.12 Let σ be a (V, ℓ)-perspectivity of $\mathrm{PG}(2, q), q = p^h$. Since σ is a collineation of $\mathrm{PG}(2, q)$, by the fundamental theorem of field planes (see Theorem 4.27), it has an equation of type $\rho X' = AX^\alpha$, where α is an automorphism of $GF(q)$. Recall that every automorphism of $\mathrm{GF}(q)$ is of type $a \mapsto a^{p^i}, i = 0, 1, \ldots, (h-1)$. Show that α is the identity automorphism.

Solution. Without loss of generality, take ℓ to be the line of equation $x = 0$. Let the (V, ℓ)-perspectivity σ have the equation

$$X \mapsto AX^\alpha = \rho X',$$

where α is an automorphism of F and

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \quad a_{ij} \in F,\ |A| \neq 0.$$

Since ℓ is fixed pointwise by σ,

$$(0, 0, 1)^\sigma = (0, 0, 1) \quad \text{and} \quad (0, 1, u)^\sigma = (0, 1, u).$$

Therefore,

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}^\alpha = \rho \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \text{for some non-zero } \rho \in F.$$

Since $1^\alpha = 1$, we have $a_{13} = a_{23} = 0$ and $a_{33} \neq 0$.
For all $u \in F$, we have:

$$\begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ u \end{bmatrix}^\alpha = \rho \begin{bmatrix} 0 \\ 1 \\ u \end{bmatrix}, \quad \text{for some non-zero } \rho \in F.$$

Put $u = 0$, and we have $a_{12} = a_{32} = 0$ and $a_{22} \neq 0$.
So, for any non-zero $u \in F$, $a_{22}/(a_{33}u^\alpha) = 1/u$, and therefore $a_{22}u = a_{33}u^\alpha$.

Suppose $\alpha\colon a \mapsto a^{p^i}, i = 1, 2, \ldots,$ or $(h-1)$. Then the polynomial $a_{22}x = a_{33}x^{p^i}$ is of degree at most p^{h-1} but is satisfied by the $p^h - 1$ non-zero elements of GF(q). In other words, the polynomial has more zeros than its degree, making it the zero polynomial. Therefore $a_{33} = a_{22} = 0$, contradicting earlier results. Hence α is the identity automorphism. Moreover, $a_{33} = a_{22}$. □

Note 35 *In the above solution, the matrix A of σ is*

$$A = \begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & 0 & a_{22} \end{bmatrix}, \quad a_{ij} \in F, \ |A| \neq 0.$$

As noted before, the eigenvectors of A are the fixed points of σ.

Case 1: $a_{11} \neq a_{22}$. *The eigenvalues are a_{11} and a_{22}. The eigenvalue a_{11} gives rise to the axis $x = 0$; the eigenvalue a_{22} gives rise to the vertex $(a_{11} - a_{22}, a_{21}, a_{31})$, which does not lie on the axis since $a_{11} \neq a_{22}$. Thus, σ is a homology.*

Case 2: $a_{11} = a_{22}$. *There is only one eigenvalue, namely a_{11}. The axis ℓ is the line $x = 0$, as in Case 1. To find the vertex V of σ, let m be the line joining $(1, 0, 0)$ to $(1, 0, 0)^\sigma = (a_{11}, a_{21}, a_{31})$. Then, $V = m \cap \ell = (0, a_{21}, a_{31})$. In this case, σ is an elation.*

Exercise 6.4

1. Let σ_1, σ_2 be two homographies of PG$(2, F)$ whose matrices are

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & k \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 0 & a \\ 0 & 1 & b \\ 0 & 0 & 1 \end{bmatrix},$$

respectively ($k \neq 0, 1$ and $a, b, k \in F$).

(a) Find their fixed points and their fixed lines.

(b) Verify that σ_1 is a homology and σ_2 is an elation.

2. Let σ be a central collineation with vertex P. Suppose σ also fixes two lines ℓ, m not on P. Show that σ is the identity.

3. Let σ be a homography of the field plane PG$(2, F)$ with matrix

$$A = \begin{bmatrix} 2 & 1 & -2 \\ 0 & 0 & 4 \\ 0 & -1 & 4 \end{bmatrix}.$$

Find the fixed points of σ and hence determine whether σ is a central collineation, and if so whether σ is an elation or a homology.

4. Let σ be a collineation that fixes the point P linewise. Show that if a line ℓ not through P is fixed, then ℓ is fixed pointwise. Hence deduce that if ℓ, m are two lines not through P which are fixed by σ, then σ is the identity.

5. Let σ be a homography of the field plane π_F with matrix A where

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Determine the action of σ on the lines of $\mathrm{PG}(2, F)$. Hence or otherwise determine whether σ is an elation or a homology of $\mathrm{PG}(2, F)$.

6. Let σ be a homography of $\mathrm{PG}(2, F)$. Show that the line $[0, 0, 1]$ is fixed pointwise if and only if σ has matrix

$$A = \begin{bmatrix} 1 & 0 & r \\ 0 & 1 & s \\ 0 & 0 & t \end{bmatrix},$$

for some $r, s, t \in F$ with $t \neq 0$. Find the vertex of σ in the case $t \neq 1$.

7. Let σ be a homography of $\mathrm{PG}(2, F)$. Let the matrix of σ be

$$A = \begin{bmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1, \end{bmatrix}, \quad a, c \neq 0.$$

Show that σ has exactly one fixed point and one fixed line (and hence σ is not a perspectivity).

8. Let V be the vertex and ℓ the axis of a homology σ of $\mathrm{PG}(2, F)$. Let P be any point not fixed by σ. Let $VP \cap \ell = Q$. Prove that the cross-ratio $(V, Q; P, P^\sigma)$ is constant.

(*Hint:* consider the homology σ_1 of Question 1.)

9. Let Γ be the collineation group of a projective plane π. For points V, P, Q, lines ℓ, k, let

   (a) $\Gamma_{(V,\ell)}$ be the set of all (V, ℓ)-perspectivities of π.

   (b) $\Gamma_{(k,\ell)}$ be the set of all (V, ℓ)-perspectivities of π with $V \in k$.

   (c) $\Gamma_{(P,Q)}$ be the set of all (P, ℓ)-perspectivities of π, for all lines ℓ through Q.

   (d) $\Gamma_{(\ell)}$ be the set of all (V, ℓ)-perspectivities of π, for all points V of π.

   (e) $\Gamma_{(P)}$ be the set of all (P, ℓ)-perspectivities of π, for all lines ℓ of π.

   Show that for all choices of P, Q, ℓ, k

   (i) $\Gamma_{(k,\ell)}, \Gamma_{(P,Q)}, \Gamma_{(\ell)}, \Gamma_{(P)}$ are subgroups of Γ.

   (ii) $\Gamma_{(\ell,\ell)} \lhd \Gamma_{(\ell)}, \Gamma_{(P,P)} \lhd \Gamma_{(P)}$.

   (*Note*: It is sufficient to prove $\Gamma_{(k,\ell)}, \Gamma_{(\ell)}$ are subgroups and $\Gamma_{(\ell,\ell)} \lhd \Gamma_{(\ell)}$. The rest follows by duality.)

(*Hint*: To prove $\Gamma_{(k,\ell)}$ is a group, consider the cases $k = \ell$ and $k \neq \ell$ separately. Take $\alpha, \beta \in \Gamma_{(k,\ell)}$, what can you say about $\alpha\beta^{-1}$? To prove $\Gamma_{(\ell,\ell)} \lhd \Gamma_{(\ell)}$, what can you say about $\gamma^{-1}\Gamma_{(x,\ell)}\gamma$, for $X \in \ell$ and $\gamma \in \Gamma_{(\ell)}$?)

10. Let ℓ be a given line of a projective plane π. Suppose
    (a) A, B are two points on ℓ;
    (b) m is another line through B;
    (c) $|\Gamma_{(A,\ell)}| > 1$ and $|\Gamma_{(B,m)}| > 1$.

    Prove that $|\Gamma_{(B,\ell)}| > 1$.

    (*Hint*: Take $\alpha \in \Gamma_{(A,\ell)}, \beta \in \Gamma_{(B,m)}$ with $\alpha \neq i$ and $\beta \neq i$. Consider the collineation $\beta\alpha\beta^{-1}\alpha^{-1}$ and its effect on a point X of ℓ, and also its effect on a line n through B. You need also to prove that $\beta\alpha\beta^{-1}\alpha^{-1}$ is not the identity i.)

Note 36 *The following theorem also holds: If*

(a) *ℓ is a given line of a projective plane π of order q;*
(b) *$|\Gamma_{(\ell,\ell)}| > q$;*

then

(a) *$|\Gamma_{(A,\ell)}| > 1$ for all points $A \in \ell$;*
(b) *q is a prime power.*

Group of elations

In what follows, an elation which is distinct from the identity i will be referred to as *non-trivial*.

Theorem 6.13 *Let A, B be two distinct points of a line ℓ of a projective plane π. Then*

1. *With respect to composition of functions, the set of all (A, ℓ)-elations forms a group G_A.*
2. *If σ is a non-trivial (A, ℓ)-elation and τ is a non-trivial (B, ℓ)-elation then $\sigma\tau$ is a (C, ℓ)-elation with $C \neq A, B$.*
3. *With respect to composition of functions, the set of all elations with ℓ as axis is an Abelian group G called the* **translation group with axis ℓ**.
4. *Every non-trivial element of the above group G is either of infinite order, or of the same prime order.*

Proof.

1. The proof of 1 is left as an easy exercise for the reader.

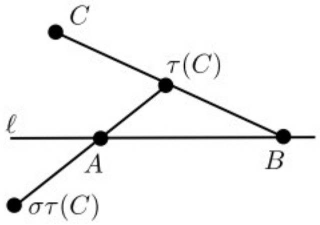

2. Let σ be a non-trivial (A, ℓ)-elation and τ a non-trivial (B, ℓ)-elation. Now $\sigma\tau \neq i$ since otherwise $\sigma = \tau^{-1}$ would imply that $A = B$. Also $\sigma\tau$ fixes ℓ pointwise and hence by Theorem 6.5 fixes some point C linewise. Suppose $C \notin \ell$. Then $\tau(C)$ is a point of BC distinct from B and C. Thus $\sigma(\tau(C))$ is a point on the line AC^{τ}. Therefore $\sigma\tau(C) \neq C$. This contradiction shows that $C \in \ell$. Let P be any point of $\pi\backslash\{\ell\}$. Then $\{P, \tau(P), B\}, \{\tau(P), \sigma(\tau(P)), A\}$, and $\{P, \sigma\tau(P), C\}$ are triples of collinear points. Hence $C \neq A, B$.

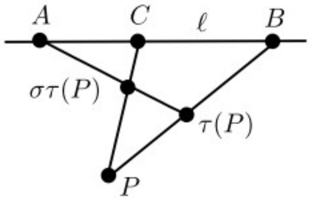

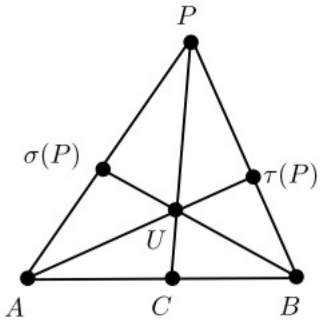

3. We first prove that $\sigma\tau = \tau\sigma$, where σ is a non-trivial (A, ℓ)-elation and τ is a non-trivial (B, ℓ)-elation. Let P be any point of π. If $P \in \ell$, then $\sigma\tau(P) = \sigma(P) = P$ and $\tau\sigma(P) = \tau(P) = P$. Suppose $P \notin \ell$, then $\{P, \sigma(P), A\}$ and $\{P, \tau(P), B\}$ are two triads of collinear points. Let $U = BP^{\sigma} \cap AP^{\tau}$. Since $\tau(P) = BP \cap AU$, we have

$$\begin{aligned} \sigma\tau(P) &= (BP \cap AU)^{\sigma} \\ &= (BP)^{\sigma} \cap (AU)^{\sigma} \\ &= BP^{\sigma} \cap AU^{\sigma} \\ &= U, \quad \text{since } AU^{\sigma} = AU. \end{aligned}$$

Similarly, $\tau\sigma(P) = U$. Therefore, $\tau\sigma(P) = \sigma\tau(P)$ for all points P in π. Therefore, $\sigma\tau = \tau\sigma$.

We next show that if α_1, α_2 are both non-trivial elations with the same vertex D and axis ℓ, then $\alpha_1\alpha_2 = \alpha_2\alpha_1$. Without loss of generality, suppose $D \neq B$. Then for $j = 1, 2$, $\tau\alpha_j \neq i$ and $\tau\alpha_j$ is a non-trivial (E_j, ℓ)-elation with $E_j \neq D, B$. But then, since non-trivial elations with distinct vertices commute, we have:

$$\begin{aligned} \alpha_1(\tau\alpha_2) &= (\tau\alpha_2)\alpha_1, \\ (\alpha_1\tau)\alpha_2 &= \tau\alpha_2\alpha_1, \\ (\tau\alpha_1)\alpha_2 &= \tau(\alpha_2\alpha_1). \end{aligned}$$

Therefore,

$$\tau(\alpha_1\alpha_2) = \tau(\alpha_2\alpha_1),$$

and so

$$\alpha_1\alpha_2 = \alpha_2\alpha_1, \quad \text{since } \tau \text{ has an inverse.}$$

It now follows that, with respect to composition of functions, the set of all elations with ℓ as axis is an Abelian group G.

4. Suppose there exists an elation of finite order in G. Then by Sylow's Theorem, G contains an element of order p for some prime p. Let β be an (E,ℓ)-elation of order p. Let δ be any non-trivial (X,ℓ)-elation, $E \neq X$. Then $\beta\delta$ is an (F,ℓ)-elation with $F \neq E, X$.

Hence $(\beta\delta)^p$ is an (F,ℓ)-elation. But since G is abelian,

$$\begin{aligned}(\beta\delta)^p &= \beta^p\delta^p \\ &= \delta^p, \quad \text{as } \beta^p = i.\end{aligned}$$

But δ^p is an (X,ℓ)-elation. This is a contradiction unless $\delta^p = i$. □

Exercise 6.5

1. In a projective plane π, let A, B be points on a line ℓ and let m be another line through B. Let α be a non-trivial elation with axis ℓ and vertex A, and let β be a non-trivial elation with axis m and vertex B.
   (a) Prove that $\beta\alpha\beta^{-1}\alpha^{-1}$ is an elation with axis ℓ and vertex B.
   (b) Prove that $\beta\alpha\beta^{-1}\alpha^{-1}$ is not the identity.
   (*Hint:* Consider the effect of $\beta\alpha\beta^{-1}\alpha^{-1}$ on a point $X \in m$.)

(V,ℓ)-transitive and (V,ℓ)-Desarguesian planes

Definition 6.14 1. *Let V be a point and ℓ a line of a projective plane π. Then π is* **(V,ℓ)-transitive** *if given any pair of points A, B of π such that*
   (a) *A, B, V are collinear and distinct;*
   (b) *$A, B \notin \ell$,*

   then there exists a (V,ℓ)-perspectivity σ with $A^\sigma = B$.

2. *If π is (X,ℓ)-transitive for all points X on a line m, then π is said to be* **(m,ℓ)-transitive**.
3. *If π is* **(ℓ,ℓ)-transitive**, *then π is said to be a* **translation plane** *with respect to ℓ; often we write π^ℓ is a translation plane. The line ℓ is called a* **translation line**.

Example 6.15 A field plane $\mathrm{PG}(2,F)$ is a translation plane with respect to any of its lines.

Definition 6.16 *A* **dual translation plane** *π is a plane π^L with a special point L, such that π^L is (L, m)-transitive for all lines m through the point L (called a* **translation point***). Thus* a dual translation plane may not be a translation plane.

Theorem 6.17 *Let A, B be distinct points of a line ℓ of a projective plane π. If π is (A, ℓ)-transitive and (B, ℓ)-transitive, then π^ℓ is a translation plane.*

Proof. Let C be any point of $\ell = AB$. Let X, Y be two points not on ℓ but collinear with C. We have to show that there exists a (C, ℓ)-elation σ such that $X^\sigma = Y$.
Let $Z = AX \cap BY$. Then $Z \neq X, Y$. Since π is both (A, ℓ)-transitive and (B, ℓ)-transitive, there exists an (A, ℓ)-elation α and a (B, ℓ)-elation β with $X^\alpha = Z$ and $Z^\beta = Y$. Put $\sigma = \beta\alpha$. By Theorem 6.13, σ is a (V, ℓ)-elation. Now

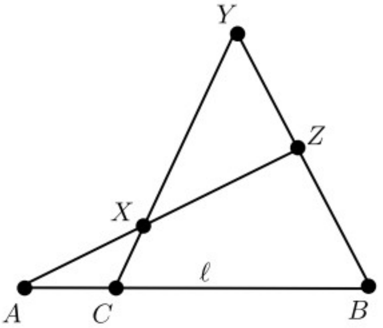

$$X^\sigma = (X^\alpha)^\beta = Z^\beta = Y \quad \text{and} \quad V = XY \cap \ell = C. \qquad \square$$

Example 6.18 Let ℓ be a translation line of a projective plane π and let α be a collineation of π. Prove that ℓ^α is also a translation line of π.

Solution. Let X, Y be any two distinct points of π not on ℓ^α. Let $XY \cap \ell^\alpha = C$. We have to show that there exists a (C, ℓ^α)-elation σ with $X^\sigma = Y$.

Since X, Y, C are collinear, with $C \in \ell^\alpha$, it follows that $X^{\alpha^{-1}}, Y^{\alpha^{-1}}, C^{\alpha^{-1}}$ are collinear, with $C^{\alpha^{-1}} \in \ell$. Since ℓ is a translation line of π, there exists a $(C^{\alpha^{-1}}, \ell)$-elation β such that

$$\beta\alpha^{-1}(X) = \alpha^{-1}(Y).$$

Therefore,

$$\alpha\beta\alpha^{-1}(X) = Y.$$

By Example 6.9, $\sigma = \alpha\beta\alpha^{-1}$ is a $((C^{\alpha^{-1}})^\alpha, \ell^\alpha)$-elation. Thus ℓ^α is a translation line of π. $\square$

Exercise 6.6 Let π be a projective plane.

1. Prove that if ℓ, m are two translation lines of π, then so is every line through $\ell \cap m$.
2. Prove that if π has three non-concurrent translation lines, then every line of π is a translation line (in which case the plane π is called a **Moufang plane**).

Definition 6.19 *Let V be a point, and ℓ be a line of a projective plane π. Let T, T' be two triangles in perspective from V, with the property that two pairs of corresponding sides of T, T' intersect on ℓ. Then π is said to be* **(V, ℓ)-Desarguesian** *if the third pair of corresponding sides necessarily intersect on ℓ.*

Example 6.20 A field plane is (V, ℓ)-Desarguesian for all points V and all lines ℓ.

Theorem 6.21 *(***Baer***) Let V be a point and ℓ be a line of a projective plane π. Then π is (V, ℓ)-transitive if and only if π is (V, ℓ)-Desarguesian.*

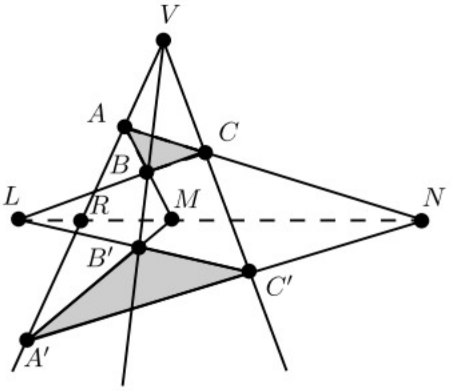

Proof. ($\Longrightarrow$) Suppose π is (V, ℓ)-transitive. Let ABC, $A'B'C'$ be two triangles in perspective from V with $M = AB \cap A'B'$ and $N = AC \cap A'C'$ both on ℓ. Since π is (V, ℓ)-transitive, there exists a (V, ℓ)-perspectivity σ, with $A^\sigma = A'$. Since V is the vertex of σ, B^σ lies on VB. Since ℓ is the axis of σ, $(ABM)^\sigma = A'B^\sigma M$. Therefore, $B^\sigma = VB \cap A'M$. Therefore, $B^\sigma = B'$. Similarly, $C^\sigma = C'$. Let $BC \cap \ell = L$. Since L, B, and C are collinear, their images (under σ) $L^\sigma = L$, $B^\sigma = B'$ and $C^\sigma = C'$ are collinear. Thus $L = BC \cap B'C'$.

($\Longleftarrow$) Suppose π is (V, ℓ)-Desarguesian. Let A, A' be two points collinear with V, but not on ℓ. We need to prove that there exists a (V, ℓ)-perspectivity σ, with $A^\sigma = A'$. Let B be any point not on the lines VAA' and ℓ. Let $M = AB \cap \ell$. Define the permutation α of

$$\pi \backslash AA' \cup \{V, AA' \cap \ell\}$$

by

$$\begin{aligned} \alpha \colon\ & P \mapsto P, \quad \text{if } P = V \text{ or } P \in \ell, \\ & B \mapsto B' = A'M \cap VB, \quad \text{if } B \neq V \text{ or } B \notin \ell. \end{aligned}$$

Having obtained a pair $\{B, B' = B^\alpha\}$, use them to define the permutation β of

$$\pi \backslash BB' \cup \{V, BB' \cap \ell\}$$

by

$$\begin{aligned} \beta\colon\ & P \mapsto P, \quad \text{if } P = V \text{ or } P \in \ell \\ & P \mapsto P' = RB' \cap VP, \quad \text{if } P \neq V \text{ or } P \notin \ell, \text{ where } R = PB \cap \ell. \end{aligned}$$

Thus, α and β agree on V and on points of ℓ.

We next prove that α and β agree on all points not on AA' and BB'. Let C be a point not on AA' and BB'. If $C \in AB$, then $C^\alpha = VC \cap A'B' = C^\beta$. If $C \notin AB$, let $AC \cap \ell = N$ and $BC \cap B'C' = L$. Since π is (V, ℓ)-Desarguesian, we have: $L \in \ell$. Hence, $C^\alpha = C^\beta$.

Define σ to be α on $\pi \backslash \{AA'\}$ and β on $\pi \backslash \{BB'\}$. Thus σ is a well-defined permutation of π.

Lastly, we prove that σ is a (V, ℓ)- perspectivity. By definition, σ leaves ℓ fixed pointwise, and V fixed linewise. Let r be any line not through V, and distinct from ℓ. Without loss of generality, let $r \cap \ell = L$, and B, C be any two points of r distinct from L. Then, $L = L^\sigma, B' = B^\sigma, C' = C^\sigma$ are collinear. Thus σ preserves collinearity. Hence, σ is a (V, ℓ)-perspectivity. □

Note 37 1. *A projective plane π is Desarguesian if and only if it is (V, ℓ)-transitive for all choices of a point V and a line ℓ of π.*

2. *(See Exercise 6.6(2)). A Moufang plane is (V, ℓ)-Desarguesian, for all choices of a point V and a line ℓ, with $V \in \ell$.*

3. *Recall Theorem 5.11. If in a projective plane π, the Little Desargues Axiom holds for triangles in perspective from any point P of a line XY, with XY as axis, then π can be coordinatised by a quasifield γ.*

   Thus π can be coordinatised by a quasifield if and only if the line XY is a translation line of π.

4. *Recall Definition 5.12(2) of a nearfield. It can be proved that a projective plane π can be coordinatised by a nearfield if and only if, in the notation of Chapter 5 and Definition 6.14, (see Theorem 5.11),*
   (a) *π is (XY, XY)-transitive;*
   (b) *π is (V, m)-transitive, with $V = X$, and m is a line through Y* or *$V = Y$, and m contains the point X.*

5. *Let π be a translation plane with $XY = \ell_\infty$ as translation line; let π be coordinatised by a quasifield γ. If two other points of ℓ_∞ are chosen to play the role of X and Y, the coordinatising ternary ring is a quasifield γ', but* γ and γ' need not be isomorphic.

6. A finite Moufang plane is Desarguesian. *This is a consequence of a theorem by Artin and Zorn:* a finite alternative division ring is a field.

Exercise 6.7

1. Let π be a projective plane which is coordinatised by the set γ, as explained in Chapter 5. Prove that π is a translation plane with respect to XY and a dual translation plane with respect to Y if and only if γ is a semifield.
2. Let π be a projective plane which is coordinatised by the set γ, as explained in Chapter 5. Prove that γ is a nearfield if and only if π is (XY, XY)-, (X, Y)-, and (Y, X)-transitive.
3. Use the above two questions to prove: Let π be a projective plane which is coordinatised by the set γ, as explained in Chapter 5. Then π is Desarguesian if and only if γ is a skew field.
4. The points and lines in the following diagram are those of a projective plane π. Assume that π is (V, ℓ)-Desarguesian. Prove that L, B, and C are collinear.

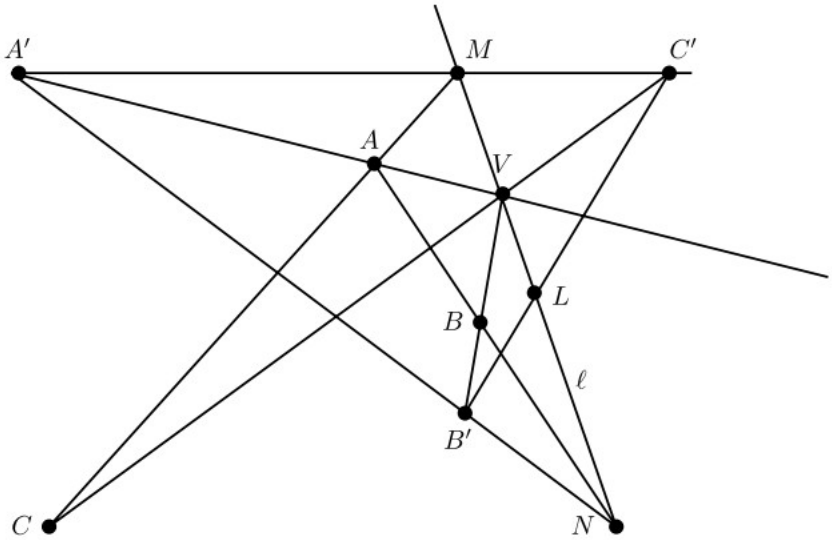

The Hughes plane of order q^2, q an odd prime power

This projective plane was first discovered by Veblen and Wedderburn, in the case $q = 3$. For q any odd prime power, Hughes (1959) defined a nearfield N of order q^2 which has special properties. This nearfield N is then used to construct a projective plane of order q^2. Such a projective plane is called a Hughes plane of order q^2 and is denoted by $\mathcal{H}(q^2)$.

$\mathcal{H}(q^2)$ is given here as an example of a projective plane, which is neither a translation plane nor a dual translation plane. Its construction depends on the following two facts:

1. There exists a 3×3 matrix A over GF(q) such that $A^{q^2+q+1} = kI, k \in$ GF(q), but no smaller power of A has this property. (See Singer's Theorem, Theorem 4.41, where such a matrix is shown to exist.)
2. For all odd prime powers q, there exists a nearfield N of order q^2, which is not a field.

In order to show the existence of such a nearfield N, we need the following lemma.

Lemma 6.22 *Let* $F = GF(q^2), q$ *odd. Then:*

$$f = x^q - \alpha x - \beta \quad \in F[x], \quad \alpha \textit{ a non-square},$$

has precisely one zero in F*, namely* $(-\beta^q - \alpha^q\beta)(\alpha^{q+1} - 1)^{-1}$.

Proof. We first prove that $\alpha^{q+1} \neq 1$. Let ω be a generator of GF(q^2). Then the $(q+1)$th roots of 1 are $1, \omega^{q-1}, \omega^{2(q-1)}, \ldots, \omega^{q(q-1)}$. Since $q-1$ is even, all of these $(q+1)$th roots of 1 are squares, and therefore none of them can be equal to α, as α is a non-square. Therefore, $\alpha^{q+1} \neq 1$. Let $X \in$ GF(q^2) be a zero of f. Then

$$\begin{aligned}
\alpha X + \beta = X^q &\implies \alpha^q X^q + \beta^q = X^{q^2} = X,\\
&\implies \alpha^q(\alpha X + \beta) + \beta^q = X,\\
&\implies (\alpha^{q+1} - 1)X = -\beta^q - \alpha^q\beta,\\
&\implies X = (-\beta^q - \alpha^q\beta)(\alpha^{q+1} - 1)^{-1}, \quad \text{since } \alpha^{q+1} \neq 1. \qquad \square
\end{aligned}$$

Example 6.23 Let $(F, +, .)$ be GF(q^2), q an odd prime power. Define $(N, +, \circ)$ as follows:

1. F, N have the same elements, and same addition.
2. Multiplication $\circ$ in N is given by

$$\begin{aligned}
x \circ y &= x.y, && \text{if } y \text{ is a square in } F,\\
&= \overline{x}.y, && \text{if } y \text{ is not a square in } F, \text{ where } \overline{x} = x^q.
\end{aligned}$$

Prove the following:

(a) N is a nearfield of order q^2.

(b) Every element of N commutes with every element of GF(q) $\subset N$.

(c) N is not a field.

Solution.

(a) (i) $(N,+)=(F,+)$ is an abelian group with 0 as identity.

(ii) $(N\backslash\{0\},\circ)$ is a loop with 1 as identity. We leave this as an exercise for the reader.

(iii) Let $a,b,m\in N$.
If m is a square, then

$$(a+b)\circ m=(a+b).m=a.m+b.m=a\circ m+b\circ m.$$

If m is not a square, then

$$(a+b)\circ m=\overline{(a+b)}.m=(\overline{a}+\overline{b}).m=\overline{a}.m+\overline{b}.m=a\circ m+b\circ m.$$

Thus,

$$(a+b)\circ m=a\circ m+b\circ m,\qquad \text{for all } a,b,m\in N.$$

(iv) Suppose $a\neq b$. We need to show that the equation

$$x\circ a=x\circ b+c$$

has unique solution x, for all $a,b,c\in N$.
If both a and b are squares, the unique solution is $c(a-b)^{-1}$.
If both a and b are non-squares, then

$$\begin{aligned} x\circ a=x\circ b+c &\implies \overline{x}.a=\overline{x}.b+c\\ &\implies \overline{x}.(a-b)=c\\ &\implies x=\overline{c}.(\overline{a}-\overline{b})^{-1}.\end{aligned}$$

If one of a or b is a square, and the other a non-square, we need to solve either $ax^q=bx+c$ or $ax=bx^q+c$. In both cases, we need to solve an equation of type

$$f=x^q-\alpha x-\beta=0,$$

where α, being either $a^{-1}b$ or $b^{-1}a$, is a non-square. By Lemma 6.22, there exists a unique solution.

(v) N has associative multiplication. We leave this as an exercise for the reader.

(b) We next prove that $a\circ x=x\circ a$, for all $x\in \mathrm{GF}(q)$, for all $a\in N$. For all $x\in\mathrm{GF}(q)\subset\mathrm{GF}(q^2)$, we have $\overline{x}=x$, and x is a square. Therefore, for all $a\in N, x\in\mathrm{GF}(q)\subset N$,

$$\begin{aligned} a\circ x&=a.x\\ x\circ a&=x.a, \qquad \text{whether } a \text{ is a square or not, as } \overline{x}=x.\\ &=a.x\\ &=a\circ x.\end{aligned}$$

(c) Let ω be a generator of $\mathrm{GF}(q^2)$. Then ω is a non-square and ω^2 is a square. Then, $\omega \circ \omega^2 = \omega^3$ and $\omega^2 \circ \omega = \omega^{2q+1} \neq \omega^3 = \omega \circ \omega^2$. Thus N is not a field. □

From now till the end of this section, we write xy for $x \circ y$.

Exercise 6.8 Let N be the nearfield of order q^2 as defined in Example 6.23. Prove that if $a, b, a', b' \in K = GF(q)$, and $n \in N \backslash K$, then

1. $a + bn = a' + b'n \iff a = a', b = b'$.
2. $a + nb = 0 \iff a = b = 0$.

Construction of the Hughes plane. Let N be the nearfield of order q^2, as constructed in Example 6.23. Let A be a 3×3 matrix over $\mathrm{GF}(q)$ such that $A^{q^2+q+1} = kI$, $k \in \mathrm{GF}(q)$, but no smaller power of A has this property. Let $V = \{(x, y, z) \mid x, y, z \in N\}$. On V, define addition and right scalar multiplication by: for all (x, y, z), $(a, b, c) \in V$, and for all $k \in N$

$$(x, y, z) + (a, b, c) = (x + a, y + b, z + c),$$
$$(x, y, z)k = (xk, yk, zk).$$

Definition 6.24 *Let N be the nearfield of order q^2, as constructed in Example 6.23. The* **Hughes plane** *$\mathcal{H}(q^2)$ is the triple $(\mathcal{P}, \mathcal{L}, \mathcal{I})$, (where the elements of $\mathcal{P}$ are called* points, *the elements of $\mathcal{L}$, are subsets of $\mathcal{P}$ to be called* lines, *and $\mathcal{I}$ is an incidence relation), defined as follows:*

1. *The set of* **points** *of $\mathcal{H}(q^2)$ is*

$$\mathcal{P} = \{(x, y, z) \mid x, y, z \in N, \text{ not all zero}\},$$

with the proviso that

$$\text{for all } k \in N\backslash\{0\}, (x, y, z)k \text{ refers to the same point.}$$

2. *Let $\alpha = 1$, or $\alpha \in N \backslash GF(q)$. For each α, let $L(\alpha)$ (to be called a* **line** *of $\mathcal{L}$) be the set of points (x, y, z) satisfying $x + \alpha y + z = 0$. Then the set of* **lines** *of $\mathcal{H}(q^2)$ is*

$$\mathcal{L} = \{A^n L(\alpha) \mid 0 \leq n \leq q^2 + q\},$$

where for each n, $A^n L(\alpha)$ is the set of points $v(A^n)^t$ with $v \in L(\alpha)$.

3. *$\mathcal{I}$: the point $v = (x, y, z)$ is incident with the line $A^n L(\alpha)$ if and only if $v = w(A^n)^t$, for some $w \in L(\alpha)$.*

Example 6.25 Find the number of points and lines in $\mathcal{H}(q^2)$. Show that each line of $\mathcal{H}(q^2)$ has $q^2 + 1$ points.

Solution.

1. $|\mathcal{P}| = \dfrac{|V|-1}{q^2-1} = \dfrac{q^6-1}{q^2-1} = q^4+q^2+1.$

2. As $0 \le n \le q^2+q$, $\alpha = 1$, or $\alpha \in N \backslash F$, there are q^2+q+1 choices for n, and q^2-q+1 choices for α. Hence
$$|\mathcal{L}| = (q^2+q+1)(q^2-q+1) = q^4+q^2+1.$$

3. For given α, if any two of x, y, z are given, the third can be uniquely determined by the equation $x+\alpha y+z=0$. Therefore, the number of non-zero solutions of this equation is $(q^2)^2-1$. Since $(x,y,z) \equiv (x,y,z)k, k \in N\backslash\{0\}$, the number of points on line $L(\alpha)$ is $(q^4-1)/(q^2-1) = q^2+1$. Thus, each line of $\mathcal{H}(q^2)$ has q^2+1 points. □

Theorem 6.26 *$\mathcal{H}(q^2)$ is a projective plane of order q^2.*

Proof. In view of Example 6.25, we need only show that every pair of lines of $\mathcal{H}(q^2)$ intersect in a point. This is left as an exercise for the reader. □

Example 6.27 Show that a line of $\mathcal{H}(q^2)$ is represented by an equation of the form
$$ax+by+cz+\alpha(a'x+b'y+c'z)=0, \quad \alpha = 1, \text{ or } \alpha \in N\backslash \mathrm{GF}(q),$$
where a, a', b, b', c, c' are elements of $\mathrm{GF}(q)$, uniquely determined by A^{-n}.

Solution. Given n and α, let $A^{-n} = [a_{ij}]$. Since the matrix A is over $\mathrm{GF}(q)$, the a_{ij}'s are necessarily elements of $\mathrm{GF}(q)$. By definition, the point
$$\begin{aligned} v=(x,y,z) \in A^nL(\alpha) &\implies v = w(A^n)^t, \quad \text{for some } w \in L(\alpha),\\ &\implies w = v(A^{-n})^t \in L(\alpha)\\ &\qquad = \begin{bmatrix} x & y & z\end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13}\\ a_{21} & a_{22} & a_{23}\\ a_{31} & a_{32} & a_{33}\end{bmatrix}^t \in L(\alpha).\end{aligned}$$

Therefore, $(a_{11}x+a_{12}y+a_{13}z)+\alpha(a_{21}x+a_{22}y+a_{23}z)+(a_{31}x+a_{32}y+a_{33}z)=0$. Thus the line $A^nL(\alpha)$ is represented by an equation of the form
$$ax+by+cz+\alpha(a'x+b'y+c'z)=0, \quad \alpha = 1, \text{ or } \alpha \in N\backslash \mathrm{GF}(q),$$

where $a = a_{11} + a_{31}, a' = a_{21}$, etc, are all elements of GF(q), uniquely determined by A^{-n}. □

Note 38 *In particular, corresponding to* $\alpha = 1$, *we have* $q^2 + q + 1$ *lines with equations of type*

$$ax + by + cz = 0, \quad a, b, c \in GF(q).$$

Thus PG$(2, q)$ is a Desarguesian subplane (a Baer subplane) of $\mathcal{H}(q^2)$.

Example 6.28 In $\mathcal{H}(q^2)$, let $O = (0, 0, 1)$, $X = (1, 0, 0)$, $Y = (0, 1, 0)$, $C = (0, b, 1)$, $A = (a, a, 1)$, $B = (b, b, 1)$, $U = (1, 1, 1)$, $I = (1, 1, 0)$, $M = (1, m, 0)$. Find the coordinates of

1. $YA \cap XB$;
2. $CI \cap YA$;
3. $YA \cap OM$.

Solution.

1. The line $A^n L(\alpha)$ has an equation of type

$$ax + by + cz + \alpha(a'x + b'y + c'z) = 0, \quad \alpha = 1, \text{ or } \alpha \in N\backslash\text{GF}(q),$$

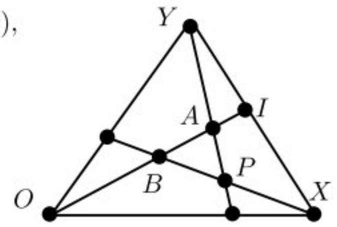

where the coefficients are all elements of GF(q), uniquely determined by A^{-n}. In particular, the line OU has equation $x - y = 0$. Thus,

$$OU = \{(a, a, 1) \mid a \in N\} \cup \{(1, 1, 0)\}.$$

Now, YA has equation

$$a_1 x + (-a_1 a)z + \alpha(a'_1 x + (-a'_1 a)z) = 0,$$

for some $a_1, a'_1 \in$ GF(q), and $\alpha = 1$, or $\alpha \in N\backslash$GF(q). The equation of XB is

$$b_1 y + (-b_1 b)z + \beta(b'_1 y + (-b'_1 b)z) = 0,$$

for some $b_1, b'_1 \in$ GF(q), and $\beta = 1$, or $\beta \in N\backslash$GF(q). Therefore, $YA \cap XB = (a, b, 1)$.

2. Let the equation of CI be

$$a_1 x + b_1 y + c_1 z + \alpha(a'_1 x + b'_1 y + c'_1 z) = 0, \quad \alpha = 1, \text{ or } \alpha \in N\backslash\text{GF}(q).$$

Since $(1,1,0) \in CI$, we have

$$a_1 + b_1 + \alpha(a_1' + b_1') = 0.$$

If $\alpha \in N \backslash \mathrm{GF}(q)$, then

$$a_1 + b_1 = 0 = a_1' + b_1',$$

since $a_1, a_1', b_1, b_1' \in \mathrm{GF}(q)$.
Thus, for $\alpha = 1$, or $\alpha \in N \backslash \mathrm{GF}(q)$, CI has equation of type

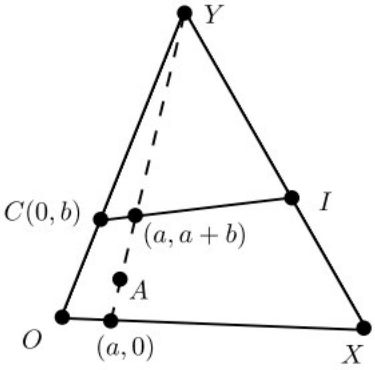

$$a_1 x - a_1 y + c_1 z + \alpha(a_1' x - a_1' y + c_1' z) = 0,$$

where

$$-a_1 b_1 + c_1 + \alpha(-a_1' b_1 + c_1') = 0. \tag{1}$$

Now the point $(a, a+b, 1)$ lies on YA and on CI, since

$$\begin{aligned} &a_1 a - a_1(a+b) + c_1 + \alpha(a_1' a - a_1'(a+b) + c_1') \\ &= a_1 a - a_1 a - a_1 b + c_1 + \alpha(a_1' a - a_1' a - a_1' b + c_1'), \quad \text{since } a_1, a_1', \in \mathrm{GF}(q) \\ &= -a_1 b + c_1 + \alpha(-a_1' b + c_1') \\ &= 0, \quad \text{by Equation (1).} \end{aligned}$$

Thus $CI \cap YA = (a, a+b, 1)$.

3. Similarly, the equation of OM is

$$a_1 x + b_1 y + \alpha(a_1' x + b_1' y) = 0,$$

where $a, a', a_1', b_1 \in \mathrm{GF}(q), \alpha = 1$, or $\alpha \in N \backslash \mathrm{GF}(q)$, and

$$a_1 + b_1 m + \alpha(a_1' + b_1' m) = 0. \tag{2}$$

The point $(a, ma, 1)$ lies on YA and on OM, since

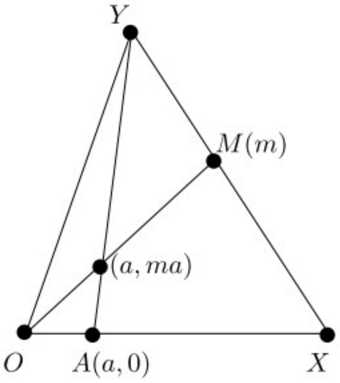

$$\begin{aligned} a_1 a + b_1 ma + \alpha(a_1' a + b_1' ma) &= \{a_1 + b_1 m + \alpha(a_1' + b_1' m)\} a \\ &= 0a, \quad \text{by Equation (2)} \\ &= 0. \end{aligned}$$

□

Note 39 *Let the point* $v = (x, y, z)$ *of* $\mathcal{H}(q^2)$ *be incident with the line* $A^n L(\alpha)$. *Then* $v = w(A^n)^t$ *for some* $w \in L(\alpha)$. *Therefore,* $vA^t = w(A^n)^t A^t = w(A^{n+1})^t$ *for some* $w \in L(\alpha)$, *and so the point* vA^t *is incident with the line* $A^{n+1}L(\alpha)$. *Thus the map* $\sigma\colon v \mapsto vA^t$ *is a collineation of* $\mathcal{H}(q^2)$, *mapping the line* $A^n L(\alpha)$ *to the line* $A^{n+1}L(\alpha)$, *which is distinct from* $A^n L(\alpha)$. *Suppose that* $\mathcal{H}(q^2)$ *is a translation plane with respect to the line* $A^n L(\alpha)$, *for some* n *and* α. *Then, by Example 6.18,* $\sigma(A^n L(\alpha))$ *is also a translation line of* $\mathcal{H}(q^2)$. *In fact the* q^2+q+1 *distinct lines* $\sigma^i(A^n L(\alpha)), i = 1, 2, \dots, q^2+q+1$, *are all translation lines. These* q^2+q+1 *distinct lines cannot all pass through the same point, since there are exactly* q^2+1 *lines through each point of* $\mathcal{H}(q^2)$. *Therefore, at least three of these* q^2+q+1 *translation lines are non-concurrent. By Exercise 6.6 it follows that every line of* $\mathcal{H}(q^2)$ *is a translation line, making the plane a Moufang plane. It has been remarked before that a finite Moufang plane is a field plane.*

Thus, if $\mathcal{H}(q^2)$ *were a translation plane, it would be isomorphic to* $PG(2, q^2)$.

Theorem 6.29 $\mathcal{H}(q^2)$ *is neither a translation plane, nor a dual translation plane.*

Proof. Example 6.28 part (1) shows that N is a coordinatising set for $\mathcal{H}(q^2)$, as defined in Chapter 5. Let the addition defined (as in Chapter 5) on N be $\oplus$, and the multiplication be $\odot$.

Suppose that $\mathcal{H}(q^2)$ is a translation plane. Then, as observed before, the coordinatising ring $(N, \oplus, \odot)$ is a field of order q^2. By Example 6.28, parts (2) and (3),

$$\begin{aligned}
a + b &= a \oplus b, && \text{for all } a,\ b\ \in N.\\
&= b \oplus a, && \text{since } (N, \oplus, \odot) \text{ is a field.}\\
&= b + a\\
m \circ a &= a \odot m, && \text{for all } a,\ m\ \in N\\
&= m \odot a, && \text{since } (N, \oplus, \odot) \text{ is a field.}\\
&= a \circ m\\
(a + b) \circ m &= (a \oplus b) \odot m\\
&= (a \odot m) \oplus (b \odot m), && \text{since } (N, \oplus, \odot) \text{ is a field.}\\
&= a \circ m + b \circ m.
\end{aligned}$$

The remaining axioms for a field are easily seen to hold in $(N, +, \circ)$. This is a contradiction, since $(N, +, \circ)$, by construction, is not a field. Hence, $\mathcal{H}(q^2)$ is not a translation plane.

It is similarly proved that $\mathcal{H}(q^2)$ is not a dual translation plane. □

The Hall planes

Given a quasifield γ, the plane it coordinatises is a translation plane. The *Hall* quasifields (which we will soon define) are examples of quasifields which are not skew fields; the planes they coordinatise are non-Desarguesian translation planes.

Definition 6.30 *Let γ be a quasifield. The* **kernel** *$K(\gamma)$ of γ is the set of all elements k of γ such that, for all $a, b \in \gamma$,*

$$k(ab) = (ka)b \qquad \text{and} \qquad k(a+b) = ka + kb.$$

Theorem 6.31 *Let γ be a quasifield. The kernel $K(\gamma)$ of γ is a skew field, and γ is a (left) vector space over $K(\gamma)$.*

Note 40 *By left vector space V over a skew field K we mean*

- *If $v \in V, k \in K$, then $kv \in V$.*
- *All the vector space axioms hold, but with scalar multiplication always on the left.*

A right vector is similarly defined.

Proof. For any $a, b \in \gamma$, and any $k, l \in K(\gamma)$, we have

$$\begin{aligned}(kl)(a+b) &= k\{l(a+b)\}\\ &= k\{la+lb\}\\ &= k(la)+k(lb)\\ &= (kl)a+(kl)b.\end{aligned}$$

Thus, $K(\gamma)$ is closed under multiplication.

Similarly, the other axioms of a skew field can be proved to hold in $K(\gamma)$.

Using the addition of γ as vector addition, it follows that γ is a (left) vector space over its kernel. □

This leads to the following important theorem:

Theorem 6.32 *Let $(\gamma, +, .)$ be a quasifield of finite order. Then,*

1. *$q = p^r$ for some prime p, some integer r,*
2. *$(\gamma, +)$ is isomorphic to $(GF(q), +)$.*

Proof.

1. Since γ is finite, its kernel $K(\gamma)$ is a finite skew field. By Wedderburn's Theorem, $K(\gamma)$ is a finite field, and therefore its order is p^h, for some prime p, and positive integer h. Let the dimension of γ, as a vector space over its kernel, be n. Then γ has order $q = p^{hn}$.
2. Both $\mathrm{GF}(q)$ and γ are vector spaces of dimension n over $\mathrm{GF}(p^h)$. Therefore, $(\gamma, +)$ is isomorphic to $(\mathrm{GF}(q), +)$. □

Exercise 6.9

1. Let π be a translation plane of prime order p. Prove that π is isomorphic to $\mathrm{PG}(2,p)$.
2. Give a complete proof of Theorem 6.31.

In view of Theorems 6.31 and 6.32, it is not surprising that every construction of a quasifield γ starts with a Galois field F, and a subfield K of F, retains the addition of F, retains the multiplication of K, and alters the multiplication in $F\backslash K$.

Definition 6.33 1. *A* **finite Hall quasifield** *is a quasifield $\mathcal{H}(q^2,f)$ constructed as follows.*

Let $K = GF(q)$, and let $f(x) = x^2 - rx - s$ be an irreducible quadratic over K. Let λ be a zero of f in $F = GF(q^2)$. Let $\mathcal{H}(q^2,f)=\{a+b\lambda \mid a,b \in K\}$. The addition $\oplus$ in $\mathcal{H}(q^2,f)$ is defined by

$$(a+b\lambda)\oplus(c+d\lambda) = (a+c)+(b+d)\lambda.$$

The multiplication $\odot$ in $\mathcal{H}(q^2,f)$ is defined by

$$\begin{aligned}(a+b\lambda)\odot c &= ac+(bc)\lambda.\\ (a+b\lambda)\odot(c+d\lambda) &= (ac-bd^{-1}f(c))+(ad-bc+br)\lambda, \qquad \textit{for } d\neq 0.\end{aligned}$$

2. *A* **finite Hall plane** *is a plane coordinatised by a Hall quasifield.*

The proof that $\mathcal{H}(q^2,f)$ is indeed a quasifield can be found in M. Hall's book 'The theory of groups'.

Exercise 6.10 Concerning the Hall quasifield $\mathcal{H}(q^2,f)$, as defined above,

1. Verify that every element α of $\mathcal{H}(q^2,f)\ \backslash K$ satisfies $f(\alpha)=0$.
2. Verify that every element of K commutes with every element of $\mathcal{H}(q^2,f)$.
3. Show that $\mathcal{H}(9,f)$, where $f = x^2+1$, is associative, making it a nearfield.
4. Show that $\mathcal{H}(9,f)$, where $f = x^2-x-1$, is not associative.
5. Show that $\mathcal{H}(q^2,f)$ is distributive if and only if $q=2$.
6. Write out the multiplication table for the Hall quasifield $\mathcal{H}(16,f)$, for an appropriate polynomial f (irreducible over $\mathrm{GF}(4)$).

It is interesting to note that the properties, enunciated in questions 1 and 2 above, characterise Hall quasifields. In *Projective Planes* (Trans Amer Math Soc 54 (1943), 229–277), M. Hall used the following definition, which is stated here as a theorem.

Theorem 6.34 *Let* $(H, +, .)$ *be a quasifield, with the field* $K = GF(q)$ *contained in both its centre and its kernel such that* H *is two-dimensional over* K. *If there exists an irreducible quadratic* f *over* K *with* $f(a) = 0$ *for all* $a \in H \backslash K$, *then* H *is a Hall quasifield.*

Remark. Recall the *Bruck–Bose construction* (see Exercise 4.5(10b)), where projective planes are constructed from spreads of projective spaces. It can be shown that every translation plane gives rise to a spread, and from every spread, the Bruck–Bose construction gives rise to a translation plane.

The problem of discovering all quasifields, or of all spreads, is not yet solved. Non-isomorphic quasifields can yield the ‘same’ plane. Criteria, weaker than isomorphism, are needed to decide if two quasifields yield the same plane. One such ‘weak’ criterion is that of isotopism. (See *Hughes and Piper: Projective Planes.*) Unfortunately, while isotopic quasifields give rise to the same plane, there exist translation planes which can be coordinatised by non-isotopic quasifields. For example, $\mathcal{H}(9, f)$, with $f = x^2 + 1$ is a nearfield, while $\mathcal{H}(9, g)$, with $g = x^2 - x - 1$ is not a nearfield. As a consequence, $\mathcal{H}(9, f)$ and $\mathcal{H}(9, g)$ are not isotopic. They both, however, give rise to the unique non-Desarguesian translation plane of order 9.

In fact, it has been proved that there are exactly four projective planes of order 9, namely: PG$(2, 9)$, the Hall translation plane of order 9, the Hall dual translation plane of order 9, and the Hughes plane of order 9.

data 11.04.2020

A) 5 Coordinatising a projective plane

Pag. 99 (112)

6 Non-Desarguesian Planes

Pag. 110 (123)

Pag. 112 (125)

Example 6.4

Prove that if a collineation σ of a projective plane π is such that

1. some line l is fixed pointwise;
2. two points $P, Q \in l$ are fixed,

then $\sigma = 1$.

B) Books_67_2020/
Mathematics of Shape Description/
Pijush K. Ghosh _ Mathematics of Shape Description ..pdf

Pag. 77 (96)

3. Algebraic Structures for Shape Description

5. Arithmetics of Geometric Shape
Pag. 165 (184)

/ 3" ")*+,)))-) #% 2
; 3 3 " 9 5!%@ F G 7 H "<:

/)*+,)))-) #% < I 4
3 " 3 = * 3 33 3 " 3 3 !@
G 7 3"

(&& 56 # * / 3 ; 3

(6 !% * 4 3

5.1 The Motivation for a Shape Arithmetic

5.1.1 Does Negative Shape Exist?

Our previous chapter ended up posing a simple question: “Is it possible to do addition and subtraction (or multiplication and division) with geometric shapes as we do in ordinary arithmetic with numbers?” In other words, for a geometric shape, does its inverse – that is, *negative shape* – exist? If this were possible, then we would obtain a remarkable insight into analyzing and synthesizing shapes, just as we have in the case of numbers. Unfortunately, the precise answer to this question is “No,” because the *algebraic structures* of various systems with geometric shapes are, in general, *weaker* than those of systems with numbers. For example, consider some of the most familiar algebraic systems with numbers, such as $(\mathbb{Z}, +)$, $(\mathbb{Q}, +)$, and $(\mathbb{R}, +)$, where $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$ denote the sets of integers, rational numbers, and real numbers, respectively, and the composition operation “$+$” denotes the ordinary addition operation. All of these systems are groups. If we exclude the number 0 from all these sets and consider ordinary multiplication “$\cdot$” as the composition operation, the corresponding systems, $(\mathbb{Z} - \{0\}, \cdot)$, $(\mathbb{Q} - \{0\}, \cdot)$, and $(\mathbb{R} - \{0\}, \cdot)$, again turn out to be groups.

On the other hand, take some of the common systems with geometric shapes, such as $(\mathcal{G}, \cup)$, $(\mathcal{G}, \cap)$, $(\mathcal{G}, \oplus)$, or $(\mathcal{G}, \ominus)$, where $\mathcal{G}$ denotes the set of all geometric shapes (more formally, the set of all subsets of real Euclidean space R^d), and $\cup$, $\cap$, $\oplus$, and $\ominus$ denote, respectively, set union, set intersection, and the Minkowski (vector) addition and decomposition of two sets of points. All of these systems are *monoids* – but not groups. We may recall that a group has a *stronger* structure than a monoid, since a restriction imposed on the elements of a monoid – namely, the existence of an inverse for each element – results in a group.

The following question then naturally arises: “Is it possible to extend our notion of geometric shapes so that with a convenient composition operation we can form an algebraic group-like structure?”

This is indeed a very fundamental question. In this chapter, we shall attempt a much modest task by restricting our domain of shapes to the polygons in E^2 and the polyhedra in E^3, where E denotes Euclidean space, and approaching the problem in a more intuitive way, rather than becoming strictly mathematical.

5

Arithmetics of Geometric Shape

5.1 The Motivation for a Shape Arithmetic

5.1.1 Does Negative Shape Exist?

Our previous chapter ended up posing a simple question: "Is it possible to do addition and subtraction (or multiplication and division) with geometric shapes as we do in ordinary arithmetic with numbers?" In other words, for a geometric shape, does its inverse – that is, *negative shape* – exist? If this were possible, then we would obtain a remarkable insight into analyzing and synthesizing shapes, just as we have in the case of numbers. Unfortunately, the precise answer to this question is "No," because the *algebraic structures* of various systems with geometric shapes are, in general, *weaker* than those of systems with numbers. For example, consider some of the most familiar algebraic systems with numbers, such as $(\mathbb{Z}, +)$, $(\mathbb{Q}, +)$, and $(\mathbb{R}, +)$, where $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$ denote the sets of integers, rational numbers, and real numbers, respectively, and the composition operation "$+$" denotes the ordinary addition operation. All of these systems are groups. If we exclude the number 0 from all these sets and consider ordinary multiplication "$\cdot$" as the composition operation, the corresponding systems, $(\mathbb{Z} - \{0\}, \cdot)$, $(\mathbb{Q} - \{0\}, \cdot)$, and $(\mathbb{R} - \{0\}, \cdot)$, again turn out to be groups.

On the other hand, take some of the common systems with geometric shapes, such as $(\mathcal{G}, \cup)$, $(\mathcal{G}, \cap)$, $(\mathcal{G}, \oplus)$, or $(\mathcal{G}, \ominus)$, where $\mathcal{G}$ denotes the set of all geometric shapes (more formally, the set of all subsets of real Euclidean space R^d), and $\cup$, $\cap$, $\oplus$, and $\ominus$ denote, respectively, set union, set intersection, and the Minkowski (vector) addition and decomposition of two sets of points. All of these systems are *monoids* – but not groups. We may recall that a group has a *stronger* structure than a monoid, since a restriction imposed on the elements of a monoid – namely, the existence of an inverse for each element – results in a group.

The following question then naturally arises: "Is it possible to extend our notion of geometric shapes so that with a convenient composition operation we can form an algebraic group-like structure?"

This is indeed a very fundamental question. In this chapter, we shall attempt a much modest task by restricting our domain of shapes to the polygons in E^2 and the polyhedra in E^3, where E denotes Euclidean space, and approaching the problem in a more intuitive way, rather than becoming strictly mathematical.

Mathematics of Shape Description Pijush K. Ghosh and Koichiro Deguchi
© 2008 John Wiley & Sons (Asia) Pte Ltd

To devise a possible approach, we can draw an analogy with the lessons of number systems, as described in the previous chapters. We can start with the system $(\mathbb{N}, +)$, which is a monoid, where $\mathbb{N}$ denotes the set of natural numbers; that is, $\mathbb{N} = \{0, 1, 2, \cdots\}$. Let us take a small subset of $\mathbb{N}$, say $A = \{0, 1, 2\}$, and a composition operation $+$ redefined by the following table:

+	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

It is evident that the system $(A, +)$ is a group, since each of the elements of A has a unique inverse. You must have realized that we are essentially using *modulo* 3 arithmetic. It is well known that the relation "$a = b$ (*modulo* m)" is an *equivalence relation* that partitions $\mathbb{N}$ into m nonoverlapping subclasses, called *equivalence classes*.The equivalence classes may allow us to form a group structure, as was done in the above example. Since $(A, +)$ is a group, it is possible to solve an equation such as

$$x + 2 = 1$$

using this small system.

Now, instead of taking a small subset of $\mathbb{N}$, such as A, if we consider the whole of $\mathbb{N}$ and try to solve an equation of the above form, we are led to the concept of negative integers, $-1, -2, \cdots$. The set of natural numbers $\mathbb{N}$ augmented by the set of negative integers forms the set of integers $\mathbb{Z}$, and the system $(\mathbb{Z}, +)$ is now obtained as a group where both additions and subtractions become possible.

5.1.2 What Form Must Negative Shapes Take?

Now, let us confirm the theories on the extension of the geometric shape concept that we have constructed so far in this book.

We already have the next fundamental theory that tells us about this extension:

> *Extension of Algebraic Systems:* If an algebraic system $(A, \circ)$ where $\circ$ is a binary operation on the element in a set A has a *unit (identity) element e* which satisfies
>
> $$a \circ e = e \circ a = a$$
>
> for any element $a \in A$, and has next three properties that,
>
> 1. $a \circ b = b \circ a$, (*Commutativity*)
> 2. $(a \circ b) \circ c = a \circ (b \circ c)$, (*Associativity*)
> 3. If $a \circ b = a \circ c$, then $b = c$, (*Injectivity*)
>
> for any elements $a, b, c \in A$, then A can be extended to $A' \supseteq A$ so that all elements $a' \in A'$ have their own respective inverses in A'; that is, $\exists b' \in A'$, for $\forall a' \in A$ and $a' \circ b' = e$.
> Bourbaki, *Algebra I* [11]; see also Section 3.2

Let us recall our system of $(\mathcal{G}, \oplus)$. A single point, say $\{o\}$ (the origin), is a unit element of this system, because, for any geometric shape S, $S \oplus \{o\} = \{o\} \oplus S = S$. (It should be noted that the origin $\{o\}$ is not a unique unit element for the operation of $\oplus$. Actually, every single point $\{p\}$ can be a unit element. This is another substantial weak point in our theory.)

Now, we denote the set of all the convex polygons by $\mathcal{C}$; that is, $\mathcal{C} \subset \mathcal{G}$. Then, all of the above properties 1, 2, and 3 hold in the system $(\mathcal{C}, \oplus)$, as was shown in the previous chapter. This means that the set of all convex polygonal shapes $\mathcal{C}$ can be extended to include the inverses of the shapes. For example, for a rectangular shape R, the above fundamental theory says that we could consider its inverse shape R^{-1} so that $R \oplus R^{-1} = \{o\}$ (or a single point). Also, a circular shape C could have its inverse shape C^{-1}.

So, the first problem here is: What, in reality, are the inverse shapes R^{-1} and C^{-1}? The next question is: "How is the inverse operation actually defined to generate R^{-1} and C^{-1} from R and C, respectively?" Especially, how does the inverse operation of $\oplus$ relate to $\ominus$? And the third problem is: "If we have a proper extension of the set of the convex polygons $\mathcal{C}$, which includes their negative shapes, can it be further extended for all the geometric shapes in $\mathcal{G}$?"

Just as we said at the beginning of this section, the general answer to the third question is "No." Therefore, the question must be rewritten as follows: "To what extent can the set of the convex polygons $\mathcal{C}$ can be defined such that each of the element shapes has its own inverse?"

We present solutions to some of these problems in this chapter, where negative shapes in terms of Minkowski addition are defined as being closely related to Minkowski decomposition. We will only treat convex objects in this chapter: the results will be extended to nonconvex objects in the next chapter.

5.2 Morphology and the Theory of Numbers

5.2.1 Morphology for High-Level Vision

The two operations of Minkowski addition $\oplus$ (also called dilation) and Minkowski decomposition $\ominus$ (also known as erosion), introduced in the previous chapter, form the kernel of all other morphological operations.

To this day, morphological operations are mostly used for low-level image processing; that is, for early processing of binary or gray-scale discrete images [90]. Their applications in representing or understanding two- or three-dimensional geometric objects for the purpose of high-level object recognition are very limited. This is certainly an enigmatic situation, considering the fact that these operations are essentially functions of the form $f : \mathcal{G}(E^d) \rightarrow \mathcal{G}(E^d)$, where $\mathcal{G}(E^d)$ denotes the power set of real d-dimensional Euclidean space E^d; their definitions do not discriminate between whether the underlying sets are discrete images or continuous geometric objects. Moreover, a number of researchers have pointed out the relevance of these operations in high-level vision applications such as geometric modeling, spatial planning, description and understanding of biological forms, crystallography and textured object modeling, and so on [27, 28, 29, 60, 84]. The close resemblance between the representation of a generalized cylinder and the Minkowski addition operation has also been observed.

One factor that may account for this situation has to do with the computational problems associated with Minkowski operations on continuous objects. Note that any binary image can be modeled as a set consisting of a finite number of discrete points. Therefore, in computing the Minkowski sum $A \oplus B$ of two binary images A and B, we can directly use the set operations of Minkowski addition:

$$A \oplus B = B \oplus A = \bigcup_{b \in B} A_b = \bigcup_{a \in A} B_a.$$

Similarly, Minkowski decomposition $A \ominus B$ of a binary image A can be computed by means of

$$A \ominus B = \bigcap_{-b \in \check{B}} A_{-b}.$$

On the other hand, a continuous "well-formed" object (such as a polygon, circle, ellipse, etc. in two dimensions, or a polyhedron, sphere, ellipsoid, etc. in three dimensions) cannot be specified as a collection of a finite number of points. In general, such an object, say A, is specified by its oriented boundary, say ∂A. This necessitates computation of the boundary $\partial(A \oplus B)$ or $\partial(A \ominus B)$ of the products from the boundaries ∂A and ∂B of the operands. But it can immediately be seen that $\partial(A \oplus B) \neq \partial A \oplus \partial B$ and also $\partial(A \ominus B) \neq \partial A \ominus \partial B$.

In general, $\partial(A \oplus B) = f(\partial A, \partial B)$, and also $\partial(A \ominus B) = g(\partial A, \partial B)$, where "$f$" and "$g$" denote some complicated functions whose computations turn out to be quite nontrivial.

5.2.2 The Resemblance between Morphology and the Theory of Numbers

We can observe a remarkable similarity between some number-theoretic results and morphological theorems, particularly in the domain of convex objects. Let us consider the number system $(\mathbb{N}, \cdot, /)$, where $\mathbb{N}$ denotes the set of natural numbers $\{0, 1, 2, \ldots\}$ and "$\cdot$" and "$/$" denote the multiplication and division operations, respectively. Since, within $\mathbb{N}$, the division operation is not an exact inverse, but only a "restricted" inverse of the multiplication operation, we can define the division of a number m by another number n as $\lfloor m/n \rfloor$; the floor function notation $\lfloor x \rfloor$ means the greatest integer less than or equal to x. This number system can be compared with the geometric system $(\mathcal{K}, \oplus, \ominus)$, where $\mathcal{K}$ denotes the set of all compact convex subsets of E^d. We refer to Table 5.1, where $A, B, C \in \mathcal{K}$ and $m, n, p \in \mathbb{N}$.

From the table, it appears as though $A \oplus B$ in the convex domain is similar to $m \cdot n$, whereas $A \ominus B$ is similar to $\lfloor m/n \rfloor$. In this chapter, we do not attempt to seek why such a resemblance exists. Instead, we adopt the computational guideline that can be derived from this resemblance.

Table 5.1 The resemblance between the morphological system and the integer number system

	System $(\mathcal{K}, \oplus, \ominus)$	System $(\mathbb{N}, \cdot, /)$
1.	$A \oplus B = B \oplus A$	$m \cdot n = n \cdot m$
2.	$A \oplus (B \oplus C) = (A \oplus B) \oplus C$	$m \cdot (n \cdot p) = (m \cdot n) \cdot p$
3.	$A \supseteq B$ implies $C \ominus A \subseteq C \ominus B$	$m > n$ implies $\lfloor p/m \rfloor \leq \lfloor p/n \rfloor$
4.	$A \subseteq B \ominus C$ iff $B \supseteq A \oplus C$	$m \leq \lfloor n/p \rfloor$ iff $n \geq m \cdot p$
5.	$A \oplus B = ((A \oplus B) \ominus B) \oplus B$	$m \cdot n = \lfloor (m \cdot n)/n \rfloor \cdot n$
6.	$A \ominus B = ((A \ominus B) \oplus B) \ominus B$	$\lfloor m/n \rfloor = \lfloor (\lfloor m/n \rfloor \cdot n)/n \rfloor$
7.	$(A \ominus B) \ominus C = A \ominus (B \oplus C)$	$\lfloor \lfloor m/n \rfloor / p \rfloor = \lfloor (m/(n \cdot p) \rfloor$
8.	$A \oplus (B \ominus C) \subseteq (A \oplus B) \ominus C$	$m \cdot \lfloor n/p \rfloor \leq \lfloor (m \cdot n)/p \rfloor$
9.	$(A \ominus B) \oplus C \subseteq (A \oplus C) \ominus B$	$\lfloor m/n \rfloor \cdot p \leq \lfloor (m \cdot p)/n \rfloor$
10.	$(A \ominus B) \oplus B \subseteq A$	$\lfloor m/n \rfloor \cdot n \leq m$
11.	$(A \ominus B) \oplus B$ $= (((A \ominus B) \oplus B) \ominus B) \oplus B$	$\lfloor m/n \rfloor \cdot n$ $= \lfloor (\lfloor m/n \rfloor \cdot n)/n \rfloor \cdot n$
12.	If $A \oplus B = A \oplus C$, then $B = C$	If $m \cdot n = m \cdot p$, then $n = p$

The value of m/n is, in general, a real number that consists of the "integer part" $\lfloor m/n \rfloor$ and the "fractional part" $m/n - \lfloor m/n \rfloor$. In the integer number domain, at the time of division we discard that fractional part and take only the integer part. We find that the morphological operations can also be treated in a similar way. The idea is to devise an operation that is the "exact" inverse of the Minkowski addition operation, such that the application of this operation on two convex objects will produce a generalized geometric object (like producing a rational number by the division of two integers) that has a physically "realizable part" (analogous to the integer part) and a "nonrealizable part" (analogous to the fractional part). The Minkowski decomposition $A \ominus B$ can then be thought of as the realizable part of the object after the nonrealizable part is discarded.

The nonrealizable part of a generalized object will be referred to as the *negative* part. (We will avoid using terms such as "nonrealizable" or "fractional" part, since they have some widely accepted literal meanings.) The realizable part – that is, any ordinary geometric object – can be called *positive*.

In our subsequent discussions, we will show how such a computational strategy can be devised.

5.3 Boundary Representation by Support Functions for Morphological Operations

5.3.1 The Support Function Representation

Our starting point in tackling this problem is a basic result given by Shephard [93], which is concerned with the boundary addition of d-dimensional convex polytopes. Informally, a convex polytope in E^d is the d-dimensional analogue of a convex polygon in E^2 or a convex polyhedron in E^3. Mathematically, a polytope is the convex hull of a finite set of points in E^d. Shephard's result is stated in terms of the faces of polytopes. In this subsection, we briefly discuss the formal notion of the faces of a d-dimensional polytope, and then we state the main result in the next subsection.

A face of a convex polytope A in E^d is defined in terms of the *supporting function* and the *supporting hyperplane* of the polytope. The representation of a convex body by means of its support function is a classical approach [8,34] that was introduced by Minkowski in 1903, and has been widely used by mathematicians since that time.

Let $A \in E^d$ be a convex body (i.e., a nonempty compact convex set) in real Euclidean d-dimensional space. The supporting function $H(A, v)$ of A for all $v \in E^d$ (provided that $v \neq 0$ – that is, v – is an arbitrary vector different from the origin O) is defined by

$$H(A, v) = \sup\{ \langle a, v \rangle \mid a \in A\}, \tag{5.1}$$

where $\langle a, v \rangle$ denotes the scalar (or inner) product of the vectors a and v, and "sup" denotes the "supremum' or "least upper bound."

Since $H(A, \lambda v) = \lambda H(A, v)$ for any real number $\lambda > 0$, the support function $H(A, v)$ is completely determined by its value on the unit sphere $\|v\| = 1$. Thus, if u denotes a unit vector – that is, $u \in S^{d-1}$, where S^{d-1} is the unit sphere in E^d with its center at the origin – it is most convenient to use the function $H(A, u)$ as the support function of A. $H(A, u)$ is a *complete representation* of the convex body A, since the values of $H(A, u)$ for all $u \in S^{d-1}$ completely specify A in the following way:

$$A = \{x \in E^d \mid \; < x, u > \; \leq H(A, u) \text{ for all } u \in S^{d-1}\}, \tag{5.2}$$

which, in words, means that A is the intersection of all the half-spaces $< x, u > \leq H(A, u)$.

Although a large part of the theory of convex bodies in mathematics uses $H(A, u)$ as the standard representation of a convex body A, in the fields of computer vision, graphics, and image processing, the support function representation is less frequently used. One primary reason is that the function $H(A, u)$ is, in general, a continuous function of u, whose closed-form specification may not be readily available. It may also appear that the $H(A, u)$-representation is less intuitive than the boundary representation or half-space representation of convex bodies.

5.3.2 The Support Function is a Signed Distance

If, for some $u \neq 0$, we have $H(A, u) < \infty$ (this condition ensures that A is bounded), then the hyperplane

$$L(A, u) = \left\{ p \in E^d \mid \langle p, u \rangle = H(A, u) \right\} \tag{5.3}$$

is called the *supporting hyperplane* of A with *outer/outerward normal* u. In E^2, the supporting hyperplane becomes the supporting line of a convex polygon A with outer normal u, while in E^3 it is the supporting plane of a convex polyhedron A.

It is easy to see in Figure 5.1 that the support function $H(A, u)$ is precisely the "signed distance" from the origin O to the supporting hyperplane $L(A, u)$. This distance is to be considered positive if A and the origin lie on the same side of the supporting hyperplane, negative if A and the origin are separated by the supporting hyperplane, and zero if the origin lies in the supporting hyperplane. It is, therefore, convenient for all practical purposes to assume that the origin lies in the interior of A, so that the function $H(A, u)$ is positive for every u.

To provide examples, we consider three simple two-dimensional convex figures – a unit circle having its center at the origin, a triangle, and an ellipse – and show their corresponding $H(A, u)$'s in Figure 5.2. Note that for the unit circle whose center is at O, $H(A, u) = 1$ for all u.

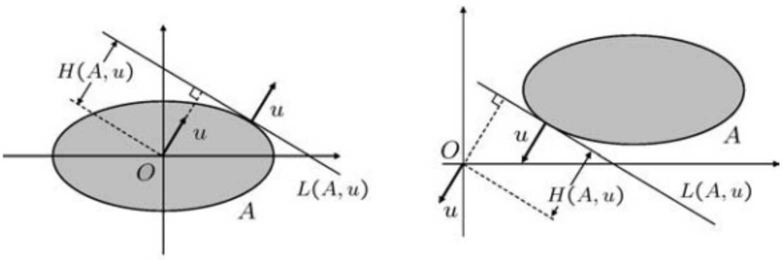

(a) $H(A, u) > 0$ (b) $H(A, u) < 0$

Figure 5.1 The support function $H(A, u)$ is the signed distance from the origin O to the hyperplane $L(A, u)$. It is (a) positive ($H(A, u) > 0$)if A and the origin lie on the same side of the supporting hyperplane, and (b) negative ($H(A, u) < 0$) if A and the origin are separated by the supporting hyperplane

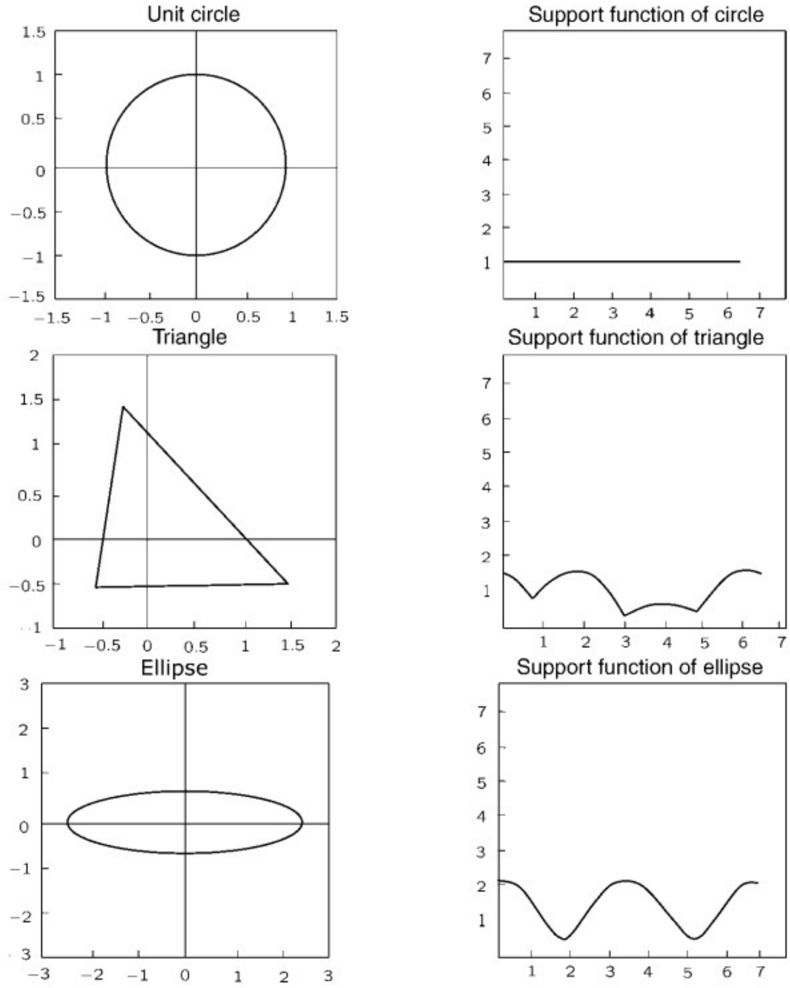

Figure 5.2 The support function representations of some typical convex figures; since in E^2 a unit vector $u = (\cos\theta, \sin\theta)$, it is specified in the graph by the angle θ (in radians) along the x-axis and the corresponding value of $H(A, u)$ along the y-axis

In fact, for some simple convex bodies, we can obtain closed-form representations of the support functions as follows:

1. For a singleton point set $\{a\}$ in E^d, $H(a, v) =< a, v >$.
2. For a ball B_α having radius α and center O, $H(B_\alpha, v) =< \alpha \frac{v}{\|v\|}, v >= \alpha\|v\|$.
3. For a line segment L_{ab} joining points a and b,

$$H(L_{ab}, v) = \max(< a, v >, < b, v >).$$

5.3.3 From Support Function Representation to Boundary Representation and Vice Versa

Assume that the support function $H(A, u)$ of a convex body A is given for all $u \in S^{d-1}$. How do we determine the boundary points of A? We answer this question by following the approach given in [8]. The boundary points of A, where the outer normal is either u or parallel to u, are precisely the set of points

$$F(A, u) = L(A, u) \cap A. \tag{5.4}$$

$F(A, u)$ is termed the face of A having outer normal u. Obviously, the dimension of $F(A, u)$ is at most $d - 1$.

Figure 5.3 shows examples of faces in E^2 and E^3 with their supporting functions and supporting hyperplanes.

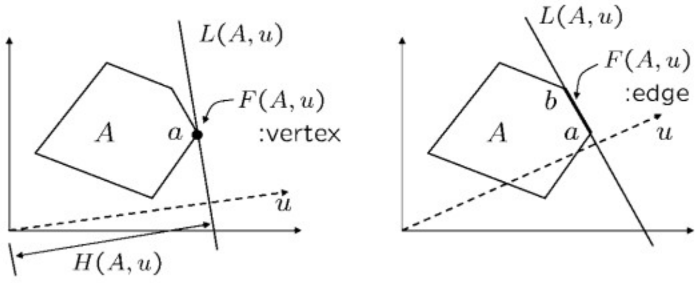
L(A, u)
F(A, u)
:vertex
A
a
u
H(A, u)
L(A, u)
F(A, u)
:edge
b
A
a
u

(a) $F(A, u)$ for 2-dimensional polygon A

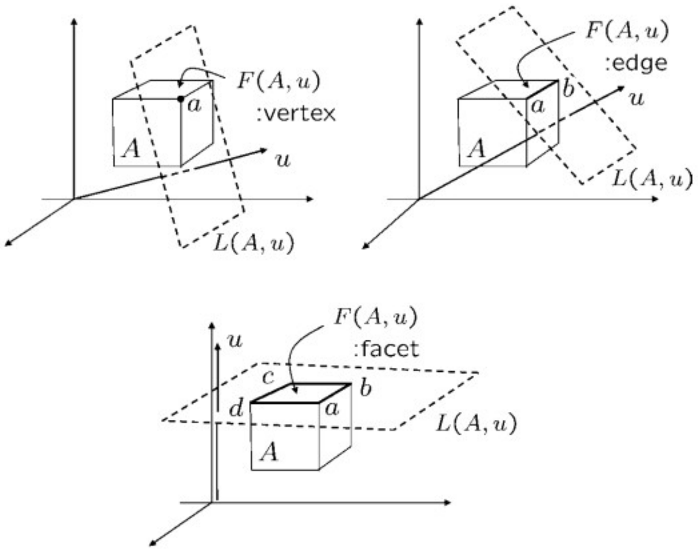
F(A, u)
:vertex
a
A
u
L(A, u)
F(A, u)
:edge
b
a
u
A
L(A, u)
F(A, u)
:facet
u
c
b
d
a
A
L(A, u)

(b) $F(A, u)$ for 3-dimensional polyhedron A

Figure 5.3 The faces $F(A, u)$ of (a) a two-dimensional polygon and (b) a three-dimensional polyhedron

Now let u_k be some fixed direction, and let $F(A, u_k)$ be the face of A that has the outer normal u_k. Our task is to determine $F(A, u_k)$. Since $F(A, u_k)$ is a subset of the supporting hyperplane $L(A, u_k)$, for any point $x \in F(A, u_k)$ it must satisfy (because of (5.3))

$$< x, u_k > \; = \; H(A, u_k). \tag{5.5}$$

Moreover, $F(A, u_k)$ is also a subset of the convex body A. So a point x of $F(A, u_k)$ must also satisfy, for all u, all of the following inequalities (see (5.2)):

$$< x, u > \; \leq \; H(A, u). \tag{5.6}$$

Now, if w is a unit vector in an arbitrary direction and $u = u_k + \lambda w$, where $\lambda > 0$, then using (5.5) and (5.6) we can write

$$< x, w > \; \leq \; \frac{H(A, u_k + \lambda w) - H(A, u_k)}{\lambda}. \tag{5.7}$$

Therefore, letting $\lambda \to 0$,

$$< x, w > \; \leq \; H'_w(A, u_k), \tag{5.8}$$

where $H'_w(A, u_k)$ is the directional derivative of $H(A, u)$ at $u = u_k$, in the direction of the unit vector w.

If we consider all of the w's (i.e., unit vectors in all directions), then the inequalities of (5.8) represent the intersection of the corresponding half-spaces, and that intersection is precisely the face $F(A, u_k)$.

Thus we arrive at the following proposition.

Proposition 5.1. *If $H(A, u)$ is the support function of a convex body A, then the face $F(A, u_k)$ of A having the outer normal u_k has the support function $H'_u(A, u_k)$.*

5.3.4 Necessary and Sufficient Conditions for a Function to be a Support Function

The support function $H(A, v)$ is a scalar function of the vector v and hence, in the present case, a mapping from E^d to R. A natural question is: "Which functions from E^d to R could be characterized as support functions?" This question is particularly important to us since, from the next section onward, we will be concerned as to whether or not a function $H(A, v) * H(B, v)$ resulting from some operation $*$ on the given support functions $H(A, v)$ and $H(B, v)$ is a valid support function. Consider, for example, the situation shown in Figure 5.4, where no object could have all of the support function values $H(A, u_1) * H(B, u_1)$, $H(A, u_2) * H(B, u_2)$, $H(A, u_3) * H(B, u_3)$, $H(A, u_4) * H(B, u_4)$, since the supporting line corresponding to the value $H(A, u_3) * H(B, u_3)$ is a redundant line.

For characterization of the support function, we can state the following result.

Proposition 5.2. *(a) Every real-valued function $F(v)$ defined for all $v \in E^d$ and satisfying the properties*

(i) $F(o) = 0$,
(ii) $F(\lambda v) = \lambda F(v)$, *for* $\lambda > 0$, *and*

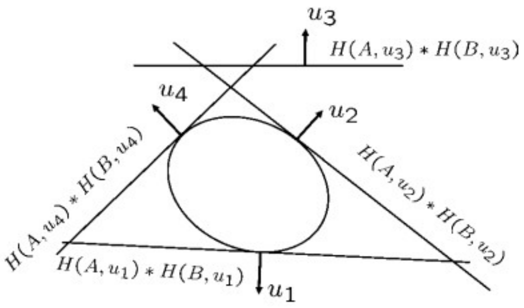

Figure 5.4 An example where the resulting function $H(A, v) * H(B, v)$ cannot be a valid support function; the supporting line corresponding to the value $H(A, u_3) * H(B, u_3)$ is a redundant line

(iii) $F(v + w) \leq F(v) + F(w)$*, is a support function of a convex body.*
(b) If $F(v)$ *is a support function, then the convex body that it represents must be the intersection of all the half-spaces* $< x, v > \leq F(v)$*.*

Part (a) is a classical result, the proof of which can be found in [8]. Part (b) follows easily from (5.2). For a review on the topic of consistency checking of the support function, see also [48]. Condition (iii) in the proposition – that is, $F(v + w) = F(v) + F(w)$ – appears to be the most nonintuitive compared to the other two conditions. However, its connection with the question of redundant half-spaces can be easily demonstrated. You are recommended to visualize this condition for yourself.

5.4 Geometric Operations by Means of Support Functions

We assume that the support functions $H(A, u)$ and $H(B, u)$ of two convex bodies A and B are given. We will now show that some simple algebraic operations on the support functions result in fairly complicated geometric operations involving the bodies A and B.

5.4.1 MAX and MIN Operations (Convex Hull and Intersection)

5.4.1.1 The MAX Operation

The MAX operation is defined as

$$\max\{H(Au),\ H(B, u)\},$$

for every $u \in S^{d-1}$, where "$\max\{\alpha, \beta\}$" specifies the maximum of the two real numbers α and β.

It is not difficult to prove that the MAX operation results in the *convex hull* operation of the union of A and B. Toward that end, we need the following proposition.

Proposition 5.3. *If* $H(A, u)$ *and* $H(B, u)$ *are the support functions of two convex bodies* A *and* B*, then the inequality*

$$H(A, u) \leq H(B, u), \qquad \text{for all } u,$$

holds if and only if $A \subseteq B$.

The proof of the proposition follows immediately from (5.1) and (5.2) (see also [8]).

We can now state and prove the main result concerning the MAX operation.

Proposition 5.4. *(a) The function* $\max\{H(A, u), H(B, u)\}$ *is a support function.*

(b) $\max\{H(A, u), H(B, u)\} = H(C, u)$*, where* $C = conv(A \cup B)$*. (Here,* $conv(X)$ *denotes the* convex hull *of the set* X*.)*

Proof: (a) See Proposition 5.2. Conditions (i) and (ii) obviously hold for $\max\{H(A, u), H(B, u)\}$, since they hold for both $H(A, u)$ and $H(B, u)$. Only condition (iii) – that is, the subadditivity condition – needs to be proved.

Let us write $F(u) = \max\{H(A, u), H(B, u)\}$, and assume that u_1, u_2 are two arbitrary unit vectors. Then,

$$\begin{aligned} F(u_1) + F(u_2) &= \max\{H(A, u_1), H(B, u_1)\} + \max\{H(A, u_2), H(B, u_2)\} \\ &\geq H(A, u_1) + H(A, u_2) \\ &\geq H(A, u_1 + u_2), \end{aligned}$$

since $H(A, u)$ is a support function.

Similarly, $F(u_1) + F(u_2) \geq H(B, u_1 + u_2)$.

Furthermore, since $F(u_1 + u_2) = \max\{H(A, u_1 + u_2), H(B, u_1 + u_2)\}$, $F(u_1 + u_2)$ is either equal to $H(A, u_1 + u_2)$ or equal to $H(B, u_1 + u_2)$.

Therefore, $F(u_1) + F(u_2) \geq F(u_1 + u_2)$, and hence $\max\{H(A, u), H(B, u)\}$ is a support function of some convex body – say, C.

(b) Let $\max\{H(A, u), H(B, u)\} = H(C, u)$, where C is some convex body. Since $H(C, u) \geq H(A, u)$ and also $H(C, u) \geq H(B, u)$ for all u, according to Proposition 5.3, $C \supseteq A \cup B$.

Let us assume that the convex hull $\text{conv}(A \cup B)$ is not C, but that $\text{conv}(A \cup B) = C'$, and that C' is strictly smaller than C; that is, $C' \subset C$. Therefore, according to Proposition 5.3, $H(C', u) \leq H(C, u)$, for all u. Let us, for some u, say that u_1, $H(C', u)$ is strictly smaller than $H(C, u)$; that is, $H(C', u_1) < \max\{H(A, u_1), H(B, u_1)\}$. But that is not possible, which means that $C = C'$. □

We show two convex polygons A and B (in the xy-space) in Figure 5.5(a) and their support function representations (i.e., A and B in the θ_ρ-space) in Figure 5.5(b). The MAX operation of the support functions and the resulting polygon, which is equal to $\text{conv}(A \cup B)$ in the θ_ρ-space, are shown in Figure 5.5(c) (heavy curves). The supporting lines corresponding to each respective curve are shown in Figure 5.5(d), as heavy lines, and then the resulting MAX convex polygon of A and B is given in Figure 5.5(e).

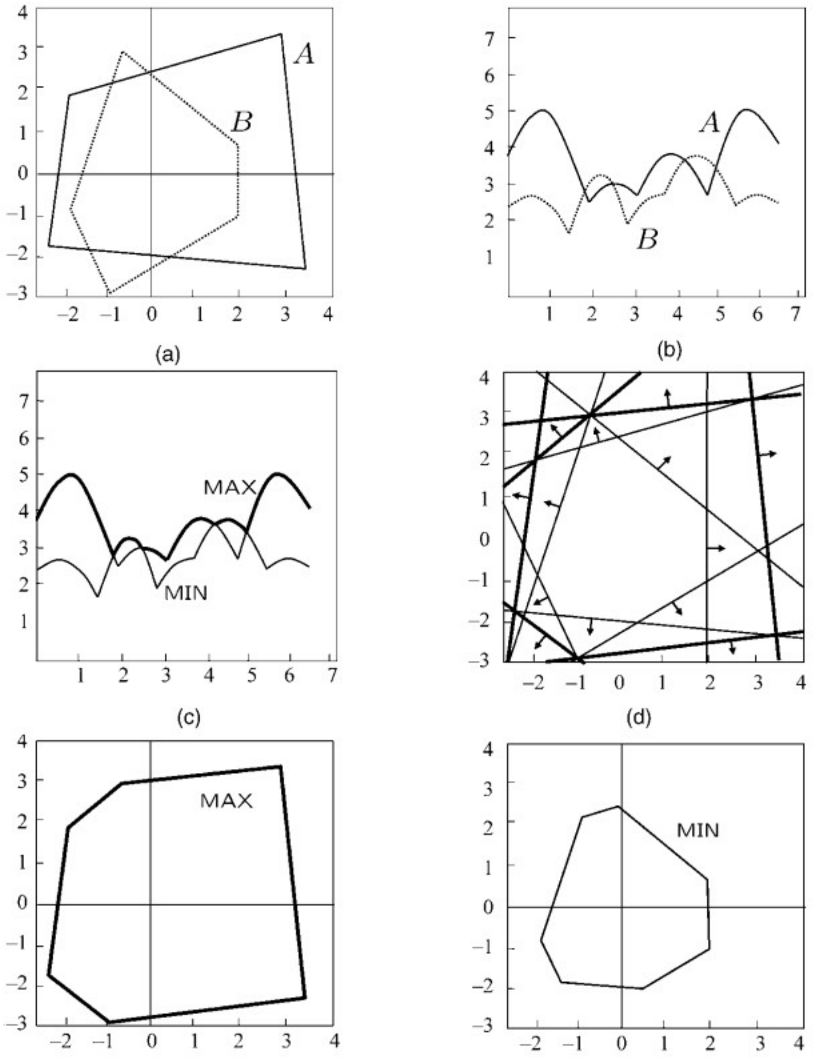

(a) (b)

(c) (d)

(e) (f)

Figure 5.5 A demonstration of the MAX and MIN operations: (a) the input convex polygons A and B (in the xy-space); (b) their support function representations (the θ_ρ-space); (c)$\max\{H(A, u), H(B, u)\}$ (heavy curves) and $\min\{H(A, u), H(B, u)\}$ (light curves) (the θ_ρ-space); (d) the supporting lines corresponding to $\max\{H(A, u), H(B, u)\}$ (heavy lines), and $\min\{H(A, u), H(B, u)\}$ (light lines); (e) the resulting MAX convex polygon (in the xy-space) represented by $\max\{H(A, u), H(B, u)\}$; and (f) the MIN convex polygon (in the xy-space) represented by $\min\{H(A, u), H(B, u)\}$

5.4.1.2 The MIN Operation

The MIN operation is similarly defined as

$$\min\{H(A, u), H(B, u)\},$$

for every $u \in S^{d-1}$, where "$\min(\alpha, \beta)$" denotes the minimum of the two real numbers α and β.

It is not difficult to show that the MIN operation performs the intersection operation $A \cap B$. But unlike the previous case, $\min\{H(A, u), H(B, u)\}$ is not a support function. As a result, some of the supporting hyperplanes defined by $\min\{H(A, u), H(B, u)\}$ may be redundant. But the common intersection of all the half-spaces (some of which may be redundant, corresponding to the redundant hyperplanes) defined by the function $\min\{H(A, u), H(B, u)\}$ will result in $A \cap B$.

To give an example, we consider the same two polygons as presented in Figure 5.5(a), and show the function $\min\{H(A, u), H(B, u)\}$ in Figure 5.5(c) (light curves). The corresponding supporting lines are depicted in Figure 5.5(d) (light lines). Note that some of the supporting lines are redundant. However, the common intersection of the half-spaces defined by all the supporting lines is $A \cap B$, which is shown in Figure 5.5(f).

5.4.2 Morphological Operations in Boundary Representation

The following result concerning Minkowski addition of convex polytopes can now be derived by using the boundary representation formalized in the previous subsection.

Theorem 5.5. *If A and B are two convex polytopes in E^d, then, for every $u \in E^d$,*

$$H(A \oplus B, u) = H(A, u) + H(B, u) \tag{5.9}$$

and

$$F(A \oplus B, u) = F(A, u) \oplus F(B, u). \tag{5.10}$$

Equation (5.9) comes directly from the definition of Minkowski addition $\oplus$; $A \oplus B = \{a + b \mid a \in A, b \in B\}$. Then, substituting (5.9) into (5.3) and (5.4), we have (5.10). A detailed proof is given either by Grünbaum [34] or Kelly and Weiss [50], but it is not difficult, and we leave it to the readers.

Let us denote the boundary of S by ∂S. Then, because ∂S consists of all the faces of $F(S, u)$,

$$\partial S = \bigcup_{u \in S^{d-1}} F(S, u).$$

This implies that, from (5.10),

$$\partial(A \oplus B) = \bigcup_{u \in S^{d-1}} \left(F(A, u) \oplus F(B, u)\right). \tag{5.11}$$

You may wonder why the $\oplus$ operation appears both on the left- and right-hand sides of (5.11). We shall, however, argue that the computation of the right-hand side finally reduces to two simple operations on real numbers: sorting of real numbers, and addition of real numbers. We will show this first for polygons – that is, in E^2 – and then for polyhedra; that is, E^3.

5.5 Morphological Operations on Convex Polygons

5.5.1 Computation by Means of Support Function Vectors

The boundary of a convex polygon A in E^d can be precisely defined by means of the supporting function of the polygon. From Theorem 5.5, we can immediately infer the following. For the addition of convex polygons, it is not necessary to compute the supporting function $H(A \oplus B, u)$ for every $u \in S^1$ (S^1 in E^2 is simply a unit circle). Since a convex polygon is completely specified by its oriented edges, it is sufficient to consider only those u's whose directions are the same as the outer normal directions of the edges of the summand polygons.

Let us, therefore, consider the class of convex polygons whose edges have the same outer normal vectors. In other words, any two polygons belong to this class have pairwise parallel and similarly directed edges. Note that every convex polygon in E^2 can be included in this class by means of introducing edges of zero length. Let $\mathcal{C}(U)$ denote this class, where U is the ordered set of outer normal vectors of the edges; that is, $U = \{u_1, \cdots, u_n\}$. In other words, if any convex polygon $A \in \mathcal{C}(U)$, then

$$A = \left\{ p \in E^2 \mid \langle p, u_i \rangle \leq \eta_i \; (i = 1, \cdots, n) \right\} \tag{5.12}$$

for some $\eta_i \in R \; (i = 1, \cdots, n)$, where R denotes the set of real numbers. Note that η_i is simply the value of the supporting function of A in the direction of the outer normal u_i.

As we know, each η_i along with u_i specifies a closed half-space (in two dimensions, a "half-space" is generally called a "half-plane"), and the intersections of all these half-spaces for $i = 1, \cdots, n$, in turn, specify all of the edges of A. Therefore, once U is given, A is specified completely by the vector $(\eta_1, \cdots, \eta_n)$. We call this the *supporting function vector* of A, and denote it by $h(A)$.

From Theorem 5.5, we can now easily derive the following result concerning convex polygons.

Proposition 5.6. *If two convex polygons $A, B \in \mathcal{C}(U)$ are represented by the supporting function vectors $(\eta_1^A, \cdots, \eta_n^A)$ and $(\eta_1^B, \cdots, \eta_n^B)$, respectively, then their Minkowski sum $A \oplus B$ is specified completely by the vector $(\eta_1^A + \eta_1^B, \cdots, \eta_n^A + \eta_n^B)$.*

Proposition 5.6 implies that Minkowski addition of two convex polygons can be seen as the vector addition $h(A) + h(B)$ of two points $h(A)$ and $h(B)$ in an n-dimensional space.

Let us clarify the idea described so far by means of an example (Figure 5.6). As is evident from the example, Minkowski addition of two convex polygons is quite straightforward by means of the supporting function vector $h(A) + h(B)$.

From this observation, we can now define the "exact" inverse of Minkowski addition. In computing $A \ominus B$, we shall first compute the vector $h(A) + (-h(B))$. This vector, like the vector $h(A) + h(B)$, again specifies n half-spaces having outer normals u_i's, but some of the half-spaces may become redundant in this case. It then turns out that the Minkowski decomposition $A \ominus B$ discards those redundant half-spaces and considers the rest. Let us state this result as the following proposition.

Proposition 5.7. *If two convex polygons $A, B \in \mathcal{C}(U)$ are represented by the supporting function vectors $(\eta_1^A, \cdots, \eta_n^A)$ and $(\eta_1^B, \cdots, \eta_n^B)$, respectively, then their Minkowski decomposition*

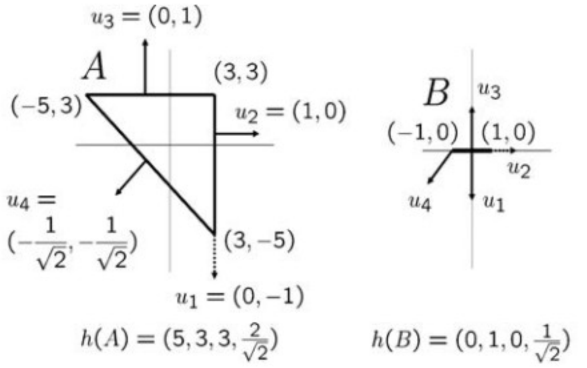
$u_3 = (0,1)$
A
$(3,3)$
$(-5,3)$
$u_2 = (1,0)$
$u_4 =$
$(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$
$(3,-5)$
$u_1 = (0,-1)$
$h(A) = (5,3,3,\frac{2}{\sqrt{2}})$
B
u_3
$(-1,0)$
$(1,0)$
u_2
u_4
u_1
$h(B) = (0,1,0,\frac{1}{\sqrt{2}})$

(a) Convex polygons A and B whose $u = (u_1, u_2, u_3, u_4)$

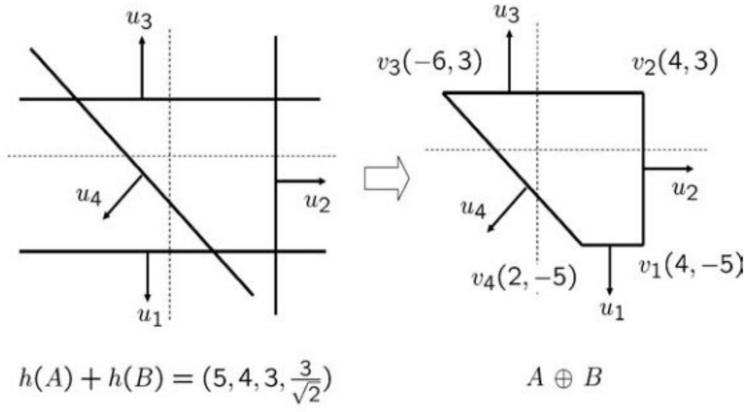
u_3
u_4
u_2
u_1
$h(A) + h(B) = (5,4,3,\frac{3}{\sqrt{2}})$
u_3
$v_3(-6,3)$
$v_2(4,3)$
u_2
u_4
$v_4(2,-5)$
$v_1(4,-5)$
u_1
$A \oplus B$

(b) Realization of Minkowski sum

Figure 5.6 Minkowski addition by means of supporting function vectors: (a) convex polygons A and B whose $u = (u_1, u_2, u_3, u_4)$; (b) the realization of the Minkowski sum from its supporting vector specification in terms of intersection of half-spaces, and then conversion of the sum polygon in terms of vertices and edges

$A \ominus B$ can be obtained by discarding the redundant half-spaces specified by the vector $(\eta_1^A - \eta_1^B, \cdots, \eta_n^A - \eta_n^B)$.

The novelty of our approach lies in our treatment of *not discarding* those redundant half-spaces, but retaining them as essentials.

We will show the principle of computation of the Minkowski decomposition by using the supporting function vectors in Figure 5.7. In computing the decomposition $A \ominus B$ as $h(A) + (-h(B)) = (\eta_1^A - \eta_1^B, \cdots, \eta_n^A - \eta_n^B)$ (in Figure 5.7(a)), note that one of the half-planes – namely, the half-plane corresponding to the outer normal u_1 – is redundant. The normal

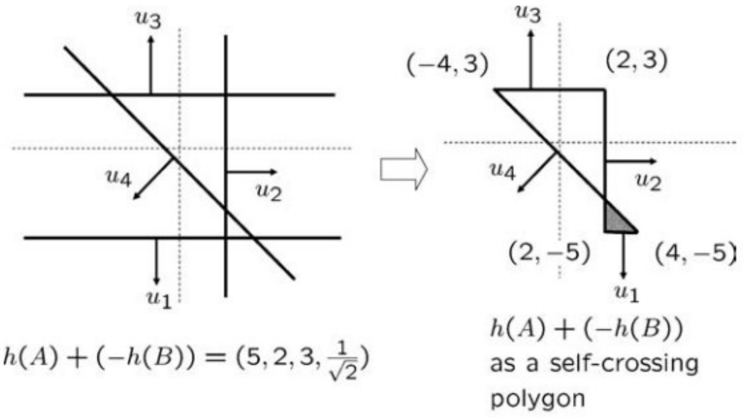
u_3
u_4
u_2
u_1
$h(A) + (-h(B)) = (5, 2, 3, \frac{1}{\sqrt{2}})$
$(-4, 3)$
$(2, 3)$
$(2, -5)$
$(4, -5)$
$h(A) + (-h(B))$
as a self-crossing polygon

(a) Support function vector $h(A) + (-h(B))$

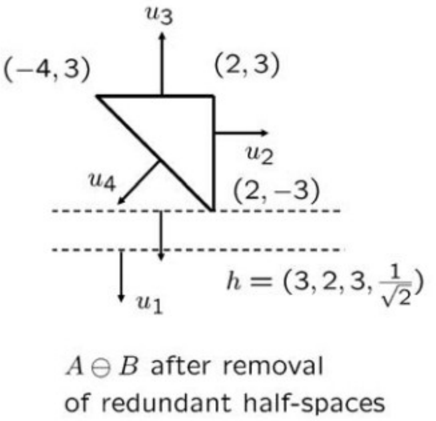
u_3
$(-4, 3)$
$(2, 3)$
u_2
u_4
$(2, -3)$
u_1
$h = (3, 2, 3, \frac{1}{\sqrt{2}})$
$A \ominus B$ after removal of redundant half-spaces

(b) Removal of redundant half-spaces

Figure 5.7 Minkowski decomposition by means of supporting function vectors: (a) the support function vector $h(A) + (-h(B))$ in terms of half-spaces, and then conversion of the half-spaces in terms of vertices and edges (which produces a self-crossing polygon); (b) the removal of redundant half-spaces – A and B are given in Figure 5.6

custom is to abandon this redundant half-plane altogether in further considerations. This effectively means an appropriate reduction of the corresponding value of the supporting function in the supporting function vector $h(A) + (-h(B))$. In our example, the corresponding supporting function value $\eta_1^A - \eta_1^B = 5$ has been reduced to 3 (see Figure 5.7(b)). This clearly explains why, in general, $(A \ominus B) \oplus B \subseteq A$.

5.5.2 *Computation by Means of Edges: The Emergence of the Boundary Addition Operation* $\uplus$

Representing a convex polygon A by means of its supporting function vector $h(A)$ is not a very common practice. A more popular representation is the *edge length* representation, where the boundary ∂A is represented by means of a starting vertex v_s^A and the length ι_i^A of the edge corresponding to every u_i; that is, $\partial A = \left\{ \{v_s^A\}, \{\iota_1^A, \cdots, \iota_n^A\} \right\}$. From the $h(A)$-representation, it is easy to arrive at the edge length representation by finding the intersection points of adjacent half-spaces (see the right-hand plots in Figures 5.6(b) or 5.7(a)).

If the other summand B is also represented in a similar way; that is, $\partial B = \left\{ \{v_s^B\}, \{\iota_1^B, \cdots, \iota_n^B\} \right\}$, then by a straightforward computation we get

$$h(A) + h(B) = \left\{ \{v_s^A + v_s^B\}, \{\iota_1^A + \iota_1^B, \cdots, \iota_n^A + \iota_n^B\} \right\}. \tag{5.13}$$

Equation (5.13) states that the boundary $\partial(A \oplus B)$ of the Minkowski sum of two convex polygons can easily be computed when the boundaries ∂A and ∂B are represented in their edge length forms. The computation essentially involves *addition of the lengths of the corresponding edges of the summands.* This operation may be termed the boundary addition operation and denoted by the symbol "$\uplus$." Equation (5.13) can, therefore, be rewritten as

$$\begin{aligned} \partial(A \oplus B) &= \partial A \uplus \partial B \\ &= \left\{ \{v_s^A + v_s^B\},\ \{\iota_1^A + \iota_1^B, \cdots, \iota_n^A + \iota_n^B\} \right\}. \end{aligned} \tag{5.14}$$

For computing Minkowski decomposition, we convert the support function vector $-h(B)$ into the equivalent edge length representation, which turns out to be

$$-h(B) = \left\{ \{-v_s^B\},\ \{-\iota_1^B, \cdots, -\iota_n^B\} \right\}.$$

The polygon represented by $-h(B)$, if drawn pictorially, looks like a hole or a *negative region*, without any positive region surrounding the hole (Figure 5.8(b)). This is because, for any ordinary convex polygon (Figure 5.8(a)), all the outer normals u_i's "diverge outward," whereas the normals u_i's of $-h(B)$ appear to "converge inward"; but the *directions* of the normals remain exactly the same in both cases. We can distinguish these two cases by introducing the concept of the "sense" of the outer normals. The outer normals of ordinary convex objects, which diverge outward, can be thought of as having a "positive" sense; whereas the normals of $-h(B)$ have a "negative" sense. We can consider the polygon represented by $-h(B)$ as the additive inverse of the polygon B, and we can denote it by the symbol B^{-1}.

As far as the geometric shape is concerned, B^{-1} appears exactly like the symmetrical set $\check{B}$. But note the differences among the objects B, $\check{B}$, and B^{-1}. The sense of every outer normal of both B and $\check{B}$ is the same – that is, positive – while the directions of the outer normals at the corresponding faces of the two objects are exactly opposite. Because of the positive sense of the outer normals, we consider both B and $\check{B}$ as positive objects. On the contrary, the directions of the outer normals at the corresponding faces of B and B^{-1} are exactly the same, but the senses are opposite. Because of the negative sense of the outer normals, B^{-1} can be considered as a negative object.

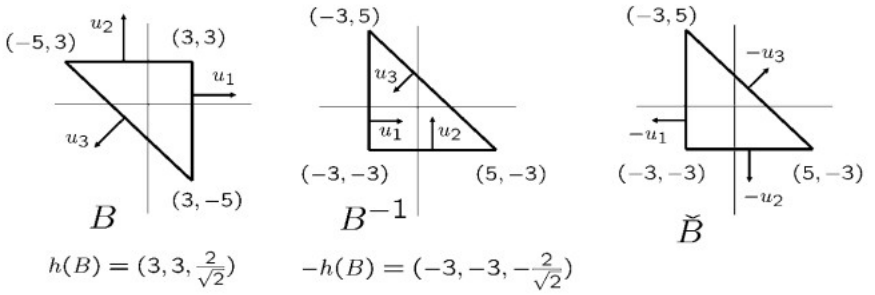

(a) Ordinary B polygon

(b) Additive inverse B^{-1} polygon

(c) $\breve{B}$: Symmetrical set of B

Figure 5.8 (a) The ordinary B polygon; (b) a geometric interpretation of the inverse shape B^{-1}; and (c) for comparison, the symmetrical set of B, $\breve{B}$

Using the notation of boundary addition, we can write

$$h(A) + (-h(B)) \equiv \partial A \uplus \partial B^{-1} = \left\{ \{v_s^A - v_s^B\}, \{\iota_1^A - \iota_1^B, \cdots, \iota_n^A - \iota_n^B\} \right\}. \tag{5.15}$$

Since $h(A) + (-h(B))$ may contain redundant half-spaces, the polygon represented by $\partial A \uplus \partial B^{-1}$ will be, in general, a self- crossing polygon (see the right-hand plot in Figure 5.7(a)). With regard to the senses of the outer normals, a self-crossing polygon may appear to have both a positive part and a negative part (the negative part is shown shaded in Figure 5.7(a)). Since we obtain $A \ominus B$ from $h(A) + (-h(B))$ by discarding the redundant half-spaces, equivalently $\partial(A \ominus B)$ is obtained from $\partial A \uplus \partial B^{-1}$ by discarding the negative part of the polygon $\partial A \uplus \partial B^{-1}$, and considering only its positive part. This can be expressed symbolically as

$$\partial(A \ominus B) = \mathrm{Pos}(\partial A \uplus \partial B^{-1}), \tag{5.16}$$

where $\mathrm{Pos}(X)$ denotes a unary operation, which extracts the positive portion of a self-crossing object X.

Note: In the case that we are only interested in the "shapes" of the objects, and not their positions in the plane, then the starting vertices v_s^A, v_s^B, and so on (which appeared in (5.14) and (5.15)) can be ignored.

5.5.3 Computation by Means of Slope Diagrams: The Unification of Minkowski Addition and Decomposition

The *slope diagram* representation of a polygon is essentially the edge length representation in a pictorial form. Since all the outer normals u_i's of a convex polygon lie on a unit circle S^1, in slope diagram representation we consider a unit circle as the basis. The representation scheme is as follows (see Figure 5.9):

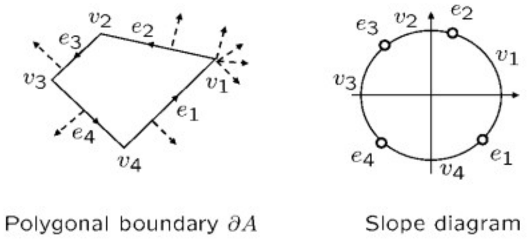

Polygonal boundary ∂A Slope diagram

(a) A typical convex polygon A and its slope diagram

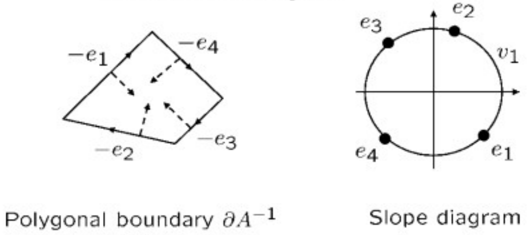

Polygonal boundary ∂A^{-1} Slope diagram

(b) The inverse polygon A^{-1} and its slope diagram

Figure 5.9 (a) A typical convex polygon A and its slope diagram; (b) the inverse polygon A^{-1} and its slope diagram

1. The outer normal direction at each edge of the polygon is represented by the corresponding point on a unit circle, which is called an *edge point*. (By "corresponding point," we mean that point on the unit circle where the outer normal direction is the same as the outer normal direction of the edge.)
2. At each vertex of the polygon, it is possible to draw innumerably many outer normals filling an angle (supplementary to the interior angle at the vertex). This set of outer normal directions at the vertex is represented by the corresponding arc on the unit circle, which is called a *vertex arc*.
3. Apart from the direction of an outer normal, the sense of the outer normal must also be indicated. If the sense of an outer normal is negative, it will be shown by heavy black points or arcs, while an outer normal that has a positive sense will be drawn using light lines.
4. The length of each edge is associated with its corresponding edge point, like a label.

5.5.4 The Computation of Boundary Addition

The computation of $\partial A \uplus \partial B$ or $\partial A \uplus \partial B^{-1}$ (see (5.14) or (5.15); we will ignore v_s^A, v_s^B, etc. at present) by means of slope diagrams becomes quite straightforward:

1. *Merging of slope diagrams.* Merge the slope diagrams of the two operands (given in Figure 5.10(a)) into a single one.

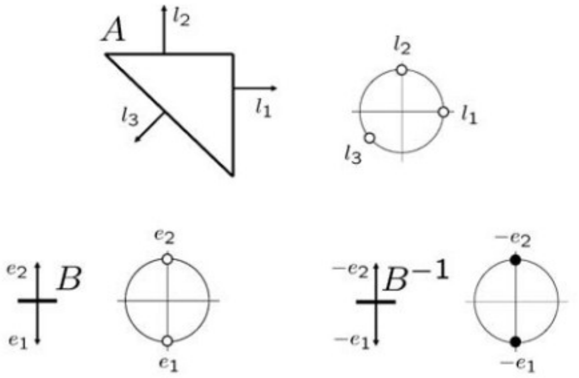

(a) Operands A, B and its additive inverse, and the corresponding slope diagrams

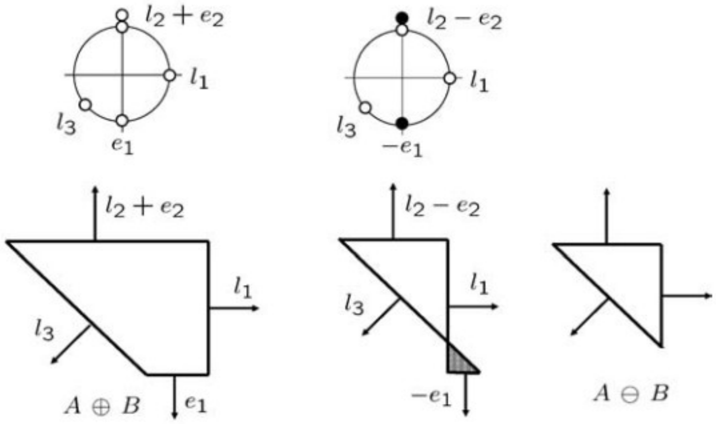

(b) Merged slope diagram of A and B and realization of $\partial A \uplus \partial B$

(c) Merged slope diagram of A, B^{-1} and $\partial A \uplus \partial B^{-1}$, and Minkowski decomposition $A \ominus B$

Figure 5.10 Minkowski addition and decomposition by means of slope diagrams – in this case, $\partial A \uplus \partial B = \partial(A \oplus B)$, but $\partial(A \ominus B) = \mathrm{Pos}(\partial A \uplus \partial B^{-1})$: (a) the operands A, B, and the additive inverse B^{-1}, and the corresponding slope diagrams; (b) the merged slope diagram of A and B and the realization of $\partial A \uplus \partial B$; (c) the merged slope diagram of A, B^{-1}, and $\partial A \uplus \partial B^{-1}$, and Minkowski decomposition $A \ominus B$

2. *Realization of the boundary sum.* From the merged slope diagram, realize the polygon that it represents. The term "realization" means "concatenation" of the edges (that is, joining the end-point of one edge to the start-point of the next edge) in the sequence in which they appear in the merged slope diagram (Figures 5.10(b) or (c)).

Another example of Minkowski addition and decomposition by means of slope diagrams is given in Figure 5.11. The realization/concatenation process may need some further clarification.

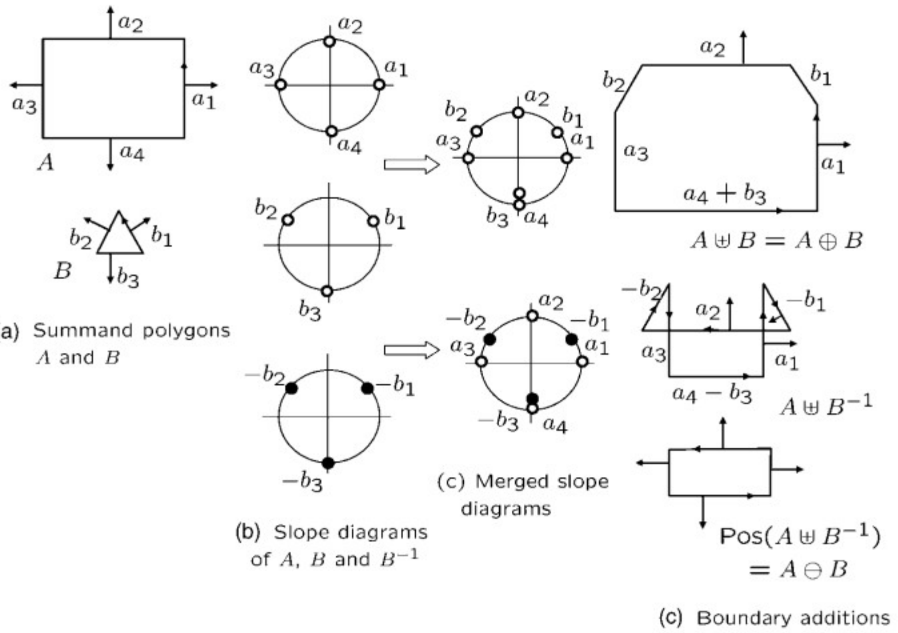

(a) Summand polygons A and B

(b) Slope diagrams of A, B and B^{-1}

(c) Merged slope diagrams

(c) Boundary additions

Figure 5.11 Another example of Minkowski addition and decomposition by means of slope diagrams: (a) summand polygons A and B; (b) the slope diagrams of A, B, and B^{-1}; (c) the merged slope diagrams; (d) the boundary additions

First, see (5.14), concerning Minkowski addition. In the merged slope diagram, two distinct cases may arise: (1) If two edge points of the summands occur at the same position of the unit circle, this means that both of the edges have the same outer normal direction. Therefore, concatenation of these two edges effectively means the addition of the lengths $\iota_i^A + \iota_i^B$ of the edges. (2) If one edge point of a summand lies on a vertex arc of the other summand, this means that the edge of the second summand has zero length at that outer normal direction. The addition of zero length simply involves consideration of the edge of the first summand, and that is precisely what we achieve by concatenation. Next, see (5.15), concerning Minkowski decomposition. Here, everything remains as before, except that some of the edge points (resulting from the B^{-1} polygon) in the merged slope diagram are black; that is, have negative sense. If the sense is negative, we have to subtract the corresponding length of the edge of B from that of A. Clearly, the subtraction of one directed edge from another simply involves reversing the direction of the former and then concatenating with the latter.

This slope diagrammatic approach clearly illustrates the following:

- Just as in real number arithmetic, instead of treating the division m/n as a separate operation, we can consider it as multiplication with the multiplicative inverse $\frac{1}{n}$, in a similar way, the Minkowski decomposition $\partial(A \ominus B)$ operation can also be viewed as addition with the *additive inverse* ∂B^{-1}.

- Just as in the system $(\mathbb{N}, \cdot, /)$, we do not obtain *closure* under the division operation. Therefore, we need to extend the number domain from $\mathbb{N}$ to the set of all positive rational numbers $\mathbb{Q}^+$ for the purpose of closure; and, similarly, to obtain closure under Minkowski decomposition it is necessary to extend the domain $\mathcal{K}$ of the ordinary convex objects – it is done here by introducing the concept of negative shape.
- Both Minkowski addition and decomposition can be reduced to a single operation $\uplus$: Minkowski addition involves boundary addition $\uplus$ of two positive (ordinary) objects, while Minkowski decomposition involves boundary addition of one positive and one negative object. Boundary addition, as we have shown, is essentially just addition of real numbers (representing the lengths of the edges of the operand polygons; those lengths may be positive or negative).
- Just as $\lfloor m/n \rfloor$ is obtained by discarding the fractional part of m/n, similarly $\partial(A \ominus B)$ is obtained by discarding the negative part of $\partial A \uplus \partial B^{-1}$.

5.6 In the Domain of Convex Polyhedra

One of the basic theses of this chapter is that Minkowski operations on three- or higher-dimensional boundary-represented objects in E^d eventually reduce to Minkowski operations on polygons in E^2, which, in turn, boil down to addition of real numbers. At present, we shall consider the case of polyhedral objects, though our approach is general enough to extend beyond three dimensions. The approach is to find relations for convex polyhedra that will be similar to (5.14) and (5.15), and then to resort to the slope diagrammatic technique, as we have done for convex polygons.

5.6.1 Computation by Means of Faces

The boundary of a convex polyhedron can be represented by means of vertices, edges, and facets. Our approach demands that these concepts are defined more precisely.

If, for some $u \neq 0$, we have $H(A, u) < \infty$ (this condition ensures that A is bounded), then the hyperplane

$$L(A, u) = \{\, p \in E^d \mid (p, u) = H(A, u)\}$$

is called the supporting hyperplane of A with *outer normal* u. In E^2, the supporting hyperplane becomes the supporting line of a convex polygon A with outer normal u, while in E^3 it is the supporting plane of a convex polyhedron A.

A face of A with outer normal u, denoted by $F(A, u)$, is then defined as

$$F(A, u) = L(A, u) \cap A.$$

Thus $F(A, u)$ is precisely that set of boundary points of A where the outer normal is either u or parallel to u. Now, considering all the directions of the E^d space as the directions of the outer normal u, the collection of the corresponding faces will describe the entire boundary of A. That is, the entire boundary of A can be described as $\partial A = \bigcup_{u \in S^{d-1}} F(A, u)$.

This notion of $F(A, u)$ is related to our conventional concept of the boundary of an object in terms of vertices, edges, faces, and so on in the following way. If A is a convex d-dimensional object, then $L(A, u)$ is a $(d-1)$-dimensional hyperplane. Therefore, $F(A, u)$ may have dimensions $0, 1, \cdots, (d-1)$. Normally, a face $F(A, u)$ of dimension r ($r = 0, 1, \cdots, (d-1)$) is called a *r-face* of A. A maximal proper face of A – that is, a $(d-1)$-dimensional face – is called a *facet* of A. Clearly, if A is a two-dimensional convex polygon, then $F(A, u)$ is either a 0-face (*vertex*) or an 1-face (*edge*). Since an edge of a polygon is a maximal proper face, it can also be called a facet of the polygon. When A is a three-dimensional convex polyhedron, in addition to being either a vertex or an edge, $F(A, u)$ may also be a 2-face (*facet*). (Note that, for three-dimensional objects, instead of "facet," the term planar "face" is more frequently used.)

From Theorem 5.5, it is not difficult to derive the following result [34, 50].

Theorem 5.8. *Let A and B be two convex polytopes in* E^d. *Then, for every* $u \in S^{d-1}$,

$$F(A \oplus B, u) = F(A, u) \oplus F(B, u). \tag{5.17}$$

Therefore, in terms of faces of the summands, the boundary $\partial(A \oplus B)$ can be expressed as

$$\begin{aligned} \partial(A \oplus B) &= \bigcup_{u \in S^{d-1}} F(A \oplus B, u) \\ &= \bigcup_{u \in S^{d-1}} \left(F(A, u) \oplus F(B, u)\right). \end{aligned} \tag{5.18}$$

Equation (5.18) will be our basis for computing Minkowski operations on convex polyhedra. We shall first argue that it is not necessary to compute $(F(A, u) \oplus F(B, u))$ for every $u \in S^2$ in the three-dimensional space. Just as a convex polygon is completely specified by its oriented edges, similarly a convex polyhedron is completely specified by its oriented facets (planar faces). Therefore, it is sufficient if we compute only the facets of $A \oplus B$, but not every $F(A \oplus B, u)$'s. As is evident from (5.18), the facets of $A \oplus B$ can be obtained in the following way:

1. *Minkowski addition of two facets.* Adding a facet of A to a facet of B; that is, $facet_A \oplus facet_B$.
2. *Minkowski addition of a facet and an edge.* Adding a facet of one of the two summands to an edge of the other; that is, $facet_A \oplus edge_B$ or $edge_A \oplus facet_B$.
3. *Minkowski addition of a facet and a vertex.* Adding a facet of one of the two summands to a vertex of the other; that is, $facet_A \oplus vertex_B$ or $vertex_A \oplus facet_B$.
4. *Minkowski addition of two nonparallel edges.* Adding nonparallel edges of A and B; that is, $edge_A \oplus edge_B$.

Here, the facets, edges, and vertices added in this way lie in supporting planes, with parallel outer normals.

Now, Minkowski addition of two facets that lie in supporting planes having parallel outer normals is equivalent to Minkowski addition of those facets lying in the same plane. That, in turn, is simply Minkowski addition of two convex polygons in E^2. The same is true for Minkowski addition of a facet and an edge, or the addition of two edges.

We depict a typical such addition in Figure 5.12. (In a similar way, we can continue further and show that Minkowski addition of two convex polytopes in E^d reduces to Minkowski addition operations in $E^{d-1}, \ldots,$ and eventually reduces to the Minkowski addition of convex polygons in E^2. This means that it finally reduces to addition of real numbers.)

In computing the Minkowski decomposition $A \ominus B$, exactly as in the polygonal case, the first step is to compute the boundary addition $\partial A \uplus \partial B^{-1}$, and then to determine its positive

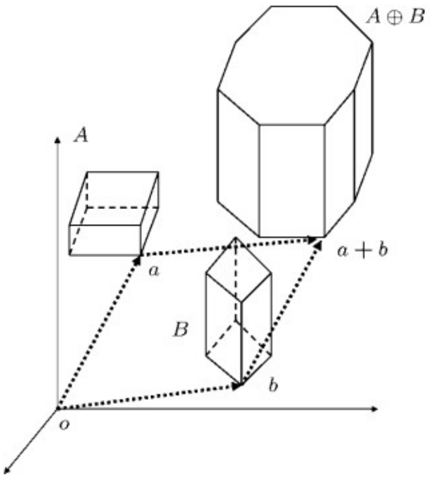

Figure 5.12 The computation of $\bigcup_{u\in S^{d-1}} (F(A,u) \oplus F(B,u))$ for two convex polyhedra A and B in E^3

portion $\mathrm{Pos}(\partial A \uplus \partial B^{-1})$. In terms of the faces of the operand polyhedra, the first step is to compute $\bigcup_{u\in S^{d-1}}(F(A,u) \oplus F(B^{-1},u))$. Therefore, according to our previous discussion, it will reduce to $\mathit{facet}_A \oplus \mathit{facet}_{B^{-1}}$, $\mathit{facet}_A \oplus \mathit{edge}_{B^{-1}}$ or $\mathit{edge}_A \oplus \mathit{facet}_{B^{-1}}$, $\mathit{facet}_A \oplus \mathit{vertex}_{B^{-1}}$ or $\mathit{vertex}_A \oplus \mathit{facet}_{B^{-1}}$, and $\mathit{edge}_A \oplus \mathit{edge}_{B^{-1}}$ (where they lie in supporting planes with parallel outer normals). As we have argued previously, each of these additions can first be reduced to the boundary addition of a positive and a negative convex polygon in E^2, which, in turn, reduces to subtraction of real numbers.

We present an example of a negative polyhedron, B^{-1}, in Figure 5.13(a). In Figure 5.13(b), we show the Minkowski decomposition of two convex polyhedra by means of boundary addition $\uplus$ and, thereafter, the Pos operation.

5.6.2 The Slope Diagram Representation of a Convex Polyhedron

The essential goal of the slope diagram representation of an object is to capture the behavior of the outer normals of the object in an explicit way. The behavior of the directions of the outer normals at various faces (that is, at the facets, edges, and vertices) of a convex polyhedron can be described as follows:

1. At each interior point of a facet of a polyhedron, there can be drawn only one outer normal.
2. At each point of an edge (different from a vertex), it is possible to draw an infinite number of normals filling a plane angle, which will be supplementary to the corresponding interior angle. (In some of the literature, this interior angle is called the *dihedral angle* corresponding to the edge. The exact definition is as follows: the dihedral angle corresponding to an edge is the angle between the planes of its two adjacent facets.)
3. At each vertex, it is possible to draw infinitely many outer normals that fill a solid angle (interpolating between the incident facet normals).

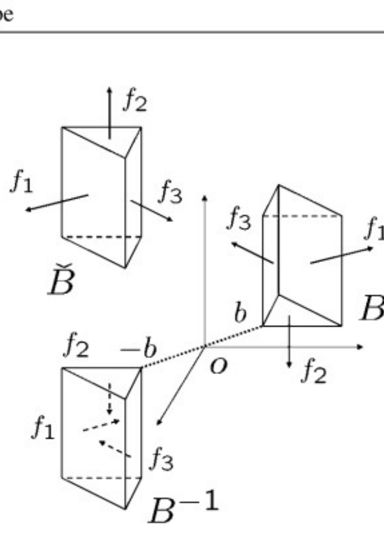
f_2
f_1
f_3
$\breve{B}$
f_3
f_1
b
B
f_2
$-b$
o
f_2
f_1
f_3
B^{-1}

(a) Polyhedron B and corresponding B^{-1} and $\breve{B}$ polyhedra

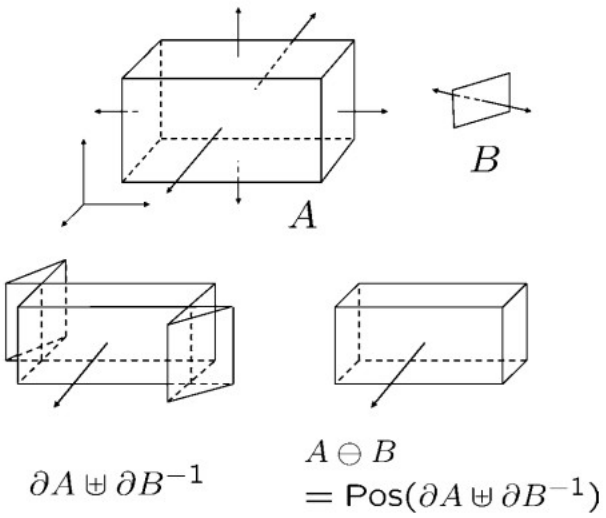
B
A
$\partial A \uplus \partial B^{-1}$
$A \ominus B$
$= \mathrm{Pos}(\partial A \uplus \partial B^{-1})$

(b) The input A and B are convex polyhedra

Figure 5.13 (a) The geometric representation of a negative polyhedron B^{-1} – polyhedron B and the corresponding B^{-1} and $\breve{B}$ polyhedra; (b) the computation of $A \ominus B$ using the boundary addition operation – the input A and B are convex polyhedra

For the slope diagram of a polyhedron, we have to start with a unit sphere S^2.

(a) *Facet representation.* As in the polygonal case, each facet can be represented by the corresponding point on the unit sphere, which is referred to as a *facet point*.

(b) *Edge representation.* Each edge of the polyhedron, we claim, can be represented by the arc of the great circle joining the two facet points corresponding to the two adjacent facets of the edge. We call such an arc an *edge arc* of the polyhedron.

Note: The intersection of the surface of a sphere by a plane is called a great circle if the plane passes through the center of the sphere. Clearly, only one great circle can be drawn through two given points on the surface of the sphere, except when the points are the extremities of a diameter of the sphere. By the "arc of a great circle," we generally mean the shorter of the two arcs joining the two points.

(c) *Vertex representation.* According to the scheme, it easy to see that the directions of the outer normals at any vertex v of the polyhedron will be represented by a region on the unit sphere. This region is bounded by the edge arcs corresponding to the edges incident at v, and the vertices of this region are the facet points corresponding to the facets of the polyhedron incident at v. We call this a *vertex region*.

(d) To denote the sense of an outer normal and the length of an edge, we use the same conventions as adopted for polygons.

In Figure 5.14, we show the slope diagram representations of two convex polyhedra. The heavy lines and solid black dots in the diagrams indicate the negative sense of the outer normals.

Note: In Mount and Silverman [75] and Guibas and Seidel [36], we find notions that are similar in some ways to the idea of the slope diagram of a convex polyhedron. However, these authors did not take into account the notion of "sense" of a outer normal, which is crucial in our approach.

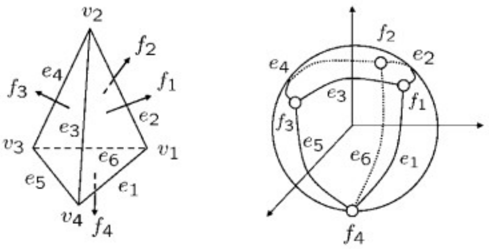

(a) Positive convex polyhedron

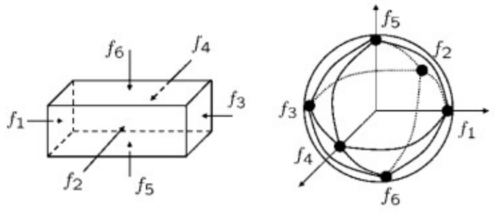

(b) Negative convex polyhedron

Figure 5.14 Two convex polyhedra and their slope diagram representations (not to scale): (a) a positive convex polyhedron; (b) a negative convex polyhedron

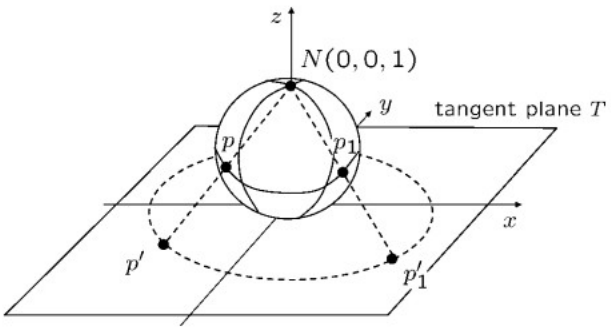

Figure 5.15 Transforming a three-dimensional slope diagram by stereographic projection on two dimensions

Since the slope diagram of a polyhedron is a three-dimensional figure, it may not be a very convenient representation to deal with for the purpose of intuitive understanding and visualization. Therefore, we transform this representation into an equivalent two-dimensional form by means of the *stereographic projection.* As shown in Figure 5.15, the stereographic projection is a projection of a sphere from one of the points N onto the plane T tangent to the sphere at the diametrically opposite point N'. The point N is called the *projection center.*

The transformation equations for the stereographic projection are given in many standard texts, such as [88]. Consider a unit sphere whose center o is at the origin, and the projection center N for which is located on the $o - z$ axis (see Figure 5.15). In this case, the point N has the coordinates (0, 0, 1) and the projection plane T is the plane $z = -1$. Let the point $p(x, y, z)$ of the sphere be projected stereographically into the point $p'(x', y', -1)$ of the plane T. Then the coordinates of the point p, corresponding to the point p', are equal to

$$x = \frac{4x'}{x'^2 + y'^2 + 4},$$

$$y = \frac{4y'}{x'^2 + y'^2 + 4},$$

$$z = \frac{x'^2 + y'^2 - 4}{x'^2 + y'^2 + 4}.$$

The inverse mapping – that is, the coordinates of the point p', corresponding to the point p – is equal to

$$x' = \frac{2x}{1 - z}, \quad y' = \frac{2y}{1 - z}, \quad z' = -1.$$

From the above equations, we find that if p is the projection center itself – that is, $p = (0, 0, 1)$ – then we have difficulty in computing x' and y'. This difficulty is circumvented by extending the plane T by an "ideal" point, which is called the *point at infinity*.

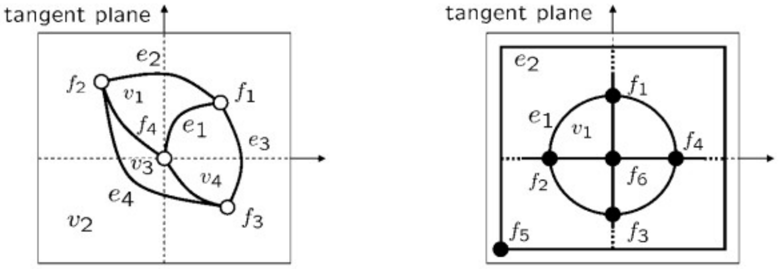

(a) Projected slope diagram of a positive tetrahedron

(the regions denote vertices, the arcs denote edges, and the vertex points denote facets).

(b) Projected slope diagram of a negative parallele-piped

(black dots denote facets ; the faset point f_5 is projected on the point at infinity).

Figure 5.16 Stereographic projections of the slope diagrams of the convex polyhedra depicted in Figure 5.14 (not to scale): (a) the projected slope diagram of a positive tetrahedron (the regions denote vertices, the arcs denote edges, and the vertex points denote facets); (b) the projected slope diagram of a negative parallelepiped (black dots denote facets – the facet point f_5 is projected on the point at infinity)

In the stereographic projection of a slope diagram, any facet point at (x_1, y_1, z_1) will be projected onto the point (x_1', y_1') of the T-plane; the point (x_1', y_1') can be determined using the above equations. Any edge arc will be projected either as a straight line or a circular arc. An edge arc joining two facet points (x_1, y_1, z_1) and (x_2, y_2, z_2) will be projected on the plane T as $(x_1y_2 - x_2y_1)(x'^2 + y'^2) + 4(y_1z_2 - y_2z_1)x' + 4(z_1x_2 - z_2x_1)y' - 4(x_1y_2 - x_2y_1) = 0$. This is the equation of a circular arc – unless $x_1y_2 = x_2y_1$, when it degenerates into a straight-line segment.

In Figure 5.16, we present the stereographic projections of the slope diagrams shown in Figure 5.14.

5.6.3 Computation by Means of Slope Diagrams

Assume that two convex polyhedra, A and B, are given in the form of their slope diagram representations. Exactly as in the polygonal case, the boundary addition computation involves two steps: (a) merging of slope diagrams of the operands, and (b) realization of the boundary sum from the merged slope diagram.

If we *merge* these two slope diagrams (that is, overlay both of the slope diagrams on the same unit sphere S^2), we can immediately identify the corresponding $F(A, u)'$s and $F(B, u)'$s (or $F(B^{-1}, u)'$s in the case of decomposition) that have the same outer normal direction u, because they will occupy the same positions on the sphere. This means that *wherever there are intersections between two slope diagrams, the corresponding faces of the operands need to be added (or subtracted).*

To *realize* the boundary sum $\partial A \uplus \partial B$ or $\partial A \uplus \partial B^{-1}$ from the merged slope diagram, we have to identify only the facet points of the boundary sum. Any facet point will be created in one of the following ways: (a) the intersection of a facet point of one operand with that of the

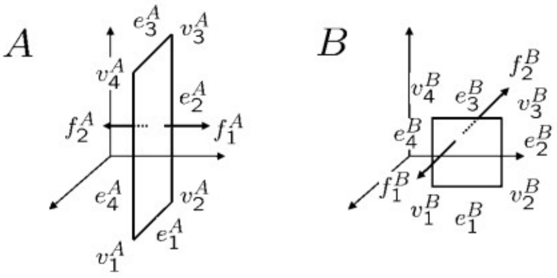

A
e_3^A
v_3^A
v_4^A
e_2^A
f_2^A
f_1^A
e_4^A
v_2^A
e_1^A
v_1^A
B
f_2^B
v_4^B
e_3^B
v_3^B
e_4^B
e_2^B
f_1^B
v_2^B
v_1^B
e_1^B

(a) Operands A and B ; facets edges, and vertices are denoted by f's, e's, v's respectively.

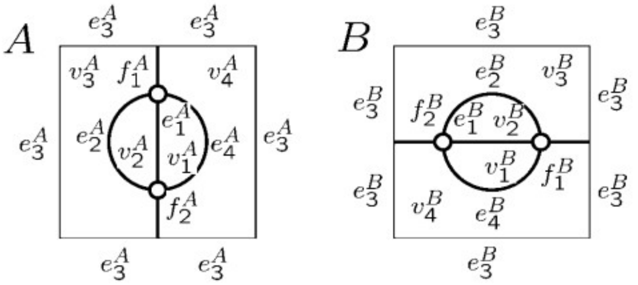

A
e_3^A
e_3^A
v_3^A
f_1^A
v_4^A
e_3^A
e_2^A
e_1^A
v_2^A
v_1^A
e_4^A
e_3^A
f_2^A
e_3^A
e_3^A
B
e_3^B
e_2^B
v_3^B
e_3^B
e_3^B
f_2^B
e_1^B
v_2^B
v_1^B
f_1^B
e_3^B
e_3^B
v_4^B
e_4^B
e_3^B

(b) Slope diagram representations of the operands.

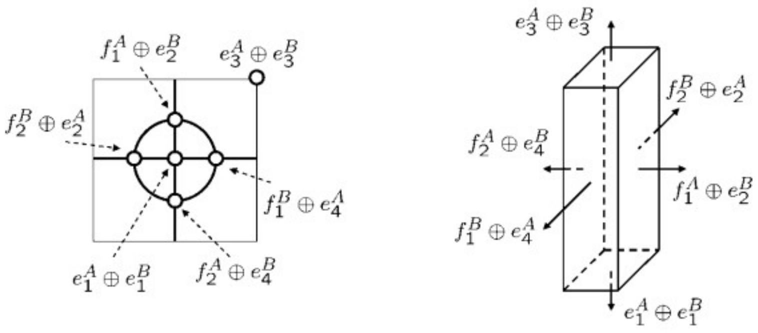

$f_1^A \oplus e_2^B$
$e_3^A \oplus e_3^B$
$f_2^B \oplus e_2^A$
$f_1^B \oplus e_4^A$
$e_1^A \oplus e_1^B$
$f_2^A \oplus e_4^B$
$e_3^A \oplus e_3^B$
$f_2^B \oplus e_2^A$
$f_2^A \oplus e_4^B$
$f_1^A \oplus e_2^B$
$f_1^B \oplus e_4^A$
$e_1^A \oplus e_1^B$

(c) Merged slope diagram: two new face points are created by non-parallel edges, i.e., $e_1^A \oplus e_1^B$ and $e_3^A \oplus e_3^B$.

(d) $\partial A \uplus \partial B$ since both A and B are positive convex polyhedra, and $\partial(A \oplus B) = \partial A \uplus \partial B$.

Figure 5.17 The Minkowski addition of two polyhedra using their slope diagram representations: (a) the operands A and B – the facets, edges, and vertices are denoted by the f's, e's, and v's, respectively; (b) the slope diagram representations of the operands; (c) the merged slope diagram, in which two new faces are created by nonparallel edges – that is, $e_1^A \oplus e_1^B$ and $e_3^A \ominus e_3^B$; (d) $\partial A \uplus \partial B$, since both A and B are convex polyhedra, and $\partial(A \oplus B) = \partial A \uplus \partial B$

other, which means *addition of two facets*; (b) the intersection of a facet point of one operand with an edge arc of the other, which means *addition of a facet and an edge*; (c) the intersection of a facet point of one operand with a vertex region of the other, which means, *addition of a facet and a vertex*; or (d) the intersection of an edge arc of one operand with a nonparallel edge arc of the other, which means *addition of two nonparallel edges*.

In the case of Minkowski decomposition, if $\partial A \uplus \partial B^{-1}$ turns out to be a self-crossing polyhedron (as shown in Figure 5.13(b)), we have to discard the negative portion and consider only the positive portion.

In Figure 5.17, we demonstrate computation by means of slope diagrams through an example of addition of a rectangular plane (A) that is parallel to the xz-plane to another rectangular plane (B) parallel to the yz-plane. The resulting sum shape $A \oplus B$ will be a rectangular parallelepiped.

11.04.2020

A) 5 Coordinatising a projective plane

Pag. 99 (112)

6 Non-Desarguesian Planes

Pag. 110 (123)

Pag. 112 (125)

Example 6.4

Prove that if a collineation σ of a projective plane π is such ~~that~~

1. some line l is fixed pointwise;
2. two points $P, Q \in l$ are fixed,

then $\sigma = 1$.

B) Books_67_2020/
Mathematics of Shape Description/
Pijush K. Ghosh_L. Mathematics of Shape Description...pdf

Pag. 77 (96)

3. Algebraic Structures for Shape Description

5. Arithmetics of Geometric Shape
Pag. 165 (184)

data 11.04.2020
S T Q Q S [✓ S] D

5.3 Boundary Representation by Support Functions for Morphological Operations

5.3.1 The Support Function Representation

Pag. 169(188)

Convex polytope A in E^d is defined in terms of the supporting function and the supporting hyperplane of the polytope.

5.3.2 The Support Function is a Signed Distance

Pag. 170(189)

Figure 5.1

Figure 5.2

Figure 5.3 a) two-dimensional polygon
b) three-dimensional polyhedron

① Books_67_2020/
Visual Quantum Mechanics/
Bernd Thaller- Advanced Visual Quantum Mechanics (2004...).pdf

Pag. 38(49) Figure 1.9

Possible trajectories of classical particles on a sphere. For all trajectories, the angular-momentum vector has $|L| = \sqrt{l(l+1)}$, with $l = 6$ and $L_z = 6$.

)*+,)))-) #% . " 2 *": "
, %@ 3 =20B- 3"
(#! %5 9 5

Here, α is the angle between the directions (ϑ, φ) and (ϑ', φ'). Denoting the corresponding unit vectors by $\boldsymbol{\omega}$ and $\boldsymbol{\omega}'$, we have $\cos\alpha = \boldsymbol{\omega} \cdot \boldsymbol{\omega}'$.

In Cartesian coordinates (x, y, z), with $r = (x^2+y^2+z^2)^{1/2}$, the spherical harmonics are (for $m \geq 0$)

$$Y_\ell^m(x, y, z) = \sqrt{\frac{2\ell+1}{4\pi}\frac{(\ell-m)!}{(\ell+m)!}} \left(\frac{x+\mathrm{i}y}{\sqrt{x^2+y^2}}\right)^m P_\ell^m\left(\frac{z}{r}\right). \tag{1.112}$$

For example,

$$Y_0^0 = \frac{1}{\sqrt{4\pi}}, \quad Y_1^0 = \sqrt{\frac{3}{4\pi}}\,\frac{z}{r}, \quad Y_1^{\pm 1} = \mp\sqrt{\frac{3}{8\pi}}\,\frac{x \pm \mathrm{i}y}{r}. \tag{1.113}$$

1.7. Particle on a Sphere

1.7.1. Classical particle on a sphere

Here, we consider a quantum mechanical system that is only able to perform a rotational motion. Imagine a particle with mass[4] m whose motion is restricted to the surface of a sphere. There are no other forces. This particle has two degrees of freedom, its position on the sphere is given by two angular coordinates, the longitude and the latitude, described by the angles ϑ and φ. A classical particle subject to these constraints will move along the geodesics of the sphere (great circles) with constant (angular) velocity. Its kinetic energy is given by

$$E = \frac{1}{2}\mathrm{m}v^2 = \frac{1}{2}\mathrm{m}(r\omega)^2 = \frac{1}{2}I\omega^2, \tag{1.114}$$

where we have introduced the angular speed $\omega = v/r$ and the moment of inertia

$$I = \mathrm{m}r^2. \tag{1.115}$$

The angular momentum of a particle moving with velocity v on a sphere of radius r is $L = \mathrm{m}vr$, or $v = L/\mathrm{m}r$. Insert this into the expression for the energy to obtain

$$E = \frac{L^2}{2\mathrm{m}r^2} = \frac{L^2}{2I}. \tag{1.116}$$

CD 1.2.1, CD 1.5.4, and CD 1.6.3 are all visualizations of the classical motion at a fixed distance from the coordinate origin.

---

[4] It is general custom to denote the mass of the particle and the eigenvalue of L_3 by the same letter. Usually there is little danger of confusion. To be on the safe side, we denote the mass by the roman letter m and use the italic letter m for the eigenvalue of L_3.

1.7.2. The rigid rotator

The particle on a sphere is a mathematical model for the rigid rotator. Consider two point-like particles of mass m_1 and m_2 that are connected by a weightless, rigid rod of length r. Denote by $\mathbf{x} = \mathbf{x}_2 - \mathbf{x}_1$ the vector pointing from m_1 to m_2. The position of the center of mass is given by

$$\mathbf{X} = \frac{\mathrm{m}_1\mathbf{x}_1 + \mathrm{m}_2\mathbf{x}_2}{\mathrm{m}_1 + \mathrm{m}_2} \tag{1.117}$$

The distances of the particles from the center of mass are therefore

$$r_1 = |\mathbf{x}_1 - \mathbf{X}| = \frac{\mathrm{m}_2\, r}{\mathrm{m}_1 + \mathrm{m}_2}, \quad r_2 = |\mathbf{x}_2 - \mathbf{X}| = \frac{\mathrm{m}_1\, r}{\mathrm{m}_1 + \mathrm{m}_2}. \tag{1.118}$$

If we are only interested in the internal motion of the two-particle system, we can choose a coordinate system that has its origin at the center of mass. In this coordinate system, the moment of inertia of the rotator is thus

$$I = \mathrm{m}_1 r_1^2 + \mathrm{m}_2 r_2^2 = \frac{\mathrm{m}_1\mathrm{m}_2}{\mathrm{m}_1 + \mathrm{m}_2}\, r^2. \tag{1.119}$$

This is precisely the moment of inertia of a particle with mass

$$\mu = \frac{\mathrm{m}_1\mathrm{m}_2}{\mathrm{m}_1 + \mathrm{m}_2} \tag{1.120}$$

at a distance r from the origin. Thus, we can replace the two-particle system with an effective one-particle system. This particle has the so-called *reduced mass* μ and the fixed distance r from the origin, that is, it is a particle on a sphere with radius r.

This is a good model for a diatomic molecule. It consists of two atoms that can oscillate along the line connecting their centers and rotate around the center of mass. The vibrational motion is much faster than the rotational motion. So, the vibrations belong to much higher energies. Quantum mechanically, oscillation states have a quantized energy. As long as the diatomic molecule is in its vibrational ground state, it is a rigid rotator.

1.7.3. Transition to quantum mechanics

In order to define the quantum mechanical Hamiltonian for a particle that is constrained to the surface of a sphere, we start with the classical expression (1.116) for the energy. From this, the Hamiltonian operator of the rigid rotator is obtained by replacing L^2 with the quantum mechanical angular-momentum operator. Note that r, the radius of the sphere, is treated as a fixed parameter. Using the expression (1.91) for the angular momentum in spherical coordinates, we arrive at

$$\hat{H} = \frac{1}{2\mathrm{m}\, r^2}\, \hat{L}^2 = \frac{1}{2I}\, \hat{L}^2. \tag{1.121}$$

The operator $\hat{L}^2$ has a discrete spectrum of eigenvalues, therefore the same is true for the energy of the rotator.

Eigenvalues of the rigid rotator:

A particle with mass m on a sphere with radius r can only have the energies

$$E_\ell = \frac{\hbar^2}{2\mathrm{m}r^2}\,\ell(\ell+1), \quad \ell = 0, 1, 2, 3, \ldots. \tag{1.122}$$

Each eigenvalue E_ℓ has the multiplicity $2\ell+1$, that is, there are $2\ell+1$ orthogonal wave functions all belonging to the eigenvalue E_ℓ:

$$\psi_{\ell,m}(\vartheta,\varphi) = \frac{1}{r}\,Y_\ell^m(\vartheta,\varphi), \quad m = -\ell, -\ell+1, \ldots, \ell. \tag{1.123}$$

It follows from (1.109) that

$$\int_{S_r^2} \bigl|\psi_{\ell,m}(\vartheta,\varphi)\bigr|^2\, r^2\, \sin\vartheta\, d\vartheta\, d\varphi = 1. \tag{1.124}$$

The factor $1/r$ in (1.123) thus guarantees that the eigenfunctions are normalized on the sphere with radius r. We conclude that the eigenfunctions form an orthonormal basis in the Hilbert space $L^2(S_r^2)$ of square integrable functions on the sphere of radius r.

We want to stress that $\hat{L}^2$ is part of any single-particle Hamiltonian, because it appears in the angular part of the kinetic energy in spherical coordinates (see (1.150) below). For a particle constrained to the surface of a sphere, the operator $\hat{L}^2$ plays the same role as the operator $-\hbar^2\Delta$ for a particle in $\mathbb{R}^3$: It is proportional to the kinetic energy and it generates the free time evolution. (In fact, $\hat{L}^2/\hbar^2$ is the so-called Laplace-Beltrami operator of the sphere.) The action of the free time evolution $\exp(-\mathrm{i}\hat{L}^2 t/2I)$ on a sphere will be discussed next.

1.7.4. Dynamics of the rigid rotator

The time evolution of an eigenfunction of the rigid rotator Hamiltonian is rather trivial. If $\psi_{\ell,m}(\vartheta,\varphi)$ is an eigenfunction belonging to the eigenvalue E_ℓ, then the function

$$\psi_{\ell,m}(t,\vartheta,\varphi) = \exp(-\mathrm{i}E_\ell\, t)\,\psi_{\ell,m}(\vartheta,\varphi) \tag{1.125}$$

is a solution of the time-dependent Schrödinger equation. The ground state of the rigid rotator has energy $E_0 = 0$, and hence the corresponding wave function does not depend on time at all.

The fact that the Schrödinger equation is linear means that we can consider linear combinations of eigenfunctions. For example, for the initial function $\psi_0(\vartheta,\varphi) = \sum_{\ell,m} c_{\ell,m}\,\psi_{\ell,m}(\vartheta,\varphi)$, the time evolution is

$$\psi(t,\vartheta,\varphi) = \sum_{\ell,m} c_{\ell,m}\,\exp(-\mathrm{i}E_\ell\,t)\,\psi_{\ell,m}(\vartheta,\varphi) \tag{1.126}$$

and this clearly gives us the solution for any initial function that is square integrable on the sphere, because the set of functions $\psi_{\ell,m}$ forms an orthonormal basis in $L^2(S^2_r)$.

Given any initial function ψ_0, the expansion coefficient $c_{\ell,m}$ can be found by an integration,

$$c_{\ell,m} = \int_{S^2_r} \overline{\psi_{\ell,m}(\vartheta,\varphi)}\,\psi_0(\vartheta,\varphi)\,r^2\,d\Omega \tag{1.127}$$

(with $d\Omega = \sin\vartheta\,d\vartheta\,d\varphi$). It is remarkable that all nonzero energies $E_\ell = \ell(\ell+1)/2I$ are integer multiples of $E_1 = 1/I$, hence all time-dependent factors $\exp(-\mathrm{i}E_\ell\,t)$ have the same basic time period. Hence, any state of the rigid rotator is periodic in time.

For the unit sphere ($r = 1$), we summarize our results in the following box.

Time evolution of the rigid rotator:

For a particle with mass m = 1 on the unit sphere S^2 the time evolution of any square-integrable initial function $\psi_0(\vartheta,\varphi)$ is given by the formula

$$\psi(t,\vartheta,\varphi) = \sum_{\ell,m} c_{\ell,m}\,\exp(-\mathrm{i}E_\ell\,t)\,Y_\ell^m(\vartheta,\varphi) \tag{1.128}$$

with

$$c_{\ell,m} = \int_{S^2} \overline{Y_\ell^m(\vartheta,\varphi)}\,\psi_0(\vartheta,\varphi)\,d\Omega. \tag{1.129}$$

The time evolution of the wave function is periodic in time,

$$\psi(t+T,\vartheta,\varphi) = \psi(t,\vartheta,\varphi) \quad \text{with period} \quad T = 2\pi. \tag{1.130}$$

CD 1.9–CD 1.12 is a collection of several movies showing various time-dependent states of the rigid rotator. CD 1.13 visualizes the time evolution of initially well-localized (Gaussian) states. In CD 1.14, the initial state has roughly the shape of the letter Ψ. This illustrates that we can indeed compute the time evolution of any square-integrable initial function on the sphere.

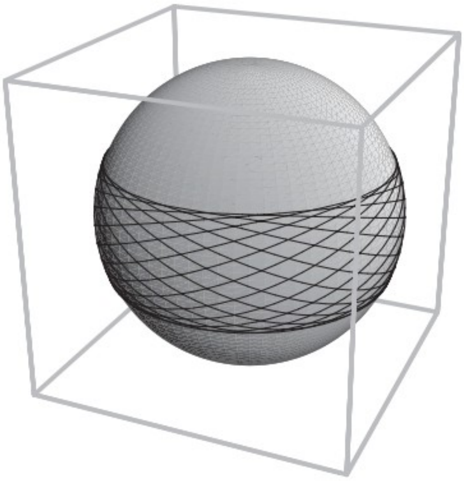

FIGURE 1.9. Possible trajectories of classical particles on a sphere. For all trajectories, the angular-momentum vector has $|\mathbf{L}| = \sqrt{\ell(\ell+1)}$, with $\ell = 6$ and $L_3 = 6$.

1.8. Quantization on a Sphere

1.8.1. Comparison of classical and quantum probability densities

Figure 1.9 shows a set of classical trajectories of particles for which the angular momenta have a fixed absolute value $|\mathbf{L}| = L$ and a fixed value of the third component L_3. The admissible positions of classical particles with these angular momenta form the classically allowed region on the sphere. This region is determined by the polar angle of the classical angular-momentum vector. We denote the polar angle of $\mathbf{L}$ by θ in order to distinguish it from the polar angle ϑ of the particle's position vector, see Figure 1.10.

Assuming $L_3 = m$, we find that the polar angle of $\mathbf{L}$ is

$$\theta = \arccos\left(\frac{m}{L}\right) \tag{1.131}$$

and hence the classically allowed region for the position of the particle is between the polar angles

$$\vartheta_{\min} = \pi/2 - \theta, \qquad \vartheta_{\max} = \pi/2 + \theta. \tag{1.132}$$

All the classical trajectories are grand circles of the sphere. The particles move on these circles with a constant angular velocity. When a particle circles around the sphere, the polar angle ϑ performs a periodic motion

' #
13" ")*+,)))-) # 0" " A
9 !@ 3 3"
)*+,)))-) # A 7 & 3"
(5 # # 9 / A

3

Force between Nucleons

3.1 INTRODUCTION

Considering the mass of a nucleus proportional to the number of nucleons in it and the radius equal to $r_0A^{1/3}$, it is easy to calculate the density for the nuclear matter of a spherical nucleus to be of the order of 10^{17} kg/m^3. In addition, the density of the densest metal, mercury, is of the order of 10^{10} kg/m^3. The very large difference, of the order of 10^7, between the density of nuclear matter and the ordinary matter indicates two facts: (1) the ordinary matter contains large volumes of empty spaces, (2) the force responsible for the binding nucleons inside a nucleus is very much stronger than and may be very different from the forces responsible for the binding of ordinary matter.

3.2 INADEQUACY OF CLASSICAL FORCES

On the first point, it may be mentioned that the ordinary matter is made of molecules and atoms. There is large vacuum in an atom around the nucleus in which electrons revolve and the intra-atomic spaces in a molecule are also empty. With regard to the second point, the forces that bind atoms and molecules in ordinary matter are largely electrical forces and to a much lesser extent gravitational and magnetic forces. It may be remembered that the range, the distance up to which the effect of the force can be felt, of all the three forces (also called classic forces) extends up to very large distances essentially up to infinity. Scattering and other experiments have indicated that the range of the force responsible for the binding of nucleons in nucleus is very small of the order of 2×10^{-15} m or 2 F. Very short-range nuclear force is a major point of difference between the classical and nuclear forces. Another difference is the manner in which atomic electrons are bound in an atom and the binding of nucleons in a nucleus. In an atom, the central positively charged nucleus produces an electric field spreading up to a large distance and electrons, relatively far away from the centre of force, are held in stationary orbits. Electrons are held at relatively weaker part of the central electric field, which is apparent from the fact that the binding energies of outer (valence) electrons are only of the order of few electron volts. The inner electrons in an atom are also not very strongly bound, their binding energies being of the order of few keV. By contrast, in a nucleus there is no source that may produce a central field; nucleons exert force on each other to bind themselves strongly, B/A of the order of 8 MeV. Such a large binding force between two nucleons cannot be provided by classical forces, as the gravitational attraction between two nucleons (a proton and a neutron) at a separation of 2 F may give a potential energy which is at least 10^{36} orders of magnitude smaller than the binding energy of nucleons. There is no question of electrostatic force of attraction between nucleons as neutron has no electric charge and protons will repel each other. A force of magnetic nature may come into play between two nucleons because of their magnetic dipole moments. However, this force will have opposite nature when the two nucleons have parallel and anti-parallel spins. Also, critical analysis of the data on the separation energies of the last nucleon from various nuclei has indicated that the nuclear force between nucleons is attractive for both the anti-parallel and parallel spin orientations. In view of this, classical forces are totally inadequate to provide the strong binding to nucleons in a nucleus.

One way to understand nuclear structure is to start with the force between two nucleons, find a potential for the nucleon–nucleon interaction, develop a structure for the nucleon consistent with the potential, and extend it to construct the nuclear structure. The first step in this approach is to find the properties of nucleon–nucleon force.

Exercise p-3.1: Calculate the gravitational potential energy between two protons at 2 F apart.

Exercise p-3.2: Calculate the electrical potential energy between two protons held 2 F apart.

3.3 TWO-BODY NUCLEON–NUCLEON FORCE

If it is assumed that nucleons exert force on each other in pairs, then the following nucleon pairs are possible: (n–n), (p–p), and (n–p). As both neutron and the proton have intrinsic spins S_n and S_p $(=\frac{1}{2}\hbar)$, there are two possibilities for the spin orientations of the two nucleons of the pair. The first possibility is that the spins are parallel making the total angular momentum (j) of the pair 1 $\hbar$ with three possible values of $m_j = 1, 0, -1$. The parallel spin state is, therefore, called the *triplet state* and is denoted by superscript 3. The second possibility is that the spins of the nucleons in the pair are anti-parallel, making the j of the pair zero, and only single possible value for $m_j = 0$. The system with anti-parallel spins is called *singlet state* and is denoted by superscript 1. In view of this, there are six possible states of a two-nucleon system: ^{1}(n–n), ^{1}(p–p), ^{1}(n–p), ^{3}(n–n), ^{3}(p–p), ^{3}(n–p). States ^{3}(n–n) and ^{3}(p–p) are disallowed by Pauli exclusion principle, which states that no two identical fermions can have identical quantum numbers. It may, however, be noted that ^{3}(n–p) state is allowed. (Can you tell which quantum number is different for neutron and proton?) As such only four pairs, three singlet, and one triplet states are left for consideration. The forces corresponding to these states may be denoted by $F_{1(n-n)}$, $F_{1(p-p)}$, $F_{1(n-p)}$, and $F_{3(n-p)}$. A word of caution about $F_{1(p-p)}$ is that it is the specific nuclear force between two protons and does not contain the Coulomb repulsive force, which is in addition to the nuclear force but of very small magnitude as compared with it.

Information about the nucleon–nucleon (n–n) force may be obtained from (1) the systematic of the separation energies of last nucleon in light nuclei, (2) level scheme of mirror nuclei, (3) study of the properties of deuteron, (4) nucleon–nucleon scattering data, and the (5) constant density of nuclear matter.

3.3.1 Separation Energy of the Last Nucleon in Light Nuclei

Energy required to take out the outermost neutron (or proton) from the nucleus is called the separation energy of last neutron (or proton) and is denoted by S_n (or S_p).

$$S_n(Z, N) = [M(Z, N-1) + M_n - M(Z, N)]c^2 \quad \text{and} \quad S_p(Z, N) = [M(Z-1, N) + M_p - M(Z, N)]c^2$$

Here, M represents the mass of the nucleus having neutrons and protons as specified in the bracket, M_n and M_p, respectively, the neutron and the proton masses. The multiplication with c^2 is done to convert mass into energy. The c^2 term can be dropped if masses are given in the unit of energy (MeV) and the nuclear mass may be replaced by the corresponding atomic mass without much error as electron mass ultimately cancels out. Replacement of nuclear masses by atomic masses only brings in some negligible error due to the difference in the electron binding energies in the two atoms. Alternatively, S_n and S_p can also be given in terms of the binding energies of the two nuclei.

$$S_n(Z, N) = B(Z, N) - B(Z, N-1) \text{ and}$$
$$S_p(Z, N) = B(Z, N) - B(Z, N-1)$$

To confine a particle within a given space, the de Broglie wavelength of the particle must be of the order of the size of the space. This condition puts a limit on the minimum energy of the particle for confinement in the limited space. For example, to confine a nucleon inside a nucleus of size of the order of 10^{-14} m, the minimum kinetic energy T of the nucleon must be of the order of 20 MeV. This implies that inside the nucleus a nucleon is in a potential well of depth V larger than 20 MeV as shown in figure 3.1. The separation energy $S = (V - T)$. The depth of the potential V must not be less than ≈30 MeV, considering the average binding energy of nucleon to be around 8 MeV.

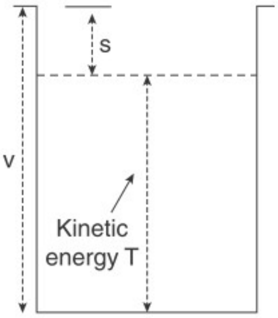

Figure 3.1 *Nuclear potential well and the nucleon separation energy*

An analogous quantity in atomic case is the binding energy of the last valance electron. The binding energy of valance electron in an atom shows cyclic variation. This is attributed to the electronic shell closure. Once a given shell is completely filled, the binding of the next electron becomes very weak. Similar cyclic variation is exhibited by S_n and S_p. This and some other systematic trends observed in stable nuclei, such as the extra stability associated with nuclei having the number of neutrons or/and protons equal to the magic numbers 2, 8, 20, 28, 50, 82, 126, and so on, indicates that nucleons inside the nucleus move in some sort of shell structure. Shell structure in atomic case is the outcome of confining particle in a central field. In case of the nucleus, there is no central field as such. To explain the apparent shell structure, it is assumed that all the nucleons inside a nucleus produce a mean central field in which each nucleon moves independently. This is the basis of the independent or single-particle model of the nucleus. The motion of nucleons in single particle approach may be grouped into states of different angular moment l. The experimental data on S_n and S_p states the following:

1. Only nucleons in the same l state (have no relative angular momentum relative to each other with even parity) exert force on each other. Interaction between nucleons in states of different orbital angular momentum is insignificant.
2. In the case of lightest nuclei, ${}^2_1$H, ${}^3_1$H, and ${}^4_2$He, the S_n increases from 2.2 to 5.4 to ≈20 MeV. Similarly, for ${}^6_3$Li, ${}^7_3$Li, and ${}^8_4$Be, S_n increases from 5.4 to 7.2 to ≈18.1 MeV. Similar trends are also observed for S_p. For example, for ${}^2_1$H, ${}^3_2$He, and ${}^4_2$He, S_p increases from 2.2 to 5.2 to 20 MeV. Systematic observations of the above-mentioned type shows that in light

nuclei neutron is less bound in case N is odd, neutron binding increases if the number of neutrons become even and it becomes maximum when the both number of neutron and proton are even. Similarly, S_p is less for odd Z, nucleus is larger in even–odd nucleus and attains maximum value for even–even nuclei. The tight binding of the last neutron or proton when the number of neutrons or protons in the nucleus is even provided the concept of some additional energy associated with the pair of the same kind. This is called the pairing energy and is denoted by δ_n or δ_p, respectively, for the neutron and the protons. Neutron and proton pairing energies are defined as

$$\delta_n = \frac{1}{2}[S_n^{(\text{even-}N)} - S_n^{(\text{odd-}N)}] \quad \text{and} \quad \delta_p = \frac{1}{2}[S_p^{(\text{even-}P)} - S_p^{(\text{odd-}P)}]$$

The magnitudes of S_n and S_p in odd-N and even-Z, and in even-N and odd-Z light nuclei are almost equal. In addition, the magnitudes of δ_n and δ_p are also nearly equal in light nuclei. In the case of heavy nuclei, similar trend in S_n and S_p values have been observed, particularly for nuclei near magic numbers, but they are less pronounced.

These close similarities in separation energies and pairing energies for neutrons and protons suggest that fundamentally the force between any pair of nucleons is independent of the charge.

Solved example S-3.1

Calculate the minimum kinetic energy of a proton that is confined in a nucleus of $^{27}{}_{13}\text{Al}$. Consider the value of r_0 as 1.2 F.

Solution. The radius of the aluminium nucleus

$$r = r_0(27)^{\frac{1}{3}} = 3.6 \times 10^{-15} \text{ m}.$$

Considering the de Broglie wavelength λ of the proton to be equal

$$\lambda = 2r = 2 \times 3.6 \times 10^{-15} \text{ m}.$$

Let M_p denote the mass of the proton and E be its kinetic energy. The momentum p of the proton is related to its energy E by the expression $p = \sqrt{2M_pE}$. The de Broglie wavelength λ of a proton of energy E is $\lambda = \frac{h}{p} = \frac{h}{\sqrt{2M_pE}}$. If the proton is confined in a nucleus of radius 3.6 F, then λ should either be smaller or at the most equal to $r = 3.6$ F. Therefore,

$$\left(2\times3.6\times10^{-15}\right)^2 = \frac{h^2}{2M_pE} \quad \text{or} \quad E = \frac{h^2}{2M_p}\frac{1}{\left(2\times3.6\times10^{-15}\right)^2}$$

Substituting the value of $M_pM_p = 1.67 \times 10^{-27}$ kg and $h = 6.62\times10^{-34}$ Js, we get

$$E = \frac{\left(6.62\times10^{-34}\right)^2}{2\times1.67\times10^{-27}}\frac{1}{51.84\times10^{-30}} = 0.25 \times 10^{-11} \text{ J}$$

$$= \frac{0.25\times10^{-11}}{1.6\times10^{-16}} \text{MeV} = 15.78 \text{ MeV}$$

It may be noted that the minimum energy required for confining a particle in given distance *d* is inversely proportional to the square of the distance. Therefore, if the nucleon is confined within a distance *r* (=3.6 F), the minimum energy will be around 64 MeV. Also, the average kinetic energy of nucleons inside a nucleus of average size is considered as 20 or 25 MeV.

Ans: The minimum kinetic energy of proton should be 15.78 MeV.

Solved example S-3.2

The mass defect ($M_{(Z,A)} - A$) for ${}^{16}_{8}O$, ${}^{15}_{8}O$, ${}^{1}_{0}N$, and, ${}^{1}_{1}P$ in μ u are, respectively, −5085, +3070, +8665, and, +7825.

Calculate the binding energies of ^{16}O and ^{15}O nuclei and the separation energies of last neutron from ^{16}O nucleus.

Solution. Mass defects in the problem are given in units of micro-unified mass unit. The masses of the respective nuclides in the unit of u are

$$M_{16O} = 16 - 0.005085 = 15.994915 \text{ u}; M_{15O} = 15 + 0.003070 = 15.003070 \text{ u}$$

$$M_N = 1 + 0.008665 = 1.008665 \text{ u}; M_P = 1 + 0.007825 = 1.007825 \text{ u}.$$

(i) Binding energy of ^{16}O, $B_{16O} = 8\,(M_N + M_P) - M_{16O} = 8(1.008665 + 1.007825) - 15.994915 = 16.13192 - 15.994915 = 0.137005 \text{ u} = 0.137005 \times 931.48 = 127.62$ MeV.

Binding energy per nucleon in $^{16}O = 127.62/16 = 7.97$ MeV

(ii) Binding energy of ^{15}O, $B^{15}{}_O = (7 \times M_N + 8 \times M_P) - M^{15}{}_O$

$$= (7.060655 + 8.0626) - 15.003070 = 0.120185 \text{ u} = 111.95 \text{ MeV}$$

Binding energy per nucleon in $^{15}O = 111.95/15 = 7.43$ MeV

(iii) Separation energy of last neutron from ^{16}O, $S_n = (M^{15}{}_O + M_N) - M^{16}{}_O$

$$= (15.003070 + 1.008665) - 15.994915 = 0.01682 \text{ u} = 15.67 \text{ MeV}$$

The binding energy of the last neutron may also be calculated from the difference of the binding energies of ^{16}O and ^{15}O nuclei

$$= 127.62 - 111.95 = 15.67 \text{ MeV}.$$

It may be observed from the calculations that the average binding energy per nucleon in ^{16}O is larger than for ^{15}O nucleus. Further the separation energy of the last neutron from the nucleus ^{16}O is much larger (15.67 MeV) than the average binding energy per nucleon (7.97 MeV). This is because ^{16}O is a doubly magic nucleus and is very strongly bound.

Exercise p-3.3: A target nucleus ${}^{A}_{Z}X$ is bombarded by energetic deuterons to produce the following (*d*, *p*) reaction:

$${}^{A}_{Z}X + {}^{2}_{1}D \rightarrow {}^{(A+1)}_{\;\;\;Z}X + {}^{1}_{1}H + Q_{(d,p)}$$

Show that the separation energy S_n of the last neutron from ${}^{(A+1)}_{Z}X$ is given by

$S_n = Q_{(d,p)} +$ Binding energy of deuteron.

Exercise p-3.4: The mass excess $(M - A)$ for the four stable isotopes ${}^{110}_{48}\mathrm{Cd}$, ${}^{111}_{48}\mathrm{Cd}$, ${}^{112}_{48}\mathrm{Cd}$, ${}^{113}_{48}\mathrm{Cd}$ and for neutron and proton are, respectively, −95,072, −95,811, −97,237, −95,591, +8665, and +7825 μu. Calculate the increase in the total nuclear binding energy when one neutron is added each to ^{110}Cd, ^{111}Cd, and ^{112}Cd nuclei.

3.3.2 Energy Levels of Mirror Nuclei

In several pairs of mirror nuclei, such as ${}^{7}_{3}\mathrm{Li}$, ${}^{7}_{4}\mathrm{Be}$; ${}^{11}_{5}\mathrm{B}$, ${}^{11}_{6}\mathrm{C}$; and ${}^{13}_{6}\mathrm{C}$, ${}^{13}_{7}\mathrm{N}$, the higher charge isobar decays into the lower charge isobar by emitting β^+ and neutrino. As the atomic mass number A for the two members of a mirror pair is same, they have same radius and same volumes. In the higher charge isobar, the charge Ze and in the lower charge isobar $(Z - 1)e$ charge is distributed in the same volume.

The schemes of low-lying level for the mirror pairs ${}^{7}_{3}\mathrm{Li}$, ${}^{7}_{4}\mathrm{Be}$ and ${}^{13}_{6}\mathrm{C}$, ${}^{13}_{7}\mathrm{N}$ are shown in figure 3.2a,b. It may be clearly seen in these diagrams that: (1) the angular moment and parity of the levels in the members of a pair are same. (2) The sequence of levels for the two members of a pair is identical, but (3) each level of the higher charge isobar of the pair is shifted in energy by the amount ΔU with respect to the corresponding level of lower charge isobar. The two members of each isobar-pair have two differences; first, their electric charge differs by one unit and second the higher charge isobar has one more proton and one less neutron compared with the lower charge isobar. Except these differences, the two members of a pair have all other properties such as radius, volume,

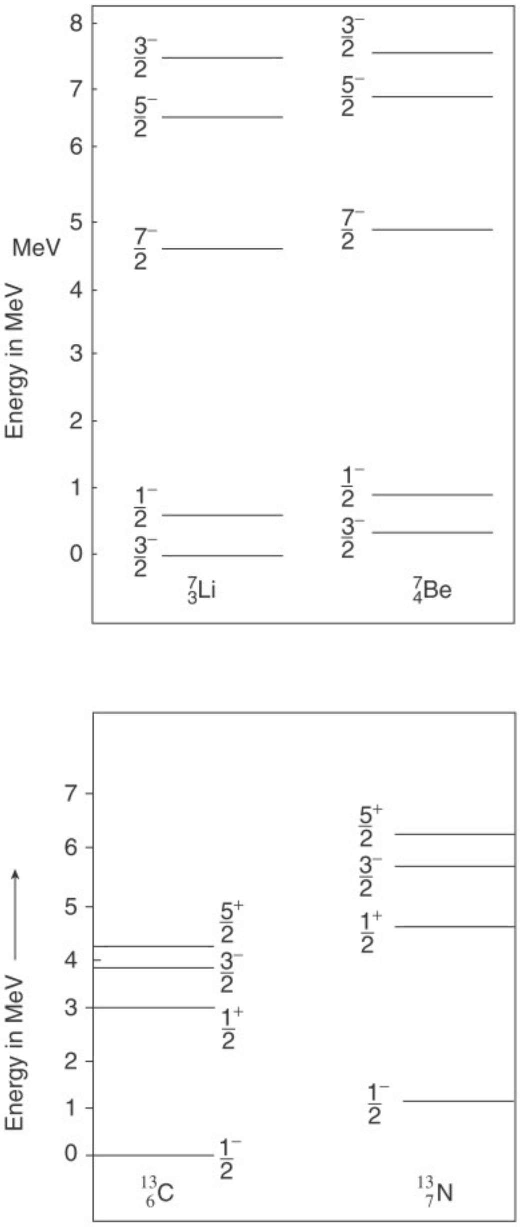

Figure 3.2 *Energy levels of (a)* ${}^{7}_{3}Li$ *and* ${}^{7}_{4}Be$ *and (b)* ${}^{13}_{6}C$ *and* ${}^{13}_{7}N$ *mirror pairs*

and total number of nucleons same. It is obvious that above-mentioned differences are responsible for the energy difference ΔU. ΔU will contain contributions from three effects:

1. Difference in the electrostatic energies of the two members of the pair ΔU_{Cou}
2. Neutron–proton mass difference ΔU_{mass}
3. Difference in the (nuclear) binding energies of the two members of the pair ΔU_{nuc}

Now, the electrostatic Coulomb energy U_{Cou} of a charge q, which is uniformly distributed in the volume of a sphere of radius r, is given by

$$U_{\text{Cou}} = \frac{3}{5}\left(\frac{q^2}{4\pi\varepsilon_0 r}\right)$$

Therefore,

$$\Delta U_{\text{Cou}} = \frac{3}{5}\left(\frac{e^2}{4\pi\varepsilon_0 r}\right)\left\{Z^2 - (Z-1)^2\right\} = \frac{3}{5}\left(\frac{e^2}{4\pi\varepsilon_0 r}\right)\{2Z-1\}$$

However, $r = r_0 A^{1/3}$, $r_0 = 1.5 \times 10^{-15}$ m, and $(2Z - 1) = A$, which gives

$$\Delta U_{\text{Cou}} = \frac{3}{5}\left(\frac{e^2}{4\pi\varepsilon_0 r_0}\right)\left\{A^{2/3}\right\}$$

The value of ΔU_{Cou} for any pair can be easily calculated from the above-mentioned equation.

Apart from the difference in the electrostatic Coulomb energy, the mass difference between a neutron and a proton in the parent and daughter nucleus also contributes to the total energy difference between the two isobars. The $\Delta U_{\text{mass}} = (M_{\text{n}} - M_{\text{p}})c^2 = 1.29$ MeV. The nuclear-binding energies of the two members of a pair may be different if the nuclear-binding force for a neutron is different from the binding force for a proton. The term ΔU_{nuc} will arise from that difference.

Putting all these terms together, we get

$$\Delta U = \Delta U_{\text{Cou}} + \Delta U_{\text{mass}} + \Delta U_{\text{nuc}} = \frac{3}{5}\,(e^2/4\pi\varepsilon_0 r_0)\,\{A^{2/3}\} - 1.29 + \Delta U_{\text{nuc}} \tag{3.1}$$

Experimentally, the magnitude of ΔU can be obtained from the decay products. The higher charge isobar of the pair decays into other members by β^+ (positron) emission. Neutrino is also emitted along with the positron. When the positron is emitted with maximum energy $E\beta_{\text{max}}$ the neutrino has zero kinetic energy. Since neutrino has zero rest mass the sum of $E\beta_{\text{max}}$ and the rest mass energy ($m_e c^2$) of the positron must be equal to ΔU.

$$\Delta U = E\beta_{\text{max}} + m_e c^2 \tag{3.2}$$

where m_{e} is the rest mass of the positron (or the electron).

A comparison of ΔU value calculated using Eq. (3.1) with the experimental values obtained from Eq. (3.2) or from other reaction data for each mirror pair, gave ΔU_{nuc} to be zero.

The fact that the shift in the energy of corresponding levels of mirror nuclei can be fully explained by the sum of the electrostatic Coulomb energy difference and the difference in the masses of neutron and proton clearly shows that the replacement of a neutron by a proton (or vice versa) does not change the nuclear binding energy. This simply means that nuclear forces do not differentiate between neutron and proton. In nuclear physics, the equivalence of the neutron–neutron force and the proton–proton force ($F_{1(n-n)} = F_{1(p-p)}$) is called charge symmetry and the more general equivalence of neutron–neutron, proton–proton, and the neutron–proton force [$F_{1(n-n)} = F_{1(p-p)} = F_{1(n-p)}$] is referred to as charge independence.

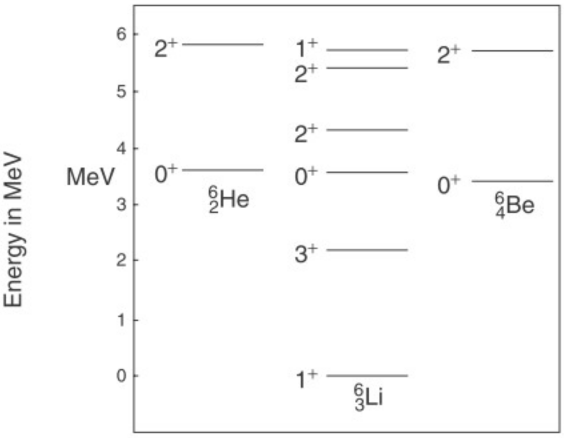

Figure 3.3 *Energy levels of ${}^6_2$He, ${}^6_4$Be, and ${}^6_3$Li*

Next, let us consider mirror nuclei that differ by two unit of charge, such as ${}^6_2$He and ${}^6_4$Be. In these nuclei, the total number of nucleons is same and hence their radii are equal. As may be seen in figure 3.3, the levels of the two nuclei are almost identical except a little energy shift. It is, however, important to note that in this pair, ${}^6_2$He has two (p–n) bonds and one (n–n) bond while in case of ${}^6_4$Be there are two (p–n) bonds and one (p–p) bond. Clearly, if there is any difference in the nuclear binding energies of the two nuclei, it must be due to the difference in the binding energies of (n–n) and (p–p) bonds. Calculations have shown that the slight energy difference in corresponding levels of the mirror pair can be fully accounted for by the differences in the Coulomb electrostatic energy and the energy corresponding to the difference in the masses of neutron and proton. Once again this shows that the nuclear force between two neutrons is equal to the force between two protons, proving the charge symmetry of nuclear force.

It is interesting to look at the energy levels of the nucleus ${}^6_3$Li (figure 3.3), which also have six nucleons and same radius as the other two nuclei. Some levels in figure 3.3 correspond to the levels in ^{6}He and ^{6}Be as expected; however, there are some additional levels as well. These additional levels in ^{6}Li appear because in the case of ^{6}He and ^{6}Be nuclei, some (p–p) or (n–n) states are restricted and are not populated due to the Pauli exclusion principle. In addition, additional (p–n) states in which exclusion principle does not apply are populated in ^{6}Li creating additional levels.

> *In short, the similarity of the low-lying levels of mirror nuclei and the fact that the energy shift between the corresponding levels is fully accounted by the electrostatic energy difference and the difference in the neutron proton masses states that fundamentally nucleon−nucleon force is charge independent.*

3.3.3 Difference in the Binding Energies of ${}^3_1$H and ${}^3_2$He

The mirror nucleus ${}^3_1$H has one proton and two neutrons and, therefore, the possible nucleon–nucleon interaction states are ^{3}(n−p) + ^{1}(n−p) + ^{1}(n−n), ^{3}(n−n) state is forbidden by Pauli exclusion principle. In ${}^3_2$He, the possible states are ^{3}(n−p) + ^{1}(n−p) + ^{1}(p−p); ^{3}(p−p) state is excluded because of exclusion principle. Any difference in the binding energies of the pair will come from the Coulomb energy difference and the binding energy difference of ^{1}(n–n) and ^{1}(p–p) states. The difference in the binding energies ≈0.76 MeV could be fully accounted by the difference of the Coulomb energies of the two nuclei. This again indicates that *nuclear force is charge symmetric*, $F^1(\text{n–n}) = F^1(\text{p–p})$.

Exercise p-3.5: Determine the Coulomb energy difference between the mirror isobars $^{13}_{7}$N and $^{13}_{6}$C and hence the maximum energy of positrons emitted by ^{13}N. Consider $r_0 = 1.4$ F.

3.3.4 Study of Deuteron Properties

Deuteron, $^2_1$H, is the smallest stable nucleus and has one neutron and one proton. The nucleons in deuteron revolve round the centre of mass and their relative separation vary with time. The square root of the mean of the squares of the distance between the centre of mass and the proton is called the r.m.s. electromagnetic radius and is denoted by r_d. Experimental values of some properties of deuteron are: (1) binding energy $B_d = 2.22$ MeV. (2) $r_d = 2.1$ F. (3) spin or the angular momentum $J = 1\hbar$. (4) Parity = even or positive. (5) Magnetic dipole moment $\mu_d = 0.857393$ nm. (6) Electric quadrupole moment $Q_d = +0.00282$ barn.

As the spin of deuteron is 1, the neutron and proton spins are parallel and, therefore, deuteron is an example of triplet neutron–proton state represented as a ^{3}(n–p) state. The deuteron has been analysed under quantum mechanical approach, in which a potential well is assumed to represent the system. In principle potential well of any shape may be used, but the simplest one is a square or rectangular well with two parameters, the depth V and the width b. High-energy scattering experiments have already suggested that there is a hard core of infinite height at a distance c from the centre of a nucleus such that nucleons cannot come closer than the distance c. This is also required for nuclear stability; otherwise nucleons will all collapse into each other under the strong attractive nuclear force between them. The radius of the core c has been estimated to be of the order of 0.4 F. A pictorial representation of the square-well approach is shown in figure 3.4. The appropriate well parameters should reproduce the observed properties of the deuteron, particularly, the binding energy, and the r.m.s. radius. The analysis shows that no unique values for the

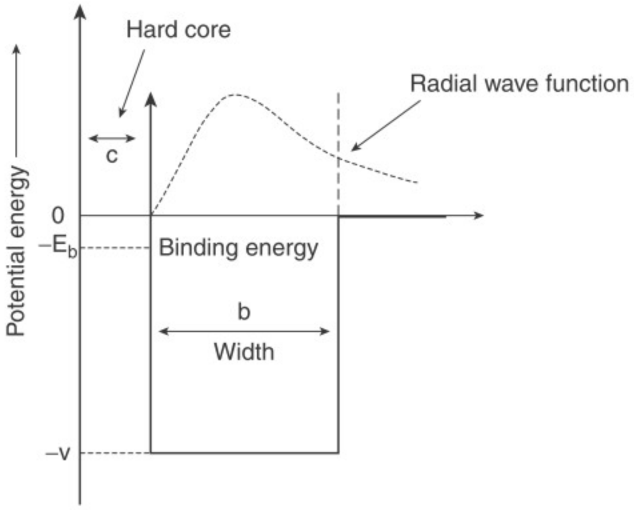

Figure 3.4 *Pictorial representation of square-well approach to deuteron and the shape of the radial wave function*

depth V and the width b could be obtained, rather several combinations of V and b could reproduce the observed binding energy. For example, a square-well potential of depth $V \approx 35$ MeV and width $b \approx 2$ F or $V = 73$ MeV and $b = 1.33$ F can give the correct binding energy of the deuteron. Detailed quantum mechanical analysis of deuteron has given the following information about the two body nucleon–nucleon forces.

1. Triplet (n–p) force $F_{3(n-p)}$ is just strong enough to bind deuteron, while the singlet force $F_{1(n-p)}$ is not enough strong to provide a bound singlet state. It means $F_{3(n-p)} > F_{1(n-p)}$.
2. In deuteron, the two nucleons remain outside the range of the nuclear force for a considerable period of time, resulting in small binding energy of deuteron. This is possible only in quantum mechanical systems where particles may undergo tunnelling, which is not possible in classical systems.
3. To reproduce the quadrupole moment of deuteron, an admixture of S and D states is necessary. The $F_{3(n-p)}$, thus, mixes states.
4. Further, the prolate shape of the deuteron suggests that the force between the neutron and the proton in deuteron is attractive when they are one above the other, and less attractive or even repulsive when they are side by side as shown in figure 3.5.

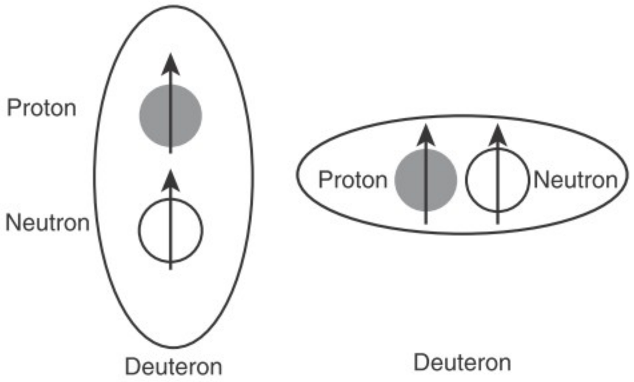

Figure 3.5 *The force between neutron and proton is attractive when they are one above the other and less attractive or even repulsive when they are side by side. That is why the shape of deuteron is prolate. Tensor force depends on the angle between the spins and the line joining the centres of nucleons*

This shows that the force also depends on the angles between the spin directions and the line joining the centres of nucleons. Such force is called a tensor force. The tensor dependence in radial nucleon–nucleon potential is generally represented by the term

$$S_{12} = (1/r^2)\,[3(\sigma_1 r)\,(\sigma_2 r) - (\sigma_1 \sigma_2) r^2]$$

Here, σ_1 and σ_2 are the spin vectors for the two nucleons. The term has different values for the cases when σs are parallel and perpendicular to the radius vector r, even for the same value of $\sigma_{\text{total}} = \sigma_1 + \sigma_2$.

Exercise p-3.6: Calculate the value of the tensor term S_{12} for the two configurations shown in figure 3.5 and for the same configurations in case of anti-parallel spins.

3.3.5 Nucleon–Nucleon Scattering

3.3.5.1 Neutron–proton scattering

As there is no bound singlet state of n–p system (system such as deuteron but with spin zero), it is not possible to study the properties of $F_{1(n-p)}$ directly. These un-bound states are, however, formed in scattering of neutrons with protons and their effect can be seen in scattering cross-sections. It may be noted that in scattering both triplet ^{3}(n–p) and singlet ^{1}(n, p) states are simultaneously populated as there is no restriction on any state from Pauli exclusion principle. The triplet state has three sub-states and the singlet state has one. As such the total number of sub-states is four. If it is assumed that each sub-state is populated with same probability, the total scattering cross-section will be equal to the weighted average of the sub-states, which will be given as

$$\sigma^{\text{total}} = \frac{3}{4}\sigma^{\text{triplet}} + \frac{1}{4}\sigma^{\text{singlet}} \tag{3.3}$$

Large number of precise experiments using neutrons from nuclear reactors and from secondary neutron-producing reactions initiated by accelerated charged particles has been conducted to measure the n–p scattering cross-sections. Polyethylene, organic material with carbon and hydrogen, is often used as target. In some experiments at very low energy, hydrogen gas has also been used as target. The measured total (n, p) cross-section between 0.01 eV and 15 MeV is shown in figure 3.6. As may be noticed in this figure, the curve can be divided into three distinct parts: (1) from 0.01 to 10 eV neutron energy, (2) between 10 and 10^4 eV, and (3) from 10^4 eV to 15 MeV. In the large energy range of 10–10^4 eV, the cross-section is nearly constant at a value of about

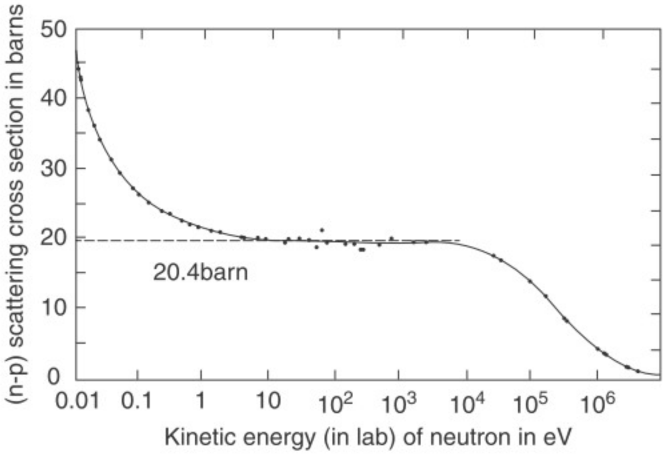

Figure 3.6 *(n, p) cross-section between 0.01 eV and 15 MeV*

20 barn. At lower energies, the scattering cross-section increases rapidly while at higher energies $>10^4$ eV, it decreases slowly. Analysis of the experimental data has been carried out assuming that the protons are fixed at their location and are free. At energies >10 eV, this assumption is true; however, at lower energies where mostly gaseous targets have been used, the protons were neither free nor fixed. At lower energies, the de Broglie wavelength of neutron becomes large and, therefore, the neutron interacts with the hydrogen molecule rather than the proton. The thermal motion of gas molecules also increases the apparent target size. Moreover, the molecular binding energy becomes comparable to the energy of incident neutron. All these effects enhance the measured scattering cross-section at lower energies. On applying corrections for these effects in the measured cross-section values, it turns out that in the lower energy range also the actual scattering cross-section is of the order of 20 barn.

In the quantum mechanical framework, the scattering problem is treated by decomposing the incoming neutron wave into partial waves of different orbital angular momentum l with respect to a fixed target. In elastic scattering, the target potential changes the phases of the incident partial waves without affecting the flux. The net effect of scattering potential is a shift in the phase of the l-th partial wave by an angle d_l. The scattering cross-section for each partial wave then becomes proportional to $\sin^2 d_l$. At low energies below 15 MeV, only $l = 0$ or S-wave contributes to scattering because the higher partial waves lie beyond the range of the nuclear field of the target proton. As such for low energy, the total (n−p) scattering cross-section $s^{s}_{(n,p)} = \{4p \sin^2 d_0/k^2\}$, where $k = \frac{\sqrt{2mE}}{\hbar}$ and m is the reduced mass = $M_n/2$ and E the centre of mass energy of incident neutron. In the zero energy limit $k \rightarrow 0$, the scattering cross-section can be given in terms of the parameters a_t and a_s, which are the scattering lengths for the triplet and singlet scattering events. Scattering lengths is the intercept of the tangent to the radial wave function at the outer boundary of the potential well. Scattering length a_t is positive for the bound ^{3}(n, p) state and a_s is negative for an unbound state like ^{1}(n, p). Precise values for a_t and a_s determined from coherent and incoherent scattering of neutrons by protons give $a_t = 5.38$ F and $a_s = -23.7$ F. The experimental value of scattering length for triplet state is consistent with the deuteron potential well parameters that give correct binding energy and r.m.s. radius of deuteron.

The negative value of the scattering length for singlet potential is consistent with the fact that the singlet state is not bound. If same well width $b = 1.33$ F that gives correct binding energy for deuteron is used with the singlet scattering length = −23.7 F, the singlet potent depth for zero energy approximation comes out to be 54.7 MeV, while the minimum depth of the well to produce a bound system is about 57 MeV. As such it may be stated that the singlet state narrowly misses to become a bound state by only about 3 MeV or so.

In summary, it may be said that a rectangular well with appropriate width and depth parameters can satisfactorily reproduce the binding energy, r.m.s. radius, scattering lengths for triplet and singlet states obtained from measured scattering cross-sections at low energy and for zero-energy approximation. Also, for higher energies, the cross-sections for both triplet and singlet state scatterings can be represented in terms of two additional parameters, the effective range r_{ot} and r_{os}, respectively, for the triplet and the singlet states. With the parameters $a_t = 5.28$ F, $a_s = -23.7$ F, $r_{ot} = 1.7$ F and $r_{os} = 2.4$ F, the measured total scattering data between 10 eV and 15 MeV can be satisfactorily explained. This shows the great success of quantum mechanical treatment of both the deuteron and the scattering process.

3.3.5.2 Proton–proton scattering

Proton–proton scattering has also been extensively studied. The major points of difference between n–p and p–p scattering are: (1) Triplet ^{3}(p–p) state in not possible due to exclusion principle. (2) Coulomb force is always present at all angles and for all partial waves. To overcome the interference due to Coulomb scattering differential scattering cross-section, $\left(\frac{d\sigma}{d\Omega}\right)$ is measured and the experimental value is corrected for the contribution from Coulomb scattering. (3) As the target and the incident particles are same, it becomes difficult to differentiate the incident particle and the scattered target. As a result, it becomes impossible to differentiate between small-angle and large-angle scattering events. (4) According to Dirac's theory, the vacuum is filled with negative energy electrons. In the presence of strong electrostatic field around proton, the wave functions of negative energy electrons get distorted resulting in the polarization of space and at very small distance from the proton the $1/r^2$ dependence of Coulomb field does not hold. However, at larger distances there is not much effect. Proton–proton scattering has been successfully treated quantum mechanically and vacuum polarization and interference between Coulomb and nuclear scattering terms have been clearly observed. Although the p–p scattering phase shift and n–p scattering phase shifts at the same energy cannot be compared but it is possible to calculate back the potential well parameters for p–p scattering. These calculations have shown that the potential well parameters for singlet p–p scattering are nearly equal to the parameters for singlet n–p scattering, once again confirming that nuclear forces are charge independent.

Nucleon scattering at higher energies confirmed the presence of a hard core. Further, the concept of phase shift has been successfully applied at high energies and consistent set of phase shift parameters have been obtained that can explain all the quantities that are observed in nucleon–nucleon scattering.

The study of deuteron and of the nucleon–nucleon scattering has revealed the following:

1. Nuclear forces are charge independent. The same inference was also drawn from the study of separation energies of last nucleon and the level schemes of mirror nuclei.
2. Scattering at high energies show that there is hard or repulsive core at the centre of the nuclear potential.
3. Nuclear force is spin dependent and of tensor nature.

Exercise p-3.7: In n–p scattering cross-section, triplet state contributes three times more than the singlet state. Explain the reason for this larger contribution of the triplet state.

Exercise p-3.8: The experimentally measured (n, p) scattering cross-section is found to increase with the decrease of energy, in the very low energy region. How can this be explained?

Exercise p-3.9: Clearly bring out the differences in (n, p) and (p, p) scattering.

3.3.6 Constant Density of Nuclear Matter: Exchange Forces

The density of the nuclear matter can be readily calculated by dividing the mass of a nucleus by its volume. Mass of a nucleus with A nucleons = $M_n A$, where M_n is the nucleon mass and the volume = $4/3\pi r^3$, where r is the radius. The radius $r = r_0 A^{1/3}$, where the unit radius r_0 (=1.5 F) is constant. The density of nuclear matter, therefore, is $\frac{M_n}{\frac{3}{4}\pi r_0^3}$ (where M_n is the nucleon mass) a constant value for all nuclei. As all nuclei big or small have same density, they may be compared with the big and small drops of a liquid. The similarity between a liquid drop and a nucleus may be looked in another way as well. In case of a liquid drop, each molecule of the liquid is bound with only few molecules around it and not to all the molecules present in the liquid. As a result, the total binding energy of a liquid drop becomes proportional to the number of molecules in the drop. In a similar way, each nucleon inside a nucleus is bound only to few other nucleons around it and not to all other nucleons present in the nucleus. This results in the total binding energy of the nucleus being proportional to A, the number of nucleons in it. [If each nucleon in the nucleus was interacting and was bound to all other $(A - 1)$ nucleons, then the binding energy of the nucleus should have been proportional to the number of all the bonds which are $A\ (A - 1)$. For heavy nuclei, $A\ (A - 1) \approx A^2$ and, therefore, the total binding energy of heavy nuclei would have been proportional to A^2 instead of A].

The property of the nucleon–nucleon force that restricts the interaction of a nucleon to only a few around it is called the saturation property of the nuclear force in analogy to the saturation of the force binding molecules in liquids. The analogy between the nucleus and the liquid drop may be further explored to understand the binding between the nucleons in a nucleus.

Let us consider a drop of liquid hydrogen. In this drop, two atoms of hydrogen are chemically bound through homopolar binding to make an H_2 molecule. A third atom of hydrogen is not attracted now to join with a molecule of H_2. This is called the saturation of chemical binding force. In homopolar binding, this saturation is achieved by assuming that the electrons continuously exchange their position from one atom to the other in the molecule. Mathematically, this exchange of electrons between the atoms of a molecule can be achieved by using exchange operators.

Extending the concept of exchange operator to the case of nucleon, there may be three types of operators: (1) operator that may exchange the space coordinates of two nucleons, (2) operator that may exchange the spin coordinates of two nucleons, and (3) the operator that may exchange both the spin and space coordinates. Apart from these exchange operators, there may also be a non-exchange operator that does not exchange any property. In simple terms, operators of a force are generally multiplicative factors. For example, if an attractive force is operated by the operator -1, the force becomes repulsive. Operators of exchange forces change the nature of the force between nucleon from attractive to repulsive depending on the value of the relative angular momentum between the nucleons.

1. **Majorana force:** The Majorana force exchanges the position coordinates of the two nucleons. It simply means that the Majorana force operator only interchanges the positions of the two interacting nucleons, while their spins and other properties remain same. Majorana force is attractive for the two nucleons if their relative angular momentum l is zero or even repulsive for odd l states. For example, in S ($l = 0$) and D ($l = 2$) states it will give attraction and for the P state ($l = 1$) it will provide repulsion between the two nucleons.
2. **Bartlett force:** The operator of Bartlett force interchanges the spins of the two interacting nucleons, leaving all other parameters including their positions unchanged. For example, if the first nucleon has its spin up and the other spin down, then as a result of Bartlett interaction the spin of the first nucleon will turn down and that of the other up. (*What will be the effect of Bartlett interaction on two parallel spin nucleons?*). For zero or even angular momentum states, Bartlett force is attractive in triplet and repulsive in singlet states. For odd l states, Bartlett force is repulsive for triplet and attractive for singlet states.

3. **Heisenberg force:** This force exchanges both the spin and the space coordinates of the two interacting nucleons. Naturally, the operator produces the same effect that is produced by the successive operations of Majorana and Bartlett force operators. The Heisenberg force is attractive for even-l triplet and odd-l singlet states. It is repulsive for the even-l singlet and odd-l triplet states.
4. **Wigner force:** Wigner force is not a force that provides saturation. This force provides only binding to the two interacting nucleons. Its operator is +1 for states of all l values.

Table 3.1 *Exchange operators for different type of forces*

Type of Force	Even-l Singlet	Odd-l Singlet	Even-l Triplet	Odd-l Triplet
Majorana	+1	−1	+1	−1
Bartlett	−1	−1	+1	+1
Heisenberg	−1	+1	+1	−1
Wigner	+1	+1	+1	+1

The actual force between two nucleons cannot be totally of Wigner type as it will not be able to provide the saturation. Also, it cannot be totally of Heisenberg type because Heisenberg force is attractive for triplet state if nucleons are in the state of $l = 0$ (S-state). The deuteron fulfils these conditions, and should have been the most tightly bound system if nucleon–nucleon force was totally of Heisenberg type. We know that the binding energy of deuteron is only 2.23 MeV. Alpha particle, the nucleus of helium, is the most tightly bound nucleus with binding energy of 28 MeV. Hence, Heisenberg force alone cannot be the nucleon–nucleon force. As the exchange of π meson between two nucleons corresponds to the exchange of their positions, Majorana force must be a significant component of the total nucleon–nucleon force. The two nucleon system can be represented as a mixture of Majorana and Wigner forces.

3.4 MEDIATION OF NUCLEAR FIELD

How does two nucleons stick together to make a nucleus? The answer to the question is obtained by drawing an analogy from the interaction of two charges. In quantum electrodynamics, a well-understood and developed theory of Coulomb field, two charged bodies interact through Coulomb field. Let us, therefore, first review the mechanism of interaction between two charges.

Quantization of electric field generates photon, which communicates the electric field between charged bodies through their exchange. This mechanism of interaction brings in the concept of the speed of the field and the time delay in action at a distance. In case of electric charges, the speed of the field is the speed of light, as photon, the quanta of the field travels with the speed of light. The mechanism also brings in the concept of virtual photons in case of electric field and virtual field particles in general.

Virtual photons are quite different from the real photons. These are called virtual because they exist within the constraints imposed by the uncertainty principle. According to this principle, an uncertainty ΔE in energy can stay for a time duration Δt such that $\Delta E \times \Delta t \sim \hbar$. As such, if a virtual photon of energy E is created, it will live for a time $\Delta t \sim \hbar/E$. A virtual photon travelling with the speed of light c will cover a distance $r = c\Delta t$ in time Δt. Hence, $r = c(\hbar/E)$. A virtual photon of energy E can be created for time duration $\Delta t = \hbar/E$ and can make a trip of distance $r = \hbar c/E$ during the time interval Δt. After this time interval, the photon must disappear to conserve the energy.

This fluctuation in energy is allowed only within the limits of time Δt, allowed by the uncertainty principle. Hence, a virtual photon of energy E is emitted by an electron, it carries a momentum $p = E/c$. The virtual photon is absorbed by the other charged particle changing its momentum by the amount of p. As such the total change of momentum during the process of exchange (emission and absorption) of virtual photon between two charged particles is $2p$. Now, according to the law of mechanics the force F responsible for the change of momentum (for the exchange of virtual photons in this case) is equal to the rate of change of momentum $= 2p/\Delta t$. As such

$$F = 2p/\Delta t = (2E/c)/(\hbar/E) = 2E^2/\hbar c = [2(\hbar c/r)^2)/\hbar c] = (2\hbar c/e^2)\,(e^2/r^2)$$

This is how the electrostatic force between two charges may be derived from the concept of the virtual photon exchange. The electromagnetic force is an infinite range force and this is due to the exchange of mass-less particles, that is photons. However, these photons are not real but virtual. These virtual photons exist only in association with charges and currents and are not to be confused with real photons. However, the 'virtual' photons have real affects and, in that respect, are real. For example, a real photon moving in free space can be represented as in figure 3.7a where the photon dissociates in to a virtual $e^{\pm}$ pair which heals up (or recombines) after a time $\Delta t = \hbar/\text{mc}^2$. Similarly, (see figure 3.7b) an electron passing in free space may be represented as a particle that emits a virtual photon, which dissociates into a virtual $e^{\pm}$ pair that eventually heals and emits a virtual photon, which is reabsorbed. The free e^- is thus surrounded by charges (virtual) in the vacuum and the e^- of e pair is repelled while e^+ of $e^{\pm}$ pair is attracted by the real e^-. This is called vacuum polarization and affects the inverse square law in the immediate vicinity of charges. The effect of vacuum polarization is felt in p–p scattering and in other processes.

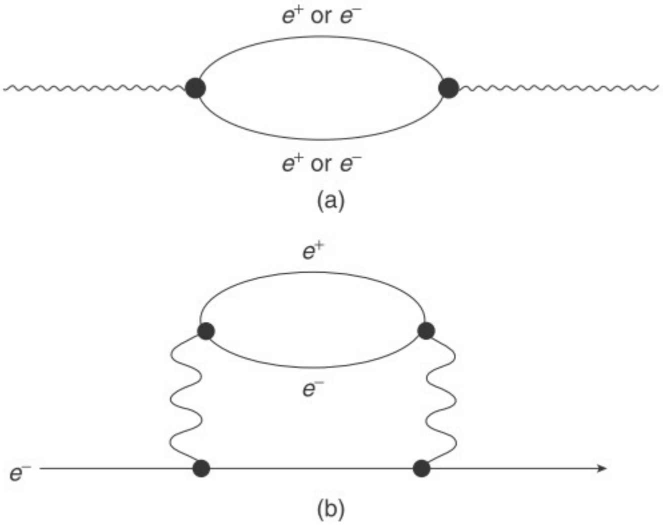

Figure 3.7 *(a) A real photon moving in free space may dissociates into a virtual e^+ and e^- pair. (b) An electron may emit a virtual photon that may create a virtual electron–positron pair which recombine to produce a virtual photon that is absorbed by the electron*

(Virtual photon emitted by a charged particle, an electron in the figure, creates a pair of virtual electron and positron which live for a short time and recombine (heal up) to give back virtual photon which in turn is reabsorbed by the electron)

This analogy to the electrodynamics was used by Yukawa to predict the emission and absorption of a virtual particle for mediation of nuclear force, now known as the p -meson or, pion. The nuclear force that has finite range, obviously, has to be mediated by the exchange of finite mass particle. Its range r = ħc/E = ħ/mc is controlled by the mass of mesons such as pion.

Yukawa considered the following scalar potential equation of electrodynamics

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)U = 0 \tag{3.4}$$

which reduces to$\nabla\ ^2U = 0$, for the static field and on integration gives the static solution $U = q/(4\pi\varepsilon_0 r)$. It may be observed that Eq. (3.4) can be obtained from the equation giving energy of a mass less particle by substituting the quantum mechanical operator $(-\hbar^2\nabla^2)$ for the square of the momentum (p^2) and $(-i\hbar\ \partial/\partial t)$ for the energy.

For a mass-less particle:

$$E^2 = p^2c^2 \text{ or } -p^2 + E^2/c^2 = 0 \tag{3.5}$$

On substituting the values of p and E in terms of the operators in Eq. (3.5), one obtains Eq. (3.4). The equation for the energy of a particle of rest mass m is

$$E^2 = p^2c^2 + m^2c^4 \quad \text{or} \quad -p^2 - m^2c^2 + E^2/c^2 = 0 \tag{3.6}$$

Yukawa argued that if quantum mechanical operators for momentum and energy are substituted in Eq. (3.6), it should give the equation for the scalar potential u for the mass m particle that may be exchanged between nucleons. On substitution of operators, we get

$$(\nabla^2 - m^2c^2/\hbar^2 - 1/c^2\ \partial^2/\partial t^2)\ u = 0.$$

which in case of static field (time independent case) reduces to

$$(\nabla^2 - \frac{m^2c^2}{\hbar^2})\ u = 0 \text{ or } (\nabla^2 - \mu^2)\ u \tag{3.7}$$

where $\mu = mc/\hbar$. A solution of Eq. (3.7) is $u = g\left(\frac{e^{-\mu r}}{r}\right)$ (3.8)

Here, g is a constant, which plays the same role as is played by charge in electrostatic. g depends on the properties of the source, the nucleon in the present case. In electrostatic case, the mechanical potential resulting from the scalar potential U is $V = q^2/4\pi\varepsilon_0 r$. Extending the analogy, the mechanical potential in the case of two nucleons each of strength g is

$$V = g^2\,\frac{e^{-\mu r}}{r} \tag{3.9}$$

3.4.1 Virtual Mesons

Virtual particles (mesons) are continuously exchanged between nucleons giving binding and saturation. These particles of mass m and energy $E = mc^2$ are virtual because they are created out of the uncertainty principle and live for a time $\Delta t = \hbar/mc^2$. In the interval of time Δt, the virtual particle if travelling with the maximum speed, the speed of light may travel a maximum distance $d = \Delta tc = \hbar/mc$. Thus, d gives the maximum range of the field. Substituting the range of nuclear field $d \approx 2\,\text{F}$, the mass m of the virtual particle may be calculated to be around 200 m_e, m_e being the mass of electron. It may be remarked that the range of the virtual particle that communicate the field is inversely proportional to the mass of the virtual particle.

Soon after the prediction by Yukawa, a particle named μ-meson (presently called muon) was discovered in cosmic rays. Initially, it was thought that μ-meson is Yukawa's particle. However, it was soon recognized that μ-meson could not be the right candidate for two reasons. First, it weakly interacts with matter and second, it has a spin $\frac{1}{2}\hbar$, therefore, cannot be absorbed or emitted from a nucleon without changing its spin. As a matter of fact, all field quantas that mediate fields, such as photon, π-meson, gluon, W and Z bosons, and so on are bosons with zero or integer intrinsic spins.

3.4.2 Yukawa Particle: π-meson

Soon the right Yukawa particle, the π-meson (also called pion) with a mass about 280 m_e was found, again in cosmic rays. There are three types of pions π^+, π^0, and π^-. Charged virtual pions are exchanged between proton and neutron, while neutral pions between neutron–neutron and proton–proton. Substituting the mass of π-meson as 300 m_e the range of the force comes out to be about 0.7 F. There is one problem with π-meson as well. It is that the parity of π-meson is negative. If the virtual π-meson is emitted or absorbed by a nucleon in $l = 0$ (S-state), then the parity of the nucleon will change. Therefore, the π-mesons are emitted or absorbed in $l = 1$ state or in other higher odd l states. This leaves the parity of nucleon unchanged but their spins are flipped, that is changes from 1/2 pointing up to 1/2 pointing down. In case of deuterons, the exchange of charged virtual pions between proton and neutron in $l = 1$ state automatically makes the wave function for the system non-spherical. Thus, tensor force term, which makes deuteron shape prolate and gives rise to additional attraction when the nucleons are one over the other as compared to when they are side by side, evolves naturally. It is also possible that not one but two virtual pions are emitted and absorbed simultaneously. The two-pion potential is of shorter range and varies as $e^{-2\mu r}/r$. Other heavy pions may also be exchanged but only when the nucleons are much closer. Figure 3.8 shows the emission and absorption of virtual pions. A very crude sketch of nucleon–nucleon potential showing relative distances where one-, two-, and three-pion exchanges become important is shown in figure 3.9.

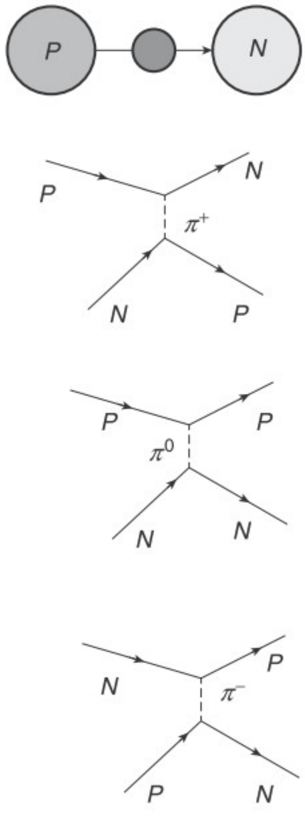

Figure 3.8 *Exchange of virtual pions between nucleons*

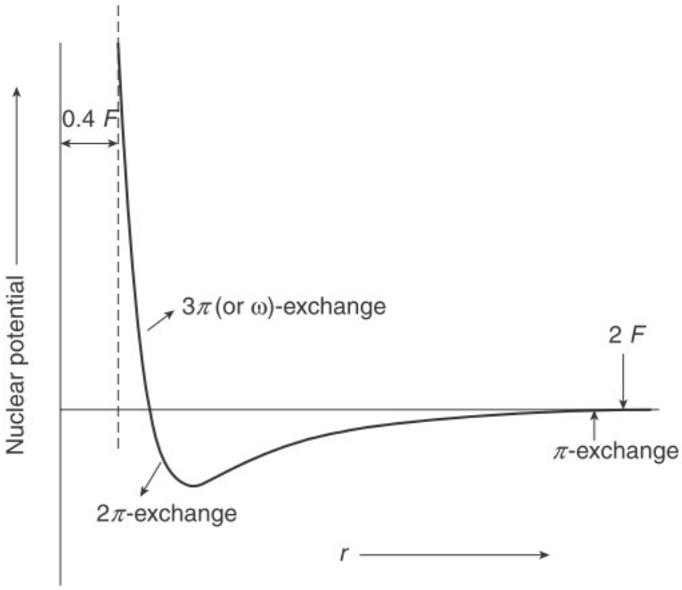

Figure 3.9 *A crude sketch of nucleon–nucleon potential showing distances where one-, two-, and three- pion exchanges become important*

3.5 SPIN–ORBIT DEPENDENCE OF NUCLEAR FORCE

The occurrence of magic numbers suggested that nucleons inside a nucleus move in some sort of shells just like the electrons in an atom. The cyclic behaviour of separation energy for the last neutron and proton further confirm the shell structure. Nuclear shell closures at lower magic numbers Z and/or $N = 2, 8, 20$ could be easily explained on the basis of shell model. However, to achieve the closure of shells at higher neutron or proton numbers 28, 50, 82, and 126 it was necessary to invoke strong coupling between the orbital and the spin motions of each individual nucleon**.** Detailed study of shell model of the nucleus revealed that the nuclear force also depends on the relative directions of the orbital angular momentum l and the spin angular momentum s of the nucleon. The force is attractive if l and s are parallel and repulsive if anti-parallel. This gives rise to what is called the spin orbit coupling term in nuclear potential. As the orbital angular momentum of a nucleon depends on its velocity of motion, the spin-orbit dependent force is also called the velocity-dependent force.

3.6 NUCLEON–NUCLEON POTENTIAL

Before proceeding further let us once again recapitulate the basic information about nucleon–nucleon force obtained from different sources.

1. They are of short range, charge symmetric, and charge independent.
2. Force is spin dependent. Triplet (n, p) force is strong enough to form a bound state but the singlet (n, p) force is just not strong enough to form a bound state. Singlet (n, p), (n, n), and (p, p) forces are of equal or nearly equal magnitude.

3. Nucleon–nucleon force is of tensor nature that means the force between two nucleons not only depends on the spins of the two nucleons but also on the angles which their spins make with the line joining their centres.
4. Nucleon–nucleon force is saturated, which means that a nucleon interacts with only a few nucleons around it. Saturation of force is brought about by exchange of pions between nucleons. Exchange force can explain saturation as it may also become repulsive depending on the property that is exchanged.
5. The nucleon–nucleon force also depends on the relative orientation of their angular momentum and spins. This is called the velocity dependence of the force.
6. High-energy scattering experiments have demonstrated that there is a hard repulsive core (≈0.4 F) such that two nucleons cannot came closer than this distance.
7. Since total electric charge is conserved in nuclear processes and the isotopic spin is related to charge, it is also conserved like the total angular momentum. All properties shown by angular momentum are expected to be shown by the isotopic spin as well.

As has been indicated earlier, a nucleon–nucleon potential now needs to be developed that may incorporate the above-mentioned properties of nucleon–nucleon force. A very crude shape of the potential is shown in figure 3.9 in which the distance where one, two, and three pion transfer becomes important is also shown. The potential may then be used in the Schrödinger equation of the given system to get its important parameters. The simplest potential is the square-well or rectangular-well potential. Other potential forms such as Woods–Saxon (figure 3.10), Yukawa (figure 3.11), Harmonic oscillator (figure 3.12), and many other have been tried each with limited success.

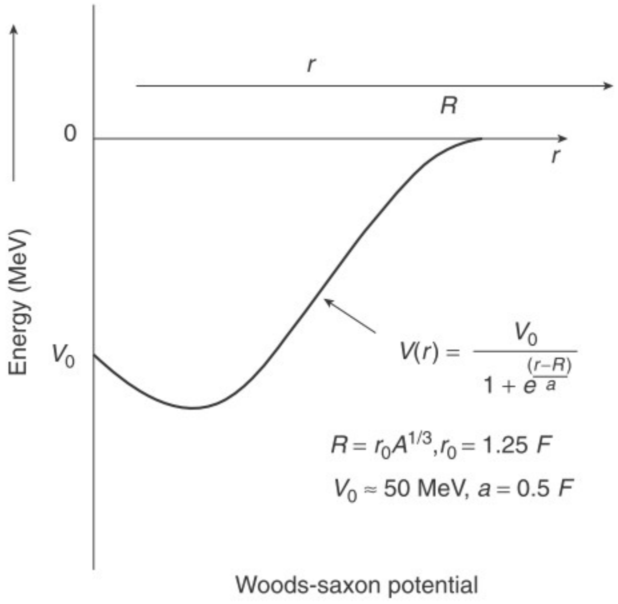
Energy (MeV)
r
R
0
r
V_0
$V(r) = \frac{V_0}{1 + e^{\frac{(r-R)}{a}}}$
$R = r_0 A^{1/3}, r_0 = 1.25\ F$
$V_0 \approx 50$ MeV, $a = 0.5\ F$
Woods-saxon potential

Figure 3.10 *Woods–Saxon potential*

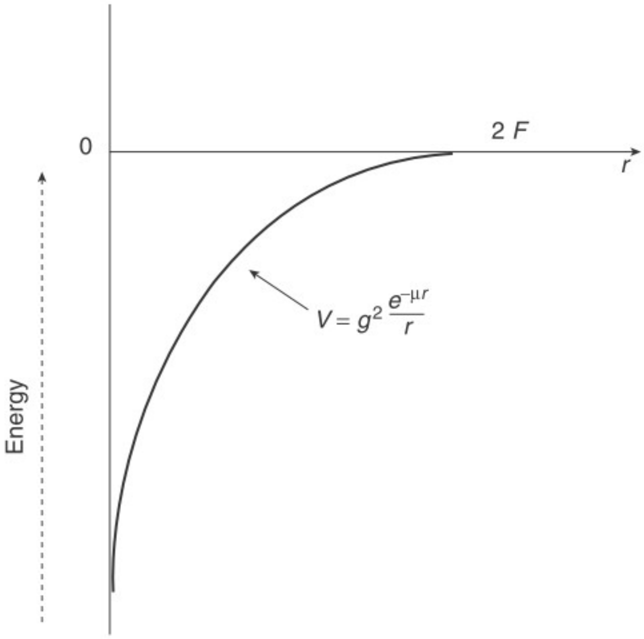

Figure 3.11 *Yukawa potential*

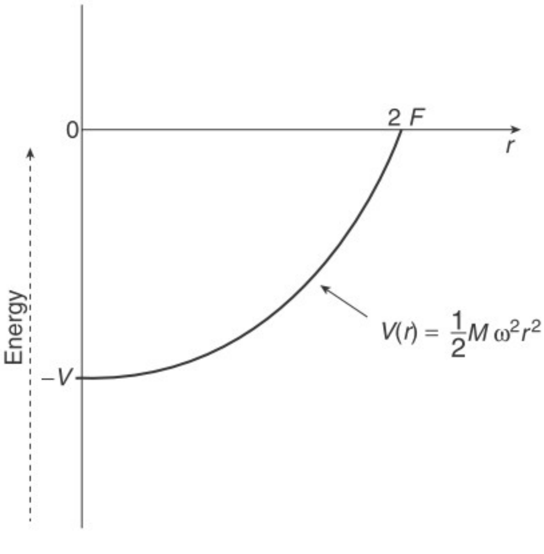

Figure 3.12 *Harmonic oscillator potential*

One form of total potential between two nucleons, in modern parlance called AV8 potential is as follows.

$$V_{12} = V(r) + V_{\sigma}(r)\,(\sigma_1\sigma_2) + V_{\varsigma}(r)\,(\zeta_1\zeta_2) + V_{\mathrm{T}}(r)S_{12} + V_{\mathrm{T}\varsigma}(r)S_{12}(\zeta_1\zeta_2) + V_{\mathrm{T}\sigma}(r)\,(\sigma_1\sigma_2)\,(\zeta_1\zeta_2) + V_{\mathrm{ls}}(r)\,(ls) + V1_{\varsigma}(r)\,(ls)\,(\zeta_1\zeta_2)$$

Here, $V(r)$s are radial potentials which may have square well or Woods–Saxon or any other form for each component of the total potential. Terms I, II, III …VIII of the potential represent, respectively, pure radial, spin–spin, isospin–isospin, tensor, tensor–isospin, tensor spin–spin, spin–orbit, and isospin–orbit interactions. As already indicated, the tensor term S_{12} is generally given as:

$$S_{12} = 1/r^2[3\,(\sigma_1 r)\,(\sigma_2 r) - (\sigma_1\sigma_2)\,r^2]$$

where σ_1 and σ_2 are the spins of the two nucleons and ζ_1 and ζ_2 their isospins. It may be observed that the first term in the expression for S_{12} depends on the angles between the spins and the radius vector while the second term on the relative orientation of the spins of the two nucleons.

3.6.1 Local and Global Nuclear Potentials

Instead of using a two-nucleon potential, attempts have been made to write the potential for the whole nucleus. Optical potential – which contains two terms, the real and the imaginary – has been quite successful in explaining the scattering and nuclear reactions. Nuclear potentials are generally classified as local – which is applicable to a small range of energy and/or the nuclear mass. Global nuclear potentials can be applied for a broader range of energy/nuclear mass. None of the potentials developed so far is applicable to all the stable nuclei. It has also been observed that not only two-body force but three-body forces are also important in case of the nucleus.

3.7 QUARK AND GLUON

It is now well established that all matter in the universe is made of two types of entities: quarks and leptons. Quarks and their anti-particles, anti-quarks, may group together to make particles that may stay free, such as nucleons and π-mesons. Particles formed by the grouping of quarks are called HADROS. Leptons may interact with quarks and with their groupings but they do not group or pair among themselves. Electrons, muons, and their neutrinos are leptons. There are two types (called flavours) of quarks; the up-quark denoted by u has an electric charge of $+2/3e$ and the down-quark denoted by d has an electric charge of $-1/3e$, e being the charge of proton. Their antiparticles denoted by $\bar{\mathrm{u}}$ and $\bar{\mathrm{d}}$ have $-3/2e$ and $+1/3e$ charges. The quark structure of some of the particles is as follows:

Proton = (uud); neutron = (udd); π^+ meson = $(\mathrm{ud}\bar{\mathrm{d}})$; π^- meson = $(\mathrm{u}\bar{\mathrm{u}}\mathrm{d})$

While Yukawa showed that the strong interaction between nucleons is communicated by the exchange of π-mesons, it has now become clear that the interaction between quarks is more fundamental. The strong force between quarks is communicated through the exchange of a particle called gluon and not by the exchange of π-mesons. The force that is communicated by the exchange of π-meson is now treated as a residual force of gluons, such as the van der Waals force in chemical binding is the residual electrostatic force between the positively charged atomic nucleus and the negatively charged electrons. It may, however, be pointed out that the Yukawa theory of meson exchange still remains valid for the nuclear two-body residual interaction, although the more fundamental strong nuclear force is communicated by gluons between quarks.

The quark structure of particles though quite successful; however, there was no apparent reason why three quarks group up to form a nucleon or a quark and an anti-quark join to make a π-meson. There was another problem: two very short lived particles Δ^{++} and Δ^{-} with quark structures (uuu) and (ddd) were identified. This was in violation with Pauli exclusion principle. To overcome this difficulty, Moo Young Han of Duke University, Yoichiro Nambu (University of Chicago), and Oscar Greenberg (University of Maryland) suggested that a new quantum number called 'colour' may be assigned to each quark. The quantum number colour was assumed to have three components, 'red', 'green', and 'blue' denoted, respectively, by subscripts 'r', 'g', and 'b'. As such there can be six types of quarks: u_r, u_g, u_b, d_r, d_g, and d_b along with their six anti-quarks with inverse colours. These colours were assigned with a specific purpose, red, green, and blue colours on mixing produce white colour – neutrality of colour, in a similar way it was demanded that particles that may stay freely must have a combination of quarks such that the total system is colourless or white. As such Δ^{++} with the structure ($u_r u_g u_d$) does not violate Pauli principle as each quark is in a different quantum state because of its colour quantum number. Further, the total structure is colour neutral or achromatic. Pions also may stay freely because they make a colourless combination of quarks. In Yukawa theory, the pions carry charge from one nucleon to the other. In case of the quark, the quanta of strong nuclear field gluon exchange colour between quarks. Colour is regarded as the equivalent of electrical charge in quark–gluon physics and the atomic physics where stability demands electrical charge neutrality, the particle physics demands colour neutrality or chromatic neutrality.

Exercise p-3.10: What are virtual particles? Clearly distinguish between a real and a virtual pion and explain how virtual pions mediate between nucleons.

Exercise p-3.11: Briefly describe how Yukawa developed the concept of the mediation of nuclear interaction between nucleons through the exchange of π-meson.

Exercise p-3.12: Using the range of the nuclear field as 2 F, show that the mass of the virtual particle that mediate interaction between nucleons should be around 168×10^{-27} kg.

Multiple choice questions

Note: In some of the following questions more than one alternative may be true. Tick all the correct alternatives in such cases for the complete answer.

Exercise M-3.1: A system of two half-integer spin particles has repulsive singlet-state for all values of relative angular momentum. The most likely exchange force is

(a) Majorana (b) Bartlett (c) Heisenberg (d) Wigner

Exercise M-3.2: The range of a virtual particle of mass m is of the order of

(a) $\frac{\hbar}{mc^2}$ (b) $\frac{\hbar}{mc}$ (c) $\frac{\hbar c}{m}$ (d) $\frac{mc}{\hbar}$

Exercise M-3.3: The mass excess ($M - A$), defined as the difference between the atomic mass M and the atomic mass number A of the nucleus, for four isobars X, Y, M, and N is, respectively, −5432−3248, +3269, and +5671 μm (micro mass unit). The most and the least bound nuclei are, respectively,

(a) N and X (b) M and Y (c) X and N (d) Y and M

Exercise M-3.4: The maximum Coulomb potential energy (in MeV) between $^{27}_{13}$Al and $^{64}_{29}$Cu nuclei is about ($r_0 = 1.2$ F)

(a) 15 (b) 25 (c) 35 (d) 45

Exercise M-3.5: The life time of a virtual particle of mass m cannot be larger than

(a) $\frac{\hbar}{mc^2}$ (b) $\frac{\hbar}{mc}$ (c) $\frac{\hbar c}{m}$ (d) $\frac{mc}{\hbar}$

Exercise M-3.6: Which type of nuclear force is always attractive?

(a) Majorana (b) Bartlett (c) Heisenberg (d) Wigner

Exercise M-3.7: Arrange the nuclides $^{14}_{7}N$, $^{16}_{8}O$, and $^{17}_{9}F$ in the decreasing order of neutron separation energies.

(a) $^{14}_{7}N$, $^{16}_{8}O$, and $^{17}_{9}F$ (b) $^{17}_{9}F$, $^{14}_{7}N$, and $^{16}_{8}O$

(c) $^{17}_{9}F$, $^{16}_{8}O$, and $^{14}_{7}N$ (d) $^{16}_{8}O$, $^{17}_{9}F$, and $^{14}_{7}N$

Exercise M-3.8: Deuteron is considered to be an admixture of S and D states. Deuteron could not be an admixture of S and P states because such an admixture will not reproduce the

(a) mass of deuteron (b) spin of deuteron

(c) quadrupole moment of deuteron (d) parity of deuteron

Exercise M-3.9: Deuteron could not be an admixture of S and F($\ell = 4$) states because such an admixture will not reproduce the

(a) mass of deuteron (b) spin of deuteron

(c) quadrupole moment of deuteron (d) parity of deuteron

Exercise M-3.10: Strong spin orbit coupling is required to explain

(a) stability of nuclei (b) higher magic numbers

(c) n-p scattering data (d) deuteron quadrupole moment

data 13.04.2020

135) A) Books-68-2020/Nuclear Physics.../
Sanet.st-Nuclear Physics Pearson 2017.pdf

Pág. 116(129)

4.2 Schrödinger Equations for a System of two spinless particles in spherical polar coordinates

Fig 4.1 Relation between the spherical polar and Cartesian coordinates.

B) Books_68_2020/
Orbital Mechanics for Engineering Students/
Orbital Mechanics for Engineering Students... Curtis(z-lib.org).pdf
(2020)

Pág. 393(400)

Sphere of Influence

Fig 8.5 Decrease of gravitational force with distance from a planet's surface.

Fig. 8.6 Relative position and gravitational force vectors among the three bodies.

Pág. 292(299) → Fig 6.4

4

Quantum Mechanical Analysis of Some Nuclear Systems

Information about the nuclear force that binds nucleons in the nucleus is obtained essentially from four types of studies: (1) the systematic of stable nuclei, (2) properties of the ground state and excited states of stable and long-lived radioactive nuclei, (3) nucleon–nucleon scattering, and (4) nuclear reactions. The other possible approach is to assume a certain nature for the nuclear force and theoretically generate a nucleus with given numbers of neutrons and protons and then to compare the properties of the theoretically generated nucleus with the one actually existing in nature. Refinements in the assumptions made about the nuclear force may then be carried out to reproduce the experimental data. Both these approaches have been followed in nuclear physics. However, from the perspective of developing a theoretical model of the nucleus, light stable nuclei with only a few nucleons are best suited. The lightest nucleus is the nucleus of hydrogen atom that contains a single proton. There is no excited state of hydrogen nucleus and, therefore, not much information about the nature of nuclear force can be obtained from it. Electron scattering experiments have indicated that the positive charge distribution in proton has a radius of about 0.8 F. Deuteron is the next stable nucleus that has one neutron and one proton. Deuteron does not have any excited state as well. However, the measured values of the spin, magnetic dipole moment, electric quadrupole moment, and the electromagnetic radius of deuteron have given valuable information about the nature of the nuclear force.

Successful application of quantum mechanical approach to the hydrogen atom by Bohr and subsequent development of detailed quantum mechanical frame work by Heisenberg, Schrödinger, and others established beyond doubt its power of describing subatomic systems. In this chapter, elementary quantum mechanical treatment of deuteron, nucleon–nucleon scattering, and barrier transmission is developed.

4.1 DEUTERON AND ITS PROPERTIES

Deuteron is the only stable system with two nucleons, a proton, and a neutron that is found in a free state in nature. Study of the deuteron is, therefore, of considerable value in getting information about the nature and properties of nucleon–nucleon force. Both di-neutron (a system with two neutrons) and di-proton are unstable and not found in nature. The presently accepted values of deuteron parameters along with brief notes on the method of their measurement and significance are as follows:

1. Deuteron represented as ${}^{2}_{1}\mathrm{D}$ or ${}^{2}_{1}\mathrm{H}$ is found only in ground state and its excited states are unstable.
2. **Mass:** $3.34358325 \times 10^{-27}$ kg, or 2.01355532 u; or 1875.61282 MeV.
   It is determined by penning trap technique.
3. **Binding energy:** 2.225 MeV.
   It is determined from the energy of gamma rays emitted in the radiative capture reaction ${}^{1}_{0}\mathrm{n} + {}^{1}_{1}\mathrm{p} = {}^{2}_{1}\mathrm{D} + \gamma$ (also written in short as n (p, γ) D). It may be noted that the binding energy per nucleon (B/A) in deuteron is only 1.11 MeV, which is much smaller than the average binding energy per nucleon of ≈8 MeV in the case of heavy nuclei. This indicates that the deuteron is a loosely bound system.
4. **Spin (or angular momentum) $J = 1$.**
   The spin of the deuteron has been determined by counting the number of components in hyperfine multiplet.

   Deuteron contains one neutron and one proton, each with spin ½. A spin value of 1 for deuteron is possible when the spins of the proton and the neutron are parallel and point in the same direction. As the spin $J = 1$ can have three possible values of the magnetic quantum number m_J, the deuteron is a triplet state. The singlet state of a neutron and a proton system, where the spins of the two particles are opposite to each other giving a resultant $J = 0$, does not exist. It means that ^{3}(n, p) state (deuteron) is a bound stable state and the ^{1}(n, p)-singlet state is unbound.
5. **Magnetic dipole moment μ_d:** 0.857393 nm (nuclear magneton).
   It is determined using magnetic resonance method.
   It may be noted that the magnetic dipole moment of deuteron is nearly, *but not exactly equal to* the sum of the magnetic moments of neutron and proton.

   $$\mu_n = -1.913 \text{ nm}; \mu_p = 2.792 \text{ nm. Hence, } \mu_n + \mu_p = 0.879 \text{ nm} \neq \mu_d.$$

   The fact that the magnetic moment of deuteron is nearly equal to the sum of the inherent magnetic moments of neutron and proton suggests that there is almost no contribution to the magnetic moment μ_d by the orbital motion of the proton. That means in deuteron, neutron, and proton are essentially in a state of a relative motion for which the relative orbital angular momentum ℓ is zero. The state of motion for which $\ell = 0$, is called the S-state. In deuteron, the neutron and the proton are in ${}^{3}\mathrm{S}_{1}$-state, the superscript 3 giving the multiplicity of the state and subscript 1 indicating the J value. The ${}^{3}\mathrm{S}_{1}$-state is spherically symmetric.
6. **Electric quadrupole moment Q_d:** 2.86×10^{-3} b (b = barn; 1b = 10^{-24} cm^{2}).
   The electric quadrupole moment of deuteron has been determined from the shift in the energy of lines in the rotational band spectrum. As Q_d is positive, the deuteron has a positive charge distribution of prolate shape.

   Now, there is a contradiction. As mentioned earlier, from the magnitude of the magnetic moment it appears that deuteron is a ${}^{3}\mathrm{S}_{1}$-state. However, S-state is spherically symmetric and, therefore, the quadrupole moment of deuteron should have been zero. As will be discussed later, the deuteron is not a pure S-state. It is an admixture of ${}^{3}\mathrm{S}_{1}$ and ${}^{3}\mathrm{D}_{1}$ states.

If it is assumed that for about 96% time of the motion the neutron and proton remain in $^3S_1$-state but for a very small time, about 4% of the time period, they are in the $^3D_1$-state for which the relative angular momentum $\ell = 2$, but $J = 1$, then the magnitudes of both Q_d and μ_d can be satisfactorily reproduced. When in $^3D_1$-state, the magnetic moment of proton due to orbital motion also contributes to the magnetic moment of the deuteron along with the inherent magnetic moments of neutron and the proton. This explains why $\mu_d \neq \mu_p + \mu_n$. Also, when deuteron changes from $^3S_1$ to $^3D_1$, the spin of the deuteron flips to keep the value of $J = (\ell + s) = 1$. This means that the directions of ℓ and s are parallel but opposite. As such, the magnetic moment due to the orbital motion of proton is negative. This explains why $\mu_d < (\mu_n + \mu_p)$.

7. **Radius of the deuteron:** The root mean square electromagnetic radius, which is defined by the relation $[<r^2>]^{1/2} = [\int r^2\rho(r)d^3r/\rho(r)d^3r]^{1/2}$, where $\rho(r)$ is the charge density at a distance r from the nuclear centre, has been determined from electron scattering and is about 2.1 F for deuteron.
8. **Parity of deuteron:** Parity of deuteron, deduced from deuteron reactions and theoretical consideration, is even.
9. **Statistics:** Deuteron with two nucleons (in $J = 1$ state) follows Bose–Einstein statistics.

4.2 SCHRÖDINGER EQUATIONS FOR A SYSTEM OF TWO SPINLESS PARTICLES IN SPHERICAL POLAR COORDINATES

We consider the general case of a system of two *spinless* particles of mass M_1 and M_2 with total energy W and potential energy U. The two-body problem can be reduced to a one-body problem by using the centre of mass (CM) coordinates. In CM system, one considers the motion of only one particle of reduced mass M, where $M = \frac{M_1 M_2}{M_1 + M_2}$ moving with the velocity V with respect to the centre of mass. Here, V is the relative velocity of M_1 and M_2. The Schrödinger equation for the particle of reduced mass M in Cartesian coordinates is

$$\nabla^2\psi + \left(\frac{2M}{\hbar^2}\right)[W - U]\psi = 0 \tag{4.1}$$

Here, ψ is the space part of the wave function. Next step is to write Eq. (4.1) in spherical polar coordinates. Spherical polar coordinates are better suited for problems related to the nucleus because of the spherical symmetry. The three spherical polar coordinates r, θ, and φ are related to the Cartesian coordinates x, y, and z through the relations $x = r \sin\theta \cos\varphi$; $y = r \sin\theta \sin\varphi$; and $z = r \cos\theta$, as shown in figure 4.1, the polar axis is taken in z-direction.

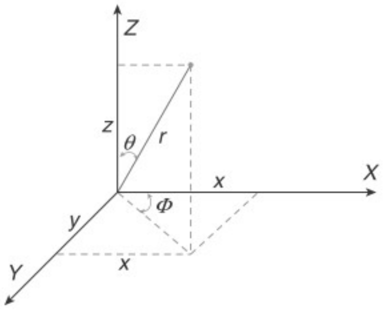

Figure 4.1 *Relation between the spherical polar and Cartesian coordinates*

Schrödinger Eq. (4.1) in spherical polar coordinates is

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial \psi}{\partial r}\right)+\frac{1}{r^2\sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta\frac{\partial \psi}{\partial \theta}\right)+\frac{1}{r^2 sin^2\theta}\frac{\partial^2 \psi}{\partial \varphi^2}+\frac{2M}{\hbar^2}[W-U]\psi=0 \qquad (4.2)$$

We now make another assumption that the potential U is a function of r alone and does not depend on θ and φ, that is $U = U(r)$, then the wave function $\psi(r, \theta, \varphi)$ may be written in two parts: $\chi(r)$, a function of r only and $Y(\theta, \varphi)$ a function of (θ, φ) only. Therefore, substituting $\psi(r, \theta, \varphi) = \chi(r)Y(\theta, \varphi)$, and dividing by ψ, we get

$$\frac{1}{\chi(r)}\frac{d}{dr}\left(r^2\frac{d}{dr}\chi(r)\right)+\frac{2Mr^2}{\hbar^2}[W-U(r)]=-\frac{1}{Y(\theta,\varphi)}\left[\frac{1}{\sin\theta}\frac{\partial}{\partial \boldsymbol{\theta}}\{Y(\theta,\varphi)\}+\frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\varphi^2}\{Y(\theta,\varphi)\}\right] \qquad (4.3)$$

In Eq. (4.3), LHS is a function of r alone and RHS is a function only of θ and φ. As r, θ, and φ are all independent variables, Eq. (4.3) will hold only if the two sides are equal to some constant. Let the constant be $\ell\,(\ell + 1)$. Thus, the radial part of Eq. (4.3) may be written as

$$\frac{1}{\chi(r)}\frac{d}{dr}\left(r^2\frac{d}{dr}\chi(r)\right)+\frac{2Mr^2}{\hbar^2}[W-U(r)]=\ell(\ell+1)$$

or
$$\frac{1}{\chi(r)}\frac{d}{dr}\left(r^2\frac{d}{dr}\chi(r)\right)+\frac{2Mr^2}{\hbar^2}[W-U(r)-\{(\ell(\ell+1)\hbar^2/2Mr^2\}]=0 \qquad (4.4a)$$

And the angular part as

$$-\frac{1}{Y(\theta,\varphi)}\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\{Y(\theta,\varphi)\}+\frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\varphi^2}\{Y(\theta,\varphi)\}\right]=\ell(\ell+1) \qquad (4.4b)$$

The constant of separation $\ell(\ell + 1)$ has been chosen with a purpose. Let us recall that Eq. (4.4) in classical framework represents the motion of a particle of mass M, total energy W, and potential energy $U(r)$ moving with a velocity V in the centre of mass. V is the relative velocity of the particles. The kinetic energy of the particle is $\frac{1}{2}MV^2$. The total energy W is the sum of the kinetic and potential energy. Therefore,

$$W=\left(\frac{1}{2}\right)MV^2+U(r) \qquad (4.5)$$

As $U(r)$ is positive, it is derived from a force of repulsion between the two particles. A rough sketch of the classical motion of the particle of mass M around the centre of mass C is shown in figure 4.2.

Let the distance of closest approach of the particle to the centre of mass be x when the particle is at point P. Now, at point P the velocity V of the particle is normal to x and hence the angular momentum J of the particle is given by

$$J = MVx \quad \text{or} \qquad (4.6)$$

$$V = J/Mx \qquad (4.7)$$

Substituting the value of V in Eq. (4.5), for the total energy of the particle, we get

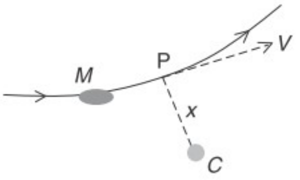

Figure 4.2 *Classical motion of the particle of reduced mass M about the centre of mass C*

$$
\begin{aligned}
W\,(\text{at P}) &= \frac{1}{2}(MJ^2/M^2x^2) + U(r = x) \\
&= (J^2/2Mx^2) + U(r = x) = (J^2/2I) + U(r = x) \qquad (4.8)
\end{aligned}
$$

Here, I represents the moment of inertia of the particle of mass M about an axis passing through the centre of mass. In Eq. (4.8), the kinetic energy of the linear motion of the particle is replaced by the energy of rotational motion with angular momentum J along an axis perpendicular to the polar axis and passing through the centre of mass. As no torque is present in the system, the angular momentum of the system is conserved. It means that the value of J remains same irrespective of the instantaneous position of the particle. Hence, Eq. (4.8) will remain valid for all values of r. In addition, in going from classical to quantum mechanical frame work $J^2 = \ell(\ell + 1)$ and therefore at any instance, the total energy of the particle may be written as

$$W = [\{(\ell(\ell + 1)\,\hbar^2/2Mr^2\} + U(r)] \qquad (4.9)$$

From Eq. (4.9), it may be observed that the sign of both the terms $\{(\ell(\ell + 1)\hbar^2/2Mr^2\}$ and $U(r)$ is the same. A potential with positive sign is derived from a repulsive force and, therefore, the first term in Eq. (5.9) also represents a potential that corresponds to a repulsive force. The first term, which has the dimensions of energy, is called the centrifugal potential because like the classical centrifugal force it has the tendency to pull the particle away from the centre of mass. Larger the value of the angular momentum ℓ more will be the centrifugal force.

Coming back to Eq. (4.4a), for the ground state of the system, which is the lowest energy state, ℓ should be zero. *It may, therefore, be said that for the ground state of a two-body system of spinless particles in a central potential, the orbital angular momentum of the relative motion of the particles is zero and the system is said to be in S state.* States of a system with relative angular momentum $\ell = 0, 1, 2, 3 \ldots$ are, respectively, designated as S, P, D, … states.

Finally, the Schrödinger equation for the ground state of a system of two spinless particles of reduced mass M, total energy W, and in a central potential $U(r)$, which depends only on r is

$$\frac{1}{\chi(r)}\frac{d}{dr}\left(r^2\frac{d}{dr}\chi(r)\right) + \frac{2Mr^2}{\hbar^2}[W - U(r)] = 0 \qquad (4.10)$$

To simplify the calculations, we substitute

$$u(r) = r\chi(r) \qquad (4.11)$$

The function $u(r)$ is called the reduced radial wave function. With this substitution, Eq. (4.10) reduces to

$$\left(\frac{d^2u(r)}{dr^2}\right) + \frac{2Mr^2}{\hbar^2}[W - U(r)]\,u(r) = 0 \qquad (4.12)$$

Equation (4.12) is the *reduced radial equation* for the ground state of a system of two spinless particles of reduced mass M, total energy W, and potential energy $U(r)$, which is a function of r alone. Eq. (4.12) is quite general, the only conditions imposed so far are (i) the particles are spinless and (ii) that $U(r)$ is a spherically symmetric function of r. It may be remarked that in Eq. (4.12), if W is positive it refers to a system of two particles that are free and not bound. A positive value of $U(r)$ means a repulsive force between the two particles. A negative value of

W means a bound system and a negative value of $U(r)$ refers to a force of attraction between the particles.

Next, we consider the angular part of the wave function described by Eq. (4.4b)). If it is further assumed that the function $Y(\theta, \varphi)$ may be written as the multiplication of two independent functions $\Theta(\theta)$ and $\Phi(\varphi)$, which are functions only of θ and φ, respectively, then on substituting $Y(\theta, \varphi) = \Theta(\theta)\, \Phi(\varphi)$ in Eq. (4.4b)), we get

$$\sin^2\theta\left[\frac{1}{\Theta\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right)+\ell(\ell+1)\right]=-\frac{1}{\Phi}\frac{d^2\Phi}{d\phi^2}$$

$$= m^2 \text{ (separation constant)} \tag{4.4c}$$

In Eq. (4.4c), the LHS and the RHS are independent of each other and are, therefore, equal to some constant of separation, for instance, m^2. Eq. (4.4c) gives two equations, one polar and the other azimuthal as

$$\left[\frac{1}{\Phi\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Phi}{d\theta}\right)+\ell\left(\ell+1\right)\right]-\frac{m^2}{\sin^2\theta}=0 \tag{4.4d}$$

And
$$\frac{d^2\Phi}{d\phi^2}+m^2\Phi=0 \tag{4.4e}$$

Equation (4.4d) is the well-known Legendre's equation. Being a second-order differential equation, it has two independent solutions. Except for the values of the separation constant $\ell(\ell+1) = 0, 2, 6 \ldots$ for which ℓ has the values 0, 1, 2, 3, … both the solutions of Legendre's equation becomes infinite for $\cos\theta = \pm 1$ and, therefore, not acceptable. It can be shown that for $\ell = 0$ and for other positive integer values and for $\ell \geq |m|$, one of the solution of Legendre's equation remains finite, single valued, and continuous, that is a well-behaved function.

If $m = 0$, from Eq. (4.4e), $\Phi(\phi)$ becomes constant. Under these conditions, the acceptable Legendre's solutions are called Legendre's polynomials and are represented by $P_\ell(\cos\theta)$. The expressions for the first few Legendre's polynomials are as follows:

$$P_0(\cos\theta) = 1$$

$$P_1(\cos\theta) = \cos\theta;$$

$$P_2(\cos\theta) = \frac{1}{2}(3\cos^2\theta - 1) = \frac{1}{4}(3\cos 2\theta + 1)$$

$$P_3(\cos\theta) = \frac{1}{2}(5\cos^3\theta - 3\cos\theta) = \frac{1}{8}(5\cos 3\theta + 3\cos\theta)$$

The azimuthal Eq. (4.4e) has the solution

$$\Phi_{\mathrm{m}}(\varphi) = e^{\pm m\varphi} \tag{4.4f}$$

As an increase in φ by 2π should bring back to the starting position in space and must give the same value of Φ, the allowed values of m, the magnetic quantum number, may be negative or positive integers.

In the case when the magnetic quantum number m is not zero but $|m| \leq \ell$, solution of Legendre's equation are called associate Legendre's polynomials and are represented by $P^{\mathrm{l}}_{\mathrm{m}}(\cos\theta)$.

$$P^{\ell}{}_{\mathrm{m}}(\cos\theta) = \left[(1-cos^2\theta)^{\frac{|m|}{2}}\right] \frac{d^{|m|}(P_\ell \cos\theta)}{d\cos\theta^{|m|}}$$

The angular part of the space wave function $Y = \Theta(\theta)\Phi(\varphi)$ depends on l, m, θ, and φ and may be written as

$$Y_{\ell\mathrm{m}}(\theta, \varphi) = N_{\ell\mathrm{m}}\, P_\ell^m(\cos\,\theta)\, e^{im\varphi}$$

Here, $N_{\ell\mathrm{m}}$ is the normalization constant. $Y_{\ell\mathrm{m}}(\theta, \varphi)$ are called the spherical harmonics.

4.3 SIMPLE QUANTUM MECHANICAL DESCRIPTION OF DEUTERON

Now, we move to the specific case of the deuteron and rewrite Eq. (4.12). In the case of deuteron, the reduced mass $M = \{(M_\mathrm{n}M_\mathrm{p})/(M_\mathrm{n} + M_\mathrm{p})\}$, where M_n and M_p are, respectively, the masses of neutron and proton. As the masses of neutron and proton are nearly equal, the reduced mass M is nearly equal to half the mass of a neutron or a proton. The deuteron is a bound system and, therefore, $W = -E_\mathrm{b}$, where E_b is the binding energy of the ground state. As in deuteron the nuclear force between neutron and proton is attractive, we represent it by a negative square-well (also called rectangular-well) potential of depth V_0 and width b (figure 4.3).

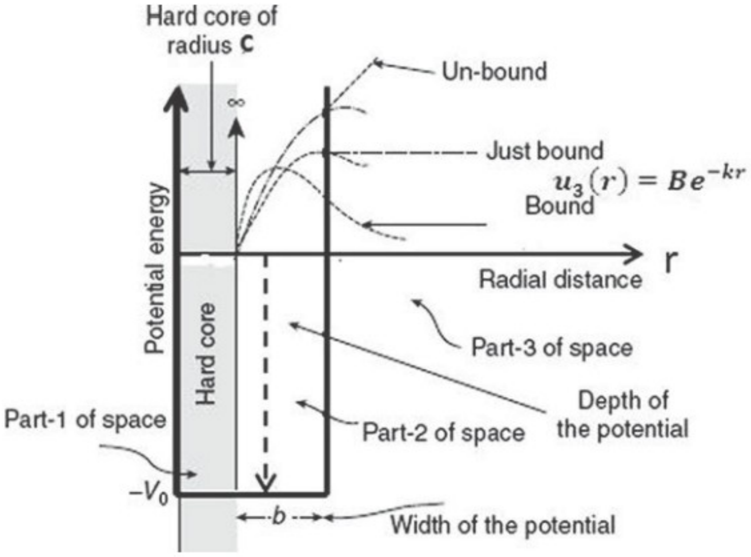
Hard core of
radius C
∞
Un-bound
Just bound
$u_3(r) = Be^{-kr}$
Bound
r
Radial distance
Potential energy
Hard core
Part-3 of space
Depth of
the potential
Part-1 of space
Part-2 of space
$-V_0$
b
Width of the potential

Figure 4.3 *Rectangular deuteron potential*

In figure 4.3, the distance c from the origin is the nuclear hard core. High-energy scattering experiments have indicated that in a nucleus two nucleons cannot come closer than c. The value of c is around 0.4 F. We chose a square-well potential to represent the deuteron as it is the simplest potential. The potential shown in figure 4.3 may be mathematically represented as

/ /)*+,)))-) # - /
0 " @ " 0" * 3 0 5 3"
/)*+,)))-) # - / /7 * 0 + 7 3"
/#)*+,)))-) # - / 9 0
" /7 J " ; 3"
(#5# % 3 '

The phase angle between earth and Mars when the spacecraft reaches Mars is given by Eq. (8.13).

$$\phi_f = \pi - n_{\text{earth}} t_{12} = \pi - 0.017202 \cdot 258.82 = -1.3107\,\text{rad}$$

Since $n_{\text{earth}} > n_{\text{Mars}}$, we choose Eq. (8.16) to find the wait time

$$t_{\text{wait}} = \frac{-2\phi_f - 2\pi N}{n_{\text{Mars}} - n_{\text{earth}}} = \frac{-2(-1.3107) - 2\pi N}{0.0091327 - 0.017202} = 778.65N - 324.85\,\text{days}$$

$N = 0$ yields a negative value, which we cannot accept. Setting $N = 1$, we find

$$\boxed{t_{\text{wait}} = 453.8\,\text{days}}$$

This is the minimum wait time (1.24 years). Obviously, we could set $N = 2, 3, \ldots$ to obtain longer wait times.

For a spacecraft to depart on a mission to Mars by means of a Hohmann (minimum energy) transfer, the phase angle between earth and Mars must be that given by Eq. (8.12). Using the results of Example 8.2, we find it to be

$$\phi_0 = \pi - n_{\text{Mars}} t_{12} = \pi - 0.0091327 \cdot 258.82 = 0.7778\,\text{rad} = 44.57°$$

This opportunity occurs once every synodic period, which we found to be 2.13 years in Example 8.1. In Example 8.2, we found that the time to fly to Mars is 258.8 days, followed by a wait time of 453.8 days, followed by a return trip time of 258.8 days. Hence, the minimum total time for a manned Mars mission, using Hohmann transfers is

$$t_{\text{total}} = 258.8 + 453.8 + 258.8 = 971.4\,\text{days} = 2.66\,\text{years}$$

8.4 SPHERE OF INFLUENCE

The sun, of course, is the dominant celestial body in the solar system. It is over 1000 times more massive than the largest planet, Jupiter, and has a mass of over 300,000 earths. The sun's gravitational pull holds all the planets in its grasp according to Newton's law of gravity (Eq. 1.31). However, near a given planet, the influence of its own gravity exceeds that of the sun. For example, at its surface the earth's gravitational force is over 1600 times greater than the sun's. The inverse-square nature of the law of gravity means that the force of gravity F_g drops off rapidly with distance r from the center of attraction. If F_{g0} is the gravitational force at the surface of a planet with radius r_0, then Fig. 8.5 shows how rapidly the force diminishes with distance. At 10 body radii, the force is 1% of its value at the surface. Eventually, the force of the sun's gravitational field overwhelms that of the planet.

To estimate the radius of a planet's gravitational sphere of influence, consider the three-body system comprising a planet p of mass m_p, the sun s of mass m_s, and a space vehicle v of mass m_v illustrated in Fig. 8.6. The position vectors of the planet and spacecraft relative to an inertial frame centered at the sun are $\mathbf{R}$ and $\mathbf{R}_v$, respectively. The position vector of the space vehicle relative to the planet is $\mathbf{r}$. (Throughout this chapter we will use uppercase letters to represent position, velocity, and acceleration measured relative to the sun and lowercase letters when they are measured relative to a planet.) The gravitational force exerted on the vehicle by the planet is denoted $\mathbf{F}_p^{(v)}$, and that exerted by the

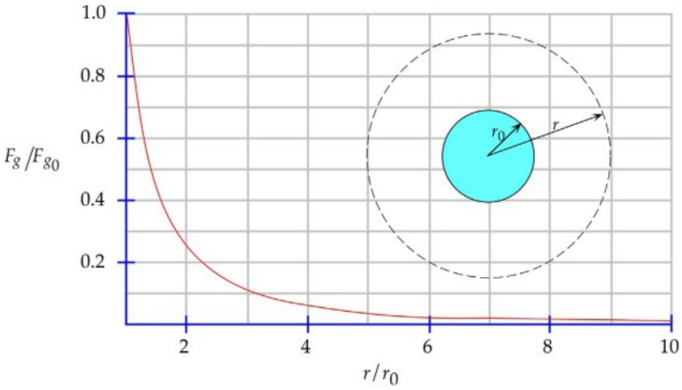

FIG. 8.5

Decrease of gravitational force with distance from a planet's surface.

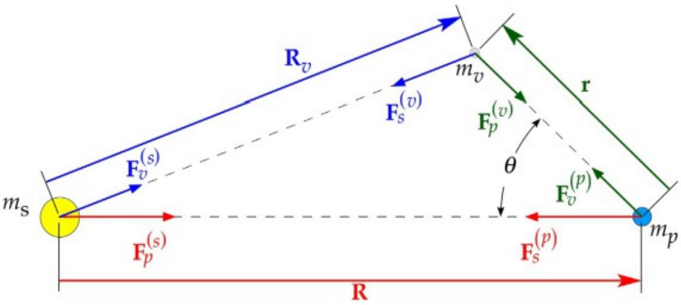
$\mathbf{R}_v$
m_v
$\mathbf{F}_s^{(v)}$
$\mathbf{F}_p^{(v)}$
$\mathbf{r}$
$\mathbf{F}_v^{(s)}$
θ
$\mathbf{F}_v^{(p)}$
m_s
m_p
$\mathbf{F}_p^{(s)}$
$\mathbf{F}_s^{(p)}$
$\mathbf{R}$

FIG. 8.6

Relative position and gravitational force vectors among the three bodies.

sun is $\mathbf{F}_s^{(v)}$. Likewise, the forces on the planet are $\mathbf{F}_s^{(p)}$ and $\mathbf{F}_v^{(p)}$, whereas on the sun we have $\mathbf{F}_v^{(s)}$ and $\mathbf{F}_p^{(s)}$. According to Newton's law of gravitation (Eq. 2.10), these forces are

$$\mathbf{F}_p^{(v)} = -\frac{Gm_v m_p}{r^3}\mathbf{r} \tag{8.18a}$$

$$\mathbf{F}_s^{(v)} = -\frac{Gm_v m_s}{{R_v}^3}\mathbf{R}_v \tag{8.18b}$$

$$\mathbf{F}_s^{(p)} = -\frac{Gm_p m_s}{R^3}\mathbf{R} \tag{8.18c}$$

Observe that

$$\mathbf{R}_v = \mathbf{R} + \mathbf{r} \tag{8.19}$$

From Fig. 8.6 and the law of cosines we see that the magnitude of $\mathbf{R}_v$ is

$$R_v = \left(R^2 + r^2 - 2Rr\cos\theta\right)^{1/2} = R\left[1 - 2(r/R)\cos\theta + (r/R)^2\right]^{1/2} \tag{8.20}$$

We expect that within the planet's sphere of influence, $r/R \ll 1$. In that case, the terms involving r/R in Eq. (8.20) can be neglected, so that, approximately,

$$R_v = R \tag{8.21}$$

The equation of motion of the spacecraft relative to the sun-centered inertial frame is

$$m_v \ddot{\mathbf{R}}_v = \mathbf{F}_s^{(v)} + \mathbf{F}_p^{(v)}$$

Solving for $\ddot{\mathbf{R}}_v$ and substituting the gravitational forces given by Eqs. (8.18a) and (8.18b), we get

$$\ddot{\mathbf{R}}_v = \frac{1}{m_v}\left(-\frac{Gm_v m_s}{{R_v}^3}\mathbf{R}_v\right) + \frac{1}{m_v}\left(-\frac{Gm_v m_p}{r^3}\mathbf{r}\right) = -\frac{Gm_s}{{R_v}^3}\mathbf{R}_v - \frac{Gm_p}{r^3}\mathbf{r} \tag{8.22}$$

Let us write this as

$$\ddot{\mathbf{R}}_v = \mathbf{A}_s + \mathbf{P}_p \tag{8.23}$$

where

$$\mathbf{A}_s = -\frac{Gm_s}{{R_v}^3}\mathbf{R}_v \qquad \mathbf{P}_p = -\frac{Gm_p}{r^3}\mathbf{r} \tag{8.24}$$

$\mathbf{A}_s$ is the primary gravitational acceleration of the vehicle due to the sun, and $\mathbf{P}_p$ is the secondary or perturbing acceleration due to the planet. The magnitudes of $\mathbf{A}_s$ and $\mathbf{P}_p$ are

$$A_s = \frac{Gm_s}{R^2} \qquad P_p = \frac{Gm_p}{r^2} \tag{8.25}$$

where we made use of the approximation given by Eq. (8.21). The ratio of the perturbing acceleration to the primary acceleration is, therefore,

$$\frac{P_p}{A_s} = \frac{\dfrac{Gm_p}{r^2}}{\dfrac{Gm_s}{R^2}} = \frac{m_p}{m_s}\left(\frac{R}{r}\right)^2 \tag{8.26}$$

The equation of motion of the planet relative to the inertial frame is

$$m_p \ddot{\mathbf{R}} = \mathbf{F}_v^{(p)} + \mathbf{F}_s^{(p)}$$

Solving for $\ddot{\mathbf{R}}$, noting that $\mathbf{F}_v^{(p)} = -\mathbf{F}_p^{(v)}$, and using Eqs. (8.18b) and (8.18c) yields

$$\ddot{\mathbf{R}} = \frac{1}{m_p}\left(\frac{Gm_v m_p}{r^3}\mathbf{r}\right) + \frac{1}{m_p}\left(-\frac{Gm_p m_s}{R^3}\mathbf{R}\right) = \frac{Gm_v}{r^3}\mathbf{r} - \frac{Gm_s}{R^3}\mathbf{R} \tag{8.27}$$

Subtracting Eq. (8.27) from Eq. (8.22) and collecting terms, we find

$$\ddot{\mathbf{R}}_v - \ddot{\mathbf{R}} = -\frac{Gm_p}{r^3}\mathbf{r}\left(1 + \frac{m_v}{m_p}\right) - \frac{Gm_s}{{R_v}^3}\left[\mathbf{R}_v - \left(\frac{R_v}{R}\right)^3\mathbf{R}\right]$$

Recalling Eq. (8.19), we can write this as

$$\ddot{\mathbf{r}} = -\frac{Gm_p}{r^3}\mathbf{r}\left(1+\frac{m_v}{m_p}\right) - \frac{Gm_s}{{R_v}^3}\left\{\mathbf{r}+\left[1-\left(\frac{R_v}{R}\right)^3\right]\mathbf{R}\right\} \tag{8.28}$$

This is the equation of motion of the vehicle relative to the planet. By using Eq. (8.21) and the fact that $m_v \ll m_p$, we can write this in approximate form as

$$\ddot{\mathbf{r}} = \mathbf{a}_p + \mathbf{p}_s \tag{8.29}$$

where

$$\mathbf{a}_p = -\frac{Gm_p}{r^3}\mathbf{r} \quad \mathbf{p}_s = -\frac{Gm_s}{R^3}\mathbf{r} \tag{8.30}$$

In this case $\mathbf{a}_p$ is the primary gravitational acceleration of the vehicle due to the planet and $\mathbf{p}_s$ is the perturbation caused by the sun. The magnitudes of these vectors are

$$a_p = \frac{Gm_p}{r^2} \quad p_s = \frac{Gm_s}{R^3}r \tag{8.31}$$

The ratio of the perturbing acceleration to the primary acceleration is

$$\frac{p_s}{a_p} = \frac{Gm_s\dfrac{r}{R^3}}{\dfrac{Gm_p}{r^2}} = \frac{m_s}{m_p}\left(\frac{r}{R}\right)^3 \tag{8.32}$$

For motion relative to the planet, the ratio p_s/a_p is a measure of the deviation of the vehicle's orbit from the Keplerian orbit arising from the planet acting by itself ($p_s/a_p = 0$). Likewise, P_p/A_s is a measure of the planet's influence on the orbit of the vehicle relative to the sun. If

$$\frac{p_s}{a_p} < \frac{P_p}{A_s} \tag{8.33}$$

then the perturbing effect of the sun on the vehicle's orbit around the planet is less than the perturbing effect of the planet on the vehicle's orbit around the sun. We say that the vehicle is therefore within the planet's sphere of influence. Substituting Eqs. (8.26) and (8.32) into Eq. (8.33) yields

$$\frac{m_s}{m_p}\left(\frac{r}{R}\right)^3 < \frac{m_p}{m_s}\left(\frac{R}{r}\right)^2$$

which means

$$\left(\frac{r}{R}\right)^5 < \left(\frac{m_p}{m_s}\right)^2$$

or

$$\frac{r}{R} < \left(\frac{m_p}{m_s}\right)^{2/5}$$

Let r_{SOI} be the radius of the sphere of influence. Within the planet's sphere of influence, defined by

$$\frac{r_{\text{SOI}}}{R} = \left(\frac{m_p}{m_s}\right)^{2/5} \tag{8.34}$$

the motion of the spacecraft is determined by its equations of motion relative to the planet (Eq. 8.28). Outside the sphere of influence, the path of the spacecraft is computed relative to the sun (Eq. 8.22).

The sphere of influence radius presented in Eq. (8.34) is not an exact quantity. It is simply a reasonable estimate of the distance beyond which the sun's gravitational attraction dominates that of a planet. The spheres of influence of all the planets and the earth's moon are listed in Table A.2.

EXAMPLE 8.3

Calculate the radius of the earth's sphere of influence.

Solution

In Table A.1 we find

$$m_{\text{earth}} = 5.974\left(10^{24}\right)\text{kg}$$

$$m_{\text{sun}} = 1.989\left(10^{30}\right)\text{kg}$$

$$R_{\text{earth}} = 149.6\left(10^{6}\right)\text{km}$$

Substituting these data into Eq. (8.34) yields

$$r_{\mathbf{SOI}} = 149.6 \times 10^{6}\left(\frac{5.974 \times 10^{24}}{1.989 \times 10^{24}}\right)^{2/5} = \boxed{925 \times 10^{6}\,\text{km}}$$

Since the radius of the earth is 6378 km,

$$r_{\text{SOI}} = 145\,\text{earth radii}$$

Relative to the earth, its sphere of influence is very large. However, relative to the sun it is tiny, as illustrated in Fig. 8.7.

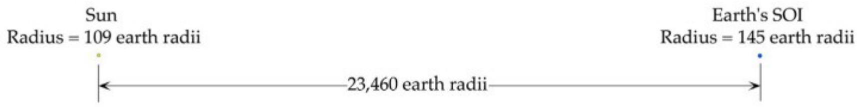

FIG. 8.7

The earth's sphere of influence and the sun, drawn to scale.

8.5 METHOD OF PATCHED CONICS

"Conics" refers to the fact that two-body, or Keplerian, orbits are conic sections with the focus at the attracting body. To study an interplanetary trajectory, we assume that when the spacecraft is outside the sphere of influence of a planet it follows an unperturbed Keplerian orbit around the sun. Because interplanetary distances are so vast, for heliocentric orbits we may neglect the size of the spheres of influence and consider them, like the planets they surround, to be just points in space coinciding with the planetary centers. Within each planetary sphere of influence, the spacecraft travels an unperturbed Keplerian path about the planet. While the sphere of influence appears as a mere speck on the scale of the solar system, from the point of view of the planet it is very large indeed and may be considered to lie at infinity.

To analyze a mission from planet 1 to planet 2 using the method of patched conics, we first determine the heliocentric trajectory, such as the Hohmann transfer ellipse discussed in Section 8.2, that will

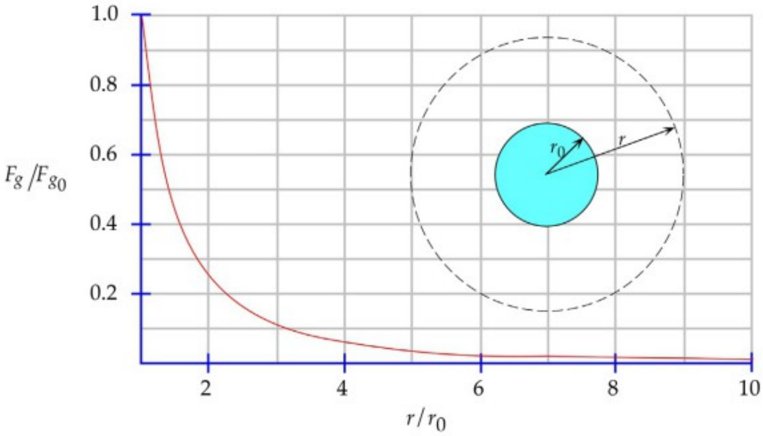
1.0
0.8
0.6
0.4
0.2
F_g / F_{g_0}
2
4
6
8
10
r / r_0
r_0
r

; : " 3 K

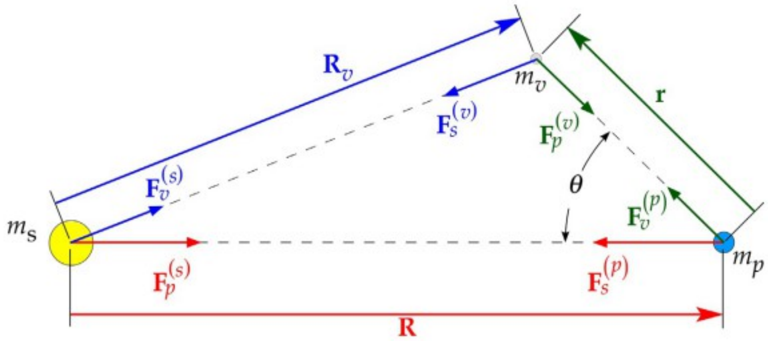
$\mathbf{R}_v$
m_v
$\mathbf{F}_s^{(v)}$
$\mathbf{F}_p^{(v)}$
$\mathbf{r}$
$\mathbf{F}_v^{(s)}$
θ
$\mathbf{F}_v^{(p)}$
m_s
m_p
$\mathbf{F}_p^{(s)}$
$\mathbf{F}_s^{(p)}$
$\mathbf{R}$

+ : 3 " : : / "

(5 55 9 6 %

6.3 HOHMANN TRANSFER

Walter Hohmann (1880–1945) was a German engineer whose interest in early rocketry led him to discover the orbital transfer maneuver that now bears his name. In a book published in 1925 (Hohmann, 1925), he showed that the Hohmann transfer is the most energy-efficient two-impulse maneuver for transferring between two coplanar circular orbits sharing a common focus. The Hohmann transfer is an elliptical orbit tangent to both circles on its apse line, as illustrated in Fig. 6.2. The periapsis and apoapsis of the transfer ellipse are the radii of the inner and outer circles, respectively. Obviously, only one-half of the ellipse is flown during the maneuver, which can occur in either direction, from the inner to the outer circle, or vice versa.

It may be helpful in sorting out orbit transfer strategies to use the fact that the energy of an orbit depends only on its semimajor axis a. Recall that for an ellipse (Eq. 2.80), the specific energy is negative,

$$\varepsilon = -\frac{\mu}{2a}$$

Increasing the energy requires reducing its magnitude, to make ε less negative. Therefore, the larger the semimajor axis, the more energy the orbit has. In Fig. 6.2, the energies increase as we move from the inner circle to the outer circle.

Starting at A on the inner circle, a velocity increment Δv_A in the direction of flight is required to boost the vehicle onto the higher energy elliptical trajectory. After coasting from A to B, another forward velocity increment Δv_B places the vehicle on the still higher energy, outer circular orbit. Without the latter delta-v burn, the spacecraft would, of course, remain on the Hohmann transfer ellipse and return to A. The total energy expenditure is reflected in the total delta-v requirement, $\Delta v_{\text{total}} = \Delta v_A + \Delta v_B$.

The same total delta-v is required if the transfer begins at B on the outer circular orbit. Since moving to the lower energy inner circle requires lowering the energy of the spacecraft, the Δv's must be accomplished by retrofires. That is, the thrust of the maneuvering rocket is directed opposite to the flight direction to act as a brake on the motion. Since Δv represents the same propellant expenditure regardless of the direction the thruster is aimed, when summing up Δv's, we are concerned only with their magnitudes.

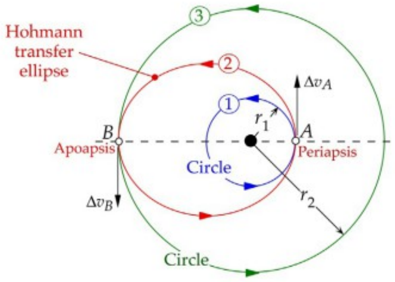

FIG. 6.2

Hohmann transfer.

Recall that the eccentricity of an elliptical orbit is found from its radius to periapsis r_p and its radius to apoapsis r_a by means of Eq. (2.84),

$$e = \frac{r_a - r_p}{r_a + r_p}$$

The radius to periapsis is given by Eq. (2.50),

$$r_p = \frac{h^2}{\mu}\frac{1}{1+e}$$

Combining these last two expressions yields

$$r_p = \frac{h^2}{\mu}\frac{1}{1+\dfrac{r_a - r_p}{r_a + r_p}} = \frac{h^2}{\mu}\frac{r_a + r_p}{2r_a}$$

Solving for the angular momentum h, we get

$$h = \sqrt{2\mu}\sqrt{\frac{r_a r_p}{r_a + r_p}} \tag{6.2}$$

This is a useful formula for analyzing Hohmann transfers, because knowing h we can find the apsidal velocities from Eq. (2.31). Note that for circular orbits ($r_a = r_p$), Eq. (6.2) yields

$$h = \sqrt{\mu r} \quad \text{(circular orbit)}$$

Alternatively, one may prefer to compute the velocities by means of the energy equation (Eq. 2.81) in the form

$$v = \sqrt{2\mu}\sqrt{\frac{1}{r} - \frac{1}{2a}} \tag{6.3}$$

This of course yields Eq. (2.63) for circular orbits.

EXAMPLE 6.1

A 2000-kg spacecraft is in a 480 km by 800 km earth orbit (orbit 1 in Fig. 6.3). Find

(a) The Δv required at perigee *A* to place the spacecraft in a 480 km by 16,000 km transfer ellipse (orbit 2).

(b) The Δv (apogee kick) required at *B* of the transfer orbit to establish a circular orbit of 16,000 km altitude (orbit 3).

(c) The total required propellant if the specific impulse is 300 s.

Solution

Since we know the perigee and apogee of all three of the orbits, let us first use Eq. (6.2) to calculate their angular momenta.

Orbit 1: $r_p = 6378 + 480 = 6858\,\text{km}$ $\quad r_a = 6378 + 800 = 7178\,\text{km}$

$$\therefore h_1 = \sqrt{2 \cdot 398{,}600}\sqrt{\frac{7178 \cdot 6858}{7178 + 6858}} = 52{,}876.5\,\text{km/s}^2 \tag{a}$$

Orbit 2: $r_p = 6378 + 480 = 6858\,\text{km}$ $\quad r_a = 6378 + 16{,}000 = 22{,}378\,\text{km}$

$$\therefore h_2 = \sqrt{2 \cdot 398{,}600}\sqrt{\frac{22{,}378 \cdot 6858}{22{,}378 + 6858}} = 64{,}689.5\,\text{km/s}^2 \tag{b}$$

Orbit 3: $r_a = r_p = 22{,}378\,\text{km}$

$$\therefore h_3 = \sqrt{398{,}600 \cdot 22{,}378} = 94{,}445.1\,\text{km/s}^2 \tag{c}$$

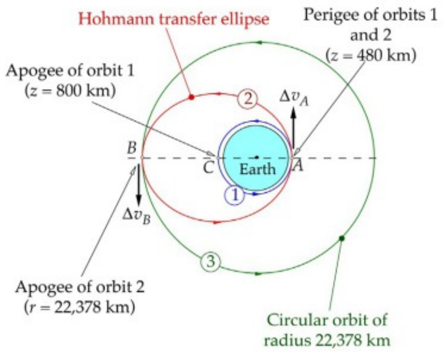
Hohmann transfer ellipse
Perigee of orbits 1 and 2 (z = 480 km)
Apogee of orbit 1 (z = 800 km)
②
Δv_A
B
C
Earth
A
①
Δv_B
③
Apogee of orbit 2 (r = 22,378 km)
Circular orbit of radius 22,378 km

FIG. 6.3

Hohmann transfer between two earth orbits.

(a) The speed on orbit 1 at point *A* is

$$v_A)_1 = \frac{h_1}{r_A} = \frac{52,876}{6858} = 7.71019\text{km/s}$$

The speed on orbit 2 at point *A* is

$$v_A)_2 = \frac{h_2}{r_A} = \frac{64,689.5}{6858} = 9.43271\text{km/s}$$

Therefore, the delta-v required at point *A* is

$$\Delta v_A = v_A)_2 - v_A)_1 = \boxed{1.7225\text{km/s}}$$

(b) The speed on orbit 2 at point *B* is

$$v_B)_2 = \frac{h_2}{r_B} = \frac{64,689.5}{22,378} = 2.89076\text{km/s}$$

The speed on orbit 3 at point *B* is

$$v_B)_3 = \frac{h_3}{r_B} = \frac{94,445.1}{22,378} = 4.22044\text{km/s}$$

Hence, the apogee kick required at point *B* is

$$\Delta v_B = v_B)_3 - v_B)_2 = \boxed{1.3297\text{km/s}}$$

(c) The total delta-v requirement for this Hohmann transfer is

$$\Delta v_{\text{total}} = |\Delta v_A| + |\Delta v_B| = 1.7225 + 1.3297 = 3.0522\text{km/s}$$

According to Eq. (6.1) (converting velocity to m/s),

$$\frac{\Delta m}{m} = 1 - e^{-\frac{3052.2}{300 \cdot 9.807}} = 0.64563$$

Therefore, the mass of propellant expended is

$$\boxed{\Delta m = 0.64563 \cdot 2000 = 1291.3\text{kg}}$$

In the previous example the initial orbit of the Hohmann transfer sequence was an ellipse rather than a circle. Since no real orbit is perfectly circular, we must generalize the notion of a Hohmann transfer to include two-impulse transfers between elliptical orbits that are coaxial (i.e., share the same apse line), as shown in Fig. 6.4. The transfer ellipse must be tangent to both the initial and target ellipses 1 and 2. As can be seen, there are two such transfer orbits, 3 and 3′. It is not immediately obvious which of the two requires the lowest energy expenditure.

To find out which is the best transfer orbit in general, we must calculate the individual total delta-v requirement for orbits 3 and 3′. This requires finding the velocities at A, A', B, and B' for each pair of orbits having those points in common. We employ Eq. (6.2)to evaluate the angular momentum of each of the four orbits in Fig. 6.4.

$$h_1 = \sqrt{2\mu}\sqrt{\frac{r_A r_{A'}}{r_A + r_{A'}}} \quad h_2 = \sqrt{2\mu}\sqrt{\frac{r_B r_{B'}}{r_B + r_{B'}}} \quad h_3 = \sqrt{2\mu}\sqrt{\frac{r_A r_B}{r_A + r_B}} \quad h_{3'} = \sqrt{2\mu}\sqrt{\frac{r_{A'} r_{B'}}{r_{A'} + r_{B'}}}$$

From these we obtain the velocities,

$$\begin{array}{ll} v_A)_1 = \dfrac{h_1}{r_A} & v_A)_3 = \dfrac{h_3}{r_A} \\ v_B)_2 = \dfrac{h_2}{r_B} & v_B)_3 = \dfrac{h_3}{r_B} \\ v_{A'})_1 = \dfrac{h_1}{r_{A'}} & v_{A'})_{3'} = \dfrac{h_{3'}}{r_{A'}} \\ v_{B'})_2 = \dfrac{h_2}{r_{B'}} & v_{B'})_{3'} = \dfrac{h_{3'}}{r_{B'}} \end{array}$$

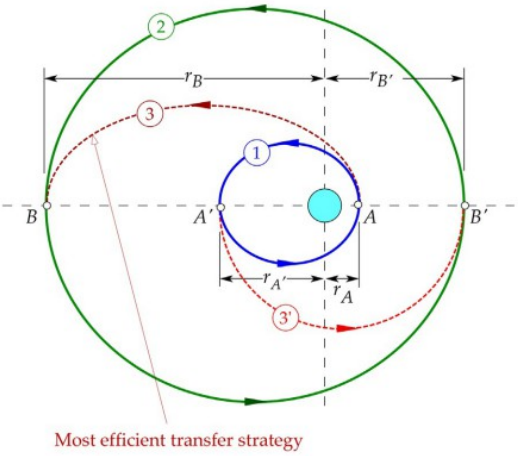

FIG. 6.4

Hohmann transfers between coaxial elliptical orbits. In this illustration, $r_{A'}/r_A = 3$, $r_B/r_A = 8$, and $r_{B'}/r_A = 4$.

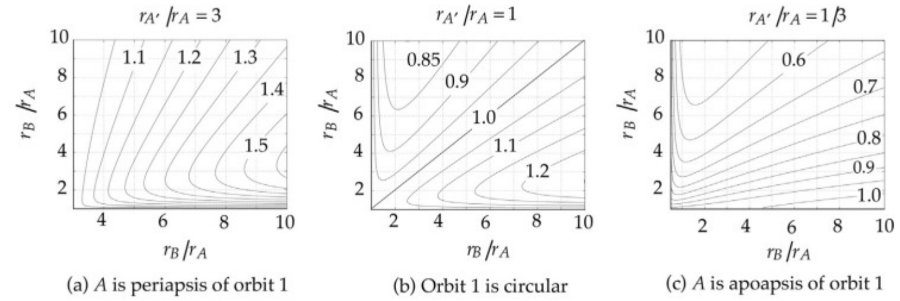

(a) A is periapsis of orbit 1 (b) Orbit 1 is circular (c) A is apoapsis of orbit 1

FIG. 6.5

Contour plots of $\Delta v_{\text{total}})_{3'}/\Delta v_{\text{total}})_3$ for different relative sizes of the ellipses in Fig. 6.4. Note that $r_B > r_{A'}$ and $r_{B'} > r_A$.

These lead to the delta-v's

$$\Delta v_A = \left| v_A)_3 - v_A)_1 \right| \quad \Delta v_B = \left| v_B)_2 - v_B)_3 \right| \quad \Delta v_{A'} = \left| v_{A'})_{3'} - v_{A'})_1 \right| \quad \Delta v_{B'} = \left| v_{B'})_2 - v_{B'})_{3'} \right|$$

and, finally, to the total delta-v requirement for the two possible transfer trajectories,

$$\Delta v_{\text{total}})_3 = \Delta v_A + \Delta v_B \qquad \Delta v_{\text{total}})_{3'} = \Delta v_{A'} + \Delta v_{B'}$$

If $\Delta v_{\text{total}})_{3'}/\Delta v_{\text{total}})_3 > 1$, then orbit 3 is the most efficient. On the other hand, if $\Delta v_{\text{total}})_{3'}/\Delta v_{\text{total}})_3 < 1$, then orbit 3′ is more efficient than orbit 3.

Three contour plots of $\Delta v_{\text{total}})_{3'}/\Delta v_{\text{total}})_3$ are shown in Fig. 6.5, for three different shapes of inner orbit 1 of Fig. 6.4. Fig. 6.5a is for $r_{A'}/r_A = 3$, which is the situation represented in Fig. 6.4, in which point A is the periapsis of the initial ellipse. In Fig. 6.5b $r_{A'}/r_A = 1$, which means the starting ellipse is a circle. Finally, in Fig. 6.5c $r_{A'}/r_A = 1/3$, which corresponds to an initial orbit of the same shape as orbit 1 in Fig. 6.4, but with point A being the apoapsis instead of periapsis.

Fig. 6.5a, for which $r_{A'} > r_A$, implies that if point A is the periapsis of orbit 1, then transfer orbit 3 is the most efficient. Fig. 6.5c, for which $r_{A'} < r_A$, shows that if point A' is the periapsis of orbit 1, then transfer orbit 3′ is the most efficient. Together, these results lead us to conclude that it is most efficient for the transfer orbit to begin at the periapsis on inner orbit 1, where its kinetic energy is greatest, regardless of the shape of the outer target orbit. If the starting orbit is a circle, then Fig. 6.5b shows that transfer orbit 3′ is the most efficient if $r_{B'} > r_B$. That is, from an inner circular orbit, the transfer ellipse should terminate at apoapsis of the outer target ellipse, where the speed is slowest.

If Hohmann transfer is in the reverse direction (i.e., to a lower energy inner orbit), the above analysis still applies, since the same total delta-v is required whether the Hohmann transfer runs forward or backward. Thus, from an outer circle or ellipse, to an inner ellipse the most energy-efficient transfer ellipse terminates at periapsis of the inner target orbit. If the inner orbit is a circle, the transfer ellipse should start at apoapsis of the outer ellipse.

We close this section with an illustration of the careful planning required for one spacecraft to rendezvous with another at the end of a Hohmann transfer.

data 13.04.2020

135) A) Books_68_2020/Nuclear Physics... /
Sanet.st - Nuclear Physics Pearson
2017.pdf

Pág. 116(129)

4.2 Schrödinger Equations for a System of two spinless particles in spherical polar coordinates

Fig 4.1 Relation between the spherical polar and Cartesian coordinates.

B) Books_68_2020/
Orbital Mechanics for Engineering Students/
Orbital Mechanics for Engineering Students... Curtis (z-lib.org).pdf
(2020)

Pág. 393(400)

Sphere of Influence

Fig 8.5 Decrease of gravitational force with distance from a planet's surface.

Fig. 8.6 Relative position and gravitational force vectors among the three bodies.

Pág. 292(299) → Fig 6.4

data 13.04.2020

Fig 6.4 Hohmann transfers between
coaxial elliptical orbits.

C) Books_68_2020/
Einstein in Matrix Form Exact
Derivation of the theory of Special
and General Relativity.../
Einstein_Matrix...pdf

Pág. 34(47)

1.8.2 Lorentz Transformation of
the Maxwell's Equations

Look (1.146) $P(v) = \left(\begin{array}{c|c} & \\ \hline & \end{array}\right)$

136) 20200414

A) Books_68_2020/
Problèmes de Géométrie Analytique
Kletenik pdf/
cours 3_et_4_niv2_rew.pdf

other source: Books_68_2020/Important

Pág. 245 Espace physique et
espace vectoriel

Look: Pag. 255 Réponse ex. 1-11
veja o gráfico (desenho geométrico)

obs: Use este livro.

)*+,)))-) # 0 > 9 = 0>
; : 2 7 3 " 4 + : 7 2 3"
(#% %& !

are gathered and summarized in the following equations:

$$
\begin{array}{rrrrcl}
 & -\dfrac{\partial b_x}{\partial x} & -\dfrac{\partial b_y}{\partial y} & -\dfrac{\partial b_z}{\partial z} & = & 0,\\
+\dfrac{1}{c}\dfrac{\partial b_x}{\partial t} & & -\dfrac{\partial e_z}{\partial y} & +\dfrac{\partial e_y}{\partial z} & = & 0,\\
+\dfrac{1}{c}\dfrac{\partial b_y}{\partial t} & +\dfrac{\partial e_z}{\partial x} & & -\dfrac{\partial e_x}{\partial z} & = & 0,\\
+\dfrac{1}{c}\dfrac{\partial b_z}{\partial t} & -\dfrac{\partial e_y}{\partial x} & +\dfrac{\partial e_x}{\partial y} & & = & 0,
\end{array}
\tag{1.126}
$$

or in matrix form,

$$
\underbrace{\begin{pmatrix} 0 & -\boldsymbol{b}^{\mathsf{T}} \\ \boldsymbol{b} & \boldsymbol{E}_{\times} \end{pmatrix}}_{\overset{\text{def}}{=}\boldsymbol{F}_{E,b}} \gamma \underbrace{\begin{pmatrix} -\frac{1}{c}\frac{\partial}{\partial t} \\ \nabla \end{pmatrix}}_{\vec{\nabla}} = \boldsymbol{o} \in \mathbb{R}^4. \tag{1.127}
$$

The matrix $\boldsymbol{F}_{E,b}$ is called the Maxwell's matrix.

So (1.125) and (1.127) contain the Maxwell's equations in a new four-dimensional form:

$$
\boxed{\boldsymbol{F}_{B,e}\vec{\nabla} = \frac{1}{c}\vec{\boldsymbol{j}} \quad \text{and} \quad \boldsymbol{F}_{E,b}\vec{\nabla} = \boldsymbol{o}.} \tag{1.128}
$$

This form has a great advantage of being invariant when transitioning to another reference system by a Lorentz transformation, i.e. in every inertial system its keeps the same external form. But the quantities appearing in it take different values in each reference system. This is shown in the following.

1.8.2 Lorentz Transformation of the Maxwell's Equations

Multiplying (1.125) from the left with $\boldsymbol{L}^{-1}$ and inserting $\boldsymbol{L}^{-1}\boldsymbol{L} = \boldsymbol{I}$, we obtain

$$
\underbrace{\boldsymbol{L}^{-1}\boldsymbol{F}_{B,e}\boldsymbol{L}^{-1}}_{\boldsymbol{F}'_{B',e'}}\boldsymbol{L}\vec{\nabla} = \frac{1}{c}\underbrace{\boldsymbol{L}^{-1}\vec{\boldsymbol{j}}}_{\vec{\boldsymbol{j}}'}. \tag{1.129}
$$

For the new Faraday's matrix $\boldsymbol{F}'_{B',e'}$ we get

$$
\begin{aligned}
\boldsymbol{F}'_{B',e'} &= \begin{pmatrix} 0 & \boldsymbol{e}'^{\mathsf{T}} \\ -\boldsymbol{e}' & \boldsymbol{B}'_{\times} \end{pmatrix} \\
&= \begin{pmatrix} \gamma_v & \frac{\gamma_v}{c}\boldsymbol{v}^{\mathsf{T}} \\ \frac{\gamma_v}{c}\boldsymbol{v} & \boldsymbol{A}(\boldsymbol{v}) \end{pmatrix} \begin{pmatrix} 0 & \boldsymbol{e}^{\mathsf{T}} \\ -\boldsymbol{e} & \boldsymbol{B}_{\times} \end{pmatrix} \begin{pmatrix} \gamma_v & \frac{\gamma_v}{c}\boldsymbol{v}^{\mathsf{T}} \\ \frac{\gamma_v}{c}\boldsymbol{v} & \boldsymbol{A}(\boldsymbol{v}) \end{pmatrix}.
\end{aligned}
\tag{1.130}
$$

After multiplying the three matrices, we obtain for the bottom left corner vector $\boldsymbol{e}'$ with $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{v})$ and $\gamma = \gamma_v$:

$$\begin{aligned}
-\boldsymbol{e}' &= \frac{\gamma}{c}\boldsymbol{A}\boldsymbol{B}_\times \boldsymbol{v} + \frac{\gamma^2}{c^2}\boldsymbol{v}\boldsymbol{e}^\mathsf{T}\boldsymbol{v} - \gamma\boldsymbol{A}\boldsymbol{e} \\
&= \frac{\gamma}{c}\left(\boldsymbol{B}_\times \boldsymbol{v} + (\gamma-1)\frac{\boldsymbol{v}\overbrace{\boldsymbol{v}^\mathsf{T}\boldsymbol{B}_\times \boldsymbol{v}}^{0}}{v^2}\right) + \frac{\gamma^2}{c^2}\boldsymbol{v}\boldsymbol{e}^\mathsf{T}\boldsymbol{v} - \gamma\boldsymbol{e} - \frac{\gamma(\gamma-1)}{v^2}\boldsymbol{v}\boldsymbol{v}^\mathsf{T}\boldsymbol{e} \\
&= \frac{\gamma}{c}\boldsymbol{B}_\times \boldsymbol{v} - \left(\underbrace{-\frac{\gamma^2}{c^2} + \frac{\gamma^2}{v^2}}_{\frac{1}{v^2}} - \frac{\gamma}{v^2}\right)\boldsymbol{v}^\mathsf{T}\boldsymbol{e}\boldsymbol{v} - \gamma\boldsymbol{e}, \\
\boldsymbol{e}' &= \gamma\left(\boldsymbol{e} - \frac{1}{c}\boldsymbol{B}_\times \boldsymbol{v}\right) + \frac{(1-\gamma)}{v^2}\left(\boldsymbol{v}^\mathsf{T}\boldsymbol{e}\right)\boldsymbol{v},
\end{aligned} \tag{1.131}$$

i.e.

$$\boxed{\boldsymbol{e}' = \gamma\left(\boldsymbol{e} + \frac{1}{c}\boldsymbol{v} \times \boldsymbol{b}\right) + (1-\gamma)\frac{\boldsymbol{v}^\mathsf{T}\boldsymbol{e}}{v^2}\boldsymbol{v}.} \tag{1.132}$$

In the matrix $\boldsymbol{F}'_{B',e'}$ in (1.130), for the right bottom 3×3 sub-matrix $\boldsymbol{B}'_\times$ after performing the matrix multiplication we obtain

$$\boldsymbol{B}'_\times = \boldsymbol{B}_\times + \frac{(\gamma-1)}{v^2}\left(\boldsymbol{B}_\times \boldsymbol{v}\boldsymbol{v}^\mathsf{T} + \boldsymbol{v}\boldsymbol{v}^\mathsf{T}\boldsymbol{B}_\times\right) + \frac{\gamma}{c}\left(\boldsymbol{v}\boldsymbol{e}^\mathsf{T} - \boldsymbol{e}\boldsymbol{v}^\mathsf{T}\right). \tag{1.133}$$

From this matrix equation, the components of the magnetic induction vector $\boldsymbol{b}'$ can be filtered out with the help of the row or column vectors of the unit matrix ($\boldsymbol{i}_j$ is the jth column of the 3×3 identity matrix $\boldsymbol{I}_3$) as follows:

$$\boldsymbol{B}'_\times \overset{\text{def}}{=} \begin{pmatrix} 0 & -b'_z & b'_y \\ b'_z & 0 & -b'_x \\ -b'_y & b'_x & 0 \end{pmatrix}. \tag{1.134}$$

We thus obtain

$$b'_x = \boldsymbol{i}_3^\mathsf{T}\boldsymbol{B}'_\times \boldsymbol{i}_2, \tag{1.135}$$

$$b'_y = \boldsymbol{i}_1^\mathsf{T}\boldsymbol{B}'_\times \boldsymbol{i}_3, \tag{1.136}$$

$$b'_z = \boldsymbol{i}_2^\mathsf{T}\boldsymbol{B}'_\times \boldsymbol{i}_1. \tag{1.137}$$

Using (1.135) for (1.133) results in

$$\begin{aligned}
b'_x = \boldsymbol{i}_3^\mathsf{T}\boldsymbol{B}'_\times \boldsymbol{i}_2 = b_x &+ \frac{(\gamma-1)}{v^2}\left[(-b_y v_1 + b_x v_2)v_2 + v_3(-b_z v_1 + b_x v_3)\right] \\
&+ \frac{\gamma}{c}(v_3 e_y - e_z v_2).
\end{aligned} \tag{1.138}$$

Accordingly, one obtains for the two remaining components

$$\begin{aligned} b'_y = \boldsymbol{i}_1^{\mathsf{T}} \boldsymbol{B}'_{\times} \boldsymbol{i}_3 &= b_y + \frac{(\gamma - 1)}{v^2}\big[(-b_z v_2 + b_y v_3)v_3 + v_1(b_y v_1 - b_x v_2)\big] \\ &\quad + \frac{\gamma}{c}(v_1 e_z - e_x v_3) \end{aligned} \tag{1.139}$$

and

$$b'_z = \boldsymbol{i}_2^{\mathsf{T}} \boldsymbol{B}'_{\times} \boldsymbol{i}_1 = b_z + \frac{(\gamma - 1)}{v^2}\big[(b_z v_1 - b_x v_3)v_1 + v_2(b_z v_2 - b_y v_3)\big] + \frac{\gamma}{c}(v_2 e_x - e_y v_1). \tag{1.140}$$

Summarizing the last terms in (1.138), (1.139) and (1.140) results in the vector product

$$\frac{\gamma}{c}\boldsymbol{e} \times \boldsymbol{v}. \tag{1.141}$$

For the second to last summands one obtains the vector

$$\frac{(\gamma - 1)}{v^2}\begin{pmatrix} (-b_y v_1 + b_x v_2)v_2 + v_3(-b_z v_1 + b_x v_3) \\ (-b_z v_2 + b_y v_3)v_3 + v_1(b_y v_1 - b_x v_2) \\ (b_z v_1 - b_x v_3)v_1 + v_2(b_z v_2 - b_y v_3) \end{pmatrix},$$

and using $0 = b_x v_1^2 - b_x v_1^2$ in the first component, $0 = b_y v_2^2 - b_y v_2^2$ in the second, and $0 = b_z v_3^2 - b_z v_3^2$ in the third,

$$\begin{aligned} &\frac{(\gamma - 1)}{v^2}\begin{pmatrix} (-b_x v_1^2 - b_y v_1 v_2 - b_z v_1 v_3) + (b_x v_1^2 + b_x v_2^2 + b_x v_3^2) \\ (-b_x v_1 v_2 - b_y v_2^2 - b_z v_2 v_3) + (b_y v_1^2 + b_y v_2^2 + b_y v_3^2) \\ (-b_x v_1 v_3 - b_y v_2 v_3 - b_z v_3^2) + (b_z v_1^2 + b_z v_2^2 + b_z v_3^2) \end{pmatrix} \\ &= \frac{(\gamma - 1)}{v^2}\big(-\big(\boldsymbol{b}^{\mathsf{T}}\boldsymbol{v}\big)\boldsymbol{v} + v^2 \boldsymbol{b}\big). \end{aligned} \tag{1.142}$$

Equations (1.141) and (1.142), together with (1.138), (1.139) and (1.140), yield the final result

$$\boxed{\boldsymbol{b}' = \gamma\left(\boldsymbol{b} - \frac{1}{c}\boldsymbol{v} \times \boldsymbol{e}\right) + (1 - \gamma)\frac{\boldsymbol{v}^{\mathsf{T}}\boldsymbol{b}}{v^2}\boldsymbol{v}.} \tag{1.143}$$

One gets the same result for $\boldsymbol{e}'$ and $\boldsymbol{b}'$ by a Lorentz transformation of $\boldsymbol{F}_{E,b}$!

The formulas (1.132) and (1.143) are simplified considerably for small velocities $v \ll c$ because then $\gamma \approx 1$ and one obtains

$$\boldsymbol{e}' = \boldsymbol{e} + \frac{1}{c}\boldsymbol{v} \times \boldsymbol{b}$$

and

$$\boldsymbol{b}' = \boldsymbol{b} - \frac{1}{c}\boldsymbol{v} \times \boldsymbol{e}.$$

The decomposition of the electromagnetic field in an electric and magnetic field has no absolute significance. If there exists, for example, in a reference system $\mathcal{X}$ a purely electrostatic field with $\boldsymbol{b} = \boldsymbol{o}$, then, due to (1.143), a magnetic field $\boldsymbol{b}' = \frac{\gamma}{c}\boldsymbol{v} \times \boldsymbol{e} \neq \boldsymbol{o}$ exists in a reference system $\mathcal{X}'$ that moves with the speed $\boldsymbol{v}$ relative to the reference system $\mathcal{X}$. Physically this means that all charges rest in $\mathcal{X}$. But these charges move relative to $\mathcal{X}'$ with the velocity $\boldsymbol{v}$. So there exists a current in $\mathcal{X}'$ which generates a magnetic field in $\mathcal{X}'$.

The equations (1.132) and (1.143) can be summarized in one equation:

$$\boxed{\begin{pmatrix} \boldsymbol{b}' \\ \boldsymbol{e}' \end{pmatrix} = \underbrace{\left(\begin{array}{c|c} \gamma\boldsymbol{I} + (1-\gamma)\frac{\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}}}{v^2} & -\frac{\gamma}{c}\boldsymbol{V}_{\times} \\ \hline \frac{\gamma}{c}\boldsymbol{V}_{\times} & \gamma\boldsymbol{I} + (1-\gamma)\frac{\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}}}{v^2} \end{array}\right)}_{\boldsymbol{P}(\boldsymbol{v})} \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{e} \end{pmatrix} \in \mathbb{R}^6,} \tag{1.144}$$

with

$$\boldsymbol{V}_{\times} \overset{\text{def}}{=} \begin{pmatrix} 0 & -v_3 & v_2 \\ v_3 & 0 & -v_1 \\ -v_2 & v_1 & 0 \end{pmatrix}.$$

The symmetric 6×6-matrix $\boldsymbol{P}(\boldsymbol{v})$ occurring in (1.144) has a formal similarity with the Lorentz matrix $\boldsymbol{L}$!

It is, of course, true that $\boldsymbol{P}(-\boldsymbol{v}) = \boldsymbol{P}^{-1}(\boldsymbol{v})$, for, with

$$\boldsymbol{V}_{\times}^2 = \boldsymbol{v}\boldsymbol{v}^{\mathsf{T}} - v^2\boldsymbol{I}, \tag{1.145}$$

one obtains easily $\boldsymbol{P}(\boldsymbol{v})\boldsymbol{P}(-\boldsymbol{v}) = \boldsymbol{I}$.

In the special case when only one component of the velocity vector in the x-direction is nonzero, i.e. $\boldsymbol{v} = [v, 0, 0]^{\mathsf{T}}$, one obtains the matrix

$$\boldsymbol{P}(\boldsymbol{v}) = \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \gamma & 0 & 0 & 0 & \frac{\gamma v}{c} \\ 0 & 0 & \gamma & 0 & -\frac{\gamma v}{c} & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{\gamma v}{c} & 0 & \gamma & 0 \\ 0 & \frac{\gamma v}{c} & 0 & 0 & 0 & \gamma \end{array}\right). \tag{1.146}$$

This special symmetric 6×6-matrix can also be found in [33].

1.8.3 Electromagnetic Invariants

For the electromagnetic field quantities $\boldsymbol{e}$ and $\boldsymbol{b}$ one can form *invariants* which do not change during the transition to another inertial system. For the scalar product of the electric field strength $\boldsymbol{e}$ and the magnetic induction $\boldsymbol{b}$ we obtain

$$\begin{aligned}\boldsymbol{e}'^{\mathsf{T}}\boldsymbol{b}' &= \left(\gamma\left(\boldsymbol{e}^{\mathsf{T}} - \frac{1}{c}(\boldsymbol{v}\times\boldsymbol{b})^{\mathsf{T}}\right) + (1-\gamma)\frac{\boldsymbol{v}^{\mathsf{T}}\boldsymbol{e}}{v^2}\boldsymbol{v}^{\mathsf{T}}\right)\left(\gamma\left(\boldsymbol{b} + \frac{1}{c}\boldsymbol{v}\times\boldsymbol{e}\right)\right. \\ &\qquad \left. + (1-\gamma)\frac{\boldsymbol{v}^{\mathsf{T}}\boldsymbol{b}}{v^2}\boldsymbol{v}\right) \\ &= \gamma^2\boldsymbol{e}^{\mathsf{T}}\boldsymbol{b} + \left[2\gamma(1-\gamma) + \left(1-\gamma^2\right)\right]\frac{\boldsymbol{e}^{\mathsf{T}}\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{b}}{v^2} - \frac{\gamma^2}{c^2}\underbrace{(\boldsymbol{v}\times\boldsymbol{b})^{\mathsf{T}}(\boldsymbol{v}\times\boldsymbol{e})}_{\boldsymbol{b}^{\mathsf{T}}\boldsymbol{V}_{\times}^{\mathsf{T}}\boldsymbol{V}_{\times}\boldsymbol{e}}.\end{aligned}$$

With $\boldsymbol{V}_{\times}^{\mathsf{T}} = -\boldsymbol{V}_{\times}$ one obtains due to (1.145)

$$\boldsymbol{V}_{\times}^{\mathsf{T}}\boldsymbol{V}_{\times} = -\boldsymbol{V}_{\times}^2 = v^2\boldsymbol{I} - \boldsymbol{v}\boldsymbol{v}^{\mathsf{T}}.$$

This gives

$$\underline{\underline{\boldsymbol{e}'^{\mathsf{T}}\boldsymbol{b}'}} = \gamma^2\boldsymbol{e}^{\mathsf{T}}\boldsymbol{b} - \frac{\gamma^2}{c^2}\left(\boldsymbol{v}^{\mathsf{T}}\boldsymbol{e}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{b} + v^2\boldsymbol{b}^{\mathsf{T}}\boldsymbol{e} - \boldsymbol{b}^{\mathsf{T}}\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{e}\right) = \left(\gamma^2 - \frac{\gamma^2 v^2}{c^2}\right)\boldsymbol{e}^{\mathsf{T}}\boldsymbol{b} = \underline{\underline{\boldsymbol{e}^{\mathsf{T}}\boldsymbol{b}}},$$

so the scalar product of the electric field intensity $\boldsymbol{e}$ and the magnetic induction $\boldsymbol{b}$ is invariant with respect to a Lorentz transformation!

Using a slightly long calculation, one can show that the difference of the squares of the field intensities is also invariant:

$$\underline{\underline{b'^2 - e'^2 = b^2 - e^2}}. \tag{1.147}$$

You can also arrive at electromagnetic invariants in another way, namely with the help of the invariant Faraday's matrix

$$\boldsymbol{F}_{B,e} = \begin{pmatrix} 0 & \boldsymbol{e}^{\mathsf{T}} \\ -\boldsymbol{e} & \boldsymbol{B}_{\times} \end{pmatrix}$$

and the invariant Maxwell's matrix

$$\boldsymbol{F}_{E,b} = \begin{pmatrix} 0 & -\boldsymbol{b}^{\mathsf{T}} \\ \boldsymbol{b} & \boldsymbol{E}_{\times} \end{pmatrix}.$$

With

$$\boldsymbol{F}^*_{B,e} \overset{\text{def}}{=} \boldsymbol{M}\boldsymbol{F}_{B,e}\boldsymbol{M} = \begin{pmatrix} 0 & -\boldsymbol{e}^{\mathsf{T}} \\ \boldsymbol{e} & \boldsymbol{B}_{\times} \end{pmatrix},$$

one gets, e.g.

$$\boldsymbol{F}^*_{B,e}\boldsymbol{F}_{B,e} = \begin{pmatrix} 0 & -\boldsymbol{e}^{\mathsf{T}} \\ \boldsymbol{e} & \boldsymbol{B}_{\times} \end{pmatrix}\begin{pmatrix} 0 & \boldsymbol{e}^{\mathsf{T}} \\ -\boldsymbol{e} & \boldsymbol{B}_{\times} \end{pmatrix} = \begin{pmatrix} \boldsymbol{e}^{\mathsf{T}}\boldsymbol{e} & -\boldsymbol{e}^{\mathsf{T}}\boldsymbol{B}_{\times} \\ -\boldsymbol{B}_{\times}\boldsymbol{e} & \boldsymbol{e}\boldsymbol{e}^{\mathsf{T}} + \boldsymbol{B}_{\times}\boldsymbol{B}_{\times} \end{pmatrix}$$

$$= \begin{pmatrix} e^2 & \boldsymbol{s}^\mathsf{T} \\ \boldsymbol{s} & \begin{pmatrix} e_x^2 - b_z^2 - b_y^2 & \cdots & \cdots \\ \cdots & e_y^2 - b_z^2 - b_x^2 & \cdots \\ \cdots & \cdots & e_z^2 - b_y^2 - b_x^2 \end{pmatrix} \end{pmatrix},$$

with the Poynting vector $\boldsymbol{s} \overset{\text{def}}{=} \boldsymbol{e} \times \boldsymbol{b}$. By forming the trace of this matrix product, i.e. taking the sum of the matrix elements on the main diagonal, one obtains

$$\operatorname{trace}\bigl(\boldsymbol{F}^*_{B,e} \boldsymbol{F}_{B,e}\bigr) = 2e^2 - 2b^2,$$

i.e. the above invariant in the form

$$\boxed{-\frac{1}{2}\operatorname{trace}\bigl(\boldsymbol{F}^*_{B,e} \boldsymbol{F}_{B,e}\bigr) = b^2 - e^2.} \tag{1.148}$$

One obtains the second invariant $\boldsymbol{e}^\mathsf{T} \boldsymbol{b}$ above by taking the trace of the product of the modified Faraday's matrix $\boldsymbol{F}^*_{B,e}$ and the Maxwell's matrix $\boldsymbol{F}_{E,b}$:

$$\boxed{-\frac{1}{4}\operatorname{trace}\bigl(\boldsymbol{F}^*_{B,e} \boldsymbol{F}_{E,e}\bigr) = \boldsymbol{e}^\mathsf{T} \boldsymbol{b}.} \tag{1.149}$$

Making invariants of this type later plays a role in the consideration of the singularities of the Schwarzkopf's solution of General Relativity equations.

1.8.4 Electromagnetic Forces

We want to determine the force acting on a charged particle with the charge q, which is in an electromagnetic field moving with the velocity $\boldsymbol{u}$ relative to an inertial frame $\mathcal{X}$. Let $\mathcal{X}'$ be the inertial frame in which the particle rests at the moment. In this system, due to $\boldsymbol{u}' = \boldsymbol{o}$ and $\boldsymbol{u}' \times \boldsymbol{b}' = \boldsymbol{o}$,

$$m_0 \frac{\mathrm{d}\boldsymbol{u}'}{\mathrm{d}t'} = q\boldsymbol{e}' \in \mathbb{R}^3. \tag{1.150}$$

Generally, due to the relativity principle, the Lorentz force in $\mathcal{X}$ is

$$\boldsymbol{f} = q\left(\boldsymbol{e} + \frac{1}{c}\boldsymbol{u} \times \boldsymbol{b}\right) = \frac{q}{c}[\boldsymbol{e} \mid -\boldsymbol{B}_\times] \begin{pmatrix} c \\ \boldsymbol{u} \end{pmatrix} \in \mathbb{R}^3. \tag{1.151}$$

The following law expresses how an electromagnetic field acts on a stationary charge q and current $\boldsymbol{j}$:

$$\boldsymbol{f} = q\boldsymbol{e} + \frac{1}{c}\boldsymbol{j} \times \boldsymbol{b} = \frac{1}{c}[\boldsymbol{e} \mid -\boldsymbol{B}_\times] \begin{pmatrix} cq \\ \boldsymbol{j} \end{pmatrix} \in \mathbb{R}^3. \tag{1.152}$$

data 13.04.2020

Fig 6.4 Hohmann transfers between coaxial elliptical orbits.

C) Books_68_2020/
Einstein in Matrix Form Exact Derivation of the theory of Special and General Relativity.../
Einstein"Matrix...pdf

Pág. 34(47)

1.8.2 Lorentz Transformation of the Maxwell's Equations

Look (1.146) $P(v)=\left(\begin{array}{c|c} & \\ \hline & \end{array}\right)$

136) 20200414

A) Books_68_2020/
Problèmes de Géométrie Analytique Kletenik pdf/
cours 3_et_4_niv2_rev.pdf

other source: Books_68_2020/Important

Pág. 245 Espace physique et espace vectoriel

Look: Pag. 255 Réponse ex. 1-11
veja o gráfico (desenho geométrico)

obs: Use este livro.

' #6
)*+,)))-) #6 #= =%= : = 3"
(% 0 3 3 7 E 3 :

1 Espace physique et espace vectoriel

Une notion fondamentale en géométrie est celle de **déplacement**. Sans elle, il n'y a pas de moyen pour dire que ces deux triangles sont égaux (isométriques). En géométrie, les déplacements sont appelés des **translations**.

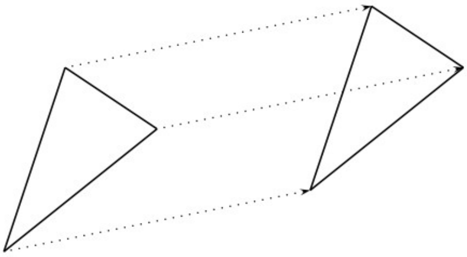

On peut considérer que l'espace physique qui nous entoure est un ensemble de points $\mathscr{E}$ qui admet un ensemble de translations $\mathscr{T}$. Chaque translation T est une bijection qui envoie chaque point A de $\mathscr{E}$ sur un autre point B de $\mathscr{E}$ de telle sorte que les **segments orientés** reliant les points à leur image soient tous de même direction, de même sens et de même longueur : une translation sera notée $T_{\overrightarrow{AB}}$ et on a $T_{\overrightarrow{AB}}(A) = B$, ou encore, d'après la figure ci-contre, $T_{\overrightarrow{AB}}(C) = D$.

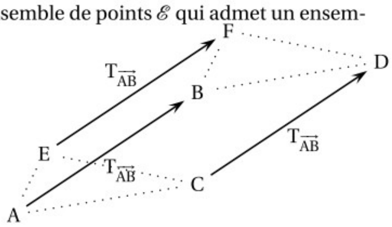
F
D
$T_{\overrightarrow{AB}}$
B
$T_{\overrightarrow{AB}}$
E
$T_{\overrightarrow{AB}}$
C
A

Cette ensemble $\mathscr{T}$ a les propriétés suivantes :

(i) une translation est complètement déterminée si on connaît son action sur un seul point A, c'est-à-dire si on connaît le point B associé au point A. Ainsi donc la translation peut être caractérisée en traçant le segment [AB] et en mettant une flèche sur $\overrightarrow{AB}$, la pointe de la flèche étant en B pour indiquer que l'on va de A vers B. En langage mathématique, on dira

$$\forall X, Y \in \mathscr{E} \quad \exists \text{ une unique translation } T_{\overrightarrow{XY}} \in \mathscr{T} \text{ telle que } T_{\overrightarrow{XY}}(X) = Y$$

(ii) les translations peuvent être additionnées entre elles et multipliées par un nombre réel, de telle sorte que la somme de deux translations est encore une translation et le produit d'une translation par un nombre réel est aussi une translation :

$$T_1 + T_2 \in \mathscr{T} \quad \text{et} \quad \alpha T \in \mathscr{T}\text{, } \alpha \text{ est un nombre réel}$$

Ces opérations ont les propriétés suivantes qui sont celles d'un espace vectoriel.

Définition 1 - 1

Soit $T_1, T_2, T_3 \in \mathscr{T}$ et $\alpha, \beta \in \mathbb{R}$, alors si les éléments d'un ensemble $\mathscr{T}$ ont les propriétés

P1 $(T_1 + T_2) + T_3 = T_1 + (T_2 + T_3)$ (associativité)

P2 il existe un élément neutre 0 tel que

$$0 + T = T + 0 = T \quad \text{pour tout élément } T \in \mathscr{T}$$

P3 étant donné un élément T, il existe un élément −T tel que

$$T + (-T) = 0$$

P4 $T_1 + T_2 = T_2 + T_1$ (commutativité)

P5 $\alpha(T_1 + T_2) = \alpha T_1 + \alpha T_2$

P6 $(\alpha + \beta)T = \alpha T + \beta T$

P7 $(\alpha\beta)T = \alpha(\beta T)$

P8 pour toute translation T, on a $1 \cdot T = T$.

on dira que $\mathscr{T}$ forme **un espace vectoriel**.

Remarque

Habituellement, un espace vectoriel est désigné par la lettre V.
Si on est dans le plan, ce sera V^2 et dans l'espace V^3.
On vérifie aisément que les translations ont ces propriétés (illustrations).

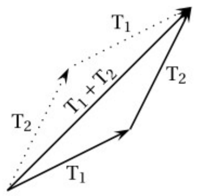

illustration de P4

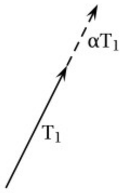

multiplication avec $\alpha > 1$

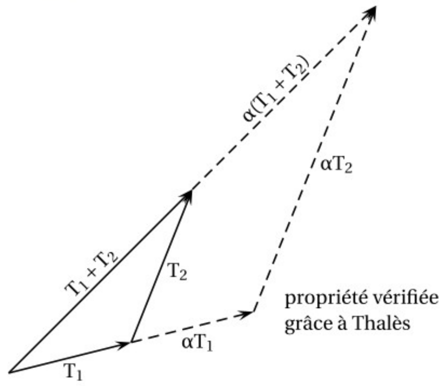

illustration de **P5**

Remarques

a. la figure illustrant l'addition de deux translations montre bien qu'elle est commutative. En effet, T_1+T_2 et T_2+T_1 constituent un parallélogramme (2 paires de côtés parallèles de même longueur). C'est la raison pour laquelle on parle de la **règle du parallélogramme** pour l'addition de deux translations (ou vecteurs, cf. plus tard);

b. certaines translations sont équivalentes. Par exemple, si une translation amène un point A sur B et un point C sur D, alors les translations $T_{\overrightarrow{AB}}$ et $T_{\overrightarrow{CD}}$ sont équivalentes;

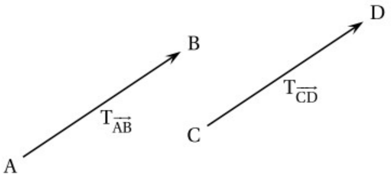

$T_{\overrightarrow{AB}} \equiv T_{\overrightarrow{CD}}$

c. si on choisit un point $O \in \mathscr{E}$ (une origine dans l'espace physique des points), alors à chaque point correspond une unique translation. En effet, soit un point A, on peut définir pour ce point une translation qui amène O sur A, notée $T_{\overrightarrow{OA}}$. Ainsi, il est possible de faire correspondre à chaque point $A \in \mathscr{E}$ une unique translation T, et réciproquement (on a une bijection entre $\mathscr{E}$ et V^3)

d. toutes les translations parallèles de même direction, de même sens et de même longueur sont équivalentes et représentent donc le même vecteur; on les appelle des **représentants de ce vecteur**. En particulier, pour chaque vecteur $\vec{v}$ il existe un unique représentant de $\vec{v}$ dont le point d'application est l'origine : $\overrightarrow{OA}$;

e. une autre manière d'exprimer la même idée est de dire que $\overrightarrow{AB}$ et $\overrightarrow{CD}$ représente le même vecteur si et seulement si ABDC est un parallélogramme;

f. un ensemble d'objets qui vérifie les propriétés P1-P8 forme un **espace vectoriel** et les éléments de cet espace, appelés des **vecteurs** sont généralement désignés par une lettre en caractère gras $\boldsymbol{v}$, ou par une lettre surmontée d'une flèche $\vec{v}$ ou encore par un couple de points représentant les extrémités d'un segment et surmonté aussi d'une flèche $\overrightarrow{AB}$. La pointe de la flèche étant en B pour indiquer que l'on va de A vers B : A étant le point d'application et B l'extrémité du vecteur;

g. ces propriétés signifient aussi que tout vecteur $\vec{v}$ obtenu à partir de deux autres $\vec{a}$ et $\vec{b}$ de la manière suivante

$$\vec{v} = \alpha \cdot \vec{a} + \beta \cdot \vec{b} \qquad \text{avec } \alpha \text{ et } \beta \text{ deux réels quelconques}$$

est un vecteur qui est dans le même espace. On dira que $\vec{v}$ est une **combinaison linéaire** des vecteurs $\vec{a}$ et $\vec{b}$.

(iii) *Dans l'espace*, il existe trois éléments T_1, T_2, T_3 qui sont **linéairement indépendants**, c.-à-d. l'équation

$$\alpha T_1 + \beta T_2 + \gamma T_3 = 0$$

n'a de solution que pour $\alpha = \beta = \gamma = 0$ (cf. activité 1 et 2). Ainsi, V est de dimension 3 et on le note V^3.

Quatre translations sont toujours linéairement dépendantes dans l'espace.

Dans le plan, on a le même type de propriétés, mais dans une dimension inférieure : il existe dans le plan seulement 2 éléments linéairement indépendants, mais pas 3.

Activité

Vecteurs générateurs du plan et de l'espace

Imaginer un robot se déplaçant dans le plan et ne pouvant se déplacer que dans deux directions ; par exemple, nord-est et sud-est. Il peut avancer à chaque fois d'un pas ou d'un multiple réel de pas (comme un demi-pas, ou trois pas, ou moins un tiers de pas). Est-ce que ce robot peut atteindre tous les points du plan dans lequel il évolue ? Quelle combinaison de pas doit-il faire pour se retrouver dans la même position ?
Si on considère le même type de problème, mais dans l'espace, combien de directions différentes un robot-oiseau doit-il pouvoir prendre pour atteindre tous les points de l'espace ?

Activité

Vecteurs linéairement dépendants

Dénotons un pas dans la direction ouest-est par le vecteur $\vec{i}$, un pas dans la direction sud-nord par $\vec{j}$ et un pas dans la direction sud-est par $\vec{l}$. Est-il possible d'exprimer un de ces trois déplacements en fonction des deux autres. Faire un dessin pour se représenter les différents déplacements !

Définition 1 - 2

Un ensemble de vecteurs sont dits linéairement dépendants si l'un des vecteurs peut être exprimé comme combinaison linéaire des autres.

Définition 1 - 3

Cas particuliers :

- Si deux vecteurs sont linéairement dépendants, on dira qu'ils sont **colinéaires**, c'est-à-dire l'un un est multiples de l'autre : $\vec{a} = \lambda \cdot \vec{b}$.
- Si trois vecteurs dans V^3 sont linéairement dépendants, on dira qu'ils sont **coplanaires**, c'est-à-dire l'un des vecteurs est une combinaison linéaire des deux autres : $\vec{c} = \alpha \cdot \vec{a} + \beta \cdot \vec{b}$.

Remarque

Si trois vecteurs $\vec{a}$, $\vec{b}$ et $\vec{c}$ sont linéairement dépendants, il doit donc être possible d'en écrire un en fonction des deux autres, par exemple

$$\vec{c} = \alpha \cdot \vec{a} + \beta \cdot \vec{b}$$

Cette égalité peut être réécrite sous la forme

$$\vec{0} = \vec{c} - \alpha \cdot \vec{a} + \beta \cdot \vec{b}$$

Inversement, si on peut écrire

$$\vec{0} = \alpha \cdot \vec{a} + \beta \cdot \vec{b} + \gamma \cdot \vec{c} \qquad \text{avec } \alpha,\ \beta,\ \gamma \text{ non tous nuls}$$

alors ces vecteurs sont linéairement dépendants, car en considérant le coefficient non nul (supposons que c'est β), il est possible d'écrire

$$\vec{b} = -\frac{\alpha}{\beta} \cdot \vec{a} - \frac{\gamma}{\beta} \cdot \vec{c}.$$

Donc, on peut poser la définition alternative pour la dépendance linéaire :

Définition 1 - 4

Un ensemble de vecteurs sont dits linéairement dépendants si on peut écrire

$$\vec{0} = \alpha \cdot \vec{a} + \beta \cdot \vec{b} + \gamma \cdot \vec{c} \qquad \text{avec } \alpha,\ \beta,\ \gamma \text{ non tous nuls}$$

de telle sorte que le terme au coefficient non nul puisse être isolé.

Du même coup, on peut donner la définition de l'**indépendance linéaire**

Définition 1 - 5

Un ensemble de trois vecteurs sont dits linéairement indépendants si l'équation

$$\vec{0} = \alpha \cdot \vec{a} + \beta \cdot \vec{b} + \gamma \cdot \vec{c}$$

n'a de solution que pour $\alpha = \beta = \gamma = 0$.

Jann Weiss, Licence Creative Commons, 2015-2016

Si l'un des coefficients pouvait être différent de 0, alors il serait possible d'isoler le terme correspondant et d'exprimer un vecteur en fonction des deux autres. Plus tard, quand les vecteurs seront donnés par leurs composantes, on verra comment tester si trois vecteurs sont linéairement indépendants.

2 Notions de base et de coordonnées

Lemme 1 - 1

Soient $\vec{a}, \vec{b}, \vec{c} \in \mathrm{V}^3$, 3 vecteurs linéairement indépendants.
Pour tout vecteur $\vec{p} \in \mathrm{V}^3$, il existe $\alpha, \beta, \gamma \in \mathbb{R}$ uniques tels que $\vec{p} = \alpha\vec{a} + \beta\vec{b} + \gamma\vec{c}$

Preuve.

$\vec{a}, \vec{b}, \vec{c}, \vec{p}$ sont linéairement dépendants (propriétés (iii))
Il existe donc $\alpha_o, \beta_o, \gamma_o, \delta_o$ non tous nuls avec

$$\alpha_o\vec{a} + \beta_o\vec{b} + \gamma_o\vec{c} + \delta_o\vec{p} = \vec{0}$$

$\delta_o \neq 0$ car si $\delta_o = 0$, alors

$$\alpha_o\vec{a} + \beta_o\vec{b} + \gamma_o\vec{c} = \vec{0}$$

et $\alpha_o, \beta_o, \gamma_o$ non tous nuls. Ce qui est impossible, car $\vec{a}, \vec{b}, \vec{c}$ sont linéairement indépendants.
Ainsi

$$\vec{p} = -\underbrace{\frac{\alpha_o}{\delta_o}}_{\alpha}\vec{a} - \underbrace{\frac{\beta_o}{\delta_o}}_{\beta}\vec{b} - \underbrace{\frac{\gamma_o}{\delta_o}}_{\gamma}\vec{c}$$

On montre encore que α, β, γ sont uniques. Supposons le contraire et qu'il existe α', β', γ' tels que

$$\begin{aligned} \vec{p} &= \alpha\vec{a} + \beta\vec{b} + \gamma\vec{c} \\ &= \alpha'\vec{a} + \beta'\vec{b} + \gamma'\vec{c} \end{aligned}$$

Alors

$$(\alpha - \alpha')\vec{a} + (\beta - \beta')\vec{b} + (\gamma - \gamma')\vec{c} = \vec{0}$$

Comme $\vec{a}, \vec{b}, \vec{c}$ sont linéairement indépendants, on a :

$$\begin{array}{lcl} \alpha - \alpha' = 0 & \text{donc} & \alpha = \alpha' \\ \beta - \beta' = 0 & \text{donc} & \beta = \beta' \\ \gamma - \gamma' = 0 & \text{donc} & \gamma = \gamma' \end{array}$$

□

Définition 1 - 6

Un ensemble de trois vecteurs linéairement indépendants de V^3 s'appelle une **base**.
α, β, γ sont les coordonnées de $\vec{p}$ dans la base en question et on peut écrire $\vec{p} = (\alpha, \beta, \gamma)$

Le choix d'une base pour V^3 permet d'identifier notre espace vectoriel avec $\mathbb{R}^3$.

Remarque

Les vecteurs qui constituent une base ne sont pas nécessairement orthogonaux entre eux, comme on peut le voir dans plusieurs exercices (exercice 1 - 10 et suivants).

Une fois qu'un espace vectoriel est muni d'une base et qu'ainsi chaque vecteur peut être caractérisé par ses coordonnées dans cette base, il reste à voir comment l'addition de vecteurs et la multiplication d'un vecteur par un scalaire (un nombre) se traduisent au niveau de ces coordonnées

Grâce aux exemples qui vont suivre, on vérifie que si $\vec{p} = (\alpha, \beta, \gamma)$ et $\vec{p'} = (\alpha', \beta', \gamma')$, alors

$$\vec{p} + \vec{p'} = (\alpha + \alpha', \beta + \beta', \gamma + \gamma')$$

et si $\nu \in \mathbb{R}$, alors

$$\nu\vec{p} = (\nu\alpha, \nu\beta, \nu\gamma)$$

mais présentons d'abord une base particulière qui est aussi la plus habituelle.

Jann Weiss, Licence Creative Commons , 2015-2016

Base standard

Plan

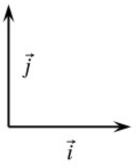

Espace

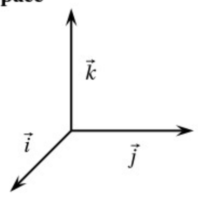

Vecteurs dans $\mathbb{R}^2$ Dans le graphique ci-contre, on lit, pour les différents vecteurs présentés, les coordonnées suivantes dans la base $\left(\vec{i}\,;\vec{j}\right)$

$\overrightarrow{AB} = 4\vec{i} = (4\,;0)$

$\overrightarrow{CD} = 3\vec{i} + 2\vec{j} = (3\,;2)$

$\overrightarrow{GH} = 4.5\vec{i} + 3\vec{j} = (4.5\,;3) = 1.5\overrightarrow{CD}$

$\overrightarrow{OE} = -3\vec{i} - 2\vec{j} = (-3\,;-2)$

$\overrightarrow{PQ} = 2\vec{i} - 3\vec{j} = (2\,;-3)$

$\overrightarrow{OF} = 2\vec{i} - 3\vec{j} = (2\,;-3) = \overrightarrow{PQ}$

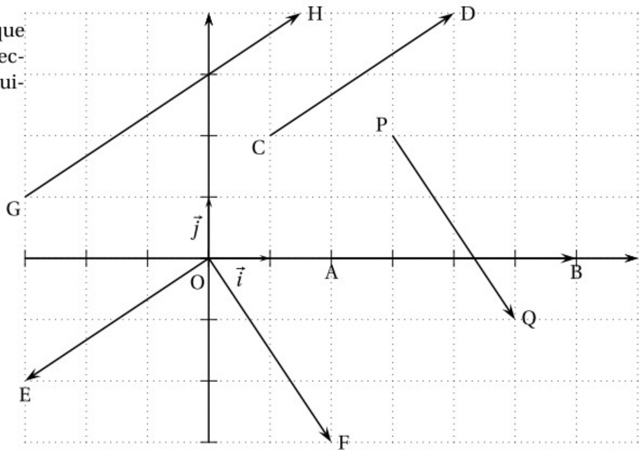

Multiplication d'un vecteur dans $\mathbb{R}^2$ par un réel Le produit d'un vecteur $\vec{p}$ par un nombre $\lambda \in \mathbb{R}$ donne un nouveau vecteur de même direction que l'original, mais dont la longueur a été multipliée par λ. Sur la figure ci-contre, on voit que par Thalès, si on multiplie les composantes du vecteur $\vec{p}$ par λ, on obtient le vecteur $\lambda\vec{p}$.

Dans les deux exemples ci-contre, on reconnaît les deux configurations de Thalès. Dans la seconde, $\vec{p} = (3\,;1)$ et $\lambda = -0.5$; ce dernier nombre représente le facteur de proportionnalité entre les côtés des deux triangles semblables, ainsi : $\lambda\vec{p} = (-1.5\,;-0,5)$.
Les propriétés de la multiplication d'un vecteur par un nombre (on dit aussi par « un scalaire ») ont été données plus haut sous **P5-P8**.

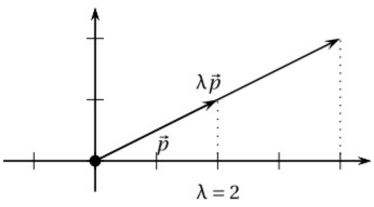

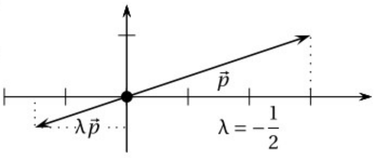

Addition de deux vecteurs dans $\mathbb{R}^2$ On sait que la somme de deux vecteurs s'obtient en effectuant successivement les translations $\vec{p}$ puis $\vec{q}$ (l'ordre n'est pas important). On obtient ainsi un parallélogramme, d'où le nom de « règle du parallélogramme » pour désigner l'addition de vecteurs. Il suffit de vérifier que cette procédure revient à additionner les composantes des vecteurs.

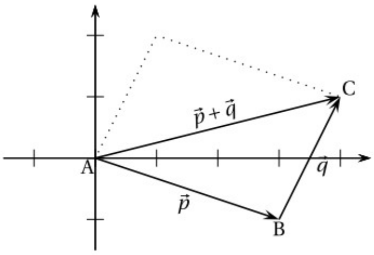

Dans l'exemple de la figure ci-contre

$$\left.\begin{aligned}\vec{p} &= (3\,;-1)\\ \vec{q} &= (1\,;2)\end{aligned}\right\}\quad \vec{p}+\vec{q} = (4\,;1)$$

ou,

$$\left.\begin{aligned}\vec{p} &= 3\vec{i}-\vec{j}\\ \vec{q} &= \vec{i}+2\vec{j}\end{aligned}\right\}\quad \begin{aligned}\vec{p}+\vec{q} &= 3\vec{i}-\vec{j}+\vec{i}+2\vec{j}\\ &= 4\vec{i}+\vec{j} = (4\,;1)\end{aligned}$$

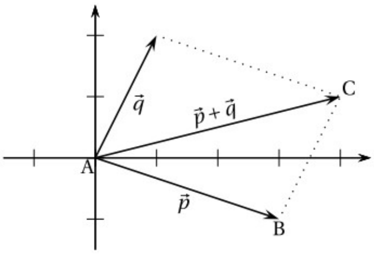

En termes de représentants, cela signifie que l'égalité

$\boxed{\overrightarrow{AB}+\overrightarrow{BC}=\overrightarrow{AC}}$ (relation de Chasles) est vraie pour $\forall A, B, C \in \mathbb{R}$.

Remarques

- **a.** $\overrightarrow{AA}$ est le **vecteur nul** : $\vec{0} = (0\,;0)$. Il n'a pas de direction prédéfinie.
- **b.** Deux vecteurs sont opposés s'ils ont la même direction, même longueur, mais sont de sens opposés : $\overrightarrow{BA}$ est l'opposé de $\overrightarrow{AB}$. Ou encore : $\vec{p}$ et $\vec{q}$ sont opposés $\Longrightarrow \vec{p}+\vec{q}=\vec{0}$.
  Notation : $-\vec{p}$ est l'opposé de $\vec{p}$.
- **c.** cf. plus haut pour les propriétés de l'addition vectorielle **P1-P4**

Il est clair que toutes ces considérations faites dans $\mathbb{R}^2$ sont aussi valables dans $\mathbb{R}^3$. Le seul changement est que, à la différence de $\mathbb{R}^2$ où les points sont des couples, les points dans $\mathbb{R}^3$ sont des triplets : (a_1, a_2, a_3). Mais, comme il n'y a aucune raison de s'arrêter là, tout ce qui vient d'être dit est tout aussi valable pour des espaces de dimensions supérieures n où les points sont représentés par des n-uples $(a_1, \ldots, a_n)$.

2 1 Espace physique, espace vectoriel V^3 et $\mathbb{R}^3$: résumé

Nous avons vu que l'espace physique avec ses translations et une origine présente une identité de structure avec un espace vectoriel. Chaque translation $T_{\overrightarrow{AB}}$ représente un vecteur $\vec{v} = \overrightarrow{AB}$. Ce vecteur a un unique représentant dont le point d'application est l'origine : $\vec{p} = \overrightarrow{OP}$. Ainsi, chaque point P de $\mathscr{E}$ correspond à un vecteur et réciproquement :

$$\begin{aligned}\mathscr{E} &\longleftrightarrow V^3\\ P &\longleftrightarrow \text{l'unique vecteur } \vec{p} \text{ tel que } T_{\vec{p}}(O) = P\end{aligned}$$

En choisissant une base $(\vec{i}, \vec{j}, \vec{k})$ pour V^3, il devient possible d'écrire un vecteur $\vec{p}$ en fonction de cette base et ainsi de l'identifier en fonction de ses composantes qui forment un triplet

$$\vec{p} = \alpha\vec{i}+\beta\vec{j}+\gamma\vec{k} \quad \text{d'où } \vec{p} = (\alpha, \beta, \gamma)$$

Ce qui permet d'identifier V^3 avec $\mathbb{R}^3$

$$\begin{aligned}V^3 &\longleftrightarrow \mathbb{R}^3\\ \vec{p} &\longleftrightarrow (\alpha, \beta, \gamma)\end{aligned}$$

Et en mettant le tout ensemble

$$\begin{array}{ccccc}\mathscr{E} & \longleftrightarrow & V^3 & \longleftrightarrow & \mathbb{R}^3\\ P & \longleftrightarrow & \vec{p} = \overrightarrow{OP} & \longleftrightarrow & (\alpha, \beta, \gamma)\end{array}$$

De cette manière, en généralisant à un espace de dimension quelconque, un n-uple représente soit un point P soit un vecteur $\vec{v} = \overrightarrow{OP}$, selon l'interprétation qu'on a en tête.

Pour terminer, établissons encore le résultat suivant.

$$\boxed{\text{Deux vecteurs } \overrightarrow{\mathrm{AB}} \text{ et } \overrightarrow{\mathrm{CD}} \text{ sont équivalents si } \mathrm{B}-\mathrm{A}=\mathrm{D}-\mathrm{C}}$$

Soient deux points A(a_1, a_2) et B(b_1, b_2). Le vecteur $\overrightarrow{\mathrm{AB}}$ est équivalent à un vecteur $\vec{p}$ dont le point d'application est l'origine. L'abscisse de l'extrémité de $\vec{p}$ s'obtient en diminuant b_1 de a_1, et son ordonnée, en diminuant b_2 de a_2, c.-à-d. que son extrémité est le point B − A et les composantes du vecteur $\overrightarrow{\mathrm{AB}} = \overrightarrow{\mathrm{O(B-A)}}$ sont $(b_1 - a_1, b_2 - a_2)$.
Ainsi, si $\overrightarrow{\mathrm{CD}}$ est un vecteur équivalent à $\overrightarrow{\mathrm{AB}}$, alors on a le résultat souhaité : B − A = D − C.

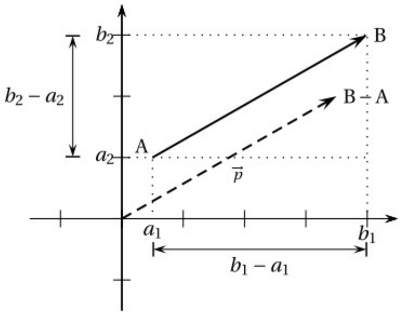

3 Exercices

1 - 1

Dessiner un représentant de

$\vec{a}+\vec{b}$

$\vec{b}-\vec{a}$

$\vec{a}-\vec{b}$

$\vec{a}+\vec{b}+\vec{c}$

$2\vec{a}-\vec{b}+\vec{c}$

$\vec{b}-2\vec{c}-2\vec{a}$

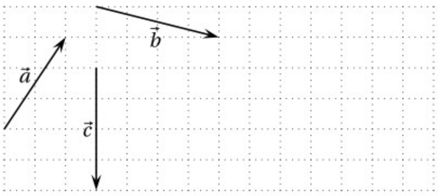

1 - 2

CLIN et OEIL sont des parallélogrammes. En utilisant les seuls points de la figure (il faut faire un dessin), écrivez plus simplement :

$\overrightarrow{\mathrm{OE}}+\overrightarrow{\mathrm{OL}}+\overrightarrow{\mathrm{NC}}$

$\overrightarrow{\mathrm{OE}}+\overrightarrow{\mathrm{IN}}+\overrightarrow{\mathrm{NC}}$

$\overrightarrow{\mathrm{IN}}+\overrightarrow{\mathrm{EO}}$

$\overrightarrow{\mathrm{OE}}+\overrightarrow{\mathrm{OL}}+\overrightarrow{\mathrm{LC}}+\overrightarrow{\mathrm{EO}}$

Remarque

Il existe plusieurs représentants pour un vecteur donné, ainsi $\overrightarrow{\mathrm{CN}}=\overrightarrow{\mathrm{IL}}=\overrightarrow{\mathrm{OE}}$

$\overrightarrow{\mathrm{CL}}=\overrightarrow{\mathrm{NI}}$

$\overrightarrow{\mathrm{CL}}=-\overrightarrow{\mathrm{IN}}$

1 - 3

$\overrightarrow{\mathrm{OA}}$ est un représentant du vecteur $\vec{a}$. $\overrightarrow{\mathrm{OB}}$ est un représentant du vecteur $\vec{b}$. Quel est le vecteur dont un représentant a son point d'application sur A *et son extrémité sur* B *?*

Remarque

On peut tirer de cet exercice une règle générale $\overrightarrow{\mathrm{AB}}=\ldots$

et si, à la place de O, on prend un point P quelconque $\overrightarrow{\mathrm{AB}}=\ldots$

... vecteur pointant sur l'extrémité – vecteur pointant sur l'origine ...

1 - 4

Exprimer le vecteur $\vec{c}$ comme combinaison linéaire de $\vec{a}$ et de $\vec{b}$ (géométriquement !) :

a)

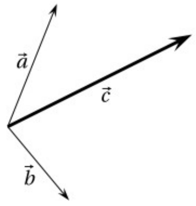

b)

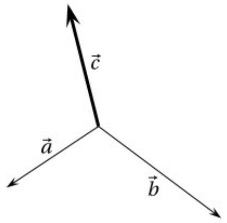

c)

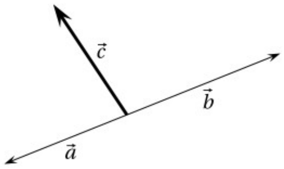

d)

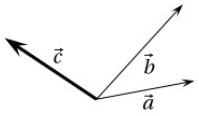

e)

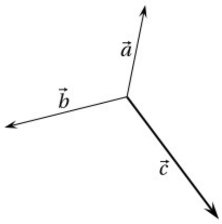

f)

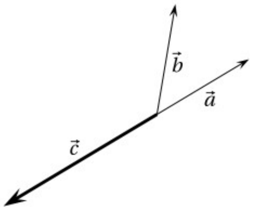

1 - 5

On considère les points $\mathrm{A} = (-3\,;1)$ $\mathrm{B} = (1\,;3)$ $\mathrm{C} = (4\,;1)$ $\mathrm{D} = (2\,;0)$
Démontrer que $(\mathrm{AB}) /\!\!/ (\mathrm{CD})$.

Remarque

$\overrightarrow{\mathrm{AB}}$ et $\overrightarrow{\mathrm{CD}}$ sont alors dits **colinéaires** (cf. les remarques de la section 1.4)

1 - 6

A, B, C, D *sont quatre points du plan. Démontrer que* $\overrightarrow{\mathrm{AB}} + \overrightarrow{\mathrm{CD}} = \overrightarrow{\mathrm{AD}} + \overrightarrow{\mathrm{CB}}$.

1 - 7

On considère les points $\mathrm{A} = \left(\frac{7}{4}\,;-2\right)$ $\mathrm{B} = \left(\frac{1}{2}\,;\frac{9}{4}\right)$ $\mathrm{C} = (3\,;-1)$

(a) *trouver les coordonnées de* E *défini par* $\overrightarrow{\mathrm{AB}} = \overrightarrow{\mathrm{CE}}$ *;*

(b) *trouver les coordonnées de* D *tel que* ABCD *soit un parallélogramme ;*

(c) *quel est le milieu de* [ED] *?*

(d) H *est le point tel que* C *est le milieu de* [BH]
K *est le point tel que C est le milieu de* [AK]
quelles sont les coordonnées de K *et de* H *?*

(e) *écrire tous les vecteurs de la figure égaux respectivement à* $\overrightarrow{\mathrm{AB}}$, *à* $\overrightarrow{\mathrm{BE}}$, *à* $\overrightarrow{\mathrm{EK}}$.

1 - 8

Le vecteur $\vec{c}$ *défini par* $\vec{c} = 2\vec{a} + 3\vec{b}$ *est une combinaison linéaire de* $\vec{a}$ *et de* $\vec{b}$*, c'est-à-dire une somme de « multiples » (par des scalaires) des vecteurs* $\vec{a}$ *et* $\vec{b}$ *:*

a) *calculer* $\vec{c}$ *si* $\vec{a} = \begin{pmatrix} 3 \\ -1 \end{pmatrix}$ *et* $\vec{b} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ *;*

b) *le vecteur* $\vec{d} = \begin{pmatrix} 5 \\ 3 \end{pmatrix}$ *est-il une combinaison linéaire de* $\vec{a}$ *et de* $\vec{b}$ *?*

c) *même question pour les vecteurs* $\vec{v} = \begin{pmatrix} 3 \\ -8 \end{pmatrix}$ *et* $\vec{w} = \begin{pmatrix} 10 \\ 8 \end{pmatrix}$.

1 - 9

Existe-t-il deux vecteurs $\vec{a}$ *et* $\vec{b}$ *tels que* $\vec{c} = \begin{pmatrix} 10 \\ 8 \end{pmatrix}$ *ne soit pas combinaison linéaire de* $\vec{a}$ *et* $\vec{b}$ *?*

1 - 10

Soient quatre vecteurs $\overrightarrow{i}$, $\overrightarrow{j}$, $\overrightarrow{u}$ et $\overrightarrow{v}$. *Déterminer graphiquement, au dixième près*

(a) *les réels* α *et* β *tels que* $\overrightarrow{u} = \alpha\overrightarrow{i} + \beta\overrightarrow{j}$

(b) *les réels* γ *et* δ *tels que* $\overrightarrow{v} = \gamma\overrightarrow{i} + \delta\overrightarrow{j}$

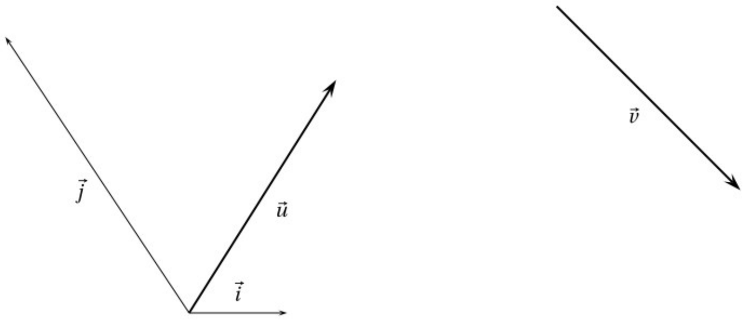

Remarque

Les vecteurs $\overrightarrow{i}$ et $\overrightarrow{j}$ forment une base du plan à l'aide de laquelle on exprime les coordonnées des vecteurs $\overrightarrow{u}$ et $\overrightarrow{v}$.

1 - 11

On donne les points O, I, J, U, V, A, B.

Soient $\overrightarrow{OI} = \overrightarrow{i}$, $\overrightarrow{OJ} = \overrightarrow{j}$, $\overrightarrow{OU} = \overrightarrow{u}$, $\overrightarrow{OV} = \overrightarrow{v}$.

1. (a) *Déterminer graphiquement, au dixième près, les coordonnées des points* A *et* B *dans le repère* $(O, \overrightarrow{i}, \overrightarrow{j})$.

   (b) *Exprimer le vecteur* $\overrightarrow{AB}$ *à l'aide des vecteurs* $\overrightarrow{i}$ et $\overrightarrow{j}$

2. (a) *Déterminer graphiquement, au dixième près, les coordonnées des points* A *et* B *dans le repère* $(O, \overrightarrow{u}, \overrightarrow{v})$.

   (b) *Exprimer le vecteur* $\overrightarrow{AB}$ *à l'aide des vecteurs* $\overrightarrow{u}$ et $\overrightarrow{v}$.

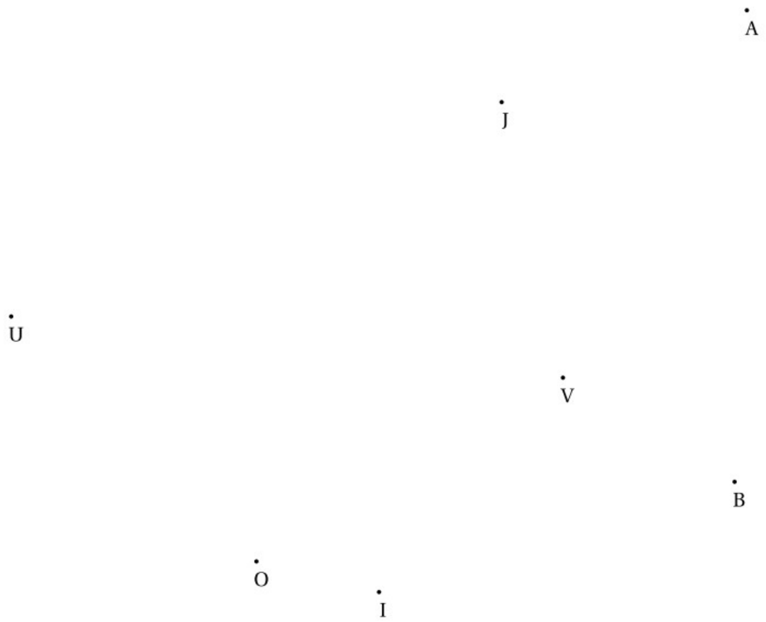

1 - 12

Soient O, I, J *trois points non-alignés. On appellera* $\overrightarrow{i}$ *le vecteur* $\overrightarrow{\mathrm{OI}}$, $\overrightarrow{j}$ *le vecteur* $\overrightarrow{\mathrm{OJ}}$

(a) *Construire les points* A, B, C *tels que*

$$\overrightarrow{\mathrm{OA}} = -\overrightarrow{i} + \overrightarrow{j}, \quad \overrightarrow{\mathrm{OB}} = -2\overrightarrow{i} - \overrightarrow{j}, \quad \overrightarrow{\mathrm{OC}} = \overrightarrow{i} - 3\overrightarrow{j}$$

(b) *Soit* $\overrightarrow{u} = 2\overrightarrow{i} - 3\overrightarrow{j}$. *Construire les points* A′, B′, C′ *tels que*

$$\overrightarrow{\mathrm{AA}'} = \overrightarrow{\mathrm{BB}'} = \overrightarrow{\mathrm{CC}'} = \overrightarrow{u}$$

(c) *Exprimer les vecteurs* $\overrightarrow{\mathrm{OA}'}$, $\overrightarrow{\mathrm{OB}'}$, $\overrightarrow{\mathrm{OC}'}$ *en fonction de* $\overrightarrow{i}$ *et* $\overrightarrow{j}$.

(d) *Exprimer les vecteurs* $\overrightarrow{\mathrm{AB}}$, $\overrightarrow{\mathrm{AC}}$, $\overrightarrow{\mathrm{BC}}$ *en fonction de* $\overrightarrow{i}$ *et* $\overrightarrow{j}$.

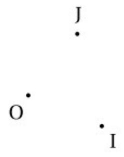

1 - 13

Est-ce que $\vec{c} = (2\,;2\,;1)$ *(vecteur de l'espace* $\mathbb{R}^3$ *!) est combinaison linéaire de* $\vec{a} = (1\,;0\,;3)$ *et de* $\vec{b} = (0\,;-2\,;5)$ *? Même question pour* $\vec{d} = (1\,;4\,;8)$.

1 - 14

Quels sont les vecteurs de $\mathbb{R}^3$ *qui sont combinaison linéaire de* $\vec{a} = (1\,;0\,;0)$ *et* $\vec{b} = (0\,;1\,;0)$ *? (On dit aussi que ces vecteurs sont engendrés par* $\vec{a}$ *et* $\vec{b}$*)*

1 - 15

Combien faut-il de vecteurs pour engendrer l'espace $\mathbb{R}^3$ (c'est-à-dire pour engendrer n'importe quel vecteur de cet espace) ? Y a-t-il des conditions à imposer à ces vecteurs ?

3 1 Réponses partielles

Réponse ex. 1 - 4

a)

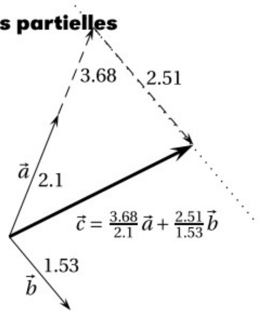
3.68
2.51
$\vec{a}$
2.1
$\vec{c} = \frac{3.68}{2.1}\vec{a} + \frac{2.51}{1.53}\vec{b}$
1.53
$\vec{b}$

b)

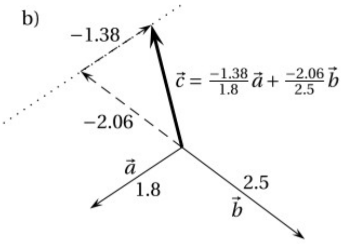
−1.38
$\vec{c} = \frac{-1.38}{1.8}\vec{a} + \frac{-2.06}{2.5}\vec{b}$
−2.06
$\vec{a}$
1.8
2.5
$\vec{b}$

c) $\vec{c}$ ne peut être une combinaison linéaire de $\vec{a}$ et $\vec{b}$

d)

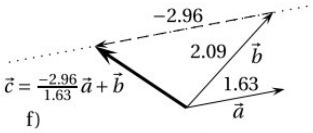
−2.96
2.09
$\vec{b}$
$\vec{c} = \frac{-2.96}{1.63}\vec{a} + \vec{b}$
1.63
$\vec{a}$

e)

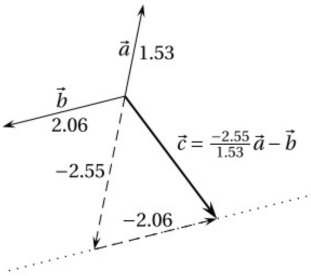
$\vec{a}$
1.53
$\vec{b}$
2.06
$\vec{c} = \frac{-2.55}{1.53}\vec{a} - \vec{b}$
−2.55
−2.06

f)

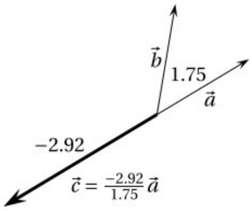
$\vec{b}$
1.75
$\vec{a}$
−2.92
$\vec{c} = \frac{-2.92}{1.75}\vec{a}$

Réponse ex. 1 - 11

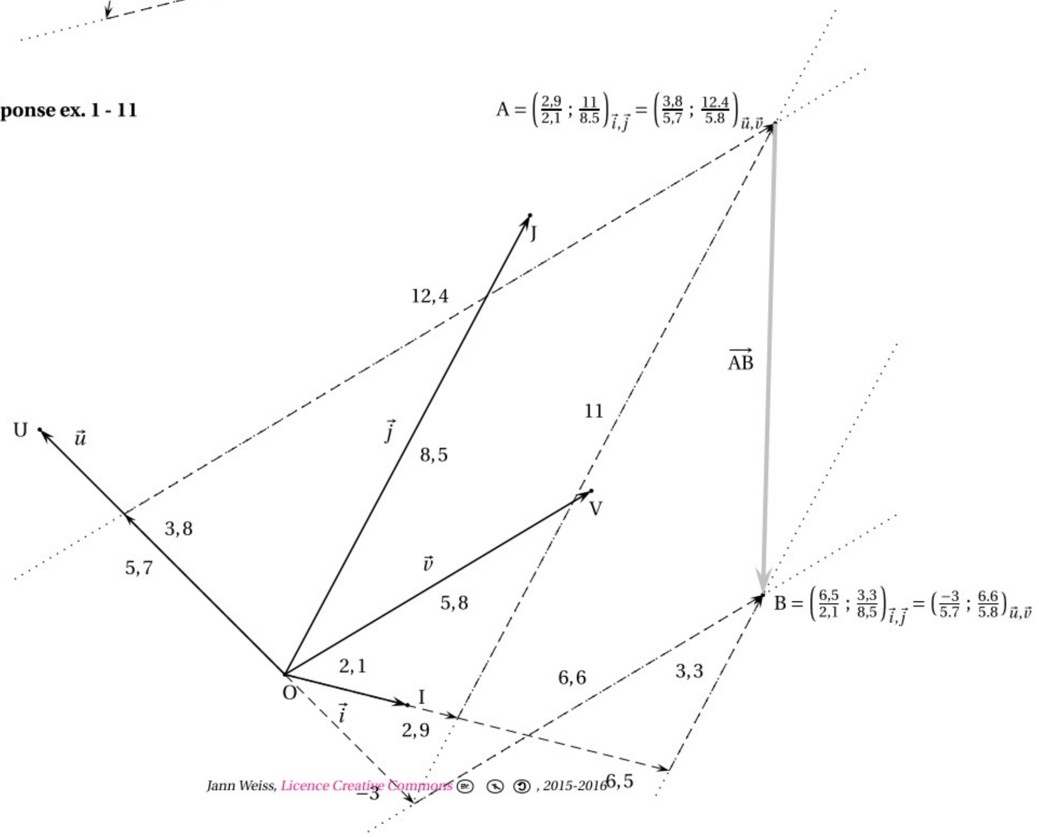
$\mathrm{A} = \left(\frac{2{,}9}{2{,}1} ; \frac{11}{8.5}\right)_{\vec{i},\vec{j}} = \left(\frac{3{,}8}{5{,}7} ; \frac{12.4}{5.8}\right)_{\vec{u},\vec{v}}$
J
12, 4
$\overrightarrow{\mathrm{AB}}$
11
U
$\vec{u}$
$\vec{j}$
8, 5
V
3, 8
5, 7
$\vec{v}$
5, 8
2, 1
O
I
$\vec{i}$
2, 9
6, 6
3, 3
$\mathrm{B} = \left(\frac{6{,}5}{2{,}1} ; \frac{3{,}3}{8{,}5}\right)_{\vec{i},\vec{j}} = \left(\frac{-3}{5.7} ; \frac{6.6}{5.8}\right)_{\vec{u},\vec{v}}$
−3
6, 5

1) (a) A $\approx (1,4\,;1,3)_{\vec{i},\vec{j}}$ et B $\approx (3,1\,;0,4)_{\vec{i},\vec{j}}$

   (b) $\overrightarrow{AB} \approx \begin{pmatrix} 1,7 \\ -0,9 \end{pmatrix}_{\vec{i},\vec{j}}$

2) (a) A $\approx (0,7\,;2,1)_{\vec{u},\vec{v}}$ et B $\approx (-0,5\,;1,1)_{\vec{u},\vec{v}}$

   (b) $\overrightarrow{AB} \approx \begin{pmatrix} -1,2 \\ -1 \end{pmatrix}_{\vec{u},\vec{v}}$

4 Quelques considérations remarquables

(a) Soit A et B deux points distincts, α un réel et M encore un autre point. La relation $\boxed{\overrightarrow{AM} = \alpha\overrightarrow{AB}}$ traduit les propriétés suivantes :

   1) les points M, A et B sont alignées

   2) le point M a pour abscisse α dans le repère $(O, \overrightarrow{AB})$ de la droite (AB)

   Les deux vecteurs $\overrightarrow{AM}$ et $\overrightarrow{AB}$ sont aussi dits **colinéaires** : ils ont la même direction et les droites (AM) et (AB) sont parallèles.

(b) On peut aussi montrer que deux vecteurs $\vec{u}$ et $\vec{v}$ sont colinéaires si, et seulement si,

$$x_u y_v - x_v y_u = 0$$

où (x_u, y_u) et (x_v, y_v) sont les coordonnées respectives de $\vec{u}$ et $\vec{v}$ dans une base $\{\vec{i}, \vec{j}\}$. En effet, $\vec{u}$ et $\vec{v}$ sont colinéaires, si par définition, il existe un $\alpha \in \mathbb{R}$, tel que $\vec{v} = \alpha \cdot \vec{u}$, c'est-à-dire

$$\begin{pmatrix} x_v \\ y_v \end{pmatrix} = \alpha \begin{pmatrix} x_u \\ y_u \end{pmatrix} \Leftrightarrow \begin{cases} x_v &= \alpha \cdot x_u \\ y_v &= \alpha \cdot y_u \end{cases} \Leftrightarrow \begin{cases} x_v \cdot y_u &= \alpha \cdot x_u \cdot y_u \\ x_u \cdot y_v &= \alpha \cdot x_u \cdot y_u \end{cases} \Leftrightarrow x_v \cdot y_u = x_u \cdot y_v \Leftrightarrow x_v \cdot y_u - x_u \cdot y_v = 0$$

Définition 1 - 7

Le réel $x_u y_v - x_v y_u$ est appelé le **déterminant** des vecteurs $\vec{u}$ et $\vec{v}$ dans une base $\{\vec{i}, \vec{j}\}$.

Notation : $\text{dét}(\vec{u}, \vec{v}) = \begin{vmatrix} x_u & x_v \\ y_u & y_v \end{vmatrix}$ et $\begin{vmatrix} x_u & x_v \\ y_u & y_v \end{vmatrix} = x_u y_v - x_v y_u.$

Théorème 1 - 1

Deux vecteurs dans V^2 sont colinéaires si et seulement si $\text{dét}(\vec{u}, \vec{v}) = 0$.

Attention ! Ce théorème n'est valable que dans V^2 et on parle dans ce cas de « déterminant 2 × 2 ».

(c) Dans l'espace, avec trois vecteurs, il existe un opérateur similaire appelé « déterminant 3 × 3 »,

Définition 1 - 8

Le **déterminant** des vecteurs $\vec{u}$, $\vec{v}$ et $\vec{w} \in V^3$ dans une base $\{\vec{i}, \vec{j}, \vec{k}\}$ est le nombre réel :

$$\begin{aligned} \text{dét}(\vec{u}, \vec{v}, \vec{w}) = \begin{vmatrix} x_u & x_v & x_w \\ y_u & y_v & y_w \\ z_u & z_v & z_w \end{vmatrix} &= x_u \cdot \begin{vmatrix} y_v & y_w \\ z_v & z_w \end{vmatrix} - y_u \cdot \begin{vmatrix} x_v & x_w \\ z_v & z_w \end{vmatrix} + z_u \cdot \begin{vmatrix} x_v & x_w \\ y_v & y_w \end{vmatrix} \\ &= x_u y_v z_w - x_u y_w z_v - (y_u x_v z_w - y_u x_w z_v) + z_u x_v y_w - z_u x_w y_v \\ &= x_u y_v z_w + y_u z_v x_w + z_u x_v y_w - z_u y_v x_w - y_u x_v z_w - x_u z_v y_w \end{aligned}$$

avec un théorème équivalent.

Théorème 1 - 2

Trois vecteurs sont linéairement dépendants (ou coplanaires) ssi leur déterminant est nul :

$$\text{dét}(\vec{u}, \vec{v}, \vec{w}) = 0.$$

Et, si le déterminant n'est pas nul, alors les trois vecteurs sont linéairement indépendants.

Cet outil facilite grandement le calcul pour déterminer si 3 vecteurs sont linéairement dépendants ou indépendants.
En principe, pour le savoir en utilisant directement la définition, il est nécessaire de résoudre un système d'équations. Par exemple, si on a les vecteurs

$$\vec{a} = \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix} \quad \vec{b} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \quad \vec{c} = \begin{pmatrix} 1 \\ 2 \\ 10 \end{pmatrix}$$

le système à résoudre est

$$\alpha \cdot \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix} + \beta \cdot \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \gamma \cdot \begin{pmatrix} 1 \\ 2 \\ 10 \end{pmatrix} = 0 \quad \Leftrightarrow \begin{array}{c} \text{①} \\ \text{②} \\ \text{③} \end{array} \left\{ \begin{array}{l} 2\alpha + \beta + \quad \gamma = \vec{0} \\ 3\alpha \quad + \ 2\gamma = 0 \\ \ \alpha + \beta + 10\gamma = 0 \end{array} \right.$$

① – ③ donne une nouvelle équation ④ : $\alpha - 9\gamma = 0$.
② – 3 · ③ donne : $29\gamma = 0$, d'où $\gamma = 0$ et en substituant cette valeur dans les équations précédentes, on obtient $\alpha = \beta = 0$. Ce système a pour unique solution $\alpha = \beta = \gamma = 0$, les vecteurs sont donc linéairement indépendants.

Cet exemple était facile à traiter, mais souvent les calculs donnent plus de fil à retordre. Avec le déterminant,

$$\text{dét}\left(\begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 10 \end{pmatrix}\right) = 2 \cdot (0-2) - 3 \cdot (10-1) + 1 \cdot (2-0) = -4 - 27 + 2 = -29 \quad \Rightarrow \text{vecteurs lin. indépendants}$$

c'est toujours très rapide.

Remarque

Le lien entre la résolution d'un système d'équations linéaires et le déterminant est assuré par la règle dite de Cramer. Cela mérite quelques explications. Commençons par données quelques propriétés du déterminant.

(a) $\textbf{dét}(\vec{a} + \vec{b}, \vec{c}, \vec{d}) = \textbf{dét}(\vec{a}, \vec{c}, \vec{d}) + \textbf{dét}(\vec{b}, \vec{c}, \vec{d})$

Pour démontrer cette propriété, il suffit de développer le déterminant

$$\begin{aligned} \text{dét}(\vec{a} + \vec{b}, \vec{c}, \vec{d}) &= (a_1 + b_1) \cdot \begin{vmatrix} c_2 & d_2 \\ c_3 & d_3 \end{vmatrix} - (a_2 + b_2) \cdot \begin{vmatrix} c_1 & d_1 \\ c_3 & d_3 \end{vmatrix} + (a_3 + b_3) \begin{vmatrix} c_1 & d_1 \\ c_2 & d_2 \end{vmatrix} \\ = a_1 \cdot \begin{vmatrix} c_2 & d_2 \\ c_3 & d_3 \end{vmatrix} - a_2 \cdot \begin{vmatrix} c_1 & d_1 \\ c_3 & d_3 \end{vmatrix} + a_3 \cdot \begin{vmatrix} c_1 & d_1 \\ c_2 & d_2 \end{vmatrix} &+ b_1 \cdot \begin{vmatrix} c_2 & d_2 \\ c_3 & d_3 \end{vmatrix} - b_2 \cdot \begin{vmatrix} c_1 & d_1 \\ c_3 & d_3 \end{vmatrix} + b_3 \cdot \begin{vmatrix} c_1 & d_1 \\ c_2 & d_2 \end{vmatrix} \\ &= \text{dét}(\vec{a}, \vec{c}, \vec{d}) + \text{dét}(\vec{b}, \vec{c}, \vec{d}) \end{aligned}$$

On a de manière semblable

$$\text{dét}(\vec{a}, \vec{b} + \vec{c}, \vec{d}) = \text{dét}(\vec{a}, \vec{b}, \vec{c}) + \text{dét}(\vec{a}, \vec{c}, \vec{d}) \qquad \text{et} \quad \text{dét}(\vec{a}, \vec{b}, \vec{c} + \vec{d}) = \text{dét}(\vec{a}, \vec{b}, \vec{c}) + \text{dét}(\vec{a}, \vec{b}, \vec{d})$$

(b) $\textbf{dét}(\alpha \cdot \vec{a}, \vec{b}, \vec{c}) = \alpha \cdot \textbf{dét}(\vec{a}, \vec{b}, \vec{c})$

On procède de manière analogue pour démontrer cette propriété

$$\begin{aligned} \text{dét}(\alpha \cdot \vec{a}, \vec{b}, \vec{c}) &= \alpha a_1 \cdot \begin{vmatrix} b_2 & c_2 \\ b_3 & c_3 \end{vmatrix} - \alpha a_2 \cdot \begin{vmatrix} b_1 & c_1 \\ b_3 & c_3 \end{vmatrix} + \alpha a_3 \begin{vmatrix} b_1 & c_1 \\ b_2 & c_2 \end{vmatrix} \\ &= \alpha \left(a_1 \cdot \begin{vmatrix} b_2 & c_2 \\ b_3 & c_3 \end{vmatrix} - a_2 \cdot \begin{vmatrix} b_1 & c_1 \\ b_3 & c_3 \end{vmatrix} + a_3 \cdot \begin{vmatrix} b_1 & c_1 \\ b_2 & c_2 \end{vmatrix} \right) = \alpha \cdot \text{dét}(\vec{a}, \vec{b}, \vec{c}) \end{aligned}$$

On a de manière semblable

$$\text{dét}(\vec{a}, \alpha \cdot \vec{b}, \vec{c}) = \alpha \cdot \text{dét}(\vec{a}, \vec{b}, \vec{c}) \qquad \text{et} \quad \text{dét}(\vec{a}, \vec{b}, \alpha \cdot \vec{c}) = \alpha \cdot \text{dét}(\vec{a}, \vec{b}, \vec{c})$$

Ces deux premières propriétés font que le déterminant est multi-**linéaire**.

(c) **$\text{dét}(\vec{a},\vec{b},\vec{c}) = -\text{dét}(\vec{b},\vec{a},\vec{c})$ (propriété d'antisymétrie)**
On le démontre en développant le membre de gauche et de droite séparément :

$$\begin{aligned}
\text{dét}(\vec{a},\vec{b},\vec{c}) &= a_1\cdot\begin{vmatrix} b_2 & c_2 \\ b_3 & c_3 \end{vmatrix} - a_2\cdot\begin{vmatrix} b_1 & c_1 \\ b_3 & c_3 \end{vmatrix} + a_3\cdot\begin{vmatrix} b_1 & c_1 \\ b_2 & c_2 \end{vmatrix}\\
&= a_1b_2c_3 - a_1b_3c_2 - a_2b_1c_3 + a_2b_3c_1 + a_3b_1c_2 - a_3b_2c_1\\
-\text{dét}(\vec{b},\vec{a},\vec{c}) &= -\left(b_1\cdot\begin{vmatrix} a_2 & c_2 \\ a_3 & c_3 \end{vmatrix} - b_2\cdot\begin{vmatrix} a_1 & c_1 \\ a_3 & c_3 \end{vmatrix} + b_3\cdot\begin{vmatrix} a_1 & c_1 \\ a_2 & c_2 \end{vmatrix}\right)\\
&= -(b_1a_2c_3 - b_1a_3c_2 - b_2a_1c_3 + b_2a_3c_1 + b_3a_1c_2 - b_3a_2c_1)\\
&= a_1b_2c_3 - a_1b_3c_2 - a_2b_1c_3 + a_2b_3c_1 + a_3b_1c_2 - a_3b_2c_1
\end{aligned}$$

On a aussi

$$\text{dét}(\vec{a},\vec{b},\vec{c}) = -\text{dét}(\vec{a},\vec{c},\vec{b})$$

et on en tire

$$\text{dét}(\vec{c},\vec{a},\vec{b}) = -\text{dét}(\vec{a},\vec{c},\vec{b}) = -\Big(-\text{dét}(\vec{a},\vec{b},\vec{c})\Big) = \text{dét}(\vec{a},\vec{b},\vec{c})$$

(d) **$\text{dét}(\vec{a},\vec{a},\vec{c}) = 0$ (le déterminant est dit alterné)**
On le découvre en posant dans la propriété précédente (c) $\vec{b} = \vec{a}$:
$\text{dét}(\vec{a},\vec{a},\vec{c}) = -\text{dét}(\vec{a},\vec{a},\vec{c})$, d'où $\text{dét}(\vec{a},\vec{a},\vec{c}) = 0$

(e) **$\text{dét}(\vec{0},\vec{b},\vec{c}) = 0$**
On le démontre en posant dans la propriété (b) $\alpha = 0$:

$$0 = 0\cdot\text{dét}(\vec{a},\vec{b},\vec{c}) = \text{dét}(0\cdot\vec{a},\vec{b},\vec{c}) = \text{dét}(\vec{0},\vec{b},\vec{c})$$

Remarque

Dépendance linéaire et déterminant nul : ces propriétés nous permettent d'aborder la notion de dépendance linéaire avec ce nouvel outil qu'est le déterminant.

Soit 3 vecteurs linéairement dépendants $\vec{a},\vec{b},\vec{c}$ de $\mathbb{R}^3$; un des vecteurs peut donc être exprimé en fonction des 2 autres, par exemples : $\vec{c} = \alpha\cdot\vec{a} + \beta\cdot\vec{b}$. Nous avons alors :

$$\begin{aligned}
\text{dét}(\vec{a},\vec{b},\vec{c}) &= \text{dét}(\vec{a},\vec{b},\alpha\cdot\vec{a}+\beta\cdot\vec{b})\\
&= \text{dét}(\vec{a},\vec{b},\alpha\cdot\vec{a}) + \text{dét}(\vec{a},\vec{b},\beta\cdot\vec{b}) = \alpha\cdot\text{dét}(\vec{a},\vec{b},\vec{a}) + \beta\cdot\text{dét}(\vec{a},\vec{b},\vec{b}) = 0+0 = 0
\end{aligned}$$

Nous avons ainsi : **si 3 vecteurs de $\mathbb{R}^3$ sont linéairement dépendants, leur déterminant est nul**.

Nous aimerions aussi avoir la réciproque. À partir de là, avec les contraposées, on pourra affirmer « trois vecteurs sont linéairement indépendants ssi le déterminant est non-nul ».

Déterminant et indépendance linéaire, les règles de Cramer : il nous faut revenir aux systèmes d'équations linéaires pour terminer. Pour tester l'indépendance linéaire de trois vecteurs dans $\mathbb{R}^3$ à partir de la définition, il fallait poser le système

$$\alpha\vec{a} + \beta\vec{b} + \gamma\vec{c} = \vec{0}$$

qui représente un système d'équations linéaires homogène avec pour variables α,β,γ. Nous allons partir de la situation générale et prendre x, y, z comme variables. Soit le système de 3 équations à 3 inconnues

$$\vec{a}x + \vec{b}y + \vec{c}z = \vec{d} \quad\Leftrightarrow\quad \begin{aligned} a_1x + b_1y + c_1z &= d_1\\ a_2x + b_2y + c_2z &= d_2\\ a_3x + b_3y + c_3z &= d_3 \end{aligned} \qquad (\blacklozenge)$$

Nous avons

$$\begin{aligned}
\text{dét}(\vec{d},\vec{b},\vec{c}) &= \text{dét}(\vec{a}x+\vec{b}y+\vec{c}z,\vec{b},\vec{c})\\
&= x\cdot\text{dét}(\vec{a},\vec{b},\vec{c}) + y\cdot\text{dét}(\vec{b},\vec{b},\vec{c}) + z\cdot\text{dét}(\vec{c},\vec{b},\vec{c}) = x\cdot\text{dét}(\vec{a},\vec{b},\vec{c})
\end{aligned} \qquad (\bigstar)$$

D'où $x = \dfrac{\text{dét}(\vec{d},\vec{b},\vec{c})}{\text{dét}(\vec{a},\vec{b},\vec{c})}$ pour autant que $\text{dét}(\vec{a},\vec{b},\vec{c}) \neq 0$.

On trouvera de manière analogue : $y = \dfrac{\text{dét}(\vec{a},\vec{d},\vec{c})}{\text{dét}(\vec{a},\vec{b},\vec{c})}$ et $z = \dfrac{\text{dét}(\vec{a},\vec{b},\vec{d})}{\text{dét}(\vec{a},\vec{b},\vec{c})}$. (★★)

La règle de Cramer donne ainsi les solutions d'un système linéaire 3×3. En l'appliquant à un système homogène comme $\alpha\vec{a}+\beta\vec{b}+\gamma\vec{c}=\vec{0}$, avec $\vec{a},\vec{b},\vec{c}\in\mathbb{R}^3$, nous allons obtenir le résultat attendu, car tout système homogène a pour solution $x=y=z=0$ (cela se vérifie immédiatement), ici $\alpha=\beta=\gamma=0$, mais il peut admettre encore d'autres solutions. La solution est unique seulement dans le cas où $\text{dét}(\vec{a},\vec{b},\vec{c})\neq 0$, car elle est alors donnée par les formules (★★) qui suppose un déterminant non nul.

En effet, si $\text{dét}(\vec{a},\vec{b},\vec{c})=0$, la règle de Cramer (★) donne

$$\text{dét}(\vec{0},\vec{b},\vec{c})=\alpha\cdot 0 \Leftrightarrow 0=\alpha\cdot 0,$$

ce qui permet de choisir de multiples valeurs pur $\alpha\neq 0$.

Remarque

Nous pouvons maintenant affirmer que **si le déterminant des vecteurs $\vec{a},\vec{b},\vec{c}$ est non nul, alors les vecteurs sont linéairement indépendants** ou, en prenant la contraposée, si les vecteurs sont linéairement dépendants, alors le déterminant est nul.

Pour terminer, revenons au système générale (♦). Si le $\text{dét}(\vec{a},\vec{b},\vec{c})\neq 0$, ce qui est la condition pour que les 3 vecteurs soient linéairement indépendants, alors ceux-ci forment une base pour $\mathbb{R}^3$ et le vecteur $\vec{d}$ peut alors être exprimer de manière unique à l'aide de ces 3 vecteurs : $\vec{d}=\vec{a}x+\vec{b}y+\vec{c}z$.
Si par contre le déterminant est nul, alors les vecteurs $\vec{a},\vec{b},\vec{c}$ ne sont pas linéairement indépendants et ne forment donc pas une base de $\mathbb{R}^3$. Les vecteurs sont coplanaires. On a alors deux cas possibles

(a) $\vec{d}$ appartient au même plan et peut donc aussi être généré par $\vec{a},\vec{b},\vec{c}$. Toutefois, il est coplanaire avec aussi seulement deux des 3 vecteurs, c'est-à-dire que, par exemple, $\text{dét}(\vec{d},\vec{b},\vec{c})=0$. Le système a alors de multiples solutions puisque dans (★), on obtient $0=\alpha 0$, $0=\beta 0$ et $0=\gamma 0$.

(b) $\vec{d}$ n'appartient pas au même plan que $\vec{a},\vec{b},\vec{c}$ et les déterminants $A=\text{dét}(\vec{a},\vec{b},\vec{d})$, $B=\text{dét}(\vec{a},\vec{d},\vec{c})$ et $C=\text{dét}(\vec{d},\vec{b},\vec{c})=0$ sont non nuls et la règle (★) donne

$$A=\alpha\cdot 0, B=\beta\cdot 0, C=\gamma\cdot C,\quad \text{avec } A,B,C\neq 0$$

... équations qui sont sans solution !

(d) Soit $\mathscr{D}$ une droite et A et B deux points distincts tels que (AB) et $\mathscr{D}$ soient parallèles. Autrement dit, (AB) et $\mathscr{D}$ ont la même direction : nous dirons que $\overrightarrow{AB}$ est un **vecteur directeur** de $\mathscr{D}$.

(e) Soit [AB] un segment et I un point. Chacune des égalités ci-après caractérise le fait que I **est le milieu de** [AB] :

$$\boxed{\overrightarrow{AI}=\overrightarrow{IB} \quad \text{ou} \quad \overrightarrow{IA}+\overrightarrow{IB}=\vec{0} \quad ou \quad \overrightarrow{AI}=\frac{1}{2}\overrightarrow{AB}.}$$

(f) La relation de Chasles peut aussi s'écrire

$$\boxed{\overrightarrow{AB}=\overrightarrow{OB}-\overrightarrow{OA}}$$

ou encore

$$\boxed{\overrightarrow{AB}=\begin{pmatrix}b_1\\b_2\end{pmatrix}-\begin{pmatrix}a_1\\a_2\end{pmatrix}=\begin{pmatrix}b_1-a_1\\b_2-a_2\end{pmatrix}}$$

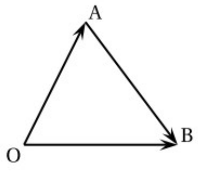

(g) M est le milieu de [AB] si et seulement si

$$\overrightarrow{OM}=\frac{1}{2}(\overrightarrow{OA}+\overrightarrow{OB})$$

ou encore

$$\overrightarrow{OM}=\left(\frac{a_1+b_1}{2}\,;\,\frac{a_2+b_2}{2}\right)$$

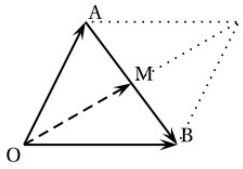

(h) Il y a quelques configurations de bases qui sont données ci-dessous

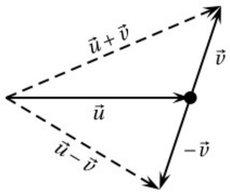

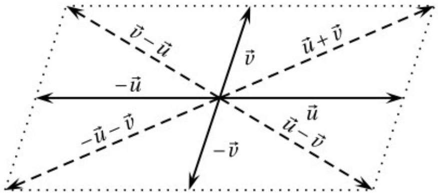

(i) Il y a deux autres configurations importantes

« le milieu I **d'un segment** [AB] **et un autre point** M **»**

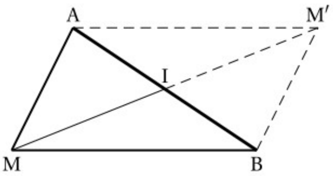

Quel que soit M :

$$\boxed{\begin{array}{l}\overrightarrow{MA}+\overrightarrow{MB}=2\overrightarrow{MI}\\ \overrightarrow{MI}=\dfrac{1}{2}(\overrightarrow{MA}+\overrightarrow{MB})\end{array}}$$

Le calcul est aisé :

$$\begin{aligned}\overrightarrow{MA}+\overrightarrow{MB}&=(\overrightarrow{MI}+\overrightarrow{IA})+(\overrightarrow{MI}+\overrightarrow{IB})\\&=2\overrightarrow{MI}+\overrightarrow{IA}+\overrightarrow{IB}\end{aligned}$$

Comme I est le milieu de [AB], $\overrightarrow{IA}+\overrightarrow{IB}=\overrightarrow{0}$ et on a

$$\overrightarrow{MA}+\overrightarrow{MB}=2\overrightarrow{MI}$$

On peut aussi utiliser le fait que AMBM′ forme un parallélogramme

- $\overrightarrow{MA}+\overrightarrow{MB}=\overrightarrow{MM'}$ (règle du parallélogramme) ;
- $\overrightarrow{MM'}=2\overrightarrow{MI}$ (I est aussi le milieu de [MC])

« le théorème des milieux »

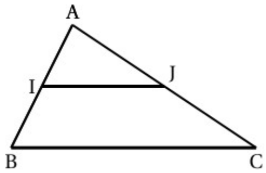

I et J sont les milieux des côtés.
On a :

$$\boxed{\overrightarrow{BC}=2\overrightarrow{IJ}}$$

(Cette relation signifie

- l'égalité BC = 2IJ
- le parallélisme de (BC) et (IJ).)

La démonstration est simple :

$$\begin{aligned}\overrightarrow{BC} = \overrightarrow{AC} - \overrightarrow{AB} &= 2\overrightarrow{AJ} - 2\overrightarrow{AI} \\ &= 2(\overrightarrow{AJ} - \overrightarrow{AI}) = 2\overrightarrow{IJ}\end{aligned}$$

1 - 16

Soient les points $A(1;-2)$, $B(-3;6)$ *et* $C(5;2)$ *dans le plan.*

(a) *Calculer les coordonnées des points* M, N, P, *milieux respectifs de* $[BC]$, $[AC]$, $[AB]$.

(b) *Calculer les coordonnées du point* G *tel que* $\overrightarrow{OG} = \frac{1}{3}(\overrightarrow{OA} + \overrightarrow{OB} + \overrightarrow{OC})$.

1 - 17

Déterminer dans chacun des cas suivants si les vecteurs $\overrightarrow{u}$ *et* $\overrightarrow{v}$ *sont colinéaires :*

(a) $\overrightarrow{u} = 3\overrightarrow{i} + 4\overrightarrow{j}$; $\overrightarrow{v} = 2\overrightarrow{i} + 3\overrightarrow{j}$

(b) $\overrightarrow{u} = 4\overrightarrow{i} - 6\overrightarrow{j}$; $\overrightarrow{v} = 6\overrightarrow{i} - 9\overrightarrow{j}$

(c) $\overrightarrow{u}(2;5)$; $\overrightarrow{v}(-3;4)$

(d) $\overrightarrow{u}(-\frac{2}{5};\frac{8}{5})$; $\overrightarrow{v}(\frac{3}{5};-\frac{12}{5})$

1 - 18

Déterminer dans chacun des cas suivants si les vecteurs sont linéairement dépendants :

1) $\vec{a} = \begin{pmatrix}1\\0\\0\end{pmatrix}, \vec{b} = \begin{pmatrix}0\\1\\0\end{pmatrix}$ *et* $\vec{c} = \begin{pmatrix}0\\0\\3\end{pmatrix}$

2) $\vec{a} = \begin{pmatrix}1\\2\\3\end{pmatrix}, \vec{b} = \begin{pmatrix}4\\5\\6\end{pmatrix}$ *et* $\vec{c} = \begin{pmatrix}7\\8\\9\end{pmatrix}$

3) $\vec{a} = \begin{pmatrix}1\\2\\3\end{pmatrix}, \vec{b} = \begin{pmatrix}4\\5\\6\end{pmatrix}$ *et* $\vec{c} = \begin{pmatrix}7\\-8\\9\end{pmatrix}$

4) $\vec{a} = \begin{pmatrix}1\\-1\\2\end{pmatrix}, \vec{b} = \begin{pmatrix}0\\0\\1\end{pmatrix}$ *et* $\vec{c} = \begin{pmatrix}2\\-2\\-1\end{pmatrix}$

5) $\vec{a} = \begin{pmatrix}1\\1\\1\end{pmatrix}, \vec{b} = \begin{pmatrix}0\\0\\1\end{pmatrix}$ *et* $\vec{c} = \begin{pmatrix}1\\0\\1\end{pmatrix}$

1 - 19

Soient les points $A(-4;5;3)$, $B(-1;1;5)$, $C(5;5;4)$ *et* $D(2;9;2)$.

(a) *Montrer que ces points sont coplanaires.*

(b) *Montrer que* ABCD *est un parallélogramme.*

(c) *Trouver les coordonnées de l'intersection des diagonales de ce parallélogramme.*

1 - 20

On donne les points $C(-3;0)$, $D(2,5;2)$, $E(1,5;-2)$ *et* $F(-3,5;-4,5)$.
Le quadrilatère CDEF *est-il un trapèze ?*

1 - 21

Soit A, A', D, D' *quatre points du plan. De plus,* I *et* J *sont les milieux respectifs de* $[AA']$ *et* $[DD']$.
Montrer que :

(a) $\overrightarrow{AD'} + \overrightarrow{A'D} = \overrightarrow{AD} + \overrightarrow{A'D'}$

(b) $\overrightarrow{AD} + \overrightarrow{A'D'} = 2 \cdot \overrightarrow{IJ}$

data 14 04 2020

B) Books_68_2020/
Mathematics Form and Function/
eng Peng 43523 - Janet. St. pdf

Pág. 107 (118)

7. Vector Geometry

3. Geometry

Pág. 61 (72)

Pág. 64 (75)

Figure 1 Do the circles meet?

Pág. 109 (120)

8. Analytic Geometry

Pág. 111 (122)

Figure 1

$P(\cos\theta, \sin\theta)$

Obs: Use este livro.

I) Books_68_2020/
Recent Advances on Elliptic and
Parabolic Issues/
Michel Chipot.m.pdf

Obs: Ver este livro

/)*+,)))-) #6 7 7%# # 3"
)*+,)))-) #6 " B =
9 " 9 5! @ 3 3"
(& ! & 4 7

construction, contain both the standard reals and actual infinitesimals; they can be used to prove theorems (for the ordinary reals) (Lightstone and Robinson [1975]).

Other approaches restrict attention to those real numbers which are explicitly "constructible" in some specific sense, perhaps by means of recursive functions of natural numbers. An efficient version of constructive analysis is due to Bishop [1967]; the corresponding real numbers are described in Myhill [1973]. Earlier, L.E.J. Brouwer developed intuitionism, with emphasis on real numbers given by choice sequences (Troelstra [1977]; see also Dummett [1977]). In my view, such requirements of constructivity tend to emphasize arithmetic at the expense of geometry. Since the real numbers should subsume *both* arithmetic and geometric aspects of mensuration, these practical uses seem better formalized by the standard axioms for the field of real numbers.

7. Vector Geometry

The real numbers provide a one-dimensional geometry—a line with a chosen origin and unit. The higher dimensional analog is vector geometry. Thus in the Euclidean plane, choose an origin 0 and a segment $0U$ representing the unit of distance. Then each point P in the plane is represented by the directed segment $0P$, called a *vector* v. Vectors $0P$ and $0Q$ can be added by the usual parallelogram law: Form the parallelogram with edges $0P$ and $0Q$, and take the diagonal vector $0R$ as the sum (Figure 1). Under this addition, the vectors form an abelian group; it is just the group of all translations of the plane (§III.7). Using the chosen unit segment $0U$, each real number r can be laid off along the line $0U$. The geometric construction of multiplication by r (Figure 3.1) then defines an operation of multiplying the vector $v = 0R$ by the real number (scalar) r to give the vector rv along the line of $0R$. If $r > 0$ it is in the direction of v; if $r < 0$, it is in the opposite direction, in both cases with $|r|$ times

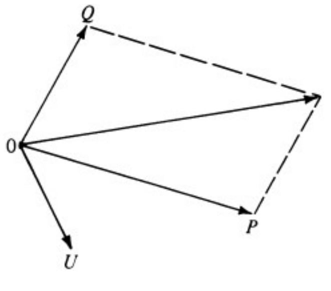

Figure 1

the length. This operation of multiplication of a vector by a scalar has the following formal properties

$$r(v + w) = rv + rw, \qquad 1v = v, \tag{1}$$

$$(r + s)v = rv + sv, \qquad (rs)v = r(sv). \tag{2}$$

for all vectors v and w and all scalars $r, s \in \mathbf{R}$. In view of these properties of addition and of scalar multiplication, we say that the vectors form a *vector space* over the real numbers.

Many geometric facts can be established readily by the use of vector algebra. For example, this algebra gives an easy proof of the theorem that the three medians of a triangle meet in a point.

Vectors may also be introduced in 3-dimensional spaces (and in higher dimensional spaces); moreover, exactly the same formulas (1) and (2) apply, whatever the dimension. More generally, a *vector space* V over a field F is an abelian group (under addition) with another operation $F \times V \to V$, written $(r,v) \mapsto rv$ for elements $r \in F$, $v \in V$, which satisfies with addition the four laws (1) and (2) above. One can also describe formally the "dimension" of such a space. In the plane, if u_1 and u_2 are two non-collinear vectors, every vector v in the plane can be written as a "linear combination" $v = x_1u_1 + x_2u_2$ with unique choices for the scalars x_1 and x_2. In three dimensions, we need three such vectors u_1, u_2 and u_3, along three "axes". More generally, vectors $u_1, \ldots, u_n$ in V are said to form a *basis* for the vector space V if every vector v in V has a unique expression

$$v = x_1u_1 + \cdots + x_nu_n ;$$

then the x_i are the *coordinates* of v relative to the basis u_i. A first theorem of vector geometry ($\doteq$ linear algebra) states that any two (finite) bases for a given vector space must have the same number of elements. This number n is the *dimension* of the space.

Vectors also provide a convenient description for certain functions transforming the plane into itself (and leaving the origin fixed): Transformations such as expansions from the origin ($v \mapsto rv$) for a fixed r, compressions on a line through the origin, and shears. All of these transformations preserve the vector operations. More generally, a *linear transformation* on V is a function t: $V \to V$ which preserves both addition of vectors and multiplication of vectors by scalars. This means that the equations

$$t(v + w) = tv + tw, \qquad t(rv) = r(tv) \tag{3}$$

hold for all vectors v and w and all scalars r. For example, a rotation about the chosen origin 0 is a linear transformation (it maps the parallelogram of addition into a parallelogram). The same notion of linear transformation applies to higher dimensional vector spaces, and is the cornerstone of the study of linear algebra and matrix theory (Chapter VII).

The general idea that some geometrical facts can be handled well by appropriate algebraic operations has now appeared in the use of real numbers for one-dimensional geometry and of vectors for higher dimensions. The addition of vectors was possible only with the choice of an origin 0. Without a choice of origin one still has some algebraic operations such as the construction of the midpoint, written $(1/2)P + (1/2)Q$, of two points P and Q or of the point $(2/3)P + (1/3)Q$ which is one third of the way from P to Q. The general such operation is the formation of the *weighted average*

$$P = w_0P_0 + w_1P_1 + \cdots + w_nP_n \tag{4}$$

of $n + 1$ points with weights real numbers w_i with sum $w_0 + \cdots + w_n = 1$. A transformation preserving such averages is called an *affine* transformation; the resulting geometry (in any number of dimensions) is *affine geometry*. It is possible to write a complete set of axioms for the weighted average operations (4), but they are cumbersome (*Algebra*, first edition). Even the simplest ternary operation $P_0 - P_1 + P_2$ of this type has not proven to be very useful. It seems more efficient to reduce to vector algebra by choosing an origin. Addition is a very convenient operation! Binary operations are much handier than ternary ones.

8. Analytic Geometry

Another and earlier reduction of plane geometry to algebra is provided by the familiar method of cartesian coordinates. Given an orientation, a choice of origin and unit, and two perpendicular coordinate axes, each point P is represented by its coordinates, a pair (x,y) of real numbers. Each line in the plane may be described as the set of those points whose coordinates satisfy a linear equation, while the distance between two points is given by the familiar formula in coordinates, derived from the Pythagorean theorem. In this way, all sort of geometric facts about the plane are handled by algebraic machinery; in effect, the plane is reduced to the cartesian product $\mathbf{R} \times \mathbf{R} = \mathbf{R}^2$ of two copies of the real line $\mathbf{R}$. However, this reduction does depend on the choice of origin *and* of axes, and one must betimes verify that truly geometric facts are independent of the choice of coordinates.

Geometry and coordinates arise first in dimensions 2 and 3. The need for higher dimensional geometry is motivated by phenomena which need specification of more than three coordinates: Events in space-time need four (position and time); in dynamics, the initial conditions for a particle need six, three for position and three for velocity. The use of such a six-dimensional *phase space* is a first example of the importance of mechanics as a motivation for mathematical developments (Chapter IX). To be sure, this example could be (but usually is not) described without explicit use of coordinates, as the product of a three-dimensional position space and another three-dimensional velocity space. A general n-dimensional Euclidean space may be constructed as the n-fold product $\mathbf{R}^n$ of real lines, with points the n-tuples of real numbers $(x_1, \ldots, x_n)$. This is a vector space over $\mathbf{R}$, with basis the n "unit vectors" $(0, \ldots, 0,1,0, \ldots, 0)$. Such use of coordinates allows for the efficient treatment of higher-dimensional phenomena in dimensions where the corresponding geometric intuitions are weak or non-existent. It serves to extend geometric ideas beyond the ordinary three dimensions.

Much of mathematical physics deals with phenomena in three dimensions and so often requires a formulation in triads of equations, one for each coordinate in $\mathbf{R}^3$. When $\mathbf{R}^3$ is regarded as a vector space, many of these equations can be written as single vector equations. This accounts for the popularity of vector analysis in Physics. It includes the use of inner products, to which we now turn.

9. Trigonometry

Trigonometry is essentially a procedure for turning angular measures into linear measures. This appear directly in the definition of the two basic trigonometric functions $\sin \theta$ and $\cos \theta$ of the angle θ. In the oriented plane take a point P on the unit circle with center at the origin 0 so that the segment $0P$ makes the given angle θ with the x coordinate axis. Then $\cos \theta$ and $\sin \theta$ are defined to be the x and y coordinates of P; since the circle has radius 1, this immediately gives the identity

$$\cos^2\theta + \sin^2\theta = 1. \tag{1}$$

This defines $\sin \theta$ and $\cos \theta$ for angles θ of all sizes (not just for an angle θ in the first quadrant, as displayed in Figure 1).

When angles are measured (in radians) by numbers t we can also think of the sine and the cosine as functions of the *number* t, so that

$$\operatorname{Sin} t = \sin(\text{angle of } t \text{ radians}). \tag{2}$$

Thus there are really two legally different functions: The Sine of a *number*, here with capital S, and the sine of an angle, with lower case s. This pedantic (but real!) difference is usually ignored. It implicitly

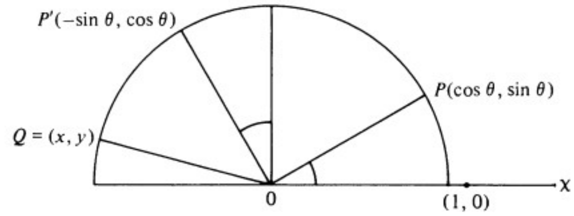

Figure 1

involves the wrapping function θ_t of §2 above. Recall that this function sends each real t to the point $P = (\cos\theta_t, \sin\theta_t)$ on the circle such that the length of the counterclockwise circular arc from (1,0) to P is congruent to t, modulo 2π. Then the definition (2) reads

$$\text{Sin } t = \sin(\theta_t), \qquad \text{Cos } t = \cos(\theta_t): \tag{3}$$

it really amounts to a composite function $t \to \theta_t \to \sin(\theta_t)$. Since the wrapping function has $\theta_{t+2\pi} = \theta_t$ we get

$$\text{Sin}(t + 2\pi) = \text{Sin}(t), \qquad \text{Cos}(t + 2\pi) = \text{Cos}(t). \tag{4}$$

In other words, Sin and Cos are periodic functions, of period 2π, as in the familiar graph of Figure 2. Now that this has been stated, we drop the S in Sin and the function θ_t; they would just get in the way of trigonometric manipulations, but we emphasize that the wrapping function accounts for radian measure and the periodicity (4) of the trigonometric functions. Indeed, the whole study of periodic functions, as carried on in Fourier analysis, concerns the expression of more or less arbitrary functions $f(t)$ of period 2π in terms of the periodic functions $\sin nt$ and $\cos nt$ for the natural numbers n.

A rotation of the circle of Figure 1 about the origin through the angle θ will carry the point (1,0) to our point P and the point (0,1) to a corresponding point P'; thus this rotation has the effect

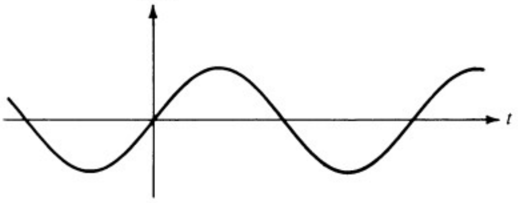

Figure 2

$$(1,0) \mapsto (\cos\theta, \sin\theta), \qquad (0,1) \mapsto (-\sin\theta, \cos\theta) \tag{5}$$

Since (1,0) and (0,1) form a basis for the 2-dimensional vector space, any point Q with coordinates (x,y) can be written as the linear combination $(x,y) = x(1,0) + y(0,1)$ of the two basis vectors. Now rotation is, as already noted, a linear transformation, so preserves linear combinations and thus by (5) has the effect

$$(x,y) \mapsto (x\cos\theta - y\sin\theta,\ x\sin\theta + y\cos\theta).$$

When one writes (x',y') for the coordinates of the rotated point, this yields the customary equations in coordinates

$$\begin{aligned} x' &= x\cos\theta - y\sin\theta, \\ y' &= x\sin\theta + y\cos\theta \end{aligned} \tag{6}$$

for a rotation $(x,y) \mapsto (x',y')$ of the plane about the origin. Also, if one takes $Q = (x,y)$ to be that point Q on the unit circle of Figure 1 for which $0Q$ makes the angle ϕ with the positive x-axis, then $x = \cos\phi$, $y = \sin\phi$. Now the rotation by θ clearly carries Q to a point Q' where $0Q'$ makes the angle $\phi + \theta$ with the positive x-axis. The coordinates (x',y') of Q' are then $\cos(\phi + \theta)$, $\sin(\phi + \theta)$ and the equations (6) for rotation become

$$\begin{aligned} \cos(\phi + \theta) &= \cos\phi\cos\theta - \sin\phi\sin\theta, \\ \sin(\phi + \theta) &= \cos\phi\sin\theta + \sin\phi\cos\theta. \end{aligned} \tag{7}$$

These "addition formulas" for the trigonometric functions are thus another expression of the relation of angular to linear measure.

This relation also accounts for the use of trigonometric functions in calculating the connections between the sides and the angles of a triangle, as suggested by the Euclidean congruence theorems. In a triangle $0AB$ with sides a, b and c, as illustrated in Figure 3, drop a perpendicular from A to the opposite side at M. If θ and ϕ are the angles of the triangle at 0 and

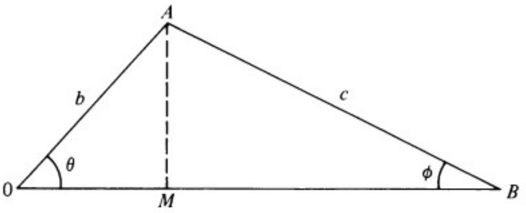

Figure 3

at B, the definition of the sine function shows that the length AM can be expressed either as $b \sin \theta$ or as $c \sin \phi$; this gives the *law of sines*: $b/c = \sin \phi / \sin \theta$. On the other hand, one may compute the length c of the side AB from the Pythagorean theorem for the right triangle AMB to be

$$c^2 = (b \sin \theta)^2 + (a - b \cos \theta)^2, \tag{8}$$

$$c^2 = a^2 + b^2 - 2ab \cos \theta.$$

This is the *law of cosines*. When $u = 0B$ and $v = 0A$ are regarded as vectors of the respective lengths $|0B| = a$, $|0A| = b$, the expression $ab \cos \theta$ which appears in (8) is called the *inner product* $u \cdot v$ of the two vectors; it is a scalar (a real number). If A has coordinates (x_1, y_1) and B coordinates (x_2, y_2), the lengths a, b, and c in (8) can be written in terms of these coordinates by the Pythagorean theorem; the formula (8) then becomes a formula for the inner product $u \cdot v$:

$$(x_1, y_1) \cdot (x_2, y_2) = x_1 x_2 + y_1 y_2. \tag{9}$$

This again is a linearization of angular phenomena.

With these ideas and formulas we have presented all of the essential concepts of trigonometry in less than 6 pages. It is curious (and troublesome) that the standard elementary presentations of trigonometry have been inflated to inordinate lengths–an inflation hardly justified by the inevitable subtlety wrapped up in the comparison of linear and angular magnitude.

In (9), the inner product is expressed in terms of a choice of coordinates. However, as in other cases, the inner product can be described in a more invariant way by suitable axioms. It has three characteristic properties: *Linearity*, *symmetry*, and *positive definiteness*, as expressed by the three equations

$$u \cdot (t_1 v_1 + t_2 v_2) = t_1 (u \cdot v_1) + t_2 (u \cdot v_2), \tag{10}$$

$$u \cdot v = v \cdot u, \tag{11}$$

$$u \cdot u \geqslant 0, \qquad u \cdot u = 0 \quad \text{only if} \quad u = 0, \tag{12}$$

valid for all vectors u, v, v_1, and v_2 and all scalars t_1 and t_2. All the properties of the inner product follow from these equations. Indeed, given these equations–for a product $u \cdot v$ of vectors in a two dimensional vector space over **R**–one can prove that there is a choice of basis and corresponding coordinates for which the given inner product is expressed by the formula (9). Such a basis is called a *normal orthogonal basis* (for the given inner product).

We will extend these ideas to higher dimensions in §VII.5.

(6 &

CHAPTER III

Geometry

This chapter is concerned with the origins of elementary plane geometry, its close relation to groups of transformations, and the notions of non-Euclidean geometry. Subsequently, n-dimensional geometry will appear with linear algebra in Chapter VII, manifolds and spaces with curvature in Chapter VIII, and topology, in its connections with complex analysis, in Chapter X. These and other aspects of geometry pervade Mathematics.

1. Spatial Activities

Geometrical figures and facts arise from a wide array of activities and observations. Some involve motion; thus in watching motion we see objects falling in vertical lines, water waves expanding in circles from a dropped stone, duneland grasses moved by the wind to describe semicircles on the sand, tree branches oscillating back and forth in the wind, long straight lines of ocean waves approaching a beach, and the like. There are also motions which we initiate and then watch–the thrown ball falling, the log rolling downhill, the circle turned into a wheel for a cart, and so on. Some activity involves construction. Thus a post or a column will stay upright in balance if it is set perpendicular to the floor as a vertical; a three-legged stool is more likely to be steady than one constructed with four legs; rods joined to form a triangle stay rigid, but rods joined in a square may wobble. Also, pieces of a board cut apart will fit together again to make up the same size board. Mapping a labyrinth or painting a scene calls for reproducing a variety of shapes, each on a smaller scale. Fitting an object inside another highlights differences in shapes. To check ahead of time that a fit will be possible may require a measurement of length or of circumference. Estimating separation at a distance requires lines of sight, angles, and triangles. These are but a few of the many activities that go into the formation of geometrical ideas.

From these activities one may disengage various figures: Circles and ovals; lines, horizontal, vertical, and perpendicular; triangles, equilateral or otherwise; squares and rectangles; cubes, cylinders, as with logs or rods, and spheres, as with oranges or balloons. One also discovers a variety of facts about these figures: A triangle is rigid because three sides (or two sides and the included angle) determine the triangle up to rigid motion. A right triangle can come in various shapes, but in every case the square constructed on the longest side is the sum of the squares on the two shorter legs (a fact said to have been useful in re-establishing property lines after the Nile had been in flood). Two different perpendiculars to the same line would never meet. Two triangles with corresponding angles equal have corresponding sides proportional—and so on.

After a considerable array of facts about geometrical figures are at hand, some connections between these facts come to light. Once the area of a triangle can be computed from the lengths of its sides, it is possible to fit four congruent triangles together so as to prove the Pythagorean theorem, by adding up the five areas in Figure 1. Or, the Pythagorean theorem can be deduced from facts about ratios of sides in similar triangles.

Explicitly (Figure 2), the perpendicular from the right angle divides the hypotenuse c in Figure 2 into pieces h and k with $h/a = a/c$ and $k/b = b/c$. Clearing of fractions, $hc = a^2$ and $kc = b^2$, so that $(h + k)c = c^2 = a^2 + b^2$, q.e.d. Many more such logical relations between geometrical facts gradually appear.

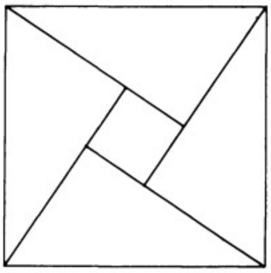

Figure 1. The Pythagorean theorem.

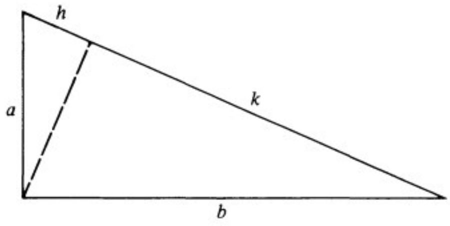

Figure 2

Nevertheless, it is truly remarkable that our experiences of spatial extent and of motion through space can be organized so completely. By taking certain figures–line, plane, segment, angle, and circle–as basic ones, and by assuming certain simple facts as axioms, other geometric figures can be constructed and other geometric facts deduced from the assumed axioms. Some of the resulting facts are surprising; for example, the three medians of a triangle meet in a point, as do the three altitudes. The resulting deductive structure of Euclidean geometry is a model of Mathematical method.

2. Proofs Without Figures

In elementary Euclidean geometry, proofs of various facts about figures are developed from the axioms, with occasional implicit use of intuitively evident properties of the figures concerned. For instance, to find the perpendicular bisector of a segment AB one constructs two circles with centers A and B and both with radius AB; these circles obviously intersect on opposite sides of the line AB, ... But what is meant by a "side" of AB and why must the circles intersect? (See Figure 1.) If *all* facts are to be deduced from axioms, then these plausible observations must also be proved from the axioms, rather than from inspection of the figure. Once this austere requirement was recognized, it turned out to be possible to add suitable new axioms to those of Euclid so that all arguments could be free of any intuitive or pictorial content. This meant that geometry was not really "about" the figures themselves, but was about certain corresponding notions defined exclusively in terms of basic notions to be taken as primitive or undefined. This austere program was systematically carried out by David Hilbert in a book, *Grundlagen der Geometrie* (B. G. Teubner, first edition 1899, twelfth edition 1977). His axioms were for both plane and solid geometry; we will examine only those for the plane. He presented the axioms in the five following groups:

Group I (Incidence Axioms). These axioms involve only three primitive notions: "point", "line", and "the point P lies on the line l". They require:

(1) Two distinct points P and Q lie on one and only one line. (Except for the meticulous logical formulation, this is the familiar requirement that "two points determine a line".)
(2) There are at least two points on every line.
(3) There are three points P, Q, R not all on a line.

With these axioms, one can define a *triangle* (the figure formed by three non-collinear points). However, one cannot yet define "right angle".

)*+,)))-) #6 3 @ J ? D A 7
+ ": 3 " 3 / = 3 " % F 3
= L @ D "@ 6 ; / % 6@ G " 3"
9 # $ $ #
M> ' #& (%5 " " M> E :
5 "

data 15.04.2020

137) Pi

BBP-Type formulas for Mathematical Constants

Bailey

dhbailey@lbl.gov

Source:

www.davidhbailey.com//dhbpapers/bbp-formulas.pdf

others:

www.davidhbailey.com/dhbpapers/pi-formulas.pdf

—

www.maa.org/sites/default/files/pdf/pubs/amm_supplements/Monthly_Reference_9.pdf

—

mathscholar.org/2019/10/pi-as-the-limit-of-n-sided-circumscribed-and-inscribed-polygons/

—

data 15 04 2020

New paper proves 80-year-old-
approximation conjecture
mathscholar.org/2019/09/
new-paper-proves-80-year-
old-approximation-conjecture/

—

www.davidhbailey.com//dhbpapers/

—

138) 20200416

A) Dancing with a black hole

Source: www.nytimes.com/2020/04/16/
science/black-hole-sagittarius-
a.html

Einstein → gravity to a warping of the geometry of space and time.

In our solar system, astronomers had long known that the planet Mercury behaves like this; its perihelion advances around the sun by the microscopic amount of 43 seconds of arc a century.

Astronomers couldn't explain why. In November 1915, Einstein tried out his new theory on Mercury's motion.

data 15 04 2020

New paper proves 80-year-old-
approximation conjecture
mathscholar.org/2019/08/
new-paper-proves-80-year-
old-approximation-conjecture/

www.davidhbailey.com//dhbpapers/

138) 20200416

A) Dancing with a black hole
Source: www.nytimes.com/2020/04/16/
science/black-hole-sagittarius-
a.html

Einstein → gravity to a warping of
the geometry of space and
time.

In our solar system, astronomers had long known that the planet Mercury behaves like this; its perihelion advances around the sun by the microscopic amount of 43 seconds of arc a century.

Astronomers couldn't explain why.
In November 1915, Einstein tried out his new theory on Mercury's motion.

* $$ + $ $ ($,$ $- .

-

data 16.04.2020

B) A star orbiting the Milky Way's* giant black hole confirms Einstein was right.

Source: www.sciencenews.org/article/star-orbiting-milky-way-giant-black-hole-confirms-einstein-was-right/

by Emily Conover
20200416

⇒ Decades of observations revealed the rotation of the star's elliptical orbit.

1915 - Einstein realized that his newly formulated general theory of relativity explained a weird quirk in the orbit of Mercury.

Star S2.

The researchers previously had used observations of S2 to identify a different effect of general relativity, the reddening of the star's light due to what's called gravitational redshift (SN 7/26/18).

Emily Conover

www.sciencenews.org/article/star-orbiting-black-hole-einstein-gravity-general-relativity

data 16.04.2020

Now, they've determined that* the ellipse rotates over time, what's known as Schwarzschild precession. That precession is the result of the warping of spacetime caused by massive objects, according to general relativity.

c) Stephen Wolfram's hypergraph project aims for a fundamental theory of physics

Source: www.sciencenews.org/article/stephen-wolfram-hypergraph-project-fundamental-theory-physics
April 14, 2020
by Tom Siegfried

→ causal invariance.

Thanks to causal invariance Wolfram's hypergraphs reproduce many of the consequences of various physical theories, such as Einstein's special theory of relativity. Traveling rapidly slows down time (as special relativity says) because hypergraph structures corresponding to moving objects make an angle through the hypergraph that extends the distance between updates (or time steps). The speed of light is a

- /012/ / /3 / #& 0 - .+ 4 +
- . 5 " 6 7 *
8*" "
/012/ / /3 / #& 0 *
9* " + 7 *

* $$ $ $ - .+ + - .

* $$ $ $ * *

+* *

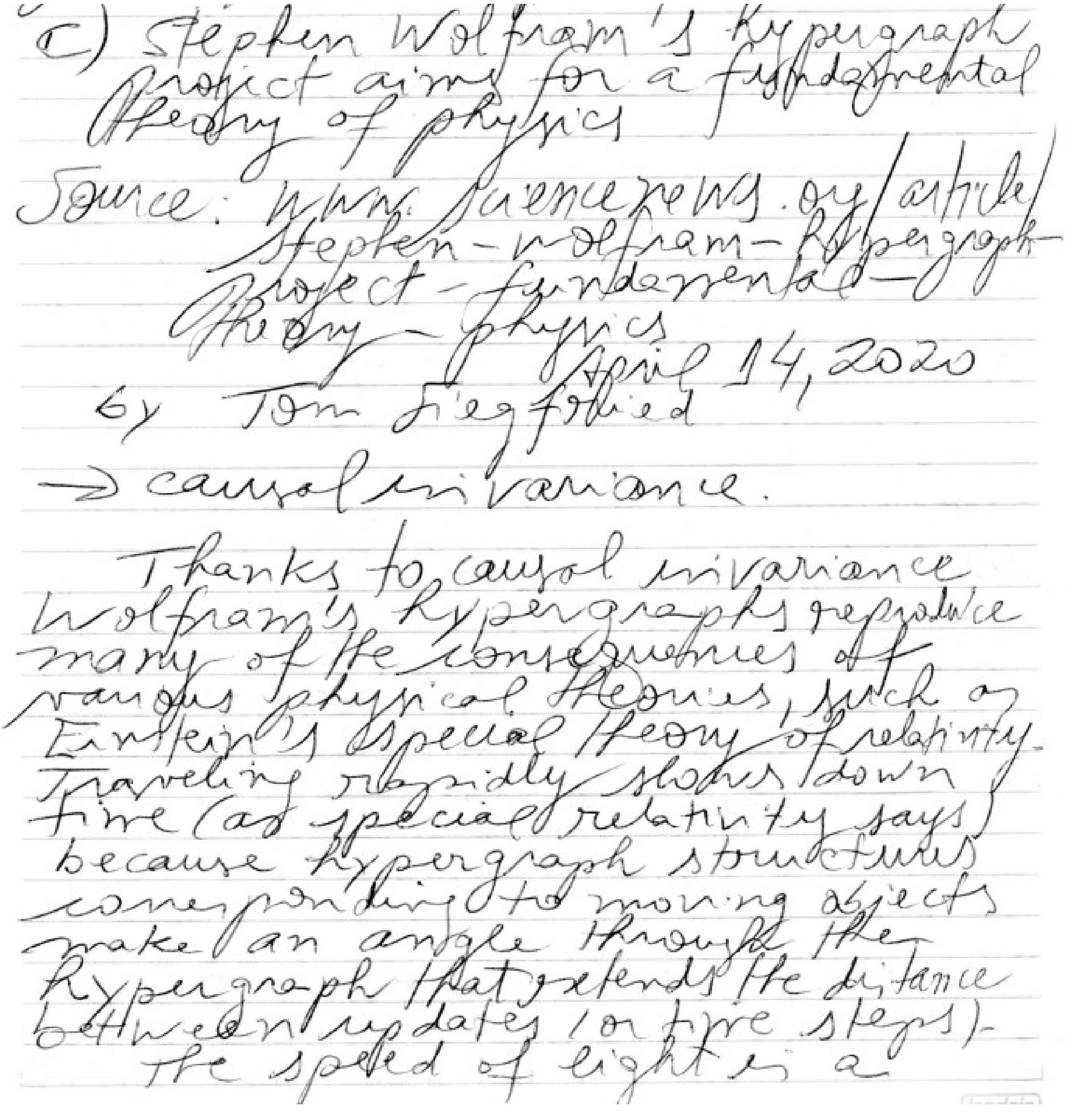

c) Stephen Wolfram's hypergraph
project aims for a fundamental
theory of physics

Source: www.sciencenews.org/article/
stephen-wolfram-hypergraph-
project-fundamental-
theory-physics
April 14, 2020
by Tom Siegfried

→ causal invariance.

Thanks to causal invariance
Wolfram's hypergraphs reproduce
many of the consequences of
various physical theories, such as
Einstein's special theory of relativity.
Traveling rapidly slows down
time (as special relativity says)
because hypergraph structures
corresponding to moving objects
make an angle through the
hypergraph that extends the distance
between updates (or time steps).
The speed of light is a

16.04.2020

Now, they've determined that* the ellipse rotates over time, what's known as Schwarzschild precession. That precession is the result of the warping of spacetime caused by massive objects, according to general relativity.

c) Stephen Wolfram's hypergraph project aims for a fundamental theory of physics

Source: www.sciencenews.org/article/stephen-wolfram-hypergraph-project-fundamental-theory-physics
April 14, 2020
by Tom Siegfried

→ causal invariance.

Thanks to causal invariance Wolfram's hypergraphs reproduce many of the consequences of various physical theories, such as Einstein's special theory of relativity. Traveling rapidly slows down time (as special relativity says) because hypergraph structures corresponding to moving objects make an angle through the hypergraph that extends the distance between updates (or time steps). The speed of light is a

16.04.2020

maximum velocity, as relativity states, because it represents the maximum rate that information can spread through the hypergraph as it updates. And gravity — described by Einstein's general theory of relativity — emerges in the relationships between features in the hypergraph that can be interpreted as matter particles. (Particles would be small sets of linked points that persist as the hypergraph updates, something like "little lumps of space" with special properties.)

supersmall unknown particles, which he calls oligons

Such oligons, subject only to gravity, would now be hanging out in and around galaxies utterly unnotified — except for their gravitational impact.

Oligons might therefore explain why astronomers infer the existence of vast amounts of invisible "dark matter" in space.

"right rule"

That "right rule" would generate a hypergraph with our universe's specific properties:

data 16.04.2020

- three (apparent) dimensions* of space,
- the right cosmic expansion rate,
- the right repertoire of elementary particles with the correct charges and masses, and other features.

Source:

- www.wolframphysics.org/technical-introduction/
- technical-introduction/limiting-behavior-and-emergent-geometry/hyperbolic-space/

I) Wolfram-ModelsForPhysics.pdf

Source:
www.wolframphysics.org/technical-introduction/inc/Wolfram-ModelsForPhysics.pdf

Pag. 105 – Curvature

$$2\pi a^2\left(1-\cos\left(\frac{r}{a}\right)\right) =$$

$$\pi r^2\left(1-\frac{r^2}{12a^2}+\frac{r^4}{360a^4}-\dots\right)$$

Volume:

$$\frac{\pi^{d/2}}{(d/2)!} r^d\left(1-\frac{r^2}{6(d+2)}R+O(r^4)\right)$$

" * $$ * + $ " $ $4 " : + *"
/012/ / /3 / #& 4 " : + *"
' (& 5

4.7 Curvature

In ordinary plane geometry, the area of a circle is πr^2. But if the circle is drawn on the surface of a sphere of radius a, the area of the spherical region enclosed by the circle is instead:

$$2\pi a^2 (1 - \cos(\frac{r}{a})) = \pi r^2 (1 - \frac{r^2}{12 a^2} + \frac{r^4}{360 a^4} - \ldots)$$

In other words, curvature in the underlying space introduces a correction to the growth rate for the area of the circle as a function of radius. And in general there is a similar correction for the volume of a d-dimensional ball in a curved space (e.g. [24][1:p1050]):

$$\frac{\pi^{d/2}}{(d/2)!} r^d (1 - \frac{r^2}{6(d+2)} R + O(r^4))$$

where here R is the Ricci scalar curvature of the space [25][26][27]. (For example, for the d-dimensional surface of a $(d+1)$-dimensional sphere of radius a, $R = \frac{(d-1)\,d}{a^2}$.)

Now consider the sequence of “sphere” graphs:

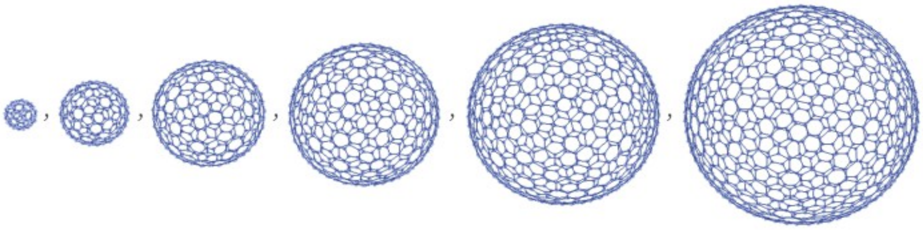

We can compute V_r for each of these graphs. Here are the log differences $\Delta(r)$ (the error bars come from the different neighborhoods associated with hexagonal and pentagonal “faces” in the graph):

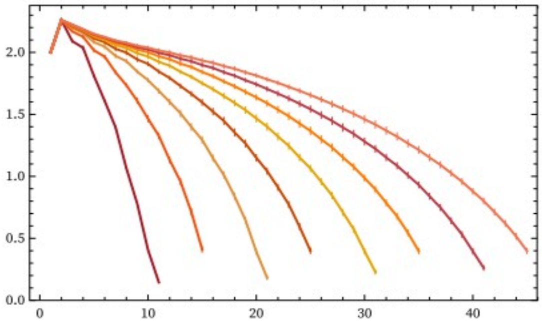

data 16.04.2020

- three (apparent) dimensions of space,
- the right cosmic expansion rate,
- the right repertoire of elementary particles with the correct charges and masses, and other features.

Source:

- www.wolframphysics.org/technical-introduction/
- technical-introduction/limiting-behavior-and-emergent-geometry/hyperbolic-space/

D) Wolfram-ModelsForPhysics.pdf
Source:
www.wolframphysics.org/technical-introduction/ine/Wolfram-ModelsForPhysics.pdf

Pag. 105 - Curvature

$$2\pi a^2\left(1-\cos\left(\frac{r}{a}\right)\right) =$$

$$\pi r^2\left(1-\frac{r^2}{12a^2}+\frac{r^4}{360a^4}-\dots\right)$$

Volume:

$$\frac{\pi^{d/2}}{(d/2)!} r^d\left(1-\frac{r^2}{6(d+2)}R+O(r^4)\right)$$

16.04.2020

Where here R is the Ricci scalar curvature of the space.

Pag. 371 $E = mc^2 + \frac{1}{2} m v^2$

E) Some Relativistic and Gravitational Properties of the Wolfram
April 12, 2020

Source:
www.wolframcloud.com/obj/wolframphysics/Documents/some-relativistic-and-gravitational-properties-of-the-wolfram-model.pdf

Pag. 31 Figure 16:

Two-dimensional spatial hypergraphs with different Ollivier-Ricci scalar curvatures, with the presence of pentagons indicating regions of positive scalar curvature and the presence of heptagons indicating regions of negative scalar curvature.

Figure 19 (Pag. 40)

Geodesic trajectories of test particles (starting on initially parallel paths) through discrete spaces containing a spherical mass distribution. Adapted from S. Wolfram, A New Kind of Science, pg 531.

Einstein–Hilbert action—thereby leading to the conclusion that $R_{\mu\nu}$ must satisfy exactly the usual vacuum Einstein equations [114][115][75][116]:

$$R_{\mu\nu} - \frac{1}{2} R\, g_{\mu\nu} = 0$$

A full derivation of this is given in [113]. Causal invariance plays a crucial role, ensuring for example that timelike directions t^i associated with different foliations give invariant results. Much like in the derivation of continuum fluid behavior from microscopic molecular dynamics (e.g. [110]), one also needs to take a variety of fairly subtle limits, and one needs sufficient intrinsic generation of effective randomness [1:7.5] to justify the use of certain statistical averages.

But there is a fairly simple interpretation of the result above. Imagine all the geodesics that start at a particular point in the causal graph. The further we go, the more possible geodesic paths there will be in the graph. To achieve a power law corresponding to a definite dimension, the geodesics must in a sense just "stream outwards", evenly distributed in direction.

But the Ricci tensor specifically measures the rate at which bundles of geodesics change their cross-sectional area. And as soon as this change is nonzero, it will inevitably change the local density of geodesics and eventually grow to disrupt the power law. And so the only way a fixed limiting dimension can be achieved is for the Ricci curvature to vanish, just as it does according to the vacuum Einstein equations. (Note that higher-order terms, involving for example the Weyl tensor and other components of the Riemann tensor, yield changes in the shape of bundles of geodesics, but not in their cross-sectional area, and are therefore not constrained by the requirement of fixed limiting dimension.)

8.8 Matter, Energy and Gravitation

In our models, not only space, but also everything "in space", must be represented by features of our evolving hypergraphs. There is no notion of "empty space", with "matter" in it. Instead, space itself is a dynamic construct created and maintained by ongoing updating events in the hypergraph. And what we call "matter"—as well as things like energy—must just correspond to features of the evolving hypergraph that somehow deviate from the background activity that we call "space".

Anything we directly observe must ultimately have a signature in the causal graph. And a potential hypothesis about energy and momentum is that they may simply correspond to excess "fluxes" of causal edges in time and space. Consider a simple causal graph in which we have marked spacelike and timelike hypersurfaces:

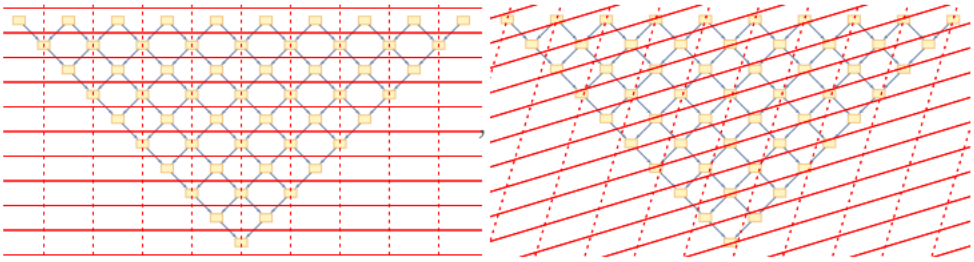

The basic idea is that the number of causal edges that cross spacelike hypersurfaces would correspond to energy, and the number that cross timelike hypersurfaces would correspond to momentum (in the spatial direction defined by a given hypersurface). Inevitably the results one gets would depend on the hypersurfaces one chooses, and so would differ from one observer to another.

And one important feature of this identification of energy and momentum is that it would explain why they follow the same relativistic transformations as time and space. In effect space and time are probing distances between nodes in the causal graph (as measured relative to a particular foliation), while momentum and energy are probing a directly dual property: the density of edges.

There is additional subtlety here, though, because causal edges are needed just to maintain the structure of spacetime—and whatever we measure as energy and momentum must just be some excess in the density of causal edges over the "background" corresponding to space. But even to know what we mean by density we have to have some notion of volume, but this is also itself defined in terms of edges in the causal graph.

But as a rough idealized picture, we might imagine that we have a causal graph that maintains the same overall structure, but adds some extra connections:

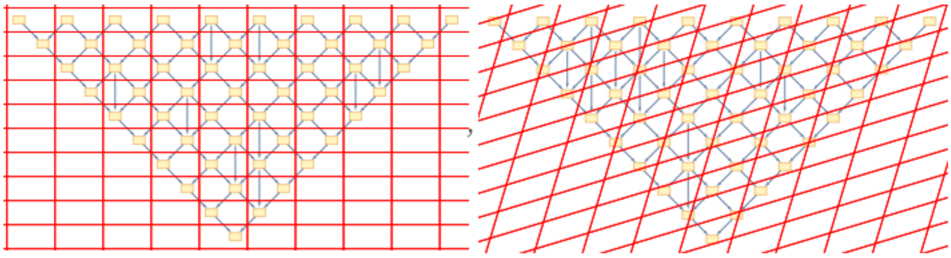

In our actual models, the causal graphs one gets are considerably more complicated. But one can still identify some features from the simple idealization. The basic concept is that energy and momentum add "extra causal connections" that are not "necessary" to define the basic structure of spacetime. In a sense the core thing that defines the structure of spacetime is the way that "elementary light cones" are knitted together.

Consider a causal graph like:

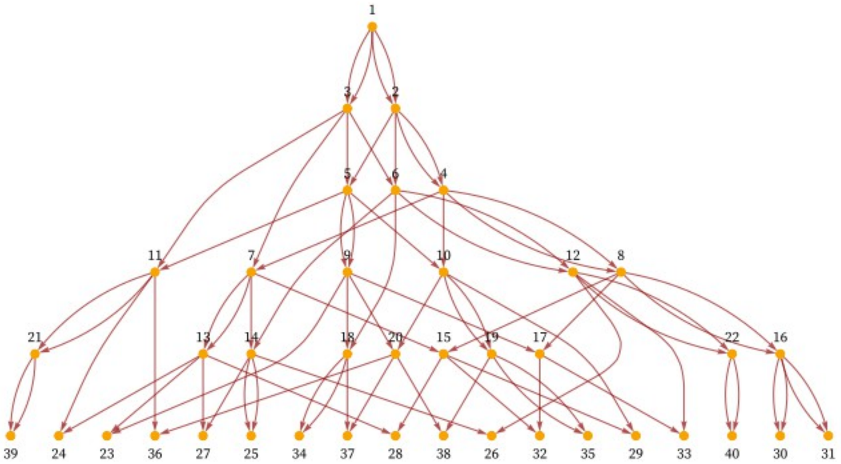

One can think of a set of edges like the ones indicated as in effect “outlining” the causal graph. But then there are other edges that add “extra connections”. The edges that “outline the graph” in effect maximally connect spatially separated regions—or in a sense transmit causal information at a maximum speed. The other edges one can think of as having slower speeds—so they are typically drawn closer to vertical in a rendering like the one above.

But now let us return to our simple grid idealization of the causal graph—with additional vertical edges added. Now do foliations like the ones we used above to represent inertial frames, parametrized by a velocity ratio β relative to the maximum speed (taken to be 1). Define $E(\beta)$ to be the density of causal edge crossing of the spacelike hypersurfaces, and $p(\beta)$ the corresponding quantity for timelike hypersurfaces. Then for speed 1 edges, we have (up to an overall multiplier) (cf. [111][112]):

$$E(\beta) = p(\beta) = \frac{1+\beta}{\sqrt{1-\beta^2}}$$

But in general for edges with speed α we have

$$E(\beta) = \frac{1-\alpha\,\beta}{\sqrt{1-\beta^2}},\;\; p(\beta) = \frac{\alpha-\beta}{\sqrt{1-\beta^2}}$$

which means that for any β

$$E\,(\beta)^2 - p\,(\beta)^2 \;=\; 1-\alpha^2$$

thus showing that our crossing densities transform like energy and momentum for a particle with mass $\sqrt{1-\alpha^2}$. In other words, we can potentially identify edges that are not maximum speed in the causal graph as corresponding to “matter” with nonzero rest mass. Perhaps not surprisingly, this whole setup is quite analogous to thinking about world lines of massive particles in elementary treatments of relativity.

But in our context, all of this must emerge from underlying features of the evolving hypergraph. Causal connections that transfer information at maximum speed can be thought of as arising from updating events that involve maximally separate nodes, and that are somehow always entraining "fresh" nodes. But causal connections that transfer information more slowly are associated with sequences of updating events that in effect reuse nodes. So in other words, rest mass can be thought of as being associated with local collections of nodes in the hypergraph that allow repeated updating events to occur without the involvement of other nodes.

Given this setup, it is possible to derive other features of energy, momentum and mass by methods similar to those used in typical discussions of relativity. It is first helpful to include units in the quantities we have introduced. If an elementary light cone has timeline extent T then we can consider its spacelike extent to be $c\,T$, where c is the speed of light. Within the light cone let us say that there effectively μ causal edges oriented in the timelike direction. With the inertial frame foliations used above, the contribution of these causal edges to energy and momentum will be (the factor c in the energy case comes from the spacelike extent of the light cone):

$$E(\beta) = c\,\frac{\mu}{\sqrt{1-\beta^2}} = c\,\mu\,(1+\frac{\beta^2}{2}+O(\beta^4))$$

$$p(\beta) = \mu\,\frac{\beta}{\sqrt{1-\beta^2}} = \mu\,\beta\,(\,1+O(\beta^2)\,)$$

But if we define the mass m as $\frac{\mu}{c}$ and substitute $\beta = \frac{v}{c}$, we get the standard formulas of special relativity [111][112], or to first order

$$E = m\,c^2 + \frac{1}{2}\,m\,v^2$$

$$p = m\,v$$

establishing in our model the relation $E = m\,c^2$ between energy and rest mass.

We should note that with our identification for energy and momentum, the conservation of energy becomes essentially the statement that the overall density of events in the causal network does not change as we progress through successive spacelike surfaces. And, as we will discuss later, if in effect the whole hypergraph is in some kind of dynamic equilibrium, then we can reasonably expect that this will be the case. Expansion (or, more specifically, non-uniform expansion) will lead to effective violations of energy conservation, much as it does for an expanding universe in the traditional formalism of general relativity [117][75].

In the previous subsection, we discussed the overall structure of spacetime, and we used the growth rate of the spacetime volume $C_t(X)$ as a way to assess this. But now let us ask about specific values of $C_t(X)$, complete with their "constant" multipliers. We can think of these multipliers as probing the local density of the causal graph. But deviations in this are what we have now identified as being associated with matter.

To compute $C_t(X)$ we ultimately need to be able to precisely count events in the causal graph. If the causal graph is somehow “uniform”, then it cannot contain what can be considered to be “matter”. In the setup we have defined, the presence of matter is effectively associated with “fluxes” of causal edges that reflect the non-uniform “arrangement” of nodes in the causal graph. To represent this, take $\rho\,(X)$ to be the “local density” of nodes in the causal graph. We can make a series expansion to probe deviations from uniformity in $\rho\,(X)$. And formally we can write

$$\rho(X) = \rho_0\,(\,1 + \sigma\,\delta t^\mu\,\delta t^\nu\,T_{\mu\nu}\,\ldots)$$

where the t^μ are timelike vectors used in the definition of C_t and now $T_{\mu\nu}$ is effectively a tensor that represents “fluxes of edges” in the causal graph. But these fluxes are what we have identified as energy and momentum, and when we think about how causal edges traverse spacelike and timelike hypersurfaces, $T_{\mu\nu}$ turns out to correspond exactly to the standard energy-momentum tensor of general relativity.

So now we can combine our formula for the effect of local density with our formula for the effect of curvature from the previous section to get:

$$C_t(X) = \rho_0(1 + \sigma\,\delta t^i\,\delta t^j\,T_{ij} + \ldots\,)\;t^{d+1}(1 - \frac{1}{6}\,\delta t^i\,\delta t^j\,R_{ij}\,+ \ldots)$$

But if we apply the same argument as in the previous subsection, then to maintain limiting fixed dimension we get the condition

$$R_{\mu\nu}\; - \frac{1}{2}\,R\,g_{\mu\nu} = \sigma'\,T_{\mu\nu}$$

which has exactly the form of Einstein’s equations in the presence of matter [114][115][75][116].

Just as we interpreted the curvature part of these equations in the previous subsection in terms of the change in area of geodesic bundles, we can interpret the “matter” part in terms of the change of geodesics associated with additional local connections. As an example, consider starting with a 2D hexagonal grid. Now imagine adding edges at each node. Doing this creates additional connections and additional geodesics, eventually producing something like the hyperbolic space examples in 4.2. So what the equation says is that any such effect, which would lead to negative curvature, must be compensated by positive curvature in the “background” spacetime—just as general relativity suggests.

data 16.04.2020

where here R is the Ricci scalar*
curvature of the space.

Pag. 371 $E = mc^2 + \frac{1}{2} m v^2$

E) Some Relativistic and Gravitational Properties of the Wolfram
April 12, 2020

Source:
www.wolframcloud.com/obj/wolframphysics/Documents/some-relativistic-and-gravitational-properties-of-the-wolfram-model.pdf

Pag. 31 Figure 16:

Two-dimensional spatial hypergraphs with different Olliver-Ricci scalar curvatures, with the presence of pentagons indicating regions of positive scalar curvature and the presence of heptagons indicating regions of negative scalar curvature.

Figure 19 (Pag. 40)

Geodesic trajectories of test particles (starting on initially parallel paths) through discrete spaces containing a spherical mass distribution. Adapted from S. Wolfram, A New Kind of Science, pg 531.

data 16.04.2020

F) Some Quantum Mechanical Properties of the Wolfram Model

Source:
www.wolframcloud.com/obj/wolframphysics/Documents/some-quantum-mechanical-properties-of-the-wolfram-model.pdf

Pág. 39 Figure 11

The Fano plane represented using the Steiner system $S(2,3,7)$ (historically known as a steiner triple system, since $K = 3 = t + 1$); the 7 blocks correspond to the 7 lines of the plane, each of which contains exactly 3 points, and with the property that every pair of points lies on exactly one line.

Pág. 48 Gullstrand–Painlevé coordinates

(t_r, r, θ, ϕ)

Jandaia

data 16.04.2020

$$g = \left(1 - \frac{2GM}{c^2 r}\right) dt_r^2 - 2\sqrt{\frac{2GM}{c^2 r}}\, dt_r\, dr - dr^2 - r^2 d\theta^2 - r^2 \sin^2(\theta)\, d\phi^2, \quad (149)$$

G) Wolfram Physics Project Launch
www.youtube.com/watch?
V=rbfFt2uNEyQ

H) Technical Documents
www.wolframphysics.org/
technical-documents/

I) How we got here: the Backstory of the Wolfram Physics Project
Source: https://writings.
stephenwolfram.com/
2020/04/
how-we-got-here-the-
backstory-of-the-wolfram-
physics-project/

ref. https://www.wolframscience.com/
nks/p479--space-as-a-
network/

→

data 16.04.2020

πr^2

$4/3 \pi r^3$,

the volume of a four-dimensional hypersphere is

$1/2 \pi^2 r^4$.

J) Uncommon Knowledge

by Al McDowell

Source:

www.spheritons.com

Obs: The Cause of Gravity

www.spheritons.com/cause-of-gravity.html

K) Relativity - the Special and General theory

Source:

www.ibiblio.org/ebooks/Einstein/Einstein_Relativity.pdf

Obs.: Pág. 184 (do todo)

L) Finally We May Have a Path to the Fundamentals of Physics... and It's Beautiful.

data 16.04.2020

https://writings.stephenwolfram.com/
2020/04/
finally-we-may-have-a-path-
to-the-fundamental-theory-of-
physics-and-its-beautiful/

139) 20200416 —

A) Books_68_2020/
Lagrangian Interaction An
Introduction TO Relativistic
Symmetry in Electrodynamics
And Gravitation/
Lagrangian Interaction...pdf

obs: Usar este livro.

B) Books_68_2020/
Geometric Analysis/

Jost Jurgen Jost - Riemannian
Geometry and Geometric Analysis.
pdf

Pág. 23(13)

1.4 Riemannian Metrics

Curvature: 8.8 Pág. 491
8.9 Pág. 498
8.10 Pág. 519 _

C) Books_68_2020/
Geometric Analysis/
Geometric function theory

% #!
/012/ / /3 / #! = % < 0
% " > 1 5 + + % 6 "+ 0 " ? 5 !! @ 4 5
*" = 5 > " AAA % * AAA

- /012/ / /3 / #! B @ B B 1
? + " ? 0 + @ * *"
' # #

(1

1.4 Riemannian Metrics

We now want to introduce metric structures on differentiable manifolds. Again, we shall start from infinitesimal considerations. We would like to be able to measure the lengths of and the angles between tangent vectors. Then, one may, for example, obtain the length of a differentiable curve by integration. In a vector space such a notion of measurement is usually given by a scalar product. We thus define

Definition 1.4.1 A *Riemannian metric* on a differentiable manifold M is given by a scalar product on each tangent space T_pM which depends smoothly on the base point p. A *Riemannian manifold* is a differentiable manifold, equipped with a Riemannian metric.

In order to understand the concept of a Riemannian metric, we again need to study local coordinate representations and the transformation behavior of these expressions.

Thus, let $x = (x^1, \dots, x^d)$ be local coordinates. In these coordinates, a metric is represented by a positive definite, symmetric matrix

$$(g_{ij}(x))_{i,j=1,\dots,d}$$

Proof. Let $v \in T_qX$, (φ, U) a chart on X with $q \in U$. Let γ be any smooth curve in $\varphi(U)$ with $\gamma(0) = \varphi(q)$, $\dot{\gamma}(0) := \frac{d}{dt}\gamma(t)_{|t=0} = d\varphi(v)$, for example $\gamma(t) = \varphi(q) + td\varphi(v)$. $c := \varphi^{-1}(\gamma)$ then is a curve in X with $\dot{c}(0) = v$. Because of $X = f^{-1}(p)$,

$$f \circ c(t) = p \ \forall \ t,$$

hence $df(q) \circ \dot{c}(0) = 0$, and consequently $v = \dot{c}(0) \in \ker df(q)$. Since also $T_qX = \dim \ker df(q) = m - n$, the claim follows. □

For our example S^n, we may choose

$$f : \mathbb{R}^{n+1} \to \mathbb{R}, f(x) = |x|^2.$$

Then

$$T_xS^n = \ker df(x) = \{v \in \mathbb{R}^{n+1} : x \cdot v (= x^i v^i) = 0\}.$$

Perspectives. H. Whitney (1936) showed that any d-dimensional differentiable manifold can be embedded into $\mathbb{R}^{2d+1}$. Thus, the class of abstract differentiable manifolds is the same as the class of submanifolds of Euclidean space. Nevertheless, the abstract and intrinsic point of view offers great conceptual and technical advantages over the approach of submanifold geometry of Euclidean spaces.

1.4 Riemannian Metrics

We now want to introduce metric structures on differentiable manifolds. Again, we shall start from infinitesimal considerations. We would like to be able to measure the lengths of and the angles between tangent vectors. Then, one may, for example, obtain the length of a differentiable curve by integration. In a vector space such a notion of measurement is usually given by a scalar product. We thus define

Definition 1.4.1 A *Riemannian metric* on a differentiable manifold M is given by a scalar product on each tangent space T_pM which depends smoothly on the base point p. A *Riemannian manifold* is a differentiable manifold, equipped with a Riemannian metric.

In order to understand the concept of a Riemannian metric, we again need to study local coordinate representations and the transformation behavior of these expressions.

Thus, let $x = (x^1, \ldots, x^d)$ be local coordinates. In these coordinates, a metric is represented by a positive definite, symmetric matrix

$$(g_{ij}(x))_{i,j=1,\ldots,d}$$

(i.e. $g_{ij} = g_{ji}$ for all i, j, $g_{ij}\xi^i\xi^j > 0$ for all $\xi = (\xi^1, \ldots, \xi^d) \neq 0$), where the coefficients depend smoothly on x. The transformation formula (1.4.3) below will imply that this smoothness does not depend on the choice of coordinates. Therefore, smooth dependence on the base point as required in def. 1.4.1 can be expressed in local coordinates.

The product of two tangent vectors $v, w \in T_pM$ with coordinate representations $(v^1, \ldots, v^d)$ and $(w^1, \ldots, w^d)$ (i.e. $v = v^i \frac{\partial}{\partial x^i}, w = w^j \frac{\partial}{\partial x^j}$) then is

$$\langle v, w \rangle := g_{ij}(x(p))v^i w^j. \tag{1.4.1}$$

In particular, $\langle \frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j} \rangle = g_{ij}$.

Similarly, the length of v is given by

$$\|v\| := \langle v, v \rangle^{\frac{1}{2}}.$$

We now want to study the transformation behavior. Let $y = f(x)$ define different local coordinates. In these coordinates, v and w have representations $(\tilde{v}^1, \ldots, \tilde{v}^d)$ and $(\tilde{w}, \ldots, \tilde{w}^d)$ with $\tilde{v}^j = v^i \frac{\partial f^j}{\partial x^i}, \tilde{w}^j = w^i \frac{\partial f^j}{\partial x^i}$. Let the metric in the new coordinates be given by $h_{k\ell}(y)$.

It follows that

$$h_{k\ell}(f(x))\tilde{v}^k \tilde{w}^\ell = \langle v, w \rangle = g_{ij}(x)v^i w^j, \tag{1.4.2}$$

hence

$$h_{k\ell}(f(x)) \frac{\partial f^k}{\partial x^i} \frac{\partial f^\ell}{\partial x^j} v^i w^j = g_{ij}(x) v^i w^j,$$

and since this holds for all tangent vectors v, w,

$$h_{k\ell}(f(x)) \frac{\partial f^k}{\partial x^i} \frac{\partial f^\ell}{\partial x^j} = g_{ij}(x). \tag{1.4.3}$$

Formula (1.4.3) gives the transformation behavior of a metric under coordinate changes.

The simplest example of a Riemannian metric of course is the Euclidean one. For $v = (v^1, \ldots, v^d), w = (w^1, \ldots, w^d) \in T_x\mathbb{R}^d$, the Euclidean scalar product is simply

$$\delta_{ij} v^i w^j = v^i w^i,$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases}$$

is the standard Kronecker symbol.

Theorem 1.4.1 *Each differentiable manifold may be equipped with a Riemannian metric.*

Proof. Let $\{(x_\alpha, U_\alpha) : \alpha \in A\}$ be an atlas, $(\varphi_\alpha)_{\alpha \in A}$ a partition of unity subordinate to $(U_\alpha)_{\alpha \in A}$ (see Lemma 1.1.1) (for simplicity of notation, we use the same index set for (φ_α) and (U_α); this may be justified by replacing the original covering (U_α) by a locally finite refinement).

For $v, w \in T_pM$ and $\alpha \in A$ with $p \in U_\alpha$ let the coordinate representations be $(v_\alpha^1, \ldots, v_\alpha^d)$ and $(w_\alpha^1, \ldots, w_\alpha^d)$. Then we put

$$\langle v, w \rangle := \sum_{\substack{\alpha \in A \\ \text{with } p \in U_\alpha}} \varphi_\alpha(p) v_\alpha^i w_\alpha^i \, .$$

This defines a Riemannian metric. (The metric is simply obtained by piecing the Euclidean metrics of the coordinate images together with the help of a partition of unity.) □

Let now $[a, b]$ be a closed interval in $\mathbb{R}, \gamma : [a, b] \to M$ a smooth curve, where "smooth", as always, means "of class C^∞".
The *length of* γ then is defined as

$$L(\gamma) := \int_a^b \| \frac{d\gamma}{dt}(t) \| dt$$

and the *energy* of γ as

$$E(\gamma) := \frac{1}{2} \int_a^b \| \frac{d\gamma}{dt}(t) \|^2 dt \, .$$

(In physics, $E(\gamma)$ is usually called "action of γ" where γ is considered as the orbit of a mass point.) Of course, these expressions can be computed in local coordinates. Working with the coordinates $(x^1(\gamma(t)), \ldots, x^d(\gamma(t)))$ we use the abbreviation

$$\dot{x}^i(t) := \frac{d}{dt}(x^i(\gamma(t))).$$

Then

$$L(\gamma) = \int_a^b \sqrt{g_{ij}(x(\gamma(t)))\dot{x}^i(t)\dot{x}^j(t)} dt$$

and

$$E(\gamma) = \frac{1}{2} \int_a^b g_{ij}(x(\gamma(t)))\dot{x}^i(t)\dot{x}^j(t) dt.$$

We also remark for later technical purposes that the length of a (continuous and) piecewise smooth curve may be defined as the sum of the lengths of the smooth pieces, and the same holds for the energy.

On a Riemannian manifold M, the *distance* between two points p, q can be defined:

$$d(p,q) := \inf\{L(\gamma) : \gamma : [a,b] \to M \text{ piecewise smooth curve with } \gamma(a) = p, \gamma(b) = q\}.$$

We first remark, that any two points $p, q \in M$ can be connected by a piecewise smooth curve, and $d(p,q)$ therefore is always defined. Namely, let

$$E_p := \{q \in M : p \text{ and } q \text{ can be connected by a piecewise smooth curve.}\}$$

With the help of local coordinates one sees that E_p is open. But then also $M \backslash E_p = \bigcup\limits_{q \notin E_p} E_q$ is open. Since M is connected and $E_p \neq \emptyset \, (p \in E_p)$, we conclude $M = E_p$.

The distance function satisfies the usual axioms:

Lemma 1.4.1

(i) $d(p,q) \geq 0$ *for all* p, q, *and* $d(p,q) > 0$ *for all* $p \neq q$

(ii) $d(p,q) = d(q,p)$

(iii) $d(p,q) \leq d(p,r) + d(r,q)$ *(triangle inequality) for all points* $p, q, r \in M$.

Proof. (ii) and (iii) are obvious. For (i), we only have to show $d(p,q) > 0$ for $p \neq q$. For this purpose, let $x : U \to \mathbb{R}^d$ be a chart with $p \in U$. Then there exists $\varepsilon > 0$ with

$$D_\varepsilon(x(p)) := \{y \in \mathbb{R}^d : |y - x(p)| \leq \varepsilon\} \subset x(U)$$

(the bars denote the Euclidean absolute value) and

$$q \notin x^{-1}(D_\varepsilon(x(p))). \tag{1.4.4}$$

Let the metric be represented by $(g_{ij}(x))$ in our chart. Since $(g_{ij}(x))$ is positive definite and smooth, hence continuous in x and $D_\varepsilon(x(p))$ is compact, there exists $\lambda > 0$ with

$$g_{ij}(y)\xi^i\xi^j \geq \lambda|\xi|^2 \tag{1.4.5}$$

for all $y \in D_\varepsilon(x(p)), \xi = (\xi^1, \ldots, \xi^d) \in \mathbb{R}^d$. Therefore, for any curve $\gamma : [a,b] \to M$ with $\gamma(a) = p, \gamma(b) = q$

$$\begin{aligned} L(\gamma) &\geq L(\gamma \cap x^{-1}(D_\varepsilon(x(p))) \\ &\geq \lambda\varepsilon > 0, \end{aligned} \tag{1.4.6}$$

because $x(\gamma)$ by (1.4.4) has to contain a point $z \in \partial D_\varepsilon(x(p))$, i.e. a point whose Euclidean distance from $x(p)$ is ε. By (1.4.5), z then has distance from $x(p)$ at least $\lambda\varepsilon$ w.r.t. the metric (g_{ij}). □

Corollary 1.4.1 *The topology on M induced by the distance function d coincides with the original manifold topology of M.*

Proof. It suffices to show that in each chart the topology induced by d coincides with the one of $\mathbb{R}^d$, i.e. the one induced by the Euclidean distance function. Now for every x in some chart, there exists $\varepsilon > 0$ for which $D_\varepsilon(x)$ is contained in the same chart, and positive constants λ, μ with

$$\lambda|\xi|^2 \leq g_{ij}(y)\xi^i\xi^j \leq \mu|\xi|^2 \text{ for all}\, y \in D_\varepsilon(x), \xi \in \mathbb{R}^d.$$

Thus

$$\lambda|y-x| \leq d(y,x) \leq \mu|y-x| \text{ for all}\, y \in D_\varepsilon(x),$$

and thus each Euclidean distance ball contains a distance ball for d, and vice versa (with

$$B(z,\delta) := \{y \in M : d(z,y) \leq \delta\}$$

we have

$$\overset{\circ}{D}_{\lambda\delta}(x) \subset \overset{\circ}{B}(x,\delta) \subset \overset{\circ}{D}_{\mu\delta}(x), \text{ if } \mu\delta \leq \varepsilon).$$ □

We now return to the length and energy functionals.

Lemma 1.4.2 *For each smooth curve $\gamma : [a,b] \to M$*

$$L(\gamma)^2 \leq 2(b-a)E(\gamma), \tag{1.4.7}$$

and equality holds if and only if $\|\frac{d\gamma}{dt}\| \equiv$ const.

Proof. By Hölder's inequality

$$\int_a^b \|\frac{d\gamma}{dt}\|dt \leq (b-a)^{\frac{1}{2}} \left(\int_a^b \|\frac{d\gamma}{dt}\|^2 dt\right)^{\frac{1}{2}}$$

with equality precisely if $\|\frac{d\gamma}{dt}\| \equiv$ const. □

Lemma 1.4.3 *If $\gamma : [a,b] \to M$ is a smooth curve, and $\psi : [\alpha,\beta] \to [a,b]$ is a change of parameter, then*

$$L(\gamma \circ \psi) = L(\gamma).$$

Proof. Let $t = \psi(\tau)$.

$$\begin{aligned} L(\gamma\circ\psi) &= \int_\alpha^\beta \left(g_{ij}(x(\gamma(\psi(\tau))))\dot{x}^i(\psi(\tau))\dot{x}^j(\psi(\tau)) \left(\frac{d\psi}{d\tau}\right)^2\right)^{\frac{1}{2}} d\tau \\ &\quad \text{(by the chain rule)} \\ &= L(\gamma) \quad \text{(by change of variables).} \end{aligned}$$ □

Lemma 1.4.4 *The Euler-Lagrange equations for the energy E are*

$$\ddot{x}^i(t) + \Gamma^i_{jk}(x(t))\dot{x}^j(t)\dot{x}^k(t) = 0, \quad i = 1, \dots, d \tag{1.4.8}$$

with

$$\Gamma^i_{jk} = \frac{1}{2} g^{i\ell}(g_{j\ell,k} + g_{k\ell,j} - g_{jk,\ell}),$$

where

$$(g^{ij})_{i,j=1,\dots,d} = (g_{ij})^{-1} \quad (i.e.\ g^{i\ell}g_{\ell j} = \delta_{ij})$$

and

$$g_{j\ell,k} = \frac{\partial}{\partial x^k} g_{j\ell}.$$

The expressions Γ^i_{jk} are called Christoffel symbols.

Proof. The Euler-Lagrange equations of a functional

$$I(x) = \int_a^b f(t, x(t), \dot{x}(t))dt$$

are given by

$$\frac{d}{dt}\frac{\partial f}{\partial \dot{x}^i} - \frac{\partial f}{\partial x^i} = 0, \quad i = 1, \dots, d.$$

In our case, we therefore get $(E(\gamma) = \frac{1}{2}\int g_{jk}(x(t))\dot{x}^j\dot{x}^k dt)$

$$\frac{d}{dt}(g_{ik}(x(t))\dot{x}^k(t) + g_{ji}(x(t))\dot{x}^j(t)) - g_{jk,i}(x(t))\dot{x}^j(t)\dot{x}^k(t) = 0 \text{ for } i = 1, \dots, d,$$

hence

$$g_{ik}\ddot{x}^k + g_{ji}\ddot{x}^j + g_{ik,\ell}\dot{x}^\ell\dot{x}^k + g_{ji,\ell}\dot{x}^\ell\dot{x}^j - g_{jk,i}\dot{x}^j\dot{x}^k = 0.$$

Renaming some indices and using the symmetry $g_{ik} = g_{ki}$, we get

$$2g_{\ell m}\ddot{x}^m + (g_{\ell k,j} + g_{j\ell,k} - g_{jk,\ell})\dot{x}^j\dot{x}^k = 0 \quad \ell = 1, \dots, d \tag{1.4.9}$$

and from this

$$g^{i\ell}g_{\ell m}\ddot{x}^m + \frac{1}{2}g^{i\ell}(g_{\ell k,j} + g_{j\ell,k} - g_{jk,\ell})\dot{x}^j\dot{x}^k = 0 \quad i = 1, \dots, d.$$

Because of

$$g^{i\ell}g_{\ell m} = \delta_{im}, \text{ and thus } g^{i\ell}g_{\ell m}\ddot{x}^m = \ddot{x}^i$$

we obtain (1.4.8) from this. □

Definition 1.4.2 A smooth curve $\gamma = [a, b] \to M$, which satisfies (with $\dot{x}^i(t) = \frac{d}{dt}x^i(\gamma(t))$ etc.)

$$\ddot{x}^i(t) + \Gamma^i_{jk}(x(t))\dot{x}^j(t)\dot{x}^k(t) = 0 \text{ for } i = 1, \dots, d$$

is called a *geodesic*.

Thus, geodesics are the critical points of the energy functional. By Lemma 1.4.3, the length functional is invariant under parameter changes. As in the Euclidean case, one easily sees that regular curves can be parametrized by arc length. We shall attempt to minimize the length within the class of regular smooth curves, and we shall succeed and complete the program in Corollary 1.4.2 below. As the length is invariant under reparametrization by Lemma 1.4.3, therefore, if one seeks curves of shortest length, it suffices to consider curves that are parametrized by arc length. For such curves, by Lemma 1.4.2 one may minimize energy instead of length. Conversely, every critical point of the energy functional, i.e. each solution of (1.4.8), i.e. each geodesic, is parametrized proportionally to arc length.

Namely, for a solution of (1.4.8)

$$\begin{aligned}
&\frac{d}{dt}\langle \dot{x}, \dot{x}\rangle = \frac{d}{dt}(g_{ij}(x(t))\dot{x}^i(t)\dot{x}^j(t))\\
&= g_{ij}\ddot{x}^i\dot{x}^j + g_{ij}\dot{x}^i\ddot{x}^j + g_{ij,k}\dot{x}^i\dot{x}^j\dot{x}^k\\
&= -(g_{jk,\ell} + g_{\ell j,k} - g_{\ell k,j})\dot{x}^\ell\dot{x}^k\dot{x}^j + g_{\ell j,k}\dot{x}^k\dot{x}^\ell\dot{x}^j\\
&\qquad\qquad \text{by formula (1.4.9) which is equivalent to (1.4.8)}\\
&= 0, \text{ since } g_{jk,\ell}\dot{x}^\ell\dot{x}^k\dot{x}^j = g_{\ell k,j}\dot{x}^\ell\dot{x}^k\dot{x}^j\\
&\qquad\qquad \text{by interchanging the indices } j \text{ and } \ell
\end{aligned}$$

Consequently $\langle \dot{x}, \dot{x}\rangle \equiv$ const., and hence the curve is parametrized proportionally to arc length. We have shown

Lemma 1.4.5 *Each geodesic is parametrized proportionally to arc length.*

□

Theorem 1.4.2 *Let M be a Riemannian manifold, $p \in M, v \in T_pM$. Then there exist $\varepsilon > 0$ and precisely one geodesic*

$$c : [0, \varepsilon] \to M$$

with $c(0) = p, \dot{c}(0) = v$. In addition, c depends smoothly on p and v.

Proof. (1.4.8) is a system of second order ODE, and the Picard-Lindelöf Theorem yields the local existence and uniqueness of a solution with prescribed initial values and derivatives, and this solution depends smoothly on the data.

□

We note that if $x(t)$ is a solution of (1.4.8), so is $x(\lambda t)$ for any constant $\lambda \in \mathbb{R}$. Denoting the geodesic of Theorem 1.4.2 with $c(0) = p, \dot{c}(0) = v$ by c_v, we obtain

$$c_v(t) = c_{\lambda v}(\frac{t}{\lambda}) \text{ for } \lambda > 0, t \in [0, \varepsilon].$$

In particular, $c_{\lambda v}$ is defined on $[0, \frac{\varepsilon}{\lambda}]$.

Since c_v depends smoothly on v, and $\{v \in T_pM : \|v\| = 1\}$ is compact, there exists $\varepsilon_0 > 0$ with the property that for $\|v\| = 1$ c_v is defined at least on $[0, \varepsilon_0]$. Therefore, for any $w \in T_pM$ with $\|w\| \leq \varepsilon_0$, c_w is defined at least on $[0, 1]$.

Definition 1.4.3 Let M be a Riemannian manifold, $p \in M$,

$$\begin{aligned} V_p :=& \{v \in T_pM : c_v \text{ is defined on } [0,1]\} \\ \exp_p :& V_p \to M \\ & v \mapsto c_v(1) \end{aligned}$$

is called the *exponential map* of M at p.

By the preceding considerations, the domain of definition of the exponential map always at least contains a small neighborhood of $0 \in T_pM$. In general, however, V_p is not all of T_pM, as is already seen in the example of a proper, open subset of $\mathbb{R}^d$, equipped with the Euclidean metric. Nevertheless, we shall see in Theorem 1.4.7 below that for a compact Riemannian manifold, $\exp_p$ can be defined on all of T_pM.

Theorem 1.4.3 *The exponential map* $\exp_p$ *maps a neighborhood of* $0 \in T_pM$ *diffeomorphically onto a neighborhood of* $p \in M$.

Proof. Since T_pM is a vector space, we may identify T_0T_pM, the tangent space of T_pM at $0 \in T_pM$, with T_pM itself. The derivative of $\exp_p$ at 0 then becomes a map from T_pM onto itself:

$$d\exp_p(0) : T_pM \to T_pM.$$

With this identification of T_0T_pM and T_pM, for $v \in T_pM$

$$\begin{aligned} d\exp_p(0)(v) &= \frac{d}{dt} c_{tv}(1)_{|t=0} \\ &= \frac{d}{dt} c_v(t)_{|t=0} \\ &= \dot{c}_v(0) \\ &= v \end{aligned}$$

Hence

$$d\exp_p(0) = id_{|T_pM}. \tag{1.4.10}$$

In particular, $d\exp_p(0)$ has maximal rank, and by the inverse function theorem, there exists a neighborhood of $0 \in T_pM$ which is mapped diffeomorphically onto a neighborhood of $p \in M$. □

Let now $e_1, e_2, \ldots, e_d$ $(d = \dim M)$ be a basis of T_pM which is orthonormal w.r.t. the scalar product on T_pM defined by the Riemannian metric. Writing for each vector $v \in T_pM$ its components w.r.t. this basis, we obtain a map

$$\begin{aligned} \Phi &: T_pM \to \mathbb{R}^d \\ v &= v^i e_i \mapsto (v^1, \ldots, v^d). \end{aligned}$$

For the subsequent construction, we identify T_pM with $\mathbb{R}^d$ via Φ. By Theorem 1.4.3, there exists a neighborhood U of p which is mapped by $\exp_p^{-1}$ diffeomorphically onto a neighborhood of $0 \in T_pM$, hence, with our identification $T_pM \cong \mathbb{R}^d$, diffeomorphically onto a neighborhood Ω of $0 \in \mathbb{R}^d$. In particular, p is mapped to 0.

Definition 1.4.4 The local coordinates defined by the chart $(\exp_p^{-1}, U)$ are called (Riemannian) *normal coordinates* with center p.

Theorem 1.4.4 *In normal coordinates, we have for the Riemannian metric*

$$g_{ij}(0) = \delta_{ij} \tag{1.4.11}$$

$$\Gamma^i_{jk}(0) = 0 \text{ (and also } g_{ij,k}(0) = 0) \textit{ for all } i, j, k\,. \tag{1.4.12}$$

Proof. (1.4.11) directly follows from the fact that the above identification $\Phi : T_pM \cong \mathbb{R}^d$ maps an orthonormal basis of T_pM w.r.t. the Riemannian metric onto an Euclidean orthonormal basis of $\mathbb{R}^d$.

For (1.4.12), we note that in normal coordinates, the straight lines through the origin of $\mathbb{R}^d$ (or, more precisely, their portions contained in the chart image) are geodesic. Namely, the line $tv, t \in \mathbb{R}, v \in \mathbb{R}^d$, is mapped (for sufficiently small t) onto $c_{tv}(1) = c_v(t)$, where $c_v(t)$ is the geodesic, parametrized by arc length, with $\dot{c}_v(0) = v$.

Inserting now $x(t) = tv$ into the geodesic equation (1.4.8), we obtain because of $\ddot{x}(t) = 0$

$$\Gamma^i_{jk}(tv)v^j v^k = 0, \text{ for } i = 1, \ldots, d\,. \tag{1.4.13}$$

In particular at 0, i.e. for $t = 0$,

$$\Gamma^i_{jk}(0)v^j v^k = 0 \text{ for all } v \in \mathbb{R}^d, i = 1, \ldots, d\,. \tag{1.4.14}$$

We put $v = \frac{1}{2}(e_\ell + e_m)$ and obtain because of the symmetry $\Gamma^i_{jk} = \Gamma^i_{kj}$

$$\Gamma^i_{\ell m}(0) = 0 \text{ for all } i.$$

Since this holds for all ℓ, m, all $\Gamma^i_{jk}(0)$ vanish. By definition of Γ^i_{jk}, we obtain at $0 \in \mathbb{R}^d$

$$g^{i\ell}(g_{j\ell,k} + g_{k\ell,j} - g_{jk,\ell}) = 0 \;\forall\, i, j, k,$$

hence also

$$g_{jm,k} + g_{km,j} - g_{jk,m} = 0 \; \forall \, j, k, m.$$

Adding now the relation (obtained by cyclic permutation of the indices)

$$g_{kj,m} + g_{mj,k} - g_{km,j} = 0,$$

we obtain (with $g_{kj} = g_{jk}$)

$$g_{jm,k}(0) = 0, \text{ for all } j, k, m. \qquad \square$$

Later on (in Chapter 3), we shall see that in general the *second* derivatives of the metric cannot be made to vanish at a given point by a suitable choice of local coordinates. The obstruction will be given by the curvature tensor.

Further properties of Riemannian normal coordinates may best be seen by using *polar coordinates*, instead of the Euclidean ones (obtained from the map Φ). We therefore introduce on $\mathbb{R}^d$ the standard polar coordinates

$$(r, \varphi^1, \dots, \varphi^{d-1}),$$

where $\varphi = (\varphi^1, \dots, \varphi^{d-1})$ parametrizes the unit sphere S^{d-1} (the precise formula for φ will be irrelevant for our purposes), and we then obtain polar coordinates on T_pM via Φ again. We express the metric in polar coordinates and write g_{rr} instead of g_{11}, because of the special role of r. We also write $g_{r\varphi}$ instead of $g_{1\ell}, \ell \in \{2, \dots, d\}$, and $g_{\varphi\varphi}$ as abbreviation for $(g_{k\ell})_{k,\ell=2,\dots,d}$. In particular, in these coordinates at $0 \in T_pM$ (this point corresponds to $p \in M$)

$$g_{rr}(0) = 1, g_{r\varphi}(0) = 0 \tag{1.4.15}$$

by (1.4.11) and since this holds for Euclidean polar coordinates.

After these preparations, we return to the analysis of the geodesic equation (1.4.8). The lines $\varphi \equiv$ const. are geodesic when parametrized by arc length. They are given by $x(t) = (t, \varphi_0)$, φ_0 fixed, and from (1.4.8)

$$\Gamma^i_{rr} = 0 \text{ for all } i$$

(we have written Γ^i_{rr} instead of Γ^i_{11}), hence

$$g^{i\ell}(2g_{r\ell,r} - g_{rr,\ell}) = 0, \text{ for all } i,$$

thus

$$2g_{r\ell,r} - g_{rr,\ell} = 0, \text{ for all } \ell. \tag{1.4.16}$$

For $\ell = r$, we conclude

$$g_{rr,r} = 0,$$

and with (1.4.15) then

$$g_{rr} \equiv 1. \tag{1.4.17}$$

Inserting this in (1.4.16), we get

$$g_{r\varphi,r} = 0,$$

and then again with (1.4.15)

$$g_{r\varphi} \equiv 0. \tag{1.4.18}$$

We have shown

Theorem 1.4.5 *For the polar coordinates, obtained by transforming the Euclidean coordinates of* $\mathbb{R}^d$*, on which the normal coordinates with centre* p *are based, into polar coordinates, we have*

$$g_{ij} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & g_{\varphi\varphi}(r,\varphi) & \\ 0 & & & \end{pmatrix},$$

where $g_{\varphi\varphi}(r,\varphi)$ *is the* $(d-1)\times(d-1)$ *matrix of the components of the metric w.r.t. angular variables* $(\varphi^1,\dots,\varphi^{d-1}) \in S^{d-1}$*.* □

The polar coordinates of Theorem 1.4.5 are often called *Riemannian polar coordinates.* The situation is the same as for Euclidean polar coordinates: For example in polar coordinates on $\mathbb{R}^2$, the Euclidean metric is given by $\begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix}$. We point out once more that in contrast to Theorem 1.4.4, Theorem 1.4.5 holds not only at the origin $0 \in T_pM$, but in the whole chart.

Corollary 1.4.2 *For any* $p \in M$*, there exists* $\rho > 0$ *such that Riemannian polar coordinates may be introduced on* $B(p,\rho) := \{q \in M : d(p,q) \le \rho\}$*. For any such* ρ *and any* $q \in \partial B(p,\rho)$*, there is precisely one geodesic of shortest length (*$= \rho$*) from* p *to* q*, and in polar coordinates, this geodesic is given by the straight line* $x(t) = (t,\varphi_0)$*,* $0 \le t \le \rho$*, where* q *is represented by the coordinates* $(\rho,\varphi_0), \varphi_0 \in S^{d-1}$*. Here, "of shortest length" means that the curve is the shortest one among all curves in* M *from* p *to* q*.*

Proof. The first claim follows from Corollary 1.4.1 (and its proof) and Theorem 1.4.3. For the second claim, let $c(t) = (r(t),\varphi(t)), 0 \le t \le T$, be an arbitrary curve from p to q. $c(t)$ need not be entirely contained in $B(p,\rho)$ and may leave our coordinate neighborhood. Let

$$t_0 := \inf\{t \le T : d(x(t),p) \ge \rho\}.$$

Then $t_0 \le T$, and the curve $c_{|[0,t_0]}$ is entirely contained in $B(p,\rho)$. We shall show $L(c_{|[0,t_0]}) \ge \rho$ with equality only for a straight line in our polar coordinates. This will then imply the second claim. The proof of this inequality goes as follows:

$$\begin{aligned}
L(c_{|[0,t_0]}) &= \int_0^{t_0} (g_{ij}(c(t))\dot{c}^i\dot{c}^j)^{\frac{1}{2}}\,dt \\
&\geq \int_0^{t_0} (g_{rr}(c(t))\dot{r}\dot{r})^{\frac{1}{2}}\,dt && \text{by (1.4.18) and} \\
& && \text{since } g_{\varphi\varphi} \text{ is positive definite} \\
&= \int_0^{t_0} |\dot{r}|dt \ \geq \ \int_0^{t_0} \dot{r}dt && \text{by (1.4.17)} \\
&= \ r(t_0) \ = \ \rho && \text{by definition of } t_0,
\end{aligned}$$

and equality holds precisely if $g_{\varphi\varphi}\dot{\varphi}\dot{\varphi} \equiv 0$, in which case $\dot{\varphi}(t)$ is constant and $\dot{r} \geq 0$ and $c(t)$ thus is a straight line through the origin. □

In particular, under the assumptions of Corollary 1.4.2, the Euclidean ball

$$d_\rho(0) := \{y \in \mathbb{R}^d : |y| \leq \rho\} \subset T_pM$$

is mapped under $\exp_p$ diffeomorphically onto the Riemannian ball with the same radius,

$$B(p, \rho).$$

Corollary 1.4.3 *Let M be a compact Riemannian manifold. Then there exists $\rho_0 > 0$ with the property that for any $p \in M$, Riemannian polar coordinates may be introduced on $B(p, \rho_0)$.*

Proof. By Corollary 1.4.2, for any $p \in M$, there exists $\rho > 0$ with those properties. By Theorem 1.4.2, $\exp_p$ is smooth in p. If thus $\exp_p$ is injective and of maximal rank on a closed ball with radius ρ in T_pM, there exists a neighborhood U of p such that for all $q \in U, \exp_q$ is injective and of maximal rank on the closed ball with radius ρ in T_qM.

Since M is compact, it can be covered by finitely many such neighborhoods and we choose ρ_0 as the smallest such ρ. □

Corollary 1.4.4 *Let M be a compact Riemannian manifold. Then there exists $\rho_0 > 0$ with the property that any two points $p, q \in M$ with $d(p, q) \leq \rho_0$ can be connected by precisely one geodesic of shortest length. This geodesic depends continuously on p and q.*

Proof. ρ_0 from Corollary 1.4.3 satisfies the first claim by Corollary 1.4.2. Moreover, by the last claim of Corollary 1.4.2, the shortest geodesic from p to $q \in B(p, \rho_0)$ depends continuously on p. Exchanging the roles of p and q yields the continuous dependence on p as well. □

We explicitly point out that for any compact Riemannian manifold there is always more than one geodesic connection between any two points (This will be discussed in Chapter 5.). Only the shortest geodesic is unique, provided p and q are sufficiently close.

We now want to show the existence of shortest (geodesic) connections between any two points of arbitrary distance on a given compact Riemannian manifold.

We recall the notion of homotopy between curves:

Definition 1.4.5 Two curves γ_0, γ_1 on a manifold M with common initial and end points p and q, i.e. two continuous maps

$$\gamma_0, \gamma_1 : I = [0,1] \to M$$

with $\gamma_0(0) = \gamma_1(0) = p$, $\gamma_0(1) = \gamma_1(1) = q$, are called *homotopic* if there exists a continuous map

$$\Gamma : I \times I \to M$$

with

$$\begin{array}{lll} \Gamma(0,s) = p, & \Gamma(1,s) = q & \text{for all } s \in I \\ \Gamma(t,0) = \gamma_0(t), & \Gamma(t,1) = \gamma_1(t) & \text{for all } t \in I \end{array}$$

Two closed curves c_0, c_1 in M, i.e. two continuous maps

$$c_0, c_1 : S^1 \to M,$$

are called *homotopic*, if there exists a continuous map

$$c : S^1 \times I \to M$$

with

$$c(t,0) = c_0(t), c(t,1) = c_1(t) \quad \text{for all } t \in S^1$$

(S^1, as usual, is the unit circle parametrized by $[0, 2\pi)$.).

Lemma 1.4.6 *The concept of homotopy defines an equivalence relation on the set of all curves in M with fixed initial and end points as well as on the set of all closed curves in M.*

The *proof* is elementary. □

With the help of this concept, we now want to show the existence of geodesics:

Theorem 1.4.6 *Let M be a compact Riemannian manifold, $p, q \in M$. Then there exists a geodesic in every homotopy class of curves from p to q, and this geodesic may be chosen as a shortest curve in its homotopy class. Likewise, every homotopy class of closed curves in M contains a curve which is shortest and geodesic.*

Proof. Since the proof is the same in both cases, we shall only consider the case of closed curves.

As a preparation, we shall first show

Lemma 1.4.7 *Let M be a compact Riemannian manifold, $\rho_0 > 0$ as in Corollary 1.4.4. Let $\gamma_0, \gamma_1 : S^1 \to M$ be curves with*

$$d(\gamma_0(t), \gamma_1(t)) \leq \rho_0 \quad \textit{for all}\, t \in S^1.$$

Then γ_0 and γ_1 are homotopic.

Proof. For any $t \in S^1$ let $c_t(s) : I \to M$ be the unique shortest geodesic from $\gamma_0(t)$ to $\gamma_1(t)$ (Corollary 1.4.4), as usual parametrized proportionally to arc length. Since c_t depends continuously on its end points by Corollary 1.4.4, hence on t,

$$\Gamma(t, s) := c_t(s)$$

is continuous and yields the desired homotopy. □

Proof of Theorem 1.4.6: Let $(\gamma_n)_{n\in\mathbb{N}}$ be a minimizing sequence for arc length in the given homotopy class. Here and in the sequel, all curves are parametrized proportionally to arc length. We may assume w.l.o.g. that the curves γ_n are piecewise geodesic; namely, for each curve, we may find $t_0 = 0 < t_1 < t_2 < \ldots < t_m < t_{m+1} = 2\pi$ with the property that

$$L(\gamma_{n|[t_{j-1},t_j]}) \leq \rho_0/2$$
$$(\rho_0 \text{ as in Corollary 1.4.4) for } j = 1, \ldots, m+1 \text{ with}\, t_{m+1} := 2\pi).$$

Replacing $\gamma_{n|[t_{j-1},t_j]}$ by the shortest geodesic arc between $\gamma_n(t_{j-1})$ and $\gamma_n(t_j)$, we obtain a curve which is homotopic to and not longer than γ_n (the same argument also shows that each homotopy class does contain curves of finite length).

We may thus assume that for any γ_n there exist points $p_{0,n}, \ldots, p_{m,n}$ for which $d(p_{j-1,n}, p_{j,n}) \leq \rho_0$ ($p_{m+1,n} := p_{0,n}, j = 1, \ldots, m+1$) and for which γ_n contains the shortest geodesic arc between $p_{j-1,n}$ and $p_{j,n}$. Since the lengths of the γ_n are bounded as they constitute a minimizing sequence, we may also assume that m is independent of n. After selection of a subsequence, by the compactness of M, the points $p_{0,n}, \ldots, p_{m,n}$ converge to points $p_0, \ldots, p_m$, for $n \to \infty$. The segment of γ_n between $p_{j-1,n}$ and $p_{j,n}$ then converges to the shortest geodesic arc between p_{j-1} and p_j, for example by Corollary 1.4.4. The union of these geodesic segments yields a curve γ. By Lemma 1.4.7, γ is homotopic to the γ_n, and

$$L(\gamma) = \lim_{n\to\infty} L(\gamma_n),$$

and since the curves γ_n are minimizing sequence for the length in their homotopy class, γ is a shortest curve in this class. Therefore, γ has to be geodesic. Namely, otherwise, there would exist points p and q on γ for which one of the two segments of γ between p and q would have length at most ρ_0, but would not be geodesic. By Corollary 1.4.4, γ could then be shortened by replacing this segment by the shortest geodesic arc between p and q. By the argument

of Lemma 1.4.7, this does not change the homotopy class, and we obtain a contradiction to the minimizing property of γ.
γ thus is the desired closed geodesic. □

Corollary 1.4.5 *On any compact Riemannian manifold* M_1, *any two points* p, q *can be connected by a curve of shortest length, and this curve is geodesic.*

Proof. Minimize over all curves between p and q (and not only over those in a fixed homotopy class) as in the proof of Theorem 1.4.6. □

We also show

Theorem 1.4.7 *Let* M *be a compact Riemannian manifold. Then for any* $p \in M$, *the exponential map* $\exp_p$ *is defined on all of* T_pM, *and any geodesic may be extended indefinitely in each direction.*

Proof. For $v \in T_pM$, let

$$\Lambda := \{t \in \mathbb{R}^+ : c_v \text{ is defined on } [-t, t]\},$$

where c_v is, as usual, the geodesic with $c_v(0) = p, \dot{c}_v(0) = v$. It follows from $c_v(-t) = c_{-v}(t)$ that e_v may also be defined for negative t, at the moment at least for those with sufficiently small absolute value. Theorem 1.4.2 implies $\Lambda \neq \emptyset$. The compactness of M implies the closedness of Λ. We shall now show openness of Λ : Let c_v be defined on $[-t, t]$; for example $\dot{c}_v(t) = w \in T_{c_v(t)}M$. By Theorem 1.4.2 there exists a geodesic $\gamma_w(s)$ with $\gamma_w(0) = c_v(t), \dot{\gamma}_w(0) = \dot{c}_v(t)$, for $s \in [0, \varepsilon]$ and $\varepsilon > 0$. Putting $c_v(t+s) = \gamma_w(s)$ for $s \in [0, \varepsilon]$, we have extended c_v to $[-t, t + \varepsilon]$. Analogously, c_v may be extended in the direction of negative t. This implies openness of Λ, hence $\Lambda = \mathbb{R}^+$. The claims follow easily. □

We now want to address the question whether the results of Theorem 1.4.7 continue to hold for a more general class of Riemannian manifolds than the compact ones. Obviously, they do hold for Euclidean space which is not compact, but they do not hold for any proper open subset of Euclidean space, essentially since such a set is not complete. It will turn out that completeness will be the right condition for extending Theorem 1.4.7.

Definition 1.4.6 A Riemannian manifold M is *geodesically complete* if for all $p \in M$, the exponential map $\exp_p$ is defined on all of T_pM, or, in other words, if any geodesic $c(t)$ with $c(0) = p$ is defined for all $t \in \mathbb{R}$.

We can now state the **Theorem of Hopf-Rinow.**

Theorem 1.4.8 *Let* M *be a Riemannian manifold. The following statements are equivalent:*

(i) *M is complete as a metric space (or equivalently, it is complete as a topological space w.r.t. its underlying topology, see Corollary 1.4.1).*

(ii) *The closed and bounded subsets of M are compact.*

(iii) *There exists $p \in M$ for which $\exp_p$ is defined on all of T_pM.*

(iv) *M is geodesically complete, i.e. for every $p \in M, \exp_p$ is defined on all of T_pM.*

Furthermore, each of the statements (i) – (iv) implies

(v) *Any two points $p, q \in M$ can be joined by a geodesic of length $d(p, q)$, i.e. by a geodesic of shortest length.*

Proof. We shall first prove that if $\exp_p$ is defined on all of T_pM, then any $q \in M$ can be connected with p by a shortest geodesic. In particular, this will show the implication (iv) ⇒ (v).

For this purpose, let

$$r := d(p, q),$$

and let $\rho > 0$ be given by Corollary 1.4.2, let $p_0 \in \partial B(p, \rho)$ be a point where the continuous function $d(q, \cdot)$ attains its minimum on the compact set $\partial B(p, \rho)$. Then $p_0 = \exp_p \rho V$, for some $V \in T_pM$. We consider the geodesic

$$c(t) := \exp_p tV,$$

and we want to show that

$$c(r) = q. \tag{1.4.19}$$

$c_{|[0,r]}$ will then be a shortest geodesic from p to q.

For this purpose, let

$$I := \{t \in [0, r] : d(c(t), q) = r - t\}.$$

(1.4.19) means $r \in I$, and we shall show $I = [0, r]$ for that purpose. I is not empty, as it contains 0 by definition of r, and it is closed for continuity reasons. $I = [0, r]$ will therefore follow if we can show openness of I.

Let $t_0 \in I$. Let $\rho_1 > 0$ be the radius of Corollary 1.4.2 corresponding to the point $c(t_0) \in M$. W.l.o.g. $\rho_1 \leq r - t_0$. Let $p_1 \in \partial B(c(t_0), \rho_1)$ be a point where the continuous function $d(q, \cdot)$ assumes its minimum on the compact set $\partial B(c(t_0), \rho_1)$. Then

$$d(p, p_1) \geq d(p, q) - d(q, p_1). \tag{1.4.20}$$

Now for every curve γ from $c(t_0)$ to q, there exists some

$$\gamma(t) \in \partial B(c(t_0), \rho_1).$$

Hence

$$\begin{aligned} L(\gamma) &\geq d(c(t_0), \gamma(t)) + d(\gamma(t), q) \\ &= \rho_1 + d(\gamma(t), q) \\ &\geq \rho_1 + d(p_1, q) \quad \text{because of the minimizing property of } p_1. \end{aligned}$$

Hence also

$$d(q, c(t_0)) \geq \rho_1 + d(p_1, q) \tag{1.4.21}$$

and by the triangle inequality, we then actually must have equality. Inserting (1.4.21) into (1.4.20) and recalling $d(q, c(t_0)) = r - t_0$ gives

$$d(p, p_1) \geq r - (r - t_0 - \rho_1) = t_0 + \rho_1.$$

On the other hand, there exists a curve from p to p_1 of length $t_0 + \rho_1$; namely one goes from p to $c(t_0)$ along c and then takes the geodesic from $c(t_0)$ to p_1 of length ρ_1. That curve thus is shortest and therefore has to be geodesic as shown in the proof of Theorem 1.4.6. By uniqueness of geodesics with given initial values, it has to coincide with c, and then

$$p_1 = c(t_0 + \rho_1).$$

Since we observed that equality has to hold in (1.4.21), we get

$$d(q, c(t_0 + \rho_1)) = r - (t_0 + \rho_1),$$

hence

$$t_0 + \rho_1 \in I,$$

and openness of I follows, proving our claim.

It is now easy to complete the proof of Theorem 1.4.8:

(iv) $\Rightarrow$ (iii) is trivial

(iii) $\Rightarrow$ (ii) Let $K \subset M$ be closed and bounded. Since bounded, $K \subset B(p, r)$ for some $r > 0$. By what we have shown in the beginning, any point in $B(p, r)$ can be connected with p by a geodesic (of length $\leq r$). Hence, $B(p, r)$ is the image of the compact ball in T_pM of radius r under the continuous map $\exp_p$. Hence, $B(p, r)$ is compact itself. Since K is assumed to be closed and shown to be contained in a compact set, it must be compact itself.

(ii) $\Rightarrow$ (i) Let $(p_n)_{n \in \mathbb{N}} \subset M$ be a Cauchy sequence. It then is bounded, and, by (ii), its closure is compact. It therefore contains a convergent subsequence, and being Cauchy, it has to converge itself. This shows completeness of M.

(i) $\Rightarrow$ (iv) Let c be a geodesic in M, parametrized by arc length, and being defined on a maximal interval I. I then is nonempty, and by Theorem 1.4.2, it is also open. To show closedness, let $(t_n)_{n \in \mathbb{N}} \subset I$ converge to t.

Since

$$d(c(t_n), c(t_m)) \leq |t_n - t_m|$$

as c is parametrized by arc length, $c(t_n)$ is a Cauchy sequence, hence has a limit $p \in M$, because we assume M to be complete. Let $\rho > 0$ be as in Corollary 1.4.2. Then $B(p, \rho)$ is compact, being the image of the compact ball of radius r in T_pM under the continuous map $\exp_p$. Therefore, the argument of Corollary 1.4.3 and Corollary 1.4.4 applies to show that there exists $\rho_0 > 0$ with the property that for any point $q \in B(p, \rho)$ any geodesic starting from q can be extended at least up to length ρ_0.

Since $c(t_n)$ converges to p, for all sufficiently large m, n

$$d(c(t_n), c(t_m)) \leq |t_n - t_m| \leq \rho_0/2$$

and

$$d(c(t_n), p), d(c(t_m), p) \leq \rho_0.$$

Therefore, the shortest geodesic from $c(t_n)$ to $c(t_m)$ can be defined at least on the interval $[-\rho_0, \rho_0]$. This shortest geodesic, however, has to be a subarc of c, and c thus can be defined up to the parameter value $t_n + \rho_0$, in particular for t, showing closedness of I. □

Let now M be a differentiable submanifold of the Riemannian manifold N. The Riemannian metric of N then induces a Riemannian metric on M, by restricting the former one to $T_pM \subset T_pN$ for $p \in N$. Thus, M also becomes a Riemannian manifold.

In particular, $S^n \subset \mathbb{R}^{n+1}$ obtains a Riemannian metric. We want to compute this metric in the local chart of 1.1, namely

$$\begin{aligned} f(x^1, \ldots, x^{n+1}) &= \left(\frac{x^1}{1 - x^{n+1}}, \ldots, \frac{x^n}{1 - x^{n+1}} \right) \quad \text{for } x^{n+1} \neq 1 \\ &=: (y^1, \ldots, y^n) \in \mathbb{R}^n\,. \end{aligned}$$

In the sequel, a Latin index occuring twice in a product has to be summed from 1 to $n+1$, a Greek one from 1 to n. We compute

$$1 = x^i x^i = y^\alpha y^\alpha (1 - x^{n+1})^2 + x^{n+1} x^{n+1}$$

hence

$$x^{n+1} = \frac{y^\alpha y^\alpha - 1}{y^\alpha y^\alpha + 1}$$

and then

$$x^i = \frac{2y^i}{1 + y^\alpha y^\alpha} \quad (i = 1, \ldots, n).$$

For $g := f^{-1}$ then

$$\begin{aligned} \frac{\partial g^j}{\partial y^k} &= \frac{2\delta_{jk}}{1 + y^\alpha y^\alpha} - \frac{4y^j y^k}{(1 + y^\alpha y^\alpha)^2} \quad \text{for } j = 1, \ldots, n,\ k = 1, \ldots, n \\ \frac{\partial g^{n+1}}{\partial y^k} &= \frac{4y^k}{(1 + y^\alpha y^\alpha)^2}. \end{aligned}$$

Let a tangent vector to S^n be represented by $w = (w^1, \dots, w^n)$ in our chart. Then

$$\begin{aligned}\langle w, w\rangle &= dg(w)\cdot dg(w),\ \text{where the point denotes the Euclidean}\\ &\qquad\qquad\qquad\text{scalar product of } \mathbb{R}^{n+1}\\ &= \frac{1}{(1+y^\alpha y^\alpha)^4}\left\{4(1+y^\alpha y^\alpha)^2 w^\beta w^\beta - 16(1+y^\alpha y^\alpha)y^\beta w^\beta y^\gamma w^\gamma\right.\\ &\qquad\qquad\left. + 16 y^\beta y^\beta y^\gamma w^\gamma y^\delta w^\delta + 16 y^\beta w^\beta y^\gamma w^\gamma\right\}\\ &= \frac{4}{(1+y^\alpha y^\alpha)^2} w^\beta w^\beta.\end{aligned}$$

Thus, the metric in our chart is given by

$$g_{ij}(y) = \frac{4}{(1+|y|^2)^2}\delta_{ij}.$$

Definition 1.4.7 A diffeomorphism $h : M \to N$ between Riemannian manifolds is an *isometry* if it preserves the Riemannian metric. Thus, for $p \in M, v, w \in T_pM$, and if $\langle\cdot,\cdot\rangle_M$ and $\langle\cdot,\cdot\rangle_N$ denotes the scalar products in T_pM and $T_{h(p)}N$, resp., we have

$$\langle v, w\rangle_M = \langle dh(v), dh(w)\rangle_N.$$

A differentiable map $h : M \to N$ is a *local isometry* if for every $p \in M$ there exists a neighborhood U for which $h_{|U} : U \to h(U)$ is an isometry, and $h(U)$ is open in N.

If $(g_{ij}(p))$ and $(\gamma_{\alpha\beta}(h(p))$ are the coordinate representations of the metric, an isometry has to satisfy

$$g_{ij}(p) = \gamma_{\alpha\beta}(h(p))\frac{\partial h^\alpha(p)}{\partial x^i}\frac{\partial h^\beta(p)}{\partial x^j}.$$

A local isometry thus has the same effect as a coordinate change. Isometries leave the lengths of tangent vectors and therefore also the lengths and energies of curves invariant. Thus, critical points, i.e. geodesics, are mapped to geodesics.

With this remark, we may easily determine the geodesics of S^n. The orthogonal group $\mathrm{O}(n+1)$ operates isometrically on $\mathbb{R}^{n+1}$, and since it maps S^n into S^n, it also operates isometrically on S^n. Let now $p \in S^n, v \in T_pS^n$. Let E be the two dimensional plane through the origin of $\mathbb{R}^{n+1}$, containing v. We claim that the geodesic c_v through p with tangent vector v is nothing but the great circle through p with tangent vector v (parametrized proportionally to arc length), i.e. the intersection of S^n with E. For this, let $S \in \mathrm{O}(n+1)$ be the reflection across that E. Together with c_v, Sc_v is also a geodesic through p with tangent vector v. The uniqueness result of Theorem 1.4.2 implies $c_v = Sc_v$, and thus the image of c_v is the great circle, as claimed.

As another example, we consider the torus T^2 introduced in 1.1. We introduce a metric on T^2 by letting the projection π be a local isometry. For each chart of the form $(U, (\pi_{|U})^{-1})$, we use the Euclidean metric on $\pi^{-1}(U)$. Since the translations

$$z \mapsto z + m_1 w_1 + m_2 w_2 \quad (m_1, m_2 \in \mathbb{Z})$$

are Euclidean isometries, the Euclidean metrics on the different components of $\pi^{-1}(U)$ (which are obtained from each other by such translations) yield the same metric on U. Hence, the Riemannian metric on T^2 is well defined.

Since π is a local isometry, Euclidean geodesics of $\mathbb{R}^2$ are mapped onto geodesics of T^2. The global behavior of geodesics on such a torus is most easily studied in the case where T^2 is generated by the two unit vectors $w_1 = (1, 0)$ and $w_2 = (0, 1)$: A straight line in $\mathbb{R}^2$ which is parallel to one of the coordinate axes then becomes a geodesic on T^2 that closes up after going around once. More generally, a straight line with rational slope becomes a closed, hence periodic geodesic on T^2, while the image of one with irrational slope lies dense in T^2.

Before ending this paragraph, we want to introduce the following important notion:

Definition 1.4.8 Let M be a Riemannian manifold, $p \in M$. The *injectivity radius* of p is

$$i(p) := \sup\{\rho > 0 : \exp_p \text{ is defined on } d_\rho(0) \subset T_pM \text{ and injective}\}.$$

The *injectivity radius* of M is

$$i(M) := \inf_{p \in M} i(p).$$

For example, the injectivity radius of the sphere S^n is π, since the exponential map of any point p maps the open ball of radius π in T_pM injectively onto the complement of the antipodal point of p.

The injectivity radius of the torus just discussed is $\frac{1}{2}$, since here the exponential map is injective on the interior of a square with centre $0 \in T_pM$ and side length 1.

Perspectives. As the name suggests, the concept of a Riemannian metric was introduced by B. Riemann, in his habilitation address [254]. He also suggested to consider more generally metrics obtained by taking metrics on the tangent spaces that are not induced by a scalar product. Such metrics were first systematically investigated by Finsler and are therefore called Finsler metrics.

For a general metric space, a geodesic is defined as a curve which realizes the shortest distance between any two sufficiently close points lying on it. Those metric spaces that satisfy the conclusion of the Hopf-Rinow theorem that any two points can be connected by a shortest geodesic are called geodesic length spaces, and they are amenable to geometric constructions as demonstrated by the school of Alexandrov. See e.g. [194], [15].

A Lorentz metric on a differentiable manifold of dimension $d+1$ is given by an inner product of signature $(1, d)$ on each tangent space T_pM depending smoothly on p. A Lorentz manifold is a differentiable manifold with a Lorentz metric. The prototype is Minkowski space, namely $\mathbb{R}^{d+1}$ equipped with the inner product

$$\langle x, y\rangle = -x^0y^0 + x^1y^1 + \ldots + x^dy^d$$

for $x = (x^0, x^1, \ldots, x^d), y = (y^0, y^1, \ldots, y^d)$. Lorentz manifolds are the spaces occuring in general relativity. Let us briefly discuss some concepts. Tangent vectors V with negative, positive, vanishing $\|V\|^2 = \langle V, V\rangle$ are called time-like, space-like, and light-like, resp. Length and energy of a curve may be defined formally as in the Riemannian case, and we again obtain geodesic equations. Geodesics whose tangent vectors all have norm zero are called null geodesics. They describe the paths of light rays. (Note that in our above description of the Minkowski metric, the conventions have been chosen so that the speed of light is 1.) Submanifolds of Lorentz manifolds whose tangent vectors are all space-like are ordinary Riemannian manifolds w.r.t. the induced metric. For treatments of Lorentzian geometry, an introduction is [209]. Deeper aspects are treated in Hawking and Ellis[113].

J. Nash proved that every Riemannian manifold M can be isometrically embedded into some Euclidean space $\mathbb{R}^k$. For the proof of this result, he developed an implicit function theorem in Fréchet spaces and an iteration technique that have found other important applications. A simpler proof was found by Günther[107].

Although on a conceptual level, Nash's theorem reduces the study of Riemannian manifolds to the study of submanifolds of Euclidean spaces, in practice the intrinsic point of view has proved to be preferable (see Perspectives on 1.3).

In our presentation, we only consider finite dimensional Riemannian manifolds. It is also possible, and often very useful, to introduce infinite dimensional Riemannian manifolds. Those are locally modeled on Hilbert spaces instead of Euclidean ones. The lack of local compactness leads to certain technical complications, but most ideas and constructions of Riemannian geometry pertain to the infinite dimensional case. Such infinite dimensional manifolds arise for example naturally as certain spaces of curves on finite dimensional Riemannian manifolds. A thorough treatment is given in [159].

1.5 Vector Bundles

Definition 1.5.1 A *(differentiable) vector bundle* of rank n consists of a total space E, a base M, and a projection $\pi : E \to M$, where E and M are differentiable manifolds, π is differentiable, each "fiber" $E_x := \pi^{-1}(x)$ for $x \in M$, carries the structure of an n-dimensional (real) vector space, and the following local triviality requirement is satisfied: For each $x \in M$, there exist a neighborhood U and a diffeomorphism

$$\varphi : \pi^{-1}(U) \to U \times \mathbb{R}^n$$

with the property that for every $y \in U$

$$\varphi_y := \varphi_{|E_y} : E_y \to \{y\} \times \mathbb{R}^n$$

is a vector space isomorphism, i.e. a bijective linear map. Such a pair (φ, U) is called a *bundle chart.*

In the sequel, we shall omit the word "differentiable" for a vector bundle. Often, a vector bundle will simply be denoted by its total space.

It is important to point out that a vector bundle is by definition locally, but not necessarily globally a product of base and fiber. A vector bundle which is isomorphic to $M \times \mathbb{R}^n$ ($n =$ rank) is called trivial.

A vector bundle may be considered as a family of vector spaces (all isomorphic to a fixed model $\mathbb{R}^n$) parametrized (in a locally trivial manner) by a manifold.

Let now (E, π, M) be a vector bundle of rank n, $(U_\alpha)_{\alpha \in A}$ a covering of M by open sets over which the bundle is trivial, and $\varphi_\alpha : \pi^{-1}(U_\alpha) \to U_\alpha \times \mathbb{R}^n$ be the corresponding local trivializations. For nonempty $U_\alpha \cap U_\beta$,we obtain *transition maps*

$$\varphi_{\beta\alpha} : U_\alpha \cap U_\beta \to \mathrm{Gl}\,(n, \mathbb{R})$$

by

$$\varphi_\beta \circ \varphi_\alpha^{-1}(x, v) = (x, \varphi_{\beta\alpha}(x)v) \quad \text{for}\, x \in U_\alpha \cap U_\beta, v \in \mathbb{R}^n, \tag{1.5.1}$$

where $\mathrm{Gl}(n, \mathbb{R})$ is the general linear group of bijective linear self maps of $\mathbb{R}^n$. The transition maps express the transformation behavior of a vector in the fiber under a change of local trivialization.

The transition maps satisfy

$$\varphi_{\alpha\alpha}(x) = \mathrm{id}_{\mathbb{R}^n} \quad \text{for} \quad x \in U_\alpha \tag{1.5.2}$$
$$\varphi_{\alpha\beta}(x)\varphi_{\beta\alpha}(x) = \mathrm{id}_{\mathbb{R}^n} \quad \text{for} \quad x \in U_\alpha \cap U_\beta \tag{1.5.3}$$
$$\varphi_{\alpha\gamma}(x)\varphi_{\gamma\beta}(x)\varphi_{\beta\alpha}(x) = \mathrm{id}_{\mathbb{R}^n} \quad \text{for} \quad x \in U_\alpha \cap U_\beta \cap U_\gamma\,. \tag{1.5.4}$$

These properties are direct consequences of (1.5.1).

A vector bundle can be reconstructed from its transition maps.

Theorem 1.5.1

$$E = \coprod_{\alpha \in A} U_\alpha \times \mathbb{R}^n / \sim\,,$$

where $\amalg$ *denotes disjoint union, and the equivalence relation* $\sim$ *is defined by*

$$(x, v) \sim (y, w) :\iff x = y \text{ and } w = \varphi_{\beta\alpha}(x)v \quad (x \in U_\alpha, y \in U_\beta,\, v, w \in \mathbb{R}^n)$$

The *proof* of this fact is a straightforward verification of the properties required in Definition 1.5.1. A reader who does not want to carry this out him/herself may consult D. Husemoller, Fibre bundles, Springer, GTM 20, 1975. □

Definition 1.5.2 Let G be a subgroup of $\mathrm{Gl}(n, \mathbb{R})$, for example $\mathrm{O}(n)$ or $\mathrm{SO}(n)$, the orthogonal or special orthogonal group. We say that a vector bundle has the *structure group* G if there exists an atlas of bundle charts for which all transition maps have their values in G.

Definition 1.5.3 Let (E, π, M) be a vector bundle. A *section* of E is a differentiable map $s : M \to E$ with $\pi \circ s = \mathrm{id}_M$. The space of sections of E is denoted by $\Gamma(E)$.

We have already seen an example of a vector bundle above, namely the tangent bundle TM of a differentiable manifold M.

Definition 1.5.4 A section of the tangent bundle TM of M is called a *vector field* on M.

Let now $f : M \to N$ be a differentiable map, (E, π, N) a vector bundle over N. We want to pull back the bundle via f, i.e. construct a bundle f^*E, for which the fiber over $x \in M$ is $E_{f(x)}$, the fiber over the image of x.

Definition 1.5.5 The *pulled back bundle* f^*E is the bundle over M with bundle charts $(\varphi \circ f, f^{-1}(U))$, where (φ, U) are bundle charts of E.

We now want to extend some algebraic concepts and constructions from vector spaces to vector bundles by performing them fiberwise. For example:

Definition 1.5.6 Let (E_1, π_1, M) and (E_2, π_2, M) be vector bundles over M. Let the differentiable map $f : E_1 \to E_2$ be fiber preserving, i.e.

$$\pi_2 \circ f = \pi_1,$$

and let the fiber maps $f_x : E_{1,x} \to E_{2,x}$ be linear, i.e. vector space homomorphisms. Then f is called a *bundle homomorphism.*

Definition 1.5.7 Let (E, π, M) be a vector bundle of rank n. Let $E' \subset E$, and suppose that for any $x \in M$ there exists a bundle chart (φ, U) with $x \in U$ and

$$\varphi(\pi^{-1}(U) \cap E') = U \times \mathbb{R}^m (\subset U \times \mathbb{R}^n, m \le n).$$

The resulting vector bundle $(E', \pi_{|E'}, M)$ is called *subbundle* of E of rank m.

Let us discuss an example: $S^1 = \{x \in \mathbb{R}^2 : |x|^2 = 1\}$ is a submanifold of $\mathbb{R}^2$. If we restrict the tangent bundle $T\mathbb{R}^2$ of $\mathbb{R}^2$ to S^1, we obtain a bundle E over S^1 that is isomorphic to $S^1 \times \mathbb{R}^2$. The tangent bundle of S^1 has fiber $T_x S^1 = \{y \in \mathbb{R}^2 : x \cdot y = 0\} \subset \mathbb{R}^2$ (where the dot $\cdot$ denotes the Euclidean scalar product). TS^1 is a subbundle of $T\mathbb{R}^2|S^1$; the reader is invited to write down explicit bundle charts.

Definition 1.5.8 Let (E_1, π_1, M) and (E_2, π_2, M) be vector bundles over M. The *Cartesian product* of E_1 and E_2 is the vector bundle over M with fiber $E_{1,x} \times E_{2,x}$ and bundle charts $(\varphi_\alpha \times \psi_\beta, U_\alpha \cap V_\beta)$, where $(\varphi_\alpha, U_\alpha)$ and (ψ_β, V_β) are bundle charts for E_1 and E_2 resp., and

$$(\varphi_\alpha \times \psi_\beta)(x, (v, w)) := (\varphi_\alpha(x, v), \psi_\beta(x, w)) \quad (v \in E_{1,x}, w \in E_{2,x}).$$

Thus, the *product bundle* is simply the bundle with fiber over $x \in M$ being the product of the fibers of E_1 and E_2 over x. By this pattern, all constructions for vector spaces can be extended to vector bundles. Of particular importance for us will be dual space, exterior and tensor product. Let us briefly recall the definition of the latter:

Let V and W be vector spaces (as always over $\mathbb{R}$) of dimension m and n, resp., and let $(e_1, \dots, e_m)$ and $(f_1, \dots, f_n)$ be bases. Then $V \otimes W$ is the vector space of dimension mn spanned by the basis $(e_i \otimes f_j)_{\substack{i=1,\dots,m \\ j=1,\dots,n}}$. There exists a canonical bilinear map

$$L : V \times W \to V \otimes W$$

mapping $(a^i e_i, b^j f_j)$ onto $a^i b^j e_i \otimes f_j$

One may then also define the *tensor product* of more than two vector spaces in an associative manner.

Definition 1.5.9 Let M be a differentiable manifold, $x \in M$. The vector space dual to the tangent space $T_x M$ to $\mathbb{R}$ is called the *cotangent space* of M at the point x and denoted by $T_x^* M$. The vector bundle over M whose fibers are the cotangent spaces of M is called the *cotangent bundle* of M and denoted by $T^* M$. Elements of $T^* M$ are called *cotangent vectors,* sections of $T^* M$ are 1-*forms.*

We now want to study the transformation behavior of cotangent vectors. Let $(e_i)_{i=1,\dots,d}$ be a basis of $T_x M$ and $(\omega^j)_{j=1,\dots,d}$ the dual basis of $T_x^* M$, i.e.

$$\omega^j(e_i) = \delta_i^j = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases} .$$

Moreover, let $v = v^i e_i \in T_x M$, $\eta = \eta_j \omega^j \in T_x^* M$. We have $\eta(v) = \eta_i v^i$. Let the bases (e_i) and (ω^j) be given by local coordinates, i.e.

$$e_i = \frac{\partial}{\partial x^i}, \quad \omega^j = dx^j.$$

Let now f be a coordinate change. v is transformed to

$$f_*(v) := v^i \frac{\partial f^\alpha}{\partial x^i} \frac{\partial}{\partial f^\alpha}.$$

η then has to be transformed to

$$f^*(\eta) := \eta_j \frac{\partial x^j}{\partial f^\beta} \, df^\beta$$

because in this case

$$f^*(\eta)(f_*(v)) = \eta_j \frac{\partial x^j}{\partial f^\alpha} v^i \frac{\partial f^\alpha}{\partial x^i} = \eta_i v^i = \eta(v).$$

Thus a tangent vector transforms with the functional matrix of the coordinate change whereas a cotangent vector transforms with the transposed inverse of this matrix. This different transformation behavior is expressed by the following definition:

Definition 1.5.10 A p times *contravariant* and q times *covariant* tensor on a differentiable manifold M is a section of

$$\underbrace{TM \otimes \ldots \otimes TM}_{p \text{ times}} \quad \otimes \quad \underbrace{T^*M \otimes \ldots \otimes T^*M}_{q \text{ times}} .$$

Actually, one should speak of a *tensor field,* because “tensor” often also means an element of the corresponding fibers, in the same manner, as a (tangent) vector is an element of T_xM and a vector field a section of TM.

If f is a coordinate change, a p times contravariant and q times covariant tensor is transformed p times by the matrix (df) and q times by the matrix $(df^{-1})^t$.

Lemma 1.5.1 *A Riemannian metric on a differentiable manifold M is a two times covariant (and symmetric and positive definite) tensor on M.*

Proof. From the formula (1.4.3) for the transformation behavior of a Riemannian metric. □

A Riemannian metric thus is a section of $T^*M \otimes T^*M$. We consequently write the metric in local coordinates as

$$g_{ij}(x) dx^i \otimes dx^j.$$

Theorem 1.5.2 *The tangent bundle of a Riemannian manifold M of dimension d has structure group $\mathrm{O}(d)$.*

Proof. Let (f, U) be a bundle chart for TM,

$$f : \pi^{-1}(U) \to U \times \mathbb{R}^d.$$

Let $e_1, \dots, e_d$ be the canonical basis vectors of $\mathbb{R}^d$, and let $v_1, \dots, v_d$ be the sections of $\pi^{-1}(U)$ with $f(v_i) = e_i, i = 1, \dots, d$. Applying the Gram-Schmidt orthogonalization procedure to $v_1(x), \dots, v_d(x)$ for each $x \in U$ we obtain sections $w_1, \dots, w_d$ of $\pi^{-1}(U)$ for which $w_1(x), \dots, w_d(x)$ are an orthonormal basis w.r.t. the Riemannian metric on T_xM, for each $x \in U$. By

$$\begin{aligned} f' : \pi^{-1}(U) &\to U \times \mathbb{R}^d \\ \lambda^i w_i(x) &\mapsto (x, \lambda^1, \dots, \lambda^d) \end{aligned}$$

we then get a bundle chart which maps the basis $w_1(x), \dots, w_d(x)$, i.e. an orthonormal basis w.r.t. the Riemannian metric, for each $x \in U$ onto an Euclidean orthonormal basis of $\mathbb{R}^d$. We apply this orthonormalization process for each bundle chart and obtain a new bundle atlas whose transition maps always map an Euclidean orthonormal basis of $\mathbb{R}^d$ into another such basis, and are hence in $\mathrm{O}(d)$. □

We want to point out, however, that in general there do not exist local coordinates for which $w_i(x) = \frac{\partial}{\partial x^i}$ for $i = 1, \dots, d$.

Corollary 1.5.1 *The tangent bundle of an oriented Riemannian manifold of dimension d has structure group* SO *(d).*

Proof. The orientation allows to select an atlas for which all transition maps have positive functional determinant. From this, one sees that we also may obtain transition functions for the tangent bundle with positive determinant. The orthonormalization process of Theorem 1.5.2 preserves the positivity of the determinant, and thus, in the oriented case, we obtain a new bundle atlas with transition maps in $\mathrm{SO}(d)$. □

Definition 1.5.11 Let (E, π, M) be a a vector bundle. A *bundle metric* is given by a family of scalar products on the fibers E_x, depending smoothly on $x \in M$.

In the same manner as Theorem 1.5.2, one shows

Theorem 1.5.3 *Each vector bundle (E, π, M) of rank n with a bundle metric has structure group $\mathrm{O}(n)$. In particular, there exist bundle charts $(f, U), f : \pi^{-1}(U) \to U \times \mathbb{R}^n$, for which for all $x \in U$, $f^{-1}(x, (e_1, \dots, e_n))$ is an orthonormal basis of E_x ($e_1, \dots, e_n$ is an orthonormal basis of $\mathbb{R}^n$).* □

Definition 1.5.12 The bundle charts of Theorem 1.5.3 are called *metric.*

In the same manner as Theorem 1.4.1, one shows

Theorem 1.5.4 *Each vector bundle can be equipped with a bundle metric.* □

It will be more important for us, however, that a Riemannian metric automatically induces bundle metrics on all tensor bundles over M. The metric of the cotangent bundle is given in local coordinates by

$$\langle \omega, \eta \rangle = g^{ij} \omega_i \eta_j \quad \text{for } \omega = \omega_i dx^i, \eta = \eta_i dx^i. \tag{1.5.5}$$

(We recall that (g^{ij}) is the matrix inverse to (g_{ij})).

Namely, this expression has the correct transformation behavior under coordinate changes: If $w \mapsto x(w)$ is a coordinate change, we get

$$\omega_i dx^i = \omega_i \frac{\partial x^i}{\partial w^\alpha} dw^\alpha =: \tilde{\omega}_\alpha dw^\alpha,$$

while g^{ij} is transformed into

$$h^{\alpha\beta} = g^{ij} \frac{\partial w^\alpha}{\partial x^i} \frac{\partial w^\beta}{\partial x^j},$$

and

$$h^{\alpha\beta} \tilde{\omega}_\alpha \tilde{\eta}_\beta = g^{ij} \omega_i \eta_j.$$

Moreover, we get

$$\|\omega(x)\| = \sup\{\omega(x)(v) : v \in T_x M, \|v\| = 1\}.$$

A Riemannian metric also induces an identification between TM and T^*M :

$$v = v^i \frac{\partial}{\partial x^i} \text{ corresponds to } \omega = \omega_j dx^j$$

$$\text{with} \quad \omega_j = g_{ij} v^i$$

$$\text{or} \quad v^i = g^{ij} \omega_j \, .$$

(1.5.5) may also be justified as follows:

Under this identification, to $v \in T_x M$ there corresponds a 1-form $\omega \in T_x^* M$ via

$$\omega(w) := \langle v, w \rangle \quad \text{for all } w$$

and (1.5.5) means then that

$$\|\omega\| = \|v\|.$$

For example, on $TM \otimes TM$, the metric is given by

$$\langle v \otimes w, \xi \otimes \eta \rangle = g_{ij} v^i \xi^j g_{k\ell} w^k \eta^l \tag{1.5.6}$$

($v = v^i \frac{\partial}{\partial x^i}$ etc. in local coordinates).

Definition 1.5.13 A local orthonormal basis of $T_x M$ of the type obtained in Theorem 1.5.3 is called an *(orthonormal) frame field.*

We put

$$\Lambda^p(T_x^*M) := \underbrace{T_x^*M \wedge \ldots \wedge T_x^*M}_{p \text{ times}} \quad \text{(exterior product).}$$

On $\Lambda^p(T_x^*M)$, we have two important operations: First, the exterior product by $\eta \in T_x^*M = \Lambda^1(T_x^*M)$:

$$\begin{aligned} \Lambda^p(T_x^*M) &\to \Lambda^{p+1}(T_x^*M) \\ \omega &\longmapsto \epsilon(\eta)\omega := \eta \wedge \omega. \end{aligned}$$

Second, the interior product or contraction by an element $v \in T_xM$:

$$\begin{aligned} \Lambda^p(T_x^*M) &\to \Lambda^{p-1}(T_x^*M) \\ \omega &\longmapsto \iota(v)\omega \end{aligned}$$

with

$$\begin{aligned} (\iota(v)\omega(v_1, \ldots, v_{p-1}) &:= \omega(v, v_1, \ldots, v_{p-1}) \\ &\qquad \text{for } v, v_1, \ldots, v_{p-1} \in T_xM. \end{aligned}$$

In fact, such constructions may be carried out with any vector space W and its dual W^* in place of T_x^*M and T_xM. This will be relevant in §1.8.

The vector bundle over M with fiber $\Lambda^p(T_x^*M)$ over x is then denoted by $\Lambda^p(M)$.

Definition 1.5.14 The space of sections of $\Lambda^p(M)$ is denoted by $\Omega^p(M)$, i.e. $\Omega^p(M) = \Gamma(\Lambda^p(M))$. Elements of $\Omega^p(M)$ are called *(exterior) p-forms.*

A p-form thus is a sum of terms of the form

$$\omega(x) = \eta(x) dx^{i_1} \wedge \ldots \wedge dx^{i_p}$$

where $\eta(x)$ is a smooth function and $(x^1, \ldots, x^d)$ are local coordinates.

Definition 1.5.15 The *exterior derivative* $d : \Omega^p(M) \to \Omega^{p+1}(M)$ $(p = 0, \ldots, d = \dim M)$ is defined through the formula

$$d(\eta(x) dx^{i_1} \wedge \ldots \wedge dx^{i_p}) = \frac{\partial \eta(x)}{\partial x^j} dx^j \wedge dx^{i_1} \wedge \ldots \wedge dx^{i_p}$$

and extended by linearity to all of $\Omega^p(M)$.

Lemma 1.5.2 *If* $\omega \in \Omega^p(M), \vartheta \in \Omega^q(M)$, *then* $d(\omega \wedge \vartheta) = d\omega \wedge \vartheta + (-1)^p \omega \wedge d\vartheta$.

Proof. This easily follows from the formula $\omega \wedge \vartheta = (-1)^{pq} \vartheta \wedge \omega$ and the definition of d. □

Let $f : M \to N$ be a differentiable map,

$$\omega(z) = \eta(z) dz^{i_1} \wedge \ldots \wedge dz^{i_p} \in \Omega^p(N).$$

We then define

$$f^*(\omega)(x) = \eta(f(x)) \frac{\partial f^{i_1}}{\partial x^{\alpha_1}} dx^{\alpha_1} \wedge \ldots \wedge \frac{\partial f^{i_p}}{\partial x^{\alpha_p}} dx^{\alpha_p}.$$

This obviously is the correct transformation formula for p-forms.

Lemma 1.5.3

$$d(f^*(\omega)) = f^*(d\omega).$$

Proof. This easily follows from the transformation invariance

$$\frac{\partial \eta(z)}{\partial z^j} dz^j = \frac{\partial \eta(f(x))}{\partial z^j} \frac{\partial f^j}{\partial x^\alpha} dx^\alpha = \frac{\partial \eta(f(x))}{\partial x^\alpha} dx^\alpha. \qquad \square$$

Corollary 1.5.2 *d is independent of the choice of coordinates.*

Proof. Apply Lemma 1.5.3 to a coordinate transformation f. $\square$

Theorem 1.5.5 $d \circ d = 0$.

Proof. By linearity of d, it suffices to check the asserted identity on forms of the type

$$\omega(x) = f(x) dx^{i_1} \wedge \ldots \wedge dx^{i_p}.$$

Now

$$\begin{aligned} & d \circ d(\omega(x)) \\ &= d\big(\frac{\partial f}{\partial x^j} dx^j \wedge dx^{i_1} \wedge \ldots \wedge dx^{i_p}\big) \\ &= \frac{\partial^2 f}{\partial x^j \partial x^k} dx^k \wedge dx^j \wedge dx^{i_1} \wedge \ldots \wedge dx^{i_p} \\ &= 0, \end{aligned}$$

since $\frac{\partial^2 f}{\partial x^j \partial x^k} = \frac{\partial^2 f}{\partial x^k \partial x^j}$ (f is assumed to be smooth) and

$$dx^j \wedge dx^k = -dx^k \wedge dx^j. \qquad \square$$

Let now M be a differentiable submanifold of the Riemannian manifold $N; \dim M = m, \dim N = n$. We saw already that M then also carries a Riemannian metric. For $x \in M$, we define

$$T_x^\perp M \subset T_x N$$

by

$$T_x^{\perp}M := \{v \in T_xN : \forall\, w \in T_xM : \langle v, w\rangle = 0\},$$

where $\langle .,.\rangle$, as usual, is the scalar product given by the Riemannian metric.

The spaces $T_x^{\perp}M$ are the fibers of a vector bundle $T^{\perp}M$ over M, and TM and $T^{\perp}M$ are both subbundles of $TN_{|M}$, the restriction of TN to M (in a more complicated manner: $TN_{|M} = i^*TN$, where $i : M \to N$ is the differentiable embedding of M as a submanifold of N). In order to see this, one may choose the first m basis vectors $v_1, \dots, v_m$ of $TN_{|M}$ in the orthonormalization procedure of the proof of Theorem 1.5.2 in such a manner that they locally span TM.

Then TM is also locally spanned by $w_1, \dots, w_m$ (notation as in the proof of Theorem 1.5.2), and the remaining basis vectors then span $T^{\perp}M$, and we have

$$\langle w_i, w_\alpha\rangle = 0 \text{ for } i = 1, \dots, m, \alpha = m+1, \dots, n.$$

Thus, $T^{\perp}M$ is the orthogonal complement of TM in $TN_{|M}$.

Definition 1.5.16 $T^{\perp}M$ is called the *normal bundle* of M in N.

For our example of the submanifold S^1 of $\mathbb{R}^2$, $T^{\perp}S^1$ is the subbundle of $T\mathbb{R}^2_{|S^1}$, the restriction of $T\mathbb{R}^2$ to S^1, with fiber $T_x^{\perp}S^1 = \{\lambda x : \lambda \in \mathbb{R}\} \subset \mathbb{R}^2$.

We conclude this section with a consideration of the complex case – again, we remind the reader that is needed only in particular places, like 5.2.

Definition 1.5.17 A vector bundle E over a differentiable manifold M is called a *complex vector bundle* if each fiber $E_z = \pi^{-1}(z)$ is a complex vector space, i.e., isomorphic to $z \times \mathbb{C}^k$, and if that complex structure varies smoothly, that is, the local trivializations are of the form

$$\varphi : \pi^{-1}(U) \to U \times \mathbb{C}^k.$$

We thus have transition maps

$$\varphi_{\beta\alpha} : U_\alpha \cap U_\beta \to \mathrm{Gl}\,(k, \mathbb{C}).$$

Here, in contrast to the Definition 1.1.5 of a complex manifold, we neither require that the base M be complex nor that these transition maps be holomorphic. If, however, these conditions are satisfied, that is, M is a complex manifold and the transition maps are ho lomorphic, then we have a *holomorphic vector bundle* .

On a complex manifold M, in local holomorphic coordinates, we have the 1-forms

$$dz^j := dx^j + i dy^j,\ dz^{\bar{k}} := dx^j - i dy^j$$

(recall (1.1.2)). We can then decompose the space Ω^k of k-forms into subspaces $\Omega^{p,q}$ with $p+q=k$. Namely, $\Omega^{p,q}$ is locally spanned by forms of the type

$$\omega(z) = \eta(z) dz^{i_1} \wedge \ldots \wedge dz^{i_p} \wedge dz^{\bar{j}_1} \wedge \ldots \wedge dz^{\bar{j}_q}.$$

Thus

$$\Omega^k(M) = \sum_{p+q=k} \Omega^{p,q}(M). \tag{1.5.7}$$

We can then let the differential operators

$$\partial = \frac{1}{2}\left(\frac{\partial}{\partial x^j} - i\frac{\partial}{\partial y^j}\right)(dx^j + idy^j) \quad \text{and} \quad \bar{\partial} = \frac{1}{2}\left(\frac{\partial}{\partial x^j} + i\frac{\partial}{\partial y^j}\right)(dx^j - idy^j) \tag{1.5.8}$$

operate on such a form by

$$\partial\omega = \frac{\partial\eta}{\partial z^i} dz^i \wedge dz^{i_1} \wedge \ldots \wedge dz^{i_p} \wedge dz^{\bar{j}_1} \wedge \ldots \wedge dz^{\bar{j}_q} \tag{1.5.9}$$

and

$$\bar{\partial}\omega = \frac{\partial\eta}{\partial z^{\bar{j}}} dz^{\bar{j}} \wedge dz^{i_1} \wedge \ldots \wedge dz^{i_p} \wedge dz^{\bar{j}_1} \wedge \ldots \wedge dz^{\bar{j}_q}. \tag{1.5.10}$$

The following important relations link them with the exterior derivative d:

Lemma 1.5.4 *The exterior derivative d satisfies*

$$d = \partial + \bar{\partial}. \tag{1.5.11}$$

Moreover,

$$\partial\partial = 0,\ \bar{\partial}\bar{\partial} = 0, \tag{1.5.12}$$

$$\partial\bar{\partial} = -\bar{\partial}\partial\,. \tag{1.5.13}$$

Proof. We have

$$\begin{aligned}\partial + \bar{\partial} =& \frac{1}{2}\left(\frac{\partial}{\partial x^j} - i\frac{\partial}{\partial y^j}\right)(dx^j + idy^j) + \frac{1}{2}\left(\frac{\partial}{\partial x^j} + i\frac{\partial}{\partial y^j}\right)(dx^j - idy^j)\\ =& \frac{\partial}{\partial x^j} dx^j + \frac{\partial}{\partial y^j} dy^j = d.\end{aligned}$$

Therefore,

$$0 = d^2 = (\partial + \bar{\partial})(\partial + \bar{\partial}) = \partial^2 + \partial\bar{\partial} + \bar{\partial}\partial + \bar{\partial}^2$$

and decomposing this into types yields (1.5.12) and (1.5.13). One may verify these relations also by direct computation, e.g.

$$\partial\bar{\partial} = \frac{\partial^2}{\partial z^j \partial z^{\bar{k}}} dz^j \wedge dz^{\bar{k}} = -\frac{\partial^2}{\partial z^{\bar{k}} \partial z^j} dz^{\bar{k}} \wedge dz^j = -\bar{\partial}\partial.$$

□

/012/ / /3 / #! ? + "
+ @ 39 "8 5 + *"
' # :

2

CONFORMAL MAPPINGS

This chapter is intended as a gentle introduction to the basic notions of conformal geometry. In this way we are able to introduce and discuss some concepts from differential geometry which we will need later. We meet Möbius transformations and then prove Liouville's celebrated rigidity theorem of 1850, that conformal diffeomorphisms of $\mathbf{R}^n$, $n \geq 3$, are Möbius transformations. This proof serves as a model for our subsequent proof of this theorem in much greater generality.

2.1 The Cauchy–Riemann system

Let Ω be a domain in $\mathbf{R}^n$, $n \geq 2$. A diffeomorphism $f = (f^1, f^2, \ldots, f^n)\colon \Omega \to \mathbf{R}^n$ is said to be *conformal* at a point $a \in \Omega$ if f preserves the angle at a between any pair of smooth curves passing through a. That is the angle between the image curves at $f(a)$ is the same as that between the curves at a. The map f is conformal in Ω if it is conformal at each $a \in \Omega$. A classical result one meets in any first course in complex analysis, for example §2.3 in [6], is that a conformal orientation-preserving diffeomorphism of a planar domain is holomorphic in Ω. Equivalently, f has a complex derivative at a and satisfies the Cauchy–Riemann equations,

$$\begin{aligned}\frac{\partial f^1}{\partial x_1} &= \frac{\partial f^2}{\partial x_2}\\ \frac{\partial f^1}{\partial x_2} &= -\frac{\partial f^2}{\partial x_1}\end{aligned} \tag{2.1}$$

It is our purpose in this chapter to investigate similar equations in higher dimensions. We shall see a surprising rigidity phenomenon (the Liouville theorem) which lies at the heart of multidimensional geometric function theory.

Let us now find the governing equations for conformal diffeomorphisms. Thus let γ, η: $[-1, 1] \to \Omega$ be smooth curves normalized so that $\gamma(0) = \eta(0) = a$. The tangent vectors $\zeta = \frac{d}{dt}\gamma|_{t=0}$ and $\xi = \frac{d}{dt}\eta|_{t=0}$ determine the angle θ between the two curves γ and η by the rule

$$\cos(\theta) = \frac{\langle \zeta, \xi \rangle}{|\zeta||\xi|}$$

The diffeomorphism f preserves this angle, which means that the angle between the two curves $f \circ \gamma$ and $f \circ \eta$ at $f(a)$ remains unchanged. The chain rule enables us to compute the tangent vectors to these two curves at $f(a)$:

$$\frac{d}{dt}(f \circ \gamma)|_{t=0} = Df(a)\zeta \quad \text{and} \quad \frac{d}{dt}(f \circ \eta)|_{t=0} = Df(a)\xi$$

where here and in the sequel $Df(a)$ is the *Jacobian matrix* of f,

$$Df(a) = \left[\frac{\partial f^i}{\partial x_j}(a)\right] \quad i,\ j = 1, 2, \ldots, n \tag{2.2}$$

Hence f is conformal at a if

$$\frac{\langle Df(a)\zeta,\ Df(a)\xi\rangle}{|Df(a)\zeta||Df(a)\xi|} = \frac{\langle \zeta,\ \xi\rangle}{|\zeta||\xi|} \text{ for all } \zeta,\ \xi \in \mathbb{R}^n \setminus \{0\} \tag{2.3}$$

As a matter of fact we shall only exploit this identity for orthogonal vectors via the following lemma.

Lemma 2.1.1 *Let A be a nonsingular matrix such that for all vectors ζ and ξ with $\langle \zeta, \xi\rangle = 0$, we have $\langle A\zeta, A\xi\rangle = 0$. Then $A = |\det(A)|^{1/n} O$ for some orthogonal matrix O.*

Proof We have $\langle A^t A\zeta, \xi\rangle = 0$ whenever $\langle \zeta, \xi\rangle = 0$. Fixing ζ and varying ξ in the space perpendicular to ζ, we see $A^t A\zeta$ is parallel to ζ. That is, $A^t A\zeta = \lambda(\zeta)\zeta$ for some scalar function λ. The linearity of the left-hand side of this equation shows λ to be constant. As ζ is an arbitrary non-zero vector, we conclude that $A^t A = \lambda \mathbf{I}$. Then the constant λ is easily computed by taking the determinant of both sides. □

Returning to equation (2.3), we find the hypothesis of Lemma 2.1.1 fulfilled by $Df(a)$. Denote by $J(a, f)$ the determinant of the matrix $Df(a)$, which is referred to as the *Jacobian* of f.

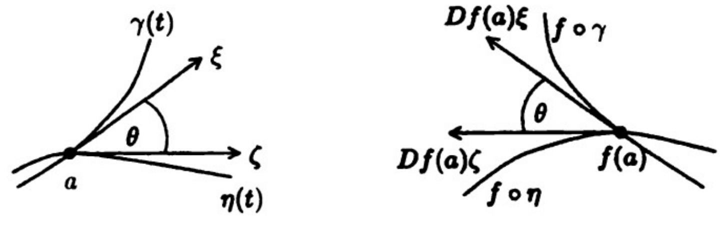

Figure 2.1 The angle between curves.

We have therefore established that an orientation-preserving conformal diffeomorphism of Ω satisfies the differential equation

$$D^t f(x) Df(x) = J(x, f)^{2/n}\mathbf{I} \quad x \in \Omega \tag{2.4}$$

It is a nice exercise to show that when $n = 2$ this system of equations reduces to the equations (2.1). It is for this reason that the system (2.4) is known in higher dimensions as the Cauchy–Riemann system.

A principal focus of this book is to study the solution to this and related equations. There are two distinctions between the higher-dimensional ($n \geq 3$) theory of these equations and the planar theory. The first is non-linearity (though strictly speaking the equations are only equivalent to a linear system in the plane) and the second is that the system is highly overdetermined. Indeed, the Cauchy–Riemann system in dimension n consists of $\frac{1}{2}(n-1)(n+2)$ independent equations for the n unknowns $f^1, f^2, \dots, f^n$ forming the components of f. That is partly why, for $n \geq 3$, the family of all solutions to this equation depends on only a finite number of real parameters. This number, as already recognized by Liouville in 1850 [204], is exactly equal to the dimension of the topological group $\mathbf{M\ddot{o}b}(n)$ of Möbius transformations of $\bar{\mathbb{R}}^n$. This dimension is $n(n+5)/2$.

2.2 The Möbius group

Let us look at some well-known solutions to the Cauchy–Riemann system. First, there are the obvious solutions $g(x) = \lambda O x$, where $\lambda \neq 0$ and O is an orthogonal transformation. Such mappings are known as *similarities*. Of course $Dg = \lambda O$. More interesting examples are furnished by reflections.

The *reflection* in a sphere $\mathbb{S}^{n-1}(a, r)$ is the mapping $g\colon \bar{\mathbb{R}}^n \to \bar{\mathbb{R}}^n$ defined by

$$g(x) = \begin{cases} a + \frac{r^2(x-a)}{|x-a|^2} & x \notin \{a, \infty\} \\ \infty & x = a \\ a & x = \infty \end{cases} \tag{2.5}$$

It is clear that $g \circ g = \mathbf{I}$ and $g(x) = x$ if and only if $x \in \mathbb{S}^{n-1}(a, r)$. For $x \notin \{a, \infty\}$ we compute the Jacobian matrix of g to be

$$Dg(x) = \frac{r^2}{|x-a|^2}\left(I - 2\frac{x-a}{|x-a|} \otimes \frac{x-a}{|x-a|}\right) \tag{2.6}$$

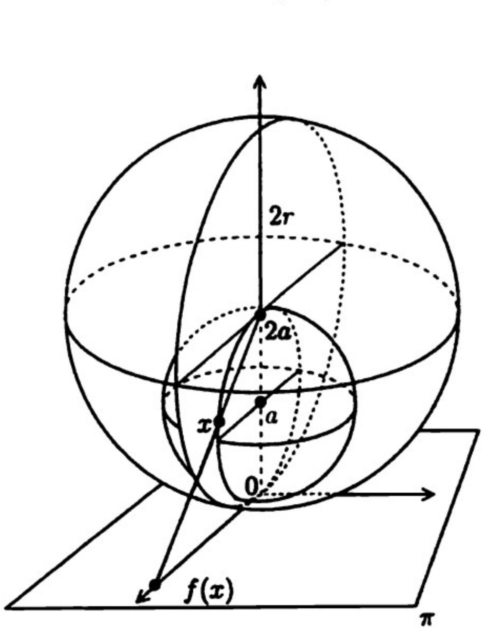

Figure 2.2 The stereographic projection of the sphere $\mathbb{S}^{n-1}(a, r)$, $r = |a|$, is the restriction of the inversion in $\mathbb{S}^{n-1}(2a, 2r)$.

In particular, the Jacobian determinant is negative:

$$J(x, g) = -r^{2n}|x - a|^{-2n} \tag{2.7}$$

and

$$D^t g(x) Dg(x) = \frac{r^4}{|x - a|^4}\mathbf{I} \tag{2.8}$$

Thus every reflection in a sphere is orientation-reversing and conformal. Recall that the reflection h in a hyperplane $\Pi = \{x \in \mathbb{R}^n : \langle x, a\rangle = t\}$ of $\mathbb{R}^n$ is an orientation-reversing isometry and that the Möbius group $\mathbf{M\ddot{o}b}(n)$ is generated by reflections in spheres and hyperplanes. Every reflection g satisfies the Cauchy–Riemann system (2.4) as is easy to see from (2.7) and (2.8). The following observation is a simple consequence of the chain rule:

Suppose f: $\Omega_1 \to \Omega_2$ and g: $\Omega_2 \to \Omega_3$ are C^1 solutions to the Cauchy–Riemann system. Then $f \circ g$: $\Omega_1 \to \Omega_3$ is also a C^1 solution. If in addition f is a diffeomorphism, then f^{-1}: $\Omega_2 \to \Omega_1$ is a solution as well.

Thus the space of diffeomorphic solutions to the Cauchy–Riemann system on domains in $\mathbb{R}^n$ form a pseudo-group which contains the Möbius transformations, since $\mathbf{M\ddot{o}b}(n)$ is generated by reflections. We now begin to study all solutions to the Cauchy–Riemann system.

2.3 The Liouville theorem (smooth case)

Later we will discuss very weak solutions of the differential equations (2.4). For now we will have to be satisfied with the following theorem, [162].

Theorem 2.3.1 *Every solution* $f \in C^3(\Omega, \mathbb{R}^n)$, $n \geq 3$, *to the Cauchy–Riemann system* $D^t f(x) Df(x) = |J(x, f)|^{2/n}\mathbf{I}$, *where* $J(x, f)$ *does not change sign in* Ω, *has the form*

$$f(x) = b + \frac{\alpha\, A(x-a)}{|x-a|^{\epsilon}} \tag{2.9}$$

In this formula the arbitrary parameters are $a \in \mathbb{R}^n \backslash \Omega$, $b \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$, A *is an orthogonal matrix, and* ϵ *is either* 0 *or* 2.

The idea of the proof is to differentiate the Cauchy–Riemann system twice to obtain an equation for the Jacobian determinant. It is instructive to place the computation within the more global framework of Riemannian geometry. We do this with a restricted degree of generality but with the more general system of equations

$$D^t f(x) Df(x) = G(x) \tag{2.10}$$

where $G(x) = [g_{ij}(x)]$ is a twice continuously differentiable function with values in the space of positive definite symmetric $n \times n$ matrices. The reader familiar with Riemannian geometry may recognize G as none other than the pullback of the Euclidean metric tensor (represented in standard coordinates) via the diffeomorphism f. We shall put off a precise discussion of this topic for the moment.

2.4 Curvature

Following the general philosophy of Cartan in dealing with overdetermined nonlinear systems, we want to eliminate f from the system of equations (2.10) by differentiation and substitution to obtain equations for G only. These of course turn out to be the curvatures of G. System (2.10) can be rewritten as

$$Df(x) G^{-1}(x) D^t f(x) = \mathbf{I} \tag{2.11}$$

where $G^{-1}(x) = [g^{ij}(x)]$ is the inverse of G.

Henceforth we follow Einstein's summation convention. Accordingly terms with repeated indices are to be summed from 1 *to* n.

In terms of the coordinate functions $f = (f^1, f^2, \ldots, f^n)$ both systems (2.10) and (2.11) can be written as

$$f_i^\alpha f_j^\alpha = g_{ij} \quad \text{or} \quad g^{\nu\mu} f_\nu^i f_\mu^j = \delta_{ij} \tag{2.12}$$

where δ_{ij} is the usual Kronecker symbol. Here we have reserved the use of lower indices on f to indicate partial differentiation with respect to a given variable, thus

$$Df = \left[\frac{\partial f^i}{\partial x_j}\right] = \left[f^i_j\right]$$

and with the further notation $f^i_{jk} = \frac{\partial^2 f^i}{\partial x_j \partial x_k}$ and so forth.

Differentiate the first system of equations at (2.12) with respect to the variable x_k and then permute the indices to obtain the three equations

$$\begin{aligned} f^\alpha_{ik} f^\alpha_j + f^\alpha_i f^\alpha_{jk} &= \frac{\partial g_{ij}}{\partial x_k} \\ f^\alpha_{ij} f^\alpha_k + f^\alpha_i f^\alpha_{jk} &= \frac{\partial g_{ik}}{\partial x_j} \\ -f^\alpha_{ij} f^\alpha_k - f^\alpha_j f^\alpha_{ik} &= -\frac{\partial g_{ij}}{\partial x_i} \end{aligned}$$

which we sum to obtain

$$2 f^\alpha_i f^\alpha_{jk} = \frac{\partial g_{ij}}{\partial x_k} + \frac{\partial g_{ik}}{\partial x_j} - \frac{\partial g_{jk}}{\partial x_i} \tag{2.13}$$

To compute the second-order derivatives of f we multiply by $f^l_\nu g^{\nu i}$ and use the second system of equations in (2.12) to find

$$f_{jk} = \Gamma^\nu_{jk} f_\nu \tag{2.14}$$

where Γ^ν_{jk} are the familiar *Christoffel symbols* of G, that is

$$2\Gamma^\nu_{jk} = g^{i\nu}\left(\frac{\partial g_{ij}}{\partial x_k} + \frac{\partial g_{ik}}{\partial x_j} - \frac{\partial g_{jk}}{\partial x_i}\right) \tag{2.15}$$

Notice that one immediate consequence of this is that if G is a constant matrix, then f is a linear transformation. In particular, conformal mappings with constant determinant are linear transformations, in fact scalar multiples of an orthogonal transformation. Next we differentiate equation (2.14) with respect to the variable x_l, $l = 1, 2, \ldots, n$, to obtain

$$\begin{aligned} f_{jkl} &= f_\nu \frac{\partial \Gamma^\nu_{jk}}{\partial x_l} + f_{\nu l} \Gamma^\nu_{jk} \\ &= \left(\frac{\partial \Gamma^\nu_{jk}}{\partial x_l} + \Gamma^\mu_{jk} \Gamma^\nu_{\mu l}\right) f_\nu \end{aligned}$$

where we have used (2.12) once again. If f is three times continuously differentiable, then f_{jkl} does not change with the permutation of indices. Permute k and l to see that

$$\left(\frac{\partial \Gamma^{\nu}_{jk}}{\partial x_l} - \frac{\partial \Gamma^{\nu}_{jl}}{\partial x_k} + \Gamma^{\mu}_{jk}\Gamma^{\nu}_{\mu l} - \Gamma^{\mu}_{jl}\Gamma^{\nu}_{\mu k}\right) f_{\nu} = 0 \tag{2.16}$$

Finally, we observe that the vectors f_{ν}, $\nu = 1, 2, \ldots, n$, are linearly independent. Hence

$$R^{\nu}_{jlk} = 0 \tag{2.17}$$

where R^{ν}_{jlk} is the *Riemannian curvature tensor*

$$R^{\nu}_{jlk} = \frac{\partial \Gamma^{\nu}_{jk}}{\partial x_l} - \frac{\partial \Gamma^{\nu}_{jl}}{\partial x_k} + \Gamma^{\mu}_{jk}\Gamma^{\nu}_{\mu l} - \Gamma^{\mu}_{jl}\Gamma^{\nu}_{\mu k} \tag{2.18}$$

We now recall two other important differential expressions from Riemannian geometry: the *Ricci curvature tensor*,

$$R_{ij} = R^{\nu}_{i\nu j} \tag{2.19}$$

and the *scalar curvature*

$$R = g^{\nu\mu} R_{\nu\mu} \tag{2.20}$$

The matrix function $G = G(x)$ is said to be *conformally Euclidean* if there is a scalar function $\lambda = \lambda(x)$ such that

$$g_{ij}(x) = e^{\lambda(x)}\delta_{ij} \tag{2.21}$$

The curvatures of a conformally Euclidean matrix are computed as follows. First the Christoffel symbols,

$$2\Gamma^{l}_{ij} = \lambda_i\delta_{jl} + \lambda_j\delta_{il} - \lambda_l\delta_{ij} \tag{2.22}$$

where the lower indices on λ denote partial differentiation. Next the curvature tensor

$$\begin{aligned} 4R^{l}_{ijk} = {} & 2\lambda_{ij}\delta_{kl} - 2\lambda_{lj}\delta_{ik} + 2\lambda_{lk}\delta_{ij} - 2\lambda_{ik}\delta_{jl} \\ & - \lambda_i\lambda_j\delta_{kl} + \lambda_l\lambda_j\delta_{ik} - \lambda_l\lambda_k\delta_{ij} + \lambda_i\lambda_k\delta_{jl} \\ & + \lambda_\nu\lambda_\nu\delta_{jk}\delta_{il} - \lambda_\nu\lambda_\nu\delta_{ik}\delta_{jl} \end{aligned}$$

and the Ricci tensor

$$4R_{ik} = 2(2-n)\lambda_{ik} - 2\lambda_{\nu\nu}\delta_{ik} + (n-2)\lambda_i\lambda_k - (n-2)\lambda_\nu\lambda_\nu\delta_{ik}$$

This yields the scalar curvature

$$\frac{4e^{\lambda}}{n-1}R = (2-n)|\nabla\lambda|^2 - 4\Delta\lambda \tag{2.23}$$

2.5 Computing the Jacobian

We are now in a position to compute the Jacobian determinant of any C^3 solution to the Cauchy–Riemann system. There is no loss of generality in assuming that $J(x, f) \geq 0$ in Ω for we may otherwise compose f with an orientation-reversing orthogonal transformation.

Lemma 2.5.1 *In each dimension $n \geq 3$ we have that either $J(x, f)$ is a constant function, or that*

$$J(x, f) = r^{2n}|x-a|^{-2n} \tag{2.24}$$

for some $r > 0$ and some $a \in \mathbb{R}^n \setminus \Omega$.

Proof Suppose that $J(x, f)$ is not constant. Fix one of the components of the open set $\{x \in \Omega : J(x, f) > 0\}$ and denote it by Ω^+. Our first aim is to show that formula (2.24) holds in Ω^+ and therefore that Ω^+ is a relatively closed subset of Ω, from which we will deduce $\Omega^+ = \Omega$. Let us set

$$\lambda(x) = \frac{2}{n}\log J(x, f) \quad x \in \Omega^+ \tag{2.25}$$

Let us now write the Cauchy–Riemann system as

$$D^t f(x) Df(x) = e^{\lambda(x)}\mathbf{I} \quad x \in \Omega^+$$

As the matrix function $G(x) = e^{\lambda(x)}\mathbf{I}$ is conformally flat its Ricci curvature vanishes. In particular, the scalar curvature $R = 0$. When $R = 0$ we see $4\Delta\lambda = (2-n)|\nabla\lambda|^2$ and we may simplify the formula for the Ricci tensor to

$$8R_{ij} = (n-2)\left[2\frac{\partial\lambda}{\partial x_i}\frac{\partial\lambda}{\partial x_j} - 4\frac{\partial^2\lambda}{\partial x_i\partial x_j} - |\nabla\lambda|^2\delta_{ij}\right] \tag{2.26}$$

Note the consequence that in two dimensions the conformally flat matrix $G = e^{\lambda}\mathbf{I}$ has vanishing scalar curvature if and only if λ is harmonic. Supposing now that $n \geq 3$, the assumption that G is Ricci flat, $R_{ij} = 0$, reads as

$$4\frac{\partial^2\lambda}{\partial x_i\partial x_j} = 2\frac{\partial\lambda}{\partial x_i}\frac{\partial\lambda}{\partial x_j} - |\nabla\lambda|^2\delta_{ij} \tag{2.27}$$

" /012/ / /3 / #! ? * ? *
" C " + @ 8 *"
3- = 5 " AAA % * AAA
? 5 " 65
? " 5 "
' (& &(> # * D *
' () & > # : D *

/012/ / /3 / #! = ? 0 + @
- " 8 5 + *"
' # ((= * *
E 0

4
Laplacian comparison theorem

Let M be a complete m-dimensional manifold. Suppose p is a fixed point in M and let us consider the distance function $r_p(x) = r(p, x)$ to p. When there is no ambiguity, the subscript will be deleted and we will simply write $r(x)$. The distance function in general is not smooth due to the presence of cut-points. However, it can be seen that it is a Lipschitz function with Lipschitz constant 1. In particular, we have

$$|\nabla r|^2 = 1$$

almost everywhere on M. Though r might not be a C^2 function, one can still estimate its Laplacian in the sense of distribution.

Theorem 4.1 *Let M be a complete m-dimensional Riemannian manifold with Ricci curvature bounded from below by*

$$\mathcal{R}_{ij} \geq (m-1)K$$

for some constant K. Then the Laplacian of the distance function satisfies

$$\Delta r(x) \leq \begin{cases} (m-1)\sqrt{K}\cot(\sqrt{K}r) & \text{for} \quad K > 0, \\ (m-1)r^{-1} & \text{for} \quad K = 0, \\ (m-1)\sqrt{-K}\coth(\sqrt{-K}r) & \text{for} \quad K < 0 \end{cases}$$

in the sense of distribution.

Proof For a smooth function f, let $x \in M$ be a point such that $\nabla f(x) \neq 0$. Then locally the level surface N of f through the point x is a smooth hypersurface. Let $\{e_1, \ldots, e_{m-1}\}$ be an orthonormal frame tangent to N, and let e_m be the unit normal vector to N. The Laplacian of f at x is defined to be

$$\begin{aligned}
\Delta f(x) &= \sum_{i=1}^{m-1}(e_i e_i - \nabla_{e_i} e_i) f(x) + (e_m e_m - \nabla_{e_m} e_m) f(x) \\
&= \sum_{i=1}^{m-1}(e_i e_i - (\nabla_{e_i} e_i)^t) f(x) \\
&\quad - \sum_{i=1}^{m-1}(\nabla_{e_i} e_i)^n f(x) + (e_m e_m - \nabla_{e_m} e_m) f(x) \\
&= \Delta_N f(x) + \overrightarrow{H} f(x) + (e_m e_m - \nabla_{e_m} e_m) f(x) \\
&= \overrightarrow{H} f(x) + (e_m e_m - \nabla_{e_m} e_m) f(x),
\end{aligned}$$

where Δ_N denotes the Laplacian of N with respect to the induced metric and $\overrightarrow{H}$ denotes the mean curvature vector of N. If we take $f = r$, then for a point x which is not in the cut-locus of p, we have $N = \partial B_p(r)$. Moreover, the unit normal vector $e_m = \partial/\partial r$. Hence, we have

$$\Delta r(x) = H(x),$$

where $H(x)$ is the mean curvature of $\partial B_p(r)$ with respect to $\partial/\partial r$. Therefore according to Corollary 2.2 the theorem is true for those points that are not in the cut-locus of p.

Using the same notation as in Theorem 2.1, let us denote

$$\bar{H}(r) = \begin{cases} (m-1)\sqrt{K}\cot(\sqrt{K}r) & \text{for} \quad K > 0, \\ (m-1)r^{-1} & \text{for} \quad K = 0, \\ (m-1)\sqrt{-K}\coth(\sqrt{-K}r) & \text{for} \quad K < 0. \end{cases}$$

To show that Δr has the desired estimate in the sense of distribution, it suffices to show that for any nonnegative compactly supported smooth function ϕ we have

$$\int_M (\Delta\phi)\, r \le \int_M \phi\, \bar{H}(r).$$

In terms of polar normal coordinates at p, we can write

$$\int_M \phi\, \bar{H}(r) = \int_0^\infty \int_{C(r)} \phi\, \bar{H}(r)\, J(\theta, r)\, d\theta\, dr.$$

On the other hand, for each $\theta \in S_p(M)$ if we let $R(\theta)$ be the maximum value of $r > 0$ such that the geodesic $\gamma(s) = \exp_p(s\theta)$ minimizes up to $s = r$, then by Fubini's theorem we can write

$$\int_0^\infty \int_{C(r)} \phi\, \bar{H}(r)\, J(\theta, r)\, d\theta\, dr = \int_{S_p(M)} \int_0^{R(\theta)} \phi\, \bar{H}(r)\, J(\theta, r)\, dr\, d\theta.$$

However, for $r < R(\theta)$, we have

$$\begin{aligned} \bar{H}(r)\, J(\theta, r) &\geq H(\theta, r)\, J(\theta, r) \\ &= \frac{\partial}{\partial r} J(\theta, r). \end{aligned}$$

Therefore,

$$\begin{aligned} \int_M \phi\, \bar{H}(r) &\geq \int_{S_p(M)} \int_0^{R(\theta)} \phi\, \frac{\partial J}{\partial r}\, dr\, d\theta \\ &= -\int_{S_p(M)} \int_0^{R(\theta)} \frac{\partial \phi}{\partial r}\, J\, dr\, d\theta + \int_{S_p(M)} [\phi\, J]_0^{R(\theta)}\, d\theta \\ &= -\int_M \frac{\partial \phi}{\partial r} + \int_{S_p(M)} \phi(\theta, R(\theta))\, J(\theta, R(\theta))\, d\theta \\ &\geq -\int_M \langle \nabla \phi, \nabla r \rangle, \end{aligned}$$

where we have used the facts that both ϕ and J are nonnegative and $J(\theta, 0) = 0$. On the other hand, since r is Lipschitz, we conclude that

$$-\int_M \langle \nabla \phi, \nabla r \rangle = \int_M (\Delta \phi) r,$$

which proves the theorem. □

We are now ready to prove a structural theorem for manifolds with nonnegative Ricci curvature. Let us first define the notions of a line and a ray in a Riemannian manifold.

Definition 4.2 A line is a normal geodesic $\gamma : (-\infty, \infty) \longrightarrow M$ such that any of its finite segments $\gamma|_a^b$ is a minimizing geodesic.

Definition 4.3 A ray is a half-line $\gamma_+ : [0, \infty) \longrightarrow M$, which is a normal minimizing geodesic.

Theorem 4.4 (Cheeger–Gromoll [CG2]) *Let M be a complete, m-dimensional, manifold with nonnegative Ricci curvature. If there exists a* *line in M, then M is isometric to $\mathbb{R} \times N$, the product of a real line and an $(m-1)$-dimensional manifold N with nonnegative Ricci curvature.*

Proof Let $\gamma_+ : [0, \infty) \longrightarrow M$ be a ray in M. One defines the Buseman function β_+ with respect to γ_+ by

$$\beta_+(x) = \lim_{t\to\infty} (t - r(\gamma_+(t), x)).$$

We observe that β_+ is a Lipschitz function with Lipschitz constant 1. Moreover, by the Laplacian comparison theorem,

$$\begin{aligned}\Delta\beta_+(x) &\geq -\lim_{t\to\infty} (m-1) r(\gamma_+(t), x)^{-1}\\ &= 0\end{aligned}$$

in the sense of distribution. If γ is a line, then $\gamma_+(t) = \gamma(t)$ and $\gamma_-(t) = \gamma(-t)$ for $t \geq 0$ are rays. The corresponding Buseman functions β_+ and β_- are subharmonic in the sense of distribution. In particular, $\beta_- + \beta_+$ is also subharmonic on M. On the other hand, since γ is a line, the triangle inequality implies that

$$\begin{aligned}2t &= r(\gamma(-t), \gamma(t))\\ &\leq r(\gamma(-t), x) + r(\gamma(t), x).\end{aligned}$$

Hence

$$t - r(\gamma_-(t), x) + t - r(\gamma_+(t), x) \leq 0,$$

and by taking the limit as $t \to \infty$ we have

$$\beta_-(x) + \beta_+(x) \leq 0.$$

Moreover, it is also clear that

$$\beta_-(x) + \beta_+(x) = 0$$

for all x on γ. However, by the strong maximum principle, since the subharmonic function $\beta_- + \beta_+$ has an interior maximum, it must be identically constant. In particular, both β_- and β_+ are harmonic and $\beta_- \equiv -\beta_+$. By regularity theory, β_+ is a smooth harmonic function with $|\nabla\beta_+| \leq 1$, and $|\nabla\beta_+| \equiv 1$ on the geodesic γ. For simplicity, let us write $\beta = \beta_+$. The Bochner formula gives

$$\begin{aligned}\Delta|\nabla\beta|^2 &= 2\beta_{ij}^2 + 2\mathcal{R}_{ij}\,\beta_i\,\beta_j + 2\langle\nabla\beta, \nabla\Delta\beta\rangle\\ &\geq 0. \qquad (4.1)\end{aligned}$$

Hence by the fact that $|\nabla\beta|^2$ achieves its maximum in the interior of M, the maximum principle for subharmonic functions again implies that $|\nabla\beta|^2 \equiv 1$ on M. Substituting this into (4.1) yields $\beta_{ij} \equiv 0$, and $\nabla\beta$ is a parallel vector field on M. This implies that M must split, which proves the theorem. □

Corollary 4.5 *Let M be a complete m-dimensional Riemannian manifold with nonnegative Ricci curvature. If M has at least two ends, then there exists a compact $(m-1)$-dimensional manifold N of nonnegative Ricci curvature such that*

$$M = \mathbb{R} \times N.$$

Proof The assumption that M has at least two ends implies that there exists a compact set $D \subset M$ such that $M \setminus D$ has at least two unbounded components. Hence there are two unbounded sequences of points $\{x_i\}_{i=1}^{\infty}$ and $\{y_i\}_{i=1}^{\infty}$, such that the minimal geodesics γ_i joining x_i to y_i must intersect D. By the compactness of D, if $p_i \in \gamma_i \cap D$ and $v_i = \gamma_i'(p_i)$, then by passing through a subsequence we have $p_i \to p$ and $v_i \to v$ for some $p \in D$ and $v \in T_pM$. We now claim that the geodesic $\gamma : (-\infty, \infty) \longrightarrow M$ given by the initial conditions $\gamma(0) = p$ and $\gamma'(0) = v$ is a line.

To see this, let us consider an arbitrary segment $\gamma|_{[a,b]}$ of γ. By the continuity of the initial conditions of the second order ordinary differential equation, we know that $\gamma_i|_{[a,b]} \to \gamma|_{[a,b]}$ because $(p_i, v_i) \to (p, v)$. However, by the assumption, $\gamma_i|_{[a,b]}$ are minimizing geodesics, hence $\gamma|_{[a,b]}$ is also minimizing, and γ is a line.

Theorem 4.4 now implies that $M = \mathbb{R} \times N$. The compactness of N follows from the assumption that M has at least two ends. □

Another application of the Laplacian comparison theorem is the eigenvalue comparison theorem of Cheng.

Theorem 4.6 (Cheng [Cg1]) *Let M be a compact Riemannian manifold of dimension m and ∂M be the boundary of M. Assume that the Ricci curvature of M is bounded by*

$$\mathcal{R}_{ij} \geq (m-1)K$$

for some constant K. Let us consider $\bar{M}$ to be the simply connected space form with constant curvature K. Let $\mu_1(M)$ be the first nonzero eigenvalue of the Dirichlet Laplacian on M and i be the inscribe radius of M. If $\bar{B}(i)$ is a *geodesic ball in $\bar{M}$ with radius i and $\mu_1(\bar{B}(i))$ is its first Dirichlet eigenvalue, then*

$$\mu_1(M) \leq \mu_1(\bar{B}(i)).$$

When $\partial M = \emptyset$, let $\lambda_1(M)$ be the first nonzero eigenvalue of the Laplacian and d be the diameter of M. If $\bar{B}(d/2)$ is a geodesic ball in $\bar{M}$ with radius $d/2$ and $\mu_1\left(\bar{B}\,(d/2)\right)$ is its first Dirichlet eigenvalue, then

$$\lambda_1(M) \leq \mu_1\left(\bar{B}\left(\frac{d}{2}\right)\right).$$

Proof Let us first consider the case when M has boundary. Let $B_p(i)$ be an inscribed ball in M. By the monotonicity of eigenvalues, it suffices to show that

$$\mu_1(B_p(i)) \leq \mu_1(\bar{B}(i)).$$

Let $\bar{u}$ be the first Dirichlet eigenfunction on $\bar{B}(i)$. By the uniqueness of $\bar{u}$, we may assume that $\bar{u} \geq 0$. If we let $\bar{p} \in \bar{B}(i)$ be its center and $\bar{r}$ be the distance function to $\bar{p}$, then $\bar{u}(\bar{r})$ must be a function of $\bar{r}$ alone. By the facts that $\bar{u} \geq 0$ and $\bar{u}(i) = 0$, the strong maximum principle implies that $(\partial\bar{u}/\partial\bar{r})(i) < 0$. If there is some value of $\bar{r} < i$ such that $\partial\bar{u}/\partial\bar{r} > 0$, then this would imply that $\bar{u}$ has a interior local minimum. However, this violates the strong maximum principle. Hence, $\bar{u}' = \partial\bar{u}/\partial\bar{r} \leq 0$.

Let us define a Lipschitz function on $B_p(i)$ by $u(r) = \bar{u}(r)$, where r denotes the distance function to p. Clearly, u satisfies the Dirichlet boundary condition. Computing the Laplacian of u gives

$$\begin{aligned}\Delta u &= \bar{u}'\Delta r + \bar{u}''|\nabla r|^2 \\ &\geq \bar{u}'\Delta r + \bar{u}''.\end{aligned}$$

On the other hand, if we let $\bar{\Delta}$ be the Laplacian on $\bar{M}$, then

$$\begin{aligned}-\mu_1(\bar{B}(i))\,\bar{u} &= \bar{\Delta}\bar{u} \\ &= \bar{u}'\bar{\Delta}\bar{r} + \bar{u}''.\end{aligned}$$

By Theorem 4.1 and the fact that $\bar{u}' \leq 0$, we conclude that

$$\Delta u \geq -\mu_1(\bar{B}(i))\,u$$

in the sense of distribution. Hence, by the Rayleigh principle for eigenvalues, we conclude that

$$\begin{aligned}\mu_1(B_p(i)) &\leq \frac{\int_{B_p(i)} |\nabla u|^2}{\int_{B_p(i)} u^2} \\ &= \frac{-\int_{B_p(i)} u\Delta u}{\int_{B_p(i)} u^2} \\ &\leq \mu_1(\bar{B}(i)).\end{aligned}$$

To prove the upper bound for the case when M has no boundary, we consider the disjoint balls $B_p(d/2)$ and $B_q(d/2)$ centered at a pair of points p and q which realize the diameter. By the above estimate,

$$\mu_1\left(B_p\left(\frac{d}{2}\right)\right) \leq \mu_1\left(\bar{B}\left(\frac{d}{2}\right)\right)$$

and

$$\mu_1\left(B_q\left(\frac{d}{2}\right)\right) \leq \mu_1\left(\bar{B}\left(\frac{d}{2}\right)\right).$$

We now claim that $\lambda_1(M) \leq \max\{\mu_1(M_1), \mu_1(M_2)\}$ for any disjoint pair of open subsets $M_1, M_2 \subset M$. This will establish the theorem by setting $M_1 = B_p(d/2)$ and $M_2 = B_q(d/2)$.

To prove the claim, let u_1 and u_2 be nonnegative first Dirichlet eigenfunctions on M_1 and M_2 respectively. By multiplying u_i by a constant, we may assume that $\int_{M_i} u_i = 1$ for $i = 1, 2$. Let us define a Lipschitz function on M by

$$u = \begin{cases} u_1 & \text{on} \quad M_1, \\ -u_2 & \text{on} \quad M_2, \\ 0 & \text{on} \quad M \setminus (M_1 \cup M_2). \end{cases}$$

Clearly, $\int_M u = 0$, hence after applying the Rayleigh principle, we have

$$\begin{aligned}\lambda_1(M)\left(\int_{M_1} u_1^2 + \int_{M_2} u_2^2\right) &= \lambda_1(M)\int_M u^2 \\ &\leq \int_M |\nabla u|^2 \\ &= \int_{M_1} |\nabla u_1|^2 + \int_{M_2} |\nabla u_2|^2\end{aligned}$$

$$= \mu_1(M_1) \int_{M_1} u_1^2 + \mu_1(M_2) \int_{M_2} u_2^2$$

$$\leq \max\{\mu_1(M_1), \mu_1(M_2)\} \left(\int_{M_1} u_1^2 + \int_{M_2} u_2^2 \right).$$

This establishes the claim. □

2

Volume comparison theorem

In this chapter, we will develop a volume comparison theorem originally proved by Bishop (see [BC]). Let $p \in M$ be a point in a complete Riemannian manifold of dimension m. In terms of polar normal coordinates at p, we can write the volume element as

$$J(\theta, r)dr \wedge d\theta,$$

where $d\theta$ is the area element of the unit $(m-1)$-sphere. The Gauss lemma asserts that the area element of submanifold $\partial B_p(r)$, which is the boundary of the geodesic ball of radius r, is given by $J(\theta, r)d\theta$. By the first and second variational formulas (1.5) and (1.6), if $x = (\theta, r)$ is not in the cut-locus of p, we have

$$\begin{aligned} J'(\theta, r) &= \frac{\partial J}{\partial r}(\theta, r) \\ &= H(\theta, r)\, J(\theta, r) \end{aligned} \tag{2.1}$$

and

$$\begin{aligned} J''(\theta, r) &= \frac{\partial^2 J}{\partial r^2}(\theta, r) \\ &= -\sum_{i,j=1}^{m-1} h_{ij}^2(\theta, r)\, J(\theta, r) - \mathcal{R}_{rr}(\theta, r)\, J(\theta, r) + H^2(\theta, r)\, J(\theta, r), \end{aligned} \tag{2.2}$$

where $\mathcal{R}_{rr} = \mathcal{R}(\partial/\partial r, \partial/\partial r)$, $H(\theta, r)$, and $(h_{ij}(\theta, r))$ denote the Ricci curvature in the radial direction, the mean curvature and the second fundamental form of $\partial B_p(r)$ at the point $x = (\theta, r)$ with respect to the unit normal vector $\partial/\partial r$, respectively.

Using the inequalities

$$\begin{aligned}\sum_{i,j=1}^{m-1} h_{ij}^2 &\geq \sum_{i=1}^{m-1} h_{ii}^2 \\ &\geq \frac{\left(\sum_{i=1}^{m-1} h_{ii}\right)^2}{m-1} \\ &= \frac{H^2}{m-1} \end{aligned} \tag{2.3}$$

and (2.1), we can estimate (2.2) by

$$\begin{aligned} J'' &\leq \frac{m-2}{m-1} H^2\, J - \mathcal{R}_{rr}\, J \\ &= \frac{m-2}{m-1} (J')^2\, J^{-1} - \mathcal{R}_{rr}\, J. \end{aligned} \tag{2.4}$$

Since any smooth metric is locally Euclidean, we have the initial conditions

$$J(\theta, r) \sim r^{m-1}$$

and

$$J'(\theta, r) \sim (m-1) r^{m-2}$$

as $r \to 0$. Let us point out that if M is a simply connected constant curvature space form with constant sectional curvature K, then all the above inequalities become equalities. In particular (2.4) becomes

$$J'' = \frac{m-2}{m-1} (J')^2\, J^{-1} - (m-1) K\, J.$$

Theorem 2.1 (Bishop [BC]) *Let M be a complete Riemannian manifold of dimension m, and p be a fixed point of M. Let us assume that the Ricci curvature tensor of M at any point x is bounded below by $(m-1)K(r(p,x))$ for some function K depending only on the distance from p. If $J(\theta, r)\, d\theta$ is the area element of $\partial B_p(r)$ as defined above and $\bar{J}(r)$ is the solution of the ordinary differential equation*

$$\bar{J}'' = \frac{m-2}{m-1} (\bar{J}')^2\, \bar{J}^{-1} - (m-1) K\, \bar{J}$$

with initial conditions

$$\bar{J}(r) \sim r^{m-1}$$

and

$$\bar{J}'(r) \sim (m-1)r^{m-2},$$

as $r \to 0$, *then within the cut-locus of* p *the function* $J(\theta, r)/\bar{J}(r)$ *is a nonincreasing function of* r. *Also, if* $\bar{H}(r) = \bar{J}'/\bar{J}$, *then* $H(\theta, r) \leq \bar{H}(r)$ *whenever* (θ, r) *is within the cut-locus of* p. *In particular, if* K *is a constant, then* $\bar{J}d\theta$ *corresponds to the area element of the sphere of radius* r *in the simply connected space form of constant curvature* K.

Proof By setting $f = J^{1/(m-1)}$, (2.1) and (2.4) can be written as

$$f' = \frac{1}{m-1} H\, f$$

and

$$\begin{aligned} f'' &\leq \frac{-1}{m-1} \mathcal{R}_{rr}\, f \\ &\leq -K\, f. \end{aligned}$$

The initial conditions become

$$f(\theta, 0) = 0$$

and

$$f'(\theta, 0) = 1.$$

Let $\bar{f} = \bar{J}^{1/(m-1)}$ be the corresponding function defined using $\bar{J}$. The function $\bar{f}$ satisfies

$$\begin{aligned} \bar{f}'' &= -K\, \bar{f}, \\ \bar{f}(0) &= 0, \end{aligned}$$

and

$$\bar{f}'(0) = 1.$$

Observe that when K is a constant, the function $\bar{f} > 0$ for all values of $r \in (0, \infty)$ when $K \leq 0$, and for $r \in (0, \pi/\sqrt{K})$ when $K > 0$. In general, $\bar{f} > 0$

on an interval $(0, a)$ for some $a > 0$. At those values of r we can define

$$F(\theta, r) = \frac{f(\theta, r)}{\bar{f}(r)}.$$

We have

$$F' = \bar{f}^{-2}\,(f'\,\bar{f} - f\,\bar{f}')$$

and

$$\begin{aligned} F'' &= \bar{f}^{-1}\,f'' - 2\bar{f}^{-2}\,f'\,\bar{f}' - \bar{f}^{-2}\,f\,\bar{f}'' + 2\bar{f}^{-3}\,f\,(\bar{f}')^2 \\ &\le -2\bar{f}^{-1}\,\bar{f}'\,F', \end{aligned}$$

hence

$$\begin{aligned} (\bar{f}^2 F')' &= \bar{f}^2(F'' + 2\bar{f}^{-1}\,\bar{f}'\,F') \\ &\le 0. \end{aligned}$$

Integrating from ϵ to r yields

$$\begin{aligned} F'(r) &\le F'(\epsilon)\,\bar{f}^2(\epsilon)\,\bar{f}^{-2}(r) \\ &= (\bar{f}(\epsilon)\,f'(\epsilon) - f(\epsilon)\,\bar{f}'(\epsilon))\,\bar{f}^{-2}(r). \end{aligned}$$

Letting $\epsilon \to 0$, the initial conditions of f and $\bar{f}$ imply that

$$F'(r) \le 0.$$

In particular, $\bar{f}\,f' - \bar{f}'\,f \le 0$, implying

$$H(\theta, r) \le \bar{H}(r).$$

Moreover, that F is a nonincreasing function of r implies that $J(\theta, r)/\bar{J}(r)$ is also a nonincreasing function of r. □

By computing the area element and the mean curvature of the constant curvature space form explicitly, we have the following corollary.

Corollary 2.2 *Under the assumption of Theorem 2.1, if K is a constant, then*

$$H \le \begin{cases} (m-1)\sqrt{K}\cot\left(\sqrt{K}r\right) & \textit{for} \quad K > 0, \\ (m-1)r^{-1} & \textit{for} \quad K = 0, \\ (m-1)\sqrt{-K}\coth\left(\sqrt{-K}r\right) & \textit{for} \quad K < 0, \end{cases}$$

and

$$\frac{J(\theta, r)}{\bar{J}(r)}$$

is a nonincreasing function of r, where

$$\bar{J}(r) = \begin{cases} \left(\frac{1}{\sqrt{K}}\right)^{m-1} \sin^{m-1}\left(\sqrt{K}r\right) & \text{for} \quad K > 0, \\ r^{m-1} & \text{for} \quad K = 0, \\ \left(\frac{1}{\sqrt{-K}}\right)^{m-1} \sinh^{m-1}\left(\sqrt{-K}r\right) & \text{for} \quad K < 0. \end{cases}$$

Let us take this opportunity to point out that this estimate implies that when $K > 0$, there must be a cut-point along any geodesic which has length $\pi/\sqrt{K}$. In particular, this proves Myers' theorem.

Corollary 2.3 (Myers) *Let M be an m-dimensional complete Riemannian manifold with Ricci curvature bounded from below by*

$$\mathcal{R}_{ij} \geq (m-1)K$$

for some constant $K > 0$. Then M must be compact with diameter d bounded from above by

$$d \leq \frac{\pi}{\sqrt{K}}.$$

Corollary 2.4 *Let M be an m-dimensional complete Riemannian manifold with Ricci curvature bounded from below by a constant $(m-1)K$. Suppose $\bar{M}$ is an m-dimensional simply connected space form with constant sectional curvature K. Let $A_p(r)$ be the area of the boundary of the geodesic ball $\partial B_p(r)$ centered at $p \in M$ of radius r and $\bar{A}(r)$ be the area of the boundary of a geodesic ball $\partial \bar{B}(r)$ of radius r in $\bar{M}$. Then for $0 \leq r_1 \leq r_2 < \infty$, we have*

$$A_p(r_1)\,\bar{A}(r_2) \geq A_p(r_2)\,\bar{A}(r_1). \tag{2.5}$$

If we let $V_p(r)$ and $\bar{V}(r)$ be the volumes of $B_p(r)$ and $\bar{B}(r)$, respectively, then for $0 \leq r_1 \leq r_2, r_3 \leq r_4 < \infty$ we have

$$\left(V_p(r_2) - V_p(r_1)\right)\left(\bar{V}(r_4) - \bar{V}(r_3)\right) \geq \left(V_p(r_4) - V_p(r_3)\right)\left(\bar{V}(r_2) - \bar{V}(r_1)\right). \tag{2.6}$$

Proof Let us define $C(r)$ to be a subset of the unit tangent sphere $S_p(M)$ at p such that for all $\theta \in C(r)$ the geodesic given by $\gamma(s) = \exp_p(s\theta)$ is minimizing up to $s = r$. Clearly for $r_1 \leq r_2$ we have $C(r_2) \subset C(r_1)$. By Theorem 2.1, we have

$$J(\theta, r_1)\, \bar{J}(r_2) \geq J(\theta, r_2)\, \bar{J}(r_1)$$

for $\theta \in C(r_2)$. Integrating over $C(r_2)$ yields

$$\begin{aligned} \int_{C(r_2)} J(\theta, r_1)\, d\theta\, \bar{J}(r_2) &\geq \int_{C(r_2)} J(\theta, r_2)\, d\theta\, \bar{J}(r_1) \\ &= A_p(r_2)\, \bar{J}(r_1). \end{aligned}$$

On the other hand,

$$\begin{aligned} A_p(r_1) &= \int_{C(r_1)} J(\theta, r_1)\, d\theta \\ &\geq \int_{C(r_2)} J(\theta, r_1)\, d\theta. \end{aligned}$$

Taking this together with the fact that

$$\bar{A}(r) = \alpha_{m-1}\, \bar{J}(r)$$

with α_{m-1} being the area of the unit $(m-1)$-sphere, we conclude (2.5).

To see (2.6), we first assume that $r_1 \leq r_2 \leq r_3 \leq r_4$ and we simply integrate the inequality

$$A_p(t_1)\, \bar{A}(t_2) \geq A_p(t_2)\, \bar{A}(t_1)$$

over $r_1 \leq t_1 \leq r_2$ and $r_3 \leq t_2 \leq r_4$. For the case when $r_1 \leq r_3 \leq r_2 \leq r_4$, we write

$$\begin{aligned} &\big(V_p\,(r_2) - V_p(r_1)\big)\big(\bar{V}(r_4) - \bar{V}(r_3)\big) \\ &\quad = \big(V_p(r_3) - V_p(r_1)\big)\big(\bar{V}(r_2) - \bar{V}(r_3)\big) \\ &\qquad + \big(V_p(r_3) - V_p(r_1)\big)\big(\bar{V}(r_4) - \bar{V}(r_2)\big) \\ &\qquad + \big(V_p(r_2) - V_p(r_3)\big)\big(\bar{V}(r_2) - \bar{V}(r_3)\big) \\ &\qquad + \big(V_p(r_2) - V_p(r_3)\big)\big(\bar{V}(r_4) - \bar{V}(r_2)\big) \end{aligned}$$

$$\begin{aligned}
&\geq \big(V_p(r_2) - V_p(r_3)\big)\big(\bar{V}(r_3) - \bar{V}(r_1)\big) \\
&\quad + \big(V_p(r_4) - V_p(r_2)\big)\big(\bar{V}(r_3) - \bar{V}(r_1)\big) \\
&\quad + \big(V_p(r_2) - V_p(r_3)\big)\big(\bar{V}(r_2) - \bar{V}(r_3)\big) \\
&\quad + \big(V_p(r_4) - V_p(r_2)\big)\big(\bar{V}(r_2) - \bar{V}(r_3)\big) \\
&= \big(V_p(r_4) - V_p(r_3)\big)\big(\bar{V}(r_2) - \bar{V}(r_1)\big).
\end{aligned} \qquad \square$$

Let us point out that equality in (2.6) holds if and only if $C(r_1) = C(r_4)$ and $J(\theta, r) = \bar{J}(r)$ for all $0 \leq r \leq r_4$ and $\theta \in C(r_1)$. In particular, if $r_1 = 0$, then $J(\theta, r) = \bar{J}(r)$ for all $r \leq r_4$ and $\theta \in S_p(M)$. This implies that $B_p(r_4)$ is isometric to $\bar{B}(r_4)$.

Theorem 2.5 *Let M be an m-dimensional complete Riemannian manifold with nonnegative Ricci curvature. Then the volume growth of M must satisfy the following estimates:*

(1) *(Bishop) If α_{m-1} is the area of the unit $(m-1)$-sphere, then*

$$V_p(\rho) \leq \frac{\alpha_{m-1}}{m}\,\rho^m$$

for all $p \in M$ and $\rho \geq 0$.

(2) *(Yau [Y2]) For all $p \in M$, there exists a constant $C(m) > 0$ depending only on m, such that*

$$V_p(\rho) \geq C\, V_p(1)\, \rho$$

for all $\rho > 2$.

Proof Applying (2.6) to $r_1 = 0 = r_3$ and $r_4 = r$, we have

$$V_p(r_2)\,\bar{V}(r) \geq V_p(r)\,\bar{V}(r_2).$$

Observing that

$$\lim_{r_2 \to 0} \frac{V_p(r_2)}{\bar{V}(r_2)} = 1,$$

the upper bound follows.

To prove the lower bound, let $x \in \partial B_p(1+\rho)$. Then (2.6) and the curvature assumption imply that

$$V_x(2+\rho) - V_x(\rho) \leq V_x(\rho)\,\frac{(2+\rho)^m - \rho^m}{\rho^m}. \tag{2.7}$$

However, since the distance between p and x is $r(p,x)=1+\rho$, we have $B_p(1)\subset(B_x(2+\rho)\setminus B_x(\rho))$, hence

$$V_p(1)\le V_x(2+\rho)-V_x(\rho). \tag{2.8}$$

Since $B_x(\rho)\subset B_p(1+2\rho)$, we have

$$V_x(\rho)\le V_p(1+2\rho),$$

therefore combining this with (2.7) and (2.8), we conclude that

$$V_p(1)\le V_p(1+2\rho)\,\frac{(2+\rho)^m-\rho^m}{\rho^m}.$$

The lower bounded follows by observing that

$$\frac{(2+\rho)^m-\rho^m}{\rho^m}=C\,\rho^{-1}$$

for $\frac{1}{2}\rho\to\infty$. □

We would like to remark that if we assume that for a sufficiently small $\epsilon>0$ the Ricci curvature has a lower bound of the form

$$\mathcal{R}_{ij}(x)\ge-\epsilon(1+r(p,x))^{-2},$$

then one can show that M must have infinite volume. On the other hand, if the Ricci curvature is bounded from below by

$$\mathcal{R}_{ij}(x)\ge-C_0(1+r(p,x))^{-2-\delta}$$

for some constants $C_0,\delta>0$, then the upper bound is also valid and is of the form

$$V_p(r)\le Cr^m,$$

where $C(C_0,\delta,m)>0$ is a constant depending on C_0, δ and m.

It is also a good exercise to show that if a complete manifold has Ricci curvature bounded from below by

$$\mathcal{R}_{ij}\ge\epsilon r(p,x)^{-2}$$

for some constant $\epsilon>\frac{1}{4}$ and for all $r>1$, then M must be compact.

The next theorem shows that when the upper bound of the diameter given by Myers' theorem is achieved, then the manifold must be isometrically a sphere.

Theorem 2.6 (Cheng [Cg1]) *Let M be a complete m-dimensional Riemannian manifold with Ricci curvature bounded from below by*

$$\mathcal{R}_{ij} \geq (m-1)K$$

for some constant $K > 0$. If the diameter d of M satisfies

$$d = \frac{\pi}{\sqrt{K}},$$

then M is isometric to the standard sphere of radius $1/\sqrt{K}$.

Proof By scaling, we may assume that $K = 1$. Let p and q be a pair of points in M which realize the diameter. The volume comparison theorem implies that

$$V_p(d) \leq V_p\left(\frac{d}{2}\right) \frac{\bar{V}(d)}{\bar{V}\,(d/2)}.$$

The assumption that $d = \pi/\sqrt{K}$ implies that

$$\frac{\bar{V}(d)}{\bar{V}\,(d/2)} = 2,$$

hence

$$V_p(d) \leq 2V_p\left(\frac{d}{2}\right).$$

Similarly, we have

$$V_q(d) \leq 2V_q\left(\frac{d}{2}\right).$$

However, by the triangle inequality and the fact that $d = r(p,q)$, we have $B_p(d/2) \cap B_q(d/2) = \emptyset$. Therefore,

$$\begin{aligned} 2V(M) &= V_p(d) + V_q(d) \\ &\leq 2\left(V_p\left(\frac{d}{2}\right) + V_q\left(\frac{d}{2}\right)\right) \\ &\leq 2V(M), \end{aligned}$$

where $V(M)$ denotes the volume of M. This implies that the inequalities in the volume comparison theorem are in fact equalities. Hence by the remark following Corollary 2.4 M must be the standard sphere. □

: /012/ / /3 / #! 1 ; F "- .
* 9 + ? + @ 7 F " *"
' , 2 ** 6 " *

? /012/ / /3 / #! 5 ? ; G ?
0 + H ; " #@ * *"
' , * &
5 H

' &)(> H

F /012/ / /3 / #! I8 5 9 J K F +-
* 9 ? +< 0 % " (@ * *"

' ! (> E 5 *

2.4 The Projective Space

The aim of this section is to present certain explicit calculations for the projective space $\mathbb{P}^n$. We shall describe sections of the holomorphic line bundles $\mathcal{O}(k)$ on $\mathbb{P}^n$ and show that the canonical bundle $K_{\mathbb{P}^n}$ of the projective space $\mathbb{P}^n$ is isomorphic to the line bundle $\mathcal{O}(-n-1)$. It will turn out that everything can be expressed in polynomials, which turns complex geometry for $\mathbb{P}^n$ into complex algebraic geometry.

In order to determine all holomorphic sections of $\mathcal{O}(k)$, let us first recall that $\mathcal{O}(-1)$ is the holomorphic line bundle which is naturally embedded into the trivial vector bundle $\mathcal{O}^{\oplus n+1}$ such that the fibre of $\mathcal{O}(-1)$ over $\ell \in \mathbb{P}^n$ is $\ell \subset \mathbb{C}^{n+1}$ (cf. Proposition 2.2.6).

We will also use the following observation. If E is a holomorphic subbundle of F, then the inclusion $E \subset F$ naturally defines an inclusion of all tensor products $E^{\otimes k} \subset F^{\otimes k}$. In particular, for $k > 0$ the line bundle $\mathcal{O}(-k)$ is a subbundle of $\mathbb{P}^n \times (\mathbb{C}^{n+1})^{\otimes k}$.

Any homogeneous polynomial $s \in \mathbb{C}[z_0, \ldots, z_n]_k$ of degree k defines a linear map $(\mathbb{C}^{n+1})^{\otimes k} \to \mathbb{C}$. This gives rise to a holomorphic map $\mathbb{P}^n \times (\mathbb{C}^{n+1})^{\otimes k} \to \mathbb{C}$ which is linear on any fibre of the projection to $\mathbb{P}^n$. The restriction to $\mathcal{O}(-k)$ thus provides a holomorphic section of $\mathcal{O}(k)$. This way we associate to any homogeneous polynomial s of degree k a global holomorphic section of $\mathcal{O}(k)$, which will also be called s (cf. Exercise 2.2.8).

Proposition 2.4.1 *For $k \geq 0$ the space $H^0(\mathbb{P}^n, \mathcal{O}(k))$ is canonically isomorphic to the space $\mathbb{C}[z_0, \ldots, z_n]_k$ of all homogeneous polynomials of degree k.*

Proof. Note that the section associated to a non-trivial polynomial $0 \neq s \in \mathbb{C}[z_0, \ldots, z_n]_k$ is not trivial. Indeed, the composition $\mathcal{O}(-1) \subset \mathbb{P}^n \times \mathbb{C}^{n+1} \to \mathbb{C}^{n+1}$ is surjective and, therefore, any polynomial s that induces the trivial section of $\mathcal{O}(k)$ defines a trivial map $\mathbb{C}^{n+1} \to \mathbb{C}$, i.e. $s = 0$. Since the map $\mathbb{C}[z_0, \ldots, z_n]_k \to H^0(\mathbb{P}^n, \mathcal{O}(k))$ is clearly linear, it suffices to show surjectivity.

Let $0 \neq t \in H^0(\mathbb{P}^n, \mathcal{O}(k))$. Choose a non-trivial $0 \neq s_0 \in H^0(\mathbb{P}^n, \mathcal{O}(k))$ induced by a homogeneous polynomial of degree k and consider the meromorphic function $F := t/s_0 \in K(\mathbb{P}^n)$ (cf. Remark 2.3.24). Composing with the projection $\mathbb{C}^{n+1} \setminus \{0\} \to \mathbb{P}^n$ yields a meromorphic map $\tilde{F} \in K(\mathbb{C}^{n+1} \setminus \{0\})$. Moreover, $G := s_0 \cdot \tilde{F}$ is a holomorphic function $\mathbb{C}^{n+1} \setminus \{0\} \to \mathbb{C}$, which, due to Hartogs' theorem 1.1.4, can be extended to a holomorphic function $G : \mathbb{C}^{n+1} \to \mathbb{C}$. By definition of $\tilde{F}$ and G, this function is homogeneous of degree k, i.e. $G(\lambda \cdot z) = \lambda^k \cdot G(z)$ for $\lambda \in \mathbb{C}$ and $z \in \mathbb{C}^{n+1}$.

Using the power series expansion of G in $z = 0$ this proves that G is a homogeneous polynomial s of degree k. Clearly, the section of $\mathcal{O}(k)$ induced by G equals $t \in H^0(\mathbb{P}^n, \mathcal{O}(k))$. □

It should be obvious from the proof that the isomorphism described above is compatible with the multiplicative structure, i.e. we obtain a ring isomorphism

$$R(\mathbb{P}^n, \mathcal{O}(1)) = \bigoplus_{k\geq 0} H^0(\mathbb{P}^n, \mathcal{O}(k)) \cong \mathbb{C}[z_0, \ldots, z_n].$$

Applying Exercise 2.2.5 the proposition also yields

Corollary 2.4.2 *For $k < 0$ the line bundle $\mathcal{O}(k)$ admits no global holomorphic sections, i.e. $H^0(\mathbb{P}^n, \mathcal{O}(k)) = 0$.* □

The higher cohomology groups of $\mathcal{O}(k)$ will be computed in Example 5.2.5 by applying the Kodaira vanishing theorem.

Let us next construct an isomorphism $K_{\mathbb{P}^n} \cong \mathcal{O}(-n-1)$. This can be done by a direct cocycle calculation or by using the Euler sequence. We shall discuss both. Using the adjunction formula, the isomorphism $K_{\mathbb{P}^n} \cong \mathcal{O}(-n-1)$ will enable us to compute the Kodaira dimension of hypersurfaces and complete intersections.

First, let us recall the standard open covering of $\mathbb{P}^n$ (see page 56). Let U_i be the open set

$$U_i := \{(z_0 : \ldots : z_n) \mid z_i \neq 0\} \subset \mathbb{P}^n.$$

Then $\mathbb{P}^n = \bigcup_{i=0}^n U_i$ and the maps

$$\varphi_i : U_i \longrightarrow \mathbb{C}^n, \quad (z_0 : \ldots : z_n) \longmapsto \left(\tfrac{z_0}{z_i}, \ldots, \widehat{\tfrac{z_i}{z_i}}, \ldots, \tfrac{z_n}{z_i}\right)$$

are biholomorphic. For $i > j$ the transition map

$$\varphi_{ij} = \varphi_i \circ \varphi_j{}^{-1} : \{(w_1, \ldots, w_n) \mid w_i \neq 0\} \longmapsto \{(w_1, \ldots, w_n) \mid w_j \neq 0\}$$

is the composition of

$$\tilde{\varphi}_{ij}(w_1, \ldots, w_n) = (w_i^{-1}w_1, \ldots, w_i^{-1}w_{i-1}, w_i^{-1}, w_i^{-1}w_{i+1}, \ldots, w_i^{-1}w_n)$$

and the permutation $(j+1, i)$, the signature of which is $(-1)^{i-j-1}$.

Proposition 2.4.3 *The canonical bundle $K_{\mathbb{P}^n}$ is isomorphic to $\mathcal{O}(-n-1)$.*

Proof. By definition $K_{\mathbb{P}^n} \cong \det(\mathcal{T}^*_{\mathbb{P}^n})$, where the holomorphic tangent bundle $\mathcal{T}_{\mathbb{P}^n}$ is uniquely determined by its cocycle $\{J(\varphi_{ij}) \circ \varphi_j : U_i \cap U_j \to \mathrm{Gl}(n, \mathbb{C})\}$. Thus, it suffices to show that the line bundle associated to the cocycle $\{\det(J(\varphi_{ij})) \circ \varphi_j\}$ is isomorphic to $\mathcal{O}(n+1)$. Using the above remark, we have

$$\det(J(\varphi_{ij})) = (-1)^{i-j-1}\det\left(\frac{\partial \tilde{\varphi}^k_{ij}}{\partial w_\ell}\right)_{k,\ell},$$

where $\tilde{\varphi}_{ij} = (\tilde{\varphi}^1_{ij}, \ldots, \tilde{\varphi}^n_{ij})$.

The matrix $\left(\frac{\partial\bar{\varphi}_{ij}^k}{\partial w_\ell}\right)_{k,\ell}$ is easily identified as

$$\begin{pmatrix}
w_i^{-1} & 0 & \cdot\,\cdot & 0 & -\frac{w_1}{w_i^2} & 0 & \cdot\quad\cdot & 0 \\
0 & w_i^{-1} & \cdot\,\cdot & 0 & -\frac{w_2}{w_i^2} & 0 & \cdot\quad\cdot & 0 \\
\cdot & & & & & & & \cdot \\
\cdot & & & & & & & \cdot \\
0 & \cdot & \cdot\,\cdot & w_i^{-1} & -\frac{w_{i-1}}{w_i^2} & 0 & \cdot\quad\cdot & 0 \\
 & & & & -\frac{1}{w_i^{-2}} & 0 & \cdot\quad\cdot & 0 \\
0 & \cdot & \cdot\,\cdot & 0 & -\frac{w_{i+1}}{w_i^2} & w_i^{-1} & \cdot\quad\cdot & 0 \\
 & & & & & & & \\
0 & \cdot & \cdot\,\cdot & 0 & -\frac{w_n}{w_i^2} & 0 & \cdot\quad\cdot & w_i^{-1}
\end{pmatrix}$$

Hence, $\det\left(\frac{\partial\bar{\varphi}_{ij}^k}{\partial w_\ell}\right) = -w_i^{-(n+1)}$ and, therefore,

$$\det(J(\varphi_{ij}))\circ\varphi_j = (-1)^{i-j}w_i^{-(n+1)}\circ\varphi_j = (-1)^{i-j}(z_i/z_j)^{-(n+1)}.$$

On the other hand, $\mathcal{O}(-1)$ is the line bundle $\{(\ell, z) \mid z \in \ell\} \subset \mathbb{P}^n \times \mathbb{C}^{n+1}$, to which is associated the cocycle $\{z_i/z_j\}$ (cf. Proposition 2.2.6). Thus, $\mathcal{O}(n+1)$ corresponds to the cocycle $\{(z_j/z_i)^{n+1}\}$. The two cocyles $\{(z_j/z_i)^{n+1}\}$ and $\{(-1)^{i-j}(z_i/z_j)^{-(n+1)}\}$ define isomorphic line bundles, for they only differ by the boundary of the cocycle $\{\eta_i \equiv (-1)^i : U_i \to \mathbb{C}^* \subset \mathcal{O}^*\}$. □

Recall that any short exact sequence $0 \to E \to F \to G \to 0$ of holomorphic vector bundles (or locally free sheaves), induces a canonical isomorphism $\det(F) \cong \det(E)\otimes\det(G)$ (cf. viii) of Example 2.2.4). Thus, the above proposition is also a consequence of the following one.

Proposition 2.4.4 (Euler sequence) *On $\mathbb{P}^n$ there exists a natural short exact sequence of holomorphic vector bundles*

$$\begin{tikzcd} 0 \arrow[r] & \mathcal{O} \arrow[r] & \bigoplus_{j=0}^n \mathcal{O}(1) \arrow[r] & \mathcal{T}_{\mathbb{P}^n} \arrow[r] & 0. \end{tikzcd}$$

Proof. The inclusion $\mathcal{O} \subset \bigoplus_{j=0}^n \mathcal{O}(1)$ is obtained from the natural inclusion $\mathcal{O}(-1) \subset \bigoplus_{j=0}^n \mathcal{O}$ (cf. Proposition 2.2.6) by twisting with $\mathcal{O}(1)$ (cf. Exercise 2.2.4). It suffices to show that the kernel of the dual map $\bigoplus_{j=0}^n \mathcal{O}(-1) \to \mathcal{O}$ is canonically isomorphic to $\Omega_{\mathbb{P}^n}$. (Here we use that a sequence of holomorphic vector bundles $0 \to E \to F \to G \to 0$ is exact if and only if $0 \to G^* \to F^* \to E^* \to 0$ is exact.)

The proof of the identification of $\Omega_{\mathbb{P}^n}$ with the kernel is a little tedious. Using the standard trivialization of $\bigoplus_{j=0}^n \mathcal{O}(-1)|_{U_i}$ we will identify $\Omega_{\mathbb{P}^n}|_{U_i}$ in a natural way with the kernel of the surjection $\bigoplus_{j=0}^{n}\mathcal{O}(-1)\to\mathcal{O}$ over U_i. The proof concludes by showing that these inclusions glue.

First note that for any local holomorphic function $f: U\to\mathbb{C}$, where $U\subset X$ is an open subset of an arbitrary manifold X the differential $df=\partial f$ is a section of $\Omega_X|_U$ (cf. Exercise 2.2.10 or Section 2.6). Moreover, if $\varphi: U_i\cong\varphi(U_i)\subset\mathbb{C}^n$ is a local coordinate system, then $\Omega_X|_U=\bigoplus_{k=1}^{n}(d\varphi^k)\cdot\mathcal{O}_U$. In our situation $\Omega_{\mathbb{P}^n}|_{U_i}=\bigoplus_{i\neq k=0}^{n}d\left(\frac{z_k}{z_i}\right)\cdot\mathcal{O}_{U_i}$.

Since $\mathcal{O}(-1)|_{U_i}$ is canonically trivialized by $(\ell,z)\mapsto z_i$, one has a natural diagonal trivialization $\psi_i(\ell,z^0,\ldots,z^n)=(\ell,z_i^0,\ldots,z_i^n)$ of $\bigoplus_{k=0}^{n}\mathcal{O}(-1)|_{U_i}$. In this trivialization the surjection $\varphi:\bigoplus\mathcal{O}(-1)\to\mathcal{O}$ is the map

$$\varphi_i(\ell,w_0,\ldots,w_n)=\sum_{k=0}^{n}w_k\frac{z_k}{z_i}.$$

Hence, the map

$$\phi_i:\Omega_{\mathbb{P}^n}|_{U_i}=\bigoplus_{k=0}^{n}d\left(\tfrac{z_k}{z_i}\right)\cdot\mathcal{O}_{U_i}\longrightarrow\bigoplus_{k=0}^{n}\mathcal{O}_{U_i},\quad d\left(\tfrac{z_k}{z_i}\right)\longmapsto e_k-\tfrac{z_k}{z_i}e_i,$$

where e_k is the k-th unit vector in $\bigoplus_{k=0}^{n}\mathcal{O}$, identifies $\Omega_{\mathbb{P}^n}|_{U_i}$ with the kernel of φ_i.

Let us now check that $\phi_i(d\left(\frac{z_k}{z_i}\right))=(\psi_i\psi_j^{-1})(\phi_j(d\left(\frac{z_k}{z_i}\right)))$. Here we use

$$d\left(\frac{z_k}{z_i}\right)=d\left(\frac{z_k}{z_j}\right)\frac{z_j}{z_i}-\frac{z_k}{z_i}\frac{z_j}{z_i}d\left(\frac{z_i}{z_j}\right).$$

Thus,

$$\phi_j\left(d\left(\frac{z_k}{z_i}\right)\right)=\frac{z_j}{z_i}\left(e_k-\frac{z_k}{z_j}e_j\right)-\frac{z_k}{z_i}\frac{z_j}{z_i}\left(e_i-\frac{z_i}{z_j}e_j\right)=\frac{z_j}{z_i}\left(e_k-\frac{z_k}{z_i}e_i\right).$$

Since $\psi_i\psi_j^{-1}$ is given by multiplication with $\frac{z_i}{z_j}$, this concludes the proof. ☐

Remark 2.4.5 The dual of the Euler sequence twisted by $\mathcal{O}(1)$ takes the form

$$0\longrightarrow\Omega_{\mathbb{P}^n}(1)\longrightarrow\bigoplus_{j=0}^{n}\mathcal{O}\longrightarrow\mathcal{O}(1)\longrightarrow 0. \tag{2.1}$$

Let us think of $\mathbb{P}^n$ more invariantly as $\mathbb{P}(V)$ with V a complex vector space of dimension $n+1$. In particular, V^* is naturally identified with the space of homogeneous linear forms on V and, therefore, $V^*=H^0(\mathbb{P}(V),\mathcal{O}(1))$. The exact sequence (2.1) can in this context be written as

$$0\longrightarrow\Omega_{\mathbb{P}(V)}(1)\longrightarrow V^*\otimes\mathcal{O}\longrightarrow\mathcal{O}(1)\longrightarrow 0,$$

where the surjection $V^*\otimes\mathcal{O}\to\mathcal{O}(1)$ is the evaluation map.

The advantage of writing the Euler sequence coordinate free is that it lends itself to important generalizations in the relative context, but we only mention this in passing:

If E is a holomorphic vector bundles over X, any fibre $\pi^{-1}(x)$ of the projection $\pi : \mathbb{P}(E) \to X$ is a complex projective space $\mathbb{P}(E(x))$ which comes with the Euler sequence

$$\begin{tikzcd} 0 \arrow[r] & \Omega_{\mathbb{P}(E(x))}(1) \arrow[r] & E(x)^* \otimes \mathcal{O} \arrow[r] & \mathcal{O}(1) \arrow[r] & 0. \end{tikzcd}$$

It turns out that these sequences can be glued to the *relative Euler sequence*

$$\begin{tikzcd} 0 \arrow[r] & \Omega_\pi \otimes \mathcal{O}_\pi(1) \arrow[r] & \pi^* E^* \arrow[r] & \mathcal{O}_\pi(1) \arrow[r] & 0, \end{tikzcd}$$

where $\mathcal{O}_\pi(1)$ is the (dual of the) relative tautological bundle (see Exercise 2.4.9) and $\mathcal{T}_\pi = \Omega_\pi^*$ is the relative tangent bundle defined as the kernel of $\mathcal{T}_{\mathbb{P}(E)} \to \pi^*\mathcal{T}_X$.

Corollary 2.4.6 *One has* $\mathrm{kod}(\mathbb{P}^n) = -\infty$.

Proof. Since the line bundles $\mathcal{O}(d)$ have global sections for $d \geq 0$, none of the powers $K_{\mathbb{P}^n}^{\otimes m} = \mathcal{O}(-m(n+1))$ for $m > 0$ admits non-trivial global sections (Corollary 2.4.2). Thus, the canonical ring $R(\mathbb{P}^n)$ is isomorphic to $\mathbb{C}$. □

This yields in particular an example of a complex compact manifold with $\mathrm{kod}(X) < a(X)$. Other examples are provided by projective complex tori.

In order to compute the canonical bundle of hypersurfaces in $\mathbb{P}^n$, or more generally of complete intersections, we will need one more result. The following works for any complex manifold Y and not just the projective space.

Proposition 2.4.7 *Let X be a smooth hypersurface of a complex manifold Y defined by a section $s \in H^0(Y, L)$ of some holomorphic line bundle L on Y. Then $\mathcal{N}_{X/Y} \cong L|_X$ and thus $K_X \cong (K_Y \otimes L)|_X$.*

Proof. First observe the following fact: Let $\varphi = (\varphi^1, \ldots, \varphi^n)$ be a holomorphic map $U \to \mathbb{C}^n$ of some connected open neighbourhood $U \subset \mathbb{C}^n$ of the origin, such that $\varphi^n(z_1, \ldots, z_{n-1}, 0) = 0$ for all $(z_1, \ldots, z_{n-1}, 0) \in U$. Then $\varphi^n(z) = z_n \cdot h(z_1, \ldots, z_n)$, where $h(z_1, \ldots, z_n)$ is a power series in $z_1, \ldots, z_n$. Hence,

$$\frac{\partial \varphi^n}{\partial z_k}(z_1, \ldots, z_{n-1}, 0) = \begin{cases} 0 & \text{for } k = 1, \ldots, n-1 \\ h(z_1, \ldots z_{n-1}, 0) & \text{for } k = n \end{cases}$$

Let us apply this to our situation. We may fix local coordinates $\varphi_i : U_i \cong \varphi_i(U_i) \subset \mathbb{C}^n$, such that $\varphi_i(U_i \cap X) = \{(z_1, \ldots, z_n) \in \varphi_i(U_i) \mid z_n = 0\}$. Thus, the transition maps $\varphi_{ij} : \varphi_j(U_i \cap U_j) \cong \varphi_i(U_i \cap U_j)$ have the above property.

This yields

$$J(\varphi_{ij})|_X = \begin{pmatrix} J(\varphi_{ij}|_X) & * \\ 0 & \frac{\partial \varphi_{ij}^n}{\partial z_n}|_X \end{pmatrix}$$

By Proposition 2.2.17, the normal bundle $\mathcal{N}_{X/Y}$ corresponds to the cocycle $\{\frac{\partial \varphi_{ij}^n}{\partial z_n} \circ \varphi_j|_X\}$.

On the other hand, due to Proposition 2.3.18, one has $L \cong \mathcal{O}(X)$ and, therefore, L corresponds to the cocycle $\{\frac{s_i}{s_j}\}$, where s_i and s_j are local equations of X on U_i respectively U_j. Thus, $s_i = \varphi_i^n$. Hence, for $x \in X$ with $\varphi_j(x) = (z_1, \ldots z_{n-1}, 0)$ one has

$$\begin{aligned} \frac{s_i}{s_j}(x) &= \frac{\varphi_i^n}{\varphi_j^n}(x) = \frac{(\varphi_{ij} \circ \varphi_j)^n}{\varphi_j^n}(x) \\ &= \left(\frac{\varphi_{ij}^n}{z_n} \circ \varphi_j\right)(x) = h(z_1, \ldots, z_{n-1}, 0) \\ &= \frac{\partial \varphi_{ij}^n}{\partial z_n}(z_1, \ldots, z_{n-1}, 0) = \left(\frac{\partial \varphi_{ij}^n}{\partial z_n} \circ \varphi_j\right)(x). \end{aligned}$$

Therefore, both line bundles $\mathcal{O}(X)|_X$ and $\mathcal{N}_{X/Y}$ are given by the same cocycle.

The last assertion follows from the adjunction formula 2.2.17. □

Remark 2.4.8 Note that the isomorphism $\mathcal{N}_{Y/X} \cong \mathcal{O}(X)|_X$ can locally be described as the map induced by $\mathcal{T}_Y \twoheadrightarrow \mathcal{O}(X)|_X$, $\frac{\partial}{\partial z} \mapsto \frac{\partial s}{\partial z_i}|_X$.

Corollary 2.4.9 *If $X \subset \mathbb{P}^n$ is a smooth hypersurface of degree d, i.e. defined by a section $s \in H^0(\mathbb{P}^n, \mathcal{O}(d))$, then $K_X \cong \mathcal{O}(d-n-1)|_X$.* □

Exercises

2.4.1 Show that the canonical bundle K_X of a complete intersection $X = Z(f_1) \cap \ldots \cap Z(f_k) \subset \mathbb{P}^n$ is isomorphic to $\mathcal{O}(\sum \deg(f_i) - n - 1))|_X$. What can you deduce from this for the Kodaira dimension of X?

2.4.2 Are there holomorphic vector fields on $\mathbb{P}^n$, i.e. global sections of $\mathcal{T}_{\mathbb{P}^n}$, which vanish only in a finite number of points? If yes, in how many?

2.4.3 Compute the Kodaira dimension of the following smooth hypersurfaces: i) $Z\left(\sum_{i=0}^2 x_i^2\right) \subset \mathbb{P}^2$, ii) $Z\left(\sum_{i=0}^2 x_i^3\right) \subset \mathbb{P}^2$, iii) $Z\left(\sum_{i=0}^3 x_i^3\right) \subset \mathbb{P}^3$, and iv) $Z\left(\sum_{i=0}^4 x_i^5\right) \subset \mathbb{P}^4$.

2.4.4 Show that $H^0(\mathbb{P}^n, \Omega_{\mathbb{P}^n}^q) = 0$ for $q > 0$.

2.4.5 The surface $\Sigma_n = \mathbb{P}(\mathcal{O}_{\mathbb{P}^1} \oplus \mathcal{O}_{\mathbb{P}^1}(n))$ is called the n-th *Hirzebruch surface.* Show that Σ_n is isomorphic to the hypersurface $Z(x_0^n y_1 - x_1^n y_2) \subset \mathbb{P}^1 \times \mathbb{P}^2$, where $(x_0 : x_1)$ and $(y_0 : y_1 : y_2)$ are the homogeneous coordinates of $\mathbb{P}^1$ respectively $\mathbb{P}^2$.

2.4.6 Describe the tangent, cotangent, and canonical bundle of $\mathbb{P}^n \times \mathbb{P}^m$.

2.4.7 Determine all complete intersections $X = H_1 \cap \ldots \cap H_{n+m-2} \subset \mathbb{P}^n \times \mathbb{P}^m$ with $H_i \in |p_1^*\mathcal{O}(d_i) \otimes p_2^*\mathcal{O}(e_i)|$, $d_i, e_i > 0$ with trivial canonical bundle.

2.4.8 Let $Y \subset X$ be a smooth hypersurface in a complex manifold X of dimension n and let α be a meromorphic section of K_X with at most simple poles along Y. Locally one can write $\alpha = h \cdot \frac{dz_1}{z_1} \wedge dz_2 \wedge \ldots \wedge dz_n$ with z_1 defining Y. One sets $\mathrm{Res}_Y(\alpha) = (h \cdot dz_2 \wedge \ldots \wedge dz_n)|_Y$.

a) Show that $\mathrm{Res}_Y(\alpha)$ is well defined and that it yields an element in $H^0(Y, K_Y)$.

b) Consider α as an element in $H^0(X, K_X \otimes \mathcal{O}(Y))$ and compare the definition of the residue with the adjunction formula: $K_Y \cong (K_X \otimes \mathcal{O}(Y))|_Y$.

c) Consider a smooth hypersurface $Y \subset \mathbb{P}^n$ defined by a homogeneous polynomial $f \in H^0(\mathbb{P}^n, \mathcal{O}(n+1))$. Prove that $\alpha := \sum (-1)^i z_i f^{-1} dz_0 \wedge \ldots \wedge \widehat{dz_i} \wedge \ldots \wedge dz_n$ is a meromorphic section of $K_{\mathbb{P}^n}$ with simple poles along Y. Furthermore, show that $\mathrm{Res}_Y(\alpha) \in H^0(Y, K_Y)$ defines a holomorpic volume form on Y, i.e. a trivializing section of K_Y.

2.4.9 Let E be a holomorphic vector bundle on a complex manifold X. Construct the relative tautological holomorphic line bundle $\mathcal{O}_\pi(-1)$ on $\mathbb{P}(E)$. Here, $\pi : \mathbb{P}(E) \to X$ denotes the projection and the fibre $\mathcal{O}_\pi(-1)(\ell)$ for a line $\ell \subset E(x)$ representing a point $\ell \in \pi^{-1}(x)$ should be, by construction, identified with ℓ.

2.4.10 This exercise generalizes the Euler sequence to Grassmannians. Show that on $\mathrm{Gr}_k(V)$ there exists a natural short exact sequence (the tautological sequence) of holomorphic vector bundles

$$\begin{tikzcd} 0 \arrow[r] & S \arrow[r] & \mathcal{O} \otimes V \arrow[r] & Q \arrow[r] & 0 \end{tikzcd}$$

such that over a point $w \in \mathrm{Gr}_k(V)$ corresponding to $W \subset V$ the inclusion $S(w) \subset (\mathcal{O} \otimes V)(w) = V$ is just $W \subset V$. Moreover, prove that $\mathrm{Hom}(S, Q)$ is isomorphic to the holomorphic tangent bundle of $\mathrm{Gr}_k(V)$. Observe that this generalize what has been said about the Euler sequence on $\mathbb{P}^n$ which corresponds to the case $k = 1$

2.4.11 Let X be a complex manifold and let $\mathcal{O}_X \otimes V \twoheadrightarrow E$ be a surjection of vector bundles where V is a vector space and E is a vector bundle of rank k. Show that there exists a natural morphism $\varphi : X \to \mathrm{Gr}_k(V)$ with $\varphi^*(\mathcal{O} \otimes V \to Q) = \mathcal{O}_X \otimes V \to E$. Proposition 2.3.26 is the special case $k = 1$. Compare the Euler sequence with the tautological sequence introduced in the previous exercise via the Plücker embedding.

4.3 Curvature

In the previous paragraph we introduced the notion of a connection generalizing the exterior differential to sections of vector bundles. Indeed, by definition a connection satisfies the Leibniz rule. However, in general a connection ∇ need not satisfy $\nabla^2 = 0$, i.e. ∇ is not a differential. The obstruction for a connection to define a differential is measured by its curvature. This concept will be explained now.

The reader familiar with the basic concepts in Riemannian geometry will find the discussion similar to the treatment of the curvature of a Riemannian manifold. As before, we shall also study what happens in the presence of additional structures, e.g. hermitian and holomorphic ones.

Let E be a vector bundle on a manifold M endowed with a connection $\nabla : \mathcal{A}^0(E) \to \mathcal{A}^1(E)$. Then a natural extension

$$\nabla : \mathcal{A}^k(E) \longrightarrow \mathcal{A}^{k+1}(E)$$

is defined as follows: If α is a local k-form on M and s is a local section of E then

$$\nabla(\alpha \otimes s) = d(\alpha) \otimes s + (-1)^k \alpha \wedge \nabla(s).$$

Observe that for $k = 0$ this is just the Leibniz formula (4.2) which also ensures that the extension is well-defined, i.e. $\nabla(\alpha \otimes (f \cdot s)) = \nabla(f\alpha \otimes s)$ for any local function f. Moreover, a generalized Leibniz rule also holds for this extension, i.e. for any section t of $\mathcal{A}^\ell(E)$ and any k-form β one has

$$\nabla(\beta \wedge t) = d(\beta) \wedge t + (-1)^k \beta \wedge \nabla(t).$$

Indeed, if $t = \alpha \otimes s$ then

$$\begin{aligned} \nabla(\beta \wedge t) &= \nabla((\beta \wedge \alpha) \otimes s) = d(\beta \wedge \alpha) \otimes s + (-1)^{k+\ell}(\beta \wedge \alpha) \otimes \nabla(s) \\ &= d(\beta) \wedge t + (-1)^k((\beta \wedge d(\alpha)) \otimes s + (-1)^\ell(\beta \wedge \alpha) \otimes \nabla(s)) \\ &= d(\beta) \wedge t + (-1)^k \beta \wedge \nabla(t) \end{aligned}$$

Definition 4.3.1 The *curvature* F_∇ of a connection ∇ on a vector bundle E is the composition

$$F_\nabla := \nabla \circ \nabla : \mathcal{A}^0(E) \longrightarrow \mathcal{A}^1(E) \longrightarrow \mathcal{A}^2(E).$$

Usually, the curvature F_∇ will be considered as a global section of $\mathcal{A}^2(\mathrm{End}(E))$, i.e. $F_\nabla \in \mathcal{A}^2(M, \mathrm{End}(E))$. This is justified by the following result.

Lemma 4.3.2 *The curvature homomorphism* $F_\nabla : \mathcal{A}^0(E) \to \mathcal{A}^2(E)$ *is* $\mathcal{A}^0$-*linear.*

Proof. For a local section s of E and a local function f on M one computes

$$\begin{aligned} F_\nabla(f\cdot s) &= \nabla(\nabla(f\cdot s)) = \nabla\left(df\otimes s + f\cdot\nabla(s)\right)\\ &= \underbrace{d^2(f)}_{0}\otimes s\underbrace{-df\wedge\nabla(s)+df\wedge\nabla(s)}_{0}+f\cdot\nabla(\nabla(s))\\ &= f\cdot F_\nabla(s)\end{aligned}$$

□

Examples 4.3.3 i) Let us compute the curvature of a connection on the trivial bundle $M\times\mathbb{C}^r$. If ∇ is the trivial connection, i.e. $\nabla = d$, then $F_\nabla = 0$.

Any other connection is of the form $\nabla = d + A$, where A is a matrix of one-forms. For a section s one obtains $F_\nabla(s) = (d+A)(d+A)(s) = (d+A)(d(s)+A\cdot s) = d^2(s)+d(A\cdot s)+A\cdot d(s)+A(A(s)) = d(A)(s)+(A\wedge A)(s)$, i.e.

$$F_\nabla = d(A) + A\wedge A.$$

ii) For a line bundle this calculation yields $F_\nabla = d(A)$. In this case, the curvature is an ordinary two-form.

For $a\in\mathcal{A}^1(M,\mathrm{End}(E))$ the two-form $a\wedge a\in\mathcal{A}^2(M,\mathrm{End}(E))$ is given by exterior product in the form part and composition in $\mathrm{End}(E)$. Using this, the example is easily generalized to

Lemma 4.3.4 *Let ∇ be a connection on a vector bundle E and let $a\in\mathcal{A}^1(M,\mathrm{End}(E))$. Then $F_{\nabla+a} = F_\nabla + \nabla(a) + a\wedge a$.* □

Any connection ∇ on a vector bundle E induces a natural connection on $\mathrm{End}(E)$ which will also be called ∇. In particular, this connection can be applied to the curvature $F_\nabla\in\mathcal{A}^2(M,\mathrm{End}(E))$ of the original connection on E. Here is the next remarkable property of the curvature:

Lemma 4.3.5 (Bianchi identity) *If $F_\nabla\in\mathcal{A}^2(M,\mathrm{End}(E))$ is the curvature of a connection ∇ on a vector bundle E, then*

$$0 = \nabla(F_\nabla)\in\mathcal{A}^3(M,\mathrm{End}(E)).$$

Proof. This follows from $\nabla(F_\nabla)(s) = \nabla(F_\nabla(s)) - F_\nabla(\nabla(s)) = \nabla(\nabla^2(s)) - \nabla^2(\nabla(s)) = 0$. Here we use Exercise 4.3.1. □

Example 4.3.6 For the connection $\nabla = d + A$ on the trivial bundle the Bianchi identity becomes $dF_\nabla = F_\nabla\wedge A - A\wedge F_\nabla$.

The curvature of induced connections on associated bundles can usually be expressed in terms of the curvature of the given connections. For the most frequent associated bundles they are given by the following proposition.

Proposition 4.3.7 *Let* E_1, E_2 *be vector bundles endowed with connections* ∇_1 *and* ∇_2, *respectively.*

i) *The curvature of the induced connection on the direct sum* $E_1 \oplus E_2$ *is given by*

$$F = F_{\nabla_1} \oplus F_{\nabla_2}.$$

ii) *On the tensor product* $E_1 \otimes E_2$ *the curvature is given by*

$$F_{\nabla_1} \otimes 1 + 1 \otimes F_{\nabla_2}.$$

iii) *For the induced connection* ∇^* *on the dual bundle* E^* *one has*

$$F_{\nabla^*} = -F_{\nabla}^{\mathrm{t}}.$$

iv) *The curvature of the pull-back connection* $f^*\nabla$ *of a connection* ∇ *under a differentiable map* $f : M \to N$ *is*

$$F_{f^*\nabla} = f^* F_{\nabla}.$$

Proof. Let us prove ii). This is the following straightforward calculation

$$\begin{aligned}
& F_{\nabla}(s_1 \otimes s_2) \\
&= \nabla(\nabla(s_1 \otimes s_2)) = \nabla(\nabla_1(s_1) \otimes s_2 + s_1 \otimes \nabla_2(s_2)) \\
&= \nabla_1^2(s_1) \otimes s_2 - \nabla_1(s_1) \otimes \nabla_2(s_2) + \nabla_1(s_1) \otimes \nabla_2(s_2) + s_1 \otimes \nabla_2^2(s_2) \\
&= F_{\nabla_1}(s_1) \otimes s_2 + s_1 \otimes F_{\nabla_2}(s_2).
\end{aligned}$$

The sign appears, because ∇ is applied to the one-form $\nabla_1(s_1) \otimes s_2$. We leave i) and iii) to the reader (cf. Exercise 4.3.2).

iv) follows from the local situation, where ∇ is given as $d + A$. Then $F_{f^*\nabla} = F_{d+f^*A} = d(f^*A) + f^*(A) \wedge f^*(A) = f^*(d(A) + A \wedge A) = f^* F_{\nabla}$. □

Next we will study the curvature of the special connections introduced in Section 4.2.

Proposition 4.3.8 i) *The curvature of an hermitian connection* ∇ *on an hermitian vector bundle* (E, h) *satisfies* $h(F_{\nabla}(s_1), s_2) + h(s_1, F_{\nabla}(s_2)) = 0$, *i.e.*

$$F_{\nabla} \in \mathcal{A}^2(M, \mathrm{End}(E, h)).$$

ii) *The curvature* F_{∇} *of a connection* ∇ *on a holomorphic vector bundle* E *over a complex manifold* X *with* $\nabla^{0,1} = \bar{\partial}$ *has no* $(0,2)$*-part, i.e.*

$$F_{\nabla} \in (\mathcal{A}^{2,0} \oplus \mathcal{A}^{1,1})(X, \mathrm{End}(E)).$$

iii) *Let* E *be a holomorphic bundle endowed with an hermitian structure* h. *The curvature of the Chern connection* ∇ *is of type* $(1,1)$, *real, and skew-hermitian, i.e.* $F_{\nabla} \in \mathcal{A}_{\mathbb{R}}^{1,1}(X, \mathrm{End}(E, h))$. *(Recall that* $\mathrm{End}(E, h)$ *is only a real vector bundle.)*

Proof. i) A local argument goes as follows: Choose a trivialization such that (E, h) is isomorphic to the trivial bundle $M \times \mathbb{C}^r$ endowed with the constant standard hermitian product. Then, $\nabla = d + A$ with $\bar{A}^{\mathrm{t}} = -A$. For its curvature $F_\nabla = d(A) + A \wedge A$ one obtains

$$\begin{aligned}\bar{F}_\nabla^{\mathrm{t}} &= d(\bar{A}^{\mathrm{t}}) + (\bar{A} \wedge \bar{A})^{\mathrm{t}} = d(\bar{A}^{\mathrm{t}}) - \bar{A}^{\mathrm{t}} \wedge \bar{A}^{\mathrm{t}} \\ &= d(-A) - (-A) \wedge (-A) = -F_\nabla.\end{aligned}$$

For a global argument one first observes that the assumption that ∇ is hermitian yields for form-valued sections $s_i \in \mathcal{A}^{k_i}(E)$ the equation

$$dh(s_1, s_2) = h(\nabla(s_1), s_2) + (-1)^{k_1} h(s_1, \nabla(s_2)).$$

Recall that $h(\alpha_1 \otimes t_1, \alpha_2 \otimes t_2) = (\alpha_1 \wedge \bar{\alpha}_2) h(t_1, t_2)$ for local forms α_1, α_2 and sections t_1, t_2.

Hence, for $s_1, s_2 \in \mathcal{A}^0(E)$ one has

$$\begin{aligned}dh(\nabla(s_1), s_2) &= h(F_\nabla(s_1), s_2) - h(\nabla(s_1), \nabla(s_2)) \\ dh(s_1, \nabla(s_2)) &= h(\nabla(s_1), \nabla(s_2)) + h(s_1, F_\nabla(s_2))\end{aligned}$$

and on the other hand

$$dh(\nabla(s_1), s_2)) + dh(s_1, \nabla(s_2)) = d(dh(s_1, s_2)) = 0.$$

This yields the assertion.

ii) Here, one first observes that the extension $\nabla : \mathcal{A}^k(E) \to \mathcal{A}^{k+1}(E)$ splits into a $(1,0)$-part and a $(0,1)$-part, the latter of which is $\bar{\partial}$. Then one computes $\nabla^2 = (\nabla^{1,0})^2 + \nabla^{1,0} \circ \bar{\partial} + \bar{\partial} \circ \nabla^{1,0}$, as $\bar{\partial}^2 = 0$. Hence, $F_\nabla(s) \in (\mathcal{A}^{2,0} \oplus \mathcal{A}^{1,1})(E)$ for all $s \in \mathcal{A}^0(E)$.

Locally one could argue as follows: Since $\nabla = d + A$ with A of type $(1,0)$, the curvature $d(A) + A \wedge A = \bar{\partial}(A) + (\partial A + A \wedge A)$ is a sum of a $(1,1)$-form and a $(2,0)$-form.

iii) Combine i) and ii). By comparing the bidegree of F_∇^{t} and F_∇ in local coordinates we find that F_∇ is of type $(1,1)$.

More globally, due to ii) one knows that $h(F_\nabla(s_1), s_2)$ and $h(s_1, F_\nabla(s_2))$ are of bidegree $(2,0) + (1,1)$ respectively $(1,1) + (0,2)$. Recall the convention $h(s_1, \alpha_2 \otimes s_2) = \bar{\alpha}_2 h(s_1, s_2)$. Using i), i.e. $h(F_\nabla(s_1), s_2) + h(s_1, F_\nabla(s_2)) = 0$, and comparing the bidegree shows that $h(F_\nabla(s_1), s_2)$ has trivial $(2,0)$-part for all sections s_1, s_2. Hence, F_∇ is of type $(1,1)$. ☐

Examples 4.3.9 i) Suppose E is the trivial vector bundle $M \times \mathbb{C}^r$ with the constant standard hermitian structure. If $\nabla = d + A$ is an hermitian connection, then i) says $d(A + \bar{A}^{\mathrm{t}}) + (A \wedge A + \overline{(A \wedge A)}^{\mathrm{t}}) = 0$. For $r = 1$ this means that the real part of A is constant.

ii) If (L, h) is an hermitian holomorphic line bundle, then the curvature F_∇ of its Chern connection is a section of $\mathcal{A}_{\mathbb{R}}^{1,1}(X, \mathrm{End}(L, h))$, which can be identified with the imaginary $(1,1)$-forms on X. Indeed, $\mathrm{End}(L, h)$ is the purely imaginary line bundle $i \cdot \mathbb{R} \times X$ (cf. page 176).

iii) If the hermitian structure on the holomorphic vector bundle E is locally given by the matrix H, then the Chern connection is of the form $d+\bar{H}^{-1}\partial(\bar{H})$. Hence, the curvature is

$$F = \bar{\partial}(\bar{H}^{-1}\partial(\bar{H})).$$

Indeed, a priori $F = \bar{\partial}(\bar{H}^{-1}\partial(\bar{H})) + \partial(\bar{H}^{-1}\partial(\bar{H}) + (\bar{H}^{-1}\partial(\bar{H})) \wedge (\bar{H}^{-1}\partial(\bar{H}))$, but the sum of the last two terms must vanish, as both are of type $(2,0)$.

If E is a line bundle, then the hermitian matrix is just a positive real function h. In this case one writes $F = \bar{\partial}\partial\log(h)$. Once again, we see that F can be considered as a purely imaginary two-form

iv) Let (X,g) be an hermitian manifold. Then the tangent bundle is naturally endowed with an hermitian structure. The curvature of the Chern connection on $\mathcal{T}_X$ is called the *curvature of the hermitian manifold* (X,g). In Section 4.A we shall see that the curvature of a Kähler manifold (X,g) is nothing but the usual curvature of the underlying Riemannian manifold.

For the Chern connection on a holomorphic hermitian bundle the Bianchi identity yields

$$0 = (\nabla(F_\nabla))^{1,2} = \bar{\partial}(F_\nabla),$$

i.e. F_∇ as an element of $\mathcal{A}^{1,1}(X, \mathrm{End}(E))$ is $\bar{\partial}$-closed. Thus, in this case the curvature yields a natural Dolbeault cohomology class $[F_\nabla] \in H^1(X, \Omega_X \otimes \mathrm{End}(E))$ for any holomorphic vector bundle E. A priori, this cohomology class might depend on the chosen hermitian structure or, equivalently, on the connection. That this is not the case is an immediate consequence of the following description of it as the Atiyah class of E.

Proposition 4.3.10 *For the curvature F_∇ of the Chern connection on an hermitian holomorphic vector bundle (E,h) one has*

$$[F_\nabla] = A(E) \in H^1(X, \Omega_X \otimes \mathrm{End}(E)).$$

Proof. The comparison of Dolbeault and Čech cohomology can be done by chasing through the following commutative diagram of sheaves:

$$\begin{tikzcd}
\Omega_X \otimes \mathrm{End}(E) \arrow[r] \arrow[d] & \mathcal{C}^0(\{U_i\}, \Omega_X \otimes \mathrm{End}(E)) \arrow[r] \arrow[d] & \mathcal{C}^1(\{U_i\}, \Omega_X \otimes \mathrm{End}(E)) \arrow[d, "i"] \\
\mathcal{A}^{1,0}(\mathrm{End}(E)) \arrow[r] \arrow[d] & \mathcal{C}^0(\{U_i\}, \mathcal{A}^{1,0}(\mathrm{End}(E))) \arrow[r, "\delta_1"] \arrow[d, "\bar{\partial}"] & \mathcal{C}^1(\{U_i\}, \mathcal{A}^{1,0}(\mathrm{End}(E))) \\
\mathcal{A}^{1,1}(\mathrm{End}(E)) \arrow[r, "\delta_0"] & \mathcal{C}^0(\{U_i\}, \mathcal{A}^{1,1}(\mathrm{End}(E))) &
\end{tikzcd}$$

Here, $X = \bigcup U_i$ is an open covering of X trivializing E via $\psi_i : E|_{U_i} \cong U_i \times \mathbb{C}^r$. With respect to ψ_i the hermitian structure h on U_i is given by an hermitian matrix H_i. Then the curvature F_∇ of the Chern connection on the holomorphic bundle E on U_i is given by $F_\nabla|_{U_i} = \psi^{-1}(\bar\partial(\bar H_i^{-1}\partial \bar H_i))\psi_i$. Thus,

$$\begin{aligned}\delta_0(F_\nabla) &= \{U_i, \psi_i^{-1}\circ(\bar\partial(\bar H_i^{-1}\partial\bar H_i))\circ\psi_i\}\\ &= \bar\partial\{U_i, \psi_i^{-1}\circ(\bar H_i^{-1}\partial\bar H_i)\circ\psi_i\},\end{aligned}$$

as the maps ψ_i are holomorphic trivializations.

Hence, it suffices to show that

$$i\{U_{ij}, \psi_j^{-1}\circ(\psi_{ij}^{-1}d\psi_{ij})\circ\psi_j\} = \delta_1\{U_i, \psi_i^{-1}\circ(\bar H_i^{-1}\partial\bar H_i)\circ\psi_i\},$$

because the cocycle $\{U_i, \psi_i^{-1}\circ(\psi_{ij}^{-1}d\psi_{ij})\circ\psi_j\}$ represents by definition the Atiyah class of E. Using the definition of δ_1 one shows that the term on the right hand side equals

$$\{U_{ij}, \psi_j^{-1}\circ(\bar H_j^{-1}\partial\bar H_j)\circ\psi_j - \psi_i^{-1}\circ(\bar H_i^{-1}\partial\bar H_i)\circ\psi_i)\}.$$

Since

$$\begin{aligned}&\psi_j^{-1}\circ(\bar H_j^{-1}\partial\bar H_j)\circ\psi_j - \psi_i^{-1}\circ(\bar H_i^{-1}\partial\bar H_i)\circ\psi_i\\ &= \psi_j^{-1}\circ(\bar H_j^{-1}\partial\bar H_j - \psi_{ij}^{-1}\circ(\bar H_i^{-1}\partial\bar H_i)\circ\psi_{ij})\circ\psi_j,\end{aligned}$$

it suffices to prove

$$\bar H_j^{-1}\partial\bar H_j - \psi_{ij}^{-1}\circ(\bar H_i^{-1}\partial\bar H_i)\circ\psi_{ij} = \psi_{ij}^{-1}\partial\psi_{ij}.$$

The latter is a consequence of the compatibility $\psi_{ij}^{\mathrm t}H_i\bar\psi_{ij} = H_j$ or, equivalently, $\bar\psi_{ij}^{\mathrm t}\bar H_i\psi_{ij} = \bar H_j$ and the chain rule

$$\begin{aligned}\bar H_j^{-1}\partial\bar H_j &= \psi_{ij}^{-1}\bar H_i^{-1}(\bar\psi_{ij}^{\mathrm t})^{-1}\left(\bar\psi_{ij}^{\mathrm t}(\partial\bar H_i)\psi_{ij} + \bar\psi_{ij}^{\mathrm t}\bar H_i\partial\psi_{ij}\right)\\ &= \psi_{ij}^{-1}(\bar H_i^{-1}\partial\bar H_i)\psi_{ij} + \psi_{ij}^{-1}\partial\psi_{ij},\end{aligned}$$

where we have used $\partial\bar\psi_{ij} = 0$. □

Remark 4.3.11 Here is a more direct argument to see that $[F_\nabla]$ does not depend on the connection ∇. Indeed, any other connection is of the form $\nabla + a$ and $F_{\nabla+a} = F_\nabla + \nabla(a) + a\wedge a$. If both connections ∇ and $\nabla + a$ are Chern connections with respect to certain hermitian structures, then F_∇ and $F_{\nabla+a}$ are $(1,1)$-forms and $a \in \mathcal{A}^{1,0}(\mathrm{End}(E))$. Thus, $a\wedge a \in \mathcal{A}^{2,0}(\mathrm{End}(E))$ and, therefore, $\nabla(a) + a\wedge a = (\nabla(a) + a\wedge a)^{1,1} = \nabla(a)^{1,1} = \bar\partial(a)$, i.e. $F_{\nabla+a} = F_\nabla + \bar\partial(a)$ and hence

$$[F_{\nabla+a}] = [F_\nabla].$$

We have used here that the induced connection on the endomorphism bundle is again compatible with the holomorphic structure.

Example 4.3.12 Let us consider the standard homogeneous linear coordinates $z_0, \ldots, z_n$ on $\mathbb{P}^n$ as sections of $\mathcal{O}(1)$. According to Example 4.1.2 one associates with them a natural hermitian metric $h(t) = |\psi(t)|^2/(\sum |\psi(z_i)|^2)$ on $\mathcal{O}(1)$.

We claim that the curvature F of the Chern connection on the holomorphic line bundle $\mathcal{O}(1)$ endowed with this hermitian metric is

$$\frac{i}{2\pi}F = \omega_{\mathrm{FS}},$$

where ω_{FS} is the Fubini–Study Kähler form (see Example 3.1.9, i)).

This can be verified locally. On a standard open subset $U_i \subset \mathbb{P}^n$ one has $\omega_{\mathrm{FS}} = \frac{i}{2\pi}\partial\bar\partial\log(1+\sum|w_i|^2)$.

The hermitian structure of $\mathcal{O}(1)|_{U_i}$ with respect to the natural trivialization is given by the scalar function $h = (1+\sum|w_i|^2)^{-1}$. By Example iii) in 4.3.9 we have $F = \bar\partial\partial\log(h) = \partial\bar\partial\log(1+\sum|w_i|^2)$.

Remark 4.3.13 In the general situation, i.e. of a holomorphic line bundle L on a complex manifold X, the hermitian structure $h_{\{s_i\}}$ induced by globally generating sections $s_0, \ldots, s_n$ is the pull-back of the hermitian structure on $\mathcal{O}(1)$ of the previous example under the induced morphism $\varphi : X \to \mathbb{P}^n$. Thus, for the curvature F_∇ of the Chern connection ∇ on L one has: $(i/2\pi)F_\nabla = \varphi^*\omega_{\mathrm{FS}}$, where ω_{FS} is the Fubini–Study form on $\mathbb{P}^n$.

Suppose now that L is not only holomorphic, but also endowed with an hermitian structure h. What is the relation between h and $h_{\{s_i\}}$? In general, they are not related at all, but if the sections $s_0, \ldots, s_n$ are chosen such that they form an orthonormal base of $H^0(X, L)$ then one might hope to approximate h by $h_{\{s_i\}}$. (Here, $H^0(X, L)$ is equipped with the hermitian product defined in 4.1.11.) This circle of questions is intensively studied at the moment and there are many open questions.

Definition 4.3.14 A real $(1,1)$-form α is called *(semi-)positive* if for all holomorphic tangent vectors $0 \neq v \in T^{1,0}X$ one has

$$-i\alpha(v, \bar v) > 0 \text{ (resp. } \geq).$$

At a point $x \in X$ any semi-positive $(1,1)$-form is a positive linear combination of forms of the type $i\beta\wedge\bar\beta$, where β is a $(1,0)$-form. The standard example of a positive form is provided by a Kähler form $\omega = \frac{i}{2}\sum h_{ij}dz_i \wedge d\bar z_j$. In this case (h_{ij}) is a positive hermitian matrix and thus $v^{\mathrm{t}}(h_{ij})\bar v > 0$ for all non-zero holomorphic tangent vectors. Together with Example 4.3.12 this shows that iF_∇ of the Chern connection ∇ on $\mathcal{O}(1)$ (with the standard hermitian structure) is positive.

Clearly, the pull-back of a semi-positive form is again semi-positive. For the curvature of a globally generated line bundle L this implies

$$F_\nabla(v, \bar v) \geq 0.$$

We next wish to generalize this observation to the higher rank case. Let (E, h) be an hermitian vector bundle over a complex manifold X and let ∇ be an hermitian connection that satisfies $F_\nabla \in \mathcal{A}^{1,1}(X, \mathrm{End}(E))$.

Definition 4.3.15 The curvature F_∇ is (Griffiths-)*positive* if for any $0 \neq s \in E$ one has

$$h(F_\nabla(s), s)(v, \bar{v}) > 0$$

for all non-trivial holomorphic tangent vectors v. Semi-positivity (negativity, semi-negativity) is defined analogously.

Remark 4.3.16 Let L be an hermitian holomorphic line bundle and let ∇ be the Chern connection. Its curvature F_∇ is an imaginary $(1,1)$-form. Then the curvature F_∇ is positive in the sense of Definition 4.3.15 if and only if the real $(1,1)$-form iF_∇ is positive in the sense of Definition 4.3.14. The extra factor i is likely to cause confusion at certain points, but these two concepts of positivity (and others) are met frequently in the literature.

Before proving the semi-positivity of the curvature of any globally generated vector bundle, we need to relate the curvature and the second fundamental forms of a split vector bundle $E = E_1 \oplus E_2$. Let ∇ be a connection on E. We denote the induced connections by ∇_i and the second fundamental forms by b_i, $i = 1, 2$. Thus,

$$\nabla = \begin{pmatrix} \nabla_1 & b_2 \\ b_1 & \nabla_2 \end{pmatrix}$$

which immediately shows

Lemma 4.3.17 *The curvature of the induced connection ∇_1 on E_1 is given by $F_{\nabla_1} = \mathrm{pr}_1 \circ F_\nabla - b_2 \circ b_1$.* □

Now let E_1 be a holomorphic subbundle of the trivial holomorphic vector bundle $E = \mathcal{O}^{\oplus r}$ endowed with the trivial constant hermitian structure. The curvature of the Chern connection ∇ on E is trivial, as ∇ is just the exterior differential. Hence, $F_{\nabla_1} = -b_2 \circ b_1$. This leads to

Proposition 4.3.18 *The curvature F_{∇_1} of the Chern connection ∇_1 of a subbundle $E_1 \subset E = \mathcal{O}^{\oplus r}$ (with the induced hermitian structure) is semi-negative.*

Proof. By the previous lemma we have $F_{\nabla_1} = -b_2 \circ b_1$. Thus, if h_2 is the induced hermitian structure on the quotient E/E_1 then

$$h_1(F_{\nabla_1}(s), s) = -h_1(b_2(b_1(s)), s) = -h_2(b_1(s), b_1(s)).$$

Here we use properties of the second fundamental form proved in Section 4.2.

More precisely, one has

$$\begin{aligned} h_1(b_2(\beta\otimes t),s) &= h_1(-\beta\otimes b_2(t),s) = -\beta h_1(b_2(t),s) \\ &= \beta h_2(t,b_1(s)) = h_2(\beta\otimes t, b_1(s)) \end{aligned}$$

for any one-form β.

Now, it suffices to show that $h(\alpha,\alpha)\geq 0$ for any $\alpha\in\mathcal{A}^{1,0}(E)$. Fix an orthonormal basis s_i of E and write $\alpha=\sum\alpha_i s_i$. Then $h(\alpha,\alpha)=\sum\alpha_i\wedge\bar{\alpha}_i$. But $\alpha_i\wedge\bar{\alpha}_i$ is semi-positive due to

$$(\alpha_i\wedge\bar{\alpha}_i)(v,\bar{v}) = \alpha_i(v)\cdot\bar{\alpha}_i(\bar{v}) = \alpha_i(v)\cdot\overline{\alpha_i(v)}\geq 0.$$

(Note that we did use that b_1 is of type $(1,0)$.) ☐

A holomorphic vector bundle E is *globally generated* if there exist global holomorphic sections $s_1,\ldots,s_r\in H^0(X,E)$ such that for any $x\in X$ the values $s_1(x),\ldots,s_r(x)$ generate the fibre $E(x)$. In other words, the sections $s_1,\ldots,s_r$ induce a surjection $\mathcal{O}^{\oplus r}\twoheadrightarrow E$. The standard constant hermitian structure on $\mathcal{O}^{\oplus r}$ induces via this surjection an hermitian structure on E and dualizing this surjection yields an inclusion of vector bundles $E^*\subset\mathcal{O}^{\oplus r}$.

Using Exercise 4.3.3 the proposition yields

Corollary 4.3.19 *The curvature of a globally generated vector bundle is semi-positive.* ☐

Here, the curvature is the curvature of the Chern connection with respect to a hermitian structure on E defined by the choice of the globally generating holomorphic sections.

Example 4.3.20 The Euler sequence on $\mathbb{P}^n$ twisted by $\mathcal{O}(-1)$ is of the form

$$\begin{tikzcd} 0 \arrow[r] & \mathcal{O}(-1) \arrow[r] & \mathcal{O}^{\oplus n+1} \arrow[r] & \mathcal{T}_{\mathbb{P}^n}(-1) \arrow[r] & 0\,. \end{tikzcd}$$

Hence, $\mathcal{T}_{\mathbb{P}^n}(-1)$ admits an hermitian structure such that the curvature of the Chern connection is semi-positive. Twisting by $\mathcal{O}(1)$ yields a connection on $\mathcal{T}_{\mathbb{P}^n}$ with positive curvature. In fact, this is the curvature of the Fubini–Study metric on $\mathbb{P}^n$. See Example 4.1.5.

Remark 4.3.21 As for line bundles, one could have first studied the universal case. Recall that on the Grassmannian $\mathrm{Gr}_r(V)$ there exists a universal short exact sequence

$$\begin{tikzcd} 0 \arrow[r] & S \arrow[r] & \mathcal{O}\otimes V \arrow[r] & Q \arrow[r] & 0. \end{tikzcd}$$

By definition, the fibre of S over a point of $\mathrm{Gr}(V)$ that corresponds to $W\subset V$ is naturally isomorphic to W. Fixing an hermitian structure on V induces hermitian structures on S and Q (and therefore on the tangent bundle $\mathcal{T}$ of $\mathrm{Gr}_r(V)$, which is $\mathrm{Hom}(S,Q)$).

If E is any holomorphic vector bundle of rank r on a complex manifold X generated by global sections spanning $V \subset H^0(X,E)$, then there exists a morphism $\varphi : X \to \mathrm{Gr}_r(V)$ with $\varphi^*Q = E$ (see Exercise 2.4.11). Again, choosing an hermitian product on V induces an hermitian structure on E which coincides with the pull-back of the hermitian structure on Q under φ. This shows that the curvature of E and Q can be compared, namely $\varphi^*F_Q = F_E$.

Thus, showing the positivity of F_Q proves the semi-positivity for any globally generated vector bundle E. Moreover, if $\varphi : X \to \mathrm{Gr}_r(V)$ is an embedding, then the curvature F_E is positive.

Exercises

4.3.1 Show that $\nabla^2 : \mathcal{A}^k(E) \to \mathcal{A}^{k+2}(E)$ is given by taking the exterior product with the form part of the curvature $F_\nabla \in \mathcal{A}^2(M, \mathrm{End}(E))$ and applying its endomorphism part to E.

4.3.2 Prove i) and iii) of Proposition 4.3.7. Compute the connection and the curvature of the determinant bundle.

4.3.3 Let (E_1, h_1) and (E_2, h_2) be two hermitian holomorphic vector bundles endowed with hermitian connections ∇_1, ∇_2 such that the curvature of both is (semi-)positive. Prove the following assertions.

i) The curvature of the induced connection ∇^* on the dual vector bundle E_1^* is (semi-)negative.

ii) The curvature on $E_1 \otimes E_2$ is (semi-)positive and it is positive if at least one of the two connections has positive curvature.

iii) The curvature on $E_1 \oplus E_2$ is (semi-)positive.

4.3.4 Find an example of two connections ∇_1 and ∇_2 on a vector bundle E, such that F_{∇_1} is positive and F_{∇_2} is negative.

4.3.5 Let L be a holomorphic line bundle on a complex manifold. Suppose L admits an hermitian structure whose Chern connection has positive curvature. Show that X is Kähler. If X is in addition compact prove $\int_X A(L)^n > 0$.

4.3.6 Show that the canonical bundle of $\mathbb{P}^n$ comes along with a natural hermitian structure such that the curvature of the Chern connection is negative.

4.3.7 Show that the curvature of a complex torus $\mathbb{C}^n/\Gamma$ endowed with a constant hermitian structure is trivial. Is this true for any hermitian structure on $\mathbb{C}^n/\Gamma$?

4.3.8 Show that the curvature of a the natural metric on the ball quotient introduced in Example 3.1.9, iv) is negative. The one-dimensional case is a rather easy calculation.

4.3.9 Let X be a compact Kähler manifold with $b_1(X) = 0$. Show that there exists a unique flat connection ∇ on the trivial holomorphic line bundle $\mathcal{O}$ with $\nabla^{0,1} = \bar{\partial}$. Moreover, up to isomorphism $\mathcal{O}$ is the only line bundle with trivial Chern class $c_1 \in H^2(X, \mathbb{Z})$.

4.3.10 Let ∇ be a connection on a (complex) line bundle L on a manifold M. Show that L locally admits trivializing parallel sections if and only if $F_\nabla = 0$.

This is the easiest case of the general fact that a connection on a vector bundle is flat if and only if parallel frames can be found locally. (Frobenius integrability.)

Comments: - Various notions for the positivity of vector bundles can be found in the literature. Usually they all coincide for line bundles, but the exact relations between them is not clear for higher rank vector bundles. Positivity is usually exploited to control higher cohomology groups. This will be illustrated in Section 5.2. We shall also see that, at least for line bundles on projective manifolds, an algebraic description of bundles admitting an hermitian structure whose Chern connection has positive curvature can be given. For a more in depth presentation of the material we refer to [35, 100] and the forthcoming book [83].

- The problem alluded to in Remark 4.3.13 is subtle. See [37, 107].

- The curvature of Kähler manifold will be dealt with in subsequent sections, in particular its comparison with the well-known Riemannian curvature shall be explained in detail.

- In [75, Prop. 1.2.2] one finds a cohomological version of the Bianchi identity in terms of the Atiyah class.

% /012/ / /3 /#! 0 - ? + 5
* 9 7 - @ * 7 L . *"
' & ! # HMG D >

Exercises

1.2.2. Prove the convergence of Newton's method in the ring of formal power series $\mathbb{C}[[x]]$, where $\phi_n \to 0$ if and only if the degree of its leading term $\to \infty$. Note that this ring is equipped with the x-adic topology, where the ideals (x^N) form a fundamental system of neighborhoods of 0.

1.2.3. Prove that Fermat's curve $x_0^n + x_1^n + x_2^n = 0$ in $\mathbb{P}^2$ is nonsingular.

1.3 Bézout's Theorem

An important feature of the projective plane is that any two lines meet. In fact, it has a much stronger property:

Theorem 1.3.1 (Weak Bézout's theorem). *Any two algebraic curves in* $\mathbb{P}^2$ *intersect.*

We give an elementary classical proof here using resultants. Given two monic polynomials

$$f(y) = y^n + a_{n-1}y^{n-1} + \cdots + a_0 = \prod_{i=1}^{n}(y - r_i),$$

$$g(y) = y^m + b_{m-1}y^{m-1} + \cdots + b_0 = \prod_{j=1}^{m}(y - s_j),$$

their resultant is the expression

$$\mathrm{Res}(f,g) = \prod_{ij}(r_i - s_j).$$

It is obvious that $\mathrm{Res}(f,g) = 0$ if and only if f and g have a common root. From the way we have written it, it is also clear that $\mathrm{Res}(f,g)$ is a polynomial of degree mn in $r_1,\ldots,r_n,s_1,\ldots,s_m$ that is symmetric separately in the r's and s's. So it can be rewritten as a polynomial in the elementary symmetric polynomials in the r's and s's. In other words, $\mathrm{Res}(f,g)$ is a polynomial in the coefficients a_i and b_j. Standard formulas for it can be found, for example, in [76].

Proof. Assume that the curves are given by homogeneous polynomials $F(x,y,z)$ and $G(x,y,z)$ respectively. After translating the line at infinity if necessary, we can assume that the polynomials $f(x,y) = F(x,y,1)$ and $g(x,y) = G(x,y,1)$ are both nonconstant in x and y. Treating these as polynomials in y with coefficients in $\mathbb{C}[x]$, the resultant $\mathrm{Res}(f,g)(x)$ can be regarded as a nonconstant polynomial in x. Since $\mathbb{C}$ is algebraically closed, $\mathrm{Res}(f,g)(x)$ must have a root, say a. Then $f(a,y) = 0$ and $g(a,y) = 0$ have a common solution. $\square$

It is worth noting that this argument is entirely algebraic, and therefore applies to any algebraically closed field, such as the field of algebraic numbers $\overline{\mathbb{Q}}$. So as a bonus, we get the following arithmetic consequence.

Corollary 1.3.2. *If the curves are defined by equations with coefficients in* $\overline{\mathbb{Q}}$*, then the there is a point of intersection with coordinates in* $\overline{\mathbb{Q}}$*.*

Suppose that the curves $C, D \subset \mathbb{P}^2$ are irreducible and distinct. Then it is not difficult to see that $C \cap D$ is finite. We can ask how many points are in the intersection. To get a more refined answer, we can assign a multiplicity to the points of intersection. If the curves are defined by polynomials $f(x,y)$ and $g(x,y)$ with a common isolated zero at the origin $O = (0,0)$, then define the *intersection multiplicity* at O by

$$i_O(C,D) = \dim \mathbb{C}[[x,y]]/(f,g),$$

where $\mathbb{C}[[x,y]]$ is the ring of formal power series in x and y. The ring of convergent power series can be used instead, and it would lead to the same result. The multiplicities can be defined at other points by a similar procedure. While this definition is concise, it does not give us much geometric insight. Here is an alternative: $i_p(D,E)$ is the number of points close to p in the intersection of small perturbations of these curves. More precisely, we have the following:

Lemma 1.3.3. $i_p(D,E)$ *is the number of points in* $\{f(x,y) = \varepsilon\} \cap \{g(x,y) = \eta\} \cap B_\delta(p)$ *for small positive* $|\varepsilon|, |\eta|, \delta$*, where* $B_\delta(p)$ *is a* δ*-ball around* p*.*

Proof. This follows from [42, 1.2.5e]. □

There is another nice interpretation of this number worth mentioning. If $K_1, K_2 \subset S^3$ are disjoint knots, perhaps with several components, their linking number is roughly the number of times one of them passes through the other. A precise definition can be found in any basic book on knot theory (e.g., [98]).

Theorem 1.3.4. *Given a small sphere* S^3 *about* p*,* $i_p(D,E)$ *is the linking number of* $D \cap S^3$ *and* $E \cap S^3$*.*

Proof. See [42, 19.2.4]. □

We can now state the strong form of Bézout's theorem. We will revisit this in Corollary 11.2.7.

Theorem 1.3.5 (Bézout's theorem). *Suppose that* C *and* D *are algebraic curves with no common components. Then the sum of intersection multiplicities at points of* $C \cap D$ *equals the product of degrees* $\deg C \cdot \deg D$*, where* $\deg C$ *and* $\deg D$ *are the degrees of the defining polynomials.*

Corollary 1.3.6. *The cardinality* $\#C \cap D$ *is at most* $\deg C \cdot \deg D$*.*

Exercises

1.3.7. Show that the vector space $\mathbb{C}[[x,y]]/(f,g)$ considered above is finite-dimensional if $f=0$ and $g=0$ have an isolated zero at $(0,0)$.

1.3.8. Suppose that $f=y$. Using the original definition show that $i_O(C,D)$ equals the multiplicity of the root $x=0$ of $g(x,0)$. Now prove Bézout's theorem when C is a line.

1.4 Cubics

We now turn our attention to the very rich subject of cubic curves. In the degenerate case, the polynomial factors into a product of a linear and quadratic polynomial or three linear polynomials. Then the curve is a union of a line with a conic or three lines. So now assume that X (1.1.2) is defined by an irreducible cubic polynomial. It is called an *elliptic curve* because of its relationship to elliptic functions and integrals.

Lemma 1.4.1. *After a projective linear transformation, an irreducible cubic can be transformed into the projective closure of an affine curve of the form $y^2 = p(x)$, where $p(x)$ is a cubic polynomial. This is nonsingular if and only if $p(x)$ has no multiple roots.*

Proof. See [105, III §1]. $\square$

We note that nonsingular cubics are very different from conics, even topologically.

Proposition 1.4.2. *A nonsingular cubic X is homeomorphic to a torus $S^1 \times S^1$.*

There is a standard way to visualize this (see Figure 1.3). Mark four points $a,b,c,d=\infty$ on $\mathbb{P}^1$, where the first three are the roots of $p(x)$. Join a to b and c to d by nonintersecting arcs α and β. The preimage of the complement $Y=\mathbb{P}^1-(\alpha\cup\beta)$ in X should fall into two pieces both of which are homeomorphic to Y. So in other words, we can obtain X by first taking two copies of the sphere, slitting them along α and β, and then gluing them along the slits to obtain a torus.

Perhaps that was not very convincing. Instead, we will use a parameterization by elliptic functions to verify Proposition 1.4.2 and more. By applying a further projective linear transformation, we can put our equation for X into Weierstrass form

$$y^2 = 4x^3 - a_2x - a_3 \tag{1.4.1}$$

with discriminant $a_2^3 - 27a_3^2 \neq 0$. The idea is to parameterize the cubic by the elliptic integral

data 16.04.2020
S T Q [✓ Q] S S D

Huybrechts/
[Universitext] Daniel Huybrechts-
Complex Geometry... pdf

Pag. 91 (102)
2.4 The Projective Space

Pag. 182 (192)
4.3 Curvature

Pag. 217 (227)
4.B Hermite-Einstein and
Kähler-Einstein Metrics

I) Books_68_2020/
Complex Geometry An
Inthoduction Huybrechts/
Algebraic Geometry over the
Complex Numbers (2012...).pdf

Pag. 07 (19)
✡ 1.3 Bezout's Theorem
Important:

data 16.04.2020

An important feature of the projective plane is that any two lines meet.

In fact, it has a much stronger property:

Theorem 1.3.1 (Weak Bézout's theorem)

Any two algebraic curves in $\mathbb{P}^2$ intersect

Obs: Usar no meu gráfico

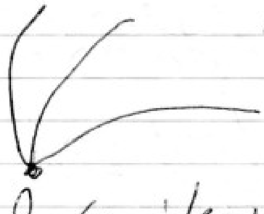

(intersect).

Pág. 70(79)

3.6 Line Bundles on Projective Space.

Pág. 117(125)

Chapter 6

Riemann Surfaces

Pág. 189(193)

Chapter 11

A Little Algebraic Surface

data 16 04 2020

Theory.

Euler characteristic,

$e = b_0 - b_1 + b_2 - b_3 + b_4 =$

$2 - 4g + 2p_g + h^{11}$

J) 20200416 — ✱

Important:

J1) Books_69_2020/ Differential Geometry and Its Applications/

The Geometry of the Octonions by Tevian Dray... pdf

Look at: capa

refere-se ao item

138, letra F

Pág. 39 Figure 11

The Fano Plane (Steiner System)

3 points = 1 line

J2) Books_69_2020/ Euclidean Geometry and Transformations —

' & &! # , = H " E 5 *

3.6 Line Bundles on Projective Space

In this section we define and study certain key examples of sheaves that are of fundamental importance in algebraic geometry. Fix a positive integer n. Let k be an algebraically closed field, $\mathbb{P} = \mathbb{P}^n_k$, and $V = k^{n+1}$. Let $R = k[x_0, \ldots, x_n] = \bigoplus R_d$, where R_d is the subspace of homogeneous polynomials of degree d. Given an integer d, define the subsheaf $\mathscr{O}_{\mathbb{P}}(d)$ of the constant sheaf of rational functions on $\mathbb{P}$ by

$$\mathscr{O}_{\mathbb{P}}(d)(U) \cong \left\{ \frac{f}{g} \mid \exists e, g \in R_e, f \in R_{e+d}, g \text{ nowhere zero on } U \right\}. \qquad (3.6.1)$$

Of course, we have $\mathscr{O}_{\mathbb{P}}(0) = \mathscr{O}_{\mathbb{P}}$ immediately from the definition. We have products

$$\mathscr{O}_{\mathbb{P}}(a) \otimes \mathscr{O}_{\mathbb{P}}(b) \to \mathscr{O}_{\mathbb{P}}(a+b).$$

given by

$$\frac{f}{g} \otimes \frac{f'}{g'} \mapsto \frac{ff'}{gg'}.$$

In particular, this shows that $\mathscr{O}_{\mathbb{P}}(d)$ is an $\mathscr{O}_{\mathbb{P}}$-module.

Lemma 3.6.1.

(a) When $d \geq 0$, the space of global sections of $\mathscr{O}_{\mathbb{P}}(d)$ is R_d.
(b) When $d < 0$, $\mathscr{O}_{\mathbb{P}}(d)$ has no nonzero global sections.
(c) Each $\mathscr{O}_{\mathbb{P}}(a)$ is locally free of rank one.
(d) There are natural isomorphisms $\mathscr{O}_{\mathbb{P}}(a) \otimes \mathscr{O}_{\mathbb{P}}(b) \cong \mathscr{O}_{\mathbb{P}}(a+b)$ and $\mathscr{O}_{\mathbb{P}}(a)^* \cong \mathscr{O}_{\mathbb{P}}(-a)$.

Proof. Statements (a) and (b) are immediate consequences of the definition (3.6.1), since the denominator g is necessarily a nonzero constant. From the definition, we can also identify

$$\mathscr{O}_{\mathbb{P}}(d)(U_i) = x_i^d \mathscr{O}_{\mathbb{P}}(U_i), \qquad (3.6.2)$$

where U_i denotes the complement of $x_i = 0$. The remaining statements follows easily from this. ☐

Part of the importance of these sheaves stems from the following fact:

Lemma 3.6.2. *The ideal sheaf of any degree-$d > 0$ hypersurface in $\mathbb{P}$ (see Example 3.5.4) is isomorphic to $\mathscr{O}_{\mathbb{P}}(-d)$.*

Proof. Given a nonzero degree-d homogeneous polynomial $h \in R_d$, we can see from (3.6.1) that we have an embedding $\mathscr{O}_{\mathbb{P}}(-d) \to \mathscr{O}_{\mathbb{P}}$ given by $f/g \mapsto fh/g$. The image is precisely the ideal sheaf defined by h. ☐

We claimed earlier that locally free sheaves arise from vector bundles. We can see this explicitly in these cases. We treat degree -1 here, and the leave the remaining cases as exercises. Recall that the tautological line bundle (Example 2.6.5) on projective space $\mathbb{P}$ is given by

$$L = \{(x,\ell) \in V \times \mathbb{P} \mid x \in \ell\}$$

with its natural projection to $\mathbb{P}$.

Lemma 3.6.3. *$\mathscr{O}_{\mathbb{P}}(-1)$ is isomorphic to the sheaf $\mathscr{L}$ of sections of L.*

Proof. By definition,

$$\mathscr{L}(U) = \{H : U \to k^{n+1} \mid H \text{ is regular and } H(\ell) \in \ell\}.$$

Thus the components h_i of H are regular functions satisfying

$$h_i(x_0,\ldots,x_n) = \lambda(x_0,\ldots,x_n)x_i$$

for some function λ. Rewriting $\lambda = h_i/x_i$, we see that it lies in $\mathscr{O}_{\mathbb{P}}(-1)(U)$. This process can be reversed. $\square$

When $k = \mathbb{C}$, we can pass to the associated complex analytic line bundle associated with L. Instead of trying to mimic equation (3.6.1) in the analytic category, we simply define $\mathscr{O}_{\mathbb{P}_{\mathrm{an}}}(-1)$ as the sheaf of holomorphic sections of L, and

$$\mathscr{O}_{\mathbb{P}_{\mathrm{an}}}(d) = \begin{cases} \mathscr{O}_{\mathbb{P}_{\mathrm{an}}}(-1) \otimes \cdots \otimes \mathscr{O}_{\mathbb{P}_{\mathrm{an}}}(-1) & (|d|\text{-times}) \quad \text{if } d \le 0, \\ \mathscr{O}_{\mathbb{P}_{\mathrm{an}}}(-1)^* \otimes \cdots \otimes \mathscr{O}_{\mathbb{P}_{\mathrm{an}}}(-1)^* & (d\text{-times}) \quad \text{otherwise.} \end{cases}$$

Note that in the sections where we will be working exclusively in the analytic category, we usually suppress the subscript an.

Exercises

3.6.4. Show that $\mathscr{O}_{\mathbb{P}}(a) \cong \mathscr{O}_{\mathbb{P}}(b)$ implies that $a = b$. Using the notation of Exercise 3.5.21, conclude that $\mathbb{Z} \subset \mathrm{Pic}(\mathbb{P})$.

3.6.5. Let $U = \mathbb{P}^{n+1} - \{[0,\ldots,0,1]\}$ and let $\pi : U \to \mathbb{P}$ denote projection. Show that this is a line bundle, and that $\mathscr{O}_{\mathbb{P}}(1)$ is isomorphic to the sheaf of its sections.

3.6.6. Let $v : \mathbb{P} \to \mathbb{P}^N = \mathbb{P}(S^dV)$ denote the degree-d Veronese map. Show that $\{(x,\ell) \in S^dV \times \mathbb{P} \mid x \in v(\ell)\}$ determines a line bundle over $\mathbb{P}$, and that $\mathscr{O}_{\mathbb{P}}(-d)$ is its sheaf of sections.

3.6.7. Choose a point $p \in \mathbb{P}^1$. Prove that there is an exact sequence

$$0 \to \mathscr{O}_{\mathbb{P}^1_{\mathrm{an}}}(d-1) \to \mathscr{O}_{\mathbb{P}^1_{\mathrm{an}}}(d) \to \mathbb{C}_p \to 0,$$

where $\mathbb{C}_p(U) = \mathbb{C}$ if $p \in U$ and 0 otherwise. Use this to prove that $\dim \Gamma(\mathscr{O}_{\mathbb{P}^1_{\mathrm{an}}}(d)) = d+1$ when $d \ge 0$.

3.7 Direct and Inverse Images

It is useful to transfer a sheaf from one topological space to another via a continuous map $f : X \to Y$. In fact, we have already done this in special cases. We start by explaining how to push a sheaf on X down to Y.

Definition 3.7.1. Given a presheaf $\mathscr{F}$ on X, the direct image $f_*\mathscr{F}$ is a presheaf on Y given by $f_*\mathscr{F}(U) = \mathscr{F}(f^{-1}U)$ with restrictions given by

$$\rho_{f^{-1}Uf^{-1}V} : \mathscr{F}(f^{-1}U) \to \mathscr{F}(f^{-1}V).$$

Lemma 3.7.2. *Direct images of sheaves are sheaves.*

Proof. Suppose that $f : X \to Y$ is a continuous map and $\mathscr{F}$ is a sheaf on X. Let $\{U_i\}$ be an open cover of $U \subseteq Y$, and $s_i \in f_*\mathscr{F}(U_i)$ a collection of sections that agree on the intersections. Then $\{f^{-1}U_i\}$ is an open cover of $f^{-1}U$, and we can regard $s_i \in \mathscr{F}(f^{-1}U_i)$ as a compatible collection of sections for it. Thus we can patch s_i to get a uniquely defined $s \in f_*\mathscr{F}(U) = \mathscr{F}(f^{-1}U)$ such that $s|_{U_i} = s_i$. This proves that $f_*\mathscr{F}$ is a sheaf. □

Now we want to consider the opposite direction. Suppose that $\mathscr{G}$ is a sheaf on Y. We would like to pull it back to X. We will denote this by $f^{-1}\mathscr{G}$, since f^* is reserved for something related to be defined later on. Naively, we can simply try to define

$$f^{-1}\mathscr{G}(U) = \mathscr{G}(f(U)).$$

However, this does not yet make sense unless $f(U)$ is open. So as a first step, given any subset $S \subset Y$ of a topological space and a presheaf $\mathscr{G}$, define

$$\mathscr{G}(S) = \varinjlim \mathscr{G}(V) \tag{3.7.1}$$

as V ranges over all open neighborhoods of S. When S is a point, $\mathscr{G}(S)$ is just the stalk. An element of $\mathscr{G}(S)$ can be viewed as germ of a section defined in a neighborhood of S, where two sections define the same germ if their restrictions agree in a common neighborhood. If $S' \subset S$, there is a natural restriction map $\mathscr{G}(S) \to \mathscr{G}(S')$ given by restriction of germs. So our naive attempt can now be made precise.

Definition 3.7.3. If $\mathscr{G}$ is a presheaf on Y, the presheaf inverse image $f^P\mathscr{G}$ is a presheaf on X given by $f^P\mathscr{G}(U) = \mathscr{G}(f(U))$ with restrictions as above. The sheaf inverse image is the sheafification $f^{-1}\mathscr{G} = (f^P\mathscr{G})^+$.

Although the definition of f^{-1} seems much more complicated compared to f_*, it is in fact just as natural. The picture becomes clearer if we view the inverse image in terms of its stalks. We then have

$$(f^{-1}\mathscr{G})_x \cong \mathscr{G}_{f(x)}, \tag{3.7.2}$$

and there is more. As explained in Exercise 3.1.22, the sheaf $\mathscr{G}$ determines the étalé space $\pi : \mathrm{Et}(\mathscr{G}) = \bigcup \mathscr{G}_y \to Y$. The étalé space for $f^{-1}\mathscr{G}$ is the fiber product

$$\mathrm{Et}(f^{-1}\mathscr{G}) = \{(x,g) \in X \times \mathrm{Et}(\mathscr{G}) \mid f(x) = \pi(g)\},$$

so that $f^{-1}\mathscr{G}$ is the sheaf of continuous cross sections of the projection $\mathrm{Et}(f^{-1}\mathscr{G}) \to X$.

If $f : X \to Y$ is the inclusion of a closed set, we also call $f_*\mathscr{F}$ the extension of $\mathscr{F}$ by 0 and $f^{-1}\mathscr{G}$ the restriction of $\mathscr{G}$. We often identify $\mathscr{F}$ with its extension, and denote the restriction by $\mathscr{G}|_X$.

These operations extend to functors $f_* : \mathrm{Ab}(X) \to \mathrm{Ab}(Y)$ and $f^{-1} : \mathrm{Ab}(Y) \to \mathrm{Ab}(X)$ in an obvious way. While these operations are generally not inverses, there is a relationship. There are canonical morphisms, called adjunctions, $\alpha : \mathscr{F} \to f_*f^{-1}\mathscr{F}$ and $\beta : f^{-1}f_*\mathscr{G} \to \mathscr{G}$ induced by the restriction maps

$$\begin{aligned} \mathscr{F}(U) \to \mathscr{F}(f(f^{-1}(U))) = f_*f^P\mathscr{F}(U) \to f_*f^{-1}\mathscr{F}(U), \quad U \subseteq Y, \\ f^Pf_*\mathscr{G}(V) = \mathscr{G}(f^{-1}(f(V))) \to \mathscr{G}(V), \quad V \subseteq X. \end{aligned}$$

Lemma 3.7.4. *There is a natural isomorphism*

$$\mathrm{Hom}_{\mathrm{Ab}(X)}(f^{-1}\mathscr{F}, \mathscr{G}) \cong \mathrm{Hom}_{\mathrm{Ab}(Y)}(\mathscr{F}, f_*\mathscr{G})$$

given by $\eta \mapsto (f_*\eta) \circ \alpha$. *The inverse is given by* $\xi \mapsto \beta \circ (f^{-1}\xi)$.

In the language of category theory, this says that the functor f^{-1} is left adjoint to f_*. The proof will be left as an exercise.

The collection of ringed spaces will form a category. To motivate the definition of morphism, observe that from a morphism F of concrete k-spaces $(X,\mathscr{R}) \to (Y,\mathscr{S})$, we get a morphism of sheaves of rings $\mathscr{S} \to F_*\mathscr{R}$ given by $s \mapsto s \circ F$.

Definition 3.7.5. A morphism of (k-)ringed spaces $(X,\mathscr{R}) \to (Y,\mathscr{S})$ is a continuous map $F : X \to Y$ together with a morphism of sheaves of rings (or algebras) $F' : \mathscr{S} \to F_*\mathscr{R}$.

By Lemma 3.7.4, to give F' is equivalent to giving the adjoint map $F^{-1}\mathscr{S} \to \mathscr{R}$. Therefore we have the induced ring homomorphism $\mathscr{S}_{F(x)} \to \mathscr{R}_x$ for each $x \in X$.

Definition 3.7.6. If $(X,\mathscr{R})$ and $(Y,\mathscr{S})$ are locally ringed, a morphism $F : (X,\mathscr{R}) \to (Y,\mathscr{S})$ is called a morphism of locally ringed spaces if it satisfies the additional requirement that the induced maps $\mathscr{S}_{F(x)} \to \mathscr{R}_x$ be local homomorphisms. A morphism of schemes is simply a morphism of locally ringed spaces

For affine schemes, we have the following description.

Proposition 3.7.7. *Let R and S be affine algebras. There is a bijection*

$$\mathrm{Hom}_{\mathrm{Schemes}}((\mathrm{Spec}_m S, \mathscr{O}_{\mathrm{Spec}_m S}), (\mathrm{Spec}_m R, \mathscr{O}_{\mathrm{Spec}_m R})) \cong \mathrm{Hom}_{k\text{-alg}}(R,S).$$

Therefore there is an antiequivalence between the category of affine schemes and affine algebras.

Proof. Let $X = \mathrm{Spec}_m S$ and $Y = \mathrm{Spec}_m R$. Given a homomorphism $f : R \to S$, we just construct the corresponding morphism $F = f^* : X \to Y$. If $m \subset S$ is a maximal ideal, then $f^{-1}m$ is again a maximal ideal by Hilbert's Nullstellensatz. So we can define $F(m) = f^{-1}m$. It is easy to see that F is continuous. The induced homomorphisms $R[1/r] \to S[1/f(r)]$ yield morphisms $\mathscr{O}_Y \to F_*\mathscr{O}_X$ of sheaves of rings on basic open sets, and therefore on their extensions. The maps on stalks are the induced local homomorphisms $f_m : R_{F(m)} \to S_m$. Therefore F is a morphism of schemes.

In the opposite direction, a morphism $F : X \to Y$ induces a homomorphism $R = \mathscr{O}(Y) \to \mathscr{O}(X) = S$. This yields the inverse. The remaining details will be left as an exercise. □

The simplest example of a scheme that is not a variety is the fat point $\mathrm{Spec}_m k[\varepsilon]/(\varepsilon^2)$. This has the following interesting property.

Lemma 3.7.8. *If X is a scheme, the set of morphisms from $\mathrm{Spec}_m k[\varepsilon]/(\varepsilon^2)$ to X is in one-to-one correspondence with the set of vectors in the disjoint union of tangent spaces $\cup T_x$.*

Proof. The scheme $D = \mathrm{Spec}_m k[\varepsilon]/(\varepsilon^2)$ has a unique point $m = (\varepsilon)$. We show that the set of morphisms $i : D \to X$ with $i(m) = x$ is in one to one correspondence with T_x. To give such a morphism is equivalent to giving a morphism of D to a neighborhood of x. Therefore, we may replace X by an affine scheme $\mathrm{Spec}_m R$. By Proposition 3.7.7, $D \to \mathrm{Spec}_m R$ corresponds to an algebra homomorphism $i^* : R \to k[\varepsilon]/(\varepsilon^2)$ such that $(i^*)^{-1}m = m_x$. The map i^* factors through R/m_x^2. We can see that any homomorphism

$$h : R/m_x^2 = k \oplus m_x/m_x^2 \to k[\varepsilon]/(\varepsilon^2) = k \oplus k\varepsilon$$

is given by $h(a,b) = a + v(b)\varepsilon$ for $v \in T_x = (m_x/m_x^2)^*$. The argument can be reversed to show that conversely, any vector in $v \in T_x$ determines an algebra map $i^* : R \to k[\varepsilon]/(\varepsilon^2)$ for which $(i^*)^{-1}m = m_x$. □

Suppose that $f : (X, \mathscr{R}) \to (Y, \mathscr{S})$ is a morphism of ringed spaces. Given an $\mathscr{R}$-module $\mathscr{M}$, $f_*\mathscr{M}$ is naturally an $f_*\mathscr{R}$-module, and hence an $\mathscr{S}$-module by restriction of scalars via $\mathscr{S} \to f_*\mathscr{R}$. Similarly, given an $\mathscr{S}$-module $\mathscr{N}$, $f^{-1}\mathscr{N}$ is naturally an $f^{-1}\mathscr{S}$-module. We define the $\mathscr{R}$-module

$$f^*\mathscr{N} = \mathscr{R} \otimes_{f^{-1}\mathscr{S}} f^{-1}\mathscr{N},$$

where the $\mathscr{R}$ is regarded as an $f^{-1}\mathscr{S}$-module under the adjoint map $f^{-1}\mathscr{S} \to \mathscr{R}$. When f is injective, we often write $\mathscr{N}|_X$ instead of $f^*\mathscr{N}$.

The inverse image of a locally free sheaf is easily seen to be locally free. This has an interpretation in the context of vector bundles. If $\pi : V \to Y$ is a vector bundle, the pullback $f^*V \to X$ is the vector bundle given set theoretically as the projection

$$f^*V = \{(v,x) \,|\, \pi(v) = f(x)\} \to X.$$

Then

$$f^*(\mathscr{V}) \cong (\text{sheaf of sections of } f^*V). \tag{3.7.3}$$

A special case is that in which $X = \{y\} \to Y$ is the inclusion of a point. Then the fiber $\pi^{-1}y$ can be identified with $f^*(\mathscr{V}) = (\mathscr{R}_y/m_y) \otimes \mathscr{V}_y$.

Exercises

3.7.9. Verify (3.7.2).

3.7.10. Prove Lemma 3.7.4.

3.7.11. Give examples where $\mathscr{F} \to f_*f^{-1}\mathscr{F}$ and $f^{-1}f_*\mathscr{G} \to \mathscr{G}$ are not isomorphisms.

3.7.12. Complete the proof of Proposition 3.7.7.

3.7.13. Given an inclusion of an open set $j : U \to X$ and a sheaf $\mathscr{F}$ on U, define $j_!\mathscr{F}$ to be the sheafification of the presheaf

$$\mathscr{G}(V) = \begin{cases} \mathscr{F}(V) & \text{if } V \subseteq U, \\ 0 & \text{otherwise.} \end{cases}$$

If $\mathscr{S}$ is a sheaf on X, show that this fits into an exact sequence

$$0 \to j_!j^{-1}\mathscr{S} \to \mathscr{S} \to i_*i^{-1}\mathscr{S} \to 0,$$

where $i : X - U \to X$ is the inclusion of the complement.

3.7.14. Generalize Lemma 3.2.5 to show that an exact sequence $0 \to \mathscr{A} \to \mathscr{B} \to \mathscr{C} \to 0$ of sheaves gives rise to an exact sequence $0 \to f_*\mathscr{A} \to f_*\mathscr{B} \to f_*\mathscr{C}$. In other words, f_* is left exact.

3.7.15. Prove that f^{-1} is exact, i.e., that it takes exact sequences to exact sequences.

3.7.16. Verify the assertion given in (3.7.3).

3.7.17. Suppose we have gluing data (R_i, ϕ_{ij}) and (R'_i, ϕ'_{ij}) and homomorphisms $f_i : R_i \to R'_i$ such that $f_i\phi'_{ij} = \phi_{ij}f_j$. Show that there is an induced morphism of glued schemes extending $f_i^* : \mathrm{Spec}_m R'_i \to \mathrm{Spec}_m R_i$. Use this to justify Remark 3.4.11.

3.7.18. Let L_d be obtained by gluing $\mathrm{Spec}_m k[x,y]$ to $\mathrm{Spec}_m k[x^{-1}, x^d y]$. Show that L_d is a line bundle over $\mathbb{P}^1_k$ whose sheaf of sections is isomorphic to $\mathscr{O}(\pm d)$.

3.7.19. Let $f : \mathbb{P}^{n-1}_k \to \mathbb{P}^n_k$ be a linear embedding, given for example by setting $x_n = 0$. Show that $f^*\mathscr{O}_{\mathbb{P}^n}(d) = \mathscr{O}_{\mathbb{P}_{n-1}}(d)$.

3.7.20. Let $v : \mathbb{P}^n_k \to \mathbb{P}^N_k$ be the dth Veronese embedding. Show that $v^*\mathscr{O}_{\mathbb{P}^N}(1) = \mathscr{O}_{\mathbb{P}^n}(d)$.

3.7.21. An n-jet on a scheme X is a morphism from $\operatorname{Spec} k[\varepsilon]/(\varepsilon^{n+1})$ to X. Give an interpretation of these similar to Lemma 3.7.8.

3.8 Differentials

With basic sheaf theory in hand, we can now construct sheaves of differential forms on manifolds and varieties in a unified way. In order to motivate things, let us start with a calculation. Suppose that $X = \mathbb{R}^n$ with coordinates $x_1, \ldots, x_n$. Given a C^∞ function f on X, we can develop a Taylor expansion about $(y_1, \ldots, y_n)$:

$$f(x_1, \ldots, x_n) = f(y_1, \ldots, y_n) + \sum \frac{\partial f}{\partial x_i}(y_1, \ldots, y_n)(x_i - y_i) + O((x_i - y_i)^2).$$

Thus the differential is given by

$$df = f(x_1, \ldots, x_n) - f(y_1, \ldots, y_n) \mod (x_i - y_i)^2.$$

We can view $x_1, \ldots, x_n, y_1, \ldots, y_n$ as coordinates on $X \times X = \mathbb{R}^{2n}$, so that $x_i - y_i = 0$ defines the diagonal Δ. Then df lies in the ideal of Δ modulo its square.

Let X be a C^∞ or complex manifold or an algebraic variety over a field k. We take $k = \mathbb{R}$ or $\mathbb{C}$ in the first two cases. We have a diagonal map $\delta : X \to X \times X$ given by $x \mapsto (x, x)$, and projections $p_i : X \times X \to X$. Let $\mathscr{I}_\Delta$ be the ideal sheaf of the image of δ, and let $\mathscr{I}_\Delta^2 \subseteq \mathscr{I}_\Delta$ be the sub–ideal sheaf locally generated by products of pairs of sections of $\mathscr{I}_\Delta$. Then we define the sheaf of 1-forms by

$$\Omega^1_X = (\mathscr{I}_\Delta / \mathscr{I}_\Delta^2)|_\Delta.$$

This has two different $\mathscr{O}_X$-module structures; we pick the first one, where $\mathscr{O}_X$ acts on $\mathscr{I}_\Delta$ through $p_1^{-1}\mathscr{O}_X \to \mathscr{O}_{X \times X}$. We define the sheaf of p-forms by $\Omega^p_X = \wedge^p \Omega^1_X$. We define a morphism $d : \mathscr{O}_X \to \Omega^1_X$ of sheaves (but not $\mathscr{O}_X$-modules) by $df = p_1^*(f) - p_2^*(f)$. This has the right formal properties because of the following:

Lemma 3.8.1. *d is a k-linear derivation.*

Proof. By direct calculation,

$$d(fg) - fdg - gdf = [p_1^*(f) - p_2^*(f)][p_2^*(g) - p_2^*(g)] \in \mathscr{I}_\Delta^2. \qquad \square$$

Differentials behave contravariantly with respect to morphisms. Given a morphism of manifolds or varieties $f : Y \to X$, we get a morphism $Y \times Y \to X \times X$

' & * , 1

Chapter 6
Riemann Surfaces

Riemann surfaces are the same thing as one-dimensional complex manifolds or nonsingular complex curves. Although the basic concept goes back to Riemann, the rigorous definition appears to be due to Weyl [122]. We already considered these briefly in the first chapter. But now we are in a position to make a more thorough study using the geometric and homological tools introduced in the intervening chapters. Nevertheless, we will only scratch the surface of this rich subject. Further details can be found in [5, 20, 49, 38] in addition to Weyl's book.

6.1 Genus

The coarsest classification is topological. A Riemann surface can be regarded as a manifold of real dimension 2. It has a canonical orientation: if we identify the real tangent space at any point with the complex tangent space, then for any nonzero vector v, we declare the ordered basis (v, iv) to be positively oriented. Let us now forget the complex structure and consider the purely topological problem of classifying these surfaces up to homeomorphism.

Given two connected 2-dimensional topological manifolds X and Y with points $x \in X$ and $y \in Y$, we can form a new topological manifold $X\#Y$, called the connected sum. To construct this, choose open disks $D_1 \subset X$ and $D_2 \subset Y$. Then $X\#Y$ is obtained by gluing $X - D_1 \cup S^1 \times [0,1] \cup Y - D_2$ appropriately. The homeomorphism class is independent of the choices made. Figure 6.1 depicts the connected sum of two tori.

Theorem 6.1.1. *A compact connected orientable 2-dimensional topological manifold is classified, up to homeomorphism, by a nonnegative integer called the genus. A genus-*0 *surface is homeomorphic to the 2-sphere* S^2*. A manifold of genus* $g > 0$ *is homeomorphic to a connected sum of the 2-torus and a surface of genus* $g - 1$*.*

Proof. An equivalent formulation is that every such manifold is homeomorphic to a sphere with g handles. A classical reference for this is [99, p. 145]. ☐

D. Arapura, *Algebraic Geometry over the Complex Numbers*, Universitext,
DOI 10.1007/978-1-4614-1809-2_6, © Springer Science+Business Media, LLC 2012

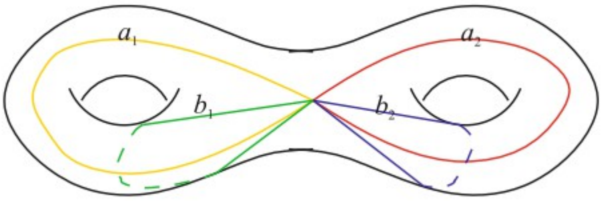

Fig. 6.1 Genus-2 surface.

There is another standard model for these surfaces [99] that is also quite useful (for instance for computing the fundamental group). A genus-g surface can be constructed by gluing the sides of a $2g$-gon. It is probably easier to visualize this in reverse. After cutting the genus-2 surface of Figure 6.1 along the indicated curves, it can be opened up to an octagon (see Figure 6.2).

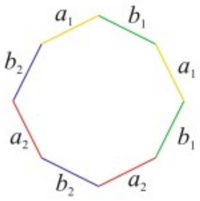

Fig. 6.2 Genus-2 surface cut open.

The topological Euler characteristic of the space X is

$$e(X) = \sum(-1)^i \dim H^i(X,\mathbb{R}).$$

From Exercise 4.5.5, we have the following lemma:

Lemma 6.1.2. *If X is a union of two open sets U and V, then $e(X) = e(U) + e(V) - e(U \cap V)$.*

Corollary 6.1.3. *If X is a manifold of genus g, then $e(X) = 2 - 2g$, and the first Betti number is given by $\dim H^1(X,\mathbb{R}) = 2g$.*

Proof. This will be left for the exercises. □

When $g = 2$, this gives $\dim H^1(X,\mathbb{R}) = 4$. We can find explicit generators by taking the fundamental classes of the curves a_1, a_2, b_1, b_2 in Figure 6.1, after choosing orientations. To see that these generate, $H^1(X,\mathbb{R})$, it suffices to prove that they are linearly independent. For this, consider the pairing

$$(\alpha,\beta) \mapsto \int_X \alpha \wedge \beta$$

on $H^1(X,\mathbb{R})$. This restricts to the intersection pairing on fundamental classes. After orienting the curves suitably, their intersection matrix is

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix},$$

and this shows independence. A similar basis with intersection matrix

$$\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$$

can be found for any g. These basis vectors will generate a lattice inside $H^1(X,\mathbb{R})$ that can be identified with $H^1(X,\mathbb{Z})$.

To answer the obvious question, we show that every genus occurs by direct construction. A much harder question is, what does a typical genus g look like? Many examples of compact Riemann surfaces can be given explicitly as nonsingular projective algebraic curves, and we start with this approach. Later, we will see that all compact Riemann surfaces arise in this way.

Example 6.1.4. Let $f(x,y,z)$ be a homogeneous polynomial of degree d. Suppose that the partials of f have no common zeros in $\mathbb{C}^3$ except $(0,0,0)$. Then the curve $V(f) = \{f(x,y,z) = 0\}$ in $\mathbb{P}^2_{\mathbb{C}}$ is nonsingular. We will see later that the genus is $\frac{(d-1)(d-2)}{2}$. In particular, not every genus occurs for these examples.

Example 6.1.5. Given a collection of homogeneous polynomials $f_i \in \mathbb{C}[x_0,\dots,x_n]$ such that $X = V(f_1, f_2, \dots) \subset \mathbb{P}^n_{\mathbb{C}}$ is a nonsingular algebraic curve, then X will be a complex submanifold of $\mathbb{P}^n_{\mathbb{C}}$ and hence a Riemann surface. By a generic projection argument, $n = 3$ suffices to give all such examples.

Example 6.1.6. Choose $2g+2$ distinct points in $a_i \in \mathbb{C}$. Consider the affine curve $C_1 \subset \mathbb{C}^2$ defined by

$$y^2 = \prod (x - a_i).$$

We can compactify this to obtain a nonsingular projective curve C by taking the projective closure and normalizing. Here is an alternative description of the same curve. Consider the affine curve C_2:

$$Y^2 = \prod (1 - a_i X).$$

Glue this to C_1 by identifying $X = x^{-1}$ and $Y = yx^{-g-1}$. Since C is nonsingular, it can be viewed as a Riemann surface. By construction, C comes equipped with a morphism $f : C \to \mathbb{P}^1_{\mathbb{C}}$ that is 2-to-1 except at the branch points $\{a_i\}$. In the exercises, it will be shown that the genus of X is g. Curves of this form are called hyperelliptic (if $g > 1$), and they are very nice to work with. However, "most" curves are not hyperelliptic.

From a more analytic point of view, we can construct many examples as quotients of $\mathbb{C}$ or the upper half-plane. In fact, the *uniformization theorem* [38] tells us that all examples other than $\mathbb{P}^1$ arise in this way.

Example 6.1.7. Let $L \subset \mathbb{C}$ be a lattice, i.e., an abelian subgroup generated by two $\mathbb{R}$-linearly independent numbers. The quotient $E = \mathbb{C}/L$ can be made into a Riemann surface (Exercise 2.2.20) called an elliptic curve. Since this is topologically a torus, the genus is 1. Conversely, any genus-1 curve is of this form.

The surface $E = \mathbb{C}/L$ can be realized as an algebraic curve in a couple of ways. As we saw in the first chapter, we can embed it into $\mathbb{P}^2$ as the cubic curve

$$zy^2 = 4x^3 - g_2(L)xz^2 - g_3(L)z^3 \tag{6.1.1}$$

by $z \mapsto [\wp(z), \wp'(z), 1]$. Alternatively, $\wp : E \to \mathbb{P}^1$ realizes this directly as a two-sheeted cover branched at four points. We can assume without loss of generality that these are $0, 1, \infty, t$. Then E is given by Legendre's equation

$$y^2z = x(x-z)(x-tz). \tag{6.1.2}$$

Let us now consider quotients of the upper half-plane $H = \{z \mid Im(z) > 0\}$. The group $\mathrm{SL}_2(\mathbb{R})$ acts on H by fractional linear transformations:

$$z \mapsto \frac{az+b}{cz+d}.$$

Note that $-I$ acts trivially, so the action factors through $\mathrm{PSL}_2(\mathbb{R}) = \mathrm{SL}_2(\mathbb{R})/\{\pm I\}$. The action of subgroup $\Gamma \subset \mathrm{SL}_2(\mathbb{R})$ on H is *properly discontinuous* if every point has a neighborhood V such that $\gamma V \cap V \neq \emptyset$ for all but finitely many $\gamma \in \Gamma$. A point $x \in V$ is called a *fixed point* if its stabilizer in $\Gamma/\{\pm I\}$ is nontrivial. The action is *free* if it has no fixed points.

Proposition 6.1.8. *If Γ acts properly discontinuously on H, the quotient $X = H/\Gamma$ becomes a Riemann surface. If $\pi : H \to X$ denotes the projection, the structure sheaf is defined by $f \in \mathscr{O}_X(U)$ if and only if $f \circ \pi \in \mathscr{O}_H(\pi^{-1}U)$.*

When Γ acts freely with compact quotient, then it has genus $g > 1$. The quickest way to see this is by applying the Gauss–Bonnet theorem [56], which says that $e(H/\Gamma) = 2 - 2g$ can be computed by integrating the Gaussian curvature for any Riemannian metric. The hyperbolic metric on H, which descends to the quotient, has negative curvature. A *fundamental domain* for this action is a region $R \subset H$ such that $\cup_\gamma \gamma \bar{R} = H$ and such that two translates of $\bar{R}$ can meet only at their boundaries. The fundamental domain in this case can be chosen to be the interior of a geodesic $2g$-gon.

The modular group $\mathrm{SL}_2(\mathbb{Z})$ is a particularly important example where the action is not free.

Theorem 6.1.9. *The quotient $H/\mathrm{SL}_2(\mathbb{Z})$ is isomorphic to $\mathbb{C}$, and its points correspond to isomorphism classes of elliptic curves (which are genus-one curves with a distinguished point).*

Proof. We sketch the idea. The details can be found in [32, 106] or [103]. A fundamental domain for the action is given in Figure 6.3. The marked points in the diagram are $\rho = e^{2\pi\imath/3}, i$, and $\rho + 1$.

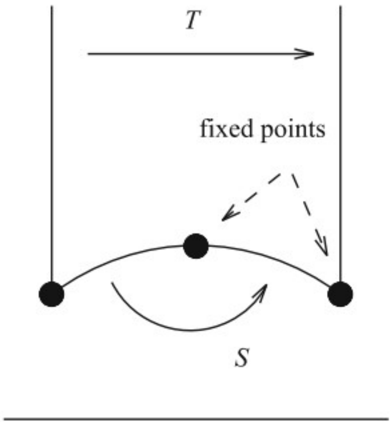

Fig. 6.3 Fundamental domain of $\mathrm{SL}_2(\mathbb{Z})$.

After identifying the sides of the domain by the transformations

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

as indicated in the diagram, one sees that that the quotient $H/\mathrm{SL}_2(\mathbb{Z})$ is homeomorphic to $\mathbb{C}$. By the Riemann mapping theorem, $H/\mathrm{SL}_2(\mathbb{Z})$ is isomorphic to either $\mathbb{C}$ or H. To see that it is $\mathbb{C}$, we show that the one-point compactification $H/\mathrm{SL}_2(\mathbb{Z}) \cup \{*\}$ can also be made into a Riemann surface, necessarily isomorphic to $\mathbb{P}^1$. Any continuous function f defined in a neighborhood of $*$ can be pulled back to an $\mathrm{SL}_2(\mathbb{Z})$-invariant function $\tilde{f}$ on H. Invariance under T implies that $\tilde{f}$ is periodic; hence it can be expanded in a Fourier series

$$\tilde{f}(z) = \sum_{n=-\infty}^{\infty} a_n e^{2\pi i n z}.$$

We declare f to be holomorphic if all the a_n are zero for $n < 0$. To put this another way, $q = e^{2\pi i z}$ is taken as a local analytic coordinate at $*$.

The relation to elliptic curves is given in the exercises. ☐

Exercises

6.1.10. Prove Corollary 6.1.3.

Chapter 11
A Little Algebraic Surface Theory

Let us return to geometry armed with what we have learned so far. We have already looked at Riemann surfaces (which will be referred to as complex curves from now on) in some detail. So we consider the next step up. A nonsingular complex surface is a two-dimensional complex manifold. By an algebraic surface, we will mean a two-dimensional nonsingular projective variety. So in particular, they are Kähler manifolds. In this chapter, we will present a somewhat breezy account of surface theory, concentrating on topics that illustrate the general theorems from the previous chapters.

Much more systematic introductions to algebraic surface theory can be found in the books by Barth, Peters, and Van de Ven [9] and Beauville [10].

11.1 Examples

The basic discrete invariant of a curve is its genus. For algebraic surfaces, there are several numbers that play a similar role. We can use the Hodge numbers. By the symmetry properties considered earlier, there are only three that matter: h^{10}, h^{20}, h^{11}. The first two are traditionally called (and denoted by) the *irregularity* ($q = h^{10}$) and *geometric genus* ($p_g = h^{20}$). The basic topological invariants are the Betti numbers, which by Poincaré duality and the Hodge decomposition theorem can be expressed as

$$b_1 = b_3 = 2q, \quad b_2 = 2p_g + h^{11}.$$

It is also convenient to consider the Euler characteristic,

$$e = b_0 - b_1 + b_2 - b_3 + b_4 = 2 - 4q + 2p_g + h^{11}.$$

Let us now start our tour.

D. Arapura, *Algebraic Geometry over the Complex Numbers*, Universitext,
DOI 10.1007/978-1-4614-1809-2_11, © Springer Science+Business Media, LLC 2012

Example 11.1.1. The most basic example is the projective plane $X = \mathbb{P}^2 = \mathbb{P}^2_{\mathbb{C}}$. We computed the Betti numbers $b_1 = 0, b_2 = 1$ in Section 7.2. Therefore $q = p_g = 0$ and $h^{11} = 1$.

Example 11.1.2. If $X = C_1 \times C_2$ is a product of two nonsingular curves of genus g_1 and g_2, then by Künneth's formula (Theorem 5.3.6 and Corollary 10.2.6),

$$q = h^{10}(X) = h^{10}(C_1)h^{00}(C_2) + h^{00}(C_1)h^{10}(C_2) = g_1 + g_2.$$

Similarly, $p_g = g_1 g_2$ and $h^{11} = 2g_1 g_2 + 2$.

Example 11.1.3. As a special case of the previous example, when we have a product of a curve C with $\mathbb{P}^1$, the invariants are $q = g$, $p_g = 0$, and $h^{11} = 2$. More generally, we can consider ruled surfaces over C, which are $\mathbb{P}^1$-bundles that are locally isomorphic to $U_i \times \mathbb{P}^1$, for a Zariski open cover $\{U_i\}$ of C. (See Section 14.5 for a bit more explanation of what this means.) We will see shortly that the invariants are the same as above, although they are not generally products. When $C = \mathbb{P}^1$, there are, up to isomorphism, countably many ruled surfaces. Here is a simple description. Let $C_n \subset \mathbb{P}^n$ be the closure of $\{(t, t^2, \dots, t^n) \mid t \in \mathbb{C}\}$. Choose a point $p_0 \in \mathbb{P}^n - C_n$. Let F_n be the set of pairs $(q, p) \in \mathbb{P}^n \times C_n$ such that q lies on the line connecting p_0 to q.

Example 11.1.4. Let $X \subset \mathbb{P}^3$ be a smooth surface of degree d. Then $q = 0$. We will list the first few values of the remaining invariants:

d	p_g	h^{11}
2	0	2
3	0	7
4	1	20
5	4	45
6	10	86

These can be calculated using formulas given later (17.3.4).

A method of generating new examples from old is by blowing up. We start by describing the blowup of $\mathbb{C}^2$ at 0:

$$\mathrm{Bl}_0\mathbb{C}^2 = \{(x, \ell) \in \mathbb{C}^2 \times \mathbb{P}^1 \,|\, x \in \ell\}.$$

The projection $p_1 : \mathrm{Bl}_0\mathbb{C}^2 \to \mathbb{C}^2$ is one-to-one away from $0 \in \mathbb{C}^2$. This can be generalized to yield the blowup $\mathrm{Bl}_pX \to X$ of a surface X at the point p. Let $B \subset X$ be a coordinate ball centered at p. After identifying B with a ball in $\mathbb{C}^2$ centered at 0, we can let Bl_0B be the preimage of B in $\mathrm{Bl}_0\mathbb{C}^2$. The boundary of Bl_0B can be identified with the boundary of B. Thus we can glue $X - B \cup \mathrm{Bl}_0B$ to form Bl_pX. When X is algebraic, Bl_pX is again algebraic by Exercise 2.4.23.

Let us compute $H^*(\mathrm{Bl}_pX, \mathbb{Z})$. Set $Y = \mathrm{Bl}_pX$ and compare Mayer–Vietoris sequences:

$$\begin{tikzcd}
H^i(X) \arrow[r] \arrow[d] & H^i(X-B')\oplus H^i(B) \arrow[r] \arrow[d] & H^i(X-B'\cap B) \arrow[d, "="] \\
H^i(Y) \arrow[r] & H^i(Y-\mathrm{Bl}_pB')\oplus H^i(\mathrm{Bl}_pB) \arrow[r] & H^i(Y-\mathrm{Bl}_pB'\cap \mathrm{Bl}_pB)
\end{tikzcd}$$

where $B' \subset B$ is a smaller ball. We thus have the following result:

Lemma 11.1.5. *$H^1(\mathrm{Bl}_pX) \cong H^1(X)$ and $H^2(\mathrm{Bl}_pX) = H^2(X) \oplus \mathbb{Z}$.*

Corollary 11.1.6. *q and p_g are invariant under blowing up. $h^{11}(\mathrm{Bl}_pX) = h^{11}(X)+1$.*

Proof. The lemma implies that $b_1 = 2q$ is invariant and $b_2(Y) = b_1(X)+1$. Since $b_2 = 2p_g + h^{11}$, the only possibilities are $h^{11}(Y) = h^{11}(X)+1$, $p_g(Y) = p_g(X)$, and $p_g(Y) < p_g(X)$. The last inequality means that there is a nonzero holomorphic 2-form on X that vanishes on $X-p$, but this is impossible. ☐

A *birational map* $\kappa : X \dashrightarrow Y$ is simply an isomorphism in the category of varieties and rational maps. In more explicit terms, it is given by an isomorphism of Zariski open sets $X \supset U \cong V \subset Y$. Blowups and their inverses ("blowdowns") are examples of birational maps. Two varieties are birationally equivalent if a birational map exists between them. For example, any two ruled surfaces over $\mathbb{P}^1$ are birationally equivalent to each other and to $\mathbb{P}^2$, because they all contain $\mathbb{A}^2$.

For surfaces, the structure of birational maps is explained by the following theorem:

Theorem 11.1.7 (Castelnuovo). *Any birational map between algebraic surfaces is given by a finite sequence of blowups and blowdowns.*

Proof. See [60, 9]. ☐

Corollary 11.1.8. *The numbers q and p_g depend only on the birational equivalence class of the surface.*

This implies that $p_g = g, q = 0$ for ruled surfaces over a genus-g curve, as claimed above.

Blowing up of singular points figures in the proof of the next important theorem.

Theorem 11.1.9 (Zariski). *Given a singular algebraic surface Y, there exist a nonsingular surface X and a morphism $\pi : X \to Y$, called a resolution of singularities, that is an isomorphism away from the singular points.*

Proof. See [9]. ☐

Corollary 11.1.10 (Zariski). *If $f : X \dashrightarrow V$ is a rational function from an algebraic surface to a variety V, then there is a finite sequence of blowups $Y \to X$ such that f extends to a holomorphic map $f' : Y \to V$.*

Proof. We can construct Y by resolving singularities of the closure of the graph of f. $\square$

Analogues of Zariski's and Castelnuovo's theorems in higher dimensions have been established by Hironaka and Włodarczyk respectively. These are much harder.

An *elliptic surface* is a surface X that admits a surjective morphism $f : X \to C$ to a smooth projective curve such that all nonsingular fibers are elliptic curves.

Example 11.1.11. A simple example of an elliptic surface is given as follows: choose two distinct nonsingular cubics $E_0, E_1 \subset \mathbb{P}^2$ defined by $f_0(x,y,z)$ and $f_1(x,y,z)$. These generate a pencil of cubics $E_t = V(tf_1 + (1-t)f_0)$ with $t \in \mathbb{P}^1$. Define

$$X = \{(p,t) \in \mathbb{P}^2 \times \mathbb{P}^1 \mid p \in E_t\}.$$

Projection to $\mathbb{P}^1$ makes this an elliptic surface. X can be identified with the blowup of $\mathbb{P}^2$ at the nine points $E_0 \cap E_1$. So $q = p_g = 0$ and $h^{11} = 10$.

Example 11.1.12. Consider the family of elliptic curves in Legendre form

$$\mathscr{E} = \{([x,y,z],t) \in \mathbb{P}^2 \times \mathbb{C} - \{0,1\} \mid y^2z - x(x-z)(x-tz) = 0\} \to \mathbb{C}.$$

The above equation is meaningful if $t = 0, 1$, and it defines a rational curve with a single node. By introducing $s = t^{-1}$, we get an equation

$$sy^2z - x(x-z)(sx-z) = 0$$

that defines a union of lines when $s = 0$. In this way, we can extend $\mathscr{E}$ to a surface $\mathscr{E}' \to \mathbb{P}^1$. Unfortunately, $\mathscr{E}'$ is singular, and it is necessary to resolve singularities to get a nonsingular surface $\bar{\mathscr{E}}$ containing $\mathscr{E}$ (we can take the minimal desingularization, which for our purposes means that $b_2(\bar{\mathscr{E}})$ is chosen as small as possible).

Exercises

11.1.13. Finish the proof of Lemma 11.1.5.

11.1.14. Given a ruled surface X over a curve C, check that $e(X) = e(\mathbb{P}^1)e(C)$, and use this to verify that $h^{11}(X) = 2$.

11.1.15. Show that there is a nonsingular quartic $X \subset \mathbb{P}^3$ containing a line ℓ, which we can assume to be $x_2 = x_3 = 0$. Show that the map $\mathbb{P}^3 \dashrightarrow \mathbb{P}^1$ defined by $[x_0, \ldots, x_3] \mapsto [x_0, x_1]$ determines a morphism $X \to \mathbb{P}^1$ that makes it an ellptic surface.

11.1.16. Given a smooth projective curve C, the symmetric product is given by $S^2C = C \times C/\sigma$, where σ is the involution interchanging factors. This has the structure of a smooth algebraic surface such that $H^i(S^2C, \mathbb{Q}) = H^i(C \times C, \mathbb{Q})^\sigma$. Compute the Betti and Hodge numbers for S^2C.

E /012/ / /3 / #! > ? + 3 -+
> 5 K +@ 0 G - *"
= . * #)@ : ' #! : > : +
* N 3- = 5 > " AAA + + ? *

E /012/ / /3 / #! + 4 K " 6 "
? + " > (@ K 5 - * -
3- = 5 " AAA

% (
H33; < < <O< < # &G/H33; < < /H . <, < /K * 5
? +/

- H33; < < <O< < # &G/H33; < < /H . <,,< /:
% : K < < /

H33; < < <O< < # &G/H33; < < /H . <,,< / - P "
?M M 0 + C /

H33; < < <O< < # &G/H33; < < /H . <,&< / * >
0 + " ? /

/Q% /H33; < < <O< < # &G/H33; < < /H . <,&< / 5
* - +/

% (
/012/ / /3 / ( )#!(& *"
' ! (# :) 0 - * " -+

Theorem 1.11.6. *A knot which is the image of a C^1 function is ambient isotopic to a polygonal knot.*

PROOF (SKETCH). We shall describe an isotopy that deforms the knot so that it becomes the set of cone axes. Outside the cones, the isotopy is the identity. Inside each cone, the isotopy behaves on each level disc as indicated in Figure 1.17: the disc is transformed so that the single point of intersection of the knot with the disc is pulled into the centre. ☐

The converse is much easier.

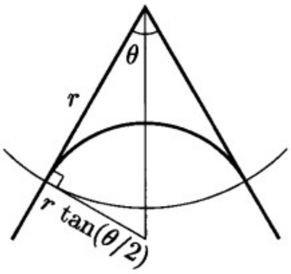

Figure 1.18. A corner can be replaced by a circular arc.

2 A Topologist's Toolkit

Most mathematics texts start from a specified base of mathematical knowledge, add some new ideas, and build on what has been previously studied to develop a familiar topic to a deeper level, or in a new direction. All the stages in the development of the theory are carefully proved and explained with nothing being assumed or hidden. If this pattern were strictly adhered to, knot theory would not be studied at undergraduate level.

To make the study of knot theory mathematically rigorous requires a fair amount of topological sophistication. We have seen that, if we are not careful, our definitions can lead to unexpected results. Trying to understand and exclude wild behaviour was one of the central questions for topologists in the middle of the twentieth century. Many strange spaces and embeddings were created to provide counterexamples to various conjectures. Eventually this work resulted in a firm foundation for geometric topology.

The fact that these foundations have not been studied should not inhibit progress. We shall proceed as working topologists do and just quote the useful results, safe in the knowledge that they have been established elsewhere. Just because we do not understand how a house is constructed that does not mean we cannot live in it.

The results we need are a set of 'tameness' conditions: they provide the reassurance that the topological world behaves as we expect it to. This means imposing a piecewise linear or differentiable structure on our objects of study. Because it is much easier to describe and manipulate piecewise linear objects, we shall work in that category.

With this in mind, this chapter contains a mixture of definitions, examples, facts, and techniques that we shall need throughout the book. The facts are theorems that we shall not prove. Some of them, like the classification of surfaces [71], you may have seen before. The others are topics for your further studies.

2.1 The 3-sphere

Knots and links live in 3-dimensional space. Normally we think of this space as being modelled by $\mathbb{R}^3$. Occasionally, however, we shall find it more convenient to think of links embedded in a compact space and for this we use the 3-dimensional sphere. Here are some descriptions of S^3.

1. Geometrically, the standard 3-sphere is the set of points in $\mathbb{R}^4$ that are one unit away from the origin: $S^3 = \{x \in \mathbb{R}^4 : |x| = 1\}$.

2. The 3-sphere is often thought of as the *1-point compactification* of $\mathbb{R}^3$. This means that we take $\mathbb{R}^3$ and an additional point denoted by ∞ and give it the following topology: a set U in $\mathbb{R}^3 \cup \{\infty\}$ is open if either U is an open set in $\mathbb{R}^3$, or $U = V \cup \{\infty\}$ where V is the complement of a closed and bounded set in $\mathbb{R}^3$.

   To show that $\mathbb{R}^3 \cup \{\infty\}$ is equivalent to the 3-sphere we use stereographic projection (the 2-dimensional case is shown in Figure 2.1). Let $\mathbb{R}^3 \subset \mathbb{R}^4$ be the subspace obtained by setting the fourth coordinate to -1, and let X be the unit 3-sphere in $\mathbb{R}^4$ minus the point $(0,0,0,1)$. There is a homeomorphism $h : \mathbb{R}^3 \longrightarrow X$: the unique line from $(0,0,0,1)$ to a point $p \in \mathbb{R}^3$ cuts X in a single point q and we define $h(p) = q$. We can extend h to a homeomorphism $\mathbb{R}^3 \cup \{\infty\} \longrightarrow S^3$ by setting $h(\infty) = (0,0,0,1)$. Check h is continuous in the given topology.

   If we are working in S^3, pictures of knots in infinite arcs (like the one

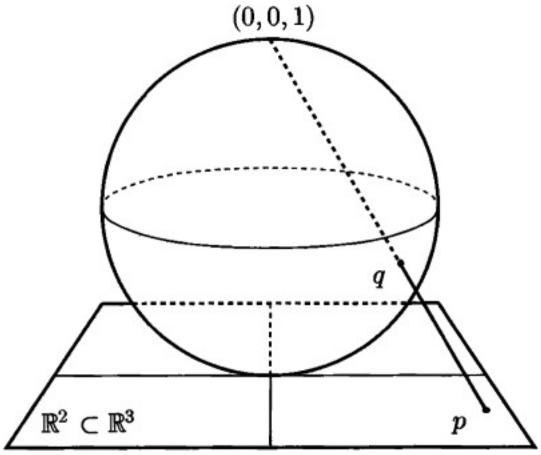

Figure 2.1. Stereographic projection of the plane onto a sphere.

shown in Figure 2.2) also represent loops: the lines extending left and right meet at ∞.

3. The (topological) 3-sphere can be constructed as the union of two balls glued together. This is analogous to the constructions in lower dimensions: a circle (S^1) can be formed by gluing the endpoints of two line segments together; a sphere (S^2) is formed by gluing the boundaries of two discs together.

   This decomposition of S^3 is easily seen through the previous example. The lower hemisphere $\{(x_1, \dots, x_4) \in S^3 \subset \mathbb{R}^4 : x_4 \leqslant 0\}$ is sent by stereographic projection to a 3-ball of radius 2 centred on $(0, 0, 0, -1)$. By symmetry, the upper hemisphere is also (homeomorphic to) a ball.

4. It is possible to construct a piecewise linear 3-sphere. An equilateral triangle is composed of 3 equal line segments meeting 2 at each vertex, and is topologically a circle (S^1). A regular tetrahedron is composed of 4 triangular discs glued edge to edge, meeting 3 at each vertex, and is topologically a sphere (S^2). Proceeding by analogy, we can take 5 solid tetrahedra, glue them face to face with 4 meeting at each vertex: this produces a combinatorial object that is topologically a 3-sphere.

From the topological viewpoint, it does not matter whether we work in $\mathbb{R}^3$ or S^3: a link that passes through ∞ can be moved a little to an isotopic position lying entirely in $\mathbb{R}^3$. However, there are some geometric problems in which differences do show up. In Figure 1.12 we saw that the figure-8 knot can be arranged so that a rigid motion of $\mathbb{R}^3$ (a rotation) carries the knot to its mirror image. In this position, it has an axis of rotation–reflection symmetry where the mirror plane is orthogonal to the axis. For some knots, such a transformation is possible in S^3 but not in $\mathbb{R}^3$. Figure 2.2 shows the oriented knot 8_{17} embedded in S^3 with the line extending left and right to ∞. This oriented knot is not + amphicheiral in S^3 or $\mathbb{R}^3$. However, it is clear from the figure that a 2-fold rotation of S^3 carries the knot to the reverse of its mirror image, showing that it is − amphicheiral. This symmetry can

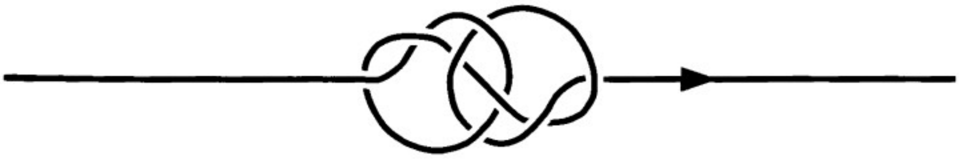

Figure 2.2. The knot 8_{17} is rigidly amphicheiral in S^3 but not in $\mathbb{R}^3$.

be shown topologically in $\mathbb{R}^3$ but it is not possible to demonstrate it using an axis of rotation–reflection in $\mathbb{R}^3$ [21, 70].

2.2 Manifolds

All the topological spaces we are interested in are n-dimensional *manifolds*, usually shortened to n-manifolds. A creature living inside a manifold can not distinguish its environment from the standard space $\mathbb{R}^n$ until it starts to move around. Any neighbourhood looks just like a neighbourhood of $\mathbb{R}^n$; it is only on a global scale that differences between manifolds can be discovered.

The upper half-space in $\mathbb{R}^n$ is the set $\mathbb{R}^n_+ = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_n \geqslant 0\}$. Its boundary, the subset given by $x_n = 0$, is a copy of $\mathbb{R}^{n-1}$.

A topological space X is an n-manifold if it is Hausdorff,* has a countable basis, and each point $x \in X$ has a neighbourhood $U \ni x$ which is homeomorphic to $\mathbb{R}^n$ or to $\mathbb{R}^n_+$. If $x \in U$ maps to a point in the boundary of $\mathbb{R}^n_+$ then x is a boundary point of X. The set of all boundary points, denoted by ∂X, is an $(n-1)$-manifold and is called the *boundary* of X.

It is clear that a manifold can be covered with open sets, and that the covering may contain uncountably many sets. A manifold X is *compact* if, when given any covering of X by open sets, we can find a finite subcover. A compact manifold with no boundary, that is $\partial X = \varnothing$, is said to be *closed.*

Up to homeomorphism, there are only four different connected 1-manifolds, and an example of each type is given in Figure 2.3. The real line $\mathbb{R}$ itself is clearly a manifold. It has no boundary and is not compact. The half-line has a single boundary component, the point 0. The edge skeleton of a cube is not a 1-manifold: the vertices do not have the required neighbourhoods.

All the 1-manifolds have natural analogues in all other dimensions. The

Example		Boundary	Compact
Line	$\{x \in \mathbb{R}\}$	×	×
Half-line	$\{x \in \mathbb{R} : x \geqslant 0\}$	✓	×
Interval	$\{x \in \mathbb{R} : x^2 \leqslant 1\}$	✓	✓
Circle	$\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$	×	✓

Figure 2.3. Examples of 1-manifolds and their properties.

---

* A space is Hausdorff if any two points can be placed in disjoint open sets.

% (
/012/ / /3 / (0 - % 5 = . @
4 " "E5
' ,&) *) = . *
H +D

% (#
/012/ / /3 / (# ? K % 5 ; "
= . @ * 7 " *"
')

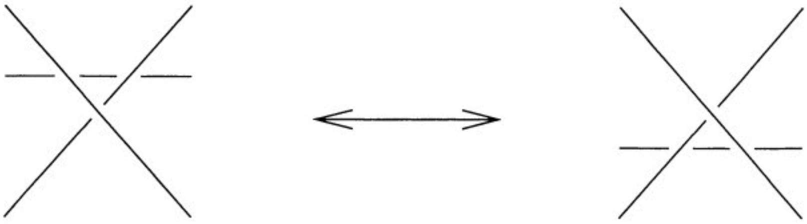

Fig. 2: Reidemeister move of type III

Definition 1.1 A *smoothing* of a crossing p is the procedure

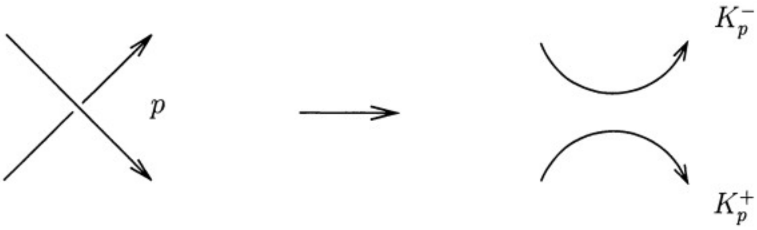

$$i(C,p) := \frac{1}{2\pi i}\oint \frac{1}{z-p}\mathrm{d}z\ .$$

Chapter 1

The space of diagrams

Let pr : $F^2 \times \mathbb{R} \to F^2$ denote the standard projection. A *knot* K is the oriented image of a smooth embedding of S^1 in $F^2 \times \mathbb{R}$. We call K a *diagram* if pr : $K \to F^2$ is an immersion.

Let M_K denote the *space of all diagrams* which are isotopic to a given diagram K. Viewing diagrams as differentiable submanifolds of $F^2 \times \mathbb{R}$ we can introduce a natural topology on M_K. It turns out that this space is never connected. Its connected components consist of diagrams which are isotopic through a smooth sequence of diagrams. This is usually called *regular isotopy*.

As in the case of $\mathrm{F}^2 = \mathbb{R}^2$ two diagrams are regular isotopic if and only if they can be transformed one into the other by finite series of unoriented Reidemeisters moves of type II and III together with isotopies which leave the curve pr(K) fixed. (The proof is the same as in the case of $\mathbb{R}^2$ because Reidemeisters theorem is just the characterization of the generic singularities in a 1-parameter family of diagrams.)

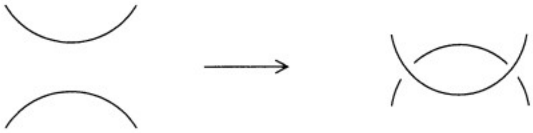

Fig. 1: Reidemeister move of type II

T. Fiedler, *Gauss Diagram Invariants for Knots and Links*
© Springer Science+Business Media Dordrecht 2001

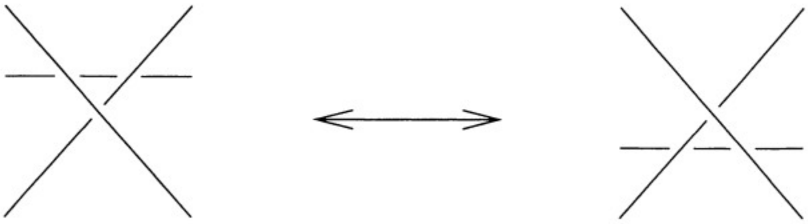

Fig. 2: Reidemeister move of type III

We introduce some basic notions.

Definition 1.1 A *smoothing* of a crossing p is the procedure

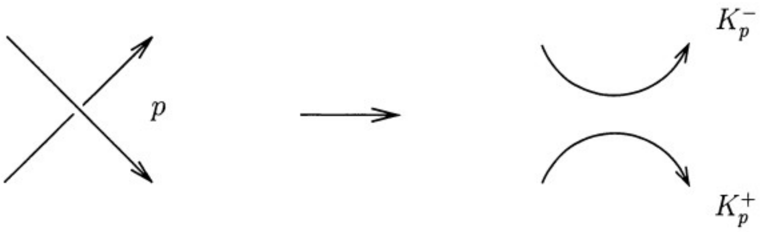

Fig. 3

where K_p^+ denotes the component, where the undercrossing is smoothed to the overcrossing.

Note that in addition to the link diagram which results after this operation, we have the trace of p in its complement.

We will characterize the components of M_K only in the special case of planar surfaces, i.e., the surface F^2 is embedded in the oriented plane $\mathbb{R}^2 = \mathbb{C}$. In this case the following notions are well defined:
The *winding index* of a curve $C \subset F^2 \subset \mathbb{C}$ around a point $p \notin C$ is

$$i(C,p) := \frac{1}{2\pi i} \oint \frac{1}{z-p} \mathrm{d}z \ .$$

Pictorially it measures how many times the curve walks around p.

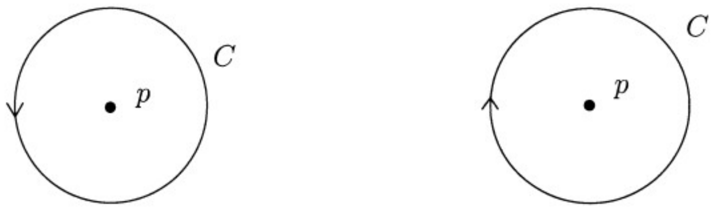

$i(C,p) = +1$ $i(C,p) = -1$

Fig. 4

The *Whitney index* $n(C)$ of a curve C is the degre of the map

$$\frac{C'}{||C'||} \; : \; S^1 \to S^1$$

where C' denotes the vector consisting of the derivatives. The *Whitney index of a diagram* K is the Whitney index of the curve pr(K).

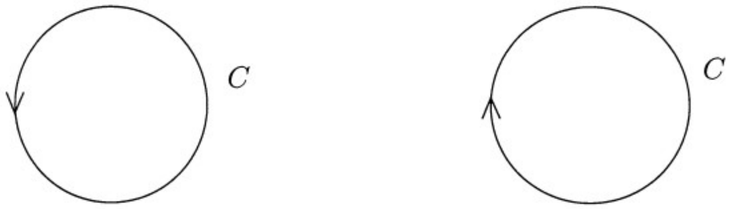

$n(C) = +1$ $n(C) = -1$

Fig. 5

Note that $n(c)$ and $i(C,p)$ are the only invariants of the couple up to regular homotopy $(C, p \notin C)$ if $F^2 = \mathbb{R}^2$ (see [A]).

The *writhe* $w(K)$ of a diagram is the sum of the writhes $w(p)$ of all its crossings p. (This is well defined in the case of *any* oriented surface F^2, because the lines $\mathbb{R}$ are always assumed to be oriented, and the writhe $w(p)$ depends only on the orientation of $F^2 \times \mathbb{R}$.)

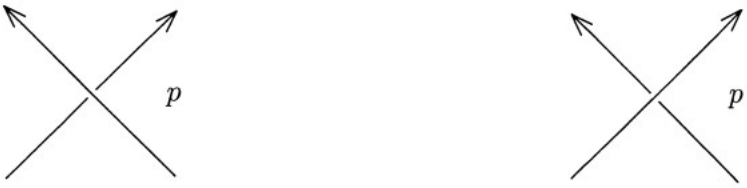

$w(p) = -1$ $w(p) = +1$

Fig. 6

The lines $\mathbb{R}$ are always oriented from the sheet of paper to the reader's eye. For each crossing p we define $i_p^{\pm} := i(K_p^{\pm}, p)$, where p means the trace of p as described above. In the following it will be more convenient to use $i_p := i_p^+ + i_p^-$ and $\sigma_p := i_p^+ - i_p^-$ instead of $i_p^{\pm}$.

The following lemma is well known (folklore) but of great importance for us.

Lemma 1.1 *Let F^2 be a planar surface. A diagram K' is in the component M_K if and only if K' is isotopic to K (they represent the same knot type) and $n(K') = n(K)$, $w(K') = w(K)$.*

Proof. Evidently $n(K)$ and $w(K)$ are invariants of M_K. We have to prove that besides the knot type these are the only invariants.

Let K_t, $t \in [0, 1]$, be an isotopy which connectes K with K'. It is a regular isotopy besides for a finite number of values of the parameter t_1,... where a Reidemeister move of type I is performed:

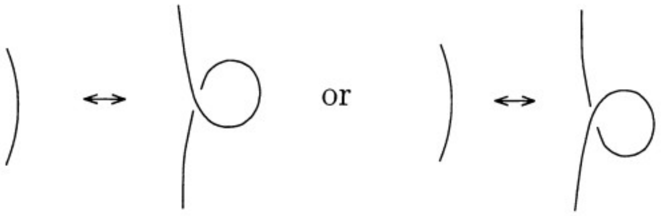

Fig. 7

There are exactly 4 types of oriented moves of type I plus their inverses:

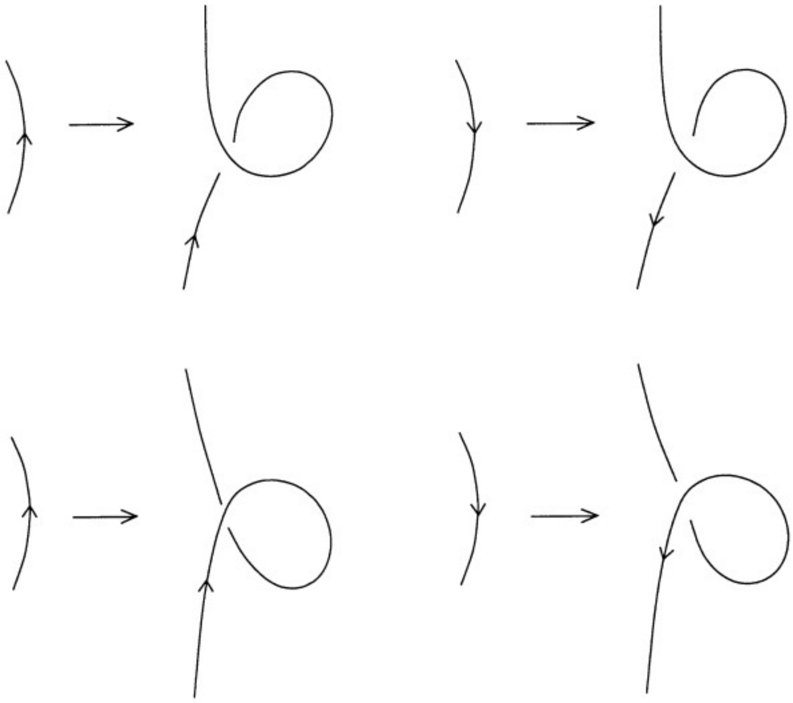

Fig. 8

Note that the four cases are distiguished exactly by the changings of n and w. We replace each of the moves in Fig. 8 by a regular isotopy using the 'Whitney trick' (Fig. 9).

We do not perform all the inverses of the moves in Fig. 8 in our isotopy K_t. We have lots of small curls like in Fig.8 in our isotopy now. They do not disturb the rest of the isotopy, and by moving them along K' we can collect them in a small arc of K. Performing the Whitney trick in the other direction we can cancel them two by two exactly if $n(K') = n(K)$ and $w(K') = w(K)$. ■

From now on M_K will always denote the connected component which contains K. Evidently for each knot K we have $n(K) + w(K) \equiv 1(\text{mod})2$ and perfoming moves as in Fig. 8 we see that this is the only restriction on $n(K)$ and $w(K)$.

Given two diagrams K, K' we add small curls to K' to obtain a diagram

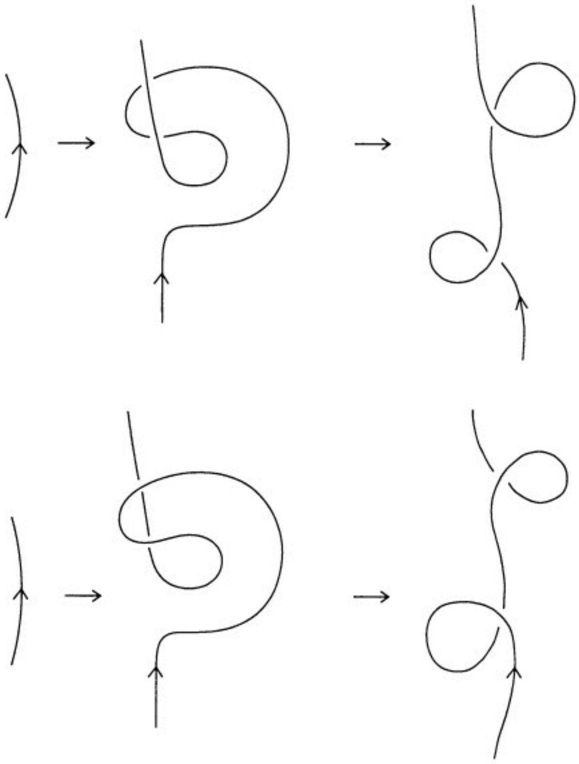

Fig. 9

K'' such that $n(K'') = n(K)$ and $w(K'') = w(K)$. This does not change the knot type. Hence K' is isotopic to K if and only if $K'' \in M_K$. So it is sufficient to find invariants of M_K.

Remark. In the case of a general surface there are two possibilities. Either an invariant does not change by a Reidemeister move of type I; hence each invariant of M_K is a knot invariant; we call these *stable invariants*. Or the invariant changes; it is then only an invariant of regular isotopy, sometimes called an invariant of *framed knots*.

It turns out that our Gauss sums for knots K and their generalizations for families of knots K_t, $t \in S^1$, are a very powerful tool for studying M_K.

16.04.2020

8.1 Bailey's Theorem

Let L be a 2-component link exterior X and group $\pi = \pi L$.

Source:

[Bai] Bailey, J. L.
Alexander Invariants of Links
Ph.D. Thesis, University of British Columbia (1977).

143) 20200416 -

Books_69_2020/
Knots and Links/
Gauss Diagram Invariants for Knots and Links (2001...).pdf
Pag. 2(18)

The winding index of a curve $C \subset \mathbb{R}^2 \subset \mathbb{C}$ around a point $p \notin C$ is

$$i(C,p) := \frac{1}{2\pi i} \oint \frac{1}{z-p} \, dz.$$

Pictorially it measures how many times the curve walks around p.

data 16.04.2020

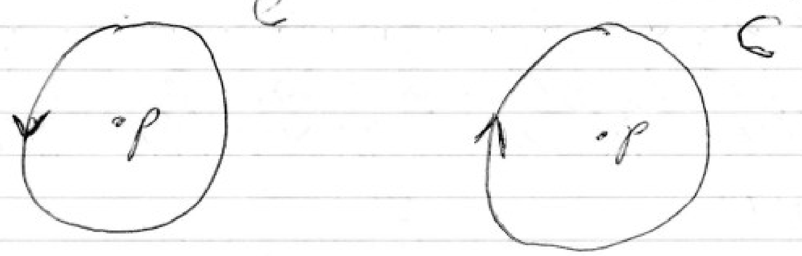

$i(C,p)=+1$ $i(C,p)=-1$

Figure 4

The Whitney index $n(C)$ of a curve C is the degree of the map

$$\frac{C'}{\|C'\|} : S^1 \to S^1$$

where C' denotes the vector consisting of the derivatives.

The Whitney index of a diagram K is the Whitney index of the curve $pr(K)$.

Obs: Usar este do mo gauss.

ref. Books_69_2020T ✡
Gauss – Werke

Gauss C.F – Werke Theoretische Astronomie Band 7 (1906). djvu

data 16.04.2020

Book: Books_69_2020/
Gauss_Werke/
Werke Fünfter Band by
Carl Friedrich Gauss ... pdf

→ Sphaeroidicorum ellipticorum
homogeneorum.

144) 2020041F

A) Books_69_2020/
Mechanisms for Generating
Mathematical Curves/
Sanet.st_3030421678.pdf

Pag. 63(74)

Chapter 4
Mechanisms for Generating Conical
Curves.

Pag. 64(75)

Fig. 4.1 Geometric construction
of the pedal Γ of the
ellipse C.

Shows the pedal of an ellipse with
the focal points F and F'. The pedal
curve results from the orthogonal
projection of a fixed on the tangent
lines of a given curve.
The locus of point M resulted by

H . /012/ / /3 / (# 4 . :R H " -+
: " ? *" S * " *

% (($ $) ((
/012/ / /3 / ((# # (,&) *"
' ,# &(* (: (

Chapter 4
Mechanisms for Generating Conical Curves

Abstract Starting from geometry, the equivalent mechanisms that trace ellipses, hyperbolas, or parabolas are built. These mechanisms are studied structurally, and their kinematic analysis is developed, resulting in relations, successive positions, diagrams, tables, and curves traced. Curves are traced either as trajectories of points or by enveloping. On the basis of projective geometry problems, a new mechanism that traces parabolas by enveloping is presented. Starting from an analytical geometry problem, synthesis of a new equivalent mechanism, which traces parabolas by enveloping, is also created. Next, we study a new mechanism, created on a locus problem, reaching the hyperbolas obtained by enveloping. Another problem in geometry allows the synthesis of the mechanism that traces a hyperbola as a trajectory of a point. A locus problem in 1896 allowed the synthesis of a new mechanism that traces ellipses, parabolas, and hyperbolas.

4.1 Mechanism Ellipsograph

4.1.1 Introduction

Ellipsograph mechanisms have been studied for a long time, being closely linked to the development of the Mechanism and Machine Theory [2, 22]. Scientists, especially mathematicians, have studied various mathematical curves [3, 4, 24]. They have established their geometric properties, and they have drawn it with a ruler and a compass. Afterward, they designed mechanisms to trace them, making models that have verified the correctness of calculations [19]. Taimina studied in detail the evolution of research in the field of the mechanisms generating facet curves [23]. Different mechanisms are described, including conicographs, throughout the evolution of the Theory of Mechanism, from the thirteenth century to the twentieth century. Reuleaux designed and built many mechanisms for curves generating [18, 19]. Artobolevskii also obtained many different mechanisms that draw different mathematical curves, including ellipsographs, parabolographs, and hyperbolographs [1, 2]. All these mechanisms have been designed based on geometrical considerations.

© Springer Nature Switzerland AG 2020

I. Popescu et al., *Mechanisms for Generating Mathematical Curves*,
Springer Tracts in Mechanical Engineering,
https://doi.org/10.1007/978-3-030-42168-7_4

Fig. 4.1 Geometric construction of the pedal Γ of the ellipse C

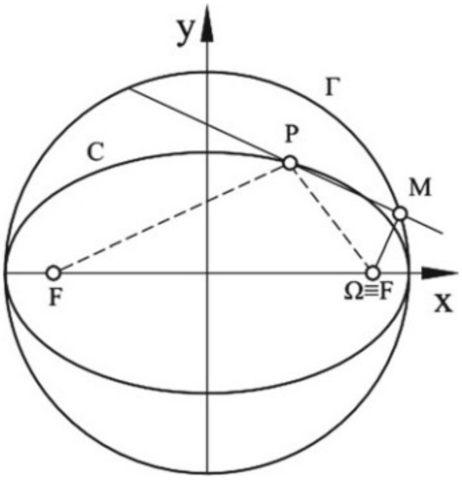

Figure 4.1 [4] shows the pedal of an ellipse with the focal points F and F'. The pedal curve results from the orthogonal projection of a fixed point on the tangent lines of a given curve.

The locus of point M resulted by bringing a perpendicular from F to the ellipse tangent, in point P. The searched pedal is the Γ circle.

4.1.2 The Mechanism Synthesis

Based on the geometric construction of the pedal (Fig. 4.1), we proposed the inverted problem, that is finding an ellipsograph mechanism [17, 20]. In this way, the point M is taken on the Γ circle, then M is joined to F and further MP is constructed perpendicular on FM, which is the envelope of the ellipse.

Figure 4.2 shows the mechanism which is built based on this method.

Fig. 4.2 Mechanism obtained

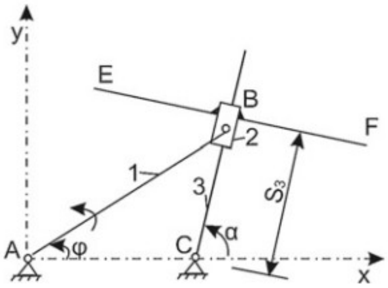

' # * > "

Chapter 5
Mechanisms for Generating Some Plane Mathematical Curves

Abstract Based on geometrical problems, the synthesis of original mechanisms that trace different mathematical curves is made. Then the structure is analyzed and the calculations are made for the positions of the mechanisms, successive positions, diagrams of the curve generating point coordinates, diagrams of the displacements for some sliders, and the traced curves. Further curves generated by these mechanisms are also studied for different initial input data. In this way, with two original mechanisms, the cissoids (of the circle and of the straight line) are traced, in some cases in particular generating ellipses, parabolas and hyperbolas. Next Berard's curve is studied, where a point of the mechanism traces a branch of the curve, and another point, the other branch. By modifying some dimensions, other curves generated by the mechanism are obtained. We also made the synthesis of a new mechanism that generates egg-shaped curve. Analysis of the mechanism led to the generation of this curve. We changed some data of the mechanism resulting in similar or modified curves, but positioned in different trigonometric quadrants. Another original mechanism, also based on a geometry problem, describes double egg curve. By changing the initial data, similar curves are obtained, some of them being incomplete. Another original mechanism, based on a geometry problem, describes Bernoulli quartic. In this case, an additional kinematic chain is used to provide the midpoint of a straight segment of variable length at the movement of the mechanism. The chapter continues with a mechanism based also on geometrical principles, which traces Maclaurin's trisectrix. By changing some initial data, the mechanism traces completely different curves from the initial ones. There has also been a synthesis of some original mechanisms that trace ophiuride. In this case, other curves are obtained by modifying some dimensions of the mechanism. Finally, an original mechanism that traces the Pascal's snail is studied. Different snails are obtained by changing some dimensions of the mechanism.

© Springer Nature Switzerland AG 2020

I. Popescu et al., *Mechanisms for Generating Mathematical Curves*,
Springer Tracts in Mechanical Engineering,
https://doi.org/10.1007/978-3-030-42168-7_5

5.1 Mechanisms for Generating Cissoids

5.1.1 The Cissoid of a Circle

5.1.1.1 Introduction

On a straight line d, which is the system's ordinate, there are taken two fixed points O and Ω, and on the curve C, it is taken an M point (Fig. 5.1a–c). The locus of point P, where the MΩ secant intersects the perpendicular from O on OM is the cissoid of C curve [1]. The cissoid of a circle is called the cissoid of Diocles.

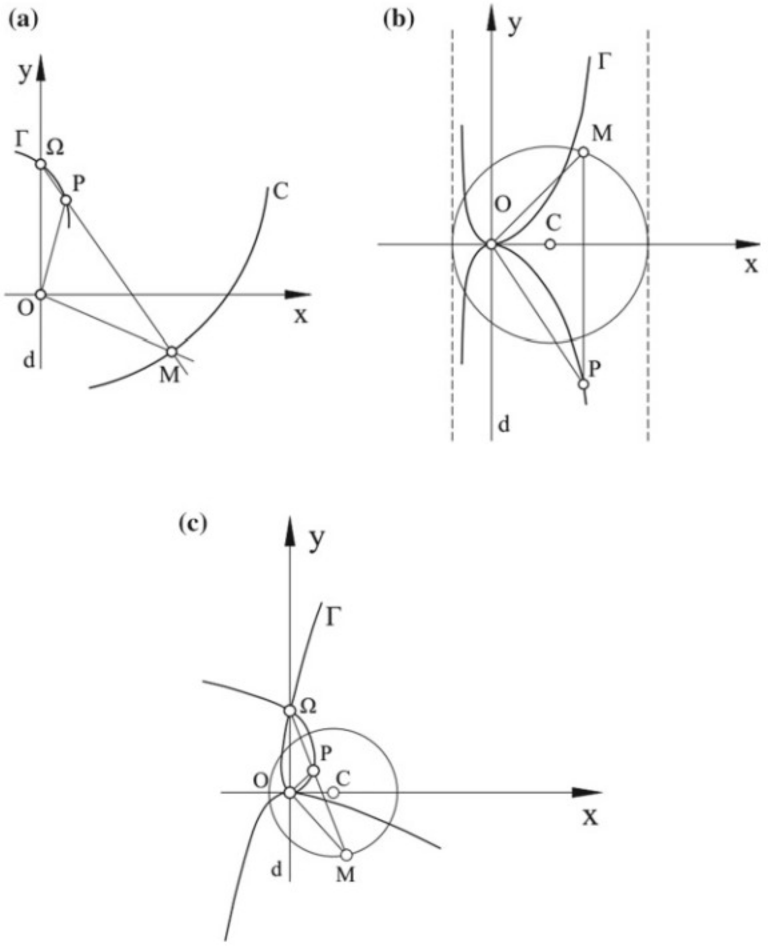

Fig. 5.1 Geometrical construction of the cissoid

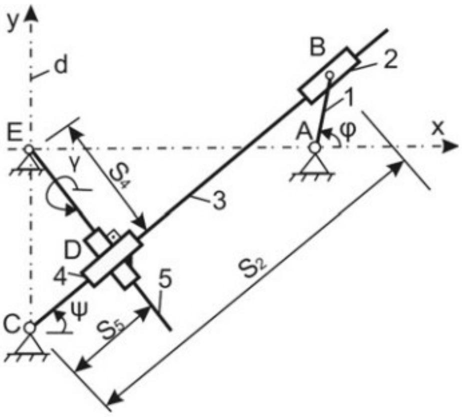

Fig. 5.2 Obtained mechanism

5.1.1.2 The Mechanism Synthesis

In Fig. 5.2, it is presented the mechanism which is constructed based on Fig. 5.1 [2].

Curve C becomes the circle of *AB* radius; the MΩ straight line becomes the *CB* element with variable length, that is why at *B* was mounted slider 2.

Point *O* shown in Fig. 5.1 becomes *E* in Fig. 5.2 and the perpendicular from *E* to *CB* was materialized into two perpendicular welded sliders. The generating point of the cissoid is the point *D*.

5.1.1.3 The Mechanism Analysis

Based on Fig. 5.2, with the contours method, Eqs. (5.1)–(5.5) can be written:

$$x_B = AB\ \cos\varphi + x_A = S_2\ \cos\psi \tag{5.1}$$

$$y_B = AB\ \sin\varphi + y_A = y_C + S_2\ \sin\psi \tag{5.2}$$

$$x_D = S_5\ \cos\psi + x_C = S_4\ \cos\gamma + x_E \tag{5.3}$$

$$y_D = y_C + S_5 \sin\psi = S_4 \sin\gamma + y_E \tag{5.4}$$

$$\gamma = \psi + 270 \tag{5.5}$$

from which the unknowns S_2, S_4, S_5, ψ are calculated.

5.1.1.4 Results Obtained

We used the following initial data: $XA = 40$, $AB = 40$, $YC = -50$; $XC = \mathrm{XE} = 0$ (mm). The mechanism for one position is given in Fig. 5.3. The mechanism for different positions is shown in Fig. 5.4.

It is seen from Fig. 5.4a that, although B makes a complete rotation, the EDC area is situated under the x-axis, and the generating point D is not passing above the x-axis. Because of this, the mechanism draws only a branch of the cissoid (Fig. 5.5a), by taking $YC = -120$ mm.

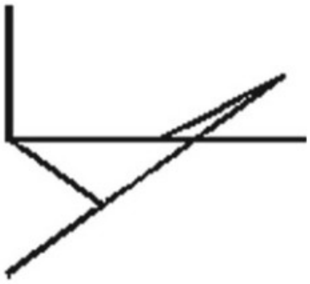

Fig. 5.3 Mechanism for a given position

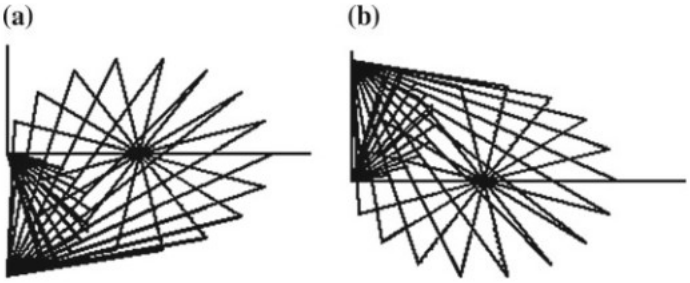

Fig. 5.4 Mechanism for different positions

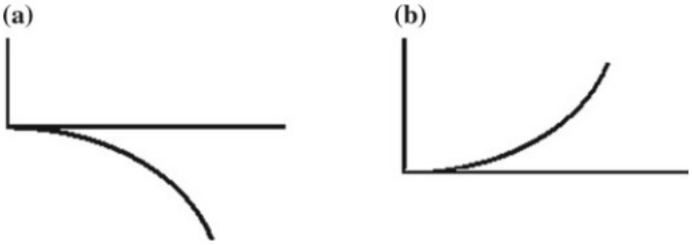

Fig. 5.5 Different branches of the cissoid

If it is assumed $YC = +120$ mm, then the other branch of the cissoid (Fig. 5.5b) is obtained. The successive positions are given in Fig. 5.4b.

5.1.2 The Straight Line Cissoid

5.1.2.1 Introduction

When the C curve of Fig. 5.1a is a straight line Δ, then the cissoid is a conic curve: an ellipse (Fig. 5.6a), a parabola (Fig. 5.6b), and a hyperbola (Fig. 5.6c) [3]. The ellipse is obtained when M is on the right side of Δ, the parabola when M is on Δ, the hyperbola when M is between Δ and OY, respectively.

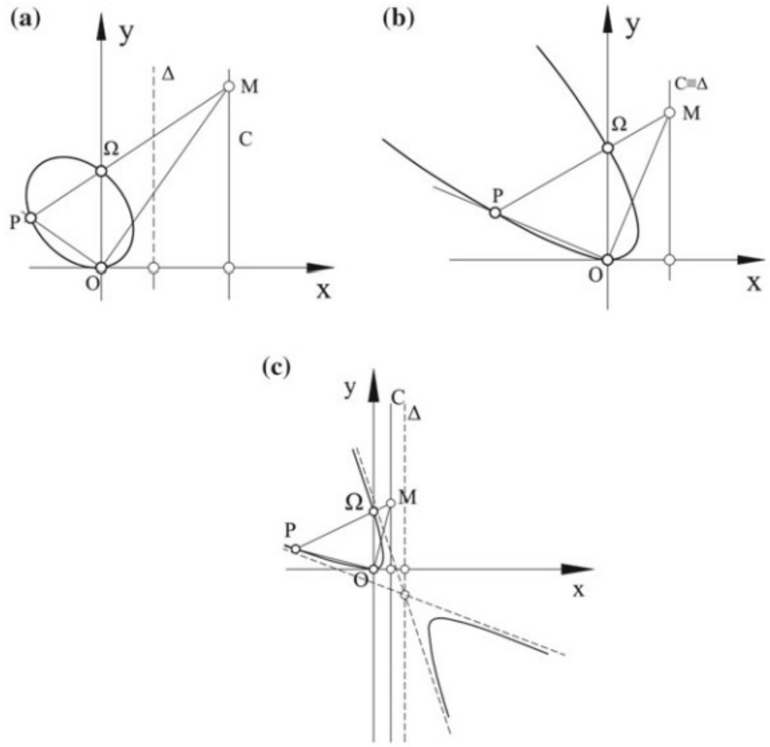

Fig. 5.6 Geometrical constructions of straight line cissoids

3- > = "

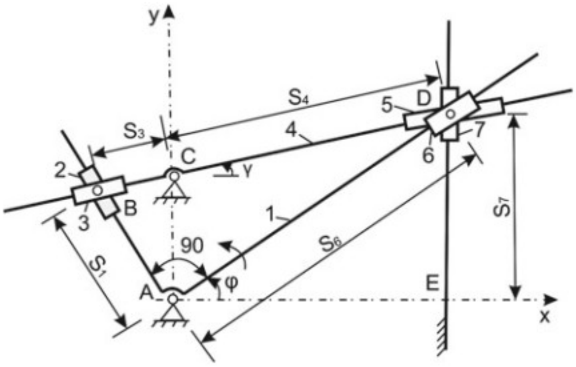

Fig. 5.7 Mechanism obtained

5.1.2.2 The Mechanism Synthesis

The mechanism that undergoes these conditions was found by synthesis, and given in Fig. 5.7 [4]. The element 1 has the BAD angle of 90°, the Δ straight line is noted as *ED*, which is defined with *XE*. At *E*, there are three sliders moving on the elements 4, 1, 0. At *B* there are two sliders which are connected by a torque in such way that between the elements 1 and 4 exists a rotational movement.

5.1.2.3 The Mechanism Analysis

In order to analyze the positions, based on the contours method we write the Eqs. (5.6)–(5.13):

$$x_D = S_6 \cos\varphi = x_E = \text{const.} \tag{5.6}$$

$$y_D = S_6 \sin\varphi = S_7 \tag{5.7}$$

$$x_B = S_1 \cos(\varphi + 90) \tag{5.8}$$

$$y_B = S_1 \sin(\varphi + 90) \tag{5.9}$$

$$x_C + S_4 \cos\gamma = x_D \tag{5.10}$$

$$y_C + S_4 \sin\gamma = y_D \tag{5.11}$$

$$x_B + S_3 \cos\gamma = x_C \tag{5.12}$$

$$y_B + S_3 \sin\gamma = y_C \tag{5.13}$$

5.1.2.4 Results Obtained

The values for *YC* and *XE* are given as input data. For the values of $YC = 50$ mm and $XE = 80$ mm the mechanism was obtained in one position, which is given in Fig. 5.8. The generating ellipse can be observed in Fig. 5.9 and the successive positions of the mechanism are given in Fig. 5.10.

Referring to Fig. 5.11, it is shown that *XB* varies by a sinusoid type curve, it also moves to the left of *y*-axis and *YB* is always positive.

For the computer program a jump instruction was provided for the cases when *S*7 advances above 200 mm, because it went beyond the drawing borders, which is

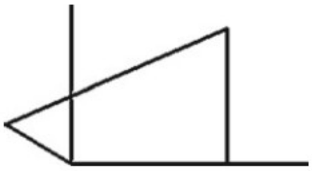

Fig. 5.8 Mechanism in one position

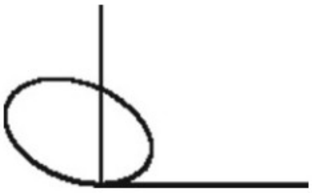

Fig. 5.9 Resulting ellipse

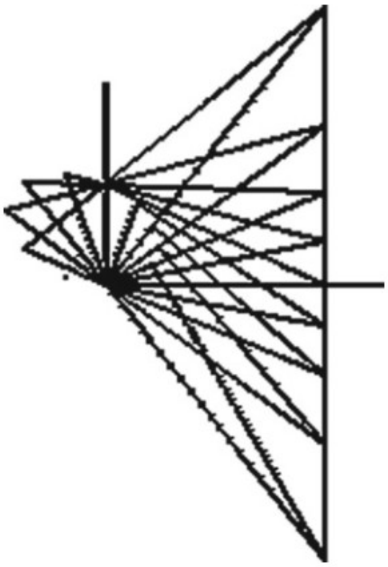

Fig. 5.10 Successive positions of the mechanism

Fig. 5.11 Variations of the coordinates y_B, x_B

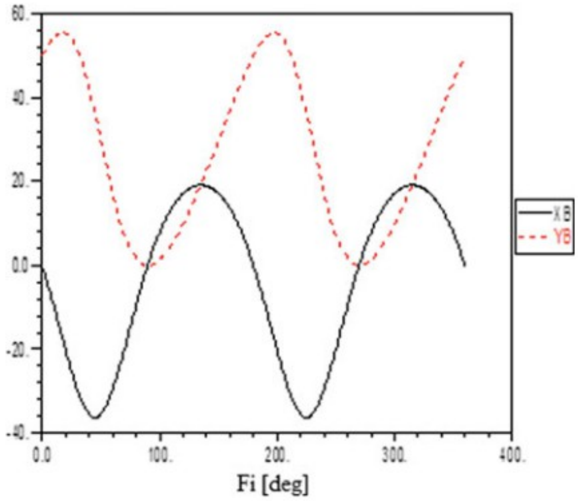

found in Fig. 5.12 where jumps occur at $S7$ and $S6$, for φ around values of 90° and 270°.

There were obtained ellipses (Fig. 5.13) for other sets of values, as follows:

1. for $YC = 100$ mm, $XE = 100$ mm;
2. for $YC = -50$ mm, $XE = 50$ mm;
3. for $YC = 50$ mm, $XE = -50$ mm;
4. for $YC = 100$ mm, $XE = -100$ mm;
5. for $YC = 80$ mm, $XE = 50$ mm.

Fig. 5.12 Curves of the displacements $S4$, $S6$, $S1$, $S7$, $S3$

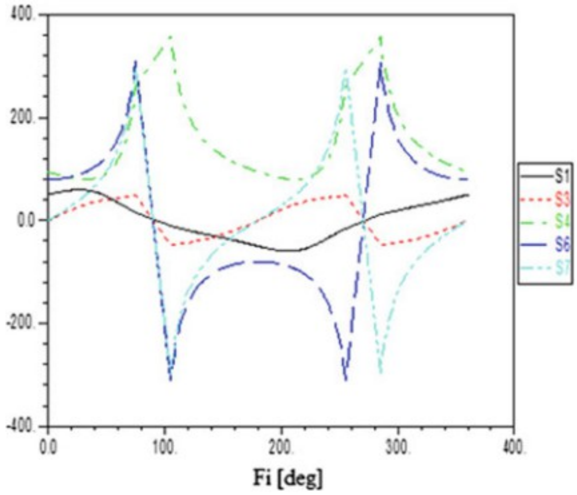

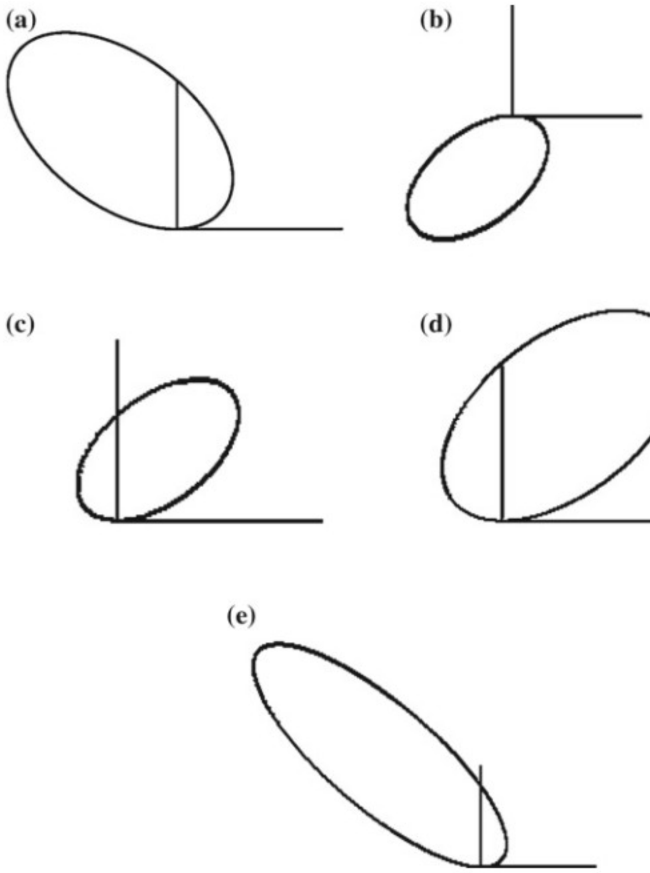

Fig. 5.13 Resulting ellipses

Changing the initial data led to resizing the ellipses dimensions and their axis orientation. When $YC < 0$, the ellipse is in quadrant III, and for $XE < 0$, the axis of the ellipse is placed in quadrant I.

We have also obtained the parabolas in Fig. 5.14:

1. for $YC = 100$ mm, $XE = 40$ mm;
2. for $YC = 100$ mm, $XE = -50$ mm.

For $YC = 80$ mm, $XE = -45$ mm, it resulted the ellipse in Fig. 5.15

In Fig. 5.16 there are presented the generating hyperbolas, for the values:

1. $YC = 100$ mm, $XE = 30$ mm;
2. $YC = 80$ mm, $XE = 25$ mm.

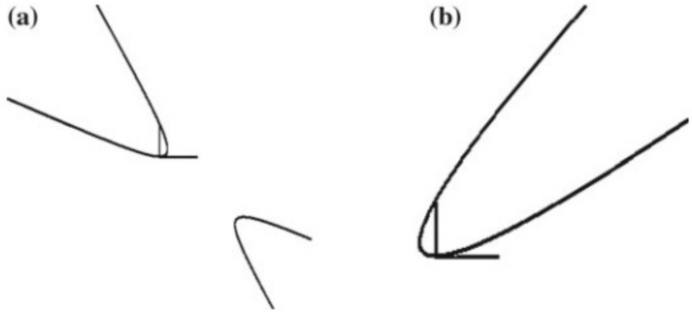

Fig. 5.14 Generated parabolas

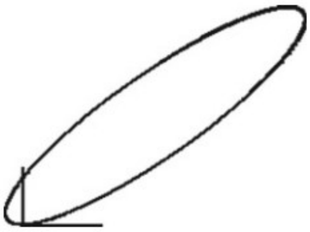

Fig. 5.15 Elipse for $y_C = 80$ mm, $x_E = –45$ mm

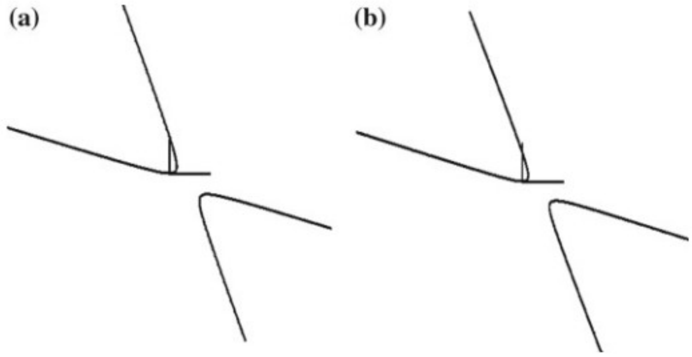

Fig. 5.16 Generated hyperbolas

5.2 Mechanism for Generating Berard's Curve

5.2.1 The Mechanism Synthesis

The rod curves of the rod-crank mechanism were studied by mathematicians since the invention of the steam engine [5]. Thus, in 1820, Bérard obtained such curves.

Fig. 5.17 Mechanism obtained

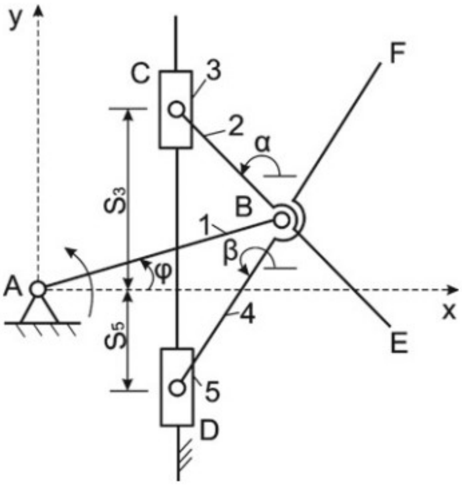

Bérard has obtained and defined them as locus of points which are situated on a straight line that goes from one end on a circle and the other end on another straight line. In [6] are given the equations and the particular cases of these curves. The movement of some straight lines is also designed, obtaining particular curves.

On the basis of these geometric considerations, the mechanism of Fig. 5.17 was conceived, which, by the *E* and *F* points, describes trajectories which are connected to each other, forming the curves of Berard. Structurally, the mechanism is R-RRP-RRP [7].

5.2.2 The Mechanism Analysis

On the basis of Fig. 5.17, Eqs. (5.14)–(5.23) are written:

$$x_B = AB\cos\varphi \tag{5.14}$$

$$y_B = AB\sin\varphi \tag{5.15}$$

$$x_C = AB\cos\varphi + BC\cos\alpha = \text{const.} \tag{5.16}$$

$$y_C = AB\sin\varphi + BC\sin\alpha = S_3 \tag{5.17}$$

$$x_E = x_B + BE\cos(\alpha + \pi) \tag{5.18}$$

$$y_E = y_B + BE\sin(\alpha + \pi) \tag{5.19}$$

- /012/ / /3 / ((K ? + "
0** -+ B 3* *"
(

1

The Geometry of Curves

1.1 Introduction

The world of Euclidean geometry is inhabited by lines and planes. If we wish to go beyond this flat world to a universe of curvature, we need to understand more general types of curves and surfaces. This chapter is devoted to exploring curves. Later we shall see that a basic understanding of the geometry of curves is essential to understanding the nature of surfaces as well.

A *curve* in 3-space $\mathbb{R}^3$ is a continuous mapping $\alpha\colon I \to \mathbb{R}^3$ where I is some type of interval (e.g., $(0, 1)$, (a, b), $[a, b]$, $(-\infty, a]$, $[0, 1]$ etc.) on the real line $\mathbb{R}$. Because the range of α is $\mathbb{R}^3$, α's output has three coordinates. We then write, for $t \in I$, a *parametrization* for α,

$$\alpha(t) = (\alpha^1(t),\ \alpha^2(t),\ \alpha^3(t))$$

where the α^i are themselves functions $\alpha^i\colon I \to \mathbb{R}$. A useful way to think about curves is to consider t to be time and $\alpha(t)$ to be the path of a particle in space. We say α is *differentiable* or *smooth* if each coordinate function α^i is differentiable as an ordinary real-valued function of $\mathbb{R}$. In fact, later, in order to define curvature and torsion, we will need each α^i to be at least 3-times differentiable. In practice, we will not find this condition to be much of a constraint.

The *velocity vector* of α at t_0 is defined to be

$$\alpha'(t_0) = \left(\left.\frac{d\alpha^1}{dt}\right|_{t=t_0},\ \left.\frac{d\alpha^2}{dt}\right|_{t=t_0},\ \left.\frac{d\alpha^3}{dt}\right|_{t=t_0} \right)$$

where $d\alpha^i/dt$ is the ordinary derivative and $|_{t=t_0}$ denotes evaluation of the derivative at $t = t_0$. We shall also write $\frac{d\alpha^i}{dt}(t_0)$ for this evaluation when it is convenient. In order to interpret α' geometrically, we introduce the

Example 1.1.1 (First Example of a Curve: a Line in $\mathbb{R}^3$).
We know that two points determine a line. In the plane $\mathbb{R}^2$, this leads to the usual slope-intercept equation which gives an algebraic description of the line. Unfortunately, in $\mathbb{R}^3$ we don't have a good notion of slope, so we can't expect exactly the same type of algebraic description to work.

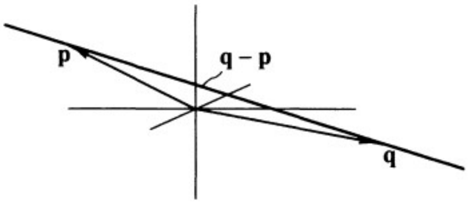

Figure 1.1. A parametrized line

For two points, **p** and **q**, the line l joining them may be described as follows. To attain the line, add the vector **p**. To travel along the line, use the direction vector $\mathbf{q} - \mathbf{p}$ since this is the direction from **p** to **q**. A *parameter* t tells exactly how far along $\mathbf{q} - \mathbf{p}$ to go. Putting these steps together produces a curve which is a *parametrized line* (see Figure 1.1) in $\mathbb{R}^3$,

$$\alpha(t) = \mathbf{p} + t(\mathbf{q} - \mathbf{p}).$$

For instance, if $\mathbf{p} = (1, 2, 3)$ and $\mathbf{q} = (-1, 4, -7)$, then $\mathbf{q} - \mathbf{p} = (-2, 2, -10)$, so the line through **p** and **q** is given by the curve

$$\begin{aligned} \alpha(t) &= (1, 2, 3) + t(-2, 2, -10) \\ &= (1 - 2t, 2 + 2t, 3 - 10t). \end{aligned}$$

Exercise* 1.1.2. What is the parametrization of the line through $(-1, 0, 5)$ and $(3, -1, -2)$?

Exercise* 1.1.3. In $\mathbb{R}^4$, what is the parametrization of the line through $(-1, 6, 5, 0)$ and $(0, 1, -3, 9)$?

So we see that, given a line $\alpha(t) = \mathbf{p} + t(\mathbf{q} - \mathbf{p})$, $\alpha'(t) = \mathbf{q} - \mathbf{p}$. Since this *direction vector* $\mathbf{q} - \mathbf{p}$ is then the velocity vector, we often write

$$\alpha(t) = \mathbf{p} + t\mathbf{v}.$$

Now that we understand lines, we can ask what the picture of the velocity vector is for any curve $\alpha(t)$? The definitions give

$$\begin{aligned} \alpha'(t_0) &= \left(\left.\frac{d\alpha^1}{dt}\right|_{t=t_0}, \left.\frac{d\alpha^2}{dt}\right|_{t=t_0}, \left.\frac{d\alpha^3}{dt}\right|_{t=t_0} \right) \\ &= \left(\lim_{t\to t_0} \frac{\alpha^1(t) - \alpha^1(t_0)}{t - t_0}, \lim_{t\to t_0} \frac{\alpha^2(t) - \alpha^2(t_0)}{t - t_0}, \lim_{t\to t_0} \frac{\alpha^3(t) - \alpha^3(t_0)}{t - t_0} \right) \\ &= \lim_{t\to t_0} \left(\frac{\left(\alpha^1(t), \alpha^2(t), \alpha^3(t)\right) - \left(\alpha^1(t_0), \alpha^2(t_0), \alpha^3(t_0)\right)}{t - t_0} \right) \\ &= \lim_{t\to t_0} \frac{\alpha(t) - \alpha(t_0)}{t - t_0}. \end{aligned}$$

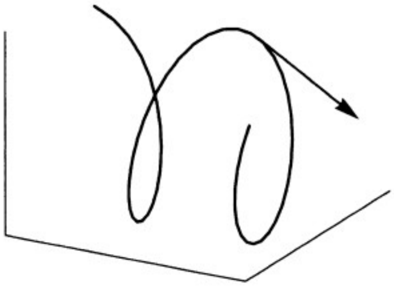

Figure 1.2. A curve with tangent vector

This looks much like our usual "slope" definition of the derivative. Note that the approximating vectors inside the limit always point in the direction of increasing t. For $t < t_0$, this follows since $t - t_0$ is negative. Now, $t \to t_0$, so the vector moves closer and closer to being a tangent vector (see Figure 1.2) to the curve at $\alpha(t_0)$. Hence, the velocity vector $\alpha'(t_0)$ to α at t_0 is a tangent vector to α at $\alpha(t_0)$. Notice that we obtain a vector with a precise length, not a line. Remember that a vector $\mathbf{v} = (v^1, v^2, v^3) \in \mathbb{R}^3$, has a *length* given by the Pythagorean theorem,

$$|\mathbf{v}| = \sqrt{(v^1)^2 + (v^2)^2 + (v^3)^2}.$$

Exercise∗ 1.1.4. Show that $|\mathbf{v}|$ is the distance of (v^1, v^2, v^3) from $(0, 0, 0)$ by two applications of the Pythagorean theorem.

A curve $\alpha(t)$ is *regular* if $\alpha'(t) \neq 0$ for all $t \in I$. In general, if a curve is not regular, then we split it into regular pieces and consider each of these separately. The assumption of regularity ensures that we keep going in the same direction along the curve. This will be important when we consider arclength shortly.

Example 1.1.5 (The Cusp $\alpha(t) = (t^2, t^3)$).
Note that $x = t^2$, $y = t^3$, so $y = \pm x^{3/2}$ with $\alpha'(t) = (2t, 3t^2)$ and $\alpha'(0) = (0, 0)$. Thus α is *not* regular, but the pieces above and below the x-axis *are* regular.

In our situation, $|\alpha'(t)| = \sqrt{(d\alpha^1/dt)^2 + (d\alpha^2/dt)^2 + (d\alpha^3/dt)^2}$ is simply the *speed* of α. Again, thinking of α as the path of a particle and t as time, we see that the length of the velocity vector is precisely the speed of the particle at the given time. For a line, $\alpha(t) = \mathbf{p} + t\mathbf{v}$, the speed is simply $|\mathbf{v}|$, the length of the direction vector. Note that we have taken lines to have constant speed $\mathbf{v}$. We will see later that every *regular* curve can be re-parametrized to have constant speed, so this is no real restriction. However, a given parametrization may disguise the fact that the underlying curve is a line until re-parametrized to have constant speed. We will keep our convention that lines are parametrized with constant speed because we want to characterize them by Proposition 1.1.7.

If $\alpha(t) = \big(\alpha^1(t), \alpha^2(t), \alpha^3(t)\big)$ is a curve in $\mathbb{R}^3$, then its acceleration vector is given by

$$\alpha''(t) = \left(\frac{d^2}{dt^2}\alpha^1(t),\ \frac{d^2}{dt^2}\alpha^2(t),\ \frac{d^2}{dt^2}\alpha^3(t)\right).$$

Exercise⁎ 1.1.6. Newton asked the question, what is the curve which, when revolved about an axis, gives a surface offering the least resistance to motion through a "rare" fluid (e.g., air)? Newton's answer is the following curve (see Chapter 7).

$$(x(t), y(t)) = \left(\frac{\lambda}{2}\left[\frac{1}{t} + 2t + t^3\right], \frac{\lambda}{2}\left[\ln\left(\frac{1}{t}\right) + t^2 + \frac{3}{4}t^4\right] - \frac{7}{8}\lambda\right),$$

where λ is a positive constant and $x \geq 2\lambda$. Graph this curve and compute its velocity and acceleration vectors. What angle does the curve make with the x-axis at the intersection point $(2\lambda, 0)$?

We now come to the first (and simplest nontrivial) instance of calculus imposing a constraint on the geometry of a curve. In a real sense, this is what the differential geometry of curves is all about and we will see many more examples of it later.

Proposition 1.1.7. *The curve α is a (constant speed) straight line if and only if $\alpha'' = 0$.*

Proof. If $\alpha(t) = \mathbf{p} + t\mathbf{v}$ is a line, then $\alpha'(t) = \mathbf{v}$ (which is a constant vector), so $\alpha''(t) = 0$.

If $\alpha''(t) = 0$ for all t, then $d^2\alpha^i(t)/dt^2 = 0$, for each coordinate function $\alpha^i(t)$. But a zero 2nd derivative simply means that $d\alpha^i(t)/dt = v^i$ is a constant. We may integrate with respect to t to obtain $\alpha^i(t) = p^i + v^i t$ where p^i is a constant of integration. Then

$$\alpha(t) = (p^1 + tv^1, p^2 + tv^2, p^3 + tv^3) = p + t\mathbf{v}$$

with $\mathbf{p} = (p^1, p^2, p^3)$ and $\mathbf{v} = (v^1, v^2, v^3)$. Hence the curve α may be parametrized as a line. □

This easy result indicates how we will use calculus to detect geometric properties. Just to see how far we can get using these elementary ideas, let's consider the following question. What is the shortest distance between two points $\mathbf{p}$ and $\mathbf{q}$ in $\mathbb{R}^3$? We have been taught since we were children that the answer is a line, but now we can see why our intuition is correct. Calculus tells us that distance is simply the integral of speed. Therefore, the integral

$$L(\alpha) = \int_a^b |\alpha'(t)|\, dt$$

calculates the *arclength* (i.e., the distance travelled) of a regular curve $\alpha\colon [a, b] \to \mathbb{R}^3$ from $\alpha(a)$ to $\alpha(b)$. In order to see this, consider a small piece (i.e., "an infinitesimal piece") of arclength approximated as in Figure 1.3. The Pythagorean theorem then says that the hypotenuse of the right triangle with sides dx and dy has length $l = \sqrt{dx^2 + dy^2}$. Suppose both x and y are parametrized by t. Then $\alpha(t) = (x(t), y(t))$ and $dx^2 = (dx/dt)^2\, dt^2$ and $dy^2 = (dy/dt)^2\, dt^2$. Finally, we have

$$l = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt$$

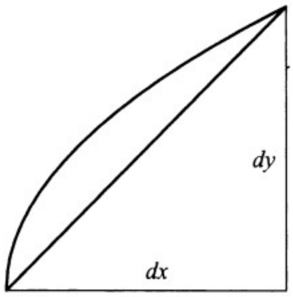

Figure 1.3. An infinitesimal piece of curve

expressing the relation, distance equals rate times time. Now we add up all the small pieces of the curve by integration to get,

$$\begin{aligned} L(\alpha) &= \int_a^b \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt \\ &= \int_a^b \left|\alpha'(t)\right| \, dt. \end{aligned}$$

The three-dimensional version of this intuitive description simply uses our earlier exercise on the Pythagorean theorem in three dimensions. Now, the length of the vector $\mathbf{v}$, $|\mathbf{v}|$, may be written $\sqrt{\mathbf{v} \cdot \mathbf{v}}$, where $\cdot$ denotes the *dot product* of vectors. Recall that, in general, $\mathbf{v} \cdot \mathbf{w} = v^1 w^1 + v^2 w^2 + v^3 w^3$, for $\mathbf{v} = (v^1, v^2, v^3)$ and $\mathbf{w} = (w^1, w^2, w^3)$. We also have the following

Proposition 1.1.8. *The dot product* $\mathbf{v} \cdot \mathbf{w}$ *may be calculated by* $\mathbf{v} \cdot \mathbf{w} = |\mathbf{v}|\,|\mathbf{w}| \cos\theta$, *where* θ *is the angle between the vectors* $\mathbf{v}$ *and* $\mathbf{w}$.

Proof. From Figure 1.4, we see that the usual properties of vectors give $\mathbf{u} = \mathbf{v} - \mathbf{w}$. Hence

$$\begin{aligned} |\mathbf{u}|^2 &= \mathbf{u} \cdot \mathbf{u} = (\mathbf{v} - \mathbf{w}) \cdot (\mathbf{v} - \mathbf{w}) \\ &= \mathbf{v} \cdot \mathbf{v} - 2\mathbf{v} \cdot \mathbf{w} + \mathbf{w} \cdot \mathbf{w} \\ &= |\mathbf{v}|^2 + |\mathbf{w}|^2 - 2\mathbf{v} \cdot \mathbf{w}. \end{aligned}$$

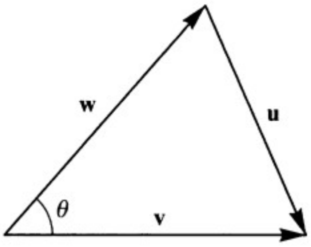

Figure 1.4.

But the Law of Cosines for this triangle says $|\mathbf{u}|^2 = |\mathbf{v}|^2 + |\mathbf{w}|^2 - 2|\mathbf{v}|\,|\mathbf{w}|\,\cos\theta$. Equating the two right-hand sides and eliminating like terms produces

$$\mathbf{v} \cdot \mathbf{w} = |\mathbf{v}|\,|\mathbf{w}| \cos\theta.$$

□

Corollary 1.1.9 (Schwarz's Inequality). *The dot product obeys the inequality,*

$$|\mathbf{v} \cdot \mathbf{w}| \leq |\mathbf{v}|\,|\mathbf{w}|.$$

Proof. Simply note that $|\cos\theta| \leq 1$. □

We note here that, with respect to the derivative, the dot product behaves just as ordinary multiplication — namely, the Leibniz (or product) rule holds. In particular,

Proposition 1.1.10. *If $\alpha(t)$ and $\beta(t)$ are two curves in $\mathbb{R}^3$, then*

$$\frac{d(\alpha \cdot \beta)}{dt} = \frac{d\alpha}{dt} \cdot \beta + \alpha \cdot \frac{d\beta}{dt}.$$

Proof. We use the ordinary product rule on the component functions,

$$\begin{aligned}
\frac{d(\alpha \cdot \beta)}{dt} &= \frac{d}{dt}(\alpha^1\beta^1 + \alpha^2\beta^2 + \alpha^3\beta^3) \\
&= \sum_i \left(\frac{d\alpha^i}{dt}\beta^i + \alpha^i\frac{d\beta^i}{dt}\right) \\
&= \sum_i \left(\frac{d\alpha^i}{dt}\beta^i + \sum_i \alpha^i\frac{d\beta^i}{dt}\right) \\
&= \frac{d\alpha}{dt} \cdot \beta + \alpha \cdot \frac{d\beta}{dt}.
\end{aligned}$$

□

When the parameter t is understood, we will write the product rule as $(\alpha \cdot \beta)' = \alpha' \cdot \beta + \alpha \cdot \beta'$. We are now able to answer the question of which (smooth) path between two points gives the shortest distance.

Theorem 1.1.11. *In $\mathbb{R}^3$, a line is the curve of least arclength between two points.*

Proof. Consider two points $\mathbf{p}, \mathbf{q} \in \mathbb{R}^3$. The line between them may be parametrized by $l(t) = \mathbf{p} + t(\mathbf{q} - \mathbf{p})$, where $\mathbf{q} - \mathbf{p}$ is the vector in the direction from $\mathbf{p}$ to $\mathbf{q}$. Then $l'(t) = \mathbf{q} - \mathbf{p}$ and $|l'(t)| = |\mathbf{q} - \mathbf{p}|$, a constant. Therefore,

$$L(l) = \int_0^1 |l'(t)|\,dt = |\mathbf{q} - \mathbf{p}| \int_0^1 dt = |\mathbf{q} - \mathbf{p}|$$

3- > = - G * " "

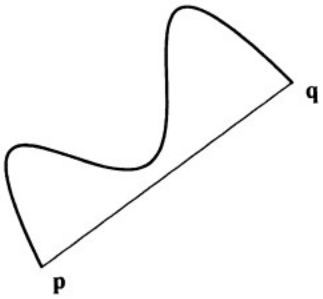

Figure 1.5. Line and comparison curve

and the length of the line segment (or direction vector) from **p** to **q** is the distance from **p** to **q** (as of course we expected). Now consider another curve α which joins **p** and **q** (see Figure 1.5).

We want to show that $L(\alpha) > L(l)$ and, since α is arbitrary, this will say that the straight line minimizes distance. Now, why should α be longer than l? One intuitive explanation is to say that α starts off in the wrong direction. That is, $\alpha'(a)$ is not "pointing toward" **q**. How can we measure this deviation? The angle between the unit vector in the direction of $\mathbf{q}-\mathbf{p}$, $\mathbf{u} = (\mathbf{q}-\mathbf{p})/|\mathbf{q}-\mathbf{p}|$, and $\alpha'(a)$, the tangent vector of α at **p**, may be determined using the dot product $\alpha'(a)\cdot\mathbf{u}$. The total deviation may be added up by integration to give us an idea of why $L(\alpha) > L(l)$ should hold. Precisely, we compute $\int_a^b \alpha'(t)\cdot\mathbf{u}\,dt$ in two ways to obtain the inequality. Now, we have $(\alpha(t)\cdot\mathbf{u})' = \alpha'(t)\cdot\mathbf{u} + \alpha(t)\cdot\mathbf{u}' = \alpha'(t)\cdot\mathbf{u}$ since **u** is a constant vector. Also, the Fundamental Theorem of Calculus gives $\int_a^b df/dt\,dt = f(b) - f(a)$, so

$$\begin{aligned}
\int_a^b \alpha'(t)\cdot\mathbf{u}\,dt &= \int_a^b (\alpha(t)\cdot\mathbf{u})'\,dt = \alpha(b)\cdot\mathbf{u} - \alpha(a)\cdot\mathbf{u} \\
&= \mathbf{q}\cdot\mathbf{u} - \mathbf{p}\cdot\mathbf{u} \qquad \text{since } \alpha(a) = \mathbf{p},\ \alpha(b) = \mathbf{q} \\
&= (\mathbf{q}-\mathbf{p})\cdot\mathbf{u} \\
&= \frac{(\mathbf{q}-\mathbf{p})\cdot(\mathbf{q}-\mathbf{p})}{|\mathbf{q}-\mathbf{p}|} \\
&= \frac{|\mathbf{q}-\mathbf{p}|^2}{|\mathbf{q}-\mathbf{p}|} \\
&= |\mathbf{q}-\mathbf{p}| \\
&= L(l),
\end{aligned} \tag{1}$$

the straight line distance from **p** to **q**. We also have,

$$\begin{aligned}
\int_a^b \alpha'(t)\cdot\mathbf{u}\,dt &\le \int_a^b |\alpha'(t)|\cdot|\mathbf{u}|\,dt \qquad \text{by the Schwarz inequality} \\
&= \int_a^b |\alpha'(t)|\,dt \qquad \text{since } |\mathbf{u}| = 1 \\
&= L(\alpha).
\end{aligned} \tag{2}$$

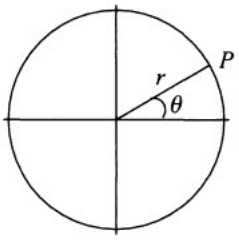

Figure 1.6. Circle of radius r

Combining (1) and (2), we have $L(l) \leq L(\alpha)$. Note also that $\alpha'(t) \cdot \mathbf{u} = |\alpha'(t)| \cdot |\mathbf{u}|$ only when $\cos\theta = 1$, or $\theta = 0$. That is, $\alpha'(t)$ must be parallel to $\mathbf{q} - \mathbf{p}$ for all t. In this case α is the line from $\mathbf{p}$ to $\mathbf{q}$. Therefore, we have the strict inequality $L(l) < L(\alpha)$ unless α is a line. ☐

So, using the first notions of calculus, we have seen how to identify lines in $\mathbb{R}^3$ and to determine exactly why lines give the shortest distance between two points. Of course, $\mathbb{R}^3$ is governed by Euclidean Geometry. One of the main goals of differential geometry is to develop analogous techniques for curved geometries as well.

While curves described parametrically may be defined at the whim of the definer, in fact they often arise from geometric or physical considerations. Several of the examples below illustrate this.

Example 1.1.12 (The Circle of Radius r Centered at $(0, 0)$).
The circle is one of the most useful and familiar curves in geometry. Of course, the question is, how can such a curve be parametrized? The definitions of $\sin\theta$ and $\cos\theta$ as the vertical and horizontal sides of the right triangle with angle θ formed by dropping a vertical line from P to the horizontal axis give us a way of assigning coordinates to the point P when $r = 1$. Namely, the point P has coordinates $(\cos\theta,\ \sin\theta)$ (see Figure 1.6). For a circle of radius r, by the fundamental property of similar triangles, the vertical and horizontal sides of the triangle must be $r\sin\theta$ and $r\cos\theta$ respectively. Hence, the coordinates of the point on the circle are $(r\cos\theta,\ r\sin\theta)$. Therefore, the circle may be parametrized by the angle θ. Again thinking of the circle as the path of a particle, we write

$$\alpha(t) = (r\cos t,\ r\sin t), \qquad \text{for } 0 \leq t < 2\pi.$$

There are two important observations to remember. First, we have $\alpha'(t) = (-r\sin t, r\cos t)$, so $\alpha(t) \cdot \alpha'(t) = 0$ for all t. In other words, the tangent vector is perpendicular to the radius. Second, $\alpha''(t) = (-r\cos t,\ -r\sin t) = -\alpha(t)$, so the acceleration vector points toward the center of the circle. Finally, we can compute the arclength (i.e., circumference) of the circle,

$$\begin{aligned} L(\alpha) &= \int_0^{2\pi} |\alpha'(t)|\, dt = \int_0^{2\pi} \sqrt{r^2\sin^2 t + r^2\cos^2 t}\, dt \\ &= \int_0^{2\pi} r\sqrt{\sin^2 t + \cos^2 t}\, dt = \int_0^{2\pi} r\, dt \\ &= r\,t\,\Big|_0^{2\pi} = 2\pi\, r. \end{aligned}$$

3- 1 "

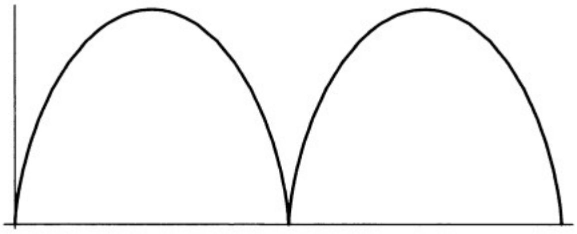

Figure 1.7. A cycloid

Exercise 1.1.13. Suppose a circle of radius a sits on the x-axis making contact at $(0, 0)$. Let the circle roll along the positive x-axis. Show that the path α followed by the point originally in contact with the x-axis is given by

$$\alpha(t) = (a\,(t - \sin t),\ a\,(1 - \cos t))$$

where t is the angle formed by the (new) point of contact with the axis, the center and the original point of contact. This curve is called a *cycloid* (see Figure 1.7) and we will meet it again in Chapter 7. Hints: Recall that $s = a\,t$ where s is arclength. Draw a picture after the circle has rolled s units. Where is the old point of contact?

Exercise 1.1.14. Consider a cycloid of the form

$$(x(t), y(t)) = (A + a\,(t - \sin t),\ B - a\,(1 - \cos t)).$$

Graph this cycloid to see that it is an inverted form of the one found in the previous exercise. Suppose a unit mass particle starts at rest at a point $(\bar{x}, \bar{y})$ on the cycloid corresponding to an angle $\bar{t}$ in the parametrization above. Under the influence of gravity (and assuming no friction), show that, no matter what initial $\bar{t}$ is chosen, it always takes a time of

$$T = \sqrt{\frac{a}{g}}\,\pi\,, \qquad \text{where } g \text{ is the gravitational constant}$$

for the particle to slide down to the bottom of the cycloid (i.e., $t = \pi$). This property was used by C. Huygens to make clocks without pendula (which he hoped would allow for accurate timekeeping on ships at sea and, thereby, improve navigation). He named the cycloid the *tautochrone* since the latter means "same time". We will revisit the cycloid in Chapter 7 where we will see it has yet another intriguing property and name. Hints: potential energy is turned into kinetic energy by $v^2/2 = g(\bar{y} - y)$. Time equals distance divided by speed, so

$$\begin{aligned} T &= \frac{1}{\sqrt{2g}} \int_{\bar{x}}^{x_{bot}} \frac{\sqrt{1 + y'^2}}{\sqrt{\bar{y} - y}}\,dx \\ &= \sqrt{\frac{a}{g}} \int_{\bar{t}}^{\pi} \sqrt{\frac{1 - \cos t}{\cos \bar{t} - \cos t}}\,dt \\ &= \sqrt{\frac{a}{g}} \left(-2 \arcsin\left[\frac{\sqrt{2}\cos\frac{t}{2}}{\sqrt{1 + \cos \bar{t}}} \right] \right) \Bigg|_{\bar{t}}^{\pi} \end{aligned}$$

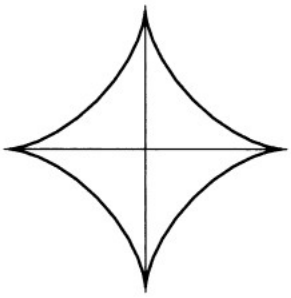

Figure 1.8. An astroid

using $y' = (dy/dt)\big/(dx/dt) = -\sin t/(1-\cos t)$, $\bar{y} = B - a(1-\cos\bar{t})$, $y = B - a(1-\cos t)$ and simplifying. Verify the final step by differentiation and compute T. Now see Section 1.7 for a Maple approach.

Example 1.1.15 (The Astroid).
Consider the astroid $\alpha(t) = (a\cos^3 t,\ a\sin^3 t)$ for $0 \leq t \leq 2\pi$ (see Figure 1.8). The definition of the astroid (which was discovered by people searching for the best form of gear teeth) is very similar to that of the cycloid. For the astroid, however, a circle is rolled, not on a line, but inside another circle. More precisely, let a circle of radius $a/4$ roll inside a large circle of radius a (centered at $(0,0)$ say). For concreteness, suppose we start the little circle at $(a, 0)$ and follow the path of the point originally in contact with $(a, 0)$ as the circle rolls up. Roll the circle a little bit and notice that a piece of arclength s is used up. The key to understanding this situation is the fact that, for t, the angle from the center of the large circle to the new contact point and θ, the angle through which the small circle has rolled (measured from the point of contact), we have $s = at$ and $s = a\theta/4$. Hence, $\theta = 4t$.

Now we can do two things to parametrize the path of the rolling point. First, we can parametrize the path *with respect to the center of the little circle*. This is easy because we are just moving an angle θ around a circle. There is a small problem because we measure θ from the point of contact, but this is handled easily as in Figure 1.9. Note that we make real use of the relation $\theta = 4t$ here. The coordinates with respect to the center then become $x = (a/4)\cos 3t$ and $y = (a/4)\sin 3t$.

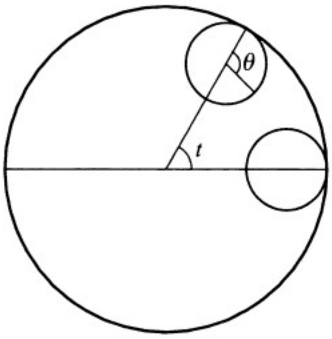

Figure 1.9.

3- 0 " " C T C 0* .. , (
* $$ $ $,$ C * .. , (

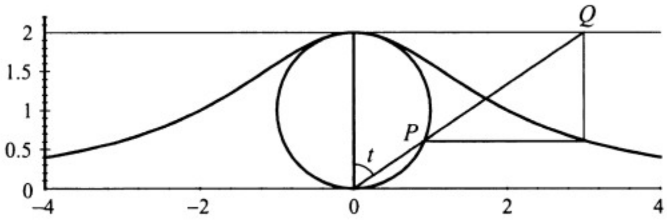

Figure 1.10. The Witch of Agnesi

Secondly, the center of the little circle always remains on the circle of radius $3a/4$ centered at $(0, 0)$, so that the center has moved to $((3a/4)\cos t, (3a/4)\sin t)$. Hence, with respect to the origin $(0, 0)$, the rolling point has moved to

$$\alpha(t) = \left(\frac{3a}{4}\cos t + \frac{a}{4}\cos 3t,\ \frac{3a}{4}\sin t - \frac{a}{4}\sin 3t\right).$$

Exercise 1.1.16. Verify the statements above and show that the formula for the astroid may be reduced to

$$\alpha(t) = (a\cos^3 t,\ a\sin^3 t)$$

with implicit form $x^{2/3} + y^{2/3} = a^{2/3}$.

Exercise 1.1.17. From the origin draw a line through any point P of the circle of radius a centered at $(0, a)$. Find the intersection Q of this line with the horizontal line $y = 2a$. Drop a vertical line from Q and intersect it with a horizontal line passing through P. This intersection is a point on the curve known as the *witch of Agnesi*.[1] See Figure 1.10. Let t denote the angle between the vertical axis and the line through $(0, 0)$, P and Q. Show that the witch $W(t)$ is given in terms of t by

$$W(t) = \left(2a\tan t,\ 2a\cos^2 t\right).$$

Example 1.1.18 (The Helix).
Consider the parametrization $\alpha(t) = (a\cos t,\ a\sin t,\ bt)$, for $0 \leq t < \infty$. Notice that the the first two coordinates provide circular motion while the third coordinate lifts the curve out of the plane. The resulting *helix* is then seen in Figure 1.11. We calculate $\alpha'(t) = (-a\sin t, a\cos t, b)$ and $\alpha''(t) = (-a\cos t, -a\sin t, 0)$. Note that $\alpha''(t)$ points toward the z-axis.

Example 1.1.19 (The Suspension Bridge).
Consider the diagram of a suspension bridge (Figure 1.12) where the cable supports a uniformly distributed load. This means that, over any interval $[a, b]$, the weight supported is given by

---

[1] The 'witch' was named by Maria Agnesi in the middle 1700s — apparently as the result of a mistranslation! The curve had been studied previously by Grandi who had used the Italian word "versorio", meaning "free to move in any direction", in connection with it. Agnesi mistakenly thought Grandi had used the Italian "versiera", which means "Devil's wife" or "witch" and so Agnesi's witch got "her" name [Yat74].

data 17 04 2020

bringing a perpendicular from F to the ellipse tangent in point P the searched pedal is the r circle.

Pág. 121 (132)

Chapter 5
Mechanisms for Generating some Plane Mathematical Curves

Pág. 122 (133)

5.1.1 the Cissoid of a Circle

Off: Cissóide de Diócles
→ curva de terceira ordem.

Fig 5.1 Geometrical construction of the cissoid

itens: a), b) e c)

B) Books-69-2020/
Differential Geometry and its Applications/
Differential Geometry and its applications ... pdf

Pág. 02 (24)

Fig 1.1 A parametrized line

Pág. 05 (27) →

data 17.04.2020

Figure 1.3 An infinitesimal piece of curve.)

The three-dimensional version of this intuitive description simply uses our earlier exercise on the Pythagorean theorem in three dimensions.

Now, the length of the vector v, $|v|$, may be written $\sqrt{v \cdot v}$, where $\cdot$ denotes the dot product of vectors. Recall that in general,

$$v \cdot w = v^1 w^1 + v^2 w^2 + v^3 w^3,$$

for $v = (v^1, v^2, v^3)$ and $w = (w^1, w^2, w^3)$.

Pag. 11 (33)

Figure 1.10 the Witch of Agnesi

Exercise 1.1.17

Pag. 33(55)

Figure 1.20 Neil's parabola

Obs.: Veja item 144 A), Pag. 122(133)

Cissoid of Circle.

⇒ As imagens são próximas (estudar!)

✡ Pag. 63 (85) →

data 17 04 2020

Figure 1.39

Viviani's curve as the intersection of sphere and cylinder.

Hence, Viviani's curve is truly spherical

Exercise 1.7.11

Show that the following curve is spherical

$(a \cos^2(t), a \sin(t) \cos(t), a \sin(t))$

$$x^2 + y^2 - \frac{b^2}{a^2} z^2 = \boxed{\frac{4b^2}{a^4}}$$

checar!

Look: Formula Bhaskara

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Outra pesquisa: (relacionada)

B1) Books_69_2020/
Complex Numbers and Geometry/
[Spectrum series] Liang...
Complex numbers and geometry... djvu

→

```
>  recreate3dview(kap3d1,tau3d1,0,7.8,0,1.5,0,1,0,1.2,
73,75);
```

```
>  kap3d2:=t->cos(t);
```

$$kap3d2 := \cos$$

```
>  recreate3dview(kap3d2,tau3d1,0,60,-3,3,-2,4,-3.5,2,
100,95);
```

Exercise 1.7.10. Recreate the rectifying curves of Theorem 1.5.11. Hint: take, say, $\kappa = s^2 + 2s$ and $\tau = s$.

We have seen in the re-creation of curves above how Maple's "dsolve" command can be used to numerically solve differential equations. In fact, however, "dsolve" can often provide explicit solutions also. Let's look at an example where we see how the Laplace transform can be used to determine the tautochrone of Exercise 1.1.14, the curve having the property that a bead sliding (without friction) to the bottom arrives in the same time no matter where it starts on the curve. Write the curve as $x = x(y)$ since the potential due to gravity involves y. Suppose that the bead begins at a point (x_1, y_1) on the curve and the bottom is at $(0, 0)$. Conservation of energy says that we have $mg\,y_1 = mv^2/2 + mg\,y$ at any point (x, y) along the slide, so the bead's speed is given by $v = \sqrt{2g(y_1 - y)}$. The arclength along an infinitesimal portion of the curve is given by $ds = \sqrt{(\frac{dx}{dy})^2 + 1}\,dy$ and $D = R\,T$, so the total time it takes to the bottom is given by $T = \int_{y_1}^{0} \sqrt{(\frac{dx}{dy})^2 + 1}/\sqrt{2g(y_1 - y)}\,dy$. We want this to be a constant independent of y_1. Write

$$\sqrt{2g}\,T = \int_{y_1}^{0} \frac{\sqrt{(\frac{dx}{dy})^2 + 1}}{\sqrt{y_1 - y}}\,dy,$$

$f(y_1) = 1/\sqrt{y_1}$ and $h(y_1) = \sqrt{(\frac{dx}{dy_1})^2 + 1}$. This means that the integral is a convolution integral of f and h, so $\sqrt{2g}\,T = -f(y) * h(y)$, where we have switched the variable to y for convenience. We will now apply the Laplace transform to both sides, ending up with the transform of $h(y)$ alone. We will then use the inverse transform to find a condition on h which will determine the curve $x = x(y)$. The necessary Maple package is "inttrans" which stands for "integral transform".

```
>  with(inttrans):
```

Sometimes we have to tell Maple certain things to assume about a problem. Here is how to do that.

```
>  assume(T,constant); additionally(g,constant);
> laplace(sqrt(2*g)*T,y,s);
```

$$\frac{\sqrt{2}\,\sqrt{g}\,T}{s}$$

```
> f:=1/sqrt(y);
```

$$f := \frac{1}{\sqrt{y}}$$

It is fortunate that Maple knows how to take this Laplace transform, because it very difficult by hand. Indeed, it involves the Γ function and its properties.

```
> laplace(f,y,s);
```

$$\sqrt{\frac{\pi}{s}}$$

Since $\sqrt{2\,g}\,L(T) = -L(f)\,L(h)$, we can find $L(h)$ by dividing as follows.

```
> Lh:= simplify(laplace(sqrt(2*g)*T,y,s)/laplace(f,y,s),
symbolic);
```

$$Lh := \frac{\sqrt{2}\,\sqrt{g}\,T}{\sqrt{\pi}\,\sqrt{s}}$$

Therefore, h(y) can be found using the inverse laplace transform.

```
> h:=simplify(invlaplace(Lh,s,y),symbolic);
```

$$h := \frac{\sqrt{2}\,\sqrt{g}\,T}{\pi\,\sqrt{y}}$$

But we know that $h(y) = \sqrt{(\frac{dx}{dy})^2 + 1}$, so this gives us a differential equation for $x = x(y)$. Here is how to solve for $x = x(y)$.

```
> solve(sqrt(diff(x(y),y)^2+1)=
2^(1/2)*g^(1/2)*T/Pi/y^(1/2),diff(x(y),y));
```

$$\frac{\sqrt{-y\,(\pi^2\,y - 2\,g\,T^2)}}{\pi\,y},\ -\frac{\sqrt{-y\,(\pi^2\,y - 2\,g\,T^2)}}{\pi\,y}$$

```
> simplify(1/Pi/y*(-y*(Pi^2*y-2*g*T^2))^(1/2), symbolic);
```

$$\frac{\sqrt{\pi^2\,y - 2\,g\,T^2}\,I}{\pi\,\sqrt{y}}$$

Note that Maple simplifies by introducing the complex number $I = \sqrt{-1}$. This is just notation however. We now use "dsolve" to solve the differential equation.

```
> dsolve(diff(x(y),y)=1/Pi/sqrt(y)*
(-(Pi^2*y-2*g*T^2))^(1/2),x(y));
```

$$x(y) = \frac{\sqrt{-\pi^2 y + 2 g T^2}\,\sqrt{y}}{\pi} + \frac{g T^2 \sqrt{(-\pi^2 y + 2 g T^2)\, y} \cdot \arctan\left(\frac{\pi \left(y - \frac{g T^2}{\pi^2}\right)}{\sqrt{-\pi^2 y^2 + 2 y g T^2}}\right)}{\pi^2 \sqrt{-\pi^2 y + 2 g T^2}\,\sqrt{y}} + \_C1$$

This doesn't give much information, so we substitute cleverly to see what is going on.

```
> x(y)=Int(1/Pi/y^(1/2)* (2*g*T^2-Pi^2*y)^(1/2),y=0..y1);
```

$$x(y) = \int_0^{y1} \frac{\sqrt{-\pi^2 y + 2 g T^2}}{\pi \sqrt{y}}\, dy$$

In order to change variables in an integral easily, we need to use the "student" package.

```
> with(student):
> x(y)=changevar(y=2*g*T^2/Pi^2*sin(w)^2,
Int(1/Pi/y^(1/2)*(2*g*T^2-Pi^2*y)^(1/2),y=0..y1),w);
```

$$x(y) = \int_0^{\arcsin\left(\frac{\sqrt{2}\,\pi \sqrt{\frac{y1}{g T^2}}}{2}\right)} \frac{2\sqrt{2}\sqrt{-2 g T^2 \sin(w)^2 + 2 g T^2}\; g T^2 \sin(w)\cos(w)}{\pi^3 \sqrt{\frac{g T^2 \sin(w)^2}{\pi^2}}}\, dw$$

Let's try to simplify the integrand.

```
> simplify(2/Pi^3*2^(1/2)/(g*T^2/Pi^2* sin(w)^2)^(1/2)*
(-2*g*T^2*sin(w)^2+2*g*T^2)^(1/2)*g*T^2*sin(w)*cos(w),trig);
```

$$\frac{4\sqrt{g T^2 \cos(w)^2}\; g T^2 \sin(w)\cos(w)}{\pi^2 \sqrt{-g T^2 (-1 + \cos(w)^2)}}$$

Now, Maple is never perfect at simplifying. As humans, we can see that certain trigonometric identities hold, but Maple doesn't always recognize them. We also note that $\cos(w)$ is positive here, but Maple won't take the square root without us telling it to do so explicitly. In practice, we must use our own skills with Maple's to achieve the best results in simplification.

```
> x(y)=int(4*g*T^2*cos(w)^2/Pi^2,w);
```

$$\mathrm{x}(y) = \frac{4\,g\,T^2\left(\frac{1}{2}\cos(w)\sin(w) + \frac{w}{2}\right)}{\pi^2}$$

```
> x(y)=combine(4*g*T^2/Pi^2* (1/2*cos(w)*sin(w)+1/2*w),w);
```

$$\mathrm{x}(y) = \frac{g\,T^2\sin(2\,w) + 2\,g\,T^2\,w}{\pi^2}$$

Therefore, we have a parametric description of the curve $x = \mathrm{x}(y)$ in terms of the parameter w (recall that $y = 2\,g\,T^2\sin(w)^2/\pi^2$ and that the half angle formula simplifies this to $y = g\,T^2\,(1 - \cos(2\,w))/\pi^2$. Substitute $2\,w = \theta$.

```
> xtheta:=(g*T^2/Pi^2)*(sin(theta)+theta);
ytheta:=(g*T^2/Pi^2)*(1-cos(theta));
```

$$xtheta := \frac{g\,T^2\,(\sin(\theta) + \theta)}{\pi^2}$$

$$ytheta := \frac{g\,T^2\,(1 - \cos(\theta))}{\pi^2}$$

This is a cycloid! So Maple (with a little human intervention) discovers the tautochrone to be a cycloid. Let's forget about the constant factors and plot the parametric curve in Figure 1.37 and Figure 1.38.

```
> plot([theta+sin(theta),1-cos(theta), theta=0..4*Pi],
scaling=constrained,view=[0..12,0..2],color=blue);
> plot([theta+sin(theta),1-cos(theta), theta=Pi..2*Pi],
scaling=constrained,view=[3..6.3,0..2],color=magenta,
ytickmarks=4);
```

Now let's change course a bit by plotting and analyzing Viviani's curve from Exercise 1.3.26.

```
> viv:=<a*(1+cos(t))|a*sin(t)|a*2*sin(t/2)>;
```

$$viv := \left[a\,(1 + \cos(t)),\ a\sin(t),\ 2\,a\sin\left(\frac{t}{2}\right)\right]$$

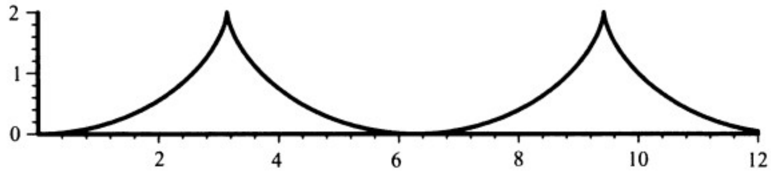

Figure 1.37. The tautochrone

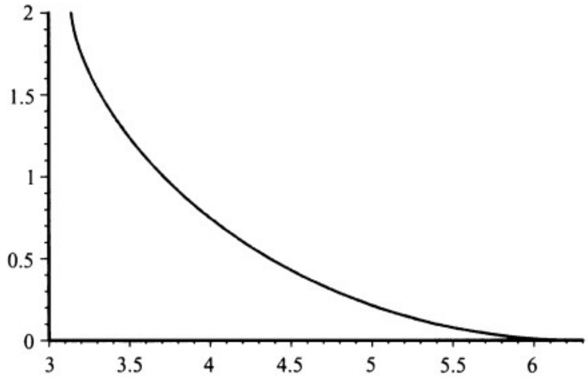

Figure 1.38. The tautochrone: a closer view

We can plot the curve in two ways. Again, we can save plots and display them all at once. We take Viviani's curve, a sphere and an intersecting cylinder. Viviani is the intersection!

```
> viv1:=spacecurve(convert(subs(a=1,viv), list),t=-Pi..Pi,
color=black):
> viv2:=tubeplot(convert(subs(a=1,viv),list), t=-2*Pi..2*Pi,
radius=0.05,color=black):
> sphere:=plot3d([2*cos(u)*cos(v), 2*sin(u)*cos(v),
2*sin(v)],u=0..2*Pi,v=-Pi/2..Pi/2,shading=XY,lightmodel
=light2):
> cyl:=plot3d([cos(u)+1,sin(u),v], u=0..2*Pi,v=-2..2,
shading=Z,lightmodel=light1):
```

Now we display the plots. Note that we use viv2 because "tubeplot" allows us to see the curve of intersection better. See Figure 1.39.

```
> display({viv2,sphere,cyl },scaling=constrained,
orientation= [-27,69]);
```

Now let's verify that Viviani's curve is a spherical curve. We use the criterion for spherical curves given in Exercise 1.3.23–Exercise 1.3.26.

```
> kk:=curv(viv);tt:=tor(viv);
```

$$kk := \frac{1}{2}\,\frac{\sqrt{3\cos\left(\frac{t}{2}\right)^2+5}}{a\left(1+\cos\left(\frac{t}{2}\right)^2\right)^{(3/2)}}$$

$$tt := \frac{3\cos\left(\frac{t}{2}\right)}{\left(3\cos\left(\frac{t}{2}\right)^2+5\right)a}$$

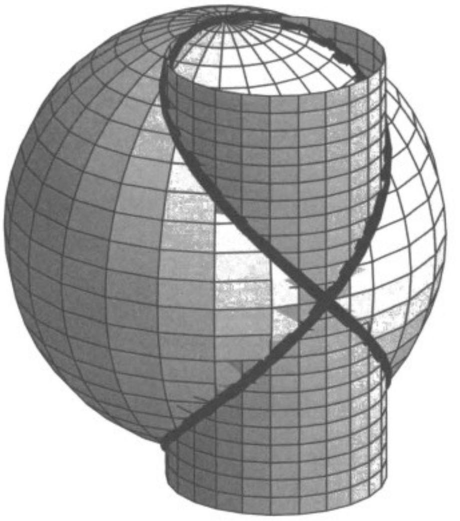

Figure 1.39. Viviani's curve as the intersection of sphere and cylinder

We now check the criterion. We need the following to be a constant.

```
> zz:=simplify(1/kk^2+(1/tt^2)* diff(1/kk,t)^2/Norm(
map(diff,viv,t),Euclidean,conjugate=false)^2,symbolic);
```

$$zz := 4\,a^2$$

Hence, Viviani's curve is truly spherical.

Exercise 1.7.11. Show that the following curve is spherical.

$$(a\cos^2(t),\, a\sin(t)\cos(t),\, a\sin(t))$$

In Exercise 1.3.15, curves of constant precession were discussed. The recreate procedures above may be used to create curves of constant precession by choosing $\kappa(s) = -a\sin(bs)$ and $\tau(s) = a\cos(bs)$ for constants $a > 0$ and b. However, in [Sco95], an explicit parametrization for such curves is derived. Furthermore, it is shown that any such curve lies on a hyperboloid of one sheet (which we shall consider several times in this book) with the formula

$$x^2 + y^2 - \frac{b^2}{a^2}\,z^2 = \frac{4b^2}{a^4}.$$

The explicit parametrization is given by

$$\begin{aligned}
x(s) &= \frac{c+b}{2c(c-b)}\,\sin([c-b]s) - \frac{c-b}{2c(c+b)}\,\sin([c+b]s)\\
y(s) &= -\frac{c+b}{2c(c-b)}\,\cos([c-b]s) + \frac{c-b}{2c(c+b)}\,\cos([c+b]s)\\
z(s) &= \frac{a}{bc}\,\sin(bs)
\end{aligned}$$

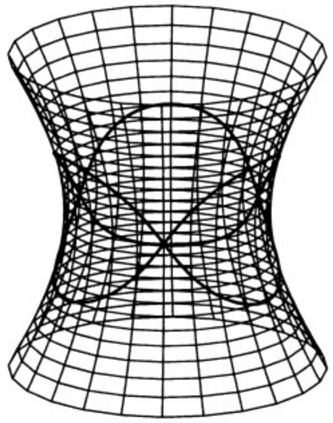

Figure 1.40. Constant precession curve with $a = 4$ and $b = 3$

where $c = \sqrt{a^2 + b^2}$. The following procedure graphs curves of constant precession on the associated hyperboloids.

```
> constprec:=proc(a,b)
local c,alpha,a1,a2,hyp,pl1,pl2;
c:=sqrt(a^2+b^2);
a1:=c+b;
a2:=c-b;
alpha:=[a1/(2*c*a2)*sin(a2*s)-a2/(2*c*a1)*sin(a1*s),
-a1/(2*c*a2)*cos(a2*s)+a2/(2*c*a1)*cos(a1*s),
a/(b*c)*sin(b*s)];
hyp:=[2*b/a^2*cosh(r)*cos(q),2*b/a^2*cosh(r)*sin(q),
2/a*sinh(r)];
pl1:=tubeplot(alpha,s=0..2*Pi,radius=0.01,color=black,
numpoints=200);
pl2:=plot3d(hyp,r=-1..1,q=0..2*Pi);
display({pl1,pl2},scaling=constrained,style=wireframe,
shading=XY);
end:
```

Let's take a curve of constant precession with $a = 4$ and $b = 3$. See Figure 1.40.

```
> constprec(4,3);
```

Exercise 1.7.12. Graph the curves of constant precession with $a = 15$, $b = 8$ and with $a = \sqrt{6}$, $b = \sqrt{2}$. Can you make a conjecture about when curves of constant precession are closed curves? See [Sco95].

Exercise 1.7.13. Graph the unit tangent vector $T(s)$ (i.e., the *tantrix* or *tangent indicatrix*) for curves of constant precession with $a = 15$, $b = 8$ and with $a = \sqrt{6}$, $b = \sqrt{2}$. Use "display" to graph $T(s)$ on the unit sphere. Can you make a conjecture about the number of cusps of such curves? For a discussion of tangent indicatrices of spherical curves, see [Sol96].

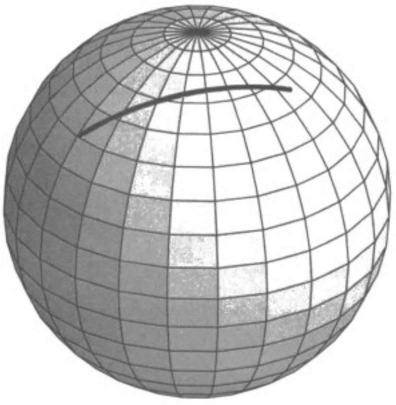

Figure 1.41. New York to Moscow

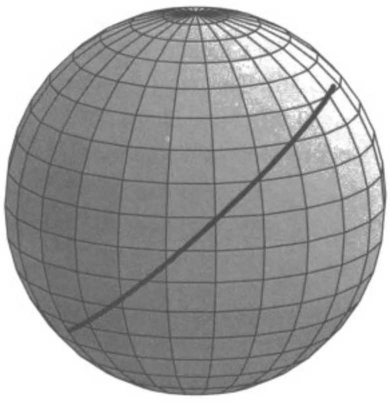

Figure 1.42 Cleveland to Sydney

Exercise 1.7.14. The following procedure draws the great circle arc between two points A and B on Earth and gives the distance between the points (assuming the radius of the Earth to be 3970 miles). A and B are given by longitude and latitude coordinates. For example A=[100,25] means that the point A has longitude 100 degrees and latitude 25 degrees. We take positive longitude to be east of Greenwich, England and negative longitude to be west of it. We also assume that the Earth is a sphere with parametrization (see Chapter 2) ($R\cos(u)\cos(v)$, $R\sin(u)\cos(v)$, $R\sin(v)$), R being the radius of the Earth.

Explain every line of the procedure. This should be a good test of the vector calculus used in this chapter.

```
>  with(LinearAlgebra):with(plots):
>  great_circle:=proc(place1,place2,theta,phi)
local point1,point2,angle,A,B,sphere,circ,alpha,unitcp,
alp1,alp2;
A:=map(evalf,[Pi/180*place1[1],Pi/180*place1[2]]);
B:=map(evalf,[Pi/180*place2[1],Pi/180*place2[2]]);
point1:=<cos(A[1])*cos(A[2])|sin(A[1])*cos(A[2])|sin(A[2])>;
point2:=<cos(B[1])*cos(B[2])|sin(B[1])*cos(B[2])|sin(B[2])>;
angle:=evalf(arccos(DotProduct(point1,point2,conjugate=
false)));
print(‘Distance is‘,3970*angle,‘miles‘);
unitcp:=1/sin(angle)*CrossProduct(point1,point2);
alp1:=ScalarMultiply(point1,cos(t));
alp2:=ScalarMultiply(CrossProduct(unitcp,point1),sin(t));
alpha:=alp1+alp2;
sphere:=plot3d([cos(u)*cos(v),sin(u)*cos(v),sin(v)],u=0..
2*Pi,v=-Pi/2..Pi/2, shading=XY,lightmodel=light2,grid=
[25,19]):
circ:=tubeplot(convert(alpha,list),t=0..angle,radius=0.01,
color=navy):
display({sphere,circ},scaling=constrained,orientation=
[theta,phi]);
end:
```

3- 6 " 5 -M /012/ / /3 / ((
< 0 !!&@ > 0 0 *"

data 17 04 2020

Figure 1.39

Viviani's curve as the intersection of sphere and cylinder.

Hence, Viviani's curve is truly spherical

Exercise 1.7.11

Show that the following curve is spherical

$(a\cos^2(t), a\sin(t)\cos(t), a\sin(t))$

$$x^2+y^2-\frac{b^2}{a^2}z^2=\boxed{\frac{4b^2}{a^4}}$$

checar!

Look: Formula Bhaskara

$$x=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$$

Outra pesquisa: (relacionada)

B1) Books_69_2020/
Complex Numbers and Geometry/
[Spectrum Series] Liang...
Complex numbers and geometry .. djvu

→

17/04/2020

Estude a capa do livro.
(desenho geométrico)

Na pág. 04 deste livro, observamos várias referências aos seguintes livros:

B2) Circles: A Mathematical View by Dan Pedoe;

B3) Journey into Geometries by Marta Sved;

B4) The Trisectors by Underwood Dudley.

obs.: Estes livros estão na mesma pasta.

(I) Books_69_2020/
Complex Numbers and Geometry/
✡ Circles_A Mathematical View
(1997...).pdf

Pág. 09(45)

3. Feuerbach's theorem

The nine-point circle of a triangle touches the incircle and excircles of the triangle.

Obs: Important.

Use-o com os demais para verificação de cálculos.

- /012/ / /3 / ((< 0
!!&@ > 0 0 *"

-# /012/ / /3 / ((B + !!&@
0 0 "E5

-(/012/ / /3 / ((8 " " K " + 0 H "
> @ * *"

3
- /012/ / /3 / ((I * J K " + 8 " "
. !! @ 0 0 "E5

/012/ / /3 / ((< 0 !!&@
> 0 0 *"
' ! (# : - D

> * " 9

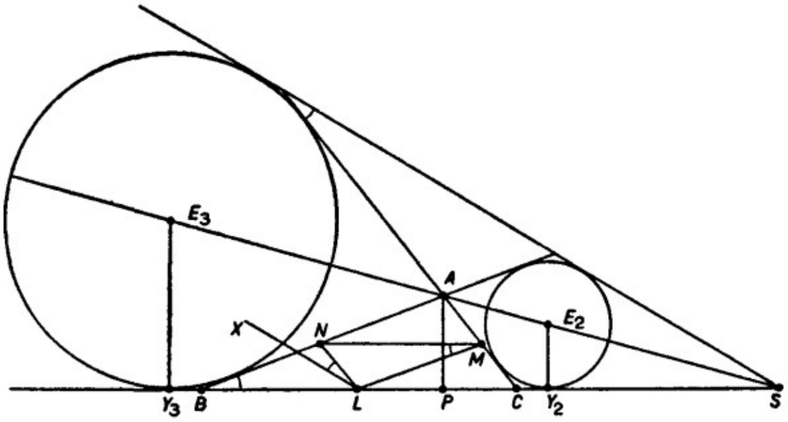

E_3
A
E_2
X
N
M
Y_3
B
L
P
C
Y_2
S

: (

3 *

> " -+ 5

' & (#

CHAPTER I

THE properties of circles discussed in this chapter are those which have the habit of appearing in other branches of mathematics.

1. The nine-point circle

This circle is the first really exciting one to appear in any course on elementary geometry. It is a circle which is found to pass through nine points intimately connected with a given triangle ABC. Three proofs are given. Each has its own peculiar merits.

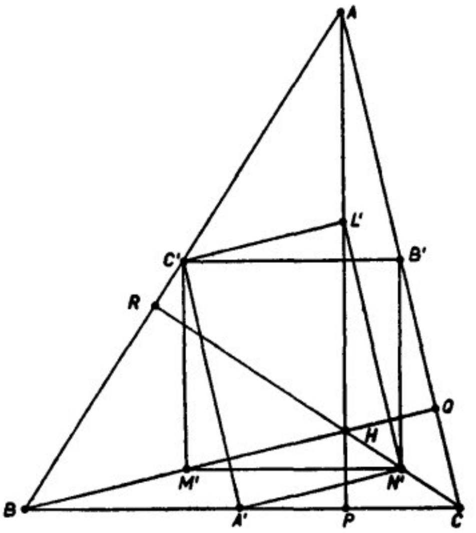

FIG. 1

(i) H is the orthocentre of triangle ABC, the intersection of the altitudes AP, BQ, CR. The midpoints of BC, CA and AB are A', B' and C' respectively. L', M', N' are the midpoints of HA, HB, HC. We show that a circle passes through the nine points A', B', C', L', M', N', P, Q, R. We use the theorem that the line joining the midpoints of two sides of a triangle is parallel to the third side.

Both $B'C'$ and $N'M'$ are $\parallel$ to BC, and both $B'N'$ and $C'M'$ are $\parallel$ to AH. Hence $B'C'M'N'$ is a rectangle. Similarly $C'A'N'L'$ is a rectangle.

Hence $A'L'$, $B'M'$ and $C'N'$ are three diameters of one circle. Since $A'P$ and $L'P$ are $\perp$, this circle passes through P. Similarly it passes through Q and R.

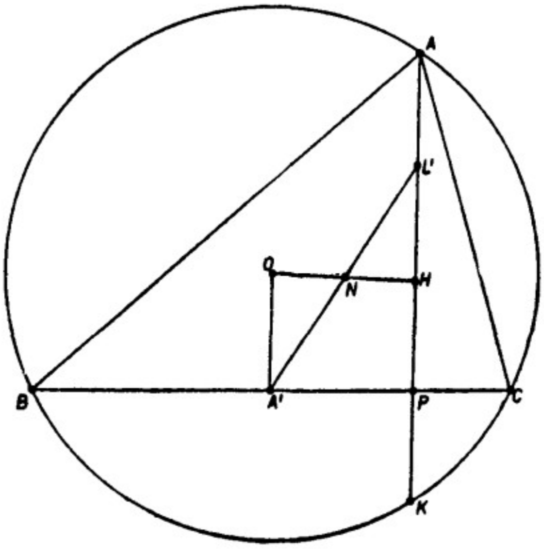

FIG. 2

(ii) Let O be the circumcentre of ABC and R the circumradius. We know that $AH = 2OA'$. Since L' bisects AH, $L'H = OA'$, and therefore OH and $L'A'$ bisect each other, in N, say.

Now $NL' = NA' = NP$ (since $\angle APB = 90°$). We also know that $HP = PK$, and since $ON = NH$, we have

$$NP = \tfrac{1}{2}OK = \tfrac{1}{2}R.$$

Therefore the circle centre N and radius $\frac{1}{2}R$ passes through P, A', L'. Similarly it passes through Q, B', M' and R, C', N'.

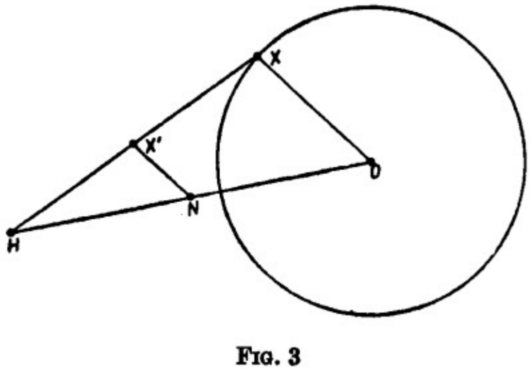

FIG. 3

(iii) The third proof uses a lemma from the theory of similar figures. If X is a point on a given circle of centre O and radius R, and H is any given fixed point, the locus of X', the midpoint of HX, is a circle centre N, where N bisects HO, and radius $\frac{1}{2}R$. The proof is clear.

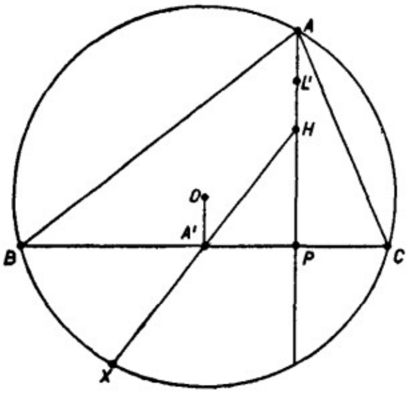

FIG. 4

Applying this lemma to the case of the circumcircle of ABC, H being the orthocentre, we see that a circle centre N and radius $\frac{1}{2}R$ passes through P, Q, R, L', M', N'. We must now show that this circle also passes through A', B', C'. It is sufficient to show that if HX contains A', then HX is bisected at A'.

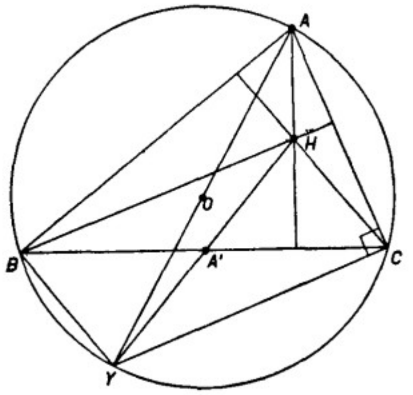

FIG. 5

Let AO meet the circumcircle in Y. Then $\angle ACY = 90°$, so that $YC \parallel BH$. Since also $\angle ABY = 90°$, $YB \parallel CH$. Therefore $YBHC$ is a parallelogram, and the diagonals YH and BC

bisect each other. Hence Y is the same point as X in the previous diagram, and the theorem is proved.

The methods used above are elementary, and at least one of the three proofs will be familiar to most readers of this book. In chapter II we shall need to assume some theorems in inversion and coaxal systems of circles. As fashions in mathematical teaching change, it cannot be assumed that all readers will be familiar with these theories, and a brief (but adequate) outline of each will now be given.

2. Inversion

This is a one-to-one transformation of the points of the plane by means of a given circle, which we shall call Σ, of radius k and centre O. To obtain the transform X' of any given point X, or the *inverse* of X in the circle Σ, we join X to O, and find X' on OX such that $OX.OX' = k^2$. The point O itself is excluded from the points of the plane which may be transformed. It is clear that

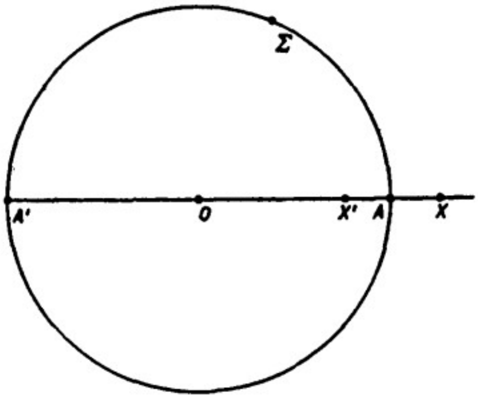

FIG. 6

(i) X is the inverse of X';
(ii) points on Σ transform into themselves;
(iii) if A, A' are the ends of the diameter of Σ through X, then X, X' are harmonic conjugates with respect to A, A'.

All circles through X and X' cut Σ orthogonally, since the square of the tangent from O to such a circle $= OX.OX' = k^2$.

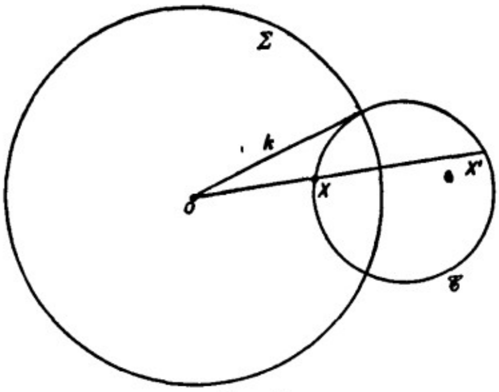

FIG. 7

As a corollary to this remark, we see that if $\mathscr{C}$ is any circle orthogonal to Σ, all points on $\mathscr{C}$ invert into points on $\mathscr{C}$, for $OX.OX' = k^2$. It is sometimes necessary to consider the process of inversion when Σ, the circle of inversion, is a line. If we keep A fixed, and let A' move off to infinity, we see that X' moves towards the geometrical image of X in the resulting line.

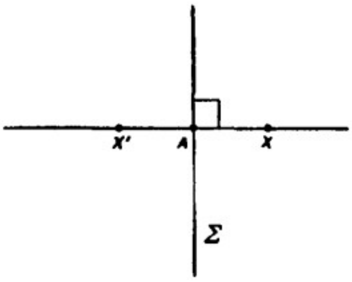

FIG. 8

For future application we now give a construction, using only a

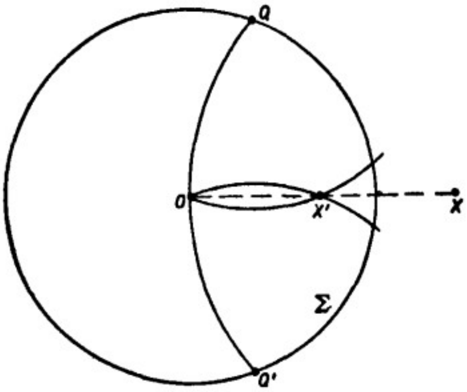

FIG. 9

pair of compasses, by which we can find the inverse of a point X in a circle Σ of given centre O. Let the circle centre X and radius XO cut Σ in Q and Q', and let O and X' be the intersections of the two circles with centres Q and Q' and radii QO and $Q'O$ respectively. Then X' is the inverse of X in Σ.

For QOX and $X'OQ$ are isosceles triangles with a common base angle. They are therefore similar, and $OX : OQ = OQ : OX'$. Hence $OX . OX' = OQ^2$.

This construction lies at the basis of the discussion on *compass geometry* in §11.

We are interested in the locus of inverse points X' as X describes a given curve $\mathscr{C}$. This locus is called the *inverse of the curve* $\mathscr{C}$ with respect to Σ. When $\mathscr{C}$ is a circle orthogonal to Σ, we have seen that the inverse is $\mathscr{C}$ itself. The general theorem is:

The inverse of a circle is a straight line or a circle.

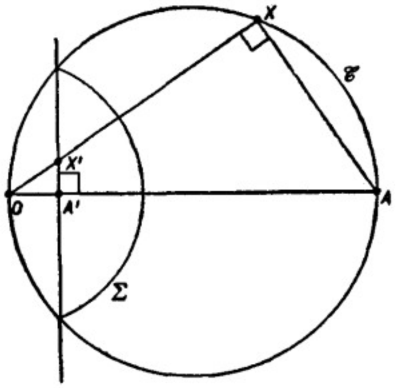

FIG. 10

(i) The centre of inversion O is on $\mathscr{C}$:

Let OA be the diameter of $\mathscr{C}$ through O, and let A' be the inverse of A. Then $OA . OA' = k^2$. If X is any point on $\mathscr{C}$, and X' its inverse,

$$OX . OX' = k^2 = OA . OA'.$$

Hence A, X, X', A' are concyclic points, and since $\angle AXO = 90°$, $\angle X'A'A = 90°$. Therefore X' describes the line through $A' \perp$ to OA.

We note that this line is $\|$ to the tangent to $\mathscr{C}$ at O. It naturally passes through the intersections of $\mathscr{C}$ and Σ, since points on Σ invert into themselves.

Corollary. The inverse of a straight line is a circle through the centre of inversion.

(ii) The centre of inversion O is not on $\mathscr{C}$. Let C be the centre of $\mathscr{C}$. Since

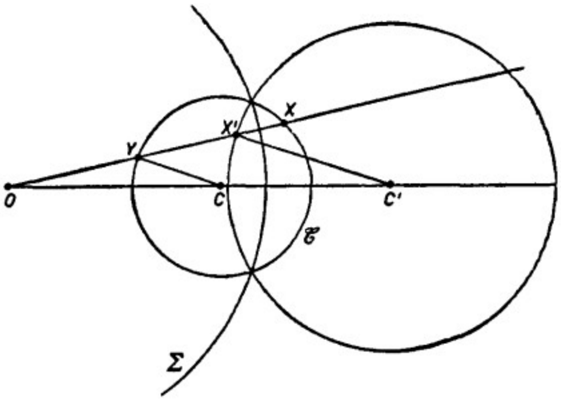

FIG. 11

$$OX.OX' = k^2,$$

and
$$OX.OY = t^2,$$

where t is the tangent[†] from O to $\mathscr{C}$, by division we have

$$\frac{OX'}{OY} = \frac{k^2}{t^2}.$$

Since this ratio is constant, it follows that if C' is a point on OC such that $C'X' \parallel CY$, then C' is a fixed point, and $C'X'$ is constant. Hence X' moves on a circle centre C'.

Note that C' is not necessarily the inverse of C. We shall soon find what point the centre of $\mathscr{C}$ does invert into.

We need only one further theorem from the theory of inversion:

The angle of intersection of two circles is unaltered by inversion.

In fact a more general theorem is true, but this one will suffice for our subsequent needs.

[†] We shall see later that t^2 may be negative. Then O is inside $\mathscr{C}$.

(i) If both circles $\mathscr{C}$ and $\mathscr{D}$ pass through the centre O of inversion, they transform into straight lines parallel to the tangents to $\mathscr{C}$ and $\mathscr{D}$ at O, so that the theorem is evident.

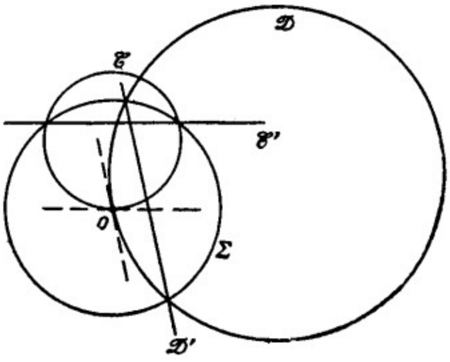

FIG. 12

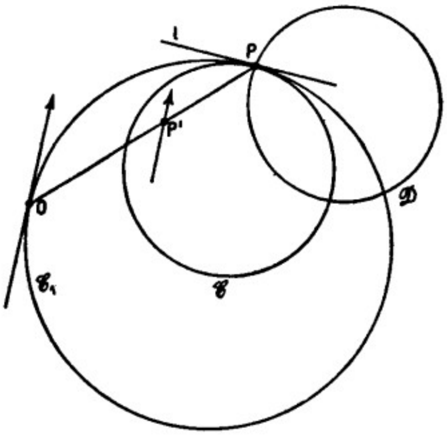

FIG. 13

(ii) If at least one of the circles $\mathscr{C}$, $\mathscr{D}$ does not pass through O, let P be a point of intersection of the two circles, and let l, m be the tangents at P to $\mathscr{C}$, $\mathscr{D}$ respectively. We may draw a circle $\mathscr{C}_1$ through O which touches l at P, and a circle $\mathscr{D}_1$ through O which touches m at P. Then the angle between $\mathscr{C}_1$ and $\mathscr{D}_1$ is the same as that between $\mathscr{C}$ and $\mathscr{D}$.

Now if P' is the inverse of P, the tangent at P' to the inverse of $\mathscr{C}$ is the inverse of $\mathscr{C}_1$. Hence the angle between the circles inverse to $\mathscr{C}$ and to $\mathscr{D}$ is equal to the angle between $\mathscr{C}_1$ and $\mathscr{D}_1$ (by (i) above), and this is equal to the angle between $\mathscr{C}$ and $\mathscr{D}$.

With the help of this theorem we can now find the inverse of the centre of a circle $\mathscr{C}$. All lines through the centre C cut $\mathscr{C}$ orthogonally. These lines invert into circles through the centre of inversion O and the inverse of C. Since all circles through O and the inverse of C cut the inverse of $\mathscr{C}$ orthogonally, the inverse of C must be *the inverse of O in the inverse of* $\mathscr{C}$.

If O lies on $\mathscr{C}$, so that the inverse of $\mathscr{C}$ is a straight line, the centre of $\mathscr{C}$ inverts into the geometrical image of O in this line.

We now give some applications of the theory of inversion.

3. Feuerbach's theorem

The nine-point circle of a triangle touches the incircle and excircles of the triangle.

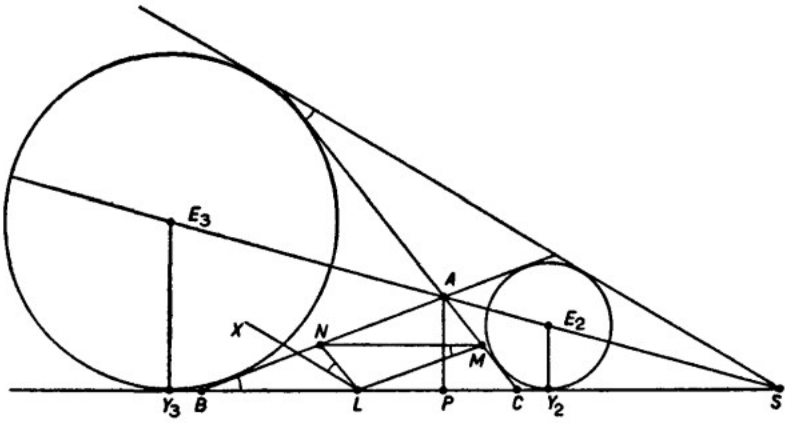

FIG. 14

We consider the excircles opposite B and C. Let their centres be E_2 and E_3, and their points of contact with BC Y_2 and Y_3. The points A, S divide E_2E_3 internally and externally in the same ratio, that of the radii of the two excircles. Dropping perpendiculars on BC, it follows that P and S also divide Y_2Y_3 internally and externally in the same ratio. That is, P and S are harmonic conjugates with respect to Y_2 and Y_3.

Let L, M, N be midpoints of BC, CA, AB. We note that $BY_2 = \frac{1}{2}(BC + CA + AB) = CY_3$, so that L is also the midpoint of Y_2Y_3. From the harmonic relation just proved we see that

$$LP \, . \, LS = LY_2{}^2 = LY_3{}^2.$$

Now let LX be the tangent at L to the circle LMN, where X is on the same side of BC as N. Then

$$\angle XLN = \angle NML = \angle ABC.$$

Hence the angle between LX and $CA = \angle ABC$. It follows that LX is $\parallel$ to the fourth common tangent of the two excircles, the common tangent through S other than BC.

Now since $LP.LS = LY_2^2 = LY_3^2$, the nine-point circle LMN is the inverse of this fourth common tangent with respect to the circle centre L and radius LY_2. For this nine-point circle contains P, and inverts into the line through S parallel to the tangent to the circle at L.

Also since the circle of inversion cuts both excircles orthogonally, each excircle inverts into itself. We have seen that the inverse of the nine-point circle touches these excircles. It follows that the nine-point circle touches the excircles centres E_2, E_3. In precisely the same way we can show that the nine-point circle touches the incircle and the excircle opposite A.

4. Extension of Ptolemy's theorem

If A, B, C, D are any four points in a plane, then

$$AB.CD + AD.BC > AC.BD,$$

unless A, B, C, D lie, in the order $ABCD$, on a circle or a straight line. In the latter case, the inequality becomes an equality.

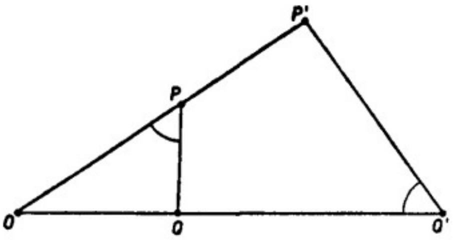

FIG. 15

This is proved by noting the influence of inversion on the length of a segment. If the points P, Q invert into the points P', Q', with O as centre of inversion, then OPQ and $OQ'P'$ are similar triangles. Hence

$$\frac{P'Q'}{PQ} = \frac{OP'}{OQ} = \frac{OP.OP'}{OP.OQ} = \frac{k^2}{OP.OQ},$$

so that

$$P'Q' = \frac{k^2\,PQ}{OP.OQ}.$$

Now, given the four points A, B, C, D, we invert with respect to A. Let B', C', D' be the respective inverses of B, C, D. Then $B'C' + C'D' > B'D'$ unless C' lies on the line $B'D'$ between B' and D′. In the latter case we have

$$B'C' + C'D' = B'D'.$$

The first inequality becomes

$$\frac{BC}{AB.AC} + \frac{CD}{AC.AD} > \frac{BD}{AB.AD},$$

or

$$AB.CD + AD.BC > AC.BD,$$

unless C' lies on the line $B'D'$ between B' and D', in which case A, B, C, D lie on a circle in the order $ABCD$, or on a straight line.

5. Fermat's problem

A, B, C are any three points in a plane. To find a point P such that $PA + PB + PC$ shall be least.

Let B and C be the acute angles of the triangle ABC. On BC, and away from A, describe an equilateral triangle BCD. Then by the extension of Ptolemy's theorem, unless P lies on the circle BCD so that the order of the points is $BPCD$, we have

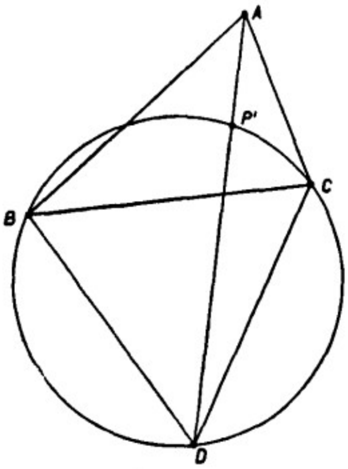

FIG. 16

$$BP.CD + PC.DB > PD.BC,$$

or

$$PB + PC > PD,$$

since $CD = DB = BC$. Therefore

$$PA + PB + PC > PA + PD.$$

Now, unless P lies on AD, we have $PA + PD > AD$. Hence, unless P is at P' (the other intersection of AD with the circle BCD), we have $PA + PB + PC > AD$.

But if P is at P', both the above inequalities become equalities; so that $P'A + P'B + P'C = AD$.

Hence $P'A + P'B + P'C < PA + PB + PC$.

Therefore P' is the required point.

If $\angle BAC = 120°$, $A = P'$, and A is the required point.

If $\angle BAC > 120°$, A is still the required point.

6. The centres of similitude of two circles

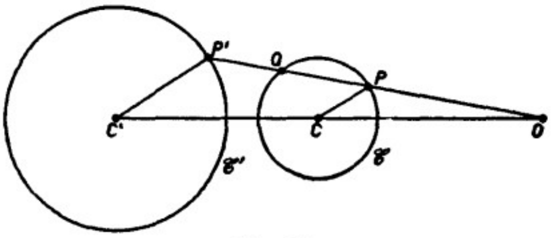

FIG. 17

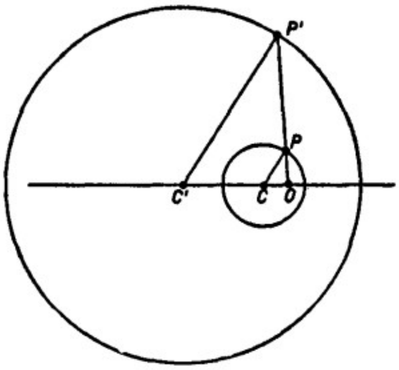

FIG. 18

These points have already appeared in some of our proofs. We now define them.

External centre of similitude. Let $\mathscr{C}$ and $\mathscr{C}'$ be two circles of unequal radii, with centres C and C'. Let CP and $C'P'$ be parallel radii drawn in the *same* sense. Then it is clear that $P'P$ cuts $C'C$ in a fixed point O, which divides $C'C$ externally in the ratio of the radii of the two circles. This is the external centre of similitude. If the ratio of the radii $= k$, then $OP' : OP = k$. We also have $OP . OQ =$ constant. Therefore

$$OP' . OQ = \text{constant}.$$

It follows that with centre of inversion O, and a suitable radius of inversion, $\mathscr{C}$ can be inverted into $\mathscr{C}'$.

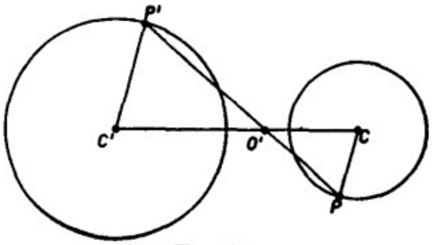

FIG. 19

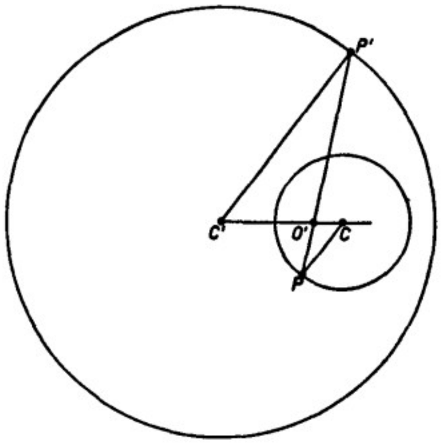

FIG. 20

Internal centre of similitude. Here the radii of the circles $\mathscr{C}$ and $\mathscr{C}'$ may be equal. CP and $C'P'$ are parallel radii in opposite senses. $P'P$ cuts $C'C$ in a fixed point O', the internal centre of similitude, which divides $C'C$ internally in the ratio of the radii.

Direct common tangents of $\mathscr{C}$ and $\mathscr{C}'$, if they exist, cut the line of centres in O, and transverse common tangents, if they exist, cut the line of centres in O'.

If $\mathscr{C}$ and $\mathscr{C}'$ are concentric, the common centre is both the internal and external centre of similitude.

If $\mathscr{C}$ and $\mathscr{C}'$ touch externally, that is, if they lie on opposite sides of the tangent at the point of contact, this point is the *internal* centre of similitude. If the circles touch internally, the point of contact is the *external* centre.

If we have three circles, no two of which have equal radii, and the centres of which form a triangle, the six centres of similitude

3- % 5 *

lie in threes on four straight lines. This is easily proved, using the theorem of Menelaus.

7. Coaxal systems of circles

These are best introduced by using the methods of coordinate geometry.

The equation of any line passing through the origin of coordinates in the (x,y)-plane may be written as

$$\lambda x + \mu y = 0.$$

These lines are said to form a *pencil* of lines, and the equation of the system depends *linearly* on two distinct lines of the pencil, given by $x = 0$ and $y = 0$.

Similarly, if $u \equiv ax + by + c = 0$ and $v \equiv a'x + b'y + c' = 0$ are any two distinct lines, the line

$$\lambda u + \mu v = 0$$

represents, for suitable λ, μ, any line through the intersection of $u = 0$ and $v = 0$.

When we consider two distinct circles $\mathscr{C}$ and $\mathscr{C}'$, given by the equations

$$C \equiv x^2 + y^2 + 2gx + 2fy + c = 0,$$
$$C' \equiv x^2 + y^2 + 2g'x + 2f'y + c' = 0,$$

the system $\lambda C + \mu C' = 0$, derived linearly from them, is a system of circles, since it may also be written as

$$x^2 + y^2 + \frac{2x(\lambda g + \mu g')}{\lambda + \mu} + \frac{2y(\lambda f + \mu f')}{\lambda + \mu} + \frac{\lambda c + \mu c'}{\lambda + \mu} = 0.$$

If (x',y') is a point of intersection of $\mathscr{C}$ and $\mathscr{C}'$,

$$C(x',y') = C'(x',y') = 0,$$

that is, the point satisfies the equations of both circles. Since also, for all values of λ, μ,

$$\lambda C(x',y') + \mu C'(x',y') = 0,$$

this same point lies on all circles of the system $\lambda C + \mu C' = 0$. Such a system of circles therefore passes through the points of intersection (if any) of $\mathscr{C}$ and $\mathscr{C}'$. The system may be described as a *pencil* of circles, but for reasons which will soon become clear, we give it the more normal title of a *coaxal system* of circles.

A geometrical meaning may be given to the ratio of the parameters $\lambda : \mu$. Let $P\,(x',y')$ be any point in the plane, and let any line through P cut $\mathscr{C}$ in Q and R. Then it is elementary that the product

$$PQ.PR = x'^2 + y'^2 + 2gx' + 2fy' + c.$$

17 04 2020

Estude a capa do livro.
(desenho geométrico)

Na pág. 04 deste livro, observamos várias referências aos seguintes livros:

B2) Circles: A Mathematical View by Dan Pedoe;

B3) Journey into Geometries by Marta Sved;

B4) The Trisectors by Underwood Dudley.

obs.: Estes livros estão na mesma pasta.

(I) Books_69_2020/
Complex Numbers and Geometry/
✡ Circles_A Mathematical View
(1997...).pdf

Pág. 09(45)

3. Feuerbach's theorem

The nine-point circle of a triangle touches the incircle and excircles of the triangle.

Obs: Important.

Use-o com os demais para verificação de cálculos.

data 17.04.2020
D S T Q Q [✓ S] S

Use juntamente com o item
nº 138 (20200416),
letra F que se refere:
a) Fano Plane (Steiner System);
b) 7 lines of the plane
3 points - one line
c) Pappu's theorem.

Obs: Source deste book:
ARQ-IMPORTANT/
Cinderella.de/
the Cinderella.2 Manual.pdf
Pág. 42(57)
3.1 Pappu's theorem
(forma de construção).

other: (same book)
Pág. 10(46)
4. Extension of Ptolemy's theorem

Jandaia

/012/ / /3 / ((> " *"
' (& # ** D > " UT

- *Scripting the Seeds of a Sunflower* (p. 58)*:* Here you will learn how to perform simple drawing operations with CindyScript. You will also learn how to combine CindyScript with the geometry part of Cinderella.

3.1 Pappus's Theorem

Pappus's theorem is one of the most fundamental theorems in projective geometry. In a certain sense it is the smallest example of an incidence theorem. In this step-by-step example we will construct an instance of the theorem that can be selected and dragged.

3.1.1 Drawing Your First Point

When you start *Cinderella.2*, the first window that appears is a "Euclidean view." This window has a large toolbar that is the key to most of *Cinderella's* functionality. Below this toolbar is the drawing surface on which you perform operations by constructing and dragging the elements you need.

You may notice that in the toolbar one button is slightly darker. This indicates the current mode of *Cinderella*.

All mouse actions in the drawing surface refer to the selected mode. For drawing a single point you have to switch to the add a point (p. 74) mode by clicking it. Now *move the mouse over the drawing surface and click the left button.* A new point is added and labeled with a capital letter.

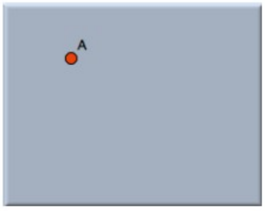

Figure 1: Your first point

Before you add a second point, continue reading. If you hold the left mouse button down while adding a point, you are able to move the point around. The definition of the point will be adapted to the geometric situation at the current location of the mouse pointer. (So far, there is not a lot of geometry in our drawing, but this will change soon.) *Move the mouse over the drawing surface, press the left button, and hold it. Now move the mouse.* You will notice that the new point sticks to the mouse pointer. Coordinates that indicate the current position are shown. When you approach an existing point (*try it*), the new point snaps to the old point. Only after you *release the button* is the new point added to the construction. If you release the mouse over an old point, no new point is added. *Play with these features and add a few new points.*

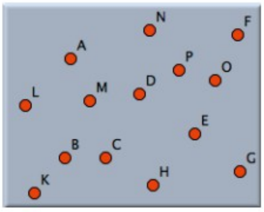

Figure 2: Many points

3.1.2 Undoing an Operation

Your screen may look a bit crowded now. There is an undo operation that inverts the actions you have performed. By pressing the undo button, you will cause the last point you added to disappear. *Click the undo button until exactly two points remain on the drawing surface*. You can use the undo button whenever you make a mistake. You can undo as many consecutive operations as you want.

3.1.3 Moving a Point

We want to continue with our construction of Pappus's theorem. You should first move the two remaining points A and B to approximately the positions shown in Figure 3.

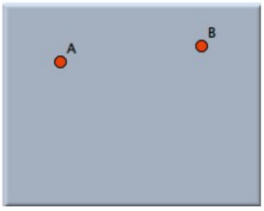

Figure 3: Move to this situation.

Most probably, your points are not already there. To move them, *select move mode* (p. 71) *by pressing the "move mode" button* *in the toolbar*. Now mouse actions in the drawing surface no longer add points; instead, you can select points and move them around. *Move the mouse pointer over the point you want to move. Press the left mouse button. Hold it and drag the mouse.* The point follows the mouse pointer. You will also notice that in this mode a point does not snap to other points. When you *release the mouse button*, the point is placed. In general, move mode allows free elements of a construction to be moved. The rest of the construction will change accordingly.

3.1.4 Adding a Line

We are now going to add a line from *A* to *B*. *Switch to add a line* (p. 76) *mode by pressing the button* . In this mode you can add a line between two points using a press–drag–release sequence with the mouse. *Move over point A. Press the left mouse button. Hold it and move the mouse over point B. Release the mouse.*

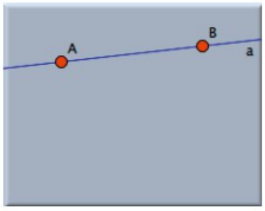

Figure 4: Add a line

This action should have created the desired line. You might have noticed a few things: When you pressed the mouse over point *A*, the point became highlighted. This means that you selected this point as the starting point of the line. While you drag the mouse, a second point is always present at the mouse position. When you release the mouse button, this second point will be added at the release point. When you approach another point, it will be highlighted, and the second point will snap to it. This is the action you used for attaching the line to point *B*.

If you did not notice these things or made a mistake, undo your operations and try again. Your final picture should look like Figure 4 before you proceed.

3.1.5 Adding More Lines

Now you are ready to add three more lines, in order to end up with Figure 5. You can do this with just three mouse operations. *First, move over point B and press–drag to the position of the not-yet-existing point C. Do another press–drag–release from C to D, and finally from D to E.*

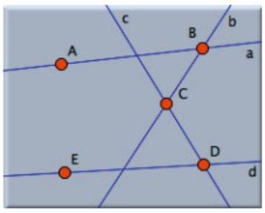

Figure 5: More lines

Observe that you have done all this without leaving "add a line" mode. You did some unnecessary work when you added the points *A* and *B* in add a point (p. 74) mode. You could have done this directly in add a line (p. 76) mode, since not only will the second point be created if necessary, but also the first one. The lines created so far were automatically labeled with the lowercase letters *a* to *d*.

3.1.6 Creating Points of Intersection

We will stay in add a line (p. 76) mode. Now we want to draw a line from point *E* to the intersection of the lines *a* and *c*. *Move the mouse over E. Press the button. Hold it and move to the intersection of a and c. Release the mouse.*

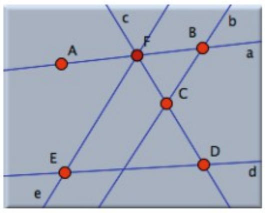

Figure 6: Adding an intersection

When you approached the intersection of the lines, they were highlighted and the endpoint of the new line snapped to the intersection. When you released the mouse, the new point was defined to be the intersection of the lines. If you move any one of the points *A* to *E*, the new line will follow the moves accordingly.

The new point *F* is drawn slightly darker. This means that it is not possible to move *F* freely in move mode. Thus, *F* is a "dependent" point. The other points in the construction are "free."

Two things are worth mentioning: With the same procedure you could also have added a "half-free" point that is bound to only one line. To do so, you need to release the mouse when you are over only one line. All these operations of creating intersections (dependent points), free, and half-free points work similarly in add a point (p. 74) mode. Many other modes, the interactive modes (p. 74), also come with this feature. Browse the reference guide of this manual for further information.

3.1.7 Finishing the Drawing

Now it should be easy for you to finish the drawing by adding four more lines to achieve the final configuration shown in Figure 7. *First add a line from A to the intersection of b and d. Then draw two lines from A to D and from B to E. Finally, draw a line from C to the intersection of e and f.*

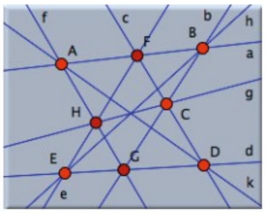

Figure 7: Pappus's theorem

You have created a configuration of eight points and nine lines. If you look at the configuration, you will notice that if everything was done correctly, the lines g, h, and k meet in a point. That this will always be the case in such a construction is the content of Pappus's theorem. *Switch to move mode* (p. 71) *by clicking* *in the toolbar, and drag around the free points of the construction to convince yourself of the truth of this theorem.* This theorem was known to the ancient Greek geometers and is named for Pappus of Alexandria, who flourished in the fourth century C.E. Later, the theorem turned out to be of fundamental importance for the theory of *projective geometry*.

3.1.8 Selecting and Changing the Appearance of a Drawing

You may not like the appearance of a drawing. The lines may be too thin, and perhaps you want to emphasize the final conclusion of the theorem. This can be changed by using the inspector (p. 153) of *Cinderella.2*. For this *choose the menu item "Edit/Information"*. The following window will pop up:

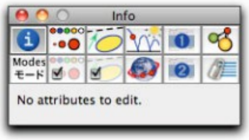

No attributes to edit.

Figure 8: The inspector

There are several tabs in the inspector that refer to different aspects of the construction and the selected objects. To change the appearance of objects, go to the appearance tab by *clicking the button* . If no elements are selected, the inspector shows that there are no attributes to edit.

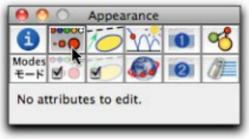

No attributes to edit.

Figure 9: Selecting the appearance tab

As soon as an element is selected, the editable attributes of this elements are shown. *Now select all lines by clicking the button* *"Select all lines"*. The inspector window then looks as follows:

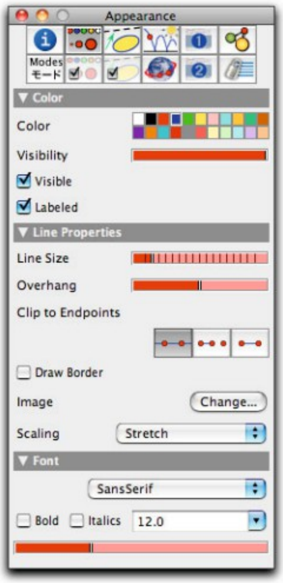

▼ Color

Color

Visibility

☑ Visible

☑ Labeled

▼ Line Properties

Line Size

Overhang

Clip to Endpoints

☐ Draw Border

Image Change...

Scaling Stretch

▼ Font

SansSerif

☐ Bold ☐ Italics 12.0

Figure 10: The appearance tab

Changes in the inspector are immediately applied to selected elements. If you now *move the slider "Line Size" in the inspector to its third tick*, all lines become thicker. After this, *switch to move mode* . Now, when you click over one or more elements, those elements will be selected. If you hold the shift key while you click, the selection state of the element will be toggled.

Click over line k, and then hold the shift key and click over lines h and g. The lines *g*, *h*, and *k* should be highlighted. *Click the red color box in the inspector's color palette.* The color of the three lines changes from blue to red. You can also change the size of these lines by moving the "Line Size" slider to its fourth tick. Finally, *deselect everything* by clicking the button .

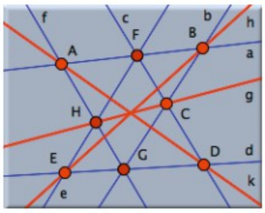

Figure 11: Changed appearance

3.1.9 Adding a Final Point and Proving a Theorem

Before continuing, open *Cinderella's* information window by choosing the menu item *Views/Information Window*. A console window pops up. *Cinderella.2* uses this console to report nontrivial facts about a configuration.

We will now create a final point at the intersection of the three red lines. This time we will add a point by *changing to intersection* (p. 92) *mode* . This mode expects you to mark two lines. A new point will be added at the intersection of these two lines. *Click two of the red lines, say g and h*. Their intersection point will be added.

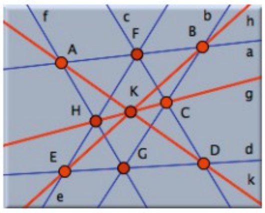

Figure 12: The point of intersection

Note that the newly added point *K* will always be incident to line *k* because of Pappus's theorem. The information console reports this remarkable fact automatically.

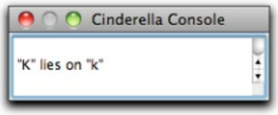

```
"K" lies on "k"
```

Figure 13: A proven theorem

You may wonder how this "theorem proving" works. *Cinderella* does not use symbolic methods to create a formal proof, but a technique called "randomized theorem checking." First the conjecture, "it seems that line *k* always passes through point *K*," is generated. Then the configuration is moved into many different positions, and for each of these it is checked whether the conjecture holds. It may sound ridiculous, but generating enough (!) random (!) examples in which the theorem holds is at least as convincing as a computer-generated symbolic proof. *Cinderella* uses this method over and over to keep its own data structures clean and consistent.

3.1.10 Moving Points to Infinity

Let us explore the symmetry properties of Pappus's theorem. For a nicer picture, we want to get rid of the line labels and make the points and lines a bit smaller. *First select all lines using . Turn off the labeling by pressing the corresponding button in the inspector. Change the line size to "thin" (use the slider in the inspector). Then select all points using . Make them smaller using the "Point Size" slider in the* *inspector (observe that the inspector shows always information about the selected elements). Now choose the menu option "Views/Spherical View"*. What you get will not look very instructive at first glance.

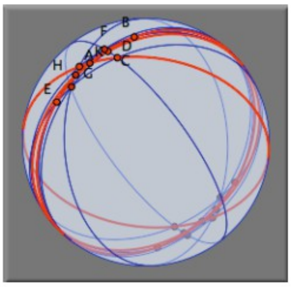

Figure 14: A spherical view

The spherical view shows a central projection of the plane to the surface of a ball. The projection point is the center of the sphere (see Figure 15). Every point is mapped to an antipodal pair of points, and each line is mapped to a great circle (a circle on the sphere dividing it into two equal hemispheres) of the ball.

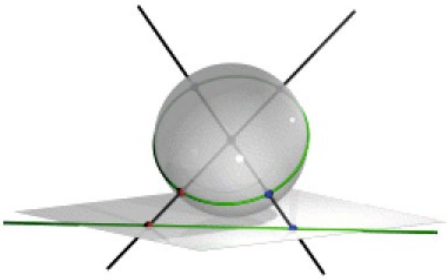

Figure 15: Projection from the plane to the ball

In spherical view there is a little red slider for controlling the distance of the ball to the original plane. Moving this slider corresponds to a kind of "zooming" operation on the sphere. Please *move the slider from its original position much more to the right.* Then the situation on the ball should become clearer. If you have adjusted the slider correctly, it should look somewhat like Figure 16. You should clearly recognize the original construction, now drawn on the surface of this ball.

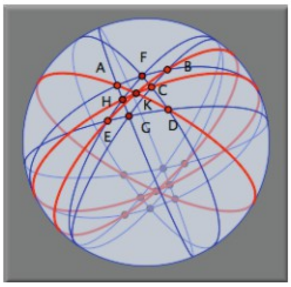

Figure 16: Pappus's theorem on a ball

The picture needs some explanation: For each point in the plane consider the line spanned by the point and the center of the ball. The intersection of this line with the surface of the ball gives the pair of antipodal points. For each line consider the plane spanned by the line and the center of the ball. The intersection of this plane and the surface of the ball is the great circle that represents the line in spherical view.

On the sphere there are points that have no correspondence in the Euclidean plane. If the ball touches the Euclidean plane at the "south pole," then the equator of the ball corresponds to the "points at infinity" in the usual Euclidean plane. In *Cinderella* it is possible to make manipulations in any currently open view. Therefore you can also move points in spherical view. The changes will be instantly reported to the Euclidean view. In particular, you can grab a point in spherical view and move it to what is really infinity in the Euclidean view.

We will do that now, in order to observe that our configuration has a nice threefold symmetry. *Select point A (in spherical view) and move it to the 11 o'clock position of the boundary.* This point is now really located "at infinity." Notice that in Euclidean view the lines passing through *A* became parallel, corresponding to the notion that "parallel lines meet at infinity." In a similar way, *move point E to the 7 o'clock position on the boundary and move point C to the 3 o'clock position.* Your spherical picture should then look like the ball in Figure 17. In Euclidean view you find three bundles of parallel lines, representing a kind of Euclidean specialization of Pappus's theorem. If you like, you can try to figure out the corresponding statement in terms of parallels and incidences.

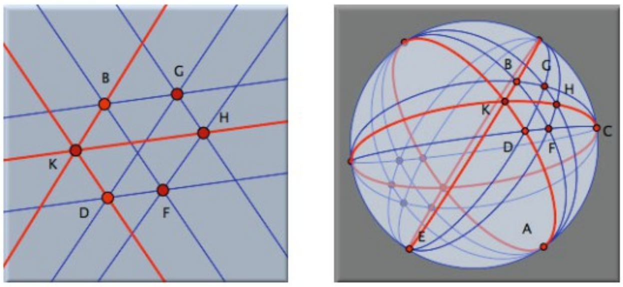

Figure 17: A threefold symmetry

It may happen that the drawing in Euclidean view is too big or too small. You can use the zooming tools to change this. There is a “Zoom-In” and a “Zoom-Out” tool below the Euclidean view. With a press–drag–release sequence you can mark the region to be zoomed in or out.

3.2 A Three-Bar Linkage

In the second tutorial we want to explore the dynamics of a little mechanical linkage. In particular, you will learn how to make bars of a fixed length. You will also learn how to produce a geometric locus of a point.

3.2.1 *Making a Bar*

After starting *Cinderella* or erasing a previous configuration, *switch to circle by radius* (p. 82) *mode by pressing the button* . This mode is used for working with circles whose radii remain constant when their centers move. *Move the mouse pointer over the view. Press the left mouse button. Hold the button while dragging the mouse. Release the mouse.* With these actions you have generated a circle. Switch to move mode (p. 71) to see how the circle behaves. When you select the center you can move the position of the circle, and the radius stays constant. You can also select the circle itself and move it. Then its center stays fixed and the radius of the circle changes.

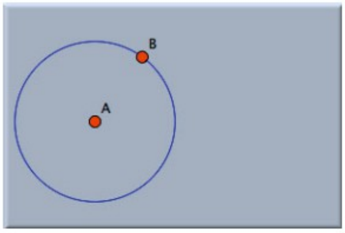

Figure 1: A first circle

We now want to add a point that is bound to the boundary of the circle. *Select add a point* (p. 74) *mode. Move the mouse over the view. With the button pressed, move the mouse pointer to the boundary of the circle. When the circle is highlighted, release the mouse.* You have now created a point that is on the boundary of the circle. You could alternatively have clicked the mouse over the circle boundary. However, that gives you a bit less control over whether you actually hit the circle. Your configuration should now look approximately like that in Figure 1.

Switch to move mode (p. 71) *to see how this new point behaves.* If you select it and drag the mouse, the point will never leave the circle; it attempts to remain as close to the mouse position as possible. When you move the center of the circle or change the circle's radius, the point also stays on the circle. The point furthermore keeps its relative angle to the center. *Choose the interactive add a segment* (p. 84) *mode by pressing the and draw a segment from the center to the point on the perimeter* (see Figure 2).

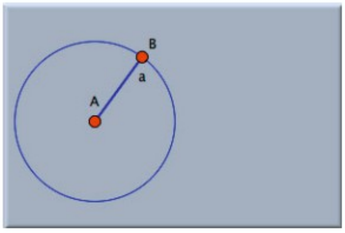

Figure 2: A bar

This line now behaves like a bar of fixed length, given by the radius of the circle. The only way to change its length is to change the radius of the circle.

3.2.2 Adding Two More Bars

We now want to add two more bars to complete the three-bar linkage that was presented in the introduction (p. 5). The linkage is a chain of three consecutive bars that is pinned down at each of its ends. *Switch to circle by radius* (p. 82) *mode again and add a second circle to the right of the first one. Do not let the two circles intersect.*

The radius of the second circle represents the length of the third bar in the row (we have not created the second one yet). Before finishing the construction we should analyze the situation. If point *C* in our construction is to be linked to point *B* of the construction by two bars of given lengths, then there is not much freedom for the point that is common to those two bars. In fact, there are exactly two possible positions for this point. They are the two intersections of two circles: one circle around *C*, already drawn, and another, yet to be drawn, circle around *B*.

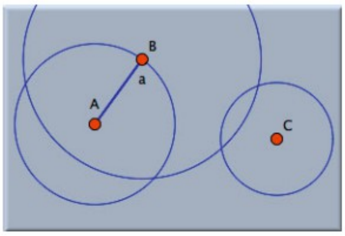

Figure 3: Two more circles

Therefore, as the next step, *draw a circle around point B in circle by radius* (p. 82) *mode. Make sure that its radius is large enough so the circle intersects the circle*

The *projective plane* consists of the points of the usual, Euclidean, plane together with one additional "infinite" point for every possible direction. The lines of the projective plane are the Euclidean lines together with one special "line at infinity." All infinite points lie on the line at infinity. The following nicely symmetric relations between points and lines hold:

- Every two distinct points have a unique connecting line (their *join*).
- Every two distinct lines have a unique point of intersection (their *meet*).

The first person who formalized these rules, around 1822, was *Victor Poncelet*, a student of Monge, who today may be considered the "father of projective geometry." In projective geometry there is no need to consider parallels as something special. They still have a point of intersection; it just happens to lie at infinity. For a readable introduction to projective geometry we refer to the books of H. S. M. Coxeter on that topic [4, 3].

2.3 Homogeneous Coordinates

On a computer we unfortunately do not have geometric objects as primitive data types. A point or a line has to be represented by numbers: the coordinates. Usually, a point in the plane is described by its (x,y) coordinates. A line may be given by the three parameters (a,b,c) of its defining equation $ax+by+c=0$. However, when we want to do projective geometry, this turns out to be impractical. Each pair of coordinates (x,y) represents a finite point, and there is no representation for the points at infinity. A solution to this problem gradually became clear in the first half of the nineteenth century. It started with Möbius's *barycentric coordinates*, via the refined setup of *homogeneous coordinates* given by Plücker, and finally led to Grassmann's setup of *multilinear algebra*.

The way out of the dilemma is as follows. For every point we use three instead of two coordinates, thereby introducing a third dimension. Consider the following scenario: The plane is embedded parallel to the *xy*-plane of three-space at a height of $z = 1$, so it does not pass through the origin (0, 0, 0).

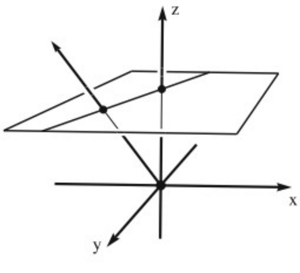

Embedding the plane in space

Every point (x, y) is represented by its three-dimensional coordinates $(x, y, 1)$. These coordinates are the *homogeneous coordinates* of the point. What happens with the rest of the points in the three-dimensional space? Almost all of them will be interpreted as points in the original plane: we identify all three-dimensional points that differ by a nonzero multiple. For instance, (4, 6, 2) and (2, 3, 1) describe the same point. In general, a point with spatial coordinates (x, y, z) is identified with the point $(x/z, y/z, 1)$ of the original plane. This process is called dehomogenization. In a way, every point of the original plane corresponds to the line spanned by this point and the origin in three-space.

However, there are points in three-space that do not correspond to points in the original plane. The points of the form $(x, y, 0)$ cannot be dehomogenized in the above way, since we would have to divide by zero. These points correspond precisely to the "points at infinity" of projective geometry. To see this, we study the behavior of a point that gradually moves to infinity in the original plane.

Assume that the moving point has coordinates (r, r). As r becomes larger and larger, this point gradually approaches a point at infinity in the 45° direction. Looking at its homogeneous coordinates, we see that they have the form $(r, r, 1) \sim (1, 1, 1/r)$. As r increases, the contribution of the first two coordinates dominates the last coordinate more and more. In the limiting case of r equals "infinity," the homogeneous coordinates are given by (1, 1, 0), an infinite point. You can also try to imagine the line through this point and the three-dimensional origin. As the point approaches infinity, this line becomes more and more horizontal, until, in the limiting case, it is contained entirely in the xy-plane.

A similar representation can be given for lines. For the line $ax + by + c = 0$ we take the parameters (a, b, c) as the homogeneous coordinates of the line. As in the case of points, we identify nonzero multiples of such coordinates, since they do not alter the solution space of the corresponding equation. There is one set of parameters, (0, 0, 1), that does not correspond to a finite line. This is the line at infinity. The vector (a, b, c) of a line is orthogonal to the plane spanned by the corresponding line and the origin of three-space. In particular, the vector (0, 0, 1) is orthogonal to the xy-plane, the "line at infinity."

In fact, the algebraic notion of homogeneous coordinates provides a complete symmetry between points and lines. Each point or line is represented by three homogeneous coordinates. A point (x, y, z) lies on a line (a, b, c) if and only if the scalar product $ax + by + cz$ is zero, which is simply rewriting the equation of the line. Geometrically, this means that the two corresponding vectors are orthogonal in three-space.

2.4 Complex Numbers

It is not only geometry that has been extended throughout the centuries. A similar process happened to numbers. Probably the first numerical concept considered by mankind was that of the positive integers: 1, 2, 3, ... From that beginning, it was reasonable to extend the number system gradually to encompass more powerful concepts. The *negative numbers*, the *rational numbers*, and the *real numbers* had to be invented to achieve a useful and self-contained system. The observation that there must be numbers that cannot be represented as quotients of two integers is of a geometric nature and dates back to approximately 600 B.C.E. It was observed by the Pythagoreans that there is no rational number measuring the length of a diagonal of a square with sides of length one. Through application of the Pythagorean theorem, this task is equivalent to finding a number x such that $x^2 = 2$. This discovery led to a deep crisis in the foundations of ancient geometry.

However, the story of extending the number system does not stop at that point. One of the extensions, with perhaps the most drastic consequences, was the introduction of *complex numbers*. It was *Geronimo Cardano* in his 1545 work *Ars Magna* who was first to explicitly propose such an extension of the real numbers. He was led to his conclusions by the study of solutions of polynomials of degree three. Based on the work of other contemporary mathematicians, he discovered that a complete systematic representation of these solutions can be given only with the help of hitherto unknown values involving square roots of negative numbers.

A complex number is a number of the form $a + ib$, where the symbol i satisfies the equation $i^2 = -1$, and a and b are real numbers. Clearly, the number i cannot be a real number, since the square of a real number can never be negative. The system of complex numbers is, like the real numbers, closed under the four basic arithmetic operations of addition, subtraction, multiplication, and division (excluding, of course, division by zero). In other words, the sum and the product of two complex numbers, as well as their difference and quotient, can again be written in the form $a + ib$ for suitable parameters a and b. However, unlike the real numbers, the system of complex numbers is also closed under the operation of finding solutions of polynomials. For instance, consider the polynomial

$$x^2 - 6x + 13 = 0$$

As you can easily check, it has no real solutions, but the complex numbers $3 + 2i$ and $3 - 2i$ do solve this equation. In fact, the following beautiful result is true: *Every polynomial equation with arbitrary real or complex coefficients has all its solutions again in the field of complex numbers.*

In a sense, the discovery of complex numbers is the starting point for most modern mathematics. Many mathematical theories find their broadest, most elegant, and most economical setting when they are formulated over the complex numbers. This happens also with geometry. Consider the situation of two circles. Depending on their position, they can have two, one, or no points of intersection.

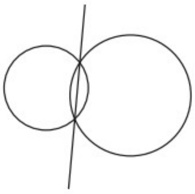

Circles can intersect...

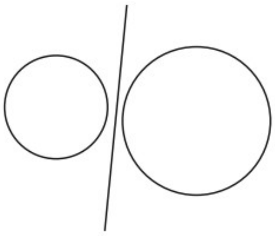

...or not.

Finding the coordinates of the points of intersection amounts to solving a quadratic equation. Over the real numbers, this equation might have no solution. In this case, the circles do not intersect. Over the complex numbers, however, a solution always exists. So in the case of visually nonintersecting circles, we say that the intersections still exist, but that they have complex coordinates, and therefore we cannot see them in the real plane.

Cinderella's mathematical kernel is implemented entirely over the complex numbers. So when intersections visually vanish, *Cinderella* does not have to deal with special cases, and it can still continue calculating. The solutions just have complex coordinates.

What happens if two complex points are connected by a line? In general, this line will also have complex coordinates. However, if the points are so-called *complex conjugates*, which means that they differ only by the sign of their complex part, then their join is again real. Since the solutions of a quadratic equation with real coefficients are always complex conjugates, it follows that the intersection points of two circles are always complex conjugates. This is the reason why the line joining them is a real line, no matter where in the plane the circles are located, even if the intersection points are complex and therefore invisible. *Cinderella* will correctly calculate and draw this line, independent of the position of the circles. It may take a while to get used to the fact that intermediate results can disappear while some constructions depending on them remain visible. However, this is exactly what you should expect. Consider the case in which the circles have the same radius. The line is then the *perpendicular bisector* of the segment joining the two midpoints. If you include the complex situations, you are not compelled to consider so many special cases.

Another example of a theorem in which intermediate steps disappear is the following statement about three circles. Construct the line joining the two points of intersection of each pair of circles. The three chords that you obtain in this way meet in a point, no matter whether the circles intersect or not.

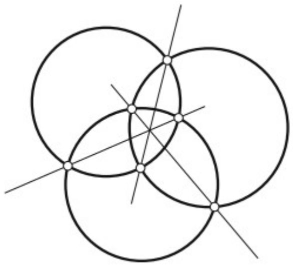

Intermediate points can be complex...

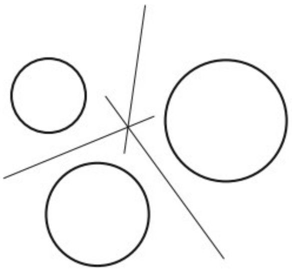

...but theorems remain true.

So in *Cinderella* each point and each line is represented by *complex homogeneous coordinates*. This means that altogether, any point or line has a six-dimensional(!) internal representation in the mathematical kernel. This may sound crazy, but it is the most natural thing to do.

2.5 Measurements and Complex Numbers

If we were satisfied with projective incidence theory, then the system presented so far would be fairly complete. However, we want to be able to measure distances and angles, too. Measurements are in a sense the most fundamental geometrical operations. Unfortunately, at first glance, projective geometry does not appear to be capable of measuring, since under perspective transformations distances can change. Actually, for a long time mathematicians considered projective geometry a "nice toy" for doing incidence geometry, but not appropriate for the real stuff: measuring.

History proved them wrong. With the right setup, projective geometry is *the* universal system for doing measurements. This system unifies and explains different kinds of measurements. For example, it explains the relationship between Euclidean and hyperbolic geometry. However, it took a long time to finally find the algebraic setting in which Projective Geometry develops its full power. The key objects are called "Cayley-Klein geometries" in modern terms. It is an elegant and consistent mathematical approach to measurements that combines projective geometry and complex numbers.

2.5.1 Euclidean and Non-Euclidean Geometry

One part of this development started with the discovery of non-Euclidean geometries. Our everyday geometry is, with relatively great accuracy, described by Euclid's five postulates. He used these postulates to axiomatize geometry. This happened almost 2000 years ago. The last postulate, the so-called parallel axiom, plays a special role in the development of geometry. A way to formulate it is this: *"Whenever there is a line l in the plane and a point P not on l, then there is exactly one line through P that does not meet l."*

Euclid was very cautious about using the parallel postulate. Large parts of Euclid's elaborations, for instance the complete theory of congruence of triangles, were done without the explicit need for this axiom. Today we are relatively sure that Euclid himself believed that this axiom was a consequence of the other four axioms. But he could not prove this. After Euclid, many other mathematicians tried to do so, and some of them even presented proofs. But all these proofs were incorrect.

In the sixteenth to eighteenth centuries, mathematicians also found many equivalent formulations for the parallel postulate. One of the most prominent formulations is "The interior angles in a triangle sum up to 180°." If this statement could be derived from Euclid's first four axioms, then this would prove the dependence of the parallel postulate.

Proving that an axiom is dependent can be done by assuming its contrary and drawing conclusions until a contradiction is shown. Many people tried this, among them *C.F. Gauss*, *J. Bolyai*, and *N. Lobachevski*. They drew conclusion after conclusion, but to their surprise, instead of arriving at a contradiction, they found themselves developing a beautiful geometric system: *hyperbolic geometry*. There Euclid's parallel postulate is modified in the following way: *"Whenever there is a line l in the plane and a point P not on l, then there will be more than one line through P that does not meet l."* A consequence of this assumption is that the angle sum in a triangle

data 17.04.2020

Use juntamente com o item
nº 138 (20200416),
letra F que se refere:
a) Fano Plane (Steiner System);
b) 7 lines of the plane
3 points - one line
c) Pappu's theorem.
Obs: Source deste book:
ARQ-IMPORTANT/
Cinderella.de/
the Cinderella.2 Manual.pdf
Pág. 42(57)
3.1 Pappu's theorem
(forma de construção).
Other: (same book)
Pág. 10(46)
4. Extension of Ptolemy's theorem

Jandaia

data 17 04 2020

Pág. 11(47)

5. Fermat's problem

↳ use too: extension of Ptolemy's theorem.

A, B, C
are any three points in a plane.
To find a point P such that PA + PB + PC shall be least.

Pág. 14(50)

7. Coaxal systems of circles

Look: Pag. 16(52)
the radical axis of C_1 and C_2 is

$$2x[\ \ldots \qquad = 0$$

looks

Pág. 21(57)

10. Problem of Apollonius

Given three circles C_1, C_2 and

17.04.2020

C3, to draw a circle touching all three.

Fig. 27.

D) Books_63_2020/
Differential Geometry and
Its Applications//
Differential Geometry and
its applications...pdf

Pag. 416(442)

8.5 Curvature

In Section 3.4 we developed for Gauss curvature which depended only on the metric.

Section. 3.4

Pag. 123(145)

3.4 A Formula for Gauss Curvature.

Obs.: Pag. 122(144)

Figure 3.5 A pseudosphere.

IMPORTANT:

The pseudosphere is actually determined by the requirement of constant curvature and certain initial

17.04.2020

conditions. Suppose we start off by requiring that a surface of revolution M have constant Gauss curvature K = -1.

E) Books_69_2020/
Differential Geometry and Its Applications/
Lokenath... Introduction to Hilbert Spaces... pdf.

F) Books_69_2020/
Nature and Numbers A Mathematical Photo Shooting/
www... Nature-and-Numbers.pdf

obs: Estudar com atenção!

G) Books_69_2020/
Topology Illustrated/
Saveliev Topology Illustrated...pdf

obs: Estudar em detalhes!!!

a) Pág. 452 (459)
4.2 The Euler-Poincaré Formula

b) Pág. 541 (548)
4. Metric Tensor
4.1 The need for measuring in calculus

" /012/ / /3 / ((K ? + "
0** -+ B 3* *"
' (, (( ) 5

since $\nabla^{\mathbb{R}^{n+1}} = \nabla + (\nabla^{\mathbb{R}^{n+1}})^N$ and, at $t = 0$, we are on M. Also, for $t = 0$, U and the ∂_i are orthogonal and the coordinates are normal, so

$$\begin{aligned}\frac{d}{dt}\left(\frac{\sqrt{g_t}}{\sqrt{g}}\right)\bigg|_{t=0} &= \sum_{i=1}^{k} -\langle U, \left(\nabla^{\mathbb{R}^{n+1}}_{\partial_i}\partial_i\right)^N\rangle \\ &= -\langle U, \sum_{i=1}^{k}\left(\nabla^{\mathbb{R}^{n+1}}_{\partial_i}\partial_i\right)^N\rangle \\ &= -\langle U, H\rangle\end{aligned}$$

where H is the mean curvature vector field of M in $\mathbb{R}^{n+1}$. After integration, we have the

Theorem 8.4.12. *The derivative of the area functional at zero is given by*

$$\frac{dA(t)}{dt}\bigg|_{t=0} = \int_M -\langle U, H\rangle\, dA$$

where H is the mean curvature vector field and U is the normal variational direction.

If M (i.e., $t = 0$) is a critical point for all variations, then we must have $H = 0$. This follows because any nonzero part of H would produce a nonzero $\langle U, H\rangle$ for some U and this then could be localized by a bump function to produce a nonzero integral, contradicting the assumption that M is a critical point of $A(t)$. Thus, we have the higher-dimensional version of Theorem 4.3.4,

Corollary 8.4.13. *If M is least area, then the mean curvature vector field is zero.*

8.5 Curvatures

In Section 3.4 we developed a formula for Gauss curvature which depended only on the metric. This led of course to Gauss's Theorem Egregium. If we look closely at how the formula was derived, we see that we used the fact that $\mathbf{x}_{uuv} - \mathbf{x}_{uvu} = 0$. Written in our "covariant" notation (and noting that the covariant derivative with respect to a parameter curve is simply partial differentiation), we have

$$\nabla^{\mathbb{R}^{n+1}}_{\mathbf{x}_v}\nabla^{\mathbb{R}^{n+1}}_{\mathbf{x}_u}\mathbf{x}_u - \nabla^{\mathbb{R}^{n+1}}_{\mathbf{x}_u}\nabla^{\mathbb{R}^{n+1}}_{\mathbf{x}_v}\mathbf{x}_u = 0.$$

Furthermore, the derivation also reveals that we made essential use of the basis descriptions of U_u and U_v, where U is the unit normal of the surface. In fact, it was these terms which provided the $(ln - m^2)/EG = K$ in the formula. Now, however, we must use M^k's covariant derivative ∇ without the benefit of a unit normal. Therefore, we should not expect the corresponding quantity

$$\nabla_{\partial_j}\nabla_{\partial_i}\partial_s - \nabla_{\partial_i}\nabla_{\partial_j}\partial_s$$

to be zero. Indeed, analogous to the derivation of the two-dimensional formula (but without the ingredients U_u and U_v), we may tentatively define a vector field $R(\partial_i, \partial_j)\partial_s = \sum_l R^l_{ijs}\partial_l$ by

$$R(\partial_i, \partial_j)\partial_s = \nabla_{\partial_j}\nabla_{\partial_i}\partial_s - \nabla_{\partial_i}\nabla_{\partial_j}\partial_s.$$

The object $R(\cdot, \cdot)$ will be called the Riemann curvature, *but the formula on the right-hand side does not yet describe it in general*. We want R to possess certain qualities. Namely, we want R to be bilinear with respect to smooth functions. By this we mean that $R(fX,Y)=fR(X,Y)$, $R(X+Z,Y)=R(X,Y)+R(Z,Y)$, $R(X,fY)=fR(X,Y)$ and $R(X,Y+Z)=R(X,Y)+R(X,Z)$ for any smooth function $f\colon M\to\mathbb{R}$. But, as the reader is asked to check in the following exercise, as yet, this is *not the case*! Moreover, we want $R(\cdot,\cdot)$ to be a linear map on vector fields as well.

Exercise 8.5.1. Use the properties of covariant derivatives to show that

$$R(\boldsymbol{\partial}_i,\boldsymbol{\partial}_j)f\boldsymbol{\partial}_s = fR(\boldsymbol{\partial}_i,\boldsymbol{\partial}_j)\boldsymbol{\partial}_s. \tag{1}$$

$$\begin{aligned} R(f\boldsymbol{\partial}_i,\boldsymbol{\partial}_j)\boldsymbol{\partial}_s &= \boldsymbol{\partial}_j f\nabla_{\boldsymbol{\partial}_i}\boldsymbol{\partial}_s + fR(\boldsymbol{\partial}_i,\boldsymbol{\partial}_j)\boldsymbol{\partial}_s \\ &= -\nabla_{[f\boldsymbol{\partial}_i,\boldsymbol{\partial}_j]}\boldsymbol{\partial}_s + fR(\boldsymbol{\partial}_i,\boldsymbol{\partial}_j)\boldsymbol{\partial}_s. \end{aligned} \tag{2}$$

where the last line follows from Proposition 8.3.2 (ii) and Exercise 8.3.4 (4). Hence, R is not linear in the first (or second) variable.

The last formula of the exercise gives a hint as to how to redefine R so as to achieve bilinearity. Namely, for tangent vector fields X, Y and Z, we have

Definition 8.5.2. The *Riemann curvature* is defined to be

$$R(X,Y)Z = \nabla_{[X,Y]}Z + \nabla_Y\nabla_X Z - \nabla_X\nabla_Y Z.$$

Exercise 8.5.3. Show that

$$\begin{aligned} R(\boldsymbol{\partial}_i,\boldsymbol{\partial}_j)\boldsymbol{\partial}_s &= \sum_l R^l_{ijs}\boldsymbol{\partial}_l \\ &= \sum_l \left(\frac{\partial\Gamma^l_{is}}{\partial u^j} - \frac{\partial\Gamma^l_{js}}{\partial u^i} + \sum_m \Gamma^m_{is}\Gamma^l_{jm} - \Gamma^m_{js}\Gamma^l_{im}\right)\boldsymbol{\partial}_l. \end{aligned}$$

Since, as we showed previously, all Christoffel symbols vanish in $\mathbb{R}^{n+1}$, the expression above says that the Riemann curvature of $\mathbb{R}^{n+1}$ is zero everywhere. We say that $\mathbb{R}^{n+1}$ is *flat*. Further, show that

$$\begin{aligned} R_{ijsr} &\overset{\text{def}}{=} \sum_l R^l_{ijs}g_{lr} \\ &= \sum_l \left(\frac{\partial\Gamma^l_{is}}{\partial u^j} - \frac{\partial\Gamma^l_{js}}{\partial u^i} + \sum_m \Gamma^m_{is}\Gamma^l_{jm} - \Gamma^m_{js}\Gamma^l_{im}\right)g_{lr} \end{aligned}$$

and calculate R_{1212} for a two-dimensional patch $\mathbf{x}(u^1,u^2)=\mathbf{x}(u,v)$ with $g_{11}=E, g_{12}=g_{21}=0$ and $g_{22}=G$. Show that

$$R_{1212} = EG\left[\frac{E_uG_u}{4E^2G} - \frac{1}{E}\left(\frac{E_v}{2G}\right)_v - \frac{E_vG_v}{4EG^2} + \frac{E_vE_v}{4E^2G} - \frac{1}{E}\left(\frac{G_u}{2G}\right)_u - \frac{G_uG_u}{4EG^2}\right]$$

and compare this with the formula above Exercise 3.4.5 to see that

$$K = \frac{R_{1212}}{g}$$

where K is Gauss curvature and $g = \det(g_{ij}) = EG$. Thus the Riemann curvature generalizes the Gauss curvature.

Exercise 8.5.4. Suppose $M^n \subset \mathbb{R}^{n+1}$ is a hypersurface with unit normal U and shape operator S. Show that, for tangent vector fields X, Y and Z,

$$R(X, Y)Z = \langle S(X), Z\rangle\, S(Y) - \langle S(Y), Z\rangle\, S(X).$$

Hints: (1) write

$$R(X, Y) = \left(\nabla^{\mathbb{R}^{n+1}}_{[X,Y]} - \mathcal{S}_{[X,Y]}\right) + \left(\nabla^{\mathbb{R}^{n+1}}_{Y} - \mathcal{S}_Y\right)\left(\nabla^{\mathbb{R}^{n+1}}_{X} - \mathcal{S}_X\right) - \left(\nabla^{\mathbb{R}^{n+1}}_{X} - \mathcal{S}_X\right)\left(\nabla^{\mathbb{R}^{n+1}}_{Y} - \mathcal{S}_Y\right)$$

where $\mathcal{S}_Y(Z) = \langle S(Y), Z\rangle$. (2) Use the facts that $\nabla^{\mathbb{R}^{n+1}}_{[X,Y]} + \nabla^{\mathbb{R}^{n+1}}_{Y}\nabla^{\mathbb{R}^{n+1}}_{X} - \nabla^{\mathbb{R}^{n+1}}_{X}\nabla^{\mathbb{R}^{n+1}}_{Y} = 0$ (since $\mathbb{R}^{n+1}$ is flat) and $\langle \nabla^{\mathbb{R}^{n+1}}_{Y} U, U\rangle = 1/2\, Y\langle U, U\rangle = 0$ (since $\langle U, U\rangle = 1$ is constant).

It is clear that Riemann curvature is a complicated object to deal with. We can gain some intuition about it however by considering vector fields in pairs. First, let us introduce the notation

$$R(X, Y, Z, W) \overset{\text{def}}{=} \langle R(X, Y)Z, W\rangle.$$

Then we define the *sectional curvature* of the two-dimensional subspace of T_pM spanned by $X(p)$ and $Y(p)$ to be

$$K(X, Y) = \frac{R(X, Y, X, Y)}{\langle X, X\rangle\langle Y, Y\rangle - \langle X, Y\rangle^2}.$$

As the notation suggests, the sectional curvature is a Gauss curvature of a two-dimensional submanifold of M through p with tangent plane $\langle X, Y\rangle$. This generalizes Exercise 8.5.3, where $K = R_{1212}/g = R(\partial_1, \partial_2, \partial_1, \partial_2)/g$. Although we will not prove it, it is a fact that *the sectional curvature actually determines the Riemann curvature*. Of course the reverse is true by definition. Also note that if X and Y are orthonormal, then $K(X, Y) = R(X, Y, X, Y)$.

Example 8.5.5 (Sectional Curvature of the Sphere S^n of Radius R). In Example 8.3.8 we saw that the shape operator of S^n (of radius R) is given by $S(X) = -X/R$, where R is the radius of the sphere. From the formula of Exercise 8.5.4, we obtain

$$\begin{aligned} R(X, Y)Z &= \langle S(X), Z\rangle\, S(Y) - \langle S(Y), Z\rangle\, S(X) \\ &= \frac{1}{R^2}(\langle X, Z\rangle Y - \langle Y, Z\rangle X). \end{aligned}$$

The sectional curvature has an even more perspicacious appearance — especially when we take X and Y to be orthonormal. For then we have

$$\begin{aligned} K(X, Y) &= R(X, Y, X, Y) \\ &= \frac{1}{R^2}\langle \langle X, X\rangle Y - \langle Y, X\rangle X, Y\rangle \\ &= \frac{1}{R^2} \end{aligned}$$

since $\langle Y, X\rangle = 0 = \langle X, Y\rangle$ and $\langle X, X\rangle = 1 = \langle Y, Y\rangle$. Thus $K = 1/R^2$ is the *constant* sectional curvature of S^n; that is, completely analogous to the constant Gauss curvature of S^2.

Exercise 8.5.6. Show that the Riemann curvature obeys the following symmetry relations:

(1) $R(X, Y, Z, W) = -R(Y, X, Z, W)$;

(2) $R(X, Y, Z, W) = -R(X, Y, W, Z)$;

(3) $R(X, Y, Z, W) = R(Z, W, X, Y)$.

Hints: (1) $\langle \nabla_Y \nabla_X V, V\rangle = Y\langle \nabla_X V, V\rangle - \langle \nabla_X V, \nabla_Y V\rangle$ and (2) $\langle \nabla_{[X,Y]} V, V\rangle = 1/2\,[X, Y]\langle V, V\rangle$, both by (iv) of Proposition 8.3.2. For part (3), use the first Bianchi identity below.

Exercise 8.5.7. Prove the first Bianchi identity:

$$R(X, Y, Z, W) + R(Y, Z, X, W) + R(Z, X, Y, W) = 0.$$

Hints: W is superfluous, so drop it from the notation. Write out the definitions of the terms and group them to make use of the identity $\nabla_X Y - \nabla_Y X = [X, Y]$ etc. Then group the terms according to the respective Lie brackets which arise and apply the identity again to end up with only brackets in the formula. Apply the Jacobi identity (Exercise 8.3.4 (3)).

There is a *second Bianchi identity* which will prove very convenient a bit later. This identity involves covariant derivatives of the Riemann curvature, so we must understand how these are defined. The covariant derivative of R is given by

$$\begin{aligned}\nabla R(X, Y, Z, W, V) &= \nabla_V R(X, Y, Z, W)\\ &= V[R(X, Y, Z, W)] - R(\nabla_V X, Y, Z, W) - R(X, \nabla_V Y, Z, W)\\ &\quad - R(X, Y, \nabla_V Z, W) - R(X, Y, Z, \nabla_V W).\end{aligned}$$

Proposition 8.5.8 (The second Bianchi identity).

$$\nabla R(X, Y, Z, W, V) + \nabla R(X, Y, W, V, Z) + \nabla R(X, Y, V, Z, W) = 0.$$

Proof. Choose normal coordinates (see Theorem 8.4.10) about a point $p \in M$ and test the identity at this arbitrary point. Let X, Y, Z, W and V be in the orthonormal coordinate basis with $\nabla_X Y(p) = 0$ etc. Also note that brackets of these vector fields vanish as well by $[X, Y] = \nabla_X Y - \nabla_Y X = 0 - 0 = 0$. Then

$$\begin{aligned}\nabla R(X, Y, Z, W, V) &= \nabla_V R(X, Y, Z, W)\\ &= \nabla_V R(Z, W, X, Y)\\ &= V[R(Z, W, X, Y)] \qquad \text{since } \nabla_V(\cdot) = 0\\ &= \langle \nabla_V R(Z, W)X, Y\rangle\end{aligned}$$

by Proposition 8.3.2 (iv) and $\nabla_V Y = 0$. Similarly, $\nabla R(X, Y, W, V, Z) = \langle \nabla_Z R(W, V)X, Y\rangle$ and $\nabla R(X, Y, V, Z, W) = \langle \nabla_W R(V, Z)X, Y\rangle$. Hence, X and Y are superfluous, so we remove them from the notation. Then

$$\begin{aligned}\nabla_V R(Z, W) + \nabla_Z R(W, V) + \nabla_W R(V, Z) = {}&(\nabla_V\nabla_W - \nabla_W\nabla_V)\nabla_Z\\ &+ (\nabla_Z\nabla_V - \nabla_V\nabla_Z)\nabla_W + (\nabla_W\nabla_Z - \nabla_Z\nabla_W)\nabla_V\end{aligned}$$

and, evaluating on X, we get $(\nabla_V R(Z, W) + \nabla_Z R(W, V) + \nabla_W R(V, Z))(X) = R(W, V, \nabla_Z X) + R(V, Z, \nabla_W X) + R(Z, W, \nabla_V X)$ since brackets are zero. But then, since $\nabla_Z X = 0$ etc., we have

$$(\nabla_V R(Z, W) + \nabla_Z R(W, V) + \nabla_W R(V, Z))(X) = 0.$$

Hence, the result is proved. □

While the Bianchi identities may seem rather esoteric, they *do* simplify various calculations considerably as we will see below.

Higher-dimensional differential geometry offers us both the challenge of understanding nonvisualizable geometric phenomena and the opportunity to create new tools with which to study such phenomena. Some of these new tools are new types of curvatures which, in two dimensions, become the Gauss curvature. We have already seen that sectional curvature can tell us information somewhat obscured by the Riemann curvature. Sectional curvature, however, cannot be the final answer since, as we mentioned previously, it determines Riemann curvature. Instead, there is a general method called *contraction* which is available to us in higher dimensions and which allows us to isolate more tractable portions of the Riemann curvature. In order to do this, we will generalize the notion of a *frame field* which was used in Chapter 6. If $\mathcal{E}_1, \dots, \mathcal{E}_k$ are vector fields defined on a neighborhood of a point $p \in M^k$ with

$$\langle \mathcal{E}_i, \mathcal{E}_j\rangle = \delta_i^j$$

at every point in the neighborhood, then the collection $\{\mathcal{E}_1, \dots, \mathcal{E}_k\}$ is called a *frame* about p. One way to obtain a frame is to choose normal coordinates around p (Theorem 8.4.10) and then parallel translate the corresponding vector fields along geodesics passing through p. By Exercise 8.4.8, parallel translation is an isometry, so the orthonormality of the frame is preserved as it is translated to other points in the normal coordinate patch. The one bothersome thing about using a frame is that we lose many of the coordinate formulas we obtained earlier. Nevertheless, we shall see that the advantages of a frame outweigh the disadvantages, so we assume that we have a frame about p in what follows.

Definition 8.5.9. The *Ricci curvature* is defined to be

$$\operatorname{Ric}(X, Y) = \sum_{i=1}^{k} \langle R(X, \mathcal{E}_i)Y, \mathcal{E}_i\rangle$$

where X and Y are tangent vector fields on M^k. The *scalar curvature* is defined to be

$$\begin{aligned}\kappa &= \sum_{j=1}^{k} \operatorname{Ric}(\mathcal{E}_j, \mathcal{E}_j)\\ &= \sum_{i,j=1}^{k} \langle R(\mathcal{E}_j, \mathcal{E}_i)\mathcal{E}_j, \mathcal{E}_i\rangle.\end{aligned}$$

The first thing we should note is that these definitions are independent of the frame field we choose. The reason is this. Any two frames are related at a point by an orthogonal matrix; that is, a matrix A with $A^t = A^{-1}$. This is equivalent to saying that the rows of A form an orthonormal set of vectors. This, together with the symmetry relations of Exercise 8.5.6 suffice to prove the invariance of the definition. Rather than give a proof of this in general, we will simply concentrate on the $k = 2$ case. So, suppose that $\{\mathcal{F}_j\}$ is another frame field which is related to $\{\mathcal{E}_i\}$ by a 2×2 orthogonal matrix $A = (a_{st})$ as follows:

$$\mathcal{E}_i = a_{1i}\mathcal{F}_1 + a_{2i}\mathcal{F}_2.$$

Now we can just compute using the fact that the rows of A are orthonormal. That is, $a_{11}a_{21} + a_{12}a_{22} = 0$, $a_{11}^2 + a_{12}^2 = 1$ and $a_{21}^2 + a_{22}^2 = 1$.

$$\begin{aligned}\operatorname{Ric}(X, Y) &= \langle R(X, \mathcal{E}_1)Y, \mathcal{E}_1\rangle + \langle R(X, \mathcal{E}_2)Y, \mathcal{E}_2\rangle\\ &= \langle R(X, a_{11}\mathcal{F}_1 + a_{21}\mathcal{F}_2)Y, a_{11}\mathcal{F}_1 + a_{21}\mathcal{F}_2\rangle\\ &\qquad + \langle R(X, a_{12}\mathcal{F}_1 + a_{22}\mathcal{F}_2)Y, a_{12}\mathcal{F}_1 + a_{22}\mathcal{F}_2\rangle\\ &= a_{11}^2\langle R(X, \mathcal{F}_1)Y, \mathcal{F}_1\rangle + a_{11}a_{21}\langle R(X, \mathcal{F}_1)Y, \mathcal{F}_2\rangle\\ &\quad + a_{21}a_{11}\langle R(X, \mathcal{F}_2)Y, \mathcal{F}_1\rangle + a_{21}^2\langle R(X, \mathcal{F}_2)Y, \mathcal{F}_2\rangle\\ &\quad + a_{12}^2\langle R(X, \mathcal{F}_1)Y, \mathcal{F}_1\rangle + a_{12}a_{22}\langle R(X, \mathcal{F}_1)Y, \mathcal{F}_2\rangle\\ &\quad + a_{22}a_{12}\langle R(X, \mathcal{F}_2)Y, \mathcal{F}_1\rangle + a_{22}^2\langle R(X, \mathcal{F}_2)Y, \mathcal{F}_2\rangle\\ &= a_{11}^2\langle R(X, \mathcal{F}_1)Y, \mathcal{F}_1\rangle + 2a_{11}a_{21}\langle R(X, \mathcal{F}_1)Y, \mathcal{F}_2\rangle + a_{21}^2\langle R(X, \mathcal{F}_2)Y, \mathcal{F}_2\rangle\\ &\quad + a_{12}^2\langle R(X, \mathcal{F}_1)Y, \mathcal{F}_1\rangle + 2a_{12}a_{22}\langle R(X, \mathcal{F}_1)Y, \mathcal{F}_2\rangle + a_{22}^2\langle R(X, \mathcal{F}_2)Y, \mathcal{F}_2\rangle\end{aligned}$$

using the symmetry relations of Exercise 8.5.6. Then

$$\begin{aligned}\operatorname{Ric}(X, Y) &= (a_{11}^2 + a_{12}^2)\langle R(X, \mathcal{F}_1)Y, \mathcal{F}_1\rangle\\ &\qquad + 2(a_{11}a_{21} + a_{12}a_{22})\langle R(X, \mathcal{F}_1)Y, \mathcal{F}_2\rangle + (a_{21}^2 + a_{22}^2)\langle R(X, \mathcal{F}_2)Y, \mathcal{F}_2\rangle\\ &= \langle R(X, \mathcal{F}_1)Y, \mathcal{F}_1\rangle + \langle R(X, \mathcal{F}_2)Y, \mathcal{F}_2\rangle\end{aligned}$$

by the equalities $a_{11}a_{21} + a_{12}a_{22} = 0$, $a_{11}^2 + a_{12}^2 = 1$ and $a_{21}^2 + a_{22}^2 = 1$. Hence, the Ricci curvature is well defined.

The second thing to note is that both Ricci and scalar curvatures come about through a "trace-like" process. Of course, to take the trace of a matrix $A = (a_{ij})$, we form

$$\operatorname{tr} A = \sum_i a_{ii} = \sum_i \langle A(\mathcal{E}_i), \mathcal{E}_i\rangle$$

for an orthonormal basis $\{\mathcal{E}_1, \dots, \mathcal{E}_k\}$. Here, for Ric, we take two slots of the Riemann curvature and sum over the same elements of the frame. Similarly, and even more closely analogous to trace, to form κ, we sum over the only available slots in Ric. This process of summing over two slots with the same frame elements is called *contraction* and is sometimes denoted $\mathcal{C}$. Thus, $\mathcal{C}_{24}R = \mathrm{Ric}$ and $\mathcal{C}_{12}\,\mathrm{Ric} = \kappa$ since, in the first case, we sum over the second and fourth slots and, in the second case, we sum over the first and second slots. We note here that some authors order the subscripts on R_{ijlr} differently, so contraction subscripts may have to be adjusted to correspond to ours. As the reader might expect, there is a more formal and precise notion of "contraction", but we shall have no need of it here.

Example 8.5.10 (Contraction of the Metric).
(i) For a k-manifold M^k, let us compute the (only available) contraction of the metric $\langle \cdot, \cdot \rangle$. For a frame as above, since $\langle \mathcal{E}_i, \mathcal{E}_i \rangle = \delta_i^j$,

$$\begin{aligned} \mathcal{C}(\langle \cdot, \cdot \rangle) &= \sum_{i=1}^{k} \langle \mathcal{E}_i, \mathcal{E}_i \rangle \\ &= k \\ &= \dim M^k. \end{aligned}$$

(ii) Contraction can also be done by coordinates. Without going into details, for coordinates $u^1, \dots, u^k$ and basis tangent vector fields $\boldsymbol{\partial}_1, \dots, \boldsymbol{\partial}_k$ with $\langle R(\boldsymbol{\partial}_i, \boldsymbol{\partial}_j)\boldsymbol{\partial}_s, \boldsymbol{\partial}_r \rangle = R_{ijsr}$, let

$$R_{is} = \sum_{j,r} R_{ijsr} g^{jr}.$$

These are then the components of Ricci curvature. Thus, by Exercise 8.5.3,

$$\begin{aligned} R_{is} &= \sum_{j,r} \sum_{l} \left(\frac{\partial \Gamma_{is}^{l}}{\partial u^j} - \frac{\partial \Gamma_{js}^{l}}{\partial u^i} + \sum_m \Gamma_{is}^{m} \Gamma_{jm}^{l} - \Gamma_{js}^{m} \Gamma_{im}^{l} \right) g_{lr} g^{jr} \\ &= \sum_{j,l} \left(\frac{\partial \Gamma_{is}^{l}}{\partial u^j} - \frac{\partial \Gamma_{js}^{l}}{\partial u^i} + \sum_m \Gamma_{is}^{m} \Gamma_{jm}^{l} - \Gamma_{js}^{m} \Gamma_{im}^{l} \right) \sum_r g_{lr} g^{jr} \\ &= \sum_{l} \left(\frac{\partial \Gamma_{is}^{l}}{\partial u^l} - \frac{\partial \Gamma_{ls}^{l}}{\partial u^i} + \sum_m \Gamma_{is}^{m} \Gamma_{lm}^{l} - \Gamma_{ls}^{m} \Gamma_{im}^{l} \right) \end{aligned}$$

since $\sum_r g_{lr} g^{jr} = \delta_j^l$.

In order to see the power of these curvatures, we offer the following theorems without proof. For a proof of the the first, see [GHL90] for example.

Theorem 8.5.11 (Myers's Theorem). *Let M^k be a complete k-manifold and suppose that the Ricci curvature on all of M^k is strictly bounded away from zero; that is,*

$$\mathrm{Ric}(X, X) \geq \frac{k-1}{r^2} > 0.$$

Then M^k is compact and the diameter of M^k is less than or equal to πr.

Compare this result with Bonnet's result, Theorem 6.8.16. Scalar curvature also has the power to constrain the type of a manifold. Compare the following result (due to Ros [Ros87]) to Alexandrov's result, Theorem 4.4.6.

Theorem 8.5.12. *A compact hypersurface $M^n \subset \mathbb{R}^{n+1}$ of constant scalar curvature is a sphere S^n with metric induced by that of $\mathbb{R}^{n+1}$.*

Exercise 8.5.13. For a surface defined by an orthogonal patch in $\mathbb{R}^3$, show that Ricci and scalar curvatures are given by

$$\operatorname{Ric}(X, Y) = K\,\langle X, Y\rangle \qquad \text{and} \qquad \kappa = 2K$$

where K is Gauss curvature. Hint: use Exercise 8.5.3. The orthogonal patch condition is, in fact, unnecessary, but Exercise 8.5.3 assumed it for simplicity.

Exercise 8.5.14. Show that the following formula holds relating the covariant derivative, the metric and the bracket.

$$2\langle X, \nabla_Z Y\rangle = Z\,\langle X, Y\rangle + Y\,\langle X, Z\rangle - X\,\langle Y, Z\rangle + \langle Y, [X, Z]\rangle + \langle Z, [X, Y]\rangle - \langle X, [Y, Z]\rangle.$$

Exercise 8.5.15. Here is a first small step toward getting away from the induced metric of $\mathbb{R}^{n+1}$. A metric $\widetilde{\langle \cdot, \cdot\rangle}$ is a *multiple* of another metric $\langle \cdot, \cdot\rangle$ if

$$\widetilde{\langle X, Y\rangle} = \lambda\,\langle X, Y\rangle$$

for all X and Y, where λ is a constant. If λ is allowed to be a smooth function on M, then the metrics are *conformal*. Denoting the covariant derivative and curvatures of the associated metric $\widetilde{\langle \cdot, \cdot\rangle}$ by $\widetilde{\nabla}$, $\tilde{R}$, $\widetilde{\operatorname{Ric}}$, $\tilde{\kappa}$, show that the following properties hold for a "multiple" metric.

(1) $\widetilde{\nabla}_X Y = \nabla_X Y$. (Use the formula from Exercise 8.5.14.)

(2) $\tilde{R}(X, Y, Z, W) = \lambda\, R(X, Y, Z, W)$. (Use (1).)

(3) $\tilde{K}(X, Y) = \dfrac{1}{\lambda}\, K(X, Y)$. (Use (2).)

(4) $\widetilde{\operatorname{Ric}}(X, Y) = \operatorname{Ric}(X, Y)$. (Use (2).)

(5) $\tilde{\kappa} = \dfrac{1}{\lambda}\,\kappa$. (Use (4).)

Exercise 8.5.16. Let $P^k = \{(u^1, \ldots, u^k) \in \mathbb{R}^k \mid u^k > 0\}$ be the "upper half space" analogous to the two-dimensional Poincaré plane. Define an analogous metric on P^k by

$$\langle \mathbf{v}, \mathbf{w}\rangle = \frac{\mathbf{v}\cdot\mathbf{w}}{(u^k)^2}.$$

In our present language, this means that

$$g_{ij} = \langle \mathbf{e}_i, \mathbf{e}_j\rangle = \frac{\mathbf{e}_i\cdot\mathbf{e}_j}{(u^k)^2} = \frac{\delta^i_j}{(u^k)^2}$$

where the $\mathbf{e}_i$ are the standard orthonormal $\mathbb{R}^k$ basis vectors. Hence, the metric is diagonal. The sectional curvature of P^k is known to be constant at -1. The following steps lead to a partial understanding of this result. For further and complete details, including formulas for general conformal metrics, see [dC92, Chapter 8.3].

(1) Show that the Christoffel symbols are

$$\begin{aligned}\Gamma^r_{ij} &= \frac{1}{2}\sum_l g^{lr}\left(\partial_i g_{jl} + \partial_j g_{li} - \partial_l g_{ij}\right)\\ &= \frac{1}{2} g^{rr}\left(\partial_i g_{jr} + \partial_j g_{ri} - \partial_r g_{ij}\right)\end{aligned}$$

and that the only non-zero ones are

$$\Gamma^j_{kj} = -\frac{1}{u^k} \quad \text{for } j \neq k, \qquad \Gamma^k_{ii} = \frac{1}{u^k} \quad \text{for } i \neq k, \qquad \Gamma^k_{kk} = -\frac{1}{u^k}.$$

(2) Now show that the components of Riemann curvature relevant to sectional curvature are

$$\begin{aligned}R_{ijij} &= \sum_l \left(\frac{\partial \Gamma^l_{ii}}{\partial u^j} - \frac{\partial \Gamma^l_{ji}}{\partial u^i} + \sum_m \Gamma^m_{ii}\Gamma^l_{jm} - \Gamma^m_{ji}\Gamma^l_{im}\right) g_{lj}\\ &= \left(\frac{\partial \Gamma^j_{ii}}{\partial u^j} - \frac{\partial \Gamma^j_{ji}}{\partial u^i} + \sum_m \Gamma^m_{ii}\Gamma^j_{jm} - \Gamma^m_{ji}\Gamma^j_{im}\right)\left(\frac{1}{(u^k)^2}\right)\end{aligned}$$

and the only non-zero components are $R_{kjkj} = R_{ikik} = R_{ijij} = -1/(u^k)^4$. Here we assume $k \neq j$ in R_{kjkj}, $i \neq k$ in R_{ikik} and $i \neq j$, $i \neq k$ and $j \neq k$ in R_{ijij}.

(3) Using the calculation of (2), show that the components of sectional curvature $K_{ij} = R_{ijij}/(g_{ii} \cdot g_{jj})$ for the cases above are constantly equal to -1. This exercise indicates how the conformal metrics of Chapter 5 may be generalized to higher dimensions.

Exercise* 8.5.17. A manifold M^k is said to have nonnegative sectional curvature if $K(X, Y) \geq 0$ for all X and Y. The Ricci curvature is said to be nonnegative when $\text{Ric}(X, X) \geq 0$ for all X. Show that nonnegative sectional curvature on M implies nonnegative Ricci curvature on M and, in turn, this implies nonnegative scalar curvature on M. Hint: First prove the formula

$$\text{Ric}(X, X) = \langle X, X\rangle \sum_{j=1}^{k-1} K(X, \mathcal{E}_j) = \langle X, X\rangle \sum_{j=1}^{k-1} K(\mathcal{E}_k, \mathcal{E}_j)$$

for a frame $\{\mathcal{E}_j\}$ spanning $X^\perp$, the space of vector fields orthogonal to X, and $\mathcal{E}_k = X/|X|$.

Exercise 8.5.18. Here is an exercise which at once goes beyond our definitions and tests them. A *Lie Group* G is a smooth manifold with a smooth associative multiplication $G \times G \to G$ having an identity e and inverses. Take G to be compact and connected. For example, the special orthogonal group $SO(n)$ is a compact connected Lie group. We may define certain vector fields on G by fixing a vector $X(e) \in T_eG$ and then taking $X(g) = g_*(X(e)) \in T_gG$. Here, g_* is the map induced on tangent vectors by the multiplication $(g, h) \mapsto g \cdot h$ for all $h \in G$. These types of vector fields are called *left invariant vector fields*. Lie groups are quite symmetric in the sense that the covariant derivative (which may be defined without reference to an ambient Euclidean space, but which has all the properties of Proposition 8.3.2) and curvature at the identity determine the covariant derivatives and curvatures at all other points of G. Furthermore, it is known that

$\nabla_X X = 0$ for X any left invariant vector field. Thus, curves whose tangent vectors belong to left invariant vector fields are geodesics. Also, for G as above, the following identity is known to hold: $\langle [X, Y], Z\rangle = -\langle Y, [X, Z]\rangle$ for left invariant vector fields X, Y and Z. Here, $\langle \cdot, \cdot \rangle$ is a metric which is first defined on T_eG and then transported around G by the Lie group multiplication.

Now, for a compact, connected Lie group G with left invariant vector fields X, Y and Z, show that

(1) $\nabla_X Y = \frac{1}{2}[X, Y]$. (Consider $\nabla_{X-Y}(X - Y)$ and use Proposition 8.3.2 (v).)

(2) $R(X, Y)Z = \frac{1}{4}[[X, Y], Z]$.
(Start with the definition of R, then use (1) to express R completely in terms of brackets. Finally, use the Jacobi identity for the bracket (Exercise 8.3.4 (3)).)

(3) If X and Y are orthonormal, then $K(X, Y) = \frac{1}{4}\left|[X, Y]\right|^2$.
Thus, sectional curvature, and hence, Ricci and scalar curvatures, are nonnegative for a compact, connected Lie group. (Use the definition of K, (1), Proposition 8.3.2 (iv) and the identity $\langle [X, Y], Z\rangle = -\langle Y, [X, Z]\rangle$ with $Z = Y$.)

The *Killing form* of a Lie group G is defined to be $b(X, Y) = \operatorname{tr}(\mathrm{ad}_X\, \mathrm{ad}_Y)$, where $\mathrm{ad}_Z(W) = [Z, W]$ is a linear transformation of the vector space of left invariant vector fields. It can be shown that b is symmetric, bilinear and invariant under all automorphisms of G. A compact Lie group with nondegenerate Killing form is said to be *semisimple*. The form b is nondegenerate if $b(V, W) = 0$ for all W implies $V = 0$. In this case b is a negative definite form (i.e., $b(V, V) < 0$ for all $V \neq 0$) and $-b$ is actually the (bi-invariant) metric $\langle \cdot, \cdot \rangle$ on G. That is, $\langle X, Y\rangle = -b(X, Y)$. For instance, $SO(n)$ is semisimple. Using (2) above and the definition of ad_Z, show that a semisimple Lie group with metric b has Ricci curvature

$$\mathrm{Ric}(X, Y) = -\frac{1}{4}\, b(X, Y).$$

To end this section, we wish to make several calculations using the definitions above to try to understand a tiny bit of modern physics. Einstein's general theory of relativity is based on a principle that energy and momentum distort the geometry of spacetime and, as a consequence, are responsible for various physical phenomena such as light "bending" around the Sun (along a spacetime geodesic) and the precession of the perihelion of Mercury. Although this principle is simple to state in our naive discussion, the precise form of the relationship is far from clear. If we are to believe in the relationship at all, then perhaps the simplest such formula is

$$G = c\,T$$

where G is some type of curvature (called the *Einstein curvature*) and T is a quantity (called the *stress-energy tensor*) which depends on the amount of energy-momentum in a particular region of spacetime. In particular, in a vacuum we have $T = 0$ — but interesting physics (such as the bending of light) can still occur. If G were Riemann curvature R, then the formula $G = c\,T$ together with the assumption $T = 0$ would imply the vanishing of Riemann curvature. But, as we have said earlier, $R = 0$ implies that a manifold is flat. Hence, spacetime would be forced to be flat and this would contradict the interesting geometrically induced physics observed. So, G cannot then be the Riemann curvature. The next logical choice to try is the Ricci curvature, so we take $G = \mathrm{Ric}$ provisionally. But we must ask, what else do we desire in a physical quantity? Certainly, the law of conservation of energy-momentum must be obeyed, so we obtain the vanishing of some sort of divergence of T. This means that we should also require G to be divergence free as well. Let us consider this now.

If $V = (V^1, V^2, V^3)$ is a vector field in $\mathbb{R}^3$ with coordinates x^1, x^2 and x^3, then we know

$$\operatorname{div} V = \frac{\partial V^1}{\partial x^1} + \frac{\partial V^2}{\partial x^2} + \frac{\partial V^3}{\partial x^3} = \sum_i \langle \nabla^{\mathbb{R}^3}_{\mathbf{e}_i} V, \mathbf{e}_i \rangle$$

where the $\mathbf{e}_i$ are the usual Euclidean orthonormal basis vectors.

Exercise 8.5.19. Verify the last equality

$$\frac{\partial V^1}{\partial x^1} + \frac{\partial V^2}{\partial x^2} + \frac{\partial V^3}{\partial x^3} = \sum_i \langle \nabla^{\mathbb{R}^3}_{\mathbf{e}_i} V, \mathbf{e}_i \rangle.$$

In the same way, on a manifold M^k, we can define the divergence of vector fields and bilinear quantities with two slots (i.e., 2-tensors) such as the metric $\langle \cdot, \cdot \rangle$ and Ric. Let $A(\cdot, \cdot)$ be such a quantity and define the *divergence of* A, $\operatorname{div} A(\cdot)$, to be a linear quantity with one slot (i.e., a 1-form; see Section 8.6),

$$\begin{aligned}\operatorname{div} A(X) &\overset{\text{def}}{=} \sum_i \nabla_{\mathcal{E}_i} A(\mathcal{E}_i, X)\\ &= \sum_i \mathcal{E}_i A(\mathcal{E}_i, X) - A(\nabla_{\mathcal{E}_i}\mathcal{E}_i, X) - A(\mathcal{E}_i, \nabla_{\mathcal{E}_i} X)\end{aligned}$$

where $\{\mathcal{E}_i\}_{i=1}^k$ is a frame and we use a definition for ∇A similar to that for ∇R in the second Bianchi identity Proposition 8.5.8. There are several things to note here. First, the divergence produces a "tensor" which is completely determined at each point by the values of its constituents at that point. Also, although we have used a frame to define the divergence, this is in fact only a convenient way to compute — the divergence may be defined without reference to a particular frame used for computation. With this in mind, consider

Example 8.5.20 (The Metric $\langle \cdot, \cdot \rangle$ has Zero Divergence).

To see this, take normal coordinates about an arbitrary point $p \in M$ and extend to a frame $\{\mathcal{E}_1, \ldots, \mathcal{E}_k\}$ by parallel translation along geodesics emanating from p. In particular, we have $\nabla_{\mathcal{E}_i}\mathcal{E}_j = 0$ at p. Now, because the divergence is determined by the values of the frame etc. at p, the same calculation as below may be made for a normal coordinate frame about every point in M. At p we have

$$\begin{aligned}\operatorname{div} \langle \cdot, \cdot \rangle(\mathcal{E}_i) &= \sum_j \nabla_{\mathcal{E}_j} \langle \mathcal{E}_j, \mathcal{E}_i \rangle\\ &= \sum_j \mathcal{E}_j \langle \mathcal{E}_j, \mathcal{E}_i \rangle - \langle \nabla_{\mathcal{E}_j}\mathcal{E}_j, \mathcal{E}_i \rangle - \langle \mathcal{E}_j, \nabla_{\mathcal{E}_j}\mathcal{E}_i \rangle\\ &= \sum_j \mathcal{E}_j \langle \mathcal{E}_j, \mathcal{E}_i \rangle - 0 - 0\end{aligned}$$

since we have normal coordinates at p. But the first term vanishes as well because $\langle \mathcal{E}_i, \mathcal{E}_j \rangle = \delta_i^j$ on the coordinate neighborhood and the $\mathcal{E}_j$-derivative of a constant is zero. Therefore,

$$\operatorname{div} \langle \cdot, \cdot \rangle(\mathcal{E}_i) = 0.$$

Since this is true for all $\mathcal{E}_i$ and the process is linear, then it holds for all vector fields as well.

Exercise* 8.5.21. For a smooth function $f : M \to \mathbb{R}$, show that

$$\operatorname{div}(f \langle \cdot, \cdot \rangle) = df$$

where df is the differential of f defined by $df(X) = X[f]$ for a vector field X. Thus, the divergence of conformal metrics (i.e., $f > 0$ at every p) depends on the conformality factor alone. Note that this makes sense in light of Example 8.5.10 since the divergence depends on the frame chosen. Verify that, for metric $f \langle \cdot, \cdot \rangle$, the frame $\{(1/\sqrt{f})\, \mathcal{E}_i\}$ gives $\operatorname{div}(f \langle \cdot, \cdot \rangle) = 0$.

We will now apply the divergence to Ricci curvature to find a remarkable relationship.

Theorem 8.5.22. *For Ricci curvature* Ric *and scalar curvature* κ*,*

$$d\kappa = 2 \operatorname{div}(\mathrm{Ric}).$$

Proof. Again choose a frame derived from parallel translation along geodesics of normal coordinates about an arbitrary point $p \in M$. Hence, at p, $\nabla_{\mathcal{E}_i} \mathcal{E}_j = 0$ for all i and j. Then

$$\begin{aligned}
2 \operatorname{div}(\mathrm{Ric})(\mathcal{E}_s) &= 2 \sum_j \nabla_{\mathcal{E}_j} \mathrm{Ric}(\mathcal{E}_j, \mathcal{E}_s) \\
&= 2 \sum_j \mathcal{E}_j \,\mathrm{Ric}(\mathcal{E}_j, \mathcal{E}_s) \\
&= 2 \sum_{i,j} \mathcal{E}_j \langle R(\mathcal{E}_j, \mathcal{E}_i)\mathcal{E}_s, \mathcal{E}_i \rangle \\
&= \sum_{i,j} \mathcal{E}_j \langle R(\mathcal{E}_j, \mathcal{E}_i)\mathcal{E}_s, \mathcal{E}_i \rangle + \sum_{i,j} \mathcal{E}_i \langle R(\mathcal{E}_i, \mathcal{E}_j)\mathcal{E}_s, \mathcal{E}_j \rangle
\end{aligned}$$

where the second term is obtained by switching i and j. Then, the symmetry relations for R imply

$$\begin{aligned}
2 \operatorname{div}(\mathrm{Ric})(\mathcal{E}_s) &= \sum_{i,j} \mathcal{E}_j \langle R(\mathcal{E}_j, \mathcal{E}_i)\mathcal{E}_s, \mathcal{E}_i \rangle + \sum_{i,j} \mathcal{E}_i \langle R(\mathcal{E}_j, \mathcal{E}_i)\mathcal{E}_j, \mathcal{E}_s \rangle \\
&= \sum_{i,j} \nabla R(\mathcal{E}_j, \mathcal{E}_i, \mathcal{E}_s, \mathcal{E}_i, \mathcal{E}_j) + \nabla R(\mathcal{E}_j, \mathcal{E}_i, \mathcal{E}_j, \mathcal{E}_s, \mathcal{E}_i) \\
&= -\sum_{i,j} \nabla R(\mathcal{E}_j, \mathcal{E}_i, \mathcal{E}_i, \mathcal{E}_j, \mathcal{E}_s)
\end{aligned}$$

by the second Bianchi identity. Then, interchanging the first two coordinates,

$$\begin{aligned}
2\operatorname{div}(\mathrm{Ric})(\mathcal{E}_s) &= \sum_{i,j} \nabla R(\mathcal{E}_i, \mathcal{E}_j, \mathcal{E}_i, \mathcal{E}_j, \mathcal{E}_s) \\
&= \sum_{i,j} \mathcal{E}_s R(\mathcal{E}_i, \mathcal{E}_j, \mathcal{E}_i, \mathcal{E}_j) \\
&= \mathcal{E}_s \sum_{i,j} R(\mathcal{E}_i, \mathcal{E}_j, \mathcal{E}_i, \mathcal{E}_j) \\
&= \mathcal{E}_s[\kappa] \\
&= d\kappa(\mathcal{E}_s).
\end{aligned}$$

Since this is true on a basis, it is true in general. ☐

Exercise* 8.5.23. The manifold $(M^k, \langle \cdot, \cdot \rangle)$ with metric $\langle \cdot, \cdot \rangle$ is an *Einstein manifold* if $\mathrm{Ric}(X, Y) = \lambda \langle X, Y \rangle$ for a fixed constant λ and all vector fields X and Y.

(1) Show that, if $\mathrm{Ric}(X, Y)(p) = \lambda(p) \langle X, Y \rangle(p)$ for a smooth function $\lambda \colon M^k \to \mathbb{R}$ with $k \geq 3$, then M^k is Einstein. Hints: (1) first compute κ and (2) then use Theorem 8.5.22.

(2) Show that a semisimple Lie group with Killing form metric $-b$ is an Einstein manifold. Hint: use the last part of Exercise 8.5.18.

Now, what has been the point of all this? The argument we presented earlier suggested that the Einstein and Ricci curvatures should be identified and that the Einstein curvature should be divergence free. But, as Theorem 8.5.22 shows, in general, the Ricci curvature has a nonzero divergence. So what is to be done to rescue the idea? By Exercise 8.5.21, we know

$$\operatorname{div}(\kappa \langle \cdot, \cdot \rangle) = d\kappa$$

so that $\operatorname{div}(\kappa \langle \cdot, \cdot \rangle) = 2\operatorname{div}(\mathrm{Ric})$ and, therefore,

$$\operatorname{div}\left(\mathrm{Ric} - \frac{1}{2}\kappa \langle \cdot, \cdot \rangle\right) = 0.$$

This calculation then says that, if we define the *Einstein curvature* to be $G = \mathrm{Ric} - \frac{1}{2}\kappa \langle \cdot, \cdot \rangle$, then we obtain a type of curvature which is divergence free. Also, we have

Theorem 8.5.24. *The Einstein and Ricci curvatures determine each other. In particular,* $G = 0$ *if and only if* $\mathrm{Ric} = 0$.

Proof. Denoting contraction by $\mathcal{C}$, we may write

$$\begin{aligned}
G &= \mathrm{Ric} - \frac{1}{2}\kappa \langle \cdot, \cdot \rangle \\
&= \mathrm{Ric} - \frac{1}{2}\mathcal{C}(\mathrm{Ric}) \langle \cdot, \cdot \rangle
\end{aligned}$$

since the contraction of the Ricci curvature is scalar curvature by the discussion preceding Example 8.5.10. Of course, Example 8.5.10 itself shows that $\mathcal{C}(\langle\cdot,\cdot\rangle) = \dim M$, so we also have

$$\begin{aligned} \mathcal{C}(G) &= \mathcal{C}(\mathrm{Ric}) - \frac{1}{2}\kappa\, \mathcal{C}(\langle\cdot,\cdot\rangle) \\ &= \kappa - \frac{1}{2}\kappa \dim M \\ &= \left(1 - \frac{\dim M}{2}\right)\kappa. \end{aligned}$$

Now, spacetime has four dimensions — time and $\mathbb{R}^3$ — so $\dim M = 4$. Hence, $\mathcal{C}(G) = -\kappa$ and

$$\begin{aligned} \mathrm{Ric} &= G + \frac{1}{2}\kappa\, \langle\cdot,\cdot\rangle \\ &= G - \frac{1}{2}\mathcal{C}(G)\, \langle\cdot,\cdot\rangle. \end{aligned}$$

Thus, Ric and G determine each other. □

Therefore, it seems that the Einstein curvature is a good choice for the Einstein field equation $G = 8\pi\, T$ (where the constant 8π is determined by taking the Newtonian limit). Indeed, in a vacuum, where $T = 0$, it is generally more convenient to solve the equations $\mathrm{Ric} = 0$ rather than $G = 0$, so the relationship with Ricci curvature is important. Perhaps we should mention what the word "solve" means here. What is known and what is unknown? Generally, we try to determine the geometric structure of spacetime from a hypothesis on the Ricci curvature and a knowledge of some facet of the relevant geometry. In Example 8.5.10 we showed that the Ricci curvature is given in terms of Christoffel symbols and their derivatives. These, in turn, are given in terms of derivatives of the metric. Hence, the expressions

$$R_{is} = 0$$

form a system of second-order partial differential equations *in the metric*. Therefore, when we solve this system, we are *determining the metric in that region of spacetime*. Since the metric determines everything, it can be safely said then that we understand the geometry of spacetime in that particular region.

Example 8.5.25 (The Schwarzschild Solution). The solution (i.e., the metric coefficients) for the vacuum field equations $R_{is} = 0$ outside a spherically symmetric body of mass M is given by

$$g_{00} = 1 - \frac{2M}{r},\ g_{11} = -\left(1 - \frac{2M}{r}\right)^{-1},\ g_{22} = -r^2,\ g_{33} = -r^2 \sin^2\phi,$$

where time t is usually given the index 0. This is the Schwarzschild solution discovered by Schwarzschild in 1916. Note that this metric is different from our usual ones in the sense that it is not positive definite. Spacetime is, in fact, based on the flat (Minkowski) metric of special relativity which has the form $g_{00} = 1$, $g_{11} = -1$, $g_{22} = -1$ and $g_{33} = -1$. These metrics are said to be *semi-Riemannian metrics* and much of the theory of ordinary Riemannian geometry carries over to these metrics as well.

since the contraction of the Ricci curvature is scalar curvature by the discussion preceding Example 8.5.10. Of course, Example 8.5.10 itself shows that $\mathcal{C}(\langle\cdot,\cdot\rangle) = \dim M$, so we also have

$$\begin{aligned}\mathcal{C}(G) &= \mathcal{C}(\mathrm{Ric}) - \frac{1}{2}\kappa\,\mathcal{C}(\langle\cdot,\cdot\rangle)\\ &= \kappa - \frac{1}{2}\kappa\dim M\\ &= \left(1 - \frac{\dim M}{2}\right)\kappa.\end{aligned}$$

Now, spacetime has four dimensions — time and $\mathbb{R}^3$ — so $\dim M = 4$. Hence, $\mathcal{C}(G) = -\kappa$ and

$$\begin{aligned}\mathrm{Ric} &= G + \frac{1}{2}\kappa\,\langle\cdot,\cdot\rangle\\ &= G - \frac{1}{2}\mathcal{C}(G)\,\langle\cdot,\cdot\rangle.\end{aligned}$$

Thus, Ric and G determine each other. ☐

Therefore, it seems that the Einstein curvature is a good choice for the Einstein field equation $G = 8\pi\, T$ (where the constant 8π is determined by taking the Newtonian limit). Indeed, in a vacuum, where $T = 0$, it is generally more convenient to solve the equations $\mathrm{Ric} = 0$ rather than $G = 0$, so the relationship with Ricci curvature is important. Perhaps we should mention what the word "solve" means here. What is known and what is unknown? Generally, we try to determine the geometric structure of spacetime from a hypothesis on the Ricci curvature and a knowledge of some facet of the relevant geometry. In Example 8.5.10 we showed that the Ricci curvature is given in terms of Christoffel symbols and their derivatives. These, in turn, are given in terms of derivatives of the metric. Hence, the expressions

$$R_{is} = 0$$

form a system of second-order partial differential equations *in the metric*. Therefore, when we solve this system, we are *determining the metric in that region of spacetime*. Since the metric determines everything, it can be safely said then that we understand the geometry of spacetime in that particular region.

Example 8.5.25 (The Schwarzschild Solution). The solution (i.e., the metric coefficients) for the vacuum field equations $R_{is} = 0$ outside a spherically symmetric body of mass M is given by

$$g_{00} = 1 - \frac{2M}{r},\ g_{11} = -\left(1 - \frac{2M}{r}\right)^{-1},\ g_{22} = -r^2,\ g_{33} = -r^2\sin^2\phi,$$

where time t is usually given the index 0. This is the Schwarzschild solution discovered by Schwarzschild in 1916. Note that this metric is different from our usual ones in the sense that it is not positive definite. Spacetime is, in fact, based on the flat (Minkowski) metric of special relativity which has the form $g_{00} = 1$, $g_{11} = -1$, $g_{22} = -1$ and $g_{33} = -1$. These metrics are said to be *semi-Riemannian metrics* and much of the theory of ordinary Riemannian geometry carries over to these metrics as well.

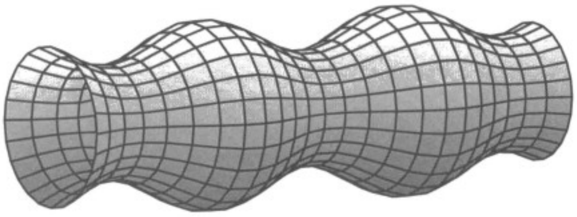

Figure 3.2. A surface of revolution

3.3 Surfaces of Revolution

We have already seen that many interesting examples of surfaces arise from revolving curves about an axis. These surfaces of revolution (see Figure 3.2) have patches of the form $\mathbf{x}(u, v) = (g(u), h(u)\cos v, h(u)\sin v)$. The curves on the surface which are circles formed by revolving a single point about the axis are called *parallels* (denoted π). The curves which look exactly like the original curve (but rotated out) are called *meridians* (denoted μ). Also, recall that $g(u)$ is the distance along the axis of revolution and $h(u)$ is the radius of a parallel. We have, $\mathbf{x}_u = (g', h'\cos v, h'\sin v)$, $\mathbf{x}_v = (0, -h\sin v, h\cos v)$ and $\mathbf{x}_u \times \mathbf{x}_v = (hh', -g'h\cos v, -g'h\sin v)$. Hence,

$$U = \frac{(h', -g'\cos v, -g'\sin v)}{\sqrt{g'^2 + h'^2}}.$$

Also, the second partial derivatives are $\mathbf{x}_{uu} = (g'', h''\cos v, h''\sin v)$, $\mathbf{x}_{uv} = (0, -h'\sin v, h'\cos v)$, and $\mathbf{x}_{vv} = (0, -h\cos v, -h\sin v)$. Hence,

$$E = g'^2 + h'^2, \qquad F = 0, \qquad G = h^2,$$
$$l = \frac{(g''h' - h''g')}{\sqrt{g'^2 + h'^2}}, \qquad m = 0, \qquad n = \frac{hg'}{\sqrt{g'^2 + h'^2}}.$$

Finally, the Gauss curvature is computed to be

$$K = \frac{g'(g''h' - h''g')}{h(g'^2 + h'^2)^2}.$$

Exercise 3.3.1. Verify the computations above.

If the original curve $\alpha(t) = (g(t), h(t), 0)$ has $g'(t) \neq 0$ for all t, then g is strictly increasing. Thus, it is one-to-one and has an inverse function g^{-1} which is differentiable as well. Such an inverse allows us to reparametrize the curve α. Define $f = h \circ g^{-1}$ and get

$$\begin{aligned}\overline{\alpha}(u) &= \alpha \circ g^{-1}(u) = (gg^{-1}(u), hg^{-1}(u), 0)\\ &= (u, f(u), 0).\end{aligned}$$

Thus, our calculations become somewhat easier. For instance, the formula for Gauss curvature becomes $K = -f''/f(1 + f'^2)^2$. To avoid confusion, we will still write $\alpha(u) = (u, h(u), 0)$ and $K = -h''/h(1 + h'^2)^2$.

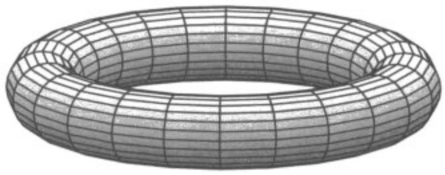

Figure 3.3. A torus

Example 3.3.2 (Torus Curvature).

$$\mathbf{x}(u,v) = (\underbrace{(R + r\cos u)}_{h(u)}\cos v,\ (R + r\cos u)\sin v,\ \underbrace{r\sin u}_{g(u)})$$

$$K = \frac{\cos u}{r(R + r\cos u)}.$$

Observe that, for $-\frac{\pi}{2} < u < \frac{\pi}{2}$, we have $\cos u > 0$. Hence $K > 0$ on the outer half of the torus. For $u = -\frac{\pi}{2}, \frac{\pi}{2}$ we have $\cos u = 0$, so $K = 0$ on the top and bottom of the torus. Finally, for $\frac{\pi}{2} < u < \frac{3\pi}{2}$ we have $\cos u < 0$ and $K < 0$ on the inner half of the torus. Note that the maximum K occurs at $u = 0$,

$$K(0) = \frac{1}{r(R + r)}$$

on the outermost circle, while the minimum K occurs at $u = \pi$,

$$K(\pi) = -\frac{1}{r(R - r)}$$

on the innermost circle.

Exercise 3.3.3. Verify the formula for K above.

Exercise∗ 3.3.4. For a surface of revolution parametrized as above, the v-parameter curves are parallels and the u-parameter curves are meridians. The normal curvatures in the directions of $\mathbf{x}_v$ and $\mathbf{x}_u$ are denoted k_π and k_μ respectively. Show that $k_\mu = l/E$ and $k_\pi = n/G$ by calculating the eigenvectors of S as $\mathbf{x}_u$ and $\mathbf{x}_v$ with eigenvalues l/E, n/G respectively. This shows directly that meridians and parallels are lines of curvature. Hint: $S(\mathbf{x}_u) = a\mathbf{x}_u + b\mathbf{x}_v$. Use $F = 0, m = 0$.

Exercise 3.3.5. For a general surface of revolution, show that

$$k_\mu = \frac{g''h' - g'h''}{(g'^2 + h'^2)^{3/2}}, \qquad k_\pi = \frac{g'}{h(g'^2 + h'^2)^{1/2}}$$

and obtain formulas for K and H. What is k_μ on a torus? What is the geometric reason for this?

Exercise 3.3.6. Revolve the catenary $y = c\cosh(\frac{x}{c})$ around the x-axis to obtain the *catenoid*. Show the following:

$$k_\mu = -\frac{1}{c\cosh^2\frac{u}{c}}, \qquad k_\pi = -k_\mu, \qquad K = -\frac{1}{c^2\cosh^4\frac{u}{c}}, \qquad H = 0.$$

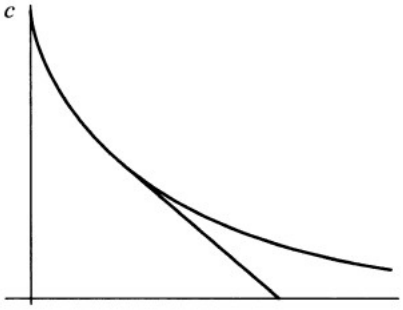

Figure 3.4. Tangent line segment of length c

Can you think of another surface of revolution which is minimal?

Exercise∗ 3.3.7. For the following surfaces of revolution, find K and describe where $K > 0$, $K < 0$, $K = 0$.

(a) Revolve $\alpha(u) = (u, e^{-u^2/2}, 0)$ about the x-axis.

(b) *Elliptical Torus.* Revolve the ellipse $\dfrac{(x-R)^2}{a^2} + \dfrac{y^2}{b^2} = 1$ about the y-axis.

Exercise 3.3.8. If a surface of revolution is generated by a *unit speed curve* $\alpha(u) = (g(u), h(u), 0)$, then show (a) $E = 1,\ F = 0,\ G = h^2$ and (b) $K = -h''/h$.

Example 3.3.9 (The Pseudosphere).
As another example of our computational techniques, we will consider a surface of revolution which is analogous to the sphere in the sense that it has *constant Gauss curvature*. This surface is called the *Pseudosphere* and it is defined by means of a condition on the generating curve. Namely, let α be a curve which begins at $(0, c)$ and traces a path so that its tangent line at any point reaches the x-axis after running a distance exactly equal to c. This means the curve α must decrease and flatten. The curve α is called a *tractrix* and is physically represented by the path a ship would follow when starting at $(0, c)$ and is being pulled by a tugboat which moves along the x-axis (see Figure 3.4).

Write $\alpha(u) = (u, h(u))$ and note $h' < 0$ and $g(u) = u$. Now, $\alpha'(u) = (1, h'(u))$ and a tangent line at $(u, h(u))$ has the equation

$$l(t) = \alpha(u) + t\alpha'(u).$$

The y-coordinate of $l(t)$ is then $h + th'$, so for $t = 0$ we are at $\alpha(u)$ and for $t = -h/h'$ we are on the x-axis. We want the length of this segment to be c, so we have

$$\begin{aligned} c &= |\alpha(u) - (\alpha(u) - \frac{h}{h'}\alpha'(u))| \\ &= \frac{h}{|h'|}|\alpha'(u)| \\ &= \frac{h}{|h'|}\sqrt{1 + h'^2}. \end{aligned}$$

Then $c^2 = \frac{h^2}{h'^2}(1 + h'^2) = \frac{h^2}{h'^2} + h^2$. Solving for h' and then taking another derivative, we get

$$h' = -\frac{h}{\sqrt{c^2 - h^2}} \qquad h'' = -\frac{h'c^2}{(c^2 - h^2)^{3/2}}.$$

Exercise 3.3.10. Show that, for the surface of revolution obtained by revolving α about the x-axis, we obtain (a) $k_\mu = h'/c$ (b) $k_\pi = -1/ch'$ and (c) $K = -1/c^2$.

Note that the Gauss curvature of the pseudosphere is a constant. However, in contrast to a sphere's constant positive curvature, the pseudosphere's curvature is negative. For $c = 1$, we obtain a surface at each point of which the Gauss curvature is -1. Later we will meet up with more geometrically natural surfaces of this kind.

Exercise* 3.3.11. Solve the separable differential equation

$$h' = -\frac{h}{\sqrt{1 - h^2}}$$

in two ways to get (1) $u = \ln|\frac{\sqrt{1-h^2}}{h} + \frac{1}{h}| - \sqrt{1 - h^2}$ and (2) $h(u) = \operatorname{sech} w$, $u = w - \tanh w$. For the first, turn the picture over so we revolve the tractrix about the z-axis and the pseudosphere (with $c = 1$) lies over the disk of radius 1. If we write things in cylindrical coordinates (r, θ, z), then the old u becomes z and the old h becomes r. By substitution in the formula above, we get

$$z = -\sqrt{1 - r^2} - \ln(r) + \ln(1 + \sqrt{1 - r^2}).$$

Of course, rectangular coordinates may now be introduced by $r = \sqrt{x^2 + y^2}$. Use this formula to graph the pseudosphere via Maple. For the second, we obtain a patch for the pseudosphere

$$\mathbf{x}(w, v) = (w - \tanh w,\ \operatorname{sech} w \cos v, \operatorname{sech} w \sin v).$$

Plot this patch with Maple also. Compare the two plots. Hints: For the first, use trigonometric substitution with $h = \cos w$. For the second, use hyperbolic trigonometric substitution.

The pseudosphere (see Figure 3.5) is actually determined by the requirement of constant curvature and certain initial conditions. Suppose we start off by requiring that a surface of

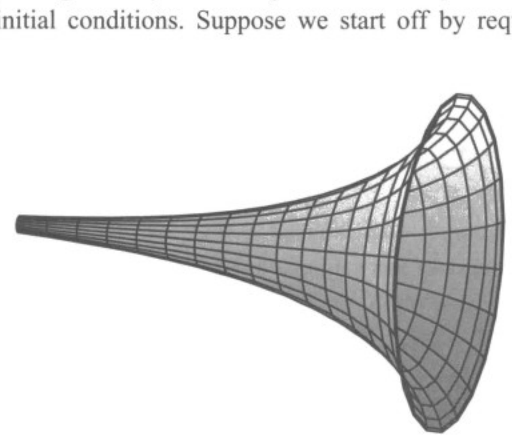

Figure 3.5. A pseudosphere

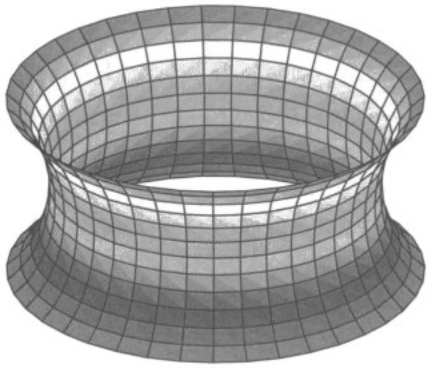

Figure 3.6. A constant negative curvature surface of revolution

revolution M have constant Gauss curvature $K = -1$. We can suppose also that M is parametrized by $\mathbf{x}(u, v) = (g(u), h(u)\cos v, h(u)\sin v)$ with $g'(u)^2 + h'(u)^2 = 1$ (i.e., the profile curve is unit speed). Then, as we have seen, the formula for Gauss curvature reduces to $K = -h''/h$, so the requirement of constant curvature $K = -1$ gives us a linear differential equation to solve,

$$h'' = h.$$

The solution to this differential equation is easy. We obtain

$$h(u) = A\,e^{u} + B\,e^{-u}.$$

Then, plugging $h(u)$ into the the unit speed relation $g'(u)^2 + h'(u)^2 = 1$ and solving for $g(u)$ gives

$$g(u) = \pm \int_{u_0}^{u} \sqrt{1 - (A\,e^{t} - B\,e^{-t})^2}\,dt.$$

In choosing A and B, new surfaces of constant Gauss curvature $K = -1$ are created. See Figure 3.6.

Exercise 3.3.12. Let $A = 1$ and $B = 0$. Show that this surface of constant Gauss curvature $K = -1$ is the pseudosphere with $c = 1$. Hint: let $e^{t} = \operatorname{sech} w$.

Exercise 3.3.13. Carry out the same procedure for the case of positive Gauss curvature $K = +1$. What should A and B be to get the unit sphere? See Exercise 3.7.8 and Section 3.8.

3.4 A Formula for Gauss Curvature

So far, we have seen that Gauss curvature tells us a great deal about the geometry of a surface. However, since we compute K using the unit normal U, it seems that the precise way in which we situate a surface in $\mathbb{R}^3$ may well change K while leaving the essential geometry unaltered. This would mean that Gauss curvature would not be an invariant of the geometry of the surface. We shall now give a formula for K which obviates the use of U and shows that K depends only on E, F, and G, the so-called *metric* of the surface. In fact, although a more general formula exists (see Exercise 3.4.7 below), we shall restrict ourselves to the case where $F = \mathbf{x}_u \cdot \mathbf{x}_v = 0$.

That is, we shall assume that the u and v-parameter curves always meet at right angles. In this case we prove

Theorem 3.4.1. *The Gauss curvature depends only on the metric E, F and G. When $F = 0$, Gauss curvature is given by*

$$(*) \qquad K = -\frac{1}{2\sqrt{EG}}\left(\frac{\partial}{\partial v}\left(\frac{E_v}{\sqrt{EG}}\right) + \frac{\partial}{\partial u}\left(\frac{G_u}{\sqrt{EG}}\right)\right).$$

Here, we have used the notation

$$E_v = \frac{\partial}{\partial v}E = \frac{\partial}{\partial v}(\mathbf{x}_u \cdot \mathbf{x}_u) \quad \text{and} \quad G_u = \frac{\partial}{\partial u}G = \frac{\partial}{\partial u}(\mathbf{x}_v \cdot \mathbf{x}_v).$$

Of course, to prove the theorem, we will show that our usual formula $K = (ln - m^2)/(EG - F^2)$ reduces to the one above. Notice that the formula of the theorem does not depend on the particular patch used, but only on the metric coefficients E, G (and implicitly F). We will consider this a bit more carefully later when we discuss isometries of surfaces, but suffice it to say that, intuitively, a surface may be "bent" without stretching and still have the same Gauss curvature. The prime example of this phenomenon is the usual right circular cylinder. If the cylinder is cut along a ruling it can be made to lie flat on a plane without stretching. This means that, locally, the cylinder and the plane have the same geometry. This further explains, geometrically, the vanishing of the cylinder's Gauss curvature. The dependence of K on the metric alone is known as *Gauss's Theorem Egregium* (meaning "remarkable theorem").

We point out two things. First, this result *does not mean* that two surfaces with the same Gauss curvature are *the same geometrically*. Surfaces may have the same formulas for Gauss curvature while still having fundamentally different geometries encoded in their individual metrics. To make this precise we need the notion of isometry, so for the moment we give

Exercise 3.4.2. Show that the following two patches have the same Gauss curvature:

$$\mathbf{x}(u, v) = (u\cos v, u\sin v, v), \qquad \mathbf{y}(u, v) = (u\cos v, u\sin v, \ln u).$$

These surfaces are not isometric as will be shown in Exercise 5.5.2.

Secondly, the Theorem Egregium holds for Gauss curvature, *but not for mean curvature*. For instance, as we have mentioned, a cylinder and a plane are geometrically the same with the same Gauss curvature $K = 0$, but their mean curvatures are quite different. The mean curvature of the plane is zero while that of the cylinder is one-half the curvature of the base circle. Thus, mean curvature is not truly an invariant of *the surface itself*, but depends on the way in which the surface sits in $\mathbb{R}^3$.

We now turn to the derivation of the formula in Theorem 3.4.1. While the details of the calculations are quite tedious, keep in mind that we are simply finding coefficients for a vector in terms of a particular basis. Because $l = \mathbf{x}_{uu} \cdot U$, $m = \mathbf{x}_{uv} \cdot U$ and $n = \mathbf{x}_{vv} \cdot U$, we need expressions for

$\mathbf{x}_{uu}$, $\mathbf{x}_{uv}$, and $\mathbf{x}_{vv}$ in terms of the basis for 3-space $\{\mathbf{x}_u, \mathbf{x}_v, U\}$. Write

$$\begin{aligned}\mathbf{x}_{uu} &= \Gamma^u_{uu}\mathbf{x}_u + \Gamma^v_{uu}\mathbf{x}_v + lU \quad (\text{since } \mathbf{x}_{uu}\cdot U = l \text{ etc.})\\ \mathbf{x}_{uv} &= \Gamma^u_{uv}\mathbf{x}_u + \Gamma^v_{uv}\mathbf{x}_v + mU\\ \mathbf{x}_{vv} &= \Gamma^u_{vv}\mathbf{x}_u + \Gamma^v_{vv}\mathbf{x}_v + nU. \end{aligned} \tag{**}$$

Our job is to find the Γ's. These are just coefficients in a basis expansion, but they are known by the name *Christoffel Symbols*. We use what we know about dot products to determine the Γ's.

$$\begin{aligned}\mathbf{x}_{uu}\cdot\mathbf{x}_u &= \Gamma^u_{uu}\mathbf{x}_u\cdot\mathbf{x}_u + 0 + 0\\ &= \Gamma^u_{uu}E \qquad \text{by definition of } E.\end{aligned}$$

If we can compute $\mathbf{x}_{uu}\cdot\mathbf{x}_u$, then we will know Γ^u_{uu}. This is simply the product rule.

$$E = \mathbf{x}_u\cdot\mathbf{x}_u, \text{ so } E_u = \mathbf{x}_{uu}\cdot\mathbf{x}_u + \mathbf{x}_u\cdot\mathbf{x}_{uu} = 2\mathbf{x}_{uu}\cdot\mathbf{x}_u.$$

Thus

$$\mathbf{x}_{uu}\cdot\mathbf{x}_u = \frac{E_u}{2} \quad\text{and}\quad \Gamma^u_{uu} = \frac{E_u}{2E}.$$

Further, $\mathbf{x}_u\cdot\mathbf{x}_v = 0$ so taking the partial with respect to u gives

$$0 = \mathbf{x}_{uu}\cdot\mathbf{x}_v + \mathbf{x}_u\cdot\mathbf{x}_{uv} \quad\text{or}\quad \mathbf{x}_{uu}\cdot\mathbf{x}_v = -\mathbf{x}_u\cdot\mathbf{x}_{uv}.$$

Also, $E = \mathbf{x}_u\cdot\mathbf{x}_u$, so taking the partial with respect to v gives $E_v = 2\mathbf{x}_u\cdot\mathbf{x}_{uv}$ and, consequently, $E_v/2 = \mathbf{x}_u\cdot\mathbf{x}_{uv} = -\mathbf{x}_{uu}\cdot\mathbf{x}_v$. Moreover,

$$\Gamma^v_{uu} = (\mathbf{x}_{uu}\cdot\mathbf{x}_v)/G = -E_v/2G \quad\text{and}\quad \Gamma^u_{uv} = \mathbf{x}_{uv}\cdot\mathbf{x}_u/E = E_v/2E.$$

Continuing on, $G = \mathbf{x}_v\cdot\mathbf{x}_v$, so $G_u/2 = \mathbf{x}_{uv}\cdot\mathbf{x}_v$. Then, since $0 = \mathbf{x}_v\cdot\mathbf{x}_u$, we have

$$-\mathbf{x}_v\cdot\mathbf{x}_{uv} = \mathbf{x}_{vv}\cdot\mathbf{x}_u$$

with

$$\Gamma^v_{uv} = \mathbf{x}_{uv}\cdot\mathbf{x}_v/G = G_u/2G \quad\text{and}\quad \Gamma^u_{vv} = \mathbf{x}_{vv}\cdot\mathbf{x}_u/E = -G_u/2E.$$

Finally, $\mathbf{x}_v\cdot\mathbf{x}_v = G$, so $\mathbf{x}_{vv}\cdot\mathbf{x}_v = G_v/2$ and $\Gamma^v_{vv} = \mathbf{x}_{vv}\cdot\mathbf{x}_v/G = G_v/2G$. We end up with the following

Formula(s) 3.4.3 (Fundamental Acceleration Formulas: $F = 0$ Case).

$$\begin{aligned}\mathbf{x}_{uu} &= \frac{E_u}{2E}\mathbf{x}_u - \frac{E_v}{2G}\mathbf{x}_v + lU\\ \mathbf{x}_{uv} &= \frac{E_v}{2E}\mathbf{x}_u + \frac{G_u}{2G}\mathbf{x}_v + mU\\ \mathbf{x}_{vv} &= -\frac{G_u}{2E}\mathbf{x}_u + \frac{G_v}{2G}\mathbf{x}_v + nU\\ U_u &= -\frac{l}{E}\mathbf{x}_u - \frac{m}{G}\mathbf{x}_v\\ U_v &= -\frac{m}{E}\mathbf{x}_u - \frac{n}{G}\mathbf{x}_v\end{aligned}$$

Exercise∗ 3.4.4. Compute U_u and U_v as above. Hint: $U_u \overset{\text{def}}{=} \nabla_{\mathbf{x}_u} U = A\mathbf{x}_u + B\mathbf{x}_v$. Find A by $\nabla_{\mathbf{x}_u} U \cdot \mathbf{x}_u = A\mathbf{x}_u \cdot \mathbf{x}_u = AE$ and $0 = \mathbf{x}_u(U \cdot \mathbf{x}_u) = \nabla_{\mathbf{x}_u} U \cdot \mathbf{x}_u + U \cdot \mathbf{x}_{uu}$.

We know that mixed partial derivatives are equal no matter the order of differentiation, so $\mathbf{x}_{uuv} = \mathbf{x}_{uvu}$, or $\mathbf{x}_{uuv} - \mathbf{x}_{uvu} = 0$. This means that the coefficients of $\mathbf{x}_u$, $\mathbf{x}_v$ and U are *all zero* when $\mathbf{x}_{uuv} - \mathbf{x}_{uvu}$ is written in this basis. Let's concentrate on the $\mathbf{x}_v$-term. (We again use the product rule repeatedly.)

$$\mathbf{x}_{uuv} = \left(\frac{E_u}{2E}\right)_v \mathbf{x}_u + \frac{E_u}{2E}\mathbf{x}_{uv} - \left(\frac{E_v}{2G}\right)_v \mathbf{x}_v - \frac{E_v}{2G}\mathbf{x}_{vv} + l_v U + l U_v.$$

Now replace $\mathbf{x}_{uv}$, $\mathbf{x}_{vv}$ and U_v by their basis expansions to get,

$$\mathbf{x}_{uuv} = [\ ]\mathbf{x}_u + \left[\frac{E_u G_u}{4EG} - \left(\frac{E_v}{2G}\right)_v - \frac{E_v G_v}{4G^2} - \frac{ln}{G}\right]\mathbf{x}_v + [\ ]U$$

$$\mathbf{x}_{uvu} = \left(\frac{E_v}{2E}\right)_u \mathbf{x}_u + \frac{E_v}{2E}\mathbf{x}_{uu} + \left(\frac{G_u}{2G}\right)_u \mathbf{x}_v + \frac{G_u}{2G}\mathbf{x}_{uv} + m_u U + m U_u$$

$$\mathbf{x}_{uvu} = [\ ]\mathbf{x}_u + \left[-\frac{E_v E_v}{4EG} + \left(\frac{G_u}{2G}\right)_u + \frac{G_u G_u}{4G^2} - \frac{m^2}{G}\right]\mathbf{x}_v + [\ ]U.$$

Because the $\mathbf{x}_v$-coefficient of $\mathbf{x}_{uuv} - \mathbf{x}_{uvu}$ is zero, we get,

$$0 = \frac{E_u G_u}{4EG} - \left(\frac{E_v}{2G}\right)_v - \frac{E_v G_v}{4G^2} + \frac{E_v E_v}{4EG} - \left(\frac{G_u}{2G}\right)_u - \frac{G_u G_u}{4G^2} - \frac{ln - m^2}{G}.$$

Notice the last term! Dividing by E, we have

$$\frac{ln - m^2}{EG} = \frac{E_u G_u}{4E^2 G} - \frac{1}{E}\left(\frac{E_v}{2G}\right)_v - \frac{E_v G_v}{4EG^2} + \frac{E_v E_v}{4E^2 G} - \frac{1}{E}\left(\frac{G_u}{2G}\right)_u - \frac{G_u G_u}{4EG^2}.$$

Of course, the left-hand side is K (since $F = 0$) and the right-hand side only depends on E and G. Thus, we have a formula for K which does not make explicit use of the normal U.

Exercise∗ 3.4.5. Show that the right-hand side above is

$$-\frac{1}{2\sqrt{EG}}\left(\frac{\partial}{\partial v}\left(\frac{E_v}{\sqrt{EG}}\right) + \frac{\partial}{\partial u}\left(\frac{G_u}{\sqrt{EG}}\right)\right).$$

Exercise∗ 3.4.6. Compute the curvature of the sphere of radius R using this formula. Does this agree with what the shape operator tells you?

Exercise 3.4.7. Show, by similar arguments, that, if $F \neq 0$, the following formula holds.

$$K = \frac{1}{(EG - F^2)^2}\left(\begin{vmatrix} -\frac{E_{vv}}{2} + F_{uv} - \frac{G_{uu}}{2} & \frac{E_u}{2} & F_u - \frac{E_v}{2} \\ F_v - \frac{G_u}{2} & E & F \\ \frac{G_v}{2} & F & G \end{vmatrix} - \begin{vmatrix} 0 & \frac{E_v}{2} & \frac{G_u}{2} \\ \frac{E_v}{2} & E & F \\ \frac{G_u}{2} & F & G \end{vmatrix}\right)$$

Exercise 3.4.8. Show that the U-terms of $\mathbf{x}_{uuv} - \mathbf{x}_{uvu} = 0$ and $\mathbf{x}_{vvu} - \mathbf{x}_{uvv} = 0$ give

$$\begin{aligned} l_v - m_u &= m\left(\frac{G_u}{2G} - \frac{E_u}{2E}\right) + E_v\, H \\ &= l\,\frac{E_v}{2E} + m\left(\frac{G_u}{2G} - \frac{E_u}{2E}\right) + n\,\frac{E_v}{2G} \end{aligned}$$

and

$$\begin{aligned} n_u - m_v &= m\left(\frac{E_v}{2E} - \frac{G_v}{2G}\right) + G_u\, H \\ &= l\,\frac{G_u}{2E} + m\left(\frac{E_v}{2E} - \frac{G_v}{2G}\right) + n\,\frac{G_u}{2G} \end{aligned}$$

where H is mean curvature (and F is assumed to be zero). These are the Codazzi-Mainardi equations. What do they say in the case where the surface is minimal (i.e., $H = 0$) and the parametrization satisfies $E = G$, $F = 0$?

3.5 Some Effects of Curvature(s)

In this section we will derive some geometric consequences of various conditions on Gauss and mean curvatures. Of course, the point of all this is simply to show that we have not wasted our time so far in learning how to compute K and H. These quantities really do tell us about the geometry of a surface. In the next chapter we will focus on mean curvature and see that it is intimately related to certain structures in Nature. For now, however, our goal is simply to see geometry reflected in the linear algebra and calculus represented by K and H.

Let us start by considering a completely impractical situation! Namely, suppose we know that every point on a surface is an umbilic point. Recall that this means that, at every point, the principal curvatures are equal. Although this does not mean, *a priori*, that the principal curvatures are constant functions on the surface (with the same value), this is precisely what we shall prove below. Of course, a direct verification of such a hypothesis is out of the question, yet we shall see later that this result is useful all the same. First try the following

Exercise* 3.5.1. Suppose $p \in M$ is an umbilic point (i.e., $k_1 = k_2 = k$ at p). Show that, at p,

$$k = \frac{l}{E} = \frac{n}{G}.$$

Moreover, if $F \neq 0$, then $k = \frac{m}{F}$ as well. Hint: Write $S(\mathbf{x}_u) = A\mathbf{x}_u + B\mathbf{x}_v = k\mathbf{x}_u$ by Theorem 2.4.10 since p is umbilic. Use $B = 0$ in

$$l = S(\mathbf{x}_u)\cdot \mathbf{x}_u = AE + BF, \quad m = S(\mathbf{x}_u)\cdot \mathbf{x}_v = AF + BG.$$

Solve for A. Do the same for $S(\mathbf{x}_v)$.

Theorem 3.5.2. *A surface M consisting entirely of umbilic points is contained in either a plane or a sphere.*

Proof. Although we know that $k_1 = k_2 = k$ at any $p \in M$, it is still possible that k may vary from point to point in M. Now, by Theorem 2.4.10, we see that

$$\nabla_{\mathbf{x}_u} U = -S(\mathbf{x}_u) = -k\mathbf{x}_u \quad \text{and} \quad \nabla_{\mathbf{x}_v} U = -S(\mathbf{x}_v) = -k\mathbf{x}_v.$$

Now the mixed covariant derivative may be calculated by Exercise 2.2.9.

$$\begin{aligned}\nabla_{\mathbf{x}_v}\nabla_{\mathbf{x}_u} U &= \sum_k \left\{ \sum_i \left[\left(\sum_j \frac{\partial^2 u_k}{\partial x_i \partial x_j} \frac{\partial x_i}{\partial u} \frac{\partial x_j}{\partial v} \right) + \frac{\partial u_k}{\partial x_i} \frac{\partial^2 x_i}{\partial u \partial v} \right] \right\} e_k \\ &= \nabla_{\mathbf{x}_u}\nabla_{\mathbf{x}_v} U\end{aligned}$$

so differentiating $-k\mathbf{x}_u$ by v and $-k\mathbf{x}_v$ by u gives $-k_v\mathbf{x}_u - k\mathbf{x}_{uv} = -k_u\mathbf{x}_v - k\mathbf{x}_{vu}$ and, hence, $k_v\mathbf{x}_u = k_u\mathbf{x}_v$. But $\mathbf{x}_u$ and $\mathbf{x}_v$ are linearly independent, so we must have $k_v = 0 = k_u$ and this implies that k *is a constant function*.

Case 1. Suppose $k = 0$. Then the shape operator is zero everywhere. By Theorem 2.2.17, M is contained in a plane.

Case 2. Suppose $k \neq 0$. Then consider $\mathbf{x}(u, v) + \frac{1}{k}U$ and compute,

$$\begin{aligned}\frac{\partial}{\partial u}\left[\mathbf{x}(u, v) + \frac{1}{k}U\right] &= \mathbf{x}_u + \frac{1}{k}U_u \\ &= \mathbf{x}_u + \frac{1}{k}[-k\mathbf{x}_u] \qquad \text{by above} \\ &= \mathbf{x}_u - \mathbf{x}_u \\ &= 0.\end{aligned}$$

In a similar fashion

$$\frac{\partial}{\partial v}\left[\mathbf{x}(u, v) + \frac{1}{k}U\right] = 0.$$

Hence, $\mathbf{x}(u, v) + \frac{1}{k}U = p$, a constant. Then $|\mathbf{x}(u, v) - p| = |-\frac{1}{k}| = \frac{1}{|k|}$, so every point $\mathbf{x}(u, v)$ lies a distance $1/|k|$ from p. Hence, $\mathbf{x}(u, v)$ lies on a sphere of radius $1/|k|$.

We note that we have proved the result only for a single patch. However, because M is connected, any patch must overlap some other patch. For a patch which overlaps $\mathbf{x}(u, v)$, the points in the overlap lie on the plane or sphere as above. Since the argument above produces a constant k on the new patch, it must therefore agree with the k of $\mathbf{x}(u, v)$. Hence, the geometry of $\mathbf{x}(u, v)$ propagates to overlapping patches. This then continues over the whole surface, each patch reproducing the geometry of its neighbors. □

Notice that, for an all umbilic surface, by Exercise 3.5.1 we obtain

$$K = \frac{ln - m^2}{EG - F^2} = \frac{\frac{l}{E}\frac{n}{G}(EG) - \frac{m^2}{F^2}F^2}{EG - F^2} = \frac{\frac{l^2}{E^2}(EG - F^2)}{EG - F^2} = \frac{l^2}{E^2} = k^2,$$

so $|k| = \sqrt{K}$ and the radius of the sphere is $1/\sqrt{K}$. The next result restricts the form surfaces in $\mathbb{R}^3$ may take in terms of their possible Gauss curvatures. Recall that a surface in $\mathbb{R}^3$ is *compact* if it is closed and bounded. Here, "bounded" means that the surface may be enclosed by a sphere of sufficient size. The term "closed" means that any sequence of points on the surface which converges to some point in $\mathbb{R}^3$ actually converges to a point on the surface. In the following result the hypothesis of "compactness" is used to ensure that maximums and minimums of functions are actually attained by the functions.

Theorem 3.5.3. *On every compact surface $M \subseteq \mathbb{R}^3$ there is some point p with $K(p) > 0$.*

Proof. Let $f : M \to \mathbb{R}$ be defined by $f(p) = |p|^2$. This function is continuous, so (since M is *compact*) f attains its max and its min. Let p_0 be a max for f. Now $f(p_0) = |p_0|^2$, so p_0 is the farthest point from the origin on M. If $r = |p_0|$, then M is inside the sphere of radius r as shown in Figure 3.7.

Therefore, it seems likely that $K(p_0) \geq \frac{1}{r^2}$, where the $\frac{1}{r^2}$ is the curvature of the r-sphere $S^2(r)$. We verify this and so prove the theorem. Consider a unit tangent vector $\mathbf{u} \in T_{p_0}(M)$ and take a curve α in M with $\alpha(0) = p_0$ and $\alpha'(0) = \mathbf{u}$. Of course the composition $f \circ \alpha$ still has a max at p_0 (since α lies in M), so we have the usual calculus conditions

$$\begin{aligned} \frac{d}{dt}(f \circ \alpha)\,|_{t=0} &= 0 \\ \frac{d^2}{dt^2}(f \circ \alpha)\,|_{t=0} &\leq 0. \end{aligned}$$

Now $f \circ \alpha(t) = |\alpha(t)|^2 = \alpha(t) \cdot \alpha(t)$, so $\frac{d}{dt}(f \circ \alpha) = 2\alpha \cdot \alpha'$. At $t = 0$,

$$0 = 2\alpha(0) \cdot \alpha'(0) = 2p_0 \cdot \mathbf{u}.$$

But $\mathbf{u}$ is any unit vector in $T_{p_0}M$, so $p_0 \cdot \mathbf{u} = 0$ for all $\mathbf{u} \in T_{p_0}M$. Hence, p_0 is normal to M at p_0. Also, $(d^2/dt^2)(f \circ \alpha) = 2\alpha' \cdot \alpha' + 2\alpha \cdot \alpha''$ and at $t = 0$,

$$\begin{aligned} 0 \geq \frac{d^2}{dt^2}(f \circ \alpha) &= 2\alpha'(0) \cdot \alpha'(0) + 2\alpha(0) \cdot \alpha''(0) \\ &= 2\,\mathbf{u} \cdot \mathbf{u} + 2\,p_0 \cdot \alpha''(0) \\ &= 2 + 2\,p_0 \cdot \alpha''(0) \\ -2 &\geq 2\,p_0 \cdot \alpha''(0) \\ -1 &\geq p_0 \cdot \alpha''(0). \end{aligned}$$

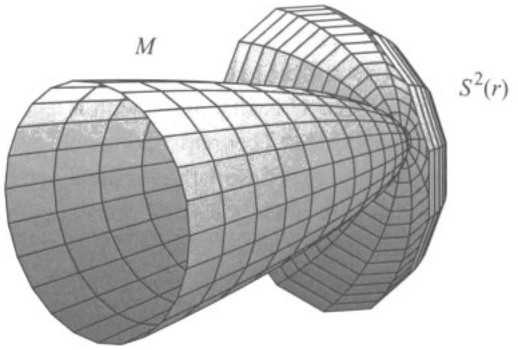

Figure 3.7. Surface at the farthest point p_0

Now, let's calculate normal curvatures at p_0. Since $p_0/r = p_0/|p_0|$ is a unit normal to M at p_0, the calculations above give

$$\begin{aligned} k(\mathbf{u}) &= S(\mathbf{u}) \cdot \mathbf{u} \\ &= S(\alpha') \cdot \alpha' \\ &= U(p_0) \cdot \alpha'' \\ &= \frac{p_0}{r} \cdot \alpha'' \\ &\leq -\frac{1}{r}. \end{aligned}$$

In particular, both k_1 and k_2 are less than or equal to $-1/r$. Thus the Gauss curvature has

$$K(p_0) = k_1(p_0)k_2(p_0) \geq \frac{1}{r} \cdot \frac{1}{r} = \frac{1}{r^2} > 0.$$

□

Corollary 3.5.4. *There are no compact surfaces in $\mathbb{R}^3$ with $K \leq 0$. In particular, no minimal surface embedded in $\mathbb{R}^3$ is compact.*

We now come to one of the main results displaying the influence of curvature on geometry. Previously, we saw that an all umbilic surface is either a plane ($K = 0$) or a sphere ($K > 0$ and constant). There is, however, a much stronger result which is based on the Gauss curvature K itself rather than its factors k_1 and k_2. This beautiful result is

Theorem 3.5.5 (H. Liebmann). *If M is a compact surface of constant Gauss curvature K, then M is a sphere of radius $1/\sqrt{K}$.*

Proof. By Theorem 3.5.3, there is some $p \in M$ with $K(p) > 0$. Since K is constant, then $K > 0$ everywhere on M. Because K is constant and $K = k_1 k_2$, where k_1 and k_2 are the principal curvatures, if p is a point where the function k_1 is a maximum, then p is also a point where k_2 is a minimum. We know that such a p exists since M is compact and k_1, k_2 are continuous functions. We now have two cases to consider.

Case 1. Suppose $k_1 = k_2$ at p. Because k_1 is a maximum and k_2 is a minimum at p, *all* normal curvatures on M must lie between these two values. Therefore, if $k_1 = k_2$ at p, the normal curvature k must be constant on M. Hence M is all umbilic with $K > 0$, so it is a sphere of radius $1/\sqrt{K}$. (This is what we meant by saying that Theorem 3.5.2 would prove useful.) We shall now show that Case 2 below is impossible, so that Case 1 proves the theorem.

Case 2. Suppose $k_1 > k_2$ at p. There is actually a small open neighborhood of p in M where $k_1 > k_2$. If not, we could take a decreasing sequence of open "balls" about p and choose points α_i in the balls with $k_1(\alpha_i) = k_2(\alpha_i)$ and $\alpha_i \to p$. Then, by continuity, we would have $k_1(p) = k_2(p)$. This contradicts our assumption, so we must have the desired neighborhood. Although we shall not prove it, on this neighborhood we may choose two principal directions which are perpendicular at each point and use these to provide a patch with $\mathbf{x}_u \cdot \mathbf{x}_u = E$, $\mathbf{x}_u \cdot \mathbf{x}_v = 0$, $\mathbf{x}_v \cdot \mathbf{x}_v = G$ and, since $\mathbf{x}_u$, $\mathbf{x}_v$ are *principal vectors*,

$$S(\mathbf{x}_u) = k_1 \mathbf{x}_u, \quad S(\mathbf{x}_v) = k_2 \mathbf{x}_v.$$

(A proof of the existence of such a patch $\mathbf{x}$ may be found in [dC76, p. 185].) Note that $m = S(\mathbf{x}_u) \cdot \mathbf{x}_v = k_1 \mathbf{x}_u \cdot \mathbf{x}_v = 0$ also. Moreover, $l = S(\mathbf{x}_u) \cdot \mathbf{x}_u = k_1 \mathbf{x}_u \cdot \mathbf{x}_u = k_1 E$, so $k_1 = \frac{l}{E}$ and, similarly, $k_2 = \frac{n}{G}$. Of course, we have the formula,

$$K = -\frac{1}{2\sqrt{EG}} \left[\frac{\partial}{\partial v} \left(\frac{E_v}{\sqrt{EG}} \right) + \frac{\partial}{\partial u} \left(\frac{G_u}{\sqrt{EG}} \right) \right],$$

and from the U-term of $\mathbf{x}_{uuv} - \mathbf{x}_{uvu} = 0$ (see Exercise 3.4.8), we get

$$\begin{aligned} 0 &= \frac{-nE_v}{2G} + l_v - \frac{E_v l}{2E} \\ l_v &= \frac{E_v}{2} \left[\frac{l}{E} + \frac{n}{G} \right] \\ &= \frac{E_v}{2}(k_1 + k_2). \end{aligned}$$

Similarly, using the U-term of $\mathbf{x}_{vvu} - \mathbf{x}_{vuv} = 0$ we get

$$n_u = \frac{G_u}{2} \left[\frac{l}{E} + \frac{n}{G} \right] = \frac{G_u}{2}(k_1 + k_2).$$

If we differentiate the formulas $k_1 = \frac{l}{E}$ and $k_2 = \frac{n}{G}$ repeatedly and use the expressions for l_v and n_u we get,

$$\begin{aligned} k_{1v} &= \frac{1}{E^2}(El_v - E_v l) = \frac{E_v}{2E}(k_2 - k_1) \\ k_{2u} &= \frac{1}{G^2}(Gn_u - G_u n) = \frac{G_u}{2G}(k_1 - k_2) \\ k_{1vv} &= \frac{E_{vv}}{2E}(k_2 - k_1) + E_v(\ldots) \\ k_{2uu} &= \frac{G_{uu}}{2G}(k_1 - k_2) + G_u(\ldots). \end{aligned}$$

Now, at p, k_1 is a max and k_2 is a min, so

$$\begin{aligned} k_{1v} &= 0 \qquad & k_{1vv} &\le 0 \\ k_{2u} &= 0 \qquad & k_{2uu} &\ge 0. \end{aligned}$$

Plugging these estimates into the equations above we get, at p,

$$\begin{aligned} E_v &= 0 \qquad & E_{vv} &\ge 0 \\ G_u &= 0 \qquad & G_{uu} &\ge 0. \end{aligned}$$

Note that, even though $E_v = 0 = G_u$ at p, the expressions

$$\frac{\partial}{\partial v}\left(\frac{E_v}{\sqrt{EG}}\right) \quad \text{and} \quad \frac{\partial}{\partial u}\left(\frac{G_u}{\sqrt{EG}}\right)$$

may be nonzero at p since the differentiation is done before the evaluation at p. These expressions then give us, at p,

$$K = -\frac{1}{2EG}\left(E_{vv} + G_{uu}\right) \leq 0.$$

But this contradicts our hypothesis that $K > 0$, so our assumption $k_1 > k_2$ must be incorrect. Hence Case 1 holds and M is a sphere. □

Notice that in the course of the proof we have discovered a famous result of David Hilbert (which will prove useful in Exercise 3.5.10).

Lemma 3.5.6 (Hilbert). *If k_1 has a* max *at p, k_2 has a* min *at p and k_1 is strictly greater than k_2 at p, then $K(p) \leq 0$.*

So far, we have concentrated on the effects of Gauss curvature on geometry. We now balance this somewhat by giving a result which indicates the effect mean curvature has on the geometry of a surface. Recall that a surface is *minimal* if the mean curvature is zero at each point of the surface. Just as for other geometric qualities, we may test the notion of minimality on classes of surfaces which we understand best. In Chapter 4 we shall consider other types of minimal surfaces, but for now let's look at the surfaces of revolution.

Theorem 3.5.7. *If a surface of revolution M is minimal, then M is contained in either a plane or a catenoid.*

Proof. Consider the profile curve of M, $\alpha(u) = (g(u), h(u))$. If $g(u)$ is a constant function, then M is a plane. If $g'(u) \neq 0$ at a point, then it is nonzero in an open interval around that point. In that interval then, $g(u)$ has an inverse function, so we could reparametrize the interval of definition so that $g(u)$ is replaced by a new variable w and the profile curve becomes $\alpha(w) = (w, h(w))$. Now, we must have one case or the other throughout the domain of M because $\alpha' = h'/g'$. Therefore, we shall concentrate on this case in what follows. Take a patch $\mathbf{x}(u, v) = (u,\ h(u)\cos v,\ h(u)\sin v)$ for M. Then

$$H = \frac{k_\mu + k_\pi}{2} = \frac{1}{2}\left(\frac{-h''}{(1+h'^2)^{3/2}} + \frac{1}{h(1+h'^2)^{1/2}}\right) = \frac{1}{2}\left(\frac{-hh'' + 1 + h'^2}{h(1+h'^2)^{3/2}}\right).$$

The surface M is minimal, so $H = 0$ and, thus, $hh'' = 1 + h'^2$. Let $w = h'$. Then

$$h'' = w' = \frac{dw}{dh}\frac{dh}{du} = \frac{dw}{dh}\,w.$$

Here we consider w as a function of h on an interval with $h'(u) \neq 0$. We can do this because, on such an interval, h has an inverse function f with $u = f(h)$. Then, taking a derivative and applying the chain rule gives

$$1 = \frac{df}{dh}\frac{dh}{du} = \frac{df}{dh}w \qquad \text{or} \qquad w = \frac{1}{\frac{df}{dh}},$$

so that w is a function of h since f and df/dh are. Thus, $hh'' = 1 + h'^2$ implies $hw\,(dw/dh) = 1 + w^2$; or

$$\frac{w}{1+w^2}dw = \frac{1}{h}dh.$$

After integration, we get,

$$\begin{aligned} \ln(\sqrt{1+w^2}) &= \ln h + C \\ \sqrt{1+w^2} &= ch \qquad \text{where } c = e^C, \\ w &= \pm\sqrt{c^2h^2-1}. \end{aligned}$$

Now, $w = \frac{dh}{du}$, so $(dh)/(\sqrt{c^2h^2-1}) = \pm du$ and integration yields (with $ch = \cosh l$, $l \geq 0$)

$$\begin{aligned} \frac{1}{c}l &= \pm u + D \\ \frac{1}{c}\cosh^{-1}(ch) &= \pm u + D \\ h &= \frac{1}{c}\cosh(cu \pm D). \end{aligned}$$

Therefore, M is part of a catenoid. □

Exercise 3.5.8. In the proof above we assumed that $h'(u) \neq 0$. What happens if $h'(u) = 0$? Explain.

Exercise* 3.5.9. Show that a flat ($K = 0$) surface of revolution is part of a cone or a cylinder. Hint: Model your proof after the theorem above. Then see Exercise 4.9.1.

Exercise 3.5.10. Show that if M is a compact, connected (oriented) surface with $K > 0$ and H constant, then M is a sphere of radius $\frac{1}{\sqrt{K}}$.
Hints: (1) Consider the function $H^2 - K = (k_1 - k_2)^2/4$ and let $p \in M$ be a max for it. (2) If $H^2 - K(p) = 0$, show K is constant and use Liebmann's Theorem (Theorem 3.5.5). (3) Suppose $H^2 - K(p) > 0$ (so $k_1(p) > k_2(p)$). Note $(k_1 - k_2)(p)$ is a max. (4) Use Hilbert's lemma, Lemma 3.5.6.

3.6 Surfaces of Delaunay

Just as we saw that surfaces of revolution of constant Gauss curvature are determined by certain integrals with chosen initial conditions, we can ask whether something similar occurs for constant mean curvature. Let's take a surface of revolution M: $\mathbf{x}(u, v)$, where the patch has the form

data 17-04-2020
S T Q Q [✓ S] S D

C3, to draw a circle touching all three.

Fig. 27.

D) Books_69_2020/
Differential Geometry and Its Applications/
Differential Geometry and its applications...pdf

Pag. 416(442)

8.5 Curvature

In Section 3.4 we developed for Gauss curvature which depended only on the metric.

Section. 3.4

Pag. 123(145)

3.4 A Formula for Gauss Curvature.

Obs.: Pag. 122(144)

Figure 3.5 A pseudosphere.

IMPORTANT:

The pseudosphere is actually determined by the requirement of constant curvature and certain initial

date 17.04.2020

conditions. Suppose we start off by
requiring that a surface of revolution
M have constant Gauss curvature
K = -1.

E) Books_69_2020/
Differential Geometry and Its
Applications/
Lokenath... Introduction to
Hilbert Spaces....pdf.

F) Books_69_2020/
Nature and Numbers A Mathematical
Photo Shooting/
www...Nature-and-Numbers.pdf

obs: Estudar com atenção!

G) Books_69_2020/
Topology Illustrated/
Sabet.Topology Illustrated...pdf

obs: Estudar em detalhes!!!

a) Pág. 452 (459)
4.2 The Euler-Poincaré Formula

b) Pág. 541 (548)
4. Metric Tensor
4.1 The need for measuring in calculus

/012/ / /3 / ((= . K - % "
F - * 0** @ > " 6" @ 6 5 0 " *"

/012/ / /3 / ((7 < "<7 - *"
3- 6 " UT AAA

/012/ / /3 / ((; E"G 5 . @ 0 % "
% " >* + " F *+ (@ 1 *"

/012/ / /3 / ((>* + % "
5 5*"
' (# (! (> 6 M :

4.2 The Euler-Poincaré Formula

The Euler characteristic of a cell complex is computed from this data:

$$c_k(K) := \dim C_k(K).$$

Notice that these numbers resemble (and in the simplest case are equal to) the Betti numbers of the cell complex:

$$\beta_k(K) := \dim H_k(K).$$

What if we evaluate an alternating sum of the latter numbers rather than the former?

Implicitly, we have already done this, in the Euler Formula. Indeed, we have:

$$\chi(\mathbf{S}^2) = 2,$$

which holds regardless of the triangulation. On the other, the alternating sum of the Betti numbers of the sphere

$$\beta_0(\mathbf{S}^2) = 1,\ \beta_1(\mathbf{S}^2) = 0,\ \beta_2(\mathbf{S}^2) = 1,\ \beta_3(\mathbf{S}^2) = 0, ...$$

also produces 2!

Is this a coincidence? Let's check out other complexes.

Example 4.11 (point). First, the point P:

$$\begin{aligned}&\chi(P) = 1 - 0 = 1,\\ &\beta_0(P) - \beta_1(P) = 1 - 0 = 0.\end{aligned}$$

Once again, the Betti numbers add up to the Euler number. $\square$

Example 4.12 (circle). Next, the circle C:

$$\begin{aligned}&\chi(C) = 1 - 1 + 0 = 0,\\ &\beta_0(C) - \beta_1(C) + \beta_2(C) = 1 - 1 + 0 = 0.\end{aligned}$$ $\square$

Example 4.13 (torus). Now the torus $\mathbf{T}^2$. We have:

- 0-cells: 1,
- 1-cells: 2,

For each $i = 0, 1, ..., n$, we have

$$\begin{array}{ll} < C_i(K) \cup C_i(L) > & = C_i(K \cup L), \\ C_i(K) \cap C_i(L) & = C_i(K \cap L). \end{array}$$

Then the alternating sum of the above equations gives us the formula. ■

The Euler characteristic behaves well under products.

Theorem 4.8 (Product Formula for Euler Characteristic). The Euler characteristic is multiplicative; i.e., for two finite cell complexes K and L, we have:

$$\chi(K \times L) = \chi(K) \cdot \chi(L).$$

Exercise 4.9. Prove the theorem. Hint: $\dim(U \oplus V) = \dim U + \dim V$.

Exercise 4.10. Compute the Euler characteristic of the n-dimensional torus $\mathbf{T}^n$.

4.2 The Euler-Poincaré Formula

The Euler characteristic of a cell complex is computed from this data:

$$c_k(K) := \dim C_k(K).$$

Notice that these numbers resemble (and in the simplest case are equal to) the Betti numbers of the cell complex:

$$\beta_k(K) := \dim H_k(K).$$

What if we evaluate an alternating sum of the latter numbers rather than the former?

Implicitly, we have already done this, in the Euler Formula. Indeed, we have:

$$\chi(\mathbf{S}^2) = 2,$$

which holds regardless of the triangulation. On the other, the alternating sum of the Betti numbers of the sphere

$$\beta_0(\mathbf{S}^2) = 1,\ \beta_1(\mathbf{S}^2) = 0,\ \beta_2(\mathbf{S}^2) = 1,\ \beta_3(\mathbf{S}^2) = 0, ...$$

also produces 2!

Is this a coincidence? Let's check out other complexes.

Example 4.11 (point). First, the point P:

$$\begin{array}{l} \chi(P) = 1 - 0 = 1, \\ \beta_0(P) - \beta_1(P) = 1 - 0 = 0. \end{array}$$

Once again, the Betti numbers add up to the Euler number. □

Example 4.12 (circle). Next, the circle C:

$$\begin{array}{l} \chi(C) = 1 - 1 + 0 = 0, \\ \beta_0(C) - \beta_1(C) + \beta_2(C) = 1 - 1 + 0 = 0. \end{array}$$ □

Example 4.13 (torus). Now the torus $\mathbf{T}^2$. We have:

- 0-cells: 1,
- 1-cells: 2,

- 2-cells: 1.

So,

$$\chi(\mathbf{T}^2) = 1 - 2 + 1 = 0.$$

Because this cell complex representation is so efficient, each of the cells also represents an element of a basis element of the corresponding homology group:

$$\begin{array}{ll} H_0(\mathbf{T}^2) = < A > & \Longrightarrow \beta_0(\mathbf{T}^2) = 1, \\ H_1(\mathbf{T}^2) = < a, b > & \Longrightarrow \beta_1(\mathbf{T}^2) = 2, \\ H_2(\mathbf{T}^2) = < \tau > & \Longrightarrow \beta_2(\mathbf{T}^2) = 1, \\ H_3(\mathbf{T}^2) = 0 & \Longrightarrow \beta_3(\mathbf{T}^2) = 0. \end{array}$$

The formula holds:

$$\chi(\mathbf{T}^2) = \beta_0(C) - \beta_1(C) + \beta_2(C) - ... \qquad \square$$

As we have been doing nothing but counting the cells, it's no surprise that the formula holds. Undoubtedly, it will always hold when there is *one cell per homology generator.* We need to show that this formula holds for all cell complexes.

Theorem 4.14 (Euler-Poincaré Formula). If K is a finite cell complex, then its Euler characteristic is equal to the alternating sum of the Betti numbers of each dimension:

$$\chi(K) = \sum_k (-1)^k \beta_k(K).$$

The proof will require a couple of facts from linear algebra of finite dimensional vector spaces.

Proposition 4.15.

- (1) If M, L are vector spaces and $A : M \to L$ is a linear operator, then

$$M/\ker A \simeq \operatorname{Im} A.$$

- (2) If Y is a subspace of vector space X, then

$$\dim X/Y = \dim X - \dim Y.$$

Exercise 4.16. Prove the proposition.

The proposition leads to the following conclusion.

Corollary 4.17. If M, L are vector spaces and $A : M \to L$ is a linear operator, then

$$\dim M - \dim \ker A = \dim \operatorname{Im} A.$$

Now, there are four vector spaces involved in the computation of the Betti numbers of complex K, for each $k = 0, 1, 2, ...$:

$$\begin{array}{lll} \text{chain group:} & C_k, & c_k := \dim C_k; \\ \text{cycle group:} & Z_k = \ker \partial_k, & z_k := \dim Z_k; \\ \text{boundary group:} & B_k = \operatorname{Im} \partial_{k+1}, & b_k := \dim B_k; \\ \text{homology group:} & H_k = Z_k/B_k, & \beta_k := \dim H_k, \end{array}$$

where

$$\partial_k : C_k \to C_{k-1}$$

is the boundary operator.

Applying the corollary to the boundary operator, we obtain:

$$\dim C_k - \dim \ker \partial_k = \dim \operatorname{Im} \partial_k .$$

In other words, we have the following.

Lemma 4.18. $c_k - z_k = b_{k-1}$.

Next, from part (2) of the proposition, we have the following.

Lemma 4.19. $\beta_k = z_k - b_k$.

Finally, the **proof** of the theorem. Suppose n is the highest dimension of a cell in K. Then the right-hand side of the formula is computed as follows:

$$\begin{array}{rllll}
= & \beta_0 & -\beta_1 & +\beta_2 - ... & +(-1)^k\beta_n \\
 & & & \text{substitute from 2nd lemma} & \\
= & (z_0-b_0) & -(z_1-b_1) & +(z_2-b_2) - ... & +(-1)^n(z_n-b_n) \\
 & & & & \\
= & z_0 & -b_0 - z_1 & +b_1+z_2-... & +(-1)^n z_n - (-1)^n b_n \\
 & & & \text{substitute from 1st lemma} & \\
= & z_0 & -(c_1-z_1)-z_1 & +(c_2-z_2)+z_2+... & +(-1)^n z_n - (-1)^n(c_{n+1}-z_{n+1}) \\
 & & & \text{cancel the z's} & \\
= & z_0 & -c_1 & +c_2 - ... & +(-1)^{n+1}c_{n+1} - (-1)^n z_{n+1} \\
= & c_0 & -c_1 & +c_2 - ... & +0 \quad - \quad 0 \\
= & \chi(K). & & &
\end{array}$$

■

The invariance of homology then gives us the following.

Corollary 4.20. The Euler characteristic is a topological invariant of polyhedra.

An interesting discovery but, because so much information is lost, the Euler characteristic is poor man's homology...

Exercise 4.21. Under what circumstances can the homology of X be derived from its Euler characteristic for (a) X a graph, (b) X a surface?

4.3 Fixed points

Imagine stretching a rubber band by moving one end to the right and the other to the left. After some experimenting one might discover that some point of the band will end up in its original position:

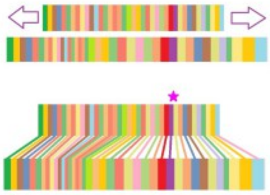

In fact, it doesn't matter if the stretching isn't uniform or if there is some shrinking, or even folding...

This problem is described mathematically as follows.

Fixed Point Problem: If X is a topological space and $f : X \to X$ is a self-map, does f have a *fixed point*, i.e., $x \in X$ such that $f(x) = x$?

The result above is, therefore, about a fixed point of a map $f : [a, b] \to [a, b]$.

Exercise 4.22. Prove this result by using the Intermediate Value Theorem.

Exercise 4.23. Demonstrate that both continuity of the map and the compactness of the interval are crucial.

Exercise 4.24. Under what circumstances is the set of fixed points closed?

A similar problem about stretching a disk or a ball will require more advanced techniques.

Theorem 4.25 (Brouwer Fixed Point Theorem). Any self-map of a closed ball has a fixed point. In other words, if $f : \mathbf{B}^n \to \mathbf{B}^n$ is a map then there is a $x \in \mathbf{B}^n$ such that $f(x) = x$.

An informal interpretation of the 2-dimensional version of this theorem is: a page and its crumpled copy placed on it will have two identical words located one exactly over the other:

Exercise 4.26. Prove the theorem. Hint:

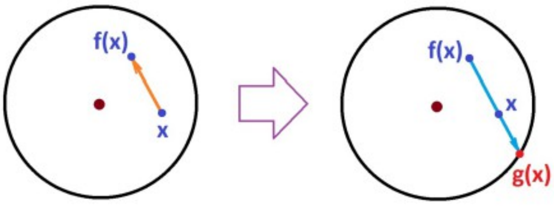

Plainly, the theorem holds if the ball is replaced with anything homeomorphic to it, such a simplex.

Though technically not a fixed point theorem, the following result is closely related to the last theorem.

Theorem 4.27 (Borsuk-Ulam Theorem). Every map $f : \mathbf{S}^n \to \mathbf{R}^n$ from an n-sphere to n-dimensional Euclidean space maps some pair of antipodal points to the same point:

$$f(x) = f(-x).$$

A common interpretation of the case $n = 1$ is that at any moment there is always a pair of antipodal points on the Earth's surface with exactly equal temperatures (or with equal barometric pressures, etc.).

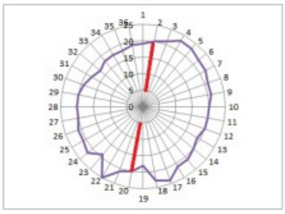

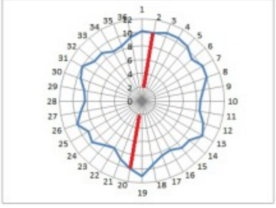

Furthermore, an interpretation for $n=2$ is that there is always a pair of antipodal points with exactly equal temperatures and equal barometric pressures, *simultaneously*.

Exercise 4.28. Explain how the theorem implies that at any moment there is always a pair of antipodal points on the Earth's surface with identical wind.

Exercise 4.29. Derive the case $n=1$ of the theorem from the Intermediate Value Theorem.

Exercise 4.30. Prove the theorem. Hint: consider this map:

$$g(x)=\frac{f(x)-f(-x)}{||f(x)-f(-x)||}.$$

4.4 An equilibrium of a market economy

Suppose we have a map $F:S\to S$. Applied consecutively, it yields a *discrete dynamical system*:

$$X_{k+1}=F(X_k),$$

where $X_k\in S$ is the current state, at time k, of the system if it started at time $k=0$ at the state $X_0\in S$. Then the meaning of the *equilibrium* of the system is simple: $F(a)=a$. It's a fixed point!

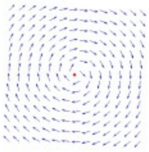

A typical interpretation of such a system is a particle moving through a stream. We will consider instead the *dynamics of prices* in a simple market economy. Is it possible for them to stay constant? In other words, does every price dynamics allow an equilibrium?

Recall that we have n commodities freely traded with possible prices that form a *price vector* $p=(p_1,...,p_n)$, at each moment of time. Then all positive, or even non-negative, prices are considered possible; i.e., the set of prices is

$$P:=\mathbf{R}^n_+.$$

The dynamics of prices is represented by a continuous function $F:P\to P$. However, we can't conclude that there is a fixed point because P isn't compact!

Let's make some additional assumptions in order to be able to guarantee an equilibrium.

How can we make the set of prices compact? One approach is to normalize the prices:

$$p\mapsto\frac{p}{||p||},$$

under some norm so that the new prices form the n-simplex σ.

Exercise 4.31. Show that this is an equivalence relation. What assumptions would guarantee that the price function remains well-defined?

A more natural (if not more realistic) assumption is that the *money supply is fixed.* Suppose there are n commodities. We assume that this is a pure exchange economy; i.e., *there is no production* of goods (other than making up for possible consumption). Then there is only a fixed amount of each of the commodities:

$$\bar{c}:=\sum_i c^i\in\mathbf{R}^n_+.$$

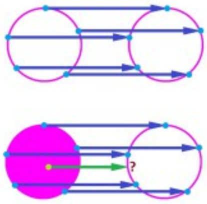

Can we find a continuous F to complete the first diagram below? With the cohomology functor, we answer: only if we can complete the last one.

$$\begin{tikzcd}
\mathbf{S}^{n-1} \arrow[r, hook] \arrow[dr, "\mathrm{Id}"'] & \mathbf{B}^n & & H^{n-1}(\mathbf{S}^{n-1}) & H^{n-1}(\mathbf{B}^n) \arrow[l] & & \mathbf{Z} \arrow[r, "H^*"] & 0 \\
& \mathbf{S}^{n-1} \arrow[u, "F=?"'] & \rightsquigarrow & & H^{n-1}(\mathbf{S}^{n-1}) \arrow[ul, "\mathrm{Id}"] \arrow[u, "?"'] & = & & \mathbf{Z} \arrow[ul, "\mathrm{Id}"] \arrow[u, "?"']
\end{tikzcd}$$

Just as in the case of homology, we see that the last diagram is impossible to complete. ☐

Exercise 3.84. Provide such analysis for the extension problem(s) for: the identity map of the circle to the torus.

Exercise 3.85. List all possible maps for each of these description based on the possible cohomology maps:
- embeddings of the circle into the torus;
- self-maps of the figure eight;
- embeddings of the circle into the sphere.

4 Metric tensor

4.1 The need for measuring in calculus

We will learn how to introduce an arbitrary geometry into the (so far) purely topological setting of cell complexes.

It is much easier to separate topology from geometry in the framework of cell complexes (and discrete calculus) because we have built it from scratch! To begin with, it is obvious that only the way the cells are attached to each other affects the matrix of the boundary operator (and the exterior derivative):

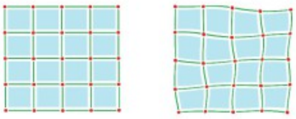

It is then clear that the sizes or the shapes of the cells are topologically irrelevant. Note that, even when those are given, the geometry of the domain will remain unspecified unless the angles are also provided:

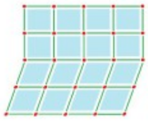

These two grids can be thought of as different (but homeomorphic) realizations of the same cell complex.

In calculus, we rely only on the topological and algebraic structures to develop the mathematics. There are two main exceptions. One needs the dot product and the norm for the following related concepts:

- the concavity,
- the arc-length, and
- the curvature. We'll need to introduce more structure – geometric in nature – to the complex.

Of course, the norm is explicitly present in the formulas for the last two. But first we will consider the role of geometry in the meaning of concavity.

The point where we start to need geometry is when we move from the first to the second derivative.

This is what we observe: shrinking, stretching, or deforming the x- or the y-axes won't change the monotonicity of a function but it will change its concavity. Below, we have deformed the x-axis and changed the concavity of the graph at $x = 1$:

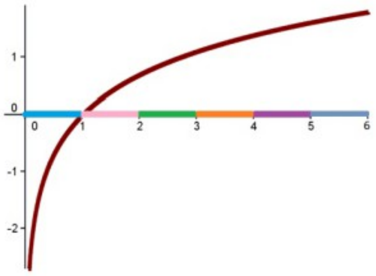

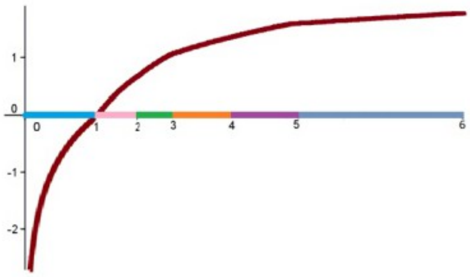

Of course, it is sufficient to know the sign of the derivative to distinguish between increasing and decreasing behavior. Therefore, this behavior depends only on the topology of the spaces, in the above sense. In fact, the chain rule implies the following elementary result.

Proposition 4.1. Monotonicity is preserved under compositions with increasing functions; i.e., given a differentiable function $f : \mathbf{R} \to \mathbf{R}$, we have:

$$f \nearrow \iff fr \nearrow,$$

for any differentiable $r : \mathbf{R} \to \mathbf{R}$ with $r' > 0$.

Exercise 4.2. Give an example of f and r as above so that f and fr have different concavity.

Exercise 4.3. Restate the proposition in the context of discrete calculus.

Similarly, in the discrete calculus we are developing, the sign of the exterior derivative will tell us the difference between increasing and decreasing behavior. Concavity is an important concept that captures the shape of the graph of a function in calculus as well as the acceleration in physics. The upward concavity of the discrete function below is obvious, which indicates a positive acceleration:

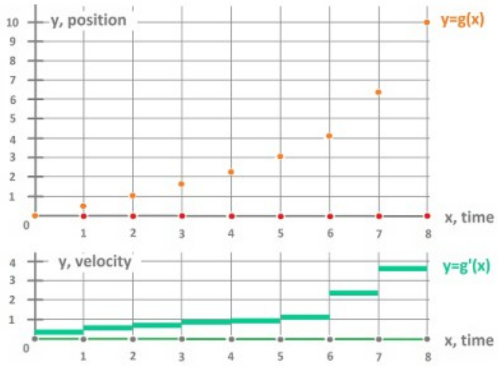

The concavity is computed and studied via the second derivative, the rate of change of the rate of change. But, with $dd=0$, we don't have the second *exterior* derivative...

We start over and look at the *change of change*. This quantity can be seen algebraically as the increase of the exterior derivative (i.e., the change):

$$g(2)-g(1) < g(3)-g(2),\ g(3)-g(2) < g(4)-g(3),\ etc.$$

But do these inequalities imply concavity? Not without assuming that the intervals have equal lengths!

Below, on the same, topologically, cell complex, we have the same discrete function with, of course, the same exterior derivative:

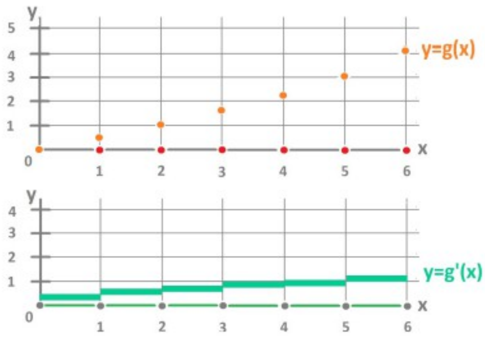

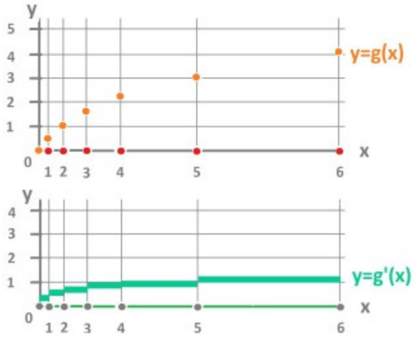

...but with the opposite concavity!

4.2 Inner products

In order to develop a complete discrete calculus we need to be able compute:

- the lengths of vectors and
- the angles between vectors.

In linear algebra, we learn how an inner product adds geometry to a vector space. We choose a more general setting.

Suppose R is our ring of coefficients and suppose V is a finitely generated free module over R.

Definition 4.4. An *inner product* on V is a function that associates a number to each pair of vectors in V:

$$\langle\cdot,\cdot\rangle : \begin{cases} V\times V \to R, \\ (u,v)\mapsto \langle u,v\rangle, \end{cases}$$

that satisfies these properties:

- 1. $\diamond$ $\langle v,v\rangle \geq 0$ for any $v \in V$ – *non-degeneracy*,
  $\diamond$ $\langle v,v\rangle = 0$ if and only if $v=0$ – *positive definiteness*;
- 2. $\langle u,v\rangle = \langle v,u\rangle$ – *symmetry* (commutativity);
- 3. $\langle ru,v\rangle = r\langle u,v\rangle$ – *homogeneity*;
- 4. $\langle u+u',v\rangle = \langle u,v\rangle + \langle u',v\rangle$ – *distributivity*.

Items 3 and 4 together make up *bilinearity*. Indeed, consider $p : V \times V \to R$ given by $p(u,v) = \langle u,v\rangle$. Then p is linear with respect to the first variable, and the second variable, separately:

- fix $v=b$, then $p(\cdot,b) : V \to R$ is linear;
- fix $u=a$, then $p(a,\cdot) : V \to R$ is linear.

Exercise 4.5. Show that this isn't the same as linearity.

It is easy to verify these axioms for the *dot product* defined on $V = \mathbf{R}^n$ or any other module with a fixed basis. For

$$u = (u_1, ..., u_n),\ v = (v_1, ..., v_n) \in \mathbf{R}^n,$$

define

$$\langle u,v\rangle := u_1v_1 + u_2v_2 + ... + u_nv_n.$$

Moreover, a *weighted dot product* on $\mathbf{R}^n$ is given by

$$\langle u,v\rangle := w_1u_1v_1 + w_2u_2v_2 + ... + w_nu_nv_n,$$

where $w_i \in \mathbf{R},\ i = 1, ..., n$, are the positive "weights", is also an inner product.

Exercise 4.6. Prove the last statement.

A module equipped with an inner product is called an *inner product space.*

When ring R is a subring of the reals $\mathbf{R}$, such as the integers $\mathbf{Z}$, we can conduct some useful computations with some familiar formulas.

Definition 4.7. The *norm of a vector* v in an inner product space V is defined as

$$\|v\| := \sqrt{\langle v,v\rangle} \in \mathbf{R}.$$

This number measures the length of the vector.

Definition 4.8. The *angle between vectors* $u,v \neq 0$ in V is defined as

$$\cos\widehat{uv} := \frac{\langle u,v\rangle}{\|u\| \cdot \|v\|}.$$

The norm satisfies certain properties that we can also use as *axioms of a normed space.*

Definition 4.9. Given a vector space V, a *norm* on V is a function

$$\|\cdot\| : V \to R$$

that satisfies

- 1. $\diamond$ $\|v\| \geq 0$ for all $v \in V$,
  $\diamond$ $\|v\| = 0$ if and only if $v = 0$;
- 2. $\|rv\| = |r|\|v\|$ for all $v \in V, r \in R$;
- 3. $\|u+v\| \leq \|u\| + \|v\|$ for all $u,v \in V$.

Exercise 4.10. Prove the propositions below.

Proposition 4.11. A normed space is a metric space with the metric $d(x,y) := \|x-y\|$.

Proposition 4.12. A normed space is a topological vector space.

Exercise 4.13. Prove that the inner product is continuous in this metric space.

This is what we know from linear algebra.

Theorem 4.14. Any inner product $\langle \cdot , \cdot \rangle$ on an n-dimensional module V can be computed via matrix multiplication

$$\langle u, v \rangle = u^T Q v,$$

where Q is some positive definite, symmetric $n \times n$ matrix.

In particular, the dot product is represented by the identity matrix I_n, while the weighted dot product is represented by the diagonal matrix:

$$Q = \left(\begin{array}{ccc} w_1 & ... & 0 \\ ... & ... & ... \\ 0 & ... & w_n \end{array} \right).$$

Below we assume that $R = \mathbf{R}$.

To learn about the eigenvalues and eigenvectors of matrix Q of the inner product, we observe that

$$||v||^2 = v^T Q v = v^T \lambda v = \lambda ||v||^2.$$

We have proven the following:

Theorem 4.15. The eigenvalues of a positive-definite matrix are real and positive.

We know from linear algebra that if the eigenvalues are real and *distinct*, the matrix is diagonalizable. As it turns out, having distinct eigenvalues isn't required for a positive-definite matrix. We state the following without proof.

Theorem 4.16. Any inner product $\langle \cdot , \cdot \rangle$ on an n-dimensional vector space V can be represented, via a choice of basis, by a diagonal $n \times n$ matrix Q with positive numbers on the diagonal. In other words, every inner product is a weighted dot product, in some basis.

Also, since the determinant $\det Q$ of Q is the product of the eigenvalues of Q, it is also positive and we have the following:

Theorem 4.17. Matrix Q that represents an inner product is invertible.

4.3 The metric tensor

What if the geometry – as described above – varies from point to point in some space X? Then the inner product of two vectors will depend on their location.

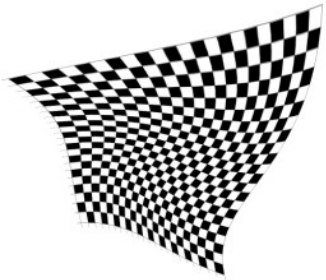

In fact, we'll need to look at a different set of vectors at each location as the angle between two vectors is meaningless unless they have the same origin.

We already have such a construction! For each vertex A in a cell complex K, the *tangent space* at A of K is a submodule of $C_1(K)$ generated by the 1-dimensional star of A:

$$T_A(K) := < \{AB \in K\} > \subset C_1(K).$$

However, the tangent bundle $T(K)$ is inadequate because this time we need to consider *two directions* at a time!

Since $V = T_A$ is a module, an inner product can be defined (so far locally). Then we have a collection of bilinear functions for each of the vertices,

$$\psi_A : T_A(K)^2 \to R, \ A \in K^{(0)}.$$

Thus the setup is as follows. We have the *space of locations* $X = K^{(0)}$, the set of all vertices of a cell complex K, and, second, to each location $A \in X$, we associate the *space of pairs of directions* $T_A(K)^2$.

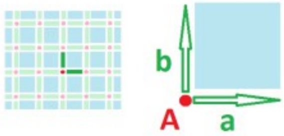

We now collect these "double" tangent spaces into one.

Definition 4.18. The *tangent bundle of degree* 2 of K is defined to be

$$T^2(K) := \bigsqcup_{A \in K} \Big(\{A\} \times T_A(K)^2 \Big).$$

Then a locally defined inner product is seen as a function on this space,

$$\psi = \{\psi_A\} : T^2(K) \to R,$$

bilinear on each tangent space, defined by

$$\psi(A, AB, AC) := \psi_A(AB, AC).$$

In other words, we have inner products "parametrized" by location.

Furthermore, we have a function that associates to each location, i.e., a vertex A in K, and each pair of directions at that location, i.e., edges $x = AB, y = AC$ adjacent to A with $B, C \neq A$, a number $\langle AB, AC \rangle$:

$$(A, AB, AC) \mapsto \langle AB, AC \rangle (A) \in R.$$

Note: we have used this notation for $f(a) = \langle f, a \rangle$, where f a form, but won't use in the rest of the book.

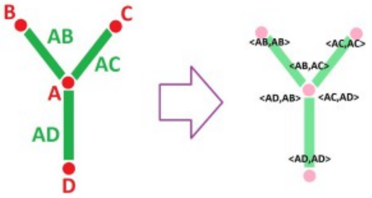

Definition 4.19. A *metric tensor* on a cell complex K is a collection of functions,

$$\psi = \{\psi_A : \ A \in X\},$$

such that each of them,

$$\psi_A : \{A\} \times T_A(K)^2 \to R,$$

satisfies the axioms of inner product, and also:

$$\psi_A(AB, AB) = \psi_B(BA, BA).$$

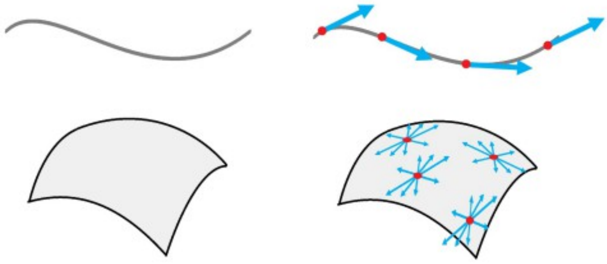

Using our notation, a metric tensor is such a function:

$$\langle \cdot , \cdot \rangle(\cdot) : T^2(K) \to R,$$

that, when restricted to any tangent space $T_A(K), \ A \in K^{(0)}$, it produces a function

$$\langle \cdot , \cdot \rangle(A) : T_A(K)^2 \to R,$$

which is an inner product on $T_A(K)$. The last condition,

$$\langle AB, AB \rangle(A) = \langle BA, BA \rangle(B),$$

ensures that these inner products match!

Note: Compare to the standard setup: the location space X is an n-manifold; then $T_AX \cong \mathbf{R}^n$ is its tangent space, when $R = \mathbf{R}$. Then we have a function that associates a number to each location in X and each pair of vectors at that location:

$$\langle \cdot , \cdot \rangle(\cdot) : \begin{cases} X \times \mathbf{R}^n \times \mathbf{R}^n & \to \mathbf{R}, \\ (x, u, v) & \mapsto \langle u, v \rangle(x). \end{cases}$$

It is an inner product that depends continuously on location $x \in X$.

Exercise 4.20. State the axioms of inner product (non-degeneracy, positive-definiteness, symmetry, homogeneity, and distributivity) for the metric tensor.

The end result is similar to a combination of 1-forms, because we have a number assigned to each edge and to each pair of adjacent edges:

- $a \mapsto \langle a, a \rangle$;
- $(a, b) \mapsto \langle a, b \rangle$.

Then the metric tensor is given by the table of its values for each vertex A:

A	AB	AC	AD	...
AB	$\langle AB, AB \rangle$	$\langle AB, AC \rangle$	$\langle AB, AD \rangle$	...
AC	$\langle AC, AB \rangle$	$\langle AC, AC \rangle$	$\langle AC, AD \rangle$	...
AD	$\langle AD, AB \rangle$	$\langle AD, AC \rangle$	$\langle AD, AD \rangle$	...
...	...	...	...	...

Whenever our ring of coefficients happens to be a subring of the reals $\mathbf{R}$, such as the integers $\mathbf{Z}$, this data can be used to extract more usable information:

- $a \mapsto \|a\|$, and
- $(a, b) \mapsto \widehat{ab}$,

where

$$\cos \widehat{ab} = \frac{\langle a, b\rangle}{\|a\|\|b\|}.$$

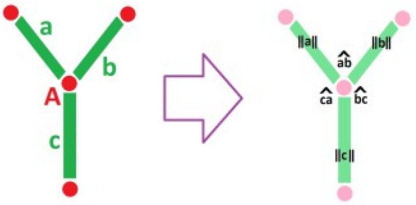

All the information we need for measuring is contained in this symmetric matrix:

A	AB	AC	AD	...
AB	$\|AB\|$	$\widehat{BAC}$	$\widehat{BAD}$	...
AC	$\widehat{BAC}$	$\|AC\|$	$\widehat{CAD}$	...
AD	$\widehat{BAD}$	$\widehat{CAD}$	$\|AD\|$	...
...	...	...	...	...

Exercise 4.21. In the 1-dimensional cubical complex $\mathbb{R}$, are 01 and 10 parallel?

Proposition 4.22. From any matrix with positive entries on the diagonal, we can (re)construct a metric tensor by the formula:

$$\langle AB, AC\rangle = \|AB\|\|AC\| \cos \widehat{BAC}.$$

With a metric tensor, we can define the two main geometric quantities of calculus in a very easy fashion. Suppose we have a *discrete curve* C in the complex K, i.e., a sequence of adjacent vertices and edges

$$C = \{A_0A_1, A_1A_2, ..., A_{N-1}A_N\} \subset K.$$

Definition 4.23. The *arc-length* of curve C is the sum of the lengths of its edges:

$$l_C := \|A_0A_1\| + \|A_1A_2\| + ... + \|A_{N-1}A_N\|.$$

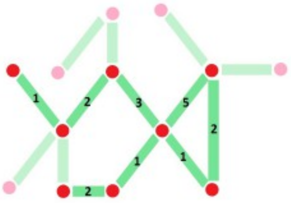

The arc-length is an example of a *line integral* of a 1-form ρ over a 1-chain a in complex K equipped with a metric tensor:

$$\begin{aligned} a = \quad & A_0A_1 + A_1A_2 + ... + A_{n-1}A_n \Longrightarrow \\ & \int_a \rho := \rho(A_0, A_0A_1)\|A_0A_1\| + ... + \rho(A_{N-1}, A_{N-1}A_N)\|A_{N-1}A_N\|. \end{aligned}$$

In particular, if we have a curve made of rods represented by edges in our complex while ρ represents the linear density of each rod, the integral gives us the *weight* of the curve.

For the curvature, the angle $\alpha := A_{s-1}\widehat{A_s}A_{s+1}$ might appear a natural choice to represent it, but, since we are to capture the change of the direction, it should be the "outer" angle β:

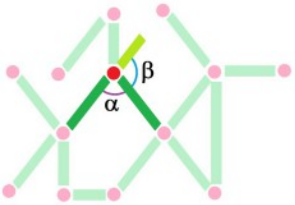

Definition 4.24. The *curvature* of curve C at a given vertex $A_s,\ 0 < s < N$, is the value of the angle

$$\kappa_C(A_s) := \pi - A_{s-1}\widehat{A_s}A_{s+1}.$$

As before, the result depends on our choice of a metric tensor.

The following is a simple but important observation.

Theorem 4.25. The set of vertices $K^{(0)}$ of a cell complex K equipped with a metric tensor is a metric space with its metric given by:

$$d(A,B) = \min\{l_C : C \text{ a curve from } A \text{ to } B\},$$

for any two vertices A, B.

Exercise 4.26. (a) Prove the theorem. (b) What is the relation between the topology of $|K|$ and this metric space?

4.4 Metric tensors in dimension 1

Let's consider examples of how adding a metric tensor turns a cell complex, which is a discrete representation of a topological space, into something more *rigid*.

Example 4.27. In the illustration below you see the following:

- a cubical complex K, a *topological* entity: cells and their boundaries also made of cells;
- complex K combined with a choice of metric tensor $\langle\cdot,\cdot\rangle$, a *geometric* entity: lengths and angles;
- a realization of K in $\mathbf{R}$ as a segment;
- a realization of K in $\mathbf{R}^2$;
- another realization of K in $\mathbf{R}^2$.

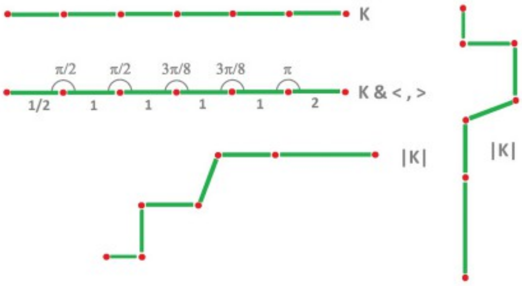

Exercise 1.6. Two polyhedra are called *dual* if the number of faces of one is equal to the number of vertices of the other (and vice versa). Find the dual polyhedra of Platonic solids and explain the relation to the Hodge duality:

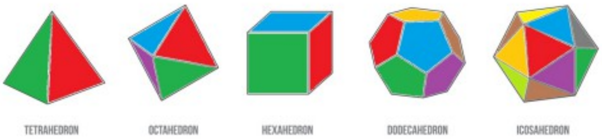

This is what the Hodge duality looks like in the case of a general 2-dimensional cell complex. The dual cell is still placed – for the sake of illustration – at the center of mass of the primal cell:

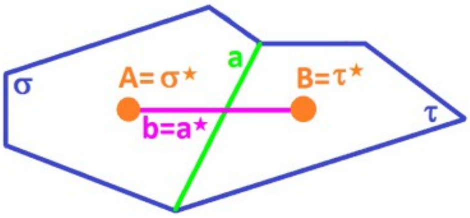

Exercise 1.7. Suppose a simplicial complex K is realized in $\mathbf{R}^n$. Describe explicitly how the dual $K^\star$ can also be realized in $\mathbf{R}^n$ side by side with $|K|$. Hint: use barycentric coordinates.

We summarize the new concept as follows.

Definition 1.8. Suppose a *dimension* n is given. Two cell complexes K and $K^\star$ with $\dim K \le n$, $\dim K^\star \le n$, are called *Hodge-dual* of each other if there is a one-to-one correspondence,

$$K \ni a \longleftrightarrow a^\star \in K^\star,$$

between their cells of complementary dimensions:

$$\dim a^\star + \dim a = n.$$

1.2 Metric complexes

We have already seen metric complexes of dimension $n = 1$ as cell complexes equipped with a geometric structure: the metric tensor of K and $K^\star$. This data allows one to find the lengths and the curvatures of curves. The Hodge star operator is then built from this data.

In this subsection, we present the necessary geometric data for the case of arbitrary dimension.

In dimension $n = 2$, we consider lengths and areas; in dimension $n = 3$, lengths, areas, and volumes, etc. In general, we need to find the "sizes" of all cells of all dimensions in the complex. These sizes are called the *k-dimensional volumes*:

dimension k	k-cell a	k-volume $\lvert a\rvert$
0	vertex	1
1	edge	length
2	face	area
3	solid	volume
...		...
n	n-cell	n-volume

Instead of trying to *derive* these quantities from the metric tensor, we simply *list* them as a part of the definition of a metric complex. In the abstract, we simply have a pair of positive numbers $|b|, |b^\star|$ assigned to each cell $b \in K$.

Definition 1.9. Given a cell complex K, the *volume function* on K defines the k-volumes of the k-cells of K:

$$|\cdot| : K \to R \setminus \{0\},$$

normalized for dimension 0:

$$|A| = 1 \text{ for all vertices } A \in K.$$

For $R = \mathbf{R}$, the definition matches the one for the case of $n = 1$ considered previously if we limit ourselves to positive volumes.

Example 1.10. The simplest choice of the volumes for the 2-cell representation of the sphere is

$$A^0, a^2 \text{ with } |A^0| = |a^2| = 1. \qquad \square$$

In general, we don't assume that $|AB|, |BC|, |CA|$ determine $|ABC|$. Even though the three sides of a triangle determine the triangle and, therefore, its area, this only works under a particular choice of geometry, such as Euclidean. We take the opposite approach and let the lengths and areas determine the geometry.

Exercise 1.11. Give a metric complex representation of a triangulation of the sphere, realized as the unit sphere in $\mathbf{R}^3$.

Example 1.12 (measure). Suppose K is a cell complex with finitely many cells at each vertex and suppose $|K|$ is its realization. Then the set of all cells (open or closed) Σ of K as subsets of $|K|$ form a σ-algebra, i.e., a collection closed under the operations of complement, countable union, and countable intersection. If, furthermore, K has a volume function, the function $\mu : \Sigma \to \mathbf{R} \cup \{\infty\}$ given by

$$\mu(a) = \begin{cases} |a| & \text{if } \dim a = \dim K, \\ 0 & \text{if } \dim a < \dim K, \end{cases}$$

is a *measure* on $|K|$; i.e., it satisfies:

- non-negativity: for all $a \in \Sigma$ we have $\mu(a) \geq 0$;
- null empty set: $\mu(\emptyset) = 0$;
- σ-additivity: for all countable collections $\{a_i\}$ of pairwise disjoint sets in Σ we have:

$$\mu\Big(\bigcup_{k=1}^{\infty} a_k \Big) = \sum_{k=1}^{\infty} \mu(a_k). \qquad \square$$

In addition to the volumes, the *angles* between every pair of dual cells will be taken into account in the new representation of the geometry.

We don't need to specify the meaning of the angle between *every choice* of a primal k-cell and its dual $(n-k)$-cell, for all k. For our purposes, it suffices to take care of the case $k = 1$. We are then in a familiar situation because the angle α between 1- and $(n-1)$-dimensional subspaces of an n-dimensional inner product space is always understood as the angle between the direction vector of the former and the normal vector of the latter.

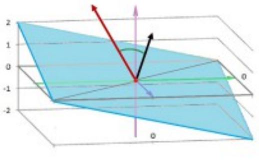

They are orthogonal when this angle is zero. It is also assumed that (for $k = 0$) a point and the whole space are orthogonal and $\cos\alpha = 1$. This assumption explains why these angles did not have to come into play in the case of a 1-dimensional complex.

For cells in the higher dimensional case, we are looking at the angle between a and the normal $b^\perp$ of $b = a^\star$, illustrated for dimension $n = 2$:

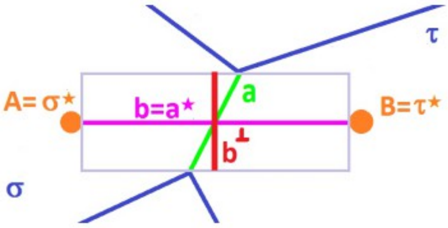

Dually, this angle can also be understood as the angle between the normal $c = a^\perp$ of a and $a^\star$, illustrated for dimension $n = 3$:

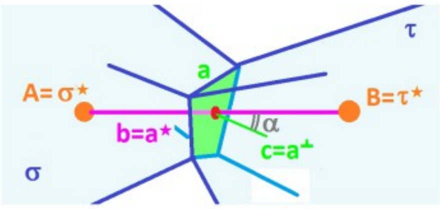

Even though these angles don't affect the volumes of either primal or dual cells, they do affect the volumes of *new* n-cells, the cells determined by each dual pair:

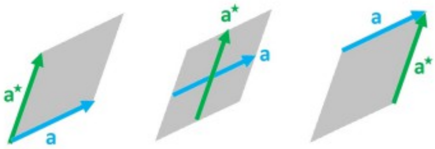

The area of this parallelogram does not depend on how a and $a^\star$ are attached to each other (they aren't) but only on the angle between them.

The way we handle angles is, of course, via the inner product: $\langle a, b\rangle \in R$ is defined for these three cases:

- adjacent $a, b \in K$,
- adjacent $a, b \in K^\star$, and now also
- $a \in K, a^\star \in K^\star$.

Definition 1.13. A *metric (cell) complex of dimension* n is a pair of Hodge-dual cell complexes K and $K^\star$ of dimension n equipped with a *joint metric tensor*, which is a combination of metric tensors for K and $K^\star$ as well as inner products of all dual pairs:

$$(a, a^\star) \mapsto \langle a, a^\star\rangle \in R.$$

Exercise 1.14. Represent a hollow cube as a metric complex. Hint:

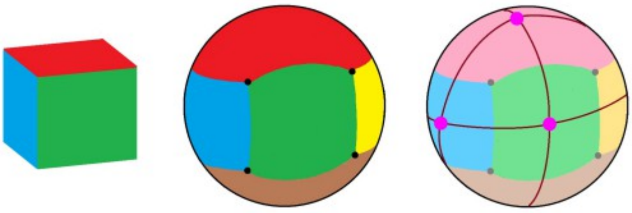

Exercise 1.15. Suppose K is a metric complex. Show how the set of its vertices K^0 becomes a metric space.

It follows from the definition that, in addition to the two volume functions of these complexes, we can also derive the *angle function*:

$$(a, a^\star) \mapsto \cos \widehat{aa^\star} = \frac{\langle a, a^\star \rangle}{|a||a^\star|}.$$

The function is well-defined whenever the expression in the right-hand belongs to our ring R. This is always the case when R is a field or $|a| = |a^\star| = 1$.

Definition 1.16. We say that K has a *rectangular geometry* if

$$\langle a, a^\star \rangle = \pm |a||a^\star|.$$

In other words, all dual pairs are orthogonal:

$$\cos \widehat{aa^\star} = \pm 1.$$

The case of $n = 1$ previously considered is rectangular.

Definition 1.17. We will say that the complex K has the *standard geometry* if it is rectangular and also fully normalized:

$$|a| = 1 \text{ for all cells } a \in K \sqcup K^\star.$$

We always assume that $\mathbb{R}^n$ has the standard geometry.

1.3 The Hodge star operator

We continue to take the concepts we have considered previously for the case of dimension 1 and define them for the general case of dimension n. We define an operator from primal to dual chains that incorporate the geometry of the complex. Its formula is based on the analysis above.

Definition 1.18. The (chain) *Hodge star operator* of a metric complex K of dimension n is the pair of homomorphisms on chains of complementary dimensions:

$$\begin{array}{ll} \star : C_k(K) \to C_{n-k}(K^\star), & k = 0, 1, ..., n, \\ \star : C_k(K^\star) \to C_{n-k}(K), & k = 0, 1, ..., n, \end{array}$$

defined by

$$\star(a) := \pm \frac{\langle a, a^\star \rangle}{|a^\star|} a^\star.$$

for any cell a, dual or primal, with the sign to be defined later.

The expression can be understood as the total flux of a across a region of area 1 on $a^\star$:

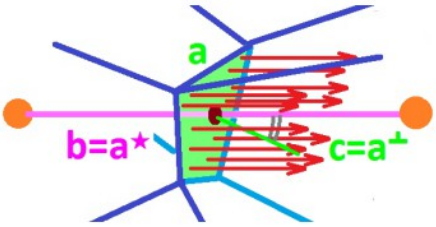

The two operators then satisfy:

$$\star(a)|a^\star| = \langle a, a^\star\rangle = \star(a^\star)|a|$$

The following statements are easy to prove.

Proposition 1.19. The matrix of the Hodge star operator $\star$ is diagonal with:

$$\star_{ii} = \pm\tfrac{|a_i|}{|a_i^\star|}\cos\widehat{a_i a_i^\star},$$

where a_i is the ith cell of K.

Theorem 1.20. The two star operators are the inverses of each other, up to a sign:

$$\star\star = \pm\,\mathrm{Id}\,.$$

Theorem 1.21. For a rectangular geometry, the Hodge star operator $\star$ preserves volumes of cells:

$$|\star(a)| = |a|.$$

Proposition 1.22. The Hodge star operator is always well-defined when

- R is a field, or
- the geometry of K is standard.

Exercise 1.23. Provide the Hodge star operator for a metric complex representation of a triangulation of the sphere realized as the hollow unit cube in $\mathbf{R}^3$.

Exercise 1.24. Provide the Hodge star operators for the examples of metric tensors presented in this and the last sections.

Next, what happens to the *forms* under Hodge duality?

Definition 1.25. In a metric complex K, the *Hodge-dual form* $\psi^\star$ of a form ψ over K is given by

$$\psi^\star(\sigma) := \psi(\sigma^\star).$$

The operation is extended from cells to chains by linearity making these diagrams commutative:

$$\begin{tikzcd}
C_k(K) \arrow[r, "\star"] \arrow[d, "\varphi"] & C_{n-k}(K^\star) \arrow[d, "\varphi^\star"] \\
R \arrow[r, phantom, "="] & R
\end{tikzcd}
\qquad
\begin{tikzcd}
C_k(K) \arrow[d, "\psi^\star"] & C_{n-k}(K^\star) \arrow[l, "\star"'] \arrow[d, "\psi"] \\
R \arrow[r, phantom, "="] & R
\end{tikzcd}$$

Then the formula is written for any chain:

$$(\star\psi)(a) := \psi(\star a),$$

where $\star$ in the right-hand side is the chain Hodge star operator.

Definition 1.26. The (cochain) *Hodge star operator* of a metric complex K is the dual of the chain Hodge star operator ($\star = \star^*$); i.e., it is the pair of homomorphisms,

$$\begin{array}{ll}\star : C^k(K) \to C^{n-k}(K^\star), & k = 0, 1, ..., n,\\ \star : C^k(K^\star) \to C^{n-k}(K), & k = 0, 1, ..., n,\end{array}$$

defined by

$$\star(\psi) := \psi^\star.$$

Proposition 1.27. The matrix of the Hodge star operator is diagonal with

$$\star_{ii} = \frac{|a_i|}{|a_i^\star|} \cos \widehat{aa^\star},$$

where a_i is the ith cell of K.

Exercise 1.28. Prove the proposition.

As before, we use in dimension $n = 1$:

$$\star^1_{ii} = \frac{|\text{point}|}{|\text{edge}|} = \frac{1}{\text{length}} = \frac{1}{\Delta x}.$$

Computing Hodge duality of forms over a complex with a non-trivial geometry is only slightly more complicated than for the standard case:

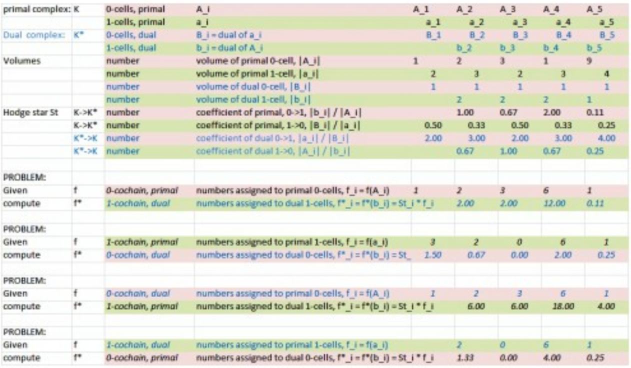

primal complex:	K	0-cells, primal	A_i	A_1	A_2	A_3	A_4	A_5
		1-cells, primal	a_i	a_1	a_2	a_3	a_4	a_5
Dual complex:	K*	0-cells, dual	B_i = dual of a_i	B_1	B_2	B_3	B_4	B_5
		1-cells, dual	b_i = dual of A_i		b_2	b_3	b_4	b_5
Volumes		number	volume of primal 0-cell, |A_i|	1	2	3	1	9
		number	volume of primal 1-cell, |a_i|	2	3	2	3	4
		number	volume of dual 0-cell, |B_i|	1	1	1	1	1
		number	volume of dual 1-cell, |b_i|		2	2	2	1
Hodge star St	K->K*	number	coefficient of primal, 0->1, |b_i| / |A_i|		1.00	0.67	2.00	0.11
	K->K*	number	coefficient of primal, 1->0, |B_i| / |a_i|	0.50	0.33	0.50	0.33	0.25
	K*->K	number	coefficient of dual 0->1, |a_i| / |B_i|	2.00	3.00	2.00	3.00	4.00
	K*->K	number	coefficient of dual 1->0, |A_i| / |b_i|		0.67	1.00	0.67	0.25
PROBLEM:								
Given	f	*0-cochain, primal*	numbers assigned to primal 0-cells, f_i = f(A_i)	*1*	*2*	*3*	*6*	*1*
compute	f*	*1-cochain, dual*	numbers assigned to dual 1-cells, f*_i = f*(b_i) = St_i * f_i		*2.00*	*2.00*	*12.00*	*0.11*
PROBLEM:								
Given	f	*1-cochain, primal*	numbers assigned to primal 1-cells, f_i = f(a_i)	*3*	*2*	*0*	*6*	*1*
compute	f*	*0-cochain, dual*	numbers assigned to dual 0-cells, f*_i = f*(b_i) = St_	*1.50*	*0.67*	*0.00*	*2.00*	*0.25*
PROBLEM:								
Given	f	*0-cochain, dual*	numbers assigned to primal 0-cells, f_i = f(A_i)	*1*	*2*	*3*	*6*	*1*
compute	f*	*1-cochain, primal*	numbers assigned to dual 1-cells, f*_i = f*(b_i) = St_i * f_i		*6.00*	*6.00*	*18.00*	*4.00*
PROBLEM:								
Given	f	*1-cochain, dual*	numbers assigned to primal 1-cells, f_i = f(a_i)		*2*	*0*	*6*	*1*
compute	f*	*0-cochain, primal*	numbers assigned to dual 0-cells, f*_i = f*(b_i) = St_i * f_i		*1.33*	*0.00*	*4.00*	*0.25*

Note: See the files online.

In dimension $n = 2$, we have the following diagonal entries for a rectangular grid:

$$\begin{array}{lllll}\star^2_{ii} & = \frac{|\text{point}|}{|\text{rectangle}|} & = \frac{1}{\text{area}} & = \frac{1}{\Delta x \Delta y},\\ \star^1_{ii} & = \frac{|\text{edge}|}{|\text{edge}|} & = \frac{\text{length}}{\text{length}} & = \frac{\Delta y}{\Delta x}.\end{array}$$

In a 2-dimensional cubical complex, combining two 1-cells may produce different orientations and, therefore, different signs. Since $a \times d = \tau$ but $d \times a = -\tau$, the duality operator is

$$\begin{array}{ll}\Big([0,1] \times \{0\}\Big)^\star & = \{0\} \times [0,1],\\ \Big(\{0\} \times [0,1]\Big)^\star & = -[0,1] \times \{0\}.\end{array}$$

In general, we are guided by the following rule.

Sign convention: The sign of the Hodge star is determined by:

$$\star \star \alpha = (-1)^{k(n-k)} \alpha,$$

for a k-form α.

Exercise 1.29. Confirm the sign convention for $\mathbb{R}^n$.

17.04.2020

conditions. Suppose we start off by requiring that a surface of revolution M have constant Gauss curvature K = -1.

E) Books_69_2020/
Differential Geometry and Its Applications/
Lokenath... Introduction to Hilbert Spaces....pdf.

F) Books_69_2020/
Nature and Numbers A Mathematical Photo Shooting/
www....Nature-and-Numbers.pdf

obs: Estudar com atenção!

G) Books_69_2020/
Topology Illustrated/
Sabet.. Topology Illustrated...pdf

obs: Estudar em detalhes!!!

a) Pág. 452(459)
4.2 The Euler-Poincaré Formula

b) Pág. 541 (548)
4. Metric Tensor
4.1 The need for measuring in calculus

data 17.04.2020
S T Q Q [✓ S] S D

c) Pág. 568(575)

1.2 Metric complexes.

Pág. 570(577)

They are orthogonal when this angle is zero.

For cells in the higher dimensional case, we are looking at the angle between a and the normal $b^\perp$ of $b = a^*$, illustrated for dimension $n = 2$

(image)

Dually, this angle can also be understood as the angle between the normal

$c = a^\perp$ of a and a^*,

illustrated for dimension $n = 3$:

(image)

Even though (apesar de) these angles don't affect the volumes of either primal or dual cells, they do affect the volumes of new n-cells, the cells determined by each dual pair. (image)

17.04.2020

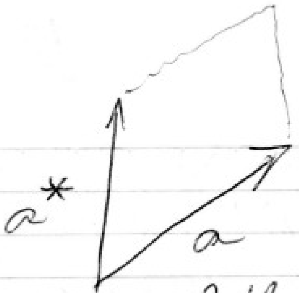

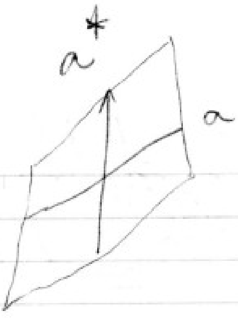

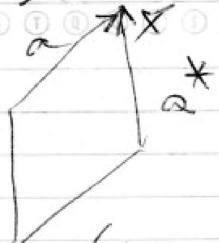

the area of this parallelogram (usar Wolfram Project) does not depend on how a and a^* are attached to each other (they aren't) but only on the angle between them.

↳ important.

(Gauss curvature → space)

↳ angle of Gauss space (179°...)

The way we handle (lidar) angle is, of course, via the inner product:

$\langle a, b \rangle \in \mathbb{R}$ is defined for these three cases:

- adjacent $a, b \in K$,
- adjacent $a, b \in K^*$, and now also
- $a \in K$, $a^* \in K^*$.

Geometric interpretation of the angle between two vectors defined using an inner product (of the vectors)

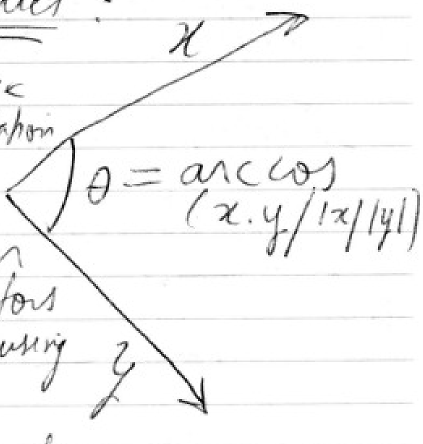

17.04.2020

Definition 1.13

A metric (cell) complex of dimension $\underline{\underline{n}}$ is a pair of Hodge-dual cell complexes K an K^* of dimension $\underline{\underline{n}}$ equipped with a joint metric tensor, which is a combination of metric tensor for K and K^* as well as inner product of dual pairs:

$$(a, a^*) \to \langle a, a^* \rangle \in R.$$

IMPORTANT: ✡

Exercise 1.14

Represent a hollow (oco) cube as a metric complex.

Image → interessante.

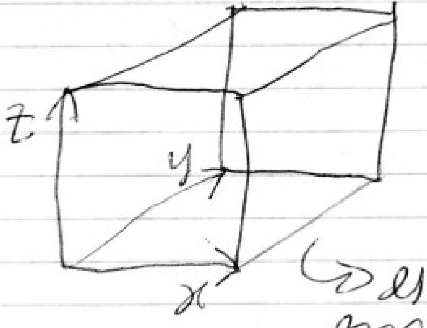

Este cubo inicial será as nossas coordenadas, usando-se coordenadas em funções algébricas

↳ este cubo será transformado em uma esfera dentro

17 09 2020

da métrica complexa.

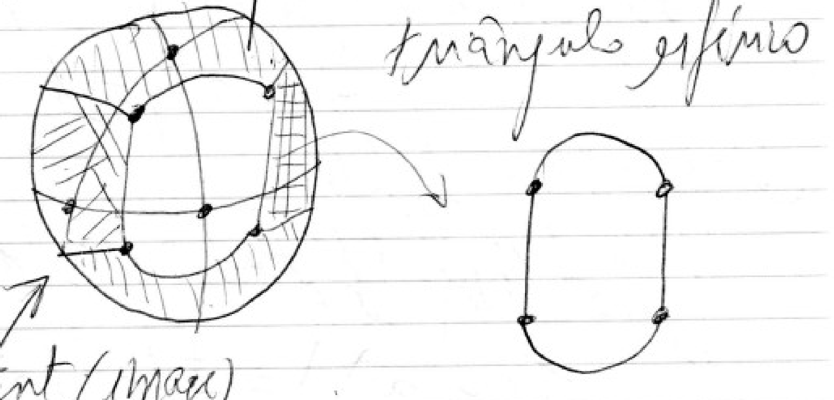

Important (Image)

Exercise 1.15

Suppose K is a metric complex. Show how the set of its vertices K^0 becomes a metric space.

It follows from the definition that, in addition to the two volume functions of these complexes, we can also derive the angle function:

$(a, a^*) \to$

$$\cos \widehat{a\, a^*} = \frac{\langle a, a^* \rangle}{|a|\,|a^*|}$$

importantíssimo

A área do paralelogramo depende do ângulo dos lados ou vetores.

The function is well-defined whenever the expression in the right-hand

17.04.2020

belongs to our ring R.
This is always the case when R is a field or $|a| = |a^*| = 1$.

Definition 1.16

We say that K has a rectangular geometry if

$$\langle a, a^* \rangle = \pm |a| |a^*| .$$

In other words, all dual pairs are orthogonal:

$$\cos \widehat{a\, a^*} = \pm 1$$

The case of $n=1$ previously considered is rectangular.

Definition 1.17

We will say that the complex K has the standard geometry if it is rectangular and also fully normalized:

$|a| = 1$ for all cells $a \in K \sqcup K^*$.

We always assume that $\mathbb{R}^n$ has the standard geometry.

H) a) Books_63_2020/
Gauss/
epdf.pub_the foundations of differential geometry.djvu

F /012/ / /3 / ((*" * -< "
" + "E5

/012/ / /3 / ((; E"G 5 . @ 0 % "
% " > * + " F *+ (@ 1 *"

- /012/ / /3 / ((*" * -< "
+ "E5

/012/ / /3 / ((*" * -< "
+ *"
3- K 5 " F -
' ! 09 6 "D 9

Making use of the notation of theorem 7, we may now say: The points A, A' are situated in space *upon one and the same side of the plane* α, and the points A, B are situated in space *upon different sides of the plane* α.

Theorem 7 gives us the most important facts relating to the order of sequence of the elements of space. These facts are the results, exclusively, of the axioms already considered, and, hence, no new space axioms are required in group II.

§ 5. GROUP III. AXIOM OF PARALLELS. (EUCLID'S AXIOM.)

The introduction of this axiom simplifies greatly the fundamental principles of geometry and facilitates in no small degree its development. This axiom may be expressed as follows:

> III. *In a plane* α *there can be drawn through any point* A, *lying outside of a straight line* a, *one and only one straight line which does not intersect the line* a. *This straight line is called the parallel to* a *through the given point* A.

This statement of the axiom of parallels contains two assertions. The first of these is that, in the plane α, there is always a straight line passing through A which does not intersect the given line a. The second states that only one such line is possible. The latter of these statements is the essential one, and it may also be expressed as follows:

> THEOREM 8. If two straight lines a, b of a plane do not meet a third straight line c of the same plane, then they do not meet each other.

For, if a, b had a point A in common, there would then exist in the same plane with c two straight lines *a* and *b* each passing through the point *A* and not meeting the straight line *c*. This condition of affairs is, however, contradictory to the second assertion contained in the axiom of parallels as originally stated. Conversely, the second part of the axiom of parallels, in its original form, follows as a consequence of theorem 8.

The axiom of parallels is a *plane axiom.*

§ 6. GROUP IV. AXIOMS OF CONGRUENCE.

The axioms of this group define the idea of congruence or displacement.

Segments stand in a certain relation to one another which is described by the word "*congruent.*"

IV, 1. *If A, B are two points on a straight line **a**, and if A' is a point upon the same or another straight line a', then, upon a given side of A'* on *the straight line a', we can always find one and only one point B' so that the segment AB (or BA) is congruent to the segment $A'B'$. We indicate this relation by writing*

$$AB \equiv A'B'.$$

Every segment is congruent to itself; that is, we always have

$$AB \equiv AB.$$

We can state the above axiom briefly by saying that every segment can be *laid off* upon a given side of a given point of a given straight line in one and and only one way.

IV, 2. ***If a segment AB is congruent to the segment $A'B'$ and also to the segment $A''B''$, then the segment $A'B'$ is congruent to the segment $A''B''$; that***

date 17.04.2020

belongs to our ring R.
This is always the case when R is a field or $|a| = |a^*| = 1$.

Definition 1.16
We say that K has a rectangular geometry if
$$\langle a, a^* \rangle = \pm |a||a^*|.$$

In other words, all dual pairs are orthogonal:
$$\cos \widehat{a\,a^*} = \pm 1$$

the case of $n=1$ previously considered is rectangular.

Definition 1.17
We will say that the complex K has the standard geometry if it is rectangular and also fully normalized:
$|a|=1$ for all cells $a \in K \sqcup K^*$.

We always assume that $\mathbb{R}^n$ has the standard geometry.

H) a) Books_63_2020/
Gauss/
epdf.pub_the foundations of differential geometry.djvu

17.04.2020

b) epdf.pub_the foundations of geometry.djvu
David Hilbert   188 pages.

obs: look:
Pág. 11 (29)
§ 5. Group III
Axiom of Parallels
(Euclid's axiom).

I) Books_69-2020/
The Mobius Strip/
Clifford A. Pickover... Mobius... djvu
other copy:
1560258268.pdf

145) 20200418 —

A) Books_70_2020/
Applications of Vector Analysis and Complex Variables in Engineering/
sanet_st_3030411672.pdf

obs. Livro todo

A1) Pág. 51 (63)
Coordinate Transformations:

3

/012/ / /3 / ((*" * -< " "
+< *"

/012/ / /3 / ((*" * -< " "
+< *"

% /012/ / /3 / ((,) ,) *"

: $ $)

V9 % (')# " " V9 C 5
"

! ! " " # $% & ' (%$
! ! ! " " # '
)*+,)))-) ' $ $ ' ./ 0!
*00 * 1 ! 0 2 3 4
-3 5 6 ! 777
(.$! 8

Chapter 5

Coordinate Transformations: Definitions of Vectors and Tensors

We consider the formal definitions of vectors and tensors, and the equations for their components in a transformed coordinate system. We restrict the analysis to straight coordinates. We first consider Cartesian coordinates and afterward non-Cartesian coordinates.

5.1 Coordinate Transformation

We consider a Cartesian system $\mathscr{B}$ with coordinate axes $(\tilde{1},\tilde{2},\tilde{3})$, representing a point in this system as $\tilde{x}_i$. The origin of $\mathscr{B}$ in terms of the original coordinate system $\mathscr{A}$ (1,2,3) is given by the coordinates b_j, see Figure 5.1. Both coordinate systems are Cartesian, and the scales in both systems are the same. The vectors $\underset{m}{\tau}^i$, defined as

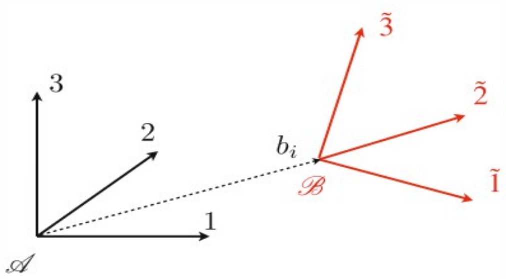

Figure 5.1: Transformation of Cartesian system $\mathscr{A}$ into Cartesian system $\mathscr{B}$

© Springer Nature Switzerland AG 2020
O. D. L. Strack, *Applications of Vector Analysis and Complex Variables in Engineering*, https://doi.org/10.1007/978-3-030-41168-8_5

$$\underset{m}{\tau}^i = \frac{\partial x_i}{\partial \tilde{x}_m} \qquad i = 1,2,3 \qquad m = 1,2,3, \tag{5.1}$$

are unit vectors tangent to the coordinate directions $\tilde{i}$ (compare (2.5)) with $\tilde{x}_m$ replacing the arc length along the curve, s. We define similar vectors that are tangent to the coordinates in system $\mathscr{A}$ in terms of the coordinates in system $\mathscr{B}$. We call these unit tangent vectors $\overset{m}{\gamma}_i$:

$$\overset{m}{\gamma}_i = \frac{\partial \tilde{x}_m}{\partial x_i}. \tag{5.2}$$

We express the original coordinates of a point x_i in terms of the new coordinates of that same point, and vice-versa, by integrating the latter two equations:

$$x_i = \underset{m}{\tau}^i \tilde{x}_m + b_i, \tag{5.3}$$

where b_i represents the origin of $\mathscr{B}$ in terms of x_i, and

$$\tilde{x}_m = \overset{m}{\gamma}_i x_i + \tilde{b}_m, \tag{5.4}$$

where $\tilde{b}_m$ represents the origin of the original system $\mathscr{A}$ in terms of the new coordinates.

The vectors $\underset{m}{\tau}^i$ are mutually orthogonal unit vectors,

$$\underset{m}{\tau}^i \underset{n}{\tau}^i = \delta_{mn}. \tag{5.5}$$

A similar equation holds for the unit vectors $\overset{m}{\gamma}_i$,

$$\overset{m}{\gamma}_i \overset{n}{\gamma}_i = \delta_{mn}. \tag{5.6}$$

We obtain a relationship between the vectors $\underset{m}{\tau}^i$ and $\overset{m}{\gamma}_i$ from (5.1) and (5.2)

$$\underset{m}{\tau}^i \overset{n}{\gamma}_i = \frac{\partial x_i}{\partial \tilde{x}_m} \frac{\partial \tilde{x}_n}{\partial x_i} = \frac{\partial \tilde{x}_n}{\partial \tilde{x}_m} = \delta_{mn}. \tag{5.7}$$

The moduli of $\underset{m}{\tau}^i$ and of $\overset{m}{\gamma}_i$ are 1; we combine this with the latter equation:

$$\underset{m}{\tau}^i = \overset{m}{\gamma}_i. \tag{5.8}$$

We introduce a matrix a_{ij},

$$a_{im} = \underset{m}{\tau}^i = \overset{m}{\gamma}_i = \frac{\partial x_i}{\partial \tilde{x}_m} = \frac{\partial \tilde{x}_m}{\partial x_i}, \tag{5.9}$$

and write (5.3) and (5.4) as

$$x_i = a_{ij} \tilde{x}_j + b_i, \tag{5.10}$$

and

$$\tilde{x}_i = a_{ji} x_j + \tilde{b}_i. \tag{5.11}$$

We observe from (5.9) that

$$a_{im} a_{in} = \frac{\partial x_i}{\partial \tilde{x}_m} \frac{\partial \tilde{x}_n}{\partial x_i} = \delta_{mn}, \tag{5.12}$$

and that

$$a_{im} a_{jm} = \frac{\partial \tilde{x}_m}{\partial x_i} \frac{\partial x_j}{\partial \tilde{x}_m} = \delta_{ij}. \tag{5.13}$$

5.2 Definition of Vectors and Tensors

A vector is a mathematical representation of some quantity that has both an orientation and a magnitude; it usually represents a physical quantity, such as the velocity of a water particle. A vector does not change its orientation and magnitude in space as a result of a coordinate transformation, which enables us to establish equations to determine if a set of three expressions represents a vector.

The differences between the coordinates of one point in space and another are the components of a vector; its length and position are invariant with respect to a coordinate transformation. Accordingly we introduce a vector A_i as

$$A_i = y_i - x_i \qquad i = 1,2,3, \tag{5.14}$$

where y_i and x_i are the coordinates of two arbitrary points in space. We use (5.11) to transform A_i into $\tilde{A}_i$:

$$\tilde{A}_i = \tilde{y}_i - \tilde{x}_i = a_{ji}(y_j - x_j) + \tilde{b}_i - \tilde{b}_i, \tag{5.15}$$

or

$$\tilde{A}_i = a_{ji} A_j. \tag{5.16}$$

We obtain the inverse transformation the same way from (5.10):

$$A_i = a_{ij} \tilde{A}_j. \tag{5.17}$$

The first index in a_{ij} corresponds to A_i in both equations. Equations (5.16) and (5.17) are the transformation formulas for a vector; any set of three components that transforms according to (5.16) and (5.17) is a vector in the sense defined above and represents a quantity that will not change due to a coordinate transformation.

5.2.1 Tensors of the Second Rank

A tensor of the second rank has nine components in three-dimensional space and has two indices, each ranging from 1 to 3. A tensor is an operator that produces a vector upon being applied to another vector. Thus, if the tensor A_{ij} operates on a vector X_j, the result is a new vector Y_i:

$$Y_i = A_{ij} X_j. \tag{5.18}$$

We obtain the transformation rules for a tensor of the second rank from those of a vector. We have, from (5.16)

$$\tilde{Y}_i = a_{ji} Y_j = a_{ji} A_{jm} X_m. \tag{5.19}$$

We use (5.17) to express X_m in terms of $\tilde{X}_m$;

$$\tilde{Y}_i = a_{ji} A_{jm} a_{mk} \tilde{X}_k = a_{ji} a_{mk} A_{jm} \tilde{X}_k. \tag{5.20}$$

We have, by definition

$$\tilde{Y}_i = \tilde{A}_{ij} \tilde{X}_j, \tag{5.21}$$

so that

$$\tilde{A}_{mn} = a_{im} a_{jn} A_{ij}. \tag{5.22}$$

The first index in the matrices a_{im} and a_{jn} refers to the original tensor components; the second index to those of the transformed one.

We obtain the inverse transformation in a similar way:

$$A_{mn} = a_{mi} a_{nj} \tilde{A}_{ij}. \tag{5.23}$$

Again, the first index in each transformation matrix refers to the original tensor.

A special tensor of the second rank is the Kronecker delta; its transformation is:

$$\tilde{\delta}_{mn} = a_{im} a_{jn} \delta_{ij} = a_{im} a_{in}, \tag{5.24}$$

or, with (5.13)

$$\tilde{\delta}_{mn} = \delta_{mn}. \tag{5.25}$$

The Kronecker delta is invariant with respect to a Cartesian coordinate transformation.

5.2.2 The Gradient Tensor

Consider a vector field A_i. We apply the operator ∂_i to this vector field to obtain a tensor, called the gradient tensor,

$$A_{ij} = \partial_i A_j \tag{5.26}$$

Problem 5.1

Demonstrate that $\partial_i A_j$ is a tensor of the second rank.

5.2.3 Principal Directions

A tensor operating on a vector X_i produces a new vector Y_i,

$$Y_i = A_{ij} X_j. \tag{5.27}$$

We choose the direction of the vector X_i such that the vector Y_i is collinear with X_i; the direction of this vector is the principal direction of the tensor. For this case, Y_i equals λX_i, where λ is some scalar quantity,

$$Y_i = \lambda X_i = A_{ij} X_j, \tag{5.28}$$

or

$$A_{ij}X_j - \lambda X_i = (A_{ij} - \lambda \delta_{ij})X_j = 0. \tag{5.29}$$

This represents a system of three homogeneous equations for the components of X_i, which has non-trivial solutions only if the determinant of the system is zero, i.e., if

$$\det(A_{ij} - \lambda \delta_{ij}) = \begin{vmatrix} A_{11} - \lambda & A_{12} & A_{13} \\ A_{21} & A_{22} - \lambda & A_{23} \\ A_{31} & A_{32} & A_{33} - \lambda \end{vmatrix} = 0. \tag{5.30}$$

This third-order equation is the characteristic equation of the tensor A_{ij}; its roots are the principal values of the tensor. The corresponding three directions of X_i are the characteristic directions, or principal directions, of the tensor. The homogenous equations can be divided by a fixed constant. If we choose for this constant the magnitude (modulus) of X_i, the result will be a unit vector, fully determined by two values. The principal values and directions are also known as the eigenvalues and eigenvectors of the tensor.

5.3 Non-Cartesian Transformations with Straight Axes

We consider straight transformed coordinates $\tilde{1},\tilde{2},\tilde{3}$ that are not orthogonal. That the coordinate axes are straight is an important restriction; differentiation along curves adds terms to derivatives, as we saw when establishing equations for the acceleration; the additional terms are the centrifugal acceleration, which is due entirely to the curvature of the path. Since we make restrictions to straight axes, such additional terms do not appear.

The definitions of the base vectors $\underset{m}{\tau}^i$ and $\overset{m}{\gamma}_i$ remain as before:

$$\underset{m}{\tau}^i = \frac{\partial x_i}{\partial \tilde{x}_m}, \tag{5.31}$$

and

$$\overset{m}{\gamma}_i = \frac{\partial \tilde{x}_m}{\partial x_i}. \tag{5.32}$$

The base vectors $\underset{m}{\tau}^i$ represent the tangent vectors to the coordinate axes $\tilde{1},\tilde{2},\tilde{3}$, but are not orthogonal. We interpret the base vectors $\overset{m}{\gamma}_i$ now in a different manner; we view them as the gradients of the functions $\tilde{x}_m(x_1,x_2,x_3)$, i.e., the base vector $\overset{1}{\gamma}_i$ points in the direction of the gradient of the function $\tilde{x}_1(x_1,x_2,x_3)$. The planes $\tilde{x}_1$=constant are, by definition, parallel to the $\tilde{2},\tilde{3}$-plane and the base vector $\overset{1}{\gamma}_i$ is normal to the $\tilde{2},\tilde{3}$-plane. The three base vectors $\overset{m}{\gamma}_i$, m=1,2,3 are thus orthogonal to the three planes $\tilde{2},\tilde{3}$, $\tilde{1},\tilde{3}$, and $\tilde{1},\tilde{2}$, respectively, see Figure 5.2.

An important difference between Cartesian and non-Cartesian coordinate systems is that the base vectors cannot all be unit vectors. We have the following relationships between the two sets of base vectors

$$\underset{m}{\tau}^i \overset{n}{\gamma}_i = \frac{\partial x_i}{\partial \tilde{x}_m} \frac{\partial \tilde{x}_n}{\partial x_i} = \delta_{mn}, \tag{5.33}$$

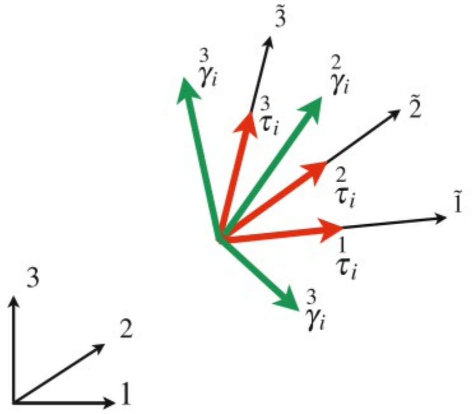

Figure 5.2: Non-orthogonal coordinates; the coordinates 1,2,3 are Cartesian; the coordinates $\tilde{1},\tilde{2},\tilde{3}$ are not. The three base vectors $\underset{m}{\tau}^i$ are tangent to the $\tilde{1},\tilde{2},\tilde{3}$ coordinates, the three gradient vectors $\overset{m}{\gamma}_i$ are orthogonal to the planes of the two coordinate axes that do not contain the number m.

and

$$\underset{m}{\tau}^i\overset{m}{\gamma}_j = \frac{\partial x_i}{\partial \tilde{x}_m}\frac{\partial \tilde{x}_m}{\partial x_j} = \delta_{ij}. \tag{5.34}$$

The dot product of tangential and gradient base vectors is zero because the tangential base vectors $\underset{m}{\tau}^i$ are normal to the gradient base vectors $\overset{n}{\gamma}_i$, except when $m = n$, in which case their dot product is one. The base vectors $\underset{m}{\tau}^i$ and $\overset{m}{\gamma}_i$ are generally not at right angles; the dot product of the base vectors is equal to

$$\underset{m}{\tau}^i\overset{n}{\gamma}_i = \underset{m}{\tau}\overset{n}{\gamma}\cos\underset{m}{\overset{n}{\alpha}} = \delta_{mn}, \tag{5.35}$$

where $\underset{m}{\overset{n}{\alpha}}$ is the angle between the base vectors. Since $\cos\underset{m}{\overset{n}{\alpha}} \neq 1$ unless $\underset{m}{\overset{n}{\alpha}} = \pi/2$, the lengths of the base vectors will generally not be equal to one, as in the Cartesian case. This is the result of the scale of measurement along the non-orthogonal coordinate axes being different from those along the Cartesian axes. In the present case the scales of measurement do not vary in space because the coordinate transformations are linear; the components of the base vectors are constants that depend upon the angles between the coordinate axes.

The components of a vector in the non-orthogonal system are not equal to their projections on the coordinate directions. There are two different sets of base vectors; we may decompose a vector in terms of components either along the tangent bases, or along the gradient bases. The former components are called the *contravariant* components and are denoted with the index in the place where normally an exponent appears, i.e., the contravariant components of a vector A_i in the transformed system are written as $\tilde{A}^i$. The Cartesian components of the vector A_i are related to the contravariant components $\tilde{A}^i$ as:

$$A_i = \underset{m}{\tau}^{i}\tilde{A}^m. \tag{5.36}$$

We may also decompose the vector in terms of components along the gradient bases. The corresponding components are the *covariant* components; we place the indices at the lower right corner of the symbol:

$$A_i = \overset{m}{\gamma}_i\tilde{A}_m \tag{5.37}$$

The two different kinds of decomposition into components are illustrated in Figure 5.3 for the two-dimensional case. The components $\tilde{A}_i$ and $\tilde{A}^i$ are not equal in length to the projections of A_i onto the base vectors; the projections consist of the dot product of $\tilde{A}^m$ and $\underset{m}{\tau}^{i}$ and that of $\tilde{A}_m$ and $\overset{m}{\gamma}_i$ and the base vectors are not of unit length. Since the covariant and contravariant components of the vectors are not equal to their projections, we distinguish them from their projections, called the physical components.

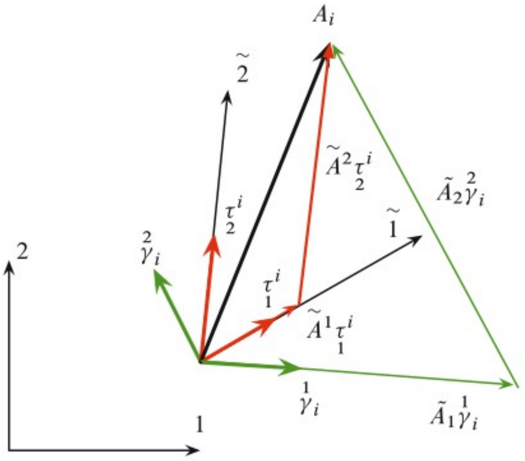
A_i
$\tilde{2}$
$\underset{2}{\tau}^{i}$
$\tilde{A}^2\underset{2}{\tau}^{i}$
$\tilde{A}_2\overset{2}{\gamma}_i$
$\tilde{1}$
$\overset{2}{\gamma}_i$
2
$\underset{1}{\tau}^{i}$
$\tilde{A}^1\underset{1}{\tau}^{i}$
$\overset{1}{\gamma}_i$
$\tilde{A}_1\overset{1}{\gamma}_i$
1

Figure 5.3: Decomposition of a vector A_i in its contra variant and covariant components

The Cartesian coordinate system is a special case, where covariant and contravariant components are equal; we place the index denoting Cartesian components at the lower right corner.

We determine a relationship between the contravariant and covariant components of the vector by first setting the right-hand sides of (5.36) and (5.37) equal to each other

$$\underset{m}{\tau}^{i}\tilde{A}^m = \overset{m}{\gamma}_i\tilde{A}_m, \tag{5.38}$$

and next contract both sides of this equation with $\overset{n}{\gamma}_i$:

$$\overset{n}{\gamma}_i\underset{m}{\tau}^{i}\tilde{A}^m = \overset{n}{\gamma}_i\overset{m}{\gamma}_i\tilde{A}_m. \tag{5.39}$$

We apply (5.34) to the left-hand side:

$$\delta_{nm}\tilde{A}^m = \tilde{A}^n = g^{nm}\tilde{A}_m. \tag{5.40}$$

The dot product of the covariant bases, g^{mn}, is called the *contravariant metric tensor*, and represents the scales of measurement in the directions normal to the axes. Indeed, the metric tensor is the dot product of the various covariant base vectors and thus indeed reflect the scales of measurement,

$$g^{nm} = \overset{n}{\gamma}_i \overset{m}{\gamma}_i. \tag{5.41}$$

If we contract both sides of (5.38) with $\underset{n}{\tau}^i$, we obtain:

$$\underset{m}{\tau}^i \underset{n}{\tau}^i \tilde{A}^m = \underset{n}{\tau}^i \overset{m}{\gamma}_i \tilde{A}_m = \delta_{nm}\tilde{A}_m = \tilde{A}_n, \tag{5.42}$$

or, introducing the *covariant metric tensor* g_{mn} as

$$g_{mn} = \underset{m}{\tau}^i \underset{n}{\tau}^i, \tag{5.43}$$

we obtain

$$\tilde{A}_n = g_{mn}\tilde{A}^m. \tag{5.44}$$

We say that contraction of the contravariant components $\tilde{A}^m$ with the covariant metric tensor g_{mn} 'brings the index down', i.e., results in producing the covariant components of the vector. Similarly, contraction of the covariant components $\tilde{A}_m$ with the contravariant metric tensor g^{mn} 'raises the index up', i.e., produces the contravariant components of the vector. We recall the definitions of the two metric tensors

$$g^{mn} = \overset{m}{\gamma}_i \overset{n}{\gamma}_i = \frac{\partial \tilde{x}_m}{\partial x_i}\frac{\partial \tilde{x}_n}{\partial x_i} \tag{5.45}$$

$$g_{mn} = \underset{m}{\tau}^i \underset{n}{\tau}^i = \frac{\partial x_i}{\partial \tilde{x}_m}\frac{\partial x_i}{\partial \tilde{x}_n}. \tag{5.46}$$

The two metric tensors are symmetrical.

We obtain the contravariant components $\tilde{A}^m$ from (5.36) by contraction with $\overset{n}{\gamma}_i$

$$A_i \overset{n}{\gamma}_i = \overset{n}{\gamma}_i \underset{m}{\tau}^i \tilde{A}^m = \tilde{A}^m \delta_{mn} = \tilde{A}^n, \tag{5.47}$$

and obtain the covariant components by contracting (5.37) with $\underset{n}{\tau}^i$

$$A_i \underset{n}{\tau}^i = \underset{n}{\tau}^i \overset{m}{\gamma}_i \tilde{A}_m = \delta_{mn}\tilde{A}_m = \tilde{A}_n. \tag{5.48}$$

The transformation rules for vectors are given by four equations in this case, two each for the contravariant and two each for the covariant components of the vector, i.e., by (5.36) with (5.47) and by (5.37) with (5.48)

Similar transformation rules apply to tensors; the rules for the contravariant components become

$$\tilde{A}^{mn} = \overset{m}{\gamma}_i \overset{n}{\gamma}_j A_{ij} \tag{5.49}$$

There exist a mix of components as well

$$\tilde{A}^m{}_n = \overset{m}{\gamma}_i \underset{n}{\tau}^j A_{ij}, \tag{5.50}$$

and

$$\tilde{A}_m{}^n = \underset{m}{\tau}^i \overset{n}{\gamma}_j A_{ij}. \tag{5.51}$$

5.3.1 Invariants

We concluded when dealing with Cartesian systems that contraction that eliminates all free indices results in a scalar quantity. This is not true for components in general coordinate systems. For example, the quantity A_{ii} is a scalar if the components of this tensor are in a Cartesian coordinate system. However, the quantities $\tilde{A}^{ii}$ and $\tilde{A}_{ii}$ are not invariants. We see this by expressing these quantities in terms of Cartesian coordinates

$$\tilde{A}^{mm} = \overset{m}{\gamma}_i \overset{m}{\gamma}_j A_{ij}, \tag{5.52}$$

which is not an invariant since the base vectors are unrelated to the components of the tensor A_{ij}. However, the quantities $\tilde{A}^m{}_m$ and $\tilde{A}_m{}^m$ are invariant scalars as follows from the transformation rules and (5.34),

$$\tilde{A}^m{}_m = \overset{m}{\gamma}_i \underset{m}{\tau}^j A_{ij} = \delta_{ij} A_{ij} = A_{ii}. \tag{5.53}$$

We obtain a similar result for $\tilde{A}_m{}^m$.

We create invariants by contracting indices in opposite positions, i.e., an index in the upper right position balanced by one in the lower right position of the contracting term. The dot product, an invariant, is

$$A_i A_i = \tilde{A}^i \tilde{A}_i, \tag{5.54}$$

which is invariant and matches A_iA_i as we see by contracting (5.36) with (5.37),

$$A_i A_i = \underset{m}{\tau}^i \tilde{A}^m \overset{n}{\gamma}_i \tilde{A}_n = \delta_{mn} \tilde{A}^m \tilde{A}_n = \tilde{A}^m \tilde{A}_m. \tag{5.55}$$

5.3.2 The Gradient

Vectors are usually expressed in terms of their contravariant components, whereas the components of the gradient of a scalar are preferably expressed in terms of their covariant components. This becomes clear by considering the definition of the gradient of some scalar function F,

$$\frac{\partial F}{\partial \tilde{x}_i} = \frac{\partial F}{\partial x_m} \frac{\partial x_m}{\partial \tilde{x}_i} = \underset{i}{\tau}^m \frac{\partial F}{\partial x_m}, \tag{5.56}$$

which conforms to the transformation rule for a covariant vector, (5.48).

Of particular interest for later use is the transformation of the Laplacian, $L_{ii} = \partial_i \partial_i F$. The Laplacian is an invariant, so that we need to transform the gradient tensor L_{ij} as follows:

$$\tilde{L}^m{}_m = \overset{m}{\gamma}_i \underset{m}{\tau}^j L_{ij} = \delta_{ij} L_{ij} = L_{ii} = \nabla^2 F. \tag{5.57}$$

We discuss in the remainder of this chapter the stress, displacement gradient, and strain tensors. We do this in terms of Cartesian coordinates.

5.4 The Stress Tensor

The stress tensor is of the second rank, and operates on the vector normal to a plane to produce the stress vector, or traction, that operates on that plane,

$$t_i = \sigma_{ji} n_j, \tag{5.58}$$

where t_i represents the stress vector, and n_i the unit normal to the plane. The unit normals to the planes in the positive coordinate directions are:

$$\overset{m}{n}_i = \delta_{mi}, \tag{5.59}$$

where m represents the coordinate axis that serves as the normal to the plane considered. We apply (5.58) to the three normal vectors of (5.59) and obtain the following for the corresponding stress vectors $\overset{m}{t}_i$:

$$\overset{m}{t}_i = \sigma_{ji} \delta_{mj} = \sigma_{mi}. \tag{5.60}$$

It appears that the components σ_{mi} for $m = 1,2,3$ represent the three stress vectors, or tractions, on the planes with normals in the $1,2,3$ directions, respectively, see Figure 5.4. The stresses that act on the planes with their normals in the negative coordinate directions are obtained from (5.58) by replacing n_i by $\overset{m}{n}_i$ with

$$\overset{m}{n}_i = -\delta_{mi}, \tag{5.61}$$

so that (5.58) becomes

$$\overset{m}{t}_i = -\sigma_{ji} \delta_{mi} = -\sigma_{mi}. \tag{5.62}$$

The expression σ_{ii} is a scalar since there is no free index, and is invariant to a coordinate transformation; this invariant is called the first invariant of the stress tensor. The isotropic stress is defined as one-third of the first invariant and is usually represented as $\overset{0}{\sigma}$

$$\overset{0}{\sigma} = \tfrac{1}{3} \sigma_{ii}. \tag{5.63}$$

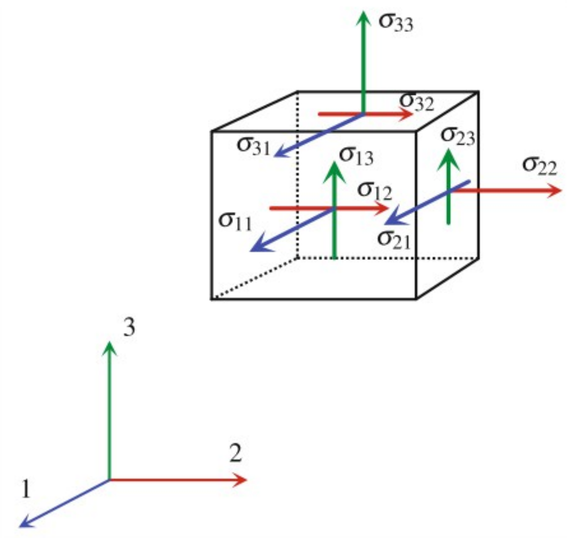

Figure 5.4: Definition of the components of the stress tensor.

5.4.1 The Normal and Shear Stress Components

We decompose the stress vector that acts on a plane into two components: the normal stress and the shear stress; the former is normal to the plane, and the latter acts in the plane. We contract t_i with n_i to obtain the normal stress:

$$\overset{n}{t} = t_i n_i = \sigma_{ij} n_i n_j. \tag{5.64}$$

The normal stress is positive if pointing in the direction of the unit normal n_i. We multiply $\overset{n}{t}$ with n_j to obtain the normal stress as a vector:

$$\overset{n}{t}_j = t_i n_i n_j = \sigma_{nm} n_n n_m n_j. \tag{5.65}$$

The tangential component of the stress vector is the difference between the stress vector t_i and its normal component:

$$\overset{s}{t}_j = t_j - t_i n_i n_j. \tag{5.66}$$

5.4.2 Principal Directions and Principal Values of the Stress Tensor in Two Dimensions.

We determine the principal directions and the principal values of the stress tensor in two dimensions for the case of plane stress, i.e., the case that there are no stresses in planes with normals in the 3-direction.

Principal Directions

The principal directions are the directions of the normals to the planes that carry only normal stresses, i.e., the planes where the shear stresses are zero. We rotate the 1,2-coordinate system in such a way that the shear stress on the plane with its normal in the $\tilde{1}$-direction is zero.

The transformed stress tensor $\tilde{\sigma}_{1j} = a_{m1}a_{nj}\sigma_{mn}$ represents the stresses on the plane normal to the $\tilde{1}$ axis and $\tilde{\sigma}_{12} = 0$,

$$\tilde{\sigma}_{12} = 0 = a_{m1}a_{n2}\sigma_{mn}, \tag{5.67}$$

where

$$a_{mj} = \underset{j}{\tau}^{m}. \tag{5.68}$$

If we represent the angle between the 1-axis and the $\tilde{1}$-axis as ψ then

$$\begin{aligned} a_{m1} &= \underset{1}{\tau}^{m} = (\cos\psi, \sin\psi) \\ a_{m2} &= \underset{2}{\tau}^{m} = (-\sin\psi, \cos\psi). \end{aligned} \tag{5.69}$$

We work out (5.67):

$$\begin{aligned} 0 &= a_{11}a_{12}\sigma_{11} + a_{11}a_{22}\sigma_{12} + a_{21}a_{12}\sigma_{21} + a_{21}a_{22}\sigma_{22} \\ &= -\sigma_{11}\sin\psi\cos\psi + \sigma_{12}\cos^2\psi - \sigma_{12}\sin^2\psi + \sigma_{22}\sin\psi\cos\psi \\ &= -\tfrac{1}{2}\sin(2\psi)(\sigma_{11} - \sigma_{22}) + \sigma_{12}\cos(2\psi). \end{aligned} \tag{5.70}$$

We solve for the tangent of 2ψ, i.e., the tangent of twice the angle that the principal direction makes with the 1-axis,

$$\tan(2\psi) = \frac{\sigma_{12}}{\frac{1}{2}(\sigma_{11} - \sigma_{22})}\ . \tag{5.71}$$

Principal Values

With the principal directions known, we determine the principal values in terms of the principal directions. We obtain an expression for the principal stress σ_{11} from $\tilde{\sigma}_{ij} = a_{mi}a_{nj}\sigma_{mn}$ by setting both i and j equal to 1:

$$\begin{aligned} \tilde{\sigma}_{11} = a_{m1}a_{n1}\sigma_{mn} &= a_{11}^2\sigma_{11} + a_{21}a_{11}\sigma_{21} + a_{11}a_{21}\sigma_{12} + a_{21}^2\sigma_{22} \\ &= \sigma_{11}\cos^2\psi + 2\sigma_{12}\sin\psi\cos\psi + \sigma_{22}\sin^2\psi \\ &= \tfrac{1}{2}\sigma_{11}[1 + \cos(2\psi)] + \tfrac{1}{2}\sigma_{22}[1 - \cos(2\psi)] + \sigma_{12}\sin(2\psi) \\ &= \tfrac{1}{2}(\sigma_{11} + \sigma_{22}) + \tfrac{1}{2}(\sigma_{11} - \sigma_{22})\cos(2\psi) + \sigma_{12}\sin(2\psi). \end{aligned} \tag{5.72}$$

The second principal stress acts on a plane that is at right angles to the plane with its normal in the direction ψ. Since the angle ψ is multiplied by a factor 2, and since the signs of both the cosine and the sine function change upon adding π to the argument, we obtain for $\tilde{\sigma}_{22}$:

$$\tilde{\sigma}_{22} = \tfrac{1}{2}(\sigma_{11} + \sigma_{22}) - \tfrac{1}{2}(\sigma_{11} - \sigma_{22})\cos(2\psi) - \sigma_{12}\sin(2\psi). \tag{5.73}$$

5.4.3 Stress Deviator

The difference between the major and minor principal stresses is the *stress deviator*, half of which we represent by the symbol λ. It is sometimes useful to express the three components of the stress tensor in terms of λ, ψ, and $\overset{0}{\sigma}$. We write the major principal stress, (5.72), as $\overset{0}{\sigma}+\lambda$, and subtract $\overset{0}{\sigma}$ from both sides of the equation,

$$\lambda = \tfrac{1}{2}(\sigma_{11}-\sigma_{22})\cos(2\psi)+\sigma_{12}\sin(2\psi). \tag{5.74}$$

We use (5.71),

$$\tfrac{1}{2}(\sigma_{11}-\sigma_{22})\sin(2\psi)-\sigma_{12}\cos(2\psi)=0, \tag{5.75}$$

multiply (5.74) by $\cos(2\psi)$ and (5.75) by $\sin(2\psi)$, and add:

$$\lambda\cos(2\psi)=\tfrac{1}{2}(\sigma_{11}-\sigma_{22}). \tag{5.76}$$

We find the expression for σ_{12} by multiplying (5.74) by $\sin(2\psi)$, multiplying (5.75) by $\cos(2\psi)$, and subtracting the results,

$$\lambda\sin(2\psi)=\sigma_{12}. \tag{5.77}$$

We first square, then add, the two sides of the latter two equations in order to obtain an expression for λ in terms of the components of the stress tensor:

$$\lambda^2=\left[\tfrac{1}{2}(\sigma_{11}-\sigma_{22})\right]^2+\sigma_{12}^2. \tag{5.78}$$

We find expressions for σ_{11} and σ_{22} from (5.76) and $\overset{0}{\sigma}=\frac{1}{2}(\sigma_{11}+\sigma_{22})$,

$$\begin{aligned}\sigma_{11}&=\overset{0}{\sigma}+\lambda\cos(2\psi)\\ \sigma_{22}&=\overset{0}{\sigma}-\lambda\cos(2\psi).\end{aligned} \tag{5.79}$$

5.4.4 Invariants

The first invariant of the stress tensor, σ_{ii}, equals twice the isotropic stress,

$$\underset{1}{I}=\sigma_{kk}=2\overset{0}{\sigma}. \tag{5.80}$$

We obtain the expression for a second invariant by contracting σ_{ij} with itself:

$$\underset{2}{I}=\sigma_{ij}\sigma_{ij}. \tag{5.81}$$

The stress deviator is also an invariant. In order to demonstrate this, we form the combination:

$$\begin{aligned}\underset{2}{I}-\underset{1}{I}^2 &=\sigma_{ij}\sigma_{ij}-\sigma_{kk}\sigma_{mm}\\ &=\sigma_{11}^2+2\sigma_{12}^2+\sigma_{22}^2-(\sigma_{11}+\sigma_{22})^2=2\sigma_{12}^2-2\sigma_{11}\sigma_{22}.\end{aligned} \tag{5.82}$$

We expand the expression for λ^2, (5.78),

$$\lambda^2 = \tfrac{1}{4}\left[\sigma_{11}^2 + \sigma_{22}^2 - 2\sigma_{11}\sigma_{22}\right] + \sigma_{12}^2 = \tfrac{1}{4}\left[\sigma_{11}^2 + \sigma_{22}^2 + 2\sigma_{11}\sigma_{22}\right] - \sigma_{11}\sigma_{22} + \sigma_{12}^2, \tag{5.83}$$

or with (5.82),

$$\lambda^2 = \tfrac{1}{4}\sigma_{kk}^2 + \tfrac{1}{2}(-I_1^2 + I_2) = -\tfrac{1}{4}I_1^2 + \tfrac{1}{2}I_2, \tag{5.84}$$

which is indeed an invariant.

Problem 5.2

Derive equations for the angle ψ between the major principal stress and the 1-axis, using (5.66), by setting the shear stress component equal to zero, expressing the traction in terms of the stress tensor and the unit normal to the plane. Compare your answer with (5.71).

5.4.5 The Equilibrium Conditions

The stresses that act on the six planes of the cube in Figure 5.4 act on opposite planes in opposite direction from one another. Thus, if the stresses are constant, they are in equilibrium. The situation changes if the stresses vary with position. We apply the condition of equilibrium to the stresses in the 2-direction on a volume of sides $\Delta x_1, \Delta x_2, \Delta x_3$ and centered at x_i, see Figure 5.5

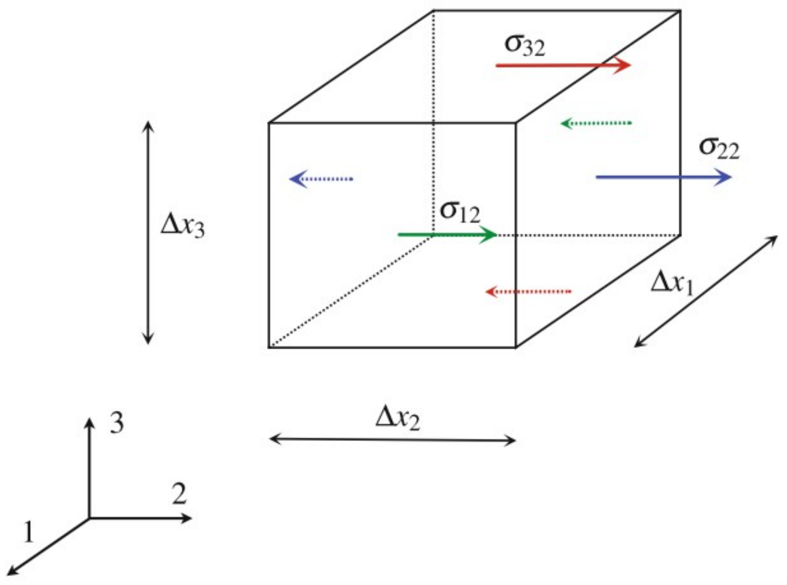

Figure 5.5: Equilibrium in the 2-direction

$$\begin{aligned}
&[\sigma_{22}(x_1,x_2+\tfrac{1}{2}\Delta x_2,x_3)-\sigma_{22}(x_1,x_2-\tfrac{1}{2}\Delta x_2,x_3)]\Delta x_1\Delta x_3\\
&+[\sigma_{12}(x_1+\tfrac{1}{2}\Delta x_1,x_2,x_3)-\sigma_{12}(x_1-\tfrac{1}{2}\Delta x_1,x_2,x_3)]\Delta x_2\Delta x_3\\
&+[\sigma_{32}(x_1,x_2,x_3+\tfrac{1}{2}\Delta x_3)-\sigma_{32}(x_1,x_2,x_3-\tfrac{1}{2}\Delta x_3)]\Delta x_1\Delta x_2\\
&+B_2\Delta x_1\Delta x_2\Delta x_3=F_2\Delta x_1\Delta x_2\Delta x_3.
\end{aligned} \tag{5.85}$$

where B_2 is the component of the body force B_i per unit volume. The force F_i, also per unit volume, is the resultant force due to the stresses and the body force B_i. This force is zero if the system is in equilibrium; if it is not, then $F_i = \rho e_i$, where ρ is the mass per unit volume, i.e., the density, and e_i is a unit vector pointing in the direction of the resultant force. We pass to the limit for $\Delta x_1 \to 0, \Delta x_2 \to 0, \Delta x_3 \to 0$ and obtain

$$\partial_i \sigma_{i2} + B_2 = F_2 = \rho a_2. \tag{5.86}$$

We generalize this equation to apply to all three coordinate directions,

$$\partial_i \sigma_{ij} + B_j = F_j = \rho a_j. \tag{5.87}$$

5.5 The Displacement Gradient and Strain Tensors

We consider a deforming body and represent the displacements of points in that body relative to some fixed initial state as u_i. We begin by examining a linear displacement, and consider displacement vectors of two points, $\overset{\mathscr{P}}{u}_i$ at $\mathscr{P}$ and $\overset{\mathscr{Q}}{u}_i$ at $\mathscr{Q}$. The vector $l\nu_i$, where ν_i is a unit vector, points from point $\mathscr{P}$ to point $\mathscr{Q}$

$$l\nu_i = \overset{\mathscr{Q}}{x}_i - \overset{\mathscr{P}}{x}_i, \tag{5.88}$$

where $\overset{\mathscr{P}}{x}_i$ and $\overset{\mathscr{Q}}{x}_i$ are the coordinates of points $\mathscr{P}$ and $\mathscr{Q}$, see Figure 5.6. The relative displacement per unit length is a_i, with

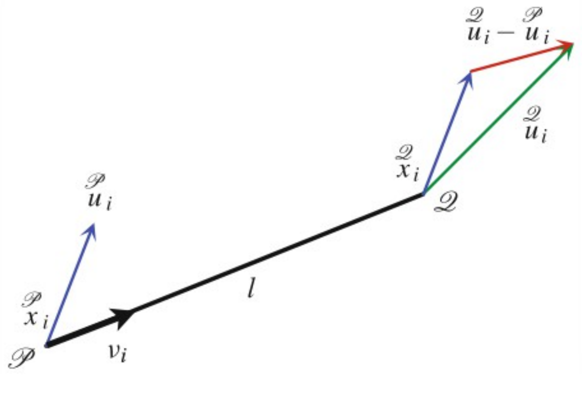
$\overset{\mathscr{Q}}{u}_i - \overset{\mathscr{P}}{u}_i$
$\overset{\mathscr{Q}}{u}_i$
$\overset{\mathscr{Q}}{x}_i$
$\mathscr{Q}$
$\overset{\mathscr{P}}{u}_i$
l
$\overset{\mathscr{P}}{x}_i$
$\mathscr{P}$
ν_i

Figure 5.6: Relative displacement of the endpoints of a filament of length l with orientation ν_i

$$a_i = \frac{\Delta u_i}{l} = \frac{\overset{\mathscr{Q}}{u}_i - \overset{\mathscr{P}}{u}_i}{l} \tag{5.89}$$

The relative displacement vector a_i is useful to describe the deformation of a body that experiences linear displacements.

5.5.1 The Displacement Gradient Tensor

The displacements generally vary with position; we use the displacement gradient tensor rather than the relative displacement vector to describe the deformation of the body:

$$\alpha_{ij} = \partial_i u_j = \frac{\partial u_j}{\partial x_i}. \tag{5.90}$$

We express the relative displacement vector a_i in terms of the displacement gradient tensor as:

$$a_j = \lim_{|\Delta x_i| \to 0} \frac{\partial u_j}{\partial x_i} \frac{\Delta x_i}{\sqrt{\Delta x_m \Delta x_m}} = \alpha_{ij} \nu_i, \tag{5.91}$$

where ν_i is a unit vector in the direction of the filament considered. This equation is similar to that obtained for the stress vector acting on an arbitrary plane.

The vector a_i represents the relative displacement of the end points of a filament of incremental length oriented in the direction ν_i. As for stresses, we define relative displacements in the direction normal to ν_i and in the direction of ν_i. The latter displacement represents the stretch of the filament, whereas the former represents rotation of the filament. The relative rotation of two filaments that are originally at right angles represents a change in angle between the filaments, the shear strain γ. The rotation of both filaments together represents local body rotation. We distinguish between the relative and absolute rotations of the filaments by separating the symmetrical and antisymmetrical parts of the displacement gradient tensor:

$$\partial_i u_j = e_{ij} + \omega_{ij}. \tag{5.92}$$

5.5.2 The Strain Tensor and the Rotation Tensor

The strain tensor e_{ij} is the symmetrical part of the displacement gradient tensor:

$$e_{ij} = \tfrac{1}{2}(\partial_i u_j + \partial_j u_i). \tag{5.93}$$

We illustrate application of the strain tensor by considering strain of two filaments at ninety degrees from one another, see Figure 5.7; the first filament has an orientation $\nu_i = \delta_{1i}$ and the second one has an orientation $\nu_i = \delta_{2i}$. The relative displacements e_i for these filaments are:

$$\begin{aligned} a_j &= e_{ij}\nu_i = e_{ij}\delta_{1i} = e_{1j} \\ a_j &= e_{ij}\nu_i = e_{ij}\delta_{2i} = e_{2j} \end{aligned} \tag{5.94}$$

These strains are shown in Figure 5.7; the strains e_{11} and e_{22} represent stretching and the strains $e_{12} = e_{21}$ represent a change in angle between the two filaments; the ends of the filaments move toward one another over a distance $e_{12} = e_{21}$.

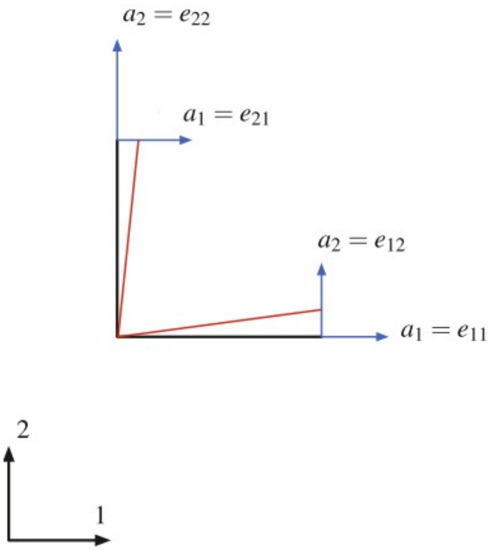

Figure 5.7: Strains in the 1,2 plane

The rotation tensor ω_{ij} represents the anti-symmetrical part of the displacement gradient tensor:

$$\omega_{ij} = \tfrac{1}{2}(\partial_i u_j - \partial_j u_i). \tag{5.95}$$

We use the relationship between the epsilon and delta tensors, (3.75) to write

$$\omega_{ij} = \tfrac{1}{2}(\partial_i u_j - \partial_j u_i) = \tfrac{1}{2}(\delta_{ip}\delta_{jq} - \delta_{iq}\delta_{jp})\partial_p u_q = \tfrac{1}{2}\varepsilon_{mij}\varepsilon_{mpq}\partial_p u_q = \tfrac{1}{2}\varepsilon_{mij}C_m, \tag{5.96}$$

where C_m is the curl of the displacement vector. It follows that the rotation tensor is fully defined by three components in three dimensions.

For the two-dimensional case, i and j range from 1 to 2 and the only non-zero component in (5.96) exists for $m = 3$,

$$\omega_{ij} = \tfrac{1}{2}\varepsilon_{3ij}C_3. \tag{5.97}$$

The rotation tensor in two dimensions is fully defined by a single scalar, ω, where $\omega = \frac{1}{2}C_3$, so that (5.96) reduces to

$$\omega_{ij} = \omega\varepsilon_{3ij}. \tag{5.98}$$

We demonstrate the effect of the rotation tensor ω_{ij} as for the strain tensor, by considering two orthogonal filaments. We obtain the following for the displacement vector:

$$\begin{aligned} v_i &= \delta_{1i} \qquad a_j = \omega_{ij}v_i = \omega_{ij}\delta_{1i} = \omega_{1j} = \tfrac{1}{2}\varepsilon_{31j}\omega_3 = (0, \omega) \\ v_i &= \delta_{2i} \qquad a_j = \omega_{ij}v_i = \omega_{ij}\delta_{2i} = \omega_{2j} = \tfrac{1}{2}\varepsilon_{32j}\omega_3 = (-\omega, 0), \end{aligned} \tag{5.99}$$

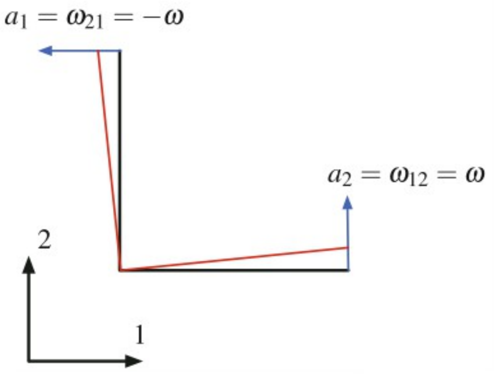

Figure 5.8: Rotation in the 1,2 plane.

or:

$$\begin{aligned}\omega_{ij} &= \tfrac{1}{2}\varepsilon_{1ij}C_1 + \tfrac{1}{2}\varepsilon_{2ij}C_2 + \tfrac{1}{2}\varepsilon_{3ij}C_3\\ \omega_{12} &= \tfrac{1}{2}\varepsilon_{312}C_3 = \tfrac{1}{2}C_3 = \omega\\ \omega_{21} &= \tfrac{1}{2}\varepsilon_{321}C_3 = -\tfrac{1}{2}C_3 = -\omega,\end{aligned} \tag{5.100}$$

The rotation tensor indeed represents true rotation, but only in an incremental sense; this is true only for small displacement gradients. That the rotation tensor represents pure rotation is in accordance with (5.96); the displacement vector is:

$$a_j = \omega_{ij}v_i = \tfrac{1}{2}\varepsilon_{mij}v_iC_m. \tag{5.101}$$

The vector a_i is the cross product of the vector v_i, which points in the direction of the filament, and the curl of the displacement vector; for the two-dimensional case, illustrated in Figure 5.8, the curl points in the 3-direction, the displacement is normal to v_i, and is always oriented the same way with respect to v_i.

As for the stress tensor, the first invariant of the strain tensor plays a special role in stress-strain relationships, for example in the theory of linear elasticity, discussed in Section 9.5. The first invariant e_{ii} is the volume strain

$$e_{ii} = \partial_i u_i, \tag{5.102}$$

which is the divergence of the displacement vector.

5.6 Integrability; The Compatibility Conditions

We consider a vector or tensor that is defined as the gradient of a scalar or a tensor of lower rank. It is generally not true that integration of the higher order tensor yields the lower order scalar or tensor, from which it was derived. This follows from the notion that the derived quantity has a greater number of components than the original one;

degrees of freedom have been added and these degrees of freedom must be subjected to constraints in order to make them compatible with the fewer number of degrees of freedom of the original quantity.

5.6.1 The Gradient of a Scalar

The conditions of integrability of a vector, defined as the gradient of a scalar,

$$v_i = -\partial_i \Phi, \tag{5.103}$$

are that the vector field is irrotational:

$$C_i = \varepsilon_{ijk}\partial_j v_k = \varepsilon_{ijk}\partial_j\partial_k \Phi = 0. \tag{5.104}$$

The three homogeneous equations (5.104), which the three components of v_i must satisfy, are sufficient to guarantee that equations (5.103) are integrable.

5.6.2 The Gradient of a Vector

A similar case applies if a tensor, for example the strain tensor e_{ij}, is derived from the gradient of a vector field. We obtain the integrability conditions as for the case of a vector. We first consider the definition of the displacement gradient tensor, in terms of the displacement vector u_i,

$$\alpha_{km} = \partial_k u_m. \tag{5.105}$$

Integrability of these equations requires that

$$\varepsilon_{ijk}\partial_j \alpha_{km} = \varepsilon_{ijk}\partial_j\partial_k u_m = 0. \tag{5.106}$$

We obtain the integrability conditions for the strains by using the definitions of the strain and rotation tensors:

$$\varepsilon_{ijk}\partial_j \alpha_{km} = \varepsilon_{ijk}\partial_j e_{km} + \varepsilon_{ijk}\partial_j \omega_{km}. \tag{5.107}$$

We use (5.96) to rewrite the trailing term as

$$\varepsilon_{ijk}\partial_j \omega_{km} = \tfrac{1}{2}\varepsilon_{ijk}\varepsilon_{kmn}\partial_j C_n = \tfrac{1}{2}(\delta_{im}\delta_{jn} - \delta_{in}\delta_{jm})\partial_j C_n = \tfrac{1}{2}\delta_{im}\partial_j C_j - \tfrac{1}{2}\partial_m C_i, \tag{5.108}$$

where C_n is the curl,

$$C_n = \varepsilon_{npq}\partial_p u_q. \tag{5.109}$$

The curl is divergence-free, as is seen from (5.109) by taking the divergence, so that (5.108) reduces to

$$\varepsilon_{ijk}\partial_j \omega_{km} = -\tfrac{1}{2}\partial_m C_i. \tag{5.110}$$

We combine (5.107) with (5.110),

$$\varepsilon_{ijk}\partial_j \alpha_{km} = \varepsilon_{ijk}\partial_j e_{km} - \tfrac{1}{2}\partial_m C_i. \tag{5.111}$$

These equations represent nine conditions for the six independent components of the strain tensor and the three components of the curl; we eliminate the curl in what follows.

Equation (5.111) represents a tensor with free indices i and m, which we call A_{im} and take the curl, $\varepsilon_{pnm}\partial_n A_{im}$, which gives:

$$\varepsilon_{pnm}\partial_n A_{im} = \varepsilon_{pnm}\varepsilon_{ijk}\partial_j\partial_n \alpha_{km} = \varepsilon_{ijk}\varepsilon_{pnm}\partial_j\partial_n\partial_k u_m = 0, \tag{5.112}$$

because $\partial_j\partial_k = \partial_k\partial_j$. We replace A_{im} by (5.111),

$$\varepsilon_{pnm}\partial A_{im} = \varepsilon_{pnm}\varepsilon_{ijk}\partial_n\partial_j\varepsilon_{km} - \tfrac{1}{2}\varepsilon_{pnm}\partial_m\partial_n C_i = 0, \tag{5.113}$$

so that:

$$\varepsilon_{pnm}\varepsilon_{ijk}\partial_n\partial_j e_{km} = \tfrac{1}{2}\varepsilon_{pnm}\partial_m\partial_n C_i = 0. \tag{5.114}$$

These equations are the nine integrability conditions that must be imposed on the components of the strain tensor. If these conditions are met, integration of the strains produces a set of unique displacements. Three of the nine equations occur twice as a result of the symmetry of both the operators $\partial_i\partial_j$ and the strain tensor e_{ij}. We demonstrate that equations (5.114) are symmetrical in terms of the indices p and i by contracting with ε_{piq} and showing that the result is zero:

$$\begin{aligned}\varepsilon_{piq}\varepsilon_{pmn}\varepsilon_{ijk}\partial_n\partial_j e_{km} &= (\delta_{im}\delta_{qn} - \delta_{in}\delta_{qm})\varepsilon_{ijk}\partial_n\partial_j e_{km} \\ &= \varepsilon_{mjk}\partial_q\partial_j e_{km} - \varepsilon_{ijk}\partial_i\partial_j e_{kq}\end{aligned} \tag{5.115}$$

This represents three equations, one for each of the three possible values of q. The first term is zero because of the symmetry of e_{km} and the second term is zero because of the symmetry of $\partial_i\partial_j$, so that

$$\varepsilon_{piq}\varepsilon_{pmn}\varepsilon_{ijk}\partial_n\partial_j e_{km} = 0. \tag{5.116}$$

The equations obtained from (5.114) are thus indeed symmetrical in terms of the indices i and p. The nine equations (5.114) for the strain tensor components,

$$\varepsilon_{pmn}\varepsilon_{ijk}\partial_n\partial_j e_{km} = 0, \tag{5.117}$$

reduce to the six independent ones, obtained by letting p and i assume the following values:

$$\begin{array}{ll} p = 1 & i = 1, 2, 3 \\ p = 2 & i = 2, 3 \\ p = 3 & i = 3. \end{array} \tag{5.118}$$

These equations are the compatibility conditions for the strain tensor for small displacement gradients.

Problem 5.3

Work out the compatibility equations for plane strain (i.e., for the case that $e_{m3} = 0, m = 1, 2, 3$.

Problem 5.4

Work out the six compatibility equations in terms of $(x, y, z) \equiv x_i$ and compare your result with the equations as found in the literature.

5.7 The Velocity Gradient Tensor

The velocity gradient tensor plays a role in fluids similar to that of the displacement gradient tensor for solids. The velocity gradient tensor is:

$$\dot{\alpha}_{ij} = \partial_i v_j, \tag{5.119}$$

where v_j is the velocity. As for the displacements, we separate the velocity gradient tensor into its symmetrical and anti-symmetrical parts:

$$\dot{\alpha}_{ij} = \tfrac{1}{2}(\dot{\alpha}_{ij} + \dot{\alpha}_{ji}) + \tfrac{1}{2}(\dot{\alpha}_{ij} - \dot{\alpha}_{ji}). \tag{5.120}$$

The symmetrical part is $\dot{e}_{ij}$:

$$\dot{e}_{ij} = \tfrac{1}{2}(\partial_i v_j + \partial_j v_i). \tag{5.121}$$

The anti-symmetrical part is (compare (5.96))

$$\dot{\omega}_{ij} = \tfrac{1}{2}(\partial_i v_j - \partial_j v_i) = \varepsilon_{ijm} C_m, \tag{5.122}$$

where C_m is the curl. The divergence of the velocity field plays a role similar to that of the volume strain for the displacement gradient and is sometimes represented as D

$$D = \partial_i v_i. \tag{5.123}$$

17.04.2020

b) epdf.pub_the foundations of geometry.djvu
David Hilbert   188 pages.

obs: look:
Pág. 11 (29)
§ 5. Group III
Axiom of Parallels
(Euclid's axiom).

I) Books_69-2020/
The Mobius Strip/
Clifford A. Pickover... Mobius... djvu
other copy:
1560258268.pdf

145) 20200418 ✡

A) Books_70_2020/
Applications of Vector Analysis and Complex Variables in Engineering/
sanet_st-3030411672.pdf

obs. Livro todo

A1) Pág. 51 (63)
Coordinate Transformations:

data 18 04 2020

Definition of Vectors and Tensors

Figure 5.1 Coordinate Transformation

Transformation of Cartesian A into Cartesian System B.

We consider a Cartesian B with coordinate axes ($\tilde{1}, \tilde{2}, \tilde{3}$), representing a point in this system as $\tilde{x}_i$.

The origin of B in terms of the original coordinate system A (1, 2, 3) is given by the coordinates b_i.

(figure 5.1)

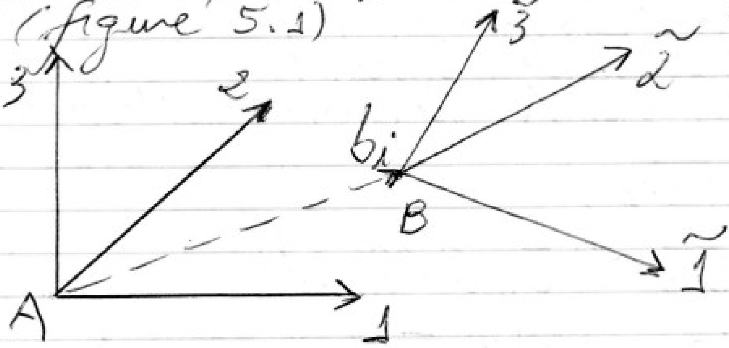

Both coordinate system are Cartesian, and the scales in both systems are the same

the vectors $\underset{m}{\tau}^i$, defined as

$$\underset{m}{\tau}^i = \frac{\partial x_1}{\partial \tilde{x}_m} \quad i = 1, 2, 3 \quad m = 1, 2, 3$$

are unit vectors tangent to the coordinate directions $\tilde{i}$ (compare (2.5))

data 18 04 2020

with $\tilde{x}_m$ replacing the arc length along the curve, s.

We define similar vectors that are tangent to the coordinates in system A in terms of the coordinates in system B.

We call these unit tangent vectors $\overset{m}{\underset{i}{\gamma}}$:

$$\overset{m}{\underset{i}{\gamma}} = \frac{\partial \tilde{x}_m}{\partial x_i}$$

We express the original coordinates of a point x_i in terms of the new coordinates of that same point and vice-versa by integrating the latter two equations:

$$x_i = \overset{i}{\underset{m}{\tau}} \tilde{x}_m + b_i,$$

where b_i represents the origin of B in terms of x_i, and

$$\tilde{x}_m = \overset{m}{\underset{i}{\gamma}} x_i + \tilde{b}_m,$$

where $\tilde{b}_m$ represents the origin of the original system A in terms of the new coordinates.

data 18.04.2020

The vectors $\underset{m}{\tau}^i$ are mutually orthogonal unit vectors,

$$\underset{m}{\tau}^i \underset{n}{\tau}^i = \delta_{mn}.$$

A similar equation holds for unit vectors $\overset{m}{\gamma}_i$,

$$\overset{m}{\gamma}_i \overset{n}{\gamma}_i = \delta_{mn}.$$

We obtain a relationship between vectors $\underset{m}{\tau}^i$ and $\overset{m}{\gamma}_i$ from (5.1) and (5.2)

$$\underset{m}{\tau}^i \overset{n}{\gamma}_i = \frac{\partial x_i}{\partial \tilde{x}_m} \frac{\partial \tilde{x}_n}{\partial x_i} = \frac{\partial \tilde{x}_n}{\partial \tilde{x}_m} = \delta_{mn}$$

The moduli of $\underset{m}{\tau}^i$ and of $\overset{m}{\gamma}_i$ are 1;

we combine this with the latter equation:

$$\underset{m}{\tau}^i = \overset{m}{\gamma}_i.$$

We introduce a matrix a_{ij},

landaia

18.04.2020
8

$$a_{im} = \tau^{i}_{m} = \gamma^{m}_{i} = \frac{\partial x_i}{\partial x_m} = \frac{\partial \tilde{x}_m}{\partial x_i}$$

and write (5.3) and (5.4) as

$$x_i = a_{ij}\tilde{x}_j + b_i,$$

and

$$\tilde{x}_i = a_{ji}x_j + \tilde{b}_i.$$

We observe from (5.9) that

$$a_{im}a_{in} = \frac{\partial x_i}{\partial \tilde{x}_m}\frac{\partial \tilde{x}_n}{\partial x_i} = \delta_{mn},$$

and that

$$a_{im}a_{jm} = \frac{\partial \tilde{x}_m}{\partial x_i}\frac{\partial x_j}{\partial \tilde{x}_m} = \delta_{ij}.$$

A2)
~~B)~~ mesmo livro
Pág. 91 (103)
8.3.2 Transformation of Vectors

A3) mesmo livro
Pág. 102 (114) →

data 18 04 2020

✡ 8.4 Irrotational and Divergence-Free Vector Fields

$$\frac{\partial w}{\partial \bar{z}} = 0 \quad \rightarrow \text{light over curve.}$$

w = velocity potential
(complex variables
z and $\bar{z}$).

Pág. 103(115)

Equations (8.133) and (8.134) are equivalent complex forms of the

Cauchy-Riemann Equations.

B) Books_70_2020/
Fundamentals of Geometry Construction the Math behind the Cad/
sanet_st_3030431304.pdf

Obs.: livro todo.

B1) Pág. 07(25)

1.2.3 Types of Coordinates Systems.

vectors are given by (5.31) and (5.32) on page 55,

$$\underset{m}{\tau}^k = \frac{\partial x_k}{\partial \tilde{x}_m} \tag{8.39}$$

$$\overset{m}{\gamma}_k = \frac{\partial \tilde{x}_m}{\partial x_k}, \tag{8.40}$$

where

$$\tilde{x}_m = (\bar{z}, z) \qquad x_m = (x, y). \tag{8.41}$$

The covariant base vectors are

$$\overset{1}{\gamma}_k = \frac{\partial \tilde{x}_1}{\partial x_i} = \left(\frac{\partial \bar{z}}{\partial x}, \frac{\partial \bar{z}}{\partial y} \right) = (1, -\mathrm{i}) \tag{8.42}$$

$$\overset{2}{\gamma}_k = \frac{\partial \tilde{x}_2}{\partial x_i} = \left(\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y} \right) = (1, \mathrm{i}). \tag{8.43}$$

The second base vector $\overset{2}{\gamma}_k$ is related to the first base vector $\overset{1}{\gamma}_k$; it is the complex conjugate of $\overset{1}{\gamma}_k$,

$$\overset{2}{\gamma}_k = \overline{\overset{1}{\gamma}}_k. \tag{8.44}$$

We obtain similar expressions for the contravariant base vectors

$$\underset{1}{\tau}^m = \frac{\partial x_m}{\partial \tilde{x}_1} = \left(\frac{\partial x}{\partial \bar{z}}, \frac{\partial y}{\partial \bar{z}} \right) = \tfrac{1}{2}(1, \mathrm{i}) \tag{8.45}$$

$$\underset{2}{\tau}^k = \frac{\partial x_m}{\partial \tilde{x}_2} = \left(\frac{\partial x}{\partial z}, \frac{\partial y}{\partial z} \right) = \tfrac{1}{2}(1, -\mathrm{i}). \tag{8.46}$$

As for the covariant bases, the second contravariant base vector is the complex conjugate of the first one:

$$\underset{2}{\tau}^k = \overline{\underset{1}{\tau}^k}. \tag{8.47}$$

The covariant and contravariant base vectors are related:

$$\overset{m}{\gamma}_k = 2\overline{\underset{m}{\tau}^k}. \tag{8.48}$$

8.3.2 Transformation of Vectors

We use the expressions for the base vectors to transform vectors and tensors. We obtain the contravariant components from

$$\tilde{v}^1 = \overset{1}{\gamma}_j v_j = v_1 - \mathrm{i} v_2 \tag{8.49}$$

$$\tilde{v}^2 = \overset{2}{\gamma}_j v_j = v_1 + \mathrm{i} v_2, \tag{8.50}$$

and observe that, as for the independent variables $\bar{z}$ and z, the one component is the complex conjugate of the other. We use a single symbol, w, to represent both

$$w = \tilde{v}^1 = v_x - \mathrm{i}v_y \tag{8.51}$$

$$\bar{w} = \tilde{v}^2 = v_x + \mathrm{i}v_y, \tag{8.52}$$

and find the expressions for the covariant components of a vector in a similar fashion,

$$\tilde{v}_1 = \underset{1}{\tau}{}^j v_j = \tfrac{1}{2}(v_x + \mathrm{i}v_y) = \tfrac{1}{2}\overline{w} \tag{8.53}$$

$$\tilde{v}_2 = \underset{2}{\tau}{}^j v_j = \tfrac{1}{2}(v_x - \mathrm{i}v_y) = \tfrac{1}{2}w. \tag{8.54}$$

The modules of the vector, $v_i v_i$, is an invariant, which transforms as

$$v_i v_i = \tilde{v}_k \tilde{v}^k = \tfrac{1}{2}w\overline{w} + \tfrac{1}{2}w\overline{w} = w\overline{w}, \tag{8.55}$$

which indeed equals $v_x^2 + v_y^2$ as follows from (8.51). The differential operator is a covariant vector;

$$\tilde{\partial}_m = \left(\frac{\partial}{\partial \bar{z}}, \frac{\partial}{\partial z}\right), \tag{8.56}$$

and its contravariant components are:

$$\tilde{\partial}^m = 2\left(\frac{\partial}{\partial z}, \frac{\partial}{\partial \bar{z}}\right). \tag{8.57}$$

As for non-Cartesian coordinate systems, the natural way to write the gradient of a scalar function is in terms of its covariant components,

$$\frac{\partial F}{\partial \bar{z}} = \tfrac{1}{2}\left(\frac{\partial F}{\partial x} + \mathrm{i}\frac{\partial F}{\partial y}\right) \tag{8.58}$$

$$\frac{\partial F}{\partial z} = \tfrac{1}{2}\left(\frac{\partial F}{\partial x} - \mathrm{i}\frac{\partial F}{\partial y}\right). \tag{8.59}$$

We are now in a position to transform the divergence of the velocity vector v_i by contracting the covariant components of the gradient with the contravariant components of the velocity vector; this is necessary to obtain the invariant, i.e., to write an expression in terms of the new coordinates in such a way that it is equivalent to the corresponding one in the Cartesian system,

$$\partial_j v_j = \tilde{\partial}_j \tilde{v}^j = \frac{\partial w}{\partial \bar{z}} + \frac{\partial \bar{w}}{\partial z} = -\gamma. \tag{8.60}$$

8.3.3 Transformation of Tensors

We transform a tensor A_{mn} into a purely contravariant tensor:

$$\tilde{A}^{pq} = \overset{p}{\gamma}_m \overset{q}{\gamma}_n A_{mn}. \tag{8.61}$$

There are only two independent components of a tensor of the second rank; the component $\tilde{A}^{21}$ is the complex conjugate of the component $\tilde{A}^{12}$; we apply (8.44):

$$\tilde{A}^{21} = \overset{2}{\gamma}_m \overset{1}{\gamma}_n A_{mn} = \overline{\overset{1}{\gamma}_m \overset{2}{\gamma}_n} A_{mn} = \overline{\tilde{A}^{12}}. \tag{8.62}$$

The component $\tilde{A}^{22}$ is the complex conjugate of $\tilde{A}^{11}$:

$$\tilde{A}^{22} = \overset{2}{\gamma}_m \overset{2}{\gamma}_n A_{mn} = \overline{\overset{1}{\gamma}_m \overset{1}{\gamma}_n} A_{mn} = \overline{\tilde{A}^{11}}. \tag{8.63}$$

The purely covariant tensor components are related to the contravariant ones; we use (8.48),

$$\tilde{A}_{mn} = \underset{m}{\tau}^p \underset{n}{\tau}^q A_{pq} = \tfrac{1}{4} \overline{\overset{m}{\gamma}_p \overset{n}{\gamma}_q} A_{pq} = \tfrac{1}{4} \overline{\tilde{A}^{mn}}. \tag{8.64}$$

We obtain similar relations between the mixed transformed tensor components

$$\tilde{A}^m{}_n = \overset{m}{\gamma}_p \underset{n}{\tau}^q A_{pq} = \overline{\underset{m}{\tau}^p \overset{n}{\gamma}_q} A_{pq} = \overline{\tilde{A}_m{}^n}. \tag{8.65}$$

Only two independent tensor components exist, $\tilde{A}^{11}$ and $\tilde{A}^{12}$; we obtain all the other ones from these two by applying equations (8.61) through (8.65).

Problem 8.1
Demonstrate that $\tilde{A}^1{}_1 = \tilde{A}_{21}$ and $\tilde{A}^1{}_2 = \tilde{A}_{22}$.

8.3.4 A Combined Kronecker Delta and Alternating Tensor: The Combination Matrix

Working in complex space with vectorial properties such as divergence and curl, requires that the Kronecker delta and the epsilon tensor be transformed. There are four possible choices for each tensor; we begin by obtaining the transformed tensor $\tilde{\delta}^m{}_n$. Application of the transformation rule gives

$$\tilde{\delta}^m{}_n = \overset{m}{\gamma}_j \underset{n}{\tau}^k \delta_{jk} = \overset{m}{\gamma}_j \underset{n}{\tau}^j = \delta_{mn}. \tag{8.66}$$

We find, similarly

$$\tilde{\delta}_m{}^n = \underset{m}{\tau}^j \overset{n}{\gamma}_k \delta_{jk} = \underset{m}{\tau}^j \overset{n}{\gamma}_j = \delta_{mn}. \tag{8.67}$$

The transformed Kronecker delta with both components contravariant becomes

$$\tilde{\delta}^{mn} = \overset{m}{\gamma}_k \overset{n}{\gamma}_j \delta_{ij} = \overset{m}{\gamma}_k \overset{n}{\gamma}_k = g^{mn} = \begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix}, \tag{8.68}$$

and with both components covariant:

$$\tilde{\delta}_{mn} = \underset{m}{\tau}^k \underset{n}{\tau}^j \delta_{ij} = \underset{m}{\tau}^j \underset{n}{\tau}^j = g_{mn} = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}. \tag{8.69}$$

The transformed Kronecker delta is called the metric tensor; the metric tensor equals the Kronecker delta for the case of mixed covariant/contravariant components.

We transform the epsilon tensor in a similar manner. Again, there are four possible combinations; we consider them in the same sequence as for the Kronecker delta. We obtain for $\tilde{\varepsilon}^m{}_n$

$$\tilde{\varepsilon}^m{}_n = \overset{m}{\gamma}_j \underset{n}{\tau}^k \varepsilon_{jk}. \tag{8.70}$$

The individual components are:

$$\begin{aligned} \tilde{\varepsilon}^1{}_1 &= \overset{1}{\gamma}_1 \underset{1}{\tau}^2 - \overset{1}{\gamma}_2 \underset{1}{\tau}^1 = \mathrm{i}/2 - (-\mathrm{i})/2 = \mathrm{i} \\ \tilde{\varepsilon}^1{}_2 &= \overset{1}{\gamma}_1 \underset{2}{\tau}^2 - \overset{1}{\gamma}_2 \underset{2}{\tau}^1 = -\frac{1}{2}\mathrm{i} + \frac{1}{2}\mathrm{i} = 0 \\ \tilde{\varepsilon}^2{}_2 &= \overline{\tilde{\varepsilon}^1{}_1} = -\mathrm{i}. \end{aligned} \tag{8.71}$$

or

$$\tilde{\varepsilon}^m{}_n = \begin{pmatrix} \mathrm{i} & 0 \\ 0 & -\mathrm{i} \end{pmatrix}. \tag{8.72}$$

We obtain the components of $\tilde{\varepsilon}_m{}^n$ similarly:

$$\tilde{\varepsilon}_m{}^n = \underset{m}{\tau}^j \overset{n}{\gamma}_k \varepsilon_{jk} = \overline{\tilde{\varepsilon}^m{}_n}, \tag{8.73}$$

which gives

$$\begin{aligned} \tilde{\varepsilon}_1{}^1 &= -\mathrm{i}/2 + (-\mathrm{i})/2 = -\mathrm{i} \\ \tilde{\varepsilon}_1{}^2 &= \tilde{\varepsilon}_2{}^1 = 0 \\ \tilde{\varepsilon}_2{}^2 &= \mathrm{i}, \end{aligned} \tag{8.74}$$

or

$$\tilde{\varepsilon}^n{}_m = \begin{pmatrix} -\mathrm{i} & 0 \\ 0 & \mathrm{i} \end{pmatrix}. \tag{8.75}$$

The purely covariant components of the epsilon tensor are

$$\tilde{\varepsilon}_{mn} = \underset{m}{\tau}^i \underset{n}{\tau}^j \varepsilon_{ij}, \tag{8.76}$$

which gives

$$\begin{aligned} \tilde{\varepsilon}_{11} &= \overline{\tilde{\varepsilon}_{22}} = \underset{1}{\tau}^1 \underset{1}{\tau}^2 - \underset{1}{\tau}^2 \underset{2}{\tau}^1 = 0 \\ \tilde{\varepsilon}_{12} &= \underset{1}{\tau}^1 \underset{2}{\tau}^2 - \underset{1}{\tau}^2 \underset{2}{\tau}^1 = -\frac{1}{4}\mathrm{i} - \frac{1}{4}\mathrm{i} = -\mathrm{i}/2 \\ \tilde{\varepsilon}_{21} &= \overline{\tilde{\varepsilon}_{12}} = \mathrm{i}/2, \end{aligned} \tag{8.77}$$

or

$$\tilde{\varepsilon}_{mn} = \begin{pmatrix} 0 & -\mathrm{i}/2 \\ \mathrm{i}/2 & 0 \end{pmatrix}. \tag{8.78}$$

The purely contravariant components $\tilde{\varepsilon}^{mn}$ are:

$$\tilde{\varepsilon}^{mn} = \overset{m}{\gamma}_k \overset{n}{\gamma}_j \varepsilon_{ij} = 4\overline{\tilde{\varepsilon}_{mn}}, \tag{8.79}$$

which gives

$$\begin{aligned} \tilde{\varepsilon}^{11} &= \tilde{\varepsilon}^{22} = 0 \\ \tilde{\varepsilon}^{12} &= 2\mathrm{i} \\ \tilde{\varepsilon}^{21} &= -2\mathrm{i}, \end{aligned} \tag{8.80}$$

or

$$\tilde{\varepsilon}^{mn} = \begin{pmatrix} 0 & 2\mathrm{i} \\ -2\mathrm{i} & 0 \end{pmatrix}. \tag{8.81}$$

The Combination Matrix

We construct a useful matrix by combining the Kronecker delta and the epsilon tensor, which we call the combination matrix. There are two possible representations for this matrix in Cartesian space, κ_{mn} and K_{mn}, that we define as:

$$\kappa_{mn} = \delta_{mn} - \mathrm{i}\varepsilon_{mn}, \tag{8.82}$$

and

$$K_{mn} = \delta_{mn} + \mathrm{i}\varepsilon_{mn} = \bar{\kappa}_{mn}. \tag{8.83}$$

We add to the transformed Kronecker delta the corresponding epsilon tensor times i for κ and $-\mathrm{i}$ for K. The mixed contravariant/covariant components are

$$\tilde{\kappa}^{m}{}_{n} = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} \qquad \tilde{K}^{m}{}_{n} = \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}, \tag{8.84}$$

the mixed covariant/contravariant ones are

$$\tilde{\kappa}_{m}{}^{n} = \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix} \qquad \tilde{K}_{m}{}^{n} = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}, \tag{8.85}$$

and the purely contravariant components are

$$\tilde{\kappa}^{mn} = \begin{pmatrix} 0 & 4 \\ 0 & 0 \end{pmatrix} \qquad \tilde{K}^{mn} = \begin{pmatrix} 0 & 0 \\ 4 & 0 \end{pmatrix}. \tag{8.86}$$

The components, finally, of the purely covariant components are

$$\tilde{\kappa}_{mn} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \qquad \tilde{K}_{mn} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}. \tag{8.87}$$

The combination matrices are not tensors; they do not obey the transformation rules (8.62) through (8.65), even though δ_{ij} and ε_{ij} do.

$ (' %'& !96 : : !

Questions:

- Determine the function w that has the divergence and curl given by ω.
- Determine the areal integral of ω over area $\mathscr{A}$ by integration.
- Determine the same integral, but now by applying (8.114).
- Determine the tangential and normal components of w along $\mathscr{C}$.

8.4 Irrotational and Divergence-Free Vector Fields

If the vector field is both irrotational and divergence-free, (8.23) reduces to

$$\frac{\partial w}{\partial \bar{z}} = 0, \tag{8.125}$$

so that w is a function of z only, i.e.,

$$w = w(z). \tag{8.126}$$

Because the flow is both irrotational and divergence-free, we can represent the complex velocity function either as the gradient of a real potential, or as the curl of a real scalar stream function. If we express w in terms of a velocity potential, we write the gradient of the potential in terms of derivatives with respect to the complex variables z and $\bar{z}$

$$\begin{aligned} w = v_x - \mathrm{i}v_y = -\left[\frac{\partial \Phi}{\partial x} - \mathrm{i}\frac{\partial \Phi}{\partial y}\right] = - & \left[\frac{\partial \Phi}{\partial z}\frac{\partial z}{\partial x} + \frac{\partial \Phi}{\partial \bar{z}}\frac{\partial \bar{z}}{\partial x}\right] \\ & + \mathrm{i}\left[\frac{\partial \Phi}{\partial z}\frac{\partial z}{\partial y} + \frac{\partial \Phi}{\partial \bar{z}}\frac{\partial \bar{z}}{\partial y}\right], \end{aligned} \tag{8.127}$$

or, with $\partial z/\partial x = 1, \partial z/\partial y = \mathrm{i}$ and $\partial \bar{z}/\partial x = 1, \partial \bar{z}/\partial y = -\mathrm{i}$

$$w = -2\frac{\partial \Phi}{\partial z}. \tag{8.128}$$

As an alternative, we write the velocity vector as minus the curl of a scalar stream function,

$$v_k = -\varepsilon_{ij}\partial_j \Psi, \tag{8.129}$$

so that

$$w = v_x - \mathrm{i}v_y = -\frac{\partial \Psi}{\partial y} - \mathrm{i}\frac{\partial \Psi}{\partial x} = -\mathrm{i}\left[\frac{\partial \Psi}{\partial x} - \mathrm{i}\frac{\partial \Psi}{\partial y}\right]. \tag{8.130}$$

Comparing the latter equation with (8.127) and (8.128), we see that we may express w in terms of the derivative of Ψ with respect to z,

$$w = -2\mathrm{i}\frac{\partial \Psi}{\partial z}. \tag{8.131}$$

The representations of w in terms of the derivatives of Φ and Ψ with respect to z gives us several possibilities to represent w:

$$w = w(z) = -2\frac{\partial\Phi}{\partial z} = -2\mathrm{i}\frac{\partial\Psi}{\partial z} = -\left(\frac{\partial\Phi}{\partial z} + \mathrm{i}\frac{\partial\Psi}{\partial z}\right). \tag{8.132}$$

The final expression in the latter equation distributes the contribution to w equally between a real potential and a real stream function. We observe from the third and fourth expressions in the latter equation that

$$\frac{\partial\Phi}{\partial z} = \mathrm{i}\frac{\partial\Psi}{\partial z}, \tag{8.133}$$

and take the complex conjugate of both sides of this equation. The real functions Φ and Ψ are not affected, but i changes to $-\mathrm{i}$ and z to $\bar{z}$:

$$\frac{\partial\Phi}{\partial\bar{z}} = -\mathrm{i}\frac{\partial\Psi}{\partial\bar{z}}. \tag{8.134}$$

Equations (8.133) and (8.134) are equivalent complex forms of the Cauchy-Riemann equations; the latter equations are obtained in real form by separating the real and imaginary parts of these equations; both equations lead to the same two real equations:

$$\begin{aligned}\frac{\partial\Phi}{\partial x} &= \frac{\partial\Psi}{\partial y}\\ \frac{\partial\Phi}{\partial y} &= -\frac{\partial\Psi}{\partial x}.\end{aligned} \tag{8.135}$$

Integration of (8.134) with respect to $\bar{z}$ gives

$$\Phi + \mathrm{i}\Psi = f(z), \tag{8.136}$$

where $f(z)$ is a function of z only, which we call the complex potential and represent by the letter Ω,

$$\Omega = \Omega(z) = \Phi + \mathrm{i}\Psi. \tag{8.137}$$

We represent w as minus the derivative with respect to z of this complex potential:

$$w = -\frac{d\Omega}{dz}. \tag{8.138}$$

This is an ordinary derivative; Ω is a function of z only, in contrast to both the potential Φ and the stream function Ψ, which are real functions and must be functions of both z and $\bar{z}$.

We observe from (8.135) that the gradients of the velocity potential and the stream function are mutually orthogonal; their dot product is zero. Since the gradient of the velocity potential is parallel to the velocity, the gradient of the stream function is at right angles to the flow. Both the contours $\Psi = const$ and the velocity are normal to the gradient of the stream function; the velocity vector must be tangent to the contours $\Psi = const$.

8.4.1 Summary

The real and imaginary parts of a complex potential $\Omega(z)$, which is a function of z only, i.e., does not contain $\bar{z}$ in its expression, produces a complex derivative $d\Omega/dz$, which is minus the complex form of a velocity field, w. This velocity field is both divergence-free and irrotational. The real and imaginary parts of the complex potential are harmonic functions, i.e., they satisfy Laplace's equation, as follows from (8.95) and (8.102).

8.5 Functions of a Single Complex Variable

Functions of a single complex variable that can be differentiated an infinite number of times in a certain region are called *analytic* or *holomorphic* inside that region. We use the term holomorphic to identify this property for functions of a single complex variable; we use the term analytic if the function depends on both z and $\bar{z}$ and can be differentiated an infinite number of times with respect to either of the independent variables z and $\bar{z}$. Note that we call these variables independent, which they are only in the strict sense of a coordinate transformation; the special property of the number i, however, makes it possible to determine the second variable once the first is known.

We allow the divergence to be non-zero at isolated points and along lines, and possibly inside isolated domains in the area considered. In such cases the stream function cannot be single-valued, and will be singular at isolated points. For very special cases rotation may be permitted at isolated points, along lines, and in isolated domains. In this case the potential function will not be single-valued at the points or singular at isolated points. A singularity of a holomorphic function is defined as a point where the function cannot be differentiated an infinite number of times without producing an infinite value at the singular point.

Working with complex variables requires certain formulas to understand the meaning of the various forms of complex functions. We present a brief summary of such formulas below without giving formal proof of the equations; the reader is referred to the multitude of texts on this topic.

The complex variable z can be represented in terms of Cartesian coordinates (x,y) and in terms of radial coordinates (r,θ),

$$z = x + \mathrm{i}y = r(\cos\theta + \mathrm{i}\sin\theta). \tag{8.139}$$

A useful formula was developed by Euler and is obtained from the Taylor series expansion of the exponential function of $(\mathrm{i}\theta)$ about $\theta = 0$:

$$\mathrm{e}^{\mathrm{i}\theta} = 1 + \mathrm{i}\theta - \frac{\theta^2}{2!} - \mathrm{i}\frac{\theta^3}{3!} + \frac{\theta^4}{4!} + \mathrm{i}\frac{\theta^5}{5!} + \cdots . \tag{8.140}$$

We separate the real and imaginary parts:

$$\Re\mathrm{e}^{\mathrm{i}\theta} = 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} + \cdots , \tag{8.141}$$

data 18.04.2020
S T Q Q S [✓ S] D

✡ 8.4 Irrotational and Divergence-Free Vector Fields

$$\frac{\partial w}{\partial \bar{z}} = 0$$

light over curve.

w = velocity potential (complex variables z and $\bar{z}$).

Pág. 103 (115)

Equations (8.133) and (8.134) are equivalent complex forms of the

Cauchy-Riemann Equations.

B) Books_70_2020/
Fundamentals of Geometry Construction the Math behind the Cad/
sanet_st_3030431304.pdf

Obs.: Livro todo.

B1) Pág. 07(25)

1.2.3 Types of Coordinates Systems.

Jandaia

data 18.04.2020

IMPORTANT:

Criar o meu sistema de coordenadas
(funções representam cada vetor)

→ CURVE VECTOR (provisório)

→ MUSICAL CURVE VECTOR.

voltando:

- Polar Coordinates,
- Cylindrical coordinates,
- spherical coordinates,
- Absolute coordinates e
- world coordinates (x, y, z) versus relative coordinates (x', y', z').

B2) Pag. 13 (31)
Vectors

B3) Matrices
Pag. 17 (35)

B4) Pag. 80 (96)
3.1.3 Distance of a Point to a Plane
(cartesiano)

;)*+,)))-) ' $ $ '$ $ ' 0!
: ! < 1
(/ $ 810 ! 1

Similar to the 2D rectangle that was constructed, a rectangular parallelepiped is created using the 3D coordinate system by establishing coordinate values for each vertex, and then inserting the appropriate edges, as shown in Fig. 1.8.

This coordinate system is used in multiview drawings and 3D modeling, using CAD tools.

1.2.2 *Right-Hand Rule*

The right-hand rule is used to determine the positive direction of the axes; it defines the X-, Y-, and Z-axes as well as the positive and negative directions of rotation on each axis.

The simplest way to remember the right-hand rule is to first make a fist with your right hand—hence the rule name—ensuring that your thumb points upward. The direction in which your thumb points is the positive direction of the Z-axis. Now straighten your index finger so that it points straight at right angles with your thumb, thereby defining the X-axis. Finally, straighten your middle finger at right angles with both your thumb and your index finger, thereby defining the Y-axis.

It is noteworthy that, opposite to the right-hand rule, there exists a left-hand rule, which defines all of the previous axes in the same way as the right-hand rule, only *reflected*. While the left-hand rule is used in some situations to describe the coordinate axes, in this book we will only employ the conventional right-hand rule for the sake of consistency, simplicity, and while following established convention.

1.2.3 *Types of Coordinate Systems*

1.2.3.1 Polar Coordinates

Polar coordinates are used to locate points in the plane; they specify a distance ρ from a prescribed point, denoted the origin $O(0, 0)$, and an angle θ from a prescribed line, usually denoted the X-axis. Figure 1.9 shows a point P in the XY-plane, a distance ρ from the origin O, segment $\overline{OP}$ making an angle of θ with the X-axis, measured ccw[1] in the positive direction. Polar coordinates are commonly used by CAD systems to locate points because of their inherent simplicity.

The relations between polar and Cartesian coordinates are straightforward:

$$\rho = \sqrt{x^2 + y^2}, \quad \theta = \arctan(x, y) \tag{1.1a}$$

$$x = r\cos\theta, \quad y = r\sin\theta \tag{1.1b}$$

where $\arctan(x, y)$ is the *single-valued* arctan function, which returns the unique value of θ whose tangent is x/y, the correct quadrant being indicated by the sign of each x and y.

---

[1] Usual abbreviation for "counterclockwise"

Fig. 1.9 Polar coordinates

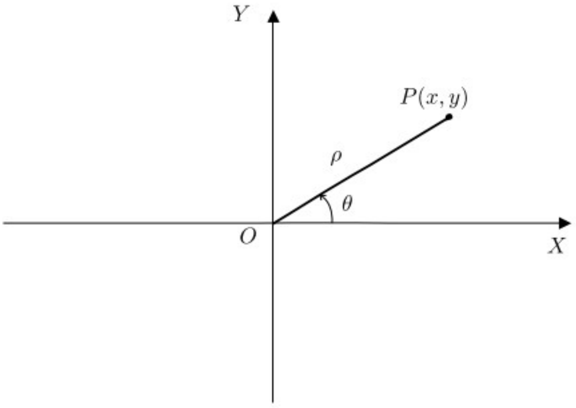

Fig. 1.10 Cylindrical coordinates

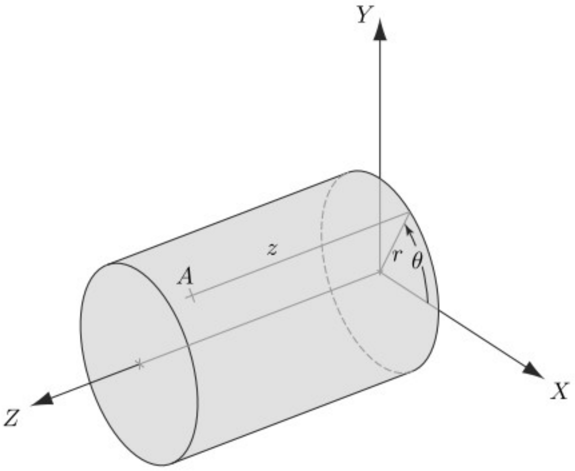

1.2.3.2 Cylindrical Coordinates

Cylindrical coordinates locate a point on the surface of a cylinder by specifying a distance from the origin and an angle from the X-axis in the XY-plane, besides the distance in the direction of Z. In Fig. 1.10, for example, point A is a distance z from the XY-plane, a distance r from the origin as measured on a line that makes an angle θ with the X-axis and lying in the XY-plane. Notice that the radius of the cylinder corresponds to the value of the cylindrical coordinate in question, with the point located on the surface of the cylinder itself.

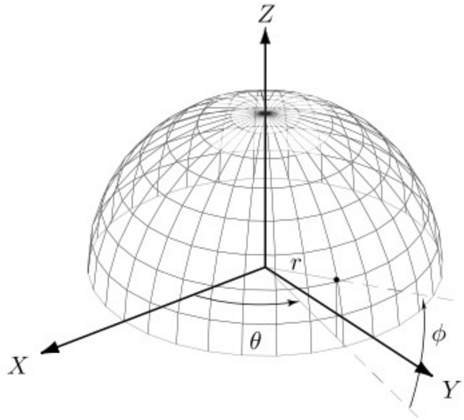

Fig. 1.11 Spherical coordinates

In general, cylindrical coordinates are used in designing axially symmetric shapes. The relations between cylindrical and Cartesian coordinates are given below:

$$r = \sqrt{x^2 + y^2}, \quad \theta = \arctan(x, y), \quad z = z \tag{1.2a}$$

$$x = r\cos\theta, \quad y = r\sin\theta, \quad z = z \tag{1.2b}$$

1.2.3.3 Spherical Coordinates

Spherical coordinates locate a point P on the surface of a sphere, as depicted in Fig. 1.11, by means of three *spherical coordinates*: the radius r of the sphere on which P lies; the *longitude* of the point, measured by angle θ; and the *latitude* of the point, measured by angle ϕ. In the figure, apparently, the X–Y plane defines the *equator*, the X–Z plane the *Greenwich meridian*.

The relations between Cartesian and spherical coordinates are readily derived from Fig. 1.11:

$$r = \sqrt{x^2 + y^2 + z^2}, \quad \phi = \arcsin\left(\frac{z}{r}\right), \quad \theta = \arccos\left(\frac{x}{\sqrt{x^2 + y^2}}\right) \tag{1.3a}$$

$$x = r\cos\phi\cos\theta, \quad y = r\cos\phi\sin\theta, z = r\sin\phi \tag{1.3b}$$

the ranges of ϕ and θ being

$$-\frac{\pi}{2} \le \phi \le \frac{\pi}{2}, \quad 0 \le \theta \le 2\pi \tag{1.3c}$$

That is, the sign of ϕ is the same as that of z, while the quadrant of θ is uniquely determined from the signs of x and y.

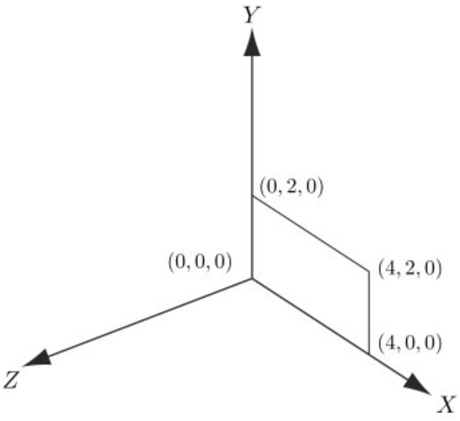

Fig. 1.12 Absolute coordinates

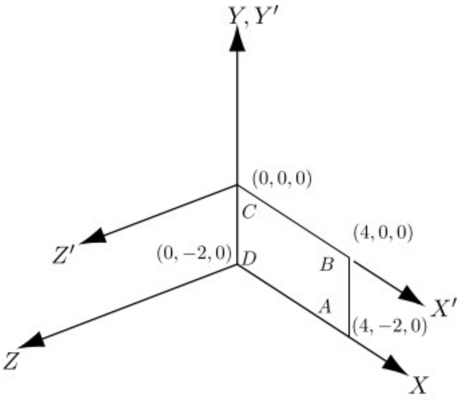

Fig. 1.13 World coordinates (x, y, x) versus relative coordinates (x', y', z')

1.2.3.4 Absolute/Relative/World Coordinate Systems

As illustrated in Fig. 1.12, *absolute coordinates* always refer to the origin (0, 0, 0).

Relative coordinates are always referenced to a previously defined location and are sometimes referred to as Δ-coordinates, as pertaining to the $\{X',\ Y',\ Z'\}$ frame, and shown in Fig. 1.13. Thus, you can have several different coordinate systems within one larger coordinate system; each object may have its own local (relative) system, but each local system is referenced to an encompassing "world" coordinate system.

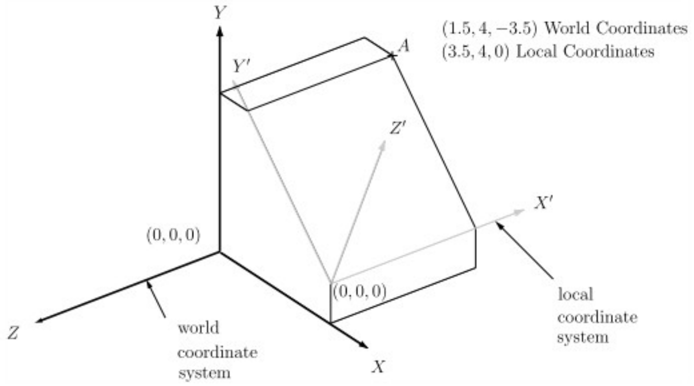

Fig. 1.14 World and local coordinates

As illustrated in Fig. 1.14, the *world coordinate system* uses a set of three numbers (x, y, z) located on three mutually perpendicular axes and measured from the origin $(0, 0, 0)$; the local coordinate system is a moving system that can be positioned anywhere in 3D space to assist in the construction of geometric objects.

1.2.4 Homogeneous Coordinates

One of the many purposes of *homogeneous coordinates* is to capture the concept of infinity. In the Euclidean coordinate system, infinity is something that does not exist. Mathematicians have discovered that many geometric concepts and computations can be greatly simplified if the concept of infinity is used. This will become apparent when we move to curve and surface design. Without the use of a homogeneous-coordinate system, it would be difficult to design certain classes of frequently used curves and surfaces in computer graphics and computer-aided design software, as well as perform transformations on these curves and surfaces. The concept of *point at infinity* is fundamental in geometry, but, as said above, difficult to represent with the aid of Euclidean geometry. *Projective geometry* comes to the rescue. In this concept, a point at infinity can be regarded as one whose distance from the observer is so large that its distance does not matter, only its *line of sight*, as in the case of stars, as far as the casual stargazer is concerned.

Any point P in the XY-plane has two coordinates. Adding a third component whose value is 1 to the coordinates of this point leads to the corresponding homogeneous coordinates. Thus, the homogeneous coordinates of any point P in the said plane, represented in the three-dimensional array $\boldsymbol{p}$, are

$$\boldsymbol{p} = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{1.4}$$

The distance d of P from the origin is known to be $d = \sqrt{x^2 + y^2}$. If we factor d out from the $\boldsymbol{p}$ array of Eq. (1.4), we obtain

$$\boldsymbol{p} = d \begin{bmatrix} \cos\theta \\ \sin\theta \\ 1/d \end{bmatrix}, \quad \theta = \cos^{-1}\left(\frac{y}{x}\right) \tag{1.5}$$

A point P_∞ lying at infinity, with a line of sight making an angle θ with the X-axis, has the homogeneous coordinates stored in the array $\boldsymbol{p}_\infty$ given below[2]:

$$\boldsymbol{p}_\infty = \begin{bmatrix} \cos\theta \\ \sin\theta \\ 0 \end{bmatrix} \tag{1.6}$$

which results from the array of Eq. (1.5) upon letting $d \to \infty$.

Similarly, the homogeneous coordinates of a point in 3D are defined as

$$\boldsymbol{p} = \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{1.7}$$

Likewise, the homogeneous coordinates of a point P_∞ lying at infinity in 3D space, with a line of sight of direction cosines (λ, μ, ν), as formally defined in Eq. (1.19), are stored in an array $\boldsymbol{p}_\infty$ given as

$$\boldsymbol{p}_\infty = \begin{bmatrix} \lambda \\ \mu \\ \nu \\ 0 \end{bmatrix} \tag{1.8}$$

Homogeneous coordinates have been traditionally used instead of ordinary Cartesian coordinates in computer graphics and geometric modeling. The representation of points in homogeneous coordinates provides a unified approach to the description of geometric transformations and allows these transformations to be represented as simple matrix operations.

---

[2] Only the array is of interest, its coefficient, d in Eq. (1.5), being left out.

($ $

We will use homogeneous coordinates in Chap. 4 when we will study *affine transformations*. As we will see, homogenous coordinates simplify (and in many cases, make possible) the mathematics needed to represent the desired transformations and manipulations.

This follows from $\boldsymbol{v} = \boldsymbol{v}_x + \boldsymbol{v}_y + \boldsymbol{v}_z$, where $\boldsymbol{v}_x = v_x \boldsymbol{i}$, $\boldsymbol{v}_y = v_y \boldsymbol{j}$, and $\boldsymbol{v}_z = v_z \boldsymbol{k}$. The same vector $\boldsymbol{v}$ can be described as a *column array*:

$$\boldsymbol{v} = \begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} \tag{1.11}$$

where $v_x, v_y, and v_z$ are the *components* of $\boldsymbol{v}$. The components may be negative, depending on the direction of the vector.

1.3.3 Basic Properties

1.3.3.1 Magnitude of a Vector

The magnitude of a vector, also known as the *Euclidean vector norm*, is non-negative and vanishes only when the vector itself does. The magnitude of $\boldsymbol{v}$ is thus a non-negative scalar quantity, denoted by $\|\boldsymbol{v}\|$ and given by

$$\|\boldsymbol{v}\| = \sqrt{v_x^2 + v_y^2 + v_z^2} \tag{1.12}$$

which is a simple application of the Pythagorean theorem used to find the length of the diagonal of a parallelepiped of sides with lengths v_x, v_y, v_z, such as the one shown in Fig. 1.8. Hence,

$$\|\boldsymbol{v}\|^2 = v_x^2 + v_y^2 + v_z^2 \tag{1.13}$$

1.3.3.2 Unit Vector

We define a unit vector as any vector whose magnitude is equal to unity, regardless of its direction. As we saw, $\boldsymbol{i}$, $\boldsymbol{j}$, and $\boldsymbol{k}$ are special cases of unit vectors, with specific directions assigned to them. An arbitrary unit vector in the direction of $\boldsymbol{v}$ can be obtained as follows:

$$\boldsymbol{w} = \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} \tag{1.14}$$

with

$$\|\boldsymbol{w}\| = 1 \tag{1.15}$$

A unit vector can be also written in the form

$$\boldsymbol{w} = \begin{bmatrix} v_x/\|\boldsymbol{v}\| \\ v_y/\|\boldsymbol{v}\| \\ v_z/|\boldsymbol{v}\| \end{bmatrix} \equiv \frac{1}{\|\boldsymbol{v}\|} \begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} \tag{1.16}$$

We can make this more concise with the definitions:

$$w_x = \frac{v_x}{\|\boldsymbol{v}\|},\ w_y = \frac{v_y}{\|\boldsymbol{v}\|},\ w_z = \frac{v_z}{\|\boldsymbol{v}\|} \tag{1.17}$$

so that

$$\boldsymbol{w} = \begin{bmatrix} w_x \\ w_y \\ w_z \end{bmatrix} \tag{1.18}$$

Note that if α, β, and γ are the angles between $\boldsymbol{v}$ and the X-, Y-, and Z-axes, respectively, then

$$w_x = \frac{v_x}{\|\boldsymbol{v}\|} = \cos\alpha \qquad w_y = \frac{v_y}{\|\boldsymbol{v}\|} = \cos\beta \qquad w_z = \frac{v_z}{\|\boldsymbol{v}\|} = \cos\gamma \tag{1.19}$$

which indicates that w_x, w_y, and w_z are the *direction cosines* of $\boldsymbol{v}$, whose squares add up to unity.

1.3.3.3 Multiplication by a Scalar

Multiplying any vector $\boldsymbol{v}$ by a scalar k produces a new vector $k\boldsymbol{v}$:

$$k\boldsymbol{v} = \begin{bmatrix} kv_x \\ kv_y \\ kv_z \end{bmatrix} \tag{1.20}$$

If k is positive, then $\boldsymbol{v}$ and $k\boldsymbol{v}$ are in the same direction; if k is negative, then $\boldsymbol{v}$ and $k\boldsymbol{v}$ are in opposite directions. The magnitude of $k\boldsymbol{v}$ is

$$\|k\boldsymbol{v}\| = \sqrt{k^2v_x^2 + k^2v_y^2 + k^2v_z^2} \tag{1.21}$$

so that

$$\|k\boldsymbol{v}\| = |k|\|\boldsymbol{v}\| \tag{1.22}$$

1.3.3.4 Vector Addition

Given $\boldsymbol{a} = [a_x\ a_y\ a_z]^T$ and $\boldsymbol{b} = [b_x\ b_y\ b_z]^T$, the sum of these two vectors is defined as

$$\boldsymbol{a} + \boldsymbol{b} = \begin{bmatrix} a_x + b_x \\ a_y + b_y \\ a_z + b_z \end{bmatrix} \tag{1.23}$$

Given vectors $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$ and scalars k and l, vector addition and scalar multiplication obey the properties below:

1. $\boldsymbol{a} + \boldsymbol{b} = \boldsymbol{b} + \boldsymbol{a}$,
2. $\boldsymbol{a} + (\boldsymbol{b} + \boldsymbol{c}) = (\boldsymbol{a} + \boldsymbol{b}) + \boldsymbol{c}$,
3. $k(l\boldsymbol{a}) = kl\boldsymbol{a}$,
4. $(k + l)\boldsymbol{a} = k\boldsymbol{a} + l\boldsymbol{a}$, and
5. $k(\boldsymbol{a} + \boldsymbol{b}) = k\boldsymbol{a} + k\boldsymbol{b}$.

1.3.4 Scalar Product

The scalar product, also known as the *dot product*, of two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is the sum of the products of their corresponding components:

$$\boldsymbol{a} \cdot \boldsymbol{b} = a_x b_x + a_y b_y + a_z b_z \tag{1.24}$$

which returns a scalar quantity, not another vector. An alternative form representing the scalar product is

$$\boldsymbol{a}^T \boldsymbol{b} = a_x b_x + a_y b_y + a_z b_z \tag{1.25}$$

with superscript T indicating *transposition*, i.e., $\boldsymbol{a}^T$ is the transpose of $\boldsymbol{a}$. Since a vector array was defined in Sect. 1.2.4 as a *column array*, $\boldsymbol{a}^T$ is a *row array*.

In addition, we can readily show that the scalar product is commutative, meaning that:

$$\boldsymbol{a} \cdot \boldsymbol{b} = \boldsymbol{b} \cdot \boldsymbol{a} \tag{1.26}$$

Alternatively, the scalar product may be calculated using the angle θ between the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$:

$$\boldsymbol{a} \cdot \boldsymbol{b} = \|\boldsymbol{a}\| \|\boldsymbol{b}\| \cos\theta \tag{1.27}$$

Moreover, the following two statements are equivalent: for any $\boldsymbol{a}, \boldsymbol{b} \neq \boldsymbol{0}$,

$$\boldsymbol{a} \cdot \boldsymbol{b} = 0 \Longleftrightarrow \boldsymbol{a} \text{ and } \boldsymbol{b} \text{ are perpendicular.} \tag{1.28}$$

In summary, the scalar product has the following properties:

1. $\boldsymbol{a} \cdot \boldsymbol{b} = \|\boldsymbol{a}\| \|\boldsymbol{b}\| \cos\theta$, where θ is the angle between $\boldsymbol{a}$ and $\boldsymbol{b}$;
2. $\boldsymbol{a} \cdot \boldsymbol{a} = \|\boldsymbol{a}\|^2$;
3. $\boldsymbol{a} \cdot \boldsymbol{b} = \boldsymbol{b} \cdot \boldsymbol{a}$, *commutativity*;
4. $\boldsymbol{a} \cdot (\boldsymbol{b} + \boldsymbol{c}) = \boldsymbol{a} \cdot \boldsymbol{b} + \boldsymbol{a} \cdot \boldsymbol{c}$, *distributivity*;
5. $(k\boldsymbol{a}) \cdot \boldsymbol{b} = \boldsymbol{a} \cdot (k\boldsymbol{b}) = k(\boldsymbol{a} \cdot \boldsymbol{b})$, *associativity*; and
6. $\boldsymbol{a}$ is perpendicular to $\boldsymbol{b} \Longleftrightarrow \boldsymbol{a} \cdot \boldsymbol{b} = 0$.

1.3.5 *Inequalities*

In connection with vector norms, there are two important inequalities that arise, as given below:

- Cauchy–Schwartz inequality:

$$(\boldsymbol{v} \cdot \boldsymbol{w})^2 \leq \|\boldsymbol{v}\|^2 \|\boldsymbol{w}\|^2 \tag{1.29}$$

- Triangle inequality:

$$\|\boldsymbol{v} + \boldsymbol{w}\| \leq \|\boldsymbol{v}\| + \|\boldsymbol{w}\| \tag{1.30}$$

These properties are quite useful for the derivation of other identities and inequalities, and are fundamental to the understanding and application of vectors in computer graphics.

1.4 Matrices

1.4.1 *Definition*

A matrix is a rectangular array of numbers arranged in m rows and n columns, namely,

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix} \tag{1.31}$$

1.4.2 *Special Matrices*

- Square matrix:
  A square matrix has an equal number of rows and columns ($m = n$), e.g.,

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \tag{1.32}$$

- Row matrix:
  A row matrix has one row:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \end{bmatrix} = \boldsymbol{a}^T \tag{1.33}$$

That is, a row matrix is a *transposed* vector.

- Column matrix:
  A column matrix has one column. In our choice of notation, this column matrix becomes a vector array, namely,

$$\boldsymbol{b} = \begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \end{bmatrix} \tag{1.34}$$

  Note the use of a bold *lowercase* font in Eqs. (1.33) and (1.34), indicating that this array is a vector.
- Diagonal matrix:
  A square matrix that has zero entries everywhere except on its main diagonal is termed *diagonal*:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & 0 & 0 & 0 & \dots & 0 \\ 0 & a_{22} & 0 & 0 & \dots & \vdots \\ 0 & 0 & a_{33} & 0 & \dots & \vdots \\ 0 & 0 & 0 & a_{44} & \dots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & \dots & \dots & \dots & 0 & a_{mm} \end{bmatrix} \tag{1.35}$$

  Diagonal matrices satisfy: $a_{ij} = 0$ if $i \neq j$.
- Identity matrix:
  This is a special diagonal matrix with unit elements on the main diagonal, and zero elsewhere. This matrix is denoted by the symbol **1**. For example, the 3 × 3 identity matrix is defined as

$$\mathbf{1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{1.36}$$

  The elements of **1** are sometimes represented by the *Kronecker delta*, namely,

$$\begin{aligned} \delta_{ij} &= 0 \quad \text{if } i \neq j \\ \delta_{ij} &= 1 \quad \text{if } i = j \end{aligned} \tag{1.37}$$

- Zero matrix:
  The $m \times n$ zero matrix has all its elements equal to zero. We will denote it by $\boldsymbol{O}$. For example, the 3 × 3 zero matrix is

$$\boldsymbol{O} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{1.38}$$

- Symmetric matrix:
  Symmetric matrices are symmetric about the main diagonal: $a_{ij} = a_{ji}$, i.e.,

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ \vdots & a_{22} & a_{23} \\ \text{symmetric} & \ldots & a_{33} \end{bmatrix} \tag{1.39}$$

- Skew-symmetric matrix:
  A matrix is skew-symmetric if $a_{ij} = -a_{ji}$, namely,

$$\boldsymbol{A} = \begin{bmatrix} 0 & a_{12} & a_{13} \\ -a_{12} & 0 & a_{23} \\ -a_{13} & -a_{23} & 0 \end{bmatrix} \tag{1.40}$$

  Notice that the diagonal entries of a skew-symmetric matrix vanish necessarily.
- Triangular matrix: An *upper triangular* matrix has all its entries below the diagonal equal to zero, namely,

$$\boldsymbol{U} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ 0 & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & a_{nn} \end{bmatrix} \tag{1.41}$$

  A *lower triangular* matrix is defined correspondingly. Furthermore, while upper (lower) triangular matrices can be rectangular, *symmetric* and *skew-symmetric* matrices are necessarily square.

1.4.3 Properties

- Matrix equality: Let us assume that matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ have the same numbers of rows and columns. Then,

$$\boldsymbol{A} = \boldsymbol{B} \Longleftrightarrow a_{ij} = b_{ij}, \text{ for all } i, j \tag{1.42}$$

- Matrix addition:
  Adding two $m \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ produces a third $m \times n$ matrix $\boldsymbol{C}$, whose entries are equal to the sum of the corresponding entries of $\boldsymbol{A}$ and $\boldsymbol{B}$. Thus,

$$\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{C} \tag{1.43}$$

  and

$$a_{ij} + b_{ij} = c_{ij} \tag{1.44}$$

Of course, we can add or subtract two matrices if and only if they have the same numbers of rows and columns. Moreover, matrix addition is commutative, i.e., $\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{B} + \boldsymbol{A}$.

- Matrix multiplication by a scalar:
  Multiplying a matrix $\boldsymbol{A}$ by a scalar k produces a new matrix $\boldsymbol{B}$ with the same number of rows and columns. Each entry of $\boldsymbol{B}$ is obtained by multiplying the corresponding entry of $\boldsymbol{A}$ by the scalar k:

$$k\boldsymbol{A} = \boldsymbol{B} \tag{1.45}$$

and

$$ka_{ij} = b_{ij} \tag{1.46}$$

- Matrix product:
  The product $\boldsymbol{AB}$ of two matrices is another matrix $\boldsymbol{C}$. This operation is possible if and only if the number of columns of the first matrix is equal to the number of rows of the second matrix. In general, the product of two matrices is not commutative:

$$\boldsymbol{AB} \neq \boldsymbol{BA} \tag{1.47}$$

The product $\boldsymbol{C}$ of two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ in terms of their entries is

$$\boldsymbol{C} = \boldsymbol{AB} \quad \Rightarrow \quad c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} \tag{1.48}$$

- Transpose matrix
  By interchanging the rows and columns of a matrix $\boldsymbol{A}$, we obtain its transpose $\boldsymbol{A}^T$, so that $a_{ij}^T = a_{ji}$, where a_{ij}^T is the (i, j) entry of the transpose of $\boldsymbol{A}$, i.e.,

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \quad \Rightarrow \quad \boldsymbol{A}^T = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \end{bmatrix} \tag{1.49}$$

1.4.4 The Vector Product

The vector product of two 3D vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, also known as the *cross product*, is defined as

$$\boldsymbol{a} \times \boldsymbol{b} = (a_y b_z - a_z b_y)\boldsymbol{i} - (a_x b_z - a_z b_x)\boldsymbol{j} + (a_x b_y - a_y b_x)\boldsymbol{k} \tag{1.50}$$

In array form,

$$\boldsymbol{a} \times \boldsymbol{b} = \begin{bmatrix} a_y b_z - a_z b_y \\ a_z b_x - a_x b_z \\ a_x b_y - a_y b_x \end{bmatrix} \tag{1.51}$$

One *mnemonic* means to compute the vector product relies on the expansion of a determinant (cofactor expansion, as outlined in Sect. 1.4.6), namely,

$$\begin{aligned} \boldsymbol{a} \times \boldsymbol{b} &= \det \begin{bmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \\ a_x & a_y & a_z \\ b_x & b_y & b_z \end{bmatrix} \\ &= (a_y b_z - a_z b_y)\boldsymbol{i} - (a_x b_z - a_z b_x)\boldsymbol{j} + (a_x b_y - a_y b_x)\boldsymbol{k} \end{aligned} \tag{1.52}$$

The properties of the vector product are given below:

- If $\boldsymbol{c} = \boldsymbol{a} \times \boldsymbol{b}$, then $\boldsymbol{c}$ is perpendicular to both $\boldsymbol{a}$ and $\boldsymbol{b}$. Consequently, $\boldsymbol{c}$ is also perpendicular to the plane defined by $\boldsymbol{a}$ and $\boldsymbol{b}$.
- The vector product is *skew-symmetric*: $\boldsymbol{a} \times \boldsymbol{b} = -(\boldsymbol{b} \times \boldsymbol{a})$.
- $\boldsymbol{a}$ and $\boldsymbol{b}$ are parallel $\Longleftrightarrow \boldsymbol{a} \times \boldsymbol{b} = \boldsymbol{0}$.

Also note that, if $\boldsymbol{c} = \boldsymbol{a} \times \boldsymbol{b}$ and θ denotes the angle from $\boldsymbol{a}$ to $\boldsymbol{b}$, in this direction, when the tails of the two vectors coincide, then

$$\boldsymbol{c} = (\|\boldsymbol{a}\| \|\boldsymbol{b}\| \sin\theta)\boldsymbol{n} \tag{1.53}$$

where $\boldsymbol{n}$ is the unit vector normal to both $\boldsymbol{a}$ and $\boldsymbol{b}$, so that $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{n})$ is a right-hand triad.

The *geometric interpretation* of the vector product is now straightforward: As depicted in Fig. 1.15, the expression in parentheses in Eq. (1.53) is nothing but *twice* the area of the triangle AOB, i.e., the area of the parallelogram $OADB$.

1.4.5 The 2D Form of the Vector (Cross) Product

The vector product, or cross product, of two 3D vectors was defined in Sect. 1.4.4. This product exists only in three dimensions. However, in 2D geometry one is confronted frequently with the calculation of the cross product. To ease the solution of 2D geometric problems involving the cross product, we introduce below a *2D form of the cross product*.

Let $\boldsymbol{E}$ be an *orthogonal matrix* that rotates vectors in the plane through an angle of 90° counterclockwise (ccw), namely,

$$\boldsymbol{E} \equiv \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \tag{1.54}$$

With this definition, we can readily prove that

$$\boldsymbol{E}^T \boldsymbol{E} = \boldsymbol{E}\boldsymbol{E}^T = \boldsymbol{1} \tag{1.55}$$

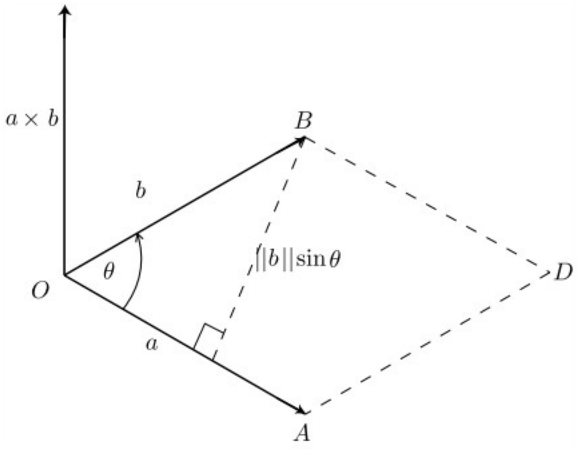

Fig. 1.15 Geometric interpretation of the vector product

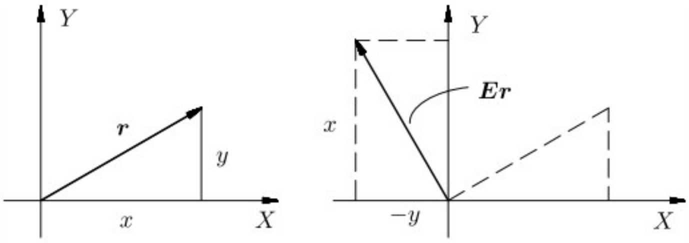

Fig. 1.16 Vector $\boldsymbol{r}$ and its image under $\boldsymbol{E}$

in which $\mathbf{1}$ is the 2×2 identity matrix. Moreover, note that $\boldsymbol{E}$ is *skew-symmetric*, i.e., $\boldsymbol{E} = -\boldsymbol{E}^T$, and hence $\boldsymbol{E}^2 = -\mathbf{1}$.

Also note that, given any vector $\boldsymbol{r} = \begin{bmatrix} x & y \end{bmatrix}^T$ in a plane Π, its image under $\boldsymbol{E}$ is given by

$$\boldsymbol{Er} = \begin{bmatrix} -y \\ x \end{bmatrix} \tag{1.56}$$

as illustrated in Fig. 1.16.

Now, let us compute the cross product $\boldsymbol{a} \times \boldsymbol{b}$, where $\boldsymbol{a} = A\boldsymbol{k}$, and $\boldsymbol{k}$ is a unit vector normal to Π, pointing toward the viewer, $\boldsymbol{a}$ thus being a 3D vector normal to Π, of magnitude $|A|$. Moreover, we assume that $\boldsymbol{b}$ lies in Π, its Z-component thus vanishing. The cross product of interest thus takes the form[4].

$$\boldsymbol{a} \times \boldsymbol{b} = \det \begin{bmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \\ 0 & 0 & A \\ b_x & b_y & 0 \end{bmatrix} = -Ab_y\boldsymbol{i} + Ab_x\boldsymbol{j} \tag{1.57}$$

---

[4] "det" is abbreviation for *determinant*. A concept formally defined in Subsection 1.4.6.

where we have recalled that the unit vectors $\boldsymbol{i}$ and $\boldsymbol{j}$ are parallel to the X- and Y-axes, respectively. The two-dimensional form of the foregoing product, then, becomes

$$(\boldsymbol{a} \times \boldsymbol{b})_{2D} = A \begin{bmatrix} -b_y \\ b_x \end{bmatrix} \equiv A\boldsymbol{E}\overline{\boldsymbol{b}} \tag{1.58}$$

where we have recalled Eq. (1.56) and $\overline{\boldsymbol{b}}$ denotes the 2D version of the three-dimensional $\boldsymbol{b}$, i.e.,

$$\overline{\boldsymbol{b}} = \begin{bmatrix} b_x \\ b_y \end{bmatrix}$$

Likewise, the cross product $\boldsymbol{b} \times \boldsymbol{c}$, for both $\boldsymbol{b}$ and $\boldsymbol{c}$ in the plane Π, is a vector perpendicular to this plane, of *signed* magnitude,[5] $\|\boldsymbol{b}\|\|\boldsymbol{c}\| \sin(\boldsymbol{b},\ \boldsymbol{c})$, where $(\boldsymbol{b},\ \boldsymbol{c})$ denotes the angle between these two vectors, measured from $\boldsymbol{b}$ to $\boldsymbol{c}$.

Thus, if $\sin(\boldsymbol{b},\ \boldsymbol{c})$ is positive, the cross-product vector points in the direction of $\boldsymbol{k}$; otherwise, in the direction of $-\boldsymbol{k}$.

More concretely, let $\boldsymbol{b}$ be defined as before, $\boldsymbol{c}$ being defined, in turn, as

$$\boldsymbol{c} \equiv \begin{bmatrix} c_x \\ c_y \\ 0 \end{bmatrix} \tag{1.59}$$

Hence,

$$\boldsymbol{b} \times \boldsymbol{c} = \det \begin{bmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \\ b_x & b_y & 0 \\ c_x & c_y & 0 \end{bmatrix} = (b_x c_y - b_y c_x)\boldsymbol{k} \ \equiv\ C\boldsymbol{k} \tag{1.60}$$

where C is a real number that can be positive, negative, or even 0. Since we know the direction of $\boldsymbol{b} \times \boldsymbol{c}$, i.e., perpendicular to the plane Π, all we need is the quantity C above, which can be readily recognized as the dot product of the two-dimensional vectors $\boldsymbol{E}\overline{\boldsymbol{b}}$, as given in Eq. (1.58), and $\overline{\boldsymbol{c}}$, the 2D counterpart of $\boldsymbol{c}$, i.e.,

$$C = \begin{bmatrix} c_x & c_y \end{bmatrix}^T \begin{bmatrix} -b_y \\ b_x \end{bmatrix} = \overline{\boldsymbol{c}}^T \boldsymbol{E}\overline{\boldsymbol{b}} \equiv (\boldsymbol{E}\overline{\boldsymbol{b}})^T \overline{\boldsymbol{c}} = -\overline{\boldsymbol{b}}^T \boldsymbol{E}\overline{\boldsymbol{c}} \tag{1.61}$$

Therefore, the sign of C depends on whether the 3D cross product of Eq. (1.60) points toward the reader or not. C vanishes, of course, if the two factors, $\overline{\boldsymbol{b}}$ and $\overline{\boldsymbol{c}}$, are parallel.

Next we recall a concept of analytic geometry: a *directed segment* $\overrightarrow{\mathcal{L}}$ is a line $\mathcal{L}$ with a direction. In Fig. 1.17, two directed segments $\overrightarrow{\mathcal{L}}_1$ and $\overrightarrow{\mathcal{L}}_2$ are shown, each given by a line $\mathcal{L}_i$ and a unit vector $\mathbf{e}_i$, for $i = 1, 2$.

---

[5] The signed magnitude of a vector is a real number, positive, negative, or zero, whose absolute value is identical to the magnitude of the vector.

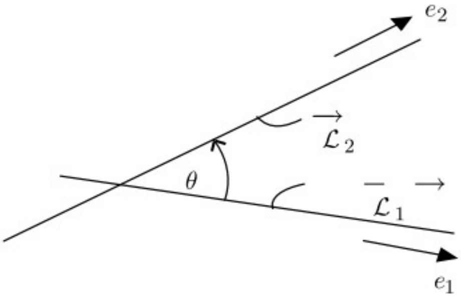
e_2
$\overrightarrow{\mathcal{L}}_2$
θ
$\overrightarrow{\mathcal{L}}_1$
e_1

Fig. 1.17 The angle between two directed segments, measured from $\overrightarrow{\mathcal{L}}_1$ to $\overrightarrow{\mathcal{L}}_2$

Now we introduce a practical application of the foregoing concepts in solving a recurrent problem of planar geometry:

Problem Given two directed segments $\overrightarrow{\mathcal{L}}_1$ and $\overrightarrow{\mathcal{L}}_2$, find the angle θ, for $0 \leq \theta \leq 2\pi$, that $\overrightarrow{\mathcal{L}}_2$ makes with $\overrightarrow{\mathcal{L}}_1$, while measuring θ ccw.

Solution Let $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ be 2D unit vectors indicating the direction of segments $\overrightarrow{\mathcal{L}}_1$ and $\overrightarrow{\mathcal{L}}_2$, respectively. Obviously, $\cos\theta$ can be derived from the scalar product P1 of $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$, namely,

$$\mathrm{P1} = \cos\theta = \boldsymbol{e}_1^T\boldsymbol{e}_2 \quad \text{or} \quad \boldsymbol{e}_1 \cdot \boldsymbol{e}_2 \tag{1.62}$$

However, the foregoing value does not determine uniquely θ, for, if $\mathrm{P1} > 0$, then θ may lie in either the first or the fourth quadrant, thus leaving us with an ambiguity. Ditto if $\mathrm{P1} < 0$, in which case θ may lie in either the second or the third quadrant.

To destroy the ambiguity, we need $\sin\theta$, which can be derived from the vector product of $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ when regarded as 3D vectors, as per Eq. (1.53). In light of the 2D form of the vector product, however, we need not work out of the plane of the two given lines. Indeed, from Eqs. (1.53), (1.60), and (1.61), we can write

$$P2 = \sin\theta = (\boldsymbol{E}\boldsymbol{e}_1)^T\boldsymbol{e}_2 \quad \text{or} \quad (\boldsymbol{E}\boldsymbol{e}_1) \cdot \boldsymbol{e}_2 \tag{1.63}$$

Obviously, if $\mathrm{P2} > 0$, then θ may lie in either the first or the second quadrant, which leaves us with an ambiguity. Ditto if $\mathrm{P2} < 0$, in which case θ may lie in either the third or the fourth quadrant.

While each of P1 and P2 does not determine *univocally* angle θ individually, both do. In fact, we can draw the rules below:

1. If $\text{P1} > 0$ and $\text{P2} > 0$, then θ lies in the first quadrant;
2. If $\text{P1} < 0$ and $\text{P2} > 0$, then θ lies in the second quadrant;
3. If $\text{P1} < 0$ and $\text{P2} < 0$, then θ lies in the third quadrant;
4. If $\text{P1} > 0$ and $\text{P2} < 0$, then θ lies in the fourth quadrant.

1.4.6 Determinants

The *determinant* is a quantity associated with an arbitrary $n \times n$ square matrix (note that the number of rows and columns must be identical). A general definition of what a determinant actually represents is rather cumbersome; luckily, we do not need it here. We can define the determinant of a $n \times n$ matrix starting with the simplest case, i.e., $n = 2$, then $n = 3$, and hence, by induction, derive a *procedure* to compute the determinant for any arbitrary value of n.

As a matter of fact, the interest in the determinant is rather theoretical; its actual computation, which is extremely costly in terms of *floating-point operations*, or *flops*, is seldom needed. The relevance of the concept lies in that the value of the determinant indicates whether or not the matrix is *singular*; if it is so, then the *solution* to the system of linear equations associated with the matrix coefficient is not unique. We are interested, in this book, only in the case in which the determinant is *non-zero*, in which case the associated system of linear equations admits one solution, which is *unique*. In this case, the matrix is said to be *invertible*, *regular* or, most commonly, *non-singular*.

A 2×2 matrix $\boldsymbol{A}$ can be *partitioned* either column-wise or row-wise, as shown below:

$$\boldsymbol{A} \equiv \begin{bmatrix} \boldsymbol{a} & \boldsymbol{b} \end{bmatrix} \equiv \begin{bmatrix} \boldsymbol{c}^T \\ \boldsymbol{d}^T \end{bmatrix} \tag{1.64}$$

where $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, and $\boldsymbol{d}$ are all two-dimensional column vectors. Furthermore, we recall the definition of $\boldsymbol{E}$ as seen in Sect. 1.4.5. If the components of $\boldsymbol{a}$ and $\boldsymbol{b}$ are given as $\boldsymbol{a} = \begin{bmatrix} a_x & a_y \end{bmatrix}^T$ and $\boldsymbol{b} = \begin{bmatrix} b_x & b_y \end{bmatrix}^T$, then the determinant of $\boldsymbol{A}$ is defined as

$$\det(\boldsymbol{A}) = \det \begin{bmatrix} a_x & b_x \\ a_y & b_y \end{bmatrix} = a_x b_y - a_y b_x \tag{1.65}$$

which can be readily cast in the form

$$\det(\boldsymbol{A}) = \begin{bmatrix} a_x & a_y \end{bmatrix} \begin{bmatrix} b_y \\ -b_x \end{bmatrix} \equiv \begin{bmatrix} a_x & a_y \end{bmatrix} \underbrace{\left(- \begin{bmatrix} -b_y \\ b_x \end{bmatrix} \right)}_{-\boldsymbol{E}\boldsymbol{b}} \tag{1.66}$$

the first array of the foregoing product being $\boldsymbol{a}^T$, the second $-\boldsymbol{E}\boldsymbol{b}$.

3.1.3 Distance of a Point to a Plane

Given the plane Π represented by Eq. (3.3a) or, alternatively, by Eq. (3.3b), we want to compute the distance of a *given* point $Q(\xi,\ \eta,\ \zeta)$ to the plane, as illustrated in Fig. 3.1. To do this, we proceed as in Sect. 2.2.1: We first find the unit normal $\boldsymbol{n}$ to the plane. Moreover, let $\boldsymbol{p}_0$, $\boldsymbol{q}$, and $\boldsymbol{p}$ be the position vectors of P_0, Q, and $P(x,\ y,\ z)$, the latter being an arbitrary point of the plane. Since P_0 and P lie in the plane, the difference $\boldsymbol{p} - \boldsymbol{p}_0$ is perpendicular to the unit normal to the plane, $\boldsymbol{n}$, i.e.,

$$\boldsymbol{n}^T(\boldsymbol{p} - \boldsymbol{p}_0) = 0 \tag{3.7}$$

or, in expanded form,

$$\boldsymbol{n}^T\boldsymbol{p} - \boldsymbol{n}^T\boldsymbol{p}_0 = 0 \tag{3.8}$$

Now, let us divide both sides of Eq. (3.3b) by $\sqrt{A^2 + B^2 + C^2}$:

$$\frac{A}{\sqrt{A^2 + B^2 + C^2}}x + \frac{B}{\sqrt{A^2 + B^2 + C^2}}y + \frac{C}{\sqrt{A^2 + B^2 + C^2}}z + \frac{D}{\sqrt{A^2 + B^2 + C^2}} = 0 \tag{3.9}$$

Comparison of Eqs. (3.8) and (3.9) leads to

$$\boldsymbol{n} = \frac{1}{\sqrt{A^2 + B^2 + C^2}}\begin{bmatrix} A \\ B \\ C \end{bmatrix}, \quad \boldsymbol{n}^T\boldsymbol{p}_0 = -\frac{D}{\sqrt{A^2 + B^2 + C^2}} \tag{3.10}$$

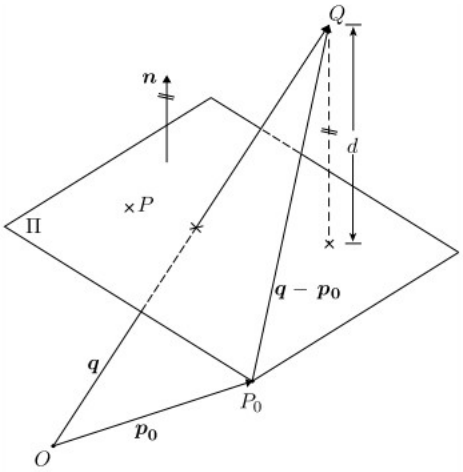

Fig. 3.1 Distance of a point to a plane

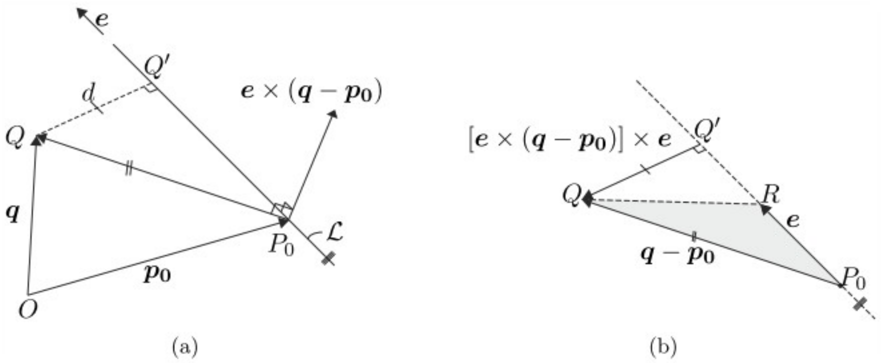

Fig. 3.2 Distance of a point to a line: **a** general layout; **b** geometric interpretation of $[\boldsymbol{e} \times (\boldsymbol{q} - \boldsymbol{p}_0)] \times \boldsymbol{e}$

From Fig. 3.1, the distance d sought is nothing but the *absolute value* of the projection of vector $\boldsymbol{q} - \boldsymbol{p}_0$ onto the unit normal $\boldsymbol{n}$, namely,

$$d = |\boldsymbol{n}^T(\boldsymbol{q} - \boldsymbol{p}_0)| \tag{3.11}$$

Interestingly, the expression for the distance of a point to a line in the 2D case, Eq. (2.9), is *formally identical* to the foregoing expression.

3.1.4 Distance of a Point to a Line

A line $\mathcal{L}$ is given by the two planes (3.6a and 3.6b). We want to find the distance of an arbitrary point $Q(\xi, \eta, \zeta)$ to $\mathcal{L}$.

First, we need a unit vector $\boldsymbol{e}$ parallel to $\mathcal{L}$ and a point P_0 of $\mathcal{L}$. If we denote by $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ the unit normal to each of the two planes that define $\mathcal{L}$, then we can obtain $\boldsymbol{e}$ as $\boldsymbol{n}_1 \times \boldsymbol{n}_2/\|\boldsymbol{n}_1 \times \boldsymbol{n}_2\|$. Moreover, $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ are produced using the expression for $\boldsymbol{n}$ displayed in Eq. (3.10):

$$\boldsymbol{n}_i = \frac{1}{\sqrt{A_i^2 + B_i^2 + C_i^2}} \begin{bmatrix} A_i \\ B_i \\ C_i \end{bmatrix}, \quad i = 1, 2 \tag{3.12}$$

We thus have the layout of Fig. 3.2a.

From Fig. 3.2a, P_0 is a point of $\mathcal{L}$, of position vector $\boldsymbol{p}_0$, while $\boldsymbol{e} \times (\boldsymbol{q} - \boldsymbol{p}_0)$ is a vector normal to the plane defined by $\mathcal{L}$ and Q, its norm $\|\boldsymbol{e} \times (\boldsymbol{q} - \boldsymbol{p}_0)\|$ being twice the area of the triangle P_0RQ depicted in Fig. 3.2b. Moreover, if we regard P_0R as the base of the triangle, d becomes its height, and hence

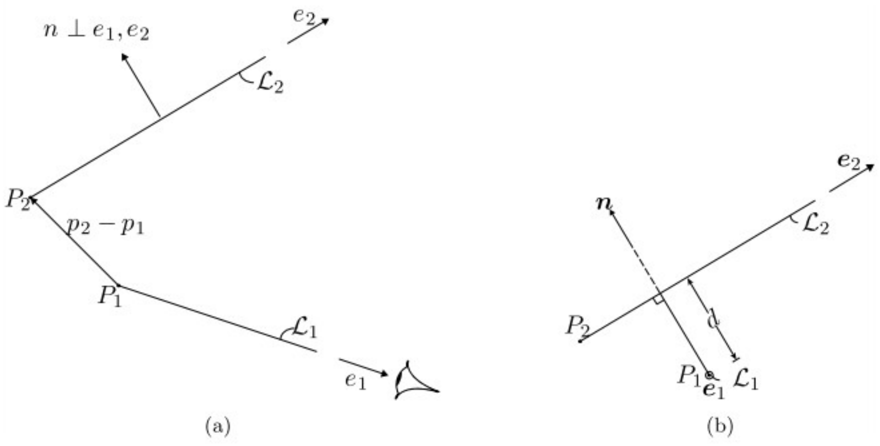

Fig. 3.3 Distance between two lines: **a** general layout; **b** view with $\mathcal{L}_1$ projected as a point, and $\boldsymbol{n}$ lying in the projection plane

$$\overline{P_0 R} d = \|\boldsymbol{e} \times (\boldsymbol{q} - \boldsymbol{p}_0)\| \tag{3.13}$$

where $\overline{P_0 R} = \|\boldsymbol{e}\| = 1$, so that

$$d = \|\boldsymbol{e} \times (\boldsymbol{q} - \boldsymbol{p}_0)\| \tag{3.14}$$

thereby computing the desired distance. Notice, moreover, that

$$\boldsymbol{q} - \boldsymbol{q}' = [\boldsymbol{e} \times (\boldsymbol{q} - \boldsymbol{p}_0)] \times \boldsymbol{e} \tag{3.15}$$

In the above equation, $\boldsymbol{q}'$ is the position vector of Q', the *orthogonal* projection of Q onto $\mathcal{L}$.

3.1.5 Distance Between Two Skew Lines

We have the general layout of Fig. 3.3, depicting two *skew* lines $\mathcal{L}_1$ and $\mathcal{L}_2$, parallel to the unit vectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ and passing through P_1 and P_2, respectively. If $\boldsymbol{n}$ denotes the unit normal to $\mathcal{L}_1$ and $\mathcal{L}_2$, then, apparently, the distance d between the two lines is nothing but the absolute value of the projection of $\boldsymbol{p}_2 - \boldsymbol{p}_1$ onto $\boldsymbol{n}$, i.e.,

$$d = |\boldsymbol{n}^T(\boldsymbol{p}_2 - \boldsymbol{p}_1)|, \quad \boldsymbol{n} \equiv \frac{\boldsymbol{e}_1 \times \boldsymbol{e}_2}{||\boldsymbol{e}_1 \times \boldsymbol{e}_2||} \tag{3.16}$$

The above relations are illustrated in Fig. 3.3b.

3.2 Surfaces

A surface is a two-dimensional set of points, extending in two directions that change from point to point, but has no thickness. We will study various types of surfaces, as described below.

The *plane*, introduced and defined in Sect. 3.1.1, is the simplest surface. That is, the plane is the *perpendicular bisector* of the *segment* defined by two given points. A plane can also be visualized as a set of lines passing through a given point and perpendicular to one given direction. In *computer graphics*, *solid objects* can be bounded by planes, forming *facets* of the solid, each facet being a polygon. In this case, the solid turns out to be a *polyhedron*. As a matter of fact, arbitrary surfaces, like airplane fuselages, are sometimes approximated, for certain computations pertaining to the solids that they enclose—volume, centroid location, etc.—by polyhedra. These polyhedra are the *finite elements* introduced to determine the stress and strain distribution on the fuselage of aircraft.[3]

In increasing order of *complexity*, the next surface is the *quadric*, namely, a surface defined, in a certain coordinate frame, by a *quadratic tri-variate polynomial*, namely,

$$\begin{aligned}F(x,\ y,\ z) = A_{11}x^2 + 2A_{12}xy + 2A_{13}xz + A_{22}y^2 + 2A_{23}yz\\ + A_{33}z^2 + B_1x + B_2y + B_3z + C = 0\end{aligned} \tag{3.17}$$

which can be cast in the compact form

$$F(\boldsymbol{p}) = \boldsymbol{p}^T\boldsymbol{A}\boldsymbol{p} + \boldsymbol{b}^T\boldsymbol{p} + C = 0 \tag{3.18a}$$

where

$$\boldsymbol{A} \equiv \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{12} & A_{22} & A_{23} \\ A_{13} & A_{23} & A_{33} \end{bmatrix}, \quad \boldsymbol{b} \equiv \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix} \tag{3.18b}$$

and $\boldsymbol{p}$ defined as in Eq. (3.1), with $\boldsymbol{A}$ being *symmetric*. The above expression can be cast in *homogeneous* array form, similar to Eq. (2.16), if we introduce homogeneous coordinates, namely,

$$F(p) = \boldsymbol{p}^T\boldsymbol{R}\boldsymbol{p} = 0 \tag{3.19a}$$

where

$$\boldsymbol{p} = \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}, \quad \boldsymbol{R} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{b} \\ \boldsymbol{b}^T & C \end{bmatrix} \tag{3.19b}$$

---

[3] In fact, some "advanced" finite elements are not polyhedra, but solids bounded by combinations of quadratic or higher order simple surfaces.

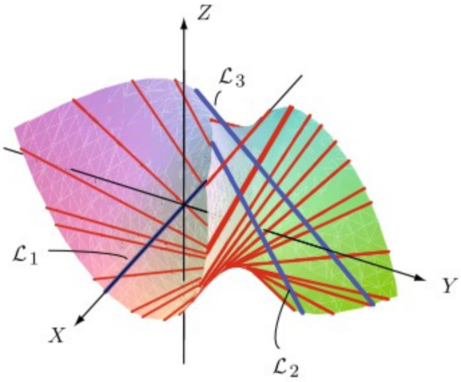

Fig. 3.4 A single-sheet hyperboloid created using three skew lines

Examples of quadrics are the *ellipsoid*, a particular case of which is the *sphere*; the *two-sheet hyperboloid*; the *single-sheet hyperboloid*; the *hyperbolic paraboloid*; and the *paraboloid*. In the same way that we have criteria based on the entries of the upper left block of matrix $\boldsymbol{R}$ of Eq. (2.17) to characterize the conics, there are criteria to characterize the quadric at hand. The reader should have noticed that matrix $\boldsymbol{R}$ of Eq. (3.19b) is *symmetric*, just like its counterpart in Eq. (2.16) in the planar case, in connection with the conic curves. The abovementioned criteria are based on matrix $\boldsymbol{A}$ of Eq. (3.19b). However, these criteria fall outside the scope of this book and will not be pursued here. The total number of quadrics[4] is 17.

Out of the foregoing surfaces, the single-sheet hyperboloid deserves special attention, as it leads, as special cases, to well-known familiar surfaces such as the cylinder and the cone. A single-sheet hyperboloid can be generated by the motion of a line constrained to intersect three given *skew lines*. Two skew lines do not intersect, which means that they do not intersect at all, not even at infinity, as do parallel lines. Shown in Fig. 3.4 is a picture of a single-sheet hyperboloid, as defined by three skew lines, $\{\mathcal{L}_i\}_1^3$.

All surfaces generated by the motion of a line belong to the class of *ruled surfaces*. The moving line generating the surface is termed the *generatrix*, which moves along a curve termed the *directrix*, the relative orientation of the generatrix with respect to the directrix being, in general, variable. Depending on the pattern of this variation, different surfaces can be obtained from the same directrix. In particular, when the orientation of the generatrix is kept constant, a *cylindrical surface* is obtained. Obviously, if the directrix is a circle and the generatrix remains normal to the plane of the circle, then a *circular cylinder* is generated. If the directrix is still a circle, but the generatrix is constrained to pass through a given point, then a *cone* is obtained. Moreover, if the given point lies on the normal to the plane of the circle passing through the center of the circle, then a *right circular cone* is generated.

Ruled surfaces are *curved* in one direction, that of the directrix, but are straight in the direction of the generatrix. For this reason, such surfaces are sometimes termed

---

[4] For the whole list, visit http://mathworld.wolfram.com/QuadraticSurface.html.

single-curve surfaces. Probably the most photographed ruled surface is the *Guggenheim Museum* in Bilbao, Spain. The architect behind this design, Frank Gehry, became well known in architects' circles because of his pioneering use of the CAD software package that was then in use for the design of aircraft fuselage. The design of the latter is much more daring in comparison, because the fuselage is a *double-curved* surface. The most general surfaces are *double curve*. The simplest examples of these are the ellipsoid and the paraboloid.

In computer graphics, certain special kinds of surfaces can be obtained from a generatrix. For example, to generate an axially symmetric surface, like the sphere, we can turn a circle around one of its diameters. Such surfaces are termed *surfaces of revolution*. Other means of generating surfaces is by *extrusion*, whereby a generatrix is translated in one fixed direction. Extrusion is studied in a more general form in Chap. 4.

A doubly curved surface is sometimes referred to as a *warped surface*. The best known example of such a surface is a horse saddle. Other occurrences are the human body, for example, and many engineering works like the aircraft fuselage mentioned above and the styled, aerodynamic shapes of sports cars.

3.3 Simple Solids

3.3.1 Cylinders

3.3.1.1 Description

A cylindrical surface is a ruled surface formed by a line segment called the cylinder *generatrix* that moves while remaining parallel to a fixed line, and touching a planar curve, called the cylinder *directrix*. Moreover, this curve lies in a plane intersecting the fixed line, called the cylinder axis, as shown in Fig. 3.5. The faces of the teeth of *spur* gears are cylindrical surfaces, as depicted in Fig. 3.6; these surfaces have a planar involute—see Fig. 2.19c—as their directrix. This is the reason why such gears are called *involute gears*, to distinguish them from other kinds, like *cycloidal gears*. The directrix of this alternative kind of gears is an *epicycloid*,[5] a curve illustrated in Fig. 2.19b.

If the above-mentioned generatrix remains perpendicular to the directrix, the cylinder is *straight*; otherwise, the cylinder is *oblique*. A right circular cylinder, whose directrix is a circle, is depicted in Fig. 3.5.

The representation of a *bounded* solid cylinder also calls, of course, for inequalities. For example, the solid cylinder of Fig. 3.5 is represented by

$$x^2 + z^2 \leq R^2, \quad 0 \leq y \leq h \tag{3.20}$$

---

[5] Only a small part of this curve is used to define the gear-tooth profile that is close to the X-axis in Fig. 2.19b.

Fig. 3.5 A bounded circular cylinder

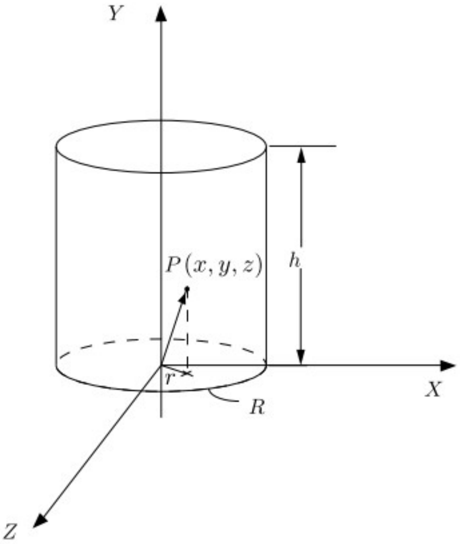

The first inequality excludes points lying a distance greater than R from the Y-axis, which is the cylinder axis; the second inequality restricts the points inside the cylinder to lie between the X–Y-plane and the plane parallel to the former a distance h above it.

In array form, the foregoing inequality becomes

$$\begin{bmatrix} x & y & z & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -R^2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \leq 0, \quad 0 \leq y \leq h \tag{3.21}$$

A pair of involute spur gears is used to transmit torque and motion between shafts of parallel axes. In order to design a spur-gear pair, two cylinders are produced with parallel axes and radii inversely proportional to the intended gear ratio. Next, an involute curve is produced around each of the bases of the circular cylinders. Now, with each of these curves as the directrix of a cylindrical surface, the flanks of the gears are generated. The reason behind the use of the involute lies in its geometric properties: upon meshing two involute profiles with a common tangent, each profile can turn about its cylinder axis, while maintaining a constant angle between the line of centers of the cylinder bases and the common normal. This angle is known as the *pressure angle*, which should be kept constant to obtain a constant gear ratio.

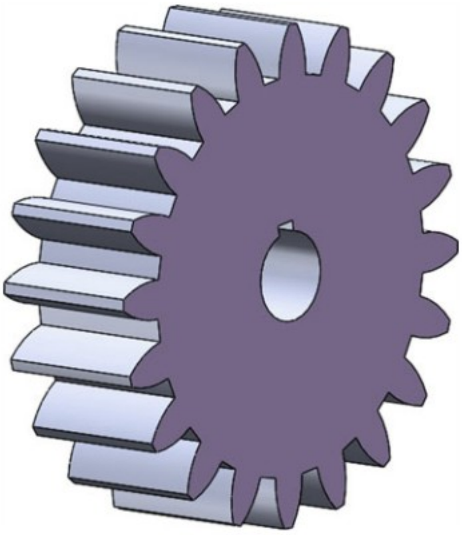

Fig. 3.6 A spur gear with its tooth flanks as instances of a cylindrical surface

3.3.2 *Cones*

3.3.2.1 Description

A conic surface is a ruled surface formed by a line, the cone *generatrix*, moving along a curved path, the cone *directrix*, such that the line always passes through one fixed point, called the *vertex*, a.k.a. the *apex*. Each position of the generatrix is called an *element* of the surface. Such surfaces, when bounded by a plane that cuts all elements, form *solid cones*. The surface defined by the teeth of a *bevel gear*, shown in Fig. 3.7, is a conic surface.

The simplest conic surfaces are those whose generatrix makes a *constant angle* with a fixed line, called its *axis*. Instances of solid cones produced from these surfaces are illustrated in Fig. 3.8, where we distinguish

- A *right cone*, shown in Fig. 3.8a, characterized by having its circular base normal to its axis.
- An *oblique cone*, as shown in Fig. 3.8b, whereby the axis is oblique to the circular base.
- A right cone to which a portion containing the apex has been cut off by means of a plane oblique to the axis gives rise to a *truncated cone*, as illustrated in Fig. 3.8c.
- A *frustum*, obtained if the foregoing cutting plane is normal to the axis, as depicted in Fig. 3.8d.

The flanks of the teeth of a bevel gear are made of conic surfaces. Commonly, the directrix of the conic surface in question is an *approximation* of the *spherical*

data 18.04.2020

IMPORTANT: ✡

Criar o meu sistema de coordenadas
(funções representam cada vetor)

→ CURVE VECTOR (provisório)

→ MUSICAL CURVE VECTOR.

voltando:

- Polar Coordinates,
- Cylindrical coordinates,
- Spherical coordinates,
- Absolute coordinates e
- World coordinates (x, y, z)
  versus relative coordinates
  (x', y', z').

B2) Pag. 13(31) –
Vectors

B3) Matrices
Pag. 17(35)

B4) Pag. 80(96) ✡
3.1.3 Distance of a Point to a Plane
(cartesiano)

data 18.04.2020

Projetá-lo sobre o espaço curvo
(veja curva de Gauss) - 179°
171...

Observation:
Use todo este livro com o livro anterior A.

C) Books_70_2020/
Geometry_Fundamentals-Belyaev/
geometry.pdf

Observation:
Utilizá-lo conjuntamente com os livros A e B.

D) Books_70_2020/
Mathematical Amazements and Surprises Fascinating Figures and noteworthy Numbers/
sanet.st_Mathematical....pdf

D1) Pág. 27 (26)
Armstrong Numbers ✡

)*+,)))-) ' 1 0!
-3 5 6 ! 777
= 3 > *2 & !

Congruence of Triangles:SSS

Lemma 1.3.10.1. *If points Z_1, Z_2 lie on opposite sides of a line a_{XY}, the congruences $XZ_1 \equiv XZ_2$, $YZ_1 \equiv YZ_2$ imply $\angle XYZ_1 \equiv XYZ_2$ (and $\angle YXZ_1 \equiv \angle YXZ_2$).*

Proof. $Z_1 a_{XY} Z_2 \Rightarrow \exists X'\ X' \in a_{XY} \,\&\, [Z_1X'Z_2]$. Observe that the lines a_{XY}, $a_{Z_1Z_2}$ meet only in X', because $Z_1 \notin a_{XY} \Rightarrow a_{Z_1Z_2} \neq a_{XY}$, and therefore for any Y' such that $Y' \in a_{XY}$, $Y' \in a_{Z_1Z_2}$, we have $X' \in a_{XY} \cap a_{Z_1Z_2} \,\&\, Y' \in a_{XY} \cap a_{Z_1Z_2} \overset{T1.1.1}{\Longrightarrow} Y' = X'$. [273] We also assume that $Y \notin a_{Z_1Z_2}$. [274] For the isosceles triangle $\triangle Z_1YZ_2$ the theorem T 1.3.3 gives $YZ_1 \equiv YZ_2 \Rightarrow \angle YZ_1Z_2 \equiv YZ_2Z_1$. On the other hand, $[Z_1X'Z_2] \overset{L1.2.11.15}{\Longrightarrow} Z_{1X'} = Z_{1Z_2} \,\&\, Z_{2X'} = Z_{2Z_1} \Rightarrow \angle YZ_1X' = \angle YZ_1Z_2 \,\&\, \angle YZ_2X' = \angle YZ_2Z_1$. Therefore, $\angle YZ_1Z_2 \equiv \angle YZ_2Z_1 \,\&\, \angle YZ_1X' = \angle YZ_1Z_2 \,\&\, \angle YZ_2Z = \angle YZ_2Z_1 \Rightarrow \angle YZ_1X' \equiv \angle YZ_2X'$.

Let Z_1, Z_2, X be collinear, i.e. $X \in a_{Z_1Z_2}$ (See Fig. 1.116, a)). Then we have $X \in a_{XY} \cap a_{Z_1Z_2} \Rightarrow X' = X$, and we can write $XZ_1 \equiv XZ_2 \,\&\, YZ_1 \equiv YZ_2 \,\&\, \angle YZ_1X' \equiv \angle YZ_2X' \overset{T1.3.4}{\Longrightarrow} \triangle XYZ_1 \equiv \triangle XYZ_2 \Rightarrow \angle XYZ_1 \equiv \angle XYZ_2 \,\&\, \angle YXZ_1 \equiv XYZ_2$. [275]

Now suppose neither X nor Y lie on $a_{Z_1Z_2}$. In this case $X' \in a_{XY} \,\&\, X' \neq X \neq Y \overset{T1.2.2}{\Longrightarrow} [X'XY] \vee [X'YX] \vee [XX'Y]$. Suppose $[X'XY]$ (See Fig. 1.116, b)). [276] Then $Y \in Z_{iY} \,\&\, X' \in Z_{iX'} \,\&\, X \in Z_{iX} \overset{L1.2.21.6,L1.2.21.4}{\Longrightarrow} Z_{iX} \subset Int\angle YZ_iX'$, where $i = 1, 2$. $[X'XY] \overset{L1.2.11.15}{\Longrightarrow} X'_X = X'_Y \Rightarrow \angle Z_1X'X = \angle Z_1X'Y \,\&\, \angle Z_2X'X = \angle Z_2X'Y$. Furthermore, arguing exactly as above, we see that $XZ_1X' = \angle XZ_1Z_2 \,\&\, \angle XZ_2X' = \angle XZ_2Z_1$, whence $\angle XZ_1X' \equiv \angle XZ_2X'$. Using T 1.3.9 we obtain $\angle XZ_1Y \equiv \angle XZ_2Y$, which allows us to write $XZ_1 \equiv XZ_2 \,\&\, YZ_1 \equiv YZ_2 \,\&\, \angle XZ_1Y \equiv \angle XZ_2Y \overset{T1.3.4}{\Longrightarrow} \triangle XZ_1Y \equiv \triangle XZ_2Y \Rightarrow \angle XYZ_1 \equiv XYZ_2 \,\&\, \angle YXZ_1 \equiv \angle YXZ_2$.

Finally, suppose $[XX'Y]$ (See Fig. 1.116, c)). Then $[XX'Y] \Rightarrow [YX'X] \overset{L1.2.11.15}{\Longrightarrow} Y_{X'} = Y_X \Rightarrow \angle XYZ_i \equiv \angle X'YZ_i$, where $i = 1, 2$. Together with $\angle X'YZ_1 \equiv \angle X'YZ_2$, this gives $\angle XYZ_1 \equiv \angle XYZ_2$. □

Theorem 1.3.10 (Third Triangle Congruence Theorem (SSS))**.** *If all sides of a triangle $\triangle ABC$ are congruent to the corresponding sides of a triangle $\triangle A'B'C'$, i.e. if $AB \equiv A'B'$, $BC \equiv B'C'$, $AC \equiv A'C'$, the triangle $\triangle ABC$ is congruent to the triangle $\triangle A'B'C'$. In other words, if a triangle $\triangle ABC$ is weakly congruent to a triangle $\triangle A'B'C'$, this implies that the triangle $\triangle ABC$ is congruent to the triangle $\triangle A'B'C'$.*

Proof. (See Fig. 1.117.) By hypothesis, $\triangle ABC \simeq \triangle A'B'C'$, i.e., $AB \equiv A'B'$, $BC \equiv B'C'$, $AC \equiv A'C'$. Using A 1.3.4, A 1.3.1, L 1.2.11.3, choose B'' so that $C'_{B''}C'_{B'}a_{A'C'}$, $\angle ACB \equiv \angle A'C'B''$, $BC \equiv B''C'$, and then choose B''' so that $C'_{B'''}a_{A'C'}C'_{B''}$.[277] Then we have $AC \equiv A'C' \,\&\, BC \equiv B''C' \,\&\, \angle ACB \equiv \angle A'C'B'' \overset{T1.3.4}{\Longrightarrow} \triangle ABC \equiv \triangle A'B''C' \Rightarrow AB \equiv A'B''$. $A'C' \equiv A'C' \,\&\, B''C' \equiv B'''C' \,\&\, \angle A'C'B'' \equiv \angle A'C'B''' \overset{T1.3.4}{\Longrightarrow} \triangle A'B''C' \equiv \triangle A'B'''C' \Rightarrow AB \equiv A'B''$. Since $AB \equiv A'B' \,\&\, AB \equiv A'B'' \,\&\, A'B'' \equiv A'B''' \,\&\, BC \equiv B'C' \,\&\, BC \equiv B''C' \,\&\, B''C' \equiv B'''C' \overset{T1.3.1}{\Longrightarrow} A'B''' \equiv A'B' \,\&\, A'B''' \equiv A'B'' \,\&\, B'''C' \equiv B'C' \,\&\, B'''C' \equiv B''C' \overset{T1.3.1}{\Longrightarrow} B''C' \equiv B'C'$, $B''B'a_{A'C'} \,\&\, B'''a_{A'C'}B'' \overset{L1.2.17.10}{\Longrightarrow} B'''a_{A'C'}B'$, we have $A'B''' \equiv A'B'' \,\&\, B'''C' \equiv B''C' \,\&\, B'''a_{A'C'}B'' \overset{L1.3.10.1}{\Longrightarrow} \angle A'C'B''' \equiv \angle A'C'B''$, $A'B''' \equiv A'B' \,\&\, B'''C' \equiv B'C' \,\&\, B'''a_{A'C'}B'$. Finally, $\angle A'C'B''' \equiv \angle A'C'B'' \,\&\, \angle A'C'B''' \equiv \angle A'C'B' \,\&\, C'_{B''}C'_{B'}a_{A'C'} \overset{A1.3.4}{\Longrightarrow} \angle A'C'B'' = \angle A'C'B' \Rightarrow C'_{B''} = C'_{B''}$, $C'B'' \equiv C'B' \,\&\, C'_{B''} = C'_{B'} \overset{A1.3.1}{\Longrightarrow} B'' = B'$. □

Congruence of Angles and Congruence of Paths as Equivalence Relations

Lemma 1.3.11.1. *If angles $\angle(h', k')$, $\angle(h'', k'')$ are both congruent to an angle $\angle(h, k)$, the angles $\angle(h', k')$, $\angle(h'', k'')$ are congruent to each other, i.e., $\angle(h', k') \equiv \angle(h'', k'')$ and $\angle(h'', k'') \equiv \angle(h', k')$.*

Proof. (See Fig. 1.118.) Denote O, O', O'' the vertices of the angles $\angle(h, k)$, $\angle(h', k')$, $\angle(h'', k'')$, respectively. Using A 1.3.1, choose $H \in h$, $K \in k$, $H'\ inh'$, $K' \in k'$, $H'' \in h''$ so that $OH \equiv O'H'$, $OK \equiv O'K'$, $OH \equiv O''H''$, $OK \equiv O''K''$, whence by T 1.3.1 $O'H' \equiv OH$, $O'K' \equiv OK$, $O''H'' \equiv OH$, $O''K'' \equiv OK$, and by L 1.2.21.1 $\angle HOK = \angle(h, k)$, $\angle H'O'K' = \angle(h', k')$, $\angle H''O''K'' = \angle(h'', k'')$. Then we have $O'H' \equiv OH \,\&\, O'K' \equiv OK \,\&\, \angle H'O'K' \equiv \angle HOK \,\&\, O''H'' \equiv OH \,\&\, O''K'' \equiv OK \,\&\, \angle H''O''K'' \overset{T1.3.4}{\Longrightarrow} \triangle H'O'K' \equiv \triangle HOK \,\&\, \triangle H''O''K'' \equiv \triangle HOK \Rightarrow K'H' \equiv KH \,\&\, K''H'' \equiv KH \overset{T1.3.1}{\Longrightarrow} K'H' \equiv K''H''$. Also, $O'H' \equiv OH \,\&\, O''H'' \equiv OH \,\&\, O'K' \equiv OK \,\&\, O''K'' \equiv$

---

[273] We take into account that, obviously, $[Z_1X'Z_2] \overset{L1.2.1.3}{\Longrightarrow} X' \in a_{Z_1Z_2}$.

[274] We can assume this without loss of generality - see next footnote.

[275] Observe that the seemingly useless fact that $\angle YXZ_1 \equiv YXZ_2$ allows us to avoid considering the case $Y \in a_{Z_1Z_2}$ separately. Instead, we can substitute X for Y and Y for X to obtain the desired result, taking advantage of the symmetry of the conditions of the theorem with respect to this substitution.

[276] Again, because of obvious symmetry with respect to substitution $X \to Y$, $Y \to X$, we do not need to consider the case when $[X'YX]$. Note that we could have avoided this discussion altogether if we united both cases $[X'XY]$, $[X'YX]$ into the equivalent $Y \in X'_X$, $Y \neq X$, but the approach taken here has the appeal of being more illustrative.

[277] To be more precise, we take a point B_0 such that $C'_{B''}a_{A'C'}B_0$, and then, using A 1.3.4, draw the angle $\angle A'C'B'''$ such that $C'_{B'''}B_0a_{A'C'}$, $\angle A'C'B'' \equiv \angle A'C'B'''$, $B''C' \equiv B'''C'$. We then have, of course, $C'_{B''}a_{A'C'}B_0 \,\&\, C'_{B'''}B_0a_{A'C'} \overset{L1.2.18.5}{\Longrightarrow} C'_{B'''}a_{A'C'}C'_{B''}$. Using jargon, as we did here, allows one to avoid cluttering the proofs with trivial details, thus saving the space and intellectual energy of the reader for more intricate points.

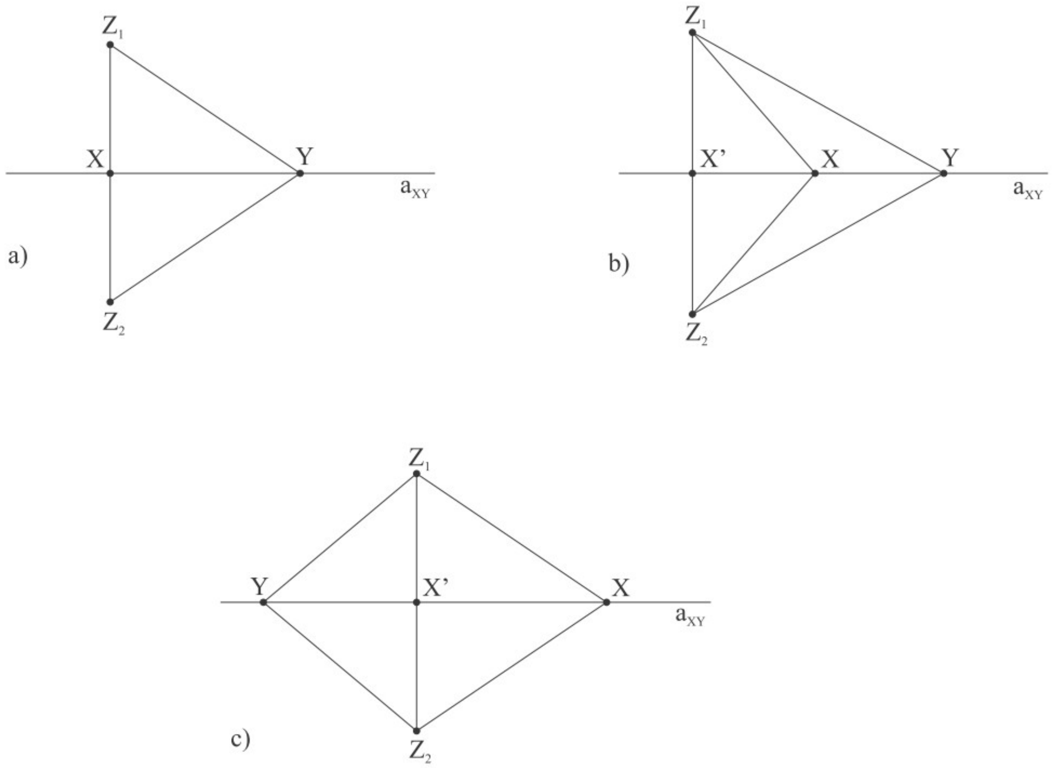

Z1
X
Y
a_XY
a)
Z2
Z1
X'
X
Y
a_XY
b)
Z2
Z1
Y
X'
X
a_XY
c)
Z2

Figure 1.116: If points Z_1, Z_2 lie on opposite sides of a line a_{XY}, the congruences $XZ_1 \equiv XZ_2$, $YZ_1 \equiv YZ_2$ imply $\angle XYZ_1 \equiv XYZ_2$. In a) Z_1, Z_2, X are collinear, i.e. $X \in a_{Z_1Z_2}$; in b), c) X, Y do not lie on $a_{Z_1Z_2}$ and $[X'XY]$ in b), $[XX'Y]$ in c).

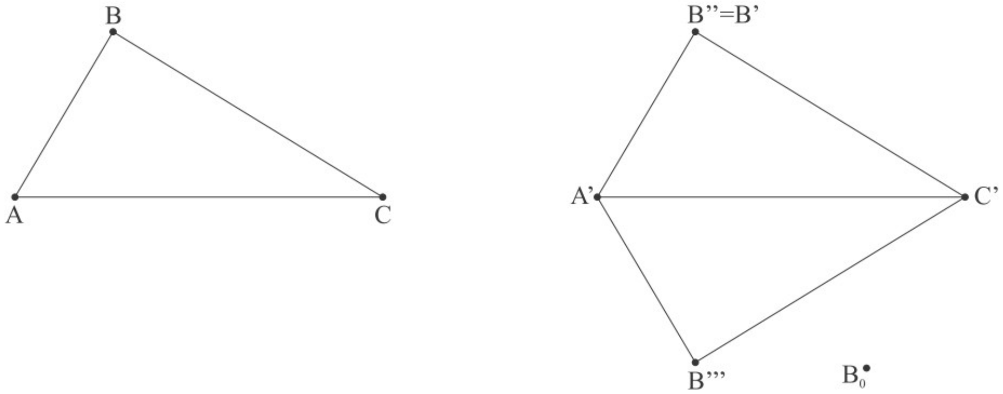

B
A
C
B''=B'
A'
C'
B'''
B0

Figure 1.117: if $AB \equiv A'B'$, $BC \equiv B'C'$, $AC \equiv A'C'$, the triangle $\triangle ABC$ is congruent to the triangle $\triangle A'B'C'$ (SSS, or The Third Triangle Congruence Theorem).

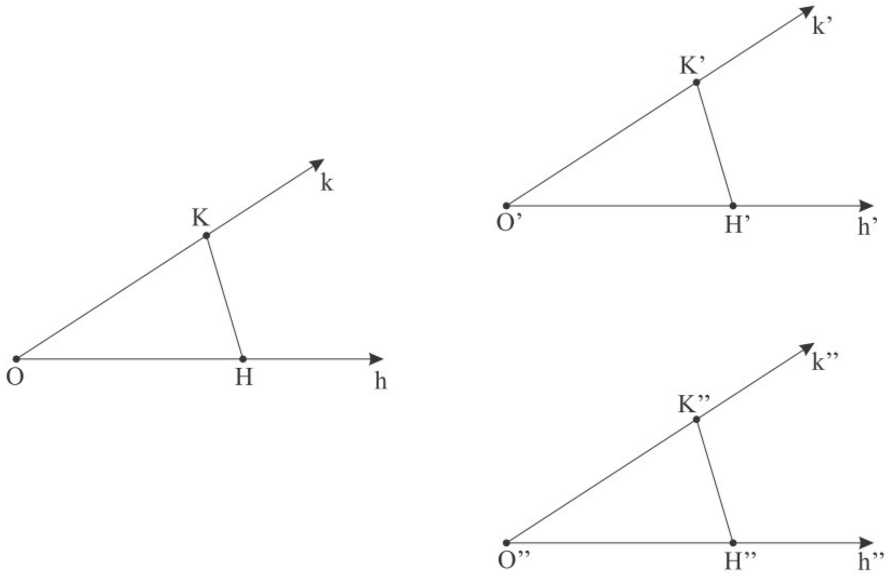

Figure 1.118: $\angle(h',k') \equiv \angle(h,k)$ and $\angle(h'',k'') \equiv \angle(h,k)$ imply $\angle(h',k') \equiv \angle(h'',k'')$ and $\angle(h'',k'') \equiv \angle(h',k')$.

$OK \overset{T1.3.1}{\Longrightarrow} O'H' \equiv O''H'' \,\&\, O'K' \equiv O''K''$. Finally, $O'H' \equiv O''H'' \,\&\, O'K' \equiv O''K'' \,\&\, K'H' \equiv K''H'' \overset{T1.3.10}{\Longrightarrow}$ $\triangle H'O'K' \equiv \triangle H''O''K'' \Rightarrow \angle H'O'K' \equiv \angle H''O''K'' \Rightarrow \angle(h',k') \equiv \angle(h'',k'')$. $\square$

Theorem 1.3.11. *Congruence of angles is a relation of equivalence on the class of all angles, i.e. it possesses the properties of reflexivity, symmetry, and transitivity.*

Proof. Reflexivity follows from A 1.3.4.

Symmetry: Let $\angle(h,k) \equiv \angle(h',k')$. Then $\angle(h',k') \equiv \angle(h',k') \,\&\, \angle(h,k) \equiv \angle(h',k') \overset{L1.3.11.1}{\Longrightarrow} \angle(h',k') \equiv \angle(h,k)$.

Transitivity: $\angle(h,k) \equiv \angle(h',k') \,\&\, \angle(h',k') \equiv \angle(h'',k'') \overset{\text{above}}{\Longrightarrow} \angle(h,k) \equiv \angle(h',k') \,\&\, \angle(h'',k'') \equiv \angle(h',k') \overset{L1.3.11.1}{\Longrightarrow}$ $\angle(h,k) \equiv \angle(h'',k'')$. $\square$

Therefore, if an angle $\angle(h,k)$ is congruent to an angle $\angle(h',k')$, we can say the angles $\angle(h,k)$, $\angle(h',k')$ are congruent (to each other).

Corollary 1.3.11.2. *Congruence of paths (in particular, of polygons) is a relation of equivalence on the class of all paths. That is, any path $A_1A_2\dots A_n$ is congruent to itself. If a path $A_1A_2\dots A_n$ is congruent to a path $B_1B_2\dots B_n$, the path $B_1B_2\dots B_n$ is congruent to the path $A_1A_2\dots A_n$. $A_1A_2\dots A_n \equiv B_1B_2\dots B_n$, $B_1B_2\dots B_n \equiv C_1C_2\dots C_n$ implies $A_1A_2\dots A_n \equiv C_1C_2\dots C_n$.*

Proof. $\square$

Again, if a path, in particular, a polygon, $A_1A_2\dots A_n$ is congruent to a path $B_1B_2\dots B_n$, we shall also say (and C 1.3.11.2 gives us the right to do so) that the paths $A_1A_2\dots A_n$ and $B_1B_2\dots B_n$ are congruent.

We are now in a position to prove theorem opposite to T 1.3.3.

Theorem 1.3.12. *If one angle, say, $\angle CAB$, of a triangle $\triangle ABC$ is congruent to another angle, say, $\angle ACB$, then $\triangle ABC$ is an isosceles triangle with $\angle ABC \equiv \angle ABC$.*

Proof. Let in a $\triangle ABC$ $\angle CAB \equiv \angle ACB$. Then by T 1.3.12 $\angle ACB \equiv \angle CAB$ and $AC \equiv \,\&\, \angle CAB \equiv \angle ACB \,\&\, \angle CAB \equiv$ $\angle ACB \overset{T1.3.11}{\Longrightarrow} \triangle CAB \equiv \triangle ACB \Rightarrow AB \equiv CB$. $\square$

Comparison of Intervals

Lemma 1.3.13.1. *For any point C lying on an open interval (AB), there are points $E, F \in (AB)$ such that $AC \equiv EF$.*

Proof. (See Fig. 1.119.) Suppose $[ACB]$. By T 1.2.1 $\exists F$ $[CFB]$. Then $[ACB] \,\&\, [CFB] \overset{L1.2.3.2}{\Longrightarrow} [ACF] \,\&\, [AFB]$. $[ACF] \,\&\, AF \equiv FA \overset{C1.3.9.2}{\Longrightarrow} \exists E\ [FEA] \,\&\, AC \equiv FE$. Finally, $[AEF] \,\&\, [AFB] \overset{L1.2.3.2}{\Longrightarrow} [AEB]$. $\square$

The following lemma is opposite, in a sense, to L 1.3.13.1

ARMSTRONG NUMBERS

Some numbers, often referred to as Armstrong numbers,[11] are each equal to the sum of their digits, each of which is taken to the power equal to the number of digits in the original number.

For example, the three-digit number 153 $=1^3+5^3+3^3$. Strangely enough, while there are no two-digit Armstrong numbers, there are four three-digit Armstrong numbers as follows (figure 1-8):

n	$100 \le n \le 999$	
153	$1^3 + 5^3 + 3^3 = 1 + 125 + 27$	= 153
370	$3^3 + 7^3 + 0^3 = 27 + 343 + 0$	= 370
371	$3^3 + 7^3 + 1^3 = 27 + 343 + 1$	= 371
407	$4^3 + 0^3 + 7^3 = 64 + 0 + 343$	= 407

Figure 1-8

There are three four-digit Armstrong numbers as seen below (figure 1-9).

n	$1{,}000 \le n \le 9{,}999$	
1,634	$1^4 + 6^4 + 3^4 + 4^4 = 1 + 1{,}296 + 81 + 256$	= 1,634
8,208	$8^4 + 2^4 + 0^4 + 8^4 = 4{,}096 + 16 + 0 + 4{,}096$	= 8,208
9,474	$9^4 + 4^4 + 7^4 + 4^4 = 6{,}561 + 256 + 2{,}401 + 256$	= 9,474

Figure 1-9

Among the five-digit numbers, there are three Armstrong numbers (figure 1-10).

n	$10{,}000 \le n \le 99{,}999$	
54,748	$5^5 + 4^5 + 7^5 + 4^5 + 8^5 = 3{,}125 + 1{,}024 + 16{,}807 + 1{,}024 + 32{,}768$	= 54,748
92,727	$9^5 + 2^5 + 7^5 + 2^5 + 7^5 = 59{,}049 + 32 + 16{,}807 + 32 + 16{,}807$	= 92,727
93,084	$9^5 + 3^5 + 0^5 + 8^5 + 4^5 = 59{,}049 + 243 + 0 + 32{,}768 + 1{,}024$	= 93,084

Figure 1-10

The only Armstrong number among the six-digit numbers is: 548,834.

$5^6 + 4^6 + 8^6 + 8^6 + 3^6 + 4^6 = 15{,}625 + 4{,}096 + 262{,}144 + 262{,}144 + 729 + 4{,}096 = 548{,}834$

---

[11] L. Deimel Jr. and M. Jones, "Finding Pluperfect Digital Invariants," *Journal of Recreational Mathematics* 42, no. 2 (1981–82): 87–107.

The following is a table (figure 1-11) that lists the Armstrong numbers of 7–10 digits.

k	**Armstrong numbers**
7	$n =$ 1,741,725; 4,210,818; 9,800,817; 9,926,315
8	$n =$ 24,678,050; 24,678,051; 88,593,477
9	$n =$ 146,511,208; 472,335,975; 534,494,836; 912,985,153
10	$n =$ 4,679,307,774

Figure 1-11

Here is a number that is 39 digits long:

115,132,219,018,763,992,565,095,597,973,971,522,401

and is equal to the sum of its digits, each of which is taken to the thirty-ninth power, thus making it an Armstrong number. Furthermore, it is the largest Armstrong number. In other words,

$$\begin{aligned}&1^{39}+1^{39}+5^{39}+1^{39}+3^{39}+2^{39}+2^{39}+1^{39}+9^{39}+0^{39}+1^{39}+8^{39}+7^{39}+6^{39}+3^{39}+9^{39}\\&\quad+9^{39}+2^{39}+5^{39}+6^{39}+5^{39}+0^{39}+9^{39}+5^{39}+5^{39}+9^{39}+7^{39}+9^{39}+7^{39}+3^{39}+9^{39}\\&\quad+7^{39}+1^{39}+5^{39}+2^{39}+2^{39}+4^{39}+0^{39}+1^{39}\\&\quad=115{,}132{,}219{,}018{,}763{,}992{,}565{,}095{,}597{,}973{,}971{,}522{,}401\\&\quad\approx 1.151322190\cdot 10^{38}\end{aligned}$$

The Armstrong Numbers

k	**Armstrong numbers of k-digit lengths**
1	0; 1; 2; 3; 4; 5; 6; 7; 8; 9
3	153; 370; 371; 407
4	1,634; 8,208; 9,474
5	54,748; 92,727; 93,084
6	548,834
7	1,741,725; 4,210,818; 9,800,817; 9,926,315
8	24,678,050; 24,678,051; 88,593,477
9	146,511,208; 472,335,975; 534,494,836; 912,985,153
10	4,679,307,774
11	32,164,049,650; 32,164,049,651; 40,028,394,225; 42,678,290,603; 44,708,635,679; 49,388,550,606; 82,693,916,578; 94,204,591,914
14	28,116,440,335,967

k	**Armstrong numbers of *k*-digit lengths**
16	4,338,281,769,391,370; 4,338,281,769,391,371
17	21,897,142,587,612,075; 35,641,594,208,964,132; 35,875,699,062,250,035
19	1,517,841,543,307,505,039; 3,289,582,984,443,187,032; 4,498,128,791,164,624,869; 4,929,273,885,928,088,826
20	63,105,425,988,599,693,916
21	128,468,643,043,731,391,252; 449,177,399,146,038,697,307
23	21,887,696,841,122,916,288,858; 27,879,694,893,054,074,471,405; 27,907,865,009,977,052,567,814; 28,361,281,321,319,229,463,398; 35,452,590,104,031,691,935,943
24	174,088,005,938,065,293,023,722; 188,451,485,447,897,896,036,875; 239,313,664,430,041,569,350,093
25	1,550,475,334,214,501,539,088,894; 1,553,242,162,893,771,850,669,378; 3,706,907,995,955,475,988,644,380; 3,706,907,995,955,475,988,644,381; 4,422,095,118,095,899,619,457,938
27	121,204,998,563,613,372,405,438,066; 121,270,696,006,801,314,328,439,376; 128,851,796,696,487,777,842,012,787; 174,650,464,499,531,377,631,639,254; 177,265,453,171,792,792,366,489,765
29	14,607,640,612,971,980,372,614,873,089; 19,008,174,136,254,279,995,012,734,740; 19,008,174,136,254,279,995,012,734,741; 23,866,716,435,523,975,980,390,369,295
31	1,145,037,275,765,491,025,924,292,050,346; 1,927,890,457,142,960,697,580,636,236,639; 2,309,092,682,616,190,307,509,695,338,915
32	17,333,509,997,782,249,308,725,103,962,772
33	186,709,961,001,538,790,100,634,132,976,990; 186,709,961,001,538,790,100,634,132,976,991
34	1,122,763,285,329,372,541,592,822,900,204,593
35	12,639,369,517,103,790,328,947,807,201,478,392; 12,679,937,780,272,278,566,303,885,594,196,922
37	1,219,167,219,625,434,121,569,735,803,609,966,019
38	12,815,792,078,366,059,955,099,770,545,296,129,367
39	115,132,219,018,763,992,565,095,597,973,971,522,400; 115,132,219,018,763,992,565,095,597,973,971,522,401

Figure 1-12

There are no Armstrong numbers for $k = 2, 12, 13, 15, 18, 22, 26, 28, 30$, and 36 (and $k > 39$). In fact, there are only eighty-nine Armstrong numbers in the decimal system.

The following is a list of the *consecutive* Armstrong numbers:

$k = 3$: 370; 371
$k = 8$: 24,678,050; 24,678,051
$k = 11$: 32,164,049,650; 32,164,049,651
$k = 16$: 4,338,281,769,391,370; 4,338,281,769,391,371
$k = 25$: 3,706,907,995,955,475,988,644,380; 3,706,907,995,955,475,988,644,381
$k = 29$: 19,008,174,136,254,279,995,012,734,740;
19,008,174,136,254,279,995,012, 734,741
$k = 33$: 186,709,961,001,538,790,100,634,132,976,990;
186,709,961,001,538,790,100, 634,132,976,991
$k = 39$: 115,132,219,018,763,992,565,095,597,973,971,522,400
115,132,219,018,763, 992,565,095,597,973,971,522,401

Incidentally, our first Armstrong number, 153, has some other amazing properties. It is a triangular number, where

$$1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10 + 11 + 12 + 13 + 14 + 15 + 16 + 17 = 153$$

The number 153 is not only equal to the sum of the cubes of its digits ($1^3 + 5^3 + 3^3 = 1 + 125 + 27 = 153$) but it is also a number that can be expressed as the sum of consecutive factorials:[12] $1! + 2! + 3! + 4! + 5! = 153$.

Can you discover other properties of this ubiquitous number 153?

MORE NUMBER CURIOSITIES

As an "extension" of the Armstrong numbers, we will present numbers where the sum of the digits of the original number—each taken to the power one more or less than the number of digits in the original number—is equal to the original number. This differs from the Armstrong numbers, where the power to which each of the digits was taken was equal to the number of digits in the original number.

To begin, consider the following four-digit numbers where each of the digits is take to the fifth power and the sum is equal to the original number.

$4{,}150 = 4^5 + 1^5 + 5^5 + 0^5$
$4{,}151 = 4^5 + 1^5 + 5^5 + 1^5$

---

[12] $n!$ reads "n factorial" and equals $n\cdot(n-1)\cdot(n-2)\cdot\ldots\cdot3\cdot2\cdot1$.

The following are some other examples of non-Armstrong numbers:

$194{,}979 = 1^5 + 9^5 + 4^5 + 9^5 + 7^5 + 9^5$
$14{,}459{,}929 = 1^7 + 4^7 + 4^7 + 5^7 + 9^7 + 9^7 + 2^7 + 9^7$

Here is a forty-one-digit number[13] that is equal to the sum of the digits, each taken to the forty-second power: 36,428,594,490,313,158,783,584,452,532,870,892,261,556

$3^{42} + 6^{42} + 4^{42} + 2^{42} + 8^{42} + 5^{42} + 9^{42} + 4^{42} + 4^{42} + 9^{42} + 0^{42} + 3^{42} + 1^{42} + 3^{42} + 1^{42} + 5^{42}$
$+ 8^{42} + 7^{42} + 8^{42} + 3^{42} + 5^{42} + 8^{42} + 4^{42} + 4^{42} + 5^{42} + 2^{42} + 5^{42} + 3^{42} + 2^{42} + 8^{42} + 7^{42}$
$+ 0^{42} + 8^{42} + 9^{42} + 2^{42} + 2^{42} + 6^{42} + 1^{42} + 5^{42} + 5^{42} + 6^{42}$
$= 36{,}428{,}594{,}490{,}313{,}158{,}783{,}584{,}452{,}532{,}870{,}892{,}261{,}556$

The relationships between numbers and powers can be fascinating. Here are a few to relish.

Notice how, in these cases, quite the opposite of the previous numbers, the powers reflect the original number and the base stays the same.

$4{,}624 = 4^4 + 4^6 + 4^2 + 4^4$
$1{,}033 = 8^1 + 8^0 + 8^3 + 8^3$
$595{,}968 = 4^5 + 4^9 + 4^5 + 4^9 + 4^6 + 4^8$
$3{,}909{,}511 = 5^3 + 5^9 + 5^0 + 5^9 + 5^5 + 5^1 + 5^1$
$13{,}177{,}388 = 7^1 + 7^3 + 7^1 + 7^7 + 7^7 + 7^3 + 7^8 + 7^8$
$52{,}135{,}640 = 19^5 + 19^2 + 19^1 + 19^3 + 19^5 + 19^6 + 19^4 + 19^0$

Then there are those numbers that differ from the previous examples and lend themselves to an even more astonishing pattern: here the powers and the bases match the original digits.[14]

$3{,}435 = 3^3 + 4^4 + 3^3 + 5^5$
$438{,}579{,}088 = 4^4 + 3^3 + 8^8 + 5^5 + 7^7 + 9^9 + 0^0 + 8^8 + 8^8$

We can also find a reversal of this pattern, namely, where the bases go in the proper order and the exponents progress in the reverse order.

$48{,}625 = 4^5 + 8^2 + 6^6 + 2^8 + 5^4$
$397{,}612 = 3^2 + 9^1 + 7^6 + 6^7 + 1^9 + 2^3$

---

[13] See L. E. Deimel Jr. and M. T. Jones, *Journal of Recreational Mathematics* 14, no. 4 (1981–82): 284.

[14] Although 0^0 is undefined, we shall value it at 0 for our purposes here.

Searching through our number system reveals that there are also some entertaining patterns where consecutive exponents are used:

$$\begin{aligned}
43 &= 4^2 + 3^3\\
63 &= 6^2 + 3^3\\
89 &= 8^1 + 9^2\\
1{,}676 &= 1^5 + 6^4 + 7^3 + 6^2
\end{aligned}$$

Yet when we formalize this a bit and use the consecutive natural numbers as exponents, we have the following amazing relationships:

$$\begin{aligned}
135 &= 1^1 + 3^2 + 5^3\\
175 &= 1^1 + 7^2 + 5^3\\
518 &= 5^1 + 1^2 + 8^3\\
598 &= 5^1 + 9^2 + 8^3\\
1{,}306 &= 1^1 + 3^2 + 0^3 + 6^4\\
1{,}676 &= 1^1 + 6^2 + 7^3 + 6^4\\
2{,}427 &= 2^1 + 4^2 + 2^3 + 7^4\\
2{,}646{,}798 &= 2^1 + 6^2 + 4^3 + 6^4 + 7^5 + 9^6 + 8^7
\end{aligned}$$

You may have noticed that 1,676 appeared in both lists of consecutive exponents.

A further charming occurrence in mathematics—in the decimal system, of course—is seen with the following illustrations. Although you may find this a bit contrived, it simply highlights the notion that there are probably limitless nuggets we can cull from our numbers system.

$$\mathbf{16}^3 + \mathbf{50}^3 + \mathbf{33}^3 = 4{,}096 + 125{,}000 + 35{,}937 = \mathbf{16\ 5{,}0\ 33} = 165{,}033$$

$$\mathbf{166}^3 + \mathbf{500}^3 + \mathbf{333}^3 = 4{,}574{,}296 + 125{,}000{,}000 + 36{,}926{,}037 = \mathbf{166,\ 500,\ 333} = 166{,}500{,}333$$

$$\mathbf{588}^2 + \mathbf{2{,}353}^2 = 345{,}744 + 5{,}536{,}609 = \mathbf{5{,}88\ 2{,}353} = 5{,}882{,}353$$

MORE UNUSUAL NUMBER RELATIONSHIPS

There are a number of unusual relationships between certain numbers—as represented in the decimal system. There is not much explanation for them. So just enjoy them and see if you can find others like them.

We are going to present pairs of numbers where the product and the sum are reversals of each other (figure 1-13).

The two numbers		**their *product***	**their *sum***
9	9	81	18
3	24	72	27
2	47	94	49
2	497	994	499

Figure 1-13

Moving along to other number curiosities, consider the number 3,025. Suppose we split the number as 30 and 25, and then add these two numbers to get $30 + 25 = 55$. Then, if we square this result, we surprisingly get back to the original number: $55^2 = 3{,}025$. The question that immediately arises is "Are there other numbers for which this little procedure also works?"

Well, here are a few more such numbers:

$9{,}801 \rightarrow 98 + 01 = 99$ and $99^2 = 9{,}801$
$2{,}025 \rightarrow 20 + 25 = 45$ and $45^2 = 2{,}025$
$088{,}209 \rightarrow 088 + 209 = 297$ and $297^2 = 88{,}209$
$494{,}209 \rightarrow 494 + 209 = 703$ and $703^2 = 494{,}209$
$998{,}001 \rightarrow 998 + 001 = 999$ and $999^2 = 998{,}001$
$99{,}980{,}001 \rightarrow 9998 + 0001 = 9{,}999$ and $9{,}999^2 = 99{,}980{,}001$

Perhaps you can now find other numbers where this amazing relationship will hold.

While we are in the mode of "splitting" numbers, let's partition numbers to get some other fascinating relationships

$1{,}233 \rightarrow 12|\,33 = 12^2 + 33^2$
$8{,}833 \rightarrow 88|\,33 = 88^2 + 33^2$
$990{,}100 \rightarrow 990|\,100 = 990^2 + 100^2$
$94{,}122{,}353 \rightarrow 9412|\,2353 = 9{,}412^2 + 2{,}353^2$
$7{,}416{,}043{,}776 \rightarrow 74160|\,43776 = 74{,}160^2 + 43{,}776^2$
$116{,}788{,}321{,}168 \rightarrow 116788|\,321168 = 116{,}788^2 + 321{,}168^2$
$221{,}859 \rightarrow 22|\,18|\,59 = 22^3 + 18^3 + 59^3$
$166{,}500{,}333 \rightarrow 166|\,500|\,333 = 166^3 + 500^3 + 333^3$

You might like to verify some of the following that can be split, where the sum of the squares of the parts will equal the original number:

$10{,}100 \rightarrow 10|100 = 10^2 + 100^2$
$5{,}882{,}353 \rightarrow 588|2353 = 588^2 + 2{,}353^2$

Here are a few more such numbers: 9,901,0|09,901, 1,765,0|38,125, 2,584,0|43,776, 999,900|010,000, 123,288|328,768, and there are still more. You might want to verify this relationship. You can also find more of these numbers.

An analogous situation—this time using differences instead of sums—also produces some interesting relationships:

$48 \rightarrow 4|8 = 8^2 - 4^2$
$3{,}468 \rightarrow 34|68 = 68^2 - 34^2$
$416{,}768 \rightarrow 416|768 = 768^2 - 416^2$
$33{,}346{,}668 \rightarrow 3334|6668 = 6{,}668^2 - 3{,}334^2$

You might like to verify some of the following that can be split, where the difference of the squares of the parts will equal the original number. 16|128, 34|188, 140|400, 484|848, 530|901, 334|668, 234|1,548, and others.

AN UNUSUAL SITUATION

Sometimes numbers can be changed by accident. In a recent publication, a computer glitch dropped all the exponents to the base line. So that 8^3 was shown as 83. Would there be a value that would not be affected by such a computer failure? You can try all the possibilities and you may stumble upon $2^5 \cdot 9^2 = 2{,}592$, which is believed to be the only such case.

FRIENDLY NUMBERS

What could possibly make two numbers friendly? Your first reaction might be that this means numbers that are friendly to you. Actually, we will be inspecting numbers that are "friendly" to each other. Mathematicians have decided that two numbers are considered "friendly" (or as often used in the more sophisticated literature, "amicable") if the sum of the proper divisors[15] of one number equals the second number *and* the sum of the proper divisors of the second number equals the first number.

---

[15] *Proper divisors* of a number *n* are divisors that are smaller than the number *n*. That means 1 is a proper divisor, but *n* is not.

Sounds complicated? It's not at all difficult to understand. Let's look at the smallest pair of friendly numbers: 220 and 284.

The divisors of **220** (other than 220 itself) are 1, 2, 4, 5, 10, 11, 20, 22, 44, 55, and 110. Their sum is 1 + 2 + 4 + 5 + 10 + 11 + 20 + 22 + 44 + 55 + 110 = **284**.

The divisors of **284** (other than 284 itself) are 1, 2, 4, 71, and 142, and their sum is 1 + 2 + 4 + 71 + 142 = **220**. This shows that the two numbers are friendly numbers.

Although this pair of friendly numbers was already known to Pythagoras about 500 BCE, a second pair of friendly numbers, first discovered in 1636 by the French mathematician Pierre de Fermat (1607–1665),[16] is 17,296 and 18,416.

$17{,}296 = 2^4 \cdot 23 \cdot 47$, and $18{,}416 = 2^4 \cdot 1{,}151$

The sum of the factors of 17,296 is
1 + 2 + 4 + 8 + 16 + 23 + 46 + 47 + 92 + 94 + 184 + 188 + 368 + 376 + 752 + 1,081 + 2,162 + 4,324 + 8,648 = $\underline{18{,}416}$

The sum of the factors of 18,416 is
1 + 2 + 4 + 8 + 16 + 1,151 + 2,302 + 4,604 + 9,208 = $\underline{17{,}296}$

Thus, they, too, are truly friendly numbers!

Two years later, the French mathematician René Descartes (1596–1650) found another pair of friendly numbers: 9,363,584 and 9,437,056. By 1747, the Swiss mathematician Leonhard Euler (1707–1783) discovered sixty pairs of friendly numbers, yet he seemed to have overlooked the second smallest pair, which was discovered in 1866 by the sixteen-year-old Nicolò Paganini:

The sum of the factors of 1,184 is 1 + 2 + 4 + 8 + 16 + 32 + 37 + 74 + 148 + 296 + 592 = 1,210

And the sum of the factors of 1,210 is 1 + 2 + 5 + 10 + 11 + 22 + 55 + 110 + 121 + 242 + 605 = 1,184

We now have identified over 363,000 pairs of friendly numbers. Here are a few of these:

2,620 and 2,924
5,020 and 5,564
6,232 and 6,368
10,744 and 10,856
12,285 and 14,595
63,020 and 76,084
111,448,537,712 and 118,853,793,424

---

[16] There is evidence that Fermat's discovery was anticipated by the Morroccan mathematician Ibn al-Banna al-Marakushi al-Azdi (1256–ca. 1321).

You might want to verify the above pairs' "friendliness"!

For the experts, the following is one method for finding friendly numbers:

$$\text{Let}\quad \begin{aligned} a &= 3\cdot 2^n - 1 \\ b &= 3\cdot 2^{n-1} - 1 \\ c &= 3^2\cdot 2^{2n-1} - 1 \end{aligned}$$

where n is an integer ≥ 2 and where a, b, and c are all prime numbers, then $2^n\cdot ab$ and $2^n\cdot c$ are friendly numbers.
(Notice that for $n = 2$, 4, and 7 that a, b, and c are all prime for $n \leq 200$.)

Here is a list of friendly numbers less than 10,000,000:

	First number	Second number	Year of discovery	Discoverer
1	220	284	-	Pythagoras
2	1,184	1,210	1860	Paganini
3	2,620	2,924	1747	Euler
4	5,020	5,564	1747	Euler
5	6,232	6,368	1747	Euler
6	10,744	10,856	1747	Euler
7	12,285	14,595	1939	Brown
8	17,296	18,416	ca. 1310/1636	Ibn al-Banna/Fermat
9	63,020	76,084	1747	Euler
10	66,928	66,992	1747	Euler
11	67,095	71,145	1747	Euler
12	69,615	87,633	1747	Euler
13	79,750	88,730	1964	Rolf
14	100,485	124,155	1747	Euler
15	122,265	139,815	1747	Euler
16	122,368	123,152	1941/42	Poulet
17	141,664	153,176	1747	Euler
18	142,310	168,730	1747	Euler
19	171,856	176,336	1747	Euler
20	176,272	180,848	1747	Euler
21	185,368	203,432	1966	Alanen/Ore/Stempel
22	196,724	202,444	1747	Euler
23	280,540	365,084	1966	Alanen/Ore/Stempel
24	308,620	389,924	1747	Euler
25	319,550	430,402	1966	Alanen/Ore/Stempel
26	356,408	399,592	1921	Mason
27	437,456	455,344	1747	Euler

	First number	Second number	Year of discovery	Discoverer
28	469,028	486,178	1966	Alanen/Ore/Stempel
29	503,056	514,736	1747	Euler
30	522,405	525,915	1747	Euler
31	600,392	669,688	1921	Mason
32	609,928	686,072	1747	Euler
33	624,184	691,256	1921	Mason
34	635,624	712,216	1921	Mason
35	643,336	652,664	1747	Euler
36	667,964	783,556	1966	Alanen/Ore/Stempel
37	726,104	796,696	1921	Mason
38	802,725	863,835	1966	Alanen/Ore/Stempel
39	879,712	901,424	1966	Alanen/Ore/Stempel
40	898,216	980,984	1747	Euler
41	947,835	1,125,765	1946	Escott
42	998,104	1,043,096	1966	Alanen/Ore/Stempel
43	1,077,890	1,099,390	1966	Lee
44	1,154,450	1,189,150	1957	Garcia
45	1,156,870	1,292,570	1946	Escott
46	1,175,265	1,438,983	1747	Euler
47	1,185,376	1,286,744	1929	Gerardin
48	1,280,565	1,340,235	1747	Euler
49	1,328,470	1,483,850	1966	Lee
50	1,358,595	1,486,845	1747	Euler
51	1,392,368	1,464,592	1747	Euler
52	1,466,150	1,747,930	1966	Lee
53	1,468,324	1,749,212	1967	Bratley/McKay
54	1,511,930	1,598,470	1946	Escott
55	1,669,910	2,062,570	1966	Lee
56	1,798,875	1,870,245	1967	Bratley/McKay
57	2,082,464	2,090,656	1747	Euler
58	2,236,570	2,429,030	1966	Lee
59	2,652,728	2,941,672	1921	Mason
60	2,723,792	2,874,064	1929	Poulet
61	2,728,726	3,077,354	1966	Lee
62	2,739,704	2,928,136	1747	Euler
63	2,802,416	2,947,216	1747	Euler
64	2,803,580	3,716,164	1967	Bratley/McKay
65	3,276,856	3,721,544	1747	Euler
66	3,606,850	3,892,670	1967	Bratley/McKay
67	3,786,904	4,300,136	1747	Euler

	First number	Second number	Year of discovery	Discoverer
68	3,805,264	4,006,736	1929	Poulet
69	4,238,984	4,314,616	1967	Bratley/McKay
70	4,246,130	4,488,910	1747	Euler
71	4,259,750	4,445,050	1966	Lee
72	4,482,765	5,120,595	1957	Garcia
73	4,532,710	6,135,962	1957	Garcia
74	4,604,776	5,162,744	1966	Lee
75	5,123,090	5,504,110	1966	Lee
76	5,147,032	5,843,048	1747	Euler
77	5,232,010	5,799,542	1967	Bratley/McKay
78	5,357,625	5,684,679	1966	Lee
79	5,385,310	5,812,130	1967	Bratley/McKay
80	5,459,176	5,495,264	1967	Lee
81	5,726,072	6,369,928	1921	Mason
82	5,730,615	6,088,905	1966	Lee
83	5,864,660	7,489,324	1967	Bratley/McKay
84	6,329,416	6,371,384	1966	Lee
85	6,377,175	6,680,025	1966	Lee
86	6,955,216	7,418,864	1946	Escott
87	6,993,610	7,158,710	1957	Garcia
88	7,275,532	7,471,508	1967	Bratley/McKay
89	7,288,930	8,221,598	1966	Lee
90	7,489,112	7,674,088	1966	Lee
91	7,577,350	8,493,050	1966	Lee
92	7,677,248	7,684,672	1884	Seelhoff
93	7,800,544	7,916,696	1929	Gerardin
94	7,850,512	8,052,488	1966	Lee
95	8,262,136	8,369,864	1966	Lee
96	8,619,765	9,627,915	1957	Garcia
97	8,666,860	10,638,356	1966	Lee
98	8,754,130	10,893,230	1946	Escott
99	8,826,070	10,043,690	1967	Bratley/McKay
100	9,071,685	9,498,555	1946	Escott
101	9,199,496	9,592,504	1929	Gerardin/Poulet
102	9,206,925	10,791,795	1967	Bratley/McKay
103	9,339,704	9,892,936	1966	Lee
104	9,363,584	9,437,056	ca. 1600/1638	Yazdi/Descartes
105	9,478,910	11,049,730	1967	Bratley/McKay
106	9,491,625	10,950,615	1967	Bratley/McKay
107	9,660,950	10,025,290	1966	Lee
108	9,773,505	11,791,935	1967	Bratley/McKay

Figure 1-14

data 18.04.2020

Projetá-lo sobre o espaço curvo
(veja curva de Gauss) - 179°
177...

Observation:
Use todo este livro com o
livro anterior A.

C) Books_70_2020/
Geometry_Fundamentals_Belyaev/
geometry.pdf

Observation:
Utilizá-lo conjuntamente
com os livros A e B.

D) Books_70_2020/
Mathematical Amazements and
Surprises Fascinating Figures and
noteworthy Numbers/
saret.st_Mathematical...pdf

D1) Pág. 27 (26)
Armstrong Numbers ✡

18 04 2020

K=3 → 3 numbers

$$153 = 1^3 + 5^3 + 3^3 = 153$$

K = 39 obtemos dos números com final par e ímpar, ambos com 39 números.

... 400

... 401

Obs. Utilizaremos estes números para dividir uma área.

Important: ✡

153 is a triangular number

$1+2+3+4+5+6+7+8+9+10+$

$11+12+13+14+15+16+17 = 153$.

Obs:

153 is also a number that can be expressed as the sum of consecutive factorials:

data 18.04.2020

$1! + 2! + 3! + 4! + 5! = 153$

D2) Pag. 147 (146)

Geometry Surprises

Pag. 152 (151)

Scalene triangle →

all three sides are of different lengths.

The medians partition the triangle into six smaller triangles, which happen to be equal in area.

We can also partition any triangle into four congruent triangles by using the midpoints of the three sides and joining them to form a medial triangle.

Point G (the point of intersection of the medians) is the center of gravity of the triangle, and is called the centroid! ✡

↳ Fascinating fact about the centroid of a triangle;

↳ IMPORTANT: ✡

! ('/ '. < 1 0

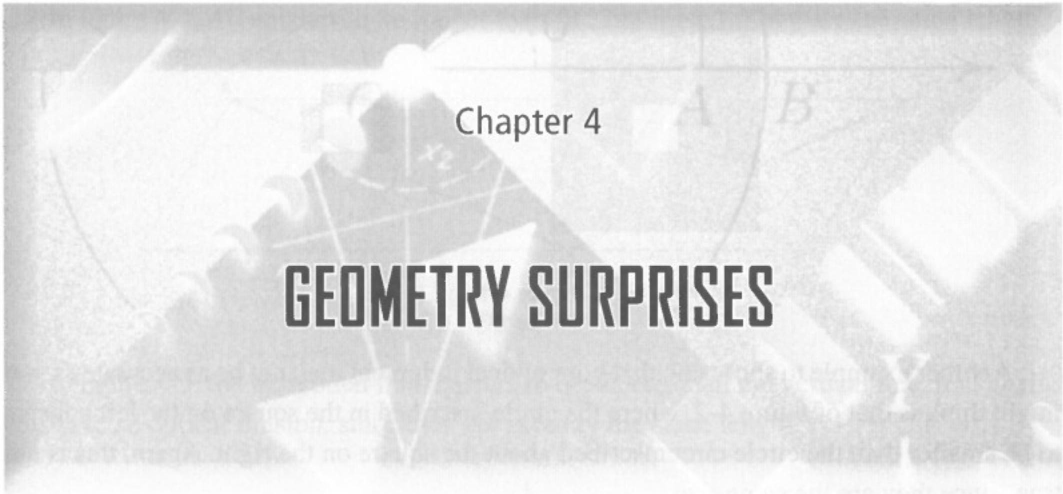

Chapter 4

GEOMETRY SURPRISES

The topic of geometry evokes, in many people, their high school memories of the subject. Usually, these recollections encompass the many proofs they were required to do to establish geometric relationships. Too often, the process of the proofs became the focus, overshadowing the actual relationships being proved and, unfortunately, leaving the beauty of geometry somewhat neglected. This chapter is intended to reverse that process. In order to demonstrate the beauty of geometry, we will present some truly awe-inspiring relationships—all of which (obviously) can be proved—but, so as not to disturb the "flow" of the presentation, we will merely refer you to sources where the proofs can be found. (See the bibliography for some suggested books.)

Join us now as we explore some truly amazing geometric relationships. However, before we embark on our journey through these geometric wonders, we must caution you that appearance alone can be misleading. Sometimes what appears to be true is not, and what appears impossible can be possible. This makes our journey through geometry that much more intriguing. To demonstrate that, we offer some optical illusions for your entertainment.

OPTICAL ILLUSIONS

In geometry, what you see is not always true. For example, in figure 4-1 you see two white squares. Most observers would say that the left-side white square is a bit smaller than the white square on the right. This is incorrect. They are both the same size. Their context and presentation make for an optical illusion.

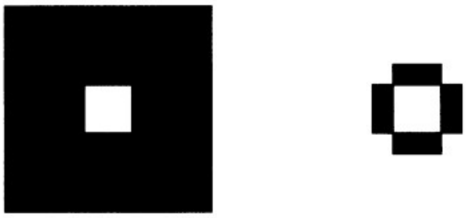

Figure 4-1

Another example to show you that your optical judgment may not be as accurate as you might think is that of figure 4-2, where the circle inscribed in the square on the left appears to be smaller than the circle circumscribed about the square on the right. Again, this is not true, since they are the same size.

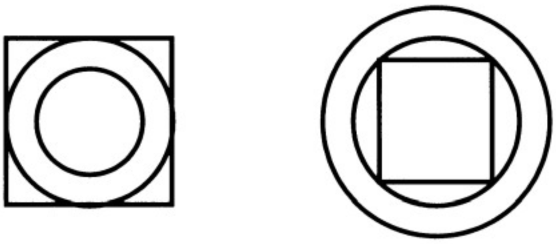

Figure 4-2

To further "upset" your sense of perception, consider the following few optical illusions. In figure 4-3, the center (black) circle on the left appears to be smaller than the center (black) circle on the right. Again, not so. They are both of equal size.

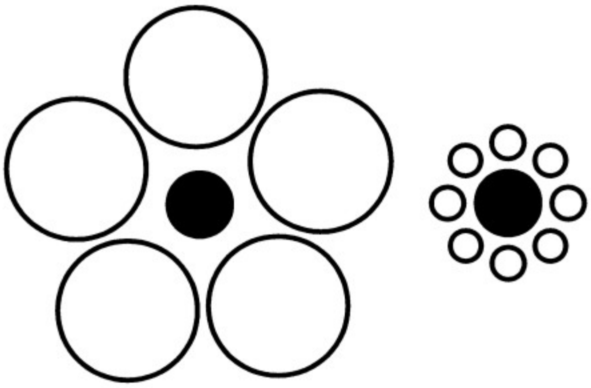

Figure 4-3

In figure 4-4, *AB* appears to be longer than *BC*, but it isn't. They are the same length.

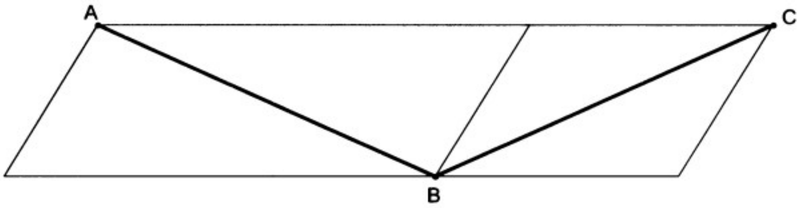

Figure 4-4

In figure 4-5, the horizontal line appears to be shorter than the vertical line, but, again, you have an optical illusion, since they are actually the same length.

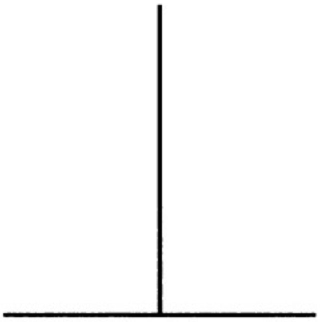

Figure 4-5

Some optical illusions are done intentionally. Take, for example, the *tribar* or *Penrose triangle* (figure 4-6). Popularized by Roger Penrose (1931–) in 1958, it was previously created in 1934 by Oscar Reutersvärd (1915–2002), a Swedish artist. It appears to be a triangle with three right angles. His invention was honored by the issuance of a postage stamp in 1982 (see figure 4-7).[1]

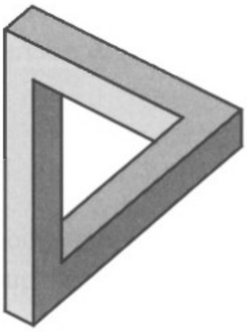

Figure 4-6

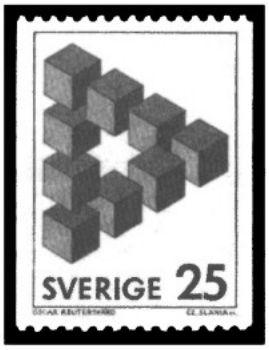

Figure 4-7

---

[1] Swedish postage stamp: 25 Öre, Sverige, February 16, 1982.

A postage stamp depicting an "irrational" cube was issued by the Republic of Austria in 1981 to commemorate the tenth International Mathematics Congress in Innsbruck (see figure 4-8).

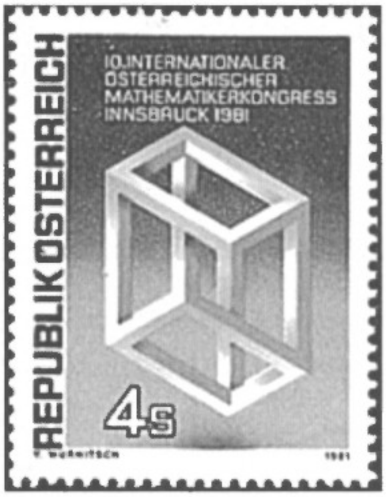

Figure 4-8

Just as there are optical illusions throughout our geometric world, so too there are "proofs" that can be fallacious, not by their errors in reasoning, but rather in their assumptions of geometric appearances.

A very popular "fallacy proof" is that every triangle can be proved to be isosceles.[2] The error in the proof is one that might have baffled Euclid, who did not properly define the concept of betweenness.[3]

Do not let these unusual optical illusions and tricks in geometry disillusion you. There is much beauty in geometry that is true. By this we are not referring to visually beautiful things, rather we are referring to amazing relationships that you would not think possible. We will present some here to rekindle your curiosity and to enhance your perception of geometry as more than just something that is optically interesting or just a series of proofs.

LOOKING AT GEOMETRY CRITICALLY

Consider the two squares in figure 4-9, where the smaller one has a side of length 3 and the larger one has a side of length 4, and where the vertex (or corner) M of square $MNPQ$ is the center of square $ABCD$.

---

[2] A triangle with two equal sides is called *isosceles*.

[3] For a complete discussion of this, see A. S. Posamentier, *Advanced Euclidean Geometry* (Hoboken, NJ: John Wiley, 2002), pp. 13–15.

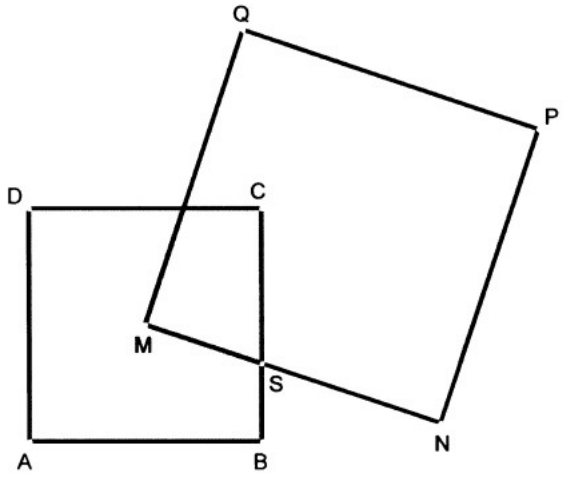

Figure 4-9

The question we raise is what is the area of the overlapping portion of the two squares? Does the exact relative position of the two squares matter? Since the exact placement of the two squares was not given—just the vertex *M*—suppose we rotate the larger square to the position shown in figure 4-10 or to that shown in figure 4-11. In either case, the "restrictions" mentioned in the original situation are still upheld, namely, that the vertex of one square is at the center of the other. From either of these new positions (figure 4-10 or 4-11), we can conclude that the overlapping portion is one-fourth of the area of the smaller square *ABCD*.

Yet this raises the question about whether this can assure us that this is not just an "extreme" case but also one that holds between the extreme or special positions.

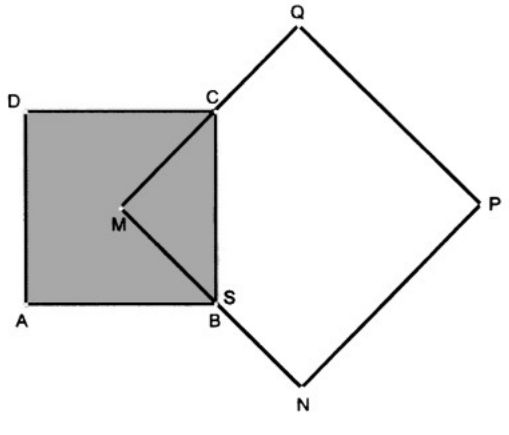

Figure 4-10

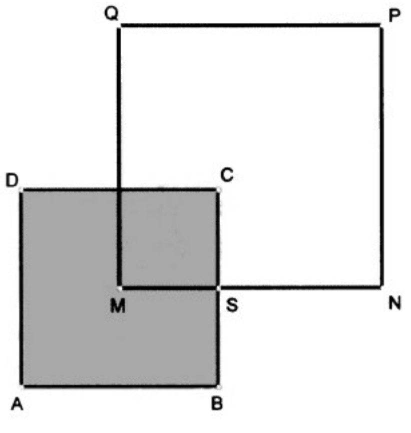

Figure 4-11

Suppose the larger square, *MNPQ*, is placed more "randomly," as shown in figure 4-12 and as it was positioned in our original question (figure 4-9). We can show (and, of course, prove) that the four regions into which square *ABCD* is partitioned all have the same area— since they can be shown to be congruent. This is a nice case of using extremes to guide us to a "conjecture" and we can then see it to be true for the random case shown in figure 4-12.

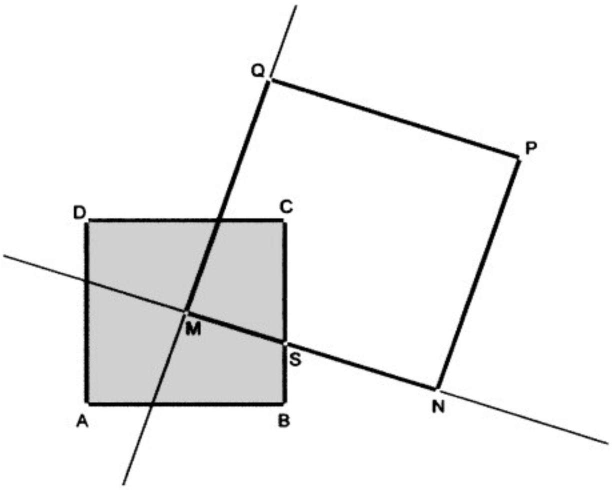

Figure 4-12

SOME TRIANGLE PROPERTIES—WITH MEDIANS

To impress you with the many wonderful, and often mind-boggling, relationships that exist in Euclidean plane geometry, we shall take you on an unfolding journey through various geometric shapes. When we refer to a triangle as being isosceles, we mean than it has two sides of equal length. A triangle with all sides of equal length is called *equilateral*. When we refer to a *scalene* triangle, it is one where all three sides are of different lengths. One way of finding some truly remarkable relationships derives from noticing that these unexpected relationships can occur in *any* triangle regardless of its shape!

We often take for granted that certain lines of a triangle behave as we would expect. For example, we know that the medians of a triangle (i.e., the lines joining a vertex with the midpoint of the opposite side) are concurrent—that is, they all go through the same point. In figure 4-13, *AF*, *BD*, and *CE* are the medians of triangle ΔABC, and *G* is their point of intersection (or the point of concurrency of the medians). This, in itself, is a fascinating reationship, since it is completely independent of the type of triangle. It holds true for *all* triangles!

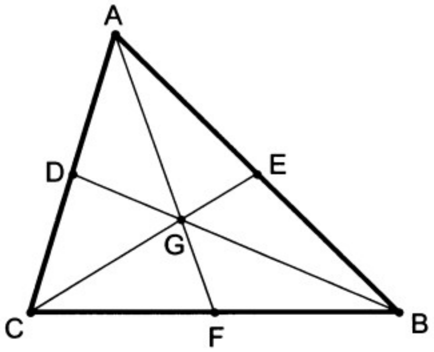

Figure 4-13

Furthermore, the medians partition the triangle into six smaller triangles, which happen to be equal in area. Again, this is true regardless of the shape of the triangle. That is, $Area\Delta ADG = Area\Delta AEG = Area\Delta BEG = Area\Delta BFG = Area\Delta CFG = Area\Delta CDG$.

We can also partition any triangle, say, ΔABC, into four congruent triangles ($\Delta ADE \cong \Delta CDF \cong \Delta FEB \cong \Delta EFD$) by using the midpoints of the three sides and joining them to form a medial triangle (see figure 4-14). Here, ΔEFD is called the *medial triangle* of ΔABC. Remember, these partitions hold true for *any* shape triangle.

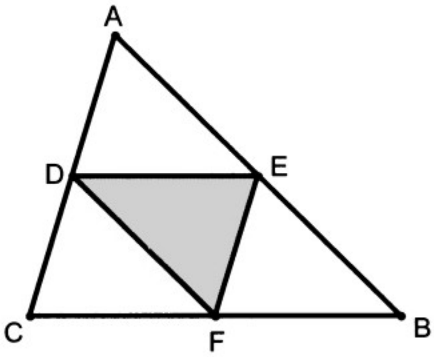

Figure 4-14

In other words, the area of triangle ΔABC is four times the area of triangle ΔEFD. That is, this relationship is a constant one that holds for all triangles.

In addition, point G (the point of intersection of the medians) is the center of gravity of the triangle, and is called the *centroid*. That is, if you want to balance a cardboard triangle on the tip of a pencil, you would need to find the centroid to place the point appropriately. We shall revisit the center of gravity a bit later. In the meantime, we can also appreciate another fascinating fact about the centroid of a triangle; namely, the centroid trisects each of the medians. In other words, in figure 4-15: $GD = \frac{1}{3}BD$, $GE = \frac{1}{3}CE$, and $GF = \frac{1}{3}AF$.

!$ (' $, ! 0

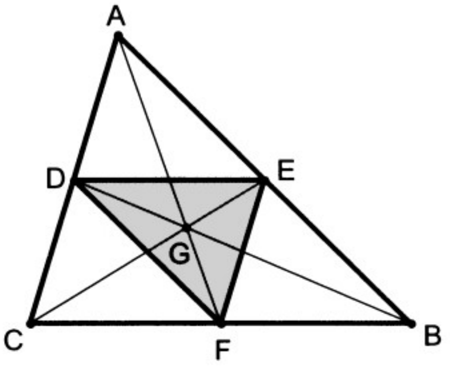

Figure 4-15

SOME QUADRILATERAL PROPERTIES

We shall now shift to inspect the quadrilateral and the midpoints of its sides. If we consecutively connect the midpoints of any quadrilateral with line segments, we always get a parallelogram (figure 4-16), which is called a *Varignon parallelogram.*[4] This is truly amazing, since it will hold true for any shape of a quadrilateral. Moreover, the Varignon parallelogram has half the area of the original quadrilateral, and its perimeter is equal to the sum of the lengths of the diagonals of the original quadrilateral. That is, in figure 4-16: the area of parallelogram *EFGH* is one-half of the area of quadrilateral *ABCD*, and the perimeter of parallelogram *EFGH* equals *AC* + *BD*.

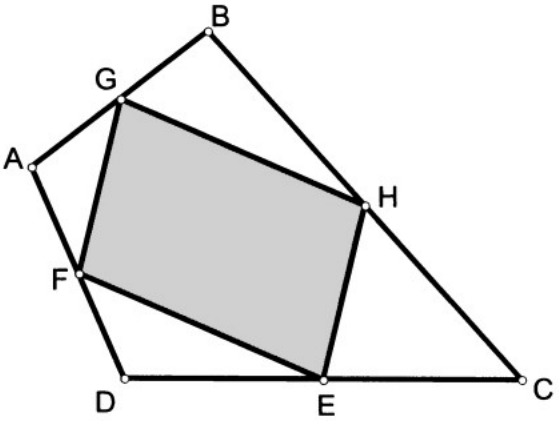

Figure 4-16

---

[4] Named for Pierre Varignon (1654–1722), a French mathematician. Although discovered in 1713, it was first published (posthumously) in 1731.

Just for curiosity's sake, let's see what happens when we reduce the length of one side of the quadrilateral, say, *AB*, to length zero. We get a triangle with one vertex having points *A*, *B*, and *G* coincide as in figure 4-17. Here, too, the resulting Varignon parallelogram still exists and is one-half the area of the original triangle (or quadrilateral with a zero-length side). You might want to relate this to our result in figure 4-14, where we found that the triangle *ABC* was partitioned into four triangles of equal area, leaving us with a similar result when we consider the parallelogram *ADFE* in figure 4-14, which is analogous to parallelogram *AFEH* in figure 4-17. The perimeter relationship (perimeter of parallelogram *EFGH* equals *AC* + *BD*) also holds true for the situation shown in figure 4-17. It is not only nice to have this consistency, it is essential!

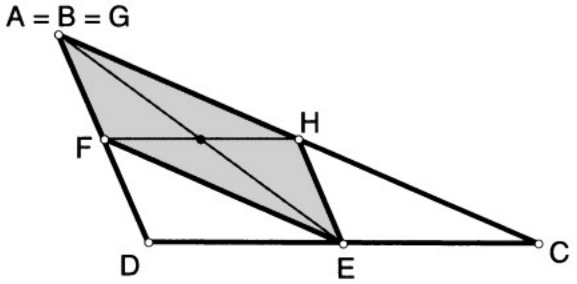

Figure 4-17

Of course, some quadrilaterals—when the midpoints of their sides are connected—will yield a special parallelogram, such as a rectangle, a square, or a rhombus (a parallelogram with all sides equal).

For example, suppose we consider a quadrilateral whose diagonals are perpendicular (see figure 4-18). When the side-midpoints of this quadrilateral are connected, a rectangle, *EFGH,* is produced.

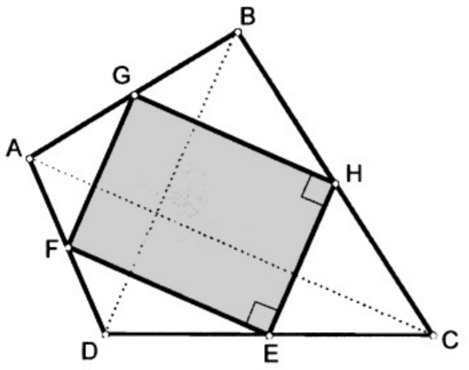

Figure 4-18

If the quadrilateral has perpendicular diagonals that are also the same length, then the quadrilateral produced by joining the side-midpoints is a special kind of rectangle that we know as a square! (See figure 4-19.)

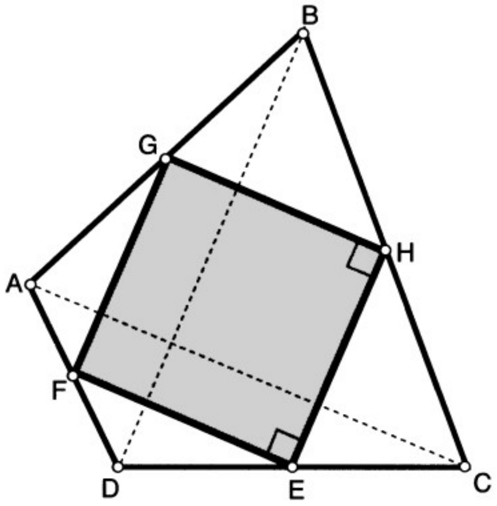

Figure 4-19

If the original quadrilateral has diagonals of equal length that are not perpendicular, then the quadrilateral formed by joining the side-midpoints will be a rhombus, as in figure 4-20.

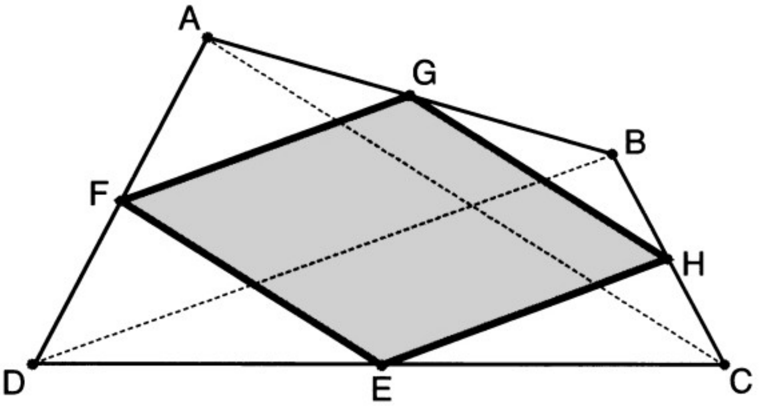

Figure 4-20

Naturally, we can continue this process of connecting consecutively the midpoints of the resulting parallelograms to get the series of parallelograms shown in figure 4-21.

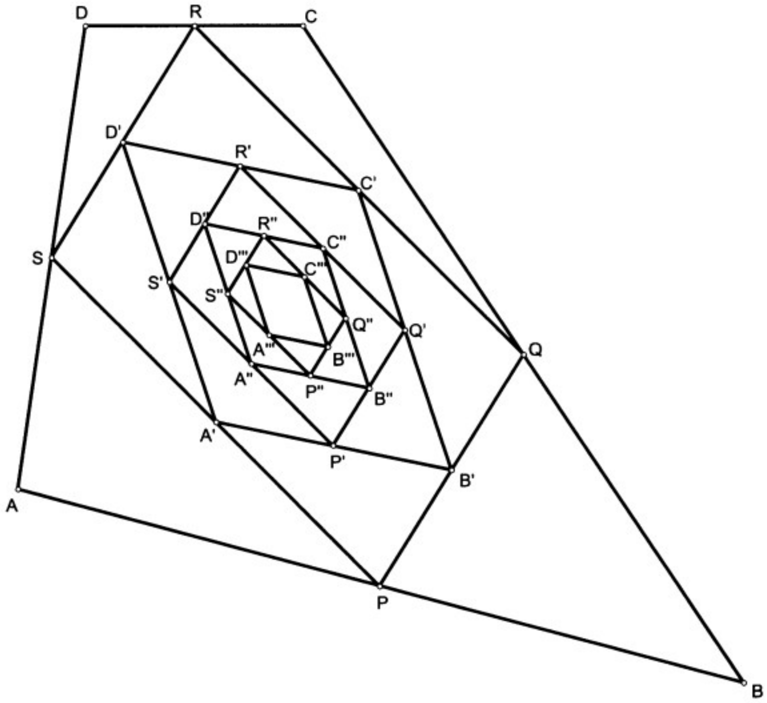

Figure 4-21

One of the prettier results from this investigation occurs when we can generate many special quadrilaterals from a randomly drawn quadrilateral. In the randomly drawn quadrilateral *ABCD* shown in figure 4-22 we connected the midpoints to form a parallelogram (*MNOP*). We then found the intersections of the bisectors of the angles of parallelogram *MNOP* to create rectangle *EFGH*. By bisecting each of the angles of this rectangle, we created the square *JKLR*. Therefore, from the randomly drawn quadrilateral *ABCD* we have created three special quadrilaterals: a parallelogram, a rectangle, and a square.

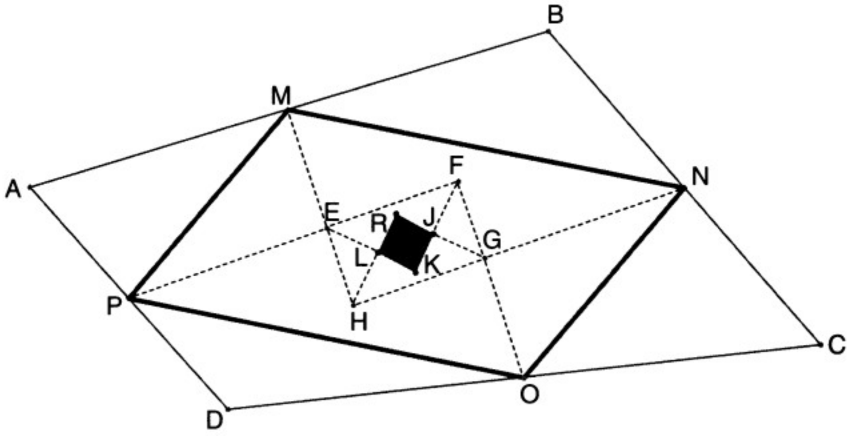

Figure 4-22

This is another example of the beauty of the geometric relationships that too often remain hidden. Their exposure is truly enlightening! You might want to try other ways of creating a square from a randomly drawn quadrilateral.

CENTER OF GRAVITY

Earlier (page 153), we located the center of gravity of a triangle (called the centroid) by finding the point of intersection of its medians. Some countries have run contests to find the "center" of their country. As you can see from figure 4-23, the contest submissions in Austria took on many creative methods. Yet the one that won the contest was for the town of Bad Aussee, where residents balanced a wooden map of the country on a point that was their town's location. Figure 4-23 is a translation of a plaque that designates the center of Austria in the town of Bad Aussee—as shown in the adjoining map (figure 4-24).

BAD AUSSEE

The geographic center in Austria
Latitude 47° 36' – Longitude 13° 47' east of border
(Dipl.-Ing. Otto Kloiber, Vienna)

MITTELPUNKT ÖSTERREICHS
BAD AUSSEE
1951

Why can Bad Aussee name itself the "Center of Austria?" Here is a brief review: In 1949 the magazine *Die Große Österreich Illustrierte* had a competition with the challenge of finding the center of Austria. Each Austrian had been invited to take part. The "winner" received a free vacation, which was a luxury then. The participants tried with different methods to find the navel of Austria. For example, the following:

THE DIAGONAL-SOLUTION:
Find the four points on the border, draw a rectangular of latitude and longitude along those points. The intersection of the diagonals would be the center.

THE CORNER-METHOD:
Find the point which is, in total length, the furthest away from the borders of Austria. The geometric solution would be the center of a circle whose circumference is tangent to the edges of other countries bordering Austria, and which reaches furthest into Austrian territory.

THE FOCAL POINT-THEORY:
Glue the map of Austria onto cardboard and cut out the map exactly along the borders. Put this Austria on the tip of a needle and move it around till the map balances.

The people of Bad Aussee started measuring and balancing and wrote to the magazine: "We are the ones!" The magazine confirmed (after a survey by the University of Vienna): "Bad Aussee is the center of Austria." The official ceremony with the actress Maria Andergast was on September 24, 1949.

In 1989 the local chamber of commerce of Bad Aussee used the 40th anniversary of the center-designation to celebrate with various festivities that the town is now the heart of Austria. As a permanent monument to remind everyone, that Bad Aussee is the center of Austria, a "center-stone" has been dedicated in the spa-gardens of the town.

The bronze plaque on the center-stone, a so-called "Omphalos" (Greek for "navel"), symbolizes the center. The inscription is in Gothic print and originated from the municipal seal of Aussee from 1505 and reads "Ause."

Furthermore, the local chamber of commerce strove to protect this title and applied to the Austrian patent office, which rendered this official designation in September 1991.

Figure 4-23

data 18.04.2020

$1! + 2! + 3! + 4! + 5! = 153$

D2) Pag. 147 (146)

Geometry Surprises

Pag. 152 (151)

Scalene triangle →

all three sides are of different lengths.

The medians partition the triangle into six smaller triangles, which happen to be equal in area.

We can also partition any triangle into four congruent triangles by using the midpoints of the three sides and joining them to form a medial triangle.

Point G (the point of intersection of the medians) is the center of gravity of the triangle, and is called the centroid. ✡

↳ Fascinating fact about the centroid of a triangle;

↳ IMPORTANT: ✡

data 18.04.2020

THE CENTROID TRISECTS EACH OF THE MEDIANS.

Figure 4.15

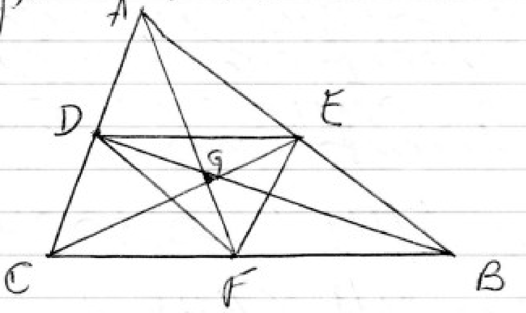

$GD = \frac{1}{3} BD$; $GE = \frac{1}{3} CE$, and $GF = \frac{1}{3} AF$

Important: ✡

estudar aplicação para trissecção do ângulo.

D3) Some Quadrilateral Properties

Pag. 154 (153)

Varignon parallelogram

If we consecutively connect the midpoint of any quadrilateral with line segments, we always get a parallelogram.

IMPORTANT: ✡

19.04.2020

off: Dia de Santo Expedito!!!

IMPORTANT:

Esta propriedade é um dos instrumentos que eu deverei usar para a checagem dos cálculos.

Use-o conjuntamente com os demais descritos no

item 144 (20200417),

subitem B4 (B4) e

The Trisectors by Dudley

subitem © (Feuerbach's theorem)

↳ que faz referência ao

item 138 (20200416).

Vou explicar melhor:

Com Varignon Parallelogram poderemos obter um paralelogramma, ou seja, 2 pares de retas paralelas que foram obtidas de um quadrilátero (que nem sempre) é regular.

Desta forma nós conseguiremos obter padrões para realizar deduções de cálculos de áreas e assim,

19.04.20

inferimos a área total deste quadrilátero irregular, cujos (~~arestas~~) lados não são paralelas e os ângulos formados entre elas (~~arestas~~) (lados) são desconhecidos.

Obs: Aresta é a reta que se origina a partir da intersecção de 2 planos que formam um ângulo.
(3D do quadrilátero irregular)

Varignon parallelogram aplica-se para todos tipos de quadrilátero: regular e irregular.

(quadrado)
↳ é um paralelograma.
↳ é um polígono que reúne os características do losango e do retângulo ao mesmo tempo.
↳ os lados são congruentes (mesma medida).
↳ um quadrado também pode ser chamado de losango.

Obs: 1. Todo paralelogramo é quadrado, retângulo ou losango;
2. Todo retângulo ou losango é também um quadrado.

IMPORTANT:

Varignon parallelogram

data 19.04.2020

VARIGNON PARALLELOGRAM

has half the area of the original quadrilateral. ✡

(mais um instrumental para calcular a área de um determinado quadrilátero).

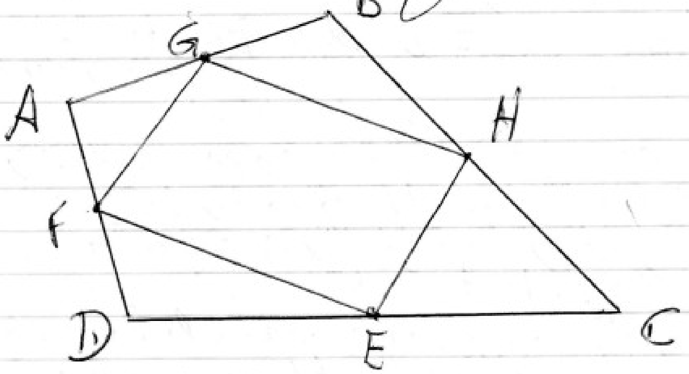

The area of parallelogram EFGH is one-half of a area of quadrilateral ABCD, and the perimeter of paralllogram EFGH equals AC + BD.

IMPORTANT:

Outro instrumento para checagem dos meus cálculos.

Quantas mais redundâncias de checagem é melhor, diminui o erro.

data 19 04 2020

Figure 4.17 (Important)

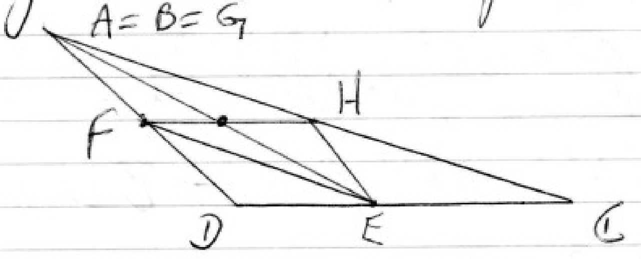

If we reduce the length of one side of the quadrilateral, AB to length zero. We get a triangle with one vertex having points A, B, and G coincide as in figure 4.17.

Here, too, the resulting Varignon Parallelogram still exists and it (is) one-half the area of the original triangle (or quadrilateral with a zero-length side).

Relate with Figure 4.14: where we found that the triangle ABC was partitioned into four triangles of equal area, leaving us with a similar result when we consider the parallelogram ADFE in figure 4.14 which is analogous to parallelogram AFEH in Figure 4.17.

→

data 19·04·2020

IMPORTANT: ✡

The perimeter relationship (perimeter of parallelogram EFGH equal AC + BD)

Observation:

AC and BD → sides of triangle

pois AB = zero.

Vejamos.

Figure 4.16 $\boxed{AC + BD}$ (não são lados)

Figure 4.17 AC + BD (são lados)

estudar a angulação!!! ✡

also holds true for the situation shown in figure 4.17.

Pág. 155 (154)

Figure 4.18

A quadrilateral whose diagonals are perpendicular. When the side midpoints of this quadrilateral are connected, a rectangulo EFGH is produced.

data 19 04 2020
X

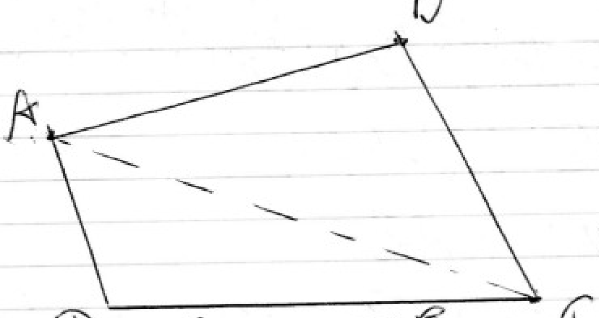

Wrong.
(no parallel diagonal)

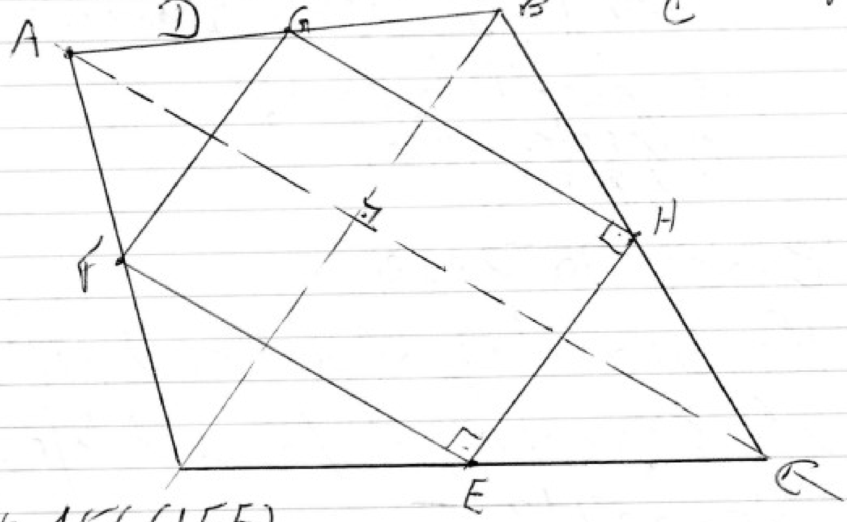

Pag. 156 (155)

✡

Figure 4.19

If the quadrilateral has perpendicular diagonals that are the same length, then the quadrilateral produced by joining the side-midpoints is a special kind of (rectangulo) rectangle that we know as a square.

Figure 4.20

If the original quadrilateral has diagonals of equal length are not perpendicular, the the quadrilateral

data 19.04.2020

formed by joining the side-midpoints *
will be a rhombus.

Pag. 157(156)

Figure 4.22

Exercise 3 special quadrilaterals:

a parallelogram,
a rectangle,
a square.

E) Books_70_2020/

The Geometry of Art and Life/

The Geometry of Art and Life.pdf

146) 20200419

A) Books_70_2020/
Microelectronics Circuit Analysis and Design, 4th/
Microelectronics Circuit Analysis Design.pdf

Pag. 77(100)

The ripple voltage is then (2.9)

$$V_r = \frac{V_M}{2fRC}$$

)*+,)))-) ' 8 < 1 * ! 5 0!
(%

rôle in crystallography; elementary crystals are close-packed and produce simple or complex space-lattices according to the different elements and different axes of symmetry. The basic configurations are the three principal systems of partition of space: cubic, hexagonal (derived from close-packed hexagonal or tetrahedral prisms) and cuboctahedric, which partakes of the two previous ones.[1] There are in all seven fundamental groups of symmetry, representing 230 possible types of point-lattices.[2] The fundamental law in crystallography is the "law of rational coefficients": the coefficients expressing (in three-dimensional analytical geometry) the relations between the different plane facets of a crystal and the directions of the three principal axes of symmetry are small integer numbers.

The different configurations of crystallized matter correspond to states of stable or relatively stable equilibrium determined by rigorous causality, the chemical reactions of the different elements being themselves explained by a tendency of the electrons to combine with each other according to more stable arrangements. The most general principle governing the states of equilibrium of physico-chemical systems is the "Criterium of Dirichlet": in order that the equilibrium of a closed system should be stable it is sufficient that its potential energy should be (or should pass through) a minimum. A corollary is Curie's Law: a body tends to take the shape presenting the minimum surface energy made possible by the directing forces. In many cases (for crystalline systems as well as for the more complex ones of organic chemistry) the minimum of superficial potential energy corresponds to the solution which for a given volume produces the smallest surface agreeing with the linking forces. (In space the sphere gives

---

[1] We have seen that common salt (Sodium Chloride) crystallizes according to the cuboctahedric lattice; so does carbon in diamond, and so do in general gold, silver, copper and aluminum. Iron at normal temperature crystallizes in the cubic system (including the centres of the cubes).

[2] A crystallized body shows a number of interlocked point-lattices corresponding to the number of atoms in each molecule.

* 1 > 5 3 !1 ! B 0 0 1
! 0 3 31 !

& '.
)*+,)))-) '.
* 1 9 0!
(//

2.1.3 Filters, Ripple Voltage, and Diode Current

If a capacitor is added in parallel with the load resistor of a half-wave rectifier to form a simple filter circuit (Figure 2.8(a)), we can begin to transform the half-wave sinusoidal output into a dc voltage. Figure 2.8(b) shows the positive half of the output sine wave, and the beginning portion of the voltage across the capacitor, assuming the capacitor is initially uncharged. If we assume that the diode forward resistance is $r_f = 0$, which means that the $r_f C$ time constant is zero, the voltage across the capacitor follows this initial portion of the signal voltage. When the signal voltage reaches its peak and begins to decrease, the voltage across the capacitor also starts to decrease, which means the capacitor starts to discharge. The only discharge current path is through the resistor. If the RC time constant is large, the voltage across the capacitor discharges exponentially with time (Figure 2.8(c)). During this time period, the diode is cut off.

A more detailed analysis of the circuit response when the input voltage is near its peak value indicates a subtle difference between actual circuit operation and the qualitative description. If we assume that the diode turns off immediately when the input voltage starts to decrease from its peak value, then the output voltage will decrease exponentially with time, as previously indicated. An exaggerated sketch of these two voltages is shown in Figure 2.8(d). The output voltage decreases at a faster rate than the input voltage, which means that at time t_1, the difference between v_I and

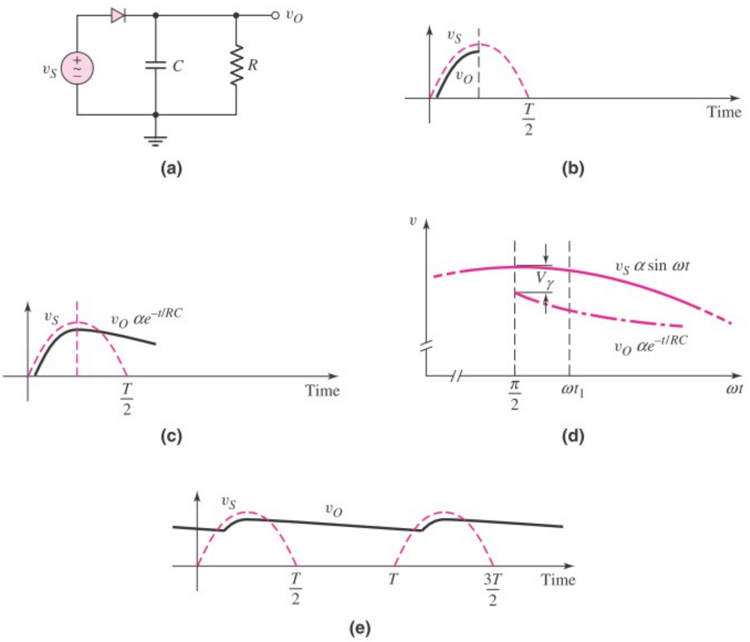

Figure 2.8 Simple filter circuit: (a) half-wave rectifier with an RC filter, (b) positive input voltage and initial portion of output voltage, (c) output voltage resulting from capacitor discharge, (d) expanded view of input and output voltages assuming capacitor discharge begins at $\omega t = \pi/2$, and (e) steady-state input and output voltages

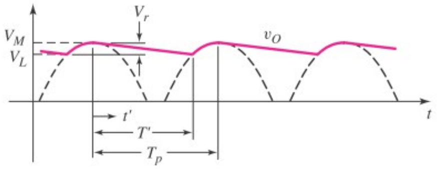

Figure 2.9 Output voltage of a full-wave rectifier with an *RC* filter showing the ripple voltage

v_O, that is, the voltage across the diode, is greater than V_γ. However, this condition cannot exist. Therefore, the diode does not turn off immediately. If the *RC* time constant is large, there is only a small difference between the time of the peak input voltage and the time the diode turns off. In this situation, a computer analysis may provide more accurate results than an approximate hand analysis.

During the next positive cycle of the input voltage, there is a point at which the input voltage is greater than the capacitor voltage, and the diode turns back on. The diode remains on until the input reaches its peak value and the capacitor voltage is completely recharged.

Since the capacitor filters out a large portion of the sinusoidal signal, it is called a **filter capacitor.** The steady-state output voltage of the *RC* filter is shown in Figure 2.8(e).

The ripple effect in the output from a full-wave filtered rectifier circuit can be seen in the output waveform in Figure 2.9. The capacitor charges to its peak voltage value when the input signal is at its peak value. As the input decreases, the diode becomes reverse biased and the capacitor discharges through the output resistance *R*. Determining the ripple voltage is necessary for the design of a circuit with an acceptable amount of ripple.

To a good approximation, the output voltage, that is, the voltage across the capacitor or the *RC* circuit, can be written as

$$v_O(t) = V_M e^{-t'/\tau} = V_M e^{-t'/RC} \tag{2.3}$$

where t' is the time after the output has reached its peak value, and *RC* is the time constant of the circuit.

The smallest output voltage is

$$V_L = V_M e^{-T'/RC} \tag{2.4}$$

where T' is the discharge time, as indicated in Figure 2.9.

The **ripple voltage** V_r is defined as the difference between V_M and V_L, and is determined by

$$V_r = V_M - V_L = V_M(1 - e^{-T'/RC}) \tag{2.5}$$

Normally, we will want the discharge time T' to be small compared to the time constant, or $T' \ll RC$. Expanding the exponential in a series and keeping only the linear terms of that expansion, we have the approximation[2]

$$e^{-T'/RC} \cong 1 - \frac{T'}{RC} \tag{2.6}$$

---

[2] We can show that the difference between the exponential function and the linear approximation given by Equation (2.6) is less than 0.5 percent for $RC = 10T'$. We need a relatively large *RC* time constant for this application.

The ripple voltage can now be written as

$$V_r \cong V_M\left(\frac{T'}{RC}\right) \qquad \textbf{(2.7)}$$

Since the discharge time T' depends on the RC time constant, Equation (2.7) is difficult to solve. However, if the ripple effect is small, then as an approximation, we can let $T' = T_p$, so that

$$V_r \cong V_M\left(\frac{T_p}{RC}\right) \qquad \textbf{(2.8)}$$

where T_p is the time between peak values of the output voltage. For a full-wave rectifier, T_p is one-half the signal period. Therefore, we can relate T_p to the signal frequency,

$$f = \frac{1}{2T_p}$$

The ripple voltage is then

$$V_r = \frac{V_M}{2fRC} \qquad \textbf{(2.9)}$$

For a half-wave rectifier, the time T_p corresponds to a full period (not a half period) of the signal, so the factor 2 does not appear in Equation (2.9). The factor of 2 shows that the full-wave rectifier has half the ripple voltage of the half-wave rectifier.

Equation (2.9) can be used to determine the capacitor value required for a particular ripple voltage.

EXAMPLE 2.3

Objective: Determine the capacitance required to yield a particular ripple voltage.

Consider a full-wave rectifier circuit with a 60 Hz input signal and a peak output voltage of $V_M = 10$ V. Assume the output load resistance is $R = 10\ \text{k}\Omega$ and the ripple voltage is to be limited to $V_r = 0.2$ V.

Solution: From Equation (2.9), we can write

$$C = \frac{V_M}{2fRV_r} = \frac{10}{2(60)(10 \times 10^3)(0.2)} \Rightarrow 41.7\ \mu\text{F}$$

Comment: If the ripple voltage is to be limited to a smaller value, a larger filter capacitor must be used. Note that the size of the ripple voltage and the size of filter capacitor are related to the load resistance R.

EXERCISE PROBLEM

Ex 2.3: Assume the input signal to a rectifier circuit has a peak value of $V_M = 12$ V and is at a frequency of 60 Hz. Assume the output load resistance is $R = 2\ \text{k}\Omega$ and the ripple voltage is to be limited to $V_r = 0.4$ V. Determine the capacitance required to yield this specification for a (a) full-wave rectifier and (b) half-wave rectifier. (Ans. (a) $C = 125\ \mu$F, (b) $C = 250\ \mu$F).

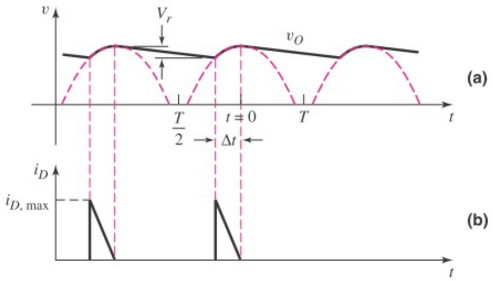

Figure 2.10 Output of a full-wave rectifier with an *RC* filter: (a) diode conduction time and (b) diode current

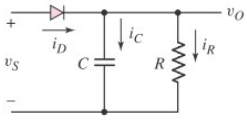

Figure 2.11 Equivalent circuit of a full-wave rectifier during capacitor charging cycle

The diode in a filtered rectifier circuit conducts for a brief interval Δt near the peak of the sinusoidal input signal. The diode current supplies the charge lost by the capacitor during the discharge time. Figure 2.10 shows the rectified output of a full-wave rectifier and the filtered output assuming ideal diodes ($V_\gamma = 0$) in the rectifier circuit. We will use this approximate model to estimate the diode current during the diode conduction time. Figure 2.11 shows the equivalent circuit of the full-wave rectifier during the charging time. We see that

$$i_D = i_C + i_R = C\frac{dv_O}{dt} + \frac{v_O}{R} \tag{2.10}$$

During the diode conduction time near $t = 0$ (Figure 2.10), we can write

$$v_O = V_M \cos \omega t \tag{2.11}$$

For small ripple voltages, the diode conduction time is small, so we can approximate the output voltage as

$$v_O = V_M \cos \omega t \cong V_M\left[1 - \frac{1}{2}(\omega t)^2\right] \tag{2.12}$$

The charging current through the capacitor is

$$i_C = C\frac{dv_O}{dt} = CV_M\left[-\frac{1}{2}(2)(\omega t)(\omega)\right] = -\omega CV_M\omega t \tag{2.13}$$

From Figure 2.10, the diode conduction occurs during the time $-\Delta t < t < 0$, so that the capacitor current is positive and is a linear function of time. We note that at $t = 0$, the capacitor current is $i_C = 0$. At $t = -\Delta t$, the capacitor charging current is at a peak value and is given by

$$i_{C,\text{peak}} = -\omega CV_M[\omega(-\Delta t)] = +\omega CV_M\omega\Delta t \tag{2.14}$$

The capacitor current during the diode charging time is approximately triangular and is shown in Figure 2.10(b).

From Equation (2.11), we can write that the voltage V_L is given by

$$V_L = V_M \cos[\omega(-\Delta t)] \cong V_M\left[1 - \frac{1}{2}(\omega\Delta t)^2\right] \tag{2.15}$$

Solving for $\omega\Delta t$, we find

$$\omega\Delta t = \sqrt{\frac{2V_r}{V_M}} \tag{2.16}$$

where $V_r = V_M - V_L$.

From Equation (2.9), we can write

$$fC = \frac{V_M}{2RV_r} \tag{2.17(a)}$$

or

$$2\pi fC = \omega C = \frac{\pi V_M}{RV_r} \tag{2.17(b)}$$

Substituting Equations (2.17(b)) and (2.16) into Equation (2.14), we have

$$i_{C,\text{peak}} = \left(\frac{\pi V_M}{RV_r}\right) V_M \left(\sqrt{\frac{2V_r}{V_M}}\right) \tag{2.18(a)}$$

or

$$i_{C,\text{peak}} = \pi \frac{V_M}{R}\sqrt{\frac{2V_M}{V_r}} \tag{2.18(b)}$$

Since the charging current through the capacitor is triangular, we have that the average capacitor current during the diode charging time is

$$i_{C,\text{avg}} = \frac{\pi}{2}\frac{V_M}{R}\sqrt{\frac{2V_M}{V_r}} \tag{2.19}$$

During the capacitor charging time, there is still a current through the load. This current is also being supplied through the diode. Neglecting the ripple voltage, the load current is approximately

$$i_L \cong \frac{V_M}{R} \tag{2.20}$$

Therefore, the peak diode current during the diode conduction time for a full-wave rectifier is approximately

$$i_{D,\text{peak}} \cong \frac{V_M}{R}\left(1 + \pi\sqrt{\frac{2V_M}{V_r}}\right) \tag{2.21}$$

and the average diode current during the diode conduction time is

$$i_{D,\text{avg}} \cong \frac{V_M}{R}\left(1 + \frac{\pi}{2}\sqrt{\frac{2V_M}{V_r}}\right) \tag{2.22}$$

The average diode current over the entire input signal period is

$$i_D(\text{avg}) = \frac{V_M}{R}\left(1 + \frac{\pi}{2}\sqrt{\frac{2V_M}{V}}\right)\frac{\Delta t}{T} \tag{2.23}$$

For the full-wave rectifier, we have $1/2T = f$, so

$$\Delta t = \frac{1}{\omega}\sqrt{\frac{2V_r}{V_M}} = \frac{1}{2\pi f}\sqrt{\frac{2V_r}{V_M}} \tag{2.24(a)}$$

Then

$$\frac{\Delta t}{T} = \frac{1}{2\pi f}\sqrt{\frac{2V_r}{V_M}}2f = \frac{1}{\pi}\sqrt{\frac{2V_r}{V_M}} \tag{2.24(b)}$$

Then the average current through the diode during the entire cycle for a full-wave rectifier is

$$i_D(\text{avg}) = \frac{1}{\pi}\sqrt{\frac{2V_r}{V_M}}\frac{V_M}{R}\left(1 + \frac{\pi}{2}\sqrt{\frac{2V_M}{V_r}}\right) \tag{2.25}$$

DESIGN EXAMPLE 2.4

Objective: Design a full-wave rectifier to meet particular specifications.

A full-wave rectifier is to be designed to produce a peak output voltage of 12 V, deliver 120 mA to the load, and produce an output with a ripple of not more than 5 percent. An input line voltage of 120 V (rms), 60 Hz is available.

Solution: A full-wave bridge rectifier will be used, because of the advantages previously discussed. The effective load resistance is

$$R = \frac{V_O}{I_L} = \frac{12}{0.12} = 100\,\Omega$$

Assuming a diode cut-in voltage of 0.7 V, the peak value of v_S is

$$v_S(\max) = v_O(\max) + 2V_\gamma = 12 + 2(0.7) = 13.4\text{ V}$$

For a sinusoidal signal, this produces an rms voltage value of

$$v_{S,\text{rms}} = \frac{13.4}{\sqrt{2}} = 9.48\text{ V}$$

The transformer turns ratio is then

$$\frac{N_1}{N_2} = \frac{120}{9.48} = 12.7$$

For a 5 percent ripple, the ripple voltage is

$$V_r = (0.05)V_M = (0.05)(12) = 0.6\text{ V}$$

The required filter capacitor is found to be

$$C = \frac{V_M}{2fRV_r} = \frac{12}{2(60)(100)(0.6)} \Rightarrow 1667\,\mu\text{F}$$

The peak diode current, from Equation (2.21), is

$$i_{D,\text{peak}} = \frac{12}{100}\left[1 + \pi\sqrt{\frac{2(12)}{0.6}}\right] = 2.50\text{ A}$$

and the average diode current over the entire signal period, from Equation (2.25), is

$$i_D(\text{avg}) = \frac{1}{\pi}\sqrt{\frac{2(0.6)}{12}}\left(\frac{12}{100}\right)\left(1 + \frac{\pi}{2}\sqrt{\frac{2(12)}{0.6}}\right) \Rightarrow 132\text{ mA}$$

Finally, the peak inverse voltage that each diode must sustain is

$$\text{PIV} = v_R(\max) = v_S(\max) - V_\gamma = 13.4 - 0.7 = 12.7\text{ V}$$

data 19.04.2020

formed by joining the side-midpoints *
will be (a) rhombus.

Pag. 157(156)

Figure 4.22

Exercise 3 special quadrilaterals:

a parallelogram,
a rectangle,
a square.

E) Books_70_2020/
The Geometry of Art and Life/
The Geometry of Art and Life.pdf

146) 20200419

A) Books_70_2020/
Microelectronics Circuit Analysis and Design, 4th/
Microelectronics Circuit Analysis Design.pdf

Pag. 77(100)

The ripple voltage is then (2.9)

$$V_r = \frac{V_M}{2fRC}$$

19.04.2020

Pág. 79 (102)

Since the charging through the capacitor is triangular, we have that the average capacitor current during the diode charging time is

$$i_{C,avg} = \frac{\pi}{2} \frac{V_M}{R} \sqrt{\frac{2V_M}{V_r}} \qquad (2.19)$$

↑ Obs.: Triangular / $\overset{\sim}{\pi}$

— Checar correlações!!!

B) Books_70_2020/
Ideas of quantum chemistry
Ideas of quantum chemistry by Lucjan Piela....pdf

Pág. 282 (317)

7.1.4 Application of the Ritz Method

We will use the Ritz variational method (see Chapter 5, p. 202) to solve the Schrödinger equation.

.... spherical harmonics

$$\{Y_l^m(\theta,\phi)\} \text{ and } \{Y_{l'}^{m'}(\theta,\phi)\}$$

$\neq$ $\neq$ $\underbrace{\Phi, \phi}_{\neq}$

Ⓗ Θ

data 19.04.2020

therefore, we may use as
variational function: 11

11. the products $Y_{l}^{m}(\Theta,\phi)$
$Y_{l'}^{m'}(\theta,\phi)$

may be used to produce linear combinations that are automatically the

EIGENFUNCTIONS of
$\hat{J}^2$ and $\hat{J}_z$, and have the proper parity (Chapter 2)

this may be achieved by using the
CLEBSCH-GORDAN coefficients

Source: D.M. Brink, G.R. Satchler,
"Angular Momentum" 1975.

(C) Books_to_2020/
Angular Momentum Brink/
Angular_momentum_1968.pdf
Pág. 05(13)

The geometrical symmetry
operations of translation and

)*+,)))-) '. * ? ? #.% !C6
)*+,)))-) '. * ? ? #.% 0!
($

Returning to symmetry transformations it is clear that the successive application of two symmetry transformations produces a third symmetry transformation, and it can be shown that the set of all symmetry transformations with this law of combination forms a group, the symmetry group of the system. For a quantum system the symmetry transformations are represented by unitary operators, and the successive application of two symmetry transformations S and T is represented by the operator product TS. From section 1.2 we see the set of symmetry transformations of a quantum system is contained in the set of all unitary operators which commute with the Hamiltonian operator of the system.

1.4. Geometrical Symmetries

An important sub-group of symmetry transformations of many dynamical systems has a geometrical origin. The space of classical physics is described by a Euclidean geometry, implying that all points of space and all directions are equivalent and only statements relating to relative position and relative orientation have a meaning independent of the coordinate system. If the Euclidean character of space is reflected in physical laws no point in space nor any directions should be distinguishable absolutely by the performance of experiments. The operations of translation and rotation applied to a physical system should therefore belong to the symmetry group of the system. Thus the geometrical character of space determines symmetries of physical laws, or, perhaps more correctly, the physical symmetries determine the geometry of space. Lorentz transformations and space reflection come also into the category of symmetries with a geometrical origin.

In practical problems often a distinction is made between a physical system and its surroundings and the effect of the surroundings is approximated by boundary conditions or by a set of known fields or forces applied from the outside. This approximate treatment may destroy the geometrical symmetries partially or completely. For example, the equations of motion of an atom in an external magnetic field are not invariant for arbitrary rotations unless the external field is also rotated.

The geometrical symmetry operations of translation and rotation can be looked at in two ways, called active and passive. In the active sense introduced above they transform one state of a system into another. A rotation actually rotates the system from one position to another. The passive approach interprets the symmetry operations as coordinate transformations, and the symmetry of the dynamical laws is expressed by stating that the equations of motion are invariant under translation or rotation of axes. Any translation or rotation of a system may, however, be induced by a coordinate transformation and the two views are equivalent. This fact can be seen most easily by considering a simple example, the rotation of a one-particle quantum system through an angle α about the z-axis. (See Fig. 1.) Rotating the system transforms its wave function to a new one so that the value of the new wave function at the point $(\phi+\alpha)$ is the same as that of the old one at the point ϕ where ϕ is the azimuthal angle. Let the original wave function be $\psi(r,\theta,\phi)$ and the rotated wave function

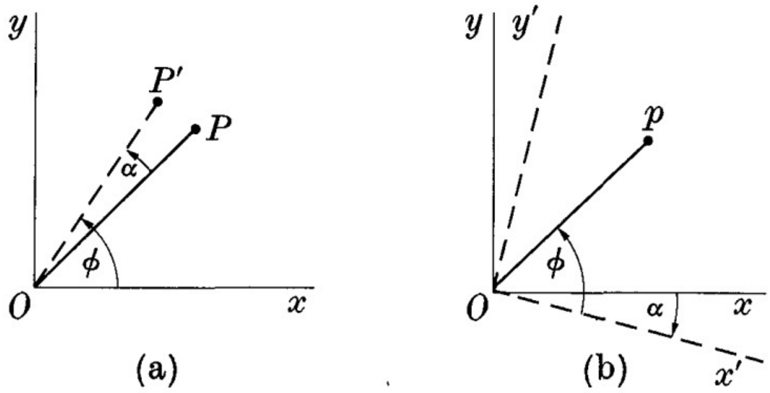

FIG. 1. (a) Represents a rotation of the system. A vector OP is carried into OP' by the rotation.
(b) Represents the equivalent rotation of axes. The vector OP remains fixed, but its coordinates are changed. The position of OP relative to the new axes in (b) is the same as that of OP' relative to the axes in (a).

$\psi' = D_\alpha \psi$ then this argument shows that

$$\psi'(r,\theta,\,\phi+\alpha) = \psi(r,\theta,\phi),$$

or

$$\psi'(r,\theta,\phi) \qquad = \psi(r,\theta,\,\phi-\alpha).$$

Thus the rotated wave function can be obtained from the original wave function by making the coordinate transformation $(r,\theta,\phi) \rightarrow (r,\theta,\phi-\alpha)$, i.e. by a rotation of the coordinate axes through an angle $-\alpha$ about the z-axis. Both these views of the rotation are important and will be used with the convention that a positive (counter clockwise) rotation refers to a positive rotation of the system, hence a negative rotation of axes.‡

1.5. Conservation Laws

There is a relation between the geometrical symmetries of a physical system and the momentum and angular momentum conservation laws, having its origin in the fact that the linear and angular momenta are respectively the generators of translations and rotations of the system. As an example consider the canonical transformation generated by L_z, the z-component of the angular momentum of a classical particle.

If F is any function of the coordinates q_i and the momenta p_i of a dynamical system and α is small then the transformation§

$$\begin{aligned} p_i' &= p_i + \alpha\{p_i, F\}, \\ q_i' &= q_i + \alpha\{q_i, F\}, \end{aligned} \tag{1.1}$$

is an infinitesimal canonical transformation in the sense that Hamilton's equations of motion retain the same form when written in terms of p_i' and q_i'. F is called the generating function of the transformation and it can be shown that if $G(p_i,\, q_i)$

---

‡ This convention agrees with Rose [54], Messiah [46]. Other authors use the opposite convention, a positive rotation referring to a positive rotation of the axes; Wigner [78], Edmonds [22]. Cf. section 2.4 and Appendix V.

§ The Poisson bracket $\{G, F\}$ of two functions of p_i and q_i is defined as

$$\{G, F\} = \sum_k \left(\frac{\partial G}{\partial q_k}\frac{\partial F}{\partial p_k} - \frac{\partial G}{\partial p_k}\frac{\partial F}{\partial q_k} \right).$$

data 19.04.2020

therefore, we may use as variational function: 11

11. the products $Y_{\ell}^{m}(\Theta,\phi)$

$$Y_{\ell'}^{m'}(\theta,\phi)$$

may be used to produce linear combinations that are automatically the

EIGENFUNCTIONS of

$\hat{J}^2$ and $\hat{J}_z$, and have the proper parity (Chapter 2)

this may be achieved by using the CLEBSCH-GORDAN coefficients

Source: D.M. Brink, G.R. Satchler, "Angular Momentum" 1975.

①) Books_70_2020/
Angular Momentum Brink/
Angular_momentum_1968.pdf
Pág. 05(13)

The geometrical symmetry operations of translation and

19.04.2020

rotation can be looked at in two ways, called active and passive.

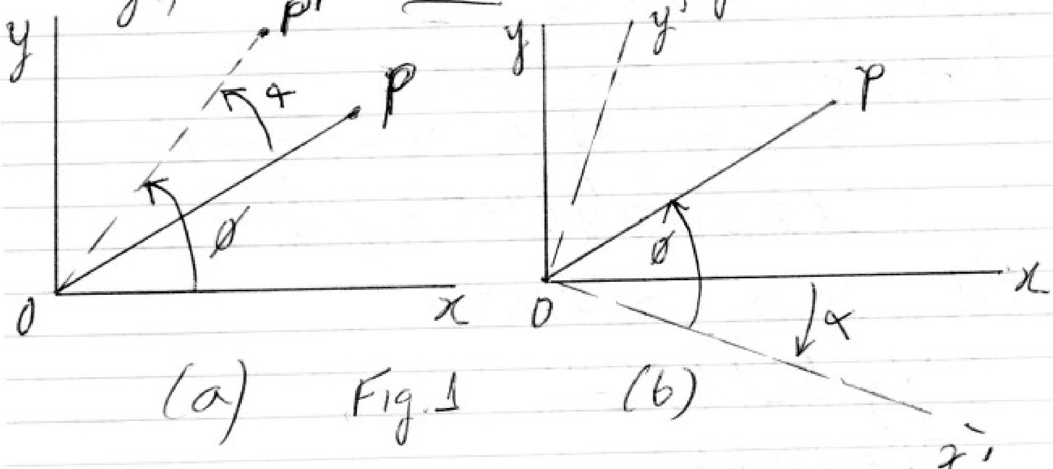

(a) Fig. 1 (b)

a) Represents a rotation of the system. A vector OP is carried into OP' by the rotation.

b) Represents the equivalent rotation of axes.

The vector OP remains fixed, but its coordinates are changed.
The position of OP relative to the new axes in (b) is the same as that OP' relative to axes in (a).

A rotation actually rotates the system from one position to another.
The passive approach interprets the symmetry operations as coordinates transformations and the symmetry of the dynamicall laws is expressed by

19 04 2020

stating that the equations of motion are invariant under translation or rotation of axes. –

147) A|Books_70_20201
Lecture_Notes_on_Complex_
Analysis/
1860946437.pdf

obs: Use todo o livro.

Pág. 111(124)

Chapter 7

Complex Integration

Pág. 113(126)

7.2. The Length of a Contour.

Polygon:

Condition:

$\gamma : [a,b] \to \mathbb{C}$ is given smooth path.

We can consider an inscribed polygon as an approximation to the path and its length will be an approximation to the length of the path.

polygon ⇒ vertices $\{z_0, z_1, \dots, z_n\}$.

& '/
)*+,)))-) '/ %. #'.'$/ 0!
(' 0 /

Chapter 7

Complex Integration

7.1 Paths and Contours

In this chapter, we consider integration along a contour in the complex plane. Recall that a path in $\mathbb{C}$ is a continuous function $\gamma : [a,b] \to \mathbb{C}$, for some $a \le b$ in $\mathbb{R}$. The set of points $\{\, z : z = \gamma(t),\ a \le t \le b \,\}$ is the trace of γ which we denote by $\operatorname{tr}\gamma$.

Example 7.1 Let $\gamma_1(t) = e^{2\pi i t}$ for $0 \le t \le 1$. Then we find that

$$\operatorname{tr}\gamma_1 = \{\, z : |z| = 1 \,\} = \text{ unit circle.}$$

Now suppose that $\gamma_2(t) = e^{-4\pi i t}$ for $0 \le t \le 1$. We see that again

$$\operatorname{tr}\gamma_2 = \{\, z : |z| = 1 \,\} = \text{ unit circle.}$$

The paths γ_1 and γ_2 have the same trace but are different paths: the path γ_1 goes round the circle anticlockwise once, whereas γ_2 goes round twice in a clockwise sense.

Remark 7.1 The convention is to take the anticlockwise sense as being positive. This is consistent with the convention that the positive direction is that with increasing polar angle.

Definition 7.1 The path $\gamma : [a,b] \to \mathbb{C}$ is said to be closed if $\gamma(a) = \gamma(b)$. The path γ is said to be simple if it does not cross itself, that is, $\gamma(s) \ne \gamma(t)$ whenever $s \ne t$ with s, t in (a,b). (The possibility that $\gamma(a) = \gamma(b)$ is allowed.)

Example 7.2 The path γ_1 in the example above is a simple closed path, whereas γ_2 is closed but not simple. The path $\gamma(t) = e^{-5\pi i t}$, $0 \le t \le 1$, has the unit circle as its trace, but γ is neither closed nor simple.

Definition 7.2 Let $\gamma : [a,b] \to \mathbb{C}$ be a path. The reverse path $\widetilde{\gamma}$ is given by $\widetilde{\gamma} : [a,b] \to \mathbb{C}$ with $\widetilde{\gamma}(t) = \gamma(a+b-t)$.

Evidently, $\widetilde{\gamma}$ is "γ in the opposite direction". Note that the parametric range for $\widetilde{\gamma}$ is the same as that for γ, namely $[a,b]$. It is also clear that $\operatorname{tr}\widetilde{\gamma} = \operatorname{tr}\gamma$.

Definition 7.3 The path $\gamma : [a,b] \to \mathbb{C}$ is smooth if the derivative $\gamma'(t)$ exists and is continuous on $[a,b]$ (with right and left derivatives at a and b, respectively). In other words, if $\gamma(t) = x(t) + iy(t)$, then γ is smooth if and only if both x and y are differentiable (real functions of a real variable) and such that the derivatives $x'(t)$ and $y'(t)$ are continuous functions of the parameter $t \in [a,b]$.

A contour is a piecewise smooth path, that is, $\gamma : [a,b] \to \mathbb{C}$ is a contour if and only if there is a finite collection $a = a_0 < a_1 < \cdots < a_n = b$ (for some $n \in \mathbb{N}$) such that each subpath $\gamma : [a_{i-1}, a_i] \to \mathbb{C}$ is smooth, $1 \le i \le n$.

We write $\gamma = \gamma_1 + \gamma_2 + \cdots + \gamma_n$, where γ_i is the restriction of γ to the subinterval $[a_{i-1}, a_i]$.

Example 7.3 According to our (reasonable) definition, the path given by $t \mapsto \gamma(t) = \cos^3(2\pi t) + i\sin^3(2\pi t)$, for $0 \le t \le 1$, is smooth. Its trace is illustrated in the figure, Fig. 7.1.

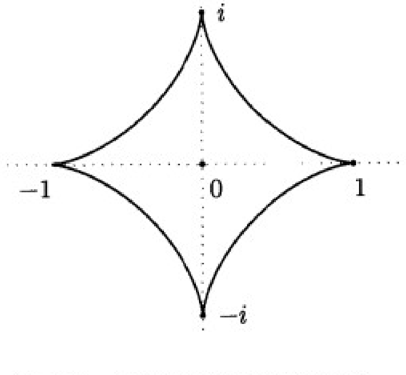

Fig. 7.1 A technically smooth path.

D / 0 0 3 E *0 BB
-3 * ! ! E F E *0 BB . '

0 "" " " ." E 0 BB . '
+)*+,))) # 0! (.

7.2 The Length of a Contour

Next, we wish to develop a means of formulating the length of a path. By way of motivation, suppose that $\gamma : [a,b] \to \mathbb{C}$ is a given smooth path. We can consider an inscribed polygon as an approximation to the path and its length will be an approximation to the length of the path. Suppose that the polygon has vertices $\{z_0, z_1, \ldots, z_n\}$. Then the length of such a polygonal path is the sum of the lengths of its straight line segments, namely, $\sum_{j=1}^{n} |z_j - z_{j-1}|$.

Suppose that the ends of the line segments correspond to values $t_0, t_1, \ldots, t_n$ of the path parameter, that is, $z_j = \gamma(t_j)$. Then

$$\begin{aligned}\sum_{j=1}^{n} |z_j - z_{j-1}| &= \sum_{j=1}^{n} |\gamma(t_j) - \gamma(t_{j-1})| \\ &= \sum_{j=1}^{n} \left| \frac{\gamma(t_j) - \gamma(t_{j-1})}{t_j - t_{j-1}} \right| (t_j - t_{j-1}),\end{aligned}$$

with $a = t_0 < t_1 < \cdots < t_n = b$. Now, the quotient $\dfrac{\gamma(t_j) - \gamma(t_{j-1})}{t_j - t_{j-1}}$ can be considered an approximation to the derivative $\gamma'(t_j)$, in which case the sum on the right hand side becomes an approximation to an integral

$$\sum_{j=1}^{n} \left| \frac{\gamma(t_j) - \gamma(t_{j-1})}{t_j - t_{j-1}} \right| (t_j - t_{j-1}) \quad \overset{\text{approx}}{\rightsquigarrow} \quad \int_a^b |\gamma'(t)| \, dt.$$

Indeed, in view of the assumed smoothness of γ, one might expect that the left hand side converges to the right hand side as $\max_j (t_j - t_{j-1}) \to 0$. Rather than prove this here, we will just take this as motivation for our definition of the length of a path, as follows.

Definition 7.4 The length of the smooth path $\gamma : [a,b] \to \mathbb{C}$ is the non-negative real number $L(\gamma)$ given by the formula

$$L(\gamma) = \int_a^b |\gamma'(t)| \, dt.$$

If $\gamma = \gamma_1 + \gamma_2 + \cdots + \gamma_n$ is a contour (with each γ_j smooth), we define the length of γ to be $L(\gamma) = L(\gamma_1) + L(\gamma_2) + \cdots + L(\gamma_n)$.

Examples 7.4

(1) Let $\gamma(t) = e^{2\pi i t}$, for $0 \le t \le 1$. We calculate

$$L(\gamma) = \int_0^1 |\gamma'(t)|\, dt = \int_0^1 \left|2\pi i e^{2\pi i t}\right|\, dt = 2\pi \int_0^1 1\, dt = 2\pi\,.$$

(2) Let $\gamma(t) = e^{-4\pi i t}$, for $0 \le t \le 1$. Then $\gamma'(t) = -4\pi i e^{-4\pi i t}$, so that

$$L(\gamma) = \int_0^1 4\pi = 4\pi\,.$$

(3) Define the path γ by

$$\gamma(t) = \begin{cases} e^{2\pi i t}, & 0 \le t < 1 \\ 1, & 1 \le t < 2 \\ e^{-2\pi i t}, & 2 \le t \le 3. \end{cases}$$

Then γ is a contour and

$$\gamma'(t) = \begin{cases} 2\pi i e^{2\pi i t}, & 0 \le t < 1 \\ 0, & 1 < t < 2 \\ -2\pi i e^{-2\pi i t}, & 2 < t \le 3. \end{cases}$$

(At $t = 1$, the left derivative is $2\pi i$, the right derivative is 0 and at $t = 2$, the left derivative is 0, the right derivative is $-2\pi i$.) We find

$$L(\gamma) = \int_0^1 2\pi\, dt + \int_1^2 0\, dt + \int_2^3 2\pi\, dt = 4\pi\,.$$

(4) Suppose that $\gamma(t) = z_0 + t(z_1 - z_0)$, $0 \le t \le 1$. Then γ is the line segment $\gamma = [z_0, z_1]$. We have $\gamma'(t) = z_1 - z_0$ and so

$$L(\gamma) = \int_0^1 |z_1 - z_0|\, dt = |z_1 - z_0|\,,$$

as it should.

Remark 7.2 We see that every contour has a well-defined length—this is because of its good behaviour. In general, paths need not have a meaningful length. For example, let $\gamma : [0,1] \to \mathbb{C}$ be the path given by

$$\gamma(t) = \begin{cases} (1-t)\, e^{2\pi i/(1-t)}, & \text{for } 0 \le t < 1 \\ 0, & \text{for } t = 1. \end{cases}$$

The path γ starts at the point $\gamma(0) = 1$ and spirals counterclockwise in towards the origin (and encircles the origin an infinite number of times). Evidently, γ is continuous at any t with $0 \le t < 1$. Furthermore, for any $0 \le t < 1$, $|\gamma(t) - \gamma(1)| = (1-t) \to 0$, as $t \uparrow 1$, and so we see that γ is also continuous at $t = 1$. Hence γ really is a path. However, for any $0 \le s < 1$, the length of γ "from 0 to s" is

$$\begin{aligned}\int_0^s |\gamma'(t)|\, dt &= \int_0^s \Big|\frac{2\pi i}{1-t} - 1\Big|\, dt \\ &\ge \int_0^s \frac{2\pi}{1-t}\, dt = 2\pi \ln\Big(\frac{1}{1-s}\Big) \to \infty\end{aligned}$$

as s increases to 1. One could say that γ has infinite length.

7.3 Integration along a Contour

One way of looking at the integral of a real function of a real variable is to think of it as the area under its graph. This is approximated by the area of suitable rectangles (typically very thin and with height determined by the values of the function at the location of the rectangle).

$$\int_a^b f(x)\, dx = \text{ area under graph } \quad \overset{\text{approx}}{=} \quad \sum f(x_i)\Delta x_i$$

where $f(x_i)\Delta x_i$ is the area (height × base) of the rectangle with height $f(x_i)$ and width Δx_i. This formula makes sense for complex variables and complex-valued functions even though the original idea of area no longer does.

Suppose $\gamma : [a,b] \to \mathbb{C}$ is a smooth path. Let $a = t_0 < t_1 < \cdots < t_n = b$ and set $z_j = \gamma(t_j)$. With the preceding comments in mind, we consider

$$\begin{aligned}\sum f(z_j)(z_j - z_{j-1}) &= \sum f(\gamma(t_j))\, \frac{\gamma(t_j) - \gamma(t_{j-1})}{t_j - t_{j-1}}\, (t_j - t_{j-1}) \\ &\overset{\text{approx ?}}{=} \int_a^b f(\gamma(t))\, \gamma'(t)\, dt\end{aligned}$$

provided that the t_js are close together. It is possible to prove convergence of such sums to integrals (provided f is continuous), but we will simply take the right hand side as our definition of the complex contour integral.

Definition 7.5 Let $\gamma : [a, b] \to \mathbb{C}$ be smooth and f a continuous complex function on $\operatorname{tr}\gamma$. The contour integral of f along γ is defined to be

$$\int_\gamma f \equiv \int_a^b f(\gamma(t))\, \gamma'(t)\, dt.$$

Sometimes the notation $\int_\gamma f(z)\, dz$ is used. If we write $f = u + iv$ and $\gamma(t) = x(t) + iy(t)$, then $\int_\gamma f$ can be written as $X + iY$, where X and Y are real integrals,

$$X = \int_a^b \{u(x(t), y(t))\, x'(t) - v(x(t), y(t))\, y'(t)\}\, dt$$

and

$$Y = \int_a^b \{u(x(t), y(t))\, y'(t) + v(x(t), y(t))\, x'(t)\}\, dt.$$

For any contour γ, we define

$$\int_\gamma f = \int_{\gamma_1} f + \int_{\gamma_2} f + \cdots + \int_{\gamma_n} f$$

where $\gamma_1, \ldots, \gamma_n$ are the smooth parts of the contour γ.

Examples 7.5

(1) Let $\gamma(t) = e^{it}$ for $a \le t \le b$, and take $f(z) = \frac{1}{z}$. Then

$$\begin{aligned}\int_\gamma f &\equiv \int_\gamma \frac{dz}{z} \\ &= \int_a^b f(e^{it})\, ie^{it}\, dt \\ &= \int_a^b \frac{1}{e^{it}}\, ie^{it}\, dt \\ &= \int_a^b i\, dt = i(b-a).\end{aligned}$$

In particular, with $a = 0$ and $b = 2\pi$, we find that $\int_\gamma \frac{dz}{z} = 2\pi i$.

(2) For given $a < b$ in $\mathbb{R}$, let $\gamma(t) = a + t(b-a)$, for $0 \le t \le 1$, that is, γ is just the line segment $[a,b]$ in $\mathbb{R}$. Then, for any f, continuous on $\operatorname{tr}\gamma$,

$$\begin{aligned}\int_\gamma f &= \int_0^1 f(\gamma(t))\,\gamma'(t)\,dt \\ &= \int_0^1 f(a+t(b-a))\,(b-a)\,dt \\ &= \int_a^b f(s)\,ds,\end{aligned}$$

by the change of variable $s = a + t(b-a)$. In particular, if f is real-valued, then $\int_{[a,b]} f$ is precisely the usual real integral $\int_a^b f(x)\,dx$.

Remark 7.3 This last observation has important consequences. Suppose that $\gamma = \gamma_1 + \cdots + \gamma_n$ is a contour where one of the smooth parts is a line segment on the real axis, $\gamma_k = [a,b]$, say, with $a < b \in \mathbb{R}$. Suppose that f is a continuous complex function on $\operatorname{tr}\gamma$ and that f is real on $\operatorname{tr}\gamma_k$, that is, f is real on the line segment $[a,b]$ on the real axis. We have

$$\int_\gamma f = \int_{\gamma_1} f + \cdots + \int_{\gamma_k} f + \cdots + \int_{\gamma_n} f$$

which expresses the real integral $\int_{\gamma_k} f = \int_a^b f(x)\,dx$ in terms of complex integrals. This may not seem much of an advance, but the point is that complex integrals can often be evaluated by general theory based only on rather general properties of the function f. This turns out to be a very powerful method for performing real integration.

Proposition 7.1 *For any contour γ and any functions, f, g, continuous on its trace* $\operatorname{tr}\gamma$*, we have*

(i) $\displaystyle\int_\gamma \alpha f + \beta g = \alpha \int_\gamma f + \beta \int_\gamma g$, *for all α, $\beta \in \mathbb{C}$;*

(ii) $\displaystyle\int_{\widetilde{\gamma}} f = -\int_\gamma f$.

Proof. The integral along a contour is given by the sum of the integrals along its smooth parts, so it is enough to consider the case of a smooth path $\gamma : [a,b] \to \mathbb{C}$. Then (i) is clear from the definition.

To prove (ii), we evaluate the left hand side.

$$\begin{aligned}\int_{\widetilde{\gamma}} f &= \int_a^b f(\widetilde{\gamma}(t))\,\widetilde{\gamma}'(t)\,dt \\ &= -\int_a^b f(\gamma(a+b-t))\,\gamma'(a+b-t)\,dt \\ &= \int_b^a f(\gamma(s))\,\gamma'(s)\,ds, \quad \text{putting } s = a+b-t, \\ &= -\int_a^b f(\gamma(s))\,\gamma'(s)\,ds \\ &= -\int_\gamma f,\end{aligned}$$

as required. □

The next result tells us that the integral along a path is somewhat insensitive to how we choose to parameterize the path.

Proposition 7.2 *The integral $\int_\gamma f$ is independent of the parametrization of the smooth path γ, i.e., if $\gamma : [a,b] \to \mathbb{C}$, $\varphi : [\alpha,\beta] \to [a,b]$ with $\varphi(\alpha) = a$, $\varphi(\beta) = b$, φ is differentiable and φ' is continuous, then*

$$\int_{\widehat{\gamma}} f = \int_\gamma f,$$

where $\widehat{\gamma}$ is the smooth path $\widehat{\gamma} = \gamma \circ \varphi : [\alpha,\beta] \to \mathbb{C}$.

Proof. This is really just a change of variable. Let $f(\gamma(t))\,\gamma'(t) = u(t) + iv(t)$, where u and v are real, and let $U(s) = \int_a^s u(t)\,dt$ and $V(s) = \int_a^s v(t)\,dt$. Put $F(s) = U(s) + iV(s)$. Then $F : [a,b] \to \mathbb{C}$ and $F'(t) = U'(t) + iV'(t) = u(t) + iv(t) = f(\gamma(t))\,\gamma'(t)$. We have

$$\begin{aligned}\int_\gamma f &= \int_a^b f(\gamma(t))\,\gamma'(t)\,dt = F(b) - F(a) = F(\varphi(\beta)) - F(\varphi(\alpha))\\ &= U(\varphi(\beta)) - U(\varphi(\alpha)) + i\big(V(\varphi(\beta)) - V(\varphi(\alpha))\big)\\ &= \int_\alpha^\beta U(\varphi(s))'\,ds + i\int_\alpha^\beta V(\varphi(s))'\,ds\\ &= \int_\alpha^\beta U'(\varphi(s))\,\varphi'(s)\,ds + i\int_\alpha^\beta V'(\varphi(s))\,\varphi'(s)\,ds\\ &= \int_\alpha^\beta F'(\varphi(s))\,\varphi'(s)\,ds = \int_\alpha^\beta f(\gamma(\varphi(s)))\,\gamma'(\varphi(s))\,\varphi'(s)\,ds\\ &= \int_\alpha^\beta f(\gamma\circ\varphi(s))\,(\gamma\circ\varphi)'(s)\,ds = \int_{\gamma\circ\varphi} f\,,\end{aligned}$$

as required. □

Example 7.6 For $0 \le t \le 1$, let $\gamma(t) = 1 + t(i-1)$, and let $\eta(t) = 1 + t^2(i-1)$. This corresponds to the reparametrization $\varphi(t) = t^2$, so that $\eta(t) = \gamma(\varphi(t))$. By way of illustration, let f be the function given by $z \mapsto f(z) = \operatorname{Re} z + 2\operatorname{Im} z$. Then

$$\begin{aligned}\int_\gamma f &= \int_0^1 (1 - t + 2t)(i-1)\,dt\\ &= i\int_0^1 (1+t)\,dt - \int_0^1 (1+t)\,dt.\end{aligned}$$

On the other hand,

$$\begin{aligned}\int_\eta f &= \int_0^1 (1 - t^2 + 2t^2)\,2t(i-1)\,dt\\ &= i\int_0^1 (1+t^2)\,2t\,dt - \int_0^1 (1+t^2)\,2t\,dt\\ &= i\int_0^1 (1+s)\,ds - \int_0^1 (1+s)\,ds, \quad \text{setting } s = t^2,\end{aligned}$$

so $\int_\eta f = \int_\gamma f$, as it should. The actual value is $\frac{3}{2}(i-1)$.

19 04 2020

stating that the equations of motion are invariant under translation or rotation of axes. –

147) A|Books_70_20201
Lecture_Notes_on_Complex_
Analysis/
1860946437.pdf

obs: Use todo o livro.

Pág. 111(124)

Chapter 7

Complex Integration

Pág. 113(126)

7.2. the Length of a Contour.

Polygon:

Condition:

$\gamma : [a, b] \to \mathbb{C}$ is given smooth path.

We can consider an inscribed polygon as an approximation to the path and its length will be an approximation to the length of the path.

polygon $\Rightarrow$ vertices $\{z_0, z_1, \ldots, z_n\}$.

19.04.2020

Then the length of such (de tal) a polygon path is the SUM of the lengths of its straight line segments, namely,

$$\sum_{j=1}^{n} |z_j - z_{j-1}|.$$

Suppose that the ends of the line segments correspond to values $t_0, t_1, \dots, t_n$ of the parameter, that is, $z_j = \gamma(t_j)$.

Then

$$\sum_{j=1}^{n} |z_j - z_{j-1}| = \sum_{j=1}^{n} |\gamma(t_j) - \gamma(t_{j-1})| =$$

$$\sum_{j=1}^{n} \left| \frac{\gamma(t_j) - \gamma(t_{j-1})}{t_j - t_{j-1}} \right| (t_j - t_{j-1}),$$

with $a = t_0 < t_1 < \dots < t_n = b$.

Now, the quotient $\dfrac{\gamma(t_j) - \gamma(t_{j-1})}{t_j - t_{j-1}}$ can be considered an approximation to the derivative $\gamma'(t_j)$, in which case the sum of the right hand side (lado direito) becomes an approximation to an integral.

data 19.04.2020

S T Q Q S S [✓ D]

$$\sum_{j=1}^{n}\left|\frac{\gamma(t_j)-\gamma(t_{j-1})}{t_j-t_{j-1}}\right|(t_j-t_{j-1})$$

$$\xrightarrow{\text{approx}} \int_a^b |\gamma'(t)|\,dt.$$

Indeed (de fato), in view of the assumed smoothness of γ one might expect that the left hand side converges to the right hand side as $\max_j (t_j - t_{j-1}) \to 0$.

Rather than (ao invés de) prove this here, we will just take this as motivation for our definition of the length of a path, as follows.

... estudar todo este capítulo.

Others:

Pág. 49(62)

3.8 Polygons and Paths in $\mathbb{C}$

Pág. 14 (27)

1.6 The Riemann Sphere and the Extended Complex Plane.

Let S^2 denote the sphere →

3.8 Polygons and Paths in $\mathbb{C}$

Definition 3.11 The line segment from z_0 to z_1 is the set

$$[z_0, z_1] = \{\, z : z = z_0 + t(z_1 - z_0),\ 0 \le t \le 1 \,\}$$

It consists of all those points in the complex plane lying on the straight line between z_0 and z_1.

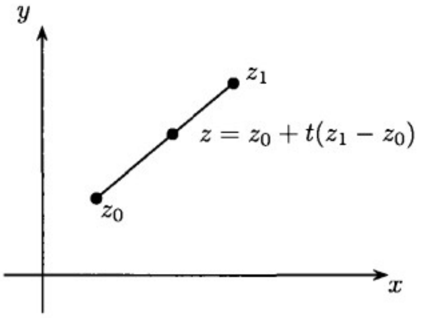

Fig. 3.5 The line segment joining z_0 to z_1.

Remark 3.12 By looking at the real and imaginary parts, we can express this in terms of (familiar) cartesian coordinates. Let $z = x+iy \in [z_0, z_1]$ and write $z_0 = x_0+iy_0$ and $z_1 = x_1+iy_1$. z can be written as $z = z_0+t(z_1-z_0)$ for some $0 \le t \le 1$, which gives the pair of equations

$$\begin{aligned} x &= x_0 + t(x_1 - x_0) \\ y &= y_0 + t(y_1 - y_0). \end{aligned}$$

Eliminating t shows that (x, y) lies on the straight line

$$y = y_0 + \frac{(y_1 - y_0)}{(x_1 - x_0)}\,(x - x_0).$$

Moreover, the restriction $0 \le t \le 1$ means that x varies between x_0 (when $t = 0$) and x_1 (when $t = 1$).

Definition 3.12 A polygon in $\mathbb{C}$ is a set of the form

$$[z_0, z_1] \cup [z_1, z_2] \cup \cdots \cup [z_{n-1}, z_n]$$

for some $n \in \mathbb{N}$ and points $z_0, z_1, \ldots, z_n$.

In other words, a polygon is a finite collection of line segments placed end to end. We say that the polygon joins the points z_0 and z_n. Notice that z_n need not be the same as z_0.

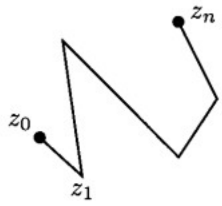

Fig. 3.6 A polygon joining z_0 to z_n.

Definition 3.13 A path in $\mathbb{C}$ is a continuous function $\varphi : [a, b] \to \mathbb{C}$, where $[a, b]$ is some closed interval in $\mathbb{R}$. We say that φ joins its initial point, $\varphi(a)$, to its final point, $\varphi(b)$. The points $\varphi(a)$ and $\varphi(b)$ are its endpoints. The trace (also called the track, or the impression) of the path φ is defined to be the set $\{ z : z = \varphi(t), t \in [a, b] \}$ and is denoted by $\operatorname{tr}\varphi$. (Note that, by definition, a map ψ from a subset S of $\mathbb{R}$ to $\mathbb{C}$ is continuous at the point $x_0 \in S$ if for any given $\varepsilon > 0$ there is $\delta > 0$ such that $|x - x_0| < \delta$, with $x \in S$, implies that $|\psi(x) - \psi(x_0)| < \varepsilon$. ψ is said to be continuous on S if it is continuous at each point of S. This is the natural definition of continuity of a complex-valued function of a real variable.)

So a path is a continuous complex-valued function of a real variable, defined on some closed interval in $\mathbb{R}$, and its track is just its range in $\mathbb{C}$.

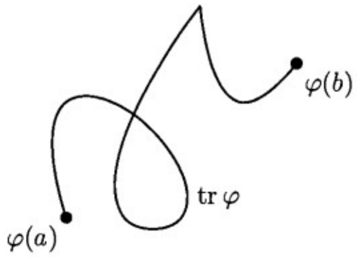

Fig. 3.7 A path joining $\varphi(a)$ to $\varphi(b)$.

Remark 3.13 A polygon can be thought of as a path, as follows. For given points (vertices) $z_0, \ldots, z_n \in \mathbb{C}$, define $\varphi : [0, n] \to \mathbb{C}$ by $\varphi(n) = z_n$ and

$$\varphi(t) = z_k + (t - k)(z_{k+1} - z_k) \quad \text{for } k \le t < k + 1$$

for $k = 0, 1, 2, \ldots, n - 1$. Then as t increases from k to $k + 1$, $\varphi(t)$ moves along the line segment $[z_k, z_{k+1}]$ from z_k towards z_{k+1}. Evidently, the trace of φ is given by $\operatorname{tr} \varphi = [z_0, z_1] \cup [z_1, z_2] \cup \cdots \cup [z_{n-1}, z_n]$.

Remark 3.14 Let $\varphi : [a, b] \to \mathbb{C}$ be a given path. Since φ is continuous, it follows that $\operatorname{tr} \varphi$ is compact in $\mathbb{C}$. To see this, let (z_n) be any sequence in $\operatorname{tr} \varphi$. Then $z_n = \varphi(t_n)$ for some $t_n \in [a, b]$. But (t_n) has a convergent subsequence $t_{n_k} \to \tau$, say, as $k \to \infty$, with $\tau \in [a, b]$. The continuity of φ implies that $\varphi(t_{n_k}) \to \varphi(\tau)$. Thus, $z_{n_k} = \varphi(t_{n_k}) \to \varphi(\tau) \in \operatorname{tr} \varphi$, as $k \to \infty$ and so $\operatorname{tr} \varphi$ is compact. We note, in particular, that any line segment $[z_0, z_1]$ is compact in $\mathbb{C}$.

Remark 3.15 Suppose that $\varphi : [a, b] \to \mathbb{C}$ and $\psi : [c, d] \to \mathbb{C}$ are paths with $\varphi(b) = \psi(c)$, i.e., ψ starts where φ ends. We can join these together into a single path as follows. Define $\gamma : [a, b + d - c] \to \mathbb{C}$, by the formula

$$\gamma(t) = \begin{cases} \varphi(t), & a \le t < b \\ \psi(c + t - b), & b \le t \le b + d - c. \end{cases}$$

We see that γ is a path starting from $\varphi(a)$ and ending at $\psi(d)$. If $c = b$, then γ is equal to φ on the subinterval $[a, b]$ of $[a, d]$ and equal to ψ on $[b, d]$. In any event, we shall write $\gamma = \varphi + \psi$.

3.9 Connectedness

Definition 3.14 A subset $S \subseteq \mathbb{C}$ is said to be pathwise connected if for any pair of points z_0, z_1 in S there is some path, φ, say, lying entirely in S (i.e., such that $\operatorname{tr} \varphi \subseteq S$) and joining z_0 to z_1

Definition 3.15 A set A in $\mathbb{C}$ is said to be polygonally connected if and only if for any two points $z', z'' \in A$ there is some polygon lying entirely within A which joins z' and z''.

Definition 3.16 A set A in $\mathbb{C}$ is said to be stepwise connected if and only if for any two points $z', z'' \in A$ there is some polygon with line segments parallel to either the real or the imaginary axes which lies entirely within A and which joins z' and z''.

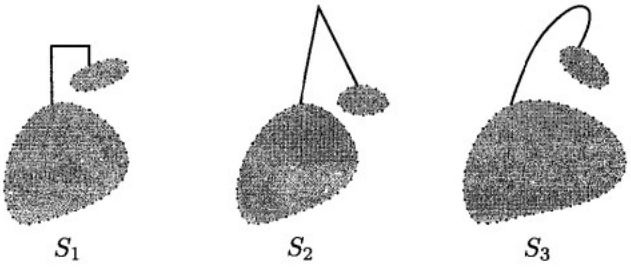

Fig. 3.8 Stepwise, polygonally and pathwise connected sets.

Figure 3.8 illustrates these three notions of connectedness. (Each set comprises two "blobs" (without their boundary) attached by a line.) Notice that S_3 is not polygonally connected, and S_2 is not stepwise connected.

Remark 3.16 The notion of connectedness is supposed to convey the idea of a set being "all in one piece". Clearly, the circle $\{ z : |z| = 1 \}$ is "all one piece" and so should qualify as being connected. However, evidently it is *not* polygonally connected—because it contains no line segments at all (it is everywhere curved!). One could therefore wonder at the usefulness of the notion of polygonal connectedness. It turns out that this is very convenient in the consideration of open sets—which is of special importance in complex analysis. (The precise point here is that pathwise connectedness is equivalent to polygonal connectedness for open sets in $\mathbb{C}$, as we will soon show.)

There is a further notion of connectedness as follows.

Definition 3.17 A subset S of $\mathbb{C}$ is said to be disconnected if there are disjoint open sets A and B (so $A \cap B = \varnothing$) such that $A \cap S \neq \varnothing$, $B \cap S \neq \varnothing$ and $S \subseteq A \cup B$. A set is said to be connected if it is not disconnected.

For open sets, the situation is a little less complicated, as the following proposition shows.

Proposition 3.8 *Let G be an open subset of $\mathbb{C}$. Then G is disconnected if and only if there are non-empty disjoint open sets G_1 and G_2, say, such that $G = G_1 \cup G_2$.*

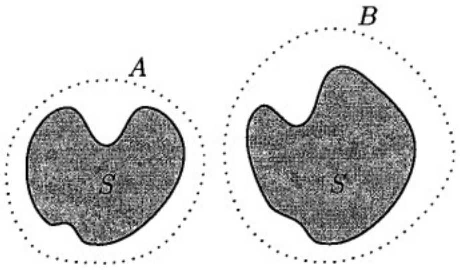

Fig. 3.9 The set S (shaded) is disconnected (the open sets A and B are dotted).

Proof. If G_1 and G_2 exist as stated, then by definition, G is disconnected.

Conversely, suppose that G is disconnected; $G \subseteq A \cup B$, with A, B disjoint, open sets such that $A \cap G \neq \varnothing$ and $B \cap G \neq \varnothing$. Setting $G_1 = A \cap G$ and $G_2 = B \cap G$ gives the open sets required. □

The next result is geometrically obvious, but nevertheless, requires proof. It rests on the completeness property of the real number system (which plays a central role in the Intermediate Value Theorem).

Theorem 3.6 *Suppose that A is a proper non-empty subset of $\mathbb{C}$. Then any path joining a point of A to a point not in A contains a point of the boundary of A. In other words, for a path to "escape" from a set, it must cross its boundary.*

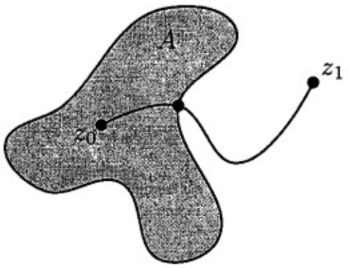

Fig. 3.10 The path must cross the boundary of A.

Proof. Suppose that $\varphi : [a, b] \to \mathbb{C}$ is a path joining $z_0 \in A$ to $z_1 \notin A$.

Let $g : [a, b] \to \mathbb{R}$ be the function

$$g(t) = \begin{cases} 0, & \varphi(t) \in A \\ 1, & \varphi(t) \notin A \,. \end{cases}$$

The function g maps $[a, b]$ onto the points $\{0, 1\}$ and so, by the Intermediate Value Theorem, g cannot be continuous. We will show that g is continuous at every point s in $[a, b]$ for which $\varphi(s) \notin \partial A$. If φ did not meet ∂A then g would be continuous on the whole of $[a, b]$, which we know to be false.

To proceed, then, suppose that $s \in [a, b]$ and that $\varphi(s) \notin \partial A$. Suppose first that $\varphi(s) \in A$, so that $g(s) = 0$. Since $\varphi(s) \notin \partial A$, there is $r > 0$ such that $D(\varphi(s), r) \subseteq A$. By continuity of φ, there is $\delta > 0$ such that

$$|t - s| < \delta, \quad t \in [a, b] \implies \varphi(t) \in D(\varphi(s), r) \subseteq A,$$

that is, $g(t) = 0 = g(s)$. Hence, for any given $\varepsilon > 0$,

$$|g(t) - g(s)| = |0 - 0| = 0 < \varepsilon$$

whenever $|t - s| < \delta$, $t \in [a, b]$. In other words, g is continuous at s.

Now suppose that $\varphi(s) \notin A$. Then $g(s) = 1$. Arguing just as before, we conclude that there is $r' > 0$ such that $D(\varphi(s), r') \subseteq \mathbb{C} \setminus A$ and that there is some $\delta' > 0$ such that

$$\varphi(t) \in D(\varphi(s), r') \subseteq \mathbb{C} \setminus A$$

whenever $|t - s| < \delta'$, $t \in [a, b]$. In other words, $g(t) = 1$ whenever we have $|t - s| < \delta'$, $t \in [a, b]$, which again shows that g is continuous at s.

We have shown that g is continuous at every point s in $[a, b]$ such that $\varphi(s) \notin \partial A$. As we have already noted, the Intermediate Value Theorem implies that g cannot be continuous on the whole of the interval $[a, b]$ and so we conclude that there must be some s for which $\varphi(s) \in \partial A$. In other words, $\operatorname{tr} \varphi \cap \partial A \neq \varnothing$ and the proof is complete. □

We can now prove the equivalence, for open sets, of the various notions of connectedness.

Theorem 3.7 *Let G be an open set. The following are equivalent.*

(i) *G is stepwise connected.*
(ii) *G is polygonally connected.*
(iii) *G is pathwise connected.*
(iv) *G is connected.*

Proof. If $G = \varnothing$ there is nothing to prove, since $\varnothing$ is "connected" in each of the four senses anyway.

Suppose that $G \neq \varnothing$. Clearly, (i) $\Longrightarrow$ (ii) $\Longrightarrow$ (iii). We first show that (iii) $\Longrightarrow$ (iv). Suppose then, that G is a non-empty pathwise connected open set in $\mathbb{C}$.

By way of contradiction, suppose that G is not connected, i.e., that G is disconnected. Then by proposition 3.8 there are disjoint, non-empty open sets A and B such that $G = A \cup B$. Let $z_0 \in A$ and suppose that $z_1 \in B$. By hypothesis, there is a path φ joining z_0 to z_1 in G. Since $z_1 \notin A$, the path φ passes through the boundary of A. That is, there is some $z^* \in \operatorname{tr} \varphi$ with $z^* \in \partial A$. Since A is open, $z^* \notin A$ and therefore $z^* \in B$. But then $D(z^*, r) \subseteq B$ for some $r > 0$. In particular, $D(z^*, r) \cap A = \varnothing$, which contradicts $z^* \in \partial A$. We deduce that G is connected, as required.

Notice that by replacing "path" by "step-path" or by "polygonal path", respectively, it follows that (iv) is a direct consequence of any one of the properties (i), (ii) or (iii).

To show that (iv) $\Longrightarrow$ (i), and hence complete the proof, suppose that G is open and connected. Let z_0 and z_1 belong to G and set

$$A = \{\, \zeta \in G : \zeta \text{ is stepwise connected to } z_0 \text{ in } G \,\}$$
$$B = G \setminus A.$$

We shall show that A is open and non-empty. Since G is open, there is $r > 0$ such that $D(z_0, r) \subseteq G$. Evidently, $D(z_0, r) \subseteq A$, so $A \neq \varnothing$.

Let $w \in A$. Then $w \in G$ and so $D(w, \rho) \subseteq G$ for some $\rho > 0$. Now, any point in $D(w, \rho)$ is stepwise connected to z_0, via w, and so $D(w, \rho) \subseteq A$ and it follows that A is open.

Next we show that B is also open (in fact, we will see that B is empty). If $z' \in B$, then $z' \in G$ and so $D(z', r') \subseteq G$, for some $r' > 0$. Any point w, say, in $D(z', r')$ is stepwise connected to z'. If w were stepwise connected to z_0, then z' would be as well (via w). We conclude that no point of $D(z', r')$ can be stepwise connected to z_0, i.e., $D(z', r') \subseteq B$, showing that B is open. But $G = A \cup B$ and G is connected, by hypothesis. It follows that $B = \varnothing$. Hence $G = A$ and the proof is complete. □

What's going on? The equivalence of these various notions of a set being all in one piece is only claimed to hold for *open* sets in $\mathbb{C}$. They are not the same in general. However, our main concern will be with open sets, so we can appeal to whichever version we find appropriate at any particular time, to suit our own convenience.

We can now easily answer the question "which sets are both open and closed"?

Theorem 3.8 *The only subsets of $\mathbb{C}$ which are simultaneously open and closed are $\varnothing$ and $\mathbb{C}$ itself.*

Proof. By way of contradiction, suppose that A is a non-empty proper subset of $\mathbb{C}$ and that A is both open and closed. It follows, in particular, that $\mathbb{C} \setminus A$, the complement of A, is open. But then we can write $\mathbb{C}$ as the disjoint union of two open sets A and $B = \mathbb{C} \setminus A$, $\mathbb{C} = A \cup B$. Since A is a proper subset of $\mathbb{C}$ it follows that B is non-empty. This means that $\mathbb{C}$ is disconnected.

However, $\mathbb{C}$ is open and evidently polygonally connected (indeed, $[z_0, z_1]$ joins the points z_0 and z_1). But this means that $\mathbb{C}$ is also connected, by theorem 3.7. This gives a contradiction and we deduce that no such non-empty proper subset A exists. □

3.10 Domains

Definition 3.18 An open connected set is called a domain (also sometimes known as a region).

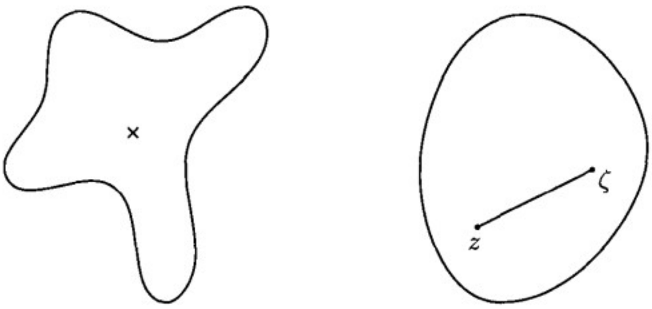

Fig. 3.11 Examples of a star-domain and a convex domain.

Definition 3.19 A domain D is (star-like or) a star-domain if there is some $z_0 \in D$ such that for each $z \in D$ the line segment $[z_0, z]$ lies in D. Any such point z_0 is called a star-centre.

What's going on? We might worry about simply asserting that $i^2 = -1$ without having said precisely what i was in the first place. However, it turns out to be all right. We can just go ahead and write any complex number as $a + ib$, where $i^2 = -1$ and not worry. It can all be justified. (But we still need to sort out square roots.)

Remark 1.5 The set of real numbers has a notion of order, defined in terms of positivity. For any real number x, precisely one of the following three statements is true; $x = 0$, $x > 0$, $-x > 0$. A property of positivity is that if $x > 0$ and $y > 0$, then $xy > 0$. It follows that $1 > 0$. (To see this, first we note that, clearly, $1 \neq 0$ (otherwise, $x = 1\,x$ would be 0 for all $x \in \mathbb{R}$). Hence either $1 > 0$ or $-1 > 0$, but not both. If $-1 > 0$ were true, then we would have $(-1)(-1) > 0$. But $(-1)(-1) = 1^2 = 1$ so that also $1 > 0$. Both $-1 > 0$ and $1 > 0$ is not allowed, so we conclude that $-1 > 0$ is false and therefore $1 > 0$.)

Is there such a notion for complex numbers which extends that for the real numbers? If this were possible then, for example, we would have either $i > 0$ or $-i > 0$ (since $i \neq 0$). If $i > 0$ were true, we would have $-1 = i^2 > 0$, which is false. Hence $-i > 0$ must be true. But then, again, this would imply $-1 = (-i)^2 > 0$, which is false. We must concede that there is no generalization of positivity extending from the real numbers to the set of complex numbers.

For real numbers x and y, the inequality $x > y$ is just a way of writing $x - y > 0$. This latter does not make sense, in general, for complex numbers, so it follows that inequalities, such as $z > \zeta$, do not make sense for complex numbers.

What's going on? Try as we might, we cannot make (useful) sense of inequalities between complex numbers.

1.6 The Riemann Sphere and the Extended Complex Plane

Let S^2 denote the sphere $\{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\}$ in $\mathbb{R}^3$ and let N denote the point $(0, 0, 1)$, the "north pole" of S^2. Think of $\mathbb{C}$ as the plane $\{(x, y, z) : z = 0\}$, containing the equator of S^2. Then given any point P in this plane, the straight line through P and N cuts the sphere S^2 in a unique point, P', say. As P varies over the plane, the corresponding point P' varies over $S^2 \setminus \{N\}$. This sets up a one-one correspondence between $\mathbb{C}$ and $S^2 \setminus \{N\}$.

We note that points far from the origin in $\mathbb{C}$ are mapped into points near the north pole (and points close to the origin are mapped into points close to the south pole of S^2, i.e., the point $(0,0,-1)$). Notice too that if (P_n) is a sequence of points in $\mathbb{C}$ which converges to some point P in $\mathbb{C}$, then the images P'_n of P_n converge in S^2 to the image P' of P. We also see that if (z_n) is a sequence in $\mathbb{C}$ such that $|z_n| \to \infty$, then the sequence (P'_n) of their images converges to N in S^2 (and vice versa). The point N is called the "point at infinity".

The extended complex plane, $\mathbb{C}_\infty$, is defined to be $\mathbb{C}$ together with one additional element, that is, $\mathbb{C}_\infty \equiv \mathbb{C} \cup \{\infty\}$, where $\{\infty\}$ is a singleton set with $\infty \notin \mathbb{C}$. It does not matter what ∞ actually is, as long as it is not already a member of the set $\mathbb{C}$. For example, we could take ∞ to be $\varnothing$, which is certainly not a complex number. (Note that a and $\{\,a\,\}$ are different mathematical objects, so, in particular, $\varnothing$ is not the same as $\{\,\varnothing\,\}$. Indeed, the objects $\{\,\varnothing\,\}$, $\{\,\varnothing, \{\,\varnothing\,\}\,\}$ and $\{\,\varnothing, \{\,\varnothing, \{\,\varnothing\,\}\,\}\,\}$ are different, as different as the numbers $1, 2, 3$.)

What's going on? The issue is how to augment a given set to give it just one new element. That is, given a set A, how does one construct a new set B such that $B \setminus A$ is a singleton set? In the case above, $A = \mathbb{C}$ and $B = \mathbb{C}_\infty$ is the set we seek. If it does not matter what the new element is, as in the case here, then the explicit construction above is just one of many possibilities.

There is then a one-one correspondence between $\mathbb{C}_\infty$ and S^2 given by $\infty \longleftrightarrow N$ together with the correspondence between $\mathbb{C}$ and $S^2 \setminus \{N\}$, as introduced above. The extended complex plane, $\mathbb{C}_\infty$, viewed in this way is referred to as the Riemann sphere.

This gives a sensible realization of "infinity". For example, the mapping $z \mapsto 1/z$ is not *a priori* defined at $z = 0 \in \mathbb{C}$. However, if we consider the extended complex plane, or the Riemann sphere, then in addition to the mapping $z \mapsto 1/z$ for $z \in \mathbb{C} \setminus \{\,0\,\}$, we can define $0 \mapsto N$ and $N \mapsto 0$, (or, in more suggestive notation, $0 \mapsto \infty$ $(1/0 \equiv \infty)$ and $\infty \mapsto 0$ $(1/\infty \equiv 0)$). This defines the map $z \mapsto 1/z$ as a mapping from $\mathbb{C}_\infty \to \mathbb{C}_\infty$. This construction is reasonable in that if $z_n \to z$, then $1/z_n \to 1/z$ even if z or any z_n is equal to 0 or to ∞.

The point here is to notice that by studying $\mathbb{C}_\infty$, rather than just $\mathbb{C}$, we can sometimes handle singularities just as ordinary points—after all, one point on a sphere is much the same as any other.

In real analysis, one considers the limits $x \to \infty$ and $x \to -\infty$. Whilst it must be stressed at the outset that this is just shorthand symbolism, nevertheless, it does invoke a kind of image of two infinities—one positive and the other negative. In view of the picture of complex numbers as points in the plane, one might wonder if it might be worth considering some kind of collection of "complex infinities", each being somewhere off in some given direction (perhaps corresponding to some "end of the rainbow" at the "end" of the ray $r(\cos\theta + i\sin\theta)$ as r becomes very large). The view of $\mathbb{C}$ as being wrapped around a sphere, as developed above, suggests that we can bundle all these "infinities" into just a single "point at infinity", namely, the north pole.

It should be stressed that whilst $\mathbb{C}$ is a field (so one can do arithmetic), this is no longer true of $\mathbb{C}_\infty$. There is no attempt to assign any meaning whatsoever to expressions such as $\infty + \infty$ or $0 \times \infty$. The operations of addition and multiplication are simply not directly applicable when ∞ is involved.

data 19.04.2020

$$\sum_{j=1}^{n} \left| \frac{\gamma(t_j)-\gamma(t_{j-1})}{t_j-t_{j-1}} \right| (t_j-t_{j-1})$$

$$\xrightarrow{\text{approx}} \int_a^b |\gamma'(t)|\,dt.$$

Indeed (de fato), in view of the assumed smoothness of γ one might expect that the left hand side converges to the right hand side as $\max_j (t_j - t_{j-1}) \to 0$.

Rather than (ao invés de) prove this here, we will just take this as motivation for our definition of the length of a path, as follows.

... estudar todo este capítulo.

Others:

Pág. 49 (62)

3.8 Polygons and Paths in $\mathbb{C}$

Pág. 14 (27)

1.6 The Riemann Sphere and the Extended Complex Plane.

Let S^2 denote the sphere →

data 19.04.2020

$\{(x,y,z)\in\mathbb{R}^3 : x^2+y^2+z^2=1\}$ in $\mathbb{R}^3$ and let N denote the point $(0,0,1)$, the "north pole" of S^2.

Think of $\mathbb{C}$ as the plane $\{(x,y,z) : z=0\}$, containing the equator of S^2.

Then given any point P in this plane, the straight line through P and N cuts the sphere S^2 in a unique place, P' say. As P varies over the plane, the corresponding point P' varies over $S^2\setminus\{N\}$.

This sets up a one-one correspondence between $\mathbb{C}$ and $S^2\setminus\{N\}$.

o o estudar

Pag. 66 (79)

4.4. Cauchy-Riemann Equations

$z=x+iy \Longleftrightarrow (x,y)\in\mathbb{R}^2$

$f(x+iy)=\mathrm{Re}\, f(x+iy)+i\,\mathrm{Im}\, f(x+iy)$

$\equiv u(x,y)+iv(x,y)$

o o o estudar

IMPORTANT:

Chapter 5 →

4.4 Cauchy-Riemann Equations

We can relate complex-valued functions of a complex variable to real-valued functions of two real variables by looking at the real and imaginary parts:

$$\begin{aligned} z = x + iy \quad &\longleftrightarrow \quad (x, y) \in \mathbb{R}^2 \\ f(x+iy) &= \operatorname{Re} f(x+iy) + i \operatorname{Im} f(x+iy) \\ &\equiv u(x,y) + iv(x,y). \end{aligned}$$

That is, given $f(z)$, we define $u, v : \mathbb{R}^2 \to \mathbb{R}$ by $u(x,y) = \operatorname{Re} f(x+iy)$ and $v(x,y) = \operatorname{Im} f(x+iy)$.

Conversely, given two functions $u(x,y)$ and $v(x,y)$, we can construct the function $f(z) = u(x,y) + iv(x,y)$, where $z = x + iy$.

The complex differentiability of f implies that u and v, the real and imaginary parts of f, are not completely independent of each other, as we shall now discuss.

Suppose that f is differentiable at $z_0 = x_0 + iy_0$. Then we know that the quotient $(f(z) - f(z_0))/(z - z_0) \to f'(z_0)$ as $z \to z_0$.

We consider two cases, the first where $z = z_0 + s$, with $s \in \mathbb{R}$. To say that $z \to z_0$ is to say that $s \to 0$. Then

$$\begin{aligned} \frac{f(z) - f(z_0)}{z - z_0} &= \frac{u(x_0 + s, y_0) - u(x_0, y_0) + i(v(x_0 + s, y_0) - v(x_0, y_0))}{s} \\ &\to f'(z_0), \end{aligned}$$

as $s \to 0$, with $s \neq 0$. This means that the real and imaginary parts of the left hand side separately have limits, which is to say that the partial derivatives u_x and v_x exist at (x_0, y_0) and, moreover,

$$u_x(x_0, y_0) + iv_x(x_0, y_0) = f'(z_0).$$

Next, consider $z = z_0 + it$ with $t \in \mathbb{R}$. Once again, we know that

$$\begin{aligned} \frac{f(z) - f(z_0)}{z - z_0} &= \frac{u(x_0, y_0 + t) - u(x_0, y_0) + i(v(x_0, y_0 + t) - v(x_0, y_0))}{it} \\ &\to f'(z_0), \end{aligned}$$

as $t \to 0$, with $t \neq 0$. Again, this means that the real and imaginary parts of the left hand side separately have limits. Hence the partial derivatives u_y and v_y exist at (x_0, y_0) and

$$-iu_y(x_0, y_0) + v_y(x_0, y_0) = f'(z_0).$$

Equating these two expressions for $f'(z_0)$ leads to the following relations.

$$\left.\begin{aligned} u_x(x_0,y_0) &= v_y(x_0,y_0) \\ u_y(x_0,y_0) &= -v_x(x_0,y_0) \end{aligned}\right\} \qquad \textbf{Cauchy-Riemann equations}$$

If we denote complex differentiation by D_z, then the discussion above tells us that

$$D_z = \partial_x = -i\partial_y$$

where it is understood that D_z is applied to $f(z)$ and the partial derivatives are applied to $u(x,y) + iv(x,y)$. Thus

$$\begin{aligned} \partial_x u + i\,\partial_x v &= D_z f = f' \\ &= -i\,\partial_y(u+iv) = \partial_y v - i\,\partial_y u. \end{aligned}$$

Equating real and imaginary parts gives the Cauchy-Riemann equations. We have proved the following theorem.

Theorem 4.1 *If the function f is differentiable at $z_0 = x_0 + iy_0$ then f satisfies the Cauchy-Riemann equations at (x_0, y_0).*

Remark 4.3 If f does *not* satisfy the Cauchy-Riemann equations at some point (x_0,y_0), then certainly f fails to be differentiable at $z_0 = x_0 + iy_0$.

Example 4.2 We have already seen that the function $f(z) = \overline{z}$ is nowhere differentiable. Let us consider the Cauchy-Riemann equations for this f.

We have $f(z) = \overline{z} = x - iy$ and so $u(x,y) = x$ and $v(x,y) = -y$. Evidently both u and v possess partial derivatives everywhere in $\mathbb{R}^2$, but $u_x = 1$, $u_y = 0$, $v_x = 0$ and $v_y = -1$. We see that u_x is never equal to v_y and so the Cauchy-Riemann equations are never valid. We can therefore conclude that f is nowhere differentiable (as we already knew).

Remark 4.4 Is the converse of this last theorem true, that is, if f satisfies the Cauchy-Riemann equations at some point (x_0,y_0), is it true that f is differentiable at $z = x_0 + iy_0$? The answer, in general, is no!

Example 4.3 Let f be the function $f(x+iy) = |xy|^{1/2}$. Then we see that $u(x,y) = |xy|^{1/2}$, $v(x,y) = 0$, and

$$u_x(0,0) = u_y(0,0) = 0 = v_x(0,0) = v_y(0,0).$$

(For example, $u_x(0,0) = \lim_{s\to 0} \dfrac{u(s,0) - u(0,0)}{s} = 0$.) From this, it follows that the Cauchy-Riemann equations hold for f at $(0,0)$.

However, consider $s + is$ as $s \to 0$. We have, for $s \neq 0$,

$$\frac{f(s+is)-f(0)}{s+is} = \frac{|s|}{s+is} = \frac{|s|\,(s-is)}{2s^2} = \frac{|s|\,(1-i)}{2s}$$
$$\to \begin{cases} \frac{1}{2}(1-i), & \text{if } s \to 0 \text{ through positive values,} \\ -\frac{1}{2}(1-i), & \text{if } s \to 0 \text{ through negative values.} \end{cases}$$

We conclude that f is not differentiable at $z = 0$ even though f satisfies the Cauchy-Riemann equations there.

Under continuity conditions on the partial derivatives, we can prove the converse. We shall use the following result from the theory of real functions of two real variables.

Lemma 4.1 *Suppose that $\psi(x, y)$ is defined in some disc around the point $(x_0, y_0) \in \mathbb{R}^2$ and has continuous partial derivatives in this disc. Then, for all sufficiently small h and k in $\mathbb{R}$*

$$\psi(x_0 + h, y_0 + k) - \psi(x_0, y_0) = h\,\psi_x(x_0, y_0) + k\,\psi_y(x_0, y_0) + R,$$

where $|R|/\sqrt{h^2+k^2} \to 0$ as $h, k \to 0$ (with $h^2 + k^2 \neq 0$).

Proof. The idea is to find an expression for R and then use the Mean Value Theorem to obtain the required estimate. Indeed, we have

$$\begin{aligned} &\psi(x_0 + h, y_0 + k) - \psi(x_0, y_0) \\ &\qquad = \psi(x_0 + h, y_0 + k) - \psi(x_0 + h, y_0) + \psi(x_0 + h, y_0) - \psi(x_0, y_0) \\ &\qquad = k\,\psi_y(x_0 + h, y_0 + \theta k) + \psi(x_0 + h, y_0) - \psi(x_0, y_0) \end{aligned}$$

for some $0 \leq \theta \leq 1$, by the Mean Value Theorem applied to the function $y \mapsto \psi(x_0 + h, y)$ for y between y_0 and $y_0 + k$,

$$= h\,\psi_x(x_0, y_0) + k\,\psi_y(x_0, y_0) + R$$

where $R = \alpha_1 + \alpha_2$ with

$$\alpha_1 = k\,(\psi_y(x_0 + h, y_0 + \theta k) - \psi_y(x_0, y_0))$$

and

$$\alpha_2 = \psi(x_0 + h, y_0) - \psi(x_0, y_0) - h\,\psi_x(x_0, y_0).$$

Now, if $h, k \to 0$, with $h^2 + k^2 \neq 0$, then

$$\Big|\frac{\alpha_1}{\sqrt{h^2+k^2}}\Big| = \underbrace{\Big|\frac{k}{\sqrt{h^2+k^2}}\Big|}_{\leq 1}\underbrace{\Big|\psi_y(x_0+h, y_0+\theta k) - \psi_y(x_0, y_0)\Big|}_{\to 0}$$

by continuity of ψ_y. Also $\alpha_2/\sqrt{h^2+k^2} = 0$ if $h = 0$, otherwise (i.e., $h \neq 0$)

$$\Big|\frac{\alpha_2}{\sqrt{h^2+k^2}}\Big| = \underbrace{\Big|\frac{h}{\sqrt{h^2+k^2}}\Big|}_{\leq 1}\Big|\underbrace{\frac{\psi(x_0+h, y_0) - \psi(x_0, y_0)}{h}}_{\to \psi_x(x_0,y_0)} - \psi_x(x_0, y_0)\Big|$$

$$\to 0$$

as $h, k \to 0$, with $h \neq 0$. The result follows. □

Theorem 4.2 *Let $f = u + iv$ and suppose that the partial derivatives u_x, u_y, v_x and v_y exist for all (x, y) with $x + iy$ in some open disc around $x_0 + iy_0$ and that these partial derivatives are continuous at (x_0, y_0). Suppose further, that f satisfies the Cauchy-Riemann equations at (x_0, y_0). Then f is complex differentiable at the point $z_0 = x_0 + iy_0$.*

Proof. Let $\lambda = u_x(x_0, y_0) = v_y(x_0, y_0)$ and $\mu = -u_y(x_0, y_0) = v_x(x_0, y_0)$. By lemma 4.1, we can write

$$u(x_0 + h, y_0 + k) - u(x_0, y_0) = h\lambda - k\mu + R_1$$

and

$$v(x_0 + h, y_0 + k) - v(x_0, y_0) = h\mu + k\lambda + R_2$$

for small h, k and where $R_1/\sqrt{h^2+k^2}$ and $R_2/\sqrt{h^2+k^2}$ tend to zero as $h, k \to 0$ (not both zero). Hence, with $\zeta = h + ik \neq 0$,

$$\begin{aligned}\frac{f(z_0+\zeta) - f(z_0)}{\zeta} &= \frac{h\lambda - k\mu + i(h\mu + k\lambda) + R_1 + iR_2}{h + ik} \\ &= \frac{(h+ik)(\lambda + i\mu) + R_1 + iR_2}{h + ik} \\ &= \lambda + i\mu + \underbrace{\frac{R_1 + iR_2}{h + ik}}_{\to 0 \text{ as } h+ik \to 0}.\end{aligned}$$

It follows that f is differentiable at $z_0 = x_0 + iy_0$ and that its derivative is given by $f'(z_0) = \lambda + i\mu$. □

data 19.04.2020

$\{(x,y,z) \in \mathbb{R}^3 : x^2+y^2+z^2=1\}$ in $\mathbb{R}^3$ and let N denote the point $(0,0,1)$, the "north pole" of S^2.

Think of $\mathbb{C}$ as the plane $\{(x,y,z) : z=0\}$, containing the equator of S^2.

Then given any point P in this plane, the straight line through P and N cuts the sphere S^2 in a unique place, P' say. As P varies over the plane, the corresponding point P' varies over $S^2 \setminus \{N\}$.

This sets up a one-one correspondence between $\mathbb{C}$ and $S^2 \setminus \{N\}$.

ou estudar

Pág. 66 (79)

4.4. Cauchy-Riemann Equations

$$z = x+iy \Longleftrightarrow (x,y) \in \mathbb{R}^2$$

$$f(x+iy) = \operatorname{Re} f(x+iy) + i \operatorname{Im} f(x+iy)$$

$$\equiv u(x,y) + i v(x,y)$$

ooo estudar

IMPORTANT:

Chapter 5 →

(%. ## . 8 *00 GGG

Corollary 5.4 *For any* $t \in \mathbb{R}$, $|\exp(it)| = 1$.

Proof. We have $\exp(it) = \cos t + i \sin t$. Both $\cos t$ and $\sin t$ are real if t is, and therefore $|\exp(it)|^2 = \cos^2 t + \sin^2 t = 1$. □

Remark 5.5 For real x, $\sin x$ and $\cos x$ are also real, and so the relation $\sin^2 x + \cos^2 x = 1$ shows that $|\sin x| \le 1$ and also $|\cos x| \le 1$. These inequalities do not extend to the complex case. For example, if $z = it$, with $t \in \mathbb{R}$, then $\sin z = \sin it = (\exp(-t) - \exp t)/2i$. If t is large (and positive), then $\exp t$ is large, $\exp(-t)$ is small and so $|\sin it|$ is large. A similar remark applies to $\cos z$.

5.6 The Appearance of π

We wish to discuss various properties of $\sin x$ and $\cos x$ for real x. In particular, we would like to introduce the number π.

Theorem 5.2 *If* $x \in (0, 2)$, *then* $\sin x > 0$.

Proof. Suppose that $0 < x \le 1$. Then

$$\begin{aligned} \cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots \\ &> 1 - \frac{x^2}{2!} - \frac{x^4}{4!} - \frac{x^6}{6!} - \dots \\ &> 1 - \frac{1}{2} - \frac{1}{2^3} - \frac{1}{2^5} - \dots, \end{aligned}$$

since $\dfrac{x^2}{2!} \le \dfrac{1}{2}$, $\dfrac{x^4}{4!} \le \dfrac{1}{4.3.2} < \dfrac{1}{2^3}, \dots$ etc.,

$$= 1 - \frac{\frac{1}{2}}{(1 - \frac{1}{4})} = \tfrac{1}{3}\,.$$

Hence $\sin' x = \cos x > \frac{1}{3}$ on $[0, 1]$ and so (by the Mean Value Theorem), $\sin x$ is strictly increasing on $[0, 1]$. But $\sin 0 = 0$ and therefore $\sin x > 0$ for $0 < x \le 1$. It follows that

$$\sin x = 2 \sin \tfrac{x}{2} \cos \tfrac{x}{2} > 0$$

whenever $0 < x \le 2$, since, in this case, $\sin \frac{x}{2} > 0$ and $\cos \frac{x}{2} > \frac{1}{3}$. □

Corollary 5.5 *The function* $\cos x$ *is strictly decreasing on* $[0, 2]$.

Proof. We have $\cos' x = -\sin x$ which is strictly negative on $(0, 2)$, by the theorem. By the Mean Value Theorem, it follows that the function $\cos x$ is strictly decreasing on $[0, 2]$. □

Theorem 5.3 $\sin 4 < 0$.

Proof. From the definition of $\sin x$,

$$\begin{aligned}\sin 4 = 4 - \frac{4^3}{3!} + \frac{4^5}{5!} - \frac{4^7}{7!} + \frac{4^9}{9!} \\ - \frac{4^{11}}{11!}\left(1 - \frac{4^2}{12.13}\right) - \frac{4^{15}}{15!}\left(1 - \frac{4^2}{16.17}\right) \\ - \frac{4^{19}}{19!}\left(1 - \frac{4^2}{20.21}\right) - \cdots \\ < 4 - \frac{4^3}{3!} + \frac{4^5}{5!} - \frac{4^7}{7!} + \frac{4^9}{9!} = -\tfrac{268}{405}\end{aligned}$$

and the result follows. □

Theorem 5.4 *There is a unique real number* π *satisfying* $0 < \pi < 4$ *such that* $\sin \pi = 0$. *Furthermore,* $\cos \frac{\pi}{2} = 0$, $\sin \frac{\pi}{2} = 1$ *and* $\cos \pi = -1$.

Proof. We have seen that $\sin x > 0$ for $x \in (0, 2)$ and so, in particular, $\sin 1 > 0$. We have also shown that $\sin 4 < 0$. Now, the map $x \mapsto \sin x$ is continuous on $\mathbb{R}$, and so, by the Intermediate Value Theorem, there is some real number, which we will denote by π, with $1 < \pi < 4$ and such that $\sin \pi = 0$.

We must now show that π is the only value in $(0, 4)$ obeying $\sin \pi = 0$. To this end, suppose that $\sin \alpha = 0$ with $0 < \alpha < 4$. Then

$$0 = \sin \alpha = 2 \sin \tfrac{\alpha}{2} \cos \tfrac{\alpha}{2}.$$

Hence either $\sin \frac{\alpha}{2} = 0$ or $\cos \frac{\alpha}{2} = 0$ (they cannot both vanish because the sum of their squares is equal to 1). But if $0 < \alpha < 4$, then $0 < \frac{\alpha}{2} < 2$ and we know that $\sin x > 0$ on $(0, 2)$, so $\sin \frac{\alpha}{2}$ cannot be zero. Hence, we must have that $\cos \frac{\alpha}{2} = 0$. In particular, $\cos \frac{\pi}{2} = 0$. However, $\cos x$ is strictly decreasing on $[0, 2]$ and so there can be at most one solution to $\cos x = 0$ in this interval. Hence $\frac{\alpha}{2} = \frac{\pi}{2}$ so that $\alpha = \pi$, and the uniqueness is established.

Next, $\sin^2 z + \cos^2 z = 1$ implies that $\sin^2 \frac{\pi}{2} = 1$, since $\cos \frac{\pi}{2} = 0$. Hence $\sin \frac{\pi}{2} = 1$ because $\sin \frac{\pi}{2} > 0$ (since $0 < \frac{\pi}{2} < 2$).

Finally, we have $\cos \pi = 2 \cos^2 \frac{\pi}{2} - 1 = -1$. □

What's going on? We have defined the trigonometric functions, out of the blue as it were, by means of power series. This approach avoids any appeal to geometry and right-angled triangles. However, having chosen this route, we must stick with it. In particular, it is necessary to get to π via these definitions rather than by drawing triangles or circles. As we have seen, this is all perfectly possible. Our view here is that the trigonometric functions are as we have defined them. Any results must be deduced as consequences of these power series definitions.

Remark 5.6 Numerical investigation yields $\pi = 3.14159\ldots$. It is known that π is irrational (in fact, transcendental). It is something of a sport (involving some fascinating numerical analysis) to calculate the value of π to a large number of decimal places and this has been done to over a million decimal places. Such programs have been used to test the computational integrity of supercomputers by checking to see whether they get these digits correct or not.

Theorem 5.5 *For any $z \in \mathbb{C}$ and $n \in \mathbb{Z}$,*

$$\begin{aligned}\sin(z + n\pi) &= (-1)^n \sin z \\ \cos(z + n\pi) &= (-1)^n \cos z.\end{aligned}$$

Proof. We use the trigonometric formulae;

$$\begin{aligned}\sin(z + \pi) &= \sin z \cos \pi + \cos z \sin \pi \\ &= -\sin z + 0,\end{aligned}$$

and

$$\begin{aligned}\cos(z + \pi) &= \cos z \cos \pi - \sin z \sin \pi \\ &= -\cos z - 0.\end{aligned}$$

For $n > 0$, the result now follows by induction. Substituting $z = w - n\pi$, the result then follows for $n < 0$. □

Remark 5.7 Particular cases of the above formulae deserve mention. If we set $z = 0$, then we see that $\sin(n\pi) = 0$ and $\cos(n\pi) = (-1)^n$, for any $n \in \mathbb{Z}$. The functions sin and cos are periodic (with period 2π). In particular, for any $x \in \mathbb{R}$ and any $n \in \mathbb{Z}$,

$$\begin{aligned}\sin(x + 2n\pi) &= \sin x \quad \text{and} \\ \cos(x + 2n\pi) &= \cos x.\end{aligned}$$

If we piece together all the information gained above, we recover the familiar picture of $\sin x$ and $\cos x$, for $x \in \mathbb{R}$, as periodic 'wavy' functions. In fact, $\sin x$ is an odd function, so it is determined by its values on $x \geq 0$. It is periodic, so it is determined by its values on $[0, 2\pi]$. But, $\sin(x + \pi) = \sin x \cos \pi + \cos x \sin \pi = -\sin x$ and so $\sin x$ is determined by its values on the interval $[0, \pi]$. Now,

$$\sin\left(\tfrac{\pi}{2} \pm x\right) = \sin \tfrac{\pi}{2} \cos x \pm \cos \tfrac{\pi}{2} \sin x = \cos x,$$

so we see that $\sin x$ is symmetric about $x = \frac{\pi}{2}$. It follows that $\sin x$ is completely determined by its values on $[0, \frac{\pi}{2}]$.

Furthermore, $\sin(x + \frac{\pi}{2}) = \cos x$ and so the graph of $\cos x$ is got by translating the graph of $\sin x$ by $\frac{\pi}{2}$ to the left.

5.7 Inverse Trigonometric Functions

From the analysis above, we see that $\sin x$ is strictly increasing on the interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$, $\cos x$ is strictly decreasing on the interval $[0, \pi]$ and is strictly increasing on the interval $[-\pi, 0]$. In particular, this means that sin is a one-one map of $[-\frac{\pi}{2}, \frac{\pi}{2}]$ onto $[-1, 1]$, cos is a one-one map of $[0, \pi]$ onto $[-1, 1]$ and also of $[-\pi, 0]$ onto $[-1, 1]$. For $t \in [-1, 1]$, let $\phi(t)$ be the unique element of $[-\frac{\pi}{2}, \frac{\pi}{2}]$ such that $\sin \phi(t) = t$, let $\psi(t)$ be the unique element of $[0, \pi]$ such that $\cos \psi(t) = t$ and let $\rho(t)$ be the unique element of $[-\pi, 0]$ with $\cos \rho(t) = t$. Thus, ϕ is the inverse of $\sin : [-\frac{\pi}{2}, \frac{\pi}{2}] \to [-1, 1]$, ψ is the inverse of $\cos : [0, \pi] \to [-1, 1]$ and ρ is the inverse of $\cos : [-\pi, 0] \to [-1, 1]$.

The standard inverse trigonometric functions $\sin^{-1}$ and $\cos^{-1}$ are given by $\sin^{-1}(t) = \phi(t)$ and $\cos^{-1}(t) = \psi(t)$ for $t \in [-1, 1]$ so that $\sin^{-1}$ takes values in $[-\frac{\pi}{2}, \frac{\pi}{2}]$ whilst $\cos^{-1}$ takes values in $[0, \pi]$.

Theorem 5.6 *Suppose that $f : [a, b] \to [c, d]$ is a strictly increasing (or decreasing) continuous map from $[a, b]$ onto $[c, d]$. Then the inverse map $f^{-1} : [c, d] \to [a, b]$ is continuous.*

Proof. The idea of the proof is straightforward, but it is a nuisance having to consider the end-points c and d of $[c, d]$ and the interior (c, d) separately. To avoid this, we shall first consider the case of $f : \mathbb{R} \to \mathbb{R}$. Suppose, then, that $f : \mathbb{R} \to \mathbb{R}$ is, say, strictly increasing and maps $\mathbb{R}$ onto $\mathbb{R}$. Let $y_0 \in \mathbb{R}$ and $\varepsilon > 0$ be given. Let $x_0 \in \mathbb{R}$ be the unique point such that $f(x_0) = y_0$, thus, $x_0 = f^{-1}(y_0)$. Set $y_1 = f(x_0 - \varepsilon)$ and $y_2 = f(x_0 + \varepsilon)$,

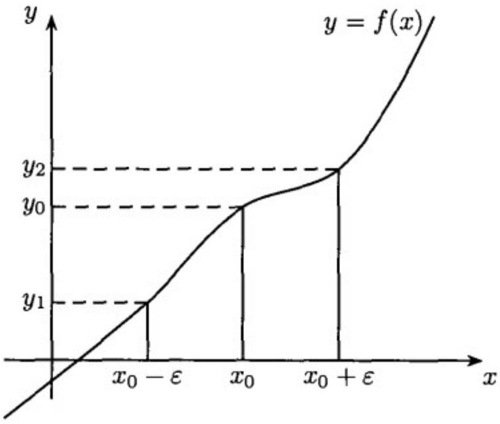

Fig. 5.1 Continuity of the inverse function.

as shown in Fig. 5.1. Then $y_1 < y_0 < y_2$. Furthermore, if $y_1 < y < y_2$, then $x_0 - \varepsilon < f^{-1}(y) < x_0 + \varepsilon$. Putting $\delta = \min\{y_0 - y_1, y_2 - y_0\}$ we see that $|y - y_0| < \delta$ implies that $\left|f^{-1}(y) - f^{-1}(y_0)\right| < \varepsilon$. That is, f^{-1} is continuous at y_0 and therefore on all of $\mathbb{R}$.

The case where f is strictly decreasing is similarly proved. Alternatively, one can note that $f^{-1}(y) = (-f)^{-1}(-y)$ and that $-f$ is increasing if f is decreasing.

Returning now to the case where f is strictly increasing and maps $[a, b]$ onto $[c, d]$, we simply extend f to the whole of $\mathbb{R}$. Define $F : \mathbb{R} \to \mathbb{R}$ by

$$F(x) = \begin{cases} x + f(a) - a, & x < a \\ f(x), & a \leq x \leq b \\ x + f(b) - b, & b < x. \end{cases}$$

Then F maps $\mathbb{R}$ onto $\mathbb{R}$, is strictly increasing, and is equal to f on $[a, b]$. As above, we see that F^{-1} is continuous. In particular, $F^{-1} = f^{-1}$ is continuous on $[c, d]$. □

Theorem 5.7 *Each of the functions ϕ, ψ, and ρ is continuous.*

Proof. This follows immediately from the preceding theorem. □

data 19.04.2020

Pag. 86(99)

5.6 The Appearance of π

Discuss various properties of sen x and cos x for real x.

In particular, we would like to introduce the number π.

→ estudar!!!

Pag. 80(93)

5.2 Complex Hyperbolic Functions

The complex hyperbolic sine and cosine functions are defined as the series

$$\operatorname{senh} z = \sum_{n=0}^{\infty} \frac{z^{2n+1}}{(2n+1)!} = z + \underbrace{\frac{z^3}{3!} + \frac{z^5}{5!} + \frac{z^7}{7!}}_{\text{ímpar}} + \cdots$$

$$\cosh z = \sum_{n=0}^{\infty} \frac{z^{2n}}{(2n)!} = 1 + \underbrace{\frac{z^2}{2!} + \frac{z^4}{4!} + \frac{z^6}{6!}}_{\text{par}} + \cdots$$

estudar!

B) Books_To_2020/ Fluid Mechanics, by Russell C. Hibbeler/

0132777622.pdf

for all $z \in \mathbb{C}$. Evidently, $\exp 0 = 1$, $\sin 0 = 0$ and $\cos 0 = 1$. Furthermore, sin is an odd function and cos is an even function, that is, $\sin(-z) = -\sin z$, and $\cos(-z) = \cos z$, for all $z \in \mathbb{C}$. If x is real, then each of $\exp x$, $\sin x$ and $\cos x$ is also real. The relationship between these three functions is not particularly transparent in the real context, but direct substitution shows that, for any $z \in \mathbb{C}$,

$$\exp iz = \cos z + i \sin z\,.$$

From this it follows that

$$\cos z = \frac{\exp iz + \exp(-iz)}{2} \quad \text{and} \quad \sin z = \frac{\exp iz - \exp(-iz)}{2i}\,.$$

5.2 Complex Hyperbolic Functions

Definition 5.2 The complex hyperbolic sine and cosine functions are defined as the series

$$\sinh z = \sum_{n=0}^{\infty} \frac{z^{2n+1}}{(2n+1)!} = z + \frac{z^3}{3!} + \frac{z^5}{5!} + \frac{z^7}{7!} + \dots$$

$$\cosh z = \sum_{n=0}^{\infty} \frac{z^{2n}}{(2n)!} = 1 + \frac{z^2}{2!} + \frac{z^4}{4!} + \frac{z^6}{6!} + \dots\,.$$

We see that

$$\sinh z = \tfrac{1}{2}\big(\exp z - \exp(-z)\big) \quad \text{and} \quad \cosh z = \tfrac{1}{2}\big(\exp z + \exp(-z)\big)$$

and, moreover, that $\sinh z = -i \sin(iz)$ and $\cosh z = \cos(iz)$. Thus, in the complex variable context, the properties of the hyperbolic functions can be readily obtained from those of the complex trigonometric functions, which shows that this is really the natural context for these functions. This is in sharp contrast to the real variable situation, where the behaviour of the hyperbolic functions on the one hand, and that of the real trigonometric functions on the other, are quite distinct. It is the extension from a real to a complex variable that exposes the otherwise hidden connections.

5.3 Properties of $\exp z$

The basic properties of the complex exponential function are given in the following proposition.

Proposition 5.1 *The exponential function has the following properties:*

(i) $\exp 0 = 1$,
(ii) $\exp(z+w) = \exp z \, \exp w$, *for any* $z, w \in \mathbb{C}$,
(iii) $\exp z \neq 0$, *for all* $z \in \mathbb{C}$,
(iv) $\exp(-z) = 1/\exp z$, *for all* $z \in \mathbb{C}$,
(v) $\exp(x+iy) = \exp x\,(\cos y + i \sin y)$, *for* $x, y \in \mathbb{C}$ *(and, in particular, for any* $x, y \in \mathbb{R}$*).*

Proof. Setting $z = 0$ immediately gives (i). To prove (ii), fix $w \in \mathbb{C}$ and set $f(z) = \exp(z+w)\,\exp(-z)$. Then we find that, for any $z \in \mathbb{C}$,

$$\begin{aligned} f'(z) &= \big(\exp(z+w)\big)'\,\exp(-z) + \exp(z+w)\,\big(\exp(-z)\big)' \\ &= \exp(z+w)\,\exp(-z) - \exp(z+w)\,\exp(-z) \\ &= 0. \end{aligned}$$

It follows that f is constant, and therefore $f(z) = f(0) = \exp w$, that is,

$$\exp w = \exp(z+w)\,\exp(-z).$$

Now set $w = a + b$ and $z = -b$. Then we obtain

$$\exp(a+b) = \exp a \, \exp b,$$

for any $a, b \in \mathbb{C}$, as required.

Using (ii) with $w = -z$, we find that

$$\exp z \, \exp(-z) = \exp 0 = 1, \text{ by (i)},$$

and so (iii) follows, and so does (iv).

To prove (v), let $z = x + iy$ and then apply (ii) to obtain

$$\begin{aligned} \exp z = \exp(x+iy) &= \exp x \, \exp(iy) \\ &= \exp x\,(\cos y + i \sin y). \end{aligned}$$

This holds for any $x, y \in \mathbb{C}$ and so, in particular, also for real x, y. □

Corollary 5.1 *Let e be the real number given by*

$$e = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \ldots = \exp 1.$$

Then $\exp n = e^n$ *for any* $n \in \mathbb{Z}$.

Proof. For each $n \in \mathbb{Z}$, let $P(n)$ be the statement that $\exp n = e^n$. For $n \in \mathbb{N}$, we shall prove the claim by induction. By definition, $\exp 1 = e$ and so $P(1)$ is true. Now let $n \in \mathbb{N}$ and suppose that $P(n)$ is true. Then

$$\begin{aligned}\exp(n+1) &= \exp n \, \exp 1, \text{ by the previous proposition,}\\ &= e^n \exp 1, \text{ by induction hypothesis,}\\ &= e^n e, \text{ since } \exp 1 = e,\\ &= e^{n+1}\end{aligned}$$

so that $P(n+1)$ is true. By induction, $P(n)$ is true for all $n \in \mathbb{N}$.

Clearly $P(0)$ is true, since both $\exp 0$ and e^0 are equal to 1 ($e^0 = 1$, by definition). Now let $n = -m$ with $m \in \mathbb{N}$. Then $\exp m = e^m$ and

$$\exp n = \exp(-m) = \frac{1}{\exp m} = \frac{1}{e^m} = e^{-m} = e^n \,,$$

as required. □

Remark 5.1 It is because of this relationship that one writes e^z for $\exp z$. This notation is often very convenient.

Remark 5.2 Suppose f is entire and satisfies $f'(z) = f(z)$ for all z and $f(0) = 1$. Then $f(z) = \exp z$. To see this, we consider the entire function $g(z) = f(z)\exp(-z)$. We see that $g'(z) = 0$ for all $z \in \mathbb{C}$ and so g is constant, $g(z) = g(0) = 1$. But then we find that $f(z) = \exp z$, as required.

For $z = x + iy$, let $f(z) = e^x(\cos y + i \sin y)$ so that $\operatorname{Re} f(x+iy) = e^x \cos y$ and $\operatorname{Im} f(x+iy) = e^x \sin y$. These functions have continuous partial derivatives throughout $\mathbb{R}^2$ and obey the Cauchy-Riemann equations and so $f(z)$ is analytic at every $z \in \mathbb{C}$. Furthermore, we know that its derivative is given by

$$f'(x+iy) = \partial_x \operatorname{Re} f(x+iy) + i\partial_x \operatorname{Im} f(x+iy) = e^x \cos y + ie^x \sin y = f(x+iy)\,.$$

Since $f(0) = 1$ we conclude that $f(z) = \exp z$.

Sometimes this approach is used to *define* the complex exponential function—assuming that the real functions e^t, $\sin\theta$ and $\cos\theta$ are somehow already known.

Proposition 5.2 *The number e is irrational.*

Proof. For any $k \in \mathbb{N}$, the power series expression for $e = \exp 1$ implies that

$$\sum_{j=0}^{k} \frac{1}{j!} < e = \sum_{j=0}^{k} \frac{1}{j!} + \frac{1}{k!}\Big(\frac{1}{k+1} + \frac{1}{(k+1)(k+2)} + \dots\Big).$$

Since $1/(k+1)(k+2)\dots(k+m) < 1/2^m$ for any $m \in \mathbb{N}$, $m > 1$, we see that the term in brackets on the right hand side above is bounded above by $\sum_{m=1}^{\infty} 1/2^m = 1$ and so

$$\sum_{j=0}^{k} \frac{1}{j!} < e < \sum_{j=0}^{k} \frac{1}{j!} + \frac{1}{k!}$$

for any $k \in \mathbb{N}$. Multiplying through by $k!$ and rearranging, we get

$$0 < k!\Big(e - \sum_{j=0}^{k} \frac{1}{j!}\Big) < 1$$

for any $k \in \mathbb{N}$. If e were rational, then we could write e as $e = p/q$ for some $p, q \in \mathbb{N}$. Setting k in the inequality above to be any integer greater than q and then taking $k!$ inside the bracket, we see that the middle expression is a positive integer strictly less than 1, which is impossible. We conclude that e is irrational. □

Remark 5.3 In fact, it is known that e is transcendental (that is, it is not a root of any real polynomial with integer coefficients, unlike $\sqrt{2}$, for example). Numerical considerations give $e = 2.71828\dots$.

5.4 Properties of $\sin z$ and $\cos z$

We turn now to basic properties of the complex trigonometric functions, sin and cos. These must be established from their power series definitions as given above.

For real α, β, we have that $\exp i(\alpha + \beta) = \cos(\alpha + \beta) + i \sin(\alpha + \beta)$. However,

$$\begin{aligned}
\exp i(\alpha + \beta) &= \exp i\alpha \, \exp i\beta \\
&= (\cos\alpha + i \sin\alpha)\,(\cos\beta + i \sin\beta) \\
&= \cos\alpha\cos\beta - \sin\alpha\sin\beta + i(\sin\alpha\cos\beta + \cos\alpha\sin\beta).
\end{aligned}$$

Equating real and imaginary parts leads to the addition rules

$$\sin(\alpha+\beta)=\sin\alpha\cos\beta+\cos\alpha\sin\beta$$

and

$$\cos(\alpha+\beta)=\cos\alpha\sin\beta-\sin\alpha\sin\beta,$$

for any $\alpha,\beta\in\mathbb{R}$. It is natural to ask whether these relationships are also valid for complex α and β. In fact they are, as we now show.

5.5 Addition Formulae

Theorem 5.1 (Addition Formulae) *For any* $a,b\in\mathbb{C}$,

$$\begin{aligned}\sin(a+b)&=\sin a\,\cos b+\cos a\,\sin b\\ \cos(a+b)&=\cos a\,\cos b-\sin a\,\sin b.\end{aligned}$$

Proof. The easiest way of proving these relations is to write the right hand sides in terms of the exponential function and do a bit of algebra. We find

$$\begin{aligned}\sin a\cos b+{}&\cos a\sin b\\ &=\frac{(e^{ia}-e^{-ia})}{2i}\,\frac{(e^{ib}+e^{-ib})}{2}\\ &\qquad+\frac{(e^{ia}+e^{-ia})}{2}\,\frac{(e^{ib}-e^{-ib})}{2i}\\ &=\frac{(e^{ia}e^{ib}+e^{ia}e^{-ib}-e^{-ia}e^{ib}-e^{-ia}e^{-ib})}{4i}\\ &\qquad+\frac{(e^{ia}e^{ib}-e^{ia}e^{-ib}+e^{-ia}e^{ib}-e^{-ia}e^{-ib})}{4i}\\ &=\frac{(e^{i(a+b)}+e^{i(a-b)}-e^{-i(a-b)}-e^{-i(a+b)})}{4i}\\ &\qquad+\frac{(e^{i(a+b)}-e^{i(a-b)}+e^{-i(a-b)}-e^{-i(a+b)})}{4i}\\ &=\frac{(e^{i(a+b)}-e^{-i(a+b)})}{2i}\\ &=\sin(a+b).\end{aligned}$$

Similarly, we calculate

$$\begin{aligned}\cos a\cos b+\sin a\sin b&=\frac{(e^{ia}+e^{-ia})}{2}\,\frac{(e^{ib}+e^{-ib})}{2}\\&\qquad+\frac{(e^{ia}-e^{-ia})}{2i}\,\frac{(e^{ib}-e^{-ib})}{2i}\\&=\frac{(e^{i(a+b)}+e^{-i(a+b)})}{2}\\&=\cos(a+b),\end{aligned}$$

as required. ☐

Corollary 5.2 *For any* $z\in\mathbb{C}$, $(\sin z)^2+(\cos z)^2=1$.

Proof. The identity $\cos(a+b)=\cos a\cos b-\sin a\sin b$, with $a=z$ and $b=-z$, gives

$$\begin{aligned}\cos 0&=\cos z\cos(-z)-\sin z\sin(-z)\\&=\cos z\cos z+\sin z\sin z,\end{aligned}$$

where we have used $\cos(-z)=\cos z$ and $\sin(-z)=-\sin z$. The result now follows since $\cos 0=1$. ☐

Remark 5.4 An alternative proof is to set $\psi(z)=\sin^2 z+\cos^2 z$ and then to calculate $\psi'(z)$. One finds that $\psi'(z)=0$, for all $z\in\mathbb{C}$, and so it follows that ψ is constant. Therefore $\psi(z)=\psi(0)=1$, as required.

In the same spirit, we can prove the addition formulae by considering the functions

$$f(z)=\sin(w-z)\cos z+\cos(w-z)\sin z$$

and

$$g(z)=\cos(w-z)\cos z-\sin(w-z)\sin z\,.$$

One finds that $f'(z)=0$ and $g'(z)=0$ for all z, so that both f and g are constant: $f(z)=f(0)=\sin w$ and $g(z)=g(0)=\cos w$. Letting $w=a+b$ and setting $z=b$, we get the required formulae.

Corollary 5.3 *For any* $z\in\mathbb{C}$, *we have*

$$\begin{aligned}&\sin 2z=2\sin z\cos z\quad\textit{and}\\&\cos 2z=\cos^2 z-\sin^2 z=2\cos^2 z-1=1-2\sin^2 z.\end{aligned}$$

Proof. We simply put $a=b=z$ in the addition formulae above and use the identity $\sin^2 z+\cos^2 z=1$. ☐

Corollary 5.4 *For any* $t \in \mathbb{R}$, $|\exp(it)| = 1$.

Proof. We have $\exp(it) = \cos t + i \sin t$. Both $\cos t$ and $\sin t$ are real if t is, and therefore $|\exp(it)|^2 = \cos^2 t + \sin^2 t = 1$. □

Remark 5.5 For real x, $\sin x$ and $\cos x$ are also real, and so the relation $\sin^2 x + \cos^2 x = 1$ shows that $|\sin x| \le 1$ and also $|\cos x| \le 1$. These inequalities do not extend to the complex case. For example, if $z = it$, with $t \in \mathbb{R}$, then $\sin z = \sin it = (\exp(-t) - \exp t)/2i$. If t is large (and positive), then $\exp t$ is large, $\exp(-t)$ is small and so $|\sin it|$ is large. A similar remark applies to $\cos z$.

5.6 The Appearance of π

We wish to discuss various properties of $\sin x$ and $\cos x$ for real x. In particular, we would like to introduce the number π.

Theorem 5.2 *If* $x \in (0, 2)$, *then* $\sin x > 0$.

Proof. Suppose that $0 < x \le 1$. Then

$$\begin{aligned}\cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots \\ &> 1 - \frac{x^2}{2!} - \frac{x^4}{4!} - \frac{x^6}{6!} - \dots \\ &> 1 - \frac{1}{2} - \frac{1}{2^3} - \frac{1}{2^5} - \dots,\end{aligned}$$

since $\dfrac{x^2}{2!} \le \dfrac{1}{2}$, $\dfrac{x^4}{4!} \le \dfrac{1}{4.3.2} < \dfrac{1}{2^3}$, ... etc.,

$$= 1 - \frac{\frac{1}{2}}{(1 - \frac{1}{4})} = \tfrac{1}{3}.$$

Hence $\sin' x = \cos x > \frac{1}{3}$ on $[0, 1]$ and so (by the Mean Value Theorem), $\sin x$ is strictly increasing on $[0, 1]$. But $\sin 0 = 0$ and therefore $\sin x > 0$ for $0 < x \le 1$. It follows that

$$\sin x = 2 \sin \tfrac{x}{2} \cos \tfrac{x}{2} > 0$$

whenever $0 < x \le 2$, since, in this case, $\sin \frac{x}{2} > 0$ and $\cos \frac{x}{2} > \frac{1}{3}$. □

data 19.04.2020

Pag. 86 (99)

5.6 The Appearance of π

Discuss various properties of sen x and cos x for real x.

In particular, we would like to introduce the number π.

→ estudar!!!

Pag. 80 (93)

5.2 Complex Hyperbolic Functions

The complex hyperbolic sine and cosine functions are defined as the series

$$\operatorname{senh} z = \sum_{n=0}^{\infty} \frac{z^{2n+1}}{(2n+1)!} = z + \underbrace{\frac{z^3}{3!} + \frac{z^5}{5!} + \frac{z^7}{7!}}_{\text{impar}} + \cdots$$

$$\cosh z = \sum_{n=0}^{\infty} \frac{z^{2n}}{(2n)!} = 1 + \underbrace{\frac{z^2}{2!} + \frac{z^4}{4!} + \frac{z^6}{6!}}_{\text{par}} + \cdots$$

estudar!

B) Books_To_2020 / Fluid Mechanics, by Russell C. Hibbeler /

0132777622.pdf

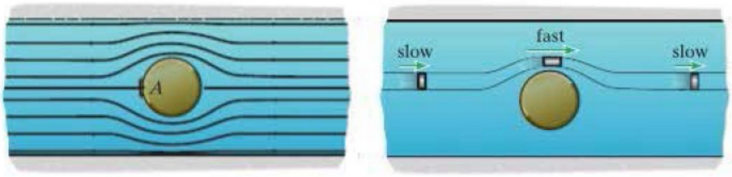

Streamlines for ideal fluid flow around a cylinder

(a)

Fluid flows faster within a narrow streamtube

(b)

Fig. 3–6

3.3 Graphical Descriptions of Fluid Flow

Several graphical methods have been devised to visualize the behavior of a flow. For analysis, these include using streamlines or streamtubes, and for experimental work, pathlines and streaklines, and optical methods are often used. We will now give each of these separate treatment.

Streamlines. A ***streamline*** is a curve that is drawn through the fluid in such a manner that it indicates the direction of the velocity of particles located on it at a particular *instant of time*. Specifically, the velocity of any particle is always *tangent* to the streamline along which it is traveling at that instant. Consequently, no fluid can flow across a streamline, only *along* the streamline. For example, the streamlines for the steady flow of an *ideal fluid* around a cylinder enclosed within a rectangular duct are shown in Fig. 3–6*a*. In particular, notice the streamline at the center of the flow field. It intersects the cylinder at A. This point is called a ***stagnation point,*** because here the velocity of any particle is *momentarily* reduced to zero when it strikes the cylinder's surface.

It is important to keep in mind that streamlines are used to represent the flow field during each *instant of time*. In the example just given, the direction of the streamlines is *maintained* as time passes and fluid particles move along them; however, sometimes the streamlines can be a function of *both* space and time. For example, this occurs when fluid flows through a rotating pipe. Here the streamlines *change position* from one instant to the next, and so the fluid particles do not move along any streamline that is fixed in position.

Since the particle's velocity is *always tangent* to its streamline, then if the velocity is known, the equation that defines a streamline at a particular instant can always be determined. For example, in the case of two-dimensional flow, the particle's velocity will have two components, u in the x direction and v in the y direction. As noted in Fig. 3–7, we require

$$\boxed{\frac{dy}{dx} = \frac{v}{u}} \qquad (3\text{–}3)$$

Integrating this equation will give the *equation of the streamlines*, $y = f(x) + C$. To find the streamline that passes through a particular point (x_0, y_0), we can substitute these coordinates into this equation to evaluate the constant C. This process is demonstrated in Examples 3.1 and 3.2.

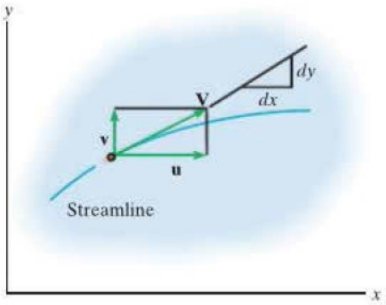

The velocity is always tangent to the streamline

Fig. 3–7

Streamtubes. For some types of analysis, it is convenient to consider a *bundle of streamlines* that surround a region of flow, Fig. 3–8. Such a circumferential grouping is called a ***streamtube***. Here the fluid flows through the streamtube as if it were contained within a curved conduit.

In two dimensions, streamtubes can be formed between any two streamlines. For example, consider the streamtube in Fig. 3–6*b* for flow around the cylinder. Notice that a small element of fluid traveling along this streamtube will move *slower* when the streamlines are *farther apart* (or the streamtube is wider), and move *faster* when they are *closer together* (or the streamtube is narrow). This is a consequence of the conservation of mass, something we will discuss in the next chapter.

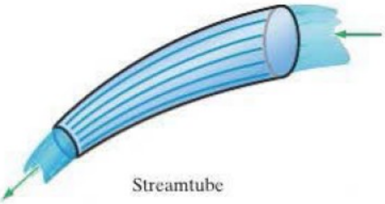

Fig. 3–8

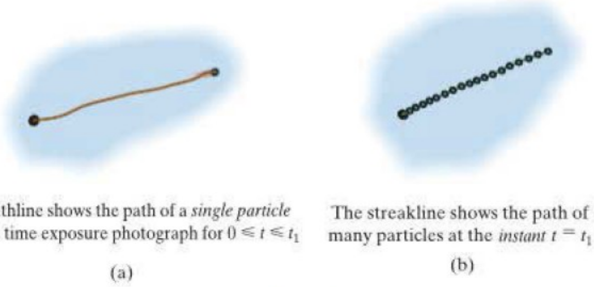
The pathline shows the path of a *single particle* using a time exposure photograph for $0 \leq t \leq t_1$

(a)

The streakline shows the path of many particles at the *instant* $t = t_1$

(b)

Fig. 3–9

This photo shows the streaklines of water particles ejected from a water sprinkler at a given instant.

Pathlines. The ***pathline*** for a fluid particle defines the "path" the particle travels over a *period of time*. To obtain the pathline experimentally, a neutrally buoyant *single particle* can be released within the flow stream and a *time-exposed photograph* taken. The line on the photograph then represents the pathline for this particle, Fig. 3–9*a*.

Streaklines. If smoke is released continuously in a gas, or colored dye is released in a liquid, a "trace" of *all the particles* will be carried along with the flow. This resulting succession of marked particles that have all come from the same point of origin is called a ***streakline***. It can be identified by taking an *instantaneous photograph* of the trace, or "streak," of *all the particles*, Fig. 3–9*b*.

As long as the flow is *steady*, the streamlines, pathlines, and streaklines will all *coincide*. For example, this occurs for the steady flow of water ejected from the *fixed* nozzle, Fig. 3–10. Here a streamline will maintain its same direction from one instant to the next, and so *every particle* coming from the nozzle will follow this same streamline, thereby producing a coincident pathline and streakline.

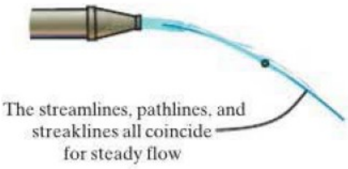
The streamlines, pathlines, and streaklines all coincide for steady flow

Fig. 3–10

Optical Methods. If the fluid is transparent, such as air and water, then the flow can be visualized indirectly using a shadowgraph. Simply put, a ***shadowgraph*** is the result of refracted or bent light rays that interact with the fluid and then cast a shadow on a screen in close proximity. Perhaps you have noticed the shadow of a heated air plume rising from a candle, or the shadow of the exhaust from a jet engine cast against a surface. In both of these cases the heat being produced changes the local density of the air, and the light passing through this air is bent. The greater the change in the fluid density, the more the light rays are bent. Shadowgraphs have been used in industry to visualize the air flow around jet aircraft and rockets, as well as to study the flow around devices that produce heat.

One other optical technique that is used to visualize the flow of a transparent fluid is **schlieren photography**. It also is based on detecting the density gradients within a fluid produced by the flow. In this case, a collimated or parallel beam of light shines on the object. The light is focused with a lens, and a knife edge is placed perpendicular to the beam at its focal point. This blocks about half the light and the fluid causes a distortion in the beam due to the density variations within the fluid. The distortion will cast light and dark regions on a nearby background surface. An example is shown in the accompanying photo of a heated source generating a flow of hot air used to suspend a ball. Schlieren photography has been used extensively in aeronautical engineering to visualize the formation of shockwaves and expansion waves formed around jet aircraft and missiles. For further details on these two optical techniques, see Ref. [5].

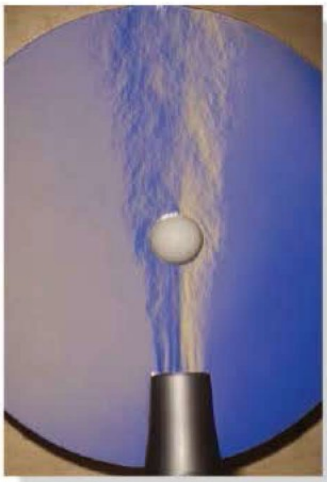

A schlieren photograph showing a ball that is suspended by an airstream produced by a jet of hot air. (© Ted Kinsman/ Science Source)

Computational Fluid Dynamics. Although the experimental techniques just discussed have played an important role in studying complicated flow patterns, they have been increasingly replaced with numerical techniques that apply the laws of fluid dynamics using high-speed computers. This is referred to as **computational fluid dynamics** (CFD). There are many types of commercially available CFD programs, and with some of them the engineer can extend an analysis to include heat transfer and multiphase changes within the fluid. Generally the output includes a plot of the streamlines or pathlines, with the velocity field displayed in a color-coded fashion. Digital printouts of steady flow can be made, or if the flow is unsteady, then a video can be produced. We will discuss this important field in greater length in Sec. 7.12.

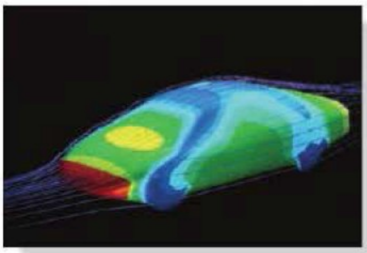

A CFD printout showing the flow around a model of a car. Based on the results, the model can be reshaped to enhance the design. (© Hank Morgan/Science Source)

Important Points

- There are two ways to describe the motion of fluid particles within a flow. A *Lagrangian description*, or system approach, has limited use since it requires tracking the location of *each particle* in the flow and reporting its motion. A *Eulerian description*, or control volume approach, is more practical, since it considers a specific region or point in the flow, and measures the motion of any fluid particles passing through this region or point.
- Fluid flow can be described in various ways. It can be classified as the result of viscous friction, which is, either laminar, transitional, or turbulent. It can be classified as nondimensional, or as one-, two-, or three-dimensional. And finally, for a space–time classification, the flow is *steady* if it does not vary with time, and *uniform* if it does not vary with location.
- A *streamline* contains particles that have velocities that are tangent to the streamline at each instant of time. If the flow is steady, then the particles move along fixed streamlines. However, if unsteady flow causes the direction of the streamlines to change, then the particles will move along streamlines that have a different orientation from one instant to the next.
- Experimentally, the flow of a *single* marked particle can be visualized using a *pathline*, which shows the path taken by the particle using a time-lapse photograph. *Streaklines* are formed when smoke or colored dye is released into the flow from the same point, and an instantaneous photo is taken of the trace of *many* particles.
- Optical methods, such as shadowgraphs and schlieren photographs, are useful for observing flow in transparent fluids caused by heat, or to visualize shock and expansion waves created in high speed flows.
- Computational fluid dynamics uses numerical analysis to apply the laws of fluid dynamics, and thereby produces data to visualize complex flow.

EXAMPLE 3.1

The velocity for the two-dimensional flow shown in Fig. 3–11 is defined by $\mathbf{V} = \{6y\mathbf{i} + 3\mathbf{j}\}\,\text{m/s}$, where y is in meters. Determine the equation of the streamline that passes through point (1 m, 2 m).

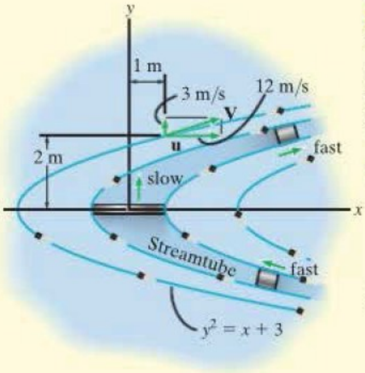

Fig. 3–11

SOLUTION

Fluid Description. This is a Eulerian description since the velocity is reported in terms of its spatial coordinates. In other words, the description reports the velocity of particles that pass through a control volume located at point (x, y) within the flow. Since time is not involved, we have steady flow, where $u = (6y)\,\text{m/s}$ and $v = 3\,\text{m/s}$.

Analysis. To find the equations of the streamlines, we must use Eq. 3–3,

$$\frac{dy}{dx} = \frac{v}{u} = \frac{3}{6y}$$

Separating the variables and integrating yields

$$\int 6y\,dy = \int 3\,dx$$

$$3y^2 = 3x + C$$

This is the equation for a parabola. Each selected constant C will produce a unique streamline. For the one passing through the point (1 m, 2 m), we require $3(2)^2 = 3(1) + C$, or $C = 9$. Therefore,

$$y^2 = x + 3 \qquad \textit{Ans.}$$

A plot of this equation (streamline) is shown in Fig. 3–11. Here particles that pass through a differential size control volume located at the point (1 m, 2 m) will have the velocity $\mathbf{V} = \{12\mathbf{i} + 3\mathbf{j}\}$ m/s, as indicated. By selecting other points to evaluate the integration constant C, we can establish the graphical representation of the entire flow field, a portion of which is also shown in Fig. 3–11. Once this is established notice how the fluid element constrained in the selected streamtube travels fast, slows down as it passes the x axis, and then speeds up again.

NOTE: If the velocity were also a function of time, then unsteady flow would occur, and the streamlines may change location from one instant to the next. To obtain them, we would first have to evaluate $\mathbf{V}$ at a particular instant of time, and then apply Eq. 3–3 in the same manner as explained here to establish the streamlines or flow field for this instant.

EXAMPLE 3.2

The velocity components of a particle in the flow field are defined by $u = 3\text{ m/s}$ and $v = (6t)\text{ m/s}$, where t is in seconds. Plot the pathline for the particle if it is released from the origin when $t = 0$. Also draw the streamline for this particle when $t = 2\text{ s}$.

SOLUTION

Fluid Description. Since the velocity is only a function of time, this is a Lagrangian description of the particle's motion. We have nonuniform flow.

Pathline. The pathline describes the location of the particle at various times. Since the particle is at (0, 0) when $t = 0$, then

$$u = \frac{dx}{dt} = 3 \qquad\qquad v = \frac{dy}{dt} = 6t$$

$$\int_0^x dx = \int_0^t 3\,dt \qquad\qquad \int_0^y dy = \int_0^t 6t\,dt$$

$$x = (3t)\text{ m} \qquad\qquad y = \left(3t^2\right)\text{ m} \tag{1}$$

Eliminating the time between these two parametric equations, we obtain our result

$$y = 3\left(\frac{x}{3}\right)^2 \quad \text{or} \quad y = \frac{1}{3}x^2 \qquad \textit{Ans.}$$

The pathline or path taken by the particle is a parabola, shown in Fig. 3–12.

Streamline. The streamline shows the direction of the velocity of a particle at a particular instant of time. At the instant $t = 2\text{s}$, using Eq. 1, the particle is located at $x = 3(2) = 6\text{m}$, $y = 3(2)^2 = 12\text{m}$. Also it has velocity components of $u = 3\text{ m/s}$, $v = 6(2) = 12\text{ m/s}$, Fig 3–12. Applying Eq. 3–3 to obtain the equation of the streamline at this instant, we have

$$\frac{dy}{dx} = \frac{v}{u} = \frac{12}{3} = 4$$

$$\int dy = \int 4\,dx$$

$$y = 4x + C$$

Since $x = 6\text{m}$ and $y = 12\text{m}$, then $C = -12\text{ m}$. Therefore

$$y = 4x - 12$$

This streamline is shown in Fig. 3–12. Note that both the streamline and pathline have the same slope at (6 m, 12 m). This is to be expected since both must give the same direction for the velocity when $t = 2\text{ s}$. At another instant, the particle moves along the pathline, and has another streamline (tangent).

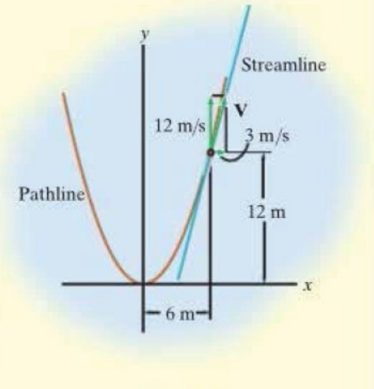

Fig. 3–12

EXAMPLE 3.3

The velocity of gas particles flowing along the center of the pipe in Fig. 3–13 is defined for $x \geq 1\,\text{m}$, by the velocity field $V = (t/x)\,\text{m/s}$, where t is in seconds and x is in meters.* If a particle is at $x = 1\,\text{m}$ when $t = 0$, determine its velocity when it is at $x = 2\,\text{m}$.

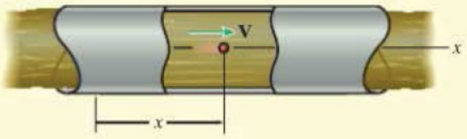

Fig. 3–13

SOLUTION

Fluid Description. Because the velocity is a function of time, the flow is unsteady, and because it is a function of position, it is nonuniform. This flow field $V = V(x, t)$ is a Eulerian description. Here *all particles* passing the location x have a velocity that changes with time, and it is measured as $V = (t/x)$.

Analysis. To find the velocity at $x = 2$ m, we must first find the time for the particle to travel from $x = 1$ m to $x = 2$ m, that is, we must find $x = x(t)$. This is a Lagrangian description since we must follow the motion of a single particle.

The particle's position can be determined from the velocity field, by realizing that for a Lagrangian description, Eq. 3–1 gives

$$V = \frac{dx}{dt} = \frac{t}{x}$$

Separating the variables and integrating yields

$$\int_1^x x\,dx = \int_0^t t\,dt$$

$$\frac{x^2}{2} - \frac{1}{2} = \frac{1}{2}t^2$$

$$x = \sqrt{t^2 + 1} \qquad (1)$$

---

*Notice that substituting t in seconds and x in meters gives units of s/m; however, here there must be a constant $1\,\text{m}^2/\text{s}^2$ (not shown) that converts these units to velocity, m/s, as stated.

The Lagrangian description of the velocity of the particle can now be determined. Here the velocity must only be a function of time.

$$V = \frac{dx}{dt} = \frac{1}{2}\left(t^2 + 1\right)^{-1/2}(2t)$$

$$= \frac{t}{\sqrt{t^2 + 1}} \qquad (2)$$

Using these results, we are now able to follow the particle as it moves along the path during each instant of time. Its position at any instant is determined from Eq. 1, and its velocity at this instant is determined from Eq. 2. Therefore, when the particle is located at $x = 2$ m, the time is

$$2 = \sqrt{t^2 + 1}$$
$$t = 1.732 \text{ s}$$

And at this time, the particle is traveling at

$$V = \frac{1.732}{\sqrt{(1.732)^2 + 1}}$$

$$V = 0.866 \text{ m/s} \qquad \textit{Ans.}$$

We can check this result, using the Eulerian description. Since we require the particle to be located at $x = 2$ m, when $t = 1.732$ s its velocity is

$$V = \frac{t}{x} = \frac{1.723}{2} = 0.866 \text{ m/s} \qquad \textit{Ans.}$$

as expected.

NOTE: In this example the velocity field has only one component, $V = u = t/x, v = 0, w = 0$. In other words, it is one-dimensional flow, and so the streamline does not change direction, rather it remains in the x direction.

3.4 Fluid Acceleration

Once the velocity field $\mathbf{V} = \mathbf{V}(x, y, z, t)$ is established, then it is possible to determine the acceleration field for the flow. One reason for doing this is to apply Newton's second law of motion, $\Sigma\mathbf{F} = m\mathbf{a}$, to relate the acceleration of the fluid particles to the forces produced by the flow. Since the velocity field is a function of *both* space and time (Eulerian description), the *time* rate of change of velocity (acceleration) must account for the changes made in *both* its space and time variables. To show how to obtain these changes, let us consider the unsteady, nonuniform flow of a fluid through the nozzle in Fig. 3–14. For simplicity, we will study the motion of a single particle as it moves through a control volume. Here the velocity is defined by $V = V(x, t)$ along the center streamline. When the particle is at position x, where the first open control surface to the control volume is located, it will have a velocity that will be *smaller* than when it is at position $x + \Delta x$, where it exits the control volume's other open control surface. This is because the nozzle constricts the flow and therefore causes the particle to have a greater speed at $x + \Delta x$. Thus the velocity of the particle will change due to its *change* Δx *in position* (nonuniform flow). If the valve is opened then the velocity of the particle can *also change within* the control volume due to a *change in time* Δt (unsteady flow). As a result, the total change in $V = V(x, t)$ will be

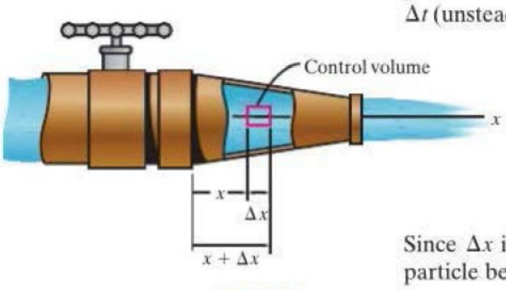

Fig. 3–14

$$\Delta V = \underbrace{\frac{\partial V}{\partial t}\Delta t}_{\substack{\text{Change of } V\\ \text{with time}\\ \text{(unsteady flow)}}} + \underbrace{\frac{\partial V}{\partial x}\Delta x}_{\substack{\text{Change of } V\\ \text{with position}\\ \text{(nonuniform flow)}}}$$

Since Δx is the distance covered in the time Δt, the acceleration of the particle becomes

$$a = \lim_{\Delta t \to 0} \frac{\Delta V}{\Delta t} = \frac{\partial V}{\partial t} + \frac{\partial V}{\partial x}\frac{dx}{dt}$$

Another way to determine this result is to use the *chain rule* of calculus. Since $V = V(x, t)$, we must use partial derivatives to find each change in V, that is, $a = \partial V/\partial t(dt/dt) + \partial V/\partial x(dx/dt)$. Realizing that $dx/dt = V$, in either case, we get

$$\boxed{a = \frac{DV}{Dt} = \underbrace{\frac{\partial V}{\partial t}}_{\substack{\text{Local}\\ \text{accel.}}} + \underbrace{V\frac{\partial V}{\partial x}}_{\substack{\text{Convective}\\ \text{accel.}}}} \qquad (3\text{–}4)$$

The notation $D(\,)/Dt$ is referred to as a ***material derivative,*** because it gives the time rate of change of a fluid property (in this case the velocity) as the fluid particle (material) passes through the control volume. Let's summarize our result.

16/04/2020 ☆

Pág. 146 (170)

3.3 Graphical Descriptions of Fluid Flow

Streamline is a curve that is drawn through the fluid in such a manner that it indicates the direction of the velocity of particles located on it at a particular instant of time. (rule)

(function)
↑↑↑
my equation

☆ The velocity of any particle is always tangent to the streamline along which it traveling at that instant.

Consequently no fluid can flow across a streamline, only along the streamline.

Stagnation point because here the velocity of any particle is momentarily reduced to zero when it strikes the cylinder's surface.

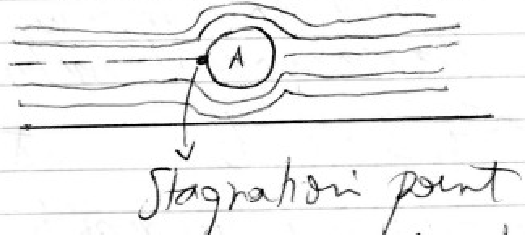

Stagnation point

☆

obs: Sometimes the streamlines can be a function of both space and time.
(Einstein).

data 19 04 2020

In the case of two-dimensional flow, the particle's velocity will have two components,

u in the x direction and

v in the y direction.

Fig. 3-7

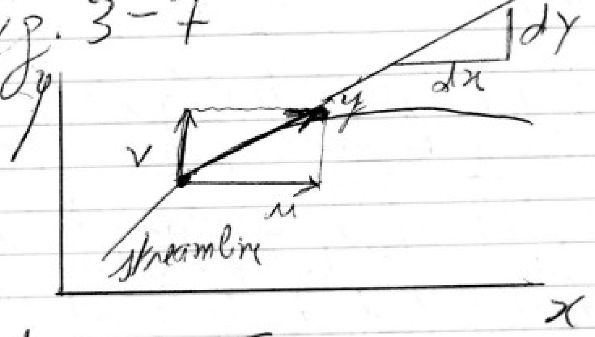

$$\frac{dy}{dx} = \frac{v}{u}$$

Integrating this equation will give the equation of the streamlines

$$Y = f(x) + \underset{\hookrightarrow \text{constant.}}{C}$$

IMPORTANT:

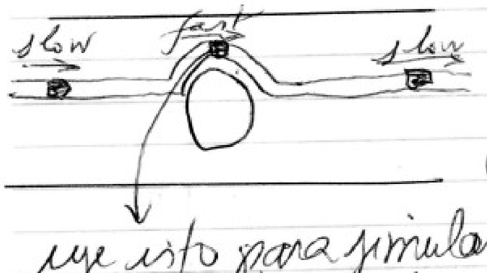

fluid flows faster within a narrow streamtube.

Fig. 3-6 ⓑ.

use isto para simular o desvio da luz quando passa perto de uma estrela como o Sol → Einstein - Teory of Relativity.

& '/
)*+,)))-) '/ 4 ? 91
% 4! H I 1 0!
('

2

KINEMATICS OF PARTICLES

CHAPTER OUTLINE

2/1 Introduction
2/2 Rectilinear Motion
2/3 Plane Curvilinear Motion
2/4 Rectangular Coordinates (x-y)
2/5 Normal and Tangential Coordinates (n-t)
2/6 Polar Coordinates (r-θ)
2/7 Space Curvilinear Motion
2/8 Relative Motion (Translating Axes)
2/9 Constrained Motion of Connected Particles
2/10 Chapter Review

2/1 INTRODUCTION

Kinematics is the branch of dynamics which describes the motion of bodies without reference to the forces which either cause the motion or are generated as a result of the motion. Kinematics is often described as the "geometry of motion." Some engineering applications of kinematics include the design of cams, gears, linkages, and other machine elements to control or produce certain desired motions, and the calculation of flight trajectories for aircraft, rockets, and spacecraft. A thorough working knowledge of kinematics is a prerequisite to kinetics, which is the study of the relationships between motion and the corresponding forces which cause or accompany the motion.

Particle Motion

We begin our study of kinematics by first discussing in this chapter the motions of points or particles. A particle is a body whose physical dimensions are so small compared with the radius of curvature of its path that we may treat the motion of the particle as that of a point. For example, the wingspan of a jet transport flying between Los Angeles and New York is of no consequence compared with the radius of curvature of its flight path, and thus the treatment of the airplane as a particle or point is an acceptable approximation.

We can describe the motion of a particle in a number of ways, and the choice of the most convenient or appropriate way depends a great deal on experience and on how the data are given. Let us obtain an overview of the several methods developed in this chapter by referring to Fig. 2/1, which shows a particle P moving along some general path in space. If the particle is confined to a specified path, as with a bead sliding along a fixed wire, its motion is said to be *constrained*. If there are no physical guides, the motion is said to be *unconstrained*. A small rock tied to the end of a string and whirled in a circle undergoes constrained motion until the string breaks, after which instant its motion is unconstrained.

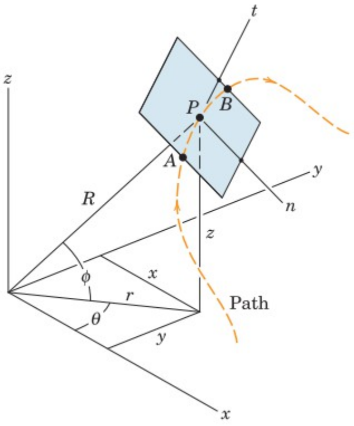

Figure 2/1

Choice of Coordinates

The position of particle P at any time t can be described by specifying its rectangular coordinates* x, y, z, its cylindrical coordinates r, θ, z, or its spherical coordinates R, θ, ϕ. The motion of P can also be described by measurements along the tangent t and normal n to the curve. The direction of n lies in the local plane of the curve.† These last two measurements are called *path variables*.

The motion of particles (or rigid bodies) can be described by using coordinates measured from fixed reference axes (*absolute-motion* analysis) or by using coordinates measured from moving reference axes (*relative-motion* analysis). Both descriptions will be developed and applied in the articles which follow.

With this conceptual picture of the description of particle motion in mind, we restrict our attention in the first part of this chapter to the case of *plane motion* where all movement occurs in or can be represented as occurring in a single plane. A large proportion of the motions of machines and structures in engineering can be represented as plane motion. Later, in Chapter 7, an introduction to three-dimensional motion is presented. We begin our discussion of plane motion with *rectilinear motion*, which is motion along a straight line, and follow it with a description of motion along a plane curve.

2/2 RECTILINEAR MOTION

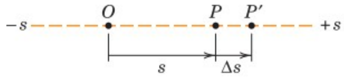

Figure 2/2

Consider a particle P moving along a straight line, Fig. 2/2. The position of P at any instant of time t can be specified by its distance s measured from some convenient reference point O fixed on the line. At time $t + \Delta t$ the particle has moved to P' and its coordinate becomes $s + \Delta s$. The change in the position coordinate during the interval Δt is called the *displacement* Δs of the particle. The displacement would be negative if the particle moved in the negative s-direction.

---

*Often called *Cartesian* coordinates, named after René Descartes (1596–1650), a French mathematician who was one of the inventors of analytic geometry.

†This plane is called the *osculating* plane, which comes from the Latin word *osculari* meaning "to kiss." The plane which contains P and the two points A and B, one on either side of P, becomes the osculating plane as the distances between the points approach zero.

date 20.04.2020
X

147) 20200420

A) Books_70_2020/
Engineering Mechanics Dynamics/
J.L. Meriam ... Engineering Mechanics.
pdf

Pag. 22(41)

Figure 2/1 Choice of Coordinates

rectangular coordinates x, y, z

its cylindrical coordinates r, θ, z, or

its spherical coordinates R, θ, ϕ.

the motion of P can also be described by measurements along the tangent t and the normal n to the curve.

the direction of n lies in the local plane of the curve.

these last two measurements are called path variables.

Motion Particles { absolute-motion analysis
{ relative-motion analysis

this plane is called the osculating plane, which the plane contains P and two points A and B, one on either side of P, becomes the osculating plane as the distances between the points approach zero.

These rules are the same as for the differentiation of scalar quantities, except for the case of the cross product where the order of the terms must be preserved. These rules are covered in Art. C/7 of Appendix C and should be reviewed at this point.

Three different coordinate systems are commonly used for describing the vector relationships for curvilinear motion of a particle in a plane: rectangular coordinates, normal and tangential coordinates, and polar coordinates. An important lesson to be learned from the study of these coordinate systems is the proper choice of a reference system for a given problem. This choice is usually revealed by the manner in which the motion is generated or by the form in which the data are specified. Each of the three coordinate systems will now be developed and illustrated.

2/4 RECTANGULAR COORDINATES (x-y)

This system of coordinates is particularly useful for describing motions where the x- and y-components of acceleration are independently generated or determined. The resulting curvilinear motion is then obtained by a vector combination of the x- and y-components of the position vector, the velocity, and the acceleration.

Vector Representation

The particle path of Fig. 2/5 is shown again in Fig. 2/7 along with x- and y-axes. The position vector $\mathbf{r}$, the velocity $\mathbf{v}$, and the acceleration $\mathbf{a}$ of the particle as developed in Art. 2/3 are represented in Fig. 2/7 together with their x- and y-components. With the aid of the unit vectors $\mathbf{i}$ and $\mathbf{j}$, we can write the vectors $\mathbf{r}$, $\mathbf{v}$, and $\mathbf{a}$ in terms of their x- and y-components. Thus,

$$\boxed{\begin{aligned}\mathbf{r} &= x\mathbf{i} + y\mathbf{j}\\ \mathbf{v} = \dot{\mathbf{r}} &= \dot{x}\mathbf{i} + \dot{y}\mathbf{j}\\ \mathbf{a} = \dot{\mathbf{v}} = \ddot{\mathbf{r}} &= \ddot{x}\mathbf{i} + \ddot{y}\mathbf{j}\end{aligned}} \qquad \mathbf{(2/6)}$$

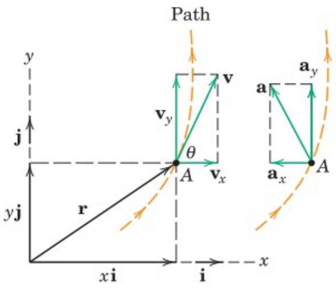

Figure 2/7

As we differentiate with respect to time, we observe that the time derivatives of the unit vectors are zero because their magnitudes and directions remain constant. The scalar values of the components of $\mathbf{v}$ and $\mathbf{a}$ are merely $v_x = \dot{x}$, $v_y = \dot{y}$ and $a_x = \dot{v}_x = \ddot{x}$, $a_y = \dot{v}_y = \ddot{y}$. (As drawn in Fig. 2/7, a_x is in the negative x-direction, so that $\ddot{x}$ would be a negative number.)

As observed previously, the direction of the velocity is always tangent to the path, and from the figure it is clear that

$$v^2 = {v_x}^2 + {v_y}^2 \qquad v = \sqrt{{v_x}^2 + {v_y}^2} \qquad \tan\theta = \frac{v_y}{v_x}$$

$$a^2 = {a_x}^2 + {a_y}^2 \qquad a = \sqrt{{a_x}^2 + {a_y}^2}$$

If the angle θ is measured counterclockwise from the x-axis to $\mathbf{v}$ for the configuration of axes shown, then we can also observe that $dy/dx = \tan\theta = v_y/v_x$.

If the coordinates x and y are known independently as functions of time, $x = f_1(t)$ and $y = f_2(t)$, then for any value of the time we can combine them to obtain $\mathbf{r}$. Similarly, we combine their first derivatives $\dot{x}$ and $\dot{y}$ to obtain $\mathbf{v}$ and their second derivatives $\ddot{x}$ and $\ddot{y}$ to obtain $\mathbf{a}$. On the other hand, if the acceleration components a_x and a_y are given as functions of the time, we can integrate each one separately with respect to time, once to obtain v_x and v_y and again to obtain $x = f_1(t)$ and $y = f_2(t)$. Elimination of the time t between these last two parametric equations gives the equation of the curved path $y = f(x)$.

From the foregoing discussion we can see that the rectangular-coordinate representation of curvilinear motion is merely the superposition of the components of two simultaneous rectilinear motions in the x- and y-directions. Therefore, everything covered in Art. 2/2 on rectilinear motion can be applied separately to the x-motion and to the y-motion.

Projectile Motion

An important application of two-dimensional kinematic theory is the problem of projectile motion. For a first treatment of the subject, we neglect aerodynamic drag and the curvature and rotation of the earth, and we assume that the altitude change is small enough so that the acceleration due to gravity can be considered constant. With these assumptions, rectangular coordinates are useful for the trajectory analysis.

For the axes shown in Fig. 2/8, the acceleration components are

$$a_x = 0 \qquad a_y = -g$$

Integration of these accelerations follows the results obtained previously in Art. 2/2*a* for constant acceleration and yields

$$\begin{aligned} v_x &= (v_x)_0 & v_y &= (v_y)_0 - gt \\ x &= x_0 + (v_x)_0 t & y &= y_0 + (v_y)_0 t - \tfrac{1}{2}gt^2 \end{aligned}$$

$$v_y{}^2 = (v_y)_0{}^2 - 2g(y - y_0)$$

In all these expressions, the subscript zero denotes initial conditions, frequently taken as those at launch where, for the case illustrated,

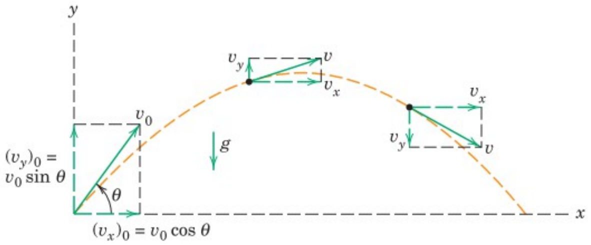

Figure 2/8

$x_0 = y_0 = 0$. Note that the quantity g is taken to be positive throughout this text.

We can see that the x- and y-motions are independent for the simple projectile conditions under consideration. Elimination of the time t between the x- and y-displacement equations shows the path to be parabolic (see Sample Problem 2/6). If we were to introduce a drag force which depends on the speed squared (for example), then the x- and y-motions would be coupled (interdependent), and the trajectory would be nonparabolic.

When the projectile motion involves large velocities and high altitudes, to obtain accurate results we must account for the shape of the projectile, the variation of g with altitude, the variation of the air density with altitude, and the rotation of the earth. These factors introduce considerable complexity into the motion equations, and numerical integration of the acceleration equations is usually necessary.

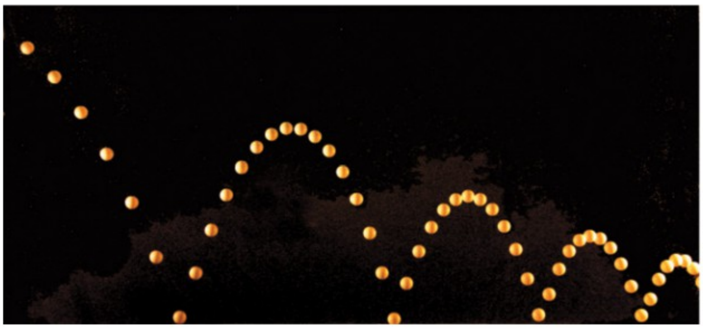

Andrew Davidhazy

This stroboscopic photograph of a bouncing ping-pong ball suggests not only the parabolic nature of the path, but also the fact that the speed is lower near the apex.

20.04.2020

Pág. 45(64)

2/4 Rectangular Coordinates

$$r = xi + yj$$

$$v = \dot{r} = \dot{x}i + \dot{y}j \quad \text{Figure 2/7}$$

$$a = \dot{v} = \ddot{r} = \ddot{x}i + \ddot{y}j$$

B) Books_70_2020/
Mathematics/Higher Level for the IB Diploma/
Mathematics for IB Diploma ... pdf
Pág. 389(400) Obs.: Todo livro.

13C Distances

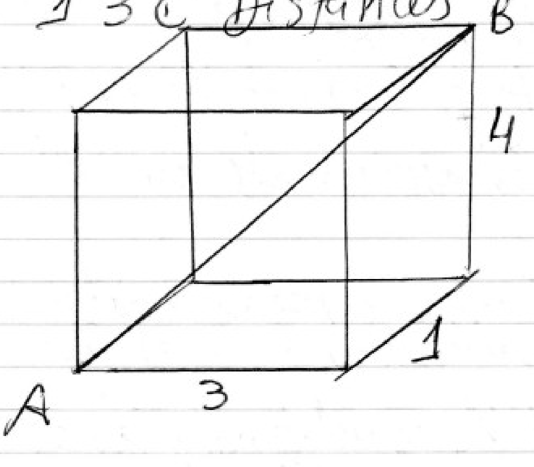

$$\vec{AB} = \begin{pmatrix} 3 \\ 1 \\ 4 \end{pmatrix}$$

$$AB = \sqrt{3^2+1^2+4^2} = \sqrt{26}$$

this quantity is called the MAGNITUDE of $\vec{AB}$, and written as $|\vec{AB}|$.

The magnitude of a vector, $a = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}$, is $|a| = \sqrt{a_1^2 + a_2^2 + a_3^2}$.

The distance between points with position vector a and b is $|b-a|$

13C Distances

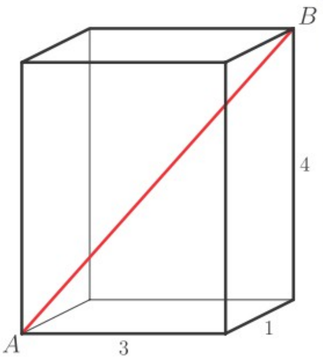

Geometry problems often involve finding distances between points. In this section we will see how to use vectors to do this.

Consider two points, A and B such that the displacement $\overrightarrow{AB} = \begin{pmatrix} 3 \\ 1 \\ 4 \end{pmatrix}$. The distance AB can be found by using Pythagoras' theorem in three dimensions: $AB = \sqrt{3^2 + 1^2 + 4^2} = \sqrt{26}$. This quantity is called the **magnitude** of $\overrightarrow{AB}$, and written as $\left|\overrightarrow{AB}\right|$.

To find the distance between A and B, using their position vectors, we first need to find the displacement vector $\overrightarrow{AB}$ and then calculate its magnitude.

Worked example 13.6

Points A and B have position vectors $\boldsymbol{a} = \begin{pmatrix} 2 \\ -1 \\ 5 \end{pmatrix}$ and $\boldsymbol{b} = \begin{pmatrix} 5 \\ 2 \\ 3 \end{pmatrix}$. Find the exact distance AB.

The distance is the magnitude of the displacement vector, so find $\overrightarrow{AB}$ first

$$\overrightarrow{AB} = \underline{b} - \underline{a}$$

$$= \begin{pmatrix} 5 \\ 2 \\ 3 \end{pmatrix} - \begin{pmatrix} 2 \\ -1 \\ 5 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \\ -2 \end{pmatrix}$$

Now use the formula for the magnitude

$$\left|\overrightarrow{AB}\right| = \sqrt{3^2 + 3^2 + 2^2} = \sqrt{22}$$

KEY POINT 13.4

The magnitude of a vector, $\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}$, is $|\boldsymbol{a}| = \sqrt{a_1^2 + a_2^2 + a_3^2}$.

The distance between points with position vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is $|\boldsymbol{b} - \boldsymbol{a}|$.

EXAM HINT

Don't forget that squaring a negative number gives a positive value.

The symbol $\gtreqless$ means 'greater than, equal to or less than'. This may appear to be a useless symbol, but it highlights an important idea in vectors – they cannot be put into order. So while it is correct to say that $|v| \gtreqless |u|$ it is not possible to say the same about the vectors themselves.

© Cambridge University Press 2012

Not for printing, sharing or distribution.

We saw in Section 13B that multiplying a vector by a scalar produces a vector in the same direction but of different magnitude. In more advanced applications of vectors it is useful to be able to use vectors of length 1, called **unit vectors**. The base vectors ***i***, ***j*** and ***k*** are examples of unit vectors.

Worked example 13.7

(a) Find the unit vector in the same direction as $\boldsymbol{a} = \begin{pmatrix} 2 \\ -2 \\ 1 \end{pmatrix}$.

(b) Find a vector of magnitude 5 parallel to ***a***.

(a) Call the required unit vector $\hat{a}$.

To produce a vector in the same direction but of different magnitude as ***a***, we need to multiply ***a*** by a scalar. We need to find the value of the scalar

Then $\hat{a} = k\hat{a}$ and $\hat{a} = 1$

$|k\underline{a}| = k|\underline{a}| = 1$

$\Rightarrow k = \dfrac{1}{|\underline{a}|}$

$|\underline{a}| = \sqrt{2^2 + 2^2 + 1^2} = 3$

$\therefore k = \dfrac{1}{3}$

Find the vector $\hat{\boldsymbol{a}}$

The unit vector is

$$\hat{a} = \frac{1}{3}\begin{pmatrix} 2 \\ -2 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{2}{3} \\ \frac{-2}{3} \\ \frac{1}{3} \end{pmatrix}$$

To get a vector of magnitude 5 we need to multiply the unit vector by 5

(b) Let $\underline{b}$ be parallel to $\underline{a}$ and $|\underline{b}| = 5$

Then $\underline{b} = 5\hat{a}$

$$\therefore \underline{b} = \begin{pmatrix} \frac{10}{3} \\ \frac{-10}{3} \\ \frac{5}{3} \end{pmatrix}$$

EXAM HINT

Note that part (b) has two possible answers, as ***b*** could be in the opposite direction. To get the second answer we would take the scalar to be –5 instead of 5.

© Cambridge University Press 2012

Not for printing, sharing or distribution.

The last example showed the general method for finding the unit vector in a given direction.

KEY POINT 13.5

The unit vector in the same direction as $\boldsymbol{a}$ is $\hat{\boldsymbol{a}} = \dfrac{1}{|\boldsymbol{a}|}\boldsymbol{a}$.

Exercise 13C

1. Find the magnitude of the following vectors in two dimensions.

$\boldsymbol{a} = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$ $\quad \boldsymbol{b} = \begin{pmatrix} -1 \\ 5 \end{pmatrix}$ $\quad \boldsymbol{c} = 2\boldsymbol{i} - 4\boldsymbol{j}$ $\quad \boldsymbol{d} = -\boldsymbol{i} + \boldsymbol{j}$

2. Find the magnitude of the following vectors in three dimensions.

$\boldsymbol{a} = \begin{pmatrix} 4 \\ 1 \\ 2 \end{pmatrix}$ $\quad \boldsymbol{b} = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$ $\quad \boldsymbol{c} = 2\boldsymbol{i} - 4\boldsymbol{j} + \boldsymbol{k}$ $\quad \boldsymbol{d} = \boldsymbol{j} - \boldsymbol{k}$

3. Find the distance between the following pairs of points in the plane.

(a) (i) $A\,(1, 2)$ and $B\,(3, 7)$ (ii) $C\,(2, 1)$ and $D\,(1, 2)$

(b) (i) $P\,(-1, -5)$ and $Q\,(-4, 2)$ (ii) $M\,(1, 0)$ and $N\,(0, -2)$

4. Find the distance between the following pairs of points in three dimensions.

(a) (i) $A\,(1, 0, 2)$ and $B\,(2, 3, 5)$

(ii) $C\,(2, 1, 7)$ and $D\,(1, 2, 1)$

(b) (i) $P\,(3, -1, -5)$ and $Q\,(-1, -4, 2)$

(ii) $M\,(0, 0, 2)$ and $N\,(0, -3, 0)$

5. Find the distance between the points with the given position vectors.

(a) $\boldsymbol{a} = 2\boldsymbol{i} + 4\boldsymbol{j} - 2\boldsymbol{k}$ and $\boldsymbol{b} = \boldsymbol{i} - 2\boldsymbol{j} - 6\boldsymbol{k}$

(b) $\boldsymbol{a} = \begin{pmatrix} 3 \\ 7 \\ -2 \end{pmatrix}$ and $\boldsymbol{b} = \begin{pmatrix} 1 \\ -2 \\ -5 \end{pmatrix}$

© Cambridge University Press 2012

Not for printing, sharing or distribution.

20.04.2020

Pág. 45(64)

2/4 Rectangular Coordinates

$$r = xi + yj$$

$$v = \dot{r} = \dot{x}i + \dot{y}j$$ Figure 2/7

$$a = \dot{v} = \ddot{r} = \ddot{x}i + \ddot{y}j$$

B) Books_70_2020/
Mathematics/Higher Level for the IB Diploma/
Mathematics for IB Diploma ... pdf

Pág. 389(400) Obs.: Todo livro.

13C Distances

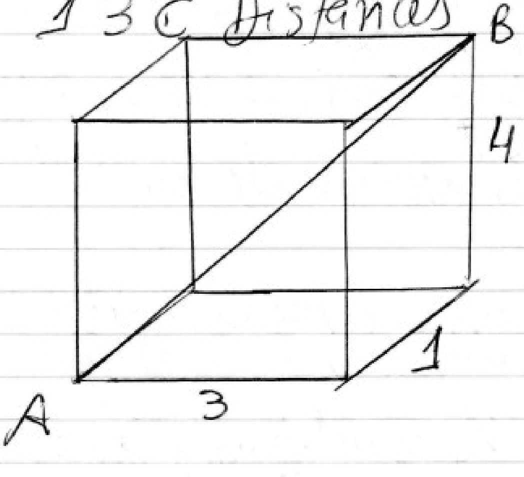

$$\overline{AB} = \begin{pmatrix} 3 \\ 1 \\ 4 \end{pmatrix}$$

$$AB = \sqrt{3^2+1^2+4^2} = \sqrt{26}$$

this quantity is called the MAGNITUDE of $\overline{AB}$, and written as $|\overline{AB}|$.

The magnitude of a vector, $a = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}$, is
$|a| = \sqrt{a_1^2 + a_2^2 + a_3^2}$.

The distance between points with position vector a and b is $|b-a|$

)*+,)))-) '/ ## .% 0!
('.. ! 6 ! 3
1 2

On a non-calculator paper we would proceed as follows.
From the trigonometrical identities we have:

$\cos 4x = \cos^2 2x - \sin^2 2x$

and $\cos^2 2x + \sin^2 2x = 1$,

giving $\cos^2 2x = 1 - \sin^2 2x$

So $\cos 4x = 1 - \sin^2 2x - \sin^2 2x$

$\Rightarrow \cos 4x = 1 - 2\sin^2 2x$

$\Rightarrow \sin^2 2x = \dfrac{1 - \cos 4x}{2}$

Hence

$$V = \pi\int_0^{\pi}\left(\frac{1 - \cos 4x}{2}\right)\mathrm{d}x$$

$$\Rightarrow V = \frac{\pi}{2}\int_0^{\pi}(1 - \cos 4x)\,\mathrm{d}x$$

$$\Rightarrow V = \frac{\pi}{2}\left[x - \frac{\sin 4x}{4}\right]_0^{\pi}$$

$$\Rightarrow V = \frac{\pi}{2}\left[\left(\pi - \frac{\sin 4\pi}{4}\right) - \left(0 - \frac{\sin 0}{4}\right)\right]$$

$$\Rightarrow V = \frac{\pi}{2}[(\pi - 0) - (0 - 0)]$$

$$\Rightarrow V = \frac{\pi^2}{2}\text{ units}^3$$

It is possible to rotate curves around a variety of different lines, but for the purposes of this syllabus it is only necessary to know how to find the volumes of solids of revolution formed when rotated about the *x*- or the *y*-axes.

Volumes of solids of revolution when rotated about the *y*-axis

The method is identical to finding the volume of the solid formed when rotating about the *x*-axis. Consider the curve $y = x^2$.

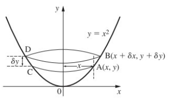

Look at the element ABCD where A is on the curve and has coordinates (x, y). If $y = x^2$, then $x = y^{\frac{1}{2}}$. Hence the coordinates of A are $(y^{\frac{1}{2}}, y)$. ABCD is approximately a cylinder with radius *x* and whose "height" is δy.

Therefore the volume of ABCD $\approx \pi x^2\,\delta y$

and the volume *V* of the entire solid $\approx \displaystyle\sum_{y=a}^{y=b}\pi x^2\,\delta y$.

The smaller δy becomes, the closer this approximation is to *V*,

i.e. $V = \lim_{\delta y \to 0}\displaystyle\sum_{y=a}^{y=b}\pi x^2\,\delta y$

$$V = \pi\int_a^b x^2\,\mathrm{d}y$$

This is the formula for a full revolution about the *y*-axis.

In this case $x = y^{\frac{1}{2}}$. To find the volume of the solid formed when the part of the curve between $y = 0$ and $y = 2$ is rotated about the *y*-axis, we proceed as follows.

$$V = \pi\int_0^2 x^2\,\mathrm{d}y$$

$$\Rightarrow V = \pi\int_0^2 y\,\mathrm{d}y$$

$$\Rightarrow V = \pi\left[\frac{y^2}{2}\right]_0^2$$

$$\Rightarrow V = \pi\left[\frac{4}{2} - 0\right]$$

$$\Rightarrow V = 2\pi$$

Example

Find the volume of the solid of revolution formed when the area bounded by the curve $xy = 2$ and the lines $x = 0, y = 3, y = 6$ is rotated about the *y*-axis.

As before, plotting the curve on a calculator or drawing a diagram first is a good idea.

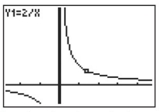

As the rotation is taking place about the *y*-axis, the required formula is:

$$V = \pi\int_3^6 x^2\,\mathrm{d}y$$

Since $xy = 2$, then $x = \dfrac{2}{y}$

So we have:

$$V = \pi\int_3^6\left(\frac{2}{y}\right)^2\mathrm{d}y$$

$$\Rightarrow V = \pi\int_3^6\frac{4}{y^2}\,\mathrm{d}y$$

$$\Rightarrow V = \pi\int_3^6 4y^{-2}\,\mathrm{d}y$$

$$\Rightarrow V = \pi\left[\frac{4y^{-1}}{-1}\right]_3^6$$

$$\Rightarrow V = \pi\left[\left(\frac{-4}{6}\right) - \left(\frac{-4}{3}\right)\right]$$

$$\Rightarrow V = \frac{2\pi}{3}$$

The calculator display is shown below.

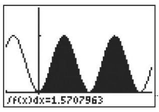

```
∫f(x)dx=1.5707963
```

Remember that this answer will need to be multiplied by π.

Hence $V = 2.09$ units^3.

Up until now it appears that volumes of solids of revolution are a theoretical application of integration. However, this is not the case. In the field of computer-aided design, volumes of solids of revolution are important. If, for example, we wanted to design a wine glass, then we could rotate the curve $y = x^2$ around the *y*-axis to give a possible shape. If we wanted a thinner wine glass, then we could rotate the curve $y = x^4$. Because this can be all modelled on a computer, and a three-dimensional graphic produced, designers can work out the shape that they want.

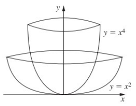

Exercise 6

1 Find the volumes generated when the following areas are rotated through 2π radians about the *x*-axis.

- **a** The area bounded by the curve $y = 3x + 2$, the *x*-axis, the *y*-axis and the line $x = 2$.
- **b** The area bounded by the curve $y = 4x - x^2$ and the *x*-axis.
- **c** The area bounded by the curve $y = x^3$, the *x*-axis and the line $x = 2$.
- **d** The area bounded by the curve $y = 1 + \sqrt{x}$, the *x*-axis, the *y*-axis and the line $x = 1$.
- **e** The area bounded by the curve $y = x^2 - 1$, the *x*-axis and the line $x = 3$.
- **f** The area bounded by the curve $y = \frac{1}{3}(2x - 1)^2$, the *x*-axis and the line $x = 5$.
- **g** The area bounded by the curve $y = 9x - x^2 - 14$ and the *x*-axis.
- **h** The area bounded by the curve $y = \sin 4x$, the *x*-axis and the lines $x = 0$ and $x = \frac{\pi}{4}$.
- **i** The area bounded by the curve $y = \tan 2x$, the *x*-axis and line $x = \frac{\pi}{6}$.
- **j** The area bounded by the curve $y = \mathrm{e}^{2x}\sin x$, the *x*-axis and the lines $x = 0.5$ and $x = 1.5$.

2 Find the volumes generated when the following areas are rotated through 2π radians about the *y*-axis.

- **a** The area bounded by the curve $y = 4 - x^2$ and the *x*-axis.
- **b** The area bounded by the curve $y = x^3$, the *y*-axis and the line $y = 2$.
- **c** The area bounded by the curve $y = \mathrm{e}^x$, the *y*-axis and the line $y = 2$.
- **d** The area bounded by the curve $y = \sin x$, the *y*-axis and the lines $y = 0.2$ and $y = 0.8$.
- **e** The area bounded by the curve $y = x^2 - 4x$, the *y*-axis, the *x*-axis and the line $y = 2$.
- **f** The area bounded by the curve $y = \sin^{-1} x$, the *y*-axis and the line $y = \frac{\pi}{3}$.
- **g** The area bounded by the curve $y = \ln(x + 1)$, the *y*-axis and the line $y = 1.5$.

3 Find a general formula for the volume generated when the area bounded by the curve $y = x^2$, the x axis and the line $x = a$ is rotated through 2π radians about the *x*-axis.

4 Find the volume obtained when the region bounded by the curve $y = 3 + \frac{4}{x}$, the *x*-axis and the lines $x = 3$ and $x = 6$ is rotated through 360° about the *x*-axis.

5 Consider the curve $y = x^{\frac{1}{2}}$. The part of the curve between $y = 2$ and $y = 5$ is rotated through 2π radians about the *y*-axis. Find the volume of the solid of revolution formed.

6 Find the volume generated when the area bounded by the curve $y = \ln x$, the *x*-axis, the *y*-axis and the line $y = 2$ is rotated through 360° about the *y*-axis.

7 Consider the curve $y = \frac{1}{5}x^2$. A volume is formed by revolving this curve through 360° about the *y*-axis. The radius of the rim of this volume is 5 cm. Find the depth of the shape and its volume.

8 Sketch the curve $y = |x^2 - 1|$ and shade the area that is bounded by the curve and the *x*-axis. This area is rotated through 2π radians about the *x*-axis. Find the volume generated. What is the volume when it is rotated through 2π radians about the *y*-axis?

9 The parabola $y = 8x^2$ is rotated through 360° about its axis of symmetry, thus forming a volume of solid of revolution. Calculate the volume enclosed between this surface and a plane perpendicular to the *y*-axis. This plane is a distance of 7 units from the origin.

10 Find the volume of the solid of revolution formed when the area included between the *x*-axis and one wavelength of the curve $y = b\sin\frac{x}{a}$ is rotated through 360° about the *x*-axis.

11 Consider the part curve $y^2 = x^2 \sin x$ which lies between $x = (n - 1)\pi$ and $x = (n + 1)\pi$ where *n* is an integer. Find the volume generated when this area is rotated through 2π radians about the *x*-axis.

12 **a** Using the substitution $x = \sin\theta$, evaluate $\int \sqrt{1 - x^2}\,\mathrm{d}x$.

b Hence or otherwise, find the volume generated when the area bounded by the curve $y^2 = 1 - x^4$, the *y*-axis and the line $y = a$, $0 < a < 1$, is rotated through 2π radians about the *y*-axis.

Review exercise

1 The region *A* is bounded by the curve $y = \sin\left(2x + \frac{\pi}{3}\right)$ and by the lines $x = 0$ and $x = \frac{\pi}{6}$. Find the exact value of the volume formed when the area *A* is rotated fully about the *x*-axis.

2 Solve the differential equation $xy\frac{\mathrm{d}y}{\mathrm{d}x} = 1 + y^2$, given that $y = 0$ when $x = 2$.

[IB Nov 00 P1 Q17]

3 A particle moves in a straight line with velocity, in metres per second, at time *t* seconds, given by $v(t) = 6t^2 - 6t$, $t \geq 0$.

Calculate the total distance travelled by the particle in the first two seconds of motion. [IB Nov 02 P1 Q11]

4 Consider the region bounded by the curve $y = \mathrm{e}^{-2x}$, the *x*-axis and the lines $x = \pm a$. Find in terms of *a*, the volume of the solid generated when this region is rotated through 2π radians about the *x*-axis.

5 Solve the differential equation $\frac{\mathrm{d}y}{\mathrm{d}x} = 5xy$ and sketch one of the solution curves which does not pass through $y = 0$.

6 The acceleration of a body is given in terms of the displacement *s* metres as $a = \frac{3s}{s^2 + 1}$. Determine a formula for the velocity as a function of the displacement given that when $s = 1$ m, $v = 2$ ms^{-1}. Hence find the exact velocity when the body has travelled 5 m.

7 The temperature *T*°C of an object in a room, after *t* minutes, satisfies the differential equation $\frac{\mathrm{d}T}{\mathrm{d}t} = k(T - 22)$ where *k* is a constant.

a Solve this equation to show that $T = A\mathrm{e}^{kt} + 22$ where *A* is a constant.
b When $t = 0, T = 100$ and when $t = 15, T = 70$.
  i Use this information to find the value of *A* and of *k*.
  ii Hence find the value of *t* when $T = 40$. [IB May 04 P1 Q4]

8 The velocity of a particle is given by the formula $v = \frac{1}{t^2\sqrt{t - 1}}$ for $t > 1$. Using the substitution $t = \sec^2\theta$, find the displacement travelled between $t = 2$ seconds and $t = T$ seconds.

9 Consider the curve $y = -\frac{r}{h}x + r$. The triangular region of this curve which occupies the first quadrant is rotated fully about the *x*-axis. Show that the volume of the cone formed is $\frac{1}{3}\pi r^2 h$.

10 A sample of radioactive material decays at a rate which is proportional to the amount of material present in the sample. Find the half-life of the material if 50 grams decay to 48 grams in 10 years. [IB Nov 01 P1 Q19]

11 The acceleration in ms^{-2} of a particle moving in a straight line at time *t*, is given by the formula $a = \sin\frac{2\pi}{3}t$. The particle starts from rest from a point where its displacement is 0.5 m from a fixed point O on the line.

a Find the velocity and displacement of the particle from O at any time *t*.
b Find the time that elapses before the particle comes to rest again.

12 **a** Let $y = \sin(kx) - kx\cos(kx)$, where *k* is a constant.

Show that $\frac{\mathrm{d}y}{\mathrm{d}x} = k^2 x \sin(kx)$.

A particle is moving along a straight line so that *t* seconds after passing through a fixed point O on the line, its velocity $v(t)$ ms^{-1} is given by $v(t) = t\sin\left(\frac{\pi}{3}t\right)$.

b Find the values of *t* for which $v(t) = 0$, given that $0 \leq t \leq 6$.
c **i** Write down a mathematical expression for the **total** distance travelled by the particle in the first six seconds after passing through O.
  ii Find this distance. [IB Nov 01 P2 Q2]

13 When air is released from an inflated balloon it is found that the rate of decrease of the volume of the balloon is proportional to the volume of the balloon. This can be represented by the differential equation $\frac{\mathrm{d}v}{\mathrm{d}t} = -kv$, where *v* is the volume, *t* is the time and *k* is the constant of proportionality.

a If the initial volume of the balloon is v_0, find an expression, in terms of *k*, for the volume of the balloon at time *t*.
b Find an expression, in terms of *k*, for the time when the volume is $\frac{v_0}{2}$.

[IB May 99 P1 Q19]

14 Show by means of the substitution $x = \tan\theta$ that

$\int_0^1 \frac{1}{(x^2 + 1)^2}\mathrm{d}x = \int_0^{\frac{\pi}{4}} \cos^2\theta\,\mathrm{d}\theta$. Hence find the exact value of the volume formed when the curve $y = \frac{1}{x^2 + 1}$ bounded by the lines $x = 0$ and $x = 1$ is rotated fully about the *x*-axis.

15 Consider the curve $y^2 = 9a(4a - x)$.

a Sketch the part of the curve that lies in the first quadrant.
b Find the exact value of the volume V_x when this part of the curve is rotated through 360° about the *x*-axis.
c Show that $\frac{V_y}{V_x} = \frac{a}{b}$ where V_y is the volume generated when the curve is rotated fully about the *y*-axis and *a* and *b* are integers.

!)*+,)))-) '/ < H - ; 3 @
+ B 0 4 .H I 1 0!
)*+,)))-) '/ < H
- ; 3 @ 0 1 + B 0 4 # 4! /H
J I 1 K H & 0!
. $ ' $

temperature variations. This introduces a major uncertainty in calculations of lift and drag. Aerodynamic forces are also affected by the flow and the pressure distribution of the rocket exhaust gases, as explained in Chapter 20.

A vehicle's flight regime in the neighborhood of Mach 1 is called the transonic flight region. Here strong unsteady aerodynamic forces often develop (due to shock-induced buffeting), which are noticeable in the steep rise and subsequent decrease of the coefficients as shown in Fig. 4–4. In some cases a vehicle's maximum load capabilities during transonic flight have been exceeded leading to structural failures.

For properly designed space launch vehicles and ballistic missiles the integrated drag losses, when expressed in terms of Δu, are typically 5 to 10% of the final ideal vehicle velocity. These relatively low values result from air densities (and thus dynamic pressures), being low at high altitudes where velocities are high, and being high at altitudes where vehicle velocities are low.

Gravitational attraction is exerted upon any flying space vehicle by all planets, stars, and moons. Gravity forces pull the vehicle in the direction of the center of mass of the attracting body. Within the immediate vicinity of the Earth, the attraction of other planets and celestial bodies (like our sun and moon) can be negligibly small compared to the Earth's gravitational force. This force or gravitational pull makes up the object's *weight*.

When Earth gravity variations with the geographical features and the oblate shape are neglected, the acceleration of gravity g varies inversely as the square of the distance from the Earth's center. If R_0 is the mean radius at the (spherical) Earth's surface at which g_0 is the acceleration of gravity, by Newton's law of gravitation, g changes with altitude (h) as

$$\begin{aligned} g &= g_0(R_0/R)^2 \\ &= g_0[R_0/(R_0 + h)]^2 \end{aligned} \tag{4–12}$$

At the equator the spherical Earth's radius is 6378.388 km and the standard value of g_0 is 9.80665 m/sec^2. At distances as far away as the moon, the Earth's gravity acceleration is only about $3.3 \times 10^{-4} g_0$. For the most accurate analyses, the value of g will vary locally with the Earth's bulge at the equator, with the height of nearby mountains, and with the local difference of the Earth's density at specific regions.

4.3. BASIC RELATIONS OF MOTION

For any vehicle that flies within the Earth's proximity, the gravitational attractions from all other heavenly bodies are neglected. Assume next that the vehicle is moving in rectilinear equilibrium flight and that all control forces, lateral forces, and moments that tend to turn the vehicle are zero. The resulting trajectory is two-dimensional and is contained in a fixed plane. Assume further that the vehicle has wings inclined to the flight path at an angle of attack α providing lift in a direction normal to the flight path. The direction of flight need not coincide with the direction of thrust as shown schematically in Figure 4–5.

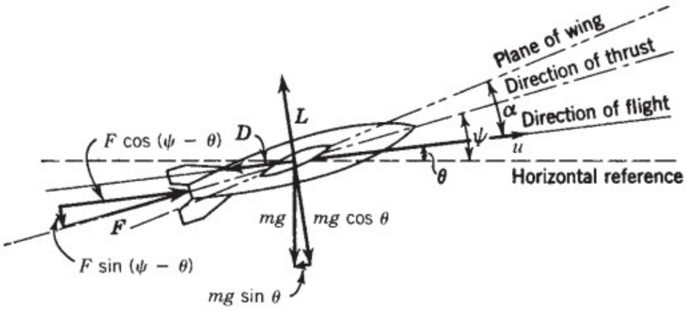

FIGURE 4–5. Two-dimensional free-body force diagram for flying vehicle with wings and fins.

Let θ be the angle of the flight path with the horizontal and ψ the angle of the direction of thrust with the horizontal. Along the flight path direction, the product of the mass and the acceleration has to equal the vector sum of the propulsive, aerodynamic, and gravitational forces:

$$m(du/dt) = F\cos(\psi - \theta) - D - mg\sin\theta \qquad (4\text{–}13)$$

The acceleration perpendicular to the flight path is $u(d\theta/dt)$; for a given value of u and at the instantaneous Earth radius R (from the Earth's center) of the flight path, it becomes u^2/R. The equation of motion in a direction normal to the flight velocity is

$$mu(d\theta/dt) = F\sin(\psi - \theta) + L - mg\cos\theta \qquad (4\text{–}14)$$

By substituting from Eqs. 4–10 and 4–11, these two basic equations can be solved for the accelerations as follows:

$$\frac{du}{dt} = \frac{F}{m}\cos(\psi - \theta) - \frac{C_D}{2m}\rho u^2 A - g\,\sin\theta \qquad (4\text{–}15)$$

$$u\frac{d\theta}{dt} = \frac{F}{m}\sin(\psi - \theta) + \frac{C_L}{2m}\rho u^2 A - g\,\cos\theta \qquad (4\text{–}16)$$

No general solution can be given to these equations, since t_p, u, C_D, C_L, ρ, θ, and/or ψ may vary independently with time, mission profile, and/or altitude. Also, C_D and C_L are functions of velocity or Mach number. In more sophisticated analyses, other factors may also be considered, such as any propellant used for nonpropulsive purposes (e.g., attitude control or flight stability). See Refs. 4–1, 4–8, 4–11, and 4–12 for background material on flight performance in some of these flight regimes. Because rocket propulsion systems are usually tailored to fit specific flight missions, different flight performance parameters are maximized (or optimized) for different rocket flight missions or flight regimes such as Δu, range, orbit height and shape, time-to-target, or altitude.

For actual trajectory analyses, navigation computation, space flight path determination, or missile-firing tables, the above two-dimensional simplified theory is not sufficiently accurate; perturbation effects, such as those listed in Section 4.4, must then be considered in addition to drag and gravity, and computer modelling is necessary to handle such complex relations. Suitable divisions of the trajectory into small elements and step-by-step or numerical integrations to define a trajectory are usually indicated. More generally, three-body theories include the gravitational attraction among three masses (for example, the Earth, the moon, and the space vehicle) and this is considered necessary in many space flight problems (see Refs. 4–2 to 4–5). The form and the solution to the given equations become further complicated when propellant flow and thrust are not constant and when the flight path is three dimensional.

For each mission or flight one can obtain actual histories of velocities and distances traveled and thus complete trajectories when integrating Eqs. 4–15 and 4–16. More general cases require six equations: three for translation along each of three perpendicular axes and three for rotation about these axes. The choice of coordinate systems of reference points can simplify the mathematical solutions (see Refs. 4–3 and 4–5) but there are always a number of trade-offs in selecting the best trajectory for the flight of a rocket vehicle. For example, for a fixed thrust the trade-off is between burn time, drag, payload, maximum velocity, and maximum altitude (or range). Reference 4–2 describes the trade-offs between payload, maximum altitude, and flight stability for sounding rockets.

Equations 4–15 and 4–16 may be further simplified for various special applications, as shown below; results of such calculations for velocity, altitude, or range using the above two basic equations are often adequate for rough design estimates. A form of these equations is also useful for determining the actual thrust or actual specific impulse during vehicle flight from accurately observed trajectory data, such as from optical or radar tracking data. Vehicle acceleration (*du*/*dt*) is essentially proportional to net thrust and, by making assumptions or measurements of propellant flow (which usually varies in a predetermined manner) and from analyses of aerodynamic forces, it is possible to determine a rocket propulsion system's actual thrust under flight conditions.

Equations 4–15 and 4–16 simplify for wingless rocket projectiles, space launch vehicles, or missiles with constant thrust and propellant flow. In Fig. 4–6 the flight direction θ is the same as the thrust direction and any lift forces for a symmetrical, wingless, stably flying vehicle are neglected at zero angle of attack. For a two-dimensional trajectory in a single plane (no wind forces) and a stationary Earth, the acceleration in the direction of flight is given below, where t_p is the operating or burn time of the propellant and ζ is the propellant mass fraction:

$$\frac{du}{dt} = \frac{c\zeta/t_p}{1 - \zeta t/t_p} - g\sin\theta - \frac{C_D \frac{1}{2}\rho u^2 A/m_0}{1 - \zeta t/t_p} \tag{4–17}$$

The force vector diagram in Fig. 4–6 also shows the net force (the addition of thrust, drag, and gravity vectors) to be at an angle to the flight path, which will therefore

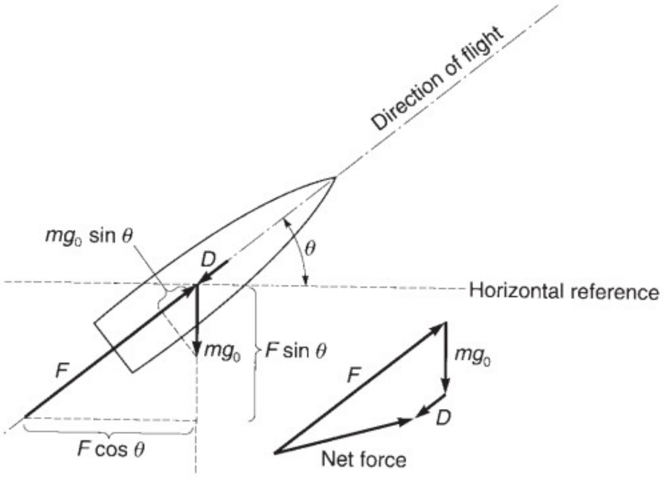

FIGURE 4–6. Simplified free-body force diagram for a vehicle without wings or fins. The force vector diagram shows the net force on the vehicle. All forces act through the vehicle's center of gravity.

be curved. These types of diagram form the basis for iterative trajectory numerical solutions.

All further relationships in this section correspond to two-dimensional flight paths, ones that lie in a single plane. If maneuvers out of that plane take place (e.g., due to solar attraction, thrust misalignment, or wind), then another set of equations will be needed—it requires both force and energy to push a vehicle out of its flight plane; Reference 4–1 describes equations for the motion of rocket projectiles in the atmosphere in three dimensions. Trajectories must be calculated accurately in order to reach any intended flight objective and today these are done exclusively with the aid of computers. Several *computer programs* for analyzing flight trajectories exist (which are maintained by aerospace companies and/or government agencies). Some are two-dimensional, relatively simple, and used for making preliminary estimates or comparisons of alternative flight paths, alternative vehicle designs, or alternative propulsion schemes. Several use a stationary flat Earth, while others use a rotating curved Earth. Three-dimensional programs used for more accurate flight path analyses may include some or all significant perturbations, orbit plane changes, or flying at angles of attack. Reference 4–4 explains their nature and complexity.

When the flight trajectory is vertical (as for a sounding rocket), then Eq. 4–17 becomes

$$\frac{du}{dt} = \frac{c\zeta/t_p}{1 - \zeta t/t_p} - g - \frac{C_D\frac{1}{2}\rho u^2 A/m_0}{1 - \zeta t/t_p} \tag{4–18}$$

The velocity at the end of burning can be found by integrating between the limits of $t = 0$ and $t = t_p$ where $u = u_0$ and then $u = u_p$. The first two terms can readily be integrated. The last term is of significance only if the vehicle spends a considerable portion of time within the lower atmosphere. It may be integrated graphically or numerically, and its value can be designated by the term BC_DA/m_0, where

$$B \equiv \int_0^{t_p} \frac{\frac{1}{2}\rho u^2}{1 - \zeta t/t_p} dt$$

The cutoff velocity or velocity at the end of propellant burning u_p then becomes

$$u_p = -\bar{c}\ln(1 - \zeta) - \bar{g}t_p - \frac{BC_DA}{m_0} + u_0 \tag{4–19}$$

where u_0 is an initial velocity such as may be given by a booster, $\bar{g}$ is an average gravitational attraction evaluated with respect to time and altitude from Eq. 4–12, and $\bar{c}$ is a time average of the effective exhaust velocity, which also depends on altitude. For nonvertical flight paths, the gravity loss becomes a function of the angle between the flight direction and the local horizontal; more specifically, the gravity loss is then given by the integral $\int (g \sin\theta)dt$.

When aerodynamic forces within the atmosphere may be neglected (or for vacuum operation) and when no booster or means for attaining an initial velocity ($u_0 = 0$) are present, the velocity at the end of the burning reached with a vertically ascending trajectory becomes simply

$$\begin{aligned} u_p &= -\bar{c}\ln(1 - \zeta) - \bar{g}t_p \\ &= -\bar{c}\ln \mathbf{MR} - \bar{g}t_p = \bar{c}\ln(1/\mathbf{MR}) - \bar{g}t_p \end{aligned} \tag{4–20}$$

The first term on the right side is usually the largest and is identical to Eq. 4–6. It is directly proportional to the effective rocket exhaust velocity and very sensitive to changes in the mass ratio. The second term is related to the Earth's gravity and is always negative during ascent, but its magnitude can be small when the burn time t_p is short or when flight is taking place at high orbits or in space where $\bar{g}$ is comparatively small.

For the simplified case given in Eq. 4–20 the net initial acceleration a_0 for vertical takeoff at sea level is

$$a_0 = (F_0 g_0/w_0) - g_0 \tag{4–21}$$

$$a_0/g_0 = (F_0/w_0) - 1 \tag{4–22}$$

where a_0/g_0 is the *initial takeoff acceleration* in multiples of the sea-level gravitational acceleration g_0, and F_0/w_0 is the thrust-to-weight ratio at takeoff. For large surface-launched vehicles, this initial-thrust-to-initial-weight ratio typically has values between 1.2 and 2.2; for small missiles (air-to-air, air-to-surface, and surface-to-air types) this ratio is usually larger, sometimes even as high as 50 or 100. *The final* or *terminal acceleration* a_f of a vehicle in vertical Earth ascent usually occurs just before the rocket engine is shut off and/or before the useable propellant is completely consumed. If drag is neglected, then the final acceleration a_f acting on the final mass m_f becomes

$$a_f = (F_f/m_f) - g \tag{4–23}$$

This equation applies when the powered flight path traverses a substantial range of altitude with g decreasing according to Eq. 4–12. For rocket propulsion systems with constant propellant flow, the final acceleration is usually also the maximum acceleration because the vehicle mass being accelerated is a minimum just before propellant termination, and for ascending rockets thrust usually increases with altitude. When this terminal acceleration is found to be too large (e.g., causes overstressing of the structure, thus necessitating an increase in structure mass), then the thrust must be reduced by redesign to a lower value for the last portion of the burning period. In manned flights, maximum accelerations are limited to the maximum g-loading that can be withstood by the crew.

Example 4–1. A simple single-stage rocket for a rescue flare has the following characteristics. Its flight path nomenclature is shown in the accompanying sketch.

Launch weight (w_0)	4.0 lbf
Useful propellant weight (w_p)	0.4 lbf
Effective specific impulse	120 sec
Launch angle θ (relative to horizontal)	80°
Burn time t_p (with constant thrust)	1.0 sec

The heavy line in the ascending trajectory designates the powered portion of the flight.

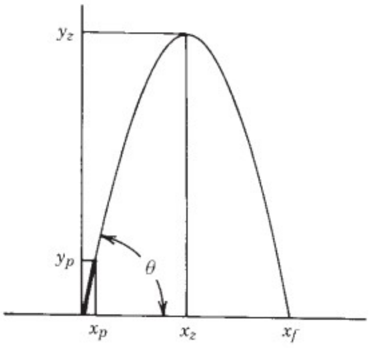

Drag may be neglected since flight velocities are low. Assume that the acceleration of gravity is unchanged from its sea-level value g_0, which then makes the propellant mass numerically equal to the propellant weight in the EE system, or 0.4 lbm. Also assume that start and stop transients are short and can be neglected.

Solve for the initial and final acceleration during powered flight, the maximum trajectory height and the time to reach maximum height, the range or horizontal distance to impact, and the angle at propulsion cutoff.

SOLUTION. We will divide the flight path into three portions: the powered flight for 1 sec, the unpowered ascent after cutoff, and the free-fall descent. The thrust is obtained from Eq. 2–5:

$$F = I_s w_p/t_p = 120 \times 0.4/1.0 = 48 \text{ lbf}$$

The initial accelerations along the x-horizontal and y-vertical directions are, from Eq. 4–22,

$$(a_0)_y = g_0[(F \sin\theta/w) - 1] = 32.17[(48/4.0)\sin 80^\circ - 1] = 348 \text{ ft/sec}^2$$

$$(a_0)_x = g_0(F \cos\theta/w) = 32.17(48/4.0)\cos 80^\circ = 67.03 \text{ ft/sec}^2$$

At thrust termination the final flight acceleration becomes

$$a_0 = \sqrt{(a_0)_x^2 + (a_0)_y^2} = 354.4 \text{ ft/sec}^2$$

The vertical and horizontal components of the velocity u_p at the end of powered flight can be obtained from Eq. 4–20. Note that the vehicle mass has been diminished by the propellant that has been consumed:

$$\begin{aligned}(u_p)_y &= c \ln(m_0/m_f) \sin\theta - g_0 t_p = 32.17 \times 120 \times \ln(4/3.6) \times 0.984 - 32.17 \times 1.0\\ &= 368 \text{ m/sec}\\ (u_p)_x &= c \ln(m_0/m_f) \cos\theta = 32.17 \times 120 \times \ln(4/3.6) \times 0.1736 = 70.6 \text{ m/sec}\end{aligned}$$

The effective exhaust velocity $c = I_s g_0$ (Eq. 2–6). The trajectory angle with the horizontal at rocket cutoff in dragless flight is

$$\tan^{-1}(368/70.6) = 79.1°$$

The final acceleration is found, using Eq. 4–22 with the final mass, as $a_f = 400 \text{ m/sec}^2$.

For the powered flight, the coordinates at propulsion burnout y_p and x_p can be calculated from the time integration of their respective velocities, using Eq. 2–6. The results are

$$\begin{aligned}y_p &= ct_p[1 - \ln(m_0/m_f)/(m_0/m_f - 1)]\sin\theta - 1/2 g_0 t_p^2\\ &= 32.17 \times 120[1 - \ln(4/3.6)/(4/3.6 - 1)] \times 0.984 - \tfrac{1}{2} \times 32.17 \times (1.0)^2 = 181 \text{ ft}\\ x_p &= ct_p[1 - \ln(m_0/m_f)/(m_0/m_f - 1)]\cos\theta\\ &= 32.17 \times 120[1 - \ln(4/3.6)/(4/3.6 - 1)] \times 0.173 = 34.7 \text{ ft}\end{aligned}$$

The unpowered part of the trajectory reaches zero vertical velocity at its zenith. The height gained in unpowered free flight may be obtained by equating the vertical kinetic energy at power cutoff to its equivalent potential energy, $g_0(y_z - y_p) = \frac{1}{2}(u_p)_y^2$

so that

$$(y_z - y_p) = \tfrac{1}{2}(u_p)_y^2/g_o = \tfrac{1}{2}(368)^2/32.17 = 2105 \text{ ft}$$

The maximum height or zenith location thus becomes $y_z = 2105 + 181 = 2259$ ft. What remains now is to solve the free-flight portion of vertical descent. The time for descent from the zenith is $t_z = \sqrt{2y_z/g_0} = 11.85$ sec and the final vertical or impact vertical velocity $(u_f)y - g_0 t_z = 381$ ft/sec.

The total horizontal range to the zenith is the sum of the powered and free-flight contributions. During free flight, the horizontal velocity remains unchanged at 70.6 ft/sec because there are no accelerations (i.e., no drag, wind, or gravity component). We now need to find the free-flight time from burnout to the zenith, which is $t = (u_p)y/g_0 = 11.4$ sec. The total free-flight time becomes $t_{ff} = 11.4 + 11.85 = 23.25$ sec.

Now, the horizontal or total range becomes $\Delta x = 34.7 + 70.6 \times 23.25 = 1676$ ft.

The impact angle would be around 79°. If drag had been included, solving this problem would have required information on the drag coefficient (C_D) and a numerical solution using Eq. 4–18. All resulting velocities and distances would turn out somewhat lower in value. A set of flight trajectories for sounding rockets is given in Ref. 4–3.

4.4. SPACE FLIGHT

Newton's law of gravitation defines the gravitational attraction force F_g between two bodies in space as:

$$F_g = Gm_1m_2/R^2 = \mu m_2/R^2 \tag{4–24}$$

Here G is the universal gravity constant ($G = 6.674 \times 10^{-11}\mathrm{m}^3/\mathrm{kg} - \mathrm{sec}^2$), m_1 and m_2 are the masses of the two attracting bodies (such as the Earth and the moon, or the Earth and a spacecraft, or the sun and a planet), and R is the distance between their centers of mass. The Earth's gravitational constant μ is the product of Newton's universal constant G and the mass of the Earth, m_1 (5.974×10^{24}kg). It is $\mu = 3.987 \times 10^{14}\mathrm{m}^3/\mathrm{sec}^2$.

Rockets offer a means for escaping the Earth's pull for lunar and interplanetary travel, for escaping our solar system, and for launching stationary or moving platforms in space. The flight velocity required to escape from the Earth can be found by equating the kinetic energy of the moving body to the work necessary to overcome gravity, neglecting the rotation of the Earth and the attraction of other celestial bodies, namely,

$$\tfrac{1}{2}mu^2 = m\int g\,dR$$

By substituting for g from Eq. 4–12 and neglecting air friction, the following relation for the Earth's escape velocity v_e is obtained:

$$v_e = R_0\sqrt{\frac{2g_0}{R_0 + h}} = \sqrt{\frac{2\mu}{R}} \tag{4–25}$$

Here, R_0 is the effective Earth mean radius (6374.2 km), h is the circular orbit's altitude above sea level, and g_0 is the acceleration of gravity at the Earth's surface (9.806 m/sec). The satellite flight radius R as measured from the earth's center is $R = R_0 + h$. The velocity of escape from the Earth's surface is 11,179 m/sec or 36,676 ft/sec and does not vary appreciably within the Earth's atmosphere, as shown by Fig. 4–7. Escape velocities for surface launches are given in Table 4–1 from the sun, the planets, and the moon. Launching from the Earth's surface at the escape velocity is not practical because as such vehicle ascends through the Earth's atmosphere, it is subject to severe aerodynamic heating and dynamic pressures. A practical launch vehicle has to traverse the lower atmosphere at relatively low velocities and then accelerate to high velocities beyond the dense atmosphere. For example, during a portion of the Space Shuttle's ascent, its main engines were actually throttled to a lower thrust. Alternatively, an Earth escape vehicle may be launched above the denser atmosphere from an orbiting space station or from an orbiting launch vehicle's upper stage.

Any rocket or spaceship may become an Earth *satellite*, revolving around the Earth in a fashion similar to that of the moon. Low Earth orbits, typically below 500 km altitude, are designated as LEO. Most orbit altitudes are above the Earth's lower atmosphere because this minimizes the energy expended to overcome the drag that

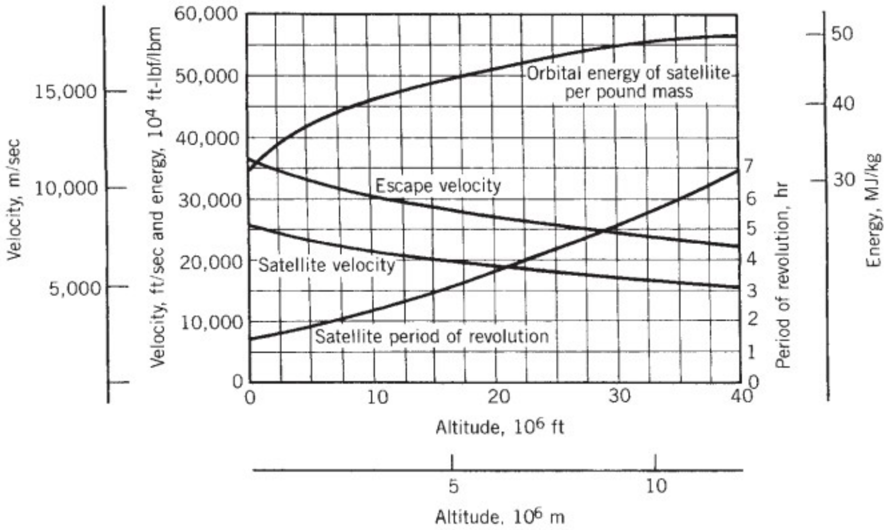

FIGURE 4–7. Orbital energy, orbital velocity, period of revolution, and Earth escape velocity for a space vehicle as a function of altitude for circular satellite orbits. Values are based on a spherical Earth and neglect the Earth's oblate shape, its rotation, and atmospheric drag.

continuously brings the vehicle's orbit closer to the Earth. However, radiation effects in the Van Allen belt on human beings and on sensitive equipment at times often necessitate the selection of Earth orbits at low altitudes.

For a satellite's circular trajectory, the velocity must be sufficient so that its centrifugal force precisely balances the Earth's gravitational attraction:

$$mu_s^2/R = mg$$

For a circular orbit, the satellite velocity u_s is found by using Eq. 4–12,

$$u_s = R_0\sqrt{g_0/(R_0 + h)} = \sqrt{\mu/R} \tag{4–26}$$

which is smaller than the escape velocity by a factor of $\sqrt{2}$. The period τ in seconds for one revolution in a circular orbit relative to a stationary Earth is

$$\tau = 2\pi(R_0 + h)/u_s = 2\pi(R_0 + h)^{3/2}/(R_0\sqrt{g_0}) \tag{4–27}$$

Neglecting drag, the energy E necessary to bring a unit of mass into a circular satellite orbit consists of its kinetic and potential energy, namely,

$$\begin{aligned} E &= \tfrac{1}{2}u_s^2 + \int_{R_0}^{R} g\,dR \\ &= \tfrac{1}{2}R_0^2\tfrac{g_0}{R_0+h} + \int_{R_0}^{R} g_0\tfrac{R_0^2}{R^2}dR = \tfrac{1}{2}R_0 g_0\tfrac{R_0+2h}{R_0+h} \end{aligned} \tag{4–28}$$

TABLE 4–1. Characteristic Data for Several Heavenly Bodies

Name	Mean Radius of Orbit (million km)	Period of Revolution	Mean Diameter (km)	Relative Mass (*Earth* = 1.0)[a]	Specific Gravity	Acceleration of Gravity at Surface (m/sec^2)	Escape Velocity at Surface (m/sec)
Sun	—	—	1,393,000	332, 950	1.41	273.4	616,000
Moon	0.383	27.3 days	3475	0.012	3.34	1.58	2380
Mercury	57.87	87.97 days	4670	0.06	5.5	3.67	4200
Venus	108.1	224.70 days	12,400	0.86	5.3	8.67	10,300
Earth	149.6	365.256 days	12,742	1.00^3	5.52	9.806	11,179
Mars	227.7	686.98 days	6760	0.15	3.95	3.749	6400
Jupiter	777.8	11.86 year	143,000	318.4	1.33	26.0	59,700
Saturn	1486	29.46 year	121,000	95.2	0.69	11.4	35,400
Uranus	2869	84.0 year	47,100	17.0	1.7	10.9	22,400
Neptune	4475	164.8 year	50,700	17.2	1.8	11.9	31,000
Pluto	5899	248 year	2368	0.00218	1.44	0.658	1229

[a] Earth mass is 5.976×10^{24} kg.

Source: In part from Refs 4–3 and 4–4.

Escape velocity, satellite velocity, satellite period, and satellite orbital energy are all shown as functions of altitude in Fig. 4–7.

A satellite moving around the Earth at an altitude of 300 miles or 482.8 km has a velocity u_s of about 7375 m/sec or 24,200 ft/sec, circles a stationary Earth in $\tau =$ 1.63 hr.; ideally it requires an energy of 3.35×10^7 J to place 1 kg of spaceship mass into its orbit. An equatorial satellite in a circular orbit at an altitude of 6.611 Earth radii (about 26,200 miles, 42,200 km, or 22,700 nautical miles) has a period of revolution of exactly 24 hr. It will, therefore, appear stationary to an observer on Earth. This is known as a *synchronous* satellite in *geo synchronous Earth orbit*, usually abbreviated as GEO. This orbit is used extensively for communications satellite and Earth observation applications. In the part of Section 4.7 on launch vehicles, we describe how the payload of any given space vehicle diminishes as the orbit's circular altitude is increased and as the inclination (angle between orbit plane and Earth equatorial plane) is changed. See Refs. 4–3, 4–4, 4–5, 4–6 and 4–9.

Elliptical Orbits

The circular orbit described above is a special case of the more general elliptical orbit shown in Fig. 4–8; here, the Earth (or any other heavenly body around which another body is moving) is located at one of the focal points of this ellipse. The relevant equations of motion come from Kepler's laws and elliptical orbits may be described as follows, when expressed in polar coordinates:

$$u = \left[\mu\left(\frac{2}{R} - \frac{1}{a}\right)\right]^{1/2} \tag{4–29}$$

where u is the velocity of the body in the elliptical orbit, R is the instantaneous radius from the center of the Earth (a vector quantity, which changes direction as well as magnitude), a is the major axis of the ellipse, and μ is the Earth's gravitational constant, 3.986×10^{14} m^3/sec^2. These symbols are defined in Fig. 4–8. From Eq. 4–29 it can be seen that the velocity has its maximum value u_p when the moving body comes closest to its focal point at its orbit's *perigee* and the minimum value u_a at its *apogee*. By substituting for R in Eq. 4–29, and by defining the ellipse's shape factor e as the *eccentricity of the ellipse*, $e = \sqrt{a^2 - b^2}/a$, the apogee and perigee velocities can be expressed as

$$u_a = \sqrt{\frac{\mu(1-e)}{a(1+e)}} \tag{4–30}$$

$$u_p = \sqrt{\frac{\mu(1+e)}{a(1-e)}} \tag{4–31}$$

Another property of an elliptical orbit is that the product of velocity and instantaneous radius remains constant for any location x or y on the ellipse, namely, $u_x R_x = u_y R_y = uR$. The exact path that a satellite takes depends on the velocity (magnitude and vector orientation) with which it is started or was injected into orbit.

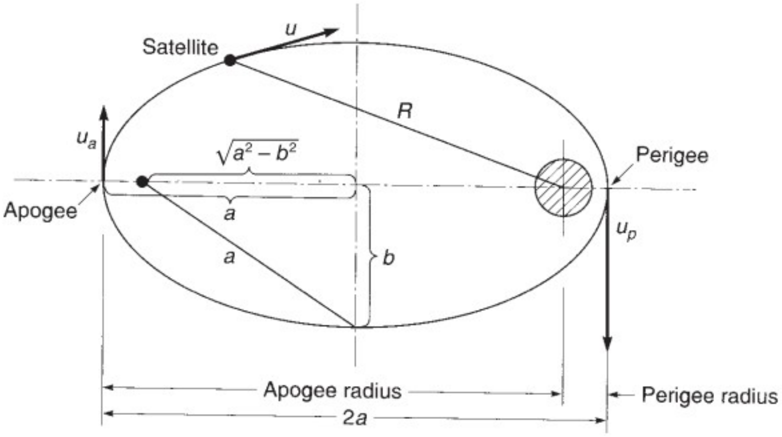
Satellite
u
R
u_a
$\sqrt{a^2 - b^2}$
Perigee
Apogee
a
a
b
u_p
Apogee radius
Perigee radius
2a

FIGURE 4–8. Elliptical orbit; the attracting body is at one of the focal points of the ellipse.

For *interplanetary transfers*, an ideal mission can be achieved with minimum energy with a simple transfer ellipse, as suggested originally by Hohmann (see Ref. 4–7). Assuming that planetary orbits about the sun are circular and coplanar, Hohmann demonstrated that the path of minimum energy is an ellipse tangent to both planetary orbits as shown in Fig. 4–9. This operation requires a velocity increment (of relatively high thrust) at the initiation (planet A at time t_1) and another at termination (planet B at time t_2): both increments equal the velocity differences between the respective circular planetary velocities and the perigee and apogee velocities which define the transfer ellipse. Thrust levels at the beginning and end maneuvers of the Hohmann ellipse must be high enough to amount to a short operating time and an acceleration of at least 0.01 g_0, but preferably more. Note that because electrical propulsion accelerations are much lower, amounting

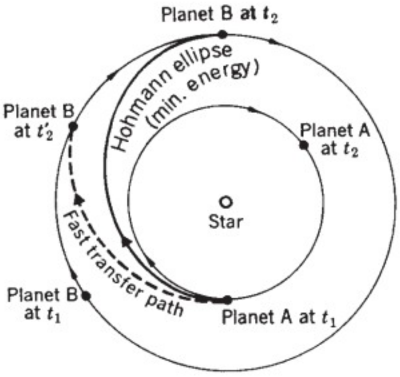
Planet B at t_2
Hohmann ellipse (min. energy)
Planet B at t_2'
Planet A at t_2
Star
Fast transfer path
Planet B at t_1
Planet A at t_1

FIGURE 4–9. Schematic of interplanetary transfer paths. These same transfer maneuvers apply when going from a low-altitude Earth satellite orbit to a higher orbit.

to about $10^{-5}g_0$, and operating times longer, weeks or months, the best transfer trajectories in electrical propulsion turn out to be much different from Hohmann ellipses; these are described in Chapter 17.

Departure dates or the *relative positions of the launch planet and the target planet* in planetary transfer missions become critical, because the spacecraft must meet with the target planet when it arrives at the target orbit. Transfer times $(t_2 - t_1)$ for Hohmann-ellipse flights starting on Earth are about 116 hours to go to the moon and about 259 days to Mars. If faster flight paths (shorter transfer times) are desired (see dashed lines in Fig. 4–9), they will require more energy than those with a Hohmann transfer ellipse. This means a larger vehicle with more propellant and a larger propulsion system, or a higher total impulse. There always is a *time window* for launching a spacecraft that will make for a successful rendezvous. For Mars missions an Earth-launched spacecraft may have a launch time window of more than two months. Hohmann transfer ellipses or faster transfer paths apply not only to planetary flight but also to Earth satellites when they go from one circular orbit to another (but within the same plane). Also, if one spacecraft goes to rendezvous with another spacecraft in a different orbit, the two have to be in the proper predetermined positions prior to any thrust application to simultaneously reach their rendezvous location.

When the launch orbit (or launch planet) is not in the same plane as the target orbit, then additional energy will be needed for applying thrust in directions normal to the launch orbit plane. More information can be found in Refs. 4–3, 4–4, 4–6, and 4–10.

Example 4–2. A satellite is launched from a circular equatorial parking orbit at an altitude of 160 km into a coplanar circular synchronous orbit by using a Hohmann transfer ellipse. Assume a homogeneous spherical Earth with a radius of 6371 km. Determine the velocity increments for entering the transfer ellipse and for achieving the synchronous orbit at 42,200 km altitude. See Fig. 4–9 for the terminology of the orbits.

SOLUTION. The two orbits are $R_A = 6.531 \times 10^6$ m; $R_B = 48.571 \times 10^6$ m. The major axis a_{te} of the transfer ellipse is

$$a_{te} = \tfrac{1}{2}(R_A + R_B) = 27.551 \times 10^6 \mathrm{m}$$

The orbit velocities of the two satellites are

$$u_A = \sqrt{\mu/R_A} = [3.986005 \times 10^{14}/6.571 \times 10^6]^{1/2} = 7788 \text{ m/sec}$$

$$u_B = \sqrt{\mu/R_B} = 2864.7 \text{ m/sec}$$

The velocities needed to enter and exit the transfer ellipse are

$$(u_{te})_A = \sqrt{\mu}[(2/R_A) - (1/a)]^{1/2} = 10{,}337 \text{ m/sec}$$

$$(u_{te})_B = \sqrt{\mu}[(2/R_B) - (1/a)]^{1/2} = 1394 \text{ m/sec}$$

The changes in velocity going from parking orbit to ellipse Δu_A and from ellipse to final orbit Δu_B are

$$\Delta u_A = |(u_{te})_A - u_A| = 2549 \text{ m/sec}$$

$$\Delta u_B = |u_B - (u_{te})_B = 1471 \text{ m/sec}$$

The total velocity change for the transfer maneuvers is

$$\Delta u_{\text{total}} = \Delta u_A + \Delta u_B = 4020 \text{ m/sec}$$

Figure 4–10 shows the elliptical transfer trajectory of a ballistic missile or a satellite launch or an ascent vehicle. During the initial powered flight the trajectory angle is adjusted by signals from the guidance system and torques from the reaction control system to an angle that will allow the vehicle to reach the apogee of its elliptical path at exactly the desired altitude. An *orbit injection velocity* increase of the space vehicle is now applied by a chemical propulsion system at this apogee, which causes the vehicle to change from an elliptical transfer flight path to a circular-orbit flight path. The horizontal arrow symbolizes this velocity increase. For an ideal satellite the simplified theory assumes that an orbit injection maneuver is essentially an instantaneous application of the total impulse when the ballistic elliptic trajectory reaches its

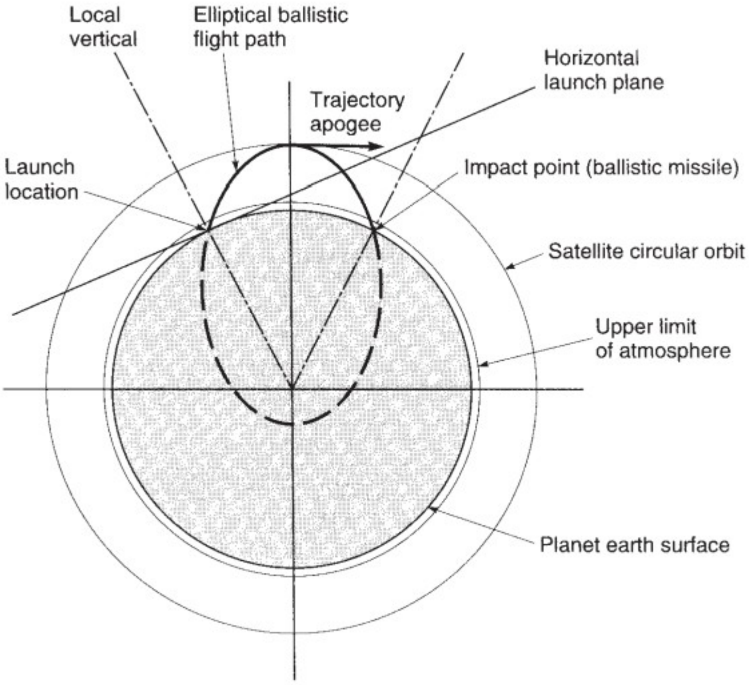
Local vertical
Elliptical ballistic flight path
Horizontal launch plane
Trajectory apogee
Launch location
Impact point (ballistic missile)
Satellite circular orbit
Upper limit of atmosphere
Planet earth surface

FIGURE 4–10. Long-range ballistic missiles follow an elliptical free flight trajectory, which is nearly drag free, with the Earth's center as one of the focal points. The surface launch is usually vertically up (not shown here) but the flight path is quickly tilted during the early powered flight to enter into an elliptic trajectory. The ballistic range is the arc distance on the Earth's surface. The same elliptical flight path can be used by launch vehicles for satellites; another powered flight period occurs (called orbit injection) just as the vehicle reaches its elliptical apogee (as indicated by the arrow), causing the vehicle to enter an orbit.

apogee or zenith. In reality, the rocket propulsion system for orbit injection operates over a finite time, during which gravity losses and changes in altitude occur.

Deep Space

Lunar and *interplanetary* missions may include circumnavigation, landing, and return flights. The energy necessary to escape from the Earth may be calculated as $\frac{1}{2}mv_e^2$ from Eq. 4–25 as 6.26×10^7 J/kg, which is more than that required to launch an Earth satellite. The gravitational attraction of various heavenly bodies and their respective escape velocities depends on their mass and diameter; approximate values are listed in Table 4–1. An idealized diagram of an interplanetary landing mission is shown in Fig. 4–11.

Escape from the solar system requires approximately 5.03×10^8 J/kg which is eight times as much energy as is required for escape from the Earth. Technology exists today to send small, unmanned probes away from the sun into outer space, but before any mission to the nearest star can be achieved some very long-duration, novel, rocket propulsion system must be introduced. The ideal trajectory for a spacecraft to escape from the sun is either a parabola (minimum energy) or a hyperbola. See Refs. 4–6 and 4–10.

The *Voyager 2 Spacecraft*, developed by NASA's Jet Propulsion Laboratory, was the first man-made object to escape from the solar system and enter interplanetary space. It was launched on August 20, 1977 for exploring the outer planets (flybys of Jupiter, Saturn, Neptune, and Uranus) and then leaving the solar system. It was not expected that Voyager 2 would continue to be operational for over 37 years. A three-axis stabilization system with gyroscopic and celestial reference instruments

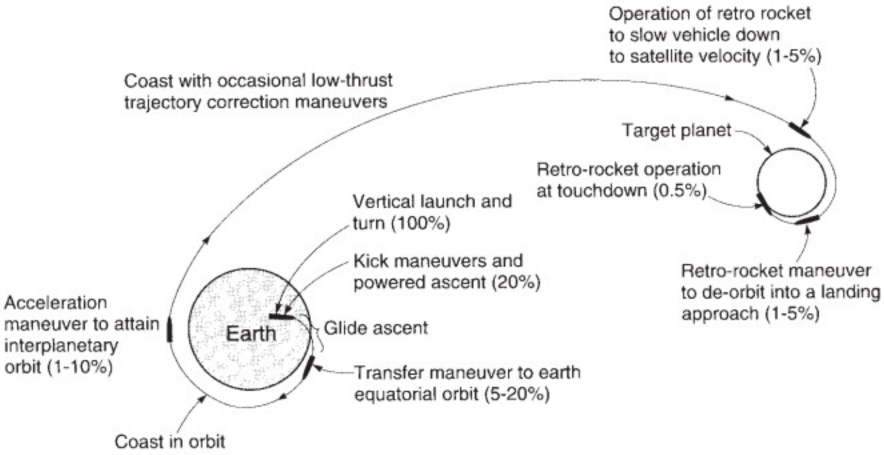
Operation of retro rocket to slow vehicle down to satellite velocity (1-5%)
Coast with occasional low-thrust trajectory correction maneuvers
Target planet
Retro-rocket operation at touchdown (0.5%)
Vertical launch and turn (100%)
Kick maneuvers and powered ascent (20%)
Retro-rocket maneuver to de-orbit into a landing approach (1-5%)
Acceleration maneuver to attain interplanetary orbit (1-10%)
Earth
Glide ascent
Transfer maneuver to earth equatorial orbit (5-20%)
Coast in orbit

FIGURE 4–11. Schematic of typical powered flight maneuvers during a hypothetical two-dimensional interplanetary mission with a landing (not drawn to scale). The numbers indicate typical thrust magnitudes of the maneuvers in percent of launch takeoff thrust. Heavier lines show powered flight path segments.

20 04 2020

C) Books_70_2020/
Mathematics/Higher Level for the IB Diploma/
0199152268.pdf

Pág. 466(251)

Volumes of solids of revolution when rotated about y-axis

obs: Ver livro todo.

D) Books_70_2020/
Rocket Propulsion Elements/
George P. Sutton .. Rocket..pdf

Pág. 106 (130)

4.3 Basic Relations of Motion

Pág. 107(131)

Figure 4-5 Two-dimensional free body force diagram for flying vehicle with wings and fins.

E) Books_70_2020/
Schaum's Outline of Trigonometry/
Robert Moyer.. Schaum... pdf

Obs. Livro todo.

F) Books_20_2020/
Complex Variables Second Edition.../
0486406792.pdf

)*+,)))-) '/ + 3 1 H : B *1
> - 8 1H ' 4! > - < = % 0!
-3 5 6 8 ! 777

)*+,)))-) '/ '%.' ./# 0!
0 2 3 :

)*+,)))-) '/ % /%.' 0!
l B 0 1 1
(' # ' +

4.1 Rotations

In classical mechanics, rotations of three-vectors can be represented by a rotation matrix $\mathbf{R}$ acting upon vectors such as $\mathbf{x} = (x, y, z)$ as

$$x_i \to x'_i = \mathbf{R}_{ij} x_j . \tag{4.1}$$

In quantum mechanics, rotation transformations are represented by unitary operators $U(\boldsymbol{\theta})$ acting upon state vectors $|\psi\rangle$ such that

$$|\psi\rangle \to |\psi'\rangle = U(\boldsymbol{\theta})|\psi\rangle, \tag{4.2}$$

where the direction of $\boldsymbol{\theta}$ is along the axis about which the rotation occurs, and its magnitude is the rotation angle. For infinitesimal rotations, the operator $U(\boldsymbol{\theta})$ can be written as $U(\boldsymbol{\theta}) \simeq 1 + \mathrm{i}\boldsymbol{\theta} \cdot \mathbf{J}$, where the Hermitian operators $\mathbf{J}$ are the rotation generators. For spinless states, J_i can be represented as differential operators ($J_k = \frac{1}{2}\epsilon_{ijk} J_{ij}$, with $J_{ij} = -\mathrm{i}(x_i \partial_j - x_j \partial_i)$), it is easy to check that the commutation relations

$$[J_i, J_j] = \mathrm{i}\epsilon_{ijk} J_k \tag{4.3}$$

are satisfied for $i, j = 1, 2, 3 \leftrightarrow x, y, z$. Finite rotations can be built up from an infinite product of infinitesimal ones, so that the operator $U(\boldsymbol{\theta}) = \exp(\mathrm{i}\boldsymbol{\theta} \cdot \mathbf{J})$. The operators $U(\boldsymbol{\theta})$ form a representation of the Lie group $SU(2)$, for which the J_i are the group generators, and where Eq. (4.3) defines the Lie algebra associated with the group $SU(2)$. Since the parameters θ_i each run over a compact domain 0 to 2π, we say that $SU(2)$ is a compact Lie group.

A Casimir operator is an operator that commutes with all of the group generators. The eigenvalues of a Casimir operator are unchanged under group transformations, so they serve as a useful tool to classify group representations. The representations of $SU(2)$ can be labelled according to the eigenvalues of the quadratic Casimir operator

4

The supersymmetry algebra

4.1 Rotations

In classical mechanics, rotations of three-vectors can be represented by a rotation matrix $\mathbf{R}$ acting upon vectors such as $\mathbf{x} = (x, y, z)$ as

$$x_i \to x'_i = \mathbf{R}_{ij} x_j. \tag{4.1}$$

In quantum mechanics, rotation transformations are represented by unitary operators $U(\boldsymbol{\theta})$ acting upon state vectors $|\psi\rangle$ such that

$$|\psi\rangle \to |\psi'\rangle = U(\boldsymbol{\theta})|\psi\rangle, \tag{4.2}$$

where the direction of $\boldsymbol{\theta}$ is along the axis about which the rotation occurs, and its magnitude is the rotation angle. For infinitesimal rotations, the operator $U(\boldsymbol{\theta})$ can be written as $U(\boldsymbol{\theta}) \simeq 1 + \mathrm{i}\boldsymbol{\theta} \cdot \mathbf{J}$, where the Hermitian operators $\mathbf{J}$ are the rotation generators. For spinless states, J_i can be represented as differential operators ($J_k = \frac{1}{2}\epsilon_{ijk} J_{ij}$, with $J_{ij} = -\mathrm{i}(x_i \partial_j - x_j \partial_i)$), it is easy to check that the commutation relations

$$[J_i, J_j] = \mathrm{i}\epsilon_{ijk} J_k \tag{4.3}$$

are satisfied for $i, j = 1, 2, 3 \leftrightarrow x, y, z$. Finite rotations can be built up from an infinite product of infinitesimal ones, so that the operator $U(\boldsymbol{\theta}) = \exp(\mathrm{i}\boldsymbol{\theta} \cdot \mathbf{J})$. The operators $U(\boldsymbol{\theta})$ form a representation of the Lie group $SU(2)$, for which the J_i are the group generators, and where Eq. (4.3) defines the Lie algebra associated with the group $SU(2)$. Since the parameters θ_i each run over a compact domain 0 to 2π, we say that $SU(2)$ is a compact Lie group.

A Casimir operator is an operator that commutes with all of the group generators. The eigenvalues of a Casimir operator are unchanged under group transformations, so they serve as a useful tool to classify group representations. The representations of $SU(2)$ can be labelled according to the eigenvalues of the quadratic Casimir operator

$J^2 = \mathbf{J} \cdot \mathbf{J}$, for which $J^2|jm\rangle = j(j+1)|jm\rangle$, with $j = 0, 1/2, 1, 3/2, 2, \ldots$ For the $j = 1/2$ representation, the operators J_i can be represented by the Pauli spin matrices $J_i = \sigma_i/2$, and the state vectors can be represented by 2-component spinors. Higher j representations can be constructed by taking direct products of lower j representations. For higher j representations of $SU(2)$, the J_i's can be represented by $(2j+1) \times (2j+1)$ matrices, and the corresponding state vectors by $2j+1$ component column matrices.

4.2 The Lorentz group

We want to build a quantum theory that is invariant under Lorentz transformations. We restrict our discussion to proper, orthochronous Lorentz transformations, i.e. boosts and rotations, and neglect parity and time reversal. In addition to rotations which mix the spatial coordinates amongst themselves, we now have boost transformations which mix the time co-ordinate x^0 with the spatial co-ordinates; e.g. a boost along the x^1 direction can be written as $x'^0 = x^0 \cosh\phi + x^1 \sinh\phi$, $x'^1 = x^0 \sinh\phi + x^1 \cosh\phi$, $x'^2 = x^2$ and $x'^3 = x^3$. The usual velocity parameter β that characterizes the boost is given in terms of the rapidity ϕ by $\beta = \tanh\phi$. The infinitesimal transformation matrix U that transforms quantum mechanical states can then be augmented to,

$$U(\boldsymbol{\theta}, \boldsymbol{\phi}) \simeq 1 + \mathrm{i}\boldsymbol{\theta} \cdot \mathbf{J} + \mathrm{i}\boldsymbol{\phi} \cdot \mathbf{K}, \tag{4.4}$$

where K_i are the boost generators and $\boldsymbol{\phi}$ points along the direction of the boost.[1] A Lorentz transformation is thus characterized by the six parameters (θ_i, ϕ_j). Since the parameters ϕ_j are not restricted to a compact interval, the Lorentz group, unlike the rotation group, is not compact.

The Lorentz group generators satisfy

$$[J_i, J_j] = \mathrm{i}\epsilon_{ijk} J_k, \quad [K_i, J_j] = \mathrm{i}\epsilon_{ijk} K_k, \quad [K_i, K_j] = -\mathrm{i}\epsilon_{ijk} J_k. \tag{4.5}$$

The first of these relations shows that rotation generators form a closed sub-algebra, so that the rotation group forms a subgroup of the Lorentz group. The commutator of two boost generators is a rotation generator (this is the origin of Thomas precession) so that the boosts, by themselves, do not form a sub-algebra.

The Lorentz algebra that we have introduced above can be written in a manifestly covariant form by writing the generators as the six components of an antisymmetric second rank tensor generator $M_{\mu\nu}$, with $M_{ij} = \epsilon_{ijk} J_k$ and $M_{0i} = -M_{i0} = -K_i$. The commutators for the Lorentz group generators can then be recast into covariant form

---

[1] For infinitesimal boosts, notice that $|\boldsymbol{\phi}| = |\boldsymbol{\beta}|$.

$$[M_{\mu\nu}, M_{\rho\sigma}] = -\mathrm{i}(g_{\mu\rho}M_{\nu\sigma} - g_{\mu\sigma}M_{\nu\rho} - g_{\nu\rho}M_{\mu\sigma} + g_{\nu\sigma}M_{\mu\rho}). \tag{4.6}$$

To find the finite-dimensional unitary representations of the Lorentz group, the generators can be alternatively written by defining $S_i = \frac{1}{2}(J_i + \mathrm{i}K_i)$ and $T_i = \frac{1}{2}(J_i - \mathrm{i}K_i)$. In this case, it is easy to check that the commutators of the generators become

$$[S_i, S_j] = \mathrm{i}\epsilon_{ijk}S_k, \quad [T_i, T_j] = \mathrm{i}\epsilon_{ijk}T_k, \quad [S_i, T_j] = 0, \tag{4.7}$$

i.e. the algebra decomposes into the product of two independent $SU(2)$ groups, for which we know the representations. For the Lorentz group, there are thus *two* Casimir operators, S^2 and T^2, with eigenvalues $s(s+1)$ and $t(t+1)$, again with $s, t = 0, 1/2, 1, \ldots$ (Note that J^2 is no longer a Casimir operator since it no longer commutes with all the group generators, e.g. $[J^2, K_1] \neq 0$.) The irreducible representations can be categorized according to values of (s, t). A Lorentz scalar transforms as the $(0, 0)$ representation while a four-vector transforms as $(1/2, 1/2)$ representation. There are *two* distinct fundamental representations $(1/2, 0)$ and $(0, 1/2)$, each of which corresponds to two-spinors. The $(1/2, 0)$ object, as we will soon see, transforms as a left-handed Weyl two-spinor whose components are usually denoted by $\psi_{\mathrm{L}A}$, with $A = 1, 2$. The $(0, 1/2)$ object transforms as a right-handed two-spinor with components $\psi_{\mathrm{R}}^{\dot{A}}$, where the dot on the index calls attention to the fact that the spinor transforms under the second of the two $SU(2)$ groups.

Exercise *Verify that the boost transformation for a state transforming as the* $(1/2, 0)$ *and* $(0, 1/2)$ *representations of the Lorentz group are respectively given by* $\psi'_{\mathrm{L,R}} = (\cosh\frac{\phi}{2} \mp \boldsymbol{\sigma}\cdot\hat{\mathbf{p}}\sinh\frac{\phi}{2})\psi_{\mathrm{L,R}}$, *where* $\hat{\mathbf{p}}$ *is a unit vector along the direction of* $\mathbf{p}$. *Recalling that* $\tanh\phi = \beta$, *show that the spinors for states with momentum* $\mathbf{p}$ *can be obtained from the corresponding rest frame states as,*

$$\psi_{\mathrm{L,R}}(\mathbf{p}) = \frac{E + m \mp \boldsymbol{\sigma}\cdot\mathbf{p}}{\sqrt{2m(E+m)}}\psi_{\mathrm{L,R}}(0).$$

Noting that $\psi_{\mathrm{L}}(0) = \psi_{\mathrm{R}}(0)$ *because there is no preferred direction in the rest frame to define the particle's handedness, show that*

$$(E \pm \boldsymbol{\sigma}\cdot\mathbf{p})\psi_{\mathrm{L,R}}(\mathbf{p}) = m\psi_{R,L}(\mathbf{p}).$$

This is just the Dirac equation in two-component notation. Notice that for $m = 0$, *we have* $(\boldsymbol{\sigma}\cdot\hat{\mathbf{p}})\psi_{\mathrm{L,R}}(\mathbf{p}) = \mp\psi_{\mathrm{L,R}}(\mathbf{p})$, *which justifies the use of our labels left and right for the states transforming as the* $(1/2, 0)$ *and* $(0, 1/2)$ *representations of the Lorentz group.*

A four-component Dirac spinor can be built out of the direct sum of two-component spinors $(1/2, 0) \oplus (0, 1/2)$ so that

$$\psi_a^D = \begin{pmatrix} \psi_{\mathrm{L}A} \\ \chi_{\mathrm{R}}^{\dot{A}} \end{pmatrix}. \tag{4.8}$$

The two spinors ψ_{L} and χ_{R} are independent. It is simple to check that the two-component spinor $-\mathrm{i}\sigma_2\psi_{\mathrm{L}}^*$ transforms as a $(0, 1/2)$ representation of the Lorentz group, i.e. it transforms as χ_{R}. We can thus construct a different four-component spinor whose right-handed piece is completely determined by its left-handed pieces via $\chi_{\mathrm{R}} = -\mathrm{i}\sigma_2\psi_{\mathrm{L}}^*$. This four-spinor would transform as the $(1/2, 0) \oplus (0, 1/2)$ representation of the Lorentz group, but would have just half as many independent components as ψ^D above. It can be expressed as

$$\psi_a = \begin{pmatrix} \psi_{\mathrm{L}A} \\ (-\mathrm{i}\sigma_2\psi_{\mathrm{L}}^*)^{\dot{A}} \end{pmatrix}. \tag{4.9}$$

This object is the Majorana spinor that we have already encountered in Chapter 3, and the relation $\chi_{\mathrm{R}} = -\mathrm{i}\sigma_2\psi_{\mathrm{L}}^*$ is simply (3.3).[2] Since the Dirac spinor contains twice as many independent components as a Majorana spinor, it can be thought of as a combination of two Majorana spinors, in much the same way that we can think of a complex number as a combination of two real numbers.

Many textbooks and review articles use the more fundamental two-component spinor notation. Here, we formulate everything in terms of four-component spinors, which are perhaps more familiar to particle physicists interested in performing phenomenological calculations.

4.3 The Poincaré group

In addition to rotations and boosts, the other spacetime transformations include translations in space and time. Translations in space and time are generated by the energy–momentum operator P_μ, which can be represented by the differential operator $P_\mu = \mathrm{i}\partial_\mu$. The Poincaré group is formed by combining rotations, boosts, and translations. We then have *ten* independent generators: the six $M_{\mu\nu}$ plus the four P_μ. It is then straightforward to work out the commutation relations for the generators, using their representation as differential operators. One finds:

---

[2] In Chapter 3 and elsewhere, $\psi_{\mathrm{L,R}}$ is also used to denote the four-component spinor $P_{\mathrm{L,R}}\psi$ which has only two non-vanishing components in the representation where the matrix γ_5 is diagonal. These non-vanishing components are just the components of the two-spinor $\psi_{\mathrm{L,R}}$ discussed in this chapter. Although this is an abuse of notation, it should be clear from the context whether we are using $\psi_{\mathrm{L,R}}$ to denote four- or two-component spinors.

$$[P_\mu, P_\nu] = 0, \tag{4.10a}$$

$$[M_{\mu\nu}, P_\lambda] = \mathrm{i}(g_{\nu\lambda} P_\mu - g_{\mu\lambda} P_\nu), \tag{4.10b}$$

$$[M_{\mu\nu}, M_{\rho\sigma}] = -\mathrm{i}(g_{\mu\rho} M_{\nu\sigma} - g_{\mu\sigma} M_{\nu\rho} - g_{\nu\rho} M_{\mu\sigma} + g_{\nu\sigma} M_{\mu\rho}). \tag{4.10c}$$

To classify the representations of the Poincaré group, we again look for Casimir operators. One of these is the operator P^2, which certainly commutes with all the group generators. Its eigenvalue operating on particle state vectors is just the squared mass $P^2|\psi\rangle = m^2|\psi\rangle$. The other Casimir operator is obtained from the Pauli–Lubanski four-vector $W_\mu = \frac{1}{2}\epsilon_{\mu\nu\rho\sigma} P^\nu M^{\rho\sigma}$, with $W^\mu P_\mu = 0$. The square of the Pauli–Lubanski vector, W^2, can be shown to commute with all the generators of the Poincaré group. Notice also that in the rest frame (of a massive state) W^i is proportional to the rotation generator J^i. The various representations of the Poincaré group were first worked out by Wigner. The physically realized unitary representations which are of interest to us are:

- $P^2 \equiv m^2 > 0$, with $W^2 = -m^2 s(s+1)$, where s denotes the spin quantum number $s = 0, \frac{1}{2}, 1, \ldots$ Thus, these states correspond to particles of definite mass and discrete spin values.
- $P^2 = 0$, $W^2 = 0$ so that $W_\mu = \lambda P_\mu$. Here λ is the state helicity value, and $\lambda = \pm s$, for $s = 0, \frac{1}{2}, 1, \ldots$ These correspond e.g. to single particle states of massless particles such as photons with $\lambda = \pm 1$ or the graviton with $\lambda = \pm 2$.
- Finally, $P^\mu \equiv 0$, corresponding to the vacuum state which is invariant under Poincaré transformations.

In the 1960s, a number of papers were written about the possibility of embedding the spacetime symmetries (i.e. the Poincaré group) into some larger *master* group such as $SU(6)$ that would serve as a more general framework for the symmetry of the laws of physics. These efforts culminated in several no-go theorems, the most general of which was the Coleman–Mandula theorem.[3] It states the following.

Theorem Let G be a connected symmetry group of the S matrix (its generators commute with the S matrix), and assume the following.

- G contains a subgroup which is locally isomorphic to the Poincaré group (Poincaré invariance).
- All particle types correspond to positive energy representations of the Poincaré group. For any finite mass m, there are only a finite number of types of particles with mass less than m.

---

[3] S. Coleman and J. Mandula, *Phys. Rev.* **159**, 1251 (1967).

- Elastic scattering amplitudes are analytic functions of the Mandelstam variables s and t in some neighborhood of the physical region, except at normal thresholds, and the S matrix is non-trivial in the sense that essentially any two one-particle momentum states scatter, except perhaps for isolated values of s.
- Finally, a technical assumption: the generators of G, considered as integral operators in momentum space, have distributions as their kernels.

Coleman and Mandula asserted that if these conditions hold, *G is locally isomorphic to the direct product of a compact symmetry group and the Poincaré group.*

Stated more simply, under a number of physically reasonable assumptions, it is not possible to form a non-trivial merger of the Poincaré symmetry with other symmetries of the S matrix into a bigger group. It is not possible to have a larger spacetime symmetry and, further, internal symmetries such as local gauge symmetries or additional global symmetries (e.g. isospin) can only be realized as a direct product of these symmetry groups with the Poincaré group. It is intriguing that all the Poincaré group symmetries of the S matrix are in fact realized in nature. It is important to recognize that Coleman and Mandula did not envisage the possibility of anticommuting spinorial charges in their analysis. It is precisely the inclusion of these that allows us to enlarge the spacetime symmetry group to include supersymmetry, as we have already seen in Chapter 3.

4.4 The supersymmetry algebra

Our investigation of the Wess–Zumino model in Chapter 3 shows that it is possible to construct a relativistic quantum field theory that is invariant under supersymmetry transformations, for which the generators are anticommuting spinorial charges Q_a. We saw that the algebra of the Q_a's (this involved anticommutators, which is how the Coleman–Mandula theorem is circumvented) closes to yield P_μ, so that the supersymmetry is, in effect, a spacetime symmetry. In this sense, supersymmetry can be looked upon as a generalization of the special theory of relativity.

We had already worked out the algebra of the spinorial generators Q_a amongst themselves and with the translation generators in Chapter 3. Since this involves anticommutators of the super-charges, it is called a graded Lie algebra. The commutator of the Lorentz generators with the super-charges Q_a is simply given by the fact that these are spin $\frac{1}{2}$ objects. We can thus write the supersymmetric extension of the Poincaré algebra, known as the super-Poincaré algebra, as

$$[P_\mu, P_\nu] = 0, \tag{4.11a}$$

$$[M_{\mu\nu}, P_\lambda] = \mathrm{i}(g_{\nu\lambda} P_\mu - g_{\mu\lambda} P_\nu), \tag{4.11b}$$

$$[M_{\mu\nu}, M_{\rho\sigma}] = -\mathrm{i}(g_{\mu\rho} M_{\nu\sigma} - g_{\mu\sigma} M_{\nu\rho} - g_{\nu\rho} M_{\mu\sigma} + g_{\nu\sigma} M_{\mu\rho}), \tag{4.11c}$$

$$[P_\mu, Q_a] = 0, \tag{4.11d}$$

$$[M_{\mu\nu}, Q_a] = -(\tfrac{1}{2}\sigma_{\mu\nu})_{ab}Q_b, \tag{4.11e}$$

$$\{Q_a, \bar{Q}_b\} = 2(\gamma^\mu)_{ab}P_\mu. \tag{4.11f}$$

Since Q is a Majorana spinor charge, we can use the last of these relations to work out the anticommutators between two Q's or $\bar{Q}$'s.

Exercise: *Verify that*

$$\{Q_a, Q_b\} = -2(\gamma^\mu C)_{ab}P_\mu,$$
$$\{\bar{Q}_a, \bar{Q}_b\} = 2(C^{-1}\gamma^\mu)_{ab}P_\mu.$$

An extension of the Coleman–Mandula type analysis that allows for spinorial charges was worked out by Haag, Lopuszanski, and Sohnius who showed that the super-Poincaré algebra above is indeed the most general extension of the Poincaré algebra, provided we have just a single spinorial charge Q.[4] These authors also showed that theories with more than one spinorial generator are possible. These are referred to as extended supersymmetry theories. Such theories do not allow chiral representations which, as we know, are crucial for phenomenology. Only theories with a single spinorial generator Q_a, known as $N = 1$ supersymmetry theories, allow chiral representations. For this reason, we restrict our attention only to $N = 1$ supersymmetry.[5] To sum up, the Haag–Lopuszanski–Sohnius theorem tells us that the most general symmetry of the S matrix is the direct product of some internal symmetry with super-Poincaré invariance.

The irreducible representations of the super-algebra can be worked out as usual by finding the relevant Casimir operators. For the SUSY algebra above, the operator P^2 again commutes with all generators, so that all particles occurring in a supermultiplet will have the same mass. However, the square of the Pauli–Lubanski pseudovector W^2 is no longer a Casimir invariant, so that supermultiplets can now contain particles of differing spins. We will not discuss the construction of a new Casimir operator for this case, but instead focus on the particle supermultiplets that furnish representations of the super-Poincaré algebra.

For the massive case $P^2 \equiv m^2 > 0$, the representations are labeled by (m, j) with $j = 0, 1/2, 1, \ldots$ For fixed m, the complete supermultiplet contains a state each corresponding to spin $s = j \pm 1/2$, and two states with spin $s = j$ (except for the case $j = 0$ where the state with $s = j - 1/2$ is absent), all of which have the same mass. Notice that the number of helicity states for the two objects with spin j $[2(2j + 1)]$ is exactly balanced by the corresponding number of helicity states for the two states with spins $j \pm \frac{1}{2}$ $[2(j+1/2)+1+2(j-1/2)+1]$. This is just the statement that the number of bosonic and fermionic helicity states are the same. If $j=0$, we have the multiplet of the Wess–Zumino model – two spin zero states and two spin half states – as we discussed in Chapter 3. If $j=1/2$, there are four bosonic degrees of freedom (three spin 1 and one spin zero) balanced by four fermionic degrees of freedom corresponding to two Majorana spin half fermions as we will see when we study spontaneously broken gauge theories.

---

[4] R. Haag, J. Lopuszanski and M. Sohnius, *Nucl. Phys.* **B88**, 257 (1975).

[5] The underlying fundamental theory could be an extended supersymmetric theory, but all but one (or none!) of the supersymmetries must somehow be broken at much higher scales, leaving an $N = 1$ supersymmetric theory as the extension of the Standard Model that could have phenomenological relevance.

For the massless case, one can show that if j is the state with largest helicity in a supermultiplet, it is always accompanied by another state with helicity $j-1/2$. Furthermore, a Lorentz invariant field theory always contains these states together with their CPT conjugates which have opposite helicities, $-j$ and $-j+1/2$ and which are also massless. These states constitute a complete massless supermultiplet. If $j=1/2$, this multiplet consists of two fermionic states with helicities $\pm 1/2$, and a pair of spin zero bosonic states. This multiplet occurs in the massless Wess–Zumino model. Such a multiplet would also describe a massless neutrino and antineutrino together with its supersymmetric partners which would be two spin zero states (which can be regarded as quanta of one massless *complex* scalar field). For $j=1$, the bosonic states would correspond to a massless gauge boson (helicities ± 1); the fermionic partner states, which have helicity $\pm 1/2$, then correspond to a Majorana fermion referred to as a *gaugino*. This "gauge multiplet" is a crucial ingredient of supersymmetric gauge theories, which form the basis of all supersymmetry phenomenology. Finally, if $j=2$, we see that the two bosonic states have helicities ± 2. These are thought to correspond to a graviton, the massless spin two quantum that mediates gravity. The fermionic partners of these $j=2$ states have helicities $\pm 3/2$ and describe a spin $\frac{3}{2}$ massless Majorana fermion referred to as a *gravitino*. This "gravity multiplet" is essential for theories in which the parameter α of SUSY transformations depends on x^{μ}. These locally supersymmetric theories necessarily involve Einsteinian gravity, and are referred to as *supergravity* theories.

20 04 2020

G) Books_70_2020/
Weak Scale Supersymmetry from Superfields...|
0521857864.pdf

Pag 41(59)

Chapter 4

the supersymmetry algebra

4.1 Rotations

In classical mechanics, rotations of three-vectors can be represented by a rotation matrix R acting upon vectors such as $x = (x, y, z)$ as

$$x_i \to x'_i = R_{ij} x_j .$$

148) 20200421

A) Books_70_2020/
Gravitation and gauge symmetries
Gravitation and gauge symmetries
(2001...).pdf

Pag. 13(26) Figure 1.2.

Vector components change under the parallel transport.

On a general surface we cannot set up

: " " ' /
L2 & '% ($' ! ! L2 E 6
!

! ! " " # # $ % #& ' $ #
! ! ! " " # #($&
)*+,)))-) #& . / ! 0
1 2 0 3 4 5! ' $ 6 4

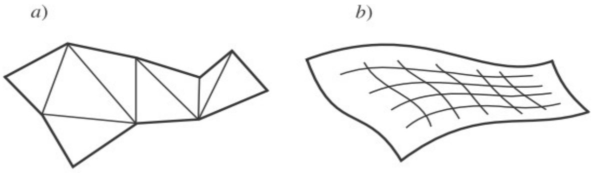

Figure 1.1. Approximate realization (a) of a curved surface (b).

trying to test spacetime geometry by physical measurements. His/her conclusions will also be valid locally for true gravitational fields, on the basis of the PE. Assume that the observer sits at the centre of the disc, and has two identical clocks: one of them is placed at the centre, the other at some point on the periphery. The observer will see that the clock on the periphery is running slower (time dilatation from SR). Consequently, clocks at various positions in the gravitational field run faster or slower, depending on the local strength of the field. There is no definition of time that is pertinent to the whole spacetime in general. The observer will also conclude that the length of a piece of line orthogonal to the radius of the disc will be shortened (length contraction from SR). Therefore, the ratio of the circumference of the circle to its radius will be smaller than 2π, so that the Euclidean geometry of space does not hold in GR.

Geometry of curved surfaces. We have seen that in spacetime, within a limited region, we can always choose a suitable reference frame, called the local inertial frame. Taking spacetime apart into locally inertial components, we can apply the laws of SR in each such component and derive various dynamical conclusions. Reconstruction of the related global dynamical picture, based on the PE (and some additional, simple geometric assumptions), gives rise to GR.

This procedure for dissecting spacetime into locally inertial components, out of which we can reconstruct, using the PE, the global structure of spacetime containing the gravitational field, can be compared geometrically with an attempt to build a curved surface, approximately, from a lot of small, plane elements ‘continuously’ bound to each other (figure 1.1). A locally flat surface is a geometric analogue of locally inertial spacetime.

GR makes essential use of curved spaces (or, more accurately curved spacetimes). Since the curved, four-dimensional spacetime cannot be visualized, let us try to understand the basic features of curved spaces by considering a two-dimensional surface X_2. We shall focus our attention on those geometric properties of X_2, which could be determined by an intelligent, two-dimensional being entirely confined to live and measure in the surface, a being that does not know anything about the embedding Euclidean space E_3. Such properties determine the *intrinsic geometry* on X_2.

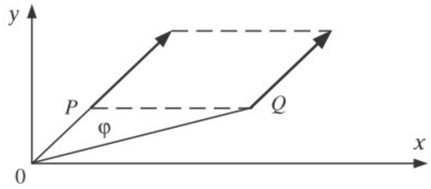

Figure 1.2. Vector components change under the parallel transport.

On a general surface we cannot set up Cartesian coordinates; the best we can do is to introduce two arbitrary families of coordinate lines and label them by $u^\alpha = (u^1, u^2)$. These curvilinear (or Gaussian) coordinates do not always have a direct geometric interpretation.

Let us now imagine a surface X_2 embedded in E_3, and consider a point P in X_2 with Cartesian coordinates (x, y, z) in E_3; since P is in X_2, these coordinates can be expressed as functions of u^α. The squared distance between two neighbouring points in X_2, $P(x, y, z)$ and $Q(x+\mathrm{d}x, y+\mathrm{d}y, z+\mathrm{d}z)$, is given by $\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2$. Going over to u^α we obtain

$$\mathrm{d}s^2 = g_{\alpha\beta}\,\mathrm{d}u^\alpha\,\mathrm{d}u^\beta \tag{1.6}$$

where the set of functions $g_{\alpha\beta}(u^1, u^2)$ defines the *metric* on X_2. The metric is an essential intrinsic property of X_2, independent of its embedding in E_3. A two-dimensional continuum X_2, equipped with a squared differential distance according to (1.6), becomes the *metric space*, $G_2 = (X_2, \boldsymbol{g})$ (it is now clear that the symbol X_2 is used to denote an abstract two-dimensional continuum, disregarding all metric properties of the real surface).

Studying the motion of particles in X_2 our observer could have arrived at the idea of the tangent vector, $\boldsymbol{a} = (a^1, a^2)$. Consider, further, two tangent vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, defined at neighbouring points P and Q. If we wish to compare these two vectors we have to know how one of them can be transported to the position of the other. This transport is called *parallel transport*. The continuum X_2 together with the parallel transport law, denoted by Γ, is called a *linearly connected* (or *affine*) space, $L_2 = (X_2, \Gamma)$.

The components of a vector in curvilinear coordinates are expected to change under parallel transport. This can be easily seen by considering the parallel transport of a vector in a plane, using polar coordinates (r, θ). As shown in figure 1.2, if the unit vector $\boldsymbol{a}$ at point P has components $(a^r = 1, a^\theta = 0)$, after parallel transport from P to Q its components become $(a^r = \cos\varphi, a^\theta = \sin\varphi)$.

A two-dimensional observer should determine the rule of parallel transport of vectors in accordance with his/her own experience. This rule is generally independent of the concept of a metric. If X_2 is equipped with a rule of

parallel transport and the independent concept of a metric, we obtain the *linearly connected metric* space $(X_2, \Gamma, \boldsymbol{g}) = (L_2, \boldsymbol{g}) = (G_2, \Gamma)$.

It is very probable that our two-dimensional observer will demand that the lengths of vectors should remain unchanged under parallel transport. After adopting this very natural property, which relates Γ and $\boldsymbol{g}$, the linearly connected metric space becomes the *Riemann–Cartan* space U_2. The usual surfaces have some additional geometric structure, and belong to the class of *Riemann* spaces V_2.

This discussion was slightly more general than strictly necessary for studying surfaces embedded in E_3, in order to emphasize that the intrinsic geometry on X_2 is defined in terms of

- the metric $\boldsymbol{g}$ and
- the rule of parallel transport Γ.

These basic concepts of the intrinsic geometry of surfaces can be extended directly to spaces with higher dimensions, although in this case we lose the intuitive geometric picture.

Any Riemann space is locally Euclidean, i.e. its metric is positive definite ($\mathrm{d}s^2 > 0$). A slight generalization of this case, that consists in admitting metrics that are pseudo-Euclidean, leads to a pseudo-Riemannian geometry or Riemannian geometry with an indefinite metric.

Relativity, covariance and Mach's ideas

Relativity and covariance. We have seen previously that inertial frames play a particularly important role in SR. Einstein considered this to be an expression of the incompleteness of SR. He wanted to generalize the relativity of inertial motions to the relativity of all motions, including accelerated ones. This idea can be formulated as the general principle of relativity:

> *General PR: The form of physical laws is the same in all reference frames.*

Equations expressing physical laws are often formulated using specific coordinates. Thus, for instance, spatial positions in an inertial laboratory L can be determined using Cartesian coordinates (x, y, z) or the coordinates obtained from these by a rotation around the z-axis for a fixed angle φ. If we allow the angle φ to change with time as $\varphi = \omega t$, new coordinates can be naturally attached to a new physical reference frame, rotating relative to L with the angular velocity ω around the z-axis. We see that coordinate transformations can be interpreted as describing transitions from one to another laboratory (the so-called active interpretation). Thus, it becomes clear that the general PR can be realized with the help of the principle of general covariance:

> *General covariance: The form of physical laws does not depend on the choice of coordinates.*

General covariance is a technical way to express the general PR.

When mathematicians study the geometric properties of space, they use certain geometric objects with properties that do not depend on the choice of coordinates. The covariance means that physical laws can be expressed in terms of these geometric objects. In classical physics such objects are tensors, so that covariant equations must be tensorial.

General covariance seems to be a powerful principle which determines acceptable forms of physical equations. Is it really so? The next example shows a certain triviality of this principle. Consider Newton's second law in an inertial frame S, $m\boldsymbol{a} = \boldsymbol{F}$, which is invariant under the group of Galilean transformations G. Going over to a non-inertial frame $\widetilde{S}$ that has linear acceleration $\boldsymbol{a}_0$ and angular velocity $\boldsymbol{\omega}$ relative to S, Newton's equation transforms into

$$m\boldsymbol{a} = \boldsymbol{F} - m\boldsymbol{a}_0 - 2m\boldsymbol{\omega} \times \boldsymbol{v} - m\dot{\boldsymbol{\omega}} \times \boldsymbol{r} - m\boldsymbol{\omega} \times (\boldsymbol{\omega} \times \boldsymbol{r}). \tag{1.7}$$

The transition from S to $\widetilde{S}$ can be described as a Galilean transformation (a spatial translation combined with a rotation) with time-dependent parameters. The set of these transformations forms a group, denoted as $\widetilde{G}$, which is larger than the Galilean group G. In this way we obtained the $\widetilde{G}$-covariant form of Newton's equation. Other physical equations can also be covariantized in a similar manner, which means that covariance implies that there are no stringent conditions on the form of the physical equations.

It is important to note that equation (1.7) is completely determined by its original form in S and the process of covariantization. The original equation has Galilean symmetry, and the covariantization 'knows' of this property, since it is based on $\widetilde{G}$ transformations. Using modern language, we can describe this procedure as an incomplete *localization* of Galilean symmetry, based on time-dependent parameters. Since the procedure does not depend on the form of the original equation but only on its global symmetry, it is clear that covariantization does not restrict the form of physically acceptable equations.

In spite of this conclusion, we know that Einstein's covariant equations have experimental consequences that are essentially different from those of Newton's theory of gravity, so that general covariance still seems to be an important principle for physics. This dilemma can be resolved in the following way. If we restrict ourselves to weak gravitational fields for simplicity, we obtain an approximate, non-covariant form of Einstein's theory, known as the Pauli–Fierz field theory in M_4. Comparing this field theory with Newton's action-at-a-distance theory we easily find significant differences; in particular, these theories have different global symmetries (Poincaré and Galilean, respectively). It is therefore quite natural that their physical contents are essentially different. This difference has nothing to do with the covariance of Einstein's equations.

After clarifying this point, we are now ready to conclude that general covariance by itself does not have any physical content. However, we should note that Einstein related covariance to the PE, so that general covariance (transition to accelerated frames) represents an important technical procedure for *introducing the gravitational field* into any physical theory. Thus, Einstein's theory of gravity is obtained by unifying the principle of general covariance with the PE:

General covariance + PE ⇒ theory of gravity.

Here, general covariance stands essentially for the general PR, while the PE tells us not only how to introduce gravity, but also the kind of physics we have in the absence of gravity.

General covariance is only a statement about the effects of gravity, it does not imply the validity of SR in locally inertial frames. Although the local validity of SR is usually assumed, theoretically it is not necessary; for instance, we could assume Galilean relativity in locally inertial frames. This is particularly clear if we think of gravity as a locally invariant theory.

Covariance and symmetry. Although the concepts of general covariance and symmetry have certain formal similarities, they are, in fact, essentially different. In order to illustrate this difference, consider the interval in $G_4 = (X_4, \boldsymbol{g})$,

$$\mathrm{d}s^2 = g_{\mu\nu}(x)\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu \tag{1.8a}$$

which is expressed in some coordinate system K. This equation is covariant with respect to general coordinate transformations $x \to x' = x'(x)$, since it has the same form in the new coordinate system K' as it had in K:

$$\mathrm{d}s^2 = g'_{\mu\nu}(x')\,\mathrm{d}x^{\mu'}\,\mathrm{d}x^{\nu'}. \tag{1.8b}$$

Here, the coordinate transformation induces a related functional change in the metric, $g_{\mu\nu} \to g'_{\mu\nu}$, where the prime denotes a new function. Now, if we restrict ourselves to those transformations that do not change the form of the metric, $g'_{\mu\nu}(x) = g_{\mu\nu}(x)$, we obtain symmetry transformations (or isometries) of the space G_4.

To see the meaning of this definition in Minkowski space, let us consider equation (1.8a) in an inertial frame S, where $g_{\mu\nu} = \eta_{\mu\nu}$ is the metric of M_4. If we look for coordinate transformations that do not change the form of the metric, we obtain Poincaré transformations. They express the symmetry of physical laws under transition from one inertial frame to another in M_4.

Mach's principle. In his attempts to build GR, Einstein was led by the idea of abolishing the privileged role of inertial frames from physics. Even if it were possible to develop a satisfactory theory of gravity within SR, he would not have stopped there—he had to go beyond SR. On this route he was greatly influenced by the ideas of the philosopher E Mach (1836–1916), who gave the first constructive critique of Newton's understanding of inertia and absolute space, and proposed a new approach to these questions.

The conclusion drawn by Newton on the basis of his experiment with a rotating vessel (that the acceleration of the water with respect to the vessel does not produce inertial forces, they are produced only by the acceleration of the water with respect to absolute space) cannot be considered as well founded. Newton's experiment only shows that the rotation of the water relative to a *small* vessel produces no noticeable inertial forces. It is not clear how the experiment would have turned out if the vessel had been more massive—these were the suspicions raised by Mach. He believed that water rotating relative to a very massive vessel would remain flat, since it would be at rest with respect to the very close and massive vessel, while the water that does not rotate would be curved, since its motion with respect to the vessel would be accelerated. In other words, inertial forces are caused by the acceleration of the water relative to the average mass distribution in the universe, not with respect to absolute space.

We shall now illustrate the difference between Newton and Mach by another example. Consider an elastic sphere that is rotating relative to some inertial frame, getting larger at the equator. How does the sphere 'know' that it is rotating and hence must deform? Newton might have said that it was accelerated relative to absolute space, which thereby causes the inertial forces that deformed it. Mach would have said that the sphere was accelerated relative to distant masses in the universe, which was the cause of deformation.

It is interesting to observe that both Newton and Mach explained the appearance of inertial forces by the *acceleration* of a body, not by its velocity. Why is that so? Mach's considerations were general in nature, he was not able to give any quantitative description of these ideas. The inertial influence of cosmic matter was clearly defined in Einstein's GR (inertia is a manifestation of the geometry of spacetime, geometry is affected by the presence of matter) which, to some extent, realizes Mach's abstract ideas in a concrete physical theory.

Mach's principle can be roughly expressed in the form of the following (not completely independent) statements:

- Space as such plays no role in physics; it is merely an abstraction from the totality of spatial relations between material objects.
- The inertial properties of every particle are determined by its interaction with all the other masses in the universe.
- Any motion is relative and can be determined only with respect to all the other masses in the universe.
- Inertial forces are caused by an acceleration of a body relative to the average mass distribution in the universe.

It is often believed that Mach's principle is so general that it cannot be checked experimentally so that, essentially, it has no physical content. Yet, there are examples showing clearly that this is not so. In spite of this, there is still no direct experimental verification (Rindler 1977).

Bearing in mind the strong influence of Mach's principle on Einstein, it is interesting to clarify to what extent this principle has been justified by GR. We can say that GR has realized only some of Mach's ideas; indeed, solutions of Einstein's field equations are determined not only by matter distribution, but also by boundary conditions. Hence, the same mass distribution may give rise to different gravitational fields. Mach wanted to abolish space in its own right and replace it with the relative configuration of matter; Einstein retained spacetime but made it non-absolute. Although Mach's principle helped stimulate the development of GR, there is no reason today to consider it as a basic principle of physics.

Perspectives of further developments

Einstein's theory of gravity predicted a number of physical effects which can be completely verified experimentally. Nevertheless, we should stress that certain properties of this theory deserve critical analysis.

GR is characterized by the existence of *singularities*, which are generic features of solutions describing both localized physical systems (black holes) and cosmology (big bang). A spacetime singularity can be intuitively characterized by an infinite growth of some physical quantities in some regions of spacetime and the analogous pathological behaviour of the related geometric objects. Both cosmology and black holes represent very interesting phenomena in gravitational physics—their features cast light on the intrinsic limitations of GR and serve as a signpost towards a new, consistent approach to gravitation.

The standard cosmological model predicts that at some finite time in the past, the universe was in a singular state (big bang). For many years it was generally believed that this prediction was due merely to the simplifying assumptions of the model. However, the singularity theorems of GR have shown that singularities are generic features of cosmological solutions. We expect that the singular behaviour can be avoided either by modifying classical GR and/or by taking quantum effects into account.

General singularity theorems also apply to black holes—gravitational objects that inevitably arise at the endpoint of the thermonuclear evolution of some stars. After a finite proper time all the matter of the star meets at one point, where the density of matter and the curvature of spacetime become infinite.

A cosmological singularity means that there existed a moment of time in the past, such that nothing before that moment, even time itself, did not have any sense, did not exist. The singularity of the black hole is related to the future, and means that for an observer falling into the black hole, there will come a moment at which not only the observer's life will end, but also time itself will go out of existence. In these extreme conditions, where the density of matter becomes very large and the predictions of GR contradict the fundamental concepts of classical

20 04 2020

G) Books_70_2020/
Weak Scale Supersymmetry from
Superfields...|

0521857864.pdf

Pág 41(59)

Chapter 4

the supersymmetry algebra

4.1 Rotations

In classical mechanics, rotations of three-vectors can be represented by a rotation matrix R acting upon vectors such as $x = (x, y, z)$ as

$$x_i \to x'_i = R_{ij} \underline{x_j}.$$

148) 20200421

A) Books_70_2020/
Gravitation and gauge symmetries
Gravitation and gauge symmetries
(2001...)...pdf

Pag. 13(26) Figure 1.2.

Vector components change under the parallel transport.

On a general surface we cannot set up

21.04.2020

Cartesian coordinates. The best we can do is to introduce two arbitrary of coordinate lines and label them by $u^\alpha = (u^1, u^2)$. These curvilinear (or Gaussian) coordinates do not always a direct geometric interpretation.

obs. Use este livro.

A2) Books_70_2020/
Gravitation and gauge symmetries/
String Gravity and Physics at the Planck Energy....pdf

Pag. 11(26)

Lectures on string theory in curved spacetimes

A3) Books_70_2020/
Gravitation and gauge symmetries/
The Mathematical theory of Cosmic ... pdf

Pag. 59(73) - Chapter 3

String dynamics in flat space

The Minkowski spacetime is the only background metric of cosmological interest in which the string equation of motion (2.14) can be integrated exactly.

)*+,)))-) #& . / 0 ! 0
8 9 0 ((61 5 : ! 5!
6 !

LECTURES ON STRING THEORY IN CURVED SPACETIMES

H.J. de Vega$^{(a)}$ **and N. Sánchez**$^{(b)}$

(a) Laboratoire de Physique Théorique et Hautes Energies, Université Pierre et Marie Curie (Paris VI) et Université Denis Diderot (Paris VII), Tour 16, 1er. étage, 4, Place Jussieu 75252 Paris, cedex 05, France. Laboratoire Associé au CNRS URA280.

(b) Observatoire de Paris, DEMIRM, 61, Avenue de l'Observatoire, 75014 Paris, France. Laboratoire Associé au CNRS URA336, Observatoire de Paris et École Normale Supérieure.

Abstract

Recent progress on string theory in curved spacetimes is reviewed. The string dynamics in cosmological and black hole spacetimes is investigated. The different methods available to solve the string equations of motion and constraints in curved spacetimes are described. That is, the string perturbation approach, the null string approach, the τ-expansion, and the construction of global solutions (for instance by inverse scattering methods).

The classical behaviour of strings in FRW and inflationary spacetimes is now understood in a large extent from the various types of explicit string solutions. Three different types of behaviour appear: **unstable, dual** to unstable and **stable**. For the unstable strings, the energy and size grow for large scale factors $R \to \infty$, proportional to R. For the dual to unstable strings, the energy and size blow up for $R \to 0$ as $1/R$. For stable strings, the energy and proper size are bounded. (In Minkowski spacetime, all string solutions are of the stable type).

Recent progress on self-consistent solutions to the Einstein equations for string dominated universes is reviewed. The energy-momentum tensor for a gas of strings is then considered as source of the spacetime geometry and from the above string behaviours the string equation of state is **derived**. The self-consistent string solution exhibits the realistic matter dominated behaviour $R \simeq T^{2/3}$ for large times and the radiation dominated behaviour $R \simeq T^{1/2}$ for early times (T being the cosmic time).

We report on the **exact integrability** of the string equations plus the constraints in de Sitter spacetime that allows to systematically find **exact** string solutions by soliton methods and the multistring solutions. **Multistring solutions** are a new feature in curved spacetimes. That is, a single world-sheet simultaneously describes many different and independent strings. This phenomenon has no analogue in flat spacetime and follows to the coupling of the strings with the geometry.

Finally, the string dynamics next and inside a Schwarzschild black hole is analyzed and their physical properties discussed.

N. Sánchez and A. Zichichi (eds.), String Gravity and Physics at the Planck Energy Scale, 11–63.
© 1996 *Kluwer Academic Publishers.*

21 04 2020

Cartesian coordinates. The best we can do is to introduce two arbitrary of coordinate lines and label them by $\mu^\alpha = (\mu^1, \mu^2)$. These curvilinear (or Gaussian) coordinates do not always a direct geometric interpretation.

obs. Use este livro.

A2) Books_70_2020/
Gravitation and gauge symmetries/
String Gravity and Physics at the
Planck Energy ....pdf

Pág. 11(26)

Lectures on string theory in curved spacetimes

A3) Books_70_2020/
Gravitation and gauge symmetries/
The Mathematical theory of
Cosmic ... pdf

Pág. 59(73) - Chapter 3

String dynamics in flat space

The Minkowski spacetime is the only background metric of cosmological interest in which the string equation of motion (2.14) can be integrated exactly.

$)*+,)))-) #& 2 2 0
1 2 0 3 4 5!
(;$ 5 $

Chapter 3

String dynamics in flat space

The Minkowski spacetime is the only background metric of cosmological interest in which the string equation of motion (2.14) can be integrated exactly. It is, therefore, not surprising that most work on string dynamics has been done in the Minkowski background, and that string dynamics in flat space is reasonably well understood. Only the Robertson–Walker background is of comparable importance and there little can be done without resort to approximations or numerical techniques (see section 5.1).

In this chapter the Minkowski version of the string equation of motion will be examined in some detail. The various laws relating to conservation of momentum, angular momentum and invariant length will be formulated explicitly and the inter-relationship between the kinks, cusps and self-intersections of a string loop will be analysed with the aid of a particular representation known as the Kibble–Turok sphere. In addition, a number of methods for generating exact periodic solutions will be presented.

3.1 The aligned standard gauge

In what follows, it is assumed that the Minkowski metric tensor $\eta_{\mu\nu}$ has its Cartesian form $\mathrm{diag}(1, -1, -1, -1)$ and that the parametrization $x^\mu = X^\mu(\tau, \sigma)$ of the world sheet has been chosen so that τ and σ are standard-gauge coordinates (see section 2.3). As mentioned in the previous chapter, there is no unique choice for the gauge coordinates in the standard gauge. One simple way to eliminate the residual gauge freedom is to identify τ with the Minkowski time coordinate $x^0 = t$. I will call this choice the *aligned standard gauge*.

Whichever form of the standard gauge is prescribed, the equation of motion (2.31) in a Minkowski background reads:

$$X^\mu{}_{,\tau\tau} = X^\mu{}_{,\sigma\sigma} \tag{3.1}$$

subject to the gauge conditions

$$X_\tau^2 + X_\sigma^2 = 0 \qquad \text{and} \qquad X_\tau \cdot X_\sigma = 0. \tag{3.2}$$

The general solution to (3.1) is

$$X^\mu(\tau,\sigma) = A^\mu(\tau+\sigma) + B^\mu(\tau-\sigma) \tag{3.3}$$

where A^μ and B^μ are arbitrary functions of their arguments, save for the gauge constraints

$$A'^2 = B'^2 = 0 \tag{3.4}$$

where a prime here denotes the derivative with respect to the relevant argument. The trajectory $X^\mu(\tau,\sigma)$ is thus a linear superposition of left- and right-moving modes which propagate along the string at the speed of light.

If the gauge coordinates (τ,σ) are not initially aligned with the time coordinate t, this can easily be achieved by defining a new pair of standard-gauge coordinates $(\overline{\tau},\overline{\sigma})$ which satisfy the equations

$$A^0(\tau+\sigma) = \tfrac{1}{2}(\overline{\tau}+\overline{\sigma}) \qquad \text{and} \qquad B^0(\tau-\sigma) = \tfrac{1}{2}(\overline{\tau}-\overline{\sigma}). \tag{3.5}$$

Then, after dropping the bars, the general solution (3.3) becomes

$$X^\mu(\tau,\sigma) = [\tau,\ \tfrac{1}{2}\mathbf{a}(\tau+\sigma) + \tfrac{1}{2}\mathbf{b}(\tau-\sigma)] \tag{3.6}$$

where **a** and **b** are 3-vector functions with the property that

$$\mathbf{a}'^2 = \mathbf{b}'^2 = 1. \tag{3.7}$$

The functions $\mathbf{a}'$ and $\mathbf{b}'$, therefore, lie on the surface of a unit sphere, first introduced by Tom Kibble and Neil Turok in 1982 [KT82] and as a result often referred to as the *Kibble–Turok sphere*[1]. The implications of the Kibble–Turok sphere for the dynamics of a string loop will be considered in more detail in sections 3.6–3.9.

If the string is closed and contains only one connected component (so that the world sheet **T** has strict cylindrical topology), then in the aligned standard gauge lines of constant τ form closed loops on **T**. The function $X^\mu(\tau,\sigma)$ is, therefore, periodic in σ with some constant gauge period L, so that

$$X^\mu(\tau,0) = X^\mu(\tau,L) \tag{3.8}$$

for all values of τ. In terms of the mode functions **a** and **b**, the periodicity condition (3.8) reads:

$$\mathbf{a}(\tau) + \mathbf{b}(\tau) = \mathbf{a}(\tau+L) + \mathbf{b}(\tau-L). \tag{3.9}$$

---

[1] It should be mentioned that many authors have chosen to use a gauge convention in which **a** is a function of $\sigma-\tau$ and **b** a function of $\sigma+\tau$. The principal drawback of this convention is that cusps occur when $\mathbf{a}'$ and $\mathbf{b}'$ are anti-parallel rather than parallel, which, in turn, complicates discussion of the Kibble–Turok representation. In what follows, I will adhere to the convention of (3.6) throughout, transforming previously published solutions into this form where necessary.

Since the tangent vector X^{μ}_{σ} is also a periodic function of σ, it follows that

$$\mathbf{a}'(\tau) - \mathbf{b}'(\tau) = \mathbf{a}'(\tau + L) - \mathbf{b}'(\tau - L) \tag{3.10}$$

as well. The most general solution of these two equations is

$$\mathbf{a}(\tau) = \widehat{\mathbf{a}}(\tau) + \mathbf{V}\tau \qquad \text{and} \qquad \mathbf{b}(\tau) = \widehat{\mathbf{b}}(\tau) + \mathbf{V}\tau \tag{3.11}$$

where $\mathbf{V}$ is a constant vector, and $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{b}}$ are periodic functions with period L satisfying the relations

$$(\widehat{\mathbf{a}}' + \mathbf{V})^2 = (\widehat{\mathbf{b}}' + \mathbf{V})^2 = 1. \tag{3.12}$$

As will be seen shortly, the vector $\mathbf{V}$ can be interpreted as the bulk velocity of the string loop, while the gauge period L is related to the energy of the loop.

In terms of the periodic mode functions $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{b}}$, the string trajectory takes the form

$$X^{\mu}(\tau, \sigma) = [\tau, \tfrac{1}{2}\widehat{\mathbf{a}}(\tau + \sigma) + \tfrac{1}{2}\widehat{\mathbf{b}}(\tau - \sigma) + \mathbf{V}\tau]. \tag{3.13}$$

Although the functions $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{b}}$ each have period L, the entire trajectory, in fact, has period $L/2$, as was first pointed out by Kibble and Turok [KT82]. To see this, note that

$$\begin{aligned} X^{\mu}&(\tau + L/2, \sigma + L/2) \\ &= [\tau + L/2, \tfrac{1}{2}\widehat{\mathbf{a}}(\tau + \sigma + L) + \tfrac{1}{2}\widehat{\mathbf{b}}(\tau - \sigma) + \mathbf{V}(\tau + L/2)] \\ &= [\tau + L/2, \tfrac{1}{2}\widehat{\mathbf{a}}(\tau + \sigma) + \tfrac{1}{2}\widehat{\mathbf{b}}(\tau - \sigma) + \mathbf{V}(\tau + L/2)] \end{aligned} \tag{3.14}$$

which is simply the original configuration (3.13) translated by an amount $L\mathbf{V}/2$.

3.2 The GGRT gauge

An alternative method of fixing a choice of standard gauge, first formulated by Goddard, Goldstone, Rebbi and Thorn in 1973 [GGRT73], is to choose a constant timelike or null vector u^{μ} and define

$$\tau = \frac{u \cdot X}{2u \cdot P} \tag{3.15}$$

where P^{μ} is the constant total 4-momentum of the string. If u^{μ} is timelike, the effect is to set τ equal to $\frac{1}{2E'}t'$, where t' is the time coordinate and E' the energy of the string in a frame with 4-velocity parallel to u^{μ}. Goddard *et al* [GGRT73] themselves preferred a gauge choice in which u^{μ} is null, as this leads to some simplifications when quantizing the relativistic string, which was their primary concern. The resulting gauge is what was originally meant by the term ‘light-cone gauge’ or (less commonly) ‘light-front gauge’. I will refer to it simply as the *GGRT gauge*.

Since the total 4-momentum P^μ of an infinite string is undefined, it is often preferable to replace the factor $2u \cdot P$ in (3.15) with some fixed constant. Following Christopher Thompson [Tho88], who revived the GGRT gauge in 1988, I will replace $2u \cdot P$ with 1 and make the choice

$$u^\mu = [1, 0, 0, -1] \tag{3.16}$$

so that

$$\tau = t + z \tag{3.17}$$

where z is the third spatial component of X^μ.

As in the aligned standard gauge, the general solution (3.3) to the equation of motion consists of a linear superposition of left- and right-moving modes, which in the present case will be written as

$$X^\mu(\tau, \sigma) = [t_{\mathrm{L}}(\tau+\sigma)+t_{\mathrm{R}}(\tau-\sigma), \mathbf{p}(\tau+\sigma)+\mathbf{q}(\tau-\sigma), z_{\mathrm{L}}(\tau+\sigma)+z_{\mathrm{R}}(\tau-\sigma)] \tag{3.18}$$

where $\mathbf{p}$ and $\mathbf{q}$ are 2-vector functions. The GGRT gauge choice $\tau = t + z$ implies that

$$t'_{\mathrm{L}} + z'_{\mathrm{L}} = \tfrac{1}{2} \qquad \text{and} \qquad t'_{\mathrm{R}} + z'_{\mathrm{R}} = \tfrac{1}{2} \tag{3.19}$$

and so, after some algebraic manipulation, the gauge constraints (3.2) read:

$$t'_{\mathrm{L}} - z'_{\mathrm{L}} = 2\mathbf{p}'^2 \qquad \text{and} \qquad t'_{\mathrm{R}} - z'_{\mathrm{R}} = 2\mathbf{q}'^2. \tag{3.20}$$

Solving the last two sets of equations simultaneously gives

$$t'_{\mathrm{L}} = \tfrac{1}{4} + \mathbf{p}'^2 \qquad t'_{\mathrm{R}} = \tfrac{1}{4} + \mathbf{q}'^2$$

and

$$z'_{\mathrm{L}} = \tfrac{1}{4} - \mathbf{p}'^2 \qquad z'_{\mathrm{R}} = \tfrac{1}{4} - \mathbf{q}'^2. \tag{3.21}$$

These restrictions, which take the place of the constraints $\mathbf{a}'^2 = \mathbf{b}'^2 = 1$ in the aligned standard gauge, have the specific advantage of being linear in the t–z mode functions. Thus, for example, a solution containing only a finite number of harmonics in $\sigma_\pm = \tau \pm \sigma$ can be generated by simply choosing the functions $\mathbf{p}$ and $\mathbf{q}$ to be finite harmonic series and integrating (3.21). The corresponding procedure in the aligned standard gauge is far more complicated (see section 3.5).

However, if the string forms a closed loop then the equations (3.21) need to be supplemented by the condition

$$\int_0^L (t'_{\mathrm{L}} - t'_{\mathrm{R}})\,\mathrm{d}\sigma = \int_0^L (z'_{\mathrm{L}} - z'_{\mathrm{R}})\,\mathrm{d}\sigma = 0 \tag{3.22}$$

or, equivalently,

$$\int_0^L \mathbf{p}'^2\,\mathrm{d}\sigma_+ = \int_0^L \mathbf{q}'^2\,\mathrm{d}\sigma_- \tag{3.23}$$

which, to a large extent, vitiates the advantage of the GGRT gauge in generating finite-harmonic solutions.

Another disadvantage of the GGRT gauge is that the coordinates τ and σ are singular at a cusp if the cusp velocity is parallel to u^μ (as $u \cdot X_\tau = 1$ everywhere but $u \cdot X_\tau = 0$ at a cusp in any non-singular gauge). Furthermore, solutions in the GGRT gauge trace out curves on the surface of the paraboloid $t - x^2 - y^2 = \frac{1}{4}$ rather than on the unit sphere. The resulting geometric description is typically harder to visualize and so has less heuristic value than the Kibble–Turok sphere.

Finally, it should be noted that for any particular choice of u^μ the GGRT gauge is incompatible with all standard-gauge solutions (3.3) in which either $u \cdot A$ or $u \cdot B$ is identically zero, as in such cases there is no reparametrization $\{\overline{\sigma}_+(\sigma_+), \overline{\sigma}_-(\sigma_-)\}$ that can transform $u \cdot X$ into a function of $\overline{\tau}$ alone. Since A'^μ and B'^μ are both null vectors the (unwelcome) possibility that either $u \cdot A = 0$ or $u \cdot B = 0$ is consistent with the gauge conditions if u^μ is null but not if u^μ is timelike. An example of this problem will be encountered in section 5.6.

Despite these caveats, the use of the GGRT gauge has helped to generate a number of simple solutions that would otherwise have remained undiscovered (see [Tho88]).

3.3 Conservation laws in flat space

Because the Minkowski spacetime has maximal symmetry, the components of the world-sheet momentum currents P^A_μ defined in section 2.4 are all separately conserved. The contravariant components of the momentum currents are

$$P^{\mu A} = \mu \gamma^{1/2} \gamma^{AB} X^\mu{}_{,B} \tag{3.24}$$

where, in the aligned standard gauge,

$$X^\mu_\tau = [1, \tfrac{1}{2}\mathbf{a}' + \tfrac{1}{2}\mathbf{b}'] \qquad X^\mu_\sigma = [0, \tfrac{1}{2}\mathbf{a}' - \tfrac{1}{2}\mathbf{b}'] \tag{3.25}$$

$$\gamma^{1/2} = X^2_\tau = \tfrac{1}{2}(1 - \mathbf{a}' \cdot \mathbf{b}') \tag{3.26}$$

and

$$\gamma^{AB} = \gamma^{-1/2}\, \mathrm{diag}(1, -1). \tag{3.27}$$

The local Lorentz factor of the string is $\lambda = \gamma^{-1/4}$, which diverges (and so indicates the formation of a cusp) whenever $\mathbf{a}' \cdot \mathbf{b}' = 1$.

An observer at rest with respect to the spacetime coordinates would measure the local 4-momentum at any point on the string to be

$$p^\mu = \lambda \mu X^\mu_\tau = \gamma^{-1/4} \mu [1, \tfrac{1}{2}\mathbf{a}' + \tfrac{1}{2}\mathbf{b}'] \tag{3.28}$$

and the element of proper distance along the string as

$$\mathrm{d}\ell = (-X^2_\sigma)^{1/2}\, \mathrm{d}\sigma = \gamma^{1/4}\, \mathrm{d}\sigma. \tag{3.29}$$

Hence, the total energy ΔE and momentum $\Delta\mathbf{p}$ of a segment of the string covering the parameter range σ to $\sigma + \Delta\sigma$ as measured by such an observer are

$$\Delta E = \int p^0 d\ell = \mu\Delta\sigma \qquad \text{and} \qquad \Delta\mathbf{p} = \int \mathbf{p}\,\mathrm{d}\ell = \tfrac{1}{2}\mu(\Delta\mathbf{a} - \Delta\mathbf{b}). \quad (3.30)$$

Note, in particular, that the local energy content of the string is proportional to σ in the aligned standard gauge.

The total 4-momentum P^μ of a string loop with parametric period L is most conveniently calculated by integrating $P^{\mu A}$ around a curve of constant τ, as in equation (2.48), so that

$$P^\mu = \mu \int_0^L X^\mu_\tau \,\mathrm{d}\sigma = \mu \int_0^L [1, \tfrac{1}{2}\widehat{\mathbf{a}}'(\tau+\sigma) + \tfrac{1}{2}\widehat{\mathbf{b}}'(\tau-\sigma) + \mathbf{V}]\,\mathrm{d}\sigma. \quad (3.31)$$

Since the mode functions $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{b}}$ are periodic with period L, they make no contribution to P^μ, and so

$$P^\mu = \mu L[1, \mathbf{V}]. \quad (3.32)$$

Hence, μL is the total energy of the string as expected, and $\mathbf{V}$ can be interpreted as its bulk velocity.

A similar conclusion can be drawn from the dynamics of the centre-of-mass 4-vector $\bar{X}^\mu$ of the loop, which is defined by

$$\bar{X}^\mu(\tau) = L^{-1} \int_0^L X^\mu \,\mathrm{d}\sigma = L^{-1} \int_0^L [\tau, \tfrac{1}{2}\widehat{\mathbf{a}}(\tau+\sigma) + \tfrac{1}{2}\widehat{\mathbf{b}}(\tau-\sigma) + \mathbf{V}\tau]\,\mathrm{d}\sigma. \quad (3.33)$$

In view of the periodicity of $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{b}}$, this equation can be expressed more compactly as

$$\bar{X}^\mu(\tau) = [\tau, \bar{\mathbf{X}}_0 + \mathbf{V}\tau] \quad (3.34)$$

where

$$\bar{\mathbf{X}}_0 = L^{-1} \int_0^L [\tfrac{1}{2}\widehat{\mathbf{a}}(\sigma) + \tfrac{1}{2}\widehat{\mathbf{b}}(-\sigma)]\,\mathrm{d}\sigma \quad (3.35)$$

is the spatial centre-of-mass at time $\tau = 0$. Thus the centre-of-mass of the loop follows an inertial trajectory with 3-velocity $\mathbf{V}$.

An additional result, due to Tom Kibble [Kib85], is that the root mean square velocity of a string loop in its centre-of-momentum frame, averaged over a period of oscillation, is $\frac{1}{\sqrt{2}}$. This follows from the fact that, if $\mathbf{r}(\tau, \sigma)$ is the spatial position vector of the string, then

$$\mathbf{r}_{,\tau\tau} = \mathbf{r}_{,\sigma\sigma} \qquad \text{and} \qquad \mathbf{r}_\sigma^2 = 1 - \mathbf{r}_\tau^2 \quad (3.36)$$

and so the period mean of $\mathbf{r}_\tau^2$ in a general Lorentz frame is

$$\langle \mathbf{r}_\tau^2 \rangle \equiv \frac{1}{L^2} \int_0^L \int_0^L \mathbf{r}_\tau^2 \,\mathrm{d}\sigma\,\mathrm{d}\tau = \frac{1}{L} \int_0^L \mathbf{V}\cdot\mathbf{r}_\tau \,\mathrm{d}\sigma - \frac{1}{L^2} \int_0^L \int_0^L \mathbf{r}\cdot\mathbf{r}_{,\tau\tau} \,\mathrm{d}\sigma\,\mathrm{d}\tau$$

$$= \mathbf{V}^2 - \frac{1}{L^2}\int_0^L\int_0^L \mathbf{r}\cdot\mathbf{r}_{,\sigma\sigma}\,\mathrm{d}\sigma\,\mathrm{d}\tau = \mathbf{V}^2 + \frac{1}{L^2}\int_0^L\int_0^L \mathbf{r}_\sigma^2\,\mathrm{d}\sigma\,\mathrm{d}\tau$$

$$= \mathbf{V}^2 + \langle 1 - \mathbf{r}_\tau^2\rangle. \tag{3.37}$$

Hence, $\langle\mathbf{r}_\tau^2\rangle = \frac{1}{2}(1+\mathbf{V}^2)$ and so, in particular, $\langle\mathbf{r}_\tau^2\rangle = \frac{1}{2}$ in the centre-of-momentum frame. Since $\langle\mathbf{r}_\tau\rangle = \mathbf{V}$, the mean square deviation of $\mathbf{r}_\tau$ from the bulk velocity $\mathbf{V}$ in a more general frame is

$$\langle(\mathbf{r}_\tau - \mathbf{V})^2\rangle = \tfrac{1}{2}(1-\mathbf{V}^2). \tag{3.38}$$

However, it should be pointed out that the integrals over σ in (3.37) are mass-weighted averages rather than ordinary spatial averages, as it is $\gamma^{1/4}\,\mathrm{d}\sigma$ rather than $\mathrm{d}\sigma$ that is the proper distance element. This is evident also from the calculation of the total 4-momentum P^μ, which indicates that the bulk velocity $\langle\mathbf{r}_\tau\rangle = \mathbf{V}$ is mass-weighted. As a result, the value of $\langle\mathbf{r}_\tau^2\rangle$ is weighted towards segments of the loop with high Lorentz factor and so, in a sense, overstates the actual mean square velocity.

Another useful conserved quantity is the total angular momentum of a string loop [Tur84]. This can be calculated in two ways. If $\mathbf{n}$ is any unit 3-vector then the rotational Killing vector about $\mathbf{n}$ is

$$k_\mu = [0, \mathbf{n}\times\mathbf{r}] \tag{3.39}$$

where $\mathbf{r}$ is again the spatial position vector. The total angular momentum of the string about $\mathbf{n}$ is, therefore,

$$J(\mathbf{n}) = \mu\int_0^L k\cdot X_\tau\,\mathrm{d}\sigma = \tfrac{1}{4}\mu\mathbf{n}\cdot\int_0^L(\mathbf{a}+\mathbf{b})\times(\mathbf{a}'+\mathbf{b}')\,\mathrm{d}\sigma. \tag{3.40}$$

Now, in terms of the periodic mode functions $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{b}}$,

$$\begin{aligned}(\mathbf{a}+\mathbf{b})\times(\mathbf{a}'+\mathbf{b}') &= \widehat{\mathbf{a}}\times\widehat{\mathbf{a}}' + \widehat{\mathbf{b}}\times\widehat{\mathbf{b}}' + \widehat{\mathbf{a}}\times\widehat{\mathbf{b}}' + \widehat{\mathbf{b}}\times\widehat{\mathbf{a}}'\\ &\quad + 2\tau\mathbf{V}\times(\widehat{\mathbf{a}}'+\widehat{\mathbf{b}}') + 2(\widehat{\mathbf{a}}+\widehat{\mathbf{b}})\times\mathbf{V}\end{aligned} \tag{3.41}$$

where, since

$$\frac{\mathrm{d}}{\mathrm{d}\sigma}[\widehat{\mathbf{b}}(\tau-\sigma)\times\widehat{\mathbf{a}}(\tau+\sigma)] = \widehat{\mathbf{b}}(\tau-\sigma)\times\widehat{\mathbf{a}}'(\tau+\sigma) + \widehat{\mathbf{a}}(\tau+\sigma)\times\widehat{\mathbf{b}}'(\tau-\sigma) \tag{3.42}$$

the third, fourth and fifth terms on the right of (3.41) all vanish on integration. Hence,

$$J(\mathbf{n}) = \tfrac{1}{4}\mu\mathbf{n}\cdot\int_0^L(\widehat{\mathbf{a}}\times\widehat{\mathbf{a}}' + \widehat{\mathbf{b}}\times\widehat{\mathbf{b}}')\,\mathrm{d}\sigma + \mathbf{n}\cdot(\mathbf{M}\times\mathbf{V}) \tag{3.43}$$

where

$$\mathbf{M} = \tfrac{1}{2}\mu\int_0^L(\widehat{\mathbf{a}}+\widehat{\mathbf{b}})\,\mathrm{d}\sigma \equiv \tfrac{1}{2}\mu L\bar{\mathbf{X}}_0 \tag{3.44}$$

is the mass moment of the loop. The angular momentum vector measured by an observer at rest with respect to the coordinate system is, therefore,

$$\mathbf{J} = \tfrac{1}{4}\mu \int_0^L (\widehat{\mathbf{a}} \times \widehat{\mathbf{a}}' + \widehat{\mathbf{b}} \times \widehat{\mathbf{b}}')\, \mathrm{d}\sigma + \mathbf{M} \times \mathbf{V}. \tag{3.45}$$

Note, in particular, that the left- and right-moving modes contribute separately to $\mathbf{J}$.

In the rest frame of the loop, $\widehat{\mathbf{a}}'$ and $\widehat{\mathbf{b}}'$ are both unit vectors, and so $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{b}}$ trace out closed curves each of length L in $\mathbb{R}^3$. Suppose that $\widehat{\mathbf{a}}$ traces a planar curve which does not intersect itself. Then

$$\mathbf{A}_a = \tfrac{1}{2} \int_0^L \widehat{\mathbf{a}} \times \widehat{\mathbf{a}}'\, \mathrm{d}\sigma \tag{3.46}$$

is a vector normal to the plane with magnitude equal to the area enclosed by the curve. In the more general case where the curve does intersect itself, $\mathbf{A}_a$ can be found by breaking the curve up into non-intersecting closed segments and adding the corresponding vector areas. Furthermore, if the curve is not planar, it is always possible to project $\widehat{\mathbf{a}}$ onto a plane normal to the direction of $\mathbf{A}_a$ and so generate a closed planar curve with length strictly less than L; the magnitude of $\mathbf{A}_a$ is then the area enclosed by this curve.

Since the area enclosed by a curve of fixed length is maximal when the curve is a circle, $|\mathbf{A}_a|$ is bounded above by $L^2/(4\pi)$. A similar argument applies to the contribution of the $\widehat{\mathbf{b}}$ mode to the angular momentum vector $\mathbf{J}$, and so

$$|\mathbf{J}| \leq \frac{1}{4\pi} \mu L^2 \tag{3.47}$$

for any loop configuration [SQSP90]. Equality holds in (3.47) if and only if the mode functions $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{b}}$ trace out co-planar circles with the same orientation. This corresponds to a trajectory in the form of a rotating doubled straight line whose ends are moving at the speed of light (see section 4.2.2).

An alternative way to characterize the angular momentum of a loop is to first construct the angular momentum currents

$$M_A^{\mu\nu} = \mu(X^\mu X^\nu{}_{,A} - X^\nu X^\mu{}_{,A}). \tag{3.48}$$

By virtue of the flat-space equation of motion $\gamma^{AB} X^\mu{}_{,AB} = 0$, these satisfy the conservation equation

$$\gamma^{AB} M_A^{\mu\nu}{}_{,B} = \mu(X^\mu \gamma^{AB} X^\nu{}_{,AB} - X^\nu \gamma^{AB} X^\mu{}_{,AB}) = 0. \tag{3.49}$$

Hence, the conserved total angular momentum tensor of the loop is

$$M^{\mu\nu} = \int_0^L M_\tau^{\mu\nu}\, \mathrm{d}\sigma = \begin{pmatrix} 0 & -M_x & -M_y & -M_y \\ M_x & 0 & J_z & -J_y \\ M_y & -J_z & 0 & J_x \\ M_z & J_y & -J_x & 0 \end{pmatrix} \tag{3.50}$$

and so can be used to reconstruct the components of both **M** and **J** in any Lorentz frame.

A further macroscopic property of interest is the invariant length of a closed loop. The parametric period L in an aligned standard gauge typically has different values in different reference frames. The minimum possible value of the parametric period occurs in the rest frame of the loop, and is

$$L_{\min} = (1 - \mathbf{V}^2)^{1/2} L \tag{3.51}$$

where L and $\mathbf{V}$ are the parametric period and bulk velocity of the loop in an arbitrary frame. Since the quantity $\mu L_{\min}$ is the magnitude of the 4-momentum P^{μ} (equation (3.32)), it is natural to interpret $L_{\min}$ as, in some sense, the rest length of the string. It was first identified by Tanmay Vachaspati and Alexander Vilenkin [VV85] in 1985, and called by them the *invariant length* of the loop.

As previously mentioned (section 2.5), the invariant length $\ell_{\mathrm{I}}(\mathfrak{p})$ of a string loop in an arbitrary background spacetime is defined to be twice the square root of the area of the subset $D(\mathfrak{p})$ whose interior is causally disconnected from the point $\mathfrak{p}$. This definition is motivated by the fact that $\ell_{\mathrm{I}}(\mathfrak{p})$ is a natural world-sheet scalar and reduces to the invariant length $L_{\min}$ at all points on the world sheet when the background is Minkowski. To see this, it is most convenient to calculate ℓ_{I} in the light-cone coordinates $\sigma_{\pm} = \tau \pm \sigma$ corresponding to an arbitrary aligned standard gauge (τ, σ). Then

$$\begin{aligned} \ell_{\mathrm{I}}(\sigma_+, \sigma_-) &= 2\left(\int_{\sigma_- - L}^{\sigma_-} \int_{\sigma_+}^{\sigma_+ + L} X_+ \cdot X_- \, \mathrm{d}\sigma_+ \, \mathrm{d}\sigma_- \right)^{1/2} \\ &= \left[\int_{\sigma_- - L}^{\sigma_-} \int_{\sigma_+}^{\sigma_+ + L} (1 - \mathbf{a}' \cdot \mathbf{b}') \, \mathrm{d}\sigma_+ \, d\sigma_- \right]^{1/2} \end{aligned} \tag{3.52}$$

where, in terms of the periodic mode functions $\widehat{\mathbf{a}}(\sigma_+)$ and $\widehat{\mathbf{b}}(\sigma_-)$,

$$1 - \mathbf{a}' \cdot \mathbf{b}' = 1 - \widehat{\mathbf{a}}' \cdot \widehat{\mathbf{b}}' - (\widehat{\mathbf{a}}' + \widehat{\mathbf{b}}') \cdot \mathbf{V} - \mathbf{V}^2. \tag{3.53}$$

Hence,

$$\ell_{\mathrm{I}}(\sigma_+, \sigma_-) = \left[\int_{\sigma_- - L}^{\sigma_-} \int_{\sigma_+}^{\sigma_+ + L} (1 - \mathbf{V}^2) \, \mathrm{d}\sigma_+ \, \mathrm{d}\sigma_- \right]^{1/2} = (1 - \mathbf{V}^2)^{1/2} L \tag{3.54}$$

as claimed.

As well as comprising a strict lower bound for the parametric period L, the invariant length $L_{\min}$ has the useful property that it forms an upper bound for the proper length of the string loop as measured in any Lorentz frame. The length ℓ of the loop on a cross-sectional surface of constant τ in the aligned standard gauge is

$$\ell(\tau) = \int_0^L (-X_\sigma^2)^{1/2} \, \mathrm{d}\sigma \le \left[L \int_0^L (-X_\sigma^2) \, \mathrm{d}\sigma \right]^{1/2} \tag{3.55}$$

where the second term on the right follows from Schwarz's inequality. Now,

$$-X_\sigma^2 = \tfrac{1}{4}(\widehat{\mathbf{a}}'^2 + \widehat{\mathbf{b}}'^2 - 2\widehat{\mathbf{a}}' \cdot \widehat{\mathbf{b}}') \leq \tfrac{1}{4}(\widehat{\mathbf{a}}'^2 + \widehat{\mathbf{b}}'^2 + 2|\widehat{\mathbf{a}}'||\widehat{\mathbf{b}}'|) \tag{3.56}$$

and so, again by virtue of Schwarz's inequality,

$$\int_0^L (-X_\sigma^2)\,\mathrm{d}\sigma \leq \tfrac{1}{4}\int_0^L \widehat{\mathbf{a}}'^2\,\mathrm{d}\sigma + \tfrac{1}{4}\int_0^L \widehat{\mathbf{b}}'^2\,\mathrm{d}\sigma + \tfrac{1}{2}\left(\int_0^L \widehat{\mathbf{a}}'^2\,\mathrm{d}\sigma\right)^{1/2}\left(\int_0^L \widehat{\mathbf{b}}'^2\,\mathrm{d}\sigma\right)^{1/2}. \tag{3.57}$$

Finally, in view of the gauge conditions (3.12) on the periodic mode functions $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{b}}$,

$$\widehat{\mathbf{a}}'^2 = 1 - 2\widehat{\mathbf{a}}' \cdot \mathbf{V} - \mathbf{V}^2 \qquad \text{and} \qquad \widehat{\mathbf{b}}'^2 = 1 - 2\widehat{\mathbf{b}}' \cdot \mathbf{V} - \mathbf{V}^2 \tag{3.58}$$

and, therefore,

$$\ell(\tau) \leq (1 - \mathbf{V}^2)^{1/2} L = L_{\min}. \tag{3.59}$$

Note that this upper bound is achievable whenever $\widehat{\mathbf{a}}'$ and $\widehat{\mathbf{b}}'$ are anti-parallel and X_σ^2 is constant at all points on the surface of constant τ. This will occur if and only if the periodic mode functions have the form

$$\widehat{\mathbf{a}}(\tau + \sigma) = (1 - \mathbf{V}^2)^{1/2}\,\widehat{\mathbf{u}}(\sigma + \tau - \tau_0) \tag{3.60}$$

and

$$\widehat{\mathbf{b}}(\tau - \sigma) = (1 - \mathbf{V}^2)^{1/2}\,\widehat{\mathbf{u}}(\sigma + \tau_0 - \tau) \tag{3.61}$$

where τ_0 is a constant, and $\widehat{\mathbf{u}}$ is a periodic function with period L and satisfies the constraints

$$\widehat{\mathbf{u}}'^2 = 1 \qquad \text{and} \qquad \widehat{\mathbf{u}}' \cdot \mathbf{V} = 0. \tag{3.62}$$

In particular, the measured length will be equal to the invariant length $L_{\min}$ in the rest frame of the string loop whenever the loop is momentarily stationary.

3.4 Initial-value formulation for a string loop

A considerable amount of effort has been invested by a number of authors in generating simple yet interesting solutions which satisfy the periodicity condition (3.8), largely for the purpose of estimating the flux of gravitational radiation from realistic string loops (see section 6.5). A common method is to decompose the trajectory into Fourier components and then impose the constraints $\mathbf{a}'^2 = \mathbf{b}'^2 = 1$, a procedure which can be simplified considerably by using a spinorial representation (see section 3.5). However, a method which has more direct physical appeal is to reconstruct a trajectory from its initial data on a surface of constant Minkowski time t.

Suppose then that on an initial surface $t = 0$ a string loop has position vector $\mathbf{r}_0(\theta)$ and velocity $\mathbf{v}_0(\theta)$, where θ is a parametric variable with range 0 to θ^*, and $\mathbf{r}_0$ and $\mathbf{v}_0$ are periodic functions of θ so that

$$\mathbf{r}_0(0) = \mathbf{r}_0(\theta^*) \qquad \text{and} \qquad \mathbf{v}_0(0) = \mathbf{v}_0(\theta^*). \tag{3.63}$$

If the trajectory of the loop is described using the aligned standard gauge, θ must be a function of the spacelike gauge coordinate σ on the initial surface, and so at $\tau = 0$ the world-sheet tangent vectors have the form

$$X^\mu_\tau = [1, \mathbf{v}_0] \tag{3.64}$$

and

$$X^\mu_\sigma = \left[0, \mathbf{r}'_0 \frac{\mathrm{d}\theta}{\mathrm{d}\sigma}\right]. \tag{3.65}$$

The gauge constraint $X_\tau \cdot X_\sigma = 0$ then requires that $\mathbf{v}_0 \cdot \mathbf{r}'_0 = 0$, so the initial velocity of the loop must be chosen to be everywhere orthogonal to $\mathbf{r}'_0$. This restriction is simply a reflection of the fact that a string is locally invariant with respect to boosts parallel to the world sheet, and so any longitudinal component in the string's velocity is undetectable. In view of the second gauge constraint

$$0 = X^2_\tau + X^2_\sigma = 1 - \mathbf{r}'^2_0 \left(\frac{\mathrm{d}\theta}{\mathrm{d}\sigma}\right)^2 - \mathbf{v}^2_0 \tag{3.66}$$

the parameter θ and the gauge coordinate σ are related by the equation

$$\frac{\mathrm{d}\theta}{\mathrm{d}\sigma} = |\mathbf{r}'_0|^{-1}\sqrt{1 - \mathbf{v}^2_0}. \tag{3.67}$$

Since according to equation (3.6), the tangent vectors on the initial surface $\tau = 0$ are just

$$X^\mu_\tau(0, \sigma) = [1, \tfrac{1}{2}\mathbf{a}'(\sigma) + \tfrac{1}{2}\mathbf{b}'(-\sigma)] \tag{3.68}$$

and

$$X^\mu_\sigma(0, \sigma) = [0, \tfrac{1}{2}\mathbf{a}'(\sigma) - \tfrac{1}{2}\mathbf{b}'(-\sigma)] \tag{3.69}$$

it follows immediately that, if $\theta(0) = 0$,

$$\mathbf{a}(\tau + \sigma) = \mathbf{r}_0(\theta(\sigma + \tau)) + \int_0^{\theta(\sigma+\tau)} \mathbf{v}_0 \frac{\mathrm{d}\sigma}{\mathrm{d}\theta}\, \mathrm{d}\theta \tag{3.70}$$

and

$$\mathbf{b}(\tau - \sigma) = \mathbf{r}_0(\theta(\sigma - \tau)) - \int_0^{\theta(\sigma-\tau)} \mathbf{v}_0 \frac{\mathrm{d}\sigma}{\mathrm{d}\theta}\, \mathrm{d}\theta. \tag{3.71}$$

In particular, if the loop is initially motionless then the full trajectory is

$$X^\mu(\tau, \sigma) = [\tau, \tfrac{1}{2}\mathbf{r}_0(\theta(\sigma + \tau)) + \tfrac{1}{2}\mathbf{r}_0(\theta(\sigma - \tau))] \tag{3.72}$$

where the function $\theta(\sigma)$ is determined implicitly by the equation

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\theta} = |\mathbf{r}_0'|. \tag{3.73}$$

In the more general case of non-static initial data, the parametric period L of the loop is given by

$$L = \int_0^{\theta^*} (1 - \mathbf{v}_0^2)^{-1/2} |\mathbf{r}_0'| \, \mathrm{d}\theta \tag{3.74}$$

while the bulk velocity of the loop is

$$\mathbf{V} = L^{-1} \int_0^{\theta^*} (1 - \mathbf{v}_0^2)^{-1/2} |\mathbf{r}_0'| \mathbf{v}_0 \, \mathrm{d}\theta. \tag{3.75}$$

Examples of solutions generated using this method will be given in section 4.6.

3.5 Periodic solutions in the spinor representation

Even though the general solution (3.6) to the string equation of motion in a Minkowski background is extremely simple, the problem of constructing periodic solutions which satisfy the gauge condition $\mathbf{a}'^2 = \mathbf{b}'^2 = 1$ is not entirely trivial. As mentioned earlier, one possible method is to begin with periodic initial data, although the integrals (3.70) and (3.71) which then result are usually very complicated. An alternative is to expand the mode functions $\mathbf{a}'$ and $\mathbf{b}'$ as a finite series of harmonics in $\sigma_+ = \tau + \sigma$ and $\sigma_- = \tau - \sigma$, respectively.

For example, if $\mathbf{a}'$ contains only zero- and first-order harmonics then it has the general form

$$\mathbf{a}' = \mathbf{V} + \mathbf{C}\cos(2\pi\sigma_+/L) + \mathbf{S}\sin(2\pi\sigma_+/L) \tag{3.76}$$

where $\mathbf{C}$ and $\mathbf{S}$ are constant vectors. The gauge condition $\mathbf{a}'^2 = 1$ then implies that

$$\mathbf{C}\cdot\mathbf{V} = \mathbf{S}\cdot\mathbf{V} = \mathbf{C}\cdot\mathbf{S} = 0 \qquad \mathbf{C}^2 = \mathbf{S}^2 \qquad \text{and} \qquad \mathbf{V}^2 + \mathbf{C}^2 = 1 \tag{3.77}$$

so that the general solution for $\mathbf{a}'$ reads:

$$\mathbf{a}' = (\mathbf{n}_1 \times \mathbf{n}_2)\cos\phi + \mathbf{n}_1 \sin\phi\cos(2\pi\sigma_+/L) + \mathbf{n}_2 \sin\phi\sin(2\pi\sigma_+/L) \tag{3.78}$$

where $\mathbf{n}_1$ and $\mathbf{n}_2$ are any two orthogonal unit vectors and ϕ is an arbitrary constant. A similar first-order solution exists for $\mathbf{b}'$.

Finding periodic solutions in this way rapidly becomes unworkable as the number of harmonics increases. If $\mathbf{a}'$ contains all harmonics up to order N, so that

$$\mathbf{a}' = \mathbf{V} + \sum_{n=1}^{N} \mathbf{C}_n \cos(2\pi n\sigma_+/L) + \sum_{n=1}^{N} \mathbf{S}_n \sin(2\pi n\sigma_+/L) \tag{3.79}$$

21.04.2020

Cartesian coordinates. the best we can do is to introduce two arbitrary of coordinate lines and label them by $u^\alpha = (u^1, u^2)$. these curvilinear (or Gaussian) coordinates do not always a direct geometric interpretation.

obs. Use este livro.

A2) Books_70_2020/
Gravitation and gauge symmetries/
String Gravity and Physics at the Planck Energy....pdf

Pág. 11(26)

Lectures on string theory in curved spacetimes

A3) Books_70_2020/
Gravitation and gauge symmetries/
the Mathematical theory of Cosmic ... pdf

Pág. 59(73) - Chapter 3

String dynamics in flat space

the Minkowski spacetime is the only background metric of cosmological interest in which the string equation of motion (2.14) can be integrated exactly.

Inductance, Inductors, and Energy

An inductor is a circuit element that produces a voltage proportional to the time rate of change of the current passing through it,

$$V = -L\,dI/dt\,,$$

where L is the inductance, measured in *henrys*, 1 H = 1 V s/A.

The inductance of a solenoid with n turns per unit length, volume $\mathcal{V}$, and core permeability μ is approximately

$$L_{\text{solenoid}} \cong n^2 \mathcal{V} \mu\,.$$

The energy stored in an inductor is $E_{\text{EM}} = LI^2/2$. This is a special case of the general expression for energy stored in magnetic fields,

$$E_{\text{EM}} = \frac{1}{2\mu} \int dV\, |\boldsymbol{B}|^2\,.$$

second circuit, proportional to the rate of change of the current in the first,

$$\mathcal{E}_2 = -M_{21} \frac{dI_1}{dt}. \tag{3.72}$$

Similarly, a time-dependent current in circuit two leads to an EMF in circuit one: $\mathcal{E}_1 = -M_{12} dI_2/dt$. It is possible to show using conservation of energy (Problem 3.23) that this relationship is symmetric: $M_{12} = M_{21} \equiv M$, where M is known as the **mutual inductance** of circuits one and two. The sign of M depends on the relative orientation of the two circuits. In general $M^2 \le L_1 L_2$, and the ideal limit $M^2 = L_1 L_2$ is reached only when all the magnetic flux loops both circuits – as it would for two coincident ideal solenoids of the same area and length. A conceptual model of such a system is shown in Figure 3.24. The symbol for a mutual inductor is , two inductors with lines denoting an iron core that channels the magnetic flux between them.

Mutual inductance enables the construction of one of the most common and useful electrical devices: the **transformer**. Consider the AC circuit shown in Figure 3.25. The left-hand circuit (the **primary**) is coupled by mutual inductance to the right-hand circuit (the **secondary**). In the ideal limit where $M^2 = L_P L_S$, all magnetic flux lines go through both inductors, so $\mathcal{E}_P = -N_P d\Phi/dt$, $\mathcal{E}_S = -N_S d\Phi/dt$, and,

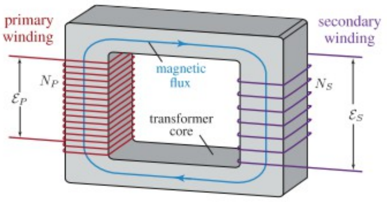

Figure 3.24 A simple transformer. Two coils are wound around a core with high magnetic permeability (typically iron), which acts as a *magnetic circuit*, channelling the magnetic flux through both coils. In the ideal limit where all magnetic flux goes through both coils, the ratio of secondary to primary voltages $\mathcal{E}_S/\mathcal{E}_P$ is given by the ratio of the number of turns, in this case roughly 1:2.

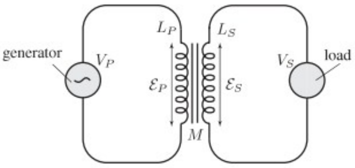

Figure 3.25 Two circuits, the *primary* (left) and the *secondary* (right), are coupled by mutual inductance. The resulting device, a *transformer*, transforms the AC voltage on the left V_P to $\sqrt{L_S/L_P}\,V_P$ on the right.

$$\frac{\mathcal{E}_S}{\mathcal{E}_P} = \frac{V_S}{V_P} = \sqrt{\frac{L_S}{L_P}} = \frac{N_S}{N_P} \tag{3.73}$$

(see Problem 3.25). Power is transmitted from the generator on the left to the load on the right.

Transformers play an essential role in the manipulation of AC power in electric power distribution networks. In the absence of imperfections such as resistance in the coils and incomplete containment of the magnetic flux, the efficiency of a transformer would be 100%. Those used in electric power distribution reach 98% or higher. We describe the role of transformers in electrical grids further in §38.

3.6 Maxwell's Equations

Up to this point, we have introduced four equations that describe the way that electric and magnetic fields interact with matter and with one another. Gauss's law (3.12) and Ampere's law (3.49) prescribe the way that electric and magnetic fields are produced by charges and currents. Two other equations – eq. (3.47), associated with the absence of magnetic charge, and Faraday's law (3.66) – constrain $\boldsymbol{E}$ and $\boldsymbol{B}$. Although we stated these laws under special conditions – for example, Gauss's law was derived only for static charges – all but one of these equations continue to hold under all known conditions.

The one equation that needs further modification is Ampere's law, which we have described only in the conditions of steady currents and constant charge distribution. As discussed earlier, Ampere's law as stated in eq. (3.49) is not consistent with charge conservation. To see this, take the divergence of eq. (3.49). Since $\nabla \cdot (\nabla \times \boldsymbol{B}) = 0$ (B.21), eq. (3.49) implies $\nabla \cdot \boldsymbol{j} = 0$, which only holds if $\partial \rho / \partial t = 0$ (see eq. (3.33)). When current and charge densities change with time, this is not true. Around 1860, the English physicist James Clerk Maxwell recognized this problem, and solved it by subtracting a term proportional to $\partial \boldsymbol{E} / \partial t$ from the left-hand side of Ampere's law. Maxwell's alteration of Ampere's law reads,

$$\nabla \times \boldsymbol{B} - \mu_0 \epsilon_0 \frac{\partial \boldsymbol{E}}{\partial t} = \mu_0 \boldsymbol{j}. \tag{3.74}$$

The demonstration that this equation is consistent with current conservation is left to Problem 3.26. Note that this modification of Ampere's law, which states that a time-varying electric field gives rise to a magnetic field, is closely parallel to Faraday's law of induction stating that a time-varying magnetic field gives rise to an electric field. Indeed, in the absence of electric charges and currents, the laws of electromagnetism are invariant under a symmetry exchanging $\boldsymbol{B}$ and $\sqrt{\mu_0 \epsilon_0}\boldsymbol{E}$.

Maxwell's modification of Ampere's law completed the unification of electricity and magnetism that began with Faraday's discovery of induction. Together with the Lorentz force law (3.45), which describes the way that electric and magnetic fields act on charges and currents, **Maxwell's equations** give a complete description of electromagnetic phenomena. These equations paved the way for Einstein's special theory of relativity, which further illuminated the unified structure of electromagnetism. Maxwell's achievement initiated great advances in understanding and application of electromagnetic theory, including the realization that light is an electromagnetic wave and that other forms of electromagnetic radiation could be produced from oscillating charges and currents.

Maxwell's Equations

Gauss's law	$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0}$
No magnetic charge	$\nabla \cdot \boldsymbol{B} = 0$
Faraday's law	$\nabla \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = 0$
Ampere–Maxwell law	$\nabla \times \boldsymbol{B} - \mu_0 \epsilon_0 \frac{\partial \boldsymbol{E}}{\partial t} = \mu_0 \boldsymbol{j}$

Discussion/Investigation Questions

3.1 Think of some devices that generally are not run on electrical power. Discuss the reasons why other power sources are favored for these devices.

3.2 Can you identify the largest component of your personal electric power consumption?

3.3 Most countries provide AC electricity at $V_{\text{RMS}} \approx 220$–240 V to residential customers. In the US and parts of South America $V_{\text{RMS}} \approx 110$–120 V is provided. How do resistive losses in household wires compare between the two? Given the result, why do you think even higher voltages are not used?

3.4 Discuss the statement that Lenz's law follows from conservation of energy. Consider, for example, the scenario of Figure 3.22(b) if the sign of eq. (3.64) were reversed.

3.5 Currents produce magnetic fields, and magnetic fields exert forces on currents. When a current flows through a wire, does the magnetic field it produces act to make the wire expand or contract? What about the force on the wires in a solenoid? Does it act to make the solenoid expand or contract? These are significant effects for wires carrying very large currents.

Problems

3.1 The electric field outside a charged conducting sphere is the same as if the charge were centered at its origin. Use this fact to calculate the capacitance of a sphere of radius R, taking the second conductor to be located at infinity. What is the most charge you can store on an otherwise isolated spherical conductor of radius R without a breakdown of the surrounding air as discussed in Example 3.2? What is the maximum energy you can store on a conducting sphere of radius 1 mm?

3.2 [**T**] Prove Gauss's law from Coulomb's law for static charge distributions by showing that the electric field of a single charge satisfies the integral form of Gauss's law and then invoking linearity.

3.3 How much energy can you store on a parallel plate capacitor with $d = 1\,\mu\text{m}$, $A = 10\,\text{cm}^2$, and $\epsilon = 100\epsilon_0$, assuming that the breakdown field of the dielectric is the same as for air?

3.4 [**T**] Starting from Gauss's law and ignoring edge effects (i.e. assume that the plates are very large and the electric field is uniform and perpendicular to the plates), derive the formula for the capacitance of a parallel plate capacitor, $C = \epsilon_0 A/d$. Refer to Figure 3.8. You can assume that the electric field vanishes outside the plates defining the capacitor. Show that the energy stored in the capacitor, $\frac{1}{2}CV^2$, can be written as the integral of $\epsilon_0|\boldsymbol{E}|^2/2$ over the region within the capacitor (as asserted in eq. (3.20)).

3.5 Suppose that a capacitor with capacitance C is charged to some voltage V and then allowed to discharge through a resistance R. Write an equation governing the rate at which energy in the capacitor decreases with time due to resistive heating. Show that the solution of this equation is $E(t) = E(0)e^{-2t/RC}$. You can ignore the internal resistance of the capacitor. Show that the heat produced in the resistor equals the energy originally stored in the capacitor.

3.6 The dielectrics in capacitors allow some *leakage current* to pass from one plate to the other. The leakage can be parameterized in terms of a *leakage resistance* R_L. This limits the amount of time a capacitor can be used to store electromagnetic energy. The circuit diagram describing an isolated capacitor, slowly leaking charge, is therefore similar to the one analyzed in Problem 3.5. The Maxwell BCAP0310 ultracapacitor (see §37.4.2) is listed as having a capacitance of 310 F with a voltage up to 2.85 V, and a maximum leakage current of 0.45 mA when fully charged. Take this to be the current at $t = 0$. What is the leakage resistance of the BCAP0310? Estimate the time scale over which the charge on the capacitor falls to $1/e$ of its initial value.

3.7 A cloud-to-ground lightning bolt can be modeled as a parallel plate capacitor discharge, with Earth's surface and the bottom of the cloud forming the two plates (see Example 3.2). A particular bolt of lightning passes to the ground from a cloud bottom at a height of 300 m. The bolt transfers a total charge of 5 C and a total energy of 500 MJ to the ground, with an average current of 50 kA. How long did the lightning bolt last? What was the electric field strength in the *cloud–earth capacitor* just before it discharged? How does this electric field compare with the breakdown field of air (3 MV/m)? (It is now known that various effects cause lightning to begin at fields that are considerably smaller than the breakdown field.)

3.8 [**T**] Consider an electric dipole composed of two charges $\pm Q$ at positions $\pm\boldsymbol{\xi}/2$. Write the exact electric field from the two charges and show that the leading term in an expansion in $1/r$ matches the $\boldsymbol{E}$-field quoted in Box 3.2. [Hint: consider the binomial expansion, eq. (B.66).]

3.9 If each of the batteries used in the flashlight in Example 3.3 has an internal resistance of 0.5 Ω (in series with the circuit), what fraction of power is lost to Joule heating within the batteries?

3.10 [**T**] Consider two resistors placed in **series**, one after the other, in an electric circuit connected to a battery with voltage V. Show that the effective resistance of the pair is $R_1 + R_2$ by using the fact that the current through both resistors is the same while voltages add. Now connect the resistors in **parallel**, so that the voltage across both resistors is V. Compute the total current and show that the effective resistance satisfies $1/R = 1/R_1 + 1/R_2$.

3.11 An appliance that uses 1000 W of power is connected by 12 gauge (diameter 2.053 mm) copper wire to a 120 V (RMS) AC outlet. Estimate the power lost per meter, dP_{lost}/dL, (in W/m) as resistive heating in the wire. (Remember that the wire to the appliance has two separate current-carrying wires in a single sheath.)

3.12 Electrical power is often used to boil water for cooking. Here are the results of an experiment: a liter of water initially at 30 °C was boiled on an electric stove top burner. The burner is rated at 6.67 A (maximum instantaneous current) and 240 V_{RMS}. It took 7 minutes 40 seconds to reach the boiling point. The experiment was repeated using an "electric kettle." The kettle is rated at 15 A (maximum instantaneous current again) and uses ordinary (US) line voltage of 120 V_{RMS}. This time it took 4 minutes and 40 seconds to reach boiling. What are the power outputs and resistances of the burner and the kettle? Compare the efficiency for boiling water of the stove top burner and the kettle. To what do you attribute the difference?

3.13 [**T**] Use the magnetic force law (3.42) and the definition of work $W = -\int_a^b d\boldsymbol{x}\cdot\boldsymbol{F}$ to show that magnetic forces do no work. [Hint: Consider the vector identity (B.6).]

3.14 [**T**] Show that a charged particle moving in the xy-plane in the presence of a magnetic field $\boldsymbol{B} = B\hat{\boldsymbol{z}}$ will move in a circle. Compute the radius of the circle and frequency of rotation in terms of the speed v of the particle. Show that adding a constant velocity $u\hat{\boldsymbol{z}}$ in the z-direction still gives a solution with no further acceleration.

3.15 [**T**] Review how magnetic fields are calculated from Ampere's law by computing (a) the magnetic field due to a straight wire and (b) the magnetic field in the interior of a very long solenoid. The contours $\mathcal{C}_1$, $\mathcal{C}_2$ shown in red in Figure 3.18 will help. You may use the fact that the magnetic field vanishes just outside the outer boundary of the solenoid in (b).

3.16 [**T**] Show that the force per unit area on the windings of an air-core solenoid from the magnetic field of the solenoid itself is of order $F/A \sim \boldsymbol{B}^2/\mu_0$. Check that the dimensions of this expression are correct and estimate F/A in pascals if $|\boldsymbol{B}| = 1\,\text{T}$.

3.17 [**T**] Derive eq. (3.52) from eq. (3.43) and eq. (3.50). Make sure you get the both the direction and magnitude.

3.18 An electric motor operates at 1000 rpm with an average torque of $\tau = 0.1\,\text{Nm}$. What is its power output? If it is running on 1.2 A of current, estimate the back-EMF from the rotor.

3.19 Consider the motor described in Box 3.4. If the resistance in the wire wrapping the rotor is $1\,\Omega$, compute the energy lost under the conditions described. What fraction of energy is lost to Joule heating in this situation? If the current is doubled but the rotation rate is kept fixed, how do the output power, Joule heating losses, and fraction of lost energy change?

3.20 [**T**] In §3.4.2 we derived the EMF on a wire loop rotating in a magnetic field using the Lorentz force law to compute the forces on the mobile charges. Although Faraday's law of induction (3.64) does not apply in a rotating reference frame, show that the same relation (3.62) follows from Faraday's law.

3.21 [**T**] Explain why the integral $\int_S d\boldsymbol{S} \cdot \boldsymbol{B}$ that appears in eq. (3.65) is independent of the choice of surface S. [Hint: make use of eq. (3.46).]

3.22 [**T**] Consider a long, hollow solenoid of volume $\mathcal{V}$. Show that, ignoring end effects, its inductance is $L = n^2\mathcal{V}\mu_0$, and that the magnetic energy it stores $E_{\text{EM}} = LI^2/2$ can be written in the form of eq. (3.71).

3.23 [**T**] Prove that the mutual inductance is a symmetric relation, $M_{12} = M_{21}$, by computing the energy stored when current I_1 is established in loop 1 and then current I_2 is established in loop 2. Then set up the same currents in the opposite order.

3.24 Design a transmission system to carry power from wind farms in North Dakota to the state of Illinois (about 1200 km). The system should handle Illinois's summertime electricity demand of 42 GW. Land for transmission towers is at a premium, so the system uses very high voltage ($V_{\text{RMS}} = 765\,\text{kV}$). Assume that the electricity is transmitted as ordinary alternating current, although in practice *three-phase power* (see §38.3.1) would be used. Assume that each tower can carry 36 aluminum cables (nine lines, each consisting of four conductors separated by non-conducting spacers). The conductors are $750\,\text{mm}^2$ in cross section. How many separate strings of transmission towers are needed if the transmission losses are to be kept below 5%? Assume a purely resistive load.

3.25 [**T**] Consider the transformer in Figure 3.25. Suppose the load is a resistor R and that the transformer is ideal, with $M^2 = L_S L_P$ and all magnetic flux lines passing through both inductors. Show that the voltage drop across the resistor is $V_S = \sqrt{L_S/L_P}\,V_P = N_S V_P/N_P$ and that the time-averaged power consumed in the resistor is $P = (L_S/L_P)(V_P^2/2R) = N_S^2 V_P^2/2N_P^2 R$.

3.26 [**T**] Take the divergence of both sides of eq. (3.74), use Coulomb's law on the left and current conservation on right to show that the equation is consistent.

CHAPTER

14

The Forces of Nature

Elementary particles and the forces that act between them are the fundamental ingredients in our most complete theoretical description of microscopic physics. Forces transform energy from one form to another when they act on matter. The study of energy thus inevitably leads to the question: What are the fundamental forces of nature, and what are the elementary particles on which these forces act? Gravity and electromagnetism are the most familiar two of the four forces that store and transform energy in our world. The other two, manifested by the *strong interactions* that hold nuclei together and the more subtle *weak interactions*, are usually the subject of more advanced physics courses.

All four forces contribute to the production and concentration of the forms of energy that we use on Earth. Gravity pulls together hydrogen atoms deep in the Sun. Strong nuclear interactions between these hydrogen atoms, facilitated by crucial weak interactions, result in the release of energy that is radiated from the Sun as electromagnetic waves, some of which hit Earth. Nuclear fission power plants rely on the interplay of strong nuclear forces, electromagnetism, and the weak interactions. Nuclear fuel is found in regions of high concentration in Earth's crust that result from a combination of gravitational pressure and chemical electromagnetic interactions. Geothermal energy, as well, arises from the combined action of the four forces. Indeed, with the exception of tidal energy, which is essentially the result of gravity and electromagnetic forces, all energy sources used by humanity rely on a happy conjunction among all four forces of nature.

The most familiar energy processes that we encounter on a day-to-day basis can be described in terms of electromagnetic and gravitational forces and interactions. The same is true of all the energy systems treated in this book so far. Much of the energy we use, however, originates in the *strong interactions*, which are not encountered in everyday experience (unless, like Homer Simpson, you work at a nuclear reactor). The drama of the strong interactions involves some relatively familiar players such as the nuclei of atoms, but it also includes a menagerie of exotic particles such as *quarks* and *gluons* that, while they are just as real as electrons and atoms, are never seen outside of the atomic nucleus. In order to explain the origins of solar, nuclear, and geothermal energy, we must first describe some of the basic properties of the strong force. Along the way, we have the opportunity to put gravity and the electromagnetic force into a somewhat broader context.

Reader's Guide

This chapter forms an "interlude" in which we survey the fundamental particles and forces that determine the structure of matter and its interactions with energy. Because they are less well known than gravity and electromagnetism, special attention is paid to the strong force that holds the nucleus together and to the weak force responsible for many nuclear decays. After discussing the four forces, we describe the properties of all the known elementary particles. We then present a simplified version of particle physics that is adequate for most of our purposes, in which only five particles, the proton, neutron, electron, photon, and antineutrino appear. The chapter closes with an introduction to β-decay and related processes that are important for understanding nuclear energy and related radiation.

Prerequisites: §2 (Mechanics), §3 (Electromagnetism), and §7 (Quantum mechanics), although much of this chapter is accessible with little technical background.

The strong and weak forces are important ingredients in the description of nuclei and nuclear energy in §17–§19, and ionizing radiation in §20. β-decay is key to understanding many nuclear decays. Gravity figures centrally in the large-scale structure of the universe (§21).

The fourth force, which is also only relevant in the microworld, is known as the *weak interaction*. Weak

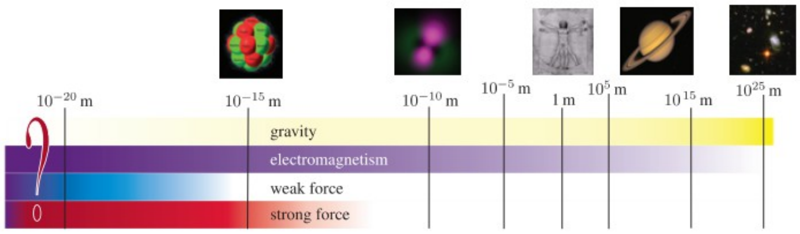

Figure 14.1 An illustration of the relative importance of the four forces over distances ranging from subnuclear (10^{-20} m) to cosmological (10^{25} m).

interactions play only a minor part in the actual generation of energy, but they act as a *facilitator* for many of the most important nuclear energy processes. Weak interactions play a role in fundamental physics a bit like fungi do in the environment. They are often overlooked; they do not contribute to building up great structures (such as nuclei and atoms); and, above all, they are the mediators of decay! Nevertheless we need to understand some of the features of the weak force because so many processes that figure significantly in energy physics would not occur without it.

Fortunately, the list of forces that affect our world stops at four: strong, electromagnetic, weak, and gravitational. There may be other forces locked away at very short distances, or that involve as-yet-undiscovered particles. Even if additional forces exist, however, we already know that they do not influence the dynamics of our familiar world in any significant way at the human scale (see Figure 14.1). We can be sure of this because we can account for essentially *all* the phenomena so far observed in nature in terms of the four known forces.[1] This may sound like an exaggeration: after all there are many common phenomena not yet fully understood by science, from the turbulent flow of liquids to the origins of life. These are generally, however, phenomena characterized by complexity, and there is no evidence that they arise from additional, as yet unknown, fundamental interactions.

The Standard Model of Particle Physics

The *Standard Model of particle physics* is a *quantum field theory* that, in concert with the classical theory of gravity, describes almost all observed phenomena to a high degree of accuracy in terms of four forces and 17 elementary particles.

If one focuses in on short distances and small systems, processes simplify, and everything that we can observe seems to be explained by a relatively simple set of fundamental equations. Our understanding is summarized in a meticulously tested and so far perfectly successful set of laws of physics, known as the **Standard Model of particle physics**. The word "*model*" connotes less certainty than "*theory*" or "*law*" – the terms typically used to designate well-established principles of physics. The Standard Model, however, which is formulated in terms of a theoretical framework known as *relativistic quantum field theory*, has described observed physical phenomena to a higher degree of precision than any other physical theory ever developed. Quantum field theory and the Standard Model are on a similar footing to Newton's and Coulomb's laws; they provide a set of simple and basic equations from which the behavior of matter can be deduced by mathematical analysis. The specific structure of the Standard Model dictates that matter is made of essentially pointlike particles with spin $1/2$ (quantum spin was discussed in §7). These particles interact with each other by emitting and absorbing photons and other massless and massive quanta that are cousins to the photon. The elementary spin-1/2 particles form families whose membership is limited by symmetry constraints and whose members have common properties. The details are beyond the scope of this book, but some aspects of the Standard Model are summarized in Tables 14.1 and 14.2 and in Figure 14.4.

Despite its successes, the Standard Model is neither a perfect nor a complete theory of nature. One glaring inadequacy is that it has about 20 parameters – numbers like the mass of the electron and the strength of the electric charge – whose origins we do not know. It also has a peculiar redundancy: each type of particle – the electron for example – comes in three versions that differ in mass, but little else. This redundancy, as well as the *raison d'etre* for the particular assortment of particles and fields that appear in nature, and other relationships in the Standard Model, are beyond our present understanding. There are some detailed aspects of the Standard Model, related to *neutrinos* and the *Higgs particle*, which are not fully settled experimentally. And, as mentioned above, it is possible that new forces or other novel phenomena are relevant at very high energies and short distance scales that have not yet been probed experimentally. Finally, and arguably at a more fundamental level, in our current understanding gravity is appended to the Standard Model only as a classical (rather than quantum) field, even though gravity also must exhibit quantum features at very short distance and length scales. Unifying gravity and quantum mechanics has challenged physicists for decades. An approach often called *string theory* provides a framework in which gravity and other forces can be described within a quantum-mechanical theory. Whether and how the Standard Model and the global structure of the universe in which we live is described within the framework of string theory is, however, still an open problem (§21).

---

[1] A few fundamental observed phenomena lie beyond the limits of our understanding. For example, *dark energy* and *dark matter*, which are discussed further in §21 (Energy in the universe), dominate the total energy budget of the universe. Important though they are at the cosmic scale, they play no role at all in the energy transformations of everyday life.

All the limitations of the Standard Model that we have just listed are irrelevant, in any case, for the purposes of this book. All known energy sources can be understood with only a subset of the structure in the Standard Model, and it is hard to imagine any way in which physics beyond the Standard Model could have practical consequences for human energy use in the forseeable future. Thus, for all intents and purposes, the fundamental physics underlying any conceivable energy system that may be relevant in this century should be described within the Standard Model.

The aim of this chapter is to take a brief tour of the Standard Model, emphasizing those bits that are most important for the physics of energy and skirting as much as possible the complexities that usually relegate the subject to graduate-level physics courses. Although the concepts discussed here are relatively remote from everyday experience, they are nevertheless relevant to practical energy systems. In fact, the basic ideas presented here are essential for understanding the origins of nuclear and solar energy and the nature of radioactivity. In §14.1 we describe the four fundamental forces in more depth. Starting at the largest distance scales, where gravity is most important, we work our way down to sub-atomic distances, encountering electromagnetic, strong, and weak interactions in succession. Next (§14.2) we meet the players in the drama of subatomic physics, the elementary particles. Before exploring the whole zoo of particles, we introduce a stripped-down version of the Standard Model that includes almost everything needed to understand energy production and flow in the observable universe at the present time. This *Standard Model "lite"* comprises the proton, neutron, electron, photon, and one potentially less familiar addition, the *(anti-)neutrino*. We briefly survey the other particles in the Standard Model and then conclude the chapter (§14.3) with a close look at β-decay, a weak interaction of particular importance in energy physics.

Non-technical introductions to particle physics can be found in many books, for example [66]. Reference [67] provides a more technical introduction, at an advanced undergraduate level.

14.1 Forces, Energies, and Distance Scales

To survey the hierarchy of strengths, distances, and energy scales that characterize the four forces, we start at astronomical distances and work our way into the interior of the atomic nucleus.

Gravitational and electromagnetic potential energies fall off like $1/r$ away from their sources. This is a special case of a more general form

$$V(r) \sim g^2 \frac{e^{-r/b}}{r}\,, \tag{14.1}$$

where b is known as the **range** of the force and g is a measure of its *strength*. Equation (14.1) defines what is known as a **Yukawa interaction**, named after the Japanese physicist Hideki Yukawa who proposed it as a model for the strong nuclear force in the 1930s. The $1/r$ potential of gravity and electromagnetism emerges when the parameter b goes to ∞. Hence, these are called *infinite range* forces. In the framework of *quantum field theory*, interactions between massive particles are the result of the exchange of particles called **force carriers** (Figure 14.2). For example, the *photon* is the force carrier for electromagnetism. In general, the exchange of a force carrier of mass m gives rise to a force of the form (14.1), where the parameter b is proportional to $1/m$. Thus, the long-range nature of the electromagnetic force arises from the fact that the photon is massless. We now go through each of the forces in turn and discuss various aspects of each.

14.1.1 Gravitational Interactions

Only gravitational forces are important at astronomical distances. Gravity alone persists at the scale of stars and galaxies not because gravity is strong – in fact gravity

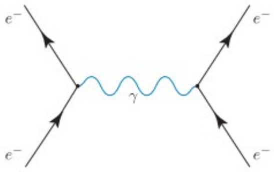

Figure 14.2 In *quantum field theory*, interactions between massive particles are mediated by the exchange of massless or massive *force carriers*. The figure depicts a *Feynman diagram* describing two electrons interacting through exchange of a *photon*, the force carrier for electromagnetic interactions.

is by far the weakest of the four forces (see Table 14.1). The reason that gravity dominates over electromagnetism on astronomical scales is that *gravity is universally attractive* and its strength is proportional to the mass (or, more generally, the total energy) of an object. In contrast, electromagnetic forces depend on the signs of the electric charges; electromagnetic attraction and repulsion between charged particles are so strong that unless we work very hard to prevent it, matter always relaxes to a state in which it is electrically neutral. So matter at macroscopic scales is essentially electromagnetically uncharged, and the force of gravity can dominate at very large scales. Thus, Newton's law of universal gravitation governs the motions of planets and stars. For very dense and very massive objects, Newton's theory of gravity must be corrected or even replaced by Einstein's General Theory of Relativity, which is described in §21.[2]

As soon as one leaves the astronomical domain, gravity subsides in importance and electromagnetic forces become equally or more important. Of course, gravity holds us here on Earth's surface – again because a very large mass is involved, namely the mass of Earth. To support the claim that gravity is unimportant on human distance scales, consider the strength of the gravitational force between, for example, two one-gram masses, separated by one meter. According to Newton, the gravitational force between them is

$$|\boldsymbol{F}_{\text{gravity}}| = G_N\,\frac{m_1 m_2}{r^2} = 6.67\times 10^{-11}\,\frac{\text{m}^3}{\text{kg}\,\text{s}^2} \times \frac{(10^{-3}\,\text{kg})^2}{1\,\text{m}^2} = 6.67\times 10^{-17}\,\text{N}\,, \quad (14.2)$$

which is far too weak to observe.

14.1.2 Electromagnetic Interactions

To gauge the strength of electromagnetic forces compared to gravity, we compute the electrostatic (Coulomb's law) force between the same two one-gram lumps of matter described above if the electric charges were not neutralized. That is, we compute the electrostatic force between the positively charged protons in the two lumps. Since about half of the mass of one gram of material comes from protons (the other half comes from neutrons; the electrons are only about 1/2000th of the mass), one gram contains about $N_A/2 \approx 3\times 10^{23}$ protons, each with charge 1.6×10^{-19} C, for a total charge of $\approx 4.8\times 10^4$ C. This generates a force of

$$|\boldsymbol{F}_{\text{EM}}| = \frac{1}{4\pi\epsilon_0}\,\frac{Q_1 Q_2}{r^2} \approx 9.0\times 10^9\,\frac{\text{Nm}^2}{\text{C}^2} \times \frac{(4.8\times 10^4\,\text{C})^2}{1\,\text{m}^2} \approx 2.1\times 10^{19}\,\text{N}. \quad (14.3)$$

Between comparable quantities of stuff, therefore, the electromagnetic force exceeds the gravitational force by $\sim$36 orders of magnitude. Of course, the repulsive Coulomb forces among the protons are so strong that there is no way we could isolate and study a gram of material with charge 4.8×10^4 C. Electromagnetic forces subside to manageable scales because – to very high accuracy – matter is neutral.

Many of the large-scale processes on Earth's surface result from the dynamic balance between electromagnetic and gravitational forces. For example, the electromagnetic repulsion between electrons in different atoms keeps matter from collapsing under huge gravitational pressures in Earth's core, and – at a more human scale – keeps you from falling through your chair. Gravitational and electromagnetic forces maintain the dynamic balance that governs the circulation of the atmosphere and oceans; the Sun's energy is absorbed and transmitted through electromagnetic forces while the overall vertical order is maintained by the pressure sustained by gravity.

There is no specific natural energy scale for the effects of *classical* gravity or electromagnetism. The larger the mass or charge, or the closer the objects, the greater the energy that can be stored in gravitational or electromagnetic potential energy. When quantum mechanics gets into the game, however, energies and charges are quantized,

---

[2] That is not to say, however, that electromagnetism can be ignored in astronomy! For example, we see distant objects only because of the light they give off. Electromagnetism is important for dynamical reasons as well. A cloud of gas cannot collapse to form a star unless it can find a way to give off some potential and kinetic energy, and radiating away energy through electromagnetic waves is the principal way that this happens.

Gravity and Electromagnetism

Gravitational and electromagnetic forces have infinite range and determine the behavior of matter over a huge range of distance scales. With minor exceptions such as natural radioactivity and geothermal energy produced by radioactive decay, natural processes on Earth's surface involve electromagnetic and gravitational interactions alone.

Maxwell's equations and quantum mechanics, along with classical gravity, are sufficient to describe basic dynamical processes from Earth-like distances down to a small fraction of the diameter of an atom.

and the quantum of energy associated with a particular force acting between elementary particles sets the fundamental scale for energies that can be absorbed, liberated, or stored by those forces acting on the quantum level. In §7 we saw that the energy of the electron in a hydrogen atom,

$$E_{\text{Bohr}} = -\frac{1}{2}\left(\frac{1}{4\pi\epsilon_0}\right)^2 \frac{me^4}{\hbar^2} = -13.6\,\text{eV}, \qquad (14.4)$$

sets a scale for the energies of atomic physics. (Remember that the minus sign in eq. (14.4) indicates that the electron in hydrogen is bound – it takes 13.6 eV of energy to remove it far away from the proton.) When a bond is formed or broken, energies of this order of magnitude are absorbed or emitted. The scale of energy available from chemical processes never strays far from the fundamental scale given in eq. (14.4), set by the mass and charge of the electron. When we manipulate materials, whether they are hydrocarbons or more exotic materials, the amount of energy we can get from them per unit mass from chemical processes is roughly the same as that available from burning coal (see Problem 14.1). To find energy sources that have much higher *energy density*, that is, more joules per kilogram, we must consider a stronger force.

14.1.3 Strong Interactions

From the human scale of meters down to the smallest distances that we can see with an optical microscopic, physics is governed by electromagnetism and gravity. Journeying inward to still shorter distances, at somewhat less than 10^{-14} m, things change rather suddenly and dramatically. The nuclei at the centers of atoms are composed of protons and neutrons (collectively referred to as **nucleons**), held together by forces known as **strong nuclear interactions**, which are a hundred times stronger than electromagnetism, and of a different nature.

The strong interactions between nucleons are complex, residual manifestations of simpler and more fundamental forces between elementary particles called **quarks**, of which the nucleons themselves are composed. The interactions among quarks, which are specified by the Standard Model, are known as **chromodynamic** forces in analogy to the *electrodynamic forces*, which they resemble in important ways.[3] Chromodynamic forces between quarks are mediated by **gluons**, massless particles that play a similar role to that of photons in electromagnetism. The forces between nucleons are to quark forces as molecular forces are to the underlying electromagnetic force between a proton and an electron. Just as the primary electric forces cancel out in neutral atoms or molecules, so the primary forces between quarks *cancel out* when the quarks group themselves together into *chromodynamically neutral* protons and neutrons. Like the residual forces between atoms and molecules (see, e.g., §9.3.1), the forces between nucleons left over after this cancellation are complicated and cannot be described accurately by a single simple mathematical formula. The name *strong interactions* was applied to these forces long before quarks and the Standard Model were discovered. Here we reserve the term "strong nuclear interactions" for these residual forces between nucleons and nuclei and distinguish them from the (in some ways even stronger, but conceptually simpler) chromodynamic forces between quarks. The strong nuclear interactions hold nuclei together and are responsible for the dynamics of fission, fusion, and radioactivity, so we explore them further here.

One way in which chromodynamics differs from electrodynamics is that the fundamental forces between quarks are so strong that quarks can never be separated from one another. Only bound states of quarks in which the primary chromodynamic forces are neutralized can exist in nature. These chromodynamically neutral **hadrons**, as they are known, include the proton and neutron and a host of other short-lived particles that can be produced in laboratory experiments. Because the chromodynamic force is so strong, it is very difficult to compute the form of the strong interaction between nucleons analytically from the theoretically simpler interactions between quarks. The best we can do at the present time is to study the nuclear force numerically or with the aid of empirical models.

---

[3] The origin of the "*chromo*" in *chromo*dynamics is explained in §14.2.3.

Chromodynamics and the Strong Nuclear Force

The *chromodynamic force* is a powerful force that binds *quarks* together into nucleons such as protons and neutrons. Residual, short-range *strong nuclear interactions* between nucleons bind nuclei together, just as residual electromagnetic interactions lead to forces between electrically neutral atoms that bind molecules together.

The strong nuclear force between nucleons depends upon their separation, their relative velocities, their relative angular momentum, and even on the orientations of their spins (protons and neutrons have spin 1/2 like the electron). This complicated force between two nucleons is often described by empirical formulae with many *fit parameters* (see, for example, [68] and references therein).

At the largest distances where it is detectable, the strong interaction potential between two nucleons at rest takes the simple form of a Yukawa interaction (14.1) with a range $b \cong 1.4 \times 10^{-15}$ m and a strength about 10 times that of electromagnetism,

$$V_{\text{strong}}(r) = -\frac{1}{4\pi\epsilon_0}\frac{g^2_{\text{strong}}}{r}e^{-r/b}\,, \tag{14.5}$$

with $g^2_{\text{strong}} \approx 10e^2$, where $e = 1.602 \times 10^{-19}$ C is the proton's electric charge.[4] The unit of distance 1 femtometer $= 10^{-15}$ m, abbreviated 1 fm, is a characteristic unit for strong interaction phenomena. It is also known as *1 Fermi* in honor of the Italian theoretical physicist Enrico Fermi. The strong interaction force between two (chromodynamically neutral) nucleons at large distances is mediated by the exchange of *pi-mesons* or **pions**, which are themselves chromodynamically neutral massive bound states of quarks and antiquarks. This is different from the long-range residual electromagnetic forces between electrically neutral atoms, which arise from exchange of multiple massless photons and have a polynomial form rather than the characteristic exponential suppression of the Yukawa potential.

Note the minus sign in eq. (14.5): the force is attractive. The form given in eq. (14.5) holds (approximately) at separations greater than about 1.5 fm. Below that, the forces between nucleons become stronger, but they do not have the simple exponential form of eq. (14.5).

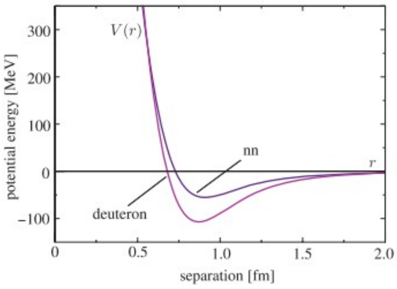

Figure 14.3 The nucleon–nucleon strong interaction potential in two configurations [68]. Roughly speaking, the curve marked *deuteron* is the potential that holds together a proton and a neutron into a bound state called a **deuteron**, while the one marked *nn* is the potential between two neutrons, which is not strong enough to bind a state. Note that while the two potential functions are quite different at ∼ 1 fm, they have the same asymptotic form (14.5) as $r \to \infty$.

Although the strength, and even the sign, of the nucleon–nucleon potential varies depending upon the nucleons' relative angular momentum, spin, and momentum, the nucleon–nucleon interaction is, on average, strong and attractive at distances greater than ∼0.7 fm. At still shorter distances, the nucleon–nucleon force turns strongly repulsive, preventing nucleons from coming too close to one another. Figure 14.3 shows the neutron–neutron potential and the proton–neutron potential; in the latter case, the configuration is the one in which the nucleons can form a bound state, the *deuteron*. Although they look quite different at 1 fm, both potentials reduce to eq. (14.5) with the same strength at large distances. At their minima near 0.8 fm, the potentials shown take values of roughly −50 MeV and −100 MeV. This means that energy in amounts measured in *millions* of electron volts per nucleon is in principle available in nuclear interactions. The depth of the potential is not, however, a good measure of the binding of nuclei. The reason for this is that the potential is very negative only in a relatively thin spherical shell. A quantum bound state in this potential has a wavefunction that is peaked at the spherical shell where $V(r)$ is negative. The second derivative of the wavefunction in the radial direction contributes substantial positive energy in the Schrödinger equation (7.7). This positive *kinetic* contribution cancels much of the 50–100 MeV of negative potential energy, reducing the net binding energy between individual nucleons to a few MeV. The two-nucleon system has a relatively small binding energy per nucleon compared to other nuclei with a few more constituent nucleons, in part since each nucleon is only bound to one other. The deuteron, composed of a proton and a neutron, is bound by only ≈ 2.2 MeV, and the two-neutron and two-proton systems are not bound at all. We investigate the systematics of nuclear binding further in §17.

---

[4] The factor $1/4\pi\epsilon_0$ is introduced so that the (appropriately scaled) strength g^2_{strong} of the strong force can be compared with electromagnetism more easily (Problem 14.3).

14.1.4 Weak Interactions

The weak interactions start to make their appearance at roughly the same distance scale as the strong interactions. The weak interactions are characterized by a Yukawa potential energy with a strength not too different from electromagnetism, but with a range that is a factor of about 500 shorter than the range of the strong interactions, so the form of the weak potential is roughly

$$V_{\text{weak}}(r) \approx \frac{g_{\text{weak}}^2}{4\pi\epsilon_0}\frac{e^{-r/b}}{r} \tag{14.6}$$

with $g_{\text{weak}} \sim e$ but $b \approx 2.5 \times 10^{-3}$ fm. Their short range makes the weak interactions far too small to be observed at nuclear distance scales, compared to both strong and electromagnetic interactions. The Yukawa form of the weak nuclear interaction arises from the fact that the fundamental particles ($W^{\pm}$ and Z *vector bosons*) mediating this interaction are massive, rather than massless like the photon. The weak interactions would be almost entirely negligible *except for the fact that they facilitate some processes that are forbidden for all the other interactions*. These processes are described in more detail in §14.3.

Are there other forces waiting to be discovered? Perhaps. Answering this question is one of the goals of modern particle physics. The *Large Hadron Collider* (LHC) facility at the European Center for Nuclear Research (CERN) in Geneva, Switzerland is looking for as yet unknown particles and forces that become important at very short distances. Whatever they may find, it will not affect the practical contemporary energy issues we are studying in this book because any new forces that may exist are locked away deep within protons and neutrons, inaccessible to our attempts so far to liberate their energies.

We have only sketched the basic properties of the forces that control physics at the subatomic scale. Some of the information about forces accumulated so far is summarized in Table 14.1 along with other facts that appear later in this chapter. Next we turn to the players in the world of subatomic physics – the elementary particles from which the world is built.

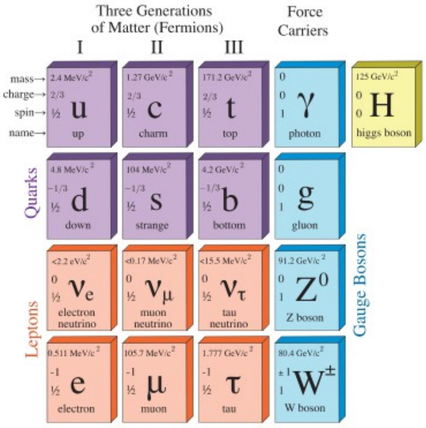

Figure 14.4 A summary of the 17 elementary particles in the Standard Model. Some of the properties of the particles are summarized in Table 14.2. The constituents of matter are shown in the first three columns, the carriers of the forces are shown in the fourth column, and the recently discovered *Higgs boson* appears in the final right-hand column. (Credit: Adapted from HolgerFiedler nach Benutzer: Murphee under CC-BY-SA 3.0 license via Wikimedia Commons)

14.2 Elementary Particles

The elementary particles in the Standard Model come in two types: the constituents of matter and the conveyers of force.[5] The archetype of the first class is the electron, and an example of the second class is the photon. It is a good idea to keep these examples in mind as the names and types of particles proliferate. For reference, Figure 14.4 gives a schematic summary of all the particles in the Standard Model, with more detail in Table 14.2.

The key to understanding the structure of Table 14.2 is the checklist of interactions in the last column. The particles that make up matter can be classified by the forces they feel. Just as electric charge indicates the extent to which a particle interacts electromagnetically, so there are "charges" for each of the other forces. *Quarks* are

---

[5] The recently discovered **Higgs particle** is the lone exception to this rule. It plays a special role in the Standard Model as the breaker of symmetries and the generator of mass. Fascinating though it is to particle physicists, we say little about the Higgs particle here since it plays no direct role in our subject.

Table 14.1 Some properties of the fundamental forces. Although energy is not explicitly listed, all the forces conserve energy. The strengths of the forces are measured relative to the chromodynamic force between quarks.

Force	Mediator	Mediator mass	Relative strength at 1 fm	Conservation laws			
				Baryon number*	Electric charge	Flavor	Parity
Gravity	Graviton	Massless	$\sim 10^{-41}$	√	√	√	√
Weak	W & Z bosons	$\sim 80\,\mathrm{GeV/c^2}$	$\sim 10^{-8}$	√	√	×	×
Electromagnetic	Photon	Massless	$\sim 10^{-2}$	√	√	√	√
Chromodynamic	Gluons	Massless	1	√	√	√	√

* Note that while all observed local processes conserve baryon number, subtle *nonperturbative* processes can violate baryon number (§21).

Table 14.2 Properties of the leptons and quarks [69]. Masses are given in units of MeV/c^2 (1 MeV/c^2= 1.783×10^{-30} kg). Note that each quark can come in three distinct *colors* corresponding to different charges under the chromodynamic force.

Name	Symbol	Mass (MeV/c^2)	Electric charge	Baryon number	Interactions			
					Strong	E&M	Weak	Gravity
Electron	e	0.510 998 92(4)	−1	0	–	√	√	√
Muon	μ	105.658 369(9)	−1	0	–	√	√	√
Tauon	τ	1776.99(29)	−1	0	–	√	√	√
Electron neutrino	ν_e	$0 < m \leq 2 \times 10^{-6}$	0	0	–	–	√	√
Mu neutrino	ν_μ	> 0	0	0	–	–	√	√
Tau neutrino	ν_τ	> 0	0	0	–	–	√	√
Up quark	u	2.3(6)	2/3	1/3	√	√	√	√
Down quark	d	4.8(4)	−1/3	1/3	√	√	√	√
Strange quark	s	95(5)	−1/3	1/3	√	√	√	√
Charm quark	c	$1.275(25) \times 10^3$	2/3	1/3	√	√	√	√
Bottom quark	b	$4.18(3) \times 10^3$	−1/3	1/3	√	√	√	√
Top quark	t	$173.2(5) \times 10^5$	2/3	1/3	√	√	√	√

"charged" with respect to chromodynamic, weak, and electromagnetic interactions. **Leptons** have no chromodynamic charge, so they do not feel the chromodynamic interactions, but all leptons interact by weak forces. Leptons are divided into **charged leptons** and **neutral leptons**, based on whether they are electromagnetically charged or neutral. All particles feel the effects of gravity.

All the particles that constitute observable matter – quarks and leptons – have spin 1/2. In §8, we discussed the division of quantum particles into *fermions* and *bosons*, the distinction being whether only one, or an arbitrary number, of particles can exist in a given quantum state. The fermionic nature of the electron is responsible for the complexity of atomic structure and the richness of chemistry. All the other leptons and all quarks are also fermions. Indeed, a fundamental theorem of relativistic quantum theory dictates that all particles with half-integral spins are fermions, while all particles with integral spins are bosons. This basic fact is discussed further in §15.5.

The particles that transmit the electromagnetic, chromodynamic, and weak forces are all spin one particles, known as **vector bosons**. The term "vector" refers to the fact that the fields associated with the gluon and the $W^\pm$ and Z are vector fields like the $\boldsymbol{E}$ and $\boldsymbol{B}$ fields associated with the photon. As mentioned earlier, the photon and gluon are massless, while the $W^\pm$ and Z bosons are massive.

The Higgs particle has spin 0, and is known as a **scalar particle** – the Higgs is the first fundamental scalar particle ever observed in nature. (The term *scalar* refers to the fact that the field of which the Higgs particle is an excitation assigns a single number – a scalar – to each point in space-time, unlike a vector field that assigns a vector to each point in space-time.) A consistent quantum theory of gravity would include a further particle associated with a quantum excitation of the gravitational field – a massless **graviton** – which would have to be a spin-2 boson.

An as yet unexplained feature of the spectrum of observed leptons and quarks is that they come in three **generations** that are almost identical except for the particle masses. The *electron* e^-, *electron neutrino* ν_e, and *up* (u) and *down* (d) quarks make up the first generation of fundamental particles. The *muon* is like a heavier version of the electron, and the *charm* (c) and *strange* (s) quarks are heavier versions of the u and d. These, together with the *muon neutrino* (ν_μ), make up the second generation. And the *top* (t) and *bottom* (b) quarks, along with the *tauon* (τ) and *τ-neutrino* (ν_τ), form a similar, but still heavier third generation. The quarks and leptons are grouped into generations in the chart of Figure 14.4.

Clearly, there is an abundance of elementary particles, many of which are difficult to produce, live only briefly before they decay, and have little if any impact on energy production or flow on Earth. Before surveying them all, we strip away the complexity and introduce the **Standard Model "lite"**, which contains the particles responsible for almost all of the phenomena we encounter in this book.

14.2.1 The Standard Model Lite: A $p\, n\, e\, \bar{\nu}_e\, \gamma$ World

For almost all practical purposes, *only five particles* play a role in the world, and one of the five usually hides in the wings, only making an occasional entrance when needed. The five key particles are the proton, neutron, electron, photon, and the (electron anti-) *neutrino*. The electron (e^-) was the first to be identified, having been discovered in the nineteenth century as the **cathode ray** in cathode-ray tubes – the same *CRT*s were later used in television sets and computer monitors. The exact date of discovery of the photon (γ) is debatable. In some sense it was "discovered" by Einstein in 1905, who introduced the concept of a quantum of the electromagnetic field when he used it to explain how light can knock electrons off the surface of metals. The proton (p) and neutron (n) were discovered during the development of nuclear physics in the first third of the twentieth century. The existence of what we now know to be the electron antineutrino ($\bar{\nu}_e$) was postulated by Austrian physicist Wolfgang Pauli in 1931, though it was not actually discovered until 1956.

In 1930, Dirac pointed out that every particle has an "opposite" or **antiparticle** with the same mass, but opposite charge. So complementing the first three particles in the *standard model lite* are the antiproton ($\bar{p}$), antineutron ($\bar{n}$), and antielectron (e^+, also known as the **positron**). The $\bar{\nu}_e$ is an *antineutrino*, so its opposite is the ν_e, the *neutrino*. The photon is unique in this list in that it is *its own antiparticle*, so there is no separate "antiphoton." Antiprotons and antineutrons play almost no role in our world. They are hard to produce and annihilate quickly into radiation when they encounter any matter. Positrons are emitted in some nuclear decays and have found use in medical diagnostics and cancer treatment. Neutrinos are produced copiously in the Sun (see §22) but pass through our world largely unnoticed. More about positrons and neutrinos can be found in §20 and §22.

The description of the world in terms of the five particles in the *Standard Model lite* is really quite simple:

Protons and neutrons bind together to form nuclei. Nucleon–nucleon forces are complicated, but nuclei are rather simple – at least at the level we encounter them. Nuclei can be described as a quantum version of "liquid drops" in which the protons and neutrons slither about. Note that while protons and neutrons are not truly fundamental particles – they are bound states of quarks and do not appear in Figure 14.4 – only experiments at very high energies can discern the structure within them, so for most practical purposes they can be thought of as primitive constituents of matter.

Electrons are attracted to nuclei by forces that are transmitted by electric and magnetic fields. Together, electrons and nuclei formed from protons and neutrons form atoms; atoms form molecules; and molecules form complex systems like us.

Photons – the quanta of the electromagnetic field – are emitted and absorbed when electrons or protons make transitions from one quantum state to another.

Electromagnetic fields play a dual role: as classical waves that can be broadcast and absorbed, and as quantized photons as already mentioned.

Antineutrinos A neutron can decay into a proton, an electron, and an electron antineutrino. This is a weak interaction known as *β-decay*. Sometimes, when protons and neutrons are bound into nuclei, this reaction is forbidden by energy conservation and instead the opposite reaction

The Standard Model "Lite"

Only five particles – the proton, neutron, electron, photon, and the electron antineutrino – are needed to provide a fairly good description of almost all phenomena in our world. Electromagnetism and gravity are the principal macroscopic forces of consequence. Weak and strong interactions operate within nuclei, which are held together by the strong nuclear interaction between nucleons (protons and neutrons).

can occur: effectively a proton can decay into a neutron, a positron, and an electron neutrino (ν_e).

Gravity acts on all these particles, even photons which are massless but carry energy.

For a short period of time – roughly the period between 1935 and 1950 – the five particles in the *Standard Model lite* were thought to be the only elementary particles in nature. In fact, this relatively complacent era in fundamental physics was over almost before it began, since the first additional particle, the *muon*, had already been detected in a cloud chamber photograph in 1937. Trying to understand the muon and the plethora of strongly interacting particles related to protons and neutrons that were made in accelerators in the 1950s led eventually to the Standard Model.

The remainder of this section provides a brief introduction to the other particles in the Standard Model.

14.2.2 Leptons

Charged leptons All six known leptons experience the weak, but not the strong interactions. Three of these leptons also carry electromagnetic charge. The electron is the lightest and most familiar of the (electromagnetically) charged leptons. We know of two other charged leptons: the **muon** (μ) with a mass of approximately 207 times the mass of the electron, and the **tauon** (τ) with a mass about 3478 times the mass of the electron. As far as we can tell, the μ and the τ are almost identical copies of the electron, except for their masses. They have the same charges, and the same kinds of interactions; they are just heavier. They are short-lived and decay into electrons, neutrinos and, occasionally, other stuff. There is good reason to believe that the electron, muon, and tauon are the only charged leptons.

Neutral leptons Each charged lepton is associated with a *neutral lepton*, called a **neutrino**. For the e there is the ν_e, for the μ, the ν_μ, and for the τ, the ν_τ. What is meant by "associated with" is explained in more detail in §14.3. Neutrinos feel only the weak force and gravity. They are very light – for many years they were thought to be massless like photons, but late last century it was discovered that they have small but nonzero masses. Their interactions are so weak that they pass through matter almost like it isn't there. To make that statement quantitative: the *mean free path* (the typical distance traveled before an interaction takes place, §20) of a 1 MeV neutrino moving through lead is roughly 1 light-year ($\sim 10^{16}$ m)!

Lepton interactions An important tool in analyzing systems of elementary particles is the existence of additional exactly or approximately *conserved charges* like the total electric charge, which remain unchanged under all, or almost all, physical processes. Each of the three lepton pairs, (e, ν_e), (μ, ν_μ), and (τ, ν_τ) is associated with a discrete *approximately* conserved quantity.[6] This quantity is called **lepton number** (L) generically: **electron number**, **muon number**, and **tauon number** (L_e, L_μ, L_τ) refer to the approximately conserved quantities associated with each of the three lepton types. The charged leptons and their associated neutrinos have lepton number $+1$, their antiparticles have lepton number -1. An example of an interaction that conserves both muon number and electron number is muon decay,

$$\mu \to e\, \bar{\nu}_e\, \nu_\mu\,. \tag{14.7}$$

The muon ($L_\mu = +1$) decays into a muon neutrino ($L_\mu = +1$), an electron ($L_e = +1$), and an electron antineutrino ($L_e = -1$). Thus the muon number is one on both sides of the equation and the total electron number is zero.

Muons are not particularly exotic. They are made by cosmic rays in the upper atmosphere and travel down well into the ground (on average), before decaying or interacting. The lifetime of the muon is about 2.2 µs, and the flux at sea level is about $200\,\text{m}^{-2}\,\text{s}^{-1}$. So hundreds have passed through you in the time it has taken you to read these words.

---

[6] A subtle phenomenon called *neutrino oscillation* allows one type of neutrino to turn into another. The phenomenon is of great interest to particle physicists, but is extremely difficult to observe and has no impact on the energy processes such as β-decay that we are interested in here. There are also more subtle effects in quantum field theory that can lead to violation of lepton number; these effects are negligible in all circumstances considered in this book.

14.2.3 Quarks and their Interactions

Quarks are massive particles that feel the chromodynamic force. While they are similar in some respects to leptons at a fundamental level, the role they play in nature is quite different. Whether or not a particle feels the chromodynamic force is determined by a "charge" – not electric charge, but similar in that particles with this charge feel the force and particles without it do not. This charge is called *color*, but it has nothing to do with visible color, which is, of course, a characteristic of electromagnetic radiation. Electric charge comes in two types, which we call *positive* and *negative*, with the property that combining positive and negative charges of equal strengths results in a system that is electrically neutral. Similarly, each type of quark comes in three versions which differ by their "color." If three quarks, one of each version, are combined, the result is neutral under chromodynamic forces. Because this resembles the rule for addition of the primary colors of light – when red, green, and blue light are combined, they yield white, or neutral light – physicists chose, somewhat whimsically, to call the chromodynamic charge **color**. The proton and the neutron are both examples of color-neutral three-quark systems. Another successful feature of the analogy is that combining a quark and antiquark also can yield a color-neutral system, in much the same way that combining a visual color with its complement gives color-neutral, i.e. white, light.

The chromodynamic forces are very strong. Furthermore, as discussed further below, gluons themselves carry color charge (unlike photons, which are electrically neutral). These effects combine in such a way that quarks can never escape from the color-neutral bound states that they form. Thus, individual quarks are never seen. Color-neutral states of three quarks are called **baryons** and bound color-neutral states of a quark and an antiquark are called **mesons**. Together, these color-neutral composite particles are known as **hadrons**.[7] Protons and neutrons are the lightest baryons. Almost all the known particles that feel the strong force fit into these two categories: baryons (or antibaryons, like the antiproton) or mesons. The exception being a handful of recently discovered and/or very short-lived states that appear to be essentially composed of two quarks and two antiquarks, or four quarks and one antiquark.

Quark Properties

Baryon number There are literally hundreds of different types of baryons, but all are unstable except for protons (and neutrons when they are bound within certain nuclei). Eventually, if you follow its life, any baryon decays into a proton or neutron plus a lot of other stuff that may include mesons, leptons, antileptons, and photons. Protons, however, do not seem to decay – ever. Searches for proton decay have come up empty, and the proton lifetime is now known to be greater than 2×10^{29} yr [69]. To account for this, physicists endow baryons with a type of "charge" called **baryon number**, which is +1 for baryons and -1 for antibaryons. So, for example, the proton and neutron have baryon number +1 and the antiproton and antineutron have baryon number -1. The stability of the proton can then be expressed as **conservation of baryon number**.[8] A baryon can decay, but among its decay products one must always find another baryon, and if you wait long enough only protons and/or neutrons will be left. Mesons, leptons, photons, and the rest have baryon number zero. A proton and an antiproton can annihilate into a cloud of mesons and photons because this conserves baryon number: $1 + (-1) = 0$.

The concept of baryon number can be attributed at a more fundamental level to quarks. Since there are three quarks in a baryon, each quark is assigned baryon number $1/3$. While quark number conservation implies conservation of baryon number, the name "baryon number" was invented long before quarks were discovered, and the name stuck. Nuclei are bound states of protons and neutrons, so they can be labeled by their baryon number. Carbon, for example, with six protons and six neutrons, has baryon number 12. The baryon number of a nucleus is thus the same thing as its *atomic mass number* in the language of chemistry. We use the terms interchangeably.

Baryon Number (Atomic Mass)

An atomic nucleus can be labeled by the number of nucleons it contains. That number is called the *baryon number* or *atomic mass*, and is usually denoted by A. A nucleus of atomic mass A contains $3A$ quarks. Baryon number is conserved in all observed processes.

---

[7] All these names are from the Greek: *hadron* = stout; *baryon* = heavy; *meson* (originally *mesotron*) = intermediate; *lepton* = slight.

[8] Some speculative ideas for physics beyond the Standard Model, such as *Grand Unified Theories*, predict a very small probability for proton decay. These ideas have not been confirmed experimentally. Also, as for leptons, subtle processes can give rise to very small violations of baryon number; while negligible for the purposes of energy physics, these processes may have been involved in the early universe in producing the presently observed asymmetry between matter and antimatter (§21).

Quark charges and colors, gluons Aside from the fact that they are never seen in isolation, perhaps at first sight the most bizarre property of quarks is that their electric charges are either $+\frac{2}{3}e$ or $-\frac{1}{3}e$, where $-e$ is the electron's charge. When quarks were originally proposed in 1964 by American physicists Murray Gell-Mann and George Zweig, the idea that a particle could have *fractional electric charge* was considered strange if not silly. Now there is plenty of evidence for these charge assignments, and the only oddity is that fractional charges are not observed directly because the allowed, color neutral, combinations of quarks and antiquarks always have charges that are integer multiples of e.

Given that they are never seen in isolation, how can we be certain that quarks have fractional charge, or indeed, that quarks exist at all? The reason is that when electrons or other particles that do not feel the strong force are scattered from protons or other nuclei, the pattern of scattering depends on the nature and distribution of the constituents of the baryon(s). In the late 1960s American physicists Jerome Friedman, Henry Kendall, and Richard Taylor scattered electrons from protons and observed scattering patterns that could only be accounted for if protons were made of pointlike spin-1/2 particles with electric charges that are third integral fractions of the electron's charge. Physicists now agree that quarks are just as "real" as electrons, even though the nature of chromodynamic force sequesters them permanently inside protons and neutrons.

The gluons that mediate the chromodynamic force couple to color charges in much the same way that a photon couples to electric charge, but in addition, gluons themselves are colored. Two beams of light pass through one another because photons do not scatter significantly from one another.[9] Gluons not only do not pass through one another, they interact with one another strongly by emitting and absorbing more gluons, and those interact by emitting and absorbing more gluons, *ad infinitum*. This is why chromodynamic forces are so complicated.

Flavor On top of all the previous complications, quarks come in six distinct types, more or less in parallel to the six types of leptons (remember (e, ν_e), (μ, ν_μ), and (τ, ν_τ)). The six types are called **up** (u), **down** (d), **charm** (c), **strange** (s), **top** (t), and **bottom** (b). They differ in electric charge: three (u, c, t) have charge $+\frac{2}{3}e$ and three (d, s, b) have charge $-\frac{1}{3}e$. The properties of the quarks, to the extent they are known, are listed in Table 14.2. Note that they all have electric charge, so they feel electromagnetic forces. They all feel the weak force as well.

These six different types of quarks are known as different **flavors** of quarks. The name "flavor" is even more fanciful than "color." It has no content except to allow physicists to discuss an abstract concept in simple words. Note that the six flavors of quarks can be divided into the three *generations* of matter as discussed above, with up and down quarks in the first generation, charm and strange in the second generation, and top and bottom in the third generation, with each generation progressively more massive than the previous.

The chromodynamic interactions of quarks conserve flavor. The number of u quarks (minus the number of u antiquarks), for example, does not change in a chromodynamic interaction. In contrast, the weak interactions allow quarks to change their flavors. Thus, for example, a weak interaction can enable a d quark to emit an electron and an electron antineutrino and turn into a u quark – a transformation not otherwise possible. These transformations are important in nuclear energy processes, and are described in §14.3.

The quark structure of protons and neutrons Protons and neutrons are the lightest baryons. The **proton** is a bound state of two u quarks and one d quark. The sum of the electric charges is $2 \times \frac{2}{3} - \frac{1}{3} = 1$. The **neutron** is a bound state of two d quarks and one u quark, with electric charge $\frac{2}{3} - 2 \times \frac{1}{3} = 0$. This structure is illustrated in Figure 14.5. The proton is lighter than the neutron because the u quark is slightly lighter than the d quark. The masses of the u and d quarks are tiny, which at first sight may seem very strange: how can objects of mass $M \approx 940\,\text{MeV}/c^2$ be made of only three particles with masses of $\sim$ 2–5 MeV/c^2 each? Where does all the mass come from? The answer lies in quantum mechanics and the concept of zero-point energy, which was briefly mentioned in §9.2.1. Recall that in each of the potentials that was studied (square-well and harmonic oscillator) in §7, the lowest energy state has energy greater than zero. The smaller the region to which a particle is confined, the greater its energy. For a region as small as 1 fm, the zero-point energy of a massless particle is ~300 MeV (Problem 14.5). The proton contains three such particles, each with a zero-point energy of about 300 MeV, so the energy of a proton at rest is ~900 MeV; by Einstein's $E = mc^2$, that energy contributes 900 MeV/c^2 to its mass.

The visible matter in the universe consists of nuclei and electrons. The nuclei in turn consist of protons and neutrons. The mass of an electron is ~1/2000th that of a

---

[9] There is a tiny quantum effect called *light-by-light scattering*, where a photon briefly materializes as an electron-positron pair, which then interacts with the other photon. This effect is so small that it has not been observed directly.

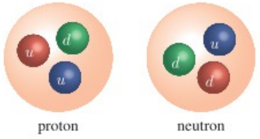

Figure 14.5 A cartoon of a proton and neutron showing the dominant quarks: uud in the proton and ddu in the neutron. The quarks are colored to remind us that all three colors must be present to make a color-neutral hadron. In actuality, protons and neutrons are described by quantum states that include fluctuating gluon fields and quark–antiquark pairs, though this complexity does not affect the nuclear reactions relevant to nuclear energy processes.

Mass in the Visible Universe

Almost all of the visible mass in the universe is found in the rest masses of protons (hydrogen nuclei), neutrons, and the nuclei they compose. This mass, in turn, is a manifestation of the quantum zero-point energy of almost-massless quarks confined to the interior of nucleons.

proton or neutron. So we conclude that most of the mass of the visible universe is due to the quantum zero-point energy of bound quarks!

Although protons and neutrons have excited states, the energies emitted and absorbed in the nuclear reactions of interest to us here are not great enough to excite the quarks in protons or neutrons. Protons and neutrons can therefore, to a good approximation, be regarded as persistent, unstructured objects whose identity does not change inside a nucleus. The internal quark substructure of protons and neutrons is manifested in the complicated structure of the strong nuclear interactions.

14.2.4 Force Carriers

As summarized earlier, the chromodynamic, weak, and electromagnetic forces each are mediated by spin one *vector bosons*. Each of the vector bosons has its own features and character. The *photon* is most visible because it is massless and carries no charge, so that it propagates freely through space mediating the long-range electromagnetic force. The gluon, while also massless, is itself *charged* under the chromodynamic interactions, so that it cannot propagate freely. Even in the absence of quarks or leptons, gluons can in principle bind together into color-free massive states known as *glueballs*. Up to now, however, no glueball has been convincingly identified experimentally. The Z^0 and $W^\pm$ vector bosons that mediate the weak interactions are massive, and thus weak interactions are intrinsically short-range forces.

14.2.5 The Higgs Boson

The final particle in the standard model is the *Higgs boson*, or *Higgs particle*. Originally predicted in 1964, this particle was found experimentally in 2013. Though this particle does not play a direct role in any energy-related processes, it is a central component of the Standard Model. It is responsible for *electroweak symmetry breaking*, the mechanism that gives rise to masses for the $W^\pm$ and Z bosons and differentiates the weak interactions from the closely related electromagnetic interaction. The Higgs particle is also associated with the phenomenon that gives rise to quark and lepton masses.

14.3 The Weak Interactions and β-decay

We have alluded several times to the crucial role that the weak interactions play in the universe. Their fundamental practical importance is that they allow protons to turn into neutrons and *vice versa*. In this section we take a more careful look at this and related processes.

14.3.1 Discovery of β-decay

In the early twentieth century physicists observed that some nuclei emit electrons and change their chemical identity. A simple, illustrative, and important example is the decay of a rare form of potassium into calcium or argon. Potassium is an atom with 19 electrons bound to a nucleus with 19 protons and usually 20 neutrons. This nucleus is stable. About one in ten thousand potassium nuclei contains 21 neutrons rather than 20, giving the nucleus a baryon number (atomic mass number) of 40. An atom with this nucleus at its core has all the chemical properties of ordinary potassium, but it weighs about 2.5% more due to the extra neutron. What is special about potassium-40 (^{40}K in standard notation, §17.1.2) is that it is unstable. It decays with a lifetime of greater than a billion years.

About 88% of the time that the nucleus of a ^{40}K atom decays, it emits an electron, thereby increasing its nuclear charge by one unit. This process turns it into a form of calcium, ^{40}Ca. At first, physicists did not know what particle was emitted in this process. They called the emitted particle a β-ray and they called the decay a **beta decay** Soon it became clear that the emitted particle was an ordinary electron. But a new puzzle arose: the electron did not always come out with the same energy. According to quantum mechanics, the decay should have taken one quantum state of fixed energy to another, and the electron's energy should have been determined by the difference of the energies of the two states. Even if the transition left the calcium nucleus in one of several excited energy basis states, the energy of the electron would have been limited to one of a few definite values. Instead the electrons' energies measured in many ^{40}K decays were found to vary continuously between zero and a maximum that equaled the difference of nuclear energies.[10] It seemed that energy was leaking away. Further experiments showed that momentum and angular momentum were not being conserved in β-decay either. In 1931, Wolfgang Pauli speculated that an *unobserved* neutral particle was being emitted *with the electron*, and it was carrying off the missing energy, momentum, and angular momentum. To have been overlooked, the neutral particle would have to have had *very weak interactions with ordinary matter*. Fermi named it "a little neutral one" or *neutrino* in Italian. This was the first time (but not the last) that a physicist predicted the existence of a new elementary particle. Further experiments confirmed that Pauli was right and that the reaction that was taking place is

$$^{40}\text{K} \to {}^{40}\text{Ca} + e^- + \bar{\nu}_e\,. \qquad (14.8)$$

Many nuclei undergo similar decays.

14.3.2 β-decay in the Standard Model

When physicists of the mid twentieth century figured out that the nucleus is a bound state of protons and neutrons, they developed a more universal description of what is happening in ^{40}K and other β-decays. In such decays, a neutron is decaying into a proton,

$$n \to p + e^- + \bar{\nu}_e\,. \qquad (14.9)$$

The proton remains bound by the strong nuclear force to the nucleus, but the electron and the neutrino are emitted. And when physicists of the late twentieth century understood that the neutron and proton are bound states of quarks, they realized that, at a deeper level still, a d quark is decaying into a u quark by a weak interaction,

$$d \to u + e^- + \bar{\nu}_e\,. \qquad (14.10)$$

Since quarks are confined to the interiors of the proton and neutron, the observed reaction is

$$n(udd) \to p(uud) + e^- + \bar{\nu}_e\,. \qquad (14.11)$$

Weak Interactions Violate Flavor Conservation

The strong interactions appear to conserve flavor absolutely, but weak interactions allow quarks of one flavor to turn into quarks of any other flavor, as long as the quark electric charge changes by ± 1 and the overall reaction conserves baryon number, electric charge, and lepton number. The transformations between u and d quarks lead to β-decay and related processes in nuclei,

$$\begin{aligned} d &\to u + e^- + \bar{\nu}_e, \\ u &\to d + e^+ + \nu_e, \\ u + e^- &\to d + \nu_e. \end{aligned}$$

The Standard Model gives quite a simple description of what is happening at the quark level. The modifications due to the details of nuclear structure, however, are very important in determining the lifetime of any particular nucleus that can undergo β-decay. Some nuclei that can β-decay have lifetimes that are longer than the age of the universe. The free neutron β-decays to the proton with a lifetime of about 10 minutes, and there are many nuclei with β-decay lifetimes of less than a microsecond. We do not attempt to develop the theory of β-decay here in any detail. Lifetimes for β-decay processes in which we are interested can be looked up online [70].

The fact that quark flavor is not conserved ($d \to u$) in β-decay confirms that the weak interactions are at work. There are many different types of weak interactions, but the only ones important for our purposes are the one given in eq. (14.10) and a few variations mentioned below. These reactions always conserve electric charge, baryon number, and lepton number as well as the other usual conserved quantities – energy, momentum, and angular momentum – but quark flavor is not conserved. The fact that weak interactions enable reactions that would otherwise be forbidden is the reason that we care about them. When they give small corrections to otherwise allowed processes, as they do under most circumstances, they are usually too small to be noticed.

14.3.3 Processes Related to β-decay

^{40}K β-decays to ^{40}Ca 88% of the time. The other 12% of the time it undergoes a related process that turns it into ^{40}Ar. In this case the nuclear charge must decrease by one unit (argon has 18 protons and 18 electrons). From what

---

[10] Actually the experiments were performed on different, more abundant nuclei with shorter lifetimes. We chose ^{40}K because it both β-decays *and electron captures* (see below) and it plays a role in geothermal energy (see §32).

Example 14.1 What Force Was Responsible For That?

Interactions among particles must satisfy the conservation laws that are respected by the type of interaction that is at work. This fact – together with the hierarchy that the strong nuclear interaction is much stronger than electromagnetism, which in turn is much stronger than the weak interactions – makes it possible to identify what force is at work in important processes. In general, the rate at which a quantum process occurs depends on the strength of the coupling, so most strong nuclear interaction processes occur much more quickly than those mediated by weak interactions. (Note that gravity is too weak to drive any measureable reaction among particles or nuclei.)

Some important examples:

- When neutrinos or antineutrinos appear, the weak interactions must be at work, because neutrinos do not couple to the strong or electromagnetic interactions.
- When photons appear, electromagnetism must be at work because photons only interact electromagnetically.
- If protons change into neutrons or *vice versa*, the weak interactions must be at work.

Consider, for example, hydrogen fusion into helium, which is known to be the Sun's primary energy source. Four protons fuse to form one helium nucleus, known also as an α-**particle**, which consists of two protons and two neutrons, as well as two positrons (needed to conserve electric charge). A first guess for the reaction might be

$$4p \to \alpha + 2e^+ \quad (?)$$

This reaction conserves baryon number and electric charge. Naively it looks like a strong nuclear interaction, which should go very quickly. One wonders why the Sun has lived so long. Closer inspection, however, reveals that two protons have turned into neutrons and that electron number is not conserved. These features indicate both that a weak interaction is involved, and that two neutrinos must be produced. A corrected version of the reaction reads

$$4p \to \alpha + 2e^+ + 2\nu_e \,.$$

The reaction can only proceed as rapidly as the weakest force involved permits, so despite the fact that protons and neutrons are involved, the time scale for hydrogen burning in the Sun is set by the weak interactions, which is a good thing for life on Earth.

we have learned about the weak interactions of quarks, we might guess that the reaction could be

$$^{40}\mathrm{K} \to \, ^{40}\mathrm{Ar} + e^+ + \nu_e \quad (?) \tag{14.12}$$

where at the microscopic level $u \to d + e^+ + \nu_e$. This is a good guess, which satisfies all the rules like lepton number and baryon number conservation. This kind of reaction is called β^{+}_{-}*decay* or *positron emission*, but it is *not* in fact how ^{40}K decays to ^{40}Ar. A quick check of the masses of the ^{40}K and ^{40}Ar nuclei reveals that there is not enough energy released in the nuclear reaction to create the rest mass of the positron, 511 keV/c^2. (We explain how to do calculations like this in §17.) Instead, the ^{40}K nucleus *captures* an electron from the surrounding cloud to produce the reaction

$$^{40}\mathrm{K} + e^- \to \, ^{40}\mathrm{Ar} + \nu_e \,, \tag{14.13}$$

which satisfies all the rules of the weak interactions. This process is called *electron capture* or *K-capture* ("*K*" because the electron is captured from the innermost atomic shell, known to chemists as the *K*-shell). This reaction can proceed even if the rest mass of ^{40}Ar is *greater* than that

Important β-decay Processes

The summary of interesting β-decay and related processes for nucleons is

β^- decay	$n \to p + e^- + \bar{\nu}_e,$
β^+ decay	$p \to n + e^+ + \nu_e,$
electron capture	$p + e^- \to n + \nu_e.$

The processes are described here in terms of protons and neutrons, but the fundamental processes involve quarks. In applications to nuclear physics, the protons and neutrons are usually bound in nuclei.

of ^{40}K, as long as the sum of the ^{40}K rest mass plus the electron's rest mass exceeds the rest mass of ^{40}Ar. Many nuclei decay by β^+-decay and/or electron capture, with lifetimes that vary from very short to longer than the age of the universe.

Discussion/Investigation Questions

14.1 Physicists searching to see if the proton decays have looked for the reaction $p \rightarrow e^+\gamma$. Which of the conservation laws explored in this chapter does this decay violate?

14.2 Gravity is a long-range force proportional to the masses of the interacting bodies. Physicists have long searched to see if there might be a long-range force that is proportional to baryon number. Think about the composition of atoms and propose a "thought experiment" that could detect such a force.

14.3 Two possible nuclear fusion reactions are $d + d \rightarrow {}^4\text{He} + \gamma$ and $d + d \rightarrow {}^3\text{He} + p$, where d is deuterium, an atom with a nucleus composed of a proton bound to a neutron. ^{3}He and ^{4}He are helium atoms with nuclei consisting of two protons and one neutron (^{3}He) or two protons and two neutrons (^{4}He). Both of these reactions satisfy energy conservation. Which do you think occurs more quickly and why? (This question and others like it arise when considering fuel for nuclear fusion on Earth.)

14.4 A *baryon* frequently encountered in particle physics experiments is known as the Λ (*"Lambda"*). Its quark content is uds. Sometimes the Λ decays by the process $\Lambda \rightarrow n\gamma$, where n is a neutron (ddu) and γ is a photon. What interactions must be active in this decay?

14.5 The tauon τ decays often to a muon μ, which in turn decays to an electron. How many neutrinos, and of which type are produced in this sequence of weak decays.

Problems

14.1 A rocket fuel should have high energy density (per kilogram) because the weight of the propellant is an important consideration. A favorite fuel is liquid hydrogen plus liquid oxygen. What is the energy density (J/kg) of a stoichiometric mixture of H_2(l) and O_2(l) (assuming they combine to form water in the gaseous state)? The strong nuclear force, in principle, allows the same stoichiometric mixture of two hydrogen atoms and one oxygen atom to fuse to form an *isotope* of neon, ^{18}Ne, giving off 4.52 MeV in the process. If this reaction were the source of power for the rocket, what would be the energy density of the fuel?

14.2 In principle, and in the absence of electromagnetic or any other non-gravitational interactions, an electron and a proton should form a "gravitational atom," bound by the force of gravity. We know that the energy of the ground state of the ordinary hydrogen atom is given by eq. (14.4). Show that the ground state energy of gravitationally bound "hydrogen" is given by $E_{\text{grav}} = -\frac{1}{2}G_N^2 m_p^2 m_e^3/\hbar^2$ and evaluate this numerically.

14.3 Show that the combination of constants $g_{\text{strong}}^2/(4\pi\epsilon_0\hbar c)$, where g_{strong} is defined in eq. (14.5), is dimensionless. It is taken as a dimensionless measure of the strength of an interaction. Compute this quantity for electromagnetism ($g \rightarrow e$), where it is known as the *fine structure constant* and denoted by α.

14.4 Which of the following interactions are consistent with the conservation laws for baryon number, electric charge, and the three lepton numbers? Don't worry about energy conservation. (A bar denotes an antiparticle. See Question 14.4 for information about the Λ particle.)

$$\begin{array}{ll}
p \rightarrow n + e + \nu_e & p \rightarrow n + \bar{e} + \bar{\nu}_e \\
\tau \rightarrow \mu + \bar{\nu}_\mu + \nu_\tau & \bar{\tau} \rightarrow \mu + \bar{\nu}_\mu + \nu_\tau \\
\nu_\mu + p \rightarrow \bar{\mu} + n & \nu_\mu + n \rightarrow \mu + p \\
\Lambda \rightarrow p + e + \bar{\nu}_e & \Lambda \rightarrow n + \nu_e + \bar{\nu}_e
\end{array}$$

14.5 Einstein's energy–momentum relation for a particle of mass m with momentum $\boldsymbol{p}$ is $E = \sqrt{\boldsymbol{p}^2c^2 + m^2c^4}$. Estimate the uncertainty in the x, y, and z components of the momentum of a quark confined in a (cubic) box of side length $L = 1$ fm using the Heisenberg uncertainty principle $\Delta p_{x,y,z} \gtrsim \hbar/L$ (Box 7.6). Approximate $\langle \boldsymbol{p}^2 \rangle$ by $(\boldsymbol{\Delta p})^2$, then show that for a quark of mass 5 MeV/c^2, $E \approx |\boldsymbol{p}|c$ and that the zero-point energy of three quarks is of order 900 MeV, the rest energy of a proton.

14.6 To get a sense of the potential practical relevance for energy purposes of one aspect of as-yet-undiscovered physics, consider proton decay. Making the most general assumptions about the nature of its decay, the lifetime of the proton is currently bounded by $\tau_{\text{proton}} > 2.1 \times 10^{29}$ y. Assume that the lifetime is in fact 3×10^{29} y, and that all of the proton's rest energy emitted in the decay can be captured and put to use. Estimate the rate of useful power output from proton decay of a cubic kilometer of hydrogen gas at NTP.

21.04.2020

B) Books_70_2020/
Lightware_Communications/
Lightware Communications.pdf

C) Books_70_2020/
The Physics of Energy/
The Physics of Energy.pdf

Pag. 52(72)

3.6 Maxwell's Equations

Pag. 263(283)

The Forces of Nature

Pag. 266(286)

Figure 14.2 Feynman diagram describing two electrons interacting exchange of a photon, the force carrier for electromagnetic interactions.

D) Books_70_2020/
The Universe of Quadrics/
3662610523.pdf

Pag. 147(155)

Quadrics in the real projective space

!)*+,)))-) #& $66 6 $ 5!
-)*+,)))-) #& 2 = / > 4
. 8 0 ! / 5 61 5 5 8 5!
' #;

and the restriction of γ_1 onto l_j with fixed points A and A_j. Since these two fixed points are corresponding under π_{l_j}, the two involutions commute (see, *e.g.*, [46, Exercise 5.4.1]). Therefore, when the points $X, Y \in l_j$ are conjugate w.r.t. π, *i.e.*, $Y = \pi_{l_j}(X)$, then $\gamma_1(Y) = \gamma_1 \circ \pi_{l_j}(X) = \pi_{l_j} \circ \gamma_1(X)$. This means that $\gamma_1(X)$ and $\gamma_1(Y)$ are also conjugate w.r.t. π or, in other words, the induced involutions of π and π' are the same.

With regard to the skew involution γ_2 with axes l and $l^* := \pi^*(l)$, we choose two different, but mutually conjugate points A and A' on l and show that γ_2 is the product of the two harmonic homologies with centers A and A' and respective axes $\pi(A)$ and $\pi(A')$.

Under the product of the two homologies, all points on $\pi(A) \cap \pi(A') = l^*$ remain fixed, but so do all points on $l = [A, A']$, since the restrictions of both harmonic homologies to the line l are involutions with the same fixed points A and A'. Furthermore, the two harmonic homologies commute since the center of one homology lies in the axial plane of the other. Consequently, the product of the two homologies is involutive and biaxial, and therefore, the skew involution with the axes l and l^*. ■

In the particular case $\pi = \pi_{\mathcal{Q}}$, we can state:

Corollary 4.2.7 *Each quadric $\mathcal{Q}$ remains fixed under all harmonic homologies with a point $A \notin \mathcal{Q}$ as center and its polar plane w.r.t. $\mathcal{Q}$ as axial plane.*

If l and its polar line l^* w.r.t. $\mathcal{Q}$ are skew, then the skew involution with axes l and l^* maps $\mathcal{Q}$ onto itself.

Collineations which keep a quadric $\mathcal{Q}$ fixed are called *automorphisms* of $\mathcal{Q}$.

Quadrics in the real projective space

Let $A_0A_1A_2A_3$ be a self-polar tetrahedron of an elliptic or hyperbolic polarity π in the real projective space $\mathbb{P}^3(\mathbb{R})$. Then, each of its triangles is self-polar w.r.t. the induced polarity in the spanned plane ε, and this planar polarity can be elliptic or hyperbolic. We recall (note [46, p. 266]):

Lemma 4.2.8 *If $A_1A_2A_3$ is a self-polar triangle of a polarity π_ε in a real projective plane ε, then, in the case of an elliptic polarity, the involution of conjugate points on each side of the triangle is elliptic, while for a hyperbolic polarity, exactly on two sides the induced involutions are hyperbolic.*

We denote these two types of self-polar triangles by (eee) and (ehh), where 'e' stands for an elliptic involution and 'h' for a hyperbolic one (see [46, p. 278, Fig. 7.14]).

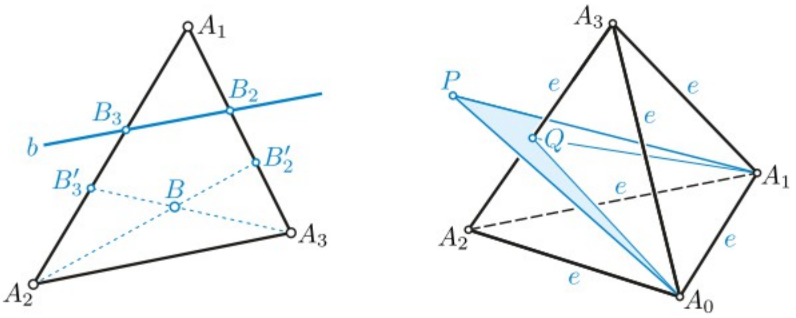

FIGURE 4.5. Left: Induced involutions of conjugate points on the sides of a self-polar triangle; Right: The point P is supposed to be self-conjugate.

Proof: We define π_ε by a self-polar triangle $A_1A_2A_3$ and an additional pair $B \mapsto b$ of pole and polar line, where $A_1, \dots, B$ is a quadrangle and $A_1, A_2, A_3 \notin b$. Then, on the side $[A_1, A_2]$, the involution of conjugate points maps A_1 onto A_2 and the intersection B_3 with b onto the point B_3' being aligned with B and A_3 (Figure 4.5, left). A positive cross-ratio $\operatorname{cr}(A_1, A_2, B_3, B_3')$ of the two pairs characterizes hyperbolic involutions, *i.e.*, real self-conjugate points on $[A_1, A_2]$. The same applies to the other sides. Due to the theorems of Ceva and Menelaos, the product of the cross-ratios on the three sides must be negative, which results in the two cases. ■

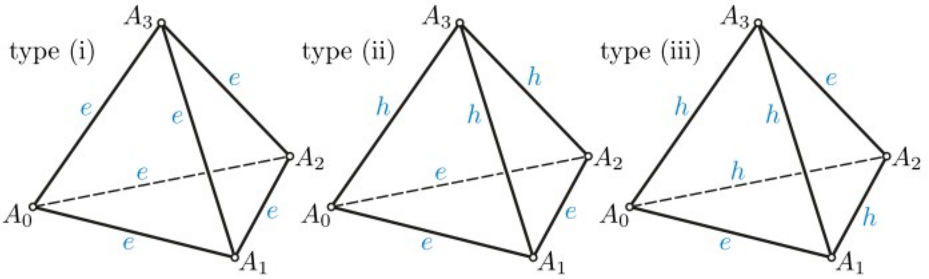

FIGURE 4.6. Induced involutions of conjugate points on the edges of a self-polar tetrahedron.

By virtue of Lemma 4.2.8, for the involutions induced by any elliptic or hyperbolic polarity π on the edges of a self-polar tetrahedron $A_0 \dots A_3$ in $\mathbb{P}^3(\mathbb{R})$, there remain three possibilities (see Figure 4.6):

type (i): all involutions are elliptic;

type (ii): the involutions are elliptic on exactly three coplanar edges;

type (iii): the involutions are elliptic on exactly two opposite edges.

If π is the polarity w.r.t. a ruled quadric $\mathcal{Q}$, then each induced polarity must be hyperbolic, since each generator of $\mathcal{Q}$ meets all planes at self-conjugate points. Therefore, each self-polar triangle must be of type (ehh), and consequently, all self-polar tetrahedra are of type (iii).

Conversely, any self-polar tetrahedron of type (iii) defines self-polar lines, because there exist self-conjugate points on opposite edges (l, l^*) and, by virtue of Lemma 4.2.2, the connection of a self-conjugate point $P \in l$ with a self-conjugate point $Q \in l^*$ is self-polar.

If π is elliptic, then the induced polarity must be elliptic in each plane, too. Therefore, each self-polar tetrahedron must be of type (i).

Conversely, the induced polarities are elliptic in the planes of any self-polar tetrahedron of type (i). Suppose that there is a self-polar point P outside the faces of $A_0 \dots A_3$ (see Figure 4.5, right). Then, the induced polarity in the plane $\varepsilon := [A_0, A_1, P]$ must be hyperbolic. Together with A_0 and A_1, the point of intersection $Q := \varepsilon \cap [A_2, A_3]$ forms a self-polar triangle of π_ε of type (ehh). This yields a self-conjugate point on the side $[Q, A_0]$ in the plane $[A_0, A_2, A_3]$, which contradicts our assumption.

What remains now is that all self-polar tetrahedra of oval quadrics must be of type (ii). We summarize:

Theorem 4.2.8 *All self-polar tetrahedra of ruled quadrics in $\mathbb{P}^3(\mathbb{R})$ are of type* (iii) *and those of oval quadrics are of type* (ii).

Let π be any elliptic or hyperbolic polarity in $\mathbb{P}^3(\mathbb{R})$ with a self-polar tetrahedron $A_0 \dots A_3$. When the vertices $A_0, \dots, A_3$ are used as base points of a coordinate frame, then the π representing matrix $\mathbf{M}$ already has diagonal form $\operatorname{diag}(a_{00}, \dots, a_{33})$ with $a_{00}, \dots, a_{33} \neq 0$, since $\pi_{\mathcal{Q}}\colon (1,0,0,0)\mathbb{F} \mapsto (1:0:0:0)\mathbb{F}$, and so on. By (4.6), the two points $X = (x_0 : x_1 : 0 : 0)$ and $X' = (x'_0 : x'_1 : 0 : 0)$ on $[A_0, A_1]$ are conjugate w.r.t. π if, and only if,

$$a_{00} x_0 x'_0 + a_{11} x_1 x'_1 = 0.$$

This symmetric bilinear form characterizes the induced involution on $[A_0, A_1]$, which, in the case $a_{00} a_{11} > 0$, is elliptic and otherwise hyperbolic. The same holds for all other edges of the tetrahedron.

If π is the polarity w.r.t. an oval quadric $\mathcal{Q}$, we can assume that the involutions on the edges through A_3 are hyperbolic. Now, we define the unit point such that the planes connecting one side of the triangle $A_0 A_1 A_2$ with E intersect the opposite edge at a self-conjugate point.

This means, *e.g.*, that the point $[A_0, A_1, E] \cap [A_2, A_3]$ with coordinates $(0 : 0 : 1 : 1)$ is self-conjugate, which implies that $a_{22} + a_{33} = 0$ and, similarly, $a_{00} + a_{33} = a_{11} + a_{33} = 0$.

If the induced involution on an edge through A_3, say on $[A_2, A_3]$, is elliptic, we specify the plane $[A_0, A_1, E]$ such that the conjugate point of $X = (0 : 0 : 1 : 1)$ is the fourth harmonic point w.r.t. A_2 and A_3, hence $X' = (0 : 0 : 1 : -1)$, which implies $a_{22} - a_{33} = 0$, and so on. Finally, we may set $a_{33} = -1$, which gives rise to the following standard forms.

Theorem 4.2.9 *For each polarity π in $\mathbb{P}^3(\mathbb{R})$, there is a coordinate frame such that the representing matrix $\mathbf{M}$ has one of the following forms:*

$$\mathbf{M} = \begin{cases} \operatorname{diag}(1,1,1,1) & \Leftrightarrow \quad \pi \text{ is elliptic,} \\ \operatorname{diag}(1,1,1,-1) & \Leftrightarrow \quad \pi = \pi_{\mathcal{Q}} \text{ and } \mathcal{Q} \text{ is oval,} \\ \operatorname{diag}(1,1,-1,-1) & \Leftrightarrow \quad \pi = \pi_{\mathcal{Q}} \text{ and } \mathcal{Q} \text{ is ruled.} \end{cases}$$

For the sake of completeness, we shall mention that even a symmetric 4×4 matrix $\mathbf{M}$ with rank $r < 4$ defines a symmetric relation called *conjugacy* between two points $\mathbf{p}\mathbb{R}$ and $\mathbf{q}\mathbb{R}$ by the condition (4.6), $\mathbf{p}^{\mathrm{T}}\mathbf{M}\,\mathbf{q} = 0$. There is a $(3-r)$-dimensional space of *singular points* $\mathbf{s}\mathbb{R}$ satisfying $\mathbf{M}\,\mathbf{s} = \mathbf{0}$. Singular points are conjugate to each point in $\mathbb{P}^3(\mathbb{R})$. When conjugacy is restricted to an $(r-1)$-dimensional space of non-singular points, then for $r = 3$, we obtain pairs of conjugate points w.r.t. a planar polarity and for $r = 2$ we obtain pairs of an elliptic or hyperbolic involution.

Theorem 4.2.10 *The set of points in $\mathbb{P}^3(\mathbb{R})$ satisfying $\mathbf{x}^{\mathrm{T}}\mathbf{M}\,\mathbf{x} = 0$ with a symmetric 4×4 matrix $\mathbf{M}$ of rank $r \leq 3$ is a singular quadric $\mathcal{Q}$.*
When $\mathbf{M}$ has rank deficiency 1, i.e., $r = 3$, this is a quadratic or imaginary cone with the apex at the singular point. In the case $r = 2$, we obtain either two real or two complex conjugate planes intersecting at the line of singular points. For $r = 1$, the set of singular points is a plane, which equals the set of self-conjugate points.
In the case of $r = 3$, it is the sets of tangent planes of a conic that are dual to the singular point quadrics; for $r = 2$, it is pairs of line bundles; and for $r = 1$, it is single line bundles.

Euclidean geometry in the projective setting

When the real projective space $\mathbb{P}^3(\mathbb{R})$ is introduced as the projective closure of the Euclidean 3-space $\mathbb{E}^3$, then it is endowed with additional structures. Below, we analyse some of them.

The orthogonality in $\mathbb{E}^3$ defines a bijectivity between bundles of parallel lines l and pencils of parallel planes ε orthogonal to l. This induces, in the plane ω at infinity, a bijective mapping $\nu\colon \mathcal{P}_\omega \to \mathcal{L}_\omega$ with $(l \cap \omega) \mapsto (\varepsilon \cap \omega)$ for $l \perp \omega$. If l runs through a pencil $\mathcal{L}_{A,\varphi}$ of lines, the corresponding planes $\varepsilon \perp l$ are parallel to lines f perpendicular to φ. Hence, ω is a correlation. Since the orthogonality $l \perp f$ between two lines is symmetric, the conjugacy between $l \cap \omega$ and $f \cap \omega$ is symmetric, too. Consequently, orthogonality in $\mathbb{E}^3$ induces a polarity ν in ω, which is called the Euclidean *absolute polarity.*

No real line can be orthogonal to itself. Therefore, the absolute polarity in ω is elliptic. In the complex extension, the self-conjugate points form the *absolute conic* j, which satisfies

$$x_0 = 0 \quad \text{and} \quad x_1^2 + x_2^2 + x_3^2 = 0, \tag{4.9}$$

when using homogeneous Cartesian coordinates $(x_0 : x_1 : x_2 : x_3) = (1 : x : y : z)$. Finite imaginary lines which meet j are called *isotropic* as well as finite planes tangent to j. Pairs of orthogonal lines in any pencil $\mathcal{L}_{A,\varepsilon}$ form an involution with fixed lines passing through the absolute circle points of ε. Hence, the absolute conic j comprises the absolute circle points of all finite planes ε in $\mathbb{E}^3$.

We recall LAGUERRE's formula from page 76: If l_1 and l_2 are two lines in the pencil $\mathcal{P}_{A,\varepsilon}$ and i and $\overline{i}$ are the two complex conjugate isotropic lines in this pencil, then the measure α of the angle enclosed by l_1 and l_2 is related to the cross ratio of the four lines by

$$\operatorname{cr}(l_1, l_2, i, \overline{i}) = e^{2\mathrm{i}\alpha}, \text{ where } \alpha = \sphericalangle\, l_1 l_2. \tag{4.10}$$

The dual formula holds for two intersecting planes $\varepsilon_1, \varepsilon_2$ and the two isotropic planes $\iota, \overline{\iota}$ through the line $\varepsilon_1 \cap \varepsilon_2$ in the form

$$\operatorname{cr}(\varepsilon_1, \varepsilon_2, \iota, \overline{\iota}) = e^{2\mathrm{i}\alpha}, \text{ where } \alpha = \sphericalangle\, \varepsilon_1 \varepsilon_2. \tag{4.11}$$

Collinear transformations κ which map the plane ω onto itself, are called *affine transformations*. If κ sends all pairs of orthogonal lines again to orthogonal lines, then κ is called a *similarity*. In this case, the restriction of κ to the plane ω keeps the absolute conic j fixed. Consequently, by virtue of LAGUERRE's formula, all angle measures between lines or planes remain fixed, too. Each triangle if mapped onto a similar triangle, *i.e.*, scaled with a certain factor, and this scaling factor is the same for all distances. In other words, for each similarity κ in $\mathbb{E}^3$, there exists a scaling factor f_κ with $\overline{\kappa(A)\kappa(B)} = f_\kappa \cdot \overline{AB}$. Similarities with scaling factor $f_\kappa = 1$ preserve all distances and are called *motions*, whether they are continuous transformations or discrete.

Important examples of motions are the *reflections in planes*. They are special harmonic homologies where the center is the ideal point orthogonal to the plane of the mirror. The *reflection in a line* g is a skew involution with axes g and g^*, where g^* is the ideal line of all planes orthogonal to g. Rotations about the axis g are examples of uniaxial collineations.

Affine and Euclidean classification of quadrics

From the viewpoint of the projective space $\mathbb{P}^3(\mathbb{F})$, the affine space $\mathbb{A}^3(\mathbb{F})$ can be defined by appointing any plane ω to be the 'plane at infinity' and focusing on elements outside ω, *i.e.*, on finite points, lines, and planes. Affine transformations are collineations which keep the plane at infinity fixed. In order to classify quadrics $\mathcal{Q}$ in the real affine space $\mathbb{A}^3(\mathbb{R})$, we have to study the relation of $\mathcal{Q}$ and $\pi_\mathcal{Q}$ to the plane ω.

If ω is self-conjugate, then the quadric $\mathcal{Q}$ is a *paraboloid*. Each finite line l through the point of contact with ω is called a *diameter line* of $\mathcal{Q}$. Its polar line l^* lies in ω, and the skew involution with the axes l and l^*, which maps $\mathcal{Q}$ onto itself, is an *affine reflection* in l. This characterizes l as an axis of affine symmetry. Each plane through l is the axial plane of an affine reflection which also leaves the paraboloid invariant.

Paraboloids can be oval or ruled and accordingly, they are called *elliptic* or *hyperbolic*. In the latter case, the paraboloid intersects ω along two lines, taken from different reguli. Therefore, all finite generators of each regulus of $\mathcal{Q}$ are parallel to a fixed plane.

If the pole M of ω is finite, then M is called the *center*, and the quadric is a *central quadric*. The harmonic homology with the center M and axis ω, which maps $\mathcal{Q}$ onto itself, is the reflection in M and shows that M is a center of symmetry. When $MA_1A_2A_3$ is a self-polar tetrahedron, then the edges through M form a *triplet of conjugate diameters.* The polar line of the diameter line $l_1 = [M, A_1]$ is $l_1^* = [A_2, A_3]$ at infinity. The affine reflection in l maps $\mathcal{Q}$ onto itself.

When the induced polarity in ω is elliptic, $\mathcal{Q}$ is an *ellipsoid.* The self-polar tetrahedra are of type (ii). Each line through M is an edge of a self-polar tetrahedron and carries two real points of $\mathcal{Q}$.

A hyperbolic polarity in ω belongs to a *hyperboloid* $\mathcal{Q}$. In this case, the tangent cone with apex M contacts the hyperboloid along its curve at infinity; this cone is called *asymptotic cone.* It depends on the type of the self-polar tetrahedra whether the hyperboloid $\mathcal{Q}$ is oval or ruled. In the first case, each triplet of conjugate diameters contains only one axis with real points of $\mathcal{Q}$, while the existence of a tetrahedron of type (iii) yields two axes with real points. Accordingly, the hyperboloids are two-sheeted or one-sheeted, respectively. Quadratic cones with an apex at infinity are, of course, quadratic *cylinders.*

It can be shown that, for each regular quadric in $\mathbb{A}^3(\mathbb{R})$, there exists a coordinate frame such that the equation can be reduced to one of the normal forms, as listed in Theorem 3.1.1, but where all coefficients α_i are equal to 1.

When classifying quadrics w.r.t. Euclidean transformations, we must additionally analyse the relation between the quadrics and the absolute polarity in ω.

In the case of a paraboloid $\mathcal{Q}$, let U denote its point of contact with ω. The absolute polar u of U has a polar line u^* w.r.t. $\mathcal{Q}$, and this is the *axis* of the paraboloid $\mathcal{Q}$, because the skew involution with axes u and u^* is the *reflection* in u^*, which maps the paraboloid onto itself. There are two cases to be distinguished:

(i) Either there are exactly two tangents t_1, t_2 at U which are polar w.r.t. $\mathcal{Q}$ and at the same time conjugate w.r.t. the absolute polarity. Then, the planes $[t_1, u^*]$ and $[t_2, u^*]$ are the only planes of symmetry. The reflection in each of these planes maps the paraboloid onto itself.

(ii) When the absolute polarity as well as $\pi_{\mathcal{Q}}$ induce the same involution in the pencil of lines $\mathcal{L}_{U,\omega}$ or, in other words, when the complex conjugate lines of intersection between an elliptic paraboloid and ω are tangents of the absolute conic, then each plane through u^* is a plane of symmetry, and $\mathcal{Q}$ is a paraboloid of revolution.

In the case of central quadrics $\mathcal{Q}$, the induced polarity in ω and the absolute polarity share at least one common self-polar triangle $A_1A_2A_3$ (see [46, p. 278]). Then, the diameter lines $[M, A_i]$ are axes of symmetry of $\mathcal{Q}$, and the self-polar tetrahedron is called the *principal self-polar tetrahedron* of $\mathcal{Q}$. In the following two cases, the axes of $\mathcal{Q}$ are not unique:

(i) When the quadric $\mathcal{Q}$, which can also be imaginary, has a two-point contact with the absolute conic, each real point A_1 on the common chord is a point of a principal self-polar tetrahedron, and $\mathcal{Q}$ is a surface of revolution, an ellipsoid or a hyperboloid.

(ii) When the induced polarity π_ω is the absolute polarity, then, in the case of a hyperbolic polarity $\pi_{\mathcal{Q}}$, the quadric $\mathcal{Q}$ is a *sphere*.

Finally, a few words about elliptic polarities π: On each line l through the pole M of ω, the induced involution is elliptic. Consequently, there exist two points $P, P' \in l$ which are conjugate w.r.t. π and symmetric w.r.t. M. The harmonic homology γ with center M and axial plane ω is the reflection in M and leaves π invariant, which means $\widehat{\gamma} \circ \pi \circ \gamma = \pi$, hence $\widehat{\gamma} \circ \pi = \pi \circ \gamma$. In this case, the product $\pi' := \pi \circ \gamma$ is also self-adjoint, since

$$\widehat{\pi}' \circ \pi' = (\widehat{\pi} \circ \widehat{\gamma}) \circ (\pi \circ \gamma) = \widehat{\pi} \circ \pi = \mathrm{id}_{\mathcal{P}}.$$

The point M is also the pole of ω w.r.t. this composed polarity π'. If, on any line l through M, the points P and P' are conjugate w.r.t. π and symmetric w.r.t. M, then $\pi' = \pi \circ \gamma$ maps P via P' to the plane $\pi(P')$ which passes through P. Hence, P is self-conjugate w.r.t. π'. This holds for each diameter line l. Therefore, the self-polar tetrahedra are of type (ii), and π' is the polarity w.r.t. an ellipsoid with center M. Thus, we proved

Theorem 4.2.11 *Each elliptic polarity in the projective extension of* $\mathbb{E}^3$ *is the commutative product of the polarity w.r.t. an ellipsoid* $\mathcal{Q}$ *and the reflection in the center of the ellipsoid. This composition is called the anti-polarity w.r.t.* $\mathcal{Q}$*.*

If the ellipsoid $\mathcal{Q}$ is given in standard form, then the respective symmetric matrices $\mathbf{M}_{\mathcal{Q}}$ of $\mathcal{Q}$ and $\mathbf{M}$ of the elliptic polarity are

$$\mathbf{M}_{\mathcal{Q}} = \operatorname{diag}\left(-1, \frac{1}{a^2}, \frac{1}{b^2}, \frac{1}{c^2}\right) \quad \text{and} \quad \mathbf{M} = \operatorname{diag}\left(1, \frac{1}{a^2}, \frac{1}{b^2}, \frac{1}{c^2}\right)$$

In the literature, the ellipsoid $\mathcal{Q}$ is sometimes called the *real representative* of the imaginary quadric, associated with the elliptic polarity.

Since the field of complex numbers is algebraically closed, in the complex extension of $\mathbb{P}^3(\mathbb{R})$, every polarity π with a real symmetric transformation matrix $\mathbf{M}$ defines a quadric as the set of self-conjugate points. With each complex point P, the complex conjugate $\overline{P}$ is also self-conjugate. In the case of an elliptic polarity π, we call this quadric *empty* or *imaginary*.

Moreover, in the complex extension, every quadric is ruled, and it contains pairs of complex conjugate generators. When the polarity π is elliptic, the generators must be complex of the second kind, since otherwise pairs of complex conjugate generators would intersect at a real self-conjugate point. This means that complex conjugate generators of an imaginary quadric belong to the same regulus. On the other hand, the generators of oval quadrics are imaginary of the first kind, and the pairs of complex conjugate generators belong to different reguli. In particular, the generators of a sphere are isotropic. The intersection of a sphere with any tangent plane is a null circle consisting of two complex conjugate isotropic lines of the first kind.

If a polarity is elliptic and the induced polarity in ω equals the absolute polarity, we call the corresponding imaginary quadric of self-conjugate points an *imaginary sphere*. Its generators are isotropic of the second kind.

- **Exercise 4.2.1** Equations of a ruled quadric.

Find a coordinate transformation which transforms the equation $x_0x_3 - x_1x_2 = 0$ of a ruled quadric, as given in Theorem 4.2.5, to the standard form $x_0^2 + x_1^2 - x_2^2 - x_3^3 = 0$ given in Theorem 4.2.9.

- **Exercise 4.2.2** Polarity w.r.t. a ruled quadric.

Prove the following statements:

(i) *Let an arbitrary line l intersect a generator e of the ruled quadric $\mathcal{Q}$. Then, also the polar line l^* of l w.r.t. $\mathcal{Q}$ intersects the generator e.*

(ii) *If the four sides of the (non-planar) quadrangle $A_1 \dots A_4$ are generators of $\mathcal{Q}$, then the two diagonals $[A_1, A_3]$ and $[A_2, A_4]$ are polar w.r.t. $\mathcal{Q}$.*

• Exercise 4.2.3 Two pairs of polar lines.

Prove the following statement concerning the complex extension of $\mathbb{P}^3(\mathbb{R})$: *Let* (g, g^*) *and* (h, h^*) *be two pairs of polar lines. Then, the four lines* g, g^*, h, h^* *belong to a regulus if, and only, if they intersect two different self-polar lines* e_1 *and* e_2.

• Exercise 4.2.4 Reducible intersection of two quadrics.

Prove the following statement: *If two quadrics* $\mathcal{Q}_1, \mathcal{Q}_2$ *in* $\mathbb{P}^3(\mathbb{R})$ *contact each other at two points* P, Q, *and the common chord* $l = [P, Q]$ *and the intersection* $l^* = \tau_P \cap \tau_Q$ *of the corresponding tangent planes are skew, then all common points of the quadrics are located in two planes through* l, *which can also coincide or be complex conjugate. This is true also when one or both quadrics are quadratic cones.*

Hint: Choose a coordinate frame with the first two base points at P and Q and the two remaining base points on l^*. Common points of $\mathcal{Q}_1$ and $\mathcal{Q}_2$ satisfy all linear combinations of the quadrics' equations. There is a linear combination of the representing symmetric matrices with zeros in the first two lines and rows (see also Theorem 6.9.2).

• Exercise 4.2.5 Proof of Theorem 2.4.4.

Prove the statements of Theorem 2.4.4 concerning cones (or cylinders) passing through two different conics c and d of a regular quadric $\mathcal{Q}$.

Hint: Let l denote the line of intersection between the planes of c and d, and select any point $C \in l$ in the exterior of $\mathcal{Q}$ as the center of a projection into the polar plane γ of C. Then in the image plane γ, there are one or two centers V of planar harmonic homologies which map the contour of $\mathcal{Q}$ onto itself and exchange the contour points of c with those of d. Each homology can be continued to a harmonic homology in the 3-space which preserves $\mathcal{Q}$ and sends c to d. This confirms that its center V is the apex of a quadratic cone through c and d.

The last statement in Theorem 2.4.4 follows by dualization.

• Exercise 4.2.6 Projective generation of quadrics in projective 3-space.

Given two different points P, Q in $\mathbb{P}^3(\mathbb{F})$, let $\kappa\colon \mathcal{L}_P \to \mathcal{E}_Q$ be a projective correlation between the bundles with respective carriers P and Q. Then, the set of points of intersection $X = x \cap \kappa(x)$ is a regular or singular quadric. Conversely, each quadric can be generated in this way.

Hint: Choose a coordinate frame with $P = (1:0:0:0)$ and $Q = (0:0:0:1)$. Note that for describing a quadric in the form $\mathbf{x}^{\mathrm{T}} \mathbf{A}\, \mathbf{x} = 0$, the matrix A need not be symmetric.

21.04.2020

B) Books_70_2020/
Lightwave_Communications/
Lightwave Communications.pdf

C) Books_70_2020/
The Physics of Energy/
the Physics of Energy.pdf

Pag.52(72)

3.6 Maxwell's Equations

Pag.263(283)

The Forces of Nature

Pag.266(286)

Figure 14.2 Feynman diagram describing two electrons interacting exchange of a photon, the force carrier for electromagnetic interactions.

D) Books_70_2020/
The Universe of Quadrics/
3662610523.pdf

Pag.147(155)

Quadrics in the real projective space

21 04 2020

Figure 4.5 a) Self-polar [illegible] triangle
b) Self-conjugate.

Proof: Ceva and Menelaos the product of the cross-ratios on the three sides must be negative, which results in the two cases.

Pag. 129(137)

Figure 4.3 Perspectivities in the projective space.

Obs: Livro muito bom!!!

D2) Books_70_2020/
The Universe of Quadrics/
The Universe of Conics ...pdf

Obs. Livro muito bom!!!

Pag. 79(87)

Curvature of conics and the construction of osculating circles.

Pag. 83(91)

Figure 3.12

Right: For the particular point P, the osculating circle op meets the ellipse c at the antipode $\overline{P}$ of P.

4 Projective and affine quadrics

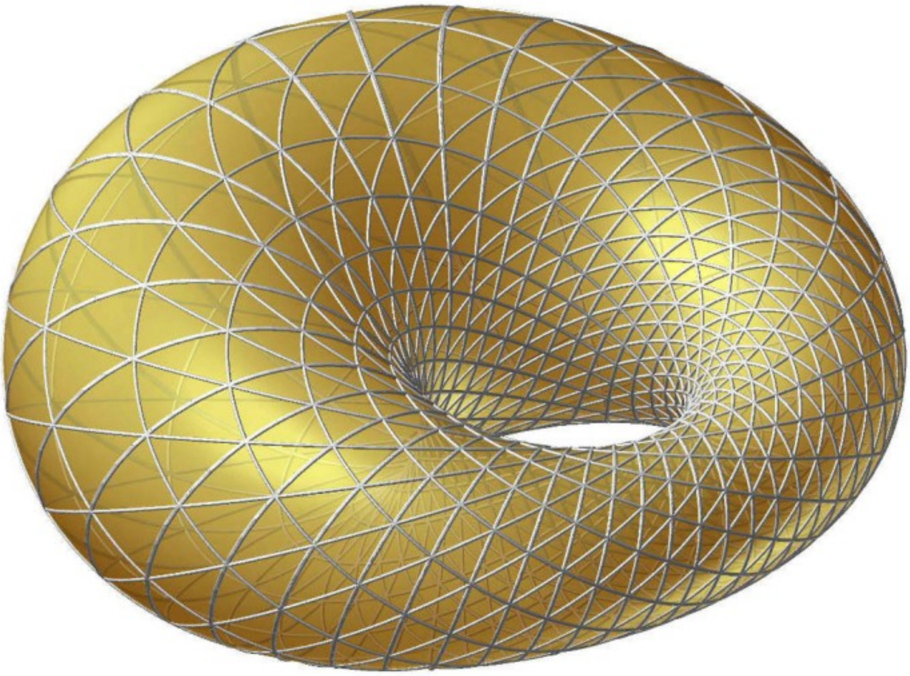

Dupin ring cyclide with a three-web consisting of Villarceau circles of both kinds and of isogonal trajectories of the circular curvature lines. In a conformal model of elliptic geometry, the cyclide represents a Clifford surface, where each kind of Villarceau circles is a family of Clifford parallels, while the isogonal trajectories play the role of coaxial helices.

© Springer-Verlag GmbH Germany, part of Springer Nature 2020

B. Odehnal et al., *The Universe of Quadrics*,
https://doi.org/10.1007/978-3-662-61053-4_4

The projective space is the proper setting for many properties of quadrics and associated polarities. It will be shown that, in the real projective 3-space, there are only two types of quadrics to be distinguished, ruled and non-ruled quadrics. With regard to polarities, there are three, apart from null polarities. The affine classification of quadrics can be done by analyzing their relative position to the plane at infinity.

In this chapter, we expect the reader to be already familiar with the basics of Projective Geometry in the plane as presented in [46, chapters 5–7].

4.1 Three-dimensional Projective Geometry

Models of projective spaces

To begin with, we give a short overview of the axiomatic approach to the projective 3-space $\mathbb{P}^3$, called projective space for short. It consists of a set $\mathcal{P}$ of *points*, a set $\mathcal{L}$ of *lines*, and a set $\mathcal{E}$ of *planes*. Lines and planes are always considered as sets of points. Sometimes we emphasize this by calling them a *range of points*. Similarly, planes are considered as sets of points and lines, and we speak sometimes of a *field of points* or *field of lines*.

The set of planes through some line g is called *pencil of planes* $\mathcal{E}_g$ with the *carrier* or *axis* g. The set of lines through some point V is called *bundle* or *star of lines* $\mathcal{L}_V$, and the set of planes through V is the *bundle* or *star of planes* $\mathcal{E}_V$ with the *carrier* or *vertex* V. For a *pencils of lines*, *i.e.*, the set of lines through some point V in some plane ε, we use the notation $\mathcal{L}_{V,\varepsilon}$. Again, the point V is the *vertex* or *carrier* and ε is the *carrier plane* of this pencil. We use the symbols $\mathcal{P}_\varepsilon$ and $\mathcal{L}_\varepsilon$ for the sets of points and lines in the plane ε, respectively.

If, for some point $P \in \mathcal{P}$ and some line $l \in \mathcal{L}$, the relation $P \in l$ holds, we say that P *is contained in* the line l, P *is incident with* l, P *lies on* l, l *goes through* P, or l *passes through* P. A similar wording is used for the *incidence* $P \in \varepsilon$ of the point P with the plane ε and for the incidence $l \subset \varepsilon$ of the line l with ε. Any set of points on the same line is called *collinear* or *aligned*. Lines or planes through the same point are called *concurrent*, points and lines in the same plane are *coplanar*, and planes through the same line are *coaxial*.

The following list of *axioms* makes the triple $(\mathcal{P}, \mathcal{L}, \mathcal{E})$ a *three-dimensional projective space* or *projective 3-space* $\mathbb{P}^3$.

(A1) Any two different points $P \neq Q$ can be joined by a unique line $l = [P, Q]$, the *connecting* or *spanned* line.

(A2) Any point P and any line l with $P \notin l$ can be joined with a unique plane $[P, l]$, the *connecting* or *spanned* plane.

(A3) For any line l and plane ε with $l \not\subset \varepsilon$, there exists a unique common point $l \cap \varepsilon$, the *point of intersection* or *piercing point.*

(A4) There exist five points such that no four of them are coplanar. We call this quintuple a *fundamental figure* of $\mathbb{P}^3$.

Two projective spaces $\mathbb{P}^3 = (\mathcal{P}, \mathcal{L}, \mathcal{E})$ and $\mathbb{P}^{3\prime} = (\mathcal{P}', \mathcal{L}', \mathcal{E}')$ are called *isomorphic* if there exist bijective mappings $\kappa : \mathcal{P} \to \mathcal{P}'$, $\kappa^* : \mathcal{L} \to \mathcal{L}'$, and $\widehat{\kappa} : \mathcal{E} \to \mathcal{E}'$ such that each incidence $P \in l$, $P \in \varepsilon$, or $l \subset \varepsilon$ in $\mathbb{P}^3$ holds if, and only if, the respective incidence $\kappa(P) \in \kappa^*(l)$, $\kappa(P) \in \widehat{\kappa}(\varepsilon)$, or $\kappa^*(l) \subset \widehat{\kappa}(\varepsilon)$, holds in $\mathbb{P}^{3\prime}$. We also say that in this case the triples $(\mathcal{P}, \mathcal{L}, \mathcal{E})$ and $(\mathcal{P}', \mathcal{L}', \mathcal{E}')$ are *models* of the same abstract projective 3-space.

Below we present two three-dimensional projective spaces. As will be shown, the first is isomorphic to a particular case of the second.

(i) **The projectively closed Euclidean 3-space.** Similarly to the projective closure of the Euclidean plane in [46, p. 180], we begin with the Euclidean 3-space $\mathbb{E}^3$. We assign to each class of parallel lines in $\mathbb{E}^3$ a unique *ideal point* or *point at infinity* and define that each line out of this class passes through this ideal point. The set of all ideal points of the 3-space is called *ideal plane* or *plane at infinity.* If we concentrate only on lines parallel to a given plane ε, the set of associated ideal points is called *ideal line* or *line at infinity* of the plane ε. Obviously, parallel planes share their line at infinity.

Now, we define $\mathcal{P}$ as the union of the set of points of $\mathbb{E}^3$ and the set of all ideal points. $\mathcal{L}$ is the union of the set of lines of $\mathbb{E}^3$ with the set of ideal lines. Finally, we obtain $\mathcal{E}$ as the set of planes of $\mathbb{E}^3$, which is extended by the plane at infinity. It is easy to verify that the triple $(\mathcal{P}, \mathcal{L}, \mathcal{E})$ satisfies all axioms (A1) to (A4). The extended space is called the *projectively closed Euclidean 3-space* or the *real projective space*, and we denote it by $\mathbb{P}^3(\mathbb{R})$. Points, lines, and planes belonging to $\mathbb{E}^3$ are called *finite* or *proper* in the projective closure $\mathbb{P}^3(\mathbb{R})$.

Connecting the finite point P with an ideal point Q means drawing a parallel through P to some line of the class associated with Q. In order to obtain the connection of two different ideal points P and Q, we choose any finite point V. Then, $[P,Q]$ is the ideal line of the plane spanned by the lines $[V,P]$ and $[V,Q]$. The point of intersection of any finite line l with the plane at infinity is the ideal point of l. The line of intersection of any finite plane ε with the plane at infinity is the ideal line of ε.

(ii) **The vector space model of a projective space.** Let $\mathbb{F}$ be an arbitrary (commutative) field and $\mathbb{F}^4$ a four-dimensional vector space over $\mathbb{F}$. It can be shown that the set $\mathcal{P}$ of one-dimensional subspaces, the set $\mathcal{L}$ of two-dimensional subspaces, and the set $\mathcal{E}$ of three-dimensional subspaces of $\mathbb{F}^4$ defines a projective 3-space. We call it the *vector space model* of a projective space and denote it by $\mathbb{P}^3(\mathbb{F})$.

If point P is spanned by the non-zero vector $\mathbf{p} \in \mathbb{F}^4$, then, instead of $P = \{\lambda\mathbf{p} \,|\, \lambda \in \mathbb{F}\}$, we write $P = \mathbf{p}\mathbb{F}$ or sometimes only $P = \mathbf{p}$ for short. The line l spanned by the points $\mathbf{p}\mathbb{F}$ and $\mathbf{q}\mathbb{F}$ is written briefly as $l = \mathbf{p}\mathbb{F} + \mathbf{q}\mathbb{F}$, which means $l = \{\lambda_1\mathbf{p} + \lambda_2\mathbf{q} \,|\, (\lambda_1, \lambda_2) \in \mathbb{F}^2\}$, provided that $\mathbf{p}$ and $\mathbf{q}$ are linearly independent, which means that the two points are different. Similarly, for three linearly independent vectors $\{\mathbf{p}, \mathbf{q}, \mathbf{r}\}$, the plane spanned by the three corresponding points is denoted $\varepsilon = \mathbf{p}\mathbb{F} + \mathbf{q}\mathbb{F} + \mathbf{r}\mathbb{F}$.

Any basis $(\mathbf{b}_0, \dots, \mathbf{b}_3)$ of $\mathbb{F}^4$ gives rise to coordinates for each vector in $\mathbb{F}^4$ and also *homogeneous coordinates* for each point $P = \mathbf{p}\mathbb{F}$ in $\mathbb{P}^3(\mathbb{F})$, according to the rule

$$P = (x_0 : x_1 : x_2 : x_3) \iff \mathbf{p} = x_0\mathbf{b}_0 + \dots + x_3\mathbf{b}_3.$$

Apparently, the coordinates of points are only unique up to a common non-vanishing scalar.

Now, we show that the choice of a basis yields also homogeneous coordinates for each plane: The set of solutions of any linear equation

$$u_0x_0 + u_1x_1 + u_2x_2 + u_3x_3 = 0 \quad \text{with} \quad (u_0, \dots, u_3) \neq (0, \dots, 0)$$

is three-dimensional, and therefore, a plane of $\mathbb{P}^3(\mathbb{F})$. Conversely, a point $\mathbf{x}\mathbb{F}$ belongs to the plane $\varepsilon = \mathbf{p}\mathbb{F} + \mathbf{q}\mathbb{F} + \mathbf{r}\mathbb{F}$ if, and only if, the vectors $\{\mathbf{p}, \mathbf{q}, \mathbf{r}, \mathbf{x}\}$ are linearly dependent. In this case, the system of linear equations

$$\begin{aligned} u_0p_0 + \ \dots \ + u_3p_3 &= 0\\ u_0q_0 + \ \dots \ + u_3q_3 &= 0\\ u_0r_0 + \ \dots \ + u_3r_3 &= 0\\ u_0x_0 + \ \dots \ + u_3x_3 &= 0 \end{aligned}$$

with the unknowns $(u_0, \dots, u_3)$ has a one-dimensional solution, and $\mathbf{x}$ must satisfy the last linear equation. We call the coefficients $(u_0 : \dots : u_3)$ *homogeneous plane coordinates* of ε and write also $\varepsilon = \mathbf{u}\mathbb{F}$, if $\mathbf{u} := (u_0, \dots, u_3)$.[1] Using the notation $\langle \cdot , \cdot \rangle$ for the standard scalar product, we can state:

Lemma 4.1.1 *Point* $\mathbf{x}\mathbb{F}$ *lies in the plane* $\mathbf{u}\mathbb{F}$ *if, and only if,* $\langle \mathbf{u}, \mathbf{x} \rangle = 0$.

A change of the basis in $\mathbb{F}^4$ yields new coordinates for each point. Due to a standard result of Linear Algebra, there is a regular 4×4 *transformation matrix* $\mathbf{T}$ such that for each point the *old* coordinates $(x_0 : \dots : x_3)$ and the *new* ones $(x'_0 : \dots : x'_3)$ are related in matrix form by

$$\begin{pmatrix} x'_0 \\ \vdots \\ x'_3 \end{pmatrix} = \mathbf{T} \begin{pmatrix} x_0 \\ \vdots \\ x_3 \end{pmatrix}.$$

Because of the homogeneity of the point coordinates, the transformation matrix $\mathbf{T}$ is unique only up to a non-vanishing scalar. Note that we still retain the convention introduced in Chapter 3 of writing coordinate vectors as columns.

Instead of choosing a basis of $\mathbb{F}^4$ for coordinatizing the projective space $\mathbb{P}^3(\mathbb{F})$, we can proceed in a more geometric way by specifying any fundamental figure $(A_0, \dots, A_3, E)$ in $\mathbb{P}^3(\mathbb{F})$:

If $A_i = \mathbf{a}_i\mathbb{F}$ for $i = 0, \dots, 3$, then $\{\mathbf{a}_0, \dots, \mathbf{a}_3\}$ is a basis of $\mathbb{F}^4$. This yields for the fifth point $E = \mathbf{e}\mathbb{F}$ a representation

$$\mathbf{e} = \lambda_0\mathbf{a}_0 + \dots + \lambda_3\mathbf{a}_3 \quad \text{with} \quad \lambda_0 \dots \lambda_3 \neq 0.$$

Now, we use $\mathbf{a}'_i = \lambda_i\mathbf{a}_i$ as the new coordinate vector of A_i for each $i \in \{0, \dots, 3\}$. The new basis $\{\mathbf{a}'_0, \dots, \mathbf{a}'_3\}$ brings about a coordinatization of $\mathbb{P}^3(\mathbb{F})$ with the following properties.

---

[1] In the language of Linear Algebra, planes are one-dimensional spaces in the dual space of $\mathbb{F}^4$, which can also be introduced in a coordinate-free way as the space of linear forms over $\mathbb{F}^4$.

Theorem 4.1.1 *For each fundamental figure* $(A_0,\dots,A_3,E)$ *in* $\mathbb{P}^3(\mathbb{F})$, *there is a coordinate frame such that* $A_0 = (1:0:0:0)$, $\dots$, $A_3 = (0:0:0:1)$ *and* $E = (1:1:1:1)$. *The points* $A_0,\dots,A_3$ *are called base points of this frame, and point* E *is the unit point.*

Finally, we show that, in the projectively closed Euclidean 3-space, we can introduce homogeneous coordinates as well. This will confirm that this projective space is isomorphic to the projective 3-space over the field of real numbers, which justifies the introduced notation $\mathbb{P}^3(\mathbb{R})$.

It is quite natural to start with a Cartesian coordinate frame in the Euclidean 3-space $\mathbb{E}^3$, though the same procedure could be performed with affine coordinates: If (x_P, y_P, z_P) are the coordinates of any (finite) point P, we extend them by adding 1 as the first coordinate, and we permit that these four coordinates can be multiplied with any factor $\lambda \in \mathbb{R}\setminus\{0\}$.[2] If Q is the point at infinity of a line with direction vector (x_Q, y_Q, z_Q), then we insert 0 as additional coordinate at the beginning. Of course, the direction vector is unique only up to a non-vanishing scalar. Thus, we assigned to each point in the projective extension of $\mathbb{E}^3$ four homogeneous coordinates.

Conversely, from any given coordinates $(x_0:\dots:x_3)$ with $(x_0,\dots,x_3) \neq (0,\dots,0)$, we can retrieve the original point: In the case $x_0 \neq 0$, we get

$$(x,y,z) = \left(\frac{x_1}{x_0}\,,\ \frac{x_2}{x_0}\,,\ \frac{x_3}{x_0}\right).$$

In the case $x_0 = 0$, the vector $(x_1,x_2,x_3) \neq (0,0,0)$ gives the direction of the point at infinity.

We call these particular homogeneous coordinates *homogeneous Cartesian coordinates*. In the language of $\mathbb{P}^3(\mathbb{R})$, the base points of this coordinate frame are the origin and the ideal points of the coordinate axes. The unit point has the Cartesian coordinates $(x,y,z) = (1,1,1)$.

First consequences of the axioms

From the four axioms of projective 3-spaces (A1)–(A4) we can deduce a sequence of statements, as listed below.

---

[2] The same extension of coordinates was already used in Chapter 3 when extended coefficient matrices of linear mappings and quadrics were introduced.

- If $P, Q \in \varepsilon$, then $[P, Q] \subset \varepsilon$, and vice versa.

- For any three non-collinear points P, Q, R, *i.e.*, for each *triangle*, there exists a unique plane ε with $P, Q, R \in \varepsilon$, which is denoted by $[P, Q, R]$ and called the *connecting* or *spanned* plane.

- Each plane contains at least a *quadrangle*, *i.e.*, four coplanar points such that no three of them are collinear.

- Any two different planes ε, ψ have exactly one line in common, the *line of intersection* $l = \varepsilon \cap \psi$.

- If two different lines g, h share a point, then they are coplanar, hence spanning a plane $[g, h]$. Conversely, any two different but coplanar lines have exactly one point in common, the *point of intersection*.

Thus, we can state:

- For each plane ε in $\mathbb{P}^3$, the incident points and lines, *i.e.*, the sets $\mathcal{P}_\varepsilon$ and $\mathcal{L}_\varepsilon$, form a projective plane (see, *e.g.*, [46, Sec. 5.1]).

Each set of four non-coplanar points $\{O, P, Q, R\}$ is called a *tetrahedron* with the six *edges* $[O, P], \dots, [Q, R]$. Let ε denote the plane $[P, Q, R]$. Then, the mapping

$$\pi\colon\ \mathbb{P}^3 \setminus \{O\} \to \mathcal{P}_\varepsilon,\ X \mapsto \pi(X) = [X, O] \cap \varepsilon,$$

is called a *projection* with *center* O and *image plane* ε.

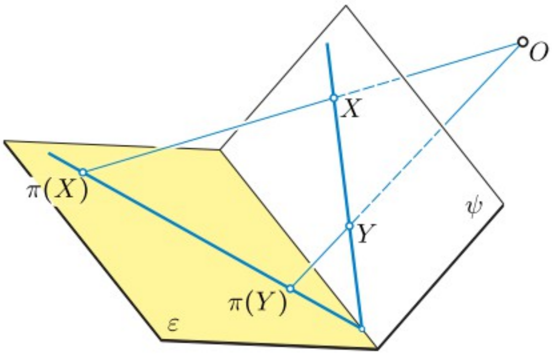

FIGURE 4.1. The projection $\pi\colon X \mapsto \pi(X)$ with center O and image plane ε.

The restriction of the projection π to any plane ψ not passing through O is bijective and preserves the collinearity of points (Figure 4.1). Hence, this is a collineation $\mathcal{P}_\psi \to \mathcal{P}_\varepsilon$. As a consequence, for any two planes ε, ψ

in $\mathbb{P}^3$, the induced projective planes are isomorphic in the sense of [46, p. 179].

- Conversely, if for two different planes ε, ψ in $\mathbb{P}^3$ there is a collineation $\lambda\colon\ \mathcal{P}_\varepsilon \to \mathcal{P}_\psi$, where each point of the line of intersection $\varepsilon \cap \psi$ remains fixed, then λ is the restriction of a projection.
- There exist pairs of *skew* lines in $\mathbb{P}^3$, *i.e.*, lines g, h having no point in common and being non-coplanar.

In each projective space $\mathbb{P}^3$, there exist five planes such that no four share a point. For example, given a fundamental figure $\{P_0, \dots, P_4\}$, the five planes $[P_0, P_1, P_2]$, $[P_1, P_2, P_3]$, $\dots$, $[P_4, P_0, P_1]$ have this property. Hence, similar to projective planes [46, p. 182], for each three-dimensional projective space $\mathbb{P}^3$, we can state the

Principle of Duality. *Any valid theorem formulated in terms of points, lines, planes, connection, intersection, and incidence remains true if we interchange the words points and planes, intersection and connection, but keep lines as lines and all incidences between points, lines, or planes.*

In other words, if $\mathbb{P}^3 = (\mathcal{P}, \mathcal{L}, \mathcal{E})$ is a projective space, then $\widehat{\mathbb{P}}^3 = (\mathcal{E}, \mathcal{L}, \mathcal{P})$ is also a projective space provided that we switch $\in$ and $\ni$ as well as $\subset$ and $\supset$. The space $\widehat{\mathbb{P}}^3$ is called the *dual space* of $\mathbb{P}^3$. When any statement is dualized, the obtained new result is called the *dual*.

The proof of the principle of duality consists in verifying that, from the axioms (A1) to (A4), all respective dual statements can be concluded, and this has already been done above.

In this spatial duality, the fields of points correspond to stars of planes, and the fields of lines (or ruled planes) to stars of lines. Therefore, for all points V, the pair $(\mathcal{E}_V, \mathcal{L}_V)$ is a projective plane, as well as $(\mathcal{L}_V, \mathcal{E}_V)$.

We recall that a projective plane is called a *Fano plane* if the axiom of FANO holds, *i.e.*, for each quadrangle the three diagonal points are not collinear. Similarly, we call a projective 3-space a *Fano space* if it contains a Fano plane.

Since any two planes are isomorphic, all quadrangles in a Fano space have a diagonal triangle. The principle of duality in the plane guarantees that, for the dual of a quadrangle, a *quadrilateral*, the three diagonals are never concurrent. It can be shown that $\mathbb{P}^3(\mathbb{F})$ is a Fano plane if, and only if, the characteristic of the field $\mathbb{F}$ is different from two (note [46, p. 184]. Otherwise, we speak of an *Anti-Fano plane*.

In this book, we will mainly focus on Fano planes without specifying so. Only in very exceptional cases, we will assume that this axiom does not hold.

For the axiomatic approach to projective planes, two other axioms are important, the axiom of DESARGUES and the axiom of PAPPUS (note [18], [46, p. 185–187], or [57]).

The first one states that if two triangles are perspective to a point, then their sides are perspective to a line. In other words, if, for two given triangles $P_1P_2P_3$ and $Q_1Q_2Q_3$, the points P_i, Q_i are collinear with a point (*perspector*) Z for $i = 1, 2, 3$, then the pairs of sides $([P_1, P_2], [Q_1, Q_2])$, $([P_2, P_3], [Q_2, Q_3])$, and $([P_3, P_1], [Q_3, Q_1])$ are concurrent with a line (*perspectrix*) a.

The axiom of Desargues guarantees the existence of perspective collineations in the plane. Furthermore, it can be proved that any Desarguesian plane is isomorphic to a projective plane $\mathbb{P}^2(\mathbb{F})$ over a field $\mathbb{F}$ which needs not be commutative (see [18]).

The situation in projective spaces is quite different:

- In each plane of a projective 3-space, the axiom of Desargues holds.

Proof: Suppose that the two triangles $\overline{P_1}\,\overline{P_2}\,\overline{P_3}$ and $Q_1Q_2Q_3$ in different planes are perspective w.r.t. $\overline{Z}$ (Figure 4.2). Then, the points $\{\overline{Z}, \overline{P_1}, \overline{P_2}, Q_1, Q_2\}$ are coplanar. Therefore, the sides $[\overline{P_1}, \overline{P_2}]$ and $[Q_1, Q_2]$ share a point A_3 on the line of intersection $a = [\overline{P_1}, \overline{P_2}, \overline{P_3}] \cap [Q_1, Q_2, Q_3]$. A cyclic permutation of the subscripts reveals that the sides of the two triangles are perspective w.r.t. a.

If two triangles $P_1P_2P_3$ and $Q_1Q_2Q_3$ belong to the same plane π and are perspective w.r.t. a point Z, we can confine ourselves to the case that the seven points $P_1, \dots, Q_1, \dots, Z$ are mutally different. Now, we interpret the scene as a projection of the spatial situation. This means that we choose a point $O \notin \pi$ and, on the line $[O, Z]$, another point $\overline{Z} \neq O, Z$. Then, there exist the points $\overline{P_i} = [O, P_i] \cap [\overline{Z}, Q_i]$ for $i = 1, 2, 3$, which do not belong to π. This results in a triangle $\overline{P_1}\,\overline{P_2}\,\overline{P_3}$ which is perspective to $Q_1Q_2Q_3$ w.r.t. $\overline{Z}$. Due to the spatial version of Desargues's theorem, the sides of these two triangles are perspective w.r.t. $a = [\overline{P_1}, \overline{P_2}, \overline{P_3}] \cap \pi$. The point of intersection $[P_1, P_2] \cap [\overline{P_1}, \overline{P_2}]$ must coincide with $A_3 = [P_1, P_2] \cap [Q_1, Q_2]$, which, after cyclic permutation of the subscripts, proves the original planar version. ■

Finally, we call a projective space a *Pappian space*, if it contains a Pappian plane. Again, the isomorphy between any two planes guarantees that each plane is Pappian.

A standard result from the theory of projective planes states that each Pappian plane is isomorphic to a projective plane $\mathbb{P}^2(\mathbb{F})$ over a commu-

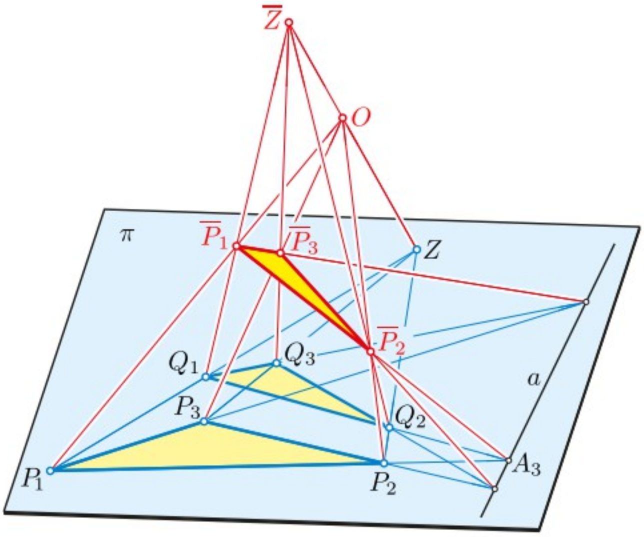

FIGURE 4.2. Proof of Desargues's theorem for the perspective triangles $P_1P_2P_3$ and $Q_1Q_2Q_3$ via the spatial version with the triangle $\overline{P}_1\overline{P}_2\overline{P}_3$.

tative field $\mathbb{F}$ (see [18]). Hence, the projective space $\mathbb{P}^3(\mathbb{F})$, as introduced above, is Pappian.

Collineations, perspectivities, and projectivities

Definition 4.1.1 A *collineation* $\kappa : \mathbb{P}^3 \to \mathbb{P}^{3\prime}$ between two projective spaces is a mapping that sends points to points and satisfies the following two conditions:

(1) Collinear points are mapped to collinear points.

(2) The mapping is one-to-one and onto.

Apparently, each isomorphism of $\mathbb{P}^3 = (\mathcal{P}, \mathcal{L}, \mathcal{E})$ to $\mathbb{P}^{3\prime} = (\mathcal{P}', \mathcal{L}', \mathcal{E}')$ defines a collineation $\kappa\colon \mathcal{P} \to \mathcal{P}'$. However, the converse is also true:

- Each collineation $\mathbb{P}^3 \to \mathbb{P}^{3\prime}$, defined as a mapping $\kappa\colon \mathcal{P} \to \mathcal{P}'$, induces an isomorphism between the two spaces, and hence the bijections $\kappa^*\colon \mathcal{L} \to \mathcal{L}'$ and $\widehat{\kappa} : \mathcal{E} \to \mathcal{E}'$. Each incidence in $\mathbb{P}^3$ is equivalent to the incidence between the corresponding elements in $\mathbb{P}^{3\prime}$.

For a proof, the reader is referred to the literature, *e.g.*, to [18].

This result reveals that the restriction of any collineation $\kappa\colon \mathbb{P}^3 \to \mathbb{P}^{3\prime}$ to each plane ε of $\mathbb{P}^3$ is a (planar) collineation $\mathcal{P}_\varepsilon \to \mathcal{P}_{\widehat{\kappa}(\varepsilon)}$. The duality guarantees that the same is true for corresponding stars $\mathcal{E}_S$ and $\mathcal{E}_{\kappa(S)}$.

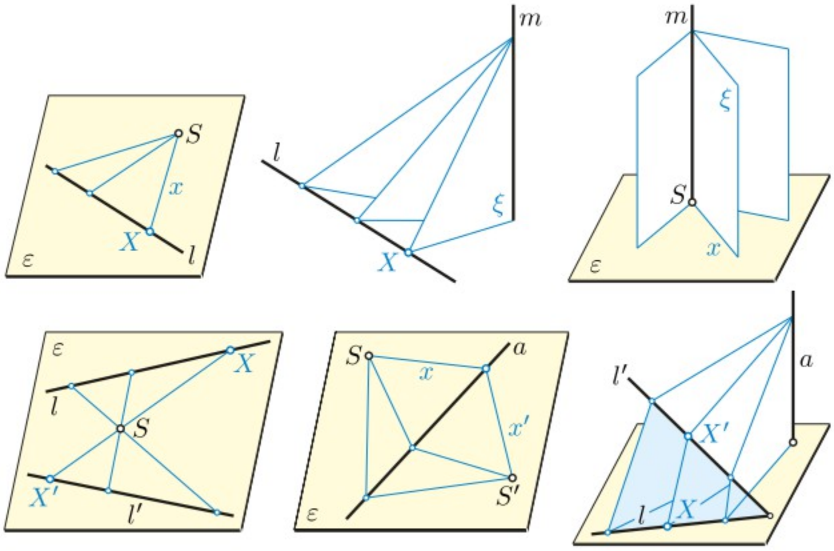

FIGURE 4.3. Perspectivities in the projective space: First row, from left to right: $l(X) \doublebarwedge S_\varepsilon(x)$, $l(X) \doublebarwedge m(\xi)$, and $S_\varepsilon(x) \doublebarwedge m(\xi)$; second row, from left to right: $l(X) \overset{S}{\doublebarwedge} l'(X')$, $S_\varepsilon(x) \overset{a}{\doublebarwedge} S'_\varepsilon(x')$, and $l(X) \overset{a}{\doublebarwedge} l'(X')$.

Now, we are going to extend perspectivities and projectivities from the projective plane to the projective 3-space.

Definition 4.1.2 Let $\mathcal{P}_l$, $\mathcal{L}_{S,\varepsilon}$ and $\mathcal{E}_m$ be, respectively, a range of points, a pencil of lines, and a pencil of planes. Any bijection between any two of these sets is called a *perspectivity* if corresponding elements are incident (Figure 4.3). We denote them for short $l \doublebarwedge S_\varepsilon$, $l \doublebarwedge m$, and $S_\varepsilon \doublebarwedge m$, respectively, or more extensively, by indicating corresponding elements, in the form

$$l(X) \doublebarwedge S_\varepsilon(x), \quad l(X) \doublebarwedge m(\xi), \quad S_\varepsilon(x) \doublebarwedge m(\xi),$$

or in opposite order.

Any composition of finitely many perspectivities is called a *projectivity* and usually expressed with the symbol $\barwedge$.

If, in particular, the lines l and l' are coplanar with S, we call the composition $l(X) \barwedge S_\varepsilon(x) \barwedge l'(X')$ a perspectivity $l \overset{S}{\barwedge} l'$ with *center* S. Similarly, we speak of a perspectivity $S_\varepsilon \overset{a}{\barwedge} S'_\varepsilon$ with the *axis* a for two coplanar pencils or pencils with a common carrier, when either $S_\varepsilon(x) \barwedge a(X) \barwedge S'_\varepsilon(x')$ or $S_\varepsilon(x) \barwedge a(\xi) \barwedge S_{\varepsilon'}(x')$, respectively. Finally, we speak of a perspectivity between two pencils of planes $l \overset{a}{\barwedge} l'$ with the axis a when $l(\xi) \barwedge a \barwedge l'(\xi')$ and the carriers l and l' are coplanar.

The perspectivities, as defined above in the projective 3-space, are extensions of those related to the projective plane. Therefore, the set of projectivities between ranges of points or pencils of lines seems to be larger than that of the planar counterparts. However, a detailed case analysis reveals that each 3-space projectivity between coplanar ranges of points or pencils of lines is also a projectivity in the sense of planar Projective Geometry. This justifies keeping the same notations. Furthermore, the *fundamental theorem* of plane geometry is still valid. This means (note [46, p. 190]):

Lemma 4.1.2 *In a Pappian projective space, any projectivity between ranges of points, or pencils of lines or planes, is uniquely defined by three pairs of corresponding elements.*

The following lemma is also an extension of a result from plane geometry.

Lemma 4.1.3 *A projectivity between two different ranges of points with coplanar carriers is a perspectivity if, and only if, the common element remains fixed.*

The same holds for projectivities between different pencils of lines that are either in a common plane or share a common center. Finally, a similar statement is valid for projectivities between two different pencils of planes with intersecting carriers.

On the other hand, we can adopt, from the vector model $\mathbb{P}^3(\mathbb{F})$, the statement (note [46, Theorem 5.4.1]):

Lemma 4.1.4 *Each projectivity between ranges of lines in $\mathbb{P}^3(\mathbb{F})$ can be described by a bijective linear mapping between the corresponding two-dimensional subspaces of $\mathbb{F}^4$.*

In this sense, the group $\mathrm{PGL}(l)$ of projective mappings of any line l onto itself in the projective space is not bigger than that in the projective plane. It equals the factor group $\mathrm{GL}(\mathbb{F},2)/\mathrm{D}(\mathbb{F})$ of the general linear group, where $\mathrm{D}(\mathbb{F})$ is the group of dilations, represented by all scalar multiples of the unit matrix $\mathbf{I}_2$ (note [46, Theorem 5.4.2]).

Definition 4.1.3 We call a collineation *projective* if its restriction to a range of points or to a pencil of lines or planes is a projectivity.

It is easy to verify that the restriction of a projective collineation to all ranges of points or pencils of lines or planes is a projectivity. Hence, the restriction to any plane is projective, too, and the projective collineations $\mathbb{P}^3 \to \mathbb{P}^3$ form a group.

From the projective plane, we recall that under projective collineations all *cross ratios* remain unchanged ([46, Theorem 5.3.1]), which in the vector model can be defined in the following way:

If, on the line $l = [A,B] = \mathbf{a}\mathbb{F} + \mathbf{b}\mathbb{F}$, two more points are given as $C = (\lambda_1\mathbf{a} + \lambda_2\mathbf{b})\mathbb{F}$ and $D = (\mu_1\mathbf{a} + \mu_2\mathbf{b})\mathbb{F}$, then

$$\mathrm{cr}(A,B,C,D) := \frac{\lambda_2}{\lambda_1}\frac{\mu_1}{\mu_2}. \tag{4.1}$$

We call $(\mu_1 : \mu_2)$ *projective coordinates* of the point D w.r.t. the coordinate frame with the *base points* A and B and the *unit point* $(\mathbf{a}+\mathbf{b})\mathbb{F}$.

Similarly, a quadrangle $A_0A_1A_2E$ defines in its spanned plane a projective coordinate frame with the *base points* $A_i = \mathbf{a}_i\mathbb{F}$, $i = 0,1,2$, and the *unit points* $E = (\mathbf{a}_0 + \mathbf{a}_1 + \mathbf{a}_2)\mathbb{F}$, which assigns to any point $P = (x_0\mathbf{a}_0 + x_1\mathbf{a}_1 + x_2\mathbf{a}_2)\mathbb{F}$ the homogeneous triplet $(x_0 : x_1 : x_2)$ as coordinates.

A fundamental theorem of Projective Geometry over commutative fields states:

Theorem 4.1.2 *Each projective collineation $\mathbb{P}^3(\mathbb{F}) \to \mathbb{P}^3(\mathbb{F})$ of a Pappian projective space onto itself can be described by a bijective linear mapping $f\colon \mathbb{F}^4 \to \mathbb{F}^4$. This mapping is unique up to a scalar.*

Conversely, any linear mapping $f : \mathbb{F}^4 \to \mathbb{F}^4$ induces a projective collineation $\kappa\colon \mathbb{P}^3(\mathbb{F}) \to \mathbb{P}^3(\mathbb{F})$ with $\mathbf{p}\mathbb{F} \mapsto f(\mathbf{p})\mathbb{F}$.

We learn from the real projective plane that, in the real projective space $\mathbb{P}^3(\mathbb{R})$, each collineation is projective.

By the same token, all collineations in $\mathbb{P}^3(\mathbb{F}) \to \mathbb{P}^3(\mathbb{F})$ are induced by bijective semilinear mappings $\mathbb{F}^4 \to \mathbb{F}^4$, *i.e.*, by the compositions of linear mappings with automorphisms of $\mathbb{F}$ (see, *e.g.*, [18, 57]). A standard example of a non-projective collineation in the complex extension $\mathbb{P}^3(\mathbb{C})$ of the real projective space $\mathbb{P}^3(\mathbb{R})$ is the *complex conjugation* ι, which replaces each coordinate by its complex conjugate value. This conjugation is involutive and exchanges each non-real point P, line g, or plane ε with its respectively *complex conjugate* element $\overline{P} = \iota(P)$, $\overline{l} = \iota^*(l)$, or $\overline{\varepsilon} = \widehat{\iota}(\varepsilon)$. Elements which remain fixed under ι are called *real*, as they are already included in $\mathbb{P}^3(\mathbb{R})$. Otherwise, we call them *non-real* or *imaginary*. Therefore, $P = \mathbf{p}\mathbb{F}$ is imaginary if, and only if, the vectors $\{\mathbf{p}, \overline{\mathbf{p}}\}$ are linearly independent. In this case, the connection of P and $\overline{P}$ is real. For the same reason, two different complex conjugate planes ε and $\overline{\varepsilon}$ contain a single real line, the line of intersection $\varepsilon \cap \overline{\varepsilon}$.

The situation for lines is slightly different: If a line contains two different real points, then it is real. However, the converse is also true: If the line $l = \mathbf{p}\mathbb{F} + \mathbf{q}\mathbb{F}$ remains fixed under the conjugation, which means that $l = \overline{\mathbf{p}}\,\mathbb{F} + \overline{\mathbf{q}}\,\mathbb{F}$, where both pairs of vectors $(\mathbf{p}, \overline{\mathbf{p}})$ and $(\mathbf{q}, \overline{\mathbf{q}})$ are linearly independent, then the two points $R = (\mathbf{p} + \overline{\mathbf{p}})\mathbb{F}$ and $S = (\mathbf{p} - \overline{\mathbf{p}})\mathbb{F}$ on l are real. Moreover, the two points R and S must be different.

For imaginary lines l and $\overline{l}$, there are two possibilities:

- Either, they intersect each other. Then, they share a real point and span a real plane. We call them *imaginary of the first kind.*
- Or, the lines l and $\overline{l}$ are skew. Then, l neither includes a real point nor does it belong to a real plane. Line l is called *imaginary of the second kind.*

Perspective collineations

Suppose that a collineation κ in the projective space $\mathbb{P}^3$ keeps all points of a plane ζ fixed. Then, this collineation is projective by Definition 4.1.3. Moreover we can prove:

Lemma 4.1.5 *If under a collineation $\kappa\colon \mathbb{P}^3 \to \mathbb{P}^3$ all points of a plane ζ remain fixed, then there exists a point Z with the property that all lines and planes through Z remain fixed, too, and vice versa.*

Proof: If a point $Z \notin \zeta$ remains fixed, *i.e.*, $\kappa(Z) = Z$, then each line g through Z is a connection of Z with another fixed point $g \cap \zeta$, and therefore, a fixed line. Furthermore, each plane through Z is the connection of two fixed lines.

There remains the case that, for all points $P \notin \zeta$, we have $\kappa(P) \neq P$. Then, the line $p = [P, \kappa(P)]$ is fixed, since its image $\kappa^*(p)$ is the connection of $\kappa(P)$ with the fixed point $p \cap \zeta$. For any other point $Q \notin p, \zeta$ the fixed line $q = [Q, \kappa(Q)]$ is coplanar with p, since $g = [P, Q]$ and $\kappa^*(g) = [\kappa(P), \kappa(Q)]$ meet at the point $g \cap \zeta$. The point $Z := p \cap q$ remains fixed as the intersection of two fixed lines. Hence, $Z \in \zeta$.

Each line x through Z must be fixed, too, since, in the case $x \notin \zeta$ and $x \neq p$, it contains any point $Q \notin \zeta, p$, and, by the same argument as used above, x equals the fixed line $[Q, \kappa(Q)] = q$.

The converse statement holds due to duality. ■

Definition 4.1.4 A collineation κ which keeps each point of a plane ζ and each plane through a point Z fixed is called *perspective collineation* with the *center* Z and the *axis* or *axial plane* ζ.

In the case $Z \in \zeta$, the collineation κ is called *elation* and otherwise *homology*.

All perspective collineations which share the center Z and the axis ζ form a group. We denote it by $\mathrm{PK}(Z, \zeta)$. A standard result of Projective Geometry states that the commutativity of this group is equivalent to the validity of the Pappus theorem.

Suppose that a perspective collineation $\kappa \in \mathrm{PK}(Z, \zeta)$, *i.e.*, with center Z and axis ζ, has another fixed point $P \neq Z$, $P \notin \zeta$. Then, each point $Q \notin [P, Z]$ remains fixed as the intersection of two fixed lines $[Q, Z]$ and $[Q, P]$. However, each point R on the fixed line $[P, Z]$ is also the intersection with another fixed line $[R, Q]$, and therefore, fixed. Hence, κ is the identity mapping $\mathrm{id}_{\mathbb{P}^3}$.

Lemma 4.1.6 *Any perspective collineation with center Z and axis ζ is uniquely defined by a pair of corresponding points P, P' collinear with Z and $P, P' \notin \zeta$.*

Proof: If there were two perspective collineations $\kappa, \lambda \in \mathrm{PK}(Z, \zeta)$ with $P \to P'$, then the product $\lambda^{-1} \circ \kappa$ keeps P fixed. This implies $\lambda^{-1} \circ \kappa = \mathrm{id}_{\mathbb{P}^3}$ and, consequently, $\kappa = \lambda$. ■

Homologies in the plane have a *characteristic cross ratio* defined by the pairs of corresponding points, the center, and the collinear point on the axis. The analogue in 3-space is also valid: For all points $X \neq Z$, $X \notin \zeta$, the cross ratio

$$\delta = \operatorname{cr}\bigl(Z, X_\zeta, X, \kappa(X)\bigr), \quad \text{where} \quad X_\zeta = \zeta \cap [Z, X],$$

is the same, since the restriction to any plane through Z has a characteristic cross ratio, and any two planes through Z share a line through Z (note Figure 4.4, left). A homology is involutive if, and only if, its characteristic cross ratio is -1. In this case, the pairs $(X, \kappa(X))$ are harmonic w.r.t. Z and the axis ζ; we speak of a *harmonic homology*.

Suppose that in a Pappian projective space $\mathbb{P}^3$, a projective collineation κ has a fundamental figure of fixed points $P_0, \dots, P_4$. Then, the restriction to the plane $\varepsilon = [P_0, P_1, P_2]$ has a quadrangle of fixed points, since $\varepsilon \cap [P_3, P_4]$ also remains fixed. Thus, this restriction is the identity, and the collineation κ is perspective with two more fixed points outside the axis ε. This yields $\kappa = \mathrm{id}_{\mathbb{P}^3}$. An immediate consequence is formulated below.

Theorem 4.1.3 *In Pappian projective spaces, for any two fundamental figures $P_0, \dots, P_4$ and $P'_0, \dots, P'_4$, there exists exactly one projective collineation with $P_i \mapsto P'_i$ for $i = 0, \dots, 4$.*

In other words, projective collineations act sharply transitive on fundamental figures. Since Pappian projective spaces are isomorphic to projective spaces over commutative fields, this result can also be concluded from Theorem 4.1.2.

After introducing a coordinate frame in $\mathbb{P}^3(\mathbb{F})$, each projective collineation $\kappa\colon (x_0 : \dots : x_3) \mapsto (x'_0 : \dots : x'_3)$ can be written in matrix form as

$$\kappa: \begin{pmatrix} x'_0 \\ \vdots \\ x'_3 \end{pmatrix} = \mathbf{M} \begin{pmatrix} x_0 \\ \vdots \\ x_3 \end{pmatrix}, \tag{4.2}$$

where $\mathbf{M}$ is a regular 4×4 matrix. Suppose that the image point lies in the plane with coordinates $(u'_0 : \dots : u'_3)$, which can be written in matrix form as $\mathbf{u}'^{\mathrm{T}}\mathbf{x}' = 0$. The coordinates $(u_0 : \dots : u_3)$ of the preimage satisfy the linear equation $\bigl(\mathbf{u}'^{\mathrm{T}}\mathbf{M}\bigr)\,\mathbf{u} = 0$. Therefore, the associated transformation of planes is given by

$$\widehat{\kappa}: \begin{pmatrix} u'_0 \\ \vdots \\ u'_3 \end{pmatrix} = \bigl(\mathbf{M}^{-1}\bigr)^{\mathrm{T}} \begin{pmatrix} u_0 \\ \vdots \\ u_3 \end{pmatrix}. \tag{4.3}$$

A point $(x_0 : \ldots : x_3)$ remains fixed under κ if, and only if, there exists $\lambda \in \mathbb{R} \setminus \{0\}$ with

$$\lambda \begin{pmatrix} x_0 \\ \vdots \\ x_3 \end{pmatrix} = \mathbf{M} \begin{pmatrix} x_0 \\ \vdots \\ x_3 \end{pmatrix}, \quad \text{and hence} \quad (\mathbf{M} - \lambda \mathbf{I}_4) \begin{pmatrix} x_0 \\ \vdots \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

with $\mathbf{I}_4$ as unit matrix. This reveals that λ must be an eigenvalue of the matrix $\mathbf{M}$. Once an eigenvalue λ is found, it depends on the rank of the matrix $(\mathbf{M} - \lambda \mathbf{I}_4)$ whether we get a fixed point, a line, or a plane consisting of fixed points only. A k-fold eigenvalue effects a rank deficiency of at least 1 and at most k.

Note that, for each eigenvalue λ of $\mathbf{M}$, the reciprocal $1/\lambda$ is an eigenvalue of $\mathbf{M}^{-1}$, since $\det(\mathbf{M} - \lambda \mathbf{I}_4) = 0$ is equivalent to $\det[\mathbf{M}^{-1}(\mathbf{M} - \lambda \mathbf{I}_4)] = \det(\mathbf{I}_4 - \lambda \mathbf{M}^{-1}) = 0$, which defines fixed planes of $\widehat{\kappa}^{-1}$ and $\widehat{\kappa}$.

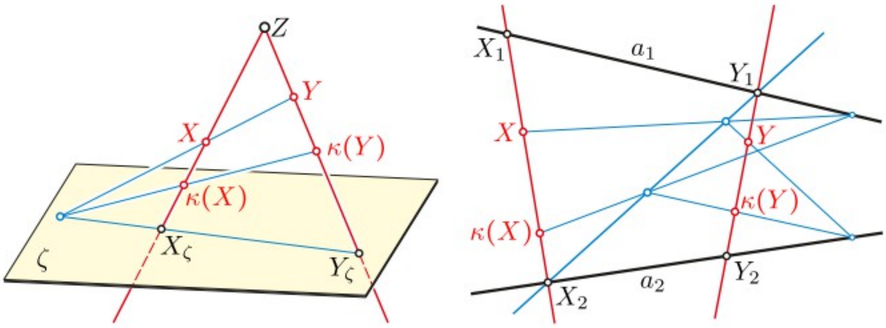

FIGURE 4.4. Left: characteristic cross ratio $\mathrm{cr}(Z, X_\zeta, X, \kappa(X))$ of a homology; Right: characteristic cross ratio $\mathrm{cr}(X_1, X_2, X, \kappa(X))$ of a biaxial collineation with axes a_1 and a_2.

In this way, we can classify projective collineations at hand of their configuration of fixed points. An appropriate tool, provided in Linear Algebra, is the Jordan normal form (see, *e.g.*, [53, 108]). For an algebraically closed field $\mathbb{F}$, there are already 13 different types to be distinguished, together with the identity $\mathrm{id}_{\mathbb{P}^3}$ (compare with (5.6) in Chapter 5). Otherwise, we obtain much more types. For example, in the real projective space, there are projective collineations without any fixed point, since the spectrum of $\mathbf{M}$ may consist of two pairs of complex conjugate eigenvalues.

Below, we concentrate on three types of projective collineations.

(i) The fixed points of a projective collineation κ can form a tetrahedron $A_0, \dots, A_3$. In this case, the restriction to each edge of the tetrahedron is a hyperbolic projectivity. If the fixed points are chosen as base points of a coordinate frame, we obtain a diagonal matrix $\mathbf{M}$ as analytic representation of κ.

(ii) If a projective collineation $\kappa \neq \mathrm{id}_{\mathbb{P}^3}$ has, besides a fixed tetrahedron, another fixed point F, then F must be located either in a face or on an edge. In the first case, κ is perspective. In the second case, κ has a line consisting of fixed points only, and the collineation is called *uniaxial*.

(iii) It is even possible that κ has two skew lines a_1 and a_2 with fixed points. We call such a collineation *biaxial* with the axes a_1 and a_2. Here all transversals, meeting both axes, remain fixed, and the restriction to each line is a hyperbolic projectivity. The restriction to any plane through one of the axes is a homology, and any plane through a_1 shares with any plane through a_2 a transversal. Therefore, the characteristic cross ratios in all these planes and on all transversals are the same. We call

$$\delta := \mathrm{cr}\left(X_1, X_2, X, \kappa(X)\right) \text{ with } X_i \in a_i,\ i = 1, 2, \tag{4.4}$$

the *characteristic cross ratio* of the biaxial collineation (Figure 4.4, right). The involutive case with $\delta = -1$ is sometimes called a *skew involution*.

• **Exercise 4.1.1** Projective collineations are products of perspective collineations.

Prove the following theorem: *In Pappian projective spaces, each projective collineation is the product of four perspective collineations at most.*

Hint: Reduce it to the two-dimensional analogue.

• **Exercise 4.1.2** Perspective collineation in vector form.

Prove that any perspective collineation κ in $\mathbb{P}^3(\mathbb{F})$ with center $\mathbf{z}\mathbb{F}$ and axis $\mathbf{u}\mathbb{F}$ can be represented as $\kappa\colon \mathbf{x}\mathbb{F} \mapsto \mathbf{x}'\mathbb{F}$, where $\mathbf{x}' = \mathbf{x} + \langle \mathbf{u}, \mathbf{x} \rangle \mathbf{z}$, provided that $1 + \langle \mathbf{u}, \mathbf{z} \rangle \neq 0$.

!)*+,)))-) #& 2 = / > 4
. 8 0 ! / 5 61 5 5 8 5!
' &$ (4 $

Metric relations between curvature radii and tangential distances

Curves can also be defined as envelopes of one-parameter families of lines. Each line is uniquely defined by a unit normal vector $\mathbf{n} = (\cos\varphi, \sin\varphi)$ and the oriented distance $d(\varphi)$ to the origin of the coordinate system such that $d\mathbf{n}$ is the position vector of the line's pedal point with respect to the origin. The function $d(\varphi)$ is called the *support function* of the curve.

The curvature and the support function of an ellipse fulfill a simple relation:

Theorem 3.2.1 *The curvature function κ and the support function d of an ellipse with semimajor axis a and semiminor axis b are related via*

$$\kappa = \frac{d^3}{a^2b^2}. \tag{3.19}$$

Proof: We shall have in mind that $x = a\cos t$ and $y = b\sin t$ (cf. (2.12)) holds for any point on the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1. \tag{3.20}$$

After eliminating $\cos t$ and $\sin t$ from the curvature function given in (3.17), we arrive at an algebraic expression for the curvature:

$$\kappa(t) = \frac{a^4b^4}{\sqrt{(b^4x^2 + a^4y^2)^3}}. \tag{3.21}$$

Via the velocity vector of the ellipse at some point $P = (\xi, \eta)$, we obtain the Hesse normal form of the tangent at P as

$$t_P : \quad \frac{-b^2\xi\, x - a^2\eta\, y + a^2b^2}{\sqrt{b^4\xi^2 + a^4\eta^2}} = 0. \tag{3.22}$$

If we insert $\xi = \eta = 0$, we obtain the signed distance of t_P to the center of the ellipse,

$$d(x,y) = \frac{a^2b^2}{\sqrt{b^4x^2 + a^4y^2}} \tag{3.23}$$

Note that x and y in (3.23) have to fulfill (3.20). Combining (3.21) and (3.23) we can eliminate the nasty square root and obtain (3.19). ■

- **Exercise 3.2.2** Analyze the curvature function of the ellipse given in (3.19) and show that the vertices of the ellipse appear where the support function is stationary.

- **Exercise 3.2.3** Derive the analogous formula to (3.19) for hyperbolas.

A deeper result relating the curvature radii with metric values is due to the French mathematician JOSEPH LIOUVILLE (1809–1882) and given in

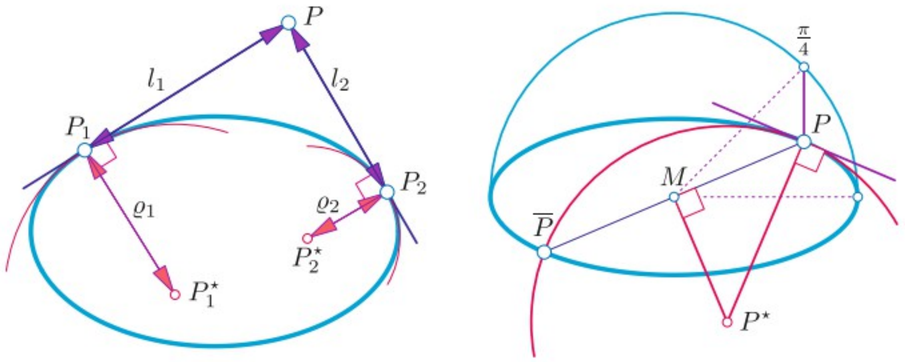

FIGURE 3.12. Left: The cube of the ratio $l_1 : l_2$ equals the ratio $\varrho_1 : \varrho_2$ of the curvatures at the two contact points P_1 and P_2. Right: For the particular point P, the osculating circle o_P meets the ellipse c at the antipode $\overline{P}$ of P.

Theorem 3.2.2 *Let P be an exterior point of a conic c. The two tangents from P to c touch at P_1 and P_2. Let $l_1 = \overline{P_1P}$ and $l_2 = \overline{P_2P}$ and let further $\varrho_1 = \kappa_1^{-1}$ and $\varrho_2 = \kappa_2^{-1}$ be the curvature radii of the conic at P_1 and P_2. Then,*

$$\frac{\varrho_1}{\varrho_2} = \left(\frac{l_1}{l_2}\right)^3$$

holds independently of the choice of P and the conic.

• **Exercise 3.2.4** Give a proof of Theorem 3.2.2 by assuming P_1 and P_2 are arbitrary points on an ellipse / hyperbola / parabola with the parametrizations

$$\text{ellipse} \quad \left(a\frac{1-t^2}{1+t^2}, b\frac{2t}{1+t^2}\right), \quad \text{hyperbola} \quad \left(a\frac{1+t^2}{1-t^2}, b\frac{2t}{1-t^2}\right), \quad \text{parabola} \quad \left(2qt^2, 2qt\right)$$

with $a, b, p \in \mathbb{R}$. Compute the intersection P of the tangents at P_1 and P_2, the distances $\overline{P_1P}$ and $\overline{P_2P}$, and the curvatures at the respective points on the conic.

In Figure 3.12 (left), the metric values mentioned in Theorem 3.2.2 are marked. Though Figure 3.12 shows only the case of an ellipse, Theorem 3.2.2 holds for hyperbolas and parabolas as well.

Figure 3.12 (right) shows another curious result. There are four points P on the ellipse (corresponding to the parameter values $t = \pm\frac{\pi}{4}$ and $t = \pm\frac{3\pi}{4}$ in the trigonometric parametrization (2.12)) whose osculating circles meet the ellipse in a pair $(P, \overline{P})$ of opposite points. In this special case, the center $P^\star$ of the osculating circle lies on the diameter line which is orthogonal to the diameter $[P, \overline{P}]$.

- **Exercise 3.2.5** Find a proof for the facts illustrated in Figure 3.12 (right).

- **Exercise 3.2.6** Vertices of a hyperbola. Show that there are two vertices $V_1 = (a, 0)$ and $V_2 = (-a, 0)$ on the hyperbola

$$h: \frac{x^2}{a^2} - \frac{y^2}{b^2} = 1 \quad \text{with} \quad ab \neq 0$$

by using the two parametrizations

$$\mathbf{h}(t) = (\pm a \cosh t, b \sinh t), \quad t \in \mathbb{R}$$

and compute the curvature (cf. the curvature plot in Figure 3.13)

$$\kappa = \mp \frac{ab}{(a^2 \sinh^2 t + b^2 \cosh^2 t)^{\frac{3}{2}}}.$$

The radius of the osculating circle at V_1 and V_2 equals $\varrho_1 = \varrho_2 = \frac{b^2}{a}$. As shown in Figure 3.14 (left), the centers $V_1^\star$ and $V_2^\star$ of the osculating (hyperosculating) circles at the hyperbola's vertices can be constructed in a simple way, provided that the vertex and at least one asymptote is known.

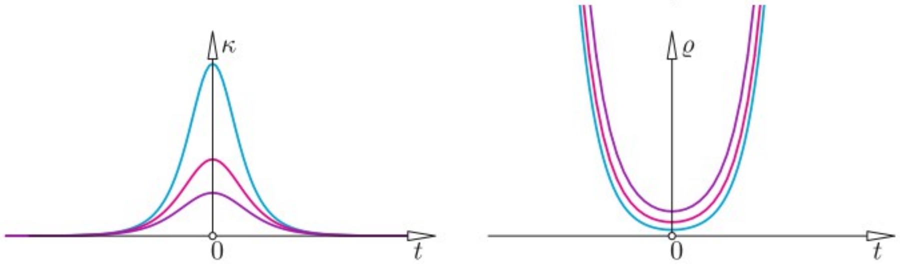

FIGURE 3.13. Curvature plot and the curvature radius function of the hyperbola for three different ratios of the semimajor and semiminor axes: 3 : 2 (blue), 1 : 1 (magenta), and 3 : 4 (violet).

- **Exercise 3.2.7** Vertex of a parabola. A parabola $p:\ x^2 = 2qy$ (with $q \in \mathbb{R} \setminus \{0\}$) can be given in parametric form as

$$\mathbf{p}(t) = (2qt, 2qt^2), \quad t \in \mathbb{R}.$$

Thus, the curvature radius function $\varrho = \frac{1}{\kappa}$ reads

$$\varrho = q(1 + 4t^2)^{\frac{3}{2}}.$$

Show that p has only one vertex $V = (0, 0)$ which corresponds to $t = 0$. The radius of the osculating circle at V equals $\varrho_0 = q$.

In Section 4.2 (Theorem 4.2.3, page 145) we learn that the *subnormals* (the orthogonal projections of the normal's segment between the curve and its principal axis onto the latter) of conics are linear functions in the abscissa. The parabola is more special: The subnormal is constant, and thus, independent of the point on the parabola as shown in Figure 3.14.

-7 - ?

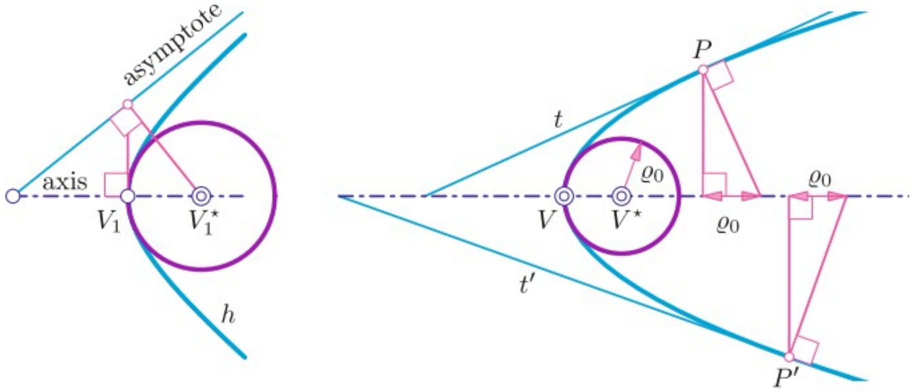

FIGURE 3.14. Left: The construction of the center $V_1^\star$ of the hyperbola's osculating circle at the vertex V_1. Right: The radius ϱ_0 of the osculating circle at a parabola's vertex V can be found if a line element, *i.e.*, a point $P \neq V$ plus incident tangent, is known.

Evolute of an ellipse

What about the osculating circle o_P for a generic point $P = \mathbf{c}(t_0)$ of an ellipse c? Assume that o is such a circle with center $P^\star$ (with coordinate vector $\mathbf{c}^\star$ and radius R). Locally, that means, in a sufficiently small neighborhood of a generic point $P = \mathbf{c}(t_0)$ on the ellipse c, the circle and the ellipse share the point, the tangent, and the curvature.

We can express these three conditions algebraically: The first condition yields

$$\langle \mathbf{c}(t) - \mathbf{c}^\star, \mathbf{c}(t) - \mathbf{c}^\star \rangle|_{t=t_0} = R^2,$$

i.e., a circle through $\mathbf{c}(t)$ (at any t) with the yet unknown center $\mathbf{c}^\star$ and radius R. Note that both, center and radius are constant at a certain point $P = \mathbf{c}(t_0)$. We differentiate once with respect to t and obtain

$$\langle \dot{\mathbf{c}}(t), \mathbf{c}(t) - \mathbf{c}^\star \rangle|_{t=t_0} = 0 \quad \Longrightarrow \quad \langle \mathbf{e}_1, \mathbf{c}(t) - \mathbf{c}^\star \rangle = 0 \tag{3.24}$$

where $(\mathbf{e}_1, \mathbf{e}_2)$ is the Frenet frame at P. The last equation tells us that the desired circle o is tangent to c at $P = \mathbf{c}(t_0)$, for its radius vector $\mathbf{c}(t) - \mathbf{c}^\star$ is perpendicular to the tangent t_P of c at P that aims in the direction $\dot{\mathbf{c}}(t)$ or $\mathbf{e}_1$. Thus, the center $P^\star$ of o is a point on the curve's normal at P.

We differentiate once again. With $\dot{\mathbf{c}} = v\kappa\mathbf{e}_2$ and $\langle \mathbf{e}_1, \mathbf{e}_1 \rangle = 1$, we find

$$v\langle \kappa\mathbf{e}_2, \mathbf{c}(t) - \mathbf{c}^\star \rangle|_{t=t_0} + v = 0 \tag{3.25}$$

which yields

$$\mathbf{c}^\star = \mathbf{c} + \frac{1}{\kappa}\mathbf{e}_2. \tag{3.26}$$

At this point we recall that, by virtue of (3.16), κ can be expressed in terms of the derivatives of the parametrization $\mathbf{c}(t)$ (with an arbitrary parameter t which does not necessarily agree with the arc length parameter).

We learned that $\mathbf{c}^\star$ is the center of the osculating circle or the *center of curvature*. The term *osculation* is explained for pairs of conics in a completely different way in Definition 6.4.1 (see Section 6.3) although the two separate approaches lead to the same circle.[10]

If t runs through the entire interval $J = [0, 2\pi[$, then $\mathbf{c}(t)$ parametrizes the ellipse and $\mathbf{c}^\star(t)$ parametrizes a curve called the *evolute of the ellipse.* The tangents of the evolute are the normals of the ellipse which follows from $\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{c}^\star = v\mathbf{e}_1 - \frac{\dot{\kappa}}{\kappa^2}\mathbf{e}_2 - \frac{v\kappa}{\kappa}\mathbf{e}_1 = -\frac{\dot{\kappa}}{\kappa^2}\mathbf{e}_2$. Figure 3.15 shows some ellipses together with their evolutes.

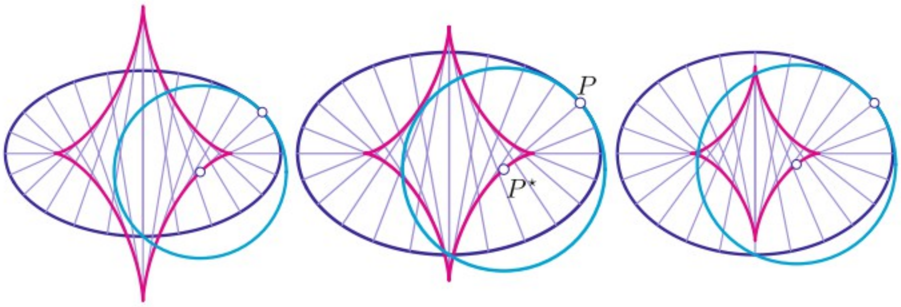

FIGURE 3.15. Evolutes of different ellipses being the envelopes of the ellipses' normals. The cusps of the evolute correspond to the vertices of the ellipse.

---

[10] Two C^r curves c and d are said to be in n-th order contact (with $n \le r$) if there exist parametrizations of both curves such that the derivatives of c and d agree up to order n. Sometimes this is called a C^n-contact. Two C^r curves are said to be in G^n-contact (geometric continuity of order $n \le r$) if the span of the first n derivative vectors of both curves agree. C^n-contact always implies G^n-contact, the converse is not true.

The projective differential geometric version deals with the first $n+1$ derivative points starting with $c(t_0)$ and $d(t_0)$ as the first derivatives.

We insert the parametrization (2.12) of the ellipse c into (3.26) and obtain a parametrization of the evolute as

$$\mathbf{c}^{\star}(t) = \frac{a^2 - b^2}{ab}\left(b\cos^3 t, -a\sin^3 t\right), \quad t \in J. \tag{3.27}$$

The evolute of an ellipse is an algebraic curve of degree six and class[11] four. The parametric representation (3.27) fulfills the algebraic equation

$$\left(\frac{x}{\alpha}\right)^{\frac{2}{3}} + \left(\frac{y}{\beta}\right)^{\frac{2}{3}} = 1 \tag{3.28}$$

with $\alpha = \frac{a^2-b^2}{a}$ and $\beta = \frac{a^2-b^2}{b}$ which is equivalent to the polynomial equation

$$\left((a^2-b^2)^2 - (a^2x^2+b^2y^2)\right)^3 - 27a^2b^2x^2y^2(a^2-b^2)^2 = 0.$$

The evolute of an ellipse has four cusps which are the centers of curvature of the four vertices, see Figure 3.15.

The scaling $(x, y) \mapsto (x', y') = \left(\frac{x}{\alpha}, \frac{y}{\beta}\right)$ transforms (3.28) into the *astroid* with the equation

$$x'^{\frac{2}{3}} + y'^{\frac{2}{3}} = 1.$$

For a generic point P on the ellipse $\mathbf{c}(t)$, there exists also an elementary construction of the center of curvature $P^{\star} = \mathbf{c}^{\star}(t)$, cf. Figure 3.16 (left):

Lemma 3.2.2 *Draw the tangent t_P and the normal n_P at $P \in c$. The normal intersects the principal axis in the point 1. The line parallel to t_P through 1 meets the ellipse's diameter through P in the point 2. The line $[2, P^{\star}]$ is parallel to the auxiliary axis of c, and $P^{\star}$ is the center of curvature which has to lie on the curve's normal n_P.*

Proof: We can verify the construction shown in Figure 3.16 (left) by a simple computation. We start with the trigonometric parametrization of c given in (2.12) and set $P = (a\cos t, b\sin t)$. For fixed t_0 we find a point P on c. Then, the tangent t_P and the normal n_P of e at P are given by the Cartesian equations

$$t_P: \ \frac{x\cos t}{a} + \frac{y\sin t}{b} = 1 \quad \text{and} \quad n_P: \ -\frac{x\sin t}{b} + \frac{y\cos t}{a} = \frac{b^2-a^2}{ab}\cos t\sin t.$$

---

[11] The class of a planar algebraic curve is, roughly speaking, the number of tangents to the curve that can be drawn from a generic point to the curve. In Section 7.5, we shall obtain the evolute of a conic as the image of the set of the conic's tangents under a quadratic Cremona transformation. Then, it will be immediately clear why the class of the evolute is four.

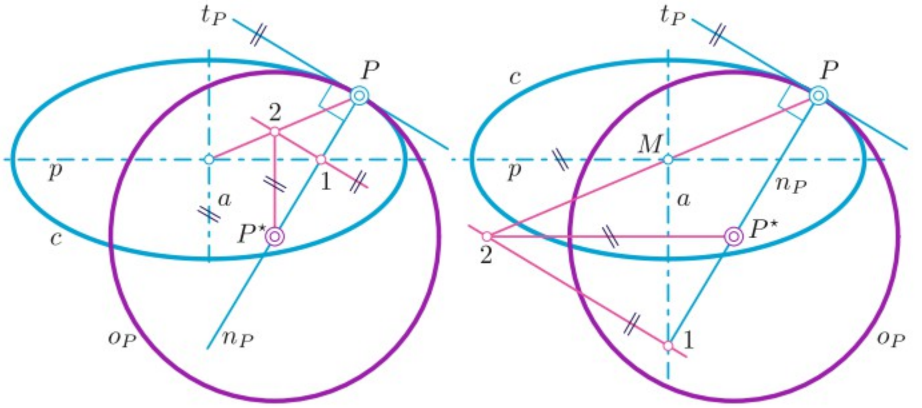

FIGURE 3.16. Two constructions of the center of curvature of an ellipse c. The construction also works if we interchange the roles of the axes.

The normal n_P and the principal axis with equation $y = 0$ meet at 1 with coordinates

$$\left(\frac{a^2 - b^2}{a}\cos t, 0\right),$$

and thus, the line l parallel to the tangent t_P has the equation

$$l:\ \frac{x\cos t}{a} + \frac{y\sin t}{b} = \frac{a^2 - b^2}{a^2}\cos^2 t.$$

The diameter through P is given by $xb\sin t - ya\cos t = 0$. Therefore, its intersection 2 with l has the same x-coordinate as $P^\star$

$$x_{P^\star} = \frac{a^2 - b^2}{a}\cos^3 t$$

which, inserted into the normal's equation, yields $y_{P^\star}$, and thus, $\mathbf{p}^\star(t)$ agrees with $\mathbf{c}^\star$ from (3.27). ■

• **Exercise 3.2.8** Alternative curvature construction. In Figure 3.16 (right), we can see that there is a second construction that leads to the center $P^\star$ of curvature for a generic point P on the ellipse c. Show that this construction leads to the same $P^\star$, *i.e.*, it is equivalent to the first construction given above.

Theorem 3.2.3 *When in the case of an ellipse or a hyperbola c the tangent t_P and the normal n_P at a generic point $P \in c$ intersect the principal axis at points 1, $1'$ and the auxiliary axis at 2, $2'$, then the signed distances to P and the center $P^\star$ of curvature satisfy*

$$\overline{P1} : \overline{P2} = \overline{P^\star 1'} : \overline{P^\star 2'}. \tag{3.29}$$

• **Exercise 3.2.9** Proof of Theorem 3.2.3. Show that the construction of $P^\star$ from Theorem 3.2.3 as shown in Figure 3.17 is equivalent to that given in Theorem 3.2.2.

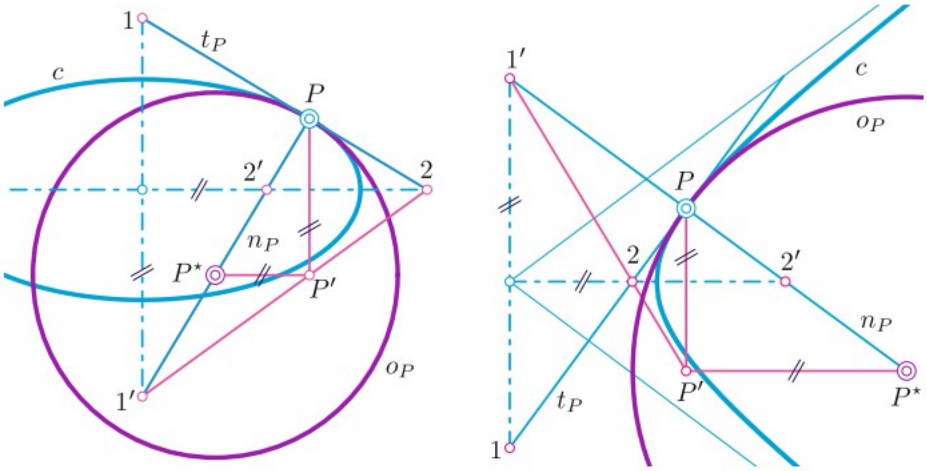

FIGURE 3.17. An alternative constructive approach to the centers of curvature of ellipses and hyperbolas. P divides the segment 12 on t_P at the same ratio as $P^\star$ does with $1'2'$ on n_P.

• **Exercise 3.2.10** Evolute of an ellipse as the envelope of the ellipse's normals. Show that the evolute of an ellipse can be found as the envelope of its normals. Start with the equations of lines depending on $t \in [0, 2\pi[$

$$n_P: \ -\frac{x \sin t}{b} + \frac{y \cos t}{a} = \frac{b^2 - a^2}{ab} \cos t \sin t$$

and intersect the lines n_P with their derivatives $\dot{n}_P$ (with respect to t). The points $n_P \cap \dot{n}_P$ are precisely the points of the ellipse's evolute.

• **Exercise 3.2.11** Evolutes of hyperbola and parabola as envelopes. Find the evolute of a hyperbola and a parabola as the envelopes of the normals.

The evolute of a parabola

Assume that we are given a parabola $p\colon\ x^2 = 2qy$ (with $q \in \mathbb{R}\setminus\{0\}$) which can be parametized as

$$\mathbf{p}(t) = (2qt, 2qt^2), \quad t \in \mathbb{R}. \tag{3.30}$$

We insert (3.30) into (3.26) and find a parametrization of its evolute

$$\mathbf{p}^\star(t) = q(-8t^3, 6t^2 + 1) \tag{3.31}$$

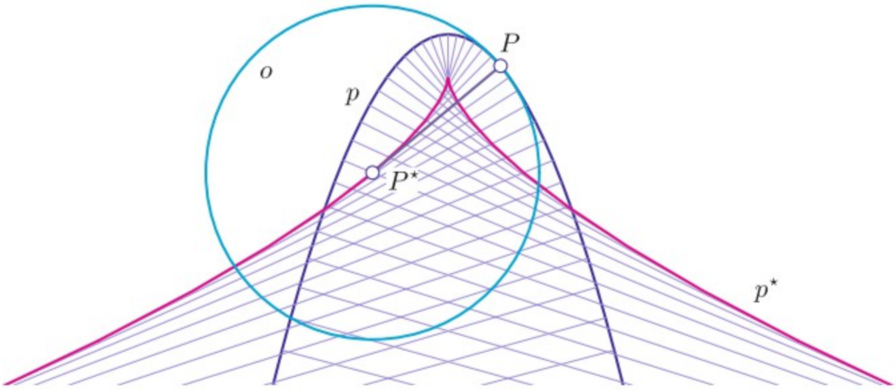

FIGURE 3.18. The evolute of a parabola as the envelope of the parabola's normals.

which is even polynomial, and thus, it can be written as a Bézier curve (see Section 8.2). This curve is sometimes called a *semicubical parabola*. An example of a parabola's evolute can be seen in Figure 3.18.

The curve given in (3.31) is algebraic and of degree three and of class three, although it is the evolute of a parabola. Its only cusp is the point $(q, 0)^{\mathrm{T}}$ which is the center of the osculating circle at the vertex. An implicit equation of the parabola's evolute from (3.31) reads

$$8(y-q)^3 - 27qx^2 = 0$$

which can be found by eliminating t from $x = -8qt^3$ and $y = 6qt^2 + q$.

• **Exercise 3.2.12** Center of curvature for a generic point on a parabola Assume $P \in p$ is a generic point (not the vertex). We can find the center $P^\star$ of p's osculating circle at P in two different ways shown in Figure 3.19:

The construction shown on the left-hand side of Figure 3.19 is the analogue to the construction of the center of curvature of the ellipse (cf. Figure 3.16). The parabola's normal n_P at P meets the axis of the parabola at a point D. There we draw the parallel to the tangent t_P. In the case of the ellipse, the point P has to be joined with the center of the curve. In the present case of the parabola, there is no center. However, there is a diameter through P, *i.e.*, the line parallel to the parabola's axis which meets the aforementioned parallel to the tangent at the point 2, such that $[2, C]$ is orthogonal to all diameters of p, especially to its axis.

Prove this by direct computation.

Theorem 3.2.4 *The normal n_P at any point P of a parabola meets the parabola's directrix at a point D. If $P^\star$ is the parabola's center of curvature*

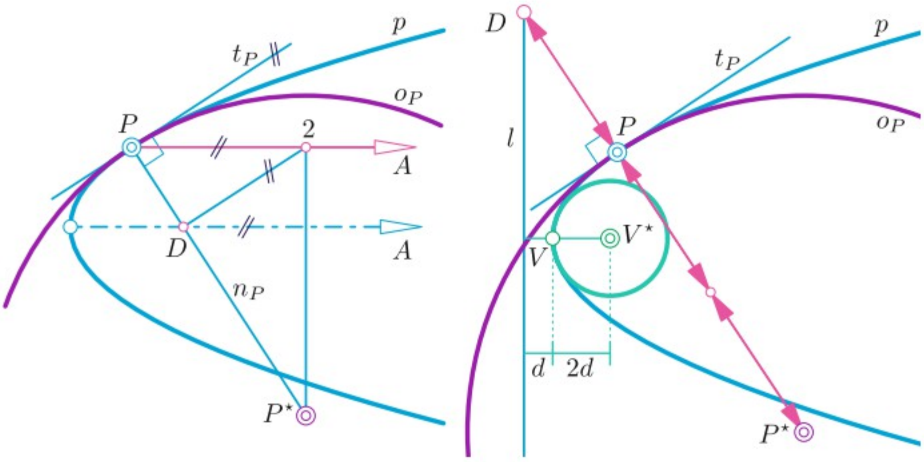

FIGURE 3.19. Two constructions of the center of curvature of parabolas.

at P, then the lengths of the segments DP and $DP^\star$ fulfill

$$\overline{DP} : \overline{DP^\star} = 1 : 3.$$

• **Exercise 3.2.13** Alternative approach to the center of curvature. The construction in Figure 3.19, right, for the center $P^\star$ of curvature of a parabola is also valid at the vertex (cf. Theorem 3.2.4). Give a proof of Theorem 3.2.4.

Evolute of a hyperbola

With the preparations from the previous sections, it is easily verified that the evolute of the hyperbola parametrized by

$$(\pm a \cosh t, b \sinh t), \quad t \in \mathbb{R}$$

admits the parametrization

$$\frac{a^2 + b^2}{ab} \left(\pm b \cosh^3 t, \mp a \sinh^3 t \right), \quad t \in \mathbb{R}.$$

An implicit equation of the evolute of the hyperbola reads either

$$\left(\frac{ax}{a^2 + b^2} \right)^{\frac{2}{3}} - \left(\frac{by}{a^2 + b^2} \right)^{\frac{2}{3}} = 1 \quad \text{or} \quad \left(1 - \frac{a^2x^2 - b^2y^2}{(a^2 + b^2)^2} \right)^3 = 27 \frac{(abxy)^2}{(a^2 + b^2)^4}.$$

Like the evolute of the ellipse, the evolute of a hyperbola is a sextic curve of class four. It has two real cusps being the centers of the osculating

21 04 2020

Figure 4.5 a) Self-polar triangle
b) Self-conjugate.

Proof: Ceva and Menelaos the product of the cross-ratios on the three sides must be negative, which results in the two cases.

Pag. 129(137)

Figure 4.3 Perspectivities in the projective space.

Obs: Livro muito bom!!!

D2) Books_70_2020/
The Universe of Quadrics/
The Universe of Conics ...pdf

Obs. Livro muito bom!!!

Pag. 79(87)

Curvature of conics and the construction of osculating circles.

Pag. 83(91)

Figure 3.12

Right: For the particular point P, the osculating circle op meets the ellipse c at the antipode $\bar{P}$ of P.

)*+,)))-) #& @ > !
! 55 1 ! 7 5!
)*+,)))-) #& (& #6 $;# 5!
' $ $6

Chapter 1

Properties of Radiation

1.1 Preamble

Early studies of the properties of electric and magnetic fields, their interplay and the resulting emission of radiation, led to formation of several fundamental laws by Carl F. Gauss (1777–1855), Michael Faraday (1791–1867), André M. Ampère (1775–1836), and James C. Maxwell (1831–1879). These were presented in a compact differential formalism by Maxwell (1865), which are referred to as Maxwell's equations:

$$\nabla \cdot \mathbf{D}(\mathbf{r}, t) = \rho(\mathbf{r}, t), \tag{1.1}$$

$$\nabla \cdot \mathbf{B}(\mathbf{r}, t) = 0, \tag{1.2}$$

$$\nabla \times \mathbf{E}(\mathbf{r}, t) = -\frac{\partial \mathbf{B}(\mathbf{r}, t)}{\partial t}, \tag{1.3}$$

$$\nabla \times \mathbf{H}(\mathbf{r}, t) = \mathbf{J}(\mathbf{r}, t) + \frac{\partial \mathbf{D}(\mathbf{r}, t)}{\partial t}. \tag{1.4}$$

The electric $\mathbf{E}(\mathbf{r}, t)$ and the magnetic fields $\mathbf{B}(\mathbf{r}, t)$, are time dependent and can be specified at every point in space and time, in which $\mathbf{r}(= x, y, z)$ denotes the position vector, t is the time, $\mathbf{J} = \sigma\mathbf{E}$, the electric current density, $q(= 1.6 \times 10^{-19}$ C) the elementary charge, $\mathbf{E}$ and $\mathbf{B}$, the electric field strength and the magnetic induction, respectively, $\mathbf{H} = \mathbf{B}/\mu$, the magnetic field, $\mathbf{D} = \epsilon\mathbf{E}$, the

High-Resolution Imaging: Detectors and Applications
Swapan K. Saha
Copyright ⓒ 2015 Pan Stanford Publishing Pte. Ltd.
ISBN 978-981-4613-27-9 (Hardcover), 978-981-4613-28-6 (eBook)
www.panstanford.com

© 2015 by Taylor & Francis Group, LLC

electric displacement, ρ, the volume density of free charge, ϵ, μ, and σ the respective dielectric permittivity, the magnetic permeability, and conductivity of the medium, "×," the cross product, and

$$\nabla = \mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z}. \tag{1.5}$$

represents a linear vector differential operator.

The unit of the electric field intensity, $\mathbf{E}$, is expressed as volt (V) m^{-1}, and that for the magnetic flux density $|\mathbf{B}|$ (in tesla; $\mathrm{T} = \mathrm{Wb\ m}^{-2}$). Equation (1.1) represents Gauss' theorem in electrostatics. Equation (1.2) tells that whatever might be the origin of $\mathbf{B}$, be it static magnetic poles or conduction current, the divergence is zero. Equation (1.3) is an outcome of Faraday's law of electromagnetic induction, which states that a changing magnetic field produces an electric field. Equation (1.4) originates from the circuital form of Ampère's law subject to satisfying the equation of continuity. What follows from circuital form of Ampère's law is

$$\nabla \times \mathbf{H} = \mathbf{J}. \tag{1.6}$$

From equation (1.6), one may obtain, $\nabla \cdot (\nabla \times \mathbf{H}) = 0$, which is an identity. That is

$$\nabla \cdot \mathbf{J} = 0, \tag{1.7}$$

which is the equation of continuity in a restricted sense of steady charge.

The equation of continuity actually is

$$\nabla \cdot \mathbf{J} + \frac{\partial \rho}{\partial t} = 0. \tag{1.8}$$

On using equation (1.1), one can rewrite equation (1.8)

$$\nabla \cdot \mathbf{J} + \frac{\partial}{\partial t}(\nabla \cdot \mathbf{D}) = 0. \tag{1.9}$$

Therefore,

$$\nabla \cdot \left(\mathbf{J} + \frac{\partial D}{\partial t}\right) = 0. \tag{1.10}$$

Using equation (1.10), Maxwell modified equation (1.6), by introducing the additional term, $\partial \mathbf{D}/\partial t$ on the right hand side (RHS), named it the displacement current density to obtain equation (1.4), which is also known as Ampère–Maxwell law.

© 2015 by Taylor & Francis Group, LLC

By taking surface integral on both sides of equation (1.4) and converting the left hand side (LHS) into a line integral, one obtains the magnetomotive force (MMF) as $\oint_1 \mathbf{H} \cdot d\mathbf{l}$. That is

$$\oint_1 \mathbf{H} \cdot d\mathbf{l} = \int_s \left(\mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \right) \cdot d\mathbf{S}. \tag{1.11}$$

The first term on the RHS indicates that MMF originates from current carrying conductor; this was postulated by Ampère and observed by Oersted. The second term indicates that time variation of the flux of electric displacement vector is a source of MMF. This reflects a symmetry of nature because it shows that as according to Faraday's law of electromagnetic induction, variation in the flux of magnetic induction gives rise to electromotive force (EMF), the variation of the flux of electric induction vector gives rise to MMF.

An exact description of electromagnetic radiation and its behavior requires a knowledge of Maxwell's electromagnetic field equations that describe the temporal and spatial dependence of electromagnetic fields and provide good agreement with observed phenomena over a wide range of frequencies in the spectrum. Equations (1.3) and (1.4) in free space (i.e., $\rho = 0$, $\mathbf{J} = 0$, $\epsilon = 0$), may be written as

$$\nabla \times \mathbf{E}(\mathbf{r}, t) = -\mu_0 \frac{\partial \mathbf{H}(\mathbf{r}, t)}{\partial t}, \tag{1.12}$$

$$\nabla \times \mathbf{H}(\mathbf{r}, t) = \epsilon_0 \frac{\partial \mathbf{E}(\mathbf{r}, t)}{\partial t}, \tag{1.13}$$

in which $\mu_0 (= 4\pi k = 4\pi \times 10^{-7}$ henrys (H).m^{-1}) is the permeability in free space (vacuum) and $\epsilon_0 (= 8.8541 \times 10^{-12}$ farads (F).m^{-1}) the permittivity in vacuum.

The curl of equation (1.12) is expressed as

$$\nabla \times \nabla \times \mathbf{E}(\mathbf{r}, t) = -\mu_0 \epsilon_0 \frac{\partial^2 \mathbf{E}(\mathbf{r}, t)}{\partial t^2}, \tag{1.14}$$

with $\nabla \times \nabla \times \mathbf{E} = \nabla(\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E}$. Since $\nabla \cdot \mathbf{E} = 0$, in free space, one may derive the wave equation as

$$\nabla^2 \mathbf{E}(\mathbf{r}, t) - \mu_0 \epsilon_0 \frac{\partial^2 \mathbf{E}(\mathbf{r}, t)}{\partial t^2} = 0, \tag{1.15}$$

$$\nabla^2 \mathbf{E}(\mathbf{r}, t) - \frac{1}{c^2} \frac{\partial^2 \mathbf{E}(\mathbf{r}, t)}{\partial t^2} = 0. \tag{1.16}$$

© 2015 by Taylor & Francis Group, LLC

Similarly, one may derive for **B**,

$$\nabla^2\mathbf{B}(\mathbf{r}, t) - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}[\mathbf{B}(\mathbf{r}, t)] = 0. \tag{1.17}$$

The permittivity constant ϵ_0 and the permeability constant μ_0 in a vacuum are related to the speed of light, c

$$c = \frac{1}{\sqrt{\mu_0\epsilon_0}} \approx 3 \times 10^8 \text{ meter (m)/second (s)}. \tag{1.18}$$

The above expressions (1.16–1.17) are known as the electromagnetic wave equations, which indicate that electromagnetic disturbances (waves) are propagated through free space. This result gives rise to Maxwell's electromagnetic theory of light. The propagation velocity, $v(= 1/\sqrt{\epsilon\mu})$, of the waves obey the wave equations; in free space $v = c = 1/\sqrt{\mu_0\epsilon_0}$.

The union of electromagnetic theory with optics began when these equations described waves traveling at the velocity of light. However, the development of quantum theory of light at the dawn of the last century, dealing with absorption and emission, cannot be described with just Maxwell's equations. While interpreting the spectrum of electromagnetic radiation emitted by a blackbody, Max Karl Ernst Ludwig Planck (1858–1947) postulated that the oscillating electric charges which gives rise to light emission can have discrete energies (Planck, 1900). The name, blackbody, comes from the assumption that the body absorbs at every frequency and hence would look black at low temperature. The radiation emitted by a body as a result of its temperature is called thermal radiation. All bodies emit thermal radiation to their surroundings and absorb such radiations from them where the rates depend upon the temperature difference.

The nature of the spectrum of thermal radiation is continuous for matter in a condensed state (i.e. solid and liquid) and is strongly dependent on temperature. Generally speaking, the detailed nature of the radiation spectrum depends somewhat on the composition of the body. However, experiments show that there is one class of hot bodies that emits thermal spectra of universal character. These are called black bodies, which have surfaces absorbing all the thermal radiation incident upon them (in the ideal case) and re-radiate energy, which is characteristic of these radiating systems.

© 2015 by Taylor & Francis Group, LLC

However, one cannot explain the radiation emitted from such bodies on the basis of the laws of classical physics. Following Planck's suggestions, Albert Einstein (1879–1955) put forward that light is not distributed in continuous waves, but travels as a finite number of points of energy quanta, called *lichtquanten*, which are localized at points in space and move without dividing and can be absorbed or generated only in complete units (Einstein, 1905a). They travel at the speed of light and have properties like energy $\hbar\omega$, zero rest mass, momentum, spin $\hbar/2$, with $\hbar = h/(2\pi)$, $h(= 6.626096 \times 10^{-34}$ joules (J) seconds (s) $= 4.136 \times 10^{-15}$ eVs) as Planck's constant. This gave way to a term, photon, introduced by a chemist Gilbert Newton Lewis (1875–1946; Lewis, 1926), which eventually led to the development of quantum theory.

In fact, the birth of quantum mechanics is intimately linked with the theories and discoveries relating to the nature of light. Sir Isaac Newton (1642–1727) had described light as a stream of particles, what was termed as corpuscular (little particles) theory. This view had been superseded by the wave picture of light (Huygens' wavefront theory), a picture that culminated in the electromagnetic theory of Maxwell.

The energy of a photon depends on the radiation frequency, ν, of the light; the shorter the wavelength, the more energy each of these quanta has, the longer the wavelength the smaller the energy of the quanta. There are photons of all energies from high-energy gamma rays to low-energy infrared and radio waves. Photons are bosons,[a] having zero mass and rest energy and one unit of spin. Planck postulated that radiation of frequency, ν, can exchange energy with matter in units of $h\nu$. In other words, radiation of frequency, ν,

---

[a] Bosons, named after Satyendra Nath Bose (1894–1974), are particles (integer spin) that are associated with integral multiples of $\hbar$. These particles account for the phenomena of laser (acronym for light amplification by stimulated emission of radiation) (see Section 3.2), superconductivity, and superfluidity. Unlike in the case of fermions, which were named after Enrico Fermi (1901–1954), the particles with spin having odd multiples of $(1/2)\hbar$ obeying Pauli's exclusion principle, bosons can go into the ground state. A photon is a boson, whereas an electron is a fermion. Pauli's exclusion principle states that no two electrons in an atom can have identical quantum number; each energy level can accommodate a maximum of two electrons, one with spin-up $(+1/2)$ and the other with spin-down $(-1/2)$.

© 2015 by Taylor & Francis Group, LLC

behaves like a particle, called a photon, of energy,

$$E = h\nu = \frac{h\omega}{2\pi} = \hbar\omega, \tag{1.19}$$

and momentum, defined by the product of the mass of a particle and its linear velocity, $\mathbf{p}$, is given by,

$$\mathbf{p} = \hbar\boldsymbol{\kappa} = m\mathbf{c}, \tag{1.20}$$

with $\kappa = \omega/c = 2\pi\nu/c = 2\pi/\lambda$ as the wave number, $\boldsymbol{\kappa}(= \kappa\hat{\mathbf{s}})$ the wave vector in medium in the direction of propagation, $\hat{\mathbf{s}}$, ω the angular frequency, λ the wavelength, and c the velocity of light in a vacuum.

Equations (1.19) and (1.20) connect particle properties $(E, \mathbf{p})$ to the corresponding wave properties (ω, κ).

Angular momentum The angular momentum of a system of particles is due to both the spin of individual particles and their orbital momentum (Marcuse, 1980). The spin angular momentum is represented by electrons rotating about their own axes; it comes in units of $\hbar/2$, with $\hbar \equiv h/2\pi$ as the Planck constant divided by 2π. The orbital angular momentum for an atomic electron may be visualized in terms of a vector model where the angular momentum vector is seen as precessing about a direction in space. The total angular momentum is written as

$$\mathbf{J} = \mathbf{L} + \mathbf{S}, \tag{1.21}$$

where $\mathbf{J} = \sqrt{j(j+1)}\hbar$,

$$j = \begin{cases} 0,\ 1,\ 2,\ \cdots & \text{boson,} \\ 1/2,\ 3/2,\ 5/2, & \text{fermion,} \end{cases}$$

which is either an integer or a half integer, $\mathbf{L}$ and $\mathbf{S}$ the orbital and spin angular momentum respectively, which are quantized according to the following rules

$$\mathbf{L} = \sqrt{l(l+1)}\hbar \qquad \mathbf{S} = \sqrt{s(s+1)}\hbar,$$

l is an integer and s takes integer or half integer values depending on the particles. Spin angular momentum consists of two orthogonal states; on the other hand, the orbital angular momentum consists of an infinite number of orthogonal states.

© 2015 by Taylor & Francis Group, LLC

Quantum theory A number of observations had been made towards the end of the nineteenth century establishing the velocity of light independent of the velocity of the source or of the observer. This result could not be explained by classical theory. Einstein (1905b) introduced the special theory of relativity. He extended and improved Maxwell's theory by modifying the foundations of classical mechanics to remove the apparent contradictions between mechanics and electrodynamics. This theory has completely changed the concepts of space and time. He showed that space and time are closely linked and affect each other. Space, time, simultaneity, mass are all relative, but speed of light is absolute being independent of the notion of the source and the observer. The mass-energy equivalence described by the equation, $E = mc^2$, arose from the theory of relativity, where he explained the varying kinetic energies of an object emitting radiation by considering the amount of the object's change in kinetic energy, as being equivalent to change in its mass.

However, in making the transition from Planck's quantized oscillators to quanta of radiation, Einstein introduced the idea of corpuscular behavior of radiation, the revival of which posed a severe conceptual problem, one of reconciling wave and particle like behavior of radiation. Quantum theory of atomic structure, proposed by Niels Henrik David Bohr (1885–1962), known as Bohr's hypothesis, in 1913 and the discovery of dual nature of electromagnetic radiation (wave-particle duality), predicted by Louis de Broglie (1892–1987), what is referred to as de Broglie's hypothesis (de Broglie, 1925), led to the development of quantum mechanics.

Wave-particle duality Wave and particle natures are dual. When one is exhibited, the other remains suppressed. If radiation can exhibit particle nature, a particle can also show wave nature. Following this argument, de Broglie stated that equations (1.19) and (1.20) would hold for a material particle, such as an electron. From wave nature of particle, de Broglie showed that the wavelength of a particle is directly related to the Planck constant and inversely proportional to the mass and velocity of the particle; the larger the object or faster it is moving, the shorter the wavelength. The

© 2015 by Taylor & Francis Group, LLC

de Broglie equations are

$$\lambda = \frac{h}{p} = \frac{h}{\gamma m_0 v} = \frac{h}{m_0 v}\sqrt{1 - \frac{v^2}{c^2}}, \tag{1.22}$$

$$\nu = \frac{E}{h} = \frac{\gamma m_0 c^2}{h} = \frac{m_0 c^2}{h}\frac{1}{\sqrt{1 - v^2/c^2}}, \tag{1.23}$$

in which m_0 designates the particle's rest mass, v the velocity of the particle, c the speed of the light in a vacuum, and $\gamma(\equiv 1/\sqrt{1 - v^2/c^2})$ the Lorentz factor.

These two equations (1.22 and 1.23) relate the wavelength, λ, of a particle to its momentum and the frequency of the wave associated with the particle to the total energy of the particle, respectively. For instance, an electron with an accurately measured value of mass and charge could undergo diffraction in a manner similar to that of light waves.

Poynting vector The energy density of a photon, w in a volume V is given by

$$\mathrm{w} = \frac{\hbar\omega}{\mathrm{V}}. \tag{1.24}$$

Equation (1.24) states that if a plane wave with an energy quantum is contained in a volume, V. The momentum density is expressed as

$$\frac{\mathbf{p}}{\mathrm{V}} = \frac{\mathbf{S}}{c^2} = \frac{\mathrm{w}}{c}\frac{\boldsymbol{\kappa}}{\kappa} = \frac{\hbar\omega/c}{\mathrm{V}}\frac{\boldsymbol{\kappa}}{\kappa} = \frac{\hbar\boldsymbol{\kappa}}{\mathrm{V}}, \tag{1.25}$$

where

$$\mathbf{S}(\mathbf{r}, t) = \mathbf{E}(\mathbf{r}, t) \times \mathbf{H}(\mathbf{r}, t), \tag{1.26}$$

is the Poynting vector, which represents the energy flux of an electromagnetic field having the units of energy per unit area per unit time (J m^{-2} s^{-1}); it is named after its inventor John Henry Poynting (1852–1914).

Uncertainty principle Planck's constant also occurs in statements of Werner Heisenberg's (1901–1976) uncertainty principle (Heisenberg, 1927) which is an important consequence of the wave-particle duality of matter and radiation and is inherent to the quantum description of nature (Beiser, 1973, and references therein). This

© 2015 by Taylor & Francis Group, LLC

provides a lower bound on the product of the standard deviations of two conjugate variables, such as position and momentum for a system or the angular momentum and angular displacement for a system or the energy and time for a moving particle. This principle applies to all conjugate variables and also to the notion of the wave packet. A pair of physical variables describing a quantum-mechanical system such that their commutator is a nonzero constant are also known as complementary variables. The precision in measurement of one of them destroys the possibility of measurement of the other. There is a minimum for the product of the uncertainties of these two measurements

$$\begin{aligned} \Delta x \Delta p &\geq \frac{\hbar}{2}, \\ \Delta L_z \Delta \phi &\geq \frac{\hbar}{2}, \\ \Delta E \Delta t &\geq \frac{\hbar}{2}, \end{aligned} \tag{1.27}$$

where the uncertainty is given as the standard deviation of the measured value from its expected value. Uncertainty should not be construed as a limitation of measurement on the measuring device. It is an inadequacy inbuilt in the process of measurement. The product of two uncertainties is always greater than, or at best, the same as $\hbar/2$.

Schrödinger's equation If there is a wave, there must be a wave equation. Erwin Schrödinger (1887–1961) developed the wave-particle duality equation (Schrödinger, 1926)

$$i\hbar \frac{\partial \Psi}{\partial t} = \hat{H}\Psi, \tag{1.28}$$

in which $i = \sqrt{-1}$, Ψ is the time-dependent wave function, and in one dimension

$$\hat{H} = \frac{-\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + V(x), \tag{1.29}$$

is the Hamiltonian operator, a quantum mechanical operator introduced by Schrödinger in quantum mechanics, representing the energy of the system, which is the sum of kinetic and potential energies.

© 2015 by Taylor & Francis Group, LLC

Operating on the wave function, Ψ, with the Hamiltonian produces Schrödinger's equation. In the time independent Schrödinger's equation, the operation may produce specific values for the energy called energy eigen values, for example

$$\hat{H}\Psi_n = E_n\Psi_n, \tag{1.30}$$

where the functions, Ψ_n, are called eigen functions (or wave functions), possessing the properties that they are single valued, finite everywhere, and also their first derivatives with respect to their variables are continuous everywhere; E the multiplying factor, called eigen value.

For each value of n, there is a energy value and corresponding wave function. Equation (1.29) is a partial differential equation elucidating how the quantum state of some physical system changes with time. The first term on the RHS is the operator associated with kinetic energy and the second term the potential energy. The most precise computations are performed quantum mechanically by calculating the potential precisely and solving this equation. However, the least precise calculations are performed classically, neglecting quantization and using Newton's laws. A semi-classical computation, in which an atom's energy levels that are associated with attraction of the electrons to the nucleus, as well as with the effects of the angular momentum of the electrons, are computed quantum mechanically, but additional effects, such as light waves, are treated classically is an intermediate case. The observations associated with blackbody radiation, which is derived quantum mechanically, the photoelectric effect (see Section 2.1), and other related phenomena associated with the emission or transformation of light could be understood if the energy of light is discontinuously distributed in space. Advantages in technology have also led to matter wave interference experiments being demonstrated successfully with electrons, neutrons, atoms, and molecules.

Quantum statistics Light is quantized, and the statistics of the photons results in uncertainties of the amplitude and the phase of the light field. A quantum state can be occupied by several particles, what is referred to as quantum statistics. This applies to indistinguishable particles. Max Born (1882–1970) formulated the now-standard interpretation of the probability density function for

© 2015 by Taylor & Francis Group, LLC

$\Psi\Psi^*$ in Schrödinger's equation of quantum mechanics (Born, 1926). A probability amplitude is, in general, a complex number whose modulus squared represents a probability or probability density; probability amplitudes may also correspond to probabilities of discrete outcomes. For example, the values taken by a normalized wave function, Ψ, are amplitudes, since $\Psi\Psi^* = |\Psi(x)|^2$ provides the probability density at position x. Probability amplitudes, defined as complex-number-valued function of position, may also correspond to probabilities of discrete outcomes. It is a quantity whose value is a definite complex number at any point in space. The probability of finding the particle described by the wave function (e.g., an electron in an atom) at that point is proportional to square of the absolute value of the probability amplitude. That the physical meaning of the wave function is probabilistic was proposed by Max Born.

1.1.1 *Planck's Law*

At the turn of the last century, Rayleigh and Jeans made a calculation of the energy density, $E_\nu(T)$, of cavity radiation using classical arguments. The radiated energy from a blackbody may be considered to be produced by standing wave or resonant modes of the cavity. By considering the radiation inside a cavity of absolute temperature T whose walls are perfect reflectors to be a series of standing electromagnetic waves. The number of independent standing waves, $G(\nu)d\nu$ in the frequency interval between ν and $d\nu$ per unit volume in the cavity is

$$G(\nu)d\nu = \frac{8\pi\nu^2 d\nu}{c^3}. \tag{1.31}$$

The average total energy, $\bar{\mathcal{E}}$, of these waves in thermal equilibrium at a temperature T can be found out using classical kinetic theory and equipartition law,[a] i.e., assigning an average energy, $\bar{\mathcal{E}}$ per standing wave

$$\bar{\mathcal{E}} = k_\mathrm{B}T, \tag{1.32}$$

---

[a] The theorem of equipartition of energy states that the mean energy per degree of freedom of an entity, for example, a molecule of an ideal gas, which is a member of a system of such entities in thermal equilibrium at the temperature, T, is $1/2k_\mathrm{B}T$.

© 2015 by Taylor & Francis Group, LLC

in which $k_B (= 1.380662 \times 10^{-23}\ \text{J K}^{-1})$ is Boltzmann's constant, to each mode of vibration.

The amount of radiation emitted in a given frequency range should be proportional to the number of modes in that range. They showed that the number of modes was proportional to the square of the frequency. The energy density is related to the spectral distribution of blackbody radiation (spectral radiancy), $B_\nu(T)$, of the spectrum emitted by the hole on the cavity by the following equation

$$\begin{aligned} E_\nu(T)d\nu = \mathcal{E}G(\nu)d\nu &= \frac{4\pi}{c}B_\nu(T)d\nu \\ &= \frac{8\pi\nu^2 k_B T}{c^3}d\nu, \end{aligned} \tag{1.33}$$

which has units of energy per unit volume per unit frequency (J m^{-3} Hz^{-1}).

The number of standing waves in the frequency interval times the average energy of the modes divided by the volume of the cavity gives the average energy per unit volume in the frequency interval ν to $\nu + d\nu$, i.e., the energy density $E_\nu(T)$. The spectral radiancy is defined in such a way that $B_\nu(T)d\nu$ equals the energy emitted per unit time in radiation of frequency in the interval of ν to $\nu + d\nu$ from a unit area of surface at absolute temperature, T. Rayleigh and Jean's result was not successful to justify the experimental results pointing out serious conflict between classical physics and experimental observations. They considered a cavity with metallic walls heated uniformly to temperature T. The walls emit electromagnetic radiation in the thermal range of frequencies. This happens due to the accelerated motions of the electrons in the metallic walls arising from the thermal agitation. In this treatment, the oscillations of the electrons were not considered, instead, attention was focused on the behavior of the standing electromagnetic waves inside the cavity. However, Rayleigh-Jeans' law laid the foundation in understanding of the equilibrium radiation from a blackbody (Rayleigh, 1900, 1905; Jeans, 1905).

Planck considered the energy as a discrete variable instead of a continuous variable. Like the speed of light, c, that sets an upper limit on any physical velocity; Planck's constant, h, turned out to be a universal constant which sets a lower limit on action (linear

© 2015 by Taylor & Francis Group, LLC

momentum multiplied by distance). This constant opened up new vistas of physics. From the assumption that the electromagnetic modes in a cavity were quantized in energy, Planck (1901) obtained a radiation formula. The average energy, $\bar{\mathcal{E}}$[a] per mode (or quantum) is the energy of the quantum times the probability that it would occupy, which is given by

$$\bar{\mathcal{E}}(\nu) = \frac{h\nu}{e^{(h\nu/k_\mathrm{B}T)} - 1}, \tag{1.34}$$

where k_B is Boltzmann's constant, T the temperature in Kelvin (K), e the base of natural logarithms $(2.718\ldots)$, and ν the frequency.

The spectral energy density in the blackbody spectrum is expressed as

$$E_\nu(T)d\nu = \frac{8\pi\nu^2}{c^3}\frac{h\nu}{e^{(h\nu/k_\mathrm{B}T)} - 1}d\nu. \tag{1.35}$$

At high frequencies, $h\nu \gg k_\mathrm{B}T$ and $e^{h\nu/k_\mathrm{B}T} \rightarrow \infty$, which means the spectral energy density, $E_\nu d\nu \rightarrow 0$. There is no more ultraviolet (UV) catastrophe. At low frequencies, $h\nu \ll K_\mathrm{B}T$ and $h\nu/K_\mathrm{B}T \ll 1$, Rayleigh-Jeans' formula holds good and is valid for radio wavelengths. Following

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots,$$

and if x is small, $e^x \approx 1 + x$, one gets

$$\frac{1}{e^{(h\nu/k_\mathrm{B}T)} - 1} \approx \frac{1}{1 + \frac{h\nu}{k_\mathrm{B}T} - 1} \approx \frac{k_\mathrm{B}T}{h\nu} \qquad h\nu \ll k_\mathrm{B}T. \tag{1.36}$$

Thus at low frequencies, equation (1.35) for Planck's law is modified to as

$$E_\nu(T)d\nu \approx \frac{8\pi k_\mathrm{B}T\nu^2}{c^3}d\nu, \tag{1.37}$$

which is Rayleigh-Jeans' formula.

---

[a] The quantity, $\bar{\mathcal{E}}$, can be evaluated from the ratio of sums

$$\bar{\mathcal{E}} = \frac{\int_{n=0}^{\infty}\mathcal{E}\mathcal{P}(\mathcal{E})d\mathcal{E}}{\int_{n=0}^{\infty}\mathcal{P}(\mathcal{E})d\mathcal{E}},$$

in which $\mathcal{P}(\mathcal{E})d\mathcal{E}$ is the probability of finding a given entity of a system with energy in the interval between $\mathcal{E}$ and $\mathcal{E} + d\mathcal{E}$, when the density of states is independent of $\mathcal{E}$.

© 2015 by Taylor & Francis Group, LLC

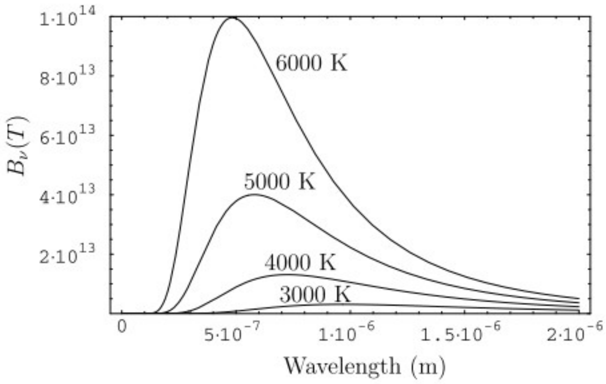

Figure 1.1 Intensity distributions of blackbody radiation at various temperatures, viz., 3000, 4000, 5000, and 6000 K. The abscissa contains the wavelengths and the ordinate depicts the intensity. The wavelength at which the curve is a maximum decreases with the increase of temperature.

A blackbody emits radiation of all wavelengths and the wavelength distribution of the radiation follows Planck's law. This pattern is characterized by

$$B_\nu(T) = \frac{2h}{c^2}\frac{\nu^3}{e^{(h\nu/k_\mathrm{B}T)} - 1}, \tag{1.38}$$

which has units of energy per unit time per unit surface area per unit solid angle per unit frequency (J s^{-1} m^{-2} Sr^{-1} Hz^{-1}).

Figure 1.1 depicts the spectral radiancy, $B_\nu(T)$, plot as function of wavelength; each curve is for a given temperature, and the peak of the curve represents the maximum emission. Stars have an overall spectrum,[a] sources of opacity both lines as well as continuum process, which is governed largely by equation (1.38).

---

[a] There are three types of spectra, such as (i) continuous spectrum that is produced by an excited dense gas radiating at all wavelengths, (ii) emission spectrum, which is produced by an excited low-density gas radiating at a specific wavelength, and (iii) absorption spectrum that results from a light comprising a continuous spectrum while passing through a cool, low-density gas. The gas absorbs some of the wavelengths of light depending on the composition of the gas and the temperature of the light source. The emission spectra, used to study nebulae, are produced when a glowing gas emits energy at specific wavelengths, characteristic of the element composing the gas.

© 2015 by Taylor & Francis Group, LLC

data 21 04 2020

E) Books_70_2020/
High Resolution Imaging Detectors
and Applications/

E1) High-resolution imaging ... pdf

Pág. 03 (36) Preamble

Exact description of electromagnetic radiation and its behavior requires a knowledge of

MAXWELL's electromagnetic field equations...

Pág. 09 (42) Preamble

Schröndinger's equation.

E2) 9814613274.pdf

Pág. 07(40) Preamble

Wave-particle duality.

Pág. 08(41) Broglie equations.

F) Books_70_2020/
Geometria_Moderna/
Geometria - Moise Downs.pdf

obs: Ver livro todo.
Bem didático.

% #(
)*+,)))-) #(B @ * 0
9 ! 5 (; 1 @ > A / 7 5!
' ($6

4. Calculus Revisited

4.1. Differentiation on Intervals

This chapter is a review of what one might call the intellectual skeleton of calculus. It is not intended to be a comprehensive review of the subject because it does not deal extensively with the techniques or applications of calculus. The focus is on the basic concepts and theorems, some of which we will need to extend beyond the level of generality customarily found in a calculus course.

Definition. *Let* I *be an interval on the real line, and let* F *be a function from* I *into* R^m. *If* x *is a point of* I, *we say that* F *is* **differentiable at** x *if the limit*

$$\lim_{t \to x} \frac{F(t) - F(x)}{t - x} \tag{4.1}$$

exists. If F *is differentiable at* x, *the limit* (4.1) *is called the* **derivative of** F **at** x *and is denoted by* $F'(x)$ *or* $(DF)(x)$. *The function* F *is called* **differentiable** *if* F *is differentiable at each point of* I.

If F is any function from I to R^m and if we fix a point $x \in I$, the difference quotient

$$\frac{F(t) - F(x)}{t - x}$$

is a function defined at all points $t \in I$ such that $t \neq x$. Thus, it makes sense to ask whether the limit (4.1) exists. Suppose it does, and let G be another function from I to R^m which is differentiable at x. From the corresponding properties of limits, we see that $F + G$ is differentiable at x

and

$$(F + G)'(x) = F'(x) + G'(x). \tag{4.2}$$

If F and G are (complex) matrix-valued functions, then FG is differentiable at x and

$$(FG)'(x) = F(x)G'(x) + F'(x)G(x). \tag{4.3}$$

The product rule (4.3) is derived by adding and then subtracting a term to the difference quotient for FG:

$$\begin{aligned}\frac{(FG)(t) - (FG)(x)}{t - x} &= \frac{F(t)G(t) - F(x)G(x)}{t - x} \\ &= F(t)\frac{G(t) - G(x)}{t - x} + \frac{F(t) - F(x)}{t - x}G(x).\end{aligned}$$

As t approaches x, $F(t)$ approaches $F(x)$ and the difference quotients approach $G'(x)$ and $F'(x)$, respectively. The result follows from the rules for limits of sums and products.

In the last paragraph we used the fact that, if F is differentiable at x, then

$$\lim_{t \to x} [F(t) - F(x)] = 0.$$

This is apparent, since in (4.1) the denominator of the difference quotient is tending to 0 and the quotient could not approach a limit unless the numerator were tending to 0 as well. What this says is that *differentiability at x implies continuity at x*. The reader should be aware that the converse is false, as the example $f(x) = |x|$ at the point $x = 0$ shows.

If F is differentiable at each point of I, then $DF = F'$ is a function on I, called the **derivative** of F, and it makes sense to ask whether DF is differentiable. If it is, we denote its derivative by D^2F or F''. The successive derivatives (if they exist) are denoted by $F^{(n)} = D^nF$. We say that the function F is **of class C^k** provided the kth derivative D^kF exists and is continuous. If F has derivatives of all orders, then F is **of class C^∞**.

If I is a semi-closed interval, then, at the closed end, the derivative is only a one-sided derivative; e.g., from within $I = [a, b)$ approach to a is possible only from the right. At times, it is useful to have left- and right-hand derivatives at interior points of the interval. Let us discuss this briefly.

Suppose F is defined on a neighborhood of x. We consider the difference quotient

$$\frac{F(x) - F(t)}{x - t}, \qquad x < t < x + \delta.$$

If that quotient has a limit at x, it is the **right-hand derivative** of F at x:

$$(D^+F)(x) = \lim_{t \to x^+} \frac{F(x) - F(t)}{x - t}. \tag{4.4}$$

The **left-hand derivative** of F at x (if it exists) is defined similarly

$$(D^{-}F)(x) = \lim_{t \to x^{-}} \frac{F(x) - F(t)}{x - t}. \tag{4.5}$$

Either of the one-sided derivatives may fail to exist. They may both exist and be different, as the example $f(x) = |x|$ at $x = 0$ shows. Evidently, *$(DF)(x)$ exists if and only if $(D^{+}F)(x)$ exists, $(D^{-}F)(x)$ exists, and $(D^{+}F)(x) = (D^{-}F)(x)$.*

Any map

$$I \xrightarrow{F} R^{m}$$

is described by an m-tuple of real-valued functions, $F = (f_1, \ldots, f_m)$. From the corresponding property of limits, it is immediate that F is differentiable at x precisely when every f_j is differentiable at x. In this case $F'(x) = (f'_1(x), \ldots, f'_m(x))$. Therefore, the entire discussion of differentiation could be carried out using only real-valued functions; however, this is not an advisable way to proceed. It is much better to learn to handle differentiation of vector-valued functions directly, without using coordinates.

This seems an appropriate time at which to discuss the special features of differentiation of real-valued functions. If f is a differentiable function, then $f'(x)$ is the slope of the tangent line to the graph of f at the point $(x, f(x))$. It measures the rate at which f is increasing at x. Thus, the tangent line to the graph is horizontal at points where f has a maximum or minimum. This simple observation has a number of important consequences.

Theorem 1. *Let* f *be a real-valued function on an interval* (a, b). *Let* x ∈ (a, b) *and suppose that*

(i) f *has a local maximum* (*or local minimum*) *at* x;
(ii) f *is differentiable at* x.

Then $f'(x) = 0$.

Proof. To say that f has a local maximum at x means that there exists $\delta > 0$ such that

$$f(t) \leq f(x), \qquad |x - t| < \delta. \tag{4.6}$$

From (4.6), it is clear that $(D^{+}f)(x) \leq 0$, if $(D^{+}f)(x)$ exists. Similarly $(D^{-}f)(x) \geq 0$ if $(D^{-}f)(x)$ exists. Hence, if they both exist and are equal, their common value must be 0.

Corollary (***Mean Value Theorem***). *If* f *is continuous on* [a, b] *and differentiable on* (a, b), *there exists an* x ∈ (a, b) *such that*

$$f'(x) = \frac{f(b) - f(a)}{b - a}. \tag{4.7}$$

Proof. Let

$$g(x) = f(b) - f(x) + \frac{f(b) - f(a)}{b - a}(x - b).$$

Then g satisfies the same conditions as f and $g(a) = g(b) = 0$. The continuous function g attains its maximum and minimum on $[a, b]$. One of those must occur at a point x in the open interval (a, b). Then $g'(x) = 0$, which gives us (4.7).

Corollary. *Suppose* f *is differentiable on* (a, b). *Then* f *is an increasing function if and only if* $f' \geq 0$; f *is constant if and only if* $f' = 0$.

Proof. The mean value theorem for derivatives.

Corollary (Chain Rule). *Let* F, g *be differentiable functions such that the image of* g *is contained in the domain of* F*:*

$$(a, b) \xrightarrow{g} (c, d) \xrightarrow{F} R^m.$$

Then the composition $F \circ g$ *is differentiable and* $(F \circ g)' = g'(F' \circ g)$, *i.e.*, $(F \circ g)'(x) = g'(x)F'(g(x))$.

Proof. Fix $x \in (a, b)$. From the definition of derivative

$$F(y) - F(g(x)) = [y - g(x)]F'(g(x)) + (y - g(x))R(y, g(x))$$

where

$$\lim_{y \to g(x)} R(y, g(x)) = 0.$$

Thus,

$$\left|\frac{F(g(t)) - F(g(x))}{t - x} - \frac{g(t) - g(x)}{t - x}F'(g(x))\right| \leq \left|\frac{g(t) - g(x)}{t - x}\right| |R(g(t), g(x))|.$$

Corollary. *If* g *is a differentiable real-valued function on the interval* (a, b) *such that* $g'(x) \neq 0$ *for each* $x \in (a, b)$, *then* g *is* 1 : 1, *the inverse function* g^{-1} *is differentiable, and*

$$(Dg^{-1})(g(x)) = \frac{1}{g'(x)}$$

Proof. If we had $g(x) = g(t)$ with $x \neq t$, the mean value theorem would provide a point c between x and t with $g'(c) = 0$. Hence g is 1 : 1, that is, g is either strictly increasing or strictly decreasing. As we observed in Chapter 3 (Theorem 9) this means that the image of (a, b) is an open interval and g^{-1} is continuous. Thus,

$$(Dg^{-1})(g(x)) = \lim_{u \to g(x)} \frac{g^{-1}(u) - g^{-1}(g(x))}{u - g(x)} = \lim_{t \to x} \frac{t - x}{g(t) - g(x)} = \frac{1}{g'(x)}.$$

EXAMPLE 1. Let A be a fixed $k \times k$ matrix with complex entries. Let

$$F(t) = e^{tA}, \qquad t \in R.$$

Then F is a differentiable function. We have

$$\frac{F(t)-F(x)}{t-x}=\frac{e^{tA}-e^{xA}}{t-x}$$
$$=\frac{e^{(t-x)A}-I}{t-x}\cdot e^{xA}$$

What we must investigate is

$$\lim_{t\to 0}\frac{e^{tA}-I}{t}.$$

Now

$$e^{tA}-I=\sum_{n=1}^{\infty}\frac{1}{n!}t^nA^n$$

so that

$$t^{-1}(e^{tA}-I)-A=\sum_{n=2}^{\infty}\frac{1}{n!}t^{n-1}A^n$$
$$|t^{-1}(e^{tA}-I)-A|\leq\sum_{n=2}^{\infty}\frac{1}{n!}|t|^{n-1}|A|^n$$
$$\leq\frac{|A|}{2}\cdot\frac{|t||A|}{1-|t||A|}.$$

Now it is clear that F is differentiable and

$$F'(x)=Ae^{xA}=AF(x). \tag{4.8}$$

Since $F'=AF$, it follows that F is a function of class C^∞.

We remark that the differential equation in (4.8) determines the exponential function. If G is a differentiable function from R into $C^{k\times k}$ and if

$$G'(x)=AG(x)$$

where A is some fixed matrix, then $e^{-xA}G(x)$ has derivative 0. Thus,

$$G(x)=e^{xA}B$$

where $B=G(0)$. Notice that, if $G(x)=Be^{xA}$, then in order that $G'=AG$ be satisfied, B must commute with A, i.e., $AB=BA$.

The most important special case is, of course, the real-valued function

$$f(t)=e^t,\qquad t\in R$$

which satisfies $f'=f$. Since $f'(t)\neq 0$, the last corollary tells us that the inverse function f^{-1}, which we defined to be the log function, satisfies

$$(D\log)(e^t)=e^{-t}\quad\text{or}\quad(D\log)(x)=\frac{1}{x}.$$

EXAMPLE 2. Let F be the complex-valued function on the real line: $F(x)=e^{ix}$. According to Example 1, this is a function of class C^∞ and $F^{(n)}(x)=i^nF(x)$. Notice that, if we did not know about sines and cosines, We could define the sine and cosine functions by:

$$\begin{aligned}\sin x &= \operatorname{Im}(F(x))\\ &= \operatorname{Im}\sum_0^\infty \frac{1}{n!}(ix)^n\\ &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots\\ &= \sum_{n=0}^\infty \frac{(-1)^n}{(2n+1)!}x^{2n+1};\\ \cos x &= \operatorname{Re}(F(x))\\ &= \operatorname{Re}\sum_{n=0}^\infty \frac{1}{n!}(ix)^n\\ &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots\\ &= \sum_{n=0}^\infty \frac{(-1)^n}{(2n)!}x^{2n}.\end{aligned}$$

These functions are of class C^∞ and

$$\begin{aligned}D(\cos) &= -\sin\\ D(\sin) &= \cos.\end{aligned}$$

Each of these functions satisfies the differential equation

$$f'' + f = 0. \tag{4.9}$$

Furthermore, any f which satisfies (4.9) on an interval is a linear combination of sin and cos. The proof of this should be very familiar to every serious student of mathematics. For convenience, suppose f is real-valued and (4.9) holds on a neighborhood of 0. Let

$$g(x) = f(x) - f(0)\cos x - f'(0)\sin x.$$

Then $g'' + g = 0$ and $g(0) = g'(0) = 0$. Thus,

$$\begin{aligned}g'g'' + gg' &= 0\\ D((g')^2 + g^2) &= 0\\ (g')^2 + g^2 &= \text{constant}.\end{aligned}$$

Since $g(0) = g'(0) = 0$, the constant is 0; so,

$$\begin{aligned}(g')^2 + g^2 &= 0\\ g' = g &= 0 \qquad \text{(since } g \text{ is real-valued)}\\ f(x) &= f(0)\cos x + f'(0)\sin x.\end{aligned}$$

EXAMPLE 3. Let f be a real-valued function on an interval. We call f a **convex function** if

$$f[cx + (1-c)u] \leq cf(x) + (1-c)f(u) \tag{4.10}$$

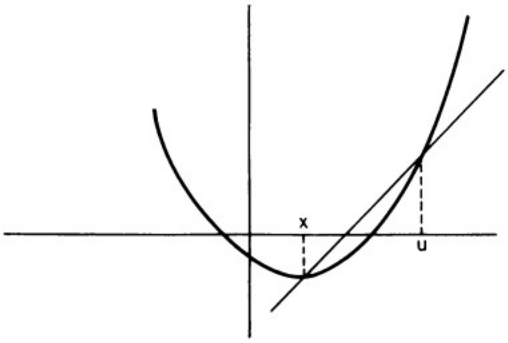

FIGURE 16

for all x, u in the domain and all c such that $0 \leq c \leq 1$. This says that, between x and u, the graph of f lies below the line joining the point $(x, f(x))$ to the point $(u, f(u))$. (See Figure 16.) Thus $f(x) = x^2$ and $f(x) = e^x$ define convex functions on the real line, while $f(x) = \sin x$ is a convex function on the interval $[-\pi, 0]$.

The defining property (4.1) can be rewritten in another useful way by observing that if $x < t < u$, then $t = cx + (1 - c)u$ where

$$c = \frac{u - t}{u - x}.$$

The condition (4.10) then says that

$$f(t) \leq \frac{u - t}{u - x} f(x) + \frac{t - x}{u - x} f(u)$$

which is more conveniently expressed as

$$\frac{f(t) - f(x)}{t - x} \leq \frac{f(u) - f(x)}{u - x}.$$

In other words, if $x < t < u$, the chord joining $(x, f(x))$ to $(u, f(u))$ has slope at least as great as the chord from $(x, f(x))$ to $(t, f(t))$. This suggests that for smooth functions convexity amounts to the fact that the derivative f' is an increasing (non-decreasing) function. This is the case, and we have left the proof to the Exercises.

As a matter of fact, the convexity property alone implies that f has a derivative at most points and that (where defined) f' is an increasing function. We sketch a proof. We know that if $x < t < u$, then

$$\frac{f(u) - f(x)}{u - x} \geq \frac{f(t) - f(x)}{t - x}.$$

Thus, unless x is the right end point of the interval of definition, the right-hand derivative $(D^+f)(x)$ exists and it is given by

$$(D^+f)(x) = \inf_{>x} \frac{f(t) - f(x)}{t - x}.$$

Similarly, unless x is the left end point, the left-hand derivative exists and

$$(D^-f)(x) = \sup_{t<x} \frac{f(t) - f(x)}{t - x}.$$

Since both $(D^+f)(x)$ and $(D^-f)(x)$ exist where they make sense, f is continuous at each such point x. More than that can be said. A little thought shows that

$$(D^+f)(t) \geq (D^-f)(t) \geq (D^+f)(x) \geq (D^-f)(x), \qquad t > x.$$

So, the left derivative D^-f is (as is D^+f) an increasing function. At any point where f fails to be differentiable, $(D^+f)(x) > (D^-f)(x)$ and so D^-f has a jump. The sum of all those jumps cannot exceed $(D^-f)(b) - (D^+f)(a)$, on the interval $[a, b]$. Therefore, there can be only a countable number of jumps. *Conclusion:* A convex function f on an open interval is continuous; it is differentiable except at a countable number of points; and the derivative Df (where defined) is an increasing function.

Exercises

1. Suppose that F is differentiable at a point x. Near x we approximate $F(t)$ by the affine function

$$A(t) = F(x) + (t - x)F'(x).$$

Show that the approximation satisfies

$$\lim_{t \to x} \frac{F(t) - A(t)}{t - x} = 0.$$

2. Let F be defined on an open interval about x. Prove that, if $(D^+F)(x)$ and $(D^-F)(x)$ both exist, then F is continuous at x.

3. Let

$$f_n(x) = \begin{cases} x^n \sin \dfrac{1}{x}, & x \neq 0 \\ 0, & x = 0. \end{cases}$$

Prove that f_1 is continuous but not differentiable at 0. Prove that f_2 is differentiable but not of class C^1. What is the corresponding fact about f_n?

4. Let $f_n(x) = x^n |x|$, $n = 0, 1, 2, \ldots$. Compute f_n' and find the largest integer k such that f_n is of class C^k.

5. Let F be a differentiable function from an interval I into R^m and suppose that $F' = 0$. Prove that F is constant.

6. Let f be the real-valued function on the real line

$$f(x) = \exp(-x^{-2}), \qquad x \neq 0$$
$$f(0) = 0.$$

Prove that f is a function of class C^∞ and that $f^{(n)}(0) = 0$ for every n.

7. Let f be a complex-valued function on an interval such that $f + f'' = 0$. Find the form of f by using the fact that $g = f - if'$ and $h = f + if'$ satisfy $g' = ig$, $h' = -ih$.

8. Let f, g be differentiable functions on R such that

$$\begin{aligned} f^2 + g^2 &= 1 \\ Df &= g \\ Dg &= -f \\ f(0) &= 0 \\ g(0) &= 1. \end{aligned}$$

Prove that f, g are what you think they are.

9. Let f be the Cantor function on $[0, 1]$, which was defined in Example 24 of Chapter 3. Prove that the derivative of f exists and is 0 at each point not in the Cantor set. Prove that f is not differentiable at any point of the Cantor set.

10. Show that the exponential function on the real line is a convex function and hence that

$$e^x \geq e^t(1 + x - t)$$

for all real numbers x and t.

11. Let f be a continuous real-valued function on an interval I which is differentiable at each interior point of I. Show that f is convex if and only if the derivative f' is an increasing (non-decreasing) function. (It is the "if" half of the result which we left unproved in the text.) What does this result tell you about the form of convex functions of class C^2?

12. If f and g are convex functions, is $\max(f, g)$ a convex function? What about $\min(f, g)$?

13. Let f be a real-valued function on an interval. The following are equivalent.

(a) f is convex.

(b) For each x, there is a line through the point $(x, f(x))$ such that the graph of f lies above that line.

(c) The set of all points in R^2 which lie above the graph of f is a convex set.

(d) f is continuous and

$$f\left(\frac{x + t}{2}\right) \leq \frac{1}{2}[f(x) + f(t)].$$

14. Let f be a continuous complex-valued function on the real line, which satisfies the functional equation

$$f(x + t) = f(x)f(t).$$

If $f \neq 0$, show that there is a complex number α such that $f(t) = e^{\alpha t}$. (You might ask first what kind of a function $\log|f|$ is.)

15. What about Exercise 14 for matrix-valued functions?

16. Let F be a differentiable function on the real line such that $|F'| \leq |F|$ and $F(0) = 0$. Prove that $F = 0$.

17. We know how to solve the differential equation $f'(x) = ax$. Suppose we have a system of differential equations

$$\begin{aligned} f_1' &= a_{11}f_1 + \cdots a_{1k}f_k \\ &\vdots \\ f_k' &= a_{k1}f_1 + \cdots + a_{kk}f_k \end{aligned}$$

involving k unknown functions f_i and k^2 known constants a_{ij}. Rewrite the system in the form

$$F' = AF$$

where A is the matrix of constants and

$$F = \begin{bmatrix} f_1 \\ \cdot \\ \cdot \\ \cdot \\ f_k \end{bmatrix}.$$

Prove that every solution for F has the form $F(t) = e^{tA}X_0$, where X_0 is some constant column matrix.

4.2. Integration on Intervals

Let F be a continuous function on a closed interval $I = [a, b]$. We want to define and discuss the integral of F over I:

$$\int_I F = \int_a^b F(x)\,dx.$$

In case F is a non-negative (and continuous) real-valued function, this integral is a *number* which measures the area under the graph of F and it is defined by a *process* known as integration, which gives a systematic method for approximating the area.

Since these ideas were first exploited systematically by Newton and Leibnitz 300 years ago, the range of applications of integration has been broadened enormously. This has been made possible in part by the modern definitions of the integral, in which the integration process is described in precise analytic terms, without appeal to any of its particular applications. In this chapter, we shall use a modified version of the formulation of integration given by Riemann. In Chapter 7 we will discuss the concept of integral given by Lebesgue, which is more general than Riemann's because it allows one to integrate many more functions. But we have no need of such generality at this point because our primary goal is to understand the basic theorems of calculus. Furthermore, it is important for every student of analysis to know that Riemann's process of integration converges for continuous functions.

In order to present the Riemann integral it is convenient to introduce two terms. A **partition** of the interval $[a, b]$ is any $(n + 1)$-tuple of the form

$$\begin{aligned} P &= (x_0, \dots, x_n) \\ a = x_0 &< x_1 < \cdots < x_n = b. \end{aligned} \tag{4.11}$$

Such a partition divides the interval into n subintervals, and the greatest of the lengths of these subintervals is called the **mesh** of the partition

$$\|P\| = \max_k (x_k - x_{k-1}) \qquad (1 \leq k \leq n). \tag{4.12}$$

Suppose F is a function from the interval $I = [a, b]$ into R^m. The integral of F over I, if it exists, will be a limit of sums

$$S(F, P, T) = \sum_{k=1}^{n} (x_k - x_{k-1})F(t_k) \tag{4.13}$$

where

$$P = (x_0, \dots, x_n)$$

is a partition of $[a, b]$ and

$$T = (t_1, \dots, t_n), \qquad t_k \in [x_{k-1}, x_k]$$

is a choice of points t_k, one from each subinterval defined by the partition. Such sums are frequently referred to as "Riemann sums", although we shall not use this terminology. The "limit" of these sums is to be taken as the mesh of the partition goes to 0; loosely speaking

$$\int_I F = \lim_{\|P\| \to 0} S(F, P, T).$$

There are so many (uncountably many) different partitions and different choices of points at which to evaluate the function that we will need to be careful how we describe the "limit" involved.

For any function which is integrable by the process we shall describe, it will always be possible (theoretically) to calculate the integral using any particular sequence of partitions $\{P_n\}$ for which the meshes tend to zero and any corresponding sequence of choices T_n. Any two such sequences will provide a sequence of vectors (approximating sums)

$$S_n = S(F, P_n, T_n)$$

which converges to the integral of F. For example, if

$$S_n = \sum_{k=1}^{n} \frac{b - a}{n} F\left(\frac{k(b - a)}{n}\right) \tag{4.14}$$

then for each integrable function F we shall have

$$\lim_n S_n = \int_a^b F(x)\, dx.$$

One might ask, therefore, why we do not use the sequence in (4.14) to define the integral, since it seems so much simpler than involving all sorts of partitions P and choices T, as in (4.13). There are two reasons, one practical and one theoretical.

(i) Even for very smooth functions F the particular sequence (4.14) may not provide the most convenient way of calculating the integral. We want to be able to use any sequence of partitions with meshes tending to 0 and know that the sums converge to the same limit. In other words, we want to be able to choose the sequence most convenient for each particular function F.

(ii) We need to deal with integrals of some functions which are not smooth and, therefore, we must be precise about the class of functions to which the integration process applies. Simple examples show that, were we to have the definition of this class of functions depend upon the values of the function at a particular sequence of points such as $(k/n)(b - a)$, we would generate chaos.

EXAMPLE 4. Let f be the real-valued function defined by

$$f(x) = \begin{cases} 0, & \text{if } x \text{ is rational} \\ 1, & \text{if } x \text{ is irrational.} \end{cases}$$

Suppose $a < b$ and let $P = (x_0, x_1, \ldots, x_n)$ be any partition of $[a, b]$. Each interval $[x_{k-1}, x_k]$ has positive length and, accordingly, it contains both rational and irrational points. For each k, choose a rational point $t_k \in [x_{k-1}, x_k]$ and an irrational point $t'_k \in [x_{k-1}, x_k]$. From the definition of f we have $f(t_k) = 0$ and $f(t'_k) = 1$ for each k. Accordingly,

$$S(F, P, T) = 0$$

$$S(F, P, T') = b - a.$$

Thus we can choose sequences (P_n, T_n) with $\|P_n\| \to 0$ for which $S(F, P_n, T_n) = 0$ and we can choose other sequences of this type for which $S(F, P_n, T_n) = b - a$. This leaves us a bit perplexed as to what the integral of f ought to be. In order to avoid this confusion, we shall define the class of functions which are "integrable" in such a way as to exclude this function f.

Definition. *Let* F *be a function from the interval* [a, b] *into* R^m. *We say that* F *is* **Riemann-integrable** *if there exists a vector* S *in* R^m *with this property: For each* $\epsilon > 0$ *there exists* $\delta > 0$ *such that if* P *is any partition of* [a, b] *with* $\|P\| \leq \delta$ *and if* T *is any choice of points in the subintervals defined by* P, *then*

$$|S - S(F, P, T)| < \epsilon.$$

Although this definition sounds somewhat complicated, it expresses a rather simple idea: The function F will be called Riemann-integrable if there exists a vector S (which will be the integral of F) to which the approximating sums $S(F, P, T)$ converge as $\|P\| \to 0$. Converge in which sense? In the sense that $S(F, P, T)$ will be close to S provided only that $\|P\|$ is small, and with no condition on T, the choice of points in the subintervals of P.

It should be clear that this basic definition can be reformulated as follows. The function F is Riemann-integrable if (and only if) there exists a vector S in R^m such that

$$\lim_n S(F, P_n, T_n) = S$$

for every sequence of partitions P_n with $\|P_n\| \longrightarrow 0$ and any corresponding choice(s) T_n of points in the subintervals defined by P_n. Thus the point of Example 4 can be described by saying that the function f which occurs there is not Riemann-integrable.

If F is Riemann-integrable, the vector S which occurs in the definition is obviously unique. It is called the **integral of F over the interval $I = [a, b]$** and is denoted by $\int_I F$ or $\int_a^b F(x)\,dx$.

A word about this notation. Part of the power of calculus as a tool derives from the very useful notation which Leibnitz invented for derivatives and integrals. For real-valued functions on an interval $[a, b]$ the integral was denoted by

$$\int_a^b f(x)\,dx$$

as a reminder of the process of integration

$$\begin{aligned}\int_a^b f(x)\,dx &= \lim_{\|P\|\to 0} \sum_{k=1}^n f(x_k)(x_k - x_{k-1})\\ &= \lim_{\max \Delta x_k \to 0} \sum_k f(x_k)\,\Delta x_k\\ &= \lim_{\Delta x \to 0} \sum f(x)\,\Delta x.\end{aligned}$$

We have not used the $\Delta x_k = x_k - x_{k-1}$ notation, in part because for vector-valued functions it would then be natural to write

$$\lim_{\|P\|\to 0} \sum_{k=1}^n \Delta x_k F(x_k) = \int_a^b dx\, F(x)$$

which is potentially confusing.

We want to direct our efforts immediately to proving that various well-behaved functions (e.g., continuous functions) are Riemann-integrable. In other words, we want to show that the process of integration "converges" for various well-behaved functions. We need a criterion for convergence analogous to the Cauchy criterion for sequences, i.e., a criterion which does not explicitly involve the limit vector.

If F is any function from $[a, b]$ into R^m, then, for each $\delta > 0$, we define the set $\Sigma_\delta = \Sigma_\delta(F)$ as follows. We consider all partitions $P = (x_0, \ldots, x_n)$ of the interval $[a, b]$ such that $\|P\| \leq \delta$ and all possible corresponding choices T of points $t_k \in [x_{k-1}, x_k]$. Each such pair (P, T) determines a sum

$$S(F, P, T) = \sum_{k=1}^n (x_k - x_{k-1})F(t_k)$$

and $\mathbf{\Sigma}_\delta(F)$ is defined to be the set of all such sums:

$$\sum_{\delta}(F) = \{S(F, P, T); \|P\| \leq \delta, t_k \in [x_{k-1}, x_k]\}. \tag{4.15}$$

If F is Riemann-integrable then, for small δ, all the sums $S(F, P, T)$ with $\|P\| \leq \delta$, i.e., all the vectors in $\mathbf{\Sigma}_\delta(F)$, will be close to the integral of F. Thus, when δ is small, $\mathbf{\Sigma}_\delta$ must be a set of small diameter clustered near the integral of F. And the integral is approximable by vectors in $\mathbf{\Sigma}_\delta$, that is, the integral lies in the closure $\bar{\mathbf{\Sigma}}_\delta$. What happens to the set $\mathbf{\Sigma}_\delta$ as we decrease δ? If $0 < \eta < \delta$, then it is immediate from the definition (4.15) that $\mathbf{\Sigma}_\eta \subset \mathbf{\Sigma}_\delta$. This is true for any F. For any integrable F it must be that, as δ tends to 0, the closed sets $\bar{\mathbf{\Sigma}}_\delta$ close down on the single vector $\int_I F$. In other words, the diameters of the sets $\mathbf{\Sigma}_\delta$ converge to 0 and $\int_I F$ is the single vector contained in the intersection

$$\bigcap_{\delta>0} \bar{\mathbf{\Sigma}}_\delta.$$

Theorem 2. *Let* F *be a function from the interval* [a, b] *into* R^m, *and define the set* $\mathbf{\Sigma}_\delta(\mathrm{F})$ *by* (4.15). *Then* F *is Riemann-integrable if and only if*

$$\lim_{\delta\to 0} \operatorname{diam} \mathbf{\Sigma}_\delta = 0. \tag{4.16}$$

Proof. It is the "if" statement which remains to be proved. Assume (4.16) and consider the intersection of the closures of the sets $\mathbf{\Sigma}_\delta$:

$$\bigcap_{\delta>0} \bar{\mathbf{\Sigma}}_\delta.$$

By the condition on the diameters, there cannot be more than one vector in this intersection. Furthermore, the intersection is nonempty for this reason. We know that $\mathbf{\Sigma}_\eta \subset \mathbf{\Sigma}_\delta$ if $\eta < \delta$. Hence the closed sets $\bar{\mathbf{\Sigma}}_\delta$ are nested:

$$\bar{\mathbf{\Sigma}}_\eta \subset \bar{\mathbf{\Sigma}}_\delta, \qquad \eta < \delta.$$

Furthermore, since the diameters of these sets tend to 0, they are bounded (sets) for all sufficiently small δ. Thus their intersection is non-empty. (See Theorem 9 of Chapter 2.) Let S be the single vector in the intersection. If $\epsilon > 0$, choose $\delta > 0$ such that $\operatorname{diam} \mathbf{\Sigma}_\delta(F) < \epsilon$. If P is any partition of $[a, b]$ with $\|P\| \leq \delta$ and if T is any choice of points in the subintervals defined by P, then the sum $S(F, P, T)$ is in the set $\mathbf{\Sigma}_\delta(F)$ and the vector S is in $\bar{\mathbf{\Sigma}}_\delta(F)$. Since $\operatorname{diam} \mathbf{\Sigma}_\delta(F) < \epsilon$,

$$|S - S(F, P, T)| < \epsilon.$$

Thus F is Riemann-integrable (and $S = \int_I F$).

In order to use the diameter criterion (4.16) to show that a given function F is Riemann-integrable, one needs to estimate, for a given δ, what the maximum distance is between any two approximating sums, formed by partitions of mesh at most δ. To carry out such estimates, we need to see what happens when we pass from a given partition to a more refined one.

Let $P = (x_0, \ldots, x_n)$ and $Q = (y_0, \ldots, y_p)$ be partitions of the interval $[a, b]$. We say that Q is a **refinement** of P if $\{x_0, \ldots, x_n\}$ is a subset of $\{y_0, \ldots, y_p\}$. In other words Q is a refinement of P if Q can be obtained by starting with the points $x_0 < x_1 < \cdots < x_n$ and interposing some $p - n$ points between them to form a finer subdivision. The fact that Q is a refinement of P guarantees among other things that $\|Q\| \leq \|P\|$, but note that it is a much stronger condition than the (mere) fact that Q has a smaller mesh than does P.

Lemma. *Let* F *be a function from* $[\mathrm{a}, \mathrm{b}]$ *into* R^{m}. *Let* $\mathrm{P} = (\mathrm{x}_0, \ldots, \mathrm{x}_{\mathrm{n}})$ *and* $\mathrm{Q} = (\mathrm{y}_0, \ldots, \mathrm{y}_{\mathrm{p}})$ *be partitions of* $[\mathrm{a}, \mathrm{b}]$ *such that* Q *is a refinement of* P. *If* T *is any choice of points* $\mathrm{t}_{\mathrm{k}} \in [\mathrm{x}_{\mathrm{k}-1}, \mathrm{x}_{\mathrm{k}}]$, $\mathrm{k} = 1, \ldots, \mathrm{n}$, *there exists points* $\mathrm{u}_1, \ldots, \mathrm{u}_{\mathrm{p}}$ *in* $[\mathrm{a}, \mathrm{b}]$ *such that*

(i) $\mathrm{y}_{\mathrm{j}-1} - \|\mathrm{P}\| \leq \mathrm{u}_{\mathrm{j}} \leq \mathrm{y}_{\mathrm{j}} + \|\mathrm{P}\|, \qquad \mathrm{j} = 1, \ldots, \mathrm{p};$

(ii) $\sum_{\mathrm{k}=1}^{\mathrm{n}} (\mathrm{x}_{\mathrm{k}} - \mathrm{x}_{\mathrm{k}-1})\mathrm{F}(\mathrm{t}_{\mathrm{k}}) = \sum_{\mathrm{j}=1}^{\mathrm{p}} (\mathrm{y}_{\mathrm{j}} - \mathrm{y}_{\mathrm{j}-1})\mathrm{F}(\mathrm{u}_{\mathrm{j}}).$

Proof. Let j be an index, $1 \leq j \leq p$. We shall define the corresponding point u_j. Since Q is a refinement of P, none of the points x_i lies in the open interval (y_{j-1}, y_j). Therefore the closed interval $[y_{j-1}, y_j]$ must lie wholly within one of the intervals $[x_{i-1}, x_i]$. In fact, since the interiors of the different intervals $[x_{i-1}, x_i]$ do not overlap, there is precisely one index i, $1 \leq i \leq n$, such that $[y_{j-1}, y_j] \subset [x_{i-1}, x_i]$. Call this index $i(j)$ and define $u_j = t_{i(j)}$.

Now observe that

$$\sum_{j=1}^{p} (y_j - y_{j-1})F(u_j) = \sum_{k=1}^{n} \sum_{i(j)=k} (y_j - y_{j-1})F(u_j)$$
$$= \sum_{k=1}^{n} \sum_{i(j)=k} (y_j - y_{j-1})F(t_k).$$

For a particular k, the indices j such that $i(j) = k$ are those for which $[y_{j-1}, y_j] \subset [x_{k-1}, x_k]$. Since Q is a refinement of P, these intervals form a partition of $[x_{k-1}, x_k]$, that is,

$$\sum_{i(j)=k} (y_j - y_{j-1}) = x_k - x_{k-1}.$$

This proves the lemma.

Theorem 3. *If* F *is a continuous function from the closed interval* $[\mathrm{a}, \mathrm{b}]$ *into* R^{m}, *then* F *is Riemann-integrable.*

Proof. The idea of the proof is to use the uniform continuity of F to estimate the diameters of the sets $\Sigma_\delta(F)$. Let $\delta > 0$. Let P, P' be partitions of $[a, b]$ of mesh at most δ and let $T = (t_1, \ldots, t_n)$, $T' = (t'_1, \ldots, t'_{n'})$

be choices of points in the subintervals determined by P and P', respectively. We want to estimate the distance from $S(F, P, T)$ to $S(F, P', T')$. To do this, we need to rewrite them as sums over the same partition. Accordingly, let $Q = (y_0, \dots, y_p)$ be the partition which is made up of the *distinct* points in the sequence $x_0, \dots, x_n, x'_0, \dots, x'_n$, arranged in increasing order. Then Q is a refinement of P and also a refinement of P'. By the preceding lemma, there exist points $u_1, \dots, u_p$ and $u'_1, \dots, u'_p$ in $[a, b]$ such that

$$S(F, P, T) = S(F, Q, U)$$
$$S(F, P', T') = S(F, Q, U').$$

We do not know that u_j and u'_j lie in the interval $[y_{j-1}, y_j]$. But, from the lemma, we do know that both u_j and u'_j lie in the interval $[y_{j-1} - \delta, y_j + \delta]$. Thus,

$$S(F, P, T) - S(F, P', T') = \sum_{j=1}^{p} (y_j - y_{j-1})[F(u_j) - F(u'_j)]$$

and $|u_j - u'_j| \leq 2\delta$, $j = 1, \dots, p$. Let

$$M = \max_j |F(u_j) - F(u'_j)|$$

and we have

$$\begin{aligned} |S(F, P, T) - S(F, P', T')| &\leq M \sum_{j=1}^{p} (y_j - y_{j-1}) \\ &= M(b - a). \end{aligned} \tag{4.17}$$

The idea now is to show that M is small whenever $\|P\|$ and $\|P'\|$ are small. The critical observation is that F is continuous on the closed interval $[a, b]$, and is therefore uniformly continuous, (Theorem 12 of Chapter 3). This guarantees that, when δ is small, each of the numbers $|F(u_j) - F(u'_j)|$ will be small, because u_j, u'_j are not more than 2δ apart. To be precise, if ω is the modulus of continuity of F (see (3.17)):

$$\omega(F, \eta) = \sup_{|x-t| \leq \eta} |F(x) - F(t)|$$

then (4.17) tells us that

$$|S(F, P, T) - S(F, P', T')| \leq (b - a)\omega(F, 2\delta).$$

Thus,

$$\operatorname{diam} \Sigma_\delta(F) \leq (b - a)\omega(F, 2\delta).$$

Now

$$\lim_{\eta \to 0} \omega(F, \eta) = 0$$

because this is what it means to say that F is uniformly continuous. We conclude that F is Riemann-integrable.

Corollary. *If* F *is a continuous function on the interval* [a, b], *if* $\mathrm{P} = (\mathrm{x}_0, \dots, \mathrm{x}_n)$ *is a partition of* [a, b] *and* $\mathrm{T} = (\mathrm{t}_1, \dots, \mathrm{t}_n)$ *any choice of points* $\mathrm{t}_k \in [\mathrm{x}_{k-1}, \mathrm{x}_k]$ *then*

$$\left|\int_a^b F(x)\,dx - \sum_{k=1}^{n} (x_k - x_{k-1})F(t_k)\right| \leq (b - a)\omega(F, 2\,||\,P\,||)$$

where ω *is the modulus of continuity of* F.

Proof. Let $\delta = ||\,P\,||$. In the proof of the theorem we verified that

$$\operatorname{diam} \Sigma_\delta(F) \leq (b, a)\omega(F, 2\delta).$$

Since the integral of F is in the closure of $\Sigma_\delta(F)$, its distance from the sum $S(F, P, T)$ cannot exceed $(b - a)\omega(F, 2\delta)$.

Exercises

1. Let F be a Riemann-integrable function on the closed interval I. Show that, if c is any real number, then cF is Riemann-integrable and

$$\int_I cF = c\int_I F.$$

(Work directly with $c\int F - S(cF, P, T)$ rather than $\Sigma_\delta(cF)$.)

2. In Exercise 1, if F is a complex-valued function and c is a complex number, is the same assertion valid? (Of course, we regard a complex-valued function as a map into R^2.)

3. Let F and G be functions from the closed interval I into R^m. Show that, if F and G are Riemann-integrable, then $F + G$ is and

$$\int_I (F + G) = \int_I F + \int_I G.$$

4. If f and g are real-valued Riemann-integrable functions on the closed interval I such that $f \geq g$, then

$$\int_I f \geq \int_I g.$$

5. If f is a non-negative continuous function on $[a, b]$ and

$$\int_a^b f(x)\,dx = 0$$

then $f = 0$.

6. Use the result of Exercise 5 to prove the following. If f is a continuous real-valued function on $[a, b]$ and

$$\int_a^b f(x)\,dx = 0$$

then there exists a point c, $a \leq c \leq b$, such that $f(c) = 0$.

7. Use the result of Exercise 6 to prove the *first mean value theorem of integral calculus*: If f is a continuous real-valued function on the interval $[a, b]$, there exists a point c, $a \leq c \leq b$, such that

$$\int_a^b f(x)\,dx = (b - a)f(c).$$

)*+,)))-) #(! * 0 9 !
5 61 5 % 7 5!

7)*+,)))-) #(C 7 <
5 2 ((#1 7 ! = / 0 5!
' #; 6 & 2 / ! . D

is the correct formula, one has only to show that the covariant derivative of the induced metric is zero and that the torsion vanishes. This follows because

$$h_{ab\|c} = (g_{ef} \mp \bar{n}_e \bar{n}_f)_{;g} h^e{}_a h^f{}_b h^g{}_c = 0,$$

and

$$f_{\|ab} = h^e{}_a h^g{}_b f_{;eg} = h^e{}_a h^g{}_b f_{;ge} = f_{\|ba}.$$

The curvature tensor $R'^a{}_{bcd}$ of the induced metric **h** can be related to the curvature tensor $R^a{}_{bcd}$ on $\theta(\mathscr{S})$ and the second fundamental form $\boldsymbol{\chi}$ as follows. If $\mathbf{Y} \in H$ is a vector field on $\theta(\mathscr{S})$, then

$$R'^a{}_{bcd} Y^b = Y^a{}_{\|dc} - Y^a{}_{\|cd}.$$

Now

$$\begin{aligned} Y^a{}_{\|dc} &= (Y^a{}_{\|d})_{\|c} = (\bar{Y}^e{}_{;f} h^g{}_e h^f{}_i)_{;k} h^a{}_g h^i{}_d h^k{}_c \\ &= \bar{Y}^e{}_{;fk} h^a{}_e h^f{}_d h^k{}_c \mp \bar{Y}^e{}_{;f} \bar{n}_e \bar{n}^g{}_{;k} h^f{}_d h^a{}_g h^k{}_c \mp \bar{Y}^e{}_{;f} \bar{n}^f \bar{n}_{i;k} h^a{}_e h^i{}_d h^k{}_c \end{aligned}$$

and

$$\bar{Y}^e{}_{;f} \bar{n}_e h^f{}_d = (\bar{Y}^e \bar{n}_e)_{;f} h^f{}_d - \bar{Y}^e \bar{n}_{e;f} h^f{}_d = -\bar{Y}^e \bar{n}_{e;f} h^f{}_d,$$

since $\bar{Y}^e \bar{n}_e = 0$ on $\theta(\mathscr{S})$, therefore

$$R'^a{}_{bcd} Y^b = (R^e{}_{bkf} h^a{}_e h^k{}_c h^f{}_d \pm \chi_{bd} \chi^a{}_c \mp \chi_{bc} \chi^a{}_d) Y^b.$$

Since this holds for all $\mathbf{Y} \in H$,

$$R'^a{}_{bcd} = R^e{}_{fgh} h^a{}_e h^f{}_b h^g{}_c h^h{}_d \pm \chi^a{}_c \chi_{bd} \mp \chi^a{}_d \chi_{bc}. \tag{2.34}$$

This is known as Gauss' equation.

Contracting this equation on a and c and multiplying by h^{bd}, one obtains the curvature scalar R' of the induced metric:

$$R' = R \mp 2R_{ab} n^a n^b \pm (\chi^a{}_a)^2 \mp \chi^{ab} \chi_{ab}. \tag{2.35}$$

One can derive another relation between the second fundamental form and the curvature tensor $R^a{}_{bcd}$ on $\theta(\mathscr{S})$ by subtracting the expressions

$$(\chi^a{}_a)_{\|b} = (\bar{n}^a{}_{;d} h^d{}_a)_{;e} h^e{}_b$$

and

$$(\chi^a{}_b)_{\|a} = (\bar{n}^c{}_{;d} h^a{}_c h^d{}_e)_{;f} h^f{}_a h^e{}_b,$$

finding

$$\chi^a{}_{b\|a} - \chi^a{}_{a\|b} = R_{ef} n^f h^e{}_b. \tag{2.36}$$

This is known as Codacci's equation.

2.8 The volume element and Gauss' theorem

If $\{\mathbf{E}^a\}$ is a basis of one-forms, one can form from it the n-form

$$\boldsymbol{\epsilon} = n!\mathbf{E}^1 \wedge \mathbf{E}^2 \wedge \ldots \wedge \mathbf{E}^n.$$

If $\{\mathbf{E}^{a'}\}$, related to $\{\mathbf{E}^{a}\}$ by $\mathbf{E}^{a'} = \Phi^{a'}{}_{a}\mathbf{E}^{a}$, is another basis, the n-form $\boldsymbol{\epsilon}'$ defined by this basis will be related to $\boldsymbol{\epsilon}$ by

$$\boldsymbol{\epsilon}' = \det(\Phi^{a'}{}_{a})\,\boldsymbol{\epsilon},$$

so this form is not unique. However, one can use the existence of the metric to define (in a given basis) the form

$$\boldsymbol{\eta} = |g|^{\frac{1}{2}}\,\boldsymbol{\epsilon}$$

where $g \equiv \det(g_{ab})$. This form has components

$$\eta_{ab\ldots d} = n!\,|g|^{\frac{1}{2}}\,\delta^{1}{}_{[a}\delta^{2}{}_{b}\ldots\delta^{n}{}_{d]}.$$

The transformation law for g will just cancel the determinant, $\det(\Phi^{a'}{}_{a})$, provided that $\det(\Phi^{a'}{}_{a}) > 0$. Therefore if $\mathscr{M}$ is orientable the n-forms $\boldsymbol{\eta}$ defined by coordinate bases of an oriented atlas will be identical, i.e. given an orientation of $\mathscr{M}$, one can define a unique n-form field $\boldsymbol{\eta}$, the *canonical n-form*, on $\mathscr{M}$.

The contravariant antisymmetric tensor

$$\eta^{ab\ldots d} = g^{ae}g^{bf}\ldots g^{dh}\eta_{ef\ldots h}$$

has components

$$\eta^{ab\ldots d} = (-)^{\frac{1}{2}(n-s)}n!\,|g|^{\frac{1}{2}}\,\delta^{[a}{}_{1}\delta^{b}{}_{2}\ldots\delta^{d]}{}_{n},$$

where s is the signature of $\mathbf{g}$ (so $\frac{1}{2}(n-s)$ is the number of negative eigenvalues of the matrix of metric components (g_{ab})). Therefore these tensors satisfy the relations

$$\eta^{ab\ldots d}\eta_{ef\ldots h} = (-)^{\frac{1}{2}(n-s)}n!\,\delta^{a}{}_{[e}\delta^{b}{}_{f}\ldots\delta^{d}{}_{h]}. \tag{2.37}$$

The Christoffel relations imply that the covariant derivatives of $\eta_{ab\ldots d}$ and $\eta^{ab\ldots d}$ with respect to the connection defined by the metric vanish, i.e.

$$\eta^{ab\ldots d}{}_{;e} = 0 = \eta_{ab\ldots d;e}.$$

Using the canonical n-form, one can define the volume (with respect to the metric $\mathbf{g}$) of an n-dimensional submanifold $\mathscr{U}$ as $\dfrac{1}{n!}\displaystyle\int_{\mathscr{U}}\boldsymbol{\eta}$. Thus $\boldsymbol{\eta}$ can be regarded as a positive definite volume measure on $\mathscr{M}$. We shall often use it in this sense, and shall denote it by $\mathrm{d}v$. Note that d is not meant to represent the exterior differential operator here; $\mathrm{d}v$ is simply a measure on $\mathscr{M}$. If f is a function on $\mathscr{M}$, one can define its integral over $\mathscr{U}$ with respect to this volume measure as

$$\int_{\mathscr{U}} f\,\mathrm{d}v = \frac{1}{n!}\int_{\mathscr{U}} f\,\boldsymbol{\eta}.$$

With respect to local oriented coordinates $\{x^a\}$, this can be expressed as the multiple integral

$$\int_{\mathscr{U}} f|g|^{\frac{1}{2}}\,\mathrm{d}x^1\,\mathrm{d}x^2\ldots\mathrm{d}x^n,$$

which is invariant under a change of coordinates.

If $\mathbf{X}$ is a vector field on $\mathscr{M}$, its contraction with $\boldsymbol{\eta}$ will be an $(n-1)$-form field $\mathbf{X}\,.\,\boldsymbol{\eta}$, where

$$(\mathbf{X}\,.\,\boldsymbol{\eta})_{b\ldots d} = X^a\eta_{ab\ldots d}.$$

This $(n-1)$-form may be integrated over any $(n-1)$-dimensional compact orientable submanifold $\mathscr{V}$. We write this integral as

$$\int_{\mathscr{V}} X^a\mathrm{d}\sigma_a = \frac{1}{(n-1)!}\int_{\mathscr{V}}\mathbf{X}\,.\,\boldsymbol{\eta},$$

where the canonical form $\boldsymbol{\eta}$ is regarded as defining a measure-valued form $\mathrm{d}\sigma_a$ on the submanifold $\mathscr{V}$. If the orientation of $\mathscr{V}$ is given by the direction of the normal form n_a, then $\mathrm{d}\sigma_a$ can be expressed as $n_a\mathrm{d}\sigma$ where $\mathrm{d}\sigma$ is a positive definite volume measure on the submanifold $\mathscr{V}$. The volume measure $\mathrm{d}\sigma$ is not unique unless the normal n_a is normalized. If n_a is normalized to unit magnitude in a metric $\mathbf{g}$ on $\mathscr{M}$, i.e. $n_a n_b g^{ab} = \pm 1$, then $\mathrm{d}\sigma$ is equal to the volume measure on $\mathscr{V}$ defined by the induced metric on $\mathscr{V}$ (to see this, simply choose an orthonormal basis with $n_a g^{ab}$ as one of the basis vectors).

Using the canonical form, one can derive Gauss' formula from Stokes' theorem: for any compact n-dimensional submanifold $\mathscr{U}$ of $\mathscr{M}$,

$$\int_{\partial\mathscr{U}} X^a\mathrm{d}\sigma_a = \frac{1}{(n-1)!}\int_{\partial\mathscr{U}}\mathbf{X}\,.\,\boldsymbol{\eta} = \frac{1}{(n-1)!}\int_{\mathscr{U}}\mathrm{d}(\mathbf{X}\,.\,\boldsymbol{\eta}).$$

But

$$\begin{aligned}(\mathrm{d}(\mathbf{X}\,.\,\boldsymbol{\eta}))_{a\ldots de} &= (-)^{n-1}(X^g\eta_{g[a\ldots d})_{;e]}\\ &= (-)^{n-1}\delta^s{}_{[a}\ldots\delta^t{}_d\delta^u{}_{e]}\eta_{gs\ldots t}X^g{}_{;u}\\ &= (-)^{(n-1)-\frac{1}{2}(n-s)}\frac{1}{n!}\eta^{s\ldots tu}\eta_{a\ldots de}\eta_{gs\ldots t}X^g{}_{;u}\\ &= \eta_{a\ldots de}\delta^s{}_{[s}\ldots\delta^t{}_t\delta^u{}_{g]}X^g{}_{;u}\\ &= n^{-1}\eta_{a\ldots de}X^g{}_{;g},\end{aligned}$$

on using relation (2.37) twice. Therefore

$$\int_{\partial\mathscr{U}} X^a\,\mathrm{d}\sigma_a = \int_{\mathscr{U}} X^g{}_{;g}\,\mathrm{d}v$$

holds for any vector field $\mathbf{X}$; this is Gauss' theorem. Note that the orientation on $\mathscr{U}$ for which this theorem is valid is that given by the normal form $\boldsymbol{\eta}$ such that $\langle \mathbf{n}, \mathbf{X} \rangle$ is positive if $\mathbf{X}$ is a vector which points out of $\mathscr{U}$. If the metric $\mathbf{g}$ is such that $g^{ab}n_a n_b$ is negative, the vector $g^{ab}n_b$ will point into $\mathscr{U}$.

2.9 Fibre bundles

Some of the geometrical properties of a manifold $\mathscr{M}$ can be most easily examined by constructing a manifold called a fibre bundle, which is locally a direct product of $\mathscr{M}$ and a suitable space. In this section we shall give the definition of a fibre bundle and shall consider four examples that will be used later: the tangent bundle $T(\mathscr{M})$, the tensor bundle $T^r_s(\mathscr{M})$, the bundle of linear frames or bases $L(\mathscr{M})$, and the bundle of orthonormal frames $O(\mathscr{M})$.

A C^k *bundle* over a C^s ($s \geqslant k$) manifold $\mathscr{M}$ is a C^k manifold $\mathscr{E}$ and a C^k surjective map $\pi: \mathscr{E} \rightarrow \mathscr{M}$. The manifold $\mathscr{E}$ is called the total space, $\mathscr{M}$ is called the base space and π the projection. Where no confusion can arise, we will denote the bundle simply by $\mathscr{E}$. In general, the inverse image $\pi^{-1}(p)$ of a point $p \in \mathscr{M}$ need not be homeomorphic to $\pi^{-1}(q)$ for another point $q \in \mathscr{M}$. The simplest example of a bundle is a *product bundle* $(\mathscr{M} \times \mathscr{A}, \mathscr{M}, \pi)$ where $\mathscr{A}$ is some manifold and the projection π is defined by $\pi(p, v) = p$ for all $p \in \mathscr{M}, v \in \mathscr{A}$. For example, if one chooses $\mathscr{M}$ as the circle S^1 and $\mathscr{A}$ as the real line R^1, one constructs the cylinder C^2 as a product bundle over S^1.

A bundle which is locally a product bundle is called a fibre bundle. Thus a bundle is a *fibre bundle* with fibre $\mathscr{F}$ if there exists a neighbourhood $\mathscr{U}$ of each point q of $\mathscr{M}$ such that $\pi^{-1}(\mathscr{U})$ is isomorphic with $\mathscr{U} \times \mathscr{F}$, in the sense that for each point $p \in \mathscr{U}$ there is a diffeomorphism ϕ_p of $\pi^{-1}(p)$ onto $\mathscr{F}$ such that the map ψ defined by $\psi(u) = (\pi(u), \phi_{\pi(u)})$ is a diffeomorphism $\psi: \pi^{-1}(\mathscr{U}) \rightarrow \mathscr{U} \times \mathscr{F}$. Since $\mathscr{M}$ is paracompact, we can choose a locally finite covering of $\mathscr{M}$ by such open sets $\mathscr{U}_\alpha$. If $\mathscr{U}_\alpha$ and $\mathscr{U}_\beta$ are two members of such a covering, the map

$$(\phi_{\alpha, p}) \circ (\phi_{\beta, p}{}^{-1})$$

is a diffeomorphism of $\mathscr{F}$ onto itself for each $p \in (\mathscr{U}_\alpha \cap \mathscr{U}_\beta)$. The inverse images $\pi^{-1}(p)$ of points $p \in \mathscr{M}$ are therefore necessarily all diffeomorphic to $\mathscr{F}$ (and so to each other). For example, the Möbius strip is a fibre bundle over S^1 with fibre R^1; we need two open sets $\mathscr{U}_1$, $\mathscr{U}_2$

)*+,)))-) #(. ! !
0 70 . 5!

)*+,)))-) #(! ! 70
E / 801 + F 7 5! ' $ #&

4

Analysis of PN Guided Missile Systems in the Frequency Domain

4.1 Introduction

The difficulty in the analysis of differential equations with time-varying coefficients describing linearized models of guidance systems does not allow researchers to obtain analytical expressions for miss distance that can be effectively utilized in practice. As mentioned in [7], "the disadvantage of the single time constant representation of a missile guidance system is that the miss distance can be seriously underestimated." The same remark can also be applied to the binomial representation.

The method of adjoints used in the previous section can be presented in the integral form as the impulse response of the adjoint system, i.e., its reaction to a unit impulse function. In the case of a single time constant guidance system, the transfer function between the target acceleration and the guidance system miss distance can be obtained from (3.21). In control theory, the method of transfer functions, as input–output characteristics of linear systems, is a foundation of frequency methods, i.e., analysis of systems in the frequency domain. Analogous to the unit step signal in the time domain, the unit sinusoidal input signal is the standard test signal in the frequency domain. The response of the system to a changing frequency is considered. The frequency response is defined as the steady-state response of the system to a sinusoidal input signal. The frequency approach is very popular among engineers because the design of a system in the frequency domain provides the designer with the control of the bandwidth of a system. It is very physical and enables researchers and designers to build realistic models and make justifiable simplifications.

The below analysis of the linearized PN guidance system models is based on the frequency response of the linearized guidance system, which corresponds to the miss distance due to a weaving target, i.e., due to a sinusoidal acceleration. As shown later, the miss due to a step target maneuver, i.e., the miss step response, can also be obtained from the frequency characteristics of the system under consideration.

The block diagram of an interceptor's main subsystems is given in Figure 4.1.

The seeker provides a guidance system with target information, which, together with the information from onboard sensors, is necessary to generate a guidance law. The guidance system generates acceleration commands for the autopilot channels to control the motion of the missile. The warhead subsystem receives a burst-hit command from the guidance system. The performance of the guided missile systems is assessed by their terminal effect. The generation and intelligent control of this "terminal effect" is one of the key requirements to missile systems.

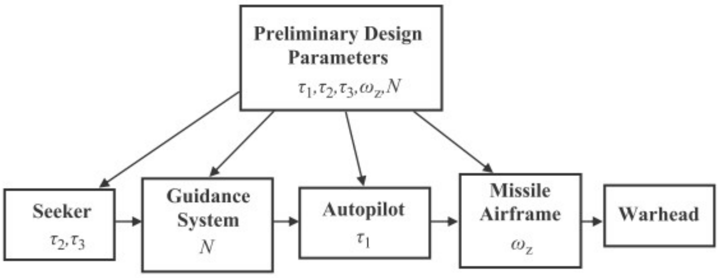

FIGURE 4.1
Block diagram of an interceptor's main subsystems.

The abovementioned subsystems are interconnected. The performance of a separate subsystem dictates requirements to the interconnected ones; e.g., the missile airframe parameters determine the airframe poles ω_z that influence significantly missile dynamics and, as a result, influence autopilot system τ_1 characteristic requirements. Higher accuracy guidance and autopilot systems can employ smaller warheads. The seeker dynamic parameters τ_2 and τ_3 influence the accuracy of the guidance system. The traditional approach to designing missile guidance and autopilot systems usually neglects the interaction between these systems and treats individual missile subsystems separately. The subsystems are designed separately and then integrated before verifying their performance. The quantification of the impact of missile parameters on the miss distance is the first important step toward integrated design of missile guidance and autopilot systems. The main factors that influence the miss distance in homing missiles are the seeker errors, aero frame characteristics, autopilot lag τ_1 and target maneuvers. An appropriate choice of the estimation system parameters (in Figure 4.1, it is combined with a seeker and presented by τ_2 and τ_3) can reduce requirements to a seeker's accuracy and a guidance law effective navigation ratio *N*. The abovementioned shows the importance of establishing an analytical relationship between the miss distance and the main parameters of the missile guidance system.

The analytical expressions for the miss distance (frequency response) and the related expressions for the missile system performance are obtained for the proportional navigation law and the guidance control system that reflects the most important characteristics of the flight control system that combines airframe and autopilot dynamics (damping, natural frequency, time constant, and airframe zero frequency). The obtained analytical expressions need to employ significantly simpler computational programs than by using the method of adjoints to analyze the influence of the basic guidance system parameters on the miss distance for step and weave maneuvers.

4.2 Adjoint Method: Generalized Model

As shown earlier, the method of adjoints is a useful tool to simulate the impulse response of the system, i.e., its response of an impulse function. The analysis in the time domain was bounded by the consideration of a step target acceleration signal. Because the frequency domain analysis utilizes a different test signal, it is convenient to consider a more general model of the adjoint system commonly used in modern control theory.

!)*+,)))-) #($#$$ $; 5!
-7 < /
)*+,)))-) #(2 5 4
2 5 5!

)*+,)))-) #(GA ! %
*55 H + ? / I @ !7 8 A ! 5 . 0 #1 5
! @ > + 5!
' &6 $& . %

38 GEOMETRIC INTERSECTION

David M. Mount

INTRODUCTION

Detecting whether two geometric objects intersect and computing the region of intersection are fundamental problems in computational geometry. Geometric intersection problems arise naturally in a number of applications. Examples include geometric packing and covering, wire and component layout in VLSI, map overlay in geographic information systems, motion planning, and collision detection. In solid modeling, computing the volume of intersection of two shapes is an important step in defining complex solids. In computer graphics, detecting the objects that overlap a viewing window is an example of an intersection problem, as is computing the first intersection of a ray and a collection of geometric solids.

Intersection problems are fundamental to many aspects of geometric computing. It is beyond the scope of this chapter to completely survey this area. Instead we illustrate a number of the principal techniques used in efficient intersection algorithms. This chapter is organized as follows. Section 38.1 discusses intersection primitives, the low-level issues of computing intersections that are common to high-level algorithms. Section 38.2 discusses detecting the existence of intersections. Section 38.3 focuses on issues related to counting the number of intersections and reporting intersections. Section 38.4 deals with problems related to constructing the actual region of intersection. Section 38.5 considers methods for geometric intersections based on spatial subdivisions.

38.1 INTERSECTION PREDICATES

GLOSSARY

Geometric predicate: A function that computes a discrete relationship between basic geometric objects.

Boundary elements: The vertices, edges, and faces of various dimensions that make up the boundary of an object.

Complex geometric objects are typically constructed from a number of primitive objects. Intersection algorithms that operate on complex objects often work by breaking the problem into a series of primitive geometric predicates acting on basic elements, such as points, lines and curves, that form the boundary of the objects involved. Examples of geometric predicates include determining whether two line segments intersect each other or whether a point lies above, below, or on a given line.

© 2004 by Chapman & Hall/CRC

Computing these predicates can be reduced to computing the sign of a polynomial, ideally of low degree. In many instances the polynomial arises as the determinant of a symbolic matrix.

Computing geometric predicates in a manner that is efficient, accurate, and robust can be quite challenging. Floating-point computations are fast but suffer from round-off errors, which can result in erroneous decisions. These errors in turn can lead to topological inconsistencies in object representations, and these inconsistencies can cause the run-time failures. Some of the approaches used to address robustness in geometric predicates include approximation algorithm that are robust to floating-point errors [SI94], computing geometric predicates exactly using adaptive floating-point arithmetic [Cla92, ABD$^+$97], exact arithmetic combined with fast floating-point filters [BKM$^+$[95, FV96], and designing algorithms that are based on a restricted set of geometric predicates [BS00].

We will concentrate on geometric intersections involving flat objects (line segments, polygons, polyhedra), but there is considerable interest in computing intersections of curves and surfaces. Predicates for curve and surface intersections are particularly challenging, because the intersection of surfaces of a given algebraic degree generally results in a curve of a significantly higher degree. Computing intersection primitives typically involves solving an algebraic system equations, which can be performed either exactly by algebraic and symbolic methods [Yap93] or approximately by numerical methods [Hof89, MC91]. See Chapter 41.

38.2 INTERSECTION DETECTION

GLOSSARY

Polygonal chain: A sequence of line segments joined end-to-end.

Self-intersecting: Said of a polygonal chain if any pair of nonadjacent edges intersects one another.

Bounding box: A rectangular box surrounding an object, usually axis-aligned (***isothetic***).

Intersection detection, the easiest of all intersection tasks, requires merely determining the existence of an intersection. Nonetheless, detecting intersections efficiently in the absence of simplifying geometric structure can be challenging. As an example, consider the following fundamental intersection problem, posed by John Hopcroft in the early 1980's. Given a set of n points and n lines in the plane, does any point lie on any line? See Figure 38.2.1. A series of efforts to solve *Hopcroft's problem* culminated in the best algorithm known for this problem to date, due to Matoušek [Mat93], which runs in $O(n^{4/3})2^{O(\log^* n)}$. There is reason to believe that this may be close to optimal; Erickson [Eri96] has shown that, in certain models of computation, $\Omega(n^{4/3})$ is a lower bound. Agarwal and Sharir [AS90] have shown that, given two sets of line segments denoted red and blue, it is possible to determine whether there is any red-blue intersection in $O(n^{4/3+\epsilon})$ time, for any positive constant ϵ.

© 2004 by Chapman & Hall/CRC

FIGURE 38.2.1
Hopcroft's Problem.

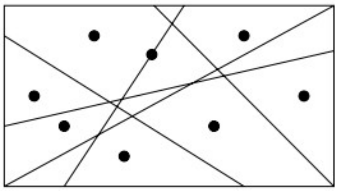

The types of objects considered in this section are polygons, polyhedra, and line segments. Let P and Q denote the two objects to be tested for intersection. Throughout, n_p and n_q denote the combinatorial complexity of P and Q, respectively, that is, the number of vertices, edges, and faces (for polyhedra). Let $n = n_p + n_q$ denote the total complexity.

Table 38.2.1 summarizes a number of results on intersection detection, which will be discussed further in this section. In the table, the terms *convex* and *simple* refer to convex and simple polygons, respectively. The notation $(s(n), q(n))$ in the "Time" column means that the solution involves preprocessing, where a data structure of size $O(s(n))$ is constructed so that intersection detection queries can be answered in $O(q(n))$ time.

TABLE 38.2.1 Intersection detection.

DIM	OBJECTS	TIME	SOURCE
2	convex-convex	$\log n$	[DK83]
	simple-simple	n	[Cha91]
	simple-simple	$(n, s\log^2 n)$	[Mou92]
	line segments	$n \log n$	[SH76]
	Hopcroft's problem	$n^{4/3}2^{O(\log^* n)}$	[Mat93]
3	convex-convex	n	[DK85]
	convex-convex	$(n, \log n_p \log n_q)$	[DK90]

INTERSECTION DETECTION OF CONVEX POLYGONS

Perhaps the most easily understood example of how the structure of geometric objects can be exploited to yield an efficient intersection test is that of detecting the intersection of two convex polygons. There are a number of solutions to this problem that run in $O(\log n)$ time. We present one due to Dobkin and Kirkpatrick [DK83].

Assume that each polygon is given by an array of vertex coordinates, sorted in counterclockwise order. The first step of the algorithm is to find the vertices of each of P and Q with the highest and lowest y-coordinates. This can be done in $O(\log n)$ time by an appropriate modification of binary search and consideration of the direction of the edges incident to each vertex [O'R98, Section 7.6]. After these vertices are located, the boundary of each polygon is split into two semi-infinite convex chains, denoted P_L, P_R and Q_L, Q_R (see Figure 38.2.2(a)). P and Q intersect if and only if P_L and Q_R intersect, and P_R and Q_L intersect.

© 2004 by Chapman & Hall/CRC

FIGURE 38.2.2
Intersection detection for two convex polygons.

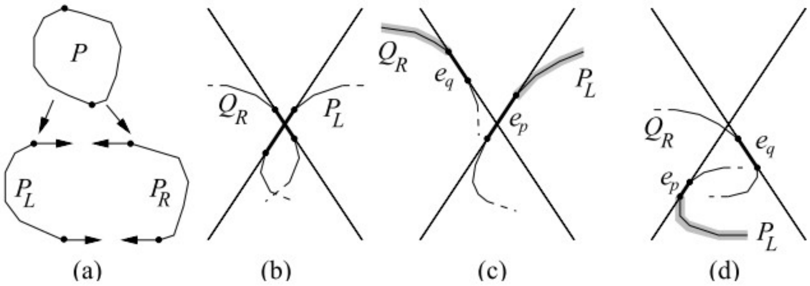

Consider the case of P_L and Q_R. The algorithm applies a variant of binary search. Consider the median edge e_p of P_L and the median edge e_q of Q_R (shown as heavy lines in the figure). By a simple analysis of the relative positions of these edges and the intersection point of the two lines on which they lie, it is possible to determine in constant time either that the polygons intersect, or that half of at least one of the two boundary chains can be eliminated from further consideration. The cases that arise are illustrated in Figure 38.2.2(b)-(d). The shaded regions indicate the portion of the boundary that can be eliminated from consideration.

SIMPLE POLYGONS

Without convexity, it is generally not possible to detect intersections in sublinear time without preprocessing; but efficient tests do exist.

One of the important intersection questions is whether a closed polygonal chain defines the edges of a simple polygon. The problem reduces to detecting whether the chain is self-intersecting. This problem can be solved efficiently by supposing that the polygonal chain is a simple polygon, attempting to triangulate the polygon, and seeing whether anything goes wrong in the process. Some triangulation algorithms can be modified to detect self intersections. In particular, the problem can be solved in $O(n)$ time by modifying Chazelle's linear-time triangulation algorithm [Cha91]. See Section 25.2.

Another variation is that of determining the intersection of two simple polygons. Chazelle observed that this can also be reduced to testing self intersections in $O(n)$ time by joining the polygons into a single closed chain by a narrow channel as shown in Figure 38.2.3.

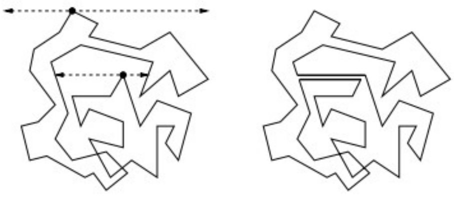

FIGURE 38.2.3
Intersection detection for two simple polygons.

© 2004 by Chapman & Hall/CRC

bounded by a given closed path C in item 2 with the law of conservation of charge, we obtain for any closed surface S,

$$\frac{d}{dt}\left(\oint_S \mathbf{D}\cdot d\mathbf{S} - \int_V \rho\, dv\right) = 0 \tag{1.25}$$

where V is the volume bounded by S. Once again, one can then argue that the quantity inside the parentheses is zero, since at an instant of time when no sources are present, it vanishes. Thus, Gauss' law for the electric field is not independent of Ampere's circuital law in view of the law of conservation of charge.

10. The cut view in Fig. 1.8 indicates that magnetic field lines are continuous, having no beginnings or endings, whereas the cut view in Fig. 1.7 indicates that electric field lines are discontinuous wherever there are charges, diverging from positive charges and converging on negative charges.

1.2.2 Maxwell's Equations in Differential Form and the Continuity Equation

From the integral forms of Maxwell's equations, one can obtain the corresponding differential forms through the use of Stoke's and divergence theorems in vector calculus, given, respectively, by

$$\oint_C \mathbf{A}\cdot d\mathbf{l} = \int_S (\nabla\times\mathbf{A})\cdot d\mathbf{S} \tag{1.26a}$$

$$\oint_S \mathbf{A}\cdot d\mathbf{S} = \int_V (\nabla\cdot\mathbf{A})\, dv \tag{1.26b}$$

where in Eq. (1.26a), S is any surface bounded by C and in Eq. (1.26b), V is the volume bounded by S. Thus, Maxwell's equations in differential form are given by

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{1.27}$$

$$\nabla\times\mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \tag{1.28}$$

$$\nabla\cdot\mathbf{D} = \rho \tag{1.29}$$

$$\nabla\cdot\mathbf{B} = 0 \tag{1.30}$$

corresponding to the integral forms Eqs. (1.18)–(1.21), respectively. These differential equations state that at any point in a given medium, the curl of the electric field intensity is equal to the time rate of decrease of the magnetic flux density, and the curl of the magnetic field intensity is equal to the sum of the current density due to flow of charges and the displacement current density (time derivative of the displacement flux density); whereas the divergence of the displacement flux density is equal to the volume charge density, and the divergence of the magnetic flux density is equal to zero.

Auxiliary to the Maxwell's equations in differential form is the differential equation following from the law of conservation of charge Eq. (1.22) through the use of Eq. (1.26b). Familiarly known as the *continuity equation*, this is given by

$$\nabla \cdot \mathbf{J} + \frac{\partial \rho}{\partial t} = 0 \tag{1.31}$$

It states that at any point in a given medium, the divergence of the current density due to flow of charges plus the time rate of increase of the volume charge density is equal to zero.

From the interdependence of the integral laws discussed in the previous section, it follows that Eq. (1.30) is not independent of Eq. (1.27), and Eq. (1.29) is not independent of Eq. (1.28) in view of Eq. (1.31).

Maxwell's equations in differential form lend themselves well for a qualitative discussion of the interdependence of time-varying electric and magnetic fields giving rise to the phenomenon of electromagnetic wave propagation. Recognizing that the operations of curl and divergence involve partial derivatives with respect to space coordinates, we observe that time-varying electric and magnetic fields coexist in space, with the spatial variation of the electric field governed by the temporal variation of the magnetic field in accordance with Eq. (1.27), and the spatial variation of the magnetic field governed by the temporal variation of the electric field in addition to the current density in accordance with Eq. (1.28). Thus, if in Eq. (1.28) we begin with a time-varying current source represented by $\mathbf{J}$, or a time-varying electric field represented by $\partial\mathbf{D}/dt$, or a combination of the two, then one can visualize that a magnetic field is generated in accordance with Eq. (1.28), which in turn generates an electric field in accordance with Eq. (1.27), which in turn contributes to the generation of the magnetic field in accordance with Eq. (1.28), and so on, as depicted in Fig. 1.11. Note that $\mathbf{J}$ and ρ are coupled, since they must satisfy Eq. (1.31). Also, the magnetic field automatically satisfies Eq. (1.30), since Eq. (1.30) is not independent of Eq. (1.27).

The process depicted in Fig. 1.11 is exactly the phenomenon of electromagnetic waves propagating with a velocity (and other characteristics) determined by the parameters of the medium. In free space, the waves propagate unattenuated with the velocity $1/\sqrt{\mu_0\varepsilon_0}$, familiarly represented by the symbol c. If either the term $\partial\mathbf{B}/\partial t$ in Eq. (1.27) or the term $\partial\mathbf{D}/\partial t$ in Eq. (1.28) is not present, then wave propagation would not occur. As already stated in the previous section, it was through the addition of the term

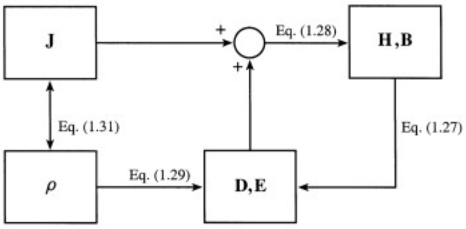

Figure 1.11 Generation of interdependent electric and magnetic fields, beginning with sources $\mathbf{J}$ and ρ.

$\partial \mathbf{D}/\partial t$ in Eq. (1.28) that Maxwell predicted electromagnetic wave propagation before it was confirmed experimentally.

Of particular importance is the case of time variations of the fields in the sinusoidal steady state, that is, the frequency domain case. In this connection, the frequency domain forms of Maxwell's equations are of interest. Using the phasor notation based on

$$A\ \cos(\omega t + \phi) = \mathrm{Re}\left[Ae^{j\phi}e^{j\omega t}\right] = \mathrm{Re}\left[\bar{A}e^{j\omega t}\right] \tag{1.32}$$

where $\bar{A} = Ae^{j\phi}$ is the phasor corresponding to the time function, we obtain these equations by replacing all field quantities in the time domain form of the equations by the corresponding phasor quantities and $\partial/\partial t$ by $j\omega$. Thus with the understanding that all phasor field quantities are functions of space coordinates, that is, $\bar{\mathbf{E}} = \bar{\mathbf{E}}(\mathbf{r})$, etc., we write the Maxwell's equations in frequency domain as

$$\nabla \times \bar{\mathbf{E}} = -j\omega\bar{\mathbf{B}} \tag{1.33}$$

$$\nabla \times \bar{\mathbf{H}} = \bar{\mathbf{J}} + j\omega\bar{\mathbf{D}} \tag{1.34}$$

$$\nabla \cdot \bar{\mathbf{D}} = \bar{\rho} \tag{1.35}$$

$$\nabla \cdot \bar{\mathbf{B}} = 0 \tag{1.36}$$

Also, the continuity equation, Eq. (1.31), transforms to the frequency domain form

$$\nabla \cdot \bar{\mathbf{J}} + j\omega\bar{\rho} = 0 \tag{1.37}$$

Note that since $\nabla \cdot \nabla \times \bar{\mathbf{E}} = 0$, Eq. (1.36) follows from Eq. (1.33), and since $\nabla \cdot \nabla \times \bar{\mathbf{H}} = 0$, Eq. (1.35) follows from Eq. (1.34) with the aid of Eq. (1.37).

Now the constitutive relations in phasor form are

$$\bar{\mathbf{D}} = \varepsilon\bar{\mathbf{E}} \tag{1.38a}$$

$$\bar{\mathbf{H}} = \frac{\bar{\mathbf{B}}}{\mu} \tag{1.38b}$$

$$\bar{\mathbf{J}}_c = \sigma\bar{\mathbf{E}} \tag{1.38c}$$

Substituting these into Eqs. (1.33)–(1.36), we obtain for a material medium characterized by the parameters ε, μ, and σ,

$$\nabla \times \bar{\mathbf{E}} = -j\omega\mu\bar{\mathbf{H}} \tag{1.39}$$

$$\nabla \times \bar{\mathbf{H}} = (\sigma + j\omega\varepsilon)\bar{\mathbf{E}} \tag{1.40}$$

$$\nabla \cdot \bar{\mathbf{H}} = 0 \tag{1.41}$$

$$\nabla \cdot \bar{\mathbf{E}} = \frac{\bar{\rho}}{\varepsilon} \tag{1.42}$$

Note however that if the medium is homogeneous, that is, if the material parameters are independent of the space coordinates, Eq. (1.40) gives

$$\nabla \cdot \bar{\mathbf{E}} = \frac{1}{\sigma + j\omega\varepsilon} \nabla \cdot \nabla \times \bar{\mathbf{H}} = 0 \tag{1.43}$$

so that $\bar{\rho} = 0$ in such a medium.

A point of importance in connection with the frequency domain form of Maxwell's equations is that in these equations, the parameters ε, μ, and σ can be allowed to be functions of ω. In fact, for many dielectrics, the conductivity increases with frequency in such a manner that the quantity $\sigma/\omega\varepsilon$ is more constant than is the conductivity. This quantity is the ratio of the magnitudes of the two terms on the right side of Eq. (1.40), that is, the conduction current density term $\sigma\bar{\mathbf{E}}$ and the displacement current density term $j\omega\varepsilon\bar{\mathbf{E}}$.

1.2.3. Boundary Conditions

Maxwell's equations in differential form govern the interrelationships between the field vectors and the associated source densities at points in a given medium. For a problem involving two or more different media, the differential equations pertaining to each medium provide solutions for the fields that satisfy the characteristics of that medium. These solutions need to be matched at the boundaries between the media by employing "boundary conditions," which relate the field components at points adjacent to and on one side of a boundary to the field components at points adjacent to and on the other side of that boundary. The boundary conditions arise from the fact that the integral equations involve closed paths and surfaces and they must be satisfied for all possible closed paths and surfaces whether they lie entirely in one medium or encompass a portion of the boundary.

The boundary conditions are obtained by considering one integral equation at a time and applying it to a closed path or a closed surface encompassing the boundary, as shown in Fig. 1.12 for a plane boundary, and in the limit that the area enclosed by the closed path, or the volume bounded by the closed surface, goes to zero. Let the quantities pertinent to medium 1 be denoted by subscript 1 and the quantities pertinent to medium 2 be denoted by subscript 2, and $\mathbf{a}_n$ be the unit normal vector to the surface and directed into medium 1. Let all normal components at the boundary in both media be directed along $\mathbf{a}_n$ and denoted by an additional subscript n and all tangential components at the boundary in both media be denoted by an additional subscript t. Let the surface charge density (C/m^2) and the surface current density (A/m) on the boundary be ρ_S and $\mathbf{J}_S$, respectively. Then, the boundary conditions corresponding to the Maxwell's equations in integral form can be summarized as

$$\mathbf{a}_n \times (\mathbf{E}_1 - \mathbf{E}_2) = \mathbf{0} \tag{1.44a}$$

$$\mathbf{a}_n \times (\mathbf{H}_1 - \mathbf{H}_2) = \mathbf{J}_S \tag{1.44b}$$

$$\mathbf{a}_n \cdot (\mathbf{D}_1 - \mathbf{D}_2) = \rho_S \tag{1.44c}$$

$$\mathbf{a}_n \cdot (\mathbf{B}_1 - \mathbf{B}_2) = 0 \tag{1.44d}$$

)*+,)))-) #(#(&; ## 6 5!
-7 < / ! JJJ

)*+,)))-) #(#(&;& # 5!
2 4 9 ! I ! 0 7
' &# 6 &

Remarks.

(1) Serendipity interpolations are obviously incomplete polynomials in ξ and η, hence have poorer local approximation compared to the local approximations based on Pascal rectangle.

(2) There is no particular theoretical basis for deriving them.

(3) In view of p-version hierarchical elements presented in the previous sections, serendipity elements are precluded and are of no practical significance.

8.11 Interpolation functions for 3D elements

For BVPs in $\mathbb{R}^3$ the domain of definition of the BVP is a subspace of $\mathbb{R}^3$, i.e. a volume. Discretizations of such domain naturally leads to 3D subdomains, i.e. 3D elements. In this section we consider hexahedron and tetrahedral families of subdomains (i.e. elements).

8.11.1 Hexahedron elements

Figure 8.50 shows hexahedron elements with distorted faces and edges and their maps in the natural coordinate space ξ, η, ζ in a two unit cube with the origin of the coordinate system located at the center of the element.

8.11.1.1 Mapping of points

In the abstract sense, the mapping of points between ξ, η, ζ and x, y, z-spaces is defined by

$$\begin{aligned} x &= x(\xi, \eta, \zeta) \\ y &= y(\xi, \eta, \zeta) \\ z &= z(\xi, \eta, \zeta) \end{aligned} \tag{8.216}$$

The inverse of the mapping is given by

$$\begin{aligned} \xi &= \xi(x, y, z) \\ \eta &= \eta(x, y, z) \\ \zeta &= \zeta(x, y, z) \end{aligned} \tag{8.217}$$

As in the case of 2D elements, (8.216) is preferable over (8.217). The explicit form of the mapping defined by (8.216) can be established. Let $\bar{N}_i(\xi, \eta, \zeta)$ be basis functions in the natural coordinate space such that $\forall \xi, \eta, \zeta \in \bar{\Omega}^{\xi\eta\zeta}$ or $\bar{\Omega}^m$ we have

$$\bar{N}_i(\xi_j, \eta_j, \zeta_j) = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases} \tag{8.218}$$

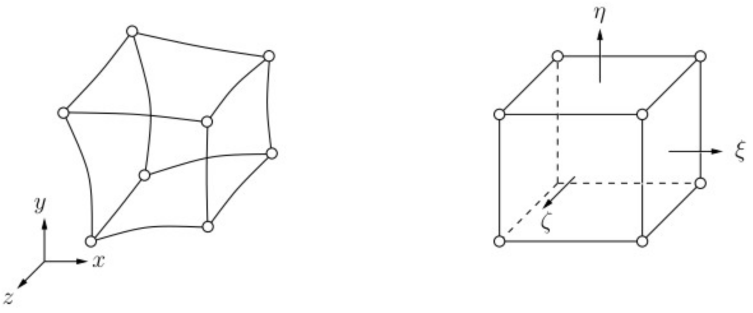

Eight-node element in x, y, z space   Eight-node element in ξ, η, ζ space

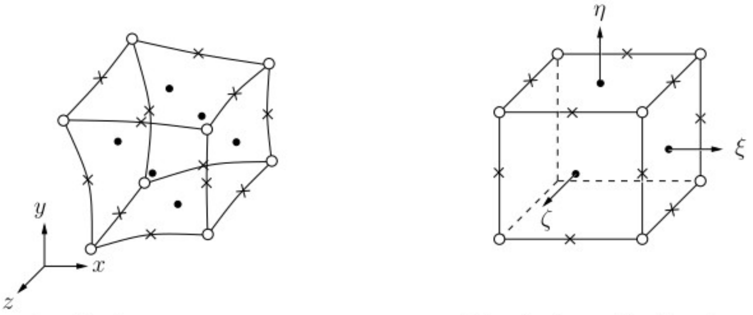

27-node element in x, y, z space   27-node element in ξ, η, ζ space

Figure 8.50: Hexahedral 3D elements

$$\sum_{i=1}^{\bar{n}} \bar{N}_i(\xi, \eta, \zeta) = 1 \tag{8.219}$$

in which $\bar{\Omega}^m$ is the map of $\bar{\Omega}^e$ in ξ, η, ζ space, i.e. mapping (8.216) is such that $\bar{\Omega}^m \to \bar{\Omega}^e$ or by (8.217), $\bar{\Omega}^e \to \bar{\Omega}^m$. Using $\bar{N}_i(\xi, \eta, \zeta)$ we can write,

$$\begin{aligned} x = x(\xi, \eta, \zeta) &= \sum_{i=1}^{\bar{n}} \bar{N}_i(\xi, \eta, \zeta)\, x_i \\ y = y(\xi, \eta, \zeta) &= \sum_{i=1}^{\bar{n}} \bar{N}_i(\xi, \eta, \zeta)\, y_i \\ z = z(\xi, \eta, \zeta) &= \sum_{i=1}^{\bar{n}} \bar{N}_i(\xi, \eta, \zeta)\, z_i \end{aligned} \tag{8.220}$$

in which (x_i, y_i, z_i) are coordinates of node i in x, y, z space. Equation (8.220) map a point (ξ^*, η^*, ζ^*) into (x^*, y^*, z^*).

8.11.1.2 Mapping of lengths

In this section we establish a relationship between length dx, dy, dz in x, y, z space and $d\xi, d\eta, d\zeta$ in ξ, η, ζ space. Based on (8.216), we can write

$$\begin{aligned} dx =& \frac{\partial x}{\partial \xi}\, d\xi + \frac{\partial x}{\partial \eta}\, d\eta + \frac{\partial x}{\partial \zeta}\, d\zeta \\ dy =& \frac{\partial y}{\partial \xi}\, d\xi + \frac{\partial y}{\partial \eta}\, d\eta + \frac{\partial y}{\partial \zeta}\, d\zeta \\ dz =& \frac{\partial z}{\partial \xi}\, d\xi + \frac{\partial z}{\partial \eta}\, d\eta + \frac{\partial z}{\partial \zeta}\, d\zeta \end{aligned} \tag{8.221}$$

or

$$\begin{Bmatrix} dx \\ dy \\ dz \end{Bmatrix} [J] \begin{Bmatrix} d\xi \\ d\eta \\ d\zeta \end{Bmatrix} \tag{8.222}$$

where

$$[J] = \begin{bmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial x}{\partial \eta} & \frac{\partial x}{\partial \zeta} \\ \frac{\partial y}{\partial \xi} & \frac{\partial y}{\partial \eta} & \frac{\partial y}{\partial \zeta} \\ \frac{\partial z}{\partial \xi} & \frac{\partial z}{\partial \eta} & \frac{\partial z}{\partial \zeta} \end{bmatrix} \tag{8.223}$$

is called the Jacobian of transformation. For the mapping to be unique, i.e. one-to-one and onto,

$$det\,[J] > 0 \quad \forall (\xi, \eta, \zeta) \in \bar{\Omega}^m \tag{8.224}$$

must hold. Equations (8.222) is the desired relationship for mapping of lengths. The elements of $[J]$ can easily be obtained using (8.220) in which $\bar{N}_i(\xi, \eta, \zeta)$ and (x_i, y_i, z_i) are known.

8.11.1.3 Mapping of volumes

In this section we derive a relationship that relates elemental volume $dx\, dy\, dz$ in x, y, z space to the elemental volume $d\xi\, d\eta\, d\zeta$ in $\xi\, \eta\, \zeta$ space. Let $\vec{i}, \vec{j}, \vec{k}$ be the unit vectors in x, y, z space and let $\vec{e}_\xi, \vec{e}_\eta, \vec{e}_\zeta$ be unit vectors in ξ, η, ζ space, then we can write

$$\begin{aligned} dx\, \vec{i} &= \frac{\partial x}{\partial \xi}\, d\xi\, \vec{e}_\xi + \frac{\partial x}{\partial \eta}\, d\eta\, \vec{e}_\eta + \frac{\partial x}{\partial \zeta}\, d\zeta\, \vec{e}_\zeta \\ dy\, \vec{j} &= \frac{\partial y}{\partial \xi}\, d\xi\, \vec{e}_\xi + \frac{\partial y}{\partial \eta}\, d\eta\, \vec{e}_\eta + \frac{\partial y}{\partial \zeta}\, d\zeta\, \vec{e}_\zeta \\ dz\, \vec{k} &= \frac{\partial z}{\partial \xi}\, d\xi\, \vec{e}_\xi + \frac{\partial z}{\partial \eta}\, d\eta\, \vec{e}_\eta + \frac{\partial z}{\partial \zeta}\, d\zeta\, \vec{e}_\zeta \end{aligned} \tag{8.225}$$

We note that

$$dx\, \vec{i} \cdot (dy\, \vec{j} \times dz\, \vec{k}) = dx\, \vec{i} \cdot dy\, dz\, \vec{i} = dx\, dy\, dz \tag{8.226}$$

Substituting for $dx\,\vec{i}, dy\,\vec{j}, dz\,\vec{k}$ from (8.225) into (8.226) and using properties of the dot product and cross products of the unit vectors in x, y, z and ξ, η, ζ spaces, we obtain

$$dx\,dy\,dz = det[J]\,d\xi\,d\eta\,d\zeta \tag{8.227}$$

Based on (8.227), $det[J] > 0$ must hold otherwise the volume in xyz space could be negative or zero if $det[J] = 0$.

8.11.1.4 Obtaining derivatives of $\phi_h^e(\xi,\eta,\zeta)$ with respect to x, y, z

With the local approximation $\phi_h^e(\xi,\eta,\zeta)$ of ϕ defined by (8.232), and assuming that $N_i(\xi,\eta,\zeta)$ are known functions, obtaining derivatives of $\phi_h^e(\xi,\eta,\zeta)$ with respect to x, y, z needed in the integral form is not directly possible due to the fact that $N_i(\cdot)$ are functions of ξ, η and ζ. First, we note that when $\phi_h^e(\xi,\eta,\zeta)$ is defined by (8.232) we have

$$\begin{aligned}\frac{\partial \phi_h^e}{\partial x} &= \sum_{i=1}^{n} \frac{\partial N_i}{\partial x}\,\delta_i^e \\ \frac{\partial \phi_h^e}{\partial y} &= \sum_{i=1}^{n} \frac{\partial N_i}{\partial y}\,\delta_i^e \\ \frac{\partial \phi_h^e}{\partial z} &= \sum_{i=1}^{n} \frac{\partial N_i}{\partial z}\,\delta_i^e\end{aligned} \tag{8.228}$$

Thus, obtaining $\frac{\partial \phi_h^e}{\partial x}$, $\frac{\partial \phi_h^e}{\partial y}$ and $\frac{\partial \phi_h^e}{\partial z}$ implies establishing $\frac{\partial N_i}{\partial x}$, $\frac{\partial N_i}{\partial y}$ and $\frac{\partial N_i}{\partial z}$; $i = 1,\dots,n$. Since $N_i = N_i(\xi,\eta,\zeta)$ and $x = x(\xi,\eta,\zeta)$, $y = y(\xi,\eta,\zeta)$ and $z = z(\xi,\eta,\zeta)$, we proceed as follows.

$$\left.\begin{aligned}\frac{\partial N_i}{\partial \xi} &= \frac{\partial N_i}{\partial x}\frac{\partial x}{\partial \xi} + \frac{\partial N_i}{\partial y}\frac{\partial y}{\partial \xi} + \frac{\partial N_i}{\partial z}\frac{\partial z}{\partial \xi} \\ \frac{\partial N_i}{\partial \eta} &= \frac{\partial N_i}{\partial x}\frac{\partial x}{\partial \eta} + \frac{\partial N_i}{\partial y}\frac{\partial y}{\partial \eta} + \frac{\partial N_i}{\partial z}\frac{\partial z}{\partial \eta} \\ \frac{\partial N_i}{\partial \zeta} &= \frac{\partial N_i}{\partial x}\frac{\partial x}{\partial \zeta} + \frac{\partial N_i}{\partial y}\frac{\partial y}{\partial \zeta} + \frac{\partial N_i}{\partial z}\frac{\partial z}{\partial \zeta}\end{aligned}\right\}, \quad i = 1,\dots,n \tag{8.229}$$

or

$$\begin{Bmatrix} \frac{\partial N_i}{\partial \xi} \\ \frac{\partial N_i}{\partial \eta} \\ \frac{\partial N_i}{\partial \zeta} \end{Bmatrix} = [J]^T \begin{Bmatrix} \frac{\partial N_i}{\partial x} \\ \frac{\partial N_i}{\partial y} \\ \frac{\partial N_i}{\partial z} \end{Bmatrix}, \quad i = 1,\dots,n \tag{8.230}$$

Therefore

$$\begin{Bmatrix} \frac{\partial N_i}{\partial x} \\ \frac{\partial N_i}{\partial y} \\ \frac{\partial N_i}{\partial z} \end{Bmatrix} = \Big[[J]^T\Big]^{-1} \begin{Bmatrix} \frac{\partial N_i}{\partial \xi} \\ \frac{\partial N_i}{\partial \eta} \\ \frac{\partial N_i}{\partial \zeta} \end{Bmatrix}, \quad i = 1,\dots,n \tag{8.231}$$

Using (8.231) derivatives of $N_i(\xi,\eta,\zeta)$ with respect to ξ, η and ζ can be transformed into the derivatives of $N_i(\xi,\eta,\zeta)$ with respect to x, y, z and, hence, the derivatives of $\phi_h^e(\xi,\eta,\zeta)$ with respect to x, y, z in (8.228) are defined.

8.11.2 Local approximation for a dependent variable ϕ over $\bar{\Omega}^m$

8.11.2.1 Hexahedron elements

Let $\phi_h^e(\xi,\eta,\zeta)$ be the local approximation of ϕ over $\bar{\Omega}^m$. Then symbolically we can write

$$\phi_h^e(\xi,\eta,\zeta) = \sum_{i=1}^{n} N_i(\xi,\eta,\zeta)\,\delta_i^e = [N]\{\delta^e\} \tag{8.232}$$

where $N_i(\xi,\eta,\zeta)$ are the local approximation functions or interpolation functions corresponding to the nodes of the element and δ_i^e are the nodal degrees of freedom. Functions $N_i(\xi,\eta,\zeta)$ are also referred to as basis functions. Explicit forms of $N_i(\xi,\eta,\zeta)$ depend upon many considerations (similar to these for 2D elements).

(1) The first important issue is the choice of nodal configuration for the element. That is, how many nodes and their locations.

(2) Means of constructing $N_i(\xi,\eta,\zeta)$ for

   (a) C^{000} local approximations of higher degree
   (b) C^{000} p-version hierarchical local approximations
   (c) C^{ijk}, $i,j,k \geq 1$ local approximations for prism family
   (d) C^{ijk}, $i,j,k \geq 1$ local approximations for distorted geometries in $\mathbb{R}^3$

(3) In the developments in (2), the choices of δ_i^e, nodal dofs are crucial and require careful considerations.

(4) The developments in (2) and (3) could be based on (i) Lagrange interpolation functions (ii) Legendre polynomials or (iii) Chebyshev polynomials. The specific details presented in the following are based on Lagrange polynomials. The extensions to the other two types are straight forward. We consider details in the following

$C^{000}(\bar{\Omega}^e)$ polynomial approximations: linear approximation

We consider local approximations of type $C^{000}(\bar{\Omega}^e)$, i.e. across the inter-element boundaries ϕ is continuous but the derivative of ϕ normal to the inter-element boundaries may be discontinuous in the physical domain. For with sides parallel to xyz axes and pointing in the same direction as xyz axes. We consider map of this element in xyz space, a prism where sides are also parallel to xyz axes and pointing in the same direction. By using tensor products of C^i, C^j and C^k one-dimensional interpolations in xyz directions, the local approximation functions and nodal variable operators for the 27 node prism family of hexahedron can be derived in a straight forward manner. These elements have same limitations and restrictions as discussed in section 8.5 for similar local approximations for 2D case in xy space. We remark that these interpolation functions can be of Lagrange, Legendre or Chebyshev type by choosing 1D interpolations based on these.

8.11.2.6 3D $C^{ijk}(\bar{\Omega}^e)$ *p*-version interpolations for distorted hexahedron elements: 27 node element

In deriving the local approximations for this family of hexahedron elements we can follow the procedure similar to that used for 2D elements (section 8.6). The decision of the choice of dofs at the corner is straight forward due to the fact that these must constitute a transformable set between $\xi\eta\zeta$ and xyz space. However, from which nodes to borrow the needed degrees of freedom is not as straight forward but can be facilitated by examining Pascal prism and ensuring that lower order admissible terms are chosen first. These interpolations can also be of Lagrange, Legendre, or Chebyshev type depending upon the choice of 1D interpolations.

8.11.2.7 Interpolation theory for 3D tetrahedron elements: basis functions of class $C^{000}(\bar{\Omega}^e)$ based on Lagrange interpolations

Parallel to Pascal prism, in case of tetrahedron elements we have Pascal pyramid (see Fig. 8.62). The locations of the terms are the locations of the nodes and the terms themselves are to be used in the linear combination for the local approximations. For example a linear tetrahedron will contain four vertex nodes with the local approximation (for a dependent variable ϕ)

$$\phi_h^e(x,y,z) = c_1 + c_2x + c_3y + c_4z \tag{8.251}$$

Using Cartesian coordinates of the vertex nodes (x_i, y_i, z_i) are the function values $\phi(x_i, y_i, z_i) = \phi_i^e$ $(i = 1, 2, \dots, 4)$. In (8.251), it is straight forward to establish

$$\phi_h^e(x,y,z) = \sum_{i=1}^{4} N_i(x,y,z)\,\phi_i^e \tag{8.252}$$

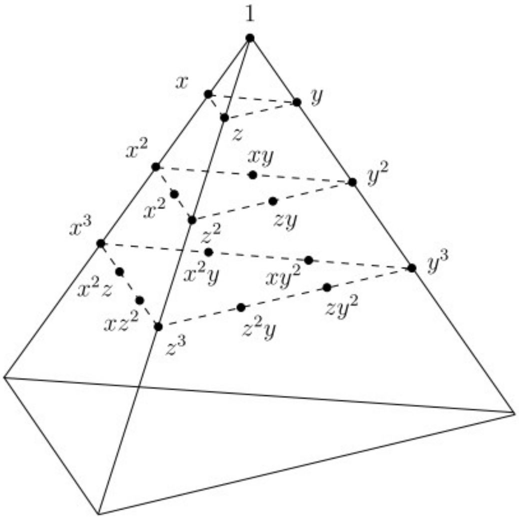

Figure 8.62: Pascal pyramid

in which $N_i(x, y, z)$ are the interpolation functions with the properties

$$
\begin{gathered}
N_i(x_j, y_j, z_j) = \begin{cases} 1, & \text{if } j = i \\ 0, & \text{if } j \neq i \end{cases} \\
\sum_{i=1}^{4} N_i(x, y, z) = 1
\end{gathered}
\tag{8.253}
$$

Using this procedure higher degree local approximations of class $C^{000}(\bar{\Omega}^e)$ can be constructed for the tetrahedron elements. As in case of hexahedron elements based on Pascal prism, here also with increasing degree of local approximations the number of nodes for an element increase and we require inverse of progressively increasing size coefficient matrices to determine $N_i(x, y, z)$.

8.11.2.8 Lagrange family C^{000} interpolations based on volume coordinates

Parallel to the area coordinates for 2D triangular elements, in case of tetrahedron elements we introduce the concept of volume coordinates. Consider a four-node tetrahedron element shown in Fig. 8.63. Consider a point P in the interior of the element. Connect point P with the vertices of the tetrahedron by straight lines. By doing so we divide the volume V into four volumes V_1, V_2, V_3 and V_4. Let the side opposite to node i be side i with volume V_i.

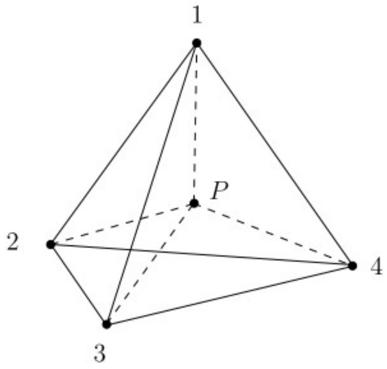

Figure 8.63: Tetrahedron element with volume coordinates

Let

$$\begin{aligned} &L_1 = \frac{V_i}{V}, \quad i = 1, 2, \dots, 4 \\ &\text{with} \quad \sum_{i=1}^{4} L_i = 1 \end{aligned} \tag{8.254}$$

be the volume coordinates. Clearly L_i is equal to 1 at node i but zero on side i. The volume coordinates can be used to define the Cartesian coordinates of any point on the element. If (x_i, y_i, z_i) are the Cartesian coordinates of the nodes of the element, then

$$\begin{aligned} x &= L_1x_1 + L_2x_2 + L_3x_3 + L_4x_4 \\ y &= L_1y_1 + L_2y_2 + L_3y_3 + L_4y_4 \\ z &= L_1z_1 + L_2z_2 + L_3z_3 + L_4z_4 \end{aligned} \tag{8.255}$$

Using (8.255) and the second equation of (8.254), we can write

$$\begin{Bmatrix} 1 \\ x \\ y \\ z \end{Bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \\ z_1 & z_2 & z_3 & z_4 \end{bmatrix} = \begin{Bmatrix} L_1 \\ L_2 \\ L_3 \\ L_4 \end{Bmatrix} = [C] \begin{Bmatrix} L_1 \\ L_2 \\ L_3 \\ L_4 \end{Bmatrix} \tag{8.256}$$

Therefore

$$\begin{Bmatrix} L_1 \\ L_2 \\ L_3 \\ L_4 \end{Bmatrix} = [C]^{-1} \begin{Bmatrix} 1 \\ x \\ y \\ z \end{Bmatrix} = \begin{Bmatrix} N_1(x,y,z) \\ N_2(x,y,z) \\ N_3(x,y,z) \\ N_4(x,y,z) \end{Bmatrix} \tag{8.257}$$

We can show that $N_i(x, y, z)$ here in (8.257) are same as these using Pascal pyramid. We note that $\det[C]$ is related to the volume of tetrahedron hence

$\det[C] > 0$ holds implying that $N_i(x,y,z)$ in (8.257) are unique. Thus for the four-node tetrahedron we can use L_1, L_2, L_3, L_4 as interpolation functions instead of $N_i(x,y,z)$; $i = 1, 2, \dots, 4$.

8.11.2.9 Higher degree C^{000} basis functions using volume coordinates

Consider the four-node tetrahedron element shown in Fig. 8.63 with vertex nodes 1, 2, 3 and 4. We draw m equally spaced planes parallel to side 1 labelled $0, 1, \dots, p \dots, m$, 0 being label for side 1. Similarly, we also draw m equally spaced planes parallel to sides 2, 3 and 4, sides 2, 3 and 4 being marked '0' and q, r, s being intermediate locations of the planes parallel to sides 2, 3, and 4. If we consider typical planes p, q, r and s (parallel to sides 1, 2, 3 and 4) then their intersection $pqrs$ defines a point within the tetrahedron element. The intersection of these planes with sides 1, 2, 3 and 4 and amongst themselves define the location of the nodes in agreement with the Pascal pyramid. Consider side 1 and planes parallel to it. We note that L_1 is one at node 1 and zero on side 1, hence we can setup Lagrange type interpolations to define functions $N_0(L_1), N_1(L_1), \dots, N_m(L_1)$ corresponding to $(m+1)$ parallel planes. We can define $N_i(L_1)$ by

$$\begin{aligned} N_i(L_1) &= \prod_{j=1}^{i} \Big(\frac{mL_1 - j + 1}{j} \Big), \quad \text{for } j \geq 1;\ i = 1, 2, \dots, m \\ &= 0 \quad \text{for } i = 0 \end{aligned} \tag{8.258}$$

Similarly in the other three directions, we setup $N_i(L_2$, $N_i(L_3)$ and $N_i(L_4)$

$$\begin{aligned} N_i(L_2) &= \prod_{j=1}^{i} \Big(\frac{mL_2 - j + 1}{j} \Big), \quad \text{for } j \geq 1;\ i = 1, 2, \dots, m \\ &= 0 \quad \text{for } i = 0 \end{aligned} \tag{8.259}$$

$$\begin{aligned} N_i(L_3) &= \prod_{j=1}^{i} \Big(\frac{mL_3 - j + 1}{j} \Big), \quad \text{for } j \geq 1;\ i = 1, 2, \dots, m \\ &= 0 \quad \text{for } i = 0 \end{aligned} \tag{8.260}$$

$$\begin{aligned} N_i(L_4) &= \prod_{j=1}^{i} \Big(\frac{mL_4 - j + 1}{j} \Big), \quad \text{for } j \geq 1;\ i = 1, 2, \dots, m \\ &= 0 \quad \text{for } i = 0 \end{aligned} \tag{8.261}$$

The 3D interpolations or basis functions for a node located at p, q, r, s i.e. $N_{pqrs}(L_1, L_2, L_3, L_4)$ can now be defined using

$$N_{pqrs}(L_1, L_2, L_3, L_4) = N_p(L_1)\, N_q(L_2)\, N_r(L_3)\, N_s(L_4) \tag{8.262}$$

25/04/2020

E) Books_71_2020/
Handbook of Engineering Electromagnetics/
[Discrete Mathematics...] Rajeev Bansal(..).pdf

Pág. 860

38. Geometric Intersection
David M. Mount

Pág. 862 Figure 38.2.1
Hopcroft's Problem

F) Books_71_2020/Handbook of Engineering
Electromagnetics/ Rajeev Bansal-Handbook....pdf

Pág. 12(22) 1.2.2 Maxwell's Equations
in Differential Form and the Continuity Equation.

G) Books_71_2020/ Chemical Engineering Fluid Mechanics/
1498724426.pdf
→ livro todo.

H) Books_71_2020/The Finite Element Method for
Boundary Value Problems Mathematics and
Computations/
1498780504.pdf

Pág. 584(611) 8.11 Interpolation functions for
3D elements.
Hexahedral 3D elements Figure 8.50
Pág. 601 Figure 8.62 Pascal Pyramid;
Pág. 602 Figure 8.63 Tetrahedron element with
Volume coordinates.

% " " (
)*+,)))-) 7 > 0
7 K #1 + 5!
' & $ $:

For the second Householder reflection ($j = 2$), the linear system becomes

$$\begin{bmatrix} -95.39 & -2.20 \\ 0 & -1.47 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} -54.51 \\ -51.10 \\ 11.91 \\ 13.64 \\ 5.36 \\ -17.91 \\ 3.81 \end{bmatrix}.$$

We can now solve the system with back substitution, starting with the first nonzero row, and the solution is

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} -0.23 \\ 34.82 \end{bmatrix}.$$

Excluding numerical round-off, this is the same solution found using the normal equations in Example 12.13.

◆

13.2 Vector Norms

Finding the magnitude or length of a 3D vector is fundamental to many geometric operations. It is also fundamental for n-dimensional vectors, even though such vectors might no longer have geometric meaning. For example, in Section 13.6, we examine iterative methods for solving linear systems, and vector length is key for monitoring improvements in the solution.

The magnitude or length of an $n-$dimensional vector $\mathbf{v}$ is typically measured by

$$\|\mathbf{v}\|_2 = \sqrt{v_1^2 + \ldots + v_n^2}. \qquad (13.2)$$

The (nonnegative) scalar $\|\mathbf{v}\|_2$ is also referred to as the *Euclidean norm* because in $\mathbb{R}^3$ it is Euclidean length. Since this is the "usual" way to measure length, the subscript 2 often is omitted.

More generally, (13.2) is a *vector norm*. Other vector norms are conceivable, for instance the 1-norm

$$\|\mathbf{v}\|_1 = |v_1| + |v_2| + \ldots + |v_n| \qquad (13.3)$$

or the ∞-norm

$$\|\mathbf{v}\|_\infty = \max_i |v_i|. \qquad (13.4)$$

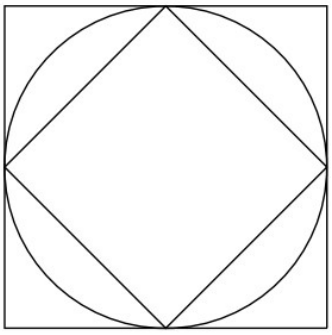

Figure 13.2.

Vector norms: outline of the unit vectors for the 2-norm is a circle, ∞-norm is a square, 1-norm is a diamond.

We can gain an understanding of these norms by studying the familiar case $n = 2$. Recall that a *unit vector* is one whose norm equals one. For the 2-norm $\|\mathbf{v}\|_2$, all unit vectors form a circle of radius one. With a bit of thinking, we can see that for the 1-norm all unit vectors form a diamond, while in the ∞-norm, all unit vectors form a square of edge length two. This geometric interpretation is shown in Figure 13.2.

All three of these norms may be viewed as members of a whole family of norms, referred to as p-norms. The p-norm of $\mathbf{v}$ is given by

$$\|\mathbf{v}\|_p = (v_1^p + v_2^p + \ldots + v_n^p)^{1/p}.$$

It is easy to see that the 1- and 2-norm fit this mold. That $p \to \infty$ gives the ∞-norm requires a little more thought; give it a try!

Example 13.4

Let

$$\mathbf{v} = \begin{bmatrix} 1 \\ 0 \\ -2 \end{bmatrix}.$$

Then

$$\|\mathbf{v}\|_1 = 3, \qquad \|\mathbf{v}\|_2 = \sqrt{5} \approx 2.24, \qquad \|\mathbf{v}\|_\infty = 2.$$

This example and Figure 13.2 demonstrate that

$$\|\mathbf{v}\|_1 \geq \|\mathbf{v}\|_2 \geq \|\mathbf{v}\|_\infty; \tag{13.5}$$

this is also true in n dimensions.

All vector norms share some basic properties:

$$\|\mathbf{v}\| \geq 0, \tag{13.6}$$

$$\|\mathbf{v}\| = 0 \quad \text{if and only if} \quad \mathbf{v} = \mathbf{0}, \tag{13.7}$$

$$\|c\mathbf{v}\| = |c|\|\mathbf{v}| \quad \text{for} \quad c \in \mathbb{R}, \tag{13.8}$$

$$\|\mathbf{v} + \mathbf{w}\| \leq \|\mathbf{v}\| + \|\mathbf{w}\|. \tag{13.9}$$

The last of these is the familiar *triangle inequality.*

By this time, we have developed a familiarity with the 2-norm, and these properties are easily verified. Recall the steps for the triangle inequality were demonstrated in (2.24).

Let's show that the vector norm properties hold for the ∞-norm. Examining (13.4), for each $\mathbf{v}$ in $\mathbb{R}^n$, by definition, $\max_i |v_i| \geq 0$. If $\max_i |v_i| = 0$ then $v_i = 0$ for each $i = 1, \ldots, n$ and $\mathbf{v} = \mathbf{0}$. Conversely, if $\mathbf{v} = \mathbf{0}$ then $\max_i |v_i| = 0$. Thus we have established that the first two properties hold. The third property is easily shown by the properties of the absolute value function:

$$\|c\mathbf{v}\|_\infty = \max_i |cv_i| = |c| \max_i |v_i| = |c|\|\mathbf{v}\|_\infty.$$

For the triangle inequality:

$$\begin{aligned}
\|\mathbf{v} + \mathbf{w}\|_\infty &= \max_i |v_i + w_i| \\
&\leq \max_i \{|v_i| + |w_i|\} \\
&\leq \max_i |v_i| + \max_i |w_i| \\
&= \|\mathbf{v}\|_\infty + \|\mathbf{w}\|_\infty.
\end{aligned}$$

From a 2D geometric perspective, consider the 1-norm: starting at the origin, move v_1 units along the $\mathbf{e}_1$-axis and then move v_2 units along the $\mathbf{e}_2$-axis. The sum of these travels determines $\mathbf{v}$'s length. This is analogous to driving in a city with rectilinear streets, and thus this norm is also called the Manhattan or taxicab norm.

The ∞-norm is also called the max-norm. Our standard measure of length, the 2-norm, can take more computing power than we might be willing to spend on a proximity problem, where we want to determine whether two points are closer than a given tolerance. Problems of this sort are often repeated many times and against many points. Instead, the max norm might be more suitable from a computing point of view and sufficient given the relationship in (13.5). This method of using an inexpensive measure to exclude many possibilities is called *trivial reject* in some disciplines such as computer graphics.

13.3 Matrix Norms

A vector norm measures the magnitude of a vector; does it also make sense to talk about the magnitude of a matrix?

Instead of exploring the general case right away, we will first give some 2D insight. Let A be a 2×2 matrix. It maps the unit circle to an ellipse—the action ellipse. In Figure 13.3 we see points on the unit circle at the end of unit vectors $\mathbf{v}_i$ and their images $A\mathbf{v}_i$, forming this action ellipse. We can learn more about this ellipse by looking at $A^{\mathrm{T}}A$. It is *symmetric* and *positive definite*, and thus has real and positive eigenvalues, which are λ'_1 and λ'_2, in decreasing value. We then define

$$\sigma_i = \sqrt{\lambda'_i}, \tag{13.10}$$

to be the *singular values* of A. (In Chapter 16, we will explore singular values in more detail.)

The singular value σ_1 is the length of the action ellipse's semi-major axis and σ_2 is the length of the action ellipse's semi-minor axis. If A

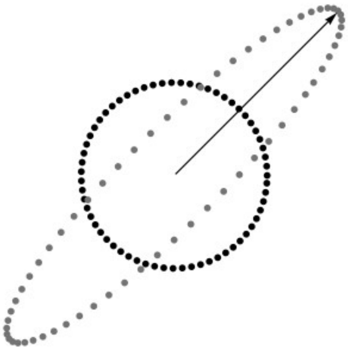

Figure 13.3.

Action of a 2×2 matrix: points on the unit circle (black) are mapped to points on an ellipse (gray) This is the *action ellipse*. The vector is the semi-major axis of the ellipse, which has length σ_1.

12

Gauss for Linear Systems

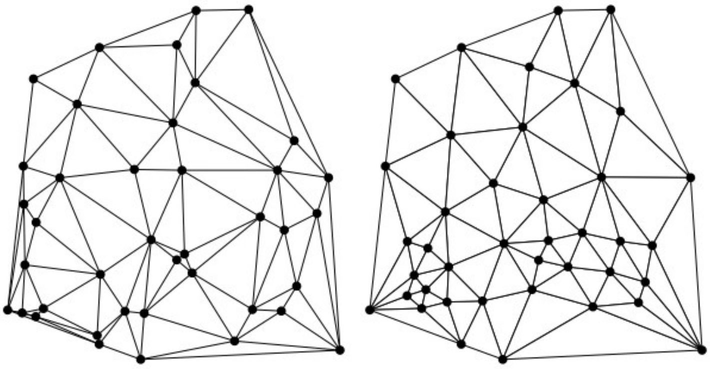

Figure 12.1.

Linear systems: the triangulation on the right was obtained from the left one by solving a linear system.

In Chapter 5, we studied linear systems of two equations in two unknowns. A whole chapter for such a humble task seems like a bit of overkill—its main purpose was really to lay the groundwork for this chapter.

Linear systems arise in virtually every area of science and engineering—some are as big as 1,000,000 equations in as many unknowns. Such huge systems require more sophisticated treatment than the methods introduced here. They *will* allow you to solve systems with several thousand equations without a problem.

Figure 12.1 illustrates the use of linear systems in the field of data smoothing. The left triangulation looks somewhat "rough"; after setting up an appropriate linear system, we compute the "smoother" triangulation on the right, in which the triangles are closer to being equilateral.

This chapter explains the basic ideas underlying linear systems. Readers eager for hands-on experience should get access to software such as Mathematica or MATLAB. Readers with advanced programming knowledge can download linear system solvers from the web. The most prominent collection of routines is LAPACK.

12.1 The Problem

A linear system is a set of equations like this:

$$\begin{aligned} 3u_1 - 2u_2 - 10u_3 + u_4 &= 0 \\ u_1 - u_3 &= 4 \\ u_1 + u_2 - 2u_3 + 3u_4 &= 1 \\ u_2 + 2u_4 &= -4. \end{aligned}$$

The unknowns are the numbers u_1, u_2, u_3, u_4. There are as many equations as there are unknowns, four in this example. We rewrite this 4×4 linear system in matrix form:

$$\begin{bmatrix} 3 & -2 & -10 & 1 \\ 1 & 0 & -1 & 0 \\ 1 & 1 & -2 & 3 \\ 0 & 1 & 0 & 2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 4 \\ 1 \\ -4 \end{bmatrix}.$$

A general $n \times n$ linear system looks like this:

$$\begin{aligned} a_{1,1}u_1 + a_{1,2}u_2 + \ldots + a_{1,n}u_n &= b_1 \\ a_{2,1}u_1 + a_{2,2}u_2 + \ldots + a_{2,n}u_n &= b_2 \\ &\vdots \\ a_{n,1}u_1 + a_{n,2}u_2 + \ldots + a_{n,n}u_n &= b_n. \end{aligned}$$

In matrix form, it becomes

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,n} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,n} \\ & & \vdots & \\ a_{n,1} & a_{n,2} & \ldots & a_{n,n} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}, \tag{12.1}$$

$$\begin{bmatrix}\mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_n\end{bmatrix}\mathbf{u} = \mathbf{b},$$

or even shorter

$$A\mathbf{u} = \mathbf{b}.$$

The *coefficient matrix* A has n rows and n columns. For example, the first row is

$$a_{1,1}, a_{1,2}, \dots, a_{1,n},$$

and the second column is

$$\begin{matrix} a_{1,2} \\ a_{2,2} \\ \vdots \\ a_{n,2}. \end{matrix}$$

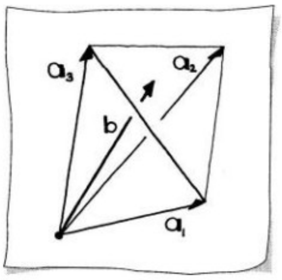

Sketch 12.1.
A solvable 3×3 system.

Equation (12.1) is a compact way of writing n equations for the n unknowns $u_1, \dots, u_n$. In the 2×2 case, such systems had nice geometric interpretations; in the general case, that interpretation needs n-dimensional linear spaces, and is not very intuitive. Still, the methods that we developed for the 2×2 case can be gainfully employed here!

Some underlying principles with a geometric interpretation are best explained for the example $n = 3$. We are given a vector $\mathbf{b}$ and we try to write it as a linear combination of vectors $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$,

$$\begin{bmatrix}\mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3\end{bmatrix}\mathbf{u} = \mathbf{b}.$$

If the $\mathbf{a}_i$ are truly 3D, i.e., if they form a tetrahedron, then a *unique solution* may be found (see Sketch 12.1). But if the three $\mathbf{a}_i$ all lie in a plane (i.e., if the volume formed by them is zero), then you cannot write $\mathbf{b}$ as a linear combination of them, unless it is itself in that 2D plane. In this case, you cannot expect uniqueness for your answer. Sketch 12.2 covers these cases. In general, a linear system is uniquely solvable if the $\mathbf{a}_i$ have a nonzero n-dimensional volume. If they do not, they span a k-dimensional *subspace* (with $k < n$)—nonunique solutions exist only if $\mathbf{b}$ is itself in that subspace. A linear system is called *consistent* if at least one solution exists.

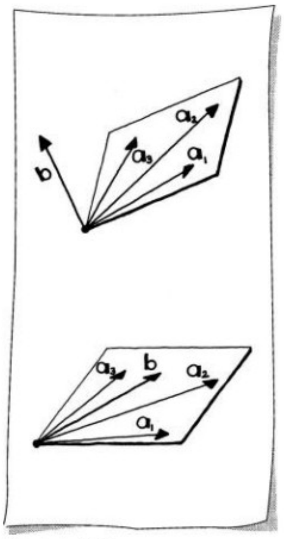

Sketch 12.2.
Top: no solution, bottom: nonunique solution.

For 2D and 3D, we encountered many problems that lent themselves to constructing the linear system in terms of a linear combination of column vectors, the $\mathbf{a}_i$. However, in Section 5.11 we looked at how a linear system can be interpreted as a problem using equations built row by row. In n-dimensions, this commonly occurs. An example follows.

Example 12.1

Suppose that at five time instances, say $t_i = 0, 0.25, 0.5, 0.75, 1$ seconds, we have associated observation data, $p(t_i) = 0, 1, 0.5, 0.5, 0$. We would like to fit a polynomial to these data so we can estimate values in between the observations. This is called *polynomial interpolation*. Five points require a degree four polynomial,

$$p(t) = c_0 + c_1 t + c_2 t^2 + c_3 t^3 + c_4 t^4,$$

which has five coefficients c_i. Immediately we see that we can write down five equations,

$$p(t_i) = c_0 + c_1 t_i + c_2 t_i^2 + c_3 t_i^3 + c_4 t_i^4, \qquad i = 0, \dots, 4,$$

or in matrix form,

$$\begin{bmatrix} 1 & t_0 & t_0^2 & t_0^3 & t_0^4 \\ 1 & t_1 & t_1^2 & t_1^3 & t_1^4 \\ & & \vdots & & \\ 1 & t_4 & t_4^2 & t_4^3 & t_4^4 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_4 \end{bmatrix} = \begin{bmatrix} p(t_0) \\ p(t_1) \\ \vdots \\ p(t_4) \end{bmatrix}.$$

Figure 12.2 illustrates the result, $p(t) = 12.667t - 50t^2 + 69.33t^3 - 32t^4$, for the given data.

◇□◆

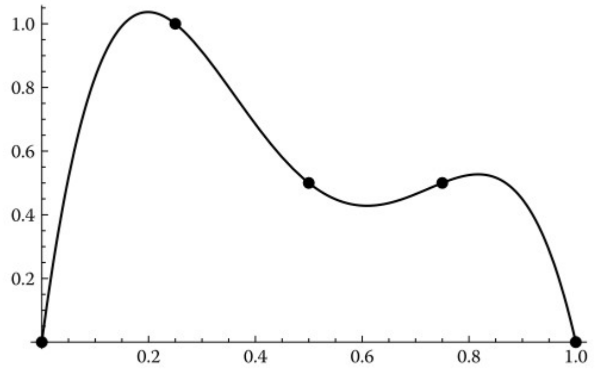

Figure 12.2.
Polynomial interpolation: a degree four polynomial fit to five data points.

12.2 The Solution via Gauss Elimination

The key to success in the 2×2 case was the application of a shear (forward elimination) so that the matrix A was transformed to *upper triangular*, meaning all entries below the diagonal are zero. Then it was possible to apply *back substitution* to solve for the unknowns. A shear was constructed to map the first column vector of the matrix onto the $\mathbf{e}_1$-axis. Revisiting an example from Chapter 5, we set

$$\begin{bmatrix} 2 & 4 \\ 1 & 6 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \end{bmatrix}. \tag{12.2}$$

The shear used was

$$S_1 = \begin{bmatrix} 1 & 0 \\ -1/2 & 1 \end{bmatrix},$$

which when applied to the system as

$$S_1 A\mathbf{u} = S_1 \mathbf{b}$$

produced the system

$$\begin{bmatrix} 2 & 4 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}.$$

Algebraically, what this shear did was to change the *rows* of the system in the following manner:

$$\mathrm{row}_1 \leftarrow \mathrm{row}_1 \quad \text{and} \quad \mathrm{row}_2 \leftarrow \mathrm{row}_2 - \frac{1}{2}\mathrm{row}_1.$$

Each of these constitutes an *elementary row operation.* Back substitution came next, with

$$u_2 = \frac{1}{4} \times 2 = \frac{1}{2}$$

and then

$$u_1 = \frac{1}{2}(4 - 4u_2) = 1.$$

The divisions in the back substitution equations are actually scalings, thus they could be rewritten in terms of a scale matrix:

$$S_2 = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/4 \end{bmatrix},$$

and then the system would be transformed via

$$S_2 S_1 A\mathbf{u} = S_2 S_1 \mathbf{b}.$$

The corresponding upper triangular matrix and system is

$$\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1/2 \end{bmatrix}.$$

Check for yourself that we get the same result.

Thus, we see the geometric steps for solving a linear system have methodical algebraic interpretations. We will be following this algebraic approach for the rest of the chapter. For general linear systems, the matrices, such as S_1 and S_2 above, are not actually constructed due to speed and storage expense. Notice that the shear to zero one element in the matrix, changed the elements in only one row, thus it is unnecessary to manipulate the other rows. This is an important observation for large systems.

In the general case, just as in the 2×2 case, *pivoting* will be used. Recall for the 2×2 case this meant that the equations were reordered such that the (pivot) matrix element $a_{1,1}$ is the largest one in the first column. A row exchange can be represented in terms of a *permutation matrix*. Suppose the 2×2 system in (12.2) was instead,

$$\begin{bmatrix} 1 & 6 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \end{bmatrix},$$

requiring pivoting as the first step. The permutation matrix that will exchange the two rows is

$$P_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

which is the identity matrix with the rows (columns) exchanged. After this row exchange, the steps for solving $P_1 A\mathbf{u} = P_1\mathbf{b}$ are the same as for the system in (12.2): $S_2 S_1 P_1 A\mathbf{u} = S_2 S_1 P_1 \mathbf{b}$.

Example 12.2

Let's step through the necessary row exchanges and shears for a 3×3 linear system. The goal is to get it in upper triangular form so we may use back substitution to solve for the unknowns. The system is

$$\begin{bmatrix} 2 & -2 & 0 \\ 4 & 0 & -2 \\ 4 & 2 & -4 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} 4 \\ -2 \\ 0 \end{bmatrix}.$$

The matrix element $a_{1,1}$ is not the largest in the first column, so we choose the 4 in the second row to be the *pivot element* and we reorder the rows:

$$\begin{bmatrix} 4 & 0 & -2 \\ 2 & -2 & 0 \\ 4 & 2 & -4 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} -2 \\ 4 \\ 0 \end{bmatrix}.$$

The permutation matrix that achieves this row exchange is

$$P_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

(The subscript 1 indicates that this matrix is designed to achieve the appropriate exchange for the first column.)

To zero entries in the first column apply:

$$\begin{aligned} \text{row}_2 &\leftarrow \text{row}_2 - \frac{1}{2}\text{row}_1 \\ \text{row}_3 &\leftarrow \text{row}_3 - \text{row}_1, \end{aligned}$$

and the system becomes

$$\begin{bmatrix} 4 & 0 & -2 \\ 0 & -2 & 1 \\ 0 & 2 & -2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} -2 \\ 5 \\ 2 \end{bmatrix}.$$

The shear matrix that achieves this is

$$G_1 = \begin{bmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}.$$

Now the first column consists of only zeroes except for $a_{1,1}$, meaning that it is lined up with the $\mathbf{e}_1$-axis.

Now work on the second column vector. First, check if pivoting is necessary; this means checking that $a_{2,2}$ is the largest in absolute value of all values in the second column that are below the diagonal. No pivoting is necessary. (We could say that the permutation matrix $P_2 = I$.) To zero the last element in this vector apply

$$\text{row}_3 \leftarrow \text{row}_3 + \text{row}_2,$$

which produces

$$\begin{bmatrix} 4 & 0 & -2 \\ 0 & -2 & 1 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} -2 \\ 5 \\ 7 \end{bmatrix}.$$

The shear matrix that achieves this is

$$G_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}.$$

By chance, the second column is aligned with $\mathbf{e}_2$ because $a_{1,2} = 0$. If this extra zero had not appeared, then the last operation would have mapped this 3D vector into the $[\mathbf{e}_1, \mathbf{e}_2]$-plane.

Now that we have the matrix in upper triangular form, we are ready for back substitution:

$$u_3 = \frac{1}{-1}(7)$$

$$u_2 = \frac{1}{-2}(5 - u_3)$$

$$u_1 = \frac{1}{4}(-2 + 2u_3).$$

This implicitly incorporates a scaling matrix. We obtain the solution

$$\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} -4 \\ -6 \\ -7 \end{bmatrix}.$$

It is usually a good idea to insert the solution into the original equations:

$$\begin{bmatrix} 2 & -2 & 0 \\ 4 & 0 & -2 \\ 4 & 2 & -4 \end{bmatrix} \begin{bmatrix} -4 \\ -6 \\ -7 \end{bmatrix} = \begin{bmatrix} 4 \\ -2 \\ 0 \end{bmatrix}.$$

It works!

◆

The example above illustrates each of the *elementary row operations* that take place during Gauss elimination:

- Pivoting results in the exchange of two rows.
- Shears result in adding a multiple of one row to another.
- Scaling results in multiplying a row by a scalar.

Gauss elimination for solving a linear system consists of two basic steps: forward elimination (pivoting and shears) and then back substitution (scaling). Here is the algorithm for solving a general $n \times n$ system of linear equations.

Gauss Elimination with Pivoting

Given: An $n \times n$ coefficient matrix A and a $n \times 1$ right-hand side $\mathbf{b}$ describing a linear system

$$A\mathbf{u} = \mathbf{b},$$

which is short for the more detailed (12.1).

Find: The unknowns $u_1, \ldots, u_n$.

Algorithm:

```
Initialize the $n \times n$ matrix $G = I$.
For $j = 1, \ldots, n-1$   ($j$ counts columns)

    Pivoting step:

    Find the element with the largest absolute value in column $j$
    from $a_{j,j}$ to $a_{n,j}$; this is element $a_{r,j}$.
        If $r > j$, exchange equations $r$ and $j$.

    If $a_{j,j} = 0$, the system is not solvable.

    Forward elimination step for column $j$:

    For $i = j+1, \ldots, n$  (elements below diagonal of column $j$)
        Construct the multiplier $g_{i,j} = a_{i,j}/a_{j,j}$
        $a_{i,j} = 0$
        For $k = j+1, \ldots, n$  (each element in row $i$ after
                                  column $j$)
            $a_{i,k} = a_{i,k} - g_{i,j}a_{j,k}$
        $b_i = b_i - g_{i,j}b_j$

At this point, all elements below the diagonal have been set to
zero. The matrix is now in upper triangular form.

    Back substitution:

    $u_n = b_n/a_{n,n}$
    For $j = n-1, \ldots, 1$
        $u_j = \frac{1}{a_{j,j}}[b_j - a_{j,j+1}u_{j+1} - \ldots - a_{j,n}u_n]$.
```

In a programming environment, it can be convenient to form an *augmented matrix*, which is the matrix A augmented with the vector $\mathbf{b}$. Here is the idea for a 3×3 linear system:

$$\begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & b_1 \\ a_{2,1} & a_{2,2} & a_{2,3} & b_2 \\ a_{3,1} & a_{3,2} & a_{3,3} & b_3 \end{bmatrix}.$$

Then the k steps would run to $n+1$, and there would be no need for the extra line for the b_i element.

As demonstrated in Example 12.2, the operations in the elimination step above may be written in matrix form. If A is the current matrix, then at step j, we first check if a row exchange is necessary. This may be achieved with a permutation matrix, P_j. If no pivoting is necessary, $P_j = I$. To produce zeroes under $a_{j,j}$ the matrix product $G_j A$ is formed, where

$$G_j = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & 1 & & & \\ & & & -g_{j+1,j} & 1 & & \\ & & & \vdots & & \ddots & \\ & & & -g_{n,j} & & & 1 \end{bmatrix}. \tag{12.3}$$

The elements $-g_{i,j}$ of G_j are the *multipliers*. The matrix G_j is called a *Gauss matrix*. All entries except for the diagonal and the entries $-g_{i,j}$ below the diagonal of the jth column are zero. We could store all P_j and G_j in one matrix

$$G = G_{n-1}P_{n-1} \cdot \ldots \cdot G_2 \cdot P_2 \cdot G_1 \cdot P_1. \tag{12.4}$$

If no pivoting is required, then it is possible to store the $g_{i,j}$ in the zero elements of A rather than explicitly setting the $a_{i,j}$ element equal to zero. Regardless, it is more efficient with regard to speed and storage not to multiply A and $\mathbf{b}$ by the permutation and Gauss matrices because this would result in many unnecessary calculations.

To summarize, Gauss elimination with pivoting transforms the linear system $A\mathbf{u} = \mathbf{b}$ into the system

$$GA\mathbf{u} = G\mathbf{b},$$

which has the same solution as the original system. The matrix GA is upper triangular, and it is referred to as U. The diagonal elements of U are the pivots.

-7 < / ! JJJ

7)*+,)))-) 9 @ !7 8
9 0 ! 5!
' # ; $ #

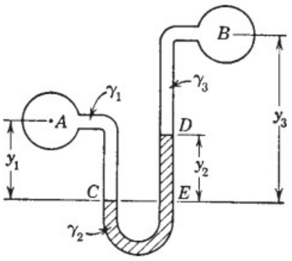

Figure 9 Differential manometer.

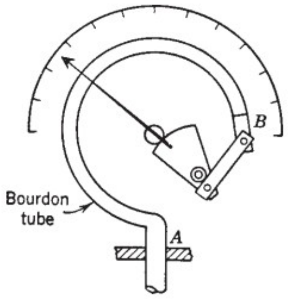

Figure 10 Essential features of Bourdon gauge.

does not vary in the horizontal direction for a fluid at rest and making use of the hydrostatic pressure equation at various points, the gauge pressure at A is calculated as

$$p_A = \gamma_2 y_2 - \gamma_1 y_1 \tag{10}$$

Figure 9 illustrates a differential manometer. The difference in pressure between A and B is calculated as

$$p_B - p_A = \gamma_1 y_1 - \gamma_2 y_2 - \gamma_3 (y_3 - y_2) \tag{11}$$

Another commonly used mechanical device for measuring both gauge pressure and absolute pressure is the Bourdon-type gauge (Fig. 10).

3.3 Fluid Forces on Plane Surfaces

The magnitude of the total fluid force exerted on a plane (flat) surface can be calculated with the equation (refer to Fig. 11)

$$F_R = p_c A \tag{12}$$

where p_c is the pressure at the centroid of the surface and A is the total area of the surface. Note that pressure always acts normal to the surface.

If the fluid is a liquid with a free surface open to the atmosphere,

$$p_c = \gamma h_c \tag{13}$$

where h_c is the distance the centroid lies below the free surface. Locations of centroids of variously shaped areas are shown in Table 5.

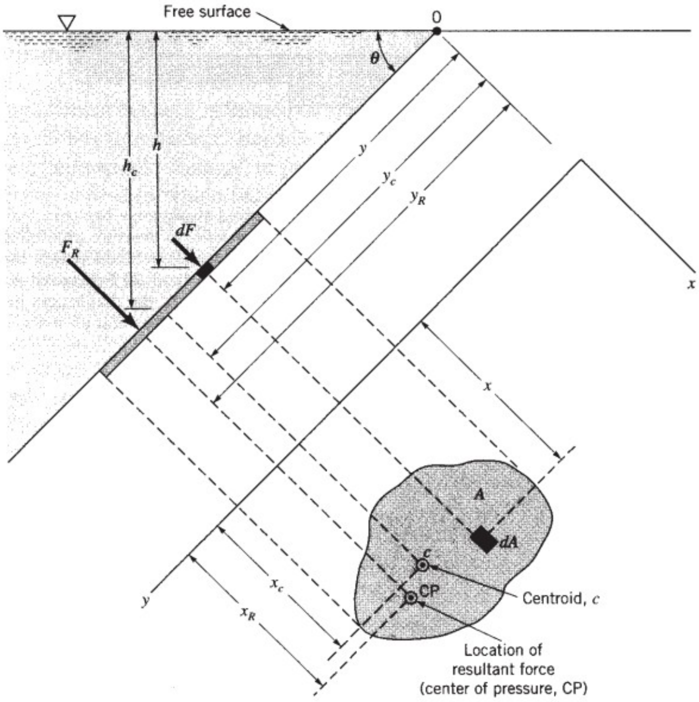

Figure 11 Notation for pressure force calculation on submerged plane of arbitrary shape. (From Ref. 2.)

The location of the resultant of the pressure forces acting on a plane area is known as the center of pressure. For a plane area lying below a liquid surface, the center-of-pressure location is computed with the equation

$$y_R - y_c = \frac{I_c}{y_c A} \tag{14}$$

where y_R is measured in the plane of the area (slant distance) from the free surface to the center of pressure, y_c is the distance (in the same plane) from the free surface to the centroid of the area, and I_c is the second moment of the area about a horizontal axis through the centroid and in the plane of the area. Values of I_c for common geometric shapes are given in Table 5.

Two special cases of this application are the horizontal-plane and the vertical-plane cases. For a horizontal plane ($\theta = 0°$ of inclination), the centroid of the surface and the center of pressure coincide since the pressure distribution over the entire surface area is uniform. For the vertical case ($\theta = 90°$ of inclination), the center of pressure will always be below the centroid of the surface since the pressure acting on the surface increases with depth. Note that for deeply submerged plane areas, the center of pressure approaches the centroid of the area regardless of the angle of inclination.

3.4 Fluid Forces on Curved Surfaces

Pressure forces exerted on curved surfaces vary in direction and must usually be divided into horizontal and vertical components to facilitate calculation. The resultants of the horizontal and

Table 5 Properties of Areas and Volumes

	Sketch	Area or Volume	Location of Centroid	I or I_c
Rectangle	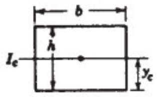	bh	$y_c = \frac{h}{2}$	$I_c = \frac{bh^2}{12}$
Triangle	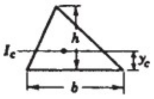	$\frac{bh}{2}$	$y_c = \frac{h}{3}$	$I_c = \frac{bh^2}{36}$
Circle	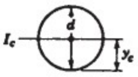	$\frac{\pi d^2}{4}$	$y_c = \frac{d}{2}$	$I_c = \frac{\pi d^4}{64}$
Semicircle[a]	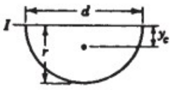	$\frac{\pi d^2}{8}$	$y_c = \frac{4r}{3\pi}$	$I = \frac{\pi d^4}{128}$
Ellipse	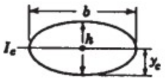	$\frac{\pi bh}{4}$	$y_c = \frac{h}{2}$	$I_c = \frac{\pi bh^2}{64}$
Semiellipse	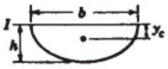	$\frac{\pi bh}{4}$	$y_c = \frac{4h}{3\pi}$	$I = \frac{\pi bh^2}{16}$
Parabola	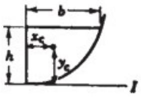	$\frac{2}{3}bh$	$y_c = \frac{3h}{5}$ $x_c = \frac{3b}{8}$	$I = \frac{2bh^2}{7}$
Cylinder	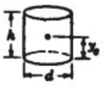	$\frac{\pi d^2 h}{4}$	$y_c = \frac{h}{2}$	
Cone	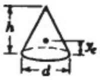	$\frac{1}{3}\left(\frac{\pi d^2 h}{4}\right)$	$y_c = \frac{h}{4}$	
Paraboloid of revolution	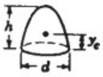	$\frac{1}{2}\left(\frac{\pi d^2 h}{4}\right)$	$y_c = \frac{h}{3}$	
Sphere	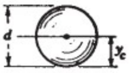	$\frac{\pi d^2}{6}$	$y_c = \frac{d}{2}$	
Hemisphere	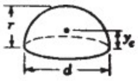	$\frac{\pi d^2}{12}$	$y_c = \frac{3r}{8}$	

[a] For the quarter-circle, the respective values are $\pi d^2/16$, $4r/3\pi$, and $\pi d^4/256$.

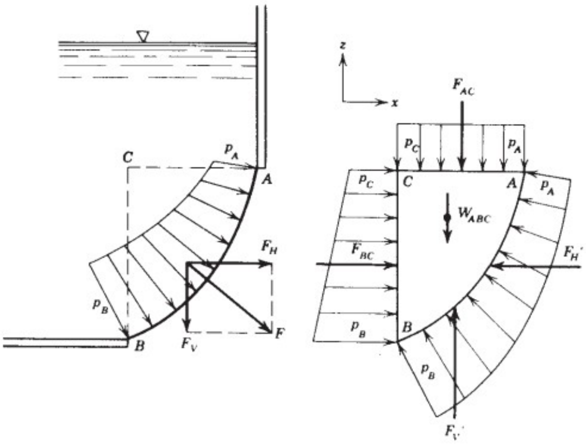

Figure 12 Forces acting on curved surface. *Source*: By permission from J. K. Vennard and R. L. Street, *Elementary Fluid Mechanics*, 5th ed., Wiley, New York, 1975.

vertical components of the pressure forces are determined with assistance from the equations developed for plane surfaces. Figure 12 illustrates a common situation. The fluid mass *ABC* is in equilibrium under the action of pressure forces; hence the vertical component of the resultant of the pressure forces is equal to the weight of the fluid mass plus the force $\mathbf{F}_{AC}$ exerted on the upper surface of the fluid mass by the overlying fluid. The horizontal component of the resultant of the pressure forces is equal to the force on a vertical projection of the curved surface. The line of action and point of application (center of pressure) for the resultant force is determined by taking moments of the contributing forces about an axis of convenience. The force $\mathbf{F}_{BC}$ is located using the formula for plane surfaces. The weight W_{ABC} is concentrated at the centroid of the area *ABC* (see Table 5).

3.5 Buoyancy and Stability

An object floating or submerged in a fluid at rest experiences pressure forces from the surrounding fluid. Consider a square block of material that is submerged in a liquid; due to hydrostatic pressure variation with depth, the pressure on the bottom surface will be greater than on the top surface (pressure on the two sides is equal and opposite). The resultant of these pressure forces is known as the buoyant force, which acts in the upward direction, and is equal in magnitude to the weight of the displaced fluid. The line of action of the buoyant force is vertical through the centroid of the volume of displaced fluid (center of buoyancy). The magnitude of the buoyant force can be found from the following equation:

$$F_B = \gamma V \tag{15}$$

where V is the volume of the displaced fluid. If the object is totally submerged in the fluid, then the volume of the displaced fluid equals the volume of the object; otherwise, the displaced volume equals only the volume of the submerged portion of the object.

The overturning stability (Fig. 13) of a floating or submerged body is affected by the relative position of the center of gravity (CG) of the body and the center of buoyancy. When the

)*+,)))-) = ! ! 2 5 0> *
% ! &1 L @ 58 = / 0 5!
' 6; ;$ 5

Chapter 5

Surfaces

On the surface of the Earth, there are two degrees of freedom. Simply fix a direction to face; the line forward and backward relative to you is one degree of freedom, while the line left and right through you determines another. Ignoring the scenery, buildings, etc., there is no topological difference between the patch of Earth you are standing on and the one near me; the Earth's surface always looks the same locally (see Figure 5.1). In fact, if we did not realize the Earth is a sphere, we could mistake any particular patch of land or sea as part of a plane,[1] or a cylinder, or even a torus. This property of indistinguishability is key in defining what we mean by a surface. In this chapter we define surfaces; introduce methods for combining, dissecting, and analyzing surfaces; and then prove half of the important theorem that classifies all compact surfaces. Near the end of the chapter we explore vector fields on surfaces other than the familiar plane.

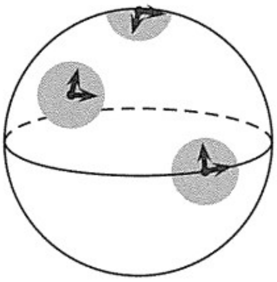

Figure 5.1: There are two degrees of freedom at any point on the surface of the Earth.

---

[1] As early as the sixth century BC, Pythagoras postulated that the Earth is a sphere, though his reasons were more philosophical than empirical. In the fourth century BC, Eratosthenes computed the circumference of the Earth using the observation that shadows were cast at different angles in two distant places at the same time. Eratosthenes's estimate was amazingly accurate.

5.1 Surfaces and Surfaces-with-Boundary

Let's begin with a formal definition of *surface*. We assume that our surfaces are connected and have no boundary edge.

Definition 5.1.1. A **surface** is a connected Hausdorff space S such that for each $x \in S$, there is an open set $U \subseteq S$ with $x \in U$, such that $U \approx B_1(\mathbf{0}) \subseteq \mathbb{R}^2$.

An open set $U \approx B_1(\mathbf{0})$ as in Definition 5.1.1 is called a **disk neighborhood** or **patch**, so we may think of a surface S as a patchwork of overlapping open disks. In this chapter we primarily work with *compact* surfaces. Recall from §4.4 that a compact space is one in which every open cover has a finite subcover. Thus, for a compact surface, it is always possible to fulfill Definition 5.1.1 with a *finite* collection of *patches*. Moreover, surfaces that have *punctures* or open edges, as illustrated in Figure 5.2, are not compact. On the other hand, there is no requirement for "smoothness" – a surface may have corners or sharp edges, as in the surface of a cube.

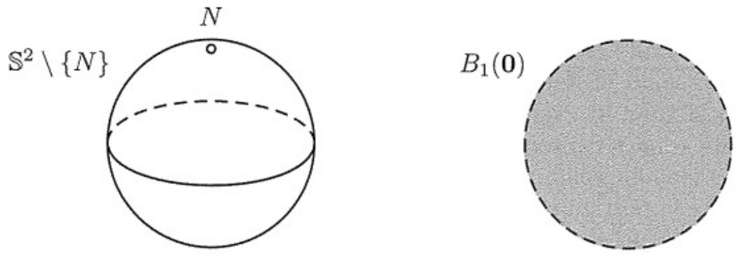

Figure 5.2: The punctured sphere, $\mathbb{S}^2 \setminus \{N\}$, and open disk are noncompact surfaces.

Example 85. Examples of compact surfaces include the sphere $\mathbb{S}^2$, the torus $\mathbb{T}$, and Klein bottle $\mathbb{K}$, among many others. Note that $\mathbb{K}$ cannot be realized in three-dimensional space without *self-intersections* (see, e.g., a typical representation of $\mathbb{K}$ in Figure 1.3). A noncompact surface may extend infinitely, such as the Euclidean plane $\mathbb{R}^2$, or the infinite cylinder, $\mathbb{S}^1 \times \mathbb{R}$, or such a surface may have a puncture and/or open boundary edge, as in the two spaces shown in Figure 5.2.

The following examples are *not* surfaces at all, according to the definition (see Figure 5.3).

- The circle $\mathbb{S}^1$. No open set of $\mathbb{S}^1$ homeomorphic to an open disk. Indeed, we may say $\mathbb{S}^1$ is one dimensional, whereas surfaces are, by definition, two dimensional. The higher-dimensional spheres $\mathbb{S}^3$, $\mathbb{S}^4$, etc., are also not surfaces.

- The closed unit disk $\mathbb{D}^2 = D_1(\mathbf{0}) \subseteq \mathbb{R}^2$. $\mathbb{D}^2$ is not a surface because there are *boundary points* in $\mathbb{D}^2$ that cannot be surrounded by an open set U homeomorphic to an open disk.
- A pair of intersecting planes. Suppose x is any point on the intersection line. Any open set U containing x must contain points that are on both planes, which implies that U is not homeomorphic to an open disk.

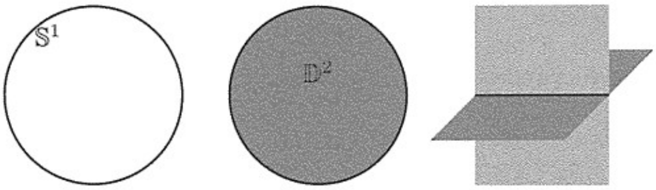

Figure 5.3: The circle, closed disk, and a pair of intersecting planes. None of these spaces is a surface according to Definition 5.1.1.

Consider two related spaces: the open unit disk, $U = B_1(\mathbf{0}) = \{x \in \mathbb{R}^2 \mid d(x, \mathbf{0}) < 1\}$, and the closed unit disk, $\mathbb{D}^2$. According to Definition 5.1.1, U is a noncompact surface, while $\mathbb{D}^2$ is a compact set that is technically *not* a surface. However, *most* points of $\mathbb{D}^2$ do fit the conditions of the definition. Only those points on the boundary edge fail to meet the requirement of belonging to an open set homeomorphic to the open disk; instead, a typical open set containing the boundary point is homeomorphic to the *half-disk* $V = \{(x, y) \in \mathbb{R}^2 \mid d((x, y), 0) < 1,\ y \geq 0\}$, as illustrated in Figure 5.4.

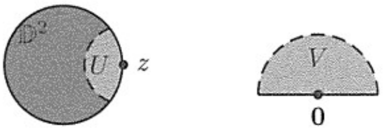

Figure 5.4: An open neighborhood $U \ni z$, for z on the boundary of the closed disk, is homeomorphic to a half-disk V.

Definition 5.1.2. A **surface-with-boundary** is a connected Hausdorff topological space S such that for each $x \in S$, there is an open set $U \subseteq S$ with $x \in U$, such that *either*:

- $U \approx B_1(\mathbf{0}) \subseteq \mathbb{R}^2$ *or*
- $U \approx V = \{(x, y) \in \mathbb{R}^2 \mid d((x, y), 0) < 1,\ y \geq 0\}$.

The terminology *surface-with-boundary* is an unfortunate misnomer, as these spaces are not actually *surfaces* according to the strict definition, unless the boundary is empty. In the literature, the term is usually *surface with boundary* (with no hyphens). While it is important to conform to the established terminology, I compromise in this textbook, erring on the side of clarity, by always writing *surface-with-boundary* (with hyphens).

Example 86. The following are examples of surfaces-with-boundary that are not surfaces.

- The closed cylinder, $\mathbb{S}^1 \times \mathbb{I}$. The boundary consists of two disjoint circles.
- The Möbius strip. The boundary consists of only one circle (see Figure 4.8).
- Suppose S is a surface and U is any open disk neighborhood of a point $x \in S$. Then $S \setminus U$ is a surface-with-boundary.

Throughout the remainder of the text, we use the notation $\mathcal{S}_c$ to denote the collection of all compact connected surfaces and surfaces-with-boundary.

Connected Sum Construction

Surfaces or surfaces-with-boundary may be combined by a process called *connected sum.* Intuitively, the construction is simple: remove an open disk from each of the two spaces, resulting in two surfaces with boundary; then glue the two pieces together along the boundary. A precise definition of this construction is tricky, as is the proof that the construction does not depend on any of the choices made along the way. (*Which open disks? In what way should the boundaries be identified?*) The fact that the connected sum construction is well defined for all compact surfaces is beyond the scope of this text,[2] and we shall take it for granted.

Definition 5.1.3. Suppose $A, B \in \mathcal{S}_c$. Choose $U_1 \subseteq A$ and $U_2 \subseteq B$ open sets, each homeomorphic to an open disk, and let C_1 and C_2 be the boundary circles of U_1 and U_2, respectively. Let $f : C_1 \to C_2$ be a homeomorphism. The **connected sum** of A and B (with respect to U_1, U_2 and f) is the quotient space,

$$A \# B = ((A \setminus U_1) \cup (B \setminus U_2)) \, / \sim,$$

where $x \sim y$ if and only if $x \in C_1$, $y \in C_2$, and $y = f(x)$.

---

[2] See Proposition 2.6.1 of Bloch [Blo97].

The connected sum operation is *commutative* and *associative*, in the sense that the resulting spaces are homeomorphic. That is:

- $A\#B \approx B\#A, \quad \forall A, B \in \mathcal{S}_c$.
- $(A\#B)\#C \approx A\#(B\#C), \quad \forall A, B, C \in \mathcal{S}_c$.

Example 87. Let's see what $\mathbb{T}\#\mathbb{T}$ looks like. As Figure 5.5 illustrates, the result is a *two-holed torus*, $2\mathbb{T}$.

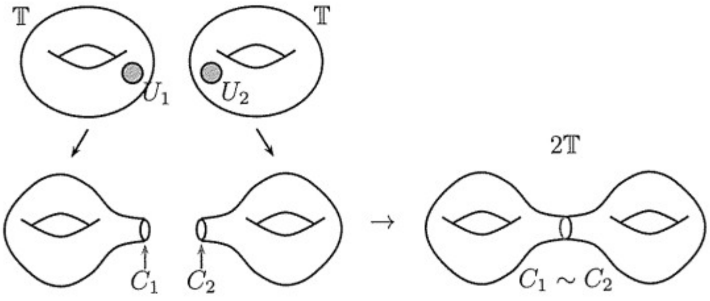

Figure 5.5: $\mathbb{T}\#\mathbb{T} = 2\mathbb{T}$.

Definition 5.1.4. The **n-holed torus**, $n\mathbb{T}$, is the connected sum of n tori. In general, if S is any surface, then nS is the connected sum of n copies of S.

See Figure 5.6 for illustrations of the n-holed torus.

Example 88. Describe each of the following:

(a) $\mathbb{T}\#\mathbb{S}^2$ (b) $4\mathbb{S}^2$ (c) $3\mathbb{T}\#2\mathbb{T}$

Solution:

(a) $\mathbb{T}\#\mathbb{S}^2 \approx \mathbb{T}$, as illustrated below.

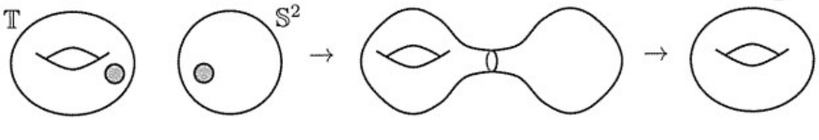

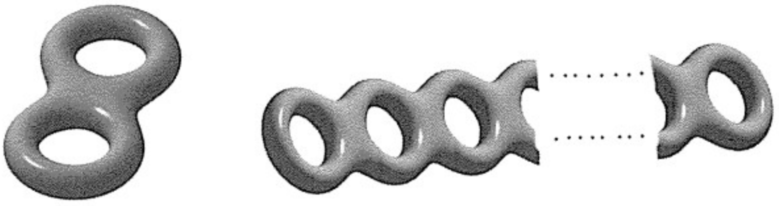

Figure 5.6: *Left*, the two-holed torus, $2\mathbb{T}$. *Right*, the n-holed torus, $n\mathbb{T}$. Images courtesy of Wikimedia Commons (modified).

(b) Just as in example (a), it should be clear that $\mathbb{S}^2\#\mathbb{S}^2 \approx \mathbb{S}^2$. Then $4\mathbb{S}^2 = \left(\left[\mathbb{S}^2\#\mathbb{S}^2\right]\#\mathbb{S}^2\right)\#\mathbb{S}^2 \approx \left(\mathbb{S}^2\#\mathbb{S}^2\right)\#\mathbb{S}^2 \approx \mathbb{S}^2\#\mathbb{S}^2 \approx \mathbb{S}^2$.

(c) $3\mathbb{T}\#2\mathbb{T} = (\mathbb{T}\#\mathbb{T}\#\mathbb{T})\#(\mathbb{T}\#\mathbb{T}) \approx 5\mathbb{T}$.

Example 89. Suppose P is a polyhedron in $\mathbb{R}^3$. Let P' be the union of the edges of P (think of this as a "wireframe"). Now suppose the edges in P' are thickened a bit, and let S be the surface of the "thickened wireframe." (That is, S is the boundary of a *tubular neighborhood* of P'; see §4.5, Exercise 8.) It can be shown that S is always homeomorphic to a connected sum of tori ($n\mathbb{T}$ for some $n \in \mathbb{N}$). Consider the tetrahedron T. Its thickened wireframe can be deformed to $3\mathbb{T}$, as illustrated in Figure 5.7. First flatten the wireframe and then arrange the holes. For what value of n is the surface in Figure 5.8 homeomorphic to $n\mathbb{T}$?

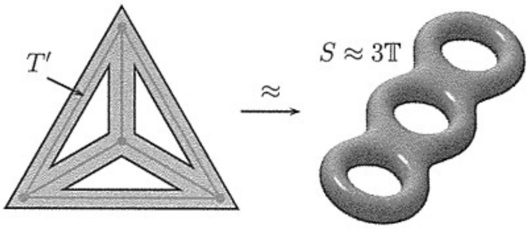

Figure 5.7: The thickened wireframe of a tetrahedron is homeomorphic to $3\mathbb{T}$.

In Example 88, we have shown that $3\mathbb{T}\#2\mathbb{T} \approx 5\mathbb{T}$. It's not too hard to see that $m\mathbb{T}\#n\mathbb{T} \approx (m+n)\mathbb{T}$ in general. It should also be clear that the sphere $\mathbb{S}^2$ acts as an *identity* for the connected sum operation. That is,

$$A\#\mathbb{S}^2 \approx \mathbb{S}^2\#A \approx A, \quad \forall A, B \in \mathcal{S}_c.$$

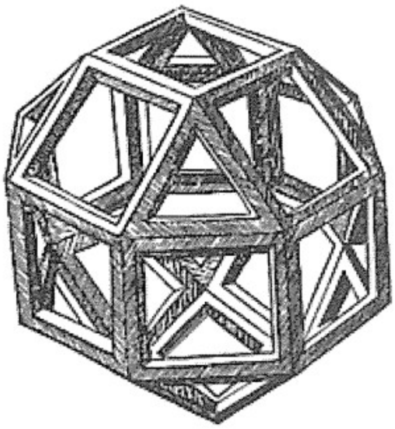

Figure 5.8: The thickened frame of a polyhedron (drawn by Leonardo da Vinci). The surface is homeomorphic to a connected sum of tori. Image courtesy of Wikimedia Commons.

Together with associativity, the existence of an identity makes $\mathcal{S}_c$ into a *monoid* under the connected sum operation. We shall have more to say about this in the next section.

Manifolds

Surfaces are essentially two-dimensional spaces, but Definition 5.1.1 can easily be extended to arbitrary dimensions. We call such spaces *manifolds.*[3]

Definition 5.1.5. Let $n \in \mathbb{N}$. A **manifold** is a connected Hausdorff space M such that for each $x \in M$, there is an open set $U \subseteq M$ with $x \in U$, such that $U \approx B_1(\mathbf{0}) \subseteq \mathbb{R}^n$. The manifold is said to be n-dimensional, or M is called an n-manifold.

There is an analogous definition of **manifold-with-boundary** (try to formulate a reasonable definition yourself, extending Definition 5.1.2).

Example 90. Because manifolds (by our definition) must be connected, the only zero-dimensional manifold is a single point. A one-dimensional manifold must be locally homeomorphic to an open interval. The whole real line $\mathbb{R}$ and the open interval (a, b) are examples of *noncompact* 1-manifolds. A closed interval

---

[3] Perhaps the better term here would be *topological* manifolds. In the literature, much more attention is paid to the more restrictive *smooth* or *differentiable* manifold theory, which falls outside the scope of this text.

$[a, b]$ is compact, but not a manifold; instead, $[a, b]$ is a manifold-with-boundary. The only compact 1-manifold (according to Definition 5.1.5) is the circle $\mathbb{S}^1$.

Example 91. Let's explore a few interesting manifolds and manifolds-with-boundary.

- Recall that $\mathbb{I}^n = \mathbb{I} \times \mathbb{I} \times \ldots \times \mathbb{I} \subseteq \mathbb{R}^n$ is the product of the unit interval with itself n times. This space is not a manifold because it has a boundary whose points cannot be surrounded by an open ball contained in $\mathbb{I}^n$. However, this space, which is called a **hypercube**, is a compact manifold-with-boundary. It is interesting to ponder what a 4-cube ($\mathbb{I}^4$) might look like. This object is called a **tesseract**,[4] and we might visualize it as a cube that has been extended through a fourth dimension (see Figure 5.9). What might $\mathbb{I}^5$ or $\mathbb{I}^6$ look like?

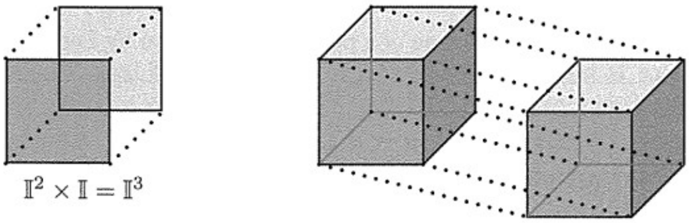

$\mathbb{I}^2 \times \mathbb{I} = \mathbb{I}^3$

$\mathbb{I}^3 \times \mathbb{I} = \mathbb{I}^4$

Figure 5.9: Just as a solid cube may be thought of as a square that has been extended through a third dimension, the tesseract is a cube extended through a fourth dimension.

- The open balls $B_\epsilon(x)$ for $x \in \mathbb{R}^n$ and $\epsilon \in \mathbb{R}^+$ are noncompact manifolds. Their closures $D_\epsilon(x) \approx \mathbb{D}^n$ are compact manifolds-with-boundary (but not manifolds). Note that $\mathbb{I}^n \approx \mathbb{D}^n$ via a deformation that "inflates" the hypercube until all points on the boundary are the same distance from the center (recall Exercise 2 from §4.5).

- The n-sphere $\mathbb{S}^n$, or **hypersphere**, was introduced in Chapter 2 by way of a particular representation of the space, as a *unit* sphere:

  $$\mathbb{S}^n = \{x \in \mathbb{R}^{n+1} \mid d(x, \mathbf{0}) = 1\} = \{(x_1, \ldots, x_{n+1}) \mid x_1^2 + \cdots x_{n+1}^2 = 1\}.$$

  There are many other useful representations of the n-sphere. It should not be hard to see that for any given $x \in \mathbb{R}^{n+1}$ and $\epsilon > 0$, we have $\partial B_\epsilon(x) = \partial \mathbb{D}^{n+1} \approx \mathbb{S}^n$; that is, the boundary of any open ball or closed disk in $\mathbb{R}^{n+1}$ is an n-sphere. Thus we also get $\mathbb{S}^n \approx \partial\mathbb{I}^{n+1}$ by way of the homeomorphism $\mathbb{I}^n \approx \mathbb{D}^n$.

---

[4] The tesseract has fascinated many science fiction authors, making an appearance in Robert Heinlein's short story "And He Built a Crooked House" [Hei41] and Madeleine L'Engle's popular children's novel *A Wrinkle in Time* [L'E62].

Now suppose H^+ and $H^- \subseteq \mathbb{S}^n$ are the "northern" and "southern" hemispheres of the unit n-sphere, respectively. That is,

$$\begin{aligned} H^+ &= \{(x_1, \ldots, x_{n+1}) \mid x_1^2 + \cdots x_{n+1}^2 = 1 \text{ and } x_{n+1} \geq 0\} \\ H^- &= \{(x_1, \ldots, x_{n+1}) \mid x_1^2 + \cdots x_{n+1}^2 = 1 \text{ and } x_{n+1} \leq 0\}. \end{aligned}$$

The quotient space H^+/H^- that collapses the entire southern hemisphere to a single point is also homeomorphic to $\mathbb{S}^n$ (and, by symmetry, $H^-/H^+ \approx \mathbb{S}^n$ as well). In fact, if we let $E = H^+ \cap H^-$, which is the equatorial sphere $\mathbb{S}^{n-1} \subseteq \mathbb{S}^n$, then we have

$$H^+/E \approx H^-/E \approx \mathbb{S}^n.$$

Example 22 (in Chapter 2) explains how $\mathbb{S}^n$ can also be regarded as $\mathbb{R}^n \cup \{\infty\}$, where ∞ is the *point at infinity.* Consider now a variation on this idea. Suppose $F : \mathbb{S}^n \setminus \{N\} \to \mathbb{R}^n$ is stereographic projection. The image $F[H^-]$ of the southern hemisphere is a closed disk $\mathbb{D}^n \subseteq \mathbb{R}^n$, as shown in Figure 5.10 (for the two-dimensional case). When all points on the boundary of that disk are identified to a single point, the resulting quotient space is again the n-sphere. That is,

$$\mathbb{D}^n/\partial\mathbb{D}^n \approx \mathbb{I}^n/\partial\mathbb{I}^n \approx \mathbb{S}^n.$$

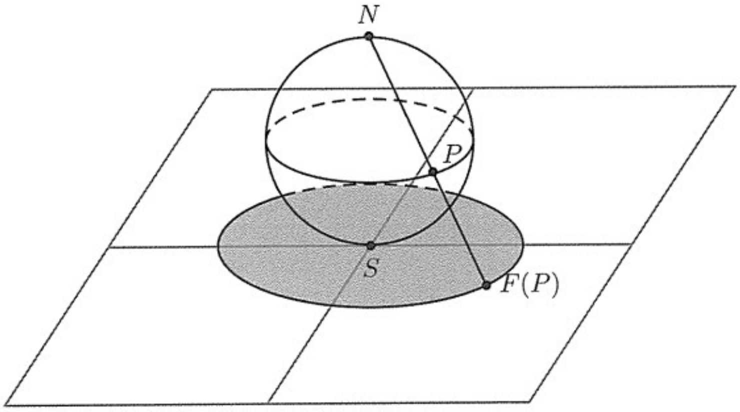

Figure 5.10: Stereographic projection to $\mathbb{R}^2$ sends the southern hemisphere of $\mathbb{S}^2$ to a disk in the plane.

- The n-**torus** $\mathbb{T}^n$ is an n-dimensional manifold defined by $\mathbb{T}^n = \mathbb{I}^n/\sim$, in which the equivalence relation is generated by identifying corresponding points on opposite facets of the hypercube. That is, $(x_1, \ldots, 0, \ldots, x_n) \sim$

$(x_1, \ldots, 1, \ldots, x_n)$, with the understanding that the 0 and 1 occur in the same position and all other entries agree. For $n = 1$, this reduces to the usual definition of $\mathbb{S}^1 = \mathbb{I}/\{0 \sim 1\}$, and for $n = 2$, this space is the familiar torus, $\mathbb{T}^2 = \mathbb{T}$.

The "n" in $\mathbb{T}^n$ is a little misleading. The n-torus is not the product of n copies of $\mathbb{T}$. In fact, the n-torus may be expressed better by $(\mathbb{S}^1)^n$, since $\mathbb{T}^n$ is actually the product of n copies of $\mathbb{S}^1$.

- **Projective space** $\mathbb{P}^n$ is defined as a certain quotient space of $\mathbb{R}^{n+1} \setminus \{\mathbf{0}\}$ as in §4.3, Example 76, and Exercise 11). In essence, every point on a line through the origin (except the origin itself) gets identified to a single point. But $\mathbb{P}^n$ can also be defined by identifying all pairs of **antipodal** points on $\mathbb{S}^n$; that is, points on the sphere that differ only by the scalar multiple -1. This means that $\mathbb{P}^n \approx \mathbb{S}^n/(x \sim -x)$, and so representatives for the equivalence classes may be taken just from the upper hemisphere, $H^+ \subseteq \mathbb{S}^n$. There are still antipodal identifications on the equator, leading to another useful way to write projective space:

$$\mathbb{P}^n \approx H^+/(x \sim -x) \approx \mathbb{D}^n/(x \sim -x, \forall x \in \partial\mathbb{D}^n). \tag{5.1}$$

Looking ahead to Figure 5.24, we can see how these identifications work in $\mathbb{P}^2 = \mathbb{P}$, the *projective plane*. Figure 5.11 shows a representation of $\mathbb{P}^3$.

The connected sum of manifolds may be defined in a manner similar to that of Definition 5.1.3, with $\mathbb{S}^n$ serving as an identity for the connected sums of n-manifolds. However, care must be taken for $n \geq 3$; the *orientation* of each manifold must be considered in order to have a well-defined operation. It would take us too far afield to explore manifolds further in this textbook. The interested reader may consult more advanced texts, such as *Introduction to Topological Manifolds* by John Marshall Lee [Lee00].

Exercises

1. Classify each of the following as surface, surface-with-boundary, or not a surface. If the space is a surface-with-boundary, describe the boundary. In each case, indicate whether the space is compact or not.
   (a) Closed square, $\mathbb{I}^2$.
   (b) Annulus, $\{(x, y) \in \mathbb{R}^2 \mid 1 \leq x^2 + y^2 \leq 2\}$.
   (c) $\mathbb{T} \setminus \{p\}$, where $p \in \mathbb{T}$ is any single point.
   (d) The union of the xy-plane, xz-plane, and yz-plane in $\mathbb{R}^3$.
   (e) Finite open cylinder $\mathbb{S}^1 \times (0, 1)$.

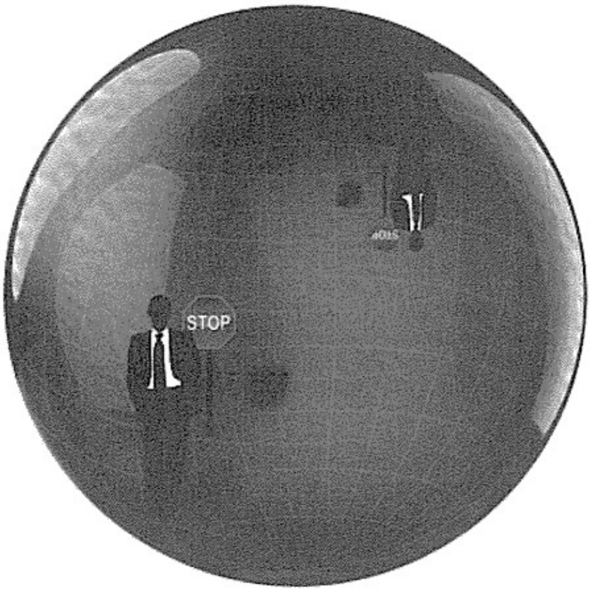

Figure 5.11: Three-dimensional projective space $\mathbb{P}^3$ may be defined as the solid sphere $\mathbb{D}^3$ with antipodal points on the surface identified. If you lived within a small model of $\mathbb{P}^3$, then you could look down the street and see yourself rotated by one half-turn. Image courtesy of Cyrus Rua – used by permission.

2. Consider the capital letters of the alphabet written in block form and thickened to three-dimensional objects (as in Exercise 1 in §1.1). For each letter, identify the value of n such that the letter is homeomorphic to $n\mathbb{T}$.

3. For each of the five Platonic solids (Figure 1.11) P, find the value n such that $P' \approx n\mathbb{T}$, where P' is the *thickened wireframe* of P (see Example 89). What about the object in Figure 5.8?

4. Recall that the Jordan Curve Theorem (Theorem 2.4.8) states that a simple closed curve C separates the plane $\mathbb{R}^2$ into two arc-connected components. Argue, with the help of stereographic projection, that the same happens on the surface of a sphere $\mathbb{S}^2$. Give an example of a simple closed curve on the torus $\mathbb{T}$ that does not separate it.

5. Write a reasonable definition for manifold-with-boundary, analogous to Definition 5.1.2.

6. Show that the connected sum of two connected surfaces is arc-connected (therefore also connected).

7. Show that the connected sum of two compact surfaces is compact. What if one or both of the surfaces are noncompact?

8. Write each connected sum below as $n\mathbb{T}$ for an appropriate value of n. In part (b), assume $k \in \mathbb{N}$.

   (a) $\mathbb{T}\#3\mathbb{S}^2\#2\mathbb{T}\#7\mathbb{S}^2$ (b) $\mathbb{T}\#2\mathbb{T}\#3\mathbb{T}\#\cdots\#k\mathbb{T}$

9. Let $S \in \mathcal{S}_c$ be arbitrary. Describe $S\#\mathbb{D}^2$.

5.2 Plane Models and Words

In this section we develop a powerful combinatorial tool that enables us to work with surfaces in an easier way, eventually allowing us to classify compact surfaces and to understand subtle structures in the surface. Let's revisit the torus, defined as a quotient space as in §4.3.

$$\mathbb{T} = \mathbb{I}^2/\{(0,y) \sim (1,y),\ (x,0) \sim (x,1)\}$$

We represent $\mathbb{T}$ as a square together with the appropriate identifications along the edges as shown in Figure 5.12, with the convention that edges labeled with the same letter should be identified in the direction indicated by the arrowheads.

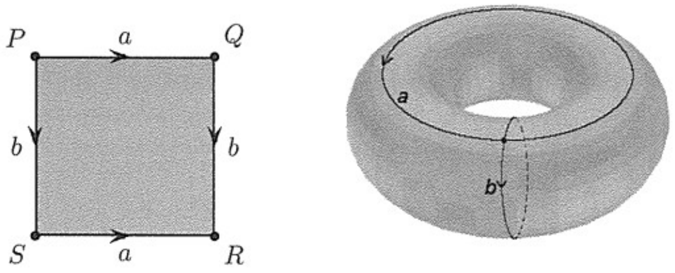

Figure 5.12: Plane model for the torus $\mathbb{T}$ and its realization. After the identifcation, all four vertices are identified to a single point, $P \sim Q \sim R \sim S$.

This type of diagram – a polygon showing certain pairs of edges identified – is called a **plane model** for the surface or surface-with-boundary, and it exhibits the space as a quotient space of a closed disk (recall $\mathbb{I}^2 \approx \mathbb{D}^2$). Any pair of edges marked with the same letter correspond to a single segment or loop in the space itself. Any segments without a label or with a unique label correspond to edges in the space, and in that case, the space may be a surface-with-boundary, but not a surface.

For each plane model, there is a way to describe the model using a sequence of letters, or a **word**. Simply pick a starting vertex; choose a direction, clockwise or counterclockwise; then list the letters in order, indicating the direction of the edge by placing an "exponent" -1 on the letter if the arrow is backward with respect to the direction you chose. There are many different words for the same plane model. For instance, there are eight equivalent words for the torus in Figure 5.12:

$$aba^{-1}b^{-1}, \qquad ba^{-1}b^{-1}a, \qquad a^{-1}b^{-1}ab, \qquad b^{-1}aba^{-1},$$
$$bab^{-1}a^{-1}, \qquad ab^{-1}a^{-1}b, \qquad b^{-1}a^{-1}ba, \qquad a^{-1}bab^{-1}.$$

Example 92. Describe the topological spaces represented by their plane models below, and find a word for the model. Of the vertices in the original diagram, indicate which ones are identified in the resulting space.

(a)

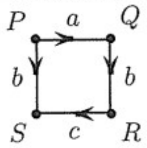

(b)

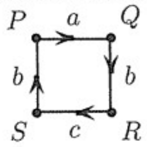

(c)

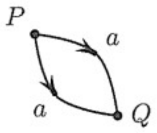

Solution:

(a) Only the left and right sides get glued together. The top and bottom remain as boundaries. This is a surface-with-boundary homeomorphic to the cylinder $\mathbb{S}^1 \times \mathbb{I}$. One word for the model is $abcb^{-1}$. The vertices are identified in pairs: $P \sim Q$ and $R \sim S$.

(b) Again the left and right sides are glued, but this time with a twist. The result is a Möbius strip, as shown in Figure 4.8. One word for the model is $abcb$. The vertices are again identified in pairs, but in the opposite sense: $P \sim R$ and $Q \sim S$.

(c) The opposite sides are identified much like the the two edges of a purse when they get zippered together. This plane model defines the sphere $\mathbb{S}^2$, with corresponding word aa^{-1}. Points P and Q are not identified and may be associated with the north and south pole of the sphere, respectively.

Plane Models and Connected Sum

Suppose we have two plane models representing spaces $A, B \in \mathcal{S}_c$. A plane model for $A\#B$ can be found by the following procedure:

1. Choose a vertex in the plane model for A, and draw a loop at A within the model. Label the loop with a letter not already used in the models for A or B, let's say q. Do the same in the plane model for B, labeling the loop with the same letter, q.

2. Remove the region inside each loop. Then separate each loop at the vertex, thereby creating plane models for $A \setminus U_1$ and $B \setminus U_2$ (where U_1 and U_2 are as in Definition 5.1.3). Now q labels a new edge of each plane model.

3. Join together the two plane models along the edge q. This edge q is now interior to the plane model and so can be erased.

Example 93. Find a plane model and corresponding word for $2\mathbb{T}$.
Solution: The steps are illustrated below. **Caution:** It is important to give distinct labels to the edges of each space.

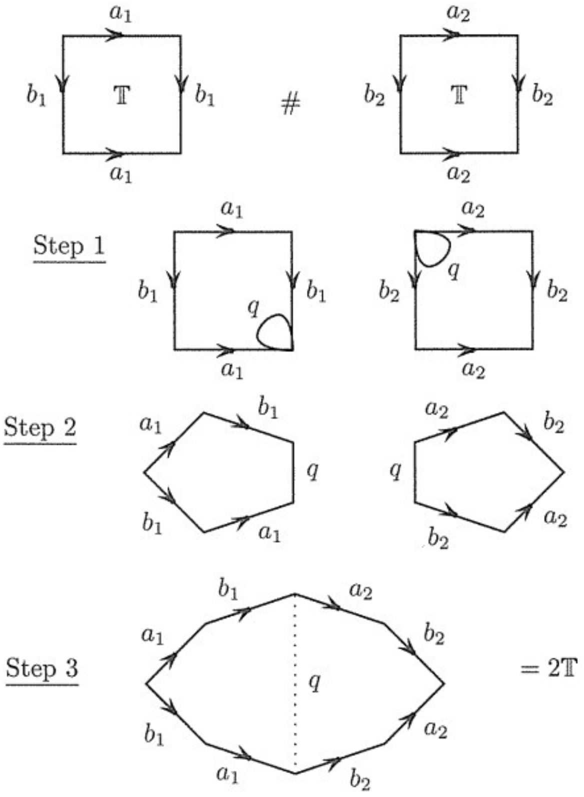

Beginning at the leftmost vertex and going clockwise, a word for $2\mathbb{T}$ is: $a_1b_1a_2b_2a_2^{-1}b_2^{-1}a_1^{-1}b_1^{-1}$ (there are many other equivalent words).

It is interesting to see how a given representation of a surface relates to the plane model for the surface. For example, how does Figure 5.6 correspond to the octagonal plane model for $2\mathbb{T}$ found in Example 93? The key idea is to find *nonseparating* loops – loops that do not divide the space into disconnected pieces. Some spaces like the sphere have no nonseparating loops, while on the torus $\mathbb{T}$, one can find two loops that can be cut simultaneously and not separate the space (e.g., the loops a and b in Figure 5.12). It turns out that the n-holed torus has a set of $2n$ simultaneously nonseparating loops.

Example 94. By cutting along nonseparating loops, a_1, a_2, b_1, and b_2, Figure 5.13 illustrates how the standard embedding of $2\mathbb{T}$ within $\mathbb{R}^3$ corresponds to its plane model. Keep in mind that the cuts are not actually changing the topology of the surface because we still keep track of the identifications.

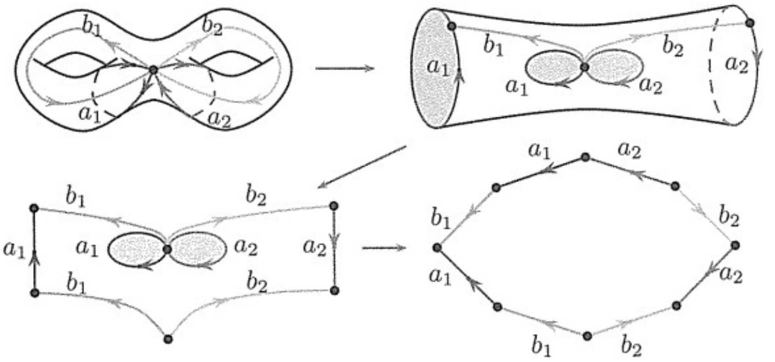

Figure 5.13: Cutting $2\mathbb{T}$ along nonseparating loops to obtain the plane model. First cut along a_1 and a_2 to create a tube with a pair of holes near the middle. Then cut lengthwise along b_1 and b_2 to unwrap the tube, finally opening up the figure into an octagon plane model for $2\mathbb{T}$.

Word Arithmetic

The construction given above for the connected sum of two plane models naturally suggests an operation on words themselves. If $abc\dots$ is a word for a space A and $zyx\dots$ is a word for B *having different letters than the word for* A, then the concatenation $abc\dots zyx\dots$ is a word for $A\#B$. Furthermore, there are certain rules for rearranging words to make equivalent words (i.e., representing the same space). We will use the notation $\equiv$ to indicate that two words are equivalent.

- For convenience, we often write aa as a^2 and $a^{-1}a^{-1}$ as a^{-2}.
- The letters in words do not **commute** ($abcd \neq bacd$); however, *cyclic permutations* are permissible. A **cyclic permutation** of a word is the result of moving one letter from the start of the word to the end, or vice versa. Thus

$$aba^{-1}b^{-1} \equiv ba^{-1}b^{-1}a \equiv a^{-1}b^{-1}ab \equiv b^{-1}aba^{-1}.$$

  This move is equivalent to picking a different starting vertex, or rotating the plane figure.

- If a letter and its inverse appear adjacent to one another, then they *cancel* each other out, *except* in the case of the sphere $\mathbb{S}^2$, whose smallest word has only two letters, aa^{-1}: canceling out these two would result in an *empty word*, which cannot be drawn as a plane model. For example,

$$abc^{-1}dd^{-1}ca^{-1}b^{-1} \equiv abc^{-1}ca^{-1}b^{-1} \equiv aba^{-1}b^{-1}.$$

- If X is a word, the inverse word X^{-1} is obtained by reversing the order and inverting each letter. For example,

$$(abc^{-1}d)^{-1} = d^{-1}cb^{-1}a^{-1}.$$

  This has the effect of choosing to read the labels in the opposite direction, or flipping over the plane model, so the resulting space is not affected: $X^{-1} \equiv X$.

- There is a *split-and-rejoin* operation, which is a useful tool in the classification of surfaces. Choose a location in the word in which to insert a new edge and its inverse, then split the word between the new letter and its inverse to obtain two words (which together still describe a single space). Any equivalence operation mentioned above may then be applied to each subword separately. Join the words again by concatenating at a letter-inverse pair.

$$\begin{array}{rll} XY \equiv Xqq^{-1}Y \equiv & Xq, & q^{-1}Y \\ & \equiv & \equiv \\ & Ua, & a^{-1}V \equiv Uaa^{-1}V \equiv UV \end{array}$$

  This operation corresponds to a cut in the plane model represented by XY by a curve that separates all edges in X from those in Y.

Example 95. The **Klein bottle** is defined as a quotient space by

$$\mathbb{K} = \mathbb{I}^2/\{(x,0) \sim (x,1),\ (0,y) \sim (1,1-y)\}.$$

Draw a plane model representing $\mathbb{K}$, and then show that the Klein bottle has a plane model whose word has the form x^2y^2.

Solution: According to the definition of $\mathbb{K}$, the top and bottom edges are identified in the same direction; however, the sides are identified with a *twist.* A word corresponding to this plane model is $aba^{-1}b$. Consider the following sequence of moves, along with their depictions in the plane model. In each picture, we have circled the starting vertex to make it easier to identify the corresponding word. The direction of the word is clockwise in each diagram.

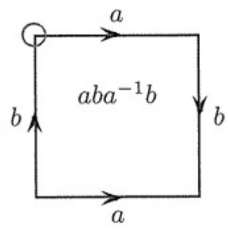

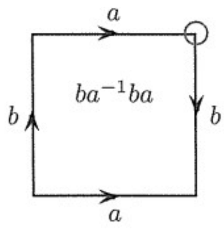

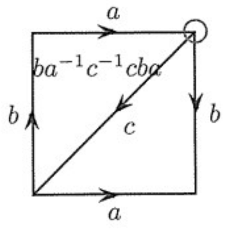

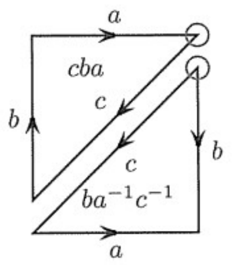

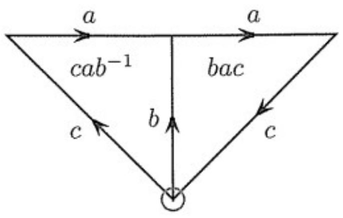

Note that in the fifth diagram above, the triangle $ba^{-1}c^{-1}$ has been flipped to create cab^{-1}, while the word cba has been cyclically permuted to bac, allowing the sides labeled b to be joined together with matching arrow directions. The result is $cab^{-1}bac \equiv caac \equiv aacc = a^2c^2$, having the desired form.

Correspondence between Plane Models and Surfaces

Not every plane model diagram corresponds to an element of $\mathcal{S}_c$. For example, if there are three or more edges labeled by the same letter, then no neighborhood of any point on the common edge would be homeomorphic to a disk. Figure 5.14 illustrates this issue in a space called the *dunce cap*. Any neighborhood of the point x fails to be homeomorphic to a disk; the neighborhood could be described as a "disk with flap."

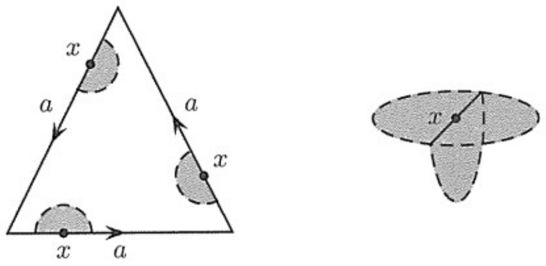

Figure 5.14: The dunce cap.

On the other hand, it can be shown that any plane model diagram in which no letter appears more than twice *does* correspond to a surface or a surface-with-boundary. For simplicity, we will prove the result when it applies to a surface;

the necessary modifications in the case of a surface-with-boundary should be clear.

Theorem 5.2.1. *Suppose X is a plane model diagram in which each letter appears exactly twice. Then X represents a compact surface.*

Proof. We must show that each point of the plane model has an open neighborhood U homeomorphic to an open disk. There are three cases to consider. It may be helpful to refer to Figures 5.15 and 5.16. In what follows, when we say *interior* to X, we mean interior to the plane model diagram before identifications are made along the boundary.

Case I. Suppose $x \in X$ is in the interior of the plane model. Then an open neighborhood U of x may be chosen small enough to lie entirely interior to X, and U is clearly homeomorphic to an open disk.

Case II. Suppose $x \in X$ is on a boundary edge of the plane model, but not at a vertex. Let the edge have label a. Then, by assumption, there is exactly one other edge labeled a. Since those two edges are identified in X, a disk neighborhood U surrounding x exists. In terms of the plane model, it is equivalent to cut and paste a small portion of the model to show how U connects across the edge a.

Case III. Suppose $x \in X$ is a vertex of the plane model. This is the trickiest case because x may be identified to many other vertices on the plane model, and we must build up the disk neighborhood from portions of the model surrounding all vertices equivalent to x. A priori, we do not know which other vertices belong to the equivalence class $[x]$, but the procedure below locates them all.

First suppose that the word for X contains no adjacent inverse pair (i.e., $\ldots aa^{-1} \ldots$ or $\ldots a^{-1}a \ldots$). If it does, then X can be simplified by zippering along the inverse pair. For each edge e, let e^- be a chosen point on that edge near its tail vertex, and let e^+ be on the edge near its head vertex. We shall take a

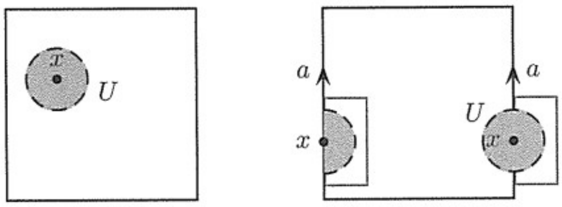

Figure 5.15: Open disk neighborhoods around an interior point and edge point of a plane model. The diagram on the right shows how a portion of the plane model can be cut and pasted to form a disk neighborhood around x.

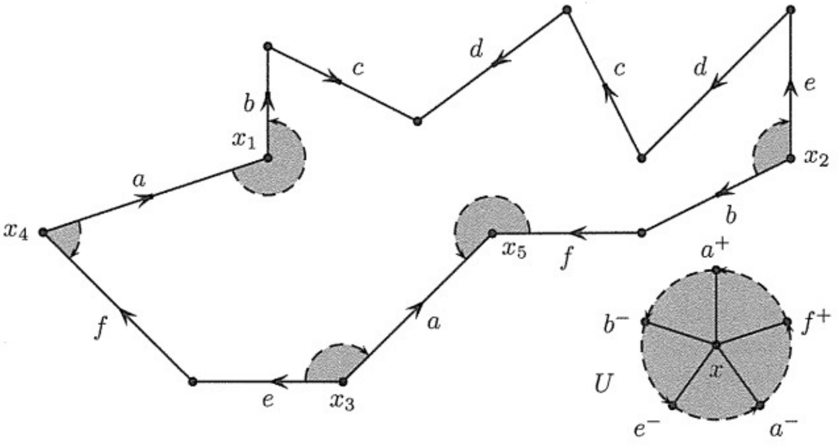

Figure 5.16: Open disk neighborhood U around a vertex x, assembled from regions surrounding all equivalent vertices. This example shows that x is enclosed by a curve containing (in order) $a^+b^-e^-a^-f^+a^+$.

tour of some (or all) of these points that eventually bounds a disk neighborhood of x.

For the purpose of this algorithm, let $x_1 = x$, and let a_1 and a_2 be the two edges meeting at x_1. If x is the tail of a_1, select a_1^-; if x is the head of a_1, select a_1^+. Do the same with respect to a_2. Then draw a small arc in X connecting the selected points. Since edge labels must come in pairs, there must be another edge labeled a_2. Let x_2 be the vertex near to the head (a_2^+) or tail (a_2^-) in the other edge, matching the sign of a_2 that was near x_1. We now know that $x_2 \sim x_1$. Now let a_3 be the label of the other edge meeting x_2. As before, select either a_3^- or a_3^+ depending on how x_2 meets a_3, and draw a small arc in X connecting the selected points. Continue this process until reaching an edge a_n that matches the first edge a_1 *and* meets x in the same way (as head or tail). This completes the tour around vertex x. By cutting and pasting all of these regions together according to the identifications in X, the union of all open regions within the arcs forms an open disk neighborhood of x. ☐

Unfortunately, what we haven't proven yet is that *every* space $X \in \mathcal{S}_c$ *has* a plane model. This depends on the existence of a *triangulation* for X, which is not a trivial thing to prove in general. We define triangulation in the next section.

Toward Triangulation

Does every surface and surface-with-boundary have a plane model? The answer is *yes*, because every space $X \in \mathcal{S}_c$ has a *triangulation*. Let's first define the term. Let $X \in \mathcal{S}_c$. A **triangulation** of X consists of:

- A finite set $V = \{v_1, v_2, \ldots, v_k\}$ of points $v_i \in X$, called the **vertices** of X.
- A finite set $E = \{e_1, e_2, \ldots, e_\ell\}$, called the **edges** of X, such that each $e_i \subseteq X$ is homeomorphic to $\mathbb{I}$, and the endpoints of each e_i are two distinct points from V. If $i \neq j$, then either $e_i \cap e_j = \emptyset$ or there is a unique $v \in V$ such that $e_i \cap e_j = v$. Each edge $e \in E$ may be described by its endpoints: $e = [v_0, v_1]$.
- A finite set $F = \{f_1, f_2, \ldots, f_m\}$, called the **faces** or **triangles** of X, such that each $f_i \subseteq X$ is homeomorphic to a closed disk, and the boundary of f_i consists of three distinct edges from E. Furthermore, if $f_i \cap f_j \neq \emptyset$ for some $i \neq j$, then $f_i \cap f_j$ is either a unique edge $e \in E$ or a unique vertex $v \in V$. Morever, the triangles *cover* X in the sense that $X = \bigcup F$. Each $f \in F$ may be described by its vertices: $f = [v_0, v_1, v_2]$.

Note that by Definition 5.1.2, every edge can belong to at most two triangles. If X is a surface (rather than surface-with-boundary), then every edge belongs to *exactly* two distinct triangles. Figure 5.17 illustrates a few ways of cutting up the torus that are not triangulations, while Figure 5.18 shows a triangulation of the torus. In fact, many different triangulations exist for a given $X \in \mathcal{S}_c$.

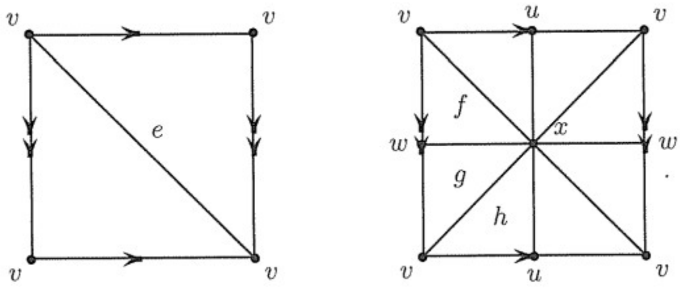

Figure 5.17: Nontriangulations of the torus. *Left*, every edge has both endpoints the same: $e = [v, v]$. *Right*, some pairs of triangles have intersections that are not a single edge or a single vertex. For example, $f \cap g = [w, x] \cup \{v\}$, and $f \cap h = \{x, v\}$.

Theorem 5.2.2. *Every $X \in \mathcal{S}_c$ has a triangulation.*

The proof of Theorem 5.2.2 is not easy. First proved rigorously in 1925 by Radó [Rad25],[5] the result relies on the Jordan Curve Theorem (Theorem 2.4.8), among other things. As a consequence of the theorem, every $S \in \mathcal{S}_c$ has a plane

---

[5] See also Ahlfors and Sario [AS60], Thomassen [Tho92], and Doyle and Moran [DM68].

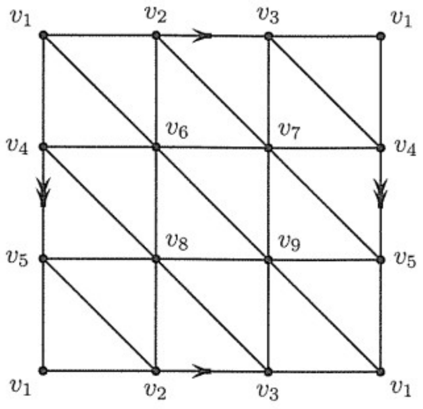

Figure 5.18: A triangulation of the torus using nine vertices. Can you find one that uses fewer vertices?

model. Given a triangulation T for S, with triangles $\{t_1, t_2, \ldots, t_n\}$, build the plane model as follows.

- For each triangle t_k, label its edges and indicate the orientation of each edge so that shared edges receive the same label and orientation.
- Start with $P_1 = t_1$.
- Choose a triangle t_2 that shares an edge with t_1 (which we may do since S is assumed to be a connected surface), and join these triangles at the common edge to form $P_2 = t_1 \cup t_2$. Label the unjoined edges.
- Having constructed $P_k = t_1 \cup t_2 \cup \cdots \cup t_k$, for $k < n$, there must be at least one such triangle t_{k+1} that shares an edge with an existing triangle in P_k. Join t_{k+1} to P_k along that edge, but do not join any other edge of t_{k+1} to P_k. This ensures that P_k remains homeomorphic to a disk.
- Repeat the procedure until all triangles are used. Labels on internal edges should be discarded.

At each step k, the model P_k is a disk with edge labels. Since each edge belongs to at most two triangles, there are no more than two edges in P_k having the same label. Thus P_n is a plane model.

Because of the rather strict rules defining a triangulation, a compact surface can be described unambiguously by a finite set of *combinatorial* data, namely, the list of triangles given in terms of their vertices. In fact, we need only identify each vertex by a number, so a triangle $f = [v_1, v_2, v_3]$ could be represented as $[1, 2, 3]$. This makes it convenient to work with triangulations in computer code, for example.

Exercises

1. Describe the surface or surface-with-boundary represented by each plane model and find a corresponding word, reducing as much as possible. Indicate which vertices are identified in the space.

(a)
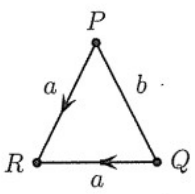

(b)
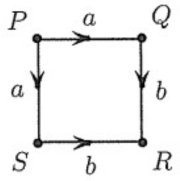

(c)
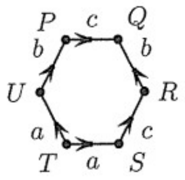

(d)
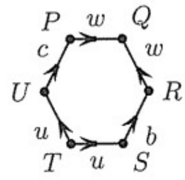

2. Find a plane model and corresponding word representing each of the following spaces.

   (a) $\mathbb{I}^2$ (b) $3\mathbb{T}$ (c) $\mathbb{T}\#\mathbb{K}$ (d) $2\mathbb{K}$

3. Redo Example 93, choosing a different vertex so that the resulting word for $2\mathbb{T}$ is $a_1b_1a_1^{-1}b_1^{-1}a_2b_2a_2^{-1}b_2^{-1}$.

4. Write a word for the n-holed torus, $n\mathbb{T}$.

5. Use words to show that there is no $Q \in \mathcal{S}_c$, such that $\mathbb{T}\#Q = \mathbb{S}^2$. This demonstrates that $\mathcal{S}_c$ is not a *group* with respect to connected sum.

6. Describe the result when any surface S is joined to the square $\mathbb{I}^2$ by connected sum. How is the word of $S\#\mathbb{I}^2$ related to the word of S?

7. Illustrate the algebraic operations of cyclic permutation and cancellation in a series of plane model pictures.

8. Write a word for the plane model shown in Figure 5.16. Show that the vertices fall into two distinct equivalence classes, $[x]$ and $[y]$. It was shown that x is enclosed by a curve labeled $a^+b^-e^-a^-f^+a^+$. Determine a curve bounding a disk neighborhood of y in this figure.

9. Suppose a compact surface has a triangulation with v vertices, e edges, and f faces. Show that $3f = 2e$. Show by example that $3f < 2e$ in a compact surface-with-boundary.

10. Show that any given plane model of a compact surface is equivalent to one in which all of the vertices belong to the same equivalence class. (*Hint:* Suppose there is a vertex y not in the same equivalence class as another vertex x, and suppose y is the head vertex of edge a. Perform the split-and-rejoin operation suggested by the picture below. Why can we assume there is another edge labeled a? How does this procedure eventually eliminate all vertices in the equivalence class of y? What happens if there is only a single vertex in the class of y?)

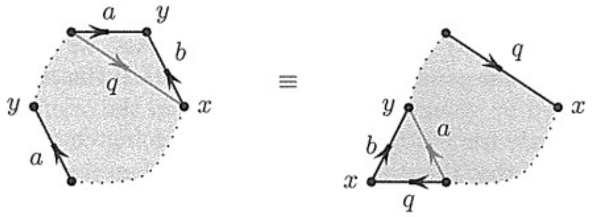

5.3 Orientability

Suppose an ant walks around on the surface of the sphere $\mathbb{S}^2$. We're talking about a three-dimensional ant whose legs touch the outside surface of the sphere as it walks. No matter where the ant walks or for how long it walks, it will never appear on the "inside" of the sphere. To be a little more precise, suppose that the ant is trailing paint as it walks – that way, we can clearly track its path. No matter how long the ant walks, or what kind of meandering path it takes, at the end of the journey we observe that there is paint only on one "side" of the surface (what we call the outside). But now consider an ant walking around on a Möbius strip M. Do this experiment for yourself: Use your pencil or pen to mark your path as you "walk" around on M. If your path is long enough, and stays in roughly the same direction, you will find that you can eventually return to your starting position, but there are now markings on both "sides" of M. This means that M really has only *one* side. The two spaces differ in that $\mathbb{S}^2$ is *orientable* while M is not.

In order to define orientability more precisely, let's first talk about what an orientation is. On the real number line $\mathbb{R}^1$, there are of course only two directions of travel, in order of increasing real numbers (*right*), or decreasing real numbers (*left*). Similarly, there are only two directions of travel on the circle $\mathbb{S}^1$. Let's assume $\mathbb{S}^1$ is represented in the xy-plane as a unit circle. Then at any point of the circle, you may go in order of either increasing angle measure with respect to the x-axis (*counterclockwise*), or decreasing angle measure (*clockwise*). Notice that in each case, we may record the direction using only one bit of information, say, $+1$ for *right* in $\mathbb{R}^1$ and -1 for *left*, or, in the circle, $+1$ for *counterclockwise* and -1 for *clockwise*.

What about an orientation on $\mathbb{R}^2$? This time there are infinitely many distinct directions that a path could take from any given starting point. Suppose

26 04 2020

Pág. 55 (71) Table 5

Properties of Areas and Volumes

C) Books_71_2020/
Introduction to Topology Second Edition (PDF)/
Understanding Topology_A Practical...
pdf

Pág. 167 (173)

Chapter 5 Surfaces.

Pág. 174 (180)

Figure 5.9 $I^2 \times I = I^3$

$I^3 \times I = I^4$

The tesseract is a cube extended through a fourth dimension.

D) Books_71_2020/
Skyrmions Topological Structures.../
Skyrmions ... pdf

Pág. 04 (25)

Figure 1 1

curvilinear polygons on the sphere.

obs: Ver todo livro.

!   )*+,)   )   )-   )   80   >   5   1
5 5   1 ! 55   ;1 +   5!
' #   4   /   5 0   -7   /   ! JJJ

topological protection cannot decay into the trivial solution, meaning that a topologically protected solution is not equivalent to the trivial one. There are many topological solitons in materials (Menzel 2011), such as screw dislocations in crystals, domain walls in ferromagnets, and other spin textures like vortices, merons, bubbles, and skyrmions. Manipulation of the topological solitons may lead to applications in magnetic storage and spintronics (Parkin et al. 2008; Pfleiderer and Rosch 2010; Jung et al. 2011; Zhang et al. 2015).

This chapter introduces different kinds of topological solitons existing in magnets and is organized as follows: Section 1.2 captures the essential features of topology, such as topological equivalence and invariants, surfaces and knots, winding numbers, and applications of homotopy theory in topological defects. Section 1.3 is devoted to the Gibbs free energy for a magnetic system. Sections 1.4 to 1.8 address different kinds of topological spin textures in magnetism, in particular a domain wall, vortex, meron, bubble, and skyrmion.

1.2 Topology

1.2.1 Topological Equivalence and Invariants

Topology studies the properties of topological spaces which are invariant under a continuous transformation, called topological equivalence. One kind of topological equivalence is homeomorphism. Suppose X and Y to be topological spaces. A function f: $X \rightarrow Y$ is called a homeomorphism when it is bijection (i.e., one-to-one and surjective), continuous, and has a continuous inverse function f^{-1} (Armstrong 1983). When such a function exists, X and Y are homeomorphic. As an example, Figure 1.1 shows the homeomorphism of a polyhedron and a sphere. For any point x on the surface of the polyhedron in Figure 1.1a, we can draw a line passing through both the point x and the center of gravity O of the polyhedron, and reach the point $f(x)$ of the sphere with its center of gravity also at O. Doing the same procedure, we can map the polyhedron to the sphere point by point, and the faces of the polyhedron project to curvilinear polygons on the sphere as shown in Figure 1.1b. Similarly, we can project the sphere to the polyhedron point by point as a continuous inverse procedure. Therefore, a polyhedron and a sphere are homeomorphic, and from a topological viewpoint they are the same.

Another example of homeomorphism is shown in Figure 1.2. In topology, these four figures are considered to be the "same space" (Armstrong 1983). A special homeomorphism from space (c) to space (d) is as follows: the points of space (c) are specified by cylindrical

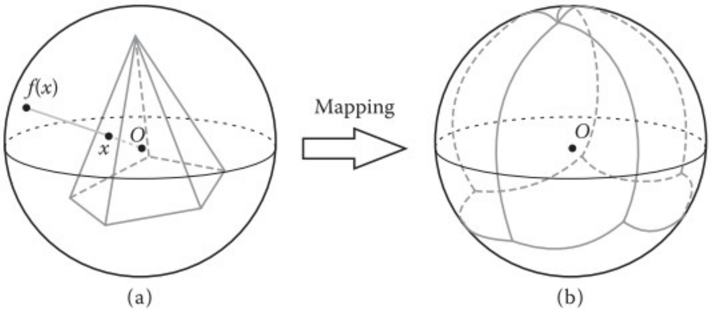

FIGURE 1.1 Continuous mapping from a regular polyhedron to a sphere. (a) The regular polyhedron with the centre of gravity at point O before mapping. (b) After mapping, the faces of the polyhedron becomes curvilinear polygons on the sphere.

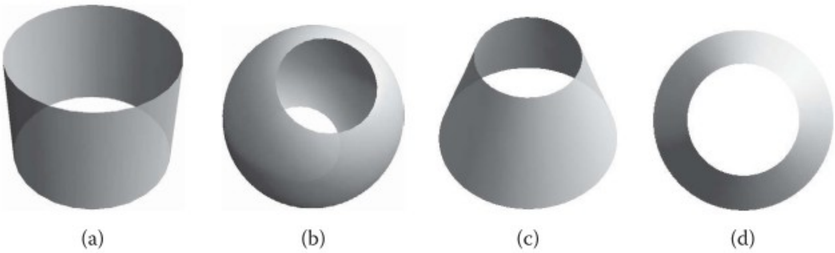

FIGURE 1.2 (a) A cylinder with finite height, excluding the two circles at the ends. (b) A sphere with two holes at the north and south poles. (c) A frustum given by the equation $z = -3\sqrt{x^2 + y^2} + 10$. (d) An annulus specified by $2 < r < 4$.

polar coordinates (r, θ, z) and those of space (d) by plane polar coordinates (r, θ). For space (c), when θ is zero, we can obtain a line $3x + z = 10$. Mapping this line to space (d) where θ is zero, we can get a line $\{ (x, y) \mid 2 < x < 4, y = 0\}$. For each value of θ varying from 0 to 2π, doing a similar procedure in a continuous manner, we can project space (c) to space (d).

Another type of topological equivalence is called homotopy equivalence. Suppose that $f, g: X \to Y$ are maps. If there is a map $F: X \times I \to Y$ such that $F(x, 0) = f(x)$ and $F(x, 1) = g(x)$ for all points $x \in X$, f is homotopic to g (Armstrong 1983), as shown in Figure 1.3a. The equivalence class of loop f is denoted by $[f]$. The set of homotopy classes of loops in X based at x forms the fundamental group under the multiplication $[f]\circ[g] \equiv [f \circ g]$ and is written as $\pi_1(X, x)$ (Braun 2012). The fundamental group may either be Abelian or non-Abelian. Figure 1.3c shows an example of a space with a non-Abelian fundamental group (Mermin 1979). Though loops f and g are freely homotopic, they are not homotopic at the base point, because one cannot deform the loop f into the loop g while holding

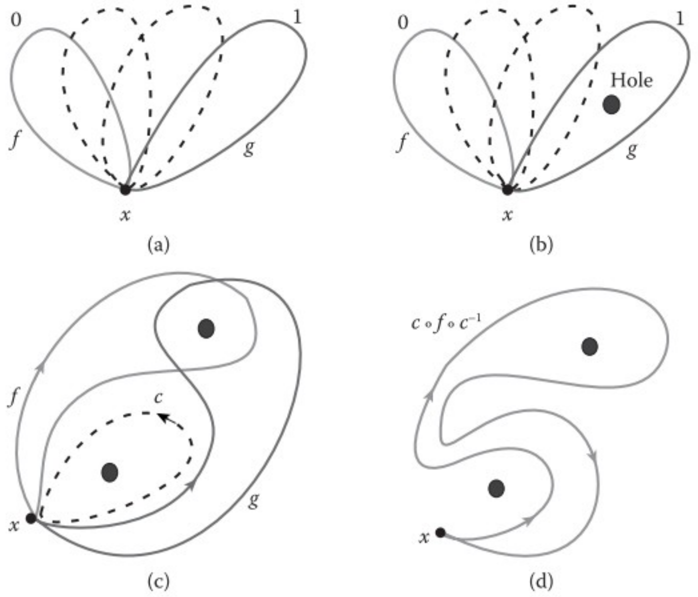

FIGURE 1.3 (a) Two loops f and g are homotopic. (b) f and g are homotopically inequivalent. (c) Loops f and g are not homotopic at the base point x even though they are freely homotopic. (d) A loop $c \circ f \circ c^{-1}$ is homotopic at x to g, where c is the dashed loop in (c).

the base point x. Figure 1.3d shows that g and $c \circ f \circ c^{-1}$ are homotopic. Since f is not homotopic at x to g, $c \circ f$ is not homotopic to $f \circ c$, thus the fundamental group is non-Abelian.

Homotopy theory is used in description and classification of topological defects. A topological defect exists when the mapping from real space S^n to an order-parameter space S^m cannot be shrunk into a single point. In other words, topological defects exist when $\pi_n(S^m) \neq 0$. The nth homotopy group of the n-sphere is isomorphic to the group of integers $\pi_n\left(S^n\right) \cong \mathbb{Z}$ for all $n \geq 1$ (Braun 2012). Therefore, topological defects corresponding to $\pi_n(S^n) \neq 0$ can be characterized by the winding number.

A topological invariant, such as a geometric property of a space, the Euler number, or a group constructed from the space, is a property of a topological space which is preserved by a homotopy (Armstrong 1983). That is, if a space X possesses a topological property, every space topologically equivalent to X possesses that property. According to the property, if a topological invariant of two spaces are calculated to be different, we can conclude that the two spaces cannot continuously be deformed into each other. In topology, a common problem is to develop ways to decide whether two topological spaces are topologically equivalent or not.

1.2.2 Surfaces and Knots

Topology has to deal with the topological equivalence of spaces. Spaces with bounded configurations occurring naturally in Euclidean space are interesting, such as the unit circle, torus, Möbius strip, Klein bottle, and so on. Figure 1.4 shows some interesting examples existing in three-dimensional (3D) space. The former two spaces are orientable, which have consistent 'orientation' over the entire surfaces, while the latter two are nonorientable.

The Möbius strip, as an important two-dimensional (2D) object in topology, has only one side and one edge and therefore arouses much attention. To prove that the Möbius strip has only one side, starting from a central point of the "edges" of the Möbius strip and drawing a line along the strip's surface until you return to the starting point without lifting the pen, you will find the line is continuous, which proves that the strip has only one side. Cutting the Möbius strip along the center line yields an amazing result: a single strip with a 2π twist, rather than two separate strips. The long strip is not a Möbius strip anymore because it has two edges. Interestingly, if you cut around the Möbius strip one-third of the way from the edge, it creates two strips: one is a Möbius strip with one-third of the width and the same length of the original one, and the other is a longer non-Möbius strip with one-third of the width and twice the length of the original strip.

A drastic change such as tearing apart or gluing parts together can destroy the topological equivalence of a space (Sarma et al. 2006). For example, gluing two Möbius strips together along their edges can produce the Klein bottle, but the Möbius strip is not topologically equivalent to the Klein bottle. This procedure cannot be done in ordinary 3D Euclidean space without creating self-intersections but can be represented without crossing itself in four-dimensional (4D) space (Martin et al. 2010).

In topology, apart from the various surfaces, knots are also interesting objects. A knot is a subspace of 3D Euclidean space which is not homeomorphic to the circle (Armstrong 1983). Figure 1.5 shows a few

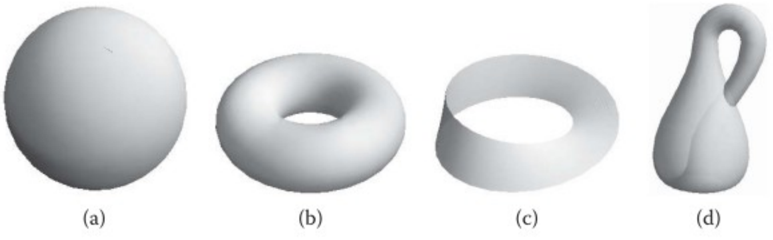
(a) (b) (c) (d)

FIGURE 1.4 Some surfaces in Euclidean space: (a) sphere; (b) torus; (c) Möbius strip; and (d) Klein bottle.

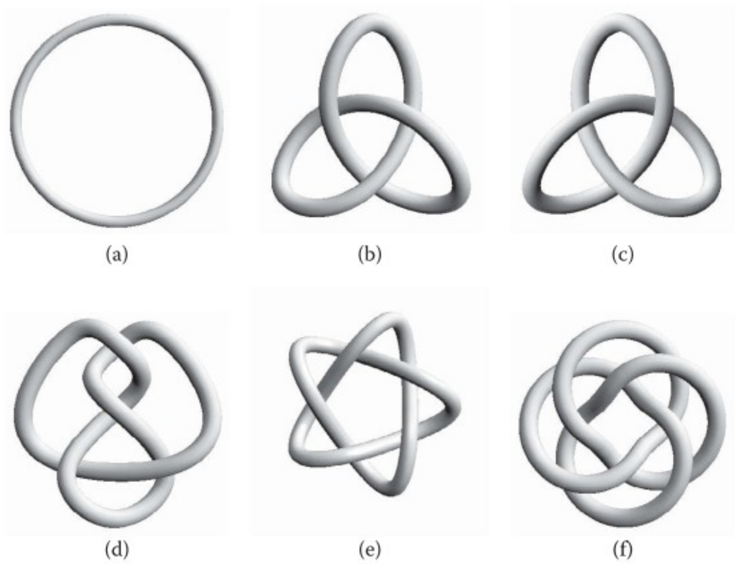

(a) (b) (c)

(d) (e) (f)

FIGURE 1.5 Different knots in three-dimensional Euclidean space: (a) unknot; (b) trefoil with left-handed chirality; (c) trefoil with right-handed chirality; (d) figure of eight; (e) pentafoil; and (f) torus knot.

examples of knots. The unit circle is called a trivial knot or unknot (Figure 1.5a). We cannot convert any of the knots in Figure 1.5b, d through f into the trivial one, nor any one into another simply by wobbling the string around (Armstrong 1983). To realize the transformation, we must cut the string and make up the knot we need, which means in some sense that these knots are all different.

Two knots k_1 and k_2 are topologically equivalent if there is a homeomorphism f of 3D Euclidean space which satisfies that $f(k_1) = k_2$ (Armstrong 1983). Reflection in a plane as a good homeomorphism can transform a knot into its mirror image, but a chiral knot is an exception which is not equivalent to its mirror image, as shown in Figure 1.5b and c. These two trefoil knots have opposite chiralities, left-handed and right-handed, which are mirror images of each other. But, these two knots are not equivalent to each other, that is, they are not amphichiral. The chirality of a knot is a knot invariant.

1.2.3 Winding Numbers

In mathematics, the winding number of a closed planar curve for a given point is the total number of times the curve travels around the point (Mcintyre and Cairns 1993), which is an integer. The sign of the

26 04 2020

Pág. 55 (71) Table 5

Properties of Areas and Volumes

C) Books_71_2020/
Introduction to Topology Second Edition (PDF)/
Understanding Topology_A Practical... pdf

Pág. 167 (173)

Chapter 5 Surfaces.

Pág. 174 (180)

Figure 5.9 $I^2 \times I = I^3$

$I^3 \times I = I^4$

The tesseract is a cube extended through a fourth dimension.

D) Books_71_2020/
Skyrmions Topological Structures.../
Skyrmions ... pdf

Pág. 04 (25)

Figure 1 1

curvilinear polygons on the sphere.

obs: Ver todo livro.

26 04 2020

Important: A Skyrmion is a tiny region of reversed magnetization – quasi particles since they are not present except in a magnetic state and also give rise to physics that cannot be described by Maxwell's equations.

E) Books_71_2020/
Introduction to Topology Second Edition (PDF)/
0486406806.pdf

Pag. 16 (24)

§4. Products of Metric Spaces

Important:

4.2 A ball, with metric

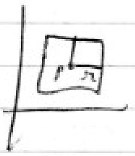

4.3 A ball, with metric

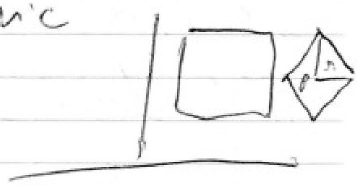

4.2 = square
4.3 = diamond

151) 20200426

A) Books_72_2020/Flat and Curved Space-Times/
Foundations of Space and Time_
Reflections on... pdf

9. Determine the interior, the closure, the limit points, and the isolated points of each of the following subsets of $\mathbb{R}$:
   (a) the interval $[0,1)$,
   (b) the set of rational numbers,
   (c) $\{m + n\pi : m \text{ and } n \text{ positive integers}\}$,
   (d) $\left\{\frac{1}{m} + \frac{1}{n} : m \text{ and } n \text{ positive integers}\right\}$.
10. Let f be a real-valued function on $\mathbb{R}$. Show that there exist $M > 0$ and a nonempty open subset U of $\mathbb{R}$ such that for any $s \in U$, there is a sequence $\{s_n\}$ satisfying $s_n \rightarrow s$ and $|f(s_n)| \leq M$, $n \geq 1$.

4. PRODUCTS OF METRIC SPACES

Let $(X_1,d_1),\ldots,(X_n,d_n)$ be metric spaces. The product set $X = X_1 \times \cdots \times X_n$ consists of all n-tuples $(x_1,\ldots,x_n)$, where $x_k \in X_k$, $1 \leq k \leq n$. There are various ways of making X into a metric space. One possible choice of a metric for X is given by

$$d(x,y) = [d_1(x_1,y_1)^2 + \cdots + d_n(x_n,y_n)^2]^{1/2}, \tag{4.1}$$

where $x = (x_1,\ldots,x_n)$ and $y = (y_1,\ldots,y_n)$. This metric on the product stems from the formula for the metric on $\mathbb{R}^n = \mathbb{R} \times \cdots \times \mathbb{R}$ in terms of the metric on the component spaces $\mathbb{R}$. An open ball in $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ with the metric (4.1) is simply a disc.

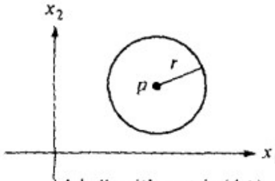

A ball, with metric (4.1)

There are other possible choices of a metric for X, which are convenient for certain problems. Two of them are given by

$$d(x,y) = \max(d_1(x_1,y_1),\ldots,d_n(x_n,y_n)), \tag{4.2}$$

$$d(x,y) = d_1(x_1,y_1) + \cdots + d_n(x_n,y_n). \tag{4.3}$$

Open balls in $\mathbb{R} \times \mathbb{R}$ with the metrics (4.2) and (4.3) are squares and diamonds, respectively. In dealing with many problems, it is not important which metric is

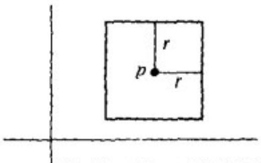

A ball, with metric (4.2)

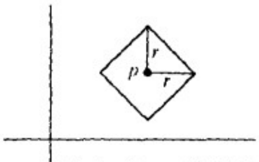

A ball, with metric (4.3)

introduced in the product space, so long as the metric behaves reasonably. One property that each of the above metrics enjoys is the following.

(4.4) A sequence $\{x^{(j)} = (x_k^{(j)})\}_{j=1}^{\infty}$ converges to $x = (x_1, \ldots, x_n)$ in X if and only if for each k the sequence of component entries $\{x_k^{(j)}\}_{j=1}^{\infty}$ converges to x_k in X_k.

The following result shows that all metrics satisfying (4.4) determine the same family of open sets in X, and it describes the structure of these open sets explicitly.

4.1 Theorem: Suppose that d is a metric on $X = X_1 \times \cdots \times X_n$ that satisfies (4.4). Then the open sets in (X,d) are the unions of product sets of the form $U_1 \times \cdots \times U_n$, where U_j is an open subset of X_j, $1 \leq j \leq n$.

Proof: First note that if E_j is a closed subset of X_j, $1 \leq j \leq n$, then $E_1 \times \cdots \times E_n$ is closed in X. Indeed, if $\{x^{(i)}\}_{i=1}^{\infty}$ is a sequence in $E_1 \times \cdots \times E_n$ that converges to $x \in X$, then the coordinate sequences $\{x_k^{(i)}\}_{i=1}^{\infty}$ converge to x_k, $1 \leq k \leq n$. Since $x_k^{(i)} \in E_k$, also $x_k \in E_k$ and $x \in E_1 \times \cdots \times E_n$. It follows that $E_1 \times \cdots \times E_n$ is closed, as asserted.

Now suppose that U_k is an open subset of X_k, $1 \leq j \leq n$. By the preceding remarks, the set

$$X_1 \times \cdots \times X_{k-1} \times (X_k \backslash U_k) \times X_{k+1} \times \cdots \times X_n$$

is closed in X. Its complement

$$V^{(k)} = X_1 \times \cdots \times X_{k-1} \times U_k \times X_{k+1} \times \cdots \times X_n$$

is then open in X. Hence

$$U_1 \times \cdots \times U_n = V^{(1)} \cap \cdots \cap V^{(n)}$$

is open in X and any union of such products is open in X.

To complete the proof, it suffices to show that every open subset of X is a union of such product sets. For this, it suffices to show that if U is open in X and if $x \in U$, then there exist open sets U_k in X_k, $1 \leq k \leq n$, such that $x \in U_1 \times \cdots \times U_n \subseteq U$. We will argue by contradiction.

Let $x \in U$ and suppose that x is contained in no such product set. In particular, for each positive integer m, the product of open balls $B(x_1;1/m) \times \cdots \times B(x_n;1/m)$ is not contained in U. Hence there exists $x^{(m)} \in X \backslash U$ such that

$$d_k(x_k^{(m)},x_k) < 1/m, \qquad 1 \leq k \leq n.$$

Evidently, the sequence $\{x_k^{(m)}\}_{m=1}^{\infty}$ converges to x_k, $1 \leq k \leq n$. By (4.4), $x^{(m)}$ converges to x. Since $X \backslash U$ is closed, x belongs to $X \backslash U$. This contradiction establishes the theorem. □

Another property of the metrics (4.1) to (4.3) is that

$$(4.5) \qquad d_k(x_k,y_k) \leq d(x,y), \qquad x,y \in X,\ 1 \leq k \leq n.$$

Condition (4.5) neither implies nor is implied by condition (4.4). It is clear though that, if a metric d on X satisfies (4.5) and if $\{x^{(m)}\}_{m=1}^{\infty}$ converges in (X,d) to x, then the component sequences $\{x_k^{(m)}\}_{m=1}^{\infty}$ converge to x_k, $1 \le k \le n$.

4.2 Theorem: Let $(X_1,d_1),\ldots,(X_n,d_n)$ be complete metric spaces. Let d be a metric on $X = X_1 \times \cdots \times X_n$ that satisfies (4.4) and (4.5). Then (X,d) is complete.

Proof: Let $\{x^{(j)}\}_{j=1}^{\infty}$ be a Cauchy sequence in X. The estimate (4.5) yields

$$d_k(x_k^{(i)},x_k^{(j)}) \le d(x^{(i)},x^{(j)}).$$

Since the right-hand side tends to 0 as $i,j \to \infty$, $\{x_k^{(i)}\}_{i=1}^{\infty}$ is a Cauchy sequence in X_k. Consequently there exists a point $x_k \in X_k$ such that $x_k^{(i)} \to x_k$ as $i \to \infty$. Set $x = (x_1,\ldots,x_n)$. By (4.4), $x^{(i)} \to x$ as $i \to \infty$. Hence X is complete. □

4.3 Corollary: The n-dimensional Euclidean space $\mathbb{R}^n$, with the usual metric

$$|x - y| = [(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2]^{1/2}, \qquad x,y \in \mathbb{R}^n,$$

is complete.

In the sequel, we shall refer to the product metric space $X = X_1 \times \cdots \times X_n$ without specifying a metric for X. One can select one's own favorite metric for the product, as long as it satisfies (4.4) and (4.5).

EXERCISES

1. Prove that the functions on $X \times X$ given by (4.1), (4.2), and (4.3) are metrics.
2. For each of the metrics (4.1), (4.2), and (4.3) on the plane $\mathbb{R}^2$, sketch the open ball $B(0;1)$. What inclusion relations hold between these balls?
3. Prove that each of the metrics (4.1), (4.2), and (4.3) has property (4.4).
4. Determine the interior, the closure, the limit points, and the isolated points of each of the following subsets of $\mathbb{R}^2$:
   (a) $\{(x,y) : 0 < x^2 + y^2 < 1\}$,
   (b) $\{(x,0) : x \in \mathbb{R}\}$,
   (c) $\{(m,n) : m \text{ and } n \text{ positive integers}\}$,
   (d) $\left\{\left(\frac{1}{m},\frac{1}{n}\right) : m \text{ and } n \text{ positive integers}\right\}$.
5. Prove that if $\tilde{X}_j$ is the completion of X_j, $1 \le j \le n$, then $\tilde{X}_1 \times \cdots \times \tilde{X}_n$ is the completion of $X = X_1 \times \cdots \times X_n$.
6. Let (X_k,d_k), $1 \le k \le n$, be metric spaces. Let d be a metric on $X = X_1 \times \cdots \times X_n$ such that the open subsets of (X,d) are precisely the unions of sets of the form $U_1 \times \cdots \times U_n$, where U_k is an open subset of X_k, $1 \le$

$k \leq n$. Show that d satisfies (4.4). *Note:* This establishes the converse to Theorem 4.1.

7. Let (X_k,d_k), $1 \leq k < \infty$, be metric spaces and let $X = \Pi_{k=1}^{\infty} X_k$ be their Cartesian product, that is, let X be the set of sequences $(x_1,x_2,\ldots)$, where $x_j \in X_j$ for $1 \leq j < \infty$.
   (a) Show that
   $$d(x,y) = \sum_{j=1}^{\infty} \frac{1}{2^j} \min(1,d_j(x_j,y_j))$$
   is a metric on X.
   (b) Show that a sequence $\{x^{(k)}\}_{k=1}^{\infty}$ converges in X if and only if $\{x_j^{(k)}\}_{k=1}^{\infty}$ converges in X_j for each $j \geq 1$.
   (c) Show that the open sets in X are precisely the unions of sets of the form $U_1 \times \cdots \times U_m \times X_{m+1} \times X_{m+2} \times \cdots$, where $m \geq 1$ and U_j is open in X_j, $1 \leq j \leq m$.

5. COMPACTNESS

In analysis it is often important to know when a property that is valid at each point of a space is in some sense uniformly valid over the space. For instance, we shall be concerned in Section 6 with the question of when a function that is continuous at each point of a space is in fact uniformly continuous on the space. The condition under which such uniformity tends to occur for sets in $\mathbb{R}^n$ is that the set be closed and bounded. However, this condition of being closed and bounded, for sets in an arbitrary metric space, does not guarantee that the set has the properties we desire. There is another condition, called *compactness,* that does guarantee that the set has these properties. There is a certain class of properties that finite sets have trivially, that are retained by compact sets. For instance, it will be seen that every continuous real-valued function on a compact space is bounded and assumes its maximum and minimum values.

Our aim in this section is to define the notion of compactness and to prove an important characterization (Theorem 5.1) of compact metric spaces. Most of the section will be devoted to the proof of Theorem 5.1, which is somewhat involved and conceptually difficult. During the course of the section, we shall return to subsets of $\mathbb{R}^n$ and show that the notion of compactness for a set in $\mathbb{R}^n$ is equivalent to the previously mentioned condition of being closed and bounded. This latter theorem is the celebrated Heine-Borel Theorem, and Theorem 5.1 may be regarded as a powerful extension of the Heine-Borel Theorem. We now formulate the appropriate ideas in a general metric space situation.

A family $\{U_\alpha\}_{\alpha \in A}$ of sets is said to *cover* a set S if S is contained in the union of the U_α's. An *open cover* of a metric space X is a family of open subsets of X that covers X. A metric space X is *compact* if every open cover has a finite subcover. In other words, X is compact if, whenever $\{U_\alpha\}_{\alpha \in A}$ is an open cover of X, there are finitely many of the U_α's, say $U_{\alpha_1},\ldots,U_{\alpha_m}$, such that

$$X \subset U_{\alpha_1} \cup \cdots \cup U_{\alpha_m}.$$

A metric space X is *totally bounded* if for each $\varepsilon > 0$, there exists a finite number of open balls of radius ε that cover X. Our main object in this section is to establish the following theorem. Both the theorem and the ideas entering into the proof are of the utmost importance.

5.1 Theorem: The following are equivalent for a metric space X:

(5.1) X is compact.

(5.2) Every sequence in X has a convergent subsequence.

(5.3) X is totally bounded and complete.

Proof: The proof of Theorem 5.1 will take most of this section. For now, we content ourselves with proving the straightforward implications (5.1) $\Rightarrow$ (5.2) $\Rightarrow$ (5.3).

Suppose that X is compact. Let $\{y_j\}_{j=1}^{\infty}$ be a sequence in X. Suppose that for each $x \in X$, there exists $\varepsilon = \varepsilon(x) > 0$ such that only finitely many terms of the sequence $\{y_j\}$ lie in $B(x;\varepsilon(x))$. The set of open balls $\{B(x;\varepsilon(x)) : x \in X\}$ forms an open cover of X. (Note the use of X as the index set for the cover!) Since X is compact, there is a finite subcover

$$X = B(x_1;\varepsilon(x_1)) \cup \cdots \cup B(x_m;\varepsilon(x_m)).$$

Since y_j belongs to $B(x_i;\varepsilon(x_i))$ for only finitely many indices j, we conclude that there is a grand total of only finitely many positive integers, an absurdity. This contradiction shows that there exists $x \in X$ such that each ball $B(x;\varepsilon)$ contains infinitely many terms of the sequence $\{y_j\}$.

Now we construct a convergent subsequence of $\{y_j\}$ as follows. Choose j_1 so that $y_{j_1} \in B(x;1)$. Since $y_j \in B(x;1/2)$ for infinitely many j, we can choose $j_2 > j_1$ such that $y_{j_2} \in B(x;1/2)$. Continuing in this manner, we find $j_1 < j_2 < \cdots$ such that $y_{j_n} \in B(x;1/n)$. Then $\{y_{j_n}\}_{n=1}^{\infty}$ is a subsequence of $\{y_j\}$ that converges to x. We have proved that (5.1) implies (5.2).

Next, suppose that (5.2) is valid. Let $\{x_j\}$ be a Cauchy sequence in X. By (5.2), $\{x_j\}$ has a convergent subsequence. By Lemma 2.2, the sequence $\{x_j\}$ itself converges. Hence X is complete.

Now assume still that (5.2) is valid and let $\varepsilon > 0$. We wish to cover X by a finite number of open balls of radius ε. Let $y_1 \in X$ be arbitrary. If $B(y_1;\varepsilon) \neq X$, let y_2 be any point in $X \backslash B(y_1;\varepsilon)$. Having chosen $y_1, \ldots, y_n$, we let y_{n+1} be any point in $X \backslash (\cup_{j=1}^{n} B(y_j;\varepsilon))$, provided the union does not coincide with X; otherwise we stop. The points $y_1, y_2, \ldots$ constructed in this manner satisfy

$$d(y_k,y_j) \geq \varepsilon, \qquad 1 \leq j < k.$$

If the procedure for selecting the y_j's does not terminate, then the sequence $\{y_j\}_{j=1}^{\infty}$ has no convergent subsequence. It follows that the choice of some y_{n+1} is impossible, so that X is covered by the n open balls of radius ε and centers $y_1, \ldots, y_n$. In particular, X is totally bounded, and (5.2) implies (5.3).

We shall return later to the proof of Theorem 5.1. First we discuss the notion of total boundedness in more detail and introduce various other notions.

A metric space X is *bounded* if there exists $b > 0$ such that $d(x,y) \leq b$ for all $x,y \in X$.

5.2 Lemma: A totally bounded metric space is bounded.

Proof: Let X be totally bounded. Choose $x_1, \ldots, x_m \in X$ such that the balls $B(x_j;1)$, $1 \le j \le m$, cover X. Let

$$b = 2 + \max\{d(x_j,x_k) : 1 \le j,k \le m\}.$$

An appropriate application of the triangle inequality shows that $d(x,y) < b$ for all $x,y \in X$, so that X is bounded. □

5.3 Lemma: Any subspace of a totally bounded metric space is totally bounded.

Proof: Let Y be a subspace of the totally bounded space X and let $\varepsilon > 0$. Since X is totally bounded, there are points $x_1, \ldots, x_n \in X$ such that the balls $B(x_j;\varepsilon/2)$ cover X. We can arrange the x_j's so that $B(x_j;\varepsilon/2) \cap Y \neq \emptyset$ for the indices $1 \le j \le m$, while $B(x_j;\varepsilon/2)$ does not meet Y for $m + 1 \le j \le n$. Then the balls $B(x_j;\varepsilon/2)$, $1 \le j \le m$, cover Y. Let y_j be any point in $B(x_j;\varepsilon/2) \cap Y$, $1 \le j \le m$. The triangle inequality shows that

$$B(x_j;\varepsilon/2) \subset B(y_j;\varepsilon).$$

Consequently the m open balls $B(y_j;\varepsilon) \cap Y$ in the subspace Y cover Y. □

5.4 Lemma: A subset E of $\mathbb{R}^n$ is totally bounded if and only if E is bounded.

Proof: The forward implication follows from Lemma 5.2. For the reverse implication, suppose that E is bounded. Then E is contained in a cube of the form $T = [-b,b] \times \cdots \times [-b,b]$ for some large $b > 0$. By Lemma 5.3, it suffices to show that T is totally bounded. For this, let $\varepsilon > 0$. Then the balls $B(\varepsilon j;\varepsilon)$ cover T, where $j = (j_1, \ldots, j_n)$ ranges over all integral lattice points of $\mathbb{R}^n$ which satisfy $\varepsilon|j_i| \le 2b$, $1 \le i \le n$. (An integral lattice point of $\mathbb{R}^n$ is by definition a point whose coordinates are integers.) Since there are only finitely many such lattice points, T is totally bounded. □

Now a subspace of $\mathbb{R}^n$ is complete if and only if it is closed, by Theorems 2.3 and 2.4. Consequently a subspace of $\mathbb{R}^n$ is totally bounded and complete if and only if it is closed and bounded. Theorem 5.1 then yields the following important theorem about subsets of $\mathbb{R}^n$.

5.5 Theorem *(Heine-Borel Theorem):* The following are equivalent for a subspace E of $\mathbb{R}^n$.

(5.1)′ E is compact.

(5.2)′ Every sequence in E has a convergent subsequence.

(5.3)′ E is closed and bounded.

Of course, the Heine-Borel Theorem is not established until the proof of Theorem 5.1 is complete.

The following theorem shows that (5.3) implies (5.2), so that the two are equivalent.

5.6 Theorem: Let X be a totally bounded metric space. Then every sequence in X has a Cauchy subsequence.

Proof: We shall use a diagonalization procedure to prove the theorem. It is important that one add this technique to one's bag of tricks.

Let $\{x_j\}_{j=1}^{\infty}$ be a sequence in X. For convenience, we rewrite the sequence in the form $\{x_{1j}\}_{j=1}^{\infty}$. By induction, we construct sequences $\{x_{kj}\}_{j=1}^{\infty}$, $k \geq 2$, with the following properties:

(i) $\{x_{kj}\}_{j=1}^{\infty}$ is a subsequence of $\{x_{k-1,j}\}_{j=1}^{\infty}$, $k \geq 2$.

(ii) $\{x_{kj}\}_{j=1}^{\infty}$ is contained in a ball of radius $1/k$, $k \geq 2$.

Indeed, suppose that $k \geq 2$ and that we already have the sequences $\{x_{ij}\}_{j=1}^{\infty}$ for $i < k$. Let $B_1, \ldots, B_n$ be a finite number of open balls of radius $1/k$ that cover X. Since there are infinitely many indices j and only finitely many balls, there must exist at least one ball, say B_m, such that $x_{k-1,j} \in B_m$ for infinitely many $j \geq 1$. Now let x_{k1} be the first of the $x_{k-1,j}$'s that belong to B_m, let x_{k2} be the second, etc. Then $\{x_{kj}\}_{j=1}^{\infty}$ has properties (i) and (ii).

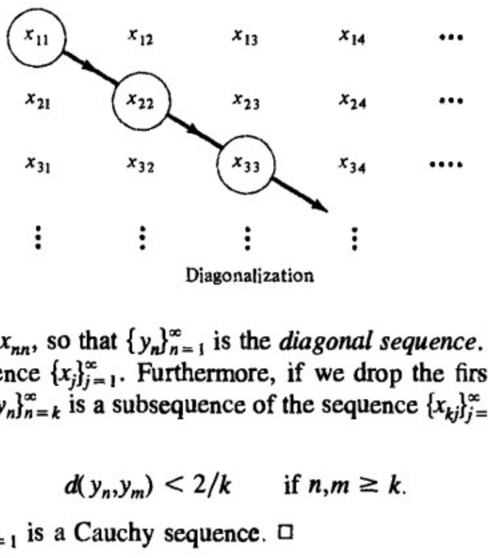

Diagonalization

Now set $y_n = x_{nn}$, so that $\{y_n\}_{n=1}^{\infty}$ is the *diagonal sequence*. It is a subsequence of the original sequence $\{x_j\}_{j=1}^{\infty}$. Furthermore, if we drop the first $k - 1$ terms, the resulting sequence $\{y_n\}_{n=k}^{\infty}$ is a subsequence of the sequence $\{x_{kj}\}_{j=1}^{\infty}$. In particular, (ii) shows that

$$d(y_n, y_m) < 2/k \qquad \text{if } n,m \geq k.$$

It follows that $\{y_n\}_{n=1}^{\infty}$ is a Cauchy sequence. □

A metric space X is *separable* if there is a dense subset of X that is countable. In other words, X is separable if and only if there is a sequence $\{x_j\}_{j=1}^{\infty}$ in X that is dense in X. (There may be repetitions among the x_j's.)

The real numbers $\mathbb{R}$ form a separable metric space. Indeed, the set $\mathbb{R}_0$ of rational numbers forms a countable dense subset of $\mathbb{R}$. Similarly, $\mathbb{R}^n$ is separable since the set of points of $\mathbb{R}^n$ with rational coordinates is countable.

5.7 Theorem: A subspace of a separable metric space is separable.

Proof: Suppose that X is a metric space with a dense subset $\{x_j\}_{j=1}^{\infty}$ and let Y be a subspace of X. Consider pairs of indices (j,n) such that

$$(*) \qquad B(x_j;1/n) \cap Y$$

is not empty. For such pairs (j,n), let y_{jn} be any point in the intersection (*). The $\{y_{jn}\}$ form a countable subset of Y. We claim that the y_{jn} are dense in Y.

Let $y \in Y$ and let $n \geq 1$. Then there exists j such that $d(x_j,y) < 1/n$. In particular, the intersection (*) is nonempty, so that y_{jn} is defined and $d(x_j,y_{jn}) < 1/n$. Then

$$d(y,y_{jn}) < d(y,x_j) + d(x_j,y_{jn}) < 2/n.$$

Letting $n \rightarrow \infty$, we find that y is adherent to the set $\{y_{jn}\}$, so that the closure of the double sequence is Y. □

5.8 Theorem: A totally bounded metric space is separable.

Proof: Let n be a positive integer. Then there exist $x_{n1},\ldots,x_{n,m_n}$ such that the open balls with centers at the x_{nj} and radii $1/n$ cover X. The family $\{x_{nj} : 1 \leq j \leq m_n, 1 \leq n < \infty\}$ is then a countable subset of X. For each $x \in X$ and each integer n, there is an x_{nj} such that $d(x_{nj},x) < 1/n$. Consequently the x_{nj} are dense in X. □

A *base of open sets* for a metric space X is a family $\mathcal{B}$ of open subsets of X such that every open subset of X is the union of sets in $\mathcal{B}$. For instance, the family of open balls forms a base of open sets for a metric space, as does the family of open balls with rational radii. Theorem 4.1 shows that the family of sets of the form $U_1 \times \cdots \times U_n$, where U_j is open in X_j, forms a base of open sets for a product metric space $X_1 \times \cdots \times X_n$.

5.9 Lemma: A family $\mathcal{B}$ of open subsets of a metric space X is a base of open sets if and only if for each $x \in X$ and each open neighborhood U of x, there exists $V \in \mathcal{B}$ such that $x \in V$ and $V \subset U$.

Proof: If $\mathcal{B}$ is a base, the set U above is the union of sets in $\mathcal{B}$, and so the condition of the lemma is evidently valid. Conversely, suppose that the condition given in the lemma holds. Let U be an open subset of X. For each $x \in U$, there is $V_x \in \mathcal{B}$ such that $x \in V_x$ and $V_x \subset U$. Then

$$U = \cup\{V_x : x \in U\}$$

represents U as a union of sets in $\mathcal{B}$, so that $\mathcal{B}$ is a base of open sets. □

A metric space satisfies the *second axiom of countability*, or is *second-countable*, if there is a base of open sets that is at most countable. In other words, X is second-countable if there is a sequence $\{U_n\}_{n=1}^{\infty}$ of open sets such that each open set U is the union of the U_n's it contains.

5.10 Theorem: A metric space is second-countable if and only if it is separable.

Proof: Suppose first that X is a separable metric space. Let $\{x_j\}_{j=1}^{\infty}$ be a dense sequence in X. Consider the family of open sets

$$\mathcal{B} = \{B(x_j;1/n) : j \geq 1,\ n \geq 1\}.$$

Let U be an open subset of X and let $x \in U$. For some $n \geq 1$, we have $B(x;2/n) \subset U$. Choose j so that $d(x_j,x) < 1/n$. Then $x \in B(x_j;1/n)$, and the triangle inequality shows that $B(x_j;1/n) \subset B(x;2/n) \subset U$. Consequently $\mathcal{B}$ is a base of open sets. Since $\mathcal{B}$ is countable, X is second-countable.

Conversely, suppose that X is second-countable. Let $\{U_n\}_{n=1}^{\infty}$ be a sequence of open sets in X that form a base. Let x_n be any point in U_n, $n \geq 1$. Then every nonempty open subset of X contains a point of the sequence $\{x_n\}$, so that the sequence is dense in X, and X is separable. □

The following theorem is the trickiest part of the circle of ideas covered in this section.

5.11 Theorem *(Lindelöf's Theorem):* Suppose the metric space X is second-countable. Then every open cover of X has a countable subcover.

Proof: Let $\{U_\alpha\}_{\alpha\in A}$ be an open cover of X, where A is some index set. Let $\mathcal{B}$ be a countable base of open sets. Let $\mathcal{C}$ be the subset of $\mathcal{B}$ consisting of those sets $V \in \mathcal{B}$ such that $V \subset U_\alpha$ for some α. We claim that $\mathcal{C}$ is a cover of X. Indeed, if $x \in X$, then there is some index α such that $x \in U_\alpha$. By Lemma 5.9, there is $V \in \mathcal{B}$ such that $x \in V$ and $V \subset U_\alpha$. In particular, $V \in \mathcal{C}$, so that $\mathcal{C}$ indeed covers X.

Now for each $V \in \mathcal{C}$, select one index $\alpha = \alpha(V)$ such that $V \subset U_{\alpha(V)}$. Then the sets $\{U_{\alpha(V)} : V \in \mathcal{C}\}$ cover X. Since $\mathcal{B}$ is countable, so is $\mathcal{C}$, so that the $U_{\alpha(V)}$'s form a countable subcover of X. □

Proof of Theorem 5.1 (concluded): As remarked earlier, Theorem 5.6 shows that (5.3) implies (5.2), so that the conditions (5.2) and (5.3) are equivalent.

Suppose that (5.2) and (5.3) are valid. It suffices to show that every open cover of X has a finite subcover.

By Theorem 5.8, X is separable, so that by Theorem 5.10, X is second-countable. By Theorem 5.11 every open cover of X has a countable subcover. It suffices to show then that if $\{U_n\}_{n=1}^{\infty}$ is a sequence of open subsets of X that cover X, then there is an integer m such that $X = U_1 \cup \cdots \cup U_m$.

Suppose that this is not the case, that is, that $U_1 \cup \cdots \cup U_m \neq X$ for all m. For each m, let x_m be any point in $X \backslash (U_1 \cup \cdots \cup U_m)$. By (5.3), the sequence $\{x_m\}$ has a subsequence that converges to some point $x \in X$. Since $x_j \in X\backslash(U_1 \cup \cdots \cup U_m)$ for all $j \geq m$ and since $X\backslash(U_1 \cup \cdots \cup U_m)$ is closed, $x \in X\backslash(U_1 \cup \cdots \cup U_m)$.

Since this is true for all m, $x \in X \setminus (\cup_{k=1}^{\infty} U_k)$ and the U_k do not cover X. This contradiction establishes the theorem. □

We state the following corollary to Theorems 5.1, 5.8, and 5.10 for emphasis.

5.12 Theorem: A compact metric space is separable and second-countable.

EXERCISES

1. Give an example of a totally bounded metric space which is not compact.
2. Give an example of a complete metric space which is not compact.
3. Show directly that a compact metric space is totally bounded.
4. Let $\{U_\alpha\}_{\alpha \in A}$ be a finite open cover of a compact metric space X.
   (a) Show that there exists $\varepsilon > 0$ such that for each $x \in X$, the open ball $B(x;\varepsilon)$ is contained in one of the U_α's.

Remark: Such an ε is called a *Lebesgue number* of the cover.

   (b) Show that if at least one of the U_α's is a proper subset of X, then there is a largest Lebesgue number for the cover.

5. Prove that the product of a finite number of compact metric spaces is compact.
6. A metric space is a *Polish space* if it is separable and if there is an equivalent metric for which it is complete. Prove that any closed subspace of a Polish space is Polish.

Remark: An important school of Polish mathematicians made deep contributions to the study of point-set topology during the period between the world wars. The terminology "Polish" recognizes the work of the Polish school on this important class of spaces. For more on Polish spaces, see Exercises 6.14 and 6.15.

7. Let $B(S)$ be the metric space of bounded real-valued functions on the set S, with the metric of uniform convergence on S (defined in Section 1). Show that $B(S)$ is separable if and only if S is finite.

   Hint: For any subset T of S, define $\chi_T \in B(S)$ by

$$\chi_T(s) = \begin{cases} 1, & s \in T, \\ 0, & s \notin T. \end{cases}$$

   Then the balls $B(\chi_T;1/2)$ are disjoint.

8. Let (X,d) be a bounded metric space and let $\mathscr{C}$ be the family of nonempty closed subsets of X. Show that

$$\rho(E,F) = \max\left(\sup_{x \in E} d(x,F),\ \sup_{y \in F} d(y,E)\right)$$

   defines a metric on $\mathscr{C}$. Show that $\mathscr{C}$ is compact whenever X is compact.

26 04 2020
X

Important: A Skyrmion is a tiny region of reversed magnetization — quasi particles since they are not present except in a magnetic state and also give rise to physics that cannot be described by Maxwell's equations.

E) Books_71_2020/
Introduction to Topology Second Edition (PDF)/
0486406806.pdf

Pág. 16 (24)

§4. Products of Metric Spaces

Important:

4.2 A ball, with metric

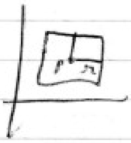

4.3 A ball, with metric

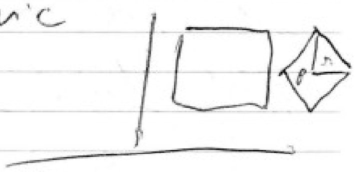

4.2 = square

4.3 = diamond

151) 20200426 —

A) Books_72_2020/ Flat and Curved Space-Times/
Foundations of Space and Time_
Reflections on... .pdf

4 " " $ $; #
MK % ' $ $! ! MK N /
!

! ! " " # $# $ % & # #
! ! ! " " # $$ $#
'()*' ' '+ ' , ! - ! . /
) * 0 1 2 3 4 ! 5 1 2 -! & 6 $

you could begin by clarifying what this is all about, before we plunge into the details? What is the roadmap, so to speak?

Me: To begin with, I will show you that the equations of motion describing gravity in *any diffeomorphism-invariant theory* can be given [2] a suggestive thermodynamic reinterpretation (Sections 2.2, 2.3). Second, taking a cue from this, I can formulate a variational principle for a suitably defined entropy functional – involving both gravity and matter – which will lead to the field equations of gravity [3,4] without varying the metric tensor as a dynamical variable (Section 2.4).

Harold: Suppose I have an action for gravity plus matter (in D dimensions)

$$A = \int d^D x \sqrt{-g} \left[L(R^{ab}_{cd}, g^{ab}) + L_{\text{matt}}(g^{ab}, q_A) \right] \tag{2.1}$$

where L is any scalar built from metric and curvature and L_{matt} is the matter Lagrangian depending on the metric and some matter variables q_A. (I will assume L does not involve derivatives of curvature tensor, to simplify the discussion.) If I vary g^{ab} in the action I will get some equations of motion (see, e.g., [5,6]), say, $2E_{ab} = T_{ab}$ where E_{ab} is[2]

$$E_{ab} = P_a^{cde} R_{bcde} - 2\nabla^c \nabla^d P_{acdb} - \frac{1}{2} L g_{ab}; \qquad P^{abcd} \equiv \frac{\partial L}{\partial R_{abcd}} \tag{2.2}$$

Now, you are telling me that (i) you can give a thermodynamic interpretation to the equation $2E_{ab} = T_{ab}$ just because it comes from a scalar Lagrangian and (ii) you can also derive it from an entropy maximization principle. I admit it is fascinating. But why should I take this approach as more fundamental, conceptually, than the good old way of just varying the total Lagrangian $L + L_{\text{matt}}$ and getting $2E_{ab} = T_{ab}$? Why is it more than a curiosity?

Me: That brings me to the third aspect of the formulation which I will discuss towards the end (Section 2.5). In my approach, I can provide a natural explanation to several puzzling aspects of gravity and horizon thermodynamics, all of which have to be thought of as mere algebraic accidents in the conventional approach you mentioned. Let me give an analogy. In Newtonian gravity, the fact that inertial mass is equal to the gravitational mass is an algebraic accident without any fundamental explanation. But in a geometrical theory of gravity based on the principle of equivalence, this fact finds a natural explanation. Similarly, I think we can make progress by identifying key facts which have no explanation in the conventional approach and providing them a natural explanation from a different perspective. You will also see that this approach connects up several pieces of conventional theory in an elegant manner.

---

[2] The signature is $-+++$ and Latin letters cover spacetime indices while Greek letters run over space indices.

26.04.2020

Pág. 03(24)

L_{matt} is the matter Lagrangian depending on the metric and some matter variables ψ_A.

(I will assume L does not involves derivatives of curvature tensor, to simplify the discussion).

B) Flat and Curved Space-Times, Second Edition (2001).pdf

Pág. 49(57)

3. Measurements in flat space-times

Pág. 122 (130)

4. The Lorentz transformation and invariant interval.

Pág. 126(134)

Fig. 4.2

↑ The relationship between the two coordinate system.

C) Relativistic cosmology (2012 Cambridge).pdf

Part 2 Relativistic Cosmological Models

4 '()*' ' '+ ' , ! 1 ! - . 3 !
7! -!
& $6 8 # -
& # $. 9 : 1 1
& ;< 6$ 1 ! -

Einstein's new theory. Using the Schwarzschild solution to describe the spherically symmetric gravitational field of the Sun, Einstein was able to determine the orbit of Mercury according to general relativity, and amazingly the prediction gave a rotation of 43 seconds, in excellent agreement with observation. This was the first experimental test of the theory and provided very compelling evidence in support of it. The test is particularly compelling because the theory was not designed specifically to meet this challenge—it just turned out that it did so, after the theory had been fashioned on the basis of fundamental considerations by Einstein on the nature of space, time, and gravitation.

More recently, the general relativistic prediction of perihelion precession has been confirmed by observations of the binary pulsar discovered by Hulse and Taylor in 1974. This system, which will be discussed in more detail in Section 5.11, consists of two very compact stars in a very tight orbit around each other. The perihelion precession is orders of magnitude larger than that of Mercury; the prediction of about 4 degrees per year agrees closely with the measured value.

Radar time-delay

A way of investigating the curvature of space produced by the Sun, say, is to measure the delay in the travel-time of a radar beam passing near it, as compared with the travel time if the space were flat. Early experiments were performed by sending a radar beam from Earth and measuring the round trip time after it was returned by a reflector on the surface of Venus or Mercury or onboard a Mariner spacecraft. As the path of the radar beam moved nearer to the Sun as the relative positions of the Earth and the reflector changed, the travel time varied (see Fig. 5.35). In a more recent experiment by Shapiro in 1976, the radar travelled to Mars and was sent back by reflectors both on the surface of the planet and in a spacecraft in orbit around it. The round trip time for signals passing near the Sun was measured and found to agree well with the values calculated from general relativity. Because of the curvature of space, the distance was found to be larger by about 37 km out of a total distance of 378 million km from Earth to Mars. The radar travel time was about 42 minutes for the round trip.

The Global Positioning System

An intriguing application of, and testing ground for, relativity lies in the setting up of the Global Positioning System (GPS). The basic idea is that an observer should be able to determine his or her position in space and time with extreme accuracy by using signals from a network of satellites. In the current system, there are twenty-four satellites in various orbits around the Earth, arranged such that four or more of them are almost always visible from any place on Earth. Each satellite carries an extremely accurate and stable atomic clock, and signals from this are emitted from the satellite. Ground-based monitoring stations collect data which is processed and re-transmitted to the satellites. The user has a small computer which uses information from the satellites to solve for its position, time,

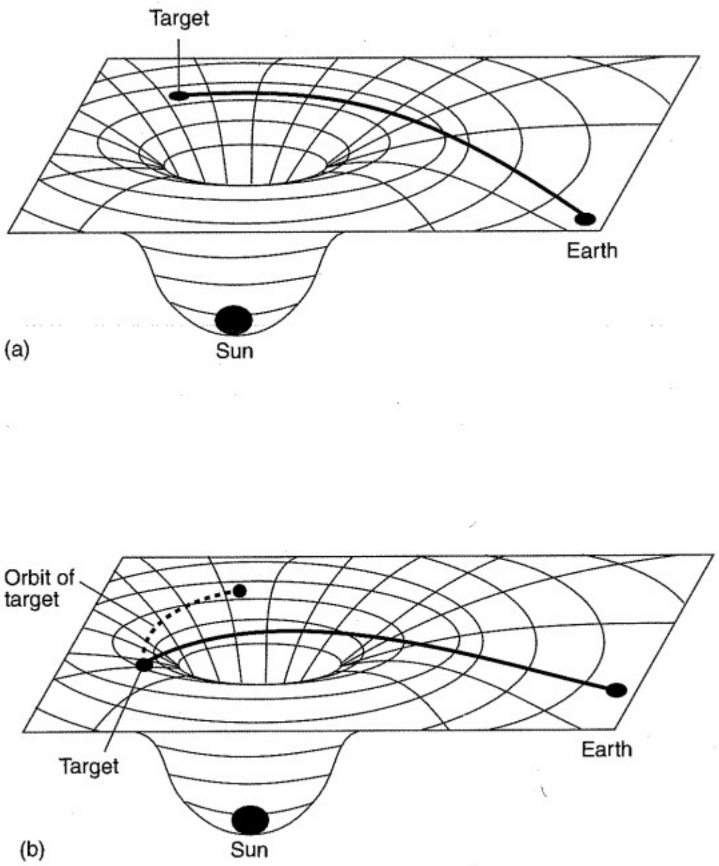

Fig. 5.35 The gravitational field of the Sun produces curvature in space, which is represented here by a 'rubber sheet' picture. (a) When the target planet is far from the Sun, the radar path is on the 'flat' part of the sheet. (b) As the target approaches the Sun, the radar has greater and greater distance to cover because of the 'dip' in the sheet, so the travel time is longer.

and velocity. The accuracy is extremely impressive, and can be as good as 5–10 cm in position.

Although the velocities of the clocks are small and the gravitational fields are weak, relativistic effects like time dilation and gravitational frequency shift would cause errors much larger than possible errors in the accuracy of the cesium clocks used, and so they need to be taken into account. A natural consequence of this is that the GPS provides a way of testing the theory of relativity. Data from the GPS satellites recorded by the TOPEX satellite (in orbit primarily to measure the height of the sea) is providing the first explicit measurements of the periodic part of the combined effect of time dilation and gravitational frequency shifts on an orbiting receiver. Preliminary analysis of the data gives an agreement between theory and experiment to within 2.5 per cent. (For more details, see 'The global positioning system' by T. A. Herring, *Scientific American*, February 1996, 32–38.)

5.10 Gravitational waves

Einstein realized more than 80 years ago that his theory predicted the existence of gravitational waves, but it is only relatively recently that any progress has been made on detecting them. In this section, we shall first look at the nature of gravitational waves and then at possible sources for their emission. In Section 5.11, we shall discuss methods of detecting them, both by experiments with bars and interferometers and also less directly through study of the energy decrease of a system which is best explained by the emission of such waves.

The nature of gravitational waves

Gravitational waves are fluctuations in the metric tensor which describes the curvature of space–time. One can think of them as ripples which travel through space–time, usually as a result of rotation or other changes in the body producing the gravitational field.

We shall look first at a particularly simple way in which Einstein's equations can have wave solutions. This is in the case when the gravitational field in empty space is weak, so that the metric tensor can be written as its flat-space value η^{ij} plus a small perturbation h^{ij}. The 'linearized field equations' are derived from Einstein's equations by ignoring terms in h^{ij}-squared and higher powers. We obtain a set of linear second-order partial differential equations for h^{ij}, and with a particular choice of coordinates called the Lorentz gauge (see Section 5.5) they take the form in empty space of the wave equation for the combination of components $\bar{h}^{ij}$ defined by $\bar{h}^{ij} = h^{ij} - \frac{1}{2}\eta^{ij}\eta_{kl}h^{kl}$. This equation has precisely the same form as the equation describing electromagnetic waves, which means that gravitational waves, like their electromagnetic counterparts, travel with the speed of light.

As in the electromagnetic case, the wave equation here has *plane wave* solutions

$$\bar{h}^{ij} = A^{ij}\sin(\omega t + \mathbf{k}.\mathbf{x}) \tag{5.10}$$

(or cosine of the same argument) where A^{ij} and $k_i = (\omega, \mathbf{k})$ are constants. There is already a restriction on A^{ij} because of the choice of the Lorentz gauge, and it can be restricted further to what is known as the *transverse traceless* gauge. Traceless means that if A^{ij} is written as a matrix, the sum of its diagonal components is zero. To understand what we mean by transverse, let us choose our axes so that the wave is travelling in the z-direction, so $k_i = (\omega, 0, 0, \omega)$. Then our gauge condi-

'()*' ' '+ ') 1 2 3
4 ! 5 1 2 -!
& 8# ;6 $ > !

4

Kinematics of cosmological models

In cosmology, the matter components allow us to make a physically motivated choice of preferred motion. For example, we could choose the CMB frame, in which the radiation dipole vanishes, or the frame in which the total momentum density of all components vanishes. Such a choice corresponds to a preferred 4-velocity field u^a that generates a family of preferred world lines. We can then make a 1+3 split relative to u^a, in order to relate the physics and geometry to the observations. In this chapter we discuss how to do this for the kinematics of cosmological models; the following chapter will consider the dynamics.

The (real or fictitious) observers are comoving with the matter-defined 4-velocity u^a, and we can call the observers and the 4-velocity ‘fundamental’. If we change our choice of fundamental 4-velocity, the kinematics and dynamics transform in a well-defined way, as discussed in the following chapter.

4.1 Comoving coordinates

To describe the spacetime geometry it is convenient to use comoving (Lagrangian) coordinates, adapted to the fundamental world lines. These are locally defined as follows.[1]

(1) Choose a surface S that intersects each world line once only (note that no unique choice is available in general). Label each world line where it intersects this surface; as the surface is three-dimensional, three labels y^i, $(i = 1,2,3)$, are required to label all the world lines.

(2) Extend this labelling off the surface S by maintaining the same labelling for the world lines at later and earlier times. Thus the y^i are *comoving coordinates*: the value of the coordinate is maintained along each world line, and in fact the world lines (and so the fundamental particles) are labelled by these coordinates.

(3) Define a time coordinate t along the fluid flow lines (it must be a function that increases along each flow line).

Then (t, y^i) are *comoving coordinates* adapted to the flow lines. Note that the surfaces $t =$const will in general not be orthogonal to the fundamental world lines; indeed, in general it is not possible to choose a time coordinate for which these surfaces are orthogonal (see Section 4.6 below).

---

[1] In this section we are not concerned about possible global problems with coordinates; these are considered in Section 6.7.

The coordinate freedoms available which preserve this form are (a) *time transformations* $t' = t'(t, y^i),\ y^{i'} = y^i$, corresponding to a new choice of time surfaces, and (b) *relabelling of the world lines* by choosing new coordinates in the initial surface: $t' = t,\ y^{i'} = y^{i'}(y^i)$.

A particular choice for t which is often convenient is the *normalized time* $s = s_0 + \tau$, where τ is proper time measured along the fundamental world lines from S (positive to the future of S, negative to the past) and s_0 is an arbitrary constant. With this choice, $x^\mu = (s, y^i)$ are *normalized comoving coordinates*, s measuring proper time from the surface S (on S, $s = s_0$) along the world lines, which lie in the intersection in spacetime of the surfaces $y^i =$const. The remaining time freedom is then $s' = s + f(y^i)$, corresponding to choice of the initial surface S. For example, the standard coordinates (t, r, θ, ϕ) in an FLRW universe model (2.65) are such normalized comoving cooordinates.

General coordinates x^μ will be related to the normalized comoving coordinates by a coordinate transformation,

$$x^\mu = x^\mu(s, y^i), \qquad \mu = 0, \ldots, 3, \qquad i = 1, \ldots, 3, \tag{4.1}$$

the spatial parts of which are similar to the transformation from Lagrangian to Eulerian coordinates in Newtonian theory. Indeed one can define Newtonian-like quasi-Eulerian ('fixed', non-comoving) coordinates in general relativity, defined by the physical (proper) distance and direction from some chosen world line in preferred space sections. They can provide a useful alternative to the more usual Lagrangian coordinates when taking the Newtonian limit (for the FLRW case, see Ellis and Rothman (1993)).

In Newtonian theory, because unique spatial sections exist in spacetime, there is a natural time coordinate t, uniquely defined (up to a constant), which measures proper time along all lines. In Newtonian cosmology, one can again choose comoving spatial ('Lagrangian') coordinates y^i as in the relativistic case, obtaining spacetime coordinates (t, y^i).

4.2 The fundamental 4-velocity

The preferred matter motion implies a preferred 4-velocity at each point. Geometrically, this can be depicted as an arrow pointing along the fundamental world lines. If the preferred world lines are given in terms of *local* coordinates x^μ by $x^\mu = x^\mu(\tau)$ where τ is proper time along the world lines, then the preferred 4-velocity is the unit timelike vector

$$u^\mu = \frac{\mathrm{d}x^\mu}{\mathrm{d}\tau} \quad \Rightarrow \quad u^\mu u_\mu = -1. \tag{4.2}$$

The implication follows by considering the integral for proper time along a world line: $\tau = \int [-(\mathrm{d}x^\mu/\mathrm{d}\tau)(\mathrm{d}x^b/\mathrm{d}\tau)g_{ab}]^{1/2}\mathrm{d}\tau = \int (-u^\mu u_\mu)^{1/2}\mathrm{d}\tau$. In normalized comoving coordinates $x^\mu = (s, y^i)$ this becomes

$$u^\mu = \delta^\mu_0 \quad \Leftrightarrow \quad \frac{\mathrm{d}s}{\mathrm{d}\tau} = 1,\ \frac{\mathrm{d}y^i}{\mathrm{d}\tau} = 0. \tag{4.3}$$

These are also sufficient conditions such that the coordinates are normalized comoving coordinates (the curve parameter is proper time, and the vector u^μ is tangent to the direction where all the y^i are constant).

In general coordinates x^μ, this vector will be given by

$$u^\mu = \left(\frac{\partial x^\mu}{\partial s}\right)_{y^i=\text{const}}. \tag{4.4}$$

This is obtained by applying the coordinate transformation (4.1) to (4.3); conversely, specializing the coordinates in (4.4) to normalized comoving coordinates, we recover (4.3).

The component of any vector X^a parallel to u^a is

$$X^a_\parallel = U^a{}_b X^b, \;\; U^a{}_b := -u^a u_b, \tag{4.5}$$

where $U^a{}_b$ is a projection tensor ($U^a{}_b U^b{}_c = U^a{}_c$) into the one-dimensional tangent line ($U^a{}_a = 1$) parallel to u^a ($U^a{}_b u^b = u^a$). For example, the fundamental 4-velocity in an FLRW universe model in the standard coordinates (2.65) is given by $u^\mu = \delta^\mu_0,\;\; u_\mu = -\delta^0_\mu$. Thus in this case $U^\mu{}_\nu = \delta^\mu_0 \delta^0_\nu$.

In Newtonian theory, a 3-velocity v^i representing the average motion of matter at each point will be defined. Lagrangian coordinates (the Newtonian version of comoving coordinates) are characterized by the condition $v^i = 0$.

Exercise 4.2.1 Show that

(a) if we use general comoving coordinates (t, y^i), then $u^\mu = v^{-1}\delta^\mu_0$, where $v(x^\alpha) = \mathrm{d}s/\mathrm{d}t$, and that $u_\mu = g_{\mu 0}/v$, $g_{00} = -v^2$, $u_0 = -v$;
(b) under a time transformation, $t' = t'(t, y^i),\;\; y^{i'} = y^i$, these relations are preserved with $v \to v' = v/(\partial t'/\partial t)$;
(c) (4.4) reduces to (4.3) on using normalized comoving coordinates;
(d) we have normalized comoving coordinates if and only if $u_\mu = g_{\mu 0}$.

4.3 Time derivatives and the acceleration vector

The time derivative of any tensor $S^{a\cdots}{}_{b\dots}$ along the fluid flow lines is

$$\dot{S}^{a\cdots}{}_{b\dots} = u^c \nabla_c S^{a\cdots}{}_{b\dots}, \tag{4.6}$$

and by (2.28) this is of the form

$$\dot{S}^{a\cdots}{}_{b\dots} = \frac{\mathrm{d}}{\mathrm{d}\tau} S^{a\cdots}{}_{b\dots} + S^{c\cdots}{}_{b\dots}\Gamma^a{}_{cd}u^d + \cdots - S^{a\cdots}{}_{c\dots}\Gamma^c{}_{bd}u^d - \cdots . \tag{4.7}$$

The first term is the apparent derivative (i.e. the value obtained by only taking the directional derivative of the components) relative to the coordinate frame along the world lines, and the others correct this apparent derivative to give the covariant derivative along the world lines. When a frame which is parallelly propagated along the world lines is used, this reduces

+4 1 - - -

to $\dot{S}^{a\cdots}{}_{b\ldots} = \mathrm{d}S^{a\cdots}{}_{b\ldots}/\mathrm{d}\tau = \partial S^{a\cdots}{}_{b\ldots}/\partial\tau + S^{a\cdots}{}_{b\ldots,i}u^i$ (the last term vanishes if comoving coordinates are used).

A particular application of time differentiation is the derivative of the 4-velocity itself in its own direction: this determines the acceleration vector,

$$\dot{u}^a = u^b\nabla_b u^a \;\Rightarrow\; \dot{u}^a u_a = 0, \tag{4.8}$$

which vanishes if and only if the flow lines are geodesics. Physically, this is the case if they represent motion under gravity and inertia alone, i.e. no non-gravitational force acts (see Section 3.1.2). From this definition,

$$\nabla_b u_a = h_a{}^c h_b{}^d \nabla_d u_c - \dot{u}_a u_b\,, \tag{4.9}$$

where the first term on the right is orthogonal to u^a, with h_{ab} defined in (4.10).

The corresponding Newtonian derivative is the 'convective derivative' (Batchelor, 1967), $\dot{T}^{ij\cdots}{}_{\ldots k\ell} = \partial T^{ij\cdots}{}_{\ldots k\ell}/\partial t + T^{ij\cdots}{}_{\ldots k\ell,m}v^m$, determining the rate of change of $T^{ij\cdots}{}_{\ldots k\ell}$ relative to the fluid. As we have just seen, this is essentially what is obtained from the general relativity equations if a parallelly propagated frame is used.

Exercise 4.3.1 Show that in normalized comoving coordinates, $\dot{u}^\mu = \Gamma^\mu{}_{00}$, so that $\dot{u}_\mu = \frac{1}{2}(2\partial g_{0\mu}/\partial t - \partial g_{00}/\partial x^\mu)$.

Exercise 4.3.2

(a) Show that the Newtonian analogue of the 'acceleration vector' is $a_i = \dot{v}_i + \Phi_{,i}$ where Φ is the Newtonian gravitational potential. Deduce that even in Newtonian theory, we are unable to separate the gravitational and inertial parts of a_i invariantly if the matter density does not go to zero at infinity. (Hint: see Heckmann and Schucking (1955), Bondi (1960), Trautman (1965)). (Note: when comparing relativistic and Newtonian equations, $\dot{u}^\mu$ should be compared with a_i not $\dot{v}_i$; the difference between the two cases can be considered to arise because in relativity covariant differentiation already has the gravitational effects coded into it.)

(b) Show that when comoving coordinates are used, $a_i = \Phi_{,i}$.

(c) Consider when this form can be obtained as a limit of the relativistic equations in the previous exercise (i.e. when does there exist a scalar potential Φ for $\dot{u}_a$?).

4.4 Projection to give three-dimensional relations

The existence of a preferred velocity at each point implies the existence of preferred rest frames at each point; locally these define surfaces of simultaneity for the fundamental observers.

4.4.1 Orthogonal projection

The (induced) effective metric tensor for these surfaces is the tensor

$$h_{ab} = g_{ab} + u_a u_b\,. \tag{4.10}$$

We see this as follows: from the above definition and (4.2),

$$h^a{}_b h^b{}_c = h^a{}_c,\ \ h^a{}_a = 3,\ \ h^a{}_b u^b = 0, \tag{4.11}$$

that is, $h^a{}_b$ is a projection tensor projecting into the three-dimensional tangent plane orthogonal to u^a.

Any vector X^a can be projected, by means of $h^a{}_b$, to its part $X^a_\perp$ orthogonal to u^a (i.e. its component in the instantaneous rest-space of an observer moving with the 4-velocity u^a):

$$X^a_\perp = h^a{}_b X^b \ \ \Rightarrow \ \ X^a_\perp u_a = 0,\ (X^a_\perp)_\perp = X^a_\perp. \tag{4.12}$$

The projection tensor is the metric tensor for the rest space; for if X^a and Y^b are any vectors orthogonal to u^a ($X^a u_a = 0 = Y^b u_b$), then $\boldsymbol{X}\cdot\boldsymbol{Y} = X^a g_{ab} Y^b = X^a h_{ab} Y^b$, that is, h_{ab} determines scalar products and so angles and magnitudes for all vectors in the rest space of u^a. It corresponds precisely to the Newtonian metric tensor h_{ij} determining magnitudes and angles in Newtonian theory.

U_{ab} and h_{ab} enable projection of any tensor into parts parallel and perpendicular to u^a. A particular example is the metric tensor itself: $g_{\perp ab} = h_a{}^d h_b{}^e g_{de} = h_{ab},\ g_{\| ab} = U_a{}^d U_b{}^e g_{de} = U_{ab}$, so its splitting into parallel and perpendicular parts is given by

$$g_{ab} = h_{ab} + U_{ab} = h_{ab} - u_a u_b\,. \tag{4.13}$$

This shows that the interval $\mathrm{d}s^2$ associated with an arbitrary displacement $x^\mu \to x^\mu + dx^\mu$ can be decomposed by

$$\mathrm{d}s^2 = g_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu = h_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu - (u_\mu \mathrm{d}x^\mu)^2 = (\delta l)^2 - (\delta t)^2 \tag{4.14}$$

into a time difference $\delta t = (U_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu)^{1/2} = (-u_\mu \mathrm{d}x^\mu)$ ($= c\,\delta t$ in units in which $c \neq 1$) and a spatial distance $\delta l = (h_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu)^{1/2}$ as measured by an observer moving with 4-velocity u^μ. This decomposition implies the usual special relativistic length contraction and time dilation formulae (as these quantities are related to the interval in the standard special relativistic way, and all the equations hold for the projection tensors associated with arbitrary 4-velocities; we will get a different decomposition of g_{ab} into U_{ab} and h_{ab} if we choose a different 4-velocity u_a).

For example, in comoving coordinates, $U^{ij} = 0$, $h_{\mu 0} = 0$ in any universe, while the components of U^{ab}, h_{ab} in an orthonormal tetrad $\{\boldsymbol{u}, \boldsymbol{e}_i\}$ are $U^{ab} = \mathrm{diag}(-1,0,0,0)$, $h_{ab} = \mathrm{diag}(0,1,1,1)$. The components of h_{ab} in an FLRW universe in the standard FLRW coordinates (2.65) are $h_{\mu\nu} = a^2(t)\,\mathrm{diag}(0,1,f^2(r),f^2(r)\sin^2\theta)$. In Newtonian theory, we have $\delta l = (h_{ij}\mathrm{d}x^i \mathrm{d}x^j)^{1/2}$, $\delta t = (-U_{ij}\mathrm{d}x^i \mathrm{d}x^j)^{1/2}$ still, but now the flat 3-space metric h_{ij} is given and fixed, and the time metric U^{ij} is also fixed. Using the preferred time coordinate t, $h_{\mu 0} = 0$, $h_{ij} = \delta_{ij}$ (in Cartesian coordinates) and $U^{\mu\nu} = -\delta^\mu_0 \delta^\nu_0$. These relations also follow from (4.13) in the slow-motion limit.

In the sequel we shall very frequently have occasion to project vectors orthogonal to u^a, and to take the traceless parts, orthogonal to u^μ, of rank two symmetric tensors – i.e. the projected symmetric tracefree, or PSTF, parts. For brevity we use angled brackets on indices to denote the PSTF parts. For convenience we use the term PSTF to include projected rank one tensors (vectors). For any V_a and tensor S_{ab}, the PSTF parts are given by

$$V_{\langle a\rangle} = h_a{}^b V_b\,, \quad S_{\langle ab\rangle} = \left\{h_{(a}{}^c h_{b)}{}^d - \tfrac{1}{3}h_{ab}h^{cd}\right\} S_{cd}\,. \tag{4.15}$$

We write equations so that all terms are manifestly PSTF. We continue to use $\perp$ to indicate the projection in other cases. One may note that a general rank two tensor can be written as

$$\begin{aligned} S_{ab} &= (h_a{}^c + U_a{}^c)(h_b{}^d + U_b{}^d)S_{cd} \\ &= \tfrac{1}{3}h_{ab}(h^{cd}S_{cd}) + S_{\langle ab\rangle} + h_a{}^c h_b{}^d S_{[cd]} \\ &\quad - (h_a{}^c S_{cd}u^d)u_b - u_a(u^c S_{cd} h_b{}^d) + u_a u_b (u^c u^d S_{cd})\,. \end{aligned} \tag{4.16}$$

This decomposition splits the tensor into parts irreducible under the rotational freedom in the hyperplane orthogonal to u^a. For the projection of time derivatives we use the notation

$$\dot{V}^{\langle a\rangle} = h^a{}_b \dot{V}^b\,, \tag{4.17}$$

and similarly for $\dot{S}^{\langle ab\rangle}$.

Any skew 2-tensor orthogonal to u^a can be written as

$$A_{ab} = A_{[ab]} = \eta_{abc}A^c \Leftrightarrow A^c = \tfrac{1}{2}\eta^{abc}A_{bc}, \tag{4.18}$$

where the projected alternating tensor η_{abc} was defined in Section 2.7.5. Then in (4.16) all terms can be expressed using u^a, scalars, projected vectors obeying $V_a = V_{\langle a\rangle}$, and PSTF 2-tensors $S_{ab} = S_{\langle ab\rangle}$. Note that such a PSTF tensor has five independent components. By evaluating components in an orthonormal tetrad aligned with u^μ it is easily found that η_{abc} has the same components as the skew object in three-dimensional vector calculus used, for instance, in defining vector products and the curl of a vector.

4.4.2 Orthogonal spatial derivatives

One can project the four-dimensional covariant derivatives to give three-dimensional derivative operators:

$$\overline{\nabla}_c S^{a\cdots}{}_{b\cdots} = h_c{}^f h^a{}_d \cdots h_b{}^e \cdots \nabla_f S^{d\cdots}{}_{e\cdots}\,. \tag{4.19}$$

Other notations for this derivative in the literature are ${}^{(3)}\nabla_a, \widehat{\nabla}_a, \vec{\nabla}_a$ and D_a.

Following Maartens (1997), we define

$$\mathrm{div}\, V = \overline{\nabla}^a V_a\,, \quad (\mathrm{div}\, S)_a = \overline{\nabla}^b S_{ab}\,, \tag{4.20}$$

$$\mathrm{curl}\, V_a = \eta_{abc}\overline{\nabla}^b V^c\,, \quad \mathrm{curl}\, S_{ab} = \eta_{cd(a}\overline{\nabla}^c S_{b)}{}^d\,. \tag{4.21}$$

The covariant div and curl preserve the PSTF property. Note that $\overline{\nabla}$, div and curl are defined as operators in a 3-manifold only if the vorticity vanishes (see below). When vorticity is non-zero, they are only operators in the tangent hyperplane at each point and not on a manifold.

One should be cautious in using three-dimensional concepts and notation in the case with non-zero vorticity – since in that case there are no hypersurfaces to which h_{ab} is everywhere tangent (see Section 4.6). This means in particular that there is in general no three-dimensional manifold in which Poincaré's Lemma can be applied to obtain scalar or vector potentials. However, when vorticity vanishes, we have a powerful covariant spatial calculus of vectors and tensors. Even with vorticity, it dramatically shortens subsequent equations, makes derivations far easier and more transparent and facilitates new insights.

With these definitions we find that

$$\overline{\nabla}_c h_{ab} = 0\,,\ \overline{\nabla}_d \eta_{abc} = 0\,,\ \dot{h}_{ab} = 2u_{(a}\dot{u}_{b)}\,,\ \dot{\eta}_{abc} = 3u_{[a}\eta_{bc]d}\dot{u}^d\,. \tag{4.22}$$

Identities obeyed by div and curl are collected in Section 4.8.

Exercise 4.4.1

(a) Show from (4.14) that the proper time $\mathrm{d}\tau$ experienced by a particle between events P and Q for which an observer O determines coordinates x^μ, $x^\mu + \mathrm{d}x^\mu$, respectively, is $\mathrm{d}\tau = \delta t/\gamma$ where $\gamma = (1 - v^2/c^2)^{-1/2}$ and $v = \delta l/\delta t$ is the velocity of the particle relative to O.

(b) Show that

$$\eta_{abcd} = 2u_{[a}\eta_{b]cd} - 2\eta_{ab[c}u_{d]}\,,\ \eta_{abc}\eta^{def} = 3!h_{[a}{}^d h_b{}^e h_{c]}{}^f\,, \tag{4.23}$$

$$(V_{\langle a\rangle})\dot{} = \dot{V}_{\langle a\rangle} + V_b\dot{u}^b u_a\,. \tag{4.24}$$

(d) Derive (4.22).

4.5 Relative position and velocity

4.5.1 Relative position vectors

Consider a curve $y^i = y^i(v)$ in a surface $S : s = s_0$, where (s, y^i) are comoving coordinates. This curve links a set of fundamental observer world lines (which we shall assume for now are also galaxy world lines) in that surface; at all later times, the same curve links the same set of world lines, that is, the curve is dragged along by the world lines from the surface S to any other surface $s =$ const.

Similarly the vector $\beta^\mu = (\mathrm{d}x^\mu/\mathrm{d}v)\delta v$ tangent to this curve, given in comoving coordinates by $\beta^\mu = (0, \delta y^i)$, where $\delta y^i = (\mathrm{d}y^i/\mathrm{d}v)\delta v$, links the same pair of world lines O: $y^i = c^i =$const and G: $y^i = c^i + \delta y^i$, $\delta y^i =$const, at all times, provided δv is small (so that the displacement represented by the vector is a good approximation to displacement along the curve). This is a *connecting vector* as described in Section 2.4, since it always joins the same pair of fundamental world lines. In general coordinates x^μ, this vector will be given by

$$\beta^\mu = \left(\frac{\partial x^\mu}{\partial y^i}\right)_{s=\text{const}} \delta y^i\ . \tag{4.25}$$

An observer on O would at all times find that the spacetime position defined by β^μ lies on the world line of the galaxy G.

26 04 2020 X

4. Kinematics of cosmological models

↳ Pág.(89) 73

Pág. 76 (92)

4.4 Projection to give three-dimensional relations.

→ Orthogonal projection

→ Orthogonal spatial derivatives

D) The large scale structure of space-time (1973 ...) ..pdf

Pág. 47 (55)

2.8 The volume element and Gauss' theorem.

Pág. 78 (86)

4. The physical significance of curvature

Pág. 117 (125)

5. Exact solutions

5.1 Minkowski space-time

E) Theoretical and observational Cosmology (1999 ...). pdf

! '()*' ' '+ ' . -
68#3 4 ! 5 1 2 -!
& $8 ; . 1 ! 0 ?

is the correct formula, one has only to show that the covariant derivative of the induced metric is zero and that the torsion vanishes. This follows because

$$h_{ab\|c} = (g_{ef} \mp \overline{n}_e \overline{n}_f)_{;g} h^e{}_a h^f{}_b h^g{}_c = 0,$$

and $$f_{\|ab} = h^e{}_a h^g{}_b f_{;eg} = h^e{}_a h^g{}_b f_{;ge} = f_{\|ba}.$$

The curvature tensor $R'^a{}_{bcd}$ of the induced metric **h** can be related to the curvature tensor $R^a{}_{bcd}$ on $\theta(\mathscr{S})$ and the second fundamental form $\boldsymbol{\chi}$ as follows. If $\mathbf{Y} \in H$ is a vector field on $\theta(\mathscr{S})$, then

$$R'^a{}_{bcd} Y^b = Y^a{}_{\|dc} - Y^a{}_{\|cd}.$$

Now

$$\begin{aligned} Y^a{}_{\|dc} &= (Y^a{}_{\|d})_{\|c} = (\overline{Y}^e{}_{;f} h^g{}_e h^f{}_i)_{;k} h^a{}_g h^i{}_d h^k{}_c \\ &= \overline{Y}^e{}_{;fk} h^a{}_e h^f{}_d h^k{}_c \mp \overline{Y}^e{}_{;f} \overline{n}_e \overline{n}^g{}_{;k} h^f{}_d h^a{}_g h^k{}_c \mp \overline{Y}^e{}_{;f} \overline{n}^f \overline{n}_{i;k} h^a{}_e h^i{}_d h^k{}_c \end{aligned}$$

and $$\overline{Y}^e{}_{;f} \overline{n}_e h^f{}_d = (\overline{Y}^e \overline{n}_e)_{;f} h^f{}_d - \overline{Y}^e \overline{n}_{e;f} h^f{}_d = -\overline{Y}^e \overline{n}_{e;f} h^f{}_d,$$

since $\overline{Y}^e \overline{n}_e = 0$ on $\theta(\mathscr{S})$, therefore

$$R'^a{}_{bcd} Y^b = (R^e{}_{bkf} h^a{}_e h^k{}_c h^f{}_d \pm \chi_{bd} \chi^a{}_c \mp \chi_{bc} \chi^a{}_d)\, Y^b.$$

Since this holds for all $\mathbf{Y} \in H$,

$$R'^a{}_{bcd} = R^e{}_{fgh} h^a{}_e h^f{}_b h^g{}_c h^h{}_d \pm \chi^a{}_c \chi_{bd} \mp \chi^a{}_d \chi_{bc}. \tag{2.34}$$

This is known as Gauss' equation.

Contracting this equation on a and c and multiplying by h^{bd}, one obtains the curvature scalar R' of the induced metric:

$$R' = R \mp 2R_{ab} n^a n^b \pm (\chi^a{}_a)^2 \mp \chi^{ab} \chi_{ab}. \tag{2.35}$$

One can derive another relation between the second fundamental form and the curvature tensor $R^a{}_{bcd}$ on $\theta(\mathscr{S})$ by subtracting the expressions

$$(\chi^a{}_a)_{\|b} = (\overline{n}^a{}_{;d} h^d{}_a)_{;e} h^e{}_b$$

and $$(\chi^a{}_b)_{\|a} = (\overline{n}^c{}_{;d} h^a{}_c h^d{}_e)_{;f} h^f{}_a h^e{}_b,$$

finding $$\chi^a{}_{b\|a} - \chi^a{}_{a\|b} = R_{ef} n^f h^e{}_b. \tag{2.36}$$

This is known as Codacci's equation.

2.8 The volume element and Gauss' theorem

If $\{\mathbf{E}^a\}$ is a basis of one-forms, one can form from it the n-form

$$\boldsymbol{\epsilon} = n!\mathbf{E}^1 \wedge \mathbf{E}^2 \wedge \dots \wedge \mathbf{E}^n.$$

If $\{\mathrm{E}^{a'}\}$, related to $\{\mathrm{E}^{a}\}$ by $\mathrm{E}^{a'} = \Phi^{a'}{}_{a}\mathrm{E}^{a}$, is another basis, the n-form $\boldsymbol{\epsilon}'$ defined by this basis will be related to $\boldsymbol{\epsilon}$ by

$$\boldsymbol{\epsilon}' = \det(\Phi^{a'}{}_{a})\,\boldsymbol{\epsilon},$$

so this form is not unique. However, one can use the existence of the metric to define (in a given basis) the form

$$\boldsymbol{\eta} = |g|^{\frac{1}{2}}\,\boldsymbol{\epsilon}$$

where $g \equiv \det(g_{ab})$. This form has components

$$\eta_{ab\ldots d} = n!\,|g|^{\frac{1}{2}}\,\delta^{1}{}_{[a}\delta^{2}{}_{b}\ldots\delta^{n}{}_{d]}.$$

The transformation law for g will just cancel the determinant, $\det(\Phi^{a'}{}_{a})$, provided that $\det(\Phi^{a'}{}_{a}) > 0$. Therefore if $\mathscr{M}$ is orientable the n-forms $\boldsymbol{\eta}$ defined by coordinate bases of an oriented atlas will be identical, i.e. given an orientation of $\mathscr{M}$, one can define a unique n-form field $\boldsymbol{\eta}$, the *canonical n-form*, on $\mathscr{M}$.

The contravariant antisymmetric tensor

$$\eta^{ab\ldots d} = g^{ae}g^{bf}\ldots g^{dh}\eta_{ef\ldots h}$$

has components

$$\eta^{ab\ldots d} = (-)^{\frac{1}{2}(n-s)}n!\,|g|^{\frac{1}{2}}\,\delta^{[a}{}_{1}\delta^{b}{}_{2}\ldots\delta^{d]}{}_{n},$$

where s is the signature of $\mathbf{g}$ (so $\frac{1}{2}(n-s)$ is the number of negative eigenvalues of the matrix of metric components (g_{ab})). Therefore these tensors satisfy the relations

$$\eta^{ab\ldots d}\eta_{ef\ldots h} = (-)^{\frac{1}{2}(n-s)}n!\,\delta^{a}{}_{[e}\delta^{b}{}_{f}\ldots\delta^{d}{}_{h]}. \tag{2.37}$$

The Christoffel relations imply that the covariant derivatives of $\eta_{ab\ldots d}$ and $\eta^{ab\ldots d}$ with respect to the connection defined by the metric vanish, i.e.

$$\eta^{ab\ldots d}{}_{;e} = 0 = \eta_{ab\ldots d;e}.$$

Using the canonical n-form, one can define the volume (with respect to the metric $\mathbf{g}$) of an n-dimensional submanifold $\mathscr{U}$ as $\dfrac{1}{n!}\displaystyle\int_{\mathscr{U}}\boldsymbol{\eta}$. Thus $\boldsymbol{\eta}$ can be regarded as a positive definite volume measure on $\mathscr{M}$. We shall often use it in this sense, and shall denote it by dv. Note that d is not meant to represent the exterior differential operator here; dv is simply a measure on $\mathscr{M}$. If f is a function on $\mathscr{M}$, one can define its integral over $\mathscr{U}$ with respect to this volume measure as

$$\int_{\mathscr{U}} f\,\mathrm{d}v = \frac{1}{n!}\int_{\mathscr{U}} f\,\boldsymbol{\eta}.$$

With respect to local oriented coordinates $\{x^a\}$, this can be expressed as the multiple integral

$$\int_{\mathscr{U}} f\,|g|^{\frac{1}{2}}\,\mathrm{d}x^1\,\mathrm{d}x^2 \dots \mathrm{d}x^n,$$

which is invariant under a change of coordinates.

If $\mathbf{X}$ is a vector field on $\mathscr{M}$, its contraction with $\boldsymbol{\eta}$ will be an $(n-1)$-form field $\mathbf{X}\,.\,\boldsymbol{\eta}$, where

$$(\mathbf{X}\,.\,\boldsymbol{\eta})_{b\dots d} = X^a \eta_{ab\dots d}.$$

This $(n-1)$-form may be integrated over any $(n-1)$-dimensional compact orientable submanifold $\mathscr{V}$. We write this integral as

$$\int_{\mathscr{V}} X^a \mathrm{d}\sigma_a = \frac{1}{(n-1)!}\int_{\mathscr{V}} \mathbf{X}\,.\,\boldsymbol{\eta},$$

where the canonical form $\boldsymbol{\eta}$ is regarded as defining a measure-valued form $\mathrm{d}\sigma_a$ on the submanifold $\mathscr{V}$. If the orientation of $\mathscr{V}$ is given by the direction of the normal form n_a, then $\mathrm{d}\sigma_a$ can be expressed as $n_a \mathrm{d}\sigma$ where $\mathrm{d}\sigma$ is a positive definite volume measure on the submanifold $\mathscr{V}$. The volume measure $\mathrm{d}\sigma$ is not unique unless the normal n_a is normalized. If n_a is normalized to unit magnitude in a metric $\mathbf{g}$ on $\mathscr{M}$, i.e. $n_a n_b g^{ab} = \pm 1$, then $\mathrm{d}\sigma$ is equal to the volume measure on $\mathscr{V}$ defined by the induced metric on $\mathscr{V}$ (to see this, simply choose an orthonormal basis with $n_a g^{ab}$ as one of the basis vectors).

Using the canonical form, one can derive Gauss' formula from Stokes' theorem: for any compact n-dimensional submanifold $\mathscr{U}$ of $\mathscr{M}$,

$$\int_{\partial\mathscr{U}} X^a \mathrm{d}\sigma_a = \frac{1}{(n-1)!}\int_{\partial\mathscr{U}} \mathbf{X}\,.\,\boldsymbol{\eta} = \frac{1}{(n-1)!}\int_{\mathscr{U}} \mathrm{d}(\mathbf{X}\,.\,\boldsymbol{\eta}).$$

But

$$\begin{aligned}(\mathrm{d}(\mathbf{X}\,.\,\boldsymbol{\eta}))_{a\dots de} &= (-)^{n-1}(X^g \eta_{g[a\dots d})_{;e]} \\ &= (-)^{n-1}\delta^s{}_{[a}\dots\delta^t{}_d\,\delta^u{}_{e]}\,\eta_{gs\dots t}X^g{}_{;u} \\ &= (-)^{(n-1)-\frac{1}{2}(n-s)}\frac{1}{n!}\eta^{s\dots tu}\eta_{a\dots de}\eta_{gs\dots t}X^g{}_{;u} \\ &= \eta_{a\dots de}\,\delta^s{}_{[s}\dots\delta^t{}_t\,\delta^u{}_{g]}X^g{}_{;u} \\ &= n^{-1}\eta_{a\dots de}X^g{}_{;g},\end{aligned}$$

on using relation (2.37) twice. Therefore

$$\int_{\partial\mathscr{U}} X^a\,\mathrm{d}\sigma_a = \int_{\mathscr{U}} X^g{}_{;g}\,\mathrm{d}v$$

holds for any vector field $\mathbf{X}$; this is Gauss' theorem. Note that the orientation on $\mathscr{U}$ for which this theorem is valid is that given by the normal form $\boldsymbol{\eta}$ such that $\langle \mathbf{n}, \mathbf{X} \rangle$ is positive if $\mathbf{X}$ is a vector which points out of $\mathscr{U}$. If the metric $\mathbf{g}$ is such that $g^{ab}n_a n_b$ is negative, the vector $g^{ab}n_b$ will point into $\mathscr{U}$.

2.9 Fibre bundles

Some of the geometrical properties of a manifold $\mathscr{M}$ can be most easily examined by constructing a manifold called a fibre bundle, which is locally a direct product of $\mathscr{M}$ and a suitable space. In this section we shall give the definition of a fibre bundle and shall consider four examples that will be used later: the tangent bundle $T(\mathscr{M})$, the tensor bundle $T^r_s(\mathscr{M})$, the bundle of linear frames or bases $L(\mathscr{M})$, and the bundle of orthonormal frames $O(\mathscr{M})$.

A C^k *bundle* over a C^s ($s \geqslant k$) manifold $\mathscr{M}$ is a C^k manifold $\mathscr{E}$ and a C^k surjective map $\pi: \mathscr{E} \to \mathscr{M}$. The manifold $\mathscr{E}$ is called the total space, $\mathscr{M}$ is called the base space and π the projection. Where no confusion can arise, we will denote the bundle simply by $\mathscr{E}$. In general, the inverse image $\pi^{-1}(p)$ of a point $p \in \mathscr{M}$ need not be homeomorphic to $\pi^{-1}(q)$ for another point $q \in \mathscr{M}$. The simplest example of a bundle is a *product bundle* $(\mathscr{M} \times \mathscr{A}, \mathscr{M}, \pi)$ where $\mathscr{A}$ is some manifold and the projection π is defined by $\pi(p, v) = p$ for all $p \in \mathscr{M}$, $v \in \mathscr{A}$. For example, if one chooses $\mathscr{M}$ as the circle S^1 and $\mathscr{A}$ as the real line R^1, one constructs the cylinder C^2 as a product bundle over S^1.

A bundle which is locally a product bundle is called a fibre bundle. Thus a bundle is a *fibre bundle* with fibre $\mathscr{F}$ if there exists a neighbourhood $\mathscr{U}$ of each point q of $\mathscr{M}$ such that $\pi^{-1}(\mathscr{U})$ is isomorphic with $\mathscr{U} \times \mathscr{F}$, in the sense that for each point $p \in \mathscr{U}$ there is a diffeomorphism ϕ_p of $\pi^{-1}(p)$ onto $\mathscr{F}$ such that the map ψ defined by $\psi(u) = (\pi(u), \phi_{\pi(u)})$ is a diffeomorphism $\psi: \pi^{-1}(\mathscr{U}) \to \mathscr{U} \times \mathscr{F}$. Since $\mathscr{M}$ is paracompact, we can choose a locally finite covering of $\mathscr{M}$ by such open sets $\mathscr{U}_\alpha$. If $\mathscr{U}_\alpha$ and $\mathscr{U}_\beta$ are two members of such a covering, the map

$$(\phi_{\alpha, p}) \circ (\phi_{\beta, p}{}^{-1})$$

is a diffeomorphism of $\mathscr{F}$ onto itself for each $p \in (\mathscr{U}_\alpha \cap \mathscr{U}_\beta)$. The inverse images $\pi^{-1}(p)$ of points $p \in \mathscr{M}$ are therefore necessarily all diffeomorphic to $\mathscr{F}$ (and so to each other). For example, the Möbius strip is a fibre bundle over S^1 with fibre R^1; we need two open sets $\mathscr{U}_1$, $\mathscr{U}_2$

4

The physical significance of curvature

In this chapter we consider the effect of space–time curvature on families of timelike and null curves. These could represent flow lines of fluids or the histories of photons. In § 4.1 and § 4.2 we derive the formulae for the rate of change of vorticity, shear and expansion of such families of curves; the equation for the rate of change of expansion (Raychaudhuri's equation) plays a central role in the proofs of the singularity theorems of chapter 8. In § 4.3 we discuss the general inequalities on the energy–momentum tensor which imply that the gravitational effect of matter is always to tend to cause convergence of timelike and of null curves. A consequence of these energy conditions is, as is seen in § 4.4, that conjugate or focal points will occur in families of non-rotating timelike or null geodesics in general space–times. In § 4.5 it is shown that the existence of conjugate points implies the existence of variations of curves between two points which take a null geodesic into a timelike curve, or a timelike geodesic into a longer timelike curve.

4.1 Timelike curves

In chapter 3 we saw that if the metric was static there was a relation between the magnitude of the timelike Killing vector and the Newtonian potential. One was able to tell whether a body was in a gravitational field by whether, if released from rest, it would accelerate with respect to the static frame defined by the Killing vector. However, in general, space–time will not have any Killing vectors. Thus one will not have any special frame against which to measure acceleration; the best one can do is to take two bodies close together and measure their relative acceleration. This will enable one to measure the gradient of the gravitational field. If one thinks of the metric as being analogous to the Newtonian potential, the gradient of the Newtonian field would correspond to the second derivatives of the metric. These are described by the Riemann tensor. Thus one would expect that the relative acceleration of two neighbouring bodies would be related to some components of the Riemann tensor.

In order to investigate this relation more precisely we shall examine the behaviour of a congruence of timelike curves with timelike unit tangent vector **V** ($g(\mathbf{V},\mathbf{V}) = -1$). These curves could represent the histories of small test particles, in which case they would be geodesics, or they might represent the flow lines of a fluid. If this were a perfect fluid, then by (3.10)

$$(\mu+p)\dot{V}^a = -p_{;b}h^{ab}, \tag{4.1}$$

where $\dot{V}^a = V^a{}_{;b}V^b$ is the acceleration of the flow lines and $h^a{}_b = \delta^a{}_b + V^aV_b$ is the tensor which projects a vector $\mathbf{X}\in T_q$ into its component in the subspace H_q of T_q orthogonal to V. One may also think of h_{ab} as the metric in H_q (cf. §2.7).

Suppose $\lambda(t)$ is a curve with tangent vector $\mathbf{Z} = (\partial/\partial t)_\lambda$. Then one may construct a family $\lambda(t,s)$ of curves by moving each point of the curve $\lambda(t)$ a distance s along the integral curves of **V**. If one now defines **Z** as $(\partial/\partial t)_{\lambda(t,s)}$ it follows from the definition of the Lie derivative (see §2.4) that $L_\mathbf{V}\mathbf{Z} = 0$ or in other words that

$$\frac{\mathrm{D}}{\partial s}Z^a = V^a{}_{;b}Z^b. \tag{4.2}$$

One may interpret **Z** as representing the separation of points equal distances from some arbitrary initial points along two neighbouring curves. If one adds a multiple of **V** to **Z** then this vector will represent the separation of points on the same two curves but at different distances along the curves. It is really only the separation of neighbouring curves that one is interested in, not the separation of particular points on these curves. One is thus concerned only with **Z** modulo a component parallel to **V**, i.e. only with the projection of **Z** at each point q into the space Q_q consisting of equivalence classes of vectors which differ only by addition of a multiple of **V**. This space can be represented as the subspace H_q of T_q consisting of vectors orthogonal to **V**. The projection of **Z** into H_q will be denoted by ${}_\perp Z^a = h^a{}_b Z^b$. In the case of a fluid one can regard ${}_\perp\mathbf{Z}$ as the distance between two neighbouring particles of the fluid as measured in their rest frame.

From (4.2) it follows that

$${}_\perp\frac{\mathrm{D}}{\partial s}({}_\perp Z^a) = V^a{}_{;b}\,{}_\perp Z^b. \tag{4.3}$$

This gives the rate of change of the separation of two infinitesimally neighbouring curves as measured in H_q. Operating again with $\mathrm{D}/\partial s$ and projecting into H_q, one finds

$$h^a{}_b \frac{\mathrm{D}}{\partial s}\left(h^b{}_c \frac{\mathrm{D}}{\partial s}{}_\perp Z^c\right) = h^a{}_b(V^b{}_{;cd\,\perp}Z^c V^d + V^b{}_{;c}V^c{}_{;d}V_e Z^e V^d$$
$$+ V^b{}_{;c}V^c V^e{}_{;d}Z_e V^d + V^b{}_{;c}h^c{}_e Z^e{}_{;d}V^d).$$

Changing the order of the derivatives in the first term and using (4.2), this reduces to

$$h^a{}_b \frac{\mathrm{D}}{\partial s}\left(h^b{}_c \frac{\mathrm{D}}{\partial s}{}_\perp Z^c\right) = -R^a{}_{bcd\,\perp}Z^c V^b V^d + h^a{}_b \dot{V}^b{}_{;c\,\perp}Z^c + \dot{V}^a \dot{V}_{b\,\perp}Z^b. \quad (4.4)$$

This equation, known as the deviation or Jacobi equation, gives the relative acceleration, i.e. the second time derivative of the separation, of two infinitesimally neighbouring curves as measured in H_q. We see that this depends only on the Riemann tensor if the curves are geodesics.

In Newtonian theory, the acceleration of each particle is given by the gradient of the potential Φ and therefore the relative acceleration of two particles with separation Z^a is $\Phi_{;ab}Z^b$. Thus the Riemann tensor term $R_{abcd}V^b V^d$ is analogous to the Newtonian $\Phi_{;ac}$. The effect of this 'tidal force' term can be seen, for example, by considering a sphere of particles freely falling towards the earth. Each particle moves on a straight line through the centre of the earth but those nearer the earth fall faster than those further away. This means that the sphere does not remain a sphere but is distorted into an ellipsoid with the same volume.

In order to investigate the deviation equation further we shall introduce dual orthonormal bases $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_2$, $\mathbf{E}_4$ and $\mathbf{E}^1$, $\mathbf{E}^2$, $\mathbf{E}^3$, $\mathbf{E}^4$ of T_q and $T^*{}_q$ at some point q on an integral curve $\gamma(s)$ of $\mathbf{V}$, with $\mathbf{E}^4 = \mathbf{V}$. One would like to propagate them along $\gamma(s)$ to obtain similar such bases at each point of $\gamma(s)$. However, if one parallelly propagates them along $\gamma(s)$ (i.e. so that $\mathrm{D}/\partial s$ of each vector is zero) $\mathbf{E}_4$ will not remain equal to $\mathbf{V}$, and $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_3$ will not remain orthogonal to $\mathbf{V}$, unless $\gamma(s)$ is a geodesic. We therefore introduce a new derivative along $\gamma(s)$ called the *Fermi derivative* $\mathrm{D_F}/\partial s$. This is defined for a vector field $\mathbf{X}$ along $\gamma(s)$ by:

$$\frac{\mathrm{D_F}\mathbf{X}}{\partial s} = \frac{\mathrm{D}\mathbf{X}}{\partial s} - g\left(\mathbf{X}, \frac{\mathrm{D}\mathbf{V}}{\partial s}\right)\mathbf{V} + g(\mathbf{X}, \mathbf{V})\frac{\mathrm{D}\mathbf{V}}{\partial s}.$$

It has the properties:

(i) $\dfrac{\mathrm{D_F}}{\partial s} = \dfrac{\mathrm{D}}{\partial s}$ if $\gamma(s)$ is a geodesic;

(ii) $\dfrac{\mathrm{D_F}\mathbf{V}}{\partial s} = 0$;

(iii) if $\mathbf{X}$ and $\mathbf{Y}$ are vector fields along $\gamma(s)$ such that

$$\frac{\mathrm{D_F}\mathbf{X}}{\partial s} = 0 = \frac{\mathrm{D_F}\mathbf{Y}}{\partial s},$$

then $g(\mathbf{X}, \mathbf{Y})$ is constant along $\gamma(s)$;

(iv) if $\mathbf{X}$ is a vector field along $\gamma(s)$ orthogonal to $\mathbf{V}$ then

$$\frac{\mathrm{D_F}\mathbf{X}}{\partial s} = {}_\perp\!\left(\frac{\mathrm{D}\mathbf{X}}{\partial s}\right).$$

(This last property shows that the Fermi derivative is a natural generalization of the derivative $\mathrm{D}/\partial s$.)

Thus, if one propagates an orthonormal basis of T_q along $\gamma(s)$ so that the Fermi derivative of each basis vector is zero, one obtains an orthonormal basis at each point of $\gamma(s)$, with $\mathbf{E}_4 = \mathbf{V}$. The vectors $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_2$ may be interpreted as giving a non-rotating set of axes along $\gamma(s)$. These could be realized physically by small gyroscopes pointing in the direction of each vector.

The definition of the Fermi derivative along $\gamma(s)$ can be extended from vector fields to arbitrary tensor fields by the usual rules:

(i) $\mathrm{D_F}/\partial s$ is a linear mapping of tensor fields of type (r, s) along $\gamma(s)$ to tensor fields of type (r, s), which commutes with contractions;

(ii) $\dfrac{\mathrm{D_F}}{\partial s}(\mathbf{K}\otimes\mathbf{L}) = \dfrac{\mathrm{D_F}\mathbf{K}}{\partial s}\otimes\mathbf{L} + \mathbf{K}\otimes\dfrac{\mathrm{D_F}\mathbf{L}}{\partial s}$;

(iii) $\dfrac{\mathrm{D_F}f}{\partial s} = \dfrac{\mathrm{d}f}{\mathrm{d}s}$, where f is a function.

From these rules it follows that the dual basis $\mathbf{E}^1$, $\mathbf{E}^2$, $\mathbf{E}^3$, $\mathbf{E}^4$ of $T^*{}_q$ is also Fermi-propagated along $\gamma(s)$. Using Fermi derivatives, (4.3) and (4.4) may be written as:

$$\frac{\mathrm{D_F}}{\partial s}\,{}_\perp Z^a = V^a{}_{;b}\,{}_\perp Z^b, \tag{4.5}$$

$$\frac{\mathrm{D^2{}_F}}{\partial s^2}\,{}_\perp Z^a = -R^a{}_{bcd}\,{}_\perp Z^c V^b V^d + h^a{}_b \dot{V}^b{}_{;c}\,{}_\perp Z^c + \dot{V}^a \dot{V}_b\,{}_\perp Z^b. \tag{4.6}$$

One may express these equations in terms of the Fermi-propagated dual bases. As $\perp\mathbf{Z}$ is orthogonal to $\mathbf{V}$ it will have components with respect to $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_2$ only. Thus it may be expressed as $Z^\alpha \mathbf{E}_\alpha$ where we adopt the convention that Greek indices take the values 1, 2, 3 only. Then (4.5) and (4.6) can be written in terms of ordinary derivatives:

$$\frac{\mathrm{d}}{\mathrm{d}s} Z^\alpha = V^\alpha{}_{;\beta} Z^\beta, \tag{4.7}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2} Z^\alpha = (-R^\alpha{}_{4\beta4} + \dot{V}^\alpha{}_{;\beta} + \dot{V}^\alpha \dot{V}_\beta)\, Z^\beta \tag{4.8}$$

where $V^\alpha{}_{;\beta}$ are the components of $V^a{}_{;b}$ for which $a = \alpha$ and $b = \beta$. As the components Z^α obey the first order linear ordinary differential equation (4.7), they can be expressed in terms of their values at some point q by:

$$Z^\alpha(s) = A_{\alpha\beta}(s)\, Z^\beta|_q, \tag{4.9}$$

where $A_{\alpha\beta}(s)$ is a 3×3 matrix which is the unit matrix at q and satisfies

$$\frac{\mathrm{d}}{\mathrm{d}s} A_{\alpha\beta}(s) = V_{\alpha;\gamma} A_{\gamma\beta}(s). \tag{4.10}$$

In the case of a fluid the matrix $A_{\alpha\beta}$ can be regarded as representing the shape and orientation of a small element of fluid which is spherical at q. This matrix can be written as

$$A_{\alpha\beta} = O_{\alpha\delta} S_{\delta\beta} \tag{4.11}$$

where $O_{\alpha\beta}$ is an orthogonal matrix with positive determinant and $S_{\alpha\beta}$ is a symmetric matrix. These will both be chosen to be the unit matrix at q. The matrix $O_{\alpha\beta}$ may be thought of as representing the rotation that neighbouring curves have undergone with respect to the Fermi-propagated basis while $S_{\alpha\beta}$ represents the separation of these curves from $\gamma(s)$. The determinant of $S_{\alpha\beta}$, which equals the determinant of $A_{\alpha\beta}$, may be thought of as representing the three-volume of the element of the surface orthogonal to $\gamma(s)$ marked out by the neighbouring curves.

At q where $A_{\alpha\beta}$ is the unit matrix, $\mathrm{d}O_{\alpha\beta}/\mathrm{d}s$ is antisymmetric and $\mathrm{d}S_{\alpha\beta}/\mathrm{d}s$ is symmetric. Thus the rate of rotation of neighbouring curves at q is given by the antisymmetric part of $V_{\alpha;\beta}$ while the rate of change of their separation from $\gamma(s)$ is given by the symmetric part of $V_{\alpha;\beta}$ and the rate of change of volume is given by the trace of $V_{\alpha;\beta}$. We therefore define the vorticity tensor as

$$\omega_{ab} = h_a{}^c h_b{}^d V_{[c;d]}, \tag{4.12}$$

the expansion tensor as

$$\theta_{ab} = h_a{}^c h_b{}^d V_{(c;d)}, \tag{4.13}$$

and the volume expansion as

$$\theta = \theta_{ab}h^{ab} = V_{a;\,b}h^{ab} = V^{a}{}_{;\,a}. \tag{4.14}$$

We further define the shear tensor as the trace free part of θ_{ab},

$$\sigma_{ab} = \theta_{ab} - \tfrac{1}{3}h_{ab}\theta, \tag{4.15}$$

and the vorticity vector as

$$\omega^{a} = \tfrac{1}{2}\eta^{abcd}V_{b}\omega_{cd} = \tfrac{1}{2}\eta^{abcd}V_{b}V_{c;\,d}. \tag{4.16}$$

The covariant derivative of the vector V may be expressed in terms of these quantities;

$$V_{a;\,b} = \omega_{ab} + \sigma_{ab} + \tfrac{1}{3}\theta h_{ab} - \dot{V}_{a}V_{b}. \tag{4.17}$$

This decomposition of the gradient of the fluid velocity vector is directly analogous to that in Newtonian hydrodynamics.

In the Fermi-propagated orthonormal basis the vorticity and expansion can be expressed in terms of the matrix $A_{\alpha\beta}$ and its inverse $A^{-1}{}_{\alpha\beta}$:

$$\omega_{\alpha\beta} = -A^{-1}{}_{\gamma[\alpha}\frac{\mathrm{d}}{\mathrm{d}s}A_{\beta]\gamma}, \tag{4.18}$$

$$\theta_{\alpha\beta} = A^{-1}{}_{\gamma(\alpha}\frac{\mathrm{d}}{\mathrm{d}s}A_{\beta)\gamma}, \tag{4.19}$$

$$\theta = (\det \mathrm{A})^{-1}\frac{\mathrm{d}}{\mathrm{d}s}(\det \mathrm{A}). \tag{4.20}$$

From the deviation equation (4.8) it follows that

$$\frac{\mathrm{d}^{2}}{\mathrm{d}s^{2}}A_{\alpha\beta} = (-R_{\alpha4\gamma4} + \dot{V}_{\alpha;\,\gamma} + \dot{V}_{\alpha}\dot{V}_{\gamma})A_{\gamma\beta}. \tag{4.21}$$

This equation enables one to calculate the propagation of the vorticity, shear and expansion along the integral curves of V if one knows the Riemann tensor.

Multiplying by $A^{-1}{}_{\beta\gamma}$ and taking the antisymmetric part, one obtains

$$\frac{\mathrm{d}}{\mathrm{d}s}\omega_{\alpha\beta} = 2\omega_{\gamma[\alpha}\theta_{\beta]\gamma} + \dot{V}_{[\alpha;\,\beta]}. \tag{4.22}$$

Thus the propagation of vorticity depends on the antisymmetric gradient of the acceleration but not the 'tidal force'. Another form of the above equation is

$$\frac{\mathrm{d}}{\mathrm{d}s}(A_{\gamma\alpha}\omega_{\gamma\delta}A_{\delta\beta}) = A_{\gamma\alpha}\dot{V}_{[\gamma;\,\delta]}A_{\delta\beta}. \tag{4.23}$$

Therefore $A_{\gamma\alpha}\omega_{\gamma\delta}A_{\delta\beta}$ is a constant matrix if the curves are geodesics; in particular, if the curves are geodesics and the vorticity vanishes at one point on a curve, it will vanish at all points on the curve. If the curves are the flow lines of a perfect fluid it follows from (4.1) that

$$\dot{V}_{[\alpha;\beta]} = -\frac{1}{\mu+p}\omega_{\alpha\beta}\frac{\mathrm{d}p}{\mathrm{d}s}.$$

If the fluid is isentropic, this implies the conservation law:

$$WA_{\gamma\alpha}\omega_{\gamma\delta}A_{\delta\beta} = \text{constant}, \tag{4.24}$$

where

$$\log W = \int\frac{\mathrm{d}p}{\mu+p}.$$

This conservation law is the relativistic form of the Newtonian vorticity conservation law. In the geodesic or pressure-free case, this takes the usual form that the magnitude of the vorticity vector is inversely proportional to the area of a cross-section orthogonal to the vorticity vector of an element of the fluid. When the pressure is non-zero, there is an extra relativistic effect arising from the fact that compression of the fluid does work on the fluid and therefore increases the mass and so the inertia of an element of the fluid (cf. (3.9)). This means that the vorticity of a fluid increases less under compression than would otherwise be expected.

Multiplying (4.21) by ${A^{-1}}_{\beta\gamma}$ and taking the symmetric part, one finds

$$\frac{\mathrm{d}}{\mathrm{d}s}\theta_{\alpha\beta} = -R_{\alpha4\beta4}-\omega_{\alpha\gamma}\omega_{\gamma\beta}-\theta_{\alpha\gamma}\theta_{\gamma\beta}+\dot{V}_{(\alpha;\beta)}+\dot{V}_{\alpha}\dot{V}_{\beta}. \tag{4.25}$$

(This equation and (4.23) can be expressed in terms of a general, non-orthonormal, non-Fermi-propagated basis by replacing the ordinary derivatives with Fermi derivatives and projecting everything into the subspace orthogonal to **V**.)

The trace of (4.25) is

$$\frac{\mathrm{d}}{\mathrm{d}s}\theta = -R_{ab}V^aV^b+2\omega^2-2\sigma^2-\tfrac{1}{3}\theta^2+{\dot{V}^a}_{;a}, \tag{4.26}$$

where

$$2\omega^2 = \omega_{ab}\omega^{ab} \geqslant 0,$$

$$2\sigma^2 = \sigma_{ab}\sigma^{ab} \geqslant 0.$$

This equation, which was discovered by Landau and independently by Raychaudhuri, will be of great importance later. From it one sees that vorticity induces expansion as might be expected by analogy with centrifugal force while shear induces contraction. By the field equations, the term $R_{ab}V^aV^b = 4\pi(\mu+3p)$ for a perfect fluid whose flow lines have tangent vectors V^a. Thus one would expect this term also to induce contraction. We shall give a general discussion of the sign of this term in §4.3.

The trace-free part of (4.25) is

$$\frac{\mathrm{D_F}}{\partial s}\sigma_{ab} = -C_{acbd}V^cV^d + \tfrac{1}{2}h_a{}^c h_b{}^d R_{cd} - \omega_{ac}\omega^c{}_b - \sigma_{ac}\sigma^c{}_b$$
$$-\tfrac{2}{3}\theta\sigma_{ab} + h_a{}^c h_b{}^d \dot{V}_{(c;d)} - \tfrac{1}{3}h_{ab}(2\omega^2 - 2\sigma^2 + \dot{V}^a{}_{;a} + \tfrac{1}{2}R_{cd}h^{cd}), \quad (4.27)$$

where C_{abcd} is the Weyl tensor. Since this tensor is trace-free it does not enter directly in the expansion equation (4.26). However since the term $-2\sigma^2$ occurs on the right of the expansion equation, the Weyl tensor produces convergence indirectly by inducing shear. The Riemann tensor can be expressed in terms of the Weyl tensor and the Ricci tensor:

$$R_{abcd} = C_{abcd} - g_{a[d}R_{c]b} - g_{b[c}R_{d]a} - \tfrac{1}{3}Rg_{a[c}g_{d]b}.$$

The Ricci tensor is given by the Einstein equations:

$$R_{ab} - \tfrac{1}{2}g_{ab}R + \Lambda g_{ab} = 8\pi T_{ab}.$$

Thus the Weyl tensor is that part of the curvature which is not determined locally by the matter distribution. However it cannot be entirely arbitrary as the Riemann tensor must satisfy the Bianchi identities:

$$R_{ab[cd;e]} = 0$$

These can be rewritten as

$$C^{abcd}{}_{;d} = J^{abc}, \quad (4.28)$$

where
$$J^{abc} = R^{c[a;b]} + \tfrac{1}{6}g^{c[b}R^{;a]}. \quad (4.29)$$

These equations are rather similar to Maxwell's equations in electrodynamics:

$$F^{ab}{}_{;b} = J^a,$$

where F^{ab} is the electromagnetic field tensor and J^a is the source current. Thus in a sense one could regard the Bianchi identities (4.28) as field equations for the Weyl tensor giving that part of the curvature at a point that depends on the matter distribution at other points. (This approach has been used to analyse the behaviour of gravitational radiation in papers by Newman and Penrose (1962), Newman and Unti (1962) and Hawking (1966*a*).)

5

Exact solutions

Any space–time metric can in a sense be regarded as satisfying Einstein's field equations

$$R_{ab} - \tfrac{1}{2}Rg_{ab} + \Lambda g_{ab} = 8\pi T_{ab}, \tag{5.1}$$

(where we use the units of chapter 3), because, having determined the left-hand side of (5.1) from the metric tensor of the space–time $(\mathcal{M}, \mathbf{g})$, one can *define* T_{ab} as the right-hand side of (5.1). The matter tensor so defined will in general have unreasonable physical properties; the solution will be reasonable only if the matter content is reasonable.

We shall mean by an *exact solution* of Einstein's equations, a space–time $(\mathcal{M}, \mathbf{g})$ in which the field equations are satisfied with T_{ab} the energy–momentum tensor of some specified form of matter which obeys postulate (*a*) ('local causality') of chapter 3, and one of the energy conditions of §4.3. In particular, one may look for exact solutions for empty space ($T_{ab} = 0$), for an electromagnetic field (T_{ab} has the form (3.7)), for a perfect fluid (T_{ab} has the form (3.8)), or for a space containing an electromagnetic field and a perfect fluid. Because of the complexity of the field equations, one cannot find exact solutions except in spaces of rather high symmetry. Exact solutions are also idealized in that any region of space–time is likely to contain many forms of matter, while one can obtain exact solutions only for rather simple matter content. Nevertheless, exact solutions give an idea of the qualitative features that can arise in General Relativity, and so of possible properties of realistic solutions of the field equations. The examples we give will show many types of behaviour which will be of interest in later chapters. We shall discuss solutions with particular reference to their global properties. Many of these global properties have only recently been discovered, although the solutions have been known in a local form for some time.

In § 5.1 and § 5.2 we consider the simplest Lorentz metrics: those of constant curvature. The spatially isotropic and homogeneous cosmological models are described in § 5.3, and their simplest anisotropic generalizations are discussed in § 5.4. It is shown that all such simple models will have a singular origin provided that Λ does not take large positive values. The spherically symmetric metrics which describe the field outside a massive charged or neutral body are examined in § 5.5, and the axially symmetric metrics describing the field outside a special class of massive rotating bodies are described in § 5.6. It is shown that some of the apparent singularities are simply due to a bad choice of coordinates. In § 5.7 we describe the Gödel universe and in § 5.8 the Taub–NUT solutions. These probably do not represent the actual universe but they are of interest because of their pathological global properties. Finally some other exact solutions of interest are mentioned in § 5.9.

5.1 Minkowski space–time

Minkowski space–time $(\mathscr{M}, \eta)$ is the simplest empty space–time in General Relativity, and is in fact the space–time of Special Relativity. Mathematically, it is the manifold R^4 with a flat Lorentz metric η. In terms of the natural coordinates (x^1, x^2, x^3, x^4) on R^4, the metric η can be expressed in the form

$$ds^2 = -(dx^4)^2 + (dx^1)^2 + (dx^2)^2 + (dx^3)^2. \tag{5.2}$$

If one uses spherical polar coordinates (t, r, θ, ϕ) where $x^4 = t$, $x^3 = r\cos\theta$, $x^2 = r\sin\theta\cos\phi$, $x^1 = r\sin\theta\sin\phi$, the metric takes the form

$$ds^2 = -dt^2 + dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2). \tag{5.3}$$

This metric is apparently singular for $r = 0$ and $\sin\theta = 0$; however this is because the coordinates used are not admissible coordinates at these points. To obtain regular coordinate neighbourhoods one has to restrict the coordinates, e.g. to the ranges $0 < r < \infty$, $0 < \theta < \pi$, $0 < \phi < 2\pi$. One needs two such coordinate neighbourhoods to cover the whole of Minkowski space.

An alternative coordinate system is given by choosing advanced and retarded null coordinates v, w defined by $v = t + r$, $w = t - r$ $(\Rightarrow v \geqslant w)$. The metric becomes

$$ds^2 = -dv\, dw + \tfrac{1}{4}(v-w)^2 (d\theta^2 + \sin^2\theta\, d\phi^2), \tag{5.4}$$

where $-\infty < v < \infty$, $-\infty < w < \infty$. The absence in the metric of terms in dv^2, dw^2 corresponds to the fact that the surfaces $\{w = \text{constant}\}$, $\{v = \text{constant}\}$ are null (i.e. $w_{;a} w_{;b} g^{ab} = 0 = v_{;a} v_{;b} g^{ab}$); see figure 12.

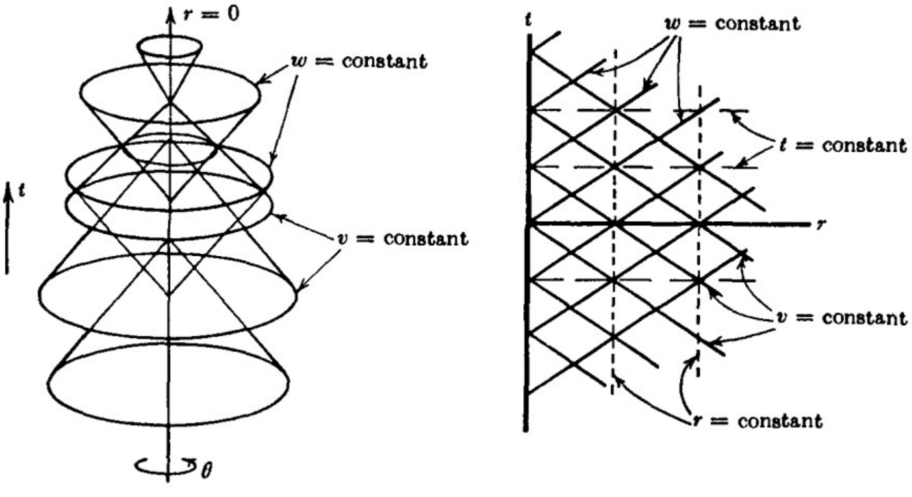

(i) (ii)

FIGURE 12. Minkowski space. The null coordinate $v(w)$ may be thought of as incoming (outgoing) spherical waves travelling at the speed of light; they are advanced (retarded) time coordinates. The intersection of a surface $\{v = \text{constant}\}$ with a surface $\{w = \text{constant}\}$ is a two-sphere.
(i) The v, w coordinate surfaces (one coordinate is suppressed).
(ii) The (t, r) plane; each point represents a two-sphere of radius r.

In a coordinate system in which the metric takes the form (5.2), the geodesics have the form $x^a(v) = b^a v + c^a$ where b^a and c^a are constants. Thus the exponential map $\exp_p\colon T_p \to \mathscr{M}$ is given by

$$x^a(\exp_p \mathbf{X}) = X^a + x^a(p),$$

where X^a are the components of $\mathbf{X}$ with respect to the coordinate basis $\{\partial/\partial x^a\}$ of T_p. Since exp is one–one and onto, it is a diffeomorphism between T_p and $\mathscr{M}$. Thus any two points of $\mathscr{M}$ can be joined by a unique geodesic curve. As exp is defined everywhere on T_p for all p, $(\mathscr{M}, \eta)$ is geodesically complete.

For a spacelike three-surface $\mathscr{S}$, the future (past) Cauchy development $D^+(\mathscr{S})$ $(D^-(\mathscr{S}))$ is defined as the set of all points $q \in \mathscr{M}$ such that each past-directed (future-directed) inextendible non-spacelike curve through q intersects $\mathscr{S}$, cf. § 6.5. If $D^+(\mathscr{S}) \cup D^-(\mathscr{S}) = \mathscr{M}$, i.e. if every inextendible non-spacelike curve in $\mathscr{M}$ intersects $\mathscr{S}$, then $\mathscr{S}$ is said to be a Cauchy surface. In Minkowski space–time, the surfaces $\{x^4 = \text{constant}\}$ are a family of Cauchy surfaces which cover the whole of $\mathscr{M}$. One can however find inextendible spacelike surfaces which are not Cauchy surfaces; for example the surfaces

$$\mathscr{S}_\sigma\colon \{-(x^4)^2 + (x^1)^2 + (x^2)^2 + (x^3)^2 = \sigma = \text{constant}\},$$

where $\sigma < 0, x^4 < 0$, are spacelike surfaces which lie entirely inside the past null cone of the origin O, and so are not Cauchy surfaces (see figure 13). In fact the future Cauchy development of $\mathscr{S}_\sigma$ is the region bounded by $\mathscr{S}_\sigma$ and the past light cone of the origin. By lemma 4.5.2, the timelike geodesics through the origin O are orthogonal to the surfaces $\mathscr{S}_\sigma$. If $r \in D^+(\mathscr{S}_\sigma) \cup D^-(\mathscr{S}_\sigma)$ then the timelike geodesic through r and O is the longest timelike curve between r and $\mathscr{S}_\sigma$. If however r does not lie in $D^+(\mathscr{S}_\sigma) \cup D^-(\mathscr{S}_\sigma)$ there is no longest timelike curve between r and $\mathscr{S}_\sigma$: either r lies in the region $\sigma \geqslant 0$, in which case there is no timelike geodesic through r orthogonal to $\mathscr{S}_\sigma$, or r lies in the region $\sigma < 0$, $x^4 \geqslant 0$, in which case there is a timelike geodesic through r orthogonal to $\mathscr{S}_\sigma$ but this geodesic is not the longest curve between r and $\mathscr{S}_\sigma$ as it contains a conjugate point to $\mathscr{S}_\sigma$ at O (cf. figure 13).

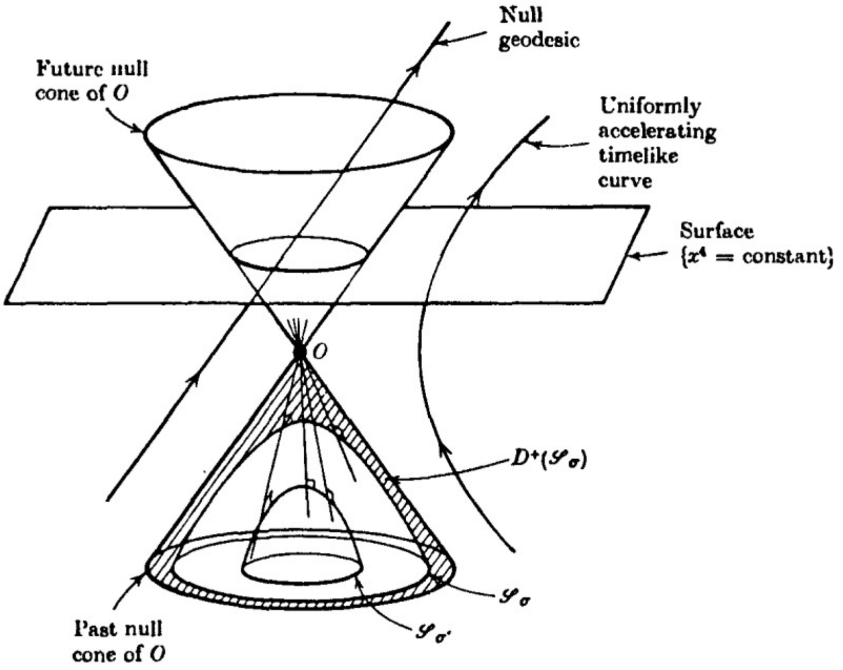
Null geodesic
Future null cone of O
Uniformly accelerating timelike curve
Surface $\{x^4 = \text{constant}\}$
O
$D^+(\mathscr{S}_\sigma)$
$\mathscr{S}_\sigma$
$\mathscr{S}_{\sigma'}$
Past null cone of O

FIGURE 13. A Cauchy surface $\{x^4 = \text{constant}\}$ in Minkowski space-time, and spacelike surfaces $\mathscr{S}_\sigma, \mathscr{S}_{\sigma'}$ which are not Cauchy surfaces. The normal geodesics to the surfaces $\mathscr{S}_\sigma, \mathscr{S}_{\sigma'}$ all intersect at O.

To study the structure of infinity in Minkowski space-time, we shall use the interesting representation of this space-time given by Penrose. From the null coordinates v, w, we define new null coordinates in which the infinities of v, w have been transformed to finite values; thus we define p, q by $\tan p = v$, $\tan q = w$ where $-\frac{1}{2}\pi < p < \frac{1}{2}\pi$, $-\frac{1}{2}\pi < q < \frac{1}{2}\pi$ (and $p \geqslant q$). Then the metric of $(\mathscr{M}, \boldsymbol{\eta})$ takes the form

$$\mathrm{d}s^2 = \sec^2 p \sec^2 q(-\mathrm{d}p\,\mathrm{d}q + \tfrac{1}{4}\sin^2(p-q)(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)).$$

The physical metric $\boldsymbol{\eta}$ is therefore conformal to the metric $\bar{\mathbf{g}}$ given by

$$\mathrm{d}\bar{s}^2 = -4\mathrm{d}p\,\mathrm{d}q + \sin^2(p-q)(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2). \tag{5.5}$$

This metric can be reduced to a more usual form by defining

$$t' = p+q, \quad r' = p-q,$$

where
$$-\pi < t'+r' < \pi, \quad -\pi < t'-r' < \pi, \quad r' \geqslant 0; \tag{5.6}$$

(5.5) is then

$$\mathrm{d}\bar{s}^2 = -(\mathrm{d}t')^2 + (\mathrm{d}r')^2 + \sin^2 r'(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2). \tag{5.7}$$

Thus the whole of Minkowski space–time is given by the region (5.6) of the metric

$$\mathrm{d}s^2 = \tfrac{1}{4}\sec^2(\tfrac{1}{2}(t'+r'))\sec^2(\tfrac{1}{2}(t'-r'))\,\mathrm{d}\bar{s}^2$$

where $\mathrm{d}\bar{s}^2$ is determined by (5.7); the coordinates t, r of (5.3) are related to t', r' by

$$2t = \tan(\tfrac{1}{2}(t'+r')) + \tan(\tfrac{1}{2}(t'-r')),$$
$$2r = \tan(\tfrac{1}{2}(t'+r')) - \tan(\tfrac{1}{2}(t'-r')).$$

Now the metric (5.7) is locally identical to that of the Einstein static universe (see § 5.3), which is a completely homogeneous space–time. One can analytically extend (5.7) to the whole of the Einstein static universe, that is one can extend the coordinates to cover the manifold $R^1 \times S^3$ where $-\infty < t' < \infty$ and r', θ, ϕ are regarded as coordinates on S^3 (with coordinate singularities at $r' = 0$, $r' = \pi$ and $\theta = 0$, $\theta = \pi$ similar to the coordinate singularities in (5.3); these singularities can be removed by transforming to other local coordinates in a neighbourhood of points where (5.7) is singular). On suppressing two dimensions, one can represent the Einstein static universe as the cylinder $x^2+y^2 = 1$ imbedded in a three-dimensional Minkowski space with metric $\mathrm{d}s^2 = -\mathrm{d}t^2 + \mathrm{d}x^2 + \mathrm{d}y^2$ (the full Einstein static universe can be imbedded as the cylinder $x^2+y^2+z^2+w^2 = 1$ in a five-dimensional Euclidean space with metric $\mathrm{d}s^2 = -\mathrm{d}t^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 + \mathrm{d}w^2$, cf. Robertson (1933)).

One therefore has the situation: the whole of Minkowski space–time is conformal to the region (5.6) of the Einstein static universe, that is, to the shaded area in figure 14. The boundary of this region may therefore be thought of as representing the conformal structure of infinity of Minkowski space–time. It consists of the null surfaces $p = \frac{1}{2}\pi$ (labelled $\mathscr{I}^+$) and $q = -\frac{1}{2}\pi$ (labelled $\mathscr{I}^-$) together with points $p = \frac{1}{2}\pi$, $q = \frac{1}{2}\pi$ (labelled i^+), $p = \frac{1}{2}\pi$, $q = -\frac{1}{2}\pi$ (labelled i^0) and $p = -\frac{1}{2}\pi$, $q = -\frac{1}{2}\pi$ (labelled i^-). Any future-directed timelike geodesic in Minkowski space approaches i^+ (i^-) for indefinitely large positive (negative) values of its affine parameter, so one can regard any timelike geodesic as originating at i^- and finishing at i^+ (cf. figure 15(i)). Similarly one can regard null geodesics as originating at $\mathscr{I}^-$ and ending at $\mathscr{I}^+$, while spacelike geodesics both originate and end at i^0. Thus one may regard i^+ and i^- as representing future and past timelike infinity, $\mathscr{I}^+$ and $\mathscr{I}^-$ as representing future and past null infinity, and i^0 as representing spacelike infinity. (However non-geodesic curves do not

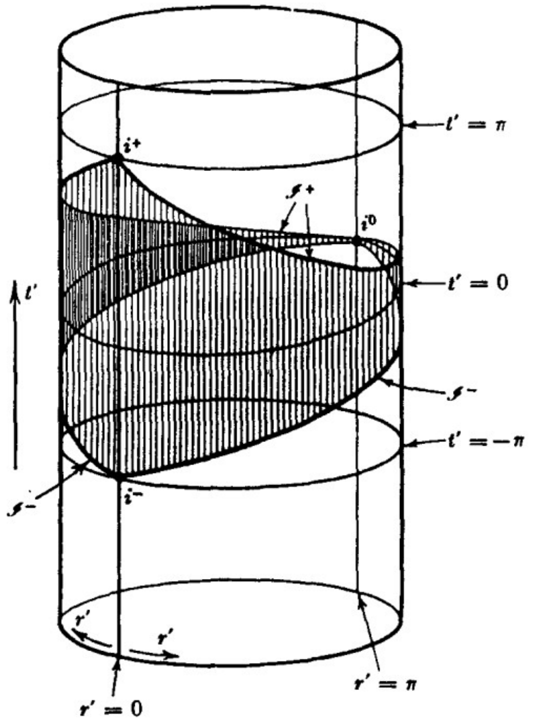

FIGURE 14. The Einstein static universe represented by an imbedded cylinder; the coordinates θ, ϕ have been suppressed. Each point represents one half of a two-sphere of area $4\pi \sin^2 r'$. The shaded region is conformal to the whole of Minkowski space–time; its boundary (part of the null cones of i^+, i^0 and i^-) may be regarded as the conformal infinity of Minkowski space–time.

obey these rules; e.g. non-geodesic timelike curves may start on $\mathscr{I}^-$ and end on $\mathscr{I}^+$.) Since any Cauchy surface intersects all timelike and null geodesics, it is clear that it will appear as a cross-section of the space everywhere reaching the boundary at i^0.

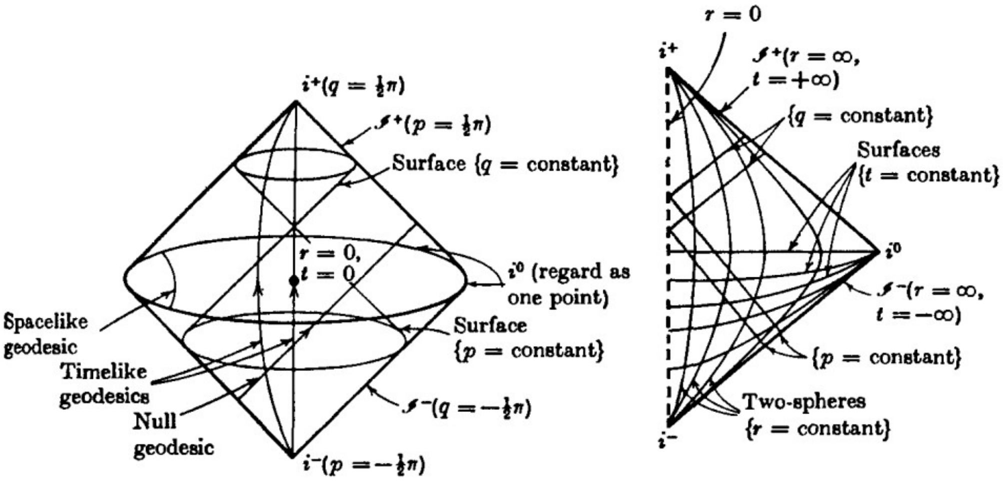
$i^+(q = \frac{1}{2}\pi)$
$\mathscr{I}^+(p = \frac{1}{2}\pi)$
Surface $\{q = \text{constant}\}$
$r = 0,$
$t = 0$
i^0 (regard as one point)
Spacelike geodesic
Surface $\{p = \text{constant}\}$
Timelike geodesics
Null geodesic
$\mathscr{I}^-(q = -\frac{1}{2}\pi)$
$i^-(p = -\frac{1}{2}\pi)$
$r = 0$
i^+
$\mathscr{I}^+(r = \infty, t = +\infty)$
$\{q = \text{constant}\}$
Surfaces $\{t = \text{constant}\}$
i^0
$\mathscr{I}^-(r = \infty, t = -\infty)$
$\{p = \text{constant}\}$
Two-spheres $\{r = \text{constant}\}$
i^-

(i) (ii)

FIGURE 15

(i) The shaded region of figure 14, with only one coordinate suppressed, representing Minkowski space–time and its conformal infinity.

(ii) The Penrose diagram of Minkowski space–time; each point represents a two-sphere, except for i^+, i^0 and i^-, each of which is a single point, and points on the line $r = 0$ (where the polar coordinates are singular).

One can also represent the conformal structure of infinity by drawing a diagram of the (t', r') plane, see figure 15 (ii). As in figure 12 (ii), each point of this diagram represents a sphere S^2, and radial null geodesics are represented by straight lines at ± 45°. In fact, the structure of infinity in any spherically symmetric space–time can be represented by a diagram of this sort, which we shall call a *Penrose diagram*. On such diagrams we shall represent infinity by single lines, the origin of polar coordinates by dotted lines, and irremovable singularities of the metric by double lines.

The conformal structure of Minkowski space we have described is what one would regard as the 'normal' behaviour of a space–time at infinity; we shall encounter different types of behaviour in later sections.

Finally, we mention that one can obtain spaces locally identical to $(\mathscr{M}, \eta)$ but with different (large scale) topological properties by identifying points in $\mathscr{M}$ which are equivalent under a discrete isometry without a fixed point (e.g. identifying the point (x^1, x^2, x^3, x^4) with the point $(x^1, x^2, x^3, x^4 + c)$, where c is a constant, changes the topological structure from R^4 to $R^3 \times S^1$, and introduces closed timelike lines into the space–time). Clearly, $(\mathscr{M}, \eta)$ is the universal covering space for all such derived spaces, which have been studied in detail by Auslander and Markus (1958).

5.2 De Sitter and anti-de Sitter space–times

The space–time metrics of constant curvature are locally characterized by the condition $R_{abcd} = \frac{1}{12}R(g_{ac}g_{bd} - g_{ad}g_{bc})$. This equation is equivalent to $C_{abcd} = 0 = R_{ab} - \frac{1}{4}Rg_{ab}$; thus the Riemann tensor is determined by the Ricci scalar R alone. It follows at once from the contracted Bianchi identities that R is constant throughout space–time; in fact these space–times are homogeneous. The Einstein tensor is

$$R_{ab} - \tfrac{1}{2}Rg_{ab} = -\tfrac{1}{4}Rg_{ab}.$$

One can therefore regard these spaces as solutions of the field equations for an empty space with $\Lambda = \frac{1}{4}R$, or for a perfect fluid with a constant density $R/32\pi$ and a constant pressure $-R/32\pi$. However the latter choice does not seem reasonable, as in this case one cannot have both the density and the pressure positive; in addition, the equation of motion (3.10) is indeterminate for such a fluid.

The space of constant curvature with $R = 0$ is Minkowski space–time. The space for $R > 0$ is *de Sitter space–time*, which has the topology $R^1 \times S^3$ (see Schrödinger (1956) for an interesting account of this space). It is easiest visualized as the hyperboloid

$$-v^2 + w^2 + x^2 + y^2 + z^2 = \alpha^2$$

in flat five-dimensional space R^5 with metric

$$-\mathrm{d}v^2 + \mathrm{d}w^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 = \mathrm{d}s^2$$

(see figure 16). One can introduce coordinates (t, χ, θ, ϕ) on the hyperboloid by the relations

$$\alpha \sinh(\alpha^{-1}t) = v, \quad \alpha \cosh(\alpha^{-1}t) \cos\chi = w,$$

$$\alpha \cosh(\alpha^{-1}t) \sin\chi \cos\theta = x, \quad \alpha \cosh(\alpha^{-1}t) \sin\chi \sin\theta \cos\phi = y,$$

$$\alpha \cosh(\alpha^{-1}t) \sin\chi \sin\theta \sin\phi = z.$$

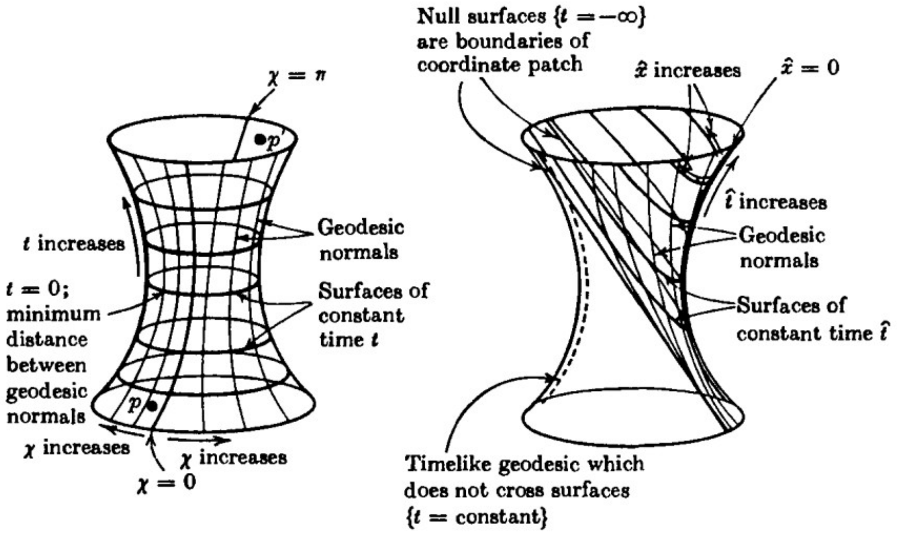
Null surfaces $\{t = -\infty\}$
are boundaries of
coordinate patch
$\chi = \pi$
p'
$\hat{x}$ increases
$\hat{x} = 0$
$\hat{t}$ increases
t increases
Geodesic
normals
Geodesic
normals
$t = 0$;
minimum
distance
between
geodesic
normals
Surfaces of
constant
time t
Surfaces of
constant time $\hat{t}$
p
χ increases
χ increases
$\chi = 0$
Timelike geodesic which
does not cross surfaces
$\{t = \text{constant}\}$

(i) (ii)

FIGURE 16. De Sitter space–time represented by a hyperboloid imbedded in a five-dimensional flat space (two dimensions are suppressed in the figure).

(i) Coordinates (t, χ, θ, ϕ) cover the whole hyperboloid; the sections $\{t = \text{constant}\}$ are surfaces of curvature $k = +1$.

(ii) Coordinates $(\hat{t}, \hat{x}, \hat{y}, \hat{z})$ cover half the hyperboloid; the surfaces $\{\hat{t} = \text{constant}\}$ are flat three-spaces, their geodesic normals diverging from a point in the infinite past.

In these coordinates, the metric has the form

$$ds^2 = -dt^2 + \alpha^2 . \cosh^2(\alpha^{-1}t) . \{d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta\, d\phi^2)\}.$$

The singularities in the metric at $\chi = 0$, $\chi = \pi$ and at $\theta = 0$, $\theta = \pi$, are simply those that occur with polar coordinates. Apart from these trivial singularities, the coordinates cover the whole space for $-\infty < t < \infty$, $0 \leqslant \chi \leqslant \pi$, $0 \leqslant \theta \leqslant \pi$, $0 \leqslant \phi \leqslant 2\pi$. The spatial sections of constant t are spheres S^3 of constant positive curvature and are Cauchy surfaces. Their geodesic normals are lines which contract monotonically to a minimum spatial separation and then re-expand to infinity (see figure 16 (i)).

One can also introduce coordinates

$$\hat{t} = \alpha \log \frac{w+v}{\alpha}, \quad \hat{x} = \frac{\alpha x}{w+v}, \quad \hat{y} = \frac{\alpha y}{w+v}, \quad \hat{z} = \frac{\alpha z}{w+v}$$

on the hyperboloid. In these coordinates, the metric takes the form

$$ds^2 = -d\hat{t}^2 + \exp(2\alpha^{-1}\hat{t})(d\hat{x}^2 + d\hat{y}^2 + d\hat{z}^2).$$

However these coordinates cover only half the hyperboloid as $\hat{t}$ is not defined for $w+v \leqslant 0$ (see figure 16 (ii)).

The region of de Sitter space for which $v+w > 0$ forms the space–time for the *steady state* model of the universe proposed by Bondi and Gold (1948) and Hoyle (1948). In this model, the matter is supposed to move along the geodesic normals to the surfaces $\{\hat{t} = \text{constant}\}$. As the matter moves further apart, it is assumed that more matter is continuously created to maintain the density at a constant value. Bondi and Gold did not seek to provide field equations for this model, but Pirani (1955), and Hoyle and Narlikar (1964) have pointed out that the metric can be considered as a solution of the Einstein equations (with $\Lambda = 0$) if in addition to the ordinary matter one introduces a scalar field of negative energy density. This 'C'-field would also be responsible for the continual creation of matter.

The steady state theory has the advantage of making simple and definite predictions. However from our point of view there are two unsatisfactory features. The first is the existence of negative energy, which was discussed in § 4.3. The other is the fact that the space–time is extendible, being only half of de Sitter space. Despite these aesthetic objections, the real test of the steady state theory is whether its predictions agree with observations or not. At the moment it seems that they do not, though the observations are not yet quite conclusive.

de Sitter space is geodesically complete; however, there are points in the space which cannot be joined to each other by any geodesic. This is in contrast to spaces with a positive definite metric, when geodesic completeness guarantees that any two points of a space can be joined by at least one geodesic. The half of de Sitter space which represents the steady state universe is not complete in the past (there are geodesics which are complete in the full space, and cross the boundary of the steady state region; they are therefore incomplete in that region).

To study infinity in de Sitter space–time, we define a time coordinate t' by

$$t' = 2\arctan(\exp \alpha^{-1} t) - \tfrac{1}{2}\pi,$$

where

$$-\tfrac{1}{2}\pi < t' < \tfrac{1}{2}\pi. \tag{5.8}$$

Then

$$ds^2 = \alpha^2 \cosh^2(\alpha^{-1}t') \, . \, d\bar{s}^2,$$

where $d\bar{s}^2$ is given by (5.7) on identifying $r' = \chi$. Thus the de Sitter space is conformal to that part of the Einstein static universe defined by (5.8) (see figure 17 (i)). The Penrose diagram of de Sitter space is accordingly as in figure 17 (ii). One half of this figure gives the Penrose

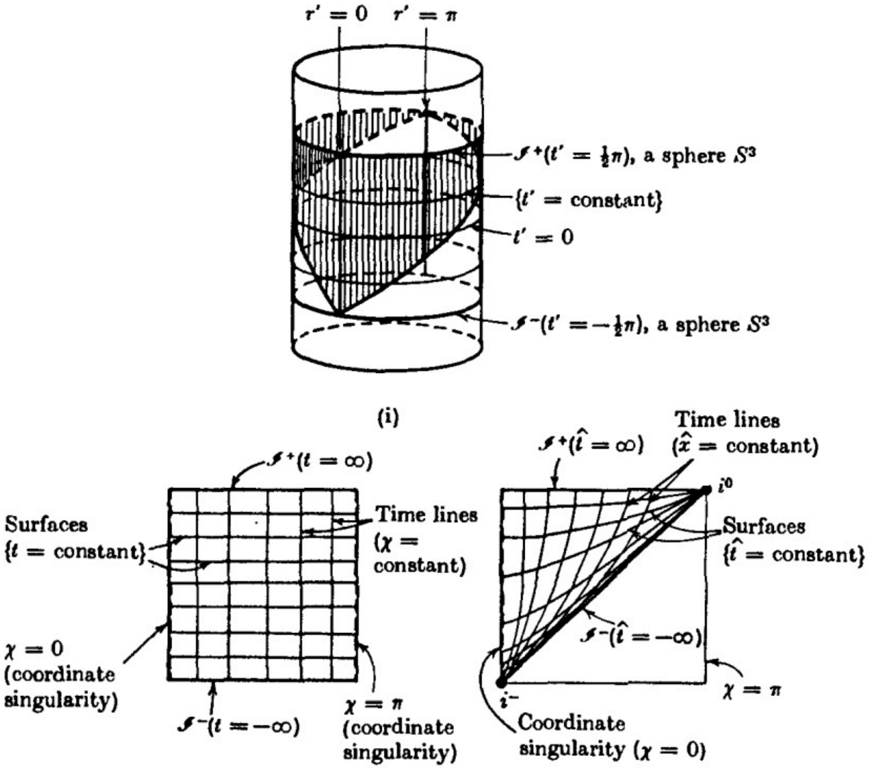
$r' = 0$
$r' = \pi$
$\mathscr{I}^+(t' = \frac{1}{2}\pi)$, a sphere S^3
$\{t' = \text{constant}\}$
$t' = 0$
$\mathscr{I}^-(t' = -\frac{1}{2}\pi)$, a sphere S^3
(i)
$\mathscr{I}^+(t = \infty)$
Surfaces $\{t = \text{constant}\}$
Time lines ($\chi =$ constant)
$\chi = 0$ (coordinate singularity)
$\mathscr{I}^-(t = -\infty)$
$\chi = \pi$ (coordinate singularity)
$\mathscr{I}^+(\hat{t} = \infty)$
Time lines ($\hat{x} =$ constant)
i^0
Surfaces $\{\hat{t} = \text{constant}\}$
$\mathscr{I}^-(\hat{t} = -\infty)$
$\chi = \pi$
i^-
Coordinate singularity ($\chi = 0$)

(ii) (iii)

FIGURE 17

(i) De Sitter space-time is conformal to the region $-\frac{1}{2}\pi < t' < \frac{1}{2}\pi$ of the Einstein static universe. The steady state universe is conformal to the shaded region.

(ii) The Penrose diagram of de Sitter space-time.

(iii) The Penrose diagram of the steady state universe.

In (ii), (iii) each point represents a two-sphere of area $2\pi \sin^2 \chi$; null lines are at 45°. $\chi = 0$ and $\chi = \pi$ are identified.

diagram of the half of de Sitter space-time which constitutes the steady state universe (figure 17 (iii)).

One sees that de Sitter space has, in contrast to Minkowski space, a spacelike infinity for timelike and null lines, both in the future and the past. This difference corresponds to the existence in de Sitter space-time of both particle and event horizons for geodesic families of observers.

In de Sitter space, consider a family of particles whose histories are timelike geodesics; these must originate at the spacelike infinity $\mathscr{I}^-$ and end at the spacelike infinity $\mathscr{I}^+$. Let p be some event on the world-

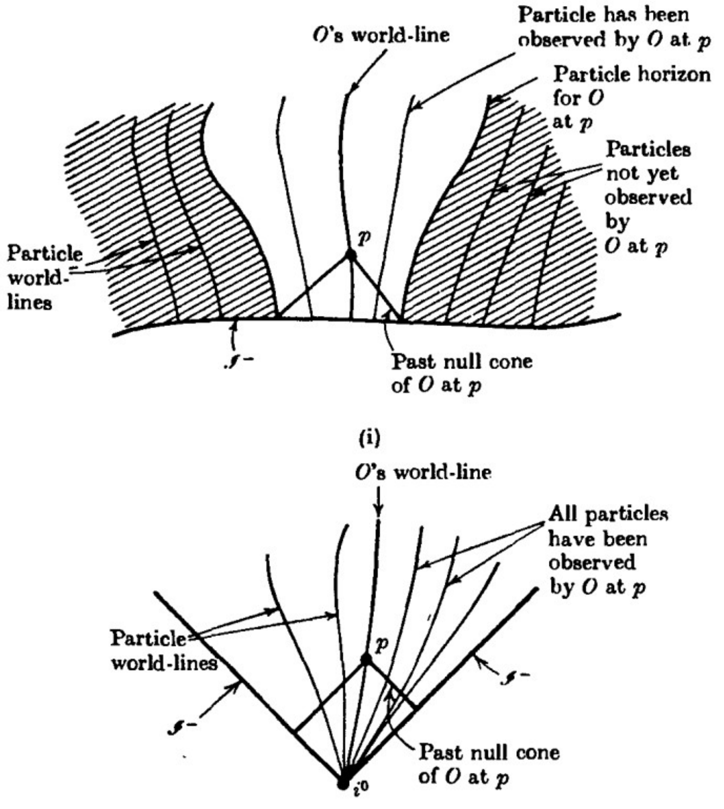
O's world-line
Particle has been observed by O at p
Particle horizon for O at p
Particles not yet observed by O at p
p
Particle world-lines
$\mathscr{I}^-$
Past null cone of O at p
O's world-line
All particles have been observed by O at p
Particle world-lines
p
$\mathscr{I}^-$
$\mathscr{I}^-$
Past null cone of O at p
i^0

(i)

(ii)

FIGURE 18

(i) The particle horizon defined by a congruence of geodesic curves when past null infinity $\mathscr{I}^-$ is spacelike.

(ii) Lack of such a horizon if $\mathscr{I}^-$ is null.

line of a particle O in this family, i.e. some time in its history (proper time measured along O's world-line). The past null cone of p is the set of events in space–time which can be observed by O at that time. The world-lines of some other particles may intersect this null cone; these particles are visible to O. However, there can exist particles whose world-lines do not intersect this null cone, and so are not yet visible to O. At a later time O can observe more particles, but there still exist particles not visible to O at that time. We say that the division of particles into those seen by O at p and those not seen by O at p, is the *particle horizon* for the observer O at the event p; it represents the history of those particles lying at the limits of O's vision. Note that it is determined only when the world-lines of all the particles in the family are known. If some particle lies on the horizon, then the event p is the event at which the particle's creation light cone intersects O's world-line. In Minkowski space, on the other hand, all the other particles are visible at any event p on O's world-line if they move on timelike geodesics. As long as one considers only families of geodesic observers, one may think of the existence of the particle horizon as a consequence of past null infinity being spacelike (see figure 18).

All events outside the past null cone of p are events which are not, and never have been, observable by O up to the time represented by the event p. There is a limit to O's world-line on $\mathscr{I}^+$. In de Sitter space–time, the past null cone of this point (obtained by a limiting process in the actual space–time, or directly from the conformal space–time) is a boundary between events which will at some time be observable by O, and those that will never be observable by O. We call this surface the *future event horizon* of the world-line. It is the boundary of the past of the world-line. In Minkowski space–time, on the other hand, the limiting null cone of any geodesic observer includes the whole of space–time, so there are no events which a geodesic observer will never be able to see. However if an observer moves with uniform acceleration his world-line may have a future event horizon. One may think of the existence of a future event horizon for a geodesic observer as being a consequence of $\mathscr{I}^+$ being spacelike (see figure 19).

Consider the event horizon for the observer O in de Sitter space–time and suppose that at some proper time (event p) on his world-line, his light cone intersects the world-line of the particle Q. Then Q is always visible to O at times after p. However there is on Q's world-line an event r which lies on O's future event horizon; O can never see later events on Q's world-line than r. Moreover an infinite proper time elapses on O's world-line from any given point till he observes r, but a finite proper time elapses along Q's world-line from any given event to r, which is a perfectly ordinary event on his world-line. Thus O sees a finite part of Q's history in an infinite time; expressed more physically, as O observes Q he sees a redshift which approaches infinity as O observes points on Q's world-line which approach r. Correspondingly, Q never sees beyond some point on O's world-line, and sees nearby points on O's world-line only with a very large redshift.

At any point on O's world-line, the future null cone is the boundary of the set of events in space–time which O can influence at and after that time. To obtain the maximal set of events in space–time that O could at any time influence, we take the future light cone of the limit

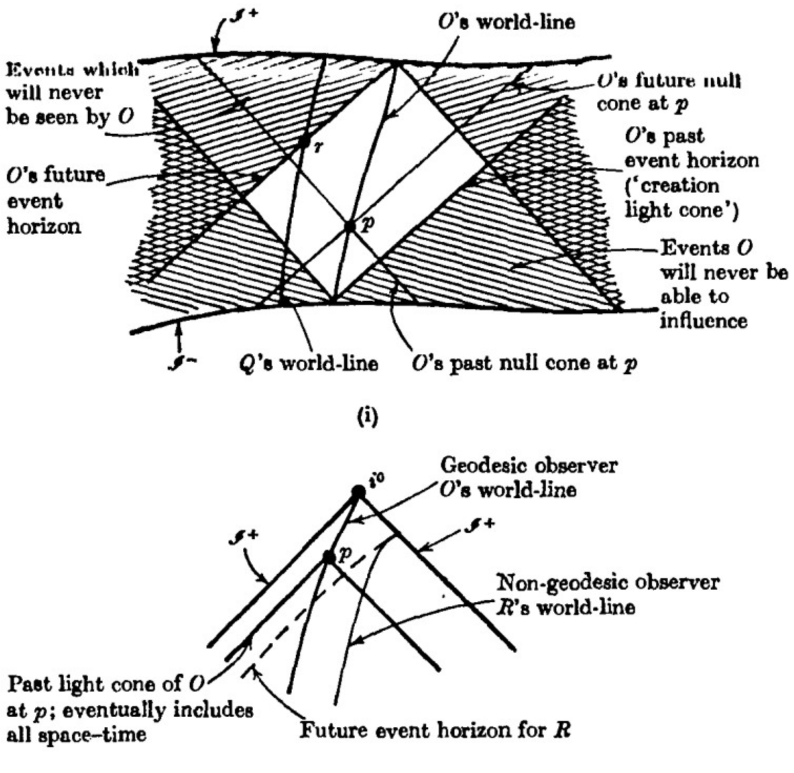
$\mathscr{I}^+$
O's world-line
Events which will never be seen by O
O's future null cone at p
O's past event horizon ('creation light cone')
O's future event horizon
r
p
Events O will never be able to influence
$\mathscr{I}^-$
Q's world-line
O's past null cone at p
(i)
i^0
Geodesic observer O's world-line
$\mathscr{I}^+$
p
$\mathscr{I}^+$
Non-geodesic observer R's world-line
Past light cone of O at p; eventually includes all space–time
Future event horizon for R
(ii)

FIGURE 19

(i) The future event horizon for a particle O which exists when future infinity $\mathscr{I}^+$ is spacelike; also the past event horizon which exists when past infinity $\mathscr{I}^-$ is spacelike.

(ii) If future infinity consists of a null $\mathscr{I}^+$ and i^0, there is no future event horizon for a geodesic observer O. However an accelerating observer R may have a future event horizon.

point of O's world-line on past infinity $\mathscr{I}^-$; that is, we take the boundary of the future of the world-line (which can be regarded as O's creation light cone). This has a non-trivial existence for a geodesic observer only if the past infinity $\mathscr{I}^-$ is spacelike (and is in fact then O's past event horizon). It is clear from the above discussion that in the steady state universe, which has a null past infinity for timelike and null geodesics and a spacelike future infinity, any fundamental observer has a future event horizon but no past particle horizon.

One can obtain other spaces which are locally equivalent to the de Sitter space, by identifying points in de Sitter space. The simplest such identification is to identify antipodal points p, p' (see figure 16) on the hyperboloid. The resulting space is not time orientable; if time increases in the direction of the arrow at p, the antipodal identification implies it must increase in the direction of the arrow at p', but one cannot continuously extend this identification of future and past half null cones over the whole hyperboloid. Calabi and Markus (1962) have studied in detail the spaces resulting from such identifications; they show in particular that an arbitrary point in the resulting space can be joined to any other point by a geodesic if and only if it is not time orientable.

The space of constant curvature with $R < 0$ is called *anti-de Sitter space*. It has the topology $S^1 \times R^3$, and can be represented as the hyperboloid

$$-u^2 - v^2 + x^2 + y^2 + z^2 = 1$$

in the flat five-dimensional space R^5 with metric

$$\mathrm{d}s^2 = -(\mathrm{d}u)^2 - (\mathrm{d}v)^2 + (\mathrm{d}x)^2 + (\mathrm{d}y)^2 + (\mathrm{d}z)^2.$$

There are closed timelike lines in this space; however it is not simply connected, and if one unwraps the circle S^1 (to obtain its covering space R^1) one obtains the universal covering space of anti-de Sitter space which does not contain any closed timelike lines. This has the topology of R^4. We shall in future mean by 'anti-de Sitter space', this universal covering space.

It can be represented by the metric

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + \cos^2 t\,\{\mathrm{d}\chi^2 + \sinh^2\chi(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\}. \quad (5.9)$$

This coordinate system covers only part of the space, and has apparent singularities at $t = \pm\frac{1}{2}\pi$. The whole space can be covered by coordinates $\{t', r, \theta, \phi\}$ for which the metric has the static form

$$\mathrm{d}s^2 = -\cosh^2 r\,\mathrm{d}t'^2 + \mathrm{d}r^2 + \sinh^2 r(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2).$$

In this form, the space is covered by the surfaces $\{t' = \text{constant}\}$ which have non-geodesic normals.

To study the structure at infinity, define the coordinate r' by

$$r' = 2\arctan(\exp r) - \tfrac{1}{2}\pi, \quad 0 \leqslant r' < \tfrac{1}{2}\pi.$$

Then one finds $\mathrm{d}s^2 = \cosh^2 r\,\mathrm{d}\bar{s}^2$, where $\mathrm{d}\bar{s}^2$ is given by (5.7); that is, the whole of anti-de Sitter space is conformal to the region $0 \leqslant r' < \frac{1}{2}\pi$ of the Einstein static cylinder. The Penrose diagram is shown in figure 20; null and spacelike infinity can be thought of as a timelike surface in this case. This surface has the topology $R^1 \times S^2$.

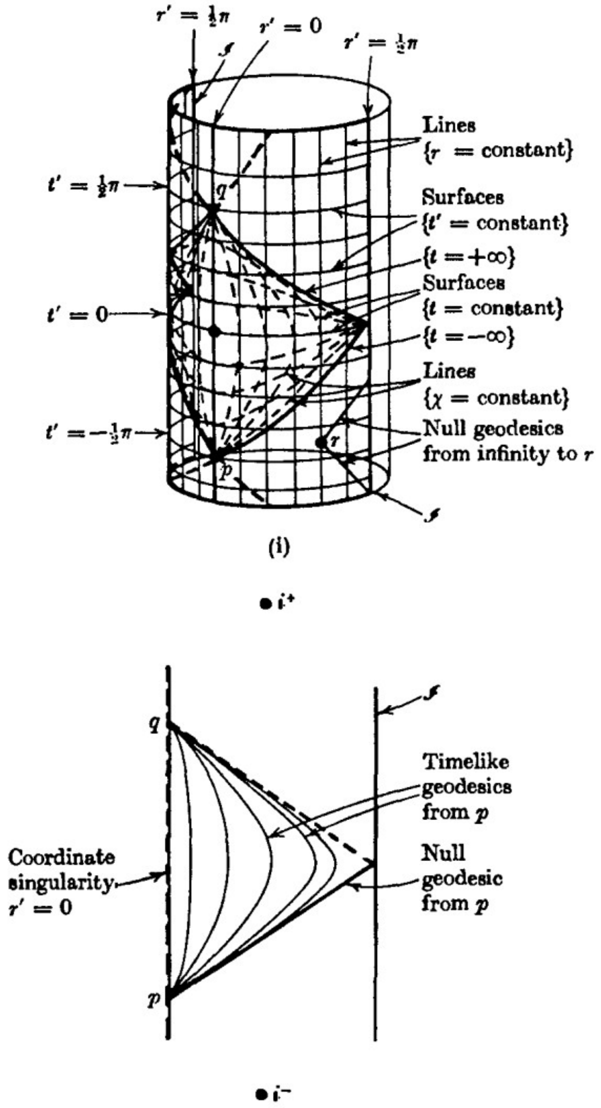
$r' = \frac{1}{2}\pi$
$r' = 0$
$r' = \frac{1}{2}\pi$
$\mathscr{I}$
Lines $\{r = \text{constant}\}$
Surfaces $\{t' = \text{constant}\}$
$\{t = +\infty\}$
Surfaces $\{t = \text{constant}\}$
$\{t = -\infty\}$
Lines $\{\chi = \text{constant}\}$
Null geodesics from infinity to r
$\mathscr{I}$
$t' = \frac{1}{2}\pi$
$t' = 0$
$t' = -\frac{1}{2}\pi$
q
p
r
i^+
$\mathscr{I}$
q
Timelike geodesics from p
Null geodesic from p
Coordinate singularity $r' = 0$
p
i^-

(i)

(ii)

FIGURE 20

(i) Universal anti-de Sitter space is conformal to one half of the Einstein static universe. While coordinates (t', r, θ, ϕ) cover the whole space, coordinates (t, χ, θ, ϕ) cover only one diamond-shaped region as shown. The geodesics orthogonal to the surfaces $\{t = \text{constant}\}$ all converge at p and q, and then diverge out into similar diamond-shaped regions.

(ii) The Penrose diagram of universal anti-de Sitter space. Infinity consists of the timelike surface $\mathscr{I}$ and the disjoint points i^+, i^-. The projection of some timelike and null geodesics is shown.

One cannot find a conformal transformation which makes timelike infinity finite without pinching off the Einstein static universe to a point (if a conformal transformation makes the time coordinate finite it also scales the space sections by an infinite factor), so we represent timelike infinity by the disjoint points i^+, i^-.

The lines $\{\chi, \theta, \phi \text{ constant}\}$ are the geodesics orthogonal to the surfaces $\{t = \text{constant}\}$; they all converge to points q (respectively, p) in the future (respectively, past) of the surface, and this convergence is the reason for the apparent (coordinate) singularities in the original metric form. The region covered by these coordinates is the region between the surface $t = 0$ and the null surfaces on which these normals become degenerate.

The space has two further interesting properties. First, as a consequence of the timelike infinity, there exists no Cauchy surface whatever in the space. While one can find families of spacelike surfaces (such as the surfaces $\{t' = \text{constant}\}$) which cover the space completely, each surface being a complete cross-section of the space-time, one can find null geodesics which never intersect any given surface in the family. Given initial data on any such surface, one cannot predict beyond the Cauchy development of the surface; thus from the surface $\{t = 0\}$, one can predict only in the region covered by the coordinates t, χ, θ, ϕ. Any attempt to predict beyond this region is prevented by fresh information coming in from the timelike infinity.

Secondly, corresponding to the fact that the geodesic normals from $t = 0$ all converge at p and q, all the past timelike geodesics from p expand out (normal to the surfaces $\{t = \text{constant}\}$) and reconverge at q. In fact, all the timelike geodesics from any point in this space (to either the past or future) reconverge to an image point, diverging again from this image point to refocus at a second image point, and so on. The future timelike geodesics from p therefore never reach $\mathscr{I}$, in contrast to the future null geodesics which go to $\mathscr{I}$ from p and form the boundary of the future of p. This separation of timelike and null geodesics results in the existence of regions in the future of p (i.e. which can be reached from p by a future-directed timelike line) which cannot be reached from p by any geodesic. The set of points which can be reached by future-directed timelike lines from p is the set of points lying beyond the future null cone of p; the set of points which can be reached from p by future-directed timelike geodesics is the interior of the infinite chain of diamond-shaped regions similar to that covered by coordinates (t, χ, θ, ϕ). One notes that all points in the Cauchy development of the surface $t = 0$ can be reached from this surface by a unique geodesic normal to this surface, but that a general point outside this Cauchy development cannot be reached by any geodesic normal to the surface.

5.3 Robertson–Walker spaces

So far, we have not considered the relation of exact solutions to the physical universe. Following Einstein, we can ask: can one find space–times which are exact solutions for some suitable form of matter and which give a good representation of the large scale properties of the observable universe? If so, we can claim to have a reasonable 'cosmological model' or model of the physical universe.

However we are not able to make cosmological models without some admixture of ideology. In the earliest cosmologies, man placed himself in a commanding position at the centre of the universe. Since the time of Copernicus we have been steadily demoted to a medium sized planet going round a medium sized star on the outer edge of a fairly average galaxy, which is itself simply one of a local group of galaxies. Indeed we are now so democratic that we would not claim that our position in space is specially distinguished in any way. We shall, following Bondi (1960), call this assumption the *Copernican principle*.

A reasonable interpretation of this somewhat vague principle is to understand it as implying that, when viewed on a suitable scale, the universe is approximately spatially homogeneous.

By spatially homogeneous, we mean there is a group of isometries which acts freely on $\mathscr{M}$, and whose surfaces of transitivity are spacelike three-surfaces; in other words, any point on one of these surfaces is equivalent to any other point on the same surface. Of course, the universe is not exactly spatially homogeneous; there are local irregularities, such as stars and galaxies. Nevertheless it might seem reasonable to suppose that the universe is spatially homogeneous on a large enough scale.

While one can build mathematical models fulfilling this requirement of homogeneity (see next section), it is difficult to test homogeneity directly by observation, as there is no simple way of measuring the separation between us and distant objects. This difficulty is eased by the fact that we can, in principle, fairly easily observe *isotropies* in extragalactic observations (i.e. we can see if these observations are the same in different directions, or not), and isotropies are closely con-

'()*' ' '+ ' . ! +4 1
2 6663 - A ! -!
& 8 6 8 0 4

modes will be the solutions in a small neighbourhood about the saddle point permitted by the linearised equations (given existence of solutions to the non-linear equations, linearisation will not prevent corresponding linearised solutions existing).)

The point is that these modes can exist as linearisations of the FLRW model; if they do not occur, then initial data has been chosen to set these modes precisely to zero (rather than being made very small), which requires very special initial conditions. Thus these modes will occur in almost all almost-FLRW universes. Hence, if one believes in generality arguments, they will occur in the real universe. When they occur, they will at early and late times grow until the model is very far from a FLRW geometry (while being arbitrarily close to an FLRW model for a very long time, as per the previous theorem).

Exercise: Most studies of CBR anisotropies and nucleosynthesis are carried out for the Bianchi types that allow FLRW models as special cases (see Figure 3). Show that Bianchi models can approximate FLRW models for extended periods even if they do not belong to those types. What kinds of CBR anisotropies can occur in these models? (See e.g. [51].)

7. Almost-FLRW universes

The real universe is *not* FLRW because of all the structure it contains, and (because of the non-linearity of the EFE) the other exact solutions we can attain have higher symmetry than the real universe. Thus in order to obtain realistic models we can compare with detailed observations we need to approximate, aiming to obtain 'almost-FLRW' models representing a universe that is FLRW-like on a large scale but allowing for generic inhomogeneities on a small scale.

7.1. GAUGE PROBLEM

The major problem in studying perturbed models is the *gauge problem*, due to the fact that there is *no* identifiable fixed background model in General Relativity. One can start with a unique FLRW universe model with metric $\overline{g}_{ab}$ in some coordinate system, and perturb it to obtain a more realistic model: $\overline{g}_{ab} \to g_{ab} = \overline{g}_{ab} + \delta g_{ab}$, but then the process has no unique inverse: the background model $\overline{g}_{ab}$ is not uniquely determined by the lumpy universe model g_{ab} (no unique tensorial averaging process has been defined that will recover $\overline{g}_{ab}$ from g_{ab}). Many choices can be made. However, the usual variables describing perturbations depend on the way the (fictitious) background model $\overline{g}_{ab}$ is fitted to the real universe g_{ab}; these variables can be given any values one wants by changing this correspondence.[26] For example, the *dimensionless density contrast* δ representing a density perturbation is usually defined by

$$\delta(x) = \frac{\mu(x) - \overline{\mu}(x)}{\mu(x)} , \tag{184}$$

where $\mu(x)$ is the actual value of the density at the point x, while $\overline{\mu}(x)$ is the (fictitious) background value there, determined by the chosen mapping of the background model into the realistic lumpy model (Ellis and Bruni [197]). This quantity can be given any value we desire by altering that map; we can for example set it to zero by choosing the real surfaces of constant density to be the background surfaces of constant time and, hence, of constant density. Consequently, perturbation equations written in terms of this variable have as solution both physical modes and gauge modes, the latter corresponding to variation of gauge choice rather than to physical variation.

One way to solve this is by very carefully keeping track of the gauge used and the resulting gauge freedom; see Ma and Bertschinger [198] and Prof. Bertschinger's lectures here. The alternative is to use gauge-invariant variables. A widely used and fundamentally important set of such variables are those introduced by Bardeen [199], and used for example by Bardeen, Steinhardt and Turner [200]. Another possibility is use of Gauge-Invariant and 1+3 Covariant ('GIC') variables, i.e. variables that are gauge-invariant and also $1 + 3$ covariantly defined so that they have a clear geometrical meaning, and can be examined in any desired coordinate system. That is what will be pursued here.

In more detail: our aim is to examine perturbed models by using $1 + 3$ covariant variables defined in the real space-time (not the background), deriving exact equations for these variables in that space-time, and then approximating by linearising about a RW geometry to get the linearised equations describing the evolution of density inhomogeneities in almost-FLRW universes. How do we handle gauge invariance in this approach? We rely on the

> **Gauge Invariance Lemma** (Stewart and Walker [201]): If a quantity $T^{\cdots}{}_{\cdots}$ vanishes in the background space-time, then it is gauge-invariant (to all orders).

---

[26] This is often represented implicitly rather than explicitly, by assuming that points with the same coordinate values in the background space and more realistic model map to each other; then the gauge freedom is contained in the coordinate freedom available in the realistic universe model.

[The proof is straightforward : If $\overline{T}^{\cdots}{}_{\cdots} = 0$, then $\delta T^{\cdots}{}_{\cdots} = T^{\cdots}{}_{\cdots} - \overline{T}^{\cdots}{}_{\cdots} = T^{\cdots}{}_{\cdots}$, which is manifestly independent of the mapping Φ from $\overline{S}$ to S (it does not matter how we map $\overline{T}^{\cdots}{}_{\cdots}$ from $\overline{S}$ to S when $\overline{T}^{\cdots}{}_{\cdots}$ vanishes).] The application to almost-FLRW models follows (see Ellis and Bruni [197]), where we use an order-of-magnitude notation as follows: Given a smallness parameter ϵ, $\mathcal{O}[n]$ denotes $\mathcal{O}(\epsilon^n)$, and $A \approx B$ means $A - B = \mathcal{O}[2]$ (i.e. these variables are equivalent to $\mathcal{O}[1]$). When $A \approx 0$ we shall regard A as vanishing when we linearise (for it is zero to the accuracy of relevant first-order calculations). Then,

- Zero-order variables are μ, p, Θ, and their time derivatives, $\dot{\mu}$, $\dot{p}$, $\dot{\Theta}$,
- First-order variables are $\dot{u}^a$, σ_{ab}, ω^a, q^a, π_{ab}, E_{ab}, H_{ab}, X_a, Z_a, and their time and space derivatives.

As these first-order variables all vanish in exact FLRW universes, provided u^a is uniquely defined in the realistic (lumpy) almost-FLRW universe model, they are all uniquely defined GIC variables. Thus this set of variables provides what we wanted: $1+3$ covariant variables characterising departures from a FLRW geometry (and, in particular, the spatial inhomogeneity of a universe) that are gauge-invariant when the universe is almost-FLRW. Because they are tensors defined in the real space-time, we can evaluate them in any coordinate system we like in that space-time.

7.1.1. *Key variables*

Two simple gauge-invariant quantities give us the information we need to discuss the time evolution of density fluctuations. The basic quantities we start with are the orthogonal projections of the energy density gradient, i.e. the vector $X_a \equiv \tilde{\nabla}_a \mu$, and of the expansion gradient, i.e. the vector $Z_a \equiv \tilde{\nabla}_a \Theta$. The first can be determined (a) from virial theorem estimates and large-scale structure observations (as, e.g., in the POTENT programme), (b) by observing gradients in the numbers of observed sources and estimating the mass-to-light ratio (Kristian and Sachs [98], Eq. (39)), and (c) by gravitational lensing observations. However, these do not directly correspond to the quantities usually calculated; but two closely related quantities do. The first is the *comoving fractional density gradient*:

$$\mathcal{D}_a \equiv S \frac{X_a}{\mu} , \tag{185}$$

which is gauge-invariant and dimensionless, and represents the spatial density variation over a fixed comoving scale. Note that S, and so $\mathcal{D}_a$, is defined only up to a constant by equation (25); this allows it to represent the density variation between any neighbouring worldlines. The vector $\mathcal{D}_a$ can be separated into a magnitude $\mathcal{D}$ and direction e_a

$$\mathcal{D}_a = \mathcal{D}\, e_a \ , \ e_a e^a = 1 \ , \ e_a u^a = 0 \quad \Rightarrow \quad \mathcal{D} = (\mathcal{D}_a \mathcal{D}^a)^{1/2} \ . \tag{186}$$

The magnitude $\mathcal{D}$ is the gauge-invariant variable[27] that most closely corresponds to the intention of the usual $\delta = \delta\mu/\mu$ given in (184). The crucial difference from the usual definition is that $\mathcal{D}$ represents a (real) spatial fluctuation, rather than a (fictitious) time fluctuation, and does so in a GIC manner. An important auxiliary variable in what follows is the *comoving spatial expansion gradient:*

$$\mathcal{Z}_a \equiv S\, Z_a \ . \tag{187}$$

The issue now is, can we find a set of equations determining how these variables evolve? Yes we can; they follow from the $1+3$ covariant exact equations of the previous sections.

7.2. DYNAMICAL EQUATIONS

We can determine exact propagation equations along the fluid flow lines for the quantities defined in the previous section, and then linearise these to the almost-FLRW case. The basic linearised equations are given by Hawking [22] (see his equations (13) to (19)); we add to them the linearised propagation equations for the gauge-invariant spatial gradients defined above [197].

7.2.1. *Growth of inhomogeneity*

Taking the spatial gradient of the equation of energy conservation (37) (for the case of a perfect fluid), we find [197]

$$\tilde{\nabla}_a(\dot{\mu}) + \Theta\, \tilde{\nabla}_a(\mu + p) + (\mu + p)\, \tilde{\nabla}_a\Theta = 0 \ ,$$

i.e.

$$h^b{}_a \nabla_b(u^c\nabla_c\mu) + \Theta\,(X_a + \tilde{\nabla}_a p) + (\mu + p)\, Z_a = 0 \ .$$

Using Leibniz' Rule and changing the order of integration in the second-derivative term (and noting that the pressure-gradient term cancels on using the momentum conservation equation (38)), we obtain the *fundamental equation for the growth of inhomogeneity*:

$$\dot{X}^{\langle a\rangle} + \tfrac{4}{3}\,\Theta\, X^a \;=\; -\,(\mu + p)\, Z^a - \sigma^a{}_b\, X^b + \eta^{abc}\,\omega_b\, X_c \ , \tag{188}$$

with source term Z^a. On taking the gradient of the Raychaudhuri equation (29) we find the companion equation for that source term:

$$\begin{aligned} \dot{Z}^{\langle a\rangle} + \Theta\, Z^a \;=\; & -\tfrac{1}{2}\, X^a - \sigma^a{}_b\, Z^b + \eta^{abc}\,\omega_b\, Z_c + \dot{u}^a\, \mathcal{R} - 2\, \tilde{\nabla}^a(\sigma^2 - \omega^2) \\ & + \tilde{\nabla}^a(\tilde{\nabla}_b \dot{u}^b + \dot{u}_b \dot{u}^b) \ , \end{aligned} \tag{189}$$

---

[27] Or, equivalently in the linear case, the spatial divergence of $\mathcal{D}_a$.

26.04.2020

Pag. 71(91)

7.1 Gauge Problem

Pag. 72(92)

Gauge Invariance Lemma

Pag. 73(93) ✡

[The proof is straight forward

Pág. 185(205) —

4 Gravitational lenses in Cosmology

4.3 The linearized equation of geometric optics

Pág. 189(209)

F) The renaissance of general relativity and cosmology...pdf

Pag. 287(297)

Topology and topology change in general relativity.

Pág. 288(298)

2.1 The Einstein-Rosen Bridge

Pág. 289(299)

and constraints on the content of our halo with low massive compact objects have been put.

4. Gravitational lenses in Cosmology

The extension of the lens equations to a cosmological context raises some technical difficulties because the background in which the objects are embedded is not flat. The aim of this section is to clarify these points. However, readers that are not familiar with cosmology can jump to section 5.

The basic equations, that describe jointly the evolution of the expansion parameter and the mean density, are the following,

$$\begin{aligned} 3\,\ddot{a} &= -4\pi G\,\rho\,a + \Lambda\,a; && (15)\\ \dot{a}^2 &= \frac{1}{a}\left[\frac{8\pi\,G}{3}\,\rho\,a^3 - k\,a + \frac{\Lambda\,a^3}{3}\right]. && (16)\end{aligned}$$

where Λ is a possible cosmological constant and k a possible curvature term. The Hubble constant reads,

$$H = \frac{\dot{a}}{a} = \sqrt{\frac{8\pi\,G}{3}\,\rho - \frac{k}{a^2} + \frac{\Lambda\,a^3}{3}}. \tag{17}$$

To simplify the discussions the reduced quantities are introduced,

$$\Omega = \frac{8\pi\,G}{3\,H^2}\,\rho \quad\text{and}\quad \lambda = \frac{\Lambda}{3\,H^2}. \tag{18}$$

They have an index 0 when they are taken at present time.

4.1. THE ANGULAR DISTANCES

We consider an object of size l (either because of its proper size or because of a peculiar physical process) at redshift z. When this size l is seen under an angle α, then by definition,

$$\alpha = \frac{l}{\mathcal{D}_0}, \tag{19}$$

where $\mathcal{D}_0$ is the angular distance. This is the distance at which this object would be in an Euclidean metric. What is then the relationship between $\mathcal{D}_0$ and z? The Friedmann-Robertson-Walker metric is given by,

$$\mathrm{d}s^2 = -c^2\,\mathrm{d}t^2 + a^2(t)\left(\frac{\mathrm{d}x^2}{1-k\,x^2} + x^2\mathrm{d}\theta^2 + x^2\,\sin^2\theta\,\mathrm{d}\varphi^2\right). \tag{20}$$

The fact that the size of this object is l means that it takes a time interval $\mathrm{d}t = l/c$ for light to travel to one end to the other. The corresponding angle $\alpha = \mathrm{d}\theta$ can be obtained by writing $\mathrm{d}s = 0$ with $\mathrm{d}x = 0, \mathrm{d}\varphi = 0$, which gives

$$l = a(t)\, x\, \alpha. \tag{21}$$

The angular *comoving* distance is thus given by $\mathcal{D}_0(z) = x(z)$. The expression of $x(z)$ can then be computed by the relation $\mathrm{d}s = 0$ along the line of sight with $\mathrm{d}\theta = 0$ and $\mathrm{d}\varphi = 0$,

$$\int_{t_0}^{t_1} \frac{c\,\mathrm{d}t}{a} = \int_0^{x(z)} \frac{\mathrm{d}x}{\sqrt{1-kx^2}} \equiv \chi(z). \tag{22}$$

For an open Universe, $k < 0$, and we have,

$$x(z) = \mathcal{D}_0(z) = \frac{1}{\sqrt{-k}} \sinh[\sqrt{-k}\ \chi(z)]. \tag{23}$$

Obviously when $k = 0$, $x = \chi$.

The relation $\chi(z)$ depends on the function $a(t)$ for a given cosmology, and therefore on the matter content of the Universe, on Λ and k . For instance, for an Einstein-de Sitter Universe, in the matter dominated era, we have

$$a(t) = \left(\frac{t}{t_0}\right)^{2/3}, \tag{24}$$

so that,

$$x(z) = \int_{t_0}^{t_1} t^{-2/3}\mathrm{d}t = 3(t_0^{1/3} - t_1^{1/3})t_0^{2/3}, \quad H_0 = \frac{2}{3\,t_0}, \tag{25}$$

and eventually,

$$\mathcal{D}_0(z) = \chi(z) = \frac{c}{H_0}\left(2 - \frac{2}{\sqrt{1+z}}\right). \tag{26}$$

This is the comoving angular distance for an Einstein-de Sitter Universe. More generally we have an explicit solution if $\Lambda = 0$ only. Finally the lens equation also requires the angular distance between two different redshifts z_1 and z_2. The calculation is actually quite simple. This distance is given formally by $\mathcal{D}_0(z_2)$ when it is calculated at a time when the observer is at redshift z_1. k being time independent we should formally have,

$$\mathcal{D}_0(z_1, z_2) = \frac{1}{\sqrt{-k}} \sinh[\sqrt{-k}\ (\chi(z_2, z_1))], \tag{27}$$

and to compute $\chi(z_2, z_1)$ one only needs to remark,

$$\chi(z_2, z_1) \equiv \int_{t_1}^{t_2} \frac{c\,\mathrm{d}t}{a} = \chi(z_2) - \chi(z_1), \tag{28}$$

which gives the expression of the angular distance we need. Eventually, it is fruitful to notice that,

$$\frac{c\,\mathrm{d}t}{a} = -\frac{c\,\mathrm{d}z}{H}, \tag{29}$$

which gives,

$$\mathcal{D}_0(z_1, z_2) = \frac{c}{H_0\sqrt{1-\Omega_0-\lambda_0}} \sinh\left[H_0\sqrt{1-\Omega_0-\lambda_0}\ \int_{z_1}^{z_2} \frac{\mathrm{d}z}{H(z)}\right], \tag{30}$$

and

$$\mathcal{D}_0(z) = \frac{c}{H_0\sqrt{1-\Omega_0-\lambda_0}} \sinh\left[H_0\sqrt{1-\Omega_0-\lambda_0}\ \int_{0}^{z} \frac{\mathrm{d}z'}{H(z')}\right]. \tag{31}$$

The whole geometrical part of the lens equation is thus established.

4.2. GEOMETRIC OPTICS IN A WEAKLY INHOMOGENEOUS UNIVERSE

What is now the source term for the deflection angle? We should first notice that in absence of lenses the light rays follow the geodesics of the Friedmann-Robertson-Walker metric. And in the applications we are interested in, the metric fluctuations are always weak. These fluctuations are given by, $G\,M/(R\,c^2)$. For instance, for

- 1 star: $M = 1\,M_\odot$, $R = 7\,10^5$km, $\delta\phi \approx 10^{-6}$;
- 1 galaxy cluster: $M = 10^{15}\,M_\odot$, $R = 1$Mpc $= 3\,10^{19}$km, $\delta\phi \approx 10^{-5}$.

The metric inhomogeneities are thus always extremely weak, even in the most extreme cosmological situations.

Following[3] Sachs (1961), we consider two nearby geodesics, $\mathcal{L}$ and $\mathcal{L}'$, in a light bundle in an FRW Universe with small metric fluctuations. We denote α_i the bi-dimensional angular distance between $\mathcal{L}$ and $\mathcal{L}'$ as it is seen by the observer. This is the distance in the image plane, that is the difference between the angular coordinates with which the photons arrive. We denote $\xi_i(z)$ the real distance between $\mathcal{L}$ and $\mathcal{L}'$ at redshift z (see Fig. 3). It implies that the geodesics are straight enough so that light always travels towards the observer. We also assume that the deflections are small enough so that it is possible to make the small angle approximation,

$$\xi_i(z) = a\ \mathcal{D}_{ij}(z)\ \alpha_j, \tag{32}$$

---

[3] See Misner, Thorne and Wheeler for an exhaustive presentation of the geometric optics.

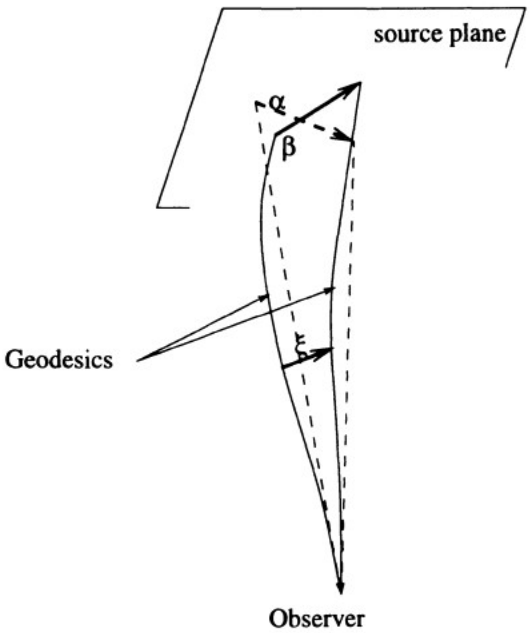

Figure 3. Sketch showing the geometrical quantities that appear in Eq. 32

that is that we assume the the position vector ξ_i can be obtained by a simple linear transform of the angular coordinates. For an homogeneous space $\mathcal{D}_{ij}(z)$ is simply given by $\mathcal{D}_0(z)\,\delta^K_{ij}$ where δ^K_{ij} is the Kronecker symbol. Obviously $\mathcal{D}_{ij}$ changes as a function of redshift along the trajectories. The "virtual" angular position in the source plane is then given by the ratio of the real distance (at time of light emission for instance) by the angular distance of the emitter in an homogeneous space,

$$\vec{\beta} = \frac{\vec{\xi}(z)}{a\mathcal{D}_0(z)}. \tag{33}$$

The amplification matrix, or rather its inverse, $\mathcal{A}^{-1}$, is then given by,

$$\mathcal{A}^{-1}(z) = \frac{\mathcal{D}_{ij}(z)}{a\mathcal{D}_0(z)}. \tag{34}$$

for a source plane at redshift z.

Sachs (1961) gave the master equation which governs the evolution of the distance between the geodesics. The derivation of this equation goes beyond these lecture notes and I give only the final answer (as it has been given by Seitz, Schneider and Ehlers 1994),

$$\frac{\mathrm{d}^2[a\ \mathcal{D}_{ij}(\vec{\beta}, z)]}{\mathrm{d}\eta^2} = a(z)\mathcal{R}_{ik}(\vec{\beta}, z)\ \mathcal{D}_{kj}(\vec{\beta}, z) \tag{35}$$

where the derivatives are taken with respect to η,

$$\mathrm{d}\eta = -\frac{\mathrm{d}a}{H(a)} = -a\ \mathrm{d}t, \quad \eta(z=0) = 0. \tag{36}$$

with the boundary conditions,

$$(\mathcal{D}_{ij})_{z=0} = 0; \quad \left(\frac{\mathrm{d}\mathcal{D}_{ij}}{\mathrm{d}\eta}\right)_{z=0} = \frac{c}{H_0}. \tag{37}$$

The matrix $\mathcal{R}_{ij}$ represents the tidal effects. It can by written in terms of the gravitational potential ϕ given by,

$$\Delta_x \phi = 4\pi\, G\, \overline{\rho}\, a^2\, \delta_{\mathrm{mass}}. \tag{38}$$

The Laplacian is taken with respect of the comoving angular distances. We have

$$\mathcal{R}_{ij} = -\frac{4\pi\, G\, \overline{\rho}}{H_0^2\, a^2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \frac{2}{H_0^2\, a^2} \begin{pmatrix} \phi_{,11} & \phi_{,12} \\ \phi_{,21} & \phi_{,22} \end{pmatrix}. \tag{39}$$

Since $8\pi\, G\, \overline{\rho}\, a^3 = 3H_0^2\, \Omega_0$, (e.g. Eq. 18) for an homogeneous Universe we have,

$$\mathcal{R}_{ij}^{(0)} = -\frac{3}{2}\,(1+z)^5\, \Omega_0\, \delta_{ij}^K \tag{40}$$

(the superscript (0) means here that it is the value of $\mathcal{R}$ without perturbations). In this case the matrix $\mathcal{D}_{ij}$ is proportional to δ_{ij}^K and we have,

$$\frac{\mathrm{d}^2[a\ \mathcal{D}_0(z)]}{\mathrm{d}\eta^2} = -\frac{3}{2}(1+z)^4\, \Omega_0\, \mathcal{D}_0(z). \tag{41}$$

We recover in fact the comoving angular distance the expression of which we know,

$$\mathcal{D}_0(z) = \frac{c}{H_0\sqrt{1-\Omega_0-\lambda_0}} \sinh\left[\sqrt{1-\Omega_0-\lambda_0} \int_0^z \frac{\mathrm{d}z'}{E(z')}\right], \tag{42}$$

with

$$E(z) = \frac{H(z)}{H_0} = \sqrt{\lambda_0 + (1+z)^2(1-\Omega_0-\lambda_0) + (1+z)^3\Omega_0}. \tag{43}$$

This integral has a closed form only when the cosmological constant, Λ, is zero.

4.3. THE LINEARIZED EQUATION OF GEOMETRIC OPTICS

We can remark that the equation (35) is not linear since $\mathcal{D}$ is not simply proportional to $\mathcal{R}$. This expresses the fact that the deformation of the angular distance is made all along the light trajectory by multiple deflections.

The general resolution of Eq. (35) is in general very complicated. It can however handled when it is linearized. Let's assume we can expand $\mathcal{D}_{ij}$ with respect of the local density contrast,

$$\mathcal{D}_{ij}(z) = \mathcal{D}_0 + \mathcal{D}_{ij}^{(1)} + \dots \tag{44}$$

It implies that, at first order,

$$\frac{\mathrm{d}^2[a\ \mathcal{D}_{ij}^{(1)}(\vec{\beta}, z)]}{\mathrm{d}\eta^2} - a(z)\mathcal{R}_{ik}^{(0)}(\vec{\beta}, z)\mathcal{D}_{kj}^{(1)}(\vec{\beta}, z) = -3\Omega_0(1+z)^4\mathcal{D}_0(z)\varphi_{,ij}(\vec{\beta}, z) \tag{45}$$

with,

$$\left(\mathcal{D}_{ij}^{(1)}\right)_{z=0} = 0, \quad \left(\frac{\mathrm{d}\mathcal{D}_{ij}^{(1)}}{\mathrm{d}\eta}\right)_{z=0} = 0, \tag{46}$$

and we define the field φ so that,

$$\Delta_x\varphi = \delta_{\mathrm{mass}}(\vec{\beta}, z) = \frac{\Delta_x\phi}{4\pi G\,\overline{\rho}\,a^2}. \tag{47}$$

To solve this differential equation it is easier to write it with the variable z. It then reads,

$$\begin{aligned} &\frac{\mathrm{d}^2[a\ \mathcal{D}_{ij}^{(1)}(\vec{\beta}, z)]}{\mathrm{d}z^2} + \frac{1}{E(z)}\frac{\mathrm{d}E(z)}{\mathrm{d}z}\frac{\mathrm{d}\mathcal{D}_{ij}^{(1)}(\vec{\beta}, z)}{\mathrm{d}z} - \\ &-\frac{1}{1+z}\frac{1}{E(z)}\frac{\mathrm{d}E(z)}{\mathrm{d}z}\mathcal{D}_{ij}^{(1)}(\vec{\beta}, z) + \frac{3}{2}\frac{\Omega_0(1+z)}{E^2(z)}\mathcal{D}_{ij}^{(1)}(\vec{\beta}, z) = \\ &\qquad -3\frac{\Omega_0(1+z)}{E^2(z)}\mathcal{D}_0(\vec{\beta}, z)\varphi_{,ij}(\vec{\beta}, z). \end{aligned} \tag{48}$$

The differential homogeneous equation it is associated with has two known solutions. One describes the angular distance $\mathcal{D}_0$, the other is given by,

$$\mathcal{U}_0(z) = \frac{1}{\sqrt{1-\Omega_0-\lambda_0}}\cosh\left[\sqrt{1-\Omega_0-\lambda_0}\int_0^z \frac{\mathrm{d}z'}{E(z')}\right], \tag{49}$$

The general solution of 48 then reads,

$$\mathcal{D}_{ij}^{(1)}(\vec{\beta}, z) = -3\Omega_0\int_0^z \mathrm{d}z'\frac{(1+z')\mathcal{D}_0(z')\varphi_{,ij}(z')}{E^2(z)}\frac{\mathcal{U}_0(z)\mathcal{D}_0(z') - \mathcal{U}_0(z')\mathcal{D}_0(z)}{\mathcal{U}_0'(z')\mathcal{D}_0(z') - \mathcal{U}_0(z')\mathcal{D}_0'(z')}, \tag{50}$$

which, after elementary mathematical transforms, gives,

$$\mathcal{D}_{ij}^{(1)}(\vec{\beta}, z) = -3\,\Omega_0 \int_0^z \frac{\mathrm{d}z'}{E(z)} \frac{1}{\sqrt{1-\Omega_0-\lambda_0}} \times$$

$$\times \sinh\left[\sqrt{1-\Omega_0-\lambda_0}\int_z^{z'} \frac{\mathrm{d}z''}{E(z'')}\right](1+z')\mathcal{D}_0(z')\varphi_{,ij}(z'). \qquad (51)$$

It can be rewritten by introducing the physical distance χ along the line of sight. We eventually have,

$$\mathcal{A}^{-1}(z) = \mathrm{Id} - \frac{3\,\Omega_0}{(c/H_0)^2}\int_0^{\chi(z)} \mathrm{d}\chi' \frac{\mathcal{D}_0(z', z)\mathcal{D}_0(z')}{\mathcal{D}_0(z)}(1+z')\varphi_{,ij}(z'), \qquad (52)$$

where the angular distances $\mathcal{D}_0(z)$ and $\mathcal{D}_0(z, z')$ are comoving. This equation actually gives the expression of the amplification matrix for a non-trivial background. We find that the amplification matrix is given by the superposition of lens effects of the different mass layers. We can remark that the lens term is given by the gravitational potential, ϕ, that is by the potential the source term of which is given by the density contrast.

Note finally that this equation is valid in two limit cases, either for a single lens plane with an arbitrary strength or the superposition of any number of weak lenses. This equation naturally extents the previous result, (10), obtained for a single lens in an Euclidean background. The higher orders of Eq. (35) give the intrinsic lens coupling effects (i.e. their nonlinear parts). We will not consider them here.

5. Galaxy clusters as gravitational lenses

The study of galaxy clusters has become a very active field since the discovery of the first gravitational arc by Soucail et al. (1988) in Abell cluster A370. Galaxy clusters give the most dramatic example of gravitational lens effects in a cosmological context. The difficulty is however to describe the shape of their mass distribution.

5.1. THE ISOTHERMAL PROFILE

For an isothermal profile we assume that the local density $\rho(r)$ behaves like,

$$\rho(r) = \rho_0 \left(\frac{r}{r_0}\right)^{-2}. \qquad (53)$$

With such a density profile the total mass is not finite. So this is not a realistic description but it is a good starting point for the central part of

'()*' ' '+ ' . 1 2
! 2 / 1 2 4 < 4 ! 2 B 66#3 4 !
5 1 2 -!
& ;8 68 . - 2 ! . - 2 0) 1 2

Topology and Topology Change in General Relativity

GARY W. GIBBONS

1 INTRODUCTION

The organisers have asked us to review the progress of some aspect of general relativity and cosmology in which they have a particular interest and to introduce their remarks by describing its relation to their interaction with Dennis Sciama. It is a great pleasure for me to do so and also to pay tribute to the inspiration that he, and his style of doing physics, has been to me over the years. In particular I have tried to follow his example by asking simple physical questions and trying to answer them with the simplest appropriate tools available. For that reason in what follows I shall not give extensive mathematical details but refer the reader to the references. Moreover because of the personal nature of the review I have made no attempt to include in those references every paper on the subject, especially where the story is widely known and can be read up in standard textbooks. For the same reason I have perhaps erred in including too many papers of my own.

I became a Research student of Dennis Sciama in October 1969 after being enthralled by his marvelously lucid and exciting Part III lectures on General Relativity. When Dennis left for Oxford a year later I transferred to Stephen Hawking, himself a former student of Dennis. I can thus be said to be the answer to a riddle in that I am the academic son of my father's son. The problem that Dennis originally suggested to me was to give an upper bound to the proportion of the rest mass of a gravitating system's rest which could be lost in gravitational radiation - a question which was at that time very topical because of Weber's claims to have observed substantial amounts of gravitational radiation coming from the galactic centre.

I failed to make much headway on that particular problem, unlike my next supervisor who spectacularly solved one aspect of it - that involving black holes, about which I knew almost literally nothing at that time - with his celebrated Area Theorem. However, as all good Ph.D. problems should, it did set me thinking about numerous things including my subject today, topology and topology change in general relativity, and I would like to describe some of my thoughts over the years on this subject, at the same time trying to put them in a more general context.

I shall begin by considering the classical theory and then pass on to the necessarily more speculative quantum theory. I will restrict myself to the case of four spacetime dimensions, and I will concentrate on two particular theories: pure general relativity (with no matter) and Einstein-Maxwell theory. Much of what I have to say can be readily extended to more complicated theories, as I shall indicate briefly at the end, but these two simple theories already exhibit in essence the basic ideas and at the same time indicate the great richness of possible physical effects one can expect.

2 THE CLASSICAL THEORY

2.1 The Einstein-Rosen Bridge

The most direct way of answering the question Dennis had posed seemed at first to be to examine the motion of 'point particles' à la E.I.H. I read many papers on this subject and understood very little of them. However the one thing that did strike me was that the point particle picture seemed rather inappropriate in general relativity because it involved extrapolating the gravitational field of the particle inside its Schwarzschild radius. I found far more appealing the view, espoused in the famous paper of Einstein and Rosen (1935) in which they first described the "Einstein-Rosen Bridge" in the Schwarzschild solution, that one should look for everywhere non-singular solutions of the non-linear field equations representing a particle. Nowadays we call such solutions "solitons" and their use has become commonplace: the topological defects such as magnetic monopoles we now consider in cosmology are of course just solutions of this sort. Einstein and Rosen were greatly impressed by the non-trivial (spatial) topology of their throat, which seemed to be responsible for the existence of the solution and absence of singularities. Sadly however the singularities were still there lurking, as it were, in the full Kruskal spacetime rather than in space at one time. Indeed this conflict between the Hamiltonian and Lagrangian viewpoints, i.e. between space and spacetime is characteristic of our present subject, as it is of a great deal else in gravity.

2.2 Geometrodynamics

Einstein and Rosen's idea was taken up by Wheeler and his group at Princeton and their main results were summarised in that marvelously inspiring book 'Geometrodynamics' (1961). From it I came away with three ideas which are important for our story.

1) That there exist very many non-singular initial data sets which are topologically non-trivial both for pure general relativity and for Einstein-Maxwell theory.
2) In the latter case the non-trivial topology can be responsible for the phenomenon

of "Charge without Charge".

3) The complete duality symmetry between electric and magnetic fields, a symmetry which can even be extended to massless spinor fields using chiral rotations involving γ_5.

2.3 Causal Selection Rules

The existence of topologically non-trivial initial data sets of this sort naturally raises questions about their evolution in time. Can there be transitions for example between spaces of different spatial topologies? One may ask such questions at the most basic kinematical level: what kind of evolutions are possible?, and at the dynamical level: what kind of evolutions will arise if field equations are imposed?

Perhaps the best known results at the purely kinematical level are those of Geroch (1967) which despite their great elegance and beauty have had the somewhat unfortunate effect until very recently of discouraging work on topology change. Let us assume the existence of an everywhere non-singular time-orientable and space-orientable Lorentzian metric g_L on a compact 4-manifold M with a closed but not necessarily connected spacelike boundary $\partial M = \Sigma$. Reinhart and later Geroch had shown that there exist many such spacetimes $\{M, g_L\}$ for every boundary Σ but Geroch also showed that if Σ consists of two components, Σ_i and Σ_f in the past and in the future respectively, and if Σ_i and Σ_f do not have the same topology then the spacetime M must contain closed timelike curves. If on the other hand M does not have closed timelike curves it is topologically the product of a closed time interval and some 3-manifold of fixed topology. Because people have, until recently, been loathe to consider closed timelike curves as physically possible, Geroch's theorem has usually been interpreted as a selection rule forbidding topology change at the classical level. But like most No Go Theorems in physics, its basic assumptions have come to be questioned, and a number of physicists have begun to take seriously closed timelike curves and time machines and even to ask whether the laws of physics forbid in principle the construction of wormholes in the laboratory and their subsequent use as time machines.

2.4 Spinorial Selection Rules

In my opinion a potentially much worse pathology for a spacetime is for it not to admit an $SL(2, C)$ spinor structure, since it is difficult according to our present understanding of matter to imagine a world without fermions described by fundamental spinor fields. If spacetime M is a product of a closed time interval and a compact 3-manifold it must necessarily admit an $SL(2, C)$ spinor structure but if it is not a product it need not. Bichteler (1968) and Geroch (1968,1970) pointed out that the necessary and sufficient condition for the existence of a spinor structure in this case is the vanishing of the second Stiefel Whitney class of the 4-manifold M. Stephen Hawking and I (1992a, 1992b) have recently obtained an interesting extension of Geroch's results. We find that the Lorentzian spacetime M will admit an $SL(2, C)$ spinor structure if and only if a certain mod 2 topological invariant $u(\Sigma)$ of the spacelike boundary vanishes. Since the invariant is additive under disjoint union:

$$u(\Sigma_1 \cup \Sigma_2) = u(\Sigma_1) + u(\Sigma_2) \mod 2$$

one may interpret our result as giving a topological conservation law for 3-manifolds. The invariant takes the value one for S^3 and zero for the wormhole, $S^1 \times S^2$. It follows that it is not possible even in principle to construct a wormhole in the laboratory in order, for example, to build a time machine. However under connected sum our invariant satisfies:

$$u(\Sigma_1 \# \Sigma_2) = u(\Sigma_1) + u(\Sigma_2) + 1 \mod 2.$$

Thus it takes the value zero for the connected sum of two wormholes. Our selection rule therefore allows the construction of a pair of wormholes in the laboratory. On the other hand it does not allow the birth of a single S^3 universe from nothing or a trouser-leg spacetime in which one S^3 universe splits into two, even though there are perfectly good Lorentzian spacetimes of this sort without spinor structure . An example would be CP^2 with three points removed.

Our invariant may be calculated in terms of the mod 2 homology groups of the boundary manifold Σ. It is called by mathematicians the mod 2 Kervaire semi-characteristic. The formula is:

$$u(\Sigma) = \dim_{Z_2}(H_0(\Sigma; Z_2) \oplus H_1(\Sigma; Z_2)) \mod 2.$$

The first term counts the connected components of Σ and the second term counts, roughly speaking, the number of wormholes or handles.

2.5 Singularities

So far we have been ignoring the field equations. If we take them into account we soon encounter more difficulties. The topologically non-trivial initial data sets we have been considering will in general possess apparent horizons and closed trapped surfaces. In the time-symmetric case for example, it is easy to see (Gibbons 1972) that the minimal 2-spheres (they must be topologically spheres if they are truly minimal by a simple application of the formula for the second variation of the area and the Gauss-Bonnet theorem) are also apparent horizons. Thus one expects, by the Singularity Theorems of Penrose and Hawking, that the future evolution will contain spacetime singularities. If our usual beliefs about Cosmic Censorship and Black Hole

Physics are true these singularities should be hidden inside event horizons outside of which the metric should settle down to one of the stationary solutions allowed by the Black Hole Uniqueness Theorems. If this is true, and I have no reason to doubt it, it means that dynamically, apart from providing the apparent horizons, the topology has no real effect on the dynamics. After all the *same* black holes could have been formed from a topologically trivial initial data set just consisting of gravitational waves or from some topologically extremely exotic initial data set, even a non-orientable initial 3-manifold or one allowing "spin without spin". According to our present views all details of any complicated non-trivial topology should in the classical theory of general relativity ultimately be lost in the singularity, and certainly hidden from our view by the event horizon. All of this is in marked contrast to conventional flat space theories admitting solitons where the influence of topology can play a decisive role in the dynamics.

At the classical level therefore it seems that we can conclude at least that:

1) It is difficult to incorporate topology change into the classical theory even at the purely *kinematic* level.
2) If the Einstein Equations hold then dynamics is no respecter of topology.

3 THE QUANTUM THEORY

3.1 Hawking Evaporation and Central Charges

These preliminary conclusions would seem to be reinforced when we turn to the (semi-classical) quantum theory. The classically stable Black Hole final states are no longer stable in the quantum theory but rather undergo Hawking evaporation, ultimately shrinking down to a size below which we cannot reasonably calculate because of our ignorance of the correct theory of quantum gravity. If it is really true that neutral black holes evaporate completely leaving no remnant then it seems reasonable to conclude that topological invariants, unless associated with some conserved charge, are simply not "good quantum numbers" . If the black holes are electrically or magnetically charged however the situation may be different and this is where Wheeler's idea of charge without charge comes in. If a black hole carries a charge which is not carried by the fundamental fields of the theory it can't be radiated away by Hawking evaporation, and it may thus stabilise the hole. In flat space theories with solitons charges of this kind, usually topological in origin, are often called "central" because in the quantum theory they commute with the other observables of the theory.

3.2 Supersymmetry

In supersymmetric theories with N spinorial supercharges Q_i, $i = 1,2,..N$, they may be central charges in the technical sense that central charges appear in the supersymmetry algebra:

$$\{\overline{Q}_i, Q_j\} = \gamma_\mu P^\mu \delta_{ij} + Z_{ij}$$

where $Z_{ij} = -Z_{ji}$ are central charges, and P^μ is the total energy momentum vector of the system. In the simplest case, $N = 2$ when $Z_{12} = -Z_{21} = Z$, it follows that the mass M and central charge Z satisfy the Bogomolnyi bound:

$$M \geq |Z|$$

with equality if and only if the soliton is partially supersymmetric in the sense that it is invariant under half the maximum number of supersymmetries.

The obvious model theory to which to apply these ideas is $N = 2$ supergravity theory, that is the supersymmetric extension of Einstein Maxwell theory.The fundamental degrees of freedom make up a supermultiplet consisting of the graviton, the photon and two (electrically neutral) Majorana gravitini. These three particles carry neither electric nor magnetic charge so either type of "charge without charge" is a candidate as a "central charge" (Gibbons, 1981). Moreover the $N = 2$ supergravity theory shares the duality invariance of its purely bosonic sector provided one extends the symmetry to the spin $\frac{3}{2}$ gravitini by means of precisely the type of chiral rotation to be found in "Geometrodynamics" applied to the spin $\frac{1}{2}$ neutrino field. Gratifyingly the whole story works out rather well. Witten's proof of the Positive Energy Theorem may be extended to the Einstein-Maxwell theory (Gibbons and Hull 1982, Gibbons, Hawking, Horowitz and Perry 1983) to bound from below the mass M of any initial data set regular outside an apparent horizon and carrying total electric charge Q and magnetic charge P consistent with the Cosmic Censorship Hypothesis:

$$M \geq \frac{\sqrt{(Q^2 + P^2)}}{\sqrt{(4\pi G)}}.$$

If one has equality, then any regular solution must be one of the static multi-black hole solutions of Papapetrou and Majumdar admitting 4 out of a possible maximum of 8 Majorana Killing spinors, that is (considered as a Supergravity solution) it has half the maximum number of supersymmetries. It is illuminating to view this from the point of view of the supersymmetric ground states or vacua of the $N = 2$ supergravity theory (Gibbons 1984)

A minimal requirement of a vacuum is that it be spacetime homogeneous, and to be fully supersymmetric it should admit the maximum number of supersymmetries, i.e. 8 Majorana Killing spinors. For the $N = 2$ theory there are two such ground states:

1) Flat spacetime with vanishing electromagnetic field.

2) The Robinson–Bertotti solution, which is homogeneous but anisotropic and whose electromagnetic field is homogeneous in space and time.

The Robinson-Bertotti solution may be thought of as a compactified state of the theory, since it is a metric product of the 2-sphere and 2-dimensional anti-de-Sitter spacetime. Both of these ground states appear to be stable against small perturbations subject to suitable boundary conditions. Now the typical behaviour of a supersymmetric soliton is that it admits half the maximum number of supersymmetries and spatially interpolates between two distinct vacua or ground states of the theory. The extreme Reissner-Nordstrom solution does just that. It is asymptotically flat but in the extreme limit the Einstein-Rosen throat becomes infinitely long and the geometry near the horizon is well approximated by the Robinson-Bertotti solution. In common with other supersymmetric solitons the extreme black holes may be fitted into a so-called shortened supermultiplet (Gibbons 1981, Gibbons and Hull 1982, Gibbons 1982a). One also has solutions in which the asymptotically flat region is replaced by another asymptotically Robinson-Bertotti like region (Whitt 1985). Brill (1992) has recently interpreted these as mediating the splitting of black holes, but in view of the stability of Robinson-Bertotti ground state this interpretation may encounter difficulties, particularly in a supersymmetric theory. The creation of pairs of black holes from a Melvin background, which will be described later, seems to me physically more plausible.

3.3 Solitons and Moduli

As mentioned above, in addition to the single soliton solution, i.e. the extreme Reissner-Nordstrom solution, there are multi-soliton solutions. These describe an arbitrary number, k, say of extreme black holes placed at k arbitrary positions and having k arbitrary masses. The moduli space is given by k distinct points in R^3, possibly quotiented by the symmetric group on k objects if the masses are all equal and the black holes are regarded as being identical. The black holes are in equipoise because, at rest, their gravitational attraction is just balanced by their electromagnetic repulsion. If they are given a small amount of kinetic energy the subsequent motion may be described as a one–parameter family of static solutions or as a curve in the moduli space. The curve is in fact a geodesic with respect to a certain non-flat metric on the moduli space (Gibbons and Ruback 1986, Eardley and Ferrell 1987, 1989) which has been calculated fully in the case $k = 2$ and in the limit of large separations for general k. Although far from the case of astrophysical interest suggested originally by Dennis, and in an approximation in which gravitational radiation is neglected, I feel that it is note-worthy that the original programme of Einstein and Rosen can be carried sufficiently far that equations of motion for solitons can can actually be worked out in considerable detail. One may even consider the non-relativistic quantum mechanics of the solitons by studying a Schroedinger equation on the moduli space (Gibbons and Ruback 1986, Ferrell and Traschen 1992). The quantum mechanics of these solitons from the supergravity point of view has been treated in great detail in a series of papers by the Viennese group (Aichelburg and Embacher 1988).

3.4 Quantum Corrections

I want now to turn to the quantum properties of these supersymmetric solitons. The first and most obvious point to reiterate is that since they have zero surface gravity and hence zero Hawking temperature and because the graviton, photon and gravitini carry no charge, they are stable against thermal evaporation in. In fact a general non-extreme Reissner-Nordstrom black hole can only reduce its mass but not its charge by emitting Hawking radiation. If the ratio M^2/Z^2 exceeds 3/4 it will get hotter as it does so but as soon as this ratio falls below 3/4 the specific heat at constant charge becomes positive and thereafter the evolution slows down and the hole coasts towards the extreme state.

Another question concerns the quantum corrections to the geometry and whether or not the formula for the entropy for example, which may be calculated using a functional integral approach in which one evaluates the partition function for quantum gravity, is still given by one quarter of the area of the event horizon. Since the Papapetrou-Majumdar solutions belong to the general class of Israel-Wilson metrics which include the self-dual multi-centre metrics and all of these metrics admit Killing spinors, it always seemed to me quite likely that the quantum corrections should vanish. Indeed one motivation for studying this type of black hole, despite its lack of obvious astrophysical application, has always been that in supergravity theories the divergence difficulties which make it so difficult to make any kind of definitive prediction about the ultimate outcome of the thermal evaporation process in pure general relativity might to some extent be obviated. This optimism has been recently been reinforced by the work of Renata Kallosh (1992) who has shown, using superfield techniques, that indeed the quantum corrections do cancel in rather the same fashion that they do around self-dual backgrounds.

3.5 Black Hole Pair Creation

This still leaves open the question of why one should not simply ignore solitons of this sort. One might feel that since they appear to be absolutely stable, if the world contained none to begin with they would be irrelevant. The answer to this argument in conventional flat space theories is well known, and is that if it is possible to pair–produce soliton–anti–soliton pairs during high energy collisions or in extremely strong external fields, then it would be *inconsistent* to ignore their existence. In particular closed loops describing virtual soliton-anti-soliton creation and annihilation should in principle be included as a contribution to all processes. This is not necessarily true for neutral black holes which it seems do not correspond to stable states and should not therefore be included in a complete set of intermediate states. Moreover it appears that they cannot be created in pairs in pure general relativity with a vanishing cosmological term, although there are some indications that this is possible if there is a positive cosmological term (Gibbons 1992). Therefore one should ask whether one can pair–produce electrically or magnetically charged black holes using strong electromagnetic fields.

3.6 Real Tunneling Geometries

To answer this question and others about the possibility of topology change via quantum tunneling, I would like to recall the main features of an approach to such tunneling in the path integral approach to quantizing gravity. The basic idea is that one should be able to approximate the path integral by classical paths, i.e. solutions of the classical equations of motion, which have a Lorentzian portion in the future but which may in general be complex. A particularly simple class of complex paths which seem to include all the cases they have arisen "in practice" are those which Hartle and I (Gibbons and Hartle 1990) have called "Real Tunneling Geometries" .

A Real Tunneling Geometry may be regarded as a triple $\{M_R, \Sigma, M_L\}$ consisting of a complete Riemannian manifold M_R with a totally geodesic boundary $\Sigma = \partial M_R$ and a Lorentzian manifold M_L for which Σ is a Cauchy surface with vanishing second fundamental form. By glueing two copies of the Riemannian manifold M_R across the boundary Σ to obtain the double $2M_R$ one obtains a a particular kind of Gravitational Instanton, one admitting a reflection map θ stabilizing the surface Σ. The simplest and most frequently encountered case (at least in the pages of some physics journals if not in the real world) is when M_R is half of the 4-sphere with its standard round metric and M_L is half of de-Sitter spacetime to the future of a surface of time symmetry. This is the basis for many discussions for example of the "creation of the universe from nothing" considered as some sort of tunneling event, with Σ playing the role of the beginning of time. It seems reasonable to take the existence of a solution of the classical equations of motion which has the form of a Real Tunneling Geometry for which the classical action of the the Riemannian portion M_R is finite, as some sort of *a priori* evidence that a tunneling process can take place in which a spacetime or portion of spacetime with spatial topology given by Σ suddenly materializes. By the same token, the non-existence of a solution for a particular topology for Σ may be interpreted as quantum mechanical selection rule forbidding that particular topological transition at the lowest semi -classical level.

3.7 The Ernst-Melvin Instanton

The question that we must now ask is whether there exists a Real Tunneling Geometries involving pairs of black holes. My first thought (Gibbons 1986) was to turn to the C-metrics. These represent uniformly accelerating black holes, charged or neutral, and possess both black hole horizons and acceleration horizons. They have two hypersurface orthogonal Killing fields, one $\frac{\partial}{\partial\phi}$ with closed spacelike orbits corresponding to rotations about the direction of motion. The other $\frac{\partial}{\partial t}$ is timelike near infinity but spacelike behind the two types of horizons. Near infinity it behaves like a boost Killing vector rather than a timelike translation. In general the surface gravities of the two types of horizon are different. However in the special case that the mass M and central charge $Z = \sqrt{(Q^2 + P^2)}/\sqrt{(4\pi G)}$ are equal, the two surface gravities are also equal (and non-zero).

It is not possible to choose a range for the angular coordinate ϕ so as to obtain a metric which is everywhere non-singular outside the horizons. No matter what range one picks there always remain conical singularities which may be interpreted as a cosmic string or a cosmic strut depending upon whether the deficit angle is positive or negative. The strut runs from the acceleration horizon to the south pole of the black hole, where south means rear-most with respect to the direction of the acceleration in this context. The string runs from the north-pole of the black hole horizon to infinity. It is possible, by appropriate choice of the range of ϕ, to eliminate either the strut or the string but not both. This is not surprising physically, and may be seen immediately by noting that the solutions has vanishing ADM mass by virtue of the boost invariance at infinity. By the positive mass theorem, generalized to include horizons (Gibbons, Hawking, Horowitz and Perry 1983), there can be no solution which is non-singular everywhere outside the horizons. If there were, we might try to construct an Instanton or Real Tunneling Geometry by adjusting the surface gravities of the acceleration horizon and the black hole horizon to be equal. We would then have a solution which could be analytically continued to purely imaginary values of the time coordinate t. Its existence would have indicated that flat Minkowski spacetime is unstable to the creation of pairs of charged black holes.

Because of the conical singularities, pair creation of black holes from empty Minkowski spacetime is forbidden. It seems more reasonable to expect that they could be created from some strong external field, so we must turn to solutions generalizing the C-metrics representing charged black holes accelerated by an an external electromagnetic field. Such solutions have been constructed by Ernst. If no black hole is present they reduce to the everywhere non-singular static, cylindrically symmetric Melvin solution which is the metric due to a self-gravitating Maxwell flux-tube. Both the metric and the spacetime are also invariant under boosts in the direction of the flux-tube. In fact it may be considered as a rather degenerate limiting case of the solution representing a super-massive Nielsen-Olesen cosmic string in which the Higgs field has decoupled because its electric charge vanishes.

Ernst was able to show that by an appropriate choice of the strength of the appended electromagnetic field all conical singularities could be eliminated for any allowed ratio of charge to mass. This gives a Lorentzian solution of the Einstein-Maxwell equations, non-singular outside the black hole and acceleration horizons, representing a uniformly accelerating black hole. From the point of view of the original programme of Einstein and Rosen this is a remarkable achievement. More significantly for our present purposes it allows us, by choosing the case $M = \sqrt{(Q^2 + P^2)}/\sqrt{(4\pi G)}$ to obtain a Real Tunneling Geometry with a complete Riemannian portion M_R. The surface Σ has the topology of a Wheeler wormhole, i.e. $S^1 \times S^3$. The solution has two arbitrary constants, the charge or mass of the hole and the strength of the external electromagnetic field.

The amplitude for the creation of pairs depends on the classical action of the Riemannian portion of the complex path. In the limit when the black hole is very small compared with the scale set by the background electromagnetic field one obtains a tunneling pre-factor for the amplitude which agrees with what one would expect if one thought of black holes as a point particles and applied the well known formula due to Schwinger for the rate of their creation (Garfinkle and Strominger 1991).

I believe that the results I have described above clearly indicate that although topology appears not to play an important role in pure general relativity, its effects can be significant in theories like Einstein-Maxwell theory admitting central charges. Moreover the phenomenon of pair creation means that it is inconsistent to ignore it at the quantum level. I have discussed this process above using some of the ideas of what is often called Euclidean Quantum Gravity, but physical intuition strongly suggests that if the semi-classical picture makes any sense at all the this will continue to be true regardless of what formalism one quantizes with.

3.8 Dilatons and Strings

The picture may change radically however if one includes the effects of massless scalar fields such as the dilaton and the axion which arise in supergravity theories, Kaluza-Klein theories, and superstring theory. Currently there is a great deal of activity in this area. I will confine myself to a few remarks which will only hint at the richness and variety of the new considerations that come into play. Because of the coupling of the dilaton and axion to the various abelian gauge fields in these theories, the scalar no-hair theorems no longer hold in the literal sense that the dilaton and axion fields must be spacetime independent. There are still central charges, and if these are non-zero then a black hole must have an associated scalar or pseudoscalar field but its strength is entirely determined by the central charges. There are no independent degrees of freedom connected with the scalars (Gibbons 1982b). It turns out that the thermodynamic properties of this more general class of black holes can be very different from those of Einstein-Maxwell black holes. Moreover they are rather sensitive to the coupling constant governing the interaction between the dilaton and the abelian gauge fields. The value that arises in superstring theory is a critical case (Gibbons and Maeda 1988). This is particularly intriguing because one expects in any case to have to turn to string theory in order to make sense of the quantum theory. This indicates at the very least that its effects on the issue of topology change may well prove to be decisive but that is a topic for some future celebration looking back over the even more distinguished and influential career of Dennis Sciama.

REFERENCES

Aichelburg, P.C. & Embacher, F. (1988) *Physical Review* **D37** 338–348, 911-917, 1436-1443, 2132-2141

Bichteler, K. (1968) *Journal of Mathematical Physics* 9 813–815

Brill, D. (1992) Maryland preprint

Eardley, D.M. & Ferrell, R.C.(1987) *Physical Review Letters* 59 1617–1620

Eardley, D.M. & Ferrell, R.C. (1989) in *Frontiers in Numerical Relativity* eds. C.R. Evans, L.S. Finn and D.W. Habill (Cambridge University Press)

Einstein, A. & Rosen, N. (1935) *Physical Review* **48** 73

Ferrell, R.C. & Traschen, J. (1992) *Physical Review* **D45** 2628–2638

Garfinkle, D. & Strominger, A. (1991) *Physics Letters* **B256** 146–149

Geroch, R.P. (1967) *Journal of Mathematical Physics* 8 782–786

Geroch, R.P. (1968) *Journal of Mathematical Physics* 9 1739–1744

Geroch, R.P. (1970) *Journal of Mathematical Physics* **11** 343–348

Gibbons, G.W. (1972) *Communications in Mathematical Physics* **27** 87

Gibbons, G.W. (1981) *Springer Lecture Notes in Physics* **160** 145– 151

Gibbons, G.W. & Hull, C.M. (1982) *Physics Letters* **B109** 337–349

Gibbons, G.W. (1982a) in *Quantum Structure of Space and Time* eds. M.J. Duff and C.J. Isham (Cambridge University Press)

Gibbons,G W. (1982b) *Nuclear Physics* **B207** 337–349

Gibbons G.W., Hawking, S.W., Horowitz, G. & Perry, M.J. (1983) *Communications in Mathematical Physics* **88** 295–308

Gibbons, G.W.(1984) in *Supersymmetry, Supergravity and Related Topics* eds.F. del Aguila, J.A. de Azcarraga and L.E. Ibanez (World Scientific, Singapore)

Gibbons, G.W. & Ruback, P.J. (1986) *Physical Review Letters* **57** 1492–1495

26 04 2020

Pag. 71(91)

7.1 Gauge Problem

Pag. 72(92)

Gauge Invariance Lemma

Pag. 73(93) ✡

[The proof is straightforward

Pág. 185(205) —

4 Gravitational lenses in Cosmology

4.3 The linearized equation of geometric optics

Pág. 189(209)

F) The renaissance of general relativity and cosmology...pdf

Pag. 287(297)

Topology and topology change in general relativity.

Pág. 288(298)

2.1 The Einstein-Rosen Bridge

Pág. 289(299)

26.04.2020

Lorentzian metric g_L

→ Lorentzian metric g_L on a compact 4-manifold M...

G) Books_72_2020/
Spacecraft Trajectory Optimization.../
Bruce Conway - Spacecraft Trajectory optimization...pdf

Pag. 40 (54)

3.2 Transcription Methods

3.2.1 A basic collocation Method (Using Hermite Polynomials)

Figure 3.1 Trajectory optimization problem broken into phases.

152) 20200427

A) Books_72_2020/
Advanced Engineering Mathematics/
H.K. Dass - Advanced Engineering Mathematics...pdf

Pag. 128 (144)

2.16 Surface Area.

B) Books_72_2020/
Dirichlet Branes and Mirror Symmetry/
Dirichlet Branes...pdf

'()*' ' '+ ' C 2 - . D 2
+- : 4 ! (- -!
& $ $

of coordinates, for example polar or Cartesian coordinates, conventional elliptic elements, or Delaunay variables, for a given problem. It also requires discussion of control modeling and parameterization, grid refinement, and transformations of coordinates required for interplanetary flight when a spacecraft leaves the sphere of influence (SOI) of one body and enters another. Another important topic is the successful generation of an initial guess of the solution, that is, a guess of the optimal trajectory that is sufficiently accurate that the NLP solver can improve on it and converge to the local minimum of the objective.

3.2 Transcription Methods

3.2.1 A Basic Collocation Method (Using Hermite Polynomials)

This classic approach transforms an optimal control problem into a nonlinear programming problem. To do this, the state functions $x(t)$ and control functions $u(t)$ are represented by piecewise polynomials and collocation is used to satisfy the differential equations of motion. The first step with this approach is to subdivide the problem into a sequence of smaller trajectory arcs, which we will call phases. The length of each phase is defined as $T_{pi} = E_{i+1} - E_i$. For each phase, the interval $[E_i, E_{i+1}]$ is further subdivided into N segments, as shown in Figure 3.1.

Let the ratio of the length of the jth segment to T_{pi} be denoted as τ_{ji}. Thus, the length of the jth segment in the ith stage is $\tau_{ji}\ T_{pi}$. Using Hermite interpolation, cubic polynomials are defined for each state on each segment using values of the states at the nodes (the boundaries at the segments) and the state time derivatives, as defined by the equations, at the nodes. The values of the states are then selected (by nonlinear programming) to force the interpolated derivatives to agree with the differential equations at the center of the segment. This procedure is illustrated in Figure 3.2.

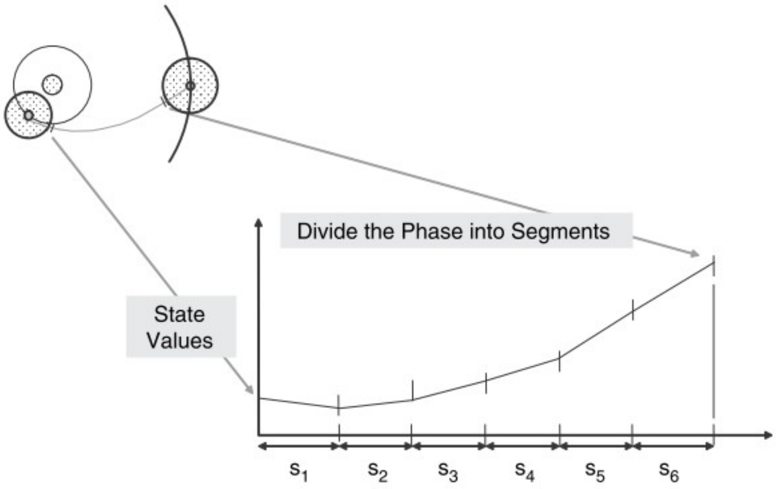

Figure 3.1. Trajectory optimization problem broken into phases.

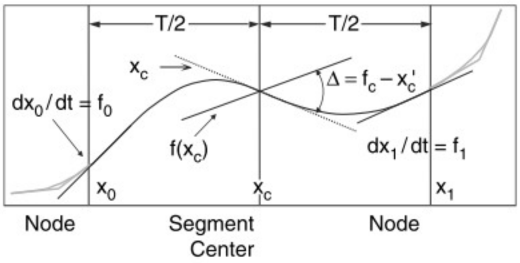

Figure 3.2. Illustration of implicit integration.

The basic procedure can be derived for any polynomial order. The classic cubic derivation is as follows. Let the states x be represented on each segment by cubics of the form

$$x = C_0 + C_1 S + C_2 S^2 + C_3 S^3 \tag{3.1}$$

where to simplify the discussion, the segment length S is transformed such that $S \in [0, 1]$. Let $x(0) = x_0$, $x(l) = x_1$, $dx/ds(0) = x_0'$, $dx/ds(l) = x_1'$. Differentiating Equation (3.1) and evaluating at $S = 0$ and $S = 1$ yields

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} C_0 \\ C_1 \\ C_2 \\ C_3 \end{bmatrix} = \begin{bmatrix} x_0 \\ x_0' \\ x_1 \\ x_1' \end{bmatrix}. \tag{3.2}$$

Inverting the 4 × 4 matrix yields

$$\begin{bmatrix} C_0 \\ C_1 \\ C_2 \\ C_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -3 & -2 & 3 & -1 \\ 2 & 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} x_0 \\ x_0' \\ x_1 \\ x_1' \end{bmatrix}. \tag{3.3}$$

Now using Equation (3.3), evaluating Equation (3.1) at $S = 1/2$, and transforming to segment length $T = g\tau_{ji}\, T_{pi}$, we see that the interpolated value of x at the center of the segment is

$$x_c = (x_0 + x_1)/2 + T(f_0 + f_1)/8 \tag{3.4}$$

where f_i, is the time rate of change as defined by the equations of motion evaluated at x_i.

(Note $x_0' = Tf[x_0, u(t_0), t_0, \omega]$.) In the same way, the slope at the center

$$x_c' = -3(x_0 - x_1)/2T - (f_0 - f_1)/8$$

is obtained. Evaluating the equations of motion at x_c gives f_c. We can then define the defect (the error in representing the dynamics) at the center of the segment as

$$\begin{aligned} \Delta &= f_c - x \\ &= f_c + 3(x_0 - x_1)/2T + (f_0 + f_1)/4 \end{aligned} \tag{3.5}$$

x_0 and x_1 are varied to enforce $\Delta = 0$. If the cubic polynomial is capable of representing the solution on the given segment, then selecting x_0 and x_1 to drive Δ to zero will produce an accurate approximation to the solution of the state dynamics. The defects for each state evaluated at the center of each segment constitute a set of nonlinear algebraic equations that are a function of the states and controls at each node, the events E, and the design parameters ω. The controls are defined at $S = [0, 1/2, 1]$. The boundary conditions and the path constraints evaluated at both the nodes and centers of the segments provide additional equations (constraints) to be satisfied. All of the independent variables may be collected into a single vector P defined by

$$P^T = [Z^T, E^T, \omega^T] \tag{3.6}$$

Where

$$Z^T = [x_0^T, u_0^T, x_1^T, u_1^T, \ldots\ldots x_{N+1}^T, u_{N+1}^T). \tag{3.7}$$

Collecting all of the nonlinear equations into a single vector equation yields

$$C^T = (\Delta^T, B_{N,}^T H_N^T) \tag{3.8}$$

with

$\Delta^T = (\Delta_{00}, \Delta_{01}, \ldots, \Delta_{\mathrm{ij}} \ldots)$ where $\Delta_{\mathrm{ij}} =$ defect for ith state at jth node
$B_N =$ collection of all boundary conditions
$H_N =$ collection of all path constraints

The performance index J is just a function of P. The trajectory optimization problem stated above can be expressed as: minimize $\Phi(P)$, subject to

$$l \leq \begin{Bmatrix} P \\ AP \\ C(P) \end{Bmatrix} \leq u\,. \tag{3.9}$$

AP is composed of all the linear relationships from Equation (3.8), l and u are the upper and lower bounds for the parameter vector. For equality constraints $l = u$. The upper and lower bounds for the great majority of the nonlinear constraints are usually set to zero, because this forces the solver to choose values for the parameters that satisfy the EOMs. (There may be a small number of additional nonlinear constraints, for example boundary conditions).

System (3.6)–(3.9) constitutes a nonlinear programming problem.

The extension of the preceding analysis to higher order integrators is straightforward. Again, a phase is subdivided into segments. The segments do not need to be of equal length. The rule of thumb, analogous to multistep explicit integrators, is that more segments should be placed where there are rapid changes in the states, controls or constraints.

Within a segment, the values of the states and controls are defined at discrete time points called nodes. For convenience, we define the placement of the nodes relative to the normalized segment lengths. An example set of nodes within a segment is shown in Figure 3.3.

The placement of nodes within a segment can be arbitrary; however, experience has shown benefits to placing the nodes at the Legendre-Gauss-Lobatto (LGL) points. This is accomplished by mapping the segment onto the interval −1 to 1 and using the LGL points in "standard" form. The odd points are called cardinal nodes (τ_1, τ_3, τ_5), while the even points (τ_2, τ_4) are labeled interior nodes. Typically, the number of cardinal nodes per segment is constant within a phase.

The state values specified at the cardinal nodes will be used as parameters for the resulting nonlinear programming problem, while the state values at the interior nodes will be obtained by interpolation. The equations of motion are evaluated to provide time derivative values at the cardinal nodes. Making use of both states values and their time derivatives at the cardinal nodes, Hermite interpolation is used to construct a polynomial representation of the states. This polynomial representation is then evaluated at the interior nodes to compute interpolated values of the states, which in turn are used to compute the state time derivatives at the interior nodes by evaluating the equations of motion. The polynomial representation is also differentiated and evaluated at the interior nodes to compute yet another set of time derivatives. The differences between the slopes as defined by the state representation and the equations of motion at the interior nodes are called the defects. The control functions are represented by values at both the cardinal and interior nodes. This procedure is illustrated in Figure 3.4.

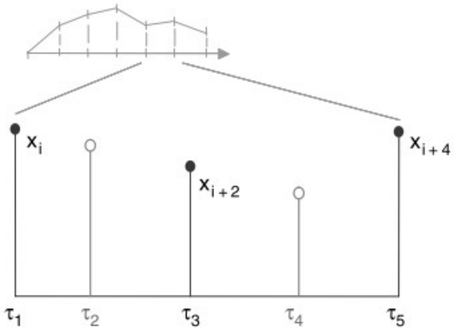

Figure 3.3. Cardinal node segment structure.

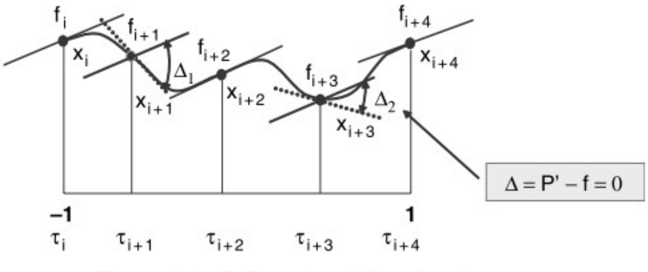

Figure 3.4. Collocation defect development.

This procedure can also be summarized as:
Using a Hermite interpolating polynomial, for each state, we can write

$$x(\tau) = \{u(\tau)\}^T \{x\} + \{v(\tau)\}^T \{f\} = \sum_{i=1}^{(n+1)/2} u_i(\tau) x(\tau_i) + \sum_{i=1}^{(n+1)/2} v_i(\tau) f(\tau_i). \tag{3.10}$$

The state values at the interior nodes can then be computed as

$$\{x_*\} = [E]\{x\} + \frac{T}{2}[F]\{f\}. \tag{3.11}$$

The matrices $[E]$ and $[F]$ are determined from the Hermite polynomial; T is segment length in "dimensional" time. This also allows

$$x'(\tau) = \{u'(\tau)\}^T \{x\} + \{v'(\tau)\}^T \{f\} = \sum_{i=1}^{(n+1)/2} u_i'(\tau) x(\tau_i) + \sum_{i=1}^{(n+1)/2} v_i'(\tau) f(\tau_i) \tag{3.12}$$

Applying our definition of defect, the vector $\{\Delta\}$ of defects at the interior nodes is

$$\{\Delta\} = [A]\{x\} + \frac{T}{2}[B]\{f\} + \frac{T}{2}[C]\{f_*\}. \tag{3.13}$$

with

$\{x\}$	states at the cardinal nodes
$\{f\}$	slopes from the equations of motion at the cardinal nodes
$\{f_*\}$	slopes from the equations of motion at the interior nodes, x_*
T	segment length (needed to convert from "τau" space to "dimensional" time
$[A]$	Matrix of constants
$[B]$	Matrix of constants
$[C]$	Identity Matrix multiplied by -1

The elements of $[A]$ and $[B]$ are just

$$a_{i,j} = u_j'(\tau_i)$$
$$b_{i,j} = v_j'(\tau_i)$$

which are easily built by applying the product rule of differentiation to ACM algorithm 211 [16] (used to compute $u \& v$)

$$u_j(\tau) = \left[\prod_{\substack{i=1 \\ i \neq j}}^{(n+1)/2} \left(\frac{(\tau - \tau_i)}{(\tau_j - \tau_i)} \right)^2 \right] \cdot \left(1 - (\tau - \tau_j) \cdot 2 \cdot \sum_{\substack{i=1 \\ i \neq j}}^{(n+1)/2} \frac{1}{(\tau_j - \tau_i)} \right)$$

$$v_j(\tau) = \left[\prod_{\substack{i=1 \\ i \neq j}}^{(n+1)/2} \left(\frac{(\tau - \tau_i)}{(\tau_j - \tau_i)} \right)^2 \right] \cdot (\tau - \tau_j). \tag{3.14}$$

Notice that the elements of $[A]$, $[B]$, $[C]$, $[D]$, $[E]$, and $[F]$ only depend on the values of τ (the node locations), which are fixed. This means that the matrices need only to be computed once and stored. Combining the segments results in the matrices taking on a banded structure, as shown below for a 2 segment, 3 cardinal node case

$$\begin{Bmatrix} \Delta_1 \\ \Delta_2 \\ \Delta_3 \\ \Delta_4 \end{Bmatrix} = \begin{bmatrix} -1.6654 & 1.4963 & 0.1701 & 0 & 0 \\ -0.1701 & 1.4963 & 1.6654 & 0 & 0 \\ 0 & 0 & -1.6654 & 1.4963 & 0.1701 \\ 0 & 0 & -0.1701 & 1.4963 & 1.6654 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_3 \\ x_5 \\ x_7 \\ x_9 \end{Bmatrix}$$

$$+ \frac{T}{2} \begin{bmatrix} -0.1386 & -0.6530 & -0.0450 & 0 & 0 \\ -0.0450 & -0.6530 & -0.1386 & 0 & 0 \\ 0 & 0 & -0.1386 & -0.6530 & -0.0450 \\ 0 & 0 & -0.0450 & -0.6530 & -0.1386 \end{bmatrix} \begin{Bmatrix} f_1 \\ f_3 \\ f_5 \\ f_7 \\ f_9 \end{Bmatrix}$$

$$+ \frac{T}{2} \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \begin{Bmatrix} f_2 \\ f_4 \\ f_6 \\ f_8 \end{Bmatrix} \tag{3.15}$$

Again we combine the state values at the cardinal nodes, along with the event times and design parameters, in a single array Z. A constraint array is assembled from the boundary conditions, defects, and path constraint. This results in a tractable nonlinear programming problem.

3.2.2 Pseudospectral Methods

Pseudospectral techniques have gained popularity within the last several years [17] [18] [19] [20]. They are very similar to the method of collocation. The major differences are with the manner in which the representative polynomial is constructed and the computation of the defects. When the pseudospectral method is employed, a polynomial representation of the states is formed and a number of constraints are enforced so that the slopes of the polynomials agree with the differential equations at a finite number of points. Values of the states are established at the cardinal and interior nodes. Making use of the state values, Lagrange interpolation is used to construct a polynomial interpolation of the states. This polynomial is differentiated and evaluated at all the nodes to compute interpolated time derivative values of the states. The equations of motion are evaluated to provide the values of the physical time derivatives. The difference between the two sets of time derivatives forms the defects. The procedure is shown in Figure 3.5.

data 26.04.2020

X

Lorentzian metric g_L

→ Lorentzian metric g_L on a compact 4-manifold M...

G) Books_72_2020/
Spacecraft Trajectory Optimization.../
Bruce Conway - Spacecraft Trajectory optimization...pdf

Pag. 40 (54)

3.2 Transcription Methods

3.2.1 A basic collocation Method (Using Hermite Polynomials)

Figure 3.1 Trajectory optimization problem broken into phases.

152) 20200427

A) Books_72_2020/
Advanced Engineering Mathematics/
H.K. Dass - Advanced Engineering Mathematics...pdf

Pag. 128 (144)

2.16 Surface Area.

B) Books_72_2020/
Dirichlet Branes and Mirror Symmetry/
Dirichlet Branes...pdf

'()*' ' '+ ' E > B (!1 ! 7
83 ! F 9 ! -!
'()*' ' '+ ' ! (!1 !
-! & ; $$ < (

5. Find the volume of the ellipsoid $\frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z^2}{c^2}=1$ **Ans.** $\frac{4\pi abc}{3}$

6. Find the volume bounded by the cylinder $x^2+y^2=a^2$ and the planes $y+z=2a$ and $z=0$.
(*M.U. II Semester 2000, 02, 06*) **Ans.** $2\pi a^3$

7. Find the volume bounded by the cylinder $x^2+y^2=a^2$ and the planes $z=0$ and $y+z=b$.
Ans. $\pi a^2 b$

8. Find the volume of the region bounded by $z=x^2+y^2$, $z=0$, $x=-a$, $x=a$ and $y=-a$, $y=a$.
Ans. $\frac{8}{3}a^4$

9. Find the volume enclosed by the cylinder $x^2+y^2=9$ and the planes $x+z=5$ and $z=0$.
Ans. $45\pi-36$

10. Compute the volume of the solid bounded by $x^2+y^2=z$, $z=2x$.(*A.M.I.E., Summer 2000*)
Ans. 2π

11. Find the volume cut from the paraboloid $4z=x^2+y^2$ by plane $z=4$.
(*U.P. I Semester, Dec. 2005*) **Ans.** 32π

12. By using triple integration find the volume cut off from the sphere $x^2+y^2+z^2=16$ by the plane $z=0$ and the cylinder $x^2+y^2=4x$. **Ans.** $\frac{64}{9}(3\pi-4)$

13. The sphere $x^2+y^2+z^2=a^2$ is pierced by the cylinder $x^2+y^2=a^2(x^2-y^2)$.

Prove that the volume of the sphere that lies inside the cylinder is $\frac{8}{3}\left[\frac{\pi}{4}+\frac{5}{3}-\frac{4\sqrt{2}}{3}\right]a^3$.

14. Find the volume of the solid bounded by the surfaces $z=0$, $3z=x^2+y^2$ and $x^2+y^2=9$.
(*A.M.I.E.T.E., Summer 2005*) **Ans.** $\frac{27\pi}{2}$

15. Obtain the volume bounded by the surface $z=c\left(1-\frac{x}{a}\right)\left(1-\frac{y}{b}\right)$ and a quadrant of the elliptic cylinder $\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$, $z>0$ and where a, $b>0$. **Ans.** πabc (*A.M.I.E.T.E., Dec. 2005*)

16. Find the volume of the paraboloid $x^2+y^2=4z$ cut off by the plane $z=4$. **Ans.** 32π

17. Find the volume bounded by the cone $z^2=x^2+y^2$ and the paraboloid $z=x^2+y^2$. **Ans.** $\frac{\pi}{6}$

18. Find the volume enclosed by the cylinders $x^2+y^2=2ax$ and $z^2=2ax$. **Ans.** $\frac{128a^3}{15}$

19. Find the volume of the solid bounded by the plane $z=0$, the paraboloid $z=x^2+y^2+2$ and the cylinder $x^2+y^2=4$. **Ans.** 16π

20. The triple integral $\iiint dx\,dy\,dz$ gives
(*a*) Volume of region (*b*) Surface area of region T
(*c*) Area of region *T* (*d*) Density of region *T*. (*A.M.I.E.T.E., Dec. 2006, 2002*) **Ans.** (*a*)

2.16 SURFACE AREA

Let $z=f(x,y)$ be the surface S. Let its projection on the x-y plane be the region A. Consider an element $\delta x.\ \delta y$ in the region A. Erect a cylinder on the element $\delta x.\ \delta y$ having its generator parallel to OZ and meeting the surface S in an element of area δs.

$$\therefore \qquad \delta x\ \delta y=\delta s\cos\gamma,$$

Where γ is the angle between the xy-plane and the tangent plane to S at P, *i.e.,* it is the angle between the Z-axis and the normal to S at P.

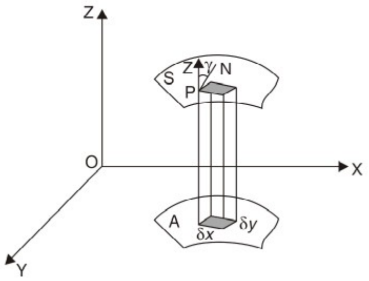

Created with Print2PDF. To remove this line, buy a license at: http://www.software602.com/

The direction cosines of the normal to the surface $F(x, y, z) = 0$ are proportional to

$$\frac{\partial F}{\partial x}, \frac{\partial F}{\partial y}, \frac{\partial F}{\partial z}$$

$\therefore$ The direction of the normal to $S\ [F = f(x, y) - z]$ are proportional to $-\frac{\partial z}{\partial x}, -\frac{\partial z}{\partial y}, 1$ and those of the Z-axis are 0, 0 , 1.

$$\text{Direction cosines} = \frac{-\frac{\partial z}{\partial x}}{\sqrt{\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2 + 1}}, \frac{-\frac{\partial z}{\partial y}}{\sqrt{\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2 + 1}}, \frac{1}{\sqrt{\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2 + 1}},$$

Hence
$$\cos\gamma = \frac{1}{\sqrt{\left[\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2 + 1\right]}} \quad (\cos\theta = l_1 l_2 + m_1 m_2 + n_1 n_2)$$

$$\delta S = \frac{\delta x\, \delta y}{\cos\gamma} = \sqrt{\left[\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2 + 1\right]}\, \delta x\, \delta y\ ; \quad S = \iint_A \sqrt{\left[\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2 + 1\right]}\, dx\, dy$$

Example 55. *Find the surface area of the cylinder* $x^2 + z^2 = 4$ *inside the cylinder* $x^2 + y^2 = 4$.

Solution. $x^2 + y^2 = 4$

$$2x + 2z\frac{\partial z}{\partial x} = 0 \text{ or } \frac{\partial z}{\partial x} = -\frac{x}{z}, \ \frac{\partial z}{\partial y} = 0$$

$$\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2 + 1 = \frac{x^2}{z^2} + 1 = \frac{x^2 + z^2}{z^2} = \frac{4}{4 - x^2}$$

Hence, the required surface area $= 8\int_0^2 \int_0^{\sqrt{4-x^2}} \sqrt{\left[\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2 + 1\right]}\, dx\, dy$

$$= 8\int_0^2 \int_0^{\sqrt{4-x^2}} \frac{2}{\sqrt{4-x^2}}\, dx\, dy = 16\int_0^2 \frac{1}{\sqrt{4-x^2}}\, [y]_0^{\sqrt{4-x^2}}\, dx = 16\int_0^2 \frac{1}{\sqrt{4-x^2}}\, [\sqrt{4-x^2}]\, dx$$

$= 16\int_0^2 dx = 16\,(x)_0^2 = 32$ **Ans.**

Example 56. *Find the surface area of the sphere* $x^2 + y^2 + z^2 = 9$ *lying inside the cylinder* $x^2 + y^2 = 3y$.

Solution.
$$x^2 + y^2 + z^2 = 9$$
$$2x + 2z\frac{\partial z}{\partial x} = 0, \quad \frac{\partial z}{\partial x} = -\frac{x}{z}$$
$$2x + 2z\frac{\partial z}{\partial y} = 0, \quad \frac{\partial z}{\partial y} = -\frac{y}{z}$$

$$\left[\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2 + 1\right] = \frac{x^2}{z^2} + \frac{y^2}{z^2} + 1 = \frac{x^2 + y^2 + z^2}{z^2} = \frac{9}{9 - x^2 - y^2} = \frac{9}{9 - r^2} \quad \begin{bmatrix} x = r\cos\theta \\ y = r\sin\theta \end{bmatrix}$$

$$x^2 + y^2 = 3y \text{ or } r^2 = 3r\sin\theta \text{ or } r = 3\sin\theta.$$

Hence, the required surface area

$$= \iint \sqrt{\left[\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2 + 1\right]}\, dx\, dy = 4\int_0^{\pi/2}\int_0^{3\sin\theta} \frac{3}{\sqrt{9 - r^2}}\, r\, d\theta\, dr = 12\int_0^{\pi/2} d\theta \int_0^{3\sin\theta} \frac{r\, dr}{\sqrt{9 - r^2}}$$

$$= 12\int_0^{\pi/2} d\theta\, [-\sqrt{9 - r^2}]_0^{3\sin\theta} = 12\int_0^{\pi/2} [-\sqrt{9 - 9\sin^2\theta} + 3]\, d\theta$$

Created with Print2PDF. To remove this line, buy a license at: http://www.software602.com/

$$= 36 \int_0^{\pi/2} (-\cos\theta + 1)\, d\theta = 36\,(-\sin\theta + \theta)_0^{\pi/2} = 36\left(-1 + \frac{\pi}{2}\right) = 18\,(\pi - 2)$$ **Ans.**

Example 57. *Find the surface area of the section of the cylinder $x^2 + y^2 = a^2$ made by the plane $x + y + z = a$.*

Solution.
$$x^2 + y^2 = a^2 \qquad \ldots (1)$$
$$x + y + z = a \qquad \ldots (2)$$

The projection of the surface area on *xy*-plane is a circle
$x^2 + y^2 = a^2$

$$1 + \frac{\partial z}{\partial x} = 0 \quad \text{or} \quad \frac{\partial z}{\partial x} = -1$$
$$1 + \frac{\partial z}{\partial y} = 0 \quad \text{or} \quad \frac{\partial z}{\partial y} = -1$$
$$\sqrt{\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2 + 1} = \sqrt{(-1)^2 + (-1)^2 + 1} = \sqrt{3}$$

Hence the required surface area

$$= 4 \int_0^a \int_0^{\sqrt{a^2 - x^2}} \sqrt{\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2 + 1}\; dx\, dy = 4 \int_0^a \int_0^{\sqrt{a^2 - x^2}} \sqrt{3}\, dx \cdot dy$$
$$= 4\sqrt{3} \int_0^a [y]_0^{\sqrt{a^2 - x^2}}\, dx = 4\sqrt{3} \int_0^a \sqrt{a^2 - x^2}\, dx$$
$$= 4\sqrt{3}\left[\frac{x}{2}\sqrt{a^2 - x^2} + \frac{a^2}{2}\sin^{-1}\frac{x}{a}\right]_0^a = 4\sqrt{3}\left[0 + \frac{a^2}{2}\,\frac{\pi}{2}\right] = 4\sqrt{3}\left(\frac{a^2\pi}{4}\right) = \sqrt{3}\,\pi\, a^2$$ **Ans.**

Example 58. *Find the area of that part of the surface of the paraboloid of the paraboloid $y^2 + z^2 = 2\,ax$, which lies between the cylinder, $y^2 = ax$ and the plane $x = a$.*

Solution.
$$y^2 + z^2 = 2\,ax \qquad \ldots (1)$$
$$y^2 = ax \qquad \ldots (2)$$
$$x = a \qquad \ldots (3)$$

Differentiating (1), we get

$$2\,z\,\frac{\partial z}{\partial x} = 2\,a, \quad \frac{\partial z}{\partial x} = \frac{a}{z}$$
$$2y + 2\,z\,\frac{\partial z}{\partial y} = 0, \quad \frac{\partial z}{\partial y} = -\frac{y}{z}$$
$$\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2 + 1 = \frac{a^2}{z^2} + \frac{y^2}{z^2} + 1 = \frac{a^2 + y^2}{z^2} + 1 \qquad \left[\begin{array}{l} y^2 + z^2 = 2\,ax \\ z^2 = 2\,ax - y^2 \end{array}\right]$$
$$= \frac{a^2 + y^2}{2\,a\,x - y^2} + 1 = \frac{a^2 + y^2 + 2\,a\,x - y^2}{2\,a\,x - y^2} = \frac{a^2 + 2\,a\,x}{2\,a\,x - y^2}$$
$$S = \int_0^a \int_{-\sqrt{ax}}^{\sqrt{ax}} \sqrt{\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2 + 1}\; dx\, dy = \int_0^a \int_{-\sqrt{ax}}^{\sqrt{ax}} \sqrt{\frac{a^2 + 2\,ax}{2\,ax - y^2}}\, dx\, dy \qquad \left[\begin{array}{c} y^2 = ax \\ y = \pm\sqrt{ax} \end{array}\right]$$
$$= \sqrt{a} \int_0^a \int_{-\sqrt{ax}}^{\sqrt{ax}} \sqrt{\frac{a + 2\,x}{2\,ax - y^2}}\, dx\, dy = \sqrt{a} \int_0^a \sqrt{a + 2\,x}\, dx \int_{-\sqrt{ax}}^{\sqrt{ax}} \frac{1}{\sqrt{2\,ax + y^2}}\, dy$$

Created with Print2PDF. To remove this line, buy a license at: http://www.software602.com/

$$= \sqrt{a}\int_0^a \sqrt{a+2x}\,dx\left[\sin^{-1}\frac{y}{\sqrt{2ax}}\right]_{-\sqrt{ax}}^{\sqrt{ax}}$$

$$= \sqrt{a}\int_0^a \sqrt{a+2x}\,dx\left[\sin^{-1}\frac{1}{\sqrt{2}}-\sin^{-1}\left(-\frac{1}{\sqrt{2}}\right)\right] = \sqrt{a}\int_0^a \sqrt{a+2x}\,dx\left[\frac{\pi}{4}-\left(\frac{\pi}{4}\right)\right]$$

$$= \sqrt{a}\,\frac{\pi}{2}\int_0^a \sqrt{a+2x}\,dx = \frac{\pi}{2}\cdot\frac{\sqrt{a}}{2}\cdot\frac{2}{3}[(a+2x)^{3/2}]_0^a$$

$$= \frac{\pi\sqrt{a}}{6}[(3a)^{3/2}-a^{3/2}] = \frac{\pi a^2}{6}[3\sqrt{3}-1] \qquad \textbf{Ans.}$$

EXERCISE 2.11

1. Find the surface area of sphere $x^2 + y^2 + z^2 = 16$. **Ans.** 64π
2. Find the surface area of the portion of the cylinder $x^2 + y^2 = 4y$ lying inside the sphere $x^2 + y^2 + z^2 = 16$. **Ans.** 64.
3. Show that the area of surfaces cz = xy intercepted by the cylinder $x^2 + y^2 = b^2$ is $\iint_A \frac{\sqrt{c^2+x^2+y^2}}{c}\,dx\,dy$, where A is the area of the circle $x^2 + y^2 = b^2$, $z = 0$ **Ans.** $\frac{2}{3}\frac{\pi}{c}\left[(c^2+b^2)^{\frac{1}{2}}-c^2\right]$
4. Find the area of the portion of the sphere $x^2 + y^2 + z^2 = a^2$ lying inside the cylinder $x^2 + y^2 = ax$. **Ans.** $2(\pi - 2)a^2$
5. Find the area of the surface of the cone $z^2 = 3(x^2 + y^2)$ cut out by the paraboloid $z = x^2 + y^2$ using surface integral. **Ans.** 6π

2.17 CALCULATION OF MASS

We have,

Volume = $\iint_V dx\,dy\,dz$ Density = $\rho = f(x, y, z)$

[Density = Mass per unit volume] Mass = Volume × Density

Mass = $\iint_V dx\,dy\,dz$ $\boxed{Mass = \iint_V f(x, y, z)\,dx\,dy\,dz}$

Example 59. Find the mass of a plate which is formed by the co-ordinate planes and the plane $\frac{x}{a}+\frac{y}{b}+\frac{z}{c}=1$, the density is given by $\rho = k\,x\,y\,z$. (*U.P., I Semester, Dec.*, 2003)

Solution. The plate is bounded by the planes $x = 0$, $y = 0$, $z = 0$ and $\frac{x}{a}+\frac{y}{b}+\frac{z}{c}=1$.

$$\text{Mass} = \iiint dx\,dy\,dz\,\rho = \int_0^c\int_0^{b\left(1-\frac{z}{c}\right)}\int_0^{a\left(1-\frac{y}{b}-\frac{z}{c}\right)} dx\,dy\,dz\,(k\,xyz)$$

$$= k\int_0^c z\,dz\int_0^{b\left(1-\frac{z}{c}\right)} y\,dy\int_0^{a\left(1-\frac{y}{b}-\frac{z}{c}\right)} x\,dx = k\int_0^c z\,dz\int_0^{b\left(1-\frac{z}{c}\right)} y\,dy\left(\frac{x^2}{2}\right)_0^{a\left(1-\frac{y}{b}-\frac{z}{c}\right)}$$

$$= k\int_0^c z\,dz\int_0^{b\left(1-\frac{z}{c}\right)} y\,dy\,\frac{a^2}{2}\left(1-\frac{y}{b}-\frac{z}{c}\right)^2 = \frac{k\,a^2}{2}\int_0^c z\,dz\int_0^{b\left(1-\frac{z}{c}\right)} y\left[\left(1-\frac{z}{c}\right)-\frac{y}{b}\right]^2 dy$$

$$= \frac{k\,a^2}{2}\int_0^c z\,dz\int_0^{b\left(1-\frac{z}{c}\right)}\left[y\left(1-\frac{z}{c}\right)^2+\frac{y^3}{b^2}-\frac{2y^2}{b}\left(1-\frac{z}{c}\right)\right]dy$$

Created with Print2PDF. To remove this line, buy a license at: http://www.software602.com/

$$= \frac{k\,a^2}{2}\int_0^c z\,dz\left[\frac{y^2}{2}\left(1-\frac{z}{c}\right)^2+\frac{y^4}{4\,b^2}-\frac{2\,y^3}{3\,b}\left(1-\frac{z}{c}\right)\right]_0^{b\left(1-\frac{z}{c}\right)}$$

$$= \frac{k\,a^2}{2}\int_0^c z\,dz\left[\frac{b^2}{2}\left(1-\frac{z}{c}\right)^4+\frac{b^4}{4\,b^2}\left(1-\frac{z}{c}\right)^4-\frac{2}{3}\cdot\frac{b^3}{b}\left(1-\frac{z}{c}\right)^4\right]$$

$$= \frac{k\,a^2}{2}\int_0^c z\left[\frac{b^2}{2}+\frac{b^2}{4}-\frac{2b^2}{3}\right]\left(1-\frac{z}{c}\right)^4 dz = \frac{k\,a^2}{2}\,\frac{b^2}{12}\int_0^c\left(1-\frac{z}{c}\right)^4 dz \qquad [\text{Put } z = c\sin^2\theta]$$

$$= \frac{k\,a^2\,b^2\,c^2}{12}\int_0^{\frac{\pi}{2}} c\sin^2\theta\,(1-\sin^2\theta)^4\,(2\,c\sin\theta\cos\theta\,d\theta)$$

$$= \frac{k^2\,a^2\,b^2\,c^2}{12}\int_0^{\pi/2}\sin^2\theta\,(\cos^8\theta)\sin\theta\cos\theta\,d\theta = \frac{k^2\,a^2\,b^2\,c^2}{12}\int_0^{\pi/2}\sin^3\theta\cos^9\theta\,d\theta$$

$$= \frac{k^2\,a^2\,b^2\,c^2}{12}\,\frac{\Gamma\left(\frac{3+1}{2}\right)\Gamma\left(\frac{9+1}{2}\right)}{2\,\Gamma\left(\frac{3+9+2}{2}\right)} = \frac{k\,a^2\,b^2\,c^2}{12}\cdot\frac{\Gamma 2\,\Gamma 5}{2\,\Gamma 7} = \frac{k\,a^2\,b^2\,c^2}{12}\,\frac{(1)\,(\Gamma 5)}{2\times 6\times 5\,\Gamma 5} = \frac{k\,a^2\,b^2\,c^2}{720} \quad \textbf{Ans.}$$

2.18 CENTRE OF GRAVITY

$$\bar{x} = \frac{\iiint x\,\rho\,dx\,dy\,dz}{\iiint \rho\,dx\,dy\,dz},\ \bar{y} = \frac{\iiint y\,\rho\,dx\,dy\,dz}{\iiint \rho\,dx\,dy\,dz},\ \bar{z} = \frac{\iiint z\,\rho\,dx\,dy\,dz}{\iiint \rho\,dx\,dy\,dz}$$

Example 60. *Find the co-ordinates of the centre of gravity of the positive octant of the sphere* $x^2 + y^2 + z^2 = a^2$, *density being given* $= k\,xyz$.

Solution. $\bar{x} = \dfrac{\iiint_V x\,\rho\,dx\,dy\,dz}{\iiint_V \rho\,dx\,dy\,dz} = \dfrac{\iiint z\,\rho\,dx\,dy\,dz}{\iiint \rho\,dx\,dy\,dz} = \dfrac{\iiint_V x^2 yz\,dx\,dy\,dz}{\iiint_V xyz\,dx\,dy\,dz}$

Converting into polar co-ordinates, $x = r\sin\theta\cos\phi$, $y = r\sin\theta\sin\phi$, $z = r\cos\theta$,
$dx\,dy\,dz = r^2\sin\theta\,dr\,d\theta\,d\phi$

$$\bar{x} = \frac{\int_0^{\pi/2}\int_0^{\pi/2}\int_0^a (r\sin\theta\cos\phi)^2\,(r\sin\theta\sin\phi)\,(r\cos\theta)\,(r^2\sin\theta\,dr\,d\theta\,d\phi)}{\int_0^{\pi/2}\int_0^{\pi/2}\int_0^a (r\sin\theta\cos\phi)\,(r\sin\theta\sin\phi)\,(r\cos\theta)\,(r^2\sin\theta\,dr\,d\theta\,d\phi)}$$

$$= \frac{\int_0^{\pi/2}\int_0^{\pi/2}\int_0^a r^6\sin^4\theta\cos\theta\sin\phi\cos^2\phi\,dr\,d\theta\,d\phi}{\int_0^{\pi/2}\int_0^{\pi/2}\int_0^a r^5\sin^3\theta\cos\theta\sin\phi\cos\phi\,dr\,d\theta\,d\phi}$$

$$= \frac{\int_0^{\pi/2}\sin\phi\cos^2\phi\,d\phi\int_0^{\pi/2}\sin^4\theta\cos\theta\,d\theta\int_0^a r^6\,dr}{\int_0^{\pi/2}\sin\phi\cos\phi\,d\phi\int_0^{\pi/2}\sin^3\theta\cos\theta\,d\theta\int_0^a r^5\,dr}$$

$$= \frac{\left[-\frac{\cos^3\phi}{3}\right]_0^{\pi/2}\left[\frac{\sin^5\theta}{5}\right]_0^{\pi/2}\left[\frac{r^7}{7}\right]_0^a}{-\left[\frac{\cos^2\phi}{2}\right]_0^{\pi/2}\left[\frac{\sin^4\theta}{4}\right]_0^{\pi/2}\left[\frac{r^6}{6}\right]_0^a} = \frac{\left(\frac{1}{3}\right)\left(\frac{1}{5}\right)\left(\frac{a^7}{7}\right)}{\left(\frac{1}{2}\right)\left(\frac{1}{4}\right)\left(\frac{a^6}{6}\right)} = \frac{16\,a}{35}$$

Similarly, $\bar{y} = \bar{z} = \dfrac{16\,a}{35}$; Hence, C.G. is $\left(\dfrac{16\,a}{35}, \dfrac{16\,a}{35}, \dfrac{16\,a}{35}\right)$ **Ans.**

Created with Print2PDF. To remove this line, buy a license at: http://www.software602.com/

4 '()*' ' '+ ' B C !
2 2 63 (2 -!
& < < < < ; $. - 1

We can think of $\mathrm{Tw}(\mathcal{A})$ as being the derived category of $\mathcal{A}$, and can write it as $\mathrm{D}^b(\mathcal{A})$.

Unfortunately, this is still not quite the correct category for us. The derived category $\mathrm{D}^b(X)$ of coherent sheaves on X has another property which $\mathrm{Fuk}(\check{X})$ or $\mathrm{D}^b(\mathrm{Fuk}(\check{X}))$ may not have. Namely, if $\mathcal{F}$ is an object in $\mathrm{D}^b(X)$, and $p : \mathcal{F} \to \mathcal{F}$ a morphism with $p^2 = p$, then one can think of the image of p as another object in $\mathrm{D}^b(X)$. Such a morphism p is called *idempotent*, and $\mathrm{D}^b(X)$ is *split-closed*, essentially saying that $\mathrm{D}^b(X)$ contains all images of idempotent morphisms.

In general, given a linear category, there is a way of constructing a larger category, the *Karoubi completion*, which contains all images of idempotent morphisms. This can be generalized to A_∞-categories, yielding from an A_∞-category $\mathcal{A}$ a split-closed category. We omit the details of this construction—the interested read can consult [**421**], §4. The construction has a similar flavour to the ones we have seen above. In particular, we denote by $\mathrm{D}^\pi(\mathcal{A})$ this process applied to $\mathrm{D}^b(\mathcal{A})$. This can be shown to be a split-closed triangulated A_∞-category. Once one does this, the hoped-for rigorous statement of Homological Mirror Symmetry becomes:

CONJECTURE 8.17. *Let X and $\check{X}$ be a mirror pair of Calabi-Yau manifolds. Then $\mathrm{D}^b_\infty(X)$ and $\mathrm{D}^\pi(\mathrm{Fuk}(\check{X}))$ are quasi-equivalent as A_∞-categories.*

8.4. The elliptic curve

We will now show a limited form of HMS for the elliptic curve. For simplicity, we will not try to compare the full derived and Fukaya categories, though what we do here constitutes most of the work. The case of the elliptic curve has already been studied in detail in [**397, 396**]; we will take a somewhat different approach here inspired by Abouzaid in [**2**] and developed in [**1**] which uses the Čech approach to defining the A_∞ structure on the derived category.

We will do this by building a degeneration of elliptic curves, a family $\mathcal{X} \to D$, where D can be viewed either as the unit disk with coordinate q or, in the category of schemes, $D = \mathrm{Spec}\,\mathbb{C}[[q]]$. This family will be a compactification of the family considered in §6.2.4. The advantage of working with such a family is that the parameter q then corresponds to the Kähler parameter q appearing in the Fukaya category of the mirror, i.e., the variable q in the Novikov ring. In particular, one should view either side as giving a family of categories parametrized by q. Furthermore, by working with an explicit degeneration, the methods introduced here fit well with the philosophy introduced by Gross and Siebert in [**205, 206**]. Although we do not elaborate on that philosophy here, it is helpful in this situation.

We will in fact build the degeneration $\mathcal{X} \to D$ along with a line bundle $\mathcal{L}$ which is ample (positive) when restricted to each fiber. We will then compute all higher multiplication maps involving powers, both positive and negative, of $\mathcal{L}$.

8.4.1. The degeneration: The Tate curve. The method of constructing our degeneration is a special case of a construction of Mumford for degenerations of abelian varieties [**373**]. This degeneration of elliptic curves is usually called the Tate curve. We start with the following data. Pick a degree d for the choice of line bundle, and consider a continuous piecewise linear function $\varphi : \mathbb{R} \to \mathbb{R}$ which has slope i on the interval $[i, i+1]$ for $i \in \mathbb{Z}$. We can choose this to take the value 0 on $[0, 1]$ to be specific; in this case,

$$\varphi(x) = ix - \tfrac{i(i+1)}{2} \text{ for } x \in [i, i+1]$$

and φ satisfies a periodicity condition

$$\varphi(x + d) = \varphi(x) + d \cdot x + \frac{d(d-1)}{2}. \tag{8.12}$$

Let

$$\Delta = \{(x, y) \in \mathbb{R}^2 | y \geq \varphi(x)\}.$$

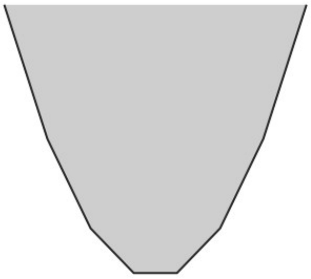

FIGURE 24. The graph of φ. The shaded area is Δ.

Just as a convex polytope determines a projective toric variety, an unbounded polytope can also determine a toric variety. This can be described via the *normal fan* of Δ. For each face $\sigma \subseteq \Delta$, define

$$\check{\sigma} = \{n \in (\mathbb{R}^2)^\vee : n|_\sigma = \text{constant}, \langle n, m\rangle \geq \langle n, m'\rangle \quad \forall m \in \Delta, m' \in \sigma\},$$

the *normal cone* to Δ at σ. We define the normal fan of Δ to be the collection of cones $\{\check{\sigma} | \sigma \subseteq \Delta \text{ a face}\}$. This is a fan living in $(\mathbb{R}^2)^\vee$. It is easy to see that if σ is the interval with endpoints $(i, \varphi(i))$ and $(i+1, \varphi(i+1))$, then $\check{\sigma}$ is the ray generated by $(-i, 1)$. Thus the normal fan Σ of Δ is as depicted in Figure 25. This violates a standard notion of fan in that it contains an infinite number of cones, but we should not be bothered by this: this fan still defines a toric variety X_Σ, covered by an *infinite* number of affine toric charts coming from the cones of Σ. Furthermore, $(0,1) \in \mathbb{R}^2$ takes the value 1 on the primitive generators of the rays of Σ, so if we denote by q the monomial corresponding to $(0,1)$, q is in fact a regular function on X_Σ, defining a map $g : X_\Sigma \to \mathbb{A}^1$.

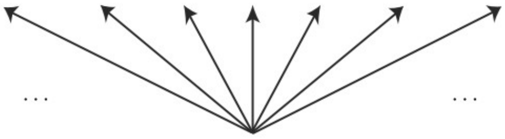

FIGURE 25. The normal fan of Δ.

Since q vanishes to order 1 on each toric divisor of X_Σ (corresponding to a ray of Σ) the fiber $g^{-1}(0)$ consists of an infinite chain of $\mathbb{P}^1$'s. On the other hand, it is not hard to see that any other fiber of g is a copy of $\mathbb{C}^*$.

Like any toric variety, X_Σ has an open cover indexed by the cones in the fan Σ, or equivalently, by the faces of Δ. In particular, we have an open cover $\{U_w | w \text{ is a vertex of } \Delta\}$, where $U_w = \operatorname{Spec} \mathbb{C}[\check{w}^\vee \cap \mathbb{Z}^2]$. This is the open subset of X_Σ corresponding to the cone $\check{w}$ of the fan Σ; alternatively $\check{w}^\vee$ can be viewed as just the tangent wedge to Δ at w, see Figure 26. Similarly, if σ is an edge of Δ, we obtain an open subset U_σ of X_Σ with $U_\sigma = \operatorname{Spec} \mathbb{C}[\check{\sigma}^\vee \cap \mathbb{Z}^2]$.

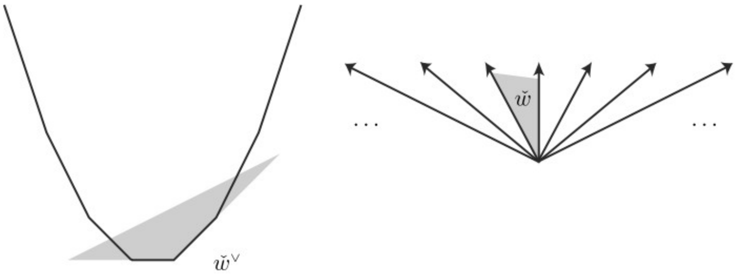

FIGURE 26

Now $\mathbb{Z}$ acts on X_Σ. We can describe this action by describing a $\mathbb{Z}$-action on the fan Σ. The generator of the action is given by the matrix $\left(\begin{smallmatrix} 1 & d \\ 0 & 1 \end{smallmatrix}\right)$ on $(\mathbb{R}^2)^\vee$: this takes a ray $\mathbb{R}_{\geq 0}(i,1)$ to $\mathbb{R}_{\geq 0}(i+d,1)$. This has the effect of acting on the big torus orbit $(\mathbb{C}^*)^2 \subseteq X_\Sigma$, with coordinates z, q corresponding to the standard basis $(1,0), (0,1)$ of $\mathbb{R}^2$, via

$$(z,q) \mapsto (zq^d, q).$$

Hence if we fix a non-zero value of q, the action on the $\mathbb{C}^*$ parametrized by z is $z \mapsto z \cdot q^d$. For $0 < |q| < 1$, the quotient of this fiber by the $\mathbb{Z}$-action is $\mathbb{C}^*/q^{\mathbb{Z}}$. This quotient coincides with

$$\mathbb{C}/\langle 1, \frac{1}{2\pi\sqrt{-1}} \log q\rangle;$$

note that $\exp(2\pi\sqrt{-1}\cdot)$ identifies $\mathbb{C}/\mathbb{Z}$ with $\mathbb{C}^*$. Clearly this is an elliptic curve as long as $|q| \neq 1$.

On the other hand, for $q = 0$, the action on the fiber shifts the infinite chain of $\mathbb{P}^1$'s d places, so the quotient by this $\mathbb{Z}$-action is a cycle of d rational curves, as depicted in Figure 27.

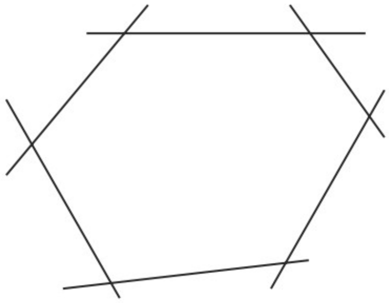

FIGURE 27

We can now take

$$\mathcal{X} = g^{-1}(D)/\mathbb{Z},$$

where D is the open unit disk. Then $g : \mathcal{X} \to D$, induced by q, is a degeneration of elliptic curves to a so-called Kodaira type I_d fiber, i.e., a cycle of d rational curves.

For our purposes, just as in the Fukaya category where we worked with the Novikov ring, a formal power series ring, it is more convenient for us to work with a formal version of this quotient. This is a rather technical point, but it allows us essentially to work entirely with the topology of the singular fiber, and also allows us to take the $\mathbb{Z}$ quotient in a more algebraic category.

Indeed, it is difficult to divide algebraic varieties by groups such as $\mathbb{Z}$. In this situation, we observe first of all that $g^{-1}(0)/\mathbb{Z}$ makes sense as an algebraic variety; as we saw, it is a cycle of d rational curves. On the other hand, let X_Σ^k denote the subscheme of X_Σ defined by the equation $q^{k+1} = 0$, so that X_Σ^k is a "kth order thickening" of $g^{-1}(0)$. Then in fact one can check (the details aren't so important for us) that $X_\Sigma^k/\mathbb{Z}$ also makes sense as a scheme, as a "kth order thickening" of $g^{-1}(0)/\mathbb{Z}$.

We can then construct the "formal scheme"

$$\hat{X}_\Sigma = \varprojlim X^k_\Sigma.$$

See [**222**], II §9, for the definition of a formal scheme. This is a ringed space whose underlying topological space is that of X^k_Σ (these coincide for all k), or equivalently, it has underlying topological space $g^{-1}(0)$. The sheaf of rings on $\hat{X}_\Sigma$ is given on an open set $U \subseteq g^{-1}(0)$ by

$$\Gamma(U, \mathscr{O}_{\hat{X}_\Sigma}) = \varprojlim \Gamma(U, \mathscr{O}_{X^k_\Sigma}).$$

(See, e.g., [**25**] for the definition of inverse limit and more details about completions of rings.) Essentially what this does is it allows us to consider functions in a "formal neighbourhood" of $g^{-1}(0)$ which are formal power series in q. For example, if U_w is the open set of X_Σ defined above, then $\hat{U}_w = g^{-1}(0) \cap U_w$ is topologically an open subset of $g^{-1}(0)$, and

$$\Gamma(\hat{U}_w, \mathscr{O}_{\hat{X}_\Sigma}) = \varprojlim \mathbb{C}[\check{w}^\vee \cap \mathbb{Z}^2]/(q^n) =: \mathbb{C}[\check{w}^\vee \cap \mathbb{Z}^2] \hat{\otimes}_{\mathbb{C}[q]} \mathbb{C}[[q]],$$

where the completed tensor product on the right is defined by the inverse limit. Then

$$\hat{\mathcal{X}} := \hat{X}_\Sigma/\mathbb{Z}$$

makes sense as a formal scheme also, and there is a map of formal schemes $\hat{g} : \hat{\mathcal{X}} \to \hat{\mathbb{A}}^1$ induced by q. Here $\hat{\mathbb{A}}^1$ is the ringed space consisting of the point 0 and the ring $\mathbb{C}[[q]]$. This should be viewed as a formal version of $g : \mathcal{X} \to D$.

We will actually calculate the desired multiplication maps on $\hat{\mathcal{X}}$; all answers will be formal power series in q. However, we first need to construct the desired line bundles on $\hat{\mathcal{X}}$.

To do so, we choose a relatively ample line bundle $\mathcal{L}$ on $\hat{\mathcal{X}}$. In fact, the choice of $\mathcal{L}$ is already present in Δ: a polytope determines not just a toric variety but a line bundle on it. In this case there is a line bundle $\mathcal{L} = \mathscr{O}_\Delta(1)$ on X_Σ which has a basis of sections which is in one-to-one correspondence with integral points of Δ. We can describe $\mathcal{L}$ via its transition maps in the usual way for line bundles on toric varieties, by saying that $\mathcal{L}|_{U_w}$ is trivial, identified naturally with $\mathscr{O}_{U_w} \cdot z^w$. Thus on $U_w \cap U_{w'}$, the transition map from $\mathscr{O}_{U_w}$ to $\mathscr{O}_{U_{w'}}$ is given by multiplication by $z^{w-w'}$. One sees easily that

$$\Delta = \bigcap_{w \text{ a vertex of } \Delta} \check{w}^\vee + w$$

and any integral point of $m \in \Delta$ then gives a well-defined section z^m of $\mathcal{L}$. Note that if $\mathcal{L}$ is restricted to any irreducible component of $g^{-1}(0)$, we get the line bundle $\mathscr{O}_{\mathbb{P}^1}(1)$.

We want $\mathcal{L}$ to induce a line bundle $\hat{\mathcal{L}}$ on $\hat{\mathcal{X}}$. In order to do this, it is not enough to know how $\mathbb{Z}$ acts on $\mathcal{X}$, but how this $\mathbb{Z}$-action lifts to an action on $\mathcal{L}$. For this purpose, it is enough to describe a $\mathbb{Z}$-action on

$$\bigoplus_{n=0}^{\infty} H^0(X_\Sigma, \mathcal{L}^{\otimes n}).$$

This space of sections has a canonical basis given by the integral points of $C(\Delta) \subseteq \mathbb{R}^3$, where

$$C(\Delta) = \overline{\{(rm, r) \mid m \in \Delta, r \in [0, \infty)\}}.$$

Note that the integral points with third coordinate an integer n correspond to elements of a basis for $H^0(X_\Sigma, \mathcal{L}^{\otimes n})$. We want an action which satisfies the following properties:

(1) It is linear on $\mathbb{Z}^3 \subseteq \mathbb{R}^3$, as we want the action of $\mathbb{Z}$ on $\mathcal{L}$ to act by toric automorphisms of the total space of $\mathcal{L}$.
(2) This action preserves the third component, i.e., does not change the degree of a section.
(3) As the action on $\mathcal{L}$ should be compatible with the action on X_Σ, the action on $\mathbb{Z}^2 \times \{0\} \subseteq \mathbb{Z}^3$ should be generated by
$$\begin{pmatrix} 1 & 0 \\ d & 1 \end{pmatrix}$$
(the transpose of the action on $(\mathbb{R}^2)^\vee$).
(4) The action should identify $C(\Delta)$ with itself.

The first three conditions tell us that the generator of the action must be

$$\begin{pmatrix} 1 & 0 & * \\ d & 1 & * \\ 0 & 0 & 1 \end{pmatrix}$$

From the fourth condition, the action must preserve the boundary of $C(\Delta)$, and it follows from the periodicity condition (8.12) that the only choice for this generator is

$$T = \begin{pmatrix} 1 & 0 & d \\ d & 1 & \frac{d(d-1)}{2} \\ 0 & 0 & 1 \end{pmatrix}.$$

For a given $n \geq 0$, we will write this generator as an affine transformation

$$T_n(x, y) := (x + nd, xd + y + \frac{d(d-1)}{2} n),$$

the first two components of $T(x, y, n)$. This describes the action on sections of $\mathcal{L}^{\otimes n}$, a basis being represented by integral points of $n\Delta$.

As before, we can now divide $\mathcal{L}$ out by $\mathbb{Z}$, and this only makes sense on the formal neighbourhood $\hat{X}_\Sigma$ of $g^{-1}(0)$. We then get a line bundle $\hat{\mathcal{L}}$ on $\hat{\mathcal{X}}$.

The easiest way to describe what we have constructed is to write down a $\mathbb{C}[[q]]$-basis of sections for $H^0(\hat{\mathcal{X}}, \hat{\mathcal{L}}^{\otimes n})$. To do so, one writes down a basis of $\mathbb{Z}$-invariant sections of $H^0(X_\Sigma, \mathcal{L}^{\otimes n})$. This basis is as follows: pick any $p \in \frac{1}{n}\mathbb{Z}$. Then set

$$\vartheta_{n,p} := \sum_{s=-\infty}^{\infty} z^{T_n^s(np, n\varphi(p))}.$$

Note that on any given open set U_w, this makes sense formally, i.e.,

$$\vartheta_{n,p} \in (\mathbb{C}[\check{w}^\vee \cap \mathbb{Z}^2]\hat{\otimes}_{\mathbb{C}[q]}\mathbb{C}[[q]]) \cdot z^{nw}.$$

To see why this makes sense, one can check that for any given $k > 0$, all but a finite number of terms in the sum defining $\vartheta_{n,p}$ are zero in $(\mathbb{C}[\check{w}^\vee \cap \mathbb{Z}^2] \otimes_{\mathbb{C}[q]} \mathbb{C}[q]/(q^k)) \cdot z^{nw}$. The functions $\vartheta_{n,p}$ are known as *theta functions*. Note also that

$$\vartheta_{n,p} = \vartheta_{n,p+d}$$

as

$$T_n(np, n\varphi(p)) = (n(p+d), n\varphi(p+d)).$$

It is not difficult to see that the $\vartheta_{n,p}$'s generate the space of $\mathbb{Z}$-invariant sections of $\mathcal{L}^{\otimes n}|_{\hat{X}_\Sigma}$ as a $\mathbb{C}[[q]]$-module. Note however that we need to allow formal power series in q in order for there to exist $\mathbb{Z}$-invariant sections: $\mathcal{L}^{\otimes n}$ on X_Σ has no such sections. This is another way of seeing why we need to pass to the completion before we can take the quotient. Note more generally that if we have any $(x, y) \in \mathbb{Z}^2$ with $(x, y) \in n\Delta$, i.e., $y \geq n\varphi(x/n)$, then

$$\sum_{s=-\infty}^{\infty} z^{T_n^s(x,y)} = q^{y-n\varphi(x/n)}\vartheta_{n,x/n}. \tag{8.13}$$

We define

$$\operatorname{ord}_n(x, y) = y - n\varphi(x/n) \tag{8.14}$$

for any $n \in \mathbb{Z}$, $(x, y) \in \mathbb{Z}^2$. (8.13) is the motivation for this definition.

Note also that elements of this basis of $\mathbb{Z}$-invariant sections are in one-to-one correspondence with points on the integral affine manifold (see §6.1.2) $B = \mathbb{R}/d\mathbb{Z}$ with coordinates in $\frac{1}{n}\mathbb{Z}$, a point which we shall return to in a moment. We write this set of points as $B((1/n)\mathbb{Z})$.

At this point, a reader unhappy with the vague description of the quotient construction we have given may proceed as follows. Consider the $\mathbb{C}[[q]]$-algebra

$$\hat{R} = \bigoplus_{n=0}^{\infty} H^0(\hat{\mathcal{X}}, \hat{\mathcal{L}}^{\otimes n}).$$

One can define this simply by knowing the set of $\mathbb{C}[[q]]$-module generators,

$$\{1\} \cup \{\vartheta_{n,p} | p \in B((1/n)\mathbb{Z}), n \geq 1\},$$

and the multiplication map, i.e., the structure constants of the algebra, something which we will explain shortly. We can then consider the scheme $\operatorname{Proj} \hat{R}$, which is a scheme defined over $\operatorname{Spec} \mathbb{C}[[q]]$ (see for example [**222**], II,

§2). This is a genuine scheme, not a formal scheme. The fact that one can obtain a genuine scheme from a formal scheme in the presence of an ample line bundle is Grothendieck's existence theorem, see EGA III, 5.4.5, [**211**]. So the construction above can be viewed as a special case.

8.4.2. Multiplication of theta functions and triangles. Let us now turn to a geometric interpretation for theta functions and their multiplication. We have just seen an affine manifold B emerge from our description of the basis of theta functions. So it is natural to try to fit what we are doing into the context of Chapter 6.

We will in fact view a fiber of the map $\mathcal{X} \to D$ as $\check{X}(B)$, see §6.2.1. Indeed, the fiber comes along with the restriction of the line bundle $\mathcal{L}$, which is necessarily of degree d. On the other hand, $\check{X}(B)$ comes with a canonical symplectic form ω with $\int_{\check{X}(B)} \omega = d$, so ω represents $c_1(\mathcal{L})$.

The reader may object that this goes against the philosophy of §6.3: namely, one considers line bundles on $X(B)$ and Lagrangians on $\check{X}(B)$. This is indeed the case, but there is a reason for this switch. When we discuss the Fukaya category, we will be able to work, at least in the elliptic curve case, with a fixed almost complex structure, namely, the complex structure $X(B)$. It will be easy to describe holomorphic disks on $X(B)$. By choosing to work on $X(B)$ on the Fukaya side, we are making it easy to describe disks but difficult to describe Lagrangians. Similarly, by working on $\check{X}(B)$ for the derived category side, it is difficult to describe sheaves, but we get a canonically given cohomology class from the symplectic form. As a result, we restrict attention to line bundles with first Chern class represented by integral multiples of ω, and more specifically, powers of $\mathcal{L}$. On $X(B)$, we restrict attention to Lagrangians of a very special sort, namely sections which are induced by multiples of the developing map (§6.1.2), with trivial local system.

Let us be more explicit now and describe, in the spirit of §6.3, the Lagrangians on $X(B)$ which will be mirror dual to the line bundles $\mathcal{L}^{\otimes n}$. First, we need to specify a symplectic form on $X(B)$: by Proposition 6.14, this can be done by specifying a convex multi-valued function on B; the function $K = y^2/2$ will do nicely, with $\omega = dy \wedge dx$. Here we are briefly using the convention of Chapter 6 and writing y as the coordinate on B, and x the fiber coordinate on $X(B)$. Now any real curve contained in $X(B)$ is Lagrangian, so the only role ω plays is in calculating areas of holomorphic disks.

For any $n \in \mathbb{Z}$, let L_n be the image in $X(B)$ of the graph of $x = -ny$, oriented in the direction of increasing y. The minus sign has to do with the choice of sign conventions in defining the Fukaya category. Since x is a periodic coordinate with period 1, and y is a periodic coordinate with period d, this makes sense (see Figure 28). One first notices that if $f : X(B) \to B$

data 26.04.2020
X

Lorentzian metric g_L

→ Lorentzian metric g_L on a compact
4-manifold M...

G) Books_72_2020/
Spacecraft Trajectory Optimization.../
Bruce Conway - Spacecraft
Trajectory Optimization...pdf

Pag. 40 (54)

3.2 Transcription Methods

3.2.1 A basic collocation Method
(Using Hermite Polynomials)

Figure 3.1 Trajectory optimization
problem broken into phases.

---

152) 20200427

A) Books_72_2020/
Advanced Engineering Mathematics/
H.K. Dass - Advanced Engineering
Mathematics...pdf

Pag. 128 (144)

2.16 Surface Area.

B) Books_72_2020/
Dirichlet Branes and Mirror Symmetry/
Dirichlet Branes...pdf

27.04.2020

Pág. 616(626)

8.4 The elliptic curve.

C) Books_72_2020/
Complex Variables and Numerical Methods/
Ravish R Singh - Complex Variables.pdf

Pág. 20

1.3 Geometrical Representation of Complex Numbers (Argand's Diagram)

Obs: Ver livro todo!

D) Books_72_2020/
Nonlinear Optics, 4th Edition/
Nonlinear Optics, 4th ed.pdf

Pág. 11(34)

Self-Focusing

Look: Figure 1.2.6.

E) Books_72_2020/
Formigeo_fai_edu/
Tratado0508.pdf

Pág. 15 Bissectrizes dos ângulos de um triângulo.

'()*' ' '+ ') 1) / G C
- @ 4 ! A ! 0 D . 5 1 2 8 0 E
7! 8 -!
& #

1.2 COMPLEX NUMBERS

A complex number z is an ordered pair (x, y) of real numbers x and y. It is written as $z = (x, y)$ or $z = x + iy$, where $i = \sqrt{-1}$ is known as the imaginary unit. Here, x is called the real part of z and is written as "**Re** (z)", and y is called the imaginary part of z and is written as "**Im** (z)".

If $x = 0$ and $y \neq 0$ then $z = 0 + iy = iy$ which is purely imaginary.

If $x \neq 0$ and $y = 0$ then $z = x + i\,0 = x$ which is real.

Hence, z is **purely imaginary**, if its real part is zero, and is **real** if its imaginary part is zero.

This shows that every real number can be written in the form of a complex number by taking its imaginary part as zero. Hence, the set of real numbers is contained in the set of complex numbers.

Two complex numbers are equal if and only if their corresponding real and imaginary parts are equal.

If $z = x + iy$ is a complex number then its conjugate or complex conjugate is defined as $\overline{z} = x - iy$.

The even power of i is either 1 or –1 and odd power of i is either i or $-i$.

$$i^2 = i.i = -1$$
$$i^3 = i^2.i = -i$$
$$i^4 = (i^2)^2 = (-1)^2 = 1$$
$$i^5 = i.\,i^4 = i, \text{ etc.}$$

1.3 GEOMETRICAL REPRESENTATION OF COMPLEX NUMBERS (ARGAND'S DIAGRAM)

Any complex number $z = x + iy$ can be represented as a point $P(x, y)$ in the xy-plane with reference to the rectangular x and y axes.

The plot of a given complex number $z = x + iy$, as the point $P(x, y)$ in the xy-plane is known as **Argand's diagram** (Fig. 1.1). The x-axis is called the real axis, y-axis is called the imaginary axis, and the xy-plane is called the complex plane.

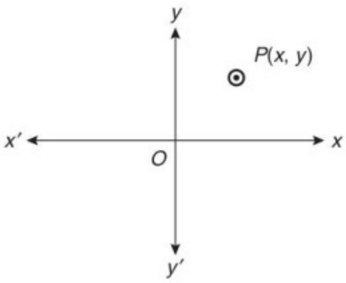

Fig. 1.1

1.4 ALGEBRA OF COMPLEX NUMBERS

Let $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$ be two complex numbers.

(a) *Addition:* $$z_1 + z_2 = (x_1 + iy_1) + (x_2 + iy_2) = (x_1 + x_2) + i\,(y_1 + y_2)$$

(b) *Subtraction:* $z_1 - z_2 = (x_1 + iy_1) - (x_2 + iy_2)$

$$= (x_1 - x_2) + i\,(y_1 - y_2)$$

(c) *Multiplication:* $z_1 z_2 = (x_1 + iy_1)\,(x_2 + iy_2)$

$$= (x_1 x_2 - y_1 y_2) + i\,(x_2 y_1 + y_2 x_1) \qquad [\because\; i^2 = -1]$$

(d) *Division:*

$$\frac{z_1}{z_2} = \frac{x_1 + iy_1}{x_2 + iy_2}$$

$$= \frac{(x_1 + iy_1)}{(x_2 + iy_2)} \cdot \frac{(x_2 - iy_2)}{(x_2 - iy_2)}$$

$$= \frac{x_1x_2 + y_1y_2}{x_2^2 + y_2^2} + i\left(\frac{y_1x_2 - x_1y_2}{x_2^2 + y_2^2}\right)$$

1.5 DIFFERENT FORMS OF COMPLEX NUMBERS

1.5.1 Cartesian or Rectangular Form

If x and y are real numbers then $z = x + iy$ is called the Cartesian form of the complex number.

1.5.2 Polar Form

The complex number $z = x + iy$ can be represented by the point P whose cartesian co-ordinates are (x, y) (Fig. 1.2). If polar coordinates of the point P are (r, θ) then $x = r\cos\theta$ and $y = r\sin\theta$.

Hence, polar form of z is

$$z = r\cos\theta + ir\sin\theta$$

$$= r(\cos\theta + i\sin\theta)$$

Polar form can also be written as $r \angle \theta$.

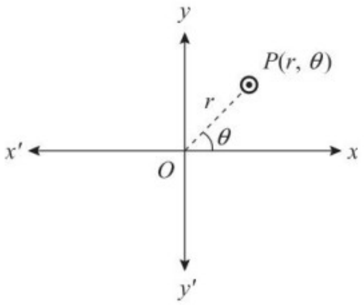

Fig. 1.2

1.5.3 Exponential Form

Polar form of z is

$$z = r(\cos\theta + i\sin\theta)$$

$$\cos\theta + i\sin\theta = e^{i\theta} \qquad \text{[Using Expansion]}$$

Hence, exponential form of z is

$$z = re^{i\theta}$$

Note $e^{i\theta} = \cos\theta + i\sin\theta$, $e^{-i\theta} = \cos\theta - i\sin\theta$.

Hence, $\cos\theta = \frac{1}{2}(e^{i\theta} + e^{-i\theta})$ and $\sin\theta = \frac{1}{2i}(e^{i\theta} - e^{-i\theta})$

1.6 MODULUS AND ARGUMENT (OR AMPLITUDE) OF COMPLEX NUMBERS

Let
$$\begin{aligned} z &= x + iy \\ &= r(\cos\theta + i\sin\theta) \end{aligned}$$

Here 'r' is called the modulus or absolute value of z and is denoted by $|z|$ or mod (z) and θ is called argument or amplitude of z and is denoted by arg (z) or amp (z).

Hence,
$$|z| = r = \sqrt{x^2 + y^2}$$
$$\arg(z) = \theta = \tan^{-1}\left(\frac{y}{x}\right)$$

Note The value of θ lying between $-\pi$ and π is called the **principal value** of argument. The argument of z is the value of θ which lies in the quadrant of point (x, y).

1.7 PROPERTIES OF COMPLEX NUMBERS

Let $z = x + iy$ and $\bar{z} = x - iy$.

(a) $\text{Re}(z) = x = \dfrac{1}{2}(z + \bar{z})$

(b) $\text{Im}(z) = y = \dfrac{1}{2i}(z - \bar{z})$

(c) $(\overline{z_1 + z_2}) = \bar{z}_1 + \bar{z}_2$

(d) $(\overline{z_1 z_2}) = \bar{z}_1 \bar{z}_2$

(e) $\overline{\left(\dfrac{z_1}{z_2}\right)} = \dfrac{\bar{z}_1}{\bar{z}_2}$

(f) $z\bar{z} = |z|^2 = |\bar{z}|^2 \qquad \left[\because |z| = |\bar{z}| = \sqrt{x^2 + y^2}\right]$

(g) $|z_1 z_2| = |z_1|\,|z_2|$

and $\arg(z_1 z_2) = \arg(z_1) + \arg(z_2)$

(h) $\left|\dfrac{z_1}{z_2}\right| = \dfrac{|z_1|}{|z_2|}$

and $\arg\left(\dfrac{z_1}{z_2}\right) = \arg(z_1) - \arg(z_2)$

Example 1

Arg $(z_1, z_2) = $ *Arg* $(z_1) + $ *Arg* (z_2)? *Justify.* **[Summer 2015]**

Solution

Yes, $Arg\,(z_1\, z_2) = Arg\,(z_1) + Arg\,(z_2)$

Let $z_1 = r_1 e^{i\theta_1},\ z_2 = r_2 e^{i\theta_2}$

$$\begin{aligned} z_1 z_2 &= (r_1\, e^{i\theta_1})\,(r_2\, e^{i\theta_2}) \\ &= r_1\, r_2\, e^{i(\theta_1 + \theta_2)} \end{aligned}$$

$$\therefore \qquad \begin{aligned} Arg\,(z_1\, z_2) &= \theta_1 + \theta_2 \\ &= Arg\,(z_1) + Arg(z_2) \end{aligned}$$

Example 2

Find the principal argument Arg z when $z = \dfrac{-2}{1+\sqrt{3}\,i}$. **[Winter 2014]**

Solution

$$\begin{aligned} z &= \frac{-2}{1+\sqrt{3}\,i} \\ &= \frac{-2}{1+\sqrt{3}i} \cdot \frac{\left(1-\sqrt{3}\,i\right)}{\left(1-\sqrt{3}\,i\right)} \\ &= \frac{-2+2\sqrt{3}\,i}{1+3} \\ &= -\frac{1}{2} + i\,\frac{\sqrt{3}}{2} \\ Arg\ z &= \tan^{-1} \frac{\left(\dfrac{\sqrt{3}}{2}\right)}{\left(-\dfrac{1}{2}\right)} \\ &= \tan^{-1}\left(-\sqrt{3}\right) \\ &= \frac{2\pi}{3} \end{aligned}$$

Example 3

Find the modulus and principal value of the arguments of the following complex numbers:

(i) $(4+2i)\left(-3+\sqrt{2}i\right)$

(ii) $\dfrac{2+6\sqrt{3}i}{5+\sqrt{3}i}$

Solution

(i)

$$\begin{aligned}
z &= (4+2i)\left(-3+\sqrt{2}i\right)\\
&= \left(-12+4\sqrt{2}i-6i-2\sqrt{2}\right)\\
&= -\left(12+2\sqrt{2}\right)+i\left(4\sqrt{2}-6\right)\\
|z| &= \sqrt{\left(12+2\sqrt{2}\right)^2+\left(4\sqrt{2}-6\right)^2}\\
&= 2\sqrt{36+2+12\sqrt{2}+8+9-12\sqrt{2}}\\
&= 2\sqrt{55}\\
\arg(z) &= \tan^{-1}\left[\frac{4\sqrt{2}-6}{-\left(12+2\sqrt{2}\right)}\right]\\
&= \tan^{-1}\left(\frac{3-2\sqrt{2}}{6+\sqrt{2}}\right)
\end{aligned}$$

Aliter

$$\begin{aligned}
z &= (4+2i)\left(-3+\sqrt{2}i\right) = z_1 z_2, \text{ say}\\
r &= |z|\\
&= |z_1 z_2|\\
&= |z_1||z_2|\\
&= |4+2i|\left|-3+\sqrt{2}i\right|\\
&= \left(\sqrt{16+4}\right)\left(\sqrt{9+2}\right)\\
&= \sqrt{220}\\
&= 2\sqrt{55}\\
\theta &= \arg(z)\\
&= \arg(z_1 z_2)\\
&= \arg(z_1)+\arg(z_2)\\
&= \arg(4+2i)+\arg\left(-3+\sqrt{2}i\right)
\end{aligned}$$

$$= \tan^{-1}\left(\frac{2}{4}\right) + \tan^{-1}\left(\frac{\sqrt{2}}{-3}\right)$$

$$= \tan^{-1}\left(\frac{1}{2}\right) - \tan^{-1}\left(\frac{\sqrt{2}}{3}\right)$$

$$= \tan^{-1}\left(\frac{\frac{1}{2} - \frac{\sqrt{2}}{3}}{1 + \frac{1}{2} \cdot \frac{\sqrt{2}}{3}}\right)$$

$$= \tan^{-1}\left(\frac{3 - 2\sqrt{2}}{6 + \sqrt{2}}\right)$$

(ii)

$$z = \frac{2 + 6\sqrt{3}i}{5 + \sqrt{3}i}$$

$$= \left(\frac{2 + 6\sqrt{3}i}{5 + \sqrt{3}i}\right)\left(\frac{5 - \sqrt{3}i}{5 - \sqrt{3}i}\right)$$

$$= \frac{28 + 28\sqrt{3}i}{28}$$

$$= 1 + \sqrt{3}i$$

$$|z| = \sqrt{(1)^2 + \left(\sqrt{3}\right)^2} = 2$$

$$\arg(z) = \tan^{-1}\left(\frac{\sqrt{3}}{1}\right) = \frac{\pi}{3}$$

Example 4

Evaluate $\left[(1+i)^{100} + (1-i)^{100}\right]$.

Solution

Let $1 + i = re^{i\theta}$

$$r = |1 + i| = \sqrt{1+1} = \sqrt{2}$$

$$\theta = \tan^{-1}\left(\frac{1}{1}\right) = \tan^{-1} 1 = \frac{\pi}{4}$$

∴ $1 + i = \sqrt{2}e^{i\frac{\pi}{4}}$

and $1 - i = \sqrt{2}e^{-i\frac{\pi}{4}}$

$$\begin{aligned}(1+i)^{100}+(1-i)^{100} &= \left(\sqrt{2}e^{i\frac{\pi}{4}}\right)^{100}+\left(\sqrt{2}e^{-i\frac{\pi}{4}}\right)^{100}\\ &= 2^{50}(e^{25\pi i}+e^{-25\pi i})\\ &= 2^{50}(2\cos 25\pi)\\ &= 2^{51}(-1)\\ &= -2^{51}\end{aligned}$$

Example 5

Express in polar form:

(i) $\left(\frac{2+i}{3-i}\right)^2$ *(ii)* $1+\sin\alpha+i\cos\alpha$

Solution

(i)

$$\begin{aligned}\left(\frac{2+i}{3-i}\right)^2 &= \frac{4+i^2+4i}{9+i^2-6i}\\ &= \frac{3+4i}{8-6i}\\ &= \frac{3+4i}{8-6i}\cdot\frac{8+6i}{8+6i}\\ &= \frac{i}{2}\\ &= 0+i\cdot\frac{1}{2}\end{aligned}$$

Let $$0+i\cdot\frac{1}{2}=r(\cos\theta+i\sin\theta)$$

$$r=\sqrt{0^2+\left(\frac{1}{2}\right)^2}=\frac{1}{2}$$

and $$\theta=\tan^{-1}\left(\frac{\frac{1}{2}}{0}\right)=\tan^{-1}\infty=\frac{\pi}{2}$$

Hence, the polar form is

$$\left(\frac{2+i}{3-i}\right)^2=\frac{1}{2}\left(\cos\frac{\pi}{2}+i\sin\frac{\pi}{2}\right)$$

(ii) $$1+\sin\alpha+i\cos\alpha = 1+\cos\left(\frac{\pi}{2}-\alpha\right)+i\sin\left(\frac{\pi}{2}-\alpha\right)$$

$$= 2\cos^2\left(\frac{\pi}{4}-\frac{\alpha}{2}\right)+2i\sin\left(\frac{\pi}{4}-\frac{\alpha}{2}\right)\cos\left(\frac{\pi}{4}-\frac{\alpha}{2}\right)$$

$$\left[\because\ 1+\cos\theta = 2\cos^2\frac{\theta}{2},\ \sin\theta = 2\sin\frac{\theta}{2}\cos\frac{\theta}{2}\right]$$

$$= 2\cos\left(\frac{\pi}{4}-\frac{\alpha}{2}\right)\left[\cos\left(\frac{\pi}{4}-\frac{\alpha}{2}\right)+i\sin\left(\frac{\pi}{4}-\frac{\alpha}{2}\right)\right]$$

Comparing with the polar form $r(\cos\theta + i\sin\theta)$,

$$r = 2\cos\left(\frac{\pi}{4}-\frac{\alpha}{2}\right)$$

$$\theta = \frac{\pi}{4}-\frac{\alpha}{2}$$

Hence, the polar form is

$$1+\sin\alpha+i\cos\alpha = 2\cos\left(\frac{\pi}{4}-\frac{\alpha}{2}\right)\left[\cos\left(\frac{\pi}{4}-\frac{\alpha}{2}\right)+i\sin\left(\frac{\pi}{4}-\frac{\alpha}{2}\right)\right]$$

Example 6

Find the value of $\sqrt{-5+12i}$.

Solution

Let $$\sqrt{-5+12i} = x+iy$$

$$-5+12i = (x+iy)^2$$

$$-5+12i = (x^2-y^2)+i(2xy)$$

Comparing real and imaginary parts,

$$x^2-y^2 = -5 \qquad \text{... (1)}$$

and $$2xy = 12$$

$$xy = 6$$

Putting $y = \frac{6}{x}$ in Eq. (1),

$$x^2 - \frac{36}{x^2} = -5$$

$$x^4 + 5x^2 - 36 = 0$$

$$(x^2 + 9)(x^2 - 4) = 0$$

$$x^2 = -9, x^2 = 4$$

Since x is real, $x = \pm 2$

and $y = \pm 3$

Hence, $\sqrt{-5+12i} = \pm 2 \pm 3i$

Example 7

If x and y are real, solve the equation $\dfrac{iy}{ix+1} - \dfrac{3y+4i}{3x+y} = 0$.

Solution

$$\frac{iy}{ix+1} - \frac{3y+4i}{3x+y} = 0$$

$$\frac{iy(3x+y)-(3y+4i)(ix+1)}{(ix+1)(3x+y)} = 0$$

$$(-3y+4x)+i(3xy+y^2-3xy-4) = 0+i0$$

$$(-3y+4x)+i(y^2-4) = 0+i0$$

Comparing real and imaginary parts,

$$-3y + 4x = 0$$

and $y^2 - 4 = 0$

$$y = \pm 2$$

$\therefore$ $x = \pm \dfrac{3}{2}$

Hence, the solution is

$$x = \pm \frac{3}{2}, \qquad y = \pm 2$$

Example 8

Solve the equation $z^2 + (2i - 3)z + 5 - i = 0$. **[Summer 2015]**

Solution

$$z^2 + (2i - 3)z + 5 - i = 0$$

27.04.2020

Pág. 616(626)

8.4 The elliptic curve.

C) Books_72_2020/
Complex Variables and Numerical Methods/
Ravish R Singh_Complex Variables.pdf

Pág. 20

1.3 Geometrical Representation of Complex Numbers (Argand's Diagram)

Obs: Ver livro todo!

D) Books_72_2020/
Nonlinear Optics, 4th Edition/
Nonlinear Optics, 4th ed.pdf

Pág. 11(34)

Self-Focusing

Look: Figure 1.2.6.

E) Books_72_2020/
Fomingeo_fau_edu/
Tratado0508.pdf

Pág. 15 Bissectrizes dos ângulos de um triângulo.

1.2.6 Third-Order Nonlinear Optical Processes

We next consider the third-order contribution to the nonlinear polarization

$$\tilde{P}^{(3)}(t) = \epsilon_0 \chi^{(3)} \tilde{E}(t)^3. \tag{1.2.14}$$

For the general case in which the field $\tilde{E}(t)$ is made up of several different frequency components, the expression for $\tilde{P}^{(3)}(t)$ is very complicated. For this reason, we first consider the simple case in which the applied field is monochromatic and is given by*

$$\tilde{E}(t) = \mathcal{E} \cos \omega t. \tag{1.2.15}$$

Then, through use of the identity $\cos^3 \omega t = \frac{1}{4}\cos 3\omega t + \frac{3}{4}\cos \omega t$, we can express the nonlinear polarization as

$$\tilde{P}^{(3)}(t) = \frac{1}{4}\epsilon_0 \chi^{(3)} \mathcal{E}^3 \cos 3\omega t + \frac{3}{4}\epsilon_0 \chi^{(3)} \mathcal{E}^3 \cos \omega t. \tag{1.2.16}$$

The significance of each of the two terms in this expression is described briefly below.

1.2.7 Third-Harmonic Generation

The first term in Eq. (1.2.16) describes a response at frequency 3ω that is created by an applied field at frequency ω. This term leads to the process of third-harmonic generation, which is illustrated in Fig. 1.2.5. According to the photon description of this process, shown in part (b) of the figure, three photons of frequency ω are destroyed and one photon of frequency 3ω is created in this process.

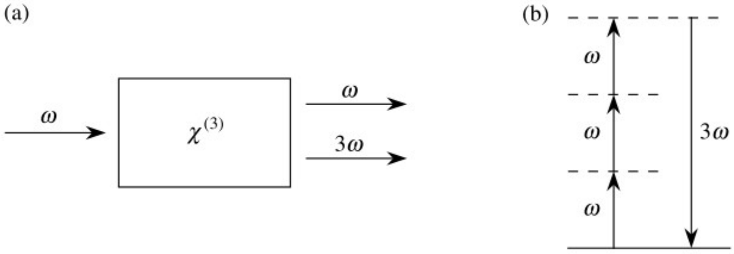

FIGURE 1.2.5: Third-harmonic generation. (a) Geometry of the interaction. (b) Energy-level description.

---

* Here we are describing the field in terms of purely real quantities for the pedagogical value of introducing such a notation.

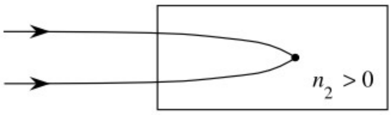

FIGURE 1.2.6: Self-focusing of light.

1.2.8 Intensity-Dependent Refractive Index

The second term in Eq. (1.2.16) describes a nonlinear contribution to the polarization at the same frequency as that of the incident field; this term hence leads to a nonlinear contribution to the refractive index experienced by a wave at frequency ω. We shall see in Section 4.1 that the refractive index in the presence of this type of nonlinearity can be represented as

$$n = n_0 + n_2 I, \tag{1.2.17a}$$

where n_0 is the usual (i.e., linear or low-intensity) refractive index, where

$$n_2 = \frac{3}{4n_0^2\epsilon_0 c}\chi^{(3)} \tag{1.2.17b}$$

is an optical constant that characterizes the strength of the optical nonlinearity, and where $I = \frac{1}{2}n_0\epsilon_0 c\mathcal{E}^2$ is the intensity of the incident wave.

Self-Focusing

One of the processes that can occur as a result of the intensity-dependent refractive index is self-focusing, which is illustrated in Fig. 1.2.6. This process can occur when a beam of light having a nonuniform transverse intensity distribution propagates through a material for which n_2 is positive. Under these conditions, the material effectively acts as a positive lens, which causes the rays to curve toward each other. This process is of great practical importance because the intensity at the focal spot of the self-focused beam is usually sufficiently large to lead to optical damage of the material. The process of self-focusing is described in greater detail in Section 7.1.

1.2.9 Third-Order Interactions (General Case)

Let us next examine the form of the nonlinear polarization

$$\tilde{P}^{(3)}(t) = \epsilon_0\chi^{(3)}\tilde{E}(t)^3 \tag{1.2.18a}$$

14.1.05

Básico - Construção de Triângulos (III)

Desafio:
Reconstruir um triângulo de que se conhece um vértice e as rectas que contêm as suas bissectrizes.

Sugestão.
Desenhe um triângulo e as suas bissectrizes (que se intersectam no incentro)dos ângulos internos. Pense em usar a perpendicular à bissectriz tirada pelo vértice do ângulo. Pode confirmar o que se pretende com a nossa construção:

Sugestão: resultado com bissectrizes de um triângulo

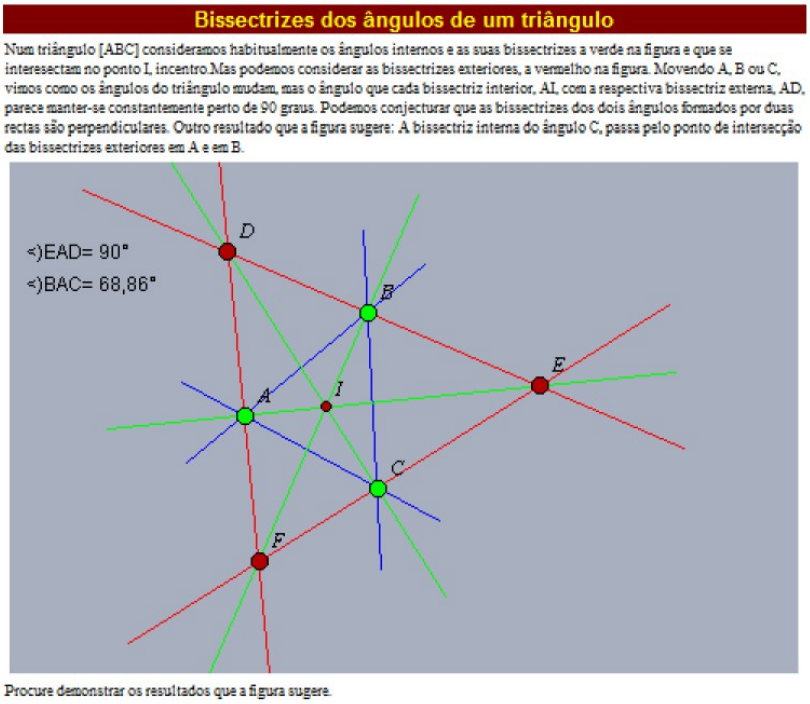

Bissectrizes dos ângulos de um triângulo

Num triângulo [ABC] consideramos habitualmente os ângulos internos e as suas bissectrizes a verde na figura e que se interesectam no ponto I, incentro.Mas podemos considerar as bissectrizes exteriores, a vermelho na figura. Movendo A, B ou C, vimos como os ângulos do triângulo mudam, mas o ângulo que cada bissectriz interior, AI, com a respectiva bissectriz externa, AD, parece manter-se constantemente perto de 90 graus. Podemos conjecturar que as bissectrizes dos dois ângulos formados por duas rectas são perpendiculares. Outro resultado que a figura sugere: A bissectriz interna do ângulo C, passa pelo ponto de intersecção das bissectrizes exteriores em A e em B.

Procure demonstrar os resultados que a figura sugere.

Resposta ao desafio:
Se não tiver reconstruído o triângulo, pode ver a nossa construção com Cinderella.

Construção dinâmica de um triângulo dados um vértice e as três bissectrizes

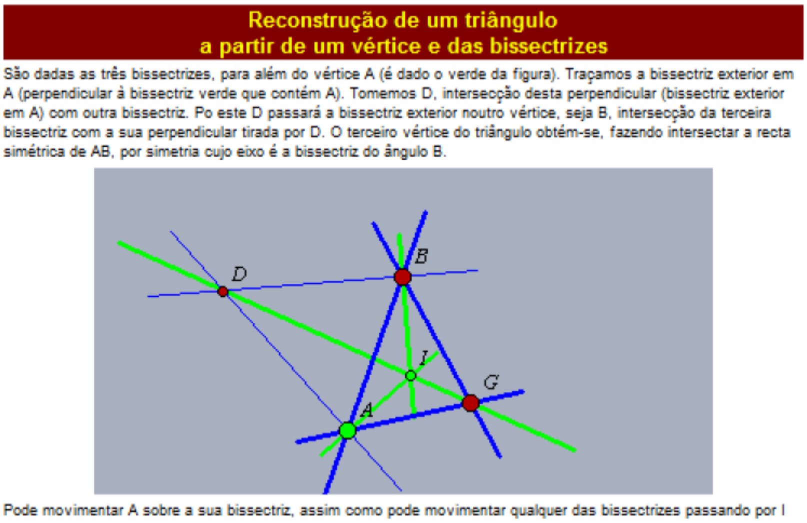

Reconstrução de um triângulo
a partir de um vértice e das bissectrizes

São dadas as três bissectrizes, para além do vértice A (é dado o verde da figura). Traçamos a bissectriz exterior em A (perpendicular à bissectriz verde que contém A). Tomemos D, intersecção desta perpendicular (bissectriz exterior em A) com outra bissectriz. Po este D passará a bissectriz exterior noutro vértice, seja B, intersecção da terceira bissectriz com a sua perpendicular tirada por D. O terceiro vértice do triângulo obtém-se, fazendo intersectar a recta simétrica de AB, por simetria cujo eixo é a bissectriz do ângulo B.

Pode movimentar A sobre a sua bissectriz, assim como pode movimentar qualquer das bissectrizes passando por I

27.04.2020
X

Pág. 616(626)

8.4 The elliptic curve.

C) Books_72_2020/
Complex Variables and Numerical Methods/
Ravish R Singh_Complex Variables.pdf

Pág. 20

1.3 Geometrical Representation of Complex Numbers (Argand's Diagram)

Obs: Ver livro todo!

D) Books_72_2020/
Nonlinear Optics, 4th Edition/
Nonlinear Optics, 4th ed.pdf

Pág. 11(34)

Self-Focusing

Look: Figure 1.2.6.

E) Books_72_2020/
Fomingeo_fau_edu/
Tratado0508.pdf

Pág. 15 Bissectrizes dos ângulos de um triângulo.

27.04.2020
X

F) Books_72_2020/
Forumgeo_fau_edu/
FG2004Volume4.pdf

Pag. 5 Antiparallels and Concurrent Euler Lines.

Figure 1 Thébault's theorem on the ~~[illegible]~~ concurrency of Euler lines.

Lost: Figure 2
Figure 3
Q(K) and the second Lemoine circle.

Obs: Ler o artigo todo. (Pag. 24

153) 20200427

A) Books_72_2020/
Projective_Geometry/
Geometry of Cuts and Metrics.pdf

Pág. 86(96

6.4 An Example: the Spherical Distance Space

$x, y \in S$

'()*' ' '+ ' ,0 $1 $ -!
& (- ! 7 9

Forum Geometricorum
Volume 4 (2004) 1–20.

FORUM GEOM
ISSN 1534-1178

Antiparallels and Concurrent Euler Lines

Nikolaos Dergiades and Paul Yiu

Abstract. We study the condition for concurrency of the Euler lines of the three triangles each bounded by two sides of a reference triangle and an antiparallel to the third side. For example, if the antiparallels are concurrent at P and the three Euler lines are concurrent at Q, then the loci of P and Q are respectively the tangent to the Jerabek hyperbola at the Lemoine point, and the line parallel to the Brocard axis through the inverse of the deLongchamps point in the circumcircle. We also obtain an interesting cubic as the locus of the point P for which the three Euler lines are concurrent when the antiparallels are constructed through the vertices of the cevian triangle of P.

1. Thébault's theorem on Euler lines

We begin with the following theorem of Victor Thébault [8] on the concurrency of three Euler lines.

Theorem 1 (Thébault)**.** *Let $A'B'C'$ be the orthic triangle of ABC. The Euler lines of the triangles $AB'C'$, $BC'A'$, $CA'B'$ are concurrent at the Jerabek center.*[1]

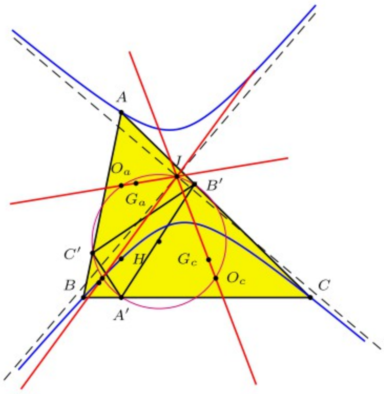

Figure 1. Thébault's theorem on the concurrency of Euler lines

We shall make use of homogeneous barycentric coordinates. With reference to triangle ABC, the vertices of the orthic triangle are the points

$$A' = (0 : S_C : S_B), \qquad B' = (S_C : 0 : S_A), \qquad C' = (S_B : S_A : 0).$$

Publication Date: February 3, 2004. Communicating Editor: Antreas P. Hatzipolakis.

[1] Thébault [8] gave an equivalent characterization of this common point. See also [7].

These are the traces of the orthocenter $H = (S_{BC} : S_{CA} : S_{AB})$.

The centroid of $AB'C'$ is the point

$$(S_{AA} + 2S_{AB} + 2S_{AC} + 3S_{BC} : S_A(S_C + S_A) : S_A(S_A + S_B)).$$

The circumcenter of $A'BC$, being the midpoint of AH, has coordinates

$$(S_{CA} + S_{AB} + 2S_{BC} : S_{AC} : S_{AB}).$$

It is straightforward to verify that these two points lie on the line

$$S_{AA}(S_B - S_C)(x+y+z) = (S_A+S_B)(S_{AB}+S_{BC}-2S_{CA})y - (S_C+S_A)(S_{BC}+S_{CA}-2S_{AB})z, \tag{1}$$

which is therefore the Euler line of triangle $AB'C'$. Furthermore, the line (1) also contains the point

$$J = (S_A(S_B - S_C)^2 : S_B(S_C - S_A)^2 : S_C(S_A - S_B)^2),$$

which is the center of the Jerabek hyperbola.[2] Similar reasoning gives the equations of the Euler lines of triangles $BC'A'$ and $A'B'C$, and shows that these contain the same point J. This completes the proof of Thébault's theorem.

2. Triangles intercepted by antiparallels

Since the sides of the orthic triangles are antiparallel to the respective sides of triangle ABC, we consider the more general situation when the residuals of the orthic triangle are replaced by triangles intercepted by lines ℓ_1, ℓ_2, ℓ_3 antiparallel to the sidelines of the reference triangle, with the following intercepts on the sidelines

	BC	CA	AB
ℓ_1		B_a	C_a
ℓ_2	A_b		C_b
ℓ_3	A_c	B_c	

These lines are parallel to the sidelines of the orthic triangle $A'B'C'$. We shall assume that they are the images of the lines $B'C'$, $C'A'$, $A'B'$ under the homotheties $\mathsf{h}(A, 1-t_1)$, $\mathsf{h}(B, 1-t_2)$, and $\mathsf{h}(C, 1-t_3)$ respectively. The points B_a, C_a etc. have homogeneous barycentric coordinates

$$\begin{aligned} B_a &= (t_1S_A + S_C : 0 : (1-t_1)S_A), & C_a &= (t_1S_A + S_B : (1-t_1)S_A : 0), \\ C_b &= ((1-t_2)S_B : t_2S_B + S_A : 0), & A_b &= (0 : t_2S_B + S_C : (1-t_2)S_B), \\ A_c &= (0 : (1-t_3)S_C : t_3S_C + S_B), & B_c &= ((1-t_3)S_C : 0 : t_3S_C + S_A). \end{aligned}$$

---

[2] The point J appears as X_{125} in [4].

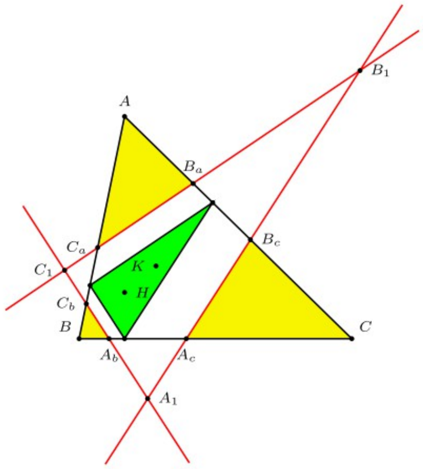

Figure 2. Triangles intercepted by antiparallels

2.1. *The Euler lines $\mathcal{L}_i$, $i = 1, 2, 3$.* Denote by $\mathbf{T}_1$ the triangle AB_aC_a intercepted by ℓ_1; similarly $\mathbf{T}_2$ and $\mathbf{T}_3$. These are oppositely similar to ABC. We shall study the condition of the concurrency of their Euler lines.

Proposition 2. *With reference to triangle ABC, the barycentric equations of the Euler lines of $\mathbf{T}_i$, $i = 1, 2, 3$, are*

$$\begin{aligned}(1-t_1)S_{AA}(S_B-S_C)(x+y+z) &= c^2(S_{AB}+S_{BC}-2S_{CA})y - b^2(S_{BC}+S_{CA}-2S_{AB})z,\\(1-t_2)S_{BB}(S_C-S_A)(x+y+z) &= a^2(S_{BC}+S_{CA}-2S_{AB})z - c^2(S_{CA}+S_{AB}-2S_{BC})x,\\(1-t_3)S_{CC}(S_A-S_B)(x+y+z) &= b^2(S_{CA}+S_{AB}-2S_{BC})x - a^2(S_{AB}+S_{BC}-2S_{CA})y.\end{aligned}$$

Proof. It is enough to establish the equation of the Euler line $\mathcal{L}_1$ of $\mathbf{T}_1$. This is the image of the Euler line $\mathcal{L}_1'$ of triangle $AB'C'$ under the homothety $\mathsf{h}(A, 1-t_1)$. A point $(x : y : z)$ on $\mathcal{L}_1$ corresponds to the point $((1-t_1)x - t_1(y+z) : y : z)$ on $\mathcal{L}_1'$. The equation of $\mathcal{L}_1$ can now by obtained from (1). □

From the equations of these Euler lines, we easily obtain the condition for their concurrency.

Theorem 3. *The three Euler lines $\mathcal{L}_i$, $i = 1, 2, 3$, are concurrent if and only if*

$$t_1a^2(S_B-S_C)S_{AA} + t_2b^2(S_C-S_A)S_{BB} + t_3c^2(S_A-S_B)S_{CC} = 0. \qquad (2)$$

Proof. From the equations of $\mathcal{L}_i$, $i = 1, 2, 3$, given in Proposition 2, it is clear that the condition for concurrency is

$$(1-t_1)a^2(S_B-S_C)S_{AA} + (1-t_2)b^2(S_C-S_A)S_{BB} + (1-t_3)c^2(S_A-S_B)S_{CC} = 0.$$

This simplifies into (2) above. □

2.2. *Antiparallels with given common point of $\mathcal{L}_i$, $i = 1, 2, 3$.* We shall assume triangle ABC scalene, *i.e.*, its angles are unequal and none of them is a right angle. For such triangles, the Euler lines of the residuals of the orthic triangle and the corresponding altitudes intersect at finite points.

Theorem 4. *Given a point Q in the plane of a scalene triangle ABC, there is a unqiue triple of antiparallels ℓ_i, $i = 1, 2, 3$, for which the Euler lines $\mathcal{L}_i$, $i = 1, 2, 3$, are concurrent at Q.*

Proof. Construct the parallel through Q to the Euler line of $AB'C'$ to intersect the line AH at O_a. The circle through A with center O_a intersects AC and AB at B_a and C_a respectively. The line B_aC_a is parallel to $B'C'$. It follows that its Euler line is parallel to that of $AB'C'$. This is the line O_aQ. Similar constructions give the other two antiparallels with corresponding Euler lines passing through Q. □

We make a useful observation here. From the equations of the Euler lines given in Proposition 2 above, the intersection of any two of them have coordinates expressible in linear functions of t_1, t_2, t_3. It follows that if t_1, t_2, t_3 are linear functions of a parameter t, and the three Euler lines are concurrent, then as t varies, the common point traverses a straight line. In particular, $t_1 = t_2 = t_3 = t$, the Euler lines are concurrent by Theorem 3. The locus of the intersection of the Euler lines is a straight line. Since this intersection is the Jerabek center when $t = 0$ (Thébault's theorem), and the orthocenter when $t = -1$,[3] this is the line

$$\mathcal{L}_\text{c} : \qquad \sum_{\text{cyclic}} S_{AA}(S_B - S_C)(S_{CA} + S_{AB} - 2S_{BC})x = 0.$$

We give a summary of some of the interesting loci of common points of Euler lines $\mathcal{L}_i$, $i = 1, 2, 3$, when the lines ℓ_i, $i = 1, 2, 3$, are subjected to some further conditions. In what follows, $\mathbf{T}$ denotes the triangle bounded by the lines ℓ_i, $i = 1, 2, 3$.

Line	Construction	Condition	Reference
$\mathcal{L}_\text{c}$	HJ	$\mathbf{T}$ homothetic to orthic triangle at X_{25}	
$\mathcal{L}_\text{q}$	Remark below	ℓ_i, $i = 1, 2, 3$, concurrent	§3.2
$\mathcal{L}_\text{t}$	KX_{74}	ℓ_i are the antiparallels of a Tucker hexagon	§6
$\mathcal{L}_\text{f}$	X_5X_{184}	$\mathcal{L}_i$ intersect on Euler line of $\mathbf{T}$	§7.2
$\mathcal{L}_\text{r}$	GX_{110}	$\mathbf{T}$ and ABC perspective	§8.3

Remark. $\mathcal{L}_\text{q}$ can be constructed as the line parallel to the Brocard axis through the intersection of the inverse of the deLongchamps point in the circumcircle.

---

[3] For $t = 1$, this intersection is the point X_{74} on the circumcircle, the isogonal conjugate of the infinite point of the Euler line.

3. Concurrent antiparallels

In this section we consider the case when the antiparallels ℓ_1, ℓ_2, ℓ_3 all pass through a point $P = (u : v : w)$. In this case,

$$\begin{aligned}
B_a &= ((S_C + S_A)u - (S_B - S_C)v : 0 : (S_A + S_B)v + (S_C + S_A)w),\\
C_a &= ((S_A + S_B)u + (S_B - S_C)w : (S_A + S_B)v + (S_C + S_A)w : 0),\\
C_b &= ((S_B + S_C)w + (S_A + S_B)u : (S_A + S_B)v - (S_C - S_A)w : 0),\\
A_b &= (0 : (S_B + S_C)v + (S_C - S_A)u : (S_B + S_C)w + (S_A + S_B)u),\\
A_c &= (0 : (S_C + S_A)u + (S_B + S_C)v : (S_B + S_C)w - (S_A - S_B)u),\\
B_c &= ((S_C + S_A)u + (S_B + S_C)v : 0 : (S_C + S_A)w + (S_A - S_B)v).
\end{aligned}$$

For example, when $P = K$, these are the vertices of the second cosine circle.

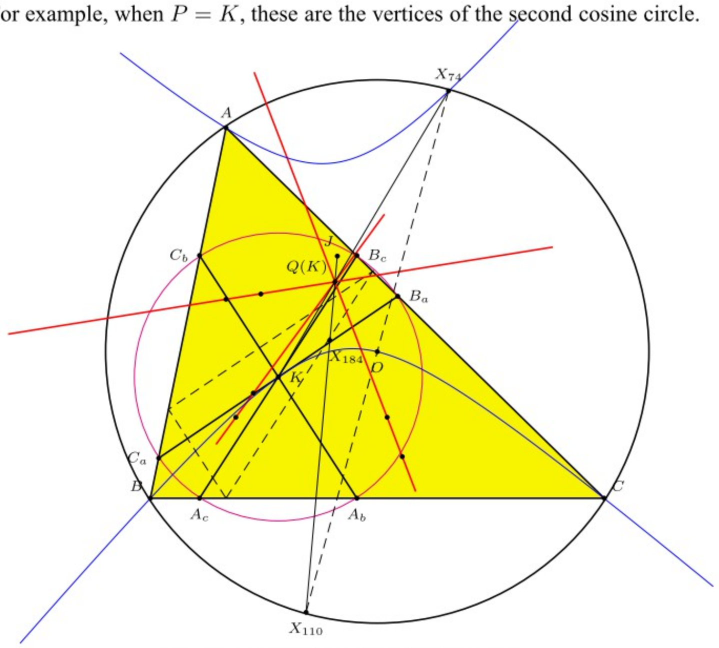

Figure 3. $Q(K)$ and the second Lemoine circle

Proposition 5. *The Euler lines of triangles* $\mathbf{T}_i$, $i = 1, 2, 3$, *are concurrent if and only if* P *lies on the line*

$$\mathcal{L}_{\rm p}: \qquad \frac{S_A(S_B - S_C)}{a^2}x + \frac{S_B(S_C - S_A)}{b^2}y + \frac{S_C(S_A - S_B)}{c^2}z = 0.$$

When P *traverses* $\mathcal{L}_{\rm p}$, *the intersection* Q *of the Euler lines traverses the line*

$$\mathcal{L}_{\rm q}: \qquad \sum_{\text{cyclic}} \frac{(b^2 - c^2)(a^2(S_{AA} + S_{BC}) - 4S_{ABC})}{a^2}x = 0.$$

For a point P on the line $\mathcal{L}_{\rm p}$, we denote by $Q(P)$ the corresponding point on $\mathcal{L}_{\rm q}$.

Proposition 6. *For points* P_1, P_2, P_3 *on* $\mathcal{L}_{\rm p}$, $Q(P_1)$, $Q(P_2)$, $Q(P_3)$ *are points on* $\mathcal{L}_{\rm q}$ *satisfying*

$$Q(P_1)Q(P_2) : Q(P_2)Q(P_3) = P_1P_2 : P_2P_3.$$

3.1. *The line* $\mathcal{L}_{\rm p}$. The line $\mathcal{L}_{\rm p}$ contains K and is the tangent to the Jerabek hyperbola at K. See Figure 4. It also contains, among others, the following points.

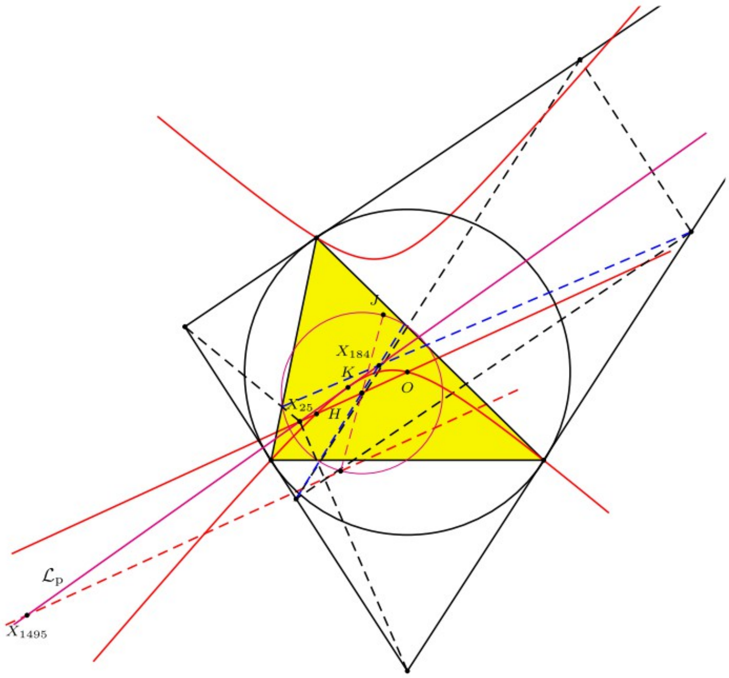

Figure 4. The line $\mathcal{L}_{\rm p}$

(1) $X_{25} = \left(\frac{a^2}{S_A} : \frac{b^2}{S_B} : \frac{c^2}{S_C}\right)$ which is on the Euler line of ABC, and is the homothetic center of the orthic and the tangential triangles,[4]

(2) $X_{184} = (a^4S_A : b^4S_B : c^4S_C)$ which is the homothetic center of the orthic triangle and the medial tangential triangle,[5]

---

[4] See also §4.1.

[5] For other interesting properties of X_{184}, see [6], where it is named the procircumcenter of triangle ABC.

(3) $X_{1495} = (a^2(S_{CA}+S_{AB}-2S_{BC}) : \cdots : \cdots)$ which lies on the parallel to the Euler line through the antipode of the Jerabek center on the nine-point circle. [6]

3.2. *The line* $\mathcal{L}_{\mathrm{q}}$. The line $\mathcal{L}_{\mathrm{q}}$ is parallel to the Brocard axis. See Figure 5. It contains the following points.

(1) $Q(K) = (a^2S_A(b^2c^2(S_{BB}-S_{BC}+S_{CC})-2a^2S_{ABC}) : \cdots : \cdots)$. It can be constructed as the intersection of the lines joining K to X_{74}, and J to X_{110}. See Figure 3 and §6 below. The line $\mathcal{L}_{\mathrm{q}}$ can therefore be constructed as the parallel through this point to the Brocard axis.

(2) $Q(X_{1495}) = (a^2S_A(a^2S^2 - 6S_{ABC}) : \cdots : \cdots)$, which is on the line joining O to X_{184} (on $\mathcal{L}_{\mathrm{p}}$).

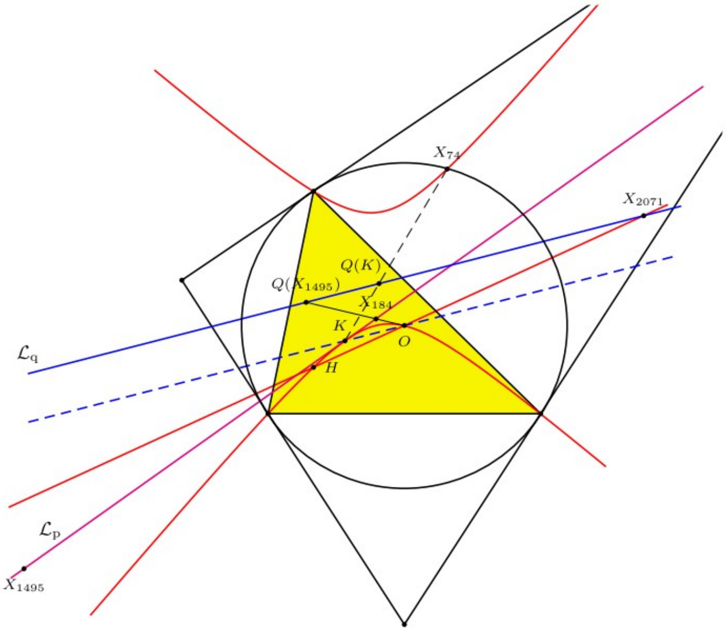

Figure 5. The line $\mathcal{L}_{\mathrm{q}}$

The line $\mathcal{L}_{\mathrm{q}}$ intersects the Euler line of ABC at the point

$$X_{2071} = (a^2(a^2S_{AAA} + S_{AA}(S_{BB} - 3S_{BC} + S_{CC}) - S_{BBCC}) : \cdots : \cdots),$$

---

[6] This is the point X_{113}.

which is the inverse of the de Longchamps point in the circumcircle. This corresponds to the antiparallels through

$$P_{2071}=(a^4((a^2S_{AAA}+S_{AA}(S_{BB}-3S_{BC}+S_{CC})-S_{BBCC}):\cdots:\cdots)$$

on the line $\mathcal{L}_{\text{p}}$. This point can be constructed by a simple application of Theorem 4 or Proposition 6. (See also Remark 2 following Theorem 12).

3.3. *The intersection of* $\mathcal{L}_{\text{p}}$ *and* $\mathcal{L}_{\text{q}}$. The lines $\mathcal{L}_{\text{p}}$ and $\mathcal{L}_{\text{q}}$ intersect at the point

$$M=(a^2S_A(S_{AB}+S_{AC}+S_{BB}-4S_{BC}+S_{CC}):\cdots:\cdots).$$

(1) $Q(M)$ is the point on $\mathcal{L}_{\text{q}}$ with coordinates

$$(a^2S_A(S_{AA}(S_{BB}+S_{CC})+a^2S_A(S_{BB}-3S_{BC}+S_{CC})+S_{BC}(S_B-S_C)^2):\cdots:\cdots).$$

(2) The point P on $\mathcal{L}_{\text{p}}$ for which $Q(P)=M$ has coordinates

$$(a^2(a^2(2S_{AA}-S_{BC})+2S_A(S_{BB}-3S_{BC}+S_{CC})):\cdots:\cdots).$$

4. The triangle $\mathbf{T}$ bounded by the antiparallels

We assume the line ℓ_i, $i=1,2,3$, nonconcurrent so that they bound a nondegenerate triangle $\mathbf{T}=A_1B_1C_1$. Since these lines have equations

$$\begin{aligned}
-t_1S_A(x+y+z) &= -S_Ax+S_By+S_Cz,\\
-t_2S_B(x+y+z) &= \ \ S_Ax-S_By+S_Cz,\\
-t_3S_C(x+y+z) &= \ \ S_Ax+S_By-S_Cz,
\end{aligned}$$

the vertices of $\mathbf{T}$ are the points

$$\begin{aligned}
A_1 =&(-a^2(t_2S_B+t_3S_C):2S_{CA}+t_2b^2S_B+t_3S_C(S_C-S_A)\\
&:2S_{AB}+t_2S_B(S_B-S_A)+t_3c^2S_C),\\
B_1 =&(2S_{BC}+t_3S_C(S_C-S_B)+t_1a^2S_A:-b^2(t_3S_C+t_1S_A)\\
&:2S_{AB}+t_3c^2S_C+t_1S_A(S_A-S_B))\\
C_1 =&(2S_{BC}+t_1a^2S_A+t_2S_B(S_B-S_C):2S_{CA}+t_1S_A(S_A-S_C)+t_2b^2S_B\\
&:-c^2(t_1S_A+t_2S_B)).
\end{aligned}$$

4.1. *Homothety with the orthic triangle* . The triangle $\mathbf{T}=A_1B_1C_1$ is homothetic to the orthic triangle $A'B'C'$. The center of homothety is the point

$$P(\mathbf{T})=\left(\frac{t_2S_B+t_3S_C}{S_A}:\frac{t_3S_C+t_1S_A}{S_B}:\frac{t_1S_A+t_2S_B}{S_C}\right), \tag{3}$$

and the ratio of homothety is

$$1+\frac{t_1a^2S_{AA}+t_2b^2S_{BB}+t_3c^2S_{CC}}{2S_{ABC}}.$$

Proposition 7. *If the Euler lines* $\mathcal{L}_i$, $i=1,2,3$, *are concurrent, the homothetic center* $P(\mathbf{T})$ *of* $\mathbf{T}$ *and the orthic triangle lies on the line* $\mathcal{L}_{\text{p}}$.

Proof. If we write $P(\mathbf{T}) = (x : y : z)$. From (3), we obtain

$$t_1 = \frac{-xS_A + yS_B + zS_C}{2S_A}, \quad t_2 = \frac{-yS_B + zS_C + xS_A}{2S_B}, \quad t_3 = \frac{-zS_C + xS_A + yS_B}{2S_C}.$$

Substitution in (2) yields the equation of the line $\mathcal{L}_{\text{p}}$. □

For example, if $t_1 = t_2 = t_3 = t$, $P(\mathbf{T}) = X_{25} = \left(\frac{a^2}{S_A} : \frac{b^2}{S_B} : \frac{c^2}{S_C}\right)$.[7] If the ratio of homothety is 0, triangle $\mathbf{T}$ degenerates into the point X_{25} on $\mathcal{L}_{\text{p}}$. The intersection of $\mathcal{L}_{\text{c}}$ and $\mathcal{L}_{\text{q}}$ is the point

$$\begin{aligned} Q(X_{25}) =& (a^2S_A(b^4S_B^4 + c^4S_C^4 + a^2S_{AAA}(S_B - S_C)^2 \\ & - S_{ABC}(4a^2S_{BC} + 3S_A(S_B - S_C)^2)) : \cdots : \cdots). \end{aligned}$$

Remark. The line $\mathcal{L}_{\text{p}}$ is also the locus of the centroid of $\mathbf{T}$ for which the Euler lines $\mathcal{L}_i$, $i = 1, 2, 3$, concur.

4.2. *Common point of $\mathcal{L}_i$, $i = 1, 2, 3$, on the Brocard axis.* We consider the case when the Euler lines $\mathcal{L}_i$, $i = 1, 2, 3$, intersect on the Brocard axis. A typical point on the Brocard axis, dividing the segment OK in the ratio $t : 1 - t$, has coordinates

$$(a^2(S_A(S_A + S_B + S_C) + (S_{BC} - S_{AA})t) : \cdots : \cdots).$$

This point lies on the Euler lines $\mathcal{L}_i$, $i = 1, 2, 3$, if and only if we choose

$$\begin{aligned} t_1 =& \frac{-(S_A + S_B + S_C)(S^2 - S_{AA}) + b^2c^2(S_B + S_C - 2S_A)t}{2S_{AA}(S_A + S_B + S_C)}, \\ t_2 =& \frac{-(S_A + S_B + S_C)(S^2 - S_{BB}) + c^2a^2(S_C + S_A - 2S_B)t}{2S_{BB}(S_A + S_B + S_C)}, \\ t_3 =& \frac{-(S_A + S_B + S_C)(S^2 - S_{CC}) + a^2b^2(S_A + S_B - 2S_C)t}{2S_{CC}(S_A + S_B + S_C)}. \end{aligned}$$

The corresponding triangle $\mathbf{T}$ is homothetic to the orthic triangle at the point

$$(a^2(-(S_A + S_B + S_C) \cdot a^2S_A + t(-(2S_A + S_B + S_C)S_{BC} + b^2S_{CA} + c^2S_{AB}) : \cdots : \cdots),$$

which divides the segment $X_{184}K$ in the ratio $2t : 1 - 2t$. The ratio of homothety is $-\frac{a^2b^2c^2}{4S_{ABC}}$. These triangles are all directly congruent to the medial tangential triangle of ABC. We summarize this in the following proposition.

Proposition 8. *Corresponding to the family of triangles directly congruent to the medial tangential triangle, homothetic to orthic triangle at points on the line $\mathcal{L}_{\text{p}}$, the common points of the Euler lines of $\mathcal{L}_i$, $i = 1, 2, 3$, all lie on the Brocard axis.*

---

[7] See also §3.1(1). The tangential triangle is $\mathbf{T}$ with $t = 1$.

5. Perspectivity of $\mathbf{T}$ with ABC

Proposition 9. *The triangles* $\mathbf{T}$ *and* ABC *are perspective if and only if*

$$\sum_{\text{cyclic}}(S_B - S_C)(t_1S_{AA} - t_2t_3S_{BC}) = 0. \tag{4}$$

Proof. From the coordinates of the vertices of $\mathbf{T}$, it is straightforward to check that $\mathbf{T}$ and ABC are perspective if and only if

$$t_1a^2S_{AA} + t_2b^2S_{BB} + t_3c^2S_{CC} + 2S_{ABC} = 0$$

or (4) holds. Since the area of triangle $\mathbf{T}$ is

$$\frac{(t_1a^2S_{AA} + t_2b^2S_{BB} + t_3c^2S_{CC} + 2S_{ABC})^2}{a^2b^2c^2S_{ABC}}$$

times that of triangle ABC, we assume $t_1a^2S_{AA}+t_2b^2S_{BB}+t_3c^2S_{CC}+2S_{ABC} \neq 0$ and (4) is the necessary and sufficient condition for perspectivity. $\square$

Theorem 10. *If the triangle* $\mathbf{T}$ *is nondegenerate and is perspective to* ABC*, then the perspector lies on the Jerabek hyperbola of* ABC*.*

Proof. If triangles $A_1B_1C_1$ and ABC are perspective at $P = (x : y : z)$, then

$$A_1 = (u + x : y : z), \quad B_1 = (x : v + y : z), \quad C_1 = (x : y : w + z)$$

for some u, v, w. Since the line B_1C_1 is parallel to $B'C'$, which has infinite point $(S_B - S_C : -(S_C + S_A) : S_A + S_B)$, we have

$$\begin{vmatrix} S_B - S_C & -(S_C + S_A) & S_A + S_B \\ x & y + v & z \\ x & y & z + w \end{vmatrix} = 0,$$

and similarly for the other two lines. These can be rearranged as

$$\begin{aligned}
\frac{(S_C + S_A)x - (S_B - S_C)y}{v} - \frac{(S_B - S_C)z + (S_A + S_B)x}{w} &= S_B - S_C, \\
\frac{(S_A + S_B)y - (S_C - S_A)z}{w} - \frac{(S_C - S_A)x + (S_B + S_C)y}{u} &= S_C - S_A, \\
\frac{(S_B + S_C)z - (S_A - S_B)x}{u} - \frac{(S_A - S_B)y + (S_C + S_A)z}{v} &= S_A - S_B.
\end{aligned}$$

Multiplying these equations respectively by

$$S_A(S_B + S_C)yz, \qquad S_B(S_C + S_A)zx, \qquad S_C(S_A + S_B)xy$$

and adding up, we obtain

$$\left(1 + \frac{x}{u} + \frac{y}{v} + \frac{z}{w}\right)\sum_{\text{cyclic}} S_A(S_{BB} - S_{CC})yz = 0.$$

Since the area of triangle $\mathbf{T}$ is

$$uvw\left(1 + \frac{x}{u} + \frac{y}{v} + \frac{z}{w}\right)$$

times that of triangle ABC, we must have $1+\frac{x}{u}+\frac{y}{v}+\frac{z}{w}\neq 0$. It follows that

$$\sum_{\text{cyclic}} S_A(S_{BB}-S_{CC})yz=0.$$

This means that P lies on the Jerabek hyperbola. □

We shall identify the locus of the common points of Euler lines in §8.3 below. In the meantime, we give a construction for the point Q from the perspector on the Jerabek hypebola.

Construction. Given a point P on the Jerabek hyperbola, construct parallels to $A'B'$ and $A'C'$ through an arbitrary point A_1' on the line AP. Let M_1 be the intersection of the Euler lines of the triangles formed by these antiparallels and the sidelines of ABC. With another point A_1'' obtain a point M_2 by the same construction. Similarly, working with two points B_1' and B_1'' on BP, we construct another line M_3M_4. The intersection of M_1M_2 and M_3M_4 is the common point Q of the Euler lines corresponding the antiparallels that bound a triangle perspective to ABC at P.

6. The Tucker hexagons and the line $\mathcal{L}_t$

It is well known that if the antiparallels, together with the sidelines of triangle ABC, bound a Tucker hexagon, the vertices lie on a circle whose center is on the Brocard axis. If this center divides the segment OK in the ratio $t:1-t$, the antiparallels pass through the points dividing the symmedians in the same ratio. The vertices of the Tucker hexagon are

$$\begin{aligned} B_a &= (S_C+(1-t)c^2:0:tc^2), & C_a &= (S_B+(1-t)b^2:tb^2:0),\\ C_b &= (ta^2:S_A+(1-t)a^2:0), & A_b &= (0:S_C+(1-t)c^2:tc^2),\\ A_c &= (0:tb^2:S_B+(1-t)b^2), & B_c &= (ta^2:0:S_A+(1-t)a^2). \end{aligned}$$

In this case,

$$1-t_1=\frac{t\cdot b^2c^2}{S_A(S_A+S_B+S_C)},\quad 1-t_2=\frac{t\cdot c^2a^2}{S_B(S_A+S_B+S_C)},\quad 1-t_3=\frac{t\cdot a^2b^2}{S_C(S_A+S_B+S_C)}.$$

It is clear that the Euler lines $\mathcal{L}_i$, $i=1,2,3$, are concurrent. As t varies, this common point traverses a straight line $\mathcal{L}_t$. We show that this is the line joining K to $Q(K)$.

(1) For $t=1$, this Tucker circle is the second Lemoine circle with center K, the triangle $\mathbf{T}$ degenerates into the point K. The common point of the Euler lines is therefore the point $Q(K)$. See §3.2 and Figure 3.

(2) For $t=\frac{3}{2}$, the vertices of the Tucker hexagon are

$$\begin{aligned} B_a &= (a^2+b^2-2c^2:0:3c^2), & C_a &= (c^2+a^2-2b^2:3b^2:0),\\ C_b &= (3a^2:b^2+c^2-2a^2:0), & A_b &= (0:a^2+b^2-2c^2:3c^2),\\ A_c &= (0:3b^2:c^2+a^2-2b^2), & B_c &= (3a^2:0:b^2+c^2-2a^2). \end{aligned}$$

The triangles $\mathbf{T}_i$, $i = 1, 2, 3$, have a common centroid K, which is therefore the common point of their Euler lines. The corresponding Tucker center is the point X_{576} (which divides OK in the ratio $3 : -1$).

From these, we obtain the equation of the line

$$\mathcal{L}_{\text{t}} : \qquad \sum_{\text{cyclic}} b^2c^2S_A(S_B - S_C)(S_{CA} + S_{AB} - 2S_{BC})x = 0.$$

Remarks. (1) The triangle $\mathbf{T}$ is perspective to ABC at K. See, for example, [5].

(2) The line $\mathcal{L}_{\text{t}}$ also contains X_{74} which we may regard as corresponding to $t = 0$.

For more about Tucker hexagons, see §8.2.

7. Concurrency of four or more Euler lines

7.1. *Common point of $\mathcal{L}_i$, $i = 1, 2, 3$, on the Euler line of ABC.* We consider the case when the Euler lines $\mathcal{L}_i$, $i = 1, 2, 3$, intersect on the Euler line of ABC. A typical point on the Euler line axis divides the segment OH in the ratio $t : 1 - t$, has coordinates

$$(a^2S_A - (S_{CA} + S_{AB} - 2S_{BC})t) : \cdots : \cdots).$$

This lies on the Euler lines $\mathcal{L}_i$, $i = 1, 2, 3$, if and only if we choose

$$\begin{aligned} t_1 =& \frac{-(S^2 - S_{AA}) + (S^2 - 3S_{AA})t}{2S_{AA}}, \\ t_2 =& \frac{-(S^2 - S_{BB}) + (S^2 - 3S_{BB})t}{2S_{BB}}, \\ t_3 =& \frac{-(S^2 - S_{CC}) + (S^2 - 3S_{CC})t}{2S_{CC}}. \end{aligned}$$

Independently of t, the corresponding triangle $\mathbf{T}$ is always homothetic to the medial tangential triangle at the point P_{2071} on the line $\mathcal{L}_{\text{p}}$ for which $Q(P_{2071}) = X_{2071}$, the intersection of $\mathcal{L}_{\text{q}}$ with the Euler line. See the end of §3.2 above. The ratio of homothety is $1 + t - \frac{8S_{ABC}}{a^2b^2c^2}t$. We summarize this in the following proposition.

Proposition 11. *Let P_{2071} be the point on $\mathcal{L}_{\text{p}}$ such that $Q(P_{2071}) = X_{2071}$. The Euler lines $\mathcal{L}_i$, $i = 1, 2, 3$, corresponding to the sidelines of triangles homothetic at P_{2071} to the medial tangential triangle intersect on the Euler line of ABC.*

7.2. *The line $\mathcal{L}_{\text{f}}$.* The Euler line of triangle $\mathbf{T}$ is the line

$$\begin{aligned} &(x + y + z) \sum_{\text{cyclic}} t_1a^2S_{AA}(S_B - S_C)(S^2 + S_{BC})(S^2 - S_{AA}) \\ =&2S_{ABC} \sum_{\text{cyclic}} (S^2 + S_{CA})(S^2 + S_{AB})x. \end{aligned} \tag{5}$$

Theorem 12. *The Euler lines of the four triangles* $\mathbf{T}$ *and* $\mathbf{T}_i$*,* $i = 1, 2, 3$*, are concurrent if and only if*

$$\begin{aligned} t_1 &= -\frac{16S^2 \cdot S_{ABC} + t(a^2b^4c^4 - 4S_{ABC}(3S^2 - S_{AA}))}{4S_{AA}(a^2b^2c^2 + 4S_{ABC})}, \\ t_2 &= -\frac{16S^2 \cdot S_{ABC} + t(a^4b^2c^4 - 4S_{ABC}(3S^2 - S_{BB}))}{4S_{BB}(a^2b^2c^2 + 4S_{ABC})}, \\ t_3 &= -\frac{16S^2 \cdot S_{ABC} + t(a^4b^4c^2 - 4S_{ABC}(3S^2 - S_{CC}))}{4S_{CC}(a^2b^2c^2 + 4S_{ABC})}, \end{aligned}$$

with $t \neq \frac{-24a^2b^2c^2S_{ABC}}{(a^2b^2c^2-8S_{ABC})(3(S_A+S_B+S_C)S^2+S_{ABC})}$*. The locus of the common point of the four Euler lines is the line* $\mathcal{L}_\mathrm{f}$ *joining the nine-point center of* ABC *to* X_{184}*, with the intersection with* $\mathcal{L}_\mathrm{q}$ *deleted.*

Proof. The equation of the Euler line $\mathcal{L}_i$, $i = 1, 2, 3$, can be rewritten as

$$\begin{aligned} & t_1S_A(S_B - S_C)(x + y + z) + S_{AA}(S_B - S_C)x \\ & +(S_{AB}(S_B - S_C) - (S_{AA} - S_{BB})S_C)y + (S_{AC}(S_B - S_C) + (S_{AA} - S_{CC})S_B)z = 0, \quad (6) \\ & t_2S_A(S_B - S_C)(x + y + z) + S_{BB}(S_C - S_A)y \\ & +(S_{BA}(S_C - S_A) + (S_{BB} - S_{AA})S_C)x + (S_{BC}(S_C - S_A) - (S_{BB} - S_{CC})S_A)z = 0, \quad (7) \\ & t_3S_C(S_A - S_B)(x + y + z) + S_{CC}(S_A - S_B)z \\ & +(S_{CA}(S_A - S_B) - (S_{CC} - S_{AA})S_B)x + (S_{CB}(S_A - S_B) + (S_{CC} - S_{BB})S_A)y = 0. \quad (8) \end{aligned}$$

Multiplying (4), (5), (6) respectively by

$$a^2S_A(S^2+S_{BC})(S^2-S_{AA}), \quad b^2S_B(S^2+S_{CA})(S^2-S_{BB}), \quad c^2S_C(S^2+S_{AB})(S^2-S_{CC}),$$

and adding, we obtain by Theorem 10 the equation of the line

$$\mathcal{L}_\mathrm{f}: \qquad \sum_{\text{cyclic}} (S_B - S_C)(S^2(2S_{AA} - S_{BC}) + S_{ABC} \cdot S_A)x = 0$$

which contains the common point of the Euler lines of $\mathbf{T}_i$, $i = 1, 2, 3$, if it also lies on the Euler line $\mathcal{L}$ of $\mathbf{T}$. The line $\mathcal{L}_\mathrm{f}$ contains the nine-point center X_5 and $X_{184} = (a^4S_A : b^4S_B : c^4S_C)$. Let Q_t be the point which divides the segment $X_{184}X_5$ in the ratio $t : 1 - t$. It has coordinates

$$\begin{aligned} &((1 - t)4S^2 \cdot a^4S_A + t(a^2b^2c^2 + 4S_{ABC})(S_{CA} + S_{AB} + 2S_{BC}) \\ &: (1 - t)4S^2 \cdot b^4S_B + t(a^2b^2c^2 + 4S_{ABC})(2S_{CA} + S_{AB} + S_{BC}) \\ &: (1 - t)4S^2 \cdot c^4S_C + t(a^2b^2c^2 + 4S_{ABC})(S_{CA} + 2S_{AB} + S_{BC})). \end{aligned}$$

The point Q_t lies on the Euler lines $\mathcal{L}_i$, $i = 1, 2, 3$, respectively if we choose t_1, t_2, t_3 given above.

If Q lies on $\mathcal{L}_{\rm q}$, then $Q_t = Q(P)$ for some point P on $\mathcal{L}_{\rm p}$.[8] In this case, the triangle $\mathbf{T}$ degenerates into the point $P \neq Q$ and its Euler line is not defined. It should be excluded from $\mathcal{L}_{\rm f}$. The corresponding value of t is as given in the statement above. □

Here are some interesting points on $\mathcal{L}_{\rm f}$.

(1) For $t = 0$, $\mathbf{T}$ is perspective with ABC at X_{74}, and the common point of the four Euler lines is X_{184}. The antiparallels are drawn through the intercepts of the trilinear polars of $X_{186} = \left(\frac{a^2}{S_A(S^2-3S_{AA})} : \cdots : \cdots\right)$, the inversive image of the orthocenter in the circumcircle.

(2) For $t = 1$, this common point is the nine-point of triangle ABC. The triangle $\mathbf{T}$ is homothetic to the orthic triangle at X_{51} and to the medial tangential triangle at the point P_{2071} in §3.2.

(3) $t = -\frac{a^2b^2c^2}{4S_{ABC}}$ gives X_{156}, the nine-point center of the tangential triangle.

In these two cases, we have the concurrency of five Euler lines.

(4) The line $\mathcal{L}_{\rm f}$ intersects the Brocard axis at X_{569}. This corresponds to $t = \frac{2a^2b^2c^2}{3a^2b^2c^2+4S_{ABC}}$.

Proposition 13. *The triangle* $\mathbf{T}$ *is perspective with* ABC *and its Euler line contains the common point of the Euler lines of* $\mathbf{T}_i$, $i = 1, 2, 3$ *precisely in the following three cases.*

(1) $t = 0$, *with perspector* X_{74} *and common point of Euler line* X_{184}.

(2) $t = \frac{-12a^2b^2c^2S_{ABC}}{a^4b^4c^4-12a^2b^2c^2S_{ABC}-16(S_{ABC})^2}$, *with perspector* K.

Remarks. (1) In the first case,

$$t_1 = \frac{k}{S_{AA}}, \qquad t_2 = \frac{k}{S_{BB}}, \qquad t_3 = \frac{k}{S_{CC}}$$

for $k = -\frac{4S^2 \cdot S_{ABC}}{a^2b^2c^2+4S_{ABC}}$. The antiparallels pass through the intercepts of the trilinear polar of X_{186}, the inversive image of H in the circumcircle.

(2) In the second case, the antiparallels bound a Tucker hexagon. The center of the Tucker circle divides OK in the ratio $t : 1-t$, where

$$t = \frac{S^2(S_A+S_B+S_C)(a^2b^2c^2-16S_{ABC})}{a^4b^4c^4-12a^2b^2c^2S_{ABC}-16(S_{ABC})^2}.$$

It follows that the common point of the Euler lines is the intersection of the lines $\mathcal{L}_{\rm f} = X_5X_{184}$ and $\mathcal{L}_{\rm t}$.

8. Common points of $\mathcal{L}_i$, $i = 1, 2, 3$, when $\mathbf{T}$ is perspective

If the Euler lines $\mathcal{L}_i$, $i = 1, 2, 3$, are concurrent, then, according to (2) we may put

$$t_1 = \frac{k(\lambda+S_A)}{a^2S_{AA}}, \qquad t_2 = \frac{k(\lambda+S_B)}{b^2S_{BB}}, \qquad t_3 = \frac{k(\lambda+S_C)}{c^2S_{CC}}$$

---

[8] This point is the intersection of $\mathcal{L}_{\rm p}$ with the line joining the Jerabek center J to X_{323}, the reflection in X_{110} of the inversive image of the centroid in the circumcircle.

for some λ and k. If, also, the $\mathbf{T}$ is perspective, (4) gives

$$k(k\lambda+S_{ABC})(\lambda+S_A+S_B+S_C)(k(3\lambda+S_A+S_B+S_C)+2S_{ABC})=0.$$

If $k=0$, $\mathbf{T}$ is the orthic triangle. We consider the remaining three cases below.

8.1. *The case* $k(S_A+S_B+S_C+3\lambda)+2S_{ABC}=0$. In this case,

$$\begin{aligned}t_1&=-\frac{2S_{ABC}+k(S_B+S_C-2S_A)}{3a^2S_{AA}},\\ t_2&=-\frac{2S_{ABC}+k(S_C+S_A-2S_B)}{3b^2S_{BB}},\\ t_3&=-\frac{2S_{ABC}+k(S_A+S_B-2S_C)}{3c^2S_{CC}}.\end{aligned}$$

The antiparallels are concurrent.

8.2. *The case* $k\lambda+S_{ABC}=0$. In this case,

$$t_1=\frac{k-S_{BC}}{a^2S_A},\qquad t_2=\frac{k-S_{CA}}{b^2S_B},\qquad t_3=\frac{k-S_{AB}}{c^2S_C}.$$

In this case, the perspector is the Lemoine point K. The antiparallels bound a Tucker hexagon. The locus of the common point of Euler lines is the line $\mathcal{L}_t$. Here are some more interesting points on this line.

(1) For $k=0$, we have

$$t_1=-\frac{S_{BC}}{S_A(S_B+S_C)},\qquad t_2=-\frac{S_{CA}}{S_B(S_C+S_A)},\qquad t_3=-\frac{S_{AB}}{S_C(S_A+S_B)}.$$

This gives the Tucker hexagon with vertices

$$\begin{aligned}B_a&=(S_{CC}:0:S^2), & C_a&=(S_{BB}:S^2:0),\\ C_b&=(S^2:S_{AA}:0), & A_b&=(0:S_{CC}:S^2),\\ A_c&=(0:S^2:S_{BB}), & B_c&=(S^2:0:S_{AA}).\end{aligned}$$

These are the pedals of A', B', C' on the sidelines. The Tucker circle is the Taylor circle. The triangle $\mathbf{T}$ is the medial triangle of the orthic triangle. The corresponding Euler lines intersect at X_{974}, which is the intersection of $\mathcal{L}_t=KX_{74}$ with X_5X_{125}. See [2].

(2) For $k=\frac{S_{ABC}}{S_A+S_B+S_C}$, we have

$$t_1=-\frac{S_{BC}}{S_A(S_A+S_B+S_C)},\qquad t_2=-\frac{S_{CA}}{S_B(S_A+S_B+S_C)},\qquad t_3=-\frac{S_{AB}}{S_C(S_A+S_B+S_C)}.$$

The Tucker circle is the second Lemoine circle, considered in §6.

(3) The line $\mathcal{L}_t$ intersects the Euler line at

$$X_{378}=\left(\frac{a^2(S^2+3S_{AA})}{S_A}:\cdots:\cdots\right).$$

The corresponding Tucker circle has center

$$(S^2(S_B+S_C)(S_C-S_A)(S_A-S_B)+3(S_A+S_B)(S_B+S_C)(S_C+S_A)S_{BC}:\cdots:\cdots)$$

which is the intersection of the Brocard axis and the line joining the orthocenter to X_{110}.

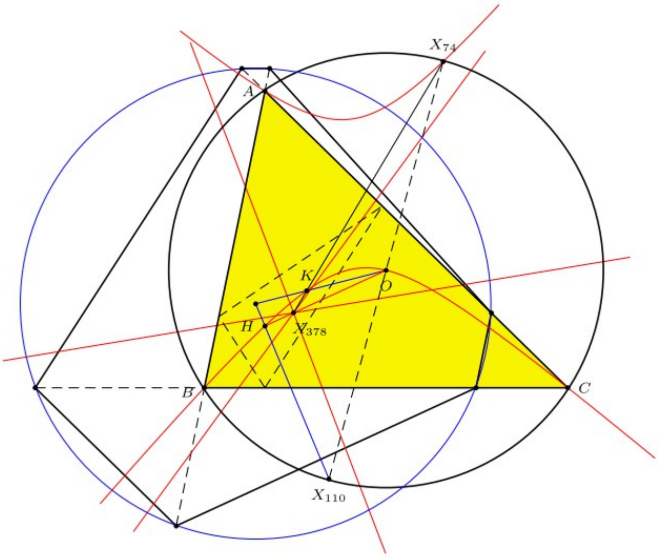

Figure 6. Intersection of 4 Euler lines at X_{378}

8.3. *The case* $\lambda = -(S_A+S_B+S_C)$. In this case, we have

$$t_1 = -\frac{k}{S_{AA}}, \qquad t_2 = -\frac{k}{S_{BB}}, \qquad t_3 = -\frac{k}{S_{CC}}.$$

In this case, the perspector is the point

$$\left(\frac{1}{2S_{ABC}\cdot S_A - k(b^2c^2-2S_{BC})}:\cdots:\cdots\right)$$

on the Jerabek hyperbola. If the point on the Jerabek hyperbola is the isogonal conjugate of the point which divides OH in the ratio $t:1-t$, then

$$k = \frac{4tS^2\cdot S_{ABC}}{a^2b^2c^2(1+t)+4t\cdot S_{ABC}}.$$

The locus of the intersection of the Euler lines $\mathcal{L}_i$, $i=1,2,3$, is clearly a line. Since this intersection is the Jerabek center for $k=0$ (Thébault's theorem) and the centroid for $k = \frac{S^2}{3}$, this is the line

$$\mathcal{L}_{\mathrm{r}} : \qquad \sum_{\text{cyclic}} (S_B - S_C)(S_{BC} - S_{AA})x = 0.$$

This line also contains, among other points, X_{110} and X_{184}. We summarize the general situation in the following theorem.

Theorem 14. *Let P be a point on the Euler line other than the centroid G. The antiparallels through the intercepts of the trilinear polar of P bound a triangle perspective with ABC (at a point on the Jerabek hyperbola). The Euler lines of the triangles $\mathbf{T}_i$, $i = 1, 2, 3$, are concurrent (at a point Q on the line L_{r} joining the centroid G to X_{110}).*

Here are some interesting examples with P easily constructed on the Euler line.

P	Perspector	Q
H	H	X_{125}
O	$X_{64} = X_{20}^*$	X_{110}
X_{30}	X_{2071}^*	G
X_{186}	X_{74}	X_{184}
X_{403}	$X_{265} = X186^*$	X_{1899}
X_{23}	$X_{1177} = X_{858}^*$	X_{182}
X_{858}		X_{1352}
X_{1316}		X_{98}

Remarks. (1) X_{186} is the inversive image of H in the circumcircle.
(2) X_{403} is the midpoint between H and X_{186}.
(3) X_{23} is the inversive image of G in the circumcircle.
(4) X_{858} is the inferior of X_{23}.
(5) X_{182} is the midpoint of OK, the center of the Brocard circle.
(6) X_{1352} is the reflection of K in the nine-point center.
(7) X_{1316} is the intersection of the Euler line and the Brocard circle apart from O.

9. Two loci: a line and a cubic

We conclude this paper with a brief discussion on two locus problems.

9.1. *Antiparallels through the vertices of a pedal triangle.* Suppose the antiparallels ℓ_i, $i = 1, 2, 3$, are constructed through the vertices of the pedal triangle of a finite point P. Then the Euler lines $\mathcal{L}_i$, $i = 1, 2, 3$, are concurrent if and only if P lies on the line

$$\sum_{\text{cyclic}} S_A(S_B - S_C)(S_{AA} - S_{BC})x = 0.$$

This is the line containing H and the Tarry point X_{98}. For $P = H$, the common point of the Euler line is

$$X_{185} = (a^2 S_A(S_A(S_{BB} + S_{CC}) + a^2 S_{BC}) : \cdots : \cdots).$$

9.2. *Antiparallels through the vertices of a cevian triangle.* If, instead, the antiparallels ℓ_i, $i = 1, 2, 3$, are constructed through the vertices of the cevian triangle of P, then the locus of P for which the Euler lines $\mathcal{L}_i$, $i = 1, 2, 3$, are concurrent is the cubic

$$\mathcal{K}: \quad \frac{S_A + S_B + S_C}{S_{ABC}} xyz + \sum_{\text{cyclic}} \frac{x}{S_A(S_B - S_C)} \left(\frac{S_A + S_B}{S_C} y^2 - \frac{S_C + S_A}{S_B} z^2 \right) = 0.$$

This can also be written in the form

$$\begin{aligned}&(\sum_{\text{cyclic}} (S_B + S_C)yz)(\sum_{\text{cyclic}} S_A(S_B - S_C)(S_B + S_C - S_A)x)\\ =&(\sum_{\text{cyclic}} S_A(S_B - S_C)x)(\sum_{\text{cyclic}} S_A(S_B + S_C)yz).\end{aligned}$$

From this, we obtain the following points on $\mathcal{K}$:

- the orthocenter H (as the intersection of the Euler line and the line $\sum_{\text{cyclic}} S_A(S_B - S_C)(S_B + S_C - S_A)x) = 0$),
- the Euler reflection point X_{110} (as the "fourth" intersection of the circumcircle and the circumconic $\sum_{\text{cyclic}} S_A(S_B + S_C)yz = 0$ with center K),
- the intersections of the Euler line with the circumcircle, the points X_{1113} and X_{1114}.

Corresponding to $P = X_{110}$, the Euler lines $\mathcal{L}_i$, $i = 1, 2, 3$, intersect at the circumcenter O. On the other hand, X_{1113} and X_{1114} are the points

$$(a^2 S_A + \lambda(S_{CA} + S_{AB} - 2S_{BC}) : \cdots : \cdots)$$

for $\lambda = -\frac{abc}{\sqrt{a^2b^2c^2 - 8S_{ABC}}}$ and $\lambda = \frac{abc}{\sqrt{a^2b^2c^2 - 8S_{ABC}}}$ respectively. The antiparallels through the traces of each of these points correspond to

$$t_1 = t_2 = t_3 = \frac{\lambda - 1}{\lambda + 1}.$$

This means that the corresponding intersections of Euler lines lie on the line $\mathcal{L}_c = HJ$ in §2.2.

9.3. *The cubic* $\mathcal{K}$. The infinite points of the cubic $\mathcal{K}$ can be found by rewriting the equation of $\mathcal{K}$ in the form

$$\begin{aligned}&(\sum_{\text{cyclic}} S_A(S_B - S_C)(S_B + S_C)yz)(\sum_{\text{cyclic}} (S_B + S_C)x)\\ =&(x + y + z)(\sum_{\text{cyclic}} (S_B + S_C)(S_B - S_C)(S_A(S_A + S_B + S_C) - S_{BC})yz)\end{aligned}$$

They are the infinite points of the Jerabek hyperbola and the line $(S_B + S_C)x + (S_C + S_A)y + (S_A + S_B)z = 0$. The latter is $X_{523} = (S_B - S_C : S_C - S_A : S_A - S_B)$. The asymptotes of $\mathcal{K}$ are

- the parallels to the asymptotes of Jerabek hyperbola through the antipode the Jerabek center on the nine-point circle, *i.e.*,

$$X_{113} = ((S_{CA} + S_{AB} - 2S_{BC})(b^2 S_{BB} + c^2 S_{CC} - a^2 S_{AA} - 2S_{ABC}) : \cdots : \cdots),$$

- the perpendicular to the Euler line (of ABC) at the circumcenter O, intersecting $\mathcal{K}$ again at

$$Z = \left(\frac{S_{CA} + S_{AB} - 2S_{BC}}{b^2 S_{BB} + c^2 S_{CC} - a^2 S_{AA} - 2S_{ABC}} : \cdots : \cdots \right),$$

  which also lies on the line joining H to X_{110}. See Figure 7. [9]

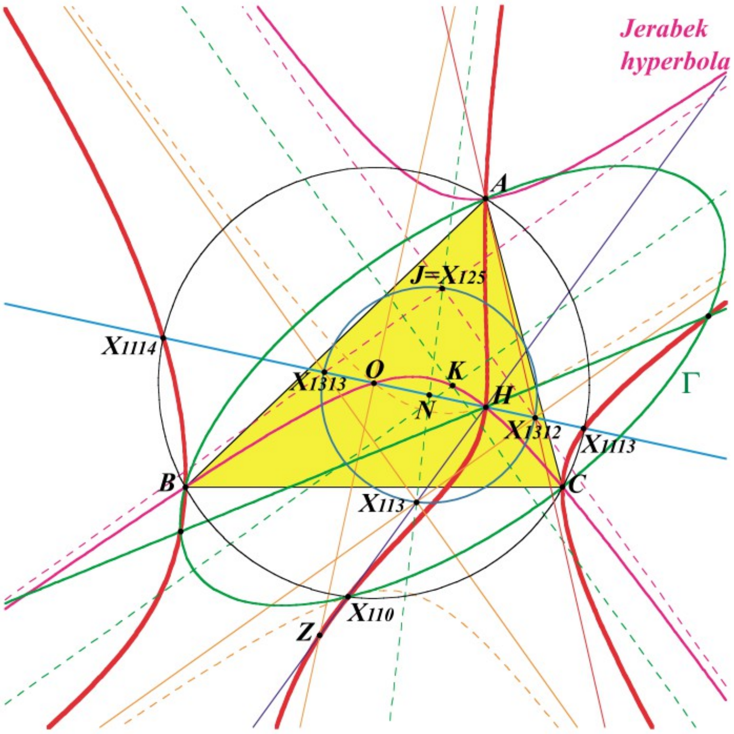
Jerabek hyperbola
A
J=X125
X1114
X1313
O
K
H
N
X1312
X1113
Γ
B
C
X113
X110
Z

Figure 7. The cubic $\mathcal{K}$

---

[9] We thank Bernard Gibert for providing the sketch of $\mathcal{K}$ in Figure 7.

Remark. The asymptotes of $\mathcal{K}$ and the Jerabek hyperbola bound a rectangle inscribed in the nine-point circle. Two of the vertices and $J = X_{125}$ and its antipode X_{113}. The other two are the points X_{1312} and X_{1313} on the Euler line.

References

[1] N. Dergiades, Hyacinthos 7777, September 4, 2003.
[2] J.-P. Ehrmann, Hyacinthos 3693, September 1, 2001.
[3] D. Grinberg, Hyacinthos 7781, September 4, 2003.
[4] C. Kimberling, *Encyclopedia of Triangle Centers*, available at http://faculty.evansville.edu/ck6/encyclopedia/ETC.html.
[5] F. M. van Lamoen, Concurrencies from Tucker hexagons, *Forum Geom.*, 2 (2002) 5–13.
[6] A. Myakishev, On the procircumcenter and related points, *Forum Geom.*, 3 (2003) 29–34.
[7] V. Thébault, Concerning the Euler line of a triangle, *Amer. Math. Monthly*, 54 (1947) 447–453.
[8] V. Thébault, O. J. Rammler, and R. Goormaghtigh, Problem 4328, *Amer. Math. Monthly*, 56 (1949) 39; solution, (1951) 45.

Nikolaos Dergiades: I. Zanna 27, Thessaloniki 54643, Greece
E-mail address: `ndergiades@yahoo.gr`

Paul Yiu: Department of Mathematical Sciences, Florida Atlantic University, Boca Raton, Florida, 33431-0991, USA
E-mail address: `yiu@fau.edu`

Forum Geometricorum
Volume 4 (2004) 21–22.

FORUM GEOM
ISSN 1534-1178

Another 5-step Division of a Segment in the Golden Section

Kurt Hofstetter

Abstract. We give one more 5-step division of a segment into golden section, using ruler and compass only.

Inasmuch as we have given in [1, 2] 5-step constructions of the golden section we present here another very simple method using ruler and compass only. It is fascinating to discover how simple the golden section appears. For two points P and Q, we denote by $P(Q)$ the circle with P as center and PQ as radius.

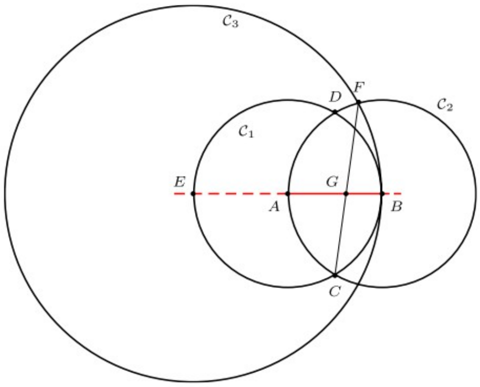

Figure 1

Construction. Given a segment AB, construct

(1) $\mathcal{C}_1 = A(B)$,
(2) $\mathcal{C}_2 = B(A)$, intersecting $\mathcal{C}_1$ at C and D,
(3) the line AB to intersect $\mathcal{C}_1$ at E (apart from B),
(4) $\mathcal{C}_3 = E(B)$ to intersect $\mathcal{C}_2$ at F (so that C and F are on opposite sides of AB),
(5) the segment CF to intersect AB at G.

The point G divides the segment AB in the golden section.

Publication Date: February 10, 2004. Communicating Editor: Paul Yiu.

Proof. Suppose AB has unit length. It is enough to show that $AG = \frac{1}{2}(\sqrt{5}-1)$.

Extend BA to intersect $\mathcal{C}_3$ at H. Let CD intersect AB at I, and let J be the orthogonal projection of F on AB. In the right triangle HFB, $BH = 4$, $BF = 1$. Since $BF^2 = BJ \times BH$, $BJ = \frac{1}{4}$. Therefore, $IJ = \frac{1}{4}$. It also follows that $JF = \frac{1}{4}\sqrt{15}$.

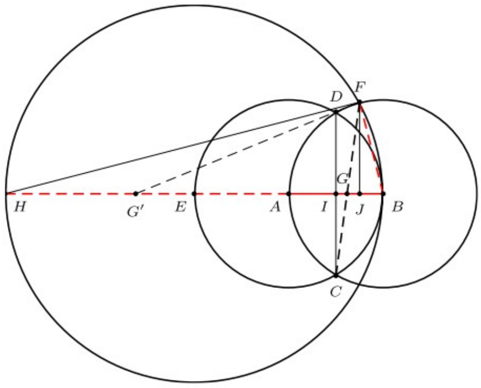

Figure 2

Now, $\frac{IG}{GJ} = \frac{IC}{JF} = \frac{\frac{1}{2}\sqrt{3}}{\frac{1}{4}\sqrt{15}} = \frac{2}{\sqrt{5}}$. It follows that $IG = \frac{2}{\sqrt{5}+2} \cdot IJ = \frac{\sqrt{5}-2}{2}$, and $AG = \frac{1}{2} + IG = \frac{\sqrt{5}-1}{2}$. This shows that G divides AB in the golden section. $\square$

Remark. If FD is extended to intersect AH at G', then G' is such that $G'A : AB = \frac{1}{2}(\sqrt{5}+1) : 1$.

After the publication of [2], Dick Klingens and Marcello Tarquini have kindly written to point out that the same construction had appeared in [3, p.51] and [4, S.37] almost one century ago.

References

[1] K. Hofstetter, A simple construction of the golden section, *Forum Geom.*, 2 (2002) 65–66.
[2] K. Hofstetter, A 5-step division of a segment in the golden section, *Forum Geom.*, 3 (2003) 205–206.
[3] E. Lemoine, *Géométrographie ou Art des Constructions Géométriques*, C. Naud, Paris, 1902.
[4] J. Reusch, *Planimetricsche Konstruktionen in Geometrographischer Ausführung*, Teubner, Leipzig, 1904.

Kurt Hofstetter: Object Hofstetter, Media Art Studio, Langegasse 42/8c, A-1080 Vienna, Austria
E-mail address: `pendel@sunpendulum.at`

27.04.2020

F) Books_72_2020/
Forumgeo_fau_edu/
FG2004Volume4.pdf
Pag. 5 Antiparallels and Concurrent Euler Lines.
Figure 1 Thébault's theorem on the ~~[illegible]~~ concurrency of Euler lines.
Look: Figure 2
Figure 3
Q(K) and the second Lemoine circle.
Obs: Ler o artigo todo. (Pag. 24

153) 20200427
A) Books_72_2020/
Projective_Geometry/
Geometry of Cuts and Metrics.pdf
Pág. 86 (96
6.4 An Example: the Spherical Distance Space
$x, y \in S$

% # " " < # 6
'()*' ' '+ ' # 0 2 ! 6683
- C E ! 4 -!
& ;< 6< < $. - B -

Theorem 6.3.7. *Let P be a polytope in $\mathbb{R}^n$ that is combinatorially equivalent to an n-dimensional hypercube, let $G(P) := (V_P, E_P)$ denote its 1-skeleton graph, and let D_P denote the set of diagonals of P. Then,*

$$\sum_{xy \in E_P} (\| x - y \|_2)^2 \geq \sum_{xy \in D_P} (\| x - y \|_2)^2,$$

with equality if and only if P is an n-parallelotope.

Proof. The proof is by induction on n. It relies essentially on the following identity:

$$\sum_{i,j=1}^{n} (\| x_i - y_j \|_2)^2 - \sum_{1 \leq i < i' \leq n} (\| x_i - x_{i'} \|_2)^2 + (\| y_i - y_{i'} \|_2)^2 = n^2 (\| p - q \|_2)^2$$

for any points $x_1, \ldots, x_n, y_1, \ldots, y_n$ and setting $p := \frac{1}{n} \sum_{i=1}^{n} x_i$, $q := \frac{1}{n} \sum_{i=1}^{n} y_i$. See Deza and Maehara [1994] for details. ∎

6.4 An Example: The Spherical Distance Space

We describe here a classical example of ℓ_1-embeddable distance space, namely, the spherical distance space. This is the distance space defined on a sphere S, taking as distance between two points $x, y \in S$ the quantity

$$(6.4.1) \qquad r \cdot \arccos \left(\frac{(x - c)^T (x - y)}{r^2} \right),$$

where c denotes the center of S and r its radius. The quantity (6.4.1) is known as the *spherical distance* (or *great circle metric*) between x and y. (This is the geodesic distance on the sphere S between the points x and y, which coincides with the angle between the two vectors from c to x and from c to y.) We may clearly suppose that the sphere S is centered at the origin and, up to scaling of the distances, that it has radius 1. Let S_m denote the m-dimensional unit sphere, i.e.,

$$S_m := \{x \in \mathbb{R}^{m+1} \mid \sum_{i=1}^{m+1} x_i^2 = 1\}.$$

We let $\mathcal{S}_m$ denote the distance space (S_m, d_S), where d_S is the distance defined by (6.4.1), i.e.,

$$(6.4.2) \qquad d_S(x, y) := \arccos(x^T y) \text{ for all } x, y \in S_m.$$

Similarly, we let $\mathcal{S}_{m,r}$ denote the distance space $(S_{m,r}, d_S)$, where $S_{m,r}$ is the m-dimensional sphere centered at the origin with radius r and d_S is defined by (6.4.1). So, $\mathcal{S}_m = \mathcal{S}_{m,1}$. The distance space $\mathcal{S}_{m,r}$ is called a *spherical distance space.* We refer to Blumenthal [1953] for a detailed study of the metric properties of these distance spaces. We mention here some properties that are most relevant to our treatment.

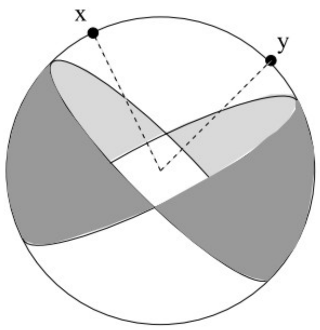

Figure 6.4.3: $H(x)\triangle H(y)$ is the shaded region

A first basic property of spherical distance spaces is that they are ℓ_1-embeddable. For this, let us consider for a point $x \in S_m$ the *hemisphere* $H(x)$ containing x; that is, $H(x)$ consists of the points $y \in S_m$ such that $d_S(x,y) \leq \frac{\pi}{2}$ or, equivalently, $x^Ty \geq 0$. We also consider the measure μ on S_m defined by

$$\mu(A) := \frac{\mathrm{vol}(A)}{\mathrm{vol}(S_m)} \quad \text{for } A \subseteq S_m.$$

(Here, $\mathrm{vol}(A)$ denotes the m-dimensional volume of A.) Hence, S_m equipped with the measure μ is a probability space. Then,

$$\text{(6.4.4)} \qquad \mu(H(x)\triangle H(y)) = \frac{1}{\pi}\arccos(x^Ty) \quad \left(= \frac{1}{\pi}d_S(x,y)\right) \quad \text{for } x,y \in S_m.$$

This relation can be easily verified; it was first observed by Kelly [1970b]. For an illustration in the case $m = 2$, see Figure 6.4.3. The result from Theorem 6.4.5[6] [7] below follows as an immediate consequence, using Proposition 4.2.1.

Theorem 6.4.5. *Given* $u_1, \ldots, u_n \in S_m$, *the vector* $(\frac{1}{\pi}\arccos(u_i^Tu_j))_{1\leq i<j\leq n}$ *belongs to the cut polytope* $\mathrm{CUT}_n^{\square}$. *Therefore, the spherical distance space* $\mathcal{S}_{m,r}$

---

[6] In other words, Theorem 6.4.5 indicates how to derive inequalities valid for the pairwise angles among a set of n unit vectors; namely, by considering valid inequalities for the cut polytope $\mathrm{CUT}_n^{\square}$. This fact will be reminded in Section 31.2.

[7] There is no analogue of Theorem 6.4.5 for the closely related elliptic distance spaces. The *elliptic distance space* $\mathcal{E}\ell_m$ is obtained from the spherical distance space $\mathcal{S}_m$ by identifying antipodal points; thus taking $\min(d_S(x,y), \pi - d_S(x,y))$ as distance between $x,y \in S_m$. The elliptic distance space $\mathcal{E}\ell_m$ is a semimetric space (Blumenthal [1953]) but it is not hypermetric (in fact, not 5-gonal (Kelly [1970b])). One more divergence between spherical and elliptic distance spaces lies in the determination of their congruence orders. As we see in Theorem 6.4.8 the order of congruence of the spherical distance space $\mathcal{S}_m$ can be easily determined; on the other hand the order of congruence of the elliptic distance space $\mathcal{E}\ell_m$ is known only in dimension $m \leq 2$ (see Seidel [1975] and precise references therein).

is isometrically ℓ_1*-embeddable for every* $r > 0$, $m \geq 1$. ∎

In fact, isometric subspaces of spherical distance spaces and of Euclidean distance spaces are very closely related notions, as the results below indicate.

Let d be a distance on the set $V_n := \{1, \ldots, n\}$. Consider the symmetric matrix[8] $A = (a_{ij})_{i,j=1}^n$ defined by

$$a_{ij} := \cos(d_{ij}) \quad \text{for } i, j = 1, \ldots, n.$$

Hence, the diagonal entries of A are all equal to 1. Moreover, let $\tilde{d}$ denote the distance on the set $V_n \cup \{x_0\}$ (x_0 is a new element not belonging to V_n) defined by

$$\begin{array}{ll} \tilde{d}(x_0, i) = 1 & \text{for } i \in V_n, \\ \tilde{d}(i, j) = 2 - 2\cos(d(i,j)) = 4\sin^2(\frac{d(i,j)}{2}) & \text{for } i, j \in V_n. \end{array}$$

Proposition 6.4.6. *Let* $m \geq 1$ *be an integer and* d, $\tilde{d}$, *and* A *be defined as above. The following assertions are equivalent.*

(i) *The distance space* (V_n, d) *is an isometric subspace of the* m*-dimensional spherical distance space* $\mathcal{S}_m$.

(ii) $d_{ij} \in [0, \pi]$ *for all* $i, j \in V_n$ *and the matrix* A *is positive semidefinite with* rank $A \leq m + 1$.

(iii) $d_{ij} \in [0, \pi]$ *for all* $i, j \in V_n$ *and the distance space* $(V_n \cup \{x_0\}, \tilde{d})$ *can be isometrically embedded into* ℓ_2^{m+1}.

Proof. The equivalence (i) $\Longleftrightarrow$ (ii) is clear. Indeed, (V_n, d) is an isometric subspace of $\mathcal{S}_m$ if and only if there exist $u_1, \ldots, u_n \in S_m$ such that $\cos(d(i,j)) = u_i^T u_j$ for all $i, j \in V_n$; that is, if the matrix A is positive semidefinite with rank $\leq m + 1$. The equivalence (ii) $\Longleftrightarrow$ (iii) follows from Corollary 6.2.5 and the fact that A coincides with the matrix $P_{x_0}(V_n \cup \{x_0\}, \tilde{d})$. ∎

Corollary 6.4.7. *A distance space* (V_n, d) *embeds isometrically in* $\mathcal{S}_{m,r}$ *(for some* $m \geq 1$*) if and only if* $d_{ij} \in [0, \pi r]$ *for all* $i, j \in V_n$ *and the matrix* $A := (\cos(\frac{d_{ij}}{r}))_{i,j=1}^n$ *is positive semidefinite. Then, the smallest* m *such that* (V_n, d) *embeds into* $\mathcal{S}_{m,r}$ *is* $m =$ rank $A - 1$. ∎

Using Proposition 6.4.6, one can derive a compactness result for spherical distance spaces analogue to the compactness result for Euclidean spaces from Theorem 6.2.13; this result can be found in Blumenthal [1953].

Theorem 6.4.8. *Given* $m \geq 1$, *a distance space* (X, d) *can be isometrically embedded into the* m*-dimensional spherical distance space* $\mathcal{S}_m$ *if and only if the*

---

[8] We will study in detail in Section 31.3 the set of positive semidefinite matrices with all diagonal entries equal to 1. We will use, in particular, Theorem 6.4.5.

same holds for every subspace (Y, d) *where* $Y \subseteq X$ *with* $|Y| = m + 3$. ▮

In particular, the order of congruence of the spherical distance space $\mathcal{S}_n$ is equal to $m + 3$. To see it, consider the distance space (X, d), where $|X| \geq m + 3$ and $d_{ij} := \arccos(-\frac{1}{m+1})$ for all $i \neq j \in X$. Then, (X, d) does not embed in $\mathcal{S}_m$ (because the matrix $(\cos(d_{ij}))_{i,j \in X}$ is not positive semidefinite). On the other hand, every subspace (Y, d) with $|Y| = m + 2$ embeds in $\mathcal{S}_m$ (as $(\cos(d_{ij}))_{i,j \in Y}$ is positive semidefinite with rank $m + 1$).

There are further intimate links between the spherical and Euclidean distances. In fact, as Schoenberg [1935] observed, every set of affinely independent vectors equipped with the Euclidean distance can be isometrically embedded into some spherical distance space. And, if the vectors are not independent, their Euclidean distances can be realized asymptotically as limits of spherical distances. This permits to derive the well-known implication: "ℓ_2-embeddable $\Longrightarrow \ell_1$-embeddable".

Proposition 6.4.9. *Let* $u_0, u_1, \ldots, u_n \in \mathbb{R}^n$ *be affinely independent. Then, the distance space* $(\{u_0, u_1, \ldots, u_n\}, d_{\ell_2})$ *is an isometric subspace of* $\mathcal{S}_{n,r}$ *for some* r *large enough.*

Proof. We can suppose without loss of generality that $u_0 := 0$. By assumption, the matrix $P = (p_{ij} := u_i^T u_j)_{i,j=1}^n$ is positive definite. Let $A(r)$ be the $(n + 1) \times (n + 1)$ symmetric matrix with entries $a_{ij} := \cos(\frac{d_{ij}}{r})$ for $i, j = 0, 1, \ldots, n$. In view of Corollary 6.4.7, it suffices to show that $A(r) \succ 0$ for r large enough. In what follows, we write $A \sim B$ for two matrices A and B if $A \succ 0 \Longleftrightarrow B \succ 0$. Clearly, $A(r) \sim B(r)$, where the 00-th entry of $B(r)$ is $b_{00} := 1$, its $0i$-th entry is $b_{0i} := a_{0i} - 1$ (for $i = 1, \ldots, n$) and its ij-th entry is $b_{ij} := a_{ij} - a_{0i} - a_{0j} + 1$ (for $i, j = 1, \ldots, n$) (to see it, subtract the row indexed by 0 in $A(r)$ to every other row and, then, the column indexed by 0 to every other column). We now use the fact that $\cos x = 1 - \frac{1}{2}x^2 + o(x^4)$ when $x \longrightarrow 0$. Hence, each entry of $B(r)$ can be expressed as $b_{0i} = -\frac{1}{2r^2}p_{ii} + o(\frac{1}{r^4})$, $b_{ij} = \frac{1}{r^2}p_{ij} + o(\frac{1}{r^4})$. Therefore, after suitably scaling $B(r)$, we obtain that $A(r) \sim B(r) \sim C(r)$, where the 0-th entry of $C(r)$ is $4r^2$, its $0i$-th entry is $p_{ii} + o(\frac{1}{r^2})$, and its ij-th entry is $p_{ij} + o(\frac{1}{r^2})$. One can now easily verify that $C(r) \succ 0$ for any r large enough. (Note here that it suffices to check that the $n+1$ principal subdeterminants of $C(r)$ disposed along the diagonal are positive, which holds for r large enough by the assumption that $P \succ 0$.) ▮

Remark 6.4.10. Proposition 6.4.9 extends to an arbitrary set of vectors in the case when they all lie on a common line. But, it does not extend to an arbitrary set of vectors in dimension ≥ 2. As counterexample, take for $u_1, \ldots, u_{n+1}$ the vertices of an equilateral simplex and for u_0 the barycentrum of the simplex. Then, the distance space $(\{u_0, u_1, \ldots, u_{n+1}\}, d_{\ell_2})$ does not embed in any spherical distance space. ▮

Let us now indicate how the Euclidean distance can be approximated by spherical distances[9]. For this, let $u_1, \ldots, u_n \in \mathbb{R}^m$. We are interested in evaluating their mutual Euclidean distances. So, set $d_{ij} := \| u_i - u_j \|_2$ for $i, j = 1, \ldots, n$. We show how to express d as a limit of distances $d^{(r)}$ (when $r \longrightarrow \infty$) where each $d^{(r)}$ can be isometrically embedded in $\mathcal{S}_{m,r}$. The idea for this is intuitively very simple. Namely, for $r > 0$ consider the sphere $S^{(r)}$ in the space $\mathbb{R}^{m+1}$ with center $c := (0, \ldots, 0, r)$ and radius r. One can visualize $S^{(r)}$ as a sphere lying on top of $\mathbb{R}^m$, being viewed as the hyperplane $x_{m+1} = 0$ in $\mathbb{R}^{m+1}$. Every vector $u \in \mathbb{R}^m$ with $\| u \|_2 \leq r$ can be 'lifted' to a point $u^{(r)} \in S^{(r)}$, by setting

$$u^{(r)} := \left(u, r - \sqrt{r^2 - (\| u \|_2)^2} \right).$$

Let $r \geq \max_{i=1}^n \| u_i \|_2$. For $i = 1, \ldots, n$, let $u_i^{(r)} \in S^{(r)}$ be the 'lifting' of u_i as defined above. Set

$$d^{(r)}(i,j) := r \cdot \arccos\left(\frac{(u_i - c)^T (u_j - c)}{r^2} \right) \quad \text{for } i, j = 1, \ldots, n.$$

So, $d^{(r)}(i,j)$ represents the spherical distance between $u_i^{(r)}$ and $u_j^{(r)}$ on the sphere $S^{(r)}$. Clearly, this length converges to the Euclidean distance $\| u_i - u_j \|_2$ as r tends to infinity, i.e.,

$$\lim_{r \longrightarrow \infty} d^{(r)}(i,j) = \| u_i - u_j \|_2 \, . \tag{6.4.11}$$

In a precise way, one can estimate $d^{(r)}(i,j)$ for large r as follows:

$$\begin{aligned} d^{(r)}(i,j) &= r \cdot \arccos\left(\tfrac{\sqrt{r^2 - (\|u_i\|_2)^2} \sqrt{r^2 - (\|u_j\|_2)^2} + u_i^T u_j}{r^2} \right) \\ &= r \cdot \arccos\left(\sqrt{1 - \tfrac{(\|u_i\|_2)^2}{r^2}} \sqrt{1 - \tfrac{(\|u_j\|_2)^2}{r^2}} + \tfrac{u_i^T u_j}{r^2} \right) \\ &\approx r \cdot \arccos\left(1 - \tfrac{(\|u_i - u_j\|_2)^2}{2r^2} \right) \approx \| u_i - u_j \|_2 \, . \end{aligned}$$

(Here, we use the fact that $\sqrt{1-x} \approx 1 - \frac{1}{2}x$ and $\arccos(1 - x^2) \approx \sqrt{2}x$ as $x \longrightarrow 0$.) ∎

Proposition 6.4.12. *For a distance space (X, d),*

d is isometrically ℓ_2-embeddable $\Longrightarrow$ d is isometrically ℓ_1-embeddable.

Proof. This follows from the above observations, (6.4.11) and Theorem 6.4.5. ∎

---

[9] This fact was already observed in Kelly [1975].

27.04.2020
X

F) Books_72_2020/
Forumgeo_fau_edu/
FG2004Volume4.pdf

Pag. 5 Antiparallels and Concurrent Euler Lines.

Figure 1 Thébault's theorem on the concurrency of Euler lines.

Look: Figure 2
Figure 3
Q(K) and the second Lemoine circle.

Obs: Ler o artigo todo. (Pag. 24

153) 20200427

A) Books_72_2020/
Projective_Geometry/
Geometry of Cuts and Metrics.pdf

Pag. 86/96

6.4 An Example: the Spherical Distance Space

$x, y \in S$

data 27/04/2020 &

$$r \cdot \arccos\left(\frac{(x-c)^T(x-y)}{r^2}\right)$$

(6.4.1)

↳ The quantity (6.4.1) is known as the spherical distance (or great circle metric) between x and y.

Pág 87(97)

Figure 6.4.3

$H(x) \Delta H(y)$ is the shaded region.

Pág. 90(100)

Let us now indicate the EUCLIDEAN distance can be approximated by spherical distance.[9]

9 → This fact was already observed in Kelly (1975).

B) Algorithms in Real Algebraic Geometry (2003...)...pdf

Pág. 430(434)

4 '()*' ' '+ ' # ( ) (4
0 2 #3 - C E ! 4 -!
& $# $#$ < . - 2 1 I 2 ! 1 ! -

Exercise 11.10. Prove that the complexity of Algorithm 11.17 (Stratifying Cylindrical Decision) $(s\,d)^{O(1)^k}$, where s is a bound on the number of elements of $\mathcal{P}$ and d is a bound on the degrees of the polynomials in $\mathcal{P}$. When $\mathrm{D}=\mathbb{Z}$, and the bitsizes of the coefficients of $\mathcal{P}$ are bounded by τ, prove the bitsizes of the intermediate computations and the output are bounded by $\tau d^{O(1)^{k-1}}$.

Using cylindrical decomposition, it is possible to design algorithms for computing various topological informations.

Remark 11.19.

a) It is possible to compute a semi-algebraic description of the semi-algebraically connected components of a $\mathcal{P}$-semi-algebraic set with complexity $(s\quad d)^{O(1)^k}$ as follows, using the proofs of Theorem 5.21 and Theorem 5.38 and the complexity analysis of Algorithm 11.17 (Stratifying Cylindrical Decomposition) (Exercise 11.10). Hint: Consider a $\mathcal{P}$-semi-algebraic set, compute a stratifying family adapted to $\mathcal{P}$ and a description of the cells associated to the corresponding Stratifying Cylindrical Decomposition. Determine the adjacency relations between cells from the list of non-empty sign conditions, using Proposition 5.39 (Generalized Thom's Lemma). Use these adjacencies to describe the connected components of S. We do not give details since we are going to see in Chapters 15 and 16 much better algorithms for the description of connected components of semi-algebraic sets.

b) A triangulation of a closed and bounded semi-algebraic set, as well as its homology groups, can be computed with complexity $(s\,d)^{O(1)^k}$, using the proof of Theorem 5.43, the definition of the homology groups of bounded and closed semi-algebraic sets in Chapter 6, and the complexity analysis of Algorithm 11.17 (Stratifying Cylindrical Decomposition) (Exercise 11.10). We do not give details either, even though we currently have no better algorithm for computing the homology groups.

c) A bound on the finite number of topological types of algebraic sets defined by polynomials of degree d follows from Algorithm 11.17 (Stratifying Cylindrical Decomposition). The bound is polynomial in d and doubly exponential in the number $\binom{d+k}{k}$ of monomials of degree d in k variables. Again, we do not give details for this quantitative version of Theorem 5.47. □

11.6 Topology of Curves

In this section, D is an ordered integral domain contained in a real closed field R and $\mathrm{C}=\mathrm{R}[i]$.

The simplest situation where the Cylindrical Decomposition Algorithm can be performed is the case of one single non-zero polynomial bivariate polynomial $P(X,Y)\in\mathrm{D}[X,Y]$.

The zero set of this polynomial is an algebraic subset $\mathrm{Zer}(P, \mathrm{R}^2) \subset \mathrm{R}^2$ contained in the plane R^2 and distinct from R^2.

Since an algebraic set is semi-algebraic, there are three possible cases:

- $\mathrm{Zer}(P, \mathrm{R}^2)$ is an algebraic set of dimension 1, which is called a **curve**.
- $\mathrm{Zer}(P, \mathrm{R}^2)$ is an algebraic set of dimension 0, i.e. a finite set of points.
- $\mathrm{Zer}(P, \mathrm{R}^2)$ is empty.

Typical examples of these three situations are the unit circle defined by $X^2 + Y^2 - 1 = 0$, the origin defined by $X^2 + Y^2 = 0$, and the empty set defined by $X^2 + Y^2 + 1 = 0$.

Let us consider a polynomial $P(X, Y)$ in two variables, monic and separable, of degree d as a polynomial in Y. By separable we mean in this section that any gcd of P and $\partial P / \partial Y$ is an element of $\mathrm{R}(X)$. This is not a big loss of generality since it is always possible to make a polynomial monic by a change of variables of the form $X + a\,Y$, $a \in \mathbb{Z}$, by Lemma 4.73. Replacing P by a separable polynomial can be done using Algorithm 10.1 (Gcd and gcd-free part), and does not modify the zero set. The zero set of this polynomial is an algebraic set $\mathrm{Zer}(P, \mathrm{R}^2)$ contained in the plane.

The cylindrifying family of polynomials associated to P consists of the subdiscriminants of P with respect to Y (see page 102 and Proposition 4.27), which are polynomials in the variable X, denoted by $\mathrm{sDisc}_j(X)$ for j from 0 to $d - 1$ (since P is monic in Y, sDisc_d and sDisc_{d-1} are constant). Denote by sDisc the list sDisc_d, ...sDisc_0. Note that $\mathrm{Disc}(X) = \mathrm{sDisc}_0(X)$ is the discriminant, and is not identically 0 since P is separable (see Equation (4.4) and Corollary 4.2). On intervals between the roots of $\mathrm{Disc}_0(X)$, the number of roots of $P(x, Y)$ in R is fixed (this is a special case of Theorem 5.16). The number of roots of $P(x, Y)$ can be determined by the signs of the other signed subresultant coefficients and is equal to PmV(sDisc) according to Theorem 4.33. Note that on an interval between two roots of $\mathrm{Disc}(X)$ in R, the signs of the subdiscriminants may change but PmV(sDisc) is fixed.

A cylindrical decomposition of the plane adapted to P can thus be obtained as follows: the cells of R are the roots of $\mathrm{Disc}(X)$ in R and the intervals they determine. Above each root of $\mathrm{Disc}(X)$ in R, $\mathrm{Zer}(P, \mathrm{R}^2)$ contains a finite number of points. These points and intervals between them are cells of R^2. Above each interval determined by the roots of $\mathrm{Disc}(X)$, there are 1 dimensional cells, called **curve segments**, which are graphs of functions from the interval to R, and 2 dimensional cells which are bands between these graphs. Finally, $\mathrm{Zer}(P, \mathrm{R}^2)$ is the union of the points and curve segments of this cylindrical decomposition and consists of a finite number of points projecting on the roots of $\mathrm{Disc}(X)$ and a finite number of curve segments homeomorphic to segments of R and projecting on intervals determined by the roots of $\mathrm{Disc}(X)$. If above every interval determined by the roots of $\mathrm{Disc}(X)$ there are no curve segments, $\mathrm{Zer}(P, \mathrm{R}^2)$ is at most a finite number of points. or empty. If moreover above every root of $\mathrm{Disc}(X)$ there are no points $\mathrm{Zer}(P, \mathrm{R}^2)$ is empty.

The purpose of the algorithm we are going to present is to compute exactly the topology of the curve $\mathrm{Zer}(P, \mathrm{R}^2)$, i.e. to determine a planar graph homeomorphic to $\mathrm{Zer}(P, \mathrm{R}^2)$. After performing, if necessary, a linear change of coordinates, this will be done by indicating adjacencies between points of the curve $\mathrm{Zer}(P, \mathrm{R}^2)$ above roots of $\mathrm{Disc}(X)$ and curve segments on intervals between these roots. In this study, the notions of critical points of the projection and of curves in generic position will be useful.

The **critical points** of the projection of $\mathrm{Zer}(P, \mathrm{C}^2)$ to the X-axis are the points $(x, y) \in \mathrm{Zer}(P, \mathrm{C}^2)$ such that y is a multiple root of $P(x, Y)$.

The critical points of the projection of $\mathrm{Zer}(P, \mathrm{C}^2)$ to the X-axis are of two kinds

- **singular points** of $\mathrm{Zer}(P, \mathrm{C}^2)$, i.e. points of $\mathrm{Zer}(P, \mathrm{C}^2)$ where
$$\partial P / \partial X(x, y) = \partial P / \partial Y(x, y) = 0,$$
- **ordinary critical points** points where the tangent to $\mathrm{Zer}(P, \mathrm{C}^2)$ is well defined and parallel to the Y-axis, i.e. points $(x, y) \in \mathrm{C}^2$ where
$$\partial P / \partial Y(x, y) = 0, \partial P / \partial X(x, y) \neq 0.$$

In both cases, the first coordinate of a critical point of the projection of $\mathrm{Zer}(P, \mathrm{C}^2)$ to the X-axis is a root of the discriminant $\mathrm{Disc}(X)$ of P considered as a polynomial in Y.

Computing the topology will be particularly easy for a curve in generic position, which is the notion we define now. Indeed, the critical points of the projection of $\mathrm{Zer}(P, \mathrm{C}^2)$ on the X_1-axis are easy to characterize in this case.

Let P be polynomial of degree d in $\mathrm{R}[X, Y]$ that is separable. The set $\mathrm{Zer}(P, \mathrm{C}^2)$ is in **generic position** if the following two conditions are satisfied:

- $\deg(P) = \deg_Y(P)$,
- for every $x \in \mathrm{C}$, $\gcd(P(x, Y), \partial P / \partial Y(x, Y))$ is either a constant or a polynomial of degree j with exactly one root of multiplicity j. In other words, there is at most one critical point (x, y) of the projection of $\mathrm{Zer}(P, \mathrm{C}^2)$ to the X_1-axis above any $x \in \mathrm{C}$.

Note that above an element x_1 of R which is a root of $\mathrm{Disc}(X)$, the unique root of $\gcd(P(x, Y), \partial P / \partial Y(x_{,Y}))$ is necessarily in R. So there is exactly a critical point with coordinates in R above each root of $\mathrm{Disc}(X)$ in R.

Example 11.20. If $P = (X^2 - Y + 1)(X^2 + Y^2 - 2Y)$, the set $\mathrm{Zer}(P, \mathrm{C}^2)$ is not in generic position since there are two critical points of the projection on X above the point 0, namely $(0, 0)$ and $(0, 1)$. □

The output of the algorithm computing the topology of a curve in generic position will be the following:

- the number r of roots of $\mathrm{Disc}(X)$ in R. We denote these roots by $x_1 < \cdots < x_r$, and by $x_0 = -\infty, x_{r+1} = +\infty$.

- the number m_i of roots of $P(x,\, Y)$ in R when x varies on $(x_i,\, x_{i+1})$, for $i = 0, \ldots, r$.
- the number n_i of roots of $P(x, Y)$ in R. We denote these roots by $y_{i,j}$, for $j = 1, \ldots, n_i$.
- a number $c_i \le n_i$ such that if $C_i = (x_i, z_i)$ is the unique critical point of the projection of $\mathrm{Zer}(P, \mathrm{C}^2)$ on the X-axis above x_i, $z_i = y_{i,c_i}$.

  More precisely, the output is

$$[m_0, [n_1, c_1], \ldots, m_{r-1}, [n_r, c_r], m_r]].$$

It is clear that a graph homeomorphic to $\mathrm{Zer}(P, \mathrm{R}^2) \subset \mathrm{R}^2$ can be drawn using the output since if $m_i \geq n_i$ (resp. $m_i \geq n_{i+1}$), C_i belong to the closure of $m_i - n_i$ (resp. $m_i - n_{i+1}$) curve segments above (x_i, x_{i+1}) getting glued at C_i (resp. C_{i+1}) and if $m_i = n_i - 1$ (resp. $m_i = n_{i+1} - 1$) the point C_i does not belong to the closure of a curve segment above (x_i, x_{i+1}).

The critical points of the projection of $\mathrm{Zer}(P,\, \mathrm{C}^2)$ on the X-axis in the generic position case are easy to determine.

We denote

$$\begin{aligned} \mathrm{sDiscP}_j(X, Y) &= \mathrm{sResP}_j(P(X, Y), \partial P / \partial Y(X, Y)) \\ \mathrm{sDisc}_j(X) &= \mathrm{sRes}_j(P(X, Y), \partial P / \partial Y(X, Y)) \end{aligned}$$

(P is considered as a polynomial in Y).

Proposition 11.21. *Let P be a polynomial of degree d in $\mathrm{R}[X, Y]$, separable and in generic position. If (x, y) is a critical point of the projection of $\mathrm{Zer}(P,\, \mathrm{C}^2)$ on the X-axis, and j is the multiplicity of y as a root of $P(x, Y)$ (considered as a polynomial in X_2), then*

$$\mathrm{Disc}(x) = \mathrm{sDisc}_1(x) = 0, \ldots, \mathrm{sDisc}_{j-1}(x) = 0, \mathrm{sDisc}_j(x) \neq 0,$$

and

$$y = -\frac{1}{j} \frac{\mathrm{sDisc}_{j,j-1}(x)}{\mathrm{sDisc}_j(x)},$$

where $sDisc_{j,j-1}(X)$ is the coefficient of Y^{j-1} in $\mathrm{sDiscP}_j(X, Y)$.

Proof: Since $(x, y) \in \mathrm{R}^2$ is a critical point of the projection of $\mathrm{Zer}(P, \mathrm{C}^2)$ on the X-axis, $j > 0$. Since P is in generic position, $\mathrm{sDiscP}_j(x, Y)$ is a degree j polynomial with only one root, y, which implies that

$$\mathrm{sDiscP}_j(x, Y) = \mathrm{sDisc}_j(x)(Y - y)^j, \tag{11.4}$$

and

$$y = -\frac{1}{j} \frac{\mathrm{sDisc}_{j,j-1}(x)}{\mathrm{sDisc}_j(x)},$$

identifying coefficients of degree $j - 1$ on both sides of (11.4). □

We can also count the number of points of the curve above the critical points using the following result.

Proposition 11.22. *Let* $P(X,Y)\in \mathrm{D}[X,Y]$ *be a degree* d *separable polynomial such that* $\mathrm{lcof}_Y(P)\in \mathrm{D}$ *and* $x\in \mathrm{R}$. *Let*

$$R(X,Y,Z)=(Y-Z)\,\partial P/\partial Y(X,Y)-d\,P(X,Y).$$

We denote

$$T_j(X,Z)=\mathrm{sRes}_j(P(X,Y),R(X,Y,Z))$$

(considered as polynomials in Y*) and* $T(X,Z)$ *the list* $T_j(X,Z)$, j *from* 0 *to* $d-1$. *Let* $r=\mathrm{PmV}(\mathrm{sDisc}(x))$.

a)

$$\#\{y\in \mathrm{R}\mid P(x,y)=0\}=r$$

b) If $P(x,z)\neq 0$,

$$\#\{y\in \mathrm{R}\mid P(x,y)=0\ \wedge\ z<y\}=(r+\mathrm{PmV}(T(x,z)))/2.$$

c) If $P(x,z)=0$,

$$\#\{y\in \mathrm{R}\mid P(x,y)=0\ \wedge\ z<y\}=(r+\mathrm{PmV}(T(x,z))-1)/2.$$

Proof: We denote

$$\begin{aligned}
Z_x &= \{y\in \mathrm{R}\mid P(x,y)=0\},\\
\mathrm{TaQ}(1,Z_x) &= \#(Z_x)\\
&= \#(\{y\in \mathrm{R}\mid P(x,y)=0\})\\
\mathrm{TaQ}(Y-z,Z_x) &= \#(\{y\in \mathrm{R}\mid P(x,y)=0\wedge y>z\})\\
&- \#(\{y\in \mathrm{R}\mid P(x,y)=0\wedge y<z\}),\\
\mathrm{TaQ}((Y-z)^2,Z_x) &= \#\{y\in \mathrm{R}\mid P(x,y)=0\wedge y\neq z\}.
\end{aligned}$$

a) follows from Theorem 4.33. b) and c) follow from Equation (2.6) (see page 65) applied to $P(x,Y)$ and $Y-z$:

$$\begin{bmatrix} 1 & 1 & 1\\ 0 & 1 & -1\\ 0 & 1 & 1\end{bmatrix}\cdot\begin{bmatrix} c(Y=z,P(x,Y)=0)\\ c(Y>z,P(x,Y)=0)\\ c(Y<z,P(x,Y)=0)\end{bmatrix}=\begin{bmatrix}\mathrm{TaQ}(1,P(x,Y))\\ \mathrm{TaQ}(Y-z,P(x,Y))\\ \mathrm{TaQ}((Y-z)^2,P(x,Y))\end{bmatrix}.$$

If $P(x,z)\neq 0$, we have $c(Y-z=0,P(x,Y)=0)=0$. Thus

$$c(Y>z,P(x,Y)=0)=(\mathrm{TaQ}(1,P(x,Y))+\mathrm{TaQ}(Y-z,P(x,Y)))/2.$$

If $P(x,z)=0$, we have $c(Y-z=0,P(x,Y)=0)=1$. Thus

$$c(Y>z,P(x,Y))=(\mathrm{TaQ}(1,P(x,Y))+\mathrm{TaQ}(Y-z,P(x,Y))-1)/2. \qquad \square$$

Next we show that it is always possible to perform a linear change of variables such that in the new coordinates the curve is in generic position. The idea is to maximize the number of distinct roots of the discriminant, which are the X-coordinates of the critical points.

Let P be a separable polynomial in $\mathrm{D}[X,\ Y]$, U a new variable and $Q(U,X,Y)$ the polynomial defined by:

$$Q(U,X,Y)=P(X+UY,Y).$$

If $a\in\mathbb{Z}$, let $\mathrm{Zer}(P_a,\mathrm{R}^2)$ denote the curve defined by the polynomial

$$P_a(X,Y)=Q(a,X,Y)=0.$$

We denote

$$\begin{aligned}\mathrm{sDiscP}_j(U,X,Y) &= \mathrm{sResP}_j(Q(U,X,Y),\partial Q/\partial Y(U,X,Y))\\ \mathrm{sDisc}_j(U,X) &= \mathrm{sRes}_j(Q(U,X,Y),\partial Q/\partial Y(U,X,Y))\end{aligned}$$

Note that $\mathrm{Disc}(U,\ X)$=$\mathrm{sDisc}_0(U,\ X)$ is the discriminant of $Q(U,\ X,\ Y)$ with respect to the variable Y. We denote by $\mathrm{sDisc}(U,\ X)$ the list of the subdiscriminants $\mathrm{sDisc}_j(U,X)$. Let $\Delta(U)$ be the non-zero subdiscriminant of smallest possible index of $\mathrm{Disc}(U,X)$ with respect to the variable X.

Proposition 11.23. *Let a be an integer number such that*

$$\deg_{X_2}(Q(a,X,Y))=\deg(Q(a,X,Y)) \quad \textit{and} \quad \Delta(a)\neq 0.$$

Then P_a is in generic position.

Proof: Suppose that $\deg_{X_2}(P_a(X,\ Y))=\deg(P_a(X,\ Y))$ and P_a is not in generic position. Let δ be a new variable and consider the field $\mathrm{C}\langle\delta\rangle$ of algebraic Puiseux series in δ. We are going to prove the following property (P): the number of distinct roots of $\mathrm{Disc}(a+\delta,X)$ in $\mathrm{C}\langle\delta\rangle$, which is the discriminant of $P_{a+\delta}(X,\ Y)$ with respect to the variable Y, is bigger than the number of distinct roots of $\mathrm{Disc}(a,\ X)$ in C, which is the discriminant of $P_a(X,Y)$ with respect to the variable Y. Thus, by definition of Δ, $\Delta(a)=0$, and the statement is proved.

The end of the proof is devoted to proving property (P).

We first study the number of critical points of the projection of $\mathrm{Zer}(P_\delta,\mathrm{C}^2)$ on the X-axis close to a critical point of the projection of $\mathrm{Zer}(P,\mathrm{C}^2)$ to the X-axis.

- If (x,y) is a singular point of $\mathrm{Zer}(P_a,\mathrm{C}^2)$, $(x+\delta y,y)$ is a singular point of $\mathrm{Zer}(P_{a+\delta},\mathrm{C}^2)$.
- If (x,y) is an ordinary critical point of the projection of $\mathrm{Zer}(P_a,\mathrm{C}^2)$ on the X_1-axis, $(x+\delta y)$ is not a critical point of the projection of $\mathrm{Zer}(P_{a+\delta},\mathrm{C}^2)$ on the X_1-axis. However we are going to prove that there is an ordinary critical point of the projection of $\mathrm{Zer}(P_{a+\delta},\mathrm{C}^2)$ on the X_1-axis of the form $(x+\delta y+u,y+v)$, $o(u)>1,o(v)>0$.

If (x,y) is an ordinary critical point of the projection of $\mathrm{Zer}(P_a,\mathrm{C}^2)$ on the X-axis, there exists r such that

$$\begin{gathered}P_a(x,y)=0,\partial P_a/\partial X(x,y)=b\neq 0,\\ \partial P_a/\partial Y(x,y)=\cdots=\partial^{r-1}P_a/\partial Y^{r-1}(x,y)=0,\partial^r P_a/\partial Y^r(x,y)=r!\,c\neq 0.\end{gathered}$$

Writing Taylor's formula for P_a in the neighborhood of (x, y), we have

$$P_a(X, Y) = b(X - x) + c(Y - y)^r + A(X, Y),$$

with $A(X, Y)$ a linear combination of monomials multiple of $(X - y)$ or $(Y - y)^r$. We consider a new variable ε and a point of $\mathrm{Zer}(P_a, \mathrm{C}\langle\varepsilon\rangle^2)$ with coordinates $x + \xi, y + \varepsilon$, infinitesimally close to x, y. In other words we consider a solution ξ of

$$b\xi + c\varepsilon^r + A(x + \xi, y + \varepsilon) = 0 \tag{11.5}$$

which is infinitesimal. Using the proof of Theorem 2.91, there is a solution of Equation (11.5) of the form $\xi = -(c/b)\varepsilon^r + w$, $0(w) > r$. Moreover

$$\begin{array}{rcl} \partial P_a/\partial X(x + \xi, y + \varepsilon) & = & b + w_1, o(w_1) > 1, \\ \partial P_a/\partial Y(x + \xi, y + \varepsilon) & = & c\, r\, \varepsilon^{r-1} + w_2, o(w_2) > r - 1. \end{array}$$

Thus, if $(\partial P_a/\partial X(x + \xi', y + \varepsilon'), \partial P_a/\partial Y(x + \xi', y + \varepsilon'))\}$ is proportional to $(1, \delta)$, with ε' and ξ' infinitesimal in $\mathrm{C}\langle\delta\rangle$, we have

$$\begin{array}{rcl} \varepsilon' & = & d\delta^{1/(r-1)} + w_3 \\ \xi' & = & -(1/r)\, d\, \delta^{r/(r-1)} + w_4 \end{array}$$

with $d^{r-1} = b/c\ r$, $o(w_3) > 1/(r - 1)$, $o(w_4) > r/(r - 1)$. Thus there is a point $(x + \xi', y + \varepsilon')$ of $\mathrm{Zer}(P_a, \mathrm{C}\langle\delta\rangle^2)$ with gradient proportional to $(1, \delta)$, and $o(\varepsilon') > 0, o(\xi') > 1$. In other words, $(x + \delta y + \xi' + \delta\varepsilon', y + \varepsilon')$ is a critical point of the projection of $\mathrm{Zer}(P_{a+\delta}, \mathrm{C}\langle\delta\rangle^2)$ on the X-axis.

Suppose that (x, y_1) and (x, y_2) are two distinct critical point of the projection of $\mathrm{Zer}(P_a, \mathrm{C}^2)$ to the X_1-axis. According to the preceding discussion, there are two critical point of the projection of $\mathrm{Zer}(P_{a+\delta}, \mathrm{C}^2)$ to the X-axis, with first coordinates $x + \delta\, y_1 + u_1$ and $x + \delta\, y_2 + u_2$, with $o(u_1) > 1, o(u_2) > 1$.Note that $x + \delta\, y_1 + u_1$ and $x + \delta\, y_2 + u_2$ are distinct, since $y_1 - y_2$ is not infinitesimal. We have proved that the number of distinct roots of $\mathrm{Disc}(a + \delta, X)$ in $\mathrm{C}\langle\delta\rangle$ is strictly bigger that the number of distinct roots of $\mathrm{Disc}(a, X)$ in C. □

The Topology of a Curve Algorithm can now be described. We perform the computation as if the curve was in generic position. If it is not, the algorithm detects it and then a new linear change of coordinates is performed.

Algorithm 11.18. **[Topology of a Curve]**

- **Structure:** an ordered integral domain D contained in a real closed field R (resp. the field of real numbers $\mathbb{R}$).
- **Input:** a separable polynomial P in $\mathrm{D}[X, Y]$ of degree d.
- **Output:** the topology of the curve $\mathrm{Zer}(P, \mathrm{R}^2)$, described by
  - An integer a such that $P_a(X, Y) = P(X + a\,Y, Y)$ is in general position.
  - The number r of roots of $\mathrm{Disc}(a,\ X)$ in R. We denote these roots by $x_1 < \cdots < x_r$, and by $x_0 = -\infty, x_{r+1} = +\infty$.
  - The number m_i of roots of $P_a(x, Y)$ in R when x varies on (x_i, x_{i+1}), $i = 0, \ldots, r$.

  - The number n_i of roots of $P_a(x_i, Y)$ in R. We denote these roots by $y_{i,j}$, $j = 1, \ldots, n_i$.
  - An index $c_i \leq n_i$ such that if $C_i = (x_i, z_i)$ is the unique critical point of the projection of $\mathrm{Zer}(P_a, \mathrm{C}^2)$ on the X_1-axis above x_i, $z_i = y_{i,c_i}$.

  More precisely, the output is $[a, [m_0, [n_1, c_1], \ldots, m_{r-1}, [n_r, c_r], m_r]]$.
- **Complexity:** $O(d^{11} \log_2(d))$, where d is a bound on the degrees of P.
- **Procedure:**
  - Take $a := 0$.
  - $(\star)$ Define $P_a(X, Y) := P(X + a\,X, Y)$.
  - If P_a is not quasi-monic with respect to Y, the curve is not in generic position. Take $a := a + 1$, go to $(\star)$.
  - Otherwise, compute the subdiscriminants $\mathrm{sDisc}(a,\ X)$ of $P_a(X,\ Y)$ (considered as a polynomial in Y). Characterize the roots $x_1, \ldots, x_r$ of $\mathrm{Disc}(a, X)$ using Algorithm 10.14 (Thom encoding) (resp. Algorithm 10.4 (Real Root Isolation)).
  - For every $1 \leq i \leq r$, determine $j(i)$ such that

    $$\mathrm{Disc}_0(a, x_i) = 0, \ldots, \mathrm{sDisc}_{j(i)-1}(a, x_i) = 0, \mathrm{sDisc}_{j(i)}(a, x_i) \neq 0,$$

    and compute the sign ε_i of $\mathrm{sDisc}_{j(i)}(a, x_i)$ using Algorithm 10.15 (Sign at the Roots in a real closed field) (resp. Algorithm 10.6 (Sign at a Real Root)).
  - Define

    $$z_i = -\frac{1}{j(i)} \frac{\mathrm{sDisc}_{j(i), j(i)-1}(a, x_i)}{\mathrm{sDisc}_{j(i)}(a, x_i)}.$$

  - Check whether, for every $i = 1, \ldots, r$, z_i is a root of multiplicity $j(i)$ of $\mathrm{sDisc}_{j(i)}(x_i, Y)$ using Algorithm 10.15 (Sign at the Roots in a real closed field) (resp. Algorithm 10.6 (Sign at a Real Root)).
  - If there exists i such that z_i is not a root of multiplicity $j(i)$ of $\mathrm{sDisc}_{j(i)}(a,\ x_i,\ Y)$, the curve is not in generic position. Take $a := a + 1$, go to $(\star)$.
  - If for every i, z_i is a root of multiplicity $j(i)$ of $\mathrm{sDisc}_{j(i)}(a, x_i, Y)$, the curve is in generic position.
  - For every $0 \leq i \leq r$ choose an intermediate point $t_i \in (x_i, x_{i+1})$ (with the convention $x_0 = -\infty, x_{r+1} = +\infty$) using Algorithm 10.18 (Intermediate Points) and Remark 10.77, and compute the number m_i of roots of $P_a(t_i,\ Y)$ evaluating the signs of the signed subresultant coefficients at t_i using Algorithm 10.11 (Sign Determination) (resp. choose an intermediate rational point $t_i \in (x_i, x_{i+1})$ using the isolating intervals for the x_i and compute the number m_i of roots of $P_a(t_i, Y)$ using Algorithm 10.4 (Real Root Isolation)).
  - Compute, for every $i \leqslant r$,

    $$\begin{aligned} S &= \varepsilon_i\,(j(i)\,\mathrm{sDisc}_{j(i)}(a, X)\,Y + \mathrm{sDisc}_{j(i), j(i)-1}(X)\,(\partial P_a / \partial Y)) \\ R &= S - \varepsilon_i\, j(i)\, d\ \mathrm{sDisc}_{j(i)}(a, X)\ P_a \end{aligned}$$

and the list

$$T_i(X) \;=\; \mathrm{sRes}(P_a, R).$$

- Evaluate the signs of the elements of $T_i(X)$ at x_i to determine the number of real roots of the polynomial $P_a(x_i, Y)$ which are strictly bigger than z_i using Proposition 11.22 and Algorithm 10.15 (Sign at the Roots in a real closed field) (resp. Algorithm 10.6 (Sign at a Real Root)).
- Decide whether $\mathrm{Zer}(P, \mathrm{R}^2)$ is empty or a finite number of points.

Proof of correctness: By Lemma 4.74, there are only a finite number of values of a such that $\deg_{X_2}(Q(a, X, Y)) \neq \deg(Q(a, X, Y))$. Moreover there are only a finite number of zeros of $\Delta(X)$. So by Proposition 11.23, there are only a finite number of values of a such that the curve is not in generic position. The correctness of the algorithm follows from Proposition 11.21, and the correctness of Algorithm 10.14 (Thom encoding) (resp. Algorithm 10.4 (Real Root Isolation)), Algorithm 10.15 (Sign at the Roots in a real closed field) (resp. Algorithm 10.6 (Sign at a Real Root)), Algorithm 10.18 (Intermediate Points), and Algorithm 10.11 (Sign Determination). □

Complexity analysis: We estimate the complexity of the computation in a general real closed field.

Let d be the degree of $P(X, Y)$ a separable polynomial in $\mathrm{D}[X, Y]$.

The degree of $\mathrm{Disc}(U, X)$ is $O(d^2)$, and the degree of δ is $O(d^4)$. So there are at most $O(d^4)$ values of a to try.

For each value of a, we compute the subdiscriminant sequence of $P_a(X, Y)$ (considered as polynomial in Y): this requires $O(d^2)$ arithmetic operations in $\mathrm{D}[X]$ by Algorithm 8.21 (Signed subresultant). The degrees of the polynomials in X produced in the intermediate computations of the algorithm are bounded by $2\, d^2$ by Proposition 8.45. The complexity in D for this computation is $O(d^6)$.

The Thom encoding of the roots of $\mathrm{Disc}(a,\ X)$ takes $O(d^8\ \log_2(d))$ arithmetic operations using the complexity analysis of Algorithm 10.14 (Thom encoding). Checking whether $Q(X,\ Y)$ is in generic position takes also $O(d^8 \log_2(d))$ arithmetic operations using the complexity analysis of Algorithm 10.15 (Sign at the Roots in a real closed field), since there are at most $O(d^2)$ calls to this algorithm with polynomials of degree d^2.

Since there are at most $O(d^4)$ values of a to try, the complexity to reach generic position is $O(d^{11} \log_2(d))$.

The choice of the intermediate points, the determination of all m_i, the determination of $j(i)$ and ε_i takes again $O(d^8 \log_2(d))$ arithmetic operations using the complexity analysis of Algorithm 10.18 (Intermediate Points) and Algorithm 10.15 (Sign at the Roots in a real closed field).

The computation of one list T_i by Algorithm 8.21 (Signed subresultant) requires $O(d^2)$ arithmetic operations in $\mathrm{D}[X]$. The degree of the polynomials in X produced in the intermediate computations of the algorithm are bounded by $O(d^4)$ by Proposition 8.45. The complexity in D for the computation of the list $\bar{T}$ is $O(d^{10})$.

So the total complexity for computing the various T_i is $O(d^{11})$.

The only remaining computations are the determination of the signs taken by the polynomials of T_i evaluated at the roots of $\mathrm{Disc}(a,\, X)$. The degrees of the polynomials involved are $O(d^4)$ and and their number is $O(d^2)$. This takes $O(d^{10})$ operations using the complexity analysis of Algorithm 10.15 (Sign at the Roots in a real closed field).

So the total complexity for determining the topology of $\mathrm{Zer}(P,\ \mathrm{R}^2)$ is $O(d^{11}\log_2(d))$.

If $\mathrm{D} = \mathbb{Z}$ and the coefficients of P are of bitsizes bounded by τ, the coefficients of $P_a(X,Y)$ are $\tau + O(d)\,\nu$ where ν is the bitsize of d, since there are at most $O(d^4)$ values of a to try. For each value of a, the bitsizes of the integers produced in the intermediate computations of the subresultantnt sequence of $P_a(X,Y)$ are bounded by $O(d\,(\tau + d\nu))$ by the complexity analysis of Algorithm 8.21 (Signed subresultant).

The univariate sign determinations performed in the last steps of the algorithm produce integers of bitsizes $O(d^2\,(\tau + d\,\nu)\log_2(d)))$, according to the complexity analysis of Algorithm 10.14 (Thom encoding), Algorithm 10.15 (Sign at the Roots in a real closed field), Algorithm 10.18 (Intermediate Points). So the maximum bitsizes of the integers in the computation determining the topology of $\mathrm{Zer}(P, \mathrm{R}^2)$ is $O(d^2\,(\tau + d\,\nu)\log_2(d)))$. □

Example 11.24. Let us consider the real algebraic curve defined by the monic polynomial

$$P = 2\,Y^3 + (3\,X - 3)\,Y^2 + (3\,X^2 - 3\,X)\,Y + X^3.$$

The discriminant of P (considered as polynomial in Y) is

$$-\,27X^2(X^4 + 6X^2 - 3).$$

whose real roots are given by

$$x_1 = -\sqrt{-\,3 + 2\sqrt{3}}, \qquad x_2 = 0, \qquad x_3 = \sqrt{-\,3 + 2\sqrt{3}}.$$

Using the signed subresultant sequence, we determine that above each $x_{1,i}$, for $i = 1, \ldots, 3$, $P(x_i, Y)$ has two real roots, and only one of these roots is double. Thus the curve is already in generic position. The multiplicity of the unique critical point (x_i, z_i) as a root of $P(x_i, Y)$ is 2, and is given by the equation

$$z_i = -\frac{x_i(x_i^2 + 2x_i - 1)}{2\,x_i^2 - 2}.$$

We also determine that before x_1 and after x_3, $P(x, Y)$ has only one real root, while between x_1 and x_2, and between x_2 and x_3 $P(x, Y)$ has three real roots.

For $i = 1, 3$, there is one root of $P(x_i, Y)$ under z_i, while for $i = 2$ there is one root of $P(x_i, Y)$ above z_i.

Finally, the topology of the curve is given by the following graph

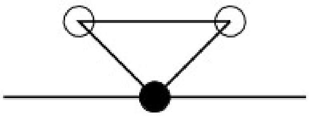

Fig. 11.3. Topology of the curve

where the points represent the critical points of the projection on the X-axis. The white points are critical points of the projection on the X-axis, while the black one is singular.

The topology can be described by 0 which is the value of a (the curve was given in generic position) and the list

$$[1, [2, 2], 3, [2, 1], 3, [2, 2], 1]$$

which can be read as follows: there are three roots of $\mathrm{Disc}(a, X)$, the number of branches above (x_0, x_1) is 1, the number of points above x_1 is 2 and the index of the critical point is 2, the number of branches above (x_1, x_2) is 3, the number of points above x_2 is 2 and the index of the critical point is 1, the number of branches above (x_2, x_3) is 3, the number of points above x_3 is 2, and the index of the critical point is 2, the number of branches above (x_3, x_4) is 1. □

11.7 Restricted Elimination

A variant of Algorithm 11.1 (Elimination) will be useful in the last chapters of the book. In this variant, we are interested at the signs of a family of polynomials at the roots of a polynomial, rather that at all the sign conditions of a family of polynomials.

We need some preliminary work.

We first prove a result similar to Theorem 5.12 and Proposition 5.13 which is adapted to the situation we are interested in studying.

Proposition 11.25. *Let P and Q be in $\mathrm{R}[X_1, \ldots X_k]$, and let S be a semi-algebraically connected semi-algebraic subset of R^{k-1} such that P is not identically 0 on S, and such that $\deg(P(x', X_k))$, the number of distinct roots of P in C, and $\deg(\gcd(P(x', X_k), Q(x', X_k)))$ are constant over S. Then there are ℓ continuous semi-algebraic functions $\xi_1 < \cdots < \xi_\ell$: $S \to \mathrm{R}$ such that, for every $x' \in S$, the set of real roots of $P(x', X_k)$ is exactly $\{\xi_1(x'), \ldots, \xi_\ell(x')\}$. Moreover, for $i = 1, \ldots, \ell$, the multiplicity of the root $\xi_i(x')$ is constant for $x' \in S$ and so is the sign of $Q(\xi_i(x'))$.*

'()*' ' '+ ' # E ! - 2
<3 5 -!
+4 9 1 ! ===

! '()*' ' '+ ' # 0 4 (- - @ 0 2
<3 - -!
& 8

Complex Surfaces of General Type: Some Recent Progress

Ingrid C. Bauer[1], Fabrizio Catanese[2], and Roberto Pignatelli[3,⋆]

[1] Mathematisches Institut, Lehrstuhl Mathematik VIII, Universitätstraße 30, D-95447 Bayreuth, Germany `Ingrid.Bauer@uni-bayreuth.de`

[2] Mathematisches Institut, Lehrstuhl Mathematik VIII, Universitätstraße 30, D-95447 Bayreuth, Germany `Fabrizio.Catanese@uni-bayreuth.de`

[3] Dipartimento di Matematica, Università di Trento, via Sommarive 14, I-38050 Povo (TN), Italy `Roberto.Pignatelli@unitn.it`

Introduction

In this article we shall give an overview of some recent developments in the theory of complex algebraic surfaces of general type.

After the *rough* or *Enriques-Kodaira* classification of complex (algebraic) surfaces, dividing compact complex surfaces in four classes according to their Kodaira dimension $-\infty$, 0, 1, 2, the first three classes nowadays are quite well understood, whereas even after decades of very active research on the third class, the class of surfaces *of general type*, there is still a huge number of very hard questions left open. Of course, we made some selection, which is based on the research interest of the authors and we claim in no way completeness of our treatment. We apologize in advance for omitting various very interesting and active areas in the theory of surfaces of general type as well as for not being able to mention all the results and developments which are important in the topics we have chosen.

Complex surfaces of general type come up with certain (topological, birational) invariants, topological as for example the topological Euler number e and the self intersection number of the canonical divisor K^2 of a minimal surface, which are linked by several (in-) equalities. In the first chapter we will summarize the classically known inequalities, which force surfaces of general type in a certain region of the plane having K^2 and e as coordinates, and we shall briefly comment on the so-called *geography* problem, whether,

---

[⋆] The present work was performed in the realm of the SCHWERPUNKT "Globale Methoden in der komplexen Geometrie", and was also supported by a VIGONI-DAAD Program. A first draft of this article took origin from the lectures by the second author at the G.A.C. Luminy Meeting, october 2005: thanks to the organizers!

given numerical invariants lying in the admissible range, i.e., fulfilling the required inequalities, does there exist a surfaces having these invariants. We shall however more broadly consider the three classical invariants K^2, p_g, q, which determine the other invariants $\chi := 1 - q + p_g, e = 12\chi - K^2$.

An important new inequality, which Severi tried without success to establish, and which has been attacked for many years with partial results by several authors, asserts that a surface of maximal Albanese dimension satisfies the inequality $K^2 \geq 4\chi$. We will report on Pardini's surprisingly simple proof of this so-called Severi's conjecture (cf. [Par05]).

The study of the pluricanonical maps is an essential technique in the classification of surface of general type. The main results concerning the m-canonical maps with $m \geq 3$ go back to an earlier period and we refer to [Cat87b] for a report on them.

We will report in the second chapter on recent developments concerning the bicanonical map; we would like to mention Ciliberto's survey (cf. [Cil97]) on this topic for the state of art ten years ago. Here instead, we combine a discussion of this topic with the closely intertwined problem of classification of surfaces with low values of the numerical invariants.

In the third chapter we report on surfaces of general type with geometric genus p_g equal to four, a class of surfaces whose investigation was started by Federigo Enriques (cf. chapter VIII of his book 'Le superficie algebriche', [Enr49]).

By Gieseker's theorem we know that for fixed K^2 and χ there exists a quasi projective coarse moduli space $\mathcal{M}_{K^2,\chi}$ for the birational equivalence classes of surfaces of general type. It is a very challenging problem to understand the geometry of these moduli spaces even for low values of the invariants. The case $p_g = 4$ is studied via the behaviour of the canonical map. While it is still possible to divide the moduli space into various locally closed strata according to the behaviour of the canonical map, it is very hard to decide how these strata patch together.

Using certain presentations of Gorenstein rings of codimension 4 introduced by M. Reid and D. Dicks, which arrange the defining equations as Pfaffians of certain matrices with many symmetries in such a way that these equations behave well under deformation, it is possible to exhibit explicit deformations, which allow to "connect" certain irreducible components of the moduli space.

Inspired by a construction of A. Beauville of a surface with $K^2 = 8$, $p_g = q = 0$, the second author defined Beauville surfaces as surfaces which are rigid and which admit an unramified covering which is isomorphic to a product of curves of genus at least 2. In this case the moduli space of surfaces orientedly homeomorphic to a given surface consists either of a unique real point, or of a pair of complex conjugate points corresponding to complex conjugate surfaces.

These surfaces, and the more general surfaces isogenous to a product, not only provide cheap counterexamples to the Friedman-Morgan speculation (which will be treated more extensively in the sixth section of this article), but provide also a wide class of surfaces quite manageable in order to test conjectures, and offer also counterexamples to various problems. The ease with which one can handle these surfaces is based on the fact that these surfaces are determined by "discrete" combinatorial data.

Beauville surfaces, their relations to group theory and to Grothendieck's theory of 'Dessins d'enfants' will be discussed in the fourth chapter.

It is a very difficult and very intriguing problem to decide whether two algebraic surfaces, which are not deformation equivalent, are in fact diffeomorphic.

The theory of Lefschetz fibrations provides an algebraic tool to prove that two surfaces are diffeomorphic. By a theorem of Kas (which holds also in the symplectic context) two Lefschetz fibrations are diffeomorphic if and only if their corresponding factorizations of the identity in the mapping class group are equivalent under the equivalence relation generated by Hurwitz moves and by simultaneous conjugation. We outline the theory, which was used with success in [CW04] in chapter five, which we end with a brief report on the status of two very old conjectures by Chisini concerning cuspidal curves and algebraic braids.

As already mentioned before, one of the fundamental problems in the theory of surfaces of general type is to understand their moduli spaces, in particular the connected components which parametrize the deformation equivalence classes of minimal surfaces of general type. By a classical result of Ehresmann, two deformation equivalent algebraic varieties are diffeomorphic. The other direction, i.e., whether two diffeomorphic minimal surfaces of general type are indeed in the same connected component of the moduli space, was an open problem since the eighties. We discuss in the last chapter the various counterexamples to the Friedman-Morgan speculation, who expected a positive answer to the question (unlike the second author, cf. [Kat83]).

Moreover, we briefly report on another equivalence relation introduced by the second author, the so-called quasi étale-deformation (Q.E.D.) equivalence relation, i.e., the equivalence relation generated by birational equivalence, by quasi étale morphisms and by deformation equivalence. For curves and surfaces of special type two varieties are Q.E.D. equivalent if and only if they have the same Kodaira dimension, whereas there are infinitely many surfaces of general type, which are pairwise not Q.E.D. equivalent.

1 Old and New Inequalities

1.1 Invariants of Surfaces

Let X be a compact complex manifold and let Ω^n_X be its canonical bundle, i.e., the line bundle of holomorphic $n-$forms (usually denoted by ω_X, since it is a dualizing sheaf in the sense of Serre duality). A corresponding canonical divisor is usually denoted by K_X.

To X one associates its *canonical ring*

$$R(X) := \oplus_{m \geq 0} H^0(\omega_X^{\otimes m}).$$

The trascendency degree over $\mathbb{C}$ of this ring leads to

- the *Kodaira dimension* $\kappa(X) := tr(R(X)) - 1$,

if $R(X) \neq \mathbb{C}$, otherwise $\kappa(X) := -\infty$. The Kodaira dimension is invariant under deformation (by Siu's theorem [Siu02], generalizing Iitaka's theorem for surfaces) and can assume the values $-\infty, 0, \dots, n = \dim X$.

Definition 1. *X is said to be* of general type *if the Kodaira dimension is maximal, $\kappa(X) = \dim X$.*

We are interested in the case of *surfaces*, i.e., of manifolds of dimension 2, of general type.

The three principal invariants under deformations for the study of these surfaces are

- the self intersection of the canonical class K_S^2 of a minimal model,
- the geometric genus $p_g := h^0(\omega_X)$ and
- the irregularity $q := h^1(\mathcal{O}_S) = h^0(\Omega^1_S)$.

The equality $h^1(\mathcal{O}_S) = h^0(\Omega^1_S)$ follows by Hodge theory since every algebraic surface is projective.

The invariants we have introduced, with the exception of K_S^2, are not only deformation invariants but also birational invariants.

Definition 2. *A smooth surface S is called* minimal *(or* a minimal model*) iff it does not contain any exceptional curve E of the first kind (i.e. $E \cong \mathbb{P}^1$, $E^2 = -1$).*

Every surface can be obtained by a minimal one (its "minimal model") after a finite sequence of blowing ups of smooth points; this model is moreover unique if $\kappa(S) \geq 0$ (see III.4.4, III.4.5 and III.4.6 of [BHPV04]). Thus, every birational class of surfaces of general type contains exactly one minimal surface, and one classifies surfaces of general type by studying their minimal models. To each minimal surface of general type we will associate its numerical

- type (K_S^2, p_g, q),

a triple of integers given by the three invariants introduced above.

In fact these determine all other classical invariants, as

- the Euler-Poincaré characteristic of the trivial sheaf $\chi(\mathcal{O}_S) = 1 - q + p_g$;
- the topological Euler characteristic $e(S) = c_2(S) = 12\chi(\mathcal{O}_S) - K_S^2$;
- the plurigenera $P_m(S) := h^0(\omega_X^{\otimes m}) = \chi(\mathcal{O}_S) + \binom{m}{2} K_S^2$.

The epression for c_2 is a classical theorem of M. Noether, and the expression for the plurigenera follows by Riemann-Roch and by Mumford's vanishing theorem.

By the theorems on pluricanonical maps (cf. [Bom73]), minimal surfaces S of general type with fixed invariants are birationally mapped to normal surfaces X in a fixed projective space of dimension $P_5(S) - 1$. X is uniquely determined, is called the *canonical model* of S, and is obtained contracting to points all the (-2)-curves of S (curves $E \cong \mathbb{P}^1$, with $E^2 = -2$).

Let us recall Gieseker's theorem

Theorem 1 (Gieseker [Gie77]). *There exists a quasi-projective coarse moduli scheme for canonical models of surfaces of general type S with fixed K_S^2 and $c_2(S)$.*

In particular, we can consider the subscheme $\mathcal{M}_{K_S^2, p_g, q}$ corresponding to minimal surfaces of general type of type (K_S^2, p_g, q). By the above theorem, it is a quasi projective scheme, in particular, it has finitely many irreducible components.

It is a dream ever since to completely describe $\mathcal{M}_{K_S^2, p_g, q}$ for as many types as possible.

1.2 Classical Inequalities and Geography

Obviously the first question is: for which values of (K_S^2, p_g, q) is $\mathcal{M}_{K_S^2, p_g, q}$ non empty?

For example, it is clear that $p_g(S)$ and $q(S)$ are always nonnegative, since they are dimensions of vector spaces.

In fact much more is known. In the following table we collect the well known classical inequalities holding among the invariants of minimal surfaces of general type:

$$\begin{array}{lll}
 & K_S^2 \geq 1 & \qquad\qquad\qquad \chi \geq 1 \\
(N) & K_S^2 \geq 2p_g - 4 & \text{or the weaker } K_S^2 \geq 2\chi(\mathcal{O}_S) - 6 \\
(D) & \text{if } q > 0,\ K_S^2 \geq 2p_g & \text{or the weaker if } q > 0,\ K_S^2 \geq 2\chi(\mathcal{O}_S) \\
(MY) & K_S^2 \leq 9\chi &
\end{array}$$

We have labeled by (N)= Noether, (D) = Debarre, (MY) = Miyaoka-Yau the rows, corresponding to the names of the inequalities ([Deb82], [Deb83], [Miy77], [Yau78], see also [BHPV04], chap. 7).

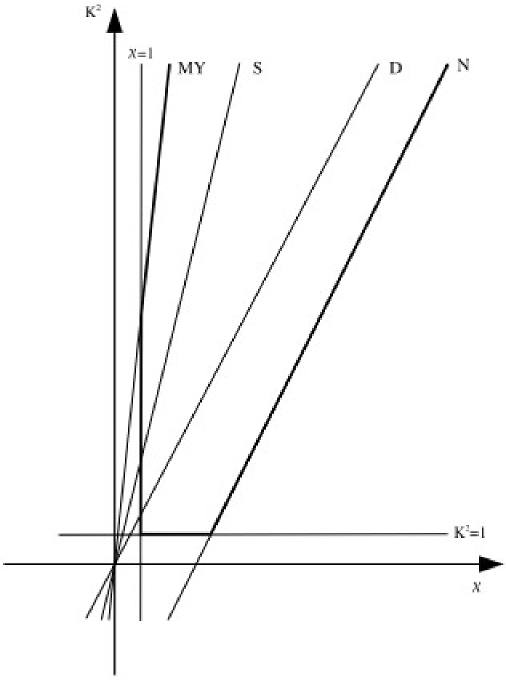

Fig. 1. The geography of minimal surfaces of general type

In figure 1 we have drawn the limit lines (i.e., where equality holds) of the various inequalities in the (χ, K_S^2)-plane.

The above listed inequalities show that the pair of invariants χ, K_S^2 of a minimal surface of general type gives a point with integral coordinates in the convex region limited by the "bold" piecewise linear curve. Moreover, if $q > 0$ this point cannot be at the "right" of the line D.

We drew one more line in our picture, labeled by S. This is the Severi line $K^2 = 4\chi$, i.e., the equality case of the Severi inequality $K^2 \geq 4\chi \Leftrightarrow K^2 \geq \frac{1}{2}e$, which will be discussed in detail at the end of this section.

1.3 Surfaces Fibred over a Curve

An important method for the study of surfaces of general type is to consider relatively minimal fibrations of surfaces over curves $f : S \to B$.

Definition 3. *A* fibration $f : S \to B$ *is a surjective morphism with connected fibres. We are interested in the case of fibrations* of surfaces to curves*, meaning that in this paper* S *and* B *will always be smooth compact complex manifolds of respective dimensions* 2 *and* 1.

The fibration is said to be relatively minimal *if* f *does not contract any rational curve of self intersection* -1 *to a point.*

One denotes

- by b the genus of the base curve B;
- by g the genus of a general fibre.

To avoid confusion, let us point out that a fibration is called rational or irrational according to the genus b of the base being 0 or > 0. On the other hand, the genus of the fibration is the genus g of the fibre. For example, if we say f is a genus 2 rational fibration, we intend that $g = 2$ and $b = 0$.

The classical way of saying: a genus b pencil of curves of genus g is however still the most convenient way to describe a fibration.

To a relatively minimal fibration f one associates

- its relative canonical bundle $\omega_{S|B} := \omega_S \otimes f^*(\omega_B^\vee)$ and
- the sheaves $(\forall n \geq 0)$ $V_n := f_*(\omega_{S|B}^{\otimes n})$.

The sheaves V_n are vector bundles (i.e., locally free sheaves) with very nice properties.

Theorem 2 (Fujita [Fuj78a], [Fuj78b]). *The vector bundles V_n are* semipositive, *i.e., every locally free quotient of it has nonnegative degree.*

To be more precise, V_1 is a direct sum of an ample vector bundle with $q(S) - b$ copies of the trivial bundle and with some undecomposable stable degree 0 vector bundle without global sections. Zucconi [Zuc97] proved moreover that if one of those stable bundles has rank 1, then it is a torsion line bundle.

For $n \geq 2$ we have:

Theorem 3 (Esnault-Viehweg [EV90]). *$\forall n \geq 2$ the vector bundle V_n is ample unless f has constant moduli, which means that all the smooth fibres are isomorphic.*

Since $R^1 f_* \omega_{S|B} = \mathcal{O}_B$ by relative duality, and $R^1 f_* \omega_{S|B}^{\otimes n} = 0 \ \forall n \geq 2$ by the assumption of relative minimality, one can compute the Euler characteristic of V_n by Riemann-Roch, and consequently its degree.

We introduce the following invariants of the fibration f:

- the self intersection of the relative canonical divisor

$$K_f^2 := \omega_{S|B} \cdot \omega_{S|B} = K_S^2 - 8(g-1)(b-1),$$

- the Euler characteristic of the relative canonical divisor

$$\chi_f = \chi(\omega_{S|B}) = \chi(\mathcal{O}_S) - (g-1)(b-1),$$

- its slope $\lambda(f) := K_f^2 / \chi_f$.

The slope is clearly defined only for $\chi_f \neq 0$, or equivalently (as we will see soon) if the fibration is not a holomorphic bundle.

The above mentioned computation gives

$$\deg V_n = \chi_f + \frac{n(n-1)}{2} K_f^2$$

and since by Fujita's theorem these numbers are nonnegative this gives the two inequalities $K_f^2 \geq 0$ and $\chi_f \geq 0$ respectively known as Arakelov's inequality (cf. [Ara71]) and Beauville's inequality (cf. [Bea82]).

In fact, we have the following list of inequalities

$$\begin{array}{ll} (A) & K_f^2 \geq 0 \ , \ i.e., \ K_S^2 \geq 8(g-1)(b-1), \\ (B) & \chi_f \geq 0 \ , \ i.e., \ \chi(\mathcal{O}_S) \geq (g-1)(b-1), \\ (ZS) & c_2(S) \geq 4(b-1)(g-1), \\ (NN) & q \leq b+g, \\ (X) & 4 - \frac{4}{g} \leq \lambda(f) \leq 12. \end{array}$$

Here the meaning of the labeling is the following: (A) = Arakelov's inequality, (B) = Beauville' inequality, (X) = Xiao's inequality (also known as slope inequality), (NN) = no name's inequality, (ZS) = Zeuthen-Segre. A proof of those inequalities can be found in [Bea82] with the exception of the slope inequality, proved in [Xia87] (see also [CH88] in the semistable case).

The equality cases of the first 4 inequalities are well described:

- if equality holds in (A), f has constant moduli;
- equality holds in (B) $\Leftrightarrow$ f has constant moduli and is smooth;
- for $g \geq 2$, equality holds in (ZS) $\Leftrightarrow$ f is smooth;
- $q = b + g \Leftrightarrow f$ is birationally equivalent to the projection of a product $B \times F$ to the first factor.

In particular, we see that the slope is defined whenever the fibration is not a holomorphic bundle, since the denominator χ_f vanishes iff equality holds in Beauville's inequality.

An important consequence is the following

Theorem 4 (Beauville). *If X is a minimal surface of general type, then $p_g \geq 2q - 4$. Moreover, if $p_g = 2q - 4$, then S is a product of a curve of genus 2 with a curve of genus $q - 2$.*

Note for later use (see next section) the following

Corollary 1. *If $p_g = q$ (i.e., if $\chi(\mathcal{O}_S) = 1$), then $p_g = q \leq 4$. Moreover, minimal surfaces of general type with $p_g = q = 4$ are exactly the products of two genus 2 curves.*

Proof of theorem 4. The standard wedge product on 1−forms induces a natural map

$$\wedge : \Lambda^2 H^0(\Omega^1_S) \to H^0(\Omega^2_S)$$

Recall that $q = \dim H^0(\Omega^1_S)$, $p_g = \dim H^0(\Omega^2_S)$. Let us assume $p_g \leq 2q - 4$.

By a dimension count, if $p_g \leq 2q - 4$, the projective linear subspace of $\mathbb{P}(\Lambda^2 H^0(\Omega^1_S))$ corresponding to the kernel of the above map must intersect the Plücker embedding of the Grasmannian $G_2(H^0(\Omega^1_S))$ (which has dimension $2q - 4$), and therefore there are two linearly independent $1-$forms ω_1 and ω_2 such that the following holomorphic two form is identically zero: $\omega_1 \wedge \omega_2 \equiv 0$.

By the theorem of Castelnuovo-De Franchis there is a fibration $f : S \to B$ with base of genus $b \geq 2$, and two holomorphic $1-$forms $\alpha_1,\ \alpha_2 \in H^0(\Omega^1_B)$ such that $f^*\alpha_i = \omega_i$. Since S is of general type, also $g \geq 2$.

Then

$$\chi_f \geq 0 \Rightarrow \chi(\mathcal{O}_S) \geq (b-1)(g-1) = (b-2)(g-2) + b + g - 3 \geq q - 3.$$

So we have $1 - q + p_g \geq q - 3 \Leftrightarrow p_g \geq 2q - 4$.

If $p_g = 2q - 4$, all inequalities are equalities and then, since $q = b + g$ and $(b-2)(g-2) = 0$, S is a product of two curves of genus at least 2, and one of the two must have genus exactly 2.

□

1.4 Severi's Inequality

We recall that the Albanese variety $Alb(X)$ of a compact Kähler manifold X is the cokernel of the natural map

$$\int : H_1(X, \mathbb{Z}) \to H^0(\Omega^1(X))^\vee$$

defined by integrating $1-$forms on 1-cycles.

The Albanese morphism

$$\alpha : X \to Alb(X)$$

is defined (up to translations in $Alb(X)$) by fixing a point $p_0 \in X$, and by associating to each point $p \in X$ the class in $Alb(X)$ of $\int_{p_0}^{p}$, where the integral is taken along any path between p_0 and p.

Recall that, if X is projective (as any surface of general type), $Alb(X)$ is an abelian variety (of dimension q).

The Albanese morphism is a powerful tool for studying *irregular* surfaces ($q > 0$) and in particular:

Definition 4. *A variety X is called of* maximal Albanese dimension *if the image of the Albanese morphism has the same dimension as X.*

This is the general case for surfaces, since otherwise the Albanese morphism is a fibration onto a smooth curve of genus q. We see then that for surfaces maximal Albanese dimension is equivalent to the non existence of a genus q pencil.

We can now state the theorem known as *Severi's inequality*

Special Families of Curves, of Abelian Varieties, and of Certain Minimal Manifolds over Curves

Martin Möller[1], Eckart Viehweg[1] and Kang Zuo[2]*

[1] Universität Duisburg-Essen, Mathematik, 45117 Essen, Germany
`martin.moeller@uni-essen.de`
`viehweg@uni-essen.de`
[2] Universität Mainz, Fachbereich 17, Mathematik, 55099 Mainz, Germany
`kzuo@mathematik.uni-mainz.de`

Introduction

Let $f : X \to Y$ be a surjective morphism from an $n+1$ dimensional complex projective manifold X to a curve Y. In this article we want to present some recent results about the structure of f, in particular about its discriminant locus S, i.e. the set of points s in Y with $f^{-1}(s)$ singular. So

$$f : V = f^{-1}(U) \longrightarrow U = Y \setminus S$$

will always be smooth. We will usually require $\Delta = f^*S$ to be a normal crossing divisor.

If $Y = \mathbb{P}^1$ and if X is a curve of genus $g(X) > 0$ the Hurwitz formula implies that $\#S \geq 3$. In the extremal case $\#S = 3$, X is obviously defined over a number field. G.V. Belyĭ has shown in [Be79] that the existence of f with $\#S = 3$ characterizes curves defined over a number field.

As conjectured by F. Catanese and M. Schneider, the first part can be generalized for $n > 0$.

Theorem 0.1 ([VZ01]). *Let X be a complex projective manifold of Kodaira dimension $\kappa(X) \geq 0$ and let $f : X \to \mathbb{P}^1$ be a surjective morphism. Then f has at least 3 singular fibres.*

It is more reasonable to assume that the general fibre F of f is connected, or as we will say, that $f : X \to Y$ is a family of n-dimensional manifolds, and to use the Hurwitz formula for the Stein factorization to get hold of the general

---

* This work has been supported by the "DFG-Schwerpunktprogramm Globale Methoden in der Komplexen Geometrie", and by the DFG-Leibniz program.

case. Y will be again a curve of arbitrary genus. Putting together results due to Parshin-Arakelov, Migliorini, Kovács, Bedulev-Viehweg, Oguiso-Viehweg and Viehweg-Zuo (see [VZ01] and the references given there) one has:

Theorem 0.2. *Let $f : X \to Y$ be a non-isotrivial family of n-folds, with general fibre F. Assume either*

a. $\kappa(F) = \dim(F)$, or
b. ω_F semiample.

Then $\deg(\Omega^1_Y(\log S)) = 2g(Y) - 2 + \#S > 0$.

Here f is isotrivial, if over some finite covering $Y' \to Y$ the pullback $X \times_Y Y'$ is birational to $F \times Y'$.

As a byproduct, the proof of 0.2 in [VZ01] (see also [Kov02]) gives some explicit lower bounds for the degree of $\deg(\Omega^1_Y(\log S))$. Writing δ for the number of singular fibres of f, which are not semistable, i.e. not reduced normal crossing divisors, one finds constants ν and e, depending only on the Hilbert polynomial h of some polarization of the family, with

$$\frac{\deg(f_*\omega^\nu_{X/Y})}{\operatorname{rank}(f_*\omega^\nu_{X/Y})} \leq (n \cdot (2g(Y) - 2 + \#S) + \delta) \cdot \nu \cdot e.$$

In Section 2 we will use [VZ05, 3.4] to get a similar bound for arbitrary semistable families.

Theorem 0.3. *Assume that $f : X \to Y$ is a semistable family of n-folds. If $Y = \mathbb{P}^1$ assume in addition that $\#S \geq 2$. Then for all $\nu \geq 1$ with $f_*\omega^\nu_{X/Y} \neq 0$*

$$\frac{\deg(f_*\omega^\nu_{X/Y})}{\operatorname{rank}(f_*\omega^\nu_{X/Y})} \leq \frac{n \cdot \nu}{2} \cdot (2g(Y) - 2 + \#S) = \frac{n \cdot \nu}{2} \cdot \deg(\Omega^1_Y(\log S)).$$

The assumption $\#S \geq 2$ is just added to guaranty that $\deg(\Omega^1_Y(\log S))$ is non-negative, and that the fundamental group of $Y \setminus S$ sufficiently large. Of course it will always hold true if one declares one or two smooth fibres to be singular.

There are families of Abelian varieties and of Calabi-Yau manifolds, with

$$\deg(f_*\omega^\nu_{X/Y}) = \frac{\deg(f_*\omega^\nu_{X/Y})}{\operatorname{rank}(f_*\omega^\nu_{X/Y})} = \frac{n \cdot \nu}{2} \cdot \deg(\Omega^1_Y(\log S)).$$

We do not know, whether there are families of manifolds of higher Kodaira dimension with such an equality.

Theorem 0.3 implies Theorem 0.2. In fact, if $\deg(\Omega^1_Y(\log S)) \leq 0$ the open set U is either an elliptic curve or it contains $\mathbb{C}^*$. In the first case the family is smooth, in the second one one finds a new family over an étale covering $\mathbb{C}^* \to \mathbb{C}^*$, which is semistable. For a semistable family $f : X \to Y$ Theorem

0.3 implies that $\deg(f_*\omega^\nu_{X/Y}) \leq 0$. On the other hand, the assumptions a) or b) in Theorem 0.2 imply that for $\nu \geq 2$

$$\deg(f_*\omega^\nu_{X/Y}) \geq 0,$$

and that $\deg(f_*\omega^\nu_{X/Y}) = 0$ if and only if f is isotrivial. The same argument is used to show Theorem 0.1:

If there is a morphism with 2 singular fibres, one uses the Hurwitz formula to show that the Stein factorization of $X \to \mathbb{P}^1$ is $\mathbb{P}^1$, and that the induced morphism has again at most 2 singular fibres. Here the positivity of $\deg(f_*\omega^\nu_{X/Y})$ follows from $\kappa(X) \geq 0$, and as above one obtains a contradiction to the inequality stated in Theorem 0.3.

In addition, the bound in Theorem 0.3 implies some weak analogue of the Shafarevich conjecture on the finiteness of smooth families of curves over a fixed curve U.

Corollary 0.4. *Fix (Y, S) and a polynomial h. Then up to deformation there are only finitely many non-isotrivial families $f : X \to Y$ smooth over $U = Y \setminus S$, such that:*

1. *ω_F is ample with Hilbert polynomial h.*
2. *$\omega_F^\rho = \mathcal{O}_F$ for some $\rho > 0$, and the smooth part $V \to U$ allows a polarization with Hilbert polynomial h.*

In fact, consider the moduli scheme M_h parameterizing either canonically polarized manifolds, or polarized manifolds with a torsion canonical divisor, and with Hilbert polynomial h. In both cases one can show (see [Vi05]) that for a given $\nu \geq 2$ with $h(\nu) \neq 0$ there is a projective compactification $\bar{M}_h$ and for some p an invertible sheaf $\lambda_\nu^{(p)}$ on $\bar{M}_h$ with:

1. $\lambda_\nu^{(p)}$ is numerically effective, and ample with respect to M_h.
2. If $f : X \to Y$ is a semistable family over a curve Y, whose general fibre belongs to the moduli functor, and if $\varphi : Y \to \bar{M}_h$ is the extension of the induced morphism to M_h, then

$$\varphi^*\lambda_\nu^{(p)} = \det(f_*\omega^\nu_{X/Y})^p.$$

Here an invertible sheaf λ is "ample with respect to M_h" if there exists some modification $\tau : \bar{M}'_h \to \bar{M}_h$, some $\mu \gg 1$ and an effective divisor E on $\bar{M}'_h$, supported in $\tau^{-1}(\bar{M}_h \setminus M_h)$, such that $\tau^*\lambda \otimes \mathcal{O}_{\bar{M}'_h}(-E)$ is ample.

So Theorem 0.3 says that for given Y and S the degree of $\varphi^*\lambda_\nu^{(p)}$ is bounded, which implies Corollary 0.4. Using a different argument (see [VZ02, 6.2]) one can extend 0.4 to families of minimal models of arbitrary (non-negative) Kodaira dimension.

As we will see in Section 2 the Theorems 0.1, 0.2 and 0.3 follow from the study of variations of Hodge structures of certain families obtained as cyclic coverings of $X \to Y$.

It remains the question, what is special about families with few singular fibres. Here the only results are for families controlled by their variation of Hodge structures $R^n f_* \mathbb{C}_V$ (or $R^1 f_* \mathbb{C}_V$ for families of Abelian varieties). Moreover, we will assume that the family has semistable reduction, or slightly weaker, that the local monodromies in all $s \in S$ are unipotent. Before stating the results, and before giving "few singular fibres" a precise meaning, we will need some notations.

Let $\mathbb{V} \subset R^k \mathbb{C}_V$ be a $\mathbb{C}$ sub variation of Hodge structures. $\mathbb{V} \otimes_C \mathcal{O}_U$ extends to a locally free sheaf $\mathcal{H}$ on Y in such a way that the Gauß-Manin connection acquires logarithmic singularities. We choose for $\mathcal{H}$ the Deligne extension, i.e. an extension such that the real part of the local residues are zero. The Hodge filtration extends to a holomorphic filtration on $\mathcal{H}$, and the extended Gauß-Manin connection defines on the associated graded bundle the structure of a logarithmic Higgs bundle (F, τ), i.e. a locally free sheaf F together with a collection of maps

$$\tau_{p,q} : F^{p,q} \longrightarrow F^{p-1,q+1} \otimes \Omega^1_Y(\log S), \qquad p+q=k.$$

The bundle maps $\tau_{p,q}$ can be iterated to obtain

$$\tau^{(\ell)} : F^{k,0} \longrightarrow F^{k-1,1} \otimes \Omega^1_Y(\log S) \longrightarrow \cdots \longrightarrow F^{k-\ell,\ell} \otimes S^\ell(\Omega^1_Y(\log S)).$$

Definition 0.5.

i. We call $\tau^{(k)} : F^{k,0} \to F^{0,k} \otimes S^k(\Omega^1_Y(\log S))$ *the Griffiths-Yukawa coupling of* $\mathbb{V}$ *(or of* f *in case* $\mathbb{V} = R^k f_* \mathbb{C}_V$*).*

ii. The Higgs field is strictly maximal, if $F^{k,0} \neq 0$ *and if the* $\tau_{p,q}$ *are all isomorphisms.*

As it will turn out, the property ii) is of numerical nature.

Lemma 0.6 ([VZ05]). *Assume that* $\mathbb{V}$ *is a variation of polarized complex Hodge structures of weight* n *with unipotent local monodromy in all* $s \in S$*, and with logarithmic Higgs bundle* $(\bigoplus F^{p,q}, \tau_{p,q})$*. If* $F^{k,0} \neq 0$ *one has the Arakelov inequality*

$$\deg(F^{k,0}) \leq \frac{k}{2} \cdot \operatorname{rank}(F^{k,0}) \cdot \deg(\Omega^1_Y(\log S)), \tag{1}$$

and (1) is an equality if and only if one has a decomposition $\mathbb{V} = \mathbb{V}_1 \oplus \mathbb{V}_2$ *where the Higgs field of* $\mathbb{V}_1$ *is strictly maximal and where* $\mathbb{V}_2$ *is a variation of polarized complex Hodge structures, zero in bidegree* $(k, 0)$.

For families of curves, the inequality (1) is due to Arakelov, and for families of Abelian varieties it has been shown by Faltings in [Fa83].

If $\#S$ is even, [VZ03, 3.4] gives a more precise description of $\mathbb{V}_1$. Choose a logarithmic theta characteristic, i.e. an invertible sheaf $\mathcal{L}$ with $\mathcal{L}^2 = \Omega^1_Y(\log S)$. The existence of $\mathbb{V}$ in 0.6 implies that $\deg(\Omega^1_Y(\log S)) > 0$. Then the Higgs bundle $\mathcal{L} \oplus \mathcal{L}^{-1}$ with Higgs field

$$\mathcal{L} \xrightarrow[\simeq]{\tau} \mathcal{L}^{-1} \otimes \Omega^1_Y(\log S)$$

is stable. As recalled in Section 1 it must be the Higgs bundle of a polarized variation of Hodge structures $\mathbb{L}$ of weight one. $\mathbb{L}$ is unique up to the tensor product with a unitary rank one local system, induced by a two-division point of $\mathrm{Pic}^0(Y)$.

Lemma 0.7. *Assume that* $\#S$ *is even, and that* $\mathbb{V}_1$ *is a variation of Hodge structures with a strictly maximal Higgs field and with unipotent local monodromies. Then there exists a unitary local system* $\mathbb{T}$ *on* Y*, regarded as a variation of Hodge structures in bidegree* $(0,0)$*, with* $\mathbb{V}_1 = S^k(\mathbb{L}) \otimes_C \mathbb{T}|_U$.

So "special families" or "families with few singular fibres" will be those where the variation of Hodge structures $R^k\mathbb{C}_V$ contains an irreducible sub variation $\mathbb{V}$ with a strictly maximal Higgs field, or equivalently a sub variation $\mathbb{V}$ for whose Higgs bundle the equality

$$\deg(F^{k,0}) = \frac{k}{2} \cdot \mathrm{rank}(F^{k,0}) \cdot \deg(\Omega^1_Y(\log S)) \tag{2}$$

holds. Here $k = n$, except for families of Abelian varieties, where we usually choose $k = 1$.

Let us first consider a semistable family $f : X \to Y$ of Abelian varieties, smooth over $U = Y \setminus S$, and $V = f^{-1}(U)$.

Theorem 0.8 ([VZ04]). *Assume that each irreducible and non-unitary sub variation* $\mathbb{V}$ *of Hodge structures in* $R^1 f_*\mathbb{C}_V$ *has a strictly maximal Higgs field. Then (replacing* U *by an étale covering and* V *by the pullback family)* U *is a Shimura curve of Hodge type, and* $f : V \to U$ *the corresponding universal family.*

The construction of Shimura curves will be recalled in Section 3. Let us just mention here, that a Shimura curve U is an étale covering of a certain moduli space of Abelian varieties with prescribed Mumford-Tate group and a suitable level structure. So whenever we talk about Shimura curves, this is a property up to étale coverings of U. This allows in particular to assume that $\#S$ is even.

As we will recall in Section 4 the converse of Theorem 0.8 was shown in [Moe3]:

Theorem 0.9. *If* $f : V \to U$ *is the universal family over a Shimura curve its Higgs field is strictly maximal.*

Because of Kollár's decomposition [Kol87] (see Lemma 1.9 below) one can write $f_*\Omega^1_{X/Y}(\log S)$ as a direct sum of an ample sheaf $\mathcal{A}$ and a subsheaf $\mathcal{B}$, flat for the Gauß-Manin connection. Correspondingly on the local system side one has a decomposition $R^1 f_*\mathbb{C}_V = \mathbb{W} \oplus \mathbb{U}$, where $\mathbb{U}$ is unitary and invariant under complex conjugation, and where the $(1,0)$ component of $\mathbb{W}$ is ample.

Hence the Higgs bundle corresponding to $\mathbb{W}$ is of the form $(\mathcal{A} \oplus \mathcal{A}^\vee, \theta)$ and using 0.6 one can restate the Theorem 0.8 as:

If $\deg(\mathcal{A}) = \frac{1}{2} \cdot \mathrm{rank}(\mathcal{A}) \cdot \deg(\Omega^1_Y(\log S))$, *then* U *is a Shimura curve of Hodge type, and* $f : V \to U$ *the corresponding universal family.*

The Lemma 0.7 implies that the Arakelov equality forces $\mathbb{W} = \mathbb{L} \otimes_{\mathbb{C}} \mathbb{T}$ for a unitary system $\mathbb{T}$. In particular, if $S \neq \emptyset$, replacing Y by an étale covering, one can assume that $\mathbb{T} = \mathbb{C}_U^{\oplus r}$ and $\mathbb{U} = \mathbb{C}_U^{\oplus 2 \cdot n - 2 \cdot r}$. As in [VZ04] one deduces the next Corollary.

Corollary 0.10. *If* $S \neq \emptyset$ *consists of an even number of points, and if*

$$\deg(\mathcal{A}) = \frac{1}{2} \cdot \mathrm{rank}(\mathcal{A}) \cdot \deg(\Omega^1_Y(\log S)),$$

then (replacing Y *by an étale covering)* $f : X \to Y$ *is isogenous to*

$$E \times_Y \cdots \times_Y E \times B,$$

where B *is an Abelian variety of dimension* $n - r$ *and where* $E \to Y$ *is a modular family of elliptic curves.*

The proof of Corollary 0.10 is easy if one assumes that the maximal unitary subsystem of $R^1 f_* \mathbb{C}_V$ is defined over $\mathbb{Q}$, and it will be sketched in Section 3. This condition holds true if $S \neq \emptyset$. For $S = \emptyset$ the proof of Theorem 0.8 in [VZ04] also gives back the known classification of Shimura curves.

As a supplement to 0.8 we will show in Section 3:

Corollary 0.11. *If the maximal unitary local subsystem* $\mathbb{U}$ *of* $R^1 f_* \mathbb{C}_V$ *is trivial, then the family* $f : V \to U$ *is rigid. In particular it is defined over* $\bar{\mathbb{Q}}$.

Let us consider next families of curves of genus $g \geq 2$, and let us return to the local systems $\mathbb{L}$, defined by logarithmic theta characteristics $\mathcal{L}$. Obviously $\mathbb{L}$ has a strictly maximal Higgs field.

Theorem 0.12 ([Moe2]). *Let* $f : X \to Y$ *be a semistable family of curves of genus* $g \geq 2$, *smooth over* $U = Y \setminus S$, *and* $V = f^{-1}(U)$.

1. $R^1 f_* \mathbb{C}_V$ *contains a rank two sub variation of Hodge structures* $\mathbb{L}$ *with a strictly maximal Higgs field if and only if* U *is a Teichmüller curve, and* $V \to U$ *the corresponding universal family.*
2. *Teichmüller curves are defined over number fields.*

The definition of a Teichmüller curve will be given in Section 4. Roughly speaking, one considers geodesics in the Teichmüller space, constructed by an $\mathrm{Sl}(2, \mathbb{R})$-action on the real and imaginary part of a given holomorphic differential form. If the quotient by a suitable lattice in $\mathrm{Sl}(2, \mathbb{R})$ is an algebraic curve, it is called a Teichmüller curve in M_g. As for Shimura curves, we always allow ourselves to replace the lattice by a smaller one, hence the Teichmüller curve by an étale cover. We say that $f : V \to U$ is the universal family if the morphism $U \to M_g$ is induced by f.

It is a striking fact, that in spite of the differential geometric definition of Teichmüller curves, they can be characterized analytically, or even by a numerical property of the Higgs bundle. This has a number of applications in the theory of Teichmüller curves (see [Moe2], [Moe3] and [BM05]), and it allows to construct new examples of such special curves in M_g.

At the same time, Theorem 0.12 allows to use differential geometric properties of Teichmüller curves to study variations of Hodge structures for families of curves. For example, at the present moment the only known proof of the next Corollary relies on Theorem 0.12, and on differential geometric methods.

Corollary 0.13. *In Theorem 0.12 the variation of Hodge structures $R^1 f_* \mathbb{C}_V$ can not contain two different sub variation of Hodge structures $\mathbb{L}_1$ and $\mathbb{L}_2$, both with a strictly maximal Higgs field.*

There remains the question whether there are higher rank irreducible sub variations of Hodge structures $\mathbb{V}$ with a strictly maximal Higgs field. By Lemma 0.7 those are of the form $\mathbb{L} \otimes_{\mathbb{C}} \mathbb{T}$, where $\mathbb{T}$ is an irreducible unitary system, at least after replacing U by an étale covering of degree two. As we will see in Section 1 $\operatorname{rank}(\mathbb{T}) \geq 2$ implies that $S = \emptyset$, and squeezing the methods used to prove 0.2 a bit more, one can show that this is not possible (see [VZ05]). So putting everything together one obtains:

Corollary 0.14. *Let $f : X \to Y$ be a semistable family of curves of genus $g \geq 2$, smooth over $U = Y \setminus S$. Then $R^1 f_* \mathbb{C}_V$ does not contain a sub variation of Hodge structures $\mathbb{V}$ with a strictly maximal Higgs field and with* $\operatorname{rank}(\mathbb{V}) > 2$.

If $S = \emptyset$ then $R^1 f_* \mathbb{C}_V$ does not contain any sub variation of Hodge structures $\mathbb{V}$ with a strictly maximal Higgs field.

As we will see in Sect. 4, the rank two local subsystem $\mathbb{L} \subset R^1 f_* \mathbb{C}_V$ can be defined over a number field K, i.e. it is of the form $\mathbb{L}_K \otimes_K \mathbb{C}$ for a local subsystem $\mathbb{L}_K \in R^1 f_* K_V$, but in general it will not be defined over $\mathbb{Q}$. Replacing K by its Galois hull, for $\sigma \in \operatorname{Gal}(K/\mathbb{Q})$ the conjugate local system $\mathbb{L}^\sigma$ can only have a strictly maximal Higgs field, if $\mathbb{L} = \mathbb{L}^\sigma$. The defect in the inequality (1) for $\mathbb{L}^\sigma$ has been studied in [BM05].

In general things are changing, if one replaces $f : X \to Y$ by its family of Jacobians. Passing from a family of curves to its Jacobian, one might replace the discriminant locus S by a subset S'. Singular fibres with a compact Jacobian, for example those with two components meeting in one point, are not contributing to the discriminant locus of the family of Jacobians. If $f : V \to U$ is the universal family over a Teichmüller curve, then the existence of the local subsystem $\mathbb{L}$ together with the inequality (1) applied to the family of Jacobians, shows that $S' = S$, hence that all singular fibres of f have a non-compact Jacobian. One can ask, whether there are any Teichmüller curves

$U \subset M_g$ in the moduli space of curves of genus g which are Shimura curves in the moduli space A_g of Abelian varieties (with a suitable level structure). By [Moe3] the answer is "no" for $g \geq 4$ and for $g = 2$, whereas for $g = 3$ there exists one example, essentially unique.

On the other hand, by the Corollaries 0.13 and 0.14 we know, that the only curves $U \subset M_g$, which are Shimura curves in A_g have to be induced by a family $f : X \to Y$ with $R^1 f_* \mathbb{C}_V = \mathbb{L} \oplus \mathbb{U}$ with $\operatorname{rank}(\mathbb{L}) = 2$ and with $\mathbb{U}$ unitary. This implies that U is a Teichmüller curve, and that U is not projective.

Theorem 0.15.

a. For $g \geq 2$ the moduli space M_g does not contain any compact Shimura curve.

b. M_g contains a non-compact Shimura curve U if and only if $g = 3$. The Shimura curve $U \in M_3$ is a Teichmüller curve, essentially unique.

In this article, we will explain the main ingredients used in the introduction, and we will sketch the proofs of some of the results mentioned above. In Section 1 we explain the Simpson correspondence for variations of Hodge structures over curves, and we will prove Lemma 0.6 and Lemma 0.7 for $k = 1$. For $k > 1$, Lemma 0.6 will be shown under the additional assumption that $\operatorname{rank}(F^{k,0}) = 1$.

In Section 2 we prove Theorem 0.3, and we give some hints, how one obtains Corollary 0.14 from Corollary 0.13.

The characterization of Shimura curves will be discussed in Section 3. We prove Theorem 0.8 for families where $R^1 f_* \mathbb{C}_V$ itself has a strictly maximal Higgs field, hence we exclude unitary direct factors. We sketch the proofs of Corollaries 0.10 and 0.11 and we also reproduce the proof of Theorem 0.9 from [Moe3].

Section 4 gives an introduction to the theory of Teichmüller curves. In particular we sketch the proof of Theorem 0.12 and of the Corollary 0.13.

Finally the non-existence of Teichmüller curves in M_g, for $g = 2$ and $g \geq 4$ and the proof of Theorem 0.15, b), will be discussed in Section 5.

1 Higgs Bundles over Curves and Arakelov Inequalities

We will frequently use C. Simpson's correspondence between polystable logarithmic Higgs bundles of degree zero and representations of the fundamental group $\pi_1(U, *)$. Recall that a logarithmic Higgs bundle is a locally free sheaf E on Y together with an $\mathcal{O}_Y$ linear morphism $\theta : E \to E \otimes \Omega^1_Y(\log S)$ with $\theta \wedge \theta = 0$. The usual definitions of stability (and semistability) for locally free sheaves extend to Higgs bundles, by requiring that

$$\mu(F) = \frac{\deg(F)}{\operatorname{rank}(F)} < \mu(E) = \frac{\deg(E)}{\operatorname{rank}(E)}$$

(or $\mu(F) \leq \mu(E)$) for all subsheaves F with $\theta(F) \subset F \otimes \Omega^1_Y(\log S)$.

Theorem 1.1 (C. Simpson [Si90]). *There exists a natural equivalence between the category of direct sums of stable filtered regular Higgs bundles of degree zero, and of direct sums of stable filtered local systems of degree zero.*

We will not recall the definition of a "filtered regular" Higgs bundle [Si90, page 717], and just remark that for a Higgs bundle corresponding to a local system $\mathbb{V}$ with unipotent monodromy around the points in S the filtration is trivial, and automatically $\deg(\mathbb{V}) = 0$.

In general it is impossible to describe the Higgs bundle (E, θ) explicitly in terms of the corresponding local system $\mathbb{V}$, with two exceptions:

Example 1.2. If $\mathbb{T}$ is a unitary local system on U with unipotent monodromy operators in $s \in S$, then the corresponding Higgs bundle is the Deligne extension N of $N_0 = \mathbb{T} \otimes \mathcal{O}_U$ and the Higgs field θ is the zero map.

The second example already occurred in the Introduction.

Example 1.3. Let $\mathbb{V}$ be a polarized $\mathbb{C}$ variation of Hodge structures of weight k and with unipotent local monodromy operators. The $\mathcal{F}$-filtration on $F_0 = \mathbb{V} \otimes_{\mathbb{C}} \mathcal{O}_U$ extends to a locally splitting filtration on the Deligne extension F

$$\mathcal{F}^{k+1} \subset \mathcal{F}^k \subset \cdots \subset \mathcal{F}^0.$$

We will usually assume that $\mathcal{F}^{k+1} = 0$ and $\mathcal{F}^0 = F$. The Griffiths transversality condition for the Gauß-Manin connection ∇ says that

$$\nabla(\mathcal{F}^p) \subset \mathcal{F}^{p-1} \otimes \Omega^1_Y(\log S),$$

and hence ∇ induces a $\mathcal{O}_Y$ linear map

$$\theta_{p,k-p} : F^{p,k-p} = \mathcal{F}^p/\mathcal{F}^{p+1} \longrightarrow F^{p-1,k-p+1} = \mathcal{F}^{p-1}/\mathcal{F}^p \otimes \Omega^1_Y(\log S).$$

So

$$\big(F = \bigoplus_p F^{p,k-p}, \theta = \bigoplus \theta_{p,k-p}\big)$$

is a Higgs bundle, and it is the image of $\mathbb{V}$ under the Simpson correspondence in Theorem 1.1.

Properties 1.4. Let $\mathbb{V}$ be a direct sum of irreducible $\mathbb{C}$ local system and let (F, θ) be the corresponding Higgs bundle.

1. $F = \bigoplus_{p+q=k} F^{p,q}$ with $\theta(F^{p,q}) \subset F^{p-1,q+1} \otimes \Omega^1_Y(\log S)$ if and only if $\mathbb{V}$ is a local system underlying a complex polarized variation of Hodge structures.
2. Under the equivalent conditions in 1) let $\mathbb{T}$ be a local subsystem of $\mathbb{V}$ with Higgs bundle

$$\big(N = \bigoplus_p N^{p,k-p}, 0\big) \ \subset \ \big(F = \bigoplus_p F^{p,k-p}, \theta = \bigoplus \theta_{p,k-p}\big).$$

Then $\mathbb{T}$ is unitary.

Proof. The condition on (F,θ) in 1) is just saying, that it is a system of Hodge bundles, as defined in [Si88]. So the first part follows from [Si88] and [Si90].

Let $\Theta(N,h)$ denote the curvature of the Hodge metric h on F restricted to N, then by [Gr84], chapter II we have

$$\Theta(N,h|_N) = -\theta_N \wedge \bar{\theta}_N - \bar{\theta}_N \wedge \theta_N = 0.$$

This means that $h|_N$ is a flat metric. Hence, $\mathbb{T}$ is a unitary local system. □

The Simpson correspondence does not say anything about the field of definition for $\mathbb{V}$. Here we say that $\mathbb{V}$ is defined over a subfield K of $\mathbb{C}$ if there is a K-local system $\mathbb{V}_K$ with $\mathbb{V} = \mathbb{V}_K \otimes_K \mathbb{C}$. In different terms, for $\mu = \mathrm{rank}(\mathbb{V})$ the representation

$$\gamma_{\mathbb{V}} : \pi_0(U,*) \longrightarrow \mathrm{Gl}(\mu,\mathbb{C})$$

is conjugate to one factoring like

$$\gamma_{\mathbb{V}} : \pi_0(U,*) \longrightarrow \mathrm{Gl}(\mu,K) \longrightarrow \mathrm{Gl}(\mu,\mathbb{C}).$$

If $\mathbb{V}$ is defined over K, and if $\sigma : K \to K'$ is an isomorphism, we will write $\mathbb{V}_K^\sigma$ for the local system defined by

$$\gamma_{\mathbb{V}} : \pi_0(U,*) \longrightarrow \mathrm{Gl}(\mu,K) \xrightarrow{\sigma} \mathrm{Gl}(\mu,K'),$$

and $\mathbb{V}^\sigma = \mathbb{V}_K^\sigma \otimes_{K'} \mathbb{C}$.

By [De71] (for $K = \mathbb{Q}$) and by [De87] the category of polarized K-variations of Hodge structures is semisimple. By [De87, Proposition 1.13] one has:

Lemma 1.5. *A local system $\mathbb{V}$, underlying a polarized variation of Hodge structures, decomposes as*

$$\mathbb{V} = \bigoplus_{i=1}^{r} (\mathbb{V}_i \otimes W_i),$$

where $\mathbb{V}_i$ are pairwise non-isomorphic irreducible $\mathbb{C}$-local systems and W_i are non-zero $\mathbb{C}$-vector spaces.

Moreover the $\mathbb{V}_i$ and the W_i carry polarized variations of Hodge structures, whose tensor product and sum gives back the Hodge structure on $\mathbb{V}$. The Hodge structure on the $\mathbb{V}_i$ (and W_i) is unique up to a shift of the bigrading.

Suppose that $\mathbb{W}$ is a local system defined over a number field L. The local system $\mathbb{W}_L$ is given by a representation $\rho : \pi_1(U,*) \to \mathrm{Gl}(W_L)$ for the fibre W_L of $\mathbb{W}_L$ over the base point $*$.

Fixing a positive integer $r < \mu$ let $\mathcal{G}(r,\mathbb{W})$ denote the set of all rank-r local subsystems of $\mathbb{W}$ and let $\mathrm{Grass}(r,W_L)$ be the Grassmann variety of r-dimensional subspaces. Then $\mathcal{G}(r,\mathbb{W})$ is the subvariety of

$$\mathrm{Grass}(r,W_L) \times_{\mathrm{Spec}(L)} \mathrm{Spec}(\mathbb{C})$$

consisting of the $\pi_1(U,*)$-invariant points. In particular, it is a projective variety defined over L. A K-valued point of $\mathcal{G}(r,\mathbb{W})$ corresponds to a local subsystem of $\mathbb{W}_K = \mathbb{W}_L \otimes_L K$. One obtains the following well known property.

Lemma 1.6. *If* $[\mathbb{V}] \in \mathcal{G}(r,\mathbb{W})$ *is an isolated point, then* $\mathbb{V}$ *is defined over* $\bar{\mathbb{Q}}$.

Lemma 1.7. *Let* $\mathbb{W}$ *be a polarized variation of Hodge structures defined over* L*, and let* $\mathbb{V} \subset \mathbb{W}$ *be an irreducible local subsystem of rank* r *defined over* $\mathbb{C}$,. *Then* $\mathbb{V}$ *can be deformed to a local subsystem* $\mathbb{V}_t \subset \mathbb{W}$*, which is isomorphic to* $\mathbb{V}$ *and which is defined over a finite extension of* L.

Proof. By Lemma 1.5 $\mathbb{W}$ is completely reducible over $\mathbb{C}$. Hence we have a decomposition $\mathbb{W} = \mathbb{V} \oplus \mathbb{V}'$.

The space $\mathcal{G}(r,\mathbb{W})$ of rank r local subsystems of $\mathbb{W}$ is defined over L and the subset

$$\{\mathbb{V}_t \in \mathcal{G}(r,\mathbb{W});\ \text{the composite } \mathbb{V}_t \subset \mathbb{V} \oplus \mathbb{V}' \xrightarrow{pr_1} \mathbb{V} \text{ is non zero }\}$$

forms a Zariski open subset. So there exists some $\mathbb{V}_t$ in this subset, which is defined over some finite extension of L. Since $p : \mathbb{V}_t \to \mathbb{V}$ is non zero, since $\mathrm{rank}(\mathbb{V}_t) = \mathrm{rank}(\mathbb{V})$, and since $\mathbb{V}$ is irreducible, p is an isomorphism. □

The Lemmata 1.6 and 1.7 are just the starting points to show, that certain local subsystems of $R^k f_*\mathbb{C}_V$ are defined over number fields (see [VZ04]). Let us give a typical example:

Lemma 1.8. *Let* $\mathbb{W}$ *be defined over a (real) number field, and let* $\mathbb{W} = \mathbb{V} \oplus \mathbb{U}$ *be a decomposition such that the Higgs field of* $\mathbb{V}$ *is maximal, and such that* $\mathbb{U}$ *is unitary.*

1. *Then* $\mathbb{V}$ *and* $\mathbb{U}$ *are defined over a (real) number field, as well as the decomposition.*
2. *If* $S \neq \emptyset$*, then* $\mathbb{U}$ *is defined over* $\mathbb{Q}$ *and it trivializes over an étale covering of* Y.

Proof. Consider a family $\mathbb{V}_t$ of local subsystems of $\mathbb{W}$ for t in a small disk Δ, and with $\mathbb{V}_0 = \mathbb{V}$. Since the Higgs field of $\mathbb{V}_0$ is maximal, one may assume that the one of $\mathbb{V}_t$ is maximal for all $t \in \Delta$. Then the projection $\mathbb{V}_t \to \mathbb{U}$ must be zero. Otherwise, the complete reducibility of local systems coming from variations of Hodge structures in 1.5 implies that $\mathbb{V}_t$ and $\mathbb{U}$ have a common factor, obviously a contradiction.

So $\mathbb{V}$ is rigid as a local subsystem, hence defined over $\bar{\mathbb{Q}}$. The same argument works for $\mathbb{U}$ instead of $\mathbb{V}$.

If $\mathbb{W}$ is defined over $\mathbb{R}$, then $\bar{\mathbb{V}} \to \mathbb{U}$ again has to be the zero map. So $\mathbb{V}$ and $\mathbb{U}$ have to be defined over $\mathbb{R}$.

Assume now that $S \neq \emptyset$. By (1) we know that the decomposition is defined over a number field. For the local monodromy operators "unitary and unipotent" implies that the nilpotent part of the monodromy is zero. This is invariant under conjugation by $\sigma \in \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$, and $\mathbb{U}^\sigma = \mathbb{U}$.

Finally, since $\mathbb{U}$ is a local subsystem of a variation of Hodge structures, it carries a $\mathbb{Z}$-structure. Then the image of the corresponding monodromy representation is finite. □

Lemma 1.5 together with Theorem 1.1 implies that the Higgs bundle (F, θ) of a variation of Hodge structures $\mathbb{V}$ is polystable (as a Higgs bundle) and of degree zero. As a first application one obtains Kollár's decomposition of the sheaf $F^{k,0}$, mentioned in the introduction.

Lemma 1.9. *Let* $(F = F^{k,0} \oplus \cdots \oplus F^{0,k}, \theta)$ *be the logarithmic Higgs bundle of a variation of Hodge structures* $\mathbb{V}$ *with unipotent local monodromies. Then* $F^{k,0} = \mathcal{A} \oplus \mathcal{B}$ *with* $\mathcal{A}$ *ample and with* $\mathcal{B}$ *flat.*

Proof. Replacing $\mathbb{V}$ by one of its irreducible direct factors, one has to show that $F^{k,0}$ is ample, provided that $\theta \neq 0$.

A quotient sheaf $\mathcal{N}$ of $F^{k,0}$ gives rise to a quotient Higgs bundle $(\mathcal{N}, 0)$ of (F, θ), hence 1.1 implies that $\deg(\mathcal{N}) \geq 0$. If $\deg(\mathcal{N}) = 0$, then it corresponds to a local subsystem, which is excluded by the irreducibility of $\mathbb{V}$. Hence all quotients of $F^{k,0}$ have strictly positive degree, which implies that $F^{k,0}$ is ample. □

Let us consider for a moment the case $k = 1$ and $\mathbb{V} = R^1 f_* \mathbb{C}_V$. Then $(\mathcal{B} \oplus \mathcal{B}^\vee, 0)$ is a sub Higgs bundle, hence by Lemma 1.4, 2), it corresponds to a unitary subbundle of $\mathbb{V}$. The latter is defined over $\mathbb{Q}$ and by Lemma 1.8, (1), this decomposition is defined over $\bar{\mathbb{Q}} \cap \mathbb{R}$.

The Arakelov inequality stated in Lemma 0.6 follows from the polystability of the Higgs bundles (see [VZ03] and [VZ05]). We will indicate the proof just in two simple cases, the one of weight one variations of Hodge structures, and the one where the $(k, 0)$ part is one dimensional.

Proof of Lemma 0.6 for weight one. Let $\mathcal{A} \subset F^{1,0}$ be a subsheaf, and let $\mathcal{C} \otimes \Omega^1_Y(\log S)$ be its image under $\theta_{1,0}$. Then $\mathcal{A} \oplus \mathcal{C}$ is a Higgs subbundle of $F^{1,0} \oplus F^{0,1}$, and 1.1 implies that $\deg(\mathcal{A}) + \deg(\mathcal{C}) \leq 0$. Since $(\ker(\theta|_{\mathcal{A}}), 0)$ is a sub Higgs bundle, one has

$$\begin{aligned}\deg(\mathcal{A}) &\leq \deg(\mathcal{C}) + \mathrm{rank}(\mathcal{C}) \cdot \deg(\Omega^1_Y(\log S)) \\ &\leq \deg(\mathcal{C}) + \mathrm{rank}(\mathcal{A}) \cdot \deg(\Omega^1_Y(\log S)) \\ &\leq -\deg(\mathcal{A}) + \mathrm{rank}(\mathcal{A}) \cdot \deg(\Omega^1_Y(\log S)), \end{aligned} \tag{3}$$

and

$$\frac{\deg(\mathcal{A})}{\mathrm{rank}(\mathcal{A})} \leq \frac{1}{2} \deg(\Omega^1_Y(\log S)).$$

For $\mathcal{A} = F^{1,0}$ one obtains the inequality (1) in 0.6.

If (1) is an equality, all the inequalities in (3) are equalities. This implies that $\mathrm{rank}(\mathcal{A}) = \mathrm{rank}(\mathcal{C})$, hence that $\theta|_{\mathcal{A}}$ is injective. Furthermore $\deg(\mathcal{A}) + \deg(\mathcal{C}) = 0$, hence $\mathcal{A} \oplus \mathcal{C}$ corresponds to a sub-local system of $\mathbb{V}$ with a maximal Higgs field. □

In fact we have shown a bit more. Assume that (1) in Lemma 0.6 is an equality. If $\mathbb{V}$ is irreducible

$$\frac{\deg(\mathcal{A})}{\mathrm{rank}(\mathcal{A})} < \frac{1}{2}\deg(\Omega^1_Y(\log S)) = \frac{\deg(F^{1,0})}{g_0},$$

except for $\mathcal{A} = F^{1,0}$. Hence $F^{1,0}$ is stable. By duality one obtains the same for $F^{0,1}$.

If $\mathbb{V}$ is not irreducible, applying this argument to all the direct factors of $\mathbb{V}$ one finds:

Lemma 1.10. *If $(F^{1,0} \oplus F^{0,1}, \theta)$ is the logarithmic Higgs bundle of a variation of Hodge structures $\mathbb{V}$ with unipotent local monodromies and with a strictly maximal Higgs field, then the sheaves $F^{1,0}$ and $F^{0,1}$ are polystable and the Higgs field $\theta : F^{1,0} \to F^{0,1} \otimes \Omega^1_Y(\log S)$ is a morphism between polystable sheaves of the same slope.*

Although in this survey we only consider families over curves, let us make one remark about families over a higher dimensional basis. One can hope that the polystability of the Hodge bundles $F^{1,0}$ and $F^{0,1}$ is again enforced by a numerical condition on the degrees with respect to $\omega_Y(S)$, and that it forces $U = Y \setminus S$ to be a Shimura variety. Some first results in this direction have been obtained in [VZ05b].

Proof of Lemma 0.7 for $k = 1$. By assumption $\#S$ is even, so we can choose a logarithmic theta characteristic $\mathcal{L}$ and an isomorphism

$$\tau : \mathcal{L} \to \mathcal{L}^{-1} \otimes \Omega^1_Y(\log S),$$

giving us the $\mathbb{C}$-variation of Hodge structures $\mathbb{L}$. By Lemma 1.10, the sheaf $\mathcal{T} = F^{1,0} \otimes \mathcal{L}^{-1}$ is polystable of degree zero, hence isomorphic to $\mathbb{T} \otimes \mathcal{O}_Y$ for a unitary local system $\mathbb{T}$. The isomorphism

$$\mathcal{T} \otimes \mathcal{L} = F^{1,0} \xrightarrow[\simeq]{\theta} F^{0,1} \otimes \Omega^1_Y(\log S) \xrightarrow[\simeq]{} F^{0,1} \otimes \mathcal{L}^2$$

induces an isomorphism $\phi : \mathcal{T} \otimes \mathcal{L}^{-1} \xrightarrow{\simeq} F^{0,1}$, such that $\theta = \phi \circ (\mathrm{id}_{\mathcal{T}} \otimes \tau)$. Hence the Higgs bundles $(F^{1,0} \oplus F^{0,1}, \theta)$ and $(\mathcal{T} \otimes (\mathcal{L} \oplus \mathcal{L}^{-1}), \mathrm{id}_{\mathcal{T}} \otimes \tau)$ are isomorphic, and $\mathbb{V} \simeq \mathbb{T} \otimes_{\mathbb{C}} \mathbb{L}$. □

Proof of Lemma 0.6 under the additional assumption $\mathrm{rank}(F^{k,0}) = 1$. For later use, let us consider a slightly more general situation:

$\mathbb{V}$ is a variation of polarized complex Hodge structures and with unipotent local monodromy in $s \in S$, and with logarithmic Higgs bundle $(\bigoplus F^{k,0}, \theta)$. Let $\mathcal{H} \subset F^{k,0}$ be an invertible subsheaf, and let $E \subset F$ be the sub Higgs bundle generated by $\mathcal{H}$. Writing $T_Y(-\log S)$ for the dual of $\Omega^1_Y(\log S)$, for some q_0

$$E^{k-q,q} = F^{k-q,q} \cap E = \begin{cases} \mathcal{H} \otimes T_Y(-\log S)^q & \text{for } \ q \leq q_0 \\ 0 & \text{for } \ q > q_0. \end{cases} .$$

Theorem 1.1 implies that

$$\deg(E) = (q_0 + 1) \cdot \deg(\mathcal{H}) - \frac{q_0 \cdot (q_0 + 1)}{2} \deg(\Omega^1_Y(\log S)) \leq 0,$$

hence

$$\deg(\mathcal{H}) \leq \frac{q_0}{2} \deg(\Omega^1_Y(\log S)) \leq \frac{k}{2} \deg(\Omega^1_Y(\log S)). \tag{4}$$

In Lemma 0.6 we choose $\mathcal{H} = F^{k,0}$, and we obtain the Arakelov inequality. If this is an equality, then $q_0 = k$ and E corresponds to a local subsystem, by construction with a strictly maximal Higgs field. □

2 Coverings and Bounds for Subsheaves of the Direct Images

Let us recall from [VZ05] the proof of the following proposition, as a first application of the theory of Higgs bundles.

Proposition 2.1. *Let $f : X \to Y$ be a semistable family of n-folds over a curve Y, and smooth over $U = Y \setminus S$. If $Y = \mathbb{P}^1$ assume that $\#S \geq 2$, and consider for $\nu \geq 1$ an invertible subsheaf $\mathcal{H}$ of $f_*\omega^{\nu}_{X/Y}$. Then*

$$\deg \mathcal{H} \leq \frac{n \cdot \nu}{2} \cdot \deg(\Omega^1_Y(\log S)).$$

Before sketching the proof of the proposition, let us show that it implies Theorem 0.3.

Proof of Theorem 0.3. Since either $g(Y) \geq 1$, or $Y = \mathbb{P}^1$ and $\#S \geq 2$, one can always find a covering $Y' \to Y$, which is unramified over U and which has a prescribed ramification order in $s \in S$. Since f is semistable, the inequality in Theorem 0.3 is compatible with replacing Y by Y' and f by a desingularization of the pullback family.

For $r = \operatorname{rank}(f_*\omega^{\nu}_{X/Y})$ consider the family $f^{(r)} : X^{(r)} \to Y$, obtained as a desingularization of the total space of the r fold product

$$X^r = X \times_Y \cdots \times_Y X \longrightarrow Y.$$

Remark that X^r is normal, Gorenstein with at most rational singularities. Hence flat base change implies that

4 Teichmüller Curves and their Variation of Hodge Structures

We start with the assumptions of Theorem 0.12: Let $f : X \to Y$ be a semistable family of curves of genus $g \geq 2$, smooth over $U = Y \setminus S$ and let $V = f^{-1}(U)$. Suppose that $R^1 f_* \mathbb{C}_V$ contains a sub variation of Hodge structures $\mathbb{L}$ of rank two with a strictly maximal Higgs field. Let us for the moment admit the following:

Lemma 4.1. $\mathbb{L}$ *is defined over* $\mathbb{R}$

By Theorem 0.2 the universal covering of U is the upper half plane $\mathcal{H}$. Let $\widetilde{\varphi} : \mathcal{H} \to \mathcal{H}$ the period map for $\mathbb{L}$. The strict maximality of the Higgs field is now equivalent to $\widetilde{\varphi}$ being an isomorphism (compare [VZ04] Lemma 2.1). In different terms $\widetilde{\varphi}$ is an isometry for the Kobayashi metric on $\mathcal{H}$. The choice of a Teichmüller marking (see below) on one of the smooth fibres of f defines a lift of the moduli map $m : U \to M_g$ to Teichmüller space. We denote this lift by $\widetilde{m} : \mathcal{H} \to T_g$ and we let $j : T_g \to \mathcal{H}_g$ the natural map to the Siegel half space, which associates to the curve its Jacobian and to the Teichmüller marking the choice of a symplectic basis. Finally we let $p_{11} : \mathcal{H}_g \to \mathcal{H}$ the projection of the matrix in $\mathcal{H}_g$ to its $(1,1)$-entry. If we have chosen the symplectic basis B suitably, i.e. such that the first pair in B spans the fibres of $\mathbb{L}$, then

$$\widetilde{\varphi} = p_{11} \circ j \circ \widetilde{m}.$$

Since the composite map $\widetilde{\varphi}$ is a Kobayashi isometry, also $\widetilde{m}$ is an isometry, if we provide the image with the restriction of the Kobayashi metric on T_g. By a theorem of Royden the Kobayashi metric on T_g coincides with the Teichmüller metric. To sum up:
The image of U in M_g is an algebraic curve, whose lift in T_g is a complex geodesic for the Teichmüller metric. These objects are called *Teichmüller curves*. We have thus proved one implication of Theorem 0.12.1.

Before showing the converse implication we need some background on geodesics for the Teichmüller metric. Details may be found in [IT92].
One of the equivalent definitions of Teichmüller space is the following. Let T_g be the space of pairs (R, h) of a Riemann surface plus a quasi-conformal mapping $h : R_0 \to R$ up to isotopy (the *Teichmüller marking*) from a 'reference surface' R_0 to R. By Teichmüller's theorem the isotopy class of h contains a unique representative with minimal maximal dilatation. Its Beltrami coefficient (a $(-1,1)$-form) $\mu = h_{\overline{z}}/h_z$ is of the form $\mu = t|q|/q$ for some quadratic differential $q \in \Gamma(R_0, (\Omega^1_{R_0})^{\otimes 2})$ and $|t| < 1$. Conversely, one can solve the Beltrami equation for any μ of with $|\mu|_\infty < 1$. In particular for each $\mu = t|q|/q$ there is a quasi-conformal map $h : R_0 \to R$ for some Riemann surface R whose Beltrami coefficient is μ.

A quadratic differential q on a Riemann surface R_0 determines an atlas of R minus the zero set of q_0 such that the transition functions are translations maybe composed by multiplication with ±1: The charts $\psi_i : U_i \to \mathbb{C}$ are given by integrating by a square root of q that exists locally. The condition for h to have minimal maximal dilatation may be rephrased by saying that h is affine in the charts ψ_i for some q on R_0 and some ('terminal') quadratic differential on R. The Teichmüller geodesic from id to h is given by the conformal maps with Beltrami coefficient $ct|q|/q$ for $c \in [0,1]$. Thus for each q, the set

$$\{\mu_q(t) := t|q|/q,\ |t| < 1\}$$

defines totally geodesic subset of T_g.

We briefly recall the construction of the complex structure on T_g that makes $T_g \to M_g$ holomorphic. Given (R,h) or rather its Beltrami coefficient μ on R_0, pull it back to universal covering $\mathcal{H}$ of R_0 and extend it by zero to obtain a Beltrami coefficient $\tilde{\mu}$ on $\mathbb{C}$. Solving the Beltrami equation for μ gives a quasi-conformal mapping $\tilde{h} : \mathbb{C} \to \mathbb{C}$, conformal on the lower half plane $\mathcal{H}^*$. The Schwarzian derivative of $\tilde{h}|_{\mathcal{H}^*}$ is a quadratic differential on $\mathcal{H}^*$. Moreover, since μ was pulled back from $R = \mathcal{H}/\Gamma_R$, the Schwarzian derivative descends from $\mathcal{H}^*$ to a quadratic differential

$$q^*_\mu \in H^0(\mathcal{H}^*/\Gamma_R, (\Omega^1_{\mathcal{H}^*/\Gamma_R})^{\otimes 2}).$$

The map $\mu \mapsto q^*_\mu$ factors through Teichmüller space and the complex structure on $H^0(\mathcal{H}^*/\Gamma_R, (\Omega^1_{\mathcal{H}^*/\Gamma_R})^{\otimes 2})$ is the one we want.

The map $\mu \mapsto q^*_\mu$ has a section locally around 0 (due to Bers) given by $q^*(z) \mapsto -2\mathrm{Im}(z) q^*(\bar{z})$. This implies that a neighborhood of 0 of

$$\{\mu_q(t),\ |t| < 1\}$$

is a holomorphic submanifold of T_g. Since the complex structure on T_g is independent of the base point, all of $\{\mu_q(t),\ |t| < 1\}$ is a holomorphic submanifold.

To sum up: For a fixed Riemann surface R_0 and $q \in \Gamma(R_0, (\Omega^1_{R_0})^{\otimes 2})$ fixed, the set $t|q|/q$, $|t| < 1$ is a totally geodesic and holomorphic submanifold of T_g, a *Teichmüller disc*.

In order to characterize which (few!) Teichmüller discs descend to Teichmüller curves in M_g we provide another description of a Teichmüller disc. There is a natural $\mathrm{Sl}(2,\mathbb{R})$-action on the bundle of triples (R,φ,q) over T_g. Post-compose local charts ψ_i of R with $A \in \mathrm{Sl}(2,\mathbb{R})$ acting on $\mathbb{C} \cong \mathbb{R}^2$. Since overlapping charts differ only by translations and multiplication by ±1, the compositions

$$\widetilde{\psi_i} : U_i \longrightarrow \mathbb{C} \cong \mathbb{R}^2 \xrightarrow{A} \mathbb{R}^2 \cong \mathbb{C}$$

determine a new complex structure on R. We let this new Riemann surface be $A \cdot R$. There is a unique quadratic differential $A \cdot q$ on $A \cdot R$ whose integration charts are $\widetilde{\psi_i}$.

The action of $\mathrm{SO}(2,\mathbb{R})$ does not change the complex structure of R. Hence the orbit $\mathrm{Sl}(2,\mathbb{R}) \cdot (R,\varphi,q)$ projects to a disc in T_g. If we lift

$$\begin{aligned} \mathcal{H} &\to \mathrm{Sl}(2,\mathbb{R}) \\ \tau \ &\mapsto \begin{pmatrix} 1 \ \mathrm{Re}\,\tau \\ 0 \ \mathrm{Im}\,\tau \end{pmatrix} \end{aligned}$$

one verifies that the Beltrami coefficient of the composition $\widetilde{\psi}_i$ is $\frac{i-\tau}{i+\tau}\frac{|q|}{q}$. Hence the image of the $\mathrm{Sl}(2,\mathbb{R})$-orbit is a Teichmüller disc.

Proposition 4.2. *A Teichmüller disc descends to a Teichmüller curve, if and only if the setwise stabilizer of the disc in the mapping class group is a lattice in* $\mathrm{Aut}(\mathcal{H}) = \mathrm{Sl}(2,\mathbb{R})$. *This stabilizer coincides up to conjugation with a quotient (by a finite group) of the group of orientation-preserving diffeomorphisms that are affine in the charts provided by* q.

Proof. Proposition 3.2 and Proposition 3.3 in [McM03]. □

When referring to a Teichmüller curve in the sequel we always assume that it is *generated by* $q = \omega^2$ *the square of a holomorphic one-form.* If a Teichmüller curve is generated by a quadratic differential on R, take the double covering of R where it admits a square root. The above criterion gives an easy way to check that this pair generates again a Teichmüller curve.

Proof of Corollary 0.13. This is [McM03] Theorem 4.2 in a different language. Suppose the contrary was true. Then $U \to M_g$ defines a Teichmüller curve. Over the universal cover $\mathcal{H}$ of U we may choose sections $\omega_j = \omega_j(\tau)$ for $j = 1, 2$ generating the $(1,0)$-part of $\mathbb{L}_j$. We may moreover choose a symplectic basis $\{a_k, b_k\}$ for $k = 1, \ldots g$ of $R^1(f_{\mathcal{H}})_*\mathbb{Z}_V$ such that $\int_{b_k} \omega_j(\tau) = \delta_{j,k}$ for $\tau \in \mathcal{H}$. We consider the maps

$$F_j : \mathcal{H} \to \mathcal{H},\ \tau \mapsto \int_{a_j} \omega_j(\tau) \quad \text{for} \quad j = 1, 2.$$

Since the local systems $\mathbb{L}_j$ are strictly maximal Higgs both F_1 and F_2 are isometries.

Let R be the fibre of f over $i \in \mathcal{H}$. Since $\mathcal{H} \to T_g$ is the Teichmüller disc generated by ω_1 (say), Ahlfors' variational formula implies that

$$\left.\frac{dF_j}{dt}\right|_{t=i} = \int_R \omega_j(i) \frac{\overline{\omega_1(i)}}{\omega_1(i)}.$$

Hence the norm of dF_j at $\tau = i$ in the hyperbolic metric equals

$$||dF_j|| = \left| \int_R \omega_j(i) \frac{\overline{\omega_1(i)}}{\omega_1(i)} \right| \ / \, \mathrm{Im} F_j(i) = \left| \int_R \omega_j(i) \frac{\overline{\omega_1(i)}}{\omega_1(i)} \right| \ \ / \int_R |\omega_(i)|^2.$$

Since F_j is an isometry, this implies by the Schwarz Lemma that $||dF_j|| = 1$. Since ω_2 is not proportional to ω_1 this violates the Cauchy–Schwarz inequality. □

Sketch of the proof of Lemma 4.1. Since $\overline{\mathbb{L}}$ is also strictly maximal Higgs we have $\mathbb{L} \cong \overline{\mathbb{L}}$ by Corollary 0.13. It remains to check that $\mathbb{L}$ can indeed be defined over $\mathbb{R}$, see [Moe2], proof of Theorem 5.3, or repeat the argument used in the proof of Lemma 1.8. □

Proof of Theorem 0.12. We fix a fibre R of f and $\omega \in \Gamma(R, \Omega^1_R)$ that generates the Teichmüller curve. By its construction as $\mathrm{Sl}(2,\mathbb{R})$-orbit the fundamental group of U maps to a lattice Γ in $\mathrm{Sl}(2,\mathbb{R})$ and the subspace

$$\langle \mathrm{Re}\,\omega, \mathrm{Im}\,\omega \rangle \subset H^1(R,\mathbb{R})$$

is invariant under the action of the fundamental group. Let $\mathbb{L}$ be the corresponding irreducible rank two $\mathbb{C}$-local system. As in Lemma 1.8 one finds that $\mathbb{L}$ is defined over a number field, whose Galois closure we denote by $\tilde{K}$. Moreover we let K the number field generated by the traces of Γ. If we apply Deligne's semisimplicity (Lemma 1.5) to the polarized $\mathbb{C}$-variation of Hodge structures $R^1 f_* \mathbb{C}_V$ the local system $\mathbb{L}$ and all its Galois conjugates will appear. Moreover we may read the above argument of [VZ04] Lemma 2.1 backwards to conclude that $\mathbb{L}$ is strictly maximal Higgs. This proves 1.

By Corollary 0.13 the local system $R^1 f_* \mathbb{C}_V$ contains no other local subsystem isomorphic to $\mathbb{L}$. We claim that

$$R^1 f_* \mathbb{C}_V = \mathbb{W} \oplus \mathbb{M}, \quad \mathbb{W}_{\tilde{K}} = \bigoplus_{\sigma \in \mathrm{Gal}(\tilde{K}/\mathbb{Q})/\mathrm{Gal}(\tilde{K}/K)} \mathbb{L}^\sigma,$$

where the summands $\mathbb{W}$ and $\mathbb{M}$ are defined over $\mathbb{Q}$ and $\mathbb{L}^\sigma$ are pairwise non-isomorphic rank two local systems. The fact that L and L^σ are isomorphic if and only if σ fixes K is not needed here, but see the remark at the end of this section.

We now prove 2. The $\mathbb{Q}$-sub variation $\mathbb{W}$ determines, after choosing a $\mathbb{Z}$-lattice and fixing a polarization of type δ, a family of Abelian varieties $A \to U$ of dimension r. By [Fa83] the tangent space to the space of deformations of the moduli map $U \to A_{r,\delta}$ is a subspace of the global sections of the local system $\mathrm{End}(\mathbb{W}_{\mathbb{C}})$ of bidegree $(-1,1)$. Using above decomposition of $\mathbb{W}_{\mathbb{C}}$ into irreducible summands, one checks ([Moe2] Lemma 3.3) that this tangent space is trivial. Hence $U \to A_{r,\delta}$ is rigid and U is defined over a number field. □

The above way of characterizing Teichmüller curves inverts the history of these objects: The first examples, starting with Veech ([Ve89]), of Teichmüller curves were shown to have this property by checking that the stabilizer of $\mathcal{H} \hookrightarrow T_g$ is a lattice. Next, in [McM03], Teichmüller curves were constructed as the image in the moduli space M_2 of the intersection of two higher-dimensional $\mathrm{Sl}(2,\mathbb{R})$-invariant loci in the bundle of one-forms over M_2. Only recently ([BM05]) Theorem 0.12 was used to construct an infinite series of Teichmüller curves, including the examples of Veech.

We address once again the decomposition of the variation of Hodge structures of a Teichmüller curves. Obviously, if $\mathbb{L} \cong \mathbb{L}^\sigma$ then σ fixes the trace field.

The converse is needed in 5.1. We reproduce the full proof, since it is, besides Corollary 0.13, the only argument in this context that relies essentially on Teichmüller theory.

By Proposition 4.2, elements in the fundamental group of U may be represented by diffeomorphisms that are affine in the ω-charts. We want to show that if σ does not fix the trace field, then $\mathbb{L} \not\cong \mathbb{L}^\sigma$. Choose a hyperbolic element γ in the fundamental group whose trace $t = \mathrm{tr}\gamma$ is not fixed by σ. Lift γ to a diffeomorphism ϕ of some smooth fibre R of f. Let ϕ^* be the induced diffeomorphism of $H^1(R, \mathbb{R})$. The endomorphism $\phi^* + (\phi^*)^{-1}$ acts on $\mathbb{L}|_R$ by multiplication by $t = \mathrm{tr}\gamma$. We need to show that $\mathrm{Ker}(\phi^* + (\phi^*)^{-1} - t \cdot \mathrm{id})$ has dimension two. It suffices to show that the eigenvalues λ^+ (resp. λ^-) of ϕ^* of maximal (resp. minimal) absolute values are unique and that these are the eigenvalues of ϕ^* acting on $\mathbb{L}|_R$. This is shown in [McM03] Theorem 5.3:

The diffomorphism ϕ is not of finite order nor does it fix a closed loop in R. Hence ϕ is pseudo-Ansov ([FLP79]). For such a diffeomorphism there are two transverse measured foliations μ^+, μ^- on R, one is expanded by λ^+ and one is contracted by $\lambda^- = (\lambda^+)^{-1}$. These two foliations represent real cohomology classes in $\langle \mathrm{Re}\,\omega, \mathrm{Im}\,\omega \rangle = \mathbb{L}|_R$. It remains to show that λ^+ is simple and of largest absolute value. The corresponding statement for λ^- follows by considering ϕ^{-1}.

The cohomology $H^1(R, \mathbb{R})$ is spanned by $\mathbb{R}$-linear combinations of the Poincaré duals $C^\vee$ to simple closed curves. We know how φ acts on $C^\vee$: Since C is stretched in the direction of μ^+ and contracted in the direction of μ^-, we have for $n \to \infty$ that $(\lambda^+)^{-n}(\phi^*)^n C^\vee \to \alpha\mu^+$ for some $\alpha \in \mathbb{R}$. This proves the nonexistence of an eigenspace for ϕ^* in $H^1(R, \mathbb{R})$ with larger eigenvalue than λ^+ and the simplicity of λ^+.

5 The only Teichmüller-Shimura Curve

This section is a comparison between Teichmüller- and Shimura curves. Recall that in the proof of Theorem 0.12 we defined the family Abelian varieties $A \to U$ corresponding to the local system $\mathbb{W}$. It is determined up to isogeny by the Teichmüller curve and coincides with the Jacobian of the universal family over the Teichmüller curve in case $r = g$.

Theorem 5.1. *The family $A \to U$ has real multiplication by K. The locus of real multiplication is the smallest Shimura subvariety of A_r that contains the image of the moduli map $U \to A_r$.*

Proof. For each $a \in K$ the cycle

$$C_a := \bigoplus_{\mathrm{Gal}(\tilde{K}/\mathbb{Q})/\mathrm{Gal}(\tilde{K}/K)} \sigma(a) \cdot \mathrm{id}_{\mathbb{L}^\sigma} \in \mathrm{End}(\mathbb{W}_{\tilde{K}}) = H^0(U, \mathrm{End}(\mathbb{W}_{\tilde{K}}))$$

is defined over $\mathbb{Q}$ and of bidegree $(0, 0)$. It is hence an endomorphism of A/U ([De71] Remark 4.4.6). Since K is a trace field of a lattice in $\mathrm{Sl}(2, \mathbb{R})$ it is real.

By the classification of endomorphisms of Abelian varieties it is totally real. This proves the first claim.

For the second statement we have to show that the local systems $\mathbb{W}^{\otimes m} \otimes (\mathbb{W}^{\vee})^{\otimes m'}$ do not contain Hodge cycles other than products and tensor powers of the cycles C_a. This is shown in [Moe2] Lemma 3.3. The spirit of this Lemma is similar to the decomposition into Schur functors of exterior powers of the variation of Hodge structures over a Shimura curve in Section 3. The difference consists of an essential use of the uniqueness of the strictly maximal Higgs local system at some point. □

We now address the question of classifying non-compact Shimura curves that lie entirely inside M_g for $g \geq 2$. By Theorem 0.8 the variation of Hodge structures consists of unitary and strictly maximal Higgs local subsystems. By Corollary 0.14 the strictly maximal Higgs part splits off a rank two local subsystem. By Theorem 0.12 such a curve is automatically also a Teichmüller curve. We may now look at the decomposition of the variation of Hodge structures in the proof of Theorem 0.12. Since the Galois conjugate of a non-cocompact lattice in $\mathrm{Sl}(2,\mathbb{R})$ still contains non-trivial parabolic elements none of the local systems $\mathbb{L}^{\sigma}$ is unitary. By Theorem 0.8 again there is no $\mathbb{L}^{\sigma}$ for $\sigma \neq \mathrm{id}$. To sum up:

Lemma 5.2. *Suppose $U \to M_g$ is a non-compact Shimura curve. Then there is an unramified cover $U' \to U$ such that the pullback of the universal family over M_g to U' has a family of Jacobians $\mathrm{Jac(f)} : J \to U'$ with fixed part of dimension $g-1$, i.e. there is an Abelian variety A of dimension $g-1$ such that $A \times U'$ injects into J.*

Proof. The only thing left to remark is that the unitary parts in the variation of Hodge structures can be trivialized after an unramified cover $U' \to U$ (see Lemma 1.8). □

The possible dimensions d of a fixed part in a family of Jacobians of dimension g over a one-dimensional base have been studied by Xiao. He proves in [Xi87] that

$$d \leq \frac{5g+1}{6}.$$

This implies that a fixed part of dimension $g-1$ can only occur for $g \leq 7$. Families of curves of genus $g=2$, $g=3$ and $g-4$ with $g-1$-dimensional fixed part are known to exist, but for $g = 5, 6, 7$ existence is still an open question.

Theorem 0.15 may be considered as a non-existence statement for such families under the additional assumption 'Shimura curve'.

Sketch of the proof of Theorem 0.15, b. There are four main ingredients: First, the large fixed part limits the number of possible singular fibres the family $f : X \to Y$.

Second, the projection to $J \to J/A$ defines a covering $\pi : V \to E$ of a family of elliptic curves. By [Moe1] we may assume that $U = X(d)$ is a modular curve and $d = \deg(\pi)$.

Third, all invariants, in particular some intersection numbers, of the complex surface E (or rather its completion) are known. We factor $\pi = i \circ \pi_{\mathrm{opt}}$, where i is an isogeny and π_{opt} does not factor through a non-trivial isogeny. The knowledge of intersection number on V will then be used to bound the degree d_{opt} of π_{opt}.
Finally, the translation structure on a fibre R of f given by the one-form $\omega \in H^0(R, \Omega^1_R)$ that generates the Teichmüller curve is shown to have very special properties. In a case-by-case discussion depending on d_{opt} and the number of zeros of ω we show that these properties are absurd, except for the case of the Legendre family and the family

$$f_W : y^4 = x(x-1)(x-t)$$

over $X(2) \cong \mathbb{P}^1 \setminus \{0, 1, \infty\}$ with local coordinate t in genus three.

We now give more details on the first step. Let $P_1, \ldots, P_s \in R$ be the zeros of ω and let k_i ($i = 1, \ldots, s$) be the multiplicity of the zero. Since $U \to M_g$ is a Teichmüller curve, we may suppose by [GJ00] Theorem 5.5 that $\pi|_R$ is ramified over one point $O \in E_0 := |\pi(R)$ only. Equivalently, the translation structure determined by ω is such that $\pi|_{R \setminus \{P_1, \ldots, P_s\}}$ is a square-tiled covering (see e.g. loc. cit.). One can deduce from this that singular fibres of f do not contain separating nodes ([Moe3] Proposition 2.3) and, more precisely, that the dual graph of a singular fibre is a ring, say of length $\ell(y)$ for $y \in S = Y \setminus U$. (loc. cit. Lemma 2.1).

Concerning the second step, we recall that we constantly consider Teichmüller and Shimura curves up to unramified cover outside S, since this does not change the property 'strictly maximal Higgs'. We may hence replace U by $X(d)$ for $d \geq 3$ sufficiently large and post-compose π by an isogeny to maintain the property $d = \deg(\pi)$.

Concerning the third step, we remark that the genus of $X(d)$, the number of cusps and the self-intersection of the zero-section of E are well-known as functions of d. From the construction of a Teichmüller curve as $\mathrm{Sl}(2, \mathbb{R})$-orbit one deduces that the zeros P_i extend to sections $p_i : U \to V$ of f for $i = 1, \ldots, s$. We denote the closure of the image of p_i in X by $\overline{P_i}$. Since we assume X to be a minimal semistable model, we have

$$\omega_{X/Y} = \mathcal{O}_X \left(\sum \overline{P_i} + d\Delta_d \cdot R + D \right),$$

where R is a fibre of f, D is a divisor supported in the singular fibres of f and

$$\Delta_d = \frac{d^2}{24} \prod_{p|d} \left(1 - \frac{1}{d^2} \right).$$

Using the geometry of the singular fibres one checks ([Moe3] Lemma 3.13) that in fact $D = 0$. From the intersection numbers on E one deduces that on X we have

$$\overline{P_i} = -d\Delta_d/(k_i+1).$$

Since $\mathbb{L}$ satisfies the Arakelov equality and since the remaining local subsystems of $R^1 f_* \mathbb{C}$ are trivial, we have

$$\deg f_* \omega_{X/Y} = d\Delta_d.$$

We claim that for each of the singular fibres R_y for $y \in S$ of f we have

$$\Delta\chi_{\mathrm{top}}(R_y) := \chi_{\mathrm{top}}(R_y) - \chi_{\mathrm{top}}(R) = d/d_{\mathrm{opt}}.$$

Putting all these data in the Noether formula

$$12 \deg f_* \omega_{X/Y} - \sum_{y \in S} \Delta\chi_{\mathrm{top}}(R_y) = \omega^2_{X/Y}$$

we obtain

$$d_{\mathrm{opt}} = \frac{12}{\left(\sum_{i=1}^{s} \frac{k_1^2}{k_1+1}\right) + 16 - 4g}.$$

Since by definition $\sum_{i=1}^{s} k_i = 2g-2$ there is only a finite number of integer solution to this equation. We thus obtain the desired bound on d_{opt}.

To prove the claim, notice first that $\Delta\chi_{\mathrm{top}}(R_y)$ coincides with the number of nodes of the dual graph of R_y or equivalently with the local component group of the Néron model of the family of Jacobians around at y. The size of the component group at y is known to be d for the universal family of elliptic curves over $X(d)$. The details how to related the two groups are in [Moe3] Proposition 3.15.

The reader is referred to [Moe3] for details on the last step. □

We remark that the exceptional family f_W is generated by the degree 8 covering $\pi|_R$ of the once-punctured torus with Galois group the quaternion group. The differential ω is the pullback of a holomorphic differential on E_0. This covering $\pi|_R$ can also be obtained by a double covering of an elliptic curve ramified precisely over all four 2-torsion points post-composed by the multiplication by two.

References

[An92] André, Y.: Mumford-Tate groups of mixed Hodge structures and the theorem of the fixed part. Compositio Math. **82**, 1–24 (1992)

[Be79] Belyĭ, G.V.: Galois extensions of a maximal cyclotomic field. Izv. Akad. Nauk SSSR Ser. Mat. **43**, 267–276 (1979)

[BM05] Bouw, I.I., Möller, M.: Teichmüller curves, triangle groups, and Lyapunov exponents. Preprint (2005)

[De71] Deligne, P.: Théorie de Hodge II. I.H.É.S. Publ. Math. **40**, 5–57 (1971)

[De72] Deligne, P.: La conjecture de Weil pour les surfaces $K3$. Invent. math. **15**, 206–226 (1972)

[De79] Deligne, P.: Variétés de Shimura: Interprétation modulaire, et techniques de construction de modèles canoniques. Proc. Symp. Pure Math. **33**, part II, 247–289, (1979)

[De82] Deligne, P.: Hodge cycles on abelian varieties. (Notes by J. S. Milne). Springer Lecture Notes in Math. **900**, 9–100 (1982)

[De87] Deligne, P.: Un théorème de finitude pour la monodromie. Discrete Groups in Geometry and Analysis, Birkhäuser, Progress in Math. 67, 1–19 (1987)

[EV92] Esnault, H., Viehweg, E.: Lectures on Vanishing Theorems. DMV-Seminar **20**, Birkhäuser, Basel-Boston-Berlin (1992)

[Fa83] Faltings, G.: Arakelov's theorem for abelian varieties. Invent. math. **73**, 337–348 (1983)

[FH91] Fulton, W., Harris, J.: Representation Theory. A first course. Graduate Texts in Math. **129**, Springer-Verlag, New-York (1991)

[FLP79] Fahti, A., Laudenbach, F., Poénaru, V.: Traveaux de Thurston sur les surfaces. Astérique **66-67** (1979)

[GJ00] Gutkin, E., Judge, C.: Affine mappings of translation surfaces. Duke Math. J. **103** (2000)

[Gr84] Griffths, P.: Topics in transcendental algebraic geometry. Ann of Math. Stud. **106**, Princeton Univ. Press. Princeton, N.J. (1984)

[IT92] Imayhosi, Y., Taniguchi, M.: An introduction to Teichmüller spaces. Springer Tokyo (1992)

[Kol87] Kollár, J.: Subadditivity of the Kodaira Dimension: Fibres of general type. In: Algebraic Geometry, Sendai, 1985. Advanced Studies in Pure Mathematics **10**, 361–398 (1987)

[Kov02] Kovács, S.: Logarithmic vanishing theorems and Arakelov-Parshin boundedness for singular varieties. Compositio Math. **131**, 291–317 (2002)

[McM03] McMullen, C.: Billiards and Teichmüller curves on Hilbert modular surfaces. Journal of the AMS **16**, 857–885 (2003)

[Moe1] Möller, M.: Maximally irregularly fibred surfaces of general type. Manusc. Math. **116**, 71–92 (2005)

[Moe2] Möller, M.: Variations of Hodge structures of Teichmüller curves, JAMS **19**, 327–344 (2006)

[Moe3] Möller, M.: Shimura and Teichmüller curves, preprint (2005)

[Moo98] Moonen, B.: Linearity properties of Shimura varieties. Part I. J. Algebraic Geom. **7**, 539–567 (1998)

[Mu66] Mumford, D.: Families of Abelian varieties. Proc. Sympos. Pure Math. **9**, 347–351 (1966)

[Mu69] Mumford, D.: A note of Shimura's paper: Discontinuous groups and Abelian varietes. Math. Ann. **181**, 345–351 (1969)

[Sch96] Schoen, C.: Varieties dominated by product varieties. Int. J. Math. **7**, 541–571 (1996)

[Si88] Simpson, C.: Constructing variations of Hodge structure using Yang-Mills theory and applications to uniformization. Journal of the AMS **1**, 867–918 (1988)

[Si90] Simpson, C.: Harmonic bundles on noncompact curves. Journal of the AMS **3**, 713–770 (1990)

[Si92] Simpson, C.: Higgs bundles and local systems. Publ. Math. I.H.E.S **75** (1992), 5–95 (1992)

[Ve89] Veech, W.: Teichmüller curves in moduli space, Eisenstein series and an application to triangular billiards. Invent. Math. 97, 533–583 (1989)

[Vi95] Viehweg, E.: Quasi-projective Moduli for Polarized Manifolds. Ergebnisse der Mathematik, 3. Folge **30**, Springer, Berlin-Heidelberg-New York (1995)

[Vi05] Viehweg, E.: Compactifications of smooth families and of moduli spaces of polarized manifolds. Preprint 2005, final version 2006

[VZ01] Viehweg, E., Zuo, K.: On the isotriviality of families of projective manifolds over curves. J. Alg. Geom. **10**, 781–799 (2001)

[VZ02] Viehweg, E., Zuo K.: Base spaces of non-isotrivial families of smooth minimal models. In: Complex Geometry (Collection of Papers dedicated to Hans Grauert), 279–328 Springer, Berlin Heidelberg New York (2002)

[VZ03] Viehweg, E., Zuo, K.: Families over curves with a strictly maximal Higgs field. Asian J. of Math. **7**, 575 - 598 (2003)

[VZ04] Viehweg, E., Zuo, K.: A characterization of certain Shimura curves in the moduli stack of abelian varieties. J. Diff. Geom. **66**, (2004) 233–287 (2004)

[VZ05] Viehweg, E., Zuo, K.: Numerical bounds for semistable families of curves or of certain higher dimensional manifolds. Preprint 2005, J. Alg. Geom. to appear

[VZ05b] Viehweg, E., Zuo, K.: Arakelov inequalities and the uniformization of certain rigid Shimura varieties. Preprint 2005

[Xi87] Xiao, G.: Fibred algebraic surfaces with low slope. Math. Ann. 276, 449–466 (1987)

27.04.2020

11.6 Topology of Curves

↳ Cylindrical Decomposition Algorithm

C) Heights in diophantine geometry..pdf

Obs: Ver livro todo!!!

D) Global Aspects of Complex Geometry (2006-).l.pdf

Pág. 01(07)

Complex Surfaces of General type: Some recent progress

Pág. 04(10)

Definition 2. A smooth surface $\underline{S}$ is called minimal (or a minimal model) if it does not contain any exceptional curve E of the first kind (i.e. $E \cong \mathbb{P}^1$, $E^2 = -1$).

Pág. 417(419)

Special Families of Curves, of Abelian Varieties, and of Certain Minimal Manifolds over Curves.

Pág. 441(443)

4 Teichmüller Curves and

27 04 2020

their Variation of Hodge Structures.

Pág. 445(447)

5 the only Teichmüller - Shimura Curve

E) Alain Connes - Noncommutative geometry (1994...)..pdf

Pág. 34 (34)

the metric aspect of noncommutative geometry.

Pág. 552 (552)

Chapter 6 the metric aspect of noncommutative geometry.

Pág. 40

1. Heisenberg and the Noncommutative Algebra off Physical Quantities.

Figure 1. Hydrogen Spectrum.

F) IMPORTANT: ✡

Books_72_2020/ /
Projective Geometry/

Four-Manifolds, geometries

'()*' ' '+ ' # (A 1
2 66$3 (! -!
& #$. - 2

the problem of completing the computation of the character of this K-cycle for discrete subgroups of higher rank Lie groups. The second example is the supersymmetric Wess-Zumino model of quantum field theory. We prove there the new unexpected result that, even for the free theory, the K-theory of the algebra $\mathcal{A}(\mathcal{O})$ of observables localized in an open subset $\mathcal{O}$ of space-time is very nontrivial, of infinite rank as an abelian group, and pairs nontrivially with the K-homology class given by the supercharge operator Q of the theory.

We have already explained that Chapter V is a detailed account of the theory of von Neumann algebras. In its last section we explain our joint work with J.-B. Bost, motivated by earlier work of B. Julia, on a natural system of quantum statistical mechanics related to the arithmetic of $\mathbb{Z}$ and the Riemann zeta function. We show that our system, formulated as a C^*-dynamical system, undergoes a phase transition with spontaneous symmetry breaking with symmetry group the Galois group of the field of roots of unity.

5. The metric aspect of noncommutative geometry

In the last chapter of this book we shall develop operator theoretic ideas about the metric aspect of geometry and then apply these ideas to a fundamental example: space-time.

Given a (not necessarily commutative) $*$-algebra $\mathcal{A}$ corresponding to the functions on a “space X”, our basic data in order to develop geometry on X is simply a K-cycle $(\mathcal{H}, D)$ over $\mathcal{A}$ in the above sense (cf. D). A first example is provided by ordinary Riemannian geometry. Given a Riemannian manifold (K-oriented for convenience) we let its algebra $\mathcal{A}$ of functions act by multiplication in the Hilbert space $\mathcal{H}$ of L^2 spinors on X, and we let D be the Dirac operator in $\mathcal{H}$. One thus obtains a K-cycle over $\mathcal{A}$ which represents the fundamental class of X in K-homology. We shall give (Section 1 of Chapter VI) four simple formulas showing how to recover, from the K-cycle $(\mathcal{H}, D)$ the following classical geometric notions on X:

1) The geodesic distance d on X

2) The integration against the volume form $\det(g_{\mu\nu})^{1/2}\ dx^1 \wedge \ldots \wedge dx^n$

3) The affine space of gauge potentials

4) The Yang-Mills action functional.

The essential fact is that all these notions will continue to make sense in our general framework, and in particular will apply nontrivially even to the case of finite spaces X.

The operator theoretic formula 1) for the geodesic distance is in essence the dual of the usual formula in terms of the length of paths γ in X. Recall that given two points p and q in a Riemannian manifold X, their geodesic distance $d(p, q)$ is given as the infimum of the lengths of paths γ from p to q. The latter length is computed by means of an integral along γ of the square root of $g_{\mu\nu}\, dx^\mu\, dx^\nu$. Our formula for the same quantity $d(p,q)$ is different in nature in that it is a supremum (instead of an infimum) and it invokes as a variable a *function* f on X instead of a path γ in X. It is the following:

$$d(p,q) = \mathrm{Sup}\,\{|f(p) - f(q)| \;;\; f \in \mathcal{A}\;,\; \|[D,f]\| \leq 1\}\,.$$

The norm involved in the right-hand side is the norm of operators in Hilbert space. Note also that the points p and q of X are used to convert the element f of $\mathcal{A}$ into a scalar, namely they are used, in conformity with Gelfand's theorem, as characters of the algebra $\mathcal{A}$.

There are two important features of this formula. The first is that since it does not invoke paths but functions it continues to be meaningful and gives a finite nonzero result when the space X is no longer arcwise connected. Indeed, one can take a space X with two points $a \neq b$: the algebra $\mathcal{A}$ of functions on X is then simply $\mathcal{A} = \mathbb{C}\oplus\mathbb{C}$, the direct sum of two copies of $\mathbb{C}$, and a natural K-cycle is given by the diagonal action of $\mathcal{A}$ in $\ell^2(X)$ with operator

$$D = \begin{bmatrix} 0 & \mu \\ \mu & 0 \end{bmatrix}$$

One then checks that with our formula the distance between a and b is given by $\ell = \mu^{-1}$.

An equally important feature of our formula is that, since it uses functions instead of paths, it is directly compatible with the formalism of quantum mechanics in which particles are described by wave functions rather than by any classical path.

The next step is to recover the tools of Riemannian geometry which go beyond the mere "metric space" attributes of such spaces such as 3) and 4). First, the volume form and the corresponding integral are recovered by the Dixmier trace, already mentioned in the discussion of Chapter IV. In fact the case already covered in Chapter IV of the Hausdorff measure on Julia sets in terms of the Dixmier trace was far more difficult. In Section 1 we shall develop the formalism of gauge theory (commutative or not, of course) and the Yang-Mills action functional from our operator theoretic data. The main mathematical result of Section 2 is the analogue of the basic inequality between the second Chern number of a Hermitian vector bundle and the minimum of the Yang-Mills functional on compatible connections on this bundle. This is proved in Section 2 using the main theorem of Chapter IV Section 2. In Section 3 we determine the Yang-Mills connections on the noncommutative torus, a joint result of M. Rieffel and myself, after defining the metric structure of that simple noncommutative space in great detail. As it turns out, even though the space we start with is non commutative, the moduli space of Yang-Mills connections is nicely behaved and finite-dimensional.

The end of the chapter (and the book) is occupied by the analysis, with the above new geometric tool, of the structure of space-time. The idea can be simply stated as follows:

It was originally through the Maxwell equations that, by the well-known steps due to Lorentz, Michelson, Morley, Poincaré and Einstein, the model of Minkowski's space-time was elaborated. In more modern terms the Maxwell-Dirac Lagrangian accounts perfectly for all phenomena involving only the electromagnetic interaction. However, a century has passed since then and essential discoveries, both experimental and theoretical, have shown that in order to account for weak and strong forces a number of modifications in the above Lagrangian were necessary. (See the small chronological table below for the early development of weak forces.)

Chronological table

- 1896 H. Becquerel discovers radioactivity
- 1898 M. Curie shows that radioactivity is an intrinsically atomic property
- 1901 E. Rutherford shows the inhomogeneity of uranic rays and isolates the β rays (i.e. the weak interaction part)
- 1902 W. Kaufman shows that the β rays are formed of electrons
- 1907-14 J. Chadwick shows that the spectrum of β rays is continuous
- 1930 W. Pauli introduces the neutrino to account for the above continuous spectrum
- 1932 W. Heisenberg introduces the concept of isospin (in the other context of strong forces)
- 1934 E. Fermi gives a phenomenological theory of weak interactions
- 1935 H. Yukawa introduces the idea of heavy bosons mediating short range forces.

Among the next steps which led finally to the Glashow-Weinberg-Salam Lagrangian of the standard model let us mention the discovery of the vector-axial form of the weak currents, and the intermediate boson hypothesis which replaces the 4-fermion interaction of the Fermi theory by a renormalizable interaction. This requires a considerable mass for the intermediate-boson in order to get a possible unification of the weak and electromagnetic forces using the ideas of nonabelian gauge theory of Yang and Mills. There was a major difficulty at this point since a nonabelian pure Yang-Mills field is necessarily massless. This difficulty was resolved by the theoretical discovery of the Higgs mechanism, allowing for the generation, by spontaneous symmetry breaking, of nontrivial masses for fermions and intermediate bosons. The Higgs field thus introduced appears in three of the five natural terms of the Lagrangian and completely spoils their geometric significance.

Now our idea is quite simple. We have at our disposal a more flexible notion of geometric space in which the continuum (i.e. manifolds say) and the discrete (cf. the two-point space example above) are treated on the same footing, and in which the Maxwell-Dirac Lagrangian does make sense: its bosonic part is the already treated

Yang-Mills action and the fermionic part is easy to get since we are given the Dirac operator to start with. Thus we can keep track of the above modifications of the Lagrangian as modifications of the geometry of space-time. In other words, we look for a geometry in our sense, i.e. a triple $(\mathcal{A}, \mathcal{H}, D)$ as above, such that the associated Maxwell-Dirac action functional produces the standard model of electroweak and strong interactions with all its refinements dictated by experimental results.

The result that we obtain is canonically derived from the standard model considered as a *phenomenological* model. What we find is a geometric space which is neither a continuum nor a discrete space but a mixture of both. This space is the product of the ordinary continuum by a discrete space with only two points, which, for reasons which will become clear later, we shall call L and R. The geometry, in our sense, of the finite space $\{L, R\}$ is given by a finite-dimensional Hilbert space representation $\mathcal{H}$ and a selfadjoint operator D in $\mathcal{H}$. These have direct physical meaning since the operator D is given by the Yukawa coupling matrix which encodes the masses of the elementary fermions as well as their mixing, i.e. the Kobayashi-Maskawa mixing parameters. Computing the distance between the two points L and R thus gives a number of the order of 10^{-16}cm, i.e. the inverse mass of the heaviest fermion.

The naive picture that emerges is that of a double space-time, i.e. the product of ordinary space-time by a very tiny discrete two-point space. By construction, purely left-handed particles such as neutrinos live on the left-handed copy X_L while electrons involve both X_L and X_R in $X = X_L \cup X_R$. Note that each point p of X_L is extremely close to a corresponding point p' of X_R, and when computing the differential of a function f on $X = X_L \cup X_R$ we shall find that it invokes not only the ordinary differential of f on X_L or on X_R but also the finite difference $f(p) - f(p')$ of the values of f on corresponding points of X_L and X_R. It is this finite difference, occurring as a new "transverse" component of the gauge potential, which yields the famous Higgs fields which appear in the three non-geometric terms of the standard model Lagrangian (cf. Chapter VI) $\mathcal{L} = \mathcal{L}_G + \mathcal{L}_f + \mathcal{L}_\varphi + \mathcal{L}_Y + \mathcal{L}_V$.

Note that there is no symmetry whatsoever between the points L and R, and in fact the natural vector bundle used over that space has fiber $\mathbb{C}^2$ over L and $\mathbb{C}$ over R. All this works very well for the electroweak sector of the standard model, but in order to account for the (already essentially geometric) strong structure much more work was necessary. It has been done in collaboration with J. Lott and is described in the last two Sections 4 and 5. It hinges on the fundamental problem of defining correctly what is a "manifold" in the setup of noncommutative geometry. Our proposed answer, which is essentially dictated by the $SU(2)\times SU(3)$ structure of up-down quark iso-doublets, is given by Poincaré duality discussed in Section 4. It fits remarkably well with the work of Sullivan on the fundamental role of Poincaré duality in K-homology for ordinary manifolds (cf. Section 4).

In Section 5 we discuss the full standard model, and also account for the hypercharge assignment of particles from a general unimodularity condition. One should understand this work as pure phenomenology, at the classical level, interpreting the detailed structure of the best phenomenological model, as the *fine structure* of space-time rather than as a long list of new particles. Applications of this point of view in quantum field theory, say as a model building device, are still ahead of us.

CHAPTER 6

The metric aspect of noncommutative geometry

The geometric spaces of Gauss and Riemann are defined as manifolds in which the metric is given by the formula

$$\delta(p,q) = \text{ infimum of length of paths } \gamma \text{ from } p \text{ to } q \tag{6.1}$$

where the length of a path γ is computed as the integral of the square root of a quadratic form in the differential of the path

$$\text{Length of } \gamma = \int_p^q (g_{\mu\nu}\, dx^\mu \ \ dx^\nu)^{1/2} \tag{6.2}$$

These geometric spaces form a relevant class of metric spaces, inasmuch as:

α) They are general enough to include numerous examples ranging from non-Euclidean geometries through surfaces embedded in $\mathbb{R}^3$ to space like hypersurfaces in general relativity.

β) They are special enough to deserve the name "geometry", since, being determined by local data, all the tools of differential and integral calculus are available to analyse them.

We have developed in Chapter IV a differential and integral calculus of "infinitesimals", given a Fredholm module $(\mathcal{H}, F)$ over the algebra $\mathcal{A}$ of coordinates on a possibly noncommutative space X. The Fredholm module $(\mathcal{H}, F)$ over $\mathcal{A}$ specifies the calculus on X but not the metric structure. For instance, the construction of $(\mathcal{H}, F)$ in the manifold

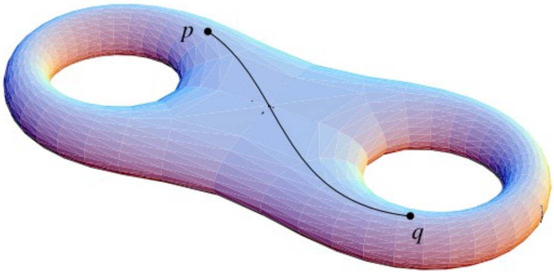

FIGURE 1. Geodesic

case (Section IV.4) only used the conformal structure. In fact, in the example of Section IV.3, where $X = \mathbb{S}^1$ and $(\mathcal{H}, F)$ is the Hilbert transform, the quantum differential expression

$$\mathrm{d}Z \; \mathrm{d}\overline{Z} = [F, Z][F, \overline{Z}] \quad , \quad Z : \mathbb{S}^1 \to \mathbb{C} \tag{6.3}$$

where Z is the boundary value of a univalent map, yields an infinitesimal unit of length intimately tied up with the metric on $Z(\mathbb{S}^1)$ induced by the usual Riemannian metric $\mathrm{d}z\mathrm{d}\overline{z}$ of $\mathbb{C}$. If we vary Z, even the *dimension* of $\mathbb{S}^1$ for the "metric" (3) will change (cf. Section IV.3).

Let $\mathcal{A}$ be an involutive algebra and $(\mathcal{H}, F)$ a Fredholm module over $\mathcal{A}$. To define a "unit of length" in the corresponding space X, we shall consider an operator of the form

$$G = \sum_{1}^{q} (\mathrm{d}x^\mu)^* \; g_{\mu\nu} (\mathrm{d}x^\nu) \tag{6.4}$$

where $\mathrm{d}x = [F, x]$ for any $x \in \mathcal{A}$, the x^μ are elements of $\mathcal{A}$ and where $g = (g_{\mu\nu})_{\mu,\nu=1,\ldots,q}$ is a positive element of the matrix algebra $M_q(\mathcal{A})$.

We want to think of G as the $\mathrm{d}s^2$ of Riemannian geometry. It is by construction a positive "infinitesimal", i.e. a positive compact operator on $\mathcal{H}$. The unit of length is its positive square root

$$\mathrm{d}s = G^{1/2}. \tag{6.5}$$

To measure distances in the possibly noncommutative space X we first replace the points $p, q \in X$ by the corresponding pure states φ, ψ on the C^*-algebra closure of $\mathcal{A}$

$$\varphi(f) = f(p) \; , \; \psi(f) = f(q) \quad \forall f \in \mathcal{A}. \tag{6.6}$$

We then dualise the basic formula (1) as follows

$$\mathrm{dist}(p, q) = \mathrm{Sup}\,\{|f(p) - f(q)| \; ; \; f \in \mathcal{A} \; , \; \|\mathrm{d}f/\mathrm{d}s\| \leq 1\} \tag{6.7}$$

which only involves p, q through the associated pure states (6). Since we are in the noncommutative set up we need to deal with the ambiguity in the order of the terms in an expression such as $\mathrm{d}f/\mathrm{d}s$ which can be either $\mathrm{d}f(\mathrm{d}s)^{-1}$ or $(\mathrm{d}s)^{-1}\,\mathrm{d}f$ or $(\mathrm{d}s)^{-\alpha}\,\mathrm{d}f(\mathrm{d}s)^{-(1-\alpha)}$ for instance. Instead of handling this problem directly we shall assume that G commutes with F, i.e. that $\mathrm{d}G = 0$, a condition similar to the Kähler condition, and introduce the following selfadjoint operator

$$D = FG^{-1/2} = F\;\mathrm{d}s^{-1}, \tag{6.8}$$

whose existence assumes that G is nonsingular, i.e. $\ker G = 0$. We shall then formulate (7) as follows

$$\mathrm{d}(p, q) = \mathrm{Sup}\,\{|f(p) - f(q)| \; ; \; f \in \mathcal{A} \; , \; \|[D, f]\| \leq 1\}\,. \tag{6.9}$$

Now the operator F is by construction the sign of D, while G is obtained from D by the formula

$$G = D^{-2}. \tag{6.10}$$

Thus it is more economical to take as our basic data the triple $(\mathcal{A}, \mathcal{H}, D)$ consisting of a Hilbert space $\mathcal{H}$, an involutive algebra $\mathcal{A}$ of operators on $\mathcal{H}$ and an unbounded selfadjoint operator D on $\mathcal{H}$. The conditions satisfied by such triples are

$$\begin{gathered}[D, a] \text{is bounded for any} a \in \mathcal{A} \\ (D-\lambda)^{-1} \text{is compact for any} \lambda \notin \mathbb{R}\end{gathered} \tag{6.11}$$

and were already formalised in Chapter IV Definition 2.11. In the present chapter we shall begin a systematic investigation of those geometric spaces. Besides Riemannian manifolds (see below) and spaces of non integral Hausdorff dimension (Section IV.3) the following are examples of geometric spaces described by our data:

a) Discrete spaces
b) Duals of discrete subgroups of Lie groups
c) Configuration space in supersymmetric quantum field theory.

We shall deal with Example a in Section 3 below. We have described already in great detail the triples $(\mathcal{A}, \mathcal{H}, D)$ corresponding to b) and c) in Section IV.9.

Our first task in this chapter will be to show that the Riemannian spaces are special cases of the above notion of geometric spaces. This will be done using an elliptic differential operator of order one, the Dirac operator (or the signature operator in the non-spin case). We shall first see that formula (9) applied to the triple (algebra of functions, Hilbert space of spinors, Dirac operator) readily gives back the geodesic distance (1) on the Riemannian manifold. Our next task will be to develop the analogue of the Lagrangian formulation of electrodynamics involving matter fields and gauge bosons for our more general geometric spaces. This will be done using the tools of the quantized calculus developed in Chapter IV Section 2. As mentioned above the commutator $[D, f]$, $f \in \mathcal{A}$ will play the role of the differential quotient $\mathrm{d}f/\mathrm{d}s$. As a central result we shall prove the inequality between the second Chern number of a "vector bundle" and the minimum of the Yang-Mills action on vector potentials. We shall see that our new notion of geometric space treats on an equal footing the continuum and the discrete, while the action for electrodynamics on the simplest mixture of continuum and discrete–the product of 4-dimensional continuum by a discrete 2-point space–gives the Glashow-Weinberg-Salam model for leptons. The notion of manifold in noncommutative geometry will be reached only after an understanding of Poincaré duality, i.e., that the K-homology cycle $(\mathfrak{H}, D)$ yields the fundamental class of the space under consideration. The notion of manifold obtained is directly inspired by the work of D. Sullivan [**543**] who discovered the basic role played by the K-homology fundamental class of a manifold.

The main example of a space to which all these considerations will be applied is Euclidean space-time in physics, i.e., space-time but with imaginary time. What we shall give is a geometric interpretation of the now experimentally confirmed effective low-energy model of particle physics, namely the Glashow–Weinberg–Salam standard model. This model is a gauge theory model with gauge group $U(1) \times SU(2) \times SU(3)$ and a pair of complex Higgs fields providing masses by the symmetry-breaking mechanism. We interpret this model geometrically as a pure gauge theory, i.e. electrodynamics, but on a more elaborate space-time $E' = E \times F$, the product of ordinary Euclidean space-time by a finite space F. The geometry of this finite space is specified by a pair $(\mathfrak{H}, D)$ as above, where $\mathfrak{H}$ is finite-dimensional and the selfadjoint operator D encodes the nine fermion masses and the four Kobayashi–Maskawa mixing parameters of the standard model.

The values of the hypercharges do not have to be fitted artificially to their experimental values but come out right from a simple unimodularity condition on the space E'.

Our analysis is limited to the classical context and does not at the moment address the questions related to renormalization, such as the existence of relations between coupling constants or the naturalness problem. Nevertheless, our more geometric and conceptual interpretation of the standard model gives a clear indication that particle physics is not so much a long list of elementary particles as the unveiling of the fine geometric structure of space-time.

The content of this Chapter Is organized as follows:

1. Riemannian geometry and the Dirac operator.
2. Positivity in Hochschild cohomology and inequalities for the Yang Mills action.
3. Product of continuum by discrete and the symmetry breaking mechanism.
4. The commutant and Poincaré duality.
5. The standard $U(1) \times SU(2) \times SU(3)$ model.

1. Riemannian Manifolds and the Dirac Operator

Let M be a compact Riemannian spin manifold, and let $D = \partial_M$ be the corresponding Dirac operator (cf. [**227**]). Thus, D is an unbounded selfadjoint operator acting in the Hilbert space $\mathfrak{H}$ of L^2-spinors on the manifold M.

We shall give four formulas below that show how to reconstruct the *metric space* (M, d), where d is the geodesic distance, the *volume measure* $\mathrm{d}v$ on M, the space of *gauge potentials*, and, finally, the *Yang–Mills* action functional, from the purely operator-theoretic data

$$(\mathcal{A}, \mathfrak{H}, D),$$

where D is the Dirac operator on the Hilbert space $\mathfrak{H}$ and where $\mathcal{A}$ is the abelian von Neumann algebra of multiplication by bounded measurable functions on M.

Thus, $\mathcal{A}$ is an abelian von Neumann algebra on $\mathfrak{H}$, and knowing the pair $(\mathfrak{H}, \mathcal{A})$ yields essentially no information (cf. Chapter V) except for the multiplicity, which is here the constant $2^{[d/2]}$, where $d = \dim M$. Similarly, the mere knowledge of the operator D on $\mathfrak{H}$ is equivalent to giving its list of eigenvalues $(\lambda_n)_{n\in\mathbb{N}}$, $\lambda_n \in \mathbb{R}$, and is an impractical point of departure for reconstructing M. The growth of these eigenvalues, i.e., the behavior of $|\lambda_n|$ as $n \to \infty$, is again governed by the dimension d of M, namely, $|\lambda_n| \sim Cn^{1/d}$ as $n \to \infty$.

What is relevant is the triple $(\mathcal{A}, \mathfrak{H}, D)$. Elements of $\mathcal{A}$ other than the constants do not commute with D, and the *boundedness* of the commutator $[D, a]$ already implies the following regularity condition on a measurable function a:

Lemma 1. *If a is a bounded measurable function on M then the densely defined operator $[D, a]$ is bounded if and only if a is almost everywhere equal to a Lipschitz function f, $|f(p) - f(q)| \leq Cd(p, q)$ $(\forall p, q \in M)$.*

Here, d is the geodesic distance in M. The operator $[D, a]$ should be viewed in effect as a quadratic form

$$\xi, \eta \mapsto \langle a\xi, D\eta\rangle - \langle D\xi, a^*\eta\rangle,$$

which is well-defined for ξ and η in the domain of D; its boundedness means an inequality

$$|\langle a\xi, D\eta\rangle - \langle D\xi, a^*\eta\rangle| \leq c\|\xi\|\|\eta\| \ \forall \xi, \eta \in \operatorname{dom} D.$$

The proof of the lemma follows immediately from results in [**202**].

Now, every Lipschitz function on M is continuous and the algebra of Lipschitz functions is norm-dense in the algebra of continuous functions on M; it follows that the C^*-algebra $C(M)$ of continuous functions on M is identical to the *norm closure* A in $\mathcal{L}(\mathfrak{H})$ of

By Gel'fand's theorem (Chapter II), we recover the compact topological space M as the spectrum of A. Thus, a point p of M is a $*$-homomorphism $\rho : A \to \mathbb{C}$,

$$\rho(ab) = \rho(a)\rho(b) \ \forall a, b \in A.$$

Any such homomorphism ρ is given by evaluation of a at p for some point $p \in M$,

$$\rho(a) = a(p) \in \mathbb{C}.$$

All of this is still qualitative; we now come to the first interesting formula, giving us a natural distance function, which turns out in this case to be the geodesic distance

Formula 1. *For any pair of points $p, q \in M$, their geodesic distance is given by the formula*

$$d(p, q) = \sup\{|a(p) - a(q)|;\ a \in \mathcal{A},\ \|[D, a]\| \leq 1\}.$$

The proof is straightforward, but it is relevant to go through it to see what is involved. The operator $[D, a]$, which by Lemma 1 is bounded iff a is Lipschitz, is then given by the Clifford multiplication $i^{-1}\gamma(\mathrm{d}a)$ by the gradient $\mathrm{d}a$ of a. This gradient is ([**202**]) a bounded measurable section of the cotangent bundle T^*M of M, and we have

$$\|[D, a]\| = \operatorname{ess\,sup} \|\mathrm{d}a\| = \text{the Lipschitz norm of } a.$$

It follows at once that the right-hand side of Formula 1 is less than or equal to the geodesic distance $d(p, q)$. However, fixing the point p and considering the function $a(q) = d(q, p)$, one checks that a is Lipschitz with constant 1, so that $\|[D, a]\| \leq 1$, which yields the desired equality. Note that Formula 1 is in essence dual to the original formula

$$(*) \qquad d(p, q) = \text{infimum of the length of paths } \gamma \text{ from } p \text{ to } q,$$

in the sense that, instead of involving arcs, namely copies of $\mathbb{R}$ inside the manifold M, it involves *functions* a, that is, maps from M to $\mathbb{R}$ (or to $\mathbb{C}$).

This is an essential point for us since, in the case of discrete spaces or of noncommutative spaces X, there are no interesting arcs in X but there are plenty of *functions*, namely, the elements $a \in \mathcal{A}$ of the defining algebra. We note at once that the right-hand side of Formula 1 is meaningful in that general context and it defines a metric on the space of *states* of the C^*-algebra A, the norm closure of $\mathbf{A} = \{a \in \mathcal{A};\ [D, a] \text{ is bounded}\}$

$$d(\varphi, \psi) = \sup\{|\varphi(a) - \psi(a)|;\ \|[D, a]\| \leq 1\}.$$

Finally, we also note that, although both Formula 1 and the formula $(*)$ give the same result for Riemannian manifolds, they are of quite different nature if we try to use them in actual measurements of distances. The formula $(*)$ uses the idealized notion of a path, and quantum mechanics teaches us that there is nothing like "the path followed by a particle". Thus, for measurements of very small distances, it is more natural to use wave functions and Formula 1.

We have now recovered from our original data $(\mathcal{A}, \mathfrak{H}, D)$ the metric space (M, d), where d is the geodesic distance. Let us now deal with the tools of differential and integral calculus, the first obvious example being the measure given by the volume form

$$f \mapsto \int_M f \mathrm{d}v,$$

where, in local coordinates $x^\mu, g_{\mu\nu}$, we have

$$\mathrm{d}v = \big(\det(g_{\mu\nu})\big)^{1/2} |\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n|.$$

This takes us to our second formula, which is nothing more than a restatement of H. Weyl's theorem about the asymptotic behavior of elliptic differential operators

([**247**], [**227**]). It does, however, involve a new tool, *the Dixmier trace* Tr_ω (cf. Chapter IV.2), which, unlike asymptotic expansions, makes sense in full generality in our context and is the correct operator-theoretic substitute for integration:

Formula 2. *For every* $f \in A$, *we have* $\int_M f \mathrm{d}v = c(d)\ \mathrm{Tr}_\omega(f|D|^{-d})$, *where* $d = \dim M$, $c(d) = 2^{(d-[d/2])}\ \pi^{d/2}\ \Gamma\left(\frac{d}{2}+1\right)$.

By convention we let D^{-1} be equal to 0 on the finite-dimensional subspace Ker$D$.

Let us refer to Section IV.2 for the detailed definition and properties of the Dixmier trace Tr_ω. We can interpret the right-hand side of the equality as the limit of the sequence

$$\frac{1}{\log N}\sum_{j=0}^{N}\lambda_j,$$

where the λ_j are the eigenvalues of the compact operator $f|D|^{-d}$, or, equivalently, as the residue, at the point $s = 1$, of the function

$$\zeta(s) = \mathrm{Trace}(f|D|^{-ds})\ (\Re s > 1).$$

The crucial fact for us is that the Dixmier trace makes sense independently of the context of pseudodifferential operators and that all properties of the integral $\int_M f \mathrm{d}v$, such as positivity, finiteness, covariance, etc., become obvious corollaries of the general properties of the Dixmier trace:

A) *Positivity*: $\mathrm{Tr}_\omega(T) \geq 0$ if T is a positive operator.
B) *Finiteness*: $\mathrm{Tr}_\omega(T) < \infty$ if the eigenvalues of $|T|$ satisfy $\sum_0^N \mu_n(T) = O(\log N)$.
C) *Covariance*: $\mathrm{Tr}_\omega(UTU^*) = \mathrm{Tr}_\omega(T)$ for every unitary U.
D) *Vanishing*: $\mathrm{Tr}_\omega(T) = 0$ if T is of trace class.

Property D is the counterpart of locality in our framework; it shows that the Dixmier trace of an operator is unaffected by a finite-rank perturbation, and allows many identities to hold, as we have seen in Chapter IV.

Now, setting up the integral of functions, i.e., the Riemannian volume form, is a good indication but quite far from the full story. In particular, many distinct Riemannian metrics yield the same volume form. Since our aim is to investigate physical spacetime at the scale of elementary particle physics, we shall now make a deliberate choice: instead of focusing on the intrinsic Riemannian curvature, which would drive us towards general relativity, we shall concentrate on the measurement (using $(\mathfrak{H}, D)$) of the curvature of connections on vector bundles, and on the Yang–Mills functional, which takes us to the theory of matter fields. This line is of course easier since it does not involve derivatives of the $g_{\mu\nu}$.

Let us state our aim clearly: to recover the Yang–Mills functional on connections on vector bundles, making use of only the following data (IV.2.11):

Definition 2. *A* K-cycle $(\mathfrak{H}, D)$, *over an algebra* $\mathcal{A}$ *with involution* $*$, *consists of a* $*$-*representation of* $\mathcal{A}$ *on a Hilbert space* $\mathfrak{H}$ *together with an unbounded selfadjoint operator* D *with compact resolvent, such that* $[D, a]$ *is bounded for every* $a \in \mathcal{A}$.

We shall assume that $\mathcal{A}$ is unital and that the unit $1 \in \mathcal{A}$ acts as the identity on $\mathfrak{H}$ (cf. Remark 12 for the nonunital case).

If the eigenvalues λ_n of $|D|$ are of the order of $n^{1/d}$ as $n \to \infty$, we say that the K-cycle is (d, ∞)-summable (cf. Section 2 of Chapter IV). On the algebra of functions on a compact Riemannian spin manifold, the Dirac operator determines a K-cycle that is (d, ∞)-summable, where $d = \dim M$. Finer regularity of functions, such as infinite differentiability, is easily expressed using the domains of powers of the derivation δ, $\delta(a) = [|D|, a]$.

We shall not be too specific about the choice of regularity; our discussion applies to any degree of regularity higher than Lipschitz.

The value of the following construction is that it will also apply when the $*$-algebra $\mathcal{A}$ is noncommutative, or when D is no longer the Dirac operator (cf. Section 3). The reader can have in mind both the Riemannian case and the slightly more involved case where the algebra $\mathcal{A}$ is the $*$-algebra of matrices of functions on a Riemannian manifold, just to bear in mind that the notion of exterior product no longer makes sense over such an algebra.

We shall begin with the notion of connection on the trivial bundle, i.e., the case of "electrodynamics", and define vector potentials and the Yang–Mills action in that case. We shall then treat the general case of arbitrary Hermitian bundles, i.e., in algebraic terms, of arbitrary Hermitian, finitely generated projective modules over $\mathcal{A}$.

We wish to define k-forms over $\mathcal{A}$ as operators on $\mathfrak{H}$ of the form

$$\omega = \sum a_0^j [D, a_1^j] \cdots [D, a_k^j],$$

where the a_i^j are elements of $\mathcal{A}$ represented as operators on $\mathfrak{H}$. This idea arises because, although the operator D fails to be invariant under the representation on $\mathfrak{H}$ of the unitary group $\mathcal{U}$ of $\mathcal{A}$,

$$\mathcal{U} = \{u \in \mathcal{A};\ u^* u = u u^* = 1\},$$

the following equality shows that the failure of invariance is governed by a 1-form in the above sense: by $\omega_u = u[D, u^*]$, that is,

$$u D u^* = D + \omega_u.$$

Note that ω_u is selfadjoint as an operator on $\mathfrak{H}$. Thus, it is natural to adopt the following definition:

Definition 3. *A* vector potential V *is a selfadjoint element of the space of* 1-*forms* $\sum a_0^j [D, a_1^j]$, *where* $a_k^j \in \mathcal{A}$.

One can immediately check that in the basic example of the Dirac operator on a spin Riemannian manifold, a vector potential in the above sense is exactly a 1-form ω on the manifold M and that this form is imaginary, the corresponding operator in the space of spinors being given by the Clifford multiplication:

$$V = i^{-1}\gamma(\omega) \ (i = \sqrt{-1}).$$

The action of the unitary group $\mathcal{U}$ on vector potentials is such that it replaces the operator $D + V$ by $u(D + V)u^*$; thus it is given by the algebraic formula

$$\gamma_u(V) = u[D, u^*] + uVu^* \ (u \in \mathcal{U}).$$

We now need only define the curvature or field strength θ for a vector potential, and use the analogue of the above Formula 2 to integrate the square of θ: the formula

$$\mathrm{YM}\,(V) = \mathrm{Tr}_\omega(\theta^2|D|^{-d})$$

should give us the Yang–Mills action.

The formula for θ should be of the form $\theta = \mathrm{d}V + V^2$; the only difficulty is in defining properly the "differential" $\mathrm{d}V$ of a vector potential, as an operator on $\mathfrak{H}$.

Let us examine what happens; the naive formulation is

$$\text{If } V = \sum a_0^j[D, a_1^j] \text{ then } \mathrm{d}V = \sum [D, a_0^j][D, a_1^j].$$

Before we point out what the difficulty is, let us check that if we replace V by $\gamma_u(V)$, where

$$\gamma_u(V) = u[D, u^*] + \sum u a_0^j[D, a_1^j]u^*,$$

then the curvature is transformed covariantly:

$$\mathrm{d}(\gamma_u(V)) + \gamma_u(V)^2 = u(\mathrm{d}V + V^2)u^*.$$

As this computation is instructive, we shall carry it out in detail. First, in order to write $\gamma_u(V)$ in the same form as V, we use the equality

$$[D, a_1^j]u^* = [D, a_1^j u^*] - a_1^j[D, u^*].$$

Thus, $\gamma_u(V) = u[D, u^*] + \sum u a_0^j[D, a_1^j u^*] - \sum u a_0^j a_1^j[D, u^*]$, and we have

$$\mathrm{d}\gamma_u(V) = [D, u][D, u^*] + \sum [D, u a_0^j][D, a_1^j u^*] - \sum [D, u a_0^j a_1^j][D, u^*].$$

We now claim that the following operators on $\mathfrak{H}$ are indeed equal:

$$\mathrm{d}\gamma_u(V) + \gamma_u(V)^2 = u(\mathrm{d}V + V^2)u^*.$$

We start with the left-hand side; it is equal to

$$\begin{aligned} & \mathrm{d}\gamma_u(V) + (u[D,u^*] + uVu^*)^2 \\ = \; & \mathrm{d}\gamma_u(V) + u[D,u^*]u[D,u^*] + u[D,u^*]uVu^* \\ & + uVu^*u[D,u^*] + uV^2u^* \\ = \; & \mathrm{d}\gamma_u(V) - [D,u][D,u^*] - [D,u]Vu^* + uV[D,u^*] + uV^2u^* \\ = \; & \sum [D, ua_0^j][D, a_1^j u^*] - \sum [D, ua_0^j a_1^j][D,u^*] \\ & - [D,u]Vu^* + uV[D,u^*] + uV^2u^* \\ = \; & u\mathrm{d}Vu^* + uV^2u^*, \end{aligned}$$

where the last equality follows from

$$\begin{aligned} & \sum [D,u]a_0^j[D,a_1^j u^*] - \sum [D,u]a_0^j a_1^j[D,u^*] = [D,u]Vu^*, \\ & \sum u[D,a_0^j][D,a_1^j u^*] - \sum u[D,a_0^j]a_1^j[D,u^*] = u\mathrm{d}Vu^*, \\ & \sum ua_0^j[D,a_1^j][D,u^*] = uV[D,u^*]. \end{aligned}$$

The difficulty that we overlooked is the following: the same vector potential V might be written in several ways as $V = \sum a_0^j[D,a_1^j]$, so that the definition of $\mathrm{d}V$ as

$$\mathrm{d}V = \sum [D,a_0^j][D,a_1^j]$$

is ambiguous.

To understand the nature of the problem, let us introduce some algebraic notation. We let $\Omega^*\mathcal{A}$ be the reduced universal differential graded algebra over $\mathcal{A}$ (Chapter III.1). It is by definition equal to $\mathcal{A}$ in degree 0 and is generated by symbols $\mathrm{d}a$ ($a \in \mathcal{A}$) of degree 1 with the following presentation:

α) $\mathrm{d}(ab) = (\mathrm{d}a)b + a\mathrm{d}b$ ($\forall a, b \in \mathcal{A}$),
β) $\mathrm{d}1 = 0$.

One can check that $\Omega^1\mathcal{A}$ is isomorphic as an $\mathcal{A}$-bimodule to the kernel $\ker(m)$ of the multiplication mapping $m : \mathcal{A} \otimes \mathcal{A} \to \mathcal{A}$, the isomorphism being given by the mapping

$$\sum a_i \otimes b_i \in \ker(m) \mapsto \sum a_i \mathrm{d}b_i \in \Omega^1\mathcal{A}.$$

The involution $*$ of $\mathcal{A}$ extends uniquely to an involution on Ω^* with the rule

$$(\mathrm{d}a)^* = -\mathrm{d}a^*.$$

The differential d on $\Omega^*\mathcal{A}$ is defined *unambiguously* by

$$\mathrm{d}(a^0\mathrm{d}a^1 \cdots \mathrm{d}a^n) = \mathrm{d}a^0\mathrm{d}a^1 \cdots \mathrm{d}a^n \;\forall a^j \in \mathcal{A},$$

and it satisfies the relations

$\mathrm{d}^2\omega = 0 \ \forall \omega \in \Omega^*\mathcal{A}$,
$\mathrm{d}(\omega_1\omega_2) = (\mathrm{d}\omega_1)\omega_2 + (-1)^{\partial\omega_1}\omega_1\mathrm{d}\omega_2 \ \forall \omega_j \in \Omega^*\mathcal{A}$.

Proposition 4.

1) *The following equality defines a $*$-representation π of the reduced universal algebra $\Omega^*(\mathcal{A})$ on $\mathfrak{H}$:*
$$\pi(a^0\mathrm{d}a^1\cdots\mathrm{d}a^n) = a^0[D,a^1]\cdots[D,a^n] \ \forall a^j \in \mathcal{A}.$$
2) *Let $J_0 = \ker\pi \subset \Omega^*$ be the graded two-sided ideal of Ω^* given by $J_0^{(k)} = \{\omega \in \Omega^k;\ \pi(\omega) = 0\}$; then $J = J_0 + \mathrm{d}J_0$ is a graded differential two-sided ideal of $\Omega^*(\mathcal{A})$.*

The first statement is obvious; let us discuss the second. By construction, J_0 is a two-sided ideal but it is not, in general, a *differential* ideal, i.e., if $\omega \in \Omega^k(\mathcal{A})$ and $\pi(\omega) = 0$, one does not in general have $\pi(\mathrm{d}\omega) = 0$. This is exactly the reason why the above definition of $\sum[D,a_0^j][D,a_1^j]$ as the differential of $\sum a_0^j[D,a_1^j]$ was ambiguous.

Let us show, however, that $J = J_0 + \mathrm{d}J_0$ is still a two-sided ideal. Since $\mathrm{d}^2 = 0$ it is obvious that J is then a differential ideal. Let $\omega \in J^{(k)}$ be a homogeneous element of J; then ω is of the form $\omega = \omega_1 + \mathrm{d}\omega_2$, where $\omega_1 \in J_0 \cap \Omega^k$, $\omega_2 \in J_0 \cap \Omega^{k-1}$. Let $\omega' \in \Omega^{k'}$, and let us show that $\omega\omega' \in J^{(k+k')}$. We have
$$\begin{aligned} \omega\omega' &= \omega_1\omega' + (\mathrm{d}\omega_2)\omega' = \omega_1\omega' + \mathrm{d}(\omega_2\omega') - (-1)^{k-1}\omega_2\mathrm{d}\omega' \\ &= \big(\omega_1\omega' + (-1)^k\omega_2\mathrm{d}\omega'\big) + \mathrm{d}(\omega_2\omega'). \end{aligned}$$
But, the first term belongs to $J_0 \cap \Omega^{k+k'}$ and $\omega_2\omega' \in J_0 \cap \Omega^{k+k'-1}$.

Using 2) of Proposition 4, we can now introduce the graded differential algebra
$$\Omega_D^* = \Omega^*(\mathcal{A})/J.$$
Let us first investigate Ω_D^0, Ω_D^1 and Ω_D^2.

We have $J \cap \Omega^0 = J_0 \cap \Omega^0 = \{0\}$ provided that we assume, as we shall, that $\mathcal{A}$ is a *subalgebra* of $\mathcal{L}(\mathfrak{H})$. Thus, $\Omega_D^0 = \mathcal{A}$.

Next, $J \cap \Omega^1 = J_0 \cap \Omega^1 + \mathrm{d}(J_0 \cap \Omega^0) = J_0 \cap \Omega^1$; thus Ω_D^1 is the quotient of Ω^1 by the kernel of π, and it is thus exactly the $\mathcal{A}$-bimodule $\pi(\Omega^1)$ of operators ω of the form
$$\omega = \sum a_j^0[D,a_j^1] \ (a_j^k \in \mathcal{A}).$$
Finally, $J \cap \Omega^2 = J_0 \cap \Omega^2 + \mathrm{d}(J_0 \cap \Omega^1)$ and the representation π gives an isomorphism
$$(*) \qquad \Omega_D^2 \cong \pi(\Omega^2)/\pi\big(\mathrm{d}(J_0 \cap \Omega^1)\big).$$
More precisely, this means that we can view an element ω of Ω_D^2 as a class of elements ρ of the form
$$\rho = \sum a_j^0[D,a_j^1][D,a_j^2] \ (a_j^k \in \mathcal{A})$$

modulo the sub-bimodule of elements of the form

$$\rho_0 = \sum [D, b_j^0][D, b_j^1] \quad ; \quad b_j^k \in \mathcal{A}, \ \sum b_j^0 [D, b_j^1] = 0.$$

It is now clear that since we work modulo this subspace $\pi\big(\mathrm{d}(J_0 \cap \Omega^1)\big)$, the question of ambiguity in the definition of $\mathrm{d}\omega$ for $\omega \in \pi(\Omega^1)$ no longer arises.

The equality $(*)$ makes sense for all k,

$$(*) \qquad \Omega_D^k \cong \pi(\Omega^k)/\pi\big(\mathrm{d}(J_0 \cap \Omega^{k-1})\big),$$

and allows us to define the following inner product on Ω_D^k: for each k let $\mathfrak{H}_k$ be the Hilbert space completion of $\pi(\Omega^k)$ with the inner product

$$\langle T_1, T_2 \rangle_k = \mathrm{Tr}_\omega(T_2^* T_1 |D|^{-d}) \ \forall T_j \in \pi(\Omega^k).$$

Let P be the orthogonal projection of $\mathfrak{H}_k$ onto the orthogonal complement of the subspace $\pi\big(\mathrm{d}(J_0 \cap \Omega^{k-1})\big)$. By construction, the inner product $\langle P\omega_1, \omega_2 \rangle = \langle P\omega_1, P\omega_2 \rangle$ for $\omega_j \in \pi(\Omega^k)$ depends only on their classes in Ω_D^k. We denote by Λ^k the Hilbert space completion of Ω_D^k for this inner product; it is, of course, equal to $P\mathfrak{H}_k$.

Proposition 5.

1) *The actions of* $\mathcal{A}$ *on* Λ^k *by left and right multiplication define commuting unitary representations of* $\mathcal{A}$ *on* Λ^k.
2) *The functional* $\mathrm{YM}(V) = \langle \mathrm{d}V + V^2, \mathrm{d}V + V^2 \rangle$ *is positive, quartic and invariant under gauge transformations,*

$$\gamma_u(V) = u\mathrm{d}u^* + uVu^* \ \forall u \in \mathcal{U}(\mathcal{A}).$$

3) *The functional* $I(\alpha) = \mathrm{Tr}_\omega(\theta^2 |D|^{-d})$, with $\theta = \pi(\mathrm{d}\alpha + \alpha^2)$, *is positive, quartic and gauge invariant on* $\{\alpha \in \Omega^1(\mathcal{A});\ \alpha = \alpha^*\}$.
4) *One has* $\mathrm{YM}(V) = \inf\{I(\alpha);\ \pi(\alpha) = V\}$.

Let us say a few words about the easy proof. First, the left and right actions of $\mathcal{A}$ on $\mathfrak{H}_k$ are unitary. The unitarity of the right action of $\mathcal{A}$ follows from the equality $\mathrm{Tr}_\omega(Ta|D|^{-d}) = \mathrm{Tr}_\omega(aT|D|^{-d}) \quad \forall T \in \mathcal{L}(\mathcal{H}),\, a \in \mathcal{A}$. Since $\pi\big(\mathrm{d}(J_0 \cap \Omega^{k-1})\big) \subset \pi(\Omega^k)$ is a sub-bimodule of $\pi(\Omega^k)$ it follows that P is a bimodule morphism:

$$P(a\xi b) = aP(\xi)b \ \forall a, b \in \mathcal{A},\ \xi \in \mathfrak{H}_k.$$

Thus 1) follows. As for 2), one merely notes that by the above calculation, with $\mathrm{d}V$ now unambiguous, $\theta = \mathrm{d}V + V^2$ is covariant under gauge transformations, whence the result. For 3), one again uses the above calculation to show that $\mathrm{d}\alpha + \alpha^2$ transforms covariantly under gauge transformations.

Finally, 4) follows from the property of the orthogonal projection P: as an element of Λ^2, $\mathrm{d}V + V^2$ is equal to $P\big(\pi(\mathrm{d}\alpha + \alpha^2)\big)$ for any α with $\pi(\alpha) = V$, and since the ambiguity in $\pi(\mathrm{d}\alpha)$ is exactly $\pi\big(\mathrm{d}(J_0 \cap \Omega^1)\big)$ one gets 4).

Stated in simpler terms, the meaning of Proposition 5 is that the ambiguity that we met above in the definition of the operator curvature $\theta = \mathrm{d}V + V^2$ can be ignored by taking the infimum

$$\mathrm{YM}(V) = \inf \mathrm{Tr}_\omega(\theta^2 |D|^{-d})$$

over all possibilities for $\theta = \mathrm{d}V + V^2$, $\mathrm{d}V = \sum [D, a_j^0][D, a_j^1]$ being ambiguous. The action obtained is nevertheless quartic by 2).

We shall now check that in the case of Riemannian manifolds with the Dirac K-cycle, the graded differential algebra Ω_D^* is canonically isomorphic to the *de Rham algebra of ordinary forms on M with their canonical pre-Hilbert space structure*. The whole point is that Propositions 4 and 5 give us these concepts in far greater generality, and the formula in 4) will allow extending to this generality, in the case $d = 4$, the inequality between the topological action and the Yang–Mills action YM (cf. Section 2). We refer the skeptical reader to the examples of Section 3.

We now specialize to the Riemannian case, where $\mathcal{A}$ is the algebra of functions (with some regularity) on the compact spin manifold M, and $D = \partial_M$ is the Dirac operator in the Hilbert space $L^2(M, S)$ of spinors. We let C be the bundle over M whose fiber at each $p \in M$ is the complexified Clifford algebra $\mathrm{Cliff}_\mathbb{C}(T_p^*(M))$ of the cotangent space at $p \in M$. Any bounded measurable section ρ of C defines a bounded operator $\gamma(\rho)$ on $\mathfrak{H} = L^2(M, S)$. For any $f^0, \dots, f^n \in \mathcal{A}$ one has

$$\pi(f^0 \mathrm{d}f^1 \cdots \mathrm{d}f^n) = i^{-n} \gamma(f^0 \mathrm{d}_\mathrm{c} f^1 \cdot \mathrm{d}_\mathrm{c} f^2 \cdots \mathrm{d}_\mathrm{c} f^n),$$

where the usual differential $\mathrm{d}_\mathrm{c} f$ is regarded as a section of C, and $\cdot$ denotes the product in C.

For each $p \in M$, the Clifford algebra C_p has a $\mathbb{Z}/2$ grading given by the parity of the number of terms ξ_j, $\xi_j \in T_p^*(M)$ in a product $\xi_1 \cdot \xi_2 \cdots \xi_n$, and a filtration, where $C_p^{(k)}$ is the subspace spanned by products of $n \leq k$ elements of $T_p^*(M)$. The associated graded algebra is canonically isomorphic to the complexified exterior algebra $\wedge_\mathbb{C}\big(T_p^*(M)\big)$ and $\sigma_k : C^{(k)} \to \wedge_\mathbb{C}^k(T_p^*)$ is the quotient mapping.

Using the canonical inner product on C given by the trace in the spinor representation, one can also identify $\bigwedge_\mathbb{C}^k$ with the orthogonal complement of $C^{(k-1)}$ in $C^{(k)}$; equivalently, if we let C^k be the subspace of $C^{(k)}$of elements of the same parity as k, then $\bigwedge_\mathbb{C}^k = C^k \ominus C^{k-2}$.

The differential algebra Ω_D^* is determined by the following lemma:

Lemma 6. *Let $(\mathfrak{H}, D)$ be the Dirac K-cycle on the algebra $\mathcal{A}$ of functions on M and let $k \in \mathbb{N}$. Then a pair T_1, T_2 of operators on $\mathfrak{H}$ is of the form $T_1 = \pi(\omega)$, $T_2 = \pi(\mathrm{d}\omega)$ for some $\omega \in \Omega^k(\mathcal{A})$ if and only if there exist sections ρ_1 and ρ_2 of C^k and C^{k+1} such that*

$$T_j = \gamma(\rho_j) \ (j = 1, 2), \ \mathrm{d}_\mathrm{c} \sigma_k(\rho_1) = i \sigma_{k+1}(\rho_2).$$

Here $\sigma_k(\rho_1)$ is an ordinary k-form on M and $\mathrm{d_c}$ is the classical differential. Note that for $k > \dim M$ one has $\sigma_k(\rho) = 0$. The proof is straightforward.

We can now easily determine the graded differential algebra Ω_D^*. First, let us identify $\pi(\Omega^k)$ with the space of sections of C^k; Lemma 6 then shows that

$$\pi\big(\mathrm{d}(J_0 \cap \Omega^{k-1})\big) = \mathrm{Ker}\sigma_k.$$

(If ρ is a section of C^k with $\sigma_k(\rho) = 0$ then the pair of sections $\rho_1 = 0$, $\rho_2 = \rho$ of C^{k-1} and C^k fulfills the condition of Lemma 6, so that $\rho = \pi(\mathrm{d}\omega)$ for some ω, $\pi(\omega) = 0$.) Thus, σ_k is an isomorphism, $\Omega_D^k \cong$ sections of $\bigwedge_{\mathbb{C}}^k(T^*)$, which, again by Lemma 6, commutes with the differential. We can then state:

Formula 3. *The mapping* $a^0\mathrm{d}a^1 \ldots \mathrm{d}a^n \mapsto a^0\mathrm{d_c}a^1 \cdots \mathrm{d_c}a^n$ *for* $a^j \in \mathcal{A}$ *extends to a canonical isomorphism of the differential graded algebra* Ω_D^* *with the de Rham algebra of differential forms on* M*. Under this isomorphism, the inner product on* Ω_D^k *is* $c(d)^{-1}$ *times the Riemannian inner product of* k*-forms*

$$\langle \omega, \omega' \rangle = \int_M \omega \wedge *\omega'.$$

The last equality follows from the computation of the Dixmier trace for the operator on $\mathfrak{H} = L^2(M, S)$ associated with a section ρ of the bundle C of Clifford algebras (cf. Section IV.2):

$$\mathrm{Tr}_\omega(\rho|D|^{-d}) = 2^{-d}\Gamma\left(\tfrac{d}{2}+1\right)^{-1}\ \pi^{-d/2}\int_M \mathrm{trace}\big(\rho(p)\big)\mathrm{d}v(p).$$

As an immediate corollary of Formula 3, we have

$$\mathrm{YM}(V) = c(d)^{-1}\int \|\mathrm{d}V\|^2\mathrm{d}v.$$

Let us now extend the definition of the action YM to connections on arbitrary Hermitian vector bundles.

First of all, we need to express in algebraic terms — i.e., using only the involutive algebra $\mathcal{A}$ — the notion of a Hermitian vector bundle over M. A vector bundle E is entirely characterized by the right $\mathcal{A}$-module $\mathcal{E}$ of sections of E with the same regularity as the elements of $\mathcal{A}$; the local triviality of E and the finite-dimensionality of its fibers translate algebraically into the statement that $\mathcal{E}$ is a direct summand of a free module $\mathcal{A}^N$ for some finite N, or, in fancier terms, that $\mathcal{E}$ is a finitely generated projective module over $\mathcal{A}$.

The Hermitian structure on E, that is, the inner product $\langle \xi, \eta \rangle_p$ on each fiber E_p, yields a sesquilinear map

$$\langle\ ,\ \rangle : \mathcal{E} \times \mathcal{E} \to \mathcal{A},$$

given by $\langle \xi, \eta \rangle(p) = \langle \xi(p), \eta(p) \rangle_p$. The mapping $\langle\ ,\ \rangle$ satisfies the following conditions:

1) $\langle \xi a, \eta b\rangle = a^* \langle \xi, \eta\rangle b$ $(\forall \xi, \eta \in \mathcal{E},\ a, b \in \mathcal{A})$,
2) $\langle \xi, \xi\rangle \geq 0$ $(\forall \xi \in \mathcal{E})$,
3) $\mathcal{E}$ is self-dual for $\langle\ ,\ \rangle$.

Thus, the Hermitian vector bundles over M correspond bijectively to the Hermitian, finitely generated projective modules over $\mathcal{A}$ in the following sense:

Definition 7. *Let $\mathcal{A}$ be a unital $*$-algebra and let $\mathcal{E}$ be a finitely generated projective module over $\mathcal{A}$. A* Hermitian structure *on $\mathcal{E}$ is given by a sesquilinear mapping $\langle\ ,\ \rangle : \mathcal{E} \times \mathcal{E} \to \mathcal{A}$ satisfying the above conditions* 1, 2 *and* 3.

We shall use this notion only in the case where $\mathcal{A}$ is a subalgebra stable under the holomorphic functional calculus in a C^*-algebra, in which case all reasonable notions of positivity coincide in $\mathcal{A}$.

In this case, all Hermitian structures on a given finitely generated projective module $\mathcal{E}$ over $\mathcal{A}$ are isomorphic to each other and are thus obtained as follows: one writes $\mathcal{E}$ as a direct summand $\mathcal{E} = e\mathcal{A}^N$ of a free module $\mathcal{E}_0 = \mathcal{A}^N$, where the idempotent $e \in M_N(\mathcal{A})$ is *self-adjoint*, and one then restricts to $\mathcal{E}$ the Hermitian structure on $\mathcal{A}^N$ given by

$$\langle \xi, \eta\rangle = \sum \xi_i^* \eta_i \in \mathcal{A}\ \forall \xi = (\xi_i), \eta = (\eta_i) \in \mathcal{A}^N.$$

The algebra $\mathrm{End}_{\mathcal{A}}(\mathcal{E})$ of endomorphisms of a Hermitian, finitely generated projective module $\mathcal{E}$ has a natural involution, given by

$$\langle T^* \xi, \eta\rangle = \langle \xi, T\eta\rangle\ \forall \xi, \eta \in \mathcal{E}.$$

With this involution, $\mathrm{End}_{\mathcal{A}}(\mathcal{E})$ is isomorphic to the reduced $*$-algebra $eM_N(\mathcal{A})e$.

As above, we now let $(\mathfrak{H}, D)$ be a K-cycle over $\mathcal{A}$, and Ω_D^1 the $\mathcal{A}$-bimodule of operators in $\mathfrak{H}$ of the form $V = \sum a_i[D, b_i]$, with $a_i, b_i \in \mathcal{A}$.

Definition 8. *Let $\mathcal{E}$ be a Hermitian, finitely generated projective module over $\mathcal{A}$. A* connection *on $\mathcal{E}$ is given by a linear mapping $\nabla : \mathcal{E} \to \mathcal{E} \otimes_{\mathcal{A}} \Omega_D^1$ such that*

$$\nabla(\xi a) = (\nabla \xi)a + \xi \otimes \mathrm{d}a\ \forall \xi \in \mathcal{E},\ a \in \mathcal{A}.$$

A connection ∇ is compatible *(with the metric) if and only if*

$$\langle \xi, \nabla \eta\rangle - \langle \nabla \xi, \eta\rangle = \mathrm{d}\langle \xi, \eta\rangle\ \forall \xi, \eta \in \mathcal{E}.$$

The last equality has a clear meaning in Ω_D^1. In the computations, one should remember that $(\mathrm{d}a)^* = -\mathrm{d}a^*$ $(\forall a \in \mathcal{A})$, and if $\nabla \xi = \sum \xi_i \otimes \omega_i$, $\omega_i \in \Omega_D^1$, then $\langle \nabla \xi, \eta\rangle = \sum \omega_i^* \langle \xi_i, \eta\rangle$.

Such connections always exist; for $\mathcal{E}$ expressed as $e\mathcal{A}^N$ as above, one may take ∇ as the "Grassmannian" connection

$$(\nabla_0 \xi) = e\eta,\ \text{where } \eta_j = \mathrm{d}\xi_j.$$

Two compatible connections ∇ and ∇' on $\mathcal{E}$ can only differ by an element $\Gamma \in \mathrm{Hom}_{\mathcal{A}}(\mathcal{E}, \mathcal{E} \otimes_{\mathcal{A}} \Omega^1_D)$.

As in Proposition 5, we shall now give two equivalent definitions of the action functional $\mathrm{YM}(\nabla)$ on the affine space $C(\mathcal{E})$ of compatible connections.

The group $\mathcal{U}(\mathcal{E})$ of unitary automorphisms of $\mathcal{E}$, $\mathcal{U}(\mathcal{E}) = \{u \in \mathrm{End}_{\mathcal{A}}(\mathcal{E});\ uu^* = u^*u = 1\}$, acts by conjugation $\gamma_u(\nabla) = u\nabla u^*$ on the space $C(\mathcal{E})$. To define the curvature θ of a connection ∇, one first extends ∇ to a unique linear mapping $\widetilde{\nabla}$ from $\widetilde{\mathcal{E}}$ to $\widetilde{\mathcal{E}}$, $\widetilde{\mathcal{E}} = \mathcal{E} \otimes_{\mathcal{A}} \Omega^*_D$, such that

$$\widetilde{\nabla}(\xi \otimes \omega) = (\nabla\xi)\omega + \xi \otimes \mathrm{d}\omega \ \forall \xi \in \mathcal{E},\ \omega \in \Omega^*_D,$$

and one checks that this mapping satisfies

$$\tilde{\nabla}(\eta\omega) = (\widetilde{\nabla}\eta)\omega + (-1)^{\partial\eta}\eta\mathrm{d}\omega$$

for every homogeneous $\eta \in \widetilde{\mathcal{E}}$ and $\omega \in \Omega^*_D$. It then follows that $\theta = \widetilde{\nabla}^2$ is an endomorphism of the right Ω^*_D-module $\widetilde{\mathcal{E}}$; it is determined by its restriction to $\mathcal{E}$, again denoted θ,

$$\theta \in \mathrm{Hom}_{\mathcal{A}}(\mathcal{E}, \mathcal{E} \otimes_{\mathcal{A}} \Omega^2_D).$$

Next, using the inner product on Ω^2_D and the Hermitian structure on $\mathcal{E}$, one has a natural inner product on

$$\mathrm{Hom}_{\mathcal{A}}(\mathcal{E}, \mathcal{E} \otimes_{\mathcal{A}} \Omega^2_D).$$

Using this, we make the following definition.

Definition 9. $\mathrm{YM}(\nabla) = \langle \theta, \theta \rangle$.

By construction, this action is gauge invariant, positive and quartic. It is moreover obvious from the above Formula 3 in the case of the Dirac K-cycle on a Riemannian spin manifold that one has:

Formula 4. *Let M be a Riemannian spin manifold with its Dirac K-cycle $(\mathfrak{H}, D)$. Then, the notion of connection* (Definition 8) *is the usual one, and*

$$\mathrm{YM}(\nabla) = c(d)^{-1} \int_M \|\theta\|^2_{\mathrm{HS}} \mathrm{d}v,$$

where θ is the usual curvature of ∇.

Thus, we recover in this case the usual Yang–Mills action. For computational purposes, and also to see the curvature as an operator in $\mathfrak{H}$, we shall now mention the easy adaptation of Proposition 5, 4) to the general case.

First of all, any compatible connection in the sense of Definition 8 is the composition with π of a *universal compatible connection*,

Definition 10. *Let $\mathcal{E}$ be a Hermitian, finitely generated projective module over $\mathcal{A}$. Then a* universal compatible connection *on $\mathcal{E}$ is a linear mapping ∇ of $\mathcal{E}$ to $\mathcal{E} \otimes_{\mathcal{A}} \Omega^1$ such that:*

a) $\nabla(\xi a) = (\nabla\xi)a + \xi \otimes \mathrm{d}a \ (\forall \xi \in \mathcal{E},\ a \in \mathcal{A})$,

b) $\langle \xi, \nabla\eta\rangle - \langle \nabla\xi, \eta\rangle = \mathrm{d}\langle \xi, \eta\rangle \ (\forall \xi, \eta \in \mathcal{E})$.

To see the surjectivity of the mapping $\pi : CC(\mathcal{E}) \to C(\mathcal{E})$, where $CC(\mathcal{E})$ is the space of universal compatible connections, it is enough to check that the special Grassmannian connection ∇_0 is of this form and that π is a surjection of Ω^1 onto $\pi(\Omega^1)$. Next (cf. Chapter III Section 3), a universal compatible connection extends uniquely as a linear mapping

$$\widetilde{\nabla} : \mathcal{E} \otimes_{\mathcal{A}} \Omega^* \to \mathcal{E} \otimes_{\mathcal{A}} \Omega^*$$

such that $\widetilde{\nabla}$ is equal to ∇ on $\mathcal{E} \otimes 1$ and such that

$$\widetilde{\nabla}(\eta\omega) = (\widetilde{\nabla}\eta)\omega + (-1)^{\deg \eta}\eta \mathrm{d}\omega$$

for every homogeneous η in $\mathcal{E} \otimes_{\mathcal{A}} \Omega^*$ and $\omega \in \Omega^*$.

The curvature $\theta = \tilde{\nabla}^2$ is then an endomorphism of the induced module $\tilde{\mathcal{E}} = \mathcal{E} \otimes_{\mathcal{A}} \Omega^*$ over Ω^*, and $\pi(\theta)$ makes sense as a bounded operator in the Hilbert space $\mathcal{E} \otimes_{\mathcal{A}} \mathfrak{H}$, as does the following operator D_∇ (cf. Section 3):

$$D_\nabla(\xi \otimes \eta) = \xi \otimes D\eta + \big((1 \otimes \pi)\nabla\xi\big)\eta \ \forall \xi \in \mathcal{E},\ \eta \in \mathfrak{H};$$

the analogue of the action I of Proposition 5 3) is then given by

$$I(\nabla) = \mathrm{Tr}_\omega\big(\pi(\theta)^2 |D_\nabla|^{-d}\big).$$

One then proves in the same way that

Proposition 11. *For any compatible connection $\nabla \in C(\mathcal{E})$, one has*

$$\mathrm{YM}(\nabla) = \inf\{I(\nabla_1);\ \pi(\nabla_1) = \nabla\}.$$

Remark 12. All the above considerations apply to noncompact spaces as well, i.e. the $*$-algebra $\mathcal{A}$ is no longer unital. The summability hypothesis $(D-\lambda)^{-1} \in \mathcal{L}^{(d,\infty)}$ $\forall \lambda \notin \mathbb{R}$ is being replaced by $a(D-\lambda)^{-1} \in \mathcal{L}^{(d,\infty)}$ $\forall \lambda \notin \mathbb{R}, a \in \mathcal{A}$.

Definition 10. *Let $\mathcal{E}$ be a Hermitian, finitely generated projective module over $\mathcal{A}$. Then a* universal compatible connection *on $\mathcal{E}$ is a linear mapping ∇ of $\mathcal{E}$ to $\mathcal{E} \otimes_{\mathcal{A}} \Omega^1$ such that:*

a) $\nabla(\xi a) = (\nabla\xi)a + \xi \otimes \mathrm{d}a \ (\forall \xi \in \mathcal{E},\ a \in \mathcal{A})$,
b) $\langle \xi, \nabla\eta\rangle - \langle \nabla\xi, \eta\rangle = \mathrm{d}\langle \xi, \eta\rangle \ (\forall \xi, \eta \in \mathcal{E})$.

To see the surjectivity of the mapping $\pi : CC(\mathcal{E}) \to C(\mathcal{E})$, where $CC(\mathcal{E})$ is the space of universal compatible connections, it is enough to check that the special Grassmannian connection ∇_0 is of this form and that π is a surjection of Ω^1 onto $\pi(\Omega^1)$. Next (cf. Chapter III Section 3), a universal compatible connection extends uniquely as a linear mapping

$$\widetilde{\nabla} : \mathcal{E} \otimes_{\mathcal{A}} \Omega^* \to \mathcal{E} \otimes_{\mathcal{A}} \Omega^*$$

such that $\widetilde{\nabla}$ is equal to ∇ on $\mathcal{E} \otimes 1$ and such that

$$\widetilde{\nabla}(\eta\omega) = (\widetilde{\nabla}\eta)\omega + (-1)^{\deg \eta}\eta \mathrm{d}\omega$$

for every homogeneous η in $\mathcal{E} \otimes_{\mathcal{A}} \Omega^*$ and $\omega \in \Omega^*$.

The curvature $\theta = \tilde{\nabla}^2$ is then an endomorphism of the induced module $\tilde{\mathcal{E}} = \mathcal{E} \otimes_{\mathcal{A}} \Omega^*$ over Ω^*, and $\pi(\theta)$ makes sense as a bounded operator in the Hilbert space $\mathcal{E} \otimes_{\mathcal{A}} \mathfrak{H}$, as does the following operator D_∇ (cf. Section 3):

$$D_\nabla(\xi \otimes \eta) = \xi \otimes D\eta + \big((1 \otimes \pi)\nabla\xi\big)\eta \ \forall \xi \in \mathcal{E},\ \eta \in \mathfrak{H};$$

the analogue of the action I of Proposition 5 3) is then given by

$$I(\nabla) = \mathrm{Tr}_\omega\big(\pi(\theta)^2 |D_\nabla|^{-d}\big).$$

One then proves in the same way that

Proposition 11. *For any compatible connection $\nabla \in C(\mathcal{E})$, one has*

$$\mathrm{YM}(\nabla) = \inf\{I(\nabla_1);\ \pi(\nabla_1) = \nabla\}.$$

Remark 12. All the above considerations apply to noncompact spaces as well, i.e. the $*$-algebra $\mathcal{A}$ is no longer unital. The summability hypothesis $(D-\lambda)^{-1} \in \mathcal{L}^{(d,\infty)}\ \forall \lambda \notin \mathbb{R}$ is being replaced by $a(D-\lambda)^{-1} \in \mathcal{L}^{(d,\infty)}\ \ \forall \lambda \notin \mathbb{R}, a \in \mathcal{A}$.

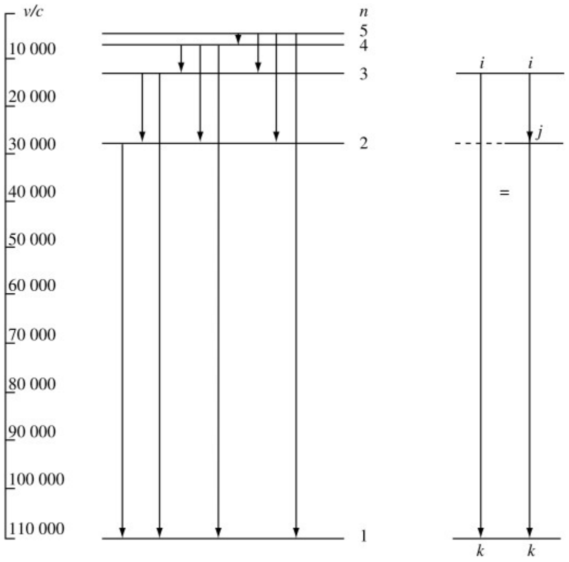

FIGURE 2. The vertical arrows represent the transitions. They are indexed by a pair of indices (i, j).

thus essential to find the regularities that appear in these configurations or *atomic spectra*. It is hydrogen that, in conformity with Mendeleev's table, has the simplest spectrum.

The numerical expression of the regularity of the lines $H_\alpha, H_\beta, H_\gamma, \ldots$ was obtained by Balmer in 1885 in the form

$$H_\alpha = \frac{9}{5}L,\ H_\beta = \frac{16}{12}L,\ H_\gamma = \frac{25}{21}L,\ H_\delta = \frac{36}{32}L,$$

where the value of the length L is approximately 3645.6×10^{-8} cm. In other words, the wavelengths of the lines in Figure 1 are of the form

$$\lambda = \frac{n^2}{n^2 - 4}\,L\,, \text{ where } n \text{ is an integer equal to 3, 4, 5 or 6.}$$

Around 1890, Rydberg showed that for a complex atom the lines of the spectrum can be classified into series, each of them being of the form

$$1/\lambda = R/m^2 - R/n^2 \text{ with } n \text{ and } m \text{ integers, } m \text{ fixed.}$$

Here $R = 4/L$ is Rydberg's constant. From this experimental discovery one deduces that, on the one hand, the frequency $\nu = c/\lambda$ is a more natural parameter than the wavelength λ for indexing the lines of the spectrum and, on the other hand, the

spectrum is a set of differences of frequencies, that is, there exists a set I of frequencies such that the spectrum is the set of differences $\nu_{ij} = \nu_i - \nu_j$ of arbitrary pairs of elements of I. This property shows that one can combine two frequencies ν_{ij} and ν_{jk} to obtain a third, $\nu_{ik} = \nu_{ij} + \nu_{jk}$. This important corollary is the Ritz-Rydberg combination principle: the spectrum is naturally endowed with a partially defined law of composition; the sum of certain frequencies of the spectrum is again a frequency of the spectrum (Figure 2).
Now, these experimental results could not be explained in the framework of the theoretical physics of the 19th century, based as it was on Newton's mechanics and Maxwell's electromagnetism. If one applies the classical conception of mechanics to the microscopic level, then an atom is described mathematically by the phase space and the Hamiltonian, and its interaction with radiation is described by Maxwell's theory. As we shall see, this theory predicts that the set of frequencies forms a subgroup Γ of $\mathbb{R}$.

Let us begin with the phase space. In the model of classical mechanics, to determine the later trajectory of a particle it is necessary to know both its initial position and its velocity. The initial data thus forms a set with six parameters, which are the three coordinates of position and the three coordinates of velocity, or rather of the momentum $p = mv$. If one is interested in a number n of particles, it is necessary to know the position and momentum of each of them. One is thus dealing with a set with $6n$ parameters, called the phase space of the mechanical system under consideration.

Starting with a function on this space which is called the Hamiltonian and which measures energy, classical mechanics prescribes the differential equations that determine the trajectory from the initial data. The natural structure of the phase space is that of a symplectic manifold X whose points are the "states" of the system. The Hamiltonian H is a function on X that intervenes to specify the evolution of every observable physical quantity, that is, of every function f on X, by the equation

$$\dot{f} = \{H, f\},$$

where $\{\ ,\ \}$ denotes the Poisson bracket and $\dot{f} = \mathrm{d}f/\mathrm{d}t$.

In the good cases such as, for example, the planetary model of the hydrogen atom, the dynamical system obtained is totally integrable. This means that there are sufficiently many "constants of motion" so that, on specifying them, the system is reduced to an almost periodic motion. The description of such a system is very simple. For, on the one hand, the algebra of observable quantities is the commutative algebra of almost periodic series

$$q(t) = \sum q_{n_1 \dots n_k} \exp(2\pi \mathrm{i} \langle n, \nu \rangle t),$$

where the n_i are integers, the ν_i are positive real numbers called the fundamental frequencies, and $\langle n, \nu \rangle = \sum n_i \nu_i$. On the other hand, the evolution in time is given by translation of the variable t.

The interaction between a classical atom and the electromagnetic field is described by Maxwell's theory. Such an atom emits an electromagnetic wave whose radiative part is calculated by superposing the plane waves W_n, $n = (n_1, \ldots, n_k)$, with frequencies $\langle n, \nu \rangle = \sum n_i \nu_i$, and whose amplitude and polarization are calculated simply from the fundamental observable that is the dipole moment. The dipole moment Q has three components Q_x, Q_y and Q_z, each of which is an observable quantity,

$$Q_x(t) = \sum q_{x,n} \exp(2\pi \mathrm{i} \langle n, \nu \rangle t)\,,$$

and which give the intensity of the emitted radiation of frequency $\langle n, \nu \rangle$ by the equation

$$I = \frac{\mathrm{d}E}{\mathrm{d}t} = \frac{2}{3c^3}(2\pi \langle n, \nu \rangle)^4 (|q_{x,n}|^2 + |q_{y,n}|^2 + |q_{z,n}|^2)\,,$$

where c denotes the speed of light. It follows, in particular, that the set of frequencies of the emitted radiations is an additive subgroup, $\Gamma \subset \mathbb{R}$, of the real numbers. Thus, to each frequency emitted there are associated all of its integral multiples or harmonics.

In fact, spectroscopy and its numerous experimental results show that this last theoretical result is contradicted by experiment. The set of frequencies emitted by an atom does not form a group, and it is false that the sum of two frequencies of the spectrum is again one. What experiment dictates is the Ritz-Rydberg combination principle, which permits indexing the spectral lines by the set Δ of all pairs (i, j) of elements of a set I of indices. The frequencies ν_{ij} and $\nu_{k\ell}$ only combine when $j = k$ to yield $\nu_{i\ell} = \nu_{ij} + \nu_{j\ell}$. The theory of Bohr, by artificially discretizing the angular momentum of the electron, succeeds in predicting the frequencies of the radiations emitted by the hydrogen atom, but it is unable to predict their intensities and polarizations.

It is by a fundamental calling into question of classical mechanics that Heisenberg arrived at this goal and went well beyond his predecessors. This questioning of classical mechanics runs approximately as follows: in the classical model, the algebra of observable physical quantities can be directly read from the *group* Γ of emitted frequencies; it is the convolution algebra of this group of frequencies. Since Γ is a commutative group, the convolution algebra is commutative. Now, in reality one is not dealing with a group of frequencies but rather, due to the Ritz-Rydberg combination principle, with a groupoid $\Delta = \{(i, j);\ i, j \in I\}$ having the composition rule $(i, j) \cdot (j, k) = (i, k)$. The convolution algebra still has meaning when one passes from a group to a groupoid, and the convolution algebra of the groupoid Δ is none other than *the algebra of matrices* since the convolution product may be written

$$(ab)_{(i,k)} = \sum_j a_{(i,j)} b_{(j,k)}\,,$$

which is identical with the product rule for matrices. On replacing the commutative convolution algebra of the group Γ by the noncommutative convolution algebra of the groupoid Δ dictated by experimental results, Heisenberg replaced classical mechanics, in which the observable quantities commute pairwise, by *matrix mechanics*, in which observable quantities as important as position and momentum no longer commute. Heisenberg's rules of algebraic calculation were imposed on him by the experimental results of spectroscopy. However, Heisenberg did not understand right away that the algebra he was working with was already known to mathematicians and was called the algebra of matrices. It was Jordan and Born who noticed this. In fact, Jordan had remarked that the conditions which, in Heisenberg's formalism, correspond to the Bohr-Sommerfeld quantization rules, signified that the diagonal elements of the matrix $[p,q]$ were equal to $-i\hbar$. In Heisenberg's matrix mechanics, an observable physical quantity is given by its coefficients $q_{(i,j)}$, indexed by elements $(i,j) \in \Delta$ of the groupoid Δ. The time evolution of an observable is given by the homomorphism $(i,j) \in \Delta \mapsto \nu_{ij} \in \mathbb{R}$ of Δ into $\mathbb{R}$ which associates with each spectral line its frequency. One has

$$q_{(i,j)}(t) = q_{(i,j)} \exp(2\pi \mathrm{i} \nu_{ij} t) \,. \tag{1}$$

This formula is the analogue of the classical formula

$$q_{n_1 \dots n_k}(t) = q_{n_1 \dots n_k} \exp(2\pi \mathrm{i} \langle n, \nu \rangle t) \,.$$

To obtain the analogue of Hamilton's law of evolution

$$\frac{dq}{dt} = \{H, q\} \,,$$

one defines a particular physical quantity H that plays the role of the classical energy and is given by its coefficients $H_{(i,j)}$, with

$$H_{(i,j)} = 0 \text{ if } i \neq j \,,\ H_{(i,i)} = h\nu_i \,, \text{ where } \nu_i - \nu_j = \nu_{ij} \ \forall\, i,j \in I,$$

where h is Planck's constant, a factor that converts frequencies into energies. One sees that H is defined uniquely, up to the addition of a multiple of the identity matrix. Moreover, the above formula (1) is equivalent to

$$\frac{dq}{dt} = \frac{2\pi i}{h}(Hq - qH) \,. \tag{2}$$

This equation is similar to the one of Hamilton that uses the Poisson brackets. It is in fact simpler, since it only uses the product of the observables, and more precisely the commutator $[A,B] = AB - BA$, which plays the role that the Poisson bracket plays in Hamiltonian mechanics. By analogy with classical mechanics, one requires the observables q of position and p of momentum to satisfy $[p,q] = -i\hbar$, where $\hbar = h/2\pi$. The simple algebraic form of the classical energy as a function of p and q then gives the equation of Schrödinger for determining the set $\{\nu_i;\ i \in I\}$, or the spectrum of H.

The quantum system thus described is much simpler and more rigid than its classical analogue. One thus obtains a nonnegligible payoff for abandoning the commutativity of classical mechanics. Though less intuitive, quantum mechanics is more directly accessible by virtue of its simplicity and its contact with spectroscopy.

In fact, the results of the theory of $*$-products show that a *symplectic* structure on a manifold such as the phase space is none other than the indication of the existence of a deformation with one parameter (h here) of the algebra of functions into a noncommutative algebra. I refer the reader to the literature ([**13**], [**37**], [**38**], [**169**], [**182**], [**220**], [**368**], [**474**], [**570**]) for a description of the results of this theory.

2. Statistical State of a Macroscopic System and Quantum Statistical Mechanics

A cubic centimeter of water contains on the order of $N = 3\times 10^{22}$ molecules of water agitated by an incessant movement. The detailed description of the motion of each molecule is not necessary, any more than is the precise knowledge of the microscopic state of the system, to determine the results of macroscopic observations. In classical statistical mechanics, a microscopic state of the system is represented by a point of the phase space, which is of dimension $6N$ for N point molecules. A statistical state is described not by a point of the phase space but by a measure μ on that space, a measure that associates with each observable f its mean value

$$\int f\,\mathrm{d}\mu\,.$$

For a system that is maintained at fixed temperature by means of a thermostat, the measure μ is called the Gibbs canonical ensemble; it is given by a formula that involves the Hamiltonian H of the system and the Liouville measure that arises from the symplectic structure of the phase space. One sets

$$\mathrm{d}\mu = \frac{1}{Z}\,e^{-\beta H}\cdot \text{Liouville measure}\,, \tag{3}$$

where $\beta = 1/kT$, T being the absolute temperature and k the Boltzmann constant, whose value is approximately 1.38×10^{-23} joules per degree Kelvin, and where Z is a normalization factor.

The thermodynamic quantities, such as the entropy or the free energy, are calculated as functions of β and a small number of macroscopic parameters introduced in the formula that gives the Hamiltonian H. For a finite system, the free energy is an analytic function of these parameters. For an infinite system, some discontinuities appear that correspond to the phenomenon of phase transition. The rigorous proof,

27 04 2020

their Variation of Hodge Structures.

Pág. 445(447)

5 the only Teichmüller - Shimura Curve

E) Alain Connes - Noncommutative geometry (1994...)..pdf

Pág. 34 (34)

the metric aspect of noncommutative geometry.

Pág. 552 (552)

Chapter 6 the metric aspect of noncommutative geometry.

Pág. 40

1. Heisenberg and the Noncommutative algebra off Physical Quantities.

Figure 1. Hydrogen Spectrum.

F) IMPORTANT. ✡

Books_72_2020/
Projective Geometry/
Four-Manifolds, geometries

27/04/2020

and Knots (2002....)....pdf

Pag. 246 (260)

12.8 The case $\pi = \mathbb{Z}/4\mathbb{Z}$

Pag. 247 (261)

Theorem 12.6

Let $M = S^2 \times S^2 / \langle \tau(I, -I)$.

Then the natural map from π to $Aut_{\pi}(\Pi)$ is an isomorphism, while $G_{\#}(M)$ has order ≤ 2.

G) The geometry of four-manifolds...pdf

Pag. 375 (389)

10. The Differential Topology of Algebraic Surfaces

154) ~~20~~ 20200428

A) Books_73_2020/
A Visual Introduction to the Fourth Dimension.../

A1) Geometry Proofs Essential Practice Problems.I.pdf

'()*' ' '+ ' # , ! 3 ! G
3 . 5 1 2 ! 2 -!
& $< < ; . J K"$K

The image of $[M;G/PL]$ in $[M;G/TOP]$ is a subgroup of index 2 (see Section 15 of [Si71]). It follows that if M is the total space of an S^2-bundle over RP^2 any homotopy equivalence $f : N \to M$ where N is also PL is homotopic to a homeomorphism. (For then $S_{TOP}(M) \cong (Z/2Z)^2$, and the nontrivial element of the image of $S_{PL}(M)$ is represented by a self homotopy equivalence of M. The case $M = S^2 \times RP^2$ was treated in [Ma79]. See also [Te97] for the cases with $\pi = Z/2Z$ and $w_1(M) = 0$.) This is also true if $M = S^4$, RP^4, CP^2, $S^2 \times S^2$ or $S^2 \tilde{\times} S^2$.

If $\pi \cong Z/4Z$ or $(Z/2Z)^2$ then $\chi(M) = 1$, and so M is nonorientable. As the $\mathbb{F}_2$-Hurewicz homomorphism is 0 in these cases Lemma 6.5 does not apply to give any exotic self homotopy equivalences.

If $\pi \cong (Z/2Z)^2$ then $[M;G/TOP] \cong (Z/2Z)^4$ and the surgery obstruction groups are $L_5((Z/2Z)^2,-) = 0$ and $L_4((Z/2Z)^2,-) = Z/2Z$, by Theorem 3.5.1 of [Wl76]. Since $w_1(M)$ is a split epimorphism $L_4(w_1(M))$ is an isomorphism, so the surgery obstruction is detected by the Kervaire-Arf invariant. As $w_1(M)^2 \neq 0$ we find that $S_{TOP}(M) \cong (Z/2Z)^3$. Thus there are between six and 56 homeomorphism classes of closed 4-manifolds with $\pi \cong (Z/2Z)^2$ and $\chi = 1$, of which half are not stably smoothable.

If $\pi \cong Z/4Z$ then $[M;G/TOP] \cong (Z/2Z)^2$ and the surgery obstruction groups $L_4(Z/4Z,-)$ and $L_5(Z/4Z,-)$ are both 0, by Theorem 3.4.5 of [Wl76]. Hence $S_{TOP}(M) \cong (Z/2Z)^2$. In the next section we show that self homotopy equivalences of $S^2 \times S^2/\langle\tau(I,-I)\rangle$ are homotopic to homeomorphisms. Hence there are four homeomorphism classes of closed 4-manifolds with $\pi \cong Z/4Z$ and $\chi = 1$. In all cases the orientable double covering space has trivial Kirby-Siebenmann invariant and so is homeomorphic to $(S^2 \times S^2)/(-I,-I)$. In the final section we give a candidate for another PL 4-manifold in this homotopy type. (This is from [HH17].)

12.8 The case $\pi = Z/4Z$

Let X be a connected cell complex, and let $G_\#(X)$ be the group of based self homotopy equivalences of X which induce the identity on all homotopy groups. Let $P_n(X)$ be the n^{th} stage of the Postnikov tower for X. This may be constructed by adjoining cells of dimension $\geq n+2$ to X. Then $G_\#(P_2(X)) \cong H^2(\pi;\pi_2(X))$, and there are exact sequences

$$H^n(P_{n-1}(X);\pi_n(X)) \to G_\#(P_n(X)) \to G_\#(P_{n-1}(X)),$$

for $n > 2$, by Theorem 2.2 and Proposition 1.5 of [Ts80], respectively. (The image on the right is the stabilizer in $G_\#(P_{n-1}(X))$ of the n^{th} k-invariant.)

In particular, if $H^k(P_{k-1}(X);\pi_k(X)) = 0$ for $2 \leq k \leq n$ then self homotopy equivalences of $P_n(X)$ are detected by their actions on the homotopy groups.

Let $M = S^2 \times S^2/\langle \tau(I,-I)\rangle$ be the geometric manifold with $\pi = \pi_1(M) \cong Z/4Z$ and $\chi = 1$. Since $\widetilde{M} = S^2 \times S^2$, $\pi_k(M) = \pi_k(S^2) \oplus \pi_k(S^2)$, for all $k \geq 2$. Let $\Pi = \pi_2(M)$, considered as a $\mathbb{Z}[\pi]$-module. Fix a basepoint $e = R_1(e)$ on the equator of S^2 and let $[s,t]$ be the image in M of $(s,t) \in S^2 \times S^2$. We shall take (e,e) and $[e,e]$ as basepoints for $S^2 \times S^2$ and M, respectively.

Theorem 12.6 *Let $M = S^2 \times S^2/\langle \tau(I,-I)\rangle$. Then the natural map from π to $Aut_\pi(\Pi)$ is an isomorphism, while $G_\#(M)$ has order ≤ 2.*

Proof The group $\pi = Z/4Z$ acts on $\Pi = \mathbb{Z}^2$ via $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$, and so $\Pi \cong \Lambda/(t^2+1) = \mathbb{Z}[i]$. Hence the natural map from π to $Aut_\pi(\Pi) \cong \mathbb{Z}[i]^\times = \langle i \rangle$ is an isomorphism. Similarly, π acts on $\pi_4(M) = (Z/2Z)^2$ via swapping the summands.

The Hopf maps corresponding to the factors of $\widetilde{M}$ generate $\pi_3(M) \cong \mathbb{Z}^2$. Hence $\pi = Z/4Z$ acts on $\pi_3(M)$ via $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. Simple calculations using the standard periodic resolution of the augmentation module $\mathbb{Z}$ give $H^0(\pi;\Pi) \cong H^0(\pi;\pi_3(M)) \cong \mathbb{Z}$, while $H^{2k-1}(\pi;\Pi) = H^{2k}(\pi;\pi_3(M)) = Z/2Z$ and $H^{2k}(\pi;\Pi) = H^{2k-1}(\pi;\pi_3(M)) = 0$, for all $k > 0$.

The homotopy fibre of the classifying map from $P_2(M)$ to $K(\pi,1)$ is $K(\Pi,2)$. Since $K(\Pi,2)$ has no cohomology in odd degrees, while $H^2(K(\Pi,2);\pi_3(M)) \cong \pi_3(M)^2$, all the terms with $p+q$ odd in the Leray-Serre spectral sequence

$$H^p(\pi;H^q(K(\Pi,2);\pi_3(M)) \Rightarrow H^{p+q}(P_2(M);\pi_3(M))$$

are 0. Hence the spectral sequence collapses, so $H^3(P_2(M);\pi_3(M)) = 0$ and $H^4(P_2(M);\pi_3(M)) \cong \mathbb{Z} \oplus T$, where T has order 4. Therefore $G_\#(P_2(M)) = G_\#(P_3(M)) = 0$.

We may assume that $P_3(M)$ has a single 5-cell, attached along a map which factors through one of the S^2 factors of $\widetilde{M}$. Therefore the connecting homomorphism from $H^4(M;\pi_4(M))$ to $H^5(P_3(M),M;\pi_4(M))$ is 0, and restriction from $H^4(P_3(M);\pi_4(M))$ to $H^4(M;\pi_4(M))$ is an isomorphism. Since $H^4(M;\pi_4(M)) \cong \overline{H_0(M;\pi_4(M))} = Z/2Z$, by Poincaré duality, we see that $G_\#(P_4(M))$ has order at most 2.

Every self homotopy equivalence of M extends to a self-homotopy equivalence of $P_n(M)$, for all $n > 0$. Conversely, if $n \geq 3$ then every self-map f of $P_n(M)$ restricts to a self-map of M, by cellular approximation, and if f is a self-homotopy equivalence then so is the restriction, by duality in the universal cover $\widetilde{M} = S^2 \times S^2$ and the Whitehead theorems. If, moreover, $n \geq 4$ then homotopies of self maps of $P_n(M)$ restrict to homotopies of self-maps of M. Thus $G_\#(M) = G_\#(P_4(M))$, and so has order ≤ 2. □

Is $G_\#(M)$ trivial? We may circumvent our ignorance of the answer to this question by the argument of the next lemma.

Let $M_o = \overline{M \setminus D^4}$, and let $j_o : M_o \to M$ be the natural inclusion.

Lemma 12.7 *Let h be a based self homotopy equivalence of M which induces the identity on π and Π. Then h is based homotopic to a self-homeomorphism of M.*

Proof The map h induces the identity on all homotopy groups, since $\pi_k(M) = \pi_k(S^2) \oplus \pi_k(S^2)$, for all $k \geq 2$. Let $P_j(h)$ be the extension of h to a self-homotopy equivalence of $P_j(M)$. Since $P_3(h)$ induces the identity on π, Π and $\pi_3(M)$, there is a homotopy H_t from $P_3(h)$ to the identity, by Theorem 12.6 and the exact sequences of [Ts80]. We may assume that $H(M_o \times [0,1])$ has image in M. Thus $h|_{M_o} \sim j_o$, and so we may assume that $h|_{M_o} = j_o$, by the HEP. We then see that h may be obtained by the pinch construction; $h = id_M \vee \gamma$ for some $\gamma \in \pi_4(M)$. Now $\gamma = \mu\eta^2$ for some $\mu \in \Pi$, since $\pi_4(S^2) = Z/2Z$ is generated by η^2. Since $H_2(M;\mathbb{F}_2) = H_2(\pi;\mathbb{F}_2)$, the Hurewicz homomorphism from Π to $H_2(M;\mathbb{F}_2)$ is 0. Therefore $kerv(h) = 0$. Hence h has trivial normal invariant, by Theorem 5.1 of [CH90]. Thus h is homotopic to a self-homeomorphism of M. □

Theorem 12.8 *Every based self-homotopy equivalence of M is based homotopy equivalent to a homeomorphism.*

Proof Interchange of factors and reflection across the equator of S^2 may be used to define basepoint preserving homeomorphisms r and s of M, with $r([x,y]) = [y,x]$ and $s([x,y]) = [R_1x, R_1y]$, for all $[x,y] \in M$. Let c be the equatorial arc from $c(0) = e$ to $c(1) = -e$ in S^2. Then $\gamma(t) = [c(t), e]$ represents a generator of π, while $r\gamma(t) = \gamma(1-t)$ and $s\gamma(t) = \gamma(t)$, for all $0 \leq t \leq 1$. Therefore $\pi_1(r) = -1$ and $\pi_1(s) = 1$. Clearly $\pi_2(r)$ and $\pi_2(s)$ have matrices $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ and $-I$, respectively.

Let h be the homeomorphism which drags the basepoint $\star$ around a loop representing a generator g of π. Then $\pi_1(h) = id_\pi$, since π is abelian, while $\pi_2(h)$

acts through g, and so has matrix $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ or its inverse. It is now clear that if f is a based self-homotopy equivalence of M there is a based self-homeomorphism F such that $\pi_i(f) = \pi_i(F)$, for $i = 1$ and 2. Since $F^{-1}f$ induces the identity on π and Π, the theorem follows from Lemma 12.7. □

Corollary 12.8.1 *There are four homeomorphism types of manifolds homotopy equivalent to* M*.* □

12.9 A smooth fake version of $S^2 \times S^2/\langle\tau(I,-I)\rangle$?

Let $M = S^2 \times S^2/\langle\tau(I,-I)\rangle$, as in the previous section, and let M^+ be its orientable double cover. Let $\Delta = \{(s,s) \mid s \in S^2\}$ be the diagonal in $S^2 \times S^2$. We may isotope Δ to a nearby sphere which meets Δ transversely in two points, by rotating the first factor, and so Δ has self-intersection ± 2. The diagonal is invariant under $(-I,-I)$, and so $\delta = \Delta/\langle\sigma^2\rangle \cong RP^2$ embeds in $M^+ = S^2 \times S^2/\langle(-I,-I)\rangle$ with an orientable regular neighbourhood. Since $\sigma(\Delta) \cap \Delta = \emptyset$ this also embeds in M. We shall see that the complementary region also has a simple description.

We shall view S^2 here as the unit sphere in the space of purely imaginary quaternions. The standard inner product on this space is given by $v \bullet w = \mathfrak{Re}(v\bar{w})$, for v, w purely imaginary quaternions. We shall also identify S^3 with the unit quaternions $\mathbb{H}_1$. Let $C_x = \{(s,t) \in S^2 \times S^2 \mid s \bullet t = x\}$, for all $x \in [-1,1]$. Then $C_1 = \Delta$ and $C_{-1} = \tau(I,-I)(\Delta)$, while $C_x \cong C_0$ for all $|x| < 1$. The map $p : S^3 \to C_0$ given by $p(q) = (q\mathbf{i}q^{-1}, q\mathbf{j}q^{-1})$ for all $q \in S^3$ is a 2-fold covering projection, and so $C_0 \cong RP^3$.

It is easily seen that $N = \cup_{x \geq \varepsilon} C_x$ and $\tau(I,-I)(N)$ are regular neighbourhoods of Δ and $\tau(I,-I)(\Delta)$, respectively, while $C = \cup_{x \in [-\varepsilon,\varepsilon]} C_x \cong C_0 \times [-\varepsilon,\varepsilon]$. In particular, N and $\tau(I,-I)(N)$ are each homeomorphic to the total space of the unit disc bundle in T_{S^2}, and $\partial N \cong C_0 \cong RP^3$. The subsets C_x are invariant under $(-I,-I)$. Hence $N(\delta) = N/\langle(-I,-I)\rangle$ is the total space of the tangent disc bundle of RP^2. In particular, $\partial N(\delta) \cong L(4,1)$ and δ represents the nonzero element of $H_2(M;\mathbb{F}_2)$, since it has self-intersection 1 in $\mathbb{F}_2$. (It is not hard to show that any embedded surface representing the nonzero element of $H_2(M;\mathbb{F}_2)$ is non-orientable but lifts to M^+, and so has an orientable regular neighbourhood.)

We also see that $C/\langle(-I,-I)\rangle \cong L(4,1) \times [-\varepsilon,\varepsilon]$. The self map of S^3 defined by right multiplication by $\frac{1}{\sqrt{2}}(\mathbf{1}+\mathbf{k})$ lifts $\tau(I,-I)$, since $p(q.\frac{1}{\sqrt{2}}(\mathbf{1}+\mathbf{k})) = \tau(I,-I)(p(q))$. Hence $C_0/\langle\tau(I,-I)\rangle = S^3/\langle\frac{1}{\sqrt{2}}(\mathbf{1}+\mathbf{k})\rangle = L(8,1)$, and so $MC =$

$C/\langle\sigma\rangle$ is the mapping cylinder of the double cover $L(4,1)\to L(8,1)$. Since $S^2\times S^2=N\cup C\cup\tau(I,-I)(N)$ it follows that $M=N(\delta)\cup MC$.

This construction suggests a candidate for another smooth 4-manifold in the same (simple) homotopy type. Let $M'=N(\delta)\cup MC'$, where MC' is the mapping cylinder of the double cover $L(4,1)\to L(8,5)$. Then $\pi_1(M')\cong Z/4Z$ and $\chi(M')=1$, and so there is a homotopy equivalence $h:M'\simeq M$. In this case, $kerv(h)$ is a complete invariant. The difficulty is in finding an easily analyzed explicit choice for h, for which we can compute $kerv(h)$.

Is there a computable homeomorphism (or diffeomorphism) invariant that can be applied here? Most readily computable invariants are invariants of homotopy type. These manifolds have $w_2\neq 0$ but $w_1^2=0$ and $w_1w_2=0$, so admit Pin^c-structures. Is it possible to distinguish them by consideration of the associated Arf invariants?

If M and M' are not homeomorphic then every closed 4-manifold with $\pi\cong Z/4Z$ and $\chi=1$ is homeomorphic to one of M, M', $*M$ or $*M'$.

'()*' ' '+ ' # . 0 2 , !
66 3 +@ ! 5 1 2 3 5 (-!
& #8 #;6 . B . - 2 (4

10

THE DIFFERENTIAL TOPOLOGY OF ALGEBRAIC SURFACES

Let us now review our position with regard to the overall aims of this book. In Chapter 1 we set up the twin problems in four-manifold theory: deciding the existence and uniqueness of manifolds with given intersection forms. In Chapter 8 we have shown how arguments with Yang–Mills moduli spaces give strong non-existence results, and in Chapter 9 we have developed invariants of four-manifolds aimed at the complementary question of uniqueness. However we have not yet given any full-blooded example of the application of these invariants, i.e. shown that they can distinguish differentiable four-manifolds with the same intersection form, and this is the purpose of the present chapter. We will describe the main ideas in the proof of the general Theorem (1.3.5), and give a complete proof of the special result (1.3.6).

We have seen in Chapter 9 that for most four-manifolds which admit connected sum decompositions the new invariants are trivial. We can construct connected sums which realize many homotopy types, so finding an example of the kind we want is essentially equivalent to showing that the invariants are not always trivial. In this chapter we shall see that, for rather basic reasons, the invariants are indeed non-trivial for complex algebraic surfaces. This general theory is explained in Section 10.1.

Another problem taken up in this chapter is that of the calculation of the invariants in any kind of generality. While we have defined invariants for a large class of four-manifolds, we have given little indication as to how they may be calculated in practice. The definition involves the solution of the non-linear ASD partial differential equations, and this can certainly not be done in general in any explicit form. However, for the particular class of complex surfaces, we have seen in Chapter 6 that the ASD connections can be identified with stable holomorphic bundles, and in this chapter we continue the same line of ideas by describing techniques which can be applied fairly generally to analyse moduli spaces of stable bundles. The basic construction is reviewed, within a differential-geometric framework in Section 10.2, and in Section 10.3 we illustrate its application in a particular case: moduli spaces of bundles over a ‘double plane’ (branched cover). These geometric calculations are then applied in Section 10.4 to calculate some of the Yang–Mills invariants for this four-manifold.

10.1 General theory

10.1.1 Statement of results

Let S be a smooth, simply connected, complex algebraic surface. As we have seen in Section 1.1.7, the underlying differentiable four-manifold has a standard orientation and $b^+(S) = 1 + 2p_g(S)$. We will henceforth assume that $p_g(S) > 0$, so $b^+ \geq 3$ and the various different polynomial invariants on $H_2(S)$ of Chapter 9 are defined. Moreover the complex structure fixes a natural orientation on the moduli space and so there are no ambiguities of sign (see Section 6.4.2). Here we want to have in mind primarily $SU(2)$-bundles, but our results apply equally to $SO(3)$-bundles where w_2 is the reduction of the first Chern class of a holomorphic line bundle over S (i.e. an integral (1, 1) class). In particular we can consider bundles with $w_2 = c_1(K_S) \bmod 2$. (The point of this condition is that we can then lift up to rank-two vector bundles over S, as explained in Section 6.1.4.)

The surface S can be holomorphically embedded in projective space. Such an embedding defines a 'hyperplane class' h in $H^2(S)$, the restriction of the standard generator for $H^2(\mathbb{CP}^N)$. Geometrically, this is the class realized by the intersecton of S with a general hyperplane or, in other language, a complex curve C in the linear system $|\mathcal{O}(1)|$ over S. The main general theorem we have is then:

Theorem (10.1.1). *For any simply connected complex algebraic surface S with $p_g(S) > 0$ and any hyperplane class h in $H_2(S)$ there is a $k_0 = k_0(S, h)$ such that for $k \geq k_0$ and any Stiefel–Whitney class α which is the reduction of a (1, 1) class $\hat{\alpha}$, we have $q_{k,\hat{\alpha}}(h) > 0$.*

This obviously implies Theorem (1.3.5), and combined with the results of Chapter 9 gives

Corollary (10.1.2). *No simply connected complex algebraic surface S can be written as a smooth connected sum $X_1 \# X_2$ with $b^+(X_1)$ and $b^+(X_2)$ both positive.*

From this we immediately obtain many examples of distinct smooth four-manifolds with the same intersection forms. Indeed we have:

Corollary (10.1.3) {= (1.3.7)}. *For any simply connected complex algebraic surface S with $b^+(S) \geq 5$, there is a smooth four-manifold $X(S)$, homotopy equivalent to S but not diffeomorphic to S, nor to any complex surface.*

For manifolds with odd forms this is immediate; by the classification of forms we can take

$$X(S) = \mathbb{CP}^2 \# \mathbb{CP}^2 \# \mathbb{CP}^2 \# \ldots \# \mathbb{CP}^2 \# \overline{\mathbb{CP}}^2 \# \ldots \# \overline{\mathbb{CP}}^2 \# \overline{\mathbb{CP}}^2$$

(and we need only assume $b^+ \geq 3$). In the other case, we consider a connected sum of the form $X = lK \# m(S^2 \times S^2)$ or $l\bar{K} \# m(S^2 \times S^2)$, where K is the $K3$ surface and $\bar{K}$ is the same manifold with reversed orientation. By Rohlin's theorem and the classification of even forms we can arrange X to have the same homotopy type as S so long as S satisfies the '11/8 inequality',

$$b^+ + b^- \leq (11/8)|b^+ - b^-|.$$

This inequality can be deduced from standard results about surfaces (an observation of Moishezon). One easily reduces to minimal surfaces of general type, for which $c_1^2 \geq 0$, and $c_1^2 \leq 3c_2$. These give $b^+ - 2b^- \leq 1$ and $5b^+ - b^- \geq -4$ (using the formulae of Section 1.1.7) which, together with the special consideration of small values of b^+, b^-, imply the result.

In any case, S cannot be diffeomorphic to $X(S)$, by Corollary (10.1.3). Of course, as well as these general results we have many explicit examples, for example the hypersurfaces S_d ($d \geq 5$) and the branched covers R_p ($p \geq 4$) of Section 1.1.7.

Notice that, combining (10.1.2) with the non-existence results of Chapter 8, in any smooth connected sum splitting of a simply connected algebraic surface one of the summands must have intersection form $\mathrm{diag}(-1, -1, -1, \ldots, -1)$, i.e. the intersection form of a connected sum of $\overline{\mathbb{CP}}^2$s. Conversely, this situation certainly can occur: we can take S to be the multiple blow up of another complex surface S'; then, as we have seen in Chapter 1:

$$S = S' \# \overline{\mathbb{CP}}^2 \# \overline{\mathbb{CP}}^2 \ldots \# \overline{\mathbb{CP}}^2.$$

Now let us go back to our remarks about Wall's theorem on the stable classification of four-manifolds from Chapter 1. We know that for any simply connected complex surface S and sufficiently large integers m, n, the manifold $S \# m\mathbb{CP}^2 \# n\overline{\mathbb{CP}}^2$ is diffeomorphic to $X(S) \# m\mathbb{CP}^2 \# n\overline{\mathbb{CP}}^2$. But we see now that we cannot take $m = 0$, however large n may be, since the connected sum of S with $n\overline{\mathbb{CP}}^2$ is still an algebraic surface. In the reverse direction it has been shown by Mandelbaum (1980) and Moishezon (1977) that for many surfaces S, for example the hypersurfaces in $\mathbb{CP}^3$, we can take $n = 0$ and $m = 1$; i.e. adding a single $\mathbb{CP}^2$ kills the 'exotic' property of the differentiable structure on S. (Such surfaces are called 'almost completely decomposable' by Mandelbaum and Moishezon, and they conjecture that in fact all surfaces have this property.) Thus we see that there is a radical difference between the differential-topological effect of the addition of $\mathbb{CP}^2$s and $\overline{\mathbb{CP}}^2$s. Of course this is quite in line with our methods, based on the ASD equation which involves a definite choice of orientation. Indeed we see that for almost completely decomposable surfaces S with b^+, b^- both odd the ASD Yang–Mills invariants of $\bar{S}$ are all zero (by (9.3.14)) while those of S are not.

10.1.2 The main idea

Let S be a complex algebraic surface as above and fix a Kähler metric on S, for simplicity a 'Hodge metric', compatible with a projective embedding. Thus the de Rham cohomology class of the metric 2-form ω is Poincaré dual to the hyperplane section class ω. (For example we could take the pull-back of the Fubini–Study metric on $\mathbb{CP}^N$.) We shall now explain why (10.1.1) is true in a favourable (but rare) case when we have a moduli space M of ASD connections, defined relative to this Kähler metric, which is non-empty, compact and regular, i.e. $H_A^2 = 0$ for all points in M. The argument is very simple: the number $q(h)$ is defined by pairing the relevant power of the cohomology class $\mu(h)$ with the fundamental cycle of M:

$$q(h) = \langle \mu(h)^d, [M] \rangle. \tag{10.1.4}$$

Now we know on the one hand, from (5.2.19), that $\mu(h)$ can be represented by the 2-form on M,

$$\Omega(a, b) = \frac{1}{8\pi^2} \int_S \operatorname{Tr}(a \wedge b) \wedge \omega,$$

since the metric form ω is self-dual. Thus, using de Rham cohomology,

$$q(h) = \int_M \Omega^d. \tag{10.1.5}$$

On the other hand we know that M is a complex manifold and that Ω is the two-form of the natural Kähler metric on M (see Section 6.5.3). It follows then that Ω^d is $(d-1)!$ times the Riemannian volume element on M, so

$$q(h) = (d-1)! \operatorname{Vol}(M) > 0,$$

as required.

This proof does not use the fact that the metric is a Hodge metric; it applies to any Kähler surface. However in the algebraic case we can carry through a parallel algebro-geometric argument. We know that $-2\pi i\Omega$ is the curvature form of a line bundle $\mathscr{L}$ over M, the determinant line bundle of the restriction of the connections to a (real) surface representing h. Now since $\mathscr{L}$ is a 'positive' line bundle, we know by Kodaira's embedding theorem that the holomorphic sections of some positive power $\mathscr{L}^{\otimes n}$ give a projective embedding $j: M \to \mathbb{CP}^r$. So the image $j(M)$ is a complex algebraic subvariety of $\mathbb{CP}^r$, and the restriction of the Hopf line bundle on $\mathbb{CP}^r$ to $j(M)$ is isomorphic to $\mathscr{L}^n$. It follows then that $q(h)$ is $1/n^d$ times the *degree* of $j(M)$: the degree being the number of points in the intersection of $j(M)$ with d general hyperplanes in $\mathbb{CP}^r$. Since the degree of a non-empty variety is always positive we reach the same conclusion. Notice that if we can construct the projective embedding directly, without recourse to Kodaira's theorem, we obtain an independent algebro-geometric proof, and we shall use this approach in Section 10.1.4 to handle the technical difficulties of singularities and non-compact moduli spaces which occur in realistic problems. First we digress to describe projective embeddings of moduli spaces of bundles, beginning with bundles over curves.

10.1.3 Gieseker's projective embedding

Let C be an algebraic curve (compact Riemann surface) of genus g and fix a line bundle $\mathcal{O}(1)$ of degree one over C. Let p be a positive integer, to be fixed below, and consider two moduli spaces W_0, W_1 of rank-two stable holomorphic bundles $\mathcal{E}$ over C with determinant $\mathcal{O}(2p)$ (the tensor power $\mathcal{O}(1)^{2p}$) and $\mathcal{O}(2p+1)$ respectively. The operation of tensoring with $\mathcal{O}(1)$ shows that these are independent of p. Each can be described in terms of projectively flat unitary connections; the space W_0 is the same as the space denoted W_C in Section 6.1.4. Both W_0 and W_1 are complex manifolds of complex dimension $3g-3$, and W_1 is also compact. We want to construct projective embeddings of W_0 and W_1 by sections of the determinant line bundle $\mathcal{L}$. In the case of W_1 we get such an embedding by Kodaira's theorem, since we know that $\mathcal{L}$ is a positive line bundle, but as it stands this does not cover the non-compact space W_0. We will describe here an algebro-geometric construction due to Gieseker which is rather more explicit and has the advantage that it works equally well in the non-compact case. In this construction we shall see another way in which the notion of stability for bundles can be fitted into the general theory described in Section 6.5, involving quotients by linear actions. As in the case of the ADHM construction of instantons, the relevant symmetry group is the automorphism group of the cohomology of a bundle.

Recall that, topologically, any bundle can be induced from a universal bundle U over a Grassman manifold by a suitable map. Gieseker's construction takes as starting point the analogue of this for holomorphic bundles. Suppose a rank-two bundle $\mathcal{E}$ over C is *generated by its global sections*, i.e. for each point x in C the evaluation map $e_x\colon H^0(\mathcal{E}) \to \mathcal{E}_x$ is surjective. If we write H for $H^0(\mathcal{E})$ and $\underline{H}$ for the corresponding trivial bundle over C we get a surjective bundle map

$$e\colon \underline{H} \longrightarrow \mathcal{E}.$$

Thus the bundle $\mathcal{E}$ is completely described by giving a family of two-dimensional quotients of the fixed vector space H, or equivalently by a family of two-dimensional subspaces of the dual space H^*. We get a canonical map to the Grassmannian of two-planes:

$$f\colon C \longrightarrow \mathrm{Gr}_2(H^*), \tag{10.1.6}$$

with $f(x)$ the annihilator of $\ker e_x$. The universal bundle U over the

Grassmannian (the dual of the tautological bundle) is holomorphic, and $\mathscr{E}$ is canonically isomorphic to the pull-back $f^*(U)$.

We next give a rather different description of f. First we apply the standard Plücker embedding i of a Grassmannian. This maps $\mathrm{Gr}_2(H^*)$ to the projective space $\mathbb{P}(\Lambda^2 H^*)$, with

$$i(\mathrm{Span}(\varepsilon_1, \varepsilon_2)) = [\varepsilon_1 \wedge \varepsilon_2].$$

Thus we have a composite map

$$g = if : C \longrightarrow \mathbb{P}(\Lambda^2 H^*).$$

Now, maps to a projective space correspond to linear systems of sections of line bundles. The pull back by g of the Hopf line bundle over the projective space is canonically isomorphic to the line bundle $\Lambda^2\mathscr{E}$ over C. The map g must therefore be induced from the universal map,

$$u : C \longrightarrow \mathbb{P}(H^0(\Lambda^2\mathscr{E})^*)$$

by a linear mapping:

$$\sigma^T : H^0(\Lambda^2\mathscr{E})^* \longrightarrow (\Lambda^2 H)^*.$$

More precisely, any such linear map induces a rational map on the projective spaces and this becomes a well-defined mapping on the one-dimensional curve C (provided σ^T is not identically zero). So we have a commutative diagram:

$$\begin{tikzcd}
C \arrow[r, "f"] \arrow[dr, "u"'] & \mathrm{Gr}_2(H^*) \arrow[r, "i"] & \mathbb{P}(\Lambda^2 H^*) \\
& \mathbb{P}(H^0(\Lambda^2\mathscr{E}))^* \arrow[ur, "\mathbb{P}(\sigma^T)"'] &
\end{tikzcd}$$

Lemma (10.1.7). *The map* σ^T *is the transpose of the tautological map* $\sigma : \Lambda^2 H = \Lambda^2(H^0(\mathscr{E})) \to H^0(\Lambda^2\mathscr{E})$.

This is just a matter of checking definitions, and we can safely omit the proof.

It follows then that our bundle is completely determined by the associated map σ. To apply this to moduli problems we consider families of bundles $\mathscr{E}$ with the same determinant $\Lambda^2\mathscr{E}$, and with a fixed dimension, N say, of $H^0(\mathscr{E})$. Then to each bundle we can associate an orbit in the vector space $\mathrm{Hom}(\Lambda^2\mathbb{C}^N, H^0(\Lambda^2\mathscr{E}))$ under the natural action of $GL(N, \mathbb{C})$. For bundles which are generated by their global sections, this orbit determines the bundle up to isomorphism.

Lemma (10.1.8). *A stable bundle* $\mathscr{E}$ *over* C *with* $\deg(\mathscr{E}) \geq 4g + 2$ *is generated by its global sections and has* $\dim H^0(\mathscr{E}) = \deg(\mathscr{E}) - 2(g - 1)$.

Proof. For any point x in C we have an exact sequence:

$$0 \longrightarrow \mathscr{E} \otimes [-x] \longrightarrow \mathscr{E} \underset{e_x}{\longrightarrow} \mathscr{E}_x \longrightarrow 0,$$

where $[-x]$ denotes the line bundle of degree -1 defined by x. The long exact cohomology sequence shows that e_x is surjective if $H^1(\mathscr{E} \otimes [-x]) = 0$. But this space is dual to $H^0(\mathscr{E}^* \otimes K_C \otimes [x])$, i.e. to the maps from $\mathscr{E}$ to the line bundle $K_C \otimes [x]$, which has degree $2(g-1)+1$. By the definition of stability, such maps exist only if $\frac{1}{2}\deg \mathscr{E} < 2(g-1)+1$. Similarly, when $\deg \mathscr{E} \geq 4g-4$ we have $H^1(\mathscr{E}) = 0$ and $\dim H$ is given by the Riemann–Roch formula.

We thus obtain the following proposition, in which we let W denote either of the moduli spaces W_0, W_1, we write $\mathbb{C}^q$ for $H^0(\mathcal{O}(2p))$, $H^0(\mathcal{O}(2p+1))$ and put $N = 2p - 2(g-1)$, $(2p+1) - 2(g-1)$ respectively, where p is any integer bigger than $2g-1$.

Proposition (10.1.9). *There is a natural injection γ of W into the set of orbits of $SL(N, \mathbb{C})$ in the projective space $\mathbb{P}(\mathrm{Hom}(\Lambda^2\mathbb{C}^N, \mathbb{C}^q))$.*

Gieseker's approach thus arrives squarely in the class of problems we discussed in Section 6.5.2. If we restrict the $SL(N, \mathbb{C})$ action to the open set of stable orbits in the projective space, the quotient U^* becomes a complex manifold (or orbifold) and the invariant polynomials, of fixed large degree s say, induce an embedding of U^* into a projective space $\mathbb{CP}^r$. What needs to be shown is that the injection γ maps W to the subset U^*, i.e. that stable bundles $\mathscr{E}$ have stable maps $\sigma_{\mathscr{E}}$. We omit the proof of this and refer to the very readable account by Gieseker (1977). In outline, one applies the Hilbert criterion to show that if $\sigma_{\mathscr{E}}$ is not stable and is destabilized by the one-parameter subgroup associated with weights w_i and a basis s_i for $H^0(\mathscr{E})$, then for all pairs with $w_i + w_j > 0$, s_i and s_j lie in a common line sub-bundle. Then the Riemann–Roch formula can be used to show that the sub-line bundle associated to the largest weight destabilizes $\mathscr{E}$ as a bundle. The conclusion is that there is a projective embedding j: $W \to \mathbb{CP}^r$ obtained by composing γ with the embedding of U^*. (Similarly one finds that semi-stable bundles $\mathscr{E}$ have semi-stable maps $\sigma_{\mathscr{E}}$.)

We want now to show that the line bundle associated to this Gieseker embedding j can be identified with the power $\mathscr{L}^s$ of the determinant line bundle over W. We consider the even case for simplicity. Recall that we defined the determinant line bundle over the moduli space of bundles V with $\Lambda^2 V$ trivial by the index of the Dirac operator, which could be identified with the $\bar{\partial}$ operator after twisting by a square root of the canonical bundle. We will now clarify the role of this twisting. Observe first that the moduli space of bundles with $\Lambda^2 V$ trivial, up to bundle isomorphism, is identical with the moduli space of pairs (V, ψ) where ψ is an isomorphism from $\Lambda^2 V$ to $\mathcal{O}$. This is because the scalars act transitively on the maps ψ. So when classifying such bundles we may suppose we have a definite trivialization of Λ^2. Now for any line bundle θ over C we can assign to the pair (V, ψ) the line obtained from the cohomology of $V \otimes \theta$. If V is stable, say, the only automorphisms of (V, ψ) are ± 1, and since the numerical Euler characteristic of V is even this acts trivially on the determinant line. So, as in Chapter 5, we obtain a line bundle, $\mathscr{L}_\theta$ say, over the moduli space of stable holomorphic bundles with a trivialization of Λ^2. At first sight it may seem that we are obtaining different line bundles in this way, since the cohomology groups of $V \otimes \theta$ are not isomorphic for different θ. In reality, however, all these line bundles are isomorphic. To see this we represent line bundles by divisors on C. Suppose for example that $\theta = [P]$ is the line bundle associated with a point P in C. Then we have an exact sequence:

$$0 \longrightarrow V \longrightarrow V \otimes \theta \longrightarrow V \otimes \theta|_P \longrightarrow 0. \qquad (10.1.10)$$

The resulting long exact cohomology sequence, and the usual property of the 'multiplicative Euler characteristic' in exact sequences, gives a natural isomorphism between the determinant line of the cohomology of $V \otimes \theta$ and the tensor product of the determinant line of the cohomology of V and $\Lambda^2(V \otimes \theta)_P$. Fix a trivialization of the fibre of θ over P. Then the trivialization of $\Lambda^2 V$ gives us an isomorphism between the two determinant lines, and hence an isomorphism between $\mathscr{L}_\theta$ and $\mathscr{L}_0$. Repeating this procedure, we see that all the line bundles $\mathscr{L}_\theta$ are isomorphic. So we can use any twist to describe our basic line bundle $\mathscr{L}$ over the moduli space W_0.

With this preliminary observation we return to the line bundle induced by Gieseker's embedding—the pull-back by j of L_U, the quotient of the Hopf line bundle over $\mathbb{P}(\mathrm{Hom}(\Lambda^2\mathbb{C}^N, \mathbb{C}^q))$. We let $\mathscr{E} = V \otimes \mathscr{O}(p)$, and by the remarks above we suppose we have a fixed isomorphism between $\Lambda^2\mathscr{E}$ and $\mathscr{O}(2p)$. Now we introduce the moduli space of triples (V, ψ, f) where ψ is a trivialization of $\Lambda^2 V$ and f is an isomorphism:

$$f: H^0(V \otimes \mathscr{O}(p)) \longrightarrow \mathbb{C}^N.$$

Given such a triple we get a point $\sigma_{V,\psi,f}$ in $\mathrm{Hom}(\Lambda^2\mathbb{C}^N, \mathbb{C}^q)$ whose orbit is $j([V])$. Changing f to ζf, where ζ is in $\mathbb{C}^*$, changes $\sigma_{V,\psi,f}$ to $\zeta^{-1}\sigma_{V,\psi,f}$. Here we come to a point which we hurried past in Section 6.5. The centre $\mathbb{Z}/N$ of $SL(N, \mathbb{C})$ acts trivially on the projective space, so the Hopf line bundle need not descend to the quotient. However its N'th power will do, i.e. the line bundle L_U over the quotient lifts to this power of the Hopf bundle. So, by the discussion above, the choice of f gives a basis element in the fibre of the pull-back of L_U, and multiplication of f by ζ multiplies this basis element by ζ^N. On the other hand, the choice of f gives a basis element in the determinant line $\det H^0(V \otimes \mathscr{O}(p))$, and multiplication of f by ζ changes this by ζ^{-N}.

We deduce that there is a natural isomorphism between the line bundles $j^*(L_U)$ and the $\mathscr{L}_{\mathscr{O}(p)}$, the latter being the bundle with fibres $\det H^0(V \otimes \mathscr{O}(p))$ (since the other cohomology group is zero). In sum then we obtain a projective embedding j of W_0 by holomorphic sections of some high power sN of the determinant line bundle $\mathscr{L}$ over W_0.

There is one last fact to mention. The points of W_0 represent stable bundles and the holomorphic sections which give Gieseker's projective embedding can be viewed as holomorphic sections of a line bundle $\mathscr{L}$ over an open subset $\mathscr{A}_s$ of the space of connections ($\bar{\partial}$ operators) over C, invariant under the natural action of the complexified gauge group $\mathscr{G}^c$. However, these sections can all be extended holomorphically over the whole space $\mathscr{A}$. One can see this by examining the definition of the sections carefully or, more directly, by using the fact that (if $g > 0$) the complement of $\mathscr{A}_s$ has complex co-dimension two or more. We know then that for any stable point in $\mathscr{A}$ there is a holomorphic, invariant, section which does not vanish at that point. The last fact we need is that the same is true for the trivial connection or, more generally, for any connection which defines a semi-stable holomorphic bundle over C. The point $\sigma_{V \otimes \mathcal{O}(p)}$ defined by such a connection is semi-stable for the $SL(N, \mathbb{C})$ action, so this follows from the corresponding piece of theory for the finite-dimensional quotient.

10.1.4 Technical facts about moduli spaces

We will now begin our description of the detailed proof of (10.1.1) which will rely on a number of general facts about moduli spaces, which we will marshall in this section. The first point is that the ASD moduli spaces are *non-empty* when k is sufficiently large. This is true for any Riemannian four-manifold and was proved by Taubes (1984) for $SU(2)$-bundles and (as a special application of the results) by Taubes (1989) for $SO(3)$-bundles. In the algebro-geometric context the existence of stable holomorphic $SL(2, \mathbb{C})$ bundles was proved by Gieseker (1988); results on general two-plane bundles can be obtained using the Serre construction discussed in Section 10.2.

The next point concerns the dimension of the moduli space of stable bundles, i.e. the ASD connections relative to a Kähler metric. We know that for *generic* Riemannian metrics the moduli space of irreducible connections is regular, and is therefore smooth and of the proper 'virtual' dimension. But the Kähler metrics are not generic and it certainly may happen that we encounter moduli spaces with singularities of various kinds, or with components which have dimension larger than the virtual dimension. In the latter case the argument we have given above for the positivity of $q(h)$ is certainly not valid. The situation is quite analogous to the familiar intersection theory of subvarieties: if P, Q are complex subvarieties of an ambient compact complex manifold V, with $\dim V = \dim P + \dim Q$, and if the intersection of P and Q consists of isolated points, then the topological intersection number $[P].[Q]$ is positive. In fact $[P].[Q]$ is at least the number of intersection points. But if $P \cap Q$ has dimension one or more it may be that $[P].[Q]$ is negative. For example, we could take $P = Q$ to be an 'exceptional curve' on a surface, with self-intersection -1.

To avoid this difficulty we look at bundles with k large. Then it can be proved that each component of the moduli space has the proper dimension.

Precisely, let us define for the fixed Hodge metric:

$$\Sigma_{k,\alpha} = \{[A] \in M_{k,\alpha} | H_A^2 \neq 0\}.$$

Σ is a complex-analytic subspace of the complex space M_k and we have:

Proposition (10.1.11). *There are constants B_1, B_2 such that for all k and all reductions α of* (1, 1) *classes:*

$$\dim \Sigma_{k,\alpha} \leq 3k + B_1 k^{1/2} + B_2.$$

This is proved by Donaldson (1990*a*) for $SU(2)$ bundles, but the argument extends easily to the general case.

Thus the dimension of Σ grows more slowly than the virtual complex dimension $4k - 3(1 + p_g)$ of M_k. On the other hand, if the dimension of a component of M exceeds the virtual dimension, then this component is contained in Σ. It follows that for large k the moduli space does indeed have the proper dimension. Moreover the singular set has large codimension in M_k.

The third fact we want is that the moduli spaces, viewed as moduli spaces of stable bundles on our projective algebraic surface S, are themselves naturally complex varieties, in the sense of abstract algebraic geometry. This is the algebro-geometric analogue of the theory we studied from the transcendental point of view in Section 6.4, that the moduli spaces are Hausdorff complex spaces. It is a generalization of the theorem that holomorphic bundles over S are necessarily algebraic. What is asserted is, first, that the local versal deformations of algebraic vector bundles over S can be realized as algebraic vector bundles over $S \times V_\lambda$, where V_λ is a 'quasi-affine' variety (the difference $V \backslash W$ of affine varieties), and that the gluing maps

$$\begin{tikzcd}
\phi_{\lambda\mu}: V_{\lambda\mu} \arrow[r] \arrow[d, phantom, "\cap" rotate=0] & V_{\mu\lambda} \arrow[d, phantom, "\cap"] \\
V_\lambda & V_\mu
\end{tikzcd}$$

are represented by rational functions. Second, the statement that M_k is a complex variety asserts that M_k has 'finite type', i.e. that it can be covered by a finite number of quasi-affine patches V_α. This is closely related to our compactness theorem for moduli spaces of ASD connections. For proofs of these assertions see Maruyama (1977) and Gieseker (1977).

Next we want to study the restriction of stable holomorphic bundles on S to curves. The main theorem we need is the following result from Mehta and Ramanathan (1984).

Proposition (10.1.12) (Mehta and Ramanathan). *For any stable bundle $\mathscr{E}$ on S there is an integer p_0 such that, for $p \geq p_0$ and generic curves C in the linear system $|\mathscr{O}(p)|$ on S, the restriction of $\mathscr{E}$ to C is also stable.*

Now the stability condition is itself open, so for each (smooth) curve C in a linear system $|\mathcal{O}(p)|$ we have a Zariski open subset $U_C \subset M_k$ consisting of bundles whose restriction to C is stable. The proposition asserts that the union of the U_C, over all p and C, is the whole of M_k. But the finite type condition asserts that M_k is compact in the Zariski topology, so we can find a finite cover:

$$M = \bigcup_{i=1}^{R} U_i, \quad U_i = U_{C_i}. \qquad (10.1.13)$$

Moreover we can suppose that the curves C_i are in the same linear system $|\mathcal{O}(p)|$ and are in general position, with all triple intersections empty. Similarly we can suppose that for all pairs $[\mathcal{E}]$, $[\mathcal{F}]$ in M_k the cohomology group $H^1(\mathrm{Hom}(\mathcal{E}, \mathcal{F})(-p))$ is zero. Finally, replacing the given projective embedding by that defined by $|\mathcal{O}(p)|$, we may as well suppose that $p = 1$ and the curves C_i are hyperplane sections.

10.1.5 Restriction to curves

We can now construct a projective embedding of M_k, using restriction to complex curves in the surface S. We first consider the situation from the point of view of determinant line bundles. Suppose first that S is spin, so there is a square root $K_S^{1/2}$ of the canonical bundle. Let C be a curve in the linear system $|\mathcal{O}(2d)|$. There is an induced square root of K_C and an exact sequence:

$$0 \longrightarrow \mathcal{O}(-d) \otimes K_S^{1/2} \longrightarrow \mathcal{O}(d) \otimes K_S^{1/2} \longrightarrow K_C^{1/2} \longrightarrow 0. \qquad (10.1.14)$$

For any bundle $\mathcal{E}$ over S we get, just as for (10.1.10), an induced isomorphism of determinant lines:

$$\{\det H^*(\mathcal{E} \otimes K_S^{1/2} \otimes \mathcal{O}(d))\}\{\det H^*(\mathcal{E} \otimes K_S^{1/2} \otimes \mathcal{O}(-d))\}^{-1}$$
$$= \det H^*(\mathcal{E}|_C \otimes K_C^{1/2}).$$

The term on the right is the fibre of the determinant line bundle defined by restriction to C, whereas the expression on the left is independent of C. We deduce that restriction to different curves gives isomorphic line bundles over the ASD moduli space. This is an algebro-geometric version of Proposition (7.1.16). It is easy to remove the spin condition, and the requirement that the homology class of C be even, by introducing twisting factors just as in Section 10.1.3.

To sum up then, for each i we have a smooth moduli space $W_i = W(C_i)$, a determinant line bundle $\mathcal{L}_i$ over W_i and a projective embedding:

$$J_i \colon W_i \longrightarrow \mathbb{P}_i = \mathbb{P}(G_i^*)$$

associated with a vector space G_i of sections of $\mathcal{L}_i^N$. Now the restriction map

$$r_i \colon U_i \longrightarrow W_i$$
$$r_i(\mathcal{E}) = \mathcal{E}|_{C_i}$$

is regular, so we get composite maps $J_i r_i \colon U_i \to \mathbb{P}_i$. On M_k we have the fixed line bundle $\mathscr{L}$, and a holomorphic isomorphism from $\mathscr{L}$ to $\mathscr{L}_i$. So we can regard the G_i as spaces of sections of $\mathscr{L}^N$ over U_i, and these extend holomorphically over M_k by the remark at the end of Section 10.1.3. Hence we now have a space $G = \oplus\, G_i$ of sections of $\mathscr{L}^N$ over M_k, and for each point $\mathscr{E}$ in M_k there is a section in G not vanishing at $\mathscr{E}$. But this just means that we have an induced map,

$$J \colon M_k \longrightarrow \mathbb{P}(G^*) \quad \text{with } J^*(\mathcal{O}(l)) = \mathscr{L}^N.$$

This also holds with $k = 0$, when we have the moduli space with one point representing the trivial bundle, by the remark at the end of Section 10.1.3. The maps $J_i r_i$ are of course the composites of J with the projection maps to the individual factors. They are rational maps on M_k, regular on U_i.

To see that J is an embedding we use the vanishing of the cohomology groups $H^1(\mathrm{Hom}(\mathscr{E}, \mathscr{F})(-1))$. This implies that any two bundles $\mathscr{E}, \mathscr{F}$ which become isomorphic when restricted to some C_i are already isomorphic over S. Then the fact that J is injective follows from the corresponding property of the J_i. Similarly, taking $\mathscr{E} = \mathscr{F}$, we see that J is an immersion. As a final technicality, we can arrange the given set of surfaces to have the desired properties for all the finite number of moduli spaces M_j for $j \leq k$. So we have embeddings, which we will still denote by J, of M_j in the same projective space, and similarly $J_i \colon M_j \to \mathbb{P}_i$.

10.1.6 The detailed argument

By examining Gieseker's algebraic construction one sees that the embedding J of the algebraic variety is defined by a rational map. It is a general fact that the image of an abstract complex variety under a projective embedding by a rational map is a *quasi-projective* variety, i.e. the difference $V_1 \setminus V_2$ of projective varieties. For any quasi-projective variety $Y \subset \mathbb{CP}^r$ we define the degree $\deg(Y)$ to be the degree of the projective variety $\bar{Y}$; it is the number of intersection points of a generic $\overline{\mathbb{CP}^q}$ $(q = r - \dim \Omega)$ with Y, and if Y is non-empty the degree is always strictly positive. So the proof of our main theorem (10.1.1) is completed by the next lemma:

Lemma (10.1.15).

(i) *Let* $J \colon M_k \to \mathbb{P}(G)$ *be the projective embedding, defined by sections of* $\mathscr{L}^N$, *constructed above. Then the degree of the image* $J(M_k)$ *is independent of the choice of curves* C_i.

(ii) *For sufficiently large* k, $q_k([C_i]) = (1/N^d)\deg(J(M_k))$.

Proof. To prove (i) it suffices to show that if we extend a given collection of curves $C_1, \ldots, C_L$ by another one C_{L+1} then the degree of the resulting projective embeddings is unchanged. But this follows immediately from the general fact that if a projection $\mathbb{P}^{n+s} \to \mathbb{P}^n$ (a rational map) restricts to a subvariety $Q \subset \mathbb{P}^{n+s}$ to give an embedding $\pi: Q \to \mathbb{P}^n$, then the degrees of Q and its image $\pi(Q)$ are equal. To prove (ii) we trace through the definition of our polynomial invariant q_k using restriction to surfaces, and make an appropriate choice of the curves C_i. Let us suppose inductively that we have chosen curves C_i ($i = 1, \ldots, l$) in general position, and sections g_i in G_i so that the common zero set

$$Z_l^{(k)} = Z(g_1, \ldots, g_l) = \{[\mathscr{E}] \in M_k \,|\, g_i(\mathscr{E}) = 0, \quad i = 1, \ldots, l\}$$

has the following good properties:

(i) $Z_l^{(k)}$ has the correct dimension $d - l$,

(ii) $Z_l^{(k)} \cap \Sigma$ has the correct dimension $\dim \Sigma - l$.

(iii) On a dense open subset of $(M_k \backslash \Sigma) \cap Z_l^{(k)}$ the zeros of g_i are transverse.

(iv) For all the lower moduli spaces M_i the zero sets

$$Z_l^{(i)} = \{[\mathscr{E}] \in M_i \mid g_j(\mathscr{E}) = 0, \quad j = 1, \ldots, l\}$$

have dimension $\dim M_i - l$.

To pass from l to $l + 1$ we note that, as quasi-projective varieties, all of the $Z_l^{(i)}$ and $Z_l^{(k)} \cap \Sigma$ have finitely many components. We choose a point in each component, i.e. a finite set of bundles $\mathscr{E}_\lambda$. Then we can, by (10.1.12), choose a curve C_{l+1} such that all the $\mathscr{E}_\lambda$ are stable on C_{l+1}. Thus the generic hyperplane section of $\mathbb{P}_{l+1}$ induces a section g_{l+1} which does not vanish at any of the $\mathscr{E}_\lambda$; then the zero sets $Z_{l+1}^{(i)}$ do not contain any component of the $Z_l^{(i)}$ and it follows that they have the proper dimension. So, inductively, we can choose a set of curves $C_1, \ldots, C_d$ and sections $g_1, \ldots, g_d$ such that the common zero set in M_k is a finite set of points $\{E_\mu\}$, none contained in Σ, and with the general position properties (i)–(iv) with respect to the lower moduli spaces.

Now on the one hand we can extend this collection of curves, if necessary, to get a projective embedding of M_k. The g_i represent hyperplane sections of $J(M_k)$ and the number of points in the intersection, counted with appropriate multiplicities, represents the degree of $J(M_k)$ *provided* there are no 'zeros at infinity', i.e. no common zeros of all the g_i in $\overline{J(M_k)} \backslash J(M_k)$. On the other hand, from the point of view of the general set-up in Chapters 5 and 9, each g_i represents a section of the determinant line bundle $\mathscr{L}_i$, and its zero set represents a codimension-two submanifold V_{C_i} in the space of connections. The intersection $M_k \cap V_{C_1} \cap \ldots \cap V_{C_d}$ is by construction a finite set of points $\{A_\mu\}$. To prove the equality of $q_k(\alpha)$ and $\deg(J(M_k))$ we have to check three things:

(*a*) That the multiplicities with which we count the points, regarded either as intersections in the projective space P or in the space of connections, agree. That is, if we perturb the Kähler metric to a nearby generic metric and perturb the V_{C_i} to be transverse, as in the definition of Section 9.2.1, the intersections A_μ split up into the correct number of nearby transverse intersection points.

(*b*) That when we perturb the metric and V_{C_i} as in Section 10.1.1, all the new intersection points are close to the A_μ.

(*c*) That in the projective space $\mathbb{P}$ there are no common zeros of the g_i in $\overline{J(M_k)}\backslash J(M_k)$.

Of these (*a*) is quite straightforward; it is just the assertion that the local multiplicity of the zero set equals the multiplicity of any transverse deformation. Points (*b*) and (*c*) are more interesting and the proof is much the same in each case. We begin with (*c*), and suppose that on the contrary there is a zero in $\overline{J(M_k)}\backslash J(M_k)$. Then we can find a sequence of holomorphic bundles $\mathscr{E}_\alpha$ over S without convergent subsequences but with all $g_i(\mathscr{E}_\alpha)$ tending to zero as α tends to ∞. Regarded as ASD connections A_α we can apply our compactness theorem and, without loss of generality, suppose that A_α tends to a limiting connection A_∞ on the complement of a finite set of points $\{x_\nu\}$. The limiting connection has Chern class $l < k$ say, and there are at most $k - l$ points x_ν. We now apply our familiar argument: for any curve C_i which does not contain any of the points x_ν the restriction of the connections converges in C^∞; it follows that for any such i, $g_i(A_\infty)$ is zero. But we have arranged that all the multiple zeros of the g_i in the lower moduli spaces are of the appropriate dimension (property (iv) above), so if d_l is the complex dimension of M_l we must have:

$$d_l + 2(k - l) \geq d = 4k - 3(1 + p_g). \tag{10.1.16}$$

In the present situation we do not know that all the lower moduli spaces have the correct (virtual) dimension. But we do know that this holds for large enough values, $j \geq k_0$, of the Chern class, by (10.1.11). We put

$$D = \max_{j \leq k_0} \dim_{\mathbf{C}} M_j.$$

Then if $2k > D + 3(1 + p_g)$ (say) the inequality (10.1.16) cannot hold, thus verifying property (*c*). Precisely the same argument shows that for a sequence of metrics $g^{(\alpha)} \to g$ and sections $g_i^{(\alpha)} \to g_i$, the common zeros in the perturbed moduli space $M_k(g^{(\alpha)})$ converge to the A_μ, verifying property (*b*).

10.2 Construction of holomorphic bundles

10.2.1 Extensions

In this section we will describe general techniques for constructing holomorphic bundles out of linear data. Consider first a complex manifold Z and

, " " 8 #
L@ % $ & ### ! ! L@ M 1
!

! " " # # $! %$ & ' ! !
" " " # # $! (
)*+,)))-) % . / 0
1 2 31 3 4 1/ 5 1 6"
)*+,)))-) % 0 078&*9 -:0+7 ;< &
"/ . " 2 31 3 4 1/ 5 1 6"
' !

Relativistic Mass

As the relative speed between two inertial reference frames gets closer to the speed of light, the effects of special relativity—including time dilation and length contraction—become more pronounced. In the limit that the relative speed approaches the speed of light, time slows down to a complete stop and length contracts to zero. However, you **can't** actually reach this limit. An object that has mass can be accelerated to nearly light speed (like 0.99c or 0.999c), but can never reach the speed of light exactly.

In Galilean relativity, it would be very easy to accelerate an object faster than the speed of light. According to **Newton's second law** of motion, the net force acting on an object equals the object's mass times its acceleration:

$$\sum \vec{\mathbf{F}} = m\vec{\mathbf{a}}.$$

(This equation applies to objects that have constant mass, but that's not a problem for Galilean relativity, where ordinarily an object's mass isn't expected to change while it accelerates.) The mass of an electron is 9.1×10^{-31} kg. If you applied a force of just 1 N (one Newton) to an electron, according to Galilean relativity, the electron would experience an acceleration of

$$a = \frac{F}{m} = \frac{1\ \mathrm{N}}{9.1\times10^{-31}\mathrm{kg}} = 1.1 \times 10^{30}\mathrm{m/s^2}.$$

What does this **acceleration** mean? It means that starting from rest, after a just 1 s (one second), an electron would have a speed of $v = 1.1 \times 10^{30}$ m/s (since acceleration describes the rate at which velocity increases). That's way, way faster than the speed of light in vacuum, which is $c = 2.9979 \times 10^{8}$ m/s.

In the laboratory, it doesn't happen that way. Although it is very easy to accelerate electrons to very high speeds, once the speed of an electron reaches the neighborhood of the speed of light in vacuum, it becomes increasingly harder to accelerate the electron. We can accelerate electrons up to 0.9c or even 0.99c, but trying to reach 0.99999c is extremely difficult. Why? According to Einstein's theory of relativity, mass isn't constant: The faster an electron travels, the more mass the electron has (relative to the laboratory). We call this **relativistic mass**.

Recall that mass is a measure of **inertia** in the following sense: The more mass an object has, the more difficult it is to overcome the object's inertia in order to accelerate the object. The relativistic mass of an object describes the object's inertia. As the object travels closer to the speed of light (relative to an inertial reference frame), the greater its relativistic mass (and relativistic inertia): It becomes harder and harder to accelerate the object.

There are two types of mass:

* 1= > / " 1= ?
1 ? 1 ? " 1= 1
@ * 1= > ? / " 1= ?
1 ? 1= ? 6 "

1)*+,)))-) % * +
. / !A 6 B ; " 1 6"
' !! ! ?

i.e. points so that all chords through them have constant length. The problem is studied extensively in note 6.3, chapter 6 of Gardner 1995 [548]. Do not miss the recent revolution in the theory of plane curves: see the end of §1.4.

1.3.2 Curvature

We now look at the outer geometry of plane curves. The inner geometry of curves does not differ from that of straight lines, but the geometry is radically different if we look at the way a curve sits in the plane. We are going to introduce a concept of *curvature* which measures how much a curve differs from a straight line. For a curve seen as a kinematic motion, curvature is directly linked with the *acceleration* vector $c''(t)$. For a geometric curve, we can introduce it by looking at the variation of the length of the curves drawn at a constant distance, called the *parallel curves* (recall the first variation equation 1.2 on page 4). If the absolute value of the infinitesimal change of length of these equidistant curves, close to a point $m = c(t)$, is the same as for a circle of radius r we say that the curve c has *radius of curvature* r at $m = c(t)$, and that its curvature is $K = 1/r$. The formula to compute curvature is

$$K(t) = \frac{1}{r(t)} = \frac{|\det(c'(t), c''(t))|}{\|c'(t)\|^3}$$

where t is any parameter, and boils down to

$$K(t) = \frac{1}{r(t)} = \|c''(t)\|$$

when t is a unit speed parameterization (i.e. arc length parameterization). Of course a circle of radius R has constant curvature equal to $1/R$. Another way to look at it is the following: the circle $C(t)$ which is defined as tangent to the curve at the point $c(t)$ and has radius $r(t)$ is the circle which has the most intimate contact with the curve (technically, the contact is third order, meaning that the curve and the circle have the same first three derivatives at that point). This circle is called the *osculating circle* to the curve at $m = c(t)$.

To rest a little, we mention an interesting fact rarely found in textbooks. When at a point m of the curve c, the curvature varies, that is to say $dK/dt(m) \neq 0$, then the osculating circle at m crosses the curve (guess where it is inside, and where outside). A baffling consequence is that the osculating circles of of a curve whose curvature is never critical never intersect one another. See the picture in figure 1.15. This picture gives two counterexamples. First, consider the unit tangent vectors to this family of circles. We get a continuous vector field, alias a differential equation in the plane. Now at every point of the curve there are at least *two* different integral curves of that vector field: namely the curve and the circle. The moral is that uniqueness of solutions of ordinary differential equations does not follow only from continuity. Second, consider the envelope of this family of circles. In many books it

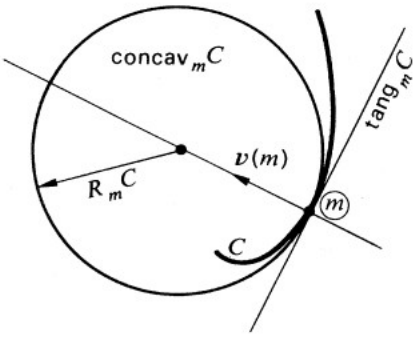

Fig. 1.15. Finding tangent line, concavity and curvature

is "proven" that the envelope is the limit of the so-called characteristic point, the point of intersection of close-by curves. This is wrong (unless we invoke "complex circles").

There is another geometric definition of the curvature. Parameterize the curve by arc length, so that $\|c'(t)\| = 1$ identically. Then $c'(t)$ is a unit vector, running through the unit circle of the plane. The curvature is the speed of its rotation. More precisely, one needs first to orient the plane, and choose an oriented direction in the plane. Then one can define an angle $\alpha(t)$ for $c'(t)$ which is a real number, defined only up to integer multiples of 2π. The curvature is $K(t) = d\alpha/dt$.

To get rid of the 2π ambiguity, there is an important fact which we are going to use below. For regular curves, when t runs through the interval of definition, one can follow $\alpha(t)$ by continuity and then define a map $t \mapsto \alpha(t)$ into $\mathbb{R}$. Note that if the curve is closed with period T then $\alpha(t+T)$ will in general differ from $\alpha(t)$ by an integral multiple of 2π.

We saw above that straight lines are the shortest paths between two points. Let us look at this again, now in a more sophisticated way. We start with a curve c with ends $p = c(a)$ and $q = c(b)$. Suppose it is as short as any curve can be with these extremities. Consider any one parameter family of curves c_μ with the same ends, with $c = c_0$, and compute their lengths. The fact that the length should be a minimum implies in particular that the first derivative of the length as a function of the parameter μ has to vanish. Computation yields

$$\left.\frac{d\,\mathrm{length}\,(c_\mu)}{d\mu}\right|_{\mu=0} = -\int_a^b \langle f(s), c''(s)\rangle \; ds \tag{1.3}$$

where the vector valued function $f(s)$ (which we can choose to be orthogonal to the curve everywhere) estimates the derivative of the normal displacement of the family of curves. This is also a first variation formula. Note that this formula is nothing but the integrated version of the definition we previously gave for curvature. A classical calculus trick (which is not obvious) proves that if we want all of these derivatives to vanish, namely the above integral

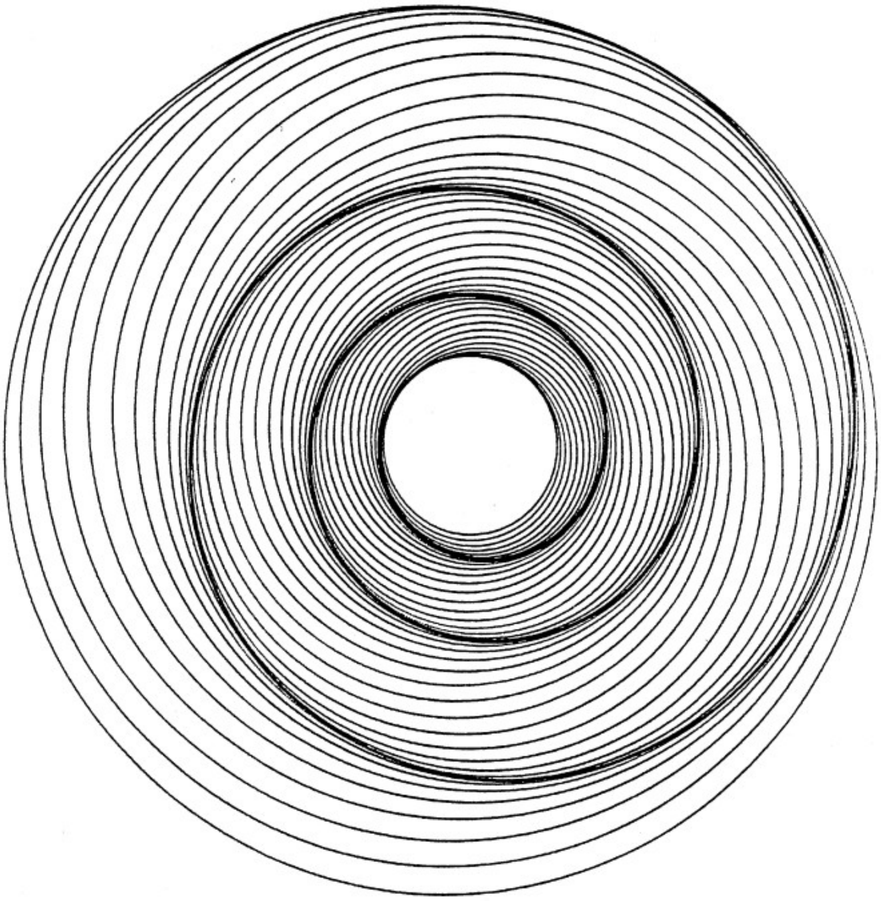

Fig. 1.16. Picture drawn for us by Etienne Ghys

to be 0 whatever the function f vanishing at $s = a$ and $s = b$, then $c''(t) = 0$ identically, and in particular our curve should have everywhere vanishing curvature. Curves with everywhere vanishing curvature are, of course, straight lines. Beware: what we have proven up to now by this method is that shortest paths can only be found among straight lines. Not every zero of the derivative of a function is a minimum. For a local minimum, a sufficient condition is to have positive second derivative. But we still have to search for an absolute minimum. The above proof is interesting because it is the simplest model of a field called *the calculus of variations*, and gives a general scheme to start searching for shortest paths. We will use it soon in §1.6 but principally in chapter 6. To conclude this story for the Euclidean plane, we know that if shortest paths exist, they should be straight lines. Then we check directly as above.

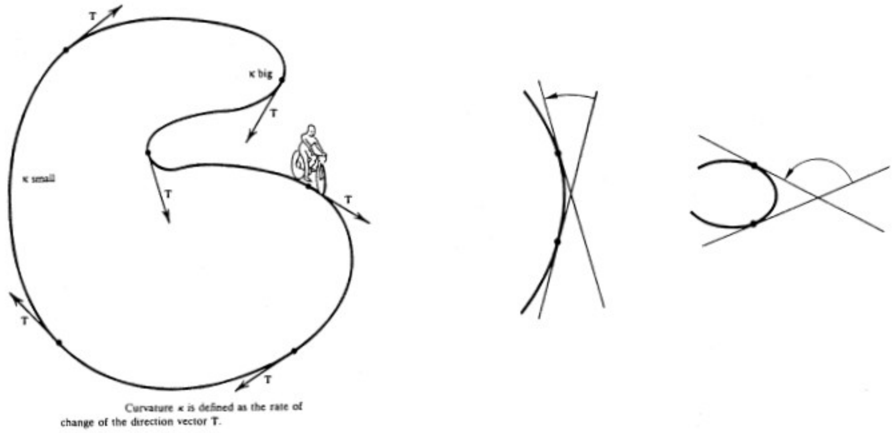
T
κ big
T
κ small
T
T
T
T
Curvature κ is defined as the rate of change of the direction vector T.

Fig. 1.17. Different definitions of curvature

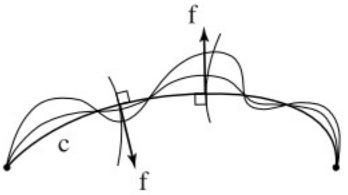

Fig. 1.18. Planar calculus of variations

There is a notion of *algebraic curvature* k as opposed to the scalar curvature K, linked by $K = |k|$. This is defined as soon as the curve is regular and oriented and the plane is also oriented, or (equivalently) if one side of the geometric curve is chosen, or (equivalently) if one of the two possible unit normal vectors is chosen. This side need not be the concavity side. Denote by $n(t)$ the unit normal vector you have chosen. Remember now that $c''(t)$ is an invariant. The algebraic curvature of c at t is the real number $k(t)$ so that

$$c''(t) = k(t)n(t)$$

A nice theorem about algebraic curvature:

Theorem 1 (More tightly curved curves) *If two curves c_1, c_2 are arc length parameterized, and have respective curvatures satisfying $k_1(t) \geq k_2(t)$ and start at the same initial point in the same direction, then the curve c_1 lies entirely inside the curve c_2.*

(because it is "more curved" as intuition tells you). For a proof see Chern 1989 [366] and exercise 7, section 5-7 of do Carmo 1976 [451]. Beware that you need simple curves if you want a global result. If not, you just get a local one. The

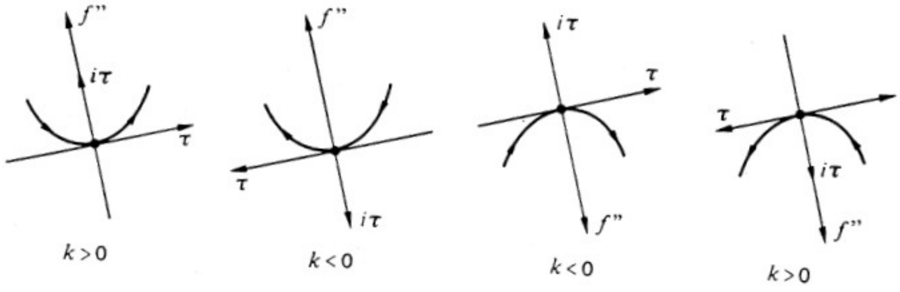

Fig. 1.19. Picturing algebraic curvatures

above theorem is one of the simplest ones in *control theory*. You will see many, many others later on. Note the following amusing consequences. Consider a simple closed plane curve, and the point where the curvature is maximum. Then the osculating circle at that point can keep rolling inside the curve all the way around. Similarly, the curve itself can roll inside its osculating circle of minimum curvature.

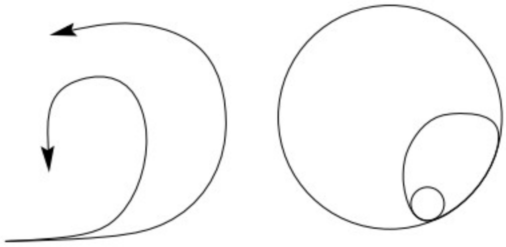

Fig. 1.20. (a) One is more curved than the other **(b)** The little circle can roll inside the closed curve, and the closed curve can roll inside the big circle

Let us make more precise our claim above that curvature is the only invariant of plane curves. More precisely the algebraic curvature is a characteristic invariant: if in the above inequality for curvatures we have equality ($k_1(t) = k_2(t)$ for every t) then the curves c_1 and c_2 are the same. Moral: if you know the value of the curvature as a function of the arc length then you know the curve completely. If the curves do not have the same initial point and direction, this just means that they are deduced one from the other by a Euclidean motion. The simplest example: a curve with non-zero constant curvature is part (or the whole) of a circle.

A spectacular application of this property of arcs of circle is the manufacture of a perfect straightedge. Take two pieces (of anything suitable for the job) and rub one against the other. The above theorem tells you that such a rubbing will eventually carve the two bounding curves into perfect circles. Now take three pieces, and rub them two by two. The boundaries will be circles, but the pieces will have to exhibit opposite convexity, i.e. equal algebraic curvature with opposite sign. For three, this is not possible unless the common curvature is zero. Naively, one might imagine that a laser can draw straight lines, but this is impractical. Today, lasers are used only to check the accuracy of the straightedge, after it is manufactured following the above recipe; once the laser has reported its evaluation of the straightedge, corrections are carried out by skilled hands using emery powder. Before the invention of the laser, accuracy was tested with interferences.

The same technique is still the only one available to make "perfect" planes: rub three surfaces against one another, and the result has to be a plane. To prove it, use the fact which we will quote on page 47 that pieces of a surface, all of whose points are umbilics, are pieces of spheres. To check accuracy, one doesn't use a laser, but drags a straightedge around on the surface, or a comparison plane, or applies an interferometer.[3] Rubbing only two pieces will yield pieces of a sphere, and is still the only method to sculpt spherical lenses.

In the same spirit, to fashion "perfect" balls, (which have thousands of practical applications, for example as ball bearings), one method was to put approximate balls inside a box, and stir them. Proof that the limit object is a sphere is not as elementary as proofs of the efficacy of the previous production methods, and was only recently established in Andrews 1999 [52]. Since 1907, balls for ball bearings have been produced by rolling rough balls between two plates with toroidal concentric groves and random indentations. See Berger 2003 [173] or technical books for more details.

For the story of curves in $\mathbb{E}^3$ see Berger and Gostiaux 1988 [175], do Carmo 1976 [451], or Stoker 1989 [1160]. We will just mention briefly what happens there. This time there are two characteristic invariants, curvature and torsion, but the curves need to be biregular, i.e. the acceleration should never vanish. Things also extend similarly to any $\mathbb{E}^d$, this time with $d-1$ invariants: see Spivak 1979 [1155]. But the first invariant is always the curvature and is defined simply as the norm $\|c''(t)\|$ for any arc-length parameterization. Only straight lines have everywhere zero curvature. For the kinematician this is the old fact that points with no force applied to them move along straight lines and at constant speed.

Take helices for example. They can be characterized as the curves with constant curvature and torsion. A practical application, in the spirit of the one just above: graduate a straightedge to make ruler; this is equivalent to making a protractor, dividing a circle in equal parts, etc. It is clear that you can do it if you can produce a screw (a helix). This is even simpler than a straightedge. Just start with an approximately good pair of a bolt and a nut and rub one against the other. At the end we will get a perfect bolt, and a perfect nut, because the common boundary will be made up of a union of curves with necessarily constant metric invariants, hence easily seen to be helices.

---

[3] An *interferometer* is a device which uses the interference of two waves (radio, acoustic, or light waves will do) to make very precise distance measurements.

1.4 Global Theory of Closed Plane Curves

1.4.1 "Obvious" Truths About Curves Which are Hard to Prove

Systematic references for what follows are Stoker 1989 [1160], do Carmo 1976 [451], Berger and Gostiaux 1988 [175], among others. See also the new events in Arnol′d 1994 [65] and other references in §§1.4.6.

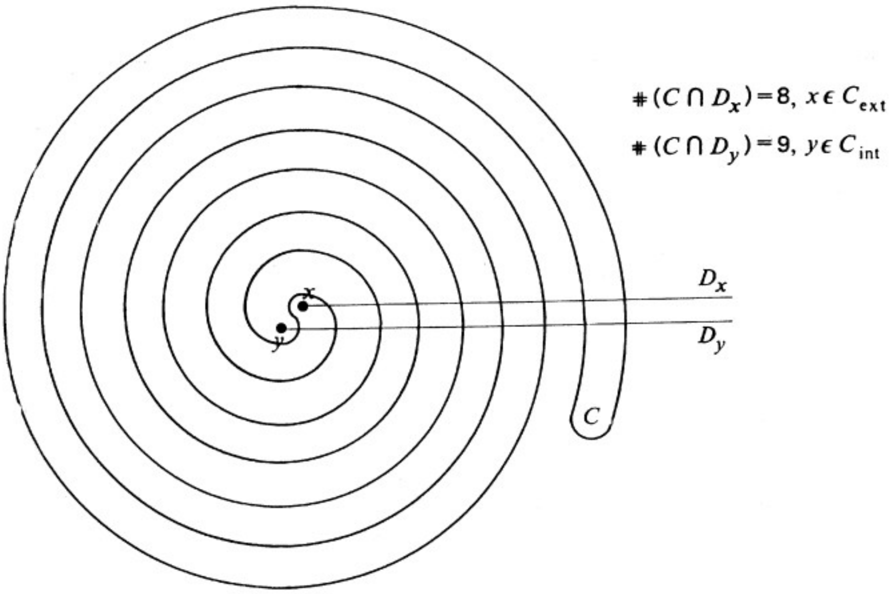
$\#(C \cap D_x) = 8$, $x \in C_{\text{ext}}$
$\#(C \cap D_y) = 9$, $y \in C_{\text{int}}$
D_x
D_y
x
y
C

Fig. 1.21. A trick to find when a point belongs to the interior or to the exterior of a curve

Theorem 2 (Jordan curve theorem) *A simple closed plane curve has an interior and an exterior.*

This implies the more common statement that a simple plane closed curve separates the plane in two regions. More precisely the plane with the curve

27.04.2020

and Knots (2002...)...pdf

Pag. 246 (260)

12.8 The case $\pi = \mathbb{Z}/4\mathbb{Z}$

Pag. 247 (261)

Theorem 12.6

Let $M = S^2 \times S^2/\langle \tau(I,-I)$.

Then the natural map from π to $Aut_{\pi}(\Pi)$ is an isomorphism, while $G_{\#}(M)$ has order ≤ 2.

G) The geometry of four-manifolds...pdf

Pag. 375 (389)

10. The Differential Topology of Algebraic Surfaces

154) ~~20~~ 20200428

A) Books_73_2020/
A Visual Introduction to the Fourth Dimension.../

A1) Geometry Proofs Essential Practice Problems. ... pdf

28.04.2020

A21 Essential Modern Physics + Study Guide...pdf

Pag. 35 Relativistic Mass

B) Books_73_2020/
Geometry II (Universiteit)/

A Panoramic View of Riemannian Geometry (2003...)...pdf

Pag. 12(33)

1.3.2 Curvature

Osculating circle
to the curve at m = c(t).

Pag. 13(34)

Fig. 1.15 Finding tangent line, concavity and curvature.

Pag. 15(36)

Fig. 1.18 Planar calculus of variations.

C) Books_73_2020/
Plasma Astrophysics _ Fundamentals and Practice _ 2 Volumes/

)*+,)))-) % %C %% D C 6"
' %D $ 4 D

Chapter 8
Coulomb Collisions in Astrophysical Plasma

Abstract Binary collisions of particles with the Coulomb potential of interaction are typical for physics of collisional plasmas in space and especially for gravitational systems. Coulomb collisions of fast particles with plasma particles determine momentum and energy losses of fast particles, the relaxation processes in astrophysical plasma.

8.1 Close and Distant Collisions

8.1.1 The Collision Parameters

Binary interactions of particles, described by the Coulomb potential

$$\varphi(r) = \frac{e}{r}\,, \tag{8.1}$$

have been studied in mechanics (see Landau and Lifshitz (1976), Chap. 4, Sect. 19). Considering binary interactions as *collisions*, we are interested only in their final result, the duration of the interaction and the actual form of particle trajectories being neglected. Thus in the centre-of-mass system, each particle is deflected through an angle χ defined by the relation

$$\tan\frac{\chi}{2} = \frac{e_1 e_2}{m v^2\, l} \tag{8.2}$$

or

$$l\,(\chi) = \frac{e_1 e_2}{m v^2}\,\cot\frac{\chi}{2}\,. \tag{8.3}$$

Here

$$m = \frac{m_1 m_2}{m_1 + m_2} \tag{8.4}$$

B.V. Somov, *Plasma Astrophysics, Part I: Fundamentals and Practice*,
Astrophysics and Space Science Library 391, DOI 10.1007/978-1-4614-4283-7_8,
© Springer Science+Business Media New York 2013

Fig. 8.1 The trajectory of a light particle with mass m_1 near a heavy particle with mass m_2

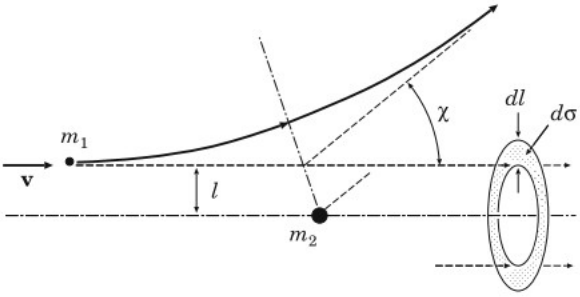
m_1
$\mathbf{v}$
l
m_2
χ
dl
$d\sigma$

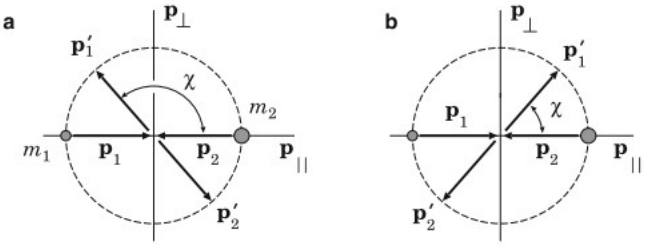
a
$\mathbf{p}_\perp$
$\mathbf{p}_1'$
χ
m_2
m_1
$\mathbf{p}_1$
$\mathbf{p}_2$
$\mathbf{p}_{||}$
$\mathbf{p}_2'$
b
$\mathbf{p}_\perp$
$\mathbf{p}_1'$
$\mathbf{p}_1$
χ
$\mathbf{p}_2$
$\mathbf{p}_{||}$
$\mathbf{p}_2'$

Fig. 8.2 Close (**a**) and distant (**b**) collisions of particles in the momentum space in the centre-of-mass system

is the reduced mass, v is the relative particle velocity at infinity, l is the "*impact parameter*". The last is the closest distance of the particle's approach, were it not for their interaction as shown in Fig. 8.1.

For particles deflected through a right angle

$$l\left(\frac{\pi}{2}\right) \equiv l_\perp = \frac{e_1 e_2}{m v^2}\,, \tag{8.5}$$

so the initial formula (8.2) is conveniently rewritten as

$$\boxed{\tan\frac{\chi}{2} = \frac{l_\perp}{l}\,.} \tag{8.6}$$

The collisions are called *close* if

$$\pi/2 \le \chi \le \pi\,, \quad \text{i.e.} \quad 0 \le l \le l_\perp\,. \tag{8.7}$$

Correspondingly, for *distant* collisions $l > l_\perp$ and $0 \le \chi < \pi/2$. Both cases are shown in Fig. 8.2.

)*+,)))-) % %C %% (%E 6"
' $ (4 %

On the other hand, in the *Bastille day flare* on 2000 July 14 (see Chaps. 6 and 7) as well as in some other large solar flares, it was possible to detect the real changes in the sunspot structure just after a flare.

> The outer penumber fields became more vertical due to magnetic reconnection in the corona during a flare

(Liu et al. 2005; Wang et al. 2005). One can easily imagine such changes by considering, for example, Fig. 4.1 between sunspots N and s.

Sudol and Harvey (2005) have used the *Global Oscillation Network Group* (GONG) magnetograms to characterize the changes in the photospheric vertical component of magnetic field during 15 large solar flares. An abrupt, significant, and persistent change in the magnetic field occurred in at least one location within the flaring active region during each event after its start. Among several possible interpretations for these observations, Sudoh and Harvey favor one in which

> the magnetic field changes result from the penumber field relaxing upward by reconnecting magnetic field above the photosphere.

This interpretation is very similar to than one given by Liu et al. (2005) and Wang et al. (2005).

The topological model (Somov et al. 2008) self-consistently explains the following observational effects in the photosphere during and after the large flare on 2000 July 14 (see Chap. 6): (a) the enhancement of the horizontal and vertical magnetic fields at the center of an active region (AR) and the weakening of the horizontal field on the AR periphery; (b) the decrease in the AR magnetic flux as a whole; and (c) the convergence of the center-of-mass positions of opposite magnetic polarities. Therefore the entire set of these observations can be interpreted as the result of a magnetic-field restructuring in the corona caused by magnetic reconnection.

As for the second objection (2) to the hypothesis of accumulation of energy in the form of magnetic field of *slowly-reconnecting* current layers in the solar atmosphere, the rapid dissipation of the field necessary for the flare is naturally explained by the theory of current layers presented in what follows, especially in Chaps. 8 and 9).

4.2 The Simple Topological Model of an Active Region

4.2.1 The Potential Magnetic Field Approximation

Gorbachev and Somov (1989, 1990) have developed a three-dimensional (3D) model for a *potential* magnetic field in the active region AR 2776 with an extended solar flare on 1980 November 5. Before discussing the active region and the flare (see Sect. 4.3.1), let us consider, at first, the general properties of this class of models called *topological*.

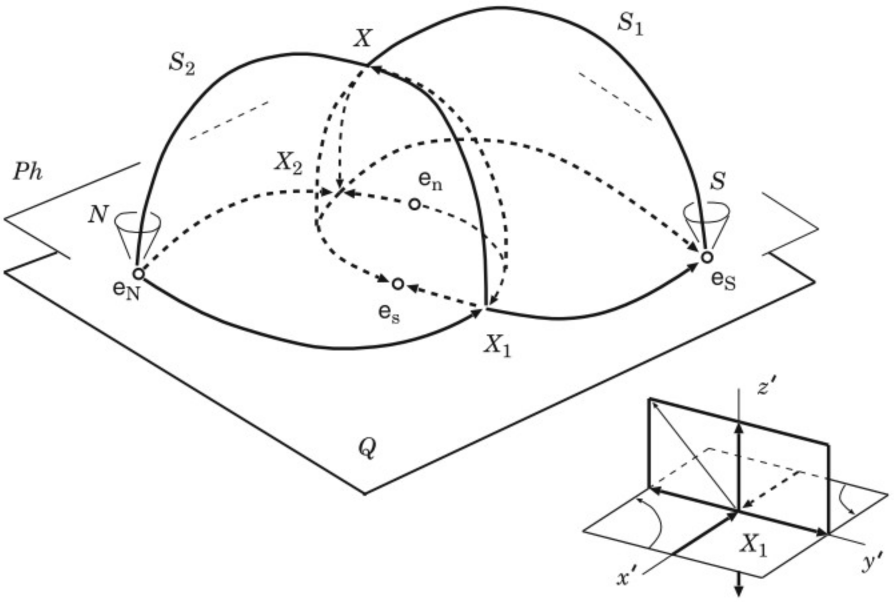

Fig. 4.2 The model for the magnetic field of four sunspots of pairwise opposite polarity. The sunspots N and S in the photospheric plane Ph. The separatrices S_1 and S_2 cross at the separator X_1XX_2 above the plane Q of the effective magnetic 'charges' e_{N}, e_{S}, etc

In the simplest variant of the model (Gorbachev and Somov1988), four magnetic field sources – the magnetic 'charges' e_{N} and e_{S}, e_{n} and e_{s}, located in the plane Q under the photosphere Ph (Fig. 4.2) – are used to reproduce the main features of the observed field in the photosphere related to the four most important sunspots: N, S, n and s. As a consequence, the *quadrupole* model reproduces only the large-scale features of the actual field in the corona related to these sunspots. As a minimum, the four sources are necessarily to describe two interacting magnetic fluxes having the two sources per each. The larger number of sources are not necessarily much better.

The main features are two magnetic surfaces, the boundary surfaces called the *separatrices* S_1 and S_2 (Fig. 4.2). They divide the whole space above the under-photospheric plane Q into four regions and, correspondingly, the whole magnetic field into four magnetic fluxes having different linkages. The field lines are grouped into four regions according to their termini. The separatrices of the potential magnetic field are formed from field lines beginning or ending at magnetic zeroth points X_1 and X_2. For example, the field lines originating at the point X_1 form a separatrix surface S_1 (for more detail see Gorbachev et al. 1988).

There is a topologically singular field line X_1XX_2 lying at the intersection of the two separatrices, it belongs to all four fluxes (two reconnecting and two reconnected fluxes) that interact at this line – the 3D magnetic *separator*. So

> the separator separates the interacting magnetic fluxes by the separatrices above the plane of magnetic sources

(cf. Fig. 1.4a; see also Sweet 1969; Lau 1993). Detection of a separator, as a field line that connects two zeroth points in the Earth magnetotail, by the four *Cluster* spacecraft (Xiao et al. 2007) provides an important step towards establishing an observational framework of 3D reconnection.

4.2.2 Classification of Zeroth Points

Let us clarify the basic properties of the zeroth points of the magnetic field discussed above. In the vicinity of a zeroth point X_i located at $\mathbf{r} = \mathbf{r}_i$ in the plane Q, the vector of magnetic field can be represented in the form

$$B_\alpha(\mathbf{r}) = \frac{\partial \psi}{\partial r_\alpha} = M_{\alpha\beta}(\mathbf{r}_i)\,\xi_\beta\,. \tag{4.2}$$

Here $\psi(\mathbf{r})$ is the potential of the field, the vector $\xi = \mathbf{r} - \mathbf{r}_i$, the index $i = 1, 2$; the Greek indices $\alpha, \beta = 1, 2, 3$ or x, y, z. The symmetric matrix

$$M_{\alpha\beta}(\mathbf{r}_i) = \frac{\partial^2 \psi\,(\mathbf{r}_i)}{\partial r_\alpha\, \partial r_\beta} = \begin{vmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{vmatrix} \tag{4.3}$$

in the frame of coordinates r'_α related to the eigenvectors $\mathbf{e}'_\alpha$, λ_α are the eigenvalues of the matrix. Because the magnetic field is potential, all the λ_α are real numbers (e.g., Dubrovin et al. 1986). Positive eigenvalues λ_α correspond to the magnetic-field lines emerging from a zeroth point and negative λ_α correspond to the field lines arriving at a point $\mathbf{r}_i$.

If the determinant

$$\det M_{\alpha\beta}(\mathbf{r}_i) = \lambda_1 \lambda_2 \lambda_3 \neq 0\,, \tag{4.4}$$

then the zeroth point is called *non-degenerate*. Such a point is *isolated*, i.e., the magnetic field in its vicinity does not vanish.

Since div $\mathbf{B} = 0$, we obtain a condition

$$\lambda_1 + \lambda_2 + \lambda_3 = 0\,. \tag{4.5}$$

Therefore there exist only three following classes of zeroth points of magnetic field (see Gorbachev et al. 1988).

Type A. All eigenvalues $\lambda_\alpha \neq 0$. This corresponds to two zeroth points shown in Fig. 4.2. At the first point X_1

$$\lambda_1 < 0\,, \quad \lambda_2 > 0\,, \quad \lambda_3 > 0\,. \tag{4.6}$$

28 04 2020

C1) 1461442826.pdf (Volume 1)
Pág. 148(170) ✡
Fig. 8.1 The trajectory of a light particle with mass m_1 near a heavy particle with mass m_2.

C2) 146144294X.pdf (Volume 2)
Pág. 72(92)
Fig. 4.2 The model for the magnetic field of four sunspots of pairwise opposite polarity.

D) Books_73_2020/
The Geometry of Hamilton and Lagrange Spaces/

The Geometry of Hamilton and Lagrange Spaces...pdf

E) Books_73_2020/
Pleasures of Pi, e and Other Interesting Numbers/
Pleasures of Pi, e ....pdf ✡

F) Books_73_2020/
The Bloch-Kato Conjecture for the Riemann Zeta Function/
The Bloch-Kato Conjecture...pdf

")*+,)))-) % 8 . / ; "
9 6 A 6 7 " 6"

)*+,)))-) % A " - &
7 1 6"

)*+,)))-) % 8 B F =
+ G 4 6"

)*+,)))-) % / " -? " "
/ : 0H 8 & * " *66
$ 6"
' !!(!%C 6 = ? " /

METRIC CONNECTIONS IN PROJECTIVE DIFFERENTIAL GEOMETRY*

MICHAEL EASTWOOD† AND VLADIMIR MATVEEV‡

(In memory of Thomas Branson)

Abstract. We search for Riemannian metrics whose Levi-Civita connection belongs to a given projective class. Following Sinjukov and Mikeš, we show that such metrics correspond precisely to suitably positive solutions of a certain projectively invariant finite-type linear system of partial differential equations. Prolonging this system, we may reformulate these equations as defining covariant constant sections of a certain vector bundle with connection. This vector bundle and its connection are derived from the Cartan connection of the underlying projective structure.

Key words. Projective differential geometry, metric connection, tractor.

AMS(MOS) subject classifications. Primary 53A20; Secondary 58J70.

1. Introduction. We shall always work on a smooth oriented manifold M of dimension n. Suppose that ∇ is a torsion-free connection on the tangent bundle of M. We may ask whether there is a Riemannian metric on M whose geodesics coincide with the geodesics of ∇ as unparameterised curves. We shall show that there is a linear system of partial differential equations that precisely controls this question.

To state our results, we shall need some terminology, notation, and preliminary observations. Two torsion-free connections ∇ and $\hat{\nabla}$ are said to be projectively equivalent if they have the same geodesics as unparameterised curves. A projective structure on M is a projective equivalence class of connections. In these terms, we are given a projective structure on M and we ask whether it may be represented by a metric connection. Questions such as this have been addressed by many authors. Starting with a metric connection, Sinjukov [9] considered the existence of other metrics with the same geodesics. He found a system of equations that controls this question and Mikeš [7] observed that essentially the same system pertains when starting with an arbitrary projective structure.

We shall use Penrose's abstract index notation [8] in which indices act as markers to specify the type of a tensor. Thus, ω_a denotes a 1-form whilst X^a denotes a vector field. Repeated indices denote the canonical

---

*This work was undertaken during the 2006 Summer Program at the Institute for Mathematics and its Applications at the University of Minnesota. The authors would like to thank the IMA for hospitality during this time. The first author is supported by the Australian Research Council.

†Department of Mathematics, University of Adelaide, SA 5005, Australia (`meastwoo@member.ams.org`).

‡Mathematisches Institut, Fakultät für Mathematik und Informatik, Friedrich-Schiller-Universität Jena, 07737 Jena, Germany (`matveev@minet.uni-jena.de`).

pairing between vectors and co-vectors. Thus, we shall write $X^a\omega_a$ instead of $X \lrcorner\, \omega$. The tautological 1-form with values in the tangent bundle is denoted by the Kronecker delta $\delta_a{}^b$.

As is well-known [5], the geometric formulation of projective equivalence may be re-expressed as

$$\hat{\nabla}_a X^b = \nabla_a X^b + \Upsilon_a X^b + \delta_a{}^b \Upsilon_c X^c \tag{1.1}$$

for an arbitrary 1-form Υ_a. We shall also adopt the curvature conventions of [5]. In particular, it is convenient to write

$$(\nabla_b\nabla_a - \nabla_a\nabla_b)X^b = R_{ab}X^b,$$

where R_{ab} is the usual Ricci tensor, as

$$(\nabla_b\nabla_a - \nabla_a\nabla_b)X^b = (n-1)\mathrm{P}_{ab}X^b - \beta_{ab}X^b \quad \text{where } \beta_{ab} = \mathrm{P}_{ba} - \mathrm{P}_{ab}.$$

If a different connection is chosen in the projective class according to (1.1), then

$$\hat{\beta}_{ab} = \beta_{ab} + \nabla_a\Upsilon_b - \nabla_b\Upsilon_a.$$

Therefore, as a 2-form β_{ab} changes by an exact form. On the other hand, the Bianchi identity implies that β_{ab} is closed. Thus, there is a well-defined de Rham cohomology class $[\beta] \in H^2(M, \mathbb{R})$ associated to any projective structure.

PROPOSITION 1.1. *The class $[\beta] \in H^2(M, \mathbb{R})$ is an obstruction to the existence of a metric connection in the given projective class.*

Proof. The Ricci tensor is symmetric for a metric connection. □

In searching for a metric connection in a given projective class, we may as well suppose that the obstruction $[\beta]$ vanishes. For the remainder of this article we suppose that this is the case and we shall consider only representative connections with symmetric Ricci tensor. In other words, all connections from now on enjoy

$$(\nabla_b\nabla_a - \nabla_a\nabla_b)X^b = (n-1)\mathrm{P}_{ab}X^b \quad \text{where } \mathrm{P}_{ab} = \mathrm{P}_{ba}. \tag{1.2}$$

A convenient alternative characterisation of such connections as follows.

PROPOSITION 1.2. *A torsion-free affine connection has symmetric Ricci tensor if and only if it induces the flat connection on the bundle of n-forms.*

Proof. If $\epsilon^{pqr\cdots s}$ has n indices and is totally skew then

$$(\nabla_a\nabla_b - \nabla_b\nabla_a)\epsilon^{pqr\cdots s} = \kappa_{ab}\epsilon^{pqr\cdots s}$$

for some 2-form κ_{ab}. But, by the Bianchi symmetry,

$$(\nabla_a\nabla_b - \nabla_b\nabla_a)\epsilon^{abr\cdots s} = -2R_{ab}\epsilon^{abr\cdots s},$$

which vanishes if and only if R_{ab} is symmetric. □

Having restricted our attention to affine connections that are flat on the bundle of n-forms, we may as well further restrict to connections ∇_a for which there is a volume form $\epsilon_{bc\cdots d}$ (necessarily unique up to scale) with $\nabla_a\epsilon_{bc\cdots d} = 0$. We shall refer to such connections as special. The freedom in special connections within a given projective class is given by (1.1) where $\Upsilon_a = \nabla_a f$ for an arbitrary smooth function f. Following [5], the full curvature of a special connection may be conveniently decomposed:

$$(\nabla_a\nabla_b - \nabla_b\nabla_a)X^c = W_{ab}{}^c{}_dX^d + \delta_a{}^c\mathrm{P}_{bd}X^d - \delta_b{}^c\mathrm{P}_{ad}X^d, \qquad (1.3)$$

where $W_{ab}{}^c{}_d$ is totally trace-free and P_{ab} is symmetric. The tensor $W_{ab}{}^c{}_d$ is known as the Weyl curvature and is projectively invariant.

2. A linear system of equations. In this section we present, as Proposition 2.1, an alternative characterisation of the Levi-Civita connection. The advantage of this characterisation is that it leads, almost immediately, to a system of linear equations that controls the metric connections within a given projective class. The precise results are Theorems 2.1 and 2.2.

PROPOSITION 2.1. *Suppose g^{ab} is a metric on M with volume form $\epsilon_{bc\cdots d}$. Then a torsion-free connection ∇_a is the metric connection for g^{ab} if and only if*

- $\nabla_a g^{bc} = \delta_a{}^b\mu^c + \delta_a{}^c\mu^b$ *for some vector field μ^a*
- $\nabla_a\epsilon_{bc\cdots d} = 0$.

Proof. Write D_a for the metric connection of g^{ab}. Then

$$\nabla_a\omega_b = D_a\omega_b - \Gamma_{ab}{}^c\omega_c \qquad (2.1)$$

for some tensor $\Gamma_{ab}{}^c = \Gamma_{ba}{}^c$. We compute

$$\epsilon^{bc\cdots d}\nabla_a\epsilon_{bc\cdots d} = -n\epsilon^{bc\cdots d}\Gamma_{ab}{}^e\epsilon_{ec\cdots d} = -n!\,\Gamma_{ab}{}^b$$

and so $\Gamma_{ab}{}^b = 0$. Similarly,

$$\nabla_a g^{bc} = \Gamma_{ad}{}^b g^{dc} + \Gamma_{ad}{}^c g^{bd}$$

and so

$$\Gamma_{ad}{}^b g^{dc} + \Gamma_{ad}{}^c g^{bd} = \delta_a{}^b\mu^c + \delta_a{}^c\mu^b. \qquad (2.2)$$

Let g_{ab} denote the inverse of g^{ab} and contract (2.2) with g_{bc} to conclude that

$$2\Gamma_{ab}{}^b = 2\mu_a \quad \text{where } \mu_a = g_{ab}\mu^b$$

and hence that $\mu^a = 0$. If we let $\Gamma_{abc} = \Gamma_{ab}{}^d g_{cd}$, then (2.2) now reads

$$\Gamma_{acb} + \Gamma_{abc} = 0.$$

Together with $\Gamma_{abc} = \Gamma_{bac}$, this implies that $\Gamma_{abc} = 0$. From (2.1) we see that $\nabla_a = D_a$, which is what we wanted to show. □

THEOREM 2.1. *Suppose ∇_a is a special torsion-free connection and there is a metric tensor σ^{ab} such that*

$$\nabla_a\sigma^{bc} = \delta_a{}^b\mu^c + \delta_a{}^c\mu^b \quad \text{for some vector field } \mu^a. \tag{2.3}$$

Then ∇_a is projectively equivalent to a metric connection.

Proof. Consider the projectively equivalent connection

$$\hat\nabla_a X^b = \nabla_a X^b + \Upsilon_a X^b + \delta_a{}^b\Upsilon_c X^c \quad \text{where } \Upsilon_a = \nabla_a f$$

for some function f. If we let $\hat\sigma^{ab} \equiv e^{-2f}\sigma^{ab}$, then

$$\begin{aligned}\hat\nabla_a\hat\sigma^{bc} &= e^{-2f}\left(-2\Upsilon_a\sigma^{bc} + \nabla_a\sigma^{bc} + 2\Upsilon_a\sigma^{bc} + \delta_a{}^b\Upsilon_d\sigma^{dc} + \delta_a{}^c\Upsilon_d\sigma^{bd}\right)\\ &= e^{-2f}\left(\delta_a{}^b\mu^c + \delta_a{}^c\mu^b + \delta_a{}^b\Upsilon_d\sigma^{dc} + \delta_a{}^c\Upsilon_d\sigma^{bd}\right)\end{aligned}$$

and so

$$\hat\nabla_a\hat\sigma^{bc} = \delta_a{}^b\hat\mu^c + \delta_a{}^c\hat\mu^b \quad \text{where } \hat\mu^a = e^{-2f}\left(\mu^a + \Upsilon_b\sigma^{ab}\right). \tag{2.4}$$

Similarly, if we choose a volume form $\epsilon_{bc\cdots d}$ killed by ∇_a and let $\hat\epsilon_{bc\cdots d} \equiv e^{(n+1)f}\epsilon_{bc\cdots d}$, then

$$\hat\nabla_a\hat\epsilon_{bc\cdots d} = e^{(n+1)f}\left(\nabla_a\epsilon_{bc\cdots d} + \Upsilon_{[a}\epsilon_{bc\cdots d]}\right) = e^{(n+1)f}\nabla_a\epsilon_{bc\cdots d} = 0. \tag{2.5}$$

Define

$$\det(\sigma) \equiv \epsilon_{a\cdots b}\epsilon_{c\cdots d}\sigma^{ac}\cdots\sigma^{bd}$$

and compute

$$\begin{aligned}\widehat{\det}(\hat\sigma) &= \hat\epsilon_{a\cdots b}\hat\epsilon_{c\cdots d}\hat\sigma^{ac}\cdots\hat\sigma^{bd}\\ &= e^{2(n+1)f}e^{-2nf}\epsilon_{a\cdots b}\epsilon_{c\cdots d}\sigma^{ac}\cdots\sigma^{bd} = e^{2f}\det(\sigma).\end{aligned}$$

Therefore, if we take

$$f = -\tfrac{1}{2}\log\det(\sigma),$$

then we have arranged that $\widehat{\det}(\hat\sigma) = 1$. This is precisely the condition that $\hat\epsilon_{bc\cdots d}$ be the volume form for the metric $\hat\sigma^{ab}$. With (2.4) and (2.5) we are now in a position to use Proposition 2.1 to conclude that $\hat\nabla_a$ is the metric connection for $\hat\sigma^{ab}$. We have shown that our original connection ∇_a is projectively equivalent to the Levi-Civita connection for the metric $g^{ab} \equiv \det(\sigma)\,\sigma^{ab}$. □

Evidently, the equations (2.3) precisely control the metric connections within a given special projective class. Precisely, if g_{ab} is a Riemannian metric with associated Levi-Civita connection ∇_a, then

$$\hat\nabla_a\hat g^{bc} = \delta_a{}^b\hat\mu^c + \delta_a{}^c\hat\mu^b,$$

where $\hat{\nabla}_a$ is projectively equivalent to ∇_a according to (1.1) with $\Upsilon_a = \nabla_a f$ and where $\hat{g}^{bc} = e^{-2f} g^{bc}$. In other words, we have shown (cf. [7, 9]):–

THEOREM 2.2. *There is a one-to-one correspondence between solutions of* (2.3) *for positive definite* σ^{bc} *and metric connections that are projectively equivalent to* ∇_a.

3. Prolongation. Let us consider the system of equations (2.3) in more detail. It is a linear system for any symmetric contravariant 2-tensor σ^{bc}. Specifically, we may write (2.3) as

$$\text{the trace-free part of } (\nabla_a \sigma^{bc}) = 0 \tag{3.1}$$

or, more explicitly, as

$$\nabla_a \sigma^{bc} - \tfrac{1}{n+1} \delta_a{}^b \nabla_d \sigma^{cd} - \tfrac{1}{n+1} \delta_a{}^c \nabla_d \sigma^{bd} = 0.$$

According to [2], this equation is of finite-type and may be prolonged to a closed system as follows. According to (1.2) and (1.3) we have

$$(\nabla_a \nabla_b - \nabla_b \nabla_a) \sigma^{bc} = W_{ab}{}^c{}_d \sigma^{bd} + \delta_a{}^c \mathrm{P}_{bd} \sigma^{bd} - n \mathrm{P}_{ad} \sigma^{cd}.$$

On the other hand, from (2.3) we have

$$(\nabla_a \nabla_b - \nabla_b \nabla_a) \sigma^{bc} = (n+1) \nabla_a \mu^c - \nabla_b (\delta_a{}^b \mu^c + \delta_a{}^c \mu^b) = n \nabla_a \mu^c - \delta_a{}^c \nabla_b \mu^b.$$

We conclude that

$$n \nabla_a \mu^c = \delta_a{}^c \left(\nabla_b \mu^b + \mathrm{P}_{bd} \sigma^{bd} \right) - n \mathrm{P}_{ad} \sigma^{cd} + W_{ab}{}^c{}_d \sigma^{bd}$$

or, equivalently, that

$$\nabla_a \mu^c = \delta_a{}^c \rho - \mathrm{P}_{ad} \sigma^{cd} + \tfrac{1}{n} W_{ab}{}^c{}_d \sigma^{bd}, \tag{3.2}$$

for some function ρ. To complete the prolongation, we use (1.2) to write

$$(\nabla_c \nabla_a - \nabla_a \nabla_c) \mu^c = (n-1) \mathrm{P}_{ac} \mu^c$$

whereas from (3.2) we also have

$$\begin{aligned}(\nabla_c \nabla_a - \nabla_a \nabla_c) \mu^c = {} & \nabla_c \left(\delta_a{}^c \rho - \mathrm{P}_{ad} \sigma^{cd} + \tfrac{1}{n} W_{ab}{}^c{}_d \sigma^{bd} \right) \\ & - \nabla_a \left(n\rho - \mathrm{P}_{cd} \sigma^{cd} \right).\end{aligned}$$

Therefore,

$$\begin{aligned}(n-1) \mathrm{P}_{ac} \mu^c = {} & \nabla_c \left(-\mathrm{P}_{ad} \sigma^{cd} + \tfrac{1}{n} W_{ab}{}^c{}_d \sigma^{bd} \right) \\ & -(n-1) \nabla_a \rho + \nabla_a (\mathrm{P}_{cd} \sigma^{cd}).\end{aligned} \tag{3.3}$$

The terms involving Weyl curvature

$$\nabla_c(W_{ab}{}^c{}_d\sigma^{bd}) = (\nabla_c W_{ab}{}^c{}_d)\sigma^{bd} + W_{ab}{}^c{}_d\nabla_c\sigma^{bd}$$

may be dealt with by (2.3) and a Bianchi identity

$$\nabla_c W_{ab}{}^c{}_d = (n-2)(\nabla_a \mathrm{P}_{bd} - \nabla_b \mathrm{P}_{ad}).$$

We see that

$$\nabla_c(W_{ab}{}^c{}_d\sigma^{bd}) = (n-2)(\nabla_a \mathrm{P}_{bd} - \nabla_b \mathrm{P}_{ad})\sigma^{bd}$$

and (3.3) becomes

$$\begin{aligned}(n-1)\mathrm{P}_{ac}\mu^c = {} & \tfrac{n-2}{n}(\nabla_a \mathrm{P}_{bd} - \nabla_b \mathrm{P}_{ad})\sigma^{bd} - \nabla_c(\mathrm{P}_{ad}\sigma^{cd}) \\ & -(n-1)\nabla_a\rho + \nabla_a(\mathrm{P}_{cd}\sigma^{cd})\end{aligned}$$

or, equivalently,

$$\begin{aligned}\mathrm{P}_{ac}\mu^c = {} & \tfrac{2}{n}(\nabla_a \mathrm{P}_{bd} - \nabla_b \mathrm{P}_{ad})\sigma^{bd} - \tfrac{1}{n-1}\mathrm{P}_{ad}\nabla_c\sigma^{cd} \\ & -\nabla_a\rho + \tfrac{1}{n-1}\mathrm{P}_{cd}\nabla_a\sigma^{cd}.\end{aligned} \tag{3.4}$$

Again, we substitute from (2.3) to rewrite

$$\mathrm{P}_{cd}\nabla_a\sigma^{cd} - \mathrm{P}_{ad}\nabla_c\sigma^{cd} = \mathrm{P}_{cd}(\delta_a{}^c\mu^d + \delta_a{}^d\mu^c) - (n+1)\mathrm{P}_{ad}\mu^d = -(n-1)\mathrm{P}_{ad}\mu^d$$

and (3.4) becomes

$$\mathrm{P}_{ac}\mu^c = \tfrac{2}{n}(\nabla_a \mathrm{P}_{bd} - \nabla_b \mathrm{P}_{ad})\sigma^{bd} - \mathrm{P}_{ad}\mu^d - \nabla_a\rho,$$

which we may rearrange as

$$\nabla_a\rho = -2\mathrm{P}_{ab}\mu^b + \tfrac{2}{n}(\nabla_a \mathrm{P}_{bd} - \nabla_b \mathrm{P}_{ad})\sigma^{bd}.$$

Together with (2.3) and (3.2), we have a closed system, essentially as in [7, 9]:

$$\begin{aligned}\nabla_a\sigma^{bc} &= \delta_a{}^b\mu^c + \delta_a{}^c\mu^b \\ \nabla_a\mu^b &= \delta_a{}^b\rho - \mathrm{P}_{ac}\sigma^{bc} + \tfrac{1}{n}W_{ac}{}^b{}_d\sigma^{cd} \\ \nabla_a\rho &= -2\mathrm{P}_{ab}\mu^b + \tfrac{4}{n}Y_{abc}\sigma^{bc}\end{aligned} \tag{3.5}$$

where $Y_{abc} = \frac{1}{2}(\nabla_a \mathrm{P}_{bc} - \nabla_b \mathrm{P}_{ac})$, the Cotton-York tensor. The three tensors σ^{bc}, μ^b, and ρ may be regarded together as a section of the vector bundle

$$\mathcal{T} = \bigodot^2 TM \oplus TM \oplus \mathbb{R}$$

where $\bigodot$ denotes symmetric tensor product and $\mathbb{R}$ denotes the trivial bundle. We have proved:

THEOREM 3.1. *If we endow $\mathcal{T}$ with the connection*

$$\begin{pmatrix} \sigma^{bc} \\ \mu^b \\ \rho \end{pmatrix} \longmapsto \begin{pmatrix} \nabla_a\sigma^{bc} - \delta_a{}^b\mu^c - \delta_a{}^c\mu^b \\ \nabla_a\mu^b - \delta_a{}^b\rho + \mathrm{P}_{ac}\sigma^{bc} - \frac{1}{n}W_{ac}{}^b{}_d\sigma^{cd} \\ \nabla_a\rho + 2\mathrm{P}_{ab}\mu^b - \frac{4}{n}Y_{abc}\sigma^{bc} \end{pmatrix} \tag{3.6}$$

then there is a one-to-one correspondence between covariant constant sections of $\mathcal{T}$ and solutions σ^{bc} of (2.3).

4. Projective invariance. The equation (2.3) is projectively invariant in the following sense. Following [5], let $\mathcal{E}^{(ab)}(w)$ denote the bundle of symmetric contravariant 2-tensors of projective weight w. Thus, in the presence of a volume form $\epsilon_{bc\cdots d}$, a section $\sigma^{ab} \in \Gamma(M, \mathcal{E}^{(ab)}(w))$ is an ordinary symmetric contravariant 2-tensor but if we change volume form

$$\epsilon_{bc\cdots d} \mapsto \hat{\epsilon}_{bc\cdots d} = e^{(n+1)f}\epsilon_{bc\cdots d} \quad \text{for any smooth function } f,$$

then we are obliged to rescale σ^{ab} according to $\hat{\sigma}^{ab} = e^{wf}\sigma^{ab}$. Equivalently, we are saying that $\mathcal{E}^{(ab)}(w) = \bigodot^2 TM \otimes (\Lambda^n)^{-w/(n+1)}$, where Λ^n is the line-bundle of n-forms on M. The projectively weighted irreducible tensor bundles are fundamental objects on a manifold with projective structure.

PROPOSITION 4.1. *The differential operator*

$$\mathcal{E}^{(ab)}(-2) \longrightarrow \text{the trace-free part of } \mathcal{E}_a{}^{(bc)}(-2) \tag{4.1}$$

defined by (3.1) *is projectively invariant.*

Proof. This is already implicit in the proof of Theorem 2.1. Explicitly, however, we just compute from (1.1):

$$\hat{\nabla}_a\hat{\sigma}^{bc} = \nabla_a\hat{\sigma}^{bc} + 2\Upsilon_a\hat{\sigma}^{bc} + \delta_a{}^b\Upsilon_d\hat{\sigma}^{dc} + \delta_a{}^c\Upsilon_d\hat{\sigma}^{bd},$$

where $\Upsilon_a = \nabla_a f$ whilst

$$\nabla_a\hat{\sigma}^{bc} = \nabla_a(e^{-2f}\sigma^{bc}) = e^{-2f}\left(\nabla_a\sigma^{bc} - 2\Upsilon_a\sigma^{bc}\right) = \widehat{\nabla_a\sigma^{bc}} - 2\Upsilon_a\hat{\sigma}^{bc}.$$

It follows that

$$\hat{\nabla}_a\hat{\sigma}^{bc} = \widehat{\nabla_a\sigma^{bc}} + \text{trace terms},$$

which is what we wanted to show. □

In hindsight, it is not too difficult to believe that (3.1) should control the metric connections within a given projective class. There are very few projectively invariant operators. In fact, there are precisely two finite-type first order invariant linear operators on symmetric 2-tensors. One of them is (4.1) and the other is

$$\mathcal{E}_{(ab)}(4) \to \mathcal{E}_{(abc)}(4) \quad \text{given by} \quad \sigma_{ab} \mapsto \nabla_{(a}\sigma_{bc)}. \tag{4.2}$$

In two dimensions, (4.2) and (4.1) coincide. In higher dimensions, however, being in the kernel of (4.2) for positive definite σ_{ab} corresponds to having a metric g_{ab} and a totally trace-free tensor Γ_{abc} with

$$\Gamma_{abc} = \Gamma_{bac} \quad \text{and} \quad \Gamma_{abc} + \Gamma_{bca} + \Gamma_{cab} = 0$$

such that the connection

$$\omega_b \longmapsto D_a\omega_b - \Gamma_{ab}{}^c\omega_c$$

belongs to the projective class of ∇_a, where D_a is the Levi-Civita connection of g_{ab}. The available tensors Γ_{abc} for a given metric have dimension $n(n+2)(n-2)/3$.

5. Relationship to the Cartan connection. On a manifold with projective structure, it is shown in [5] how to associate vector bundles with connection to any irreducible representation of $\mathrm{SL}(n+1,\mathbb{R})$. These are the tractor bundles following their construction by Thomas [10]. Equivalently, they are induced by the Cartan connection [4] of the projective structure. The relevant tractor bundle in our case is induced by $\bigodot^2\mathbb{R}^{n+1}$ where $\mathbb{R}^{n+1}$ is the defining representation of $\mathrm{SL}(n+1,\mathbb{R})$. It has a composition series

$$\mathcal{E}^{(BC)} = \mathcal{E}^{(bc)}(-2) + \mathcal{E}^b(-2) + \mathcal{E}(-2)$$

and in the presence of a connection is simply the direct sum of these bundles. Under projective change of connection according to (1.1), however, we decree that

$$\widehat{\begin{pmatrix} \sigma^{bc} \\ \mu^b \\ \rho \end{pmatrix}} = \begin{pmatrix} \sigma^{bc} \\ \mu^b + \Upsilon_c\sigma^{bc} \\ \rho + 2\Upsilon_b\mu^b + \Upsilon_b\Upsilon_c\sigma^{bc} \end{pmatrix}. \tag{5.1}$$

Following [5], the tractor connection on $\mathcal{E}^{(AB)}$ is given by

$$\nabla_a \begin{pmatrix} \sigma^{bc} \\ \mu^b \\ \rho \end{pmatrix} = \begin{pmatrix} \nabla_a\sigma^{bc} - \delta_a{}^b\mu^c - \delta_a{}^c\mu^b \\ \nabla_a\mu^b - \delta_a{}^b\rho + \mathrm{P}_{ac}\sigma^{bc} \\ \nabla_a\rho + 2\mathrm{P}_{ab}\mu^b \end{pmatrix}.$$

Therefore, we have proved:

THEOREM 5.1. *The solutions of* (2.3) *are in one-to-one correspondence with solutions of the following system:*

$$\nabla_a \begin{pmatrix} \sigma^{bc} \\ \mu^b \\ \rho \end{pmatrix} - \frac{1}{n}\begin{pmatrix} 0 \\ W_{ac}{}^b{}_d\sigma^{cd} \\ 4Y_{abc}\sigma^{bc} \end{pmatrix} = 0. \tag{5.2}$$

COROLLARY 5.1. *There is a one-to-one correspondence between solutions of* (5.2) *for positive definite* σ^{bc} *and metric connections that are projectively equivalent to* ∇_a.
Notice that the extra terms in (5.2) are projectively invariant as they should be. Specifically, it is observed in [5] that

$$\hat{Y}_{abc} = Y_{abc} + \tfrac{1}{2} W_{ab}{}^{d}{}_{c} \Upsilon_d$$

and so

$$4\hat{Y}_{abc}\sigma^{bc} = 4Y_{abc}\sigma^{bc} + 2\Upsilon_b W_{ac}{}^{b}{}_{d}\sigma^{cd}$$

in accordance with (5.1).

It is clear from Theorem 3.1 that, generically, (2.3) has no solutions. Indeed, this is one reason why the prolonged from is so helpful. More generally, we should compute the curvature of the connection (3.6) and the form (5.2) is useful for this task. A model computation along these lines is given in [5]. In our case, the tractor curvature is given by

$$(\nabla_a\nabla_b - \nabla_b\nabla_a)\begin{pmatrix} \sigma^{cd} \\ \mu^c \\ \rho \end{pmatrix} = \begin{pmatrix} W_{ab}{}^{c}{}_{e}\sigma^{de} + W_{ab}{}^{d}{}_{e}\sigma^{ce} \\ W_{ab}{}^{c}{}_{d}\mu^d + 2Y_{abd}\sigma^{cd} \\ 4Y_{abc}\mu^c \end{pmatrix}$$

and we obtain:

PROPOSITION 5.1. *The curvature of the connection* (3.6) *is given by*

$$\begin{pmatrix} \sigma^{cd} \\ \mu^c \\ \rho \end{pmatrix} \mapsto \begin{pmatrix} W_{ab}{}^{c}{}_{e}\sigma^{de} + W_{ab}{}^{d}{}_{e}\sigma^{ce} \\ W_{ab}{}^{c}{}_{d}\mu^d + 2Y_{abd}\sigma^{cd} \\ 4Y_{abc}\mu^c \end{pmatrix} + \frac{1}{n}\begin{pmatrix} \delta_a{}^{c}U_b{}^{d} + \delta_a{}^{d}U_b{}^{c} - \delta_b{}^{c}U_a{}^{d} - \delta_b{}^{d}U_a{}^{c} \\ * \\ * \end{pmatrix},$$

where $U_b{}^d = W_{be}{}^{d}{}_{f}\sigma^{ef}$ *and* $*$ *denotes expressions that we shall not need.*

COROLLARY 5.2. *The curvature of the connection* (3.6) *vanishes if and only if the projective structure is flat.*

Proof. Let us suppose that $n \geq 3$. The uppermost entry of the curvature is given by

$$\sigma^{cd} \longmapsto \text{the trace-free part of } (W_{ab}{}^{c}{}_{e}\sigma^{de} + W_{ab}{}^{d}{}_{e}\sigma^{ce})$$

and it is a matter of elementary representation theory to show that if this expression is zero for a fixed $W_{ab}{}^{c}{}_{d}$ and for all σ^{cd}, then $W_{ab}{}^{c}{}_{d} = 0$. Specifically, the symmetries of $W_{ab}{}^{c}{}_{d}$, namely

$$W_{ab}{}^{c}{}_{d} + W_{ba}{}^{c}{}_{d} = 0, \qquad W_{ab}{}^{c}{}_{d} + W_{bd}{}^{c}{}_{a} + W_{da}{}^{c}{}_{b} = 0, \qquad W_{ab}{}^{a}{}_{d} = 0, \quad (5.3)$$

constitute an irreducible representation of $\mathrm{SL}(n,\mathbb{R})$. Hence, the submodule

$$\{W_{ab}{}^c{}_d \text{ s.t. the trace-free part of } (W_{ab}{}^c{}_e\sigma^{de} + W_{ab}{}^d{}_e\sigma^{ce}) = 0,\ \forall\, \sigma^{cd}\}$$

must be zero since it is not the whole space. We have shown that if the curvature of the connection (3.6) vanishes, then $W_{ab}{}^c{}_d = 0$. For $n \geq 3$ this is exactly the condition that the projective structure be flat. For $n = 2$, the Weyl curvature $W_{ab}{}^c{}_d$ vanishes automatically since the symmetries (5.3) are too severe a constraint. Instead, a similar calculation shows that $Y_{abc} = 0$ and this is the condition that the projective structure be flat. □

Following Mikeš [7], the dimension of the space of solutions of (2.3) is called the degree of mobility of the projective structure. Theorem 3.1 implies that the degree of mobility is bounded by $(n+1)(n+2)/2$ and Corollary 5.2 implies that this bound is achieved only for the flat projective structure. Of course, the flat projective structure may as well be represented by the flat connection $\nabla_a = \partial/\partial x^a$ on $\mathbb{R}^n$, which is the Levi-Civita connection for the standard Euclidean metric. In this case, we may use (3.5) find the general solution of (2.3):

$$\sigma^{ab} = s^{ab} + x^a m^b + x^b m^a + x^a x^b r. \tag{5.4}$$

This form is positive definite near the origin if and only if s^{ab} is positive definite. We conclude that the general projectively flat metric near the origin in $\mathbb{R}^n$ is

$$g^{ab} = \det(\sigma)\,\sigma^{ab},$$

where σ^{ab} is as in (5.4) for some positive definite quadratic form s^{ab}. In fact, these metrics are constant curvature. Rather than prove this by calculation, there is an alternative as follows. As already observed, the Weyl curvature $W_{ab}{}^c{}_d$ corresponds to an irreducible representation of $\mathrm{SL}(n,\mathbb{R})$ characterised by (5.3). In the presence of a metric g_{ab}, however, we should decompose $W_{ab}{}^c{}_d$ further under $\mathrm{SO}(n)$.

PROPOSITION 5.2. *In the presence of a metric* g_{ab}

$$\begin{aligned} W_{ab}{}^c{}_d = C_{ab}{}^c{}_d &+ \tfrac{1}{(n-1)(n-2)}\left(\delta_a{}^c\Phi_{bd} - \delta_b{}^c\Phi_{ad}\right) \\ &+ \tfrac{1}{n-2}\left(\Phi_a{}^c g_{bd} - \Phi_b{}^c g_{ad}\right) \end{aligned} \tag{5.5}$$

where $C_{ab}{}^c{}_d$ *is the Weyl part of the Riemann curvature tensor and* Φ_{ab} *is the trace-free part of the Ricci tensor.*

Proof. According to (1.3),

$$R_{ab}{}^c{}_d = W_{ab}{}^c{}_d + \delta_a{}^c\mathrm{P}_{bd} - \delta_b{}^c\mathrm{P}_{ad} \tag{5.6}$$

but the Riemann curvature decomposes according to

$$R_{abcd} = C_{abcd} + g_{ac}Q_{bd} - g_{bc}Q_{ad} + Q_{ac}g_{bd} - Q_{bc}g_{ad}, \tag{5.7}$$

where Q_{ab} is the Schouten tensor

$$Q_{ab} = \tfrac{1}{n-2}\Phi_{ab} + \tfrac{1}{2n(n-1)}Rg_{ab}.$$

Comparing (5.6) and (5.7) leads, after a short computation, to (5.5). □

COROLLARY 5.3. *A projectively flat metric is constant curvature.*

Proof. If $n \geq 3$ and the projective Weyl tensor vanishes then the only remaining part of the Riemann curvature tensor is the scalar curvature. As usual, a separate proof based on Y_{abc} is needed for the case $n = 2$. □

This corollary is usually stated as follows. If a local diffeomorphism between two Riemannian manifolds preserves geodesics and one of them is constant curvature, then so is the other. This is a classical result due to Beltrami [1].

6. Concluding remarks. Results such as Theorem 2.2 and Theorem 5.1 are quite common in projective, conformal, and other parabolic geometries. It is shown in [5], for example, that the Killing equation in Riemannian geometry is projectively invariant and its solutions are in one-to-one correspondence with covariant constant sections of the tractor bundle $\mathcal{E}_{[AB]}$ equipped with a connection that is derived from (but not quite equal to) the tractor connection. The situation is completely parallel for conformal Killing vectors in conformal geometry and, more generally, for the infinitesimal automorphisms of parabolic geometries [3]. It is well-known that having an Einstein metric in a given conformal class is equivalent to having a suitably positive covariant constant section of the standard tractor bundle $\mathcal{E}^A$ equipped with its usual tractor connection. Gover and Nurowski [6] use this observation systematically to find obstructions to the existence of an Einstein metric within a given conformal class. We anticipate a similar use for Theorem 5.1 in establishing obstructions to the existence of a metric connection within a given projective class.

REFERENCES

[1] E. BELTRAMI, *Rizoluzione del problema: riportare i punti di una superficie sopra un piano in modo che le linee geodetiche vengano rappresentate da linee rette*, Ann. Mat. Pura Appl. (1865), **7**: 185–204.

[2] T.P. BRANSON, A. ČAP, M.G. EASTWOOD, AND A.R. GOVER, *Prolongations of geometric overdetermined systems*, Int. Jour. Math. (2006), **17**: 641–664.

[3] A. ČAP, *Infinitesimal automorphisms and deformations of parabolic geometries*, preprint ESI 1684 (2005), Erwin Schrödinger Institute, available at `http://www.esi.ac.at`.

[4] E. CARTAN, *Sur les variétés à connexion projective*, Bull. Soc. Math. France (1924), **52**: 205–241.

[5] M.G. EASTWOOD, *Notes on projective differential geometry*, this volume.

[6] A.R. GOVER AND P. NUROWSKI, *Obstructions to conformally Einstein metrics in n dimensions*, Jour. Geom. Phys. (2006), **56**: 450–484.

[7] J. MIKEŠ, *Geodesic mappings of affine-connected and Riemannian spaces*, Jour. Math. Sci. (1996), **78**: 311–333.

[8] R. Penrose and W. Rindler, *Spinors and Space-time, Vol. 1*, Cambridge University Press 1984.

[9] N.S. Sinjukov, *Geodesic mappings of Riemannian spaces* (Russian), "Nauka," Moscow 1979.

[10] T.Y. Thomas, *Announcement of a projective theory of affinely connected manifolds*, Proc. Nat. Acad. Sci. (1925), **11**: 588–589.

28 04 2020

C1) 1461442826.pdf (Volume 1)

Pág. 148(170) ✡

Fig. 8.1 The trajectory of a light particle with mass m_1 near to heavy particle with mass m_2.

C2) 146144294X.pdf (Volume 2)

Pág. 72(92)

Fig. 4.2 The model for the magnetic field of four sunspots of pairwise opposite polarity.

D) Books_73_2020/
The Geometry of Hamilton and Lagrange Spaces/

The Geometry of Hamilton and Lagrange Spaces...pdf

E) Books_73_2020/
Pleasures of Pi, e and Other Interesting Numbers/
Pleasures of Pi, e....pdf ✡

F) Books_73_2020/
The Bloch-Kato Conjecture for the Riemann Zeta Function/
The Bloch-Kato Conjecture...pdf

28.04.2020

G) Books_73-2020/
Symmetries of Partial Differential Equations/
Symmetries and Overdetermined Systems of Partial Differential Equations...pdf
Pág. 339(346)
Metric connections in projective differential geometry

H) Books_73_2020/
Geometry from Isometries to Special Relativity/
3030421007.pdf
Obs. Livro todo!
Pág. 23(35)
Chapter 2 Sphere
2.1. The Sphere S^2 in $\mathbb{R}^3$

$$S^2 = \{(x,y,z) \in \mathbb{R}^3 / x^2+y^2+z^2=1\}.$$

Pág. 24(36)
Fig. 2.1 Spherical distance $d_{S^2}(p_1,p_2)$ between the two points p_1 and p_2 on S^2.

)*+,)))-) % !! % $ 6"
-1 9 ? " III
' !! 6 6

Chapter 2
Sphere

"The description of right lines and circles, upon which geometry is founded, belongs to mechanics. Geometry does not teach us to draw these lines, but requires them to be drawn."

Isaac Newton (1642–1727)

"Geometry is the art of correct reasoning from incorrectly drawn figures."

Henri Poincaré (1854–1912)

Spherical geometry is almost as old as Euclidean geometry. In fact, the word geometry means 'measurement of the Earth', and the Earth is (more or less) a sphere. The ancient Greek geometers knew that the Earth was spherical. Navigation motivated the study of spherical geometry because, even 2000 years ago, the fact that the earth is curved had a noticeable effect on cartography. In spherical geometry, the 'points' are points on the surface of the sphere. We are not concerned with the 'inside' of the sphere.

2.1 The Sphere $\mathbb{S}^2$ in $\mathbb{R}^3$

The (unit) sphere is

$$\mathbb{S}^2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}.$$

© Springer Nature Switzerland AG 2020

N.-H. Lee, *Geometry: from Isometries to Special Relativity*, Undergraduate Texts in Mathematics, https://doi.org/10.1007/978-3-030-42101-4_2

A *great circle* is a circle on the sphere that divides the sphere into two equal hemispheres. A great circle can also be defined as the intersection of the sphere and a plane that goes through the origin **0**. Two points that are diametrically opposite on the sphere are called *antipodal points*. In spherical geometry, two points determine a great circle unless they are antipodal points, in which case there are infinitely many great circles joining them.

For two points $p_1 = (x_1, y_1, z_1)$ and $p_2 = (x_2, y_2, z_2)$ in $\mathbb{S}^2$, the distance between them in $\mathbb{R}^3$ is

$$d(p_1, p_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}.$$

The *spherical distance* $d_{\mathbb{S}^2}(p_1, p_2)$ is the arc length of the shortest path on $\mathbb{S}^2$ from p_1 to p_2, which is a segment of the great circle through the points p_1 and p_2 (Figure 2.1).

Note that

$$d_{\mathbb{S}^2}(p_1, p_2) = \angle p_1 \mathbf{0} p_2 = 2\theta = 2 \arcsin\left(\frac{1}{2} d(p_1, p_2)\right), \tag{2.1}$$

where $\theta = \frac{1}{2}\angle p_1 \mathbf{0} p_2$ (Figure 2.2).

A plane in $\mathbb{R}^3$ is a set of points defined by a linear equation,

$$ax + by + cz = d,$$

where a, b, c, and d are constants with $(a, b, c) \neq (0, 0, 0)$. Similar to a line in the Euclidean plane, a plane in $\mathbb{R}^3$ can be regarded as the set of points equidistant from two distinct points p_1 and p_2 in $\mathbb{R}^3$:

$$P_{p_1, p_2} = \{p \in \mathbb{R}^3 \mid d(p_1, p) = d(p_2, p)\}.$$

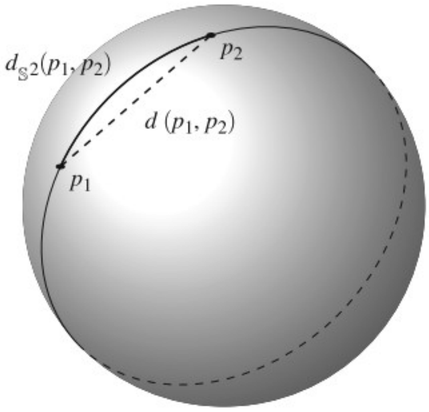

Fig. 2.1 Spherical distance $d_{\mathbb{S}^2}(p_1, p_2)$ between the two points p_1 and p_2 on $\mathbb{S}^2$

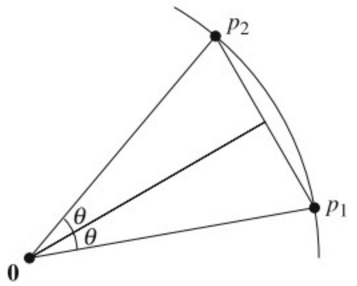

Fig. 2.2 $d_{\mathbb{S}^2}(p_1, p_2) = \angle p_1 \mathbf{0} p_2 = 2\theta = 2 \arcsin\left(\frac{1}{2} d(p_1, p_2)\right)$

Let

$$p_1 = (a_1, b_1, c_1) \text{ and } p_2 = (a_2, b_2, c_2).$$

A point $p = (x, y, z)$ belongs to P_{p_1, p_2} if and only if $d(p_1, p) = d(p_2, p)$, i.e.,

$$\sqrt{(a_1 - x)^2 + (b_1 - y)^2 + (c_1 - z)^2} = \sqrt{(a_2 - x)^2 + (b_2 - y)^2 + (c_2 - z)^2}$$

or

$$(a_1 - a_2)x + (b_1 - b_2)y + (c_1 - c_2)z - d = 0,$$

which is a linear equation, where $d = \frac{1}{2}((a_1^2 + b_1^2 + c_1^2) - (a_2^2 + b_2^2 + c_2^2))$. For a point $p = (x, y, z)$ in $\mathbb{R}^3$, the norm of p is defined by

$$\|p\| = \sqrt{x^2 + y^2 + z^2}.$$

Then, we obviously have $d(p, q) = \|p - q\|$.

Now let us define the notion of maps that preserve the geometric properties of figures on the sphere.

Definition 2.1. A bijective map $\phi : \mathbb{S}^2 \to \mathbb{S}^2$ is called an *isometry* of $\mathbb{S}^2$ if it preserves the spherical distance, i.e.,

$$d_{\mathbb{S}^2}(\phi(p_1), \phi(p_2)) = d_{\mathbb{S}^2}(p_1, p_2)$$

for any two points $p_1, p_2 \in \mathbb{S}^2$.

The main purpose of this chapter is to classify the isometries of the sphere as we did for the Euclidean plane in the previous chapter. In doing so, we will be able to gain considerable knowledge about the spherical geometry.

A bijective map $\phi : \mathbb{R}^3 \to \mathbb{R}^3$ is called an *isometry* of $\mathbb{R}^3$ if the distance is preserved, i.e.,

$$d(\phi(p_1), \phi(p_2)) = d(p_1, p_2)$$

for any two points $p_1, p_2 \in \mathbb{R}^3$.

According to (2.1),

$$d_{\mathbb{S}^2}(\phi(p_1), \phi(p_2)) = d_{\mathbb{S}^2}(p_1, p_2)$$

if and only if

$$2 \arcsin\left(\frac{1}{2} d(\phi(p_1), \phi(p_2))\right) = 2 \arcsin\left(\frac{1}{2} d(p_1, p_2)\right),$$

i.e.,

$$d(\phi(p_1), \phi(p_2)) = d(p_1, p_2).$$

In summary,

$$d_{\mathbb{S}^2}(\phi(p_1), \phi(p_2)) = d_{\mathbb{S}^2}(p_1, p_2) \quad \Leftrightarrow \quad d(\phi(p_1), \phi(p_2)) = d(p_1, p_2). \tag{2.2}$$

Therefore, isometries of the sphere and isometries of $\mathbb{R}^3$ are closely related.

Proposition 2.2. *If an isometry of $\mathbb{R}^3$ fixes the origin $\mathbf{0}$, then it induces an isometry of $\mathbb{S}^2$. In other words, if an isometry ϕ of $\mathbb{R}^3$ satisfies*

$$\phi(\mathbf{0}) = \mathbf{0},$$

then the map $\psi : \mathbb{S}^2 \to \mathbb{S}^2$, given by $\psi(p) = \phi(p)$, is well defined and an isometry of the sphere.

Proof. First, we must verify that $\psi(p)$ belongs to the sphere if the point p does. This is shown by the following:

$$d(\psi(p), \mathbf{0}) = d(\phi(p), \mathbf{0}) = d(\phi(p), \phi(\mathbf{0})) = d(p, \mathbf{0}) = 1.$$

Since ϕ^{-1} is also an isometry of $\mathbb{R}^3$, the map $\psi' : \mathbb{S}^2 \to \mathbb{S}^2$ by $\psi'(p) = \phi^{-1}(p)$ is also well defined. Then, ψ' is the inverse map of ψ, and so ψ is bijective. Finally, by (2.2), ψ preserves the spherical distance because ϕ preserves the distance. □

We denote the set of all the isometries of $\mathbb{S}^2$ by $\mathrm{Iso}(\mathbb{S}^2)$. In Chapter 1, we saw that reflections in lines play important roles when studying isometries. We define a *spherical line* for two distinct points p_1 and p_2 in $\mathbb{S}^2$ (Figure 2.3) as follows:

$$G_{p_1,p_2} = \{p \in \mathbb{S}^2 \mid d_{\mathbb{S}^2}(p_1, p) = d_{\mathbb{S}^2}(p_2, p)\}.$$

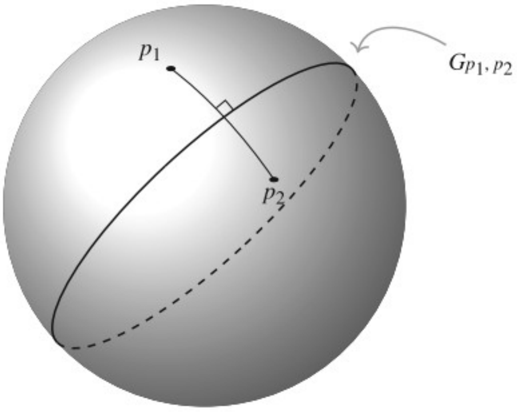

Fig. 2.3 Spherical line G_{p_1,p_2} with respect to points p_1 and p_2 on $\mathbb{S}^2$

Proposition 2.3. *A spherical line is a great circle.*

Proof. For two distinct points p_1 and p_2 on $\mathbb{S}^2$, consider a spherical line

$$G_{p_1,p_2} = \{p \in \mathbb{S}^2 \mid d_{\mathbb{S}^2}(p_1, p) = d_{\mathbb{S}^2}(p_2, p)\}.$$

By (2.2),

$$\begin{aligned} G_{p_1,p_2} &= \{p \in \mathbb{S}^2 \mid d(p_1, p) = d(p_2, p)\} \\ &= \mathbb{S}^2 \cap \{p \in \mathbb{R}^3 \mid d(p_1, p) = d(p_2, p)\} \\ &= \mathbb{S}^2 \cap P_{p_1,p_2}, \end{aligned}$$

which is a intersection of the sphere and a plane P_{p_1,p_2}. The plane P_{p_1,p_2} passes through the origin because

$$d(p_1, \mathbf{0}) = 1 = d(p_2, \mathbf{0}),$$

i.e., $d(p_1, \mathbf{0}) = d(p_2, \mathbf{0})$. □

Points on the sphere are said to be *collinear* if they lie on a great circle.

Lemma 2.4. *An isometry of the sphere maps non-collinear points p_1, p_2, and p_3 to non-collinear points.*

Proof. Let $\phi : \mathbb{S}^2 \to \mathbb{S}^2$ be an isometry. Suppose that the points $\phi(p_1)$, $\phi(p_2)$, and $\phi(p_3)$ are collinear, then they lie on a great circle $G_{p,q}$. Thus, we have

$$\begin{aligned} d_{\mathbb{S}^2}(p, \phi(p_1)) &= d_{\mathbb{S}^2}(q, \phi(p_1)), \\ d_{\mathbb{S}^2}(p, \phi(p_2)) &= d_{\mathbb{S}^2}(q, \phi(p_2)) \end{aligned}$$

and

$$d_{\mathbb{S}^2}(p, \phi(p_3)) = d_{\mathbb{S}^2}(q, \phi(p_3)).$$

However, then,

$$d_{\mathbb{S}^2}(\phi^{-1}(p), p_1) = d_{\mathbb{S}^2}(\phi^{-1}(q), p_1),$$

$$d_{\mathbb{S}^2}(\phi^{-1}(p), p_2) = d_{\mathbb{S}^2}(\phi^{-1}(q), p_2)$$

and

$$d_{\mathbb{S}^2}(\phi^{-1}(p), p_3) = d_{\mathbb{S}^2}(\phi^{-1}(q), p_3),$$

which imply that the points p_1, p_2, and p_3 lie on the great circle $G_{\phi^{-1}(p),\phi^{-1}(q)}$. Because this is a contradiction, the points $\phi(p_1)$, $\phi(p_2)$, and $\phi(p_3)$ are non-collinear. $\square$

In Theorem 1.4, we saw that an isometry of $\mathbb{R}^2$ is determined by how it maps three non-collinear points. The following theorem is a version of Theorem 1.4 for the spherical geometry.

Theorem 2.5 (Three points theorem for the sphere). *Let ϕ and ψ be isometries of the sphere. If*

$$\phi(p_1) = \psi(p_1), \phi(p_2) = \psi(p_2) \text{ and } \phi(p_3) = \psi(p_3)$$

for some set of non-collinear points p_1, p_2, and p_3, then $\phi = \psi$.

Proof. According to Lemma 2.4, the three points $\phi(p_1)$, $\phi(p_2)$, and (p_3) are also non-collinear.

Suppose that $\phi \neq \psi$. Then, there exists a point p such that

$$\phi(p) \neq \psi(p),$$

and we can define a great circle $G = G_{\phi(p),\psi(p)}$. Note that

$$\begin{aligned} d_{\mathbb{S}^2}(\phi(p), \phi(p_1)) &= d_{\mathbb{S}^2}(p, p_1) && (\because\ \phi \text{ is an isometry}) \\ &= d_{\mathbb{S}^2}(\psi(p), \psi(p_1)) && (\because\ \psi \text{ is an isometry}) \\ &= d_{\mathbb{S}^2}(\psi(p), \phi(p_1)), \end{aligned}$$

i.e., $d_{\mathbb{S}^2}(\phi(p), \phi(p_1)) = d_{\mathbb{S}^2}(\psi(p), \phi(p_1))$. So $\phi(p_1) \in G$. Similarly,

$$\phi(p_2) \in G \text{ and } \phi(p_3) \in G.$$

However, the three points $\phi(p_1)$, $\phi(p_2)$, and (p_3) are then collinear, which is a contradiction. $\square$

Exercises

2.1. Let $p_1 = \frac{1}{\sqrt{2}}(1, 1, 0)$ and $p_2 = \frac{1}{\sqrt{2}}(1, 0, 1)$. Calculate the following:

(a) $d(p_1, p_2)$ (b) $d_{\mathbb{S}^2}(p_1, p_2)$.

2.2. (a) Show that there is no map $f : \mathbb{R}^2 \to \mathbb{S}^2$ such that

$$d(p, q) = d_{\mathbb{S}^2}\,(f(p), f(q))$$

for all points p and q in $\mathbb{R}^2$.

(b) Show that there is no map $g : \mathbb{S}^2 \to \mathbb{R}^2$ such that

$$d_{\mathbb{S}^2}(p, q) = d\,(g(p), g(q))$$

for all points p and q in $\mathbb{S}^2$.

2.3. A spherical circle C of radius ρ with its center $p \in \mathbb{S}^2$ is defined by

$$C = \{x \in \mathbb{S}^2 \mid d_{\mathbb{S}^2}(p, x) = \rho\}.$$

Determine its circumference $\Pi(\rho)$. Show that $\Pi(\rho) < 2\pi\rho$ and

$$\lim_{\rho \to 0^+} \frac{\Pi(\rho)}{\rho} = 2\pi.$$

2.2 Isometries of the Sphere $\mathbb{S}^2$

The reflection $\bar{r}_{P_{p_1,p_2}}$ in the plane P_{p_1,p_2} is an isometry of $\mathbb{R}^3$ such that the line segment from the given point p in $\mathbb{R}^3$ to the point $\bar{r}_{P_{p_1,p_2}}(p)$ intersects the plane P_{p_1,p_2} orthogonally at its midpoint. For example, if P is the yz-plane, then

$$\bar{r}_P(x, y, z) = (-x, y, z).$$

Similar to a reflection in a line, a reflection in a plane satisfies the following:

(a)

$$\bar{r}_{P_{p_1,p_2}}(p_1) = p_2, \bar{r}_{P_{p_1,p_2}}(p_2) = p_1.$$

28.04.2020

G) Books_73_2020/
Symmetries of Partial Differential
Equations/
Symmetries and Overdetermined
Systems of Partial Differential
Equations....pdf

Pág. 339(346)

Metric connections in projective
differential geometry

H) Books_73_2020/
Geometry from Isometries to
Special Relativity/
3030421007.pdf

obs. Livro todo!

Pág. 23(35)

Chapter 2 Sphere

2.1. The Sphere $\underline{\underline{S^2}}$ in $\mathbb{R}^3$

$$S^2=\{(x,y,z)\in\mathbb{R}^3 / x^2+y^2+z^2=1\}.$$

Pág. 24(36)

Fig. 2.1 Spherical distance $d_{S^2}(p_1,p_2)$
between the two points p_1 and p_2 on S^2.

&
)*+,)))-) 0 0H J / "
* 6 . + ? / 6"
1)*+,)))-) ! ! !C C 6"
' D 8

for $m \in \mathbb{N}_0$, $s \in (m, m+1)$, $\theta = s - [s]$ and $1 < p < \infty$, often called Sobolev-Slobodeckij spaces,

(f) and the complex interpolation spaces

$$H^{s,p}(\Omega) = \left[W^{m,p}(\Omega), W^{m+1,p}(\Omega)\right]_\theta$$

for $1 \leq p \leq \infty$, $0 < \theta < 1$, $s = (1-\theta)m + \theta(m+1) = m + \theta$, defined on open sets $\Omega \subset \mathbb{R}^n$. We do not give precise definitions of these spaces here and refer the reader e.g. to the a classical monograph by Triebel [149] or the recent one by Amann [12].

1.2 The Vector Valued Setting

As before, we start this section by introducing basic facts on distributions and Fourier transforms, now however in the *vector-valued setting*.

Banach Space Valued Distributions

We note that the proofs of the assertions listed below on distributions are only straightforward modifications of the ones for the scalar case described in Sect. 1.1.

Let X be a Banach space. We denote by $\mathcal{S}(\mathbb{R}^n; X)$ the Schwartz space of smooth rapidly decreasing X-valued functions on $\mathbb{R}^n$. Then

$$\mathcal{D}(\mathbb{R}^n; X) \overset{d}{\hookrightarrow} \mathcal{S}(\mathbb{R}^n; X) \overset{d}{\hookrightarrow} \mathcal{E}(\mathbb{R}^n; X),$$

where $\mathcal{D}(\mathbb{R}^n; X)$ is the space of all X–valued C^∞-functions on $\mathbb{R}^n$ with compact supports, as usual equipped with the inductive limit topology, and $\mathcal{E}(\mathbb{R}^n; X) := C^\infty(\mathbb{R}^n; X)$ equipped with the bounded convergence topology.

The space $\mathcal{S}'(\mathbb{R}^n; X)$ of X–valued temperate distributions is defined by $\mathcal{S}'(X) := \mathcal{L}(\mathcal{S}(\mathbb{R}^n); X)$; the spaces $\mathcal{E}'(\mathbb{R}^n; X)$ and $\mathcal{D}'(\mathbb{R}^n; X)$ are defined analogously. Then

$$\mathcal{E}'(\mathbb{R}^n; X) \hookrightarrow \mathcal{S}'(\mathbb{R}^n; X) \hookrightarrow \mathcal{D}'(\mathbb{R}^n; X).$$

Finally, the space $\mathcal{O}_M(\mathbb{R}^n; X)$ of all X–valued slowly increasing smooth functions on $\mathbb{R}^n$ consists of all $u \in \mathcal{E}(\mathbb{R}^n; X)$ such that, given $\alpha \in \mathbb{N}_0^n$, there exists $m_\alpha \in \mathbb{N}$ and $C_\alpha > 0$ such that

$$\|\partial^\alpha u(x)\| \leq C_\alpha (1 + |x|^2)^{m_\alpha}, \quad x \in \mathbb{R}^n.$$

Observe that $\mathcal{S}(\mathbb{R}^n; X) \hookrightarrow \mathcal{O}_M(\mathbb{R}^n; X) \hookrightarrow \mathcal{S}'(\mathbb{R}^n; X)$. We also write $\mathcal{S}(\mathbb{R}^n) := \mathcal{S}(\mathbb{R}^n; \mathbb{K})$, $\mathcal{S}'(\mathbb{R}^n) := \mathcal{S}'(\mathbb{R}^n; \mathbb{K})$ and $\mathcal{O}_M(\mathbb{R}^n) := \mathcal{O}_M(\mathbb{R}^n; \mathbb{K})$, where $\mathbb{K} \in \{\mathbb{C}, \mathbb{R}\}$.

We next turn our attention to the Bochner integral. To this end, given an interval I in $\mathbb{R}$, bounded or unbounded, or a rectangle I in $\mathbb{R}^2$, we call a function $g : I \to X$ a *simple* function if it is of the form $g(t) = \sum_{j=1}^n x_j \chi_{S_j}(t)$ for some $n \in \mathbb{N}$, $x_j \in X$

and Lebesgue measurable sets $S_j \subset I$ with finite Lebesgue measure $\mu(S_j)$; g is called a *step function* if S_j can be chosen to be an interval. A function $f : I \to X$ is called *measurable* if there exists a sequence of simple functions g_n such that $f(t) = \lim_{n\to\infty} g_n(t)$ for almost all $t \in I$. For a simple function $g : I \to X$ of the form $g = \sum_{j=1}^{n} x_j \chi_{S_j}$ we define

$$\int_I g(t)dt := \sum_{j=1}^{n} x_j \mu(S_j).$$

A function $f : I \to X$ is then called *Bochner integrable* if there exist simple functions g_n such that $g_n \to f$ pointwise almost everywhere and $\lim_{n\to\infty} \int_I \|f(t) - g_n(t)\| dt = 0$. In this case, the *Bochner integral of f on I* is given by

$$\int_I f(t)dt := \lim_{n\to\infty} \int_I g_n(t)dt.$$

The class of Bochner integrable functions can be characterized as follows.

Proposition 1.2.1 (Bochner) *A function $f : I \to X$ is Bochner integrable if and only if f is measurable and $\|f\|$ is integrable. If f is Bochner integrable, then*

$$\Big\| \int_I f(t)dt \Big\| \le \int_I \|f(t)\| dt.$$

For $1 \le p < \infty$, we denote by $L^p(I; X)$ the space of all measurable functions $f : I \to X$ such that

$$\|f\|_p := \Big(\int_I \|f(t)\|^p dt \Big)^{1/p} < \infty.$$

Furthermore, we set $L^\infty(I; X)$ to be the space of all measurable functions $f : I \to X$ such that

$$\|f\|_\infty := \operatorname*{ess\,sup}_{t\in I} \|f(t)\| < \infty.$$

With the usual identifications, $L^p(I; X)$ becomes a Banach space for all $1 \le p \le \infty$. We note that the above theory of integration works in the same way, when the interval I is replaced by a measurable set in $\mathbb{R}^n$. It follows from Fubini's theorem that for $1 \le p < \infty$ there is an isometric isomorphism between $L^p(I \times \Omega; X)$ and $L^p(I; L^p(\Omega; X))$ for any measurable set $\Omega \subset \mathbb{R}^n$ given by $f \mapsto g$, where

$$(g(s))(t) := f(s, t).$$

We further note that for $1 \le p < \infty$ the norm of a function $f \in L^p(\mathbb{R}^n; X)$ can be calculated by

$$\begin{aligned} &\|f\|_{L^p(\mathbb{R}^n;X)} \\ &\quad = \sup \Big\{ \Big| \int_{\mathbb{R}^n} < g(x), f(x) > dx \Big| : g \in L^{p'}(\mathbb{R}^n; X'), \|g\|_{L^{p'}(\mathbb{R}^n;X')} \le 1 \Big\}. \end{aligned} \tag{1.2.1}$$

Given $u \in L^1(\mathbb{R}^n; X)$ we define the *Fourier transform* of u by

$$\mathcal{F}u(\xi) := \hat{u}(\xi) := \int_{\mathbb{R}^n} e^{-i<x,\xi>} u(x)dx, \quad \xi \in \mathbb{R}^n,$$

where as above $< x, \xi >= \sum_{j=1}^n x_j \xi_j$. The *Fourier inversion theorem* guarantees that the Fourier transform $\mathcal{F}$ is an isomorphism on $\mathcal{S}(\mathbb{R}^n; X)$ and that

$$\mathcal{F}^{-1}u = (2\pi)^{-n} \check{\hat{u}} = (2\pi)^{-n} \hat{\check{u}}, \quad u \in \mathcal{S}(\mathbb{R}^n; X) \tag{1.2.2}$$

where $\check{u}(x) := u(-x)$ for $x \in \mathbb{R}^n$ denotes the reflection of u. The Fourier transform $\hat{u} := \mathcal{F}u$ of $u \in \mathcal{S}'(\mathbb{R}^n; X)$ is defined by

$$\hat{u}(\varphi) := u(\hat{\varphi}), \quad \varphi \in \mathcal{S}(\mathbb{R}^n).$$

Defining $\check{u}(\varphi) := u(\check{\varphi})$ for $u \in \mathcal{D}'(\mathbb{R}^n; X)$ and $\varphi \in \mathcal{D}(\mathbb{R}^n)$ it follows that $\mathcal{F}$ is an isomorphism on $\mathcal{S}'(\mathbb{R}^n; X)$ and that (1.2.2) holds for $u \in \mathcal{S}'(\mathbb{R}^n; X)$.

Let $\varphi \in \mathcal{S}(\mathbb{R}^n)$ and $u \in \mathcal{S}'(\mathbb{R}^n; X)$. Then the convolution $u * \varphi$ of u and φ is defined by

$$(u * \varphi)(x) := u(\tau_x \check{\varphi}), \quad x \in \mathbb{R}^n,$$

where $\tau_a \varphi(x) := \varphi(x - a)$ for $x \in \mathbb{R}^n$ and $a \in \mathbb{R}^n$. Moreover, the convolution theorem states that

$$(u * \varphi)\hat{} = \hat{u}\hat{\varphi}.$$

Finally, let H be a Hilbert space. Then $L^2(\mathbb{R}^n; H)$ is a Hilbert space with respect to the inner product

$$(u|v)_2 := (u|v)_{L^2(\mathbb{R}^n;H)} := \int_{\mathbb{R}^n} u \cdot \overline{v}\, dx.$$

We note that Plancherel's theorem carries over to the situation of Hilbert space valued functions. More precisely, the following holds.

Proposition 1.2.2 (Plancherel's Theorem in Hilbert Spaces) *Let H be a Hilbert space and $u, v \in \mathcal{S}(\mathbb{R}^n; H)$. Then*

$$(\hat{u}|\hat{v})_2 = (2\pi)^n (u|v)_2$$

Moreover, $(2\pi)^{-n/2}\mathcal{F}$ is a unitary operator on $L^2(\mathbb{R}^n; H)$.

Singular Integrals: Classical Theory

Singular integral operators will be an important tool in the analysis of the Stokes equation in the following sections. We now state two results on the L^p-boundedness of singular integral operators in the vector valued setting.

Consider a kernel operator of the form

$$(Tf)(x) = \int_{\mathbb{R}^n} K(x, y) f(y) d\mu(y), \tag{1.2.3}$$

where the kernel K is singular near $x = y$, and (1.2.3) is meaningful only in some limiting sense. More precisely, we assume that for $f \in L^2(\mathbb{R}^n; X)$ with compact support, the integral in (1.2.3) converges absolutely for a. a. $x \in (\operatorname{supp} f)^C$ and that (1.2.3) holds for these x, where the kernel K is assumed to belong to $L^1_{loc}(\mathbb{R}^n \times \mathbb{R}^n, \mathcal{L}(X, Y))$.

Theorem 1.2.3 *Let X and Y be Banach spaces and let T be defined as in (1.2.3). Assume that there exists a constant $C > 0$ such that*

(i) $\|Tf\|_{L^2(\mathbb{R}^n;Y)} \le C\|f\|_{L^2(\mathbb{R}^n;X)}$ *for all* $f \in L^2(\mathbb{R}^n; X)$
(ii) There exists a constant $r > 1$ such that for all $y_1 \in \mathbb{R}^n$, $\delta > 0$

$$\int_{B(y_1, r\delta)^C} \left\| K(x, y_1) - K(x, y_2) \right\|_{\mathcal{L}(X,Y)} d\mu(x) \le C \text{ if } y_2 \in B(y_1, \delta)$$

Then, for all $1 < p < 2$, there exists a constant C_p (depending only on C and p) such that

$$\|Tf\|_{L^p(\mathbb{R}^n;Y)} \le C_p \|f\|_{L^p(\mathbb{R}^n;X)}, \text{ for } f \in L^p(\mathbb{R}^n; X).$$

We note that the proof of the scalar-valued version of Theorem 1.2.3 given e.g. in [140] extends without any difficulties to the situation considered above.

Remark 1.2.4 It is worthwhile to point out the following variant of Theorem 1.2.3. Assume that T defined as in (1.2.3) satisfies the condition

$$\|Tf\|_{L^r(\mathbb{R}^n;Y)} \le C\|f\|_{L^r(\mathbb{R}^n;X)}, \qquad f \in L^r(\mathbb{R}^n; X)$$

for some $r \in (1, \infty)$ and condition (ii) of Theorem 1.2.3. Then T admits a bounded extensions to L^p for $1 < p < r$. The proof follows the lines of the proof given in [140] and is left to the reader.

Consider now the special case of convolution operators. More specifically, let X and Y be Banach spaces and suppose that $K \in L^1_{loc}(\mathbb{R}^n \setminus \{0\}, \mathcal{L}(X, Y))$. Then the integral

$$Tf(x) := \int_{\mathbb{R}^n} K(x - y)f(y)dy \tag{1.2.4}$$

is well defined for $f \in L^\infty(\mathbb{R}^n; X)$ with compact support and $x \in (suppf)^C$. For operators of the form (1.2.4), the condition (ii) of Theorem 1.2.3 is equivalent to the *Hörmanders's condition*

$$\int_{|x|>2|y|} \big\|K(x - y) - K(x)\big\|dx \leq C < \infty, \quad y \in \mathbb{R}^n\setminus\{0\}. \tag{1.2.5}$$

The following result due to Benedek, Calderón and Panzone says that operators of the form (1.2.4) are bounded operators on $L^p(\mathbb{R}^n; X)$, for $1 < p < \infty$, provided this holds for some $r \in (1, \infty)$ and condition (1.2.5) is satisfied.

Theorem 1.2.5 (Benedek, Calderón, Panzone) *[16] Suppose that* $T \in \mathcal{L}(L^r(\mathbb{R}^n; X)\ L^r(\mathbb{R}^n; Y))$ *for some* $r \in (1, \infty)$. *Assume that* T *may be represented by (1.2.4) for* $f \in L^\infty(\mathbb{R}^n; X)$ *with compact support and* $x \in (\,supp\ f)^C$, *and that* (1.2.5) *is satisfied. Then* T *admits a bounded extension to* $L^p(\mathbb{R}^n; X)$ *for* $p \in (1, \infty)$. *Moreover, there exists a constant C such that*

$$\|Tf\|_{L^p(\mathbb{R}^n;Y)} \leq C\|f\|_{L^p(\mathbb{R}^n;X)}, \quad 1 < p < \infty.$$

Proof We note that for $p \in (1, r)$ the assertion follows from Remark 1.2.4 since for translation invariant operators of the form (1.2.4), the condition ii) of Theorem 1.2.3 is equivalent to (1.2.5). The remaining case where $p \in (r, \infty)$ follows by a duality argument. □

Hilbert Transform and UMD-Spaces

An interesting application of Theorem 1.2.5 concerns the Hilbert transform on $L^p(\mathbb{R}; X)$. We already saw in the previous Sect. 1.1 that the Hilbert transform acts boundedly on $L^p(\mathbb{R})$. As in the scalar case, we define for $u \in \mathcal{S}(\mathbb{R}^n; X)$ the *Hilbert transform* Hu of u by

$$Hu := \frac{1}{\pi} pv(1/t) * u, \quad u \in \mathcal{S}(\mathbb{R}^n; X).$$

For the time being, assume that for some $p \in (1, \infty)$ we have $\|Hu\|_{L^p(\mathbb{R};X)} \leq C\|u\|_{L_p(\mathbb{R};X)}$ for $u \in \mathcal{S}(\mathbb{R}; X)$. Then there exists a unique extension of H to a bounded operator on $L^p(\mathbb{R}; X)$. This extension, also denoted by H, is called the *Hilbert transform* on $L^p(\mathbb{R}; X)$. Since $\widehat{Hu} = -isign(\cdot)\hat{u}$, it follows from Plancherel's Theorem 1.2.2 that

$$(Hu|Hv)_2 = \frac{1}{2\pi}(\widehat{Hu}|\widehat{Hv})_2 = \frac{1}{2\pi}(\hat{u}|\hat{v})_2 = (u|v)_2$$

for $u, v \in \mathcal{S}(\mathbb{R}; X)$. By density, H is an isomorphism on $L^2(\mathbb{R}; X)$ satisfying $H^{-1}v = -Hv$ for $v \in L^2(\mathbb{R}; X)$. We summarize these observations in part (a) of the following proposition.

Proposition 1.2.6 *Let X be a Banach space.*

(a) The Hilbert transform is a unitary operator on $L^2(\mathbb{R}; X)$ provided X is a Hilbert space.
(b) Suppose that the Hilbert transform is bounded on $L^p(\mathbb{R}; X)$ for some $p \in (1, \infty)$. Then it is bounded on $L^q(\mathbb{R}; X)$ for all $q \in (1, \infty)$.

Proof Note that assertion (b) is an immediate consequence of Theorem 1.2.5 since

$$\int_{|x|>2|y|} \|K(x-y) - K(x)\| dx = \int_{|x|>2|y|} \Big|\frac{1}{x-y} - \frac{1}{x}\Big| dx \le C.$$

□

Let us remark that the Hilbert transform is not bounded on $L^p(\mathbb{R}; X)$, in general. Banach spaces for which the Hilbert transform is, however, bounded for some (and then all) $p \in (1, \infty)$ play an important role in the regularity theory of parabolic evolution equations. We finish this section with the definition of such spaces.

Definition 1.2.7 A Banach space X is called a *UMD-space* if the Hilbert transform is bounded on $L^p(\mathbb{R}; X)$ for some (and then all) $p \in (1, \infty)$.

The word UMD stands for the property of *unconditional martingale differences* and goes back to the work of Burkholder [24] and Bourgain [20]. For a comprehensive treatment of these spaces we refer to the monographs [87] and [8]. The following remark collects some basic facts and examples of UMD-spaces.

Remarks 1.2.8

(a) Let us remark here that every finite dimensional Banach space or any Hilbert space is a UMD-space.
(b) UMD spaces are reflexive. The converse is not true.
(c) If X is a UMD space and (Ω, μ) a σ-finite measure space, then $L^p(\Omega, \mu; X)$ has the UMD-property provided $1 < p < \infty$.
(d) All closed linear subspaces of UMD-spaces are UMD spaces.
(e) $L^1(\Omega, \mu)$ or spaces of continuous functions $C(K)$ do not have the UMD property.

)*+,)))-) .
+ " J *66 6 " H A B 3 K B 6"
' $ 6 ? 6L "

12

Mechanical Curves

In this chapter we shall describe mechanical curves from the Lagrangian and Hamiltonian point of view. In this way, many geometric properties of these curves will be derived from the variational formalism.

A mechanical curve is a curve described by a particle on which acts an exterior force. For instance the circle, cycloid, hypocycloid, astroid, etc are models of particle trajectories under some exterior forces. A particle on which acts a central force of constant magnitude describes a circle. A point on a circle which rolls on a line, without slipping, describes a cycloid. A point on a circle tangent interior to another circle, which rotates without slipping in the interior of the large circle, describes a hypocycloid.

12.1 The areal velocity

Suppose an object moves in the plane from the point A to point B along a continuous arc $\widehat{AB}$. Let $\mathcal{A}$ be the area swept by the vectorial radius $\overrightarrow{OX}$ with $X \in \widehat{AB}$. We shall consider positive the orientation given by the clock-wise rotation of $\overrightarrow{OX}$. An elementary calculus formula in polar coordinates yields

$$\mathcal{A} = \frac{1}{2}\int_0^{\phi} r^2\, d\phi, \tag{12.1.1}$$

where $r = r(\phi)$ is the length of the vectorial radius and the argument angle $\phi = \angle AOX$.

Written in differential form, we have $d\mathcal{A} = \dfrac{1}{2} r^2\, d\phi$. Let t be the time parameter. Then

$$\frac{d\mathcal{A}}{dt} = \frac{1}{2} r^2 \frac{d\phi}{dt} = \frac{1}{2} r^2 \dot{\phi}. \tag{12.1.2}$$

The derivative $d\mathcal{A}/dt$ is called *areal velocity*.

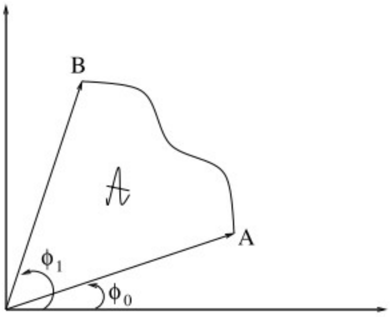

Figure 12.1: The area swept by the vectorial radius between two points.

12.1.0.1 Areal velocity as an angular momentum

Using polar coordinates $x = r\ \cos\phi,\ y = r\ \sin\phi$, the areal velocity becomes

$$\begin{aligned}\frac{d\mathcal{A}}{dt} &= \frac{1}{2}r^2\dot{\phi} = \frac{1}{2}\big(r\cos\phi\, r\cos\phi\,\dot{\phi} - r\sin\phi\,(-r\cos\phi)\,\dot{\phi}\big)\\ &= \frac{1}{2}(x\dot{y} - y\dot{x}).\end{aligned}$$

The expression $x\dot{y} - y\dot{x} = \langle (x, y), (\dot{x}, -\dot{y})\rangle$ is called *angular momentum.*

If $x\dot{y} - y\dot{x}$ is constant, the particle moves such that equal areas are described in equal amounts of time *i.e.*, the vectorial radius sweeps out equal areas in equal time. This happens for instance, in the case of a particle in uniform motion on a circle or in the case of a planet in the revolution motion around the sun (Kepler's second law).

12.2 The circular motion

Consider a particle in the (x, y)-plane which is described by the Lagrangian

$$L(x, y, \dot{x}, \dot{y}) = \frac{1}{2}(\dot{x}^2 + \dot{y}^2) + (x\dot{y} - y\dot{x}) \tag{12.2.3}$$

i.e., the particle moves on a trajectory which is an extremizer for the action

$$S = \int L(x, y, \dot{x}, \dot{y}). \tag{12.2.4}$$

The Lagrangian L is the sum of the kinetic energy and the angular momentum.

Theorem 12.1. *The Euler–Lagrange system associated with the Lagrangian* (12.2.3) *is*

$$\begin{cases}\ddot{x} - 2\dot{y} &= 0,\\ \ddot{y} + 2\dot{x} &= 0.\end{cases} \tag{12.2.5}$$

The solutions of the system (12.2.5) *with the boundary conditions*

$$x(0) = \mathbf{x}_0, \qquad x(\tau) = \mathbf{x}, \qquad y(0) = \mathbf{y}_0, \qquad y(\tau) = \mathbf{y}, \tag{12.2.6}$$

with $0 < \tau < \pi$, *are*

$$\begin{aligned} x(s) &= \pm C \sin s\ \sin(s + \alpha_0) + \mathbf{x}_0, \\ y(s) &= \pm C \sin s\ \cos(s + \alpha_0) + \mathbf{y}_0, \end{aligned}$$

with $C = \sqrt{\frac{E}{2}}$ *and the energy* E *given by*

$$E = \frac{(\mathbf{x} - \mathbf{x}_0)^2 + (\mathbf{y} - \mathbf{y}_0)^2}{2 \sin^2 \tau}. \tag{12.2.7}$$

Proof. Using

$$\begin{aligned} &\frac{\partial L}{\partial \dot{x}} = \dot{x} - y, \qquad \frac{d}{dt}\frac{\partial L}{\partial \dot{x}} = \ddot{x} - \dot{y}, \qquad \frac{\partial L}{\partial x} = \dot{y}, \\ &\frac{\partial L}{\partial \dot{y}} = \dot{y} + x, \qquad \frac{d}{dt}\frac{\partial L}{\partial \dot{y}} = \ddot{y} + \dot{x}, \qquad \frac{\partial L}{\partial \dot{y}} = -\dot{x}, \end{aligned}$$

it is easy to see that the Euler–Lagrange system becomes (12.2.5).

Multiplying the first equation of (12.2.5) by $\dot{x}$ and the second by $\dot{y}$ and adding yields $\dot{x}\ddot{x} + \dot{y}\ddot{y} = 0$, therefore

$$\frac{d}{dt}(\dot{x}^2 + \dot{y}^2) = 0 \Longrightarrow \dot{x}^2 + \dot{y}^2 = C^2 \tag{12.2.8}$$

where C is a constant along the trajectory. Let $E = \frac{1}{2}(\dot{x}^2 + \dot{y}^2)$ denote the first integral of energy. Using (12.2.8) there is a smooth function $\alpha = \alpha(s)$ such that

$$\begin{aligned} \dot{x}(s) &= \pm C \sin\alpha(s) \Longrightarrow \ddot{x}(s) = \pm C \cos\alpha(s)\, \dot{\alpha}(s), \\ \dot{y}(s) &= \pm C \cos\alpha(s) \Longrightarrow \ddot{y}(s) = \mp C \sin\alpha(s)\, \dot{\alpha}(s). \end{aligned}$$

Substituting back in the system (12.2.5) yields

$$\begin{aligned} \pm \cos^2\alpha(s)\, \dot{\alpha}(s) &= \pm 2 \cos^2\alpha(s), \\ \mp \sin^2\alpha(s)\, \dot{\alpha}(s) &= \mp 2 \sin^2\alpha(s). \end{aligned}$$

Subtracting we get

$$\dot{\alpha}(s) = 2 \Longrightarrow \alpha(s) = 2s + \alpha_0, \tag{12.2.9}$$

with α_0 constant. Hence

$$\begin{aligned}\dot{x}(s) &= \pm C\sin(2s+\alpha_0) \Longrightarrow x(s) = \pm C\int_0^s \sin(2u+\alpha_0)\,du + \mathbf{x}_0\\ &= \pm C\frac{1}{2}(-\cos w)\Big|_{\alpha_0}^{2s+\alpha_0} + \mathbf{x}_0 = \pm\frac{C}{2}\Big(\cos\alpha_0 - \cos(2s+\alpha_0)\Big) + \mathbf{x}_0\\ &= \pm C\sin s\sin(s+\alpha_0) + \mathbf{x}_0,\end{aligned}$$

where we used

$$\begin{aligned}\cos\alpha_0 - \cos(2s+\alpha_0) &= -2\sin\frac{\alpha_0+2s+\alpha_0}{2}\sin\frac{\alpha_0-2s-\alpha_0}{2}\\ &= 2\sin(s+\alpha_0)\sin s.\end{aligned}$$

Substituting $\alpha(s)$ in the formula for $\dot{y}(s)$ yields

$$\begin{aligned}\dot{y}(s) &= \pm C\cos\alpha(s) = \pm C\cos(2s+\alpha_0)\\ \Longrightarrow y(s) &= \pm C\int_0^s \cos(2u+\alpha_0)\,du + \mathbf{y}_0\\ &= \pm\frac{1}{2}C\sin w\Big|_{\alpha_0}^{2s+\alpha_0} + \mathbf{y}_0\\ &= \pm\frac{1}{2}C\Big(\sin(2s+\alpha_0) - \sin\alpha_0\Big) + \mathbf{y}_0\\ &= \pm C\sin s\cos(s+\alpha_0) + \mathbf{y}_0,\end{aligned}$$

where we used

$$\begin{aligned}\sin(2s+\alpha_0) - \sin\alpha_0 &= 2\sin\frac{2s+\alpha_0-\alpha_0}{2}\cos\frac{2s+\alpha_0+\alpha_0}{2}\\ &= 2\sin s\cos(s+\alpha_0).\end{aligned}$$

Hence we have arrived at

$$x(s) = \pm C\sin s\ \sin(s+\alpha_0) + \mathbf{x}_0, \tag{12.2.10}$$

$$y(s) = \pm C\sin s\ \cos(s+\alpha_0) + \mathbf{y}_0, \tag{12.2.11}$$

where α_0 is a constant. We shall show that energy $E = C^2/2$ does not depend on α_0. Making $s = \tau$ in (12.2.10), (12.2.11) yields

$$\mathbf{x} = \pm C\sin\tau\ \sin(\tau+\alpha_0) + \mathbf{x}_0, \tag{12.2.12}$$

$$\mathbf{y} = \pm C\sin\tau\ \cos(\tau+\alpha_0) + \mathbf{y}_0, \tag{12.2.13}$$

hence it follows that

$$\Big(\frac{\mathbf{x}-\mathbf{x}_0}{\sin\tau}\Big)^2 = C^2\sin^2(\tau+\alpha_0), \tag{12.2.14}$$

$$\Big(\frac{\mathbf{y}-\mathbf{y}_0}{\sin\tau}\Big)^2 = C^2\cos^2(\tau+\alpha_0). \tag{12.2.15}$$

Adding yields

$$2E = C^2 = \frac{(\mathbf{x}-\mathbf{x}_0)^2 + (\mathbf{y}-\mathbf{y}_0)^2}{\sin^2\tau},$$

which is (12.2.7). ∎

Remark 12.2 *Let* $d = \sqrt{(\mathbf{x}-\mathbf{x}_0)^2 + (\mathbf{y}-\mathbf{y}_0)^2}$ *denote the Euclidean distance between* $(\mathbf{x}_0, \mathbf{y}_0)$ *and* $(\mathbf{x}, \mathbf{y})$. *We note the fact that the energy* $E = \dfrac{d^2}{2\sin^2\tau}$ *is not Euclidean. Replacing* $\sin^2\tau$ *by* τ^2 *we obtain the Euclidean energy.*
Let δ *be the Riemannian distance in which the solutions of the Euler–Lagrange equations become geodesics. Then* $E = \dfrac{\delta^2}{2\tau^2}$. *Then* $\delta^2 = \left(\dfrac{\tau}{\sin\tau}\right)^2 d^2$, *and hence* d *and* δ *are homothetic.*

The action

The action $S = S(\mathbf{x}_0, \mathbf{y}_0, \mathbf{x}, \mathbf{y}, \tau)$ satisfies the Hamilton–Jacobi equation

$$\begin{aligned} \frac{\partial S}{\partial \tau} &= -E = -\frac{d^2}{2\sin^2\tau} = \frac{d^2}{2}\frac{\partial}{\partial\tau}(\cot\tau) \\ \Longrightarrow 0 &= \frac{\partial}{\partial\tau}\Big(S - \frac{d^2}{2}\cot\tau\Big) \Longrightarrow S = S_0 + \frac{d^2}{2}\cot\tau. \end{aligned} \tag{12.2.16}$$

Proposition 12.3 *The Hamiltonian associated with the Lagrangian* (12.2.3) *is*

$$H(x, y, p_1, p_2) = \frac{1}{2}\big(p_1 + y\big)^2 + \frac{1}{2}\big(p_2 - x\big)^2. \tag{12.2.17}$$

Proof. The Hamiltonian system for the Hamiltonian (12.2.17) yields

$$\begin{aligned} \dot{x} &= H_{p_1} = p_1 + y \Longrightarrow p_1 = \dot{x} - y, \\ \dot{y} &= H_{p_2} = p_2 - x \Longrightarrow p_2 = \dot{y} + x. \end{aligned}$$

Using the Legendre transform we have

$$\begin{aligned} L &= p_1\dot{x} + p_2\dot{y} - H \\ &= (\dot{x} - y)\dot{x} + (\dot{y} + x)\dot{y} - \frac{1}{2}\big(p_1 + y\big)^2 - \frac{1}{2}\big(p_2 - x\big)^2 \\ &= (\dot{x} - y)\dot{x} + (\dot{y} + x)\dot{y} - \frac{1}{2}\dot{x}^2 - \frac{1}{2}\dot{y}^2 \\ &= \frac{1}{2}(\dot{x}^2 + \dot{y}^2) + x\dot{y} - y\dot{x}. \end{aligned}$$

∎

We note that the Hamiltonian (12.2.17) is the principal symbol of the operator

$$P = \frac{1}{2}\big(\partial_x + y\big)^2 + \frac{1}{2}\big(\partial_y - x\big)^2$$
$$= \frac{1}{2}(\partial_x^2 + \partial_y^2) + y\partial_x - x\partial_y + \frac{1}{2}(x^2 + y^2),$$

which describes the circular motion.

12.3 The astroid

The trajectory of a point P on the unit circle which rolls without slipping in the interior of a circle of radius 4 is a hypocycloid with four cuspidal points. This curve is called *astroid*. The equation of the astroid is

$$x^{2/3} + y^{2/3} = 1. \tag{12.3.18}$$

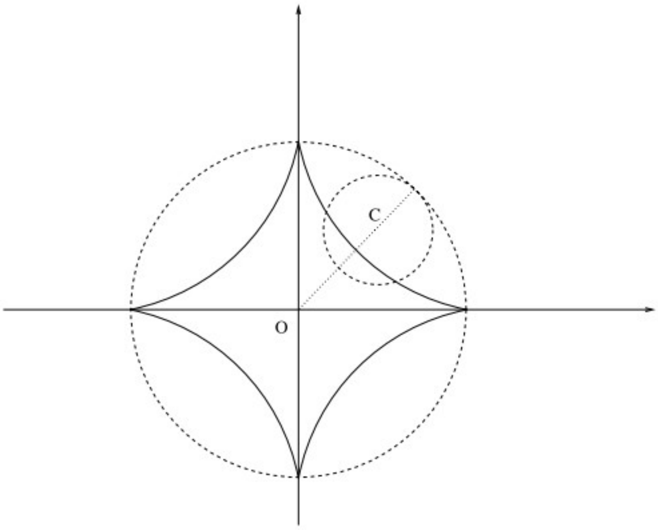

Figure 12.2: The astroid.

If P starts at the cuspidal point (4, 0) and s denotes the angle argument of the center C, we have

$$x(s) = \cos^3 s, \qquad y(s) = \sin^3 s, \tag{12.3.19}$$

which are equivalent with

$$x(s) = 3\cos s + \cos 3s, \qquad y(s) = 3\sin s - \sin 3s. \tag{12.3.20}$$

A simple computation shows that (12.3.20) is the solution of the system

$$\begin{cases} \ddot{x} - 2\dot{y} + 3x = 0, \\ \ddot{y} + 2\dot{x} + 3y = 0, \end{cases} \tag{12.3.21}$$

with initial conditions

$$x(0) = 4, \quad \dot{x}(0) = \dot{y}(0) = y(0) = 0. \tag{12.3.22}$$

Standard ODE techniques show that the solution (12.3.20) is unique.

Proposition 12.4 *The system* (12.3.21) *is the Euler–Lagrange system associated with the Lagrangian*

$$L(x, y, \dot{x}, \dot{y}) = \frac{1}{2}(\dot{x}^2 + \dot{y}^2) + x\dot{y} - \dot{x}y - \frac{3}{2}(x^2 + y^2). \tag{12.3.23}$$

Proof. We have

$$\frac{\partial L}{\partial \dot{x}} = \dot{x} - y, \quad \frac{d}{dt}\frac{\partial L}{\partial \dot{x}} = \ddot{x} - \dot{y}, \quad \frac{\partial L}{\partial x} = \dot{y} - 3x,$$

$$\frac{\partial L}{\partial \dot{y}} = \dot{y} + x, \quad \frac{d}{dt}\frac{\partial L}{\partial \dot{y}} = \ddot{y} + \dot{x}, \quad \frac{\partial L}{\partial y} = -\dot{x} - 3y.$$

Then

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{x}} = \frac{\partial L}{\partial x}, \quad \frac{d}{dt}\frac{\partial L}{\partial \dot{y}} = \frac{\partial L}{\partial y}$$

yields the system (12.3.21). ■

12.3.0.2 Noether's Theorem

The Lagrangian (12.3.23) is invariant under rotations centered at the origin. The vector field associated with this rotation at the point (x, y) is $(-y, x)$. Noether's theorem yields a first integral of motion given by

$$\begin{aligned} I &= \langle \left(\frac{\partial L}{\partial \dot{x}}, \frac{\partial L}{\partial \dot{y}}\right), (-y, x) \rangle = (\dot{x} - y)(-y) + (\dot{y} + x)x \\ &= -\dot{x}y + y^2 + \dot{y}x + x^2 = x^2 + y^2 + x\dot{y} - \dot{x}y \\ &= r^2 + 2\frac{d\mathcal{A}}{ds}, \end{aligned}$$

where $x = r\cos\phi$, $y = r\sin\phi$. We have arrived at the following result.

Proposition 12.5 *For any solution of the system* (12.3.21) *there is a constant C such that*

(i)
$$r^2 + 2\frac{d\mathcal{A}}{ds} = C,$$

(ii)
$$\phi(s) = C\int_0^s \frac{du}{r^2(u)} - s$$

along the solution.

Proof. (i) It clearly follows from the fact that the first integral is constant along the solutions.
(ii) We have

$$\begin{aligned} C &= r^2 + x\dot{y} - \dot{x}y \\ &= r^2 + r^2\dot{\phi} \\ &= r^2(1+\dot{\phi}) \\ \Longrightarrow \frac{d\phi}{ds} &= \frac{C}{r^2} - 1. \end{aligned}$$

Integrating yields the desired result. ■

We can get the same result if we write the Euler–Lagrange system in polar coordinates. See Exercise 2.

As the astroid is a solution of the system (12.3.21), the above proposition applies to it. In this case the constant C is obtained by taking the value at $s = 0$,

$$\begin{aligned} C &= x^2(0) + y^2(0) + x(0)\dot{y}(0) - \dot{x}(0)y(0) \\ &= 16. \end{aligned}$$

Proposition 12.6 *The Hamiltonian associated with the Lagrangian* (12.3.23) *is*

$$H(p_1, p_2, x, y) = \frac{1}{2}[(p_1 + y)^2 + (p_2 - x)^2] + \frac{3}{2}(x^2 + y^2). \tag{12.3.24}$$

Proof. The momenta are $p_1 = \dfrac{\partial L}{\partial \dot{x}} = \dot{x} - y, \quad p_2 = \dfrac{\partial L}{\partial \dot{y}} = \dot{y} + x$, and then

$$\dot{x} = p_1 + y, \quad \dot{y} = p_2 - x. \tag{12.3.25}$$

Using (12.3.25), the Legendre transform yields

$$\begin{aligned} H(p_1, p_2, x, y) &= p_1\dot{x} + p_2\dot{y} - \frac{1}{2}(\dot{x}^2 + \dot{y}^2) - x\dot{y} + \dot{x}y + \frac{3}{2}(x^2 + y^2) \\ &= p_1(p_1 + y) + p_2(p_2 - x) - \frac{1}{2}\Big((p_1 + y)^2 + (p_2 - x)^2\Big) \\ &\quad -x(p_2 - x) + (p_1 + y)y + \frac{3}{2}(x^2 + y^2) \\ &= \frac{1}{2}(p_1^2 + p_2^2) + p_1 y - p_2 x + 2(x^2 + y^2) \\ &= \frac{1}{2}\Big[(p_1 + y)^2 + (p_2 - x)^2\Big] + \frac{3}{2}(x^2 + y^2). \end{aligned}$$

■

12.3.0.3 The first integral of energy

As

$$\frac{\partial H}{\partial t} = 0 \quad \text{and} \quad \frac{\partial H}{\partial t} = \frac{dH}{dt},$$

it follows that H is preserved along the trajectory. The value of H along the trajectory is called *the total energy*. In x, y, $\dot{x}$, $\dot{y}$ coordinates the energy takes the form

$$E = \frac{1}{2}(\dot{x}^2 + \dot{y}^2) + \frac{3}{2}(x^2 + y^2). \tag{12.3.26}$$

Note that E does not depend on the angular momentum as the Lagrangian does. It depends only on the magnitude of the velocity and the distance to the origin.

12.3.0.4 Physical interpretation

The speed of a particle described by a solution of the Euler–Lagrange system is

$$v = \sqrt{\dot{x}^2 + \dot{y}^2}.$$

If $r = \sqrt{x^2 + y^2}$ denotes the distance from the origin to the point (x, y), formula (12.3.26) yields

$$v^2 = 2E - 3r^2.$$

In the case of the astroid with the initial conditions (12.3.26) we have $E = 24$. Thus $v = \sqrt{3(16 - r^2)}$ with $r \in [0, 4]$. The speed on the astroid is zero iff $r = 4$, which occurs only at the cuspidal points.

12.4 The cycloid

Consider a particle described by a Lagrangian, which is the sum of the kinetic, angular momentum and potential energy in the x-direction

$$L(x, y, \dot{x}, \dot{y}) = \frac{1}{2}(\dot{x}^2 + \dot{y}^2) + \frac{1}{2}(x\dot{y} - \dot{x}y) + x. \tag{12.4.27}$$

The Euler–Lagrange system of equations associated with the Lagrangian (12.4.27) is

$$\begin{cases} \ddot{x} - \dot{y} = 1, \\ \ddot{y} + \dot{x} = 0. \end{cases} \tag{12.4.28}$$

If we consider the initial conditions

$$x(0) = 0, \quad \dot{x}(0) = 0, \quad y(0) = 0, \quad \dot{y}(0) = 0$$

+ =)*+,)))
6" ' M $ 6 6 1 H *6 33
-1 * " " H N H *6 33 C %

6 ## # # C# H 6 33 C %
+)*+,))) (6"

)*+,)))-)
. J J: J0H J J J& 6"
' $ %

2 Calculus on surfaces

In this section we discuss basic concepts of differential geometry. We start in Section 2.1 by describing the parametric representation of γ via charts. This classical point of view is critical to introduce the first fundamental form $\mathbf{g}$, the area element q, and the unit normal $\boldsymbol{\nu}$ of γ. We present in Section 2.2 the tangential operators (gradient ∇_γ, divergence div_γ, and Laplace–Beltrami Δ_γ) as well as the Weingarten map; we also discuss H^2-regularity for Δ_γ on surfaces γ of class C^2. We introduce the signed distance function d in Section 2.3 and derive several important properties of it; this intrinsic approach avoids parametrizations and allows for implicit representations of γ. We devote Section 2.4 to the second fundamental form of γ and its principal curvatures using both parametric and intrinsic approaches.

of functions $\widetilde{u} \in H^2(\gamma)$ provided γ is of class C^2. This turns out to be essential for our later error analysis of the trace and narrow band methods for C^2 surfaces. In Section 4 we give a self-contained exposition of parametric FEMs for surfaces of class $C^{1,\alpha}$ and C^2, including a priori and a posteriori error analyses. In Section 5 we describe the trace method and conclude in Section 6 with the narrow band method. Both discussions assume C^2 regularity of γ.

2 Calculus on surfaces

In this section we discuss basic concepts of differential geometry. We start in Section 2.1 by describing the parametric representation of γ via charts. This classical point of view is critical to introduce the first fundamental form $\mathbf{g}$, the area element q, and the unit normal $\boldsymbol{\nu}$ of γ. We present in Section 2.2 the tangential operators (gradient ∇_γ, divergence div_γ, and Laplace–Beltrami Δ_γ) as well as the Weingarten map; we also discuss H^2-regularity for Δ_γ on surfaces γ of class C^2. We introduce the signed distance function d in Section 2.3 and derive several important properties of it; this intrinsic approach avoids parametrizations and allows for implicit representations of γ. We devote Section 2.4 to the second fundamental form of γ and its principal curvatures using both parametric and intrinsic approaches.

2.1 Parametric surfaces

We assume that γ is a closed, compact, orientable manifold of class $C^{1,\alpha}$, $0 < \alpha \leq 1$, and co-dimension 1 in $\mathbb{R}^{n+1}$. It can be represented parametrically by an atlas $\{(\mathcal{V}_i, \mathcal{U}_i, \boldsymbol{\chi}_i)\}_{i\in I}$, where the individual charts $\boldsymbol{\chi}_i : \mathcal{V}_i \to \mathcal{U}_i \cap \gamma \subset \mathbb{R}^{n+1}$ are isomorphisms of class $C^{1,\alpha}$ compatible with the orientation of γ; the open connected sets $\mathcal{V}_i \subset \mathbb{R}^n$ are the parametric domains. Unless stated otherwise, it will be often sufficient to consider a single chart and resort to a partition of the unity. We thus drop the index i for convenience. For $\mathbf{x} \in \mathcal{U} \cap \gamma$, we set $\mathbf{y} := \boldsymbol{\chi}^{-1}(\mathbf{x}) \in \mathcal{V}$.

Let $\partial_j \boldsymbol{\chi}(\mathbf{y})$ be the column vector of jth partial derivatives of $\boldsymbol{\chi}(\mathbf{y})$ for $1 \leq j \leq n$ at $\mathbf{y} \in \mathcal{V}$. By definition, the rank of $D\boldsymbol{\chi}(\mathbf{y}) = (\partial_j \boldsymbol{\chi}(\mathbf{y}))_{j=1}^n \in \mathbb{R}^{(n+1)\times n}$ is n (full rank). This implies that $\{\partial_j \boldsymbol{\chi}(\mathbf{y})\}_{j=1}^n$ are linearly independent and span the tangent hyperplane to γ at $\mathbf{x}$.

The *first fundamental form* is the symmetric and positive definite matrix $\mathbf{g} \in \mathbb{R}^{n\times n}$ defined by

$$\mathbf{g}(\mathbf{y}) := D\boldsymbol{\chi}(\mathbf{y})^t D\boldsymbol{\chi}(\mathbf{y}) \quad \forall \mathbf{y} \in \mathcal{V}. \tag{1}$$

If $\mathbf{g} = (g_{ij})_{i,j=1}^n$, then the components g_{ij} read

$$g_{ij} = \partial_i \boldsymbol{\chi}^t \partial_j \boldsymbol{\chi} = \partial_i \boldsymbol{\chi} \cdot \partial_j \boldsymbol{\chi},$$

which depends on the choice of parametrization. A normal vector $\mathbf{N}(\mathbf{y})$ to γ at $\mathbf{x}$ can be written as $\mathbf{N}(\mathbf{y}) = \sum_{j=1}^{n+1} A_j(\mathbf{y})\mathbf{e}_j$, where $A_j := \det(\mathbf{e}_j, D\boldsymbol{\chi})$ and $\{\mathbf{e}_j\}_{j=1}^{n+1}$ is the canonical basis of $\mathbb{R}^{n+1}$. In fact, since

$$\mathbf{N}\cdot\partial_i\boldsymbol{\chi} = \sum_{j=1}^{n+1}\mathbf{e}_j\cdot\partial_i\boldsymbol{\chi}\det(\mathbf{e}_j, D\boldsymbol{\chi}) = \det\left(\sum_{j=1}^{n+1}(\mathbf{e}_j\cdot\partial_i\boldsymbol{\chi})\mathbf{e}_j, D\boldsymbol{\chi}\right) = \det(\partial_i\boldsymbol{\chi}, D\boldsymbol{\chi}) = 0,$$

and $A_j \neq 0$ for at least one j because $D\boldsymbol{\chi}$ has rank n, we deduce that

$$\boldsymbol{\nu}(\mathbf{y}) := \frac{\mathbf{N}(\mathbf{y})}{|\mathbf{N}(\mathbf{y})|} \quad \forall \mathbf{y}\in\mathcal{V} \tag{2}$$

is a well-defined unit normal vector to γ. Therefore, the matrix

$$\mathbf{T}(\mathbf{y}) := [D\boldsymbol{\chi}(\mathbf{y}), \boldsymbol{\nu}(\mathbf{y})] \in \mathbb{R}^{(n+1)\times(n+1)} \quad \forall \mathbf{y}\in\mathcal{V}$$

has rank $n + 1$ and so is invertible. We write its inverse as

$$\mathbf{T}^{-1} = \begin{bmatrix}\mathbf{B}\\ \mathbf{v}^t\end{bmatrix}, \quad \mathbf{B}\in\mathbb{R}^{n\times(n+1)}, \mathbf{v}\in\mathbb{R}^n,$$

and note that

$$\mathbf{I}_{(n+1)\times(n+1)} = \mathbf{T}^{-1}\mathbf{T} = \begin{bmatrix}\mathbf{B}\,D\boldsymbol{\chi} & \mathbf{B}\mathbf{v}\\ \mathbf{v}^t\ D\boldsymbol{\chi} & \mathbf{v}^t\boldsymbol{\nu},\end{bmatrix}$$

whence

$$\mathbf{B}\,D\boldsymbol{\chi} = \mathbf{I}_{n\times n}, \quad \mathbf{v}^t D\boldsymbol{\chi} = 0, \quad \mathbf{v}^t\,\boldsymbol{\nu} = 1.$$

The last two equalities imply $\mathbf{v} = \boldsymbol{\nu}$. Reversing the order of multiplication yields

$$\mathbf{I}_{(n+1)\times(n+1)} = \mathbf{T}\,\mathbf{T}^{-1} = D\boldsymbol{\chi}\,\mathbf{B} + \boldsymbol{\nu}\boldsymbol{\nu}^t,$$

whence the *projection matrix* $\Pi\in\mathbb{R}^{(n+1)\times(n+1)}$ on the tangent hyperplane to γ has the form

$$\Pi := \mathbf{I} - \boldsymbol{\nu}\otimes\boldsymbol{\nu} = D\boldsymbol{\chi}\,\mathbf{B}. \tag{3}$$

To obtain an explicit expression for $\mathbf{B}$ note that

$$D\boldsymbol{\chi} = (\mathbf{I} - \boldsymbol{\nu}\otimes\boldsymbol{\nu})^t D\boldsymbol{\chi} = \mathbf{B}^t D\boldsymbol{\chi}^t D\boldsymbol{\chi} = \mathbf{B}^t\mathbf{g} \quad\Rightarrow\quad \mathbf{B} = \mathbf{g}^{-1}D\boldsymbol{\chi}^t.$$

This leads to the following useful expression of Π defined in (3):

$$\Pi = D\boldsymbol{\chi}\,\mathbf{g}^{-1}D\boldsymbol{\chi}^t. \tag{4}$$

The *area element* $q(\mathbf{y})$ is the ratio of the infinitesimal volume at $\mathbf{y}\in\mathcal{V}$ and area of γ at $\mathbf{x} = \boldsymbol{\chi}(\mathbf{y})$, namely the volume of the parallelotope in the tangent plane to γ spanned by the vectors $\{\boldsymbol{\chi}_j\}_{j=1}^n$:

$$q(\mathbf{y}) := \det([\boldsymbol{\nu}(\mathbf{y}), D\boldsymbol{\chi}(\mathbf{y})]) \quad \forall \mathbf{y}\in\mathcal{V}. \tag{5}$$

To obtain a more familiar form of q we argue as follows:

$$q=\frac{1}{|N|}\det([\mathbf{N}, D\boldsymbol{\chi}])=\frac{1}{|N|}\det\left([\mathbf{N}, D\boldsymbol{\chi}]^t[\mathbf{N}, D\boldsymbol{\chi}]\right)^{\frac{1}{2}}=\sqrt{\det\mathbf{g}}, \tag{6}$$

because $\det([\mathbf{N}, D\boldsymbol{\chi}]^t[\mathbf{N}, D\boldsymbol{\chi}])=|\mathbf{N}|^2\det(D\boldsymbol{\chi}^tD\boldsymbol{\chi})=|\mathbf{N}|^2\det\mathbf{g}$. Moreover, exploiting that $|\mathbf{N}|^2=\sum_{j=1}^{n+1}A_j\det([\mathbf{e}_j, D\boldsymbol{\chi}])=\det([\mathbf{N}, D\boldsymbol{\chi}])$, we deduce

$$q=|\mathbf{N}|. \tag{7}$$

An integrable function $v:\mathcal{V}\to\mathbb{R}$ induces an integrable function $\widetilde{v}:\gamma\to\mathbb{R}$ by composition $v=\widetilde{v}\circ\boldsymbol{\chi}$, or equivalently $\widetilde{v}(\mathbf{x})=v(\mathbf{y})$ for all $\mathbf{y}\in\mathcal{V}$. The area element allows for integration over γ via the formula

$$\int_\gamma\widetilde{v}=\int_{\mathcal{V}}vq \quad \forall v\in L_1(\mathcal{V}). \tag{8}$$

This definition does not depend on the parametrization: if $\boldsymbol{\chi}_1, \boldsymbol{\chi}_2$ are parametrizations of γ, then $\boldsymbol{\chi}_1=\boldsymbol{\chi}_2\circ(\boldsymbol{\chi}_2^{-1}\circ\boldsymbol{\chi}_1)$ and $D\boldsymbol{\chi}_1=D\boldsymbol{\chi}_2D(\boldsymbol{\chi}_2^{-1}\circ\boldsymbol{\chi}_1)$ whence

$$q_1=\left|\det\left(D(\boldsymbol{\chi}_2^{-1}\circ\boldsymbol{\chi}_1)\right)\right|q_2 \quad\Rightarrow\quad \int_{\mathcal{V}_1}vq_1=\int_{\mathcal{V}_2}vq_2.$$

2.2 Differential operators

If a function $v:\mathcal{V}\to\mathbb{R}$ is of class C^1, we can define the *tangential (or surface) gradient* of the corresponding function $\widetilde{v}:\gamma\to\mathbb{R}$ as a vector tangent to γ that satisfies the chain rule

$$\nabla v(\mathbf{y})=D\boldsymbol{\chi}(\mathbf{y})^t\nabla_\gamma\widetilde{v}(\mathbf{x}) \quad \forall\mathbf{y}\in\mathcal{V}. \tag{9}$$

Since $\nabla_\gamma\widetilde{v}$ is spanned by $\{\partial_j\boldsymbol{\chi}\}_{j=1}^n$, we get $\nabla_\gamma\widetilde{v}=D\boldsymbol{\chi}w$ for some $w\in\mathbb{R}^n$ whence $w=\mathbf{g}^{-1}\nabla v$ and

$$\nabla_\gamma\widetilde{v}(\mathbf{x})=D\boldsymbol{\chi}(\mathbf{y})\mathbf{g}(\mathbf{y})^{-1}\nabla v(\mathbf{y}) \quad \forall\mathbf{y}\in\mathcal{V}. \tag{10}$$

If $\widetilde{\mathbf{v}}=(\widetilde{v}_i)_{i=1}^{n+1}:\gamma\to\mathbb{R}^{n+1}$ is a vector field of class C^1, we define its tangential differential $D_\gamma\widetilde{\mathbf{v}}\in\mathbb{R}^{(n+1)\times(n+1)}$ as a matrix whose ith row is $(\nabla_\gamma\widetilde{v}_i)^t$. If γ is of class C^2, then the unit normal vector $\boldsymbol{\nu}$ is of class C^1 and its differential

$$\mathbf{W}(\mathbf{x})=D_\gamma\boldsymbol{\nu}(\mathbf{x}) \quad \forall\mathbf{x}\in\gamma \tag{11}$$

is called the *Weingarten* (or shape) map of γ. In addition, the *tangential divergence* of $\widetilde{\mathbf{v}}$ is the trace of $D_\gamma\widetilde{\mathbf{v}}$

$$\operatorname{div}_\gamma\widetilde{\mathbf{v}}(\mathbf{x})=\operatorname{trace}\left(D_\gamma\widetilde{\mathbf{v}}(\mathbf{x})\right)=\sum_{i,j=1}^n g^{ij}(\mathbf{y})\,\partial_i\boldsymbol{\chi}(\mathbf{y})\cdot\partial_j\mathbf{v}(\mathbf{y}) \quad \forall\mathbf{y}\in\mathcal{V}, \tag{12}$$

date 29.04.2020
S T [✓ Q] Q S S D

155) 20200429

A) Einstein Math

Books_73_2020/
Einstein Equations Physical and Mathematical Aspects of General Relativity.../
Einstein Equations....pdf

Obs: Livro todo!

B) Books_73_2020/
Mathematical Analysis of the Navier-Stokes Equations.../
3030362256.pdf
Pág. 11 (18)
The Vector Value Setting.

C) Books_73_2020/
Geometric Partial Differential Equations/
→ Diretório inteiro!

C1) Geometric Mechanics on Riemannian manifold...pdf
Pág. 251 (257)

29 04 2020

Chapter 12 Mechanical Curves
(capítulo todo)

C2) Geometric_Partial_Differential_Equations_-_Part_I.pdf

Pág. 07(14)

2. Calculus on surfaces

D) Books_73_2020/
Engineering Mathematics for Semesters/

D1) 9339219643.pdf

Pág. 359

Chapter 7

Vector Differential and Integral Calculus

D2) 9385880500.pdf

Pág. 127 - 256

Chapter 3 Complex Variables and Calculus

Chapter 14 Calculus of Variations (831-900)

obs: Dois livros bons para consultas!

— 202004280040

7

Vector Differential and Integral Calculus

7.1 INTRODUCTION

Vector calculus (or analysis) is a branch of mathematics concerned with differentiation and integration of vector fields. Vector calculus plays an important role in differential geometry and in the study of partial differential equations. It is used extensively in physics and engineering, especially in the description of electromagnetic fields, fluid flow, and gravitational fields. In this chapter, we shall study the vector differential and integral calculus. We often call this study as *vector field theory* or *vector analysis*. We first introduce the some concepts.

(i) Scalar Function

A scalar function $f(x, y, z)$ is a function defined at each point in a certain domain D in space. Its value is real and depends only on the point $P(x, y, z)$ in space, but not on any particular coordinate system being used, for example, the distance function in 3D space which defines the distance between the points $P(x, y, z)$ and $Q(x_0, y_0, z_0)$, then

$$D = f(P) = \sqrt{(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2} \text{ defines a scalar field.}$$

(ii) Vector Function

A function $\vec{F} = \vec{F}(P) = F_1\,\hat{i} + F_2\,\hat{j} + F_3\,\hat{k}$ defined at each point $P \in D$ is called a vector function. We say that a vector field is defined in D. In Cartesian coordinates, we can write

$$\vec{F} = F_1(z, y, z)\,\hat{i} + F_2(z, y, z)\,\hat{j} + F_3(z, y, z)\,\hat{k}$$

If the scalar and vector fields depend on time also then we denote them as $f(P, t)$ and $\vec{F}\,(P, t)$, respectively. Both the fields are independent of the choice of the coordinate systems.

(iii) Level Surfaces

If $f(x, y, z)$ be a single-valued continuous scalar function defined at every point $P(x, y, z)$ in the domain D. Then $f(x, y, z) = c$, a constant defines the equation of a surface and is called *a level surface of the function*. For the different values of C, we obtain the different surfaces, no two of which intersect.

7.2 PARAMETRIC REPRESENTATION OF VECTOR FUNCTIONS

The two-dimensional curve C can be parametrized by $x = x(t)$, $y = y(t)$, $a \le t \le b$. Then, the position vector of a point P on the curve C can be written as follows:

$$\vec{r}(t) = x(t)\hat{i} + y(t)\hat{j}$$

Similarly, a three-dimensional curve can be parametrized as

$$\vec{r}(t) = x(t)\hat{i} + y(t)\hat{j} + z(t)\hat{k}; \quad a \le t \le b$$

Parametric Forms of the Standard Curves

(i) Straight Line $\vec{r}(t) = \vec{a} + \vec{b}t$

The parametric form is

$$\vec{r}(t) = t\,\hat{i} + (1-t)\,\hat{j} \qquad 0 \le t \le 1$$

(ii) Circle $x^2 + y^2 = a^2$, the parametric form of the circle is

$$\vec{r}(t) = a\cos t\,\hat{i} + a\sin t\,\hat{j}; \qquad 0 \le t \le 2\pi$$

(iii) Parabola Consider the parabola $y^2 = 4ax$.

Then we assume $y = t$ as a parameter and the parametric form of the parabola as

$$\vec{r}(t) = \left(\frac{t^2}{4a}\right)\hat{i} + t\,\hat{j}; \qquad -\infty < t < \infty$$

(iv) Ellipse The parametric form of the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ can be written as

$$\vec{r}(t) = a\cos t\,\hat{i} + b\sin t\,\hat{j}; \qquad 0 \le t \le 2\pi$$

7.3 LIMIT, CONTINUITY, AND DIFFERENTIABILITY OF A VECTOR FUNCTION

A vector function $\vec{F}(t)$ is said to a limit 'l', when t tends to t_0, if for any given positive number $\in$, however small, there corresponds a positive number δ such that

$$|\vec{F}(t) - l| < \varepsilon, \text{ whenever } |t - t_0| < \delta$$

If $\vec{F}(t)$ tends to a limit 'l' as t tends to t_0,

$$\therefore \qquad \lim_{t \to t_0} \vec{F}(t) = l$$

A vector point function $\vec{F}(t)$ is said to be continuous for a value t_0 of t if $\vec{F}(t_0)$ is defined and for any given positive number ε, however small, there exists a positive number δ such that

$$|\vec{F}(t) - \vec{F}(t_0)| < \varepsilon, \text{ whenever } |t - t_0| < \delta.$$

A vector function $\vec{F}(t)$ is said to be differentiable at $t = t_0$, if the limit $\lim_{t_0 \to 0} \frac{\overline{F}(t+t_0) - \overline{F}(t)}{t_0}$ exists.

3

Complex Variables and Calculus

3.1 COMPLEX NUMBER

An ordered pair of real numbers x and y to be written as $z = (x, y)$ is called a complex number. Also, we may write $z = (x, y) = x + iy$, where imaginary unit i(iota) is defined as $i = (0, 1)$. Here x is called real part of z and y is called imaginary part of z. The real part of z is denoted by Re(z) and the imaginary part of z by Im(z).

Set of Complex Numbers

The set of all complex numbers is denoted by C, i.e.,

$$C = \{x + iy \mid x, y \in R\}.$$

Since a real number 'x' can be written as $x + oi$,

$\therefore$ Every real number is a complex number. Hence $R \subset C$, where R is the set of all real numbers.

3.2 EQUALITY OF COMPLEX NUMBERS

Two complex numbers $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$ are equal if $x_1 = x_2$ and $y_1 = y_2$, i.e. $\text{Re}(z_1) = \text{Re}(z_2)$ and $\text{Im}(z_1) = \text{Im}(z_2)$.

3.3 FUNDAMENTAL OPERATIONS WITH COMPLEX NUMBERS

Let $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$ be two complex numbers, then

(i) Addition of Complex Numbers

Addition of two complex numbers z_1 and z_2 is denoted by $z_1 + z_2$ and is defined as

$$z_1 + z_2 = (x_1 + x_2) + i(y_1 + y_2)$$

i.e., $\text{Re}(z_1 + z_2) = \text{Re}(z_1) + \text{Re}(z_2)$

and $\text{Im}(z_1 + z_2) = \text{Im}(z_1) + \text{Im}(z_2)$.

Properties of Addition of Complex Numbers

(a) Addition is commutative for any two complex numbers z_1 and z_2, we have

$$z_1 + z_2 = z_2 + z_1.$$

(b) *Addition is associative* For any three complex numbers z_1, z_2, z_3, we have

$$(z_1 + z_2) + z_3 = z_1 + (z_2 + z_3).$$

(c) *Existence of additive identity* The complex number $0 = 0 + 0i$ is an identity element for addition, i.e. $z + 0 = z = 0 + z$ for al $z \in C$.

(d) *Existence of additive inverse* For every non-zero complex number $z = x + iy$ and there exists an additive inverse $z = -(x + iy)$, then

$$\begin{aligned} z + (-z) &= (x + iy) - (x + iy) \\ &= 0 + 0i \\ &= 0 \end{aligned}$$

(ii) Subtraction of Complex Numbers

The subtraction of z_2 from z_1 is denoted by $z_1 - z_2$ and is defined as the addition of z_1 and z_2.
Thus, $z_1 +(-z_2) = z_1 - z_2 = (x_1 + iy_1) - (x_2 + iy_2)$

$$= (x_1 - x_2) + i(y_1 - y_2).$$

(iii) Multiplication of Complex Numbers

The multiplication of z_1 with z_2 is denoted by z_1z_2 and is defined as the complex number

$$(x_1x_2 - y_1y_2) + i(x_1y_2 + x_2y_1).$$

Thus,

$$\begin{aligned} z_1z_2 &= (x_1 + i\, y_1) \cdot (x_2 + \mathrm{i}y_2) \\ &= (x_1x_2 - y_1y_2) + i(x_1y_2 + x_2y_1). \end{aligned}$$

$\Rightarrow$

$$z_1z_2 = [\mathrm{Re}(z_1)\,\mathrm{Re}(z_2) - \mathrm{Im}(z_1)\,\mathrm{Im}(z_2)] + i\,[\mathrm{Re}(z_1)\,\mathrm{Im}(z_2) + \mathrm{Re}(z_2)\,\mathrm{Im}(z_1)]$$

Properties of Multiplication

(a) *Multiplication is commutative* For any two complex numbers z_1 and z_2, we have

$$z_1z_2 = z_2z_1.$$

(b) *Multiplication is associative* For any three complex numbers z_1, z_2, z_3, we have

$$(z_1\, z_2)\, z_3 = z_1(z_2\, z_3).$$

(c) *Existence of identity element for multiplication* The complex number $1 = 1 + oi$ is the identity element for multiplication, i.e. for every complex number z, we have
$z \cdot 1 = z = 1 \cdot z$.

(d) *Existence of multiplicative inverse* For every non-zero complex number $z = x + iy$, there exists a complex number $z_1 = x_1 + iy_1$. Such that

$$z\, z_1 = 1 = z_1\, z.$$

(e) *Multiplication of complex numbers is distributive over addition of complex numbers* For any three complex numbers z_1, z_2, z_3, we have
(i) $z_1(z_2 + z_3) = z_1z_2 + z_1z_3$ (Left distributivity)
(ii) $(z_2 + z_3)z_1 = z_2z_1 + z_3z_1$ (Right distributivity)

3.4 DIVISION OF COMPLEX NUMBERS

The division of a complex number z_1 by a non-zero complex number z_2 is defined as the multiplication of z_1 by the multiplicative inverse of z_2 and is denoted by $\dfrac{z_1}{z_2} = z_1 \cdot (z_2)^{-1} = z_1 \cdot \left(\dfrac{1}{z_2}\right)$.

14

Calculus of Variations

14.1 INTRODUCTION

Calculus of variations plays an important role in modern developments in analysis, geometry and physics. Originating as a study of certain maximum and minimum problems not treatable by the methods of elementary calculus, variational calculus in its present form provides powerful methods for the treatment of differential equations, theory of invariants, existence theorems in geometric function theory, variational principles in mechanics. It is also important to the applications to boundary value problems in partial differential equations and in the numerical calculation of many types of problems which can be stated in variational form. No literature representing these diverging viewpoints is to be found among standard texts on calculus of variations, and in this course an attempt will be made to do justice to this variety of problems. The subject matter with which calculus of variations is concerned is a class of extremum (i.e., maximum or minimum) problems which can be considered as an extension of the familiar class of extremum problems dealt with by elementary differential calculus. In the elementary problems, one seeks extremal values of a function of one or more (but in any case a finite number) real variables in the more general problems considered by calculus of variations, the functions to be extremized, sometimes called functionals, have functions as independent variables. In this chapter, we will study various method of variational with fixed boundaries

14.2 FUNCTION

Let X and Y be any two non-empty sets and there be a correspondence or association between the elements of X and Y such that for every elements of X and Y, there exists a unique element $y \in Y$ written as $y = f(x)$ then we say that y is a mapping or function from X to Y and written as $f: X \rightarrow Y$ such that $y = f(x)\ \forall\ x \in X, y \in X$.

14.3 FUNCTIONAL

Let there be functional belonging to class of functions. Then a variable quantity denoted by $I[y(x)]$ is a functional if to each function belonging to the class of functions there is definite value of I.

Thus functional is a function between the set of functions and set variable quantity. Obviously, domain of the functional is a set of functions for which the functional have been defined as following:

(i) $\int_{x_1}^{x_2} \sqrt{\left[1+\left(\frac{dy}{dx}\right)^2\right]}\,dx$ determines the length of arc between the points (x_1, y_1) and (x_2, y_2) on the curve $y = f(x)$. Length of the arc is determined by the choice of functions. Thus, we see corresponding to functions through (x_1, y_1) and (x_2, y_2) determine the length of arc between these points. Therefore it is a clear example of functional. Thus we note $I\left[y(x)\right]=\int_{x_1}^{x_2}\sqrt{\left[1+\left(\frac{dy}{dx}\right)^2\right]}dx.$ We can also write it as $I\left[y(x)\right]=\int_{x_1}^{x_2} f\left[x, y(x), y'(x)\right]dx.$

(ii) Area S of a surface bounded by a given curve C is determined by the surface $z = z(x, y)$ and given by $\iint_D \sqrt{\left(1+\left(\frac{\partial z}{\partial x}\right)^2+\left(\frac{\partial z}{\partial y}\right)^2\right)}dxdy$ where D is the projection of area bounded by curve on xy-plane.

14.4 DIFFERENCE BETWEEN FUNCTION AND FUNCTIONAL

S.No	Function	Functional
1.	The variable y is called a function of a variable x, in writing as $y = y(x)$.	The variable I is called a functional depending on a function $y = y(x)$, in writing as $I = I[y(x)]$.
2.	If to each value of x from a certain domain there corresponds a certain value of y, to a given number x, there corresponds a number y.	If to each function $y(x)$, from a certain class of function there corresponds a certain value of I, to a given function $y(x)$, there corresponds a number I.
3.	The increment Δx of the argument x of a function $y(x)$ is the difference of two values of this argument $\Delta x = x - x_1$. If x is the independent variable, then the differential of x coincides with its increment $dx = \Delta x$.	The increment or the variation δy of the argument $y(x)$ of a functional $I[y(x)]$ is the difference of two functions $\delta y = y(x) - y_1(x)$. It is assumed that the argument is $y(x)$ runs through a certain class of functions.
4.	A function $y(x)$ is said to be a continuous function, if small variations of x always lead to small variations of the functions $y(x)$. 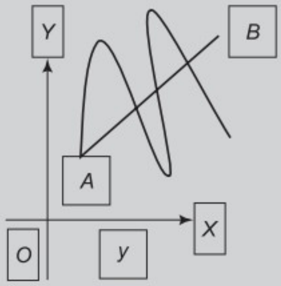	A functional $I[y(x)]$ is said to be a continuous function, if small variations of $y(x)$ always lead to small variations of the functional $I[y(x)]$. 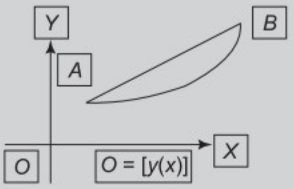

5.	The function $l(x)$ is called linear if $l(cx) = cl(x)$.	The functional $L(x)$ is called linear, if $L(cx) = cL(x)$.
6.	If the increment $\Delta y = y = y(x + \Delta x) - y(x)$ is of the form $\Delta y = A(x)\Delta x + \beta(x, \Delta x).\Delta x$.	If the increment $\Delta I = I\left[y(x)+\delta y)\right]-I\left[y(x)\right]$ of a functional is of the form $\Delta I = L\left[y(x),\delta x\right]+\beta(y(x),\delta y).\max\left\|\delta y\right\|$.
7.	The differential of a function $y(x)$ is given by $\left(\frac{\partial}{\partial\alpha} y(x+\alpha.\Delta x)\right)_{\alpha=0}$.	The variation of a functional $I[y(x)]$ is given by $\left(\frac{\partial}{\partial\alpha} I\left[y(x+\alpha.\delta y)\right]\right)_{\alpha=0}$.

14.5 CLOSENESS OF CURVES

In the calculus of real variable we have defined the continuity of function at $x = a$. As the variable x is becoming closer to a, then $f(x)$ becomes closer to $f(a)$.

Here in case of functional, domain is the class of functions and here idea of closeness of $y(x)$ and $y_1(x)$ is to say that the absolute value of their difference, i.e. $|y(x) - y_1(x)|$ is small for all x for which $y(x)$ and $y_1(x)$ are defined.

If it happens then we say $y(x)$ is close to $y_1(x)$ in the sense of zero order proximity. $I\left[y(x)\right]=\int_{x_1}^{x_2} f\left[x, y(x), y'(x)\right]dx$. Curve $y(x)$ and $y_1(x)$ are said to be close in the sense of n^{th} order proximity if $|y(x) - y_1(x)|$, $\left|y'(x)-y_1{}'(x)\right|\cdots\left|y^n(x)-y_1^n(x)\right|$ are small for the values of for x which these functions are defined.

14.6 CONTINUITY OF FUNCTIONAL

Domain of functional is the set of functions, therefore continuity of functional at some function is defined as following: Functional $I[y(x)]$ is said to be continuous at $y = y_0(x)$ in the sense of $n^{\underline{\text{th}}}$ order proximity if given for any positive number ε, $\exists\ \delta > 0$ such that $\left|I\left[y(x)\right]-I\left[y_0(x)\right]\right| < \varepsilon$

for $\left|y(x)-y_0(x)\right| < \delta, \left|y'(x)-y_0{}'(x)\right| < \delta \cdots \left|y^n(x)-y_0^n(x)\right| < \delta$

Linear Property Let the functional $I[y(x)]$ be defined in the linear space M of the functions $y(x)$. Functional $I[y(x)]$ is said to be linear if it satisfies the following conditions

(a) $I\left[c\ y(x)\right] = c\ I\left[y(x)\right]$ where c is any arbitrary constant.

(b) $I\left[\ y_1(x)+y_2(x)\right]=I\left[y_1(x)\right]+I\left[y_2(x)\right]$ where $y_1(x)$ and $y_2(x) \in M$

Example 1 Functional $I\left[y(x)\right]=\int_a^b\left(y'+2y\right)dx$ defined in the space which consists of the continuous function possessing continuous derivatives of first order.

29 04 2020

Chapter 12 Mechanical Curves
(capitulo todo)

C2) Geometric_Partial_Differential_Equations_-_Part_I.pdf

Pág. 07(14)

2. Calculus on surfaces

D) Books_73_2020/
Engineering Mathematics for Semesters/

D1) 9339219643.pdf

Pág. 359

Chapter 7

Vector Differential and Integral Calculus

D2) 9385880500.pdf

Pág. 127 – 256

Chapter 3 Complex Variables and Calculus

Chapter 14 Calculus of Variations (831-900)

obs: Dois livros bons para consultas!

— 202004280040

& C !# # D %
)*+,)))-) C
&" J J, J /J? J 6"
' % $! ! 8 ; 1 / 6 6

> The indeterminism appears however only in the physical space, while in the space of all states (Hilbert space) everything is perfectly deterministic. The wave function evolves in a deterministic way according to the time-dependent Schrödinger equation (1.10).

The puzzling way in which indeterminism operates will be shown below.

1.3 The Heisenberg uncertainty principle

Consider two mechanical quantities A and B, for which the corresponding Hermitian operators (constructed according to Postulate II), $\hat{A}$ and $\hat{B}$, give the commutator $\left[\hat{A}, \hat{B}\right] = \hat{A}\hat{B} - \hat{B}\hat{A} = i\hat{C}$, where $\hat{C}$ is a Hermitian operator.[52] This is what happens for example for $A = x$ and $B = p_x$. Indeed, for any differentiable function ϕ one has $[\hat{x}, \hat{p_x}]\phi = -xi\hbar\phi' + i\hbar(x\phi)' = i\hbar\phi$, and therefore the operator $\hat{C}$ in this case means simply multiplication by $\hbar$.

> From axioms of quantum mechanics one can prove that a product of errors (in the sense of standard deviation) of measurements of two mechanical quantities is greater than or equal to $\frac{1}{2}\left\langle[\hat{A}, \hat{B}]\right\rangle$, where $\left\langle[\hat{A}, \hat{B}]\right\rangle$ is the mean value of the commutator $[\hat{A}, \hat{B}]$.

This is known as the Heisenberg uncertainty principle.

Werner Heisenberg did not provide any formal proof; instead he carried out a *Gedankenexperiment* (an ideal thought experiment) with an electron interacting with an electromagnetic wave ("Heisenberg's microscope").

The formal proof goes as follows.

---

[52] This is guaranteed. Indeed, $\hat{C} = -i\left[\hat{A}, \hat{B}\right]$ and then the Hermitian character of $\hat{C}$ is shown by the following chain of transformations: $\left\langle f|\hat{C}g\right\rangle = -i\left\langle f|\left[\hat{A}, \hat{B}\right]g\right\rangle = -i\left\langle f|(\hat{A}\hat{B} - \hat{B}\hat{A})g\right\rangle = -i\left\langle(\hat{B}\hat{A} - \hat{A}\hat{B})f|g\right\rangle = \left\langle -i(\hat{A}\hat{B} - \hat{B}\hat{A})f|g\right\rangle = \left\langle\hat{C}f|g\right\rangle$.

Werner Karl Heisenberg (1901–1976) was born in Würzburg (Germany), attended high school in Munich, and then (with his friend Wolfgang Pauli) studied physics at Munich University under Sommerfeld's supervision. In 1923 he defended his doctoral thesis on turbulence in liquids. Reportedly, during the doctoral examination he had problems writing down the chemical reaction in lead batteries. He joined the laboratory of Max Born at Göttingen (following his friend Wolfgang) and in 1924 the Institute of Theoretical Physics in Copenhagen, working under the supervision of Niels Bohr. A lecture delivered by Niels Bohr decided the future direction of his work. Heisenberg later wrote: "*I was taught optimism by Sommerfeld, mathematics in Göttingen, physics by Bohr*." In 1925 (only a year after being convinced by Bohr) Heisenberg developed a formalism, which became the first successful quantum theory. Then, in 1926 Heisenberg, Born, and Jordan elaborated the formalism, which resulted in a coherent theory ("matrix mechanics"). In 1927 Heisenberg obtained a chair at Leipzig University, which he held until 1941 (when he became director of the Kaiser Wilhelm Physics Institute in Berlin). Heisenberg received the Nobel Prize in 1932 "*for the creation of quantum mechanics, the application of which has, inter alia, led to the discovery of the allotropic forms of hydrogen*." In 1937 Werner Heisenberg was at the height of his powers. He was nominated professor and got married. However, just after returning from his honeymoon, the rector of the university called him, saying that there was a problem. In the SS weekly, an article by Professor Johannes Stark (a Nobel Prize winner and faithful Nazi) was about to appear claiming that Professor Heisenberg is not such a good patriot as he pretends, because he socialized in the past with Jewish physicists. Soon Professor Heisenberg was invited to the SS headquarters at Prinz Albert Strasse in Berlin. The interrogation took place in the basement. On the raw concrete wall the scoffing slogan "*Breathe deeply and quietly*" was written. One of the questioners was a PhD student from Leipzig, who had once been examined by Heisenberg. The terrified Heisenberg told his mother about the problem. She recalled that in her youth she had made the acquaintance of Heinrich Himmler's mother. Frau Heisenberg paid a visit to Frau Himmler and asked her to pass a letter from her son to Himmler. At the beginning Himmler's mother tried to separate her maternal feelings for her beloved son from politics. She was finally convinced after Frau Heisenberg said "*we mothers should care about our boys*." After a certain time, Heisenberg received a letter from Himmler saying that his letter "coming through unusual channels" has been examined especially carefully. He promised to stop the attack. In the *post scriptum* there was a precisely tailored phrase: "*I think it best for your future, if for the benefit of your students, you would carefully separate scientific achievements from the personal and political beliefs of those who carried them out. Your faithfully, Heinrich Himmler*" (after D. Bodanis, "$E = mc^2$," Fakty, Warsaw, 2001, p. 130) ©The Nobel Foundation.

Recall the definition of the variance, or the square of the standard deviation $(\Delta A)^2$, of measurements of the quantity A:

$$(\Delta A)^2 = \langle \hat{A}^2 \rangle - \langle \hat{A} \rangle^2, \tag{1.22}$$

where $\langle X \rangle$ means the mean value of many measurements of the quantity X. The *standard deviation* ΔA *represents the width of the distribution of* A, i.e., *the error made.* Eq. (1.22) is equivalent to[53]

$$(\Delta A)^2 = \langle (\hat{A} - \langle \hat{A} \rangle)^2 \rangle. \tag{1.23}$$

Consider the product of the standard deviations for the operators $\hat{A}$ and $\hat{B}$, taking into account that $\langle \hat{u} \rangle$ denotes (Postulate IV) the integral $\langle \Psi | \hat{u} | \Psi \rangle$ according to (1.20). One obtains (denoting $\hat{\mathcal{A}} = \hat{A} - \langle \hat{A} \rangle$ and $\hat{\mathcal{B}} = \hat{B} - \langle \hat{B} \rangle$; of course, $[\hat{\mathcal{A}}, \hat{\mathcal{B}}] = [\hat{A}, \hat{B}]$)

$$(\Delta A)^2 \cdot (\Delta B)^2 = \langle \Psi | \hat{\mathcal{A}}^2 \Psi \rangle \langle \Psi | \hat{\mathcal{B}}^2 \Psi \rangle = \langle \hat{\mathcal{A}} \Psi | \hat{\mathcal{A}} \Psi \rangle \langle \hat{\mathcal{B}} \Psi | \hat{\mathcal{B}} \Psi \rangle,$$

where the Hermitian character of the operators $\hat{\mathcal{A}}$ and $\hat{\mathcal{B}}$ is used. Now, let us use the Schwarz inequality (Appendix B) $\langle f_1 | f_1 \rangle \langle f_2 | f_2 \rangle \geq |\langle f_1 | f_2 \rangle|^2$:

$$(\Delta A)^2 \cdot (\Delta B)^2 = \langle \hat{\mathcal{A}} \Psi | \hat{\mathcal{A}} \Psi \rangle \langle \hat{\mathcal{B}} \Psi | \hat{\mathcal{B}} \Psi \rangle \geq |\langle \hat{\mathcal{A}} \Psi | \hat{\mathcal{B}} \Psi \rangle|^2.$$

Next,

$$\begin{aligned} \langle \hat{\mathcal{A}} \Psi | \hat{\mathcal{B}} \Psi \rangle &= \langle \Psi | \hat{\mathcal{A}} \hat{\mathcal{B}} \Psi \rangle = \langle \Psi | \{[\hat{\mathcal{A}}, \hat{\mathcal{B}}] + \hat{\mathcal{B}} \hat{\mathcal{A}}\} \Psi \rangle = i \langle \Psi | \hat{C} \Psi \rangle + \langle \Psi | \hat{\mathcal{B}} \hat{\mathcal{A}} \Psi \rangle \\ &= i \langle \Psi | \hat{C} \Psi \rangle + \langle \hat{\mathcal{B}} \Psi | \hat{\mathcal{A}} \Psi \rangle = i \langle \Psi | \hat{C} \Psi \rangle + \langle \hat{\mathcal{A}} \Psi | \hat{\mathcal{B}} \Psi \rangle^*. \end{aligned}$$

Hence,

$$i \langle \Psi | \hat{C} \Psi \rangle = 2i \ \mathrm{Im}\{\langle \hat{\mathcal{A}} \Psi | \hat{\mathcal{B}} \Psi \rangle\}.$$

This means that $\mathrm{Im}\{\langle \hat{\mathcal{A}} \Psi | \hat{\mathcal{B}} \Psi \rangle\} = \frac{\langle \Psi | \hat{C} \Psi \rangle}{2}$, which gives $|\langle \hat{\mathcal{A}} \Psi | \hat{\mathcal{B}} \Psi \rangle| \geq \frac{\left|\langle \Psi | \hat{C} \Psi \rangle\right|}{2}$. Hence,

$$(\Delta A)^2 \cdot (\Delta B)^2 \geq |\langle \hat{\mathcal{A}} \Psi | \hat{\mathcal{B}} \Psi \rangle|^2 \geq \frac{|\langle \Psi | \hat{C} \Psi \rangle|^2}{4} \tag{1.24}$$

or, taking into account that $|\langle \Psi | \hat{C} \Psi \rangle| = |\langle \Psi | [\hat{A}, \hat{B}] \Psi \rangle|$, we have[54]

$$\Delta A \cdot \Delta B \geq \frac{1}{2} |\langle \Psi | [\hat{A}, \hat{B}] \Psi \rangle|. \tag{1.25}$$

---

[53] Because $\langle (\hat{A} - \langle \hat{A} \rangle)^2 \rangle = \langle \hat{A}^2 - 2\hat{A}\langle \hat{A} \rangle + \langle \hat{A} \rangle^2 \rangle = \langle \hat{A}^2 \rangle - 2\langle \hat{A} \rangle^2 + \langle \hat{A} \rangle^2 = \langle \hat{A}^2 \rangle - \langle \hat{A} \rangle^2$.

[54] This particular result was used later by Mandelshtam and Tamm to derive the energy-time uncertainty – an analogy to the Heisenberg uncertainty principle; see p. 100.

There are two important special cases.

(a) $\hat{C} = 0$, i.e., the operators $\hat{A}$ and $\hat{B}$ commute. We have $\Delta A \cdot \Delta B \geq 0$, i.e., *the errors can be arbitrarily small*. Both quantities therefore *can be* measured simultaneously without error.

(b) $\hat{C} = \hbar$, as in the case of $\hat{x}$ and $\hat{p_x}$. Then, $(\Delta A) \cdot (\Delta B) \geq \frac{\hbar}{2}$.

In particular, for $\hat{A} = \hat{x}$ and $\hat{B} = \hat{p}_x$, if quantum mechanics is valid, *one cannot determine the exact position and the exact momentum of a particle*. The sense of this statement is the following: even if it *may happen* that we get as a result of a measurement the exact position and the exact momentum of the particle, by repeating many times the measurement (in order to be sure about the error made) we are unable to get zero standard deviation (error) for position and energy measurements. When the precision with which x is measured increases, the particle's momentum has so wide a distribution that the error of determining p_x is huge.[55] Fig. 1.13 shows why the Heisenberg principle does not represent any mystery, but is a natural consequence of the wave nature of particles. Also, the principle does not rely on any putative perturbation of position or momentum of the particle during measurement.

The power of the Heisenberg uncertainty principle is seen when it is used for estimation, just from scratch, of the size of some systems.

Example 1 (Size of the hydrogen atom). How on earth this can be estimated from virtually no information? Let us see.

We assume that the electron moves, while the nucleus does not.[56] Whatever the electron does in the hydrogen atom, it has to conform to the uncertainty principle $\Delta x \cdot \Delta p_x \geq \frac{\hbar}{2}$, i.e., in the most compact state we may expect the *equality* $\Delta x \cdot \Delta p_x = \frac{\hbar}{2}$. We may estimate Δx as the radius r of the atom, while $\Delta p_x = \sqrt{\langle p_x^2 \rangle - \langle p_x \rangle^2} = \sqrt{\langle p_x^2 \rangle - 0} = \sqrt{\langle p_x^2 \rangle}$. Therefore, we have an estimation $r \cdot \sqrt{\langle p_x^2 \rangle} = \frac{\hbar}{2}$, or $\sqrt{\langle p_x^2 \rangle} = \frac{\hbar}{2r}$. The total energy may be estimated as the sum of the kinetic and potential energies: $E = \frac{\langle p^2 \rangle}{2m} - \frac{e^2}{r} = \frac{\langle p_x^2 + p_y^2 + p_z^2 \rangle}{2m} - \frac{e^2}{r} = \frac{3\hbar^2}{8mr^2} - \frac{e^2}{r}$. Now, let us find the minimum of $E(r)$ as its probable value: $\frac{\mathrm{d}E}{\mathrm{d}r} = 0 = -2\frac{3\hbar^2}{8mr^3} + \frac{e^2}{r^2}$, or $e^2 = \frac{3}{4}\frac{\hbar^2}{mr}$.

---

[55] There is an apocryphal story about a *Polizei* patrol stopping Professor Heisenberg for speeding. The very serious man asks: "*Do you know how fast you were going when I stopped you?*" Professor Heisenberg answered: "*I have absolutely no idea, Herr Oberleutnant, but I can tell you precisely where you stopped me.*"

[56] Note that by the momentum conservation law the nucleus moves 1840 times slower than the electron. This practically means the electron moves in the electric field of the immobilized nucleus.

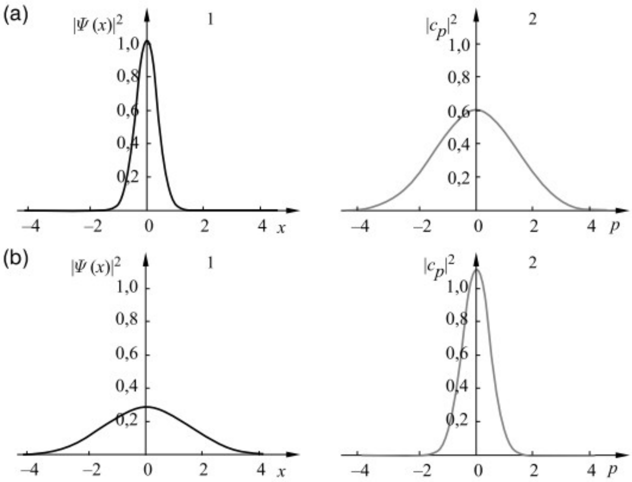

Fig. 1.13. Illustration of the Heisenberg uncertainty principle in case of a single particle moving on the x axis. (a1) $|\Psi(x)|^2$ (i.e., the probability density distribution of finding the particle) as function of coordinate x. The width of this distribution is related to the expected error of determining x. The curve is narrow, which means the error in determining the particle's position is expected to be small. The wave function $\Psi(x)$ can be expanded in the infinite series $\Psi(x) = \sum_p c_p \exp(ipx)$, where p denotes the momentum. Note that each individual function $\exp(ipx)$ is an eigenfunction of momentum, and therefore, if $\Psi(x) = \exp(ipx)$, a measurement of momentum would give exactly p. If, however, $\Psi(x) = \sum_p c_p \exp(ipx)$, then such a measurement yields a given p with the probability $|c_p|^2$. (a2) $|c_p|^2$ as function of p. As one can see a broad range of p (large uncertainty of momentum) ensures a sharp $|\Psi(x)|^2$ distribution (small uncertainty of position in (a1)). Simply the waves $\exp(ipx)$ to obtain a sharp peak of $\Psi(x)$ should exhibit a perfect constructive interference in a small region and a perfect destructive interference elsewhere. This requires a lot of different p's, i.e., a broad momentum distribution. This means that a small error in determining the particle's position is necessarily associated with a large error in determining the particle's momentum (Heisenberg uncertainty principle). (b1, b2) The same is shown, but this time a narrow p distribution gives a broad x distribution. The Heisenberg uncertainty principle has nothing mysterious in it and does not represent a unique feature of quantum mechanics. If one wanted to construct from the ocean waves (of various directions and wave lengths) a tall water pole in the middle of the still ocean, one would be forced to use many wave lengths to get the constructive interference in one spot and destructive interference in all the others.

Hence, we have an estimation, $r = \frac{3}{4}\frac{\hbar^2}{me^2} = 0.75 \cdot a_0$, where $a_0 = 0.529$ Å (as will be shown in Chapter 4, Eq. (4.41), p. 232) is known as the "Bohr first orbit radius of the hydrogen atom." Thus, just from the Heisenberg uncertainty principle we got a value, which has the correct order of magnitude!

We have an important result.

The Heisenberg uncertainty principle prevents atoms from collapsing (the electron falling down to the nucleus, $x = 0$, $\Delta x = 0$, its energy would be equal to $-\infty$), which would mean a catastrophe in the Universe (emission of an infinite amount of energy). The principle also gives a correct order of magnitude for the density of matter.

Example 2 (Size of nucleus). How large is an atomic nucleus? Well, again we may estimate its size from the Heisenberg uncertainty principle knowing only one thing: that the binding energy per nucleon is of the order of 8 MeV. We are now in the realm of nuclear forces acting among nucleons. It seems therefore our task is extremely difficult. We just ignore all these extremely complex forces, considering instead *a single nucleon moving in a mean potential, bound with the energy* $E = 8$ MeV, *and therefore its kinetic energy has to be of the same order of magnitude*. We focus on this kinetic energy[57] of the nucleon assuming simply that $E = \frac{\langle p^2\rangle}{2m_N} = \frac{\langle(\Delta p)^2\rangle}{2m_N}$. From this we calculate $\sqrt{\langle(\Delta p)^2\rangle} = \Delta p = \sqrt{2m_N E}$ and, using the Heisenberg uncertainty principle, we have the uncertainty of the nucleon's position as $a = \frac{\hbar}{2\Delta p} = \frac{\hbar}{2\sqrt{2m_N E}}$. Calculating this in atomic units ($\hbar = 1$, $e = 1$, $m = 1$, where m is the electron mass) we get $a = \frac{\hbar}{2\sqrt{2m_N E}} = \frac{1}{2\sqrt{2\cdot 1840\cdot 8\cdot 10^6\cdot 3.67516\cdot 10^{-2}}} = 0.0000152$ a.u. $\approx 10^{-13}$ cm, 100,000 times smaller than the hydrogen atom. Experiments confirm our number: this is indeed the order of magnitude of the sizes of atomic nuclei!

1.4 The Copenhagen interpretation of the world[58]

The picture of the world that emerged from quantum mechanics was "diffuse" with respect to classical mechanics. In classical mechanics one could measure a particle's position and momentum with a desired accuracy,[59] whereas the Heisenberg uncertainty principle states that this is simply *impossible*.

Bohr presented a philosophical interpretation of the world, which at its foundation had in a sense a *nonreality of the world*.

---

[57] We are interested in the order of its magnitude only.

[58] Schrödinger did not like the Copenhagen interpretation. Once, Bohr and Heisenberg invited him for a Baltic Sea cruise and tried to indoctrinate him so strongly that Schrödinger became ill and stopped participating in their discussions.

[59] This is an exaggeration. Classical mechanics also has its own problems with uncertainty. For example, obtaining the same results for a game of dice would require a perfect reproduction of the initial conditions, which is never feasible.

According to Bohr, before a measurement on a particle is made *nothing* can be said about the value of a given mechanical quantity, unless the wave function represents an eigenfunction of the operator of this mechanical quantity. Moreover, except this case, the particle *does not have* any fixed value of mechanical quantity at all.

A measurement gives a value of the mechanical property (A). Then, according to Bohr, after the measurement is completed, the state of the system changes (the so-called *wave function collapse* or *decoherence*) to the state described by *an eigenfunction of the corresponding operator* $\hat{A}$, and as the *measured value one obtains the eigenvalue corresponding to the wave function*. According to Bohr, there is no way to foresee which eigenvalue one will get as the result of the measurement. However, one can calculate the *probability* of getting a particular eigenvalue. This probability may be computed as the square of the overlap integral (cf. p. 29) of the initial wave function and the eigenfunction of $\hat{A}$.

1.5 How to disprove the Heisenberg principle? The Einstein–Podolsky–Rosen's recipe

The Heisenberg uncertainty principle came as a shock. Many scientists felt a strong imperative to prove that the principle is false. One of them was Albert Einstein, who used to play with ideas by performing some (as he used to say) ideal thought experiments (in German *Gedankenexperiment*) in order to demonstrate internal contradictions in theories. Einstein believed in the reality of our world. With his colleagues Podolsky and Rosen ("EPR team") he designed a special *Gedankenexperiment*.[60] It represented an attempt to disprove the Heisenberg uncertainty principle and to show that one *can* measure the position and momentum of a particle *without any error*. To achieve this, the gentlemen invoked the help of a second particle.

The key statement of the whole reasoning, given in the EPR paper, was the following: "*If, without in any way disturbing a system, we can predict with certainty (i.e., with probability equal to unity) the value of a physical quantity, then there exists an element of physical reality corresponding to this physical quantity*." In other words, the quantity to be measured *has* a particular, albeit maybe unknown, value. EPR considered a coordinate system fixed in space and two particles: particle 1 with coordinate x_1 and momentum p_{x1} and particle 2 with coordinate x_2 and momentum p_{x2}, the total system being in a well-defined total momentum, $P = p_{x1} + p_{x2}$, and a well-defined relative position, $x = x_1 - x_2$. The meaning of the words "well-defined" is that, according to quantum mechanics, there is a possibility of the *exact measurement* of the two quantities (x and P), because the two operators $\hat{x}$ and $\hat{P}$ do commute.[61] At this point, Einstein and his colleagues and the great interpreters of quantum theory agreed.

---

[60] A. Einstein, B. Podolsky, N. Rosen, *Phys. Rev.*, 47(1935)777.

We now come to the crux of the real controversy.

The particles interact, then separate and fly far away (at any time we are able to measure exactly both x and P). When they are extremely far from each other (e.g., one close to us, the other one millions of light years away), we begin to suspect that each of the particles may be treated as free. Then, we decide to measure p_{x1}. However, after we do it, we know *with absolute certainty the momentum of the second particle* $p_{x2} = P - p_{x1}$, and this knowledge has been acquired *without any perturbation of particle* 2. According to the above cited statement, one has to admit that p_{x2} represents an element of physical reality. So far so good. However, we *might have decided* with respect to particle 1 to measure its *coordinate* x_1. If this happened, then we would know *with absolute certainty the position of the second particle*, $x_2 = x - x_1$, without perturbing particle 2 at all. Therefore, x_2, as p_{x2}, is an element of physical reality. The Heisenberg uncertainty principle says that it is *impossible* for x_2 and p_{x2} to be exactly measurable quantities. Conclusion: the Heisenberg uncertainty principle is wrong, and quantum mechanics is at least incomplete.

A way to defend the Heisenberg principle was to treat the two particles as an indivisible total system and reject the supposition that the particles are independent, even if they are millions of light years apart. This is how Niels Bohr desperately defended himself against Einstein (and his two colleagues). He said that the state of the total system in fact never fell apart into particles 1 and 2, and still is in what is known as an *entangled quantum state*[62] of the system of particles 1 and 2, and

> any measurement influences the state of the system as a whole, *independently of the distance of particles* A and B.

This reduces to the statement that measurement manipulations on particle 1 influence the results of measurements on particle 2. This correlation between measurements on particles 1 and 2 has to take place *immediately, regardless of the space that separates them.*[63] This is a shocking and nonintuitive feature of quantum mechanics. This is why it is often said that quantum mechanics cannot be understood. One can apply it successfully and obtain an excellent agreement with experiment, but there is something strange in its foundations. This represents a challenge: an excellent theory, but based on some unclear foundations.

---

[61] Indeed, $\hat{x}\hat{P} - \hat{P}\hat{x} = (\hat{x}_1 - \hat{x}_2)(\hat{p}_{x1} + \hat{p}_{x2}) - (\hat{p}_{x1} + \hat{p}_{x2})(\hat{x}_1 - \hat{x}_2) = [\hat{x}_1, \hat{p}_{x1}] - [\hat{x}_2, \hat{p}_{x2}] + [\hat{x}_1, \hat{p}_{x2}] - [\hat{x}_2, \hat{p}_{x1}] = -i\hbar + i\hbar + 0 - 0 = 0.$

[62] To honor Einstein, Podolsky, and Rosen the entanglement of states is sometimes called the EPR effect.

[63] Nevertheless, the correlation is not quite clear. One may pose some questions. The statement about the instantaneous correlation between particles 1 and 2 in the EPR effect cannot be correct, because the measurements are separated in the space–time manifold and the simultaneity is problematic (see Chapter 3). What is the laboratory fixed coordinate system? How is information about particle 2 transferred to where we carry out the measurement on particle 1? This takes time. After that time particle 1 is elsewhere. Is there anything to say about the separation time? In which coordinate system is the separation time measured?

In the following, some precise experiments will be described, in which it is shown that quantum mechanics is right, however absurd it looks. Before that let us consider a cat.

1.6 The life and death of Schrödinger's cat

The abovementioned paper by Einstein, Podolsky, and Rosen represented a severe critique of quantum mechanics in the form it has been presented by its fathers. After the appearance of the article, Erwin Schrödinger published a series of works[64] showing some other problematic issues in quantum mechanics. In particular, he described a *Gedankenexperiment*, later known as the Schrödinger's cat paradox. According to Schrödinger it shows some absurd consequences of quantum mechanics.

Here is the paradox. There is a cat closed in an isolated steel box (filled with air). Together with the cat there is a Geiger counter in the box (protected from the cat). We put in the counter a bit of a radioactive substance, carefully prepared in such a way that every hour two events happen with the same probability: either a radioactive nucleus decays or no radioactive nucleus decays. If a nucleus really decays, it causes ionization and electric discharge in the counter tube, which in turn results in a hammer hitting a glass capsule with hydrocyanic acid (HCN) gas, which kills the cat. If one puts the cat in for an hour, then *before* opening the box, one cannot say the cat is dead or alive. According to the Copenhagen interpretation, this is reflected by a proper wave function Ψ (for the box with everything in it), which is a superposition of two states with equal probability amplitudes: one state corresponds to the cat being alive, the other to the dead cat. The function Ψ describes therefore a cat in an intermediate state, neither alive nor dead, just in the middle between life and death, which, according to Schrödinger, would represent a totally absurd description. Einstein joined Schrödinger enthusiastically in this mockery from the Copenhagen interpretation, adding in his style "an even better" idea about how to kill the cat (gun powder instead of cyanide). All in all, in this way an atomic-scale phenomenon (decay of a nucleus) can result in some drastic events in the macroscopic world. If, after an hour, an observer opens the box ("measurement" of the state of the content of the box), then according to the Copenhagen interpretation, this will result in the collapse of the corresponding wave function and the observer will find the cat *either* alive *or* dead. This is the crux of the paradox.

Since the early days of quantum mechanics many scholars were trying to rationalize the paradox, always relying on some particular interpretation of quantum mechanics. In one of the interpretations, quantum mechanics does not describe a single system, but rather an infinite set of systems. We have therefore plenty of cats and the same number of boxes, each of them with the same macabre gear inside. Then, the paradox disappears, because, after the boxes are open, in 50% of cases the cats will be alive and in 50% they will be dead. In another interpretation it is criticized that Schrödinger treats the box as a quantum system, while the observer is treated classically. In this interpretation not only Schrödinger plays the role of the observer, but also the cat, and even the box itself (since it may contain a camera). What happened may be described differently by each of the observers, depending on what information they have about the whole system. For example, in the cat (alive or dead) there is information what has happened even *before* the box is open. The human observer does not have this information. Therefore, the collapse of the wave function happened earlier for the cat than for the observer! Only after the box is open it will turn out for the cat and for the observer that the collapse happened to the same state.

---

[64] E. Schrödinger, *Naturwissenschaften*, 23(1935)807, 823, 844.

Hugh Evereth stepped out with another, truly courageous, interpretation. According to Evereth we do not have our single Universe, as most of us might think naively, but a plethora of universes coexist, and in each of them some other things happen! When the box is open, the state of the box is entangled with the state of the observer, but then a collapse of the wave function happens, and in one of the universes the cat is alive, but in the other one it is dead (bifurcation). These two universes evolve independently ("parallel universes"), and they do not know about each other. Similar bifurcations happen massively in other events; hence, according to Evereth, the number of parallel Universes is astronomic.

The quantity of possible interpretations, large differences among them, and desperate character of some of them indicate that the problem of understanding quantum mechanics is still unsolved...

1.7 Bilocation

Assume that the world (stars, Earth, Moon, me, table, proton, electron, etc.) exists objectively. One may suspect this from everyday observations. For example, the Moon is seen by many people, who describe it in a similar way.[65] Instead of the Moon, let us begin with something simpler: how is it with electrons, protons, or other elementary particles? This is an important question because the world as we know it – including the Moon – is mainly composed of protons.[66] Here one encounters a mysterious problem. I will try to describe it by reporting results of several experiments.

---

[65] This *may* indicate that the Moon exists independently of our observations and overcome importunate suspicions that the Moon ceases to exist when we do not look at it. Besides, there are people who claim to have seen the Moon from very close and even touched it (admittedly through a glove) and this slightly strengthens our belief in the Moon's existence. First of all, one has to be cautious. For example, some chemical substances, hypnosis, or an ingenious set of mirrors may cause some people to be convinced about the reality of some phenomena, while others do not see them. Yet, would it help if even everybody saw? We should not verify serious things by voting. The example of the Moon also intrigued others; cf. D. Mermin, "*Is the Moon there, when nobody looks?*", *Phys. Today*, 38(1985)38.

Following Richard Feynman,[67] imagine two slits in a wall. Every second (the time interval has to be large enough to be sure that we deal with properties of a *single* particle) we send an electron towards the slits. There is a screen behind the two slits, and when an electron hits the screen, there is a flash (fluorescence) at the point of collision. Nothing special happens. Some electrons will not reach the screen at all, but traces of others form a pattern, which seems quite chaotic. The experiment looks monotonous and boring. Just a flash here, and another there. One cannot predict where a particular electron will hit the screen. But suddenly we begin to suspect that there is some regularity in the traces (Fig. 1.14).

A strange pattern appears on the screen: regions with high concentrations of traces are separated by regions of low concentration. This resembles the interference of waves, e.g., a stone thrown into water causes interference behind two slits: an alternation of high and low amplitudes of water level. Well, but what has an electron in common with a wave on the water surface? The interference on water was possible, because there were two sources of waves (the Huygens principle) – two slits.

Common sense tells us that nothing like this could happen with the electron, because, firstly, the electron could not pass through *both* slits, and, secondly, unlike the waves, the electron has hit a tiny spot on the screen (transferring there its energy). Let us repeat the experiment with a *single* slit. The electrons go through the slit and make flashes on the screen here and there, but there is only a single major concentration region (just facing the slit) fading away from the center (with some minor minima).

This result should make you feel faint. Why? You would like the Moon, a proton, or an electron to be solid objects, wouldn't you? All investigations made so far indicate that the electron is a point-like elementary particle. If, in the experiments we have considered, the electrons were to be divided into two classes, i.e., those that passed through slit 1 and those that passed through slit 2, then the electron patterns would be different. The pattern with the two slits *had to be* the

---

[66] In the darkest communist times a colleague of mine came to my office. Conspiratorially, very excited, and occasionally looking over his shoulder, he whispered: "*The proton decays!!!*" He just read in a government newspaper that the lifetime of a proton turned out to be finite. When asked about the lifetime, he gave an astronomical number, something like 10^{30} years or so. I said: "*Why do you look so excited then and why all this conspiracy?*" He answered: "*The Soviet Union is built of protons, and therefore is bound to decay as well!*"

[67] After Richard Feynman, "The Character of Physical Law," MIT Press, 1967.

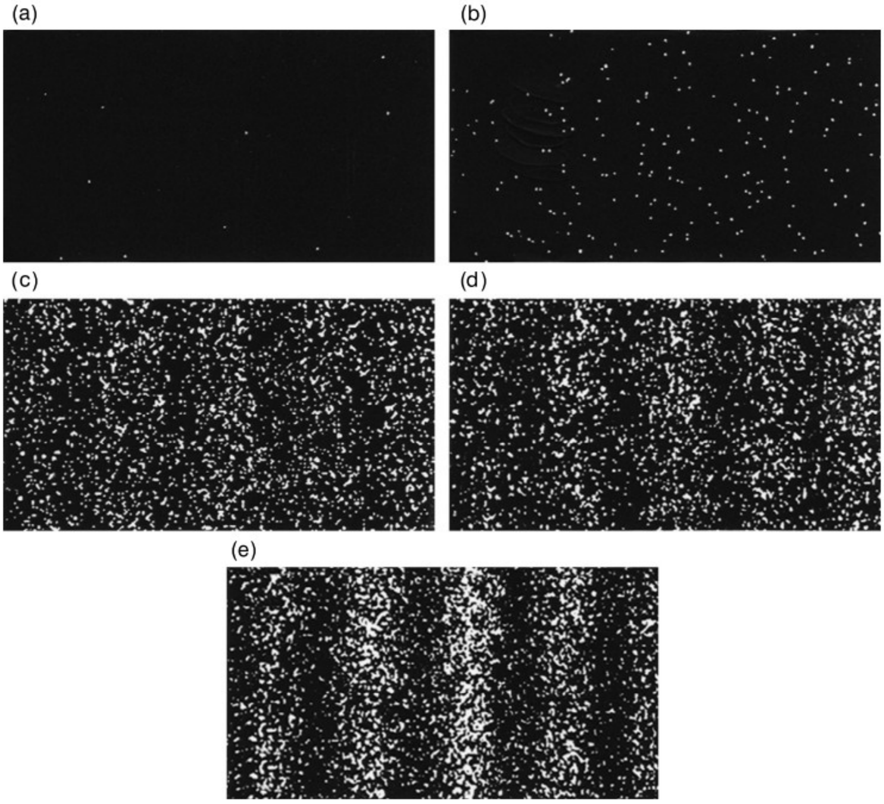

Fig. 1.14. Two-slit electron interference pattern registered by Akira Tonomura. (a) 10 electrons. (b) 100 electrons. (c) 3000 electrons – one begins to suspect something. (d) 20 000 electrons – no doubt, we will have a surprise. (e) 70 000 electrons – here it is! Conclusion: there is only one possibility – each electron went through the two slits (according to J. Gribbin, "*Q is for Quantum: An Encyclopedia of Particle Physics*," Weidenfeld and Nicolson, 1998).

sum of the patterns corresponding to only one open slit (facing slit 1 and slit 2). *We do not have that picture*.

> The *only* explanation for this interference of the electron with itself is that with the two slits open it went through *both*.

Clearly, the two parts of the electron united somehow and caused the flash at a single point on the screen. The quantum world is really puzzling. Despite the fact that the wave function is delocalized, the measurement gives its single point position (decoherence). How could an electron pass simultaneously through two slits? We do not understand this, but this is what happens.

Maybe it is possible to pinpoint the electron passing through two slits? Indeed, one may think of the Compton effect: a photon collides with an electron and changes its direction, and this can be detected (“a flash on the electron”). When one prepares two such ambushes at the two open slits, it turns out that the flash is *always on a single slit, not on both*. This cannot be true! If it were true, then the pattern would be of a *noninterference* character (and had to be the sum of the two one-slit patterns), but we have the interference. No. There is no interference. *Now*, the pattern does not show the interference. The interference was when the electrons were not observed. When we observe them, there is no interference... Somehow we perturb the electron’s momentum (the Heisenberg principle) and the interference disappears.

We have to accept that the electron passes through two slits. Maybe it only pertains to the electron, maybe the Moon is something completely different? A weak hope. The same thing happens to protons. Sodium atoms were also found to interfere.[68] A sodium atom, of diameter of a few Å, looks like an ocean liner when compared to a child’s toy boat of a tiny electron (almost 42 000 times lighter). And this ocean liner passed through two slits separated by thousands of Å. At the end of 1999 similar interference was observed for fullerene,[69] a giant C_{60} molecule, about one million times heavier than the electron. It is worth noting that after such adventure the fullerene molecule remained intact: somehow all its atoms, with the details of their chemical bonds, preserved their nature. There is something intriguing in this.

1.8 The magic of erasing the past

John Archibald Wheeler was not completely satisfied by the description of the hypothetical two-slit experiment, if it were performed at an astronomically large scale. Indeed, there was something puzzling in it. Suppose we have the two slits far away from us. A photon goes through the slit region, then flies towards us for a long time, and finally arrives to our screen. At large numbers of such photons, we obtain *an interference picture on the screen that witnesses each photon went through two slits*. However, when a photon passed the slit region, leaving the slits behind and heading towards us, we had plenty of time (say, of the order of a billion of years) to think. In particular, we might have an idea to replace the screen by two telescopes, each directed on one slit. In a similar situation we never observe half of the photon in one telescope and half in the other one: there is always a photon seen in *one* telescope only. *The*

---

[68] To observe such phenomena the slit distance has to be of the order of the de Broglie wave length, $\lambda = h/p$, where h is the Planck constant and p is the momentum. Cohen-Tannoudji lowered the temperature to such an extent that the momentum was close to 0, and λ could be of the order of thousands of Å.

[69] M. Arndt, O. Nairz, J. Voss-Andreae, C. Keller, G. van der Zouw, A. Zeilinger, *Nature*, 401(1999)680.

telescopes allow us to identify unambiguously which slit the particular photon passed through. Therefore, we may divide all such photons into two classes (depending on the slit they were through), and their distribution cannot be interference-like; it has to be bullet-like (it is a sum of distributions of both classes). And now we have a paradox.

> *Quantum eraser*
>
> Our decision (taken *after* the photons passed the slits) to replace the screen by the telescopes visibly changed the way the photons have been passing the slits. The problem is that the photon's decision must have been anticipating our action by some billion years or so.

This strange behavior (for a much shorter time scale) has been confirmed experimentally[70] in 2007. Well, this is a kind of drama. How may such a thing happen, however absurd it looks? Nobody knows. Let us put that clearly: it seems there is something fundamentally wrong in our current understanding of the world. Therefore, it is perhaps more safe to limit ourselves to strict physical *measurements* rather than telling pure science-fiction-like stories about traveling particles and their decisions anticipating situations that will happen in some billion years. Note also that we have "changed the past" only in a very special sense, since we are speaking here about two separate experiments: one with the telescopes and afterwards the other one with the screen. This is not like undoing an airplane catastrophe that has already taken place, but rather like changing something before the catastrophe takes place in order not to have it come.

1.9 A test for common sense: the Bell inequality

John Stuart Bell (1928–1990), Irish mathematician at Centre Européen de la Recherche Nucleaire (CERN) in Geneva. In the 1960s Bell attacked an old controversy of locality versus nonlocality, hidden variables, etc., apparently exhausted after exchange of ideas between Einstein and Bohr.

John Bell proved a theorem in 1964 that pertains to the results of measurements carried out on particles and some of the inequalities they have to fulfill. The theorem pertains to the basic logic of the measurements and is valid independently of the kind of particles and of the nature of their interaction. The theorem soon became abstract Heisenberg algebra. Schrödinger proved the equivalence of both theories, gave the solution for the hydrogen atom, and introduced the variational principle. ★ The time-dependent perturbation theory described in this chapter was developed by Paul Dirac in 1926. Twenty years later, Enrico Fermi, lecturing at the University of Chicago, coined the term "The Golden Rule" for these results. Since then, they are known as the Fermi Golden Rule.

---

[70] V. Jacques, E. Wu, F. Grosshans, F. Treussart, Ph. Grangier, A. Aspect, J.-F. Roch, *Science*, 315(2007)966.

* * *

2.1 Symmetry of the nonrelativistic Hamiltonian and the conservation laws

From classical mechanics it follows that for an isolated system (and assuming the forces to be central and obeying the action-reaction principle), its *energy*, *momentum*, and *angular momentum* are conserved.

Emmy Noether (1882–1935), German mathematician, informally professor, formally only the assistant of David Hilbert at the University of Göttingen (in the first quarter of the 20th century, women were not allowed to be professors in Germany). Her outstanding achievements in mathematics meant nothing to the Nazis, because Noether was Jewish (people should be reminded of such problems) and in 1933 Noether was forced to emigrate to the USA (Institute for Advanced Study in Princeton).

Imagine a well-isolated space ship observed in a space-fixed coordinate system. Its *energy is preserved*, its center of mass moves along a straight line with constant velocity (*the total, or center-of-mass, momentum vector is preserved*), and it preserves its *total angular momentum*.[2] The same is true for a molecule or atom, but the conservation laws have to be formulated in the language of quantum mechanics.

---

[2] That is, its *length and direction*. Think of a skater performing a spin: extending the arms sideways slows down the rotation, while stretching them along the axis of rotation results in faster rotation. *But all the time the total angular momentum vector is the same*. Well, what happens to the angular momentum, when the dancer finally stops rotating due to the friction? The angular momentum attains zero? No way. This is simply impossible. When the dancer increases its angular velocity, the Earth's axis changes a bit its direction to preserve the previous angular momentum of the total system (the Earth plus the dancer). When the dancer stops, the Earth's axis comes back towards its previous position, but not completely, because a part of the angular momentum is hidden in the rotation of molecules caused by the friction. Whatever happens, the total angular momentum has to be preserved! If the space ship captain wanted to stop the rotation of the ship which is making the crew sick, he could either throw something (e.g., gas from a steering jet) away from the ship, or spin a well-oriented body fast inside the ship. But even the captain is unable to change the *total* angular momentum.

Where do the conservation laws come from? Emmy Noether proved that they follow from the symmetry operations, with respect to which the equation of motion is invariant.[3]

> Thus, it turned out that invariance of the equation of motion with respect to an arbitrary:
> – *translation in time* (time homogeneity) results in the *energy conservation principle*;
> – *translation in space* (space homogeneity) gives the *total momentum conservation principle*;
> – *rotation in space* (space isotropy) implies the *total angular momentum conservation principle*.

These may be regarded as the foundations of science.[4] The homogeneity of time allows one to expect that repeating experiments give the same results. The homogeneity of space makes it possible to compare the results of the same experiments carried out in two different laboratories. Finally, the isotropy of space allows one to reject any suspicion that a different orientation of our laboratory bench changes the result.

Conservation laws represent most precious information about our system. It is not important what happens to the *isolated* system, what it is composed of, how complex the processes taking place in it are, whether they are slow or violent, whether there are people in the system or not, or whether they think about how to cheat the conservation laws; *nothing can violate the conservation of the energy, momentum, or angular momentum.*

Now, let us try to incorporate this into quantum mechanics.

---

[3] In case of a one-parameter family of operations $\hat{S}_\alpha \hat{S}_\beta = \hat{S}_{\alpha+\beta}$, e.g., translation ($\alpha$, β stand for the translation vectors), rotation (α, β are rotational angles), etc. Some other operations may not form such families and then the Noether theorem is no longer valid. This was an important discovery. Symmetry of a theory is much more fundamental than symmetry of an object. *The symmetry of a theory means that phenomena are described by the same equations no matter what laboratory coordinate system is chosen.*

[4] Well, to some extent. For example, the Universe does not show exact isotropy, because the matter there does not show spherical symmetry. Moreover, even if only one object were in the Universe, this very object would itself destroy the anisotropy of the Universe. We should rather think of this as a kind of idealization (approximation of the reality).

All symmetry operations (e.g., translation, rotation, reflection in a plane) are isometric, i.e., $\hat{U}^{\dagger} = \hat{U}^{-1}$ and $\hat{U}$ does not change distances between points of the transformed object (Figs. 2.2 and 2.3).

> The operator $\hat{U}$ acting in three-dimensional Cartesian space corresponds to the operator $\hat{\mathcal{U}}$ acting in the Hilbert space; cf. Eq. (C.2), p. 608. Thus the function $f(\mathbf{r})$ transforms to $f' = \hat{\mathcal{U}} f = f(\hat{U}^{-1}\mathbf{r})$, while the operator $\hat{A}$ transforms to $\hat{A}' = \hat{\mathcal{U}} \hat{A} \hat{\mathcal{U}}^{-1}$ (Fig. 2.3). The formula for $\hat{A}'$ *differs* in general from $\hat{A}$, but *when it does not*, i.e., $\hat{A}' = \hat{A}$, then $\hat{\mathcal{U}}$ commutes with $\hat{A}$.

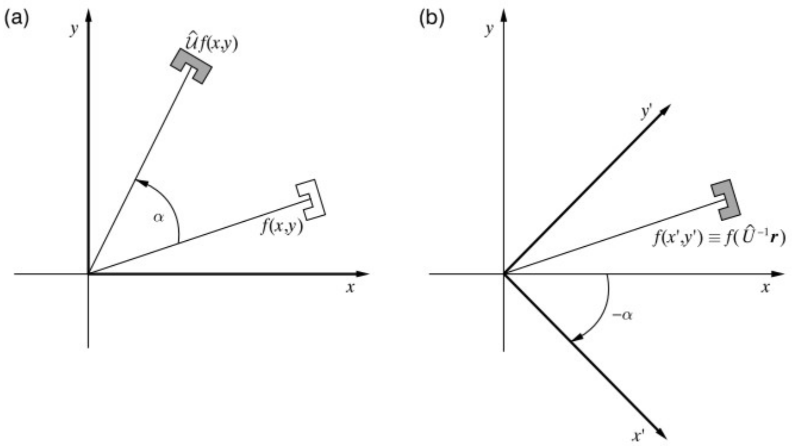
(a)
y
$\hat{\mathcal{U}} f(x,y)$
α
$f(x,y)$
x
(b)
y
y'
$f(x',y') \equiv f(\hat{U}^{-1}\boldsymbol{r})$
x
$-\alpha$
x'

Fig. 2.2. (a) An *object* is rotated by angle α. (b) The *coordinate system* is rotated by angle $-\alpha$. The new position of the object in the old coordinate system (a) is the same as the initial position of the object in the new coordinate system (b).

Indeed, then $\hat{A} = \hat{\mathcal{U}} \hat{A} \hat{\mathcal{U}}^{-1}$, i.e., one has the commutation relation $\hat{A}\hat{\mathcal{U}} = \hat{\mathcal{U}}\hat{A}$, which means that $\hat{\mathcal{U}}$ and $\hat{A}$ share their eigenfunctions (Appendix B, p. 595).

Let us take the Hamiltonian $\hat{H}$ as the operator $\hat{A}$. Before writing it down let us introduce *atomic units*. Their justification comes from something similar to laziness. The quantities one calculates in quantum mechanics are stuffed up by some constants: $\hbar = \frac{h}{2\pi}$, where h is the Planck constant, the electron charge $-e$, the electron (rest) mass m_0, etc. These constants appear in clumsy formulae with various powers, in the nominator and denominator (see Table of units, p. 763). One always knows, however, that the quantity one com-

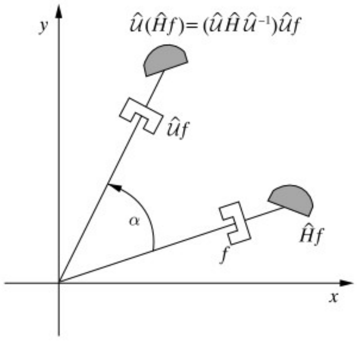
y
$\hat{\mathcal{U}}(\hat{H}f) = (\hat{\mathcal{U}}\hat{H}\hat{\mathcal{U}}^{-1})\hat{\mathcal{U}}f$
$\hat{\mathcal{U}}f$
α
$\hat{H}f$
f
x

Fig. 2.3. A schematic view from the Hilbert space of functions $f(x, y)$. The f and $\hat{H}f$ represent, in general, different functions. Rotation (by α) of function $\hat{H}f$ gives function $\hat{\mathcal{U}}(\hat{H}f)$ and, in consequence, is bound to denote the *rotation of* f (i.e., $\hat{U}f$) and the transformation $\hat{U}\hat{H}\hat{U}^{-1}$ *of the operator* $\hat{H}$. Indeed, only then $\hat{\mathcal{U}}\hat{H}\hat{\mathcal{U}}^{-1}$ when acting on the *rotated* function, i.e., $\hat{\mathcal{U}}f$ gives $\hat{\mathcal{U}}\hat{H}\hat{\mathcal{U}}^{-1}(\hat{\mathcal{U}}f) = \hat{\mathcal{U}}(\hat{H}f)$, i.e., the rotated result of the Hamiltonian action. Because of $\hat{\mathcal{U}}(\hat{H}f) = (\hat{\mathcal{U}}\hat{H})(\hat{\mathcal{U}}f)$, when verifying the invariance of $\hat{H}$ with respect to transformation $\hat{\mathcal{U}}$, it is sufficient to check whether $\hat{\mathcal{U}}\hat{H}$ has the same formula as $\hat{H}$, but expressed in the new coordinates. Only this $\hat{\mathcal{U}}\hat{H}$ will fit to f expressed in the new coordinates, i.e., to $\hat{\mathcal{U}}f$. This is how we will proceed shortly.

putes is energy, length, time, etc., and how energy, length, etc., are expressed in $\hbar$, e, m_0, etc.

ATOMIC UNITS

If one inserts $\hbar = 1$, $e = 1$, $m_0 = 1$, this gives a dramatic simplification of the formulae. One has to remember though that these units have been introduced and, whenever needed, one can evaluate the result in other units (see Table of conversion coefficients, p. 764).

The Hamiltonian for a system of M nuclei (with charges Z_I and mass m_I, $I = 1, ..., M$) and N electrons, in the nonrelativistic approximation and assuming point-like particles without any internal structure,[5] takes in atomic units (a.u.) the following form (see p. 22):

---

[5] No internal structure of the electron has yet been discovered. The electron is treated as a point-like particle. Contrary to this, nuclei have a rich internal structure and nonzero dimensions.
A clear multilevel-like structure appears (which has to a large extent forced a similar structure on the corresponding scientific methodologies):

- Level I. A nucleon (neutron, proton) consists of three (the valence) quarks, clearly seen on the scattering image obtained for the proton. Nobody has yet observed a free quark.

$$\hat{H} = \hat{T}_n + \hat{T}_e + \hat{V}, \tag{2.1}$$

where the kinetic energy operators for the nuclei and electrons (in a.u.) read as

$$\hat{T}_n = -\frac{1}{2}\sum_{I=1}^{M}\frac{1}{m_I}\Delta_I, \tag{2.2}$$

$$\hat{T}_e = -\frac{1}{2}\sum_{i=1}^{N}\Delta_i, \tag{2.3}$$

where the Laplacians are

$$\Delta_I = \frac{\partial^2}{\partial X_I^2} + \frac{\partial^2}{\partial Y_I^2} + \frac{\partial^2}{\partial Z_I^2},$$
$$\Delta_i = \frac{\partial^2}{\partial x_i^2} + \frac{\partial^2}{\partial y_i^2} + \frac{\partial^2}{\partial z_i^2},$$

and x, y, z stand for the Cartesian coordinates of the nuclei and electrons indicated by vectors $\mathbf{R}_I = (X_I, Y_I, Z_I)$ and $\mathbf{r}_i = (x_i, y_i, z_i)$, respectively.

- Level II. The strong forces acting among nucleons have a range of about 1–2 fm (1 fm $= 10^{-15}$ m). Above 0.4–0.5 fm they are attractive, at shorter distances they correspond to repulsion. One does not need to consider their quark structure when computing the forces among nucleons, but they may be treated as particles *without internal structure*. The attractive forces between nucleons practically do not depend on the nucleon's charge and are so strong that they may overcome the Coulomb repulsion of protons. Thus nuclei composed of many nucleons (various chemical elements) may be formed, which in mean-field theory exhibit a shell structure (analogous to the electronic structure, cf. Chapter 8) related to the packing of the nucleons. The motion of the nucleons is strongly correlated. A nucleus may have various energy states (ground and excited), may be distorted, may undergo splitting, etc. About 2000 nuclei are known, of which only 270 are stable. The smallest nucleus is the proton, the largest known so far is ^{209}Bi (209 nucleons). The largest observed number of protons in a nucleus is 118. Even the largest nuclei have diameters about 100 000 times smaller than the electronic shells of the atom. Even for an atom with atomic number 118, the first Bohr radius is equal to $\frac{1}{118}$ a.u., or $5 \cdot 10^{-13}$ m, still about 100 times larger than the nucleus.
- Level III. Chemists *can* neglect the internal structure of nuclei. A nucleus *can* be treated as a structureless point-like particle and using the theory described in this book, one is able to predict extremely precisely virtually all the chemical properties of atoms and molecules. Some interesting exceptions will be given at the end of Chapter 6.

The operator $\hat{V}$ corresponds to the electrostatic interaction of all the particles (nucleus–nucleus, nucleus–electron, electron–electron)[6]:

$$\hat{V} = \sum_{I=1}^{M}\sum_{J>I}^{M} \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} - \sum_{I=1}^{M}\sum_{i=1}^{N} \frac{Z_I}{|\mathbf{r}_i - \mathbf{R}_I|} + \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}, \tag{2.4}$$

or, in a simplified form,

$$\hat{V} = \sum_{I=1}^{M}\sum_{J>I}^{M} \frac{Z_I Z_J}{R_{IJ}} - \sum_{I=1}^{M}\sum_{i=1}^{N} \frac{Z_I}{r_{iI}} + \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{1}{r_{ij}}. \tag{2.5}$$

If the Hamiltonian turned out to be invariant with respect to a symmetry operation $\hat{U}$ (translation, rotation, etc.), this would imply the commutation of $\hat{U}$ and $\hat{H}$. We will check this in more detail below.

Note that the distances R_{IJ}, r_{iI}, and r_{ij} in the Coulombic potential energy (2.5) witness about assumption of instantaneous interactions in nonrelativistic theory (infinite travel speed of the interaction through space).

2.1.1 *Invariance with respect to translation*

Translation by vector $\mathbf{T}$ of function $f(\mathbf{r})$ in space means the function $\hat{\mathcal{U}} f(\mathbf{r}) = f\left(\hat{U}^{-1}\mathbf{r}\right) = f(\mathbf{r} - \mathbf{T})$, i.e., an opposite (by vector $-\mathbf{T}$) translation of the coordinate system (Fig. 2.4).

The transformation $\mathbf{r}' = \mathbf{r} + \mathbf{T}$ does not change the Hamiltonian. This is evident for the potential energy $\hat{V}$, because the translations $\mathbf{T}$ are canceled out, leaving the interparticle distances unchanged. For the kinetic energy one obtains $\frac{\partial}{\partial x'} = \sum_{\sigma=x,y,z} \frac{\partial \sigma}{\partial x'}\frac{\partial}{\partial \sigma} = \frac{\partial x}{\partial x'}\frac{\partial}{\partial x} = \frac{\partial}{\partial x}$, and all the kinetic energy operators (Eqs. (2.2) and (2.3)) are composed of the operators having this form.

> The Hamiltonian is therefore invariant with respect to any translation of the coordinate system.

There are two main consequences of translational symmetry:

---

[6] We do not include in this Hamiltonian tiny magnetic interactions of electrons and nuclei coming from their spin and orbital angular momenta, because they are of relativistic nature (see Chapter 3). In Chapter V2-4 we will be interested just in such small magnetic effects and the Hamiltonian will have to be generalized to include these interactions.

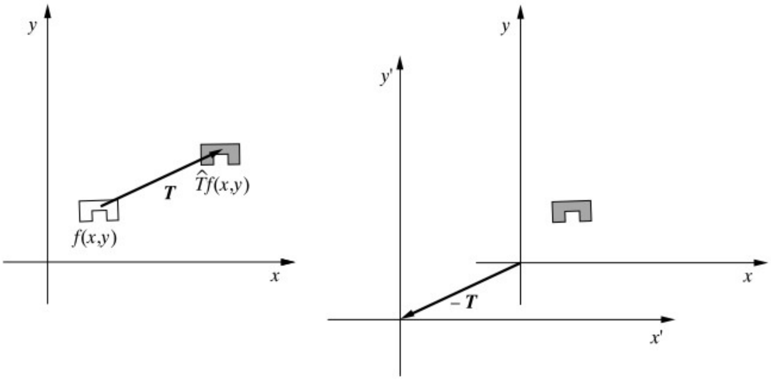

Fig. 2.4. A *function* f shifted by vector $\mathbf{T}$ (symmetry operation $\hat{T}$), i.e., $\hat{T}f(x,y)$ in the coordinate system (x,y) is the same as function $f\left(x',y'\right)$ in the *coordinate system* (x',y') shifted by $-\mathbf{T}$.

- No matter whether the coordinate system used is fixed in Trafalgar Square or in the center of mass of the system, one has to solve the same mathematical problem.
- The solution to the Schrödinger equation corresponding to the space-fixed coordinate system located in Trafalgar Square is Ψ_{pN}, whereas Ψ_{0N} is calculated in the body-fixed coordinate system (see Appendix J) located in the center of mass at $\mathbf{R}_{CM}$ that moves in the space-fixed coordinate system with the (total) momentum $\mathbf{p}_{CM}$. These two solutions are related by[7] $\Psi_{pN} = \Psi_{0N}\ \exp(i\mathbf{p}_{CM}\cdot\mathbf{R}_{CM})$. The number $N = 0, 1, 2, \ldots$ counts the energy states after the center-of-mass motion is separated.

> This means that the energy spectrum represents a continuum, because the center of mass may have any (nonnegative) kinetic energy $p_{CM}^2/(2m)$. If, however, one assumes that $p_{CM} = const$, then the energy spectrum is *discrete* for low-energy eigenvalues for atoms and molecules (see Eq. (1.13)).

This spectrum corresponds to the bound states, i.e., those states which do not correspond to any kind of dissociation (including ionization). Higher energy states lead to dissociation of the molecule, and the fragments may have any kinetic energy. Therefore, above the discrete spectrum one has a continuum of states. The states Ψ_{0N} will be called *spectroscopic states*. The bound states Ψ_{0N} are square integrable, as opposed to Ψ_{pN}, which are not because of function $exp(i\mathbf{p}\mathbf{R}_{CM})$, which describes the free motion of the center of mass.

---

[7] This follows from the separation of the center-of-mass motion (Appendix J) and noting that $\exp(i\mathbf{p}_{CM}\cdot\mathbf{R}_{CM})$ represents a solution for the motion of a free particle (Chapter 4).

2.1.2 Invariance with respect to rotation

The Hamiltonian is also invariant with respect to any rotation in space $\hat{U}$ of the coordinate system about a fixed axis. The rotation is carried out by applying an orthogonal matrix transformation[8] $\mathbf{U}$ of vector $\mathbf{r} = (x, y, z)^T$ that describes any particle of coordinates x, y, z. Therefore, all the particles undergo the same rotation and the new coordinates are $\mathbf{r}' = \hat{U}\mathbf{r} = \mathbf{U}\mathbf{r}$. Again, there is no problem with the potential energy, because a rotation does not change the interparticle distances. What about the Laplacians in the kinetic energy operators? Let us see. We have

$$\Delta = \sum_{k=1}^{3} \frac{\partial^2}{\partial x_k^2} = \sum_{k=1}^{3} \frac{\partial}{\partial x_k}\frac{\partial}{\partial x_k} = \sum_{k=1}^{3} \left(\sum_{i=1}^{3} \frac{\partial}{\partial x_i'}\frac{\partial x_i'}{\partial x_k} \right) \left(\sum_{i=1}^{3} \frac{\partial}{\partial x_i'}\frac{\partial x_i'}{\partial x_k} \right) =$$

$$\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3} \left(\frac{\partial}{\partial x_i'}\frac{\partial x_i'}{\partial x_k} \right) \left(\frac{\partial}{\partial x_j'}\frac{\partial x_j'}{\partial x_k} \right) =$$

$$\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3} \left(\frac{\partial}{\partial x_i'} U_{ik} \right) \left(\frac{\partial}{\partial x_j'} U_{jk} \right) = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3} \left(\frac{\partial}{\partial x_i'} U_{ik} \right) \left(\frac{\partial}{\partial x_j'} U_{kj}^{\dagger} \right) =$$

$$\sum_{i=1}^{3}\sum_{j=1}^{3} \left(\frac{\partial}{\partial x_i'} \right) \left(\frac{\partial}{\partial x_j'} \right) \sum_{k=1}^{3} U_{ik} U_{kj}^{\dagger} =$$

$$\sum_{i=1}^{3}\sum_{j=1}^{3} \left(\frac{\partial}{\partial x_i'} \right) \left(\frac{\partial}{\partial x_j'} \right) \delta_{ij} = \sum_{k=1}^{3} \frac{\partial^2}{\partial \left(x_k' \right)^2} = \Delta'.$$

Thus, one has invariance of the Hamiltonian with respect to any rotation about the origin of the coordinate system. This means (see p. 665) that the Hamiltonian and the operator of the square of the total angular momentum $\hat{J}^2$ (as well as of one of its components, denoted by $\hat{J}_z$) commute. One is able, therefore, to measure simultaneously the energy, the square of the total angular momentum as well as one of the components of the total angular momentum, and (as will be shown in Section 4.9) one has ($\mathbf{r}$ and $\mathbf{R}$ denote the electronic and the nuclear coordinates, respectively)

$$\hat{J}^2 \Psi_N(\mathbf{r}, \mathbf{R}) = J(J+1)\hbar^2 \Psi_N(\mathbf{r}, \mathbf{R}), \tag{2.6}$$

$$\hat{J}_z \Psi_N(\mathbf{r}, \mathbf{R}) = M_J \hbar \Psi_N(\mathbf{r}, \mathbf{R}), \tag{2.7}$$

where $J = 0, 1, 2...$ and $M_J = -J, -J+1, ... + J$.

---

[8] $\mathbf{U}^T = \mathbf{U}^{-1}$.

Any rotation matrix may be shown as a product of "elementary" rotations, each about the x, y, or z axis. For example, rotation about the y axis by angle θ corresponds to the matrix $\begin{pmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{pmatrix}$. The pattern of such matrices is simple: one has to put in some places sines, cosines, zeros, and ones with the proper signs.[9] This matrix is orthogonal,[10] i.e., $\mathbf{U}^T = \mathbf{U}^{-1}$, which you may easily check. The product of two orthogonal matrices represents an orthogonal matrix; therefore, any rotation corresponds to an orthogonal matrix.

2.1.3 Invariance with respect to permutation of identical particles (fermions and bosons)

The Hamiltonian also has permutational symmetry. This means that if someone exchanged labels numbering the identical particles, independently of how it was done, he/she would always obtain the identical mathematical expression for the Hamiltonian. This implies that any wave function has to be either symmetric (for bosons) or antisymmetric (for fermions) with respect to the exchange of labels between two identical particles (cf. p. 39).

2.1.4 Invariance of the total charge

In addition to the energy, momentum, and angular momentum, strict conservation laws exist exclusively for total electric charge and the baryon and lepton numbers (a given particle contributes $+1$, the corresponding antiparticle -1).[11] The charge conservation (whatever happens) follows from the gauge symmetry: total electric charge conservation follows from the fact that a description of the system has to be invariant with respect to the mixing of the particle and antiparticle states, which is analogous to rotation.

---

[9] Clockwise and counterclockwise rotations and two possible signs at sines cause a problem with memorizing the right combination. In order to choose the correct one, one may use the following trick. First, we decide that what moves is an *object* (e.g., a function, not the coordinate system). Then, you have to have this book in your pocket. In Fig. 2.2a one sees that the rotation of the point with coordinates $(1, 0)$ by angle $\theta = 90°$ should give the point (0, 1), and this is ensured only by the rotation matrix: $\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$.

[10] And therefore also unitary (cf. Appendix A, p. 589).

[11] For example, in the Hamiltonian (2.1) it is assumed that whatever might happen to our system, the numbers of the nucleons and electrons will remain constant.

2.1.5 Fundamental and less fundamental invariances

The conservation laws described are of a fundamental character, because they are related to the homogeneity of space and time, the isotropy of space, and the nondistinguishability of identical particles.

Besides these strict conservation laws (energy, momentum, angular momentum, permutation of identical particles, charge, and baryon and lepton numbers), there are also some approximate laws. Two of these, parity and charge conjugation, will be discussed below. They are rooted in these strict laws, but are valid only in some conditions. For example, in most experiments, not only the baryon number, but also the number of nuclei of each kind is conserved. Despite the importance of this law in chemical reaction equations, this does not represent any strict conservation law, as shown by radioactive transmutations of elements.

Some other approximate conservation laws will soon be discussed.

2.1.6 Invariance with respect to inversion – parity

There are orthogonal transformations which are not equivalent to any rotation, e.g., the matrix of inversion $\begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$, which corresponds to changing $\mathbf{r}$ to $-\mathbf{r}$ for all the particles, does not represent any rotation. If one performs such a symmetry operation, the Hamiltonian remains invariant and $|\Psi_{0N}(-\mathbf{r},-\mathbf{R})|^2 = |\Psi_{0N}(\mathbf{r},\mathbf{R})|^2$. This is evident, both for $\hat{V}$ (the interparticle distances do not change) and for the Laplacian (single differentiation changes sign, double does not). Two consecutive inversions mean identity operation: $\Psi_{0N}(\mathbf{r},\mathbf{R}) = \exp(i\alpha)\Psi_{0N}(-\mathbf{r},-\mathbf{R}) = [\exp(i\alpha)]^2\Psi_{0N}(\mathbf{r},\mathbf{R})$. Hence, $[\exp(i\alpha)]^2 = 1$, $\exp(i\alpha) = \pm 1$ and one has

$$\Psi_{0N}(-\mathbf{r},-\mathbf{R}) = \Pi\Psi_{0N}(\mathbf{r},\mathbf{R}), \quad \text{where } \Pi \in \{1,-1\}.$$

Therefore,

> the wave function of a stationary state represents an eigenfunction of the inversion operator, and the eigenvalue can be either $\Pi = 1$ or $\Pi = -1$. This property is known as parity (P).

Now the reader will be taken by surprise. From what we have said, it follows that *no molecule has a nonzero dipole moment*. Indeed, the dipole moment is calculated as the mean value of the dipole moment operator (as usual in quantum mechanics, see Postulate IV in Chapter 1), i.e.,

$\boldsymbol{\mu} = \langle \Psi_{0N} | \hat{\boldsymbol{\mu}} \Psi_{0N} \rangle = \langle \Psi_{0N} | \left(\sum_i q_i \mathbf{r}_i \right) \Psi_{0N} \rangle$, where q's represent particles' charges. This integral will be calculated very easily: the integrand is antisymmetric with respect to inversion[12] and therefore $\boldsymbol{\mu} = \mathbf{0}$.

Is therefore the very meaning of the dipole moment, a quantity often used in chemistry and physics, a fairy tale? If HCl has no dipole moment, then it is more understandable that also H_2 does not. All this seems absurd.

Let us stress that our conclusion pertains to the *total* wave function, which has to reflect the space isotropy leading to the zero dipole moment, because all orientations in space are equally probable. If one applied the transformation $\mathbf{r} \rightarrow -\mathbf{r}$ only to *some* particles in the molecule (e.g., electrons) and not to others (e.g., the nuclei), the wave function will show no parity (it would be neither symmetric nor antisymmetric). We will introduce the Hamiltonian in Chapter 6, which corresponds to immobilizing the nuclei (the adiabatic or Born–Oppenheimer approximation) in certain positions in space, and in such a case the wave function depends on the electronic coordinates only. This wave function may be neither symmetric nor antisymmetric with respect to the *partial* inversion transformation $\mathbf{r} \rightarrow -\mathbf{r}$ (for the electrons only). To give an example, let us imagine an HF molecule in a coordinate system, its origin being in the middle between the H and F nuclei. Consider a particular configuration of the 10 electrons of the molecule; all close to the fluorine nucleus in some well-defined points. One may compute the value of the wave function for this configuration of electrons. Its square gives us the probability density of finding this particular configuration of electrons. Now, imagine the (*partial*) inversion $\mathbf{r} \rightarrow -\mathbf{r}$ applied to all the electrons. Now they will all be close to the proton. If one computes the probability density for the new situation, one would obtain a completely different value (much, much smaller, because the electrons prefer the fluorine, not the hydrogen), so we have no symmetry or antisymmetry. No wonder therefore that if one computed $\boldsymbol{\mu} = \langle \Psi_{0N} | \hat{\boldsymbol{\mu}} \Psi_{0N} \rangle$ with such a function (integration is over the electronic coordinates only), the result would differ from zero. This is why chemists believe the HF molecule has a nonzero dipole moment.[13] On the other hand, if the molecule taken as the example were B_2 (also ten electrons), then the two values have to be equal, because they describe the same physical situation. This corresponds, therefore, to a wave function with definite parity (symmetric or antisymmetric), and therefore, in this case $\boldsymbol{\mu} = \mathbf{0}$. This is why chemists believe molecules such as H_2, B_2, and O_2 have no dipole moment.

---

[12] Ψ_{0N} may be symmetric or antisymmetric, but $|\Psi_{0N}|^2$ is bound to be symmetric. Therefore, since $\sum_i q_i \mathbf{r}_i$ is antisymmetric, indeed, the integrand is antisymmetric (while the integration limits are symmetric).

[13] What therefore they do measure? The answer will be given in Chapter V2-4.

Product of inversion and rotation

The Hamiltonian is also invariant with respect to some other symmetry operations, like changing the sign of the x coordinates of all particles, or similar operations which are products of inversion and rotation. If one changed the sign of all the x coordinates, it would correspond to a mirror reflection. Since rotational symmetry stems from space isotropy (which we will treat as "trivial"), the mirror reflection may be identified with parity P.

Enantiomers

A consequence of inversion symmetry is that the wave functions have to be eigenfunctions of the inversion operator with eigenvalues $\Pi = 1$, i.e., the wave function is symmetric, or $\Pi = -1$, i.e., the wave function is antisymmetric. Any asymmetric wave function corresponding to a stationary state is therefore excluded ("illegal"). However, two optical isomers (enantiomers), corresponding to an object and its mirror image, do exist (Fig. 2.5).[14]

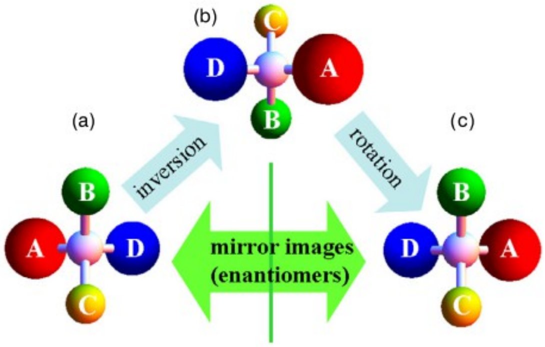

Fig. 2.5. Example of chirality. (a) A molecule's central atom with four different side groups (A, B, C, and D) in a nonplanar configuration. (b) The same molecule after applying the inversion operation (with respect to the central atom). (c) An attempt at superposing the initial and the transformed molecules by rotation, to get matching, fails. (a) and (c) as well as (a) and (b) represent an example of a pair of enantiomers. Each of these isomers, after reflection in a mirror (like, e.g., the central atom-BC plane), becomes identical to its partner.

---

[14] The property that distinguishes them is known as chirality (your hands are an example of chiral objects). The chiral molecules (enantiomers) exhibit optical activity, i.e., polarized light passing through a solution of one of the enantiomers undergoes a rotation of the polarization plane *always in the same direction* (which may be easily seen by reversing the direction of the light beam). The enantiomeric molecules have the same properties, provided one is checking this by using nonchiral objects. If the probe were chiral, one of the enantiomers would interact with it differently (for purely sterical reasons). Enantiomers (e.g., molecular associates) may be formed from chiral or nonchiral subunits.

We ask in a pharmacy for D-glucose. Strangely enough, the pharmacist is fully cooperative and does not make trouble. We pay a small sum and he gives us something which should not exist[15] – a substance with a single enantiomer. We should obtain a substance composed of molecules in their stationary states, which therefore have to have definite parity, either as a sum of the wave functions for the two enantiomers D and L ($\Pi = 1$, cf. Appendix D on p. 655, Case I), i.e., $\psi_+ = \psi_D + \psi_L$, or as the difference ($\Pi = -1$), i.e., $\psi_- = \psi_D - \psi_L$. The energies corresponding to ψ_+ and ψ_- differ, but the difference is extremely small (quasidegeneracy). The brave shopkeeper has given us something with the wave function $\psi = N(\psi_+ + \psi_-) = \psi_D$ (as result of decoherence), which therefore describes a nonstationary state.[16] As we will see in a moment (p. 107), the approximate lifetime of the state is proportional to the inverse of the integral $\left\langle \psi_D | \hat{H} \psi_L \right\rangle$. If one computed this integral, one would obtain an extremely small number.[17] It would turn out that the pharmacy could safely keep the stock of glucose for millions of years. Maybe the reason for decoherence is interaction with the rest of the Universe, maybe even interaction with a vacuum. The very existence of enantiomers, or even prevailing of one of them on Earth, does not mean breaking parity symmetry. This would happen if one of the enantiomers corresponded to a lower energy than the other.[18]

---

[15] More exactly, should be unstable.

[16] Only ψ_+ and ψ_- are stationary states.

[17] This is seen even after attempting to overlap two molecular models physically (Fig. 2.5), they do not match. The overlap of the wave functions will be small for the same reasons (the wave functions decay exponentially with distance).

[18] This is what happens in reality, although the energy difference is extremely small. Experiments with β-decay have shown that Nature breaks parity in weak interactions. The parity conservation law therefore has an approximate character.

With no preference for any of the enantiomers, one of them may *spontaneously* increase its concentration until 100%, i.e., with a complete elimination of the other one (0%)! Is something like that possible at all? It seems it contradicts common sense, since one of the enantiomers won, while the other lost, whereas their chances were exactly equal!

This phenomenon occurs in reality if autocatalysis is involved. The key information is the following: in such a system a *large random and self-augmenting fluctuation* is possible. Indeed, let us imagine just for simplicity 50 molecules of D and 50 molecules of L together with a certain number of molecules N (let us call them "neutral"), with the following possible reactions (giving equal chances to D and L):

$$\begin{gathered} \mathrm{D} + \mathrm{L} \rightarrow 2\mathrm{N} \\ \mathrm{D} + \mathrm{N} \rightarrow 2\mathrm{D} \\ \mathrm{L} + \mathrm{N} \rightarrow 2\mathrm{L}. \end{gathered}$$

The last two reactions (of autocatalytic character) represent in fact inducing an enantiomer ("forcing chirality") through interaction of N with a molecule of a given enantiomer. No preference of D or L is assumed in this induction. Nevertheless one of the enantiomers will defeat the other one! To explain this, let us assume that the elementary reactions 1, 2 or 3 of the individual molecules form a random chain in time, like, for example, 1, 2, 2, 3, 1, 3, ... If the chain starts by the first reaction, we have a situation similar to the starting one (49 : 49 instead of 50 : 50, in number of D:L molecules). If, however, the chain starts by chance from the second reaction,

REPUBLIQUE DE GUINEE
Office de la Poste Guinéenne 2008
PV
Chen Ning Yang
(楊振寧)
5000 FG.
PRIX NOBEL CHINOIS
REPUBLIQUE DE GUINEE
Office de la Poste Guinéenne 2008
Tsung-Dao Lee
(李政道)
5000 FG.
PRIX NOBEL CHINOIS

Chen Ning Yang (b. 1922) and Tsung Dao Lee (b. 1926), American physicists, professors at the Advanced Study Institute in Princeton, predicted in 1956 parity breaking in the weak interactions, which a few months later was confirmed experimentally by Madam Wu. In 1957 Yang and Lee received the Nobel Prize "*for their penetrating investigation of parity laws, which led to important discoveries regarding elementary particles*."

2.1.7 Invariance with respect to charge conjugation

If one changed the signs of the charges of all particles, the Hamiltonian would not change.

---

there will be a 51 : 50 preference of D, while when the third reaction takes place one will have 50 : 51 preference of L. In the last two cases we have a deviation from the equal chances of two enantiomers (fluctuation). Suppose we have the second case, i.e., 51 : 50 preference of D. Note that *now, when continuing the reaction chain at random, the chance for reaction 2 to happen is greater than the chance of reaction 3*. It is seen therefore that any fluctuation from an "equal-chances situation" has a tendency to self-augment, although there is always a chance that the fluctuation disappears. The racemic mixture (with the important N molecules though) represents therefore an instable system. Sooner or later there will be a transition (first mild but then more and more violent) from racemate to the absolute prevailing of one of the enantiomers. Their chances were and are equal, but just by chance one of them won.

Sometimes random processes may lead to large fluctuations. For example, Louis Pasteur was able to crystallize the crystals of one enantiomer and the crystals of the other enantiomer from a racemic solution. This is possible not only without Louis Pasteur, but also without any human, just spontaneously. Suppose that two such pieces of crystals were created where the primordial life was to be started, and that afterwards a volcano explosion took place, in which the volcanic lava covered (or even destroyed) one of the crystals, eliminating it from anything that happened later on (e.g., the chiral induction during biological evolution). This simple argument, proposed by Leszek Stolarczyk, is worth to be considered in discussions on how it comes that in living nature there is a strong preference for one chirality.

This therefore corresponds to exchanging particles and antiparticles.[19] Such a symmetry operation is called the charge conjugation and is denoted as C symmetry. This symmetry will be not marked in the wave function symbol (because, as a rule, we have to do with matter, not antimatter), but we will remember it. Sometimes it may turn out unexpectedly to be useful (see Chapter V2-5, p. V2-367). After Wu's experiment, physicists tried to save the hypothesis that what is conserved is the CP symmetry, i.e., the product of charge conjugation and inversion. However, analysis of experiments with decay of K mesons has shown that even this symmetry is approximate (although the deviation is extremely small).

2.1.8 Invariance with respect to the symmetry of the nuclear framework

In many applications the positions of the nuclei are fixed (clamped nuclei approximation, Chapter 6), often in a high-symmetry configuration (cf. Appendix C, p. 605). For example, the benzene molecule in its ground state (after minimizing the energy with respect to the positions of the nuclei) has the symmetry of a regular hexagon. In such cases the electronic Hamiltonian additionally exhibits invariance with respect to some symmetry operations and therefore the wave functions are the eigenstates of these molecular symmetry operations. Therefore, any wave function may have an additional label: the symbol of the irreducible representation[20] it belongs to.

2.1.9 Conservation of total spin

In an isolated system the *total* angular momentum **J** is conserved. However, $\mathbf{J} = \mathbf{L} + \mathbf{S}$, where **L** and **S** stand for the orbital and spin angular momenta (sum over all particles), respectively. The spin angular momentum **S**, being a sum over all particles, is not conserved.

However, the (nonrelativistic) Hamiltonian does not contain any spin variables. This means that it commutes with the operator of the square of the total spin as well as with the operator of one of the spin components (by convention the z component). Therefore, in the nonrelativistic approximation one can simultaneously measure the energy E, the square of the spin S^2, and one of its components: S_z.

---

[19] Somebody thought he had carried out computations for benzene, but he also computed antibenzene. The wave functions for benzene and antibenzene are the same.

[20] Of the symmetry group composed of the symmetry operations mentioned above, see Appendix C.

2.1.10 Indices of spectroscopic states

In summary, assumptions about the homogeneity of space and time, isotropy of space, and parity conservation lead to the following quantum numbers (indices) for the spectroscopic states:

- N quantizes energy,
- J quantizes the length of total angular momentum,
- M quantizes the z component of total angular momentum,
- Π determines parity:

$$\boxed{\Psi_{N,J,M,\Pi}\left(\mathbf{r},\mathbf{R}\right).}$$

Besides these indices following from the fundamental laws (in the case of parity it is a little too exaggerated), there may be also some indices related to less fundamental conservation laws:

- the irreducible representation index of the symmetry group of the clamped nuclei Hamiltonian (Appendix C),
- the values of S^2 (traditionally one gives the *multiplicity* $2S+1$) and S_z.

2.2 Schrödinger equation for stationary states

It may be instructive to see how Erwin Schrödinger invented his famous equation (1.13) for *stationary states* ψ of energy E ($\hat{H}$ denotes the Hamiltonian of the system),

$$\hat{H}\psi = E\psi. \tag{2.8}$$

Schrödinger surprised the contemporary quantum elite (associated mainly with Copenhagen and Göttingen) by a clear formulation of quantum mechanics as wave mechanics. January 27, 1926, when Schrödinger submitted a paper entitled "*Quantisierung als Eigenwertproblem*"[21] to *Annalen der Physik,* may be regarded as the birthday of wave mechanics.

Most probably, Schrödinger's reasoning was as follows. De Broglie discovered that what people called a particle also had a wave nature (Chapter 1). That is really puzzling. If a wave is involved, then according to Debye's suggestion at the November seminar in Zurich, it might be possible to write the standing wave equation with $\psi\left(x\right)$ as its amplitude at position x, i.e.,

$$v^2\frac{d^2\psi}{dx^2}+\omega^2\psi=0, \tag{2.9}$$

---

[21] Quantization as an eigenproblem. Well, once upon a time quantum mechanics was discussed in German. Some traces of that period remain in the nomenclature. One is the "*eigen*value problem or *eigen*problem," which is a German-English hybrid.

Isaac Newton (1643–1727), English physicist, astronomer, and mathematician, professor at Cambridge University, from 1672 member of the Royal Society of London, from 1699 Director of the Royal Mint – said to be merciless to the forgers. In 1705 Newton became a Lord. In the *opus magnum* mentioned above he developed the notions of space, time, mass, and force, gave three principles of dynamics and the law of gravity, and showed that the latter pertains to problems that differ enormously in their scale (e.g., the famous apple and the planets). Newton is also a founder of differential and integral calculus (independently from G.W. Leibnitz). In addition Newton made some fundamental discoveries in optics; among other things he was the first to think that light is composed of particles.

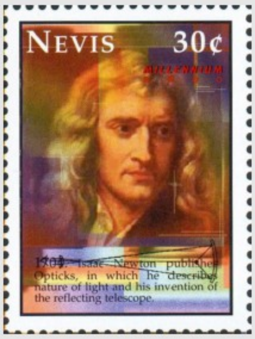
NEVIS
30¢
MILLENNIUM
1704: Isaac Newton publishes Opticks, in which he describes nature of light and his invention of the reflecting telescope.

The first portrait of Newton (at the age of 46) by Godfrey Kneller is shown. The readers are encouraged to read the excellent book "Isaac Newton" by G.E. Christianson, Oxford University Press, New York, 1996.

7.1 Rovibrational spectra – an example of accurate calculations: atom–diatomic molecule

One of the consequences of adiabatic approximation is the potential energy hypersurface $V(\mathbf{R})$ for the motion of nuclei. To obtain the wave function for the motion of nuclei (and then to construct the total product-like wave function for the motion of electrons and nuclei) we have to solve the Schrödinger equation with $V(\mathbf{R})$ as the potential energy. This is what this hypersurface is for. We will find rovibrational energy levels and the corresponding wave functions, which will serve us to obtain rovibrational spectra (frequencies and intensities) to compare with experimental results.

7.1.1 Coordinate system and Hamiltonian

Let us consider a diatomic molecule AB plus a weakly interacting atom C (e.g., HH$\cdots$Ar or CO$\cdots$He), the total system in its electronic ground state. Let us center the origin of the body-fixed coordinate system[2] (with the axes oriented as in the space-fixed coordinate system, see Appendix J, p. 691) in the center of mass of AB. The problem involves therefore $3 \times 3 - 3 = 6$ dimensions.

---

[2] Any coordinate system is equally good from the point of view of mathematics, but its particular choice may make the solution easy or difficult. In the case of a weak C$\cdots$AB interaction (our case) the proposed choice of the origin is one of the natural ones.

However strange it may sound, six is too much for contemporary (otherwise impressive) computer techniques. Let us subtract one dimension by assuming that no vibrations of AB occur (rigid rotator). The five-dimensional problem becomes manageable. The assumption about the stiffness of AB now also pays off because we exclude right away two possible chemical reactions C + AB→ CA + B and C + AB → CB + A, and admit therefore only some limited set of nuclear configurations – only those that correspond to a weakly bound complex C + AB. This approximation is expected to work better when the AB molecule is stiffer, i.e., has a larger force constant (and therefore vibration frequency).[3]

Carl Gustav Jacob Jacobi (1804–1851), German mathematical genius, son of a banker, graduated from school at the age of 12, then associated with the universities of Berlin and Königsberg. Jacobi made important contributions to number theory, elliptic functions, partial differential equations, and analytical mechanics.

We will introduce the *Jacobi coordinates* (Fig. 7.2): three components of vector **R** pointing to C from the origin of the coordinate system (the length R and angles Θ and Φ, both angles denoted by $\hat{\mathbf{R}}$) and the angles θ, ϕ showing the orientation $\hat{\mathbf{r}}$ of vector $\mathbf{r} = \overrightarrow{AB}$, altogether five coordinates – as there should be.

Now let us write down the Hamiltonian for the motion of the nuclei in the Jacobi coordinate system (with the stiff AB molecule with the AB equilibrium distance equal to r_{eq}).[4] We have

$$\hat{H} = -\frac{\hbar^2}{2\mu R^2}\frac{d}{dR}R^2\frac{d}{dR} + \frac{\hat{l}^2}{2\mu R^2} + \frac{\hat{j}^2}{2\mu_{AB} r_{eq}^2} + V,$$

where $\hat{l}^2$ denotes the operator of the square of the angular momentum of the atom C, $\hat{j}^2$ stands for the square of the angular momentum of the molecule AB, where

$$\hat{l}^2 = -\hbar^2\left[\frac{1}{\sin\Theta}\frac{\partial}{\partial\Theta}\sin\Theta\frac{\partial}{\partial\Theta} + \frac{1}{\sin^2\Theta}\frac{\partial^2}{\partial\Phi^2}\right],$$
$$\hat{j}^2 = -\hbar^2\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\sin\theta\frac{\partial}{\partial\theta} + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}\right],$$

---

[3] A certain measure of this might be the ratio of the dissociation energy of AB to the dissociation energy of C··· AB. The higher the ratio, the better our model will be.

[4] As proposed in S. Bratož, M.L. Martin, *J. Chem. Phys.*, 42(1965)1051.

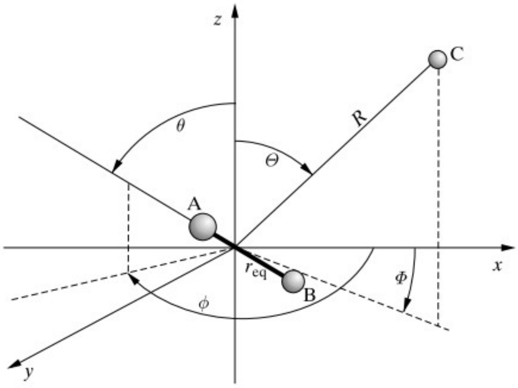

Fig. 7.2. The Jacobi coordinates for the C$\cdots$AB system. The origin is in the center of mass of AB (the distance AB is constant and equal to r_{eq}). The positions of atoms A and B are fixed by giving the angles θ,ϕ. The position of atom C is determined by three coordinates: R, Θ, and Φ. Altogether we have five coordinates: $R, \Theta, \Phi, \theta, \phi$ or (R, $\hat{\mathbf{R}}$ and $\hat{\mathbf{r}}$).

μ is the reduced mass of C and the mass of (A+B), μ_{AB} denotes the reduced mass of A and B, and V stands for the potential energy of the nuclear motion.

The expression for $\hat{H}$ is quite understandable. First of all, we have in $\hat{H}$ five coordinates, as there should be: R, two angular coordinates hidden in the symbol $\hat{\mathbf{R}}$, and two angular coordinates symbolized by $\hat{\mathbf{r}}$ – the four angular coordinates enter the operators of the squares of the two angular momenta. The first three terms in $\hat{H}$ describe the kinetic energy, V is the potential energy (the electronic ground-state energy depends on the nuclear coordinates). The kinetic energy operator describes the radial motion of C with respect to the origin (first term), the rotation of C about the origin (second term), and the rotation of AB about the origin (third term).

7.1.2 Anisotropy of the potential V

How to figure out the shape of V? Let us first make a section of V. If we freeze the motion of AB,[5] atom C would have (concerning the interaction energy) a sort of an energetic well around AB wrapping the AB molecule, caused by the C$\cdots$AB van der Waals interaction. The bottom of the well would be quite distant from the molecule (van der Waals equilibrium distance), while the shape determined by the bottom points would resemble the shape of AB, i.e., would be a little bit elongated. The depth of the well would vary depending on orientation with respect to the origin.

---

[5] That is, fixed the angles θ and ϕ.

If V were isotropic, i.e., if atom C would have C$\cdots$AB interaction energy independent[6] of $\hat{\mathbf{r}}$, then of course we might say that there is no coupling between the rotation of C and the rotation of AB. We would then have a conservation law separately for the first and the second angular momentum and the corresponding commutation rules (cf. Chapter 2 and Appendix F), i.e.,

$$\left[\hat{H}, \hat{l}^2\right] = \left[\hat{H}, \hat{j}^2\right] = 0,$$

$$\left[\hat{H}, \hat{l}_z\right] = \left[\hat{H}, \hat{j}_z\right] = 0.$$

Therefore, the wave function of the total system would be the eigenfunction of $\hat{l}^2$ and $\hat{l}_z$ as well as of $\hat{j}^2$ and $\hat{j}_z$. The corresponding quantum numbers $l = 0, 1, 2, \ldots$ and $j = 0, 1, 2, \ldots$, which determine the squares of the angular momenta for $\hat{l}^2$ and $\hat{j}^2$, as well as the corresponding quantum numbers $m_l = -l, -l+1, \ldots, l$ and $m_j = -j, -j+1, \ldots, j$, which determine the projections of the corresponding angular momenta on the z axis, would be legal[7] quantum numbers (full analogy with the rigid rotator, Chapter 4). The rovibrational levels could be labeled using pairs of quantum numbers: (l, j). In the absence of an external field (no privileged orientation in space) any such level would be $(2l+1)(2j+1)$-tuply degenerate, since this is the number of different projections of both angular momenta on the z axis.

7.1.3 Adding the angular momenta in quantum physics

However, V *is not isotropic* (although the anisotropy is small). What then? Then, of all angular momenta, only the *total* angular momentum $\mathbf{J} = \mathbf{l} + \mathbf{j}$ is conserved (the conservation law results from the very foundations of physics, cf. Chapter 2).[8] Therefore, the vectors $\mathbf{l}$ and $\mathbf{j}$, when added to $\mathbf{J}$, would make all allowed angles: from minimum angle (the quantum number $J = l + j$), through larger angles[9] and the corresponding quantum numbers $J = l + j - 1, l + j - 2$, etc., up to the maximum angle, corresponding to $J = |l - j|$). Therefore, the number of all possible values of J (each corresponding to a different energy) is equal to the number of projections of the *shorter*[10] of the vectors $\mathbf{l}$ and $\mathbf{j}$ on the longer one, i.e.,

---

[6] That is, the bottom of the well would be a sphere centered in the center of mass of AB and the well depth would be independent of the orientation.

[7] We use to say "*good*."

[8] Of course, the momentum has also been conserved in the isotropic case, but in this case the energy was identical independently of the quantum number J (resulting from different angles between $\mathbf{l}$ and $\mathbf{j}$).

[9] The projections of the angular momenta are quantized.

$$J = (l + j), (l + j - 1), \ldots, |l - j|. \tag{7.1}$$

For a given J there are $2J + 1$ projections of $\mathbf{J}$ on the z axis (because $|M_J| \le J$); without any external field all these projections correspond to identical energy.

Please check that the number of all possible eigenstates is equal to $(2l+1)(2j+1)$, i.e., exactly what we had in the isotropic case. For example, for $l = 1$ and $j = 1$ the degeneracy in the isotropic case is equal to $(2l+1)(2j+1) = 9$, while for anisotropic V we would deal with five states for $J = 2$ (all of the same energy), three states corresponding to $J = 1$ (the same energy, but different from $J = 2$), and a single state with $J = 0$ (still another value of energy), altogether nine states. This means that switching anisotropy on partially removed the degeneracy of the isotropic level (l, j) and gave the levels characterized by quantum number J.

7.1.4 Application of the Ritz method

We will use the Ritz variational method (see Chapter 5, p. 271) to solve the Schrödinger equation. What should we propose as the expansion functions? It is usually recommended that we proceed systematically and choose first a complete set of functions depending on R, then a complete set depending on $\hat{\mathbf{R}}$, and finally a complete set that depends on the $\hat{\mathbf{r}}$ variables. Next, one may create the complete set depending on all five variables (these functions will be used in the Ritz variational procedure) by taking all possible products of the three functions depending on R, $\hat{\mathbf{R}}$, and $\hat{\mathbf{r}}$. There is no problem with the complete sets that have to depend on $\hat{\mathbf{R}}$ and $\hat{\mathbf{r}}$, as these may serve the spherical harmonics (the wave functions for the rigid rotator) $\{Y_l^m(\Theta, \Phi)\}$ and $\{Y_{l'}^{m'}(\theta, \phi)\}$, while for the variable R we may propose the set of harmonic oscillator wave functions $\{\chi_v(R)\}$.[11] Therefore, we may use the following as the variational function[12]:

$$\Psi(R, \Theta, \Phi, \theta, \phi) = \sum c_{vlml'm'} \chi_v(R) Y_l^m(\Theta, \Phi) Y_{l'}^{m'}(\theta, \phi),$$

---

[10] In the case of two vectors of the same length, the role of the shorter vector may be taken by either of them.

[11] See Chapter 4. Of course, our system does not represent any harmonic oscillator, but what counts is that the harmonic oscillator wave functions form a complete set (as the eigenfunctions of a Hermitian operator).

[12] The products $Y_l^m(\Theta, \Phi) Y_{l'}^{m'}(\theta, \phi)$ may be used to produce linear combinations that are automatically the eigenfunctions of $\hat{J}^2$ and $\hat{J}_z$, and have the proper parity (see Chapter 2). This may be achieved by using the Clebsch–Gordan coefficients (D.M. Brink, G.R. Satchler, *Angular Momentum*, Clarendon, Oxford, 1975). The good news is that this way we can obtain a smaller matrix for diagonalization in the Ritz procedure, the bad news is that the matrix elements will contain more terms to be computed. The method described above will give the same result as using the Clebsch–Gordan coefficients, because the eigenfunctions of the Hamiltonian obtained within the Ritz method will automatically be the eigenfunctions of $\hat{J}^2$ and $\hat{J}_z$ and have the proper parity.

where c are the variational coefficients and the summation goes over v, l, m, l', m' indices. The summation limits have to be finite in practical applications; therefore the summations go to some maximum values of v, l, and l' (m and m' vary from $-l$ to l and from $-l'$ to $+l'$). We hope (as always in quantum chemistry) that numerical results of a demanded accuracy will not depend on these limits. Then, as usual, the Hamiltonian matrix is computed and diagonalized (see p. 703), and the eigenvalues E_J as well as the eigenfunctions ψ_{J,M_J} of the ground and excited states are found. Each of the eigenfunctions will correspond to some J, M_J and to a certain parity. The problem is solved.

7.2 Force fields (FFs)

The middle of the 20th century marked the end of a long period of determining the building blocks of chemistry: chemical elements, chemical bonds, and bond angles. The lists of these are not definitely closed, but future changes will be rather cosmetic than fundamental. This made it possible to go one step further and begin to rationalize the structure of molecular systems as well as to foresee the structural features of the compounds to be synthesized. The crucial concept is based on the Born–Oppenheimer approximation and on the theory of chemical bonds and resulted in the spatial structure of molecules. The great power of such an approach was first proved by the construction of the DNA double-helix model by Watson and Crick. The first DNA model was build from iron spheres, wires, and tubes.

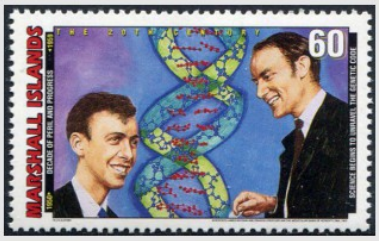
MARSHALL ISLANDS
60
THE 20TH CENTURY
DECADE OF PERIL AND PROGRESS
SCIENCE BEGINS TO UNRAVEL THE GENETIC CODE

James Dewey Watson (left), born in 1928, American biologist, professor at Harvard University. Francis Harry Compton Crick (right) (1916–2004), British physicist, professor at Salk Institute in San Diego. Both scholars won the 1962 Nobel Prize for "*their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material*." At the end of the historic paper J.D. Watson, F.H.C. Crick, *Nature*, 737(1953) (of about 800 words) the famous enigmatic but crucial sentence appears: "*It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material*." The story behind the discovery is described in a colorful and nonconventional way by Watson in his book "*Double helix: A Personal Account of the Discovery of the Structure of DNA*."

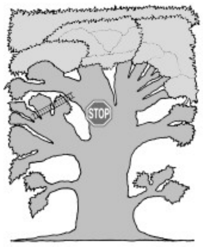

CHAPTER 8

Orbital Model of Electronic Motion in Atoms and Molecules

Everything should be made as simple as possible, but not simpler.
Albert Einstein

Where are we?

We are in the upper part of the main trunk of the TREE.

An example

What is the electronic structure of atoms? How do atoms interact in a molecule? Two *neutral* moieties (say, hydrogen atoms) attract each other with a large force *of a similar order of magnitude* as the Coulombic forces between two ions. This is quite surprising. What pulls these neutral objects to one another? These questions are at the foundations of chemistry.

What is it all about?

Hartree–Fock method – a bird's eye view (▲) **p. 463**

- Spin orbitals as the one-electron building blocks
- Variables
- Slater determinant
- What is the Hartree–Fock method all about?

Towards the optimal spin orbitals and the Fock equation (△) **p. 469**

- Dirac notation for integrals
- Energy functional to be minimized
- Energy minimization with constraints
- Slater determinant subject to a unitary transformation
- The $\hat{J}$ and $\hat{K}$ operators are invariant
- Diagonalization of the Lagrange multipliers
- Optimal spin orbitals are solutions of the Fock equation (general Hartree–Fock [GHF] method)

Ideas of Quantum Chemistry
https://doi.org/10.1016/B978-0-44-464246-2.00016-1
Copyright © 2020 Elsevier B.V. All rights reserved.

- “Unrestricted” Hartree–Fock (UHF) method
- The closed shell systems and the restricted Hartree–Fock (RHF) method
- Iterative solution: the Self-Consistent Field (SCF) method

Total energy in the Hartree–Fock method (△) **p. 490**

Computational technique: atomic orbitals as building blocks of the molecular wave function (▲) **p. 492**

- Centering of an atomic orbital
- Slater-type orbitals (STOs)
- Gaussian-type orbitals (GTOs)
- Linear combination of atomic orbitals (LCAO) method
- Basis sets of atomic orbitals
- The Hartree–Fock–Roothaan method (SCF LCAO MO)
- Some practical problems

Back to foundations (△ ⧫) **p. 510**

- When does the RHF method fail? (△)
- Fukutome classes (⧫)

RESULTS OF THE HARTREE–FOCK METHOD

Mendeleev’s periodic table (▲) **p. 521**

- All atoms are similar to the hydrogen atom – the orbital model of an atom
- Shells and subshells
- Educated guess of atomic orbitals – the Slater rules
- Atomic radii

The nature of the chemical bond – quantum makes a difference (△) **p. 531**

- The simplest chemical bond: H_2^+ in the MO picture
- Can we see a chemical bond?

Excitation energy, ionization potential, and electron affinity (RHF approach) (△) **p. 537**

- Approximate energies of electronic states
- Singlet or triplet excitation?
- Hund’s rules
- Hund’s rules for the atomic terms
- Ionization potential and electron affinity (Koopmans’ theorem)

Towards a chemical picture – localization of molecular orbitals (△) **p. 547**

- Can a chemical bond be defined in a polyatomic molecule?
- The external localization methods
- The internal localization methods
- Examples of localization

- Localization in practice – computational technique
- The chemical bonds of σ, π, δ symmetry
- Electron pair dimensions and the foundations of chemistry
- Hybridization or mixing one-center AOs

A minimal model of a molecule (▲) **p. 571**

- Valence Shell Electron Pair Repulsion (VSEPR)

The isolobal analogy (△) **p. 578**

The Born–Oppenheimer (or adiabatic) approximation is the central point of this book (note its position in the TREE). Thanks to the approximation, we can consider *separately* two *coupled* problems concerning molecules:

- the motion of the electrons at fixed positions of the nuclei (to obtain the electronic energy),
- the motion of nuclei in the potential representing the electronic energy of the molecule (see Chapter 7).

From now on we will concentrate on the *motion of the electrons at fixed positions of the nuclei* (the Born–Oppenheimer approximation, p. 320).

To solve the corresponding Eq. (6.18), we have at our disposal the variational and the perturbation methods. The latter one should have a reasonable starting point (i.e., an unperturbed system). This is not the case in the problem we want to consider at the moment. Thus, only the variational method remains. If so, a class of the trial functions should be proposed. In this chapter the trial wave function will have a very specific form, bearing significant importance for the theory. We mean here the so-called Slater determinant, which is composed of molecular orbitals (MOs). At a certain level of approximation, each MO is a "home" for two electrons. We will now learn how to get the optimum MOs (Hartree–Fock method). In spite of some quite complex formulae, which will appear below, the main idea behind them is extremely simple. It can be expressed in the following way.

Let us consider road traffic; the cars (electrons) move at fixed positions of buildings (nuclei). The safe motion of the cars looks very complex indeed (as it does for the electrons). Is there any way at all to design such a complex many-body and safe motion? Well, one may radically simplify the problem and use the so-called mean-field approximation (paying necessarily the price of lower quality). In the mean-field approximation method we focus on the motion of *one car only*, considering its motion in such a way that the car avoids *those streets that are known as most jammed*. In this chapter we will treat the electrons in a similar manner (leaving the difficulties of considering the correlation of the motions of the electrons to Chapter V2-2). Now, the electrons will not feel the true electric field of the other electrons (as they should in a precise approach), but rather their static *mean* electric field, i.e., averaged over their motions. The idea looks hopeless, but it turns out to be a very effective and fruitful approach.

Translating this into quantum mechanical language, the underlying assumptions of the mean-field method for N identical particles (here: electrons) are the following:

- there is a certain "effective" *one-particle* operator $\hat{F}(i)$ of an identical mathematical form for all particles $i = 1, 2, ...N$, which has the eigenfunctions ϕ_k, i.e., $\hat{F}\phi_k = \varepsilon_k \phi_k$, such that
- $\left\langle \Psi | \hat{H} \Psi \right\rangle \approx \left\langle \tilde{\Psi} | \hat{H}^{ef} \tilde{\Psi} \right\rangle$, where $\tilde{\Psi}$ (normalized) is a wave function that approximates the exact wave function Ψ for the total system, $\hat{H}$ is the electronic Hamiltonian (in the clamped nuclei approximation, Chapter 6), and $\hat{H}^{ef} = \sum_{i=1}^{N} \hat{F}(i)$. In such a case the eigenvalue equation $\hat{H}^{ef} \Pi_{i=1}^{N} \phi_i(i) = E_0 \Pi_{i=1}^{N} \phi_i(i)$ holds, and the approximate total energy is equal to $E_0 = \sum_{i=1}^{N} \varepsilon_k$, *as if the particles were independent*.

Any mean-field method needs to solve two problems:

- How should $\tilde{\Psi}$ be constructed using N eigenfunctions ϕ_k?
- What is the form of the one-particle effective operator $\hat{F}$?

These questions will be answered in the present chapter.

Such effectively independent, yet interacting particles, are called quasiparticles or – as we sometimes use to say – bare particles dressed up by the interaction with others.

It is worth remembering that the mean-field method bears several different names in chemistry:

- one-determinant approximation,
- one-electron approximation,
- one-particle approximation,
- molecular orbital method,
- independent-particle approximation,
- mean-field approximation,
- Hartree–Fock method,
- self-consistent field method (as regards practical solutions).

It will be shown how the mean-field method implies that milestone of chemistry: the periodic table of chemical elements.

Next, we will endeavor to understand *why* two atoms create a chemical bond, and also what affects the ionization energy and the electron affinity of a molecule.

Then, still within the MO scheme, we will show how we can reach a localized description of a molecule, with chemical bonds between *some* atoms, with the inner electronic shells, and the lone electronic pairs. The last terms are elements of a rich and very useful language commonly used by chemists.

Why this is important?

Contemporary quantum chemistry uses better methods than the mean field, described in this chapter. We will get know them in Chapters V2-2 and V2-3. Yet all these methods *start from the mean-field approximation* and in most cases they only perform cosmetic changes in energy and in electron distribution.

For example, the methods described here yield about 99% of the total energy of a system.[1] There is one more reason for why this chapter is important. Methods beyond the one-electron approximation are – computationally – very time consuming (hence they may be applied only to small systems), while the molecular orbital approach is the "daily bread" of quantum chemistry. It is a sort of standard method, and the standards have to be learned.

What is needed?

- Postulates of quantum chemistry (Chapter 1, necessary),
- operator algebra, Hermitian operators (Appendix B, p. 595, necessary),
- complete set of functions (Chapter 1, necessary),
- Hilbert space (Appendix B, p. 595, recommended),
- determinants (Appendix A, p. 589, absolutely necessary),
- Slater–Condon rules (Appendix N, p. 707, only the results are needed),
- Lagrange multipliers (Appendix O, p. 719, necessary),
- Mulliken population analysis (Appendix V2-D, p. V2-595, occasionally used).

Classical works

This chapter deals with the basic theory explaining the electronic structure of atoms and molecules. This is why we begin with Dimitryi Ivanovitch Mendeleev, who discovered in 1865, when writing his textbook "*Osnovy Khimii*" ("*Principles of Chemistry*"), St. Petersburg, Tovarishchestvo Obshchestvennaya Polza, 1869-1871, his famous periodic table of elements – one of the greatest human achievements. ★ Gilbert Newton Lewis in the paper "*The Atom and the Molecule*," published in the *Journal of the American Chemical Society*, 38(1916)762, and Walter Kossel in the article "*Über die Molekülbildung als Frage des Atombaus*," published in *Annalen der Physik*, 49(1916)229, introduced such important algorithmic theoretical tools as the octet rule and stressed the importance of the noble gas electronic configurations. ★ As soon as quantum mechanics was formulated in 1926, Douglas R. Hartree published several papers in the *Proceedings of the Cambridge Philosophical Society*, 24(1927)89, 24(1927)111, 26(1928)89, entitled "*The Wave Mechanics of an Atom with a Non-Coulomb Central Field*," containing the computations for large atoms such as Rb and Cl. These were *self-consistent ab initio*[2] computations, and the wave function was assumed to be the *product* of spin orbitals. ★ The LCAO approximation (for the solid state) was introduced by Felix Bloch in his PhD thesis "*Über die Quantenmechanik der Elektronen in Kristallgittern*," University of Leipzig, 1928, and three years later Erich Hückel used this method to describe the first molecule (benzene) in a publication "*Quantentheoretische Beitrage zum Benzolproblem. I. Die Elektronenkonfiguration des Benzols*," which appeared in *Zeitschrift für Physik*, 70(1931)203. ★ Vladimir

---

[1] In physics and chemistry we are seldom interested in the total energy. The energy differences of various states are of importance. Sometimes such precision is not enough, but the result speaks for itself.

[2] That is, derived from the first principles of (nonrelativistic) quantum mechanics! Note that these young people worked incredibly fast (no email, no PCs).

Fock introduced the antisymmetrization of the spin orbital product in his publication "*Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems*," in *Zeitschrift für Physik*, 61(1930)126 and *ibid*. 62(1930)795. ★ John Slater proposed the idea of the multiconfigurational wave function ("*Cohesion in Monovalent Metals*," *Physical Review*, 35(1930)509). ★ The Hartree–Fock method in the LCAO approximation was formulated by Clemens C.J. Roothaan in his work "*New Developments in Molecular Orbital Theory*," published in the *Reviews of Modern Physics*, 23(1951)69, and, independently, by George G. Hall in the paper "*The Molecular Orbital Theory of Chemical Valency*," in *Proceedings of the Royal Society (London) A*, 205(1951)541. ★ The physical interpretation of the orbital energies in the Hartree–Fock method was given by Tjalling C. Koopmans in his *only* quantum chemical paper "*On the Assignment of Wave Functions and Eigenvalues to the Individual Electron of an Atom*," published in *Physica*, 1(1933/1934)104. ★ The first localized orbitals (for the methane molecule) were computed by Charles A. Coulson despite the difficulties of war time (*Transactions of the Faraday Society*, 38(1942)433). ★ Hideo Fukutome, first in *Progress in Theoretical Physics*, 40(1968)998, and then in several following papers, analyzed general solutions for the Hartree–Fock equations from the symmetry viewpoint, and showed exactly eight classes of such solutions.

* * *

In the previous chapter the motion of nuclei was considered. In the Born–Oppenheimer approximation (Chapter 6) the motion of the nuclei takes place in the potential, which is the electronic energy of a system (being a function of the nuclei position, **R**, in the configurational space). The electronic energy $E_k^0(\mathbf{R})$ is an eigenvalue given in Eq. (6.8) (adapted to the polyatomic case, hence $R \rightarrow \mathbf{R}$): $\hat{H}_0\psi_k(\mathbf{r};\mathbf{R}) = E_k^0(\mathbf{R})\psi_k(\mathbf{r};\mathbf{R})$. We will now deal exclusively with this equation, i.e., we will consider the electronic motion at fixed positions of the nuclei (clamped nuclei). Thus, our goal is two-fold: we are interested in what the electronic structure looks like and in how the electronic energy depends on the positions of the nuclei.[3]

Any theoretical method applicable to molecules may be also used for atoms, although very accurate wave functions, even for simple atoms, are not easy to calculate.[4] In fact for atoms we know the solutions quite well only in the mean-field approximation, i.e., the atomic orbitals (AOs). Such orbitals play an important role as building blocks of many-electron wave functions.

---

[3] Due to tradition in the previous chapter the ground-state electronic energy $E_0^0(\mathbf{R})$ was denoted as $V(\mathbf{R})$.

[4] If an atom is considered in the Born–Oppenheimer approximation, the problem is even simpler, the electronic equation also holds; we can then take, e.g., $\mathbf{R} = \mathbf{0}$. People still try to compute correlated wave functions (i.e., beyond the mean-field approximation, see Chapter V2-2) for heavier atoms. Besides, relativistic effects (see Chapter 3) play increasingly important roles for such atoms. Starting with magnesium, they are larger than the correlation corrections. Fortunately, the relativistic corrections for atoms are largest for the inner electronic shells, which are least important for chemists.

)*+,)))-) C
&" J J, J /J? J 6"
'

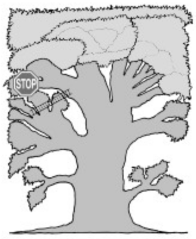

CHAPTER 1

Electronic Orbital Interactions in Periodic Systems

Beauty of style and harmony and grace and good rhythm depend on simplicity.
Plato

Where are we?

We are on the upper left branch of the TREE.

An example

Polyacetylene[1] represents a *practically* infinite polymeric chain: ⋯–CH=CH–CH=CH–CH=CH–CH=CH–⋯. There is no such a thing in Nature as a truly infinite system.[2] Yet, if we examine larger and larger portions of a homogeneous material, we come to the idea that such quantities as energy per stoichiometric unit, electron excitation energy, vibrational frequencies, etc., depend less and less on system size. This means that a boundary region (polymer ends, crystal surface) contribution to these quantities becomes negligible. Therefore, these (known as *intensive*) quantities attain limit values identical to those for an infinite system. *It pays to investigate the infinite system, because we can use its translational symmetry to simplify its description.* Well, this is what this chapter is all about.

---

[1] The discovery of conducting polymers (like polyacetylene) was highlighted by the Nobel Prize 2000 for Hideki Shirakawa, who synthesized a crystalline form of polyacetylene, as well as Allan G. MacDiarmid and Allan J. Heeger, who increased its electric conductivity by 18 orders of magnitude by doping the crystal with some electron acceptors and donors. This incredible increase is probably the largest known to humanity in any domain of experimental sciences (H. Shirakawa, E.J. Louis, A.G. MacDiarmid, C.K. Chiang, A.J. Heeger, *Chem. Soc. Chem. Commun.*, 578(1977)).

[2] That is, a macromolecule. The concept of polymer was introduced to chemistry by Herman Staudinger.

Ideas of Quantum Chemistry
https://doi.org/10.1016/B978-0-44-464248-6.00009-0
Copyright © 2020 Elsevier B.V. All rights reserved.

100th Anniversary
of the Nobel Prizes
Hermann Staudinger (Germany)
Chemistry, 1953
UGANDA 300/-

Herman Staudinger (1881–1965), German polymer chemist, professor at the University of Freiburg, received the Nobel Prize in 1953 "*for his discoveries in the field of macromolecular chemistry*." However strange it may sound now, as late as in 1926 the concept of polymers was unthinkable in chemistry. It will be encouraging for PhD students to read that a professor advised Staudinger in the late 1920s: "*Dear colleague, leave the concept of large molecules well alone: organic molecules with a molecular weight above 5000 do not exist. Purify your products, such as rubber, then they will crystallize and prove to be lower molecular substances.*"

What is it all about?

Primitive lattice (■) p. 6

Wave vector (■) p. 8

Inverse lattice (■) p. 11

First Brillouin zone (FBZ) (■) p. 14

Properties of the FBZ (■) p. 15

A few words on Bloch functions (■) p. 15

- Waves in 1D
- Waves in 2D

Infinite crystal as a limit of a cyclic system (■) p. 22

- Origin of the band structure
- Born–von Kármán condition in 1D
- *k* dependence of orbital energy

A triple role of the wave vector (■) p. 26

Band structure (■) p. 26

- Born–von Kármán boundary condition in 3D
- Crystal orbitals from Bloch functions (LCAO CO method)
- SCF LCAO CO equations
- Band width
- Fermi level and energy gap: insulators, metals, and semiconductors

Solid state quantum chemistry (■) **p. 40**

- Why do some bands go up?
- Why do some bands go down?
- Why do some bands stay constant?
- More complex behavior explainable – examples

The Hartree–Fock method for crystals (■) **p. 53**

- Secular equation
- Integration in the FBZ
- Fock matrix elements
- Iterative procedure (SCF LCAO CO)
- Total energy

Long-range interaction problem (■) **p. 58**

- Fock matrix corrections
- Total energy corrections
- Multipole expansion applied to the Fock matrix
- Multipole expansion applied to the total energy

Back to the exchange term (■) **p. 68**

Choice of unit cell (✠) **p. 71**

- Field compensation method
- The symmetry of subsystem choice

If a motif (e.g., a cluster of atoms) associated with a unit cell is *regularly translated along three different directions in space*, we obtain an infinite three-dimensional periodic structure (translational symmetry). In 2D (Fig. 1.1a) this means that having a single (special) type of tiles (unit cells) we have been successful in tiling the *complete* two-dimensional space. One of the consequences is that the five-fold symmetry axes have to be absent in the atomic arrangements in such crystals. *It turned out that a vast majority of real crystals can be reliably modeled using this idea.* There were efforts in mathematics to design some nontranslational complete tilings. In 1963 it was first shown that for a number $N = 20\,000$ of square tiles such a nontranslational tiling is possible. This number has been gradually reduced and in 1976 Roger Penrose proposed covering by $N = 2$ kinds of tiles (see Fig. 1.1b). Then, Daniel Shechtman discovered[3] that there are substances (known now as quasicrystals) that indeed show five-fold symmetry axes (also other translationally forbidden symmetry axes). The reason why quasicrystals exist in Nature is quite simple: some strong short-range interactions force unusual five-ligand complexes. It is remarkable that the discovery of quasicrystals has been preceded by ancient artists (Fig. 1.1c).

When applying the Hartree–Fock method to such periodic infinite objects one exploits the translational symmetry of the system, e.g., in calculating integrals. It would be indeed prodigal to compute the integrals many times, the equality of which is *guaranteed* by translational symmetry. When translational symmetry is taken into account, the problem reduces to the calculation of the interaction of a single unit

---

[3] He received the 2011 chemistry Nobel Prize for “the discovery of quasicrystals.”

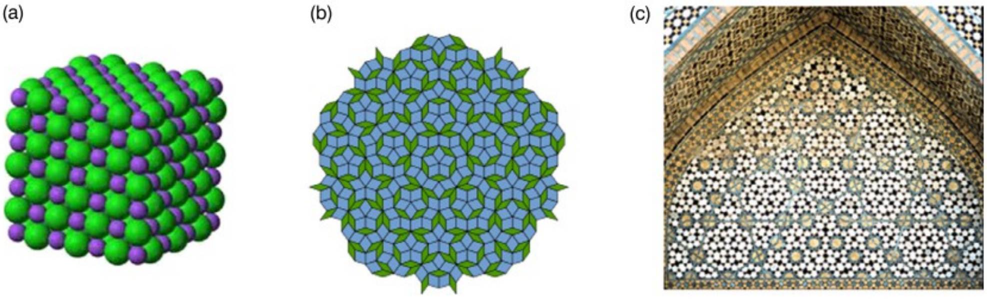

Fig. 1.1. Translational symmetry in crystals and its lack in the quasicrystals despite a perfect long-range order. (a) Translational symmetry in the NaCl crystal build of Na^+ and Cl^- ions. (b) The Penrose tiling as an example of a quasicrystal, no translational symmetry. (c) A medieval Arabian mosaic as an example of a long-range nontranslational order.

cell (reference, labeled by 0) with all other unit cells, the nearest neighbor cells being most important. The infinite size of the system is hidden in the plethora of points (to be taken into account) in what is known as the first Brillouin zone (FBZ). The FBZ represents a unit cell in what is called inverse lattice (associated with a given lattice reflecting the translational symmetry).

The electronic orbital energy becomes a function of the FBZ points and we obtain what is known as a band structure of energy levels. This band structure decides the electronic properties of the system (insulator, semiconductor, metal). We will also show how to carry out the mean-field (Hartree–Fock) computations on infinite periodic systems. The calculations require infinite summations (interaction of the reference unit cell with the infinite crystal) to be made. This creates some mathematical problems, which will be also described in the present chapter.

Why is this important?

The present chapter is particularly important for those readers who are interested in solid state physics and chemistry. Others may treat it as exotic and, if they decide they do not like exotic matter, may go directly to other chapters.

The properties of a polymer or a crystal sometimes differ very widely from those of the atoms or molecules of which they are built. *The same substance* may form *different* periodic structures, which have *different* properties (e.g., graphite and diamond). The properties of periodic structures could be computed by extrapolation of the results obtained for larger and larger clusters of the atoms from which the substance is composed. This avenue is however noneconomic. It is easier to carry out quantum mechanical calculations for an infinite system[4] than for a large cluster.[5]

---

[4] The surface effects can be neglected and the units the system is composed of are treated as equivalent.

[5] Sometimes we may be interested in a particular cluster, not in an infinite system. Then it may turn out that it is more economic to perform the calculations for the infinite system, and use the results in computations for the clusters (e.g., R.A. Wheeler, L. Piela, R. Hoffmann, *J. Am. Chem. Soc.*, 110(1988)7302).

What is needed?

- Operator algebra (Appendix V1-B, p. V1-595, necessary),
- translation operator (Appendix V1-C, p. V1-605, necessary),
- Hartree–Fock method (Chapter V1-8, necessary),
- multipole expansion (Appendix G, p. 613, advised),
- matrix diagonalization (Appendix V1-L, p. V1-703, advised).

Classical works

At the age of 23, Felix Bloch published an article "*Über die Quantenmechanik der Elektronen in Kristallgittern*" in *Zeitschrift für Physik*, 52(1928)555 (only two years after Schrödinger's historic publication) on the translational symmetry of the wave function. This has also been the first application of LCAO expansion. ★ A book appeared in 1931 by Leon Brillouin entitled *Quantenstatistik* (Springer Verlag, Berlin, 1931), in which the author introduced some of the fundamental notions of band theory. ★ The first *ab initio* calculations for a polymer were carried out by Jean-Marie André in the paper "*Self-Consistent Field Theory for the Electronic Structure of Polymers*," published in the *Journal of the Chemical Physics*, 50(1969)1536.

* * *

1.1 Primitive lattice

Let us imagine an *infinite crystal*, e.g., a system that exhibits the 3D *translational symmetry* of the charge distribution (nuclei and electrons). The translational symmetry will be fully determined by three (linearly independent) basis vectors[6]: $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, and $\boldsymbol{a}_3$, having the property that $\boldsymbol{a}_i$ beginning at any atom extends to the corresponding nearest-neighbor identical atom located in the crystal. The lengths of the basis vectors $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, and $\boldsymbol{a}_3$ are called the *lattice constants* along the three periodicity axes.[7]

A lot of such basis sets are possible. Any choice of basis vectors is acceptable from the point of view of mathematics. For economic reasons we choose one of the possible vector sets that give the *least volume parallelepiped*[8] with sides $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, and $\boldsymbol{a}_3$. This parallelepiped (arbitrarily

---

[6] Not necessarily perpendicular though; they determine the periodicity axes.

[7] As shown on p. V1-514, a symmetry of the nuclear framework does not guarantee the same symmetry of the electronic charge distribution computed using a mean-field method. We may cope with the period doubling as compared to the period of the nuclear framework (cf. BOAS, p. V1-514). If this happens, then we should choose such lattice constants that ensure the periodicity of both nuclear and electron distributions.

[8] We are, however, interested in the smallest unit cell.

Alexandr Alexandrovitch Friedmann (1888–1925), Russian mathematician and physicist, in his article in *Zeit. Phys.*, 10(1922)377 proved on the basis of Einstein's general theory of relativity that the curvature of the Universe must change, which became the basis of cosmological models of the expanding Universe. During the First World War, Friedman was a pilot in the Russian army and made bombing raids over my beloved Przemyśl.

In one of his letters he asked his friend cheerfully, the eminent Russian mathematician Steklov, for advice about the integration of equations he derived to describe the trajectories of his bombs. Later, in a letter to Steklov of February 28, 1915 he wrote: "*Recently I had an opportunity to verify my theory during a flight over Przemyśl, the bombs fell exactly in the places predicted by the theory. To get the final proof of my theory I intend to test it in flights during the next few days*."

More information can be found on http://www-groups.dcs.st-and.ac.uk/~history/Mathematicians/Friedmann.html.

2.4 Exponentially correlated Gaussian functions

In 1960, Boys[28] and Singer[29] noticed that the functions which are products of Gaussian orbitals and correlational factors of Gaussian type, $\exp(-br_{ij}^2)$, where r_{ij} is the distance between electron i and electron j, generate relatively simple integrals in quantum chemical calculations. A product of two Gaussian orbitals with positions shown by the vectors $\mathbf{A}, \mathbf{B}$ and of an exponential correlation factor is called *an exponentially correlated Gaussian geminal*,[30] i.e.,

$$g\left(\mathbf{r}_i, \mathbf{r}_j; \mathbf{A}, \mathbf{B}, a_1, a_2, b\right) = N e^{-a_1(\mathbf{r}_i-\mathbf{A})^2} e^{-a_2\left(\mathbf{r}_j-\mathbf{B}\right)^2} e^{-br_{ij}^2}.$$

A geminal represents a generalization of an orbital – there we have a one-electron function, here we have a two-electron one. A single geminal is very rarely used in computations[31]; we apply hundreds or even thousands of Gaussian geminals. When we want to find out what the optimal positions $\mathbf{A}, \mathbf{B}$ and the optimal exponents a and b are in these thousands of geminals, it turns out that nothing sure is known about them; the $\mathbf{A}, \mathbf{B}$ positions are scattered chaotically,[32] and in the $a > 0$ and $b > 0$ exponents there is no regularity either.

---

[28] S.F. Boys, *Proc. Royal Soc. A*, 258(1960)402.

[29] K. Singer, *Proc. Royal Soc. A*, 258(1960)412.

[30] This is an attempt to go beyond the two-electron systems with the characteristic for these systems approach of James, Coolidge, Hylleraas, Kołos, Wolniewicz, and others.

[31] Ludwik Adamowicz introduced an idea of the minimal basis of the Gaussian geminals (equal to the number of the electron pairs) and applied to the LiH and HF molecules; L. Adamowicz, A.J. Sadlej, *J. Chem. Phys.*, 69(1978)3992.

Edwin Powell Hubble (1889–1953), American astronomer, explorer of galaxies, found in 1929 that the distance between galaxies is proportional to the infrared shift in their spectrum caused by the Doppler effect, which is consequently interpreted as expansion of the Universe. A surprise from recent astronomical studies is that the expansion is faster and faster (for unknown reasons).

EDWIN HUBBLE
ASTRONOMER
usa 41

2.5 Electron holes

2.5.1 Coulomb hole ("correlation hole")

It is always good to count "on fingers" to make sure that everything is all right. Let us see how a *single* Gaussian geminal describes the correlation of the electronic motion. Let us begin with the helium atom with the nucleus in the position $\mathbf{A} = \mathbf{B} = \mathbf{0}$. The geminal takes the form

$$g_{He} = Ne^{-a_1 r_1^2} e^{-a_1 r_2^2} e^{-b r_{12}^2}, \tag{2.18}$$

where $N > 0$ is a normalization factor. Let us assume[33] that electron 1 is at $(x_1, y_1, z_1) = (1, 0, 0)$. Where does electron 2 prefer to be in such situation? We will find it out (Fig. 2.2) from the position of electron 2 for which g_{He} assumes the largest value.

Just to get an idea, let us try to restrict the motion of electron 2. For instance, let us demand that it moves only on the sphere of radius equal to 1 centered at the nucleus, so we insert $r_1 = r_2 = 1$. Then, $g_{He} = const \exp\left[-b r_{12}^2\right]$ and we will find out easily what electron 2 likes most. With $b > 0$ the latter factor tells us that what electron 2 likes best is just to sit on electron 1! Is it what the correlation is supposed to mean that one electron sits on the other? Here we have rather an anticorrelation. Something is going wrong. According to this analysis we should rather take the geminal of, e.g., the form

$$g_{He} = Ne^{-a_1 r_1^2} e^{-a_1 r_2^2} \left[1 - e^{-b r_{12}^2}\right].$$

---

[32] The methods in which those positions are selected at random scored great success.

[33] We use atomic units.

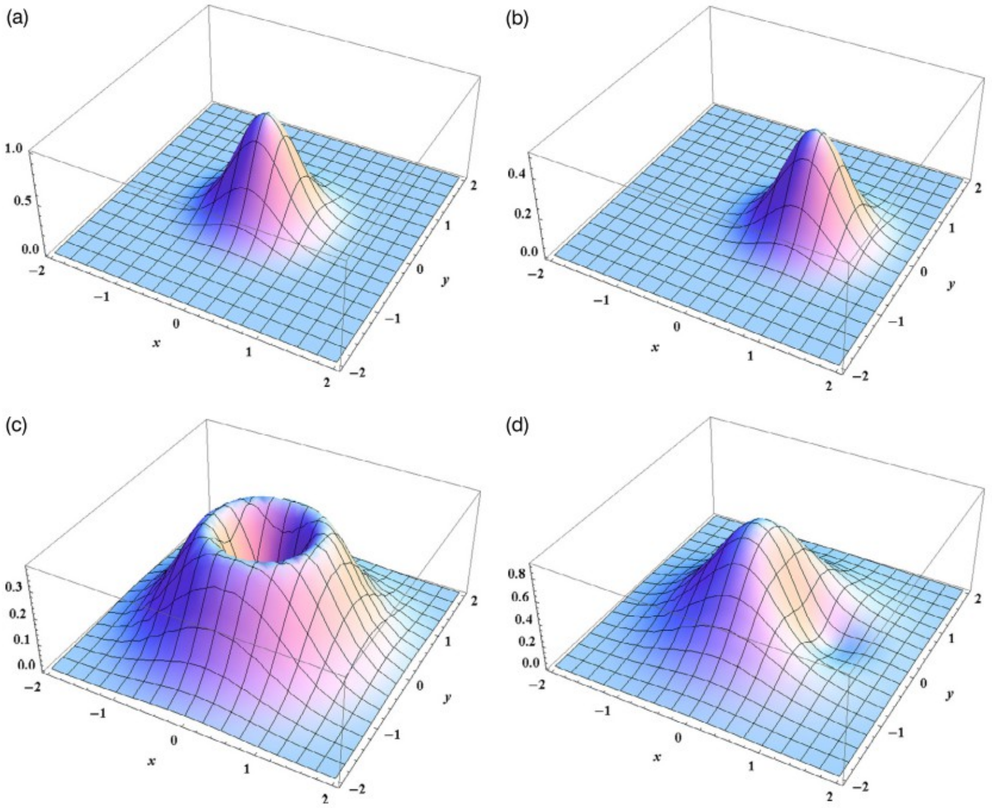

Fig. 2.2. Illustration of the correlation and anticorrelation of the electrons in a model helium atom. The machinery of the "anticorrelation" connected with the (not normalized) geminal $g_{He} = \exp\left[-r_1^2\right]\exp\left[-r_2^2\right]\exp\left[-2r_{12}^2\right]$. (a) Electron 1 has position $(0,0,0)$. (b) Electron 1 is at point $(1,0,0)$. It can be seen that electron 2 *holds on to electron 1*, i.e., it behaves in a completely unphysical manner (since the electrons repel each other). (c,d) Electron 2 responds to such two positions of electron 1 in case the wave function represents the geminal $g_{He} = \exp\left[-r_1^2\right]\exp\left[-r_2^2\right]\left[1-\exp\left[-2r_{12}^2\right]\right]$. (c) Electron 2 runs away (the hollow in the middle) from electron 1 placed at $(0,0,0)$, and we have correlation. (d) Similarly, if electron 1 is in point $(1,0,0)$, then it causes a slight depression for electron 2 in this position, and we have correlation. However, the graph is different from the one in (c). This is understandable since the nucleus is all the time in the point $(0,0,0)$. (e,f) The same displacements of electron 1, but this time the wave function is equal to $\psi\left(\mathbf{r}_1,\mathbf{r}_2\right) = \left(1+\frac{1}{2}r_{12}\right)\exp[-\left(r_1^2+r_2^2\right)]$, i.e., is similar to the wave function of the harmonic helium atom. It can be seen (particularly in (e)) that there *is* a correlation, although much less visible than in the previous examples. (g,h) The same as in (e,f), but (to amplify [artificially] the correlation effect) for the function $\psi\left(\mathbf{r}_1,\mathbf{r}_2\right) = (1+25r_{12})\exp[-\left(r_1^2+r_2^2\right)]$, which (unlike (e,f)) does not satisfy the correlation cusp condition.

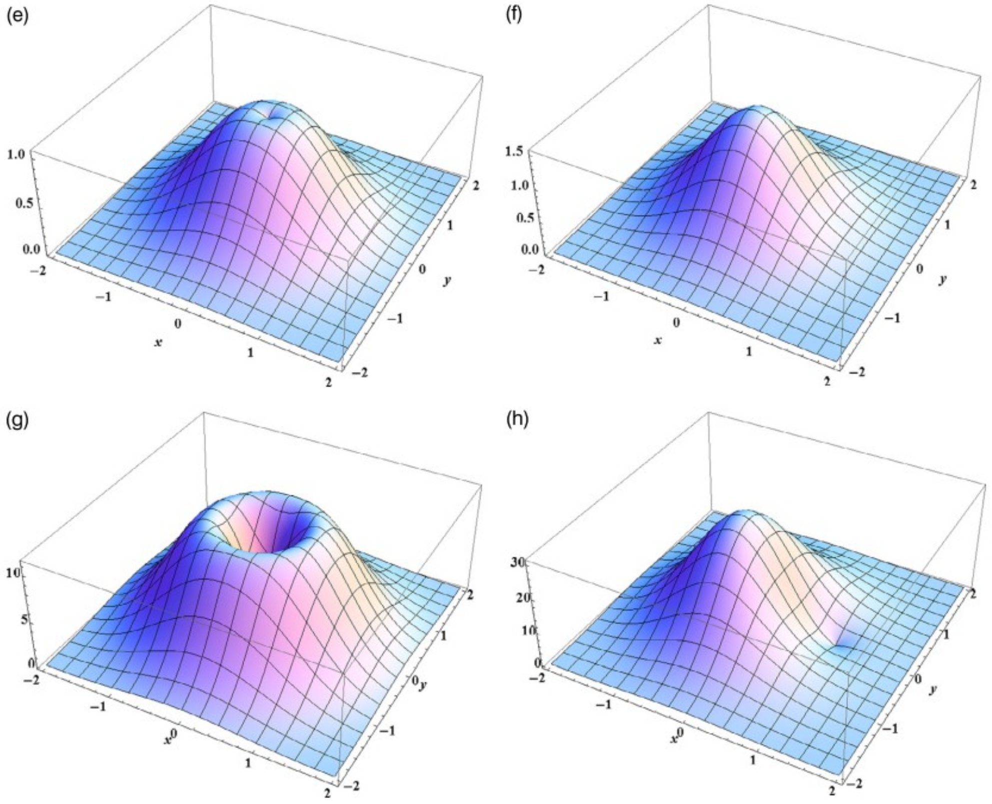

Fig. 2.2. (*continued*)

Now everything is qualitatively in order. When the interelectronic distance increases, the value of the g_{He} function also increases, which means that such a situation is *more* probable than that corresponding to a short distance. If the electrons become too agitated and begin to think that it would be better when their distance gets very large, they would be called to order by the factors $\exp\left[-a_1 r_1^2\right]\exp\left[-a_1 r_2^2\right]$. Indeed, in such a case, the distance between the nucleus and at least one of the electrons is long and the probability of such a situation is quenched by one or both exponential factors. For large r_{12} distances, the factor $\left[1-\exp\left[-b r_{12}^2\right]\right]$ is practically equal to 1. This means that for large interelectronic distances g_{He} is practically equal to $N\exp\left[-a_1 r_1^2\right]\exp\left[-a_1 r_2^2\right]$, i.e., to the product of the orbitals (no correlation of motions at long interelectronic distances and rightly so).

Around electron 1 there is a region of low probability of finding electron 2. This region is called the Coulomb hole.

The Gaussian geminals do not satisfy the correlation cusp condition (p. 92), because of the factor $\exp(-br_{ij}^2)$. It is required (for simplicity we write $r_{ij} = r$) that $(\frac{\partial g}{\partial r})_{r=0} = \frac{1}{2} g(r = 0)$, whereas the left-hand side is equal to 0, while the right-hand side $\frac{1}{2} N \exp\left[-a_1 \left(\mathbf{r}_i - \mathbf{A}\right)^2\right]$ $\exp\left[-a_2 \left(\mathbf{r}_j - \mathbf{B}\right)^2\right]$ is not equal to zero. This is not a disqualifying feature, since the region of space in which this condition should be fulfilled is very small.

The method of Gaussian geminals has been applied in extremely accurate calculations but for three- and four-electron systems only.[34]

2.5.2 Exchange hole ("Fermi hole")

The mutual avoidance of electrons in a helium atom or in a hydrogen molecule is caused by Coulombic repulsion of electrons ("Coulomb hole," see above). As we have shown in this chapter, in the Hartree–Fock method the Coulomb hole is absent, whereas methods which account for electron correlation generate such a hole. However, electrons avoid each other not only because of their charge. The Pauli principle is additional reason. One of the consequences is the fact that electrons with the same spin coordinate cannot reside in the same place (see p. V1-40). The continuity of the wave function implies that the probability density of them staying *in the vicinity* of each other is small, i.e.,

around the electron there is a NO PARKING area for other electrons with the same spin coordinate ("exchange hole" or "Fermi hole").

Let us see how such exchange holes arise. We will try to make the calculations as simple as possible.

We have shown above that the Hartree–Fock function does not include any electron correlation. We must admit, however, that we have come to this conclusion on the basis of the two-electron closed shell case. This is a special situation, since both electrons have *different* spin coordinates ($\sigma = \frac{1}{2}$ and $\sigma = -\frac{1}{2}$). Is it really true that the Hartree–Fock function does not include any correlation of electronic motion?

---

[34] W. Cencek, PhD Thesis, Adam Mickiewicz University, Poznań, 1993; also J. Rychlewski, W. Cencek, J. Komasa, *Chem. Phys. Letters*, 229(1994)657; W. Cencek, J. Rychlewski, *Chem. Phys. Letters*, 320(2000)549.

For each point on the IRC, i.e., on the curve $\mathbf{x}_{\mathrm{IRC}}(s)$, we may read the mass-weighted coordinates, and use them to calculate the coordinates of each atom. Therefore, each point on the IRC corresponds to a certain spatial and electronic structure of the system.

6.5 Reaction path Hamiltonian method

6.5.1 Energy close to IRC

A hypersurface of the potential energy represents an expensive product. We first have to calculate the potential energy for a grid of points. If we assume that ten points per coordinate is a sufficient number, then we have to perform 10^{3N-6} advanced quantum mechanical calculations; for $N = 10$ atoms this gives 10^{24} calculations, which is an impossible task. Now you see why specialists like three-atomic systems so much.

Are all the points necessary? For example, if we assume low energies, the system will in practice stay close to the IRC. Why, therefore, worry about other points? This idea was exploited by Miller, Handy, and Adams.[26] They decided to introduce the coordinates that are natural for the problem of motion in the reaction "drain-pipe." The approach corresponds to point 4 on p. 447.

The authors derived the

REACTION PATH HAMILTONIAN

an approximate expression for the energy of the reacting system in the form that stresses the existence of the IRC and of deviations from it.

This formula (*Hamilton function of the reaction path*) has the following form:

$$H(s, p_s, \{Q_k, P_k\}) = T(s, p_s, \{Q_k, P_k\}) + V(s, \{Q_k\}), \tag{6.26}$$

where T is the kinetic energy, V stands for the potential energy, s denotes the reaction coordinate along the IRC, $p_s = \frac{ds}{dt}$ represents the momentum coupled with s (mass = 1), and $\{Q_k\}$, $k = 1, 2, ...3N-7$, stand for other coordinates orthogonal to the reaction path $\mathbf{x}_{IRC}(s)$ (this is why Q_k will depend on s) and the momenta $\{P_k\}$ conjugated with them.

We obtain the coordinates Q_k in the following way. At point s of the reaction path we diagonalize the Hessian, i.e., the matrix of the second derivatives of the potential energy, and consider all the resulting normal modes ($\omega_k(s)$ are the corresponding frequencies, cf. Chapter V1-7)

---

[26] W.H. Miller, N.C. Handy, J.E. Adams, *J. Chem. Phys.,* 72(1980)99.

other than that which corresponds to the reaction coordinate s (the latter corresponds to the "imaginary"[27] frequency ω_k). The diagonalization also gives the orthonormal vectors $\mathbf{L}_k(s)$, each having a direction in the $(3N-6)$-dimensional configurational space (the mass-weighted coordinate system). The coordinate $Q_k \in (-\infty, +\infty)$ measures the displacement along the direction of $\mathbf{L}_k(s)$. The coordinates s and $\{Q_k\}$ are called the *natural coordinates*. To stress that Q_k is related to $\mathbf{L}_k(s)$, we will write it as $Q_k(s)$.

The potential energy close to the IRC can be approximated (*harmonic approximation*, cf. p. V1-218) by

$$V(s, \{Q_k\}) \cong V_0(s) + \frac{1}{2}\sum_{k=1}^{3N-7} \omega_k(s)^2 Q_k(s)^2, \tag{6.27}$$

where the term $V_0(s)$ represents the potential energy that corresponds to the bottom of the reaction "drain-pipe" at a point s along the IRC, while the second term tells us what will happen to the potential energy if we displace the point (i.e., the system) perpendicular to $\mathbf{x}_{IRC}(s)$ along all the normal oscillator coordinates. In the *harmonic approximation* for the oscillator k, the energy goes up by half the force constant × the square of the normal coordinate Q_k^2. The force constant is equal to ω_k^2, because the mass is equal to 1.

The kinetic energy turns out to be more complicated. We have

$$T(s, p_s, \{Q_k, P_k\}) = \frac{1}{2}\frac{[p_s - \sum_{k=1}^{3N-7}\sum_{k'=1}^{3N-7} B_{kk'} Q_{k'} P_k]^2}{[1 + \sum_{k=1}^{3N-7} B_{ks} Q_k]^2} + \sum_{k=1}^{3N-7} \frac{P_k^2}{2}. \tag{6.28}$$

The last term is recognized as the vibrational kinetic energy for the independent oscillations perpendicular to the reaction path (recall that the mass is treated as equal to 1). If in the first term we insert $B_{kk'} = 0$ and $B_{ks} = 0$, the term would be equal to $\frac{1}{2}p_s^2$ and, therefore, would describe the kinetic energy of a point moving as if the reaction coordinate were a straight line.

---

[27] For large $|s|$ the corresponding ω^2 is close to zero. When $|s|$ decreases (we approach the saddle point), ω^2 becomes negative (i.e., ω is imaginary). For simplicity we will call this the "imaginary frequency" for any s.

> CORIOLIS AND CURVATURE COUPLINGS
>
> - $B_{kk'}$ are called the Coriolis coupling constants. They couple the normal modes perpendicular to the IRC.
> - The B_{ks} are called the *curvature coupling constants*, because they would be equal to zero if the IRC was a straight line. They couple the translational motion along the reaction coordinate with the vibrational modes orthogonal to it.
> - All the above coupling constants B depend on s.

Therefore, in the reaction path Hamiltonian we have the following quantities that characterize the reaction "drain-pipe":

- the reaction coordinate s that measures the progress of the reaction along the "drain-pipe,"
- the value $V_0(s) \equiv V_0(\mathbf{x}_{IRC}(s))$ represents the energy that corresponds to the bottom of the "drain-pipe"[28] at the reaction coordinate s,
- the width of the "drain-pipe" is characterized[29] by a set $\{\omega_k(s)\}$,
- the curvature of the "drain-pipe" is hidden in constants B; their definition will be given later in this chapter. Coefficient $B_{kk'}(s)$ tells us how normal modes k and k' are coupled together, while $B_{ks}(s)$ is responsible for a similar coupling between reaction path $\mathbf{x}_{\mathrm{IRC}}(s)$ and vibration k perpendicular to it.

6.5.2 Vibrational adiabatic approximation

Most often when moving along the bottom of the "drain-pipe," potential energy $V_0(s)$ only changes moderately when compared to the potential energy changes the molecule undergoes when oscillating perpendicularly[30] to $\mathbf{x}_{IRC}(s)$. Simply, the valley bottom profile results from the fact that the molecule hardly holds together when moving along the reaction coordinate s, *a chemical bond breaks*, while *other bonds remain strong* and it is not so easy to stretch their strings. This suggests that there is *slow* motion along s and *fast* oscillatory motion along the coordinates Q_k.

> Since we are mostly interested in the slow motion along s, we may average over the fast motion.

---

[28] That is, the classical potential energy corresponding to the point of the IRC given by s (this gives an idea of how the potential energy changes when walking along the IRC).

[29] A small ω corresponds to a wide valley when measured along a given normal mode coordinate ("soft" vibration), a large ω means a narrow valley ("hard" vibration).

[30] That is, when moving along the coordinates Q_k, $k = 1, 2, ...3N - 7$.

-1 " ? 6

The philosophy behind the idea is that while the system moves slowly along s, it undergoes a large number of oscillations along Q_k. After such vibrational averaging the only information that remains about the oscillations are the vibrational quantum energy levels for each of the oscillators (of course, the levels will depend on s).

VIBRATIONAL ADIABATIC APPROXIMATION

The fast vibrational motions will be treated quantum mechanically and their total energy will enter the potential energy for the classical motion along s.

This approximation parallels the adiabatic approximation made in Chapter V1-6, where fast motion of electrons was separated from the slow motion of nuclei in a similar way. There after averaging over fast electronic motion, the total electronic energy became the potential energy for motion of nuclei, here after averaging over fast vibrational motion perpendicular to IRC the total vibrational energy (the energy of the corresponding harmonic oscillators in their quantum states) becomes the potential energy for the slow motion along s. This concept is called the *vibrational adiabatic approximation*.

In this approximation, to determine the stage of the reaction we give two *classical* quantities, i.e., *where* the system is on the reaction path (s) and *how fast* the system moves along the reaction path (p_s), as well as what the *quantum states* of the oscillators vibrating perpendicularly to the reaction path are (vibrational quantum number $v_k = 0, 1, 2, \ldots$ for each of the oscillators). Therefore, the potential energy for the (slow) motion along the reaction coordinate s is[31]

$$V_{adiab}(s; v_1, v_2, \ldots, v_{3N-7}) = V_0(s) + \sum_{k=1}^{3N-7} (v_k + \frac{1}{2})\hbar[\omega_k(s) - \omega_k(-\infty)], \quad (6.29)$$

where we have chosen an arbitrary additive constant in the potential as equal to the vibrational energy of the reactants (with minus sign), $-\sum_{k=1}^{3N-7}(v_k + \frac{1}{2})\hbar\omega_k(-\infty)$. Note that even though $v_k = 0$ for each of the oscillators, there is a nonzero vibrational correction to the classical potential energy $V_0(s)$, because the zero-vibrational energy changes if s changes.

---

[31] Even if (according to the vibrational adiabatic approximation) the vibrational quantum numbers were kept constant during the reaction, their energies as depending on s through ω would change.

The vibrational adiabatic potential V_{adiab} was created for a given set of the vibrational quantum numbers υ_k, fixed during the reaction process. Therefore, it is impossible to exchange energy between the vibrational modes (we assume therefore the Coriolis coupling constants $B_{kk'} = 0$), as well as between the vibrational modes and the reaction path (we assume the curvature coupling constants $B_{ks} = 0$). This would mean a change of υ_k's.

From Eq. (6.29) we may draw the following conclusion.

When during the reaction the frequency of a normal mode decreases dramatically (which corresponds to breaking of a chemical bond), the square bracket becomes negative. This means that an excitation of the bond before the reaction decreases the (vibrational adiabatic) reaction barrier and the reaction rate will increase.

As a matter of fact, this is quite obvious; a vibrational excitation that engages the chemical bond to be broken already weakens the bond before the reaction.

Why do chemical reactions proceed?

Exothermic reactions

When the reactants (products) have kinetic energy higher than the barrier and the corresponding momentum p_s is large enough, with a high probability the barrier will be overcome (cf. p. V1-205). Even if the energy is lower than the barrier there is still a nonzero probability of passing to the other side because of the tunneling effect. In both cases (passing over and under the barrier) it is easier when the kinetic energy is large.

The barrier *height* is usually different for the reactions reactants → products and products → reactants (Fig. 6.1). If the barrier height is smaller for the reactants, this *may* result in an excess of the product concentration over the reactant concentration.[32] Since the reactants have higher energy than the products, the potential energy excess will change into kinetic energy[33] of the products (which is observed as a temperature increase – the reaction is exothermic). This may happen if the system has the possibility to pump the potential (i.e., electronic) energy excess into the translational and rotational degrees of freedom or to a "third body or bodies" (through collisions, e.g., with the solvent molecules) or has the possibility to emit light quanta. If the system has no such possibilities the reaction will not take place.

---

[32] Because a lower barrier is easier to overcome.

[33] Most often rotational energy.

Endothermic reactions

The barrier height does not always decide.

> Besides the barrier height the *widths* of the entrance and exit channels also count.

For the time being let us take an example with $V_0(-\infty) = V_0(\infty)$, i.e., the barrier calculated from IRC is *the same* in both directions. Imagine a narrow entrance channel, i.e., large force constants for the oscillators, and a wide exit channel, i.e., low force constants (Figs. 6.2g,h and 6.7).

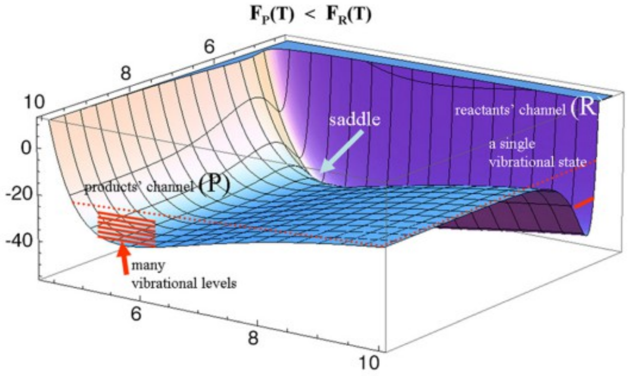

Fig. 6.7. Why do some endothermic reactions proceed spontaneously? The figure shows the reaction "drain-pipe" V. As we can see, the reactants have lower energy than the products. Yet it is not V that decides the reaction to proceed, but the free energy $F = E - TS$, where T is the temperature and S the entropy. The free energy depends on the density of the vibrational states of the reactants and products. The more numerous the low-energy vibrational levels, the larger the entropy and the lower the free energy, if $T > 0$. As we can see, the reactant vibrational levels are scarce (only one is shown), while on the product side they are densely distributed. When the energy gain related to the entropy overcomes the potential energy loss, the (endothermic) reaction will proceed spontaneously.

Note that, if the entrance channel is narrow while the exit channel is wide, the density of the vibrational states will be small in the entrance channel and large in the exit channel. Therefore, for $T > 0$ there will be a lot of possibilities to occupy the low-energy vibrational levels for the products, and only a few possibilities for the reactants. This means a high entropy of the products and a small entropy of the reactants, i.e., the products will be more stabilized by the entropy than the reactants.[34] Once again we can see that while the energy in a spontaneous endothermic reaction increases, the decisive factor is the free energy which *decreases*. The reactants → products reaction occurs “uphill” the potential energy, but “downhill” the free energy.

---

[34] It pertains to the term $-TS$ in the free energy. In such a chemistry masterpiece as biology the endothermic reactions which are not spontaneous, are often coupled with exothermic ones. In such a way the whole process may go spontaneously.

Kinetic and thermodynamic pictures

- In a macroscopic reaction carried out in a chemist’s flask we have a statistical ensemble of the systems that are in different microscopic stages of the reaction.
- The ensemble may be modeled by a reaction “drain-pipe” (we assume the barrier) with a lot of points, each representing one of the reacting systems.
- When the macroscopic reaction begins (e.g., we mix two reactants) a huge number of points appear in the entrance channel, i.e., we have the reactants only. As the reactant molecules assemble or dissociate the points appear or disappear.
- If the barrier were high (no tunneling) and temperature $T = 0$, all the reactants would be in their zero-vibrational states[35] and in their ground rotational states. This state would be stable even when the product valley corresponded to a lower energy (this would be a metastable state).
- Raising the temperature causes some of the ensemble elements in the entrance channel to acquire energy comparable to the barrier height. Those elements might have a chance to pass the barrier either by tunneling (for energy smaller than the barrier) or by going over the barrier. Not all of the elements with sufficient energies would pass the barrier, but only those with reactive trajectories (corresponding to a proper orientation and kinetic energy).
- After passing the barrier the energy is conserved, but in case of exothermic reactions the excess of energy changes into the translational, vibrational, and rotational energy of products or may be transferred to a “third body” (e.g., the solvent molecules) or changed into electromagnetic radiation. The latter if in the infrared region means a hot testtube – typical for the exotermic reactions. On the other hand, a spontaneous endotermic reaction takes the necessary energy from the neighborhood, which means cooling or even freezing (like when producing icecream).
- The probability of the reactive trajectories might be calculated in a way similar to that described in Chapter V1-4 (tunneling[36]), with additionally taking into account the initial vibrational and rotational states of the reactants as well as averaging over the energy level populations.
- The products would also have a chance to pass the barrier back to the reactant side (Nature does not know what we call reactants), but at the beginning the number of the elements

---

[35] Please note that even in this case ($T = 0$) the energy of these states would not only depend on the bottom of the valley, $V_0(s)$, but also on the valley’s width through $\omega_k(s)$, according to Eq. (6.29).

[36] See also H. Eyring, J. Walter, G.E. Kimball, “*Quantum Chemistry*,” John Wiley, New York, 1967.

passing the barrier in the reactant-to-product direction would be larger (nonequilibrium state).

- However, the higher the product concentration, the more often the products transform into the reactants. As an outcome we arrive at the *thermodynamic equilibrium state*, in which the average numbers of the elements passing the barrier per unit time in either direction are equal.
- If the barrier is high and the energies considered are low, then the stationary states of the system could be divided into those (of energy $E_{i,R}$) which have high amplitudes in the entrance channel (the reactant states) and those (of energy $E_{i,P}$) with high amplitudes in the exit channel (product states). In such a case we may calculate the partition function for the reactants as (we choose the same reference energy)

$$Z_R(T) = \sum_i g_i \exp\left(-\frac{E_{i,R} - E_{0,R}}{k_B T}\right)$$

and for the products as

$$Z_P(T) = \sum_i g_i \exp\left(-\frac{E_{i,P} - E_{0,R}}{k_B T}\right) = \sum_i g_i \exp\left(-\frac{E_{i,P} - E_{0,P} + \Delta E}{k_B T}\right),$$

where g_i stands for the degeneracy of the i-th energy level and the difference of the ground-state levels is $\Delta E = E_{0,P} - E_{0,R}$.

- Having the partition functions, we may calculate (at a given temperature and volume and a fixed number of particles[37]) the free or Helmholtz energy (F) corresponding to the entrance and exit channels (in thermodynamic equilibrium) as follows:

$$F_R(T) = -k_B T \frac{\partial}{\partial T} \ln Z_R(T), \tag{6.30}$$

$$F_P(T) = -k_B T \frac{\partial}{\partial T} \ln Z_P(T). \tag{6.31}$$

- The reaction goes in such a direction as to attain the minimum free energy F.
- The higher the density of states in a given channel (this corresponds to higher entropy), the lower F. The density of the vibrational states is higher for wider channels (see Fig. V1-4.16).

---

[37] Similar considerations may be performed for a constant pressure (instead of volume). The quantity that then attains the minimum at the equilibrium state is the Gibbs potential G.

data 29.04.2020
S T Q Q S S D

156) 20200429

A) Books_74_2020/
Ideas of Quantum Chemistry/

A1) Ideas_of_Quantum_Chemistry_
vol_1.pdf

Pag. 41(73)

1.3 The Heisenberg uncertainty principle.

Pag. 47(79)

1.5 How to disprove the Heisenberg principle? The Einstein-Podolsky-Rosen's recipe.

Pag. 72(104)

2.1 Symmetry of the nonrelativistic Hamiltonian and the conservation laws.

Pag. 401(429)

7.1.1 Coordinate system and Hamiltonian

Pag. 457(484)

Orbital Model of Electronic Motion

Jandaia

29.04.2020

in Atoms and Molecules

A2) Ideas_of_Quantum_Chemistry-Vol-2.pdf

Pag. 01(25)

Electronic Orbital Interactions in Periodic Systems

Pag. 101(125)

2.4 Exponentially correlated Gaussian functions

Pag. 463(486)

6.5 Reaction path Hamiltonian method.

B) Books_74_2020/
Field Guide to Quantum Mechanics/
Field_Guide_to_Quantum_Mechanics.pdf

Pag. 43(63)

Schrödinger, Heisenberg, and Interaction Pictures.

C) Books_74_2020/
Higher Engineering Mathematics by Bandary Ramana/

Schrödinger, Heisenberg, and Interaction Pictures

The terms **Schrödinger picture**, **Heisenberg picture**, and **interaction picture** are given to three particular frames of reference. The pictures are distinguished from one another by the specific time-dependent unitary transformations involved in defining the different reference frames.

In the Schrödinger picture, state vectors evolve in time under the action of the Hamiltonian according to the Schrödinger equation and postulate 3 on page 11. If a time evolution operator $\hat{\mathbb{U}}(t, t_0)$ and an initial state vector $|\Psi_S(t_0)\rangle$ are known for a given system, then the state vector dynamics in the Schrödinger picture are found via $|\Psi_S(t)\rangle = \hat{\mathbb{U}}(t, t_0)|\Psi_S(t_0)\rangle$. The state vectors are labeled with the subscript S here to identify that the state vectors correspond to the Schrödinger picture. Also in the Schrödinger picture, position and momentum operators have no time dependence. If not explicitly stated otherwise, the state vectors and operators of this *Field Guide* are given in the Schrödinger picture. The Schrödinger picture is the frame of reference on which the transformations to the other two pictures are based.

The Heisenberg picture (page 45) is defined by a unitary transformation operator $\hat{\mathbb{F}}(t) = \hat{\mathbb{U}}^\dagger(t, t_0)$ (see page 42) that is the adjoint of the time evolution operator of the Schrödinger picture. When this unitary transformation is applied to a Schrödinger-picture state vector $|\Psi_S(t)\rangle$, time dependence vanishes: $|\Psi_H\rangle \equiv \hat{\mathbb{U}}^\dagger(t, t_0)|\Psi_S(t)\rangle = |\Psi_S(t_0)\rangle$, where $|\Psi_H\rangle$ is the transformed ket of the Heisenberg picture. The Heisenberg picture is therefore the specific time-dependent reference frame in which the transformed state vectors are constant in time, while operators (such as the position and momentum operators) that have no time dependence in the Schrödinger picture may depend on time in the Heisenberg picture.

The interaction picture (page 46) is used when the Schrödinger-picture Hamiltonian is time dependent. In this picture, operators and state vectors generally evolve in time.

Schrödinger Picture: Expectation Value Dynamics

Position and momentum operators have no explicit time dependence in the Schrödinger picture. Nevertheless, other operators that do have explicit time dependence can be constructed. Let $\hat{A}_S(t)$ be an arbitrary operator in the Schrödinger picture that may have explicit time dependence.

For a given Schrödinger-picture state vector $|\Psi_S(t)\rangle$ evolving under a Hamiltonian $\hat{H}_S(t)$, the expectation value of $\hat{A}_S(t)$ will generally have time dependence that results from both the time dependence of $|\Psi_S(t)\rangle$ and any time dependence in $\hat{A}_S(t)$. This time-dependent expectation value may be calculated as

$$\langle \hat{A}_S(t)\rangle(t) = \langle \Psi_S(t)|\hat{A}_S(t)|\Psi_S(t)\rangle$$

Differentiating the equation above with respect to time produces the following differential equation for the dynamics of expectation values in the Schrödinger picture, noting that $\hat{A}_S$, $\hat{H}_S$ and the expectation values are generally time dependent.

Schrödinger-picture expectation values:

$$\frac{d}{dt}\langle \hat{A}_S\rangle = \frac{1}{i\hbar}\left\langle [\hat{A}_S, \hat{H}_S]\right\rangle + \left\langle \frac{\partial \hat{A}_S}{\partial t}\right\rangle$$

Ehrenfest's equations (or Ehrenfest's theorem) are obtained by replacing $\hat{A}_S(t)$ in the equation above with the position operator $\hat{\mathbf{R}} = (\hat{X}, \hat{Y}, \hat{Z})$, and also with the momentum operator $\hat{\mathbf{P}} = (\hat{P}_x, \hat{P}_y, \hat{P}_z)$. Noting that $\frac{\partial \hat{\mathbf{R}}}{\partial t} = 0$ and $\frac{\partial \hat{\mathbf{P}}}{\partial t} = 0$, and that the expectation value of the gradient of the potential energy operator $V(\hat{\mathbf{R}})$ is $\langle \nabla V(\hat{\mathbf{R}})\rangle \equiv \left(\left\langle \frac{\partial V(\hat{\mathbf{R}})}{\partial \hat{X}}\right\rangle, \left\langle \frac{\partial V(\hat{\mathbf{R}})}{\partial \hat{Y}}\right\rangle, \left\langle \frac{\partial V(\hat{\mathbf{R}})}{\partial \hat{Z}}\right\rangle\right)$, the pair of Ehrenfest's equations can be solved to determine how expectation values of position and momentum evolve for any state of a system.

Ehrenfest's equations (Ehrenfest's theorem):

$$\frac{d}{dt}\langle \hat{\mathbf{R}}\rangle = \frac{1}{m}\langle \hat{\mathbf{P}}\rangle$$

$$\frac{d}{dt}\langle \hat{\mathbf{P}}\rangle = -\langle \nabla V(\hat{\mathbf{R}})\rangle$$

Heisenberg Picture: Operators and Dynamics

The adjoint $\hat{\mathbb{U}}^{\dagger}(t, t_0)$ of the time evolution operator $\hat{\mathbb{U}}(t, t_0)$ defines a unitary transformation that transforms state vectors of the Schrödinger picture into time-independent state vectors within the Heisenberg picture. An arbitrary operator $\hat{A}_S(t)$ in the Schrödinger picture is transformed into the Heisenberg operator

$$\hat{A}_H(t) = \hat{\mathbb{U}}^{\dagger}(t, t_0)\hat{A}_S(t)\hat{\mathbb{U}}(t, t_0)$$

The expectation value of an observable is a physically meaningful quantity and must be the same in any picture, since transforming between pictures is only of mathematical utility for performing calculations. As shown below, the time-dependent expectation value of $\hat{A}_H(t)$ calculated in the Heisenberg picture is equivalent to that of $\hat{A}_S(t)$. In the table, $\hat{H}_S$ and $|\Psi_S(t)\rangle$ are the Schrödinger-picture Hamiltonian and state vector.

Heisenberg Picture Quantities and Dynamics	
$\lvert\Psi_H\rangle$	$\equiv \hat{\mathbb{U}}^{\dagger}(t, t_0)\lvert\Psi_S(t)\rangle = \lvert\Psi_S(t_0)\rangle$
$\hat{A}_H(t)$	$\equiv \hat{\mathbb{U}}^{\dagger}(t, t_0)\hat{A}_S(t)\hat{\mathbb{U}}(t, t_0)$
$\hat{H}_H$	$= \hat{H}_S$, for time-independent H_S
$\hat{H}_H(t)$	$= \hat{H}_S(t)$, for $[\hat{H}_S(t), \hat{H}_S(t')] = 0$
$\langle\hat{A}_H(t)\rangle(t)$	$= \langle\Psi_H\lvert\hat{A}_H(t)\rvert\Psi_H\rangle$
	$= \langle\Psi_S(t_0)\lvert\hat{\mathbb{U}}^{\dagger}(t, t_0)\hat{A}_S(t)\hat{\mathbb{U}}(t, t_0)\rvert\Psi_S(t_0)\rangle$
	$= \langle\Psi_S(t)\lvert\hat{A}_S(t)\rvert\Psi_S(t)\rangle = \langle\hat{A}_S(t)\rangle(t)$
$i\hbar\frac{d}{dt}\hat{A}_H(t)$	$= [\hat{A}_H(t), \hat{H}_H(t)]$
	$+ i\hbar\hat{\mathbb{U}}^{\dagger}(t, t_0)\left(\frac{d}{dt}\hat{A}_S(t)\right)\hat{\mathbb{U}}(t, t_0)$

The effective Hamiltonian of the Heisenberg picture is $\hat{H}_E = 0$, which is not identical to the transformed Hamiltonian $\hat{H}_H(t)$ (see page 42). Since $\hat{H}_E = 0$ in the Heisenberg picture, the effective Schrödinger equation $i\hbar\frac{\partial}{\partial t}|\Psi_H\rangle = 0$ is solved by $|\Psi_H\rangle = |\Psi_S(t_0)\rangle$. In the Heisenberg picture, only operators evolve in time, satisfying the differential equation in the last line of the table above. These operators are then used with $|\Psi_H\rangle$ in calculations of physically meaningful quantities.

Interaction Picture

The interaction picture is obtained with a unitary transformation of the state vectors and operators of the Schrödinger picture. This transformation removes some of the time dependence of the Schrödinger-picture state vectors, while also altering the time dependence of operators. The interaction picture is typically used with a time-dependent Schrödinger-picture Hamiltonian $\hat{H}_S(t) = \hat{H}_0 + \hat{W}(t)$, where the eigenstates of the time-independent term $\hat{H}_0$ are known, and $\hat{W}(t)$ is a time-dependent **perturbation**. $\hat{W}(t)$ can induce time-dependent dynamics and **transitions** between the eigenstates of $\hat{H}_0$, which are not stationary states of the system. These dynamics are often of primary interest when the interaction picture is used.

To transform into the interaction picture, an evolution operator $\hat{\mathbb{U}}_0(t, t_0) \equiv e^{-\frac{i}{\hbar}(t-t_0)\hat{H}_0}$ is associated with $\hat{H}_0$. Note that $\hat{\mathbb{U}}_0$ is not the time evolution operator associated with $\hat{H}_S(t)$. If $|\Psi_S(t)\rangle$ is a state vector and $\hat{A}_S(t)$ is an arbitrary operator, both specified in the Schrödinger picture, then the transformed state vector and arbitrary operator expressed in the interaction picture are

$$|\Psi_I(t)\rangle \equiv \hat{\mathbb{U}}_0^\dagger(t, t_0)|\Psi_S(t)\rangle$$
$$\hat{A}_I(t) \equiv \hat{\mathbb{U}}_0^\dagger(t, t_0)\hat{A}_S(t)\hat{\mathbb{U}}_0(t, t_0)$$

The effective Schrödinger equation in the interaction picture is

$$i\hbar\frac{\partial}{\partial t}|\Psi_I(t)\rangle = \hat{H}_E(t)|\Psi_I(t)\rangle$$

where $\hat{H}_E(t)$ is an effective Hamiltonian given by

$$\hat{H}_E(t) = \hat{\mathbb{U}}_0^\dagger(t, t_0)\hat{W}(t)\hat{\mathbb{U}}_0(t, t_0)$$

and is not identical to the transformed Hamiltonian $\hat{H}_I(t) \equiv \hat{\mathbb{U}}_0^\dagger(t, t_0)\hat{H}_S(t)\hat{\mathbb{U}}_0(t, t_0)$ (see the boxes at the bottom of pages 42 and 45). In the interaction picture, calculation of state vector dynamics that are due entirely to $\hat{W}(t)$ are emphasized, since the dynamics due to $\hat{H}_0$ are removed in the transformation. If $\hat{W}(t) = 0$, then the interaction picture reduces to the Heisenberg picture; in that case, $\hat{H}_E(t) = 0$ and $|\Psi_I(t)\rangle = |\Psi_S(t_0)\rangle$.

Exactly Solvable Problems in One Dimension

There are a few common **exactly solvable problems** consisting of a particle of mass m in a 1D time-independent potential well $V(x)$. A "problem" in this context means the identification of a specific state space and a Hamiltonian. A "solution" usually involves determining the energy eigenvalues and energy eigenstates, the relationships between these quantities, and the expression of energy eigenstates in representations that are meaningful for the given problem. Exactly solvable problems are ones for which solutions can be found analytically and without approximation methods.

For each of the exactly solvable 1D problems described on the following pages, the Hamiltonian has the form

$$\hat{H} = \frac{1}{2m}\hat{P}_x^2 + V(\hat{X})$$

In the 1D position representation, $\hat{H}$ takes the form

$$H_{\{x\}} = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + V(x)$$

If a constant energy term is added to $\hat{H}$, all energy eigenvalues will be shifted by that amount. However, the physical dynamics of the system depend only on differences between energy eigenvalues, so the dynamics remain unchanged. Energy is therefore measured and specified with respect to some implicit or specified reference value for every problem.

The examples on the following pages utilize the position representation. The problems and solutions include

- The potential $V(x)$, which defines the problem
- The energy eigenvalue equation, which defines the notation used to specify energy eigenvalues, quantum numbers, and the energy eigenkets and eigenfunctions
- Solutions to the energy eigenvalue equation
- Explanatory or illustrative notes, plots, and special features about the specific problem

Free Particle and Delta Function Potential Well

Free particle:

$$V(x) = 0$$

$$H_{\{x\}}\, \varphi_{p'}(x) = E_{p'}\, \varphi_{p'}(x)$$

Solutions:

$$E_{p'} = \frac{p'^2}{2m}; \quad \left\{ p' \in \mathbb{R} \,|\, p' : \left[\frac{\mathrm{kg} \cdot \mathrm{m}}{\mathrm{s}}\right] \right\}$$

$$\varphi_{p'}(x) = \frac{1}{\sqrt{2\pi\hbar}} e^{ixp'/\hbar}$$

The functions $\varphi_{p'}(x)$ are **plane waves**, each indexed by a precise particle momentum p'. These solutions are non-degenerate eigenkets of the momentum operator, expressed in the position representation; they are non-normalizable and are not physically realizable wavefunctions. In the momentum representation, $\varphi_{p'}(x)$ transforms (page 24) into a Dirac delta function $\tilde{\varphi}_{p'}(p) = \delta(p - p')$.

For any $p' \neq 0$, $\varphi_{p'}(x)$ and $\varphi_{-p'}(x)$ have the same energy eigenvalue $E_{p'}$, so that for scalars c_+ and c_-, the superposition

$$\psi(x) = c_+ \varphi_{p'}(x) + c_- \varphi_{-p'}(x)$$

is also an energy eigenfunction with eigenvalue $E_{p'}$.

Dirac delta function potential well:

$$V(x) = -\alpha\delta(x); \qquad \{\alpha \in \mathbb{R}^+ \,|\, \alpha : [\mathrm{J} \cdot \mathrm{m}]\}$$

$$H_{\{x\}}\, \varphi_\alpha(x) = E_\alpha\, \varphi_\alpha(x)$$

Solutions:

$$E_\alpha = -\frac{m\alpha^2}{2\hbar^2}$$

$$\varphi_\alpha(x) = \sqrt{m\alpha/\hbar^2}\, e^{-m\alpha|x|/\hbar^2}$$

$\varphi_\alpha(x)$ is the wavefunction for the ground state, which is also the system's only bound state (i.e., localized about $x = 0$).

Chapter 36

Analytical Solid Geometry

36.1 INTRODUCTION

In 1637, Rene Descartes[*] represented geometrical figures (configurations) by equations and vice versa. Analytical Geometry involves algebraic or analytic methods in geometry. Analytical geometry in three dimensions also known as Analytical solid[**] geometry or solid analytical geometry, studies geometrical objects in space involving three dimensions, which is an extension of coordinate geometry in plane (two dimensions).

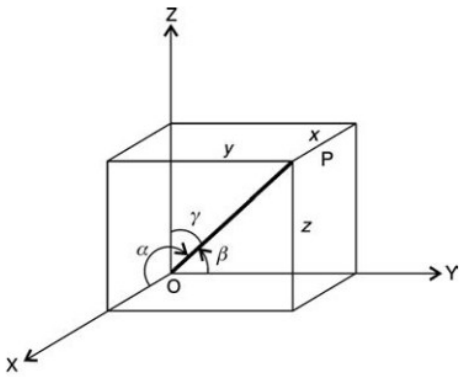

Fig. 36.1

Rectangular Cartesian Coordinates

The position (location) of a point in space can be determined in terms of its perpendicular distances (known as rectangular cartesian coordinates or simply *rectangular coordinates*) from three mutually perpendicular planes (known as **coordinate planes**). The lines of intersection of these three coordinate planes are known as *coordinate axes* and their point of intersection the *origin*.

The three axes called x-axis, y-axis and z-axis are marked positive on one side of the origin. The positive sides of axes OX, OY, OZ form a right handed system. The coordinate planes divide entire space into eight parts called *octants*. Thus a point P with coordinates x, y, z is denoted as $P(x, y, z)$. Here x, y, z are respectively the perpendicular distances of P from the YZ, ZX and XY planes. Note that a line perpendicular to a plane is perpendicular to every line in the plane.

Distance between two points $P_1(x_1, y_1, z_1)$ and $P_2(x_2, y_2, z_2)$ is $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2}$. *Distance* from origin $O(0, 0, 0)$ is $\sqrt{x_2^2 + y_2^2 + z_2^2}$. Divisions of the line joining two points P_1, P_2: The coordinates of $Q(x, y, z)$, the point on P_1P_2 dividing the line segment P_1P_2 in the ratio $m : n$ are $\left(\frac{nx_1+mx_2}{m+n}, \frac{ny_1+my_2}{m+n}, \frac{nz_1+mz_2}{m+n}\right)$ or putting k for $\frac{m}{n}$, $\left(\frac{x_1+kx_2}{1+k}, \frac{y_1+ky_2}{1+k}, \frac{z_1+kz_2}{1+k}\right); k \neq -1$. Coordinates of mid point are $\left(\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2}, \frac{z_1+z_2}{2}\right)$.

---

[*] Rene Descartes (1596–1650) French philosopher and mathematician, latinized name for Renatus Cartesius.

[**] *Not* used in the sense of "non-hollowness". By a sphere or cylinder we mean a hollow sphere or cylinder.

Direction of a line: A line in space is said to be directed if it is taken in a definite sense from one extreme (end) to the other (end).

Angle between Two Lines

Two straight lines in space may or may not intersect. If they intersect, they form a plane and are said to be coplanar. If they do not intersect, they are called *skew lines*.

Angle between two intersecting (coplanar) lines is the angle between their positive directions.

Angle between two non-intersecting (non-coplanar or skew) lines is the angle between two intersecting lines whose directions are same as those of given two lines.

36.2 DIRECTION COSINES AND DIRECTION RATIOS

Direction Cosines of a Line

Let L be a directed line OP from the origin $O(0, 0, 0)$ to a point $P(x, y, z)$ and of length r (Fig. 36.2). Suppose OP makes angles α, β, γ with the positive directions of the coordinate axes. Then α, β, γ are known as the *direction angles* of L. The cosines of these angles $\cos\alpha$, $\cos\beta$, $\cos\gamma$ are known as the *direction cosines* of the line $L(OP)$ and are in general denoted by l, m, n respectively.

Thus

$$l = \cos\alpha = \frac{x}{r}, \quad m = \cos\beta = \frac{y}{r}, \quad n = \cos\gamma = \frac{z}{r}.$$

where $\qquad r = \sqrt{x^2 + y^2 + z^2}.$

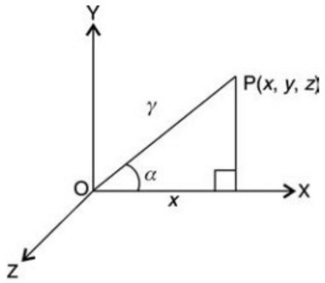

Fig. 36.2

Corollary 1: Lagrange's identity: $l^2+m^2+n^2 = 1$ i.e., sum of the squares of the direction cosines of any line is one, since $l^2 + m^2 + n^2 = \cos^2\alpha + \cos^2\beta + \cos^2\gamma = \frac{x^2}{r^2} + \frac{y^2}{r^2} + \frac{z^2}{r^2} = 1$.

Corollary 2: Direction cosines of the coordinate axes *OX*, *OY*, *OZ* are (1, 0, 0), (0, 1, 0), (0, 0, 1) respectively.

Corollary 3: The coordinates of P are (lr, mr, nr) where l, m, n are the direction cosines of OP and r is the length of OP.

Note: Direction cosines is abbreviated as DC's.

Direction Ratios

(abbreviated as DR's:) of a line L are any set of three numbers a, b, c which are proportional to l, m, n the DC's of the line L. DR's are also known as direction numbers of L. Thus $\frac{l}{a} = \frac{m}{b} = \frac{n}{c} = k$ (proportionality constant) or $l = ak, m = bk, n = ck$. Since $l^2 + m^2 + n^2 = 1$ or $(ak)^2 + (bk)^2 + (ck)^2 = 1$ or $k = \frac{\pm 1}{\sqrt{a^2+b^2+c^2}}$. Then the actual direction cosines are $\cos\alpha = l = ak = \pm\frac{a}{\sqrt{a^2+b^2+c^2}}$, $\cos\beta = m = bk = \pm\frac{b}{\sqrt{a^2+b^2+c^2}}$, $\cos\gamma = m = ck = \pm\frac{c}{\sqrt{a^2+b^2+c^2}}$ with $a^2 + b^2 + c^2 \neq 0$. Here +ve sign corresponds to positive direction and −ve sign to negative direction.

Note 1: Sum of the squares of DR's need not be one.

Note 2: Direction of line is $[a, b, c]$ where a, b, c are DR's.

Direction cosines of the line joining two points $P_1(x_1, y_1, z_1)$ and $P_2(x_2, y_2, z_2)$:

$$l = \cos\alpha = \frac{PQ}{r} = \frac{LM}{r} = \frac{OM - OL}{r} = \frac{x_2 - x_1}{r}.$$

Similarly, $m = \cos\beta = \frac{y_2 - y_1}{r}$ and $n = \cos\gamma = \frac{z^2 - z_1}{r}$. Then the DR's of P_1P_2 are $x_2 - x_1$, $y_2 - y_1$, $z_2 - z_1$

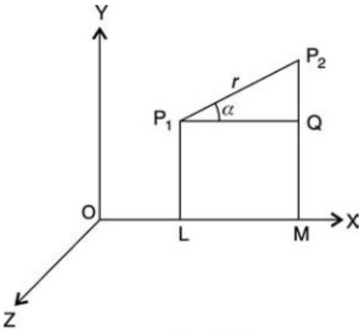

Fig. 36.3

Projections

Projection of a point P on line L is Q, the foot of the perpendicular from P to L.

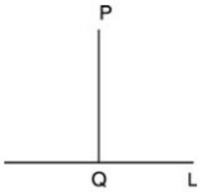

Fig. 36.4

Projection of line segment

P_1P_2 on a line L is the line segment MN where M and N are the feet of the perpendiculars from P and Q on to L. If θ is the angle between P_1P_2 and line L, then projection of P_1P_2 on $L = MN = PR = P_1P_2\cos\theta$. Projection of line segment P_1P_2 on line L with (whose) DC's l, m, n is

$$l(x_2 - x_1) + m(y_2 - y_1) + n(z_2 - z_1)$$

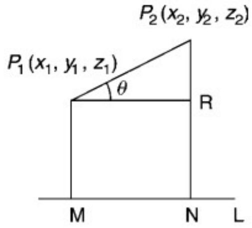

Fig. 36.5

Angle between Two Lines

Let θ be the angle between the two lines OP_1 and OP_2. Let $OP_1 = r_1, OP_2 = r_2$. Let l_1, m_1, n_1 be DC's of OP_1 and l_2, m_2, n_2 are DC's of OP_2. Then the coordinates of P_1 are l_1r_1, m_1r_1, n_1r_1 and of P_2 and l_2r_2, m_2r_2, n_2r_2.

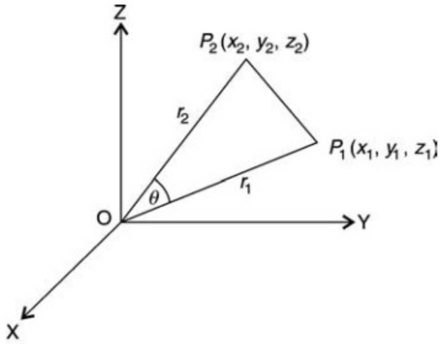

Fig. 36.6

From ΔOP_1P_2, we have

$$P_1P_2^2 = OP_1^2 + OP_2^2 - 2OP_1 \cdot OP_2 \cdot \cos\theta$$

$$(l_2r_2 - l_1r_1)^2 + (m_2r_2 - m_1r_1)^2 + (n_2r_2 - n_1r_1)^2$$
$$= \left[(l_1r_1)^2 + (m_1r_1)^2 + (n_1r_1)^2\right]$$
$$+ \left[(l_2r_2)^2 + (m_2r_2)^2 + (n_2r_2)^2\right] - 2 \cdot r_1r_2\cos\theta.$$

Using $l_1^2 + m_1^2 + n_1^2 = 1$ and $l_2^2 + m_2^2 + n_2^2 = 1$,

$$r_1^2 + r_2^2 - 2r_1r_2(l_1l_2 + m_1m_2 + n_1n_2)$$
$$= r_1^2 + r_2^2 - 2r_1r_2\cos\theta.$$

Then $\boxed{\cos\theta = l_1l_2 + m_1m_2 + n_1n_2}$

Corollary 1:

$$\sin^2\theta = 1 - \cos^2\theta = 1 - (l_1l_2 + m_1m_2 + n_1n_2)^2$$
$$= (l_1^2 + m_1^2 + n_1^2)(l_2^2 + m_2^2 + n_2^2)$$
$$-(l_1l_2 + m_1m_2 + n_1n_2)^2$$
$$= (l_1m_2 - m_1l_2)^2 + (m_1n_2 - n_1m_2)^2$$
$$+(n_1l_2 - n_2l_1)^2$$

using the Lagrange's identity. Then

$$(l_1^2+m_1^2+n_1^2)(l_2^2+m_2^2+n_2^2)-(l_1l_2+m_1m_2+n_1n_2)^2)$$
$$= (l_1m_2-l_2m_1)^2+(m_1n_2-m_2n_1)^2+(n_1l_2-n_2l_1)^2.$$

Thus $\sin\theta = \sqrt{\sum(l_1m_2 - m_1l_2)^2}$

Corollary 2: $\tan\theta = \frac{\sin\theta}{\cos\theta} = \frac{\sqrt{\sum(l_1m_2-m_1l_2)^2}}{l_1l_2+m_1m_2+n_1n_2}$.

Corollary 3: If a_1, b_1, c_1 and a_2, b_2, c_2 are DR's of OP_1 and OP_2

Then $l_1 = \frac{a_1}{\sqrt{a_1^2+b_1^2+c_1^2}}$, $m_1 = \frac{b_1}{\sqrt{a_1^2+b_1^2+c_1^2}}$, $n_1 = \frac{c_1}{\sqrt{a_1^2+b_1^2+c_1^2}}$ etc.

Then $\cos\theta = \frac{a_1a_2+b_1b_2+c_1c_2}{\sqrt{a_1^2+b_1^2+c_1^2}\sqrt{a_2^2+b_2^2+c_2^2}}$,

$$\sin\theta = \frac{\sqrt{(a_1b_2-a_2b_1)^2+(b_1c_2-b_2c_1)^2+(c_1a_2-c_2a_1)^2}}{\sqrt{a_1^2+b_1^2+c_1^2}\sqrt{a_2^2+b_2^2+c_2^2}}.$$

Corollary: Condition for perpendicularity:

The two lines are perpendicular if $\theta = 90°$. Then

$$\cos\theta = \cos 90 = 0$$

Thus $\boxed{l_1l_2 + m_1m_2 + n_1n_2 = 0}$

or $\boxed{a_1a_2 + b_1b_2 + c_1c_2 = 0}$

Corollary: Condition for parallelism:

If the two lines are parallel then $\theta = 0$. So $\sin\theta = 0$.

$$(l_1m_2 - m_1l_2)^2 + (m_1n_2 - m_2n_1)^2 + (n_1l_2 - n_2l_1)^2 = 0$$

or $$\frac{l_1}{l_2} = \frac{m_1}{m_2} = \frac{n_1}{n_2} = \frac{\sqrt{l_1^2+m_1^2+n_1^2}}{\sqrt{l_2^2+m_2^2+n_2^2}} = \frac{1}{1}.$$

Thus $\boxed{l_1 = l_2, \quad m_1 = m_2, \quad n_1 = n_2}$

or $$\frac{a_1}{a_2} = \frac{b_1}{b_2} = \frac{c_1}{c_2}.$$

WORKED OUT EXAMPLES

Example 1: Find the angle between the lines $A(-3, 2, 4)$, $B(2, 5, -2)$ and $C(1, -2, 2)$, $D(4, 2, 3)$.

Solution: DR's of AB: $2 - (-3), 5 - 2, -2 - 4 = 5, 3, -6$
DR's of CD: 3, 4, 1. Then DC's of AB are $l_1 = \cos\alpha_1 = \frac{5}{\sqrt{5^2+3^2+6^2}} = \frac{5}{\sqrt{25+9+36}} = \frac{5}{\sqrt{70}}$ and $m_1 = \cos\beta_1 = \frac{3}{70}$, $n_1 = \cos\gamma_1 = \frac{-6}{\sqrt{70}}$. Similarly, $l_2 = \cos\alpha_2 = \frac{3}{\sqrt{3^2+4^2+1^2}} = \frac{3}{\sqrt{9+16+1}} = \frac{3}{\sqrt{26}}$, and $m_2 = \cos\beta_2 = \frac{4}{\sqrt{26}}$, $n_2 = \cos\gamma_2 = \frac{1}{\sqrt{26}}$. Now

$$\begin{aligned}\cos\theta &= \cos\alpha_1\cdot\cos\alpha_2 + \cos\beta_1\cdot\cos\beta_2 + \cos\gamma_1\cdot\cos\gamma_2\\ &= l_1l_2 + m_1m_2 + n_1n_2\\ \cos\theta &= \frac{5}{\sqrt{70}}\cdot\frac{3}{\sqrt{26}} + \frac{3}{\sqrt{70}}\cdot\frac{4}{\sqrt{26}} - \frac{6}{\sqrt{70}}\cdot\frac{1}{\sqrt{26}}\\ &= 0.49225\end{aligned}$$

$$\therefore\quad \theta = \cos^{-1}(0.49225) = 60°30.7'$$

Example 2: Find the DC's of the line that is $\perp^r$ to each of the two lines whose directions are $[2, -1, 2]$ and $[3, 0, 1]$.

Solution: Let $[a, b, c]$ be the direction of the line. Since this line is $\perp^r$ to the line with direction $[2, -1, 2]$, by orthogonality

$$2a - b + 2c = 0$$

Similarly, direction $[a, b, c]$ is $\perp^r$ to direction $[3, 0, 1]$. So

$$3a + 0 + c = 0.$$

Solving $c = -3a$, $b = -4a$ or
direction $[a, b, c] = [a, -4a, -3a] = [1, -4, -3]$.
$\therefore$ DC's of the line: $\frac{1}{\sqrt{1^2+4^2+3^2}} = \frac{1}{\sqrt{26}}, \frac{-4}{\sqrt{26}}, \frac{-3}{\sqrt{26}}$.

Example 3: Show that the points $A(1, 0, -2)$, $B(3, -1, 1)$ and $C(7, -3, 7)$ are collinear.

Solution: DR's of AB: $[2, -1, 3]$, DR's of AC: $[6, -3, 9]$, DR's of BC: $[4, -2, 6]$. Thus DR's of AB, AC, BC are same. Hence A, B, C are collinear.

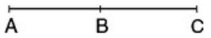

Example 4: Find the coordinates of the foot of the perpendicular from $A(1, 1, 1)$ on the line joining $B(1, 4, 6)$ and $C(5, 4, 4)$.

Solution: Suppose D divides BC in the ratio $k : 1$. Then the coordinates of D are $\left(\frac{5k+1}{k+1}, \frac{4k+4}{k+1}, \frac{4k+6}{k+1}\right)$. DR's of AD: $\frac{4k}{k+1}, 3, \frac{3K+5}{k+1}$, DR's of BC: $4, 0, -2$ AD is $\perp^r$ BC: $16k - 6k - 10 = 0$, or $k = 1$.

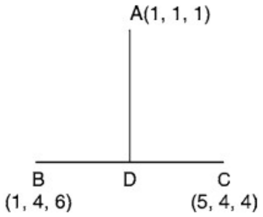

Coordinates of the foot of perpendicular are (3, 4, 5).

Example 5: Show that the points $A(1, 0, 2)$, $B(3, -1, 3)$, $C(2, 2, 2)$, $D(0, 3, 1)$ are the vertices of a parallelogram.

Fig. 36.7

Solution: DR's of AB are $[3 - 1, -1 - 0, 3 - 2] = [2, -1, 1]$. Similarly, DR's of BC are $[-1, 3, -1]$, of $CD[-2, 1, -1]$ of $DA[-1, 3, -1]$. Since DR's of AB and CD are same, they are parallel. Similarly BC and DA are parallel since DR's are same. Further AB is not $\perp^r$ to AD because

$$2(+1) + (-1)(-3) + 1(+1) = 6 \neq 0$$

Similarly, AD is not $\perp^r$ to BC because

$$2(-1) + (-1)3 + 1(-1) = -6 \neq 0.$$

Hence $ABCD$ is a parallelogram.

EXERCISE

1. Show that the points $A(7, 0, 10)$, $B(6, -1, 6)$, $C(9, -4, 6)$ form an isoscales right angled triangle.

   Hint: $AB^2 = BC^2 = 18$, $CA^2 = 36$, $AB^2 + BC^2 = CA^2$

2. Prove that the points $A(3, -1, 1)$, $B(5, -4, 2)$, $C(11, -13, 5)$ are collinear.

   Hint 1: $AB^2 = 14$, $BC^2 = 126$, $CA^2 = 224$, $AB + BC = 4\sqrt{14} = CA$

   Hint 2: DR's of $AB = 2, -3, 1$; BC: $6, -9, 3$; $AB \|^l$ to BC

3. Determine the internal angles of the triangle ABC where $A(2, 3, 5)$, $B(-1, 3, 2)$, $C(3, 5, -2)$.

   Hint: $AB^2 = 18, BC^2 = 36, AC^2 = 54$. DC's AB: $-\frac{1}{\sqrt{2}}, 0, -\frac{1}{\sqrt{2}}$; BC: $\frac{2}{3}, \frac{1}{3}, \frac{-2}{3}$; AC: $\frac{1}{3\sqrt{6}}, \frac{2}{3\sqrt{6}}, \frac{-7}{3\sqrt{6}}$.

Ans. $\cos A = \frac{1}{\sqrt{3}}$, $\cos B = 0$ i.e., $B = 90^\circ$, $\cos C = \frac{\sqrt{6}}{3}$.

4. Show that the foot of the perpendicular from $A(0, 9, 6)$ on the line joining $B(1, 2, 3)$ and $C(7, -2, 5)$ is $D(-2, 4, 2)$.

   Hint: D divides BC in $k : 1$, $D\left(\frac{7k+1}{k+1}, \frac{-2k+2}{k+1}, \frac{5k+3}{k+1}\right)$. DR's AD: $(7k + 1, -11k - 7, -k - 3)$, DR's BC: $6, -4, 2$. $AD \perp^r BC$: $k = -\frac{1}{3}$.

5. Find the condition that three lines with DC's $l_1, m_1, n_1; l_2, m_2, n_2; l_3, m_3, n_3$ are concurrent.

   Hint: Line with DC's l, m, n through point of concurrency will be $\perp^r$ to all three lines, $ll_i + mm_i + nn_i = 0$, $i = 1, 2, 3$.

Ans.
$$\begin{vmatrix} l_1 & m_1 & n_1 \\ l_2 & m_2 & n_2 \\ l_3 & m_3 & n_3 \end{vmatrix} = 0$$

6. Show that $\cos^2\alpha + \cos^2\beta + \cos^2\gamma + \cos^2\delta = \frac{4}{3}$ where $\alpha, \beta, \gamma, \delta$ are the angles which a line makes with the four diagonals of a cube.

   Hint: DC's of four diagonals are (k, k, k), $(-k, k, k)$, $(k, -k, k)$, $(k, k, -k)$ where $k = \frac{1}{\sqrt{3}}$; l, m, n are DC's of line. $\cos\alpha = l.k. + mk + nk$, $\cos\beta = (-l + m + n)k$, $\cos\gamma = (l - m + n)k$, $\cos\delta = (l + m - n)k$.

7. Show that the points $A(-1, 1, 3)$, $B(1, -2, 4)$, $C(4, -1, 1)$ are vertices of a right triangle.

   Hint: DR's $AB : [2, -3, 1]$, $BC : [3, 1, -3]$, $CA : [5, -2, -2]$. AB is $\perp^r$ BC.

8. Prove that $A(3, 1, -2)$, $B(3, 0, 1)$, $C(5, 3, 2)$, $D(5, 4, -1)$ form a rectangle.

   Hint: DR's: AB: $[0, -1, 3]$; AC: $[2, 2, 4]$, $CD[0, 1, -3]$; $AD[2, 3, 1]$; $BC[2, 3, 1]$; $AB \| CD$, $AD \| BC$, $AD \perp AB: 0 - 3 + 3 = 0$, $BC \perp DC: 0 + 3 - 3 = 0$.

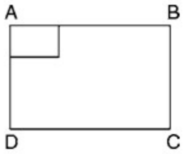

9. Find the interior angles of the triangle

   $A(3, -1, 4)$, $B(1, 2, -4)$, $C(-3, 2, 1)$.

   Hint: DC's of AB: $(-2, 3, -8)k_1$, BC: $(-4, 0, 5)k_2$, AC: $(-6, 3, -3)k_3$ where $k_1 = \frac{1}{\sqrt{77}}$, $k_2 = \frac{1}{\sqrt{41}}$, $k_3 = \frac{1}{\sqrt{54}}$.

Ans. $\cos A = \frac{15}{\sqrt{462}}$, $\cos B = \frac{32}{\sqrt{3157}}$, $\cos C = \frac{3}{\sqrt{246}}$.

10. Determine the DC's of a line $\perp^r$ to a triangle formed by $A(2, 3, 1)$, $B(6, -3, 2)$, $C(4, 0, 3)$.

Ans. $(3, 2, 0)k$ where $k = \frac{1}{\sqrt{13}}$.

   Hint: DR: AB: $[4, -6, 1]$, BC: $[-2, 3, 1]$, CA: $[2, -3, 2]$. $[a, b, c]$ of $\perp^r$ line: $4a - 6b + c = 0$, $-2a + 3b + c = 0$, $2a - 3b + 2c = 0$.

36.3 THE PLANE

Surface is the locus of a point moving in space satisfying a single condition.

Example: Surface of a sphere is the locus of a point that moves at a constant distance from a fixed point.

Surfaces are either plane or curved. Equation of the locus of a point is the analytical expression of the given condition(s) in terms of the coordinates of the point.

Exceptional cases: Equations may have locus other than a surface. Examples: (i) $x^2 + y^2 = 0$ is z-axis (ii) $x^2 + y^2 + z^2 = 0$ is origin (iii) $y^2 + 4 = 0$ has no locus.

Plane is a surface such that the straight line PQ, joining any two points P and Q on the plane, lies completely on the plane.

General equation of first degree in x, y, z is of the form

$$Ax + By + Cz + D = 0$$

Here A, B, C, D are given real numbers and A, B, C are not all zero (i.e., $A^2 + B^2 + C^2 \neq 0$)

Book Work: Show that every equation of the first degree in x, y, z represents a plane.

Proof: Let

$$Ax + By + Cz + D = 0 \tag{1}$$

be the equation of first degree in x, y, z with the condition that not all A, B, C are zero (i.e., $A^2 + B^2 + C^2 \neq 0$). Let $P(x_1, y_1, z_1)$ and $Q(x_2, y_2, z_2)$ be any two points on the surface represented by (1). Then

$$Ax_1 + By_1 + Cz_1 + D_1 = 0 \tag{2}$$

$$Ax_2 + By_2 + Cz_2 + D_2 = 0 \tag{3}$$

Multiplying (3) by k and adding to (2), we get

$$A(x_1 + kx_2) + B(y_1 + ky_2) + C(z_1 + kz_2) + D(1 + k) = 0 \tag{4}$$

Assuming that $1 + k \neq 0$, divide (4) by $(1 + k)$.

$$A\left(\frac{x_1 + kx_2}{1 + k}\right) + B\left(\frac{y_1 + ky_2}{1 + k}\right) + C\left(\frac{z_1 + kz_2}{1 + k}\right) + D = 0$$

i.e., the point $R\left(\frac{x_1+kx_2}{1+k}, \frac{y_1+ky_2}{1+k}, \frac{z_1+kz_2}{1+k}\right)$ which is point dividing the line PQ in the ratio $k : 1$, also lies on the surface (1). Thus any point on the line joining P and Q lies on the surface i.e., line PQ completely lies on the surface. Therefore the surface by definition must be a plane.

General form of the equation of a plane is

$$Ax + By + Cz + D = 0$$

Special cases:

(i) Equation of plane passing through origin is

$$Ax + By + Cz = 0 \tag{5}$$

(ii) Equations of the coordinate planes XOY, YOZ and ZOX are respectively $z = 0, x = 0$ and y=0

(iii) $Ax + By + D = 0$ plane $\perp^r$ to xy-plane
$Ax + Cz + D = 0$ plane $\perp^r$ to xz-plane
$Ay + Cz + D = 0$ plane $\perp^r$ to yz-plane.

Similarly, $Ax + D = 0$ is $\|^l$ to yz-plane, $By + D = 0$ is $\|^l$ to zx-plane, $cz + D = 0$ is $\|^l$ to xy-plane.

One point form

Equation of a plane through a fixed point $P_1(x_1, y_1, z_1)$ and whose normal CD has DC's proportional to (A, B, C): For any point $P(x, y, z)$ on the given plane, the DR's of the line P_1P are $(x - x_1, y - y_1, z - z_1)$. Since a line perpendicular to a plane is perpendicular to every line in the plane, so ML is perpendicular to P_1, P. Thus

$$A(x - x_1) + B(y - y_1) + C(z - z_1) = 0 \qquad (6)$$

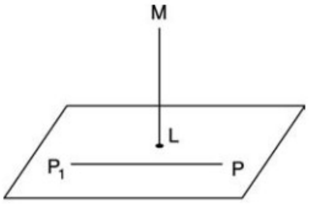

Fig. 36.8

Note 1: Rewriting (6), we get the general form of plane

$$Ax + By + Cz + D = 0 \qquad (1)$$

where $D = -ax_1 - by_1 - cz_1$

Note 2: The real numbers A, B, C which are the coefficients of x, y, z respectively in (1) are proportional to DC's of the normal ot the plane (1).

Note 3: Equation of a plane parallel to plane (1) is

$$Ax + By + Cz + D^* = 0 \qquad (7)$$

x-intercept of a plane is the point where the plane cuts the x-axis. This is obtained by putting $y = 0$, $z = 0$. Similarly, y-, z-intercepts. Traces of a plane are the lines of intersection of plane with coordinate axis.

Example: xy-trace is obtained by putting $z = 0$ in equation of plane.

Intercept form

Suppose $P(a, 0, 0)$, $Q(0, b, 0)$, $R(0, 0, c)$ are the x-, y-, z-intercepts of the plane. Then P, Q, R lies on the plane. From (1)

$$Aa + 0 + 0 + D = 0$$

or
$$A = -\frac{D}{a}.$$

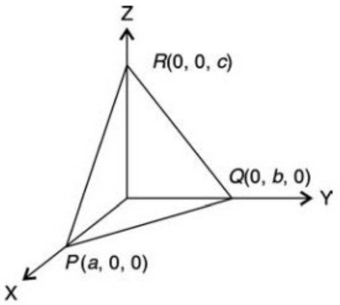

Fig. 36.9

similarly, $0 + bB + 0 + D = 0$ or $B = -\frac{D}{b}$ and $C = -\frac{D}{c}$.

Eliminating A, B, C the equation of the plane in the intercept form is

$$-\frac{D}{a}x - \frac{D}{b} - \frac{D}{c}z + D = 0$$

or
$$\frac{x}{a} + \frac{y}{b} + \frac{z}{c} = 1 \qquad (8)$$

Normal form

Let $P(x, y, z)$ be any point on the plane. Let ON be the perpendicular from origin O to the given plane. Let $ON = p$. (i.e., length of the perpendicular ON is p). Suppose l, m, n are the DC's of ON. Now ON is perpendicular to PN. Projection of OP on ON is ON itself i.e., p.

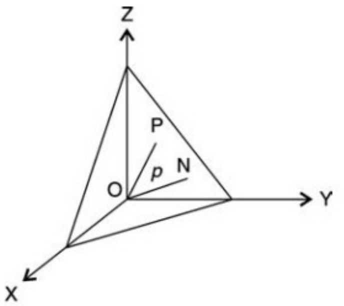

Fig. 36.10

Also the projection OP joining origin (0, 0, 0) to $P(x, y, z)$ on the line ON with DC's l, m, n is

$$l(x-0)+m(y-0)+n(z-0)$$

or
$$lx+my+nz \tag{9}$$

Equating the two projection values from (8) & (9)

$$lx+my+nz=p \tag{10}$$

Note 1: p is always positive, since p is the perpendicular distance from origin to the plane.

Note 2: Reduction from general form.

Transpose constant term to R.H.S. and make it positive (if necessary by multiplying throughout by -1). Then divide throughout by $\pm\sqrt{A^2+B^2+C^2}$. Thus the general form $Ax+By+Cz+D=0$ takes the following normal form

$$\frac{Ax}{\pm\sqrt{A^2+B^2+C^2}}+\frac{By}{\pm\sqrt{A^2+B^2+C^2}}+\frac{Cz}{\pm\sqrt{A^2+B^2+C^2}}$$
$$=\frac{-D}{\pm\sqrt{A^2+B^2+C^2}} \tag{11}$$

The sign before the radical is so chosen to make the R.H.S. in (11) positive.

Three point form

Equation of a plane passing through three given points $P_1(x_1, y_1, z_1)$, $P_2(x_2, y_2, z_2)$, $P_3(x_3, y_3, z_3)$:

Since the three points P_1, P_2, P_3 lie on the plane

$$Ax+By+Cz+D=0 \tag{1}$$

we have
$$Ax_1+By_1+Cz_1+D=0 \tag{12}$$
$$Ax_2+By_2+Cz_2+D=0 \tag{13}$$
$$Ax_3+By_3+Cz_3+D=0 \tag{14}$$

Eliminating A, B, C, D from (1), (12), (13), (14) (i.e., a non trivial solution A, B, C, D for the homogeneous system of 4 equations exist if the determinant coefficient is zero)

$$\begin{vmatrix} x & y & z & 1 \\ x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \end{vmatrix}=0 \tag{15}$$

Equation (15) is the required equation of the plane through the 3 points P_1, P_2, P_3.

Corollary 1: Coplanarity of four given points: The four points $P_1(x_1, y_1, z_1)$, $P_2(x_2, y_2, z_2)$, $P_3(x_3, y_3, z_3)$, $P_4(x_4, y_4, z_4)$ are coplanar (lie in a plane) if

$$\begin{vmatrix} x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \\ x_4 & y_4 & z_4 & 1 \end{vmatrix}=0 \tag{16}$$

Angle between Two Given Planes

The angle between two planes

$$A_1x+B_1y+C_1z+D_1=0 \tag{17}$$
$$A_2x+B_2y+C_2z+D_2=0 \tag{18}$$

is the angle θ between their normals. Here A_1, B_1, C_1 and A_2, B_2, C_2 are the DR's of the normals respectively to the planes (17) and (18). Then

$$\cos\theta=\frac{A_1A_2+B_1B_2+C_1C_2}{\sqrt{A_1^2+B_1^2+C_1^2}\sqrt{A_2^2+B_2^2+C_2^2}}$$

Condition for perpendicularity

If $\theta=0$ then the two planes are $\perp^r$ to each other. Then

$$A_1A_2+B_1B_2+C_1C_2=0 \tag{19}$$

Condition for parallelism

If $\theta=0$, the two planes are $\|^l$ to each other. Then

$$\frac{A_1}{A_2}=\frac{B_1}{B_2}=\frac{C_1}{C_2} \tag{20}$$

Note: Thus parallel planes differ by a constant.

Although there are four constants A, B, C, D in the equation of plane, essentially three conditions are required to determine the three ratios of A, B, C, D, for example plane passing through:

a. three non-collinear points

b. two given points and $\perp^r$ to a given plane

c. a given point and $\perp^r$ to two given planes etc.

Coordinate of the Foot of the Perpendicular from a Point to a Given Plane

Let $Ax + By + Cz + D = 0$ be the given plane and $P(x_1, y_1, z_1)$ be a given point. Let PN be the perpendicular from P to the plane. Let the coordinates of the foot of the perpendicular PN be $N(\alpha, \beta, \gamma)$. Then DR's of $PN(x_1 - \alpha, y_1 - \beta, z_1 - \gamma)$ are proportional to the coefficients A, B, C i.e.,

$$x_1 - \alpha = kA, \quad y_1 - B = kB, \quad z_1 - \gamma = kC$$

or $$\alpha = x_1 - kA, \quad y_1 = \beta - kB, \quad z_1 = \gamma - kC$$

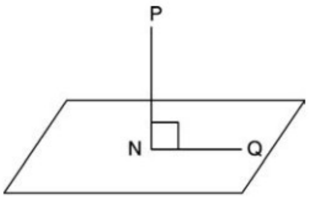

Fig. 36.11

Since N lies in the plane

$$A\alpha + B\beta + C\gamma + D = 0$$

Substituting α, β, γ,

$$A(x_1 - kA) + b(y_1 - kB) + c(z_1 - kC) + D = 0$$

Solving $$k = \frac{Ax_1 + By_1 + CZ_1 + D}{A^2 + B^2 + C^2}$$

Thus the coordinates of $N(\alpha, \beta, \gamma)$ the foot of the perpendicular from $P(x_1, y_1, z_1)$ to the plane are

$$\alpha = x_1 - \frac{A(Ax_1 + By_1 + Cz_1 + D)}{A^2 + B^2 + C^2},$$

$$\beta = y_1 - \frac{B(Ax_1 + By_1 + Cz_1 + D)}{A^2 + B^2 + C^2},$$

$$\gamma = z_1 - \frac{C(Ax_1 + By_1 + Cz_1 + D)}{A^2 + B^2 + C^2} \tag{21}$$

Corollary 1: Length of the perpendicular from a given point to a given plane:

$$PN^2 = (x_1 - \alpha)^2 + (y_1 - \beta)^2 + (z_1 - \gamma)^2$$
$$= (kA)^2 + (kB)^2 + (kC)^2$$
$$= k^2(A^2 + B^2 + C^2)$$
$$= \left[\frac{Ax_1 + By_1 + Cz_1 + D}{A^2 + B^2 + C^2}\right]^2 (A^2 + B^2 + C^2)$$
$$= \frac{(Ax_1 + By_1 + Cz_1 + D)^2}{A^2 + B^2 + C^2}$$

or $$PN = \frac{Ax_1 + By_1 + Cz_1 + D}{\pm\sqrt{A^2 + B^2 + C^2}}.$$

The sign before the radical is chosen as positive or negative according as *D* is positive or negative. Thus the numerical values of the length of the perpendicular *PN* is

$$PN = \left|\frac{Ax_1 + By_1 + Cz_1 + D}{\sqrt{A^2 + B^2 + C^2}}\right| \tag{22}$$

Note: PN is obtained by substituting the coordinates (x_1, y_1, z_1) in the L.H.S. of the Equation (1) and dividing it by $\sqrt{A^2 + B^2 + C^2}$.

Equation of a plane passing through the line of intersection of two given planes $u \equiv A_1x + B_1y + C_1z + D_1 = 0$ and $v \equiv A_2x + B_2y + C_2z + D_2 = 0$ is $u + kv = 0$ where k is any constant.

Equations of the two planes bisecting the angles between two planes are

$$\frac{A_1x + B_1y + C_1z + D_1}{\sqrt{A_1^2 + B_1^2 + C_1^2}} = \pm\frac{A_2x + B_2y + C_2z + D_2}{\sqrt{A_2^2 + B_2^2 + C_2^2}}.$$

Worked Out Examples

Example 1: Find the equation of the plane which passes through the point $(2, 1, 4)$ and is

a. Parallel to plane $2x + 3y + 5z + 6 = 0$

b. Perpendicular to the line joining $(3, 2, 5)$ and $(1, 6, 4)$

c. Perpendicular to the two planes $7x + y + 2z = 6$ and $3x + 5y - 6z = 8$

d. Find intercept points and traces of the plane in case c.

Solution:

a. Any plane parallel to the plane

$$2x + 3y + 5z + 6 = 0$$

is given by $2x+3y+5z+k=0$ (1) (differs by a constant). Since the point $(2,1,4)$ lies on the plane (1), $2(2)+3(1)+5(4)+k=0$, $k=-27$. Required equation of plane is $2x+3y+5z-27=0$.

b. Any plane through the point $(2,1,4)$ is (one point form)

$$A(x-2)+B(y-1)+C(z-4)=0 \qquad (2)$$

DC's of the line joining $M(3,2,5)$ and $N(1,6,4)$ are proportional to $2,-4,1$. Since MN is perpendicular to (2), A,B,C are proportional to $2,-4,1$. Then $2(x-2)-4(y-1)+1(z-4)=0$. The required equation of plane is $2x-4y+z-4=0$.

c. The plane through $(2,1,4)$ is

$$A(x-2)+B(y-1)+C(z-4)=0. \qquad (2)$$

This plane (2) is perpendicular to the two planes $7x+y+2z=6$ and $3x+5y-6z=8$.
Using $A_1A_2+B_1B_2+C_1C_2=0$, we have

$$7a+b+2c=0$$

$$3a+5b-6c=0$$

Solving $\frac{a}{-6-10}=\frac{-b}{-42-8}=\frac{c}{35-3}$ or $\frac{a}{1}=\frac{b}{-3}=\frac{c}{-2}$.

Required equation of plane is

$$1(x-4)-3(y-1)-2(z-4)=0$$

or
$$x-3y-2z+7=0$$

d. x-intercept: Put $y=z=0$, $\therefore x=-7$ or $(-7,0,0)$ is the x-intercept. Similarly, y-intercept is $(0,\frac{7}{3},0)$ and z-intercept is $\left(0,0,\frac{7}{2}\right)$. xy-trace is obtained by putting $z=0$. It is $x-3y+7=0$. Similarly, yz-trace is $3y+2z-7=0$ and zx-trace is $x-2z+7=0$.

Example 2: Find the equation of the plane containing the points $P(3,-1,-4)$, $Q(-2,2,1)$, $R(0,4,-1)$.

Solution: Equation of plane through the point $P(3,-1,-4)$ is

$$A(x+3)+B(y+1)+C(z+4)=0. \qquad (1)$$

DR's of PQ: $-5,3,5$; DR's of PR: $-3,5,3$. Since line PQ and PR completely lies in the plane (1), normal to (1) is perpendicular to PQ and PR. Then

$$-5A+3B+5C=0$$

$$-3A+5B+3C=0$$

Solving $A=C=1, \qquad B=0$

$$(x-3)+0+(z+4)=0$$

Equation of the plane is

$$x+z+1=0$$

Aliter: Equation of the plane by 3-point form is

$$\begin{vmatrix} x & y & z & 1 \\ 3 & -1 & -4 & 1 \\ -2 & 2 & 1 & 1 \\ 0 & 4 & -1 & 1 \end{vmatrix}=0$$

Expanding $D_1x-D_2y+D_3z-1.D_4=0$ where

$$D_1=\begin{vmatrix} -1 & -4 & 1 \\ 2 & 1 & 1 \\ 4 & -1 & 1 \end{vmatrix}=-16, \quad D_2=\begin{vmatrix} 3 & -4 & 1 \\ -2 & 1 & 1 \\ 0 & -1 & 1 \end{vmatrix}0=0$$

$$D_3=\begin{vmatrix} 3 & -1 & 1 \\ -2 & 2 & 1 \\ 0 & 4 & 1 \end{vmatrix}=-16, \quad D_4=\begin{vmatrix} 3 & -1 & -4 \\ -2 & 2 & 1 \\ 0 & 4 & -1 \end{vmatrix}=16$$

or required equation is $x+z+1=0$.

Example 3: Find the perpendicular distance between (a) The Point $(3,2,-1)$ and the plane $7x-6y+6z+8=0$ (b) between the parallel planes $x-2y+2z-8=0$ and $x-2y+2z+19=0$ (c) find the foot of the perpendicular in case (a).

Solution:
Perpendicular distance $=\left(\frac{Ax_1+By_1+Cz_1+D}{\sqrt{A^2+B^2+C^2}}\right)$

a. Point $(3,2,-1)$, plane is $7x-6y+6z+8=0$. So perpendicular distance from $(3,2,-1)$ to plane is

$$=\frac{7(3)-6(2)+6(-1)+8}{\sqrt{7^2+6^2+6^2}}=\frac{11}{-11}=|-1|=1$$

b. x-intercept point of plane $x-2y+2z-8=0$ is $(8,0,0)$ (obtained by putting $y=0, z=0$ in the equation). Then perpendicular distance from the point $(8, 0, 0)$ to the second plane $x - 2y + 2z + 19 = 0$ is $\frac{1.8-2.0+2.0+19}{\sqrt{1^2+2^2+2^2}} = \frac{27}{3} = 9$

c. Let $N(\alpha, \beta, \gamma)$ be the foot of the perpendicular from $P(3, 2, -1)$. DR's of PN: $3 - \alpha, 2 - \beta, -1 - \gamma$. DR's of normal to plane are $7, -6, 6$. These are proportional. $\frac{3-\alpha}{7} = \frac{2-\beta}{-6} = \frac{-1-\gamma}{6}$ or $\alpha = 3 - 7k, \beta = 2 + 6k, \gamma = -1 - 6k$. Now (α, β, γ) lies on the plane. $7(3 - 7k) - 6(2 + 6k) + 6(-1 - 6k) + 8 = 0$ or $k = \frac{1}{11}$.
$\therefore$ the coordinates of the foot of perpendicular are $\left(\frac{26}{11}, \frac{28}{11}, \frac{-17}{11}\right)$.

Example 4: Are the points $(2, 3, -5)$ and $(3, 4, 7)$ on the same side of the plane $x + 2y - 2z = 9$?

Solution: Perpendicular distance of the point $(2, 3, -5)$ from the plane $x + 2y - 2z - 9 = 0$ or $-x - 2y + 2z + 9 = 0$ is $\frac{-1.2-2(3)-2(-5)+9}{\sqrt{1^2+2^2+2^2}} = -\frac{9}{3} = -3$.

$\perp^r$ distance of $(3, 4, 7)$ is $\frac{-1.3-2.4+2.7+9}{\sqrt{1^2+2^2+2^2}} = \frac{12}{3} = 6$

$\perp^r$ distance from origin $(0, 0, 0)$ is $\frac{0+0+0+9}{3} = 3$

So points $(2, 3, -5)$ and $(3, 4, 7)$ are on opposite sides of the given plane.

Example 5: Find the angle between the planes $4x - y + 8z = 9$ and $x + 3y + z = 4$.

Solution: DR's of the planes are $[4, -1, 8]$ and $[1, 3, 1]$. Now

$$\cos\theta = \frac{A_1A_2 + B_1B_2 + C_1C_2}{\sqrt{A_1^2 + B_1^2 + C_1^2}\sqrt{A_2^2 + B_2^2 + C_2^2}}$$
$$= \frac{4.1 + 3\cdot(-1) + 1.8}{\sqrt{16 + 1 + 64}\sqrt{1 + 9 + 1}}$$
$$= \frac{9}{\sqrt{81}\sqrt{11}} = \frac{1}{\sqrt{11}} \quad \text{or} \quad \theta = \cos^{-1}\frac{1}{\sqrt{11}}.$$

Example 6: Find the equation of a plane passing through the line of intersection of the planes.

a. $3x + y - 5z + 7 = 0$ and $x - 2y + 4z - 3 = 0$ and passing through the point $(-3, 2, -4)$

b. $2x - 5y + z = 3$ and $x + y + 4z = 5$ and parallel to the plane $x + 3y + 6z = 1$.

Solution:

a. Equation of plane is $u + kv = 0$ i.e.,

$$(3x + y - 5z + 7) + k(x - 2y + 4z - 3) = 0.$$

Since point $(-3, 2, -4)$ lies on the intersection plane

$$[3(-3) + 1.(2) - 5(-4) + 7]$$
$$+k[1(-3) - 2(2) + 4(-4) - 3] = 0.$$

So $k = \frac{10}{13}$. Then the required plane is

$$49x - 7y - 25z + 61 = 0.$$

b. Equation of plane is $u + kv = 0$ i.e.,

$$(2x - 5y + z - 3) + k(x + y + 4z - 5) = 0$$

or $$(2+k)x+(-5+k)y+(1+4k)z+(-3-5k) = 0.$$

Since this intersection plane is parallel to $x + 3y + 6z - 1 = 0$

So $$\frac{2+k}{1} = \frac{-5+k}{3} = \frac{1+4k}{6} \quad \text{or} \quad k = -\frac{11}{2}.$$

Required equation of plane is $7x + 21y + 42z - 49 = 0$.

Example 7: Find the planes bisecting the angles between the planes $x + 2y + 2z = 9$ and $4x - 3y + 12z + 13 = 0$. Specify the angle θ between them.

Solution: Equations of the bisecting planes are

$$\frac{x + 2y + 2z - 9}{\sqrt{1 + 2^2 + 2^2}} = \pm\frac{4x - 3y + 12z + 13}{\sqrt{4^2 + 3^2 + 12^2}}$$
$$\frac{x + 2y + 2z - 9}{3} = \pm\frac{4x - 3y + 12z + 13}{13}$$

or $$25x + 17y + 62z - 78 = 0 \quad \text{and}$$
$$x + 35y - 10z - 156 = 0.$$

$$\cos\theta = \frac{25\cdot 1 + 17\cdot 35 - 62 \times 10}{\sqrt{25^2 + 17^2 + 62^2}\sqrt{1 + 35^2 + 10^2}} = 0$$

$$\therefore \quad \theta = \frac{\pi}{2}$$

i.e, angle between the bisecting planes is $\frac{\pi}{2}$.

Example 8: Show that the planes

$$7x + 4y - 4z + 30 = 0 \qquad (1)$$

$$36x - 51y + 12z + 17 = 0 \qquad (2)$$

$$14x + 8y - 8z - 12 = 0 \qquad (3)$$

$$12x - 17y + 4z - 3 = 0 \qquad (4)$$

form four faces of a rectangular parallelopiped.

Solution: (1) and (3) are parallel since $\frac{7}{14} = \frac{4}{8} = \frac{-4}{-8} = \frac{1}{2}$. (2) and (4) are parallel since $\frac{36}{12} = \frac{-51}{-17} = \frac{12}{4} = 3$. Further (1) and (2) are $\perp^r$ since

$$7 \cdot 36 + 4(-51) - 4(12) = 252 - 204 - 48 = 0.$$

EXERCISE

1. Find the equation of the plane through $P(4, 3, 6)$ and perpendicular to the line joining $P(4, 3, 6)$ to the point $Q(2, 3, 1)$.

   Hint: DR's PQ: $[2, 0, 5]$, DR of plane through $(4, 3, 6) : x - 4, y - 3, z - 6; \perp^r$: $2(x - 4) + 0(y - 3) + 5(z - 6) = 0$

Ans. $2x + 5z - 38 = 0$

2. Find the equation of the plane through the point $P(1, 2, -1)$ and parallel to the plane $2x - 3y + 4z + 6 = 0$.

   Hint: Eq. $2x - 3y + 4z + k = 0$, $(1, 2, -1)$ lies, $k = 8$.

Ans. $2x - 3y + 4z + 8 = 0$

3. Find the equation of the plane that contains the three points $P(1, -2, 4)$, $Q(4, 1, 7)$, $R(-1, 5, 1)$.

   Hint: $A(x - 1) + B(y + 2) + C(z - 4) = 0$, DR: PQ: $[3, 3, 3]$, PR: $[-2, 7, -3]$. $\perp^r$ $3A + 3B + 3C = 0$, $-2A + 7B - 3C = 0$, $A = -10B$, $C = 9B$.

   Aliter: $$\begin{vmatrix} x & y & z & 1 \\ 1 & -2 & 4 & 1 \\ 4 & 1 & 7 & 1 \\ -1 & 5 & 1 & 1 \end{vmatrix} = 0,$$

   $$D_1 x - D_2 y + D_3 z - D_4 = 0$$

   where $$D_1 = \begin{vmatrix} -2 & 4 & 1 \\ 1 & 7 & 1 \\ 5 & 1 & 1 \end{vmatrix} \text{ etc.}$$

Ans. $10x - y - 9z + 24 = 0$

4. Find the equation of the plane
   - **a.** passing through $(1, -1, 2)$ and $\perp^r$ to each of the planes $2x + 3y - 2z = 5$ and $x + 2y - 3z = 8$
   - **b.** passing through $(-1, 3, -5)$ and parallel to the plane $6x - 3y - 2z + 9 = 0$
   - **c.** passing through $(2, 0, 1)$ and $(-1, 2, 0)$ and $\perp^r$ to the plane $2x - 4y - z = 7$.

Ans. **a.** $5x - 4y - z = 7$
b. $6x - 3y - 2z + 5 = 0$
c. $6x + 5y - 8z = 4$

5. Find the perpendicular distance between
   - **a.** the point $(-2, 8, -3)$ and plane $9x - y - 4z = 0$
   - **b.** the two planes $x - 2y + 2z = 6$, $3x - 6y + 6z = 2$
   - **c.** the point $(1, -2, 3)$ and plane $2x - 3y + 2z - 14 = 0$.

Ans. (a) $\sqrt{2}$ (b) $\frac{-16}{9}$ (c) 0 i.e., lies on the plane.

6. Find the angle between the two planes
   - **a.** $x + 4y - z = 5$, $y + z = 2$
   - **b.** $x - 2y + 3z + 4 = 0$, $2x + y - 3z + 7 = 0$

Ans. (a) $\cos\theta = \frac{1}{2}$, $\theta = 60^\circ$ (b) $\cos\theta = \frac{-9}{14}$.

7. Prove that the planes $5x - 3y + 4z = 1$, $8x + 3y + 5z = 4$, $18x - 3y + 13z = 6$ contain a common line.

   Hint: $u + kv = 0$ substitute in $w = 0$, $k = \frac{1}{2}$

8. Find the coordinates of N, the foot of the perpendicular from the point $P(-3, 0, 1)$ on the plane $4x - 3y + 2z = 19$. Find the length of this perpendicular. Find also the image of P in the plane.

   Hint: $PN = NQ$ i.e., N is the mid point.

Ans. $N(1, -3, 3)$, $\sqrt{29}$, image of P is $Q(5, -6, 5)$

9. Find the equation of the plane through the line of intersection of the two planes $x - 3y + 5z - 7 = 0$ and $2x + y - 4z + 1 = 0$ and $\perp^r$ to the plane $x + y - 2z + 4 = 0$.

Ans. $3x - 2y + z - 6 = 0$

10. A variable plane passes through the fixed point (a, b, c) and meets the coordinate axes in P, Q, R. Prove that the locus of the point common to the planes through P, Q, R parallel to the coordinate plane is $\frac{a}{x} + \frac{b}{y} + \frac{c}{z} = 1$.

    Hint: $OP = x_1,\ OQ = y_1,\ OR = z_1,\ \frac{x}{x_1} + \frac{y}{y_1} + \frac{z}{z_1} = 1$, (a, b, c) lies, $\frac{a}{x_1} + \frac{b}{y_1} + \frac{c}{z_1} = 1$.

36.4 THE STRAIGHT LINE

Two surfaces will in general intersect in a curve. In particular two planes, which are not parallel, intersect in a straight line.

Example: The coordinate planes ZOX and XOY, whose equations are $y = 0$ and $z = 0$ respectively, intersect in a line the x-axis.

Straight line

The locus of two simultaneous equations of first degree in x, y, z

$$\begin{aligned} A_1x + B_1y + C_1z + D_1 = 0 \\ A_2x + B_2y + C_2z + D_2 = 0 \end{aligned} \tag{1}$$

is a straight line, provided $A_1 : B_1 : C_1 \neq A_2 : B_2 : C_2$ (i.e., not parallel). Equation (1) is known as the **general form** of the equation of a straight line. Thus the equation of a straight line or simply line is the pair of equations taken together i.e., equations of two planes together represent the equation of a line. However this representation is not unique, because many planes can pass through a given line. Thus a given line can be represented by different pairs of first degree equations.

Projecting planes

Of the many planes passing through a given line, those that are perpendicular to the coordinate planes are known as projecting planes and their traces give the **projections** of the line on the coordinate planes.

Symmetrical Form

The equation of line passing through a given point $P_1(x_1, y_1, z_1)$ and having direction cosines l, m, n is given by

$$\frac{x - x_1}{l} = \frac{y - y_1}{m} = \frac{z - z_1}{n} \tag{2}$$

since for any point $P(x, y, z)$ on the line, the DR's of PP_1: $x - x_1, y - y_1, z - z_1$ be proportional to l, m, n. Equation (2) represent two independent linear equations and are called the symmetrical (or symmetric) form of the equation of a line.

Corollary: Any point P on the line (2) is given by

$$x = x_1 + lr, \qquad y = y_1 + mr, \qquad z = z_1 + nr \tag{3}$$

for different values of r, where $r = PP_1$.

Corollary: Lines perpendicular to one of the coordinate axes:

a. $x = x_1, \frac{y-y_1}{m} = \frac{z-z_1}{n}$, ($\perp^r$ to x-axis i.e., $\|^l$ to yz-plane)

b. $y = y_1, \frac{x-x_1}{l} = \frac{z-z_1}{n}$, ($\perp^r$ to y-axis i.e., $\|^l$ to xz-plane)

c. $z = z_1, \frac{x-x_1}{l} = \frac{y-y_1}{m}$, ($\perp^r$ to z-axis i.e., $\|^l$ to xy-plane)

Corollary: Lines perpendicular to two axes

a. $x = x_1, y = y_1$ ($\perp^r$ to x- & y-axis i.e., $\|^l$ to z-axis):

b. $x = x_1, z = z_1$ ($\perp^r$ to x- & z-axis i.e., $\|^l$ to y-axis)

c. $y = y_1, z = z_1$ ($\perp^r$ to y- & z-axis i.e., $\|^l$ to x-axis)

Corollary: Projecting planes: (containing the given line)

(a) $\frac{x-x_1}{l} = \frac{y-y_1}{m}$ (b) $\frac{x-x_1}{l} = \frac{z-z_1}{n}$ (c) $\frac{y-y_1}{m} = \frac{z-z_1}{n}$.

Note: When any of the constants l, m, n are zero, the Equation (2) are equivalent to equations

$$\frac{l}{x - x_1} = \frac{m}{y - y_1} = \frac{n}{z - z_1}.$$

Example: $\frac{x}{0}=\frac{y}{2}=\frac{z}{0}$ means $\frac{0}{x}=\frac{2}{y}=\frac{0}{z}$.

Corollary: If a, b, c are the DR's of the line, then (2) takes the form $\frac{x-x_1}{a}=\frac{y-y_1}{b}=\frac{z-z_1}{c}$.

Corollary: **Two point form** of a line passing through two given points $P_1(x_1, y_1, z_1)$ and $P_2(x_2, y_2, z_2)$ is

$$\frac{x-x_1}{x_2-x_1}=\frac{y-y_1}{y_2-y_1}=\frac{z-z_1}{z_2-z_1} \tag{4}$$

since the DR's of P_1P_2 are $x_2-x_1, y_2-y_1, z_2-z_1$.

Transformation of General Form to Symmetrical Form

The general form also known as unsymmetrical form of the equation of a line can be transformed to symmetrical form by determining

(a) one point on the line, by putting say $z=0$ and solving the simultaneous equations in x and y.

(b) the DC's of the line from the fact that this line is $\perp^r$ to both normals of the given planes.

For example,

(a) by putting $z=0$ in the general form

$$\begin{aligned} A_1x+B_1y+C_1z+D_1&=0\\ A_2x+B_2y+C_2z+D_2&=0 \end{aligned} \tag{2}$$

and solving the resulting equations

$$A_1x+B_1y+D_1=0$$

$$A_2x+B_2y+D_2=0,$$

we get a point on the line as

$$\left(\frac{B_1D_2-B_2D_1}{A_1B_2-A_2B_1}, \frac{A_2D_1-A_1D_2}{A_1B_2-A_2B_1}, 0\right) \tag{5}$$

(b) Using the orthogonality of the line with the two normals of the two planes, we get

$$lA_1+mB_1+nC_1=0$$

$$lA_2+mB_2+nC_2=0$$

where $(l, m, n), (A_1, B_1, C_1)$ and (A_2, B_2, C_2) are DR's of the line, normal to first plane, normal to second plane respectively. Solving, we get the DR's l, m, n of the line as

$$\frac{l}{B_1C_2-B_2C_1}=\frac{m}{C_1A_2-C_2A_1}=\frac{n}{A_1B_2-A_2B_1} \tag{6}$$

Using (5) and (6), thus the given general form (2) of the line reduces to the symmetrical form

$$\frac{x-\frac{(B_2D_1-B_1D_2)}{A_1B_2-A_2B_1}}{B_1C_2-B_2C_1}=\frac{y-\frac{(A_2D_1-A_1D_2)}{A_1B_2-A_2B_1}}{C_1A_2-C_2A_1}= =\frac{z-0}{A_1B_2-A_2B_1} \tag{7}$$

Note 1: In finding a point on the line, one can put $x=0$ or $y=0$ instead of $z=0$ and get similar results.

Note 2: General form (2) can also be reduced to the two point form (4) (special case of symmetric form) by determining two points on the line.

Angle between a Line and a Plane

Let π be the plane whose equation is

$$Ax+By+Cz+D=0 \tag{8}$$

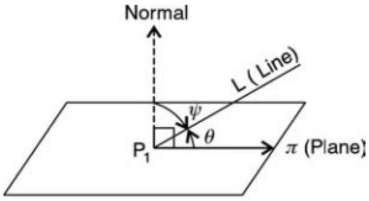

Fig. 36.12

and L be the straight line whose symmetrical form is

$$\frac{x-x_1}{l}=\frac{y-y_1}{m}=\frac{z-z_1}{n} \tag{2}$$

Let θ be the angle between the line L and the plane π. Let ψ be the angle between L and the normal to the plane π. Then

$$\cos\psi=\frac{lA+mB+nC}{\sqrt{l^2+m^2+n^2}\sqrt{A^2+B^2+C^2}}$$

$$=\cos(90-\theta)=\sin\theta \tag{9}$$

since $\psi=90-\theta$. The angle between a line L and plane π is the complement of the angle between the line L and the normal to the plane). Thus θ is determined from (9).

Corollary: Line is $\|^l$ to the plane if $\theta = 0$ then $\sin\theta = 0$ i.e.,

$$\boxed{lA + mB + nC = 0} \tag{10}$$

Corollary: Line is $\perp^r$ to the plane if $\theta = \frac{\pi}{2}$, then $\sin\theta = 1$ i.e.,

$$\boxed{\frac{l}{A} = \frac{m}{B} = \frac{n}{C}} \tag{11}$$

(i.e., DR's of normal and the line are same).

Conditions for a Line L to Lie in a Plane π

If every point of line L is a point of plane π, then line L lies in plane π. Substituting any point of the line $L : (x_1 + lr, y_1 + mr, z_1 + nr)$ in the equation of the plane (8), we get

$$A(x_1 + lr) + B(y_1 + mr) + C(z_1 + nr) + D = 0$$

or $$(Al + Bm + Cn)r + (Ax_1 + By_1 + Cz_1 + D) = 0 \tag{12}$$

This Equation (12) is satisfied for all values of r if the coefficient of r and constant term in (12) are both zero i.e.,

$$\boxed{\begin{array}{ll} Al + Bm + Cn = 0 & \text{and} \\ Ax_1 + By_1 + Cz_1 + D = 0 & \end{array}} \tag{13}$$

Thus the two conditions for a line L to lie in a plane π are given by (13) which geometrically mean that (i) line L is $\perp^r$ to the nomal ot the plne and (ii) a (any one) point of line L lies on the plane.

Corollary: General equation of a plane containing line L (2) is

$$A(x - x_1) + B(y - y_1) + C(z - z_1) = 0 \tag{14}$$

subject to

$$Al + Bm + Cn = 0$$

Corollary: Equation of any plane through the line of intersection of the two planes

$$u \equiv A_1x + B_1y + C_1z + D_1 = 0 \quad \text{and}$$
$$v \equiv A_2x + B_2y + C_2z + D_2 = 0$$

is $u + kv = 0$ or $(A_1x + B_1y + C_1z + D_1) + k(A_2x + B_2y + C_2z + D_2) = 0$ where k is a constant.

Coplanar Lines

Consider two given straight lines L_1

$$\frac{x - x_1}{l_1} = \frac{y - y_1}{m_1} = \frac{z - z_1}{n_1} \tag{15}$$

and line L_2

$$\frac{x - x_2}{l_2} = \frac{y - y_2}{m_2} = \frac{z - z_2}{n_2} \tag{16}$$

From (14), equation of any plane containing line L_1 is

$$A(x - x_1) + B(y - y_1) + C(z - z_1) = 0 \tag{17}$$

subject to

$$Al_1 + Bm_1 + Cn_1 = 0 \tag{18}$$

If the plane (17) contains line L_2 also, then the point (x_2, y_2, z_2) of L_2 should also lie in the plane (17). Then

$$A(x_2 - x_1) + B(y_2 - y_1) + C(z_2 - z_1) = 0 \tag{19}$$

But the line L_2 is $\perp^r$ to the normal to the plane (17). Thus

$$Al_2 + Bm_2 + Cn_2 = 0 \tag{20}$$

Therefore the two lines L_1 and L_2 will lie in the same plane if (17), (18), (20) are simultaneously satisfied. Eliminating A, B, C from (19), (18), (20)(i.e., homogeneous system consistent if coefficient determinant is zero), we have

$$\begin{vmatrix} x_2 - x_1 & y_2 - y_1 & z_2 - z_1 \\ l_1 & m_1 & n_1 \\ l_2 & m_2 & n_2 \end{vmatrix} = 0 \tag{21}$$

Thus (21) is the condition for coplanarity of the two lines L_1 and L_2. Now the equation of the plane containing lines L_1 and L_2 is

$$\begin{vmatrix} x - x_1 & y - y_1 & z - z_1 \\ l_1 & m_1 & n_1 \\ l_2 & m_2 & n_2 \end{vmatrix} = 0 \tag{22}$$

which is obtained by eliminating A, B, C from (17), (18), (20).

Corollary: Condition for the two lines L_1

$$\begin{aligned} & u_1 \equiv A_1x + B_1y + C_1z + D_1 = 0, \\ & u_2 \equiv A_2x + B_2y + C_2z + D_2 = 0 \\ \text{and Line } L_2 \quad & u_3 \equiv A_3x + B_3y + C_3z + D_3 = 0, \\ & u_4 \equiv A_4x + B_4y + C_4z + D_4 = 0 \end{aligned} \tag{23}$$

to be coplanar is

$$\begin{vmatrix} A_1 & B_1 & C_1 & D_1 \\ A_2 & B_2 & C_2 & D_2 \\ A_3 & B_3 & C_3 & D_3 \\ A_4 & B_4 & C_4 & D_4 \end{vmatrix} = 0 \tag{24}$$

If $P(\alpha, \beta, \gamma)$ is the point of intersection of the two lies, then P should satisfy the four Equations (23): $u_i|$ at $(\alpha, \beta, \gamma) = 0$ for $i = 1, 2, 3, 4$. Elimination of (α, β, γ) from these four equations leads to (24).

Corollary: The general form of equations of a line L_3 intersecting the lines L_1 and L_2 given by (23) are

$$u_1 + k_1u_2 = 0 \quad \text{and} \quad u_3 + k_2u_4 = 0 \tag{25}$$

where k_1 and k_2 are any two numbers.

Foot and length of the perpendicular from a point $P_1(\alpha, \beta, \gamma)$ to a given line L: $\frac{x-x_1}{l} = \frac{y-y_1}{m} = \frac{z-z_1}{n}$

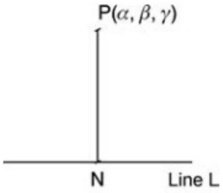

Fig. 36.13

Any point on the line L be $(x_1 + lr, y_1 + mr, z_1 + nr)$. The DR's of PN are $x_1 + lr - \alpha, y_1 + mr - \beta, z_1 + nr - \gamma$. Since PN is $\perp^r$ to line L, then

$$l(x_1+lr-\alpha)+m(y_1+mr-\beta)+n(z_1+nr-\gamma) = 0.$$

Solving

$$r = \frac{l(\alpha - x_1) + m(\beta - y_1) + n(\gamma - z_1)}{l^2 + m^2 + n^2} \tag{26}$$

The coordinates of N, the foot of the perpendicular PN is $(x_1 + lr - \alpha, y_1 + mr - \beta, z_1 + nr - \gamma)$ where r is given by (26).

The length of the perpendicular PN is obtained by distance formula between P (given) and N (found).

Line of greatest slope in a plane

Let ML be the line of intersection of a horizontal plane I with slant plane II. Let P be any point on plane II. Draw PN $\perp^r$ to the line ML. Then the line of greatest slope in plane II is the line PN, because no other line in plane II through P is inclined to the horizontal plane I more steeply than PN.

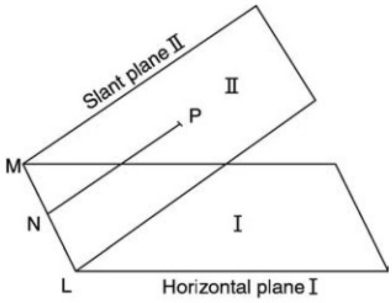

Fig. 36.14

WORKED OUT EXAMPLES

Example 1: Find the points where the line $x - y + 2z = 2$, $2x - 3y + 4z = 0$ pierces the coordinate planes.

Solution: Put $z = 0$ to find the point at which the line pierces the xy-plane: $x - y = 2$ and $2x - 3y = 0$ or $x = 6$, $y = 4$. $\therefore$ $(6, 4, 0)$.
Put $x = 0$, $-y + 2z = 2$, $-3y + 4z = 0$ or $y = 4$, $z = 3$ $\therefore$ $(0, 4, 3)$ is piercing point.
Put $y = 0$, $x + 2z = 2$, $2x + 4z = 0$ no unique solution.
Note that DR's of the line are $[2, 0, -1]$. So this line is $\perp^r$ to y-axis whose DR's are $[0, 1, 0]$ (i.e., $2 \cdot 0 + 0 \cdot 1 + (-1) \cdot 0 = 0$). Hence the given line does not pierce the xz-plane.

Example 2: Transfer the general (unsymmetrical) form $x + 2y + 3z = 1$ and $x + y + 2z = 0$ to the symmetrical form.

Solution: Put $x = 0$, $2y + 3z = 1$, $y + 2z = 0$. Solving $z = -1$, $y = 2$. So $(0, 2, -1)$ is a point on the line. Let l, m, n be the DR's of the line. Since this line is $\perp^r$ to both normals of the given two planes, we have

$$1 \cdot l + 2 \cdot m + 3 \cdot n = 0$$

$$1 \cdot l + 1 \cdot m + 2 \cdot n = 0$$

Solving $\dfrac{l}{4-3} = -\dfrac{m}{2-3} = \dfrac{n}{1-2}$ or $\dfrac{l}{1} = \dfrac{m}{1} = -\dfrac{n}{1}$

Equation of the line passing through the point $(0, 2, -1)$ and having DR's $1, 1, -1$ is

$$\frac{x-0}{1} = \frac{y-2}{2} = \frac{z+1}{-1}$$

Aliter: Two point form.

Put $y = 0$, $x + 3z = 1$, $x + 27 = 0$. Solving $z = 1$, $x = -2$ or $(-2, 0, 1)$ is another point on the line. Now DR's of the line joining the two points $(0, 2, -1)$ and $(-2, 0, 1)$ are $-2, -2, 2$. Hence the equation of the line in the two point form is

$$\frac{x-0}{-2} = \frac{y-2}{-2} = \frac{z+1}{2} \quad \text{or} \quad \frac{x}{1} = \frac{y-2}{1} = \frac{z+1}{-1}.$$

Example 3: Find the acute angle between the lines $\frac{x}{2} = \frac{y}{2} = \frac{z}{1}$ and $\frac{x}{5} = \frac{y}{4} = \frac{z}{-3}$.

Solution: DR's are $[2, 2, 1]$ and $[5, 4, -3]$. If θ is the angle between the two lines, then

$$\cos\theta = \frac{l_1 l_2 + m_1 m_2 + n_1 n_2}{\sqrt{l_1^2 + m_1^2 + n_1^2}\sqrt{l_2^2 + m_2^2 + n_2^2}}$$

$$= \frac{2 \cdot 5 + 2 \cdot 4 + 1 \cdot (-3)}{\sqrt{4+4+1}\sqrt{25+16+9}} = \frac{15}{3\sqrt{50}} = \frac{1}{\sqrt{2}}$$

$$\therefore \quad \theta = 45^\circ$$

Example 4: Find the equation of the plane containing the line $x = y = z$ and passing through the point $(1, 2, 3)$.

Solution: General form of the given line is

$$x - y = 0 \quad \text{and} \quad x - z = 0.$$

Equation of a plane containing this line is

$$(x - y) + k(x - z) = 0$$

Since point $(1, 2, 3)$ lies on this line, it also lies on the above plane. Then

$$(1 - 2) + k(1 - 3) = 0 \quad \text{or} \quad k = -\frac{1}{2}$$

Equation of required plane is

$$(x - y) - \frac{1}{2}(x - z) = 0$$

or
$$x - 2y + z = 0.$$

Example 5: Show that the lines $\frac{x}{1} = \frac{y+3}{2} = \frac{z+1}{3}$ and $\frac{x-3}{2} = \frac{y}{1} = \frac{z-1}{-1}$ intersect. Find the point of intersection.

Solution: Rewriting the equation in general form, we have

$$2x - y = 3, \qquad 3x - z = 1$$

and
$$x - 2y = 3, \qquad x + 2z = 5$$

If these four equations have a common solution, then the given two lines intersect. Solving, $y = -1$, then $x = 1$, $z = 2$. So the point of intersection is $(1, -1, 2)$.

Example 6: Find the acute angle between the lines $\frac{x}{3} = \frac{y}{1} = \frac{z}{0}$ and the plane $x + 2y - 7 = 0$.

Solution: DR's of the line: $[3, 1, 0]$. DR's of normal to the plane is $[1, 2, 0]$. If ψ is the angle between the line and the normal, then

$$\cos\psi = \frac{3 \cdot 1 + 1 \cdot 2 + 0 \cdot 0}{\sqrt{3^2 + 1^2 + 0^2}\sqrt{1^2 + 2^2 + 0^2}}$$

$$= \frac{5}{\sqrt{10}\sqrt{5}} = \frac{1}{\sqrt{2}} \qquad \text{so} \qquad \psi = 45^\circ.$$

Angle θ between the line and the plane is the complement of the angle ψ i.e., $\theta = 90 - \psi = 90 - 45 = 45^\circ$.

Example 7: Show that the lines $x + y - 3z = 0$, $2x + 3y - 8z = 1$ and $3x - y - z = 3$, $x + y - 3z = 5$ are parallel.

Solution: DR's of the first line are

$$\begin{array}{ccc} l_1 & m_1 & n_1 \\ 1 & 1 & -3 \\ 2 & 3 & -8 \end{array} \quad \text{or} \quad \frac{l_1}{1} = \frac{m_1}{2} = \frac{n_1}{1}$$

Similarly, DR's of the second line are

$$\begin{array}{ccc} l_2 & m_2 & n_2 \\ 3 & -1 & -1 \\ 1 & 1 & -3 \end{array} \quad \text{or} \quad \frac{l_2}{4}=\frac{m_2}{8}=\frac{n_2}{4} \quad \text{i.e.,} \quad \frac{l_2}{1}=\frac{m_2}{2}=\frac{n_2}{1}$$

Since the DR's of the two lines are same, they are parallel.

Example 8: Find the acute angle between the lines $2x - y + 3z - 4 = 0$, $3x + 2y - z + 7 = 0$ and $x + y - 2z + 3 = 0$, $4x - y + 3z + 7 = 0$.

Solution: The line represented by the two planes is perpendicular to both the normals of the two planes. If l_1, m_1, n_1 are the DR's of this line, then

$$\begin{array}{ccc} l_1 & m_1 & n_1 \\ 2 & -1 & 3 \\ 3 & 2 & -1 \end{array} \quad \text{or} \quad \frac{l_1}{-5} = \frac{m_1}{11} = \frac{n_1}{7}$$

Similarly, DR's of the 2nd line are

$$\begin{array}{ccc} l_2 & m_2 & n_2 \\ 1 & +1 & -2 \\ 4 & -1 & -3 \end{array} \quad \text{or} \quad \frac{l_2}{-1} = \frac{m_2}{11} = \frac{n_2}{5}$$

If θ is the angle between the lines, then

$$\cos\theta = \frac{l_1 l_2 + m_1 m_2 + n_1 n_2}{\sqrt{l_1^2 + m_1^2 + n_1^2}\sqrt{l_2^2 + m_2^2 + n_2^2}}$$

$$= \frac{5 + 121 + 35}{\sqrt{195}\sqrt{147}} = \frac{23}{3\sqrt{65}}$$

$$\therefore \qquad \text{So} \qquad \theta = 180^\circ 1.4'$$

Example 9: Prove that the line $\frac{x-4}{2} = \frac{y-2}{3} = \frac{z-3}{6}$ lies in the plane $3x - 4y + z = 7$.

Solution: The point of the line $(4, 2, 3)$ should also lie in the plane. So $3 \cdot 4 - 4 \cdot 2 + 1 \cdot 3 = 7$ satisfied. The line and normal to the plane are perpendicular. So $2 \cdot 3 + 3 \cdot (-4) + 6 \cdot 1 = 6 - 12 + 6 = 0$. Thus the given line completely lies in the given plane.

Example 10: Show that the lines $\frac{x-2}{2} = \frac{y-3}{-1} = \frac{z+4}{3}$ and $\frac{x-3}{1} = \frac{y+1}{3} = \frac{z-1}{-2}$ are coplanar. Find their common point and determine the equation of the plane containing the two given lines.

Solution: Here first line passes through $(2, 3, -4)$ and has DR's $l_1, m_1, n_1 : 2, -1, 3$. The second line passes through $(3, -1, 1)$ and has DR's $l_2, m_2, n_2 :$ $1, 3, -2$. Condition for coplanarity:

$$\begin{vmatrix} x_2-x_1 & y_2-y_1 & z_2-z_1 \\ l_1 & m_1 & n_1 \\ l_2 & m_2 & n_2 \end{vmatrix} = \begin{vmatrix} 3-2 & -1-3 & 1+4 \\ 2 & -1 & 3 \\ 1 & 3 & 2 \end{vmatrix}$$

$$= \begin{array}{c} 7+28-35 = 0 \\ \text{satisfied.} \end{array}$$

Point of intersection: Any point on the first line is $(2 + 2r_1, 3 - r_1 - 4 + 3r_1)$ and any point on the second line is $(3 + r_2, -1 + 3r_2, 1 - 2r_2)$. When the two lines intersect in a common point then coordinates on line (1) and line (2) must be equal, i.e., $2 + 2r_1 = 3 + r_2, 3 - r_1 = -1 + 3r_2$ and $-4 + 3r_1 = 1 - 2r_2$. Solving $r_1 = r_2 = 1$. Therefore the point of intersection is $(2 + 2 \cdot 1, 3 - 1, -4 + 3 \cdot 1)$ $= (4, 2, -1)$.
Equation of plane containing the two lines:

$$\begin{vmatrix} x-x_1 & y-y_1, & z-z_1 \\ l_1 & m_1 & n_1 \\ l_2 & m_2 & n_2 \end{vmatrix} = \begin{vmatrix} x-2 & y-3 & z+4 \\ 2 & -1 & 3 \\ 1 & 3 & -2 \end{vmatrix} = 0$$

Expanding $\quad -7(x-2) - (-7)(y-3) + 7(z+4) = 0$ or $x - y - z + 3 = 0$.

Example 11: Find the coordinates of the foot of the perpendicular from $P(1, 0, 2)$ to the line $\frac{x+1}{3} = \frac{y-2}{-2} = \frac{z+1}{-1}$. Find the length of the perpendicular and its equation.

Solution: Any point N on the given line is $(3r - 1, 2 - 2r, -1 - r)$. DR's of PN are $(3r - 2, 2 - 2r, -3 - r)$. Now PN is normal to line if $3(3r - 2) + (-2)(2 - 2r) + (-1)(-3 - r) = 0$ or $r = \frac{1}{2}$. So the coordinates of N the foot of the perpendicular from P to the line are $\left(3 \cdot \frac{1}{2} - 1, 2 - 2 \cdot \frac{1}{2}, -1 - \frac{1}{2}\right)$ or $\left(\frac{1}{2}, 1, -\frac{3}{2}\right)$.

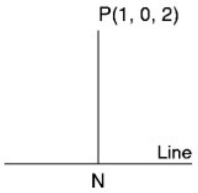

Length of the perpendicular

$$PN = \sqrt{\left(\frac{1}{2}-1\right)^2 + (1-0)^2 + \left(-\frac{3}{2}-2\right)^2}$$

$$= \sqrt{\frac{1}{4}+1+\frac{49}{4}} = \sqrt{\frac{54}{4}} = \frac{3}{2}\sqrt{6}.$$

DR's of *PM* with $r = \frac{1}{2}$ are $[3 \cdot \frac{1}{2} - 2, 2 - 2 \cdot \frac{1}{2}, -3 - \frac{1}{2}]$ i.e., DR's of *PM* are $\frac{1}{2}, -1, \frac{7}{2}$. And *PM* passes through $P(1, 0, 2)$. Therefore the equation of the perpendicular *PM*

$$\frac{x-1}{\frac{1}{2}} = \frac{y-0}{-1} = \frac{z-2}{\frac{7}{2}} \quad \text{or} \quad x - 1 = \frac{y}{-2} = \frac{z-2}{7}.$$

Example 12: Find the equation of the line of the greatest slope through the point (2, 1, 1) in the slant plane $2x + y - 5z = 0$ to the horizontal plane $4x - 3y + 7z = 0$.

Solution: Let l_1, m_1, n_1 be the DR's of the line of intersection *ML* of the two given planes. Since *ML* is $\perp^r$ to both normals,

$$2l_1 + m_1 - 5n_1 = 0, \qquad 4l_1 - 3m_1 + 7n_1 = 0.$$

Solving $\frac{l_1}{4} = \frac{m_1}{17} = \frac{n_1}{5}$. Let *PN* be the line of greatest slope and let l_2, m_2, n_2 be its DR's. Since *PN* and *ML* are perpendicular

$$4l_2 + 17m_2 + 5n_2 = 0$$

Also *PN* is perpendicular to normal of the slant plane $2x + y - 5z = 0$. So

$$2l_2 + m_2 - 5n_2 = 0$$

Solving $\frac{l_2}{3} = \frac{m_2}{-1} = \frac{n_2}{1}$.
Therefore the equation of the line of greatest slope *PN* having DR's $3, -1, 1$ and passing through $P(2, 1, 1)$ is

$$\frac{x-2}{3} = \frac{y-1}{-1} = \frac{z-1}{1}.$$

EXERCISE

1. Find the points where the line $x + y + 4z = 6$, $2x - 3y - 2z = 2$ pierce the coordinate planes.

Ans. $(0, -2, 2), (4, 2, 0), (2, 0, 1)$

2. Transform the general form $3x + y - 2z = 7$, $6x - 5y - 4z = 7$ to symmetrical form and two point form.

   Hint: $(0, 1, -3), (2, 1, 0)$ are two points on the line.

Ans. $\frac{x-2}{2} = \frac{y-1}{0} = \frac{z-0}{3}$

3. Show that the lines $x = y = z + 2$ and $\frac{x-1}{1} = \frac{y}{0} = \frac{z}{2}$ intersect and find the point of intersection.

   Hint: Solve $x - y = 0$, $y - z = 2$, $y = 0$, $2x - z = 2$ simultaneously.

Ans. $(0, 0, -2)$

4. Find the equation plane containing the line $x = y = z$ and

   a. Passing through the line $x + 1 = y + 1 = z$

   b. Parallel to the line $\frac{x+1}{3} = \frac{y}{2} = \frac{z}{-1}$.

Ans. (a) $x - y = 0$; (b) $3x - 4y + z = 0$

5. Show that the line $\frac{x+1}{1} = \frac{y}{-1} = \frac{z-2}{2}$ is in the plane $2x + 4y + z = 0$.

   Hint: $2(1) + 4(-1) + 1(2) = 0$,
   $2(-1) + 4(0) + 2 = 0$

6. Find the equation of the plane containing line $\frac{x-1}{3} = \frac{y-1}{4} = \frac{z-2}{2}$ and parallel to the line $x - 2y + 3z = 4$, $2x - 3y + 4z = 5$.

   Hint: Eq. of 2nd line $\frac{x-0}{\frac{1}{2}} = \frac{y-1}{1} = \frac{z-2}{\frac{1}{2}}$, contains 1st line: $3A + 4B + 2C = 0$. Parallel to 2nd line $A + 2B + C = 0$, $A = 0$, $B = -\frac{1}{2}C$, $D = -\frac{3}{2}C$.

Ans. $y - 2z + 3 = 0$

7. Show that the lines $x + 2y - z = 3$, $3x - y + 2z = 1$ and $2x - 2y + 3z = 2$, $x - y + z + 1 = 0$ are coplanar. Find the equation of the plane containing the two lines.

   Hint: $\frac{x-0}{3} = \frac{y-\frac{7}{3}}{-5} = \frac{z-\frac{5}{3}}{-7}$, $\frac{x-0}{1} = \frac{y-5}{+1} = \frac{z-4}{0}$.

$$\begin{vmatrix} x-0 & y-5 & z-4 \\ 3 & -5 & -7 \\ 1 & 1 & 0 \end{vmatrix} = 0, \quad \text{Expand.}$$

Ans. $7x - 7y + 8z + 3 = 0$

8. Prove that the equation of the plane through the origin containing the line $\frac{x-1}{5} = \frac{y-2}{4} = \frac{z-3}{5}$ is $x - 5y + 3z = 0$.

   Hint: $A(x-1) + B(y-2) + C(z-3) = 0$, $5A + 2B + 3C = 0$, $A + 2B + 3C = 0$,

   $$\text{Expand} \begin{vmatrix} x-1 & y-2 & z-3 \\ 5 & 4 & 5 \\ 1 & 2 & 3 \end{vmatrix} = 0$$

9. Find the image of the point $P(1, 3, 4)$ in the plane $2x - y + z + 3 = 0$.

   Hint: Line through P and $\perp^r$ to plane: $\frac{x-1}{2} = \frac{y-3}{-1} = \frac{z-4}{1}$. Image Q: $(2r + 1, -r + 3, r+4)$. Mid point L of PQ is $(r + 1, -\frac{1}{2}r + 3, \frac{1}{2}r + 4)$. L lies on plane, $r = -2$.

Ans. $(-3, 5, 2)$

10. Determine the point of intersection of the lines

    $$\frac{x-4}{1} = \frac{y+3}{-4} = \frac{z+1}{7}, \frac{x-1}{2} = \frac{y+1}{-3} = \frac{z+10}{8}$$

    Hint: General points: $(r_1 + 4, -4r_1 - 3, 7r_1 - 1)$, $(2r_2 + 1, -3r_2 - 1, 8r_2 - 10)$, Equating $r_1 + 4 = 2r_2 + 1$, $-4r_1 - 3 = -3r_2 - 1$, solving $r_1 = 1, r_2 = 2$.

Ans. $(5, -7, 6)$

11. Show that the lines $\frac{x+3}{2} = \frac{y+5}{3} = \frac{z-7}{-3}$, $\frac{x+1}{4} = \frac{y+1}{5} = \frac{z+1}{-1}$ are coplanar. Find the equation of the plane containing them.

Ans. $6x - 5y - z = 0$

12. Find the equation of the line which passes through the point $(2, -1, 1)$ and intersect the lines $2x + y = 4$, $y + 2z = 0$, and $x + 3z = 4$, $2x + 5z = 8$.

Ans. $x + y + z = 2, x + 2z = 4$

13. Find the coordinates of the foot of the perpendicular from $P(5, 9, 3)$ to the line $\frac{x-1}{2} = \frac{y-2}{3} = \frac{z-3}{4}$. Find the length of the perpendicular and its equations.

Ans. $(3, 5, 7)$, Length: 6, Equation $\frac{x-5}{-2} = \frac{y-9}{-4} = \frac{z-3}{4}$.

14. Find the equation of the line of greatest slope in the slant plane $2x + y - 5z = 12$ and passing through the point $(2, 3, -1)$ given that the line $\frac{x}{4} = \frac{y}{-3} = \frac{z}{7}$ is vertical.

Ans.

15. Find the angle between the line $\frac{x+1}{2} = \frac{y}{3} = \frac{z-3}{6}$ and the plane $3x + y + z = 7$.

    Hint: DR's of line: 2, 3, 6; DR's of normal to plane 3, 1, 1

    $$\cos(90 - \theta) = \sin\theta = \frac{2 \cdot 3 + 3 \cdot 1 + 6 \cdot 1}{\sqrt{4 + 9 + 36}\sqrt{9 + 1 + 1}}.$$

Ans. $\sin\theta = \frac{15}{7\sqrt{11}}$

16. Find the angle between the line $x + y - z = 1$, $2x - 3y + z = 2$ and the plane $3x + y - z + 5 = 0$.

    Hint: DR's of line 2, 3, 5, DR's of normal: $3, 1, -1$

    $$\cos(90 - \theta) = \sin\theta = \frac{2 \cdot 3 + 3 \cdot 1 + 5 \cdot (-1)}{\sqrt{4 + 9 + 25}\sqrt{9 + 1 + 1}}.$$

Ans. $\sin\theta = \frac{4}{\sqrt{38}\sqrt{11}}$

36.5 SHORTEST DISTANCE BETWEEN SKEW LINES

Skew lines: Any two straight lines which do not lie in the same plane are known as skew lines (or non-planar lines). Such lines neither intersect nor are parallel. **Shortest distance between two skew lines:**

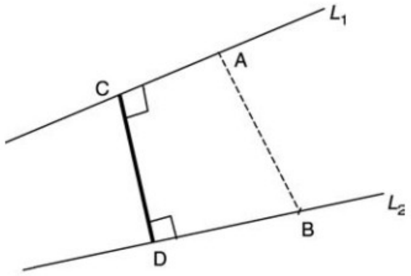

Fig. 36.15

Let L_1 and L_2 be two skew lines; L_1 passing through a given point A and L_2 through a given point

B. Shortest distance between the two skew lines L_1 and L_2 is the length of the line segment CD which is perpendicular to **both** L_1 and L_2. The equation of the shortest distance line CD can be uniquely determined since it intersects both lines L_1 and L_2 at right angles. Now $CD =$ projection of AB on $CD = AB \cos\theta$ where θ is the angle between AB and CD. Since $\cos\theta < 1$, $CD < AB$, thus CD is the shortest distance between the lines L_1 and L_2.

Magnitude (length) and the equations of the line of shortest distance between two lines L_1 and L_2:

Suppose the equation of given line L_1 be

$$\frac{x-x_1}{l_1} = \frac{y-y_1}{m_1} = \frac{z-z_1}{n_1} \tag{1}$$

and of line L_2 be

$$\frac{x-x_2}{l_2} = \frac{y-y_2}{m_2} = \frac{z-z_2}{n_2} \tag{2}$$

Assume the equation of shortest distance line CD as

$$\frac{x-\alpha}{l} = \frac{y-\beta}{m} = \frac{z-\gamma}{n} \tag{3}$$

where (α, β, γ) and (l, m, n) are to be determined. Since CD is perpendicular to both L_1 and L_2,

$$ll_1 + mm_1 + nn_1 = 0$$

$$ll_2 + mm_2 + nn_2 = 0$$

Solving

$$\frac{l}{m_1n_2 - m_2n_1} = \frac{m}{n_1l_2 - n_2l_1} = \frac{n}{l_1m_2 - l_2m_1}$$

$$= \frac{\sqrt{l^2+m^2+n^2}}{\sqrt{(m_1n_2-m_2n_1)^2+(n_1l_2-n_2l_1)^2+(l_1m_2-l_2m_1)^2}}$$

$$= \frac{1}{\sqrt{\sum(m_1n_2-m_2n_1)^2}} = \frac{1}{k}$$

where $k = \sqrt{\sum(m_1n_2 - m_2n_1)^2}$

or $l = \dfrac{m_1n_2 - m_2n_1}{k}, \quad m = \dfrac{n_1l_2 - n_2l_1}{k},$

$$n = \frac{l_1m_2 - l_2m_1}{k} \tag{4}$$

Thus the DC's l, m, n of the shortest distance line CD are determined by (4).

Magnitude of shortest distance $CD =$ projection of AB on CD where $A(x_1, y_1, z_1)$ is a point on L_1 and $B(x_2, y_2, z_2)$ is a point on L_2.

$\therefore$ shortest distance $CD =$

$$= l(x_2 - x_1) + m(y_2 - y_1) + n(z_2 - z_1) \tag{5}$$

In the determinant form,

$$\text{Shortest distance } CD = \frac{1}{k}\begin{vmatrix} x_2 - x_1 & y_2 - y_1 & z_2 - z_1 \\ l_1 & m_1 & n_1 \\ l_2 & m_2 & n_2 \end{vmatrix} \tag{5$'$}$$

Note: If shortest distance is zero, then the two lines L_1 and L_2 are coplanar.

Equation of the line of shortest distance CD: Observe that CD is coplanar with both L_1 and L_2. Let P_1 be the plane containing L_1 and CD. Equation of plane P_1 containing coplanar lines L_1 and CD is

$$\begin{vmatrix} x - x_1 & y - y_1 & z - z_1 \\ l_1 & m_1 & n_1 \\ l & m & n \end{vmatrix} = 0 \tag{6}$$

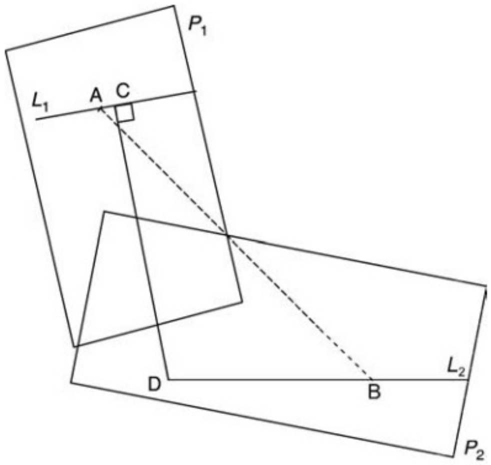

Fig. 36.16

Similarly, equation of plane P_2 containing L_2 and CD is

$$\begin{vmatrix} x - x_2 & y - y_2 & z - z_2 \\ l_2 & m_2 & n_2 \\ l & m & n \end{vmatrix} = 0 \tag{7}$$

Equations (6) and (7) together give the equation of the line of shortest distance.

Points of intersection C and D with L_1 and L_2: Any general point C^* on L_1 is

$$(x_1 + l_1 r_1, \quad y_1 + m_1 r_1, \quad z_1 + n_1 r_1)$$

and any general point D^* on L_2 is

$$(x_2 + l_2 r_2, \quad y_2 + m_2 r_2, \quad z_2 + n_2 r_2)$$

$$\text{DR's of } C^*D^*: (x_2 - x_1 + l_2 r_2 - l_1 r_1,\ y_2 - y_1 + m_2 r_2 - m_1 r_1,\ z_2 - z_1 + n_2 r_2 - n_1 r_1)$$

If C^*D^* is $\perp^r$ to both L_1 and L_2, we get two equations for the two unknowns r_1 and r_2. Solving and knowing r_1 and r_2, the coordinates of C and D are determined. Then the magnitude of CD is obtained by length formula, and equation of CD by two point formula.

Parallel planes: Shortest distance CD = perpendicular distance from any point on L_1 to the plane parallel to L_1 and containing L_2.

WORKED OUT EXAMPLES

Example 1: Find the magnitude and equation of the line of shortest distance between the lines

$$\frac{x-1}{2} = \frac{y-2}{3} = \frac{z-3}{4},$$

$$\frac{x-2}{3} = \frac{y-4}{4} = \frac{z-5}{5}.$$

Solution: Point $A(x_1, y_1, z_1)$ on first line is $(1, 2, 3)$ and $B(x_2, y_2, z_2)$ on second line is $(2, 4, 5)$. Also (l_1, m_1, n_1) are $(2, 3, 4)$ and $(l_2, m_2, n_2) = (3, 4, 5)$. Then

$$\begin{aligned} k^2 &= (m_1 n_2 - m_2 n_1)^2 + (n_1 l_2 - n_2 l_1)^2 + (l_1 m_2 - l_2 m_1)^2 \\ &= (15-16)^2 + (12-10)^2 + (8-9)^2 \\ &= 1 + 4 + 1 = 6 \quad \text{or} \quad k = \sqrt{6}. \end{aligned}$$

So DR's is of line of shortest of distance: $-\frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}}, -\frac{1}{\sqrt{6}}$.

$$\text{Shortest distance} = \frac{1}{k}\begin{vmatrix} x_2 - x_1 & y_2 - y_1 & z_2 - z_1 \\ l_1 & m_1 & n_1 \\ l_2 & m_2 & n_2 \end{vmatrix}$$

$$\begin{aligned} &= \begin{vmatrix} 1 & 2 & 2 \\ 2 & 3 & 4 \\ 3 & 4 & 5 \end{vmatrix} \frac{1}{\sqrt{6}} \\ &= \frac{(15-16) - 2(10-12) + 2(8-9)}{\sqrt{6}} \\ &= \frac{-1+4-2}{\sqrt{6}} = \frac{1}{\sqrt{6}}. \end{aligned}$$

Equation of shortest distance line:

$$\begin{vmatrix} x-1 & y-2 & z-3 \\ 2 & 3 & 4 \\ -\frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \end{vmatrix} = 0 \quad \text{or} \quad 11x + 2y - 7z + 6 = 0$$

and

$$\begin{vmatrix} x-1 & y-4 & z-5 \\ 2 & 3 & 4 \\ -\frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \end{vmatrix} = 0 \quad \text{or} \quad 7x + y - 5z + 7 = 0.$$

Example 2: Determine the points of intersection of the line of shortest distance with the two lines

$$\frac{x-3}{3} = \frac{y-8}{-1} = \frac{z-3}{1}; \ \frac{x+3}{-3} = \frac{y+7}{2} = \frac{z-6}{4}.$$

Also find the magnitude and equation of shortest distance.

Solution: Any general point C^* on first line is $(3 + 3r_1, 8 - r_1, 3 + r_1)$ and any general point D^* on the second line is $(-3 - 3r_2, -7 + 2r_2, 6 - 4r_2)$. DR's of C^*D^* are $(6 + 3r_1 + 3r_2, 15 - r_1 - 2r_2, -3 + r_1 - 4r_2)$. If C^*D^* is $\perp^r$ to both the given lines, then

$$3(6+3r_1+3r_2) - 1(15-r_1-2r_2) + 1(-3+r_1-4r_2) = 0$$

$$-3(6+3r_1+3r_2) + 2(15-r_1-2r_2) + 4(-3+r_1-4r_2) = 0$$

Solving for r_1 and r_2, $11r_1 - 7r_2 = 0, +7r_1 + 29r_2 = 0$ so $r_1 = r_2 = 0$. Then the points of intersection of shortest distance line CD with the given two lines are $C(3, 8, 3)$, $D(-3, -7, 6)$.

$$\begin{aligned} \text{Length of } CD &= \sqrt{(-6)^2 + (-15)^2 + (3)^2} \\ &= \sqrt{270} = 3\sqrt{30} \end{aligned}$$

$$\text{Equation } CD\text{:} \quad \frac{x-3}{-3-3} = \frac{y-8}{-7-8} = \frac{z-3}{6-3}$$

$$\text{i.e.,} \quad \frac{x-3}{-6} = \frac{y-8}{-15} = \frac{z-3}{3}.$$

Example 3: Calculate the length and equation of line of shortest distance between the lines

$$5x - y - z = 0, \quad x - 2y + z + 3 = 0 \qquad (1)$$

$$7x - 4y - 2z = 0, \quad x - y + z - 3 = 0 \qquad (2)$$

Solution: Any plane containing the second line (2) is

$$(7x - 4y - 2z) + \mu(x - y + z - 3) = 0$$

$$\text{or } (7 + \mu)x + (-4 - \mu)y + (-2 + \mu)z - 3\mu = 0 \quad (3)$$

DR's of first line (1) are $(l, m, n) = (-3, -6, -9)$ obtained from:

$$\begin{array}{ccc} l & m & n \\ 5 & -1 & -1 \\ l & -2 & 1 \end{array}$$

The plane (3) will be parallel to the line (1) with $l = -3, m = -6, n = -9$ if

$$-3(7 + \mu) + 6(4 + \mu) + 9(2 - \mu) = 0 \quad \text{or} \quad \mu = \frac{7}{2}$$

Substituting μ in (3), we get the equation of a plane containing line (2) and parallel to line (1) as

$$7x - 5y + z - 7 = 0 \quad (4)$$

To find an arbitrary point on line (1), put $x = 0$. Then $-y - z = 0$ or $y = -z$ and $-2y + z + 3 = 0$, $z = -1$, $y = 1$. $\therefore$ $(0, 1, -1)$ is a point on line (1). Now the length of the shortest distance = perpendicular distance of $(0, 1, -1)$ to plane (4)

$$= \frac{0 - 5(1) + (-1) - 7}{\sqrt{49 + 25 + 1}} = \left|\frac{-13}{\sqrt{75}}\right| = \frac{13}{\sqrt{75}} \quad (5)$$

Equation of any plane through line (1) is

$$5x - y - z + \lambda(x - 2y + z + 3) = 0$$

$$\text{or } (5 + \lambda)x + (-y - 2\lambda)y + (-1 + \lambda)z + 3\lambda = 0 \quad (6)$$

DR's of line (2) are $(l, m, n) = (2, 3, 1)$ obtained from

$$\begin{array}{ccc} l & m & n \\ 7 & -4 & -2 \\ 1 & -1 & 1 \end{array}$$

plane (6) will be parallel to line (2) if

$$2(5 + \lambda) + 3(-y - 2\lambda) + 1(-1 + \lambda) = 0 \quad \text{or} \quad \lambda = 2.$$

Thus the equation of plane containing line (1) and parallel to line (2) is

$$7x - 5y + z + 6 = 0 \quad (7)$$

Hence equation of the line of shortest distance is given by (6) and (7) together.

Aliter: A point on line (2) is $(0, -1, 2)$ obtained by putting $x = 0$ and solving (2). Then the length of shortest distance = perpendicular distance of $(0, -1, 2)$ to the plane (7) $= \frac{0+5+2+6}{\sqrt{75}} = \frac{13}{\sqrt{75}}$

Note: By reducing (1) and (2) to symmetric forms

$$\frac{x - \frac{1}{3}}{1} = \frac{y - \frac{5}{3}}{2} = \frac{z}{3}$$

$$\frac{x + 4}{1} = \frac{y + 7}{\frac{3}{2}} = \frac{z}{\frac{1}{2}}$$

The problem can be solved as in above worked Example 1.

Example 4: Show that the lines $\frac{x-1}{2} = \frac{y-2}{3} = \frac{z-3}{4}; \frac{x-2}{3} = \frac{y-3}{4} = \frac{z-4}{5}$ are coplanar.

Solution: Shortest distance between the two lines is

$$\begin{vmatrix} 2-1 & 3-2 & 4-3 \\ 2 & 3 & 4 \\ 3 & 4 & 5 \end{vmatrix} = \begin{vmatrix} 1 & 1 & 1 \\ 2 & 3 & 4 \\ 3 & 4 & 5 \end{vmatrix} \begin{aligned} &= (-1) - (-2) + (-1) \\ &= 0 \end{aligned}$$

$\therefore$ Lines are coplanar.

Example 5: If a, b, c are the lengths of the edges of a rectangular parallelopiped, show that the shortest distance between a diagonal and an edge not meeting the diagonal is $\frac{bc}{\sqrt{b^2+c^2}} \left(\text{or} \frac{ca}{\sqrt{c^2+a^2}} \text{ or } \frac{ab}{\sqrt{a^2+b^2}}\right)$.

Solution: Choose coterminus edges *OA*, *OB*, *OC* along the X, Y, Z axes. Then the coordinates are $A(a, 0, 0)$, $B(0, b, 0)$, $C(0, 0, c)$, $E(a, b, 0)$, $D(0, b, c)$, $G(a, 0, c)F(a, b, c)$ etc. so that $OA = a$, $OB = b$, $OC = c$.
To find the shortest distance between a diagonal *OF* and an edge *GC*. Here *GC* does not interest *OF*

$$\text{Equation of the line } OF\text{: } \frac{x - 0}{a - 0} = \frac{y - 0}{b - 0} = \frac{z - 0}{c - 0}$$

$$\text{or} \quad \frac{x}{a} = \frac{y}{b} = \frac{z}{c} \quad (1)$$

$$\text{Equation of the line } GC\text{: } \frac{x - 0}{a - 0} = \frac{y - 0}{b - 0} = \frac{z - c}{c - c}$$

$$\text{or} \quad \frac{x}{1} = \frac{y}{0} = \frac{z - c}{0} \quad (2)$$

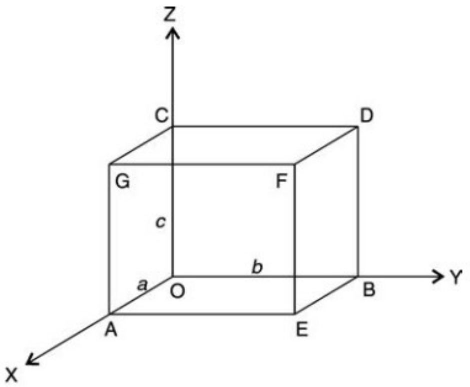

Fig. 36.17

Equation of a plane containing line (1) and parallel to (2) is

$$\begin{vmatrix} x & y & z \\ a & b & c \\ 1 & 0 & 0 \end{vmatrix} = 0 \qquad \text{or} \qquad cy - bz = 0 \qquad (3)$$

Shortest distance = Length of perpendicular drawn from a point say $C(0, 0, c)$ to the plane (3)

$$= \frac{c \cdot 0 - b \cdot c}{\sqrt{0^2 + c^2 + b^2}} = \frac{bc}{\sqrt{c^2 + b^2}}.$$

In a similar manner, it can be proved that the shortest distance between the diagonal OF and non-intersecting edges AN and AM are respectively $\frac{ca}{\sqrt{c^2+a^2}}$, $\frac{ab}{\sqrt{a^2+b^2}}$.

EXERCISE

1. Determine the magnitude and equation of the line of shortest distance between the lines. Find the points of intersection of the shortest distance line, with the given lines

   $\frac{x-8}{3} = \frac{y+9}{-16} = \frac{z-10}{7}$, $\quad \frac{x-15}{3} = \frac{y-29}{8} = \frac{z-5}{-5}$.

Ans. $14, 117x + 4y - 41z - 490 = 0, 9x - 4y - z = 14$, points of intersection $(5, 7, 3)$, $(9, 13, 15)$.

2. Calculate the length, points of intersection, the equations of the line of shortest distance between the two lines

   $\frac{x+1}{2} = \frac{y+1}{3} = \frac{z+1}{4}$, $\quad \frac{x+1}{3} = \frac{y}{4} = \frac{z}{5}$.

Ans. $\frac{1}{\sqrt{6}}, \frac{x-\frac{5}{3}}{\frac{1}{6}} = \frac{y-3}{-\frac{1}{2}} = \frac{z-\frac{15}{2}}{\frac{1}{6}}, \left(\frac{5}{3}, 3, \frac{13}{3}\right),$ $\left(\frac{3}{2}, \frac{10}{3}, \frac{25}{6}\right)$.

3. Find the magnitude and equations of shortest distance between the two lines

   $\frac{x-1}{2} = \frac{y-2}{3} = \frac{z-3}{4}$, $\quad \frac{x-2}{3} = \frac{y-4}{4} = \frac{z-5}{5}$.

Ans. $\frac{1}{\sqrt{6}}, 11x + 2y - 7z + 6 = 0, 7x + y - 5z + 7 = 0$.

4. Show that the shortest distance between the lines $\frac{x}{2} = \frac{y}{-3} = \frac{z}{1}$ and $\frac{x-2}{3} = \frac{y-1}{-5} = \frac{z+2}{2}$ is $\frac{1}{\sqrt{3}}$ and its equations are $4x + y - 5z = 0, 7x + y - 8z = 31$.

5. Determine the points on the lines $\frac{x-6}{3} = \frac{y-7}{-1} = \frac{z-4}{1}$, $\frac{x}{-3} = \frac{y+9}{2} = \frac{z-2}{4}$ which are nearest to each other. Hence find the shortest distance between the lines and find its equations.

Ans. $(3, 8, 3)$, $(-3, -7, 6)$, $3\sqrt{30}$, $\frac{x-3}{2} = \frac{y-8}{5} = \frac{z-3}{-1}$.

6. Prove that the shortest distance between the two lines $\frac{x-1}{3} = \frac{y-4}{2} = \frac{z-4}{-2}$, $\frac{x+1}{2} = \frac{y-1}{-4} = \frac{z+2}{1}$ is $\frac{120}{\sqrt{341}}$

   Hint: Equation of a plane passing through the first lines nad parallel to the second line is $6x + 7y + 16z = 98$. A point on second line is $(-1, 1, -2)$. Perpendicular distance $= \frac{6(-1)+7(1)+16(-2)}{\sqrt{6^2+7^2+16^2}}$.

7. Find the length and equations of shortest distance between the lines $x - y + z = 0, 2x - 3y + 4z = 0$; and $x + y + 2z - 3 = 0, 2x + 3y + 3z - 4 = 0$.

   Hint: Equations of two lines in symmetric form are $\frac{x}{1} = \frac{y}{2} = \frac{z}{1}$, $\frac{x-5}{-3} = \frac{y+2}{1} = \frac{z}{1}$.

Ans. $\frac{13}{\sqrt{66}}, 3x - y - z = 0, x + 2y + z - 1 = 0$.

8. Determine the magnitude and equations of the line of shortest distance between the lines $\frac{x-3}{2} = \frac{y+15}{-7} = \frac{z-9}{5}$ and $\frac{x+1}{2} = \frac{y-1}{1} = \frac{z-9}{-3}$.

Ans. $4\sqrt{3}$, $-4x + y + 3z = 0$, $4x - 5y + z = 0$ (or $x = y = z$).

9. Obtain the coordinates of the points where the line of shortest distance between the lines $\frac{x-23}{-6} = \frac{y-19}{-4} = \frac{z-25}{3}$ and $\frac{x-12}{-9} = \frac{y-1}{4} = \frac{z-5}{2}$ meets them. Hence find the shortest distance between the two lines.

Ans. $(11, 11, 31), (3, 5, 7), 26$

10. Find the shortest distance between any two opposite edges of a tetrahedron formed by the planes $x + y = 0$, $y + z = 0$, $z + x = 0$, $x + y + z = a$. Also find the point of intersection of three lines of shortest distances.

    Hint: Vertices are $(0, 0, 0)$, $(a, -a, a)$, $(-a, a, a)$, $(a, a, -a)$.

Ans. $\frac{2a}{\sqrt{6}}, (-a, -a, -a)$.

11. Find the shortest distance between the lines PQ and RS where $P(2, 1, 3)$, $Q(1, 2, 1)$, $R(-1, -2, -2)$, $S(-1, 4, 0)$.

Ans. $3\sqrt{2}$

36.6 THE RIGHT CIRCULAR CONE

Cone

A **cone** is a surface generated by a straight line (known as **generating line or generator**) passing through a fixed point (known as **vertex**) and satisfying a condition, for example, it may intersect a given curve (known as **guiding curve**) or touches a given surface (say a sphere). Thus cone is a set of points on its generators. Only cones with second degree equations known as quadratic cones are considered here. In particular, quadratic cones with vertex at origin are homogeneous equations of second degree.

Equation of cone with vertex at (α, β, γ) and the conic $ax^2 + 2hxy + by^2 + 2gx + 2fy + c = 0$, $z = 0$ as the guiding curve:

The equation of any line through vertex (α, β, γ) is

$$\frac{x-\alpha}{l} = \frac{y-\beta}{m} = \frac{z-\gamma}{n} \qquad (1)$$

(1) will be generator of the cone if (1) intersects the given conic

$$ax^2 + 2hxy + by^2 + 2gx + 2fy + c = 0, \; z = 0 \qquad (2)$$

Since (1) meets $z = 0$, put $z = 0$ in (1), then the point $\left(\alpha - \frac{l\gamma}{n}, \beta - \frac{m\gamma}{n}, 0\right)$ will lie on the conic (2), if

$$a\left(\alpha - \frac{l\gamma}{n}\right)^2 + 2h\left(\alpha - \frac{l\gamma}{n}\right)\left(\beta - \frac{m\gamma}{n}\right) + b\left(\beta - \frac{m\gamma}{n}\right)^2 + \\ + 2g\left(\alpha - \frac{l\gamma}{n}\right) + 2f\left(\beta - \frac{m\gamma}{n}\right) + c = 0 \qquad (3)$$

From (1)

$$\frac{l}{n} = \frac{x-\alpha}{z-\gamma}, \qquad \frac{m}{n} = \frac{y-\beta}{z-\gamma} \qquad (4)$$

Eliminate l, m, n from (3) using (4),

$$a\left(\alpha - \frac{x-\alpha}{z-\gamma}\cdot\gamma\right)^2 + \\ + 2h\left(\alpha - \frac{x-\alpha}{z-\gamma}\cdot\gamma\right)\left(\beta - \frac{y-\beta}{z-\gamma}\cdot\gamma\right) + \\ + b\left(\beta - \frac{y-\beta}{z-\gamma}\cdot\gamma\right)^2 + 2g\left(\alpha - \frac{x-\alpha}{z-\gamma}\cdot\gamma\right) + \\ + 2f\left(\beta - \frac{\gamma-\beta}{z-\gamma}\cdot\gamma\right) + c = 0$$

or

$$a(\alpha z - x\gamma)^2 + 2h(\alpha z - x\gamma)(\beta z - y\gamma) + \\ + b(\beta z - y\gamma)^2 + 2g(\alpha z - x\gamma)(z - \gamma) + \\ + 2f(\beta z - y\gamma)(z - \gamma) + c(z - \gamma)^2 = 0$$

or

$$a(x-\alpha)^2 + b(y-\beta)^2 + c(z-\gamma)^2 + \\ + 2f(z-\gamma)(y-\beta) + 2g(x-\alpha)(z-\gamma) + \\ + 2h(x-\alpha)(y-\beta) = 0 \qquad (5)$$

Thus (5) is the equation of the quadratic cone with vertex at (α, β, γ) and guiding curve as the conic (2).

Special case: Vertex at origin $(0, 0, 0)$. Put $\alpha = \beta = \gamma = 0$ in (5). Then (5) reduces to

$$ax^2 + by^2 + cz^2 + 2fzy + 2gxz + 2hxy = 0 \qquad (6)$$

Equation (6) which is a homogeneous and second degree in x, y, z is the equation of cone with vertex at origin.

Right circular cone

A right circular cone is a surface generated by a line (**generator**) through a fixed point (**vertex**) making a constant angle θ (**semi-vertical angle**) with the fixed line (**axis**) through the fixed point (**vertex**). Here the guiding curve is a circle with centre at c. Thus every section of a right circular cone by a plane perpendicular to its axis is a circle.

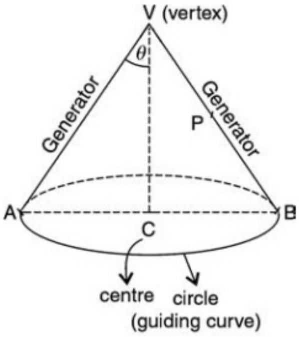

Fig. 36.18

Equation of a right circular cone: with vertex at (α, β, γ), semi vertical angle θ and equation of axis

$$\frac{x-\alpha}{l} = \frac{y-\beta}{m} = \frac{z-\gamma}{n} \tag{1}$$

Let $P(x, y, z)$ be any point on the generating line VB. Then the DC's of VB are proportional to $(x-\alpha, y-\beta, z-\gamma)$. Then

$$\cos\theta = \frac{l(x-\alpha)+m(y-\beta)+n(z-\gamma)}{\sqrt{(l^2+m^2+n^2)}\sqrt{(x-\alpha)^2+(y-\beta)^2+(z-\gamma)^2}}$$

Rewriting, the required equation of cone is

$$\left[l(x-\alpha)+m(y-\beta)+n(z-\gamma)\right]^2 =$$

$$= (l^2+m^2+n^2)\left[(x-\alpha)^2+(y-\beta)^2+(z-\gamma)^2\right]\cos^2\theta \tag{2}$$

Case 1: If vertex is origin $(0, 0, 0)$ then (2) reduces

$$(lx+my+nz)^2 = (l^2+m^2+n^2)(x^2+y^2+z^2)\cos^2\theta \tag{3}$$

Case 2: If vertex is origin and axis of cone is z-axis (with $l = 0, m = 0, n = 1$) then (2) becomes

$$z^2 = (x^2+y^2+z^2)\cos^2\theta \quad \text{or} \quad z^2\sec^2\theta = x^2+y^2+z^2$$

$$z^2(1+\tan^2\theta) = x^2+y^2+z^2$$

i.e.,
$$x^2+y^2 = z^2\tan^2\theta \tag{4}$$

Similarly, with y-axis as the axis of cone

$$x^2+z^2 = y^2\tan^2\theta$$

with x-axis as the axis of cone

$$y^2+z^2 = x^2\tan^2\theta.$$

If the right circular cone admits sets of three mutually perpendicular generators then the semi-vertical angle $\theta = \tan^{-1}\sqrt{2}$ (since the sum of the coefficients of x^2, y^2, z^2 in the equation of such a cone must be zero i.e., $1+1-\tan^2\theta = 0$ or $\tan\theta = \sqrt{2}$).

WORKED OUT EXAMPLES

Example 1: Find the equation of cone with base curve $\frac{x^2}{a^2}+\frac{y^2}{b^2} = 1, z = 0$ and vertex (α, β, γ). Deduce the case when base curve is $\frac{x^2}{16}+\frac{y^2}{9} = 1, z = 0$ and vertex at $(1, 1, 1)$.

Solution: The equation of any generating line through the vertex (α, β, γ) is

$$\frac{x-\alpha}{l} = \frac{y-\beta}{m} = \frac{z-\gamma}{n} \tag{1}$$

This generator (1) meets $z = 0$ in the point

$$\left(x = \alpha - \frac{l\gamma}{n}, \qquad y = \beta - \frac{m\gamma}{n}, \qquad z = 0\right) \tag{2}$$

Point (2) lies on the generating curve

$$\frac{x^2}{a^2}+\frac{y^2}{b^2} = 1 \tag{3}$$

Substituting (2) in (3)

$$\frac{\left(\alpha - \frac{l\gamma}{n}\right)^2}{a^2} + \frac{\left(\beta - \frac{m\gamma}{n}\right)^2}{b^2} = 1 \tag{4}$$

Eliminating l, m, n from (4) using (1),

$$\frac{\left[\alpha - \left(\frac{x-\alpha}{z-\gamma}\right)\gamma\right]^2}{a^2} + \frac{\left[\beta - \left(\frac{y-\beta}{z-\gamma}\right)\gamma\right]^2}{b^2} = 1$$

$$b^2\left[\alpha(z-\gamma)-\gamma(x-\alpha)\right]^2 + a^2\left[\beta(z-\gamma)-\gamma(y-\beta)\right]^2$$

$$= a^2b^2(z-\gamma)^2$$

Deduction: When $a=4, b=3, \alpha=1, \beta=1, \gamma=1$,

$$9\Big[(z-1)-(x-1)\Big]^2+16\Big[(z-1)-(y-1)\Big]^2$$
$$=144(z-1)^2$$
$$9x^2+16y^2-119z^2-18xz-32yz+288z-144=0.$$

Example 2: Find the equation of the cone with vertex at (1, 0, 2) and passing through the circle $x^2+y^2+z^2=4$, $x+y-z=1$.

Solution: Equation of generator is

$$\frac{x-1}{l}=\frac{y-0}{m}=\frac{z-2}{n} \qquad (1)$$

Any general point on the line (1) is

$$(1+lr, \quad mr, \quad 2+nr). \qquad (2)$$

Since generator (1) meets the plane

$$x+y-z=1 \qquad (3)$$

substitute (2) in (3)

$$(1+lr)+(mr)-(2+nr)=1$$

or $$r=\frac{2}{l+m-n}. \qquad (4)$$

Since generator (1) meets the sphere

$$x^2+y^2+z^2=4 \qquad (5)$$

substitute (2) in (5)

$$(1+lr)^2+(mr)^2+(2+nr)^2=4$$

or $$r^2(l^2+m^2+n^2)+2r(l+2n)+1=0 \qquad (6)$$

Eliminate r from (6) using (4), then

$$\frac{4}{(l+m-n)^2}(l^2+m^2+n^2)+2\frac{2}{(l+m-n)}(l+2n)+1=0$$
$$9l^2+5m^2-3n^2+6lm+2ln+6nm=0 \qquad (7)$$

Eliminate l, m, n from (7) using (1), then

$$9\left(\frac{x-1}{r}\right)^2+5\left(\frac{y}{r}\right)^2-3\left(\frac{z-2}{r}\right)^2+6\left(\frac{x-1}{r}\right)\left(\frac{y}{r}\right)+$$
$$+2\left(\frac{x-1}{r}\right)\left(\frac{z-2}{r}\right)+6\left(\frac{z-2}{r}\right)\left(\frac{y}{r}\right)=0$$

or $$9(x-1)^2+5y^2-3(z-2)^2+6y(x-1)+$$
$$+2(x-1)(x-2)+6(z-2)y=0$$

Vertex (0, 0, 0):

Example 3: Determine the equation of a cone with vertex at origin and base curve given by

a. $ax^2+by^2=2z, \qquad lx+my+nz=p$

b. $ax^2+by^2+cz^2=1, \qquad lx+my+nz=p$

c. $x^2+y^2+z^2=25, \qquad x+2y+2z=9$

Solution: We know that the equation of a quadratic cone with vertex at origin is a homogeneous equation of second degree in x, y, z. By eliminating the non-homogeneous terms in the base curve, we get the required equation of the cone.

a. $2z$ is the term of degree one and is non homogeneous. Solving

$$\frac{lx+my+nz}{p}=1$$

rewrite the equation

$$ax^2+by^2=2\cdot z(1)=2z\left(\frac{lx+my+nz}{p}\right)$$
$$apx^2+bpy^2-2nz^2-2lxz-2myz=0$$

which is the equation of cone.

b. Except the R.H.S. term 1, all other terms are of degree 2 (and homogeneous). Rewriting, the required equation of cone as

$$ax^2+by^2+cz^2=(1)^2=\left(\frac{lx+my+nz}{p}\right)^2$$
$$(ap^2-l^2)x^2+(bp^2-m^2)y^2+(cp^2-n^2)z^2-$$
$$-2lmxy-2mnyz-2lnxz=0$$

c. On similar lines

$$x^2+y^2+z^2=25=25(1)^2=25\left(\frac{x+2y+2z}{9}\right)^2$$
$$56x^2-19y^2-19z^2-100xy-200yz-100xz=0$$

Right circular cone:

Example 4: Find the equation of a right circular cone with vertex at (2, 0, 0), semi-vertical angle $\theta=30°$ and axis is the line $\frac{x-2}{3}=\frac{y}{4}=\frac{z}{6}$.

Solution: Here $\alpha = 2, \beta = 0, \gamma = 0, l = 3, m = 4, n = 6$

$$\frac{\sqrt{3}}{2} = \cos 30 = \cos\theta$$

$$= \frac{l(x-\alpha) + m(y-\beta) + n(z-\gamma)}{\sqrt{(l^2+m^2+n^2)[(x-\alpha)^2+(y-\beta)^2+(z-\gamma)^2]}}$$

$$\frac{\sqrt{3}}{2} = \frac{3(x-2)+4y+6z}{\sqrt{9+16+36}\sqrt{(x-2)^2+y^2+z^2}}$$

$$183[(x-2)^2+y^2+z^2] = 4[3(x-2)+4y+6z]^2$$

$$147x^2 + 119y^2 + 39z^2 - 192yz - 144zx - 96xy -$$

$$-588x + 192y + 288z + 588 = 0$$

Vertex (0, 0, 0):

Example 5: Find the equation of the right circular cone which passes through the line $2x = 3y = -5z$ and has $x = y = z$ as its axis.

Solution: DC's of the generator $2x = 3y = -5z$ are $\frac{1}{2}, \frac{1}{3}, -\frac{1}{5}$. DC's of axis are $\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}$. Point of intersection of the generator and axis is (0, 0, 0). Now

$$\cos\theta = \frac{\frac{1}{2}\cdot\frac{1}{\sqrt{3}}+\frac{1}{3}\cdot\frac{1}{\sqrt{3}}-\frac{1}{5}\cdot\frac{1}{\sqrt{3}}}{\sqrt{\frac{1}{3}+\frac{1}{3}+\frac{1}{3}}\sqrt{\frac{1}{4}+\frac{1}{9}+\frac{1}{25}}} = \frac{\frac{19}{30}}{\sqrt{\frac{361}{900}}}\cdot\frac{1}{\sqrt{3}} = \frac{1}{\sqrt{3}}$$

Equation of cone with vertex at origin

$$\frac{1}{\sqrt{3}} = \cos\theta = \frac{\frac{1}{\sqrt{3}}(x+y+z)}{1\sqrt{x^2+y^2+z^2}}$$

$$x^2+y^2+z^2 = (x+y+z)^2$$

$$xy + yz + zx = 0.$$

Example 6: Determine the equation of a right circular cone with vertex at origin and the guiding curve circle passing through the points (1, 2, 2), (1, −2, 2)(2, −1, −2).

Solution: Let l, m, n be the DC's of OL the axis of the cone. Let θ be the semi- vertical angle. Let $A(1, 2, 2)$, $B(1, -2, 2)$, $C(2, -1, -2)$ be the three points on the guiding circle. Then the lines OA, OB, OC make the same angle θ with the axis OL. The DC's of OA, OB, OC are proportional to (1, 2, 2)(1, −2, 2)(2, −1, −2) respectively. Then

$$\cos\theta = \frac{l(1)+m(2)+n(2)}{\sqrt{1}\cdot\sqrt{1+4+4}} = \frac{l+2m+2n}{3} \qquad (1)$$

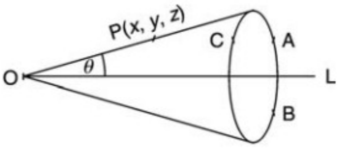

Fig. 36.19

Similarly,

$$\cos\theta = \frac{l(1)+m(-2)+n(2)}{\sqrt{1}\sqrt{1+4+4}} = \frac{l-2m+2n}{3} \qquad (2)$$

$$\cos\theta = \frac{2l-m-2n}{3} \qquad (3)$$

From (1) and (2), $4m = 0$ or $m = 0$.
From (2) and (3), $l + m - 4n = 0, l - 4n = 0$ or $l = 4n$.

DC's $\quad \frac{l}{4} = \frac{m}{0} = \frac{n}{1}$ or $\frac{l}{\frac{4}{\sqrt{17}}} = \frac{m}{0} = \frac{n}{\frac{1}{\sqrt{17}}}$.

From (1) $\quad \cos\theta = \dfrac{\frac{4}{\sqrt{17}}+2\cdot 0+2\frac{1}{\sqrt{17}}}{3} = \dfrac{2}{\sqrt{17}}$.

Equation of right circular cone is

$$(l^2+m^2+n^2)(x^2+y^2+z^2)\cos^2\theta = (lx+my+nz)^2$$

$$\left(\frac{16}{17}+0+\frac{1}{17}\right)(x^2+y^2+z^2)\frac{4}{17} = \left(\frac{4}{\sqrt{17}}x+0+\frac{1}{\sqrt{17}}z\right)^2$$

$$4(x^2+y^2+z^2) = (4x+z)^2$$

$$12x^2 - 4y^2 - 3z^2 + 8xz = 0$$

is the required equation of the cone.

EXERCISE

1. Find the equation of the cone whose vertex is (3, 1, 2) and base circle is $2x^2 + 3y^2 = 1$, $z = 1$.

Ans. $2x^2 + 3y^2 + 20z^2 - 6yz - 12xz + 12x + 6y$
$-38z + 17 = 0$

2. Find the equation of the cone whose vertex is origin and guiding curve is $\frac{x^2}{4}+\frac{y^2}{9}+\frac{z^2}{1}=1, x+y+z=1$.

Ans. $27x^2+32y^2+72(xy+yz+zx)=0$.

3. Determine the equation of the cone with vertex at origin and guiding curve $x^2+y^2+z^2-x-1=0, x^2+y^2+z^2+y-z=0$.

   Hint: Guiding curve is circle in plane $x+y=1$. Rewrite $x^2+y^2+z^2-x(x+y)-(x+y)^2=0$.

Ans. $x^2+3xy-z^2=0$

4. Show that the equation of cone with vertex at origin and base circle $x=a, y^2+z^2=b^2$ is $a^2(y^2+z^2)=b^2x^2$. Further prove that the section of the cone by a plane parallel to the XY-plane is a hyperbola.

Ans. $b^2x^2-a^2y^2=a^2c^2, z=c$ (put $z=c$ in equation of cone)

5. Find the equation of a cone with vertex at origin and guiding curve is the circle passing through the X, Y, Z intercepts of the plane $\frac{x}{a}+\frac{y}{b}+\frac{z}{c}=1$.

Ans. $a(b^2+c^2)yz+b(c^2+a^2)zx+c(a^2+b^2)xy=0$

6. Write the equation of the cone whose vertex is $(1, 1, 0)$ and base is $y^2+z^2=9, x=0$.

   Hint: Substitute $\left(0, 1-\frac{m}{l}, -\frac{n}{l}\right)$ in base curve and eliminate $\frac{m}{l}=\frac{y-1}{x-1}, \frac{n}{l}=\frac{z}{z-1}$.

Ans. $x^2+y^2+z^2-2xy=0$

Right circular cone (R.C.C.)

7. Find the equation of R.C.C. with vertex at $(2, 3, 1)$, axis parallel to the line $-x=\frac{y}{2}=z$ and one of its generators having DC's proportional to $(1, -1, 1)$.

   Hint: $\cos\theta=\frac{-1-2+1}{\sqrt{6}\sqrt{3}}, l=-1, m=2, n=1, \alpha=2, \beta=3, \gamma=1$.

Ans. $x^2-8y^2+z^2+12xy-12yz+6zx-46x++36y+22z-19=0$

8. Determine the equation of R.C.C. with vertex at origin and passes through the point $(1, 1, 2)$ and axis line $\frac{x}{2}=\frac{-y}{4}=\frac{z}{3}$.

   Hint: $\cos\theta=\frac{2-4+6}{\sqrt{6}\sqrt{29}}$, DC's of generator: 1, 1, 2, axis: 2, −4, 3

Ans. $4x^2+40y^2+19z^2-48xy-72yz+36xz=0$

9. Find the equation of R.C.C. whose vertex is origin and whose axis is the line $\frac{x}{1}=\frac{y}{2}=\frac{z}{3}$ and which has semi- vertical angle of 30°

   Hint: $\cos 30=\frac{\sqrt{3}}{2}=\frac{x(1)+y(2)+z(3)}{\sqrt{(x^2+y^2+z^2)}\sqrt{1+4+9}}$

Ans. $19x^2+13y^2+3z^2-8xy-24yz-12zx=0$

10. Obtain the equation of R.C.C. generated when the straight line $2y+3z=6, x=0$ revolves about z-axis.

    Hint: Vertex (0, 0, 2), generator $\frac{x}{0}=\frac{y}{3}=\frac{z-2}{-2}, \cos\theta=-\frac{2}{\sqrt{13}}$.

Ans. $4x^2+4y^2-9z^2+36z-36=0$

11. Lines are drawn from the origin with DC's proportional to $(1, 2, 2), (2, 3, 6), (3, 4, 12)$. Find the equation of R.C.C.

    Hint: $\cos\alpha=\frac{l+2m+2n}{3}=\frac{2l+3m+6n}{7}=\frac{3l+4m+12n}{13}$ $\frac{l}{-1}=\frac{m}{1}=\frac{n}{1}, \cos\alpha=\frac{1}{\sqrt{3}}$, DC's of axis: −1, 1, 1.

Ans. $xy-yz+zx=0$

12. Determine the equation of the R.C.C. generated by straight lines drawn from the origin to cut the circle through the three points $(1, 2, 2), (2, 1, -2)$, and $(2, -2, 1)$.

    Hint: $\cos\alpha=\frac{l+2m+2n}{3}=\frac{2l+m-2n}{3}=\frac{2l-2m+n}{3}$ $\frac{l}{5}=\frac{m}{1}=\frac{n}{1}, \cos\alpha=\frac{5+2+2}{3\sqrt{27}}=\frac{1}{\sqrt{3}}$.

Ans. $8x^2-4y^2-4z^2+5xy+5zx+yz=0$

36.7 THE RIGHT CIRCULAR CYLINDER

A cylinder is the surface generated by a straight line (known as **generator**) which is parallel to a fixed straight line (known as **axis**) and satisfies a condition; for example, it may intersect a fixed curve (known as the **guiding curve**) or touch a given surface. A **right circular cylinder** is a cylinder whose surface is generated by revolving the generator at a fixed distance (known as the **radius**) from the axis; i.e., the guiding curve in this case is a circle. In fact, the intersection of the right circular cylinder with any plane perpendicular to axis of the cylinder is a circle.

Equation of a cylinder with generators parallel to the line $\frac{x}{l}=\frac{y}{m}=\frac{z}{n}$ and guiding curve conic $ax^2+by^2+2hxy+2gx+2fy+c=0, z=0$.

Let $P(x_1, y_1, z_1)$ be any point on the cylinder. The equation of the generator through $P(x_1, y_1, z_1)$ which is parallel to the given line

$$\frac{x}{l}=\frac{y}{m}=\frac{z}{n} \tag{1}$$

is
$$\frac{x-x_1}{l}=\frac{y-y_1}{m}=\frac{z-z_1}{n} \tag{2}$$

Since (2) meets the plane $z=0$,

$$\therefore \quad \frac{x-x_1}{l}=\frac{y-y_1}{m}=\frac{0-z_1}{n}$$

or
$$x=x_1-\frac{l}{n}z_1,\ y=y_1-\frac{m}{n}z_1 \tag{3}$$

Since this point (3) lies on the conic

$$ax^2+by^2+2hxy+2gx+2fy+c=0 \tag{4}$$

substitute (3) in (4). Then

$$a\left(x_1-\frac{l}{n}z_1\right)^2+b\left(y_1-\frac{m}{n}z_1\right)^2+$$
$$+2h\left(x_1-\frac{l}{n}z_1\right)\left(y_1-\frac{m}{n}z_1\right)+2g\left(x_1-\frac{l}{n}z_1\right)+$$
$$+2f\left(y_1-\frac{m}{n}z_1\right)+c=0.$$

The required equation of the cylinder is

$$a(nx-lz)^2+b(ny-mz)^2+2h(nx-lz)(ny-mz)+$$
$$+2ng(nx-lz)+2nf(ny-mz)+cn^2=0 \tag{5}$$

where the subscript 1 is droped because (x_1, y_1, z_1) is any general point on the cylinder.

Corollary 1: The equation of a cylinder with axis parallel to z-axis is obtained from (5) by putting $l=0, m=0, n=1$ which are the DC's of z-axis: i.e.,

$$ax^2+by^2+2hxy+2gx+2fy+c=0$$

which is **free** from z.

Thus the equation of a cylinder whose axis is paralle to x-axis (y-axis or z-axis) is obtained by eliminating the variable x(y or z) from the equation of the conic.

Equation of a right circular cylinder:

a. Standard form: with z-axis as axis and of radius a. Let $P(x, y, z)$ be any point on the cylinder. Then M the foot of the perpendicular PM has $(0, 0, z)$ and $PM=a$ (given). Then

$$a=PM=\sqrt{(x-0)^2+(y-0)^2+(z-z)^2}$$
$$x^2+y^2=a^2$$

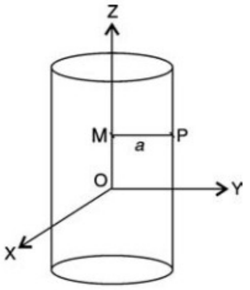

Fig. 36.20

Corollary 2: Similarly, equation of right circular cylinder with y-axis is $x^2+z^2=a^2$, with x-axis is $y^2+z^2=a^2$.

b. General form with the line $\frac{x-\alpha}{l}=\frac{y-\beta}{m}=\frac{z-\gamma}{n}$ as axis and of radius a.

Axis AB passes through the point (α, β, γ) and has DR's l, m, n. Its DC's are $\frac{l}{k}, \frac{m}{k}, \frac{n}{k}$ where $k=\sqrt{l^2+m^2+n^2}$.

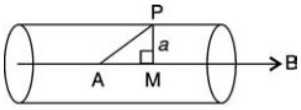

Fig. 36.21

From the right angled triangle APM

$$AP^2=PM^2+AM^2$$
$$(x-\alpha)^2+(y-\beta)^2+(z-\gamma)^2$$
$$=a^2+\left[l(x-\alpha)+m(y-\beta)+n(z-\gamma)\right]^2$$

which is the required equation of the cylinder (Here AM is the projection of AP on the line AB is equal to $l(x-\alpha)+m(y-\beta)+n(z-\gamma)$).

Enveloping cylinder of a sphere is the locus of the tangent lines to the sphere which are parallel to a given line. Suppose

$$x^2+y^2+z^2=a^2 \tag{1}$$

is the sphere and suppose that the generators are parallel to the given line

$$\frac{x}{l}=\frac{y}{m}=\frac{z}{n} \tag{2}$$

Then for any point $P(x_1, y_1, z_1)$ on the cylinder, the equation of the generating line is

$$\frac{x-x_1}{l}=\frac{y-y_1}{m}=\frac{z-z_1}{n} \tag{3}$$

Any general point on (3) is

$$(x_1+lr, \qquad y_1+mr, \qquad z_1+nr) \tag{4}$$

By substituting (4) in (1), we get the points of intersection of the sphere (1) and the generating line (3) i.e.,

$$(x_1+lr)^2+(y_1+mr)^2+(z_1+nr)^2=a^2$$

Rewriting as a quadratic in r, we have

$$(l^2+m^2+n^2)r^2+2(lx_1+my_1+nz_1)r+ +(x_1^2+y_1^2+z_1^2-a^2)=0 \tag{5}$$

If the roots of (5) are equal, then the generating line (3) meets (touches) the sphere in a single point i.e., when the discriminant of the quadratic in r is zero.

or $$4(lx_1+my_1+nz_1)^2-4(l^2+m^2+n^2)\times \times(x_1^2+y_1^2+z_1^2-a^2)=0$$

Thus the required equation of the enveloping cylinder is

$$(lx+my+nz)^2=(l^2+m^2+n^2)(x^2+y^2+z^2-a^2)$$

where the subscript 1 is droped to indicate that (x, y, z) is a general point on the cylinder.

WORKED OUT EXAMPLES

Example 1: Find the equation of the quadratic cylinder whose generators intersect the curve $ax^2+by^2+cz^2=k$, $lx+my+nz=p$ and parallel to the y-axis. Deduce the case for $x^2+y^2+z^2=1$ and $x+y+z=1$ and parallel to y-axis

Solution: Eliminate y between

$$ax^2+by^2+cz^2=k \tag{1}$$

and $$lx+my+nz=p \tag{2}$$

Solving (2) for y, we get

$$y=\frac{p-lx-nz}{m} \tag{3}$$

Substitute (3) in (1), we have

$$ax^2+b\left(\frac{p-lx-nz}{m}\right)^2+cz^2=k.$$

The required equation of the cylinder is

$$(am^2+l^2)x^2+(bn^2+m^2c)z^2-2pblx -2npbz+2blnxz+(bp^2-m^2k)=0.$$

Deduction: Put $a=1, b=1, c=1, k=1, l=m=n=p=1$

$$2x^2+2z^2-2x-2z+2xz=0$$

or $$x^2+z^2+xz-x-z=0.$$

Example 2: If l, m, n are the DC's of the generators and the circle $x^2+y^2=a^2$ in the XY-plane is the guiding curve, find the equation of the cylinder. Deduce the case when $a=4, l=1, m=2, n=3$.

Solution: For any point $P(x_1, y_1, z_1)$ on the cylinder, the equation of the generating line through P is

$$\frac{x-x_1}{l}=\frac{y-y_1}{m}=\frac{z-z_1}{n} \tag{1}$$

Since the line (1) meets the guiding curve $x^2+y^2=a^2, z=0$,

$$\frac{x-x_1}{l}=\frac{y-y_1}{m}=\frac{0-z_1}{n}$$

or $$x=x_1-\frac{lz_1}{n}, \qquad y=y_1-\frac{mz_1}{n} \tag{2}$$

This point (2) lies on the circle $x^2+y^2=a^2$ also. Substituting (2) in the equation of circle, we have

$$\left(x_1-\frac{lz_1}{n}\right)^2+\left(y_1-\frac{mz_1}{n}\right)^2=a^2$$

or $$(nx-lz)^2+(ny-mz)^2=n^2a^2$$

is the equation of the cylinder.

Deduction: Equation of cylinder whose generators are parallel to the line $\frac{x}{1}=\frac{y}{2}=\frac{z}{3}$ and pass through the curve $x^2+y^2=16, z=0$. With $a=4, l=1, m=2, n=3$, the required equation of the cylinder is

$$(3x-z)^2+(3y-2z)^2=9(16)=144$$

or $$9x^2+9y^2+5z^2-6zx-12yz-144=0.$$

Example 3: Find the equation of the right circular cylinder of radius 3 and the line $\frac{x-1}{2}=\frac{y-3}{2}=\frac{z-5}{-1}$ as axis.

Solution: Let $A(1, 3, 5)$ be the point on the axis and DR's of AB are $2, 2, -1$ or DC's of AB are $\frac{2}{3}, \frac{2}{3}, -\frac{1}{3}$. Radius $PM=3$ given. Since AM is the projection of AP on AB, we have

$$AM=\frac{2}{3}(x-1)+\frac{2}{3}(y-3)-\frac{1}{3}(z-5)$$

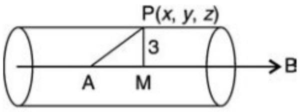

Fig. 36.22

From the right angled triangle APM

$$AP^2=AM^2+MP^2$$
$$(x-1)^2+(y-3)^2+(z-5)^2$$
$$=\left[2\frac{(x-1)}{3}+2\frac{y-3}{3}-1\frac{(z-5)}{3}\right]+9$$
$$9[x^2+1-2x+y^2+9-6y+z^2+25-10z]$$
$$=[2x+2y-z-3]^2+81$$
$$9[x^2+y^2+z^2-2x-6y-10z+35]$$
$$=[4x^2+4y^2+z^2+9+8xy-4xz-12x$$
$$-4yz-12y+6z]+81$$

is the required equation of the cylinder.

Example 4: Find the equation of the enveloping cylinder of the sphere $x^2+y^2+z^2-2y-4z-11=0$ having its generators parallel to the line $x=-2y=2z$.

Solution: Let $P(x_1, y_1, z_1)$ be any point on the cylinder. Then the equation of the generating line through P and parallel to the line $x=-2y=2z$ or $\frac{x}{1}=\frac{y}{-\frac{1}{2}}=\frac{z}{\frac{1}{2}}$ is

$$\frac{x-x_1}{1}=\frac{y-y_1}{-\frac{1}{2}}=\frac{z-z_1}{\frac{1}{2}} \qquad (1)$$

Any general point on (1) is

$$\left(x_1+r, \quad y_1-\frac{1}{2}r, \quad z_1+\frac{1}{2}r\right) \qquad (2)$$

The points of intersection of the line (1) and the sphere

$$x^2+y^2+z^2-2y-4z-11=0 \qquad (3)$$

are obtained by substituting (2) in (3).

$$(x_1+r)^2+\left(y_1-\frac{1}{2}r\right)^2+\left(z_1+\frac{1}{2}r^2\right)^2-2\left(y_1-\frac{1}{2}r\right)$$
$$-4\left(z_1+\frac{1}{2}r\right)-11=0$$

Rewriting this as a quadratic in r

$$\frac{3}{2}r^2+(2x_1-y_1+z_1-1)r$$
$$+(x_1^2+y_1^2+z_1^2-2y_1-4z_1-11)=0 \qquad (4)$$

The generator touches the sphere (3 if (4) has equal roots i.e., discriminant is zero or

$$(2x_1-y_1+z_1-1)^2$$
$$=4\cdot\frac{3}{2}\cdot(x_1^2+y_1^2+z_1^2-2y_1-4z_1-11).$$

The required equation of the cylinder is

$$2x^2+5y^2+5z^2+4xy-4xz+2yz$$
$$+4x-14y-22z-67=0.$$

EXERCISE

1. Find the equation of the quadratic cylinder whose generators intersect the curve
   a. $ax^2+by^2=2z, lx+my+nz=p$ and are parallel to z-axis.
   b. $ax^2+by^2+cz^2=1, lx+my+nz=p$ and are parallel to x-axis.
   Hint: Eliminate z

Ans. **a.** $n(ax^2+by^2)+2lx+2my-2p=0$

Hint: Eliminate x.

Ans. **b.** $(bl^2+am^2)y^2+(cl^2+an^2)z^2+2amnyz$ $-2ampy-2anpz+(ap^2-l^2)=0$

2. If l, m, n are the DC's of the generating line and the circle $x^2+z^2=a^2$ in the zx-plane is the guiding curve, find the equation of the sphere.

Ans. $(mx-ly)^2+(mz-ny)^2=a^2m^2$

Find the equation of a right circular cylinder (4 to 9)

4. Whose axis is the line $\frac{x-1}{2}=\frac{y+3}{-1}=\frac{z-2}{5}$ and radius is 2 units.

Ans. $26x^2+29y^2+5z^2+4xy+10yz-20zx+150y+30z+75=0$

5. Having for its base the circle $x^2+y^2+z^2=9, x-y+z=3$.

Ans. $x^2+y^2+z^2+xy+yz-zx-9=0$

6. Whose axis passes through the point (1, 2, 3) and has DC's proportional to (2, −3, 6) and of radius 2.

Ans. $45x^2+40y^2+13z^2+36yz-24zx+12xy-42x-280y-126z+294=0.$

7. Whose axis is the line $\frac{x-1}{2}=\frac{y-2}{1}=\frac{z-3}{2}$ and radius 2 units.

Ans. $5x^2+8y^2+5z^2-4yz-8zx-4xy+22x-16y-14z-10=0$

8. The guiding curve is the circle through the three points (1, 0, 0), (0, 1, 0)(0, 0, 1).

Ans. $x^2+y^2+z^2-xy-yz-zx=1$

9. The directing curve is $x^2+z^2-4x-2z+4=0$, $y=0$ and whose axis contains the point (0, 3, 0). Also find the area of the section of the cylinder by a plane parallel to xz-plane.

Hint: Centre of circle (2, 0, 1) radius: 1

Ans. $9x^2+5y^2+9z^2+12xy+6yz-36x-30y-18z+36=0, \pi$

10. Find the equation of the enveloping cylinder of the sphere $x^2+y^2+z^2-2x+4y=1$, having its generators parallel to the line $x=y=z$.

Ans. $x^2+y^2+z^2-xy-yz-zx-2x+7y+z-2=0.$

")*+,)))-) C !DC$! $(6"
' % B . / " 8 /

Chapter 12

Introduction to trigonometry

***Why it is important to understand:* Introduction to trigonometry**

There are an enormous number of uses of trigonometry and trigonometric functions. Fields that use trigonometry or trigonometric functions include astronomy (especially for locating apparent positions of celestial objects, in which spherical trigonometry is essential) and hence navigation (on the oceans, in aircraft, and in space), music theory, acoustics, optics, analysis of financial markets, electronics, probability theory, statistics, biology, medical imaging (CAT scans and ultrasound), pharmacy, chemistry, number theory (and hence cryptology), seismology, meteorology, oceanography, many physical sciences, land surveying and geodesy (a branch of earth sciences), architecture, phonetics, economics, electrical engineering, mechanical engineering, civil engineering, computer graphics, cartography, crystallography and game development. It is clear that a good knowledge of trigonometry is essential in many fields of engineering.

At the end of this chapter, you should be able to:

- state the theorem of Pythagoras and use it to find the unknown side of a right-angled triangle
- define sine, cosine, tangent, secant, cosecant and cotangent of an angle in a right-angled triangle
- evaluate trigonometric ratios of angles
- solve right-angled triangles
- understand angles of elevation and depression
- use the sine and cosine rules to solve non-right-angled triangles
- calculate the area of any triangle
- solve practical problems involving trigonometry

Higher Engineering Mathematics. 978-1-138-67357-1, © 2017 John Bird. Published by Taylor & Francis. All rights reserved.

12.1 Trigonometry

Trigonometry is the branch of mathematics which deals with the measurement of sides and angles of triangles, and their relationship with each other. There are many applications in engineering where knowledge of trigonometry is needed.

12.2 The theorem of Pythagoras

With reference to Fig. 12.1, the side opposite the right angle (i.e. side b) is called the **hypotenuse**. The **theorem of Pythagoras*** states:

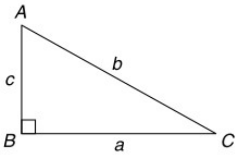

Figure 12.1

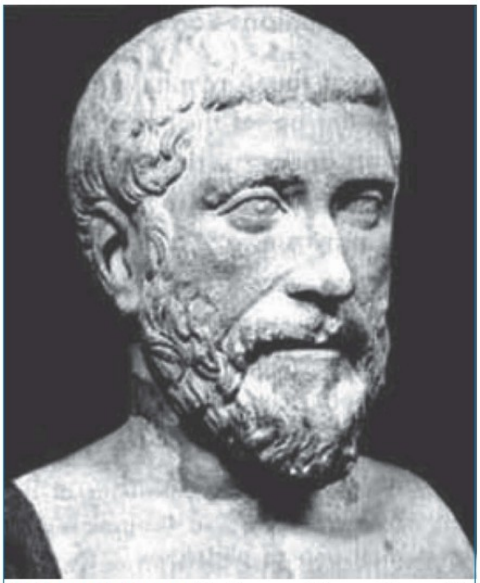

[*] **Who was Pythagoras**? **Pythagoras of Samos** (Born c. 570 BC and died about 495 BC) was an Ionian Greek philosopher and mathematician. He is best known for the Pythagorean theorem, which states that in a right-angled triangle $a^2+b^2=c^2$. To find out more go to **www.routledge.com/cw/bird**

'In any right-angled triangle, the square on the hypotenuse is equal to the sum of the squares on the other two sides.'

Hence $\boldsymbol{b^2 = a^2 + c^2}$

Problem 1. In Fig. 12.2, find the length of EF.

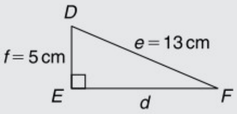

Figure 12.2

By Pythagoras' theorem:

$$e^2 = d^2 + f^2$$

Hence $\quad 13^2 = d^2 + 5^2$

$$169 = d^2 + 25$$

$$d^2 = 169 - 25 = 144$$

Thus $\quad d = \sqrt{144} = 12\,\text{cm}$

i.e. $\quad \boldsymbol{EF = 12\,\text{cm}}$

Problem 2. Two aircraft leave an airfield at the same time. One travels due north at an average speed of 300 km/h and the other due west at an average speed of 220 km/h. Calculate their distance apart after four hours.

After four hours, the first aircraft has travelled $4 \times 300 = 1200$ km, due north, and the second aircraft has travelled $4 \times 220 = 880$ km due west, as shown in Fig. 12.3. Distance apart after four hours = BC.

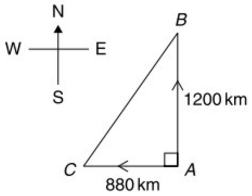

Figure 12.3

From Pythagoras' theorem:

$$BC^2 = 1200^2 + 880^2 = 1\,440\,000 + 774\,400$$

and $BC = \sqrt{(2\,214\,400)}$

Hence distance apart after four hours = 1488 km

Now try the following Practice Exercise

Practice Exercise 51 The theorem of Pythagoras (Answers on page 863)

1. In a triangle *CDE*, $D = 90°$, $CD = 14.83$ mm and $CE = 28.31$ mm. Determine the length of *DE*.
2. Triangle *PQR* is isosceles, *Q* being a right angle. If the hypotenuse is 38.47 cm find (a) the lengths of sides *PQ* and *QR*, and (b) the value of $\angle QPR$.
3. A man cycles 24 km due south and then 20 km due east. Another man, starting at the same time as the first man, cycles 32 km due east and then 7 km due south. Find the distance between the two men.
4. A ladder 3.5 m long is placed against a perpendicular wall with its foot 1.0 m from the wall. How far up the wall (to the nearest centimetre) does the ladder reach? If the foot of the ladder is now moved 30 cm further away from the wall, how far does the top of the ladder fall?
5. Two ships leave a port at the same time. One travels due west at 18.4 km/h and the other due south at 27.6 km/h. Calculate how far apart the two ships are after four hours.
6. Figure 12.4 shows a bolt rounded off at one end. Determine the dimension *h*.

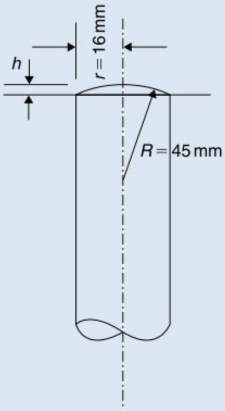

Figure 12.4

7. Figure 12.5 shows a cross-section of a component that is to be made from a round bar. If the diameter of the bar is 74 mm, calculate the dimension *x*.

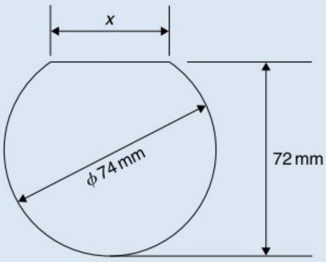

Figure 12.5

12.3 Trigonometric ratios of acute angles

(a) With reference to the right-angled triangle shown in Fig. 12.6:

(i) $$\text{sine}\,\theta = \frac{\text{opposite side}}{\text{hypotenuse}}$$

i.e. $$\boldsymbol{\sin\theta = \frac{b}{c}}$$

(ii) $$\text{cosine}\,\theta = \frac{\text{adjacent side}}{\text{hypotenuse}}$$

i.e. $$\boldsymbol{\cos\theta = \frac{a}{c}}$$

(iii) $$\text{tangent}\,\theta = \frac{\text{opposite side}}{\text{adjacent side}}$$

i.e. $$\boldsymbol{\tan\theta = \frac{b}{a}}$$

(iv) $$\text{secant}\,\theta = \frac{\text{hypotenuse}}{\text{adjacent side}}$$

i.e. $$\boldsymbol{\sec\theta = \frac{c}{a}}$$

(v) $$\text{cosecant}\,\theta = \frac{\text{hypotenuse}}{\text{opposite side}}$$

i.e. $$\boldsymbol{\text{cosec}\,\theta = \frac{c}{b}}$$

(vi) $\text{cotangent}\,\theta = \dfrac{\text{adjacent side}}{\text{opposite side}}$

i.e. $\boldsymbol{\cot\theta = \dfrac{a}{b}}$

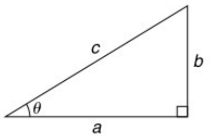

Figure 12.6

(b) From above,

(i) $$\frac{\sin\theta}{\cos\theta} = \frac{\frac{b}{c}}{\frac{a}{c}} = \frac{b}{a} = \tan\theta,$$

i.e. $$\boldsymbol{\tan\theta = \frac{\sin\theta}{\cos\theta}}$$

(ii) $$\frac{\cos\theta}{\sin\theta} = \frac{\frac{a}{c}}{\frac{b}{c}} = \frac{a}{b} = \cot\theta,$$

i.e. $$\boldsymbol{\cot\theta = \frac{\cos\theta}{\sin\theta}}$$

(iii) $$\boldsymbol{\sec\theta = \frac{1}{\cos\theta}}$$

(iv) $$\boldsymbol{\operatorname{cosec}\theta = \frac{1}{\sin\theta}}$$

(Note, the first letter of each ratio, '*s*' and '*c*' go together)

(v) $$\boldsymbol{\cot\theta = \frac{1}{\tan\theta}}$$

Secants, cosecants and cotangents are called the **reciprocal ratios**.

Problem 3. If $\cos X = \dfrac{9}{41}$ determine the value of the other five trigonometry ratios.

Fig. 12.7 shows a right-angled triangle *XYZ*.

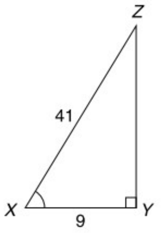

Figure 12.7

Since $\cos X = \dfrac{9}{41}$, then $XY = 9$ units and $XZ = 41$ units.

Using Pythagoras' theorem: $41^2 = 9^2 + YZ^2$ from which $YZ = \sqrt{(41^2 - 9^2)} = 40$ units.

Thus

$$\boldsymbol{\sin X = \frac{40}{41}},\ \boldsymbol{\tan X} = \frac{40}{9} \boldsymbol{= 4\frac{4}{9}},$$

$$\boldsymbol{\operatorname{cosec} X = \frac{41}{40} = 1\frac{1}{40}},$$

$$\boldsymbol{\sec X = \frac{41}{9} = 4\frac{5}{9}} \text{ and } \boldsymbol{\cot X = \frac{9}{40}}$$

Problem 4. If $\sin\theta = 0.625$ and $\cos\theta = 0.500$ determine, without using trigonometric tables or calculators, the values of $\operatorname{cosec}\theta$, $\sec\theta$, $\tan\theta$ and $\cot\theta$.

$$\operatorname{cosec}\theta = \frac{1}{\sin\theta} = \frac{1}{0.625} = \mathbf{1.60}$$

$$\sec\theta = \frac{1}{\cos\theta} = \frac{1}{0.500} = \mathbf{2.00}$$

$$\tan\theta = \frac{\sin\theta}{\cos\theta} = \frac{0.625}{0.500} = \mathbf{1.25}$$

$$\cot\theta = \frac{\cos\theta}{\sin\theta} = \frac{0.500}{0.625} = \mathbf{0.80}$$

Problem 5. Point *A* lies at co-ordinate (2, 3) and point *B* at (8, 7). Determine (a) the distance *AB*, (b) the gradient of the straight line *AB*, and (c) the angle *AB* makes with the horizontal.

(a) Points *A* and *B* are shown in Fig. 12.8(a).

In Fig. 12.8(b), the horizontal and vertical lines *AC* and *BC* are constructed.

Since *ABC* is a right-angled triangle, and $AC = (8-2) = 6$ and $BC = (7-3) = 4$, then by Pythagoras' theorem

$$AB^2 = AC^2 + BC^2 = 6^2 + 4^2$$

and $\boldsymbol{AB} = \sqrt{(6^2 + 4^2)} = \sqrt{52} = \mathbf{7.211,}$ correct to 3 decimal places.

(b) The gradient of *AB* is given by tan *A*,

i.e. $\textbf{gradient} = \tan A = \dfrac{BC}{AC} = \dfrac{4}{6} = \mathbf{\dfrac{2}{3}}$

(c) **The angle *AB* makes with the horizontal** is given by $\tan^{-1}\frac{2}{3} = \mathbf{33.69°}$

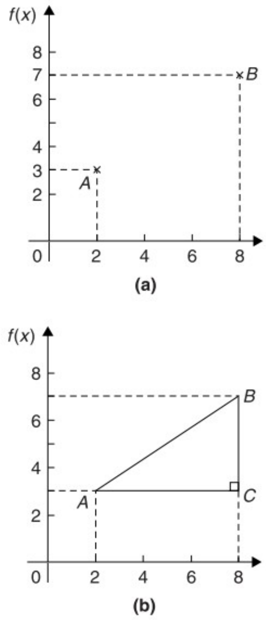

Figure 12.8

Now try the following Practice Exercise

Practice Exercise 52 Trigonometric ratios of acute angles (Answers on page 863)

1. In triangle ABC shown in Fig. 12.9, find $\sin A, \cos A, \tan A, \sin B, \cos B$ and $\tan B$.

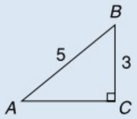

Figure 12.9

2. If $\cos A = \dfrac{15}{17}$ find $\sin A$ and $\tan A$, in fraction form.

3. For the right-angled triangle shown in Fig. 12.10, find:
   (a) $\sin\alpha$ (b) $\cos\theta$ (c) $\tan\theta$

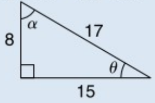

Figure 12.10

4. Point P lies at co-ordinate $(-3, 1)$ and point Q at $(5, -4)$. Determine
   (a) the distance PQ
   (b) the gradient of the straight line PQ and
   (c) the angle PQ makes with the horizontal.

12.4 Evaluating trigonometric ratios

The easiest method of evaluating trigonometric functions of any angle is by using a **calculator**.

The following values, correct to 4 decimal places, may be checked:

sine $18^\circ = 0.3090$, cosine $56^\circ = 0.5592$
sine $172^\circ = 0.1392$ cosine $115^\circ = -0.4226$,
sine $241.63^\circ = -0.8799$, cosine $331.78^\circ = 0.8811$

tangent $29^\circ = 0.5543$,
tangent $178^\circ = -0.0349$
tangent $296.42^\circ = -2.0127$

To evaluate, say, sine $42^\circ 23'$ using a calculator means finding sine $42\dfrac{23^\circ}{60}$ since there are 60 minutes in 1 degree.

$$\frac{23}{60} = 0.383\dot{3} \text{ thus } 42^\circ 23' = 42.38\dot{3}^\circ$$

Thus sine $42^\circ 23' =$ sine $42.38\dot{3}^\circ = 0.6741$, correct to 4 decimal places.

Similarly, cosine $72^\circ 38' =$ cosine $72\dfrac{38^\circ}{60} = 0.2985$, correct to 4 decimal places.

Most calculators contain only sine, cosine and tangent functions. Thus to evaluate secants, cosecants and cotangents, reciprocals need to be used. The following values, correct to 4 decimal places, may be checked:

$$\text{secant}\,32^\circ = \frac{1}{\cos 32^\circ} = 1.1792$$

$$\text{cosecant}\,75^\circ = \frac{1}{\sin 75^\circ} = 1.0353$$

$$\text{cotangent}\,41^\circ = \frac{1}{\tan 41^\circ} = 1.1504$$

$$\text{secant}\,215.12^\circ = \frac{1}{\cos 215.12^\circ} = -1.2226$$

$$\text{cosecant}\,321.62^\circ = \frac{1}{\sin 321.62^\circ} = -1.6106$$

$$\text{cotangent}\,263.59^\circ = \frac{1}{\tan 263.59^\circ} = 0.1123$$

If we know the value of a trigonometric ratio and need to find the angle we use the **inverse function** on our calculators.

For example, using shift and sin on our calculator gives $\sin^{-1}($

If, for example, we know the sine of an angle is 0.5 then the value of the angle is given by:

$$\sin^{-1} 0.5 = \mathbf{30^\circ} \quad (\text{Check that } \sin 30^\circ = 0.5)$$

(Note that $\sin^{-1} x$ does not mean $\dfrac{1}{\sin x}$; also, $\sin^{-1} x$ may also be written as $\arcsin x$)

Similarly, if $\cos\theta = 0.4371$ then

$\theta = \cos^{-1} 0.4371 = \mathbf{64.08^\circ}$

$$\text{and} \quad \text{if } \tan A = 3.5984 \text{ then } A = \tan^{-1} 3.5984 = \mathbf{74.47^\circ}$$

each correct to 2 decimal places.

Use your calculator to check the following worked examples.

Problem 6. Determine, correct to 4 decimal places, $\sin 43^\circ 39'$

$$\sin 43^\circ 39' = \sin 43\frac{39}{60}^\circ = \sin 43.65^\circ = \mathbf{0.6903}$$

This answer can be obtained using the **calculator** as follows:

1. Press sin
2. Enter 43
3. Press ° ’’’
4. Enter 39
5. Press ° ’’’
6. Press)
7. Press = Answer = **0.6902512**....

Problem 7. Determine, correct to 3 decimal places, $6\cos 62^\circ 12'$

$$6\cos 62^\circ 12' = 6\cos 62\frac{12^\circ}{60} = 6\cos 62.20^\circ = \mathbf{2.798}$$

This answer can be obtained using the **calculator** as follows:

1. Enter 6
2. Press cos
3. Enter 62
4. Press ° ’’’
5. Enter 12
6. Press ° ’’’
7. Press)
8. Press = Answer = **2.798319**....

Problem 8. Evaluate, correct to 4 decimal places: (a) $\text{sine}\,168^\circ 14'$ (b) $\text{cosine}\,271.41^\circ$ (c) $\text{tangent}\,98^\circ 4'$

(a) $\text{sine}\,168^\circ 14' = \text{sine}\,168\dfrac{14^\circ}{60} = \mathbf{0.2039}$

(b) $\text{cosine}\,271.41^\circ = \mathbf{0.0246}$

(c) $\text{tangent}\,98^\circ 4' = \tan 98\dfrac{4^\circ}{60} = \mathbf{-7.0558}$

Problem 9. Evaluate, correct to 4 decimal places: (a) $\text{secant}\,161^\circ$ (b) $\text{secant}\,302^\circ 29'$

(a) $\sec 161^\circ = \dfrac{1}{\cos 161^\circ} = \mathbf{-1.0576}$

(b) $\sec 302^\circ 29' = \dfrac{1}{\cos 302^\circ 29'} = \dfrac{1}{\cos 302\dfrac{29^\circ}{60}} = \mathbf{1.8620}$

Problem 10. Evaluate, correct to 4 significant figures: (a) $\text{cosecant}\,279.16^\circ$ (b) $\text{cosecant}\,49^\circ 7'$

(a) $\text{cosec}\,279.16^\circ = \dfrac{1}{\sin 279.16^\circ} = \mathbf{-1.013}$

(b) $\text{cosec}\,49^\circ 7' = \dfrac{1}{\sin 49^\circ 7'} = \dfrac{1}{\sin 49\dfrac{7^\circ}{60}} = \mathbf{1.323}$

Problem 11. Evaluate, correct to 4 decimal places: (a) $\text{cotangent}\,17.49^\circ$ (b) $\text{cotangent}\,163^\circ 52'$

(a) $\cot 17.49^\circ = \dfrac{1}{\tan 17.49^\circ} = \mathbf{3.1735}$

(b) $\cot 163^\circ 52' = \dfrac{1}{\tan 163^\circ 52'} = \dfrac{1}{\tan 163\dfrac{52^\circ}{60}} = \mathbf{-3.4570}$

Problem 12. Evaluate, correct to 4 significant figures: (a) $\sin 1.481$ (b) $\cos(3\pi/5)$ (c) $\tan 2.93$

(a) $\sin 1.481$ means the sine of 1.481 radians. Hence a calculator needs to be on the radian function. Hence $\sin 1.481 = \mathbf{0.9960}$

(b) $\cos(3\pi/5) = \cos 1.884955\ldots = \mathbf{-0.3090}$

(c) $\tan 2.93 = \mathbf{-0.2148}$

Problem 13. Evaluate, correct to 4 decimal places:
(a) secant 5.37 (b) cosecant $\pi/4$
(c) cotangent $\pi/24$

(a) Again, with no degrees sign, it is assumed that 5.37 means 5.37 radians.

Hence $\sec 5.37 = \dfrac{1}{\cos 5.37} = \mathbf{1.6361}$

(b) $$\operatorname{cosec}(\pi/4) = \frac{1}{\sin(\pi/4)} = \frac{1}{\sin 0.785398\ldots} = \mathbf{1.4142}$$

(c) $$\cot(5\pi/24) = \frac{1}{\tan(5\pi/24)} = \frac{1}{\tan 0.654498\ldots} = \mathbf{1.3032}$$

Problem 14. Find, in degrees, the acute angle $\sin^{-1} 0.4128$ correct to 2 decimal places.

$\sin^{-1} 0.4128$ means ‘the angle whose sine is 0.4128’

Using a calculator:

1. Press shift 2. Press sin 3. Enter 0.4128

4. Press) 5. Press = The answer 24.380848… is displayed

Hence, $\mathbf{\sin^{-1} 0.4128 = 24.38^\circ}$

Problem 15. Find the acute angle $\cos^{-1} 0.2437$ in degrees and minutes.

$\cos^{-1} 0.2437$ means ‘the angle whose cosine is 0.2437’

Using a calculator:

1. Press shift 2. Press cos 3. Enter 0.2437

4. Press) 5. Press = The answer 75.894979… is displayed

6. Press ° ’’’ and $75^\circ 53' 41.93''$ is displayed

Hence, $\mathbf{\cos^{-1} 0.2437 = 75.89^\circ = 77^\circ 54'}$ correct to the nearest minute.

Problem 16. Find the acute angle $\tan^{-1} 7.4523$ in degrees and minutes.

$\tan^{-1} 7.4523$ means ‘the angle whose tangent is 7.4523’

Using a calculator:

1. Press shift 2. Press tan 3. Enter 7.4523

4. Press) 5. Press = The answer 82.357318… is displayed

6. Press ° ’’’ and $82^\circ 21' 26.35''$ is displayed

Hence, $\mathbf{\tan^{-1} 7.4523 = 82.36^\circ = 82^\circ 21'}$ correct to the nearest minute.

Problem 17. Determine the acute angles:
(a) $\sec^{-1} 2.3164$ (b) $\operatorname{cosec}^{-1} 1.1784$
(c) $\cot^{-1} 2.1273$

(a) $$\sec^{-1} 2.3164 = \cos^{-1}\left(\frac{1}{2.3164}\right) = \cos^{-1} 0.4317\ldots = \mathbf{64.42^\circ} \text{ or } \mathbf{64^\circ 25'}$$

or **1.124 radians**

(b) $$\operatorname{cosec}^{-1} 1.1784 = \sin^{-1}\left(\frac{1}{1.1784}\right) = \sin^{-1} 0.8486\ldots = \mathbf{58.06^\circ} \text{ or } \mathbf{58^\circ 4'}$$

or **1.013 radians**

(c) $$\cot^{-1} 2.1273 = \tan^{-1}\left(\frac{1}{2.1273}\right) = \tan^{-1} 0.4700\ldots = \mathbf{25.18^\circ} \text{ or } \mathbf{25^\circ 11'}$$

or **0.439 radians**

Problem 18. Evaluate the following expression, correct to 4 significant figures:

$$\frac{4 \sec 32^\circ 10' - 2 \cot 15^\circ 19'}{3 \operatorname{cosec} 63^\circ 8' \tan 14^\circ 57'}$$

By calculator:

$$\sec 32^\circ 10' = 1.1813, \cot 15^\circ 19' = 3.6512$$

$$\operatorname{cosec} 63^\circ 8' = 1.1210, \tan 14^\circ 57' = 0.2670$$

$$\text{Hence}\quad \frac{4\sec 32^\circ 10' - 2\cot 15^\circ 19'}{3\operatorname{cosec} 63^\circ 8' \tan 14^\circ 57'}$$

$$= \frac{4(1.1813) - 2(3.6512)}{3(1.1210)(0.2670)}$$

$$= \frac{4.7252 - 7.3024}{0.8979}$$

$$= \frac{-2.5772}{0.8979} = \mathbf{-2.870}$$

correct to 4 significant figures.

Problem 19. Evaluate correct to 4 decimal places: (a) sec(–115°) (b) cosec (–95°47′)

(a) Positive angles are considered by convention to be anticlockwise and negative angles as clockwise.

Hence –115° is actually the same as 245° (i.e. 360° – 115°)

$$\text{Hence}\quad \sec(-115^\circ) = \sec 245^\circ = \frac{1}{\cos 245^\circ}$$

$$= \mathbf{-2.3662}$$

(b) $$\operatorname{cosec}(-95^\circ 47') = \frac{1}{\sin\left(-95\dfrac{47^\circ}{60}\right)} = \mathbf{-1.0051}$$

Problem 20. In triangle EFG in Fig. 12.11, calculate angle G.

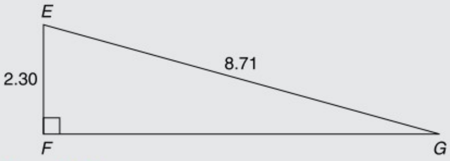

Figure 12.11

With reference to $\angle G$, the two sides of the triangle given are the opposite side EF and the hypotenuse EG; hence, sine is used,

i.e. $$\sin G = \frac{2.30}{8.71} = 0.26406429\ldots$$

from which, $$G = \sin^{-1} 0.26406429\ldots$$

i.e. $$G = 15.311360\ldots$$

Hence, $$\angle \mathbf{G} = \mathbf{15.31^\circ} \text{ or } \mathbf{15^\circ 19'}$$

Now try the following Practice Exercise

Practice Exercise 53 Evaluating trigonometric ratios (Answers on page 863)

In Problems 1 to 8, evaluate correct to 4 decimal places:

1. (a) sine 27° (b) sine 172.41° (c) sine 302°52′
2. (a) cosine 124° (b) cosine 21.46° (c) cosine 284°10′
3. (a) tangent 145° (b) tangent 310.59° (c) tangent 49°16′
4. (a) secant 73° (b) secant 286.45° (c) secant 155°41′
5. (a) cosecant 213° (b) cosecant 15.62° (c) cosecant 311°50′
6. (a) cotangent 71° (b) cotangent 151.62° (c) cotangent 321°23′
7. (a) sine $\frac{2\pi}{3}$ (b) cos 1.681 (c) tan 3.672
8. (a) sec $\frac{\pi}{8}$ (b) cosec 2.961 (c) cot 2.612

In Problems 9 to 14, determine the acute angle in degrees (correct to 2 decimal places), degrees and minutes, and in radians (correct to 3 decimal places).

9. $\sin^{-1} 0.2341$
10. $\cos^{-1} 0.8271$
11. $\tan^{-1} 0.8106$
12. $\sec^{-1} 1.6214$

13. $\text{cosec}^{-1} 2.4891$

14. $\cot^{-1} 1.9614$

15. In the triangle shown in Fig. 12.12, determine angle θ, correct to 2 decimal places.

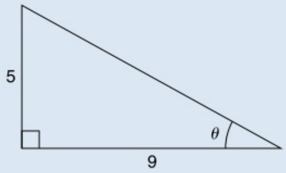

Figure 12.12

16. In the triangle shown in Fig. 12.13, determine angle θ in degrees and minutes.

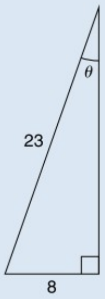

Figure 12.13

In Problems 17 to 20, evaluate correct to 4 significant figures:

17. $4\cos 56°19' - 3\sin 21°57'$

18. $\dfrac{11.5\tan 49°11' - \sin 90°}{3\cos 45°}$

19. $\dfrac{5\sin 86°3'}{3\tan 14°29' - 2\cos 31°9'}$

20. $\dfrac{6.4\,\text{cosec}\, 29°5' - \sec 81°}{2\cot 12°}$

21. Determine the acute angle, in degrees and minutes, correct to the nearest minute, given by $\sin^{-1}\left(\dfrac{4.32\sin 42°16'}{7.86}\right)$

22. If $\tan x = 1.5276$, determine $\sec x$, $\text{cosec}\, x$, and $\cot x$. (Assume x is an acute angle.)

In Problems 23 to 25 evaluate correct to 4 significant figures:

23. $\dfrac{(\sin 34°27')(\cos 69°2')}{(2\tan 53°39')}$

24. $3\cot 14°15' \sec 23°9'$

25. $\dfrac{\text{cosec}\, 27°19' + \sec 45°29'}{1 - \text{cosec}\, 27°19' \sec 45°29'}$

26. Evaluate correct to 4 decimal places:
    (a) sine $(-125°)$ (b) $\tan(-241°)$
    (c) $\cos(-49°15')$

27. Evaluate correct to 5 significant figures:
    (a) cosec $(-143°)$ (b) $\cot(-252°)$
    (c) $\sec(-67°22')$

12.5 Solution of right-angled triangles

To 'solve a right-angled triangle' means 'to find the unknown sides and angles'. This is achieved by using (i) the theorem of Pythagoras, and/or (ii) trigonometric ratios. This is demonstrated in the following problems.

Problem 21. In triangle PQR shown in Fig. 12.14, find the lengths of PQ and PR.

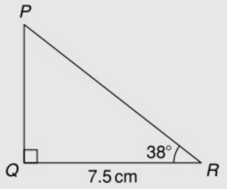

Figure 12.14

$$\tan 38° = \frac{PQ}{QR} = \frac{PQ}{7.5}$$

hence $PQ = 7.5\tan 38° = 7.5(0.7813)$

$$= \mathbf{5.860\,cm}$$

$$\cos 38° = \frac{QR}{PR} = \frac{7.5}{PR}$$

hence $PR = \dfrac{7.5}{\cos 38°} = \dfrac{7.5}{0.7880} = \mathbf{9.518\,cm}$

[Check: Using Pythagoras' theorem

$$(7.5)^2+(5.860)^2=90.59=(9.518)^2]$$

Problem 22. Solve the triangle *ABC* shown in Fig. 12.15.

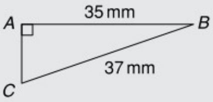

Figure 12.15

To 'solve triangle *ABC*' means 'to find the length *AC* and angles *B* and *C*'

$$\sin C=\frac{35}{37}=0.94595$$

hence $\angle \boldsymbol{C}=\sin^{-1}0.94595=71.08°=\mathbf{71°5'}$
$\angle \boldsymbol{B}=180°-90°-71°5'=\mathbf{18°55'}$ (since angles in a triangle add up to 180°)

$$\sin B=\frac{AC}{37}$$

hence $\boldsymbol{AC}=37\sin 18°55'=37(0.3242)$

$$=\mathbf{12.0\,mm}$$

or, using Pythagoras' theorem, $37^2=35^2+AC^2$, from which, $\boldsymbol{AC}=\sqrt{(37^2-35^2)}=\mathbf{12.0\,mm}$.

Problem 23. Solve triangle *XYZ* given $\angle X=90°$, $\angle Y=23°17'$ and $YZ=20.0$ mm. Determine also its area.

It is always advisable to make a reasonably accurate sketch so as to visualise the expected magnitudes of unknown sides and angles. Such a sketch is shown in Fig. 12.16.

$$\angle Z=180°-90°-23°17'=\mathbf{66°43'}$$

$$\sin 23°17'=\frac{XZ}{20.0}$$

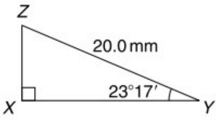

Figure 12.16

hence $\boldsymbol{XZ}=20.0\sin 23°17'$

$$=20.0(0.3953)=\mathbf{7.906\,mm}$$

$$\cos 23°17'=\frac{XY}{20.0}$$

hence $\boldsymbol{XY}=20.0\cos 23°17'$

$$=20.0(0.9186)=\mathbf{18.37\,mm}$$

[Check: Using Pythagoras' theorem

$$(18.37)^2+(7.906)^2=400.0=(20.0)^2]$$

Area of triangle *XYZ*

$$=\tfrac{1}{2}\text{ (base) (perpendicular height)}$$

$$=\tfrac{1}{2}(XY)(XZ)=\tfrac{1}{2}(18.37)(7.906)$$

$$=\mathbf{72.62\,mm^2}$$

Now try the following Practice Exercise

Practice Exercise 54 The solution of right-angled triangles (Answers on page 863)

1. Solve triangle *ABC* in Fig. 12.17(i).

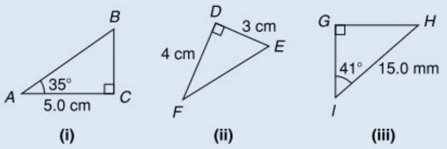

Figure 12.17

2. Solve triangle *DEF* in Fig. 12.17(ii).
3. Solve triangle *GHI* in Fig. 12.17(iii).
4. Solve the triangle *JKL* in Fig. 12.18(i) and find its area.
5. Solve the triangle *MNO* in Fig. 12.18(ii) and find its area.

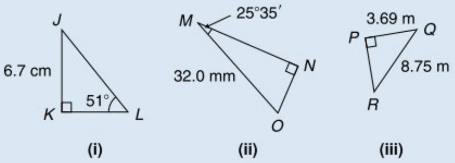

Figure 12.18

6. Solve the triangle *PQR* in Fig. 12.18(iii) and find its area.

7. A ladder rests against the top of the perpendicular wall of a building and makes an angle of 73° with the ground. If the foot of the ladder is 2 m from the wall, calculate the height of the building.

12.6 Angles of elevation and depression

(a) If, in Fig. 12.19, BC represents horizontal ground and AB a vertical flagpole, then the **angle of elevation** of the top of the flagpole, A, from the point C is the angle that the imaginary straight line AC must be raised (or elevated) from the horizontal CB, i.e. angle θ.

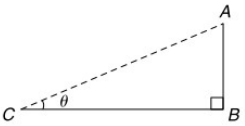

Figure 12.19

(b) If, in Fig. 12.20, PQ represents a vertical cliff and R a ship at sea, then the **angle of depression** of the ship from point P is the angle through which the imaginary straight line PR must be lowered (or depressed) from the horizontal to the ship, i.e. angle ϕ.

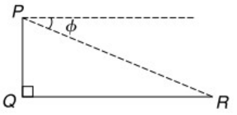

Figure 12.20

(Note, $\angle PRQ$ is also ϕ – alternate angles between parallel lines.)

Problem 24. An electricity pylon stands on horizontal ground. At a point 80 m from the base of the pylon, the angle of elevation of the top of the pylon is 23°. Calculate the height of the pylon to the nearest metre.

Figure 12.21 shows the pylon AB and the angle of elevation of A from point C is 23°

$$\tan 23° = \frac{AB}{BC} = \frac{AB}{80}$$

Hence height of pylon AB

$$= 80 \tan 23° = 80(0.4245) = 33.96\,\text{m}$$

$$= \mathbf{34\,m\ to\ the\ nearest\ metre}.$$

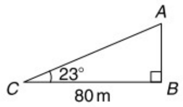

Figure 12.21

Problem 25. A surveyor measures the angle of elevation of the top of a perpendicular building as 19°. He moves 120 m nearer the building and finds the angle of elevation is now 47°. Determine the height of the building.

The building PQ and the angles of elevation are shown in Fig. 12.22.

In triangle PQS,

$$\tan 19° = \frac{h}{x+120}$$

hence $h = \tan 19°(x+120)$,

i.e. $h = 0.3443(x+120)$ (1)

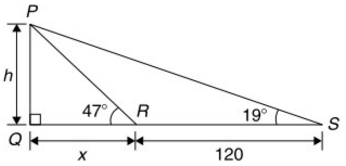

Figure 12.22

In triangle PQR, $\tan 47° = \dfrac{h}{x}$

hence $h = \tan 47°(x)$, i.e. $h = 1.0724x$ (2)

Equating equations (1) and (2) gives:

$$0.3443(x+120) = 1.0724x$$
$$0.3443x + (0.3443)(120) = 1.0724x$$
$$(0.3443)(120) = (1.0724 - 0.3443)x$$
$$41.316 = 0.7281x$$
$$x = \frac{41.316}{0.7281} = 56.74\,\text{m}$$

From equation (2), **height of building**,

$\boldsymbol{h} = 1.0724x = 1.0724(56.74) = \mathbf{60.85\,m}$

Problem 26. The angle of depression of a ship viewed at a particular instant from the top of a 75 m vertical cliff is 30°. Find the distance of the ship from the base of the cliff at this instant. The ship is sailing away from the cliff at constant speed and one minute later its angle of depression from the top of the cliff is 20°. Determine the speed of the ship in km/h.

Fig. 12.23 shows the cliff AB, the initial position of the ship at C and the final position at D. Since the angle of depression is initially 30° then $\angle ACB = 30°$ (alternate angles between parallel lines).

$$\tan 30° = \frac{AB}{BC} = \frac{75}{BC}$$

hence $$BC = \frac{75}{\tan 30°} = \frac{75}{0.5774} = \mathbf{129.9\,m}$$

$= $ **initial position of ship from base of cliff**

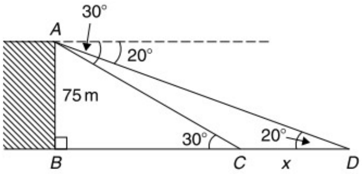

Figure 12.23

In triangle ABD,

$$\tan 20° = \frac{AB}{BD} = \frac{75}{BC+CD} = \frac{75}{129.9+x}$$

Hence $$129.9 + x = \frac{75}{\tan 20°} = \frac{75}{0.3640} = 206.0\,\text{m}$$

from which $$x = 206.0 - 129.9 = 76.1\,\text{m}$$

Thus the ship sails 76.1 m in one minute, i.e. 60 s, hence speed of ship

$$= \frac{\text{distance}}{\text{time}} = \frac{76.1}{60}\,\text{m/s}$$

$$= \frac{76.1 \times 60 \times 60}{60 \times 1000}\,\text{km/h} = \mathbf{4.57\,km/h}$$

Now try the following Practice Exercise

Practice Exercise 55 Angles of elevation and depression (Answers on page 864)

1. If the angle of elevation of the top of a vertical 30 m high aerial is 32°, how far is it to the aerial?
2. From the top of a vertical cliff 80.0 m high the angles of depression of two buoys lying due west of the cliff are 23° and 15°, respectively. How far are the buoys apart?
3. From a point on horizontal ground a surveyor measures the angle of elevation of the top of a flagpole as $18°40'$. He moves 50 m nearer to the flagpole and measures the angle of elevation as $26°22'$. Determine the height of the flagpole.
4. A flagpole stands on the edge of the top of a building. At a point 200 m from the building the angles of elevation of the top and bottom of the pole are 32° and 30° respectively. Calculate the height of the flagpole.
5. From a ship at sea, the angles of elevation of the top and bottom of a vertical lighthouse standing on the edge of a vertical cliff are 31° and 26°, respectively. If the lighthouse is 25.0 m high, calculate the height of the cliff.
6. From a window 4.2 m above horizontal ground the angle of depression of the foot of a building across the road is 24° and the angle of elevation of the top of the building is 34°. Determine, correct to the nearest centimetre, the width of the road and the height of the building.
7. The elevation of a tower from two points, one due east of the tower and the other due west of it are 20° and 24°, respectively, and the two points of observation are 300 m apart. Find the height of the tower to the nearest metre.

12.7 Sine and cosine rules

To **'solve a triangle'** means 'to find the values of unknown sides and angles'. If a triangle is **right angled**, trigonometric ratios and the theorem of Pythagoras may be used for its solution, as shown in Section 12.5.

However, for a **non-right-angled triangle**, trigonometric ratios and Pythagoras' theorem **cannot** be used. Instead, two rules, called the **sine rule** and the **cosine rule**, are used.

Sine rule

With reference to triangle *ABC* of Fig. 12.24, the **sine rule** states:

$$\frac{a}{\sin A} = \frac{b}{\sin B} = \frac{c}{\sin C}$$

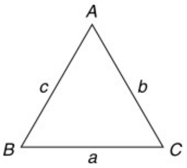

Figure 12.24

The rule may be used only when:

(i) one side and any two angles are initially given, or

(ii) two sides and an angle (not the included angle) are initially given.

Cosine rule

With reference to triangle *ABC* of Fig. 12.24, the **cosine rule** states:

$$a^2 = b^2 + c^2 - 2bc\cos A$$
$$\text{or}\quad b^2 = a^2 + c^2 - 2ac\cos B$$
$$\text{or}\quad c^2 = a^2 + b^2 - 2ab\cos C$$

The rule may be used only when:

(i) two sides and the included angle are initially given, or

(ii) three sides are initially given.

12.8 Area of any triangle

The **area of any triangle** such as *ABC* of Fig. 12.24 is given by:

(i) $\frac{1}{2}$ **× base × perpendicular height, or**

(ii) $\frac{1}{2}ab\sin C$ **or** $\frac{1}{2}ac\sin B$ **or** $\frac{1}{2}bc\sin A$**, or**

(iii) $\sqrt{[s(s-a)(s-b)(s-c)]}$**, where** $s = \frac{a+b+c}{2}$

12.9 Worked problems on the solution of triangles and finding their areas

Problem 27. In a triangle *XYZ*, $\angle X = 51°$, $\angle Y = 67°$ and $YZ = 15.2$ cm. Solve the triangle and find its area.

The triangle *XYZ* is shown in Fig. 12.25. Since the angles in a triangle add up to 180°, then $\boldsymbol{Z} = 180° - 51° - 67° = \mathbf{62°}$. Applying the sine rule:

$$\frac{15.2}{\sin 51°} = \frac{y}{\sin 67°} = \frac{z}{\sin 62°}$$

Using $\frac{15.2}{\sin 51°} = \frac{y}{\sin 67°}$ and transposing gives:

$$y = \frac{15.2\ \sin 67°}{\sin 51°} = \mathbf{18.00\,cm} = \boldsymbol{XZ}$$

Using $\frac{15.2}{\sin 51°} = \frac{z}{\sin 62°}$ and transposing gives:

$$z = \frac{15.2 \sin 62°}{\sin 51°} = \mathbf{17.27\,cm} = \boldsymbol{XY}$$

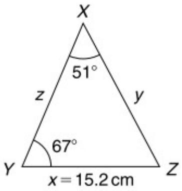

Figure 12.25

Area of triangle $\boldsymbol{XYZ} = \frac{1}{2}xy\sin Z$ $= \frac{1}{2}(15.2)(18.00)\sin 62° = \mathbf{120.8\,cm^2}$ (or area $= \frac{1}{2}xz\sin Y = \frac{1}{2}(15.2)(17.27)\sin 67° = \mathbf{120.8\,cm^2}$).

It is always worth checking with triangle problems that the longest side is opposite the largest angle, and vice-versa. In this problem, *Y* is the largest angle and *XZ* is the longest of the three sides.

Problem 28. Solve the triangle PQR and find its area given that $QR = 36.5$ mm, $PR = 29.6$ mm and $\angle Q = 36°$.

Triangle PQR is shown in Fig. 12.26.

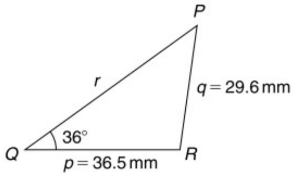

Figure 12.26

Applying the sine rule:

$$\frac{29.6}{\sin 36°} = \frac{36.5}{\sin P}$$

from which,

$$\sin P = \frac{36.5 \sin 36°}{29.6} = 0.7248$$

Hence $P = \sin^{-1} 0.7248 = 46°27'$ or $133°33'$

When $P = 46°27'$ and $Q = 36°$ then
$R = 180° - 46°27' - 36° = 97°33'$
When $P = 133°33'$ and $Q = 36°$ then
$R = 180° - 133°33' - 36° = 10°27'$

Thus, in this problem there are **two** separate sets of results and both are feasible solutions. Such a situation is called the **ambiguous case**.

Case 1. $P = 46°27'$, $Q = 36°$, $R = 97°33'$, $p = 36.5$ mm and $q = 29.6$ mm.
From the sine rule:

$$\frac{r}{\sin 97°33'} = \frac{29.6}{\sin 36°}$$

from which,

$$r = \frac{29.6 \sin 97°33'}{\sin 36°} = \mathbf{49.92\ mm}$$

$$\mathbf{Area} = \tfrac{1}{2} pq \sin R = \tfrac{1}{2}(36.5)(29.6)\sin 97°33'$$

$$= \mathbf{535.5\,mm^2}$$

Case 2. $P = 133°33'$, $Q = 36°$, $R = 10°27'$, $p = 36.5$ mm and $q = 29.6$ mm.
From the sine rule:

$$\frac{r}{\sin 10°27'} = \frac{29.6}{\sin 36°}$$

from which,

$$r = \frac{29.6 \sin 10°27'}{\sin 36°} = \mathbf{9.134\,mm}$$

$$\mathbf{Area} = \tfrac{1}{2} pq \sin R = \tfrac{1}{2}(36.5)(29.6)\sin 10°27'$$

$$= \mathbf{97.98\,mm^2}$$

Triangle PQR for case 2 is shown in Fig. 12.27.

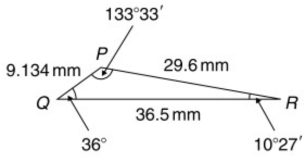

Figure 12.27

Now try the following Practice Exercise

Practice Exercise 56 Solving triangles and finding their areas (Answers on page 864)

In Problems 1 and 2, use the sine rule to solve the triangles ABC and find their areas.

1. $A = 29°$, $B = 68°$, $b = 27$ mm.
2. $B = 71°26'$, $C = 56°32'$, $b = 8.60$ cm.

In Problems 3 and 4, use the sine rule to solve the triangles DEF and find their areas.

3. $d = 17$ cm, $f = 22$ cm, $F = 26°$
4. $d = 32.6$ mm, $e = 25.4$ mm, $D = 104°22'$

In Problems 5 and 6, use the sine rule to solve the triangles JKL and find their areas.

5. $j = 3.85$ cm, $k = 3.23$ cm, $K = 36°$
6. $k = 46$ mm, $l = 36$ mm, $L = 35°$

12.10 Further worked problems on solving triangles and finding their areas

Problem 29. Solve triangle DEF and find its area given that $EF = 35.0$ mm, $DE = 25.0$ mm and $\angle E = 64°$

Triangle *DEF* is shown in Fig. 12.28.

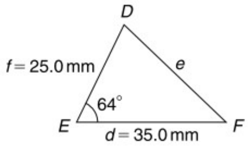

Figure 12.28

Applying the cosine rule:

$$e^2 = d^2 + f^2 - 2d\,f\cos E$$

i.e. $$e^2 = (35.0)^2 + (25.0)^2 - [2(35.0)(25.0)\cos 64°]$$

$$= 1225 + 625 - 767.1 = 1083$$

from which, $\boldsymbol{e} = \sqrt{1083} = \mathbf{32.91\,mm}$

Applying the sine rule:

$$\frac{32.91}{\sin 64°} = \frac{25.0}{\sin F}$$

from which, $$\sin F = \frac{25.0\sin 64°}{32.91} = 0.6828$$

Thus $$\angle F = \sin^{-1} 0.6828$$
$$= 43°4' \text{ or } 136°56'$$

$F = 136°56'$ is not possible in this case since $136°56' + 64°$ is greater than 180°. Thus only $\boldsymbol{F} = \mathbf{43°4'}$ is valid

$$\angle \boldsymbol{D} = 180° - 64° - 43°4' = \mathbf{72°56'}$$

Area of triangle $DEF = \frac{1}{2}\,d\,f\sin E$
$= \frac{1}{2}(35.0)(25.0)\sin 64° = \mathbf{393.2\,mm^2}$

Problem 30. A triangle *ABC* has sides $a = 9.0$ cm, $b = 7.5$ cm and $c = 6.5$ cm. Determine its three angles and its area.

Triangle *ABC* is shown in Fig. 12.29. It is usual first to calculate the largest angle to determine whether the triangle is acute or obtuse. In this case the largest angle is *A* (i.e. opposite the longest side).

Applying the cosine rule:

$$a^2 = b^2 + c^2 - 2bc\cos A$$

from which, $2bc\cos A = b^2 + c^2 - a^2$

and $$\cos A = \frac{b^2 + c^2 - a^2}{2bc} = \frac{7.5^2 + 6.5^2 - 9.0^2}{2(7.5)(6.5)}$$
$$= 0.1795$$

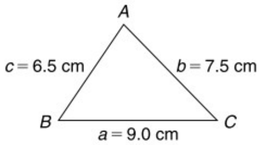

Figure 12.29

Hence $A = \cos^{-1} 0.1795 = \mathbf{79°40'}$ (or 280°20′, which is obviously impossible). The triangle is thus acute angled since cos *A* is positive. (If cos *A* had been negative, angle *A* would be obtuse, i.e. lie between 90° and 180°.)

Applying the sine rule:

$$\frac{9.0}{\sin 79°40'} = \frac{7.5}{\sin B}$$

from which,

$$\sin B = \frac{7.5\sin 79°40'}{9.0} = 0.8198$$

Hence $\boldsymbol{B} = \sin^{-1} 0.8198 = \mathbf{55°4'}$

and $\boldsymbol{C} = 180° - 79°40' - 55°4' = \mathbf{45°16'}$

$$\text{Area} = \sqrt{[s(s-a)(s-b)(s-c)]}$$

where $$s = \frac{a+b+c}{2} = \frac{9.0 + 7.5 + 6.5}{2}$$
$$= 11.5\,\text{cm}$$

Hence **area**

$$= \sqrt{[11.5(11.5 - 9.0)(11.5 - 7.5)(11.5 - 6.5)]}$$

$$= \sqrt{[11.5(2.5)(4.0)(5.0)]} = \mathbf{23.98\,cm^2}$$

Alternatively, area $= \frac{1}{2}ab\sin C$
$= \frac{1}{2}(9.0)(7.5)\sin 45°16' = \mathbf{23.98\,cm^2}$

Now try the following Practice Exercise

Practice Exercise 57 Solving triangles and finding their areas (Answers on page 864)

In Problems 1 and 2, use the cosine and sine rules to solve the triangles *PQR* and find their areas.

1. $q = 12$ cm, $r = 16$ cm, $P = 54°$
2. $q = 3.25$ m, $r = 4.42$ m, $P = 105°$

In problems 3 and 4, use the cosine and sine rules to solve the triangles *XYZ* and find their areas.

3. $x=10.0$ cm, $y=8.0$ cm, $z=7.0$ cm.
4. $x=21$ mm, $y=34$ mm, $z=42$ mm.

12.11 Practical situations involving trigonometry

There are a number of **practical situations** where the use of trigonometry is needed to find unknown sides and angles of triangles. This is demonstrated in the following problems.

Problem 31. A room 8.0 m wide has a span roof which slopes at 33° on one side and 40° on the other. Find the length of the roof slopes, correct to the nearest centimetre.

A section of the roof is shown in Fig. 12.30.

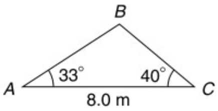

Figure 12.30

Angle at ridge, $B=180^\circ-33^\circ-40^\circ=107^\circ$
From the sine rule:

$$\frac{8.0}{\sin 107^\circ}=\frac{a}{\sin 33^\circ}$$

from which,

$$a=\frac{8.0\sin 33^\circ}{\sin 107^\circ}=4.556\,\text{m}$$

Also from the sine rule:

$$\frac{8.0}{\sin 107^\circ}=\frac{c}{\sin 40^\circ}$$

from which,

$$c=\frac{8.0\sin 40^\circ}{\sin 107^\circ}=5.377\,\text{m}$$

Hence the roof slopes are 4.56 m and 5.38 m, correct to the nearest centimetre.

Problem 32. Two voltage phasors are shown in Fig. 12.31. If $V_1=40$ V and $V_2=100$ V determine the value of their resultant (i.e. length OA) and the angle the resultant makes with V_1

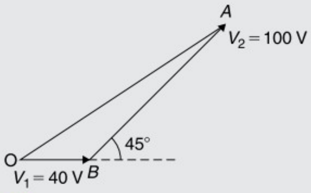

Figure 12.31

Angle $OBA=180^\circ-45^\circ=135^\circ$

Applying the cosine rule:

$$\begin{aligned}OA^2&=V_1^2+V_2^2-2V_1V_2\cos OBA\\&=40^2+100^2-\{2(40)(100)\cos 135^\circ\}\\&=1600+10\,000-\{-5657\}\\&=1600+10\,000+5657=17\,257\end{aligned}$$

The resultant

$$OA=\sqrt{(17\,257)}=131.4\,\text{V}$$

Applying the sine rule:

$$\frac{131.4}{\sin 135^\circ}=\frac{100}{\sin AOB}$$

from which, $$\sin AOB=\frac{100\sin 135^\circ}{131.4}=0.5381$$

Hence angle $AOB=\sin^{-1}0.5381=32^\circ 33'$ (or $147^\circ 27'$, which is impossible in this case).

Hence the resultant voltage is 131.4 volts at 32°33′ to V_1

Problem 33. In Fig. 12.32, PR represents the inclined jib of a crane and is 10.0 long. PQ is 4.0 m long. Determine the inclination of the jib to the vertical and the length of tie QR.

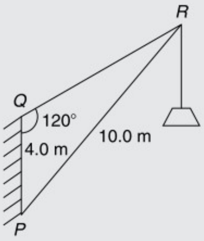

Figure 12.32

Applying the sine rule:

$$\frac{PR}{\sin 120^\circ} = \frac{PQ}{\sin R}$$

from which,

$$\sin R = \frac{PQ \sin 120^\circ}{PR} = \frac{(4.0)\sin 120^\circ}{10.0} = 0.3464$$

Hence $\angle R = \sin^{-1} 0.3464 = 20^\circ 16'$ (or $159^\circ 44'$, which is impossible in this case).

$\angle \boldsymbol{P} = 180^\circ - 120^\circ - 20^\circ 16' = \mathbf{39^\circ 44'}$, **which is the inclination of the jib to the vertical**.

Applying the sine rule:

$$\frac{10.0}{\sin 120^\circ} = \frac{QR}{\sin 39^\circ 44'}$$

from which, **length of tie**,

$$\boldsymbol{QR} = \frac{10.0 \sin 39^\circ 44'}{\sin 120^\circ} = \mathbf{7.38\,m}$$

Now try the following Practice Exercise

Practice Exercise 58 Practical situations involving trigonometry (Answers on page 864)

1. A ship P sails at a steady speed of 45 km/h in a direction of W 32° N (i.e. a bearing of 302°) from a port. At the same time another ship Q leaves the port at a steady speed of 35 km/h in a direction N 15° E (i.e. a bearing of 015°). Determine their distance apart after four hours.
2. Two sides of a triangular plot of land are 52.0 m and 34.0 m, respectively. If the area of the plot is 620 m^2 find (a) the length of fencing required to enclose the plot and (b) the angles of the triangular plot.
3. A jib crane is shown in Fig. 12.33. If the tie rod PR is 8.0 long and PQ is 4.5 m long, determine (a) the length of jib RQ and (b) the angle between the jib and the tie rod.

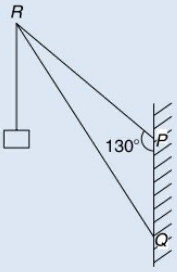

Figure 12.33

4. A building site is in the form of a quadrilateral, as shown in Fig. 12.34, and its area is 1510 m^2. Determine the length of the perimeter of the site.

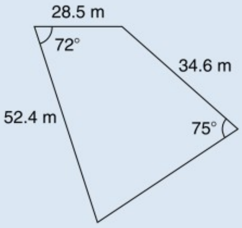

Figure 12.34

5. Determine the length of members BF and EB in the roof truss shown in Fig. 12.35.

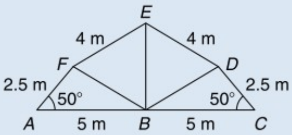

Figure 12.35

6. A laboratory 9.0 m wide has a span roof which slopes at 36° on one side and 44° on the other. Determine the lengths of the roof slopes.

12.12 Further practical situations involving trigonometry

Problem 34. A vertical aerial stands on horizontal ground. A surveyor positioned due east of the aerial measures the elevation of the top as 48°. He moves due south 30.0 m and measures the elevation as 44°. Determine the height of the aerial.

In Fig. 12.36, DC represents the aerial, A is the initial position of the surveyor and B his final position.

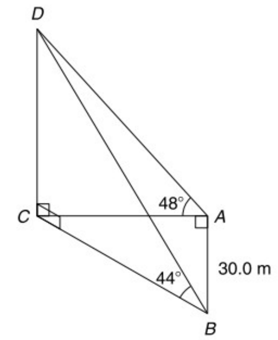

Figure 12.36

From triangle ACD, $\tan 48^\circ = \dfrac{DC}{AC}$

from which $$AC = \frac{DC}{\tan 48^\circ}$$

Similarly, from triangle BCD,

$$BC = \frac{DC}{\tan 44^\circ}$$

For triangle ABC, using Pythagoras' theorem:

$$BC^2 = AB^2 + AC^2$$

$$\left(\frac{DC}{\tan 44^\circ}\right)^2 = (30.0)^2 + \left(\frac{DC}{\tan 48^\circ}\right)^2$$

$$DC^2\left(\frac{1}{\tan^2 44^\circ} - \frac{1}{\tan^2 48^\circ}\right) = 30.0^2$$

$$DC^2(1.072323 - 0.810727) = 30.0^2$$

$$DC^2 = \frac{30.0^2}{0.261596} = 3440.4$$

Hence, height of aerial,

$$\boldsymbol{DC = \sqrt{3440.4} = 58.65\,\text{m}}$$

Problem 35. A crank mechanism of a petrol engine is shown in Fig. 12.37. Arm OA is 10.0 cm long and rotates clockwise about O. The connecting rod AB is 30.0 cm long and end B is constrained to move horizontally.

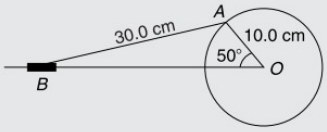

Figure 12.37

(a) For the position shown in Fig. 12.37 determine the angle between the connecting rod AB and the horizontal and the length of OB.

(b) How far does B move when angle AOB changes from 50° to 120°?

(a) Applying the sine rule:

$$\frac{AB}{\sin 50^\circ} = \frac{AO}{\sin B}$$

from which,

$$\sin B = \frac{AO \sin 50^\circ}{AB} = \frac{10.0 \sin 50^\circ}{30.0} = 0.2553$$

Hence $B = \sin^{-1} 0.2553 = 14^\circ 47'$ (or $165^\circ 13'$, which is impossible in this case).

Hence the connecting rod AB makes an angle of $14^\circ 47'$ with the horizontal.

Angle $OAB = 180^\circ - 50^\circ - 14^\circ 47' = 115^\circ 13'$.

Applying the sine rule:

$$\frac{30.0}{\sin 50^\circ} = \frac{OB}{\sin 115^\circ 13'}$$

from which,

$$\boldsymbol{OB} = \frac{30.0 \sin 115^\circ 13'}{\sin 50^\circ} = \boldsymbol{35.43\,\text{cm}}$$

(b) Fig. 12.38 shows the initial and final positions of the crank mechanism. In triangle $OA'B'$, applying the sine rule:

$$\frac{30.0}{\sin 120^\circ} = \frac{10.0}{\sin A'B'O}$$

from which,

$$\sin A'B'O = \frac{10.0 \sin 120^\circ}{30.0} = 0.2887$$

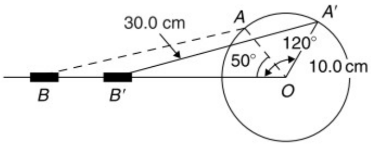

Figure 12.38

Hence $A'B'O = \sin^{-1} 0.2887 = 16^\circ 47'$ (or $163^\circ 13'$ which is impossible in this case).

Angle $OA'B' = 180^\circ - 120^\circ - 16^\circ 47' = 43^\circ 13'$

Applying the sine rule:

$$\frac{30.0}{\sin 120^\circ} = \frac{OB'}{\sin 43^\circ 13'}$$

from which,

$$OB' = \frac{30.0 \sin 43^\circ 13'}{\sin 120^\circ} = 23.72\,\text{cm}$$

Since $OB = 35.43\,\text{cm}$ and $OB' = 23.72\,\text{cm}$ then $BB' = 35.43 - 23.72 = 11.71\,\text{cm}$.

Hence *B* moves 11.71 cm when angle *AOB* changes from 50° to 120°

Problem 36. The area of a field is in the form of a quadrilateral *ABCD* as shown in Fig. 12.39. Determine its area.

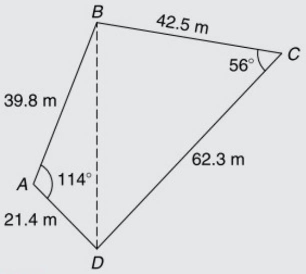

Figure 12.39

A diagonal drawn from *B* to *D* divides the quadrilateral into two triangles.

Area of quadrilateral *ABCD*

$$= \text{area of triangle } ABD + \text{area of triangle } BCD$$

$$= \tfrac{1}{2}(39.8)(21.4)\sin 114^\circ + \tfrac{1}{2}(42.5)(62.3)\sin 56^\circ$$

$$= 389.04 + 1097.5 = \mathbf{1487\,m^2}$$

Now try the following Practice Exercise

Practice Exercise 59 Practical situations involving trigonometry (Answers on page 864)

1. *PQ* and *QR* are the phasors representing the alternating currents in two branches of a circuit. Phasor *PQ* is 20.0 A and is horizontal. Phasor *QR* (which is joined to the end of *PQ* to form triangle *PQR*) is 14.0 A and is at an angle of 35° to the horizontal. Determine the resultant phasor *PR* and the angle it makes with phasor *PQ*.
2. Three forces acting on a fixed point are represented by the sides of a triangle of dimensions 7.2 cm, 9.6 cm and 11.0 cm. Determine the angles between the lines of action and the three forces.
3. Calculate, correct to 3 significant figures, the co-ordinates *x* and *y* to locate the hole centre at *P* shown in Fig. 12.40.

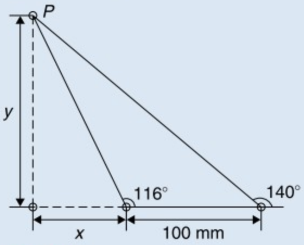

Figure 12.40

4. An idler gear, 30 mm in diameter, has to be fitted between a 70 mm diameter driving gear and a 90 mm diameter driven gear as shown in Fig. 12.41. Determine the value of angle θ between the centre lines.

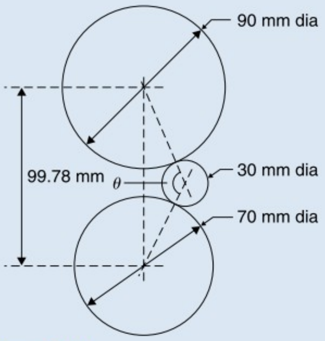

Figure 12.41

5. A reciprocating engine mechanism is shown in Fig. 12.42. The crank AB is 12.0 cm long and the connecting rod BC is 32.0 cm long. For the position shown determine the length of AC and the angle between the crank and the connecting rod.

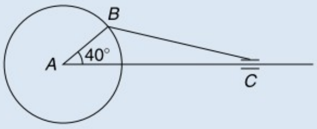

Figure 12.42

6. From Fig. 12.42, determine how far C moves, correct to the nearest millimetre, when angle CAB changes from 40° to 160°, B moving in an anticlockwise direction.

7. A surveyor, standing W 25° S of a tower, measures the angle of elevation of the top of the tower as $46°30'$. From a position E 23° S from the tower the elevation of the top is $37°15'$. Determine the height of the tower if the distance between the two observations is 75 m.

8. An aeroplane is sighted due east from a radar station at an elevation of 40° and a height of 8000 m and later at an elevation of 35° and height 5500 m in a direction E 70° S. If it is descending uniformly, find the angle of descent. Determine also the speed of the aeroplane in km/h if the time between the two observations is 45 s.

9. Sixteen holes are equally spaced on a pitch circle of 70 mm diameter. Determine the length of the chord joining the centres of two adjacent holes.

For fully worked solutions to each of the problems in Practice Exercises 51 to 59 in this chapter, go to the website:
www.routledge.com/cw/bird

Chapter 13

Cartesian and polar co-ordinates

***Why it is important to understand:* Cartesian and polar co-ordinates**

Applications where polar co-ordinates would be used include terrestrial navigation with sonar-like devices, and those in engineering and science involving energy radiation patterns. Applications where Cartesian co-ordinates would be used include any navigation on a grid and anything involving raster graphics (e.g. bitmap – a dot matrix data structure representing a generally rectangular grid of pixels). The ability to change from Cartesian to polar co-ordinates is vitally important when using complex numbers and their use in a.c. electrical circuit theory and with vector geometry.

At the end of this chapter, you should be able to:

- change from Cartesian to polar co-ordinates
- change from polar to Cartesian co-ordinates
- use a scientific notation calculator to change from Cartesian to polar co-ordinates and vice-versa

---

Higher Engineering Mathematics. 978-1-138-67357-1, © 2017 John Bird. Published by Taylor & Francis. All rights reserved.

13.1 Introduction

There are two ways in which the position of a point in a plane can be represented. These are

(a) by **Cartesian co-ordinates**, (named after **Descartes***), i.e. (x, y), and

(b) by **polar co-ordinates**, i.e. (r, θ), where r is a 'radius' from a fixed point and θ is an angle from a fixed point.

13.2 Changing from Cartesian into polar co-ordinates

In Fig. 13.1, if lengths x and y are known, then the length of r can be obtained from Pythagoras' theorem (see Chapter 12) since OPQ is a right-angled triangle. Hence $r^2 = (x^2 + y^2)$

from which, $\boldsymbol{r = \sqrt{x^2 + y^2}}$

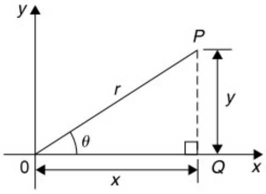

Figure 13.1

From trigonometric ratios (see Chapter 12),

$$\tan\theta = \frac{y}{x}$$

from which $\boldsymbol{\theta = \tan^{-1}\frac{y}{x}}$

$r = \sqrt{x^2 + y^2}$ and $\theta = \tan^{-1}\frac{y}{x}$ are the two formulae we need to change from Cartesian to polar co-ordinates. The angle θ, which may be expressed in degrees or radians, must **always** be measured from the positive x-axis, i.e., measured from the line OQ in Fig. 13.1. It is suggested that when changing from Cartesian to polar co-ordinates a diagram should always be sketched.

Problem 1. Change the Cartesian co-ordinates (3, 4) into polar co-ordinates.

A diagram representing the point (3, 4) is shown in Fig. 13.2.

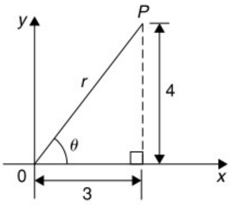

Figure 13.2

From Pythagoras' theorem, $r = \sqrt{3^2 + 4^2} = 5$ (note that -5 has no meaning in this context). By trigonometric ratios, $\theta = \tan^{-1}\frac{4}{3} = 53.13°$ or 0.927 rad.

[Note that $53.13° = 53.13 \times (\pi/180)\,\text{rad} = 0.927\,\text{rad}$.]

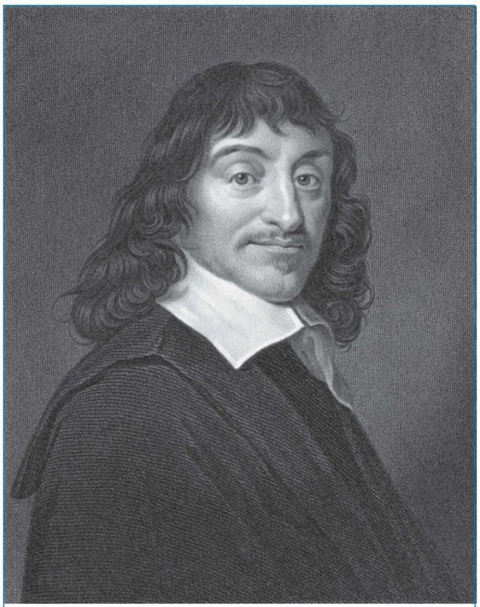

* **Who was Descartes?** **René Descartes** (31 March 1596–11 February 1650) was a French philosopher, mathematician, and writer. He wrote many influential texts including *Meditations on First Philosophy*. Descartes is best known for the philosophical statement '*Cogito ergo sum*' (I think, therefore I am), found in part IV of *Discourse on the Method*. To find out more go to **www.routledge.com/cw/bird**

Hence (3, 4) in Cartesian co-ordinates corresponds to (5, 53.13°) or (5, 0.927 rad) in polar co-ordinates.

Problem 2. Express in polar co-ordinates the position $(-4, 3)$

A diagram representing the point using the Cartesian co-ordinates $(-4, 3)$ is shown in Fig. 13.3.

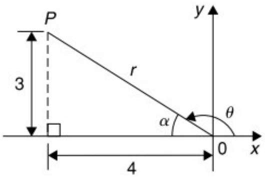

Figure 13.3

From Pythagoras' theorem, $r=\sqrt{4^2+3^2}=5$

By trigonometric ratios, $\alpha=\tan^{-1}\frac{3}{4}=36.87^\circ$ or 0.644 rad.

Hence $\theta=180^\circ-36.87^\circ=143.13^\circ$ or $\theta=\pi-0.644=2.498$ rad.

Hence the position of point *P* in polar co-ordinate form is (5, 143.13°) or (5, 2.498 rad).

Problem 3. Express $(-5, -12)$ in polar co-ordinates.

A sketch showing the position $(-5, -12)$ is shown in Fig. 13.4.

$$r=\sqrt{5^2+12^2}=13$$

and
$$\alpha=\tan^{-1}\frac{12}{5}$$
$$=67.38^\circ \text{ or } 1.176\,\text{rad}$$

Hence
$$\theta=180^\circ+67.38^\circ=247.38^\circ \text{ or}$$
$$\theta=\pi+1.176=4.318\,\text{rad}$$

Thus (−5, −12) in Cartesian co-ordinates corresponds to (13, 247.38°) or (13, 4.318 rad) in polar co-ordinates.

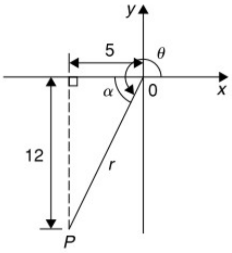

Figure 13.4

Problem 4. Express $(2, -5)$ in polar co-ordinates.

A sketch showing the position $(2, -5)$ is shown in Fig. 13.5.

$$r=\sqrt{2^2+5^2}=\sqrt{29}=5.385 \text{ correct to 3 decimal places.}$$

$$\alpha=\tan^{-1}\frac{5}{2}=68.20^\circ \text{ or } 1.190\,\text{rad}$$

Hence
$$\theta=360^\circ-68.20^\circ=291.80^\circ \text{ or}$$
$$\theta=2\pi-1.190=5.093\,\text{rad}$$

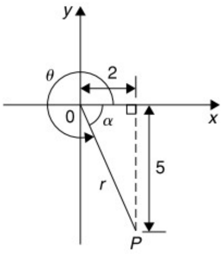

Figure 13.5

Thus (2, −5) in Cartesian co-ordinates corresponds to (5.385, 291.80°) or (5.385, 5.093 rad) in polar co-ordinates.

Now try the following Practice Exercise

Practice Exercise 60 Changing Cartesian into polar co-ordinates (Answers on page 864)

In Problems 1 to 8, express the given Cartesian co-ordinates as polar co-ordinates, correct to 2 decimal places, in both degrees and in radians.

1. (3, 5)
2. (6.18, 2.35)
3. (−2, 4)
4. (−5.4, 3.7)
5. (−7, −3)
6. (−2.4, −3.6)
7. (5, −3)
8. (9.6, −12.4)

13.3 Changing from polar into Cartesian co-ordinates

From the right-angled triangle *OPQ* in Fig. 13.6,

$$\cos\theta = \frac{x}{r} \text{ and } \sin\theta = \frac{y}{r}, \text{ from trigonometric ratios}$$

Hence $\boldsymbol{x = r\cos\theta}$ and $\boldsymbol{y = r\sin\theta}$

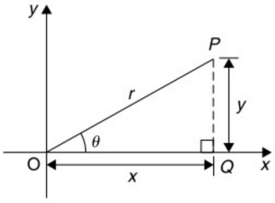

Figure 13.6

If lengths r and angle θ are known then $x = r\cos\theta$ and $y = r\sin\theta$ are the two formulae we need to change from polar to Cartesian co-ordinates.

Problem 5. Change (4, 32°) into Cartesian co-ordinates.

A sketch showing the position (4, 32°) is shown in Fig. 13.7.

Now $x = r\cos\theta = 4\cos 32^\circ = 3.39$

and $y = r\sin\theta = 4\sin 32^\circ = 2.12$

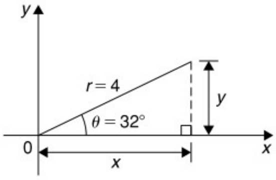

Figure 13.7

Hence (4, 32°) in polar co-ordinates corresponds to (3.39, 2.12) in Cartesian co-ordinates.

Problem 6. Express (6, 137°) in Cartesian co-ordinates.

A sketch showing the position (6, 137°) is shown in Fig. 13.8.

$$x = r\cos\theta = 6\cos 137^\circ = -4.388$$

which corresponds to length *OA* in Fig. 13.8.

$$y = r\sin\theta = 6\sin 137^\circ = 4.092$$

which corresponds to length *AB* in Fig. 13.8.

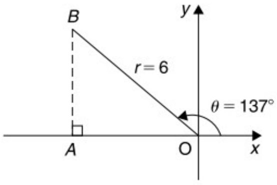

Figure 13.8

Thus (6, 137°) in polar co-ordinates corresponds to (−4.388, 4.092) in Cartesian co-ordinates.

(Note that when changing from polar to Cartesian co-ordinates it is not quite so essential to draw a sketch. Use of $x = r\cos\theta$ and $y = r\sin\theta$ automatically produces the correct signs.)

Problem 7. Express (4.5, 5.16 rad) in Cartesian co-ordinates.

A sketch showing the position (4.5, 5.16 rad) is shown in Fig. 13.9.

$$x = r\cos\theta = 4.5\cos 5.16 = 1.948$$

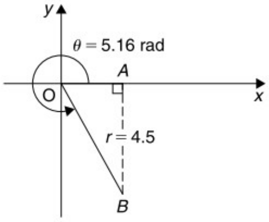

Figure 13.9

which corresponds to length OA in Fig. 13.9.

$$y = r\sin\theta = 4.5\sin 5.16 = -4.057$$

which corresponds to length AB in Fig. 13.9.

Thus (1.948, −4.057) in Cartesian co-ordinates corresponds to (4.5, 5.16 rad) in polar co-ordinates.

Now try the following Practice Exercise

Practice Exercise 61 Changing polar into Cartesian co-ordinates (Answers on page 864)

In Problems 1 to 8, express the given polar co-ordinates as Cartesian co-ordinates, correct to 3 decimal places.

1. (5, 75°)
2. (4.4, 1.12 rad)
3. (7, 140°)
4. (3.6, 2.5 rad)
5. (10.8, 210°)
6. (4, 4 rad)
7. (1.5, 300°)
8. (6, 5.5 rad)
9. Fig. 13.10 shows five equally spaced holes on an 80 mm pitch circle diameter. Calculate their co-ordinates relative to axes Ox and Oy in (a) polar form, (b) Cartesian form.
   (c) Calculate also the shortest distance between the centres of two adjacent holes.

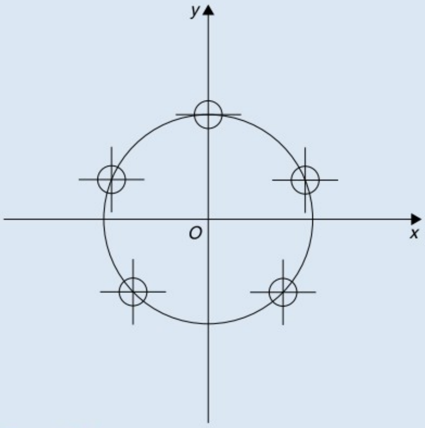

Figure 13.10

13.4 Use of Pol/Rec functions on calculators

Another name for Cartesian co-ordinates is **rectangular** co-ordinates. Many scientific notation calculators possess **Pol** and **Rec** functions. 'Rec' is an abbreviation of 'rectangular' (i.e. Cartesian) and 'Pol' is an abbreviation of 'polar'. Check the operation manual for your particular calculator to determine how to use these two functions. They make changing from Cartesian to polar co-ordinates, and vice-versa, so much quicker and easier.

For example, with the Casio fx-991ES PLUS calculator, or similar, to change the Cartesian number (3, 4) into polar form, the following procedure is adopted:

1. Press 'shift' 2. Press 'Pol' 3. Enter 3
4. Enter 'comma' (obtained by 'shift' then))
5. Enter 4 6. Press)
7. Press = The answer is: $\boldsymbol{r = 5, \theta = 53.13°}$

Hence, (3, 4) in Cartesian form is the same as $(5, 53.13°)$ in polar form.

If the angle is required in **radians**, then before repeating the above procedure press ‘shift’, ‘mode’ and then 4 to change your calculator to radian mode.

Similarly, to change the polar form number (7, 126°) into Cartesian or rectangular form, adopt the following procedure:

1. Press ‘shift’ 2. Press ‘Rec’ 3. Enter 7
4. Enter ‘comma’
5. Enter 126 (assuming your calculator is in degrees mode)
6. Press) 7. Press =

The answer is: $X = -4.11$, **and scrolling across,** $Y = 5.66$, correct to 2 decimal places.

Hence, (7, 126°) in polar form is the same as $(-4.11, 5.66)$ in rectangular or Cartesian form.

Now return to Practice Exercises 60 and 61 in this chapter and use your calculator to determine the answers, and see how much more quickly they may be obtained.

For fully worked solutions to each of the problems in Practice Exercises 60 and 61 in this chapter, go to the website:
www.routledge.com/cw/bird

Chapter 14

The circle and its properties

***Why it is important to understand:* The circle and its properties**

A circle is one of the fundamental shapes of geometry; it consists of all the points that are equidistant from a central point. Knowledge of calculations involving circles is needed with crank mechanisms, with determinations of latitude and longitude, with pendulums, and even in the design of paper clips. The floodlit area at a football ground, the area an automatic garden sprayer sprays and the angle of lap of a belt drive all rely on calculations involving the arc of a circle. The ability to handle calculations involving circles and its properties is clearly essential in several branches of engineering design.

At the end of this chapter, you should be able to:

- define a circle
- state some properties of a circle – including radius, circumference, diameter, semicircle, quadrant, tangent, sector, chord, segment and arc
- appreciate the angle in a semicircle is a right angle
- define a radian, change radians to degrees, and vice-versa
- determine arc length, area of a circle and area of a sector of a circle
- state the equation of a circle
- sketch a circle given its equation
- understand linear and angular velocity
- understand centripetal force

14.1 Introduction

A **circle** is a plain figure enclosed by a curved line, every point on which is equidistant from a point within, called the **centre**.

14.2 Properties of circles

(i) The distance from the centre to the curve is called the **radius**, r, of the circle (see OP in Fig. 14.1).

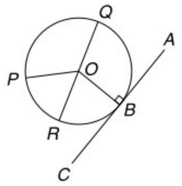

Figure 14.1

(ii) The boundary of a circle is called the **circumference**, c.

Higher Engineering Mathematics. 978-1-138-67357-1, © 2017 John Bird. Published by Taylor & Francis. All rights reserved.

(iii) Any straight line passing through the centre and touching the circumference at each end is called the **diameter**, d (see QR in Fig. 14.1). Thus $\boldsymbol{d=2r}$.

(iv) The ratio $\dfrac{\text{circumference}}{\text{diameter}}=$ a constant for any circle.

This constant is denoted by the Greek letter π (pronounced 'pie'), where $\pi=3.14159$, correct to 5 decimal places.

Hence $c/d=\pi$ or $\boldsymbol{c=\pi d}$ or $\boldsymbol{c=2\pi r}$.

(v) A **semicircle** is one half of the whole circle.

(vi) A **quadrant** is one quarter of a whole circle.

(vii) A **tangent** to a circle is a straight line which meets the circle in one point only and does not cut the circle when produced. AC in Fig. 14.1 is a tangent to the circle since it touches the curve at point B only. If radius OB is drawn, then angle ABO is a right angle.

(viii) A **sector** of a circle is the part of a circle between radii (for example, the portion OXY of Fig. 14.2 is a sector). If a sector is less than a semicircle it is called a **minor sector**, if greater than a semicircle it is called a **major sector**.

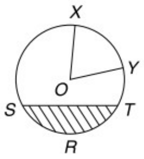

Figure 14.2

(ix) A **chord** of a circle is any straight line which divides the circle into two parts and is terminated at each end by the circumference. ST, in Fig. 14.2 is a chord.

(x) A **segment** is the name given to the parts into which a circle is divided by a chord. If the segment is less than a semicircle it is called a **minor segment** (see shaded area in Fig. 14.2). If the segment is greater than a semicircle it is called a **major segment** (see the unshaded area in Fig. 14.2).

(xi) An **arc** is a portion of the circumference of a circle. The distance SRT in Fig. 14.2 is called a **minor arc** and the distance $SXYT$ is called a **major arc**.

(xii) The angle at the centre of a circle, subtended by an arc, is double the angle at the circumference subtended by the same arc. With reference to Fig. 14.3, **Angle** $\boldsymbol{AOC=2\times}$ **angle** $\boldsymbol{ABC}$.

(xiii) The angle in a semicircle is a right angle (see angle BQP in Fig. 14.3).

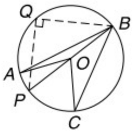

Figure 14.3

Problem 1. If the diameter of a circle is 75 mm, find its circumference.

$$\text{Circumference, } c=\pi\times\text{diameter}=\pi d$$
$$=\pi(75)=\mathbf{235.6\,mm}$$

Problem 2. In Fig. 14.4, AB is a tangent to the circle at B. If the circle radius is 40 mm and $AB=150$ mm, calculate the length AO.

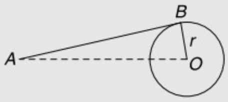

Figure 14.4

A tangent to a circle is at right angles to a radius drawn from the point of contact, i.e. $ABO=90°$. Hence, using Pythagoras' theorem:

$$AO^2=AB^2+OB^2$$
$$AO=\sqrt{(AB^2+OB^2)}=\sqrt{[(150)^2+(40)^2]}$$
$$=\mathbf{155.2\,mm}$$

Now try the following Practice Exercise

Practice Exercise 62 Properties of circles (Answers on page 864)

1. If the radius of a circle is 41.3 mm, calculate the circumference of the circle.
2. Find the diameter of a circle whose perimeter is 149.8 cm.
3. A crank mechanism is shown in Fig. 14.5, where XY is a tangent to the circle at point X. If

the circle radius OX is 10 cm and length OY is 40 cm, determine the length of the connecting rod XY.

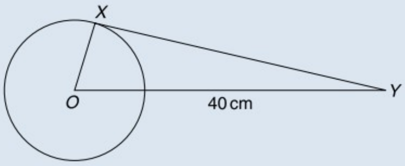

Figure 14.5

4. If the circumference of the Earth is 40 000 km at the equator, calculate its diameter.
5. Calculate the length of wire in the paper clip shown in Fig. 14.6. The dimensions are in millimetres.

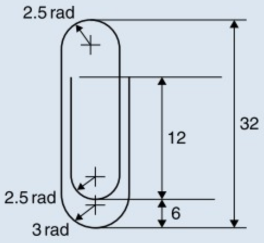

Figure 14.6

14.3 Radians and degrees

One **radian** is defined as the angle subtended at the centre of a circle by an arc equal in length to the radius.

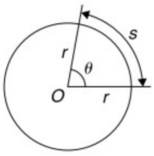

Figure 14.7

With reference to Fig. 14.7, for arc length s,

$$\theta \text{ radians} = \frac{s}{r}$$

When $s =$ whole circumference $(= 2\pi r)$ then $\theta = \frac{s}{r} = \frac{2\pi r}{r} = 2\pi$

i.e. $\mathbf{2\pi}$ **radians = 360°** or $\boldsymbol{\pi}$ **radians = 180°**

Thus, $\mathbf{1\,rad = \frac{180°}{\pi} = 57.30°}$, correct to 2 decimal places.

Since π rad $= 180°$, then $\frac{\pi}{2} = 90°, \frac{\pi}{3} = 60°, \frac{\pi}{4} = 45°$, and so on.

Problem 3. Convert to radians: (a) 125° (b) 69°47′

(a) Since $180° = \pi$ rad then $1° = \pi/180$ rad, therefore

$$125° = 125\left(\frac{\pi}{180}\right)^c = \mathbf{2.182\ rad}$$

(Note that ^{c} means 'circular measure' and indicates radian measure.)

(b) $69°47' = 69\frac{47°}{60} = 69.783°$

$$69.783° = 69.783\left(\frac{\pi}{180}\right)^c = \mathbf{1.218\,rad}$$

Problem 4. Convert to degrees and minutes: (a) 0.749 rad (b) $3\pi/4$ rad.

(a) Since π rad $= 180°$ then 1 rad $= 180°/\pi$, therefore

$$0.749 = 0.749\left(\frac{180}{\pi}\right)^{\circ} = 42.915°$$

$0.915° = (0.915 \times 60)' = 55'$, correct to the nearest minute, hence

$$\mathbf{0.749\ rad = 42°55'}$$

(b) Since 1 rad $= \left(\frac{180}{\pi}\right)^{\circ}$ then

$$\frac{3\pi}{4}\text{ rad} = \frac{3\pi}{4}\left(\frac{180}{\pi}\right)^{\circ} = \frac{3}{4}(180)^{\circ} = \mathbf{135°}$$

Problem 5. Express in radians, in terms of π, (a) 150° (b) 270° (c) 37.5°

Since $180^\circ = \pi$ rad then $1^\circ = \pi/180$, hence

(a) $150^\circ = 150\left(\frac{\pi}{180}\right)$ rad $= \mathbf{\frac{5\pi}{6}}$ **rad**

(b) $270^\circ = 270\left(\frac{\pi}{180}\right)$ rad $= \mathbf{\frac{3\pi}{2}}$ **rad**

(c) $37.5^\circ = 37.5\left(\frac{\pi}{180}\right)$ rad $= \frac{75\pi}{360}$ rad $= \mathbf{\frac{5\pi}{24}}$ **rad**

Now try the following Practice Exercise

Practice Exercise 63 Radians and degrees (Answers on page 865)

1. Convert to radians in terms of π: (a) 30° (b) 75° (c) 225°
2. Convert to radians: (a) 48° (b) $84^\circ51'$ (c) $232^\circ15'$
3. Convert to degrees: (a) $\frac{7\pi}{6}$ rad (b) $\frac{4\pi}{9}$ rad (c) $\frac{7\pi}{12}$ rad.
4. Convert to degrees and minutes: (a) 0.0125 rad (b) 2.69 rad (c) 7.241 rad.
5. A car engine speed is 1000 rev/min. Convert this speed into rad/s.

14.4 Arc length and area of circles and sectors

Arc length

From the definition of the radian in the previous section and Fig. 14.7,

arc length, $\boldsymbol{s = r\theta}$ where θ is in radians

Area of circle

For any circle, area $= \pi \times (\text{radius})^2$

i.e. $\mathbf{area} = \boldsymbol{\pi r^2}$

Since $r = \frac{d}{2}$, then $\mathbf{area} = \boldsymbol{\pi r^2}$ or $\boldsymbol{\frac{\pi d^2}{4}}$

Area of sector

Area of a sector $= \boldsymbol{\frac{\theta}{360}(\pi r^2)}$ when θ is in degrees

$$= \frac{\theta}{2\pi}(\pi r^2) = \boldsymbol{\frac{1}{2}r^2\theta}$$

when θ is in radians

Problem 6. A hockey pitch has a semicircle of radius 14.63 m around each goal net. Find the area enclosed by the semicircle, correct to the nearest square metre.

Area of a semicircle $= \frac{1}{2}\pi r^2$

When $r = 14.63$ m, area $= \frac{1}{2}\pi(14.63)^2$

i.e. **area of semicircle = 336 m^2**

Problem 7. Find the area of a circular metal plate, correct to the nearest square millimetre, having a diameter of 35.0 mm.

Area of a circle $= \pi r^2 = \frac{\pi d^2}{4}$

When $d = 35.0$ mm, area $= \frac{\pi(35.0)^2}{4}$

i.e. **area of circular plate = 962 mm^2**

Problem 8. Find the area of a circle having a circumference of 60.0 mm.

Circumference, $c = 2\pi r$

from which radius $r = \frac{c}{2\pi} = \frac{60.0}{2\pi} = \frac{30.0}{\pi}$

Area of a circle $= \pi r^2$

i.e. **area** $= \pi\left(\frac{30.0}{\pi}\right)^2 = \mathbf{286.5\,mm^2}$

Problem 9. Find the length of arc of a circle of radius 5.5 cm when the angle subtended at the centre is 1.20 rad.

Length of arc, $s = r\theta$, where θ is in radians, hence

$s = (5.5)(1.20) = \mathbf{6.60\,cm}$

Problem 10. Determine the diameter and circumference of a circle if an arc of length 4.75 cm subtends an angle of 0.91 rad.

Since $s = r\theta$ then $r = \dfrac{s}{\theta} = \dfrac{4.75}{0.91} = 5.22\,\text{cm}$

Diameter $= 2 \times$ radius $= 2 \times 5.22 = \mathbf{10.44\,cm}$

Circumference, $c = \pi d = \pi(10.44) = \mathbf{32.80\,cm}$

Problem 11. If an angle of 125° is subtended by an arc of a circle of radius 8.4 cm, find the length of (a) the minor arc, and (b) the major arc, correct to 3 significant figures.

(a) Since $180° = \pi$ rad then $1° = \left(\dfrac{\pi}{180}\right)$ rad and $125° = 125\left(\dfrac{\pi}{180}\right)$ rad.

Length of minor arc,

$$s = r\theta = (8.4)(125)\left(\frac{\pi}{180}\right) = \mathbf{18.3\,cm},$$

correct to 3 significant figures.

(b) Length of major arc

$= $ (circumference − minor arc)

$= 2\pi(8.4) - 18.3 = 34.5\,\text{cm}$,

correct to 3 significant figures.

(Alternatively, major arc $= r\theta$
$= 8.4(360 - 125)(\pi/180) = \mathbf{34.5\,cm}$.)

Problem 12. Determine the angle, in degrees and minutes, subtended at the centre of a circle of diameter 42 mm by an arc of length 36 mm. Calculate also the area of the minor sector formed.

Since length of arc, $s = r\theta$ then $\theta = s/r$

$$\text{Radius, } r = \frac{\text{diameter}}{2} = \frac{42}{2} = 21\,\text{mm}$$

$$\text{hence } \theta = \frac{s}{r} = \frac{36}{21} = 1.7143\,\text{rad}$$

$1.7143\,\text{rad} = 1.7143 \times (180/\pi)° = 98.22° = \mathbf{98°13'} =$ angle subtended at centre of circle.

Area of sector

$$= \tfrac{1}{2}r^2\theta = \tfrac{1}{2}(21)^2(1.7143) = \mathbf{378\,mm^2}$$

Problem 13. A football stadium floodlight can spread its illumination over an angle of 45° to a distance of 55 m. Determine the maximum area that is floodlit.

Floodlit area $=$ area of sector

$$= \frac{1}{2}r^2\theta = \frac{1}{2}(55)^2\left(45 \times \frac{\pi}{180}\right)$$

$$= \mathbf{1188\,m^2}$$

Problem 14. An automatic garden spray produces a spray to a distance of 1.8 m and revolves through an angle α which may be varied. If the desired spray catchment area is to be 2.5 m^2, to what should angle α be set, correct to the nearest degree.

Area of sector $= \tfrac{1}{2}r^2\theta$, hence $2.5 = \tfrac{1}{2}(1.8)^2\alpha$

from which, $\alpha = \dfrac{2.5 \times 2}{1.8^2} = 1.5432\,\text{rad}$

$1.5432\,\text{rad} = \left(1.5432 \times \dfrac{180°}{\pi}\right) = 88.42°$

Hence **angle $\alpha = 88°$**, correct to the nearest degree.

Problem 15. The angle of a tapered groove is checked using a 20 mm diameter roller as shown in Fig. 14.8. If the roller lies 2.12 mm below the top of the groove, determine the value of angle θ.

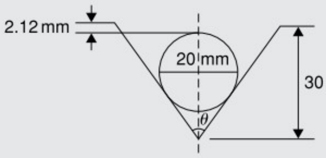

Figure 14.8

In Fig. 14.9, triangle ABC is right-angled at C (see Section 14.2 (vii)).

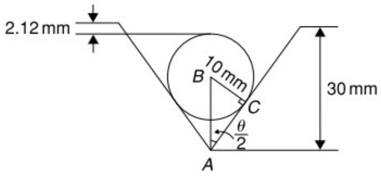

Figure 14.9

Length $BC = 10\,\text{mm}$ (i.e. the radius of the circle), and $AB = 30 - 10 - 2.12 = 17.88\,\text{mm}$ from Fig. 14.9.

Hence, $\sin\dfrac{\theta}{2} = \dfrac{10}{17.88}$ and $\dfrac{\theta}{2} = \sin^{-1}\left(\dfrac{10}{17.88}\right) = 34°$ and **angle $\theta = 68°$**

Now try the following Practice Exercise

Practice Exercise 64 Arc length and area of circles and sectors (Answers on page 865)

1. Calculate the area of a circle of radius 6.0 cm, correct to the nearest square centimetre.
2. The diameter of a circle is 55.0 mm. Determine its area, correct to the nearest square millimetre.
3. The perimeter of a circle is 150 mm. Find its area, correct to the nearest square millimetre.
4. Find the area of the sector, correct to the nearest square millimetre, of a circle having a radius of 35 mm, with angle subtended at centre of 75°
5. An annulus has an outside diameter of 49.0 mm and an inside diameter of 15.0 mm. Find its area correct to 4 significant figures.
6. Find the area, correct to the nearest square metre, of a 2 m wide path surrounding a circular plot of land 200 m in diameter.
7. A rectangular park measures 50 m by 40 m. A 3 m flower bed is made round the two longer sides and one short side. A circular fish pond of diameter 8.0 m is constructed in the centre of the park. It is planned to grass the remaining area. Find, correct to the nearest square metre, the area of grass.
8. Find the length of an arc of a circle of radius 8.32 cm when the angle subtended at the centre is 2.14 rad. Calculate also the area of the minor sector formed.
9. If the angle subtended at the centre of a circle of diameter 82 mm is 1.46 rad, find the lengths of the (a) minor arc (b) major arc.
10. A pendulum of length 1.5 m swings through an angle of 10° in a single swing. Find, in centimetres, the length of the arc traced by the pendulum bob.
11. Determine the length of the radius and circumference of a circle if an arc length of 32.6 cm subtends an angle of 3.76 rad.
12. Determine the angle of lap, in degrees and minutes, if 180 mm of a belt drive are in contact with a pulley of diameter 250 mm.
13. Determine the number of complete revolutions a motorcycle wheel will make in travelling 2 km, if the wheel's diameter is 85.1 cm.
14. The floodlight at a sports ground spreads its illumination over an angle of 40° to a distance of 48 m. Determine (a) the angle in radians, and (b) the maximum area that is floodlit.
15. Determine (a) the shaded area in Fig. 14.10 (b) the percentage of the whole sector that the area of the shaded portion represents.

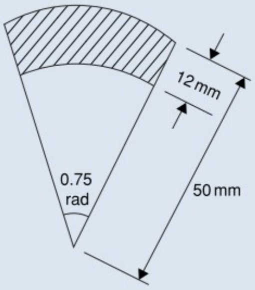

Figure 14.10

16. Determine the length of steel strip required to make the clip shown in Fig. 14.11.

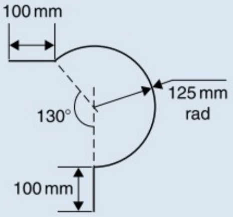

Figure 14.11

17. A 50° tapered hole is checked with a 40 mm diameter ball as shown in Fig. 14.12. Determine the length shown as x.

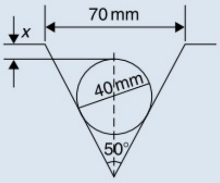

Figure 14.12

14.5 The equation of a circle

The simplest equation of a circle, centre at the origin, radius r, is given by:

$$x^2 + y^2 = r^2$$

For example, Fig. 14.13 shows a circle $x^2 + y^2 = 9$. More generally, the equation of a circle, centre (a, b), radius r, is given by:

$$(x - a)^2 + (y - b)^2 = r^2 \qquad (1)$$

Figure 14.14 shows a circle $(x - 2)^2 + (y - 3)^2 = 4$. The general equation of a circle is:

$$x^2 + y^2 + 2ex + 2fy + c = 0 \qquad (2)$$

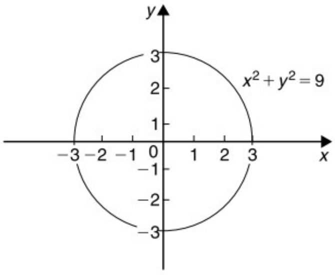

Figure 14.13

Multiplying out the bracketed terms in equation (1) gives:

$$x^2 - 2ax + a^2 + y^2 - 2by + b^2 = r^2$$

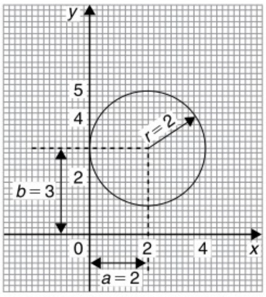

Figure 14.14

Comparing this with equation (2) gives:

$$2e = -2a, \text{ i.e. } \boldsymbol{a = -\frac{2e}{2}}$$

and $2f = -2b$, i.e. $\boldsymbol{b = -\frac{2f}{2}}$

and $c = a^2 + b^2 - r^2$,

i.e., $\boldsymbol{r = \sqrt{(a^2 + b^2 - c)}}$

Thus, for example, the equation

$$x^2 + y^2 - 4x - 6y + 9 = 0$$

represents a circle with centre $a = -\left(\frac{-4}{2}\right)$, $b = -\left(\frac{-6}{2}\right)$, i.e. at (2, 3) and radius $r = \sqrt{(2^2 + 3^2 - 9)} = 2$

Hence $x^2 + y^2 - 4x - 6y + 9 = 0$ is the circle shown in Fig. 14.14 (which may be checked by multiplying out the brackets in the equation

$$(x - 2)^2 + (y - 3)^2 = 4)$$

Problem 16. Determine (a) the radius and (b) the co-ordinates of the centre of the circle given by the equation: $x^2 + y^2 + 8x - 2y + 8 = 0$

$x^2 + y^2 + 8x - 2y + 8 = 0$ is of the form shown in equation (2),

where $a = -\left(\frac{8}{2}\right) = -4, b = -\left(\frac{-2}{2}\right) = 1$

and $r = \sqrt{[(-4)^2 + (1)^2 - 8]} = \sqrt{9} = 3$

Hence $x^2 + y^2 + 8x - 2y + 8 = 0$ represents a circle **centre (−4, 1)** and **radius 3**, as shown in Fig. 14.15.

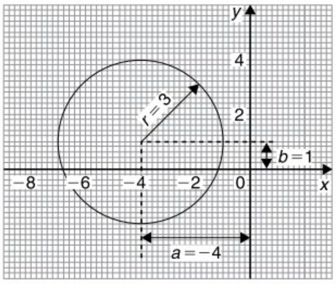

Figure 14.15

Alternatively, $x^2+y^2+8x-2y+8=0$ may be rearranged as:

$$(x+4)^2+(y-1)^2-9=0$$

i.e. $$(x+4)^2+(y-1)^2=3^2$$

which represents a circle, **centre (−4, 1)** and **radius 3**, as stated above.

Problem 17. Sketch the circle given by the equation: $x^2+y^2-4x+6y-3=0$

The equation of a circle, centre (a, b), radius r is given by:

$$(x-a)^2+(y-b)^2=r^2$$

The general equation of a circle is

$$x^2+y^2+2ex+2fy+c=0$$

From above, $a=-\frac{2e}{2}$, $b=-\frac{2f}{2}$ and

$$r=\sqrt{(a^2+b^2-c)}$$

Hence if $x^2+y^2-4x+6y-3=0$

then $a=-\left(\frac{-4}{2}\right)=2,\quad b=-\left(\frac{6}{2}\right)=-3$

and $$r=\sqrt{[(2)^2+(-3)^2-(-3)]}$$
$$=\sqrt{16}=4$$

Thus **the circle has centre (2, −3) and radius 4**, as shown in Fig. 14.16.

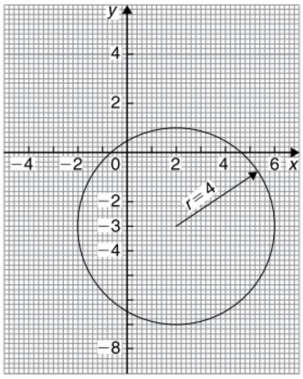

Figure 14.16

Alternatively, $x^2+y^2-4x+6y-3=0$ may be rearranged as:

$$(x-2)^2+(y+3)^2-3-13=0$$

i.e. $$(x-2)^2+(y+3)^2=4^2$$

which represents a circle, **centre (2, −3)** and **radius 4**, as stated above.

Now try the following Practice Exercise

Practice Exercise 65 The equation of a circle (Answers on page 865)

1. Determine (a) the radius, and (b) the co-ordinates of the centre of the circle given by the equation $x^2+y^2+6x-2y-26=0$
2. Sketch the circle given by the equation $x^2+y^2-6x+4y-3=0$
3. Sketch the curve $x^2+(y-1)^2-25=0$
4. Sketch the curve $x=6\sqrt{[1-(y/6)^2]}$

14.6 Linear and angular velocity

Linear velocity

Linear velocity v is defined as the rate of change of linear displacement s with respect to time t. For motion in a straight line:

$$\text{linear velocity}=\frac{\text{change of displacement}}{\text{change of time}}$$

i.e. $$v=\frac{s}{t} \qquad (1)$$

The unit of linear velocity is metres per second (m/s).

Angular velocity

The speed of revolution of a wheel or a shaft is usually measured in revolutions per minute or revolutions per second but these units do not form part of a coherent system of units. The basis in SI units is the angle turned through in one second.

Angular velocity is defined as the rate of change of angular displacement θ, with respect to time t. For an object rotating about a fixed axis at a constant speed:

$$\text{angular velocity}=\frac{\text{angle turned through}}{\text{time taken}}$$

i.e. $$\boldsymbol{\omega = \frac{\theta}{t}} \qquad (2)$$

The unit of angular velocity is radians per second (rad/s). An object rotating at a constant speed of n revolutions per second subtends an angle of $2\pi n$ radians in one second, i.e. its angular velocity ω is given by:

$$\boldsymbol{\omega = 2\pi n\,\text{rad/s}} \qquad (3)$$

From page 152, $s = r\theta$ and from equation (2) above, $\theta = \omega t$

hence $$s = r(\omega t)$$

from which $$\frac{s}{t} = \omega r$$

However, from equation (1) $v = \frac{s}{t}$

hence $$\boldsymbol{v = \omega r} \qquad (4)$$

Equation (4) gives the relationship between linear velocity v and angular velocity ω.

Problem 18. A wheel of diameter 540 mm is rotating at $\frac{1500}{\pi}$ rev/min. Calculate the angular velocity of the wheel and the linear velocity of a point on the rim of the wheel.

From equation (3), angular velocity $\omega = 2\pi n$ where n is the speed of revolution in rev/s. Since in this case $n = \frac{1500}{\pi}\,\text{rev/min} = \frac{1500}{60\pi} = \text{rev/s}$, then

$$\textbf{angular velocity}\,\omega = 2\pi\left(\frac{1500}{60\pi}\right) = \mathbf{50\,rad/s}$$

The linear velocity of a point on the rim, $v = \omega r$, where r is the radius of the wheel, i.e.

$$\frac{540}{2}\,\text{mm} = \frac{0.54}{2}\,\text{m} = 0.27\,\text{m}$$

Thus **linear velocity** $v = \omega r = (50)(0.27)$

$$= \mathbf{13.5\,m/s}$$

Problem 19. A car is travelling at 64.8 km/h and has wheels of diameter 600 mm.

(a) Find the angular velocity of the wheels in both rad/s and rev/min.

(b) If the speed remains constant for 1.44 km, determine the number of revolutions made by the wheel, assuming no slipping occurs.

(a) Linear velocity $v = 64.8\,\text{km/h}$

$$= 64.8\frac{\text{km}}{\text{h}} \times 1000\frac{\text{m}}{\text{km}} \times \frac{1}{3600}\frac{\text{h}}{\text{s}} = 18\,\text{m/s}$$

The radius of a wheel $= \frac{600}{2} = 300\,\text{mm}$

$= 0.3\,\text{m}$

From equation (5), $v = \omega r$, from which,

$$\textbf{angular velocity}\,\omega = \frac{v}{r} = \frac{18}{0.3} = \mathbf{60\,rad/s}$$

From equation (4), angular velocity, $\omega = 2\pi n$, where n is in rev/s.

Hence angular speed $n = \frac{\omega}{2\pi} = \frac{60}{2\pi}\,\text{rev/s}$

$$= 60 \times \frac{60}{2\pi}\,\text{rev/min}$$

$$= \mathbf{573\,rev/min}$$

(b) From equation (1), since $v = s/t$ then the time taken to travel 1.44 km, i.e. 1440 m at a constant speed of 18 m/s is given by:

$$\text{time } t = \frac{s}{v} = \frac{1440\,\text{m}}{18\,\text{m/s}} = 80\,\text{s}$$

Since a wheel is rotating at 573 rev/min, then in 80/60 minutes it makes

$$573\,\text{rev/min} \times \frac{80}{60}\,\text{min} = \mathbf{764\,revolutions}$$

Now try the following Practice Exercise

Practice Exercise 66 Linear and angular velocity (Answers on page 865)

1. A pulley driving a belt has a diameter of 300 mm and is turning at $2700/\pi$ revolutions per minute. Find the angular velocity of the pulley and the linear velocity of the belt assuming that no slip occurs.
2. A bicycle is travelling at 36 km/h and the diameter of the wheels of the bicycle is 500 mm. Determine the linear velocity of a point on the rim of one of the wheels of the bicycle, and the angular velocity of the wheels.

3. A train is travelling at 108 km/h and has wheels of diameter 800 mm.
   (a) Determine the angular velocity of the wheels in both rad/s and rev/min.
   (b) If the speed remains constant for 2.70 km, determine the number of revolutions made by a wheel, assuming no slipping occurs.

14.7 Centripetal force

When an object moves in a circular path at constant speed, its direction of motion is continually changing and hence its velocity (which depends on both magnitude and direction) is also continually changing. Since acceleration is the (change in velocity)/(time taken), the object has an acceleration. Let the object be moving with a constant angular velocity of ω and a tangential velocity of magnitude v and let the change of velocity for a small change of angle of θ $(=\omega t)$ be V in Fig. 14.17. Then $v_2 - v_1 = V$. The vector diagram is shown in Fig. 14.17(b) and since the magnitudes of v_1 and v_2 are the same, i.e. v, the vector diagram is an isosceles triangle.

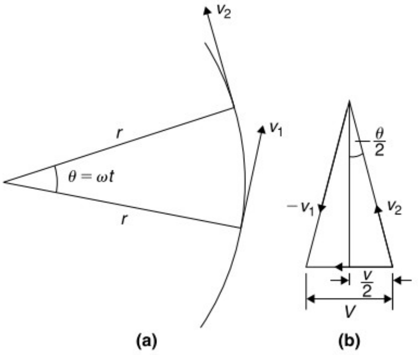

Figure 14.17

Bisecting the angle between v_2 and v_1 gives:

$$\sin\frac{\theta}{2} = \frac{V/2}{v_2} = \frac{V}{2v}$$

i.e. $$V = 2v\sin\frac{\theta}{2} \qquad (1)$$

Since $\theta = \omega t$ then

$$t = \frac{\theta}{\omega} \qquad (2)$$

Dividing equation (1) by equation (2) gives:

$$\frac{V}{t} = \frac{2v\sin(\theta/2)}{(\theta/\omega)} = \frac{v\omega\sin(\theta/2)}{(\theta/2)}$$

For small angles $\frac{\sin(\theta/2)}{(\theta/2)} \approx 1$,

hence $$\frac{V}{t} = \frac{\text{change of velocity}}{\text{change of time}}$$

$$= \text{acceleration } a = v\omega$$

However, $\omega = \frac{v}{r}$ (from Section 14.6)

thus $$v\omega = v\cdot\frac{v}{r} = \frac{v^2}{r}$$

i.e. the acceleration a is $\frac{v^2}{r}$ and is towards the centre of the circle of motion (along V). It is called the **centripetal acceleration**. If the mass of the rotating object is m, then by Newton's second law, the **centripetal force is** $\boldsymbol{\frac{mv^2}{r}}$ and its direction is towards the centre of the circle of motion.

Problem 20. A vehicle of mass 750 kg travels around a bend of radius 150 m, at 50.4 km/h. Determine the centripetal force acting on the vehicle.

The centripetal force is given by $\frac{mv^2}{r}$ and its direction is towards the centre of the circle.

$$\text{Mass } m = 750\,\text{kg},\ v = 50.4\,\text{km/h}$$
$$= \frac{50.4 \times 1000}{60 \times 60}\,\text{m/s}$$
$$= 14\,\text{m/s}$$

and radius $r = 150\,\text{m}$,

thus **centripetal force** $= \frac{750(14)^2}{150} = \mathbf{980\,N}$

Problem 21. An object is suspended by a thread 250 mm long and both object and thread move in a horizontal circle with a constant angular velocity of 2.0 rad/s. If the tension in the thread is 12.5 N, determine the mass of the object.

Centripetal force (i.e. tension in thread),

$$F = \frac{mv^2}{r} = 12.5\,\text{N}$$

Angular velocity $\omega = 2.0$ rad/s and
radius $r = 250$ mm $= 0.25$ m.

Since linear velocity $v = \omega r$, $v = (2.0)(0.25)$
$= 0.5$ m/s.

Since $F = \dfrac{mv^2}{r}$, then mass $m = \dfrac{Fr}{v^2}$,

i.e. **mass of object,** $\boldsymbol{m} = \dfrac{(12.5)(0.25)}{0.5^2} = \mathbf{12.5\ kg}$

Problem 22. An aircraft is turning at constant altitude, the turn following the arc of a circle of radius 1.5 km. If the maximum allowable acceleration of the aircraft is $2.5g$, determine the maximum speed of the turn in km/h. Take g as $9.8\,\text{m/s}^2$.

The acceleration of an object turning in a circle is $\dfrac{v^2}{r}$. Thus, to determine the maximum speed of turn, $\dfrac{v^2}{r} = 2.5g$, from which,

$$\textbf{velocity, } \boldsymbol{v} = \sqrt{(2.5gr)} = \sqrt{(2.5)(9.8)(1500)}$$
$$= \sqrt{36\,750} = 191.7\,\text{m/s}$$

and $191.7\,\text{m/s} = 191.7 \times \dfrac{60 \times 60}{1000}\,\text{km/h} = \mathbf{690\,km/h}$

Now try the following Practice Exercise

Practice Exercise 67 Centripetal force (Answers on page 865)

1. Calculate the tension in a string when it is used to whirl a stone of mass 200 g round in a horizontal circle of radius 90 cm with a constant speed of 3 m/s.
2. Calculate the centripetal force acting on a vehicle of mass 1 tonne when travelling around a bend of radius 125 m at 40 km/h. If this force should not exceed 750 N, determine the reduction in speed of the vehicle to meet this requirement.
3. A speed-boat negotiates an S-bend consisting of two circular arcs of radii 100 m and 150 m. If the speed of the boat is constant at 34 km/h, determine the change in acceleration when leaving one arc and entering the other.

For fully worked solutions to each of the problems in Practice Exercises 62 to 67 in this chapter, go to the website:
www.routledge.com/cw/bird

Revision Test 4 Trigonometry, Cartesian and polar co-ordinates and properties of circles

This Revision Test covers the material contained in Chapters 12 to 14. *The marks for each question are shown in brackets at the end of each question.*

1. A 2.0 m long ladder is placed against a perpendicular pylon with its foot 52 cm from the pylon. (a) Find how far up the pylon (correct to the nearest mm) the ladder reaches. (b) If the foot of the ladder is moved 10 cm towards the pylon, how far does the top of the ladder rise? (7)

2. Evaluate correct to 4 significant figures:

   (a) $\cos 124°13'$ (b) $\cot 72.68°$ (4)

3. From a point on horizontal ground a surveyor measures the angle of elevation of a church spire as 15°. He moves 30 m nearer to the church and measures the angle of elevation as 20°. Calculate the height of the spire. (9)

4. If secant $\theta = 2.4613$, determine the acute angle θ (4)

5. Evaluate, correct to 3 significant figures:

$$\frac{3.5 \operatorname{cosec} 31°17' - \cot(-12°)}{3 \sec 79°41'} \qquad (5)$$

6. A man leaves a point walking at 6.5 km/h in a direction E 20° N (i.e. a bearing of 70°). A cyclist leaves the same point at the same time in a direction E 40° S (i.e. a bearing of 130°) travelling at a constant speed. Find the average speed of the cyclist if the walker and cyclist are 80 km apart after five hours. (8)

7. A crank mechanism shown in Fig. RT4.1 comprises arm *OP*, of length 0.90 m, which rotates anticlockwise about the fixed point *O*, and connecting rod *PQ* of length 4.20 m. End *Q* moves horizontally in a straight line *OR*.

   (a) If $\angle POR$ is initially zero, how far does end *Q* travel in $\frac{1}{4}$ revolution.

   (b) If $\angle POR$ is initially 40° find the angle between the connecting rod and the horizontal and the length *OQ*.

   (c) Find the distance *Q* moves (correct to the nearest cm) when $\angle POR$ changes from 40° to 140°. (16)

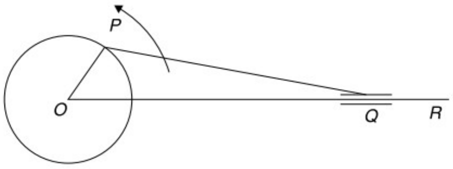

Figure RT4.1

8. Change the following Cartesian co-ordinates into polar co-ordinates, correct to 2 decimal places, in both degrees and in radians:

   (a) (−2.3, 5.4) (b) (7.6, −9.2) (10)

9. Change the following polar co-ordinates into Cartesian co-ordinates, correct to 3 decimal places: (a) (6.5, 132°) (b) (3, 3 rad) (6)

10. (a) Convert 2.154 radians into degrees and minutes.

    (b) Change 71°17′ into radians. (4)

11. 140 mm of a belt drive is in contact with a pulley of diameter 180 mm which is turning at 300 revolutions per minute. Determine (a) the angle of lap, (b) the angular velocity of the pulley, and (c) the linear velocity of the belt, assuming that no slipping occurs. (9)

12. Figure RT4.2 shows a cross-section through a circular water container where the shaded area represents the water in the container. Determine: (a) the depth, *h*, (b) the area of the shaded portion, and (c) the area of the unshaded area. (11)

29 04 2020

007063419X.pdf

Pág. 179 Chapter 5

Curve Tracing

Pág. 1228 Chapter 36

Analytical Solid Geometry

Obs: Livro todo!!!

D) Books_74_2020/
Higher Engineering Mathematics by
Bandary Ramana/

1138673579.pdf

Pág. 140 Section B

Geometry and trigonometry

Pág. 270 Section D

Complex numbers

Obs. Livro todo!!!

E) Books_74_2020/
Electric Field Analysis/

Electric Field Analysis.pdf

Pág. 61(84)

27.04.2020

Pág. 39 (62) Chapter 2
Gauss's Law and Related Topics
Pág. 59 (82) Chapter 3
Orthogonal Coordinate Systems
Pág. 219 (242) Chapter 10
Sphere or Cylinder in Uniform External Field.

F) Booty_74_2020/
La Matematica Vol. 1. I tempi e i luoghi/

La matematica....pdf

Pág. 189 (205)

5.2 La trigonometria

Teorema di Menelao.

Se CE, CF, DF, DB sono archi di cerchi massimi su una sfera.
(círculos)

Pág. 190 Teorema del seno.

Pág. 402 (418) Leibniz, consigliere dei principi

)*+,)))-) C 0 4 "* / 6"
' C D%

3.1.2 Right-Handed Convention

Because three constant coordinate surfaces are involved in determining the coordinates of any point, the question that arises is in which order the three coordinates are to be mentioned while specifying a point. In this context, the choice of the first coordinate (u_1) or surface (f_1) is arbitrary. Once f_1 is chosen, then the order in which the other two constant coordinate surfaces f_2 and f_3 is to be taken is determined by the right-handed system as governed by the definition of vector cross product. In other words, the cross product of the first and second unit vectors must be equal to the third unit vector.

This right-handed property of the coordinate system is an arbitrary convention. But once that choice is made for one coordinate system, it has to be followed for all other coordinate systems to ensure that the expressions for the laws of electric field have the same form in all the coordinate systems.

3.1.3 Differential Distance and Metric Coefficient

The constant coordinate surfaces could be of many different types. The simplest of them is constant distance surfaces for which the differential change du_i in the coordinate u_i is the same as the differential distance dl_i between the surfaces $f_i = u_i$ and $f_i = u_i + du_i$. However, there are other types of surfaces such as constant angle surfaces, in which case the differential distance is related to differential coordinate (angle) change as

$$dl_i = h_i du_i = h_i d\theta_i \tag{3.2}$$

where:
The scale factor h_i is the corresponding radius

The scale factors for each of the coordinate system are called the *metric coefficients* of the respective coordinate system. The three metric functions of any coordinate system (h_1,h_2,h_3) constitute a unique set and hence are often known as the *signature of a coordinate system*. Metric coefficients are also very important in writing the vector operators in the general orthogonal curvilinear coordinate system, as will be explained in Section 3.5.

3.1.4 Choice of Origin

A basic element of any coordinate system is the choice of origin. If any object or any arrangement is given, then any point within the object or within the whole arrangement could be chosen as origin.

From the earlier discussions, it is clear that the same point will have very different coordinates when defined in different coordinate systems. Therefore, the question that needs to be answered is that whether there is any property of the position of the point that remains the same in different coordinate systems. The answer to this question is that the distance of the point from the origin remains the same in all the coordinate systems if the origin is kept the same in all the cases. As a logical extension, it may be stated that the distance between any two points remains the same no matter how their coordinates are defined in any coordinate system.

3.2 Cartesian Coordinate System

The invention of Cartesian coordinates in the seventeenth century by the French mathematician René Descartes revolutionized mathematics by providing the first systematic link between Euclidean geometry and algebra. In this coordinate system, the three constant coordinate surfaces are defined by Equation 3.3

$$f_1 = x, f_2 = y, f_3 = z \tag{3.3}$$

Figure 3.2 shows such constant coordinate surfaces, which are three planes. Cartesian coordinate system has the unique feature that all the three constant coordinate surfaces are constant distance surfaces.

Each point P in space is then assigned a triplet of values (x,y,z), the so-called Cartesian coordinates of that point. The ranges of the values of three coordinates are $-\infty < x < \infty$, $-\infty < y < \infty$ and $-\infty < z < \infty$. Coordinates can be

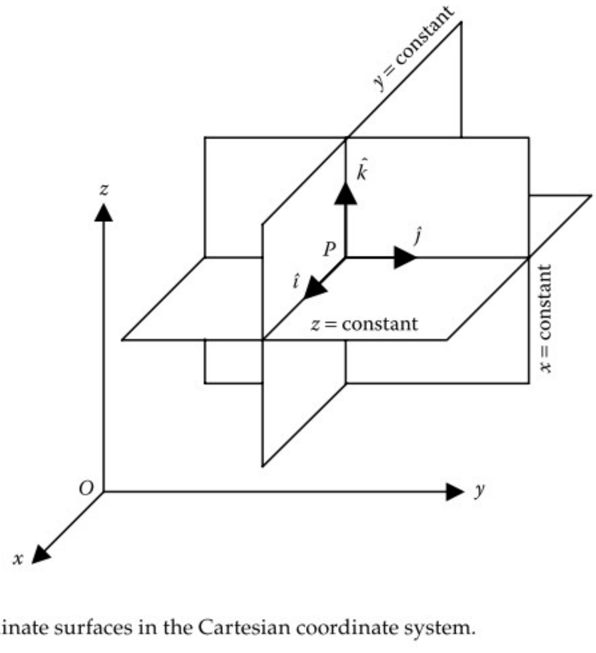

FIGURE 3.2
Constant coordinate surfaces in the Cartesian coordinate system.

defined as the positions of the normal projections of the point onto three mutually perpendicular constant coordinate surfaces passing through the origin, expressed as signed distances from the origin.

In the Cartesian coordinate system, three mutually perpendicular axes are commonly defined as follows. The intersection of two mutually perpendicular planes y = constant and z = constant passing through the origin gives a coordinate curve, which is a straight line. This line is then defined as the x-axis. Similarly, the intersections of other two pairs of mutually perpendicular constant coordinate planes passing through origin define the y- and z-axes.

The unit vector $\hat{i}$ is perpendicular to all the surfaces described by x = constant and points in the direction of increasing value of x-coordinate. In the same manner, unit vectors $\hat{j}$ and $\hat{k}$ are defined in y- and z-directions, respectively. Another unique feature of the Cartesian coordinate system is that the directions of the three unit vectors remain the same irrespective of the location of the point in space. The orthogonality of the Cartesian coordinate system is defined by Equation 3.4

$$\hat{i}\cdot\hat{j}=0,\ \hat{j}\cdot\hat{k}=0,\ \hat{k}\cdot\hat{i}=0 \tag{3.4}$$

and the right handedness is defined by Equation 3.5

$$\hat{i}\times\hat{j}=\hat{k},\ \hat{j}\times\hat{k}=\hat{i},\ \hat{k}\times\hat{i}=\hat{j} \tag{3.5}$$

As shown in Figure 3.3, the differential line element in the Cartesian coordinate system is given by Equation 3.6

$$d\vec{l}=dl_1\hat{i}+dl_2\hat{j}+dl_3\hat{k}=dx\hat{i}+dy\hat{j}+dz\bar{k} \tag{3.6}$$

As shown in Figure 3.4, the differential area element corresponding to the surface of a differential cubical volume is given by Equation 3.7

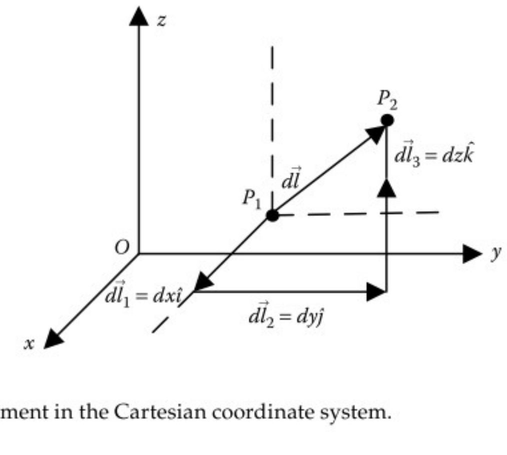

FIGURE 3.3
Differential line element in the Cartesian coordinate system.

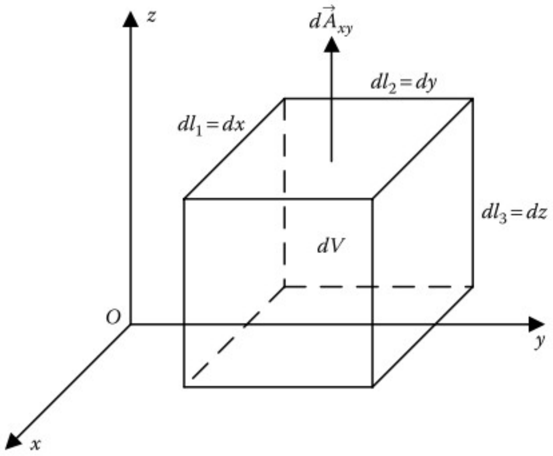

FIGURE 3.4
Differential area and volume elements in the Cartesian coordinate system.

$$d\vec{A}_{xy} = dl_1 dl_2 \hat{k} = dxdy\hat{k} \tag{3.7}$$

As shown in Figure 3.4, the differential volume element is given by Equation 3.8

$$dV = dl_1 dl_2 dl_3 = dxdydz \tag{3.8}$$

As depicted in Equation 3.6 the differential distance in x-direction is $dl_1 = dx = h_1 dx$, that is, $h_1 = 1$. Similarly, in y- and z-directions, $h_2 = h_3 = 1$. The metric coefficients of the Cartesian coordinate system are, therefore, $h_1 = 1$, $h_2 = 1$ and $h_3 = 1$.

3.3 Cylindrical Coordinate System

Cylindrical coordinates are an alternate way of describing points in a three-dimensional space. In this coordinate system, one of the rectangular coordinate planes, namely, the x–y plane, as described by the Cartesian coordinate system, is replaced by a polar plane. In the cylindrical coordinate system, everything is measured with respect to a fixed point called the *pole* and an axis called the *polar axis*. The pole is the equivalent to the origin in the Cartesian coordinate system and the polar axis corresponds to the positive direction of the x-axis. The cylindrical coordinates of a point are then the ordered triplet (r,θ,z), as defined in Figure 3.5.

As shown in Figure 3.5, r is the distance from the pole to the projection of the point P on the polar plane, that is, the x–y plane passing through the pole, θ is the azimuthal angle, that is, the angle from the polar axis spinning around the z-axis in counter-clockwise direction, and z is the vertical height from the polar plane. The ranges of the values of the three coordinates are $0 \le r < \infty$, $0 \le \theta \le 2\pi$ and $-\infty < z < \infty$.

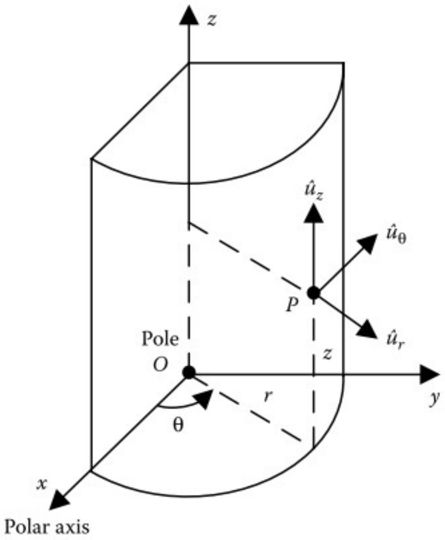

FIGURE 3.5
Depiction of cylindrical coordinates of a point.

In the cylindrical coordinate system, the three constant coordinate surfaces are defined by Equation 3.9

$$f_1 = r, f_2 = \theta, f_3 = z \tag{3.9}$$

Figure 3.6 shows the three constant coordinate surfaces in the cylindrical coordinate system. Out of these three surfaces, the first and the third surfaces, namely, $f_1 = r$ and $f_3 = z$, are constant distance surfaces, whereas the second one, that is, $f_2 = \theta$, is a constant angle surface. As shown in Figure 3.6, the surfaces θ = constant and z = constant are planes, whereas the surface r = constant is a cylindrical surface.

In this coordinate system, two unit vectors are defined on the x–y plane. The unit vector $\hat{u}_r$ points in the direction of increasing r, that is, radially outwards from the z-axis and the unit vector $\hat{u}_\theta$ points in the direction of increasing θ, that is, it points in the direction of the tangent to the circle of radius r in the counter-clockwise sense. The third unit vector $\hat{u}_z$ points in the direction of increasing z, that is, vertically upwards from the x–y plane. The unit vectors are shown in Figure 3.5. The orthogonality of cylindrical coordinate system is defined by Equation 3.10

$$\hat{u}_r \cdot \hat{u}_\theta = 0, \hat{u}_\theta \cdot \hat{u}_z = 0, \hat{u}_z \cdot \hat{u}_r = 0 \tag{3.10}$$

and the right handedness is defined by Equation 3.11

$$\hat{u}_r \times \hat{u}_\theta = \hat{u}_z, \hat{u}_\theta \times \hat{u}_z = \hat{u}_r, \hat{u}_z \times \hat{u}_r = \hat{u}_\theta \tag{3.11}$$

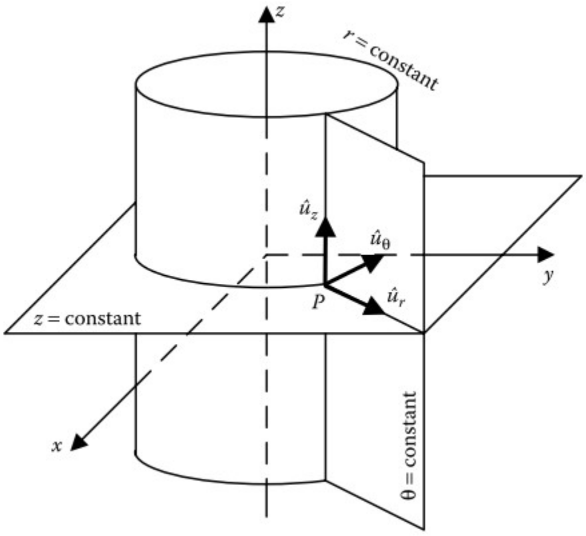
z
r = constant
$\hat{u}_z$
$\hat{u}_\theta$
y
P
$\hat{u}_r$
z = constant
x
θ = constant

FIGURE 3.6
Constant coordinate surfaces in the cylindrical coordinate system.

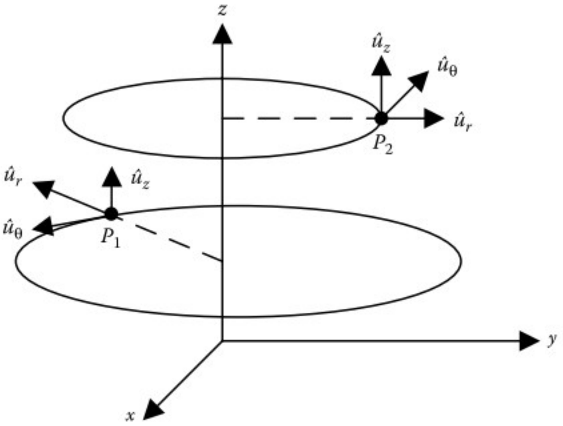
z
$\hat{u}_z$
$\hat{u}_\theta$
$\hat{u}_r$
P_2
$\hat{u}_r$
$\hat{u}_z$
$\hat{u}_\theta$
P_1
y
x

FIGURE 3.7
Unit vectors at different points in the cylindrical coordinate system.

Unlike the Cartesian system, in the cylindrical coordinate system, the directions of all the three unit vectors do not remain the same as one moves around the space. As shown in Figure 3.7, the directions of the unit vectors $\hat{u}_r$ and $\hat{u}_\theta$ get changed at different points in space keeping $\hat{u}_z$ unchanged.

As shown in Figure 3.8, the differential line element in the cylindrical coordinate system is given by Equation 3.12

$$d\vec{l} = dl_1\hat{u}_r + dl_2\hat{u}_\theta + dl_3\hat{u}_z = dr\hat{u}_r + rd_\theta\hat{u}_\theta + dz\hat{u}_z \tag{3.12}$$

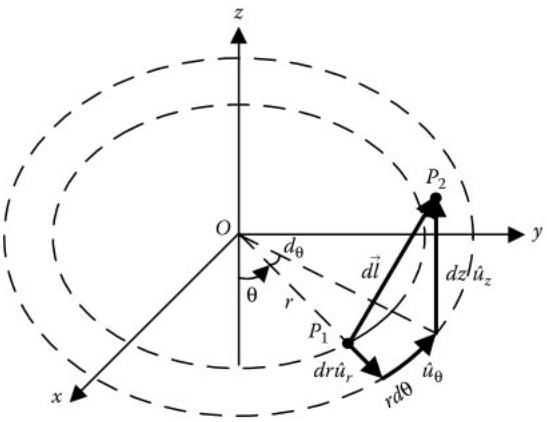
z
y
x
O
P_2
P_1
d_θ
θ
r
$d\vec{l}$
$dz\,\hat{u}_z$
$dr\hat{u}_r$
$rd\theta$
$\hat{u}_\theta$

FIGURE 3.8
Differential line element in the cylindrical coordinate system.

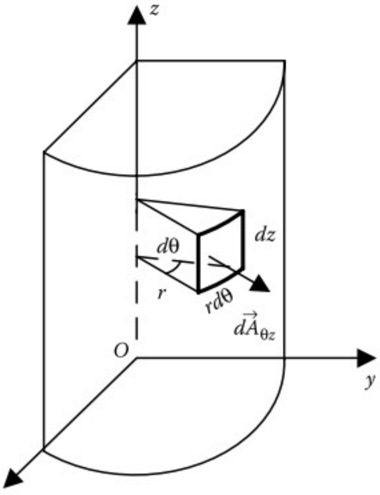
z
dz
$d\theta$
r
$rd\theta$
$d\vec{A}_{\theta z}$
O
y
x

FIGURE 3.9
Differential area element on the cylinder surface in the cylindrical coordinate system.

As shown in Figure 3.9, in the cylindrical coordinate system, the differential area element on the surface of a cylinder of radius r is given by Equation 3.13

$$d\vec{A}_{\theta z} = dl_2 dl_3 \hat{u}_r = rd\theta dz\hat{u}_r \tag{3.13}$$

and the differential area element on the surface of a disc in the x–y plane, as shown in Figure 3.10, is given by Equation 3.14

$$d\vec{A}_{r\theta} = dl_1 dl_2 \hat{u}_z = drrd\theta\hat{u}_z \tag{3.14}$$

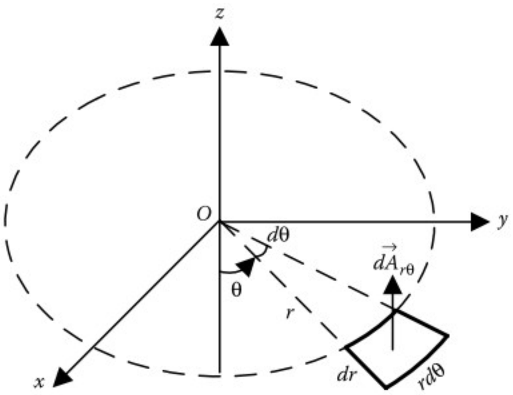

FIGURE 3.10
Differential area element on the disc surface in the cylindrical coordinate system.

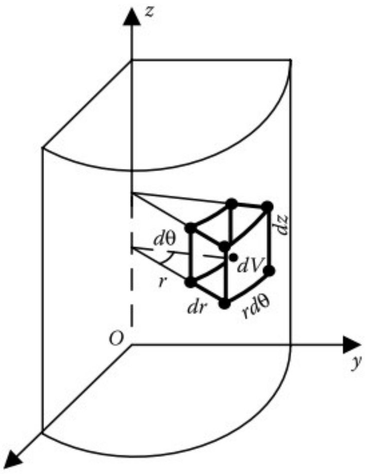

FIGURE 3.11
Differential volume element of a cylinder in the cylindrical coordinate system.

As shown in Figure 3.11, the differential volume element is given by Equation 3.15

$$dV = dl_1 dl_2 dl_3 = dr r d\theta dz \tag{3.15}$$

As depicted in Equation 3.12, the differential distance in r-direction is $dl_1 = dr = h_1 dr$, that is, $h_1 = 1$, in θ-direction, it is $dl_2 = r d\theta = h_2 d\theta$, that is, $h_2 = r$ and in z-direction, it is $dl_3 = dz = h_3 dz$, that is, $h_3 = 1$. The metric coefficients of cylindrical coordinate system are, therefore, $h_1 = 1$, $h_2 = r$ and $h_3 = 1$.

3.4 Spherical Coordinate System

Spherical coordinates are particularly useful for analyzing fields having spherical symmetry. In the spherical coordinate system, the coordinates of any point in space are the ordered triplet (r,θ,ϕ), as shown in Figure 3.12. The coordinate r measures the radial distance from the origin to the point P. The coordinate θ is the angle that the r vector makes with the positive direction of the z-axis, whereas the coordinate ϕ is the azimuthal angle with respect to the positive direction of the x-axis spinning around the z-axis in counterclockwise sense. The angle ϕ is defined on the x–y plane. The ranges of the values of the three coordinates are $0 \le r < \infty$, $0 \le \theta \le \pi$ and $0 \le \phi \le 2\pi$.

In the spherical coordinate system, the three constant coordinate surfaces are defined by Equation 3.16

$$f_1 = r,\ f_2 = \theta,\ f_3 = \phi \tag{3.16}$$

Figure 3.13 shows the three constant coordinate surfaces in the spherical coordinate system. Out of these three surfaces, the first surface, namely, $f_1 = r$, is a constant distance surface, whereas the other two surfaces, that is, $f_2 = \theta$ and $f_3 = \phi$, are constant angle surfaces. As shown in Figure 3.13, r = constant is a spherical surface, θ = constant is a conical surface and ϕ = constant is a plane.

In this coordinate system, the unit vector $\hat{u}_r$ points in the direction of increasing r, that is, radially outwards from the origin. The unit vector $\hat{u}_\theta$ points in the direction of increasing θ along the tangent to a circle of radius r in the plane containing the z-axis and the r vector. The unit vector $\hat{u}_\phi$ points

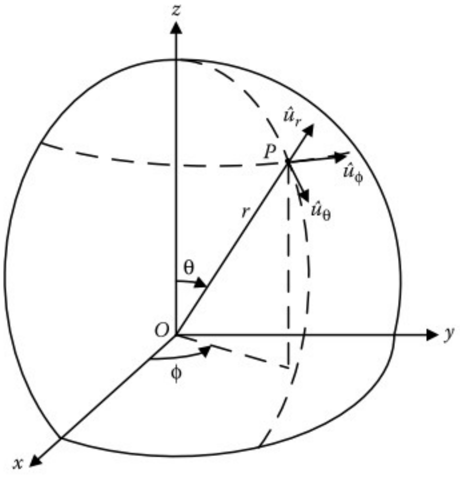

FIGURE 3.12
Depiction of spherical coordinates of a point.

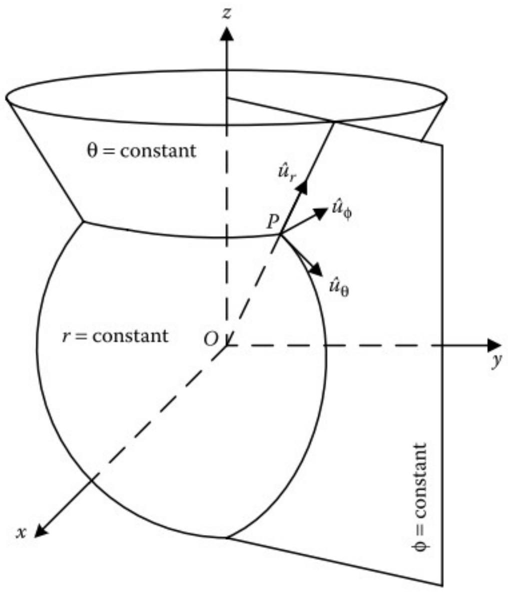

FIGURE 3.13
Constant coordinate surfaces in the spherical coordinate system.

in the direction of increasing ϕ along the tangent to a circle in the x–y plane, which is centred on the z-axis. The unit vectors are shown in Figure 3.12. The orthogonality of the spherical coordinate system is defined by Equation 3.17

$$\hat{u}_r \cdot \hat{u}_\theta = 0,\ \hat{u}_\theta \cdot \hat{u}_\phi = 0,\ \hat{u}_\phi \cdot \hat{u}_r = 0 \tag{3.17}$$

and the right handedness is defined by Equation 3.18

$$\hat{u}_r \times \hat{u}_\theta = \hat{u}_\phi,\ \hat{u}_\theta \times \hat{u}_\phi = \hat{u}_r,\ \hat{u}_\phi \times \hat{u}_r = \hat{u}_\theta \tag{3.18}$$

Similar to the cylindrical coordinate system, in the spherical coordinate system, the directions of all the three unit vectors do not remain the same as one move around the space. As shown in Figure 3.14, the directions of all the three unit vectors $\hat{u}_r$, $\hat{u}_\theta$ and $\hat{u}_\phi$ get changed at different points in space.

As shown in Figure 3.15, the differential line element in the spherical coordinate system is given by Equation 3.19

$$d\vec{l} = dl_1\hat{u}_r + dl_2\hat{u}_\theta + dl_3\hat{u}_\phi = dr\hat{u}_r + rd\theta\hat{u}_\theta + r\sin\theta d\phi\hat{u}_\phi \tag{3.19}$$

As shown in Figure 3.16, in the spherical coordinate system, the differential area element on the surface of a sphere of radius r is given by Equation 3.20

$$d\vec{A}_{\theta\phi} = dl_2dl_3\hat{u}_r = (rd\theta)(r\sin\theta d\phi)\hat{u}_r = r^2\sin\theta d\theta d\phi\hat{u}_r \tag{3.20}$$

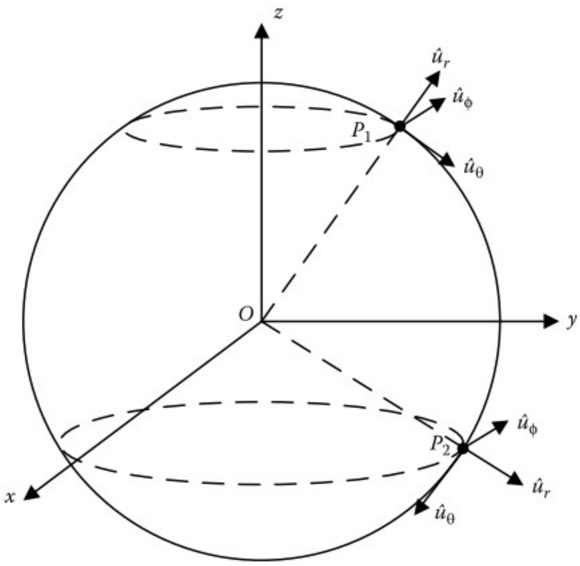
z
$\hat{u}_r$
$\hat{u}_\phi$
P_1
$\hat{u}_\theta$
O
y
$\hat{u}_\phi$
P_2
$\hat{u}_r$
x
$\hat{u}_\theta$

FIGURE 3.14
Unit vectors at different points in the spherical coordinate system.

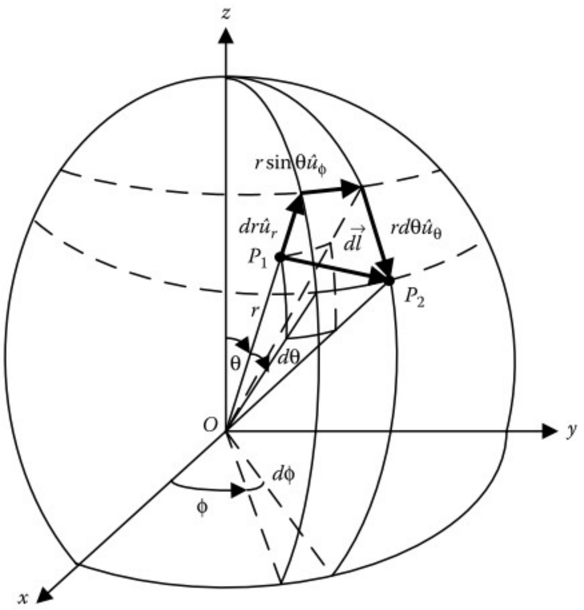
z
$r \sin\theta \hat{u}_\phi$
$dr\hat{u}_r$
$\vec{dl}$
$rd\theta\hat{u}_\theta$
P_1
P_2
r
θ
$d\theta$
O
y
$d\phi$
ϕ
x

FIGURE 3.15
Differential line element in the spherical coordinate system.

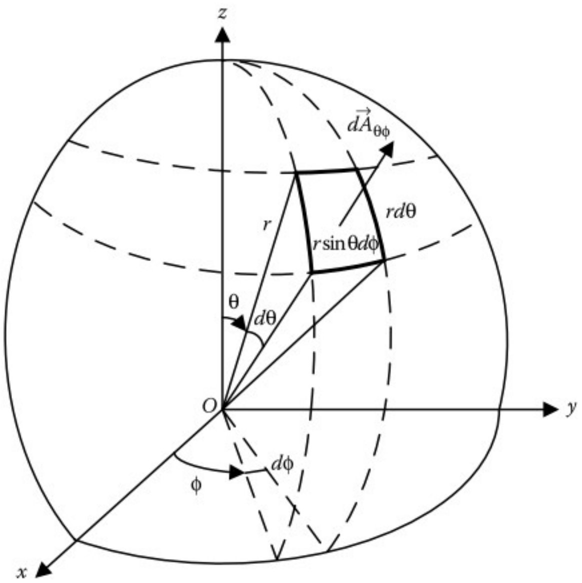

FIGURE 3.16
Differential area element on the sphere surface in the spherical coordinate system.

As shown in Figure 3.17, the differential volume element is given by Equation 3.21

$$dV = dl_1 dl_2 dl_3 = (dr)(rd\theta)(r\sin\theta d\phi) = r^2 \sin\theta drd\theta d\phi \tag{3.21}$$

As depicted in Equation 3.19, the differential distance in r-direction is $dl_1 = dr = h_1 dr$, that is, $h_1 = 1$, in θ-direction, it is $dl_2 = rd\theta = h_2 d\theta$, that is, $h_2 = r$ and in ϕ-direction, it is $dl_3 = r\sin\theta d\phi = h_3 d\phi$, that is, $h_3 = r\sin\theta$. Therefore, the metric coefficients of spherical coordinate system are $h_1 = 1$, $h_2 = r$ and $h_3 = r\sin\theta$.

3.5 Generalized Orthogonal Curvilinear Coordinate System

From the discussions of the previous sections, it may be seen that a major hindrance in treating the three commonly used orthogonal coordinate systems, namely, Cartesian, cylindrical and spherical coordinate systems, on an equal footing is that the unit vectors are not the best way to visualize these three coordinate systems. In this context, a generalized orthogonal curvilinear coordinate system provides insight for useful linkage between such orthogonal coordinate systems. Generalized orthogonal curvilinear coordinate system emphasizes on the similarities between different orthogonal coordinate systems rather than highlighting their differences. Thus, the definition and utilization of generalized

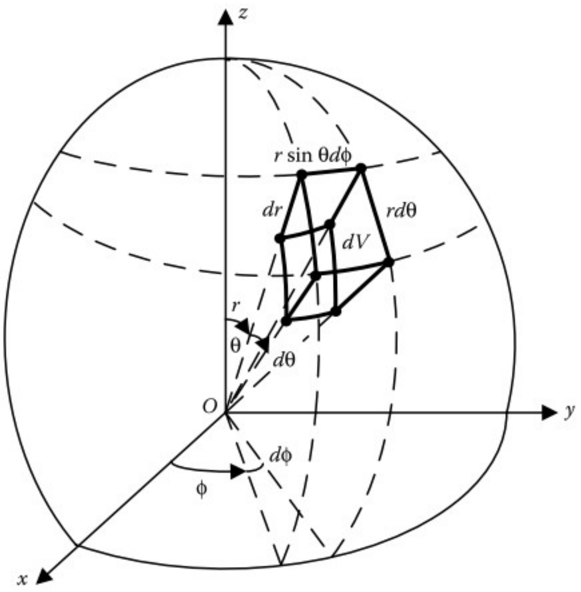

FIGURE 3.17
Differential volume element of a sphere in the spherical coordinate system.

orthogonal curvilinear coordinate system is very important for the proper understanding of the formulation and solution of electric field problems.

Consider that $f(x,y,z) = u$ is a function of three independent space variables x,y,z in the Cartesian coordinate system that specify a surface characterized by the constant parameter u. Setting such function equal to three different constant parameters u_1, u_2 and u_3 defines three different surfaces as follows:

$$f_1(x,y,z) = u_1,\ f_2(x,y,z) = u_2 \text{ and } f_3(x,y,z) = u_3 \tag{3.22}$$

Considering that these three surfaces are orthogonal, they intersect in space only at one point. In other words, any point in space could be uniquely defined by a set of values of the three parameters (u_1,u_2,u_3), which are then called the *orthogonal curvilinear coordinates* of the point being defined. The constant coordinate surfaces for a generalized orthogonal curvilinear coordinate system are shown in Figure 3.18.

As shown in Figure 3.18, the unit vectors $\hat{a}_1$, $\hat{a}_2$ and $\hat{a}_3$ are normal to the surfaces u_1 = constant, u_2 = constant and u_3 = constant, respectively, and point towards the increasing values of the coordinates u_1, u_2 and u_3, respectively. The orthogonality of the generalized curvilinear coordinate system is defined by Equation 3.23

$$\hat{a}_1 \cdot \hat{a}_2 = 0,\ \hat{a}_2 \cdot \hat{a}_3 = 0,\ \hat{a}_3 \cdot \hat{a}_1 = 0 \tag{3.23}$$

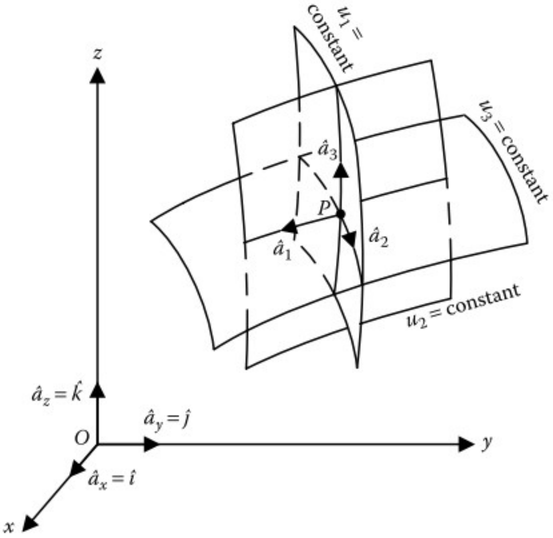

FIGURE 3.18
Constant coordinate surfaces in the generalized orthogonal curvilinear coordinate system.

and the right handedness is defined by Equation 3.24

$$\hat{a}_1 \times \hat{a}_2 = \hat{a}_3,\ \hat{a}_2 \times \hat{a}_3 = \hat{a}_1,\ \hat{a}_3 \times \hat{a}_1 = \hat{a}_2 \tag{3.24}$$

As shown in Figure 3.19, the surfaces $u = u_1$ and $u = u_1 + du_1$ are separated by a differential length element dl_1, which is normal to the surface $u = u_1$, where $dl_1 = h_1(u_1,u_2,u_3)du_1$. The scale factor or the metric coefficient h_1 is a function of the three curvilinear coordinates (u_1,u_2,u_3). The other two metric coefficients, that is, $h_2(u_1,u_2,u_3)$ and $h_3(u_1,u_2,u_3)$, are also similarly defined.

As shown in Figure 3.19, the differential line element in the generalized orthogonal curvilinear coordinate system is given by Equation 3.25

$$d\vec{l} = dl_1\hat{a}_1 + dl_2\hat{a}_2 + dl_3\hat{a}_3 = h_1du_1\hat{a}_1 + h_2du_2\hat{a}_2 + h_3du_3\hat{a}_3 \tag{3.25}$$

As shown in Figure 3.19, the differential area elements in the generalized orthogonal curvilinear coordinate system are given by Equation 3.26

$$\begin{aligned} d\vec{A}_{12} &= (h_1du_1)(h_2du_2)\hat{a}_3 = h_1h_2du_1du_2\hat{a}_3 \\ d\vec{A}_{23} &= (h_2du_2)(h_3du_3)\hat{a}_1 = h_2h_3du_2du_3\hat{a}_1 \\ d\vec{A}_{31} &= (h_3du_3)(h_1du_1)\hat{a}_2 = h_3h_1du_3du_1\hat{a}_2 \end{aligned} \tag{3.26}$$

and the differential volume element in generalized orthogonal curvilinear coordinate system is given by Equation 3.27

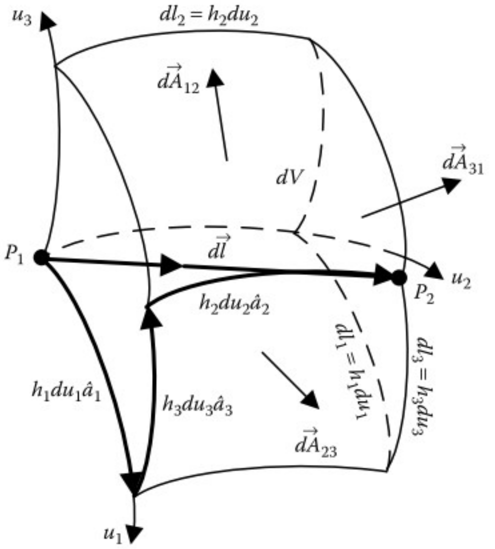

FIGURE 3.19
Differential line, area and volume elements in the generalized orthogonal curvilinear coordinate system.

$$dV = (h_1du_1)(h_2du_2)(h_3du_3) = h_1h_2h_3du_1du_2du_3 \tag{3.27}$$

In the light of the generalized orthogonal curvilinear coordinate system, the three commonly used orthogonal coordinate systems could be summarized as in Table 3.1.

Recall that the metric coefficients were earlier stated to be functions of the coordinates (u_1,u_2,u_3). Table 3.1 clearly shows such dependence of the metric coefficients on the respective coordinates in cylindrical and spherical coordinate systems. Cartesian coordinate system is unique in the sense that its metric coefficients are constant.

TABLE 3.1

Characterization of Orthogonal Coordinate Systems

Orthogonal Coordinate System	Coordinates	Unit Vectors	Metric Coefficients
Generalized	(u_1,u_2,u_3)	$(\hat{a}_1,\hat{a}_2,\hat{a}_3)$	(h_1,h_2,h_3)
Cartesian	(x,y,z)	$(\hat{i},\hat{j},\hat{k})$	$(1,1,1)$
Cylindrical	(r,θ,z)	$(\hat{u}_r,\hat{u}_\theta,\hat{u}_z)$	$(1,r,1)$
Spherical	(r,θ,ϕ)	$(\hat{u}_r,\hat{u}_\theta,\hat{u}_\phi)$	$(1,r,r\sin\theta)$

2

Gauss's Law and Related Topics

ABSTRACT Gauss's law constitutes one of the fundamental laws of electromagnetism. Gauss's law provides an easy means of finding electric field intensity or electric flux density for symmetrical charge distributions such as a point charge, an infinite line charge, an infinite cylindrical surface charge and so forth. The procedure for applying Gauss's law to calculate the electric field involves the introduction of Gaussian surfaces, which are closed surfaces on which electric field is either constant or zero. In most of the cases, electric field is either normal or tangential to the Gaussian surfaces. There are two different ways of stating Gauss's law, namely, integral form and differential form, which are employed to solve field problems depending on suitability. These two forms of Gauss's law in combination lead to *divergence theorem*. The divergence relationship between electric flux density and volume charge density as obtained from Gauss's law and the gradient relationship between electric field intensity and electric potential could be combined in the form of an elliptic partial differential equation known as Poisson's equation. In charge-free field region, Poisson's equation reduces to Laplace's equation. Both Poisson's and Laplace's equations could be solved for field quantities imposing suitable boundary conditions.

2.1 Introduction

German mathematician and physicist Karl Friedrich Gauss published his famous flux theorem in 1867, which is now well known in physics as Gauss's law. It is one of the basic laws of classical field theory. Gauss's law is a general law that can be applied to any closed surface. With the help of Gauss's law, the amount of enclosed charge could be assessed by mapping the field on a surface outside the charge distribution. By the application of Gauss's law, the electric field in many practical arrangements could be evaluated by forming a symmetric surface, commonly known as *Gaussian surface*, surrounding a charge distribution and then determining the electric flux through that surface.

In electric field theory, distinction is made between free charge and bound charge. Free charge implies a charge that is free to move over distances large as compared to atomic scale lengths. Free charge typically comes from electrons, for example, in the case of metals, or ions, or in the case of aqueous solutions. Bound charge implies charges of equal magnitude but opposite signs that are held close to each other and are free to move only through atomic scale lengths. Bound charges arise in the context of polarizable dielectric materials. Typical example of bound charge is the positive charge of an atomic nucleus and the negative charge of its associated electron cloud. The microscopic charge displacements in dielectric materials are not as dramatic as the rearrangement of charge in a conductor, but their cumulative effects account for the characteristic behaviour of dielectric materials. Typically, the detailed effect of bound charge is represented through electrical permittivity of dielectric materials, which will be discussed in Chapter 5. The Gauss's law, as discussed in this chapter, is in terms of electric flux and the free charge only.

2.2 Useful Definitions and Integrals

2.2.1 Electric Flux through a Surface

Consider that the total electric flux through the surface *S* of an area *A* needs to be evaluated, as shown in Figure 2.1. The first important point to be noted here is that the electric flux density at different locations on *S* may not be the same. Therefore, the total flux through *S* has to be computed by subdividing the entire surface into a large number of smaller surfaces, such that over each small surface area, the electric flux density vector is uniform. As shown in Figure 2.1, for any small surface area *dA*, the area vector $d\vec{A}$ and the electric flux density vector $\vec{D}$ could be along different directions. Then the electric flux through *dA* is given by $\vec{D}\cdot d\vec{A}$

Then the total flux through *S* is given by the summation $\vec{D}_1 \cdot d\vec{A}_1 + \vec{D}_2 \cdot d\vec{A}_2 + \cdots$

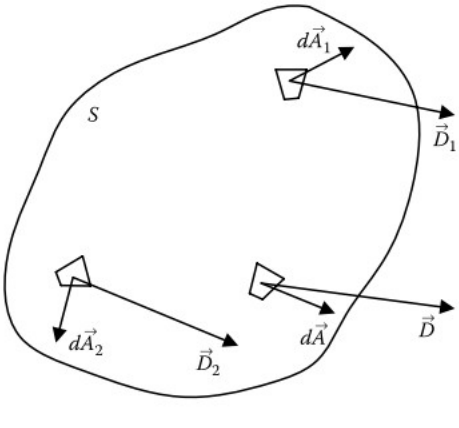

FIGURE 2.1
Evaluation of the total electric flux through a surface.

When the sizes of the individual areas become infinitesimally small, then the total flux through *S* is given by the integral

$$\int_S \vec{D}\cdot d\vec{A} = \iint \vec{D}\cdot d\vec{A} \tag{2.1}$$

2.2.2 Charge within a Closed Volume

Consider that the total charge within the volume *V* enclosed by the surface *S* needs to be evaluated, as shown in Figure 2.2. Here also, it is to be kept in mind that the distributed charge distribution may not be uniform throughout *V*. Then the entire volume needs to be subdivided into a large number of small volumes such that within each small volume, charge density is constant. Then the total charge within *V* is given by the summation $\rho_{v1}dV_1 + \rho_{v2}dV_2 + \rho_{v3}dV_3 + \cdots$

When the sizes of the individual volumes become infinitesimally small, then the total charge within *V* is given by the integral $\int_V \rho_v dV = \iiint \rho_v\, dV$.

2.2.3 Solid Angle

The concept of solid angle is a natural extension of two-dimensional plane angle to three dimensions. The solid angle subtended by an area *A* at the point *O* is measured by the area Ω on the surface of the unit sphere centred at *O*, as shown in Figure 2.3. This is the area that would be cut out on the unit sphere surface by lines drawn from *O* to every point on the periphery of *A*. It is measured in terms of the unit called *steradian*, abbreviated as sr.

Consider an elementary area *dA*, as shown in Figure 2.4. If all the points on the periphery of *dA* are joined to the point *O*, then these lines cut out an area *d*Ω on the surface of the unit sphere. In other words, *dA* subtends a solid angle *d*Ω at *O*. Because the area is infinitesimally small, all points on *dA*

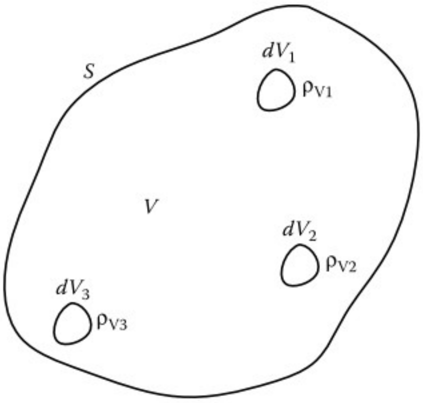

FIGURE 2.2
Evaluation of the total charge within a volume.

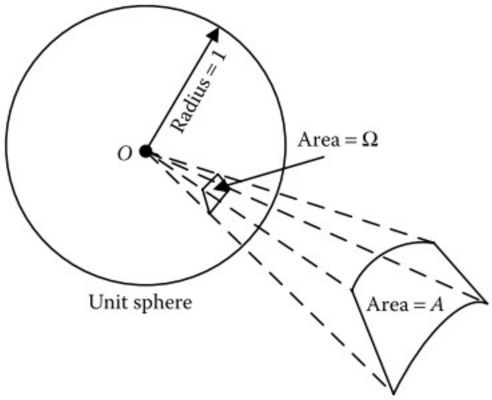

FIGURE 2.3
Solid angle subtended by an area at a point.

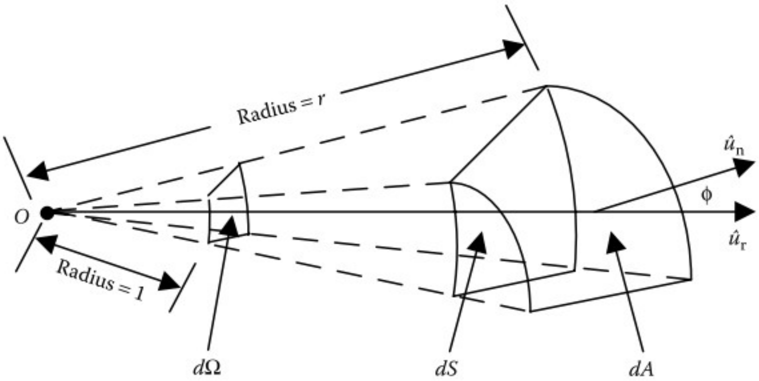

FIGURE 2.4
Evaluation of the solid angle subtended by an elementary area.

could be considered to be equidistant from O. Let the distance of dA from O be r. The unit vector in the direction of the distance vector $\vec{r}$ is $\hat{u}_r$. Because dA is infinitesimally small, it may be taken as flat for practical purposes and hence $\hat{u}_n$ is defined as the unit vector in the direction normal to dA, as shown in Figure 2.4. The angle between $\hat{u}_r$ and $\hat{u}_n$ is ϕ. Then the projection of dA on the sphere of radius r will be $dS = dA\cos\phi$.

The areas dS and $d\Omega$ have the same general shape and are related to the respective radii of the spheres having the following proportionality:

$$\frac{d\Omega}{dS} = \frac{1^2}{r^2}, \text{ or, } d\Omega = \frac{dA\cos\phi}{r^2} = \frac{\hat{u}_r \cdot d\vec{A}}{r^2} \tag{2.2}$$

When an area completely encloses O, then Ω due to that area on the unit sphere will be 4π. Hence, the solid angle subtended at a point by an area that completely encloses the point is 4π sr, which happens to be maximum possible value of the solid angle.

2.3 Integral Form of Gauss's Law

Gauss's law states that the net electric flux through any closed surface enclosing a homogeneous volume of material is equal to the net electric charge enclosed by that closed surface. In other words, the surface integral of electric flux density vector over a closed surface is equal to the volume integral of charge densities within the volume enclosed by that surface.

In integral form,

$$\int_A \vec{D} \cdot d\vec{A} = \int_V \rho_v dV \tag{2.3}$$

Gauss's law is valid for any discrete set of point charges. Nevertheless, this law is also valid when an electric field is produced by charged objects with continuously distributed charges, because any continuously distributed charge may be viewed as a combination of discrete point charges.

Figure 2.5 shows a certain volume V of homogeneous material enclosed by a surface A. This volume is continuously charged by distributed charges. Consider these distributed charges to be represented by N number of discrete point charges, as shown in Figure 2.5.

Take a point charge q_1 located within V. Then consider an elementary area dA on A, as shown in Figure 2.5. The distance of dA from q_1 is, say, r.

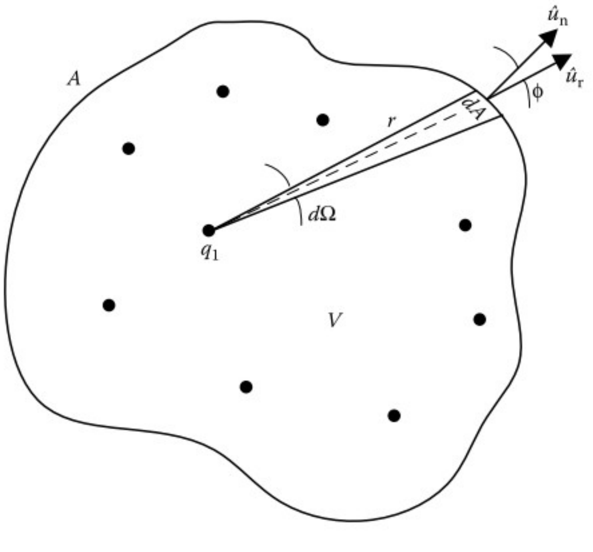

FIGURE 2.5
Pertaining to the integral form of Gauss's law.

Then at a distance r from q_1,

$$\vec{E}_1 = \frac{q_1}{4\pi\varepsilon_0 r^2}\hat{u}_\text{r} \text{ and } \vec{D}_1 = \frac{q_1}{4\pi r^2}\hat{u}_\text{r} \tag{2.4}$$

Then, the total flux through A due to q_1 could be obtained as follows:

$$\int_A \vec{D}_1 \cdot d\vec{A} = \int_A \frac{q_1}{4\pi}\frac{\hat{u}_\text{r} \cdot d\vec{A}}{r^2} = \frac{q_1}{4\pi}\int_A d\Omega, \text{ as } \frac{\hat{u}_\text{r} \cdot d\vec{A}}{r^2} = d\Omega \tag{2.5}$$

Because A completely encloses q_1, the solid angle subtended by A at the location of q_1 is 4π. Thus,

$$\int_A \vec{D}_1 \cdot d\vec{A} = \frac{q_1}{4\pi} \times 4\pi = q_1 \tag{2.6}$$

Considering, N number of point charges within V, total electric flux density at the location of dA, as shown in Figure 2.5, will be as follows:

$$\vec{D} = \vec{D}_1 + \vec{D}_2 + \vec{D}_3 + \cdots + \vec{D}_N \tag{2.7}$$

Therefore, the total flux through A will be as follows:

$$\begin{aligned} \int_A \vec{D} \cdot d\vec{A} &= \int_A \left(\vec{D}_1 + \vec{D}_2 + \vec{D}_3 + \cdots + \vec{D}_N\right) \cdot d\vec{A} \\ &= \int_A \vec{D}_1 \cdot d\vec{A} + \int_A \vec{D}_2 \cdot d\vec{A} + \int_A \vec{D}_3 \cdot d\vec{A} + \cdots + \int_A \vec{D}_N \cdot d\vec{A} \\ &= q_1 + q_2 + q_3 + \cdots + q_N \end{aligned} \tag{2.8}$$

The right-hand side (RHS) of Equation 2.8 is equal to the total electric charge enclosed within V. Considering continuously distributed charge within V,

The total charge within is

$$V = \int_V \rho_\text{v}\, dV$$

where:

ρ_v is the volume charge density within V

Hence, Equation 2.8 can be rewritten as

$$\int_A \vec{D} \cdot d\vec{A} = \int_V \rho_\text{v} dV$$

From a physical viewpoint, the above expression indicates that the sum of all sources (positive electric charges) and all sinks (negative electric charges) within a volume gives the net flux through the surface enclosing the same volume.

Gauss's law in integral form is only useful for exact solutions when the electric field has symmetry, for example, spherical, cylindrical or planar symmetry. The symmetry of the electric field follows from the symmetry of the charge distribution that is given. Therefore, the electric field of a symmetrically charged sphere will also have the spherical symmetry.

2.3.1 Gaussian Surface

A closed surface in a three-dimensional space through which the flux of a vector field is calculated is known as *Gaussian surface*. It is an arbitrary closed surface over which surface integral is performed to evaluate the total amount of enclosed source quantity, for example, electric charge. A Gaussian surface could also be used for calculating the electric field due to a given charge distribution.

Gaussian surfaces are normally carefully chosen in order to exploit symmetries to simplify the evaluation of the surface integral. If the symmetry is such that a surface can be found on which the electric field is constant, then the evaluation of electric flux can be done by multiplying the value of the field with the area of the Gaussian surface.

Two commonly used Gaussian surfaces are as follows:

1. The spherical Gaussian surface, which is chosen to be concentric with the charge distribution. It can be used to determine the electric field or electric flux due to a point charge or a spherical shell of uniform charge density or any other charge distribution with spherical symmetry, as shown in Figure 2.6.
2. The cylindrical Gaussian surface, which is chosen to be co-axial with the charge distribution. It can be used to determine the electric

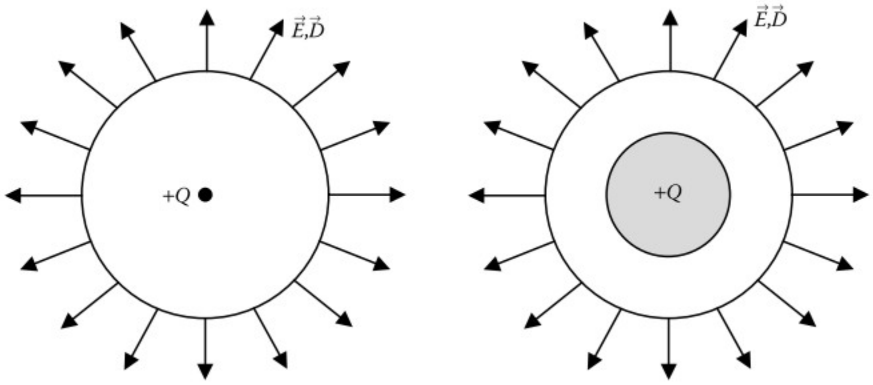

Gaussian surface for a point charge

Gaussian surface for a charged sphere

FIGURE 2.6
Spherical Gaussian surface.

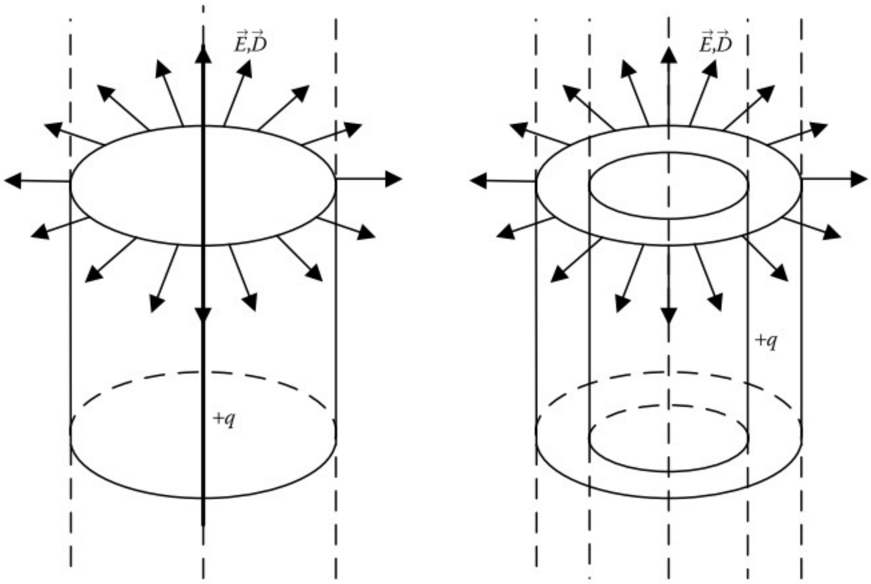

Gaussian surface due to an infinitely long line of uniform charge density

Gaussian surface due to an infinitely long cylinder of uniform charge density

FIGURE 2.7
Cylindrical Gaussian surface.

field or electric flux due to an infinitely long line of uniform charge density or an infinitely long cylinder of uniform charge density, as shown in Figure 2.7.

2.4 Differential Form of Gauss's Law

If the electric field is known everywhere, then with the help of integral form of Gauss's law, the charge in any given field region can be deduced by integrating the electric field. However, in most of the practical cases, the electric field needs to be computed when the electric charge distribution is known. This is much more difficult because even if the total flux through a given surface is known, the information about the electric field may be unknown, as electric flux could go in and out of the surface in complex patterns. In this context, the differential form of Gauss's law becomes useful.

Consider that Gauss's law is to be applied for the infinitesimal parallelepiped, as shown in Figure 2.8. The volume of the infinitesimal parallelepiped is

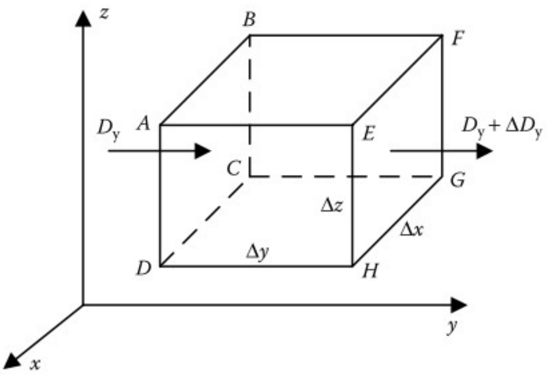

FIGURE 2.8
Application of Gauss's law to a differential volume.

$\Delta V = \Delta x \Delta y \Delta z$. Let the component of electric flux density going normally into the surface ABCD be D_y and that coming out of the surface EFGH be $D_y + \Delta D_y$. Surfaces ABCD and EFGH are lying along the x–z plane. Then the net electric flux coming out in the y-direction, which is normal, to x–z plane is given by

$$(D_y + \Delta D_y)\Delta x \Delta z - D_y \Delta x \Delta z = \Delta D_y \Delta x \Delta z = \frac{\Delta D_y}{\Delta y}\Delta x \Delta y \Delta z \approx \frac{\partial D_y}{\partial y}\Delta V \qquad (2.9)$$

Similarly, the net electric fluxes coming out in x- and z-directions are given by $(\partial D_x/\partial x)\Delta V$ and $(\partial D_z/\partial z)\Delta V$, respectively. Hence, net electric flux coming of the volume ΔV

$$= \frac{\partial D_x}{\partial x}\Delta V + \frac{\partial D_y}{\partial y}\Delta V + \frac{\partial D_z}{\partial z}\Delta V = \left(\frac{\partial D_x}{\partial x} + \frac{\partial D_y}{\partial y} + \frac{\partial D_z}{\partial z}\right)\Delta V$$

Let ρ_v be the volume charge density of the small volume ΔV. Then according to Gauss's law, the above expression for the net flux coming out of ΔV is equal to the total charge within the volume, that is, $\rho_v \Delta V$. Therefore,

$$\begin{aligned} &\left(\frac{\partial D_x}{\partial x} + \frac{\partial D_y}{\partial y} + \frac{\partial D_z}{\partial z}\right)\Delta V = \rho_v \Delta V \\ &\text{or, } \frac{\partial D_x}{\partial x} + \frac{\partial D_y}{\partial y} + \frac{\partial D_z}{\partial z} = \rho_v \end{aligned} \qquad (2.10)$$

Equation 2.10 is known as the differential form of Gauss's law. It can be expanded as follows:

$$\left(\frac{\partial}{\partial x}\hat{i}+\frac{\partial}{\partial y}\hat{j}+\frac{\partial}{\partial z}\hat{k}\right)\left(D_x\hat{i}+D_y\hat{j}+D_z\hat{k}\right)=\rho_v,\ \text{or},\ \vec{\nabla}\cdot\vec{D}=\rho_v \tag{2.11}$$

In Section 1.7.4, it has been mentioned that when the $\vec{\nabla}$ operator acts on a scalar quantity, it is denoted as gradient. As depicted in Equation 2.11, when the $\vec{\nabla}$ operator acts on a vector quantity with a dot product, then it is denoted as divergence.

The physical meaning of divergence of a vector field can be explained from the left-hand side (LHS) of Equation 2.10. From the above discussion, it may be seen that the LHS of Equation 2.10 is the net flux coming out per unit volume at a given location. Consequently, the divergence of a vector field at any location is the net flux of that vector field coming out per unit volume at that given location. If the divergence of a vector field is positive at any location, then the flux coming out of unit volume is higher than the flux going into the unit volume at that location. On the other hand, if the divergence of a vector field is negative, then the flux going into the unit volume is higher than the flux coming out of the unit volume.

2.5 Divergence Theorem

Integral form of Gauss's law is, $\int_A \vec{D}\cdot d\vec{A}=\int_V \rho_v dV$ and differential form of Gauss's law is, $\vec{\nabla}\cdot\vec{D}=\rho_v$.

By substituting ρ_v on the RHS of the integral form by the LHS of the differential form, the above two equations could be combined as follows:

$$\int_A \vec{D}\cdot d\vec{A}=\int_V (\vec{\nabla}\cdot\vec{D})dV \tag{2.12}$$

where:

The volume V is enclosed by the surface A

Equation 2.12 shows that the divergence theorem could be used to convert a volume integral over V to a surface integral over A enclosing V.

2.6 Poisson's and Laplace's Equations

A useful approach to the calculation of electric potentials is to relate electric potential to electric charge density, which causes electric potential. Because electric potential is a scalar quantity, this approach has advantages over

3

Orthogonal Coordinate Systems

ABSTRACT The purpose of this chapter is to present coordinate systems in such a way that it improves the ability of readers to use them to describe the electric field in a better way. The goal is to put forward the coordinate system as a useful language for describing, visualizing and understanding the concepts that are central to electric field analysis. Many mathematical operators related to electric field analysis have simple forms in Cartesian coordinates and are easy to remember and evaluate. In fact, all the problems could be solved using the Cartesian coordinate system. However, the resulting expressions might be unnecessarily complex. While dealing with problems that have high degree of symmetry, it is helpful to use a coordinate system that exploits the symmetry. As a result, it is useful to have a general method for expressing mathematical operators in non-Cartesian forms. Orthogonal curvilinear coordinate system is a useful approach by which the matching of the symmetry of a given physical configuration could be done and simplified mathematical models could be built to clarify the field concepts clearly.

3.1 Basic Concepts

The concept that unifies the different coordinate systems is surfaces of constant coordinates. It means that an equation of the form (coordinate = value) defines a surface. In order to specify a point uniquely in a three-dimensional space, three different types of surfaces are needed. The values associated with the three constant coordinate surfaces that intersect at a specific point are the three coordinates, which uniquely define that point.

Consider that three surfaces are defined by three functions f_1, f_2 and f_3, respectively, in a three-dimensional space. The coordinate, that is, the value, that describes a surface could be defined by the equation $f_i = u_i$. Here, u is introduced to represent a generalized coordinate. In the stated equation, u_i refers to one of the coordinates in a particular system. Thus, the three values (u_1,u_2,u_3) define a point, where the three surfaces defined by Equation 3.1 intersect. These three values (u_1,u_2,u_3) are then called the *coordinates* of that particular point.

$$f_1 = u_1,\ f_2 = u_2 \text{ and } f_3 = u_3 \tag{3.1}$$

The three functions, however, could not be chosen arbitrarily. The three functions are to be chosen in such a way that for any choice of the three coordinates (u_1,u_2,u_3), the three surfaces generated by Equation 3.1 will intersect only at one point. It will ensure that a given set of three coordinate values (u_1,u_2,u_3) always specifies only one point in space. Moreover, the constant coordinate surfaces are always perpendicular to each other, that is, orthogonal, at the point where they intersect. The intersection of any two constant coordinate surfaces defines a coordinate curve.

3.1.1 Unit Vector

An important concept associated with coordinate system is unit vector. It will be easy to understand the concept if it is explained with the help of a specific example involving the Cartesian coordinate system, as shown in Figure 3.1. Each value of z, for example, z_1, z_2 or z_3, defines a plane normal to the z-axis. The unit vector $\hat{k}$ is defined in such a way that it is normal to all such z = constant planes and points in the direction of the planes with increasing values of z, as shown in Figure 3.1. The magnitude of the vector $\hat{k}$ is taken as unity. The other two unit vectors in the Cartesian coordinate system could be visualized in the same manner.

In a generalized manner, the properties of the unit vector associated with a constant coordinate surface defined by the function f_i are as follows: (1) it is perpendicular to the surface $f_i = u_i$ at any point (u_1,u_2,u_3); (2) it points in the direction of the surfaces with increasing value of u_i and (3) it is one unit in magnitude. Orthogonality of the constant coordinate surfaces, as discussed earlier, demands that the dot product of any two different unit vectors will always be zero. Conversely, the dot product of two same unit vectors, for example, $\hat{k}\cdot\hat{k}$, will always be 1.

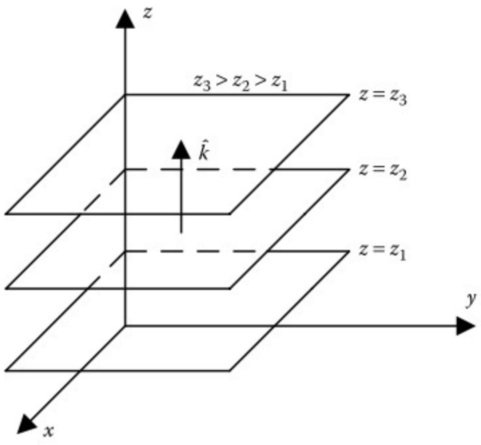

FIGURE 3.1
The unit vector $\hat{k}$.

3.1.2 Right-Handed Convention

Because three constant coordinate surfaces are involved in determining the coordinates of any point, the question that arises is in which order the three coordinates are to be mentioned while specifying a point. In this context, the choice of the first coordinate (u_1) or surface (f_1) is arbitrary. Once f_1 is chosen, then the order in which the other two constant coordinate surfaces f_2 and f_3 is to be taken is determined by the right-handed system as governed by the definition of vector cross product. In other words, the cross product of the first and second unit vectors must be equal to the third unit vector.

This right-handed property of the coordinate system is an arbitrary convention. But once that choice is made for one coordinate system, it has to be followed for all other coordinate systems to ensure that the expressions for the laws of electric field have the same form in all the coordinate systems.

3.1.3 Differential Distance and Metric Coefficient

The constant coordinate surfaces could be of many different types. The simplest of them is constant distance surfaces for which the differential change du_i in the coordinate u_i is the same as the differential distance dl_i between the surfaces $f_i = u_i$ and $f_i = u_i + du_i$. However, there are other types of surfaces such as constant angle surfaces, in which case the differential distance is related to differential coordinate (angle) change as

$$dl_i = h_i du_i = h_i d\theta_i \tag{3.2}$$

where:

The scale factor h_i is the corresponding radius

The scale factors for each of the coordinate system are called the *metric coefficients* of the respective coordinate system. The three metric functions of any coordinate system (h_1,h_2,h_3) constitute a unique set and hence are often known as the *signature of a coordinate system*. Metric coefficients are also very important in writing the vector operators in the general orthogonal curvilinear coordinate system, as will be explained in Section 3.5.

3.1.4 Choice of Origin

A basic element of any coordinate system is the choice of origin. If any object or any arrangement is given, then any point within the object or within the whole arrangement could be chosen as origin.

From the earlier discussions, it is clear that the same point will have very different coordinates when defined in different coordinate systems. Therefore, the question that needs to be answered is that whether there is any property of the position of the point that remains the same in different coordinate systems. The answer to this question is that the distance of the point from the origin remains the same in all the coordinate systems if the origin is kept the same in all the cases. As a logical extension, it may be stated that the distance between any two points remains the same no matter how their coordinates are defined in any coordinate system.

3.2 Cartesian Coordinate System

The invention of Cartesian coordinates in the seventeenth century by the French mathematician René Descartes revolutionized mathematics by providing the first systematic link between Euclidean geometry and algebra. In this coordinate system, the three constant coordinate surfaces are defined by Equation 3.3

$$f_1 = x\,,\, f_2 = y\,,\, f_3 = z \tag{3.3}$$

Figure 3.2 shows such constant coordinate surfaces, which are three planes. Cartesian coordinate system has the unique feature that all the three constant coordinate surfaces are constant distance surfaces.

Each point P in space is then assigned a triplet of values (x,y,z), the so-called Cartesian coordinates of that point. The ranges of the values of three coordinates are $-\infty < x < \infty$, $-\infty < y < \infty$ and $-\infty < z < \infty$. Coordinates can be

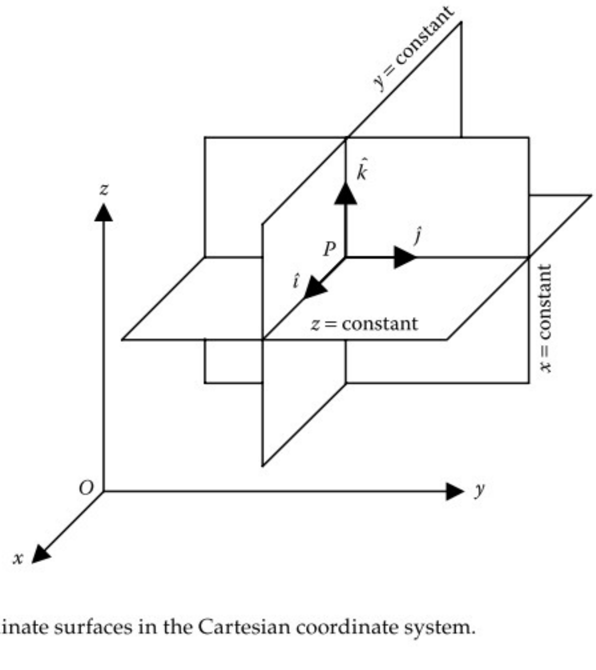

FIGURE 3.2
Constant coordinate surfaces in the Cartesian coordinate system.

defined as the positions of the normal projections of the point onto three mutually perpendicular constant coordinate surfaces passing through the origin, expressed as signed distances from the origin.

In the Cartesian coordinate system, three mutually perpendicular axes are commonly defined as follows. The intersection of two mutually perpendicular planes y = constant and z = constant passing through the origin gives a coordinate curve, which is a straight line. This line is then defined as the x-axis. Similarly, the intersections of other two pairs of mutually perpendicular constant coordinate planes passing through origin define the y- and z-axes.

The unit vector $\hat{i}$ is perpendicular to all the surfaces described by x = constant and points in the direction of increasing value of x-coordinate. In the same manner, unit vectors $\hat{j}$ and $\hat{k}$ are defined in y- and z-directions, respectively. Another unique feature of the Cartesian coordinate system is that the directions of the three unit vectors remain the same irrespective of the location of the point in space. The orthogonality of the Cartesian coordinate system is defined by Equation 3.4

$$\hat{i}\cdot\hat{j} = 0,\ \hat{j}\cdot\hat{k} = 0,\ \hat{k}\cdot\hat{i} = 0 \tag{3.4}$$

and the right handedness is defined by Equation 3.5

$$\hat{i}\times\hat{j} = \hat{k},\ \hat{j}\times\hat{k} = \hat{i},\ \hat{k}\times\hat{i} = \hat{j} \tag{3.5}$$

As shown in Figure 3.3, the differential line element in the Cartesian coordinate system is given by Equation 3.6

$$d\vec{l} = dl_1\hat{i} + dl_2\hat{j} + dl_3\hat{k} = dx\hat{i} + dy\hat{j} + dz\overleftarrow{k} \tag{3.6}$$

As shown in Figure 3.4, the differential area element corresponding to the surface of a differential cubical volume is given by Equation 3.7

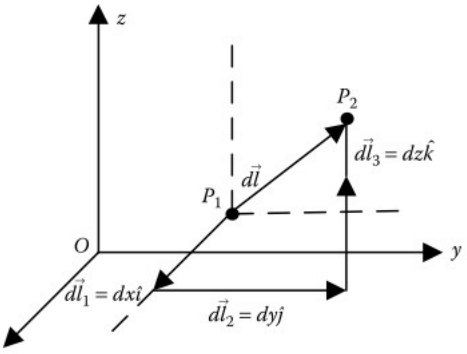

FIGURE 3.3
Differential line element in the Cartesian coordinate system.

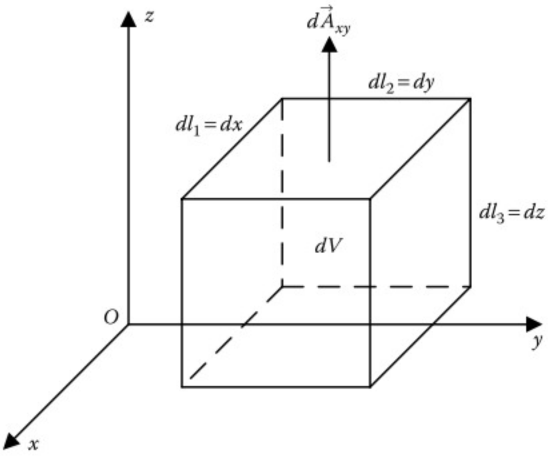

FIGURE 3.4
Differential area and volume elements in the Cartesian coordinate system.

$$d\vec{A}_{xy} = dl_1 dl_2 \hat{k} = dxdy\hat{k} \tag{3.7}$$

As shown in Figure 3.4, the differential volume element is given by Equation 3.8

$$dV = dl_1 dl_2 dl_3 = dxdydz \tag{3.8}$$

As depicted in Equation 3.6 the differential distance in x-direction is $dl_1 = dx = h_1 dx$, that is, $h_1 = 1$. Similarly, in y- and z-directions, $h_2 = h_3 = 1$. The metric coefficients of the Cartesian coordinate system are, therefore, $h_1 = 1$, $h_2 = 1$ and $h_3 = 1$.

3.3 Cylindrical Coordinate System

Cylindrical coordinates are an alternate way of describing points in a three-dimensional space. In this coordinate system, one of the rectangular coordinate planes, namely, the x–y plane, as described by the Cartesian coordinate system, is replaced by a polar plane. In the cylindrical coordinate system, everything is measured with respect to a fixed point called the *pole* and an axis called the *polar axis*. The pole is the equivalent to the origin in the Cartesian coordinate system and the polar axis corresponds to the positive direction of the x-axis. The cylindrical coordinates of a point are then the ordered triplet (r,θ,z), as defined in Figure 3.5.

As shown in Figure 3.5, r is the distance from the pole to the projection of the point P on the polar plane, that is, the x–y plane passing through the pole, θ is the azimuthal angle, that is, the angle from the polar axis spinning around the z-axis in counter-clockwise direction, and z is the vertical height from the polar plane. The ranges of the values of the three coordinates are $0 \le r < \infty$, $0 \le \theta \le 2\pi$ and $-\infty < z < \infty$.

10

Sphere or Cylinder in Uniform External Field

ABSTRACT If a conducting or dielectric object is present within the electric field caused by external sources, then the induced charges on such an object adds to the existing external field. If the size of the conducting or dielectric object is small compared to the external field region, then the electric field is modified only in the vicinity of such an object. In order to get the complete solution of the electric field in such configuration, the boundary conditions imposed by the conducting or dielectric object must also be satisfied. If the object is of well-defined geometric shape such as cylinder or sphere, and if the external field is considered to be uniform in space, then the complete solution for the electric field could be obtained analytically with the help of method of separation of variables. These analytical solutions are often used to validate numerical methodology and could also be used to estimate field modification due to small conducting or dielectric pieces present in large high-voltage insulation arrangement.

10.1 Introduction

Conducting and dielectric components are integral parts of any electrical equipment. If the size of the conducting or dielectric object is very small compared to dimensions of the field region where the object is located, then the object contributes to the field only in the domain near the object. In many cases, such objects are present as stray bodies in high-voltage insulation arrangement. As practical examples, one may cite a small piece of conductor or dielectric floating in liquid insulation of large volume in transformers, metallic dust particles floating in gaseous insulation within gas-insulated system (GIS) and so on. It is important to understand how the presence of a conducting or dielectric object modifies the external field in the vicinity of the object, because any enhancement of the electric field intensity due to the conducting or dielectric object may lead to unwanted discharge or in the worst case failure of the insulation system.

If it is assumed that the source charges (in practical arrangement, the electrodes or conductors with specific potentials) that produce the external field is located far away from the object under consideration, then they are unaffected by the presence of the object. Consequently, the field due to the source charges may be considered to be uniform at the location of the object. If the object is such that its shape is defined by well-known mathematical functions, for example, cylinders or spheres, then the complete solution for the electric field due to the source charges located at far away positions and the induced charges on the surface of the object could be obtained by solving Laplace's equation considering the field region to be free from any volume charge. However, in order to get the complete solution appropriate boundary conditions on the surface of the object, whether it is conducting or dielectric, need to be satisfied. One of the common methods of getting the analytical solution for a cylinder or sphere in a uniform external field is the method of separation of variables as described in this chapter.

10.2 Sphere in Uniform External Field

Consider a spherical object of radius a within uniform external field, as shown in Figure 10.1. Because the boundary is a sphere of r = constant, hence the system is best described in spherical coordinates, as shown in Figure 10.1. The uniform external field is given by $\vec{E}_0 = -E_0\hat{u}_z$ and the potential at any point due to the external field is given by $E_0 r\cos\theta = E_0 z$ with respect to the centre of the sphere. In order to get the complete solution for the electric field in this system, Laplace's equation in spherical coordinates, as given in Equation 10.1, needs to be solved.

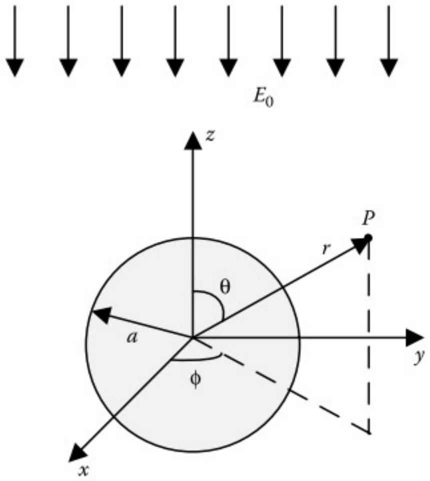

FIGURE 10.1
Sphere in uniform external field.

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial V}{\partial r}\right)+\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial V}{\partial\theta}\right)+\frac{1}{r^2\sin^2\theta}\frac{\partial^2 V}{\partial\phi^2}=0 \tag{10.1}$$

The field system has azimuthal symmetry with respect to the z-axis, that is, the field system does not change with the rotation around the z-axis. Therefore, z-axis is made the polar axis in the spherical coordinate system. Then the field is independent of coordinate φ and Laplace's equation reduces to

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial V}{\partial r}\right)+\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial V}{\partial\theta}\right)=0 \tag{10.2}$$

In order to separate the terms of the left-hand side (LHS) of Equation 10.2 into functions of only one variable, Equation 10.2 may be rewritten by multiplying r^2 as

$$\frac{\partial}{\partial r}\left(r^2\frac{\partial V}{\partial r}\right)+\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial V}{\partial\theta}\right)=0 \tag{10.3}$$

Then LHS of Equation 10.3 is the sum of two terms, which are functions of only one variable each, that is, the first term is function of r only, whereas the second term is function of θ only. The solution to Equation 10.3 can be obtained as the product of two functions of which one is dependent only on r and the other is dependent only on θ.

Let the assumed solution be

$$V(r,\theta)=M(r)N(\theta) \tag{10.4}$$

The assumed solution is convenient as the boundary lies at r = constant.

Combining Equations 10.3 and 10.4

$$\frac{\partial}{\partial r}\left[r^2\frac{\partial M(r)N(\theta)}{\partial r}\right]+\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left[\sin\theta\frac{\partial M(r)N(\theta)}{\partial\theta}\right]=0$$

$$\text{or, } N(\theta)\frac{\partial}{\partial r}\left[r^2\frac{\partial M(r)}{\partial r}\right]+M(r)\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left[\sin\theta\frac{\partial N(\theta)}{\partial\theta}\right]=0$$

Dividing by $M(r)N(\theta)$,

$$\frac{1}{M(r)}\frac{d}{dr}\left[r^2\frac{dM(r)}{dr}\right]+\frac{1}{N(\theta)}\frac{1}{\sin\theta}\frac{d}{d\theta}\left[\sin\theta\frac{dN(\theta)}{d\theta}\right]=0 \tag{10.5}$$

The partial derivatives become total derivatives in Equation 10.5, as each term is dependent on only one coordinate.

The sum of two terms of the LHS of Equation 10.5 could be zero only when the two terms are separately equal to opposite and equal constant terms as given in Equation 10.6.

Equal and opposite separation constant solution:

$$\frac{1}{M(r)}\frac{d}{dr}\left[r^2\frac{dM(r)}{dr}\right]=+p \text{ and } \frac{1}{N(\theta)}\frac{1}{\sin\theta}\frac{d}{d\theta}\left[\sin\theta\frac{dN(\theta)}{d\theta}\right]=-p \qquad (10.6)$$

where:
p is a positive constant

Another solution is obtained when the separation constant is zero.
Hence, the zero separation constant solution

$$\frac{1}{M(r)}\frac{d}{dr}\left[r^2\frac{dM(r)}{dr}\right]=0 \text{ and } \frac{1}{N(\theta)}\frac{1}{\sin\theta}\frac{d}{d\theta}\left[\sin\theta\frac{dN(\theta)}{d\theta}\right]=0 \qquad (10.7)$$

Each of the above-mentioned two solutions is to be obtained separately.

Determination of the zero separation constant solution:
The first term of Equation 10.7 is

$$\frac{1}{M(r)}\frac{d}{dr}\left[r^2\frac{dM(r)}{dr}\right]=0$$

where:
$M(r)$ is non-zero

Therefore,

$$\frac{d}{dr}\left[r^2\frac{dM(r)}{dr}\right]=0,\text{ or, } r^2\frac{dM(r)}{dr}=C_0,\text{ or, }\frac{dM(r)}{dr}=\frac{C_0}{r^2}$$

Integrating and incorporating constants of integration

$$M(r)=\frac{C_{10}}{r}+C_{20} \qquad (10.8)$$

Next the second term of Equation 10.7 is

$$\frac{1}{N(\theta)}\frac{1}{\sin\theta}\frac{d}{d\theta}\left[\sin\theta\frac{dN(\theta)}{d\theta}\right]=0$$

where:
$N(\theta)$ is non-zero

Therefore,

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\left[\sin\theta\frac{dN(\theta)}{d\theta}\right]=0,\text{ or, }\frac{d}{d\theta}\left[\sin\theta\frac{dN(\theta)}{d\theta}\right]=0$$

$$\text{or, }\sin\theta\frac{dN(\theta)}{d\theta}=A_0$$

Integrating and incorporating constant of integration

$$N(\theta) = A_{10}\ln\left(\tan\frac{\theta}{2}\right) + A_{20} \tag{10.9}$$

Equation 10.9 becomes undefined for $\theta = \pi$. But this is not feasible in the given system as potential must be a continuous function. Therefore, A_{10} should be zero in Equation 10.9. Therefore,

$$N(\theta) = A_{20} \tag{10.10}$$

Then from Equations 10.4, 10.8 and 10.10, the zero separation constant solution can be obtained as

$$V(r,\theta) = \frac{C_1}{r} + C_2 \tag{10.11}$$

where:
$C_1 = A_{20}C_{10}$ and $C_2 = A_{20}C_{20}$

Determination of the equal and opposite separation constant solution:
The first term of Equation 10.6 is

$$\frac{1}{M(r)}\frac{d}{dr}\left[r^2\frac{dM(r)}{dr}\right] = +p$$

$$\text{or, } \frac{d}{dr}\left[r^2\frac{dM(r)}{dr}\right] = +pM(r) \tag{10.12}$$

Putting $M(r) = Cr^n$ in Equation 10.12, we get

$$\frac{d}{dr}\left(r^2Cnr^{n-1}\right) = +pCr^n, \text{ or, } Cn(n+1)r^n = pCr^n, \text{ or, } n^2 + n - p = 0$$

Hence,

$$n = \frac{1}{2}\left(-1 \pm \sqrt{1+4p}\right) \tag{10.13}$$

The second term of Equation 10.6 is

$$\frac{1}{N(\theta)}\frac{1}{\sin\theta}\frac{d}{d\theta}\left[\sin\theta\frac{dN(\theta)}{d\theta}\right] = -p$$

$$\text{or, } \frac{1}{\sin\theta}\frac{d}{d\theta}\left[\sin\theta\frac{dN(\theta)}{d\theta}\right] = -pN(\theta) \tag{10.14}$$

Putting $N(\theta) = B\cos\theta$ in Equation 10.14,

$$\frac{d}{d\theta}\left(-B\sin^2\theta\right) = -pB\cos\theta\sin\theta, \text{ or, } p = 2.$$

Hence, from Equation 10.13 $n = +1, -2$.

$$\text{Therefore, } M(r) = C'r + \frac{C''}{r^{-2}} \text{ and } N(\theta) = B\cos\theta \tag{10.15}$$

From Equations 10.4 and 10.15,

$$V(r,\theta) = \left(C_3 r + \frac{C_4}{r^2}\right)\cos\theta \tag{10.16}$$

where:

$C_3 = C'B$ and $C_4 = C''B$

The complete solution for potential function is uniquely given as a linear combination of the two solutions given by Equations 10.11 and 10.16.

$$V(r,\theta) = \frac{C_1}{r} + C_2 + \left(C_3 r + \frac{C_4}{r^2}\right)\cos\theta \tag{10.17}$$

where:

The constants are determined by satisfying the boundary conditions

It is evident from Equation 10.17 that the first term corresponds to a net charge on the sphere and the second term to a finite potential.

10.2.1 Conducting Sphere in Uniform Field

Consider that the sphere is a conducting one and is isolated and uncharged. Further, consider that the potential at the location of the centre of the sphere due to the external field is V_0.

Because the perturbing action of the sphere is negligible at a large distance from the sphere, the potential at a large distance from the sphere ($r >> a$) is given by

$$V(r,\theta) = V_0 + E_0 r\cos\theta \tag{10.18}$$

If the sphere is charged with a finite amount of charge Q, then

$$V(r,\theta) = \frac{Q}{4\pi\varepsilon_0 r} + V_0 + E_0 r\cos\theta \tag{10.19}$$

In practical systems, floating metallic particles are usually not charged and hence Equation 10.18 is taken here for further discussion.

Comparing Equations 10.17 and 10.18 for $r \to \infty$, $C_2 = V_0$ and $C_3 = E_0$. C_1 will be zero for uncharged sphere.

Therefore, Equation 10.17 can be rewritten as

$$V(r,\theta) = V_0 + \left(E_0 r + \frac{C_4}{r^2} \right) \cos\theta \tag{10.20}$$

On the conductor surface, that is, for $r = a$,

$$V(a,\theta) = V_0 + \left(E_0 a + \frac{C_4}{a^2} \right) \cos\theta \tag{10.21}$$

But the conducting sphere surface is an equipotential and hence electric potential is independent of θ on the conductor surface.

Therefore, from Equation 10.21,

$$C_4 = -E_0 a^3$$

Hence, the complete solution for an electric potential in the domain $r > a$ is given by

$$V(r,\theta) = V_0 + E_0 \left(r - \frac{a^3}{r^2} \right) \cos\theta \tag{10.22}$$

The r and θ components of the electric field intensity could be obtained as follows:

$$E_{\mathrm{r}} = -\frac{\partial V}{\partial r} = -E_0 \left(1 + \frac{2a^3}{r^3} \right) \cos\theta \tag{10.23}$$

$$E_\theta = -\frac{1}{r}\frac{\partial V}{\partial \theta} = E_0 \left(1 - \frac{a^3}{r^3} \right) \sin\theta \tag{10.24}$$

On the conducting sphere surface, tangential component of the electric field intensity must be zero as it is an equipotential surface. Equation 10.24 shows that for $r = a$, E_θ is zero, which in turn validates the solution obtained.

Again, on the conducting sphere surface, E_{r} is the normal component of the electric field intensity, which is given by $E_{\mathrm{r}}\big|_{r=a} = -3E_0 \cos\theta$. Thus, the maximum value of the electric field intensity on the surface of the conducting sphere is $3E_0$, that is, three times the strength of the uniform external field.

This is the reason why metallic dust particles should be avoided at all costs for GISs. Because the presence of metallic dust particles will increase the local electric field intensity three times, which will result into partial discharge within the GIS, that is very detrimental for the GIS operation.

Induced surface charge density on the surface of the conducting sphere may be obtained as follows:

$$\frac{\sigma_s}{\varepsilon_0} = E_r\bigg|_{r=a} = -3E_0\cos\theta, \text{ or, } \sigma_s = -3\varepsilon_0 E_0\cos\theta \tag{10.25}$$

As stated earlier, the sphere may be charged with an additional charge Q, which is distributed uniformly on the sphere surface and its effect on the field could be found by superposition.

10.2.2 Dielectric Sphere in Uniform Field

In the case of dielectric sphere present in uniform external field, there will be two solutions to potential function, V_i valid for the region within the sphere having dielectric of permittivity ε_i and V_e valid for the region outside the sphere having dielectric of permittivity ε_e. Therefore, from Equation 10.17

$$V_i(r,\theta) = \frac{C_{1i}}{r} + C_{2i} + \left(C_{3i}r + \frac{C_{4i}}{r^2}\right)\cos\theta \tag{10.26}$$

and

$$V_e(r,\theta) = \frac{C_{1e}}{r} + C_{2e} + \left(C_{3e}r + \frac{C_{4e}}{r^2}\right)\cos\theta \tag{10.27}$$

The potential at large distance r ($r >> a$) from the sphere

$$V(r,\theta) = V_0 + E_0 r\cos\theta \tag{10.28}$$

where:

V_0 is the potential at the location of the centre of the sphere due to the external field

Comparing Equations 10.27 and 10.28 for $r \to \infty$, $C_{2e} = V_0$ and $C_{3e} = E_0$. C_{1e} will be zero as a dielectric sphere is not considered to have any free charge.

Hence, Equation 10.27 can be rewritten as

$$V_e(r,\theta) = V_0 + \left(E_0 r + \frac{C_{4e}}{r^2}\right)\cos\theta \tag{10.29}$$

Inside the dielectric sphere electric potential must be finite at all the points. Hence, from Equation 10.26 $C_{1i} = C_{4i} = 0$, $C_{2i} = V_0$. Hence, Equation 10.26 can be rewritten as

$$V_i(r,\theta) = V_0 + C_{3i}\, r\cos\theta \tag{10.30}$$

At $r = a$, both Equations 10.29 and 10.30 should yield the same electric potential. Therefore,

)*+,)))-) C 9 & 6
$A 0 " 6"
' D(

5.2. La trigonometria.

Pur essendo maturata entro i confini dell'astronomia, nella scienza araba la trigonometria ha rappresentato il crocevia di diverse attività matematiche. L'astronomia aveva conosciuto un importante sviluppo a partire fondamentalmente dal patrimonio indiano e greco. In principio tale sviluppo era stato favorito dalle necessità della pratica religiosa e da alcune commissioni statali, come il finanziamento delle traduzioni voluto dai califfi al-Mansūr, al-Mahdī e Hārūn ar-Rashīd e l'avvio di un esteso programma di osservazioni e misurazioni da parte di al-Ma'mūn.

Tra le varie espressioni dell'astronomia, a stimolare lo sviluppo degli strumenti della trigonometria sono stati in particolare lo studio del movimento dei pianeti, il calcolo del tempo, la concezione di strumento e soprattutto la realizzazione di centinaia di tabelle per usi di ogni tipo. Queste tavole (chiamate *zījs*) contenevano, in particolare, i calendari adattati ai bisogni di questa o quella comunità in funzione delle sue credenze, attività e modi di vivere. In base a determinati calcoli, le *zījs* stabilivano gli orari delle preghiere, le latitudini e longitudini delle città, la direzione della Mecca a partire da qualunque città dell'impero, previsioni sull'apparizione dei quarti di luna, ecc. A fianco di questa quantità di informazioni direttamente utilizzabili vi era una parte matematica, che comprendeva le tavole trigonometriche e i metodi di calcolo che permettevano di determinare il movimento degli astri e di prevedere fenomeni di vario genere, come per esempio le eclissi [Kennedy 1966; King 1973; Juškevič 1976, pp. 141-50].

Prendendo le mosse dai primi problemi trigonometrici posti dall'*Almagesto* di Tolomeo, che poggiavano sul teorema di Menelao, alcuni astronomi arabi preferivano però utilizzare per i loro calcoli espliciti la piccola tavola di seni contenuta nei *Siddhānta* indiani, che si rivelava, ai fini dei calcoli, piú economica della tavola delle corde dell'angolo doppio elaborata dalla tradizione greca. In un secondo momento gli stessi astronomi hanno considerevolmente accresciuto il volume della tavola del seno (raffinando le approssimazioni) e creato tavole del seno verso e del coseno. Poi hanno introdotto nuove linee trigonometriche come la tangente, la cotangente, la secante e la cosecante, alcune delle quali sono state utilizzate per la prima volta da Ḥabash al-Ḥāsib nella seconda metà del IX secolo. Nel medesimo periodo è stata realizzata anche la loro tabulazione, come si può vedere nell'*Iṣlāḥ al-Mājisṭī* (La revisione dell'Almagesto) di al-Battānī (m. 929) [Debarnot 1997; Nallino 1899-1907; Djebbar 2004].

Teorema di Menelao. Se CE, CF, DF, DB sono archi di cerchi massimi su una sfera, si avrà:

(i)
$$\frac{\mathrm{crd}\,(2CB)}{\mathrm{crd}\,(2BF)} = \frac{\mathrm{crd}\,(2CA)}{\mathrm{crd}\,(2AE)} \times \frac{\mathrm{crd}\,(2ED)}{\mathrm{crd}\,(2DF)}$$

(formulazione basata sulle corde dell'angolo doppio);

(ii)
$$\frac{\mathrm{sen}\,(CB)}{\mathrm{sen}\,(BF)} = \frac{\mathrm{sen}\,(CA)}{\mathrm{sen}\,(AE)} \times \frac{\mathrm{sen}\,(ED)}{\mathrm{sen}\,(DF)}$$

(formulazione basata sui seni).

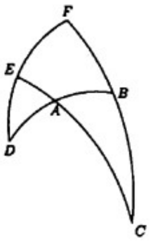

Una terza tappa è stata raggiunta nella prima metà del x secolo con la definizione delle relazioni fondamentali esistenti tra queste sei funzioni trigonometriche. Successivamente, in uno sforzo di ottimizzazione dei calcoli, sono state intraprese alcune ricerche per trovare nuove relazioni trigonometriche. Tali ricerche sono sfociate nella definizione di nuovi teoremi e di nuove formule piú facili da utilizzare. La piú celebre è quella a cui verrà dato il nome di «formula che rende superfluo». Di fatto si tratta della versione sferica del teorema del seno, ed è stata denominata in questo modo perché, facendo intervenire soltanto quattro grandezze (mentre quella di Menelao ne utilizza sei), riduceva considerevolmente i tempi di calcolo per la creazione delle tavole [Debarnot 1985].

Teorema del seno. Se a, b, g sono archi di cerchi massimi e A, B, G sono i rispettivi angoli opposti, si avrà:

$$\frac{\mathrm{sen}\,(a)}{\mathrm{sen}\,(A)} = \frac{\mathrm{sen}\,(b)}{\mathrm{sen}\,(B)} = \frac{\mathrm{sen}\,(g)}{\mathrm{sen}\,(G)}$$

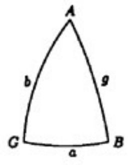

Tra i ricercatori che hanno contribuito a questo progresso, o che hanno rapidamente compreso l'importanza del teorema integrandolo nelle loro opere, figurano Abū 'l-Wafā', Ibn 'Irāq (XI secolo) e al-Bīrūnī in Oriente, Ibn Mu'ādh (XI secolo) e Jābir ibn Aflaḥ (m. 1145) in Andalusia [Villuendas 1979; Lorch 1975]. Questa ripartizione geografica rivela in particolare l'uniformazione del livello scientifico che aveva avuto luogo nei paesi dell'Islam quale risultato di una grande circolazione delle idee attraverso i libri, gli scambi epistolari e gli spostamenti degli stessi uomini di scienza, come Ibn Mu'ādh, trasferitosi al Cairo.

Il superamento della quarta e ultima tappa nella storia della trigonometria araba ha avuto inizio con la comparsa nei trattati di astronomia di capitoli specificamente dedicati agli oggetti e agli strumenti della nuova disciplina. È questo il caso, per esempio, del *Risāla fī qusiyy al-kura* (Epistola sugli archi della sfera) di Ibn 'Irāq o del *Kitāb al-Mājisṭī* (Libro dell'Almagesto) di Abū 'l-Wafā'. Il processo si è compiuto con la pubblicazione di opere interamente consacrate a questa materia, come il *Kitāb majhūlāt quisiyy al-kura* (Libro degli archi non noti della sfera) di Ibn Mu'ādh e, soprattutto, il *Kitāb ash-ṣakl al-qaṭṭā'* (Libro della figura secante) di Naṣīr ad-Dīn aṭ-Ṭūsī [Caratheodory 1891; Villuendas 1979].

5.3. L'analisi combinatoria.

È in un'opera maghrebina del XII secolo, il *Fiqh al-ḥisāb* (La scienza del calcolo) di Ibn Mun'im (m. 1228), che a quanto ci risulta compare per la prima volta nella storia della matematica un capitolo dedicato in modo specifico a un argomento puramente combinatorio [Djebbar 1985]. Ma questo evento è stato preceduto dall'esistenza di diverse pratiche di enumerazione e computazione nell'ambito di svariati campi, a volte molto lontani gli uni dagli altri, come l'astronomia, l'algebra e la linguistica.

In astronomia, processi combinatori elementari che utilizzano il computo ottenuto per mezzo di semplice enumerazione intervengono in un opuscolo di Thābit ibn Qurra e in un trattato di al-Bīrūnī [Djebbar 1980, pp. 56-60]. In algebra, procedimenti analoghi si trovano in Abū Kāmil e as-Samaw'al per il computo delle equazioni o delle soluzioni di un problema dato. Nel *Kitāb aṭ-ṭarā'if fī 'l-ḥisāb* (Libro delle rarità nel calcolo), il primo di questi studiosi effetti, la trasformazione che si andava attuando del sistema di mecenatismo individuale in un sistema di ricerca collettiva, ben presto finanziato dagli Stati, esercitò una forte pressione sugli orientamenti della ricerca e sulla scelta dei problemi da studiare. Mordechai Feingold [1984, p. 216] ha mostrato che in Inghilterra, fra il 1560 e il 1640, l'attività scientifica tradizionalmente distribuita in vari settori «venne progressivamente indirizzata in ricerche influenzate in larga parte dalle esigenze e dalle richieste dei mecenati. Il risultato immediato di questo reindirizzamento fu la proliferazione di progetti di carattere pratico che da una parte assicurarono all'Inghilterra un posto fra le nazioni piú potenti e accrebbero in maniera considerevole le fortune dei loro committenti, dall'altra portarono alla creazione di meccanismi e artifici destinati a dilettare le classi superiori». Presteremo dunque una particolare attenzione alle possibili influenze – legate ai meccanismi propri del mecenatismo – esercitate sugli interessi matematici dei protagonisti.

Sullo sfondo della nostra analisi terremo presente anche l'ipotesi formulata da Mario Biagioli nell'epilogo del suo *Galileo, Courtier*, intitolato *From Patronage to Academies. A Hypothesis* [Biagioli 1993, pp. 353-62]. Biagioli sostiene che «la comprensione del processo di legittimazione sociale e cognitiva offerto dalla Corte offre importanti elementi per definire la genealogia del sistema sociale che venne successivamente a crearsi, del suo ethos e dello stile della filosofia naturale» [p. 354]. La questione delle nuove pratiche scientifiche generate dal passaggio dal mecenatismo alle istituzioni scientifiche sarà uno dei cardini della nostra riflessione, tenendo presente che tali pratiche includevano: i protocolli di legittimazione e dunque di reclutamento; i codici di comportamento, soprattutto nel caso di controversie; le ragioni soggiacenti alla scelta dei temi di ricerca; gli approcci adottati nell'affrontare i problemi e altre analoghe questioni.

2. *Leibniz, consigliere dei principi.*

2.1. Per la gloria e il diletto di un principe.

Quando Leibniz, giurista di formazione e futuro diplomatico di professione, alla fine dei suoi studi universitari, si trovò di fronte alla necessità di guadagnarsi da vivere, scelse di mettersi al servizio di un potente, sebbene gli fosse stata offerta una cattedra all'Università di Altdorf nel 1667. Ricordiamo che Leibniz (nato a Lipsia il 1° luglio 1646, due anni prima che la pace di Westfalia mettesse fine alle guerre di religione che avevano devastato la Germania da cima a fondo) aveva studiato filosofia e diritto a Lipsia e a Jena. Perché dunque rifiutare una tranquilla carriera universitaria come quella di suo padre, Friedrich Leibnütz? Certamente il giovane Leibniz nutriva l'ambizione di influenzare la politica del suo paese e pensava di realizzarla diventando consigliere di uno dei suoi principi. In effetti è questo il concetto che esprime nel dicembre 1675 in una lettera indirizzata a Johann Carl Kahm, valletto di camera del duca Giovanni Federico di Brunswick-Lüneburg:

> Per natura io non sono avido di denaro, ma, una volta che possa contare su un reddito soddisfacente, cerco soltanto di accrescere il mio prestigio, lavorando e svolgendo dei servizi per il bene comune [...]: è soprattutto per questo motivo che desideravo trovare un'alta personalità [...] la cui protezione, l'autorità, l'aiuto e i favori potessero dare corpo a idee utili; in cambio io non avrei altra preoccupazione che lavorare per la sua gloria e per il suo diletto [Heinekamp et al. 1994, pp. 43-44].

Leibniz trovò questo patrono nella persona di Johann Christian von Boineburg, ministro del principe elettore di Magonza, Giovanni Filippo di Schönborn. Il giovane Leibniz cominciò dunque la sua carriera di giurista e diplomatico a Magonza – dove sarà poi segretario, bibliotecario e consigliere di Corte d'appello – redigendo rapporti e *memoranda* per Boineburg. Nel marzo 1672, accompagnò il nipote del principe elettore in missione diplomatica a Parigi, ma la morte di Boineburg e successivamente quella di Schönborn misero fine alla sua promettente carriera. Leibniz cercò allora di stabilirsi a Parigi per potersi introdurre nei circoli eruditi della capitale e ottenere uno stipendio dall'Académie des Sciences, senza successo: «Parigi è un posto dove è difficile emergere: si trovano, in tutti i campi scientifici, gli uomini piú brillanti dell'epoca e serve molto tempo per costruirsi una solida reputazione» [*ibid.*, p. 46]. In ottobre, lasciò Parigi per mettersi al servizio di Giovanni Federico di Brunswick-Lüneburg, con il titolo di consigliere della Corte ducale.

2.2. Leibniz a Parigi: un periodo di libertà creativa.

Il soggiorno parigino fu, per il matematico Leibniz, uno dei periodi piú fruttuosi. Fu l'occasione per incontrare studiosi della statura di Christiaan Huygens, membro dell'Académie des Sciences, conoscenza che gli fece capire quanto fosse vasta la sua ignoranza in matematica; Edme Mariotte; l'ormai anziano Roberval o il giovane Ehrenfried Walter Tschirnaus. Grazie alla loro frequentazione, poté assimilare la matematica del suo tempo e fare ampie letture nelle ricche biblioteche parigine. In questo periodo di piena maturità intellettuale, Leibniz riuscí a impadronirsi da autodidatta – in soli quattro anni e con rara intensità – di quelle tecniche matematiche che seppe poi rapidamente impiegare in maniera del tutto originale. Fece la conoscenza di Étienne Périer, nipote di Pascal, che gli permise di consultare i manoscritti dello zio, da cui Leibniz trasse il famoso *Traité des coniques*.

Nell'ottobre del 1675, inventò il calcolo differenziale mentre era impegnato nella risoluzione di un problema inverso di tangenti (cioè un problema di quadratura) e stava cercando di presentare nella maniera piú semplice e chiara il suo risultato. In seguito, nel corso di queste ricerche, Leibniz introdusse i suoi famosi simboli, una «S» stilizzata per la somma (o l'integrale) e una «d» per la differenza. Come commenta Josef E. Hofmann [1974, p. 194]: «dal momento in cui furono compiuti questi primi passi cruciali verso l'algebrizzazione dei problemi infinitesimali, si aprí una nuova prospettiva, per quest'uomo abituato a identificare gli elementi generali e caratterizzanti in mezzo a un'accozzaglia di cose simili». Leibniz pubblicò il suo nuovo calcolo solo nel 1684 negli *Acta Eruditorum* di Lipsia. La storia di questa invenzione e dei suoi sviluppi è ben nota e rinviamo chi desiderasse approfondirla alla vasta letteratura esistente.

2.3. Leibniz al servizio degli Hannover.

Leibniz si stabilí a Hannover, dove rimase per il resto dei suoi giorni (morí il 14 novembre 1716, all'età di 70 anni), nonostante si lamentasse spesso del provincialismo della cittadina, della mancanza di «eruditi» con cui discutere e della totale assenza di circoli intellettuali in cui si sviluppasse una qualche ricerca. Del resto, scrive a Thomas Burnett in una famosa lettera del 17 marzo 1696, «la vita di Corte e le conversazioni dotte si escludono vicendevolmente» [Heinekamp et al. 1994, p. 51]. Leibniz lavorò successivamente per tre esponenti della casa di Hannover: dal 1676 al 1679 serví Giovanni Federico, duca di Brunswick-Lüneburg; dal 1679 al 1698 Ernesto Augusto (fratello di Giovanni Federico) e suo figlio Giorgio Ludovico, il futuro Giorgio I d'Inghilterra. Le relazioni che Leibniz stabilí con ognuno dei suoi tre protettori furono diverse: confidenziali con il primo, che apprezzava le conversazioni filosofiche e le opinioni del suo protetto; piú formali con Ernst August che, affidando a Leibniz l'incarico di redigere la storia dei Guelfi, lo coinvolse nel suo progetto di glorificazione della casa di Hannover, progetto poi trascurato da Giorgio Ludovico, il quale comunque rimproverò Leibniz di non aver portato a termine i suoi *Annali* e lo mortificò continuamente (rifiutandogli, per l'esempio, l'autorizzazione a viaggiare).

I suoi protettori non si accontentarono di trarre vantaggio dai consigli giuridici dispensati da Leibniz, dalle sue ricerche storiche o dalla sua attività di collezionista di libri e bibliotecario, ma gli chiesero anche, secondo la piú classica tradizione del mecenatismo rinascimentale, di intervenire nella risoluzione di problemi tecnologici e nella pianificazione del territorio. Johann Friedrich affidò infatti a Leibniz il compito di migliorare le difficili tecniche di estrazione dell'argento dalle miniere dello Harz. Fedele al proprio motto «Theoria cum praxi», Leibniz propose di utilizzare l'energia eolica per drenare l'acqua dai pozzi e dalle gallerie delle miniere. Il progetto venne portato avanti per diversi anni, nel corso dei quali egli sperimentò sul campo diverse tecniche, sviluppò i calcoli necessari e cercò di prendere accordi con gli amministratori delle miniere che erano ostili alle innovazioni tecnologiche, finché l'esperienza venne bruscamente interrotta da Ernesto Augusto, che nel frattempo era succeduto al fratello, nel 1685. Poiché il nuovo duca aveva deciso di stabilire la propria residenza estiva a Herrenhausen, dove voleva creare un giardino alla francese con fontane e giochi d'acqua, si rivolse a Leibniz, il quale nel 1696 preparò un progetto di cui vennero messi in opera solo alcuni importanti elementi – il progetto infatti non venne mai realizzato nel suo complesso – che si possono ammirare ancora oggi.

Gli studi leibniziani nel campo delle matematiche erano destinati a rimanere di dominio privato. I 7300 manoscritti matematici, per la maggior parte inediti, conservati a Hannover e i carteggi fra Leibniz e numerosi matematici dell'epoca, testimoniano l'ampiezza e la varietà di queste ricerche. Alcuni carteggi, come quello con Johann Bernoulli, documentano l'attività di un vero e proprio laboratorio di matematiche, nel quale si discute, oltre che di problemi scientifici, anche dell'organizzazione sociale e dell'istituzionalizzazione della disciplina: posti, strategie di conquista di cattedre, controversie e risposte da opporre, ecc. La matematica offriva solo raramente materia di discussioni dotte alla Corte degli Hannover, anche se in un'occasione Leibniz riuscí a destare l'interesse di Rodolfo Augusto, duca di Brunswick-Wolfenbüttel, per una delle sue invenzioni matematiche.

2.4. Il sistema binario presentato al duca di Brunswick-Wolfenbüttel.

Mentre era bibliotecario della Bibliotheca Augusta di Wolfenbüttel (a partire dal 1691), Leibniz inviò il 2 gennaio 1697 i suoi auguri per l'anno nuovo a Rodolfo Augusto, in ossequio all'uso comune, come afferma egli stesso, ma anche per l'affetto che nutriva verso il duca. In quella lettera[1] spiega al duca il calcolo binario inventato alcuni anni addietro e pubblicato nei «Mémoires de l'Académie Royale des Sciences» (1703) per giustificare la sua elezione come membro straniero[2]. In occasione di un incontro precedente all'invio della lettera, Leibniz aveva già avuto modo di illustrare la sua invenzione al duca, che aveva suggerito un'analogia fra il sistema binario e la Creazione [Aiton 1985]. Leibniz fece sua questa analogia e offrí al duca il bozzetto di una medaglia in cui essa veniva rappresentata in forma emblematica. Uno dei principali misteri della fede cristiana, diceva Leibniz, è la creazione *ex nihilo* di tutte le cose da parte dell'onnipotenza divina:

---

[1] Nella pagina web www.fh-augsburg.de/ harsch/germanica/Chronologie/17Jh/Leibniz/lei bina.html si trova la riproduzione e la trascrizione della lettera originale che abbiamo parafrasato nel testo. Si veda anche G. W. Leibniz, *Sämtliche Schriften und Briefe*, 8 Reihen, Akademie-Ausgabe, Berlin, I, 13 [d'ora in poi: *Akademie-Ausgabe*].

[2] Il diplomatico Christophe Brosseau, residente di Hannover a Parigi, aveva dovuto ricordare a Leibniz i suoi doveri di cittadino della *République des lettres*. Gli scriveva infatti il 16 agosto 1700 da Parigi: «Mi figuravo, Signore, e l'avevo anche fatto sperare al Signor Abate Bignon, che con la prima delle Vostre lettere avrei ricevuto le risposte che dovete non solo a lui ma anche a Monsieur de Fontenelle, che Vi hanno scritto con l'occasione di inviarVi il vostro appannaggio di accademico onorario. Purtroppo io queste risposte non le ho trovate affatto con la Vostra che mi è arrivata il giorno 3 di questo mese. Permettetemi di essere alquanto sorpreso, poiché sembrerebbe che Voi non stiate nemmeno pensando di scrivere loro. Tuttavia, Signore, l'onestà e la buona creanza Vi obbligano ad adempiere ai Vostri doveri, pertanto Vi supplico di non serbarmi rancore se Vi esorto e se oso dirVi che si tratta di un dovere dal quale non potete dispensarVi» [*Akademie-Ausgabe*, I, 18,

> Dunque, nulla rappresenta meglio la Creazione, anzi nulla la dimostra meglio dell'origine dei numeri cosí come viene spiegata qui: tutti i numeri vengono generati in base alla loro sola espressione formata da successioni di Uno e Zero, cioè niente. Sarà difficile poter trovare in natura o in filosofia un simbolo migliore per questo mistero ed è per questo motivo che ho iscritto sulla medaglia le parole: IMAGO CREATIONIS.

L'emblema avrebbe dovuto rendere manifesto che tutto quello che è stato creato, è stato perfettamente creato. Nella usuale rappresentazione dei numeri, non c'è nessun ordine che si palesi nella successione dei caratteri, contrariamente a ciò che avviene nel sistema binario, fondato su una regola che permette di scrivere i numeri senza alcun calcolo e tanto grandi quanto si vuole: per ottenere la rappresentazione binaria dei numeri interi disposti in una colonna verticale[3], basta scrivere nella colonna di destra alternativamente: 0, 1, 0, 1, 0, 1, ecc.; nella colonna successiva, partendo da destra, si scriverà: 0, 0, 1, 1, 0, 0, 1, 1, ecc.; nella terza colonna: 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, ecc.; nella quarta 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, ecc. Lo spazio centrale della medaglia è occupato da una piccola tabella che presenta la scrittura binaria dei numeri interi da 1 a 17 e mette in evidenza questo bell'ordine univoco. Leibniz ne deduceva che

> il disordine che si immagina nelle opere di Dio è solo apparente; non appena si osservano le cose da un giusto punto di vista, come nella prospettiva, si coglie la loro simmetria. Questo ci spingerà sempre piú a lodare e ad amare la saggezza, la bontà e la bellezza del Bene Supremo, da cui deriva tutta la bontà e tutta la bellezza.

Nel prosieguo della lettera, Leibniz menzionava Claudio Filippo Grimaldi, missionario gesuita in Cina, che aveva conosciuto a Roma e che gli aveva scritto da Goa in occasione del suo viaggio di ritorno in Cina; Leibniz aveva ritenuto opportuno illustrargli la rappresentazione binaria dei numeri nella speranza che l'emblema raffigurante il mistero della Creazione potesse servire a rendere manifesta la superiorità della fede cristiana, dal momento

---

pp. 804-5]. Leibniz fece pervenire a Fontenelle il 26 febbraio 1701 il trattato dal titolo *Essay d'une nouvelle science des nombres* che non volle però veder pubblicato; al suo posto, inserí la memoria (G. W. Leibniz, *Mathematische Schriften*, a cura di C. I. Gerhardt, 7 voll., Halle, Berlin 1849-63, nuova edizione Olms, Hildesheim 1960-61, vol. VII, pp. 223-27 [d'ora in poi: *Math. Schriften*]).

[3] Ricordiamo che nel sistema binario un numero intero si scrive come somma di potenze di 2 moltiplicate per i coefficienti 0 o 1. Cosí, per esempio, $3 = 1 \times 2^1 + 1 \times 2^0$ si scrive 11; $8 = 1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 0 \times 2^0$ si scrive 1000

che Grimaldi gli aveva raccontato che l'imperatore cinese era un grande estimatore dell'arte di calcolare, anche al modo europeo, che aveva appreso da padre Ferdinand Verbiest, predecessore di Grimaldi.

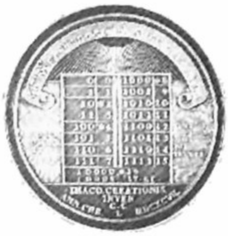

Leibniz continuava poi descrivendo gli altri elementi della medaglia e spiegando l'addizione e la moltiplicazione dei numeri binari, sebbene ritenesse che non fosse il caso di usare questi numeri nella vita quotidiana. Per fare in modo che la medaglia avesse un aspetto gradevole e la sua superficie non fosse completamente riempita di numeri, Leibniz pensava di rappresentarvi la luce e le tenebre, dal momento che le tenebre ricoprivano l'abisso e lo Spirito di Dio aleggiava sulla superficie delle acque. Allora Dio disse: «Sia la luce, e la luce fu». L'abisso e le tenebre sono assimilabili allo Zero e al nulla, lo spirito di Dio con la sua luce all'Uno onnipotente. Come motto dell'impresa Leibniz propone i versi seguenti, che mostrano bene per quale motivo l'emblema è IMAGO CREATIONIS:

2, 3, 4, 5 etc. 0
Omnibus ex nihilo ducendis, SUFFICIT UNUM

Ovvero: per dedurre il tutto dal nulla, basta l'Uno.

Al termine della lettera, Leibniz osservava che il sistema binario racchiude in sé una tale quantità di osservazioni meravigliose e utili al progresso delle scienze che la Società matematica di Amburgo avrebbe dovuto esercitare la sua intelligenza per sviluppare quello che lui, Leibniz, assicurava di poter trovare e che avrebbe contribuito alla loro gloria e a quella della nazione tedesca. Non è questa l'unica occasione in cui Leibniz si rivolge a un'istituzione per invitarla a impegnarsi in determinate ricerche. In effetti, aveva suggerito a Christophe Sturm, membro della Deutsche Akademie der Naturforscher, di effettuare in diversi luoghi dei rilievi della declinazione magnetica e di ripetere le misurazioni a intervalli di tempo regolari. A questo invito non ne erano seguiti altri fino al tentativo di coinvolgimento della Società matematica di Amburgo [Heinekamp et al. 1994, p. 85].

La straordinaria lettera che abbiamo appena parafrasato riassume da sola diverse ossessioni di Leibniz. Egli offre al suo protettore un'invenzione matematica celebrata da una medaglia che, allo stesso tempo, avrebbe glorificato il duca. Questa invenzione racchiude potenzialmente molti risultati nuovi che sarebbe opportuno mettere a disposizione di una comunità scientifica, in questo caso una delle piú antiche società matematiche d'Europa. Peraltro, raffigurando l'analogia, suggerita dal duca e ripresa da Leibniz, fra numerazione binaria e Creazione, la medaglia avrebbe potuto inoltre servire all'attività di proselitismo portata avanti dai Gesuiti in Cina, dal momento che incoraggiava i Cinesi ad accogliere la matematica europea assieme alla fede cristiana.

Forte di questa convinzione, il 15 febbraio 1701, Leibniz descrisse la sua invenzione – principî di aritmetica binaria e analogia con la Genesi – al missionario gesuita Joachim Bouvet, che riconobbe un legame (puramente formale) fra i numeri binari e i cosiddetti «esagrammi di Fohy», ovvero i diagrammi contenuti nel *Libro dei mutamenti* (*I Ching*). Si puo supporre che Leibniz abbia incluso questa citazione nella memoria che stava redigendo per l'Accademia parigina *Explication de l'arithmétique binaire [...]* per effetto dell'entusiasmo che la lettera di Bouvet gli aveva provocato [*Math. Schriften*, VII, pp. 223-27]. Non era piú questione di medaglia, né di analogia con la Creazione del mondo a partire dal nulla, ma di «perfezione della scienza dei Numeri» in termini spesso assimilabili a quelli impiegati nella lettera al duca Rodolfo Augusto. Anzi, nel titolo completo del trattato, *Explication de l'arithmétique binaire, qui se sert des seuls caractères 0 et 1 avec des remarques sur son utilité et sur ce qu'elle donne le sens des anciennes figures chinoises de Fohy*, si mettono generosamente in evidenza le «figure cinesi di Fohy» (*Fu Xi*), le quali secondo Leibniz (che in questo segue Bouvet) «sono forse il piú antico monumento scientifico che esista al mondo» [*ibid.*, pp. 226-27] e sarebbero identiche alla tavola binaria di Leibniz[4]. Conclude la sua breve memoria avanzando però il sospetto che ci fosse «nella scrittura cinese un vantaggio simile a quello che esiste necessariamente in una Caratteristica che sto progettando. Ogni ragionamento che si può dedurre dalle nozioni potrebbe essere dedotto dai loro Caratteri attraverso una specie di calcolo e questo sarebbe uno dei piú importanti mezzi per aiutare lo spirito umano» [*ibid.*, p. 227].

2.5. La storia dei Guelfi, un pretesto per viaggi scientifici.

Nel 1685 Leibniz aveva ricevuto l'incarico di redigere la storia della casa dei Guelfi e per questo gli era stato concesso un vitalizio. Nelle intenzioni di Ernesto Augusto, questa *Historia domus* era concepita come una genealogia del suo casato, finalizzata a far valere i propri diritti per l'Elettorato, ottenuto poi nel 1692. Per Leibniz, invece, si trattava di scrivere la storia del Casato del Brunswick-Lüneburg basandosi su fonti storiche attendibili e per fare questo mise insieme un impressionante *corpus* di documenti (pubblicati poi nel XVIII secolo) ma non portò mai a termine la sua storia dei Guelfi, subendo cosí vivaci rimproveri da parte del duca Giorgio Ludovico. La strategia che mise in atto per reperire le fonti storiche necessarie si articolò in due fasi: dapprima provvide ad ampliare la sua rete epistolare e successivamente si mise a viaggiare per visitare archivi e biblioteche. Dal novembre 1687 al giugno 1690 intraprese dunque un viaggio che gli avrebbe dato l'occasione per stabilire nuovi contatti con vari studiosi. André Robinet [1988] ha ampiamente documentato queste nuove relazioni, e ha descritto il soggiorno romano di Leibniz – dalla metà dell'aprile fino al 21 novembre 1689 – come un «felice momento creativo». In effetti, Leibniz venne ammesso all'Accademia fisico-matematica, i cui membri si riunivano tutte le sere nel palazzo di Giovanni Ciampini situato dietro la chiesa di Sant'Agnese. Secondo Robinet, è proprio a Roma che Leibniz avrebbe iniziato a redigere la sua *Dynamica*, i cui appunti preparatori si erano mescolati a quelli di un altro testo, il *Phoranomus*, rimasto inedito fino alla pubblicazione curata da Robinet nel 1991. Quest'ultimo documento sarebbe, secondo l'interpretazione originale e affascinante di Robinet, il rendiconto, o meglio la messa in scena, di una seduta dell'Accademia romana, nel corso della quale Leibniz avrebbe presentato per la prima volta in pubblico la sua dinamica (sotto forma di *Phoranomus*) legando «non soltanto il ricordo di Parigi al presente di Roma, ma anche al futuro dell'Accademia di Pechino» [Robinet 1988, p. 95].

---

[4] Questa interpretazione si fonda tuttavia su un malinteso. Sulla questione si veda Aiton [1985, p. 248].

In effetti, Leibniz comincia l'opera come se fosse una lettera indirizzata a Melchisédech Thévenot, membro dell'Académie des Sciences di Parigi, e si riferisce a un libello che Pierre-Daniel Huet, vescovo di Soissons, aveva appena pubblicato contro la filosofia di Descartes. In seguito, dopo aver divagato (§ 2) sulle antichità romane, Leibniz fa intervenire nella *pièce* lo stesso padre Grimaldi, di cui parlerà al duca di Wolfenbüttel nel 1697. Il gesuita era appena arrivato a Roma di ritorno dalla Cina ed era stato interpellato a proposito della possibilità di succedere a padre Verbiest come presidente del Tribunale dei matematici di Pechino. Per questo motivo tutti gli occhi si volgono verso il nuovo venuto e alcuni gli fanno domande sulla Cina. Leibniz, in particolare, gli chiede informazioni sui caratteri cinesi, sui fuochi d'artificio, sull'astronomia, su una pianta capace di dare sollievo ai dolori della gotta e sulle officine tessili che si sarebbero sviluppate in Cina. Si passa poi a parlare delle macchine che i cinesi desidererebbero importare dall'Europa.

Leibniz ne approfitta per fare l'elogio della sua macchina aritmetica, che permette «anche a un bambino di fare molto rapidamente delle moltiplicazioni e delle divisioni molto complesse, senza nessuno sforzo e quasi distrattamente, come in un gioco»’ [Robinet 1991, § 4, p. 451]. Interrogato da Grimaldi sulla differenza fra la sua macchina e quella di Pascal, Leibniz cita la testimonianza di Périer, nipote di Pascal, che, in occasione della presentazione dell'invenzione all'Académie di Parigi ne aveva riconosciuto l'originalità. Successivamente egli spiega, con un esempio, le operazioni e i movimenti necessari per ottenere il risultato. Nel seguito, la discussione si sposta sulle matematiche pratiche e si aggiunge un nuovo arrivato, Domenico Quartaroni, professore alla Sapienza e animatore di un'Accademia matematica. Grimaldi indirizza la conversazione sui principî meno certi sui quali si fondano queste matematiche e ciò spinge Antonio Baldigiani, che era entrato in scena con Grimaldi ma non aveva ancora parlato, a rivolgersi a Leibniz (§ 6) chiedendogli provocatoriamente che cosa si aspetti da tali principî. È dunque a questo professore di matematiche del Collegio Romano che tocca porre la questione chiave a cui è dedicato il *Phoranomus*: in che cosa consiste la vostra nuova fisica? Leibniz si fa pregare fino a che non vede entrare Adrien Auzout, considerato come una vera e propria autorità in matematica, in presenza del quale si decide a spiegare la nuova fisica, che si fonda sul principio di conservazione delle forze vive (mv^2). Leibniz passa quindi ad affrontare il nodo della questione.

Questa prima parte del *Phoranomus*, in cui Leibniz sceglie di far conversare un certo numero di matematici, ci fa rivivere una seduta (forse immaginaria) dell'Accademia romana. Grimaldi interviene per primo, dal momento che la sua presenza a Roma è rara e i suoi racconti sono in grado di fornire informazioni inedite sulla Cina. In virtú dei riferimenti alla Cina e di quelli all'Accademia parigina, Leibniz sembra voler conferire una dimensione internazionale all'attività matematica che descrive. È esattamente ciò che cercò di realizzare impegnandosi attivamente per sostenere l'idea delle Accademie.

2.6. I progetti accademici di Leibniz.

A partire dal 1670 e fino al termine della sua vita, Leibniz si interessò all'organizzazione internazionale delle scienze e tentò di promuovere, con tutta l'energia di cui disponeva, la creazione di «società» o accademie scientifiche. Il progetto di una *Societas eruditorum Germaniae* che presentò nel 1670, appena ventiduenne, al principe elettore di Magonza era legato alla fondazione di un periodico, il «Nucleus Librarius Semestralis» – che non ottenne il privilegio imperiale –, in cui si sarebbero dovuti recensire tutti

---

’ «Tum ego, quid ipsi de Machina quadam Arithmetica videretur, qua ego effecissem, ut sine labore ac prope sine attentione animi maximae multiplicationes ac divisiones etiam a puello velut ludendo peragi promptissime possent». In un manoscritto del 1680, *Machina arithmeticae dyadicae*, Leibniz considera una macchina che trasforma i numeri decimali in numeri binari, ma non venne mai costruita, diversamente da quella che si fonda sul sistema decimale.

i libri pubblicati nell'impero, elogiandone il contenuto o, all'occorrenza, censurandolo. Regolare il mercato dei libri, scartando quelli ridondanti e superflui, costituire una biblioteca universale, inventariarne il sapere per mezzo di indici: questi dovevano essere i compiti principali della nuova Accademia che avrebbe dovuto aver sede a Francoforte sotto la presidenza del principe elettore in persona [Heinekamp et al. 1994, pp. 121-25].

Questo progetto, cosí come tutti gli altri che si sarebbero succeduti regolarmente negli anni successivi, non ottenne l'assenso delle autorità e quindi nemmeno finanziamenti. Ciò a cui mirava Leibniz, di fatto, era la creazione di istituzioni che si occupassero non solo di scienza, ma anche di economia e di cultura in genere, stabilendo poi dei rapporti con le attività pratiche. I risultati teorici ottenuti, in altre parole, avrebbero dovuto essere utilizzati al servizio del benessere dell'intera umanità. Quest'ottica universalista e filantropica andava ben oltre quella delle Accademie esistenti e le idee leibniziane si sarebbero concretizzate, anche se molto parzialmente, solo nel 1700 con l'istituzione dell'Accademia di Berlino, finanziata dai proventi del monopolio della produzione di calendari e almanacchi. Il 19 marzo 1700 l'elettore di Brandeburgo fece istituire un'Accademia delle scienze e un osservatorio a Berlino, nominando Leibniz presidente a vita. Lo statuto di questa Sozietät der Wissenschaften di Berlino fu ratificato l'11 luglio 1700 «in modo che l'anniversario di fondazione della società coincid[esse] con il genetliaco di Monsignore l'Elettore» [*Akademie-Ausgabe*, I, 18, p. 737]. La società venne ufficialmente inaugurata solo nel 1711, senza il suo presidente che aveva intanto trovato il tempo e le energie per pubblicare nel 1710 il primo volume dei *Miscellanea Berolinensia*, di cui aveva redatto un quinto delle memorie e dove i matematici rivestivano un ruolo non trascurabile: 8 memorie su 60, di cui tre di Leibniz, alcuni contributi di Johann Bernoulli, Jacob Hermann e Philippe Naudé (cui Leibniz aveva chiesto di collaborare allo sviluppo delle applicazioni della sua aritmetica binaria). Lo sviluppo della società venne ostacolato dai ricorrenti problemi di finanziamenti e non riuscí mai a raggiungere la forma «degna della grandezza e della gloria del suo fondatore» [*Akademie-Ausgabe*, I, 18, p. 448] che Leibniz avrebbe voluto offrirgli in dono. Dopo il 1713, con la presa di potere di Federico Guglielmo I, il *Soldatenkönig* (re-sergente), l'Accademia decadde miseramente, e solo le grandi ambizioni di Federico II, grande riformatore dello Stato prussiano e della sua vita culturale, poterono in seguito farla rivivere.

3. *I matematici e la nuova Accademia di San Pietroburgo.*

3.1. L'Accademia come elemento di modernizzazione dello Stato russo.

Negli ultimi anni del XVII secolo, Leibniz si era sempre più interessato alla Russia e al suo zar, sostenitore dell'importanza del ruolo delle scienze nel processo di modernizzazione del suo immenso impero. Fin dal 1696, egli aveva offerto i propri servigi a Pietro il Grande, che incontrò in tre diverse riprese. I rapporti con lo zar rappresentavano per Leibniz un'occasione per riflettere sul modo di creare, sistematicamente e partendo da zero, istituzioni per l'insegnamento e la ricerca. Fra gli specifici programmi di ricerca che si augurava di vedere sviluppati in Russia, si ritrova l'osservazione della declinazione magnetica, che aveva già suggerito senza successo a Sturm e alla Deutsche Akademie der Naturforscher.

Dopo la morte di Leibniz, lo zar si rivolse a Christian Wolff, allora professore di filosofia a Halle, perché diventasse il suo intermediario presso gli intellettuali europei. Il 18 marzo 1723 Wolff gli mandò il suo manuale di fisica, *Vernünfftige Gedancken von den Würckungen der Natur* (Idee ragionevoli sugli effetti della natura), accompagnato da una dedica: «[...] non si può aiutare un paese a svilupparsi, se non favorendo lo studio delle scienze fondamentali e soprattutto delle scienze matematiche e fisiche» [Nagel e Verdun 2005, p. 20], ma rifiutò il posto di consigliere che lo zar gli aveva offerto. Il bibliotecario dello zar viaggiò in Francia, in Germania, in Olanda e in Inghilterra, visitandone le rispettive istituzioni scientifiche, per riuscire a stabilire rapporti di collaborazione con San Pietroburgo. Alla fine, l'8 febbraio 1724, il Senato russo approvò il progetto di Accademia proposto dal medico di corte, Laurentius Blumentrost, che ne divenne il primo presidente. L'Accademia avrebbe dovuto contare su tre elementi costitutivi: l'istituto accademico di ricerca (ispirato al modello delle Accademie di Parigi e Berlino), l'università e il ginnasio per la formazione dei giovani – ma gli ultimi due conobbero un rapido declino. Un resoconto del progetto di questa Accademia venne ampiamente diffuso presso gli ambasciatori russi e gli studiosi europei i quali, in assenza di candidati russi, accorsero numerosi a Pietroburgo[6].

Al suo esordio, l'Accademia contava ventitre membri, sette dei quali erano matematici. Wolff aveva raccomandato Jacob Hermann, allievo di Jacob Bernoulli, pupillo di Leibniz e professore di matematica a Francoforte sull'Oder e anche Johann Bernoulli, il quale declinò l'invito, ma convinse due dei suoi figli, Daniel e Nicolaus (II), a mettersi al servizio dello zar. Il 9 dicembre 1724, Wolff scriveva all'ambasciatore russo in Prussia, il conte Golovkin: «L'affiliazione [di Daniel Bernoulli] all'Accademia imperiale mi sembra importante perché abbiamo bisogno di un uomo che coltivi la *Mathesis sublimior* senza distrarsi con altre attività e che sappia applicarla, all'occorrenza, alla natura e alle arti» [Nagel e Verdun 2005, p. 22]. Sfortunatamente Pietro il Grande, morto il 18 gennaio 1725, non poté vedere compiuta la realizzazione del suo progetto, che fu portato a termine con successo dalla sua vedova, Caterina I. Al termine dell'estate un certo numero di accademici si erano già stabiliti a San Pietroburgo e le sedute di lavoro si tenevano il martedí e il venerdí[7]. Christian Goldbach, che era arrivato, inatteso, in compagnia di Hermann, fu incaricato di redigere i verbali in latino (lo storico Gerhard Müller lo definí, pensando al *secrétaire perpetuel* dell'Académie parigina, il «Fontenelle latinista») e di occuparsi delle pubblicazioni e della corrispondenza, anche se al tempo stesso era membro della classe di matematica. Jacob Hermann ricoprí la cattedra di matematica, Nicolaus Bernoulli quella di meccanica, Daniel Bernoulli fu professore di fisiologia e di anatomia fino a che nel 1731, dopo la partenza di Hermann, ottenne la cattedra di matematica. Ogni giorno insegnava le applicazioni della matematica alla medicina dalle 7 alle 8 del mattino, mentre il fratello Nicolaus insegnava dalle 8 alle 9 la meccanica. Gli studenti, tuttavia, erano poco numerosi, dato che mancava un sistema scolastico in grado di fornire le conoscenze di base necessarie alla comprensione di corsi avanzati. Joseph-Nicolas Delisle, che nel 1722 venne personalmente invitato da Pietro I, assunse la direzione del Dipartimento di Geografia e di Astronomia.

---

[6] Durante il periodo 1725-99, l'Accademia di San Pietroburgo contava 111 membri ordinari, dei quali 55 provenivano dall'impero germanico e solo 35 dai territori russi [Nagel e Verdun 2005, p. 29].

3.2. Un discorso matematico per l'inaugurazione dell'Accademia di Pietroburgo.

La cerimonia d'inaugurazione, che avrebbe dovuto essere presieduta da Caterina I, era stata programmata in coincidenza con il giorno di santa Caterina d'Alessandria (il 25 novembre), ma ebbe effettivamente luogo piú di un mese dopo, il 7 gennaio 1726, in assenza dell'imperatrice. Una seconda seduta pubblica dell'Accademia si tenne il 12 agosto in presenza di Caterina, dei membri della famiglia imperiale, dei rappresentanti del Senato, del Santo Sinodo e della Corte, nonché degli ambasciatori dei paesi stranieri. Dopo un solenne discorso di benvenuto pronunciato in tedesco dal professore di antichità romane, Jacob Hermann presentò una dissertazione in latino sull'origine e lo sviluppo della geometria (pubblicata poi nel 1735 [Le Roy s.d.]). Hermann, convinto che i geometri potessero trovare nella storia nuovi stimoli per le loro ricerche, nella prima parte del suo discorso tracciò la storia della geometria dall'antichità al XVIII secolo. Egli distingue tre epoche principali: l'infanzia della geometria, che comincia con Talete e Pitagora per culminare con Euclide, Archimede e Apollonio; dopo uno *hiatus*, ovvero un periodo di ristagno della geometria, la seconda fase, indicata con *medium ævum*, si apre con Viète ed è dominata dalla figura di Descartes; la terza epoca, inaugurata dalla pubblicazione della *Nova methodus* (1684) di Leibniz e ancora in corso, è quella della *geometria sublimior*.

---

[7] Gli stipendi concessi agli accademici potevano arrivare anche fino al triplo di una quota base. Sebbene in letteratura vengano citate cifre contraddittorie, Delisle e Hermann guadagnavano almeno il triplo di quello che veniva corrisposto a giovani esordienti come Daniel Bernoulli. Il giovanissimo Eulero venne assunto con un salario ancora inferiore.

Hermann sottolinea il rigore dei metodi degli antichi, ma giudicandoli goffi in relazione ai potenti e sbrigativi metodi del calcolo differenziale e integrale, arrivando fino a considerare la *scrupulositas in demonstrando* della geometria greca come un ostacolo all'innovazione:

> Sembra poi che la raffinata scrupolosità dimostrativa degli Antichi abbia costituito un ostacolo che impediva loro di allargare il sacro recinto della Geometria: gli Antichi preferirono infatti una sicura navigazione sotto costa piuttosto che avventurarsi su una piccola zattera nel mare aperto della scienza, a rischio di naufragio [Nagel e Verdun 2005, p. 128][8].

Grazie a Descartes, Pascal, Beaune, Van Schooten, Sluse, Sturm e Barrow, la geometria dell'età di mezzo oltrepassa le coniche apolloniane e la geometria archimedea, ma non raggiunge imprese compiute nell'età matura, quella della *geometria sublimior*, rese possibili dai risultati di Fermat, Wallis e Barrow sul calcolo delle tangenti e il metodo *de maximis et minimis*. Hermann rende omaggio ai *Principia* (1687) di Newton, ma ricorda anche che in quest'opera non c'è molto sul calcolo delle flussioni, reso pubblico solo nel 1704. Viene messa in rilievo l'evoluzione dei metodi di calcolo leibniziano, sviluppati dai fratelli Jacob e Johann Bernoulli, dei quali espone dettagliatamente i risultati. Hermann definisce l'*Analyse des infiniment petits* del marchese de l'Hôpital un «elegantissimum opus» e minimizza la polemica sulla priorità della scoperta del calcolo differenziale. Conclude affermando: «Sicuramente Leibniz ha avuto molto merito nei confronti del pubblico, dal momento che per primo diede alle stampe la sua invenzione e ha cosí offerto, al momento giusto, l'occasione di estendere il recinto sacro della scienza» [*ibid.*, p. 131][9]. Hermann ci offre qui una concezione teleologica della storia, in cui le tappe precedenti sono interpretate unicamente in funzione del livello raggiunto dalla matematica a lui contemporanea.

[8] «Dein exquisita antiquorum scrupulositas in demonstrando obstaculum fuisse videtur, quod Geometriæ pomœria non longius protenderint, maluerunt enim geometriæ ripas tuto peragrare quam ratem alto scientiæ Oceano committere, parum securi a naufragio».

[9] «Et sane Leibnitius de Publico optime meritus est, quod inventum suum primus publicavit atque pomœria Scientiarum extendi occasionem mature præbuit».

Fino alla sua partenza da San Pietroburgo nel 1731, Hermann presentò regolarmente delle memorie sulla risoluzione delle equazioni differenziali, sulla misura delle forze o la teoria dei moti. Insegnava analisi e algebra il lunedí, mercoledí, giovedí e sabato, dalle 11 alle 12; fu incaricato dal vicecancelliere dell'impero russo, Johann Heinrich Ostermann, di redigere un *Abrégé des mathématiques*, apparso nel 1728 in francese e in traduzione russa. Questo manuale, composto in forma di dialogo, doveva contribuire all'istruzione matematica dello zar Pietro II, all'epoca dodicenne, ed era articolato in tre volumi, ciascuno dedicato a un alto dignitario dell'impero. Il primo era consacrato all'aritmetica, alla geometria e alla trigonometria; il terzo alle fortificazioni. Il secondo volume, opera di Delisle, trattava l'astronomia e la geografia. Mi sembra che i meccanismi del mecenatismo che operano in questo contesto siano chiaramente identificabili: un accademico come Hermann, matematico già affermato, può dedicarsi alle proprie ricerche, ma in cambio deve prendere in carico anche l'insegnamento e redigere un manuale da offrire ai suoi protettori.

3.3. Un «paradiso degli eruditi».

Cosí Wolff definiva la situazione a Pietroburgo. E in effetti i giovani, in particolare Daniel Bernoulli (Nicolaus era morto prematuramente nel luglio 1726), Eulero e Johann Georg Krafft, seppero trarre ampio profitto dalle condizioni di lavoro che venivano loro offerte: la possibilità di dedicarsi interamente alla scienza nell'ambito di un circolo di intellettuali con cui potevano discutere e scambiare idee. Anche il giovane Eulero proveniva da Basilea come i due Bernoulli ed era allievo di Johann. Già nel 1725, appena diciottenne, provava «l'indescrivibile desiderio di intraprendere il viaggio per Pietroburgo con i Bernoulli» [Fellmann 1995, p. 12]. Di fatto, grazie alla raccomandazione dei Bernoulli e di Goldbach, dal giugno 1727 poté essere accolto nella nuova Accademia. Arrivò a Pietroburgo il 24 maggio, quando la città era in lutto per la morte recentissima (17 maggio) della zarina Caterina I. Sappiamo che il giovane Eulero andò a pranzo da Hermann e a cena da Daniel Bernoulli, presso cui avrebbe poi abitato fino a quando, nel 1731, poté occupare la cattedra di fisica.

Di che cosa avranno discusso questi due giovani la sera di quel 24 maggio 1727? Forse già di idraulica, il loro campo d'indagine preferito. Tanto Daniel quanto Eulero, buoni allievi di Johann Bernoulli, risolsero alcuni problemi di idraulica applicando il principio delle forze vive (equivalente alla conservazione dell'energia cinetica $\frac{1}{2}mv^2$). I *Commentarii Academiae Petropolitanae* degli anni 1727 e 1729 documentano gli studi di Bernoulli sul flusso d'acqua in uscita da diversi recipienti di forma irregolare, dotati di un foro d'apertura di dimensione finita (e questo rappresentava una novità, dal momento che ci si limitava a considerare fori d'apertura di dimensione infinitamente piccola). Dai verbali sappiamo che anche Eulero lavorava su problemi analoghi, tuttavia i suoi studi non vennero mai pubblicati. L'emulazione fra i due amici era tale che Eulero fu portato a testimoniare che Daniel, accusato di plagio da uno degli altri accademici[10], aveva ottenuto i suoi risultati in idraulica in maniera indipendente dalla sua. Nel giugno 1729, Bernoulli fece trasmettere al presidente dell'Accademia il seguente programma:

> ho iniziato già da 6 mesi un trattato molto completo sulle leggi del moto delle acque, che conterà circa 40 fogli; [...] quest'opera mi terrà impegnato quasi giorno e notte fino alla scadenza dei miei impegni, perché non ho avuto alcun aiuto dagli autori che, fino alla formulazione del nuovo principio di conservazione delle forze vive, non avevano potuto ridurre le suddette leggi alla pura Geometria. Un'altra ragione che ritarderà la conclusione di quest'opera è il gran numero di esperimenti che ci toccherà fare [Nagel e Verdun 2005, p. 139].

Quando partirà da Pietroburgo nel 1733, lascerà un manoscritto di 400 pagine, quello della famosa *Hydrodynamica*[11], che sarà pubblicata poi a Strasburgo. Nel corso di quattro anni, Daniel seppe approfittare della libertà di ricerca che gli venne accordata a Pietroburgo per redigere questo trattato, che divenne il fondamento di una disciplina, cui egli stesso diede il nome, come testimonia in una lettera a Goldbach del 17 luglio 1730:

> Per quanto mi riguarda, mi sono interamente immerso nelle acque che hanno costituito la mia unica occupazione e rinuncio da qualche tempo a occuparmi di tutto ciò che non sia idrostatica o idraulica [...] Ho l'onore di dirvi che, nei giorni scorsi, ho fatto qualche nuova scoperta che può essere proficuamente utilizzata nella struttura dei condotti idraulici, ma che soprattutto inaugurerà una nuova stagione per la fisiologia; fatto sta che ho studiato la statica delle acque correnti, scienza che (a quanto ne so) nessuno aveva mai considerato prima di me e che, se pur se ne fosse intrapreso lo studio, sarebbe risultato molto lacunoso, dal momento che non erano stati ancora formulati i principî di cui mi servo e che sono stati pienamente confermati da un gran numero di esperienze [*ibid.*, p. 140].

Nella *Hydrodynamica*, si trova in particolare l'equazione di Bernoulli che esprime la conservazione dell'energia (*vis viva*) lungo una linea di flusso.

[10] Si tratta del fisico Georg Bernhard Bülfinger (1693-1750), originario di Tubinga. Nei verbali dell'Accademia imperiale, sono conservate notizie di discussioni accese, sulla questione delle forze vive, fra Hermann e Bülfinger da una parte e Bernoulli ed Eulero dall'altra.

[11] *Hydrodynamica, sive de viribus et motibus fluidorum commentarii*, Dulsecker, Strasbourg 1738

29 04 2020

007063419X.pdf

Pág. 179 Chapter 5

Curve Tracing

Pág. 1228 Chapter 36

Analytical Solid Geometry

Obs: Livro todo!!!

D) Books_74_2020/
Higher Engineering Mathematics by
Banday Ramana/

1138673579.pdf

Pág. 140 Section B

Geometry and trigonometry

Pág. 270 Section D

Complex numbers

Obs. Livro todo!!!

E) Books_74_2020/
Electric Field Analysis/

Electric Field Analysis.pdf

Pág. 61(84)

27.04.2020

Pág. 39 (62) Chapter 2
Gauss's Law and Related Topics

Pág. 59 (82) Chapter 3
Orthogonal Coordinate Systems

Pág. 219 (242) Chapter 10
Sphere or Cylinder in Uniform External Field.

F) Booky_74_2020/
La Matematica Vol. 1. I tempi e i luoghi/

La matematica....pdf

Pág. 189 (205)

5.2 La trigonometria

Teorema di Menelao.

Se CE, CF, DF, DB sono archi di cerchi massimi su una sfera.
(círculos)

Pág. 190 Teorema del seno,

Pág. 402 (418) Leibniz, consigliere dei principi

! " " #$ $ # % & ' (#!#
" " " #$ $ #
)*+,)))-) ' . / /
/0 1 2 / 3 ." (

Love Agarwal

FAMILY OF CIRCLES

In this article, we will learn about family of circles, common tangents and angle of intersection of two circles. These are very important concept which are frequently asked in JEE.

Family of Circles

A collection of circles is called a system or family of circles. There can be various types of family of circles. The general equation of a circle is $x^2 + y^2 + 2gx + 2fy + c = 0$. Since, this equation involves three unknowns, *viz.*, g, f and c, so we need atleast three conditions to get a unique circle.

Theorem 1

The equation of a family of circles passing through the intersection of a circle $S \equiv x^2 + y^2 + 2gx + 2fy + c = 0$ and line

$$L \equiv lx + my + n = 0 \text{ is}$$

$$x^2 + y^2 + 2gx + 2fy + c + \lambda(lx + my + n) = 0$$

or $$S + \lambda L = 0$$

where, λ is any real number.

Example 1 The intercept on the line $y = x$ by the circle $x^2 + y^2 - 2x = 0$ is AB. Equation of the circle with AB as a diameter is

(a) $x^2 + y^2 + x + y = 0$ (b) $x^2 + y^2 - x - y = 0$
(c) $x^2 + y^2 - x + y = 0$ (d) $x^2 + y^2 + x - y = 0$

Sol. **(b)** Equation of any circle passing through the point of intersection of $x^2 + y^2 - 2x = 0$ and $y = x$ is

$$S + \lambda L = 0$$

$$\Rightarrow \quad x^2 + y^2 - 2x + \lambda(y - x) = 0$$

$$\Rightarrow \quad x^2 + y^2 - (2 + \lambda)x + \lambda y = 0$$

Now, above circle has centre $\left[\left(\frac{2 + \lambda}{2}\right), \frac{-\lambda}{2}\right]$.

For AB to be the diameter of the required circle, the centre must be on AB.

$$\therefore \quad \frac{2 + \lambda}{2} = \frac{-\lambda}{2} \Rightarrow \lambda = -1$$

Therefore, equation of the required circle is

$$x^2 + y^2 - (2 - 1)x - y = 0$$

$$\Rightarrow \quad x^2 + y^2 - x - y = 0$$

Example 2 Equation of the smallest circle passing through the intersection of the line $x+y=1$ and the circle $x^2+y^2=9$ is

(a) $x^2+y^2-x-y-8=0$
(b) $x^2+y^2+x+y+9=0$
(c) $x^2+y^2+2x+2y+4=0$
(d) $2x^2+2y^2-2x-2y-7=0$

Sol. **Key point** Smallest circle is that circle having given line as diameter.

(a) Any circle passing through the point of intersection of the given line and circle has the equation

$$x^2+y^2-9+\lambda(x+y-1)=0 \qquad \text{...(i)}$$

Its centre is $\left(\frac{-\lambda}{2}, \frac{-\lambda}{2}\right)$.

This circle is the smallest, if $\left(\frac{-\lambda}{2}, \frac{-\lambda}{2}\right)$ is on the chord $x+y-1=0$.

$$\therefore \quad \frac{-\lambda}{2}-\frac{\lambda}{2}-1=0 \Rightarrow \lambda=-1$$

On putting the value of λ in Eq. (i), the equation of the smallest circle is

$$x^2+y^2-9-1(x+y-1)=0$$
$$\Rightarrow \quad x^2+y^2-x-y-8=0$$

Theorem 2

The equation of the family of circles passing through the points $A(x_1, y_1)$ and $B(x_2, y_2)$ is

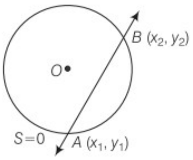

$$(x-x_1)(x-x_2)+(y-y_1)(y-y_2)+\lambda\begin{vmatrix} x & y & 1 \\ x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \end{vmatrix}=0$$

or $(x-x_1)(x-x_2)+(y-y_1)(y-y_2)+\lambda L=0$, where $L=0$ represents the line passing through $A(x_1, y_1)$ and $B(x_2, y_2)$ and $\lambda \in R$.

POINTS TO REMEMBER

1. If $x_2 \to x_1$ and $y_2 \to y_1$, then equation of the circle becomes $(x-x_1)^2+(y-y_1)^2+\lambda L=0$.
2. Also, when $x_2 \to x_1$ and $y_2 \to y_1$, then line AB becomes tangent to the circle $S=0$.

Example 3 The circle passing through $(1,-2)$ and touching the axis of X at $(3, 0)$, also passes through the point **[JEE Main 2013]**

(a) $(-5, 2)$ (b) $(2, -5)$ (c) $(5, -2)$ (d) $(-2, 5)$

Sol. **(c)** The equation of the circle touching X-axis, i.e. $y=0$ at $(3, 0)$ is

$$(x-3)^2+(y-0)^2+\lambda y=0 \qquad \text{...(i)}$$

$\because$ It passes through $(1,-2)$.

$$\therefore \quad (1-3)^2+(-2-0)^2-2\lambda=0$$
$$\Rightarrow \quad \lambda=4$$

Putting $\lambda=4$ in Eq. (i), we obtain

$$(x-3)^2+y^2+4y=0$$
$$\Rightarrow \quad x^2+y^2-6x+4y+9=0$$

Clearly, it passes through $(5,-2)$.

Example 4 If C_1, C_2 and C_3 belong to a family of circles through the points (x_1, y_1) and (x_2, y_2), then the ratio of lengths of the tangent from any point on C_1 to the circles is

(a) constant (b) cannot be determined
(c) equal to difference of x-coordinates
(d) equal to difference of y-coordinates

Sol. **(a)** Equation of the family of circles through (x_1, y_1) and (x_2, y_2) is

$$(x-x_1)(x-x_2)+(y-y_1)(y-y_2)+\lambda\begin{vmatrix} x & y & 1 \\ x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \end{vmatrix}=0$$

Let (h, k) be the point on C_1.

$$\Rightarrow \quad f(h,k)+\lambda_1 g(h,k)=0$$

where, $f(h,k)=(h-x_1)(h-x_2)+(k-y_1)(k-y_2)$

and $$g(h,k)=\begin{vmatrix} h & k & 1 \\ x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \end{vmatrix}$$

Let T_2 be the length of the tangent from (h,k) to C_2 and T_3 be the length of the tangent from (h, k) to C_3.

Then, $T_2=\sqrt{f(h,k)+\lambda_2 g(h,k)}$

and $T_3=\sqrt{f(h,k)+\lambda_3 g(h,k)}$

$$\therefore \quad \frac{T_2}{T_3}=\frac{\sqrt{f(h,k)+\lambda_2 g(h,k)}}{\sqrt{f(h,k)+\lambda_3 g(h,k)}}$$

$$\Rightarrow \quad \frac{T_2}{T_3}=\frac{\sqrt{(\lambda_2-\lambda_1)\,g(h,k)}}{\sqrt{(\lambda_3-\lambda_1)\,g(h,k)}}$$

$$\Rightarrow \quad \frac{T_2}{T_3}=\sqrt{\frac{\lambda_2-\lambda_1}{\lambda_3-\lambda_1}}=\text{Constant}$$

Theorem 3

The equation of the family of circles touching the circle $S \equiv x^2+y^2+2gx+2fy+c=0$ at point $P(x_1, y_1)$ is

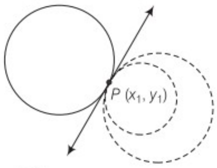

$$x^2+y^2+2gx+2fy+c + \lambda\,[xx_1+yy_1+g(x+x_1)+f(y+y_1)+c]=0$$

or $S+\lambda L=0$, where $L=0$ is the equation of the tangent to $S=0$ at (x_1, y_1) and $\lambda \in R$.

POINTS TO REMEMBER

1. The equation of a family of circles touching a fixed line $y - y_1 = m(x - x_1)$ at the fixed point (x_1, y_1) is $(x - x_1)^2 + (y - y_1)^2 + k\,[y - y_1 - m(x - x_1)] = 0$, where k is a parameter.
2. In case the line through (x_1, y_1) is parallel to Y-axis, the equation of the family of circles touching it at (x_1, y_1) becomes $(x - x_1)^2 + (y - y_1)^2 + k(x - x_1) = 0$.
3. Also, if the line parallel to X-axis, then the equation of family of circles touching it at $(x_1, y_1) = 0$ becomes $(x - x_1)^2 + (y - y_1)^2 + k(y - y_1) = 0$.

Example 5 The circle passing through the point $(-1, 0)$ and touching the Y-axis at $(0, 2)$, also passes through the point **[IIT JEE 2011]**

(a) $\left(-\frac{3}{2}, 0\right)$ (b) $\left(-\frac{5}{2}, 2\right)$ (c) $\left(-\frac{3}{2}, \frac{5}{2}\right)$ (d) $(-4, 0)$

Sol. **(d)** The equation of the family of circles touching Y-axis at $(0, 2)$ is

$$(x - 0)^2 + (y - 2)^2 + \lambda x = 0$$

It passes through $(-1, 0)$.

$$1 + 4 - \lambda = 0 \Rightarrow \lambda = 5$$

So, the equation of the circle is

$$x^2 + 5x + y^2 - 4y + 4 = 0$$

Clearly, it passes through $(-4, 0)$.

Theorem 4

The equation of a family of circles passing through the intersection of two circles

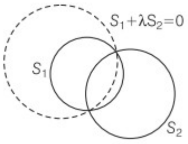

$$S_1 \equiv x^2 + y^2 + 2g_1x + 2f_1y + c_1 = 0$$

and $$S_2 \equiv x^2 + y^2 + 2g_2x + 2f_2y + c_2 = 0$$

is $$S_1 + \lambda S_2 = 0$$

i.e. $$(x^2 + y^2 + 2g_1x + 2f_1y + c_1) + \lambda\,(x^2 + y^2 + 2g_2x + 2f_2y + c_2) = 0,$$

where $\lambda\ (\neq -1)$ is an arbitrary real number.

Example 6 The equation of the circle passing through $(1, -3)$ and the points common to the two circles $x^2 + y^2 - 6x + 8y - 16 = 0$, $x^2 + y^2 + 4x - 2y - 8 = 0$ is

(a) $x^2 + y^2 - 4x + 6y + 24 = 0$
(b) $2x^2 + 2y^2 + 3x + y - 20 = 0$
(c) $3x^2 + 3y^2 - 5x + 7y - 19 = 0$
(d) None of the above

Sol. **(b)** The equation of circle is

$$(x^2 + y^2 - 6x + 8y - 16) + \lambda(x^2 + y^2 + 4x - 2y - 8) = 0$$

$\because$ It passes through $(1, -3)$.

$$\therefore \quad (1 + 9 - 6 - 24 - 16) + \lambda(1 + 9 + 4 + 6 - 8) = 0$$
$$\Rightarrow \quad -36 + \lambda(12) = 0$$
$$\Rightarrow \quad \lambda = 3$$

So, the equation of the required circle is

$$4x^2 + 4y^2 + 6x + 2y - 40 = 0$$
$$\Rightarrow \quad 2x^2 + 2y^2 + 3x + y - 20 = 0$$

Circle in Contact

Let us discuss condition for two circles touching each other. Two circles can touch each other externally or internally.

Case I When Circles Touch Externally

When two circles touch each other externally, then distance between their centres is equal to sum of their radii,

i.e. $|C_1C_2| = r_1 + r_2$

In this case, point of contact P divides the line joining C_1 and C_2 internally in the ratio $r_1 : r_2$.

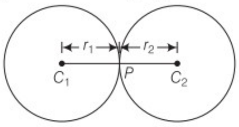

$$\Rightarrow \quad \frac{C_1P}{C_2P} = \frac{r_1}{r_2}$$

If $C_1(x_1, y_1)$ and $C_2(x_2, y_2)$, then the coordinates of P is

$$\left(\frac{r_1x_2 + r_2x_1}{r_1 + r_2}, \frac{r_1y_2 + r_2y_1}{r_1 + r_2}\right).$$

Example 7 Let C be the circle with centre $(1, 1)$ and radius 1 unit. If T is the circle centred at $(0, y)$ passing through origin and touching the circle C externally, then the radius of T is equal to **[IIT JEE 2014]**

(a) $\frac{\sqrt{3}}{\sqrt{2}}$ (b) $\frac{\sqrt{3}}{2}$ (c) $\frac{1}{2}$ (d) $\frac{1}{4}$

Sol. **(d)** Here, the radius of circle C is 1 unit and the radius of another circle T whose centre is $(0, y)$ and passing through the origin is $y = k$ (say).

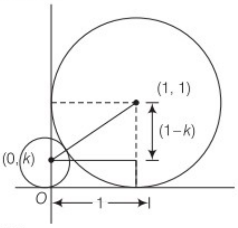

From the figure,

Distance between the centres = Sum of radii

$$\sqrt{1+(1-k)^2} = k+1$$

$$\Rightarrow \quad 1+1+k^2-2k = k^2+1+2k$$

$$\Rightarrow \quad 4k = 1 \Rightarrow k = \frac{1}{4}$$

Hence, radius of circle T is $\frac{1}{4}$.

Case II When Circles Touch Internally

When two circles touch each other internally, then distance between their centres is equal to the difference of their radii,

i.e. $\quad |C_1C_2| = |r_1 - r_2|$

In such a case, point of contact P divide the line C_1 and C_2 enternally in the ratio $r_1 : r_2$.

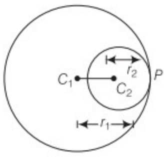

$$\Rightarrow \quad \frac{C_1P}{C_2P} = \frac{r_1}{r_2}$$

If $C_1 \equiv (x_1, y_1)$ and $C_2 \equiv (x_2, y_2)$, then the coordinates of P is $\left(\frac{r_1x_2 - r_2x_1}{r_1 - r_2}, \frac{r_1y_2 - r_2y_1}{r_1 - r_2}\right)$.

Example ⑧ Two circles $x^2 + y^2 + 2ax + c^2 = 0$ and $x^2 + y^2 + 2by + c^2 = 0$ touch each other, then $\frac{1}{a^2} + \frac{1}{b^2}$ is equal to

(a) 1 (b) $\frac{1}{c^2}$

(c) 0 (d) None of these

Sol. **(b)** Given circles are

$$x^2 + y^2 + 2ax + c^2 = 0 \qquad \text{...(i)}$$

and

$$x^2 + y^2 + 2by + c^2 = 0 \qquad \text{...(ii)}$$

Let C_1 and C_2 be the centres of circles (i) and (ii) respectively and r_1, r_2 be their radii, then $C_1(-a, 0)$, $C_2(0, -b)$, $r_1 = \sqrt{a^2 - c^2}$, $r_2 = \sqrt{b^2 - c^2}$. Since, the circles touch each other externally or internally.

$$\therefore \quad |C_1C_2| = |r_1 \pm r_2|$$

$$\Rightarrow \quad \sqrt{a^2 + b^2} = \left|\sqrt{a^2 - c^2} \pm \sqrt{b^2 - c^2}\right|$$

On squaring both the sides, we get

$$a^2 + b^2 = a^2 - c^2 + b^2 - c^2 \pm 2\sqrt{(a^2 - c^2)(b^2 - c^2)}$$

$$\Rightarrow \quad c^2 = \pm\sqrt{(a^2 - c^2)(b^2 - c^2)}$$

$$\Rightarrow \quad c^2 = \pm\sqrt{a^2b^2 - c^2(a^2 + b^2) + c^4}$$

$$\Rightarrow \quad c^4 = a^2b^2 - c^2(a^2 + b^2) + c^4$$

$$\Rightarrow \quad c^2(a^2 + b^2) = a^2b^2 \Rightarrow \frac{1}{a^2} + \frac{1}{b^2} = \frac{1}{c^2}$$

Common Tangents to Two Circles

The number of common tangents depends upon the position of the given circles as discussed below. Let the equations of the two circles be

$$(x - h_1)^2 + (y - k_1)^2 = a_1^2 \qquad \text{...(i)}$$

and

$$(x - h_2)^2 + (y - k_2)^2 = a_2^2 \qquad \text{...(ii)}$$

with centres $C_1(h_1, k_1)$ and $C_2(h_2, k_2)$, and radii a_1 and a_2, respectively. Then, the following cases of intersection of these two circles may arises.

Case I When $C_1C_2 > a_1 + a_2$

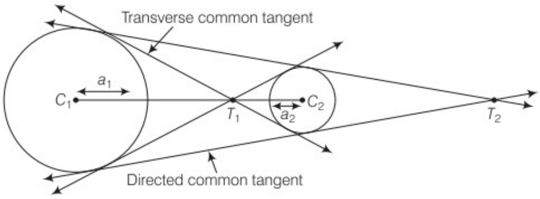

There are four common tangents that can be drawn to two circles. Two of them are direct common tangents and the other two are transverse common tangents. The points of intersection of direct and transverse common tangents always lie on the line joining the centres C_1 and C_2, and divide it internally and externally respectively in the ratio $a_1 : a_2$,

i.e. $\quad \frac{C_1T_2}{C_2T_2} = \frac{a_1}{a_2}$ (externally) and $\frac{C_1T_1}{C_2T_1} = \frac{a_1}{a_2}$ (internally).

Case II When $C_1C_2 = a_1 + a_2$

In this case, two direct tangents are real and distinct while the transverse tangents are coincident.

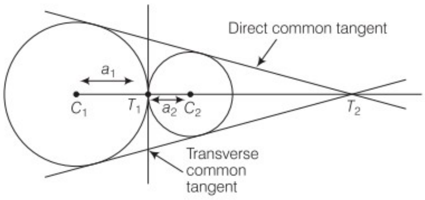

Case III When $C_1C_2 < a_1 + a_2$

In this case, the two direct common tangents are real while the transverse tangents are imaginary.

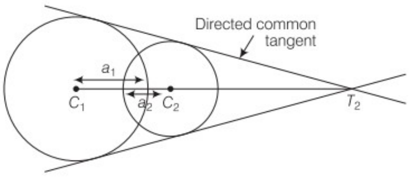

Case IV When $C_1C_2 = a_1 - a_2$

In this case, two tangents are real and coincident while the other two tangents are imaginary.

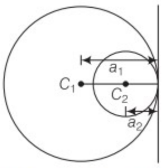

Case V When $C_1C_2 < a_1 - a_2$

In this case, all the four common tangents are imaginary.

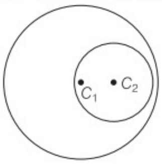

Example 9 How many common tangents can be drawn to the circles $x^2 + y^2 = 6x$ and $x^2 + y^2 + 6x + 2y + 1 = 0$?

(a) 4 (b) 3
(c) 2 (d) 1

Sol. **(a)** The coordinates of the given circles are $C_1(3, 0)$, $C_2(-3, -1)$ and corresponding radii are $r_1 = 3$ and $r_2 = 3$, respectively.

Now, $C_1C_2 = \sqrt{(3+3)^2 + (0+1)^2} = \sqrt{37}$

and $r_1 + r_2 = 3 + 3 = 6$

$\therefore$ $C_1C_2 > r_1 + r_2$

So, 4 common tangents can be drawn to the given circles.

Example 10 The common tangents to the circles $x^2 + y^2 + 2x = 0$ and $x^2 + y^2 - 6x = 0$ form a triangle which is

(a) equilateral
(b) isosceles
(c) right angled
(d) None of the above

Sol. **(b)** Clearly, the distance between the centres of the given circles is equal to the sum of their radii. So, two given circles will touch each other externally.

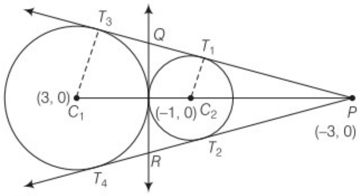

We have,

$$PT_1 = PT_2 \text{ and } PT_3 = PT_4$$

$\Rightarrow$ $T_1T_3 = T_2T_4$

$\Rightarrow$ $2T_1Q = 2T_2R$ $[\because QT_3 = QT_1 \text{ and } RT_4 = RT_2]$

$\Rightarrow$ $T_1Q = T_2R$

$\Rightarrow$ $PT_1 + T_1Q = PT_1 + T_2R$

$\Rightarrow$ $PT_1 + T_1Q = PT_2 + T_2R$

$\Rightarrow$ $PQ = PR$

Hence, PQR is an isosceles triangle.

POINTS TO REMEMBER

1. Length of direct common tangent $= \sqrt{d^2 - (r_1 - r_2)^2}$.
2. Length of transverse common tangent $= \sqrt{d^2 - (r_1 + r_2)^2}$ where, d is distance between the circles and r_1, r_2 are radii.

Common Chord

The chord joining the points of intersection of two given circles is called their common chord. The equation of the common chord of two circles

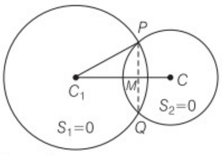

$$S_1 \equiv x^2 + y^2 + 2g_1x + 2f_1y + c_1 = 0$$

and $S_2 \equiv x^2 + y^2 + 2g_2x + 2f_2y + c_2 = 0$

is $2x(g_1 - g_2) + 2y(f_1 - f_2) + c_1 - c_2 = 0$

i.e. $S_1 - S_2 = 0$.

Length of the Common Chord

From above figure, we have

$$PQ = 2\,PM = 2\sqrt{C_1P^2 - C_1M^2}$$

where, C_1P = Radius of the circle $S = 0$

and C_1M = Length of the perpendicular from the centre C_1 to the common chord PQ

Example 11 If the circle $x^2 + y^2 + 6x - 2y + k = 0$ bisect the circumference of the circle $x^2 + y^2 + 2x - 6y - 15 = 0$, then k is equal to

(a) 21 (b) – 21
(c) 23 (d) – 23

Sol. **(d)** The equation of common chord of the given circle is

$$4x + 4y + (k + 15) = 0 \quad \text{...(i)}$$

Since, $x^2 + y^2 + 6x - 2y + k = 0$ bisects the circumference of $x^2 + y^2 + 2x - 6y - 15 = 0$. Therefore, Eq. (i) is a diameter of the circle $x^2 + y^2 + 2x - 6y - 15 = 0$.

$\therefore$ $4(-1) + 4 \times 3 + k + 15 = 0$

$\Rightarrow$ $k = -23$

Angle of Intersection of Two Circles

The angle of intersection of two circles is defined as the angle between the tangents to the two circles at their point of intersection.

Let $x^2 + y^2 + 2g_1x + 2f_1y + c_1 = 0$

and $x^2 + y^2 + 2g_2x + 2f_2y + c_2 = 0$

be two circles having their centres at $C_1(-g_1, -f_1)$ and $C_2(-g_2, -f_2)$.

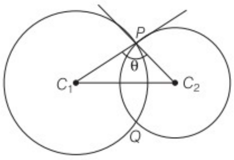

Then, $\cos\theta = \dfrac{C_1P^2 + C_2P^2 - C_1C_2^2}{2C_1P \cdot C_2P}$

Orthogonal Circles

Two circles are said to intersect orthogonally, if their angles of intersection is a right angle.

For two circles $x^2 + y^2 + 2g_1x + 2f_1y + c_1 = 0$ and $x^2 + y^2 + 2g_2x + 2f_2y + c_2 = 0$ cut orthogonally, the necessary condition is $2(g_1g_2 + f_1f_2) = C_1 + C_2$.

Example 12 The angle of intersection of the circles $x^2 + y^2 - 6x + 4y + 11 = 0$ and $x^2 + y^2 - 4x + 6y + 9 = 0$ is

(a) 30° (b) 45°
(c) 60° (d) 90°

Sol. **(b)** The equations of the circles are

$$x^2 + y^2 - 6x + 4y + 11 = 0 \quad \text{...(i)}$$

and $$x^2 + y^2 - 4x + 6y + 9 = 0 \quad \text{...(ii)}$$

The centres of these circles are $C_1(3, -2)$ and $C_2(2, -3)$, respectively. Let r_1 and r_2 be their radii.

Then, $r_1 = \sqrt{9 + 4 - 11} = \sqrt{2}$ and $r_2 = \sqrt{4 + 9 - 9} = 2$

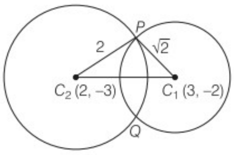

Also, $C_1C_2 = \sqrt{2}$

Suppose circles (i) and (ii) intersect at P.

Then,

$$\cos(\angle C_1PC_2) = \frac{C_1P^2 + C_2P^2 - C_1C_2^2}{2C_1P \cdot C_2P} = \frac{2 + 4 - 2}{2 \times \sqrt{2} \times 2} = \frac{1}{\sqrt{2}}$$

$$\Rightarrow \quad \angle C_1PC_2 = 45°$$

Smart Practice

1. C_1 and C_2 are circles of unit radius with centres at (0, 0) and (1, 0), respectively. C_3 is a circle of unit radius. It passes through the centres of the circles C_1 and C_2, and its centre above the X-axis. Then, the equation of the common tangent to C_1 and C_3 which does not pass through C_2, is

(a) $\sqrt{3}x + y - 2 = 0$ (b) $\sqrt{3}y + x - 2 = 0$
(c) $\sqrt{3}x - y + 2 = 0$ (d) $\sqrt{3}y - x - 2 = 0$

2. The radius of the smallest circle which touches the straight line $3x - y = 6$ at $(1, -3)$ and also touches the line $y = x$, is (compute upto one place of decimal only)

(a) 2.5 (b) 1.5 (c) 3.5 (d) 4.5

3. A rhombus is inscribed in the region common to the two circles $x^2 + y^2 - 4x - 12 = 0$ and $x^2 + y^2 + 4x - 12 = 0$ with two of its vertices on the line joining the centres of the circles. The area of the rhombus is

(a) $8\sqrt{3}$ sq units (b) $4\sqrt{3}$ sq units
(c) $6\sqrt{3}$ sq units (d) None of these

4. Two circles with radii a and b touch each other externally such that θ is the angle between the direct common tangents $(a > b \geq 2)$. Then,

(a) $\theta = 2\cos^{-1}\left(\dfrac{a - b}{a + b}\right)$

(b) $\theta = 2\tan^{-1}\left(\dfrac{a + b}{a - b}\right)$

(c) $\theta = 2\sin^{-1}\left(\dfrac{a + b}{a - b}\right)$

(d) $\theta = 2\sin^{-1}\left(\dfrac{a - b}{a + b}\right)$

5. The angle at which the circles $(x - 1)^2 + y^2 = 10$ and $x^2 + (y - 2)^2 = 5$ intersect, is

(a) $\dfrac{\pi}{6}$ (b) $\dfrac{\pi}{4}$
(c) $\dfrac{\pi}{3}$ (d) $\dfrac{\pi}{2}$

6. If the line $x\cos\theta + y\sin\theta = 2$ is the equation of a transverse common tangent to the circles $x^2 + y^2 = 4$ and $x^2 + y^2 - 6\sqrt{3}x - 6y + 20 = 0$, then the value of θ is

(a) $\dfrac{5\pi}{6}$ (b) $\dfrac{2\pi}{4}$
(c) $\dfrac{\pi}{3}$ (d) $\dfrac{\pi}{6}$

7. If $C_1 : x^2 + y^2 - 20x + 64 = 0$ and $C_2 : x^2 + y^2 + 30x + 144 = 0$, then the length of the shortest line segment PQ which touches C_1 at P and C_2 at Q, is

(a) 20
(b) 15
(c) 22
(d) 27

8. The coordinates of two points P and Q are (x_1, y_1) and (x_2, y_2) and O is the origin. If the circles are described on OP and OQ as diameters, then the length of their common chord is

(a) $\dfrac{|x_1y_2 + x_2y_1|}{PQ}$
(b) $\dfrac{|x_1y_2 - x_2y_1|}{PQ}$
(c) $\dfrac{|x_1x_2 + y_1y_2|}{PQ}$
(d) $\dfrac{|x_1x_2 - y_1y_2|}{PQ}$

9. The equation of the circle passing through the point of intersection of the circles $x^2 + y^2 - 4x - 2y = 8$ and $x^2 + y^2 - 2x - 4y = 8$ and the point $(-1, 4)$ is

(a) $x^2 + y^2 + 4x + 4y - 8 = 0$
(b) $x^2 + y^2 - 3x + 4y + 8 = 0$
(c) $x^2 + y^2 + x + y - 8 = 0$
(d) $x^2 + y^2 - 3x - 3y - 8 = 0$

10. If the length of the common chord of two circles $x^2 + y^2 + 8x + 1 = 0$ and $x^2 + y^2 + 2\mu y - 1 = 0$ is $2\sqrt{6}$, then the value of μ are

(a) ± 2
(b) ± 3
(c) ± 4
(d) None of these

11. A circle C_1 of radius b touches the circle $x^2 + y^2 = a^2$ externally and has its centre on the positive X-axis, another circle C_2 of radius c touches the circle C_1 externally and has its center on the positive X-axis. Given $a < b < c$, the three circles have a common tangent, if a, b and c are in

(a) AP
(b) GP
(c) HP
(d) None of these

12. Consider a family of circles passing through two fixed points $A(3, 7)$ and $B(6, 5)$. The chords in which the circle $x^2 + y^2 - 4x - 6y - 3 = 0$ cuts the members of the family are concurrent at a point. Then, the coordinates of this point are

(a) $\left(\dfrac{7}{2}, \dfrac{9}{2}\right)$
(b) $\left(\dfrac{9}{2}, \dfrac{7}{2}\right)$
(c) $\left(-\dfrac{7}{2}, \dfrac{9}{2}\right)$
(d) $\left(\dfrac{9}{2}, -\dfrac{7}{2}\right)$

13. If the circles $x^2 + y^2 + 2x + 2ky + 6 = 0$ and $x^2 + y^2 + 2ky + k = 0$ intersect orthogonally, then k is equal to

(a) 2 or $-\dfrac{3}{2}$
(b) -2 or $-\dfrac{3}{2}$
(c) 2 or $\dfrac{3}{2}$
(d) -2 or $\dfrac{3}{2}$

14. The number of common tangents to the circles $x^2 + y^2 - 4x - 6y - 12 = 0$ and $x^2 + y^2 + 6x + 18y + 26 = 0$ is **[JEE Main 2015]**

(a) 3
(b) 4
(c) 1
(d) 2

15. If the circles $x^2 + y^2 + 4x + 22y + c = 0$ bisects the circumference of the circle $x^2 + y^2 - 2x + 8y - d = 0$, then $c + d$ is equal to

(a) 60
(b) 50
(c) 40
(d) 56

16. The circles having radii r_1 and r_2 intersect orthogonally. The length of their common chord is

(a) $\dfrac{2r_1r_2}{\sqrt{r_1^2 + r_2^2}}$
(b) $\dfrac{\sqrt{r_1^2 + r_2^2}}{2r_1r_2}$
(c) $\dfrac{r_1r_2}{\sqrt{r_1^2 + r_2^2}}$
(d) $\dfrac{\sqrt{r_1^2 + r_2^2}}{r_1r_2}$

17. The distance from the centre of the circle $x^2 + y^2 = 2x$ to the common chord of the circles $x^2 + y^2 + 5x - 8y + 1 = 0$ and $x^2 + y^2 - 3x + 7y - 25 = 0$ is

(a) 2
(b) 4
(c) $\dfrac{34}{13}$
(d) $\dfrac{26}{17}$

18. If the radii of the circle $(x - 1)^2 + (y - 2)^2 = 1$ and $(x - 7)^2 + (y - 10)^2 = 4$ are increasing uniformly w.r.t. time as 0.3 unit/s and 0.4 unit/s, then they will touch each other at t equal to

(a) 45 s
(b) 90 s
(c) 11 s
(d) 135 s

19. A circle is given by $x^2 + (y - 1)^2 = 1$. Another circle C touches it externally and also the X-axis. Then, the locus of its centre is

(a) $\{(x, y) : x^2 = 4y\} \cup \{(x, y) : y \le 0\}$
(b) $\{(x, y) : x^2 + (y - 1)^2 = 4\} \cup \{(x, y) : y \le 0\}$
(c) $\{(x, y) : x^2 = y\} \cup \{(0, y) : y \le 0\}$
(d) $\{(x, y) : x^2 = 4y\} \cup \{(0, y) : y \le 0\}$

20. A circle S passes through the point $(0, 1)$ and is orthogonal to the circles $(x - 1)^2 + y^2 = 16$ and $x^2 + y^2 = 1$. Then,

(a) radius of S is 8
(b) radius of S is 7
(c) centre of S is $(-7, 11)$
(d) centre of S is $(-8, 1)$

Do You Know

We all know 0 was invented by Aryabhatt. As far as the invention of digits 1-9 is concerned, these are believed to be invented in Arab. These digits are also known as Arabic Numerals. The first positional numerical system was developed in Babylon in the 2nd Millenium BC.

Answers with Explanation

1. ***(c)*** The equation of any circle through (0, 0) and (1, 0) is

$$(x-1)(x-0)+(y-0)(y-0)+\lambda\begin{vmatrix} x & y & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{vmatrix}=0$$

$$\Rightarrow \qquad x^2+y^2-x+\lambda y=0$$

If it represents C_3, then its radius is 1.

$$\therefore \qquad 1=\frac{1}{4}+\frac{\lambda^2}{4} \quad \text{or} \quad \lambda=\pm\sqrt{3}$$

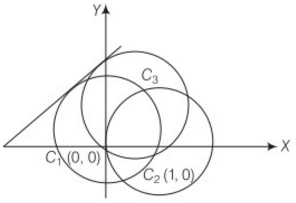

As the centre of C_3 lies above the X-axis. We take $\lambda=-\sqrt{3}$. Thus, an equation of C_3 is $x^2+y^2-x-\sqrt{3}y=0$. Since, C_1 and C_3 intersect and are of unit radius, their common tangents are parallel to the line joining their centres (0, 0) and $\left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)$.

So, let the equation of a common tangent be $\sqrt{3}x-y+k=0$.

It will touch C_1, if $\dfrac{|k|}{\sqrt{3+1}}=1 \Rightarrow k=\pm 2$

2. ***(b)*** The equation of the circle touching the line $3x-y-6=0$ at $(1,-3)$ is given by

$$(x-1)^2+(y+3)^2+\lambda(3x-y-6)=0$$

$$\Rightarrow x^2+y^2+(3\lambda-2)x+(6-\lambda)y+10-6\lambda=0 \qquad \text{...(i)}$$

$\because$ Eq. (i) touches $y=x$. Therefore,

$$2x^2+(2\lambda+4)x+10-6\lambda=0$$

i.e.

$$x^2+(\lambda+2)x+5-3\lambda=0 \qquad \text{...(ii)}$$

has equal roots. Therefore,

$$(\lambda+2)^2-4(5-3\lambda)=\lambda^2+16\lambda-16=0$$

$$\Rightarrow \qquad \lambda=\frac{1}{2}[-16\pm\sqrt{256+64}]=-8\pm\sqrt{80}$$

Now, $(\text{Radius})^2=R^2$

$$=\frac{1}{4}[(3\lambda-2)^2+(6-\lambda)^2-(10-6\lambda)4]$$

$$=\frac{1}{4}[10\lambda^2]$$

$$\Rightarrow \qquad R=\frac{\sqrt{10}\lambda}{2}=\frac{\sqrt{10}}{2}(-8\pm4\sqrt{5})$$

$$=\left|-4\sqrt{10}\pm10\sqrt{2}\right|$$

$\therefore$ Radius of smaller circle $=\left|-4\sqrt{10}+10\sqrt{2}\right|$

$$=1.5 \text{ (approx.)}$$

3. ***(a)*** The circles are with centres (2, 0) and (− 2, 0) and radius 4. Therefore, the Y-axis is their common chord.

$\therefore$ ABC is an equilateral triangle.

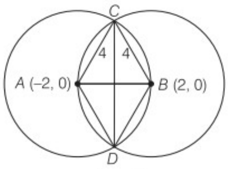

Hence, area of $ADBC=\dfrac{2\times\sqrt{3}}{4}(4)^2=8\sqrt{3}$ sq units

4. ***(d)***

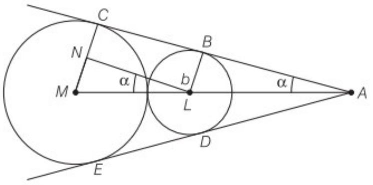

From ΔMLN,

$$\sin\alpha=\frac{a-b}{a+b}$$

$$\Rightarrow \qquad \alpha=\sin^{-1}\left(\frac{a-b}{a+b}\right)$$

$\therefore$ Angle between AB and $AD=2\alpha=2\sin^{-1}\left(\dfrac{a-b}{a+b}\right)$

5. ***(b)*** For given $r_1=\sqrt{10}$, $C_1\equiv(1,0)$

and for $\qquad r_2=\sqrt{5}$, $C_2\equiv(0,2)$

$\therefore \qquad d=C_1C_2=\sqrt{5}$

If θ is the angle between the circles, then

$$\cos\theta=\frac{|d^2-r_1^2-r_2^2|}{2r_1r_2}=\frac{|5-10-5|}{2\sqrt{10}\sqrt{5}}=\frac{1}{\sqrt{2}}$$

Hence, $\qquad \theta=\dfrac{\pi}{4}$

6. ***(d)*** $C_1C_2=r_1+r_2$

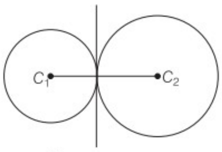

$C_1\equiv(0,0)$; $C_2\equiv(3\sqrt{3},3)$ and $r_1=2$, $r_2=4$

The circles touch each other externally.

The equation of common tangent is $\sqrt{3}x+y-4=0$.

Comparing it with $x\cos\theta+y\sin\theta=2$, we get $\theta=\dfrac{\pi}{6}$.

7. ***(a)*** The centres are $(10, 0)$ and $(-15, 0)$, and the radii are $r_1 = 6$ and $r_2 = 9$.

Also, $d = 25, r_1 + r_2 < d$

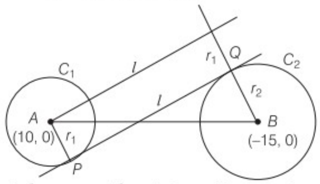

So, the circles are neither intersecting nor touching.

Therefore, $PQ = \sqrt{d^2 - (r_1 + r_2)^2}$

$= \sqrt{625 - 225} = \sqrt{400}$

$= 20$

8. ***(b)*** $OR = \dfrac{2 \times \text{Area of } \Delta OPQ}{PQ}$

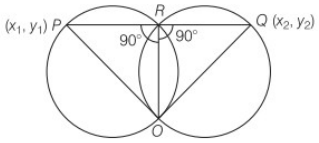

$$= \frac{2 \times \left| \frac{1}{2}(x_1y_2 - x_2y_1) \right|}{PQ}$$

$$= \frac{|x_1y_2 - x_2y_1|}{PQ}$$

9. ***(d)*** The equation of any circle through the points of intersection of the given circles is

$$x^2 + y^2 - 4x - 2y - 8 + k(x^2 + y^2 - 2x - 4y - 8) = 0 \quad \text{...(i)}$$

Since, circle (i) passes through $(-1, 4)$, we have $k = 1$

Therefore, the required circle is

$$x^2 + y^2 - 3x - 3y - 8 = 0.$$

10. ***(b)*** The common chord is $S_1 - S_2 = 0$.

i.e. $4x - \mu y + 1 = 0$

$\therefore AC = \text{Radius of first circle} = \sqrt{16}$

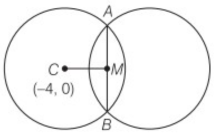

Also, $AM = \dfrac{AB}{2} = \sqrt{6}$

$\therefore \quad CM = 3$

$\Rightarrow \quad \dfrac{|-16 + 1|}{\sqrt{16 + \mu^2}} = 3$

$\Rightarrow \quad \mu = \pm 3$

11. ***(b)***

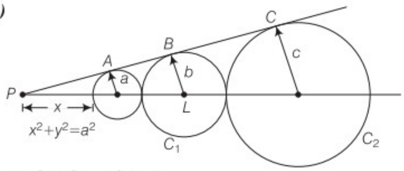

From the above figure,

$$\frac{a}{x} = \frac{b}{x + a + b} = \frac{c}{x + a + 2b + c}$$

$\therefore \quad ax + a^2 + ab = bx$ [comparing first two terms]

$\Rightarrow \quad x = \dfrac{a^2 + ab}{b - a}$

$\Rightarrow \quad ax + a^2 + 2ab + ac = cx$

[comparing first and third terms]

$\Rightarrow \quad x = \dfrac{a^2 + 2ab + ac}{c - a}$

$\therefore \quad \dfrac{a^2 + ab}{b - a} = \dfrac{a^2 + 2ab + ac}{c - a}$

$\Rightarrow a^2c - a^3 + abc - a^2b = a^2b + 2ab^2 + abc - a^3 - 2a^2 - b - a^2c$

$\Rightarrow \quad ac = b^2$

Therefore, a, b and c are in GP.

12. ***(a)*** The equation of the line passing through the points $A(3, 7)$ and $B(6, 5)$ is

$$y - 7 = -\frac{2}{3}(x - 3) \text{ or } 2x + 3y - 27 = 0$$

Also, the equation of the circle with A and B as the end points of diameter is

$$(x - 3)(x - 6) + (y - 7)(y - 5) = 0$$

Now, the equation of the family of circles through A and B is

$$(x - 3)(x - 6) + (y - 7)(y - 5) + \lambda(2x + 3y - 27) = 0 \quad \text{...(i)}$$

If the circle belonging to this family touches the X-axis.

Then, the equation

$(x - 3)(x - 6) + (0 - 7)(0 - 5) + \lambda\{2x + 3(0) - 27\} = 0$

has two equal roots for which discriminant $D = 0$.

It gives two values of λ.

The equation of the common chord of circle (i) and $x^2 + y^2 - 4x - 6y - 3 = 0$ is the radical axis which is

$$[(x - 3)(x - 6) + (y - 7)(y - 5) + \lambda(2x + 3y - 27)] - [x^2 + y^2 - 4x - 6y - 3] = 0$$

$\Rightarrow \quad (2\lambda - 5)x + (3\lambda - 6)y + (-27\lambda + 56) = 0$

$\Rightarrow \quad (-5x - 6y + 56) + \lambda(2x + 3y - 27) = 0$

This is the family of lines which passes through the point of intersection of $-5x - 6y + 56 = 0$ and $2x + 3y - 27 = 0$, i.e. $\left(2, \dfrac{23}{3}\right)$.

If circle (i) cuts $x^2 + y^2 = 29$ orthogonally, then

$$0 + 0 = -29 + 56 - 27\lambda = 0 \text{ or } \lambda = 1$$

So, the required circle is $x^2 + y^2 - 7x - 9y + 26 = 0$ and the centre is $\left(\dfrac{7}{2}, \dfrac{9}{2}\right)$.

13. ***(a)*** Given circles will intersect orthogonally, if

$$2(1 \times 0 + k \times k) = 6 + k$$

$$\Rightarrow \quad 2k^2 - k - 6 = 0$$

$$\Rightarrow \quad (2k + 3)(k - 2) = 0$$

$$\Rightarrow \quad k = 2, -\frac{3}{2}$$

14. ***(a)*** The coordinates of the centres and radii of the given circles are

$C_1 : (2, 3) \quad C_2 : (-3, -9)$

$r_1 : 5 \quad r_2 : 8$

We find $r_1 + r_2 = C_1C_2 = 13$. So, the two circles touch each other externally. Hence, there are 3 common tangents.

15. ***(b)*** The common chord of the given circles is

$$6x + 14y + c + d = 0 \qquad \text{...(i)}$$

Since, $x^2 + y^2 + 4x + 22y + c = 0$ bisects the circumference of the circle $x^2 + y^2 - 2x + 8y - d = 0$. So, the common chord (i) passes through the centre of the second circle, i.e. $(1, -4)$.

$$\therefore \quad 6 - 56 + c + d = 0 \Rightarrow c + d = 50$$

16. ***(a)*** Let $\angle AB_1B_2 = \theta$. Then,

$$AD = r_1 \sin\theta \text{ and } AD = r_2 \cos\theta$$

$$\therefore \quad AD^2\left(\frac{1}{r_1^2} + \frac{1}{r_2^2}\right) = 1$$

$$\Rightarrow \quad AD = \frac{r_1r_2}{\sqrt{r_1^2 + r_2^2}}$$

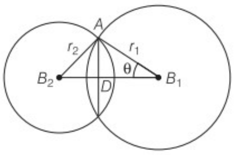

Thus, length of common chord is $\frac{2r_1r_2}{\sqrt{r_1^2 + r_2^2}}$.

17. ***(a)*** The centre of the circle $x^2 + y^2 = 2x$ is $(1, 0)$. The common chord of the other two circles is

$$8x - 15y + 26 = 0$$

The distance from $(1, 0)$ to $8x - 15y + 26 = 0$ is

$$\frac{|8 + 26|}{\sqrt{15^2 + 8^2}} = 2$$

18. ***(b)*** Given circles are

$$(x - 1)^2 + (y - 2)^2 = 1 \qquad \text{...(i)}$$

and

$$(x - 7)^2 + (y - 10)^2 = 4 \qquad \text{...(ii)}$$

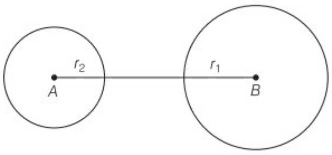

Let $A \equiv (1, 2), B \equiv (7, 10), r_1 = 1, r_2 = 2,$

$$AB = 10, r_1 + r_2 = 3$$

$$\because \quad AB > r_1 + r_2$$

Hence, the two circles are separated.

The radii of the two circles at time t are $(1 + 0.3t)$ and $(2 + 0.4t)$.

For the two circles to touch each other,

$$AB^2 = [(r_1 + 0.3t) \pm (r_2 + 0.4t)]^2$$

$$\Rightarrow \quad 100 = [(1 + 0.3t) \pm (2 + 0.4t)]^2$$

$$\Rightarrow \quad 100 = (3 + 0.7t)^2, [(0.1)t + 1]^2$$

$$\Rightarrow \quad 3 + 0.7t = \pm 10, 0.1t + 1 = \pm 10$$

$$\therefore \quad t = 10, t = 90 \qquad [\because t > 0]$$

The two circles will touch each other externally in 10 s and internally in 90 s.

19. ***(d)*** Let the centre of the circle C be (h, k).

Then, as the circle touches the X-axis, its radius is $|k|$.

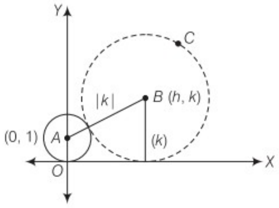

Also, it touches the given circle $x^2 + (y - 1)^2 = 1$ with centre $(0, 1)$ and radius 1, externally. Therefore,

Distance between centres = Sum of radii

$$\Rightarrow \quad \sqrt{(h - 0)^2 + (k - 1)^2} = 1 + |k|$$

$$\Rightarrow \quad h^2 + k^2 - 2k + 1 = 1 + 2|k| + k^2$$

$$\Rightarrow \quad h^2 = 2k + 2|k|$$

Therefore, the locus of (h, k) is $x^2 = 2y + 2|y|$.

Now, if $y > 0$, it becomes $x^2 = 4y$.

Also, if $y \le 0$, it becomes $x = 0$.

Combining the two, the required locus is

$$\{(x, y) : x^2 = 4y\} \cup \{(0, y) : y \le 0\}$$

20. ***(b)*** Let the circle S be $x^2 + y^2 + 2gx + 2fy + c = 0$. It is orthogonal to the circles $x^2 + y^2 - 2x - 15 = 0$ and $x^2 + y^2 - 1 = 0$.

Therefore, $2[g(-1) + f \times 0] = c - 15$

and $\quad 2[g \times 0 + f \times 0] = c - 1$

$$\Rightarrow \quad -2g = c - 15 \text{ and } c = 1$$

$$\Rightarrow \quad c = 1, g = 7$$

The circle S passes through $(0, 1)$.

$$\therefore \quad 1 + 2f + c = 0 \Rightarrow f = -1$$

Thus, the centre of the circle S is $(-7, 1)$ and radius is $\sqrt{49 + 1 - 1} = 7$.

‘0’ IS EITHER POSITIVE OR NEGATIVE

Asst. Prof. Naorem Rabei Singh
Pravabati College,
Mayang Imphal (MU)

ABOUT THE CHARACTER OF ZERO

0 is the integer immediately preceding 1. Zero is an even number because it is divisible by 2. Also, 0 is neither positive nor negative. By most definitions, 0 is number which quantifies a count or an amount of null size. In most cultures 0 was identified before the idea of negative things (quantities) that go lower than zero was accepted.

The number 0 is the smallest non-negative/non-positive integer. The natural number following 0 is 1 and no natural number precedes 0. The number 0 may or may not be considered a natural number but it is a whole number and hence a rational number and real number (as well as algebraic number and a complex number)

The number 0 is neither positive nor negative and appears in the middle of a number line.

As a new approach, the following theory is propounded.,

THEORY “0” (zero) has either positive or negative character.

PROOF Let’s discuss a simple division. Say 161 divided by 4

$$\begin{array}{l} 4\overline{\smash{)}161}\big(40 \\ \quad\underline{16} \\ \qquad 1 \\ \qquad \underline{0} \\ \qquad 1 \end{array}$$

Now, the remainder is 1 (one). Let us focus on this remainder. Generally, remainder occurs when something is taken out from something.

Mathematically, ‘Remainder’ occurs when a positive number is subtracted from a bigger positive number.

Then, 1 (one) cannot be called a remainder before subtracting 0 (zero) from 1 (one). But a question : why before or after subtracting 0 (zero) from 1 (one), 1 (one) itself exists arises. The answer to this question should be that 0 (zero) is a positive number, which is negligible. Hence, 0 (zero) has positive character.

Using “Euclid’s Division Lemma” which states as under

Let a and b be any two integers and $b > 0$. Then, there exists unique integers q and r such that $a = bq + r$ and $0 \leq r < b$. Here, q is called the quotient of a with respect to b and r is the remainder of a with respect to b. Any such type of division may be discussed.

Again, **Let us consider another case.**

We know that any positive number divided by zero is undefined.

Say $\frac{2}{0} = \infty$

Now, let us prove $\frac{2}{0} = \infty$

PROOF We get $\frac{2}{0.1} = 20$

$$\frac{2}{0.01} = 200$$

$$\frac{2}{0.001} = 2000$$

$$\frac{2}{0.0001} = 20000$$

....

$$\frac{2}{0.0001\,\ldots\ldots\ldots\,1} = 200\,\ldots\ldots\ldots$$

$$\Rightarrow \qquad \frac{2}{0} = 2000\,\ldots\ldots\ldots = \infty$$

It shows that $0.000\,\ldots\ldots\ldots\,1 = 0$

Therefore, zero is a positive number which is negligible.

Hence, 0 (zero) has positive character.

Also, we know that for any real number x having positive character, there exists number $(-x)$ having negative character. However, $0 = -0$; thereby 0 has negative character also.

Now, the concept, “0 is neither positive nor negative” is to be changed into “0 is either positive or negative”.

Let us, see the two series

(i) 0, 1, 2, 3,.......... (ii) – 0, – 1, – 2, – 3,........

Disclaimer,

This theory is propounded by N. Rabei Singh, Asst. Prof., P.B. College, Mayang Imphal (Manipur University). This is his exclusive intellectual and cannot be used, adopted and otherwise extracted/modified without prior permission in writing. Violation of this condition shall be dealt with strictly in accordance with law.

(!

CONCE
your Revi
VECTOR

VECTOR TRIPLE PRODUCT

For any three vectors **a**, **b** and **c** the vectors $\mathbf{a}\times(\mathbf{b}\times\mathbf{c})$ and $(\mathbf{a}\times\mathbf{b})\times\mathbf{c}$ are called vector triple products of **a**, **b**, **c**.

If any three vectors are **a**, **b** and **c**, then

$$\mathbf{a}\times(\mathbf{b}\times\mathbf{c})=(\mathbf{a}\cdot\mathbf{c})\,\mathbf{b}-(\mathbf{a}\cdot\mathbf{b})\,\mathbf{c}$$

AID TO MEMORY

$\mathbf{a}\times(\mathbf{b}\times\mathbf{c})=(\mathbf{a}\cdot\mathbf{c})\,\mathbf{b}-(\mathbf{a}\cdot\mathbf{b})\,\mathbf{c}$ is written as

$\text{I}\times(\text{II}\times\text{III})=(\text{I}\cdot\text{III})\,\text{II}-(\text{I}\cdot\text{II})\,\text{III}$

PROPERTIES OF VECTOR TRIPLE PRODUCT

1. The vector triple product $\mathbf{a}\times(\mathbf{b}\times\mathbf{c})$ is a linear combination of those two vectors which are within brackets.

❶ THE SCALAR OR DOT PRODUCT

Let **a** and **b** be two non-zero vectors inclined at an angle θ. Then, the scalar product of **a** with **b** is denoted by $\mathbf{a}\cdot\mathbf{b}$ and is defined as the scalar $|\mathbf{a}||\mathbf{b}|\cos\theta$.

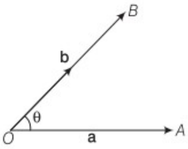

Thus, $\mathbf{a}\cdot\mathbf{b}=|\mathbf{a}||\mathbf{b}|\cos\theta$

REMARK If **a** or **b** or both is a zero vector, then θ is not defined as 0 has no direction. In this case their dot product $\mathbf{a}\cdot\mathbf{b}$ is defined as the scalar zero.

❷ PROJECTION

The projection of a vector is the scalar product of two vectors with the product of modulus of either vector and the projection of the other in its direction.

$\therefore$ Projection of **b** on $\mathbf{a}=\dfrac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}|}=\dfrac{\mathbf{a}}{|\mathbf{a}|}\cdot\mathbf{b}=\hat{\mathbf{a}}\cdot\mathbf{b}$

and projection of **a** on $\mathbf{b}=\dfrac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{b}|}=\dfrac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{b}|}=\mathbf{a}\cdot\hat{\mathbf{b}}$

❸ PROPERTIES OF SCALAR PRODUCT

Ⓐ **Property I** The scalar product of two vectors is commutative, i.e. $\mathbf{a}\cdot\mathbf{b}=\mathbf{b}\cdot\mathbf{a}$.

Ⓑ **Property II** The scalar product of vectors is distributive over vector addition, i.e.

(i) $\mathbf{a}\cdot(\mathbf{b}+\mathbf{c})=\mathbf{a}\cdot\mathbf{b}=\mathbf{a}\cdot\mathbf{c}$ [left distributive]

(ii) $(\mathbf{b}+\mathbf{c})\cdot\mathbf{a}=\mathbf{b}\cdot\mathbf{a}+\mathbf{c}\cdot\mathbf{a}$ [right distributive]

Ⓒ **Property III** Let **a** and **b** are two non-zero vectors. Then, $\mathbf{a}\cdot\mathbf{b}=0\Leftrightarrow\mathbf{a}\perp\mathbf{b}$

REMARK Since, $\hat{\mathbf{i}},\hat{\mathbf{j}}$ and $\hat{\mathbf{k}}$ are mutually perpendicular unit vectors along the coordinate axes. Therefore,

$$\hat{\mathbf{i}}\cdot\hat{\mathbf{j}}=\hat{\mathbf{j}}\cdot\hat{\mathbf{i}}=0,\ \hat{\mathbf{j}}\cdot\hat{\mathbf{k}}=\hat{\mathbf{k}}\cdot\hat{\mathbf{j}}=0,\ \hat{\mathbf{k}}\cdot\hat{\mathbf{i}}=\hat{\mathbf{i}}\cdot\hat{\mathbf{k}}=0$$

Ⓓ **Property IV** For any vector **a**, we have $\mathbf{a}\cdot\mathbf{a}=|\mathbf{a}|^2$

REMARK Since, $\hat{\mathbf{i}},\hat{\mathbf{j}}$ and $\hat{\mathbf{k}}$ are unit vectors along the coordinate axes. Therefore,

$\hat{\mathbf{i}}\cdot\hat{\mathbf{i}}=|\hat{\mathbf{i}}|^2=1,\ \hat{\mathbf{j}}\cdot\hat{\mathbf{j}}=|\hat{\mathbf{j}}|^2=1$ and $\hat{\mathbf{k}}\cdot\hat{\mathbf{k}}=|\hat{\mathbf{k}}|^2=1$

Ⓔ **Property V** If m is a scalar and **a**, **b** are any two vectors, then

$$(m\cdot\mathbf{a})\cdot\mathbf{b}=m(\mathbf{a}\cdot\mathbf{b})=\mathbf{a}\cdot(m\cdot\mathbf{b})$$

Ⓕ **Property VI** If m and n are scalars and **a**, **b** are two vectors, then

$m\mathbf{a}\cdot n\mathbf{b}=mn(\mathbf{a}\cdot\mathbf{b})=(mn\mathbf{a})\cdot\mathbf{b}=\mathbf{a}\cdot(mn\mathbf{b})$

Ⓖ **Property VII** For any two vectors **a** and **b**, we have

(i) $\mathbf{a}\cdot(-\mathbf{b})=-(\mathbf{a}\cdot\mathbf{b})=(-\mathbf{a})\cdot\mathbf{b}$

(ii) $(-\mathbf{a})\cdot(-\mathbf{b})=\mathbf{a}\cdot\mathbf{b}$

Ⓗ **Property VIII** For any two vectors **a** and **b**, we have

(i) $|\mathbf{a}+\mathbf{b}|^2=|\mathbf{a}|^2+|\mathbf{b}|^2+2\mathbf{a}\cdot\mathbf{b}$

(ii) $|\mathbf{a}-\mathbf{b}|^2=|\mathbf{a}|^2+|\mathbf{b}|^2-2\mathbf{a}\cdot\mathbf{b}$

(iii) $(\mathbf{a}+\mathbf{b})\cdot(\mathbf{a}-\mathbf{b})=|\mathbf{a}|^2-|\mathbf{b}|^2$

(iv) $|\mathbf{a}+\mathbf{b}|^2+|\mathbf{a}-\mathbf{b}|^2=2(|\mathbf{a}|^2+|\mathbf{b}|^2)$

(v) $|\mathbf{a}+\mathbf{b}|=|\mathbf{a}-\mathbf{b}|\Leftrightarrow\mathbf{a}\perp\mathbf{b}$

(vi) $|\mathbf{a}+\mathbf{b}|=|\mathbf{a}|+|\mathbf{b}|\Leftrightarrow\mathbf{a}$ is parallel to **b**.

(vii) $|\mathbf{a}+\mathbf{b}|^2=|\mathbf{a}|^2+|\mathbf{b}|^2$

$\Leftrightarrow$ **a** and **b** are orthogonal.

Ⓘ **Property IX** If **a**, **b** and **c** are three vectors, then

$$|\mathbf{a}+\mathbf{b}+\mathbf{c}|^2=|\mathbf{a}|^2+|\mathbf{b}|^2+|\mathbf{c}|^2+2(\mathbf{a}\cdot\mathbf{b}+\mathbf{b}\cdot\mathbf{c}+\mathbf{c}\cdot\mathbf{a})$$

❹ SCALAR PRODUCT IN TERMS OF COMPONENTS

Let $\mathbf{a}=a_1\hat{\mathbf{i}}+a_2\hat{\mathbf{j}}+a_3\hat{\mathbf{k}}$ and $\mathbf{b}=b_1\hat{\mathbf{i}}+b_2\hat{\mathbf{j}}+b_3\hat{\mathbf{k}}$.

Then, by using the distributivity of dot products over vector addition, we get

$$\mathbf{a}\cdot\mathbf{b}=a_1b_1+a_2b_2+a_3b_3$$

❺ ANGLE BETWEEN TWO VECTORS

Let **a** and **b** be two vectors inclined at an angle θ.

Then, $\mathbf{a}\cdot\mathbf{b}=|\mathbf{a}||\mathbf{b}|\cos\theta$

$$\Rightarrow\quad\cos\theta=\frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}|\cdot|\mathbf{b}|}$$

$$\Rightarrow\quad\theta=\cos^{-1}\left[\frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}|\cdot|\mathbf{b}|}\right]$$

If $\quad\mathbf{a}=a_1\hat{\mathbf{i}}+a_2\hat{\mathbf{j}}+a_3\hat{\mathbf{k}}$

and $\quad\mathbf{b}=b_1\hat{\mathbf{i}}+b_2\hat{\mathbf{j}}+b_3\hat{\mathbf{k}}$,

then $\mathbf{a}\cdot\mathbf{b}=a_1b_1+a_2b_2+a_3b_3$

$$|\mathbf{a}|=\sqrt{a_1^2+a_2^2+a_3^2},\ |\mathbf{b}|=\sqrt{b_1^2+b_2^2+b_3^2}$$

$$\theta=\cos^{-1}\left[\frac{a_1b_1+a_2b_2+a_3b_3}{\sqrt{a_1^2+a_2^2+a_3^2}\,\sqrt{b_1^2+b_2^2+b_3^2}}\right]$$

❻ COMPONENTS OF A VECTOR ALONG AND PERPENDICULAR TO ANOTHER VECTOR

The component of **b** along and perpendicular to **a** are

$$\left\{\frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}|^2}\right\}\mathbf{a}$$

and $\quad\mathbf{b}-\left\{\dfrac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}|^2}\right\}\mathbf{a}$, respectively.

❼ VECTOR OR CROSS PRODUCT

Let **a** and **b** are two non-zero, non-parallel vectors. Then, the vector product is defined as $\mathbf{a}\times\mathbf{b}$ in order as a vector whose magnitude is $|\mathbf{a}||\mathbf{b}|\sin\theta$.

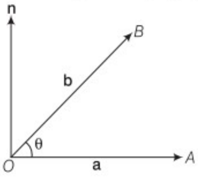

where θ is the angle between **a** and **b** and whose direction is perpendicular to the plane of **a** and **b** in such a way that **a**, **b** and this direction constitute a right handed system.

$\therefore\quad\mathbf{a}\times\mathbf{b}=|\mathbf{a}||\mathbf{b}|\sin\theta\,\hat{\mathbf{n}}$

where, $\hat{\mathbf{n}}$ is a unit vector perpendicular to the plane of **a** and **b**.

Note
- If one of **a** or **b** or both is **0**, then θ is not defined as **0** has no direction and so $\hat{\mathbf{n}}$ is not defined. In such a case, $\mathbf{a}\times\mathbf{b}=\mathbf{0}$.
- If **a** and **b** are collinear, i.e. if $\theta=0$ or π, then the direction of $\hat{\mathbf{n}}$ is not well-defined and so in this case also $\mathbf{a}\times\mathbf{b}=\mathbf{0}$.

❽ PROPERTIES OF VECTOR PRODUCT

Ⓐ **Property I** Vector product is not commutative,

i.e. $\quad\mathbf{a}\times\mathbf{b}\neq\mathbf{b}\times\mathbf{a}$

However, $\mathbf{a}\times\mathbf{b}=-(\mathbf{b}\times\mathbf{a})$

Ⓑ **Property II** If **a** and **b** are two vectors and m is a scalar, then

$$m\cdot\mathbf{a}\times\mathbf{b}=m(\mathbf{a}\times\mathbf{b})=\mathbf{a}\times m\mathbf{b}$$

Ⓒ **Property III** If **a** and **b** are two vectors and m,n are scalars, then

$$m\mathbf{a}\times n\mathbf{b}=mn(\mathbf{a}\times\mathbf{b})=m(\mathbf{a}\times n\mathbf{b})=n(m\mathbf{a}\times\mathbf{b})$$

Ⓓ **Property IV** Let **a**, **b** and **c** be any three vectors, then

(i) $\mathbf{a}\times(\mathbf{b}+\mathbf{c})=\mathbf{a}\times\mathbf{b}+\mathbf{a}\times\mathbf{c}$ [left distributivity]

(ii) $(\mathbf{b}+\mathbf{c})\times\mathbf{a}=\mathbf{b}\times\mathbf{a}+\mathbf{c}\times\mathbf{a}$ [right distributivity]

Ⓔ **Property V** For any three vectors **a**, **b** and **c**, we have

$$\mathbf{a}\times(\mathbf{b}-\mathbf{c})=\mathbf{a}\times\mathbf{b}-\mathbf{a}\times\mathbf{c}$$

Ⓕ **Property VI** The vector product of two non-zero vectors is zero vector iff they are parallel (collinear),

PT MAP
sion Tool
PRODUCT

2. The vector $\mathbf{r} = \mathbf{a} \times (\mathbf{b} \times \mathbf{c})$ is perpendicular to **a** and lies in the plane **b** and **c**.
3. The formula $\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \cdot \mathbf{c})\,\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\,\mathbf{c}$ is true only when the vector outside the bracket is on the left-most side. If it is not we first shift on left by using the properties of cross product and then apply the same formula.
$$\begin{aligned}(\mathbf{b} \times \mathbf{c}) \times \mathbf{a} &= -\{\mathbf{a} \times (\mathbf{b} \times \mathbf{c})\}\\ &= -\{(\mathbf{a} \cdot \mathbf{c})\,\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\,\mathbf{c}\}\\ &= -(\mathbf{a} \cdot \mathbf{c})\,\mathbf{b} + (\mathbf{a} \cdot \mathbf{b})\,\mathbf{c}\end{aligned}$$
4. For any three vectors **a**, **b** and **c** are the value of $\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) + \mathbf{b} \times (\mathbf{c} \times \mathbf{a}) + \mathbf{c} \times (\mathbf{a} \times \mathbf{b})$ is zero. Thus, $\mathbf{a} \times (\mathbf{b} \times \mathbf{c})$, $\mathbf{b} \times (\mathbf{c} \times \mathbf{a})$, $\mathbf{c} \times (\mathbf{a} \times \mathbf{b})$ are coplanar.

i.e. $\mathbf{a} \times \mathbf{b} = \mathbf{0} \Leftrightarrow \mathbf{a} \,||\, \mathbf{b}$, where **a** and **b** are non-zero vectors.

REMARK It follows from the above property that $\mathbf{a} \times \mathbf{a} = \mathbf{0}$ for every non-zero vector **a**.

Ⓖ **Property VII** Vector product of orthogonal triad of unit vectors $\hat{\mathbf{i}}, \hat{\mathbf{j}}$ and $\hat{\mathbf{k}}$ are $\hat{\mathbf{i}} \times \hat{\mathbf{i}} = \hat{\mathbf{j}} \times \hat{\mathbf{j}} = \hat{\mathbf{k}} \times \hat{\mathbf{k}} = \mathbf{0}$

$$\hat{\mathbf{i}} \times \hat{\mathbf{j}} = \hat{\mathbf{k}}, \hat{\mathbf{j}} \times \hat{\mathbf{k}} = \hat{\mathbf{i}}, \hat{\mathbf{k}} \times \hat{\mathbf{i}} = \hat{\mathbf{j}}$$
$$\hat{\mathbf{j}} \times \hat{\mathbf{i}} = -\hat{\mathbf{k}}, \hat{\mathbf{k}} \times \hat{\mathbf{j}} = -\hat{\mathbf{i}}, \hat{\mathbf{i}} \times \hat{\mathbf{k}} = -\hat{\mathbf{j}}$$

Ⓗ **Property VIII** Let $\mathbf{a} = a_1\hat{\mathbf{i}} + a_2\hat{\mathbf{j}} + a_3\hat{\mathbf{k}}$ and $\mathbf{b} = b_1\hat{\mathbf{i}} + b_2\hat{\mathbf{j}} + b_3\hat{\mathbf{k}}$ be two vectors.

Then, $\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix}$

Ⓘ **Property IX** Unit vectors normal to the plane containing two non-zero, non-parallel vectors **a** and **b** are given by $\pm \dfrac{\mathbf{a} \times \mathbf{b}}{|\mathbf{a} \times \mathbf{b}|}$.

Ⓙ **Property X** The area of parallelogram with adjacent sides **a** and **b** is $|\mathbf{a} \times \mathbf{b}|$.

Ⓚ **Property XI** The area of a triangle with adjacent sides **a** and **b** is $\frac{1}{2}|\mathbf{a} \times \mathbf{b}|$.

Ⓛ **Property XII** The area of a ΔABC is $\frac{1}{2}|\mathbf{AB} \times \mathbf{AC}|$ or $\frac{1}{2}|\mathbf{BC} \times \mathbf{BA}|$ or $\frac{1}{2}|\mathbf{CB} \times \mathbf{CA}|$.

Ⓜ **Property XIII** The area of a parallelogram with diagonals **a** and **b** is $\frac{1}{2}|\mathbf{a} \times \mathbf{b}|$.

Ⓝ **Property XIV** The area of a plane quadrilateral *ABCD* is $\frac{1}{2}|\mathbf{AC} \times \mathbf{BD}|$, where **AC** and **BD** are its diagonals.

⑨ SCALAR TRIPLE PRODUCT

Let **a**, **b** and **c** be three vectors. Then, the scalar $(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c}$ is called the scalar triple product of **a**, **b** and **c** and is denoted by $[\mathbf{a}\ \mathbf{b}\ \mathbf{c}]$.

Thus, $\quad [\mathbf{a}\ \mathbf{b}\ \mathbf{c}] = (\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c}$

The scalar triple product $[\mathbf{a}\ \mathbf{b}\ \mathbf{c}]$ represents the volume of the parallelopiped whose coterminous edges **a**, **b**, **c** form a right handed system of vectors.

⑩ PROPERTIES OF SCALAR TRIPLE PRODUCT

Ⓐ **Property I** If **a**, **b** and **c** are cyclically permuted, then value of scalar triple product remains same, i.e. $(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = (\mathbf{b} \times \mathbf{c}) \cdot \mathbf{a} = (\mathbf{c} \times \mathbf{a}) \cdot \mathbf{b}$

or $\quad [\mathbf{a}\ \mathbf{b}\ \mathbf{c}] = [\mathbf{b}\ \mathbf{c}\ \mathbf{a}] = [\mathbf{c}\ \mathbf{a}\ \mathbf{b}]$

Ⓑ **Property II** The change of cyclic order of vectors in scalar triple product changes the sign of the scalar triple product but not the magnitude, i.e.
$$\begin{aligned}[\mathbf{a}\ \mathbf{b}\ \mathbf{c}] &= -[\mathbf{b}\ \mathbf{a}\ \mathbf{c}]\\ &= -[\mathbf{c}\ \mathbf{b}\ \mathbf{a}]\\ &= -[\mathbf{a}\ \mathbf{c}\ \mathbf{b}]\end{aligned}$$

Ⓒ **Property III** In scalar triple product, the positions of dot and cross can be interchanged provided that the cyclic order of the vectors remain same, i.e.
$$(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$$

Ⓓ **Property IV** The scalar triple product of three vectors is zero, if any two of them are equal.

i.e. $\quad [\mathbf{a}\ \mathbf{a}\ \mathbf{c}] = 0$

Ⓔ **Property V** For any three vectors **a**, **b** and **c** and scalars λ, μ and ν, we have
$$[\lambda\mathbf{a}\ \mu\mathbf{b}\ \nu\mathbf{c}] = \lambda\mu\nu\ [\mathbf{a}\ \mathbf{b}\ \mathbf{c}]$$

Ⓕ **Property VI** The scalar triple product of three vectors is zero, if any two of them are parallel or collinear.

Ⓖ **Property VII** If **a**, **b**, **c** and **d** are four vectors, then
$$[\mathbf{a} + \mathbf{b}\ \mathbf{c}\ \mathbf{d}] = [\mathbf{a}\ \mathbf{c}\ \mathbf{d}] + [\mathbf{b}\ \mathbf{c}\ \mathbf{d}]$$

Ⓗ **Property VIII** Let $\mathbf{a} = a_1\hat{\mathbf{i}} + a_2\hat{\mathbf{j}} + a_3\hat{\mathbf{k}}$,
$$\mathbf{b} = b_1\hat{\mathbf{i}} + b_2\hat{\mathbf{j}} + b_3\hat{\mathbf{k}}$$
and $\quad \mathbf{c} = c_1\hat{\mathbf{i}} + c_2\hat{\mathbf{j}} + c_3\hat{\mathbf{k}}$

be the three vectors.

Then, we have
$$[\mathbf{a}\ \mathbf{b}\ \mathbf{c}] = \begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix}$$

Ⓘ **Property IX** The necessary and sufficient condition for three non-zero, non-collinear vectors **a**, **b**, **c** to be coplanar,

i.e. $\quad [\mathbf{a}\ \mathbf{b}\ \mathbf{c}] = 0.$

Ⓙ **Property X** Four points with position vectors **a**, **b**, **c** and **d** will be coplanar, if
$$[\mathbf{d}\ \mathbf{b}\ \mathbf{c}] + [\mathbf{d}\ \mathbf{c}\ \mathbf{a}] + [\mathbf{d}\ \mathbf{a}\ \mathbf{b}] = [\mathbf{a}\ \mathbf{b}\ \mathbf{c}]$$

Ⓚ **Property XI** If **a**, **b**, **c** are three non-coplanar vectors and **A**, **B**, **C** are three vectors such that
$$\mathbf{A} = x_1\mathbf{a} + y_1\mathbf{b} + z_1\mathbf{c},$$
$$\mathbf{B} = x_2\mathbf{a} + y_2\mathbf{b} + z_2\mathbf{c}$$
and $\quad \mathbf{C} = x_3\mathbf{a} + y_3\mathbf{b} + z_3\mathbf{c}$, then
$$\therefore \quad [\mathbf{A}\ \mathbf{B}\mathbf{C}] = \begin{vmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{vmatrix} [\mathbf{a}\ \mathbf{b}\ \mathbf{c}]$$

Ⓛ **Property XII** If **a**, **b**, **c** and **l**, **m**, **n** are two trials of vectors, then
$$[\mathbf{a}\ \mathbf{b}\ \mathbf{c}][\mathbf{l}\ \mathbf{m}\ \mathbf{n}] = \begin{vmatrix} \mathbf{a} \cdot \mathbf{l} & \mathbf{b} \cdot \mathbf{l} & \mathbf{c} \cdot \mathbf{l} \\ \mathbf{a} \cdot \mathbf{m} & \mathbf{b} \cdot \mathbf{m} & \mathbf{c} \cdot \mathbf{m} \\ \mathbf{a} \cdot \mathbf{n} & \mathbf{b} \cdot \mathbf{n} & \mathbf{c} \cdot \mathbf{m} \end{vmatrix}$$

Ⓜ **Property XIII** For any three vectors **a**, **b**, **c**,
$$[\mathbf{a}\ \mathbf{b}\ \mathbf{c}]^2 = \begin{vmatrix} \mathbf{a} \cdot \mathbf{a} & \mathbf{a} \cdot \mathbf{b} & \mathbf{a} \cdot \mathbf{c} \\ \mathbf{b} \cdot \mathbf{a} & \mathbf{b} \cdot \mathbf{b} & \mathbf{b} \cdot \mathbf{c} \\ \mathbf{c} \cdot \mathbf{a} & \mathbf{c} \cdot \mathbf{b} & \mathbf{c} \cdot \mathbf{c} \end{vmatrix}$$

Ⓝ **Property XIV** If **a**, **b**, **c**, **l**, **m** are vectors, then
$$[\mathbf{a}\ \mathbf{b}\ \mathbf{c}](\mathbf{l} \times \mathbf{m}) = \begin{vmatrix} \mathbf{a} & \mathbf{b} & \mathbf{c} \\ \mathbf{a} \cdot \mathbf{l} & \mathbf{b} \cdot \mathbf{l} & \mathbf{c} \cdot \mathbf{l} \\ \mathbf{a} \cdot \mathbf{m} & \mathbf{b} \cdot \mathbf{m} & \mathbf{c} \cdot \mathbf{m} \end{vmatrix}$$

⑪ VOLUME OF TETRAHEDRON

The volume of a tetrahedron whose three coterminous edges in the right handed system are **a**, **b** and **c** is $\frac{1}{6}[\mathbf{a}\ \mathbf{b}\ \mathbf{c}]$.

If A, B, C, D are the vertices of a tetrahedron *ABCD*, then its volume is numerically equal to
$$\frac{1}{6}[\mathbf{AB}\ \mathbf{AC}\ \mathbf{AD}]$$
or $\quad \frac{1}{6}[\mathbf{BC}\ \mathbf{BD}\ \mathbf{BA}]$

or $\quad \frac{1}{6}[\mathbf{CD}\ \mathbf{CB}\ \mathbf{CA}].$

RECIPROCAL SYSTEM OF VECTORS

Let **a**, **b** and **c** be three non-coplanar vectors and let
$$\mathbf{a}' = \frac{\mathbf{b} \times \mathbf{c}}{[\mathbf{a}\ \mathbf{b}\ \mathbf{c}]},$$
$$\mathbf{b}' = \frac{\mathbf{c} \times \mathbf{a}}{[\mathbf{a}\ \mathbf{b}\ \mathbf{c}]}$$
and $\quad \mathbf{c}' = \dfrac{\mathbf{a} \times \mathbf{b}}{[\mathbf{a}\ \mathbf{b}\ \mathbf{c}]}$

where, $\mathbf{a}', \mathbf{b}', \mathbf{c}'$ are said to form a reciprocal system of vectors for the vectors **a**, **b**, **c**.

(!!

MATHS GALAXY

Love Agarwal

CUBE ROOTS AND *n*th ROOTS OF UNITY

*n*TH ROOTS OF UNITY IS VERY IMPORTANT TOPIC OF COMPLEX NUMBER. SPECIALLY CUBE ROOTS OF UNITY (ω) IS ONE OF THE MOST IMPORTANT TOPIC FROM WHICH EVERY YEAR ATLEAST ONE QUESTION IS ASKED IN IIT JEE MAIN. THIS SECTION WILL COVER BASIC CONCEPT AND VARIOUS PROPERTIES OF CUBE ROOTS OF UNITY.

*n*th Roots of a Complex Number

Let $z = a + ib$ be a complex number and let $r(\cos\theta + i\sin\theta)$ be the polar form of z. Then, by De-Moivre's theorem,

$$z^{1/n} = r^{1/n}\left\{\cos\left(\frac{\theta}{n}\right) + i\sin\left(\frac{\theta}{n}\right)\right\}$$

where, $z^{1/n}$ has n distinct values.

Algorithm to Find the *n*th Roots of a Complex Number

Algorithm

***Step* I** Write the given complex number in polar form.

***Step* II** Add $2r\pi$ to the argument.

***Step* III** Apply De-Moivre's theorem.

***Step* IV** Put $r = 0, 1, 2, \ldots, (n-1)$, i.e. one less than the number in the denominator of the given index in the lowest form.

Example ① The number of roots of the equation $z^6 = -64$ whose real parts are non-negative, is

(a) 2 *(b)* 3 *(c)* 4 *(d)* 5

Sol. **(c)** We have, $z^6 = -64$

$$\Rightarrow \quad z = (-64)^{1/6}$$

$$\Rightarrow \quad z = 2(-1)^{1/6}$$

$$\Rightarrow \quad z = 2(\cos\pi + i\sin\pi)^{1/6}$$

$$\Rightarrow \quad z = 2\{\cos(2r\pi + \pi) + i\sin(2r\pi + \pi)\}^{1/6}, r \in Z$$

$$\Rightarrow z = 2\left\{\cos(2r+1)\frac{\pi}{6} + i\sin(2r+1)\frac{\pi}{6}\right\}; r = 0, 1, 2, 3, 4, 5$$

$$\Rightarrow z = 2\left(\frac{\sqrt{3}}{2} + \frac{i}{2}\right), 2i, 2\left(\frac{-\sqrt{3}}{2} + \frac{i}{2}\right), 2\left(\frac{-\sqrt{3}}{2} - \frac{i}{2}\right), -2i, 2\left(\frac{\sqrt{3}}{2} - \frac{i}{2}\right)$$

Clearly, four of these complex numbers have non-negative real parts.

POINT TO REMEMBER

De-Moivre's Theorem

If $n \in Z$, then $(\cos\theta + i\sin\theta)^n = \cos n\theta + i\sin n\theta$

*n*th Roots of Unity

Let $z = 1^{1/n} \Rightarrow z = (\cos 0 + i\sin 0)^{1/n}$

$\Rightarrow \quad z = (\cos 2r\pi + i\sin 2r\pi)^{1/n}$, where $r \in Z$

$\Rightarrow \quad z = \cos\dfrac{2r\pi}{n} + i\sin\dfrac{2r\pi}{n}$, where $r = 0, 1, 2, \ldots, (n-1)$

$\Rightarrow \quad z = e^{\frac{i2r\pi}{n}}; r = 0, 1, 2, \ldots, (n-1) \quad [\because e^{i\theta} = \cos\theta + i\sin\theta]$

$\Rightarrow \quad z = (e^{i2\pi/n})^r; r = 0, 1, 2, \ldots, (n-1)$

$\Rightarrow \quad z = \alpha^r$, where $\alpha = e^{\frac{i2\pi}{n}}$

Thus, nth roots of unity are

$$1\,(=\alpha^0), \alpha, \alpha^2, \alpha^3, \ldots, \alpha^{n-1},$$

where $\quad \alpha = e^{\frac{i2\pi}{n}} = \cos\left(\dfrac{2\pi}{n}\right) + i\sin\left(\dfrac{2\pi}{n}\right)$

Example 2 If z_1 and z_2 are two nth roots of unity, then $\arg\left(\dfrac{z_1}{z_2}\right)$ is a multiple of

(a) $n\pi$ (b) $\dfrac{3\pi}{n}$ (c) $\dfrac{2\pi}{n}$ (d) None of these

Sol. **(c)** We have, $z_1 = e^{\frac{i2r\pi}{n}}$ and $z_2 = e^{\frac{i2s\pi}{n}}$

where, r and s are integers lying between 0 and n including 0.

$\therefore \quad \arg\left(\dfrac{z_1}{z_2}\right) = \dfrac{2r\pi}{n} - \dfrac{2s\pi}{n} = 2(r-s)\dfrac{\pi}{n}$

Clearly, $\arg\left(\dfrac{z_1}{z_2}\right)$ is a multiple of $\dfrac{2\pi}{n}$.

Properties of *n*th Roots of Unity

Following are the properties of nth roots of unity

Property 1

nth roots of unity form a GP with common ratio $e^{i2\pi/n}$.

Property 2

Sum of nth roots of unity is always zero.

Let nth roots of unity are $1, \alpha, \alpha^2, \ldots, \alpha^{n-1}$.

$\therefore \quad 1 + \alpha + \alpha^2 + \ldots + \alpha^{n-1}$

$$= \frac{1-\alpha^n}{1-\alpha} = \frac{1-e^{i2\pi}}{1-\alpha} \qquad \left[\because \alpha = e^{\frac{i2\pi}{n}} \Rightarrow \alpha^n = e^{i2\pi}\right]$$

$$= \frac{1-1}{1-\alpha} = 0 \qquad [\because e^{i2\pi} = \cos 2\pi + i\sin 2\pi = 1]$$

Example 3 If α is a non-zero fourth root of unity, then the values of $\alpha^{4n-1} + \alpha^{4n-2} + \alpha^{4n-3}$, $n \in N$ is

(a) 0 (b) -1 (c) 3 (d) None of these

Sol. **(b)** $\alpha^{4n-1} + \alpha^{4n-2} + \alpha^{4n-3} = \alpha^{4n}(\alpha^{-1} + \alpha^{-2} + \alpha^{-3})$

$= \alpha^{-1} + \alpha^{-2} + \alpha^{-3} \qquad [\because \alpha^{4n} = 1]$

$= \alpha^3 + \alpha^2 + \alpha \qquad [\because \alpha^4 = 1]$

$= (1 + \alpha + \alpha^2 + \alpha^3) - 1$

$= 0 - 1 \qquad$ [$\because$ sum of nth roots of unity is always zero]

$= -1$

Property 3

Sum of pth powers of nth roots of unity is zero, if p is not a multiple of n.

Let nth roots of unity are $1, \alpha, \alpha^2, \ldots, \alpha^{n-1}$.

$\therefore \quad 1 + \alpha^p + (\alpha^2)^p + (r^3)^p + \ldots + (\alpha^{n-1})^p$

$$= 1 + \alpha^p + \alpha^{2p} + r^{3p} + \ldots + \alpha^{(n-1)p}$$

$$= \frac{1-(\alpha^p)^n}{1-\alpha^p} = \frac{1-(e^{i2\pi p/n})^n}{1-e^{i2\pi p/n}}$$

$$= \frac{1-e^{i2\pi p}}{1-e^{i2\pi p/n}} = \frac{1-1}{1-e^{i2\pi p/n}}$$

$$[\because e^{i2\pi p} = \cos 2\pi p + i\sin 2\pi p = 1]$$

$$= 0$$

Property 4

Sum of pth powers of nth roots of unity is n, if p is a multiple of n.

Let $p = \lambda n$, where $\lambda \in N$.

Then, $\quad \alpha^p = e^{\frac{i2\pi p}{n}} = e^{\frac{i2\pi\lambda n}{n}} = e^{i2\pi\lambda}$

$$= \cos 2\pi\lambda + i\sin 2\pi\lambda = 1$$

$\therefore \quad$ Sum of pth powers of nth roots of unity

$$= 1 + \alpha^p + \alpha^{2p} + \ldots + \alpha^{(n-1)p}$$

$$= 1 + 1 + 1 + \ldots + 1 = n$$

Property 5

Product of nth roots of unity is $(-1)^{n-1}$.

$\therefore$ Product $= 1 \cdot \alpha \cdot \alpha^2 \ldots \alpha^{n-1}$

$$= \alpha^{1+2+3+\ldots+(n-1)} = \alpha^{n(n-1)/2} = (\alpha^{n/2})^{n-1}$$

$$= (e^{i\pi})^{n-1} = (\cos\pi + i\sin\pi)^{n-1} = (-1)^{n-1}$$

Example 4 If $z_r = \cos\left(\dfrac{2r\pi}{5}\right) + i\sin\left(\dfrac{2r\pi}{5}\right)$; $r = 0, 1, 2, 3, 4$, then $z_0z_1z_2z_3z_4$ is equal to

(a) -1 (b) 0 (c) 1 (d) None of these

Sol. **(c)** Clearly, z_0, z_1, z_2, z_3, z_4 are fifth roots of unity. We know that the product of nth roots of unity is $(-1)^{n-1}$.

Therefore, $z_0z_1z_2z_3z_4 = (-1)^{5-1} = 1$

Cube Roots of Unity

Let $z = 1^{1/3}$

Then, $z^3 = 1 \quad \Rightarrow \quad z^3 - 1 = 0$

$\Rightarrow \quad (z-1)(z^2+z+1) = 0$

$\Rightarrow \quad z - 1 = 0 \text{ or } z^2 + z + 1 = 0$

$\Rightarrow \quad z = 1 \text{ or } z = \dfrac{-1 \pm \sqrt{3}i}{2}$

So, the cube roots of unity are 1, $(-1+\sqrt{3}i)/2$ and $(-1-\sqrt{3}i)/2$. Clearly, one of the roots of unity is real and other two are complex.

Note If we denote $\omega = \dfrac{-1+\sqrt{3}i}{2}$, then $\omega^2 = \dfrac{-1-\sqrt{3}i}{2}$.

Therefore, the cube roots of unity are 1, ω, ω^2.

Example 5 If ω is an imaginary cube root of unity, then the value of the expression $1(2-\omega)(2-\omega^2) + 2(3-\omega)(3-\omega^2) + \ldots + (n-1)(n-\omega)(n-\omega^2)$ is **[IIT JEE 2001]**

(a) $\left\{\dfrac{n(n+1)}{2}\right\}^2$ (b) $\left\{\dfrac{n(n+1)}{2}\right\}^2 - n$

(c) $\left\{\dfrac{n(n+1)}{2}\right\}^2 + n$ (d) None of these

Sol. **(b)** We have,

$$(z-1)(z-\omega)(z-\omega^2) = z^3 - 1$$

$$\therefore 1(2-\omega)(2-\omega^2) + 2(3-\omega)(3-\omega^2) + \ldots + (n-1)(n-\omega)(n-\omega^2)$$

$$= \sum_{r=2}^{n} (r-1)(r-\omega)(r-\omega^2)$$

$$= \sum_{r=2}^{n} (r^3 - 1) = \sum_{r=2}^{n} r^3 - \sum_{r=2}^{n} 1 = \sum_{r=1}^{n} r^3 - 1 - \sum_{r=2}^{n} 1$$

$$= \left[\frac{n(n+1)}{2}\right]^2 - 1 - (n-1) = \left\{\frac{n(n+1)}{2}\right\}^2 - n$$

Properties of Cube Roots

Following are the properties of the cube root of unity

(i) Cube roots of unity are 1, ω and ω^2, where $\omega = \dfrac{-1}{2} \pm i\dfrac{\sqrt{3}}{2} = e^{i2\pi/3}$.

(ii) $\arg(\omega) = \dfrac{2\pi}{3}$ and $\arg(\omega^2) = \dfrac{4\pi}{3}$

(iii) $z^3 - 1 = (z-1)(z-\omega)(z-\omega^2)$

(iv) ω and ω^2 are roots of the equation $z^2 + z + 1 = 0$.

(v) Cube roots of unity lie on the unit circle $|z| = 1$ and divide its circumference into three equal parts.

(vi) $1 + \omega^n + \omega^{2n} = \begin{cases} 0, & \text{if } n \text{ is not a multiple of } 3 \\ 3, & \text{if } n \text{ is a multiple of } 3 \end{cases}$

(vii) Cube roots of -1 are $-1, -\omega, -\omega^2$.

(viii) $z^3 + 1 = (z+1)(z+\omega)(z+\omega^2)$

(ix) $-\omega$ and $-\omega^2$ are roots of $z^2 - z + 1 = 0$.

(x) $a^3 + b^3 = (a+b)(a+b\omega)(a+b\omega^2)$

$a^3 - b^3 = (a-b)(a-b\omega)(a-b\omega^2)$

$a^2 + b^2 + c^2 - ab - bc - ca$
$= (a + b\omega + c\omega^2)(a + b\omega^2 + c\omega)$

$a^3 + b^3 + c^3 - 3abc$
$= (a+b+c)(a + b\omega + c\omega^2)(a + b\omega^2 + c\omega)$

Example 6 If ω is an imaginary cube root of unity, then the $(1+\omega-\omega^2)^7$ equals **[IIT JEE 1998]**

(a) 128ω (b) -128ω (c) $128\omega^2$ (d) $-128\omega^2$

Sol. **(d)** $(1+\omega-\omega^2)^7 = (-2\omega^2)^7 = -128\omega^2$

Example 7 If $z(2-2\sqrt{3}i)^2 = i(\sqrt{3}+i)^4$, then $\arg(z)$ is equal to

(a) $\dfrac{5\pi}{6}$ (b) $-\dfrac{\pi}{6}$ (c) $\dfrac{\pi}{6}$ (d) $\dfrac{7\pi}{6}$

Sol. **(b)** We have, $z(2-2\sqrt{3}i)^2 = i(\sqrt{3}+i)^4$

$$\Rightarrow \quad 4z(1-i\sqrt{3})^2 = i^5(1-i\sqrt{3})^4$$

$$\Rightarrow \quad 4z = i(1-i\sqrt{3})^2$$

$$\Rightarrow \quad 4z = i(-2\omega)^2$$

$$\Rightarrow \quad 4z = 4i\omega^2$$

$$\Rightarrow \quad z = i\omega^2 = \frac{\sqrt{3}}{2} - \frac{1}{2}i$$

Clearly, $\arg(z) = \dfrac{-\pi}{6}$

Example 8 Let $\omega\,(\neq 1)$ be a complex cube root of unity. If $(3-3\omega+2\omega^2)^{4n+3} + (2+3\omega-3\omega^2)^{4n+3} + (-3+2\omega+2\omega^2)^{4n+3} = 0$, then the set of possible value(s) of n is(are) **[JEE Advanced 2015]**

(a) N (b) $\{3k : k \in N\}$

(c) $N - \{3k : k \in N\}$ (d) $\{6k : k \in N\}$

Sol. **(c)** We have,

$$(3-3\omega+2\omega^2)^{4n+3} + (2+3\omega-3\omega^2)^{4n+3} + (-3+2\omega+3\omega^2)^{4n+3} = 0$$

$$\Rightarrow \quad (3-3\omega+2\omega^2)^{4n+3} + \{\omega(3-3\omega+2\omega^2)\}^{4n+3} + \{\omega^2(3-3\omega+2\omega^2)\}^{4n+3} = 0$$

$$\Rightarrow \quad (3-3\omega+2\omega^2)^{4n+3}(1+\omega^{4n+3}+\omega^{8n+6}) = 0$$

$$\Rightarrow \quad (3-3\omega+2\omega^2)^{4n+3}(1+\omega^{4n}+\omega^{8n}) = 0$$

$$\Rightarrow \quad (3-3\omega+2\omega^2)^{4n+3}(1+\omega^{n}+\omega^{2n}) = 0$$

$$\Rightarrow \quad 1+\omega^n+\omega^{2n} = 0$$

$\Rightarrow$ n is not a multiple of 3.

Hence, the set of values of n is $N - \{n : n = 3k, k \in N\}$

29 04 2020

G) 157) 20200429 2246

A) Books_74_2020/
Magazine/

Spectrum_Mathematics_-_November_2016.pdf

→ family of circles. - Pág. 03

properties of triangles. - Pág. 66.

Pág. 13 "0" is either positive or negative.

Pág. 42 Cube Roots and nth Roots of Unity

→ nth roots of unity is very important topic of complex numbers.

B) Books_74_2020/
Magazine/

Spectrum_Mathematics_-_July_2016.pdf

→ Tangent to a circle. Pág. 42-49.

Look pages 40 e 41.

Trigonometric Equations and Identities. (resume)

Pág. 54 - Inverse Trigonometric Functions

2)*+,)))-) ' . / /
/4 5/ 3 ."
(!

CONCE
your Revi
Trigonometric Equa

❶ ANGLE

Measure of an angle is the amount of rotation of a revolving line with respect to a fixed line. If the rotation is in clockwise sense, then the measure of the angle is negative and it is positive, if the rotation is in anti-clockwise sense.

There are three systems of measuring an angle *viz.*

(i) Sexagesimal system or English system
(ii) Circular system (iii) French system

First two of these systems are commonly used. In sexagesimal system, a right angle is divided in 90 equal parts called degrees. Further, each degree is divided into sixty equal parts called minutes and each minute is divided into sixty equal parts called seconds.

Thus, 1 right angle = 90 degrees (90°)

$1° = 60$ minutes $(60')$

$1' = 60$ seconds $(60'')$

In circular system the unit of measurement is radian. One radian is the angle made by an arc of length equal to radius of a given circle at its centre.

❷ RELATION BETWEEN DEGREE AND RADIAN

If D is the degree measure of an angle and R is its measure in radians, then $\frac{D}{90} = \frac{2R}{\pi}$

$\therefore 1 \text{ radian} = \frac{180}{\pi} \text{ degrees} = 57°17'45''$ (approximately)

and 1 degree $= \frac{\pi}{180}$ radian

❸ DOMAIN AND RANGE OF TRIGONOMETRICAL FUNCTIONS

	Domain	Range
$\sin A$	R	$[-1, 1]$
$\cos A$	R	$[-1, 1]$
$\tan A$	$R - \{(2n+1)\pi/2 : n \in Z\}$	R
$\cot A$	$R - \{n\pi : n \in Z\}$	R
$\operatorname{cosec} A$	$R - \{n\pi : n \in Z\}$	$R - (-1, 1)$
$\sec A$	$R - \{(2n+1)\pi/2 : n \in Z\}$	$R - (-1, 1)$

❹ SUM AND DIFFERENCE FORMULAE

1. $\sin(A + B) = \sin A \cos B + \cos A \sin B$
2. $\sin(A - B) = \sin A \cos B - \cos A \sin B$
3. $\cos(A + B) = \cos A \cos B - \sin A \sin B$
4. $\cos(A - B) = \cos A \cos B + \sin A \sin B$
5. $\tan(A + B) = \dfrac{\tan A + \tan B}{1 - \tan A \tan B}$
6. $\tan(A - B) = \dfrac{\tan A - \tan B}{1 + \tan A \tan B}$
7. $\cot(A + B) = \dfrac{\cot A \cot B - 1}{\cot B + \cot A}$
8. $\cot(A - B) = \dfrac{\cot A \cot B + 1}{\cot B - \cot A}$
9. $\cos(A + B)\cos(A - B) = \cos^2 A - \sin^2 B = \cos^2 B - \sin^2 A$
10. $\sin(A + B)\sin(A - B) = \sin^2 A - \sin^2 B = \cos^2 B - \cos^2 A$
11. $\sin 2\theta = 2 \sin\theta\cos\theta = \dfrac{2\tan\theta}{1 + \tan^2\theta}$
12. $\cos 2\theta = \cos^2\theta - \sin^2\theta = 2\cos^2\theta - 1 = 1 - 2\sin^2\theta = \dfrac{1 - \tan^2\theta}{1 + \tan^2\theta}$
13. $\cos^2\theta = \dfrac{1 + \cos 2\theta}{2}, \sin^2\theta = \dfrac{1 - \cos 2\theta}{2}$
14. $\tan 2\theta = \dfrac{2\tan\theta}{1 - \tan^2\theta}$
15. $\dfrac{1 - \cos\theta}{\sin\theta} = \tan\dfrac{\theta}{2}, \dfrac{1 + \cos\theta}{\sin\theta} = \cot\dfrac{\theta}{2}$
16. $\dfrac{1 - \cos\theta}{1 + \cos\theta} = \tan^2\dfrac{\theta}{2}, \dfrac{1 + \cos\theta}{1 - \cos\theta} = \cot^2\dfrac{\theta}{2}$
17. $\sin 3\theta = 3\sin\theta - 4\sin^3\theta$
18. $\cos 3\theta = 4\cos^3\theta - 3\cos\theta$
19. $\tan 3\theta = \dfrac{3\tan\theta - \tan^3\theta}{1 - 3\tan^2\theta}$
20. $\cos A \cos 2A \cos 2^2 A \ldots \cos 2^{n-1} A = \dfrac{\sin 2^n A}{2^n \sin A}$
21. $\sin\theta\sin(60° - \theta)\sin(60° + \theta) = \dfrac{1}{4}\sin 3\theta$
22. $\cos\theta\cos(60° - \theta)\cos(60° + \theta) = \dfrac{1}{4}\cos 3\theta$
23. $\tan\theta\tan(60° - \theta)\tan(60° + \theta) = \tan 3\theta$

❺ SUM AND DIFFERENCE INTO PRODUCTS

1. $\sin A + \sin B = 2\sin\left(\dfrac{A + B}{2}\right)\cos\left(\dfrac{A - B}{2}\right)$
2. $\sin A - \sin B = 2\sin\left(\dfrac{A - B}{2}\right)\cos\left(\dfrac{A + B}{2}\right)$
3. $\cos A + \cos B = 2\cos\left(\dfrac{A + B}{2}\right)\cos\left(\dfrac{A - B}{2}\right)$
4. $\cos A - \cos B = -2\sin\left(\dfrac{A + B}{2}\right)\sin\left(\dfrac{A - B}{2}\right)$

❻ MAXIMUM AND MINIMUM VALUES

We know that $-1 \le \sin x \le 1$, $-1 \le \cos x \le 1$, $-\infty < \tan x < \infty$, $|\sec x| \ge 1$ and $|\operatorname{cosec} x| \ge 1$. If there is a trigonometrical expression of the form $a\sin x + b\cos x$, then by putting $a = r\cos\theta, b = r\sin\theta$, we have

$a\sin x + b\cos x = r\cos\theta\sin x + r\sin\theta\cos x$

$\Rightarrow \quad a\sin x + b\cos x = r\sin(x + \theta)$

where, $r = \sqrt{a^2 + b^2}, \tan\theta = \dfrac{b}{a}$

We have, $-1 \le \sin(x + \theta) \le 1$ for all values of x.

$\Rightarrow \quad -r \le r\sin(x + \theta) \le r$ for all x.

$\Rightarrow \quad -\sqrt{a^2 + b^2} \le a\sin x + b\cos x \le \sqrt{a^2 + b^2}, \forall x$

Hence, the maximum and minimum values of trigonometrical expressions of the form $a\sin x + b\cos x$ are $\sqrt{a^2 + b^2}$ and $-\sqrt{a^2 + b^2}$, respectively.

❼ PROPERTIES OF TRIANGLES

In any ΔABC, the sides BC, opposite to $\angle A$ is denoted by a; CA and AB, opposite to $\angle B$ and $\angle C$ respectively are denoted by b and c, respectively. Semi-perimeter of the triangle is denoted by s and its area by Δ.

GRAPHS OF TRIGONOME

Parent Function	Graph
$y = \sin(x)$ Odd Domain : $(-\infty, \infty)$ Range : $[-1, 1]$ Period : 2π Zeros : $(\pi K, 0)$ $K \in$ Integers	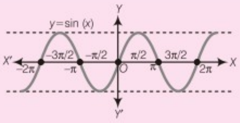
$y = \cos(x)$ Even Domain : $(-\infty, \infty)$ Range : $[-1, 1]$ Period : 2π Zeros : $\left(\dfrac{\pi}{2} + \pi K, 0\right)$	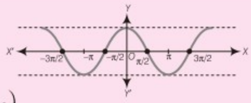
$y = \tan(x)$ Odd Domain : $x \ne \dfrac{\pi}{2} + \pi K$ Range : $(-\infty, \infty)$ Period : π Zeros : $(\pi K, 0)$	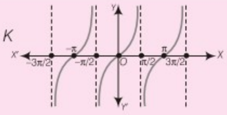

Ⓐ SINE FORMULAE

In any ΔABC, $\dfrac{\sin A}{a} = \dfrac{\sin B}{b} = \dfrac{\sin C}{c}$

i.e. the sines of the angles are proportional to the lengths of the opposite sides.

Ⓑ COSINE FORMULAE

In any ΔABC, $\cos A = \dfrac{b^2 + c^2 - a^2}{2bc}$;

$\cos B = \dfrac{c^2 + a^2 - b^2}{2ac}; \cos C = \dfrac{a^2 + b^2 - c^2}{2ab}$

Ⓒ PROJECTION FORMULAE

In any ΔABC, we have

$a = b\cos C + c\cos B;\quad b = c\cos A + a\cos C$

$c = a\cos B + b\cos A$

Ⓓ AREA OF ΔABC

In a ΔABC, its area Δ is given by

$$\Delta = \frac{1}{2}ab\sin C = \frac{1}{2}bc\sin A = \frac{1}{2}ca\sin B$$

$$\Delta = \sqrt{s(s - a)(s - b)(s - c)}, \text{ where } s = \frac{a + b + c}{2}$$

❽ CIRCLES CONNECTED TO A TRIANGLE

Ⓐ CIRCUMCIRCLE OF A TRIANGLE

The circle which passes through the angular points of a ΔABC is called its circumcircle. The centre of this circle

PT MAP

ision Tool

tions and Identities

TRIC FUNCTION

Parent Function	Graph
$y = \cot(x)$ Odd Domain : $x \neq \pi K$ Range : $(-\infty, \infty)$ Period : π Zeros : $\left(\frac{\pi}{2} + \pi K, 0\right)$	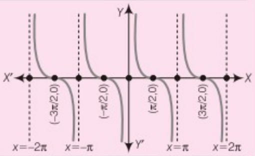
$y = \sec(x)$ Even Domain : $x \neq \frac{\pi}{2} + \pi K$ Range: $(-\infty, -1] \cup [1, \infty)$ Period : 2π Zeros : None	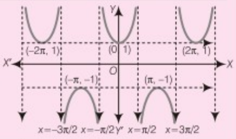
$y = \text{cosec}(x)$ Odd Domain: $x \neq \pi K$ Range: $(-\infty, -1] \cup [1, \infty)$ Period: 2π Zeros: None	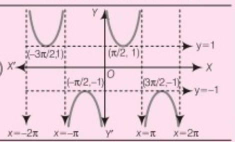

is the ABC is the point of intersection of perpendicular bisectors of the sides and is called the circumcentre. Its radius is always denoted by R.

Ⓑ CIRCUMRADIUS

The radius of the circumcircle of a ΔABC is called the circumradius given by

(i) $R = \frac{a}{2\sin A} = \frac{b}{2\sin B} = \frac{c}{2\sin c}$ (ii) $R = \frac{abc}{4\Delta}$

Ⓒ INSCRIBED CIRCLE OR INCIRCLE OF A TRIANGLE

The circle which can be inscribed within the triangle so as to touch each of its sides is called its inscribed circle or incircle. The centre of this circle is the point of intersection of bisectors of the angles of the triangle. The radius of this circle is always denoted by r. The radius of the inscribed circle of a triangle is called the inradius. It is denoted by r and is given by

(i) $r = \frac{\Delta}{s}$

(ii) $r = (s-a)\tan\frac{A}{2} = (s-b)\tan\frac{B}{2} = (s-c)\tan\frac{C}{2}$

(iii) $r = \frac{a\sin B/2 \sin C/2}{\cos A/2} = \frac{b\sin A/2 \sin C/2}{\cos B/2}$
$= \frac{c\sin B/2 \sin A/2}{\cos C/2}$

(iv) $r = 4R\sin\frac{A}{2}\sin\frac{B}{2}\sin\frac{C}{2}$

Ⓓ ESCRIBED CIRCLES OF A TRIANGLES

The circle which touches the sides BC and two sides AB and AC produced of a ΔABC is called the escribed circle opposite to $\angle A$. Its radius is denoted by r_1. Similarly, r_2 and r_3 denote the radii of the escribed circles opposite to $\angle B$ and $\angle C$, respectively.

The centres of the escribed circles are called the excentres. The centre of the escribed circle opposite to $\angle A$ is the point of intersection of the external bisectors of $\angle B$ and $\angle C$.

In ΔABC,

(i) $r_1 = \frac{\Delta}{s-a}, r_2 = \frac{\Delta}{s-b}, r_3 = \frac{\Delta}{s-c}$

(ii) $r_1 = s\tan A/2,\ r_2 = s\tan B/2,\ r_3 = s\tan C/2$

(iii) $r_1 = \frac{a\cos B/2 \cos C/2}{\cos A/2}; r_2 = \frac{b\cos C/2 \cos A/2}{\cos B/2}$

$r_3 = \frac{c\cos A/2 \cos B/2}{\cos C/2}$

(iv) $r_1 = 4R\sin\frac{A}{2}\cos\frac{B}{2}\cos\frac{C}{2}$

$r_2 = 4R\cos\frac{A}{2}\sin\frac{B}{2}\cos\frac{C}{2}$

$r_3 = 4R\cos\frac{A}{2}\cos\frac{B}{2}\sin\frac{C}{2}$

⑨ AREA OF A CYCLIC QUADRILATERAL

A quadrilateral is a cyclic quadrilateral, if its vertices lie on a circle. If s is the semi-perimeter of a cyclic quadrilateral $ABCD$ such that $AB = a, BC = b, CD = c$ and $DA = d$, then its area Δ is given by

$$\Delta = \sqrt{(s-a)(s-b)(s-c)(s-d)}$$

⑩ SOME IMPORTANT FACTS

(i) In an isosceles triangle, the median is perpendicular to the base ΔABC is isosceles and $AD \perp BC$.

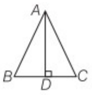

(ii) In similar triangle, the corresponding sides are proportional.

(iii) The exterior angle of a triangle is equal to the sum of interior opposite angles.

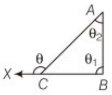

In ΔABC, $\theta = \theta_1 + \theta_2$

(iv) In a triangle, the internal bisector of an angle divides the opposite side in the ratio of the arms of the angle.

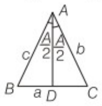

In ΔABC, AD is bisector of $\angle A$.

$\therefore \quad \frac{BD}{DC} = \frac{c}{b}$

(v) In a ΔABC, if $DE \parallel AB$, then

$\frac{AB}{DE} = \frac{BC}{DC} = \frac{AC}{EC}$

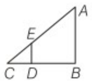

(vi) In a right angled ΔABC, the mid-point D of hypotenuse AC is equidistant from its vertices A, B and C.

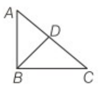

i.e. $AD = BD = CD$

(vii) ***m*-*n* Theorem** In a ΔABC, D divides BC in the ratio $m : n$, then

(a) $(m+n)\cot\theta = n\cot B - m\cot C$

(b) $(m+n)\cot\theta = m\cot\alpha - n\cot\beta$

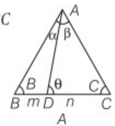

(viii) **Appolonius Theorem**

If in ΔABC, AD is median, then

$$AB^2 + AC^2 = 2(AD^2 + BD^2)$$

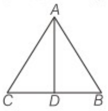

(ix) Angles in the same segment of a circle are equal.

(x) The angle subtended at the centre by an arc is double of the angle subtended at any point of the remaining arc by the same arc.

(xi) **Ptolemy's Theorem**

Incyclic quadrilateral $ABCD$,

$$AB \cdot CD + AD \cdot BC = AC \cdot BD$$

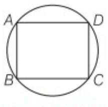

⑪ SOLUTIONS OF TRIGONOMETRIC EQUATIONS

A solution of a trigonometric equation is the value of the unknown angles that satisfies the equation. A solution generalised by means of periodicity is known as the general solution. Solving an equation means to find its general solution. Following are the general solutions of some trigonometric equations

	Equations	General solution
(i)	$\sin\theta = 0$	$\theta = n\pi, n \in Z$
(ii)	$\cos\theta = 0$	$\theta = (2n+1)\pi/2, n \in Z$
(iii)	$\tan\theta = 0$	$\theta = n\pi, n \in Z$
(iv)	$\cot\theta = 0$	$\theta = (2n+1)\pi/2, n \in Z$
(v)	$\sin\theta = \sin\alpha$	$\theta = n\pi + (-1)^n\alpha, n \in Z$
(vi)	$\cos\theta = \cos\alpha$	$\theta = 2n\pi \pm \alpha, n \in Z$
(vii)	$\tan\theta = \tan\alpha$	$\theta = n\pi + \alpha, n \in Z$
(viii)	$\sin^2\theta = \sin^2\alpha$ $\cos^2\theta = \cos^2\alpha$ $\tan^2\theta = \tan^2\alpha$	$\theta = n\pi \pm \alpha, n \in Z$

TRIGONOMETRIC RATIOS OF SOME SPECIAL AND USEFUL ANGLES

- $\sin 7\frac{1^\circ}{2} = \sqrt{\frac{4-\sqrt{2}-\sqrt{6}}{2\sqrt{2}}}, \cos 7\frac{1^\circ}{2} = \frac{\sqrt{4+\sqrt{2}+\sqrt{6}}}{2\sqrt{2}}$
  $\tan 7\frac{1^\circ}{2} = (\sqrt{3}-\sqrt{2})(\sqrt{2}-1)$
- $\sin 15^\circ = \frac{\sqrt{3}-1}{2\sqrt{2}}, \cos 15^\circ = \frac{\sqrt{3}+1}{2\sqrt{2}}, \tan 15^\circ = 2-\sqrt{3}$
- $\sin 22\frac{1^\circ}{2} = \frac{1}{2}\sqrt{2-\sqrt{2}}, \cos 22\frac{1^\circ}{2} = \frac{1}{2}\sqrt{2+\sqrt{2}}$
  $\tan 22\frac{1^\circ}{2} = \sqrt{2}-1$
- $\sin 18^\circ = \frac{\sqrt{5}-1}{4}, \cos 18^\circ = \frac{\sqrt{10+2\sqrt{5}}}{4}$
  $\tan 18^\circ = \frac{\sqrt{25-10\sqrt{15}}}{5}$
- $\sin 36^\circ = \frac{\sqrt{10-2\sqrt{5}}}{4}, \cos 36^\circ = \frac{\sqrt{5}+1}{4}, \tan 36^\circ = \sqrt{5-2\sqrt{5}}$
- $\sin 9^\circ = \frac{\sqrt{3+\sqrt{5}}-\sqrt{5-\sqrt{5}}}{4}$,
  $\cos 9^\circ = \frac{\sqrt{3+\sqrt{5}}+\sqrt{5-\sqrt{5}}}{4}$

TANGENT TO A CIRCLE

Coverage of special topics for JEE Advanced

In this section, we will learn about various equations of tangent to a circle. We will also discuss the pair of tangent drawn from a point and common tangents. In JEE Advanced, atleast one question is asked every year from this topic.

Tangent to a Circle

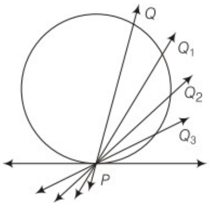

Let P be a point on a circle and let PQ be a secant. Now, as the point Q moves along the circle towards P, the line PQ rotates about P and as Q approaches to P, the line PQ coincides with the limiting line PT. This limiting position PT of secant PQ as Q tends to P is called the tangent line to the circle at point P.

Thus, the tangent at a point P is the limiting position of a secant PQ, when Q tends to P along the circle. The point P is called the point of contact of the tangent.

Different Forms of the Equations of Tangents

Following are the different forms of the equations of tangents to a circle.

1. **Slope Form**

The equation of a tangent of slope m to the circle $x^2 + y^2 = a^2$ is $y = mx \pm a\sqrt{1+m^2}$. The coordinates of the point of contact are $\left(\pm \frac{am}{\sqrt{1+m^2}}, \mp \frac{a}{\sqrt{1+m^2}}\right)$

2. **Point Form**

The equation of the tangent at the point $P(x_1, y_1)$ to a circle $x^2 + y^2 + 2gx + 2fy + c = 0$ is

$$xx_1 + yy_1 + g(x + x_1) + f(y + y_1) + c = 0$$

3. **Parametric Form**

The equation of the tangent to the circle $x^2 + y^2 = a^2$ at the point $(a\cos\theta, a\sin\theta)$ is $x\cos\theta + y\sin\theta = a$.

REMEMBER

1. The equations of the tangents of slope m to the circle $(x-a)^2 + (y-b)^2 = r^2$ are given by $y - b = m(x-a) \pm r\sqrt{1+m^2}$ and the coordinates of the point of contact are
$$\left(a \pm \frac{mr}{\sqrt{1+m^2}}, b \mp \frac{r}{\sqrt{1+m^2}}\right)$$
2. The equations of the tangents of slope m to the circle $x^2 + y^2 + 2gx + 2fy + c = 0$ are
$$y + f = m(x+y) \pm \sqrt{(g^2 + f^2 - c)(1+m^2)}$$
3. The equation of the tangent to the circle $(x-a)^2 + (y-b)^2 = r^2$ at the point $(a + r\cos\theta, b + r\sin\theta)$ is
$$(x-a)\cos\theta + (y-b)\sin\theta = r$$

Example ① The centre of circle inscribed in square formed by the lines $x^2 - 8x + 12 = 0$ and $y^2 - 14y + 45 = 0$, is

[JEE Main 2003]

(a) (4, 7)
(b) (7, 4)
(c) (9, 4)
(d) (4, 9)

Sol. **(a)** Given, circle is inscribed in square formed by the lines $x^2 - 8x + 12 = 0$ and $y^2 - 14y + 45 = 0$

$\Rightarrow x = 6, x = 2, y = 5, y = 9$

which could be plotted shown as figure where, $ABCD$ clearly forms a square.

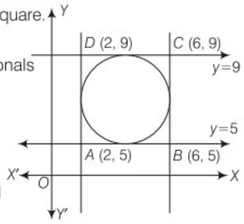

$\therefore$ Centre of inscribed circle

= Point of intersection of diagonals

= Mid-point of AC or BD

$= \left(\frac{2+6}{2}, \frac{5+9}{2}\right)$

$= (4, 7)$

Hence, the centre of inscribed circle is (4, 7).

Example 2 Let $ABCD$ be a quadrilateral with area 18, with side AB parallel to the side CD and $AB = 2CD$. Let AD be perpendicular to AB and CD. If a circle is drawn inside the quadrilateral $ABCD$ touching all the sides, then its radius, is **[JEE Main 2007]**

(a) 3 (b) 2 (c) 3/2 (d) 1

Sol. (b) $\because$ Area of trapezium $ABCD = 1/2$ (Sum of parallel sides) (Perpendicular distance between them)

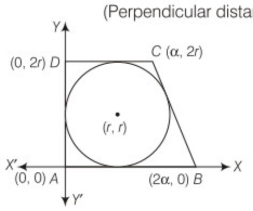

Then, $18 = \frac{1}{2}(3\alpha)(2r) \Rightarrow \alpha r = 6$ ...(i)

Also, line $y = -\frac{2r}{\alpha}(x - 2\alpha)$ is tangent to circle

$$(x - r)^2 + (y - r)^2 = r^2 \Rightarrow 2\alpha = 3r \quad \text{...(ii)}$$

From Eqs. (i) and (ii), we get

$$r = 2$$

Example 3 The lines $3x - 4y + 4 = 0$ and $6x - 8y - 7 = 0$ tangents to the same circle. The radius of this circle is .... **[JEE Main 1984]**

Sol. Since, $3x - 4y + 4 = 0$ and $3x - 4y - \frac{7}{2} = 0$ are two parallel tangents. Thus, distance between them is diameter of circle.

$$\therefore \text{Diameter} = \frac{\left|4 + \frac{7}{2}\right|}{\sqrt{3^2 + 4^2}} = \frac{15}{2 \times 5} = \frac{3}{2}$$

$$\therefore \quad \text{Radius} = \frac{3}{4}$$

Example 4 Let A be the centre of the circle $x^2 + y^2 - 2x - 4y - 20 = 0$. Suppose that, the tangents at the points $B(1, 7)$ and $D(4, -2)$ on the circle meet at the point C. Find the area of the quadrilateral $ABCD$. **[IIT JEE 1981]**

Sol. Equations of tangents at (1, 7) and (4, −2) are

$$x + 7y - (x + 1) - 2(y + 7) - 20 = 0$$

$$\Rightarrow \quad y = 7$$

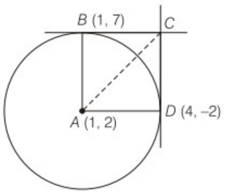

and $4x - 2y - (x + 4) - 2(y - 2) - 20 = 0$

$$\Rightarrow \quad 3x - 4y = 20$$

$\therefore$ Point C is (16, 7).

$\therefore$ Vertices of a quadrilateral are

$A(1, 2)$, $B(1, 7)$, $C(16, 7)$ and $D(4, -2)$.

$\therefore$ Area of quadrilateral $ABCD$

$$= \text{Area of } \Delta ABC + \text{Area of } \Delta ACD$$

$$= \frac{1}{2} \times 15 \times 5 + \frac{1}{2} \times 15 \times 5 = 75 \text{ sq units}$$

Example 5 Find the equations of the circles passing through $(-4, 3)$ and touching the lines $x + y = 2$ and $x - y = 2$. **[IIT JEE 1982]**

Sol. Let the equation of the required circle be

$$x^2 + y^2 + 2gx + 2fy + c = 0 \quad \text{...(i)}$$

It passes through (−4, 3).

$$\therefore \quad 25 - 8g + 6f + c = 0 \quad \text{...(ii)}$$

Since, circle touches the lines $x + y = 2$ and $x - y = 2$.

$$\therefore \quad \left|\frac{-g - f - 2}{\sqrt{2}}\right| = \left|\frac{-g + f - 2}{\sqrt{2}}\right| = \sqrt{g^2 + f^2 - c} \quad \text{...(iii)}$$

$$\text{Now,} \quad \left|\frac{-g - f - 2}{\sqrt{2}}\right| = \left|\frac{-g + f - 2}{\sqrt{2}}\right|$$

$$\Rightarrow \quad -g - f - 2 = \pm(-g + f - 2)$$

$$\Rightarrow \quad -g - f - 2 = -g + f - 2$$

$$\text{or} \quad -g - f - 2 = g - f + 2$$

$$\Rightarrow \quad g = -2 \text{ or } f = 0$$

***Case* I** When $f = 0$

From Eq. (iii), we get

$$\left|\frac{-g - 2}{\sqrt{2}}\right| = \sqrt{g^2 - c}$$

$$\Rightarrow \quad (g + 2)^2 = 2(g^2 - c)$$

$$\Rightarrow \quad g^2 - 4g - 4 - 2c = 0 \quad \text{...(iv)}$$

On putting $f = 0$ in Eq. (ii), we get

$$25 - 8g + c = 0 \quad \text{...(v)}$$

From Eqs. (iv) and (v), we get

$$g^2 - 4g - 4 - 2(8g - 25) = 0$$

$$\Rightarrow \quad g^2 - 4g - 4 - 16g + 50 = 0$$

$$\Rightarrow \quad g^2 - 20g + 46 = 0 \Rightarrow g = 10 \pm 3\sqrt{6}$$

$$\text{and} \quad c = 55 \pm 24\sqrt{6}$$

On substituting the values of g, f and c in Eq. (i), we get

$$x^2 + y^2 + 2(10 \pm 3\sqrt{6})x + (55 \pm 24\sqrt{6}) = 0$$

***Case* II** When $g = -2$

From Eq. (iii), we get

$$\Rightarrow \quad f^2 = 2(4 + f^2 - c) \Rightarrow f^2 - 2c + 8 = 0 \quad \text{...(vi)}$$

On putting $g = -2$ in Eq. (ii), we get

$$c = -6f - 41$$

On substituting the value of c in Eq. (vi), we get

$$f^2 + 12f + 90 = 0$$

This equation gives imaginary values of f. Thus, there is no circle in this case. Hence, the required equation of the circles are $x^2 + y^2 + 2(10 \pm 3\sqrt{6})x + (55 \pm 24\sqrt{6}) = 0$

Example ⑥ Lines $5x+12y-10=0$ and $5x-12y-40=0$ touch a circle C_1 of diameter 6. If the centre of C_1 lie in the first quadrant, then find the equation of the circle C_2 which is concentric with C_1 and cuts intercepts of length 8 on these lines.

Sol. Since, $5x+12y-10=0$ and $5x-12y-40=0$ are both perpendicular tangents to the circle C_1.

$\therefore$ $OABC$ forms a square.

Let the centre coordinates be (h, k), where $OC = OA = 3$ and $OB = 6\sqrt{2}$.

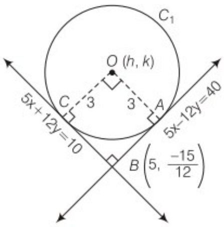

$$\Rightarrow \frac{|5h+12k-10|}{13} = 3$$

$$\text{and } \frac{|5h-12k-40|}{13} = 3$$

$$\Rightarrow \quad 5h+12k-10 = \pm 39$$

$$\text{and } \quad 5h-12k-40 = \pm 39$$

On solving above equations.

The coordinates which lie in I quadrant, are (5, 2).

$\therefore$ Centre for C_1 (5, 2).

To obtain equation of circle concentric with C_1 and making an intercept of length 8 on $5x+12y=10$ and $5x-12y=40$

So, required equation of circle C_2 has centre (5, 2) and radius 5.

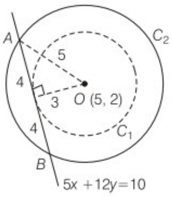

$$\Rightarrow \quad (x-5)^2 + (y-2)^2 = 5^2$$

Pairs of Tangents

From a given point, two tangents can be drawn to a circle which are real and distinct, coincident or imaginary according to as the given point lies outside, on or inside the circle. The combined equation of the pair of tangents drawn from a point $P(x_1, y_1)$ to the circle $x^2+y^2=a^2$ is

$$(x^2+y^2-a^2)\,(x_1^2+y_1^2-a^2) = (xx_1+yy_1-a^2)^2$$

or $$SS' = T^2$$

where, $S = x^2+y^2-a^2$, $S' = x_1^2+y_1^2-a^2$

$$T = xx_1+yy_1-a^2$$

Length of the Tangents

The length of the tangent from the point $P(x_1, y_1)$ to the circle $x^2+y^2+2gx+2fy+c=0$ is equal to

$$\sqrt{s_1} = \sqrt{x_1^2+y_1^2+2gx_1+2fy_1+c}$$

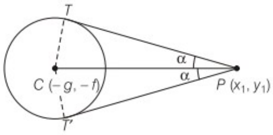

REMEMBER

If PT is the length of the tangent from a point P to a given circle, then PT^2 is called the power of the point with respect to the given circle.

Example ⑦ The angle between a pair of tangents drawn from a point P to the circle $x^2+y^2+4x-6y+9\sin^2\alpha+13\cos^2\alpha=0$ is 2α. The equation of the locus of the point P is **[IIT JEE 1996]**

(a) $x^2+y^2+4x-6y+4=0$

(b) $x^2+y^2+4x-6y-9=0$

(c) $x^2+y^2+4x-6y-4=0$

(d) $x^2+y^2+4x-6y+9=0$

Sol. (d) Centre of the circle

$x^2+y^2+4x-6y+9\sin^2\alpha+13\cos^2\alpha=0$ is $C(-2, 3)$

and radius $\sqrt{(-2)^2+(3)^2-9\sin^2\alpha-13\cos^2\alpha}$

$= \sqrt{13-9\sin^2\alpha-13\cos^2\alpha} = \sqrt{13(1-\cos^2\alpha)-9\sin^2\alpha}$

$= \sqrt{13\sin^2\alpha-9\sin^2\alpha} = \sqrt{4\sin^2\alpha} = 2\sin\alpha$

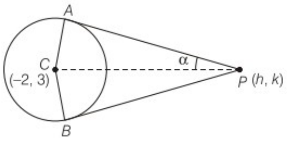

Let (h, k) be any point P and $\angle APC = \alpha$, $\angle PAC = \frac{\pi}{2}$

i.e. $\Delta\ APC$ is a right triangle.

$$\therefore \sin\alpha = \frac{AC}{PC} = \frac{2\sin\alpha}{\sqrt{(h+2)^2+(k-3)^2}}$$

$$\Rightarrow \quad (h+2)^2+(k-3)^2 = 4$$

$$\Rightarrow \quad h^2+4+4h+k^2+9-6k = 4$$

$$\Rightarrow \quad h^2+k^2+4h-6k+9 = 0$$

Thus, required equation of the locus is

$$x^2+y^2+4x-6y+9=0$$

Example ⑧ If the tangent at the point P on the circle $x^2+y^2+6x+6y=2$ meets the straight line $5x-2y+6=0$ at a point Q on the Y-axis, then the length of PQ, is **[JEE Main 2002]**

(a) 4 (b) $2\sqrt{5}$ (c) 5 (d) $3\sqrt{5}$

Sol. (c) The line $5x-2y+6=0$ meets the Y-axis at the point (0, 3) and therefore the tangent has to pass through the point (0, 3) and required length

$$= \sqrt{x_1^2+y_1^2+6x_1+6y_1-2}$$

$$= \sqrt{0^2+3^2+6(0)+6(3)-2}$$

$$= \sqrt{25}$$

$$= 5$$

Example 9 Tangents drawn from the point $P(1, 8)$ to the circle $x^2 + y^2 - 6x - 4y - 11 = 0$ touch the circle at the points A and B. The equation of the circumcircle of the ΔPAB, is **[JEE Main 2009]**

(a) $x^2 + y^2 + 4x - 6y + 19 = 0$
(b) $x^2 + y^2 - 4x - 10y + 19 = 0$
(c) $x^2 + y^2 - 2x + 6y - 29 = 0$
(d) $x^2 + y^2 - 6x - 4y + 19 = 0$

Sol. ***(b)*** For required circle, $P(1, 8)$ and $O(3, 2)$ will be the end points of its diameter.

∴ Equation of required circle is

$$(x-1)(x-3) + (y-8)(y-2) = 0$$

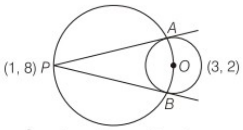

$$\Rightarrow \quad x^2 + y^2 - 4x - 10y + 19 = 0$$

Example 10 Tangents are drawn from the point $(17, 7)$ to the circle $x^2 + y^2 = 169$.

Statement I The tangents are mutually perpendicular.

Statement II The locus of the points from which a mutually perpendicular tangents can be drawn to the given circle is $x^2 + y^2 = 338$. **[IIT JEE 2007]**

(a) Statement I is true, Statement II is also true; Statement II is the correct explanation of Statement I
(b) Statement I is true, Statement II is also true; Statement II is not the correct explanation of Statement I
(c) Statement I is true, Statement II is false
(d) Statement I is false; Statement II is true

Sol. ***(a)*** Since, the tangents are perpendicular. So, locus of perpendicular tangents to circle $x^2 + y^2 = 169$ is a director circle having equation $x^2 + y^2 = 338$.

Example 11 The equations of the tangents drawn from the origin to the circle $x^2 + y^2 + 2rx + 2hy + h^2 = 0$, are **[IIT JEE 1988]**

(a) $x = 0$ (b) $y = 0$
(c) $(h^2 - r^2)x - 2rhy = 0$ (d) $(h^2 - r^2)x + 2rhy = 0$

Sol. ***(a, c)*** Since, tangents are drawn from origin. So,the equation of tangent is given by $y = mx$.

⇒ Length of perpendicular from origin = Radius

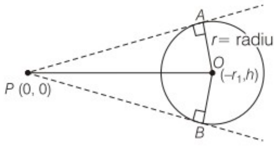

$$\Rightarrow \quad \frac{|mr + h|}{\sqrt{m^2+1}} = r \Rightarrow m^2r^2 + h^2 + 2mrh = r^2(m^2+1)$$

$$\Rightarrow \quad m = \left|\frac{r^2 - h^2}{2rh}\right|, m = \infty$$

∴ Equations of tangents are

$$y = \left|\frac{r^2 - h^2}{2rh}\right| x \text{ and } x = 0$$

∴ $(h^2 - r^2)x - 2rhy = 0$ and $x = 0$

Example 12 Let $x^2 + y^2 - 4x - 2y - 11 = 0$ be a circle. A pair of tangents from the point $(4, 5)$ with a pair of radii form a quadrilateral of area .... **[JEE Main 1985]**

Sol. Centre of circle is (2, 1) and radius

$$= \sqrt{(-2)^2 + (-1)^2 - (-11)} = 4$$

Here, length of tangent AB

$$= \sqrt{4^2 + 5^2 - 4(4) - 2(5) - 11}$$

$$= 2$$

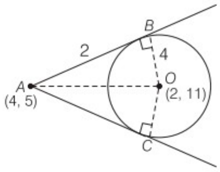

∴ Area of quadrilateral $ABOC$ = 2 × Area of ΔABO

$$= 2 \times \frac{1}{2}(AB)(OB)$$

$$= 2 \times \frac{1}{2} \times 2 \times 4$$

$$= 8 \text{ sq units}$$

Example 13 The area of the triangle formed by the tangents from the point $(4, 3)$ to the circle $x^2 + y^2 = 9$ and the line joining their points of contact is .... **[IIT JEE 1987]**

Sol. Area of triangle formed by the tangents from the point (h, k) to the circle $x^2 + y^2 = a^2$ and their chord of contact

$$= \frac{a(h^2 + k^2 - a^2)^{3/2}}{h^2 + k^2}$$

Thus, area of triangle formed by tangents from (4, 3) to the circle $x^2 + y^2 = 9$ and their chord of contact

$$= \frac{3(4^2 + 3^2 - 9)^{3/2}}{4^2 + 3^2}$$

$$= \frac{3(16 + 9 - 9)^{3/2}}{25} = \frac{3(64)}{25}$$

$$= \frac{192}{25} \text{ sq units}$$

Example 14 C_1 and C_2 are two concentric circles, the radius of C_2 being twice that of C_1. From a point P on C_2, tangents PA and PB are drawn to C_1. Prove that the centroid of ΔPAB lies on C_1. [JEE Main 1998]

Sol. Let the coordinates of point P be $(2r\cos\theta, 2r\sin\theta)$.

We have, $OA = r, OP = 2r$

Since, ΔOAP is a right angled triangle.

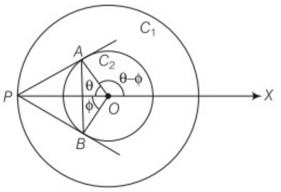

$$\cos\phi = \frac{1}{2} \Rightarrow \phi = \frac{\pi}{3}$$

$\therefore$ Coordinates of A are $\{r\cos(\theta - \pi/3), r\sin(\theta - \pi/3)\}$ and that of B are $[r\cos(\theta + \pi/3), r\sin(\theta + \pi/3)]$.

If p, q is the centroid of ΔPAB, then

$$p = \frac{1}{3}[r\cos(\theta - \pi/3) + r\cos(\theta + \pi/3) + 2r\cos\theta]$$

$$= \frac{1}{3}[r\{\cos(\theta - \pi/3) + \cos(\theta + \pi/3)\} + 2r\cos\theta]$$

$$= \frac{1}{3}\left[r\left\{\begin{matrix} 2\cos\left(\dfrac{\theta - \frac{\pi}{3} + \theta + \frac{\pi}{3}}{2}\right) \\ \cos\left(\dfrac{\theta - \frac{\pi}{3} - \theta - \frac{\pi}{3}}{2}\right)\end{matrix}\right\} + 2r\cos\theta\right]$$

$$= \frac{1}{3}[r\{2\cos\theta\cos\pi/3\} + 2r\cos\theta]$$

$$= \frac{1}{3}[r\cos\theta + 2r\cos\theta] = r\cos\theta$$

and
$$q = \frac{1}{3}\left[r\sin\left(\theta - \frac{\pi}{3}\right) + r\sin\left(\theta + \frac{\pi}{3}\right) + 2r\sin\theta\right]$$

$$= \frac{1}{3}\left[r\left\{\sin\left(\theta - \frac{\pi}{3}\right) + \sin\left(\theta + \frac{\pi}{3}\right)\right\} + 2r\sin\theta\right]$$

$$= \frac{1}{3}\left[r\left\{\begin{matrix} 2\sin\left(\dfrac{\theta - \frac{\pi}{3} + \theta + \frac{\pi}{3}}{2}\right) \\ \cos\left(\dfrac{\theta - \frac{\pi}{3} - \theta - \frac{\pi}{3}}{2}\right)\end{matrix}\right\} + 2r\sin\theta\right]$$

$$= \frac{1}{3}[r(2\sin\theta\cos\pi/3) + 2r\sin\theta]$$

$$= \frac{1}{3}[r\sin\theta + 2r\sin\theta] = r\sin\theta$$

Now, $(p, q) = (r\cos\theta, r\sin\theta)$ lies on $x^2 + y^2 = r^2$ which is called C_1.

Common Tangents to Two Circles

The number of common tangents depends upon the positions of the given circle as discussed below

Let the equations of the two circles be

$$(x - h_1)^2 + (y - k_1)^2 = a_1^2 \quad \ldots(i)$$

and
$$(x - h_2)^2 + (y - k_2)^2 = a_2^2 \quad \ldots(ii)$$

with centres $C_1(h_1, k_1)$, $C_2(h_2, k_2)$ and radii a_1, a_2 respectively. Then, the following cases of intersection of these two circles may arises.

***Case* I** When $C_1C_2 > a_1 + a_2$

In this case, the two circles do not intersect with each other and four common tangents can be drawn to two circles. Two of them are direct and other two are transverse common tangents.

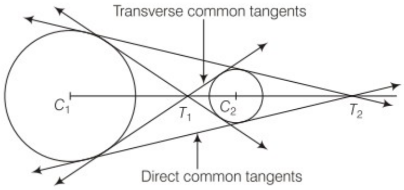

***Case* II** When $C_1C_2 = a_1 + a_2$

In this case, two direct tangents are real and distinct while the transverse tangents are coincident.

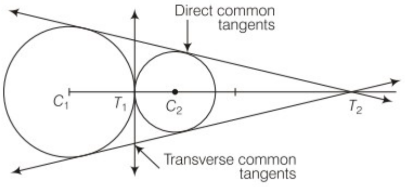

***Case* III** When $C_1C_2 < a_1 + a_2$

In this case, the two direct common tangents are real while the transverse tangents are imaginary.

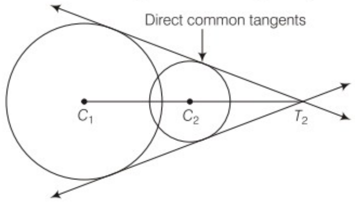

***Case* IV** When $C_1C_2 = a_1 - a_2$

In this case, two tangents are real and coincident while the other two tangents are imaginary.

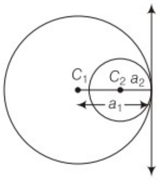

***Case* V** When $C_1C_2 < a_1 - a_2$

In this case, all the four common tangents are imaginary.

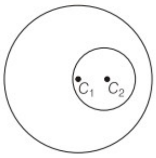

Example 15 The number of common tangents to the circles $x^2 + y^2 = 4$ and $x^2 + y^2 - 6x - 8y = 24$, is/are **[IIT JEE 1998]**

(a) 0 (b) 1 (c) 3 (d) 4

Sol. **(b)** Given, $x^2 + y^2 = 4$

Centre, $C_1 \equiv (0, 0)$ and $R_1 = 2$

Again, $x^2 + y^2 - 6x - 8y - 24 = 0$

Then, $C_2 \equiv (3, 4)$ and $R_2 = 7$

Again, $C_1C_2 = 5 = R_2 - R_1$

Therefore, the given circles touch internally such that they can have just one common tangent at the point of contact.

Example 16 A tangent PT is drawn to the circle $x^2 + y^2 = 4$ at the point $P(\sqrt{3}, 1)$. A straight line L, perpendicular to PT is a tangent to the circle $(x-3)^2 + y^2 = 1$. **[IIT JEE 2012]**

(i) A common tangent of the two circles is

(a) $x = 4$ (b) $y = 2$
(c) $x + \sqrt{3}y = 4$ (d) $x + 2\sqrt{2}y = 6$

(ii) A possible equation of L is

(a) $x - \sqrt{3}y = 1$ (b) $x + \sqrt{3}y = 1$
(c) $x - \sqrt{3}y = -1$ (d) $x + \sqrt{3}y = 5$

Sol. (i) **(d)** Here, equation of common tangent is

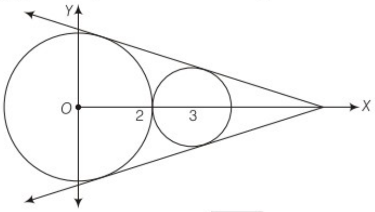

$$y = mx \pm 2\sqrt{1 + m^2}$$

which is also the tangent to

$$(x-3)^2 + y^2 = 1$$

$$\Rightarrow \frac{|3m - 0 + 2\sqrt{1 + m^2}|}{\sqrt{m^2 + 1}} = 1$$

$$\Rightarrow 3m + 2\sqrt{1 + m^2} = \pm \sqrt{1 + m^2}$$

$$\Rightarrow 3m = -3\sqrt{1 + m^2} \text{ or } 3m = -\sqrt{1 + m^2}$$

$$\Rightarrow m^2 = 1 + m^2 \text{ or } 9m^2 = 1 + m^2$$

$$\Rightarrow m \in \phi \text{ or } m = \pm \frac{1}{2\sqrt{2}}$$

$$\therefore y = \pm \frac{1}{2\sqrt{2}} x \pm 2\sqrt{1 + \frac{1}{8}}$$

$$\Rightarrow y = \pm \frac{x}{2\sqrt{2}} \pm \frac{6}{2\sqrt{2}}$$

$$\Rightarrow 2\sqrt{2}y = \pm (x + 6)$$

(ii) **(a)** Here, tangent to $x^2 + y^2 = 4$ at $(\sqrt{3}, 1)$ is $\sqrt{3}x + y = 4$.

As, L is perpendicular to $\sqrt{3}x + y = 4$.

$\therefore$ $x - \sqrt{3}y = \lambda$, which is tangent to $(x-3)^2 + y^2 = 1$

$$\Rightarrow \frac{|3 - 0 - \lambda|}{\sqrt{1 + 3}} = 1 \Rightarrow |3 - \lambda| = 2$$

$$\Rightarrow 3 - \lambda = 2, -2$$

$$\therefore \lambda = 1, 5$$

$$\Rightarrow L : x - \sqrt{3}y = 1, x - \sqrt{3}y = 5$$

Example 17 Match the conditions/expressions in Column I with statement in Column II. **[IIT JEE 2007]**

Column I	Column II
A. Two intersecting circles	p. have a common tangent
B. Two mutually external circles	q. have a common normal
C. Two circles, one strictly inside the other	r. do not have a common tangent
D. Two branches of a hyperbola	s. do not have a common normal

Sol. A $\to$ p, q; B $\to$ p, q; C $\to$ q, r; D $\to$ q, r

A. When two circles are intersecting they have a common normal and common tangent.

B. Two mutually external circles have a common normal and common tangent.

C. When one circle lies inside the other, then they have a common normal but no common tangent.

D. Two branches of a hyperbola have a common normal but no common tangent.

Example 18 Two circles, each of radius 5 units, touch each other at (1, 2). If the equation of their common tangent is $4x + 3y = 10$, then find the equations of the circles. **[IIT JEE 1991]**

Sol. We have, slope of the common tangent $= -\frac{4}{3}$

$\therefore$ Slope of $C_1C_2 = \frac{3}{4}$

If C_1C_2 makes an angle θ with X-axis,

then $\cos\theta = \frac{4}{5}$ and $\sin\theta = \frac{3}{5}$.

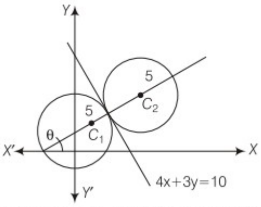

So, the equation of C_1C_2 in parametric form is

$$\frac{x-1}{4/5}=\frac{y-2}{3/5} \quad \text{...(i)}$$

Since, C_1 and C_2 are points on Eq. (i) at a distance of 5 units from P.

So, the coordinates of C_1 and C_2 are given by

$$\frac{x-1}{4/5}=\frac{y-2}{3/5}=\pm 5$$

$$\Rightarrow \quad x = 1 \pm 4 \text{ and } y = 2 \pm 3$$

Thus, the coordinates of C_1 and C_2 are (5, 5) and (–3, – 1) respectively.

Hence, the equations of the two circles are

$$(x-5)^2+(y-5)^2=5^2 \text{ and } (x+3)^2+(y+1)^2=5^2$$

Example 19 Three circles touch one another externally. The tangents at their point of contact meet at a point whose distance from a point of contact is 4. Find the ratio of the product of the radii to the sum of the radii of the circles.
[IIT JEE 1992]

Sol. Suppose the circles have centres at C_1, C_2 and C_3 with radii R_1, R_2 and R_3 respectively. Let the circles touch at A, B and C. Let the common tangents at A, B and C meet at O. We have,

$$OA = OB = OC = 4 \quad \text{[given]}$$

Now, the circle with centre at O and passing through A, B and C is the incircle of $\Delta C_1C_2C_3$ (because $OA \perp C_1C_2$).

Therefore, the inradius of $\Delta C_1C_2C_3$ is 4.

$$\text{and} \quad r=\frac{\Delta}{s} \quad \text{...(i)}$$

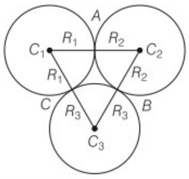

Now, perimeter of a triangle,

$$2s = R_1+R_2+R_2+R_3+R_3+R_1$$

$$\Rightarrow \quad 2s = 2(R_1+R_2+R_3)$$

$$\Rightarrow \quad s = R_1+R_2+R_3$$

$$\text{and} \quad \Delta = \sqrt{s(s-a)(s-b)(s-c)}$$

$$= \sqrt{(R_1+R_2+R_3)(R_3)(R_2)(R_1)}$$

$$\text{From Eq. (i),} \quad 4=\frac{\sqrt{R_1R_2R_3(R_1+R_2+R_3)}}{R_1+R_2+R_3}$$

$$\Rightarrow 16=\frac{R_1R_2R_3(R_1+R_2+R_3)}{(R_1+R_2+R_3)^2} \Rightarrow 16=\frac{R_1R_2R_3}{R_1+R_2+R_3}$$

Example 20 Find the coordinates of the point at which the circles $x^2-y^2-4x-2y+4=0$ and $x^2+y^2-12x-8y+36=0$ touch each other. Also, find equation of common tangents touching the circles at distinct points.
[IIT JEE 1993]

Sol. Two circles touch each other externally, if $C_1C_2=r_1+r_2$ and internally, if $C_1C_2=r_1 \sim r_2$.

Given circles are $x^2-y^2-4x-2y+4=0$, whose centre $C_1(2, 1)$ and radius, $r_1=1$.

$$\text{and} \quad x^2+y^2-12x-8y+36=0$$

whose centre C_2 (6, 4) and radius, $r_2=4$

∴ Distance between the centres

$$=\sqrt{(6-2)^2+(4-1)^2}=\sqrt{16+9}=5$$

$$\Rightarrow \quad C_1C_2=r_1+r_2$$

Therefore, the circles touch each other externally and at the point of touching the point divides the line joining the two centres internally in the ratio of their radii, i.e. 1 : 4.

$$\text{Therefore,} \quad x_1=\frac{1\times 6+4\times 2}{1+4}=\frac{14}{5}$$

$$\text{and} \quad y_1=\frac{1\times 4+4\times 1}{1+4}=\frac{8}{5}$$

Again, to determine the equations of common tangents touching the circles in distinct points, we know that, the tangents pass through a point which divides the line joining the two centres externally in the ratio of their radii, i.e. 1 : 4.

$$\text{Therefore,} \quad x_2=\frac{1\times 6-4\times 2}{1-4}=\frac{-2}{-3}=\frac{2}{3}$$

$$\text{and} \quad y_2=\frac{1\times 4-4\times 1}{1-4}=0$$

Now, let m be the slope of the tangent and this line passing through $\left(\frac{2}{3}, 0\right)$ is

$$y-0=m\left(x-\frac{2}{3}\right) \Rightarrow y-mx+\frac{2}{3}m=0$$

This is tangent to the first circle, if perpendicular distance from centre = radius.

$$\therefore \quad \frac{1-2m+(2/3)m}{\sqrt{1+m^2}}=1 \quad [\because C_1 \equiv (2, 1) \text{ and } r_1=1]$$

$$\Rightarrow 1-2m+\left(\frac{2}{3}\right)m=\sqrt{1+m^2} \Rightarrow 1-\frac{4}{3}m=\sqrt{1+m^2}$$

$$\Rightarrow 1+\frac{16}{9}m^2-\frac{8}{3}m=1+m^2 \Rightarrow \frac{7}{9}m^2-\frac{8}{3}m=0$$

$$\Rightarrow \quad m\left(\frac{7}{9}m-\frac{8}{3}\right)=0$$

$$\Rightarrow \quad m=0, m=\frac{24}{7}$$

Hence, the equations of the two tangents are

$$y=0 \quad \text{and} \quad y=\frac{24}{7}\left(x-\frac{2}{3}\right)$$

$$\Rightarrow \quad y=0 \quad \text{and} \quad 7y-24x+16=0$$

Example 21 The centres of two circles C_1 and C_2 each of unit radius are at a distance of 6 units from each other. Let P be the mid- point of the line segment joining the centres of C_1 and C_2 and C be a circle touching circles C_1 and C_2 externally. If a common tangents to C_1 and C passing through P is also a common tangent to C_2 and C, then the radius of the circle C is ..... . **[IIT JEE 2009]**

Sol. $\because (r+1)^2 = \alpha^2 + 9$

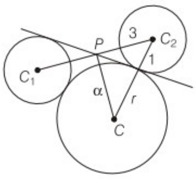

$$r^2 + 8 = \alpha^2$$

$$\Rightarrow \quad r^2 + 2r + 1 = r^2 + 8 + 9$$

$$\Rightarrow \quad 2r = 16$$

$$\Rightarrow \quad r = 8 \text{ units}$$

Example 22 Let T_1 and T_2 be two tangents drawn from $(-2, 0)$ on the circle $C: x^2 + y^2 = 1$. Determine the circles touching C and having T_1, T_2 as their pair of tangents. Further, find the equations of all possible common tangents to these circles when taken two at a time. **[IIT JEE 1999]**

Sol.

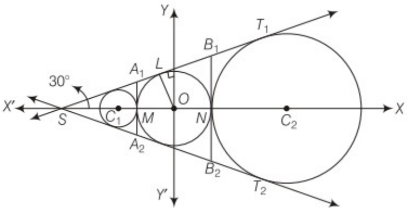

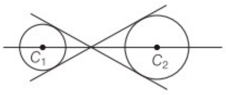

From figure, it is clear that, ΔOLS is a right angled triangle with right angle at L.

Also, $\quad OL = 1$ and $OS = 2$

$$\therefore \quad \sin(\angle LSO) = \frac{1}{2}$$

$$\Rightarrow \quad \angle LSO = 30^\circ$$

Since, $SA_1 = SA_2$, so ΔSA_1A_2 is an equilateral triangle.

The circle with centre at C_1 is a circle inscribed in ΔSA_1A_2.

Therefore, centre C_1 is centroid of ΔSA_1A_2. Thus, C_1 divides SM in the ratio 2 : 1. Therefore, coordinates of C_1 are $(-4/3, 0)$ and its radius $= C_1M = \frac{1}{3}$.

$$\therefore \text{ Its equation is } \left(x + \frac{4}{3}\right)^2 + y^2 = \left(\frac{1}{3}\right)^2 \qquad \text{...(i)}$$

The other circle touches the equilateral ΔSB_1B_2 externally. Its radius r is given by $r = \dfrac{\Delta}{s-a}$

where, $\quad B_1B_2 = a$

But $\quad \Delta = \frac{1}{2}(a)(SN) = \frac{3}{2}a$

and $\quad s = a = \frac{3}{2}a - a = \frac{a}{2}$

Thus, $\quad r = 3$

$\Rightarrow$ Coordinates of C_2 are (4, 0).

$\therefore$ Equation of circle with centre at C_2 is

$$(x-4)^2 + y^2 = 3^2 \qquad \text{...(ii)}$$

Equations of common tangents to circle (i) and circle C are

$$x = -1$$

and $\quad y = \pm \frac{1}{\sqrt{3}}(x+2) \qquad [T_1 \text{ and } T_2]$

Equations of common tangents to circle (ii) and circle C are

$$x = 1$$

and $\quad y = \pm \frac{1}{\sqrt{3}}(x+2) \qquad [T_1 \text{ and } T_2]$

Two tangents common to circles (i) and (ii) are T_1 and T_2 at O. To find the remaining two transverse tangents to circles (i) and (ii), we find a point I which divides the joint of C_1C_2 in the ratio

$$r_1 : r_2 = \frac{1}{3} : 3 = 1 : 9$$

$\therefore$ Coordinates of I are $\left(-\frac{4}{5}, 0\right)$.

Equation of any line through I is $y = m\left(x + \frac{4}{5}\right)$. It will touch circle (i), if

$$\frac{\left|m\left(-\frac{4}{3} + \frac{4}{5}\right) - 0\right|}{\sqrt{1+m^2}} = \frac{1}{3}$$

$$\Rightarrow \quad \left|-\frac{8m}{15}\right| = \frac{1}{3}\sqrt{1+m^2}$$

$$\Rightarrow \quad 64m^2 = 25(1+m^2)$$

$$\Rightarrow \quad 39\,m^2 = 25$$

$$\Rightarrow \quad m = \pm \frac{5}{\sqrt{39}}$$

Hence, these tangents are

$$y = \pm \frac{5}{\sqrt{39}}\left(x + \frac{4}{5}\right).$$

Smart Practice

Single Option Correct Type

1. The line $2x - y + 1 = 0$ is tangent to the circle at the point $(2, 5)$ and the centre of the circle lies on $x - 2y = 4$. The radius of the circle is

(a) $3\sqrt{5}$ (b) $5\sqrt{3}$ (c) $2\sqrt{5}$ (d) $5\sqrt{2}$

2. If O is the origin and OP, OQ are the tangents from the origin to the circle $x^2 + y^2 - 6x + 4y + 8 = 0$, then the circumcentre of ΔOPQ is

(a) $(3, -2)$ (b) $\left(\frac{3}{2}, -1\right)$ (c) $\left(\frac{3}{4}, -\frac{1}{2}\right)$ (d) $\left(-\frac{3}{2}, 1\right)$

3. If (α, β) is a point on the circle whose centre is on the X-axis and which touches the line $x + y = 0$ at $(2, -2)$, then the greatest value of α is

(a) $4 - \sqrt{2}$ (b) 6 (c) $4 + 2\sqrt{2}$ (d) $4 + 2$

4. There are two circles whose equations are $x^2 + y^2 = 9$ and $x^2 + y^2 - 8x - 6y + n^2 = 0$, $n \in Z$. If the two circles have exactly two common tangents, then the number of possible values of n is

(a) 2 (b) 8
(c) 9 (d) None of these

5. The equation of the tangent to the circle $x^2 + y^2 + a^2$, which makes a triangle of area a^2 with the coordinate axes, is

(a) $x \pm y = a\sqrt{2}$ (b) $x \pm y = \pm a\sqrt{2}$
(c) $x \pm y = 2a$ (d) $x \pm y = \pm 2a$

One or More than One Option Correct Type

6. The equation of the tangent to the circle $x^2 + y^2 = 25$ passing through $(-2, 11)$, is/are

(a) $4x + 3y = 25$
(b) $3x + 4y = 38$
(c) $24x - 7y + 125 = 0$
(d) $7x + 24y = 250$

7. The circles $x^2 + y^2 - 2x - 4y + 1 = 0$ and $x^2 + y^2 + 4x + 4y - 1 = 0$

(a) touch internally
(b) touch externally
(c) have $3x + 4y - 1 = 0$ as the common tangent at the point of contact
(d) have $3x + 4y + 1 = 0$ as the common tangent at the point of contact

8. The range of values of a such that angle θ between the pair of tangents drawn from $(a, 0)$ to the circle $x^2 + y^2 + 1$ satisfies $\frac{\pi}{2} < \theta < \pi$, is

(a) $(1, 2)$ (b) $(1, \sqrt{2})$
(c) $(-\sqrt{2}, -1)$ (d) $(-2, -1)$

Linked Comprehension Type

Passage (Q. Nos. 9-11) A circle C whose radius is 1 unit, touches the X-axis at point A. The centre Q of C lies in the first quadrant. The tangent from the origin O to the circle touches it at T and a point P lies on it such that ΔOAP is a right angled triangle at A and its perimeter is 8 units.

9. The length of PQ is

(a) $\frac{1}{2}$ (b) $\frac{4}{3}$
(c) $\frac{5}{3}$ (d) $\frac{7}{3}$

10. The equation of circle C is

(a) $\{x - (2 + \sqrt{3})\}^2 + (y - 1)^2 = 1$
(b) $\{x - (\sqrt{3} - \sqrt{2})\}^2 + (y - 1)^2 = 1$
(c) $(x - \sqrt{3})^2 + (y - 1)^2 = 1$
(d) None of the above

11. The equation of tangent OT is

(a) $x - \sqrt{3}\,y = 0$ (b) $x - \sqrt{2}\,y = 0$
(c) $y - \sqrt{3}\,x = 0$ (d) $y - \sqrt{2}\,x = 0$

Single Integer Type

12. The sum of the slopes of the lines tangent to both the circles $x^2 + y^2 = 1$ and $(x - 6)^2 + y^2 = 4$, is …… .

13. The length of a common internal tangent to two circles is 7 and that of a common external tangent is 11. If the product of the radii of the two circles is P, then the value of $\frac{P}{2}$ is ……… .

14. If two perpendicular tangents can be drawn from the origin to the circle $x^2 - 6x + y^2 - 2py + 17 = 0$, then the value of $|p|$ is ……… .

Match the Column

15. Match the Column I items with Column II items.

	Column I		Column II
A.	The number of circles touching the given three non-concurrent lines is	p.	1
B.	The number of circles touching $y = x$ at (2, 2) and also touching the line $x + 2y - 4 = 0$, is	q.	2
C.	The number of circles touching the lines $x \pm y = 2$ and passing through the point (4, 3), is	r.	4
D.	The number of circles intersecting the given three circles orthogonally, is	s.	infinite

Answers with **Explanation**

1. ***(a)*** We have, $2x - y + 1 = 0$ is a tangent. Slope of line $OA = -\frac{1}{2}$. The equation of line OA is $y - 5 = -\frac{1}{2}(x - 2)$ or $x + 2y = 12$

Therefore, intersection with $x - 2y = 4$ will give the coordinates of centre as (8, 2).

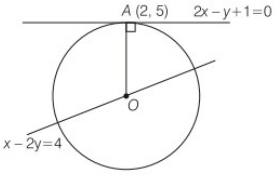

Hence,

$$r = OA = \sqrt{(8-2)^2 + (2-5)^2}$$
$$= 3\sqrt{5}$$

2. ***(b)*** Clearly, $OPCQ$ is a cyclic quadrilateral. Then, the circumcircle of ΔOPQ passes through the point C.

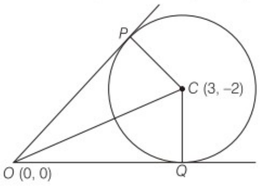

For this, OC is diameter.

Then, the circumcentre is the mid-point of OC,

i.e. $$\left(\frac{0+3}{2}, \frac{0-2}{2}\right) = \left(\frac{3}{2}, -1\right)$$

3. ***(c)*** If $(a, 0)$ is the centre C and P is $(2, -2)$, then $\angle COP = 45°$. Since, the equation of OP is $x + y = 0$, so we have $OP = 2\sqrt{2} = CP$.

Hence, $OC = \sqrt{OP^2 + CP^2} = \sqrt{(2\sqrt{2})^2 + (2\sqrt{2})^2} = 4$

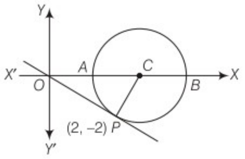

The point on the circle with the greatest x-coordinate is B.

$$\therefore \quad \alpha = OB = OC + CB = OC + PC = 4 + 2\sqrt{2}$$

4. ***(c)*** For $x^2 + y^2 = 9$, the centre is (0, 0) and the radius is 3. For $x^2 + y^2 - 8x - 6y + n^2 = 0$, the centre is (4, 3) and the radius is $\sqrt{4^2 + 3^2 - n^2}$.

Now, $4^2 + 3^2 - n^2 > 0$ or $n^2 < 5^2$ or $-5 < n < 5$. The circles should cut externally to have exactly two common tangents. So, $r_1 + r_2 > d$ (distance between centres)

$$\therefore \quad 3 + \sqrt{25 - n^2} > \sqrt{4^2 + 3^2}$$
$$\Rightarrow \quad \sqrt{25 - n^2} > 2 \Rightarrow 25 - n^2 > 4$$
$$\therefore \quad n^2 < 21 \Rightarrow -\sqrt{21} < n < \sqrt{21}$$

Therefore, the common values of n should satisfy

$$-\sqrt{21} < n < \sqrt{21}$$

But $n \in Z$. So, $n = -4, -3, \ldots, 4$

$\therefore$ The number of possible values of n is 9.

5. ***(b)*** Let the tangent be of the form $\frac{x}{x_1} + \frac{y}{y_1} = 1$ and the area of triangle formed by it with the coordinate axes be

$$\frac{1}{2}x_1y_1 = a^2 \qquad \ldots(\text{i})$$

Again, $y_1x + x_1y - x_1y_1 = 0$

On applying the conditions of tangency, we get

$$\frac{|-x_1y_1|}{\sqrt{x_1^2 + y_1^2}} = a \text{ or } x_1^2 + y_1^2 = \frac{x_1^2y_1^2}{a^2} \qquad \ldots(\text{ii})$$

From Eqs. (i) and (ii), we get the values of x_1 and y_1, which gives the equation of tangent as $x \pm y = \pm a\sqrt{2}$.

6. ***(a, c)*** The equation of any tangent to the circle $x^2 + y^2 = 25$ is of the form

$$y = mx + 5\sqrt{1 + m^2}, \text{ where } m \text{ is the slope.}$$

It passes through (−2, 11).

$$\therefore \quad 11 = -2m + 5\sqrt{1 + m^2}$$
$$\Rightarrow \quad (11 + 2m)^2 = 25(1 + m^2)$$
$$\Rightarrow \quad m = \frac{24}{7}, \frac{-4}{3}$$

Therefore, the equation of tangents are

$$24x - 7y + 125 = 0$$

and $$4x + 3y = 25$$

7. ***(b, c)*** For the given circles

$$S_1 : x^2 + y^2 - 2x - 4y + 1 = 0$$

and $$S_2 : x^2 + y^2 + 4x + 4y - 1 = 0,$$

$$C_1 \equiv (1, 2), r_1 = 2; C_2 \equiv (-2, -2), r_2 = 3$$

Now, $r_1 + r_2 = 5$ and $C_1C_2 = 5$

Hence, the circles touch externally. Also, the common tangent at the point of contact is $S_1 - S_2 = 0$ or $3x + 4y - 1 = 0$.

8. ***(b, c)*** The equation of the pair of tangents given by $SS' = T^2$ is

$$(ax + 0 - 1)^2 = (x^2 + y^2 - 1)(a^2 + 0 - 1)$$
$$\Rightarrow \quad (a^2 - 1)y^2 - x^2 + 2ax - a^2 = 0$$

If θ is the angle between the tangents, then

$$\tan\theta = \frac{2\sqrt{H^2 - AB}}{A + B} = \frac{2\sqrt{-(a^2 - 1)(-1)}}{a^2 - 2}$$
$$= \frac{2\sqrt{a^2 - 1}}{a^2 - 2}$$

If θ lies in the second quadrant, then $\tan\theta < 0$.

METICULOUS ANALYSIS

$$\therefore \quad \frac{2\sqrt{a^2-1}}{a^2-2} < 0$$

$$\Rightarrow \quad a^2 - 1 > 0 \text{ and } a^2 - 2 < 0$$

$$\Rightarrow \quad |a| > 1 \text{ and } |a| < \sqrt{2}$$

$\therefore\ a \in (-\sqrt{2}, -1) \cup (1, \sqrt{2})$

9. ***(c)*** Given, $QT = QA = 1$

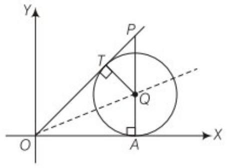

Let $\quad PQ = x$. Then, $PT = \sqrt{x^2-1}$

ΔTQP and ΔAPO are similar triangles.

Then, $\quad OT = OA = \dfrac{x+1}{\sqrt{x^2-1}}$

$$\Rightarrow \quad 1 + x + \frac{2(x+1)}{\sqrt{x^2-1}} + \sqrt{x^2-1} = 8$$

$$\therefore \quad x = \frac{5}{3}$$

10. ***(a)*** Here, $AP = \dfrac{8}{3}$ and $OP = \dfrac{16}{3}$

Let $\quad \angle AOP = 2\theta,$

then $\quad \sin 2\theta = \dfrac{1}{2}$

In ΔOAQ,

$$\tan\theta = \frac{1}{OA} \Rightarrow OA = \frac{1}{\tan\theta}$$

Also, $\quad \sin 2\theta = \dfrac{2\tan\theta}{1+\tan^2\theta} = \dfrac{1}{2}$

$$\therefore \quad \tan\theta = 2 - \sqrt{3}$$

Thus, $\quad OA = 2 + \sqrt{3}$

Hence, the equation of circle is

$$\{x - (2+\sqrt{3})\}^2 + (y-1)^2 = 1$$

11. ***(a)*** The equation of tangent OT is

$$\frac{x-0}{\cos 2\theta} = \frac{y-0}{\sin 2\theta}$$

$$\therefore \quad x - \sqrt{3}y = 0$$

12. ***(0)*** Since, both circles are symmetric about X-axis, so the sum of slopes of tangent is 0.

13. ***(9)***

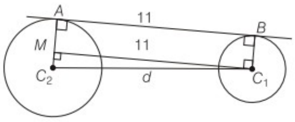

From the figure,

$$d^2 = 11^2 + (r_1 - r_2)^2 \qquad \text{...(i)}$$

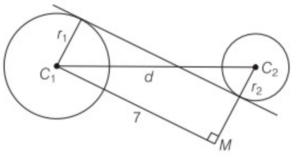

From the figure,

$$d^2 = 7^2 + (r_1 + r_2)^2 \qquad \text{...(ii)}$$

From Eqs. (i) and (ii), we get

$$4r_1r_2 = 72$$

$$\Rightarrow \quad r_1r_2 = 18 = P \qquad \text{[given]}$$

$$\therefore \quad \frac{P}{2} = \frac{18}{2} = 9$$

14. ***(5)*** The equation of the given circle is

$$x^2 + y^2 - 6x - 2py + 17 = 0$$

$$\Rightarrow \quad (x-3)^2 + (y-p)^2 = (p^2 - 8) \qquad \text{...(i)}$$

Also, (0, 0) lies outside the circle.

The equation of director circle of $S = 0$ will be

$$(x-3)^2 + (y-p)^2 = 2(p^2-8) \qquad \text{...(ii)}$$

Tangents drawn from (0, 0) to circle (i) are perpendicular to each other. Therefore, (0, 0) must lie on director circle.

$$\therefore \quad (0-3)^2 + (0-p)^2 = 2(p^2-8)$$

$$\Rightarrow \quad p^2 = 25$$

$$\Rightarrow \quad p = \pm 25$$

$$\Rightarrow \quad |p| = 5$$

15. A → r; B → q; C → q; D → p

A.

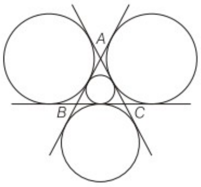

B.

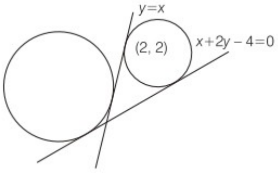

C.

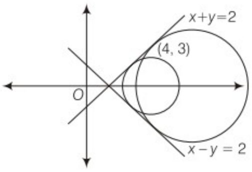

D. Only 1

29 04 2020
S T [✓ Q] Q S S D

G) 157) 2020 0429 2246

A) Books_74_2020/
Magazine/
Spectrum_Mathematics_-_November_
2016.pdf

→ family of circles. - Pág. 03

properties of triangles. - Pág. 66.

Pág. 13 "0" is either positive or negative.

Pág. 42 Cube Roots and nth Roots of Unity

→ nth roots of unity is very important topic of complex numbers.

B) Books_74_2020/
Magazine/
Spectrum_Mathematics_-_July_2016.pdf

→ Tangent to a circle. Pág. 42-49.

Look pages 40 e 41.

Trigonometric Equations and Identities. (resume)

Pág. 54 - Inverse Trigonometric Functions

& % 666
)*+,)))-) % 7 / 8
* " 9 : . 2 ."
; 1 " <<<

2)*+,)))-) % =# # 3 % ' ."
%# . & . .

>--? / / / '@)>--? / /)> A /' /)& .)

")*+,)))-) % + "@ 1 : . 9 5
* 9 " B : . &
2 ."
(3' %% . #

Chapter 3
Early Works in the Field of Contact Geometry of Surfaces

The investigation of the contact geometry of curves and surfaces can be traced back to the eighteenth century. The study of the contact of curves and surfaces was undertaken in considerable detail by *J.L. Lagrange*[1] in his "*Theorié des Fonctions Analytiques*" (1797) [1], and *A.L. Cauchy*[2] in his "*Leçons sur les Applications du Calcul Infinitésimal á la Geometrie*" (1826) [2]. Later, in the twentieth century, an investigation in the realm of contact geometry of curves and surfaces was undertaken by *J. Favard*[3] in his "*Course de Gèomètrie Diffèrentialle Locale*" (1957) [3]. A few more names of the researchers to be mentioned.

Contacts of curves and surfaces of two different kinds are distinguished. Speaking more generally, the contact of surfaces with "*different order of tangency*" is investigated in differential geometry of surfaces. As a particular case, the conditions of contact of two smooth regular part surfaces in the "*first order of tangency*" are investigated in engineering geometry of surfaces.

Let us begin the discussion from the analysis of the results of the research obtained so far in differential geometry of surfaces.

3.1 Order of Contact

Order of contact of two smooth regular surfaces can be specified in terms of the order of contact of two smooth regular planar curves. It is understood there that:

- the curves are within a common plane;
- the curves are passing through the point of contact of the surfaces; and

---

[1] *Joseph-Louis* ***Lagrange*** (January 25, 1736–April 10, 1813)—an Italian born [born *Giuseppe Lodovico (Luigi)* ***Lagrangia***] famous French mathematician and mechanician.

[2] *Augustin-Louis* ***Cauchy*** (August 21, 1789–May 23, 1857)—a famous French mathematician.

[3] *Jean* ***Favard*** (August 28, 1902–January 21, 1965)—a French mathematician.

© Springer Nature Switzerland AG 2020

S. P. Radzevich, *Geometry of Surfaces*,
https://doi.org/10.1007/978-3-030-22184-3_3

- the plane is through the direction of the common perpendicular to the contacting surfaces at a point of interest.

Under such an investigation scenario, the contact geometry of two surfaces can be reduced to investigation of the geometry of contact of the two curves when the section plane spins about the straight line along the common perpendicular to the contacting surfaces. One application of such an approach is in fitting a local model to a shape if we want the contact to be as tight as possible. This requirement can be formulated in other words, that is, if we want the highest possible degree of conformity of the contacting surfaces in a particular plane through the common perpendicular. Another application is in the interaction of probes with a shape. For instance, a light ray may graze a surface in many different ways, the degree of contact varying enormously from case to case.

Planar curves have no contact if they have no point in common. Contact of order zero is observed when two smooth regular curves have a common intersection. Then, the order-one contact of the curves occurs when—at the point in common—the tangents to the curves align with one another. Ultimately, it is said that two curves have contact of the order n if all their derivatives with respect to arc length up to the nth-order are aligned, whereas the $(n+1)$th-order is different. The order thus amounts to the order of the common initial part of a *Taylor*[4] series.[5]

Similarly, it is said for the case of two surfaces that the order of contact is n if all mixed partial derivatives up to the nth-order for some mutual parameterization $[\mathbf{r}_{P1}(U_P,\ V_P)$, and $\mathbf{r}_{P2}(U_P,\ V_P)]$ align, whereas this is not the case for the $(n+1)$th-order.

Thus, first-order contact means:

$$\frac{\partial \mathbf{r}_{P1}}{\partial U_P}(U_P,\ V_P) = \frac{\partial \mathbf{r}_{P2}}{\partial U_P}(U_P,\ V_P) \tag{3.1}$$

$$\frac{\partial \mathbf{r}_{P1}}{\partial V_P}(U_P,\ V_P) = \frac{\partial \mathbf{r}_{P2}}{\partial V_P}(U_P,\ V_P) \tag{3.2}$$

and at least one of the equalities:

$$\frac{\partial^2 \mathbf{r}_{P1}}{\partial U_P^2}(U_P,\ V_P) = \frac{\partial^2 \mathbf{r}_{P2}}{\partial U_P^2}(U_P,\ V_P) \tag{3.3}$$

$$\frac{\partial^2 \mathbf{r}_{P1}}{\partial U_P^2}(U_P,\ V_P) = \frac{\partial^2 \mathbf{r}_{P2}}{\partial U_P^2}(U_P,\ V_P) \tag{3.4}$$

$$\frac{\partial^2 \mathbf{r}_{P1}}{\partial U_P \partial V_P}(U_P,\ V_P) = \frac{\partial^2 \mathbf{r}_{P2}}{\partial U_P \partial V_P}(U_P,\ V_P) \tag{3.5}$$

---

[4] *Brook* ***Taylor*** (August 18, 1685–December 29, 1731)—a well known English mathematician.

[5] It can be found out in the book by *Favard* [3] that two curves having infinite degree of contact cannot be congruent to one another.

is not true; that is to say, the tangent planes coincide but the osculating quadrics differ.

In Eq. (3.1) through Eq. (3.5), the position vector of a point of the first of the contacting surfaces P_1 is designated $\mathbf{r}_{P1}$, and the position vector of a point of the second of the contacting surface P_2 is denoted $\mathbf{r}_{P2}$. Both the contacting surfaces P_1 and P_2 must be parameterized so as to have common "*Gaussian coordinates*" U_P and V_P. In a case the part surfaces P_1 and P_2 are expressed in different U_P- and V_P-coordinates, then it is required to re-parameterize them to common U_P- and V_P-parameters (see Appendix C).

A "*curve*" has contact of at least order n with a "*surface*" if there exists a curve in the surface, which has nth-order contact with the curve. One defines the contact of space curves in the same way as the contact of planar curves mentioned earlier.

3.2 Contact Geometry of Part Surfaces

When solving problems in engineering geometry of surfaces, an appropriate method for analytical description of the contact geometry of part surfaces—say of a part surface P_1 and a part surface P_2—is required. The problem of analytical description of the contact geometry of two smooth regular part surfaces in the first order of tangency is a sophisticated one. It can be solved on the premise of analysis of geometry of the contacting surfaces in the differential vicinity of the point of their contact.

Various methods for analytical description of the contact geometry of smooth regular surfaces have been developed. An overview of the known methods can be found in the monograph by Prof. *Radzevich* [4]. The latest achievements in the field are discussed in the papers [5, 6] and in the monograph [7]. In these publications as well as in other works, two part surfaces P_1 and P_2 that make contact at a point K are considered. Locally, each of the contacting surfaces is represented by a corresponding quadric. Owing to the fact that only differential vicinity of the contacting surfaces is considered below, that same designation P_1 and P_2 is applied to the quadrics as to the surfaces P_1 and P_2 themselves.

Generally speaking, at the point of contact K, the contacting part surfaces P_1 and P_2 may feature an arbitrary configuration. Namely, the local part surface patches P_1 and P_2 may be turned around a common perpendicular through a certain angle. Thus, it makes sense to begin the discussion from the analytical description of local relative orientation of the part surfaces P_1 and P_2 (in the differential vicinity of the point of contact of the surfaces).

3.3 Local Relative Orientation of Contacting Part Surfaces

Consider two local part surface patches P_1 and P_2, which make contact at a point K as illustrated in Fig. 3.1. At the contact point K, the surfaces P_1 and P_2 share a

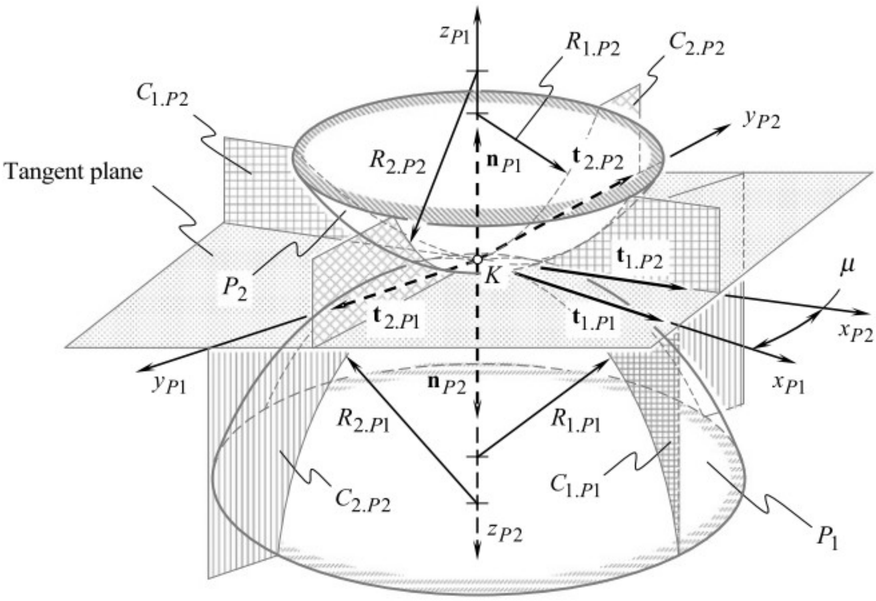

Fig. 3.1 Local configuration of two smooth regular part surface patches P_1 and P_2 at a point K of their contact

common tangent plane. Tangency of the surfaces imposes a kind of restrictions on the relative configuration (location and orientation) of these surfaces, and on their relative motion, if any. In engineering geometry of surface generation, a quantitative measure of the surfaces P_1 and P_2 relative orientation is established.

Relative orientation of the contacting part surfaces P_1 and P_2 is specified by the angle μ of the surfaces "*local relative orientation*".[6] By definition [5, 8, and 9], the angle μ is equal to the angle that the unit tangent vector $\mathbf{t}_{1.P1}$ of the first principal direction of the part surface P_1 forms with the unit tangent vector $\mathbf{t}_{1.P2}$ of the first principal direction of the part surface P_2. That same angle μ can also be specified as the angle made by the unit tangent vectors $\mathbf{t}_{2.P1}$ and $\mathbf{t}_{2.P2}$ of the second principal directions of the part surfaces P_1 and P_2 at a point K of their contact. This immediately yields the equations for the calculation of the angle μ of the surfaces local relative orientation:

$$\sin\mu = |\mathbf{t}_{1.P1} \times \mathbf{t}_{1.P2}| = |\mathbf{t}_{2.P1} \times \mathbf{t}_{2.P2}| \tag{3.6}$$

$$\cos\mu = \mathbf{t}_{1.P1} \cdot \mathbf{t}_{1.P2} = \mathbf{t}_{2.P1} \cdot \mathbf{t}_{2.P2} \tag{3.7}$$

---

[6] Orientation of the surfaces P_1 and P_2 is "*local*" in nature as it is related only to the differential vicinity of point K of contact of the surfaces.

$$\tan \mu = \frac{|\mathbf{t}_{1.P1} \times \mathbf{t}_{1.P2}|}{\mathbf{t}_{1.P1} \cdot \mathbf{t}_{1.P2}} \equiv \frac{|\mathbf{t}_{2.P1} \times \mathbf{t}_{2.P2}|}{\mathbf{t}_{2.P1} \cdot \mathbf{t}_{2.P2}} \tag{3.8}$$

Here, we define:

$\mathbf{t}_{1.P1}$, $\mathbf{t}_{2.P1}$	Are the unit vectors of principal directions on the first part surface P_1 of two contacting surfaces
$\mathbf{t}_{1.P2}$, $\mathbf{t}_{2.P2}$	Are the unit vectors of principal directions on the second part surface P_2 of two contacting surfaces.

In case of point contact of the part surfaces P_1 and P_2, the actual value of the angle μ is calculated at the point K of contact of the surfaces. If the part surfaces P_1 and P_2 are in line contact, then the actual value of the angle μ of the surfaces' local relative orientation can be calculated at every point within the line of contact.[7] The reason for why the angle μ is referred to as the angle of "*local*" surfaces' orientation is clearly illustrated with this.

The determination of the angle μ of the surfaces' P_1 and P_2 local relative orientation is illustrated in Fig. 3.2.

Unit tangent vectors of the principal directions $\mathbf{t}_{1.P1}$ and $\mathbf{t}_{1.P2}$ at a point of contact of the part surfaces P_1 and P_2 are employed for the purpose of calculation of the angle μ of the surfaces' local relative orientation.

Consider two part surfaces P_1 and P_2 in point contact that are represented in a common reference system. The surfaces make contact at point K. For further analysis, an equation of the common tangent plane to the surfaces P_1 and P_2 at the point K is necessary (Figs. 3.1 and 3.3):

$$(\mathbf{r}_{\text{tp}} - \mathbf{r}_K) \cdot \mathbf{u}_{P1} \cdot \mathbf{v}_{P1} = 0 \tag{3.9}$$

Here, we define:

$\mathbf{r}_{\text{tp}}$	Is the position vector of a point of the common tangent plane
$\mathbf{r}_K$	Is the position vector of the contact point K
$\mathbf{u}_{P1}$ and $\mathbf{v}_{P1}$	are the unit vectors tangent to U_{P1}- and V_{P1}-coordinate lines on the surface P_1 at the point K.

An equation similar to Eq. (3.9) can also be expressed in terms of parameters of geometry of the other contacting surface P_2.

The actual value of the angle ω_{P1} between the vectors $\mathbf{v}_{P1}$ and $\mathbf{v}_{P1}$ can be calculated from one of the following equations:

$$\sin \omega_{P1} = \frac{\sqrt{E_{P1} G_{P1} - F_{P1}^2}}{\sqrt{E_{P1} G_{P1}}} \tag{3.10}$$

---

[7] It is worth pointing out here that for line contact, the relative orientation of the surfaces P_1 and P_2 is predetermined in global sense. However, the actual value of the angle μ of the surfaces' P_1 and P_2 "*local*" relative orientation at different points of the line of contact is different.

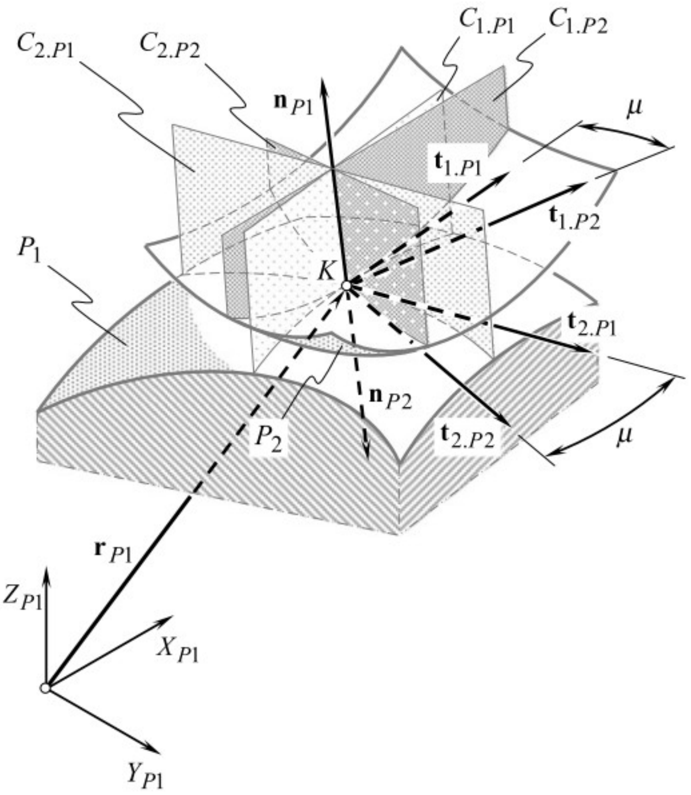
$C_{1.P1}$
$C_{1.P2}$
$C_{2.P1}$
$C_{2.P2}$
$\mathbf{n}_{P1}$
μ
$\mathbf{t}_{1.P1}$
$\mathbf{t}_{1.P2}$
P_1
K
$\mathbf{t}_{2.P1}$
$\mathbf{n}_{P2}$
P_2
$\mathbf{t}_{2.P2}$
μ
$\mathbf{r}_{P1}$
Z_{P1}
X_{P1}
Y_{P1}

Fig. 3.2 Specification of the angle μ of the surfaces P_1 and P_2 local relative orientation

$$\cos \omega_{P1} = \frac{F_{P1}}{\sqrt{E_{P1} G_{P1}}} \tag{3.11}$$

$$\tan \omega_{P1} = \frac{\sqrt{E_{P1} G_{P1} - F_{P1}^2}}{F_{P1}} \tag{3.12}$$

Equations similar to Eq. (3.9) through Eq. (3.12) are also valid for the calculation of the angle ω_{P2} of the other contacting surface P_2.

The unit tangent vectors $\mathbf{u}_{P1}$ and $\mathbf{v}_{P1}$, which are tangent to the U_{P1}- and V_{P1}-coordinate lines on the part surface P_1, and the unit tangent vectors $\mathbf{u}_{P2}$ and $\mathbf{v}_{P2}$, which are tangent to the U_{P2}- and V_{P2}-coordinate lines on the surface P_2, are specified by the angles θ and ε. The actual values of the angles θ and ε can be calculated from the following equations:

$$\cos \theta = \mathbf{u}_{P1} \cdot \mathbf{u}_{P2} \tag{3.13}$$

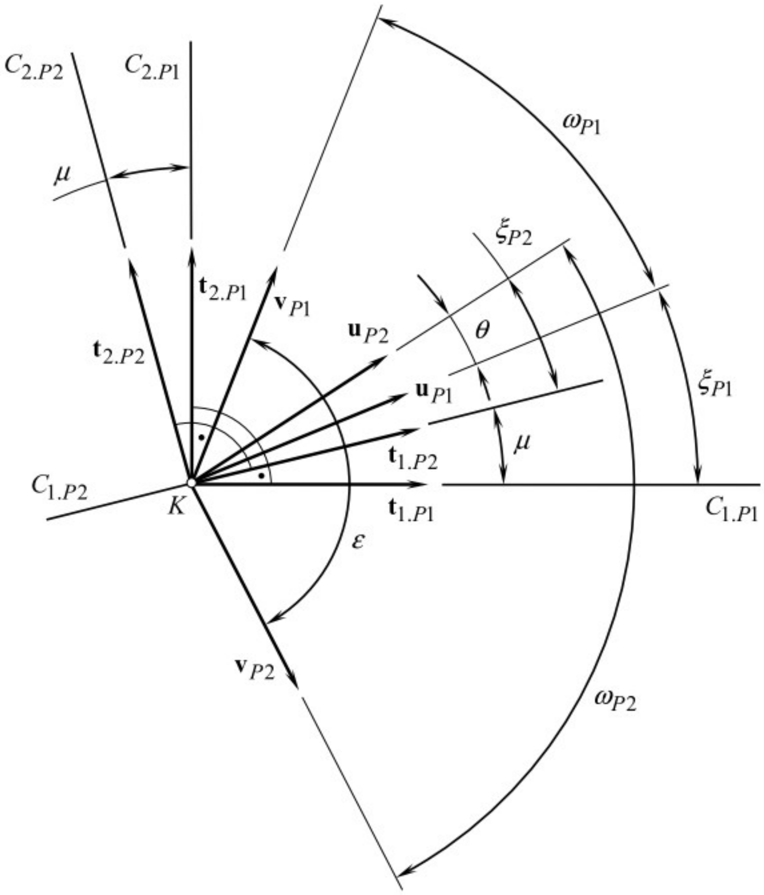
$C_{2.P2}$
$C_{2.P1}$
μ
ω_{P1}
ξ_{P2}
$\mathbf{t}_{2.P1}$
$\mathbf{v}_{P1}$
$\mathbf{t}_{2.P2}$
$\mathbf{u}_{P2}$
θ
$\mathbf{u}_{P1}$
ξ_{P1}
μ
$\mathbf{t}_{1.P2}$
$C_{1.P2}$
K
$\mathbf{t}_{1.P1}$
$C_{1.P1}$
ε
$\mathbf{v}_{P2}$
ω_{P2}

Fig. 3.3 *Local relative orientation* of the contacting part surfaces P_1 and P_2, considered in a common tangent plane

$$\cos\varepsilon = \mathbf{v}_{P1} \cdot \mathbf{v}_{P2} \tag{3.14}$$

Angle ξ_P is the angle that the first principal direction $\mathbf{t}_{1.P}$ on the surface P forms with the unit tangent vector $\mathbf{u}_P$ (see Fig. 3.3). The equation for the calculation of the actual value of the angle ξ_P is derived by Prof. *Radzevich* [5, 8, and 9]:

$$\sin\xi_{P1} = \frac{\eta_{P1}}{\sqrt{\eta_{P1}^2 - 2\eta_{P1}\cos\omega_{P1} + 1}}\sin\omega_{P1} \tag{3.15}$$

where the ratio $\partial U_{P1}/\partial V_{P1}$ is denoted by η_{P1}, ($\eta_{P1} = \partial U_{P1}/\partial V_{P1}$).

In the event $F_{P1} = 0$, the equality $\tan\xi_{P1} = \eta_{P1}$ is observed. Here, the ratio η_{P1} is calculated as the root of the quadratic equation:

$$\begin{aligned}(F_{P1}L_{P1} - E_{P1}M_{P1})\eta_{P1}^2 + (G_{P1}L_{P1} - E_{P1}N_{P1})\eta_{P1} \\ + (G_{P1}M_{P1} - F_{P1}N_{P1}) = 0\end{aligned} \tag{3.16}$$

that follows immediately from Eq. (1.68).

The equation for calculation of the actual value of the angle ξ_{P1} yields another representation. Following the chain rule, $\mathrm{d}\mathbf{r}_{P1}$ can be represented in the form:

$$\mathrm{d}\mathbf{r}_{P1} = \mathbf{U}_{P1}\mathrm{d}U_{P1} + \mathbf{V}_{P1}\mathrm{d}V_{P1} \tag{3.17}$$

By definition, $\tan\xi_{P1} = \frac{\sin\xi_{P1}}{\cos\xi_{P1}}$. The functions $\sin\xi_{P1}$ and $\cos\xi_{P1}$ yield representation in the form:

$$\sin\xi_{P1} = \frac{|\mathbf{U}_{P1} \times \mathrm{d}\mathbf{r}_{P1}|}{|\mathbf{U}_{P1}| \cdot |\mathrm{d}\mathbf{r}_{P1}|} \tag{3.18}$$

$$\cos\xi_{P1} = \frac{\mathbf{U}_{P1} \cdot \mathrm{d}\mathbf{r}_{P1}}{|\mathbf{U}_{P1}| \cdot |\mathrm{d}\mathbf{r}_{P1}|} \tag{3.19}$$

The last expressions make it possible the formula:

$$\begin{aligned}\tan\xi_{P1} &= \frac{\sin\xi_{P1}}{\cos\xi_{P1}} = \frac{|\mathbf{U}_{P1} \times \mathrm{d}\mathbf{r}_{P1}|}{\mathbf{U}_{P1} \cdot \mathrm{d}\mathbf{r}_{P1}} = \frac{|\mathbf{U}_{P1} \times \mathrm{d}\mathbf{r}_{P1}|}{\mathbf{U}_{P1} \cdot (\mathbf{U}_{P1}\mathrm{d}U_{P1} + \mathbf{V}_{P1}\mathrm{d}V_{P1})} \\ &= \frac{|\mathbf{U}_{P1} \times \mathrm{d}\mathbf{r}_{P1}| \cdot \mathrm{d}V_{P1}}{\mathbf{U}_{P1} \cdot \mathbf{U}_{P1}\mathrm{d}U_{P1} + \mathbf{U}_{P1} \cdot \mathbf{V}_{P1}\mathrm{d}V_{P1}}\end{aligned} \tag{3.20}$$

By definition,

$$\mathbf{U}_{P1} \cdot \mathbf{U}_{P1} = E_{P1} \tag{3.21}$$

$$\mathbf{U}_{P1} \cdot \mathbf{V}_{P1} = F_{P1} \tag{3.22}$$

$$\mathbf{V}_{P1} \cdot \mathbf{V}_{P1} = G_{P1} \tag{3.23}$$

$$|\mathbf{U}_{P1} \times \mathbf{V}_{P1}| = \sqrt{E_{P1}G_{P1} - F_{P1}^2} \tag{3.24}$$

Then, Eq. (3.17) through Eq. (3.24) yield the formula:

$$\tan\xi_{P1} = \frac{\sqrt{E_{P1}G_{P1} - F_{P1}^2}}{\eta_P \cdot E_{P1} + F_{P1}} \tag{3.25}$$

for calculation of the actual value of the angle ξ_{P1}.

Equations, similar to those above Eqs. (3.15) and (3.25), are also valid for the calculation of the actual value of the angle ξ_{P2} that the first principal direction $\mathbf{t}_{1.P2}$ on the part surface P_2 makes with the unit tangent vector $\mathbf{u}_{P2}$.

The analysis performed above yields the following equations for the calculation of the unit tangent vectors $\mathbf{t}_{1.P1}$, $\mathbf{t}_{2.P1}$ of the principal directions:

$$\mathbf{t}_{1.P1} = \mathbf{Rt}(\xi_{P1}, \mathbf{n}_{P1}) \cdot \mathbf{u}_{P1} \tag{3.26}$$

$$\mathbf{t}_{2.P1} = \mathbf{Rt}\Big[\Big(\xi_{P1} + \frac{\pi}{2}\Big), \mathbf{n}_{P1}\Big] \cdot \mathbf{u}_{P1} \tag{3.27}$$

for the first part surface P_1, and similar equations for the calculation of the unit tangent vectors $\mathbf{t}_{1.T}$, $\mathbf{t}_{2.T}$ of the principal directions:

$$\mathbf{t}_{1.P2} = \mathbf{Rt}(\xi_{P2},\ \mathbf{n}_{P2}) \cdot \mathbf{u}_{P2} \tag{3.28}$$

$$\mathbf{t}_{2.P2} = \mathbf{Rt}\Big[\Big(\xi_{P2} + \frac{\pi}{2}\Big),\ \mathbf{n}_{P2}\Big] \cdot \mathbf{u}_{P2} \tag{3.29}$$

for the second part surface P_2.

See Appendix B for the calculation of the operators of rotation $\mathbf{Rt}(\xi_{P1},\ \mathbf{n}_{P1})$, $\mathbf{Rt}\big[\big(\xi_{P1} + \frac{\pi}{2}\big),\ \mathbf{n}_{P1}\big]$, and $\mathbf{Rt}(\xi_{P2},\ \mathbf{n}_{P2})$, $\mathbf{Rt}\big[\big(\xi_{P2} + \frac{\pi}{2}\big),\ \mathbf{n}_{P2}\big]$ in Eq. (3.26) through Eq. (3.29) accordingly.

3.4 First-Order Analysis: Common Tangent Plane

Various methods can be implemented for analytical description of the contact geometry of two smooth regular part surfaces in the first order of tangency. The first-order analysis is the simplest among all of them.

First-order analysis is the simplest method used for the analysis of geometry of contact of two smooth regular part surfaces P_1 and P_2 at a point K of their contact. Implementation of the first-order analysis returns limited information on the contact geometry of the surfaces in the differential vicinity of the point of contact. Accurate analytical description of the geometry of contact of two smooth regular part surfaces is possible only when the first-order analysis is used together with a kind of higher-order analysis. Under such a scenario, the first-order analysis is incorporated as a first step of a higher-order analysis of the contact geometry of two smooth regular part surfaces P_1 and P_2.

The common tangent plane can be defined as a plane through the contact point K. This plane is perpendicular to the common unit normal vector $\mathbf{n}_P$ to the part surfaces P_1 and P_2 at K. For an analytical description of the common tangent plane, vector equation [see Eq. (1.15)]:

$$(\mathbf{r}_{\mathrm{tp}} - \mathbf{r}_K) \cdot \mathbf{u}_{P1} \cdot \mathbf{v}_{P1} = 0 \tag{3.30}$$

can be employed.

Any pair of unit vectors chosen from the set of vectors, namely from the set of vectors $\mathbf{u}_{P1}$, $\mathbf{v}_{P1}$, $\mathbf{u}_{P2}$, $\mathbf{v}_{P2}$, $\mathbf{t}_{1.P1}$, $\mathbf{t}_{2.P1}$, $\mathbf{t}_{1.P2}$, $\mathbf{t}_{2.P2}$ can be used as a replacement for the unit tangent vectors $\mathbf{u}_{P1}$ and $\mathbf{v}_{P1}$ in Eq. (3.30).

The first-order analysis provides very limited information about the contact geometry of two smooth regular part surfaces in first order of tangency. Hence, the first-order analysis itself has limited application. However, the importance of the first-order analysis becomes valuable as it is incorporated into the second-order analysis, and/or into a higher-order analysis.

3.5 Second-Order Analysis

For a more accurate analytical description of the contact geometry of two part surfaces P_1 and P_2, consideration of the second-order parameters is required. Elements of both of the first order and second order are taken into consideration when the second-order analysis undertaken. To perform the second-order analysis, familiarity with "*Dupin indicatrix*" is highly desirable. The "*Dupin indicatrix*" is a perfect starting point for consideration of the second-order analysis.

It makes sense to begin the discussion of the contact geometry of two smooth regular part surfaces in the first order of tangency from the assumptions those made by the founder of contact mechanics of materials, German physicist *Heinrich Hertz*. The theory proposed by *Hertz* is based on the implementation of an imaginary surface, which (after *Hertz*) is referred to as "*surface of relative curvature*."

3.5.1 Comments on Analytical Description of Local Geometry of Contacting Surfaces Loaded by a Normal Force: "Hertz Proportional Assumption"

Investigation of the geometry of interacting surfaces under an applied normal load can be traced back to the fundamental research by *H. Hertz*[8] on contact of solid elastic bodies (*Hertz*, 1881–1896).

In 1886–1889, *Hertz* published two articles on what was to become known as the field of "*contact mechanics*" of materials. His work basically summarizes how two axisymmetric objects placed in contact will behave under loading. The developed theory is based on *Hertz's* observation of elliptical "*Newton's rings*" formed upon placing a glass sphere upon a lens as the basis of assuming that the pressure exerted by the sphere follows an elliptical distribution.

---

[8] *Heinrich Rudolf* ***Hertz*** (February 22, 1857–January 1, 1894), a famous German physicist.

data 01.05.2020

S T Q Q [✓ S] S D

158) 20200501

A) Books_75_2020/
Important/
Two Millenia of Mathematics...pdf

Obs: Pi
Ler todo livro!

B) Books_75_2020/
Important/
9353162807.pdf

→ Improper integrals

C) Books_75_2020/
Important/Hagihara/ ✡

→ 5 Volumes.

D) Books_76_2020/
Generation of Surfaces Kinematic
Geometry of Surface Machining/

Radzevich, Stephen - Geometry
of Surfaces ....pdf

Pág. 67(88) Chapter 3
Early Works in the Field of Contact
Geometry of Surfaces
Curvature.

is not true; that is to say, the tangent planes coincide but the osculating quadrics differ.

In Eq. (3.1) through Eq. (3.5), the position vector of a point of the first of the contacting surfaces P_1 is designated $\mathbf{r}_{P1}$, and the position vector of a point of the second of the contacting surface P_2 is denoted $\mathbf{r}_{P2}$. Both the contacting surfaces P_1 and P_2 must be parameterized so as to have common "*Gaussian coordinates*" U_P and V_P. In a case the part surfaces P_1 and P_2 are expressed in different U_P- and V_P-coordinates, then it is required to re-parameterize them to common U_P- and V_P-parameters (see Appendix C).

A "*curve*" has contact of at least order n with a "*surface*" if there exists a curve in the surface, which has nth-order contact with the curve. One defines the contact of space curves in the same way as the contact of planar curves mentioned earlier.

3.2 Contact Geometry of Part Surfaces

When solving problems in engineering geometry of surfaces, an appropriate method for analytical description of the contact geometry of part surfaces—say of a part surface P_1 and a part surface P_2—is required. The problem of analytical description of the contact geometry of two smooth regular part surfaces in the first order of tangency is a sophisticated one. It can be solved on the premise of analysis of geometry of the contacting surfaces in the differential vicinity of the point of their contact.

Various methods for analytical description of the contact geometry of smooth regular surfaces have been developed. An overview of the known methods can be found in the monograph by Prof. *Radzevich* [4]. The latest achievements in the field are discussed in the papers [5, 6] and in the monograph [7]. In these publications as well as in other works, two part surfaces P_1 and P_2 that make contact at a point K are considered. Locally, each of the contacting surfaces is represented by a corresponding quadric. Owing to the fact that only differential vicinity of the contacting surfaces is considered below, that same designation P_1 and P_2 is applied to the quadrics as to the surfaces P_1 and P_2 themselves.

Generally speaking, at the point of contact K, the contacting part surfaces P_1 and P_2 may feature an arbitrary configuration. Namely, the local part surface patches P_1 and P_2 may be turned around a common perpendicular through a certain angle. Thus, it makes sense to begin the discussion from the analytical description of local relative orientation of the part surfaces P_1 and P_2 (in the differential vicinity of the point of contact of the surfaces).

3.3 Local Relative Orientation of Contacting Part Surfaces

Consider two local part surface patches P_1 and P_2, which make contact at a point K as illustrated in Fig. 3.1. At the contact point K, the surfaces P_1 and P_2 share a

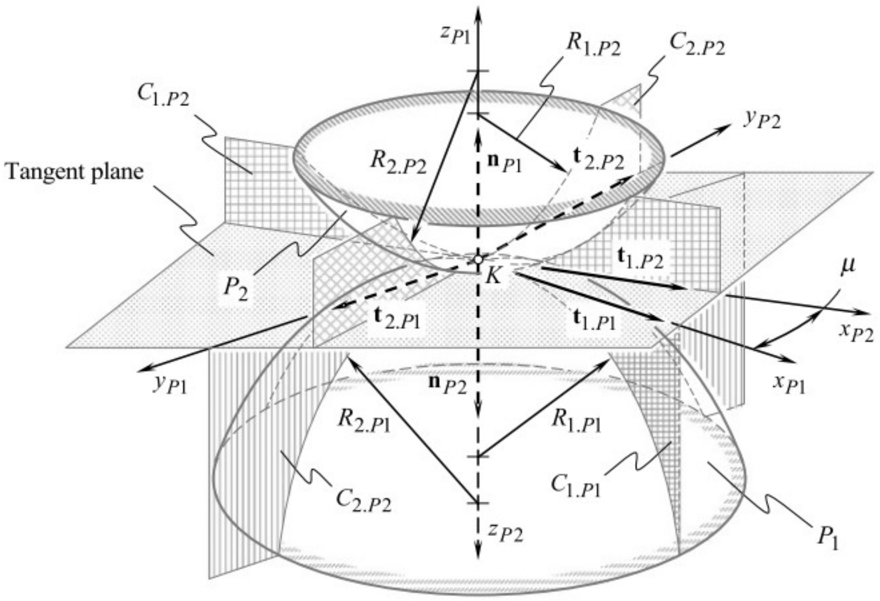

Fig. 3.1 Local configuration of two smooth regular part surface patches P_1 and P_2 at a point K of their contact

common tangent plane. Tangency of the surfaces imposes a kind of restrictions on the relative configuration (location and orientation) of these surfaces, and on their relative motion, if any. In engineering geometry of surface generation, a quantitative measure of the surfaces P_1 and P_2 relative orientation is established.

Relative orientation of the contacting part surfaces P_1 and P_2 is specified by the angle μ of the surfaces "*local relative orientation*".[6] By definition [5, 8, and 9], the angle μ is equal to the angle that the unit tangent vector $\mathbf{t}_{1.P1}$ of the first principal direction of the part surface P_1 forms with the unit tangent vector $\mathbf{t}_{1.P2}$ of the first principal direction of the part surface P_2. That same angle μ can also be specified as the angle made by the unit tangent vectors $\mathbf{t}_{2.P1}$ and $\mathbf{t}_{2.P2}$ of the second principal directions of the part surfaces P_1 and P_2 at a point K of their contact. This immediately yields the equations for the calculation of the angle μ of the surfaces local relative orientation:

$$\sin\mu = |\mathbf{t}_{1.P1} \times \mathbf{t}_{1.P2}| = |\mathbf{t}_{2.P1} \times \mathbf{t}_{2.P2}| \tag{3.6}$$

$$\cos\mu = \mathbf{t}_{1.P1} \cdot \mathbf{t}_{1.P2} = \mathbf{t}_{2.P1} \cdot \mathbf{t}_{2.P2} \tag{3.7}$$

---

[6] Orientation of the surfaces P_1 and P_2 is "*local*" in nature as it is related only to the differential vicinity of point K of contact of the surfaces.

$$\tan \mu = \frac{|\mathbf{t}_{1.P1} \times \mathbf{t}_{1.P2}|}{\mathbf{t}_{1.P1} \cdot \mathbf{t}_{1.P2}} \equiv \frac{|\mathbf{t}_{2.P1} \times \mathbf{t}_{2.P2}|}{\mathbf{t}_{2.P1} \cdot \mathbf{t}_{2.P2}} \tag{3.8}$$

Here, we define:

$\mathbf{t}_{1.P1}$, $\mathbf{t}_{2.P1}$	Are the unit vectors of principal directions on the first part surface P_1 of two contacting surfaces
$\mathbf{t}_{1.P2}$, $\mathbf{t}_{2.P2}$	Are the unit vectors of principal directions on the second part surface P_2 of two contacting surfaces.

In case of point contact of the part surfaces P_1 and P_2, the actual value of the angle μ is calculated at the point K of contact of the surfaces. If the part surfaces P_1 and P_2 are in line contact, then the actual value of the angle μ of the surfaces' local relative orientation can be calculated at every point within the line of contact.[7] The reason for why the angle μ is referred to as the angle of "*local*" surfaces' orientation is clearly illustrated with this.

The determination of the angle μ of the surfaces' P_1 and P_2 local relative orientation is illustrated in Fig. 3.2.

Unit tangent vectors of the principal directions $\mathbf{t}_{1.P1}$ and $\mathbf{t}_{1.P2}$ at a point of contact of the part surfaces P_1 and P_2 are employed for the purpose of calculation of the angle μ of the surfaces' local relative orientation.

Consider two part surfaces P_1 and P_2 in point contact that are represented in a common reference system. The surfaces make contact at point K. For further analysis, an equation of the common tangent plane to the surfaces P_1 and P_2 at the point K is necessary (Figs. 3.1 and 3.3):

$$(\mathbf{r}_{\text{tp}} - \mathbf{r}_K) \cdot \mathbf{u}_{P1} \cdot \mathbf{v}_{P1} = 0 \tag{3.9}$$

Here, we define:

$\mathbf{r}_{\text{tp}}$	Is the position vector of a point of the common tangent plane
$\mathbf{r}_K$	Is the position vector of the contact point K
$\mathbf{u}_{P1}$ and $\mathbf{v}_{P1}$	are the unit vectors tangent to U_{P1}- and V_{P1}-coordinate lines on the surface P_1 at the point K.

An equation similar to Eq. (3.9) can also be expressed in terms of parameters of geometry of the other contacting surface P_2.

The actual value of the angle ω_{P1} between the vectors $\mathbf{v}_{P1}$ and $\mathbf{v}_{P1}$ can be calculated from one of the following equations:

$$\sin \omega_{P1} = \frac{\sqrt{E_{P1}G_{P1} - F_{P1}^2}}{\sqrt{E_{P1}G_{P1}}} \tag{3.10}$$

---

[7] It is worth pointing out here that for line contact, the relative orientation of the surfaces P_1 and P_2 is predetermined in global sense. However, the actual value of the angle μ of the surfaces' P_1 and P_2 "*local*" relative orientation at different points of the line of contact is different.

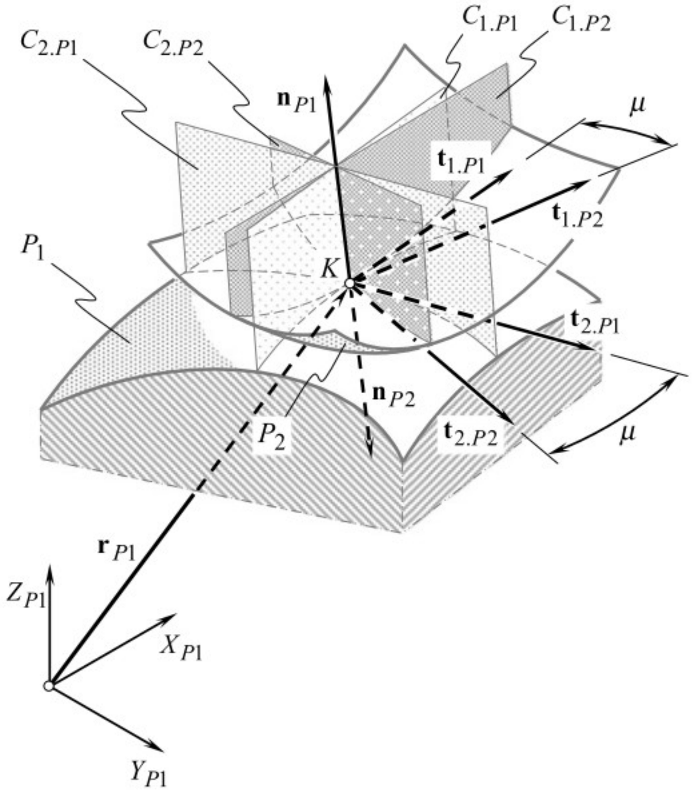
$C_{2.P1}$
$C_{2.P2}$
$C_{1.P1}$
$C_{1.P2}$
$\mathbf{n}_{P1}$
μ
$\mathbf{t}_{1.P1}$
$\mathbf{t}_{1.P2}$
P_1
K
$\mathbf{t}_{2.P1}$
$\mathbf{n}_{P2}$
P_2
$\mathbf{t}_{2.P2}$
μ
$\mathbf{r}_{P1}$
Z_{P1}
X_{P1}
Y_{P1}

Fig. 3.2 Specification of the angle μ of the surfaces P_1 and P_2 local relative orientation

$$\cos \omega_{P1} = \frac{F_{P1}}{\sqrt{E_{P1} G_{P1}}} \tag{3.11}$$

$$\tan \omega_{P1} = \frac{\sqrt{E_{P1} G_{P1} - F_{P1}^2}}{F_{P1}} \tag{3.12}$$

Equations similar to Eq. (3.9) through Eq. (3.12) are also valid for the calculation of the angle ω_{P2} of the other contacting surface P_2.

The unit tangent vectors $\mathbf{u}_{P1}$ and $\mathbf{v}_{P1}$, which are tangent to the U_{P1}- and V_{P1}-coordinate lines on the part surface P_1, and the unit tangent vectors $\mathbf{u}_{P2}$ and $\mathbf{v}_{P2}$, which are tangent to the U_{P2}- and V_{P2}-coordinate lines on the surface P_2, are specified by the angles θ and ε. The actual values of the angles θ and ε can be calculated from the following equations:

$$\cos \theta = \mathbf{u}_{P1} \cdot \mathbf{u}_{P2} \tag{3.13}$$

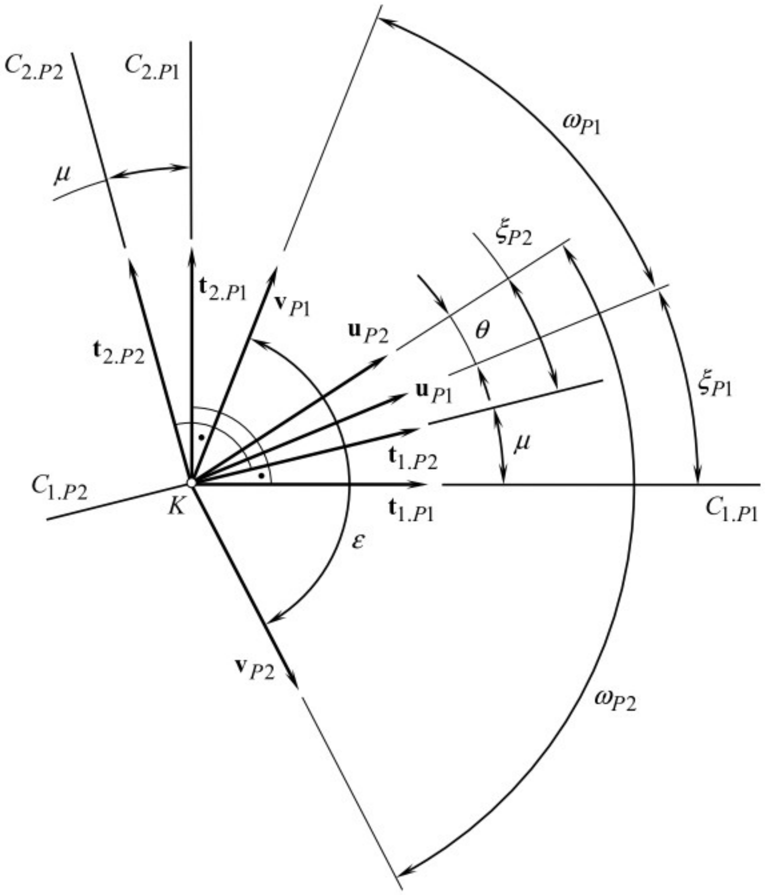
$C_{2.P2}$
$C_{2.P1}$
ω_{P1}
μ
ξ_{P2}
$\mathbf{t}_{2.P1}$
$\mathbf{v}_{P1}$
$\mathbf{u}_{P2}$
θ
$\mathbf{t}_{2.P2}$
$\mathbf{u}_{P1}$
ξ_{P1}
μ
$\mathbf{t}_{1.P2}$
$C_{1.P2}$
K
$\mathbf{t}_{1.P1}$
$C_{1.P1}$
ε
$\mathbf{v}_{P2}$
ω_{P2}

Fig. 3.3 *Local relative orientation* of the contacting part surfaces P_1 and P_2, considered in a common tangent plane

$$\cos \varepsilon = \mathbf{v}_{P1} \cdot \mathbf{v}_{P2} \tag{3.14}$$

Angle ξ_P is the angle that the first principal direction $\mathbf{t}_{1.P}$ on the surface P forms with the unit tangent vector $\mathbf{u}_P$ (see Fig. 3.3). The equation for the calculation of the actual value of the angle ξ_P is derived by Prof. *Radzevich* [5, 8, and 9]:

$$\sin \xi_{P1} = \frac{\eta_{P1}}{\sqrt{\eta_{P1}^2 - 2\eta_{P1}\cos\omega_{P1} + 1}} \sin \omega_{P1} \tag{3.15}$$

where the ratio $\partial U_{P1}/\partial V_{P1}$ is denoted by η_{P1}, ($\eta_{P1} = \partial U_{P1}/\partial V_{P1}$).

In the event $F_{P1} = 0$, the equality $\tan \xi_{P1} = \eta_{P1}$ is observed. Here, the ratio η_{P1} is calculated as the root of the quadratic equation:

$$\begin{aligned}&(F_{P1}L_{P1}-E_{P1}M_{P1})\eta_{P1}^2+(G_{P1}L_{P1}-E_{P1}N_{P1})\eta_{P1}\\&\quad+(G_{P1}M_{P1}-F_{P1}N_{P1})=0\end{aligned}\tag{3.16}$$

that follows immediately from Eq. (1.68).

The equation for calculation of the actual value of the angle ξ_{P1} yields another representation. Following the chain rule, $\mathrm{d}\mathbf{r}_{P1}$ can be represented in the form:

$$\mathrm{d}\mathbf{r}_{P1}=\mathbf{U}_{P1}\mathrm{d}U_{P1}+\mathbf{V}_{P1}\mathrm{d}V_{P1}\tag{3.17}$$

By definition, $\tan\xi_{P1}=\frac{\sin\xi_{P1}}{\cos\xi_{P1}}$. The functions $\sin\xi_{P1}$ and $\cos\xi_{P1}$ yield representation in the form:

$$\sin\xi_{P1}=\frac{|\mathbf{U}_{P1}\times\mathrm{d}\mathbf{r}_{P1}|}{|\mathbf{U}_{P1}|\cdot|\mathrm{d}\mathbf{r}_{P1}|}\tag{3.18}$$

$$\cos\xi_{P1}=\frac{\mathbf{U}_{P1}\cdot\mathrm{d}\mathbf{r}_{P1}}{|\mathbf{U}_{P1}|\cdot|\mathrm{d}\mathbf{r}_{P1}|}\tag{3.19}$$

The last expressions make it possible the formula:

$$\begin{aligned}\tan\xi_{P1}&=\frac{\sin\xi_{P1}}{\cos\xi_{P1}}=\frac{|\mathbf{U}_{P1}\times\mathrm{d}\mathbf{r}_{P1}|}{\mathbf{U}_{P1}\cdot\mathrm{d}\mathbf{r}_{P1}}=\frac{|\mathbf{U}_{P1}\times\mathrm{d}\mathbf{r}_{P1}|}{\mathbf{U}_{P1}\cdot(\mathbf{U}_{P1}\mathrm{d}U_{P1}+\mathbf{V}_{P1}\mathrm{d}V_{P1})}\\&=\frac{|\mathbf{U}_{P1}\times\mathrm{d}\mathbf{r}_{P1}|\cdot\mathrm{d}V_{P1}}{\mathbf{U}_{P1}\cdot\mathbf{U}_{P1}\mathrm{d}U_{P1}+\mathbf{U}_{P1}\cdot\mathbf{V}_{P1}\mathrm{d}V_{P1}}\end{aligned}\tag{3.20}$$

By definition,

$$\mathbf{U}_{P1}\cdot\mathbf{U}_{P1}=E_{P1}\tag{3.21}$$

$$\mathbf{U}_{P1}\cdot\mathbf{V}_{P1}=F_{P1}\tag{3.22}$$

$$\mathbf{V}_{P1}\cdot\mathbf{V}_{P1}=G_{P1}\tag{3.23}$$

$$|\mathbf{U}_{P1}\times\mathbf{V}_{P1}|=\sqrt{E_{P1}G_{P1}-F_{P1}^2}\tag{3.24}$$

Then, Eq. (3.17) through Eq. (3.24) yield the formula:

$$\tan\xi_{P1}=\frac{\sqrt{E_{P1}G_{P1}-F_{P1}^2}}{\eta_P\cdot E_{P1}+F_{P1}}\tag{3.25}$$

for calculation of the actual value of the angle ξ_{P1}.

Equations, similar to those above Eqs. (3.15) and (3.25), are also valid for the calculation of the actual value of the angle ξ_{P2} that the first principal direction $\mathbf{t}_{1.P2}$ on the part surface P_2 makes with the unit tangent vector $\mathbf{u}_{P2}$.

The analysis performed above yields the following equations for the calculation of the unit tangent vectors $\mathbf{t}_{1.P1}$, $\mathbf{t}_{2.P1}$ of the principal directions:

$$\mathbf{t}_{1.P1} = \mathbf{Rt}(\xi_{P1}, \mathbf{n}_{P1}) \cdot \mathbf{u}_{P1} \tag{3.26}$$

$$\mathbf{t}_{2.P1} = \mathbf{Rt}\Big[\Big(\xi_{P1} + \frac{\pi}{2}\Big), \mathbf{n}_{P1}\Big] \cdot \mathbf{u}_{P1} \tag{3.27}$$

for the first part surface P_1, and similar equations for the calculation of the unit tangent vectors $\mathbf{t}_{1.T}$, $\mathbf{t}_{2.T}$ of the principal directions:

$$\mathbf{t}_{1.P2} = \mathbf{Rt}(\xi_{P2},\ \mathbf{n}_{P2}) \cdot \mathbf{u}_{P2} \tag{3.28}$$

$$\mathbf{t}_{2.P2} = \mathbf{Rt}\Big[\Big(\xi_{P2} + \frac{\pi}{2}\Big),\ \mathbf{n}_{P2}\Big] \cdot \mathbf{u}_{P2} \tag{3.29}$$

for the second part surface P_2.

See Appendix B for the calculation of the operators of rotation $\mathbf{Rt}(\xi_{P1},\ \mathbf{n}_{P1})$, $\mathbf{Rt}\big[\big(\xi_{P1} + \frac{\pi}{2}\big),\ \mathbf{n}_{P1}\big]$, and $\mathbf{Rt}(\xi_{P2},\ \mathbf{n}_{P2})$, $\mathbf{Rt}\big[\big(\xi_{P2} + \frac{\pi}{2}\big),\ \mathbf{n}_{P2}\big]$ in Eq. (3.26) through Eq. (3.29) accordingly.

3.4 First-Order Analysis: Common Tangent Plane

Various methods can be implemented for analytical description of the contact geometry of two smooth regular part surfaces in the first order of tangency. The first-order analysis is the simplest among all of them.

First-order analysis is the simplest method used for the analysis of geometry of contact of two smooth regular part surfaces P_1 and P_2 at a point K of their contact. Implementation of the first-order analysis returns limited information on the contact geometry of the surfaces in the differential vicinity of the point of contact. Accurate analytical description of the geometry of contact of two smooth regular part surfaces is possible only when the first-order analysis is used together with a kind of higher-order analysis. Under such a scenario, the first-order analysis is incorporated as a first step of a higher-order analysis of the contact geometry of two smooth regular part surfaces P_1 and P_2.

The common tangent plane can be defined as a plane through the contact point K. This plane is perpendicular to the common unit normal vector $\mathbf{n}_P$ to the part surfaces P_1 and P_2 at K. For an analytical description of the common tangent plane, vector equation [see Eq. (1.15)]:

$$(\mathbf{r}_{\text{tp}} - \mathbf{r}_K) \cdot \mathbf{u}_{P1} \cdot \mathbf{v}_{P1} = 0 \tag{3.30}$$

can be employed.

Any pair of unit vectors chosen from the set of vectors, namely from the set of vectors $\mathbf{u}_{P1}$, $\mathbf{v}_{P1}$, $\mathbf{u}_{P2}$, $\mathbf{v}_{P2}$, $\mathbf{t}_{1.P1}$, $\mathbf{t}_{2.P1}$, $\mathbf{t}_{1.P2}$, $\mathbf{t}_{2.P2}$ can be used as a replacement for the unit tangent vectors $\mathbf{u}_{P1}$ and $\mathbf{v}_{P1}$ in Eq. (3.30).

The first-order analysis provides very limited information about the contact geometry of two smooth regular part surfaces in first order of tangency. Hence, the first-order analysis itself has limited application. However, the importance of the first-order analysis becomes valuable as it is incorporated into the second-order analysis, and/or into a higher-order analysis.

3.5 Second-Order Analysis

For a more accurate analytical description of the contact geometry of two part surfaces P_1 and P_2, consideration of the second-order parameters is required. Elements of both of the first order and second order are taken into consideration when the second-order analysis undertaken. To perform the second-order analysis, familiarity with "*Dupin indicatrix*" is highly desirable. The "*Dupin indicatrix*" is a perfect starting point for consideration of the second-order analysis.

It makes sense to begin the discussion of the contact geometry of two smooth regular part surfaces in the first order of tangency from the assumptions those made by the founder of contact mechanics of materials, German physicist *Heinrich Hertz*. The theory proposed by *Hertz* is based on the implementation of an imaginary surface, which (after *Hertz*) is referred to as "*surface of relative curvature*."

3.5.1 *Comments on Analytical Description of Local Geometry of Contacting Surfaces Loaded by a Normal Force:* "Hertz Proportional Assumption"

Investigation of the geometry of interacting surfaces under an applied normal load can be traced back to the fundamental research by *H. Hertz*[8] on contact of solid elastic bodies (*Hertz*, 1881–1896).

In 1886–1889, *Hertz* published two articles on what was to become known as the field of "*contact mechanics*" of materials. His work basically summarizes how two axisymmetric objects placed in contact will behave under loading. The developed theory is based on *Hertz's* observation of elliptical "*Newton's rings*" formed upon placing a glass sphere upon a lens as the basis of assuming that the pressure exerted by the sphere follows an elliptical distribution.

---

[8] *Heinrich Rudolf* **Hertz** (February 22, 1857–January 1, 1894), a famous German physicist.

01.05.2020

Pág. 70(91)

Fig. 3.1 Local configuration of two smooth regular part surfaces P_1 and P_2 at a point K of their contact

Obs: Livro todo!!!

Pág. 72(93)

Fig. 3.2 Specification of the angle μ of the surfaces P_1 and P_2 local relative orientation

Pág. 73(94)

Fig. 3.3 Local relative orientation of the contacting part surfaces P_1 and P_2, considered in a common tangent plane. ✡

Important:

Aplicar estes conceitos no estudo da curvatura do Universo devidamente projetado em uma superfície plana.

Pág. 83(104)

3.5.3 "Dupin Indicatrix" at a Point of Surface of Relative Normal Curvature ✡

It is of importance to stress here again that all of three normal curvatures $k_{\mathscr{R}}$, k_{P1}, and k_{P2} in Eq. (4.36) are taken in a common normal plane section through the point[10] of contact K of the surfaces P_1, P_2, and $\mathscr{R}$.

Based on the calculated values of the principal curvatures $k_{1.\mathscr{R}}$ and $k_{2.\mathscr{R}}$, the implicit equation of the surface of relative curvature yields representation in the form:

$$2Z_{\mathscr{R}} = k_{1.\mathscr{R}} X_{\mathscr{R}}^2 + k_{2.\mathscr{R}} Y_{\mathscr{R}}^2 \tag{3.47}$$

A characteristic surface similar to the surface $\mathscr{R}$ of relative normal curvature can be defined as the surface for which the equality $R_{\mathscr{R}} = R_{P1} - R_{P2}$ is observed. Evidently, this equality is similar in nature to the equality Eq. (4.36).

3.5.3 "Dupin Indicatrix" *at a Point of Surface of Relative Normal Curvature*

Consider an intersection of the surface of relative normal curvature by a plane, parallel to the tangent plane at the point K, but only at a small distance away from it. Then, project the intersection on the tangent plane. In the coordinate plane $x_P y_P$, principal part of the intersection will be given by the equation of "*Dupin indicatrix*"[11] [2]

$$\mathrm{Dup}(\mathscr{R}) \Rightarrow \frac{L_{\mathscr{R}}}{E_{\mathscr{R}}} x_{\mathscr{R}}^2 + \frac{2M_{\mathscr{R}}}{\sqrt{E_{\mathscr{R}} G_{\mathscr{R}}}} x_{\mathscr{R}} y_{\mathscr{R}} + \frac{N_{\mathscr{R}}}{G_{\mathscr{R}}} y_{\mathscr{R}}^2 = \pm 1 \tag{3.48}$$

which describes the distribution of normal relative curvature within the differential vicinity of the point K.

Here, $E_{\mathscr{R}}$, $F_{\mathscr{R}}$, and $G_{\mathscr{R}}$ designate fundamental magnitudes of the first order, and $L_{\mathscr{R}}$, $M_{\mathscr{R}}$, and $N_{\mathscr{R}}$ designate fundamental magnitudes of the second order at the point K of the surface $\mathscr{R}$.

If the axes $x_{\mathscr{R}}$ and $y_{\mathscr{R}}$ of the local coordinate system $x_{\mathscr{R}} y_{\mathscr{R}}$ align with the principal directions $\mathbf{t}_{1.\mathscr{R}}$ and $\mathbf{t}_{2.\mathscr{R}}$ of the surface of relative curvature $\mathscr{R}$, then Eq. (3.48) reduces to:

$$\mathrm{Dup}(\mathscr{R}) \Rightarrow k_{1.\mathscr{R}} x_{\mathscr{R}}^2 + k_{2.\mathscr{R}} y_{\mathscr{R}}^2 = \pm 1 \tag{3.49}$$

---

[10] In a case of line contact of the surfaces P_1 and P_2, point K is the point on the line of the surfaces' contact at which the normal curvatures $k_{\mathscr{R}}$, k_{P1}, and k_{P2} are required to be calculated.

[11] To be more exact, the distribution of normal relative radii of curvature $R_{\mathscr{R}}$, and not of normal relative curvature $k_{\mathscr{R}}$ itself, is reflected by the "*Dupin indicatrix*, $\mathrm{Dup}(\mathscr{R})$". Thus, it could be designated as $\mathrm{Dup}_R(\mathscr{R})$. However, the equation of the indicatrix $\mathrm{Dup}_k(\mathscr{R})$ of a surface normal curvature could also be composed.

Similarly, the corresponding equations for the normalized $\mathrm{Dup}_R(\mathscr{R})$ indicatrix of relative normal radius of curvature, and the indicatrix of normal curvature $\mathrm{Dup}_k(\mathscr{R})$ could also easily be derived.

An important intermediate conclusion follows immediately from Eq. (3.48):

Conclusion 3.2 *The direction* $\mathbf{t}_{1.\mathscr{R}}$ *for the maximum* $k_{1.\mathscr{R}}$*, and the direction* $\mathbf{t}_{2.\mathscr{R}}$ *for the minimum* $k_{2.\mathscr{R}}$ *values of normal curvature at a point of the surface of relative normal curvature are always orthogonal to one another, and, therefore, the condition* $\mathbf{t}_{1.\mathscr{R}} \perp \mathbf{t}_{2.\mathscr{R}}$ *is always observed.*

The major axes of the "*Dupin indicatrix*, Dup($\mathscr{R}$)" make the angles φ_{min} and φ_{max} with the principal directions $\mathbf{t}_{1.\mathscr{R}}$ and $\mathbf{t}_{2.\mathscr{R}}$.

3.5.4 Matrix Representation of Equation of "Dupin Indicatrix" *at a Point of Surface of Relative Normal Curvature*

Like any other quadratic form, Eq. (3.48) of the "*Dupin indicatrix*" of the surface of normal relative curvature $\mathscr{R}$ can be represented in matrix form

$$\text{Dup}(\mathscr{R}) \Rightarrow [\,x_P \;\; y_P \;\; 0 \;\; 0\,] \cdot \begin{bmatrix} \frac{L_{\mathscr{R}}}{E_{\mathscr{R}}} & \frac{2M_{\mathscr{R}}}{\sqrt{E_{\mathscr{R}}G_{\mathscr{R}}}} & 0 & 0 \\ \frac{2M_{\mathscr{R}}}{\sqrt{E_{\mathscr{R}}G_{\mathscr{R}}}} & \frac{N_{\mathscr{R}}}{G_{\mathscr{R}}} & 0 & 0 \\ 0 & 0 & \mp 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_P \\ y_P \\ 0 \\ 0 \end{bmatrix} = \pm 1 \tag{3.50}$$

In "*Darboux trihedron,*" this equation reduces to:

$$\text{Dup}(\mathscr{R}) \Rightarrow [\,x_P \;\; y_P \;\; 0 \;\; 0\,] \cdot \begin{bmatrix} L_{\mathscr{R}} & M_{\mathscr{R}} & 0 & 0 \\ M_{\mathscr{R}} & N_{\mathscr{R}} & 0 & 0 \\ 0 & 0 & \mp 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_P \\ y_P \\ 0 \\ 0 \end{bmatrix} = \pm 1 \tag{3.51}$$

It is convenient to implement matrix representation of the equation of the "*Dupin indicatrix*, Dup($\mathscr{R}$)" (see above), for instance, when multiple coordinate system transformations need to be performed.

The equation of the "*Dupin indicatrix*, $Dup(\mathscr{R})$" can be represented in the form:

$$r_{\text{Dup}}(\varphi) = \sqrt{|R_{\mathscr{R}}(\varphi)|} \cdot \text{sgn}\Phi_{2.\mathscr{R}}^{-1} \tag{3.52}$$

This last equation reveals that the position vector of a point of the indicatrix Dup($\mathscr{R}$) in any direction is equal to the square root of the radius of curvature in that same direction.[12]

---

[12] Similar to *Dupin* indicatrix Dup($\mathscr{R}$), a planar characteristic curve of another kind can be introduced as well. The equation of this characteristic curve can be postulated in the form $r_{\text{Dup}.k}(\varphi) = \sqrt{|k_{\mathscr{R}}(\varphi)|} \cdot \text{sgn}\Phi_{2.\mathscr{R}}^{-1}$. Application of curvature indicatrix in the form $r_{\text{Dup}.k}(\varphi)$ yields

3.5.5 Surface of Relative Normal Radii of Curvature

The normal curvature $k_{\mathscr{R}}$ of the surface of normal relative curvature $\mathscr{R}$, the normal curvatures k_{P1}, and k_{P2} of the part surfaces P_1 and P_2 can be represented in the form: $k_{\mathscr{R}} = R_{\mathscr{R}}^{-1}$, $k_{P1} = R_{P1}^{-1}$, and $k_{P2} = R_{P2}^{-1}$ correspondingly, where $R_{\mathscr{R}}$, R_{P1}, and R_{P2} are the corresponding radii of normal curvature of the surfaces $\mathscr{R}$, P_1, and P_2. They are also taken in a common normal plane section through the point of contact K of these surfaces $\mathscr{R}$, P_1, and P_2.

The radius of relative normal curvature is another widely known tool that is used in contemporary practice for the purposes of analytical description of the geometry of contact when performing second-order analysis. The radius $R_{\mathscr{R}}$ of relative normal curvature can be defined by the expression:

$$R_{\mathscr{R}} = R_{P1} - R_{P2} \tag{3.53}$$

In many applications, Eq. (3.53) for the radius of relative normal of curvature $R_{\mathscr{R}}$ is equivalent to Eq. (3.46) for the relative normal curvature $k_{\mathscr{R}}$.

3.5.6 Normalized Relative Normal Curvature

Usually, it is preferable to deal with unitless values when performing second-order analysis rather than with values having units. In order to eliminate unit values, it is recommended to use a normalized relative normal curvature $\bar{k}_{\mathscr{R}}$ of the contacting surfaces P_1 and P_2. The normalized relative normal curvature $\bar{k}_{\mathscr{R}}$ of the surfaces P_1 and P_2 is referred to as the value determined by the equation

$$\bar{k}_{\mathscr{R}} = \frac{k_{P1} + k_{P2}}{|k_{1.P}|} \tag{3.54}$$

Similarly, the normalized radius of relative normal curvature $\bar{R}_{\mathscr{R}}$ of the surfaces P_1 and P_2 can be introduced here based on Eq. (3.53). The normalized relative radius of normal curvature $\bar{R}_{\mathscr{R}}$ of the surfaces P_1 and P_2 is referred to as the value determined by the equation:

$$\bar{R}_{\mathscr{R}} = \frac{R_{P1} - R_{P2}}{|R_{1.P1}|} \tag{3.55}$$

Implementation of the unitless parameters $\bar{k}_{\mathscr{R}}$, $\bar{R}_{\mathscr{R}}$, as well as others makes it possible avoiding dealing with unit values. Equations comprised of unitless parameters are often more convenient in applications.

---

avoiding of uncertainty in cases of a plane. For a plane surface Dup($\mathscr{R}$) exists, while $r_{\text{Dup}.k}(\varphi)$ shrinks to the point K.

A corresponding "*Dupin indicatrix*" can be constructed for all of the above-considered characteristic surfaces. Namely for:

(a) the surface of normal relative radii of curvature $\mathrm{Dup}_R(\mathscr{R})$,
(b) the normalized surface of normal relative curvature $\mathrm{Dup}(\bar{\mathscr{R}})$, and
(c) the normalized surface of normal radii of relative curvature $\mathrm{Dup}_R(\bar{\mathscr{R}})$.

3.5.7 Curvature Indicatrix of Surface of Relative Normal Curvature

The curvature indicatrix of the surface of relative normal curvature $\mathscr{R}$ can be introduced similar to already done for a smooth regular part P. This planar characteristic curve is designated as $\mathrm{Crv}(\mathscr{R})$. The curvature indicatrix $\mathrm{Crv}(\mathscr{R})$ can be implemented to make clear whether or not two smooth regular part surfaces P_1 and P_2 make physically feasible contact.

The curvature indicatrices $\mathrm{Crv}(\mathscr{R})$ those shown in Fig. 3.6 correspond to physically feasible kinds of contact of the part surfaces. Two part surfaces P_1 and P_2 can contact one another physically if the curvature indicatrix $\mathrm{Crv}(\mathscr{R})$ is either of convex elliptic type (Fig. 3.6a), or of convex umbilic type (Fig. 3.6b), or finally of convex parabolic type (Fig. 3.6c). Under such circumstances, the curvature indicatrix $\mathrm{Crv}(\mathscr{R})$ of the surface $\mathscr{R}$ of relative normal curvature can be described analytically by the inequality:

$$\mathrm{Crv}(\mathscr{R}) \Rightarrow \frac{L_{\mathscr{R}}}{E_{\mathscr{R}}} x_P^2 + \frac{2M_{\mathscr{R}}}{\sqrt{E_{\mathscr{R}} G_{\mathscr{R}}}} x_P y_P + \frac{N_{\mathscr{R}}}{G_{\mathscr{R}}} y_P^2 > 1 \tag{3.56}$$

In all of these cases, the mean curvature $\mathscr{M}_{\mathscr{R}}$ of the surface $\mathscr{R}$ of relative curvature is of a positive value ($\mathscr{M}_{\mathscr{R}} > 0$).

In a particular case, the surface of relative normal curvature can be reduced either to a point or a line. These degenerated cases of the surface $\mathscr{R}$ are illustrated schematically in Fig. 3.7a, b correspondingly.

In case contact of two smooth regular surfaces P_1 and P_2 is physically impossible, the curvature indicatrix $\mathrm{Crv}(\mathscr{R})$ of the surface of relative normal curvature $\mathscr{R}$ is bounded by the corresponding "*Dupin indicatrix* $\mathrm{Dup}(\mathscr{R})$" position vector of a point of which has negative value [at least for one point for the characteristic curve $\mathrm{Dup}(\mathscr{R})$]. Curvature indicatrices $\mathrm{Crv}(\mathscr{R})$ for such cases are schematically illustrated in Fig. 3.8.

Curvature indicatrices $\mathrm{Crv}(\mathscr{R})$ of a surface of relative normal curvature $\mathscr{R}$ are represented by three concave local patches of the surface $\mathscr{R}$, namely:

(a) concave elliptic (Fig. 3.8a),
(b) concave umbilic (Fig. 3.8b), and
(c) concave parabolic (Fig. 3.8c)

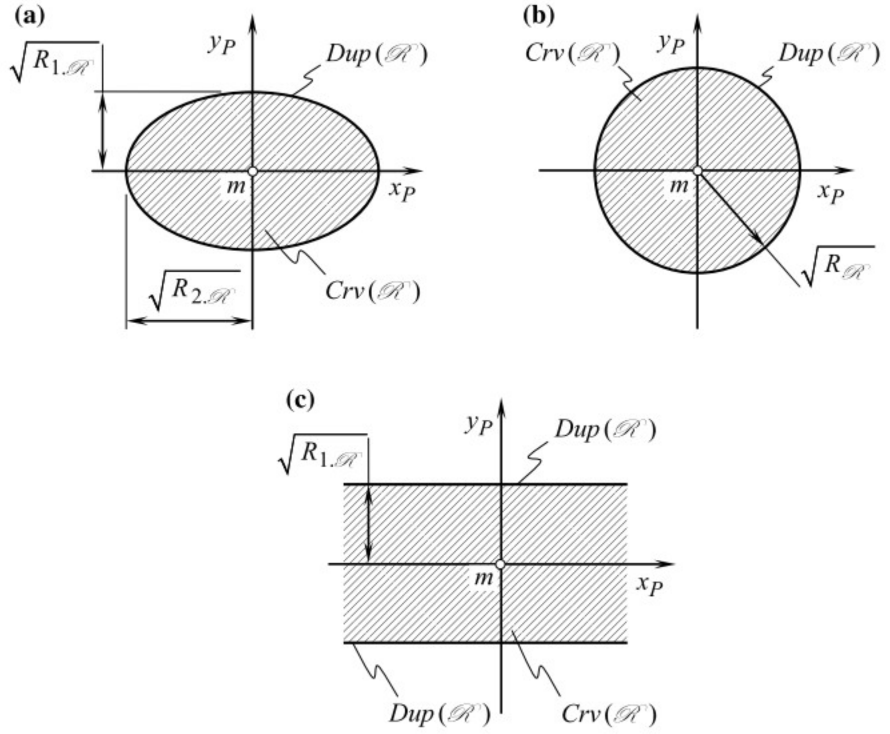
(a)
y_P
$\sqrt{R_{1.\mathscr{R}}}$
$Dup(\mathscr{R})$
m
x_P
$\sqrt{R_{2.\mathscr{R}}}$
$Crv(\mathscr{R})$
(b)
$Crv(\mathscr{R})$
y_P
$Dup(\mathscr{R})$
m
x_P
$\sqrt{R_{\mathscr{R}}}$
(c)
y_P
$Dup(\mathscr{R})$
$\sqrt{R_{1.\mathscr{R}}}$
m
x_P
$Dup(\mathscr{R})$
$Crv(\mathscr{R})$

Fig. 3.6 Three kinds of curvature indicatrices Crv($\mathscr{R}$) at a point m of a surface $\mathscr{R}$ of relative normal curvature for the cases when physical contact of two smooth regular surfaces P_1 and P_2 is feasible

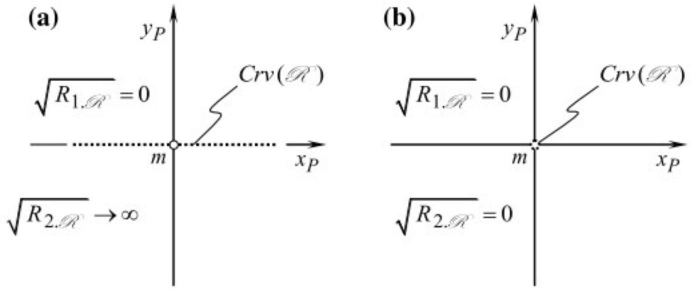
(a)
y_P
$Crv(\mathscr{R})$
$\sqrt{R_{1.\mathscr{R}}} = 0$
m
x_P
$\sqrt{R_{2.\mathscr{R}}} \to \infty$
(b)
y_P
$Crv(\mathscr{R})$
$\sqrt{R_{1.\mathscr{R}}} = 0$
m
x_P
$\sqrt{R_{2.\mathscr{R}}} = 0$

Fig. 3.7 Two kinds of curvature indicatrices Crv($\mathscr{R}$) at a point m of a surface $\mathscr{R}$ of relative normal curvature for the cases when physical contact of two smooth regular surfaces P_1 and P_2 is unidentified

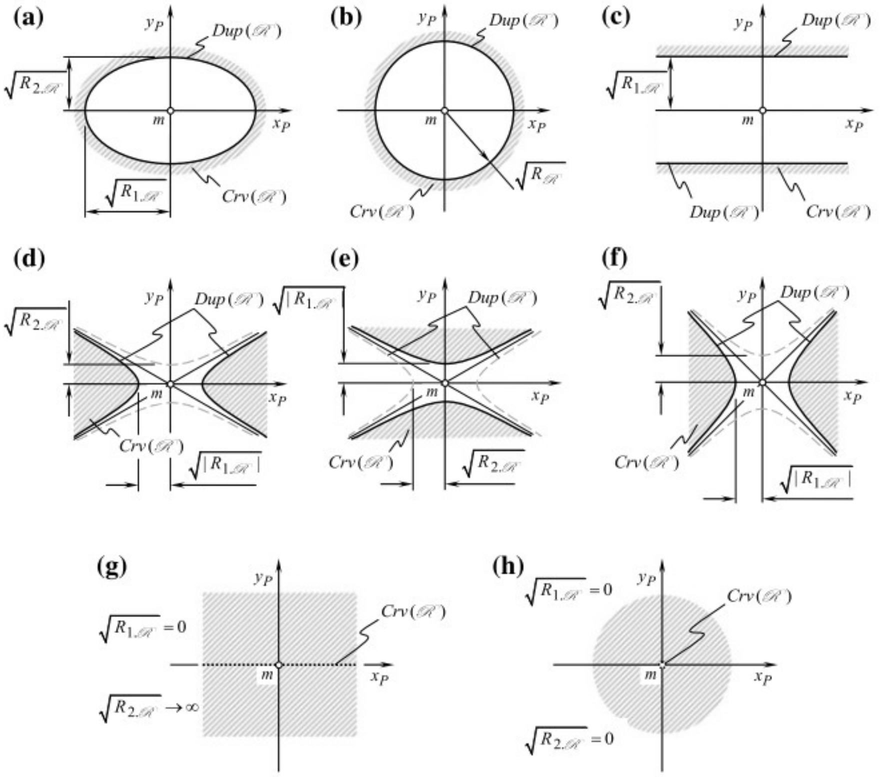
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
y_P
x_P
m
$Dup(\mathscr{R})$
$Crv(\mathscr{R})$
$\sqrt{R_{1.\mathscr{R}}}$
$\sqrt{R_{2.\mathscr{R}}}$
$\sqrt{R_{\mathscr{R}}}$
$\sqrt{|R_{1.\mathscr{R}}|}$
$\sqrt{R_{1.\mathscr{R}}} = 0$
$\sqrt{R_{2.\mathscr{R}}} \to \infty$
$\sqrt{R_{2.\mathscr{R}}} = 0$

Fig. 3.8 Curvature indicatrices Crv($\mathscr{R}$) at a point m of a surface $\mathscr{R}$ of relative normal curvature for cases when physical contact of two smooth regular surfaces P_1 and P_2 is infeasible

local surface patches.

In addition to concave local patches, three saddle-like local patches of the surface of relative normal curvature $\mathscr{R}$ are represented by three local patches having:

(d) pseudo-convex (Fig. 3.8d),
(e) minimal (Fig. 3.8e), and
(f) pseudo-concave (Fig. 3.8f)

local surface patches.

Ultimately, two degenerate kinds of the curvature indicatrices Crv($\mathscr{R}$) of a surface of relative normal curvature $\mathscr{R}$ are represented in Fig. 3.8g, h.

All the indicatrices Crv($\mathscr{R}$) in Fig. 3.8 fulfill the following inequality:

$$\mathrm{Crv}(\mathscr{R}) \Rightarrow \frac{L_{\mathscr{R}}}{E_{\mathscr{R}}} x_P^2 + \frac{2M_{\mathscr{R}}}{\sqrt{E_{\mathscr{R}} G_{\mathscr{R}}}} x_P y_P + \frac{N_{\mathscr{R}}}{G_{\mathscr{R}}} y_P^2 < 1 \tag{3.57}$$

Again, the curvature indicatrices Crv($\mathscr{R}$) shown in Fig. 3.8 indicate that the surfaces P_1 and P_2 cannot contact one another physically.

data 01 05 2020

Pág. 70(91)

Fig. 3.1 Local configuration of two smooth regular part surfaces P_1 and P_2 at a point K of their contact

Obs: Livro todo!!!

Pág. 72(93)

Fig. 3.2 Specification of the angle μ of the surfaces P_1 and P_2 local relative orientation

Pág. 73(94)

Fig. 3.3 Local relative orientation of the contacting part surfaces P_1 and P_2, considered in a common tangent plane. ✡

Important:

Aplicar estes conceitos no estudo da curvatura do Universo devidamente projetado em uma superfície plana.

Pág. 83(104)

3.5.3 "Dupin Indicatrix" at a Point of Surface of Relative Normal Curvature ✡

data 01.05.2020

Consider an intersection of the surface of relative normal curvature by a plane, parallel to the tangent plane at the point K, but only at a small distance away from it.

Veja Capitulo todo!

Pag. 131(150)

6.3 Relative Characteristic Curve.

Pag. 133(152)

Fig. 6.3 "relative indicatrix" at a point of contact K of two part surfaces P_1 and P_2.

Pag. 138(157)

Possible kinds of Contact of two smooth regular Part Surfaces in the First Order of tangency

Fig. 7.1 a) b) c) ☆

Pag. 243(257)

Coordinate Systems and Linear

curve of the second order. In the case under consideration, it is always convex with uniform change of curvature. The "$\mathscr{A}n_R(P_1)$-*indicatrix*" is also a planar smooth regular curve. However, the points of inflection are inherent to this curve. This is due to the fact that the "$\mathscr{A}n_R(P_1)$-*indicatrix*" is a curve of fourth order. The higher order enhances the capabilities of the characteristic curve $\mathscr{A}n_R(P_1)$.

Owing to the higher order of the equation, the "$\mathscr{A}n_R(P_1)$-*indicatrix*" reflects the distribution of normal radii of curvature, while the "*Dupin indicatrix*, $Dup(P_1)$" reflects the distribution of square roots of normal radii of curvature of a part surface P_1 at the surface point m. In order to make the difference clear, it is sufficient to represent the equation of the "*Dupin indicatrix*" in the form similar to that for "$\mathscr{A}n_R(P_1)$-*indicatrix*" [see Eq. (6.37)]:

$$Dup(P_1) \quad \Rightarrow \quad \mathbf{r}_{Dup}(\varphi) = \begin{bmatrix} \sqrt{|R_{P1}(\varphi)|}\cos\varphi \operatorname{sgn} R_{P1}(\varphi) \\ \sqrt{|R_{P1}(\varphi)|}\sin\varphi \operatorname{sgn} R_{P1}(\varphi) \\ 0 \\ 0 \end{bmatrix} \tag{6.39}$$

It is evident from the above consideration that Eqs. (6.37) and (6.39) are similar.

6.3 Relative Characteristic Curve

The considered properties of the "*Plücker conoid*" can be employed for the derivation of the equation of a planar characteristic curve for analytical description of the contact geometry of two smooth regular part surfaces for the needs of engineering geometry of surfaces.

6.3.1 On a Possibility of Implementation of Two "Plücker Conoids"

At first glimpse, implementation of two "*Plücker conoids*" sounds promising for the purpose of solving the problem of analytical description of contact geometry of two smooth regular part surfaces.

In order to develop an appropriate solution to the problem under consideration, the characteristic surface $\mathbf{Pl}_R(P_1/P_2)$ that reflects the "*summa*" of the corresponding normal radii of curvature at a point of contact of the part surfaces P_1 and P_2 could be introduced. The following matrix representation of the equation of the surface $\mathbf{Pl}_R(P_1/P_2)$ follows immediately:

$$\mathbf{Pl}_R(P_1/P_2) \quad \Rightarrow \quad \mathbf{R}_R^*(\varphi) = \begin{bmatrix} (R_{P1} + R_{P2}) \cos\varphi \\ (R_{P1} + R_{P2}) \sin\varphi \\ 2 \sin\varphi \cos\varphi \\ 0 \end{bmatrix} \tag{6.40}$$

The characteristic surface $\mathbf{Pl}_R(P_1/P_2)$ is referred to as the "*Plücker relative conoid*."

Because the centers $C_{1.P1}$ and $C_{2.P1}$ of the principal curvature of the part surfaces P_1, and the centers $C_{1.P2}$ and $C_{2.P2}$ of the principal curvatures of the part surface P_2 in general case do not coincide with one another, the actual reciprocation of the straight-line segment of length $(R_{P1} - R_{P2})$ could be restricted by different pairs of the limiting points $C_{1.P1}$, $C_{2.P1}$, $C_{1.P2}$, $C_{2.P2}$. Various locations of the limiting points on the axis through the unit normal vector $\mathbf{n}_{P1}$ result in "*deformation*" of the characteristic surface $\mathbf{Pl}_R(P_1/P_2)$ in its axial direction. A deformation of such type does not affect the surface appearance in the direction of $(R_{P1} + R_{P2})$ that is of critical importance for engineering geometry of surfaces.

The characteristic surface $\mathbf{Pl}_R(P_1/P_2)$ is described analytically by Eq. (6.40). This indicates that the "*Plücker relative conoid*" correctly reflects the degree of conformity at a point of contact K of the part surfaces P_1 and P_2. However, the surface $\mathbf{Pl}_R(P_1/P_2)$ itself is inconvenient for implementation in engineering geometry of surfaces. In order to fix this undesired particular problem, the "*Plücker relative indicatrix* $\mathbf{Pl}_R(P_1/P_2)$" at a point of contact K of two surfaces P_1 and P_2 is introduced. The equation of this 3*D* characteristic curve immediately follows immediately from Eq. (6.40):

$$\mathbf{Pl}_R(P_1/P_2) \quad \Rightarrow \quad \mathbf{R}_R(\varphi) = \begin{bmatrix} (R_{P1} + R_{P2}) \cos\varphi \\ (R_{P1} + R_{P2}) \sin\varphi \\ (R_{P1} + R_{P2}) \\ 0 \end{bmatrix} \tag{6.41}$$

Here, both radii of normal curvatures, namely R_{P1} and R_{P2}, are signed values. They are of positive values when the lines of intersection of the part surfaces P_1 and P_2 by a plane through the common perpendicular $\mathbf{n}_{P1}$ are convex, and they are of negative values when the lines of intersection of the part surfaces by a plane through the common perpendicular are concave.

Further, the characteristic curve $\mathbf{Pl}_R(P_1/P_2)$ can be reduced to a corresponding planar characteristic curve. Omitting the intermediate considerations, one may wish to go directly to the "*-relative indicatrix*" at a point of contact of the part surfaces P_1 and P_2.

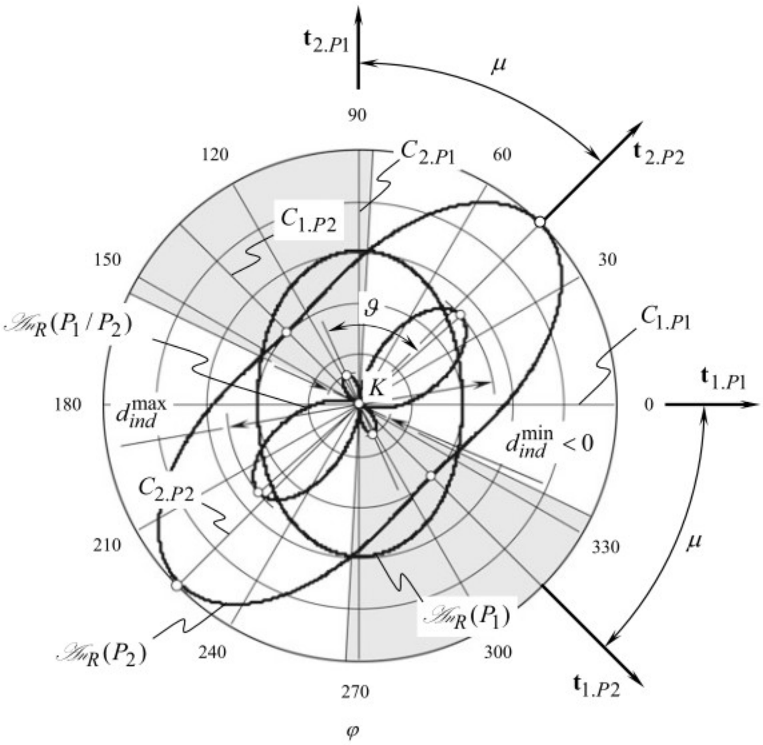
$\mathbf{t}_{2.P1}$
μ
90
120
$C_{2.P1}$
60
$\mathbf{t}_{2.P2}$
$C_{1.P2}$
150
30
$\mathscr{Pl}_R(P_1/P_2)$
ϑ
$C_{1.P1}$
$\mathbf{t}_{1.P1}$
K
180
$d_{ind}^{\max}$
0
$d_{ind}^{\min} < 0$
$C_{2.P2}$
210
330
μ
$\mathscr{Pl}_R(P_1)$
$\mathscr{Pl}_R(P_2)$
240
300
$\mathbf{t}_{1.P2}$
270
φ

Fig. 6.3 An example of the "$\mathscr{Pl}_R(P_1/P_2)$*-relative indicatrix*" at a point of contact K of two part surfaces P_1 and P_2 ($R_{1.P1} = 2\,\text{mm}$, $R_{2.P1} = 3\,\text{mm}$, $R_{1.P2} = -2\,\text{mm}$, $R_{2.P2} = -5\,\text{mm}$, $\mu = 45°$), plotted together with the corresponding "$\mathscr{Pl}_R(P_1)$*-indicatrix*" of the part surface P_1 , and the "$\mathscr{Pl}_R(P_2)$*-indicatrix*" of the part surface P_2 constructed at that same contact point K

6.3.2 "$\mathscr{Pl}_R(P_1/P_2)$**-Relative Indicatrix"** *at Point of Contact of Two Part Surfaces* **P**$_1$ *and* **P**$_2$

Aiming for further simplification of the analytical description of the contact geometry of two part surfaces P_1 and P_2, the "*Plücker relative indicatrix*, $\mathscr{Pl}_R(P_1/P_2)$" can be replaced with the planar characteristic curve of a simpler structure. The equation of the 2*D* "$\mathscr{Pl}_R(P_1/P_2)$*-relative indicatrix*" at a point of contact of the part surfaces P_1 and P_2 can be derived from Eq. (6.41):

$$\mathscr{Pl}_R(P_1/P_2) \;\Rightarrow\; \mathbf{R}_{iR}(\varphi) = \begin{bmatrix} (R_{P1} + R_{P2})\cos\varphi \\ (R_{P1} + R_{P2})\sin\varphi \\ 0 \\ 0 \end{bmatrix} \tag{6.42}$$

This planar characteristic curve is referred to as the "$\mathscr{A}n_R(P_1/P_2)$*-relative indicatrix of the first kind.*" The distribution of summa of normal radii of curvature at a point of contact K of the part surfaces P_1 and P_2 is described analytically by the "$\mathscr{A}n_R(P_1/P_2)$*-relative indicatrix of the first kind.*"

In Eq. (6.42), the radii of normal curvature R_{P1} and R_{P2} can be expressed in terms of the corresponding principal radii of curvature $R_{1.P1}$, $R_{2.P1}$, and $R_{1.P2}$, $R_{2.P2}$, and in terms of the angle μ of the part surfaces' P_1 and P_2 local relative orientation at the contact point K.

An example of the "$\mathscr{A}n_R(P_1/P_2)$*-relative indicatrix*" at a point of contact of two part surfaces P_1 and P_2 is shown in Fig. 6.3. Parameters of the characteristic curve $\mathscr{A}n_R(P_1/P_2)$ are calculated at point of contact of the convex elliptic local patch of the part surface P_1 ($R_{1.P1} = 3\,\text{mm}$, and $R_{2.P1} = 15\,\text{mm}$), and the concave elliptic local patch of the part surface P_2 ($R_{1.P2} = -2\,\text{mm}$, and $R_{2.P2} = -5\,\text{mm}$). The part surfaces P_1 and P_2 are turned around the common perpendicular $\mathbf{n}_{P1}$ with respect to one another through the angle of local relative orientation $\mu = 45°$. The corresponding "$\mathscr{A}n_R(P_1)$*-indicatrix*" as well as the "$\mathscr{A}n_R(P_1)$*-indicatrix*" is also plotted in Fig. 6.3. It is of importance to point out here that the direction of the minimum diameter $d_{ind}^{\min}$ and the direction of the maximum diameter $d_{ind}^{\max}$ of the characteristic curve $\mathscr{A}n_R(P_1/P_2)$ do not align either with the principal directions $\mathbf{t}_{1.P1}$ and $\mathbf{t}_{2.P1}$ on the part surface P_1 or with the principal directions $\mathbf{t}_{1.P2}$ and $\mathbf{t}_{2.P2}$ on the part surface P_2. The extremal directions of the "$\mathscr{A}n_R(P_1/P_2)$*-relative indicatrix*" are not orthogonal to one other. In the general case of two part surfaces in contact, they make a certain angle $\vartheta \neq 90°$.

The following statement can be made at this point:

Conclusion 6.1 *In the general case of part surfaces in contact, the directions of the extremal (i.e., of maximum minimum) degree of conformity of two surfaces P_1 and P_2 at the contact point K are not orthogonal to one another. The directions of the extremal degree of conformity of the part surface P_1 and P_2 could be orthogonal to one another only in a particular (degenerated) cases of the surfaces contact.*

The shape and parameters of the "$\mathscr{A}n_R(P_1/P_2)$*-relative indicatrix*" depend upon the algebraic values of the principal radii of curvature $R_{1.P1}$, $R_{2.P1}$ and $R_{1.P2}$, $R_{2.P2}$ at a point of contact of the part surfaces P_1 and P_2, as well as on the actual value of the angle μ of the part surfaces P_1 and P_2 local relative orientation.

The "*Dupin* indicatrix $Dup\,(\mathscr{R})$" at a point of the surface of relative curvature indicates that the directions of the extremum degree of conformity of the part surfaces P_1 and P_2 at the contact point K are always perpendicular to one another. The above consideration reveals that this is not correct and results in errors of calculations.

The characteristic curve $\mathscr{A}n_R(P_1/P_2)$ is of simpler structure rather than the "*Plücker relative indicatrix*, $\mathscr{P}l_R(P_1/P_2)$" itself. The "$\mathscr{A}n_R(P_1/P_2)$*-relative indicatrix*" is always a kind of planar curve, while the "*Plücker relative indicatrix*, $\mathscr{P}l_R(P_1/P_2)$" is always a spatial 3*D* curve. This makes use of the characteristic curve $\mathscr{A}n_R(P_1/P_2)$, preferred rather than the "*Plücker relative indicatrix*, $\mathscr{P}l_R(P_1/P_2)$."

The distribution of the differences between normal curvatures of the part surfaces P_1 and P_2 at the contact point K can be described analytically by a planar characteristic curve of another kind:

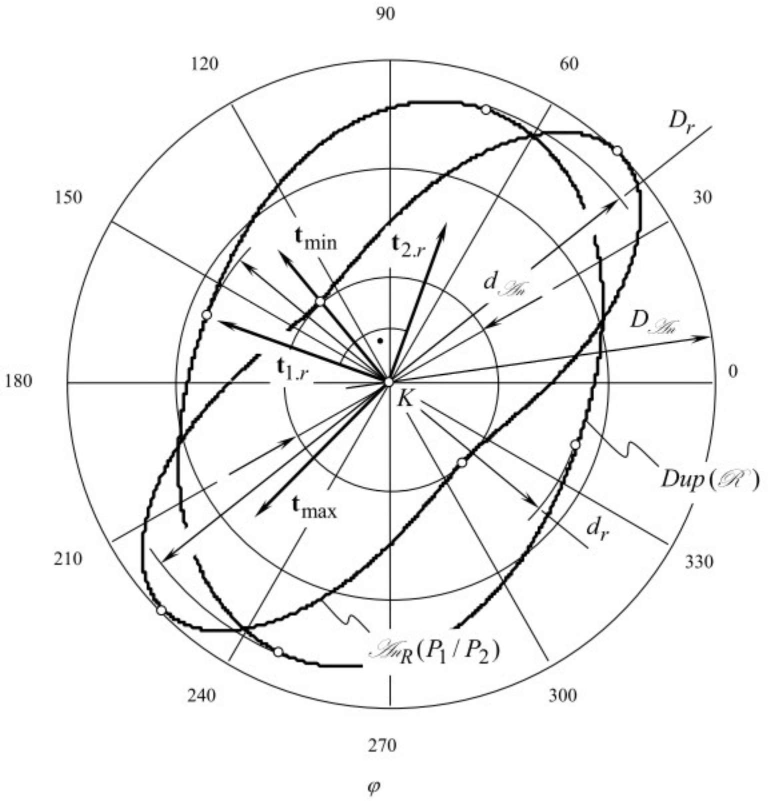
90
120
60
D_r
150
30
$\mathbf{t}_{\min}$
$\mathbf{t}_{2.r}$
$d_{\mathscr{A}n}$
$D_{\mathscr{A}n}$
180
0
$\mathbf{t}_{1.r}$
K
$Dup(\mathscr{R})$
$\mathbf{t}_{\max}$
d_r
210
330
$\mathscr{A}n_R(P_1/P_2)$
240
300
270
φ

Fig. 6.4 Comparison of the *-relative indicatrix* of two part surfaces P_1 and P_2 in contact with the *Dupin* indicatrix $Dup(\mathscr{R})$ of the surface of relative curvature for that same part surfaces P_1 and P_2

$$\mathscr{A}n_k(P_1/P_2) \quad \Rightarrow \quad \mathbf{R}_{ik}(\varphi) = \begin{bmatrix} (k_{P1} - k_{P2})\cos\varphi \\ (k_{P1} - k_{P2})\sin\varphi \\ 0 \\ 1 \end{bmatrix} \tag{6.43}$$

The characteristic curve [see Eq. (6.43)] is referred to as the "$\mathscr{A}n_k(P_1/P_2)$*-relative indicatrix of the second kind*".

Equation (6.43) of the "$\mathscr{A}n_k(P_1/P_2)$*-indicatrix*" can be expressed in terms of radii of principal curvature of the contacting part surfaces P_1 and P_2:

$$\mathscr{A}n_k(P_1/P_2) \quad \Rightarrow \quad \mathbf{R}_{ik}(\varphi) = \begin{bmatrix} [k_{1.P1}\cos\varphi + k_{2.P1}\sin\varphi - k_{1.P2}\cos(\varphi+\mu) - k_{2.P2}\sin(\varphi+\mu)]\cos\varphi \\ [k_{1.P1}\cos\varphi + k_{2.P1}\sin\varphi - k_{1.P2}\cos(\varphi+\mu) - k_{2.P2}\sin(\varphi+\mu)]\sin\varphi \\ 0 \\ 1 \end{bmatrix} \tag{6.44}$$

The difference between the "$\mathscr{I}n_R(P_1/P_2)$-*relative indicatrix of the second kind*" and between the "*Dupin indicatrix*, $Dup(\mathscr{R})$" of the surface of relative curvature is clear illustrated in Fig. 6.4.

The planar characteristic curves $\mathscr{I}n_R(P_1)$ and $\mathscr{I}n_k(P_1)$, as well as the characteristic curves $\mathscr{I}n_R(P_2)$ and $\mathscr{I}n_k(P_2)$ originate from the "*Plücker conoid*." Equations (6.37), (6.38) and (6.42), (6.43) of the corresponding indicatrices, $\mathscr{I}n_R(P_1)$, $\mathscr{I}n_k(P_1)$ and, $\mathscr{I}n_R(P_2)$, $\mathscr{I}n_k(P_2)$ are derived on the premise of Eq. (6.8) of the surface of "*Plücker conoid*" (see [8], and Sect. 4.9, pp. 257–260 in [9]).

It is proven analytically that both planar characteristic curves, say the characteristic curve $\mathscr{I}n_R(P_1/P_2)$, and the "*indicatrix of conformity* $Cnf_R(P_1/P_2)$" at a point of contact of two smooth regular part surfaces P_1 and P_2 specify that same direction $\mathbf{t}_{cnf}^{\max}$ at which the degree of conformity of the part surfaces P_1 and P_2 reaches its maximum value. Both the characteristic curves, $Cnf_R(P_1/P_2)$ and $\mathscr{I}n_R(P_1/P_2)$, are powerful tools in engineering geometry of surfaces. They are extensively used for the analysis of the geometry of contact of two smooth regular part surfaces P_1 and P_2.

References

1. Plücker, J. (1865). On a new geometry of space. *Philosophical Transactions of the Royal Society, 155,* 725–791.
2. Radzevich, S. P. (2004). A possibility of application of plücker conoid for mathematical modeling of contact of two smooth regular surfaces in the first order of tangency. *Mathematical and Computer Modeling, 42,* 999–1022.
3. von Seggern, D. (1993). *CRC standard curves and surfaces* (p. 288). Boca Raton, FL: CRC Press.
4. Fisher, G. (Ed.) (1986). *Mathematical models.* Braunachweig/Wiesbaden: Friedrich Vieweg & Sohn.
5. Gray, A. (1997). Plücker Conoid. *Modern differential geometry of curves and surfaces with mathematics* (2nd ed., pp. 435–437). Boca Raton, FL: CRC Press.
6. http://www.mathcurve.com/surfaces/plucker/plucker.shtml.
7. Struik, D. J. (1961). *Lectures on classical differential Geometry* (2nd ed., p. 232). Massachusetts: Addison-Wesley Publishing Company Inc.
8. Radzevich, S. P. (1991). *Sculptured Surface Machining on Multi-Axis NC Machine*, Monograph, Kiev, Vishcha Schola, 192 p.
9. Radzevich, S.P. (2001). *Fundamentals of surface Generation*, *Monograph*, Kiev, Rastan, 592 p.

Chapter 7

Possible Kinds of Contact of Two Smooth Regular Part Surfaces in the First Order of Tangency

Interaction of two part surfaces in contact is of tremendous importance for plurality of engineering applications. Therefore, analysis and classification of all possible kinds of contact of two smooth regular part surfaces P_1 and P_2 in the first order of tangency is a critical issue in engineering geometry of surfaces. Development of scientific classification of all possible kinds of contact of part surfaces P_1 and P_2 could be considered as the ultimate point in the analysis of the part surfaces P_1 and P_2 contact.

Prior to discussing this important issue, two more particular issues need to be mentioned.

The first issue relates to the possibility of implementation of the indicatrix of conformity for the identification of the actual kind of contact of two part surfaces P_1 and P_2.

The second issue relates to the impact of the accuracy of the computations on the desired parameters of the indicatrix of conformity $Cnf_R(P_1/P_2)$ at a point of contact of the part surfaces P_1 and P_2, that is, onto the directions of extreme degree of their conformity.

7.1 On a Possibility of Implementation of the Indicatrix of Conformity for the Purposes of Identification of the Actual Kind of Contact of Two Smooth Regular Part Surfaces

Two regular part surfaces P_1 and P_2 can make contact either at a point, or along a line, or ultimately over a surface patch. Actually, the kind of contact of two part surfaces results in certain features in the shape and the parameters of the indicatrix of conformity $Cnf_R(P_1/P_2)$ of the contacting part surfaces P_1 and P_2.

As follows from Eqs. (6.34) and (6.35), special features in the shape and the parameters of the indicatrix of conformity $Cnf_R(P_1/P_2)$ are inherited to every pos-

© Springer Nature Switzerland AG 2020
S. P. Radzevich, *Geometry of Surfaces*,
https://doi.org/10.1007/978-3-030-22184-3_7

sible kind of contact of the part surfaces P_1 and P_2. For example, when two smooth regular part surfaces P_1 and P_2 make contact:

(a) at a point K (Fig. 7.1a), then the minimum diameter $d_{cnf}^{\min}$ of the indicatrix of

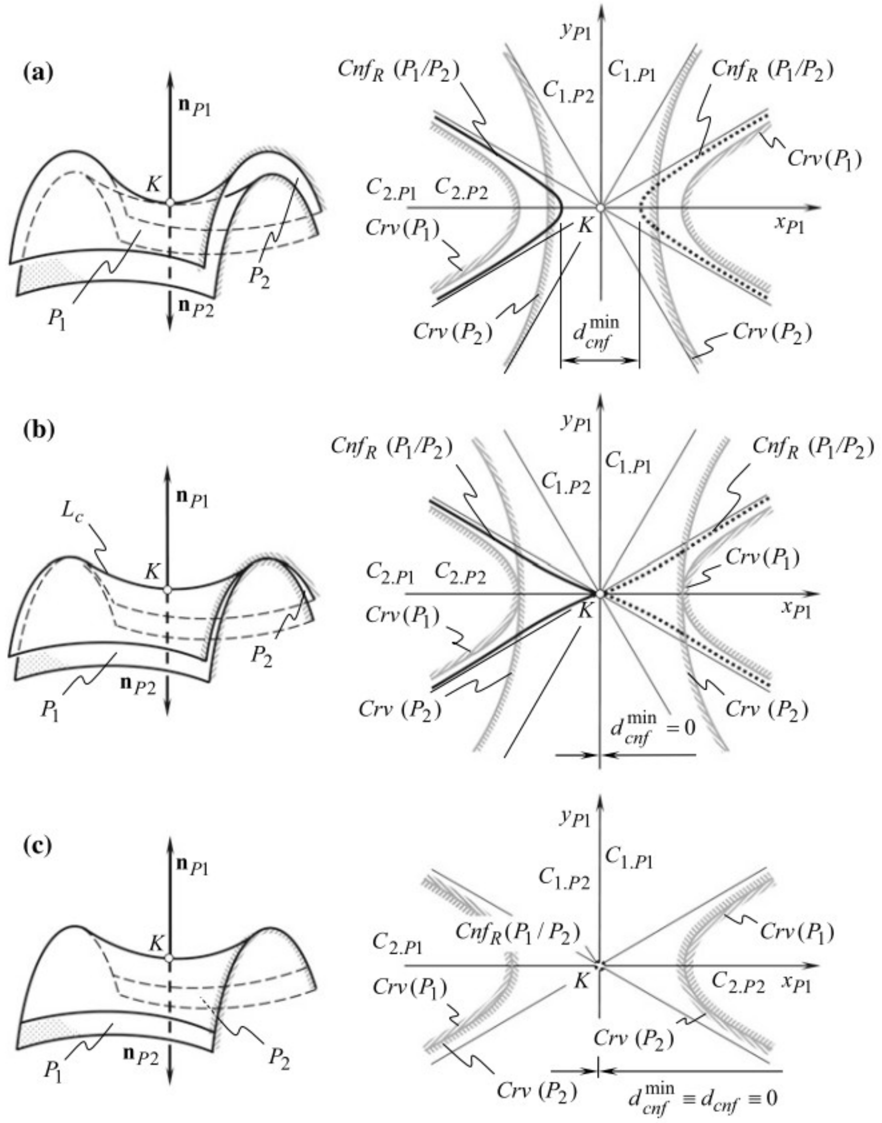
(a)
$\mathbf{n}_{P1}$
K
P_1
P_2
$\mathbf{n}_{P2}$
y_{P1}
$Cnf_R\ (P_1/P_2)$
$C_{1.P1}$
$C_{1.P2}$
$Cnf_R\ (P_1/P_2)$
$Crv(P_1)$
$C_{2.P1}$ $C_{2.P2}$
$Crv(P_1)$
K
x_{P1}
$Crv\ (P_2)$
$d_{cnf}^{\min}$
$Crv\ (P_2)$
(b)
$\mathbf{n}_{P1}$
L_c
K
P_1
P_2
$\mathbf{n}_{P2}$
y_{P1}
$Cnf_R\ (P_1/P_2)$
$C_{1.P1}$
$C_{1.P2}$
$Cnf_R\ (P_1/P_2)$
$Crv(P_1)$
$C_{2.P1}$ $C_{2.P2}$
$Crv(P_1)$
K
x_{P1}
$Crv\ (P_2)$
$d_{cnf}^{\min} = 0$
$Crv\ (P_2)$
(c)
$\mathbf{n}_{P1}$
K
P_1
$\mathbf{n}_{P2}$
P_2
y_{P1}
$C_{1.P1}$
$C_{1.P2}$
$C_{2.P1}$
$Cnf_R(P_1 / P_2)$
$Crv(P_1)$
$Crv(P_1)$
K
$C_{2.P2}$ x_{P1}
$Crv\ (P_2)$
$Crv\ (P_2)$
$d_{cnf}^{\min} \equiv d_{cnf} \equiv 0$

Fig. 7.1 Correspondence between the shape of the indicatrix of conformity $Cnf_R(P_1/P_2)$ and the kind of contact of two part surfaces P_1 and P_2: **a** point contact, **b** line contact, and **c** surface contact

conformity $Cnf_R(P_1/P_2)$ at a point of contact (as well as all other diameters of this characteristic curve) is always of a positive value ($d_{cnf}^{\min} > 0$);

(b) along a curve L_c (Fig. 7.1b), then the minimum diameter $d_{cnf}^{\min}$ of the indicatrix of conformity $Cnf_R(P_1/P_2)$ at a point of contact is always identical to zero ($d_{cnf}^{\min} \equiv 0$), while all other diameters of this characteristic curve are of a positive value ($d_{cnf} > 0$), and finally

(c) over a surface patch (Fig. 7.1c), then the indicatrix of conformity at a point of contact of the part surfaces P_1 and P_2 shrinks to a point, which coincides with the point K of contact of the part surfaces P_1 and P_2.

The above examples are worked out for the cases of the part surfaces P_1 and P_2 contact, when in the local vicinity of the contact point K both of the surfaces P_1 and P_2 are smooth regular surfaces of saddled type (having negative "*Gauss curvature*" $\mathscr{G}_{P1} < 0$, and $\mathscr{G}_{P2} < 0$). A similarity is observed for all other types of local patches of part surfaces P_1 and P_2 as well.

In a case when the part surfaces P_1 and P_2 intersect one another, that is, the interference of the surfaces P_1 and P_2 into each other is observed as illustrated schematically in Fig. 5.6, then the minimum diameter of the indicatrix of conformity $Cnf_R(P_1/P_2)$ at a point of contact of the part surfaces P_1 and P_2 is always of a negative value ($d_{cnf}^{\min} < 0$).

It is of importance to make difference between partial and full interference of the part surfaces P_1 and P_2 in the differential vicinity of the contact point K.

For instance, in the differential vicinity of the point of contact K, a convex elliptical local patch of the part surface P_2 can partially intersect a hyperbolical local patch of the part surface P_1 (Fig. 5.6a). In this case, the minimum diameter $d_{cnf}^{\min}$of the indicatrix of conformity $Cnf_R(P_1/P_2)$ is of a negative value ($d_{cnf}^{\min} < 0$). Owing to this, the indicatrix of conformity $Cnf_R(P_1/P_2)$ not only intersects itself, but intersection of each of its branches also occurs. Varying the angular parameter φ within the interval $0 \leq \varphi \leq \pi$, positive ($d_{cnf} > 0$), as well as negative ($d_{cnf} < 0$) values of the current diameter d_{cnf} of the characteristic curve $Cnf_R(P_1/P_2)$ are observed.

Another example: In the differential vicinity of the contact point K, the local patch of the part surface P_2 interferes into a local patch of the part surface P_1 totally, as shown for two hyperbolic local patches (Fig. 5.6b) and for two parabolic (Fig. 5.6c) local patches of the part surfaces P_1 and P_2. For the cases of total local interference of the part surfaces P_1 and P_2, the minimum diameter $d_{cnf}^{\min}$of the indicatrix of conformity $Cnf_R(P_1/P_2)$ is always of a negative value ($d_{cnf}^{\min} < 0$), as well as all other diameters d_{cnf} of this characteristic curve $Cnf_R(P_1/P_2)$ are also of a negative value, regardless of the actual value of the angular parameter φ.

The examples above show that every type of contact of the part surfaces P_1 and P_2 features important peculiarities in the shape and parameters of the indicatrix of conformity $Cnf_R(P_1/P_2)$ at the point K of contact of the part surfaces P_1 and P_2. The shape and the parameters of the characteristic curve $Cnf_R(P_1/P_2)$ uniquely follow from the actual kind of contact of the part surfaces P_1 and P_2 and are completely predetermined by the actual kind of contact of these surfaces. Here, it is natural to assume that the inverse statement could also be true. The question can be formulated as follows:

Question 7.1 *Are the features in the shape and parameters of the indicatrix of conformity* $Cnf_R(P_1/P_2)$ *at a point of contact of two part surfaces in contact* P_1 *and* P_2 *necessary and sufficient for drawing a conclusion regarding the kind of surfaces'* P_1 *and* P_2 *contact: contact at a point, along a curve* L_c*, or over a surface patch*?

Or, in another words, could the value and sign of the minimum diameter $d_{cnf}^{\min}$of the indicatrix of conformity $Cnf_R(P_1/P_2)$ and the features in its shape be implemented as a criterion for determining uniquely the kind of contact of the part surfaces P_1 and P_2?

The detailed investigation of this particular sub-problem enables one to make the following statement:

Conclusion 7.1 *The actual value and sign of the minimum diameter* $d_{cnf}^{\min}$*of the indicatrix of conformity* $Cnf_R(P_1/P_2)$ *at a point of contact of two part surfaces* P_1 *and* P_2*, as well as features in the shape of this characteristic curve, cannot be implemented as a sufficient criterion for determining uniquely the actual kind of contact of the part surfaces* P_1 *and* P_2.

The positive value of the minimum diameter $d_{cnf}^{\min}$of the indicatrix of conformity $Cnf_R(P_1/P_2)$ (that is, $d_{cnf}^{\min} > 0$) is a sufficient but not a necessary requirement for contact of the part surfaces P_1 and P_2 at a point.

The indicatrix of conformity $Cnf_R(P_1/P_2)$, that is shrunk to the contact point K, is not sufficient to identify the kind of contact of the part surfaces P_1 and P_2 over a surface patch. However, if two part surfaces P_1 and P_2 are congruent to each other within a certain surface patch, then the indicatrix of conformity shrinks to a point, which coincides with the point of contact K. The inverse statement is not correct. In the event of the indicatrix of conformity $Cnf_R(P_1/P_2)$ shrinking to a point, then the part surfaces P_1 and P_2 can be congruent to one another only "*locally.*" Thus, if the indicatrix of conformity $Cnf_R(P_1/P_2)$ shrinks to the contact point K, this indicates only a necessary but not a sufficient condition for the part surfaces P_1 and P_2 contact over a certain surface patch. In the case under consideration, the part surfaces P_1 and P_2 can make contact along a characteristic curve L_c and/or at a point K as well.

If the minimum diameter $d_{cnf}^{\min}$of the indicatrix of conformity $Cnf_R(P_1/P_2)$ at a point of contact of the part surfaces P_1 and P_2 is equal to zero, this does not necessarily mean that the part surfaces P_1 and P_2 make contact along a curve L_c. This requirement is only a necessary but not a sufficient condition for the line contact of two smooth regular part surfaces. In the event $d_{cnf}^{\min} = 0$, the part surfaces P_1 and P_2 can make contact at a point.

As follows from the above discussion, in the event, the part surfaces P_1 and P_2 make contact along a line L_c, then the direction along the minimum diameter $d_{cnf}^{\min}$of the indicatrix of conformity $Cnf_R(P_1/P_2)$ at a point of contact of the part surfaces P_1 and P_2 is aligned with the tangent line to the curve L_c at the contact point K. This issue is of importance for engineering geometry of surfaces because it follows directly from the statement made above. According to this statement, in the differential vicinity of the point of contact K, the direction, along which minimum diameter $d_{cnf}^{\min}$ can be measured, is aligned with the direction, along which the degree of conformity at a point of contact of the part surfaces P_1 and P_2 reaches a maximal value. Therefore, at the contact point K, the direction along the minimum diameter $d_{cnf}^{\min}$ and the direction that is tangent to the line of contact L_c align to one another. Hence, contact point K is a point of tangency of:

(a) the straight line that aligns with the direction in which the minimum diameter $d_{cnf}^{\min}$ of the indicatrix of conformity $Cnf_R(P_1/P_2)$ is measured, and
(b) the line of contact L_c of the part surfaces P_1 and P_2.

The statements above are also true for the inverse indicatrix of conformity $Cnf_k(P_1/P_2)$ at a point of contact of the part surfaces P_1 and P_2 in the first order of tangency.

It is of critical importance to stress here that some of the conclusions similar to those listed above could be made based on other known approaches. This proves indirectly that the proposed novel approach for analytical description of the geometry of contact of two smooth regular part surfaces in the first order of tangency is in perfect agreement with the earlier developed approaches in the field. In a case when the contacting part surfaces are of a simple geometry (that is, planes, spheres, cylinders of revolution), implementation of the discussed approach, as well as implementation of known approaches for analytical description of contact geometry of two part surfaces returns identical results. If the contacting part surfaces are of a complex geometry, then implementation only of the indicatrix of conformity (of both kinds), or of the relative indicatrix (of both kinds), and not of any other characteristic curve, enable correct results of computation of parameters of the geometry of contact of two surfaces.

7.1.1 Impact of Accuracy of Computation on the Parameters of the Indicatrices of Conformity $Cnf_R(P_1/P_2)$

All the calculations, whether performed manually or by computer, are subject to errors of the computation. No calculations of any kind are performed with zero error. Errors of the calculation are unavoidable for many reasons.

The accuracy of calculation affects the desired parameters of the indicatrix of conformity $Cnf_R(P_1/P_2)$ at a point of contact of the part surfaces P_1 and P_2. For a predetermined error of the calculation, the optimal parameters of the characteristic curve $Cnf_R(P_1/P_2)$ can be computed. As is known, the characteristic curve $Cnf_R(P_1/P_2)$ is a function of the design parameters of two part surfaces P_1 and P_2 and of their relative orientation. Commonly, the design parameters of the given part

data 01.05.2020

Consider an intersection of the surface of relative normal curvature by a plane, parallel to the tangent plane at the point K, but only at a small distance away from it.

Veja capitulo todo!

Pag. 131 (150)

6.3 Relative Characteristic Curve.

Pag. 133 (152)

Fig. 6.3 "relative indicatrix" at a point of contact K of two part surfaces P_1 and P_2.

Pag. 138 (157)

Possible kinds of contact of two smooth regular part surfaces in the first order of tangency

Fig. 7.1 a) b) c) ☆

Pag. 243 (257)

Coordinate Systems and Linear

Jandaia

data 01.05.2020

Transformations

Pág. 277(291)

B.5 Impact of the Coordinate System Transformations on Fundamentals Forms of the Surface.

obs: Veja livro todo!!!

E) Books_76_2020/ Geometric Folding Algorithms/

Geometric Folding Algorithms...pdf

obs: Livro todo!!!

Pág. 30(44)

3.1.3 Peaucellier Linkage

Pág. 32(45)

Fig. 3.2 ☆

Pág. 299(311) Part III Polyhedra

Pág. 301(315) 21.2 Curvature

Pág. 302(316) ☆

Figure 21.4 →

(" = >

$$\mathbf{Rs}\left(K_g \to K_p\right) = \mathbf{Rs}^{-1}\left(K_p \to K_g\right) = \mathbf{Rt}(-\varphi_z, z_p) \tag{B.64}$$

Implementation of the discussed results allows for:

(a) representation of the gear tooth flank, $\mathscr{G}$, and the pinion tooth flank, $\mathscr{P}$, of the form cutting tool, as well as their relative motion in a common coordinate system, and
(b) consideration meshing of the gear tooth flank, $\mathscr{G}$, in any desired coordinate systems that is a component of the chain and/or the closed loop of consequent coordinate system transformations.

Transition from one coordinate system to another coordinate system can be performed in both of two feasible directions, say in direct as well as in inverse directions.

B.5 Impact of the Coordinate Systems Transformations on Fundamental Forms of the Surface

Every coordinate system transformation results in a corresponding change to equation of the gear tooth flank, $\mathscr{G}$, and/or pinion tooth flank, $\mathscr{P}$. Because of this, it is often necessary to re-calculate coefficients of the first $\Phi_{1.g}$ and of the second $\Phi_{2.g}$ fundamental forms of the gear tooth flank, $\mathscr{G}$ (as many times as the coordinate system transformation is performed). This routing and time-consuming operation can be eliminated if the operators of coordinate system transformations are used directly to the fundamental forms $\Phi_{1.g}$ and $\Phi_{2.g}$.

After been calculated in an initial reference system, the fundamental magnitudes E_g, F_g, and G_g of the first, $\Phi_{1.g}$, and the fundamental magnitudes L_g, M_g, and N_g of the second, $\Phi_{2.g}$, fundamental forms can be determined in any new coordinate system using for this purpose, the operators of translation, of rotation, and of resultant coordinate system transformation [106]. Transformation of such kind of the fundamental magnitudes, $\Phi_{1.g}$ and $\Phi_{2.g}$, becomes possible due to implementation of a formula, that can be found out immediately below.

Let us consider a gear tooth flank, $\mathscr{G}$, that is given by equation $\mathbf{r}_g = \mathbf{r}_g(U_g, V_g)$, where $(U_g, V_g) \in G$.

For the analysis below, it is convenient to use the equation of the first fundamental form, $\Phi_{1.g}$, of the gear tooth flank, $\mathscr{G}$, represented in matrix form:

$$\left[\Phi_{1.g}\right] = \left[\, dU_g \quad dV_g \quad 0 \quad 0 \,\right] \cdot \begin{bmatrix} E_g & F_g & 0 & 0 \\ F_g & G_g & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} dU_g \\ dV_g \\ 0 \\ 0 \end{bmatrix} \tag{B.65}$$

Similarly, equation of the second fundamental form $\Phi_{2.g}$ of the gear tooth flank, $\mathcal{G}$, can be given by:

$$\left[\Phi_{2.g}\right] = \left[\, dU_g \quad dV_g \quad 0 \quad 0 \,\right] \cdot \begin{bmatrix} L_g & M_g & 0 & 0 \\ M_g & N_g & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} dU_g \\ dV_g \\ 0 \\ 0 \end{bmatrix} \tag{B.66}$$

The coordinate system transformation that is performed by the operator of linear transformation $\mathbf{Rs}(1 \to 2)$ transfers the equation $\mathbf{r}_g = \mathbf{r}_g(U_g, V_g)$ of the gear tooth flank, $\mathcal{G}$, initially given in $X_1Y_1Z_1$, to the equation $\mathbf{r}_g^* = \mathbf{r}_g^*(U_g^*, V_g^*)$ of that same gear tooth flank, $\mathcal{G}$, in a new coordinate system $X_2Y_2Z_2$. It is clear that $\mathbf{r}_g \neq \mathbf{r}_g^*$.

In the new coordinate system, the gear tooth flank, $\mathcal{G}$, is analytically described by the following expression:

$$\mathbf{r}_g^*(U_g^*, V_g^*) = \mathbf{Rs}(1 \to 2) \cdot \mathbf{r}_g(U_g, V_g) \tag{B.67}$$

The operator of the resultant coordinate system transformation $\mathbf{Rs}(1 \to 2)$ casts the column matrices of variables in Eqs. (B.65) and (B.66) to the form:

$$\left[\, dU_g^* \quad dV_g^* \quad 0 \quad 0 \,\right]^T = \mathbf{Rs}(1 \to 2) \cdot \left[\, dU_g \quad dV_g \quad 0 \quad 0 \,\right]^T. \tag{B.68}$$

Substitution of Eq. (B.68) into Eq. (B.65) and Eq. (B.66) makes it possible the expressions for the fundamental forms, $\Phi_{1.g}^*$ and $\Phi_{2.g}^*$, in the new coordinate system:

$$\begin{aligned} \left[\Phi_{1.g}^*\right] = {} & \left[\mathbf{Rs}(1 \to 2) \cdot \left[\, dU_g \quad dV_g \quad 0 \quad 0 \,\right]^T\right]^T \cdot \left[\Phi_{1.g}\right] \cdot \\ & \mathbf{Rs}(1 \to 2) \cdot \left[\, dU_g \quad dV_g \quad 0 \quad 0 \,\right]^T \end{aligned} \tag{B.69}$$

$$\begin{aligned} \left[\Phi_{2.g}^*\right] = {} & \left[\mathbf{Rs}(1 \to 2) \cdot \left[\, dU_g \quad dV_g \quad 0 \quad 0 \,\right]^T\right]^T \cdot \left[\Phi_{2.g}\right] \cdot \\ & \mathbf{Rs}(1 \to 2) \cdot \left[\, dU_g \quad dV_g \quad 0 \quad 0 \,\right]^T \end{aligned} \tag{B.70}$$

The following equation is valid for multiplication:

$$\left[\mathbf{Rs}(1 \to 2) \cdot \left[\, dU_g \quad dV_g \quad 0 \quad 0 \,\right]^T\right]^T = \mathbf{Rs}^T(1 \to 2) \cdot \left[\, dU_g \quad dV_g \quad 0 \quad 0 \,\right] \tag{B.71}$$

Therefore,

$$\begin{aligned} \left[\Phi_{1.g}^*\right] = {} & \left[\, dU_g \quad dV_g \quad 0 \quad 0 \,\right]^T \cdot \left\{\mathbf{Rs}^T(1 \to 2) \cdot \left[\Phi_{1.g}\right] \cdot \mathbf{Rs}(1 \to 2)\right\} \\ & \cdot \left[\, dU_g \quad dV_g \quad 0 \quad 0 \,\right] \end{aligned} \tag{B.72}$$

$$\left[\Phi^{*}_{2.g}\right] = \left[dU_g \quad dV_g \quad 0 \quad 0\right]^{T} \cdot \left\{\mathbf{Rs}^{T}(1 \to 2) \cdot \left[\Phi_{2.g}\right] \cdot \mathbf{Rs}(1 \to 2)\right\} \cdot \left[dU_g \quad dV_g \quad 0 \quad 0\right] \tag{B.73}$$

It can be easily shown that the matrices $\left[\Phi^{*}_{1.g}\right]$ and $\left[\Phi^{*}_{2.g}\right]$ in Eq. (B.72) and Eq. (B.73) represent quadratic forms with respect to dU_g and dV_g.

The operator of transformation $\mathbf{Rs}(1 \to 2)$ of the gear tooth flank, $\mathscr{G}$, having the first, $\Phi_{1.g}$, and the second, $\Phi_{2.g}$, fundamental forms from the initial coordinate system $X_1Y_1Z_1$ to the new coordinate system, $X_2Y_2Z_2$, results in that in the new coordinate system the corresponding fundamental forms are expressed in the form:

$$\left[\Phi^{*}_{1.g}\right] = \mathbf{Rs}^{T}(1 \to 2) \cdot \left[\Phi_{1.g}\right] \cdot \mathbf{Rs}(1 \to 2) \tag{B.74}$$

$$\left[\Phi^{*}_{2.g}\right] = \mathbf{Rs}^{T}(1 \to 2) \cdot \left[\Phi_{2.g}\right] \cdot \mathbf{Rs}(1 \to 2) \tag{B.75}$$

Equations (B.74) and (B.75) reveal that after the coordinate system transformation is completed, the first $\Phi^{*}_{1.g}$ and the second $\Phi^{*}_{2.g}$ fundamental forms of the gear tooth flank, $\mathscr{G}$, in the coordinate system $X_2Y_2Z_2$ are expressed in terms of the first, $\Phi_{1.g}$, and the second, $\Phi_{2.g}$, fundamental forms initially represented in the coordinate system $X_1Y_1Z_1$. In order to do that, the corresponding fundamental form (either the form, $\Phi_{1.g}$, or the form, $\Phi_{2.g}$) must be pre-multiplied by $\mathbf{Rs}(1 \to 2)$ and after that, it been post-multiplied by $\mathbf{Rs}^{T}(1 \to 2)$.

Implementation of Eqs. (B.74) and (B.75) significantly simplifies formulae transformations.

Equations similar to those above Eqs. (B.74) and (B.75) are valid with respect to pinion tooth flank, $\mathscr{P}$.

In case of use of the third, $\Phi_{3.g}$, and of the fourth, $\Phi_{4.g}$, fundamental forms, their coefficients can be expressed in terms of the fundamental magnitudes of the first and of the second order.

-2 ; 1 " <<<

3 Planar Linkage Mechanisms

In this chapter, we discuss aspects of the movements of linkage mechanisms embedded in the plane, permitting links to pass through one another, without obstacles. Our goal is to explain Kempe's universality theorem 3.2.1.

3.1 STRAIGHT-LINE LINKAGES

In preparation for presenting this universality result, we discuss the fascinating history of the design of linkages to draw straight lines, which led directly to Kempe's work.

3.1.1 Degrees of Freedom

Linkages that have a point that follows a particular curve have 1 degree of freedom, that is, the configuration space is one-dimensional. We consider two types of joints: pinned and unpinned or free. Both are universal joints, but a pinned joint is fixed to a particular point on the plane. Let a linkage have j free joints and r rods (or links). Pinned joints add no freedom, a free joint adds 2 degrees of freedom, and a rod removes 1 degree of freedom if it is not redundant (Graver 2001, p. 135). Thus, if all rods are "nonredundant" (a common situation), a linkage has $2j - r$ degrees of freedom. Thus, to follow a curve, a linkage should have $r = 2j - 1$ rods; in particular, it must have an odd number of rods.

3.1.2 Watt Parallel Motion

It was a question of considerable practical importance to design a linkage so that one point moves along a straight line, a motion needed in many machines, for example, to drive the piston rod of a steam engine. James Watt invented a simple linkage in 1784 that nearly, but not quite, achieves this.[1] Watt's "parallel motion" linkage (something of a misnomer that has stuck) is illustrated in Figure 3.1. Two pinned joints x and y are centers of equal-radii circles traced by joints a and b respectively. Here $j = 2$ and $r = 3$, so it has $2j - r = 1$ degree of freedom. The midpoint c of ab moves approximately on a vertical straight line. Joint a is pulled leftward and b rightward on their circles, with

---

[1] We rely here primarily on Kempe (1877).

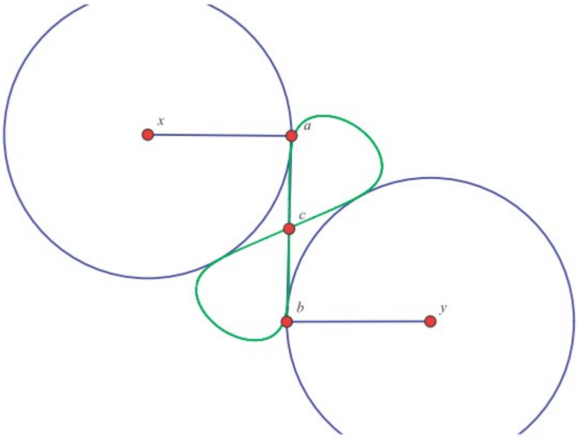

Figure 3.1. A Watt linkage. Joints x and y are pinned; joints a and b are free. The locus of point c is the figure-8 curve. [Construction from Cinderella.]

c staying approximately balanced between. Point c actually follows a "lemniscate," a figure-8 shape with one long, nearly straight section. Variations on this theme are now called "Watt linkages." Watt wrote, "I am more proud of the parallel motion than of any other mechanical invention I have ever made."[2]

3.1.3 Peaucellier Linkage

It was 80 more years before exact straight-line motion was achieved in 1864, by Charles-Nicolas Peaucellier, a captain in the French army. He received a prize, the Prix Montyon, from the Institute of France for his discovery.[3] His remarkable linkage is illustrated in Figure 3.2. Joints x and y are pinned, and rods xa and xb keep joints a and b on circle C_x and rod yc keeps joint c on circle C_y, which also includes x. There are $j = 4$ free joints and $r = 7$ rods, so the linkage has $2j - r = 1$ degree of freedom. The rhombus $abcd$ expands and narrows symmetrically about the diagonal ab. As the cell $abcd$ is rotated clockwise, the gap between the circles C_x and C_y narrows, squeezing the cell, remarkably just enough so that joint d stays on a vertical line L. It is clear that x and the tips c and d of the rhombus are collinear. It can be shown that the product of the distances from x to c and to d is a constant, and that this implies that d follows a straight line.

When Lord Kelvin worked a model of the device, he is reputed to have remarked that "it was the most beautiful thing he had ever seen" (Fergusson 1962).

Peaucellier's linkage was discussed by J. J. Sylvester in a lecture in 1874, and soon improved by Hart to a linkage of only five links, as we will describe in Section 3.3. It was

---

[2] Hunt (1978, p. 134), quoting Reuleaux from 1875.

[3] The history is tangled. A student of Chebyshev, Lippman Lipkin, independently discovered the same mechanism in 1871 (contra: Coxeter and Greitzer 1967, p. 109), as Peaucellier acknowledged in 1873 (Cabillon 1999). Peaucellier's result was first published in 1864 (Peaucellier 1864). See also Koetsier (1999) and Kempe (1877, p. 12).

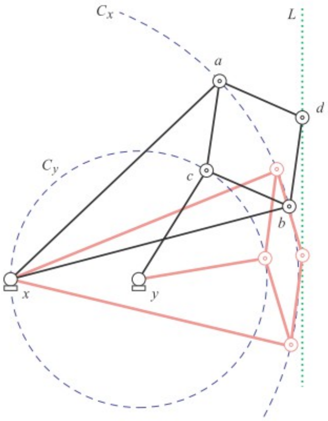

Figure 3.2. Peaucellier linkage. The dark lines show it in one position, the light lines in another.

this lecture that initiated Alfred Bray Kempe[4] to the intricacies of linkages and inspired his work toward his universality theorem.[5]

3.2 KEMPE'S UNIVERSALITY THEOREM

An attractive version of Kempe's result was formulated by Thurston in the 1970s (King 1999, Cor. 2.1):

> There is a linkage that signs your name.

As mentioned, Kempe's proof of Theorem 3.2.1 was flawed, and the subsequent history rather complex. We will not detail this history (see King 1999 or Connelly and Demaine 2004), but content ourselves by mentioning that the first complete, detailed proof is generally acknowledged to be by Kapovich and Millson (2002). In King (1999), the universality theorem is credited to Thurston and Kapovich-Millson. Because Kempe is acknowledged by all to be the source of the main ideas, we will continue to call it "Kempe's universality theorem":

Theorem 3.2.1 (Kempe 1876). *Let $\mathcal{C}$ be a bounded portion of an algebraic curve in the plane, that is, the intersection of zero-set of a real-coefficients polynomial $\phi(x, y) = 0$ with a closed disk. Then there exists a planar linkage such that the orbit of one joint is precisely $\mathcal{C}$.*

We now describe Kempe's proof, tracking his own presentation (Kempe 1876), followed by a discussion of the flaw and one route to repair.

---

[4] Kempe was a barrister by profession, but an excellent amateur mathematician. He is best known for his false proof of the four-color theorem in 1879, which stood for 11 years before Heawood found a flaw. Several ideas in his "proof" were, however, valid, and survive in the form of "Kempe chains" in the computer-aided proof of Appel and Haken a century later (Appel and Haken 1977). The story with linkages is similar: His proof was flawed, but the overall conception was brilliant.

[5] There has since been an extension to 3D (Zindler 1931).

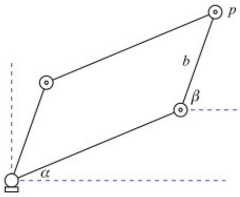

Figure 3.3. Linkage parallelogram.

3.2.1 Kempe's Proof

Let $\phi(x, y) = 0$ be the algebraic curve to be followed by some joint p. The construction begins by placing p at the apex of a parallelogram as shown in Figure 3.3. This parallelogram constrains the range of p's motion, but if we choose a and b equal and large, p is free to move anywhere within distance $2a$ of the pinned vertex O. The feature of this construction is that, if we view O as the origin $(0, 0)$, then the coordinates of $p = (x, y)$ may be expressed in terms of cosines of angles α and β:

$$x = a\cos\alpha + b\cos\beta,$$
$$y = a\cos(\alpha - \pi/2) + b\cos(\beta - \pi/2).$$

(This linkage has $j = 3$ and $r = 4$ and so $2j - r = 2$ degrees of freedom.) Substituting into $\phi(x, y)$ and repeatedly applying the trigonometric product-to-sum identity

$$\cos A \cos B = \frac{1}{2}\cos(A + B) + \frac{1}{2}\cos(A - B) \tag{3.1}$$

converts powers and products of x and y into cosines of sums of angles, resulting in the following general form for the polynomial $\phi(x, y)$:

$$\phi(x, y) = c + \sum_i c_i \cos(r_i\alpha + s_i\beta + \delta_i), \tag{3.2}$$

where c and c_i are constants, r_i and s_i are integers, and $\delta_i \in \{0, \pm\pi/2\}$. For example, the term xy^2 expands and reduces to

$$\frac{1}{4}\big[a^3\cos\alpha + 2ab^2\cos\alpha - a^3\cos(3\alpha) + ab^2\cos(\alpha - 2\beta) + a^2b\cos(2\alpha - \beta)$$
$$+ 2a^2b\cos\beta + b^3\cos\beta - b^3\cos(3\beta) - 3a^2b\cos(2\alpha + \beta) - 3ab^2\cos(\alpha + 2\beta)\big].$$

Each term of the sum in Equation (3.2) is achieved by a link of length c_i, at a suitable angle, with the whole shifted by c. Thus the problem reduces to constructing an angle of the form $r_i\alpha + s_i\beta + \delta_i$ from α and β.

Kempe shows that such a construction can be accomplished with three "gadgets": a "translator" for translating a motion, an "additor" for adding any two angles, and a "multiplicator" for multiplying an angle by a positive integer. The essence of the translator is the action of a pantograph, touched on earlier in Section 1.2.2. Note that while a multiplicator could naturally be implemented using a collection of additors (adding an angle repeatedly to itself), in fact the additor construction uses a multiplicator, and thus both are necessary. Hence, we describe these gadgets in reverse order.

Multiplicator. First, a *contraparallelogram* is a four-bar linkage formed by flipping two adjacent sides of a parallelogram across the diagonal they determine (see Figure 3.4).

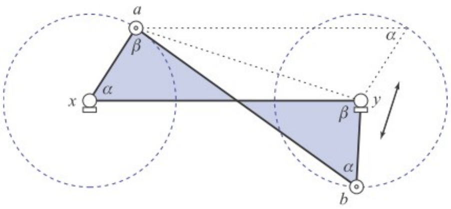

Figure 3.4. A contraparallelogram. $|xa| = |yb|$ and $|xy| = |ab|$. (Note: There is no joint at the point at which the two rods xy and ab cross.)

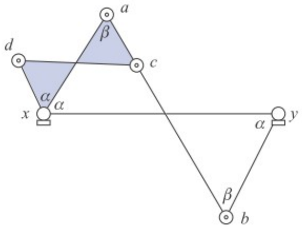

Figure 3.5. Kempe's "multiplicator": two nested, similar contraparallelograms.

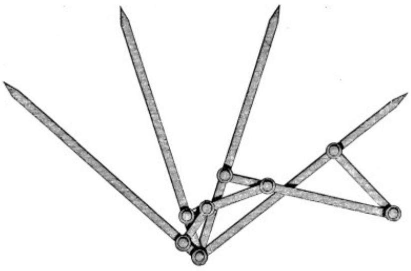

Figure 3.6. Kempe's angle trisector (Kempe 1877, p. 42).

(This structure plays a key role in Hart's linkage, as we will see in Section 3.3.) The opposite angles α and β are equal as in a parallelogram; the two shaded triangles are congruent. The linkage has 1 degree of freedom with x and y pinned[6] and so fixing either one of α or β determines the other. Kempe cleverly joined two similar contraparallelograms as shown in Figure 3.5 so that the angle β at joint a is shared. The two have their lengths in proportion: $|xy|/|xa| = |xa|/|xd|$. Because fixing β determines α in a contraparallelogram, and because of their similarity, the angle α is duplicated at x in the smaller, shaded contraparallelogram, effectively multiplying the angle α by 2. This is why Kempe called this linkage a "multiplicator." Its dynamic action is quite pleasing.[7]

This same process may be repeated, as illustrated in Figure 3.6 for $r_i = 3$, to multiply an angle by a positive integer r_i. Kempe notes that this linkage of eight bars exactly

---

[6] In this case, the xy rod is redundant. So $j = 2$ and $r = 3$, and $2j - r = 1$.

[7] See http://www.gfalop.org/I.

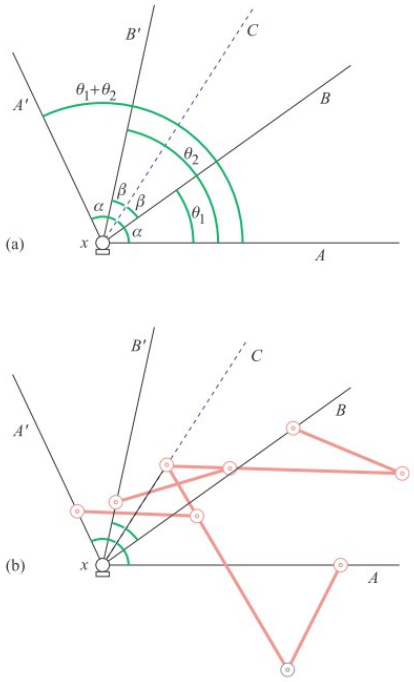

Figure 3.7. (a) A reflects to A' about C, as does B to B'; (b) the construction with superimposed reversors.

trisects an angle, a construction not possible by straight edge and compass. We will return to this point in Part II, Chapter 19 (p. 285).

Additor. Kempe's "additor" adds two angles together. This relies on the multiplicator, but only in its times-2 version in Figure 3.5, which he also called a "reversor" because it can be viewed as reflecting the angle α about the line xa. The construction for adding two angles θ_1 and θ_2 is illustrated in Figure 3.7. It consists of five links, all joined at x, as illustrated in Figure 3.7(a). A makes an angle α with C and reflects across C to A'. Similarly, B makes an angle β with C and reflects across C to B'. If these reflections are maintained, then A and A' are placed symmetrically α around C and B and B' are placed symmetrically β around C. Then the angle $\theta_1 = \alpha - \beta$ and $\theta_2 = \alpha + \beta$ so that $\theta_1 + \theta_2 = 2\alpha$, which is the angle between A and A'. Thus, with A the "base" link, we can "input" the angle θ_1 with link B and θ_2 with link B', and then A' reads off the "output" $\theta_1 + \theta_2$.

The reflections about C can be maintained by Kempe's "reversor" in Figure 3.5. This requires erecting an appropriate contraparallelogram that maintains the reflected angle between A and C, and another to maintain the reflection of B about C. These extra links and joints are shown shaded in Figure 3.7(b).

Translator. Kempe's simplest gadget is the translator, whose purpose is to copy a link at its same orientation elsewhere. His design is shown in Figure 3.8(a). Link A, making an angle α with the horizontal, is displaced to the parallel link C. It makes sense to choose $|D| = |E|$ so that C is free to be placed anywhere within a disk centered on x,

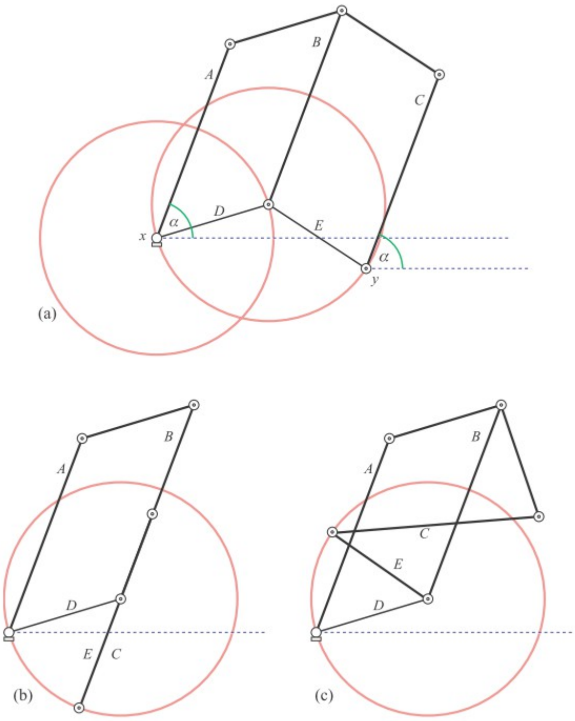

Figure 3.8. (a) Kempe's translator; (b) alignment of C with B; (c) second parallelogram flips to a contraparallelogram. [See also Figure 2.3.2 in Hopcroft et al. 1984.]

that is, the annulus of reachability for the point y (according to Theorem 5.1.2) is a disk. There is, however, a flaw in this design, to which we will return below.

Overall design. Returning to Equation (3.2), the "multiplicator" permits multiplying any angle by an integer r_i. The "additor" permits adding angles, and the "translator" permits connecting together the terms of the sum. Combining these gadgets in an appropriate arrangement yields a joint p' that stays at a horizontal distance $\phi(x, y) - c$ from the y-axis. Using a Peaucellier linkage to force p' to stay on the vertical line $x + c = 0$ in turn forces p to stay on the curve $\phi(x, y) = 0$. Thus, this complex linkage of 1 degree of freedom has the joint p follow the equation $\phi(x, y) = 0$ within the range of x values provided by the parallelogram linkage in Figure 3.3 (see Figure 3.9).

3.2.2 The Flaw and Repairs

Concerning the translator in Figure 3.8, Kempe (1876, p. 214) says: "it is clear that the link [C] is equal and parallel to [A], but is otherwise free to assume any position." What he missed is that when link C aligns with B ((b) of the figure), the second parallelogram is free to flip to a contraparallelogram, as in (c), in which case C is no longer parallel to

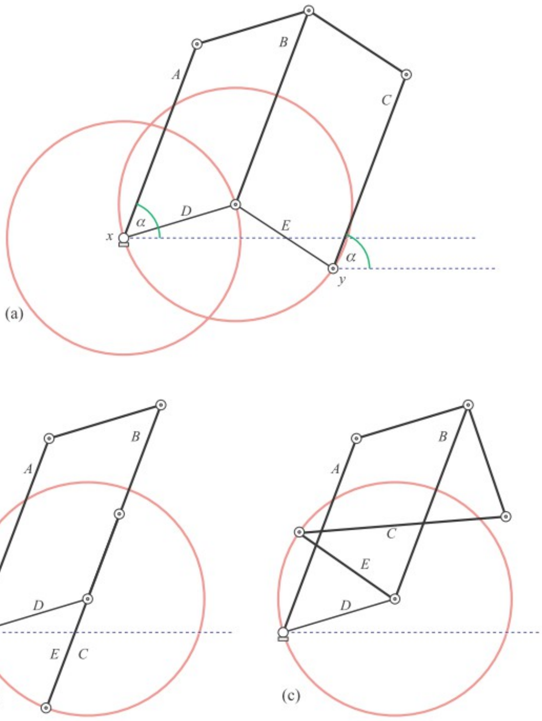

Figure 3.8. (a) Kempe's translator; (b) alignment of C with B; (c) second parallelogram flips to a contraparallelogram. [See also Figure 2.3.2 in Hopcroft et al. 1984.]

that is, the annulus of reachability for the point y (according to Theorem 5.1.2) is a disk. There is, however, a flaw in this design, to which we will return below.

Overall design. Returning to Equation (3.2), the "multiplicator" permits multiplying any angle by an integer r_i. The "additor" permits adding angles, and the "translator" permits connecting together the terms of the sum. Combining these gadgets in an appropriate arrangement yields a joint p' that stays at a horizontal distance $\phi(x, y) - c$ from the y-axis. Using a Peaucellier linkage to force p' to stay on the vertical line $x + c = 0$ in turn forces p to stay on the curve $\phi(x, y) = 0$. Thus, this complex linkage of 1 degree of freedom has the joint p follow the equation $\phi(x, y) = 0$ within the range of x values provided by the parallelogram linkage in Figure 3.3 (see Figure 3.9).

3.2.2 The Flaw and Repairs

Concerning the translator in Figure 3.8, Kempe (1876, p. 214) says: "it is clear that the link [C] is equal and parallel to [A], but is otherwise free to assume any position." What he missed is that when link C aligns with B ((b) of the figure), the second parallelogram is free to flip to a contraparallelogram, as in (c), in which case C is no longer parallel to

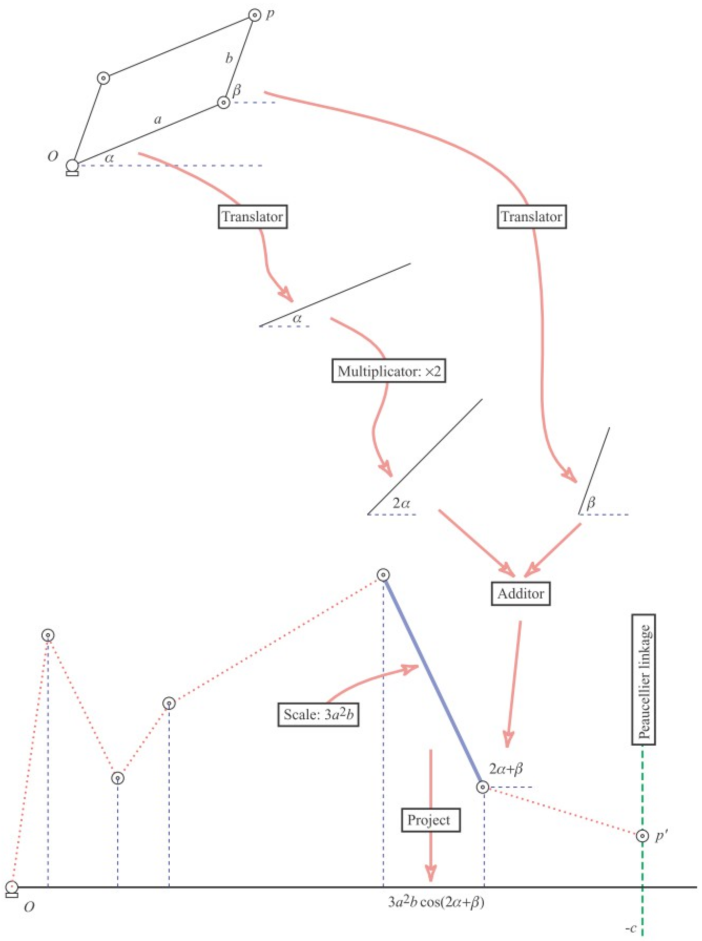
p
b
β
a
O
α
Translator
Translator
α
Multiplicator: ×2
2α
β
Additor
Peaucellier linkage
Scale: $3a^2b$
$2\alpha+\beta$
Project
p′
O
$3a^2b\cos(2\alpha+\beta)$
-c

Figure 3.9. A schematic of the overall structure of Kempe's construction. Each term of Equation (3.2) is simulated by a link whose projection on the horizontal axis is that term. The point p' is constrained to follow a particular vertical line by a Peaucellier linkage. The joints labeled O are the same joints.

A! This unwanted passage through a degeneracy to another part of the configuration space is a generic problem throughout Kempe's construction, noted by Thurston, and in Hopcroft et al. (1984), and perhaps by others. This is how the problem is phrased by Kapovich and Millson (2002):

> The main problem with Kempe's proof is that it works well only for a certain subset of the moduli space,[8] however, near certain "degenerate" configurations the moduli

---

[8] Another term for the configuration space of the linkage.

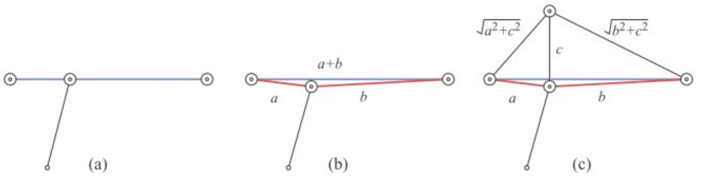

Figure 3.10. Simulating a joint in the interior of a link.

space splits into several components and the linkage fails to describe the desired polynomial function. We use the "rigidified parallelograms" to resolve this problem and get rid of undesirable configurations (p. 3).

To repair Kempe's proof, we need to "brace" both the parallelogram and the contraparallelogram to prevent flipping from one to the other. For the first bracing, we follow King's presentation (King 1999), of Kapovich and Millson's resolution. The latter authors did not need to brace the contraparallelogram, as their proof deviates from Kempe's and does not employ that mechanism. For the second bracing, we follow a recent idea of Abbott and Barton (2004).

3.2.2.1 Bracing the Parallelogram

It is convenient to be able to locate a joint in the interior of a rigid bar. One could either permit this as part of the definition of a linkage or retain the restriction that joints are only at endpoints of bars and simulate an interior joint as follows. Refer to Figure 3.10. The joint in the interior of edge *e* in (a) of the figure can be simulated with links whose length sums the length of *e*, as shown in (b). This linkage is not "infinitesimally rigid" (as defined in Section 4.4, p. 49), so one might want to add the "stiffening truss" shown in (c).

Now we show how to "rigidify" parallelograms. Consider a rhombus linkage (see Figure 3.11). With its two bottom vertices a and b pinned, it has the natural circular motion shown in (a). However, as Kapovich and Millson pointed out, it also has the degenerate motions shown in (b) and (c). So the configuration space of this linkage is three circles, each pair intersecting in a single point. Blocking of the degenerate motions is accomplished by adding a bar between the midpoints of opposite sides, as in (d). These extra links are needed at many points throughout Kempe's construction, for example, in the parallelogram that is part of the Peaucellier linkage (Figure 3.2) employed in Kempe's construction.

3.2.2.2 Bracing the Contraparallelogram

Bracing the contraparallelogram is a bit more complicated. Here we describe a bracing due to Abbott and Barton (2004).

Let the contraparallelogram be $abcd$, and let p, q, r, and s be the midpoints of its links. The linkage is braced by four extra bars to a common point x on the perpendicular bisector of the line containing those midpoints (see Figure 3.12(a)). As long as these four extra bars are chosen long enough, the contraparallelogram is prevented from flipping to the parallelogram configuration. The reason why is hinted at in Figure 3.12(b). When $abcd$ is a parallelogram, so is $pqrs$. Now, returning to (a) of the figure, we know that x is on the perpendicular bisector of sq and on the perpendicular bisector of pr. This

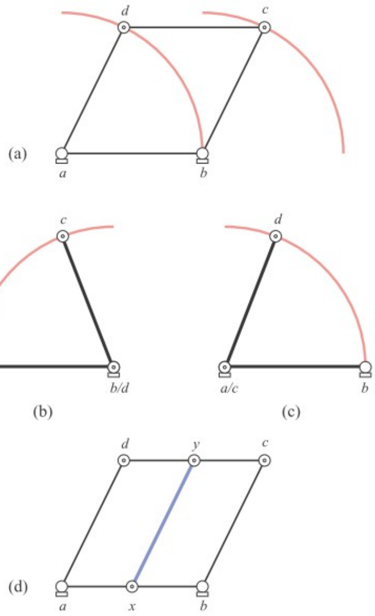

Figure 3.11. (a–c) Three circular motions of a rhombus; (d) rhombus rigidified.

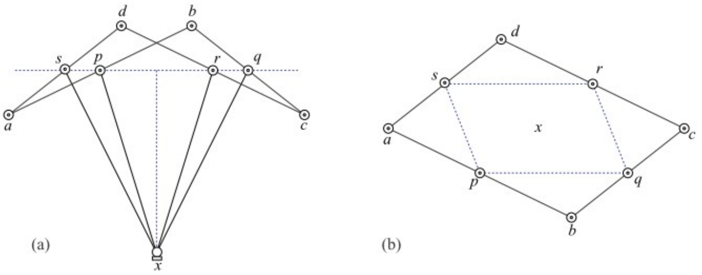

Figure 3.12. (a) Braced contraparallelogram; (b) flipping to the corresponding parallelogram is not possible.

would place x at the center of the parallelogram $pqrs$. But if the extra bars are chosen long enough, this is impossible. See Box 3.1 for details.

3.2.3 Recap

We have chosen to present Kempe's original proof with repairs, but as mentioned, the first full proof is due to Kapovich and Millson. They follow Kempe's construction in spirit but not detail. Their proof employs modified versions of many of his linkages, but avoids

Box 3.1: Braced Contraparallelogram

Let $r_1 = |xp| = |xr|$ and $r_2 = |xs| = |xq|$ be the two lengths of the bracing bars. These lengths are chosen so that $r_2^2 - r_1^2 = \frac{1}{4}(|ab|^2 - |ad|^2)$. A few facts:

1. p, q, r, and s are collinear.
2. $|sp| = |rq|$.
3. x is on the perpendicular bisector of sq and on the perpendicular bisector of pr.
4. $|sp| \cdot |sr| = \frac{1}{4}(|ab|^2 - |ad|^2)$.

This last claim is not obvious, but we defer a proof to the discussion of Hart's inversor (Box 3.2). Let L be the linkage composed of just the closed 4-chain $abcd$, and let L_b be the braced version. Using the facts above, Abbott and Barton (2004) prove that any degenerate (flat) or contraparallelogram configuration of L can be extended to a configuration of L_b, and, conversely, the bracing of L_b restricts the configurations of $L \subset L_b$ to be just the degenerate and contraparallelogram configurations.

his multiplicator in favor of a more complex multiplier mechanism (incorporating three translators, a pantograph, three inversors (modified Peaucellier linkages), and an adder) whose properties are more easily controlled. The collection of elementary linkages allow a nearly straightforward composition of a linkage for complex polynomial maps, $f : \mathbb{C}^n \to \mathbb{C}$, without recourse to the trigonometric expansions employed by Kempe. This complex linkage is then modified further to obtain a linkage realizing a real polynomial map $f : \mathbb{R}^n \to \mathbb{R}$, achieving a version of the universality theorem (Theorem 2.2.3): for any compact real algebraic set X (e.g., a bounded portion of a curve), there is a linkage L such that the map is an "analytically trivial covering"[9] of X. Thus, the formidable full proof finally firmly establishes that, indeed, there is a linkage that signs your name, even if (as King 1999, notes) your signature includes cusps and multiple strokes! And, as previously noted, this beautiful generalization of the task of creating straight-line mechanisms has led to a thorough understanding of the rich topology of the configuration space of linkages.

It is natural to ask how many links are required to trace a curve defined by a polynomial of degree n. This was first addressed in Gao et al. (2002), where an $O(n^4)$ bound was obtained on the number of links. This was subsequently improved by Abbott and Barton (2004) to $O(n^2)$, and extended to arbitrary dimension d to $O(n^d)$ bars. They also offer strong evidence that these bounds are optimal.

Two interesting open questions were recently posed on this topic:

Open Problem 3.1: Continuous Kempe Motion.[a] Given a (connected) curve C in the plane, is there is a linkage that traces exactly that curve via a continuous motion/reconfiguration?

[a] Abbott and Barton (2004).

---

[9] The map is an *analytic isomorphism*, with both f and f^{-1} restrictions of analytic maps (i.e., locally, power series).

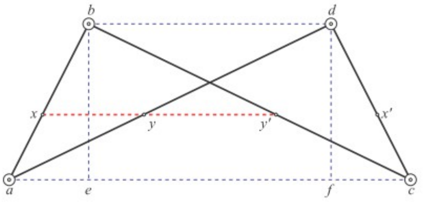

Figure 3.13. Hart's inversor. $|ab| = |cd|$ and $|bc| = |da|$, so that $abcd$ is a contraparallelogram.

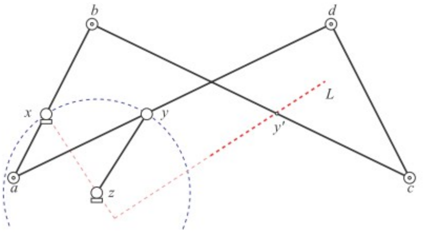

Figure 3.14. Hart's mechanism to convert circular motion of y to linear motion of y'. Joint x is pinned and the center of inversion. L is perpendicular to the line containing xz.

The point here is that although the proofs construct a linkage such that the positions of one joint are exactly C, it is possible that the whole of C cannot be traced by a continuous reconfiguration of the linkage. Perhaps it is necessary to jump to a different configuration to trace different sections of C.

The next open problem ties into linkages forbidden from self-intersection, a topic discussed at great length later on in Part I:

Open Problem 3.2: Noncrossing Linkage to Sign Your Name.[a] Is there a linkage that traces any given algebraic curve, and moves without any self-intersection throughout the tracing? If not, what is the class of curves that can be traced by a noncrossing linkage?

---

[a] Don Shimamoto, April 2004.

3.3 HART'S INVERSOR

We complete the story of linear-motion linkages by sketching Hart's 1875 5-link mechanism,[10] which "improves" upon Peaucellier's 7-link arrangement at the cost of a loss of some symmetry and the use of one more complex joint.

Start with a contraparallelogram $abcd$ as in Figure 3.13. The symmetry of the contraparallelogram guarantees that ac and bd remain parallel during movement. Hart takes four points x, y, y', x' dividing the four links in the same proportion. It is then not hard to see that these four points are collinear, and remain collinear and parallel to ac and bd throughout the free motion of the linkage. The key property – not at all obvious – is

---

[10] Hart (1875), as described in Kempe (1877, p. 18ff). See also Hunt (1978, p. 15) and http://poncelet.math.nthu.edu.tw/disk3/cabrijava/hart2.html.

21 Introduction and Overview

21.1 OVERVIEW

In 1525 the painter and printmaker Albrecht Dürer[1] published a book, *Underweysung der Messung* (later translated as "The Painter's Manual"), in which he explained the methods of perspective, which he had just learned himself on a trip to Italy[2] (see Figure 21.1).

Dürer's book includes a description of many polyhedra, which he presented as surface unfoldings, are now called "nets." For example, Figure 21.2 shows his net for a cuboctahedron. Even when he drew a more complex net, such as that for a truncated icosahedron (the shape of a soccer ball; Figure 21.3), he always chose an unfolding that avoided self-overlap. Although there is no evidence that Dürer distilled this property into a precise question, it is at least implicit in the practice of subsequent generations that every convex polyhedron may be so unfolded.[3] More precisely, define an *edge unfolding* as a development of the surface of a polyhedron to a plane, such that the surface becomes a flat polygon bounded by segments that derive from edges of the polyhedron. One may view such an unfolding as obtained by slicing the surface along a collection of edges. We would like an unfolding to possess three characteristics, enjoyed by all Dürer's drawings:

1. The unfolding is a single, simply connected piece.
2. The boundary of the unfolding is composed of (whole) edges of the polyhedron, that is, the unfolding is a union of polyhedron faces.
3. The unfolding does not self-overlap, that is, it is a "simple polygon."

We call a simple polygon that satisfies these conditions a *net* for the polyhedron.[4] This leads to the following open problem:

---

[1] 1471–1528, Nürnberg, Germany.

[2] Here we rely on George Hart's history at http://www.georgehart.com/virtual-polyhedra/durer.html.

[3] This point was first made to us by Branko Grünbaum (personal communication, 1987).

[4] Here we follow Shephard's (1975) and Grünbaum's usage (2001) in insisting that a net is derived from an edge unfolding. We will also explore polygons that fold to a polyhedron but are not a union of polyhedron faces. (Alexandrov (2005, p. 61) uses what is translated to English as "net" for another concept. What is translated to German as "das Netz" in Alexandrov (1958) is translated to English as "development" in Alexandrov (2005).) Dolbilin calls a net a "regular edgewise development" (Dolbilin 1998).

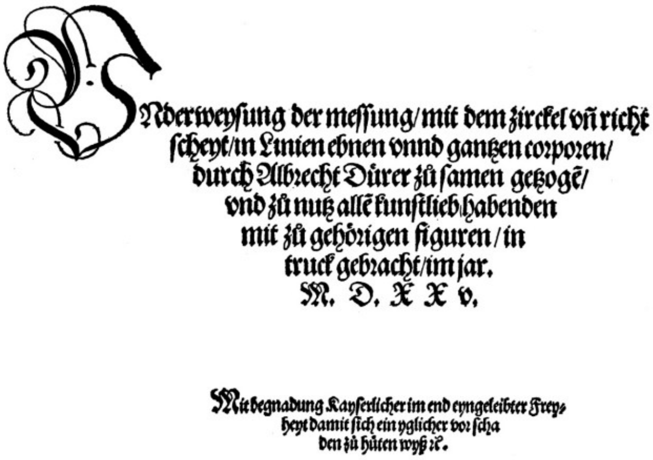

Underweyſung der meſſung/ mit dem zirckel vñ richtſcheyt/ in Linien ebnen vnnd gantzen coꝛpoꝛen/
durch Albꝛecht Dürer zů ſamen getzogē/
vnd zů nutz allē kunſtliebhabenden
mit zů gehöꝛigen figuren/ in
truck gebꝛacht/ im jar.
M. D. X X v.

Mit begnadung Kayſerlicher im end eyngeleibter Freyheyt damit ſich ein yglicher voꝛ ſchaden zů hüten wyß ꝛc.

Figure 21.1. The title page of Dürer's book: "Underweysung der messung . . . im jar MDXXV."

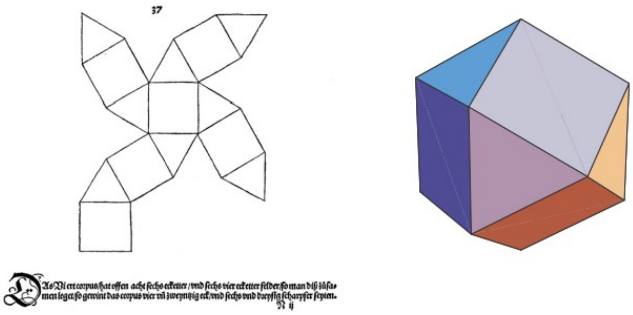

DAs Vi ert corpus/ hat offen achte ſechs ecketter/ vnd ſechs vier ecketter felder/ ſo man diß zůſamen leget/ ſo gewint das corpus vier vñ zweyntzig eck/ vnd ſechs vnd dreyſſig ſcharpffer ſeyten.
N ij

Figure 21.2. Dürer's net (left) for a cuboctahedron (right).

Open Problem 21.1: Edge-Unfolding Convex Polyhedra.[a] Does every convex polyhedron have an edge unfolding to a simple, nonoverlapping polygon, i.e., does every convex polyhedron have a net?

---

[a] Dürer (1525).

This has remained unsolved over the 475 years since Dürer's drawings. (The first explicit statement of the problem of which we are aware was formulated by Shephard in 1975 (Shephard 1975): "Does every 3-polytope possess a net?") This problem can be viewed as a central motivation for the third part of this book.

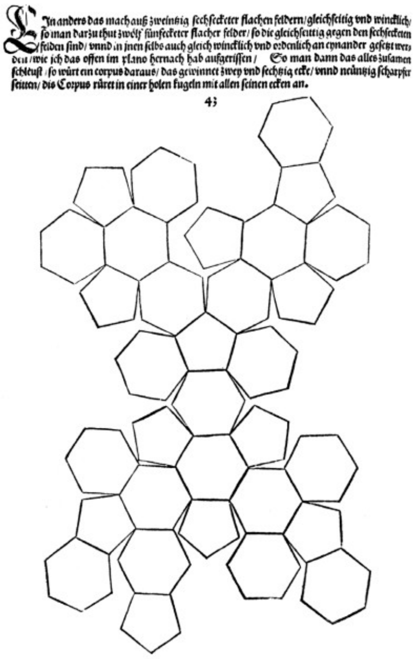

Ein anders das mach auß zweintzig ſechſeckter flachen feldern/gleichſeitig vnd wincklich/ ſo man darzu thut zwölff fünfeckter flacher felder/ ſo die gleichſeittig gegen den ſechſeckten felden ſind/ vnnd in jnen ſelbs auch gleich wincklich vnd ordenlich an eynander geſetzt werden/wie ich das offen im plano hernach hab aufgeriſſen/ So man dann das alles zuſamen ſchleuſt/ſo würt ein corpus daraus/ das gewinnet zwey vnd ſechtzig ecke/ vnnd neüntzig ſcharpffer ſeitten/ dis Corpus rüret in einer holen kugeln mit allen ſeinen ecken an.

43

Figure 21.3. Dürer's net for a truncated icosahedron.

21.2 CURVATURE

The concept of "curvature" will play an important role in this part of the book. Although we will employ it primarily for polyhedra, its origins lie in the differential geometry of smooth curves and surfaces. After describing curvature in these contexts, the definition for polyhedral surfaces will be motivated.

21.2.1 Smooth Curves

Let $c(t)$ be a smooth[5] curve in the plane, having a unit-speed parametrization by t. The curvature of $c(t)$ at a point p measures how sharply the curve turns at p by the magnitude of the acceleration $|\ddot{c}|$ felt at p by a particle (or a car) traveling at unit speed along $c(t)$. This acceleration is the same[6] as the reciprocal of the radius of the *osculating circle* at p: the unique circle that matches at p both the velocity vector (the tangent) and the acceleration vector (which is normal to the curve), in both magnitude and direction (see Figure 21.4). If $c(t)$ is straight at p, the circle radius is effectively infinite and the curvature is zero.

---

[5] For our purposes, *smooth* means C^{∞}: possessing a continuous kth derivative for all $k = 1, 2, 3, \ldots$, although usually the assumption of C^2 suffices.

[6] The same under the assumption that $|\ddot{c}| = 1$.

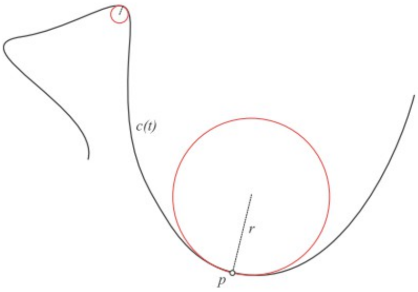

Figure 21.4. The curvature at a point is $1/r$, where r is the radius of the osculating circle.

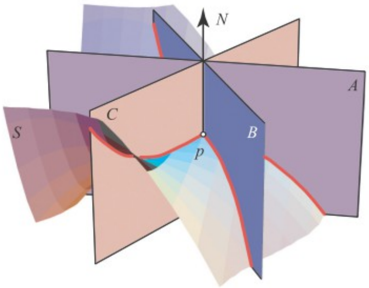

Figure 21.5. Spinning a plane about N through $p \in S$ yields the principal curvatures at p. The curvature of the curves for plane A and B are positive at p, but negative for plane C.

21.2.2 Smooth Surfaces

The curvature of a surface $S \subset \mathbb{R}^3$ at a point p cannot be captured by the osculating sphere, for this sphere does not distinguish between a spherical surface and a cylindrical surface, which intuitively have different curvatures (e.g., they have different flattening properties). Two numbers are needed at each point rather than one.

Principal curvatures. These two numbers are known as the *principal curvatures* and may be defined as follows (Lee 1997, p. 5). Let N be the unit normal vector to an oriented surface S at p, and P a plane through p containing N. Then $P \cap S$ is a one-dimensional curve whose curvature at p is defined as described above. It can be given a sign according to whether the osculating circle is locally inside S (+) or outside (−), with "outside" defined by the direction in which N points. Now imagine spinning P about N. The principal curvatures κ_1 and κ_2 at p are the minimum and maximum signed curvatures over all intersection curves obtained (see Figure 21.5). Notice that the principal curvatures do distinguish between a sphere and a cylinder: $\kappa_1 = \kappa_2 = 1/r$ for a sphere of radius r, but $\kappa_1 = 0$ and $\kappa_2 = 1/r$ for a cylinder of the same radius. There remains a flaw of sorts, however: the principal curvatures are not "intrinsic" properties of the surface, invariant under isometric embeddings. A flat piece of paper and the same paper rolled to a cylinder have the same intrinsic properties (to, say, flatlanders living in the surface), but have different principal curvatures.

Gaussian curvature. Gauss discovered in 1827 that the product $K = \kappa_1\kappa_2$ is intrinsic. He called this result the "*Theorema Egregium*" – roughly, the distinguished, extraordinary theorem – and said that it "ought to be counted among the most elegant in the theory of curved surfaces" (Spivak 1979, II: p. 114). K, known as the *Gaussian Curvature*, will be used frequently in the sequel. Note that both the flat paper and the cylinder have the same Gaussian curvature K: for the cylinder, $\kappa_1\kappa_2 = 0 \cdot 1 = 0$.

Curvature and normals. A useful extrinsic view of the Gaussian curvature is that "K measures the total spread of normal directions per unit surface area": it is "the ratio between the solid angle subtended by the normals on a small patch divided by the area" (Koenderink 1989, p. 215). Positive K spreads the normals, and negative K reverses the orientation of the spherical image of the normals.

Curvature and perimeter and area. Another equivalent – but this time intrinsic – definition of the Gaussian curvature at $p \in S$ is in terms of the perimeter or area of a small disk around p. Define a *geodesic disk* centered on p of radius r to be the locus of points of S at distance r of p, where distance to a point x is measured by the shortest path on S from p to x. (We will discuss shortest paths and geodesics in Chapter 24.) Such a disk is sometimes said to have "intrinsic radius" r. Let the circumference of the disk be $C(r)$. Bertrand and Puiseux proved in 1848 (Berger 2003, p. 110; Spivak 1979, II: p. 130) that

$$K = \lim_{r\to 0} \frac{3(2\pi r - C(r))}{\pi r^3}. \tag{21.1}$$

From this equation, we can see that points of positive curvature have a perimeter deficit with respect to the flat $2\pi r$ and those of negative curvature have a perimeter excess. Points of positive curvature are called *elliptic* and those of negative curvature *hyperbolic*.

Integrating around the perimeter of a geodesic disk leads to a similar equation for K in terms of the area deficit, discovered by Diquet:

$$K = \lim_{r\to 0} \frac{12(\pi r^2 - A(r))}{\pi r^4}, \tag{21.2}$$

where $A(p)$ is the area of the disk. Again we have the same classification: area deficit corresponds to positive curvature, and area excess to negative curvature.

21.2.3 Polyhedral Surfaces: Angle Deficit

The area-deficit viewpoint of Gaussian curvature makes the definition of Gaussian curvature on a polyhedral surface S natural, or at least plausible: the *Gaussian curvature* (or simply "the curvature") of a point p on a polyhedral surface S is the *angle deficit*[7] at p: 2π minus the total face angle incident to p. Thus, a point in the interior of a face of S has curvature zero. Less obvious is that a point on the relative interior of a polyhedron edge e also has curvature zero, for it has π face angle incident from either side of e. All the curvature of a (closed) polyhedron is concentrated at the vertices: A vertex v has curvature 2π minus the sum of the face angles (the *total angle*) incident to v.

For example, each vertex of a cube has curvature $\pi/2 = 90^\circ = 360^\circ - 3(90^\circ)$; and each vertex of a dodecahedron has curvature $\pi/5 = 36^\circ = 360^\circ - 3(108^\circ)$

---

[7] Sometimes called the "angle defect" or "jump angle." This notion of angle deficit was introduced by Descartes in his "Treatise on Polyhedra," which he wrote in the mid-1600s, a century prior to Gauss.

If the surface is cut open and flattened in the neighborhood of v, there will be an angle gap equal to the curvature at v. It is this notation of curvature that will be used in the sequel.

21.3 GAUSS–BONNET THEOREM

If a person walks around the boundary of a simple polygon, by the time they revisit their starting point, they have turned or twisted in total (with negative turnings canceling positive turnings) exactly 2π. This is an easy consequence of the fact that the sum of the internal angles of a polygon of n vertices is $(n-2)\pi$. The sum of the complete angles at the n vertices is $n(2\pi)$, and so the sum of the exterior angles is $(n+2)\pi$. The turn at each vertex is the excess of the exterior angle there over π (e.g., no turn at all corresponds to an exterior angle of π). So the total turn is $(n+2)\pi - n\pi = 2\pi$. The same holds true for any Jordan curve in the plane, with the discrete sum of angles replaced by an integral of the angle turn at each point of the curve. Call this total angle turn around a curve or polygon τ. This turn τ is sometimes called the *geodesic curvature* of the curve, as it measures the deviation from a straight geodesic.

On an arbitrary (nonflat) surface, one can define a notion of the turn of a curve on the surface, which intuitively is the turn experienced by a person walking along the curve while vertically aligned with the surface normal. Under this notion of surface angle turn, a closed curve does not in general have an angle turn of 2π. For example, consider a spherical triangle on a sphere, with a corner on the equator, a corner at the north pole, and, after a 90° turn there, a third corner on the equator. The total turn of this triangle is $\tau = 3\pi/2$. The reason τ is less than 2π on the sphere is that each curve encloses some positive curvature. There is a tradeoff between turn and enclosed curvature, which is captured precisely in the *Gauss – Bonnet formula*, one of the jewels of differential geometry[8]:

$$\tau + \gamma = 2\pi, \tag{21.3}$$

where τ is the total turn around the boundary of the curve and γ is the integral of the Gaussian curvature enclosed in the "polygon" P:

$$\int_P K dA. \tag{21.4}$$

On the plane, $\gamma = 0$, so $\tau = 2\pi$. On positively curved surfaces, polygons turn less than 2π, and on negatively curved surfaces, they turn more than 2π.

Because the equator of (or any great circle on) a sphere is a geodesic, it has $\tau = 0$. A west–east walk around the equator implies via Equation (21.3) that the northern hemisphere must contain 2π of curvature, as must the southern hemisphere via an east–west walk. Thus the total curvature of a sphere is 4π. This result generalizes to any surface homeomorphic to a sphere (i.e., of genus zero), by partitioning it into regions and applying Equation (21.3) to each region:

Theorem 21.3.1. *The total sum of the Gaussian curvature over all points of a surface homeomorphic to a sphere is* 4π.

---

[8] See, e.g., Lee (1997, Ch. 9) and Morgan (1991, Ch. 8).

Several examples:

1. *Sphere.* The curvature at each point of a sphere of radius $r = 2$ is $1/r^2 = 1/4$, and the integral over the entire surface area of $4\pi r^2 = 16\pi$ is 4π.
2. *Cube and dodecahedron.* The curvature sum for a cube is $8(\pi/2) = 4\pi$, and for a dodecahedron is $20(\pi/5) = 4\pi$.
3. *Cylinder.* Curvature is concentrated at the points on the top and bottom rim of a cylinder. If it has radius $r = 1$, each rim point has curvature 1, which, integrated over the $2\pi r = 2\pi$ length of the top and bottom rims, is 4π.

The reach of this theorem is breathtaking, as it holds for any surface, convex or nonconvex, smooth or polyhedral,[9] and connects the intrinsic, local curvature to a global constant. Intuitively it says that a dent in a surface at one location must be compensated by an equal bump at another. The theorem generalizes to surface of genus g with a sum of $2\pi(2 - 2g)$; the number $\chi = 2 - 2g = V - E + F$ is known as the *Euler characteristic* of the manifold.

One immediate consequence that we will exploit in Section 24.4 is the generalization of our observation about the great circles on the sphere: a closed geodesic (which does not turn, and so has $\tau = 0$) must enclose $\gamma = 2\pi$ of curvature.

---

[9] This result for convex polyhedra was proved by Descartes, and is sometimes known as Descartes' theorem (Phillips 1999).

data 01.05.2020
D S T Q Q [✓ S] S

Transformations

Pág. 277(291)

B.5 Impact of the Coordinate System Transformations on Fundamentals Forms of the Surface.

obs: Veja livro todo!!!

E) Books_76_2020/
Geometric Folding Algorithms/
Geometric Folding Algorithms...pdf

obs: Livro todo!!!

Pág. 30(44)
3.1.3 Peaucellier Linkage

Pág. 32(45)
Fig. 3.2 ☆

Pág. 299(311) Part III Polyhedra

Pág. 301(315) 21.2 Curvature

Pág. 302(316) ☆
Figure 21.4 →

data 01.05.2020

The curvature at a point is 1/r, where r is the radius of the OSCULATING CIRCLE.

Figure 21.5

Spinning a plane about N through p ∈ S yields the (produz) principal curvatures at p.

The curvature of the curves for plane A and B are positive at p, but negative for plate C.

Pág. 287 (301)

19.2 Huzita's Axioms and Hatori's Addition

Box 19.1: Abe's trisection

Important: three parallel lines

Estudar!!! com Pappus ✡

Fig. 19.4 Abe's trisection analyzed.

Jandaia

Box 19.1: Abe's trisection

Let a, b, and c be three points at the left ends of the bottom three parallel lines: the bottom side of the square and the first two fold lines (see Figure 19.4). Point b is the midpoint of ac by construction. Now reflection through the A6 crease line L maps the three points to a', b', and c', with again b' the midpoint of $a'c'$. Because a' lies on the horizontal line through b and because b is the midpoint of the vertical line segment ac, length $a'c$ equals length $a'a$, so $\triangle aa'c$ is isosceles. Therefore its reflection through crease line L, $\triangle aa'c'$, is also isosceles. Thus, $\angle a'ac'$ is cleverly split in half: $\angle a'ab' = \angle b'ac' = \alpha$. Because $\triangle aa'c'$ is isosceles, its base angle $\angle aa'c' = 90^\circ - \alpha$. Because a' is a reflection of the corner of the square, it is a right angle, leaving $\angle xa'a = \alpha$, and therefore by reflection $\angle xaa' = \alpha$. Thus the angle θ at a is trisected into $\alpha + \alpha + \alpha$.

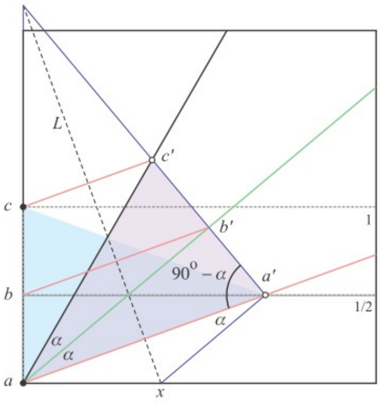

Figure 19.4. Abe's trisection analyzed.

constructions can construct rational numbers, the roots of quadratic equations with constructible coefficients, and nothing more. The key word here is "quadratic," not surprising because circles are quadratics. Axioms A1–A5 have been shown to be equivalent to straight edge and compass constructions (Alperin 2000; Geretschläger 1995; Justin 1989b). However, A6 cannot be accomplished (in general) using straight edge and compass; it adds the power to solve certain cubic equations. To see this, let us revisit A5, in Figure 19.5(a). A5 maps a point to a line over a crease. Because the point reflects across the crease, its distance to the crease is the same as the distance from the crease to the reflected point. This shows that the crease is tangent to the parabola whose focus is the point and whose directrix is the line, for a parabola is the locus of points equidistant from the focus and the directrix.

Now it should be clear that the crease in A6 is tangent to two parabolas, as shown in Figure 19.5(b). Finding a common tangent between two parabolas can be phrased as solving a cubic equation (Emert et al. 1994).[3]

---

[3] Alperin (2004) describes another geometric view of this axiom and its connections to "petals of conics."

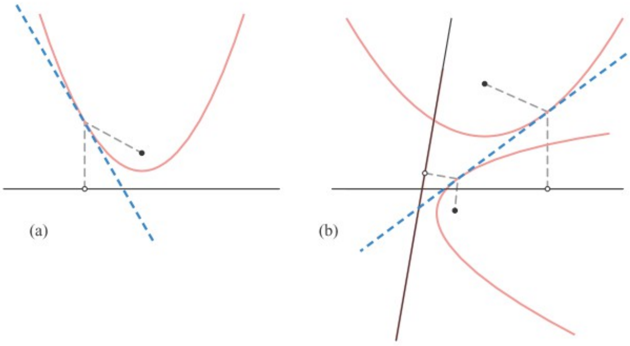

Figure 19.5. Huzita's axioms A5 and A6.

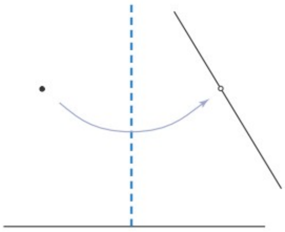

Figure 19.6. Hatori's axiom A7.

19.2.1 Hatori's Seventh Axiom

Recently Hatori suggested a seventh axiom (Hatori 2004; Hull 2003b):

A7. Given one point and two lines, one can fold a crease perpendicular to one line so that the point maps to the other line.

See Figure 19.6.

This axiom is again fundamentally quadratic, but not equivalent to any one of the others. Lang proved that these seven axioms together are complete in the sense that "these are all of the operations that define a single fold by alignment of combinations of points and finite line segments" (Lang 2004, p. 234).

19.3 CONSTRUCTIBLE NUMBERS

We mentioned that origami, as specified by the Huzita–Hatori axioms, can go beyond the quadratic limitation of straight edge and compass constructions, in having at least some cubic capabilities, leading to the ability to trisect an angle. It is then natural to ask if origami can do more than trisect. There is a characterization of what is achievable with a straight edge, compass, and a tool that only trisects a given angle: one can then

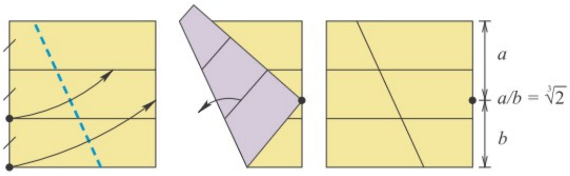

Figure 19.7. Messer's method for doubling a cube using origami.

solve a real cubic equation with constructible coefficients if and only if all its roots are real (Gleason 1988). In particular, because $z^3 - 2 = 0$ has only one real root, one cannot "double a cube," that is, construct $\sqrt[3]{2}$, using such tools.[4]

However, Messer (1986) showed in 1985 that you can double a cube using origami (see Figure 19.7).

Justin (1989b), Geretschläger (1995), and Alperin (2000) independently showed that all cubic equations with constructible coefficients can be solved with origami. Unknown to these authors, a proof of this result was published already in 1936 by Margherita Piazzolla Beloch (1936) (see also Huzita 1994). Because quartic equations can be solved with cubics, the consequence of this result is that all quartic equations can be solved with origami.

It remained an open problem for several years (Hull 1996; Murakami 1987) to determine whether origami can go beyond solving quartics. The answer is NO. This was first established by Huzita and Scimemi (1989), and independently by Alperin (2000). Here we rely on a connection to other work first noted by Demaine and Demaine (2000).

The basic origami fold that the axioms use is a "book fold," which amounts to a reflection through a line. A related geometric-construction tool with this property is a semireflective mirror allowing you to simultaneously see a reflection and see through the mirror. Originally, such a mirror was made by coating plate glass with a thin layer of evaporated metal. Today, one can use a hard thin piece of dark-colored transparent plastic with a smooth finish. Such a device is commonly referred to by the trade name Mira.[5]

Many papers have considered the geometric power of the Mira, starting in 1963 when Hochstein (1963) showed how to trisect an angle with a semireflective mirror (before the term "Mira" was coined). In 1994, Emert et al. (1994) were the first to give a full axiomatic characterization of the Mira. And it turns out that the operations are in fact equivalent to the origami axioms. Emert et al. provided a complete characterization of Mira-constructible (and hence origami-constructible) numbers; they proved that every constructible number is the root of a quadratic or cubic equation with constructible coefficients. Thus, origami can solve all quartic equations but no more.

19.4 FOLDING REGULAR POLYGONS

A classic motivation for determining the power of a particular geometric device is to characterize the constructible regular polygons. For example, it is well known that straight edge and compass can construct a regular n-gon exactly when n is of the form $2^r p_1 \cdots p_k$, where the p_i's are distinct Fermat primes (i.e., prime numbers of the form $2^u + 1$). Origami has a similar characterization, essentially by adding "3" in addition to "2," because origami can solve cubic equations in addition to quadratic equations. More precisely, origami can construct a regular n-gon exactly when n is of the form $2^r 3^s p_1 \cdots p_k$, where the p_i's are distinct primes of the form $2^u 3^v + 1$ (Emert et al. 1994). Curiously, these are precisely the same regular polygons that are constructible with straight edge, compass, and a tool that just trisects (Gleason 1988). In other words, the ability to solve general cubic equations, instead of just those with all real roots, does not help to construct any new regular polygons.

---

[4] This is one of the three classical Euclidean construction questions: given the side of a cube, construct the side of a cube that has twice the volume. If the given side has length 1, the goal is to construct $\sqrt[3]{2}$.

[5] Registered trademark of the Mira-Math Company in Willowdale, Ontario, Canada.

In comparison to straight edge and compass, origami can construct many more regular polygons. Below we list the first several unconstructible polygons for each. In particular, $n = 11$ is the smallest unconstructible regular polygon by paper folding.

Straight edge and compass: 7, 9, 11, 13, 14, 18, 19, 21, 22, 23, 25, 26, 27, 28, 29, 31, 33, 35, 36, 37, 38, 39, 41, 42, 43, 44, 45, 46, 47, 49, 50, 52, 53, 54, 55, 56, 57, 58, 59, . . .

Origami: 11, 22, 23, 25, 29, 31, 33, 41, 43, 44, 46, 47, 49, 50, 53, 55, 58, 59, 61, 62, 66, 67, 69, 71, 75, 77, 79, 82, 83, 86, 87, 88, 89, 92, 93, 94, 98, 99, 100, 101, 103, 106, 107, . . .

19.5 GENERALIZING THE AXIOMS TO SOLVE ALL POLYNOMIALS?

The results on the power of paper folding rely on the restricted set of Huzita–Hatori axioms, which only permit simple folds. It is conceivable that more powerful axioms would permit solving polynomial equations of arbitrary degree via origami. Here we give the germ of an idea from Demaine and Demaine (2000) that might lead to this result. The plan is to exploit the universality theorems from linkages (Part I, Section 3.2, p. 31) that say that any polynomial curve can be traced out by a linkage. We sketch the method first and then discuss what additional axioms might be needed.

The first step is to use the methods described in Chapter 15 to fold the paper into a shape with several *rulers*, as shown in Figure 19.8(a), making n long rulers available for manipulation, plus the shaded *construction area*. The edges of this silhouette are horizontal and vertical and have integral length, and so are easily constructible.

The second step is to "draw" (crease) in the construction area the bars (line segments) of a linkage, conceptually connected by hinges at their endpoints. The example in Figure 19.8(b) shows a 4-bar linkage, but in fact the linkage will have to follow the complete Kempe or Kapovich–Millson universality proofs. We suppose that the linkage has n bars, the same as the number of rulers we construct.

The third step is to fold each ruler into position, atop a corresponding bar in the linkage, as in Figure 19.8(c)–(d). In general, a ruler can be folded into position by first folding it vertically upward, then horizontally left (for bars with negative slope), and then at an appropriate angle to match the bar. By first folding a ruler upward, we never fold a ruler on top of a lower ruler. Thus, we can assign each ruler a "layer" that it occupies, and these layers of paper do not interfere, so the rulers can be folded flexibly and independently.

The fourth step is to reconfigure the rulers so as to simulate the linkage. Each ruler can be marked (by creasing) with the length of the corresponding bar. The resulting

data 01.05.2020

The curvature at a point is
1/r, where r is the radius of the
OSCULATING CIRCLE.

Figure 21.5
Spinning a plane about N
through $p \in S$ yields the
(produz)
principal curvatures at p.
The curvature of the curves for
plane A and B are positive
at p, but negative for
plane C.

Pág. 287 (301)
19.2 Huzita's Axioms and
Hatori's Addition
Box 19.1: Abe's trisection
Important: three parallel
lines
Estudar!!! com Pappus ☆
Fig. 19.4 Abe's trisection analyzed.

)*+,)))-) % B . 1 " > 9
+ . ."
(## . C

CHAPTER X

Two-Variable Iwasawa Theory of Elliptic Curves

Our study of the p-ordinary part of the Selmer group for E is almost at an end. However there is one last twist in the tale. Up to this point, the ground field $\mathbf{F}$ we have been working over has remained fixed, but this assumption can now be relaxed. In particular, allowing the number field $\mathbf{F}_n$ to vary in a cyclotomic $\mathbb{Z}_p$-extension adjoins an extra variable, Y say, to the ambient deformation ring.

Why is this cyclotomic variable Y important, and what net benefits are gained? The principal motivation has its origins in the Iwasawa theory of elliptic curves admitting complex multiplication. Specifically, if E has CM by an order in $\mathcal{O}_{\mathbb{Q}(\sqrt{-D})}$ then $\mathrm{Gal}\Big(\mathbb{Q}(E[p^\infty])/\mathbb{Q}(E[p])\Big) \cong \mathbb{Z}_p \times \mathbb{Z}_p$. In terms of completed group algebras

$$\mathbb{Z}_p[[G]] \;\cong\; \mathbb{Z}_p[[\Gamma^{\mathrm{wt}} \times \Gamma^{\mathrm{cy}}]][\Delta] \;\cong\; \bigoplus_{\delta\in\Delta} \mathbb{Z}_p[[X,Y]].\delta$$

where $G = \mathrm{Gal}\Big(\mathbb{Q}(E[p^\infty])/\mathbb{Q}(\sqrt{-D})\Big)$, and Δ is finite with order coprime to p. Therefore in the CM case, combining the cyclotomic deformation with the weight deformation is *equivalent* to studying the arithmetic of E over the field of definition of its p-power division points.

If E does not admit complex multiplication, then $\mathrm{Gal}\big(\mathbb{Q}(E[p^\infty])/\mathbb{Q}\big)$ is an open subgroup of $\mathrm{GL}_2(\mathbb{Z}_p)$, and its underlying Iwasawa algebra is a non-commutative power series ring in four variables. In this scenario $\mathbb{Z}_p\big[[\Gamma^{\mathrm{wt}} \times \Gamma^{\mathrm{cy}}]\big]$ represents the completed group ring of its maximal torus, i.e. the cyclotomic-weight deformation encodes the abelianisation of GL_2-Iwasawa theory.

A major goal of this chapter is to formulate three new **Main Conjectures**: the first corresponding to the vertical line $s = 1$, another for the line of symmetry $s = k/2$, and the last for the whole (s, k)-plane. These are fundamental statements reconciling the 'algebraic world' with the 'analytic world' (as a useful mnemonic). More precisely, there should be a one-to-one correspondence

$$\Big\{\text{characteristic ideals of Selmer}\Big\} \quad \xleftrightarrow{\text{Main Conj.}} \quad \Big\{p\text{-adic } L\text{-functions}\Big\}.$$

Warning: Before we forget it entirely, let's first remember to compute the leading term of the algebraic L-series $\text{Ш}_{\mathbf{F}}\Big(\mathbb{T}^{-1/2}_{\infty,E}\Big)$. Indeed this p-adic L-value is a vital constituent of the Euler characteristic over the two-variable deformation of $E[p^\infty]$. As many of the details are akin to the vertical calculation (in the previous chapter), we shall attempt to keep the exposition relatively brief.

10.1 The half-twisted Euler characteristic formula

Throughout one takes $\mathbf{F}$ to be an abelian number field, and the prime number $p \geq 3$. For simplicity assume $\rho_{\infty,E}$ is residually irreducible, and $\mathcal{R}_E$ is a Gorenstein ring. The fundamental work of Flach and Nekovář [Fl,Ne3] guarantees the existence of a (nearly) self-dual, symplectic form

$$\big\langle -,- \big\rangle^{\mathrm{Nek}}_{\mathbf{F},p,\mathbb{T}^{-1/2}_{\infty,E}} \;:\; \widetilde{H}^1_f\left(\mathbf{F}_\Sigma/\mathbf{F},\ \mathbb{T}^{-1/2}_{\infty,E}\right) \;\times\; \widetilde{H}^1_f\left(\mathbf{F}_\Sigma/\mathbf{F},\ \mathbb{T}^{-1/2}_{\infty,E}\right) \;\longrightarrow\; \mathcal{R};$$

its manifestation at weight two, yields a $\mathbb{Z}_p$-bilinear pairing

$$\big\langle -,- \big\rangle^{-1/2}_{\mathbf{F},p} : H^1_f\big(\mathbf{F}_\Sigma/\mathbf{F},\ V_pE\big)^{\mathrm{gen}} \;\times\; H^1_f\big(\mathbf{F}_\Sigma/\mathbf{F},\ V_pE\big)^{\mathrm{gen}} \;\longrightarrow\; \mathbb{Q}_p$$

on the 'generic part' of the vector-space H^1_f.

Notation: Let us denote by $\widetilde{\lambda}^{-1/2}_{\infty,E}$ the direct sum of the restriction maps

$$H^1\Big(\mathbf{F}_\Sigma/\mathbf{F},\ A^{-1/2}_{\mathbb{T}_{\infty,E}}\Big) \longrightarrow \bigoplus_{\nu\in\Sigma-S_{\mathbf{F},p}} \left(\frac{H^1\Big(\mathbf{F}_\nu, A^{-1/2}_{\mathbb{T}_{\infty,E}}\Big)}{H^1_{\mathrm{nr}}\Big(\mathbf{F}_\nu, A^{-1/2}_{\mathbb{T}_{\infty,E}}\Big)}\right) \oplus \bigoplus_{\nu\in S_{\mathbf{F},p}} \left(\frac{H^1\Big(\mathbf{F}_\nu, A^{-1/2}_{\mathbb{T}_{\infty,E}}\Big)}{H^1_+\big(\mathbf{F}_\nu\big)^{\perp}}\right).$$

The next result is effectively the half-twisted analogue of Theorem 9.18.

Theorem 10.1. *Suppose the following four conditions hold:*

(i) $\#\text{Ш}_{\mathbf{F}}(E)[p^\infty]$ *is finite;*

(ii) $\big\langle -,- \big\rangle^{-1/2}_{\mathbf{F},p}$ *is non-degenerate;*

(iii) the homomorphism $\widetilde{\lambda}^{-1/2}_{\infty,E}$ *is surjective;*

(iv) E does **not** *attain split multiplicative reduction at any place above p.*

Then $\mathrm{order}_{w=2}\text{Ш}_{\mathbf{F}}\left(\mathbb{T}^{-1/2}_{\infty,E}, w\right) \;=\; \mathrm{rank}_{\mathbb{Z}}E(\mathbf{F}) - \mathrm{corank}_{\Lambda^{\mathrm{wt}}}\mathrm{Sel}_{\mathbf{F}}\big(\mathbb{T}^{-1/2}_{\infty,E}\big) \;=\; r^{\dagger}_{\mathbf{F}}$ *say, and its leading term is given by*

$$\left.\frac{\text{Ш}_{\mathbf{F}}\big(\mathbb{T}^{-1/2}_{\infty,E}, w\big)}{(w-2)^{r^{\dagger}_{\mathbf{F}}}}\right|_{w=2} \times \prod_{\nu} \mathrm{Tam}^{(2)}_{\mathbf{F}_\nu}\left(\mathbb{T}^{-1/2}_{\infty,E}\right) \;\times\; \mathrm{covol}^{(2)}_{\mathbb{T}^{-1/2}_{\infty,E}}\big(\mathbf{F}\otimes\mathbb{Z}_p\big) \;\times\; \mathfrak{D}_{\mathbf{F},p}\big(\rho_{\infty,E}\big)$$

$$= \prod_{\mathfrak{p}\in S^{\mathrm{good}}_{\mathbf{F},p}} \#\widetilde{E}\left(\mathbb{F}_{q_{\mathfrak{p}}}\right)^2 \times \frac{\#\text{Ш}_{\mathbf{F}}(E)\prod_\nu\big[E(\mathbf{F}_\nu):E_0(\mathbf{F}_\nu)\big]}{\#E\big(\mathbf{F}\big)^2_{\mathrm{tors}}} \times \det\big\langle -,- \big\rangle^{-1/2}_{\mathbf{F},p}\Big|_{E(\mathbf{F})\times E(\mathbf{F})}$$

modulo $\mathbb{Z}_p^\times$*, where* $S^{\mathrm{good}}_{\mathbf{F},p} \;=\; S_{\mathbf{F},p} \,\cap\, \big\{\mathfrak{p}\nmid N_E\big\}$.

N.B. The quantity $\prod_{\mathfrak{p}\in S^{\mathrm{good}}_{\mathbf{F},p}}\left(\#\widetilde{E}\left(\mathbb{F}_{q_{\mathfrak{p}}}\right)\right)^2$ is trivial when $p\big|N_E$ as the set $S^{\mathrm{good}}_{\mathbf{F},p} = \emptyset$. Alternatively, if $p\nmid N_E$ then the subset $S^{\mathrm{good}}_{\mathbf{F},p}$ is the whole of $S_{\mathbf{F},p}$, and the product is therefore non-empty.

Proof: The main ingredients are identical to the demonstration of Theorem 9.18. Firstly condition (iv) implies the set of split multiplicative primes $S^{\mathrm{mult}}_{\mathbf{F},p}$ is void, in which case Conjecture 7.11 holds. As a nice consequence, the natural mapping $\mathrm{Sel}_{\mathbf{F}}\left(\mathcal{C}^{\min}\right)[p^\infty] \overset{\alpha^{-1/2}_{\infty,E}}{\longrightarrow} \mathrm{Sel}_{\mathbf{F}}\Big(\mathbb{T}_{\infty,E}\otimes\Psi^{-1/2}\Big)^{\Gamma^{\mathrm{wt}}}$ will then be a quasi-isomorphism of cofinitely-generated groups over $\mathbb{Z}_p$.

Fortunately, condition (iii) ensures we have the short exact sequence

$$0 \longrightarrow \mathrm{Sel}_{\mathbf{F}}\Big(\mathbb{T}^{-1/2}_{\infty,E}\Big) \longrightarrow H^1\Big(\mathbf{F}_\Sigma/\mathbf{F},\ A^{-1/2}_{\mathbb{T}_{\infty,E}}\Big) \overset{\widetilde{\lambda}^{-1/2}_{\infty,E}}{\longrightarrow} \bigoplus_{\nu\in\Sigma} \frac{H^1\big(\mathbf{F}_\nu, A^{-1/2}_{\mathbb{T}_{\infty,E}}\big)}{H^1_{\star,+}\big(\mathbf{F}_\nu, \mathbb{T}^{-1/2}_{\infty,E}\big)^{\perp}} \longrightarrow 0.$$

Taking its Γ^{wt}-cohomology, produces a commutative diagram with exact columns:

$$\begin{tikzcd}
0 \arrow[d] & 0 \arrow[d] \\
\mathrm{Sel}_{\mathbf{F}}\left(\mathcal{C}^{\min}\right)[p^\infty] \arrow[d] \arrow[r, "\alpha^{-1/2}_{\infty,E}"] & \mathrm{Sel}_{\mathbf{F}}\Big(\mathbb{T}^{-1/2}_{\infty,E}\Big)^{\Gamma^{\mathrm{wt}}} \arrow[d] \\
H^1\Big(\mathbf{F}_\Sigma/\mathbf{F},\ \mathcal{C}^{\min}[p^\infty]\Big) \arrow[d, "\lambda^{-1/2}_{\mathcal{C}^{\min}}"] \arrow[r, "\beta^{-1/2}_{\infty,E}"] & H^1\Big(\mathbf{F}_\Sigma/\mathbf{F},\ A^{-1/2}_{\mathbb{T}_{\infty,E}}\Big)^{\Gamma^{\mathrm{wt}}} \arrow[d, "\lambda^{-1/2}_{\infty,E}"] \\
\begin{array}{c} \displaystyle\bigoplus_{\nu\in\Sigma-S_{\mathbf{F},p}} \frac{H^1\big(\mathbf{F}_\nu,\mathcal{C}^{\min}[p^\infty]\big)}{H^1_{\mathrm{nr}}\big(\mathbf{F}_\nu,\mathcal{C}^{\min}[p^\infty]\big)} \\ \displaystyle\oplus \bigoplus_{\nu\in S_{\mathbf{F},p}} \frac{H^1\big(\mathbf{F}_\nu,\mathcal{C}^{\min}[p^\infty]\big)}{\mathcal{W}_{\mathcal{C}^{\min},+}\big(\mathbf{F}_\nu\big)^{\perp}} \end{array} \arrow[d] \arrow[r, "\oplus\delta^{-1/2}_{\infty,E,\nu}"] & \begin{array}{c} \displaystyle\bigoplus_{\nu\in\Sigma-S_{\mathbf{F},p}} \left(\frac{H^1\Big(\mathbf{F}_\nu, A^{-1/2}_{\mathbb{T}_{\infty,E}}\Big)}{H^1_{\mathrm{nr}}\Big(\mathbf{F}_\nu, A^{-1/2}_{\mathbb{T}_{\infty,E}}\Big)}\right)^{\Gamma^{\mathrm{wt}}} \\ \displaystyle\oplus \bigoplus_{\nu\in S_{\mathbf{F},p}} \left(\frac{H^1\Big(\mathbf{F}_\nu, A^{-1/2}_{\mathbb{T}_{\infty,E}}\Big)}{H^1_{+}\big(\mathbf{F}_\nu\big)^{\perp}}\right)^{\Gamma^{\mathrm{wt}}} \end{array} \arrow[d] \\
H^1_f\Big(\mathbf{F},\mathrm{Ta}_p\big(\mathcal{C}^{\min}\big)\Big)^{\vee} & \mathrm{Sel}_{\mathbf{F}}\Big(\mathbb{T}^{-1/2}_{\infty,E}\Big)_{\Gamma^{\mathrm{wt}}} \arrow[l, dashed, "\widehat{\theta}^{-1/2}_{\infty,E}"']
\end{tikzcd}$$

Figure 10.1

Under the conditions (i) and (ii), the order of vanishing of $\text{Ш}_{\mathbf{F}}\left(\mathbb{T}^{-1/2}_{\infty,E},\ w\right)$ at $w=2$ is precisely equal to the dimension of $\mathbb{Q}_p\otimes H^0\left(\Gamma^{\mathrm{wt}},\ \left(\mathrm{Sel}_{\mathbf{F}}\left(\mathbb{T}^{-1/2}_{\infty,E}\right)_{/\Lambda^{\mathrm{wt}}\text{-div}}\right)^{\vee}\right)$. In fact, the half-twisted pairing $\big\langle -,-\big\rangle^{-1/2}_{\mathbf{F},p}$ factorises through the dual map $\theta^{-1/2}_{\infty,E}$. Its discriminant is therefore intertwined with that of

$$H^0\left(\Gamma^{\mathrm{wt}}, \mathrm{Tors}_{\Lambda^{\mathrm{wt}}}\left(\mathrm{Sel}_{\mathbf{F}}\left(\mathbb{T}^{-1/2}_{\infty,E}\right)^{\vee}\right)\right) \overset{\mathrm{nat}}{\longrightarrow} H^1\left(\Gamma^{\mathrm{wt}}, \mathrm{Tors}_{\Lambda^{\mathrm{wt}}}\left(\mathrm{Sel}_{\mathbf{F}}\left(\mathbb{T}^{-1/2}_{\infty,E}\right)^{\vee}\right)\right)$$

which can now be computed via Figure 10.1.

Remarks: (a) Every horizontal arrow in the above diagram is a quasi-isomorphism.

(b) If the map $\xi_{\mathrm{div}}^{-1/2} : \mathcal{S}^{\Gamma^{\mathrm{wt}}} \longrightarrow \mathcal{S}_{\Gamma^{\mathrm{wt}}}$ where $\mathcal{S} = \mathrm{Sel}_{\mathbf{F}}\left(\mathbb{T}_{\infty,E}^{-1/2}\right)_{/\Lambda^{\mathrm{wt}}\text{-div}}$, then

$$\log_p(u_0)^{-r_{\mathbf{F}}^{\dagger}} \times \mathrm{disc}\left(\langle -,- \rangle_{\mathbf{F},p}^{-1/2}\right) \;=\; \mathcal{H}\Big(\xi_{\mathrm{div}}^{-1/2}\Big)^{-1} \times \mathcal{H}\Big(\oplus\, \delta_{\infty,E,\nu}^{-1/2}\Big) \times \mathcal{H}\Big(\beta_{\infty,E}^{-1/2}\Big)^{-1}.$$

(c) The Herbrand quotient of $\beta_{\infty,E}^{-1/2}$ was calculated in Lemma 7.12(a).

(d) The Herbrand quotient of $\delta_{\infty,E,\nu}^{-1/2}$ was calculated in Proposition 8.7(ii) if $\nu \nmid p$, and in Proposition 8.13 when $\nu \big| p$.

Combining (a)-(b)-(c)-(d) together, one obtains a formula relating the half-twisted arithmetic invariants of $\mathcal{C}^{\min}$ to the reciprocal Herbrand quotient $\mathcal{H}\Big(\xi_{\mathrm{div}}^{-1/2}\Big)^{-1}$. However the latter quantity is the leading term of $\text{Ш}_{\mathbf{F}}\left(\mathbb{T}_{\infty,E}^{-1/2}\right)$ at $X = 0$.

Finally, one may exchange the optimal curve $\mathcal{C}^{\min}$ with the strong Weil curve E using Cassels' Isogeny Theorem 1.19, which explains why the defect term $\mathfrak{D}_{\mathbf{F},p}$ rears its ugly head once more.

□

Example 10.2. Let's consider the modular curve $E = X_0(11)$ over the field $\mathbf{F} = \mathbb{Q}$. Since the analytic rank is zero, the half-twisted regulators must therefore be trivial. If $7 \leq p \leq 97$ and $p \neq 11, 19, 29$ then the elliptic curve E has good ordinary reduction at p, and our theorem implies that $\text{Ш}_{\mathbb{Q}}\left(\mathbb{T}_{\infty,E}^{-1/2},\, 2\right)$ will be a p-adic unit. It follows at all good ordinary primes $p < 100$ excluding 5, the μ-invariant over $\mathbb{Q}$ associated to the half-twisted deformation of $\mathrm{Ta}_p\big(X_0(11)\big)$ vanishes (c.f. Ex. 9.19).

The reader will notice that we have studiously avoided the prime $p = 11$ in the above example, because Theorem 10.1 does not cover split multiplicative reduction. Nevertheless in the semistable case, one has a shrewd idea what to expect.

Conjecture 10.3. *Assume that the prime number p divides N_E. Then*

$$\mathrm{order}_{w=2}\text{Ш}_{\mathbf{F}}\left(\mathbb{T}_{\infty,E}^{-1/2}, w\right) \;=\; \mathrm{rank}_{\mathbb{Z}}E(\mathbf{F}) \;+\; \#S_{\mathbf{F},p}^{\mathrm{split}} \;-\; \mathrm{corank}_{\Lambda^{\mathrm{wt}}}\mathrm{Sel}_{\mathbf{F}}\big(\mathbb{T}_{\infty,E}^{-1/2}\big);$$

moreover, the leading term of $\text{Ш}_{\mathbf{F}}^{-1/2}$ satisfies

$$\left.\frac{\text{Ш}_{\mathbf{F}}\big(\mathbb{T}_{\infty,E}^{-1/2}, w\big)}{(w-2)^{r_{\mathbf{F}}^{\dagger} + \#S_{\mathbf{F},p}^{\mathrm{split}}}}\right|_{w=2} \times \prod_{\nu} \mathrm{Tam}_{\mathbf{F}_\nu}^{(2)}\left(\mathbb{T}_{\infty,E}^{-1/2}\right) \;\times\; \mathrm{covol}_{\mathbb{T}_{\infty,E}^{-1/2}}^{(2)}\big(\mathbf{F} \otimes \mathbb{Z}_p\big) \;\times\; \mathfrak{D}_{\mathbf{F},p}$$
$$= \quad \prod_{\nu \in S_{\mathbf{F},p}^{\mathrm{split}}} \frac{\log_p\left(\mathbf{q}_{E_{/\mathbf{F}_\nu}}\right)}{\mathrm{ord}_\nu\left(\mathbf{q}_{E_{/\mathbf{F}_\nu}}\right)} \;\times\; \frac{\#\text{Ш}_{\mathbf{F}}(E)\prod_\nu \big[E(\mathbf{F}_\nu) : E_0(\mathbf{F}_\nu)\big]}{\#E(\mathbf{F})_{\mathrm{tors}}^{\,2}} \;\times\; \det\langle -,- \rangle_{\mathbf{F},p}^{-1/2}$$

again up to p-adic units.

N.B. Over the rationals, the $\mathcal{L}$-invariant $\frac{\log_p(\mathbf{q}_E)}{\mathrm{ord}_p(\mathbf{q}_E)} \neq 0$ courtesy of the result in [StÉ]. For a general number field $\mathbf{K} \neq \mathbb{Q}$ the non-vanishing is still an open problem.

Question. *When is the global-to-local map* $\widetilde{\lambda}_{\infty,E}^{-1/2}$ *surjective?*

In general, the answer to this question lies fairly deep.

If the representation $\rho_{\infty,E} : G_{\mathbb{Q}} \longrightarrow \mathrm{GL}_2\big(\mathbb{Z}_{p,\mathbb{U}_E}\langle\langle w\rangle\rangle\big)$ is of CM-type by $\mathcal{O}_{\mathbb{Q}(\sqrt{-D})}$, this reduces to the existence or otherwise of non-trivial pseudo-null submodules in the (Pontrjagin dual of) the two-variable Selmer group over $\mathbf{F}\Big(\sqrt{-D},\ E[p^\infty]\Big)$. Alternatively, if we demand that the p-adic L-function is suitably well-behaved along the line $s = k/2$, one can still make some progress:

Proposition 10.4. *Assume Greenberg's conjecture holds for the family* $\{\mathbf{f}_k\}_{k\in\mathbb{U}_E\cap\mathbb{N}}$ *with its generic rank* $r_{\mathrm{gen}}(\mathbf{F}) = 0$*, and in addition that* $\mathcal{R}_E$ *is a regular local ring. Then the homomorphism* $\widetilde{\lambda}_{\infty,E}^{-1/2}$ *is always surjective.*

Proof: Let $\mathfrak{S}_{\mathbf{F}}\Big(\mathbb{T}_{\infty,E}^{-1/2}\Big)$ denote the compact Selmer group associated to the half-twisted deformation of $\rho_{E,p}$. Its Pontrjagin dual sits inside the Poitou-Tate sequence

$$\begin{aligned} 0 \ \longrightarrow\ \mathrm{Sel}_{\mathbf{F}}\Big(\mathbb{T}_{\infty,E}^{-1/2}\Big) \ &\longrightarrow\ H^1\Big(\mathbf{F}_\Sigma/\mathbf{F},\ A_{\mathbb{T}_{\infty,E}}^{-1/2}\Big) \\ \xrightarrow{\widetilde{\lambda}_{\infty,E}^{-1/2}}\ \bigoplus_{\nu\in\Sigma} \frac{H^1\big(\mathbf{F}_\nu, A_{\mathbb{T}_{\infty,E}}^{-1/2}\big)}{H^1_{\star,+}\big(\mathbf{F}_\nu, \mathbb{T}_{\infty,E}^{-1/2}\big)^{\perp}} \ &\longrightarrow\ \mathfrak{S}_{\mathbf{F}}\Big(\mathbb{T}_{\infty,E}^{-1/2}\Big)^{\vee} \ \longrightarrow\ \ldots . \end{aligned}$$

To establish the surjectivity of $\widetilde{\lambda}_{\infty,E}^{-1/2}$, clearly it is enough to show $\mathfrak{S}_{\mathbf{F}}\Big(\mathbb{T}_{\infty,E}^{-1/2}\Big) = 0$. There are two stages:

(A) Prove that $\mathfrak{S}_{\mathbf{F}}\Big(\mathbb{T}_{\infty,E}^{-1/2}\Big)$ is Λ^{wt}-torsion, and has universally p-bounded order;

(B) Prove that all proper subgroups of $H^1\Big(\mathbf{F}_\Sigma/\mathbf{F},\ \mathbb{T}_{\infty,E}^{-1/2}\Big)\big[p^\infty\big]$ are zero.

Since (A) implies the compact Selmer group lies in the $\mathbb{Z}_p$-torsion submodule of $H^1\Big(\mathbf{F}_\Sigma/\mathbf{F},\ \mathbb{T}_{\infty,E}^{-1/2}\Big)$, any finite subgroup of $\mathfrak{S}_{\mathbf{F}}\Big(\mathbb{T}_{\infty,E}^{-1/2}\Big)$ must then be zero by (B). As a corollary, the whole Selmer group was trivial to begin with.

Remark: One can identify the p^∞-torsion submodule of $H^1\Big(\mathbf{F}_\Sigma/\mathbf{F},\ \mathbb{T}_{\infty,E}^{-1/2}\Big)$ with the 0^{th}-cohomology group

$$H^0\Big(\mathbf{F}_\Sigma/\mathbf{F},\ \mathbb{T}_{\infty,E}^{-1/2}\otimes\mathbb{Q}_p/\mathbb{Z}_p\Big) \quad\subset\quad \left(\Psi_{\mathcal{R}}^{-1/2}\otimes\mathbf{e}_{\mathcal{R}}.\varinjlim_{\pi_1^*}{}^{\mathbf{t}}\mathcal{J}_r^{\mathrm{ord}}\big[p^\infty\big]^{\vee}\otimes_{\mathbb{Z}_p}\mu_{p^\infty}\right)(\mathbf{F}).$$

The right-hand side has no finite p-subgroups, which can be seen by using the same argument to Step 3 of Theorem 9.2 (N.B. as an $\mathcal{R}_E$-module $\mathbf{e}_{\mathcal{R}}.\varinjlim_{\pi_1^*}{}^{\mathbf{t}}\mathcal{J}_r^{\mathrm{ord}}\big[p^\infty\big]^{\vee}$ is identical to $\Psi_{\mathcal{R}}^{-1/2}\otimes\mathbf{e}_{\mathcal{R}}.\varinjlim_{\pi_1^*}{}^{\mathbf{t}}\mathcal{J}_r^{\mathrm{ord}}\big[p^\infty\big]^{\vee}$, therefore the transition maps π_1^*, $\pi_{p\,*}$ must again act nilpotently on the p-torsion). It immediately follows (B) is true.

This leaves us with the task of establishing (A).

Key Claim: For almost all arithmetic weights $k \in \mathbb{U}_E \cap \mathbb{N}_{\geq 2}$, the natural map

$$\mathfrak{S}_{\mathbf{F}}\Big(\mathbb{T}_{\infty,E}^{-1/2}\Big) \otimes_{\Lambda^{\mathrm{wt}},\mathfrak{g}_k} \mathbb{Z}_p \quad \overset{\mathrm{nat}}{\longrightarrow} \quad H^1_{f,\mathrm{Spec}\mathcal{O}_{\mathbf{F}}}\Big(\mathbf{F}_\Sigma/\mathbf{F},\ \mathbb{T}_{\infty,E}^{-1/2} \otimes_{\Lambda^{\mathrm{wt}},\mathfrak{g}_k} \mathbb{Z}_p\Big)$$

has finite kernel and cokernel, bounded independently of k.

When $\mathbf{F} = \mathbb{Q}$ and $\mathcal{R}_E = \Lambda^{\mathrm{wt}}$ this statement is shown in the article [NP, Sect.4]. However the proof generalises seamlessly to Noetherian regular local rings, as all their height one prime ideals are principal. Similarly, the argument for abelian number fields $\mathbf{F}$ reduces to the analogous result over $\mathbb{Q}$, since one may decompose the Selmer group into its $X_{\mathbf{F}}$-eigencomponents.

Remark: If $k \equiv 2 \mod 2(p-1)$ then $\mathbb{T}_{\infty,E}^{-1/2} \otimes_{\Lambda^{\mathrm{wt}},\mathfrak{g}_k} \mathbb{Z}_p$ is a lattice inside $V_{\mathbf{f}_k}\big(k/2\big)$. Provided that the L-value $L\big(\mathbf{f}_k/\mathbf{F}, k/2\big)$ is non-zero, Kato has shown [Ka1, 14.2(2)] the vector-space Selmer group

$$H^1_{f,\mathrm{Spec}\mathcal{O}_{\mathbf{F}}}\Big(\mathbf{F}_\Sigma/\mathbf{F},\ V_{\mathbf{f}_k}\big(k/2\big)\Big) \ = \ 0.$$

In particular, the above specialisation $H^1_{f,\mathrm{Spec}\mathcal{O}_{\mathbf{F}}}\Big(\mathbf{F}_\Sigma/\mathbf{F},\ \mathbb{T}_{\infty,E}^{-1/2} \otimes_{\Lambda^{\mathrm{wt}},\mathfrak{g}_k} \mathbb{Z}_p\Big)$ must therefore be a finite abelian p-group.

Because we assumed that the generic rank $r_{\mathrm{gen}}(\mathbf{F}) = 0$, the L-value at $s = k/2$ is non-zero at all bar finitely many exceptional weights k. Moreover the 'Key Claim' together with Kato's result implies $\#\left(\mathfrak{S}_{\mathbf{F}}\Big(\mathbb{T}_{\infty,E}^{-1/2}\Big) \otimes_{\Lambda^{\mathrm{wt}},\mathfrak{g}_k} \mathbb{Z}_p\right) < \infty$ for such k, in which case the compact Selmer group over $\mathbf{F}$ is Λ^{wt}-torsion.

Lastly, an identical argument to Theorem 9.2 (Step 2) establishes that its order is bounded above by $p^{\nu_1^\dagger + \nu_2^\dagger}$; here $p^{\nu_1^\dagger}$ is chosen large enough to kill off the kernel of

$$\mathfrak{S}_{\mathbf{F}}\Big(\mathbb{T}_{\infty,E}^{-1/2}\Big) \otimes_{\Lambda^{\mathrm{wt}},\mathfrak{g}_k} \mathbb{Z}_p \overset{\mathrm{nat}}{\to} H^1_{f,\mathrm{Spec}\mathcal{O}_{\mathbf{F}}} \hookrightarrow \Big(\big(\mathbb{T}_{\infty,E}^{-1/2} \otimes_{\Lambda^{\mathrm{wt}},\mathfrak{g}_k} \mathbb{Z}_p\big) \otimes \mathbb{Q}/\mathbb{Z}\Big)^{G_{\mathbf{F}}}$$

and likewise $\nu_2^\dagger = \max_{\tau,\psi}\left\{\nu_{2,\tau,\psi}^\dagger\right\}$, where each integer $\nu_{2,\tau,\psi}^\dagger$ is set equal to

$$\max_{1 \leq j \leq n(\tau,\psi)} \left\{\mathrm{ord}_p\Big(1 - \psi(l)a_l(\mathbf{f}_k)l^{1-k/2} + \psi(l)^2 l^{2-k/2} < l >^{k-2}\Big) \,\Big|\, k \in D_j^{\tau,\psi},\, k \equiv \tau\right\}$$

N.B. the above notation is the same as was employed in the proof of Theorem 9.2.

The surjectivity of the homomorphism $\widetilde{\lambda}_{\infty,E}^{-1/2}$ follows immediately.

□

If $r_{\mathrm{gen}}(\mathbf{F}) > 0$ then it is unclear whether or not the global-to-local map surjects. One possibility is to mimic the approach of §6.5, and construct an embedding

$$\mathfrak{S}_{\mathbf{F}}\Big(\mathbb{T}_{\infty,E}^{-1/2}\Big) \quad \overset{?}{\hookrightarrow} \quad \mathrm{Hom}_{\Lambda^{\mathrm{wt}}}\left(\mathrm{Sel}_{\mathbf{F}}\Big(\mathbb{T}_{\infty,E}^{-1/2}\Big)^{\vee},\ \mathbb{Z}_p\langle\!\langle w\rangle\!\rangle\right)^{\square}.$$

The right-hand side is a free $\mathbb{Z}_p[\![\Gamma^{\mathrm{wt}}]\!]$-module of finite-type; one could then compute Λ^{wt}-ranks along the Poitou-Tate sequence, to deduce triviality of $\mathrm{Coker}\Big(\widetilde{\lambda}_{\infty,E}^{-1/2}\Big)$.

10.2 The p-adic height over a double deformation

It is fair to say we now possess a pretty good understanding of the arithmetic of E along the vertical line $s = 1$, and also along the line of symmetry $s = k/2$ (subject to Greenberg's Conjecture 7.5, of course). The very last piece of the jigsaw is to allow the number field $\mathbf{F}$ to vary in a cyclotomic $\mathbb{Z}_p$-extension of the ground field. The finite sub-layers in this $\mathbb{Z}_p$-extension '$\mathbf{F}^{\mathrm{cy}}$' are still abelian extensions of $\mathbb{Q}$, however they are no longer unramified at the prime p.

We'll begin by recalling the structure theory for modules over a two-variable deformation ring. Let R be a normal integral domain, finite and flat over $\mathbb{Z}_p[\![X]\!]$. An $R[\![Y]\!]$-module is called *pseudo-null* if its localisations at all height one primes in Spec $R[\![Y]\!]$ are trivial. Further, a homomorphism $\phi : M \longrightarrow N$ of $R[\![Y]\!]$-modules is said to be a *pseudo-isomorphism* if Kerϕ and Cokerϕ are both pseudo-null.

Notation: $M \overset{\mathrm{ps}\sim}{\longrightarrow} N$.

Assume that M denotes some compact, finitely-generated, torsion $R[\![Y]\!]$-module. The structure theory informs us there exists a pseudo-isomorphism

$$M \quad \overset{\mathrm{ps}\sim}{\longrightarrow} \quad \bigoplus_{i=1}^{t} \frac{R[\![Y]\!]}{\mathcal{P}_i^{e_i}}$$

where the $\mathcal{P}_i$'s are height one prime ideals of $R[\![Y]\!]$, and the e_i's are positive integers. The product $\prod_{i=1}^{t} \mathcal{P}_i^{e_i}$ is the characteristic ideal of M (it need not be principal).

Hypothesis. *The integral domain R is regular.*

As an immediate consequence, the two-variable ring $R[\![Y]\!]$ must also be regular. This condition guarantees that the ideal $\mathrm{char}_{R[\![Y]\!]}\big(M\big)$ has a principal generator. We also observe that all regular CNL rings are automatically Gorenstein.

Lemma 10.5. *Let $\mathfrak{C}_M(Y)$ denote the characteristic power series of M. Then*

$$r \;=\; \mathrm{order}_{Y=0}\Big(\mathfrak{C}_M(Y)\Big) \;\geq\; \mathrm{rank}_R\Big(M/Y.M\Big)$$

with equality iff $\alpha : M[\times Y] \longrightarrow M/Y.M$ has an R-torsion kernel and cokernel; assuming this does occur, one has the leading-term formula

$$\left.\frac{\mathfrak{C}_M(Y)}{Y^r}\right|_{Y=0} \;=\; \frac{\mathrm{char}_R\big(\mathrm{Coker}(\alpha)\big)}{\mathrm{char}_R\big(\mathrm{Ker}(\alpha)\big)} \qquad \text{modulo } R^{\times}.$$

The proof is fairly standard, and can be found in any decent algebra textbook.

Remark: In the next section we shall take $R = \mathcal{R}_E$, the universal p-ordinary deformation ring for $\rho_{E,p} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Z}_p)$. The $R[\![Y]\!]$-module M will then either be the dual of $\mathrm{Sel}_{\mathbf{F}^{\mathrm{cy}}}\Big(\mathbb{T}_{\infty,E}\Big)$, or in the half-twisted case the dual of $\mathrm{Sel}_{\mathbf{F}^{\mathrm{cy}}}\Big(\mathbb{T}_{\infty,E}^{-1/2}\Big)$.

01 05 2020

F) Books_76_2020/
Elliptic Curves and Big Galois
Representations/

Elliptic Curves and Big Galois
Representations... pdf

obs: Livro bom!

Pag. 222(233) Chapter X

Two-Variable Iwasawa Theory
of Elliptic Curves.

159) 20200502

A) "A line is considered a circle
with infinite radius and zero
curvature".

"In fact the work was of such
quality that Weingarten received
a prize for work on the lines
of curvature of a surface in 1857".

"A formula for the line integral of the
geodesic curvature along a closed
curve is known as the

Gauss Bonnet theorem".

"That is, in this case, the Riemann
curvature tensor has $4 \times 4 \times 4 \times 4 = 256$
components!"

& =

01.05.2020

F) Books_76_2020/
Elliptic Curves and Big Galois Representations/

Elliptic Curves and Big Galois Representations...pdf

obs: Livro bom!

Pag. 222(233) Chapter X

Two-Variable Iwasawa Theory of Elliptic Curves.

159) 20200502

A) "A line is considered a circle with infinite radius and zero curvature".

"In fact the work was of such quality that Weingarten received a prize for work on the lines of curvature of a surface in 1857".

"A formula for the line integral of the geodesic curvature along a closed curve is known as the Gauss Bonnet theorem".

"That is, in this case, the Riemann curvature tensor has 4 × 4 × 4 × 4 = 256 components!"

data 02.05.2020

"A deeper characterization of a helix is that is the unique curve in 3-space for which the ratio of curvature to torsion is a constant, a result know as Lancret's theorem."

"Some of his work on physical topics relates to his non-euclidean geometry for he examined how the gravitational potential as given by Newton would have to be modified in a space of negative curvature".

"The local curvature could then be computed from the coefficients of this quadratic surface".

"He showed that spaces of constant curvature could have several different topological structures".

"Surfaces of constant mean curvature or constant Gaussian curvature are now called Weingarten surfaces!"

(surfaces)

"The value assigned for each cell is the value of Gaussian curvature at the cell centroid"

(centroid).

02 05 2020

1 "Gaussian curvature is positive 1 for surfaces like a hemisphere, which are either concave or convex in all the directions"

B) B1) Books_2020_12/
Books_75_2020/
PI/

Celestial Mechanics/

B2) Books_76_2020/

Einstein the Closet Geocentrist_Videos/

B3) Books_76_2020/
A Treatise on Solid Geometry/

A Treatise on Solid Geometry.pdf

Pág. 67(37)

104. To find the distance between two points whose (de quem) equations are given

105. (Pág. 69(99))

We deduce the expression for the square of the distance.

106. (Pág. 69(99)) IMPORTANT: ⚠

2 2 > A / / $> A /' / $ &$

2 > A /'3/ $B 9 " $

2#)*+,)))-) = * 7 " 9 5 ."
(3' =' !

II. For points between BCD, and ABC, ACD, ADC produced,

$$\frac{s}{l} \text{ is negative, } \frac{s}{m},\ \frac{s}{n},\ \frac{s}{r} \text{ between 1 and } \infty.$$

III. For points within the solid angle formed by BCA, CDA, DBA produced,

$$\frac{s}{l} \text{ is between 0 and 1, } \frac{s}{m},\ \frac{s}{n},\ \frac{s}{r} \text{ negative.}$$

IV. For points between ACD, BCD, BAC, ABD produced,

$$\frac{s}{m} \text{ and } \frac{s}{l} \text{ are negative, } \frac{s}{n} \text{ and } \frac{s}{r} \text{ are positive.}$$

And similarly for the nine other compartments.

103. The results of the preceding article may be obtained, if we assume a knowledge of Tetrahedral Co-ordinates, by observing that the tetrahedral co-ordinates of the point are

$$\frac{l}{s},\ \frac{m}{s},\ \frac{n}{s},\ \frac{r}{s}. \qquad \text{(Art. 97).}$$

104. *To find the distance between two points whose equations are given.*

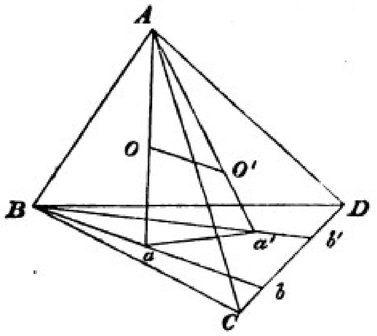

Let the equations of the two points O, O' be

$$l\alpha + m\beta + n\gamma + r\delta = 0,$$
$$l'\alpha + m'\beta + n'\gamma + r'\delta = 0,$$

and let $l + m + n + r = s$, $l' + m' + n' + r' = s'$.

Digitized by Google

Then, if AO, AO' meet BCD in a, a', and Ba, Ba' meet CD in b, b', we have from the triangles OAO', aAa',

$$\frac{AO^2+AO'^2-OO'^2}{2AO.AO'}=\cos OAO'=\frac{Aa^2+Aa'^2-aa'^2}{2Aa.Aa'}.$$

But $\dfrac{Oa}{l}=\dfrac{Aa}{s}=\dfrac{AO}{s-l}$, and $\dfrac{O'a'}{l'}=\dfrac{Aa'}{s'}=\dfrac{AO'}{s'-l'}$;

$$\therefore\ OO'^2=\left(1-\frac{l}{s}\right)^2 Aa^2+\left(1-\frac{l'}{s'}\right)^2 Aa'^2$$

$$-\left(1-\frac{l}{s}\right)\left(1-\frac{l'}{s'}\right)(Aa^2+Aa'^2-aa'^2)$$

$$=\left(1-\frac{l}{s}\right)\left(\frac{l'}{s'}-\frac{l}{s}\right)Aa^2$$

$$+\left(1-\frac{l'}{s'}\right)\left(\frac{l}{s}-\frac{l'}{s'}\right)Aa'^2+\left(1-\frac{l}{s}\right)\left(1-\frac{l'}{s'}\right)aa'^2.$$

From which form it is manifest that the final result will be an expression in terms of the squares of the edges, and we shall obtain the result by investigating the coefficient of AB^2, and deducing the rest by symmetry.

Now the only terms which can involve AB^2 are contained in Aa^2 and Aa'^2, and, as before,

$$\frac{AB^2+Ba^2-Aa^2}{2AB.Ba}=\cos ABa=\frac{AB^2+Bb^2-Ab^2}{2AB.Bb},$$

also $$\frac{ab}{m}=\frac{Bb}{s-l}=\frac{Ba}{s-l-m};$$

$$\therefore\ Aa^2=AB^2-\left(1-\frac{m}{s-l}\right)AB^2+\ldots\ldots$$

$$=\frac{m}{s-l}AB^2+\ldots\ldots$$

writing down only the terms in which AB^2 appears.

Similarly, $Aa'^2=\dfrac{m'}{s'-l'}AB^2+\ldots\ldots$

$$\therefore\ OO'^2 = \frac{m}{s}\left(\frac{l'}{s'} - \frac{l}{s}\right) AB^2 + \frac{m'}{s'}\left(\frac{l}{s} - \frac{l'}{s'}\right) AB^2 + \ldots\ldots$$
$$= \left(\frac{l'}{s'} - \frac{l}{s}\right)\left(\frac{m}{s} - \frac{m'}{s'}\right) AB^2 + \ldots\ldots$$

which gives the distance required.

105. If $\alpha, \beta, \gamma, \delta$ and $\alpha', \beta', \gamma', \delta'$ be tetrahedral co-ordinates of the points O, O' we deduce the expression for the square of the distance,

$$OO'^2 = (\alpha' - \alpha)(\beta - \beta') AB^2 + \ldots\ldots$$

which gives the distance between two points, when their tetrahedral co-ordinates are given.

106. *To find the relation between the co-ordinates of a plane.*

Let $\alpha, \beta, \gamma, \delta$ be the co-ordinates of a plane, λ, μ, ν, ρ the cosines of the angles between the direction in which the co-ordinates are measured, and the perpendiculars from A, B, C, D on the opposite faces, and let Aa be the perpendicular on BCD, ϖ_b, ϖ_c, ϖ_d those from B, C, D on the plane drawn through a parallel to the given plane, EF the trace of this plane on BCD.

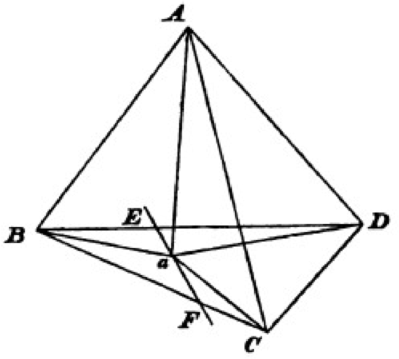

Project Aa on the normal to the given plane;

$$\left.\begin{aligned} \therefore\ p_1\lambda &= \alpha - \beta + \varpi_b \\ &= \alpha - \gamma + \varpi_c \\ &= \alpha - \delta + \varpi_d \end{aligned}\right\}. \qquad (1)$$

Digitized by Google

Now since ϖ_b, ϖ_c, ϖ_d are proportional to the perpendiculars on EF, from B, C, D,

$$\varpi_b \,.\, aCD + \varpi_c \,.\, aDB + \varpi_d \,.\, aBC = 0\,;^*$$

$$\text{and } aCD + aDB + aBC = BCD\,;$$

$\therefore$ by (1), $p_1\lambda \,.\, BCD = \alpha \,.\, BCD - \beta \,.\, aCD - \gamma \,.\, aDB - \delta \,.\, aBC$;

also, $p_1 \,.\, BCD = p_2 \,.\, ACD = p_3 \,.\, ADB = p_4 \,.\, ABC$;

$$\therefore\ \lambda = \frac{\alpha}{p_1} - \frac{\beta}{p_2}\cos(CD) - \frac{\gamma}{p_3}\cos(DB) - \frac{\delta}{p_4}\cos(BC),$$

denoting by (CD) the angle between the faces which meet in CD, and similarly for the rest.

Similar values may be obtained in the same way for μ, ν, and ρ.

If P be the foot of the perpendicular from A on the given plane, the perpendiculars from P on the faces of the tetrahedron will be

$$a\lambda - p_1,\ a\mu,\ a\nu,\ a\rho,$$

$$\therefore\ \alpha\frac{a\lambda - p_1}{p_1} + \beta\frac{a\mu}{p_2} + \gamma\frac{a\nu}{p_3} + \delta \,.\, \frac{a\rho}{p_4} = 0\,; \qquad \text{Art. (81)}$$

$$\therefore\ \frac{\alpha}{p_1}\lambda + \frac{\beta}{p_2}\mu + \frac{\gamma}{p_3}\nu + \frac{\delta}{p_4}\rho = 1,$$

$$\text{or, } \frac{\alpha^2}{p_1^{\,2}} + \frac{\beta^2}{p_2^{\,2}} + \frac{\gamma^2}{p_3^{\,2}} + \frac{\delta^2}{p_4^{\,2}}$$
$$- \frac{2\alpha\beta}{p_1p_2}\cos(CD) - \frac{2\alpha\gamma}{p_1p_3}\cos(DB) - \frac{2\alpha\delta}{p_1p_4}\cos(BC)$$
$$- \frac{2\gamma\delta}{p_3p_4}\cos(AB) - \frac{2\delta\beta}{p_4p_2}\cos(AC) - \frac{2\beta\gamma}{p_2p_3}\cos(AD) = 1,$$

which is the relation required.

---

[*] The equation of a straight line, referred to trilinear co-ordinates, is

$$\alpha\frac{\varpi_1}{p_1} + \beta\frac{\varpi_2}{p_2} + \gamma\frac{\varpi_3}{p_3} = 0,$$

where ϖ_1, ϖ_2, ϖ_3 are the perpendiculars, estimated in the same direction, from the angular points of the fundamental triangle upon the straight line, and p_1, p_2, p_3 those upon the sides. Now, taking EF for the straight line, and BCD for the fundamental triangle, for a, a point in that line, $\frac{\alpha}{p_1} : \frac{\beta}{p_2} : \frac{\gamma}{p_3} :: aCD : aDB : aBC$,

$$\therefore\ \varpi_1 \,.\, aCD + \varpi_2 \,.\, aDB + \varpi_3 \,.\, aBC = 0,$$

whence the equation in the text.

Here (αβγ) are the trilinear point coor. of the point a

Digitized by Google

107. *To find the relations between the co-ordinates of a plane at an infinite distance.*

If the plane be at an infinite distance, the difference between any two of its co-ordinates will vanish compared with either of them; whence, if α, β, γ, δ be the co-ordinates of the plane, we must have

$$\alpha = \beta = \gamma = \delta,$$

the relations required.

108. Since each of the co-ordinates is of infinite magnitude, the expression for λ in the last article will give us

$$0 = \frac{1}{p_1} - \frac{1}{p_2}\cos(CD) - \frac{1}{p_3}\cos(DB) - \frac{1}{p_4}\cos(BC),$$

which is the equation arising from the projection of the faces meeting in A upon BCD.

From this and the three similar equations, we may eliminate p_1, p_2, p_3, and p_4, and obtain the relation which subsists between the cosines of the inclinations of the faces.

109. *To find the perpendicular from a given point upon a given plane, referred to Tetrahedral Co-ordinates.*

Let $(\alpha', \beta', \gamma', \delta')$ be the co-ordinates of the point,

$$l\alpha + m\beta + n\gamma + r\delta = 0,$$

the equation of the plane.

If α_1, β_1, γ_1, δ_1 be the point-co-ordinates of the plane, the equation of the plane may be written

$$\alpha\alpha_1 + \beta\beta_1 + \gamma\gamma_1 + \delta\delta_1 = 0, \qquad \text{Art. (81)},$$

and the perpendicular required is equal to

$$\alpha'\alpha_1 + \beta'\beta_1 + \gamma'\gamma_1 + \delta'\delta_1. \qquad \text{Art. (94)}.$$

Now $\dfrac{\alpha_1}{l} = \dfrac{\beta_1}{m} = \dfrac{\gamma_1}{n} = \dfrac{\delta_1}{r} = \dfrac{\pm\sqrt{\left\{\dfrac{\alpha_1^2}{p_1^2} + \ldots\ldots - \dfrac{2\alpha_1\beta_1}{p_1p_2}\cos(CD) - \ldots\ldots\right\}}}{\sqrt{\left\{\dfrac{l^2}{p_1^2} + \ldots\ldots - \dfrac{2lm}{p_1p_2}\cos(CD) - \ldots\ldots\right\}}}$,

therefore the perpendicular required is

$$\frac{l\alpha' + m\beta' + n\gamma' + r\delta'}{\pm\sqrt{\left\{\dfrac{l^2}{p_1^2} + \ldots\ldots - \dfrac{2lm}{p_1p_2}\cos(CD) - \ldots\ldots\right\}}}.$$

Digitized by Google

It may be shewn, without much difficulty, that the denominator of this expression cannot vanish for real values of $l : m : n : r$, except when $l = m = n = r$, in which case, since $\alpha_1 = \beta_1 = \gamma_1 = \delta_1$, the plane is at an infinite distance.

110. *To find the cosines of the angles which a plane, whose co-ordinates are given, makes with the faces of the tetrahedron.*

If λ, μ, ν, ρ be these cosines, α, β, γ, δ the co-ordinates of the given plane, we have deduced in the previous article

$$\lambda = \frac{\alpha}{p_1} - \frac{\beta}{p_2}\cos(CD) - \frac{\gamma}{p_3}\cos(DB) = \frac{\delta}{p_4}\cos(BC),$$

and similarly for μ, ν, ρ.

The relation between λ, μ, ν and ρ is formed immediately from the consideration that the equation

$$\frac{\lambda\alpha}{p_1} + \frac{\mu\beta}{p_2} + \frac{\nu\gamma}{p_3} + \frac{\rho\delta}{p_4} = 1,$$

when the plane moves parallel to itself to an infinite distance, becomes

$$\frac{\lambda}{p_1} + \frac{\mu}{p_2} + \frac{\nu}{p_3} + \frac{\rho}{p_4} = 0.$$

If the plane be the face BCD,

$$\lambda = 1, \quad \mu = -\cos CD, \quad \nu = -\cos DB, \text{ and } \rho = -\cos BC;$$

$$\therefore\ \frac{1}{p_1} - \frac{1}{p_2}\cos CD - \frac{1}{p_3}\cos DB - \frac{1}{p_4}\cos BC = 0,$$

as before, Art. (108).

111. *To find the condition that two planes whose co-ordinates are given may be parallel.*

If $(\alpha, \beta, \gamma, \delta)$ $(\alpha_1', \beta_1', \gamma_1', \delta')$ be the two planes,

$$\alpha - \alpha' = \beta - \beta' = \gamma - \gamma' = \delta - \delta',$$

since the perpendicular distance between the planes is constant.

112. *To find the equation of a line.*

Let the equations of two points in the line be

$$u = l\alpha + m\beta + n\gamma + \nu\delta = 0,$$
$$v = l'\alpha + m'\beta + n'\gamma + \nu'\delta = 0,$$

Digitized by Google

02 05 2020

1

"Gaussian curvature is positive for surfaces like a hemisphere, which are either concave or convex in all the directions"

B) B1) Books_2020_12/
Books_75_2020/
PI/

Celestial Mechanics/

B2) Books_76_2020/

Einstein the Closet Geocentrist_Videos/

B3) Books_76_2020/
A Treatise on Solid Geometry/

A Treatise on Solid Geometry.pdf

Pág. 67(97)

104. To find the distance between two points whose (de quem) equations are given

105. (Pág. 69(99))

We deduce the expression for the square of the distance.

106. (Pág. 69(99)) IMPORTANT: ✡

02.05.2020

106. To find the relation between the co-ordinates of a plane

→ usar para triangular um ponto que foi projetado da superfície curvada para o plano (superfície plana).

IMPORTANT:

O ponto A é o referencial inicial para a projeção em uma superfície.

O ponto A pode estar sobre a curvatura.

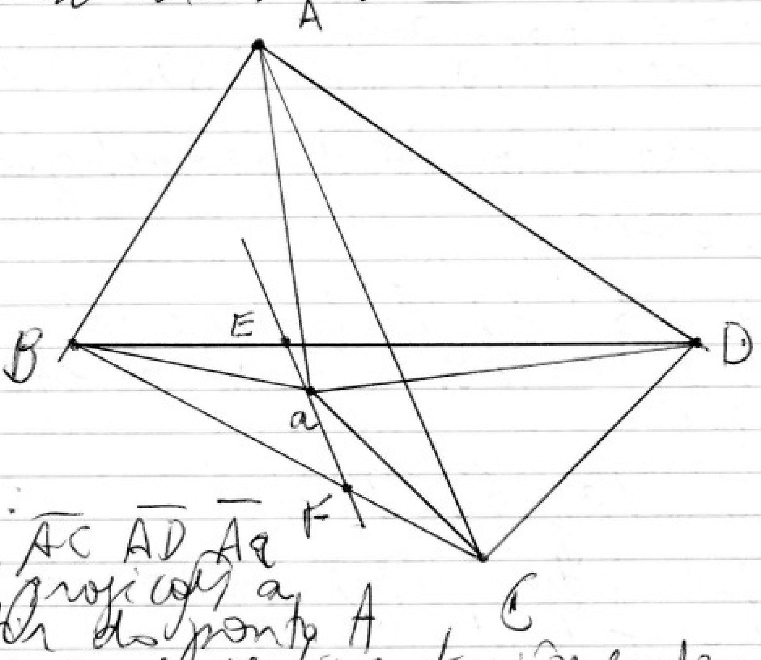

Obs.:
$\overline{AB}$ $\overline{AC}$ $\overline{AD}$ $\overline{Aa}$
são projeções a partir do ponto A para a superfície triangular $\triangle BCD$.

02 05 2020

Pág. 70 (100) Nota de rodapé

* The equation of a straight line referred to trilinear co-ordinates,) is

$$\alpha \frac{\varpi_1}{\rho_1} + \beta \frac{\varpi_2}{\rho_2} + \gamma \frac{\varpi_3}{\rho_3} = 0,$$

where

$\varpi_1, \varpi_2, \varpi_3$ are the perpendiculars, estimated in the same direction, from the angular points of the fundamental triangle upon (em cima de) the straight line, and

ρ_1, ρ_2, ρ_3 those upon the sides.

Now taking EF for the straight line, and BCD for the fundamental triangle, for a, a point in that line,

$$\frac{\alpha}{\rho_1} : \frac{\beta}{\rho_2} : \frac{\gamma}{\rho_3} :: aCD : aDB : aBC,$$

$$\therefore \varpi_1 . aCD + \varpi_2 . aDB + \varpi_3 . aBC = 0,$$

whence (de onde) the equation in the text.

data 02 05 2020
S T Q Q S [✓ S] D

Pág. 71 (101)

Item 107. To find the relations between the coordinates of plane an infinite distance.

outra fonte:

Books_77_2020/
Treatise-of-Solid_Geometry_
Frost/

→ Diretório todo!!!

outro livro (Frost)

An Elementary Treatise on Curve Tracing (2004...)..djvu.

c) Gaussian curvature

Source: youtube.com/playlist?list=PLF566D2BAC693237E

Copy: Books_76_2020/
Einstein the closet geocentrist videos/
a) Relativity_videos/
b) Gauss/ → Relativity 76. ✡

-
)*+,)))-) = 8 * 7 " 9 5
."

9 1
>--? / / / '@)>--? / /)> A /'3/)B 7
9 / ")+ 1 5/ ")

)D&)>--? / / / '@)>--? / /)> A /'3/)B 7
9 / ")+ 1 5/ ")9)

)*+,)))-) = + 1 5 '2 " 5
&& .!

" " 7 . " .

8 2 #

")*+,)))-) = E . 2
. && ."

0
. $$. $ $ = # # ."

)*+,)))-) = = # # ."

-
)*+,)))-) = = !=' ."

* +-C& *7&-0 FB G+&B 9G0G+*7+& B B7 ,EB;,EB -04B 7E+B

)*+,)))-) = E 1 H I " "
."

)*+,)))-) = E 1 H I " .
."

data 02 05 2020
S T Q Q S [✓ S] D

Pag. 71 (101)

Item 107. To find the relations between the coordinates of plane an infinite distance.

outra fonte:
Books_77_2020/
Treatise-of-Solid_Geometry_
Frost/
→ Diretório todo!!!

outro livro (Frost)
An Elementary Treatise on Curve Tracing
(2004...)..djvu.

c) Gaussian curvature
Source: youtube.com/playlist?list=PLF566b2BA6693237E

Copy: Books_76_2020/
Einstein the closet geocentrist videos/
a) Relativity_Videos/
b) Gauss/ → Relativity Z6.

02 05 2020

D) Books_2020_12/
Books_75_2020/
PI/

D1) 2002.12137.pdf
The calculation of p_n and $\pi(n)$
Simon Plouffe
Feb 23. 2020.

Source: plouffe.fr/Simon Plouffe.htm

D2) Une formule pour les nombres premier II.pdf
Le calcul de p_n et $\pi(n)$
Feb 21, 2020

Source: plouffe.fr/NEW/Une formule pour les nombres premiers II.pdf

Obs: outros arquivos:
plouffe.fr/simon/

New file: Books_77_2020/
Simon_Plouffe/
0912.0303.pdf

Source: plouffe.fr/simon/0912.0303.pdf

data 02.05.2020

On the computation of the n^{th} decimal digit of various transcendental numbers

Pag. 02

The formula for π used is

$$\pi + 3 = \sum_{n=1}^{\infty} \frac{n 2^n n!^2}{(2n)!}$$

E) Einstein Space-Time theories

Source: https://www.nasa.gov/mission_pages/gpb/

→ Nasa's Gravity Probe B confirms two Einstein Space-Time theories.

other: https://einstein.stanford.edu

https://arxiv.org/abs/1105.3456

https://iopscience.iop.org/journal/0264-9381/page/Focus-issue-on-Gravity-Probe-B

02 05 2020

On the computation of the n^{th} decimal digit of various transcendental numbers

Pág. 02

The formula for π used is

$$\pi + 3 = \sum_{n=1}^{\infty} \frac{n 2^n n!^2}{(2n)!}$$

E) Einstein Space-time theories

Source: https://www.nasa.gov/mission_pages/gpb/

→ Nasa's Gravity Probe B confirms two Einstein Space-Time theories.

other: https://einstein.stanford.edu

https://arxiv.org/abs/1105.3456

https://iopscience.iop.org/journal/0264-9381/page/Focus-issue-on-Gravity-Probe-B

02.05.2020

Press Conference:
https://www.youtube.com/watch?
v=SBiYOFn1ze4

→ Einstein Passes Tests by NASA's
Gravity Probe B

Copy: Books_76_2020/
Einstein the Closet
Geocentrist_Videos/
Einstein Passes...mp4

---

160) 20200502

A) Gaussian Quadrature Formulas

Source: Books_77_2020/
Gaussian_Curvature/

A1) Gaussian quadrature formulas...djvu

A2) Gaussian Harmonic Analysis (2019)....pdf

B) Books_77_2020/
Gravity/

B1) Beyond Einstein Gravity...pdf

B2) IMPORTANT:

J"

>--? / / / '@)>--? / /)> A /'3/)B 7
9 / ")

)*+,)))-) = B 7 25 0* *K
9 1 5 2 > .!

& 3
9 , "

> A /"/ $9 / 1 $

02 05 2020

Press Conference:
https://www.youtube.com/watch?
v=5BiYOFn1ze4
→ Einstein Passes Tests by NASA's
Gravity Probe B

Copy: Books_76_2020/
Einstein the Closet
Geocentrist_Videos/
Einstein Passes...mp4

---

160) 20200502
A) Gaussian Quadrature Formulas
Source: Books_77_2020/
Gaussian_Curvature/
A1) Gaussian quadrature formulas...djvu
A2) Gaussian Harmonic Analysis
(2019)....pdf
B) Books_77_2020/
Gravity/
B1) Beyond Einstein Gravity...pdf
B2) IMPORTANT:

02.05.2020

B2) Gravity, a Geometrical Course
2 Volumes
Pietro Giuseppe Frè

B2.1) Gravity, a Geometrical Course, Volume 1...pdf

B2.2) Gravity, a Geometrical Course, Volume 2...pdf

B3) A. Zee - Einstein gravity a nut shell (2013...) a...pdf

Pag. 25(48)

Part I From Newton to Riemann: Coordinates to Curvature

B4) Foundations of Space and Time - Reflections...pdf

B5) Equations of Motion in Relativistic Gravity (2015...)...pdf ✡

B6) Gauge_Gravity_Foundations...pdf

B7) Gravity's Shadow...pdf

B8) Important:

02 05 2020

B8) Quantum Gravity (1997...)..pdf

Pág. 542(571)

On the observable properties of cosmic background radiation in the Universe with negative curvature.

B9) Gravity and Strings (2015...)...pdf

Pág. 09(36)

1.3 Metric Spaces.

1.3.1 Riemann-Cartan spacetime Ud

1.3.2 Einstein-Weyl spacetime EWd

1.3.3 Riemann spacetime Vd

Pág. 166(193)

4.5 Gravity as a gauge theory

Pág. 323(350)

12.2 The Einstein-Maxwell System

B10) Semiclassical and Stochastic Gravity – Quantum Field Effects on Curved Spacetime (2020)....pdf

☆. Important: 615 pages.

data 02.05.2020

Pág. 337(352) Chapter 10
The Einstein-Langevin Equation

Pág. 364(379) Chapter 11
Metric Fluctuations in Minkowski Spacetime.

Obs: Os dois capítulos tem vínculo (Langevin Equation)

Pág. 467(482) Chapter 15 e seguintes
Stress-Energy Tensor Fluctuations in de Sitter Space.

Pág. 519 (534) Chapter 17
Riemann Tensor Correlator in de Sitter

B11) Quantum Field Theory in Curved Spacetime – Quantized Fields and Gravity (2009...)...pdf ✡

Obs: Livro todo!!!

Pág. 01(19) Chapter 1
Quantum fields in Minkowski spacetime

Pág. 36(54) Chapter 2
Basics of quantum fields in curved spacetimes.

02 05 2020

~~A)~~ ~~B)~~ C) Modern Physics.

Source: Books_77_2020/
Modern Physics, by KAUR AND PICKRELL/

C1) 9332901813.pdf

Obs: Livro consulta. Ler todo!

Pág. 17(43) 1.7 Lorentz Contraction (Length Contraction)

The measured distance also depends on the reference frame.

If we consider Galilean transformation, then the measured size an object remains unaffected by the relative motion of observers. But for

LORENTZ TRANSFORMATION,

length (length) varies as time interval varies.

Pág. 18 (44) 1.8 Concept of Space-Time (Minkowski Space)

Pág. 20 (46) Fig 1.9 Space-Time diagram.

C2) Geometric Mechanics on Riemannian Manifolds... pdf

→ Applications to Partial Differential Equations

02 05.2020

Pág. 251(257) Chapter 12
Mechanical Curves _

D) Joseph Edwards _ A Treatise on the Integral Calculus with Applications ... Volume 1 ... pdf
2 pdf

Source: Books_77_2020/
Calculus with Applications/

E) Books_77_2020/
Understanding Analysis/

Understanding Analysis (2015...)...pdf

Pág. 85(97) Chapter 3
Basic Topology of R
→ the Cantor Set

Pág. 215(224) Chapter 7
The Riemann Integral

F) ARQ_IMPORTANT_05/
Géométrie/

→ diretório todo!

→ Traité de Géométrie (2 Volumes)

obs. Está incompleto. Vou consertar!!! OK agora

#$ $ % # !

3
9 1
)*+,)))-) 3 9 L "
2 =33: M "N1
.
F O $> A $9 , " $

)*+,)))-) 3 9 M * 5 =:
. 2 ."

2 > A /"/ $9 1 5$

2)*+,)))-) 3 > 5 " B 9 1 5/ * 1 5
9 1 7 5 " * . 5 : . 0 " ."

2 & .

2)*+,)))-) 3 9 1 5: 9
> A M : 5 " & " . 1 5 : . ."

2)*+,)))-) 3 9 1 5: 9
> A M : 5 " & " . 1 5 : . ."

2#)*+,)))-)9 1 5: 9
> A M : 5 " & " . 1 5 : . ."
(!% &

1.1 Newton's Laws

The foundational equation of our subject

> For in those days I was in the prime of my age for invention and minded Mathematicks & Philosophy more than at any time since.
>
> —Newton describing his youth in his memoirs

Let us start with one of Newton's laws, which curiously enough is spoken as $F = ma$ but written as $ma = F$. For a point particle moving in D-dimensional space with position given by $\vec{x}(t) = (x^1(t), x^2(t), \cdots, x^D(t))$, Mr. Newton taught us that

$$m\frac{d^2x^i}{dt^2} = F^i \tag{1}$$

with the index* $i = 1, \cdots, D$. For $D \leq 3$ the coordinates have traditional "names": for example, for $D = 3$, x^1, x^2, x^3 are often called, with some affection, x, y, z, respectively.

Bad notation alert! In teaching physics, I sometimes feel, with only slight exaggeration, that students are confused by bad notation almost as much as by the concepts. I am using the standard notation of x and t here, but the letter x does double duty, as the position of the particle, which more strictly should be denoted by $x^i(t)$ or $\vec{x}(t)$, and as the space coordinates x^i, which are variables ranging from $-\infty$ to ∞ and which certainly are independent of t.

The different status between x and t in say (1) is particularly glaring if $N > 1$ particles are involved, in which case we write $m\frac{d^2{x^i}_a}{dt^2} = {F^i}_a$ or $m\frac{d^2\vec{x}_a}{dt^2} = \vec{F}_a$ with ${x^i}_a(t)$ for $a = 1, 2, \cdots, N$. But certainly t_a is a meaningless concept in Newtonian physics. In the Newtonian universe, t is the time ticked off by a universal clock, while $\vec{x}_a(t)$ is each particle's private business. We will have plenty more to say about this point. Here ${x^i}_a(t)$ are $3N$ functions of t, but there are still only 3 x^i.

---

* See appendix 2.

Some readers may feel that I am overly pedantic here, but in fact this fundamental inequality of status between x and t will come to a head when we get to the special theory of relativity. (I now drop the arrow on $\vec{x}$.) Perhaps Einstein as a student was bothered by this bad notation. One way to remedy the situation is to use q (or q_a) to denote the position of particles, as in more advanced treatments. But here I bow to tradition and continue to use x.

Have differential equation, will solve

After Newton's great insight, we "merely" have to solve some second order differential equations.

To understand Newton's fabulous equation, it's best to work through a few examples. (I need hardly say that if you do not already know Newtonian mechanics, you are unlikely to be able to learn it here.)

A priori, the force F^i could depend on any number of things, but from experience we know that in many simple cases, it depends only on x and not on t or $\frac{dx}{dt}$. As physicists unravel the mysteries of Nature, it becomes increasingly clear that fundamental forces are derived from an underlying quantum field theory and that they have simple forms. Complicated forces often merely result from some approximations we make in particular situations.

Example A

A particle in 1-dimensional space tied to a spring oscillates back and forth.

The force F is a function of space. Newton's equation

$$m\frac{d^2x}{dt^2} = -kx \tag{2}$$

is easily solved in terms of two integration constants: $x(t) = a\cos\omega t + b\sin\omega t$, with $\omega = \sqrt{\frac{k}{m}}$. The two constants a and b are determined by the initial position and initial velocity, or alternatively* by the initial position at $t = 0$ and by the final position at some time $t = T$. Energy, but not momentum, is conserved.

Example B

We kick a particle in 1-dimensional space at $t = 0$.

The force F is a function of time. This example allows me to introduce the highly useful Dirac[1] delta function, or simply delta function.[2] By the word "kick" we mean that the time scale τ during which the force acts is much less than the other time scales we are

---

* See part II.

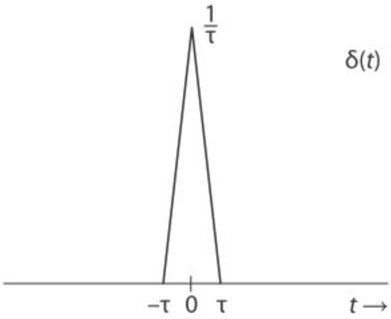

Figure 1 The delta function, which could be thought of as an infinitely sharp spike, is strictly speaking not a function, but the limit of a sequence of functions.

interested in. Thus, take $F(t) = w\delta(t)$, where the function $\delta(t)$ rises sharply just before $t = 0$, rapidly reaches its maximum at $t = 0$, and then sharply drops to 0. Because we included a multiplicative constant w, we could always normalize $\delta(t)$ by

$$\int dt\ \delta(t) = 1 \tag{3}$$

As we will see presently, the precise form of $\delta(t)$ does not matter. For example, we could take $\delta(t)$ to rise linearly from 0 at $t = -\tau$, reach a peak value of $1/\tau$ at $t = 0$, and then fall linearly to 0 at $t = \tau$. For $t < -\tau$ and for $t > \tau$, the function $\delta(t)$ is defined to be zero. Take the limit $\tau \to 0$, in which this function is known as the delta function. In other words the delta function is an infinitely sharp spike. See figure 1.

The δ function is somehow treated as an advanced topic in mathematical physics, but in fact, as you will see, it is an extremely useful function that I will use extensively in this book, for example in chapters II.1 and III.6. More properties of the δ function will be introduced as needed.

Integrating

$$\frac{d^2x}{dt^2} = \frac{w}{m}\delta(t) \tag{4}$$

from some time $t_- < 0$ to some time $t_+ > 0$, we obtain the change in velocity $v \equiv \frac{dx}{dt}$:

$$v(t_+) - v(t_-) = \frac{w}{m} \tag{5}$$

Note that in this example, neither energy nor momentum is conserved. The lack of conservation is easy to understand: (4) does not include the agent administering the kick. In general, a time-dependent force indicates that the description is not dynamically complete.

Example C

A planet approximately described as a point particle of mass m goes around its sun of mass $M \gg m$.

This is of course the celebrated problem Newton solved to unify celestial and terrestrial mechanics, previously thought to be two different areas of physics. His equation now reads

$$m\frac{d^2\vec{r}}{dt^2} = -GMm\frac{\vec{r}}{r^3} \tag{6}$$

where we use the notation $\vec{r} = (x, y, z)$ and $r = \sqrt{\vec{x} \cdot \vec{x}} = \sqrt{x^2 + y^2 + z^2}$.

John Wheeler has emphasized the interesting point that while Newton's law (1) tells us how a particle moves in space as a function of time, we tend to think of the trajectory of a particle as a curve fixed in space. For example, when we think of the motion of a planet around the sun, we think of an ellipse rather than a spiral around the time axis. Even in Newtonian mechanics, it is often illuminating to think in terms of a spacetime picture rather than a picture in space.[3]

Newton and his two distinct masses

> By thinking on it continually.
> —Newton (reply given when
> asked how he discovered
> the law of gravity)

Conceptually, in (6), m represents two distinct physical notions of mass. On the left hand side, the inertial mass measures the reluctance of the object to move. On the right hand side, the gravitational mass measures how strongly the object responds to a gravitational field. The equality of the inertial and the gravitational mass was what Galileo tried to verify in his famous apocryphal experiment dropping different objects from the Leaning Tower of Pisa. Newton himself experimented with a pendulum consisting of a hollow wooden box, which he proceeded to fill with different substances, such as sand and water. In our own times, this equality has been experimentally verified[4,5] to incredible accuracy.

That the same m appears on both sides of the equation turns out to be one of the greatest mysteries in physics before Einstein came along. His great insight was that this unexplained fact provided the clue to a deeper understanding of gravity. At this point, all we care about this mysterious equality is that m cancels out of (6), so that $\ddot{\vec{r}} = -\kappa\frac{\vec{r}}{r^3}$, with $\kappa \equiv GM$.

Celestial mechanics solved

Since the force is "central," namely it points in the direction of $\vec{r}$, a simple symmetry argument shows that the motion is confined to a plane, which we take to be the $(x\text{-}y)$ plane. Set $z = 0$ and we are left with

$$\ddot{x} = -\kappa x/r^3 \quad \text{and} \quad \ddot{y} = -\kappa y/r^3 \tag{7}$$

I have already, without warning, switched from Leibniz's notation to Newton's dot notation

$$\dot{x} \equiv \frac{dx}{dt} \quad \text{and} \quad \ddot{x} \equiv \frac{d^2x}{dt^2} \tag{8}$$

Since this is one of the most beautiful problems[6] in theoretical physics, I cannot resist solving it here in all its glory. Think of this as a warm-up before we do the heavy lifting of learning Einstein gravity. Also, later, we can compare the solution here with Einstein's solution.

Evidently, we should change from Cartesian coordinates (x, y) to polar coordinates (r, θ). We will do it by brute force to show, in contrast, the elegance of the formalism we will develop later. Differentiate

$$x = r\cos\theta \quad \text{and} \quad y = r\sin\theta \tag{9}$$

twice to obtain first

$$\dot{x} = \dot{r}\cos\theta - r\sin\theta\,\dot{\theta} \quad \text{and} \quad \dot{y} = \dot{r}\sin\theta + r\cos\theta\,\dot{\theta} \tag{10}$$

and then

$$\begin{aligned} &\ddot{x} = \ddot{r}\cos\theta - 2\dot{r}\sin\theta\,\dot{\theta} - r\cos\theta\,\dot{\theta}^2 - r\sin\theta\,\ddot{\theta} \\ \text{and}\quad &\ddot{y} = \ddot{r}\sin\theta + 2\dot{r}\cos\theta\,\dot{\theta} - r\sin\theta\,\dot{\theta}^2 + r\cos\theta\,\ddot{\theta} \end{aligned} \tag{11}$$

(Note that in each pair of these equations, the second could be obtained from the first by the substitution $\theta \to \theta - \frac{\pi}{2}$, so that $\cos\theta \to \sin\theta$, and $\sin\theta \to -\cos\theta$.)

Multiplying the first equation in (7) by $\cos\theta$ and the second by $\sin\theta$ and adding, we obtain, using (11),

$$\ddot{r} - r\dot{\theta}^2 = -\frac{\kappa}{r^2} \tag{12}$$

On the other hand, multiplying the first equation in (7) by $\sin\theta$ and the second by $\cos\theta$ and subtracting, we have

$$2\dot{r}\dot{\theta} + r\ddot{\theta} = 0 \tag{13}$$

I remind the reader again that we are doing all this in a clumsy brute force way to show the power of the formalism we are going to develop later.

After staring at (13) we recognize that it is equivalent to

$$\frac{d}{dt}(r^2\dot{\theta}) = 0 \tag{14}$$

which implies that

$$\dot{\theta} = \frac{l}{r^2} \tag{15}$$

for some constant l. Inserting this into (12), we have

$$\ddot{r} = \frac{l^2}{r^3} - \frac{\kappa}{r^2} = -\frac{dv(r)}{dr} \tag{16}$$

where we have defined

$$v(r) = \frac{l^2}{2r^2} - \frac{\kappa}{r} \tag{17}$$

Multiplying (16) by $\dot{r}$ and integrating over t, we have

$$\int dt\, \frac{1}{2}\frac{d}{dt}\dot{r}^2 = \int dt\, \dot{r}\ddot{r} = -\int dt \frac{dr}{dt}\frac{dv(r)}{dr} = -\int dr \frac{dv(r)}{dr}$$

so that finally

$$\frac{1}{2}\dot{r}^2 + v(r) = \epsilon \tag{18}$$

with ϵ an integration constant.

This describes a unit mass particle moving in the potential $v(r)$ with energy ϵ. Plot $v(r)$. Clearly, if ϵ is equal to the minimum of the potential $v_{\text{min}} = -\frac{\kappa^2}{2l^2}$, then $\dot{r} = 0$ and r stays constant. The planet follows a circular orbit of radius l^2/κ. If $\epsilon > v_{\text{min}}$ the orbit is elliptical, with r varying between r_{min} (perihelion) and r_{max} (aphelion) defined by the solutions to $\epsilon = v(r)$. For $\epsilon > 0$ the planet is not bound and should not even be called a planet.

We have stumbled across two conserved quantities, the angular momentum l and the energy ϵ per unit mass, seemingly by accident. They emerged as integration constants, but surely there should be a more fundamental and satisfying way of understanding conservation laws. We will see in chapter II.4 that there is.

Orbit closes

One fascinating apparent mystery is that the orbit closes. In other words, as the particle goes from r_{min} to r_{max} and then back to r_{min}, θ changes by precisely 2π. To verify that this is so, solve (18) for $\dot{r}$ and divide by (15) to obtain $\frac{dr}{d\theta} = \pm(r^2/l)\sqrt{2(\epsilon - v(r))}$. Changing variable from r to $u = 1/r$, we see, using (17), that $2(\epsilon - v(r))$ becomes the quadratic polynomial $2\epsilon - l^2u^2 + 2\kappa u$, which we can write in terms of its two roots as $l^2(u_{\text{max}} - u)(u - u_{\text{min}})$. Since u varies between u_{min} and u_{max}, we are led to make another change of variable from $u = u_{\text{min}} + (u_{\text{max}} - u_{\text{min}})\sin^2\zeta$ to ζ, so that ζ ranges from 0 to $\frac{\pi}{2}$. Thus, as the particle completes one round trip excursion in r, the polar angle changes by (note that $u_{\text{min}} = 1/r_{\text{max}}$ and $u_{\text{max}} = 1/r_{\text{min}}$)

$$\begin{aligned}\Delta\theta &= 2\int_{r_{\text{min}}}^{r_{\text{max}}} \frac{l\,dr}{r^2\sqrt{2(\epsilon - v(r))}} = 2\int_{u_{\text{min}}}^{u_{\text{max}}} \frac{l\,du}{\sqrt{2\epsilon - l^2u^2 + 2\kappa u}} \\ &= 2\int_{u_{\text{min}}}^{u_{\text{max}}} \frac{du}{\sqrt{(u_{\text{max}} - u)(u - u_{\text{min}})}} = 4\int_0^{\frac{\pi}{2}} d\zeta = 2\pi\end{aligned} \tag{19}$$

That this integral turns out to be exactly 2π is at this stage nothing less than an apparent miracle. Surely, there is something deeper going on, which we will reveal in chapter I.4. Note also that the inverse square law is crucial here. Incidentally, the change of variable here indicates how the Newtonian orbit* (and also the Einsteinian orbit, as we will see in part VI) could be determined. See exercise 2.

Bad notation alert! In (1), the force on the right hand side should be written as $F^i(x(t))$ (in many cases). In C, the gravitational force exists everywhere, namely $F(x)$ exists as a function, and what appears in Newton's equation is just $F(x)$ evaluated at the position of the particle $x(t)$. In contrast, in A, with a mass pulled by a spring, $F(x)$ does not make sense, only $F(x(t))$ does. The force exerted by the spring does not pervade all of space, and hence is defined only at the position of the particle $x(t)$, not at any old x. I can practically hear the reader chuckling, wondering what kind of person I could be addressing here, but believe me, I have encountered plenty of students who confuse these two basic concepts: spatial coordinates and the location of particles. I may sound awfully pedantic, but when we get to curved spacetime, it is often important to be clear that certain quantities are defined only on so-called geodesic curves, while others are defined everywhere in spacetime.

A historical digression on the so-called Newton's constant

> Wouldn't we be better off with the two eyes we now have plus a third that would tell us what is sneaking up behind? . . . With six eyes, we could have precise stereoscopic vision in all directions at once, including straight up. A six-eyed Newton might have dodged that apple and bequeathed us some levity rather than gravity.
>
> —George C. Williams[7]

Physics textbooks by necessity cannot do justice to physics history. As you probably know, in the *Principia,* Newton (1642–1727) converted his calculus-based calculations to geometric arguments,[8] which most modern readers find rather difficult to follow. Here I want to mention another curious point: Newton never did specifically define what we call his constant G. What he did with $ma = GMm/r^2$ was to compare the moon's acceleration with the apple's acceleration: $a_{\text{moon}} R^2_{\text{lunar orbit}} = GM_{\text{earth}} = a_{\text{apple}} R^2_{\text{radius of earth}}$. But to write $GM_{\text{earth}} = a_{\text{apple}} R^2_{\text{radius of earth}}$, he had to prove what is sometimes referred to as the first of Newton's two "superb theorems," namely that with the inverse square law the gravitational force exerted by a spherical mass distribution acts as if the entire mass were concentrated in a point at the center of the distribution. (See exercise 4.) Even with his abilities, Newton had to struggle for almost 20 years, the length of which contributed to the bitter priority fight he had with Hooke on the inverse square law, with Newton claiming that he had the law a long time before publication. You should be able to do it faster by a factor of $\sim 10^4$ as an exercise.

---

* On the old one pound note, a portrait of Newton together with his orbits appears on the back. Amusingly, the artist felt compelled to put the sun at the center, rather than one of the foci, of the ellipse.

Knowing the moon's period and $R_{\text{lunar orbit}}$, Newton could calculate a_{moon}. Since $R_{\text{radius of earth}}$ had been known since antiquity, he was thus able to calculate a_{apple} and obtained agreement* with Galileo's measurement of a_{apple}. This of course represents one of the most magnificent advances in physics history, with Newton unifying[9] the previously disparate subjects of celestial and terrestrial mechanics in one stroke. I don't have space to dwell on this here, but I do want to call your attention to the fact that Newton did not need to know G and M_{earth} to perform his feat.

Indeed, G was not measured until 1798 by Henry Cavendish (1731–1810) using equipment built and designed by his friend John Michell (1724–1793), now of black hole fame, who died before he could carry out the experiment.

Needless to say, what I have presented here should only be regarded as a comic book version of history.

Appendix 1: Where is hell?

You will find it in this appendix, sort of.

Curiously, contrary to what some textbooks and popular books stated, Cavendish's goal was not to measure G, but M_{earth} and hence the earth's density. Why this was of more interest to physicists of the time than G is in itself another interesting tidbit in physics history.

I mentioned that Newton had two superb theorems and that the first triggered his feud with Hooke. His second superb theorem states that there is no gravitational force inside a spherical shell.[10] Are you curious why Newton would even attack such a problem? An erroneous calculation had convinced him that the earth was much less dense than the moon, which led his friend Edmond Halley (1656–1742), who by the way published the *Principia* at his expense, to propose the hollow earth theory.[11] Witness the popularity of the idea in science fiction, notably Jules Verne's *Journey to the Center of the Earth* (1864). The idea may seem absurd to us, but at that time, a location for hell had to be found, and leading physicists gave serious thought to this pressing problem. Every epoch in physics has its own top ten problems.

So now we understand Cavendish's interest in M_{earth} and hence in the density of the earth rather than in G. Some textbooks give the impression that people easily obtained M_{earth} by multiplying the density of rock and the volume of the earth. Not so easy if you think that the earth might be hollow! We learn from Newton's second theorem that there is no gravitational force in hell, so the usual portrayal of the leaping flames can't be right!

Appendix 2: Fear of indices

Occasionally, a student or two would profess, unaccountably, a "fear of indices." In fact, there is nothing to fear.[12] At this stage, just stand back and admire how clever the invention of indices is. Instead of giving names to each coordinate axis, such as x, y, and z, we could pass fluidly between different dimensions by writing x^i, with $i = 1, 2, \cdots, D$. The length of the alphabet we use does not limit us, and we could easily go beyond 26 dimensions.

When we get to Einstein's theory, there will be a flood of indices, and we will have to distinguish between upper and lower indices. In Newtonian mechanics, there is no significance to whether we write the index as a superscript or a subscript. Have no fear: we will discuss each of these features of indices when the need arises. At this point, we merely note that a quantity can carry more than one index. In the text, we wrote ${x^i}_a$, with $i = 1, 2, \cdots, D$ labeling the different spatial directions, and $a = 1, 2, \cdots, N$ labeling the different particles. We will encounter more examples as we go along.

---

* Newton's first try did not lead to excellent agreement, because the value for the earth's equatorial radius was off. Just a reminder that physics never progresses as smoothly as textbooks say.

With only slight exaggeration, we could say that the invention of indices represents one of the really clever ideas[13] in the history of physics and mathematics, almost a "magic trick" that enables us to deal with as many particles in as many spatial dimensions as we like with the mere addition of some subscripts and superscripts.

Exercises

1 Show that for some suitably smooth function $f(x)$, the integral $\int_{-\infty}^{\infty} dx\delta(x)f(x) = f(0)$. Then show that $\delta(ax) = \delta(x)/|a|$ by evaluating the integral $\int_{-\infty}^{\infty} dx\delta(ax)f(x)$ for some smooth function $f(x)$.

2 Determine the orbit $r(\theta)$ by changing variable from r to $u = 1/r$. We will need the result of this exercise later.

3 Newton thought that light consists of "corpuscles." Calculate the deflection of light by the sun, applying what you learned in the text to the case $\epsilon > 0$. Note that the mass of these minute "particles of light" drops out in Newtonian theory anyway. We will need this result to compare with Einstein's theory later in chapter VI.3.

4 Prove Newton's first superb theorem: the gravitational force exerted by a spherical mass distribution acts as if the entire mass were concentrated in a point at the center of the distribution.

5 Prove Newton's second superb theorem.

6 Suppose engineers can build a straight tunnel connecting two cities on earth. Then we could have a free unpowered "gravity express"[14] by simply dropping a railroad car into the tunnel, allowing it to fall from one city to the other. Use Newton's two superb theorems to calculate the transit time.

Notes

1. Also introduced by Cauchy, Poisson, Hermite, Kirchoff, Kelvin, Helmholtz, and Heaviside. See J. D. Jackson, *Am. J. Phys.* 76 (2008), pp. 707–709.
2. Rigorous mathematicians go berserk at physicists' use of the word "function" here; they prefer to call it a distribution, defined as the limit of a function. But working physicists do not give a flying barf about such niceties. In any case, I do not personally know a theoretical physicist suffering any harm by calling $\delta(t)$ a function.
3. Consider a game of tennis. Compare a hard drive down the line and a soft lob high over the net. In both cases, we are to solve Newton's law $\frac{d^2x}{dt^2} = 0$, $\frac{d^2y}{dt^2} = -g$, with the boundary conditions $x(0) = 0$, $x(T) = L$, and $y(0) = y(T) = 0$. (The problem is so elementary that we won't bother to explain the notation, that y denotes the vertical direction, that $y = 0$ is the ground, that T is the time of flight before the ball hits the ground, that L is the length of the tennis court, and so on and so forth. You might want to draw your own figure.) The solution is $x = Lt/T$, $y = \frac{1}{2}g(T - t)t$. Note that the two types of tennis shots are governed by the same equation and the same L. Hence we obtain the same solution, but keep in mind that T is small in the case of the hard drive and that T is large in the case of the soft lob. Now eliminate t to obtain y as a function of x, namely $y(x) = \frac{1}{2}gT^2(1 - \frac{x}{L})\frac{x}{L}$, a parabola in both cases (of course). But compare the curvature of the two parabolas: we have $\frac{d^2y}{dx^2} = -g(T/L)^2$, very small in the case of the hard drive (small T) and very large in the case of the lob (large T). The hard drive down the line barely skimming over the net, and the soft lob climbing lazily high up into the sky, look and feel totally different pictured in space. In contrast, consider y as a function of t. We also have two parabolas (of course), namely $y(t) = \frac{1}{2}g(T - t)t$, as given earlier. Now compare the curvature of the two parabolas: we have $\frac{d^2y}{dt^2} = -g$, the same in both cases. The curvature of the ball's trajectory in spacetime is universal (universal gravity, get it?). But we tend to see in our mind's eye the two parabolas $y(x)$ in space, one for the hard drive and one for the lob, which look quite different, rather than the parabolas $y(t)$ in spacetime, which have the same curvature. I learned this long ago from John Wheeler.
4. Currently to one part in 10^{13}. The modern round of experiments started with Loránd Eötvös in 1885 and continues with the Eöt-Wash experiment led by E. Adelberger in our days.

5. The equality of the gravitational and inertial mass of the neutron has also been verified to good accuracy using neutron interferometry.
6. For Newton's letter to Halley about Hooke on the inverse square, see P. J. Nahin, *Mrs. Perkins's Electric Quilt,* Princeton University Press, 2009.
7. G. C. Williams, *The Pony Fish's Glow,* Basic Books, 1997, p. 128.
8. S. Chandrasekhar, *Newton's Principia for the Common Reader,* Oxford University Press, 2003.
9. *Fearful,* pp. 74–75.
10. For a popular account, see *Toy/Universe*.
11. N. Kollerstrom, "The Hollow World of Edmond Halley," *J. Hist. Astronomy* 23 (1992), p. 185.
12. Surely most readers are familiar with indices. My son the biologist informs me that even biologists use indices routinely; for example, on p. 20 of *Genetics and Analysis of Quantitative Traits* by M. Lynch and B. Walsh, indices appear without explanation or apology.
13. A colleague told me to mention that indices are crucial in computer programming, something that many readers can relate to.
14. *Toy/Universe,* p. xxix.

2!)*+,)))-) 3 8 " . " 7
+ , 9 1 5 : 2 " E 1 5 ."

2)*+,)))-) 3 BL + 1
9 1 5 : . ."

23)*+,)))-) 3 9 /9 1 5 F 5
8 " " *.. : 2 " E 1 5 ."

2')*+,)))-) 3 9 1 5K "
1 1 !: E 1 5 ."

2% & .

)*+,)))-) 3 , 9 1 5 ==': P "
2 ; " ."
(! '

ON THE OBSERVABLE PROPERTIES OF COSMIC BACKGROUND RADIATION IN THE UNIVERSE WITH NEGATIVE CURVATURE

V.G.Gurzadyan‡and A.A.Kocharyan#

‡ Department of Theoretical Physics, Yerevan Physics Institute, 375036 Yerevan, Armenia (permanent address); Astronomy Centre, University of Sussex, Brighton, UK

Department of Mathematics, Monash University , Clayton, Victoria 3168, Australia

Abstract.– As was shown before the statistical properties (exponential mixing) of motion of CMB photon beams in Friedmann Universe with negative curvature can have definite observable consequences including the decrease of the CMB anisotropy amplitude after the last scattering, distortion of images on CMB sky maps. Here we consider one more consequence concerning the behaviour of temperature autocorrelation function $C(\theta,\beta)$, where θ is the separation angle, and β the observing beam effective angle. We prove that in the Universe with negative curvature the temperature autocorrelation function should tend to become independent on separation angle regardless the initial form of the spectrum of perturbations at last scattering epoch. It is also predicted that the temperature autocorrelation function should increase while obtained from measurements with small beam sizes within some *mixing beam angle.*

The discovery of anisotropy of the Cosmic Microwave Background radiation (CMB) announced the beginning of far more informative studies on the evolution and the present structure of the Universe [1]. The first and most popular viewpoint of COBE and other available data (though slightly revised by later analysis) concerning in particular the temperature autocorrelation function $C(\theta,\beta)$ was the support of Harrison-Zeldovich scale-invariant spectrum of initial fluctuations and hence of inflationary scenario [2].

Below we consider another possible interpretation of the same data, namely as indication of the negative curvature of the Universe. Our approach is based on the studies on properties of photon beam motion in Friedmann-Robertson-Walker Universe with $k=-1$, which can be reduced to the problem of the behaviour of correlation function of geodesic flow on 3-manifold with negative constant curvature. In particular the study of the exponential decay of correlation functions in the context of CMB problem has led to the prediction of the following observable effects [3,4,5]: a) decrease of CMB anisotropy after the last scattering epoch depending on the value of the density parameter Ω; b) distortion of images of anisotropy structures on CMB maps; c) possibility of measuring of similar properties of CMB photon beams arriving from regions having no causal connection.

The first effect as follows from the calculations [3,4], can lead to a damping of $\delta T/T$ since last scattering epoch, $z=1000$, up to the present time by a rather significant factor depending on Ω (up to $100-1000$ at $\Omega\sim 0.1-0.3$).

The second effect can lead to complex-elongated structures of the constant temperature curves on CMB maps. As a simplified indicator, the degree of elongation, if measured on CMB maps can give the present value of Ω by the formula derived in [5]. The preliminary statistical analysis of COBE-DMR maps [6], 1993) does reveal existence of elongated shapes; if attributed to the described effect the results indicate $\Omega = 0.2 - 0.4$; special pattern-recognition code was developed for that study [7] to distinguish from other map properties [8]. The third effect, equivalent to the fact that *two ink drops are mixing in a cup of water in the same manner without requiring information on each other*, removes the mystery of the horizon problem.

All these three effects occur only in Friedmann Universe with negative curvature (i.e. they disappear if either $k = 0$ or $k = +1$), and therefore can be of particular importance for probing of this principal property of the Universe.

Evidently nothing is happening to any individual photon during free propagation after the last scattering, and these effects are purely statistical and are determined by the principal limitations of obtaining information during measurements, i.e. by the impossibility of reconstruction of the trajectory of given photon while observing within finite smoothing angle due to the overlapping of exponentially deviating geodesics in any cut of phase space. This situation of loss of information can be described via *random sequences* of Kolmogorov [9] and Chaitin [10].

Below we consider another aspect of the problem of exponential decay of correlators of geodesic flow having particular observational consequences concerning the properties of temperature autocorrelation function $C(\theta, \beta)$ [11].

Our present study is essentially determined by recent rigorous results on the correlation functions of geodesic flow on 3-manifold with negative curvature [12]. Geodesic flow on d-dimensional manifolds with negative curvature has been object of intense studies rather long time, since Hadamard (1898), Hedlund, Hopf (30ies) up to fundamental results by Anosov and others. In particular, it was proved that geodesic flow on closed (compact, without boundary) manifold with negative curvature is Anosov system [13], is possessing the strongest chaotic properties (e.g. mixing of all degrees), with positive Kolmogorov-Sinai (KS) entropy. Rather informative characteristics of any dynamical system are the correlation functions, in particular the law of their decay, though the obtaining of rigorous results for each given system usually requires great efforts and in fact for only a few classes of dynamical systems results on correlation functions exist up to date. The new result [12] being of remarkable interest for us reads: the correlation function

$$b_{A_1,A_2}(\lambda) = \int_{SM} A_1 \circ f^{\lambda} \cdot A_2 d\mu - \int_{SM} A_1 d\mu \int_{SM} A_2 d\mu$$

of the geodesic flow $\{f^{\lambda}\}$ on the unit tangent bundle SM of the Riemann 3-manifold M with negative constant curvature decreases by exponential law, namely for all functions $A_1, A_2 \in L^2(SM)$ but a finite-dimensional space in $L^2(SM)$ there exists $c > 0$ such that

$$|b_{A_1,A_2}(\lambda)| \leq c \cdot |b_{A_1,A_2}(0)| \cdot e^{-h\lambda} , \tag{1}$$

where μ is the Liouville measure and $\mu(SM) = 1$, h is the KS-entropy of the geodesic flow $\{f^{\lambda}\}$.

We now take up the question of how one describes the free motion of photons in the $4D$-World (for details see [4]). To make matters more simple we will only consider a Friedmann $4D$-World W, decomposed into a homogeneous $3D$-Universe characterized by some coordinates x, evolving with respect to the cosmic time t. Thus each event in the World can be assigned unique space-time coordinates (x, t). One can then project any trajectory γ from the World W into the Universe U simply by associating to $\gamma(\lambda) = (x(\lambda), t(\lambda))$ the curve $c(\lambda) = x(\lambda)$. It turns out that null geodesics in the World project onto geodesics in the Universe with new affine parameter:

$$\lambda(t) = \int_{t_0}^{t} \frac{ds}{a(s)} .$$

The latter relates to the problem of "internal time" of the K-systems in cosmology considered initially in [14]. As it can be easily shown by means of the formulae in [3,4,15], the degree of smoothing of anisotropy $\delta T/T$ in post-scattering epoch is determined by the KS-entropy h and affine parameter λ which depend on the parameters of the Universe. For the matter-dominated Universe one has

$$e^{h\lambda} = (1+z)^2 \left[\frac{1+\sqrt{1-\Omega}}{\sqrt{1+z\Omega}+\sqrt{1-\Omega}} \right]^4 . \tag{2}$$

For $z \gg \Omega^{-1}$ one may obtain

$$e^{h\lambda} \sim \frac{(1+\sqrt{1-\Omega})^4}{\Omega^2} . \tag{3}$$

Here our aim is to investigate the behaviour of the temperature autocorrelation function $C(\theta, \beta)$ using properties of the correlation function of the geodesic flow $\{f^\lambda\}$.

One may readily see that if

$$A_1(u) = \mathcal{T}(u) - 1 = \frac{T(u)}{\bar{T}} - 1 ,$$

and

$$A_2(u) = \frac{\chi_{\mathcal{K}(v)}(u)}{\mu(\mathcal{K}(v))} - 1 ,$$

where $T(u)$ is the real temperature at decoupling time at $u \in SM$,

$$\bar{T} = \int_{SM} T d\mu ,$$

$\mathcal{K}(v) \subset SM$ is a beam of detected photons at $v \in SM$, and $\chi_{\mathcal{K}(v)}$ is the characteristic function of the set $\mathcal{K}(v)$, then normalized "measured" temperature at point u at λ defines as follows

$$\mathcal{T}_\lambda(u) = \frac{1}{\mu(\mathcal{K}(u))} \int_{\mathcal{K}(u)} \mathcal{T} \circ f^\lambda d\mu = 1 + b_{A_1,A_2}(\lambda) . \tag{4}$$

Therefore one may using Eqs. (1), (2) arrive at the following inequality

$$|\mathcal{T}_\lambda(u) - 1| \leq c \cdot |\mathcal{T}_0(u) - 1| \cdot \frac{1}{(1+z)^2} \cdot \left[\frac{\sqrt{1+z\Omega}+\sqrt{1-\Omega}}{1+\sqrt{1-\Omega}} \right]^4 .$$

Thus we see, that for any u, i.e. at any space point in any sky direction, *the "measured" temperature tends to the constant mean temperature, i.e. the isotropic state is the final state.*

For normalized temperature autocorrelation function

$$C_\lambda(\theta, \beta) = \langle \mathcal{T}_\lambda(u) \mathcal{T}_\lambda(v) \rangle_{g(u,v)=\cos\theta} , \tag{5}$$

the following inequality holds

$$|C_\lambda(\theta, \beta) - 1| \leq c \cdot |C_0(\theta, \beta) - 1| \cdot \frac{1}{(1+z)^2} \cdot \left[\frac{\sqrt{1+z\Omega}+\sqrt{1-\Omega}}{1+\sqrt{1-\Omega}} \right]^4 .$$

The following conclusion immediately follows from this formula :

Temperature autocorrelation function $C(\theta,\beta)$ *is almost constant with respect to* θ *at the present time regardless on the initial perturbations at* $z \sim 1000$.

Numerically the effect determined by Eq. (2) as was mentioned above is large enough for CMB photons [3,4], to reveal it during the evolution of the Universe due to negativity of curvature of the Universe.

Using Eq. (4) and the fact that the geodesic flow is an Anosov system one may show that there exists an angle ϕ such that smoothing factor s of $\delta T/T(\beta)$ is almost constant if either $\beta \gg \phi$ $(s \sim e^{-h\lambda})$ or $\beta \ll \phi$ $(s \sim 1)$, and increases as β is decreased at $\beta \sim \phi$ by the law $s \sim const/\beta \cdot e^{-h\lambda}$.

One can explain this result and the content of the *smoothing angle* ϕ as follows. As we have mentioned above, the smoothing of CMB is a statistical effect due to the loss of information during the measurements.while averaging within some beam size. Therefore, the more narrow is the beam size (pencil-beam), the less information is lost at smoothing in terms of the temperature autocorrelation function. Maximal information, e.g. the anisotropy corresponding exactly the last scattering surface should be obtained while measuring by beams within some mixing angle ϕ; measurements on small angles should not influence the autocorrelation function which should be determined solely by physical conditions of the last scattering surface (its thickness, the Silk effect, etc).

Among the measurements (for review see [1]) there are those with smaller beam size than COBE, such as of MAX, ($\delta T/T \leq 4.2 \cdot 10^{-5}$, beam size 0.5°) and of ARGO ($1.6 \cdot 10^{-5} < \delta T/T < 2.6 \cdot 10^{-5}$, beam size 0.8°(52′)) indicating increase of anisotropy compared with COBE and therefore seem to support the prediction made above. Less supporting is the result of South Pole ACME experiment, $\delta T/T \leq 1.4 \cdot 10^{-5}$, with 1.5° beam size. Evidently more accuracy and statistics is needed to distinguish our prediction from, say Doppler shift and predictions of certain CDM galaxy formation models.

This prediction, therefore indicates a way to obtain the sign of the curvature of the Universe, the real anisotropy at the last scattering epoch as well as the present value of Ω.

To summarize, the effects we are considering are model-independent, i.e. they do not depend on dark matter models, initial fluctuations, etc. The effects should appear if the Universe is Friedmann - homogeneous and isotropic, and has negative curvature. One has a real possibility of getting information directly on the curvature of the Universe, since the effects disappear at $k = 0$ and $k = +1$.

Acknowledgements

We thank I.Prigogine, A.Polnarev and D.W.Sciama for stimulating discussions, P. de Bernardis, F.Melchiorri and S.Torres for explanation of observational details.

References

[1] White M., Scott D., Silk J. *Ann.Rev.Astr.Ap.*, **32**, 319 (1994).

[2] Smoot G.F. et.al. *Astrophys.J.Lett.*, **396**, L1, (1992).

[3] Gurzadyan V.G., Kocharyan A.A. *Astr.Astrophys.* **260**, 14, (1992); in: *Quantum Gravity-V*, Eds. Markov M.A., Berezin V.A., Frolov V.P., World Sci. Singapore, 689, (1991).

[4] Gurzadyan V.G., Kocharyan A.A. *Int.Journ.Mod.Phys.D.*, **2**, 97, (1993); *Paradigms of the Large-Scale Universe*, Gordon and Breach, 1994.

[5] Gurzadyan V.G., Kocharyan A.A. *Europhys.Lett*, **22**, 231, (1993).

[6] Gurzadyan V.G., Torres S. in: *The Present and Future of the Cosmic Microwave Background* (Eds. Martinez-Gonzalez E., Cayon L., Sanz J.) Lecture Notes in Physics, **429**, Springer-Verlag, 1994.

[7] Torres S. preprint ICRA 12-3-93 (1993), University of Rome.

[8] Martinez-Gonzalez E., Cayon L. in: *The Infrared and Submillimetre Sky after COBE* (Eds. Signore M., Dupraz C.), Kluwer Academic Publ., 1992.

[9] Kolmogorov A.N. in *Problems of Information Transmission* **1**, 3, (1969).

[10] Chaitin G.J. *Information-Theoretic Completeness*, Worlds Scientific, 1992.

[11] Silk J. in: *The Infrared and Submillimetre Sky after COBE* (Eds. Signore M., Dupraz C.), Kluwer Academic Publ., 1992.

[12] Pollicott M. *Journ.Stat.Phys.*, **67**, 667, (1992).

[13] Anosov D.V. *Comm. Steklov Math.Inst.*, **90**, (1967).

[14] Lockhart C.M., Misra B., Prigogine I. *Phys.Rev.*, **D25**, 921, 1982.

[15] Peebles P.J.E. *Principles of Physical Cosmology*, Princeton University Press, 1993.

2=)*+,)))-) 3 9 1 5 " :
2 " E 1 5 ."
(= #3 #

1.3 Metric spaces

To go further we need to add structure to a manifold: a *metric* in tangent space, i.e. an inner product for tangent-space vectors (symmetric, bilinear) associating a function $g(\xi, \epsilon)$ with any pair of vectors (ξ, ϵ). This corresponds to a symmetric $(0, 2)$-type tensor g symmetric in its two covariant components $g_{\mu\nu} = g_{(\mu\nu)}$:

$$\xi \cdot \epsilon \equiv g(\xi, \epsilon) = \xi^\mu \epsilon^\nu g_{\mu\nu}. \tag{1.38}$$

The norm squared of a vector is just the product of the vector with itself, $\xi^2 = \xi \cdot \xi$. The metric will be required to be non-singular, i.e.

$$g \equiv \det(g_{\mu\nu}) \neq 0, \tag{1.39}$$

and locally diagonalizable into $\eta_{\mu\nu} = \mathrm{diag}(+ - \cdots -)$ for physical and conventional reasons. Thus, in d dimensions

$$\mathrm{sign}\; g = \frac{g}{|g|} = (-1)^{d-1}. \tag{1.40}$$

As usual, a metric can be used to establish a correspondence between a vector space and its dual, i.e. between vectors and 1-forms: with each vector ξ^μ we associate a 1-form ω_μ whose action on any other vector η^μ is the product of ξ and η, $\omega(\eta) = \xi^\mu \eta^\nu g_{\mu\nu}$, which means the relation between components $\omega_\nu = \xi^\mu g_{\mu\nu}$. It is customary to denote this 1-form by ξ_μ, and the transformation from vector to 1-form is represented by lowering the index.

The inverse metric can be used as a metric in cotangent space, and its components are those of the inverse matrix and are denoted with upper indices. The operation of raising indices can be similarly defined, and the consistency of all these operations is guaranteed because the dual of the dual is the original vector space. The extension to tensors of higher ranks is straightforward.

The determinant of the metric can also be used to relate tensors and weight w tensor densities, since it transforms as a density of weight $w = 2$ and the product of a tensor and $g^{w/2}$ transforms as a density of weight w.

Furthermore, with a metric we can define the *Ricci scalar* R and the *Einstein tensor* $G_{\mu\nu}$,

$$R = R_\mu{}^\mu, \qquad G_{\mu\nu} = R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R, \tag{1.41}$$

which needs not be symmetric (just like the Ricci tensor).

So far we have two independent fields defined on our manifold: the metric and the affine connection. An L_d spacetime equipped with a metric is sometimes denoted by (L_d, g). The affine connection and the metric are related by the *non-metricity tensor* $Q_{\mu\nu\rho}$,

$$Q_{\mu\nu\rho} \equiv -\nabla_\mu g_{\nu\rho}. \tag{1.42}$$

If we take the combination $\nabla_\mu g_{\rho\sigma} + \nabla_\rho g_{\sigma\mu} - \nabla_\sigma g_{\mu\rho}$ and expand it, we find that the connection can be written as follows:

$$\Gamma_{\mu\nu}{}^\rho = \left\{ {\rho \atop \mu\,\nu} \right\} + K_{\mu\nu}{}^\rho + L_{\mu\nu}{}^\rho, \tag{1.43}$$

where

$$\left\{ {\rho \atop \mu\,\nu} \right\} = \tfrac{1}{2} g^{\rho\sigma} \{ \partial_\mu g_{\nu\sigma} + \partial_\nu g_{\mu\sigma} - \partial_\sigma g_{\mu\nu} \} \tag{1.44}$$

are the *Christoffel symbols*, which are completely determined by the metric, and K is called the *contorsion tensor* and is given in terms of the torsion tensor by

$$\begin{aligned} K_{\mu\nu}{}^{\rho} &= \tfrac{1}{2} g^{\rho\sigma} \{ T_{\mu\sigma\nu} + T_{\nu\sigma\mu} - T_{\mu\nu\sigma} \} \,, \\ K_{[\mu\nu]}{}^{\rho} &= -\tfrac{1}{2} T_{\mu\nu}{}^{\rho}, \qquad K_{\mu\nu\rho} = -K_{\mu\rho\nu}. \end{aligned} \tag{1.45}$$

Finally

$$L_{\mu\nu}{}^{\rho} = \tfrac{1}{2} \{ Q_{\mu\nu}{}^{\rho} + Q_{\nu\mu}{}^{\rho} - Q^{\rho}{}_{\mu\nu} \} \,. \tag{1.46}$$

Observe that the contorsion tensor depends on the metric whereas the torsion tensor does not. Furthermore, observe that, since the contorsion and non-metricity tensors transform as tensors, the piece responsible for the non-homogeneous term in the transformation of the affine connection is the Christoffel symbol.

With a metric it is also possible to define the length of a curve γ, $x^\mu(\xi)$, by the integral

$$s = \int_\gamma d\xi \, \sqrt{g_{\mu\nu}(x) \dot{x}^\mu \dot{x}^\nu}. \tag{1.47}$$

If we consider the above expression as a functional in the space of all curves joining two given points, we can ask which of those curves minimizes it. The answer is given by the Euler–Lagrange equations, which take the simple form

$$\ddot{x}^\mu + \dot{x}^\rho \dot{x}^\sigma \left\{ {\mu \atop \rho\,\sigma} \right\} = 0, \tag{1.48}$$

if we parametrize the curve by its proper length s. This is the *geodesic equation*, and is different from the autoparallel equation (1.23) whenever there is torsion and non-metricity.

In the standard theory of gravity metric and affine connection are not independent variables since we want to describe only the degrees of freedom corresponding to a massless spin-2 particle. To relate these two fields one imposes the *metric postulate*

$$Q_{\mu\rho\sigma} = -\nabla_\mu g_{\rho\sigma} = 0, \tag{1.49}$$

which makes the operations of raising and lowering of indices commute with the covariant derivative. A connection satisfying the above condition is said to be *metric compatible* and a spacetime (L_d, g) with a metric-compatible connection is called a *Riemann–Cartan spacetime* and is denoted by U_d.

Sometimes a weaker condition is required: the vanishing of the trace-free part of the non-metricity tensor $\hat{Q}$

$$\hat{Q}_{\mu\nu\rho} \equiv Q_{\mu\nu\rho} - \tfrac{1}{d} Q_{\mu\sigma}{}^{\sigma} g_{\nu\rho} = 0. \tag{1.50}$$

In this case, the non-metricity must take the form

$$Q_{\mu\nu\rho} = -A_\mu g_{\nu\rho}, \quad \Rightarrow \quad L^{\rho}_{\mu\nu} = A_{(\mu} g_{\nu)}{}^{\rho} - \tfrac{1}{2} g_{\mu\nu} A^{\rho}, \tag{1.51}$$

where $A_\mu = -\frac{1}{d}Q_{\mu\sigma}{}^{\sigma}$ is known as the *Weyl connection* and the corresponding space is a *Cartan–Weyl* space. The presence of a Weyl connection makes the spacetime invariant under conformal rescalings of the metric

$$g'_{\mu\nu} \equiv \Omega^2 g_{\mu\nu}, \qquad \nabla'_\mu g'_{\nu\rho} = \Omega^2 \nabla_\mu g_{\nu\rho}, \tag{1.52}$$

assuming that the Weyl connection transforms as

$$A'_\mu = A_\mu - \partial_\mu \ln \Omega^2, \tag{1.53}$$

that is, as a gauge connection with gauge group $\mathbb{R}$ (see Appendix A).[6]

Still, even if the Weyl connection vanishes, the metric postulate leaves the torsion undetermined. If we want to have a connection completely determined by the metric, left as the only independent field, one has to impose the vanishing of the torsion tensor. The torsionless, metric-compatible connection is called *Levi-Civita connection* and its components are given by the Christoffel symbols.[7] A Riemann–Cartan spacetime U_d with vanishing torsion is a *Riemann spacetime* V_d. If the torsion vanishes but there is a Weyl connection, the spacetime is an *Einstein–Weyl spacetime*.

There is another way of reducing the number of independent fields: by imposing the vanishing of the curvature tensor. In this case (for vanishing Weyl connection), both the metric and the connection are completely determined by the *Vielbein* (to be defined latter). The connection is called *Weitzenböck connection* [1243, 1244] and has torsion (also determined by the Vielbein). A Riemann–Cartan spacetime with Weitzenböck connection is a *Weitzenböck spacetime* A_d; if one includes a Weyl connection, then it is called a *Weitzenböck–Weyl space*.

If both torsion and curvature vanish, the space has to be *Minkowski spacetime* M_d since the *Minkowski metric* $g_{\mu\nu} = \eta_{\mu\nu}$ is the only one that makes the full Riemann tensor vanish in the absence of torsion. Again, adding a Weyl connection, we have a *Minkowski–Weyl space*.

The diagram in Fig. 1.1, a more complete version of those in Refs. [736, 737] taken essentially from Ref. [927], summarizes the different particular structures that we can have on an affinely connected manifold equipped with a metric. We have introduced some changes in the diagram of Ref. [927] to indicate that Weyl spaces have a Weyl connection and not just arbitrary non-metricity.

In the rest of this section we are going to study the particular properties of some of these spacetimes. The Weitzenböck spacetime will be studied after the introduction of Vielbeins in Section 1.4.

1.3.1 Riemann–Cartan spacetime U_d

As has been said, this is an affinely connected metric spacetime with a metric-compatible connection, so the non-metricity tensor vanishes, $Q_{\mu\nu\rho} = 0$. According to the general result,

---

[6] This is, as a matter of fact, how gauge symmetries were introduced in physics (see Ref. [1004] and references therein.)

[7] Sometimes (especially in the supergravity context) the Levi-Civita connection is written $\Gamma(g)$ to stress the fact that it is a function of the metric in order to distinguish it from arbitrary connections that are independent of the metric. We will do so only when necessary.

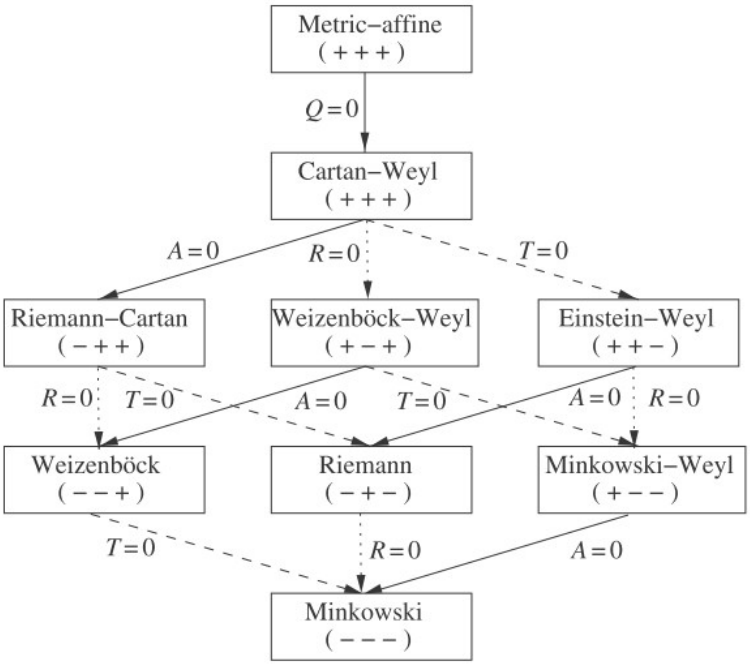
Metric–affine
(+ + +)
$Q = 0$
Cartan–Weyl
(+ + +)
$A = 0$
$R = 0$
$T = 0$
Riemann–Cartan
(− + +)
Weizenböck–Weyl
(+ − +)
Einstein–Weyl
(+ + −)
$R = 0$
$T = 0$
$A = 0$
$T = 0$
$A = 0$
$R = 0$
Weizenböck
(− − +)
Riemann
(− + −)
Minkowski–Weyl
(+ − −)
$T = 0$
$R = 0$
$A = 0$
Minkowski
(− − −)

Fig. 1.1. Particular structures in an affinely connected spacetime equipped with a metric (L_d, g) classified by their vanishing $(-)$ or non-vanishing $(+)$ Weyl connection A, curvature R, and torsion T expressed as the triplet (A, R, T). Only Weyl connections with vanishing trace-free non-metricity $\hat{Q}$ are usually considered and they are called Cartan–Weyl spacetimes. They can have arbitrary Weyl connection, Riemann curvature, and torsion, and they are denoted by $(+\ +\ +)$. A Cartan–Weyl spacetime with vanishing Weyl connection is a Riemann–Cartan spacetime $(-\ +\ +)$. If the curvature vanishes instead, we have a Weitzenböck–Weyl spacetime $(+\ -\ +)$, and, if it is the torsion that vanishes, we have an Einstein–Weyl spacetime $(+\ +\ -)$. The rest of the diagram follows the same pattern.

a metric-compatible connection of a Riemann–Cartan spacetime always has the form

$$\Gamma_{\mu\nu}{}^{\rho} = \left\{ {\rho \atop \mu\,\nu} \right\} + K_{\mu\nu}{}^{\rho}. \tag{1.54}$$

Observe that the symmetric part of the contorsion tension does not vanish, but

$$K_{(\mu\nu)\,\rho} = \tfrac{1}{2}\left(T_{\mu\rho\nu} + T_{\nu\rho\mu}\right) \neq 0. \tag{1.55}$$

This means that the presence of torsion implies not only that the connection has a non-vanishing antisymmetric part, but also that the symmetric part is not fully determined by the metric but

$$\Gamma_{(\mu\nu)}{}^{\rho} = \left\{ {\rho \atop \mu\,\nu} \right\} + K_{(\mu\nu)}{}^{\rho} \neq \left\{ {\rho \atop \mu\,\nu} \right\}. \tag{1.56}$$

The curvature, Ricci, and Einstein tensors of a metric-compatible connection satisfy further identities. On contracting the γ and σ indices in the third Bianchi identity Eqs. (1.30) and using the metric postulate, we find the so-called *contracted Bianchi identity*

$$\nabla_\alpha G_\mu{}^{\alpha} + T_{\mu\alpha\beta}R^{\beta\alpha} - \tfrac{1}{2}T_{\alpha\beta\gamma}R_\mu{}^{\gamma\alpha\beta} = 0. \tag{1.57}$$

Furthermore, by applying the Ricci identity to the metric and using the metric postulate, one can prove a fourth Bianchi identity:

$$R_{\alpha\beta(\gamma\delta)} = 0. \tag{1.58}$$

If we modify the connection according to Eq. (1.31) and Γ is metric-compatible, the Ricci scalar is

$$R(\tilde{\Gamma}) = R(\Gamma) - T_{\mu\nu}{}^{\rho}\tau_\rho{}^{\mu\nu} + 2\nabla_\mu\tau_\nu{}^{\mu\nu} + \tau_\mu{}^{\mu\lambda}\tau_\nu{}^{\nu}{}_{\lambda} + \tau_\nu{}^{\mu\rho}\tau_{\mu\rho}{}^{\nu}. \tag{1.59}$$

If $\tilde{\Gamma}$ is not a metric-compatible connection, then τ contains all the contributions of the non-metricity tensor $\tau_{\mu\nu}{}^{\rho} = L_{\mu\nu}{}^{\rho}$ and the above formula allows us to work in the framework of a Riemann–Cartan spacetime with non-metric-compatible connections. We will use this formula when we deal with Einstein–Weyl spacetimes.

If both $\tilde{\Gamma}$ and Γ are metric-compatible connections and $\tilde{\Gamma}$ has torsion but $\Gamma = \Gamma(g)$, then $\tau = \tilde{K}$, the contorsion tensor of $\tilde{\Gamma}$, and the above formula takes a simpler form:

$$R(\tilde{\Gamma}) = R[\Gamma(g)] + 2\nabla_\mu\tilde{K}_\nu{}^{\mu\nu} + (\tilde{K}_\mu{}^{\mu\lambda})^2 + \tilde{K}_\nu{}^{\mu\rho}\tilde{K}_{\mu\rho}{}^{\nu}. \tag{1.60}$$

Now, this formula allows us to work with torsion in a Riemann spacetime. Particularly interesting is the case in which the contorsion $\tilde{K}_{\mu\nu\rho}$ is a completely antisymmetric tensor (proportional to the Kalb–Ramond field strength $H_{\mu\nu\rho}$, for instance). Then we have, if

$$\tilde{K}_{\mu\nu\rho} = \frac{1}{\sqrt{12}}H_{\mu\nu\rho}, \tag{1.61}$$

$$\int d^dx\sqrt{|g|}\;R(\tilde{\Gamma}) = \int d^dx\sqrt{|g|}\left\{R[\Gamma(g)] + \frac{1}{2\cdot 3!}H_{\mu\nu\rho}H^{\mu\nu\rho}\right\}. \tag{1.62}$$

1.3.2 Einstein–Weyl spacetime EW_d

This is defined by the conditions $\hat{Q} = T = 0$ which determine the connection Γ to be the sum of the Levi-Civita connection and a Weyl connection so it is invariant under GCTs and Weyl transformations. If we compute the commutator of two covariant derivatives on the metric and use the Ricci identity, we find that the curvature tensor is not antisymmetric in the second pair of indices, and instead of Eq. (1.58) we have this Bianchi identity

$$R_{\mu\nu(\rho\sigma)} = -\tfrac{1}{2}F_{\mu\nu}(A)g_{\rho\sigma}, \qquad \text{where} \qquad F_{\mu\nu}(A) = 2\partial_{[\mu}A_{\nu]}. \tag{1.63}$$

Therefore, $R_{\mu\nu(\rho\sigma)} = 0$ if A is *pure gauge* $A_\mu = \partial_\mu \ln \Omega^2$.

The complete curvature and Ricci tensors and the Ricci scalar can be easily computed using Eqs. (1.31), (1.32), (1.36), and (1.59) with $T = 0$ and $\tau_{\mu\nu}{}^{\rho} = L_{\mu\nu}{}^{\rho}$:

$$\begin{aligned} R_{\mu\nu\rho}{}^{\sigma}(\Gamma) = {} & R_{\mu\nu\rho}{}^{\sigma}(g) + \tfrac{1}{2}F_{\mu\nu}(A)g_{\rho}{}^{\sigma} + \nabla_{[\mu|}A_{\rho}g_{|\nu]}{}^{\sigma} - \nabla_{[\mu}A^{\sigma}g_{\nu]\rho} \\ & + \tfrac{1}{2}\left\{A_{[\mu}A^{\sigma}g_{\nu]\rho} - A_{[\mu|}A_{\rho}g_{|\nu]}{}^{\sigma} - \tfrac{1}{2}A^2 g_{[\mu}{}^{\sigma}g_{\nu]\,\rho}\right\}, \end{aligned} \tag{1.64}$$

$$\begin{aligned} R_{\mu\nu}(\Gamma) = {} & R_{\mu\nu}(g) + \tfrac{1}{4}(d-2)F_{\mu\rho}(A) + \tfrac{1}{2}(d-2)\left[\nabla_{(\mu}A_{\rho)} + \frac{1}{(d-2)}\nabla_\lambda A^\lambda g_{\mu\rho}\right] \\ & - \tfrac{1}{4}(d-2)\left[A_\mu A_\rho - A^2 g_{\mu\rho}\right], \end{aligned} \tag{1.65}$$

$$R(\Gamma) = R(g) + (d-1)\nabla_\lambda A^\lambda + \tfrac{1}{4}(d-2)(d-1)A^2. \tag{1.66}$$

For $A_\mu = \partial_\mu \ln \Omega^2$ we recover the relations between the curvature and Ricci tensors of a metric before and after a Weyl rescaling in Appendix M. The Ricci scalar has an overall factor Ω^{-2} because its definition uses an inverse metric.

1.3.3 Riemann spacetime V_d

It is defined by the conditions $Q = T = 0$ which determine the connection to be the Levi-Civita connection $\Gamma(g)$ whose components in a coordinate basis are given by the Christoffel symbols. In a Riemann spacetime one can construct infinitesimal parallelograms, and autoparallel curves are also geodesics (as in flat spacetime). There are also additional interesting properties. To start with, we can write the transformation of tensors under infinitesimal GCTs (Lie derivatives) in terms of covariant derivatives alone (all torsion terms vanish). In particular, for the metric and r-forms we can write

$$\begin{aligned} \delta_\xi g_{\mu\nu} &= -2\nabla_{(\mu}\xi_{\nu)}, \\ \delta_\xi B_{\mu_1\cdots\mu_r} &= -\xi^\lambda \nabla_\lambda B_{\mu_1\cdots\mu_r} - r(\nabla_{[\mu_1|}\xi^\lambda)B_{\lambda|\mu_2\cdots\mu_r]}. \end{aligned} \tag{1.67}$$

Furthermore, we have the usual identity

$$\Gamma_{\rho\mu}{}^{\rho} = \partial_\mu \ln\left(\sqrt{|g|}\right), \tag{1.68}$$

which allows us to write the Laplacian of a scalar function f in this way:

$$\nabla^2 f = \frac{1}{\sqrt{|g|}} \partial_\mu \Big(\sqrt{|g|}\, \partial^\mu f \Big), \tag{1.69}$$

and the divergence of a completely antisymmetric tensor (k-form) in this way:[8]

$$\nabla_{\mu_1} F^{\mu_1 \mu_2 \cdots \mu_k} = \frac{1}{\sqrt{|g|}} \partial_{\mu_1} \Big(\sqrt{|g|}\, F^{\mu_1 \mu_2 \cdots \mu_k} \Big). \tag{1.70}$$

The Bianchi identities take the form

$$R_{(\alpha\beta)\gamma}{}^{\delta} = 0, \qquad R_{[\alpha\beta\gamma]}{}^{\delta} = 0, \qquad \nabla_{[\alpha} R_{\beta\gamma]\rho}{}^{\sigma} = 0, \qquad R_{\alpha\beta(\gamma\delta)} = 0. \tag{1.71}$$

The first and fourth identities imply together

$$R_{\alpha\beta\gamma\delta} = R_{\gamma\delta\alpha\beta}, \tag{1.72}$$

which in turn implies that the Ricci and Einstein tensors are symmetric. The contracted Bianchi identity says now that the Einstein tensor is divergence free:

$$\nabla_\mu G^{\mu\nu} = 0, \tag{1.73}$$

which is a crucial identity in the development of general relativity.

The number of independent components of the curvature in d dimensions after taking into account all these Bianchi identities is $(1/12)\, d^2 (d^2 - 1)$.

The four-dimensional curvature tensor can be split into different pieces which transform irreducibly under the Lorentz group: a scalar piece $D(0,0)$, which is nothing but the Ricci scalar R, a two-index symmetric, a traceless piece $R_{\mu\nu} - \frac{1}{4} g_{\mu\nu} R$ (corresponding to the representation $D(1,1)$), and a four-index tensor with the same symmetries as the Riemann tensor but traceless: the Weyl tensor $C_{\mu\nu\rho}{}^{\sigma}$ with $C_{\mu\sigma\rho}{}^{\sigma} = 0$,

$$R_{\mu\nu}{}^{\rho\sigma} = C_{\mu\nu}{}^{\rho\sigma} + 2\Big(R_{[\mu}{}^{[\rho} - \tfrac{1}{4} R g_{[\mu}{}^{[\rho} \Big) g_{\nu]}{}^{\sigma]} + \tfrac{1}{6} R g_{[\mu}{}^{\rho} g_{\nu]}{}^{\sigma}. \tag{1.74}$$

The Weyl tensor can be decomposed into its self-dual and anti-self-dual parts (with respect to the last two indices). These two complex tensors transform in the $D(2,0)$ and $D(0,2)$ representations, respectively.

In d dimensions the Weyl tensor is defined by

$$C_{\mu\nu}{}^{\rho\sigma} = R_{\mu\nu}{}^{\rho\sigma} - \frac{4}{d-2} R_{[\mu}{}^{[\rho} g_{\nu]}{}^{\sigma]} + \frac{2}{(d-1)(d-2)} R\, g_{[\mu}{}^{[\rho} g_{\nu]}{}^{\sigma]}. \tag{1.75}$$

The main property of the Weyl tensor $C_{\mu\nu\rho}{}^{\sigma}$ with the indices in these positions is that it is left invariant by Weyl rescalings of the metric (see Appendix M). Furthermore, just as the Riemann curvature vanishes only for Minkowski spacetime, the Weyl tensor vanishes only for conformally flat (Minkowski) spacetimes, i.e. spacetimes that are related to Minkowski's by a given conformal transformation.

---

[8] Observe that this implies that the second term on the r.h.s. of Eq. (1.60) times $\sqrt{|g|}$ is a total derivative.

A final property of the Levi-Civita connection that is worth mentioning is the form of its variation under an arbitrary variation of the metric:

$$\delta\Gamma_{\mu\nu}{}^{\rho}(g) = \tfrac{1}{2}g^{\rho\sigma}\{\nabla_\mu\delta g_{\nu\sigma} + \nabla_\nu\delta g_{\mu\sigma} - \nabla_\sigma\delta g_{\mu\nu}\}\,. \tag{1.76}$$

Since $\delta g_{\mu\nu}$ is a tensor, $\delta\Gamma$ is a tensor even though Γ is not.

1.4 Tangent space

So far we have considered, for a given coordinate system, only one basis in tangent space: the coordinate basis. We are now going to consider an arbitrary basis in tangent space. Such a basis is defined by a set of d contravariant vectors labeled by a *tangent-space index* a $\{e_a = e_a{}^\mu\partial_\mu\}$ and is also referred to as a *frame* or, generically, *Vielbein* basis.[9] The coordinate basis is now a particular case in which $e_a{}^\mu = \delta_a{}^\mu$. Now we can express any vector in this basis $\xi = \xi^a e_a$ and its components ξ^a will be related to the coordinate basis components by

$$\xi^\mu = \xi^a e_a{}^\mu. \tag{1.77}$$

We can immediately define the dual basis of 1-forms $\{e^a = e^a{}_\mu dx^\mu\}$ defined by

$$\langle e^a|e_b\rangle = \delta^a{}_b, \tag{1.78}$$

which implies that the matrix of components $e^a{}_\mu$ of the 1-forms in the coordinate basis is the inverse, transposed, of that of the vectors:

$$e^a{}_\mu e_b{}^\mu = \delta^a{}_b, \quad \Rightarrow e_a{}^\mu e^a{}_\nu = \delta^\mu{}_\nu. \tag{1.79}$$

We can now relate frame and world indices of any tensor using these two matrices. In particular, we can use the frame components g_{ab} of the metric,

$$g_{ab} = e_a{}^\mu e_b{}^\nu g_{\mu\nu}, \tag{1.80}$$

that can also be interpreted as the matrix of inner products of the Vielbein basis $g(e_a, e_b) = g_{ab}$. An orthonormal Vielbein basis leads to $g_{ab} = \eta_{ab}$. Frames are usually chosen in such a way as to obtain a particular g_{ab} and orthonormal frames will be particularly important in what follows.

It is easy to see that g_{ab} and its inverse g^{ab} can be consistently used to raise and lower frame indices. In particular,

$$e^a{}_\mu = g_{\mu\nu}e_b{}^\nu g^{ab}, \qquad g_{\mu\nu} = e^a{}_\mu e^b{}_\nu g_{ab}. \tag{1.81}$$

A frame is invariant under GCTs (only the components in the coordinate basis change) and, thus, frame components of any tensor are also invariant. However, we can make a change of basis. Any two Vielbein bases are related by a $\mathrm{GL}(d,\mathbb{R})$ transformation $\Lambda^a{}_b$ in tangent space at a given point of the manifold. This transformation can in fact be different at each point and thus we have to consider local frame transformations $\Lambda^a{}_b(x)$. We write their action on vectors and forms as follows:

---

[9] *Einbein* for $d = 1$, *Zweibein* for $d = 2$, *Dreibein* for $d = 3$, *Vierbein* for $d = 4$, etc. In four dimensions it is also called a *tetrad*.

$$e'_a = e_b\left(\Lambda^{-1}\right)^b{}_a, \qquad e'^{\,a} = \Lambda^a{}_b e^b. \tag{1.82}$$

The *Ricci rotation coefficients* (or *anholonomy coefficients*) $\Omega_{ab}{}^c$ are the Lie brackets

$$[e_a, e_b] = -2\Omega_{ab}{}^c e_c, \qquad \Omega_{ab}{}^c = e_a{}^\mu e_b{}^\nu \partial_{[\mu} e^c{}_{\nu]}. \tag{1.83}$$

A *non-holonomic* frame is one with non-vanishing Ωs. Observe that, given a basis of vectors $\{e_a\}$, we could try to find a new set of coordinates $y^a(x^\mu)$ such that

$$e_a y^b = e_a{}^\mu \partial_\mu y^b = \delta_a{}^b. \tag{1.84}$$

The integrability condition for the system of partial differential equations $[e_c, e_a]y^b = 0$ is precisely the vanishing of the anholonomy coefficients $\Omega_{ab}{}^c$. A holonomic basis of vectors $\{e_a\}$ is one for which these coefficients vanish and then we can trivialize them ($e_a{}^\mu = \delta_a{}^\mu$) by a change of coordinates.

Just as we defined a covariant derivative transforming world tensors into world tensors we are now going to define a derivative that transforms tangent-space tensors into tangent-space tensors transforming well under local $\mathrm{GL}(d, \mathbb{R})$ transformations associated with a connection ω. Its action on scalars, vectors, and forms is[10]

$$\begin{aligned}
\mathcal{D}_a\phi &= \partial_a\phi, & &(= e_a{}^\mu \partial_\mu\phi),\\
\mathcal{D}_a\xi^b &= \partial_a\xi^b + \omega_{ac}{}^b\xi^c, & &(= e_a{}^\mu \mathcal{D}_\mu\xi^b),\\
\mathcal{D}_a\varepsilon_b &= \partial_a\varepsilon_b - \varepsilon_c\,\omega_{ab}{}^c, & &(= e_a{}^\mu \mathcal{D}_\mu\varepsilon_b).
\end{aligned} \tag{1.85}$$

Local $\mathrm{GL}(d, \mathbb{R})$ covariance implies the inhomogeneous transformation law for the connection:

$$\omega'_{ab}{}^c = \left[\Lambda^c{}_d\omega_{ef}{}^d\left(\Lambda^{-1}\right)^f{}_b - \left(\Lambda^{-1}\right)^d{}_b\partial_e\Lambda^c{}_d\right](\Lambda^{-1})^e{}_a. \tag{1.86}$$

The curvature of this connection can be defined through the Ricci identities in the standard fashion (observe that there are no torsion terms here):

$$\begin{aligned}
[\mathcal{D}_\mu, \mathcal{D}_\nu]\phi &= 0,\\
[\mathcal{D}_\mu, \mathcal{D}_\nu]\xi^a &= R_{\mu\nu b}{}^a\xi^b,\\
[\mathcal{D}_\mu, \mathcal{D}_\nu]\varepsilon_a &= -\varepsilon_b\,R_{\mu\nu a}{}^b,
\end{aligned} \tag{1.87}$$

and then the curvature is given by[11]

$$R_{\mu\nu a}{}^b = 2\partial_{[\mu}\,\omega_{\nu]a}{}^b - 2\omega_{[\mu|a}{}^c\,\omega_{|\nu]c}{}^b. \tag{1.89}$$

---

[10] Of course, the formalism we are developing is just that of a $\mathrm{GL}(d, \mathbb{R})$ gauge theory and our notation is basically identical to that of Appendix A. Here we are dealing with vector representations of $\mathrm{GL}(d, \mathbb{R})$. The d^2 generators of its Lie algebra can be labeled by a pair of vector indices ab and they are given, for instance, by $(T_{ab})^c{}_d = -\eta_{ad}\eta_b{}^c$. Thus $A_\mu{}^I\Gamma_{\mathrm{v}}(T_I) = \omega_\mu{}^{ab}(T_{ab})^c{}_d = -\omega_{\mu\,d}{}^c$. The subgroup $\mathrm{SO}(1, d-1, \mathbb{R})$ will be treated in more detail.

[11] Observe that, with all Latin indices, $R_{abc}{}^d = e_a{}^\mu e_b{}^\nu R_{\mu\nu c}{}^d$ and, therefore,

$$R_{abc}{}^d = 2\partial_{[a}\,\omega_{b]c}{}^d - 2\omega_{[a|c}{}^e\,\omega_{|b]e}{}^d + 2\Omega_{ab}{}^e\omega_{ec}{}^d. \tag{1.88}$$

At this point we have introduced a new connection ω that is independent of the metric. In the previous section we managed to relate the connection Γ to the metric via the metric postulate. Here we are going to generalize the metric postulate first to relate the two connections (the first Vielbein postulate) and then to relate them to the metric (the second Vielbein postulate). Before we enunciate these postulates we introduce the *total covariant derivative*, covariant with respect to all the indices of the object it acts on. We denote it by ∇ again, and, for instance, acting on Vielbeins it is

$$\nabla_\mu e_a{}^\nu = \partial_\mu e_a{}^\nu + \Gamma_{\mu\rho}{}^\nu e_a{}^\rho - e_b{}^\nu \omega_{\mu a}{}^b. \tag{1.90}$$

We can motivate the first Vielbein postulate as follows: we would like to be able to convert tangent into world indices and vice versa inside the total covariant derivative, so $e^a{}_\nu \nabla_\mu \xi^\nu = \mathcal{D}_\mu \xi^a$ and $\mathcal{D}$ is just the projection of ∇ onto the Vielbein basis. To have this property we impose the *first Vielbein postulate*,

$$\nabla_\mu e_a{}^\nu = 0. \tag{1.91}$$

It is worth stressing that this does not imply the covariant constancy of the metric $\nabla_\mu g_{\nu\rho} = 0$. The above postulate implies the following relation between the connections:

$$\omega_{\mu a}{}^b = \Gamma_{\mu a}{}^b - e_a{}^\nu \partial_\mu e^b{}_\nu. \tag{1.92}$$

Furthermore, the curvatures of the two connections are now related by

$$R_{\mu\nu\rho}{}^\sigma(\Gamma) = e^a{}_\rho e_b{}^\sigma R_{\mu\nu a}{}^b(\omega). \tag{1.93}$$

The first Vielbein postulate also gives an important relation between the torsion and the Vielbein: on taking the antisymmetric part of $\nabla_\mu e^a{}_\nu = 0$, we obtain

$$2\mathcal{D}_{[\mu} e^a{}_{\nu]} = 2\left(\partial_{[\mu} e^a{}_{\nu]} - \omega_{[\mu}{}^a{}_{\nu]}\right) = -T_{\mu\nu}{}^a. \tag{1.94}$$

The significance of the torsion in this formalism, from the point of view of the gauge theory of GL$(d, \mathbb{R})$, is unclear. We can provide an interpretation in the framework of the gauge theory of the affine group IGL$(d, \mathbb{R})$ but we will do it in the more restricted context of the Lorentz and Poincaré groups in Section 4.5.

The first Vielbein postulate has allowed us to recover the structure of affinely connected spacetime (L_d, g) with only one (independent) connection, generalized to allow the use of an arbitrary basis in tangent space. Furthermore, we can recover the different particular structures that we defined in the previous section, also generalized to allow the use of arbitrary basis in tangent space. First, if Γ is a completely general connection, it is given by Eq. (1.43) and then ω (which is related to Γ by the first Vielbein postulate) is given by

$$\omega_{ab}{}^c = \omega_{ab}{}^c(e) + K_{ab}{}^c + L_{ab}{}^c, \tag{1.95}$$

where $\omega(e)$ is the (*Cartan* or even *Levi-Civita*) connection related to the Levi-Civita connection $\Gamma(g)$ by Eq. (1.92). It is completely determined by the Vielbeins through the anholonomy coefficients defined in Eq. (1.83):

$$\omega_{ab}{}^c(e) = \left\{ {c \atop a\,b} \right\} + \{-\Omega_{ab}{}^c + \Omega_b{}^c{}_a - \Omega^c{}_{ab}\}, \tag{1.96}$$

where

$$\left\{ {c \atop a\,b} \right\} = \tfrac{1}{2} g^{cd} \{ \partial_a g_{bd} + \partial_b g_{ad} - \partial_d g_{ab} \}. \tag{1.97}$$

$K_{ab}{}^{c}$ is nothing but the contorsion tensor expressed in a tangent-space basis, i.e. $K_{ab}{}^{c} = e_a{}^{\mu} e_b{}^{\nu} e^c{}_{\rho} K_{\mu\nu}{}^{\rho}$ and, similarly, $L_{ab}{}^{c} = e_a{}^{\mu} e_b{}^{\nu} e^c{}_{\rho} L_{\mu\nu}{}^{\rho}$.

Observe that

$$\omega_{a(bc)} = \tfrac{1}{2} (Q_{abc} + \partial_a g_{bc}). \tag{1.98}$$

We can impose the metric-compatibility condition Eq. (1.49), which in this context is known as the *second Vielbein postulate*, and we have a Riemann–Cartan spacetime U_d. The result is that Γ is again given by Eqs. (1.54), (1.44), and (1.45), and ω (which is related to Γ by the first Vielbein postulate) is given by

$$\omega_{ab}{}^{c} = \omega_{ab}{}^{c}(e) + K_{ab}{}^{c}. \tag{1.99}$$

If we now impose the vanishing of torsion, we obtain the Levi-Civita and Cartan connections $\Gamma(g)$ and $\omega(e)$ and we recover a Riemann spacetime V_d.

The two most important cases to which we can apply this general formalism are the following.

1. The case in which we use a coordinate basis $e_a{}^{\mu} = \delta_a{}^{\mu}$, so $g_{ab} = g_{\mu\nu}$, $\Omega = 0$, and the connections Γ and ω are identical.
2. The case in which we use an orthonormal basis $g_{ab} = \eta_{ab}$ in which

$$\left\{ {c \atop a\,b} \right\} = 0$$

and

$$\omega_{abc} = \omega_{abc}(e) + K_{abc} + L_{abc}, \qquad \omega_{abc}(e) = -\Omega_{abc} + \Omega_{bca} - \Omega_{cab}. \tag{1.100}$$

In the second case we would like to restrict ourselves to those changes of frame that preserve the form of the metric in tangent-space indices (here usually referred to as *flat indices* because they are raised and lowered with the flat space metric). By definition, these are transformations of the d-dimensional Lorentz group $\mathrm{SO}(1, d-1)$ whose gauge theory we are now led to consider. This gauge theory is developed in Section A.2.3 of Appendix A and the spinorial representations of the Lorentz group are studied in Appendix D. We are simply going to rewrite here the main formulae we have obtained, adapted to the Lorentz subgroup of $\mathrm{GL}(d, \mathbb{R})$.

The main justification for making this step is that the Lorentz group admits spinorial representations, which are necessary in order to describe fermions, whereas the diffeomorphism group of a manifold does not. This is the only known method by which to describe spinors in curved spacetime in arbitrary coordinates and, thus, the only method known to couple fermions to gravity. This formalism was pioneered by Weyl [1252].

First of all, the generators M_I of the Lorentz subgroup of $\mathrm{GL}(d,\mathbb{R})$ are just the antisymmetric combinations of those of $\mathrm{GL}(d,\mathbb{R})$ and can be labeled by two antisymmetric vector indices, i.e. M_{ab}. In this notation every generator appears twice and factors of $\frac{1}{2}$ have to be included in the right places. However, in general, the connection $\omega_\mu{}^{ab}$ is not antisymmetric in the "gauge" indices ab unless it is also metric compatible ($\mathcal{D}\eta_{ab}=0$), according to Eq. (1.98). We are going to consider only metric-compatible connections that are fully antisymmetric in the gauge indices and we will call them *spin connections*.[12]

Using the explicit form of the infinitesimal Lorentz generators in the vector representation $\Gamma_{\mathrm{v}}(M_{bc})^a{}_d$ given in Eq. (A.60) and in the spinorial representation $\Gamma_{\mathrm{s}}(M_{ab})^\alpha{}_\beta$ (we use temporarily the first few Greek letters $\alpha,\beta,\ldots$ as spinorial indices) given in Eq. (D.3), we find the following expressions for the (total) covariant derivatives of contravariant and covariant vectors and spinors:

$$\begin{aligned}
\nabla_\mu\xi^a &= \partial_\mu\xi^a - \tfrac{1}{2}\omega_\mu{}^{bc}\Gamma_{\mathrm{v}}(M_{bc})^a_d\xi^d = \partial_\mu\xi^a + \omega_{\mu b}{}^a\xi^b,\\
\nabla_\mu\varepsilon_a &= \partial_\mu\varepsilon_a + \varepsilon_d\tfrac{1}{2}\omega_\mu{}^{bc}\Gamma_{\mathrm{v}}(M_{bc})^d_a = \partial_\mu\varepsilon_a - \varepsilon_b\,\omega_{\mu a}{}^b,\\
\nabla_\mu\psi^\alpha &= \partial_\mu\psi^\alpha - \tfrac{1}{2}\omega_\mu{}^{ab}\Gamma_{\mathrm{s}}(M_{ab})^\alpha{}_\beta\psi^\beta = \partial_\mu\psi^\alpha - \tfrac{1}{4}\omega_\mu{}^{ab}(\Gamma_{ab})^\alpha{}_\beta\psi^\beta,\\
\nabla_\mu\varphi_\alpha &= \partial_\mu\varphi_\alpha + \varphi_\beta\tfrac{1}{2}\omega_\mu{}^{ab}\Gamma_{\mathrm{s}}(M_{ab})^\beta{}_\alpha = \partial_\mu\varphi_\alpha + \varphi_\beta\tfrac{1}{4}\omega_\mu{}^{ab}(\Gamma_{ab})^\beta{}_\alpha.
\end{aligned}\tag{1.101}$$

These definitions, once we impose the Vielbein postulates, are consistent with the raising and lowering of vector indices with the Minkowski metric and with Dirac conjugation of the spinors. With the postulates, the spin connection is given by Eqs. (1.100).

The Ricci identities can now be written for the total covariant derivative in this form:

$$\begin{aligned}
[\nabla_\mu,\nabla_\nu]\,\phi &= T_{\mu\nu}{}^\rho\nabla_\rho\phi,\\
[\nabla_\mu,\nabla_\nu]\,\xi^a &= R_{\mu\nu b}{}^a(\omega)\xi^b + T_{\mu\nu}{}^\rho\nabla_\rho\xi^a,\\
[\nabla_\mu,\nabla_\nu]\,\varepsilon_a &= -\varepsilon_b R_{\mu\nu a}{}^b(\omega) + T_{\mu\nu}{}^\rho\nabla_\rho\varepsilon_a,\\
[\nabla_\mu,\nabla_\nu]\;\psi &= -\tfrac{1}{4}R_{\mu\nu}{}^{ab}(\omega)\Gamma_{ab}\psi + T_{\mu\nu}{}^\rho\nabla_\rho\psi,\\
[\nabla_\mu,\nabla_\nu]\;\varphi &= +\tfrac{1}{4}\varphi R_{\mu\nu}{}^{ab}(\omega)\Gamma_{ab} + T_{\mu\nu}{}^\rho\nabla_\rho\varphi.
\end{aligned}\tag{1.102}$$

For more general tensors one has to add a curvature (ω) term for each flat index and a curvature (Γ) term for each world index. The curvatures have the same form as in Eqs. (1.27) and (1.89) but now $R_{\mu\nu}{}^{ab}$ is antisymmetric in ab.

The following expression is sometimes used:

$$R_{ab} = -\partial_a\omega_c{}^c{}_b - \partial_c\omega_{ab}{}^c + \omega_{cda}\omega^{dc}{}_b + \omega_{abd}\omega_c{}^{cd}.\tag{1.103}$$

The Vielbein formalism allows us to study the Weitzenböck spacetime defined on p. 11.

1.4.1 Weitzenböck spacetime A_d

This spacetime is defined by a metric-compatible connection that we denote by $W_{\mu\nu}{}^\rho$ and call the *Weitzenböck connection* [1243, 1244] whose Riemann curvature is identically zero,

---

[12] If we wanted to have a non-metric-compatible spin connection, we would have to modify the first Vielbein postulate.

$R_{\mu\nu\rho}{}^{\sigma}(W)=0$. Trying to solve this equation directly for $W\neq 0$ is a very difficult task. However, we can use the Vielbein formalism to find a solution. Let us denote by $W_{\mathrm{s}\,\mu}{}^{ab}$ the tangent-space connection associated with W via the first Vielbein postulate

$$\nabla_{\mu}e^{a}{}_{\nu}=\partial_{\mu}e^{a}{}_{\nu}-W_{\mu\nu}{}^{a}+W_{\mathrm{s}\,\mu\nu}{}^{a}=0. \tag{1.104}$$

The curvature of W_{s} is obviously zero on account of Eq. (1.93). Now, however, we can use the trivial solution to the equation $R_{\mu\nu}{}^{ab}(W_{\mathrm{s}})=0$, namely $W_{\mathrm{s}}=0$, because, according to the above relation, $W_{\mathrm{s}}=0$ does not imply $W=0$ but

$$W_{\mu\nu}{}^{\rho}=e_{a}{}^{\rho}\partial_{\mu}e^{a}{}_{\nu}. \tag{1.105}$$

This is the Weitzenböck connection whose curvature vanishes identically. It cannot be rewritten in terms of the metric: it is necessary to use the Vielbein formalism. Observe that, using this connection, we can write the relation between any two connections Γ and ω satisfying the first Vielbein postulate in the form

$$\Gamma_{\mu\nu}{}^{\rho}=W_{\mu\nu}{}^{\rho}+\omega_{\mu\nu}{}^{\rho}. \tag{1.106}$$

$\omega_{\mu\nu}{}^{\rho}$ is a tensor, but $\Gamma_{\mu\nu}{}^{\rho}$ is not (it is an affine connection), and responsible for this is the Weitzenböck connection $W_{\mu\nu}{}^{\rho}$. We can also write

$$\omega_{\mu}{}^{ab}=\Gamma_{\mu}{}^{ab}-W_{\mu}{}^{ab},\qquad W_{\mu}{}^{ab}=e^{a\nu}\partial_{\mu}e^{b}{}_{\nu}. \tag{1.107}$$

Now $\Gamma_{\mu}{}^{ab}$ is a $\mathrm{GL}(d,\mathbb{R})$ tensor in the upper two indices whereas $\omega_{\mu}{}^{ab}$ is not (because it is a $\mathrm{GL}(d,\mathbb{R})$ connection). Again, the Weitzenböck connection $W_{\mu}{}^{ab}$ is responsible for this.

Even though we have to search explicitly for a metric-compatible connection to find W, it is easy to check that it is indeed metric compatible. Then, it can be decomposed into the sum of the Levi-Civita connection and the contorsion tensor. The torsion tensor is

$$T_{\mu\nu}{}^{\rho}=-2\Omega_{\mu\nu}{}^{\rho}, \tag{1.108}$$

and, therefore, the contorsion tensor is given by

$$K_{\mu\nu\rho}(W)=\Omega_{\mu\nu\rho}-\Omega_{\nu\rho\mu}+\Omega_{\rho\mu\nu}=-\omega_{\mu\nu\rho}(e), \tag{1.109}$$

where $\omega(e)$ is, as usual, the Cartan connection (which is associated via the first Vielbein postulate with the Levi-Civita connection $\Gamma(g)$).

Now, if we use Eqs. (1.31) and (1.32) for $\tilde{\Gamma}=W,\Gamma=\Gamma(g)$, and $\tau=K$, we find an expression for the Riemann curvature tensor of the Levi-Civita connection in terms of the contorsion tensor of the Weitzenböck connection:

$$R_{\mu\nu\rho}{}^{\sigma}[\Gamma(g)]=-2\nabla_{[\mu}K_{\nu]\rho}{}^{\sigma}-2K_{[\mu|\lambda}{}^{\sigma}K_{|\nu]\rho}{}^{\lambda}. \tag{1.110}$$

On contracting indices, and eliminating a total derivative, we find that

$$\int d^{d}x\,\sqrt{|g|}\,R(g)=-\int d^{d}x\,\sqrt{|g|}\,\left\{K_{\mu}{}^{\mu\lambda}K_{\mu}{}^{\mu}{}_{\lambda}+K_{\nu}{}^{\mu\rho}K_{\mu\rho}{}^{\nu}\right\}, \tag{1.111}$$

which can be expressed entirely in terms of the anholonomy coefficients $\Omega_{\mu\nu\rho}$, providing an alternative form of the Einstein–Hilbert action. This is, in fact, an alternative way of deriving Palatini's identity Eq. (L.4).

It is worth stressing here that the building blocks of the Riemann curvature tensor of the Levi-Civita connection in the above expression (the anholonomy coefficients/torsion and contorsion) are tensors, whereas in the standard expression for the curvature the building blocks are the Christoffel symbols, which are not tensors.

The main property of the Weitzenböck spacetime (the vanishing of the curvature) implies that parallel transport is path independent and it is possible to define parallelism of vectors *at different spacetime points*:[13] two vectors $v^\mu(x_1), w^\mu(x_2)$ are parallel if their components in the Vielbein basis $\{e_a{}^\mu\}$ are proportional. This is a consistent definition because the components in the Vielbein basis are invariant under the parallel transport defined by the W connection associated with that Vielbein basis. Indeed, the vector $v^\mu(x)$, parallel-transported to $x^\mu + \epsilon^\mu$ is

$$v^\mu(x+\epsilon) = v^\mu(x) - \epsilon^\nu v^\rho(x) W_{\nu\rho}{}^\mu(x), \tag{1.112}$$

and its tangent-space components can be found with the inverse Vielbein basis at $x^\mu + \epsilon^\mu$:

$$\begin{aligned} e^a{}_\mu(x+\epsilon)v^\mu(x+\epsilon) &= [e^a{}_\mu(x) + \epsilon^\nu \partial_\nu e^a{}_\mu(x)][v^\mu(x) - \epsilon^\nu v^\rho(x) W_{\nu\rho}{}^\mu(x)] \\ &= e^a{}_\mu(x) v^\mu(x). \end{aligned} \tag{1.113}$$

Also, it can be shown that the vanishing of the curvature is equivalent to the existence of d vector fields (the Vielbeins) covariantly constant with respect to the W connection,

$$\overset{W}{\nabla}_\mu\, e^a{}_\nu = \partial_\mu e^a{}_\nu - W_{\mu\nu}{}^\rho e^a{}_\rho = 0. \tag{1.114}$$

1.5 Killing vectors

If, given a metric $g_{\mu\nu}$, there exists a vector field k^μ such that the Lie derivative of $g_{\mu\nu}$ with respect to it vanishes,

$$\mathcal{L}_k g_{\mu\nu} = 2\nabla_{(\mu} k_{\nu)} = 0, \tag{1.115}$$

we say that $g_{\mu\nu}$ admits the *Killing vector* k^μ. The above equation is the *Killing equation*. It means that the metric does not change along the integral curves of k^μ and it is also said that the metric possesses an *isometry* in the direction k^μ. If the metric does not change along the integral curves of a Killing vector, and we use as a coordinate the parameter of those integral curves (*adapted coordinates*), then the metric does not depend on that coordinate.

The Ricci identity implies the following consistency condition:[14]

$$\nabla_\alpha \nabla_\beta k_\nu = -R^\lambda{}_{\alpha\beta\nu} k_\lambda, \quad \Rightarrow \nabla^2 k^\mu = R^\mu{}_\nu k^\nu. \tag{1.116}$$

---

[13] Also known as *teleparallelism* or *absolute parallelism*.

[14] These equations can be obtained as follows: from the Killing equation we have

$$\nabla_\alpha \nabla_{(\beta} k_{\gamma)} = 0, \qquad \text{and} \qquad \nabla_\alpha \nabla_{(\beta} k_{\gamma)} + \nabla_\beta \nabla_{(\gamma} k_{\alpha)} - \nabla_\gamma \nabla_{(\alpha} k_{\beta)} = 0.$$

The last equation can be rewritten in the form

$$\nabla_{(\alpha} \nabla_{\beta)} k_\gamma = -\tfrac{1}{2}[\nabla_\alpha, \nabla_\gamma] k_\beta - \tfrac{1}{2}[\nabla_\beta, \nabla_\gamma] k_\alpha,$$

A weaker but also interesting property that a metric can have is a *conformal isometry*. This happens when there is a vector field c^μ along whose integral curves the metric changes only by a conformal factor,

$$\mathcal{L}_c g_{\mu\nu} = 2\nabla_{(\mu}c_{\nu)} = 2\lambda g_{\mu\nu}. \tag{1.117}$$

On taking the trace of the above equation, we find in d dimensions

$$\lambda = \frac{1}{d}\nabla_\mu c^\mu, \quad \Rightarrow \nabla_{(\mu}c_{\nu)} - \frac{1}{d}g_{\mu\nu}\nabla_\rho c^\rho = 0, \tag{1.118}$$

called the *conformal Killing equation*; c^μ is then known as a *conformal Killing vector*.

1.6 Duality operations

The antisymmetric Levi-Civita tensor is defined in d dimensions in tangent space by

$$\epsilon^{01\cdots(d-1)} = +1, \qquad \Rightarrow \epsilon_{01\cdots(d-1)} = (-1)^{d-1}, \tag{1.119}$$

and in curved indices by

$$\epsilon^{\mu_1\cdots\mu_d} = \sqrt{|g|}\, e^{\mu_1}{}_{a_1}\cdots e^{\mu_d}{}_{a_d}\epsilon^{a_1\cdots a_d}, \tag{1.120}$$

so, with upper indices, it is independent of the metric and, in curved indices, which we underline to distinguish them from the tangent-space ones,

$$\epsilon^{\underline{0}\cdots\underline{(d-1)}} = +1, \qquad \epsilon_{\underline{0}\cdots\underline{(d-1)}} = g = (-1)^{d-1}|g|. \tag{1.121}$$

The contraction of n indices of two ϵ symbols gives

$$\epsilon^{\mu_1\cdots\mu_n\rho_1\cdots\rho_{(d-n)}}\epsilon_{\mu_1\cdots\mu_n\sigma_1\cdots\sigma_{(d-n)}} = n!(d-n)!\, g\, g^{\rho_1\cdots\rho_{(d-n)}}{}_{\sigma_1\cdots\sigma_{(d-n)}}, \tag{1.122}$$

where

$$\begin{aligned} g^{\rho_1\cdots\rho_{(d-n)}}{}_{\sigma_1\cdots\sigma_{(d-n)}} &= g_{[\rho_1}{}^{\sigma_1}\cdots g_{\rho_{(d-n)}]}{}^{\sigma_{(d-n)}} \\ &= \frac{1}{(d-n)!}\begin{vmatrix} g_{\rho_1}{}^{\sigma_1} & \cdots & g_{\rho_1}{}^{\sigma_{(d-n)}} \\ \vdots & \vdots & \vdots \\ g_{\rho_{(d-n)}}{}^{\sigma_1} & \cdots & g_{\rho_{(d-n)}}{}^{\sigma_{(d-n)}} \end{vmatrix}. \end{aligned} \tag{1.123}$$

---

and, using the Ricci identities for the l.h.s.,

$$\nabla_{(\alpha}\nabla_{\beta)}k_\gamma = \tfrac{1}{2}k_\lambda R_{\alpha\gamma\beta}{}^\lambda + \tfrac{1}{2}k_\lambda R_{\beta\gamma\alpha}{}^\lambda.$$

Now, using the Bianchi identities for the metric-compatible torsionless curvature we get

$$\nabla_{(\alpha}\nabla_{\beta)}k_\gamma = -k_\lambda R^\lambda{}_{\alpha\beta\gamma} + \tfrac{1}{2}k_\lambda R^\lambda{}_{\gamma\beta\alpha}.$$

Combining this equation with the Ricci identity

$$\nabla_{[\alpha}\nabla_{\beta]}k_\gamma = -\tfrac{1}{2}k_\lambda R_{\alpha\beta\gamma}{}^\lambda = \tfrac{1}{2}k_\lambda R^\lambda{}_{\gamma\alpha\beta},$$

we obtain Eq. (1.116).

4.5 Gravity as a gauge theory

In Ref. [916] MacDowell and Mansouri formulated gravity as the gauge theory of the Poincaré group and supergravity as the gauge theory of the super-Poincaré group.[18] This approach was later extended successfully to many other situations, and it is interesting enough to review it briefly here because the similarities with and differences of gravity from the gauge theories of internal symmetries (some of which we have already mentioned) are manifest in this formulation. Here we will loosely follow Refs. [983, 565].

One of the differences we observed in Section 4.4 between the first-order formalism for gravity using Vielbeins and spin connection and a pure gauge theory is that we did not have an interpretation of the Vielbeins as gauge fields. Furthermore, our intuition tells us that, if gravity can be interpreted as a gauge theory at all, it cannot be a gauge theory of the Lorentz group alone and that at least gauge translations should be introduced into the game. We should then consider the gauging of the Poincaré group. It is worth stressing here that we are talking about the "Poincaré group of the tangent space." That is, at each point in the base manifold, which may but need not be invariant under any translational isometry, we consider inhomogeneous transformations of Lorentz vectors preserving the Minkowski metric. The relation between these gauge transformations and GCTs is one of the subtle points of this formulation of gravity.

To find the generators of the Poincaré group and their commutation relations, we can use the representation in position space (as differential operators) or alternatively we can use the following representation by $(d+1)\times(d+1)$ matrices of Poincaré transformations composed of a translation a^a and a Lorentz transformation $\Lambda^a{}_b$:

$$\begin{pmatrix} 1 \\ v'^{\,a} \end{pmatrix} = \left(\begin{array}{c|c} 1 & 0 \\ \hline a^a & \Lambda^a{}_b \end{array}\right) \begin{pmatrix} 1 \\ v^b \end{pmatrix}. \tag{4.146}$$

This representation is suggestive because of its $(d+1)$-dimensional homogeneous form. We will see later that there is a reason for its existence.

We give here again the non-vanishing commutators of the generators $\{M_{ab}, P_a\}$:

$$\begin{aligned} [M_{ab}, M_{cd}] &= -M_{eb}\Gamma_{\rm v}(M_{cd})^e{}_a - M_{ae}\Gamma_{\rm v}(M_{cd})^e{}_b, \\ [P_c, M_{ab}] &= -P_d\Gamma_{\rm v}(M_{ab})^d{}_c. \end{aligned} \tag{4.147}$$

Here $\Gamma_{\rm v}(M_{ab})^d{}_c$ is the matrix corresponding to the generator M_{ab} in the vector representation of the Lorentz group. The last commutator says that the d generators of translations P_a can be understood as the components of a Lorentz vector. Observe that P_a acts trivially on objects with Lorentz indices. It would act non-trivially on objects with a non-trivial "$(d+1)$th" index in the above representation, but by construction they do not exist.

For each generator we would introduce a gauge field: the spin connection $\omega_\mu{}^{ab}$ for the Lorentz subalgebra plus d new gauge fields for the translation subalgebra. Our theory has d Vielbein fields with Lorentz-vector indices and it is natural to try to interpret them as the gauge fields of translations; the gauge field of the Poincaré group would, tentatively, be, in some representation Γ,

$$A_\mu = \tfrac{1}{2}\omega_\mu{}^{ab}\Gamma(M_{ab}) + e_\mu{}^a\Gamma(P_a)\,. \tag{4.148}$$

---

[18] The earliest work on this subject is in Ref. [1207].

Observe that, since P_a does not act on objects with Lorentz indices, the covariant derivative contains in practice only the spin connection.

If we can reproduce Einstein's theory with these elements, we could say that Einstein's theory is the pure gauge theory of the Poincaré group. We are going to see whether this is possible. First we determine the effect of gauge transformations using the standard formalism of Appendix A. If σ^{ab} and ξ^a are the infinitesimal gauge parameters of Lorentz rotations and translations, then

$$\delta\omega_\mu{}^{ab} = -\mathcal{D}_\mu\sigma^{ab}, \qquad \delta e_\mu{}^a = -\mathcal{D}_\mu\xi^a + \sigma^a{}_b e_\mu{}^b. \tag{4.149}$$

In both cases $\mathcal{D}$ stands for the gauge-covariant derivative (no Levi-Civita connection is contained in it because, for the moment, we have no metric but a gauge field $e_\mu{}^a$). It is useful to compare the second Eq. (4.149) with the effect of an infinitesimal GCT generated by the world vector ξ^μ (unrelated in principle to the Lorentz vector ξ^a):

$$\begin{aligned} \delta_\xi x^\mu &= \xi^\mu, \\ \delta_\xi e^a{}_\mu &= -\xi^\nu\partial_\nu e^a{}_\mu - \partial_\mu\xi^\nu e^a{}_\nu = -\mathcal{D}_\mu(\xi^\nu e^a{}_\nu) + 2\xi^\nu\mathcal{D}_{[\mu}e^a{}_{\nu]} - (\xi^\nu\omega_\nu{}^a{}_b)e^b{}_\mu. \end{aligned} \tag{4.150}$$

The covariant derivative is, again, the Poincaré (Lorentz) gauge one. The effect of an infinitesimal reparametrization is identical to the effect of a P_a gauge transformation with parameter $\xi^a = \xi^\mu e^a{}_\mu$ plus a local Lorentz transformation with parameter $\sigma^{ab} = \xi^\mu\omega_\mu{}^{ab}$ *if* $\mathcal{D}_{[\mu}e^a{}_{\nu]}$ *vanishes*.

We know that this condition is equivalent to the vanishing of torsion and we know that this constraint allows us to express $\omega_\mu{}^{ab}$ in terms of $e^a{}_\mu$. If we implement this constraint in our gauge theory, it will automatically become invariant under reparametrizations.[19] It is implemented in the first-order formalism of Section 4.4, where it appears as the equation of motion of $\omega_\mu{}^{ab}$.

The next step is to construct the gauge field strength:

$$\begin{aligned} R_{\mu\nu} &= \tfrac{1}{2}R_{\mu\nu}{}^{ab}\Gamma(M_{ab}) + R_{\mu\nu}{}^a\Gamma(P_a)\,, \\ R_{\mu\nu}{}^{ab} &= 2\partial_{[\mu}\omega_{\nu]}{}^{ab} - 2\omega_{[\mu}{}^a{}_c\omega_{\nu]}{}^{cb}, \\ R_{\mu\nu}{}^a &= 2\mathcal{D}_{[\mu}e^a{}_{\nu]}. \end{aligned} \tag{4.151}$$

The last line is identically equal to $-T_{\mu\nu}{}^a$. Thus, we have just learned that torsion can be interpreted in this formalism as the part of the gauge field strength that is associated with translations.

Now the moment to construct the action arrives. As already mentioned, in order to recover the constraint $R_{\mu\nu}{}^a$, the action has to be linear in the curvature components $R_{\mu\nu}{}^{ab}$. The requirement of Lorentz invariance also makes it very difficult to build quadratic actions (different from $\mathrm{Tr}(R \wedge \star R)$, which is wrong for gravity) that are not trivial (i.e. they do not correspond to topological invariants). We are then led to the action Eq. (4.129), which we know is correct.

---

[19] In supergravity formulated as the gauge theory of a supergroup the problem is how to relate supersymmetry transformations (in general super-reparametrizations in superspace) to gauge transformations associated with the supersymmetry and translation generators.

What have we learned by considering the gauge theory of the Poincaré group? Essentially we have given a gauge-field interpretation to Vielbeins (although we have not justified why they have to be invertible) and we have found that constraints are necessary in order to relate Poincaré gauge invariance to reparametrization invariance. The construction of the action is still rather ad hoc.

A slight improvement of the situation was achieved by MacDowell and Mansouri [916] (see also Refs. [339, 330, 1153]), who used it to construct supergravity actions [565]. Working in four dimensions, they considered the anti-de Sitter group $\mathrm{SO}(2,3)$. Upon performing a Wigner–Inönü contraction [817] (which is essentially the zero-cosmological-constant limit), this group becomes the Poincaré group $\mathrm{ISO}(1,3)$ and one recovers our previous results.

More precisely, we introduce $\mathrm{SO}(2,3)$ indices $\hat{a},\hat{b},\dots=-1,0,1,2,3$. The metric is $\hat{\eta}^{\hat{a}\hat{b}}=\mathrm{diag}(++---)$, and the algebra $\mathfrak{so}(2,3)$ can be written in the general form

$$\left[\hat{M}_{\hat{a}\hat{b}},\hat{M}_{\hat{c}\hat{d}}\right]=-\hat{\eta}_{\hat{a}\hat{c}}\hat{M}_{\hat{b}\hat{d}}-\hat{\eta}_{\hat{b}\hat{d}}\hat{M}_{\hat{a}\hat{c}}+\hat{\eta}_{\hat{a}\hat{d}}\hat{M}_{\hat{b}\hat{c}}+\hat{\eta}_{\hat{b}\hat{c}}\hat{M}_{\hat{a}\hat{d}}. \tag{4.152}$$

To perform the contraction, we need to introduce a dimensional parameter. This is, naturally, g, the gauge coupling constant in gauged $d=4$, $N=2$ supergravity; g is related to the AdS_4 radius R and to the cosmological constant Λ by

$$R=1/g=\sqrt{-3/\Lambda}. \tag{4.153}$$

We can now perform a $1+4$ splitting of the indices $\hat{a}=(-1,a)$, $a=0,1,2,3$, to interpret this algebra from the point of view of the Lorentz subalgebra $\mathfrak{so}(1,3)$. On defining

$$\hat{M}_{ab}=M_{ab}, \qquad \hat{M}_{a-1}=-g^{-1}P_a, \tag{4.154}$$

we can rewrite the AdS_4 algebra as follows:

$$\begin{aligned}
[M_{ab},M_{cd}]&=-\eta_{ac}M_{bd}-\eta_{bd}M_{ac}+\eta_{ad}M_{bc}+\eta_{bc}M_{ad},\\
[P_c,M_{ab}]&=-2P_{[a}\eta_{b]c}, \qquad [P_a,P_b]=-g^2M_{ab}.
\end{aligned} \tag{4.155}$$

Taking the limit $g\to 0$, we recover the Poincaré algebra.

We could equally well have started with the four-dimensional de Sitter group $\mathrm{SO}(1,4)$. The difference is that, instead of having an extra timelike direction (which we have denoted with a -1 index), we have an extra spacelike direction (which we would denote with a 4 index). The two spaces (and groups) are related by analytic continuation $x^{-1}\to x^4$, and, in the contraction of the extra dimension, we would find that the sign of the cosmological constant is reversed ($g\to ig$). We will use a general notation and point out where differences between the two groups could arise. However, one should keep in mind that only the anti-de Sitter space is a good background for QFT and only its group can consistently be supersymmetrized.

The gauge theory of the AdS_4 group is just a particular case of the general construction in Appendix A.2.3. We can also perform the contraction in the gauge field and curvature. First, we split the indices in the connection and then we rescale the gauge fields inversely to the generators:

$$\begin{aligned}\hat{\omega}_\mu &= \tfrac{1}{2}\hat{\omega}_\mu{}^{\hat{a}\hat{b}}\Gamma\left(\hat{M}_{\hat{a}\hat{b}}\right) = \tfrac{1}{2}\hat{\omega}_\mu{}^{ab}\Gamma\left(\hat{M}_{ab}\right) + \hat{\omega}_\mu{}^{a,-1}\Gamma\left(\hat{M}_{a,-1}\right) \\ &= \tfrac{1}{2}\omega_\mu{}^{ab}\Gamma(M_{ab}) + e^a{}_\mu\Gamma(P_a)\,, \end{aligned} \tag{4.156}$$

where

$$\hat{\omega}_\mu{}^{ab} = \omega_\mu{}^{ab}, \qquad \hat{\omega}_\mu{}^{a,-1} = -g e^a{}_\mu. \tag{4.157}$$

In this scheme, linear momentum and Vierbeins are on the same footing as the rest of the generators and gauge fields. This is obviously due to the semisimple nature of the AdS_4 group. There is some resemblance between this structure and the idea of grand unification in particle physics, although there are also obvious differences.

We can also split and rescale the curvature components, expressing everything in terms of Lorentz tensors:

$$\hat{R}_{\mu\nu}{}^{ab} = R_{\mu\nu}{}^{ab} + 2g^2 e^{[a}{}_\mu e^{b]}{}_\nu, \qquad \hat{R}_{\mu\nu}{}^{a,-1} = 2g\mathcal{D}_{[\mu}e^a{}_{\nu]}. \tag{4.158}$$

Now we can address again the construction of a quadratic action for this group. To have diffeomorphism invariance, the Lagrangian has to be a 4-form that we can integrate over a four-dimensional manifold and, therefore, the exterior product of two curvature terms $R \wedge R$. We now have to saturate the $\mathfrak{so}(2,3)$ indices. If we did it with the Killing metric, we would manifest SO(2, 3) invariance but the Lagrangian would be a total derivative, as in any Yang–Mills theory. Thus, we have to give up explicit SO(2, 3) invariance. We have to keep Lorentz invariance, though, and with the only two invariant tensors of the Lorentz group $(\eta_{ab}, \epsilon_{abcd})$ we can build two terms:

$$\hat{R}^{ab} \wedge \hat{R}^{cd}\epsilon_{abcd}, \qquad \hat{R}^{a,-1} \wedge \hat{R}^{b,-1}\eta_{ab}. \tag{4.159}$$

The second term is not invariant under parity and for this reason it is discarded. The inclusion of this term would also introduce torsion and it would also lead to the existence of non-invertible Vierbeins (see the discussion in Ref. [565]).

The first term can also be given an SO(2,3)-invariant origin [1153]: by introducing a constant vector $V^{\hat{a}} = \eta^{\hat{a}}{}_{-1}$ it can be written using the invariant tensor $\hat{\epsilon}$, and we obtain the action

$$S = \alpha \int \hat{R}^{\hat{a}\hat{b}} \wedge \hat{R}^{\hat{c}\hat{d}} V^{\hat{e}}\hat{\epsilon}_{\hat{a}\hat{b}\hat{c}\hat{d}\hat{e}}. \tag{4.160}$$

This is only formally SO(2, 3) invariant because the vector would change under AdS_4 transformations. Nevertheless, this form of the action is very suggestive.

On expanding this action in terms of Lorentz tensors, we have

$$S = \alpha \int d^4x\, R_{\mu\nu}{}^{ab} R_{\rho\sigma}{}^{cd}\epsilon_{abcd}\epsilon^{\mu\nu\rho\sigma} - 16g^2\alpha \int d^4x\, e\left[R(e,\omega) + 6g^2\right]. \tag{4.161}$$

The first term is a total derivative (proportional to the Euler characteristic, a topological invariant) that does not contribute to the equations of motion, and the second term is the first-order Einstein–Hilbert action with cosmological constant $\Lambda = -3g^2$. In the

$g \to 0$ limit (provided that $\alpha \sim g^{-2}$) we recover the usual Einstein–Hilbert action plus a topological term. Observe that the variation of the action under P_a gauge transformations is proportional to torsion terms and, thus, vanishes on-shell.

This is a very attractive result, which, however, leaves some questions unanswered, such as the reason for the invertibility of the Vierbein and the value of the vector $V^{\hat{a}}$. A possible solution has been proposed by Wilczek [1255].

To finish this section, we should mention that the gauge approach has been extended to larger groups, such as the full d-dimensional affine group. A comprehensive review on these developments may be found in Ref. [738].

4.6 Teleparallelism

In this section we provide a short introduction to relativistic theories of gravity based on *teleparallelism*, i.e. theories in which there is a well-defined notion of parallelism of vectors defined at different points. In GR and other generalizations based on the Riemannian or Riemann–Cartan geometry, gravity, described by the metric or Vielbein fields, is characterized by a curvature; therefore, parallel transport is path dependent and there is no such well-defined (path-independent) notion of parallelism. Teleparallelism is based on the Weitzenböck geometry and the Weitzenböck connection $W_{\mu\nu}{}^{\rho}$ described in Section 1.4.1, which has identically vanishing curvature (but non-vanishing torsion[20]).

These theories are interesting for several reasons: first of all, GR can be viewed as a particular theory of teleparallelism and, thus, teleparallelism could be considered at the very least as a different point of view that can lead to the same results. Of course, there are teleparallel theories different from and even inconsistent with GR. Second, in this framework, one can define an energy–momentum tensor for the gravitational field that is a true tensor under all GCTs. This is the reason why teleparallelism was reconsidered[21] by Møller in 1961 [962] when he was studying the problem of defining an energy–momentum *tensor* for the gravitational field [960, 961]. The idea was taken over by Pellegrini and Plebański in Ref. [1025], which constructed the general Lagrangian for these theories. The third reason why these theories are interesting is that they can be seen as gauge theories of the translation group [735, 351] (not the full Poincaré group) and, thus, they give an alternative interpretation of GR.

The basic field in these theories is the Vielbein $e^{a}{}_{\mu}$. This field has d^2 independent components, while the metric has only $d(d+1)/2$. The extra independent components that the Vielbein field has are those of an antisymmetric $d \times d$ tensor, such as the electromagnetic-field-strength tensor $F_{\mu\nu}$, and that is why Einstein thought that these theories could describe gravitation and electromagnetism in a unified way. In the standard Vielbein formalism (Weyl's), the extra $d(d-1)/2$ independent components of the Vielbein field are removed by introducing local Lorentz invariance, with a Lorentz connection that is not an independent field but is built out of the Vielbeins. Here, we are not interested a priori in having this local invariance, and, in principle, we will construct only theories that are invariant under GCTs and global Lorentz transformations. Thus, as we will see, these theories describe in general more degrees of freedom than just those of a graviton.

---

[20] It is possible to have a non-trivial theory with vanishing curvature and torsion if the non-metricity tensor does not vanish. In Ref. [977] a theory equivalent to GR based on this geometry was constructed.

[21] Teleparallelism had originally been considered by Einstein, who studied it as a unified theory of gravitation and electromagnetism in Refs. [493, 494, 495, 496, 497] (see also Ref. [891]) until Ref. [502] showed that the particular theory considered by him was inconsistent.

The construction of the Lagrangian of these theories is fairly simple: we look for terms that have the required invariances and are, at most, quadratic in derivatives of the Vielbeins. The elementary building blocks are the Ricci rotation coefficients $\Omega_{\mu\nu}{}^{a} = \partial_{[\mu}e^{a}{}_{\nu]}$ that transform as tensors (2-forms) under GCTs and as vectors under (global) Lorentz transformations. In the context of the Weitzenböck geometry, the $-2\Omega_{\mu\nu}{}^{a}$s are the components of the torsion of Weitzenböck connection and, since there is no curvature tensor available, it is only natural to construct the Lagrangian using them.

In any dimension there are three terms with the required properties (they transform as densities under GCTs and as scalars under Lorentz transformations, being quadratic in first partial derivatives of the Vielbeins): the *Weitzenböck invariants* $I_1, \ldots, I_3$,

$$I_1 = e\,\Omega_{\mu\nu\rho}\Omega^{\mu\nu\rho}, \qquad I_2 = e\,\Omega_{\mu\nu\rho}\Omega^{\rho\nu\mu}, \qquad I_3 = e\,\Omega_{\mu\rho}{}^{\rho}\Omega^{\mu}{}_{\sigma}{}^{\sigma}. \tag{4.162}$$

There is another invariant I_4, which is quadratic only in $d = 4$:

$$I_4 = \epsilon^{\mu_1\cdots\mu_{d-3}\nu_1\nu_2\nu_3}\Omega_{\mu_1\rho_1}{}^{\rho_1}\Omega_{\mu_2\rho_2}{}^{\rho_2}\cdots\Omega_{\mu_{d-3}\rho_{d-3}}{}^{\rho_{d-3}}\Omega_{\nu_1\nu_2\nu_3}, \tag{4.163}$$

but it is not invariant under parity transformations (a further requirement) and it is usually not considered. Also, e by itself is another invariant (a cosmological-constant term) that we will not consider. Observe that all the Weitzenböck invariants involve the inverse Vielbeins $e_a{}^{\mu}$ and are, therefore, highly non-linear in the Vielbeins.

The general teleparallel Lagrangian of Pellegrini and Plebański [1025] is the integral of a linear combination of the Weitzenböck invariants with arbitrary coefficients:

$$\mathcal{L}_{\mathrm{T}} = \sum_{i=1}^{3} c^i I_i. \tag{4.164}$$

Only two of them are really independent since we can choose the overall normalization.

This general Lagrangian, written in differential-form language to relate it to the Poincaré gauge theory of gravity which is customarily written in it (see e.g. Ref. [738]), is known as the *Rumpf Lagrangian* [1088] (see also Refs. [683, 967]).

There are other ways to parametrize this Lagrangian, for instance by splitting Ω_{abc} into several pieces ${}^{(1)}\Omega$, ${}^{(2)}\Omega$, and ${}^{(3)}\Omega$ (*tentor, trator, and axitor*, respectively [683]). First we define

$$\begin{aligned} v_a &\equiv \Omega_{ab}{}^{b}, & {}^{(2)}\Omega_{abc} &= \frac{2}{1-d}\eta_{a[b}v_{c]}, \\ {}^{(3)}\Omega_{abc} &= \Omega_{[abc]}, & {}^{(1)}\Omega_{abc} &= \Omega_{abc} - {}^{(2)}\Omega_{abc} - {}^{(3)}\Omega_{abc}. \end{aligned} \tag{4.165}$$

Then, a Lagrangian equivalent to Pellegrini and Plebański's is [683]

$$\mathcal{L}_{\mathrm{T}} = e\,\Omega^{abc}\sum_{i=1}^{3} a_i\,{}^{(i)}\Omega_{abc}. \tag{4.166}$$

The relation between these two parametrizations is

$$a_1 = c_1 + \tfrac{1}{2}c_2, \quad a_2 = c_1 + \tfrac{1}{2}c_2 + \frac{d-1}{2}c_3, \quad a_3 = c_1 - c_2. \tag{4.167}$$

Another parametrization based on v_a, the tensors

$$a^{a_1\cdots a_{d-3}} = \frac{1}{3!}\epsilon^{a_1\cdots a_{d-3}b_1b_2b_3}\Omega_{b_1b_2b_3}, \qquad t_{abc} = \Omega_{a(bc)} - {}^{(2)}\Omega_{a(bc)}, \tag{4.168}$$

and the invariants $v^2, t^2,$ and a^2 can be found in Ref. [736].

4.6.1 The linearized limit

Our goal now is to try to understand which kind of theories are those defined by the Lagrangian Eq. (4.164). First, we observe with Møller [962] that, for $c_1 = 1, c_2 = 2,$ and $c_4 = -4$, this Lagrangian is identical (up to total derivatives) to the Einstein–Hilbert Lagrangian, and, therefore, gives the vacuum Einstein equations.[22] The Lagrangian turns out to be invariant under not just global but also local Lorentz transformations, and the only degrees of freedom left are (we know) those of the graviton. For general values of the parameters, the analysis is more complicated and it is convenient to start by studying the linear limit. To this end, we split the Vielbeins into their vacuum (Minkowski) values plus perturbations. Working in Cartesian coordinates for simplicity, we write

$$e^a{}_\mu = \delta^a{}_\mu + A^a{}_\mu. \tag{4.169}$$

For the inverse Vielbeins, we have

$$e_a{}^\mu = \delta_a{}^\mu - \delta_b{}^\mu\delta_a{}^\nu A^b{}_\nu + \mathcal{O}(A^2). \tag{4.170}$$

To this order we can unambiguously trade curved and flat indices, and the above formula can be rewritten:

$$e_a{}^\mu = \delta_a{}^\mu - A^\mu{}_a + \mathcal{O}(A^2), \qquad A^\mu{}_a \equiv \delta_b{}^\mu\delta_a{}^\nu A^b{}_\nu. \tag{4.171}$$

The metric perturbation that we have called $h_{\mu\nu}$ in previous chapters is given by the symmetric part of A at lowest order:

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} + \mathcal{O}(A^2), \quad h_{\mu\nu} \equiv 2A_{(\mu\nu)}, \quad b_{\mu\nu} \equiv 2A_{[\mu\nu]}, \quad A_{\mu\nu} \equiv \delta_{a\mu}A^a{}_\mu. \tag{4.172}$$

With these definitions it is straightforward to obtain, up to total derivatives, the linear limit of the action for the Lagrangian density Eq. (4.164):

$$\begin{aligned} S_{\mathrm{T}}[h,b] = \int d^dx \Big\{ & \frac{1}{16}(2c_1 + c_2)\partial_\mu h_{\nu\rho}\partial^\mu h^{\nu\rho} - \frac{1}{16}(2c_1 + c_2 - c_3)\partial_\mu h_{\nu\rho}\partial^\nu h^{\mu\rho} \\ & - \frac{1}{8}c_3\partial_\mu h\partial_\nu h^{\nu\mu} + \frac{1}{16}c_3(\partial h)^2 - \frac{1}{16}[4c_1 + 2(c_2 + c_3)]\partial_\mu h_{\nu\rho}\partial^\rho b^{\nu\mu} \\ & + \frac{1}{16}\partial_\mu b_{\nu\rho}\partial^\mu b^{\nu\rho} - \frac{1}{16}(2c_1 - 3c_2 - c_3)\partial_\mu b_{\nu\rho}\partial^\rho b^{\nu\mu} \Big\}. \end{aligned} \tag{4.173}$$

---

[22] In fact, this theory is sometimes referred to as the *teleparallel equivalent of GR*.

Table 4.1. Values of the parameters c_i in the general Lagrangian of Pellegrini and Plebański Eq. (4.164) for several theories: GR, the viable models, the Yang–Mills-type model (YM), and the von der Heyde model (vdH) [751] (which is one of the viable ones with $\lambda = 0$)

	GR	Viable	YM	vdH
c_1	1	$2 - \lambda$	2	2
c_2	2	2λ	0	0
c_3	-4	-4	0	-4

The first four terms are familiar to us: up to coefficients, they are the same terms as those that appear in the Fierz–Pauli Lagrangian Eq. (3.84). The last two terms are also well known: up to coefficients, they are exactly those that appear in the Lagrangian of the Kalb–Ramond 2-form field, which we still have not seen. The third term in the second line represents a coupling (already at the linear level) between these two fields.

Now, it is clear that it is not possible to recover solutions of the vacuum Fierz–Pauli theory if the coupling terms have a non-zero coefficient: a non-vanishing h field is a source for a non-vanishing b field and vice versa. Thus, the only theories which we expect to be phenomenologically viable are those in the family

$$2c_1 + c_2 + c_3 = 0. \tag{4.174}$$

Furthermore, both in the Kalb–Ramond and in the Fierz–Pauli cases, for certain choices of the coefficients, the action has a gauge invariance,

$$\delta_\epsilon h_{\mu\nu} = -2\partial_{(\mu}\epsilon_{\nu)}, \qquad \delta_\eta b_{\mu\nu} = 2\partial_{[\mu}\eta_{\nu]}, \tag{4.175}$$

whose existence is crucial for its consistent quantization. We also expect that only when these gauge invariances are present will the theory be consistent. It turns out that all these conditions are simultaneously met: let us eliminate c_3 using Eq. (4.174) and then, calling $c_2 = 2\lambda$, the action can be rewritten in the form

$$S_{\mathrm{T}}[h, b] = \frac{c_1 + \lambda}{2} S_{\mathrm{FP}}[h] + \frac{c_1 - \lambda}{2} S_{\mathrm{KR}}[b], \tag{4.176}$$

where $S_{\mathrm{FP}}[h]$ is the Fierz–Pauli action given in Eq. (3.84) and $S_{\mathrm{KR}}[b]$ is the Kalb–Ramond action

$$S_{\mathrm{KR}}[b] = \int d^d x \, \frac{1}{12} H^2, \qquad H_{\mu\nu\rho} \equiv 3\partial_{[\mu} b_{\nu\rho]}, \qquad H^2 = H_{\mu\nu\rho} H^{\mu\nu\rho}. \tag{4.177}$$

For $c_1 = \lambda$ ($c_2 = 2\lambda$, $c_3 = -4\lambda$), the Kalb–Ramond Lagrangian disappears. Up to an overall normalization constant and a total derivative, this teleparallel Lagrangian is completely equivalent to the Einstein–Hilbert Lagrangian, as we mentioned before. For $c_1 = -\lambda$ the Fierz–Pauli Lagrangian disappears and only the Kalb–Ramond Lagrangian remains. This theory does not describe gravity. If we are always going to keep the Fierz–Pauli Lagrangian, then it makes sense to set $c_1 = 2 - \lambda$ ($c_3 = -4,\ c_2 = 2\lambda$) and keep the one-parameter family of actions

$$S_{\rm T}[h, b] = S_{\rm FP}[h] + (1 - \lambda) S_{\rm KR}[b], \tag{4.178}$$

which represent viable models of gravity (in the sense that they fulfill the above requirements) based on teleparallelism (see Table 4.1). The case $\lambda = 0$ is the model proposed in Ref. [751].

Of course, we know that the full non-linear theory will be consistent only if additional conditions are satisfied. In particular, we know from our results in Chapter 3 that the quantization of the spin-2 field $h_{\mu\nu}$ will be consistent only if it couples to the total energy–momentum tensor, the sum of the spin-2 energy–momentum tensor and the Kalb–Ramond energy–momentum tensor, although the presence of the Kalb–Ramond field could modify this result. Checking that this is (or not) the case in the above family of theories requires an expansion to order $\mathcal{O}(A^3)$ that would be interesting to do.

It is amusing to compare these results with the linearized limit of the low-energy string effective action (see Chapter 21). The linearized actions are identical, except for the presence of the dilaton in the string case. However, the non-linear actions are quite different: in the string case, we simply have standard gravity coupled to matter (the Kalb–Ramond field) that appears only quadratically (at least to lowest order in α'), whereas, in the teleparallel case, the Kalb–Ramond field should also appear non-linearly in the full action.

It is possible to view the theories of teleparallelism as gauge theories of the group of translations [735, 351], with the Vielbeins playing the role of gauge vectors, but we will not enter into this interesting aspect.

From this point of view, if BHs can be understood as particle-like objects, looking for BHs with a well-defined scalar charge is utterly hopeless. One should look instead for "black strings" and instantons, and in due time we will do so and find them.[4]

There is another point of view concerning scalar fields: in some cases they should be interpreted not as matter fields but as "local coupling constants" (as in the case of the string-theory dilaton) or, more generally, as *moduli fields*, which we will define in Chapter 15, in which case they should be treated as backgrounds and there would be no room for the notion of scalar charge.

In conclusion, if we want to find new BH solutions, we need to couple the Einstein–Hilbert action to matter fields that have associated conserved charges. The charges must be those of point-particles or we will naturally obtain solutions describing extended objects instead of black holes. Thus, we have to consider vector fields, and the simplest one is an Abelian vector field A_μ. We are going to study in some detail the resulting system because later we will find generalizations of all the concepts and formulae developed here.

12.2 The Einstein–Maxwell system

The action for gravity coupled to an Abelian vector field A_μ is the so-called Einstein–Maxwell action[5] obtained by adding the Einstein–Hilbert and the Maxwell action with $\eta_{\mu\nu}, \partial_\mu$, and d^4x replaced by $g_{\mu\nu}, \nabla_\mu$, and d^4x:

$$\boxed{S_{\rm EM}[g_{\mu\nu}, A_\mu] = S_{\rm EH}[g] + \frac{1}{c}\int d^4x\sqrt{|g|}\,\left[-\tfrac{1}{4}F^2\right].} \tag{12.15}$$

$F_{\mu\nu}$ is the field strength of the electromagnetic vector field A_μ and is again given by

$$F_{\mu\nu} = 2\partial_{[\mu}A_{\nu]}, \qquad F^2 = F_{\mu\nu}F^{\mu\nu}, \tag{12.16}$$

since, in the absence of torsion, $\nabla_{[\mu}A_{\nu]} = \partial_{[\mu}A_{\nu]}$. The components of A_μ and $F_{\mu\nu}$ in a given coordinate system are customarily split in this way,

$$(A_\mu) = (\phi, -\vec{A}), \qquad (F_{\mu\nu}) = \begin{pmatrix} 0 & E_1 & E_2 & E_3 \\ -E_1 & 0 & -B_3 & B_2 \\ -E_2 & B_3 & 0 & -B_1 \\ -E_3 & -B_2 & B_1 & 0 \end{pmatrix}, \tag{12.17}$$

---

[4] This argument really applies to pseudoscalar fields.

[5] In this section we work in the Heaviside system of units, so the Coulomb force between two charges is

$$\frac{1}{4\pi}\frac{q_1 q_2}{r_{12}^2}. \tag{12.14}$$

In the Gaussian system we should replace $1/(4c)$ by $1/(16\pi c)$ and the factor of 4π disappears from the Coulomb force. The dimensions of the vector field A_μ are $\mathrm{M}^{1/2}\mathrm{L}^{1/2}\mathrm{T}^{-1}$ (that is, L^{-1} in natural units $\hbar = c = 1$) and the electric charge's units are $M^{1/2}L^{3/2}T^{-1}$, so it is dimensionless in natural units. At the end we will introduce another system of units, which will be the one we more often will work with, taking $c = 1$ and replacing the factor of $1/(4c)$ in front of F^2 by $1/\big(64G_{\rm N}^{(4)}\big)$.

where $\vec{E} = (E_1, E_2, E_3)$ and $\vec{B} = (B_1, B_2, B_3)$ are the electric and magnetic 3-vector fields in that coordinate system, and, thus, with $\vec{\nabla} = (\partial_{\underline{1}}, \partial_{\underline{2}}, \partial_{\underline{3}})$,

$$\begin{cases} E_i = F_{\underline{0i}}, \\ B_i = -\frac{1}{2}\epsilon_{ijk} F_{\underline{kl}}, \end{cases} \qquad \Leftrightarrow \qquad \begin{cases} \vec{E} = -\vec{\nabla}\phi - \dfrac{1}{c}\dfrac{\partial}{\partial t}\vec{A}, \\ \vec{B} = \vec{\nabla} \times \vec{A}. \end{cases} \tag{12.18}$$

The field strength (and the action) are invariant under the Abelian gauge transformations

$$A'_\mu = A_\mu + \partial_\mu \Lambda \tag{12.19}$$

with smooth, gauge parameter Λ. Depending on which gauge group we consider ($\mathbb{R}$ or $\mathrm{U}(1)$), Λ must be a single-valued or multivalued function.[6] In differential-form language

$$A = A_\mu dx^\mu, \qquad A' = A + d\Lambda, \qquad F = \tfrac{1}{2} F_{\mu\nu} dx^\mu \wedge dx^\nu = dA, \tag{12.20}$$

and the gauge invariance of F is a consequence of $d^2 = 0$. Using these differential forms, the Maxwell action can be rewritten as follows:

$$S_{\mathrm{M}}[A] = \frac{1}{8c} \int F \wedge \star F. \tag{12.21}$$

Observe that there is no matter charged with respect to A_μ in this system. This is analogous to the presence of no matter fields in the Einstein–Hilbert action. However, the Einstein–Hilbert action contains the self-coupling of gravity and therefore the presence of a coupling constant in it makes sense, whereas in the Maxwell theory there are no *direct* interactions between photons and, in principle, there is neither an electromagnetic coupling constant nor a unit of electric charge. We will see that things are a bit more complicated in the presence of gravity, through which photons do interact.

The equations of motion of $g_{\mu\nu}$ and A_μ are

$$G_{\mu\nu} - \frac{8\pi G_{\mathrm{N}}^{(4)}}{c^4} T_{\mu\nu} = 0, \tag{12.22}$$

$$\nabla_\mu F^{\mu\nu} = 0 \quad \text{(Maxwell's equation)}, \tag{12.23}$$

where

$$T_{\mu\nu} = \frac{-2c}{\sqrt{|g|}} \frac{\delta S_{\mathrm{M}}[A]}{\delta g^{\mu\nu}} = F_{\mu\rho} F_\nu{}^\rho - \tfrac{1}{4} g_{\mu\nu} F^2 \tag{12.24}$$

is the energy–momentum tensor of the vector field, which is traceless[7] in $d = 4$. The

---

[6] If the gauge group is $\mathbb{R}$, the elements of the group will be $e^{\Lambda/L}$, whereas, if it is U(1), they will be $e^{i\Lambda/L}$, where L is a constant introduced to make the exponent dimensionless because Λ is dimensionful. In the second case Λ will have to be identified with $\Lambda + 2\pi L$. When there is a unit of charge, L is related to it.

[7] This property is associated with the invariance of the Maxwell Lagrangian in curved spacetime under Weyl rescalings of the metric,

$$g'_{\mu\nu} = \Omega^2(x) g_{\mu\nu}. \tag{12.25}$$

In fact, if $\Omega = e^\sigma$, then for infinitesimal transformations $\delta_\sigma g_{\mu\nu} = 2\sigma(x) g_{\mu\nu}$ we have

$$\delta_\sigma S_{\mathrm{M}} = \frac{\delta S_{\mathrm{M}}}{\delta g_{\mu\nu}} \delta_\sigma g_{\mu\nu} \sim \sigma T^{\mu\nu} g_{\mu\nu} = 0. \tag{12.26}$$

tracelessness of the electromagnetic energy–momentum tensor implies that $R = 0$ and the Einstein equation takes the simpler form

$$R_{\mu\nu} = \frac{8\pi G_{\mathrm{N}}^{(4)}}{c^4} T_{\mu\nu}. \tag{12.27}$$

On taking the divergence of the Einstein equation and using the contracted Bianchi identity for the Einstein tensor $\nabla_\mu G^{\mu\nu} = 0$, we find

$$F_{\nu\rho} \nabla_\mu F^{\mu\rho} - \tfrac{3}{2} F^{\mu\rho} \nabla_{[\mu} F_{\rho\nu]} = 0. \tag{12.28}$$

Since the Levi-Civita connection is symmetric,

$$\nabla_{[\mu} F_{\rho\nu]} = \partial_{[\mu} F_{\rho\nu]} = 0 \quad \text{(the Bianchi identity)} \tag{12.29}$$

identically, using the definition of $F_{\mu\nu}$, and then we see that the Einstein equation implies generically the Maxwell equation. Using Eq. (1.70), the Maxwell equation can also be written in a simpler, equivalent, form:

$$\partial_\mu \left(\sqrt{|g|} F^{\mu\nu} \right) = 0. \tag{12.30}$$

The equations are written in terms of the field strength F and usually they are solved in terms of it. However, we are ultimately interested in the vector field A itself and we have to make sure that the F we obtain is such that it is related to some vector field by Eq. (12.16) or Eq. (12.20). It turns out that, locally, A exists if the electromagnetic Bianchi identity Eq. (12.29) is satisfied.[8]

---

[8] In fact, if we are given F and the Bianchi identity is satisfied, we can always find the corresponding vector potential by using, e.g., the formula

$$\boxed{A_\mu(x) = -\int_0^1 d\lambda \lambda x^\nu F_{\mu\nu}(\lambda x).} \tag{12.31}$$

To check this formula it is necessary to use the Bianchi identity: taking the curl of the l.h.s.,

$$\partial_{[\rho} A_{\mu]}(x) = -\int_0^1 d\lambda \lambda \partial_{[\rho} [x^\nu F_{\mu]\nu}(\lambda x)], \tag{12.32}$$

and operating,

$$\partial_{[\rho} \big(F_{\mu]\nu}(\lambda x) \big) = \lambda \partial_{[\rho} F_{\mu]\nu}(\lambda x) - \tfrac{1}{2} \lambda (\partial_\nu F_{\rho\mu})(\lambda x), \tag{12.33}$$

where the Bianchi identity Eq. (12.29) has been used in the last identity, we obtain

$$\begin{aligned} \partial_{[\rho} A_{\mu]}(x) &= \tfrac{1}{2} \int_0^1 d\lambda \big\{ \lambda^2 x^\nu \partial_\nu F_{\rho\mu}(\lambda x) - \lambda F_{\mu\rho}(\lambda x) \big\} = \tfrac{1}{2} \int_0^1 d\lambda \frac{d}{d\lambda} \big[\lambda^2 F_{\rho\mu}(\lambda x) \big] \\ &= \frac{1}{2} F_{\rho\mu}(x). \end{aligned} \tag{12.34}$$

This means that, up to gauge transformations, the field strength determines completely the potential in the Maxwell theory, which one can see as the simplest Abelian Yang–Mills theory. Interestingly, this is no longer true in the non-Abelian Yang–Mills case, as first noticed in Refs. [1278, 448].

The Bianchi identity can also be written in this form (by contracting Eq. (12.29) with $\epsilon^{\mu\nu\rho\sigma}$, introducing it into the partial derivative (because it is constant), and using the definition of the Hodge dual and Eq. (12.30) for the divergence):

$$\nabla_\mu \star F^{\mu\sigma} = 0. \tag{12.35}$$

In the language of differential forms, the Maxwell equation and Bianchi identity are

$$d \star F = 0, \tag{12.36}$$

$$dF = 0, \tag{12.37}$$

and the Bianchi identity is just a consequence of the definition Eq. (12.20) and $d^2 = 0$.

Then, if we work with the field strength, we find that there are two pairs of equations, (12.23) and (12.35) and (12.36) and (12.23), which are (as pairs) invariant if one replaces F by $\star F$ (by virtue of $\star \star F = -F$). This is an *electric–magnetic-duality* transformation. The name is due to the fact that this transformation interchanges the electric and magnetic fields in any given coordinate system according to

$$\vec{E}\,' = \vec{B}, \qquad \vec{B}\,' = -\vec{E}. \tag{12.38}$$

Actually, this pair of homogeneous equations (the Maxwell equation and the Bianchi identity) would be invariant under the (invertible) substitution for F of any linear combination of F and $\star F$. We would have a symmetry of all the equations of motion if the Einstein equation were also invariant under this replacement. We will see in Section 12.7 when this is the case and how the Einstein–Maxwell theory is invariant under electric–magnetic duality.

The four Maxwell equations in Minkowski spacetime can be deduced from the Maxwell equation and the Bianchi identity (two of them imply the existence of the potential A_μ and are equivalent to the latter). We have

$$\begin{aligned} \partial_\mu F^{\mu\nu} = 0 \quad &\Leftrightarrow \quad \begin{cases} \vec{\nabla} \cdot \vec{E} = 0, \\ \vec{\nabla} \times \vec{B} - \dfrac{1}{c}\dfrac{\partial}{\partial t}\vec{E} = 0, \end{cases} \\ \partial_\mu \star F^{\mu\nu} = 0 \quad &\Leftrightarrow \quad \begin{cases} \vec{\nabla} \cdot \vec{B} = 0, \\ \vec{\nabla} \times \vec{E} + \dfrac{1}{c}\dfrac{\partial}{\partial t}\vec{B} = 0. \end{cases} \end{aligned} \tag{12.39}$$

12.2.1 Electric charge

The electric charge can be defined in terms of a source coupled to the electromagnetic field (this is analogous to the energy–momentum-pseudotensor approach for the gravitational field) or in terms of the Noether current associated with the gauge invariance (the approach that leads to Komar's formula and its generalizations for the gravitational field). The two definitions are equivalent and are very closely related to each other because the gauge invariance of the free theory imposes strong constraints on the possible couplings.

Let us first introduce the electric charge using sources. A source for the Maxwell field is described by a current j^{μ}, which naturally couples to the vector field through a term in the action of the form

$$\frac{1}{c^2}\int d^4x\sqrt{|g|}\ [-A_{\mu}j^{\mu}]. \tag{12.40}$$

This additional interaction term spoils the action's gauge invariance unless the source j^{μ} is divergence free,

$$\nabla_{\mu}j^{\mu}=0 \ \Leftrightarrow\ d\star j=0 \qquad (j\equiv j_{\mu}dx^{\mu}), \tag{12.41}$$

which implies the *continuity equation* for the vector density $\mathfrak{j}^{\mu}\equiv\sqrt{|g|}\,j^{\mu}$,

$$\partial_{\mu}\mathfrak{j}^{\mu}=0. \tag{12.42}$$

The continuity equation can be used to establish the local conservation of the electric charge, as explained in Section 2.3, if the electric charge contained in a three-dimensional volume at a given time t, V^3_t, is defined by[9]

$$q(t)=-\frac{1}{c}\int_{\mathrm{V}^3_t}d^3x\,\mathfrak{j}^0, \tag{12.43}$$

or, in a more covariant form,

$$q(t)=\frac{1}{c}\int_{\mathrm{V}^3_t}\star j. \tag{12.44}$$

As explained in Section 2.3, this quantity is not constant: its variation is related to the flux of charge through the boundary of V^3_t. If V^3_t is a constant-time slice of the whole spacetime with no boundary, then the above integrals give the total charge, which will be constant in time. If we can foliate our spacetime with constant-time hypersurfaces, then we take the four-dimensional spacetime V^4 contained in between two constant-time slices $\mathrm{V}^3_{t_1}$ and $\mathrm{V}^3_{t_2}$, integrate the continuity equation over it, and use Stokes' theorem. The boundary of the four-dimensional region we have proposed is made up of the two constant-time slices with opposite orientations, so

$$0=\int_{\mathrm{V}^4}d\star j=\int_{\mathrm{V}^3_{t_1}}\star j-\int_{\mathrm{V}^3_{t_2}}\star j, \tag{12.45}$$

and the total electric charge is constant in time.

Thus, gauge invariance of the action implies that the source is divergence free, and from this the local conservation of the electric charge (and the global conservation of the total electric charge) follows.

On the other hand, in the presence of the source, the Maxwell equation is modified into

$$\nabla_{\mu}F^{\mu\nu}=\frac{1}{c}j^{\nu}, \tag{12.46}$$

---

[9] The sign is conventional.

or, equivalently,

$$d \star F = \frac{1}{c} \star j, \tag{12.47}$$

and, using the antisymmetry of $F_{\mu\nu}$ or $d^2 = 0$, it is trivial to see that, since the l.h.s. of the equation is divergence free, the r.h.s. of the equation is also, for consistency, divergence free, as we knew it had to be in order to preserve the gauge invariance of the action. This is no coincidence: the fact that the r.h.s. of the Maxwell equation is divergence free is in fact the gauge identity associated with the invariance under $\delta A_\mu = \partial_\mu \Lambda$, as we are going to see.

Finally, using the Maxwell equation (12.47), we can rewrite the definition of the total electric charge Eq. (12.44) in terms of the field strength and again use Stokes' theorem. If the boundary of a constant-time slice has the topology of a sphere, S^2, at infinity, we obtain

$$\boxed{q = \int_{\mathrm{S}^2_\infty} \star F} \tag{12.48}$$

which is a useful definition of the total electric charge of a spacetime in terms of the field strength (the electric flux) and which we will generalize further in Part III.

This is the kind of formula that we will use because in the Einstein–Maxwell system there are no fields explicitly written that act as sources for A_μ. Just as in the case of the Maxwell equations in vacuum, we can obtain solutions describing the field of charges. These solutions are singular near the place where the charge ought to be and the solution is not a solution there (there are no charges explicitly included in the system). However, the above expression allows us to calculate the charge that ought to be placed there to produce the flux of electromagnetic field that we observe.[10] We have introduced sources as a device for understanding the definition.

We could also have used the invariance of the Einstein–Maxwell action to find the conserved Noether current and define the electric charge through it.

We studied the invariance of the Maxwell action and found the corresponding Noether current in Minkowski spacetime in Section 3.2.1. The coupling to gravity introduces only minor changes and the conclusion is, again, that the electric charge can be defined by Eq. (12.48).

It is useful to consider a simple example of a source: the current associated with a particle of electric charge q and worldline γ parametrized by $X^\mu(\xi)$. In a manifestly covariant form it is given by

$$j^\mu(x) = qc \int_\gamma dX^\mu \frac{1}{\sqrt{|g|}} \delta^{(4)}[x - X(\xi)], \tag{12.49}$$

where $dX^\mu = d\xi dX^\mu/d\xi$. On making the choice $\xi = X^0$ and integrating over X^0, we obtain

---

[10] Of course, this is just a covariant generalization of the Gauss theorem that relates the flux of electric field through a closed surface to the charge enclosed by it.

$$\begin{aligned} j^\mu(x^0,\vec{x}) &= qc\int dX^0 \frac{dX^\mu}{dX^0}\frac{1}{\sqrt{|g|}}\delta^{(3)}(\vec{x}-\vec{X})\delta(x^0-X^0) \\ &= -qc\frac{dX^\mu}{dx^0}\frac{\delta^{(3)}[\vec{x}-\vec{X}(x^0)]}{\sqrt{|g|}}. \end{aligned} \tag{12.50}$$

If the particle is at rest at the origin in the chosen coordinate system, the current is

$$j^\mu(x^0,\vec{x}) = -qc\delta^{\mu 0}\frac{\delta^{(3)}(\vec{x})}{\sqrt{|g|}}, \tag{12.51}$$

and it is easy to see that q is indeed the electric charge according to the above definitions. The current j^μ is conserved:

$$\begin{aligned} \nabla_\mu j^\mu &\sim \frac{\partial}{\partial x^\mu}\left\{\sqrt{|g(x)|}\int d\xi \dot{X}^\mu \frac{1}{\sqrt{|g(X)|}}\delta^{(4)}[x-X(\xi)]\right\} \\ &= \int d\xi \dot{X}^\mu \frac{\partial}{\partial x^\mu}\delta^{(4)}[x-X(\xi)] = -\int d\xi \dot{X}^\mu \frac{\partial}{\partial X^\mu}\delta^{(4)}[x-X(\xi)] \\ &= -\int d\xi \frac{d}{d\xi}\delta^{(4)}[x-X(\xi)] = -\delta^{(4)}[x-X(\xi)]\Big|_{\xi_1}^{\xi_2} = 0, \end{aligned} \tag{12.52}$$

generically, except for the initial and final positions of the particle $X^\mu(\xi_1)$ and $X^\mu(\xi_2)$, which look like a one-particle source and a sink and can be taken to infinity.

Observe that, for the current Eq. (12.49), the interaction term Eq. (12.40) becomes the integral of the 1-form A over the worldline γ:

$$-\frac{q}{c}\int_{\gamma(\xi)} A_\mu \dot{x}^\mu d\xi = -\frac{q}{c}\int_\gamma A. \tag{12.53}$$

This term has to be added to the action of the particle, Eq. (3.255), (3.257), or (3.258), in order to obtain the worldline action of a massive electrically charged particle,

$$\boxed{S[X^\mu(\xi)] = -Mc\int d\xi\,\sqrt{g_{\mu\nu}(X)\dot{X}^\mu\dot{X}^\nu} - \frac{q}{c}\int d\xi A_\mu \dot{X}^\mu,} \tag{12.54}$$

or that of a massless one. That kind of term is known as a *Wess–Zumino* (WZ) term. In this form it is easy to see that, under a gauge transformation, the action changes by a total derivative. The integral of the total derivative vanishes exactly only for special boundary conditions, though.

This action can be used as a source, but it also describes the motion of a charged particle in a gravitational/electromagnetic background. In the special-relativistic limit, taking $\xi = X^0 = ct$, the action takes the standard form

$$S \sim \int dt\left\{-Mc^2 + \tfrac{1}{2}Mv^2 - q\phi + \frac{q}{c}\vec{A}\cdot\vec{v}\right\}. \tag{12.55}$$

If there is a point-like charge q at rest at the origin the only non-vanishing components of F are $F_{\underline{0}r}$ and they should depend only on r because of the spherical symmetry of the problem. Using the above definition of charge and working in general static spherical coordinates Eq. (11.22), we find

$$q = \int_{\mathrm{S}^2_\infty} \frac{\epsilon_{\mu\nu\rho\sigma}}{4\sqrt{|g|}} F^{\rho\sigma} dx^\mu \wedge dx^\nu = \int_{\mathrm{S}^2_\infty} d\Omega^2 r^2 F_{\underline{0}r} = \omega_{(2)} \lim_{r\to\infty} \left(r^2 F_{\underline{0}r}\right), \tag{12.56}$$

where $\omega_{(2)}$ is the volume of the 2-sphere 4π. Then, the electromagnetic field of a point-like charge must behave for large r as follows:

$$E_r = F_{\underline{0}r} \sim +\frac{1}{4\pi}\frac{q}{r^2}, \qquad \phi \sim +\frac{1}{4\pi}\frac{q}{r}. \tag{12.57}$$

(Of course, this result is exact in the absence of gravity, in Minkowski spacetime.) On plugging this result into Eq. (12.55), we find that the electrostatic force between two particles is, in this unit system, $q_1 q_2/(4\pi r^2)$, as we said.

In the units that we are using, M appears multiplied by $G_{\mathrm{N}}^{(4)}$ in the metric (as in the Schwarzschild solution) and q does not. Some simplification is achieved by using the following normalization and units that are standard in this field; we set $c = 1$ and rewrite the Einstein–Maxwell action as follows:

$$\boxed{S_{\mathrm{EM}}[g,A] = \frac{1}{16\pi G_{\mathrm{N}}^{(4)}} \int d^4x \sqrt{|g|}\, \left[R - \tfrac{1}{4}F^2\right].} \tag{12.58}$$

In these units both A_μ and $g_{\mu\nu}$ are dimensionless. The factor $16\pi G_{\mathrm{N}}^{(4)}$ disappears from the equations of motion. Furthermore, if we keep (by definition) the WZ term as in Eq. (12.54) without any additional normalization factor, the electric charge is now

$$\boxed{q = \frac{1}{16\pi G_{\mathrm{N}}^{(4)}} \int_{\mathrm{S}^2_\infty} \star F,} \tag{12.59}$$

and has dimensions of mass (energy). Finally, for a point-like charge we expect, for large r,

$$E_r = F_{0r} \sim \frac{4G_{\mathrm{N}}^{(4)} q}{r^2}, \tag{12.60}$$

which implies that the force between two charges is

$$F_{12} = 4G_{\mathrm{N}}^{(4)} \frac{q_1 q_2}{r_{12}^2}. \tag{12.61}$$

12.2.2 Massive electrodynamics

Before concluding this section it is worth considering which facts would be modified if the vector field were massive. A massive vector field in Minkowski spacetime is described by the Proca Lagrangian Eq. (3.67) and its generalization to curved spacetime is straightforward. The equation of motion is

$$\nabla_\nu F^{\nu\mu} + m^2 A^\mu = 0. \tag{12.62}$$

We immediately see that this equation is completely different from the Bianchi identity Eq. (12.35), which is also valid in the massive case, which implies that massive electrodynamics, apart from gauge invariance, has no electric–magnetic duality. This implies that, in principle, there will be no Dirac magnetic monopoles dual to the electric ones, which explains the results of Ref. [816].

If we take the divergence of this equation, we find the integrability condition

$$\nabla_\nu A^\nu = 0, \tag{12.63}$$

which removes one of the degrees of freedom described by the vector field, leaving only three that correspond to the three possible helicities of a massive spin-1 particle $(-1, 0, +1)$.

The quanta of the Proca field, being massive, will propagate at a speed smaller than 1 (c) and the interaction they mediate will be short ranged. We can see this by finding the static, spherically symmetric solution that describes the field of an electric monopole in this theory in Minkowski spacetime. On substituting the ansatz

$$A_\mu = \delta_{\mu 0} \frac{f(r)}{r} \tag{12.64}$$

into the equation of motion, we obtain the differential equation

$$f'' - m^2 f = 0, \tag{12.65}$$

whose solution is (with the boundary condition $A_\mu \rightarrow 0$ when $r \rightarrow \infty$)

$$A_\mu = Q\delta_{\mu 0} \frac{e^{-mr}}{r}, \tag{12.66}$$

where Q is an integration constant that is somehow related to the "electric charge." However, the lack of gauge invariance suggests that the "electric charge" is not conserved in this system. In fact, it is not easy to define what is meant by electric charge here. It is then useful to consider a slightly more general system with the following classically equivalent action for A_μ and a scalar auxiliary field ϕ:

$$S[A_\mu, \phi] = \int d^4x \sqrt{|g|} \left[-\tfrac{1}{4} F^2 + \tfrac{1}{2} (\partial\phi + mA)^2 \right]. \tag{12.67}$$

This action is invariant under the following *massive gauge transformations*:

$$\delta A_\mu = \partial_\mu \Lambda, \qquad \delta\phi = -m\Lambda. \tag{12.68}$$

Observe that, for consistency, the scalar ϕ has to live in the gauge group manifold: either $\mathbb{R}$ or S^1 (if the gauge group is U(1)). On fixing the gauge $\phi = 0$ we recover the Proca Lagrangian, and any solution of the equations of motion of the original system is also a solution of this one in this gauge.

It is sometimes said that the scalar ϕ is "eaten" by the vector field, which acquires a mass in the process; ϕ is then referred to as a *Stückelberg field* [1162]. Observe that the number of degrees of freedom before and after the gauge fixing are the same. Observe also that this procedure for obtaining a massive vector field is different from the standard spontaneous symmetry-breaking mechanism. There are two main differences: the scalar is real and carries no charge with respect to the vector field and there is no potential for the scalar. (Actually, there is no way to write a gauge-invariant potential with only one real scalar.)

The equations of motion corresponding to the new Lagrangian are

$$\nabla_\nu F^{\nu\mu} + m\big(\nabla^\mu \phi + m A^\mu\big) = 0, \qquad \nabla^2 \phi + m\partial_\mu A^\mu = 0. \tag{12.69}$$

To define a conserved charge, we can either introduce a source j^μ into the first equation or use the conserved Noether current associated with *constant*[11] shifts of ϕ:

$$j^\mu_{\mathrm{N}} = \partial_\nu F^{\nu\mu} + m\big(\partial^\mu \phi + m A^\mu\big). \tag{12.70}$$

The source j^μ is conserved, but only on-shell (upon use of the ϕ equation of motion) and the same applies to the Noether current, which is associated with a global symmetry. In both cases we can define the electric charge in this system by

$$q = \int d^3x \sqrt{|g|}\, j^0_{\mathrm{N}} = \int d^3x \sqrt{|g|}\, \big[\nabla_\nu F^{\nu 0} + m\big(\nabla^0 \phi + m A^0\big)\big]. \tag{12.71}$$

On applying this definition to the electric monopole solution Eq. (12.66) and using

$$\vec{\nabla}^2 \Big(\frac{e^{-mr}}{r}\Big) = -4\pi \delta^{(3)}(\vec{x}) + \frac{m^2 e^{-mr}}{r}, \tag{12.72}$$

we find $q = 4\pi Q$, as we naively expected. It should be stressed, though, that this charge is of a completely different nature from the usual electric charge since it is associated with a *global* symmetry of a different field. In principle, the no-hair conjecture should apply (negatively) to charges of this kind associated with short-range interactions and global (rather than local) symmetries.

Finally, note that neither the original Proca action nor the new one with the Stückelberg field ϕ has any duality *symmetry*. However, the new action can be dualized (i.e. written in *dual variables*), as we will see in Section 12.7.5.

12.3 The electric Reissner–Nordström solution

We are now ready to find BH-type solutions of the equations of motion derived from the Einstein–Maxwell action normalized as in Eqs. (12.58). Since the Maxwell equation is

---

[11] If we use the full gauge invariance of the theory, we recover exactly the same Noether current and Bianchi identity as in the massless case. The definition Eq. (12.59) then gives zero charge because F goes to zero too fast at infinity.

2)*+,)))-) 3 "
9 1 5 , 8 " B 1 " . : 2 " E 1 5 ."
(##' # . 7 B ; 1 BL

10

The Einstein–Langevin Equation

Whereas semiclassical gravity is based on the semiclassical Einstein equation with sources given by the expectation value of the stress-energy tensor of quantum fields, stochastic semiclassical gravity is based on the Einstein–Langevin equation, which has in addition sources due to the noise kernel. The noise kernel is the vacuum expectation value of the stress-energy bitensor which describes the fluctuations of quantum matter fields in curved spacetimes. A new improved criterion for the validity of semiclassical gravity may also be formulated from the viewpoint of this theory. In this chapter we describe the fundamentals of this theory via two approaches: the axiomatic and the functional. The axiomatic approach is useful to see the structure of the theory from the framework of semiclassical gravity, showing the link from the mean value of the stress-energy tensor to the correlation functions. The functional approach uses the Feynman–Vernon influence functional and the Schwinger-Keldysh closed time path (CTP) effective action methods.

In Sections 10.1 and 10.2 we formulate *stochastic semiclassical gravity*, or *stochastic gravity* for short, in an axiomatic way. It is introduced as an extension of semiclassical gravity motivated by the search of self-consistent equations which describe the backreaction of the quantum stress-energy fluctuations on the gravitational field. We then discuss the equivalence between the stochastic correlation functions for the metric perturbations and the quantum correlation functions in the large N expansion, i.e. the quantum theory describing the interaction of the gravitational field with N arbitrary free fields expanded in powers of $1/N$, and illustrate the equivalence with a simple toy model. Based on the stochastic formulation a test is proposed for the validity of the semiclassical theory in Section 10.3. In the second part of the chapter, in Section 10.4, semiclassical gravity as well as stochastic gravity are formulated using a functional approach based on the open quantum system paradigm, discussed in the earlier chapters, in which the system of interest (the gravitational field) interacts with an environment (the matter fields). The functional approach brings familiar quantum field theory techniques into play and provides powerful tools for explicit perturbative calculation. Finally, using this formalism an explicit form of the Einsein–Langevin equation is derived in Section 10.5.

In subsequent chapters we describe some applications of stochastic gravity. In Chapter 11 we consider metric perturbations in a Minkowski spacetime, compute the two-point correlation functions of these perturbations and prove that Minkowski spacetime is a stable solution of semiclassical gravity. In Chapter 12 we discuss the backreaction problem in cosmology when the gravitational field couples to a quantum conformal matter field, and derive the Einstein–Langevin equations describing the metric fluctuations on the cosmological background. In Chapter 13, we discuss structure formation, stochastic gravity is an alternative framework to study the generation of primordial inhomogeneities in inflationary models, which can easily incorporate effects that go beyond the linear perturbations of the inflaton field. In Chapter 14 using the Einstein–Langevin equation we discuss the backreaction of Hawking radiation and the behavior of metric fluctuations for both the quasi-equilibrium condition of a black hole in a box and the fully nonequilibrium condition of an evaporating black hole spacetime.

10.1 Semiclassical Gravity: Axiomatic Approach

Semiclassical gravity describes the interaction of a classical gravitational field with quantum matter fields. This theory can be formally derived as the leading $1/N$ approximation of quantum gravity interacting with N independent and identical free quantum fields which interact with gravity only [194, 684, 685, 686]. By keeping the value of $NG_{\rm N}$ finite, where $G_{\rm N}$ is Newton's gravitational constant, one arrives at a theory in which formally the gravitational field can be treated as a c-number field, i.e., quantized at tree level, while matter fields are fully quantized. The semiclassical theory may be summarized as follows.

Let $(\mathcal{M}, g_{ab})$ be a globally hyperbolic four-dimensional spacetime manifold $\mathcal{M}$ with metric g_{ab} and consider a real scalar quantum field ϕ of mass m propagating on that manifold; we just assume a scalar field for simplicity. The classical action $S_{\rm m}$ for this matter field is given by the functional

$$S_{\rm m}[g,\phi] = -\frac{1}{2}\int d^4x\sqrt{-g}\left[g^{ab}\nabla_a\phi\nabla_b\phi + \left(m^2+\xi R\right)\phi^2\right], \qquad (10.1)$$

where ∇_a is the covariant derivative associated to the metric g_{ab}, ξ is a dimensionless coupling parameter between the field and the scalar curvature of the underlying spacetime R, and $g = \det g_{ab}$.

The field may be quantized in the manifold using the standard canonical quantization formalism [47, 48, 49]. The field operator in the Heisenberg picture $\hat{\phi}$ is an operator-valued distribution solution of the Klein–Gordon equation,

$$(\Box - m^2 - \xi R)\hat{\phi} = 0, \qquad (10.2)$$

where $\Box = \nabla_a \nabla^a$. We may write the field operator as $\hat{\phi}[g;x)$ to indicate that it is a functional of the metric g_{ab} and a function of the spacetime point x. This notation will be used also for other operators and tensors.

The classical stress-energy tensor is obtained by functional derivation of this action in the usual way $T^{ab}(x) = (2/\sqrt{-g})\delta S_{\rm m}/\delta g_{ab}$, leading to

$$T^{ab}[g,\phi] = \nabla^a\phi\nabla^b\phi - \frac{1}{2}g^{ab}\left(\nabla^c\phi\nabla_c\phi + m^2\phi^2\right) + \xi\left(g^{ab}\Box - \nabla^a\nabla^b + G^{ab}\right)\phi^2, \tag{10.3}$$

where G_{ab} is the Einstein tensor. With the notation $T^{ab}[g,\phi]$ we explicitly indicate that the stress-energy tensor is a functional of the metric g_{ab} and the field ϕ. This is the same as the stress-energy tensor (2.59) where $(1-\xi)\xi_n$ has been replaced by ξ, so that now $\xi = 0$ and $\xi = \xi_n$ correspond to minimal and conformal couplings, respectively.

The next step is to define a stress-energy tensor operator $\hat{T}^{ab}[g;x)$. Naively one would replace the classical field $\phi[g;x)$ in the above functional by the quantum operator $\hat{\phi}[g;x)$, but this procedure involves taking the product of two distributions at the same spacetime point. This is ill-defined and a regularization procedure is needed. There are several regularization methods which one may use, one is the point-splitting or point-separation regularization method introduced by Christensen [100, 101] in which one takes a point y in a neighborhood of the point x and then uses as the regulator the vector tangent at the point x of the geodesic joining x and y; this method is discussed in a previous chapter.

Another well known method is dimensional regularization in which one works in n dimensions, where n is not necessarily an integer, and then uses as the regulator the parameter $\epsilon = n-4$; this method is implicitly used in this Section. The regularized stress-energy operator using the Weyl ordering prescription, i.e., symmetrical ordering, can be written as

$$\hat{T}^{ab}[g] = \frac{1}{2}\{\nabla^a\hat{\phi}[g]\,,\,\nabla^b\hat{\phi}[g]\} + \mathcal{D}^{ab}[g]\,\hat{\phi}^2[g], \tag{10.4}$$

where $\mathcal{D}^{ab}[g]$ is the differential operator:

$$\mathcal{D}^{ab} \equiv (\xi - 1/4)\,g^{ab}\Box + \xi\left(R^{ab} - \nabla^a\nabla^b\right). \tag{10.5}$$

Note that if dimensional regularization is used, the field operator $\hat{\phi}[g;x)$ propagates in a n-dimensional spacetime. Once the regularization prescription has been introduced a regularized and renormalized stress-energy operator $\hat{T}^{\rm R}_{ab}[g;x)$ may be defined as

$$\hat{T}^{\rm R}_{ab}[g;x) = \hat{T}_{ab}[g;x) + F^C_{ab}[g;x)\hat{I}, \tag{10.6}$$

which differs from the regularized $\hat{T}_{ab}[g;x)$ by the identity operator times some tensor counterterms $F^C_{ab}[g;x)$, which depend on the regulator and are local functionals of the metric. The field states can be chosen in such a way that for any pair of physically acceptable states, i.e., Hadamard states in the sense of [49], $|\psi\rangle$ and $|\varphi\rangle$, the matrix element $\langle\psi|T^{\mathrm{R}}_{ab}|\varphi\rangle$, defined as the limit when the regulator takes the physical value, is finite and satisfies Wald's axioms [102, 48]. These counterterms can be extracted from the singular part of a Schwinger–DeWitt series [48, 100, 101, 91]. The choice of these counterterms is not unique but this ambiguity can be absorbed into the renormalized coupling constants which appear in the equations of motion for the gravitational field.

10.1.1 Semiclassical Einstein Equation

The *semiclassical Einstein equation* for the metric g_{ab} can then be written as

$$G_{ab}[g] + \Lambda g_{ab} - 2(\alpha_d A_{ab} + \beta_d B_{ab})[g] = 8\pi G_{\mathrm{N}} \langle \hat{T}^{\mathrm{R}}_{ab}[g]\rangle, \qquad (10.7)$$

where $\langle \hat{T}^{\mathrm{R}}_{ab}[g]\rangle$ is the expectation value of the operator $\hat{T}^{\mathrm{R}}_{ab}[g,x)$ after the regulator takes the physical value in some physically acceptable state of the field on $(\mathcal{M}, g_{ab})$. Note that both the stress-energy tensor and the quantum state are functionals of the metric, hence the notation.

The parameters G_{N}, Λ, α_d and β_d are the renormalized coupling constants, respectively, the gravitational constant, the cosmological constant and two parameters which are zero in the classical Einstein equation and have dimensions of square length (we use units in which $\hbar = c = 1$). In the effective field theory description of general relativity [192] these two parameters are naturally of the order of the Planck length square $\ell_{\mathrm{P}}^2 = G_{\mathrm{N}}$ and in later Chapters will be convenient to use instead dimensionless parameters a_1 and a_2 defined by $\alpha_d \equiv 8\pi\ell_{\mathrm{P}}^2 a_1$ and $\beta_d \equiv 8\pi\ell_{\mathrm{P}}^2 a_2$. The renormalized four constants of the semiclassical equation must be understood as the result of "dressing" the bare constants which appear in the classical action before renormalization. The values of these constants must be determined by experiment.

The left-hand side of Eq. (10.7) may be derived from the gravitational action

$$S_{\mathrm{g}}[g] = \frac{1}{8\pi G_{\mathrm{N}}} \int d^4x \sqrt{-g} \left[\frac{1}{2} R - \Lambda + \alpha_d C_{abcd} C^{abcd} + \beta_d R^2 \right], \qquad (10.8)$$

where C_{abcd} is the Weyl tensor. The parameters α_d and β_d are related to the parameters a, b, c introduced in the action (2.67). By writing the Ricci and the Riemann squares in terms of Ricci scalar and Weyl squares and the Euler density G, given in (2.61), and taking into account that by the Gauss–Bonnet theorem the Euler density does not contribute to the field equations one can see that these two actions are equivalent with $\alpha_d = b/4 + c$ and $\beta_d = a/2 + (b + c)/6$.

The tensors A_{ab} and B_{ab} come from the functional derivatives with respect to the metric of the terms quadratic in the curvature in Eq. (10.8), and are explicitly given by

$$
\begin{aligned}
A^{ab} &= \frac{1}{\sqrt{-g}} \frac{\delta}{\delta g_{ab}} \int d^4 \sqrt{-g} C_{cdef} C^{cdef} \\
&= \frac{1}{2} g^{ab} C_{cdef} C^{cdef} - 2R^{acde} R^{b}_{cde} + 4R^{ac} R^{b}_{c} - \frac{2}{3} R R^{ab} \\
&\quad - 2\Box R^{ab} + \frac{2}{3} \nabla^a \nabla^b R + \frac{1}{3} g^{ab} \Box R, \qquad (10.9) \\
B^{ab} &= \frac{1}{\sqrt{-g}} \frac{\delta}{\delta g_{ab}} \int d^4 \sqrt{-g} R^2 \\
&= \frac{1}{2} g^{ab} R^2 - 2R R^{ab} + 2\nabla^a \nabla^b R - 2g^{ab} \Box R, \qquad (10.10)
\end{aligned}
$$

where R_{abcd} and R_{ab} are the Riemann and Ricci tensors, respectively. These two tensors are, like the Einstein and metric tensors, symmetric and divergenceless, $\nabla^a A_{ab} = 0 = \nabla^a B_{ab}$, and they have been introduced previously in (2.71) and (2.69) with different names: $A_{ab} = {}^{(C)}H_{ab}$ and $B_{ab} = {}^{(1)}H_{ab}$.

A solution of semiclassical gravity consists of a spacetime $(\mathcal{M}, g_{ab})$, a quantum field operator $\hat{\phi}[g]$ which satisfies the evolution Eqs. (10.2), and a physically acceptable state $|\psi[g]\rangle$ for this field, such that Eq. (10.7) is satisfied when the expectation value of the renormalized stress-energy operator is evaluated in this state.

For a free quantum field this theory is robust in the sense that it is self-consistent and fairly well understood. As long as the gravitational field is assumed to be described by a classical metric, the above semiclassical Einstein equations seems to be the only plausible dynamical equation for this metric: the metric couples to matter fields via the stress-energy tensor and for a given quantum state the only physically observable c-number stress-energy tensor that one can construct is the above renormalized expectation value. However, lacking a full quantum theory of gravity the scope and limits of the theory are not so well understood. It is assumed that the semiclassical theory should break down at Planck scales, which is when simple order of magnitude estimates suggest that the quantum effects of gravity should not be ignored because the energy of a quantum fluctuation in a Planck size region, as determined by the Heisenberg uncertainty principle, is comparable to the gravitational energy of that fluctuation.

The theory is expected to break down when the fluctuations of the stress-energy operator are large [195]. A criterion based on the ratio of the fluctuations to the mean was proposed by Kuo and Ford [196] (see also work via zeta-function methods in [197, 205]). This proposal was questioned by Hu and Phillips [198, 199, 200] because it does not contain a scale at which the theory is probed or how accurately the theory can be resolved. They suggested the use of a smearing scale or point-separation distance, for integrating over the bitensor quantities, equivalent to a stipulation of the resolution level of measurements; see also the response by Ford and Wu [206, 207]. A different criterion was suggested by Anderson et al. [208, 209] based on linear response theory.

Later, Hu et al. [190, 214] proposed a criterion for the validity of semiclassical gravity which is based on the stability of the solutions of the semiclassical Einstein equations with respect to quantum metric fluctuations. The two-point correlations for the metric perturbations can be described in the framework of stochastic gravity, which is closely related to the quantum theory of gravity interacting with N matter fields, to the next to leading order in a $1/N$ expansion. We describe these developments in the following sections.

10.2 Stochastic Gravity: Axiomatic Approach

The purpose of stochastic gravity is to extend the semiclassical theory to account for the stress-energy tensor fluctuations in a self-consistent way. A physical observable that describes these fluctuations to lowest order is the *noise kernel* bitensor, which is defined through the two-point correlation of the stress-energy operator as

$$N_{abc'd'}[g;x,y] = \frac{1}{2}\langle\{\hat{t}_{ab}[g;x], \hat{t}_{c'd'}[g;y]\}\rangle, \tag{10.11}$$

where the curly brackets mean anticommutator,

$$\hat{t}_{ab}[g;x] \equiv \hat{T}_{ab}[g;x] - \langle\hat{T}_{ab}[g;x]\rangle, \tag{10.12}$$

and we use unprimed and primed indices to emphasize that this bitensor is a tensor with respect to the first two indices at the point x and a tensor with respect to the last two indices at the point y.

The noise kernel is defined in terms of the unrenormalized stress-energy tensor operator $\hat{T}_{ab}[g;x)$ on a given background metric g_{ab}, thus a regulator is implicitly assumed on the right-hand side of Eq. (10.11). However, for a linear quantum field the above kernel – the expectation function of a bitensor – is free of ultraviolet divergences because the regularized $T_{ab}[g;x)$ differs from the renormalized $T^{\mathrm{R}}_{ab}[g;x)$ by the identity operator times some tensor counterterms, see Eq. (10.6), so that in the subtraction (10.12) the counterterms cancel. Consequently the ultraviolet behavior of $\langle\hat{T}_{ab}(x)\hat{T}_{c'd'}(y)\rangle$ is the same as that of $\langle\hat{T}_{ab}(x)\rangle\langle\hat{T}_{c'd'}(y)\rangle$, and $\hat{T}_{ab}$ can be replaced by the renormalized operator $\hat{T}^{\mathrm{R}}_{ab}$ in Eq. (10.11). The noise kernel should be thought of as a distribution function; the limit of coincidence points has meaning only in the sense of distributions.

The bitensor $N_{abc'd'}[g;x,y)$, or $N_{abc'd'}(x,y)$ for short, is real and positive semi-definite, as a consequence of $\hat{T}^{R}_{ab}$ being self-adjoint. A simple proof of this property can be given as follows. Let $|\psi\rangle$ be a given quantum state and let $\hat{Q}$ be a self-adjoint operator, $\hat{Q}^{\dagger} = \hat{Q}$, then one can write $\langle\psi|\hat{Q}\hat{Q}|\psi\rangle = \langle\psi|\hat{Q}^{\dagger}\hat{Q}|\psi\rangle = |\hat{Q}|\psi\rangle|^2 \geq 0$. Now let $\hat{t}(x)$ be a self-adjoint operator (this is shorthand for $\hat{t}_{ab}$), then if we define $\hat{Q} = \int dx f(x)\hat{t}(x)$ for an arbitrary well defined function $f(x)$, the previous inequality can be written as follows $\int dxdy f(x)\langle\psi|\hat{t}(x)\hat{t}(y)|\psi\rangle f(y) \geq 0$,

11

Metric Fluctuations in Minkowski Spacetime

Although the Minkowski vacuum is an eigenstate of the total four-momentum operator of a field in Minkowski spacetime, it is not an eigenstate of the stress-energy operator. Hence, even for those solutions of semiclassical gravity such as the Minkowski metric, for which the expectation value of the stress-energy operator can always be chosen to be zero, the fluctuations of this operator are non-vanishing. This fact leads to consider the stochastic metric perturbations induced by these fluctuations.

Here we derive the Einstein–Langevin equation for the metric perturbations in a Minkowski background. We solve this equation for the linearized Einstein tensor and compute the associated two-point correlation functions, as well as the two-point correlation functions for the metric perturbations. Even though, in this case, we expect to have negligibly small values for these correlation functions for points separated by lengths larger than the Planck length, there are several reasons why it is worth carrying out this calculation.

On the one hand, these are the first backreaction solutions of the full Einstein–Langevin equation. There are analogous solutions to a "reduced" version of this equation inspired in a "mini-superspace" model in [218, 219].

On the other hand, the results of this calculation, which confirm our expectations that gravitational fluctuations are negligible at length scales larger than the Planck length, but also predict that the fluctuations are strongly suppressed on small scales, can be considered a first test of stochastic semiclassical gravity. These results also reveal an important connection between stochastic gravity and the large N expansion of quantum gravity. In addition, they are used in Section 11.5 to study the stability of the Minkowski metric as a solution of semiclassical gravity, which constitutes an application of the validity criterion introduced in Section 10.3. This calculation requires also a discussion of the problems posed by the so called *runaway* solutions, which arise in the backreaction equations of semiclassical and stochastic gravity, and some of the methods to deal with them. As a result we conclude that Minkowski spacetime is a stable and valid solution of semiclassical gravity.

11.1 Perturbations around Minkowski Spacetime

The Minkowski metric η_{ab}, in a manifold $\mathcal{M}$ which is topologically $\mathbb{R}^4$, and the usual Minkowski vacuum, denoted as $|0\rangle$, is the simplest solution to the semiclassical Einstein equation (10.7), the so-called trivial solution of semiclassical gravity [128]. It constitutes the ground state of semiclassical gravity. In fact, we can always choose a renormalization scheme in which the renormalized expectation value of the stress-energy tensor vanishes: $\langle 0|\,\hat{T}_{\mathrm{R}}^{ab}\,[\eta]|0\rangle = 0$. Thus, Minkowski spacetime $(\mathbb{R}^4, \eta_{ab})$ and the vacuum state $|0\rangle$ are a solution to the semiclassical Einstein equation with renormalized cosmological constant $\Lambda = 0$.

The fact that the vacuum expectation value of the renormalized stress-energy operator in Minkowski spacetime should vanish was originally proposed by Wald [102] and it may be understood as a renormalization convention [48, 51]. Note that other possible solutions of semiclassical gravity with zero vacuum expectation value of the stress-energy tensor are the exact gravitational plane waves, since these are known to be vacuum solutions of Einstein equations which induce neither particle creation nor vacuum polarization [719, 720, 721].

As we have already mentioned the vacuum $|0\rangle$ is an eigenstate of the total four-momentum operator in Minkowski spacetime, but not an eigenstate of $\hat{T}^{\mathrm{R}}_{ab}[\eta]$. Hence, even in the Minkowski background, there are quantum fluctuations in the stress-energy tensor and, as a result, the noise kernel does not vanish. This fact leads to consider the stochastic corrections to this class of trivial solutions of semiclassical gravity. Since, in this case, the Wightman and Feynman functions (10.55), their values in the two-point coincidence limit, and the products of derivatives of two of such functions appearing in expressions (10.56) and (10.57) are known in dimensional regularization, we can compute the Einstein–Langevin equation using the methods outlined in Chapter 10.

To perform explicit calculations it is convenient to work in a global inertial coordinate system $\{x^\mu\}$ and in the associated basis, in which the components of the flat metric are simply $\eta_{\mu\nu} = \mathrm{diag}(-1, 1, \ldots, 1)$. In Minkowski spacetime, the components of the classical stress-energy tensor (10.3) reduce to

$$T^{\mu\nu}[\eta,\phi] = \partial^\mu\phi\partial^\nu\phi - \frac{1}{2}\,\eta^{\mu\nu}\partial^\rho\phi\partial_\rho\phi - \frac{1}{2}\,\eta^{\mu\nu}m^2\phi^2 + \xi\left(\eta^{\mu\nu}\Box - \partial^\mu\partial^\nu\right)\phi^2, \quad (11.1)$$

where $\Box \equiv \partial_\mu\partial^\mu$, and the formal expression for the components of the corresponding "operator" in dimensional regularization, see Eq. (10.4), is

$$\hat{T}_n^{\mu\nu}[\eta] = \frac{1}{2}\{\partial^\mu\hat{\phi}_n, \partial^\nu\hat{\phi}_n\} + \mathcal{D}^{\mu\nu}\hat{\phi}_n^2, \quad (11.2)$$

where $\mathcal{D}^{\mu\nu}$ is the differential operator (10.5), with $g_{\mu\nu} = \eta_{\mu\nu}, R_{\mu\nu} = 0$, and $\nabla_\mu = \partial_\mu$. The field $\hat{\phi}_n(x)$ is the field operator in the Heisenberg picture in a n-dimensional Minkowski spacetime, which satisfies the Klein–Gordon equation (10.2). We use here a stress-energy tensor which differs from the canonical one, which corresponds to $\xi = 0$, both tensors, however, define the same total momentum.

The Wightman and Feynman functions (10.55) when $g_{\mu\nu} = \eta_{\mu\nu}$, are well known:

$$G_n^+(x,y) = \Delta_n^+(x-y), \quad G_{F_n}(x,y) = \Delta_{F_n}(x-y), \tag{11.3}$$

with

$$\begin{aligned}\Delta_n^+(x) &= -2\pi i \int \frac{d^n k}{(2\pi)^n}\, e^{ik\cdot x}\, \delta(k^2+m^2)\, \theta(k^0),\\ \Delta_{F_n}(x) &= -\int \frac{d^n k}{(2\pi)^n}\, \frac{e^{ik\cdot x}}{k^2+m^2-i\epsilon}, \epsilon \to 0^+,\end{aligned} \tag{11.4}$$

where $k^2 \equiv \eta_{\mu\nu}k^\mu k^\nu$ and $k\cdot x \equiv \eta_{\mu\nu}k^\mu x^\nu$. Note that the derivatives of these functions satisfy $\partial_\mu^x \Delta_n^+(x-y) = \partial_\mu \Delta_n^+(x-y)$ and $\partial_\mu^y \Delta_n^+(x-y) = -\partial_\mu \Delta_n^+(x-y)$, and similarly for the Feynman propagator $\Delta_{F_n}(x-y)$.

To write down the semiclassical Einstein equation (10.7) in n-dimensions for this case, we need to compute the vacuum expectation value of the stress-energy operator components (11.2). Since, from (11.3), we have that $\langle 0|\hat{\phi}_n^2(x)|0\rangle = i\Delta_{F_n}(0) = i\Delta_n^+(0)$, which is a constant (independent of x), we have simply

$$\langle 0|\hat{T}_n^{\mu\nu}[\eta]|0\rangle = -i\int \frac{d^n k}{(2\pi)^n}\, \frac{k^\mu k^\nu}{k^2+m^2-i\epsilon} = \frac{\eta^{\mu\nu}}{2}\left(\frac{m^2}{4\pi}\right)^{n/2} \Gamma\left(-\frac{n}{2}\right), \tag{11.5}$$

where the integrals in dimensional regularization have been computed in the standard way and where $\Gamma(z)$ is the Euler's gamma function. The semiclassical Einstein equation (10.7) in n-dimensions before renormalization reduces now to

$$\frac{\Lambda_B}{8\pi G_B}\, \eta^{\mu\nu} = \mu^{-(n-4)} \langle 0|\hat{T}_n^{\mu\nu}[\eta]|0\rangle. \tag{11.6}$$

This equation, thus, simply sets the value of the bare coupling constant Λ_B/G_B. Note, from (11.5), that in order to have $\langle 0|\,\hat{T}_{\mathrm{R}}^{\mu\nu}\,|0\rangle[\eta] = 0$, the renormalized and regularized stress-energy tensor "operator" for a scalar field in Minkowski spacetime, see Eq. (10.6), has to be defined as

$$\hat{T}_{\mathrm{R}}^{\mu\nu}[\eta] = \mu^{-(n-4)}\, \hat{T}_n^{\mu\nu}[\eta] - \frac{\eta^{\mu\nu}}{2}\, \frac{m^4}{(4\pi)^2} \left(\frac{m^2}{4\pi\mu^2}\right)^{\frac{n-4}{2}} \Gamma\left(-\frac{n}{2}\right), \tag{11.7}$$

which corresponds to a renormalization of the cosmological constant

$$\frac{\Lambda_B}{G_B} = \frac{\Lambda}{G_{\mathrm{N}}} - \frac{2}{\pi}\, \frac{m^4}{n(n-2)}\, \kappa_n + \mathcal{O}(n-4), \tag{11.8}$$

where

$$\kappa_n \equiv \frac{1}{(n-4)} \left(\frac{e^{\gamma} m^2}{4\pi\mu^2} \right)^{\frac{n-4}{2}} = \frac{1}{n-4} + \frac{1}{2} \ln \left(\frac{e^{\gamma} m^2}{4\pi\mu^2} \right) + \mathcal{O}(n-4), \tag{11.9}$$

being γ the Euler's constant. In the case of a massless scalar field, $m^2 = 0$, one simply has $\Lambda_B / G_B = \Lambda / G_{\mathrm{N}}$. Introducing this renormalized coupling constant into Eq. (11.6), we can take the limit $n \to 4$. We find that, for $(\mathbb{R}^4, \eta_{ab}, |0\rangle)$ to satisfy the semiclassical Einstein equation, we must take $\Lambda = 0$.

We can now write down the Einstein–Langevin equations for the components $h_{\mu\nu}$ of the stochastic metric perturbation in dimensional regularization. In our case, using $\langle 0|\hat{\phi}_n^2(x)|0\rangle = i\Delta_{F_n}(0)$ and the explicit expression of Eq. (10.52) we obtain

$$\begin{aligned} \frac{1}{8\pi G_B} & \left[G^{(1)\mu\nu} + \Lambda_B \left(h^{\mu\nu} - \frac{1}{2}\, \eta^{\mu\nu} h \right) \right] (x) \\ & - \frac{4}{3}\, \alpha_B D^{(1)\mu\nu}(x) - 2\beta_B B^{(1)\mu\nu}(x) - \xi G^{(1)\mu\nu}(x) \mu^{-(n-4)} i\Delta_{F_n}(0) \\ & + \frac{1}{2} \int d^n y \mu^{-(n-4)} H_n^{\mu\nu\alpha'\beta'}(x,y) h_{\alpha'\beta'}(y) = \xi^{\mu\nu}(x). \end{aligned} \tag{11.10}$$

The indices in $h_{\mu\nu}$ are raised with the Minkowski metric, $h \equiv h^{\rho}_{\rho}$, and here a superscript (1) denotes the components of a tensor linearized around the flat metric. Note that in n-dimensions the two-point correlation functions for the field $\xi^{\mu\nu}$ is written as

$$\langle \xi^{\mu\nu}(x) \xi^{\alpha'\beta'}(y) \rangle_s = \mu^{-2(n-4)} N_n^{\mu\nu\alpha'\beta'}(x,y), \tag{11.11}$$

Explicit expressions for $D^{(1)\mu\nu}$ and $B^{(1)\mu\nu}$ are given by

$$D^{(1)\mu\nu}(x) = \frac{1}{2}\, \underset{x}{\mathcal{F}}^{\mu\nu\alpha'\beta'}\, h_{\alpha'\beta'}(x), \qquad B^{(1)\mu\nu}(x) = 2 \underset{x}{\mathcal{F}}^{\mu\nu} \underset{x}{\mathcal{F}}^{\alpha\beta} h_{\alpha\beta}(x), \tag{11.12}$$

where we introduced the differential operators $\underset{x}{\mathcal{F}}^{\mu\nu} \equiv \eta^{\mu\nu} \Box_x - \partial^{\mu}_{x} \partial^{\nu}_{x}$ and $\underset{x}{\mathcal{F}}^{\mu\nu\alpha'\beta'} \equiv 3 \underset{x}{\mathcal{F}}^{\mu(\alpha'} \underset{x}{\mathcal{F}}^{\beta')\nu} - \underset{x}{\mathcal{F}}^{\mu\nu} \underset{x}{\mathcal{F}}^{\alpha'\beta'}$. Note that to distinguish tensor indices in x from tensor indices in y is unnecessary in the Minkowski background with standard coordinates. However, we will follow the convention of keeping the last two indices primed in bitensors in order to have an homogeneous notation with previous chapters.

11.2 The Kernels in the Minkowski Background

Since the two kernels (10.54) are free of ultraviolet divergences in the limit $n \to 4$, we can deal directly with the $F^{\mu\nu\alpha'\beta'}(x-y) \equiv \lim_{n\to 4} \mu^{-2(n-4)} F_n^{\mu\nu\alpha'\beta'}$ in Eq. (10.53). Also $N^{\mu\nu\alpha'\beta'}(x,y) = \operatorname{Re} F^{\mu\nu\alpha'\beta'}(x-y)$ and $H_{\mathrm{A}}^{\mu\nu\alpha'\beta'}(x,y) = \operatorname{Im} F^{\mu\nu\alpha'\beta'}(x-y)$ are actually the components of the "physical" noise and dissipation kernels that will appear in the Einstein–Langevin equations once the renormalization procedure has been carried out. The bitensor $F^{\mu\nu\alpha'\beta'}$ can be expressed in terms of the positive Wightman function in four spacetime dimensions, according to (10.56). The different terms in this kernel can be easily computed using the integrals

$$I(p) \equiv \int \frac{d^4k}{(2\pi)^4}\, \delta(k^2+m^2)\, \theta(-k^0)\, \delta[(k-p)^2+m^2]\, \theta(k^0-p^0), \tag{11.13}$$

and $I^{\mu_1\cdots\mu_r}(p)$ which are defined as the previous one by inserting the momenta $k^{\mu_1}\dots k^{\mu_r}$ with $r=1,2,3,4$ in the integrand. All these integrals can be expressed in terms of $I(p)$; see [189] for the explicit expressions.

It is convenient to separate $I(p)$ in its even and odd parts with respect to the variables p^μ as

$$I(p) = I_{\rm S}(p) + I_{\rm A}(p), \tag{11.14}$$

where $I_{\rm S}(-p) = I_{\rm S}(p)$ and $I_{\rm A}(-p) = -I_{\rm A}(p)$. These two functions are explicitly given by

$$\begin{aligned} I_{\rm S}(p) &= \frac{1}{8\,(2\pi)^3}\, \theta(-p^2-4m^2)\, \sqrt{1+4\,\frac{m^2}{p^2}}, \\ I_{\rm A}(p) &= \frac{-1}{8\,(2\pi)^3}\, \mathrm{sign}(p^0)\theta(-p^2-4m^2)\, \sqrt{1+4\,\frac{m^2}{p^2}}. \end{aligned} \tag{11.15}$$

After some manipulations, we find

$$\begin{aligned} F^{\mu\nu\alpha'\beta'}(x) &= \frac{\pi^2}{45}\, \mathcal{F}_x^{\mu\nu\alpha'\beta'} \int \frac{d^4p}{(2\pi)^4}\, e^{-ip\cdot x} \left(1+4\frac{m^2}{p^2}\right)^2 I(p) \\ &\quad + \frac{8\pi^2}{9} \mathcal{F}_x^{\mu\nu} \mathcal{F}_x^{\alpha'\beta'} \int \frac{d^4p}{(2\pi)^4}\, e^{-ip\cdot x} \left(3\Delta\xi + \frac{m^2}{p^2}\right)^2 I(p), \end{aligned} \tag{11.16}$$

where $\Delta\xi \equiv \xi - 1/6$. The real and imaginary parts of the last expression, which yield the noise and dissipation kernels, are easily recognized as the terms containing $I_{\rm S}(p)$ and $I_{\rm A}(p)$, respectively.

To write them explicitly, it is useful to introduce the new kernels

$$\begin{aligned} N_{\rm A}(x;m^2) &\equiv \frac{1}{480\pi} \int \frac{d^4p}{(2\pi)^4}\, e^{ip\cdot x}\, \theta(-p^2-4m^2)\, \sqrt{1+4\,\frac{m^2}{p^2}} \left(1+4\,\frac{m^2}{p^2}\right)^2, \\ N_{\rm B}(x;m^2,\Delta\xi) &\equiv \frac{1}{72\pi} \int \frac{d^4p}{(2\pi)^4}\, e^{ip\cdot x}\, \theta(-p^2-4m^2) \\ &\qquad \times \sqrt{1+4\,\frac{m^2}{p^2}} \left(3\Delta\xi + \frac{m^2}{p^2}\right)^2, \end{aligned}$$

$$
\begin{aligned}
D_{\mathrm{A}}(x;m^2) &\equiv \frac{-i}{480\pi}\int \frac{d^4p}{(2\pi)^4}\, e^{ip\cdot x}\,\mathrm{sign}(p^0)\theta(-p^2-4m^2)\\
&\quad\times\sqrt{1+4\,\frac{m^2}{p^2}}\left(1+4\,\frac{m^2}{p^2}\right)^2,\\
D_{\mathrm{B}}(x;m^2,\Delta\xi) &\equiv \frac{-i}{72\pi}\int \frac{d^4p}{(2\pi)^4}\, e^{ip\cdot x}\,\mathrm{sign}(p^0)\theta(-p^2-4m^2)\\
&\quad\times\sqrt{1+4\,\frac{m^2}{p^2}}\left(3\Delta\xi+\frac{m^2}{p^2}\right)^2,
\end{aligned}
\tag{11.17}
$$

and we finally obtain

$$
\begin{aligned}
N^{\mu\nu\alpha'\beta'}(x,y) &= \frac{1}{6}\mathcal{F}_x^{\mu\nu\alpha'\beta'}N_{\mathrm{A}}(x-y;m^2)+\mathcal{F}_x^{\mu\nu}\mathcal{F}_x^{\alpha'\beta'}N_{\mathrm{B}}(x-y;m^2,\Delta\xi),\\
H_{\mathrm{A}}^{\mu\nu\alpha'\beta'}(x,y) &= \frac{1}{6}\mathcal{F}_x^{\mu\nu\alpha'\beta'}D_{\mathrm{A}}(x-y;m^2)+\mathcal{F}_x^{\mu\nu}\mathcal{F}_x^{\alpha'\beta'}D_{\mathrm{B}}(x-y;m^2,\Delta\xi).
\end{aligned}
\tag{11.18}
$$

Notice that the noise and dissipation kernels defined in (11.17) are actually real because, for the noise kernels, only the $\cos(p\cdot x)$ terms of the exponentials $e^{ip\cdot x}$ contribute to the integrals, and, for the dissipation kernels, the only contribution of such exponentials comes from the $i\sin(p\cdot x)$ terms.

The evaluation of the kernel $H_{\mathrm{S}_n}^{\mu\nu\alpha'\beta'}(x,y)$ is a more involved task. Since this kernel contains divergences in the limit $n\to 4$, we use dimensional regularization. Using Eq. (10.57), this kernel can be written in terms of the Feynman propagator (11.4) as

$$
\mu^{-(n-4)}H_{\mathrm{S}_n}^{\mu\nu\alpha'\beta'}(x,y)=\mathrm{Im}\,K^{\mu\nu\alpha'\beta'}(x-y),
\tag{11.19}
$$

where

$$
\begin{aligned}
K^{\mu\nu\alpha'\beta'}(x) \equiv -\mu^{-(n-4)}\Big\{&2\partial^{\mu}\partial^{(\alpha'}\Delta_{F_n}(x)\,\partial^{\beta')}\partial^{\nu}\Delta_{F_n}(x)\\
&+2\mathcal{D}^{\mu\nu}\left(\partial^{\alpha'}\Delta_{F_n}(x)\partial^{\beta'}\Delta_{F_n}(x)\right)+2\mathcal{D}^{\alpha'\beta'}\left(\partial^{\mu}\Delta_{F_n}(x)\,\partial^{\nu}\Delta_{F_n}(x)\right)\\
&+2\mathcal{D}^{\mu\nu}\mathcal{D}^{\alpha'\beta'}\left(\Delta_{F_n}^2(x)\right)+\Big[\eta^{\mu\nu}\partial^{(\alpha'}\Delta_{F_n}(x)\partial^{\beta')}+\eta^{\alpha'\beta'}\partial^{(\mu}\Delta_{F_n}(x)\partial^{\nu)}\\
&+\Delta_{F_n}(0)\left(\eta^{\mu\nu}\mathcal{D}^{\alpha'\beta'}+\eta^{\alpha'\beta'}\mathcal{D}^{\mu\nu}\right)\\
&+\frac{1}{4}\eta^{\mu\nu}\eta^{\alpha'\beta'}\left(\Delta_{F_n}(x)\Box-m^2\Delta_{F_n}(0)\right)\Big]\delta^n(x)\Big\}.
\end{aligned}
\tag{11.20}
$$

Let us define the integrals

$$
J_n(p)\equiv\mu^{-(n-4)}\int\frac{d^nk}{(2\pi)^n}\,\frac{1}{(k^2+m^2-i\epsilon)\left[(k-p)^2+m^2-i\epsilon\right]},
\tag{11.21}
$$

and $J_n^{\mu_1\cdots\mu_r}(p)$ obtained by inserting the momenta $k^{\mu_1}\dots k^{\mu_r}$ into the previous integral, together with

$$I_{0_n} \equiv \mu^{-(n-4)} \int \frac{d^n k}{(2\pi)^n} \, \frac{1}{(k^2+m^2-i\epsilon)}, \tag{11.22}$$

and $I_{0_n}^{\mu_1\dots\mu_r}$ which are also obtained by inserting momenta in the integrand. Then, the different terms in Eq. (11.20) can be computed. These integrals are explicitly given in [189]. It is found that $I_{0_n}^{\mu}=0$ and the remaining integrals can be written in terms of I_{0_n} and $J_n(p)$. It is useful to introduce the projector $P^{\mu\nu}$ orthogonal to p^μ and the tensor $P^{\mu\nu\alpha'\beta'}$ as

$$p^2 P^{\mu\nu} \equiv \eta^{\mu\nu} p^2 - p^\mu p^\nu, \qquad P^{\mu\nu\alpha'\beta'} \equiv 3P^{\mu(\alpha'}P^{\beta')\nu} - P^{\mu\nu}P^{\alpha'\beta'}, \tag{11.23}$$

then the action of the operator $\mathcal{F}_x^{\mu\nu}$ is simply written as $\mathcal{F}_x^{\mu\nu} \int d^n p \, e^{ip\cdot x} f(p) = -\int d^n p \, e^{ip\cdot x} f(p) \, p^2 P^{\mu\nu}$ where $f(p)$ is an arbitrary function of p^μ.

Finally after a rather long but straightforward calculation, and after expanding around $n=4$, we get,

$$\begin{aligned} K^{\mu\nu\alpha'\beta'}(x) = \frac{i}{(4\pi)^2}\Bigg\{ & \kappa_n \left[\frac{1}{90}\mathcal{F}_x^{\mu\nu\alpha'\beta'}\delta^n(x) + 4\Delta\xi^2 \mathcal{F}_x^{\mu\nu}\mathcal{F}_x^{\alpha'\beta'}\delta^n(x)\right. \\ & + \frac{2}{3}\frac{m^2}{(n-2)}\left(\eta^{\mu\nu}\eta^{\alpha'\beta'}\Box_x - \eta^{\mu(\alpha'}\eta^{\beta')\nu}\Box_x + \eta^{\mu(\alpha'}\partial_x^{\beta')}\partial_x^{\nu}\right. \\ & \left. + \eta^{\nu(\alpha'}\partial_x^{\beta')}\partial_x^{\mu} - \eta^{\mu\nu}\partial_x^{\alpha'}\partial_x^{\beta'} - \eta^{\alpha'\beta'}\partial_x^{\mu}\partial_x^{\nu}\right)\delta^n(x) \\ & \left. + \frac{4m^4}{n(n-2)}\left(2\eta^{\mu(\alpha'}\eta^{\beta')\nu} - \eta^{\mu\nu}\eta^{\alpha'\beta'}\right)\delta^n(x)\right] \\ & + \frac{1}{180}\mathcal{F}_x^{\mu\nu\alpha'\beta'}\int \frac{d^n p}{(2\pi)^n} e^{ip\cdot x}\left(1+4\,\frac{m^2}{p^2}\right)^2 \bar{\phi}(p^2) \\ & + \frac{2}{9}\,\mathcal{F}_x^{\mu\nu}\mathcal{F}_x^{\alpha'\beta'}\int \frac{d^n p}{(2\pi)^n}\, e^{ip\cdot x}\left(3\Delta\xi + \frac{m^2}{p^2}\right)^2 \bar{\phi}(p^2) \\ & - \left[\frac{4}{675}\,\mathcal{F}_x^{\mu\nu\alpha'\beta'} + \frac{1}{270}\,(60\xi - 11)\,\mathcal{F}_x^{\mu\nu}\mathcal{F}_x^{\alpha'\beta'}\right]\delta^n(x) \\ & - m^2\left[\frac{2}{135}\,\mathcal{F}_x^{\mu\nu\alpha'\beta'} + \frac{1}{27}\,\mathcal{F}_x^{\mu\nu}\mathcal{F}_x^{\alpha'\beta'}\right]\Delta_n(x)\Bigg\} + \mathcal{O}(n-4), \end{aligned} \tag{11.24}$$

where κ_n has been defined in (11.9), and $\bar{\phi}(p^2)$ and $\Delta_n(x)$ are given by

$$\begin{aligned} \bar{\phi}(p^2) &\equiv \int_0^1 d\alpha \, \ln\!\left(1+\frac{p^2}{m^2}\alpha(1-\alpha) - i\epsilon\right) \\ &= -i\pi\theta(-p^2-4m^2)\sqrt{1+4\frac{m^2}{p^2}} + \varphi(p^2), \end{aligned} \tag{11.25}$$

$$\Delta_n(x) \equiv \int \frac{d^n p}{(2\pi)^n}\, e^{ip\cdot x}\, \frac{1}{p^2}, \tag{11.26}$$

where $\varphi(p^2) \equiv \int_0^1 d\alpha \, \ln|1+\frac{p^2}{m^2}\,\alpha(1-\alpha)|$. The imaginary part of (11.24) gives the kernel components $\mu^{-(n-4)} H_{\mathrm{S}_n}^{\mu\nu\alpha'\beta'}(x,y)$, according to (11.19). It can be easily obtained multiplying this expression by $-i$ and retaining only the real part, $\varphi(p^2)$, of the function $\bar{\phi}(p^2)$.

11.3 Einstein–Langevin Equation

With the previous results for the kernels we can now write the n-dimensional Einstein–Langevin equation (11.24), previous to the renormalization. Taking also into account Eqs. (11.5) and (11.6), we may finally write:

$$\begin{aligned}
&\frac{1}{8\pi G_B} G^{(1)\mu\nu}(x) - \frac{4}{3}\,\alpha_B D^{(1)\mu\nu}(x) - 2\beta_B B^{(1)\mu\nu}(x) \\
&\quad + \frac{\kappa_n}{(4\pi)^2}\left[-4\Delta\xi\,\frac{m^2}{(n-2)}\,G^{(1)\mu\nu} + \frac{1}{90}\,D^{(1)\mu\nu}\Delta\xi^2 B^{(1)\mu\nu}\right](x) \\
&\quad + \frac{1}{2880\pi^2}\left\{-\frac{16}{15}\,D^{(1)\mu\nu}(x) + \left(\frac{1}{6} - 10\Delta\xi\right) B^{(1)\mu\nu}(x)\right. \\
&\quad + \int d^n y \int \frac{d^n p}{(2\pi)^n}\, e^{ip\cdot(x-y)}\,\varphi(p^2)\,\left[\left(1+4\frac{m^2}{p^2}\right)^2 D^{(1)\mu\nu}(y)\right. \\
&\qquad\qquad\qquad\qquad\qquad \left. +10\left(3\Delta\xi + \frac{m^2}{p^2}\right)^2 B^{(1)\mu\nu}(y)\right] \\
&\quad \left. - \frac{m^2}{3}\int d^n y\,\Delta_n(x-y)\left(8D^{(1)\mu\nu} + 5B^{(1)\mu\nu}\right)(y)\right\} \\
&\quad + \frac{1}{2}\int d^n y\,\mu^{-(n-4)} H_{\mathrm{A}_n}^{\mu\nu\alpha'\beta'}(x,y)\,h_{\alpha'\beta'}(y) + \mathcal{O}(n-4) = \xi^{\mu\nu}(x).
\end{aligned} \tag{11.27}$$

Notice that the terms containing the bare cosmological constant have cancelled. These equations can now be renormalized, that is, we can now write the bare coupling constants as renormalized coupling constants plus some suitably chosen counterterms and take the limit $n \to 4$. In order to carry out such a procedure, it is convenient to distinguish between massive and massless scalar fields.

It is convenient to introduce the two new kernels

$$\begin{aligned}
H_{\mathrm{A}}(x;m^2) \equiv \frac{1}{480\pi^2}\int \frac{d^4 p}{(2\pi)^4}\, e^{ip\cdot x}\left\{\left(1+4\,\frac{m^2}{p^2}\right)^2\right. \\
\left. \times\left[-i\pi\mathrm{sign}(p^0)\theta(-p^2-4m^2)\sqrt{1+4\frac{m^2}{p^2}} + \varphi(p^2)\right] - \frac{8}{3}\,\frac{m^2}{p^2}\right\},
\end{aligned}$$

$$H_{\rm B}(x;m^2,\Delta\xi) \equiv \frac{1}{72\pi^2}\int \frac{d^4p}{(2\pi)^4} e^{ip\cdot x}\Bigg\{\left(3\Delta\xi+\frac{m^2}{p^2}\right)^2 \times\left[-i\pi\mathrm{sign}(p^0)\theta(-p^2-4m^2)\sqrt{1+4\frac{m^2}{p^2}}+\varphi(p^2)\right]-\frac{1}{6}\frac{m^2}{p^2}\Bigg\}, \tag{11.28}$$

where $\varphi(p^2)$ is given by the restriction to $n=4$ of expression (11.25). The renormalized coupling constants $1/G_{\rm N}$, α and β are easily computed as it was done in Eq. (11.8). Substituting their expressions into Eq. (11.27), we can take the limit $n\to 4$, using the fact that, for $n=4$, $D^{(1)\mu\nu}(x)=(3/2)\,A^{(1)\mu\nu}(x)$, we obtain the corresponding semiclassical Einstein–Langevin equation.

For the massless case one needs the limit $m\to 0$ of Eq. (11.27). In this case it is convenient to separate κ_n in (11.9) as $\kappa_n=\tilde{\kappa}_n+\frac{1}{2}\ln(m^2/\mu^2)+\mathcal{O}(n-4)$, where

$$\tilde{\kappa}_n \equiv \frac{1}{(n-4)}\left(\frac{e^\gamma}{4\pi}\right)^{\frac{n-4}{2}} = \frac{1}{n-4}+\frac{1}{2}\ln\left(\frac{e^\gamma}{4\pi}\right)+\mathcal{O}(n-4), \tag{11.29}$$

and use that, from Eq. (11.25), we have

$$\lim_{m^2\to 0}\left[\varphi(p^2)+\ln(m^2/\mu^2)\right] = -2+\ln\left|\frac{p^2}{\mu^2}\right|. \tag{11.30}$$

The coupling constants are then easily renormalized. We note that in the massless limit, the Newtonian gravitational constant is not renormalized and, in the conformal coupling case, $\Delta\xi=0$, we have that $\beta_B=\beta$. Note also that, by making $m=0$ in (11.17), the noise and dissipation kernels can be written as

$$\begin{aligned} &N_{\rm A}(x;m^2=0)=N(x), \qquad N_{\rm B}(x;m^2=0,\Delta\xi)=60\Delta\xi^2 N(x),\\ &D_{\rm A}(x;m^2=0)=D(x), \qquad D_{\rm B}(x;m^2=0,\Delta\xi)=60\Delta\xi^2 D(x), \end{aligned} \tag{11.31}$$

where

$$N(x)\equiv\frac{1}{480\pi}\int\frac{d^4p}{(2\pi)^4}e^{ip\cdot x}\theta(-p^2), \qquad D(x)\equiv\frac{-i}{480\pi}\int\frac{d^4p}{(2\pi)^4}e^{ip\cdot x}\mathrm{sign}\,p^0\theta(-p^2). \tag{11.32}$$

It is also convenient to introduce the new kernel

$$\begin{aligned} H(x;\mu^2) &\equiv \frac{1}{480\pi^2}\int\frac{d^4p}{(2\pi)^4}\,e^{ip\cdot x}\left[\ln\left|\frac{p^2}{\mu^2}\right|-i\pi\,\mathrm{sign}(p^0)\theta(-p^2)\right]\\ &=\frac{1}{480\pi^2}\lim_{\epsilon\to 0^+}\int\frac{d^4p}{(2\pi)^4}\,e^{ip\cdot x}\,\ln\left(\frac{-(p^0+i\epsilon)^2+p^ip_i}{\mu^2}\right). \end{aligned} \tag{11.33}$$

This kernel is real and can be written as the sum of an even part and an odd part in the variables x^μ, where the odd part is the dissipation kernel $D(x)$. The Fourier transforms (11.32) and (11.33) can actually be computed and, thus, in this case we have explicit expressions for the kernels in position space; see for instance [722, 723, 213].

Finally, the Einstein–Langevin equation for the physical stochastic perturbations $h_{\mu\nu}$ can be written in both cases, for $m \neq 0$ and for $m = 0$, as

$$\frac{1}{8\pi G_{\mathrm{N}}} G^{(1)\mu\nu}(x) - 2\left(\bar{\alpha} A^{(1)\mu\nu}(x) + \bar{\beta} B^{(1)\mu\nu}(x)\right) + \frac{1}{4}\int d^4y \left[H_{\mathrm{A}}(x-y) A^{(1)\mu\nu}(y) + H_{\mathrm{B}}(x-y) B^{(1)\mu\nu}(y)\right] = \xi^{\mu\nu}(x), \quad (11.34)$$

where in terms of the renormalized constants α_d and β_d the new constants are $\bar{\alpha} = \alpha_d (8\pi G_{\mathrm{N}})^{-1} + (3600\pi^2)^{-1}$ and $\bar{\beta} = \beta_d (8\pi G_{\mathrm{N}})^{-1} - (1/12 - 5\Delta\xi)(2880\pi^2)^{-1}$. The kernels $H_{\mathrm{A}}(x)$ and $H_{\mathrm{B}}(x)$ are given by Eqs. (11.28) when $m \neq 0$, and $H_{\mathrm{A}}(x) = H(x;\mu^2)$, $H_{\mathrm{B}}(x) = 60\Delta\xi^2 H(x;\mu^2)$ when $m = 0$.

In the massless case, we can use the arbitrariness of the mass scale μ to eliminate one of the parameters $\bar{\alpha}$ or $\bar{\beta}$. The components of the Gaussian stochastic source $\xi^{\mu\nu}$ have zero mean value and their two-point correlation functions are given by $\langle \xi^{\mu\nu}(x) \xi^{\alpha'\beta'}(y) \rangle_s = N^{\mu\nu\alpha'\beta'}(x,y)$, where the noise kernel is given in (11.18), which in the massless case reduces to (11.31).

It is interesting to consider the massless conformally coupled scalar field, i.e., the case $\Delta\xi = 0$, of particular interest because of its similarities with the electromagnetic field, and also because of its interest in cosmology: matter fields in the standard model of particle physics become effectively conformally invariant in the very early universe. We have already mentioned that for a conformally coupled field, the stochastic source tensor must be traceless (up to first-order perturbations around semiclassical gravity), in the sense that the stochastic variable $\xi^{\mu}_{\mu} \equiv \eta_{\mu\nu}\xi^{\mu\nu}$ behaves deterministically as a vanishing scalar field. This can be directly checked by noticing, from Eqs. (11.18) and (11.31), that, when $\Delta\xi = 0$, one has $\langle \xi^{\mu}_{\mu}(x) \xi^{\alpha'\beta'}(y) \rangle_s = 0$, since $\mathcal{F}^{\mu}_{\mu} = 3\Box$ and $\mathcal{F}^{\mu\alpha}\mathcal{F}^{\beta}_{\mu} = \Box \mathcal{F}^{\alpha\beta}$. The Einstein–Langevin equations for this particular case (and generalized to a spatially flat FLRW background) were first obtained by Campos and Verdaguer [186]. We review this derivation in Chapter 12.

11.3.1 Stress-Energy Tensor

The expectation value of the renormalized stress-energy tensor for a scalar field can be obtained by comparing the previous equation (11.34) with the Einstein–Langevin equation (10.14).

For a massive scalar field, $m \neq 0$, we get

$$\begin{aligned}\langle \hat{T}^{(1)\mu\nu}_{\mathrm{R}}(x) \rangle = & \frac{1}{2880\pi^2}\left[\frac{8}{5} A^{(1)\mu\nu}(x) - \left(\frac{1}{6} - 10\Delta\xi\right) B^{(1)\mu\nu}(x)\right] \\ & - \int d^4y \left[H_{\mathrm{A}}(x-y;m^2) A^{(1)\mu\nu}(y)\right. \\ & \left. + H_{\mathrm{B}}(x-y;m^2,\Delta\xi) B^{(1)\mu\nu}(y)\right] + \mathcal{O}(h^2). \qquad (11.35)\end{aligned}$$

For a massless scalar field, $m = 0$, we get

$$\begin{aligned}\langle \hat{T}_{\mathrm{R}}^{(1)\mu\nu}(x)\rangle = \frac{1}{2880\pi^2}\left[\frac{8}{5}\,A^{(1)\mu\nu}(x) - \left(\frac{1}{6} - 10\Delta\xi\right) B^{(1)\mu\nu}(x)\right]& \\ -\int d^4y\, H(x-y;\mu^2)\left[A^{(1)\mu\nu}(y)\right.& \\ \left.+60\Delta\xi^2 B^{(1)\mu\nu}(y)\right] + \mathcal{O}(h^2),& \qquad (11.36)\end{aligned}$$

Notice that in the massive case we have chosen, as usual, a renormalization scheme such that the expectation value of the renormalized stress-energy tensor does not have local terms proportional to the metric and the Einstein tensor.

The result (11.36) for a massless field agree with the general form found by Horowitz [213, 698] using an axiomatic approach and coincides with that given in [128]. The particular cases of conformal coupling, $\Delta\xi = 0$, and minimal coupling, $\Delta\xi = -1/6$, are also in agreement with the results for this cases given in [213, 698, 724, 423, 677], modulo local terms proportional to $A^{(1)\mu\nu}$ and $B^{(1)\mu\nu}$ due to different choices of the renormalization scheme. For the case of a massive minimally coupled scalar field, $\Delta\xi = -1/6$, the result (11.35) is equivalent to that found by Tichy and Flanagan [725].

11.4 Solutions of the Einstein–Langevin Equation

Here we solve the Einstein–Langevin equations (11.34) for the components $G^{(1)\mu\nu}$ of the linearized Einstein tensor. Then we use these solutions to compute the corresponding two-point correlation functions, which give a measure of the gravitational fluctuations predicted by the stochastic semiclassical theory of gravity in the present case. Since the linearized Einstein tensor is invariant under gauge transformations of the metric perturbations, these two-point correlation functions are also gauge-invariant. Once we have computed the two-point correlation functions for the linearized Einstein tensor, we find the solutions for the metric perturbations and compute the associated two-point correlation functions. The procedure used to solve the Einstein–Langevin equation is similar to the one used by Horowitz [213], see also [128], to analyze the stability of Minkowski spacetime in semiclassical gravity.

We first note that the tensors $A^{(1)\mu\nu}$ and $B^{(1)\mu\nu}$ can be written in terms of $G^{(1)\mu\nu}$ as

$$A^{(1)\mu\nu} = \frac{2}{3}\,(\mathcal{F}^{\mu\nu} G^{(1)\alpha}_{\alpha} - \mathcal{F}^{\alpha}_{\alpha} G^{(1)\mu\nu}), \qquad B^{(1)\mu\nu} = 2\mathcal{F}^{\mu\nu} G^{(1)\alpha}{}_{\alpha}, \qquad (11.37)$$

where we have used that $3\Box = \mathcal{F}^{\alpha}_{\alpha}$. Therefore, the Einstein–Langevin equation (11.34) can be seen as a linear integro-differential stochastic equation for the components $G^{(1)\mu\nu}$. In order to find solutions to Eq. (11.34), it is convenient to Fourier transform this equation. With the convention $\tilde{f}(p) = \int d^4x e^{-ip\cdot x} f(x)$ for a given field $f(x)$, one finds, from (11.37),

$$\begin{aligned}
\tilde{A}^{(1)\mu\nu}(p) &= 2p^2\tilde{G}^{(1)\mu\nu}(p) - \frac{2}{3}p^2 P^{\mu\nu}\tilde{G}^{(1)\alpha}_{\ \ \alpha}(p), \\
\tilde{B}^{(1)\mu\nu}(p) &= -2p^2 P^{\mu\nu}\tilde{G}^{(1)\alpha}_{\ \ \alpha}(p).
\end{aligned} \tag{11.38}$$

The Fourier transform of the Einstein–Langevin equation (11.34) now reads

$$F^{\mu\nu}_{\alpha'\beta'}(p)\,\tilde{G}^{(1)\alpha'\beta'}(p) = 8\pi G_{\mathrm{N}}\,\tilde{\xi}^{\mu\nu}(p), \tag{11.39}$$

where

$$F^{\mu\nu}_{\alpha'\beta'}(p) \equiv F_1(p)\,\delta^{\mu}_{(\alpha'}\delta^{\nu}_{\beta')} + F_2(p)\,p^2 P^{\mu\nu}\eta_{\alpha'\beta'}, \tag{11.40}$$

with

$$\begin{aligned}
F_1(p) &\equiv 1 + 16\pi G_{\mathrm{N}}\,p^2\left[\frac{1}{4}\tilde{H}_{\mathrm{A}}(p) - 2\bar{\alpha}\right], \\
F_2(p) &\equiv -\frac{16}{3}\,\pi G_{\mathrm{N}}\left[\frac{1}{4}\tilde{H}_{\mathrm{A}}(p) + \frac{3}{4}\tilde{H}_{\mathrm{B}}(p) - 2\bar{\alpha} - 6\bar{\beta}\right].
\end{aligned} \tag{11.41}$$

In the Fourier transformed Einstein–Langevin equation (11.39), $\tilde{\xi}^{\mu\nu}(p)$, the Fourier transform of $\xi^{\mu\nu}(x)$, is a Gaussian stochastic source of zero average and

$$\langle\tilde{\xi}^{\mu\nu}(p)\tilde{\xi}^{\alpha'\beta'}(p')\rangle_s = (2\pi)^4\,\delta^4(p+p')\,\tilde{N}^{\mu\nu\alpha'\beta'}(p), \tag{11.42}$$

where we have introduced the Fourier transform of the noise kernel. The explicit expression for $\tilde{N}^{\mu\nu\alpha'\beta'}(p)$ is found from (11.17) and (11.18) to be

$$\begin{aligned}
\tilde{N}^{\mu\nu\alpha'\beta'}(p) = \frac{1}{720\pi}\theta(-p^2-4m^2)\sqrt{1+4\,\frac{m^2}{p^2}}\,&\left[\frac{1}{4}\left(1+4\,\frac{m^2}{p^2}\right)^2(p^2)^2P^{\mu\nu\alpha'\beta'}\right. \\
&\left.+10\left(3\Delta\xi+\frac{m^2}{p^2}\right)^2(p^2)^2P^{\mu\nu}P^{\alpha'\beta'}\right],
\end{aligned} \tag{11.43}$$

which in the massless case reduces to

$$\lim_{m\to 0}\tilde{N}^{\mu\nu\alpha'\beta'}(p) = \frac{1}{480\pi}\theta(-p^2)\left[\frac{1}{6}\,(p^2)^2\,P^{\mu\nu\alpha'\beta'} + 60\Delta\xi^2(p^2)^2P^{\mu\nu}P^{\alpha'\beta'}\right]. \tag{11.44}$$

11.4.1 Linearized Einstein Tensor Correlator

In general, we can write $G^{(1)\mu\nu} = \langle G^{(1)\mu\nu}\rangle_s + G^{(1)\mu\nu}_{\mathrm{f}}$, where $G^{(1)\mu\nu}_{\mathrm{f}}$ is a solution to Eq. (11.34) with zero average, or (11.39) in the Fourier transformed version. The average $\langle G^{(1)\mu\nu}\rangle_s$ must be a solution of the linearized semiclassical Einstein equation obtained by averaging Eq. (11.34), or (11.39). Solutions to this semiclassical equation (specially in the massless case, $m=0$) have been studied [213, 698, 726, 727, 728, 729, 730, 194, 287, 677, 128], particularly in connection with the problem of the stability of the ground state of semiclassical gravity. The two-point correlation functions for the linearized Einstein tensor are defined by

$$\begin{aligned}\mathcal{G}^{\mu\nu\alpha'\beta'}(x,x') &\equiv \langle G^{(1)\mu\nu}(x)G^{(1)\alpha'\beta'}(x')\rangle_s - \langle G^{(1)\mu\nu}(x)\rangle_s \langle G^{(1)\alpha'\beta'}(x')\rangle_s \\ &= \langle G^{(1)\mu\nu}_{\rm f}(x)G^{(1)\alpha'\beta'}_{\rm f}(x')\rangle_s. \qquad (11.45)\end{aligned}$$

Now we shall seek the family of solutions to the Einstein–Langevin equation which can be written as a linear functional of the stochastic source and whose Fourier transform, $\tilde{G}^{(1)\mu\nu}(p)$, depends locally on $\tilde{\xi}^{\alpha\beta}(p)$. Each of such solutions is a Gaussian stochastic field and, thus, it can be completely characterized by the averages $\langle G^{(1)\mu\nu}\rangle_s$ and the two-point correlation functions (11.45). For such a family of solutions, $\tilde{G}^{(1)\mu\nu}_{\rm f}(p)$ is the most general solution to Eq. (11.39) which is linear, homogeneous and local in $\tilde{\xi}^{\alpha\beta}(p)$. It can be written as

$$\tilde{G}^{(1)\mu\nu}_{\rm f}(p) = 8\pi G_{\rm N}\, D^{\mu\nu}_{\alpha'\beta'}(p)\, \tilde{\xi}^{\alpha'\beta'}(p), \qquad (11.46)$$

where $D^{\mu\nu}_{\alpha'\beta'}(p)$ are the components of a Lorentz invariant tensor field distribution in Minkowski spacetime (by "Lorentz-invariant" we mean invariant under the transformations of the orthochronous Lorentz subgroup; see [213] for more details on the definition and properties of these tensor distributions). This tensor is symmetric under the interchanges of $\alpha \leftrightarrow \beta$ and $\mu \leftrightarrow \nu$, and is the most general solution of

$$F^{\mu\nu}_{\hat{\rho}\hat{\sigma}}(p)\, D^{\hat{\rho}\hat{\sigma}}_{\alpha'\beta'}(p) = \delta^{\mu}_{(\alpha'}\delta^{\nu}_{\beta')}. \qquad (11.47)$$

In addition, we must impose the conservation condition: $p_\nu \tilde{G}^{(1)\mu\nu}_{\rm f}(p) = 0$, where this zero must be understood as a stochastic variable which behaves deterministically as a zero vector field. We can write $D^{\mu\nu}_{\alpha'\beta'}(p) = D^{\mu\nu}_{p\,\alpha'\beta'}(p) + D^{\mu\nu}_{h\,\alpha'\beta'}(p)$, where $D^{\mu\nu}_{p\,\alpha'\beta'}(p)$ is a particular solution to Eq. (11.47) and $D^{\mu\nu}_{h\,\alpha'\beta'}(p)$ is the most general solution to the homogeneous equation. Consequently, see Eq. (11.46), we can write $\tilde{G}^{(1)\mu\nu}_{\rm f}(p) = \tilde{G}^{(1)\mu\nu}_{p}(p) + \tilde{G}^{(1)\mu\nu}_{h}(p)$.

To find the particular solution, we try an ansatz of the form

$$D^{\mu\nu}_{p\,\alpha'\beta'}(p) = d_1(p)\, \delta^{\mu}_{(\alpha'}\delta^{\nu}_{\beta')} + d_2(p)\, p^2 P^{\mu\nu}\eta_{\alpha'\beta'}. \qquad (11.48)$$

Substituting this ansatz into Eq. (11.47), it is easy to see that it solves this equation if we take

$$d_1(p) = \left[\frac{1}{F_1(p)}\right]_r, \qquad d_2(p) = -\left[\frac{F_2(p)}{F_1(p)F_3(p)}\right]_r, \qquad (11.49)$$

with

$$F_3(p) \equiv F_1(p) + 3p^2F_2(p) = 1 - 48\pi G_{\rm N}\, p^2 \left[\frac{1}{4}\tilde{H}_{\rm B}(p) - 2\bar{\beta}\right], \qquad (11.50)$$

and where the notation $[\quad]_r$ means that the zeros of the denominators are regulated with appropriate prescriptions in such a way that $d_1(p)$ and $d_2(p)$

are well defined Lorentz-invariant scalar distributions. This yields a particular solution to the Einstein–Langevin equation:

$$\tilde{G}_p^{(1)\mu\nu}(p) = 8\pi G_{\rm N}\, D^{\mu\nu}_{p\,\alpha'\beta'}(p)\, \tilde{\xi}^{\alpha'\beta'}(p), \tag{11.51}$$

which, since the stochastic source is conserved, satisfies the conservation condition. Note that, in the case of a massless scalar field, $m = 0$, the above solution has a functional form analogous to that of the solutions of linearized semiclassical gravity found in the appendix of [128]. Notice also that, for a massless conformally coupled field, $m = 0$ and $\Delta\xi = 0$, the second term on the right-hand side of Eq. (11.48) will not contribute in the correlation functions (11.45), since in this case the stochastic source is traceless.

A detailed analysis given in [189] concludes that the homogeneous solution $\tilde{G}_h^{(1)\mu\nu}(p)$ gives no contribution to the correlation functions (11.45). Consequently $\mathcal{G}^{\mu\nu\alpha'\beta'}(x,x') = \langle G_p^{(1)\mu\nu}(x) G_p^{(1)\alpha'\beta'}(x')\rangle_s$, where $G_p^{(1)\mu\nu}(x)$ is the inverse Fourier transform of (11.51), and the correlation functions (11.45) are

$$\langle \tilde{G}_p^{(1)\mu\nu}(p) \tilde{G}_p^{(1)\alpha'\beta'}(p')\rangle_s = 64(2\pi)^6 G_{\rm N}^2 \delta^4(p+p') D^{\mu\nu}_{p\hat{\rho}\hat{\sigma}}(p) D^{\alpha'\beta'}_{p\bar{\lambda}\bar{\gamma}}(-p) \tilde{N}^{\hat{\rho}\hat{\sigma}\bar{\lambda}\bar{\gamma}}(p). \tag{11.52}$$

It is easy to see from the above analysis that the prescriptions $[\ \ ]_r$ in the factors D_p are irrelevant in the last expression and, thus, they can be suppressed. Taking into account that $F_l(-p) = F_l^*(p)$, with $l = 1, 2, 3$, we get from Eqs. (11.48) and (11.49)

$$\begin{aligned}
\langle \tilde{G}_p^{(1)\mu\nu}(p)\, \tilde{G}_p^{(1)\alpha'\beta'}(p')\rangle_s = {}& 64(2\pi)^6\, G_{\rm N}^2\, \frac{\delta^4(p+p')}{|F_1(p)|^2} \\
& \times \left[\tilde{N}^{\mu\nu\alpha'\beta'}(p) - \frac{F_2(p)}{F_3(p)}\, p^2 P^{\mu\nu}\, \tilde{N}^{\alpha'\beta'}{}_{\rho}^{\rho}(p) \right. \\
& - \frac{F_2^*(p)}{F_3^*(p)}\, p^2 P^{\alpha'\beta'}\, \tilde{N}^{\mu\nu}{}_{\rho'}^{\rho'}(p) \\
& \left. + \frac{|F_2(p)|^2}{|F_3(p)|^2} p^2 P^{\mu\nu} p^2 P^{\alpha'\beta'}\, \tilde{N}^{\rho}{}_{\rho}{}^{\sigma'}{}_{\sigma'}(p) \right].
\end{aligned} \tag{11.53}$$

This last expression is well defined as a bi-distribution and can be easily evaluated using Eq. (11.43). The final explicit result for the Fourier transformed correlation function for the Einstein tensor is thus

$$\begin{aligned}
\langle \tilde{G}_p^{(1)\mu\nu}(p)\, \tilde{G}_p^{(1)\alpha'\beta'}(p')\rangle_s = {}& \frac{2G_{\rm N}^2}{45}(2\pi)^5 \frac{\delta^4(p+p')}{|F_1(p)|^2} \theta(-p^2-4m^2) \sqrt{1+4\frac{m^2}{p^2}} \\
& \times \left[\frac{1}{4}\left(1+4\,\frac{m^2}{p^2}\right)^2 (p^2)^2\, P^{\mu\nu\alpha'\beta'} \right. \\
& \left. +10\left(3\Delta\xi + \frac{m^2}{p^2}\right)^2 (p^2)^2 P^{\mu\nu} P^{\alpha'\beta'} \left|1 - 3p^2 \frac{F_2(p)}{F_3(p)}\right|^2 \right].
\end{aligned} \tag{11.54}$$

To obtain the correlation functions in coordinate space, i.e., Eq. (11.45), we take the inverse Fourier transform. The final result is:

$$\mathcal{G}^{\mu\nu\alpha'\beta'}(x,x') = \frac{\pi G_{\rm N}^2}{45}\mathcal{F}_x^{\mu\nu\alpha'\beta'}\mathcal{G}_{\rm A}(x-x') + \frac{8\pi G_{\rm N}^2}{9}\mathcal{F}_x^{\mu\nu}\mathcal{F}_x^{\alpha'\beta'}\mathcal{G}_{\rm B}(x-x'), \quad (11.55)$$

with

$$\begin{aligned}\tilde{\mathcal{G}}_{\rm B}(p) &\equiv \theta(-p^2-4m^2)\sqrt{1+4\frac{m^2}{p^2}}\left(3\Delta\xi+\frac{m^2}{p^2}\right)^2\frac{1}{|F_1(p)|^2}\left|1-3p^2\frac{F_2(p)}{F_3(p)}\right|^2,\\ \tilde{\mathcal{G}}_{\rm A}(p) &\equiv \theta(-p^2-4m^2)\sqrt{1+4\,\frac{m^2}{p^2}}\left(1+4\frac{m^2}{p^2}\right)^2\frac{1}{|F_1(p)|^2},\end{aligned} \quad (11.56)$$

where $F_l(p), l=1,2,3$, are given in (11.41) and (11.50).

Notice that, for a massless field ($m=0$), we have

$$\begin{aligned}F_1(p) &= 1+4\pi G_{\rm N}p^2\tilde{H}(p;\bar{\mu}^2),\\ F_2(p) &= -\frac{16}{3}\pi G_{\rm N}\left[(1+180\Delta\xi^2)\frac{1}{4}\tilde{H}(p;\bar{\mu}^2)-6\Upsilon\right],\\ F_3(p) &= 1-48\pi G_{\rm N}p^2\left[15\Delta\xi^2\tilde{H}(p;\bar{\mu}^2)-2\Upsilon\right],\end{aligned} \quad (11.57)$$

with $\bar{\mu} \equiv \mu\exp(1920\pi^2\bar{\alpha})$ and $\Upsilon \equiv \bar{\beta}-60\Delta\xi^2\bar{\alpha}$, and where $\tilde{H}(p;\mu^2)$ is the Fourier transform of $H(x;\mu^2)$ given in (11.33).

11.4.2 Two-Point Correlations for Metric Perturbations

Starting from the solutions found for the linearized Einstein tensor, which are characterized by the two-point correlation functions (11.55) or, in terms of Fourier transforms, (11.54), we can now solve the equations for the metric perturbations. Working in the harmonic gauge, $\partial_\nu\bar{h}^{\mu\nu}=0$ (this zero must be understood in a statistical sense) where $\bar{h}_{\mu\nu} \equiv h_{\mu\nu}-(1/2)\eta_{\mu\nu}h^\alpha_\alpha$, the equations for the metric perturbations in terms of the Einstein tensor are

$$\Box\bar{h}^{\mu\nu}(x) = -2G^{(1)\mu\nu}(x), \quad (11.58)$$

or, in terms of Fourier transforms, $p^2\tilde{\bar{h}}^{\mu\nu}(p) = 2\tilde{G}^{(1)\mu\nu}(p)$. Similarly to the analysis of the equation for the Einstein tensor, we can write $\bar{h}^{\mu\nu} = \langle\bar{h}^{\mu\nu}\rangle_s+\bar{h}^{\mu\nu}_{\rm f}$, where $\bar{h}^{\mu\nu}_{\rm f}$ is a solution to these equations with zero average, and the two-point correlation functions are defined by

$$\begin{aligned}\mathcal{H}^{\mu\nu\alpha'\beta'}(x,x') &\equiv \langle\bar{h}^{\mu\nu}(x)\bar{h}^{\alpha'\beta'}(x')\rangle_s-\langle\bar{h}^{\mu\nu}(x)\rangle_s\langle\bar{h}^{\alpha'\beta'}(x')\rangle_s\\ &= \langle\bar{h}^{\mu\nu}_{\rm f}(x)\bar{h}^{\alpha'\beta'}_{\rm f}(x')\rangle_s.\end{aligned} \quad (11.59)$$

We can now seek solutions of the Fourier transform of Eq. (11.58) of the form $\tilde{\bar{h}}^{\mu\nu}_{\rm f}(p) = 2D(p)\tilde{G}^{(1)\mu\nu}_{\rm f}(p)$, where $D(p)$ is a Lorentz-invariant scalar distribution

15

Stress-Energy Tensor Fluctuations in de Sitter Space

The geometry of most inflationary models as well as the late time behavior of our expanding universe are well approximated by the de Sitter geometry. For this reason quantum metric perturbations around de Sitter and near-de Sitter spacetimes have been extensively studied in the context of inflationary models. The standard analysis is based on the linearized computation of the two-point function of the metric perturbations [954, 734, 753]. But the need for matter loop corrections was emphasized by Weinberg [427], who pointed out also the importance of testing perturbation theory in cosmology given the success of the tree-level approximation; see also Senatore and Zaldarriaga [567]. A series of works including matter loop corrections in different models have been performed using different approaches and models [955, 956, 957] including stochastic gravity [958, 203].

However, certain aspects such as the significance of possible large infrared effects, have not been completely settled yet. The standard theoretical framework also assumes that the background dynamics of de Sitter and near-de Sitter spacetimes is not affected by large quantum corrections. However, there is also a number of studies that claim that one-loop corrections could give a significant backreaction in inflationary models [959, 960, 961, 576, 577]. It has also been argued that higher-order radiative corrections from graviton two-loops could give rise to a secular screening of the cosmological constant for pure gravity in de Sitter spacetime [962, 963]. Such nonlinear effects are dominated by infrared modes which are copiously generated by the exponential expansion and should be fairly insensitive to the short-distance behavior of quantum gravity. However, whether these large effects really take place is still uncertain, mainly due to the gauge dependence of the results [964, 965, 966, 508, 575].

The existence of large quantum nonlinear effects like those mentioned above could lead to a breaking of de Sitter invariance, as suggested by Tsamis and Woodard [962, 963] or by Polyakov [967]. In order to gain insight into this possibility, it seems convenient to study quantum metric perturbations around de Sitter spacetime keeping de Sitter invariance always manifest. This can be achieved with the formalism of the de Sitter-invariant bitensors introduced by Allen and Jacobson [476]. The tree-level graviton two-point function has been computed using this formalism by different authors such as Allen and Turyn [968], Antoniadis and Mottola [969] and Higuchi and Kouris [970]. A conclusion that can be drawn from their results is that the two-point function of the linearized Riemann tensor ${R_{ab}}^{cd}$, which is gauge-invariant around de Sitter spacetime, decays like the fourth power of the physical distance between the two points; see Kouris [971].

We wish to go beyond the linear level and study the effects of matter loops on the quantum metric fluctuations. However, a first step in this direction is to study the stress-energy tensor fluctuations of the matter fields described by the noise kernel. Calculating the noise kernel in de Sitter spacetime, in a de Sitter-invariant vacuum, and with the formalism of de Sitter-invariant bitensors, is the main goal of this chapter. In subsequent chapters quantum gravitational field correlations, including one-loop corrections from matter fields, in de Sitter spacetime are studied.

We start by reviewing in Section 15.1 the basic geometric properties of de Sitter spacetime and the invariant bitensors that will be used in the rest of the chapter and in later chapters. These tools are employed in Section 15.2 to write the noise kernel for spacelike separated points in de Sitter-invariant form, and explicit expressions for the case of a free minimally coupled scalar field are derived. Furthermore, closed results in terms of elementary functions can be given for the particular cases of small masses, vanishing mass, and large separations. A massless limit discontinuity is found and is discussed in some detail. In Section 15.3 an alternative analysis of the noise kernel based on the approximate solutions of the field modes is given, which also helps to clarify the above discontinuity. Finally we discuss the implications of our results for the quantum metric fluctuations around de Sitter spacetime in Section 15.4.

15.1 De Sitter Geometry and Invariant Bitensors

The D-dimensional de Sitter spacetime, dS_D, may be defined as a submanifold embedded in $(D+1)$-dimensional Minkowski spacetime, whose points satisfy the equation

$$\eta_{AB}X^A(x)X^B(x) = H^{-2}, \tag{15.1}$$

where H is a parameter with dimensions of mass or inverse length (we use $\hbar = c = 1$), η_{AB} the Minkowski metric, and $X^A(x)$ is the position vector in Minkowski spacetime of the point x. The hyperboloid is invariant under Lorentz transformations $X^A = \Lambda^A_B X^B$ and these define the isometries σ of dS_D: $X^A(\sigma(x)) = \Lambda^A_B X^B(x)$. Thus, the isometries of de Sitter form the Lorentz group

$O(D,1)$. Another isometry that will also play a role is the antipodal isometry $\alpha(x)$ defined by $X^A(\alpha(x)) = -X^A(x)$.

One may introduce spatially flat coordinates by labelling the points in dS_D with the coordinates $\{t, x^i\}$, with $i = 1, \ldots, D-1$, defined by

$$\begin{aligned} X^0 &= \frac{1}{H}\sinh(Ht) + \frac{1}{2}He^{Ht}\delta_{ij}x^i x^j, \\ X^i &= x^i e^{Ht}, \\ X^D &= \frac{1}{H}\cosh(Ht) - \frac{1}{2}He^{Ht}\delta_{ij}x^i x^j, \end{aligned} \tag{15.2}$$

then the line element induced on the hypersurface takes the form

$$ds^2 = \eta_{AB}dX^A(x)dX^B(x) = -dt^2 + e^{2Ht}\delta_{ij}dx^i dx^j, \tag{15.3}$$

so that the hypersurfaces Σ_t of constant t are flat. These coordinates do not cover the whole de Sitter spacetime (they cover the Poincaré patch), but it is not difficult to introduce coordinates that do and which, moreover, are useful to construct a conformal diagram for de Sitter spacetime. To see this we first rewrite the previous metric using the conformal time $ad\eta = dt$, where a is the scale factor $a = \exp(Ht)$, and spherical coordinates,

$$ds^2 = \frac{1}{H^2\eta^2}\left(-d\eta^2 + dr^2 + r^2\left(d\theta^2 + \sin^2\theta d\varphi^2\right)\right), \tag{15.4}$$

where $-\infty < \eta < 0$, $0 \leq r < \infty$, $0 \leq \theta < \pi$ and $0 \leq \varphi < 2\pi$. We may then introduce new time and space coordinates $(\tilde{\eta}, \chi)$ which span a finite range as [745],

$$\eta = \frac{\sin\tilde{\eta}}{\cos\tilde{\eta} + \cos\chi}, \quad r = \frac{\sin\chi}{\cos\tilde{\eta} + \cos\chi}. \tag{15.5}$$

where $-\pi < \tilde{\eta} < 0$ and $0 \leq \chi < \pi$, and the metric now reads

$$ds^2 = \frac{1}{H^2\sin^2\tilde{\eta}}\left(-d\tilde{\eta}^2 + d\chi^2 + \sin^2\chi\left(d\theta^2 + \sin^2\theta d\varphi^2\right)\right). \tag{15.6}$$

the section of $\tilde{\eta} = \text{const}$ are three spheres. These coordinates cover the entire de Sitter spacetime. They represent a closed universe which contracts for $-\pi < \tilde{\eta} < -\pi/2$ and expands for $-\pi/2 < \tilde{\eta} < 0$. The de Sitter conformal diagram, for the conformal metric $(H^2\sin^2\tilde{\eta})\, ds^2$, may be represented in terms of these coordinates $(\chi, \tilde{\eta})$, each point of the conformal diagram represents a S^2 sphere of radius $\sin\chi$.

The metric of de Sitter spacetime is a solution of Einstein equation in vacuum with a cosmological constant

$$G_{ab} + \Lambda g_{ab} = 0, \tag{15.7}$$

where the cosmological constant is $\Lambda = [(D-1)(D-2)/2]\, H^2$. The metric can be interpreted as a cosmological FLRW metric for a perfect fluid whose pressure equals the constant energy density (vacuum energy). Because the energy density is constant any hypersurface is an hypersurface of constant energy density.

15.1.1 De Sitter Invariant Bitensors

To introduce the formalism of the de Sitter invariant bitensors it is convenient to define the basic biscalar

$$Z(x,x') = H^2 \eta_{AB} X^A(x) X^B(x'), \tag{15.8}$$

where $x, x' \in dS_D$. Note that Z is invariant under the de Sitter group; if σ is an isometry, then $Z(\sigma(x), \sigma(x')) = Z(x,x')$ and under the antipodal transformation $Z(\alpha(x), x') = -Z(x,x')$. This biscalar is related to the Minkowskian distance, defined by

$$d^2(x,x') = \eta_{AB}[X^A(x) - X^A(x')][X^B(x) - X^B(x')], \tag{15.9}$$

which is the length of the straight line in the embedding $(D+1)$-Minkowski spacetime that joins x and x'. It follows from the two previous equations that Z and d are related through

$$d^2(x,x') = 2H^{-2}(1 - Z(x,x')). \tag{15.10}$$

Therefore, $Z = 1$ when x' is on the light cone of x, $Z > 1$ when the points are timelike separated, and $Z < 1$ when the points are spacelike separated. It should also be noted that, as the squared Minkowskian distance between two points on the hyperboloid can take any value, $Z \in (-\infty, \infty)$.

Now, without loss of generality we can take the point x in the conformal diagram $(\chi = 0, \tilde{\eta} = -\pi/2)$ and consider the spacelike straight curve with $\tilde{\eta} =$ const connecting x and its antipodal $\alpha(x)$ which in that diagram is given by $(\chi = \pi, \tilde{\eta} = -\pi/2)$. This curve is a geodesic, because it is a straight line on a cylinder tangent to the hyperboloid. The Minkowskian distance between x and any other point x' in this geodesic satisfies $0 \leq d(x,x') \leq 2$. This implies

$$-1 \leq Z(x,x') \leq 1. \tag{15.11}$$

Any other spacelike geodesic passing through x can be obtained from this one by an isometry that leaves x invariant (it also leaves $\alpha(x)$ invariant). This is analogous to the case of the sphere, where any geodesic (meridian) passing through the north pole is obtained from a given one by a rotation that leaves the north pole (and south pole) invariant. This means that any point y' connected to x by a spacelike geodesic satisfies $y' = \sigma(x')$ for some x' in the above geodesic, and where σ is an isometry such that $\sigma(x) = x$. De Sitter invariance of Z then implies $Z(x,y') = Z(\sigma(x), \sigma(x')) = Z(x,x')$. Consequently, Eq. (15.11) holds for any point x' connected with x by a spacelike geodesic. In other words, spacelike separated points x, x' such that $Z(x,x') < -1$ cannot be joined by a geodesic.

17

Riemann Tensor Correlator in de Sitter

The Riemann tensor correlation function exhibits a number of interesting features. First, with appropriately raised indices it is gauge-invariant even when including contributions from loops of matter fields but excluding graviton loops, which as we have seen before can be implemented systematically within a large N expansion for N matter fields. Second, it is compatible with de Sitter invariance, in contrast with other gauge-invariant quantities more commonly used in cosmology. In fact, whether de Sitter invariance is preserved can be manifestly seen by expressing the result in terms of maximally symmetric bitensors. Thirdly, the Riemann tensor provides a complete and suitable characterization of the local geometry: geometrical observables associated with a finite region of spacetime can be obtained from it. Thus, they are insensitive to potentially infrared divergent superhorizon contributions which cannot be probed by observables involving finite spacetime regions.

In this chapter the linearized Riemann tensor correlator on a de Sitter background including one-loop corrections from conformal fields is derived, following Fröb et al. [572]. In Section 17.1 we revisit the formalism described in Section 15.1 of de Sitter-invariant bitensors, and introduce an alternative set of bitensors which are useful for this computation. The main result of the chapter is obtained in Section 17.2: by taking suitable derivatives of the metric correlator function found in Section 16.5, the connected two-point correlator function of the Riemann tensor $R^{ab}{}_{cd}$ is computed. The result is written in a manifestly de Sitter-invariant form in terms of the bitensors introduced by Allen and Jacobson. Moreover, given the decomposition of the Riemann tensor in terms of Weyl and Ricci tensors, we write the explicit results for the Weyl and Ricci tensors correlators as well as the Weyl–Ricci tensors correlator and study both their subhorizon and superhorizon behavior. These results are extended to general conformal field theories in Section 17.3.

We also derive the Riemann tensor correlator in Minkowski spacetime in a manifestly Lorentz invariant form in Section 17.4, by carefully taking the flat-space limit of our result in de Sitter. To the best of our knowledge, this result for the Riemann tensor correlator in Minkowski spacetime including loop corrections from general conformal fields had not been obtained previously to the calculation in de Sitter by Fröb et al. [572]. There were results in the literature for the metric correlator by Tomboulis [686] and Martín and Verdaguer [189], and the correlator of the Einstein tensor by Osborn and Shore [204], which is closely related to the stress-energy tensor correlator, but not for the full Riemann tensor correlator.

17.1 De Sitter-Invariant Bitensors

We have seen in Section 15.1 that any de Sitter-invariant two-point correlation function can be expressed in terms of maximally symmetric bitensors. Here we adapt the formalism to our problem. The basic de Sitter-invariant biscalar $Z(x,x')$ was defined in (15.8) The geodesic distance $\mu(x,x')$ along the shortest geodesic connecting the points x and x' is related to the biscalar $Z(x,x')$ by Eq. (15.18), which we now write re-introducing the Hubble parameter H as

$$Z(x,x') = \cos\left(H\mu(x,x')\right). \tag{17.1}$$

For points such that $Z(x,x') < -1$, i.e., points not connected by a geodesic, μ can still be defined by a suitable analytic continuation from (17.1).

As we have seen in Section 15.1 a theorem by Allen and Jacobson [476] establishes that there exists a complete set of de Sitter-invariant bitensors from which all maximally symmetric bitensors can be constructed. This set is given by the metric $\hat{g}_{ab}(x)$, the geodesic distance $\mu(x,x')$, the unit tangent vectors to the geodesic at the points x and x', $n_a(x,x')$ and $n_{a'}(x,x')$, defined by Eq. (15.14), and the propagator $\hat{g}_{ab'}(x,x')$ which parallely transports a vector from x' to x along the geodesic connecting them. Here as in Chapter 16 we use a hat over a symbol to refer to quantities which are defined on the background de Sitter space, see Eq. (16.18).

By inverting Eq. (17.1), we can write the complete set of bitensors described above in terms of covariant derivatives of $Z(x,x')$ as follows:

$$\begin{gathered} n_a = -\frac{1}{H\sqrt{1-Z^2}}\hat{\nabla}_a Z, \qquad n_{a'} = -\frac{1}{H\sqrt{1-Z^2}}\hat{\nabla}_{a'} Z, \\ H^2 \hat{g}_{ab'} = \hat{\nabla}_a \hat{\nabla}_{b'} Z - \frac{1}{1+Z}(\hat{\nabla}_a Z)(\hat{\nabla}_{b'} Z). \end{gathered} \tag{17.2}$$

In addition, as an illustration of the above theorem, see Eq. (15.17), we have that

$$\hat{\nabla}_a n_b = \frac{Z}{\sqrt{1-Z^2}} H\left(\hat{g}_{ab} - n_a n_b\right), \tag{17.3}$$

2)*+,)))-) 3 , 8 " 7 5 1 "
. / , @ " 8 " " 9 1 5 =: 2 " E 1 5 ."
-2 ; 1 " <<< (= . , " A A .

1

Quantum fields in Minkowski spacetime

The theory of quantum fields in curved spacetime is a generalization of the well-established theory of quantum fields in Minkowski spacetime. To a great extent, the behavior of quantum fields in curved spacetime is a direct consequence of the corresponding flat spacetime theory. Local entities, such as the field equations and commutation relations, are to a large extent determined by the principle of general covariance and the principle of equivalence. However, global entities which are unique in Minkowski spacetime lose that uniqueness in curved spacetime. For example, the vacuum state, which in Minkowski spacetime is determined by Poincaré invariance, is not unambiguously determined in curved spacetime. This ambiguity is closely tied to the phenomenon of particle creation by certain gravitational fields, as in the expanding universe or near a black hole.

It is logical, therefore, to review the relevant aspects of flat spacetime quantum field theory. This will serve to establish the necessary background, to fix our notation, and to highlight those aspects of the theory which can be carried over to curved spacetime, as well as those which lose their meaning in curved spacetime. We will often be brief, emphasizing concepts while omitting many derivations, and only touch on particular topics. Our discussion of the curved spacetime theory in later chapters will be more detailed.

In this initial chapter, we discuss the canonical formulation, including the Schwinger action principle and the relation between symmetry transformations and conserved currents (Schwinger 1951b, 1953). We review the dynamical descriptions known as the Heisenberg picture, the Schrödinger picture, and the interaction picture. We introduce the Fock representations, in which states are described in terms of their particle content, and the Schrödinger representation, in which the states are described by field configurations. We include discussions of the Maxwell and Yang–Mills gauge fields, as well as the Dirac field, and the definitions of spin and angular momentum.

1.1 Canonical formulation

Recall that in classical mechanics the equations of motion of a particle or system of particles having independent generalized coordinates $q_i(t)$ and velocities $\dot{q}_i(t)$ are given by the principle of stationary action. This principle states that the action

$$S = \int_{t_1}^{t_2} dt\, L(q, \dot{q}) \tag{1.1}$$

is stationary under arbitrary variations of the q_i which vanish on the boundary of the region of integration. Here L is the Lagrangian of the system. The Hamiltonian is defined by

$$H(q,p) = \sum_i p_i \dot{q}_i - L,$$

where

$$p_i = \frac{\partial L}{\partial \dot{q}_i} \tag{1.2}$$

is the momentum conjugate to q_i. The system is quantized by taking the qs and ps to be Hermitian operators acting on a Hilbert space, and by imposing the canonical commutation relations

$$[q_i, q_j] = 0, \quad [p_i, p_j] = 0,$$

$$[q_i, p_j] = i\delta_{i,j}. \tag{1.3}$$

Here $\delta_{i,j}$ is the Kronecker delta. We are using units with $\hbar = c = 1$. From (1.3) it follows that if $F(q,p)$ is a function of the coordinate and momentum operators, then (assuming F can be Taylor expanded in p)

$$[q_i, F] = i\frac{\partial F}{\partial p_i}. \tag{1.4}$$

The above commutation relations imply that the q_i are a complete set of commuting observables with continuous spectra consisting (in the absence of impenetrable walls) of all real numbers. The same can be said for the p_i. (An observable is an Hermitian operator with a complete set of eigenstates.) The eigenstates[1] of the q_i are the kets $|q'>$, where q' denotes the set of eigenvalues q'_i of the operators q_i. Thus,

$$q_i|q'> = q'_i|q'>,$$

with the normalization

$$<q'|q''> = \delta(q' - q''),$$

---

[1] We use the conventions of Dirac (1958) in distinguishing the eigenvalue of an operator from the operator by using a $'$.

where $\delta(q' - q'')$ is the Dirac δ-function. It also follows that

$$< q'|p_i|q'' >= -i\frac{\partial\delta(q' - q'')}{\partial q'_i},$$

and that

$$< q'|p' >= (2\pi)^{-n/2}\exp\left(i\sum_{i=1}^{n} p'_i q'_i\right),$$

where $|p' >$ is an eigenket of the p_i having delta function normalization, and we have taken the index i to run from 1 to n. (For derivations, see Messiah (1961, pp. 302–309).) If $F(q,p)$ is a function of the q_i and p_i with any well-defined ordering of factors, then

$$< q'|F(q,p)|q'' >= F\left(q', -i\frac{\partial}{\partial q'}\right)\delta(q' - q''), \tag{1.5}$$

where $F(q', -i(\partial/\partial q'))$ is the same ordered function with $-i(\partial/\partial q'_i)$ replacing p_i in each position.

In the Schrödinger or configuration space representation, the abstract operators are represented by matrix elements based on the $|q' >$, such as that of p_i above, and the states are represented by functions. For example, a state $|\psi >$ is represented by the Schrödinger wave function $\psi(q') =< q'|\psi >$. An example is the wave function $< q'|p' >$ above, representing a particle of definite momentum. Similarly, in the momentum-space representation the operators are represented by matrix elements formed from the $|p' >$ and the states are represented by functions such as $< p'|\psi >$.

Up to now the description has been purely kinematical, with time playing no role. The dynamical evolution of the system is governed by the Hamiltonian $H(q,p,t)$. We have allowed for the possibility that H may have explicit time dependence, as through an interaction with an external field. The time evolution may be described in several physically equivalent ways, known as “pictures.”

In the Schrödinger picture, the fundamental observables q and p do not change with time. Rather, the dynamical evolution of measurable quantities, such as expectation values of observables, is expressed through the time dependence of the ket describing the state of the system at each time. The fundamental dynamical equation is that of Schrödinger,

$$i\frac{d}{dt}|\psi(t) >= H(q,p,t)|\psi(t) >. \tag{1.6}$$

Because H may have explicitly time-dependent terms involving p and q, in general $H(t')$ and $H(t'')$ may not commute. (For brevity, we suppress the dependence of H on q and p.)

We note in passing that in the Schrödinger representation, the Schrödinger equation (1.6) becomes (using the completeness of the $|q' >$)

$$i < q'|\frac{d}{dt}|\psi(t) >= \int < q'|H(q,p,t)|q'' > dq'' < q''|\psi(t) >$$

or

$$i\frac{\partial}{\partial t}\psi(q',t) = H\left(q', -i\frac{\partial}{\partial q'}, t\right)\psi(q',t), \tag{1.7}$$

where $\psi(q',t) =< q'|\psi(t) >$.

The solution of (1.6) is

$$|\psi(t) >= U(t,t_0)|\psi(t_0) >,$$

with $U(t,t_0)$ satisfying

$$i\frac{d}{dt}U(t,t_0) = H(t)U(t,t_0), \tag{1.8}$$

with the boundary condition

$$U(t_0,t_0) = 1.$$

The evolution operator $U(t,t_0)$ preserves the norm of the state vector and is thus unitary, satisfying

$$UU^\dagger = U^\dagger U.$$

From $U(t_0,t)U(t,t_0) = U(t_0,t_0) = 1$, it then follows that $U(t,t_0)^\dagger = U(t_0,t)$.

In the Heisenberg picture, the ket describing the state of the system does not change with time, while the dynamical evolution of the system is expressed through the time dependence of the fundamental observables $q(t)$ and $p(t)$. By applying $U(t,t_0)^\dagger$ to the Schrödinger picture ket describing the state of the system, we obtain a time-independent ket, which can be taken as the ket describing the state of the system in the Heisenberg picture. Thus, denoting quantities in the Schrödinger picture by subscript S and those in the Heisenberg picture by subscript H, we have

$$|\psi_H >= U(t,t_0)^\dagger|\psi_S(t) >= |\psi_S(t_0) >. \tag{1.9}$$

In order that measurable expectation values remain the same as in the Schrödinger picture, the Heisenberg picture operator $F_H(t)$ corresponding to a Schrödinger picture operator F_S must be related by

$$F_H(t) = U(t,t_0)^\dagger F_S U(t,t_0). \tag{1.10}$$

Note that the Hamiltonian H in (1.8) is H_S. When H_S has no explicit time dependence, the solution of (1.8) is

$$U(t,t_0) = \exp[-i(t-t_0)H],$$

and it follows that H commutes with U so that $H_H = H_S$. When H_S does have an explicit time dependence, $H_H \neq H_S$, and we must use (1.10) to define H_H in terms of H_S. In general, F_H will depend on time through $q_H(t)$, $p_H(t)$, and through further explicit appearance of t. Denoting by $\partial F_H/\partial t$ the derivative only with respect to this further explicit appearance of t, it follows from (1.8) and (1.10) that

$$i\frac{d}{dt}F_H = [F_H,\, H_H] + i\frac{\partial}{\partial t}F_H. \tag{1.11}$$

This resembles the classical equation of motion with the Poisson bracket replaced by $-i$ times the commutator. Of course, we do not need to define the Schrödinger picture before the Heisenberg picture. The latter is fully defined by stating that the ket describing the state of the system is independent of time and that operators F, constructed from the q_i, p_i (dropping subscript H), and t obey the Heisenberg equation of motion (1.11).

When two systems interact through a term in the Hamiltonian, which can be regarded as a perturbative interaction, it is useful to introduce another picture of the dynamical evolution, known as the "interaction picture." In this picture, the ket describing the state of the system evolves as in the Schrödinger picture, but only under the influence of the interaction term in the Hamiltonian, while operators evolve as in the Heisenberg representation, but only under the influence of the unperturbed term in the Hamiltonian.

Let us now turn to the canonical quantization of a system of independent real fields $\phi_a(x)$, where x refers to the Minkowski space and time coordinates x^μ, and the index a includes tensor or spinor indices and internal quantum numbers of the field multiplet. We will deal here with bosons and discuss later the modification of canonical quantization required with fermions. For brevity, the index a will often be suppressed; we can usually think of ϕ as a column or row matrix, depending on where it appears in an expression. Canonical quantization proceeds as in the previously discussed quantization of a particle. One thinks of the classical field $\phi(x)$ as analogous to the classical $q_i(t)$, with the spatial coordinates $\vec{x}$ regarded as labels like i. Because we are now dealing with a continuous label, Dirac δ-functions involving $\vec{x}$ will appear where Kronecker deltas involving the label i previously appeared. As before, we assume that the system is described by an action

$$S = \int_{t_1}^{t_2} dt\, L[\phi, \partial\phi]_t, \tag{1.12}$$

where the Lagrangian L is now a functional of the field ϕ and its first derivatives $\partial\phi/\partial x^\mu \equiv \partial_\mu\phi$, which are denoted collectively by $\partial\phi$. The subscript t indicates that L is a function of t. The Lagrangian L can be expressed in terms of a Lagrangian density $\mathcal{L}$ as

$$L[\phi, \partial\phi]_t = \int dV_x\, \mathcal{L}\big(\phi(\vec{x}, t), \partial\phi(\vec{x}, t)\big),$$

where dV_x is the spatial volume element. Because L is now a functional, the momentum π conjugate to the field $\phi(\vec{x},t)$ is defined in analogy with (1.2) through the following functional derivative (regarding $\partial_0\phi$ as independent of ϕ at time t),

$$\begin{aligned}\pi(\vec{x},t) &= \delta L[\phi,\partial\phi]_t/\delta\big(\partial_0\phi(\vec{x},t)\big) \\ &= \partial\mathcal{L}\big(\phi(\vec{x},t),\partial\phi(\vec{x},t)\big)/\partial\big(\partial_0\phi(\vec{x},t)\big). \end{aligned} \tag{1.13}$$

Here, we have used the definition of the functional derivative, which states that if $F[\chi]$ is a functional of $\chi(\vec{x})$, then under a variation $\delta\chi(\vec{x})$ of χ, which vanishes sufficiently fast at spatial infinity, we have

$$\delta F[\chi] = \int dV_x\, \frac{\delta}{\delta\chi(\vec{x})} F[\chi]\delta\chi(\vec{x}). \tag{1.14}$$

It follows that if the functional F has the form

$$F[\chi] = \int dV_x\, f\big(\chi(\vec{x}),\vec{\partial}\chi(\vec{x})\big),$$

then

$$\frac{\delta}{\delta\chi(\vec{x})}F[\chi] = \frac{\partial}{\partial\chi}f\big(\chi,\vec{\partial}\chi\big) - \partial_i\left(\frac{\partial}{\partial(\partial_i\chi)}f\big(\chi,\vec{\partial}\chi\big)\right),$$

where χ is evaluated at $\vec{x}$. The result in (1.13) follows from this with $\chi \to \partial_0\phi$. Another consequence is that

$$\frac{\delta\chi(\vec{x}')}{\delta\chi(\vec{x})} = \delta(\vec{x}'-\vec{x}).$$

One can regard the action in (1.12) as a functional depending on the space and time dependence of ϕ. Then, the Euler–Lagrange field equation below will be recognized as another application of the above result, but for a functional depending on one more dimension.

The Hamiltonian is defined by

$$H[\phi,\pi]_t = \int dV_x\, \pi_a(\vec{x},t)\partial_0\phi_a(\vec{x},t) - L[\phi,\partial\phi]_t. \tag{1.15}$$

(Although we write $H[\phi,\pi]$, dependence on spatial derivatives of ϕ (or π) is permitted.) The principle of stationary action yields upon variation of the fields in (1.12), the Euler–Lagrange field equations

$$\partial_\mu\left(\frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\right) - \frac{\partial\mathcal{L}}{\partial\phi} = 0, \tag{1.16}$$

where the repeated spacetime coordinate index μ is summed over its full range of values, in accordance with the Einstein summation convention.

The field is quantized in analogy with the canonical commutators of (1.3). Thus, we postulate that

$$[\phi_a(\vec{x},t),\phi_b(\vec{x}',t)]=0,\quad [\pi_a(\vec{x},t),\pi_b(\vec{x}',t)]=0,$$

$$[\phi_a(\vec{x},t),\pi_b(\vec{x}',t)]=i\delta_{a,b}\delta(\vec{x}-\vec{x}'), \tag{1.17}$$

where, as noted earlier, $\delta(\vec{x}-\vec{x}')$ is the Dirac δ-function. (In a theory of interacting fields, if we deal directly with the renormalized fields, then (1.17) is somewhat altered by a normalization factor. We are thus dealing here with the bare fields, which by definition satisfy the field equations with unrenormalized masses and coupling constants.) The t dependence is included in the commutation relations to emphasize that in a dynamical picture like the Heisenberg picture, in which the operators ϕ and π depend on time (as do the classical fields), the canonical commutation relations must be imposed on the fields and conjugate momenta evaluated at the same time. For it follows from (1.10) that $q_H(t)$ and $p_H(t)$ evaluated at the same time do satisfy (1.3), while that is not true in general if they are evaluated at different times. Of course, in the Schrödinger picture the fields and conjugate momenta have no time dependence, and t would not appear in (1.17).

The functional analogue of (1.4) follows from the above commutators:

$$[\phi_a(\vec{x}),F[\phi,\pi]\,]=i\frac{\delta}{\delta\pi_a(\vec{x})}F[\phi,\pi], \tag{1.18}$$

where we are suppressing the time dependence, taking all fields to be at a single time t. One can now set up a Schrödinger or field representation using the eigenstates of $\phi(\vec{x})$ defined by

$$\phi(\vec{x})|\phi'>=\phi'(\vec{x})|\phi'>. \tag{1.19}$$

The ket $|\phi'>$ corresponds to a state of the system in which the field has configuration $\phi'(\vec{x})$, where ϕ' is an ordinary or c-number function, unlike the field operator ϕ. Thus, we are using the analogue of the Dirac notation in which "eigenvalues" of the operator $\phi(\vec{x})$ are functions denoted by $\phi'(\vec{x})$. (Here the prime does not denote derivative, but instead distinguishes a c-number from an operator.) In ordinary quantum mechanics, it follows from (1.5) that if $|\psi>$ is an element of the Hilbert space spanned by the eigenkets $|q'>$, then

$$<q'|F(q,p)|\psi>=F\left(q',-i\frac{\partial}{\partial q'}\right)<q'|\psi>.$$

Similarly, we can show that if $|\Psi>$ is a state in the space spanned by the eigenstates $|\phi'>$, and $F[\phi,\pi]$ is a functional formed from the field operator and conjugate momentum, then

$$<\phi'|F[\phi,\pi]|\Psi>=F\left[\phi',-i\frac{\delta}{\delta\phi'}\right]<\phi'|\Psi>. \tag{1.20}$$

Here $<\phi'|\Psi>\equiv\Psi[\phi']$ is a complex number which is a functional of ϕ'. It is interpreted as the probability amplitude for finding the field observable ϕ to have the configuration or set of values given by $\phi'(\vec{x})$ when the system is in the state described by the vector $|\Psi>$.

If we work in the Schrödinger picture, then (1.20) can be used to turn the Schrödinger equation

$$i\frac{d}{dt}|\Psi(t)>= H[\phi,\pi]|\Psi(t)>$$

into the functional differential equation

$$i\frac{\partial}{\partial t}\Psi[\phi',t] = H\left[\phi', -i\frac{\delta}{\delta\phi'}\right]\Psi[\phi',t], \tag{1.21}$$

where $\Psi[\phi',t] \equiv <\phi'|\Psi(t)>$, and ϕ, π depend on $\vec{x}$, but not on t.

On the other hand, in the Heisenberg picture the state describing the evolving system is independent of time, while a general functional F of $\phi(\vec{x},t)$ and $\pi(\vec{x},t)$ will depend on time through its dependence on ϕ and π, as well as through a possible further explicit dependence on t. Then, the Heisenberg equation of motion is

$$i\frac{d}{dt}F[\phi,\pi;t] = \big[F[\phi,\pi;t],\, H[\phi,\pi]\big] + i\frac{\partial}{\partial t}F[\phi,\pi;t]. \tag{1.22}$$

The consistency of (1.22) with the Euler–Lagrange equations (1.16) can be proved in Minkowski spacetime and in the more general curved spacetime context (Parker, 1973, see Appendix B).

Finally, we expect that when there are no time- or space-dependent external parameters in the Lagrangian, then there should exist a conserved vector observable P^μ corresponding to the total energy and momentum of the system. For such a Lagrangian, by multiplying the Euler–Lagrange equations by $\partial_\mu\phi$ and using[2]

$$\partial_\mu\mathcal{L} = \frac{\partial\mathcal{L}}{\partial(\partial_\nu\phi)}\partial_\mu\partial_\nu\phi + \frac{\partial\mathcal{L}}{\partial\phi}\partial_\mu\phi,$$

we find immediately that

$$\partial_\mu T^\mu{}_\nu = 0, \tag{1.23}$$

where

$$T^\mu{}_\nu = \frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi_a)}\partial_\nu\phi_a - \delta^\mu{}_\nu\mathcal{L}. \tag{1.24}$$

Here, summation over internal indices a of ϕ and, when the fields are treated as operators, symmetrization over fields and their conjugate momenta are understood. The tensor $T_{\mu\nu}$ is called the canonical energy-momentum or stress tensor.

We mention in passing that in order to serve as the source in the Einstein gravitational field equations, $T_{\mu\nu}$ should be symmetric under interchange of indices.

---

[2] This assumes that the Lagrangian does not have any explicit dependence on the coordinates x^μ.

(This symmetry is also required if we wish to define a conserved angular momentum in terms of the energy-momentum tensor.) However, except in the case of particular forms of $\mathcal{L}$, the expression given in (1.24) (after lowering the index ν with the Minkowski metric $\eta_{\mu\nu}$) is not a symmetric tensor. A general manifestly symmetric expression for $T_{\mu\nu}$ will be given later, when we deal with curved spacetime. (Since both of these expressions for $T^{\mu}{}_{\nu}$ will satisfy (1.23) for *any* system with no explicit time or space dependence, we expect on physical grounds that they must each yield the same conserved energy and momentum P^{μ} to within a constant.) Furthermore, a modification of the canonical $T_{\mu\nu}$ that makes it symmetric, and yields the same P_{μ} and angular momentum as the original canonical $T_{\mu\nu}$, has been given by Belinfante (1939, 1940).

From (1.23) we have

$$\int dv_x \, \partial_{\mu} T^{\mu}{}_{\nu} = 0,$$

where dv_x denotes the *spacetime* volume element, and the integration is over a spacetime volume bounded by spatial infinity and any two constant-time hypersurfaces. Assuming that matrix elements of physical interest will be between states in which the physical field configuration is of finite spatial extent, we obtain by the Gauss divergence theorem the conservation law

$$\frac{d}{dt} \int dV_x \, T^{0}{}_{\nu} = 0,$$

where dV_x denotes the *spatial* volume element and the integration is over any constant-time hypersurface. Hence,

$$P_{\nu} = \int dV_x T^{0}{}_{\nu} \tag{1.25}$$

is the conserved energy-momentum vector. The sign in this definition is chosen so that $P_0 = H$, as can be verified by comparing (1.15) with (1.24).

As a special case of the Heisenberg field equation (1.22), suppose that the functional F is an ordinary function $f(\phi(x), \partial_i \phi(x), \pi(x))$ with no explicit time dependence. Then, (1.22) can be written as

$$i\partial_0 f = [f, P_0], \tag{1.26}$$

where the partial derivative symbol is used here in the conventional manner to indicate that the x^i are held fixed. In (1.22), the partial derivative symbol denoted derivation only with respect to explicit t dependence not coming from ϕ and π. The partial derivative in (1.26) includes *all* t dependence. The result in (1.26) is the 0-component of the more general relation

$$i\partial_{\mu} f = [f, P_{\mu}], \tag{1.27}$$

The $\mu = 0$ component, as noted above, follows from (1.22). For $\mu = i$, we easily verify (1.27) for powers of ϕ and π and thus for functions which can be expanded in power series in ϕ and π. For example, suppressing the t-dependence, which is the same in all arguments, we have

$$\begin{aligned}[\pi(\vec{x})^n, P_i] &= \int dV'_x\,\pi(\vec{x}')[\pi(\vec{x})^n, \partial'_i\phi(\vec{x}')] \\ &= -i\int dV'_x\,\pi(\vec{x}')n\pi(\vec{x})^{n-1}\partial'_i\delta(\vec{x}-\vec{x}') \\ &= i\partial_i(\pi(\vec{x})^n).\end{aligned}$$

Equation (1.27) also follows from a powerful generalization of the action principle and of Noether's theorem, known as the Schwinger operator action principle (Schwinger 1951b, 1953). (For textbook treatments see Roman (1969) or Toms (2007).) The action of (1.12) is integrated over a spacetime volume v bounded by two constant-time hypersurfaces at t_1 and t_2. (As originally stated, the principle deals with arbitrary spacelike hypersurfaces, but we work with constant-time hypersurfaces for simplicity at this stage. See Section 6.2 for a discussion.) Consider arbitrary infinitesimal variations, δx^μ and $\delta_0\phi(x)$, of the coordinates and field operators,

$$x^\mu \to x'^\mu = x^\mu + \delta x^\mu, \tag{1.28}$$

$$\phi(x) \to \phi'(x) = \phi(x) + \delta_0\phi(x), \tag{1.29}$$

where $\delta_0\phi(x)$ vanishes on the spatial boundary of integration at each time (i.e., it vanishes everywhere on the boundary of v, except on the interior of the constant-time hypersurfaces that bound v at t_1 and t_2). Then, the Schwinger action principle states that the variation of the action S of (1.12) has the form

$$\delta S = G(t_2) - G(t_1), \tag{1.30}$$

where the operator $G(t)$ is the generator of the above variation of the coordinates and fields at time t.

To say that G generates the variation means the following. For an operator functional $F[\phi, \pi]$, we have

$$i\delta_0 F = [F, G], \tag{1.31}$$

where all quantities are evaluated at the same time, and $\delta_0 F$ is the infinitesimal variation of F produced by (1.28) and (1.29). That is,

$$\delta_0 F = F[\phi + \delta_0\phi, \partial(\phi + \delta_0\phi)] - F[\phi, \partial\phi].$$

One can show from (1.30) that the generator has the form

$$G(t) = \int dV_x\left[\pi_a\delta\phi_a - T^0{}_\nu\delta x^\nu\right], \tag{1.32}$$

where $T^\mu{}_\nu$ is the energy-momentum tensor of (1.24), and

$$\delta\phi(x) \equiv \phi'(x') - \phi(x) \tag{1.33}$$

is the change of the field at a given physical point in spacetime (i.e., the local variation).

The derivation is as follows. The variation δS above is defined by

$$\delta S = S' - S,$$

where

$$S' = \int dv'_x \, \mathcal{L}(\phi'(x'), \partial' \phi'(x'))$$

and

$$S = \int dv_x \, \mathcal{L}(\phi(x), \partial\phi(x)).$$

Here, both integrals are taken over the same physical volume v, with x and x' denoting the same physical point of the system in spacetime, and thus related by (1.28). This way of viewing the transformation is the so-called passive viewpoint, in which the coordinates undergo a transformation from x to x', but the physical system does not change.

Alternatively, it is possible to adopt the active viewpoint taken by Messiah (1961), in which x and x' are regarded as describing a change of the physical system relative to a single coordinate system, such that the event at x is dragged to x'. In the active viewpoint, the volume v is infinitesimally displaced to a new physical volume v', related to v by (1.28). We can write

$$\begin{aligned} S' &= \int_{v'} dv_x \, \mathcal{L}(\phi'(x), \partial\phi'(x)) \\ &= \int_{v'} dv_x \, \mathcal{L}(\phi(x) + \delta_0\phi(x), \partial\phi(x) + \partial\delta_0\phi(x)) \\ &= \int_{v'} dv_x \mathcal{L}(\phi, \partial\phi) + \int_v dv_x \left\{ \frac{\partial \mathcal{L}}{\partial \phi} \delta_0 \phi + \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \partial_\mu \delta_0 \phi \right\}. \end{aligned}$$

The volume of integration in the second integral has been changed to v because the integrand is already of first order.

Corresponding events on the boundaries of v and v' are related by the infinitesimal displacement δx^μ. Let $d\sigma^\mu$ denote $n^\mu d\sigma$, where n is an outward normal to the boundary of v at a point x, and $d\sigma$ is an element of surface area (i.e., hypersurface volume) on the boundary of v at x. Then, the scalar product $d\sigma_\mu \delta x^\mu$ is the volume of the cylindrical four-volume element capped at x by the surface area $d\sigma$ and extending along δx^μ from the boundary of v to that of v'. Then, denoting the boundary of v by ∂v, the surface integral,

$$\int_{\partial v} d\sigma_\mu \delta x^\mu \mathcal{L}(\phi, \partial\phi),$$

is equal to the following difference of volume integrals over v and v':

$$\int_{v'} dv_x \mathcal{L}(\phi, \partial\phi) - \int_v dv_x \mathcal{L}(\phi, \partial\phi),$$

because the difference is just the integral of $\mathcal{L}$ over the infinitesimal volume lying between the boundaries of v and v'.

Hence,

$$\begin{aligned}\delta S &= S' - S \\ &= \int_{\partial v} d\sigma_\mu \delta x^\mu \mathcal{L}(\phi, \partial\phi) + \int_v dv_x \left\{ \frac{\partial \mathcal{L}}{\partial \phi} \delta_0 \phi + \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \partial_\mu \delta_0 \phi \right\}.\end{aligned}$$

Converting the surface integral to a volume integral by the Gauss divergence theorem, and using the identity,

$$\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \partial_\mu \delta_0 \phi = \partial_\mu \left\{ \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \delta_0 \phi \right\} - \partial_\mu \left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \right) \delta_0 \phi$$

in the second term, we are left with the result

$$\delta S = \int_v dv_x \, \partial_\mu \left\{ \delta x^\mu \mathcal{L} + \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \delta_0 \phi \right\}.$$

Here, the Euler–Lagrange equations (1.16) have been used.[3] It is convenient to express our result in terms of the local variation $\delta\phi$ defined in (1.33). One has

$$\phi'(x) = \phi'(x' - \delta x) = \phi'(x') - \partial_\mu \phi(x) \delta x^\mu$$

to first order. Consequently, (1.29) gives

$$\delta_0 \phi = \delta \phi - (\partial_\mu \phi) \delta x^\mu.$$

Hence,

$$\delta S = \int_v dv_x \, \partial_\mu \left\{ \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \delta \phi - T^\mu{}_\nu \delta x^\nu \right\}, \tag{1.34}$$

where $T^\mu{}_\nu$ is the energy-momentum tensor of (1.24). It follows that $\delta S = G(t_2) - G(t_1)$, with $G(t)$ given by (1.32), which completes the derivation.

Under an infinitesimal translation, we have $\phi'(x') = \phi(x)$, so that $\delta\phi$ vanishes. Therefore,

$$\delta_0 \phi = -(\partial_\mu \phi) \delta x^\mu.$$

Using the fact that for a translation δx^μ is constant, we now find that the generator is

$$G = -P_\mu \delta x^\mu,$$

where P_ν is the momentum operator of (1.25). One then has (1.31) for the commutator of G with an operator functional $F[\phi, \pi]$. Let us write (1.31) for the case when F is a function f formed from ϕ and the $\partial_\mu \phi$. Then,

$$\delta_0 f = \frac{\partial f}{\partial \phi} \delta_0 \phi + \frac{\partial f}{\partial(\partial_\mu \phi)} \partial_\mu \delta_0 \phi.$$

---

[3] The Euler–Lagrange equations follow, as usual, by demanding that δS is zero for the subset of variations such that $\delta_0 \phi$ vanishes on the boundary, ∂v.

Upon using the form of $\delta_0\phi$ for a translation, we find

$$\delta_0 f = -(\partial_\mu f)\delta x^\mu.$$

Using these results in (1.31), we obtain immediately (1.27).

The transformations of (1.28) and (1.29) are said to be a symmetry or invariance of the Lagrangian density $\mathcal{L}$ if

$$\mathcal{L}(\phi'(x'), \partial'\phi'(x')) = \mathcal{L}(\phi(x), \partial\phi(x)). \tag{1.35}$$

Recall that in S and S', the integrals are taken over the same *physical* volume (x and x' refer to the same physical point in spacetime). Therefore, the action is invariant when (1.35) holds, and we have $\delta S = 0$. In that case, (1.30) implies that G is a constant of the motion. Furthermore, (1.34) implies that

$$\partial_\mu J^\mu = 0, \tag{1.36}$$

with the conserved current

$$J^\mu = \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi_a)}\delta\phi_a - T^\mu{}_\nu \delta x^\nu. \tag{1.37}$$

It follows that when a translation in the x^μ direction is a symmetry of the Lagrangian density, then the operator P_μ is a constant of the motion. In Minkowski spacetime, the Lagrangian density of a free field is symmetric under space and time translations, which implies that energy and momentum are conserved. In curved spacetime, because of the presence in the Lagrangian density of the metric, we will find that for a free field (i.e., one influenced by gravitation alone), only the component of P_μ along the direction of an isometry of the spacetime is conserved.

1.2 Particles

The Schrödinger representation based on the eigenvectors of the field operator ϕ emphasizes the field aspect of the quantized field theory. Let us now turn to a representation which emphasizes the dual particle aspect of the quantum field. One expects that since the particle and wave properties of a system are known to be complementary, rather than simultaneously measurable, the observable corresponding to particle number, for example, will be constructed from both the field ϕ and its non-commuting conjugate momentum π. Furthermore, as the concept of free particle is well understood, we must first understand how to describe a system of free particles. Then, mutually interacting particles can be described through a perturbative description in which the free particles appear at early times and emerge again at late times. Therefore, we consider for now a single Hermitian free-scalar (or pseudo-scalar) field described by the Lagrangian density

$$\mathcal{L} = \frac{1}{2}(\eta^{\mu\nu}\partial_\mu\phi\partial_\nu\phi - m^2\phi^2). \tag{1.38}$$

2

Basics of quantum fields in curved spacetimes

The successful predictions of general relativity are convincing evidence that gravitational phenomena are most clearly understood by regarding spacetime as curved. In general relativity matter exerts its gravitational influence by curving spacetime, and we study the propagation of particles and waves on this curved background. It is then natural to study the propagation of quantum fields in curved spacetimes in order to search for new effects of gravitation. At this level, the gravitational field itself is not quantized, and the methods of Minkowski spacetime quantum field theory are carried over as much as possible. As we shall see, this modest extension of quantum field theory has turned out to be richer in consequences than we could have anticipated.

Among other things, it gives rise to the physically important processes of particle creation in cosmological and black hole spacetimes. The same amplification process that creates particles in an expanding universe is responsible for creating, in the context of an early inflationary expansion, the primordial fluctuations that are now observed with astonishing accuracy in the cosmic microwave background (CMB) radiation. These same primordial fluctuations also appear responsible for the large-scale structure of the universe. The creation of particles by black holes is necessary for maintaining the second law of thermodynamics in their presence. This process of radiation and evaporation of black holes is an important facet in the fundamental search for a microscopic explanation of the entropy of black holes; a search which appears to be leading to new and exciting physics connecting gravitation and quantum theory. Quantum field theory in curved spacetime also provides a new dynamical explanation of the connection between spin and statistics, and brings out some new features of Minkowski spacetime physics, such as the excitation of accelerated detectors. This partial listing is sufficient motivation to turn now to the basis of the theory of quantized fields in curved spacetime.

2.1 Canonical quantization and conservation laws

Consider a set of fields $\phi_a(x)$ propagating in a curved spacetime with invariant line element

$$ds^2 = g_{\mu\nu}(x)dx^\mu dx^\nu. \tag{2.1}$$

The metric $g_{\mu\nu}(x)$ will be treated as a given unquantized external field. We will assume that the spacetime has a well-defined causal structure and set of Cauchy hypersurfaces. The set of fields $\phi_a(x)$ to be quantized may include linearized gravitational wave perturbations propagating on the background $g_{\mu\nu}(x)$. Let n denote the dimension of spacetime, with x^0 being the time coordinate and $x^1, \ldots, x^{n-1}$ being the spatial coordinates.

The action S is constructed from the field ϕ_a, so that it is invariant under general coordinate transformations (diffeomorphisms):

$$S[\phi'(x'), \nabla'\phi'(x'), {g'}_{\mu\nu}(x')] = S[\phi(x), \nabla\phi(x), g_{\mu\nu}(x)]. \tag{2.2}$$

The simplest way to construct such an action is to start with the Minkowski spacetime action and replace ordinary derivatives ∂_μ by covariant derivatives $\nabla_\mu, \eta^{\mu\nu}$ by $g^{\mu\nu}$, and d^nx by the invariant volume element $d^nx|g|^{1/2}$, where $g = \det(g_{\mu\nu})$. This is called the minimal coupling prescription, and is consistent with the Einstein principle of equivalence, according to which local gravitational effects are not present in a neighborhood of the spacetime origin of a locally inertial frame of reference. Occasionally, we can further increase the symmetry by the addition to the Lagrangian of a term which does not vanish at the origin of a locally inertial frame, and such a possibility will also be considered. (The action already involves $\partial_\lambda g_{\mu\nu}$ through $\nabla_\mu\phi_a$; when additional terms are included, it may also involve higher derivatives of $g_{\mu\nu}$.)

The requirement that variations of the action

$$S = \int d^nx\mathcal{L}(\phi, \nabla\phi, g_{\mu\nu}) \tag{2.3}$$

vanish with respect to variations of the fields ϕ_a, which are zero on the boundary of integration, then yields the Euler–Lagrange equations

$$\partial_\mu\left(\frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi_a)}\right) - \frac{\partial\mathcal{L}}{\partial\phi_a} = 0. \tag{2.4}$$

The general covariance of (2.4) is insured by the invariance of the action S. Note that $\mathcal{L}$ is a scalar density, transforming like $|g|^{1/2}$.

Variation of the action with respect to the external field $g_{\mu\nu}$ does not in general vanish because we have not included an additional term proportional to the scalar curvature R in (2.3). (Such a term gives rise to the geometric part of the Einstein gravitational field equations. See Misner *et al.* (1973) or Weinberg (1972) for example.) However, because of the invariance of S under general coordinate transformation, δS will be zero under the change in $g_{\mu\nu}$ induced by an infinitesimal coordinate transformation

$$x^\mu \to x'^\mu = x^\mu - \epsilon^\mu(x), \tag{2.5}$$

where x and x' refer to the same event in spacetime. Under this transformation we have

$$g_{\mu\nu}(x) \to g'_{\mu\nu}(x') = \frac{\partial x^\lambda}{\partial x'^\mu}\frac{\partial x^\sigma}{\partial x'^\nu} g_{\lambda\sigma}(x),$$

which yields the variation

$$\delta_o g_{\mu\nu}(x) \equiv g'_{\mu\nu}(x) - g_{\mu\nu}(x) = \mathcal{L}_\epsilon g_{\mu\nu}, \tag{2.6}$$

where $\mathcal{L}_\epsilon$ denotes the Lie derivative,

$$\mathcal{L}_\epsilon g_{\mu\nu} = \nabla_\mu \epsilon_\nu + \nabla_\nu \epsilon_\mu. \tag{2.7}$$

Let us assume that $\epsilon^\mu(x)$ and $\partial_\lambda \epsilon^\mu(x)$ are zero on the boundary of the region of integration defining the action S of (2.3). Then, under the above infinitesimal coordinate transformation, we have

$$\delta S = \int d^n x \frac{\delta S}{\delta g_{\mu\nu}(x)} \delta_0 g_{\mu\nu}(x),$$

because variations in S produced by the changes in the dynamical fields ϕ_a vanish as a consequence of (2.4) and the boundary conditions on ϵ_μ. With $\delta_o g_{\mu\nu}$ given by (2.6) and (2.7), the invariance of S under coordinate transformations requires that $\delta S = 0$. Hence, with $dv_x \equiv d^n x |g|^{1/2}$,

$$\delta S = -\int dv_x T^{\mu\nu} \nabla_\mu \epsilon_\nu = 0, \tag{2.8}$$

where we have defined the tensor

$$T^{\mu\nu} \equiv -2|g|^{-1/2} \frac{\delta S}{\delta g_{\mu\nu}(x)}, \tag{2.9}$$

and have used its symmetry under interchange of indices. Then from (2.8) and

$$\begin{aligned}\nabla_\mu (T^{\mu\nu}\epsilon_\nu) &= |g|^{-1/2}\partial_\mu (|g|^{1/2} T^{\mu\nu}\epsilon_\nu) \\ &= (\nabla_\mu T^{\mu\nu})\epsilon_\nu + T^{\mu\nu}\nabla_\mu \epsilon_\nu,\end{aligned}$$

it follows that

$$\int dv_x (\nabla_\mu T^{\mu\nu})\epsilon_\nu = 0,$$

and because ϵ_ν is arbitrary, we must have

$$\nabla_\mu T^{\mu\nu} = 0. \tag{2.10}$$

This is the generally covariant generalization of (1.23). Furthermore, $T^{\mu\nu}$ as defined by (2.9) is symmetric, which is not true in general for the canonical energy-momentum tensor $\Theta^{\mu\nu}$ of (1.24). (In Chapter 1, the symbol $T^{\mu\nu}$ was used for the canonical energy-momentum tensor, $\Theta^{\mu\nu}$, but from now on $T^{\mu\nu}$ will refer to the symmetric energy-momentum tensor.) From $\delta(g^{\mu\nu}g_{\mu\lambda}) = 0$, it follows that $g_{\mu\lambda}g_{\nu\sigma}\frac{\delta}{\delta g_{\lambda\sigma}} = -\frac{\delta}{\delta g^{\mu\nu}}$, so

$$T_{\mu\nu} = +\frac{2}{|g|^{1/2}}\frac{\delta S}{\delta g^{\mu\nu}}. \tag{2.11}$$

The sign convention in our definition of $T^{\mu\nu}$ in (2.9) is chosen so that T_{00} will be positive for the classical electromagnetic field. If we were to use the opposite metric signature, then the signs on the right-hand sides of (2.9) and (2.11) would also be opposite.

We can calculate the symmetric energy-momentum tensor $T^{\mu\nu}$ in curved spacetime and then go to the flat spacetime limit, thereby obtaining a symmetric energy-momentum tensor satisfying $\partial_\mu T^{\mu\nu} = 0$ in Minkowski spacetime. For any isolated system in Minkowski spacetime, both $\Theta^{\mu\nu}$ and $T^{\mu\nu}$ yield a conserved energy-momentum vector p_μ (which, as noted earlier, is unique). In curved spacetime it is the symmetric energy-momentum tensor $T^{\mu\nu}$ defined in (2.9) which describes the matter and radiation and couples to the gravitational field through the Einstein field equations. (In a general curved spacetime, the tensor density $\Theta^{\mu\nu}$ of weight $1/2$ defined by (1.24) does not satisfy $\partial_\mu \Theta^\mu{}_\nu = 0$ as it does in Minkowski spacetime, nor any simple generalization of that equation.)

The Schwinger operator action principle continues to hold in curved spacetime for an arbitrary infinitesimal transformation of the form given in (1.28) and (1.29), provided that under the transformation $\delta_0 g_{\mu\nu}(x) = 0$, or

$$g'_{\mu\nu}(x) = g_{\mu\nu}(x). \tag{2.12}$$

In that case, the derivation leading to (1.34) goes through as before (with the canonical stress tensor now called $\Theta^\mu{}_\nu$); because we have $\delta_0 g_{\mu\nu} = 0$, it is not necessary for the Euler–Lagrange equation to hold for the external field $g_{\mu\nu}$. (Note that the metric $g_{\mu\nu}$ is not included here among the fields ϕ_a of (1.34).) Thus, under a transformation of the coordinates and fields given in (1.28) and (1.29), and for which (2.12) holds, we have as in Section 1.1

$$\delta S = G(t_2) - G(t_1), \tag{2.13}$$

with

$$G(t) = \int d^{n-1}x[\pi_a \delta\phi_a - \Theta^o{}_\nu \delta x^\nu], \tag{2.14}$$

where, as noted above, $\Theta^o{}_\nu$ is defined by the right-hand side of (1.24). Here the integration is on a constant time hypersurface, and

$$\pi_a = \frac{\partial \mathcal{L}}{\partial(\partial_0 \phi_a)}. \tag{2.15}$$

As before, with suitable operator ordering and regularization, G acts as the generator of the transformation, satisfying

$$i\delta_0 F = [F, G], \tag{2.16}$$

where F is a functional of the ϕ_a and π_a.

As in Minkowski spacetime, (2.16) will hold with

$$[\phi_a(\vec{x},t), \phi_b(\vec{x}',t)] = 0, \quad [\pi_a(\vec{x},t), \pi_b(\vec{x}',t)] = 0,$$
$$[\phi_a(\vec{x},t), \pi_b(\vec{x}',t)] = i\delta_{a,b}\delta(\vec{x}-\vec{x}'), \tag{2.17}$$

which is appropriate for bosons, or

$$\{\phi_a(\vec{x},t), \phi_b(\vec{x}',t)\} = 0, \quad \{\pi_a(\vec{x},t)\pi_b(\vec{x}',t)\} = 0,$$
$$\{\phi_a(\vec{x},t), \pi_b(\vec{x}',t)\} = i\delta_{a,b}\delta(\vec{x}-\vec{x}'), \tag{2.18}$$

which is appropriate for fermions. Here $\delta(\vec{x}-\vec{x}')$ is the Dirac δ-function satisfying $\int d^{n-1}x\delta(\vec{x}-\vec{x}')f(\vec{x}) = f(\vec{x}')$ with the integral being over the spacelike hypersurface $t =$ constant. The above commutation relations are imposed on independent field components. (They may be modified when gauge conditions are present.) One can show that $\pi(\vec{x}',t)$ and $\delta(\vec{x}-\vec{x}')$ each transform as spatial scalar densities under transformations of the spatial coordinates on the constant-t hypersurface. Hence, (2.17) and (2.18) are covariant under transformations of the spatial coordinates on the hypersurface, and are therefore the spatially covariant generalization of the corresponding relations that hold in flat spacetime. One can also show that they are consistent with the equations of motion of the fields, in the sense that if they hold on one constant-t spatial hypersurface, then they also will hold on the other constant-t spatial hypersurfaces.

Furthermore, for bosons we can define a complete set of commuting Hermitian fields, and define a basis $|\phi' >$ for the Hilbert space of state vectors, as in Minkowski spacetime, through (1.19). This gives a field or Schrödinger representation. For a system of fermions, we can define a set of Hermitian commuting quantities which are bilinear in the fields and build a Hilbert space of state vectors spanned by the simultaneous eigenvectors of these bilinear operators.

There are several cases of interest when G is conserved. The simplest is when $\delta x^\mu = 0$ and $\delta_0\phi_a$ is a symmetry of $\mathcal{L}$. Then (2.12) is trivially satisfied, and the symmetry of $\mathcal{L}$ implies that $\delta S = 0$ so that G is independent of time. One also has, in that case, $\partial_\mu J^\mu = 0$, with

$$J^\mu = \frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi_a)}\delta\phi_a. \tag{2.19}$$

(Note that J^μ is a vector density of weight $1/2$, so that we have $\nabla_\mu(|g|^{-1/2}J^\mu) = 0$.) Thus, in curved spacetime the electric charge and the generators of internal symmetries continue to be conserved.

The generator G is also conserved when $\delta x^\mu \neq 0$, but is such that (2.12) holds, and the fields ϕ_a are components of a spacetime tensor. Then, invariance under coordinate transformations implies that $\delta S = 0$, and (2.13) implies that G is constant. For example, if it is possible to choose a coordinate system in which a particular coordinate, say x^λ, does not appear in $g_{\mu\nu}(x)$, then under a translation in the x^λ direction (2.12) holds, and furthermore we have $\delta\phi(x) = 0$. Then, as in Section 1.1, it follows that

$$p_\lambda = \int d^{n-1}x \Theta^0{}_\lambda \tag{2.20}$$

is constant. (Since $g^{\lambda\mu}$ and the other components of p_μ may not be constants, it does not follow that $p^\lambda = g^{\lambda\mu}p_\mu$ is constant.) A similar expression involving the symmetric energy-momentum tensor $T_{\mu\nu}$ also holds. In deriving that result, we will also work in a more covariant language.

A coordinate transformation for which (2.12) holds is called an isometry of the spacetime. For an infinitesimal isometry of the form

$$x^\mu \to x'^\mu = x^\mu - \epsilon\xi^\mu(x), \tag{2.21}$$

with $\epsilon << 1$, it follows from (2.6) and (2.12) that

$$\mathcal{L}_\xi g_{\mu\nu} \equiv \nabla_\mu\xi_\nu + \nabla_\nu\xi_\mu = 0. \tag{2.22}$$

A vector field satisfying (2.22) is called a Killing vector of the spacetime. As a consequence of the symmetry of $T^{\mu\nu}$ and (2.10) we have

$$\nabla_\mu(T^{\mu\nu}\xi_\nu) = 0. \tag{2.23}$$

But since $T^{\mu\nu}\xi_\nu$ is a vector, we have $\nabla_\mu(T^{\mu\nu}\xi_\nu) = |g|^{-1/2}\partial_\mu(|g|^{1/2}T^{\mu\nu}\xi_\nu)$, and it follows that

$$\partial_\mu(|g|^{1/2}T^{\mu\nu}\xi_\nu) = 0. \tag{2.24}$$

Hence

$$P_\xi \equiv \int dV_x T^0{}_\nu(x)\xi^\nu(x) \tag{2.25}$$

is constant, where $dV_x = d^{n-1}x|g|^{1/2}$. In the case when the coordinates are such that $g_{\mu\nu}(x)$ is independent of a particular coordinate, say x^λ, then $\xi^\nu = \delta^\nu{}_\lambda$ is a Killing vector field, and (2.25) reduces to $P_\lambda = \int dV_x T^0{}_\lambda$ being constant. In such a case, (2.20) should yield the same p_λ to within possibly an additive constant independent of the field configuration.

Thus, in a general curved spacetime, we have shown that $\nabla_\mu T^{\mu\nu} = 0$, and in a spacetime having special symmetries or isometries as implied by the existence of one or more Killing vector fields ξ^μ, we have $\partial_\mu(|g|^{1/2}T^\mu{}_\nu\xi^\nu) = 0$ and $P_\xi \equiv \int dV_x T^0{}_\nu\xi^\nu$ is constant.

Let us now consider again a general curved spacetime without special isometries. An invariance of the action which is of interest is curved spacetime conformal invariance. Certain fields, such as the electromagnetic and massless Dirac fields in curved spacetime, exhibit an invariance of the action under conformal transformations of the metric and field, as defined below. An action $S[\phi, g_{\mu\nu}]$ (suppressing derivatives for brevity) is conformally invariant if

$$S[\phi, g_{\mu\nu}] = S[\tilde{\phi}, \tilde{g}_{\mu\nu}] + \text{surface integral}, \tag{2.26}$$

where

$$\tilde{g}_{\mu\nu}(x) = \Omega^2(x) g_{\mu\nu}(x), \tag{2.27}$$

$$\tilde{\phi}(x) = \Omega^{2p}(x)\phi(x). \tag{2.28}$$

Here $\Omega(x)$ is an arbitrary function and p is a dimensionless constant. Consider an infinitesimal conformal transformation with

$$\Omega^2(x) = 1 + \lambda(x), \quad |\lambda(x)| << 1. \tag{2.29}$$

Then $\delta x^\mu = 0$, and from (2.27) and (2.28),

$$\delta_0 g_{\mu\nu}(x) = \tilde{g}_{\mu\nu}(x) - g_{\mu\nu}(x) = \lambda(x) g_{\mu\nu}(x), \tag{2.30}$$

$$\delta_0 \phi(x) = \tilde{\phi}(x) - \phi(x) = p\lambda(x)\phi(x). \tag{2.31}$$

From (2.26), if $\lambda(x)$ vanishes sufficiently rapidly on the boundary of the region of integration,

$$0 = \delta S = \int d^n x \left\{ \frac{\delta S}{\delta \phi} \delta_0 \phi + \frac{\delta S}{\delta g_{\mu\nu}} \delta_0 g_{\mu\nu} \right\}.$$

If we assume that ϕ satisfies the Euler–Lagrange equation, $\delta S/\delta\phi = 0$, then it follows that

$$0 = \int d^n x \frac{\delta S}{\delta g_{\mu\nu}(x)} g_{\mu\nu}(x)\lambda(x),$$

and because $\lambda(x)$ is arbitrary,

$$\frac{\delta S}{\delta g_{\mu\nu}(x)} g_{\mu\nu}(x) = 0,$$

from which we have

$$g_{\mu\nu} T^{\mu\nu} = 0. \tag{2.32}$$

Thus, conformal invariance of the action implies that the trace of the energy-momentum tensor is zero. We will find later that a theory based on a classical or "bare" action which is conformally invariant will in general lose its conformal invariance in the quantum theory as a result of renormalization. The energy-momentum tensor thus acquires a non-vanishing trace, known as the trace or conformal anomaly.

In this section, we have discussed the action, field equations, symmetric energy-momentum tensor, generators of field transformations, commutation or anticommutation relations, Hilbert space of state vectors in the field representation, isometries and conservation laws, and conformal invariance. Having built a foundation for our discussion, let us consider next the curved spacetime generalization of the free neutral scalar field.

2.2 Scalar field

Following the minimal coupling prescription based on the principle of equivalence and described in Section 2.1, the action of the free neutral scalar field based on the Lagrangian density of (1.38) becomes in curved spacetime

$$S = \int d^n x \mathcal{L} \tag{2.33}$$

with

$$\mathcal{L} = \frac{1}{2}|g|^{1/2}(g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi - m^2\phi^2). \tag{2.34}$$

(For the scalar field we have $\nabla_\mu\phi = \partial_\mu\phi$.) This is a special case of the more general Lagrangian density

$$\mathcal{L} = \frac{1}{2}|g|^{1/2}(g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi - m^2\phi^2 - \xi R\phi^2), \tag{2.35}$$

where ξ is a dimensionless constant and R is the scalar curvature of the spacetime, $R \equiv g^{\mu\nu}R_{\mu\nu}$. (As noted in the section on notation, we use the following conventions: metric signature -2, $R^\mu{}_{\nu\lambda\sigma} = \partial_\sigma\Gamma^\mu{}_{\nu\lambda} - \partial_\lambda\Gamma^\mu{}_{\nu\sigma} + \Gamma^\tau{}_{\nu\lambda}\Gamma^\mu{}_{\sigma\tau} - \Gamma^\tau{}_{\nu\sigma}\Gamma^\mu{}_{\lambda\tau}$, and $R_{\mu\nu} = R^\lambda{}_{\mu\lambda\nu}$, in agreement with Birrell and Davies (1982), or $(-,-,-)$ in the notation of Misner *et al.* (1973).) We will carry out our discussion for this more general Lagrangian. One reason is that when we include an interaction term such as $|g|^{1/2}\lambda\phi^4$, with λ a constant (dimensionless in four-dimensional spacetime), then a term of the form $|g|^{1/2}\xi R\phi^2$ is needed for renormalization. (This will be shown in Section 6.7.) Another reason is that when $m = 0$ and $\xi = 1/6$ (in four dimensions), the action is invariant under the curved spacetime conformal transformations. The case $\xi = 0$ is referred to as minimal coupling. The magnitude of ξ cannot be very large, because if ξ has a non-zero value, then the term in the Lagrangian that is proportional to $R\phi^2$ can cause the effective gravitational constant to vary with time and position as a result of such variations in ϕ.

Consider an infinitesimal conformal transformation of the form of (2.29)–(2.31) with $p = -1/2$:

$$g_{\mu\nu}(x) \to \tilde{g}_{\mu\nu}(x) = (1 + \lambda(x))g_{\mu\nu}(x), \tag{2.36}$$

$$\phi(x) \to \tilde{\phi}(x) = \left(1 - \frac{1}{2}\lambda(x)\right)\phi(x). \tag{2.37}$$

data 02.05.2020

Pág. 337(352) Chapter 10
The Einstein-Langevin Equation

Pág. 364(379) Chapter 11
Metric Fluctuations in Minkowski Spacetime.

Obs: Os dois capitulos tem vinculos (Langevin Equation)

Pág. 467(482) Chapter 15 e seguintes
Stress-Energy Tensor Fluctuations in de Sitter Space.

Pág. 519 (534) Chapter 17
Riemann Tensor Correlator in de Sitter

B11) Quantum Field Theory in Curved Spacetime — Quantized Fields and Gravity (2009...)...pdf

Obs, Livro todo!!!

Pág. 01 (19) Chapter 1
Quantum fields in Minkowski spacetime

Pág. 36(54) Chapter 2
Basics of quantum fields in curved spacetimes.

" 5

)*+,)))-) 3 =## = % # ."
(' !# ' ; @ ;

16 μs. Hence, the average distance travelled by a muon is 4,700 m and these are able to reach the surface of earth. Thus, the results of time dilation are in agreement with the experimental results.

CERN in Geneva conducted an experiment in 1976, in which muons were injected into large storage rings reaching speeds of about 0.9994*c*. Electrons were produced by decaying of muons. The decay lifetime of muons was measured. The lifetime of moving muons was found to be 30 times higher than that of stationary muons, in perfect agreement with the theory of relativity.

1.7 Lorentz Contraction (Length Contraction)

In the previous sections, we have discussed that time is not absolute. In other words, we can say that the time interval between two events depend on the reference frame. In the same way, the measured distance also depends on the reference frame.

If we consider Galilean transformations, then the measured size for an object remains unaffected by the relative motion of observers. But for Lorentz transformation, length varies as time interval varies. To understand this concept, we consider a stick AB at rest w.r.t. frame S and moving w.r.t. frame S′ as shown in Figure 1.7.

The rod is parallel to *X*-axis and difference between AB gives actual length, L_0.

According to reference frame S

$$L_0 = x_2 - x_1 \tag{35}$$

and from reference frame S′

$$L_o' = x_2' - x_1' \tag{36}$$

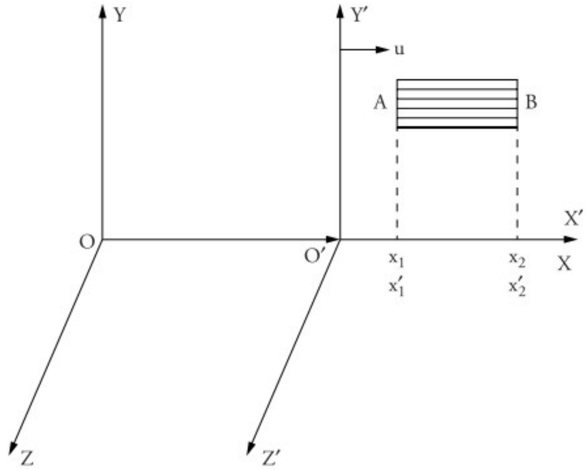

Figure 1.7 The length of rod for S = $x_2 - x_1$ and for S′ = $x_2' - x_1'$.

Substitute $x_2' = \gamma(x_2 - ut)$

$$\left.\begin{aligned} \text{or} \quad x_2 &= \gamma\left(x_2' + ut'\right) \\ \text{and} \quad x_1 &= \gamma\left(x_1' + ut'\right) \end{aligned}\right] \text{in Eqn. (35)}$$

$$L_o = \gamma\left[x_2' + \cancel{u}t' - x_1' - \cancel{u}t'\right]$$

$$L_o = \gamma\left(x_2' - x_1'\right) \tag{37}$$

$$\begin{aligned} L_o &= \gamma L_o' \\ As \quad \gamma > 1 \quad & L_o \geq L_o' \end{aligned} \tag{38}$$

The dimensions of the rod are same in *y* and *z* directions because there is no motion along these directions. L_0 is the proper length of the object. The proper length (like proper time) of an object is always measured by someone who is at rest w.r.t. the object. Hence, if the length of an object is measured from some moving reference frame, then the length appears to be small. This phenomena is called length contraction.

> Although muons are formed at a proper height of 6,000 m, the muons move towards earth at U = 0.998c, and hence the length appears to be contracted for them as $L'_o \frac{L_o}{\gamma}$ with a value of γ = 15.82. The height becomes $L_o' = \frac{6{,}000}{15.82} = 379.2$ m.

1.8 Concept of Space–time (Minkowski Space)

Einstein proposed the special theory of relativity, but the modern approach to the theory depends on the concept of four-dimensional universe as proposed by Hermann Minkowski in 1908. He proposed that if things could be arranged in time, then universe becomes four dimensional. He also emphasized that Einstein's discovery is a consequence of a four-dimensional universe. He used Pythagoras' theorem to relate space and time in four-dimensional space i.e.

$$S^2 = x^2 + y^2 + z^2 + (ict)^2 \tag{39}$$

where *x*, *y* and *z* are displacements, *c* is constant and *t* is time spanned by space–time interval S. If we do a dimensional analysis, then it becomes

Meters = (Conversion constant) × Time

c has a value of 3×10^8 m/s. This transforms Eqn. (39) as

$$S^2 = x^2 + y^2 + z^2 - (ct)^2 \tag{40}$$

If universe is four dimensional, then the space–time interval will be invariant. The constancy of velocity is also its consequence, as explained:

(i) If an object is travellling at velocity c, then the space–time interval is 0.

$$S^2 = x^2 + y^2 + z^2 - (ct)^2 \tag{41a}$$

(ii) If an object moves along x-direction for t seconds with velocity v, then

$$S^2 = (vt)^2 - (ct)^2 \tag{41b}$$

If $v = c$, then

$$S^2 = 0 \tag{42}$$

Hence, we can say that if an object is traveling at c m/s, then all observers will measure the same velocity for it, no matter how fast the observers are moving. This universal constant "c" is the velocity of light. The second consequence of space–time demonstrates that clocks will appear to go slower on objects moving relative to observer.

To illustrate this, we can take the example of Jack and Jill. Let us assume that Jack and Jill are on separate planets moving away from each other. Jack draws a graph for Jill's motion through space and time as demonstrated below (Figure 1.8): Jack and Jill are on planets and they think they are stationary and moving through time only. Both think the same way for each other.

Jack calculates Jill's space–time as

$$S^2 = (vt)^2 - (ct)^2 \tag{43}$$

whereas Jill calculates his space–time as

$$S^2 = (0t)^2 - (cT)^2 \tag{44}$$

But as the space–time interval is invariant, that is,

$$-(cT)^2 = (vt)^2 - (ct)^2$$

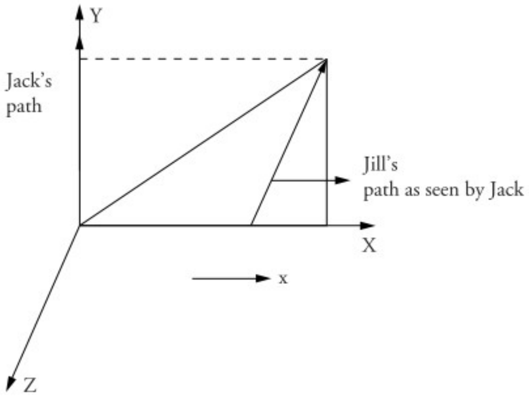

Figure 1.8 Jill thinks he is moving through time only, but Jack thinks he is moving through both space and time.

$$t = \frac{T}{\sqrt{1 - \frac{v^2}{c^2}}} \tag{45}$$

Hence, $T = 1$ between two ticks of clocks that are in rest in Jill's reference frame.

Great care must be taken while interpreting space–time diagrams. The general space–time diagram is shown in Figure 1.9. Time is plotted on a vertical axis and space coordinates on a horizontal axis. World line represents the plot of position of an object w.r.t. time. Hence, an event is a point in space–time and world line indicates a line of space–time. Line of simultaneity is the line parallel to x-axis, and it makes a difference in Galilean relativity and special relativity. In Galilean relativity, all events occur simultaneously on line of simultaneity.

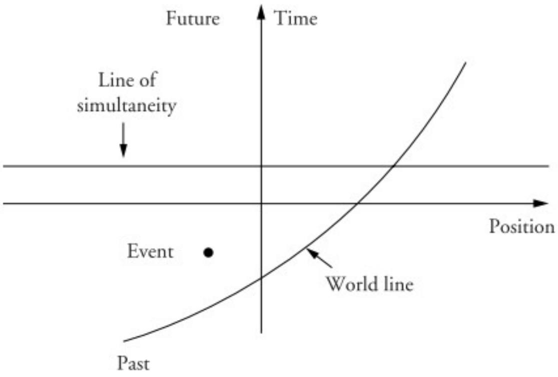

Figure 1.9 Space–time diagram.

The slope of world line has significant meaning.

$$\text{Velocity} = \frac{1}{\text{Slope of world line}} \tag{46}$$

This is valid for "+ve" and "–ve" slopes. If an object changes its velocity, then the world line is curved and velocity is represented by tangent to curve.

In conclusion, the space time as interpreted by minkowski is:

> The views of space and time which I wish to lay before you have spring from the soil of experimental physics, and therein lies their strength. They are radical. Henceforth time by itself and space by itself are doomed to fade away in mere shadows and only a union of two will preserve an independent reality. (Herman Minkowski, 1908, Assembly of German Natural Scientists and Physicians)

The motion of objects was provided by earth itself: from astronomical calculations, it was predicted to be 30,000 m/s. As this speed is 4 orders smaller than the velocity of light, the maximum difference expected in terrestrial measurements was 0.0002c.

9 + "
)*+,)))-) 3 9 +
" *.. F BL !: > A R > ."
(' . 1

12

Mechanical Curves

In this chapter we shall describe mechanical curves from the Lagrangian and Hamiltonian point of view. In this way, many geometric properties of these curves will be derived from the variational formalism.

A mechanical curve is a curve described by a particle on which acts an exterior force. For instance the circle, cycloid, hypocycloid, astroid, etc are models of particle trajectories under some exterior forces. A particle on which acts a central force of constant magnitude describes a circle. A point on a circle which rolls on a line, without slipping, describes a cycloid. A point on a circle tangent interior to another circle, which rotates without slipping in the interior of the large circle, describes a hypocycloid.

12.1 The areal velocity

Suppose an object moves in the plane from the point A to point B along a continuous arc $\widehat{AB}$. Let $\mathcal{A}$ be the area swept by the vectorial radius $\overrightarrow{OX}$ with $X \in \widehat{AB}$. We shall consider positive the orientation given by the clock-wise rotation of $\overrightarrow{OX}$. An elementary calculus formula in polar coordinates yields

$$\mathcal{A} = \frac{1}{2}\int_0^{\phi} r^2\,d\phi, \tag{12.1.1}$$

where $r = r(\phi)$ is the length of the vectorial radius and the argument angle $\phi = \angle AOX$.

Written in differential form, we have $d\mathcal{A} = \dfrac{1}{2}r^2\,d\phi$. Let t be the time parameter. Then

$$\frac{d\mathcal{A}}{dt} = \frac{1}{2}r^2\frac{d\phi}{dt} = \frac{1}{2}r^2\dot{\phi}. \tag{12.1.2}$$

The derivative $d\mathcal{A}/dt$ is called *areal velocity*.

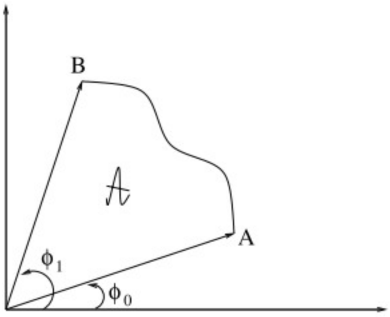

Figure 12.1: The area swept by the vectorial radius between two points.

12.1.0.1 Areal velocity as an angular momentum

Using polar coordinates $x = r\ \cos\phi,\ y = r\ \sin\phi$, the areal velocity becomes

$$\begin{aligned}\frac{d\mathcal{A}}{dt} &= \frac{1}{2}r^2\dot{\phi} = \frac{1}{2}\big(r\cos\phi\, r\cos\phi\,\dot{\phi} - r\sin\phi\,(-r\cos\phi)\,\dot{\phi}\big)\\ &= \frac{1}{2}(x\dot{y} - y\dot{x}).\end{aligned}$$

The expression $x\dot{y} - y\dot{x} = \langle (x, y), (\dot{x}, -\dot{y})\rangle$ is called *angular momentum*.

If $x\dot{y} - y\dot{x}$ is constant, the particle moves such that equal areas are described in equal amounts of time *i.e.*, the vectorial radius sweeps out equal areas in equal time. This happens for instance, in the case of a particle in uniform motion on a circle or in the case of a planet in the revolution motion around the sun (Kepler's second law).

12.2 The circular motion

Consider a particle in the (x, y)-plane which is described by the Lagrangian

$$L(x, y, \dot{x}, \dot{y}) = \frac{1}{2}(\dot{x}^2 + \dot{y}^2) + (x\dot{y} - y\dot{x}) \tag{12.2.3}$$

i.e., the particle moves on a trajectory which is an extremizer for the action

$$S = \int L(x, y, \dot{x}, \dot{y}). \tag{12.2.4}$$

The Lagrangian L is the sum of the kinetic energy and the angular momentum.

Theorem 12.1. *The Euler–Lagrange system associated with the Lagrangian* (12.2.3) *is*

$$\begin{cases}\ddot{x} - 2\dot{y} &= 0,\\ \ddot{y} + 2\dot{x} &= 0.\end{cases} \tag{12.2.5}$$

The solutions of the system (12.2.5) *with the boundary conditions*

$$x(0) = \mathbf{x}_0, \qquad x(\tau) = \mathbf{x}, \qquad y(0) = \mathbf{y}_0, \qquad y(\tau) = \mathbf{y}, \tag{12.2.6}$$

with $0 < \tau < \pi$, *are*

$$\begin{aligned} x(s) &= \pm C \sin s\, \sin(s + \alpha_0) + \mathbf{x}_0, \\ y(s) &= \pm C \sin s\, \cos(s + \alpha_0) + \mathbf{y}_0, \end{aligned}$$

with $C = \sqrt{\frac{E}{2}}$ *and the energy* E *given by*

$$E = \frac{(\mathbf{x} - \mathbf{x}_0)^2 + (\mathbf{y} - \mathbf{y}_0)^2}{2\sin^2 \tau}. \tag{12.2.7}$$

Proof. Using

$$\begin{aligned} &\frac{\partial L}{\partial \dot{x}} = \dot{x} - y, \qquad \frac{d}{dt}\frac{\partial L}{\partial \dot{x}} = \ddot{x} - \dot{y}, \qquad \frac{\partial L}{\partial x} = \dot{y}, \\ &\frac{\partial L}{\partial \dot{y}} = \dot{y} + x, \qquad \frac{d}{dt}\frac{\partial L}{\partial \dot{y}} = \ddot{y} + \dot{x}, \qquad \frac{\partial L}{\partial \dot{y}} = -\dot{x}, \end{aligned}$$

it is easy to see that the Euler–Lagrange system becomes (12.2.5).

Multiplying the first equation of (12.2.5) by $\dot{x}$ and the second by $\dot{y}$ and adding yields $\dot{x}\ddot{x} + \dot{y}\ddot{y} = 0$, therefore

$$\frac{d}{dt}(\dot{x}^2 + \dot{y}^2) = 0 \Longrightarrow \dot{x}^2 + \dot{y}^2 = C^2 \tag{12.2.8}$$

where C is a constant along the trajectory. Let $E = \frac{1}{2}(\dot{x}^2 + \dot{y}^2)$ denote the first integral of energy. Using (12.2.8) there is a smooth function $\alpha = \alpha(s)$ such that

$$\begin{aligned} \dot{x}(s) &= \pm C \sin \alpha(s) \Longrightarrow \ddot{x}(s) = \pm C \cos \alpha(s)\, \dot{\alpha}(s), \\ \dot{y}(s) &= \pm C \cos \alpha(s) \Longrightarrow \ddot{y}(s) = \mp C \sin \alpha(s)\, \dot{\alpha}(s). \end{aligned}$$

Substituting back in the system (12.2.5) yields

$$\begin{aligned} \pm \cos^2 \alpha(s)\, \dot{\alpha}(s) &= \pm 2 \cos^2 \alpha(s), \\ \mp \sin^2 \alpha(s)\, \dot{\alpha}(s) &= \mp 2 \sin^2 \alpha(s). \end{aligned}$$

Subtracting we get

$$\dot{\alpha}(s) = 2 \Longrightarrow \alpha(s) = 2s + \alpha_0, \tag{12.2.9}$$

with α_0 constant. Hence

$$\begin{aligned}\dot{x}(s) &= \pm C\sin(2s+\alpha_0) \Longrightarrow x(s) = \pm C\int_0^s \sin(2u+\alpha_0)\,du + \mathbf{x}_0\\ &= \pm C\frac{1}{2}(-\cos w)\Big|_{\alpha_0}^{2s+\alpha_0} + \mathbf{x}_0 = \pm\frac{C}{2}\Big(\cos\alpha_0 - \cos(2s+\alpha_0)\Big) + \mathbf{x}_0\\ &= \pm C\sin s\sin(s+\alpha_0) + \mathbf{x}_0,\end{aligned}$$

where we used

$$\begin{aligned}\cos\alpha_0 - \cos(2s+\alpha_0) &= -2\sin\frac{\alpha_0+2s+\alpha_0}{2}\sin\frac{\alpha_0-2s-\alpha_0}{2}\\ &= 2\sin(s+\alpha_0)\sin s.\end{aligned}$$

Substituting $\alpha(s)$ in the formula for $\dot{y}(s)$ yields

$$\begin{aligned}\dot{y}(s) &= \pm C\cos\alpha(s) = \pm C\cos(2s+\alpha_0)\\ \Longrightarrow y(s) &= \pm C\int_0^s \cos(2u+\alpha_0)\,du + \mathbf{y}_0\\ &= \pm\frac{1}{2}C\sin w\Big|_{\alpha_0}^{2s+\alpha_0} + \mathbf{y}_0\\ &= \pm\frac{1}{2}C\Big(\sin(2s+\alpha_0) - \sin\alpha_0\Big) + \mathbf{y}_0\\ &= \pm C\sin s\cos(s+\alpha_0) + \mathbf{y}_0,\end{aligned}$$

where we used

$$\begin{aligned}\sin(2s+\alpha_0) - \sin\alpha_0 &= 2\sin\frac{2s+\alpha_0-\alpha_0}{2}\cos\frac{2s+\alpha_0+\alpha_0}{2}\\ &= 2\sin s\cos(s+\alpha_0).\end{aligned}$$

Hence we have arrived at

$$x(s) = \pm C\sin s\ \sin(s+\alpha_0) + \mathbf{x}_0, \tag{12.2.10}$$

$$y(s) = \pm C\sin s\ \cos(s+\alpha_0) + \mathbf{y}_0, \tag{12.2.11}$$

where α_0 is a constant. We shall show that energy $E = C^2/2$ does not depend on α_0. Making $s = \tau$ in (12.2.10), (12.2.11) yields

$$\mathbf{x} = \pm C\sin\tau\ \sin(\tau+\alpha_0) + \mathbf{x}_0, \tag{12.2.12}$$

$$\mathbf{y} = \pm C\sin\tau\ \cos(\tau+\alpha_0) + \mathbf{y}_0, \tag{12.2.13}$$

hence it follows that

$$\Big(\frac{\mathbf{x}-\mathbf{x}_0}{\sin\tau}\Big)^2 = C^2\sin^2(\tau+\alpha_0), \tag{12.2.14}$$

$$\Big(\frac{\mathbf{y}-\mathbf{y}_0}{\sin\tau}\Big)^2 = C^2\cos^2(\tau+\alpha_0). \tag{12.2.15}$$

Adding yields

$$2E = C^2 = \frac{(\mathbf{x}-\mathbf{x}_0)^2 + (\mathbf{y}-\mathbf{y}_0)^2}{\sin^2 \tau},$$

which is (12.2.7). ■

Remark 12.2 *Let* $d = \sqrt{(\mathbf{x}-\mathbf{x}_0)^2 + (\mathbf{y}-\mathbf{y}_0)^2}$ *denote the Euclidean distance between* $(\mathbf{x}_0, \mathbf{y}_0)$ *and* $(\mathbf{x}, \mathbf{y})$. *We note the fact that the energy* $E = \dfrac{d^2}{2\sin^2 \tau}$ *is not Euclidean. Replacing* $\sin^2 \tau$ *by* τ^2 *we obtain the Euclidean energy.*
Let δ *be the Riemannian distance in which the solutions of the Euler–Lagrange equations become geodesics. Then* $E = \dfrac{\delta^2}{2\tau^2}$. *Then* $\delta^2 = \Big(\dfrac{\tau}{\sin \tau}\Big)^2 d^2$, *and hence* d *and* δ *are homothetic.*

The action

The action $S = S(\mathbf{x}_0, \mathbf{y}_0, \mathbf{x}, \mathbf{y}, \tau)$ satisfies the Hamilton–Jacobi equation

$$\begin{aligned} \frac{\partial S}{\partial \tau} &= -E = -\frac{d^2}{2\sin^2 \tau} = \frac{d^2}{2}\frac{\partial}{\partial \tau}(\cot \tau) \\ \Longrightarrow 0 &= \frac{\partial}{\partial \tau}\Big(S - \frac{d^2}{2}\cot \tau\Big) \Longrightarrow S = S_0 + \frac{d^2}{2}\cot \tau. \end{aligned} \tag{12.2.16}$$

Proposition 12.3 *The Hamiltonian associated with the Lagrangian* (12.2.3) *is*

$$H(x, y, p_1, p_2) = \frac{1}{2}\big(p_1 + y\big)^2 + \frac{1}{2}\big(p_2 - x\big)^2. \tag{12.2.17}$$

Proof. The Hamiltonian system for the Hamiltonian (12.2.17) yields

$$\begin{aligned} \dot{x} &= H_{p_1} = p_1 + y \Longrightarrow p_1 = \dot{x} - y, \\ \dot{y} &= H_{p_2} = p_2 - x \Longrightarrow p_2 = \dot{y} + x. \end{aligned}$$

Using the Legendre transform we have

$$\begin{aligned} L &= p_1\dot{x} + p_2\dot{y} - H \\ &= (\dot{x} - y)\dot{x} + (\dot{y} + x)\dot{y} - \frac{1}{2}\big(p_1 + y\big)^2 - \frac{1}{2}\big(p_2 - x\big)^2 \\ &= (\dot{x} - y)\dot{x} + (\dot{y} + x)\dot{y} - \frac{1}{2}\dot{x}^2 - \frac{1}{2}\dot{y}^2 \\ &= \frac{1}{2}(\dot{x}^2 + \dot{y}^2) + x\dot{y} - y\dot{x}. \end{aligned}$$

■

We note that the Hamiltonian (12.2.17) is the principal symbol of the operator

$$\begin{aligned} P &= \frac{1}{2}\big(\partial_x + y\big)^2 + \frac{1}{2}\big(\partial_y - x\big)^2 \\ &= \frac{1}{2}(\partial_x^2 + \partial_y^2) + y\partial_x - x\partial_y + \frac{1}{2}(x^2 + y^2), \end{aligned}$$

which describes the circular motion.

12.3 The astroid

The trajectory of a point P on the unit circle which rolls without slipping in the interior of a circle of radius 4 is a hypocycloid with four cuspidal points. This curve is called *astroid*. The equation of the astroid is

$$x^{2/3} + y^{2/3} = 1. \tag{12.3.18}$$

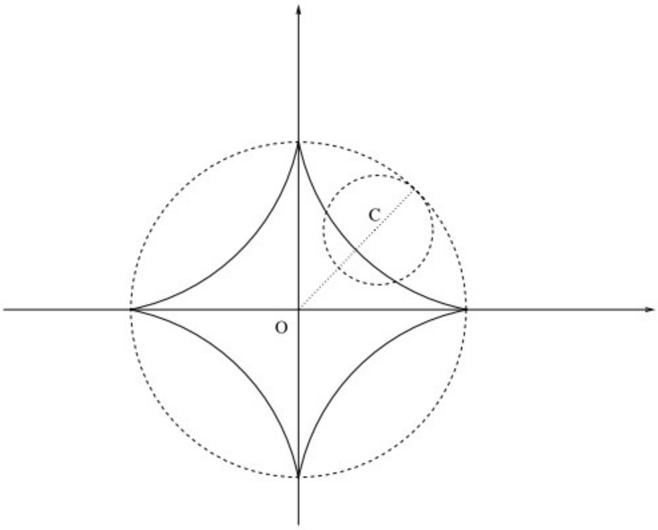

Figure 12.2: The astroid.

If P starts at the cuspidal point $(4, 0)$ and s denotes the angle argument of the center C, we have

$$x(s) = \cos^3 s, \qquad y(s) = \sin^3 s, \tag{12.3.19}$$

which are equivalent with

$$x(s) = 3\cos s + \cos\, 3s, \qquad y(s) = 3\sin s - \sin\, 3s. \tag{12.3.20}$$

A simple computation shows that (12.3.20) is the solution of the system

$$\begin{cases} \ddot{x} - 2\dot{y} + 3x = 0, \\ \ddot{y} + 2\dot{x} + 3y = 0, \end{cases} \tag{12.3.21}$$

-2 7 * " . ; A)*+,)))
." (

with initial conditions

$$x(0) = 4, \quad \dot{x}(0) = \dot{y}(0) = y(0) = 0. \tag{12.3.22}$$

Standard ODE techniques show that the solution (12.3.20) is unique.

Proposition 12.4 *The system* (12.3.21) *is the Euler–Lagrange system associated with the Lagrangian*

$$L(x, y, \dot{x}, \dot{y}) = \frac{1}{2}(\dot{x}^2 + \dot{y}^2) + x\dot{y} - \dot{x}y - \frac{3}{2}(x^2 + y^2). \tag{12.3.23}$$

Proof. We have

$$\frac{\partial L}{\partial \dot{x}} = \dot{x} - y, \quad \frac{d}{dt}\frac{\partial L}{\partial \dot{x}} = \ddot{x} - \dot{y}, \quad \frac{\partial L}{\partial x} = \dot{y} - 3x,$$

$$\frac{\partial L}{\partial \dot{y}} = \dot{y} + x, \quad \frac{d}{dt}\frac{\partial L}{\partial \dot{y}} = \ddot{y} + \dot{x}, \quad \frac{\partial L}{\partial y} = -\dot{x} - 3y.$$

Then

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{x}} = \frac{\partial L}{\partial x}, \quad \frac{d}{dt}\frac{\partial L}{\partial \dot{y}} = \frac{\partial L}{\partial y}$$

yields the system (12.3.21). ∎

12.3.0.2 Noether's Theorem

The Lagrangian (12.3.23) is invariant under rotations centered at the origin. The vector field associated with this rotation at the point (x, y) is $(-y, x)$. Noether's theorem yields a first integral of motion given by

$$\begin{aligned} I &= \left\langle \left(\frac{\partial L}{\partial \dot{x}}, \frac{\partial L}{\partial \dot{y}}\right), (-y, x) \right\rangle = (\dot{x} - y)(-y) + (\dot{y} + x)x \\ &= -\dot{x}y + y^2 + \dot{y}x + x^2 = x^2 + y^2 + x\dot{y} - \dot{x}y \\ &= r^2 + 2\frac{d\mathcal{A}}{ds}, \end{aligned}$$

where $x = r\cos\phi$, $y = r\sin\phi$. We have arrived at the following result.

Proposition 12.5 *For any solution of the system* (12.3.21) *there is a constant C such that*

(*i*)
$$r^2 + 2\frac{d\mathcal{A}}{ds} = C,$$

(*ii*)
$$\phi(s) = C\int_0^s \frac{du}{r^2(u)} - s$$

along the solution.

Proof. (i) It clearly follows from the fact that the first integral is constant along the solutions.
(ii) We have

$$\begin{aligned} C &= r^2 + x\dot{y} - \dot{x}y \\ &= r^2 + r^2\dot{\phi} \\ &= r^2(1+\dot{\phi}) \\ \Longrightarrow \frac{d\phi}{ds} &= \frac{C}{r^2} - 1. \end{aligned}$$

Integrating yields the desired result. ∎

We can get the same result if we write the Euler–Lagrange system in polar coordinates. See Exercise 2.

As the astroid is a solution of the system (12.3.21), the above proposition applies to it. In this case the constant C is obtained by taking the value at $s = 0$,

$$\begin{aligned} C &= x^2(0) + y^2(0) + x(0)\dot{y}(0) - \dot{x}(0)y(0) \\ &= 16. \end{aligned}$$

Proposition 12.6 *The Hamiltonian associated with the Lagrangian* (12.3.23) *is*

$$H(p_1, p_2, x, y) = \frac{1}{2}[(p_1 + y)^2 + (p_2 - x)^2] + \frac{3}{2}(x^2 + y^2). \tag{12.3.24}$$

Proof. The momenta are $p_1 = \dfrac{\partial L}{\partial \dot{x}} = \dot{x} - y, \quad p_2 = \dfrac{\partial L}{\partial \dot{y}} = \dot{y} + x$, and then

$$\dot{x} = p_1 + y, \quad \dot{y} = p_2 - x. \tag{12.3.25}$$

Using (12.3.25), the Legendre transform yields

$$\begin{aligned} H(p_1, p_2, x, y) &= p_1\dot{x} + p_2\dot{y} - \frac{1}{2}(\dot{x}^2 + \dot{y}^2) - x\dot{y} + \dot{x}y + \frac{3}{2}(x^2 + y^2) \\ &= p_1(p_1 + y) + p_2(p_2 - x) - \frac{1}{2}\Big((p_1 + y)^2 + (p_2 - x)^2\Big) \\ &\quad - x(p_2 - x) + (p_1 + y)y + \frac{3}{2}(x^2 + y^2) \\ &= \frac{1}{2}(p_1^2 + p_2^2) + p_1 y - p_2 x + 2(x^2 + y^2) \\ &= \frac{1}{2}\Big[(p_1 + y)^2 + (p_2 - x)^2\Big] + \frac{3}{2}(x^2 + y^2). \end{aligned}$$

∎

12.3.0.3 The first integral of energy

As

$$\frac{\partial H}{\partial t} = 0 \quad \text{and} \quad \frac{\partial H}{\partial t} = \frac{dH}{dt},$$

it follows that H is preserved along the trajectory. The value of H along the trajectory is called *the total energy*. In x, y, $\dot{x}$, $\dot{y}$ coordinates the energy takes the form

$$E = \frac{1}{2}(\dot{x}^2 + \dot{y}^2) + \frac{3}{2}(x^2 + y^2). \tag{12.3.26}$$

Note that E does not depend on the angular momentum as the Lagrangian does. It depends only on the magnitude of the velocity and the distance to the origin.

12.3.0.4 Physical interpretation

The speed of a particle described by a solution of the Euler–Lagrange system is

$$v = \sqrt{\dot{x}^2 + \dot{y}^2}.$$

If $r = \sqrt{x^2 + y^2}$ denotes the distance from the origin to the point (x, y), formula (12.3.26) yields

$$v^2 = 2E - 3r^2.$$

In the case of the astroid with the initial conditions (12.3.26) we have $E = 24$. Thus $v = \sqrt{3(16 - r^2)}$ with $r \in [0, 4]$. The speed on the astroid is zero iff $r = 4$, which occurs only at the cuspidal points.

12.4 The cycloid

Consider a particle described by a Lagrangian, which is the sum of the kinetic, angular momentum and potential energy in the x-direction

$$L(x, y, \dot{x}, \dot{y}) = \frac{1}{2}(\dot{x}^2 + \dot{y}^2) + \frac{1}{2}(x\dot{y} - \dot{x}y) + x. \tag{12.4.27}$$

The Euler–Lagrange system of equations associated with the Lagrangian (12.4.27) is

$$\begin{cases} \ddot{x} - \dot{y} = 1, \\ \ddot{y} + \dot{x} = 0. \end{cases} \tag{12.4.28}$$

If we consider the initial conditions

$$x(0) = 0, \quad \dot{x}(0) = 0, \quad y(0) = 0, \quad \dot{y}(0) = 0$$

" ")*+,)))-) 3 4 . B" " * 7 -
7 & P *.. : BQ . * " 2 = ." "

")*+,)))-) 3 4 . B" " * 7 - 7
& P *.. : BQ . * " 2 = ."

)*+,)))-) 3 E " " * 5 :
. 0 S A ."
(% =' . #

Chapter 3

Basic Topology of R

3.1 Discussion: The Cantor Set

What follows is a fascinating mathematical construction, due to Georg Cantor, which is extremely useful for extending the horizons of our intuition about the nature of subsets of the real line. Cantor's name has already appeared in the first chapter in our discussion of uncountable sets. Indeed, Cantor's proof that **R** is uncountable occupies another spot on the short list of the most significant contributions toward understanding the mathematical infinite. In the words of the mathematician David Hilbert, "No one shall expel us from the paradise that Cantor has created for us."

Let C_0 be the closed interval $[0, 1]$, and define C_1 to be the set that results when the open middle third is removed; that is,

$$C_1 = C_0 \setminus \left(\frac{1}{3}, \frac{2}{3}\right) = \left[0, \frac{1}{3}\right] \cup \left[\frac{2}{3}, 1\right].$$

Now, construct C_2 in a similar way by removing the open middle third of each of the two components of C_1:

$$C_2 = \left(\left[0, \frac{1}{9}\right] \cup \left[\frac{2}{9}, \frac{1}{3}\right]\right) \cup \left(\left[\frac{2}{3}, \frac{7}{9}\right] \cup \left[\frac{8}{9}, 1\right]\right).$$

If we continue this process inductively, then for each $n = 0, 1, 2, \dots$ we get a set C_n consisting of 2^n closed intervals each having length $1/3^n$. Finally, we define the *Cantor set* C (Fig. 3.1) to be the intersection

$$C = \bigcap_{n=0}^{\infty} C_n.$$

© Springer Science+Business Media New York 2015
S. Abbott, *Understanding Analysis*, Undergraduate Texts
in Mathematics, DOI 10.1007/978-1-4939-2712-8_3

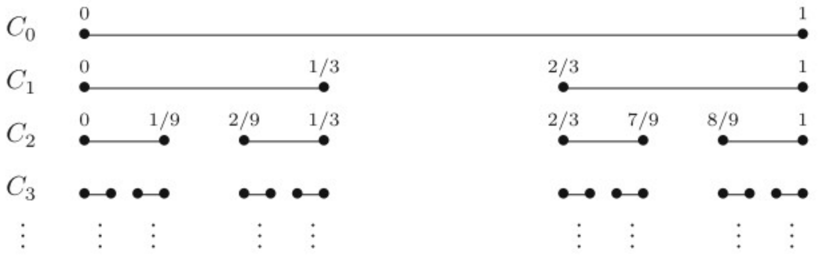

Figure 3.1: Defining the Cantor set; $C = \bigcap_{n=0}^{\infty} C_n$.

It may be useful to understand C as the remainder of the interval $[0, 1]$ after the iterative process of removing open middle thirds is taken to infinity:

$$C = [0,1] \backslash \left[\left(\frac{1}{3}, \frac{2}{3} \right) \cup \left(\frac{1}{9}, \frac{2}{9} \right) \cup \left(\frac{7}{9}, \frac{8}{9} \right) \cup \cdots \right].$$

There is some initial doubt whether anything remains at all, but notice that because we are always removing open middle thirds, then for every $n \in \mathbf{N}$, $0 \in C_n$ and hence $0 \in C$. The same argument shows $1 \in C$. In fact, if y is the endpoint of some closed interval of some particular set C_n, then it is also an endpoint of one of the intervals of C_{n+1}. Because, at each stage, endpoints are never removed, it follows that $y \in C_n$ for all n. Thus, C at least contains the endpoints of all of the intervals that make up each of the sets C_n.

Is there anything else? Is C countable? Does C contain any intervals? Any irrational numbers? These are difficult questions at the moment. All of the endpoints mentioned earlier are rational numbers (they have the form $m/3^n$), which means that if it is true that C consists of only these endpoints, then C would be a subset of $\mathbf{Q}$ and hence countable. We shall see about this. There is some strong evidence that not much is left in C if we consider the total length of the intervals removed. To form C_1, an open interval of length $1/3$ was taken out. In the second step, we removed two intervals of length $1/9$, and to construct C_n we removed 2^{n-1} middle thirds of length $1/3^n$. There is some logic, then, to defining the "length" of C to be 1 minus the total

$$\frac{1}{3} + 2\left(\frac{1}{9}\right) + 4\left(\frac{1}{27}\right) + \cdots + 2^{n-1}\left(\frac{1}{3^n}\right) + \cdots = \frac{\frac{1}{3}}{1 - \frac{2}{3}} = 1.$$

The Cantor set has *zero length.*

To this point, the information we have collected suggests a mental picture of C as a relatively small, thin set. For these reasons, the set C is often referred to as Cantor "dust." But there are some strong counterarguments that imply a very different picture. First, C is actually *uncountable*, with cardinality equal to the cardinality of $\mathbf{R}$. One slightly intuitive but convincing way to see this is to create a 1–1 correspondence between C and sequences of the form $(a_n)_{n=1}^{\infty}$, where $a_n = 0$ or 1. For each $c \in C$, set $a_1 = 0$ if c falls in the left-hand component

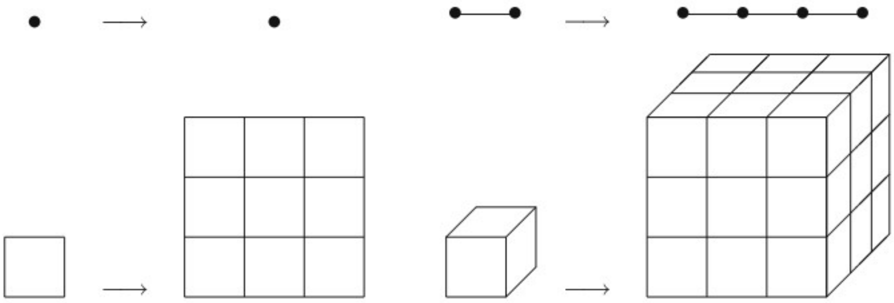

Figure 3.2: Magnifying sets by a factor of 3.

of C_1 and set $a_1 = 1$ if c falls in the right-hand component. Having established where in C_1 the point c is located, there are now two possible components of C_2 that might contain c. This time, we set $a_2 = 0$ or 1 depending on whether c falls in the left or right half of these two components of C_2. Continuing in this way, we come to see that every element $c \in C$ yields a sequence $(a_1, a_2, a_3, \ldots)$ of zeros and ones that acts as a set of directions for how to locate c within C. Likewise, every such sequence corresponds to a point in the Cantor set. Because the set of sequences of zeros and ones is uncountable (Exercise 1.6.4), we must conclude that C is uncountable as well.

What does this imply? In the first place, because the endpoints of the approximating sets C_n form a countable set, we are forced to accept the fact that not only are there other points in C but there are uncountably many of them. From the point of view of *cardinality*, C is quite large—as large as $\mathbf{R}$, in fact. This should be contrasted with the fact that from the point of view of *length*, C measures the same size as a single point. We conclude this discussion with a demonstration that from the point of view of *dimension*, C strangely falls somewhere in between.

There is a sensible agreement that a point has dimension zero, a line segment has dimension one, a square has dimension two, and a cube has dimension three. Without attempting a formal definition of dimension (of which there are several), we can nevertheless get a sense of how one might be defined by observing how the dimension affects the result of magnifying each particular set by a factor of 3 (Fig. 3.2). (The reason for the choice of 3 will become clear when we turn our attention back to the Cantor set). A single point undergoes no change at all, whereas a line segment triples in length. For the square, magnifying each length by a factor of 3 results in a larger square that contains 9 copies of the original square. Finally, the magnified cube yields a cube that contains 27 copies of the original cube within its volume. Notice that, in each case, to compute the "size" of the new set, the dimension appears as the exponent of the magnification factor.

	dim	$\times 3$	new copies
point	0	$\to$	$1 = 3^0$
segment	1	$\to$	$3 = 3^1$
square	2	$\to$	$9 = 3^2$
cube	3	$\to$	$27 = 3^3$
C	x	$\to$	$2 = 3^x$

Figure 3.3: DIMENSION OF C; $2 = 3^x \Rightarrow x = \log 2/\log 3$.

Now, apply this transformation to the Cantor set. The set $C_0 = [0, 1]$ becomes the interval $[0, 3]$. Deleting the middle third leaves $[0, 1] \cup [2, 3]$, which is where we started in the original construction except that we now stand to produce an additional copy of C in the interval $[2, 3]$. Magnifying the Cantor set by a factor of 3 yields *two* copies of the original set. Thus, if x is the dimension of C, then x should satisfy $2 = 3^x$, or $x = \log 2/\log 3 \approx .631$ (Fig. 3.3).

The notion of a noninteger or fractional dimension is the impetus behind the term "fractal," coined in 1975 by Benoit Mandlebrot to describe a class of sets whose intricate structures have much in common with the Cantor set. Cantor's construction, however, is over a hundred years old and for us represents an invaluable testing ground for the upcoming theorems and conjectures about the often elusive nature of subsets of the real line.

3.2 Open and Closed Sets

Given $a \in \mathbf{R}$ and $\epsilon > 0$, recall that the *ϵ-neighborhood* of a is the set

$$V_\epsilon(a) = \{x \in \mathbf{R} : |x - a| < \epsilon\}.$$

In other words, $V_\epsilon(a)$ is the open interval $(a - \epsilon, a + \epsilon)$, centered at a with radius ϵ.

Definition 3.2.1. A set $O \subseteq \mathbf{R}$ is *open* if for all points $a \in O$ there exists an ϵ-neighborhood $V_\epsilon(a) \subseteq O$.

Example 3.2.2. (i) Perhaps the simplest example of an open set is $\mathbf{R}$ itself. Given an arbitrary element $a \in \mathbf{R}$, we are free to pick any ϵ-neighborhood we like and it will always be true that $V_\epsilon(a) \subseteq \mathbf{R}$. It is also the case that the logical structure of Definition 3.2.1 requires us to classify the empty set $\emptyset$ as an open subset of the real line.

(ii) For a more useful collection of examples, consider the open interval

$$(c, d) = \{x \in \mathbf{R} : c < x < d\}.$$

Chapter 7

The Riemann Integral

7.1 Discussion: How Should Integration be Defined?

The Fundamental Theorem of Calculus is a statement about the inverse relationship between differentiation and integration. It comes in two parts, depending on whether we are differentiating an integral or integrating a derivative. Under suitable hypotheses on the functions f and F, the Fundamental Theorem of Calculus states that

(i) $\displaystyle\int_a^b F'(x)\,dx = F(b) - F(a)$ and

(ii) if $\displaystyle G(x) = \int_a^x f(t)\,dt$, then $G'(x) = f(x)$.

Before we can undertake any type of rigorous investigation of these statements, we need to settle on a definition for $\int_a^b f$. Historically, the concept of integration was defined as the inverse process of differentiation. In other words, the integral of a function f was understood to be a function F that satisfied $F' = f$. Newton, Leibniz, Fermat, and the other founders of calculus then went on to explore the relationship between antiderivatives and the problem of computing areas. This approach is ultimately unsatisfying from the point of view of analysis because it results in a very limited number of functions that can be integrated. Recall that every derivative satisfies the intermediate value property (Darboux's Theorem, Theorem 5.2.7). This means that any function with a jump discontinuity cannot be a derivative. If we want to define integration via antidifferentiation, then we must accept the consequence that a function as simple as

© Springer Science+Business Media New York 2015
S. Abbott, *Understanding Analysis*, Undergraduate Texts in Mathematics, DOI 10.1007/978-1-4939-2712-8_7

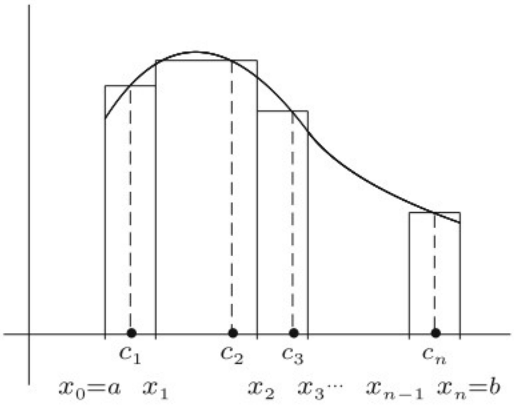

Figure 7.1: A RIEMANN SUM.

$$h(x) = \begin{cases} 1 & \text{for } 0 \le x < 1 \\ 2 & \text{for } 1 \le x \le 2 \end{cases}$$

is not integrable on the interval $[0, 2]$.

A very interesting shift in emphasis occurred around 1850 in the work of Cauchy, and soon after in the work of Bernhard Riemann. The idea was to completely divorce integration from the derivative and instead use the notion of "area under the curve" as a starting point for building a rigorous definition of the integral. The reasons for this were complicated. As we have mentioned earlier (Section 1.2), the concept of *function* was undergoing a transformation. The traditional understanding of a function as a holistic formula such as $f(x) = x^2$ was being replaced with a more liberal interpretation, which included such bizarre constructions as Dirichlet's function discussed in Section 4.1. Serving as a catalyst to this evolution was the budding theory of Fourier series (discussed in Section 8.5), which required, among other things, the need to be able to integrate these more unruly objects.

The Riemann integral, as it is called today, is the one usually discussed in introductory calculus. Starting with a function f on $[a, b]$, we partition the domain into small subintervals. On each subinterval $[x_{k-1}, x_k]$, we pick some point $c_k \in [x_{k-1}, x_k]$ and use the y-value $f(c_k)$ as an approximation for f on $[x_{k-1}, x_k]$. Graphically speaking, the result is a row of thin rectangles constructed to approximate the area between f and the x-axis. The area of each rectangle is $f(c_k)(x_k - x_{k-1})$, and so the total area of all of the rectangles is given by the *Riemann sum* (Fig. 7.1)

$$\sum_{k=1}^{n} f(c_k)(x_k - x_{k-1}).$$

Note that "area" here comes with the understanding that areas below the x-axis are assigned a negative value.

What should be evident from the graph is that the accuracy of the Riemann-sum approximation seems to improve as the rectangles get thinner. In some sense, we take the *limit* of these approximating Riemann sums as the width of the individual subintervals of the partitions tends to zero. This limit, if it exists, is Riemann's definition of $\int_a^b f$.

This brings us to a handful of questions. Creating a rigorous meaning for the limit just referred to is not too difficult. What will be of most interest to us—and was also to Riemann—is deciding what types of functions can be integrated using this procedure. Specifically, what conditions on f guarantee that this limit exists?

The theory of the Riemann integral turns on the observation that smaller subintervals produce better approximations to the function f. On each subinterval $[x_{k-1}, x_k]$, the function f is approximated by its value at some point $c_k \in [x_{k-1}, x_k]$. The quality of the approximation is directly related to the difference

$$|f(x) - f(c_k)|$$

as x ranges over the subinterval. Because the subintervals can be chosen to have arbitrarily small width, this means that we want $f(x)$ to be close to $f(c_k)$ whenever x is close to c_k. But this sounds like a discussion of continuity! We will soon see that the continuity of f is intimately related to the existence of the Riemann integral $\int_a^b f$.

Is continuity sufficient to prove that the Riemann sums converge to a well-defined limit? Is it necessary, or can the Riemann integral handle a discontinuous function such as $h(x)$ mentioned earlier? Relying on the intuitive notion of area, it would seem that $\int_0^2 h = 3$, but does the Riemann integral reach this conclusion? If so, how discontinuous can a function be before it fails to be integrable? Can the Riemann integral make sense out of something as pathological as Dirichlet's function on the interval $[0, 1]$?

A function such as

$$g(x) = \begin{cases} x^2 \ \sin(\frac{1}{x}) & \text{for } x \neq 0 \\ 0 & \text{for } x = 0 \end{cases}$$

raises another interesting question. Here is an example of a differentiable function, studied in Section 5.1, where the derivative $g'(x)$ is *not* continuous. As we explore the class of integrable functions, some attempt must be made to reunite the integral with the derivative. Having defined integration independently of differentiation, we would like to come back and investigate the conditions under which equations (i) and (ii) from the Fundamental Theorem of Calculus stated earlier hold. If we are making a wish list for the types of functions that we want to be integrable, then in light of equation (i) it seems desirable to expect this set to at least contain the set of derivatives. The fact that derivatives are not always continuous is further motivation not to content ourselves with an integral that cannot handle some discontinuities.

7.2 The Definition of the Riemann Integral

Although it has the benefit of some polish due to Darboux, the development of the integral presented in this chapter is closely related to the procedure just discussed. In place of Riemann sums, we will construct *upper sums* and *lower sums* (Fig. 7.2), and in place of a limit we will use a supremum and an infimum.

Throughout this section, it is assumed that we are working with a *bounded* function f on a closed interval $[a, b]$, meaning that there exists an $M > 0$ such that $|f(x)| \leq M$ for all $x \in [a, b]$.

Partitions, Upper Sums, and Lower Sums

Definition 7.2.1. A *partition* P of $[a, b]$ is a finite set of points from $[a, b]$ that includes both a and b. The notational convention is to always list the points of a partition $P = \{x_0, x_1, x_2, \ldots, x_n\}$ in increasing order; thus,

$$a = x_0 < x_1 < x_2 < \cdots < x_n = b.$$

For each subinterval $[x_{k-1}, x_k]$ of P, let

$$m_k = \inf\{f(x) : x \in [x_{k-1}, x_k]\} \quad \text{and} \quad M_k = \sup\{f(x) : x \in [x_{k-1}, x_k]\}.$$

The *lower sum* of f with respect to P is given by

$$L(f, P) = \sum_{k=1}^{n} m_k(x_k - x_{k-1}).$$

Likewise, we define the *upper sum* of f with respect to P by

$$U(f, P) = \sum_{k=1}^{n} M_k(x_k - x_{k-1}).$$

For a particular partition P, it is clear that $U(f, P) \geq L(f, P)$. The fact that this same inequality holds if the upper and lower sums are computed with respect to different partitions is the content of the next two lemmas.

Definition 7.2.2. A partition Q is a *refinement* of a partition P if Q contains all of the points of P; that is, if $P \subseteq Q$.

Lemma 7.2.3. *If $P \subseteq Q$, then $L(f, P) \leq L(f, Q)$, and $U(f, P) \geq U(f, Q)$.*

Proof. Consider what happens when we refine P by adding a single point z to some subinterval $[x_{k-1}, x_k]$ of P.

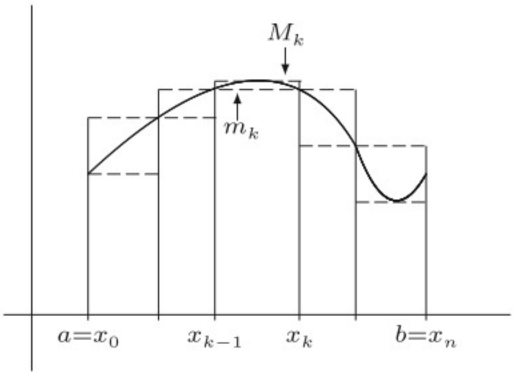

Figure 7.2: Upper and Lower Sums.

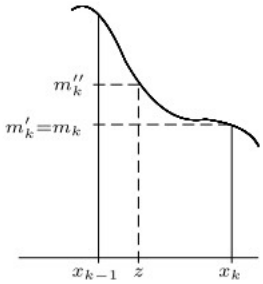

Focusing on the lower sum for a moment, we have

$$\begin{aligned} m_k(x_k - x_{k-1}) &= m_k(x_k - z) + m_k(z - x_{k-1}) \\ &\leq m_k'(x_k - z) + m_k''(z - x_{k-1}), \end{aligned}$$

where

$$m_k' = \inf\{f(x) : x \in [z, x_k]\} \quad \text{and} \quad m_k'' = \inf\{f(x) : x \in [x_{k-1}, z]\}$$

are each necessarily as large or larger than m_k.

By induction, we have $L(f,P) \leq L(f,Q)$, and an analogous argument holds for the upper sums. ☐

Lemma 7.2.4. *If P_1 and P_2 are any two partitions of $[a,b]$, then $L(f,P_1) \leq U(f,P_2)$.*

Proof. Let $Q = P_1 \cup P_2$ be the so-called *common refinement* of P_1 and P_2. Because $P_1 \subseteq Q$ and $P_2 \subseteq Q$, it follows that

$$L(f,P_1) \leq L(f,Q) \leq U(f,Q) \leq U(f,P_2).$$ ☐

Integrability

Intuitively, it helps to visualize a particular upper sum as an overestimate for the value of the integral and a lower sum as an underestimate. As the partitions get more refined, the upper sums get potentially smaller while the lower sums get potentially larger. A function is *integrable* if the upper and lower sums "meet" at some common value in the middle.

Rather than taking a limit of these sums, we will instead make use of the Axiom of Completeness and consider the *infimum* of the upper sums and the *supremum* of the lower sums.

Definition 7.2.5. Let $\mathcal{P}$ be the collection of all possible partitions of the interval $[a, b]$. The *upper integral* of f is defined to be

$$U(f) = \inf\{U(f, P) : P \in \mathcal{P}\}.$$

In a similar way, define the *lower integral* of f by

$$L(f) = \sup\{L(f, P) : P \in \mathcal{P}\}.$$

The following fact is not surprising.

Lemma 7.2.6. *For any bounded function f on $[a, b]$, it is always the case that $U(f) \geq L(f)$.*

Proof. Exercise 7.2.1. □

Definition 7.2.7 (Riemann Integrability). A bounded function f defined on the interval $[a, b]$ is *Riemann-integrable* if $U(f) = L(f)$. In this case, we define $\int_a^b f$ or $\int_a^b f(x)\,dx$ to be this common value; namely,

$$\int_a^b f = U(f) = L(f).$$

The modifier "Riemann" in front of "integrable" accurately suggests that there are other ways to define the integral. In fact, our work in this chapter will expose the need for a different approach, one of which is discussed in Section 8.1. In this chapter, the Riemann integral is the only method under consideration, so it will usually be convenient to drop the modifier "Riemann" and simply refer to a function as being "integrable."

Criteria for Integrability

To summarize the situation thus far, it is always the case for a bounded function f on $[a, b]$ that

$$\sup\{L(f, P) : P \in \mathcal{P}\} = L(f) \leq U(f) = \inf\{U(f, P) : P \in \mathcal{P}\}.$$

The function f is integrable if the inequality is an equality. The major thrust of our investigation of the integral is to describe, as best we can, the class of integrable functions. The preceding inequality reveals that integrability is really equivalent to the existence of partitions whose upper and lower sums are arbitrarily close together.

Theorem 7.2.8 (Integrability Criterion). *A bounded function f is integrable on $[a, b]$ if and only if, for every $\epsilon > 0$, there exists a partition P_ϵ of $[a, b]$ such that*

$$U(f, P_\epsilon) - L(f, P_\epsilon) < \epsilon.$$

Proof. Let $\epsilon > 0$. If such a partition P_ϵ exists, then

$$U(f) - L(f) \leq U(f, P_\epsilon) - L(f, P_\epsilon) < \epsilon.$$

Because ϵ is arbitrary, it must be that $U(f) = L(f)$, so f is integrable. (To be absolutely precise here, we could throw in a reference to Theorem 1.2.6.)

The proof of the converse statement is a familiar triangle inequality argument with parentheses in place of absolute value bars because, in each case, we know which quantity is larger. Because $U(f)$ is the greatest lower bound of the upper sums, we know that, given some $\epsilon > 0$, there must exist a partition P_1 such that

$$U(f, P_1) < U(f) + \frac{\epsilon}{2}.$$

Likewise, there exists a partition P_2 satisfying

$$L(f, P_2) > L(f) - \frac{\epsilon}{2}.$$

Now, let $P_\epsilon = P_1 \cup P_2$ be the common refinement. Keeping in mind that the integrability of f means $U(f) = L(f)$, we can write

$$\begin{aligned} U(f, P_\epsilon) - L(f, P_\epsilon) &\leq U(f, P_1) - L(f, P_2) \\ &< \left(U(f) + \frac{\epsilon}{2}\right) - \left(L(f) - \frac{\epsilon}{2}\right) \\ &= \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \end{aligned}$$

□

In the discussion at the beginning of this chapter, it became clear that integrability is closely tied to the concept of continuity. To make this observation more precise, let $P = \{x_0, x_1, x_2, \ldots, x_n\}$ be an arbitrary partition of $[a, b]$, and define $\Delta x_k = x_k - x_{k-1}$. Then,

$$U(f, P) - L(f, P) = \sum_{k=1}^{n} (M_k - m_k)\Delta x_k,$$

where M_k and m_k are the supremum and infimum of the function on the interval $[x_{k-1}, x_k]$, respectively. Our ability to control the size of $U(f, P) - L(f, P)$ hinges on the differences $M_k - m_k$, which we can interpret as the variation in the range of the function over the interval $[x_{k-1}, x_k]$. Restricting the variation of f over arbitrarily small intervals in $[a, b]$ is *precisely* what it means to say that f is uniformly continuous on this set.

Theorem 7.2.9. *If f is continuous on $[a, b]$, then it is integrable.*

Proof. Because f is continuous on a compact set, it must be bounded. It is also uniformly continuous for the same reason. This means that, given $\epsilon > 0$, there exists a $\delta > 0$ so that $|x - y| < \delta$ guarantees

$$|f(x) - f(y)| < \frac{\epsilon}{b-a}.$$

Now, let P be a partition of $[a, b]$ where $\Delta x_k = x_k - x_{k-1}$ is less than δ for every subinterval of P.

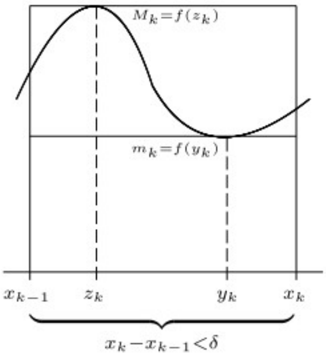

Given a particular subinterval $[x_{k-1}, x_k]$ of P, we know from the Extreme Value Theorem (Theorem 4.4.2) that the supremum $M_k = f(z_k)$ for some $z_k \in [x_{k-1}, x_k]$. In addition, the infimum m_k is attained at some point y_k also in the interval $[x_{k-1}, x_k]$. But this means $|z_k - y_k| < \delta$, so

$$M_k - m_k = f(z_k) - f(y_k) < \frac{\epsilon}{b-a}.$$

Finally,

$$U(f, P) - L(f, P) = \sum_{k=1}^{n} (M_k - m_k)\Delta x_k < \frac{\epsilon}{b-a} \sum_{k=1}^{n} \Delta x_k = \epsilon,$$

and f is integrable by the criterion given in Theorem 7.2.8. □

Exercises

Exercise 7.2.1. Let f be a bounded function on $[a, b]$, and let P be an arbitrary partition of $[a, b]$. First, explain why $U(f) \geq L(f, P)$. Now, prove Lemma 7.2.6.

Exercise 7.2.2. Consider $f(x) = 1/x$ over the interval $[1, 4]$. Let P be the partition consisting of the points $\{1, 3/2, 2, 4\}$.

(a) Compute $L(f,P)$, $U(f,P)$, and $U(f,P)-L(f,P)$.

(b) What happens to the value of $U(f,P)-L(f,P)$ when we add the point 3 to the partition?

(c) Find a partition P' of $[1,4]$ for which $U(f,P')-L(f,P')<2/5$.

Exercise 7.2.3 (Sequential Criterion for Integrability). (a) Prove that a bounded function f is integrable on $[a,b]$ if and only if there exists a sequence of partitions $(P_n)_{n=1}^{\infty}$ satisfying

$$\lim_{n\to\infty}\left[U(f,P_n)-L(f,P_n)\right]=0,$$

and in this case $\int_a^b f=\lim_{n\to\infty}U(f,P_n)=\lim_{n\to\infty}L(f,P_n)$.

(b) For each n, let P_n be the partition of $[0,1]$ into n equal subintervals. Find formulas for $U(f,P_n)$ and $L(f,P_n)$ if $f(x)=x$. The formula $1+2+3+\cdots+n=n(n+1)/2$ will be useful.

(c) Use the sequential criterion for integrability from (a) to show directly that $f(x)=x$ is integrable on $[0,1]$ and compute $\int_0^1 f$.

Exercise 7.2.4. Let g be bounded on $[a,b]$ and assume there exists a partition P with $L(g,P)=U(g,P)$. Describe g. Is g necessarily continuous? Is it integrable? If so, what is the value of $\int_a^b g$?

Exercise 7.2.5. Assume that, for each n, f_n is an integrable function on $[a,b]$. If $(f_n)\to f$ uniformly on $[a,b]$, prove that f is also integrable on this set. (We will see that this conclusion does not necessarily follow if the convergence is pointwise.)

Exercise 7.2.6. A *tagged partition* $(P,\{c_k\})$ is one where in addition to a partition P we choose a sampling point c_k in each of the subintervals $[x_{k-1},x_k]$. The corresponding *Riemann sum*,

$$R(f,P)=\sum_{k=1}^{n}f(c_k)\Delta x_k,$$

is discussed in Section 7.1, where the following definition is alluded to.

Riemann's Original Definition of the Integral: A bounded function f is *integrable* on $[a,b]$ with $\int_a^b f=A$ if for all $\epsilon>0$ there exists a $\delta>0$ such that for any tagged partition $(P,\{c_k\})$ satisfying $\Delta x_k<\delta$ for all k, it follows that

$$|R(f,P)-A|<\epsilon.$$

Show that if f satisfies Riemann's definition above, then f is integrable in the sense of Definition 7.2.7. (The full equivalence of these two characterizations of integrability is proved in Section 8.1.)

Exercise 7.2.7. Let $f:[a,b]\to\mathbf{R}$ be increasing on the set $[a,b]$ (i.e., $f(x)\leq f(y)$ whenever $x<y$). Show that f is integrable on $[a,b]$.

02.05.2020

Pág. 251(257) Chapter 12
Mechanical Curves –

D) Joseph Edwards – A Treatise on the Integral Calculus with Applications ... Volume 1 ... pdf
2 pdf

Source: Books_77_2020/
Calculus with Applications/

E) Books_77_2020/
Understanding Analysis/

Understanding Analysis (2015...)...pdf

Pág. 85(97) Chapter 3
Basic Topology of R
→ the Cantor Set

Pág. 215(224) Chapter 7
the Riemann Integral

F) ARQ_IMPORTANT_05/
Géométrie/
→ diretório todo!
→ Traité de Géométrie (2 Volumes)

obs. Está incompleto. Vou consertar!!! OK agora

* L/& . / $9I I $

-2

F O " <<<

7 I " 9I I

* "

7 I " 9I I F . 1 #

8 #$ $ = ! %
OQ & 3 (#3# " " OQ L 1
"

! ! "# # $ % "" & ' ("""
! ! !
)*+,)))-) ' .
/ "0 ! 12 &304 /+&*5 + &3 6!
-7 5 8 ! 999

7)*+,)))-) ' :%$""1%1; 6!
(% <

2 Fuzzy Controller Design

In this chapter we describe the basic definitions of fuzzy sets and operators on fuzzy sets. As they are used throughout the book it is necessary to start by introducing basic definitions of terms such as linguistic variables, fuzzy propositions, relations, implications, and inference engines. An emphasis is given to the description of the fuzzy controller structure that is most commonly used in practice, as well as to the several ways of defuzzification, that is, calculation of the crisp controller output value.

2.1 FUZZY SETS

We will use the following example to help us define a fuzzy set. Let A be a set of all integers *greater* than 10. We write

$$A = \{x\colon x \in \mathbb{N}, x > 10\} \tag{2.1}$$

Let B be a set of all integers *much greater* than 10. Mathematically, this statement can be written as

$$B = \{x\colon x \in \mathbb{N}, x \gg 10\} \tag{2.2}$$

The main difference between these two sets is that relation (2.1) completely defines set A, while relation (2.2) is not sufficient for a complete definition of set B. The reason is the vagueness of the term *much greater*. It is clear that 11, 12, 1178, and 2,075 are elements of set A. Most of the people will agree that 11,234 and 23^{10} undoubtedly belong to set B, but it is doubtful whether 15 or 50 are elements of B. The problem is how to determine the lowest integer which is *much greater* than 10.

This problem can be solved if one uses an alternative way of describing a set. According to traditional set theory, a set can be defined by its *characteristic function*. In other words, instead of individually declaring each element of a set we define a function that can take on values 1 or 0 depending on full membership or no membership of a particular element, respectively.

Definition 2.1 (Characteristic function, crisp set) Let S be a set from the domain X. A *characteristic function* of the set S attains value $\mu_S(x) = 1$ if $x \in S$, and $\mu_S(x) = 0$ if $x \notin S$, $\mu\colon X \to \{0, 1\}$. Set S with its characteristic function is called a *crisp set*.

Defined as it is, the characteristic function cannot describe set B, that is, it cannot cope with the vagueness in determining the lowest integer which would

© 2006 by Taylor & Francis Group, LLC

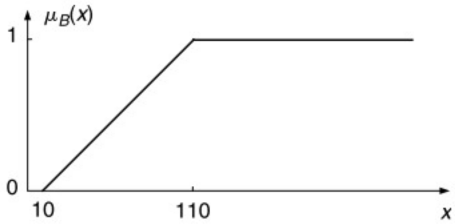

FIGURE 2.1 A graphical representation of a fuzzy set.

belong to set B. However, broadening the notion of a characteristic function offers an elegant way to define set B. Instead of determining the lowest integer belonging to set B, we may say that *all* integers greater than 10 belong to set B but with a different *membership degree*. The characteristic function, obtaining partial, or graded, values from the interval [0, 1], now becomes a *membership function*.

Definition 2.2 (Membership function and fuzzy set) Let F be a set from the domain X. A membership function $\mu_F(x)$ of set F is a function that assigns value, or *membership degree*, to every $x \in F$, $\mu\colon X \to [0,1]$. Then set F is called a *fuzzy set*.

Apparently, crisp sets may be treated as a special case of fuzzy sets since the characteristic function can assume only margin values from the interval [0, 1] on which membership function is defined.

Now we can completely define fuzzy set B as a set of pairs:

$$B = \{(\mu_B(x), x)\colon x \in \aleph\}$$

$$\mu_B(x) = \begin{cases} 0, & \text{for } x < 10 \\ \dfrac{x-10}{100}, & \text{for } 10 \le x \le 110 \\ 1, & \text{for } x > 110 \end{cases} \tag{2.3}$$

From the above definition we can see that numbers with membership degree 0 do not belong to fuzzy set B. Number 11 is an element of B with membership degree $\mu_B(11) = 0.01$, while membership degree of number 100 is $\mu_B(100) = 0.9$. Fuzzy set B is pictured in Figure 2.1.

From fuzzy set Definition 2.2, it follows that two fuzzy sets having the same elements will be equal only if their membership functions are equal. This means that each element belonging to one fuzzy set must belong to the other fuzzy set with the same membership degree.

Definition 2.3 (Fuzzy subset) Fuzzy set C is a *fuzzy subset* of fuzzy set B if

$$\forall x \in X\colon \mu_C(x) \le \mu_B(x) \tag{2.4}$$

© 2006 by Taylor & Francis Group, LLC

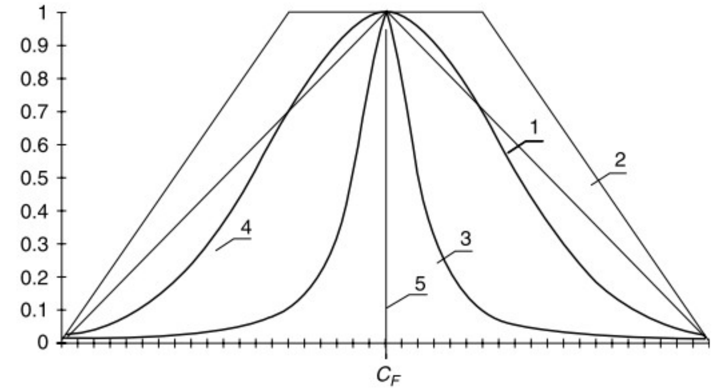

FIGURE 2.2 Typical shapes of membership functions: 1 — triangular, 2 — trapezoidal, 3 — Gaussian, 4 — bell-shaped, 5 — singleton.

In fuzzy sets theory, the range of possible quantitative values considered for fuzzy set members is called *universe of discourse*. Universe of discourse can be continuous or discrete. Discrete universe of discourse is normally bounded and contains a finite number of elements. A fuzzy set with discrete universe of discourse is called *a discrete fuzzy set*. The measure of fuzziness of each element is determined using a membership function spread either over a part or over the entire universe of discourse. As we have stated, the membership function converts the degree of fuzziness into the normalized interval [0, 1] where the boundary values 0 and 1 resemble the membership degrees of crisp set members. Membership functions can attain different forms. However, triangular, trapezoidal, Gaussian, and bell-shaped forms, shown in Figure 2.2, are used more than others:

$$\mu_F(x) = \begin{cases} 0, & \text{for } x < a \\ \dfrac{x-a}{b-a}, & \text{for } a \leq x < b \\ \dfrac{c-x}{c-b}, & \text{for } b \leq x \leq c \\ 0, & \text{for } x > c \end{cases} \quad \text{triangular}$$

$$\mu_F(x) = \begin{cases} 0, & \text{for } x < a \\ \dfrac{x-a}{b-a}, & \text{for } a \leq x < b \\ 1, & \text{for } b \leq x < c \\ \dfrac{d-x}{d-c}, & \text{for } c \leq x \leq d \\ 0, & \text{for } x > d \end{cases} \quad \text{trapezoidal}$$

$$\mu_F(x) = \mathrm{e}^{-(x-c_F)^2/w} \quad \text{Gaussian}, \qquad \mu_F(x) = \frac{1}{1+(x-c_F)^2} \quad \text{bell-shaped}$$

© 2006 by Taylor & Francis Group, LLC

These fuzzy sets are defined for variable x. Anticipating that different variables, for example, x and y, may have fuzzy sets with identical names (indices), it is convenient to introduce modified membership function notation $\mu_F^x = \mu_F(x)$. Such notation can, for example, discern notation for $\mu_F^y = \mu_F(y)$. This allows further generalizations, such as $\mu_i^x = \mu_i(x), i = 1, \ldots, l$.

Definition 2.4 (Center and nucleus of fuzzy set) A singular value $x = c_F = c_F^x \in F$ with the maximum degree of membership, $\mu_F(c_F) = 1$, is called the *center* of fuzzy set F. If there exists a set of values with the maximum degree of membership,

$$\mathrm{nuc}(F) = \{x \in X \colon \mu_F(x) = 1\} \tag{2.5}$$

then $\mathrm{nuc}(F)$ is called the *nucleus* of fuzzy set F.

The center of fuzzy set F with a nucleus is defined as $c_F^x = (x_a + x_b)/2$; where x_a and x_b are the boundaries of the nucleus. Fuzzy set F, having the center c_F^x as its only element, is a *fuzzy singleton* (Figure 2.2). In case we would like to replace a typical fuzzy set with a singleton (which is often done in the case of output fuzzy sets of fuzzy controllers), we must decide which value is representative and unique for different shapes of fuzzy sets, in order to be able to make such a replacement. Due to the fact that fuzzy sets can be either uniform or not, symmetrical or not, bounded or not — a good measure of their geometrical shape is the *centroid*, the point within a fuzzy set designating its center of gravity (COG). For example, the centroid of a triangular fuzzy set is the concurrence point of its three medians.

For a symmetrical fuzzy set the projection of its centroid on universe of discourse is equal to the center of the fuzzy set. That is why the center c_F^x is often referred to as the centroid. Fuzzy singletons are frequently used in fuzzy controller design since they reduce calculation efforts and provide shorter control intervals, which are desirable features in real-time control applications. We shall return to this issue in the chapters that follow.

Definition 2.5 (Adjacent fuzzy sets) Let two fuzzy sets, F and T, with centers c_F^x and c_T^x, $c_F^x < c_T^x$, be defined from the same universe of discourse X. If there does not exist a fuzzy set S, defined from X, with the center c_S^x, such that $c_F^x < c_S^x < c_T^x$, then F and T are called *adjacent fuzzy sets*.

The importance of adjacency will become apparent in the next paragraph where we describe properties of a set of rules formed by fuzzy sets.

Operations *union*, *intersection*, and *complement* are strictly defined on crisp sets. They are unambiguous because statements in traditional set theory are formed by *and*, *or*, and *not* operators which have well-defined semantics. In traditional set theory statement "$x \in B$ and $x \in C$" is true only if both declarations are true. In other words, a new set can be formed from sets B and C and x will belong to this new set only in case it is an element of both set B and set C. In fuzzy set theory the interpretation of statement "($x \in B$: $\mu_B(x) = 0.1$) and ($x \in C$: $\mu_C(x) = 0.3$)"

© 2006 by Taylor & Francis Group, LLC

is not that simple. Namely, it is not apparent how to determine the membership degree of x in the new fuzzy set formed by fuzzy sets B and C.

There are many different suggestions for determining the membership function of a fuzzy set that is the result of union, intersection, and complement of other fuzzy sets. Zadeh proposed the following definitions of these operations:

$$\begin{aligned}
&\mu_{B\cap C}(x) = \min(\mu_B(x), \mu_C(x)) \\
&\mu_{B\cup C}(x) = \max(\mu_B(x), \mu_C(x)) \\
&\mu_{\bar{B}}(x) = 1 - \mu_B(x)
\end{aligned} \tag{2.6}$$

According to (2.6), statement "($x \in B$: $\mu_B(x) = 0.1$) and ($x \in C$: $\mu_C(x) = 0.3$)" form a new fuzzy set $D = B \cap C$ with $\mu_D(x) = 0.1$.

It should be noted that the above definitions are also valid for crisp sets. If a membership function is replaced with a characteristic function that obtains only two values, 0 and 1, then Zadeh's operators will give the same results as standard *and*, *or*, and *not* operators.

In general, operators on fuzzy sets use *triangular norms* — a class of binary functions, which may be divided into *T-norms* (AND operators) and *S-norms* (OR operators) [1,2]. T-norms perform an *intersection* operation on fuzzy sets and have a particular importance in fuzzy logic control. T-norm is usually denoted as $T(a, b)$. S-norms represent a *union* operation denoted as $S(a, b)$.

Almost all T-norms used in fuzzy control applications can be derived from four basic T-norms listed below:

$$\begin{aligned}
&1.\quad T(\mu_B, \mu_C) = \min(\mu_B, \mu_C) \\
&2.\quad T(\mu_B, \mu_C) = \mu_B \cdot \mu_C \\
&3.\quad T(\mu_B, \mu_C) = \max(0, \mu_B + \mu_C - 1) \\
&4.\quad T(\mu_B, \mu_C) = \begin{cases} \mu_B, & \text{if } \mu_C = 1 \\ \mu_C, & \text{if } \mu_B = 1 \\ 0, & \text{if } \mu_B, \mu_C < 1 \end{cases}
\end{aligned} \tag{2.7}$$

The differences between these four T-norms are shown in the following example.

Example 2.1 Four basic T-norms.

We will now consider two fuzzy sets, B and C, which are defined as:

$$B = \{(\mu_B(x), x) \colon x \in \aleph\}, \qquad \mu_B(x) = \begin{cases} 0.1 \cdot x, & \text{for } x < 10 \\ \dfrac{20 - x}{10}, & \text{for } 10 \le x \le 19 \\ 0, & \text{for } x > 19 \end{cases}$$

© 2006 by Taylor & Francis Group, LLC

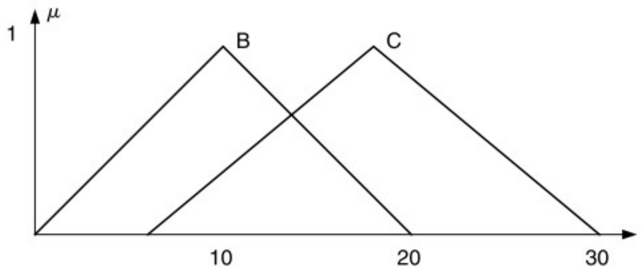

FIGURE 2.3 Fuzzy sets *B* and *C* considered in Example 2.1.

$$C = \{(\mu_C(x), x) \colon x \in \aleph\} \qquad \mu_C(x) = \begin{cases} 0, & \text{for } x < 7 \\ \dfrac{x-6}{12}, & \text{for } 7 \le x \le 18 \\ \dfrac{30-x}{12}, & \text{for } 19 \le x \le 29 \\ 0, & \text{for } x > 29 \end{cases}$$

Fuzzy sets *B* and *C* are pictured in Figure 2.3. Numerical values of the corresponding membership functions and results obtained after the application of four basic *T*-norms are given in Table 2.1.

Membership functions calculated in Table 2.1 for four different *T*-norms are represented graphically in Figure 2.4.

2.2 LINGUISTIC VARIABLES

In everyday communication we often use short sentences, which carry the same amount of information as their longer counterparts. When we say that "the car is far away" we actually mean that "the car's distance belongs to the *far away* (*very long*) category." Even if we knew that the distance was exactly 350 m, in everyday communication we would prefer saying that "the car is far away," as we would assume that there is a common understanding what a *very long distance* in traffic terms is. The term *distance* may attain two different values: numerical (350 m) and linguistic (far away). Variables, for which values are words or sentences, rather than numbers, are called *linguistic variables*. Continuing with our car analogy — the driver makes decisions about her/his actions based on imprecise linguistic qualifications of input information. Therefore, the strategy of driving a car is expressed in the form of IF–THEN rules, which contain linguistic variables: usually the names of inputs and outputs, rather than their concrete values (numbers). As Zadeh has stated, linguistic variables may assume different linguistic values over a specified universe of discourse. This means that linguistic values defined by an appropriate

© 2006 by Taylor & Francis Group, LLC

data 02 05 2020

Traité de Géométrie Descriptive
3 Volumes.

161) Books_77_2020/
Geometric_Transformations_for_3D/

A) Michael Mortenson - Geometric Transformation ....pdf

Obs. Livro todo!

B) Books_77_2020/
1Fuzzy Controller Design Theory and Applications/

084933747X.pdf

Pág. 11(24) Figure 2.2

Typical shapes of membership functions:
1- triangular, 2- trapezoidal,
3- Gaussian, 4- bell-shaped,
5- singleton.

Pág. 12(25)

Due to the fact that fuzzy sets can be either uniform or not, symmetrical or not, bounded or not - a good measure of their geometrical shape is the centroid the point within a fuzzy set designating its center of gravity (COG).

& '
* =>? # #

)*+,)))-) ' &? 0 6 >
@ 5 8 2 7 ! 4 8 A 6!
% 0

)*+,)))-) ' $$ ': 6!
B C6 ! @ 5 8

")*+,)))-) ' D ! , 6!
"$% % : 0 ! 8 . A

%)*+,)))-) ' 6!
-7 5 8 6 E 999

)*+,)))-) ' F**3> % 6!
F F !

')*+,)))-) ' * / ! . A 6!

7 * =>? > #

7 #@ F # 1"""% $; 6!

+ !

* *+,>?* & # #

)*+,)))-) ' * 6
*66 . A 0 8 7 6 7

)*+,)))-) ' *!8 ! * & !
* A 0 8 ? G 6 7

")*+,)))-) ' * A ! . A 7A
3 A 6!
($ ' : H

Let OP be drawn through the origin parallel to the given line so that l, m, n are the cosines of the angles which OP makes with OX, OY, OZ respectively. (*Refer Fig. 6*)

Let (x, y, z) be the co-ordinates of any point P on this line.

Let $\qquad OP=r.$

$\therefore \qquad x=lr,\ y=mr,\ z=nr.$

Squaring and adding, we obtain

$$x^2+y^2+z^2=(l^2+m^2+n^2)r^2.$$

But $\qquad x^2+y^2+z^2=OP^2=r^2.$

$$\therefore \qquad \mathbf{l^2+m^2+n^2=1.}$$

Cor. If a, b, c be three numbers *proportional* to the *actual* direction cosines l, m, n of a line, we have

$$\frac{l}{a}=\frac{m}{b}=\frac{n}{c}=\pm\frac{\sqrt{(l^2+m^2+n^2)}}{\sqrt{(a^2+b^2+c^2)}}=\pm\frac{1}{\sqrt{(a^2+b^2+c^2)}}.$$

$$\therefore \qquad l=\pm\frac{a}{\sqrt{(a^2+b^2+c^2)}},\quad m=\pm\frac{b}{\sqrt{(a^2+b^2+c^2)}},$$

$$n=\pm\frac{c}{\sqrt{(a^2+b^2+c^2)}},$$

where the same sign, positive or negative, is to be chosen throughout.

✓ **Direction Ratios.** From above, we see that a set of three numbers which are proportional to the actual direction cosines are sufficient to specify the direction of a line. Such numbers are called the **direction ratios.** Thus if a, b, c be the direction ratios of a line, its direction cosines are

$$\pm a/\sqrt{\Sigma a^2},\ \pm b/\sqrt{\Sigma a^2},\ \pm c/\Sigma a^2$$

Note. It is easy to see that if a line OP through the origin O makes angles α, β, γ with OX, OY, OZ, then the line OP obtained by producing OP backwards through O will make angles $\pi-\alpha$, $\pi-\beta$, $\pi-\gamma$ with OX, OY, OZ. Thus if

$$\cos\alpha=l,\ \cos\beta=m,\ \cos\gamma=n$$

are the direction cosines of OP, then

$$\cos(\pi-\alpha)=-l,\ \cos(\pi-\beta)=-m,\ \cos(\pi-\gamma)=-n$$

are the direction cosines of OP', *i.e.*, the line OP produced backwards.

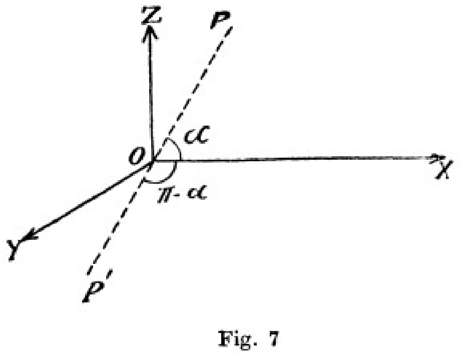

Fig. 7

Thus if we ignore the two senses of a line, we can think of the direction cosines l, m, n or $-l, -m, -n$ determining the direction of one and the same line. This explains the ambiguity in sign obtained above.

Note. The student should always make a distinction between direction cosines and direction ratios. It is only when l, m, n are direction cosines, that we have the relation

$$l^2+m^2+n^2=1.$$

Exercises

1. 6, 2, 3 are proportional to the direction cosines of a line. What are their actual values? [*Ans.* (6/7, 2/7, 3/7).

2. What are the direction cosines of lines equally inclined to the axes? How many such lines are there? [*Ans.* $(1/\sqrt{3}, \pm 1/\sqrt{3}, \pm 1/\sqrt{3})$; 4.

3. The co-ordinates of a point P are (3, 12, 4). Find the direction cosines of the line OP. [*Ans.* (3/13, 12/13, 4/13).

4. *The direction cosines l, m, n, of two lines are connected by the relations*

$$l+m+n=0 \qquad \ldots(i)$$

$$2lm+2ln-mn=0. \qquad \ldots(ii)$$

Find them?

Eliminating n between (i) and (ii); we get

$$2l^2-lm-m^2=0$$

or

$$2\left(\frac{l}{m}\right)^2-\frac{l}{m}-1=0. \qquad \ldots(iii)$$

This equation gives two values of l/m and hence there are two lines. The two roots of (iii) are 1 and $-\frac{1}{2}$.

If l_1, m_1, n_1 and l_2, m_2, n_2 be the direction cosines of two lines, we have

$$\frac{l_1}{m_1}=1 \text{ and } \frac{l_2}{m_2}=-\frac{1}{2}.$$

Also $\because l_1+m_1+n_1=0$ or $\dfrac{l_1}{m_1}+1+\dfrac{n_1}{m_1}=0$, $\quad\therefore \dfrac{n_1}{m_1}=-2,$

and $\because l_2+m_2+n_2=0$ or $\dfrac{l_2}{m_2}+1+\dfrac{n_2}{m_2}=0$, $\quad\therefore \dfrac{n_2}{m_2}=-\dfrac{1}{2}.$

$$\therefore \quad \frac{l_1}{1}=\frac{m_1}{1}=\frac{n_1}{-2}=\frac{1}{\sqrt{6}}, \text{ or } l_1=\frac{1}{\sqrt{6}}, \quad m_1=\frac{1}{\sqrt{6}}, \; n_1=-\frac{2}{\sqrt{6}}.$$

$$\text{and } \frac{l_2}{1}=\frac{m_2}{-2}=\frac{n_2}{1}=\frac{1}{\sqrt{6}}, \text{ or } l_2=\frac{1}{\sqrt{6}}, \; m_2=-\frac{2}{\sqrt{6}}, \; n_2=\frac{1}{\sqrt{6}}.$$

5. The direction cosines of two lines are determined by the relations

(i) $l-5m+3n=0, \; 7l^2+5m^2-3n^2=0$;

(ii) $l+m-n=0, \quad mn+6ln-12lm=0$;

find them?

[*Ans.* (i) $\dfrac{1}{\sqrt{14}}, \dfrac{2}{\sqrt{14}}, \dfrac{3}{\sqrt{14}}$; $-\dfrac{1}{\sqrt{6}}, \dfrac{1}{\sqrt{6}}, \dfrac{2}{\sqrt{6}}$

(ii) $\dfrac{1}{\sqrt{26}}, \dfrac{3}{\sqrt{26}}, \dfrac{4}{\sqrt{26}}$; $\dfrac{1}{\sqrt{14}}, \dfrac{2}{\sqrt{14}}, \dfrac{3}{\sqrt{14}}$.

1·8. Projection on a Straight line.

1·81. Projection of a point on a line. The foot of the perpendicular P from a given point A on a given straight line BC is called the *orthogonal projection* (or simply projection for the purpose of this book) of the point on the line and is the same point where the plane through the given point and perpendicular to the given line meets the line.

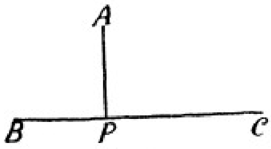

Fig. 8.

Thus in Fig. 1, page 1, A is the projection of P on X-axis ; also B and C are the projections of P on Y-axis and Z-axis respectively.

1·82. Projection of a segment of a line on another line. The projection of a segment AB of a line on any line CD is the segment $A'B'$ of CD where A', B' are the projections of A, B respectively on the line CD.

Clearly $A'B'$ is the intercept made on CD by planes through A, B each perpendicular to CD.

Ex. The co-ordinates of a point P are (x, y, z). What are the projections of OP on the co-ordinate axes ? [*Ans.* x, y, z.

Theorem. *The projection of a given segment AB of a line on any line CD is $AB \cos \theta$, where θ is the angle between AB and CD.*

Let the planes through A and B perpendicular to the line CD meet it in A', B' respectively so that $A'B'$ is the projection of AB.

Through A draw a line $AP \parallel CD$ to meet the plane through B at P.

Now, $AP \parallel CD$.

$\therefore$ $\angle PAB = \theta$.

Also BP lies in the plane which is $\perp AP$.

$\therefore$ $\angle APB = 90°$

Hence $AP = AB \cos \theta$

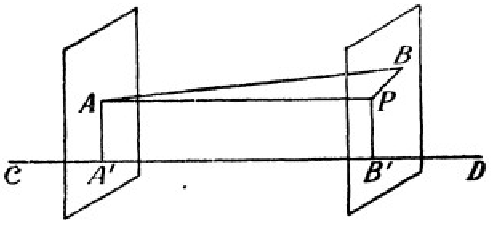

Fig. 9

Clearly $A'B'PA$ is a rectangle so that we have

$$AP = A'B'.$$

Hence $A'B' = AB \cos \theta$.

Cor. Direction cosines of the join of two points.

To find the direction cosines of the line joining the two points

$$P(x_1, y_1, z_1) \text{ and } Q(x_2, y_2, z_2).$$

Let L, M be the feet of the perpendiculars drawn from P, Q to the X-axis respectively so that

$$OL = x_1,\ OM = x_2.$$

Projection of PQ on X-axis $= LM$

$$= OM - OL$$

$$= x_2 - x_1.$$

Also if l, m, n be the direction cosines of PQ, the projection of PQ on X-axis $= l.PQ$.

$$\therefore \qquad l.PQ = x_2 - x_1.$$

Similarly projecting PQ on Y-axis and Z-axis, we get

$$m.PQ = y_2 - y_1,$$

$$n.PQ = z_2 - z_1.$$

PQ $= \sqrt{(x_2-x_1)^2+(y_2-}$

$$\therefore \qquad \frac{x_2 - x_1}{l} = \frac{y_2 - y_1}{m} = \frac{z_2 - z_1}{n} = PQ. \checkmark$$

Thus the direction cosines of the line joining the two points

(x_1, y_1, z_1) *and* (x_2, y_2, z_2)

are proportional to

$$x_2 - x_1,\ y_2 - y_1,\ z_2 - z_1.$$

Exercises

1. Find the direction cosines of the lines joining the points
   (*i*) (4, 3, −5) and (−2, 1, −8). [*Ans.* (6/7, 2/7, 3/7)
   (*ii*) (7, −5, 9) and (5, −3, 8). [*Ans.* (2/3, −2/3, 1/3)
2. Show that the points (1, −2, 3), (2, 3, −4), (0, −7, 10) are collinear.
3. The projections of a line on the axes are 12, 4, 3. Find the length and the direction cosines of the line. [*Ans.* 13 ; (12/13, 4/13, 3/13)

1·83. Projection of a broken line (*consisting of several continuous segments*). *If* $P_1, P_2, P_3, \ldots\ldots, P_n$ *be any number of points in space, then the sum of the projections of*

$$P_1P_2,\ P_2P_3, \ldots\ldots, P_{n-1}P_n$$

on any line is equal to the projection of P_1P_n *on the same line.*

Let $Q_1, Q_2, Q_3, \ldots\ldots, Q_n$ be the projections of the points

$$P_1, P_2, P_3, \ldots\ldots, P_n$$

on the given line. Then

$$Q_1Q_2 = \text{projection of } P_1P_2,$$

$$Q_2Q_3 = \quad ,, \quad ,, \ P_2P_3,$$

and so on.

Donated by
Mr. N. Sreeka
M.Sc.(Maths)

Also $\qquad Q_1Q_n = \text{projection of } P_1P_n.$

As $Q_1, Q_2, Q_3 \ldots\ldots, Q_n$ lie on the same line we have, for all relative positions of these points on the line, the relation

$$Q_1Q_2 + Q_2Q_3 + \ldots\ldots + Q_{n-1}Q_n = Q_1Q_n$$

Hence the theorem.

✓ **1·84. Projection of the join of two points on a line.** *To show that the projection of the line joining*

$$P(x_1, y_1, z_1) \text{ and } Q(x_2, y_2, z_2)$$

on a line with direction cosines l, m, n is

$$(x_2-x_1)l+(y_2-y_1)m+(z_2-z_1)n.$$

Through P, Q draw planes parallel to the co-ordinate planes to form a rectangular parallelopiped whose one diagonal is PQ. (*See Fig. 3, Page 3*).

Now

$$PA=x_2-x_1,\ AN=y_2-y_1,\ NQ=z_2-z_1.$$

The lines PA, AN, NQ are respectively parallel to x-axis, y-axis, z-axis. Therefore, their respective projections on the line with direction cosines l, m, n are

$$(x_2-x_1)l,\ (y_2-y_1)m,\ (z_2-z_1)n.$$

As the projection of PQ on any line is equal to the sum of the projections of PA, AN, NQ on that line, therefore the required projection is

$$(\mathbf{x_2}-\mathbf{x_1})\mathbf{l}+(\mathbf{y_2}-\mathbf{y_1})\mathbf{m}+(\mathbf{z_2}-\mathbf{z_1})\mathbf{n}.$$

Exercises

1. $A(6, 3, 2)$, $B(5, 1, 4)$, $C(3, -4, 7)$, $D(0, 2, 5)$ are four points. Find the projections of AB on CD and of CD on AB. [*Ans.* $-13/7$; $-13/3$.

2. Show by projection that if P, Q, R, S are the points $(6, -6, 0)$, $(-1, -7, 6)$, $(3, -4, 4,)$ $(2, -9, 2)$ respectively then $PQ \perp RS$.

✓ **1·9. Angle between two lines.** *To find the angle between lines whose direction cosines are (l_1, m_1, n_1) and (l_2, m_2, n_2).*

Let OP_1, OP_2, be lines through the origin parallel to the given lines so that the cosines of the angles which OP_1 and OP_2 make with the axes are l_1, m_1, n_1 and l_2, m_2, n_2, respectively and the angle between the given lines is the angle between OP_1 and OP_2. Let this angle be θ.

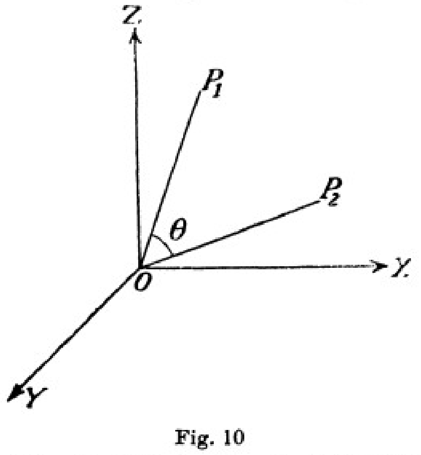

Fig. 10

Let the co-ordinates of P_2 be (x_2, y_2, z_2).

The projection of the line OP_2 joining

$$O(0, 0, 0) \text{ and } P_2(x_2, y_2, z_2)$$

on the line OP_1 whose direction cosines are

$$l_1, m_1, n_1,$$

is $$(x_2-0)l_1+(y_2-0)m_1+(z_2-0)n_1=l_1x_2+m_1y_2+n_1z_2$$

Also this projection is $OP_2 \cos\theta$.

$\therefore$ $$OP_2 \cos\theta=l_1x_2+m_1y_2+n_1z_2.$$

But $$x_2=l_2.OP_2,\ y_2=m_2.OP_2,\ z_2=n_2.OP_2. \qquad (\S 1{\cdot}61)$$

$\therefore$ $$OP_2 \cos\theta=(l_1l_2+m_1m_2+n_1n_2)OP_2$$

or $$\cos\theta=\mathbf{l_1l_2+m_1m_2+n_1n_2}.$$

Second Method. Suppose $OP_1=r_1$, $OP_2=r_2$.

Let the co-ordinates of P_1, P_2, be (x_1, y_1, z_1) and (x_2, y_2, z_2) respectively.

Then $$x_1=r_1l_1,\quad y_1=r_1m_1,\quad z_1=r_1n_1, \qquad (\S 1{\cdot}61)$$

and $$x_2=r_2l_2,\quad y_2=r_2m_2,\quad z_2=r_2n_2.$$

We have

$$\begin{aligned}P_1P_2^2&=(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2\\&=(x_2^2+y_2^2+z_2^2)+(x_1^2+y_1^2+z_1^2)-2(x_1x_2+y_1y_2+z_1z_2)\\&=r_2^2+r_1^2-2r_2r_1(l_1l_2+m_1m_2+n_1n_2). \qquad \ldots(i)\end{aligned}$$

Also from Trigonometry, we have

$$P_1P_2^2=r_1^2+r_2^2-2r_1r_2\cos\theta \qquad \ldots(ii)$$

Therefore, from (i) and (ii), we obtain

$$\begin{aligned}r_1^2+r_2^2-2r_1r_2\cos\theta&=P_1P_2^2\\&=r_1^2+r_2^2-2r_1r_2(l_1l_2+m_1m_2+n_1n_2)\end{aligned}$$

i.e., $$\cos\theta=l_1l_2+m_1m_2+n_1n_2.$$

Cor. 1. Sin θ and **tan θ**. The expressions for sin θ and tan θ in a convenient form are obtained as follows :—

$$\begin{aligned}\sin^2\theta&=1-\cos^2\theta\\&=1-(l_1l_2+m_1m_2+n_1n_2)^2\\&=(l_1^2+m_1^2+n_1^2)(l_2^2+m_2^2+n_2^2)-(l_1l_2+m_1m_2+n_1n_2)^2\\&=(l_1m_2-l_2m_1)^2+(m_1n_2-m_2n_1)^2+(n_1l_2-n_2l_1)^2.\end{aligned}$$

$\therefore$ $$\sin\theta=\pm\sqrt{[\Sigma(l_1m_2-l_2m_1)^2]}$$

and $$\tan\theta=\frac{\sin\theta}{\cos\theta}=\pm\frac{\sqrt{[\Sigma(l_1m_2-l_2m_1)^2]}}{\Sigma l_1l_2}.$$

Cor. 2. If the direction cosines of two lines be proportional to a_1, b_1, c_1, and a_2, b_2, c_2, then their actual values are

$$\pm\frac{a_1}{\sqrt{(a_1^2+b_1^2+c_1^2)}},\ \pm\frac{b_1}{\sqrt{(a_1^2+b_1^2+c_1^2)}},\ \pm\frac{c_1}{\sqrt{(a_1^2+b_1^2+c_1^2)}}\ ;$$

$$\pm\frac{a_2}{\sqrt{(a_2^2+b_2^2+c_2^2)}},\ \pm\frac{b_2}{\sqrt{(a_2^2+b_2^2+c_2^2)}},\ \pm\frac{c_2}{\sqrt{(a_2^2+b_2^2+c_2^2)}},$$

so that if θ be the angle between the given lines, we have

$$\cos\theta = \pm\frac{a_1a_2+b_1b_2+c_1c_2}{\sqrt{(a_1^2+b_1^2+c_1^2)}\sqrt{(a_2^2+b_2^2+c_2^2)}},$$

$$\sin\theta = \pm\frac{\sqrt{[(a_1b_2-a_2b_1)^2+(b_1c_2-b_2c_1)^2+(c_1a_2-c_2a_1)^2]}}{\sqrt{(a_1^2+b_1^2+c_1^2)}\sqrt{(a_2^2+b_2^2+c_2^2)}},$$

$$\tan\theta = \pm\frac{\sqrt{[\Sigma(a_1b_2-a_2b_1)^2]}}{\Sigma a_1a_2}.$$

The expression for tan θ is of the same form whether we use direction cosines or direction ratios.

Cor. 3. Conditions for perpendicularity and parallelism.

(*i*) When the given lines are perpendicular,

$$\theta=90° \text{ so that } \cos\theta=0.$$

This gives

$$\mathbf{a_1a_2+b_1b_2+c_1c_2=0}.$$

(*ii*) When the given lines are parallel,

$$\theta=0 \text{ so that } \sin\theta=0.$$

This gives

$$(a_1b_2-a_2b_1)^2+(b_1c_2-b_2c_1)^2+(c_1a_2-c_2a_1)^2=0,$$

which is true only when

$$a_1b_2-a_2b_1=0,\ b_1c_2-b_2c_1=0,\ c_1a_2-c_2a_1=0,$$

or

$$\frac{\mathbf{a_1}}{\mathbf{a_2}}=\frac{\mathbf{b_1}}{\mathbf{b_2}}=\frac{\mathbf{c_1}}{\mathbf{c_2}}.$$

This result is also otherwise evident, for, the lines through the origin drawn parallel to the parallel lines coincide and, therefore, their direction cosines must be the same and hence direction ratios proportional.

Exercises

1. Find the angles between the lines whose direction ratios are

(*i*) 5, −12, 13 ; −3, 4, 5. [*Ans.* $\cos^{-1}(1/65)$.

(*ii*) 1, 1, 2 ; $\sqrt{3}-1, -\sqrt{3}-1, 4$. [*Ans.* $\pi/3$.

2. Show that the angle between the lines whose direction cosines are given by the relations in Ex. 4, P. 9 is $\frac{1}{3}\pi$.

3. *Find the direction cosines of the line which is perpendicular to the lines with direction cosines proportional to (1, −2, −2), (0, 2, 1).*

Sol. If l, m, n be the direction cosines of the line perpendicular to the given lines, we have

$$l.1+m(-2)+n(-2)=0,\ i.e.,\ l-2m-2n=0,$$
$$l(0)+m(2)+n(1)=0,\ i.e.,\ 0l+2m+n=0.$$

These give

$$\frac{l}{2}=\frac{m}{-1}=\frac{n}{2}.$$

$$\therefore\quad l=\frac{2}{\sqrt{[2^2+(-1)^2+2^2]}}=\frac{2}{3},\ m=-\frac{1}{3},\ n=\frac{2}{3}.$$

4. Show that a line can be found perpendicular to the three lines with direction cosines proportional to (2, 1, 5), (4, −2, 2), (−6, 4, −1). Hence show that if these three lines be concurrent, they are also coplanar.

5. *l_1, m_1, n_1 ; l_2, m_2, n_2 are the direction cosines of two mutually perpendicular lines. Show that the direction cosines of the line perpendicular to them both are*

$$m_1n_2-m_2n_1,\ n_1l_2-n_2l_1,\ l_1m_2-l_2m_1.$$

Sol. If l, m, n be the direction cosines of the required line, we have

$$ll_1+mm_1+nn_1=0,$$
$$ll_2+mm_2+nn_2=0.$$

These give

$$\frac{l}{m_1n_2-m_2n_1}=\frac{m}{n_1l_2-n_2l_1}=\frac{n}{l_1m_2-l_2m_1}=\frac{\sqrt{\Sigma l^2}}{\sqrt{[\Sigma(m_1n_2-m_2n_1)^2]}}=\frac{1}{\sin\theta},$$

where θ is the angle between the given lines. As $\theta=90°$, we have $\sin\theta=1$. Hence the result.

6. l_1, m_1, n_1 and l_2, m_2, n_2 are the direction ratios of two intersecting lines. Show that lines through the intersection of these two with direction ratios

$$l_1+kl_2,\ m_1+km_2,\ n_1+kn_2$$

are coplanar with them ; k being any number whatsoever.

(Show that they all have a common perpendicular direction.)

7. Show that three concurrent lines with direction cosines

$$(l_1, m_1, n_1),\ (l_2, m_2, n_2),\ (l_3, m_3, n_3)$$

are coplanar if.

$$\begin{vmatrix} l_1, & m_1, & n_1 \\ l_2, & m_2, & n_2 \\ l_3, & m_3, & n_3 \end{vmatrix}=0.$$

8. Show that the join of points (1, 2, 3), (4, 5, 7) is parallel to the join of the points (−4, 3, −6), (2, 9, 2).

9. Show that the points

(4, 7, 8) ; (2, 3, 4) ; (−1, −2, 1) ; (1, 2, 5)

are the vertices of a parallelogram.

10. Show that the points

(5, −1, 1), (7, −4, 7), (1, −6, 10), (−1, −3, 4)

are the vertices of a rhombus.

11. Show that the points.

(0, 4, 1), (2, 3, −1), (4, 5, 0), (2, 6, 2)

are the vertices of a square.

✓**12.** $A(1, 8, 4)$, $B(0, -11, 4)$, $C(2, -3, 1)$ are three points and D is the foot of the perpendicular from A on BC. Find the co-ordinates of D.

[*Ans.* (4, 5, −2).

13. Find the point in which the join of (−9, 4, 5) and (11, 0, −1) is met by the perpendicular from the origin. [*Ans.* (1, 2, 2).

14. $A(-1, 2, -3)$, $B(5, 0, -6)$, $C(0, 4, -1)$ are three points. Show that the direction cosines of the bisectors of the angle BAC are proportional to (25, 8, 5) and (−11, 20, 23).

[*Hint.*—Find the co-ordinates of the points which divide BC in the ratio $AB : AC$.]

15. Find the angle between the lines whose direction cosines are given by the equations $3l+m+5n=0$ and $6mn-2nl+5lm=0$. [*Ans.* $\cos^{-1}\frac{1}{6}$.

16. Show that the pair of lines whose direction cosines are given by $3lm-4ln+mn=0$, $l+2m+3n=0$ are perpendicular.

17. *Show that the straight lines whose direction cosines are given by the equations*

$$al+bm+cn=0,\ ul^2+vm^2+wn^2=0$$

are perpendicular or parallel according as

$$a^2(v+w)+b^2(w+u)+c^2(u+v)=0 \quad \text{or} \quad a^2/u+b^2/v+c^2/w=0,$$

Sol. Eliminating l, between the given relations, we have

$$\frac{u(bm+cn)^2}{a^2}+vm^2+wn^2=0$$

or
$$(b^2u+a^2v)m^2+2ubcmn+(c^2u+a^2w)n^2=0 \qquad \ldots(i)$$

If the lines be parallel, their direction cosines are equal so that the two values of m/n must be equal. The condition for this is

$$u^2b^2c^2=(b^2u+a^2v)(c^2u+a^2w)$$

or

$$\frac{a^2}{u}+\frac{b^2}{v}+\frac{c^2}{w}=0.$$

Again, if l_1, m_1, n_1 and l_2, m_2, n_2 be the direction cosines of the two lines, then equation (i) gives

$$\frac{m_1}{n_1}\cdot\frac{m_2}{n_2}=\frac{m_1m_2}{n_1n_2}=\frac{c^2u+a^2w}{b^2u+a^2v}$$

or

$$\frac{m_1m_2}{c^2u+a^2w}=\frac{n_1n_2}{b^2u+a^2v}$$

or

Similarly the elimination of n, gives, (or by symmetry)

$$\frac{l_1l_2}{b^2w+c^2v}=\frac{m_1m_2}{a^2w+c^2u}$$

$$\therefore \qquad \frac{l_1l_2}{b^2w+c^2v}=\frac{m_1m_2}{a^2w+c^2u}=\frac{n_1n_2}{b^2u+a^2v}=k, \text{ say}$$

or

$$l_1l_2+m_1m_2+n_1n_2=k(b^2w+c^2v+a^2w+c^2u+b^2u+a^2v).$$

For perpendicular lines

$$l_1l_2+m_1m_2+n_1n_2=0$$

Thus the condition for perpendicularity is

$$a^2(v+w)+b^2(w+u)+c^2(u+v)=0.$$

18. Show that the straight lines whose direction cosines are given by

$$al+bm+cn=0,\ fmn+gnl+hlm=0,$$

are perpendicular if

$$f/a+g/b+h/c=0,$$

and parallel if

$$\sqrt{af}\pm\sqrt{bg}\pm\sqrt{ch}=0.$$

19. *l_1, m_1, n_1 ; l_2, m_2, n_2 are the direction cosines of two concurrent lines. Show that the direction cosines of the lines bisecting the angles between them are proportional to*

$$l_1\pm l_2,\ m_1\pm m_2,\ n_1\pm n_2.$$

Sol. Let the lines concur at the origin O and let OA, OB be the two lines. Take points A, B on the two lines such that $OA=OB=r$, say. Also take a point A' on AO produced such that $AO=OA'$. Let C, C' be the mid-points of AB and $A'B$. Then OC, OC' are the required bisectors. The result, now, follows from the fact that the co-ordinates of A, B, A' respectively are

$$(l_1r,\ m_1r,\ n_1r)\ ;\ (l_2r,\ m_2r,\ n_2r)\ ;\ (-l_1r,\ -m_1r,\ -n_1r),$$

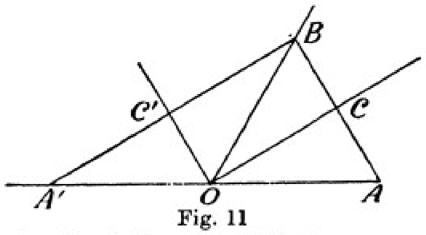

Fig. 11

20. *If the edges of a rectangular parallelopiped are a, b, c show that the angles between the four diagonals are given by*

$$\cos^{-1}[(\pm a^2 \pm b^2 \pm c^2)/(a^2+b^2+c^2)].$$

Sol. Take one of the vertices O of the parallelopiped as origin and the three rectangular faces through it as the three rectangular co-ordinate planes. (*See Fig. 1, Page 1*).

$$\text{Let } OA=a,\ OB=b,\ OC=c$$

The lines OP, AL, BM, CN are the four diagonals.

The co-ordinates of A, B, C are $(a, 0, 0)$; $(0, b, 0)$; $(0, 0, c)$.

" L, M, N are $(0, b, c)$; $(a, 0, c)$; $(a, b, 0)$.

" O, P are $(0, 0, 0)$; (a, b, c).

Direction cosines of OP are $\dfrac{a}{\sqrt{\Sigma a^2}}, \dfrac{b}{\sqrt{\Sigma a^2}}, \dfrac{c}{\sqrt{\Sigma a^2}}$;

Direction cosines of AL are $\dfrac{-a}{\sqrt{\Sigma a^2}}, \dfrac{b}{\sqrt{\Sigma a^2}}, \dfrac{c}{\sqrt{\Sigma a^2}}$;

" " " BM are $\dfrac{a}{\sqrt{\Sigma a^2}}, \dfrac{-b}{\sqrt{\Sigma a^2}}, \dfrac{c}{\sqrt{\Sigma a^2}}$;

" " " CN are $\dfrac{a}{\sqrt{\Sigma a^2}}, \dfrac{b}{\sqrt{\Sigma a^2}}, \dfrac{-c}{\sqrt{\Sigma a^2}}$.

The angle between OP and CN, therefore, is

$$\cos^{-1} \frac{a^2+b^2-c^2}{a^2+b^2+c^2}.$$

Similarly the angle between any one of the six pairs of diagonals can be found.

21. A line makes angles α, β, γ, δ, with the four diagonals of a cube ; prove that $\cos^2\alpha+\cos^2\beta+\cos^2\gamma+\cos^2\delta=4/3$. *(P.U. 1932)*

(Choose axes as in Ex. 20 above and suppose that the direction cosines of the given line are l, m, n.)

22. *O, A, B, C, are four points not lying in the same plane and such that $OA \perp BC$ and $OB \perp CA$. Prove that $OC \perp AB$. What well-known theorem does this become if four points are co-planar ?*

The result of this example may also be stated thus :—

"If two pairs of opposite edges of a tetrahedron be at right angles, then so is the third."

Take O as origin and any three mutually perpendicular lines through O as co-ordinate axes.

Let (x_1, y_1, z_1), (x_2, y_2, z_2), (x_3, y_3, z_3) be the co-ordinates of the points A, B, C respectively.

As $OA \perp BC$, we have

$$x_1(x_2-x_3)+y_1(y_2-y_3)+z_1(z_2-z_3)=0, \qquad \ldots(i)$$

As $OB \perp CA$, we have

$$x_2(x_3-x_1)+y_2(y_3-y_1)+z_2(z_3-z_1)=0. \qquad \ldots(ii)$$

Adding (i) and (ii), we obtain

$$x_3(x_2-x_1)+y_3(y_2-y_1)+z_3(z_2-z_1)=0$$

which shows that $OC \perp AB$.

23. If, in a tetrahedron $OABC$,

$$OA^2+BC^2=OB^2+CA^2=OC^2+AB^2,$$

then its pairs of opposite edges are at right angles.

24. l_1, m_1, n_1 and l_2, m_2, n_2 are two directions inclined at an angle φ, to each other. Show that the direction

$$\frac{l_1+l_2}{2\cos\frac{1}{2}\varphi}, \frac{m_1+m_2}{2\cos\frac{1}{2}\varphi}, \frac{n_1+n_2}{2\cos\frac{1}{2}\varphi}$$

bisects the angle between these two directions.

Show that these direction cosines are the actual values.

25. Show that the direction equally inclined to the three mutually perpendicular directions

$$l_1, m_1, n_1\ ;\ l_2, m_2, n_2\ ;\ l_3, m_3, n_3,$$

is given by the direction cosines

$$\frac{l_1+l_2+l_3}{\sqrt{3}},\ \frac{m_1+m_2+m_3}{\sqrt{3}},\ \frac{n_1+n_2+n_3}{\sqrt{3}}.$$

26. Show that the area of the triangle whose vertices are the origin and the points (x_1, y_1, z_1), and (x_2, y_2, z_2) is

$$\tfrac{1}{2}\sqrt{[(y_1z_2-y_2z_1)^2+(z_1x_2-z_2x_1)^2+(x_1y_2-x_2y_1)^2]}. \qquad \textit{(B.U. 1958)}$$

27. $l_1, m_1, n_1\ ;\ l_2, m_2, n_2\ ;\ l_3, m_3, n_3$

are the direction cosines of three mutually perpendicular lines ; show that

$$l_1, l_2, l_3\ ;\ m_1, m_2, m_3\ ;\ n_1, n_2, n_3$$

are also the direction cosines of three mutually perpendicular lines. Hence show that

$$\begin{aligned}
l_1^2+m_1^2+n_1^2&=1, & l_1^2+l_2^2+l_3^2&=1,\\
l_2^2+m_2^2+n_2^2&=1, & m_1^2+m_2^2+m_3^2&=1,\\
l_3^2+m_3^2+n_3^2&=1, & n_1^2+n_2^2+n_3^2&=1,\\
l_1l_2+m_1m_2+n_1n_2&=0, & l_1m_1+l_2m_2+l_3m_3&=0,\\
l_2l_3+m_2m_3+n_2n_3&=0, & m_1n_1+m_2n_2+m_3n_3&=0,\\
l_3l_1+m_3m_1+n_3n_1&=0, & n_1l_1+n_2l_2+n_3l_3&=0.
\end{aligned}$$

")+,)))-) ' * A . A 0 ! "0
7A + 6 7

" ")*+,)))-) ' * F / A 2 I + 7
*!8 ! I A 6!
(:% = "

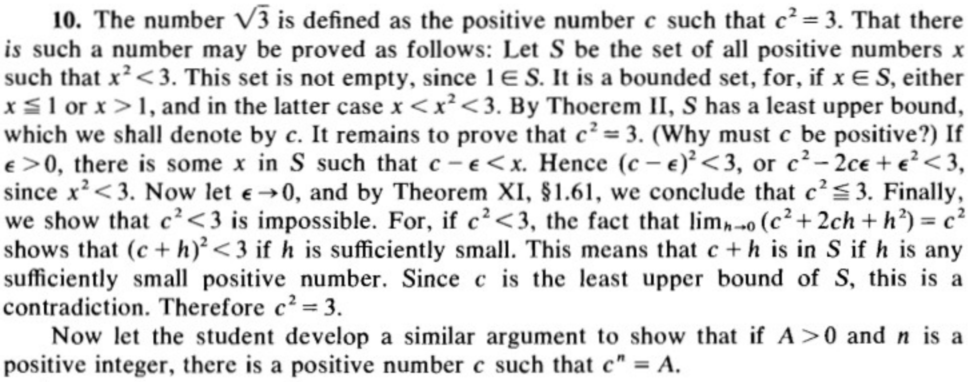

10. The number $\sqrt{3}$ is defined as the positive number c such that $c^2 = 3$. That there *is* such a number may be proved as follows: Let S be the set of all positive numbers x such that $x^2 < 3$. This set is not empty, since $1 \in S$. It is a bounded set, for, if $x \in S$, either $x \leqq 1$ or $x > 1$, and in the latter case $x < x^2 < 3$. By Thoerem II, S has a least upper bound, which we shall denote by c. It remains to prove that $c^2 = 3$. (Why must c be positive?) If $\epsilon > 0$, there is some x in S such that $c - \epsilon < x$. Hence $(c-\epsilon)^2 < 3$, or $c^2 - 2c\epsilon + \epsilon^2 < 3$, since $x^2 < 3$. Now let $\epsilon \to 0$, and by Theorem XI, §1.61, we conclude that $c^2 \leqq 3$. Finally, we show that $c^2 < 3$ is impossible. For, if $c^2 < 3$, the fact that $\lim_{h \to 0} (c^2 + 2ch + h^2) = c^2$ shows that $(c+h)^2 < 3$ if h is sufficiently small. This means that $c + h$ is in S if h is any sufficiently small positive number. Since c is the least upper bound of S, this is a contradiction. Therefore $c^2 = 3$.

Now let the student develop a similar argument to show that if $A > 0$ and n is a positive integer, there is a positive number c such that $c^n = A$.

" %)*+,)))-) ' ! 7A F
I G G 6!

)*+,)))-) ' ! * A . A2 3
F! 7A . ? / 2 + 5 < A 6!

")*+,)))-) ' 0 !
*66 ! 6 1 6!

" ')*+,)))-) ' < - 6 C 7 7A
J 6!

" 1)*+,)))-) ' . A 6!
* A ! * 7 . A

" :)*+,)))-) ' @ D J F A
> * & *66 6!

" $)*+,)))-) ' @ F
7A ? . 6!

")*+,)))-) ' & "0 . A
! ? 7 ! &&/ DFF ! *!8 ! + H G !
. @ 7A + H G ! 6!

")*+,)))-) ' G 6! > A A
! ! "! 6 8 6!

")*+,)))-) ' 7A J
6!

6
)*+,)))-) ' 5 * 7 !
. H / 4 8 A 1 6!

)*+,)))-) ' 2 5 * 7 2 !
0 < * 4 ! *66 C F! 6!

)*+,)))-) ' I @ <
" !H8

%)*+,)))-) '
* A F ! 6 6!

(" " 6 F ! 6

2

Euclidean Space

Euclidean space is a mathematical construct that encompasses the line, the plane, and three-dimensional space as special cases. Its elements are called *vectors*. Vectors can be understood in various ways: as arrows, as quantities with magnitude and direction, as displacements, or as points. However, along with a sense of what vectors are, we also need to emphasize how they interact. The axioms in Section 2.1 capture the idea that vectors can be added together and can be multiplied by scalars, with both of these operations obeying familiar laws of algebra. Section 2.2 expresses the geometric ideas of length and angle in Euclidean space in terms of vector algebra. Section 2.3 discusses continuity for functions (also called *mappings*) whose inputs and outputs are vectors rather than scalars. Section 2.4 introduces a special class of sets in Euclidean space, the *compact* sets, and shows that compact sets are preserved under continuous mappings.

2.1 Algebra: Vectors

Let n be a positive integer. The set of all ordered n-tuples of real numbers,

$$\mathbb{R}^n = \{(x_1, \ldots, x_n) : x_i \in \mathbb{R} \text{ for } i = 1, \ldots, n\},$$

constitutes **n-dimensional Euclidean space**. When $n = 1$, the parentheses and subscript in the notation (x_1) are superfluous, so we simply view the elements of $\mathbb{R}^1$ as real numbers x and write $\mathbb{R}$ for $\mathbb{R}^1$. Elements of $\mathbb{R}^2$ and of $\mathbb{R}^3$ are written (x, y) and (x, y, z) to avoid needless subscripts. These first few Euclidean spaces, $\mathbb{R}$, $\mathbb{R}^2$, and $\mathbb{R}^3$, are conveniently visualized as the line, the plane, and space itself. (See Figure 2.1.)

Elements of $\mathbb{R}$ are called **scalars**, of $\mathbb{R}^n$, **vectors**. The **origin** of $\mathbb{R}^n$, denoted $\mathbf{0}$, is defined to be

$$\mathbf{0} = (0, \ldots, 0).$$

© Springer International Publishing AG 2016

J. Shurman, *Calculus and Analysis in Euclidean Space*,
Undergraduate Texts in Mathematics, DOI 10.1007/978-3-319-49314-5_2

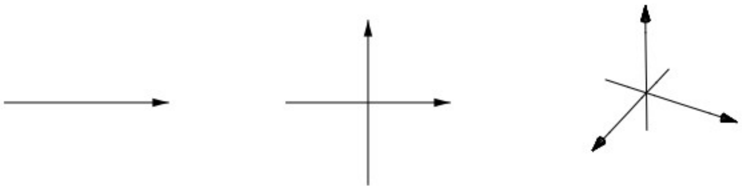

Figure 2.1. The first few Euclidean spaces

Sometimes the origin of $\mathbb{R}^n$ will be denoted $\mathbf{0}_n$ to distinguish it from other *origins* that we will encounter later.

In the first few Euclidean spaces $\mathbb{R}$, $\mathbb{R}^2$, $\mathbb{R}^3$, one can visualize a vector as a point x or as an arrow. The arrow can have its tail at the origin and its head at the point x, or its tail at any point p and its head correspondingly translated to $p+x$. (See Figure 2.2. Most illustrations will depict $\mathbb{R}$ or $\mathbb{R}^2$.)

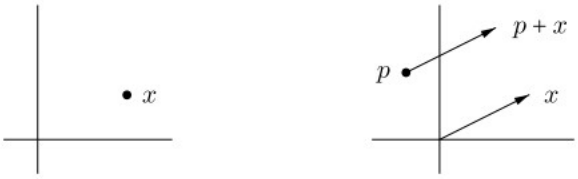

Figure 2.2. Various ways to envision a vector

To a mathematician, the word *space* doesn't connote volume but instead refers to a set endowed with some structure. Indeed, Euclidean space $\mathbb{R}^n$ comes with two algebraic operations. The first is **vector addition**,

$$+ : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}^n,$$

defined by adding the scalars at each component of the vectors,

$$(x_1, \dots, x_n) + (y_1, \dots, y_n) = (x_1 + y_1, \dots, x_n + y_n).$$

For example, $(1,2,3) + (4,5,6) = (5,7,9)$. Note that the meaning of the "+" sign is now overloaded: on the left of the displayed equality, it denotes the new operation of vector addition, whereas on the right side it denotes the old addition of real numbers. The multiple meanings of the plus sign shouldn't cause problems, because the meaning of "+" is clear from context, i.e., the meaning of "+" is clear from whether it sits between vectors or scalars. (An expression such as "$(1,2,3)+4$," with the plus sign between a vector and a scalar, makes no sense according to our grammar.)

The interpretation of vectors as arrows gives a geometric description of vector addition, at least in $\mathbb{R}^2$. To add the vectors x and y, draw them as arrows starting at $\mathbf{0}$ and then complete the parallelogram P that has x and y as two of its sides. The diagonal of P starting at $\mathbf{0}$ is then the arrow depicting the vector $x+y$. (See Figure 2.3.) The proof of this is a small argument with similar triangles, left to the reader as Exercise 2.1.2.

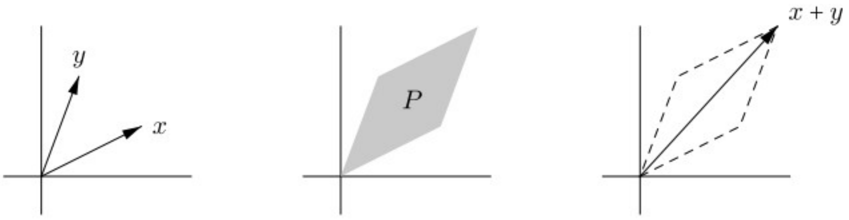

Figure 2.3. The parallelogram law of vector addition

The second operation on Euclidean space is **scalar multiplication**,

$$\cdot : \mathbb{R} \times \mathbb{R}^n \longrightarrow \mathbb{R}^n,$$

defined by

$$a \cdot (x_1, \dots, x_n) = (ax_1, \dots, ax_n).$$

For example, $2\cdot(3,4,5) = (6,8,10)$. We will almost always omit the symbol "$\cdot$" and write ax for $a\cdot x$. With this convention, juxtaposition is overloaded as "+" was overloaded above, but again this shouldn't cause problems.

Scalar multiplication of the vector x (viewed as an arrow) by a also has a geometric interpretation: it simply stretches (i.e., scales) x by a factor of a. When a is negative, ax turns x around and stretches it in the other direction by $|a|$. (See Figure 2.4.)

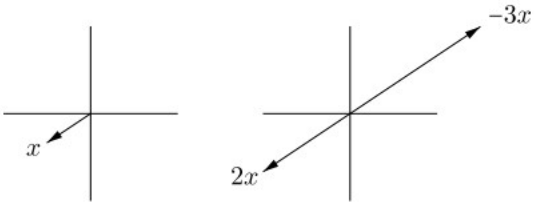

Figure 2.4. Scalar multiplication as stretching

With these two operations and distinguished element $\mathbf{0}$, Euclidean space satisfies the following algebraic laws.

Theorem 2.1.1 (Vector space axioms).

(A1) *Addition is associative:* $(x+y)+z = x+(y+z)$ *for all* $x, y, z \in \mathbb{R}^n$.
(A2) $\mathbf{0}$ *is an additive identity:* $\mathbf{0} + x = x$ *for all* $x \in \mathbb{R}^n$.
(A3) *Existence of additive inverses: for each* $x \in \mathbb{R}^n$ *there exists* $y \in \mathbb{R}^n$ *such that* $y + x = \mathbf{0}$.
(A4) *Addition is commutative:* $x + y = y + x$ *for all* $x, y \in \mathbb{R}^n$.
(M1) *Scalar multiplication is associative:* $a(bx) = (ab)x$ *for all* $a, b \in \mathbb{R}$, $x \in \mathbb{R}^n$.
(M2) 1 *is a multiplicative identity:* $1x = x$ *for all* $x \in \mathbb{R}^n$.
(D1) *Scalar multiplication distributes over scalar addition:* $(a+b)x = ax + bx$ *for all* $a, b \in \mathbb{R}$, $x \in \mathbb{R}^n$.
(D2) *Scalar multiplication distributes over vector addition:* $a(x+y) = ax + ay$ *for all* $a \in \mathbb{R}$, $x, y \in \mathbb{R}^n$.

All of these are consequences of how "+" and "·" and $\mathbf{0}$ are defined for $\mathbb{R}^n$ in conjunction with the fact that the real numbers, in turn endowed with "+" and "·" and containing 0 and 1, satisfy the field axioms (see Section 1.1). For example, to prove that $\mathbb{R}^n$ satisfies (M1), take any scalars $a, b \in \mathbb{R}$ and any vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Then

$$\begin{aligned}
a(bx) &= a(b(x_1, \ldots, x_n)) && \text{by definition of } x \\
&= a(bx_1, \ldots, bx_n) && \text{by definition of scalar multiplication} \\
&= (a(bx_1), \ldots, a(bx_n)) && \text{by definition of scalar multiplication} \\
&= ((ab)x_1, \ldots, (ab)x_n) && \text{by } n \text{ applications of (m1) in } \mathbb{R} \\
&= (ab)(x_1, \ldots, x_n) && \text{by definition of scalar multiplication} \\
&= (ab)x && \text{by definition of } x.
\end{aligned}$$

The other vector space axioms for $\mathbb{R}^n$ can be shown similarly, by unwinding vectors to their coordinates, quoting field axioms coordinatewise, and then bundling the results back up into vectors (see Exercise 2.1.3). Nonetheless, the vector space axioms do not perfectly parallel the field axioms, and you are encouraged to spend a little time comparing the two axiom sets to get a feel for where they are similar and where they are different (see Exercise 2.1.4). Note in particular that

For $n > 1$, $\mathbb{R}^n$ *is not endowed with vector-by-vector multiplication.*

Although one can define vector multiplication on $\mathbb{R}^n$ componentwise, this multiplication does not combine with vector addition to satisfy the field axioms except when $n = 1$. The multiplication of complex numbers makes $\mathbb{R}^2$ a field, and in Section 3.10 we will see an interesting noncommutative multiplication of vectors for $\mathbb{R}^3$, but these are special cases.

One benefit of the vector space axioms for $\mathbb{R}^n$ is that they are phrased **intrinsically**, meaning that they make no reference to the scalar coordinates of the vectors involved. Thus, once you use coordinates to establish the vector space axioms, your vector algebra can be intrinsic thereafter, making it lighter and more conceptual. Also, in addition to being intrinsic, the vector space axioms are general. While $\mathbb{R}^n$ is the prototypical set satisfying the vector space axioms, it is by no means the only one. In coming sections we will encounter other sets V (whose elements may be, for example, functions) endowed with their own addition, multiplication by elements of a field F, and distinguished element $\mathbf{0}$. If the vector space axioms are satisfied with V and F replacing $\mathbb{R}^n$ and $\mathbb{R}$ then we say that V is a **vector space over** F.

The pedagogical point here is that although the similarity between vector algebra and scalar algebra may initially make vector algebra seem uninspiring, in fact the similarity is exciting. It makes mathematics easier, because familiar algebraic manipulations apply in a wide range of contexts. The same symbol-patterns have more meaning. For example, we use intrinsic vector algebra to prove a result from Euclidean geometry, that the three medians of a triangle intersect. (A median is a segment from a vertex to the midpoint of the opposite edge.) Consider a triangle with vertices x, y, and z, and form the average of the three vertices,

$$p = \frac{x+y+z}{3}.$$

This algebraic average will be the **geometric center** of the triangle, where the medians meet. (See Figure 2.5.) Indeed, rewrite p as

$$p = x + \frac{2}{3}\left(\frac{y+z}{2} - x\right).$$

The displayed expression for p shows that it is two-thirds of the way from x along the line segment from x to the average of y and z, i.e., that p lies on the triangle median from vertex x to side yz. (Again see the figure. The idea is that $(y+z)/2$ is being interpreted as the midpoint of y and z, each of these viewed as a point, while on the other hand, the little mnemonic

head minus tail

helps us to remember quickly that $(y+z)/2 - x$ can be viewed as the arrow-vector from x to $(y+z)/2$.) Since p is defined symmetrically in x, y, and z, and it lies on one median, it therefore lies on the other two medians as well. In fact, the vector algebra has shown that it lies two-thirds of the way along each median. (As for how a person might find this proof, it is a matter of hoping that the geometric center $(x+y+z)/3$ lies on the median by taking the form $x + c((y+z)/2 - x)$ for some c and then seeing that indeed $c = 2/3$ works.)

The **standard basis** of $\mathbb{R}^n$ is the set of vectors

$$\{e_1, e_2, \ldots, e_n\}$$

where

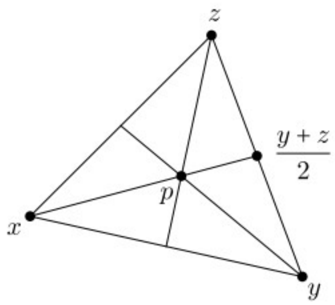

Figure 2.5. Three medians of a triangle

$$e_1 = (1,0,\dots,0), \quad e_2 = (0,1,\dots,0), \quad \dots, \quad e_n = (0,0,\dots,1).$$

(Thus each e_i is itself a vector, not the ith scalar entry of a vector.) Every vector $x = (x_1, x_2, \dots, x_n)$ (where the x_i are scalar entries) decomposes as

$$\begin{aligned} x &= (x_1, x_2, \dots, x_n) \\ &= (x_1, 0, \dots, 0) + (0, x_2, \dots, 0) + \cdots + (0, 0, \dots, x_n) \\ &= x_1(1,0,\dots,0) + x_2(0,1,\dots,0) + \cdots + x_n(0,0,\dots,1) \\ &= x_1e_1 + x_2e_2 + \cdots + x_ne_n, \end{aligned}$$

or more succinctly,

$$x = \sum_{i=1}^{n} x_i e_i. \tag{2.1}$$

Note that in equation (2.1), x and the e_i are vectors, while the x_i are scalars. The equation shows that every $x \in \mathbb{R}^n$ is expressible as a **linear combination** (sum of scalar multiples) of the standard basis vectors. The expression is unique, for if also $x = \sum_{i=1}^n x_i' e_i$ for some scalars $x_1', \dots, x_n'$ then the equality says that $x = (x_1', x_2', \dots, x_n')$, so that $x_i' = x_i$ for $i = 1, \dots, n$.

(The reason that the geometric-sounding word *linear* is used here and elsewhere in this chapter to describe properties having to do with the algebraic operations of addition and scalar multiplication will be explained in Chapter 3.)

The standard basis is handy in that it is a finite set of vectors from which each of the infinitely many vectors of $\mathbb{R}^n$ can be obtained in exactly one way as a linear combination. But it is not the only such set, nor is it always the optimal one.

Definition 2.1.2 (Basis). *A set of vectors* $\{f_i\}$ *is a* **basis** *of* $\mathbb{R}^n$ *if every* $x \in \mathbb{R}^n$ *is uniquely expressible as a linear combination of the* f_i.

For example, the set $\{f_1, f_2\} = \{(1,1),(1,-1)\}$ is a basis of $\mathbb{R}^2$. To see this, consider an arbitrary vector $(x,y) \in \mathbb{R}^2$. This vector is expressible as a linear combination of f_1 and f_2 if and only if there are scalars a and b such that

$$(x,y) = af_1 + bf_2.$$

Since $f_1 = (1,1)$ and $f_2 = (1,-1)$, this vector equation is equivalent to a pair of scalar equations,

$$\begin{aligned} x &= a + b, \\ y &= a - b. \end{aligned}$$

Add these equations and divide by 2 to get $a = (x+y)/2$, and similarly $b = (x-y)/2$. In other words, we have found that

$$(x,y) = \frac{x+y}{2}(1,1) + \frac{x-y}{2}(1,-1),$$

and the coefficients $a = (x+y)/2$ and $b = (x-y)/2$ on the right side of the equation are the only possible coefficients a and b for the equation to hold. That is, scalars a and b exist to express the vector (x,y) as a linear combination of $\{f_1, f_2\}$, and the scalars are uniquely determined by the vector. Thus $\{f_1, f_2\}$ is a basis of $\mathbb{R}^2$, as claimed.

The set $\{g_1, g_2\} = \{(1,3),(2,6)\}$ is not a basis of $\mathbb{R}^2$, because every linear combination $ag_1 + bg_2$ takes the form $(a+2b, 3a+6b)$, with the second entry equal to three times the first. The vector $(1,0)$ is therefore not a linear combination of g_1 and g_2.

Nor is the set $\{h_1, h_2, h_3\} = \{(1,0),(1,1),(1,-1)\}$ a basis of $\mathbb{R}^2$, because $h_3 = 2h_1 - h_2$, so that h_3 is a nonunique linear combination of the h_j.

See Exercises 2.1.9 and 2.1.10 for practice with bases.

Exercises

2.1.1. Write down any three specific nonzero vectors u, v, w from $\mathbb{R}^3$ and any two specific nonzero scalars a, b from $\mathbb{R}$. Compute $u+v$, aw, $b(v+w)$, $(a+b)u$, $u+v+w$, abw, and the additive inverse to u.

2.1.2. Working in $\mathbb{R}^2$, give a geometric proof that if we view the vectors x and y as arrows from $\mathbf{0}$ and form the parallelogram P with these arrows as two of its sides, then the diagonal z starting at $\mathbf{0}$ is the vector sum $x+y$ viewed as an arrow.

2.1.3. Verify that $\mathbb{R}^n$ satisfies vector space axioms (A2), (A3), (D1).

2.1.4. Are all the field axioms used in verifying that Euclidean space satisfies the vector space axioms?

2.1.5. Show that $\mathbf{0}$ is the unique additive identity in $\mathbb{R}^n$. Show that each vector $x \in \mathbb{R}^n$ has a unique additive inverse, which can therefore be denoted $-x$. (And it follows that vector subtraction can now be defined,

$$- : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}^n, \qquad x - y = x + (-y) \quad \text{for all } x, y \in \mathbb{R}^n.)$$

Show that $0x = \mathbf{0}$ for all $x \in \mathbb{R}^n$.

2.1.6. Repeat the previous exercise, but with $\mathbb{R}^n$ replaced by an arbitrary vector space V over a field F. (Work with the axioms.)

2.1.7. Show the uniqueness of the additive identity and the additive inverse using only (A1), (A2), (A3). (This is tricky; the opening pages of some books on group theory will help.)

2.1.8. Let x and y be noncollinear vectors in $\mathbb{R}^3$. Give a geometric description of the set of all linear combinations of x and y.

2.1.9. Which of the following sets are bases of $\mathbb{R}^3$?

$$\begin{aligned} S_1 &= \{(1,0,0),(1,1,0),(1,1,1)\}, \\ S_2 &= \{(1,0,0),(0,1,0),(0,0,1),(1,1,1)\}, \\ S_3 &= \{(1,1,0),(0,1,1)\}, \\ S_4 &= \{(1,1,0),(0,1,1),(1,0,-1)\}. \end{aligned}$$

How many elements do you think a basis for $\mathbb{R}^n$ must have? Give (without proof) geometric descriptions of all bases of $\mathbb{R}^2$, of $\mathbb{R}^3$.

2.1.10. Recall the field $\mathbb{C}$ of complex numbers. Define **complex n-space** $\mathbb{C}^n$ analogously to $\mathbb{R}^n$:

$$\mathbb{C}^n = \{(z_1, \dots, z_n) : z_i \in \mathbb{C} \text{ for } i = 1, \dots, n\},$$

and endow it with addition and scalar multiplication defined by the same formulas as for $\mathbb{R}^n$. You may take for granted that under these definitions, $\mathbb{C}^n$ satisfies the vector space axioms with scalar multiplication by scalars from $\mathbb{R}$, and also $\mathbb{C}^n$ satisfies the vector space axioms with scalar multiplication by scalars from $\mathbb{C}$. That is, using language that was introduced briefly in this section, $\mathbb{C}^n$ can be viewed as a vector space over $\mathbb{R}$ and also, separately, as a vector space over $\mathbb{C}$. Give a basis for each of these vector spaces.

Brief Pedagogical Interlude

Before continuing, a few comments about how to work with these notes may be helpful.

The subject-matter of Chapters 2 through 5 is largely cumulative, with the main theorem of Chapter 5 being proved with main results of Chapters 2, 3, and 4. Each chapter is largely cumulative internally as well. To acquire detailed command of so much material and also a large-scale view of how it fits together, the trick is to focus on each section's techniques while studying that section and working its exercises, but thereafter to use the section's ideas freely by reference. Specifically, after the scrutiny of vector algebra in the previous section, one's vector manipulations should be fluent from now on, freeing one to concentrate on vector geometry in the next section, after which the geometry should also be light while one is concentrating on the analytical ideas of the following section, and so forth.

Admittedly, the model that one has internalized all the prior material before moving on is idealized. For that matter, so is the model that a body of interplaying ideas is linearly cumulative. In practice, focusing entirely on the details of whichever topics are currently active while using previous ideas by reference isn't always optimal. One might engage with the details of previous ideas because one is coming to understand them better, or because the current ideas showcase the older ones in a new way. Still, the paradigm of technical emphasis on the current ideas and fluent use of the earlier material does help a person who is navigating a large body of mathematics to conserve energy and synthesize a larger picture.

2.2 Geometry: Length and Angle

The geometric notions of length and angle in $\mathbb{R}^n$ are readily described in terms of the algebraic notion of inner product.

Definition 2.2.1 (Inner product). *The* **inner product** *is a function from pairs of vectors to scalars,*

$$\langle\ ,\ \rangle : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R},$$

defined by the formula

$$\langle (x_1, \dots, x_n), (y_1, \dots, y_n) \rangle = \sum_{i=1}^{n} x_i y_i.$$

For example,

$$\begin{aligned} &\langle (1,1,\dots,1), (1,2,\dots,n) \rangle = \frac{n(n+1)}{2}, \\ &\langle x, e_j \rangle = x_j \quad \text{where } x = (x_1, \dots, x_n) \text{ and } j \in \{1, \dots, n\}, \\ &\langle e_i, e_j \rangle = \delta_{ij} \text{ (this means 1 if } i = j \text{, 0 otherwise).} \end{aligned}$$

Proposition 2.2.2 (Inner product properties).

(IP1) *The inner product is positive definite:* $\langle x, x\rangle \geq 0$ *for all* $x \in \mathbb{R}^n$*, with equality if and only if* $x = \mathbf{0}$.
(IP2) *The inner product is symmetric:* $\langle x, y\rangle = \langle y, x\rangle$ *for all* $x, y \in \mathbb{R}^n$.
(IP3) *The inner product is bilinear:*

$$\begin{aligned}\langle x + x', y\rangle &= \langle x, y\rangle + \langle x', y\rangle, \quad \langle ax, y\rangle = a\langle x, y\rangle,\\ \langle x, y + y'\rangle &= \langle x, y\rangle + \langle x, y'\rangle, \quad \langle x, by\rangle = b\langle x, y\rangle\end{aligned}$$

for all $a, b \in \mathbb{R}$*,* $x, x', y, y' \in \mathbb{R}^n$.

Proof. Exercise 2.2.4. □

The reader should be aware that:

> *In general,* $\langle x + x', y + y'\rangle$ *does not equal* $\langle x, y\rangle + \langle x', y'\rangle$.

Indeed, expanding $\langle x + x', y + y'\rangle$ carefully with the inner product properties shows that the cross-terms $\langle x, y'\rangle$ and $\langle x', y\rangle$ are present in addition to $\langle x, y\rangle$ and $\langle x', y'\rangle$.

Like the vector space axioms, the inner product properties are phrased intrinsically, although they need to be proved using coordinates. As mentioned in the previous section, intrinsic methods are neater and more conceptual than using coordinates. More importantly:

> *The rest of the results of this section are proved by reference to the inner product properties, with no further reference to the inner product formula.*

The notion of an inner product generalizes beyond Euclidean space—this will be demonstrated in Exercise 2.3.4, for example—and thanks to the displayed sentence, once the properties (IP1) through (IP3) are established for any inner product, all of the pending results in the section will follow automatically with no further work. (But here a slight disclaimer is necessary. In the displayed sentence, the word *results* does not refer to the pending graphic figures. The fact that the length and angle to be defined in this section will agree with prior notions of length and angle in the plane, or in three-dimensional space, does depend on the specific inner product formula. In Euclidean space, the inner product properties do not determine the inner product formula uniquely. This point will be addressed in Exercise 3.5.1.)

Definition 2.2.3 (Modulus). *The* **modulus** (*or* **absolute value**) *of a vector* $x \in \mathbb{R}^n$ *is defined as*

$$|x| = \sqrt{\langle x, x\rangle}.$$

Thus the modulus is defined in terms of the inner product, rather than by its own formula. The inner product formula shows that the modulus formula is

$$|(x_1, \dots, x_n)| = \sqrt{x_1^2 + \cdots + x_n^2},$$

so that some particular examples are

$$|(1,2,\dots,n)| = \sqrt{\frac{n(n+1)(2n+1)}{6}},$$
$$|e_i| = 1.$$

However, the definition of the modulus in terms of inner product combines with the inner product properties to show, with no reference to the inner product formula or the modulus formula, that the modulus satisfies the following properties (Exercise 2.2.5).

Proposition 2.2.4 (Modulus properties).

(Mod1) *The modulus is positive:* $|x| \geq 0$ *for all* $x \in \mathbb{R}^n$*, with equality if and only if* $x = \mathbf{0}$.

(Mod2) *The modulus is absolute-homogeneous:* $|ax| = |a||x|$ *for all* $a \in \mathbb{R}$ *and* $x \in \mathbb{R}^n$.

Like other symbols, the absolute value signs are now overloaded, but their meaning can be inferred from context, as in property (Mod2). When n is 1, 2, or 3, the modulus $|x|$ gives the distance from $\mathbf{0}$ to the point x, or the length of x viewed as an arrow. (See Figure 2.6.)

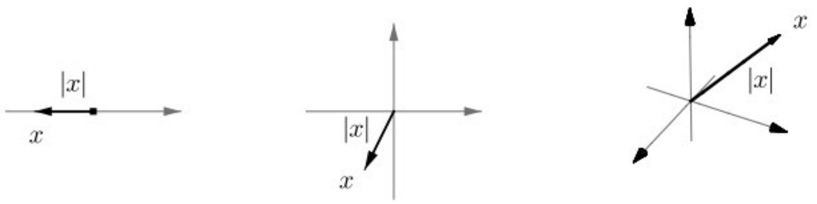

Figure 2.6. Modulus as length

The following relation between inner product and modulus will help to show that distance in $\mathbb{R}^n$ behaves as it should, and that angle in $\mathbb{R}^n$ makes sense. Since the relation is not obvious, its proof is a little subtle.

Theorem 2.2.5 (Cauchy–Schwarz inequality). *For all* $x, y \in \mathbb{R}^n$,

$$|\langle x, y\rangle| \leq |x|\,|y|,$$

with equality if and only if one of x, y *is a scalar multiple of the other.*

Note that the absolute value signs mean different things on each side of the Cauchy–Schwarz inequality. On the left side, the quantities x and y are vectors, their inner product $\langle x, y\rangle$ is a scalar, and $|\langle x, y\rangle|$ is its scalar absolute value, while on the right side, $|x|$ and $|y|$ are the scalar absolute values of vectors, and $|x|\,|y|$ is their product. That is, the Cauchy–Schwarz inequality says:

The size of the product is at most the product of the sizes.

The Cauchy–Schwarz inequality can be written out in coordinates if we temporarily abandon the principle that we should avoid reference to formulas,

$$(x_1y_1 + \cdots + x_ny_n)^2 \le (x_1^2 + \cdots + x_n^2)(y_1^2 + \cdots + y_n^2).$$

And this inequality can be proved unconceptually as follows (the reader is encouraged only to skim the following computation). Rewrite the desired inequality as

$$\Big(\sum_i x_iy_i\Big)^2 \le \sum_i x_i^2 \cdot \sum_j y_j^2,$$

where the indices of summation run from 1 to n. Expand the square to get

$$\sum_i x_i^2y_i^2 + \sum_{\substack{i,j\\ i\ne j}} x_iy_ix_jy_j \le \sum_{i,j} x_i^2y_j^2,$$

and canceling the terms common to both sides reduces it to

$$\sum_{i\ne j} x_iy_ix_jy_j \le \sum_{i\ne j} x_i^2y_j^2,$$

or

$$\sum_{i\ne j}(x_i^2y_j^2 - x_iy_ix_jy_j) \ge 0.$$

Rather than sum over all pairs (i,j) with $i \ne j$, sum over the pairs with $i < j$, collecting the (i,j)-term and the (j,i)-term for each such pair, and the previous inequality becomes

$$\sum_{i<j}(x_i^2y_j^2 + x_j^2y_i^2 - 2x_iy_jx_jy_i) \ge 0.$$

Thus the desired inequality has reduced to a true inequality,

$$\sum_{i<j}(x_iy_j - x_jy_i)^2 \ge 0.$$

So the main proof is done, although there is still the question of when equality holds.

But surely the previous paragraph is not the graceful way to argue. The computation draws on the minutiae of the formulas for the inner product and the modulus, rather than using their properties. It is uninformative, making the Cauchy–Schwarz inequality look like a low-level accident. It suggests that larger-scale mathematics is just a matter of bigger and bigger formulas. To prove the inequality in a way that is enlightening and general, we should work intrinsically, keeping the scalars $\langle x, y\rangle$ and $|x|$ and $|y|$ notated in their concise forms, and we should use properties, not formulas. The idea is that the calculation in coordinates reduces to the fact that squares are nonnegative. That is, the Cauchy–Schwarz inequality is somehow *quadratically hard*, and its verification amounted to completing many squares. The argument to be given here is guided by this insight to prove the inequality by citing facts about quadratic polynomials, facts established by completing one square back in high-school algebra at the moment that doing so was called for. Thus we eliminate redundancy and clutter. So the argument to follow will involve an auxiliary object, a judiciously chosen quadratic polynomial, but in return it will become coherent.

Proof. The result is clear when $x = \mathbf{0}$, so assume $x \neq \mathbf{0}$. For every $a \in \mathbb{R}$,

$$\begin{aligned} 0 &\le \langle ax - y, ax - y\rangle && \text{by positive definiteness} \\ &= a\langle x, ax - y\rangle - \langle y, ax - y\rangle && \text{by linearity in the first variable} \\ &= a^2\langle x, x\rangle - a\langle x, y\rangle - a\langle y, x\rangle + \langle y, y\rangle && \text{by linearity in the second variable} \\ &= a^2|x|^2 - 2a\langle x, y\rangle + |y|^2 && \text{by symmetry, definition of modulus.} \end{aligned}$$

View the right side as a quadratic polynomial in the scalar variable a, where the scalar coefficients of the polynomial depend on the generic but fixed vectors x and y,

$$f(a) = |x|^2a^2 - 2\langle x, y\rangle a + |y|^2.$$

We have shown that $f(a)$ is always nonnegative, so f has at most one root. Thus by the quadratic formula its discriminant is nonpositive,

$$4\langle x, y\rangle^2 - 4|x|^2|y|^2 \le 0,$$

and the Cauchy–Schwarz inequality $|\langle x, y\rangle| \le |x|\,|y|$ follows. Equality holds exactly when the quadratic polynomial $f(a) = |ax - y|^2$ has a root a, i.e., exactly when $y = ax$ for some $a \in \mathbb{R}$. □

Geometrically, the condition for equality in Cauchy–Schwarz is that the vectors x and y, viewed as arrows at the origin, are parallel, though perhaps pointing in opposite directions. A geometrically conceived proof of Cauchy–Schwarz is given in Exercise 2.2.15 to complement the algebraic argument that has been given here.

The Cauchy–Schwarz inequality shows that the modulus function satisfies the *triangle inequality*.

Theorem 2.2.6 (Triangle inequality). *For all* $x, y \in \mathbb{R}^n$,

$$|x + y| \le |x| + |y|,$$

with equality if and only if one of x, y is a nonnegative scalar multiple of the other.

Proof. To show this, compute

$$\begin{aligned} |x+y|^2 &= \langle x+y, x+y\rangle \\ &= |x|^2 + 2\langle x,y\rangle + |y|^2 && \text{by bilinearity} \\ &\le |x|^2 + 2|x||y| + |y|^2 && \text{by Cauchy–Schwarz} \\ &= (|x|+|y|)^2, \end{aligned}$$

proving the inequality. Equality holds exactly when $\langle x,y\rangle = |x||y|$, or equivalently when $|\langle x,y\rangle| = |x||y|$ and $\langle x,y\rangle \ge 0$. These hold when one of x, y is a scalar multiple of the other and the scalar is nonnegative. □

While the Cauchy–Schwarz inequality says that the size of the product is at most the product of the sizes, the triangle inequality says:

The size of the sum is at most the sum of the sizes.

The triangle inequality's name is explained by its geometric interpretation in $\mathbb{R}^2$. View x as an arrow at the origin, y as an arrow with tail at the head of x, and $x+y$ as an arrow at the origin. These three arrows form a triangle, and the assertion is that the lengths of two sides sum to at least the length of the third. (See Figure 2.7.)

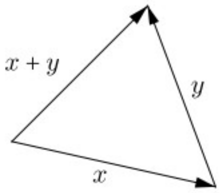

Figure 2.7. Sides of a triangle

The **full triangle inequality** says that for all $x, y \in \mathbb{R}^n$,

$$\big|\,|x| - |y|\,\big| \le |x \pm y| \le |x| + |y|.$$

The proof is Exercise 2.2.7.

A small argument, which can be formalized as induction if one is painstaking, shows that the basic triangle inequality extends from two vectors to any finite number of vectors. For example,

$$|x+y+z| \le |x+y| + |z| \le |x| + |y| + |z|.$$

The only obstacle to generalizing the basic triangle inequality in this fashion is notation. The argument can't use the symbol n to denote the number of vectors, because n already denotes the dimension of the Euclidean space where we are working; and furthermore, the vectors can't be denoted with subscripts since a subscript denotes a component of an individual vector. Thus, for now we are stuck writing something like

$$|x^{(1)} + \cdots + x^{(k)}| \le |x^{(1)}| + \cdots + |x^{(k)}| \quad \text{for all } x^{(1)}, \dots, x^{(k)} \in \mathbb{R}^n,$$

or

$$\left|\sum_{i=1}^{k} x^{(i)}\right| \le \sum_{i=1}^{k} |x^{(i)}|, \quad x^{(1)}, \dots, x^{(k)} \in \mathbb{R}^n.$$

As our work with vectors becomes more intrinsic, vector entries will demand less of our attention, and we will be able to denote vectors by subscripts. The notation-change will be implemented in the next section.

For every vector $x = (x_1, \dots, x_n) \in \mathbb{R}^n$, useful bounds on the modulus $|x|$ in terms of the scalar absolute values $|x_i|$ are as follows.

Proposition 2.2.7 (Size bounds). *For every* $j \in \{1, \dots, n\}$,

$$|x_j| \le |x| \le \sum_{i=1}^{n} |x_i|.$$

The proof (by quick applications of the Cauchy–Schwarz inequality and the triangle inequality) is Exercise 2.2.8.

The modulus gives rise to a distance function on $\mathbb{R}^n$ that behaves as distance should. Define

$$d : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$$

by

$$d(x,y) = |y - x|.$$

For example, $d(e_i, e_j) = \sqrt{2}(1 - \delta_{ij})$.

Theorem 2.2.8 (Distance properties).

(D1) *Distance is positive:* $d(x,y) \ge 0$ *for all* $x, y \in \mathbb{R}^n$, *and* $d(x,y) = 0$ *if and only if* $x = y$.
(D2) *Distance is symmetric:* $d(x,y) = d(y,x)$ *for all* $x, y \in \mathbb{R}^n$.
(D3) *Triangle inequality:* $d(x,z) \le d(x,y) + d(y,z)$ *for all* $x, y, z \in \mathbb{R}^n$.

(D1) and (D2) are clearly desirable as properties of a distance function. Property (D3) says that you can't shorten your trip from x to z by making a stop at y.

Proof. Exercise 2.2.9. □

The Cauchy–Schwarz inequality also lets us define the angle between two nonzero vectors in terms of the inner product. If x and y are nonzero vectors in $\mathbb{R}^n$, define their angle $\theta_{x,y}$ by the condition

$$\cos\theta_{x,y} = \frac{\langle x, y\rangle}{|x||y|}, \quad 0 \le \theta_{x,y} \le \pi. \tag{2.2}$$

The condition is sensible because $-1 \le \frac{\langle x,y\rangle}{|x||y|} \le 1$ by the Cauchy–Schwarz inequality. For example, $\cos\theta_{(1,0),(1,1)} = 1/\sqrt{2}$, and so $\theta_{(1,0),(1,1)} = \pi/4$. In particular, two nonzero vectors x and y are **orthogonal** when $\langle x, y\rangle = 0$. Naturally, we would like $\theta_{x,y}$ to correspond to the usual notion of angle, at least in $\mathbb{R}^2$, and indeed it does—see Exercise 2.2.10. For convenience, define *any* two vectors x and y to be orthogonal if $\langle x, y\rangle = 0$, thus making $\mathbf{0}$ orthogonal to all vectors.

Rephrasing geometry in terms of intrinsic vector algebra not only extends the geometric notions of length and angle uniformly to any dimension, it also makes some low-dimensional geometry easier. For example, vectors show in a natural way that the three altitudes of every triangle must meet. Let x and y denote two sides of the triangle, making the third side $x-y$ by the *head minus tail* mnemonic. Let q be the point where the altitudes to x and y meet. (See Figure 2.8, which also shows the third altitude.) Thus

$$q - y \perp x \quad \text{and} \quad q - x \perp y.$$

We want to show that q also lies on the third altitude, i.e., that

$$q \perp x - y.$$

To rephrase matters in terms of inner products, we want to show that

$$\begin{Bmatrix} \langle q - y, x\rangle = 0 \\ \langle q - x, y\rangle = 0 \end{Bmatrix} \implies \langle q, x - y\rangle = 0.$$

Since the inner product is linear in each of its arguments, a further rephrasing is that we want to show that

$$\begin{Bmatrix} \langle q, x\rangle = \langle y, x\rangle \\ \langle q, y\rangle = \langle x, y\rangle \end{Bmatrix} \implies \langle q, x\rangle = \langle q, y\rangle.$$

And this is immediate because the inner product is symmetric: $\langle q, x\rangle$ and $\langle q, y\rangle$ both equal $\langle x, y\rangle$, and so they equal each other as desired. The point q where the three altitudes meet is called the **orthocenter** of the triangle. In general, the orthocenter of a triangle is not the geometric center that we considered in the previous section.

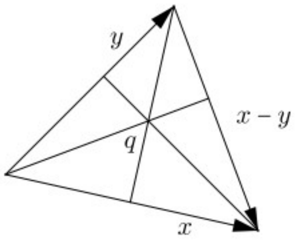

Figure 2.8. Three altitudes of a triangle

Exercises

2.2.1. Let $x = (\frac{\sqrt{3}}{2}, -\frac{1}{2}, 0)$, $y = (\frac{1}{2}, \frac{\sqrt{3}}{2}, 1)$, $z = (1,1,1)$. Compute $\langle x, x\rangle$, $\langle x, y\rangle$, $\langle y, z\rangle$, $|x|$, $|y|$, $|z|$, $\theta_{x,y}$, θ_{y,e_1}, θ_{z,e_2}.

2.2.2. Show that the points $x = (2,-1,3,1)$, $y = (4,2,1,4)$, $z = (1,3,6,1)$ form the vertices of a triangle in $\mathbb{R}^4$ with two equal angles.

2.2.3. Explain why for all $x \in \mathbb{R}^n$, $x = \sum_{j=1}^n \langle x, e_j\rangle e_j$.

2.2.4. Prove the inner product properties.

2.2.5. Use the inner product properties and the definition of the modulus in terms of the inner product to prove the modulus properties.

2.2.6. In the text, the modulus is defined in terms of the inner product. Prove that this can be turned around by showing that for every $x, y \in \mathbb{R}^n$,

$$\langle x, y\rangle = \frac{|x+y|^2 - |x-y|^2}{4}.$$

2.2.7. Prove the full triangle inequality: for every $x, y \in \mathbb{R}^n$,

$$\big|\,|x| - |y|\,\big| \le |x \pm y| \le |x| + |y|.$$

Do not do this by writing three more variants of the proof of the triangle inequality, but by substituting suitably into the basic triangle inequality, which is already proved.

2.2.8. Let $x = (x_1, \dots, x_n) \in \mathbb{R}^n$. Prove the size bounds: for every $j \in \{1, \dots, n\}$,

$$|x_j| \le |x| \le \sum_{i=1}^n |x_i|.$$

(One approach is to start by noting that $x_j = \langle x, e_j\rangle$ and recalling equation (2.1).) When can each "$\le$" be an "$=$"?

2.2.9. Prove the distance properties.

2.2.10. Working in $\mathbb{R}^2$, depict the nonzero vectors x and y as arrows from the origin and depict $x - y$ as an arrow from the endpoint of y to the endpoint of x. Let θ denote the angle (in the usual geometric sense) between x and y. Use the law of cosines to show that

$$\cos\theta = \frac{\langle x, y\rangle}{|x||y|},$$

so that our notion of angle agrees with the geometric one, at least in $\mathbb{R}^2$.

2.2.11. Prove that for every nonzero $x \in \mathbb{R}^n$, $\sum_{i=1}^{n} \cos^2\theta_{x,e_i} = 1$.

2.2.12. Prove that two nonzero vectors x, y are orthogonal if and only if $|x + y|^2 = |x|^2 + |y|^2$.

2.2.13. Use vectors in $\mathbb{R}^2$ to show that the diagonals of a parallelogram are perpendicular if and only if the parallelogram is a rhombus.

2.2.14. Use vectors to show that every angle inscribed in a semicircle is right.

2.2.15. Let x and y be vectors, with x nonzero. Define the **parallel** component of y along x and the **normal** component of y to x to be

$$y_{(\|x)} = \frac{\langle x, y\rangle}{|x|^2}x \qquad \text{and} \qquad y_{(\perp x)} = y - y_{(\|x)}.$$

(a) Show that $y = y_{(\|x)} + y_{(\perp x)}$; show that $y_{(\|x)}$ is a scalar multiple of x; show that $y_{(\perp x)}$ is orthogonal to x. Show that the decomposition of y as a sum of vectors parallel and perpendicular to x is unique. Draw an illustration.

(b) Show that

$$|y|^2 = |y_{(\|x)}|^2 + |y_{(\perp x)}|^2.$$

What theorem from classical geometry does this encompass?

(c) Explain why it follows from (b) that

$$|y_{(\|x)}| \le |y|,$$

with equality if and only if y is a scalar multiple of x. Use this inequality to give another proof of the Cauchy–Schwarz inequality. This argument gives the geometric content of Cauchy–Schwarz: *the parallel component of one vector along another is at most as long as the original vector.*

(d) The proof of the Cauchy–Schwarz inequality in part (c) refers to parts (a) and (b), part (a) refers to orthogonality, orthogonality refers to an angle, and as explained in the text, the fact that angles make sense depends on the Cauchy–Schwarz inequality. And so the proof in part (c) apparently relies on circular logic. Explain why the logic is in fact not circular.

2.2.16. Given nonzero vectors $x_1, x_2, \dots, x_n$ in $\mathbb{R}^n$, the **Gram–Schmidt process** is to set

$$\begin{aligned}
x'_1 &= x_1\\
x'_2 &= x_2 - (x_2)_{(\parallel x'_1)}\\
x'_3 &= x_3 - (x_3)_{(\parallel x'_2)} - (x_3)_{(\parallel x'_1)}\\
&\vdots\\
x'_n &= x_n - (x_n)_{(\parallel x'_{n-1})} - \cdots - (x_n)_{(\parallel x'_1)}.
\end{aligned}$$

(a) What is the result of applying the Gram–Schmidt process to the vectors $x_1 = (1,0,0)$, $x_2 = (1,1,0)$, and $x_3 = (1,1,1)$?

(b) Returning to the general case, show that $x'_1, \dots, x'_n$ are pairwise orthogonal and that each x'_j has the form

$$x'_j = a_{j1}x_1 + a_{j2}x_2 + \cdots + a_{j,j-1}x_{j-1} + x_j.$$

Thus every linear combination of the new $\{x'_j\}$ is also a linear combination of the original $\{x_j\}$. The converse is also true and will be shown in Exercise 3.3.13.

2.3 Analysis: Continuous Mappings

A **mapping** from $\mathbb{R}^n$ to $\mathbb{R}^m$ is some rule that assigns to each point x in $\mathbb{R}^n$ a point in $\mathbb{R}^m$. Generally, mappings will be denoted by letters such as f, g, h. When $m = 1$, we usually say *function* instead of mapping.

For example, the mapping

$$f : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$$

defined by

$$f(x,y) = (x^2 - y^2, 2xy)$$

takes the real and imaginary parts of a complex number $z = x + iy$ and returns the real and imaginary parts of z^2. By the nature of multiplication of complex numbers, this means that each output point has modulus equal to the square of the modulus of the input point and has angle equal to twice the angle of the input point. Make sure that you see how this is shown in Figure 2.9.

Mappings expressed by formulas may be undefined at certain points (e.g., $f(x) = 1/|x|$ is undefined at $\mathbf{0}$), so we need to restrict their domains. For a given dimension n, a given set $A \subset \mathbb{R}^n$, and a second dimension m, let $\mathcal{M}(A, \mathbb{R}^m)$ denote the set of all mappings $f : A \longrightarrow \mathbb{R}^m$. This set forms a vector space over $\mathbb{R}$ (whose *points* are functions) under the operations

$$+ : \mathcal{M}(A, \mathbb{R}^m) \times \mathcal{M}(A, \mathbb{R}^m) \longrightarrow \mathcal{M}(A, \mathbb{R}^m),$$

"0 A #

)*+,)))-) ' * 7 ! / A2 %
F! 6!

)*+,)))-) ' "0 . A2 2
! 8 ** / C 7 G 6!
('1 1: 6 8 !

")*+,)))-) ' < ! -
0 6!

%)*+,)))-) ' . A / 5 !
6!
-7 5 8 6 E 999

-
)*+,)))-) ' * * C ! 9 @ I !
? . $2 * 6 7
-7

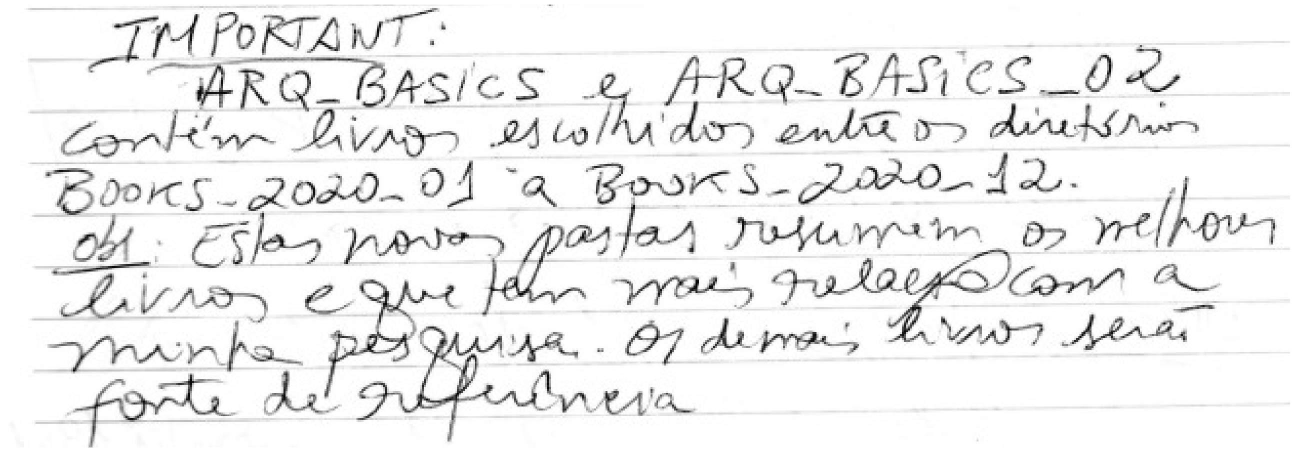

IMPORTANT:
ARQ_BASICS e ARQ_BASICS_02
contém livros escolhidos entre os diretórios
BOOKS_2020_01 a BOOKS_2020_12.
obs: Estas novas pastas resumem os melhores
livros e que têm mais relação com a
minha pesquisa. Os demais livros serão
fonte de referência

data 04.05.2020

Copy editări:

a) Advanced_Calculus_Taylor.pdf
b) Advanced_Calculus_Taylor.djvu

Obs: Advanced_Calculus_Taylor.pdf

Pag. 84 (100) → 10

The number $\sqrt{3}$ is defined as the positive number c such that $c^2 = 3$.

A3.4) Calculus the classic by Earl W. Swokowski... pdf

A3.5) Direct Methods in the Calculus of Variations (Applied...)... pdf

A3.6) Functions of Complex Variable... pdf

A3.7) Geometry.pdf

A3.8) H. Jerome Keisler... Infinitesimal... pdf

A3.9) Higher Engineering Mathematics by B.S. Grewal.pdf

A3.10) Integral Calculus 3D Geometry and Vector Booster...pdf

A3.11) kupdf.net analytical-geometry-2d-and-3d-p-n-vittal.pdf

A3.12) Vector Calculus by Krishna Series.pdf

04 05 2020

A4) Calculus and Analysis in Euclidean Space/
Saret st_CalculusAnalysisEuclidean Space.pdf

Pág. 23(35) Chapter 2 - Euclidean Space.

A5) Calculus in 3D Geometry, Vectors, and Multivariate Calculus...

A5.1) Algebra and Trigonometry...pdf

A5.2) Calculus in 3D Geometry....pdf

Pág. 67(78) - Chapter 2
Curves and Vector-Valued Functions of One Variable

A5.3) Fundamentals of Mathematical Differential Calculus.pdf

A.5.4) Geometry, the Line and the Circle.pdf

IMPORTANT:
ARQ_BASICS e ARQ_BASICS_02 contém livros escolhidos entre os diretórios BOOKS_2020_01 à BOOKS_2020_12.
obs: Estas novas pastas resumem os melhores livros e que tem mais relação com a minha pesquisa. Os demais livros serão fonte de referência

& '''

data 04.05.2020

163) Pappus D Alexandrie

Source: https://archive.org/details/PappusDAlexandrieEd.PaulVerEeckelLaCollectionMathmatique/

Obs.: 1010 pages. Pág. 400(525)

Identifier-ark

ark:/13960/t5s80978s

other: lib.arch.nmu.org.ua/handle/GenofondUA/22654

obs: Oeuvres Mathématiques D'Évariste Galois

https://gallica.bnf.fr/ark:/12148/bpt6k9800489w/f11.image

obs: Les trois livres de Porisme D'Euclide

https://gallica.bnf.fr/ark:/12148/bpt6k6214433z/f9.image

others: Historia Mathematica

Volume 8, Issue 4, November 1981,

04.05.2020

Pages 417-438.

How trisections of the angle were transmitted from Greek to islamic geometry

Jan P Hogendijk

https://doi.org/10.1016/0315-0860(81)90051-3

Source: www.sciencedirect.com/science/article/pii/0315086081900513

other: caphi.uni-nantes.fr/IMG/pdf/Descartes-Geometrie.pdf

→ Descartes La Geometrie 1637.

164) 20200505

A) Books_79_2020/
Great Physics the Life and Times.../

A1) Great Physics ....pdf

Pag. 203(215) Chapter 14

Adventure in thought
Albert Einstein

Pag. 208(220)

Doctrine of Space and Time

& '%
K ?--J " % $ 1L)?--J > > ")? G >1$>)
)*+,)))-) '% . A > / 5 ! /
5 ! A . @ G 2 -C ! 4 8 A 2 4 * 6!
(2 " 6 % *!8 / * 7 F
(: 0 6 ! /

04.05.2020

Pages 417-438.

How trisections of the angle were transmitted from Greek to islamic geometry

Jan P Hogendijk

https://doi.org/10.1016/0315-0860(81)90051-3

Source: www.sciencedirect.com/science/article/pii/0315086081900513

other: caphi.univ-nantes.fr/IMG/pdf/Descartes-Geometrie.pdf
→ Descartes La Geometrie 1637.

164) 20200505

A) Books_79 2020/Great Physics the Life and Times.../
A1) Great Physics ....pdf
Pág. 203(215) Chapter 14
Adventure in thought
Albert Einstein
Pág. 208(220)
Doctrine of Space and Time

05 05 2020

(Pag.209) Whereas Lorentz had struggled to explain the invariance of the speed of light with a constructive theory that hypothesized motion-dependent molecular forces, Einstein bypassed all the complications by simply promoting the constancy to a postulate.

For Lorentz and his contemporaries, it was a problem, for Einstein a principle.

A2) The Quantum Physicists And an Introduction to their Physics (1970)...pdf.

Pag. 30(43)

Conservation of momentum

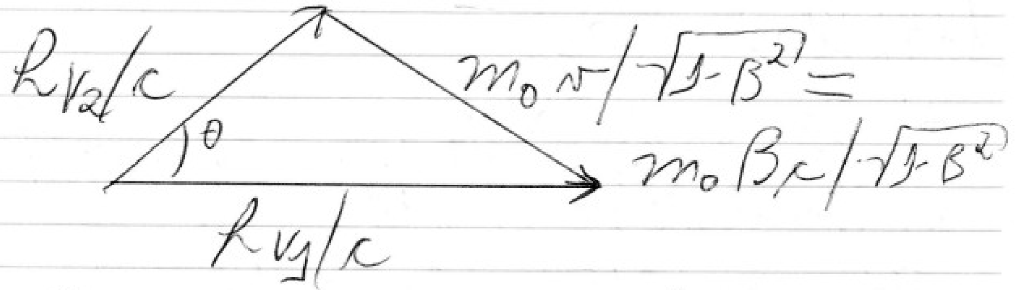

Conservation of momentum requires that the incident photon vector be equal to the vectors representing the scattered photon and the recoil electron:

and this (with the law of cosines)

)*+,)))-) '% / , A >*!
& ! / A $1 6!
(" %" 8

$$h = 6.55 \times 10^{-27} \text{ erg-sec.}\dagger$$

The precise agreement between these two calculations—which have nothing in common but the quantum postulates—is very impressive. Surely this is the strongest kind of evidence for the validity of the quantum theory in general, and for the Einstein photon concept in particular. But in 1915 it was not enough to establish any kind of general belief in photons. Even Millikan was skeptical. He admitted the empirical strength of the Einstein photoelectric equation—but no more: "I spent ten years of my life testing that 1905 equation of Einstein's, and, contrary to all my expectations, I was compelled in 1915 to assert its unambiguous experimental verification in spite of its unreasonableness. . . ." When Millikan described his results in 1916 he felt that the photon concept was hardly worthy of comment: "I shall not attempt to present the basis for such an assumption, for, as a matter of fact, it had almost none at the time." Nevertheless, he had to marvel at "the astonishing situation that these facts [the photoelectric equation] were correctly and exactly predicted nine years ago by a form of quantum theory which has now been pretty generally abandoned."

The Compton Effect

The story of the further trials of Einstein's quantum theory, and the accumulation of the mountain of evidence which finally brought photons into the world of physics, must be cut short for lack of space. We shall proceed to what was effectively the final episode in the story, the work of Arthur H. Compton on the scattering of X rays, first reported in 1923.

Compton's investigation concerned the "secondary radiation" produced by irradiating thin layers of such light elements as carbon with a "primary" beam formed in a conventional X-ray tube. Compton measured the wavelengths of the secondary, or scattered, radiation (using a Bragg crystal spectrometer) at various wavelengths of the primary radiation and at various angles with respect

---

† This calculation involves using the Planck distribution law (8) to derive an equation for the frequency ν_m or wavelength λ_m of maximum energy density and a theoretical version of the Stefan-Boltzmann equation (1). When measured values for ν_m and the Stefan constant σ are introduced into these two equations, two unknown quantities remain, h and the Boltzmann constant k. Planck's result for h is quoted; Planck obtained

$$k = 1.346 \times 10^{-16} \text{ erg deg}^{-1}.$$

to the direction of the primary beam. He found a difference in wavelength between the primary and secondary radiation. The difference was not large, but there was, unquestionably, a difference. The scattered wavelengths were always larger than those of the primary beam, and the magnitude of this wavelength was dependent only on the angle at which the secondary radiation was observed, and not at all on the scattering material or on the wavelength of the primary beam.

For some years before Compton's experiments it had been surmised that the scattering of X rays was due to interactions between the incident X rays and the electrons of the scattering material. A classical theory, based on the assumption that radiation was strictly wavelike, was available. However, when this theory was applied in a simple form to the X-ray–electron scattering interaction, it predicted that the primary and scattered radiation should have exactly the same wavelength, in direct disagreement with Compton's observations. After a futile attempt to elaborate the classical theory to fit the scattering data, Compton arrived at a highly successful quantum interpretation.

The central idea, as in Einstein's explanation of the photoelectric effect, is that the radiation beam (X rays, in this case) contains photons, each one of which carries an amount of energy $h\nu$. These are the same attractively simple radiation particles that Einstein had proposed, to no avail, almost twenty years earlier. If anything, the Compton concept of the photon is even simpler than that of Einstein. The photon-electron scattering interaction is, from the mechanical viewpoint, nothing more than a collision between two minute billiard balls. An electron belonging to an atom such as carbon is struck by an X-ray photon from the primary beam. Before the collision the electron is imagined to be stationary and free (actually it is bound by an atom, but the binding energy is negligible compared to the large energy which the photon brings to the collision), and the collision sets it in motion, billiard-ball fashion. To be sure, the electron, after the collision, is an exceedingly rapidly moving billiard ball. Its speed is so great that it must be described, from the relativistic viewpoint, as having a momentum of

$$m_0 v/\sqrt{1-\beta^2} \qquad \text{or} \qquad m_0 \beta c/\sqrt{1-\beta^2},$$

where m_0 is the electron's rest mass and β is v/c, the ratio between the electron's velocity v and the velocity of light c. The corresponding kinetic energy is $m_0 c^2(1/\sqrt{1-\beta^2} - 1)$. To represent the photon energy before and after the scattering collision, write $h\nu_1$ and $h\nu_2$, respectively; the corresponding momenta are $h\nu_1/c$ and $h\nu_2/c$.[†]

In Compton's words, the scattering event is "an X-ray particle colliding with an electron and bounding elastically from it with reduced energy, the lost energy appearing as the recoil energy of the electron." The momentum changes are shown with a vector diagram as follows:

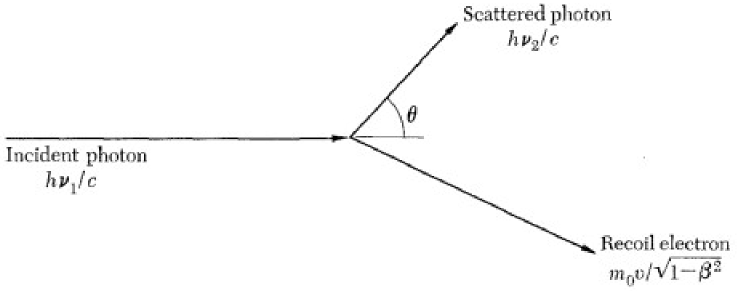

Conservation of momentum requires that the incident photon vector be equal to the vectors representing the scattered photon and the recoil electron:

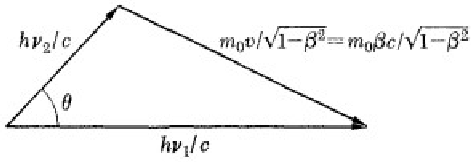

and this (with the law of cosines) leads to the equation

$$(m_0\beta c/\sqrt{1-\beta^2})^2 = (h\nu_1/c)^2 + (h\nu_2/c)^2 - 2(h\nu_1/c)(h\nu_2/c)\cos\theta.$$

Conservation of energy can also be stated, to the effect that the incident photon energy is equivalent to the total of the energy of

---

[†] That $h\nu/c$ is a measure of photon momentum can be seen by equating two Einstein equations for total energy, $E = h\nu$ and $E = mc^2$: $h\nu = mc^2$. The quantity on the left is the photon energy specified by Einstein's quantum theory and the quantity on the right is the total energy of any particle whose mass is m, as prescribed by Einstein's special theory of relativity. Since c is the velocity of the photon, mc is its momentum, and this, according to the above energy equation, is equivalent to $h\nu/c$. The situation is especially simple for the photon because it has no rest mass. For a particle with rest mass m_0 the momentum p is

$$\begin{aligned} p &= \sqrt{(mc^2)^2 - (m_0c^2)^2}/c \\ &= (\sqrt{m^2 - m_0{}^2})c. \end{aligned}$$

the scattered photon and the kinetic energy of the recoil electron:

$$h\nu_1 = h\nu_2 + m_0c^2(1/\sqrt{1-\beta^2} - 1).$$

This can be rearranged as:

$$m_0c^2/\sqrt{1-\beta^2} = h\nu_1 - h\nu_2 + m_0c^2$$
$$m_0c/\sqrt{1-\beta^2} = h\nu_1/c - h\nu_2/c + m_0c$$
$$(m_0c/\sqrt{1-\beta^2})^2 = (\quad)^2,$$

where ( ) has been written for the quantity $h\nu_1/c - h\nu_2/c + m_0c$. The last equation can be solved for β^2:

$$\beta^2 = 1 - m_0{}^2c^2/(\quad)^2,$$

so that

$$\begin{aligned}\beta^2(m_0c/\sqrt{1-\beta^2})^2 &= (\quad)^2 - m_0{}^2c^2\\ &= (h\nu_1/c - h\nu_2/c + m_0c)^2 - m_0{}^2c^2\\ &= (h\nu_1/c)^2 + (h\nu_2/c)^2 + 2m_0c(h\nu_1/c)\\ &\quad - 2m_0c(h\nu_2/c) - 2(h\nu_1/c)(h\nu_2/c).\end{aligned}$$

When this is subtracted from the momentum-balance equation, and the result is whittled down, a simple equation in the two frequencies (before and after the collision) is derived:

$$\nu_1/\nu_2 = 1 + 2(h\nu_1/m_0c^2)\sin^2(\theta/2).$$

Conversion to an equation in wavelength leads to †

$$\lambda_2/\lambda_1 = 1 + 2(h/m_0c\lambda_1)\sin^2(\theta/2), \quad \text{or} \quad \lambda_2 = \lambda_1 + (2h/m_0c)\sin^2(\theta/2).$$

The increase in wavelength $\lambda_2 - \lambda_1$ caused by the scattering of the primary X-ray beam at an angle of θ is therefore

$$(2h/m_0c)\sin^2(\theta/2) \qquad \text{or} \qquad 0.0484\sin^2(\theta/2)$$

when the wavelength is expressed in Angstroms (Å). For scattering at $\theta = 90°$ the theory calculates a wavelength increase of

$$(0.0484)/2 = 0.0242\ \text{Å},$$

regardless of the scattering material and the wavelength of the incident beam. Using X rays of wavelength 0.708 Å, Compton found an increase of 0.022 Å for the radiation scattered at 90°. When γ rays of wavelength 0.022 Å were used the scattered wavelength at 90° was 0.043 Å, an increase of 0.021 Å. In spite of the large difference between the primary X-ray wavelength and the γ-ray wavelength, the scattered wavelengths are nearly the same amount longer in the two cases, and this wavelength increase agrees reasonably well with the theoretical calculation. Compton used his basic premise to derive several other equations which could be checked experimentally. In all aspects the theory was successful.

---

† Wavelength and frequency for any electromagnetic disturbance are related by $\lambda\nu = c$, where λ is the wavelength. Thus, $\lambda_1/c = 1/\nu_1$, $\lambda_2/c = 1/\nu_2$, and $\nu_1/\nu_2 = \lambda_2/\lambda_1$.

With the publication of Compton's work, resistance to the photon concept was at last struck a mortal blow. But death came slowly and painfully. Compton wrote that his "results . . . immediately became a subject of the most lively scientific controversy that I had ever known." The chief opponent was William Duane of Harvard University, who had been willing to accept photons in other X-ray phenomena, but curiously opposed any thought of photons in the scattering process. For some reason Duane was unable to repeat Compton's scattering experiments in his own laboratory. There was a formal confrontation at a meeting of the American Physical Society, in which Compton and Duane defended their results and interpretations. Compton diplomatically decided not to press his criticism: "I might have criticized his interpretation of his results on rather obvious grounds, but thought it would be wiser to let Duane himself find the answer." † Later, Compton visited Harvard and Duane's laboratory. This visit seems to have been the occasion for some intense discussions; late one afternoon Compton was seen "disheveled, unshaven and obviously overtired." Duane's difficulty was not found then, but a new series of experiments was started which eventually turned up some spurious effects and Duane at last found himself in agreement with Compton. Compton's matter-of-fact assessment of his own work seems a good closing remark for this final scene in the story of the Einstein photon: "It may be fair to say that these experiments were the first to give, at least to physicists in the United States, a conviction of the fundamental validity of the quantum theory."

---

† After another Compton-Duane encounter, at the 1924 meeting of the British Association for the Advancement of Science, the Indian physicist C. V. Raman expressed himself to Compton with the remark: "Compton, you are good debater, but the truth is not in you." This story is told by Compton with the opinion that this same discussion may have inspired Raman to begin the investigation which led to the discovery of the "Raman effect," a scattering phenomenon involving ordinary photons which is in many ways, as it turns out, similar to the Compton X-ray scattering.

Duality

To do them justice, the enemies of the photon were disturbed by much more than the simple idea of particles of radiation. The "corpuscular" theory of radiation had long been known in physics: it was originally proposed by Newton. The difficulty was that nineteenth-century physics had very firmly established another, seemingly incompatible, theory of radiation. Such phenomena as refraction and diffraction had been studied with great care and explained to everyone's satisfaction on the assumption that radiation forms, such as light and X rays, have a wave-like structure. Among other things, there was a need to explain "interference" effects, the mixing of radiation beams to give alternating light and dark bands. The two beams seem to cancel in one part of the pattern and to reinforce in another part. As anyone knows who has watched water waves, this sort of behavior is possible where two trains of waves come together. A wave theory gives a natural and easy explanation of interference. A particle theory, on the other hand, cannot, at least apparently, explain interference patterns unless it is assumed that the particles from the two beams annihilate each other in the dark bands and produce more particles in the light bands. Such a theory is very unattractive to the physicist, who finds it advantageous to insist on the principle of conservation of energy with no complete annihilation processes and no creation of new particles.

So it was with nineteenth-century physics: a single picture, that of waves, sufficed to explain all the observed phenomena associated with radiation. Then came the quantum theory, Einstein's paper, and the unwelcome data on the photoelectric effect and X-ray scattering. Now there were observations which had simple *particle* explanations, and apparently no acceptable *wave* explanation. A strange duality had emerged: depending on how you looked at it, the same beam of light or X rays could seem to be either a shower of particles or a procession of waves. But which was it actually? Waves or particles?

Like most of the other perplexing problems of modern physics, this one was first recognized and partially resolved in the mind of Albert Einstein. His remarks in 1909 prophesied exactly the way physics would turn in about fifteen years:

> It is undeniable that there is an extensive group of data concerning radiation which shows that light has certain fundamental properties that can be understood much more readily from the

> standpoint of the Newton emission [particle] theory than from the standpoint of the wave theory. It is my opinion, therefore, that the next phase of the development of theoretical physics will bring us a theory of light that can be interpreted as a kind of fusion of the wave and emission theories.

In 1909 Einstein presented a theory which did indeed show a fusion of the wave and particle viewpoints. The central argument of this theory dealt with the statistical calculation of energy fluctuations. The elements of the calculation lie outside the scope of this chapter; it can only be said that Einstein's result showed two distinct kinds of fluctuations, one that could be traced to interference between wave trains and another that reflected deviations of a type familiar in collections of particles.†

The paradox of wave-particle duality, elaborated and extended in many different ways, and finally resolved by completely uprooting the great classical principle of determinism, became the dominant theme of the quantum theory in the 1920's. Finally, when applied to *matter*, that is, to electrons (which had previously been regarded strictly as particles, but suddenly in 1925 emerged in a wave manifestation), it brought forth the present form of the quantum theory, what is now known as "quantum mechanics." But that is a story for other chapters.

---

† This analysis made use of the full Planck radiation law, unlike Einstein's argument of the 1905 paper which was based on the Wien formula (3) and therefore restricted to low energy densities.

data 05 05 2020

(Pag.209) Whereas Lorentz had struggled to explain the invariance of the speed of light with a constructive theory that hypothesized motion-dependent molecular forces, Einstein bypassed all the complications by simply promoting the constancy to a postulate.

For Lorentz and his contemporaries, it was a problem, for Einstein a principle.

A2) The Quantum Physicists And an Introduction to their Physics (1970).pdf.

Pag. 30(43)

Conservation of momentum

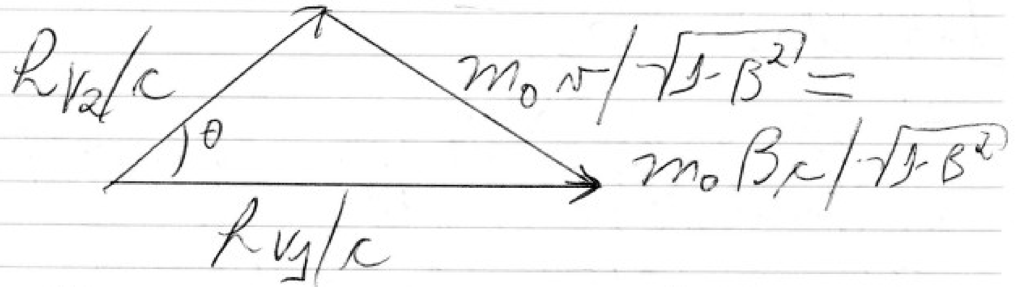

Conservation of momentum requires that the incident photon vector be equal to the vectors representing the scattered photon and the recoil electron:

and this (with the law of cosines)

05 05 2020

leads to the equation:

$$(m_0 \beta c / \sqrt{1-\beta^2})^2 =$$
$$(h\nu_1/c)^2 + (h\nu_2/c)^2 - 2(h\nu_1/c)(h\nu_2/c)\cos\theta$$

Obs:
That $h\nu/c$ is a measure of photon momentum can be seen by the equation two Einstein's equations for total energy,
$E = h\nu$ and $E = mc^2$: $h\nu = mc^2$.

B) Books_79_2020/
Traité de Coniques/

B1) Gaspard Monge Founder of "Constructive Geometry".pdf

Obs: look at: referentes
(find: gallica.bnf.fr)

B2) 517.pdf – Álgebra

B3) 522.pdf – Bolas de futebol, Arquimedes e ondeletas

B4) 523.pdf – Triângulos Belos

B5) 524.pdf – PI faraônico

B6) 525.pdf – As bolas da FIFA não são

7 ? G >1$> #/ M ! =

7)*+,)))-) '% . 6 ! < !
N 8 . AO 6!
-7 5 G

7)*+,)))-) '% 1 6!
* 7

7")*+,)))-) '% 6!
? ! 7 2 * = ! - !

7%)*+,)))-) '% " 6!
/ P ?

7)*+,)))-) '% % 6!
< Q

7')*+,)))-) '% 6!
? ! <

71)*+,)))-) '% 1 6!
* = ! ! F

7:)*+,)))-) '% : 6!
- H ! / ! 2 - H R C

7$)*+,)))-) '% $ 6!
3S ! ?

7)*+,)))-) '% " 6!
! ! - 8

7)*+,)))-) '% 1 6!
7 * S 6

No âmbito de uma colaboração entre a Gazeta e o Atractor, este é um espaço da responsabilidade do Atractor, relacionado com conteúdos interativos do seu site *www.atractor.pt* . Quaisquer reações ou sugestões serão bem-vindas para *atractor@atractor.pt*

TRIÂNGULOS BELOS

Verificaremos como a propriedade de os três lados de um triângulo plano serem iguais resulta de uma combinação harmoniosa, e única, de lados e ângulos.

Um triângulo diz-se equilátero se os quocientes entre os comprimentos dos seus lados forem iguais a 1. Que triângulos do plano têm estes três quocientes racionais? Dado um triângulo no plano com lados de comprimentos a, b e c tais que existem números naturais p_1, p_2, p_3, q_1, q_2 e q_3 que verificam $a/b = p_1/q_1$, $b/c = p_2/q_2$ e $a/c = p_3/q_3$ podemos reescalonar o triângulo e obter outro semelhante com lados racionais. Para isso, basta usar a homotetia $H_1 : (x, y) \in \mathbb{R}^2 \mapsto (x/c,\ y/c)$, que transforma o triângulo inicial num de lados $a' = (p_1\, p_2)/(q_1\, q_2)$, $b' = p_2/q_2$ e $c' = 1$. Se, de seguida, aplicarmos a homotetia $H_2 : (x, y) \in \mathbb{R}^2 \mapsto (q_1\, q_2\, x,\ q_1\, q_2\, y)$, obtemos um triângulo de lados inteiros.

Se o triângulo é retângulo, com a versão semelhante que acabámos de construir entramos no mundo vasto dos triângulos pitagóricos. A figura 1 ilustra um deles: note-se que, além de os lados terem comprimentos inteiros, os três ângulos têm cossenos racionais, pois são $\pi/2$, $\arccos(3/5)$ e $\arccos(4/5)$.

Mas há muitas outras possibilidades. Veja-se, por exemplo, o triângulo escaleno de lados com comprimentos 5, 6 e 7: neste caso, os ângulos também têm cossenos racionais pois são $\arccos(1/5)$, $\arccos(5/7)$ e $\arccos(19/35)$. Repare-se agora no outro triângulo da figura 2: os quocientes entre as amplitudes dos ângulos são racionais, mas nem todos os quocientes dos lados o são.

Se, todavia, juntarmos as condições

Q_1: *Os quocientes entre os comprimentos dos três lados são racionais.*

Q_2: *Os quocientes entre as amplitudes dos três ângulos são racionais.*

então ficamos reduzidos aos triângulos equiláteros.

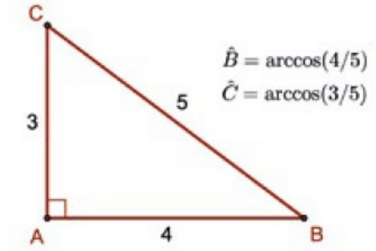

Figura 1

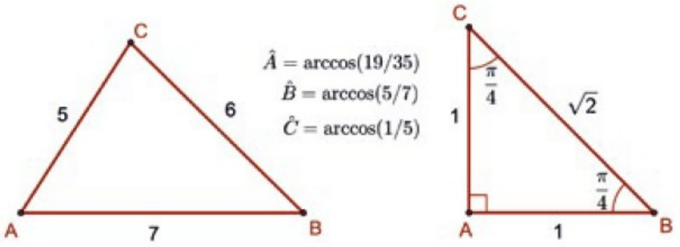

Figura 2

Obviamente, um triângulo equilátero satisfaz as duas hipóteses Q_1 e Q_2, com a particularidade de os seis quocientes em causa serem iguais a 1. Vejamos que também é válida a implicação recíproca, ou seja:

> $\mathcal{P}_1$: *Se, num triângulo, são racionais os quocientes entre os comprimentos dos três lados e entre as amplitudes dos três ângulos, então o triângulo é equilátero.*

Seja $\mathcal{T}$ um tal triângulo de lados a, b, c e ângulos $\angle A$, $\angle B$, $\angle C$. Reescalonando-o, como se explicou anteriormente, podemos supor que a, b, c são racionais (ou mesmo inteiros). Além disso, podemos reescrever $\angle A = \pi\,\alpha_A$, $\angle B = \pi\,\alpha_B$ e $\angle C = \pi\,\alpha_C$ para uma escolha adequada de reais positivos α_A, α_B, α_C. Acrescente-se que, pela hipótese Q_2, existem racionais positivos s_1, s_2 e s_3 tais que $\angle A/\angle B = s_1$, $\angle B/\angle C = s_2$ e $\angle A/\angle C = s_3$. Logo, tem-se $\alpha_A = s_3\,\alpha_C$ e $\alpha_B = s_2\,\alpha_C$, e, como a soma dos ângulos de um triângulo plano é igual a π, deduzimos que $s_3\,\alpha_C + s_2\,\alpha_C + \alpha_C = 1$. Consequentemente, $\alpha_C = 1/(1+s_2+s_3) \in \mathbb{Q}$. Em resumo: Se os quocientes das amplitudes dos ângulos de um triângulo plano são racionais, então os ângulos são múltiplos racionais de π.

Neste momento é oportuno reunirmos o que já sabemos sobre $\mathcal{T}$:

1. É semelhante a um triângulo com lados racionais (consequência de Q_1).
2. Os seus ângulos são múltiplos racionais de π(consequência de Q_2).

Conseguiremos obter mais informação sobre a forma de um tal triângulo se aliarmos as duas propriedades anteriores, o que pode ser feito através do Lei dos Cossenos. Sendo os lados a, b, c racionais, então

$$\cos(\angle A) = \frac{b^2+c^2-a^2}{2bc},\ \cos(\angle B) = \frac{a^2+c^2-b^2}{2ac} \text{ e}$$

$$\cos(\angle C) = \frac{a^2+b^2-c^2}{2ab}$$

são números racionais. Este é um dado muito bem-vindo, porque se conhecem todos os ângulos $\theta \in [0,\pi]$ que são múltiplos racionais de π e têm cosseno racional: θ é 0, π, $\pi/2$, $\pi/3$ ou $2\pi/3$, cujos cossenos são, respetivamente, ± 1, 0 e $\pm 1/2$. Com este resultado (pode ver-se uma prova em [2]), estamos em condições de estabelecer a caracterização que procurávamos para o triângulo $\mathcal{T}$. Realmente, $\mathcal{T}$ tem de ser equilátero porque é semelhante a um triângulo com lados racionais, cujos ângulos estão em $]0,\pi[$ mas nenhum deles é $\pi/2$ (caso contrário, um dos outros dois teria de ser menor ou igual a $\pi/4$ e o seu cosseno não teria um valor permitido) nem $2\pi/3$ (por razão idêntica); logo, os três ângulos são iguais a $\pi/3$.

Cada uma das propriedades $\mathcal{P}_1$ e

> $\mathcal{P}_2$: *Se $\theta \in [0,\pi]$ é um múltiplo racional de π e tem cosseno racional, então $\theta \in \{0, \pi, \pi/2, \pi/3, 2\pi/3\}$.*

pode ser deduzida a partir da outra.

Para o provar, e uma vez que $\cos(\{0,\pi,\pi/2\}) = \{1,-1,0\}$ e que $\cos(\theta) = -\cos(\pi-\theta)$, podemos restringir a análise a ângulos de $]0,\pi/2[$. Seja $\theta \in]0,\pi/2[$ um ângulo cuja amplitude é um múltiplo racional de π e tem cosseno racional. Construa-se o triângulo da figura 3, de lados 1, 1, $2\cos\theta$ e ângulos θ, θ, $\pi - 2\theta$. Observe-se que estes três ângulos são múltiplos racionais de π. Além disso, tendo em conta a hipótese de que $\cos\theta$ é um número racional, os quocientes dos comprimentos dos lados deste triângulo são racionais. Logo, podemos aplicar $\mathcal{P}_1$ e concluir que o triângulo é equilátero. E, portanto, $\theta = \pi/3$.

Note-se que, no que aqui ficou dito, fizemos uso de várias propriedades dos triângulos planos que não são válidas em geometrias não euclidianas, ainda que nelas um triângulo equilátero possa ser também equiângulo. Por exemplo, na esfera de raio $2/\pi$ podemos traçar o triângulo (pitagórico) de lados 1, 1, 2 e ângulos $\pi/2$, $\pi/2$, π (figura 4). E, nela, estão também: o triângulo com estes mesmos lados mas de ângulos $3\pi/2$, $3\pi/2$ e π; um triângulo equilátero cujos três lados se situam no equador e que tem os três ângulos iguais a π; um triângulo pitagórico e equilátero de lado 1 e com os três ângulos iguais a $\pi/2$. (Podem construir-se imagens destes triângulos em [1]). Qual será a versão esférica da caracterização dos triângulos equiláteros planos?

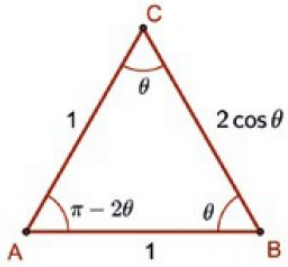

Figura 3

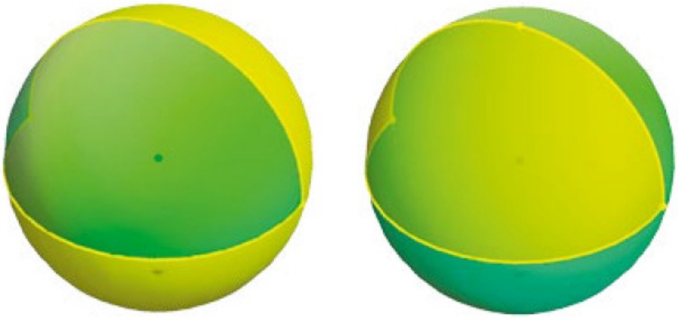

Figura 4

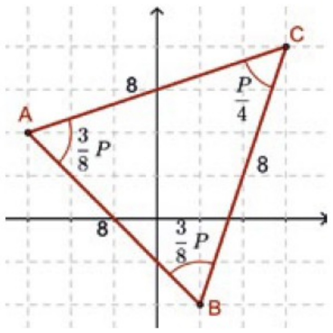

Figura 5

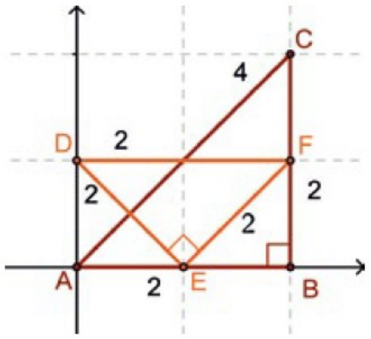

Figura 6

Seria também interessante analisar, na perspetiva anterior, as propriedades dos triângulos equiláteros planos quando optamos por outros modos de medir as distâncias em $\mathbb{R}^2$. Considere-se, por exemplo, a geometria-do-táxi. Neste modo de medir comprimentos, uma circunferência tem a forma de um quadrado da geometria euclideana com os lados fazendo 45º com os eixos coordenados; π é substituído por $P = 4$; e a soma dos ângulos de qualquer triângulo é P. Contudo, um triângulo equilátero pode não ser equiângulo. Observe-se, por exemplo, o triângulo da figura 5 cujos lados medem 8, dois dos ângulos têm amplitude $^3/_8P$ e o terceiro mede $P/4$.

E há outras diferenças. O Teorema de Pitágoras não admite uma extensão a esta métrica, como mostram os triângulos DEF e EBF da figura 6, um de lados 1, 1, 2, ângulo retângulo em B e hipotenusa medindo 2, e outro que é equilátero, retângulo em E e cuja hipotenusa também mede 2. Adicionalmente, falham testes de semelhança que foram fundamentais no argumento anterior. Por exemplo, os triângulos ABC e DEF da figura 6 indicam que um triângulo nesta geometria não está univocamente determinado se forem conhecidos dois lados e o ângulo por eles formado.

Haverá algum critério, análogo ao que vimos no plano com a métrica euclidiana, para testar se um triângulo na geometria-do-táxi é equilátero? Para explorar esta questão, o leitor é convidado a utilizar o módulo interativo em [2], que permite calcular rapidamente distâncias e ângulos nesta métrica. (Uma imagem dessa utilização consta da figura 7.)

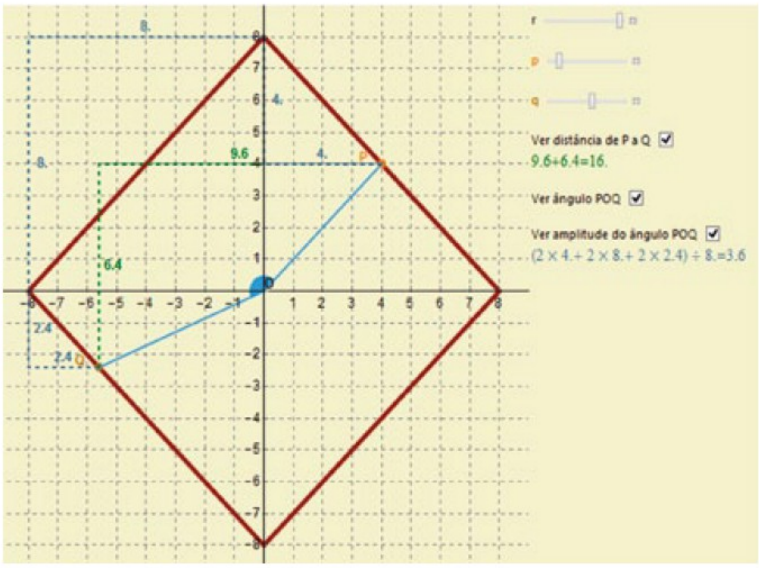

Ver distância de P a Q ☑
9.6+6.4=16.
Ver ângulo POQ ☑
Ver amplitude do ângulo POQ ☑
(2 × 4.+ 2 × 8.+ 2 × 2.4) ÷ 8.=3.6

Figura 7

REFERÊNCIAS

[1] *http://atractor.pt/mat/GeomEsf/index.htm*

[2] *http://www.atractor.pt/mat/triangulos_belos*

II FEIRA DA MATEMÁTICA
NO MUSEU NACIONAL DE HISTÓRIA NATURAL
E DA CIÊNCIA DA UNIVERSIDADE DE LISBOA

23|24 OUTUBRO 2015

MARQUE NA AGENDA!
SEXTA-FEIRA | ESPECIAL ESCOLAS
SÁBADO | PÚBLICO GERAL

EXPOSIÇÕES
WORKSHOPS
JOGOS E DESAFIOS
DEMONSTRAÇÕES
CIRCO MATEMÁTICO
PALESTRAS

muhnac
spm
APM
Ludus
MPT

PI FARAÓNICO

Jorge Nuno Silva
Universidade
de Lisboa
jnsilva@cal.berkeley.edu

O Papiro de Rhind, obra incontornável no estudo da matemática na Antiguidade, dá-nos muitas razões para nos surpreendermos com as capacidades dos egípcios de há milénios. Escrito no século XVII a.C., sendo cópia de trabalho mais antigo, é essencialmente uma colecção de problemas e soluções. Hoje falaremos um pouco do Problema 50, que, em linguagem anacrónica, pode dizer-se tratar da quadratura do círculo.

A interpretação de texto tão antigo não pode ser sempre consensual. A lacónica escrita do escriba Ahmes não facilita a vida aos estudiosos. No caso do nosso problema 50, a questão parece ser a de obter um valor para a área de um círculo de diâmetro d.

O procedimento descrito consiste em aproximar o círculo por um octógono, cuja área seja fácil de determinar.

Comecemos por considerar um quadrado circunscrito à figura.

Vamos dividir cada lado em nove partes, cada uma de comprimento $d/9$. Conservando o terço central de cada lado e unindo as extremidades, obtemos um octógono, que, como o quadriculado da figura sugere, aproxima *bem* o círculo.

O quadriculado da ilustração permite determinar com facilidade a área do polígono. Se tomarmos para unidade o quadradinho de lado $d/9$, esta área será

$$81 - 4 \times \frac{3 \times 3}{2} = 63\,.$$

Ora, 63 é um número que ... é quase 64 e este é um quadrado perfeito: $64 = 8^2$. Tomemos então 64 como valor aproximado para a área do círculo!

Em linguagem corrida: a área de um círculo de diâmetro d é aproximadamente igual à de um quadrado de lado $(8/9)d$.

Usando o diâmetro original $d = 2r$, temos, para valor da área do círculo, $A_{\bigcirc}$,

$$A_{\bigcirc} = \left(\frac{8}{9}d\right)^2 = \left(\frac{16}{9}\right)^2 r^2\,.$$

Como sabemos, a fórmula correta é $A_{\bigcirc} = \pi r^2$, pelo que a expressão deduzida do argumento de Ahmes corresponde a tomar

$$\pi = \left(\frac{16}{9}\right)^2 = 3.160493827\ldots$$

que difere do valor exato por menos de 0.02. Esta aproximação é notável em si e impressiona também pela facilidade

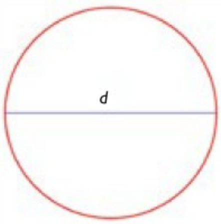

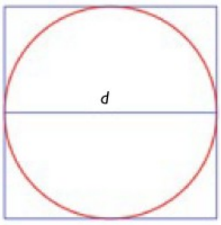

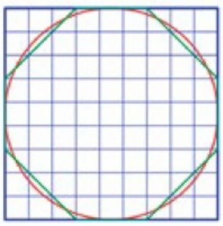

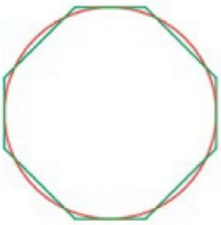

com que é obtida. Sugerimos ao leitor que tente obter outras aproximações usando métodos semelhantes a este, a ver se consegue melhorar este resultado ...

Suponhamos agora que se usava a mesma figura para aproximar o perímetro do círculo.

Tomando de novo como unidade $d/9$ e após uma aplicação do Teorema de Pitágoras, obtemos para o perímetro do octógono o valor

$$\frac{4}{3}\left(1+\sqrt{2}\right)d\,.$$

Como o valor correto para o perímetro do círculo é $P_{\bigcirc} = \pi d$, o valor aproximado baseado na aproximação pelo octógono corresponde a tomar

$$\pi = \frac{4}{3}\left(1+\sqrt{2}\right) = 3.218951\ldots$$

que difere do valor exato por menos de 0.08.

É certo que os egípcios antigos não conheciam a bela estrutura da matemática de Euclides ou Arquimedes, mas, não esqueçamos, eles construíram as impressionantes Pirâmides, testemunhos perenes do seu engenho.

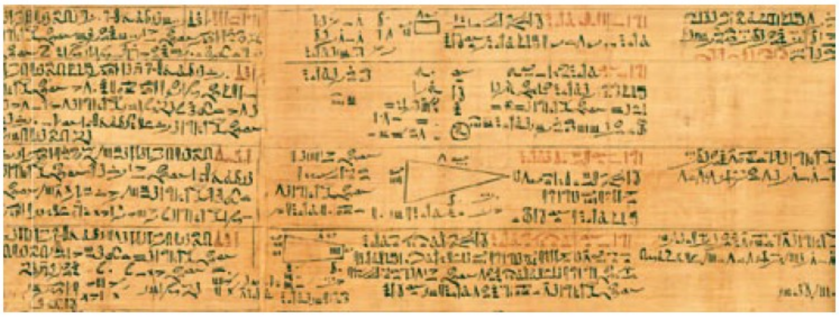

Papiro de Rhind

Já é sócio da SPM?

spm
SOCIEDADE PORTUGUESA DE MATEMÁTICA

Conheça as vantagens e saiba
como aderir em www.spm.pt
ou através do número 217 939 785

Consulte também as condições para os sócios institucionais (Departamentos, Faculdades, ESES, Politécnicos, etc.)

CANTO DÉLFICO

Eduardo Marques de Sá
Universidade
de Coimbra
emsa@mat.uc.pt

AS BOLAS DA FIFA NÃO SÃO GRANDE COISA

Trunquem um pouco mais fundo o icosaedro truncado e farão da Telstar uma bola melhor.

Negócio de bolas é assunto sério. Há laboratórios sofisticados inovando para melhor vender, e adivinham-se empresas de *marketing* com a missão de conceber ideias e ficções que a FIFA certifica. São exemplos a publicidade hiperbólica quadrienal e a caução 'científica' de inovações mal experimentadas. O caso da Jabulani, a bola do Mundial 2010, deu que falar; aos elogios dos vendedores – a mais perfeita de sempre, de aerodinâmica superlativa, nunca vista, etc. – opuseram-se os protestos de arqueiros, zagueiros e pontas-de-lança, que não a pouparam: horrorosa, dançarina, bola de supermercado. Houve mesmo quem lhe atirasse as culpas pelo fiasco do Mundial 2010: é que de entre todos os mundiais ele ocupa o penúltimo lugar na média de golos por jogo.

Mas o problema a tratar aqui é outro. Seria pouco razoável fabricar uma bola de futebol com quatro peças triangulares (de cabedal ou material sintético) cosidas de modo a formar um tetraedro. Depois de insuflado ar, a pressão arredondaria a bola de raiz tetraédrica, mas o jogo seria outro, mais imprevisível do que o râguebi. O princípio lógico é o de que quanto mais rotundo for o poliedro não inflado, mais rotunda e resistente será a bola depois de cheia.

A UEFA escolheu, nos anos 60 do século passado, o icosaedro truncado como modelo, com as faces pentagonais pintadas de negro, a que foi dado o *nom de guerre* Telstar, roubado a um então famoso satélite de telecomunicações. Anos depois, a Telstar venceu o concurso 'Miss Bola' para o Mundial do México-1970. Hoje, contas feitas, o arquimediano a preto e branco ou a cores por encomenda já leva mais de meio século de reinado.

Em 1981, J. M. Göthals e J. J. Seidel, num artigo justamente intitulado *The football*[1], mostraram que a bola podia ser melhorada. Os argumentos envolvem teoria dos invariantes, fórmulas de quadratura de Sobolev e órbitas do grupo icosaedral. A FIFA não se deixou convencer e, tanto quanto sabemos, a bola continuou a mesma.

O Canto Délfico espera ter mais sucesso junto das autoridades com uma abordagem bem mais modesta. Usando dois critérios elementares, distintos dos de 81, somos levados à mesma conclusão: a Telstar não está bem.

QUÃO *ROTUNDA* É UMA BOLA?

Expressões da linguagem quotidiana do género "este sólido é mais rotundo do que aquele" podem precisar-se por muitos métodos. Um dos mais simples começa por chamar *raio interno* de um sólido[2] ao maior dos raios das esferas nele contidas, e *raio externo* de um sólido ao menor dos raios das esferas que o contêm. Em seguida, define-se *esfericidade* do sólido como sendo a razão do raio interno pelo raio externo. Claro que

> *A esfericidade é um número positivo que não excede 1, sendo igual a 1 quando e só quando o sólido é uma esfera.*

Outro critério de avaliação da rotundidade baseia-se no chamado quociente isoperimétrico (QI) de um sólido S, que se define por

$$36\pi \frac{V^2}{A^3},$$

onde V é o volume de S e A a área da sua superfície envolvente. Este quociente satisfaz uma propriedade elementar importante: qualquer homotético de S tem o mesmo QI que S. O factor 36π destina-se exclusivamente a obter 1 como quociente isoperimétrico das esferas.

O teorema isoperimétrico tridimensional pode enunciar-se assim:

> *O quociente isoperimétrico dum sólido é um número positivo que não excede 1, sendo igual a 1 quando e só quando o sólido é uma esfera.*

Estamos perante um enunciado muito simples, com um lastro histórico notável e uma demonstração de elevado grau de dificuldade, em claro contraste com o gémeo trivial sobre a esfericidade.

Cada uma das funções, esfericidade e quociente isoperimétrico, permite comparar quaisquer sólidos quanto à sua rotundidade. Mas são critérios diferentes que podem, por isso, conduzir a conclusões díspares. Por exemplo, os cinco platónicos ficam assim ordenados, por ordem crescente dos seus quocientes isoperimétricos: tetraedro (.3023), cubo (.5236), octaedro (.6046), dodecaedro (.7547), icosaedro (.8288). Na corrida da esfericidade ganham o icosaedro e o dodecaedro, *ex-æquo* (.7947), o cubo e o octaedro ficam em terceiro lugar, *ex-æquo* (.5774), e o tetraedro em quinto (.3333), sozinho por não ter irmão dual.[3]

No caso dos poliedros arquimedianos, os dois critérios conduzem a ordenações algo diferentes; por exemplo, os quatro mais rotundos são os mesmos para os dois

Esfericidade à moda da FIFA

Na sua página digital das bolas, uma foto sugere o modo como avaliam a esfericidade: entalam a bola, sem a pressionar, entre duas placas planas paralelas e medem com grande rigor a distância entre os planos. Dizem repetir a medida para 45.000 posições da bola. O quociente do mínimo pelo máximo dessas distâncias é a esfericidade-*F*, da FIFA. Destaca o *site* que as bolas da classe FIFA PRO apresentam uma esfericidade-*F* de cerca de 98,5%.

Sugerimos que a bola oficial passe a ter a forma de castanha de revolução que a figura ilustra, com esfericidade-*F* inultrapassável.

Isoperimétrico O termo tem a sua raiz num problema bidimensional, famoso e muito antigo: *de entre as curvas planas simples fechadas com certo perímetro fixo, quais as que delimitam maior área?* O assunto foi desenvolvido na Grécia Antiga, datando desse período a intuição correcta de que as circunferências são a chave da questão. As referências remontam aos antigos pitagóricos, tendo sido, no entanto, o ateniense Zenodoro (séc. II a.C.) quem nos legou os resultados mais relevantes. Por testemunhos indirectos sabemos que ele provou teoremas isoperimétricos relativos a polígonos, como este: *de entre os polígonos fechados de perímetro fixo e um número fixo de vértices, os regulares são os que delimitam maior área.* Os *Elementos* de Euclides já continham a prova de que o quadrado tem a maior área de entre os rectângulos isoperimétricos.

Porém, só no terceiro quartel do séc. XIX, com K. Weierstrass à cabeça, surgiram as primeiras provas convincentes do famoso teorema isoperimétrico bidimensional.

---

[1] Vejam-se os artigos [1] - [3].

[2] Para tornar as coisas mais simples, *sólido* significa conjunto convexo, fechado, limitado e com, pelo menos, um ponto interior.

[3] Estes e outros números foram arredondados a quatro dígitos significativos.

critérios mas por ordens distintas, como mostra a tabela seguinte, onde *esf.* e *QI* abreviam esfericidade e quociente isoperimétrico:

	esf.	*QI*
Rombicosidodecaedro	.9246	.9390
Dodecaedro roscado[4]	.9189	.9470
Icosaedro truncado	.9150	.9032
Grande rombicosidodecaedro	.9049	.9136

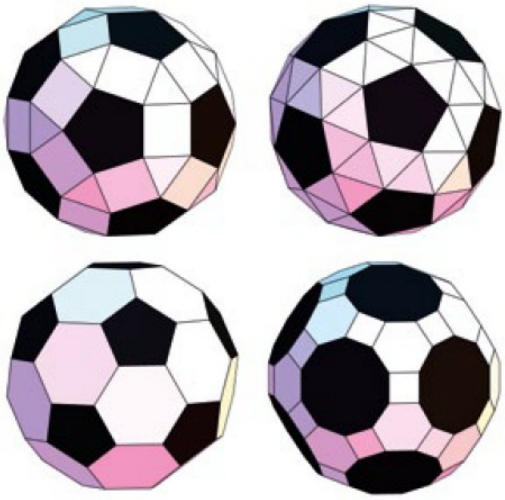

Figura 1. Os quatro mais rotundos.

Zenodoro em 3D? *De entre os poliedros com n vértices (ou faces) quais são os que optimizam o QI, ou outro critério de rotundidade?* Cada valor de n origina problemas de dificuldade em geral elevada, nos quais, contrariamente ao caso 2D de Zenodoro, a regularidade não ajuda muito. Por exemplo, se n é ímpar, platónicos e arquimedianos ficam de fora, pois todos eles têm números pares de vértices e faces.

Para $n = 4$, o tetraedro regular é optimal, como o leitor interessado facilmente adivinha e comprova. Para cinco vértices, surge o problema de haver dois esqueletos (grafos de arestas) não isomorfos, os que a figura mostra, correspondentes à pirâmide quadrangular e à bipirâmide triangular.

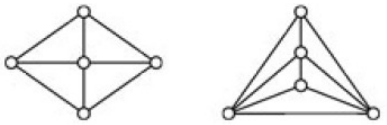

Neste caso, os poliedros mais rotundos não são os mesmos para o QI e para a esfericidade, e não têm faces regulares. Saltando para o caso de oito vértices, o cubo tem menor QI e menor esfericidade do que o anti-prisma de base quadrada; mas a questão não fica resolvida só com isto, pois existem 257 esqueletos possíveis de poliedros com oito vértices; veja-se em [4] a sequência

1, 2, 7, 34, 257, 2606, 32.300,...

A figura 1 mostra os quatro vencedores, fortes candidatos a bolas da FIFA, pintados a preceito. Como sabemos, há mais de meio século a UEFA seleccionou o icosaedro truncado como modelo das bolas de futebol. A escolha parece acertada, dadas a pouca formosura do Grande Rombi e a complexidade estrutural dos dois mais rotundos – imaginem, por exemplo, o que seria coser à mão 12 pentágonos e 80 pequenos triângulos de cabedal num total de 150 costuras!

TRUNQUEM MAIS FUNDO!

Imaginemos que a FIFA, num momento de lucidez, encomendava à *Gazeta* o desenho duma bola de futebol à moda antiga. A ideia mais imediata consiste em tomar um dos cinco platónicos e truncar-lhe os vértices quanto baste. Por motivos estéticos, dinâmicos e funcionais, a truncatura deve consistir na remoção de pequenas pirâmides regulares, uma por cada vértice e de modo igual para todos os vértices. Admitindo que o nosso platónico tem arestas de medida 1, chamamos *profundidade* da truncatura à medida t das arestas laterais das pirâmides removidas. A figura 2 mostra o mais rotundo dos platónicos – e, por isso, o melhor candidato à truncatura – truncado a diversas profundidades.

A questão matemática cuja solução oferecemos à FIFA é esta: escolhido um critério de rotundidade e um sólido platónico,

> *Que profundidade(s) de truncatura desse platónico optimiza(m) a rotundidade?*

Mais precisamente, neste Canto Délfico resolvemos o caso do icosaedro truncado e convidamos o leitor a mandar à FIFA o tratamento dos outros quatro platónicos truncados.

O primeiro critério a considerar vai ser o de mais fácil tratamento, a esfericidade. Seja $\mathfrak{I}_t$ o icosaedro truncado

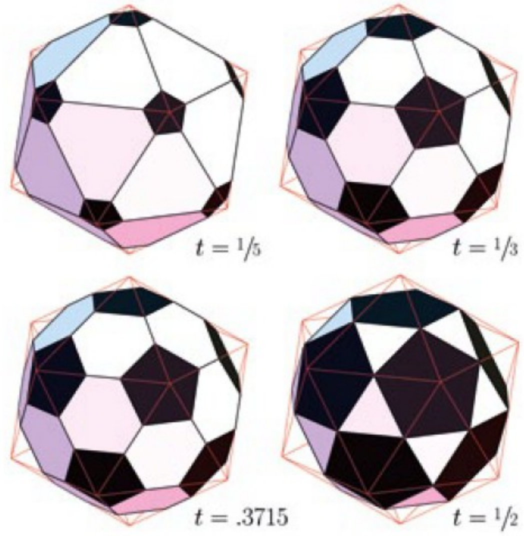

Figura 2. Truncaturas de um icosaedro.

à profundidade t; denotem-se por $r(t)$ e $R(t)$ os raios interno e externo de $\mathfrak{I}_t$. Por definição, a esfericidade de $\mathfrak{I}_t$ é

$$\epsilon(t) = \frac{r(t)}{R(t)}.$$

Note-se que t percorre o intervalo $[0, 1/2]$. Crescendo t, cada vértice de $\mathfrak{I}_t$ aproxima-se do centro de cada face triangular a que pertence, pelo que $R(t)$ é estritamente decrescente. O raio $r(t)$ é a menor das distâncias das faces hexagonais e pentagonais ao centro de $\mathfrak{I}_t$; portanto, $r(t)$mantém-se constante para t num intervalo $[0, c]$, onde c é o número *crítico* para o qual todas as faces de $\mathfrak{I}_c$ são tangentes à esfera interna do icosaedro.

Para qualquer arquimediano, as faces mais próximas do centro são as de maior número de vértices. Por escolha da FIFA, a Telstar é arquimediana, com $1/3$ de profundidade de truncatura. Portanto, $c > 1/3$, o que prova, para o critério vigente, que

> *A Telstar não é boa bola, pois pode aumentar-se a esfericidade truncando o icosaedro um pouco mais fundo.*

Não será difícil mostrar que o valor crítico de t é dado por

$$c = \frac{12\sqrt{5} - (\sqrt{54} + \sqrt{30})\sqrt{5 - \sqrt{5}}}{12(\sqrt{5} - 1)},$$

com valor aproximado .3715. A Telstar crítica está representada na figura 2; as suas faces hexagonais são irregulares, com arestas de dois comprimentos distintos, na razão de 1 para .6920. A prova da sua optimalidade e a determinação da esfericidade óptima exigem mergulhar as mãos em pormenores.

CONTAS

A determinação de uma fórmula para a esfericidade $\epsilon(t)$ requer alguns cálculos envolvendo a geometria do triângulo, do pentágono e do hexágono regulares, a indispensável percepção tridimensional e aplicações adequadas do teorema de Pitágoras. É complicado, aborrecido, mas elementar. A figura 3 mostra um corte do icosaedro (de arestas 1) em vias de ser truncado à profundidade t.

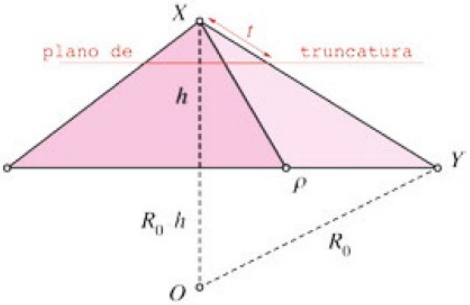

Figura 3. Corte do icosaedro.

Em destaque estão o centro O do icosaedro e dois dos seus vértices, X e Y. O polígono colorido representa a pirâmide pentagonal do icosaedro com vértice em X; o plano da base está de topo, tal como o plano de truncatura, e a aresta $[X, Y]$ está no plano da figura. Interessa calcular o raio externo ρ da base da pirâmide, a sua altura h e o raio externo R_0 do icosaedro:

$$\begin{aligned} \rho &= \sqrt{\frac{1}{10}(5 + \sqrt{5})} \\ h &= \sqrt{1 - \rho^2} \\ R_0 &= \frac{1}{2h} = \sqrt{\frac{1}{8}(5 + \sqrt{5})}. \end{aligned}$$

---

[4] Tentativa de tradução do inglês *snub*. A "snubificação" de um poliedro consiste em tomar o invólucro convexo da configuração que se obtém por expansão, *i.e.*, por translação radial das faces do poliedro, acompanhada de rotação planar de cada face em torno do seu centro.

Da figura 3 facilmente resulta que $R_0 - ht$ é a distância das faces pentagonais de $\Im_t$ ao seu centro, e que o raio interno do icosaedro é

$$\Delta = \frac{1}{4}\sqrt{3} + \frac{1}{12}\sqrt{15}.$$

O raio externo $R(t)$ também se 'vê' em verdadeira grandeza na figura. De tudo isto resulta

$$\begin{aligned} r(t) &= \min\{\Delta, R_0 - ht\} \\ R(t) &= \sqrt{R_0^2 - (1-t)t}. \end{aligned}$$

O parâmetro crítico é a solução da equação $\Delta = R_0 - ht$; calculando e simplificando obtemos o valor de c acima anunciado. Para $t \geqslant c$, a esfericidade é

$$\epsilon(t) = \frac{R_0 - ht}{\sqrt{R_0^2 - (1-t)t}}$$

com numerador e denominador decrescentes. A identidade

$$2h\frac{R(t)^4}{R_0 - ht}\epsilon(t)\epsilon'(t) = (h^2 - 1)t,$$

válida para $t \geqslant c$, mostra que $\epsilon'(t) < 0$ e que $\epsilon(t)$ é decrescente no intervalo em estudo. Portanto, $\Im_c$ é optimal. Pode determinar-se uma expressão exacta, com radicais, para a esfericidade óptima, pouco útil pela complexidade que se adivinha; o valor aproximado

$$\epsilon(c) \approx .9226$$

mostra que o icosaedro decentemente truncado produz uma bola de esfericidade superior à do dodecaedro roscado.

O *QI* DUM ICOSAEDRO TRUNCADO

Sejam V_0 e A_0 o volume do platónico que vai truncar-se e a área da sua superfície limítrofe. Considera-se, em dado vértice X, a pirâmide P que é removida quando truncamos X à profundidade $t=1$. Sejam v_P o volume de P e δ_P a diferença entre a sua área lateral e a da base. O volume e a área da superfície limítrofe do platónico truncado à profundidade t são dados por

$$V(t) = V_0 - nv_Pt^3 \quad \text{e} \quad A(t) = A_0 - n\delta_Pt^2,$$

onde n denota o número de vértices do platónico em causa. No caso do icosaedro, que está na berlinda, $n=12$ e P é a pirâmide da figura 3. (No caso do tetraedro, P é o tetraedro.) O quociente isoperimétrico do platónico truncado é, pois,

$$q(t) = \frac{V(t)^2}{A(t)^3}.$$

As identidades

$$\begin{aligned} A^6\left(\frac{V^2}{A^3}\right)' &= 2VV'A^3 - 2A^2A'V^2 \\ &= 6nVA^2t(\delta_PV - v_PAt) \\ &= 6nVA^2t(\delta_PV_0 - v_PA_0t) \end{aligned}$$

mostram que o sinal da derivada de $q(t)$ é o sinal de $\delta_PV_0 - v_PA_0t$. Portanto, a derivada de $q(t)$ anula-se para um só valor de t, nomeadamente

$$t_0 = \frac{\delta_PV_0}{v_PA_0},$$

$q'(t)$ é positiva para $t < t_0$ e negativa para $t > t_0$. Sendo assim, a truncatura de profundidade t_0 é a única que maximiza o quociente isoperimétrico. A surpresa é que t_0 é o parâmetro crítico c encontrado acima!

> *O icosaedro truncado vencedor do concurso da esfericidade é, também, o de maior QI.*

Mais um desaire para a FIFA, que assim perde em três tabuleiros. O que importa ao matemático, criatura pouco sensível a estas vaidades e que desconfia de coincidências, é encontrar motivos da surpresa mais nobres do que as contas que a ela conduziram.

Ao Professor João Queiró agradeço os muitos aperfeiçoamentos do texto.

REFERÊNCIAS

[1] J. M. Göthals e J. J. Seidel, "The Football", *Nieuw Arch. Wisk.* (3), 29(1981), pp. 50-58.

[2] J. M. Goethals, J. J. Seidel, "Cubature formulae, polytopes and spherical designs", in C. Davis *et al.* (ed.s), *The Geometric Vein: The Coxeter Festschrift*, Springer-Verlag, New York, 1981, pp. 203-218.

[3] R. Hardin e N. Sloane, "McLaren's improved snub cube and other new spherical designs in three dimensions", *Discrete Comput. Geom.*, 15(1996), pp. 429-441.

[4] "The On-Line Encyclopedia of Integer Sequences", *https://oeis.org/*, sequência *A000944*, vista em 16-6-2015.

O autor escreve de acordo com a grafia antiga.

Αρχιμηδης ο γεωμετρης

ARQUIMEDES E O VOLUME DA ESFERA

Fátima Vinagre
Escola Secundária da Azambuja
mfatimavinagre@gmail.com

Arquimedes de Siracusa (287 a.C.– 212 a.C.) é muito provavelmente um dos maiores matemáticos de sempre. Tendo realizado os seus estudos em Alexandria, no Egito, capital do mundo helénico, manteve correspondência com matemáticos alexandrinos como Eratóstenes e Aristarco.

A sua vasta obra matemática não padeceu de grande divulgação, uma vez que o próprio Arquimedes contribuiu, em parte, para isso. Os seus trabalhos dedicados à quadratura do círculo e da parábola, ao estudo da espiral, ao volume da esfera e de outros corpos redondos, ao estudo do centro de gravidade e aos problemas da flutuação eram escritos na forma de breves tratados, dirigidos aos matemáticos da escola de Alexandria. O estilo elíptico, a densidade de referências internas e a dificuldade das demonstrações não favoreceram a difusão, mesmo entre os matemáticos, apesar da grande originalidade dos seus trabalhos.

Foi apenas em 1906 que o filósofo e historiador dinamarquês Johan Ludvig Heiberg (1854-1928) estudou um palimpsesto[1], o qual continha além de textos matemáticos de Arquimedes subjacentes a um texto religioso do século X, um texto grego, "Sobre os corpos flutuantes". Mas foi entre os textos de Arquimedes que fez a sua maior descoberta, "O método sobre teoremas mecânicos", atualmente conhecido por "O Método", no qual Arquimedes apresentava a Eratóstenes, de Alexandria, o procedimento heurístico que empregava para obter as suas conclusões, que consistia na descrição de um "método mecânico" para investigar questões matemáticas, explicando várias das suas descobertas, entre as quais aquela que pode ser considerada uma das maiores descobertas matemáticas de todos os tempos, a fórmula para calcular o volume de uma esfera.

Existem diversas formas de demonstrar como se calcula o volume de uma esfera em função do seu raio. Contudo, neste artigo, pretende-se fazê-lo com base na lei da alavanca de Arquimedes de Siracusa (287 a.C.– 212 a.C.).

Arquimedes sabia que o volume de cone é igual a um terço do volume de um cilindro com a mesma altura e a mesma base e que o volume de um cilindro é dado pelo produto da área de uma das suas bases pela sua altura, e apresentou uma dedução teórica da **lei da alavanca** na sua obra "Sobre o Equilíbrio dos Planos" ou "Sobre os Centros de Gravidade das Figuras Planas", tendo-a engenhosamente aplicado na demonstração da fórmula de cálculo do volume de uma esfera.

Tenhamos, por base, o conceito de centro de gravidade de um corpo de Arquimedes:

> "Dois pesos w_1 e w_2 nas extremidades de uma barra de peso desprezável e distantes x_1 e x_2 respetivamente de um ponto de apoio estão em equilíbrio se $w_1 x_1 = w_2 x_2$."

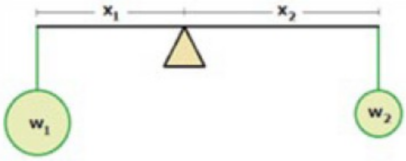

Figura 1: Ilustração do conceito de centro de gravidade de um corpo.

Segundo o geómetra, o centro de gravidade de qualquer corpo é um ponto, pertencente ao corpo ou localizado no espaço vazio, tal que, se for concebido que o corpo está suspenso por este ponto, o corpo assim sustentado permanece em repouso e preserva a sua posição original, sem se inclinar em nenhuma direção, qualquer que seja a sua orientação inicial em relação à Terra [1] e que todo o corpo, suspenso por qualquer ponto, assume um estado de equilíbrio quando o ponto de suspensão e o centro de gravidade do corpo estão ao longo de uma mesma linha vertical.[2]

Considere-se, ainda, o postulado 6 (no qual assenta a lei da alavanca) e os resultados teóricos tratados por Arquimedes na sua obra "Sobre o Equilíbrio dos Planos", relevantes para o presente artigo:

---

[1] Um pergaminho usado para nele se escrever várias vezes.

[2] Segundo Arquimedes afirma, no decorrer da demonstração da proposição 6, que comprovou o enunciado ao qual a autora se refere, num outro trabalho, hoje perdido.

Postulado 6: *Se grandezas se equilibram a certas distâncias*[3]*, então grandezas equivalentes*[4] *a estas grandezas equilibrar-se-ão, por sua vez, nas mesmas distâncias.*[5]*[1]*

Proposição 4: *Se duas grandezas iguais não possuem o mesmo centro de gravidade, o centro de gravidade da grandeza composta por estas [duas] grandezas estará no ponto médio do segmento de reta ligando os centros de gravidade das [duas] grandezas. [3]*

Proposições 6 e 7: *Se [duas] grandezas são comensuráveis [proposição 6] ou incomensuráveis [proposição 7], elas equilibrar-se-ão em distâncias inversamente proporcionais às grandezas. [3]*

Façamos, agora, algumas considerações acerca do que se entende por alavanca.

A alavanca consiste num corpo rígido, geralmente linear, capaz de girar ao redor de um eixo horizontal fixo em relação à Terra (o fulcro ou ponto de sustentação PS). O eixo de rotação é, em geral, ortogonal à alavanca, com os dois ficando usualmente no plano horizontal quando a alavanca está parada em relação à Terra.

É como se fosse uma balança, mas com a possibilidade de colocarmos pesos a distâncias diferentes do fulcro.

Chamamos braço da alavanca à distância horizontal, d, entre o ponto de apoio de um corpo sobre o travessão (ou haste da alavanca) e o plano vertical passando pelo fulcro. Para facilitar a linguagem, usaremos a expressão de distância entre o peso e o fulcro, mas em geral deve entender-se distância horizontal entre o ponto de atuação do peso na alavanca e o plano vertical passando pelo fulcro.

Quando falarmos de dois braços da alavanca, deve entender-se que estes estão em lados opostos do plano vertical que passa pelo fulcro.

Para se utilizar a lei da alavanca, pressupõe-se que a alavanca seja sensível, isto é, tenha liberdade de oscilação e de giro ao redor do fulcro; verifique as razões $\Delta P/PA$ (onde ΔP é a diferença entre os pesos que estão dos dois lados da alavanca e PA é o peso da alavanca) e d/h (onde h é a distância vertical entre o ponto de sustentação e centro de gravidade do travessão, e a distância d é o menor dos braços da alavanca).

Os dois lados do travessão da alavanca devem ter-lhes apostas diversas distâncias iguais em relação à vertical que passa pelo fulcro.

Mais, para que haja equilíbrio de dois corpos numa alavanca, apenas é relevante o seu peso e a sua distância ao fulcro.

A lei da alavanca determina, tal como Arquimedes expressou com a proposição 6 suprarreferida, que ao colocar-se um número N_A de corpos com o mesmo peso $P(P_A \equiv N_A P)$ atuando juntos num dos braços d_A de uma alavanca e um número N_B de corpos com o mesmo peso $P(P_B N_B P)$ atuando juntos no outro braço da alavanca d_B e soltando-a do repouso na horizontal, ela permanece parada na horizontal quando $d_B/d_A = P_A/P_B = N_A/N_B$.

Este princípio determina que ao se colocar N pesos $P_1, P_2, \ldots, P_N$ de um dos lados da alavanca suspensos, respetivamente, nas distâncias $d_1, d_2, \ldots, d_N$ do plano vertical que passa pelo fulcro e M pesos $P_{N+1}, P_{N+2}, \ldots, P_{N+M}$ do outro lado da alavanca suspensos, respetivamente, nas distâncias $d_{N+1}, d_{N+2}, \ldots, d_{N+M}$ do plano vertical que passa pelo fulcro e soltar-se do repouso a alavanca com esta na horizontal, a alavanca permanece em equilíbrio, verificando-se

$$\sum_{i=1}^{N} \frac{P_i}{P_0} \frac{d_i}{d_0} = \sum_{i=N+1}^{N+M} \frac{P_i}{P_0} \frac{d_i}{d_0}.$$

Para determinarmos algebricamente o centro de gravidade entre dois corpos, procedemos da seguinte forma:

Considere-se uma alavanca em equilíbrio estável[6] na horizontal sem a colocação de outros corpos. Suponha-se que essa alavanca tem uma haste homogénea tal que quando o fulcro fica ao longo de um plano vertical que divide a haste em duas partes com o mesmo comprimento, a alavanca permanece em equilíbrio ao ser solta do repouso, parada na horizontal. O centro de gravidade do travessão está verticalmente abaixo do fulcro, ou do ponto de sustentação.

Sabemos que este equilíbrio mantém-se caso sejam colocados dois corpos A e B suspensos em lados opostos da alavanca, desde que se verifique $d_B/d_A = P_A/P_B$, onde d_A e d_B são as distâncias horizontais entre os pontos de suspensão de A e de B até ao plano vertical passando pelo fulcro, sendo P_A e P_B os pesos de A e de B, respetivamente. Isto significa que o centro de gravidade destes dois corpos também se encontra no plano vertical que passa pelo fulcro.

Tomemos um eixo horizontal, x, ao longo da haste da alavanca de comprimento L e com origem (escolhida de forma arbitrária) em $x = 0$.

Suponha-se que as extremidades da haste da alavanca estão localizadas em X_E e $X_D = X_E + L$, que colocamos os corpos A e B suspensos nas posições X_A e X_B do eixo X, respetivamente, e que a alavanca permanece em equilíbrio ao ser solta do repouso com os corpos A e B

atuando nestas posições, conforme se pode observar na figura seguinte.

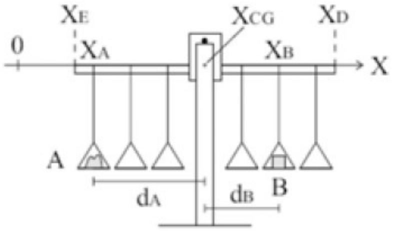

Figura 2: Encontrando a expressão algébrica para o centro de gravidade.

O centro de gravidade deste sistema tem de estar sobre o plano vertical que passa pelo fulcro quando $d_B/d_A = P_A/P_B$, uma vez que a alavanca permanece em equilíbrio.

Represente-se a localização do centro de gravidade dos corpos A e B por X_{CG}.

Então, $d_A = X_{CG}X_A$ e $d_B = X_BX_{CG}$.

Pela lei da alavanca, a posição X_{CG} do centro de gravidade deste sistema de dois corpos ao longo do eixo X é dada por:

$$\frac{X_B - X_{CG}}{X_{CG} - X_A} = \frac{P_A}{P_B},$$

ou seja,

$$X_{CG} = \frac{P_A}{P_T}X_A + \frac{P_B}{P_T}X_B,$$

onde $P_T = P_A + P_B$ é o peso total dos dois corpos.

Facilmente se observa que quando $P_A = P_B$, X_{CG} é o ponto médio entre X_A e X_B.

Feitas as considerações anteriores, estamos em condições de deduzir, utilizando a lei da alavanca de Arquimedes, a fórmula para calcular o volume de uma esfera de raio r.

Pretende-se colocar os sólidos a seguir descritos suspensos nas extremidades de uma alavanca, como mostra a figura 3.

Para tal, comecemos por considerar os três sólidos de revolução seguintes: uma esfera gerada pela rotação do seu círculo maior de raio r em torno do seu diâmetro, um cone de raio e altura iguais ao diâmetro da esfera e um cilindro de raio e altura iguais ao diâmetro da esfera cujo eixo coincide com o diâmetro da esfera, dispostos conforme a figura 4.

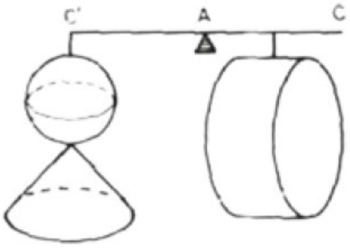

Figura 3: Alavanca com os sólidos suspensos e com fulcro em *A*.

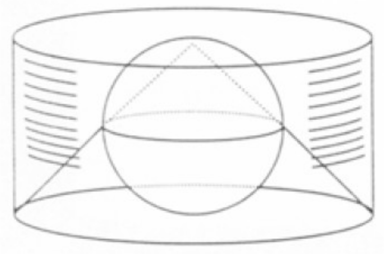

Figura 4: Posição dos sólidos.

Deixemos a cargo do leitor a justificação da importância da posição dos sólidos.

Tomemos o ponto de suspensão dos sólidos no vértice do cone. Então, os centros de gravidade dos três sólidos encontram-se numa mesma linha reta vertical.

Considere-se uma alavanca HC de braços HA e AC em equilíbrio sem os sólidos suspensos, cujo comprimento dos braços é $2r$ e com fulcro no ponto médio de HC

---

[3] Entenda-se "grandezas a certas distâncias" como "grandezas cujos centros de gravidade estão às mesmas distâncias do fulcro."

[4] Entenda-se "grandezas equivalentes" como "grandezas com o mesmo peso."

[5] Este postulado permitia a Arquimedes substituir um corpo *A* suspenso numa alavanca através do seu centro de gravidade localizado à distância horizontal *d* do plano vertical passando pelo fulcro, por outro corpo *B* conveniente que tivesse o mesmo peso que *A*, sem afetar o equilíbrio da alavanca, desde que o centro de gravidade deste corpo *B* também estivesse pendurado na alavanca à mesma distância horizontal *d* do plano vertical passando pelo fulcro.

[6] Equilíbrio estável é quando qualquer perturbação na posição do corpo fizer com que o centro de gravidade suba em relação à sua posição anterior.

(para que o plano vertical que passa pelo fulcro divida ao meio o travessão da alavanca). Seja A o fulcro.

Comece-se por considerar que o fulcro se encontra no vértice do cone e, por comodidade, faça-se um corte transversal nos três sólidos passando pelo fulcro e suponha-se que os sólidos estão suspensos neste ponto de apoio[7]. Obtém-se a figura 5.

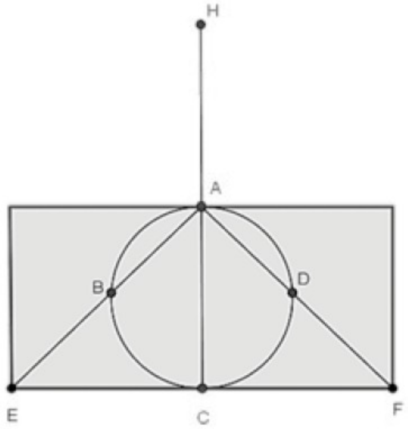

Figura 5: Secção transversal dos sólidos.

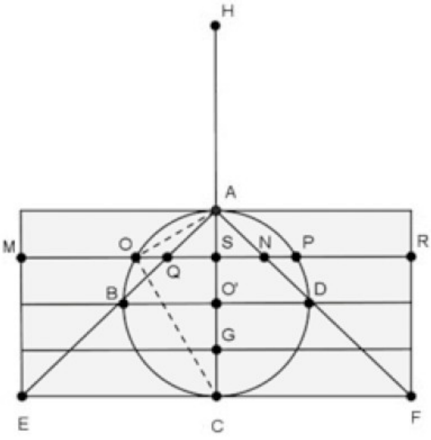

Figura 6: Designação de pontos da secção.

De seguida, através deste corte transversal, tentemos encontrar pontos de suspensão para os sólidos em ambos os braços da alavanca, que não perturbem o equilíbrio da mesma.

Para tal, façam-se três cortes nos sólidos segundo planos paralelos à base do cilindro, tais que o cilindro fique dividido em quatro partes iguais[8]. Tais planos cortam AC em quatro partes iguais.

Denominem-se por S, O' e G os pontos de interseção dos planos com o eixo de rotação do cilindro.

Centremo-nos, agora, na secção transversal dos sólidos.

Designem-se pelas letras O e P os pontos de interseção do círculo com a reta MR e Q e N os pontos de interseção da reta MR com o triângulo AEH, assinalados na figura 6 e considere-se o triângulo AOC.

Como

$$\frac{AS}{AC} = \frac{AS}{AC} \cdot \frac{AC}{AC} =$$

$$= \frac{AS \cdot AC}{AC^2} = \frac{AS \cdot (AS + SC)}{AC^2} = \frac{AS^2 + AS \cdot SC}{AC^2}$$

e $AS \cdot SC = OS \cdot SP$, dado que AC e OP são cordas da circunferência que se intersetam em S, podemos afirmar que

$$\frac{AS}{AC} = \frac{AS^2 + OS \cdot SP}{AC^2} = \frac{AS^2 + OS^2}{AC^2}$$

e, portanto,

$$AS \cdot AC^2 = AC \cdot (AS^2 + OS^2). \qquad (1)$$

Como, por construção, se sabe que $AH = AC = MS$, atendendo a que o triângulo AQS é isósceles (porque os ângulos $\angle QAS$ e $\angle AQS$ têm as mesmas amplitudes, são ambos de 45°), tem-se que $QS = AS$.

Consequentemente, a igualdade (1) é equivalente a

$$AS \cdot MS^2 = AH \cdot (QS^2 + OS^2),$$

ou seja,

$$\frac{AS}{AH} = \frac{QS^2 + OS^2}{MS^2} \Leftrightarrow \frac{AS}{AH} = \frac{\pi \cdot (QS^2 + OS^2)}{\pi \cdot MS^2}$$

$$\Leftrightarrow \frac{AS}{AH} = \frac{\pi \cdot QS^2 + \pi \cdot OS^2}{\pi \cdot MS^2}.$$

Assim, como $\pi \cdot QS^2, \pi \cdot OS^2$ e $\pi \cdot MS^2$ representam as áreas dos círculos de raios QS, OS e MS, respetivamente, e como o equilíbrio de dois corpos se mantém, se as distâncias horizontais entre os pontos de suspensão destes forem inversamente proporcionais ao seu peso, podemos concluir que o equilíbrio da alavanca não é perturbado, se colocarmos os círculos de raios QS e OS suspensos na extremidade H da alavanca e o círculo de raio MS suspenso no ponto S pertencente ao outro braço da alavanca, ou seja, a uma distância horizontal do fulcro igual a $r/2$.

Tendo em conta que todos os resultados deduzidos através da secção transversal mantêm as mesmas propriedades se considerarmos S a variar de A até O', tomemos o cilindro como união infinita dos círculos de raio igual a MS, o cone como união infinita dos círculos de raio igual a QS e a semiesfera como união infinita dos círculos de raio igual a OS, com S a variar de A até O'.[9]

Podemos concluir que o equilíbrio da alavanca não é perturbado se remontássemos os sólidos e os colocássemos suspensos no travessão da alavanca da seguinte forma: a semiesfera e o cone na extremidade H da alavanca e o cilindro no ponto de suspensão S situado no outro braço da alavanca.

Designem-se por V_S, V_C e V_i os volumes da semiesfera, do cone e do cilindro, respetivamente.

Como a alavanca se mantém equilibrada com a suspensão dos sólidos, verifica-se a relação $\frac{V_c+V_s}{V_i} = \frac{AS}{AH}$ ou seja,

$$\frac{V_c+V_s}{V_i} = \frac{r}{2}2r \Leftrightarrow \frac{V_c+V_s}{V_i} = \frac{1}{4}. \qquad (2)$$

Atendendo a que $V_c = 1/3 \cdot \pi r^2 \cdot r$ e $V_i = (2r)^2 \pi r$, substituindo em (2), vem:

$$\frac{\frac{1}{3} \cdot \pi r^2 \cdot r + V_s}{(2r)^2 \cdot \pi r} = \frac{1}{4} \Leftrightarrow \frac{1}{3} \cdot \pi r^3 + V_s = \frac{(2r)^2 \cdot \pi r}{4}$$

$$\Leftrightarrow V_s = \pi r^3 - \frac{1}{3} \cdot \pi r^3 \Leftrightarrow V_s = \frac{2}{3} \cdot \pi r^3.$$

Como o volume da esfera inicial é o dobro do volume desta semiesfera, podemos concluir que o volume da esfera inicial é $4/3 \cdot \pi r^3$.

REFERÊNCIAS

[1] A. K. T. Assis. *Archimedes, the Center of Gravity, and the First Law of Mechanics*. Apeiron, Montreal, 2008.

[2] G. Ávila. "Arquimedes, a esfera e o cilindro", *Revista do Professor de Matemática* 10, 2010.

[3] Archimedes. *The Works of Archimedes*. Dover, New York, 2002. (Translated and edited in modern notation by T. L. Heath.)

[4] A.Silva. *Arquimedes e o volume da esfera*. Belo Horizonte, 2005.

SOBRE A AUTORA

Fátima Vinagre é professora de Matemática na Escola Secundária da Azambuja, licenciada em Matemática - Ramo Educacional, pela FCT-UNL, e mestre em Matemática – Grupos Cristalográficos Tridimensionais, pela Universidade de Aveiro. Exerceu funções de professora auxiliar de Álgebra Linear e Geometria Analítica e Análise Matemática I na FCT-UNL, foi responsável pela coordenação de Cursos de Especialização Tecnológica, presidente de um Conselho Técnico-Pedagógico, orientadora de estágios – profissionalização em serviço, avaliadora de pessoal docente e não docente, subdiretora e vice-presidente da Escola Secundária de Montejunto, entre outras funções.

---

[7] Desta forma temos a garantia de que o sistema continua em equilíbrio.

[8] O cilindro é dividido em partes iguais, para que tenham pesos iguais.

[9] Tem em conta os procedimentos adotados por Arquimedes. [2].

Gazeta de Matemática
Gazeta de Matemática
Cirurgia Plástica do Ponto de Vista Matemático
Gueorgui Smirnov e Vera Sá
A Matemática do Sistema Cardiovascular

Visite o site da
Gazeta de Matemática.

www.gazeta.spm.pt

Para aceder à área reservada a assinantes, solicite o seu código de subscrição através do e-mail **gazeta@spm.pt**

data 05 05 2020

grande coisa.

B7) 527.pdf → Arquimedes e o Volume da Esfera

B8) 528.pdf → O conjunto de todos os conjuntos não existe.

B9) 529.pdf → Números de Bernoulli

B10) 530.pdf → Os "Modelos de Olivier".

B11) 531.pdf → Problemas antigos com números perfeitos.

Important:

B12) 001_Traité_de_géométrie_descriptive... Javary... pdf

obs: 337 pages.

B13) 002_Traité_de_géométrie_descriptive_Javary...pdf

obs: 816 pages.

B14) Traité de géométrie supérieure... Chasles...pdf

obs.: 699 pages.

C) Books_79_2020/
Bernhard Riemann 1826-1866.../

7)*+,)))-) '%
/ M>! > M M >! 6 8 >6 >T UD 8 A>*! >76 'G:: 1 ' 6!
-7 ""1 6

7 ")*+,)))-) '%
/ M>! > M M >! 6 8 >6 >T UD 8 A>*! >76 'G:: 1 6!
-7 : ' 6

7 %)*+,)))-) '%
/ M>! > M M > 6M >>>T U > >76 'G$' :8 6!

-

)*+,)))-) '%
/ M>! > M M >! 6 8 >6 >T U >/ M 6 >76 'G % ! 6!
-7 "' 6

)*+,)))-) '%
/ M>! > M M >! 6 8 >6 >T U >/ M 6 >76 'G$% ' %% 6!
-7 "1 6

)*+,)))-) '%
/ M>! > M M >>> 6 >T U5 A> < V * >76 'G $ % 6!

)*+,)))-) '% ? ! + : ' :">
/ 6 $$:2 ? G W ? 6!

)*+,)))-) '% ? ! + : ' :"> /
6 $$:2 ? G W ? !H8

& '

)*+,)))-) ' & !
2 ? @ >F 8 6!
(' - /

This indicates that when a tensor is invertible, there is a one-to-one mapping of vectors **a** and **b**. On the other hand, if a tensor **T** does not have an inverse, then, for a given **b**, there are in general more than one **a** that transform into **b**. This fact is illustrated in the following example.

Example 2.14.2

Consider the tensor $\mathbf{T} = \mathbf{cd}$ (the dyadic product of **c** and **d**).

(a) Obtain the determinant of **T**.

(b) Show that if $\mathbf{Ta} = \mathbf{b}$, then $\mathbf{T}(\mathbf{a} + \mathbf{h}) = \mathbf{b}$, where **h** is any vector perpendicular to the vector **d**.

Solution

(a) $[\mathbf{T}] = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} [\, d_1 \quad d_2 \quad d_3 \,] = \begin{bmatrix} c_1 d_1 & c_1 d_2 & c_1 d_3 \\ c_2 d_1 & c_2 d_2 & c_2 d_3 \\ c_3 d_1 & c_3 d_2 & c_3 d_3 \end{bmatrix}$ and $\det [\mathbf{T}] = c_1 c_2 c_3 d_1 d_2 d_3 \begin{vmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{vmatrix} = 0.$

That is, **T** is a singular tensor, for which an inverse does not exist.

(b) $\mathbf{T}(\mathbf{a}+\mathbf{h}) = (\mathbf{cd})(\mathbf{a}+\mathbf{h}) = \mathbf{c}(\mathbf{d} \cdot \mathbf{a}) + \mathbf{c}(\mathbf{d} \cdot \mathbf{h})$. Now $\mathbf{d} \cdot \mathbf{h} = 0$ (**h** is perpendicular to **d**); therefore,

$$\mathbf{T}(\mathbf{a} + \mathbf{h}) = \mathbf{c}(\mathbf{d} \cdot \mathbf{a}) = (\mathbf{cd})\mathbf{a} = \mathbf{Ta} = \mathbf{b}.$$

That is, all vectors $\mathbf{a} + \mathbf{h}$ transform into the vector **b**, and it is not a one-to-one transformation.

2.15 ORTHOGONAL TENSORS

An *orthogonal tensor* is a linear transformation under which the transformed vectors preserve their lengths and angles. Let **Q** denote an orthogonal tensor; then by definition, $|\mathbf{Qa}| = |\mathbf{a}|$, $|\mathbf{Qb}| = |\mathbf{b}|$, and $\cos(\mathbf{a},\mathbf{b}) = \cos(\mathbf{Qa}, \mathbf{Qb})$. Therefore,

$$\mathbf{Qa} \cdot \mathbf{Qb} = \mathbf{a} \cdot \mathbf{b} \tag{2.15.1}$$

for any vectors **a** and **b**.

Since by the definition of transpose, Eq. (2.11.1), $(\mathbf{Qa}) \cdot (\mathbf{Qb}) = \mathbf{b} \cdot \mathbf{Q}^{\mathrm{T}}(\mathbf{Qa})$, thus

$$\mathbf{b} \cdot \mathbf{a} = \mathbf{b} \cdot (\mathbf{Q}^{\mathrm{T}}\mathbf{Q})\mathbf{a} \quad \text{or} \quad \mathbf{b} \cdot \mathbf{Ia} = \mathbf{b} \cdot \mathbf{Q}^{\mathrm{T}}\mathbf{Qa}.$$

Since **a** and **b** are arbitrary, it follows that

$$\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}. \tag{2.15.2}$$

This means that for an orthogonal tensor, the inverse is simply the transpose,

$$\mathbf{Q}^{-1} = \mathbf{Q}^{\mathrm{T}}. \tag{2.15.3}$$

Thus [see Eq. (2.14.8)],

$$\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{Q}\mathbf{Q}^{\mathrm{T}} = \mathbf{I}. \tag{2.15.4}$$

In matrix notation, Eq. (2.15.4) takes the form:

$$[\mathbf{Q}]^{\mathrm{T}}[\mathbf{Q}] = [\mathbf{Q}][\mathbf{Q}]^{\mathrm{T}} = [\mathbf{I}], \tag{2.15.5}$$

and in subscript notation, we have

$$Q_{mi}Q_{mj} = Q_{im}Q_{jm} = \delta_{ij}. \tag{2.15.6}$$

Example 2.15.1

The tensor given in Example 2.7.2, being a reflection, is obviously an orthogonal tensor. Verify that $[\mathbf{T}][\mathbf{T}]^{\mathrm{T}} = [\mathbf{I}]$ for the $[\mathbf{T}]$ in that example. Also, find the determinant of $[\mathbf{T}]$.

Solution

Evaluating the matrix product:

$$[\mathbf{T}][\mathbf{T}]^{\mathrm{T}} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The determinant of $\mathbf{T}$ is

$$|\mathbf{T}| = \begin{vmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = -1.$$

Example 2.15.2

The tensor given in Example 2.7.3, being a rigid body rotation, is obviously an orthogonal tensor. Verify that $[\mathbf{R}][\mathbf{R}]^{\mathrm{T}} = [\mathbf{I}]$ for the $[\mathbf{R}]$ in that example. Also find the determinant of $[\mathbf{R}]$.

Solution

$$[\mathbf{R}][\mathbf{R}]^{\mathrm{T}} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$\det[\mathbf{R}] = |\mathbf{R}| = \begin{vmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{vmatrix} = 1.$$

The determinant of the matrix of any orthogonal tensor $\mathbf{Q}$ is easily shown to be equal to either $+1$ or -1. In fact, since

$$[\mathbf{Q}][\mathbf{Q}]^{\mathrm{T}} = [\mathbf{I}],$$

therefore,

$$|[\mathbf{Q}][\mathbf{Q}]^{\mathrm{T}}| = |\mathbf{Q}||\mathbf{Q}^{\mathrm{T}}| = |\mathbf{I}|.$$

Now $|\mathbf{Q}| = |\mathbf{Q}^{\mathrm{T}}|$ and $|\mathbf{I}| = 1$, therefore, $|\mathbf{Q}|^2 = 1$, thus

$$|\mathbf{Q}| = \pm 1. \tag{2.15.7}$$

From the previous examples, we can see that for a rotation tensor the determinant is +1, whereas for a reflection tensor, it is −1.

2.16 TRANSFORMATION MATRIX BETWEEN TWO RECTANGULAR CARTESIAN COORDINATE SYSTEMS

Suppose that $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ and $\{\mathbf{e}_1', \mathbf{e}_2', \mathbf{e}_3'\}$ are unit vectors corresponding to two rectangular Cartesian coordinate systems (see Figure 2.16-1). It is clear that $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ can be made to coincide with $\{\mathbf{e}_1', \mathbf{e}_2', \mathbf{e}_3'\}$ through either a rigid body rotation (if both bases are same-handed) or a rotation followed by a reflection (if different-handed). That is, $\{\mathbf{e}_i\}$ and $\{\mathbf{e}_i'\}$ are related by an orthogonal tensor $\mathbf{Q}$ through the equations below.

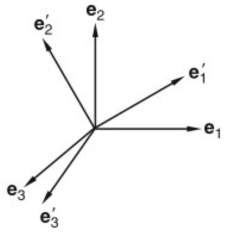

FIGURE 2.16-1

$$\mathbf{e}_i' = \mathbf{Q}\mathbf{e}_i = Q_{mi}\mathbf{e}_m, \tag{2.16.1}$$

that is,

$$\begin{aligned} \mathbf{e}_1' &= Q_{11}\mathbf{e}_1 + Q_{21}\mathbf{e}_2 + Q_{31}\mathbf{e}_3, \\ \mathbf{e}_2' &= Q_{12}\mathbf{e}_1 + Q_{22}\mathbf{e}_2 + Q_{32}\mathbf{e}_3, \\ \mathbf{e}_3' &= Q_{13}\mathbf{e}_1 + Q_{23}\mathbf{e}_2 + Q_{33}\mathbf{e}_3, \end{aligned} \tag{2.16.2}$$

where

$$Q_{im}Q_{jm} = Q_{mi}Q_{mj} = \delta_{ij}, \tag{2.16.3}$$

or

$$\mathbf{Q}\mathbf{Q}^{\mathrm{T}} = \mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}. \tag{2.16.4}$$

We note that

$$\begin{aligned} Q_{11} &= \mathbf{e}_1 \cdot \mathbf{Q}\mathbf{e}_1 = \mathbf{e}_1 \cdot \mathbf{e}_1' = \text{cosine of the angle between } \mathbf{e}_1 \text{ and } \mathbf{e}_1', \\ Q_{12} &= \mathbf{e}_1 \cdot \mathbf{Q}\mathbf{e}_2 = \mathbf{e}_1 \cdot \mathbf{e}_2' = \text{cosine of the angle between } \mathbf{e}_1 \text{ and } \mathbf{e}_2', \text{etc.} \end{aligned}$$

That is, in general, Q_{ij} = cosine of the angle between $\mathbf{e}_i$ and $\mathbf{e}_j'$, which may be written:

$$Q_{ij} = \cos(\mathbf{e}_i, \mathbf{e}_j'). \tag{2.16.5}$$

The matrix of these direction cosines, i.e., the matrix

$$[\mathbf{Q}] = \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} \\ Q_{21} & Q_{22} & Q_{23} \\ Q_{31} & Q_{32} & Q_{33} \end{bmatrix}, \tag{2.16.6}$$

is called the *transformation matrix* between $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ and $\{\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3\}$. Using this matrix, we shall obtain in the following sections the relationship between the two sets of components, with respect to these two sets of base vectors, of a vector and a tensor.

Example 2.16.1

Let $\{\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3\}$ be obtained by rotating the basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ about the $\mathbf{e}_3$ axis through 30°, as shown in Figure 2.16-2. In this figure, $\mathbf{e}_3$ and $\mathbf{e}'_3$ coincide.

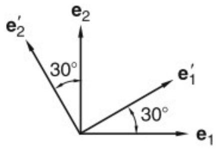

FIGURE 2.16-2

Solution

We can obtain the transformation matrix in two ways:

1. Using Eq. (2.16.5), we have

$$Q_{11} = \cos(\mathbf{e}_1, \mathbf{e}'_1) = \cos 30^\circ = \sqrt{3}/2, \quad Q_{12} = \cos(\mathbf{e}_1, \mathbf{e}'_2) = \cos 120^\circ = -1/2, \quad Q_{13} = \cos(\mathbf{e}_1, \mathbf{e}'_3) = \cos 90^\circ = 0,$$
$$Q_{21} = \cos(\mathbf{e}_2, \mathbf{e}'_1) = \cos 60^\circ = 1/2, \quad Q_{22} = \cos(\mathbf{e}_2, \mathbf{e}'_2) = \cos 30^\circ = \sqrt{3}/2, \quad Q_{23} = \cos(\mathbf{e}_2, \mathbf{e}'_3) = \cos 90^\circ = 0,$$
$$Q_{31} = \cos(\mathbf{e}_3, \mathbf{e}'_1) = \cos 90^\circ = 0, \quad Q_{32} = \cos(\mathbf{e}_3, \mathbf{e}'_2) = \cos 90^\circ = 0, \quad Q_{23} = \cos(\mathbf{e}_3, \mathbf{e}'_3) = \cos 0^\circ = 1.$$

2. It is easier to simply look at Figure 2.16-2 and decompose each of the $\mathbf{e}'_i$ into its components in the $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ directions, i.e.,

$$\mathbf{e}'_1 = \cos 30^\circ \mathbf{e}_1 + \sin 30^\circ \mathbf{e}_2 = \frac{\sqrt{3}}{2}\mathbf{e}_1 + \frac{1}{2}\mathbf{e}_2,$$
$$\mathbf{e}'_2 = -\sin 30^\circ \mathbf{e}_1 + \cos 30^\circ \mathbf{e}_3 = -\frac{1}{2}\mathbf{e}_1 + \frac{\sqrt{3}}{2}\mathbf{e}_2,$$
$$\mathbf{e}'_3 = \mathbf{e}_3.$$

Thus, by Eq. (2.16.2), we have

$$[\mathbf{Q}] = \begin{bmatrix} \sqrt{3}/2 & -1/2 & 0 \\ 1/2 & \sqrt{3}/2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

2.17 TRANSFORMATION LAW FOR CARTESIAN COMPONENTS OF A VECTOR

Consider any vector **a**. The Cartesian components of the vector **a** with respect to $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ are:

$$a_i = \mathbf{a} \cdot \mathbf{e}_i, \tag{2.17.1}$$

and its components with respect to $\{\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3\}$ are:

$$a'_i = \mathbf{a} \cdot \mathbf{e}'_i. \tag{2.17.2}$$

Now $\mathbf{e}'_i = Q_{mi}\mathbf{e}_m$ [see Eq. (2.16.1)]; therefore,

$$a'_i = \mathbf{a} \cdot Q_{mi}\mathbf{e}_m = Q_{mi}(\mathbf{a} \cdot \mathbf{e}_m), \tag{2.17.3}$$

that is,

$$a'_i = Q_{mi}a_m. \tag{2.17.4}$$

In matrix notation, Eq. (2.17.4) is

$$\begin{bmatrix} a'_1 \\ a'_2 \\ a'_3 \end{bmatrix} = \begin{bmatrix} Q_{11} & Q_{21} & Q_{31} \\ Q_{12} & Q_{22} & Q_{32} \\ Q_{13} & Q_{23} & Q_{33} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}, \tag{2.17.5}$$

or

$$[\mathbf{a}]' = [\mathbf{Q}]^{\mathrm{T}}[\mathbf{a}]. \tag{2.17.6}$$

Equation (2.17.4), or Eq. (2.17.5), or Eq. (2.17.6) is the transformation law relating components of the *same* vector with respect to different rectangular Cartesian unit bases. It is very important to note that in Eq. (2.17.6), $[\mathbf{a}]'$ denotes the matrix of the vector **a** with respect to the primed basis $\{\mathbf{e}'_i\}$, and $[\mathbf{a}]$ denotes the same vector with respect to the unprimed basis $\{\mathbf{e}_i\}$. Eq. (2.17.6) is not the same as $\mathbf{a}' = \mathbf{Q}^{\mathrm{T}}\mathbf{a}$. The distinction is that $[\mathbf{a}]'$ and $[\mathbf{a}]$ are matrices of the same vector, whereas **a** and $\mathbf{a}'$ are two different vectors—$\mathbf{a}'$ being the transformed vector of **a** (through the transformation $\mathbf{a}' = \mathbf{Q}^{\mathrm{T}}\mathbf{a}$).

If we premultiply Eq. (2.17.6) with $[\mathbf{Q}]$, we get

$$[\mathbf{a}] = [\mathbf{Q}][\mathbf{a}]'. \tag{2.17.7}$$

The indicial notation for this equation is:

$$a_i = Q_{im}a'_m. \tag{2.17.8}$$

Example 2.17.1

Given that the components of a vector **a** with respect to $\{\mathbf{e}_i\}$ are given to be [2,0,0]. That is, $\mathbf{a} = 2\mathbf{e}_1$, find its components with respect to $\{\mathbf{e}'_i\}$, where the $\{\mathbf{e}'_i\}$ axes are obtained by a 90° counter-clockwise rotation of the $\{\mathbf{e}_i\}$ axis about its $\mathbf{e}_3$ axis.

Solution

The answer to the question is obvious from Figure 2.17-1, that is,

$$\mathbf{a} = 2\mathbf{e}_1 = -2\mathbf{e}'_2.$$

To show that we can get the same answer from Eq. (2.17.6), we first obtain the transformation matrix of **Q**. Since $\mathbf{e}_1' = \mathbf{e}_2, \mathbf{e}_2' = -\mathbf{e}_1$ and $\mathbf{e}_3' = \mathbf{e}_3$, we have

$$[\mathbf{Q}] = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Thus,

$$[\mathbf{a}]' = [\mathbf{Q}]^{\mathrm{T}}[\mathbf{a}] = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ -2 \\ 0 \end{bmatrix},$$

that is,

$$\mathbf{a} = -2\mathbf{e}_2'.$$

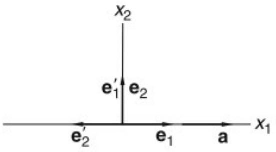

FIGURE 2.17-1

2.18 TRANSFORMATION LAW FOR CARTESIAN COMPONENTS OF A TENSOR

Consider any tensor **T**. The components of **T** with respect to the basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ are:

$$T_{ij} = \mathbf{e}_i \cdot \mathbf{T}\mathbf{e}_j. \tag{2.18.1}$$

Its components with respect to $\{\mathbf{e}_1', \mathbf{e}_2', \mathbf{e}_3'\}$ are:

$$T_{ij}' = \mathbf{e}_i' \cdot \mathbf{T}\mathbf{e}_j'. \tag{2.18.2}$$

With $\mathbf{e}_i' = Q_{mi}\mathbf{e}_m$, we have

$$T_{ij}' = Q_{mi}\mathbf{e}_m \cdot \mathbf{T}Q_{nj}\mathbf{e}_n = Q_{mi}Q_{nj}\mathbf{e}_m \cdot \mathbf{T}\mathbf{e}_n,$$

that is,

$$T_{ij}' = Q_{mi}Q_{nj}T_{mn}. \tag{2.18.3}$$

In matrix notation, the preceding equation reads:

$$\begin{bmatrix} T_{11}' & T_{12}' & T_{13}' \\ T_{21}' & T_{22}' & T_{23}' \\ T_{31}' & T_{32}' & T_{33}' \end{bmatrix} = \begin{bmatrix} Q_{11} & Q_{21} & Q_{31} \\ Q_{12} & Q_{22} & Q_{32} \\ Q_{13} & Q_{23} & Q_{33} \end{bmatrix} \begin{bmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{bmatrix} \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} \\ Q_{21} & Q_{22} & Q_{23} \\ Q_{31} & Q_{32} & Q_{33} \end{bmatrix}, \tag{2.18.4}$$

or

$$[\mathbf{T}]' = [\mathbf{Q}]^{\mathrm{T}}\,[\mathbf{T}]\,[\mathbf{Q}]. \tag{2.18.5}$$

We can also express the unprimed components in terms of the primed components. Indeed, if we premultiply the preceding equation with $[\mathbf{Q}]$ and post-multiply it with $[\mathbf{Q}]^{\mathrm{T}}$, we obtain, since

$$[\mathbf{Q}]\,[\mathbf{Q}]^{\mathrm{T}} = [\mathbf{Q}]^{\mathrm{T}}[\mathbf{Q}] = [\mathbf{I}], \tag{2.18.6}$$

$$[\mathbf{T}] = [\mathbf{Q}]\,[\mathbf{T}]'\,[\mathbf{Q}]^{\mathrm{T}}. \tag{2.18.7}$$

In indicial notation, Eq. (2.18.7) reads

$$T_{ij} = Q_{im}Q_{jn}T'_{mn}. \tag{2.18.8}$$

Equations (2.18.5) [or Eq. (2.18.3)] and Eq. (2.18.7) [or Eq. (2.18.8)] are the transformation laws relating components of the *same* tensor with respect to different Cartesian unit bases. Again, it is important to note that in Eqs. (2.18.5) and (2.18.7), $[\mathbf{T}]$ and $[\mathbf{T}]'$ are different matrices of the *same* tensor $\mathbf{T}$. We note that the equation $[\mathbf{T}]' = [\mathbf{Q}]^{\mathrm{T}}[\mathbf{T}][\mathbf{Q}]$ differs from $\mathbf{T}' = \mathbf{Q}^{\mathrm{T}}\mathbf{T}\mathbf{Q}$ in that the former relates the components of the same tensor $\mathbf{T}$ whereas the latter relates the two different tensors $\mathbf{T}$ and $\mathbf{T}'$.

Example 2.18.1

Given that with respect to the basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, the matrix of a tensor $\mathbf{T}$ is given by

$$[\mathrm{T}] = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Find $[\mathbf{T}]'$, that is, find the matrix of $\mathbf{T}$ with respect to the $\mathbf{e}'_i$ basis, where $\{\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3\}$ is obtained by rotating $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ about its $\mathbf{e}_3$-axis through 90° (see Figure 2.17-1).

Solution

Since $\mathbf{e}'_1 = \mathbf{e}_2, \mathbf{e}'_2 = -\mathbf{e}_1$ and $\mathbf{e}'_3 = \mathbf{e}_3$, by Eq. (2.7.1) we have

$$[\mathbf{Q}] = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Thus, Eq. (2.18.5) gives

$$[\mathbf{T}]' = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

that is,

$$T'_{11} = 2, \quad T'_{12} = -1, \quad T'_{13} = 0, \quad T'_{22} = 0, \quad T'_{23} = 0, \quad T'_{33} = 1.$$

Example 2.18.2

Given a tensor $\mathbf{T}$ and its components T_{ij} and T'_{ij} with respect to two sets of bases $\{\mathbf{e}_i\}$ and $\{\mathbf{e}'_i\}$. Show that T_{ii} is invariant with respect to these bases, i.e., $T_{ii} = T'_{ii}$.

Solution

The primed components are related to the unprimed components by Eq. (2.18.3):

$$T'_{ij} = Q_{mi}Q_{nj}T_{mn},$$

thus,

$$T'_{ii} = Q_{mi}Q_{ni}T_{mn}.$$

But $Q_{mi}Q_{ni} = \delta_{mn}$ [Eq. (2.15.6)], therefore,

$$T'_{ii} = \delta_{mn}T_{mn} = T_{mm} = T_{ii},$$

that is,

$$T_{11} + T_{22} + T_{33} = T'_{11} + T'_{22} + T'_{33}.$$

We see from Example 2.18.1 that we can calculate all nine components of a tensor **T** with respect to $\{\mathbf{e}'_i\}$ from the matrix $[\mathrm{T}]_{\{\mathbf{e}_i\}}$ by using Eq. (2.18.5). However, there are often times when we need only a few components. Then it is more convenient to use Eq. (2.18.1). In matrix form, this equation is written:

$$T'_{ij} = [\mathbf{e}'_i]^{\mathrm{T}}[\mathbf{T}]\,[\mathbf{e}'_j], \tag{2.18.9}$$

where $[\mathbf{e}'_i]^{\mathrm{T}}$ denote the row matrix whose elements are the components of $\mathbf{e}'_i$ with respect to the basis $\{\mathbf{e}_i\}$.

Example 2.18.3

Obtain T'_{12} for the tensor **T** and the bases $\{\mathbf{e}_i\}$ and $\{\mathbf{e}'_i\}$ given in Example 2.18.1 by using Eq. (2.18.1).

Solution

Since $\mathbf{e}'_1 = \mathbf{e}_2$ and $\mathbf{e}'_2 = -\mathbf{e}_1$, therefore,

$$T'_{12} = \mathbf{e}'_1 \cdot \mathbf{T}\mathbf{e}'_2 = \mathbf{e}_2 \cdot \mathbf{T}(-\mathbf{e}_1) = -T_{21} = -1.$$

Alternatively, using Eq. (2.18.9), we have

$$T'_{12} = [\mathbf{e}'_1]^{\mathrm{T}}[\mathbf{T}][\mathbf{e}'_2] = [0 \quad 1 \quad 0]\begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix} = [0 \quad 1 \quad 0]\begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix} = -1.$$

2.19 DEFINING TENSOR BY TRANSFORMATION LAWS

Equation (2.17.4) or (2.18.3) states that when the components of a vector or a tensor with respect to $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ are known, then its components with respect to any $\{\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3\}$ are uniquely determined from them. In other words, the components a_i or T_{ij} with respect to one set of $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ completely characterize a vector or a tensor. Thus, it is perfectly meaningful to use a statement such as "*consider a tensor* T_{ij}," meaning consider the tensor **T** whose components with respect to some set of $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ are T_{ij}. In fact, an alternative way of defining a tensor is through the use of transformation laws relating components of a tensor with respect to different bases. Confining ourselves to only rectangular Cartesian coordinate systems and using unit vectors along positive coordinate directions as base vectors, we now define Cartesian components of tensors of different orders in terms of their transformation laws in the following, where the primed quantities are

2.60 Do the previous problem for the following matrix: $[\mathbf{T}] = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 0 & 4 \\ 0 & 4 & 0 \end{bmatrix}$.

2.61 A tensor **T** has a matrix given below. Find the principal values and three mutually perpendicular principal directions.

$$[\mathbf{T}] = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

PART C: TENSOR CALCULUS

2.26 TENSOR-VALUED FUNCTIONS OF A SCALAR

Let $\mathbf{T} = \mathbf{T}(t)$ be a tensor-valued function of a scalar t (such as time). The derivative of **T** with respect to t is defined to be a second-order tensor given by:

$$\frac{d\mathbf{T}}{dt} = \lim_{\Delta t \to 0} \frac{\mathbf{T}(t+\Delta t) - \mathbf{T}(t)}{\Delta t}. \tag{2.26.1}$$

The following identities can be easily established:

$$\frac{d}{dt}(\mathbf{T}+\mathbf{S}) = \frac{d\mathbf{T}}{dt} + \frac{d\mathbf{S}}{dt}, \tag{2.26.2}$$

$$\frac{d}{dt}(\alpha(t)\mathbf{T}) = \frac{d\alpha}{dt}\mathbf{T} + \alpha\frac{d\mathbf{T}}{dt}, \tag{2.26.3}$$

$$\frac{d}{dt}(\mathbf{TS}) = \frac{d\mathbf{T}}{dt}\mathbf{S} + \mathbf{T}\frac{d\mathbf{S}}{dt}, \tag{2.26.4}$$

$$\frac{d}{dt}(\mathbf{Ta}) = \frac{d\mathbf{T}}{dt}\mathbf{a} + \mathbf{T}\frac{d\mathbf{a}}{dt}, \tag{2.26.5}$$

$$\frac{d}{dt}(\mathbf{T}^{\mathrm{T}}) = \left(\frac{d\mathbf{T}}{dt}\right)^{\mathrm{T}}. \tag{2.26.6}$$

We shall prove here only Eq. (2.26.5). The other identities can be proven in a similar way. Using the definition given in Eq. (2.26.1), we have

$$\begin{aligned}
\frac{d}{dt}(\mathbf{Ta}) &= \lim_{\Delta t \to 0} \frac{\mathbf{T}(t+\Delta t)\mathbf{a}(t+\Delta t) - \mathbf{T}(t)\mathbf{a}(t)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{\mathbf{T}(t+\Delta t)\mathbf{a}(t+\Delta t) - \mathbf{T}(t)\mathbf{a}(t) - \mathbf{T}(t)\mathbf{a}(t+\Delta t) + \mathbf{T}(t)\mathbf{a}(t+\Delta t)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{\mathbf{T}(t+\Delta t)\mathbf{a}(t+\Delta t) - \mathbf{T}(t)\mathbf{a}(t+\Delta t) + \mathbf{T}(t)\mathbf{a}(t+\Delta t) - \mathbf{T}(t)\mathbf{a}(t)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{\mathbf{T}(t+\Delta t) - \mathbf{T}(t)}{\Delta t}\mathbf{a}(t+\Delta t) + \lim_{\Delta t \to 0} \mathbf{T}(t)\frac{\mathbf{a}(t+\Delta t) - \mathbf{a}(t)}{\Delta t}.
\end{aligned}$$

Thus,

$$\frac{d(\mathbf{Ta})}{dt} = \frac{d\mathbf{T}}{dt}\mathbf{a} + \mathbf{T}\frac{d\mathbf{a}}{dt}.$$

Example 2.26.1

Show that in Cartesian coordinates, the components of $d\mathbf{T}/dt$, i.e., $(d\mathbf{T}/dt)_{ij}$ are given by the derivatives of the components dT_{ij}/dt.

Solution

From

$$T_{ij} = \mathbf{e}_i \cdot \mathbf{T}\mathbf{e}_j,$$

we have

$$\frac{dT_{ij}}{dt} = \frac{d\mathbf{e}_i}{dt} \cdot \mathbf{T}\mathbf{e}_j + \mathbf{e}_i \cdot \frac{d\mathbf{T}}{dt}\mathbf{e}_j + \mathbf{e}_i \cdot \mathbf{T}\frac{d\mathbf{e}_j}{dt}.$$

Since the base vectors are fixed, their derivatives are zero; therefore,

$$\frac{dT_{ij}}{dt} = \mathbf{e}_i \cdot \frac{d\mathbf{T}}{dt}\mathbf{e}_j = \left(\frac{d\mathbf{T}}{dt}\right)_{ij}.$$

Example 2.26.2

Show that for an orthogonal tensor $\mathbf{Q}(t)$, $\left(\frac{d\mathbf{Q}}{dt}\right)\mathbf{Q}^{\mathrm{T}}$ is an antisymmetric tensor.

Solution

Since $\mathbf{Q}\mathbf{Q}^{\mathrm{T}} = \mathbf{I}$, we have

$$\frac{d\left(\mathbf{Q}\mathbf{Q}^{\mathrm{T}}\right)}{dt} = \mathbf{Q}\frac{d\mathbf{Q}^{\mathrm{T}}}{dt} + \frac{d\mathbf{Q}}{dt}\mathbf{Q}^{\mathrm{T}} = \frac{d\mathbf{I}}{dt} = \mathbf{0}.$$

Since [see Eq. (2.26.6)] $\frac{d\mathbf{Q}^{T}}{dt} = \left(\frac{d\mathbf{Q}}{dt}\right)^{\mathrm{T}}$, therefore, the above equation leads to

$$\mathbf{Q}\left(\frac{d\mathbf{Q}}{dt}\right)^{\mathrm{T}} = -\frac{d\mathbf{Q}}{dt}\mathbf{Q}^{\mathrm{T}}.$$

Now $\mathbf{Q}\left(\frac{d\mathbf{Q}}{dt}\right)^{\mathrm{T}} = \left(\frac{d\mathbf{Q}}{dt}\mathbf{Q}^{\mathrm{T}}\right)^{\mathrm{T}}$; therefore,

$$\left(\frac{d\mathbf{Q}}{dt}\mathbf{Q}^{\mathrm{T}}\right)^{\mathrm{T}} = -\frac{d\mathbf{Q}}{dt}\mathbf{Q}^{\mathrm{T}},$$

that is, $\left(\frac{d\mathbf{Q}}{dt}\right)\mathbf{Q}^{\mathrm{T}}$ is an antisymmetric tensor.

Example 2.26.3

A time-dependent rigid body rotation about a fixed point can be represented by a rotation tensor $\mathbf{R}(t)$, so that a position vector $\mathbf{r}_{\mathrm{o}}$ is transformed through the rotation into $\mathbf{r}(t) = \mathbf{R}(t)\mathbf{r}_{\mathrm{o}}$. Derive the equation

$$\frac{d\mathbf{r}}{dt} = \boldsymbol{\omega} \times \mathbf{r}, \tag{2.26.7}$$

where $\boldsymbol{\omega}$ is the dual vector of the antisymmetric tensor $\frac{d\mathbf{R}}{dt}\mathbf{R}^{\mathrm{T}}$.

Solution

From $\mathbf{r}(t) = \mathbf{R}(t)\mathbf{r}_o$, we obtain

$$\frac{d\mathbf{r}}{dt} = \frac{d\mathbf{R}}{dt}\mathbf{r}_o = \frac{d\mathbf{R}}{dt}\mathbf{R}^{-1}\mathbf{r} = \frac{d\mathbf{R}}{dt}\mathbf{R}^{\mathrm{T}}\mathbf{r}. \tag{i}$$

But $\frac{d\mathbf{R}}{dt}\mathbf{R}^{\mathrm{T}}$ is an antisymmetric tensor (see the previous example, Example 2.26.2) so that

$$\frac{d\mathbf{r}}{dt} = \boldsymbol{\omega} \times \mathbf{r}, \tag{ii}$$

where $\boldsymbol{\omega}$ is the dual vector of $\frac{d\mathbf{R}}{dt}\mathbf{R}^{\mathrm{T}}$. From the well-known equation in rigid body kinematics, we can identify $\boldsymbol{\omega}$ as the angular velocity of the rigid body.

2.27 SCALAR FIELD AND GRADIENT OF A SCALAR FUNCTION

Let $\phi(\mathbf{r})$ be a scalar-valued function of the position vector $\mathbf{r}$. That is, for each position $\mathbf{r}$, $\phi(\mathbf{r})$ gives the value of a scalar, such as density, temperature, or electric potential at the point. In other words, $\phi(\mathbf{r})$ describes a scalar field. Associated with a scalar field is a vector field, called the *gradient* of ϕ. The gradient of ϕ at a point is defined to be a vector, denoted by grad ϕ or by $\nabla\phi$ such that its dot product with $d\mathbf{r}$ gives the difference of the values of the scalar at $\mathbf{r} + d\mathbf{r}$ and $\mathbf{r}$, i.e.,

$$d\phi = \phi(\mathbf{r} + d\mathbf{r}) - \phi(\mathbf{r}) = \nabla\phi \cdot d\mathbf{r}. \tag{2.27.1}$$

If dr denote the magnitude of $d\mathbf{r}$, and $\mathbf{e}$ the unit vector in the direction of $d\mathbf{r}$ (*Note:* $\mathbf{e} = d\mathbf{r}/dr$). Then the above equation gives, for $d\mathbf{r}$ in the $\mathbf{e}$ direction,

$$\frac{d\phi}{dr} = \nabla\phi \cdot \mathbf{e}. \tag{2.27.2}$$

That is, the component of $\nabla\phi$ in the direction of $\mathbf{e}$ gives the rate of change of ϕ in that direction (directional derivative). In particular, the components of $\nabla\phi$ in the coordinate directions $\mathbf{e}_i$ are given by

$$\frac{\partial\phi}{\partial x_i} = \left(\frac{d\phi}{dr}\right)_{\mathbf{e}_i - dir} = \nabla\phi \cdot \mathbf{e}_i. \tag{2.27.3}$$

Therefore, the Cartesian components of $\nabla\phi$ are $\partial\phi/\partial x_i$, that is,

$$\nabla\phi = \frac{\partial\phi}{\partial x_1}\mathbf{e}_1 + \frac{\partial\phi}{\partial x_2}\mathbf{e}_2 + \frac{\partial\phi}{\partial x_3}\mathbf{e}_3 = \frac{\partial\phi}{\partial x_i}\mathbf{e}_i. \tag{2.27.4}$$

The gradient vector has a simple geometrical interpretation. For example, if $\phi(\mathbf{r})$ describes a temperature field, then, on a surface of constant temperature (i.e., isothermal surface), $\phi =$ a constant. Let $\mathbf{r}$ be a point on an isothermal surface. Then, for any and all neighboring point $\mathbf{r}+d\mathbf{r}$ on the same isothermal surface, $d\phi = 0$. Thus,$\nabla\phi \cdot d\mathbf{r} = 0$. In other words, $\nabla\phi$ is a vector, perpendicular to the surface at the point $\mathbf{r}$. On the other hand, the dot product $\nabla\phi \cdot d\mathbf{r}$ is a maximum when $d\mathbf{r}$ is in the same direction as $\nabla\phi$. In other words, $\nabla\phi$ is greatest if $d\mathbf{r}$ is normal to the surface of constant ϕ and in this case, $d\phi = |\nabla\phi|dr$, or

$$\left(\frac{d\phi}{dr}\right)_{\max} = |\nabla\phi|, \tag{2.27.5}$$

for $d\mathbf{r}$ in the direction normal to the surface of constant temperature.

Example 2.27.1

If $\phi = x_1x_2 + 2x_3$, find a unit vector **n** normal to the surface of a constant ϕ passing through the point (2,1,0).

Solution

By Eq. (2.27.4),

$$\nabla\phi = \frac{\partial\phi}{\partial x_1}\mathbf{e}_1 + \frac{\partial\phi}{\partial x_2}\mathbf{e}_2 + \frac{\partial\phi}{\partial x_3}\mathbf{e}_3 = x_2\mathbf{e}_1 + x_1\mathbf{e}_2 + 2\mathbf{e}_3.$$

At the point (2,1,0), $\nabla\phi = \mathbf{e}_1 + 2\mathbf{e}_2 + 2\mathbf{e}_3$. Thus,

$$\mathbf{n} = \frac{1}{3}(\mathbf{e}_1 + 2\mathbf{e}_2 + 2\mathbf{e}_3).$$

Example 2.27.2

If **q** denotes the heat flux vector (rate of heat transfer/area), the Fourier heat conduction law states that

$$\mathbf{q} = -k\nabla\Theta, \tag{i}$$

where Θ is the temperature field and k is thermal conductivity. If $\Theta = 2(x_1^2 + x_2^2)$, find $\nabla\Theta$ at the location A (1,0) and $B\left(1/\sqrt{2}, 1/\sqrt{2}\right)$. Sketch curves of constant Θ (isotherms) and indicate the vectors **q** at the two points.

Solution

By Eq. (2.27.4),

$$\nabla\Theta = \frac{\partial\Theta}{\partial x_1}\mathbf{e}_1 + \frac{\partial\Theta}{\partial x_2}\mathbf{e}_2 + \frac{\partial\Theta}{\partial x_3}\mathbf{e}_3 = 4x_1\mathbf{e}_1 + 4x_2\mathbf{e}_2.$$

Thus,

$$\mathbf{q} = -4k(x_1\mathbf{e}_1 + x_2\mathbf{e}_2).$$

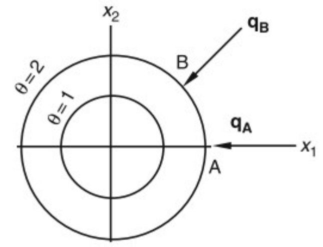

FIGURE 2.27-1

At point A,

$$\mathbf{q}_A = -4k\mathbf{e}_1,$$

and at point B,

$$\mathbf{q}_B = -2\sqrt{2}k(\mathbf{e}_1 + \mathbf{e}_2).$$

Clearly, the isotherms, Figure 2.27-1, are circles and the heat flux is an inward radial vector (consistent with heat flowing from higher to lower temperatures).

Example 2.27.3

A more general heat conduction law can be given in the following form:

$$\mathbf{q} = -\mathbf{K}\nabla\Theta,$$

where **K** is a tensor known as thermal conductivity tensor. (a) What tensor **K** corresponds to the Fourier heat conduction law mentioned in the previous example? (b) Find **q** if $\Theta = 2x_1 + 3x_2$, and

$$[\mathbf{K}] = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}.$$

Solution

(a) Clearly, $\mathbf{K} = k\mathbf{I}$, so that $\mathbf{q} = -k\mathbf{I}\nabla\Theta = -k\nabla\Theta$.

(b) $\nabla\Theta = 2\mathbf{e}_1 + 3\mathbf{e}_2$ and

$$[\mathbf{q}] = -\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}\begin{bmatrix} 2 \\ 3 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ -4 \\ 0 \end{bmatrix}$$

that is,

$$\mathbf{q} = -\mathbf{e}_1 - 4\mathbf{e}_2,$$

which is clearly not normal to the isotherm (see Figure 2.27-2).

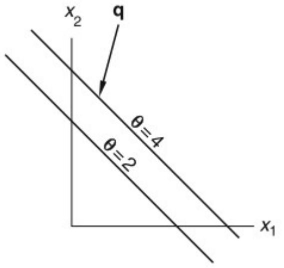

FIGURE 2.27-2

2.28 VECTOR FIELD AND GRADIENT OF A VECTOR FUNCTION

Let $\mathbf{v}(\mathbf{r})$ be a vector-valued function of position describing, for example, a displacement or a velocity field. Associated with $\mathbf{v}(\mathbf{r})$, is a tensor field, called the *gradient* of v, which is of considerable importance. The gradient of $\mathbf{v}$ (denoted by $\nabla\mathbf{v}$ or grad $\mathbf{v}$) is defined to be the second-order tensor, which, when operating on $d\mathbf{r}$, gives the difference of $\mathbf{v}$ at $\mathbf{r} + d\mathbf{r}$ and $\mathbf{r}$. That is,

$$d\mathbf{v} = \mathbf{v}(\mathbf{r} + d\mathbf{r}) - \mathbf{v}(\mathbf{r}) = (\nabla\mathbf{v})d\mathbf{r}. \tag{2.28.1}$$

Again, let dr denote $|d\mathbf{r}|$ and $\mathbf{e}$ denote $d\mathbf{r}/dr$; we have

$$\left(\frac{d\mathbf{v}}{dr}\right)_{\text{in } \mathbf{e}-\text{direction}} = (\nabla\mathbf{v})\mathbf{e}. \tag{2.28.2}$$

Therefore, the second-order tensor $\nabla\mathbf{v}$ transforms a unit vector $\mathbf{e}$ into the vector describing the rate of change of $\mathbf{v}$ in that direction. In Cartesian coordinates,

$$\left(\frac{d\mathbf{v}}{dr}\right)_{\text{in } \mathbf{e}_j-\text{direction}} = \frac{\partial\mathbf{v}}{\partial x_j} = (\nabla\mathbf{v})\mathbf{e}_j, \tag{2.28.3}$$

therefore, the components of $\nabla\mathbf{v}$ in indicial notation are given by

$$(\nabla\mathbf{v})_{ij} = \mathbf{e}_i \cdot (\nabla\mathbf{v})\mathbf{e}_j = \mathbf{e}_i \cdot \frac{\partial\mathbf{v}}{\partial x_j} = \frac{\partial(\mathbf{v}\cdot\mathbf{e}_i)}{\partial x_j} = \frac{\partial v_i}{\partial x_j}, \tag{2.28.4}$$

and in matrix form

$$[\nabla\mathbf{v}] = \begin{bmatrix} \frac{\partial v_1}{\partial x_1} & \frac{\partial v_1}{\partial x_2} & \frac{\partial v_1}{\partial x_3} \\ \frac{\partial v_2}{\partial x_1} & \frac{\partial v_2}{\partial x_2} & \frac{\partial v_2}{\partial x_3} \\ \frac{\partial v_3}{\partial x_1} & \frac{\partial v_3}{\partial x_2} & \frac{\partial v_3}{\partial x_3} \end{bmatrix}. \tag{2.28.5}$$

Geometrical interpretation of $\nabla\mathbf{v}$ will be given later in connection with the deformation of a continuum (Chapter 3).

2.29 DIVERGENCE OF A VECTOR FIELD AND DIVERGENCE OF A TENSOR FIELD

Let $\mathbf{v}(\mathbf{r})$ be a vector field. The *divergence* of $\mathbf{v}(\mathbf{r})$ is defined to be a scalar field given by the trace of the gradient of $\mathbf{v}$. That is,

$$\text{div } \mathbf{v} \equiv \text{tr}(\nabla\mathbf{v}). \tag{2.29.1}$$

In Cartesian coordinates, this gives

$$\text{div } \mathbf{v} = \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3} = \frac{\partial v_i}{\partial x_i}. \tag{2.29.2}$$

Let $\mathbf{T}(\mathbf{r})$ be a tensor field. The divergence of $\mathbf{T}(\mathbf{r})$ is defined to be a vector field, denoted by div $\mathbf{T}$, such that for any vector $\mathbf{a}$

$$(\text{div } \mathbf{T}) \cdot \mathbf{a} \equiv \text{div}\left(\mathbf{T}^{\mathrm{T}}\mathbf{a}\right) - \text{tr}\left(\mathbf{T}^{\mathrm{T}}\nabla\mathbf{a}\right). \tag{2.29.3}$$

To find the Cartesian components of the vector div $\mathbf{T}$, let $\mathbf{b} = \text{div } \mathbf{T}$, then (*Note:* $\nabla\mathbf{e}_i = \mathbf{0}$ for Cartesian coordinates), from (2.29.3), we have

$$b_i = \mathbf{b} \cdot \mathbf{e}_i = \text{div}\left(\mathbf{T}^{\mathrm{T}}\mathbf{e}_i\right) - \text{tr}\left(\mathbf{T}^{\mathrm{T}}\nabla\mathbf{e}_i\right) = \text{div}\left(T_{ij}\mathbf{e}_j\right) - 0 = \partial T_{ij}/\partial x_j. \tag{2.29.4}$$

In other words,

$$\text{div } \mathbf{T} = \left(\partial T_{ij}/\partial x_j\right)\mathbf{e}_i. \tag{2.29.5*}$$

Example 2.29.1

Let $\alpha = \alpha(\mathbf{r})$ and $\mathbf{a} = \mathbf{a}(\mathbf{r})$. Show that $\text{div}(\alpha\mathbf{a}) = \alpha\text{div } \mathbf{a} + (\nabla\mathbf{v}) \cdot \mathbf{a}$.

Solution

Let $\mathbf{b} = \alpha\mathbf{a}$. Then $b_i = \alpha a_i$, so

$$\text{div } \mathbf{b} = \frac{\partial b_i}{\partial x_i} = \alpha\frac{\partial a_i}{\partial x_i} + \frac{\partial\alpha}{\partial x_i}a_i.$$

That is,

$$\text{div } (\alpha\mathbf{a}) = \alpha \text{ div } \mathbf{a} + (\nabla\alpha) \cdot \mathbf{a}. \tag{2.29.6}$$

Example 2.29.2

Given $\alpha = \alpha(\mathbf{r})$ and $\mathbf{T} = \mathbf{T}(\mathbf{r})$, show that

$$\text{div } (\alpha\mathbf{T}) = \mathbf{T}(\nabla\alpha) + \alpha \text{ div } \mathbf{T}. \tag{2.29.7}$$

Solution

We have, from (2.29.5),

$$\text{div } (\alpha\mathbf{T}) = \frac{\partial\left(\alpha T_{ij}\right)}{\partial x_j}\mathbf{e}_i = \frac{\partial\alpha}{\partial x_j}T_{ij}\mathbf{e}_i + \alpha\frac{\partial T_{ij}}{\partial x_j}\mathbf{e}_i = \mathbf{T}(\nabla\alpha) + \alpha \text{ div } \mathbf{T}.$$

---

*We note that the Cartesian components of the third-order tensor $\mathbf{M} \equiv \nabla\mathbf{T} = \nabla\left(T_{ij}\mathbf{e}_i\mathbf{e}_j\right)$are $\partial T_{ij}/\partial x_k$. In terms of $\mathbf{M} = M_{ijk}\mathbf{e}_i\mathbf{e}_j\mathbf{e}_k$, div $\mathbf{T}$ is a vector given by $M_{ijj}\mathbf{e}_i$. More on the components of $\nabla\mathbf{T}$ will be given in Chapter 8.

2.30 CURL OF A VECTOR FIELD

Let $\mathbf{v}(\mathbf{r})$ be a vector field. The *curl* of $\mathbf{v}(\mathbf{r})$ is defined to be a vector field given by twice the dual vector of the antisymmetric part of $\nabla\mathbf{v}$. That is

$$\text{curl } \mathbf{v} \equiv 2\mathbf{t}^{\mathbf{A}}, \tag{2.30.1}$$

where $\mathbf{t}^{\mathrm{A}}$ is the dual vector of $(\nabla\mathbf{v})^{\mathrm{A}}$.

In rectangular Cartesian coordinates,

$$[\nabla\mathbf{v}]^{\mathrm{A}} = \begin{bmatrix} 0 & \frac{1}{2}\left(\frac{\partial v_1}{\partial x_2} - \frac{\partial v_2}{\partial x_1}\right) & \frac{1}{2}\left(\frac{\partial v_1}{\partial x_3} - \frac{\partial v_3}{\partial x_1}\right) \\ -\frac{1}{2}\left(\frac{\partial v_1}{\partial x_2} - \frac{\partial v_2}{\partial x_1}\right) & 0 & \frac{1}{2}\left(\frac{\partial v_2}{\partial x_3} - \frac{\partial v_3}{\partial x_2}\right) \\ -\frac{1}{2}\left(\frac{\partial v_1}{\partial x_3} - \frac{\partial v_3}{\partial x_1}\right) & -\frac{1}{2}\left(\frac{\partial v_2}{\partial x_3} - \frac{\partial v_3}{\partial x_2}\right) & 0 \end{bmatrix}. \tag{2.30.2}$$

Thus, the curl of $\mathbf{v}(\mathbf{r})$ is given by [see Eq. (2.21.3)]:

$$\text{curl } \mathbf{v} = 2\mathbf{t}^{\mathrm{A}} = \left(\frac{\partial v_3}{\partial x_2} - \frac{\partial v_2}{\partial x_3}\right)\mathbf{e}_1 + \left(\frac{\partial v_1}{\partial x_3} - \frac{\partial v_3}{\partial x_1}\right)\mathbf{e}_2 + \left(\frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2}\right)\mathbf{e}_3. \tag{2.30.3}$$

It can be easily verified that in indicial notation

$$\text{curl } \mathbf{v} = -\varepsilon_{ijk}\frac{\partial v_j}{\partial x_k}\mathbf{e}_i. \tag{2.30.4}$$

2.31 LAPLACIAN OF A SCALAR FIELD

Let $f(\mathbf{r})$ be a scalar-valued function of the position vector $\mathbf{r}$. The definition of the Laplacian of a scalar field is given by

$$\nabla^2 f = \text{div } (\nabla f) = \text{tr}(\nabla(\nabla f)). \tag{2.31.1}$$

In rectangular coordinates the Laplacian becomes

$$\nabla^2 f = \text{tr}(\nabla(\nabla f)) = \frac{\partial^2 f}{\partial x_i \partial x_i} = \frac{\partial^2 f}{\partial x_1^2} + \frac{\partial^2 f}{\partial x_2^2} + \frac{\partial^2 f}{\partial x_3^2}. \tag{2.31.2}$$

2.32 LAPLACIAN OF A VECTOR FIELD

Let $\mathbf{v}(\mathbf{r})$ be a vector field. The Laplacian of $\mathbf{v}$ is defined by the following:

$$\nabla^2\mathbf{v} = \nabla\ (\text{div } \mathbf{v}) - \text{curl } (\text{curl } \mathbf{v}). \tag{2.32.1}$$

In rectangular coordinates,

$$\nabla(\text{div }\mathbf{v}) = \frac{\partial}{\partial x_i}\left(\frac{\partial v_k}{\partial x_k}\right)\mathbf{e}_i, \quad \text{curl }\mathbf{v} = -\varepsilon_{\alpha jk}\left(\frac{\partial v_j}{\partial x_k}\right)\mathbf{e}_\alpha, \tag{2.32.2}$$

and

$$\text{curl (curl }\mathbf{v}) = -\varepsilon_{i\alpha\beta}\frac{\partial}{\partial x_\beta}\left(-\varepsilon_{\alpha jk}\frac{\partial v_j}{\partial x_k}\right)\mathbf{e}_i = \varepsilon_{i\alpha\beta}\varepsilon_{\alpha jk}\frac{\partial}{\partial x_\beta}\left(\frac{\partial v_j}{\partial x_k}\right)\mathbf{e}_i. \tag{2.32.3}$$

Now $\varepsilon_{i\alpha\beta}\varepsilon_{\alpha jk} = -\varepsilon_{\alpha i\beta}\varepsilon_{\alpha jk} = -\left(\delta_{ij}\delta_{\beta k} - \delta_{ik}\delta_{\beta j}\right)$ [see Prob. 2.12], therefore,

$$\text{curl (curl }\mathbf{v}) = -\left(\delta_{ij}\delta_{\beta k} - \delta_{ik}\delta_{\beta j}\right)\frac{\partial}{\partial x_\beta}\left(\frac{\partial v_j}{\partial x_k}\right)\mathbf{e}_i = \left\{-\frac{\partial}{\partial x_\beta}\left(\frac{\partial v_i}{\partial x_\beta}\right) + \frac{\partial}{\partial x_\beta}\left(\frac{\partial v_\beta}{\partial x_i}\right)\right\}\mathbf{e}_i. \tag{2.32.4}$$

Thus,

$$\nabla^2\mathbf{v} = \nabla\text{ (div }\mathbf{v}) - \text{curl (curl }\mathbf{v}) = \frac{\partial}{\partial x_i}\left(\frac{\partial v_k}{\partial x_k}\right)\mathbf{e}_i - \left\{-\frac{\partial}{\partial x_\beta}\left(\frac{\partial v_i}{\partial x_\beta}\right) + \frac{\partial}{\partial x_i}\left(\frac{\partial v_\beta}{\partial x_\beta}\right)\right\}\mathbf{e}_i. \tag{2.32.5}$$

That is, in rectangular coordinates,

$$\nabla^2\mathbf{v} = \frac{\partial^2 v_i}{\partial x_\beta \partial x_\beta}\mathbf{e}_i = \nabla^2 v_i \mathbf{e}_\mathrm{i}. \tag{2.32.6}$$

In long form,

$$\nabla^2\mathbf{v} = \left(\frac{\partial^2 v_1}{\partial x_1^2} + \frac{\partial^2 v_1}{\partial x_2^2} + \frac{\partial^2 v_1}{\partial x_3^2}\right)\mathbf{e}_1 + \left(\frac{\partial^2 v_2}{\partial x_1^2} + \frac{\partial^2 v_2}{\partial x_2^2} + \frac{\partial^2 v_2}{\partial x_3^2}\right)\mathbf{e}_2 + \left(\frac{\partial^2 v_3}{\partial x_1^2} + \frac{\partial^2 v_3}{\partial x_2^2} + \frac{\partial^2 v_3}{\partial x_3^2}\right)\mathbf{e}_3. \tag{2.32.7}$$

Expressions for the polar, cylindrical, and spherical coordinate systems are given in Part D.

PROBLEMS FOR PART C

2.62 Prove the identity $\frac{d}{dt}(\mathbf{T}+\mathbf{S}) = \frac{d\mathbf{T}}{dt} + \frac{d\mathbf{S}}{dt}$ using the definition of derivative of a tensor.

2.63 Prove the identity $\frac{d}{dt}(\mathbf{TS}) = \mathbf{T}\frac{d\mathbf{S}}{dt} + \frac{d\mathbf{T}}{dt}\mathbf{S}$ using the definition of derivative of a tensor.

2.64 Prove that $\frac{d\mathbf{T}^\mathrm{T}}{dt} = \left(\frac{d\mathbf{T}}{dt}\right)^\mathrm{T}$ by differentiating the definition $\mathbf{a}\cdot\mathbf{Tb} = \mathbf{b}\cdot\mathbf{T}^\mathrm{T}\mathbf{a}$, where **a** and **b** are arbitrary constant vectors.

2.65 Consider the scalar field $\phi = x_1^2 + 3x_1x_2 + 2x_3$. (a) Find the unit vector normal to the surface of constant ϕ at the origin and at (1,0,1). (b) What is the maximum value of the directional derivative of ϕ at the origin? at (1,0,1)? (c) Evaluate $d\phi/dr$ at the origin if $d\mathbf{r} = ds(\mathbf{e}_1 + \mathbf{e}_3)$.

2.66 Consider the ellipsoidal surface defined by the equation $x^2/a^2 + y^2/b^2 + z^2/b^2 = 1$. Find the unit vector normal to the surface at a given point (x, y, z).

2.67 Consider the temperature field given by $\Theta = 3x_1x_2$. (a) If $\mathbf{q} = -k\nabla\Theta$, find the heat flux at the point $A(1,1,1)$. (b) If $\mathbf{q} = -\mathbf{K}\nabla\Theta$, find the heat flux at the same point, where

$$[\mathbf{K}] = \begin{bmatrix} k & 0 & 0 \\ 0 & 2k & 0 \\ 0 & 0 & 3k \end{bmatrix}.$$

2.68 Let $\phi(x_1, x_2, x_3)$ and $\psi(x_1, x_2, x_3)$ be scalar fields, and let $\mathbf{v}\ (x_1, x_2, x_3)$ and $\mathbf{w}\ (x_1, x_2, x_3)$ be vector fields. By writing the subscripted components form, verify the following identities:

(a) $\nabla(\phi + \psi) = \nabla\phi + \nabla\psi$, sample solution:

$$[\nabla(\phi + \psi)]_i = \frac{\partial(\phi + \psi)}{\partial x_i} = \frac{\partial\phi}{\partial x_i} + \frac{\partial\psi}{\partial x_i} = \nabla\phi + \nabla\psi,$$

(b) $\text{div}(\mathbf{v} + \mathbf{w}) = \text{div}\ \mathbf{v} + \text{div}\ \ \mathbf{w}$, (c) $\text{div}(\phi\mathbf{v}) = (\nabla\phi)\mathbf{v} + \phi(\text{div}\ \mathbf{v})$ and (d) $\text{div}(\text{curl}\ \mathbf{v}) = 0$.

2.69 Consider the vector field $\mathbf{v} = x_1^2\mathbf{e}_1 + x_3^2\mathbf{e}_2 + x_2^2\mathbf{e}_3$. For the point (1,1,0), find (a) $\nabla\mathbf{v}$, (b) $(\nabla\mathbf{v})\mathbf{v}$, (c) div $\mathbf{v}$ and curl $\mathbf{v}$, and (d) the differential $d\mathbf{v}$ for $d\mathbf{r} = ds(\mathbf{e}_1 + \mathbf{e}_2 + \mathbf{e}_3)/\sqrt{3}$.

PART D: CURVILINEAR COORDINATES

In Part C, the Cartesian components for various vector and tensor operations such as the gradient, the divergence, and the Laplacian of a scalar field and tensor fields were derived. In this part, components in polar, cylindrical, and spherical coordinates for these same operations will be derived.

2.33 POLAR COORDINATES

Consider polar coordinates (r,θ), (see Figure 2.33-1) such that

$$r = \sqrt{x_1^2 + x_2^2} \quad \text{and} \quad \theta = \tan^{-1}\frac{x_2}{x_1}. \tag{2.33.1}$$

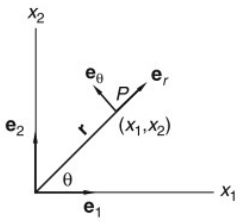

FIGURE 2.33-1

The unit base vectors $\mathbf{e}_r$ and $\mathbf{e}_\theta$ can be expressed in terms of the Cartesian base vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ as

$$\mathbf{e}_r = \cos\theta\,\mathbf{e}_1 + \sin\theta\,\mathbf{e}_2, \quad \mathbf{e}_\theta = -\sin\theta\,\mathbf{e}_1 + \cos\theta\,\mathbf{e}_2. \tag{2.33.2}$$

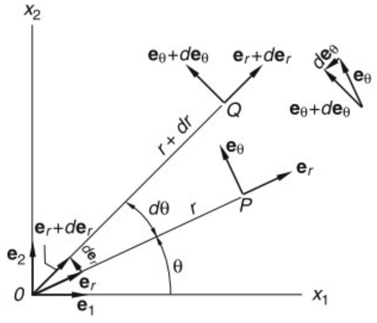
x_2
$\mathbf{e}_\theta+d\mathbf{e}_\theta$
$\mathbf{e}_r+d\mathbf{e}_r$
$d\mathbf{e}_\theta$
$\mathbf{e}_\theta$
Q
$\mathbf{e}_\theta+d\mathbf{e}_\theta$
$r+dr$
$\mathbf{e}_\theta$
$\mathbf{e}_r$
$d\theta$
r
P
$\mathbf{e}_r+d\mathbf{e}_r$
$d\mathbf{e}_r$
$\mathbf{e}_2$
θ
$\mathbf{e}_r$
0
$\mathbf{e}_1$
x_1

FIGURE 2.33-2

These unit base vectors vary in direction as θ changes. In fact, from Eqs. (2.33.2), we have

$$d\mathbf{e}_r = (-\sin\theta\mathbf{e}_1 + \cos\theta\mathbf{e}_2)d\theta = d\theta\mathbf{e}_\theta, \quad d\mathbf{e}_\theta = (-\cos\theta\mathbf{e}_1 - \sin\theta\mathbf{e}_2)d\theta = -d\theta\mathbf{e}_r. \tag{2.33.3}$$

The geometrical representation of $d\mathbf{e}_r$ and $d\mathbf{e}_\theta$ are shown in Figure 2.33-2, where one notes that $\mathbf{e}_r(P)$ has rotated an infinitesimal angle $d\theta$ to become $\mathbf{e}_r(Q) = \mathbf{e}_r(P) + d\mathbf{e}_r$ where $d\mathbf{e}_r$ is perpendicular to $\mathbf{e}_r(P)$ with a magnitude $|d\mathbf{e}_r| = (1)d\theta = d\theta$. Similarly, $d\mathbf{e}_\theta$ is perpendicular to $\mathbf{e}_\theta$ (P) but pointing in the negative $\mathbf{e}_r$ direction, and its magnitude is also $d\theta$.

Now, from the position vector

$$\mathbf{r} = r\mathbf{e}_r, \tag{2.33.4}$$

we have

$$d\mathbf{r} = dr\mathbf{e}_r + rd\mathbf{e}_r. \tag{2.33.5}$$

Using Eq. (2.33.3), we get

$$d\mathbf{r} = dr\mathbf{e}_r + rd\theta\mathbf{e}_\theta. \tag{2.33.6}$$

The geometrical representation of this equation is also easily seen if one notes that $d\mathbf{r}$ is the vector PQ in the preceding figure.

The components of ∇f, $\nabla\mathbf{v}$, div $\mathbf{v}$, div $\mathbf{T}$, $\nabla^2 f$ and $\nabla^2\mathbf{v}$ in polar coordinates will now be obtained.

(i) Components of ∇f:

Let $f(r,\theta)$ be a scalar field. By definition of the gradient of f, we have

$$df = \nabla f\cdot d\mathbf{r} = (a_r\mathbf{e}_r + a_\theta\mathbf{e}_\theta)\cdot(dr\mathbf{e}_r + rd\theta\mathbf{e}_\theta) = a_r dr + a_\theta rd\theta, \tag{2.33.7}$$

where a_r and a_θ are components of ∇f in the $\mathbf{e}_r$ and $\mathbf{e}_\theta$ direction, respectively. But from calculus,

$$df = \frac{\partial f}{\partial r}dr + \frac{\partial f}{\partial\theta}d\theta. \tag{2.33.8}$$

Since Eqs. (2.33.7) and (2.33.8) must yield the same result for all increments $dr, d\theta$, we have

$$a_r = \frac{\partial f}{\partial r}, \quad a_\theta = \frac{1}{r}\frac{\partial f}{\partial\theta}, \tag{2.33.9}$$

thus,

$$\nabla f = \frac{\partial f}{\partial r}\mathbf{e}_r + \frac{1}{r}\frac{\partial f}{\partial \theta}\mathbf{e}_\theta. \tag{2.33.10}$$

(ii) Components of $\nabla\mathbf{v}$: Let

$$\mathbf{v}(r,\theta) = v_r(r,\theta)\mathbf{e}_r + \mathrm{v}_\theta(r,\theta)\mathbf{e}_\theta. \tag{2.33.11}$$

By definition of $\nabla\mathbf{v}$, we have

$$d\mathbf{v} = \nabla\mathbf{v}d\mathbf{r}. \tag{2.33.12}$$

Let $\mathbf{T} = \nabla\mathbf{v}$. Then

$$d\mathbf{v} = \mathbf{T}d\mathbf{r} = \mathbf{T}(dr\mathbf{e}_r + rd\theta\mathbf{e}_\theta) = dr\mathbf{T}\mathbf{e}_r + rd\theta\mathbf{T}\mathbf{e}_\theta. \tag{2.33.13}$$

Now

$$\mathbf{T}\mathbf{e}_r = T_{rr}\mathbf{e}_r + T_{\theta r}\mathbf{e}_\theta \quad \text{and} \quad \mathbf{T}\mathbf{e}_\theta = T_{r\theta}\mathbf{e}_r + T_{\theta\theta}\mathbf{e}_\theta, \tag{2.33.14}$$

therefore,

$$d\mathbf{v} = (T_{rr}dr + T_{r\theta}rd\theta)\mathbf{e}_r + (T_{\theta r}dr + T_{\theta\theta}rd\theta)\mathbf{e}_\theta. \tag{2.33.15}$$

From Eq. (2.33.11), we also have

$$d\mathbf{v} = dv_r\mathbf{e}_r + v_r d\mathbf{e}_r + d\mathrm{v}_\theta\mathbf{e}_\theta + \mathrm{v}_\theta d\mathbf{e}_\theta. \tag{2.33.16}$$

Since [see Eq. (2.33.3)]

$$d\mathbf{e}_r = d\theta\mathbf{e}_\theta, \quad d\mathbf{e}_\theta = -d\theta\mathbf{e}_r, \tag{2.33.17}$$

therefore, Eq. (2.33.16) becomes

$$d\mathbf{v} = (dv_r - \mathrm{v}_\theta d\theta)\mathbf{e}_r + (v_r d\theta + d\mathrm{v}_\theta)\mathbf{e}_\theta. \tag{2.33.18}$$

From calculus,

$$dv_r = \frac{\partial v_r}{\partial r}dr + \frac{\partial v_r}{\partial \theta}d\theta, \quad d\mathrm{v}_\theta = \frac{\partial \mathrm{v}_\theta}{\partial r}dr + \frac{\partial \mathrm{v}_\theta}{\partial \theta}d\theta. \tag{2.33.19}$$

Substituting Eq. (2.33.19) into Eq. (2.33.18), we have

$$d\mathbf{v} = \left[\frac{\partial v_r}{\partial r}dr + \left(\frac{\partial v_r}{\partial \theta} - \mathrm{v}_\theta\right)d\theta\right]\mathbf{e}_r + \left[\frac{\partial \mathrm{v}_\theta}{\partial r}dr + \left(\frac{\partial \mathrm{v}_\theta}{\partial \theta} + v_r\right)d\theta\right]\mathbf{e}_\theta. \tag{2.33.20}$$

Eq. (2.33.15) and Eq. (2.33.20), then, give

$$\frac{\partial v_r}{\partial r}dr + \left(\frac{\partial v_r}{\partial \theta} - \mathrm{v}_\theta\right)d\theta = T_{rr}dr + T_{r\theta}rd\theta, \quad \frac{\partial \mathrm{v}_\theta}{\partial r}dr + \left(\frac{\partial \mathrm{v}_\theta}{\partial \theta} + v_r\right)d\theta = T_{\theta r}dr + T_{\theta\theta}rd\theta. \tag{2.33.21}$$

Eq. (2.33.21) must hold for any values of dr and $d\theta$. Thus,

$$T_{rr} = \frac{\partial v_r}{\partial r}, \quad T_{r\theta} = \frac{1}{r}\left(\frac{\partial v_r}{\partial \theta} - \mathrm{v}_\theta\right), \quad T_{\theta r} = \frac{\partial \mathrm{v}_\theta}{\partial r}, \quad T_{\theta\theta} = \frac{1}{r}\left(\frac{\partial \mathrm{v}_\theta}{\partial \theta} + v_r\right). \tag{2.33.22}$$

In matrix form,

$$[\nabla \mathbf{v}] = \begin{bmatrix} \dfrac{\partial v_r}{\partial r} & \dfrac{1}{r}\left(\dfrac{\partial v_r}{\partial \theta} - v_\theta\right) \\ \dfrac{\partial v_\theta}{\partial r} & \dfrac{1}{r}\left(\dfrac{\partial v_\theta}{\partial \theta} + v_r\right) \end{bmatrix}. \tag{2.33.23}$$

(iii) div **v**:
Using the components of $\nabla \mathbf{v}$ given in (ii), that is, Eq. (2.33.23), we have

$$\operatorname{div} \mathbf{v} = \operatorname{tr}(\nabla \mathbf{v}) = \frac{\partial v_r}{\partial r} + \frac{1}{r}\left(\frac{\partial v_\theta}{\partial \theta} + v_r\right). \tag{2.33.24}$$

(iv) Components of curl **v:**
The antisymmetric part of $\nabla \mathbf{v}$ is

$$[\nabla \mathbf{v}]^{\mathrm{A}} = \frac{1}{2}\begin{bmatrix} 0 & \dfrac{1}{r}\left(\dfrac{\partial v_r}{\partial \theta} - v_\theta\right) - \dfrac{\partial v_\theta}{\partial r} \\ -\left\{\dfrac{1}{r}\left(\dfrac{\partial v_r}{\partial \theta} - v_\theta\right) - \dfrac{\partial v_\theta}{\partial r}\right\} & 0 \end{bmatrix}. \tag{2.33.25}$$

Therefore, from the definition that curl $\mathbf{v}$ = twice the dual vector of $(\nabla \mathbf{v})^{\mathrm{A}}$, we have

$$\operatorname{curl} \mathbf{v} = \left(\frac{\partial v_\theta}{\partial r} + \frac{v_\theta}{r} - \frac{1}{r}\frac{\partial v_r}{\partial \theta}\right)\mathbf{e}_3. \tag{2.33.26}$$

(v) Components of div **T:**
The invariant definition of the divergence of a second-order tensor is

$$(\operatorname{div} \mathbf{T}) \cdot \mathbf{a} = \operatorname{div}(\mathbf{T}^{\mathrm{T}}\mathbf{a}) - \operatorname{tr}((\nabla \mathbf{a})\mathbf{T}^{\mathrm{T}}) \text{ for any } \mathbf{a}. \tag{2.33.27}$$

Take $\mathbf{a} = \mathbf{e}_r$; then the preceding equation gives

$$(\operatorname{div} \mathbf{T})_r = \operatorname{div}(\mathbf{T}^{\mathrm{T}}\mathbf{e}_r) - \operatorname{tr}((\nabla \mathbf{e}_r)\mathbf{T}^{\mathrm{T}}). \tag{2.33.28}$$

To evaluate the first term on the right-hand side, we note that

$$\mathbf{T}^{\mathrm{T}}\mathbf{e}_r = T_{rr}\mathbf{e}_r + T_{r\theta}\mathbf{e}_\theta, \tag{2.33.29}$$

so that according to Eq. (2.33.24),

$$\operatorname{div}(\mathbf{T}^{\mathrm{T}}\mathbf{e}_r) = \operatorname{div}(T_{rr}\mathbf{e}_r + T_{r\theta}\mathbf{e}_\theta) = \frac{\partial T_{rr}}{\partial r} + \frac{1}{r}\left(\frac{\partial T_{r\theta}}{\partial \theta} + T_{rr}\right). \tag{2.33.30}$$

To evaluate the second term, we first use Eq. (2.33.23) to obtain $\nabla \mathbf{e}_r$. In fact, since $\mathbf{e}_r = (1)\mathbf{e}_r + (0)\mathbf{e}_\theta$, we have, with $v_r = 1$ and $v_\theta = 0$,

$$[\nabla \mathbf{e}_r] = \begin{bmatrix} 0 & 0 \\ 0 & \dfrac{1}{r} \end{bmatrix}, \quad [\nabla \mathbf{e}_r][\mathbf{T}^{\mathrm{T}}] = \begin{bmatrix} 0 & 0 \\ \dfrac{T_{r\theta}}{r} & \dfrac{T_{\theta\theta}}{r} \end{bmatrix}, \quad \mathrm{tr}([\nabla \mathbf{e}_r][\mathbf{T}^{\mathrm{T}}]) = \frac{T_{\theta\theta}}{r}. \tag{2.33.31}$$

Thus, Eq. (2.33.28) gives

$$(\mathrm{div}\,\mathbf{T})_r = \frac{\partial T_{rr}}{\partial r} + \frac{1}{r}\frac{\partial T_{r\theta}}{\partial \theta} + \frac{T_{rr} - T_{\theta\theta}}{r}. \tag{2.33.32}$$

In a similar manner, one can derive

$$(\mathrm{div}\,\mathbf{T})_\theta = \frac{\partial T_{\theta r}}{\partial r} + \frac{1}{r}\frac{\partial T_{\theta\theta}}{\partial \theta} + \frac{T_{r\theta} + T_{\theta r}}{r}. \tag{2.33.33}$$

(vi) Laplacian of $f(\mathbf{x})$:
Given a scalar field $f(\mathbf{x})$, the Laplacian of $f(\mathbf{x})$ is given by $\nabla^2 f = \mathrm{div}(\nabla f) = \mathrm{tr}(\nabla(\nabla f))$. In polar coordinates,

$$\nabla f = \frac{\partial f}{\partial r}\mathbf{e}_r + \frac{1}{r}\frac{\partial f}{\partial \theta}\mathbf{e}_\theta. \tag{2.33.34}$$

From, $\mathrm{div}\,\mathbf{v} = \dfrac{\partial v_r}{\partial r} + \dfrac{1}{r}\dfrac{\partial v_\theta}{\partial \theta} + \dfrac{v_r}{r}$, we have

$$\nabla^2 f = \mathrm{div}\,\nabla f = \frac{\partial^2 f}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 f}{\partial \theta^2} + \frac{1}{r}\frac{\partial f}{\partial r}. \tag{2.33.35}$$

(vii) Laplacian of a vector field $\mathbf{v}(\mathbf{x})$:
Laplacian of $\mathbf{v}$ is given by: $\nabla^2 \mathbf{v} = \nabla(\mathrm{div}\,\mathbf{v}) - \mathrm{curl}\,\mathrm{curl}\,\mathbf{v}$. Now, in polar coordinates:

$$\begin{aligned} \nabla(\mathrm{div}\,\mathbf{v}) &= \frac{\partial}{\partial r}\left(\frac{\partial v_r}{\partial r} + \frac{1}{r}\frac{\partial v_\theta}{\partial \theta} + \frac{v_r}{r}\right)\mathbf{e}_\mathrm{r} + \frac{1}{r}\frac{\partial}{\partial \theta}\left(\frac{\partial v_r}{\partial r} + \frac{1}{r}\frac{\partial v_\theta}{\partial \theta} + \frac{v_r}{r}\right)\mathbf{e}_\theta \\ &= \left(\frac{\partial^2 v_r}{\partial r^2} + \frac{1}{r}\frac{\partial^2 v_\theta}{\partial r \partial \theta} - \frac{1}{r^2}\frac{\partial v_\theta}{\partial \theta} + \frac{1}{r}\frac{\partial v_r}{\partial r} - \frac{v_r}{r^2}\right)\mathbf{e}_\mathrm{r} + \left(\frac{1}{r}\frac{\partial^2 v_r}{\partial \theta \partial r} + \frac{1}{r^2}\frac{\partial^2 v_\theta}{\partial \theta^2} + \frac{1}{r^2}\frac{\partial v_r}{\partial \theta}\right)\mathbf{e}_\theta, \end{aligned} \tag{2.33.36}$$

and

$$\mathrm{curl}\,\mathbf{v} = \left(\frac{\partial v_\theta}{\partial r} + \frac{v_\theta}{r} - \frac{1}{r}\frac{\partial v_r}{\partial \theta}\right)\mathbf{e}_z. \tag{2.33.37}$$

Since [see Eq. (2.34.7)]

$$\mathrm{curl}\,\mathbf{v} = \left(\frac{1}{r}\frac{\partial v_z}{\partial \theta} - \frac{\partial v_\theta}{\partial z}\right)\mathbf{e}_r + \left(\frac{\partial v_r}{\partial z} - \frac{\partial v_z}{\partial r}\right)\mathbf{e}_\theta + \left(\frac{\partial v_\theta}{\partial r} + \frac{\mathrm{v}_\theta}{r} - \frac{1}{r}\frac{\partial v_r}{\partial \theta}\right)\mathbf{e}_z,$$

therefore,

$$(\mathrm{curl}\,\mathrm{curl}\,\mathbf{v})_r = \frac{1}{r}\frac{\partial}{\partial \theta}\left(\frac{\partial v_\theta}{\partial r} + \frac{v_\theta}{r} - \frac{1}{r}\frac{\partial v_r}{\partial \theta}\right) = \left(\frac{1}{r}\frac{\partial^2 v_\theta}{\partial \theta \partial r} + \frac{1}{r^2}\frac{\partial v_\theta}{\partial \theta} - \frac{1}{r^2}\frac{\partial^2 v_r}{\partial \theta^2}\right), \tag{2.33.38}$$

$$(\text{curl curl }\mathbf{v})_\theta = -\frac{\partial}{\partial r}\left(\frac{\partial v_\theta}{\partial r} + \frac{v_\theta}{r} - \frac{1}{r}\frac{\partial v_r}{\partial \theta}\right) = \left(-\frac{\partial^2 v_\theta}{\partial r^2} - \frac{1}{r}\frac{\partial v_\theta}{\partial r} + \frac{v_\theta}{r^2} + \frac{1}{r}\frac{\partial^2 v_r}{\partial r \partial \theta} - \frac{1}{r^2}\frac{\partial v_r}{\partial \theta}\right). \tag{2.33.39}$$

Thus,

$$(\nabla^2 \mathbf{v})_r = \frac{\partial^2 v_r}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 v_r}{\partial \theta^2} + \frac{\partial^2 v_r}{\partial z^2} + \frac{1}{r}\frac{\partial v_r}{\partial r} - \frac{2}{r^2}\frac{\partial v_\theta}{\partial \theta} - \frac{v_r}{r^2}, \tag{2.33.40}$$

and

$$(\nabla^2 \mathbf{v})_\theta = \frac{\partial^2 v_\theta}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 v_\theta}{\partial \theta^2} + \frac{1}{r}\frac{\partial v_\theta}{\partial r} + \frac{2}{r^2}\frac{\partial v_r}{\partial \theta} - \frac{v_\theta}{r^2}. \tag{2.33.41}$$

2.34 CYLINDRICAL COORDINATES

In cylindrical coordinates, the position of a point P is determined by (r, θ, z), where r and θ determine the position of the vertical projection of the point P on the xy plane (the point P' in Figure 2.34-1) and the coordinate z determines the height of the point P from the xy plane. In other words, the cylindrical coordinates is a polar coordinate (r, θ) in the xy plane plus a coordinate z perpendicular to the xy plane.

We shall denote the position vector of P by $\mathbf{R}$, rather than $\mathbf{r}$, to avoid confusion between the position vector $\mathbf{R}$ and the coordinate r (which is a radial distance in the xy plane). The unit vector $\mathbf{e}_r$ and $\mathbf{e}_\theta$ are on the xy plane and it is clear from Figure 2.34-1 that

$$\mathbf{R} = r\mathbf{e}_r + z\mathbf{e}_z, \tag{2.34.1}$$

and

$$d\mathbf{R} = dr\mathbf{e}_r + rd\mathbf{e}_r + dz\mathbf{e}_z + zd\mathbf{e}_z. \tag{2.34.2}$$

In the preceding equation, $d\mathbf{e}_r$ is given by exactly the same equation given earlier for the polar coordinates [Eq. (2.33.3)]. We note also that $\mathbf{e}_z$ never changes its direction or magnitude regardless where the point P is, thus $d\mathbf{e}_z = 0$. Therefore,

$$d\mathbf{R} = dr\mathbf{e}_r + rd\theta\,\mathbf{e}_\theta + dz\mathbf{e}_z. \tag{2.34.3}$$

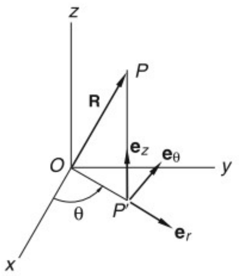

FIGURE 2.34-1

By retracing all the steps used in the previous section on polar coordinates, we can easily obtain the following results:

(i) Components of ∇f:

$$\nabla f = \frac{\partial f}{\partial r}\mathbf{e}_r + \frac{1}{r}\frac{\partial f}{\partial \theta}\mathbf{e}_\theta + \frac{\partial f}{\partial z}\mathbf{e}_z. \tag{2.34.4}$$

(ii) Components of $\nabla \mathbf{v}$:

$$[\nabla \mathbf{v}] = \begin{bmatrix} \dfrac{\partial v_r}{\partial r} & \dfrac{1}{r}\left(\dfrac{\partial v_r}{\partial \theta} - \mathrm{v}_\theta\right) & \dfrac{\partial v_r}{\partial z} \\[2ex] \dfrac{\partial v_\theta}{\partial r} & \dfrac{1}{r}\left(\dfrac{\partial v_\theta}{\partial \theta} + v_r\right) & \dfrac{\partial v_\theta}{\partial z} \\[2ex] \dfrac{\partial v_z}{\partial r} & \dfrac{1}{r}\dfrac{\partial v_z}{\partial \theta} & \dfrac{\partial v_z}{\partial z} \end{bmatrix}. \tag{2.34.5}$$

(iii) div $\mathbf{v}$:

$$\operatorname{div} \mathbf{v} = \frac{\partial v_r}{\partial r} + \frac{1}{r}\left(\frac{\partial v_\theta}{\partial \theta} + v_r\right) + \frac{\partial v_z}{\partial z}. \tag{2.34.6}$$

(iv) Components of curl $\mathbf{v}$:
The vector curl $\mathbf{v}$ = twice the dual vector of $(\nabla \mathbf{v})^{\mathrm{A}}$, thus,

$$\operatorname{curl} \mathbf{v} = \left(\frac{1}{r}\frac{\partial v_z}{\partial \theta} - \frac{\partial v_\theta}{\partial z}\right)\mathbf{e}_r + \left(\frac{\partial v_r}{\partial z} - \frac{\partial v_z}{\partial r}\right)\mathbf{e}_\theta + \left(\frac{\partial v_\theta}{\partial r} + \frac{\mathrm{v}_\theta}{r} - \frac{1}{r}\frac{\partial v_r}{\partial \theta}\right)\mathbf{e}_z. \tag{2.34.7}$$

(v) Components of div $\mathbf{T}$:

$$(\operatorname{div} \mathbf{T})_r = \frac{\partial T_{rr}}{\partial r} + \frac{1}{r}\frac{\partial T_{r\theta}}{\partial \theta} + \frac{T_{rr} - T_{\theta\theta}}{r} + \frac{\partial T_{rz}}{\partial z}, \tag{2.34.8}$$

$$(\operatorname{div} \mathbf{T})_\theta = \frac{\partial T_{\theta r}}{\partial r} + \frac{1}{r}\frac{\partial T_{\theta\theta}}{\partial \theta} + \frac{T_{r\theta} + T_{\theta r}}{r} + \frac{\partial T_{\theta z}}{\partial z}, \tag{2.34.9}$$

$$(\operatorname{div} \mathbf{T})_z = \frac{\partial T_{zr}}{\partial r} + \frac{1}{r}\frac{\partial T_{z\theta}}{\partial \theta} + \frac{\partial T_{zz}}{\partial z} + \frac{T_{zr}}{r}. \tag{2.34.10}$$

(vi) Laplacian of f:

$$\nabla^2 f = \operatorname{div} \nabla f = \frac{\partial^2 f}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 f}{\partial \theta^2} + \frac{1}{r}\frac{\partial f}{\partial r} + \frac{\partial^2 f}{\partial z^2}. \tag{2.34.11}$$

(vii) Laplacian of $\mathbf{v}$:

$$(\nabla^2 \mathbf{v})_r = \frac{\partial^2 v_r}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 v_r}{\partial \theta^2} + \frac{\partial^2 v_r}{\partial z^2} + \frac{1}{r}\frac{\partial v_r}{\partial r} - \frac{v_r}{r^2} - \frac{2}{r^2}\frac{\partial v_\theta}{\partial \theta}, \tag{2.34.12}$$

$$(\nabla^2\mathbf{v})_\theta = \frac{\partial^2 v_\theta}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 v_\theta}{\partial \theta^2} + \frac{\partial^2 v_\theta}{\partial z^2} + \frac{1}{r}\frac{\partial v_\theta}{\partial r} + \frac{2}{r^2}\frac{\partial v_r}{\partial \theta} - \frac{v_\theta}{r^2}, \tag{2.34.13}$$

$$(\nabla^2\mathbf{v})_z = \frac{\partial^2 v_z}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 v_z}{\partial \theta^2} + \frac{1}{r}\frac{\partial v_z}{\partial r} + \frac{\partial^2 v_z}{\partial z^2}. \tag{2.34.14}$$

2.35 SPHERICAL COORDINATES

In Figure 2.35-1, we show the spherical coordinates (r, θ, ϕ) of a general point P. In this figure, $\mathbf{e}_r$, $\mathbf{e}_\theta$ and $\mathbf{e}_\phi$ are unit vectors in the direction of increasing r, θ and ϕ, respectively.

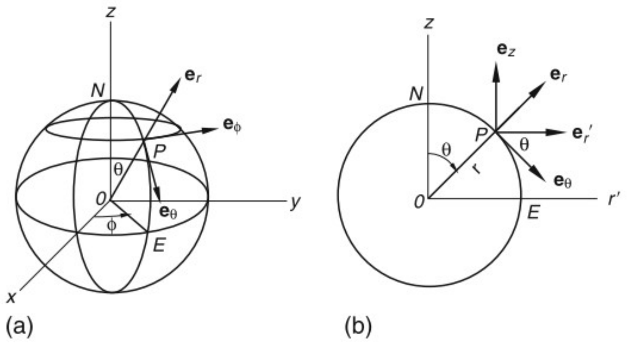

FIGURE 2.35-1

The position vector for the point P can be written as

$$\mathbf{r} = r\mathbf{e}_r, \tag{2.35.1}$$

where r is the magnitude of the vector $\mathbf{r}$. Thus,

$$d\mathbf{r} = dr\mathbf{e}_r + rd\mathbf{e}_r. \tag{2.35.2}$$

To evaluate $d\mathbf{e}_r$ we note from Figure 2.35-1(b) that

$$\mathbf{e}_r = \cos\theta\mathbf{e}_z + \sin\theta\mathbf{e}_r', \quad \mathbf{e}_\theta = \cos\theta\mathbf{e}_r' - \sin\theta\mathbf{e}_z, \tag{2.35.3}$$

where $\mathbf{e}_r'$ is the unit vector in the OE (i.e., r') direction (r' is in the xy plane). Thus,

$$d\mathbf{e}_r = -\sin\theta d\theta\mathbf{e}_z + \cos\theta d\mathbf{e}_z + \cos\theta d\theta\mathbf{e}_r' + \sin\theta d\mathbf{e}_r' = \left(-\sin\theta\mathbf{e}_z + \cos\theta\mathbf{e}_r'\right)d\theta + \sin\theta d\mathbf{e}_r',$$

that is,

$$d\mathbf{e}_r = d\theta\mathbf{e}_\theta + \sin\theta d\mathbf{e}_r'. \tag{2.35.4}$$

Now, just as in polar coordinates, due to $d\phi$,

$$d\mathbf{e}_r' = d\phi\mathbf{e}_\phi, \tag{2.35.5}$$

therefore,

$$d\mathbf{e}_r = d\theta\,\mathbf{e}_\theta + \sin\theta\, d\phi\,\mathbf{e}_\phi. \tag{2.35.6}$$

Now, from the second equation of (2.35.3), we have,

$$d\mathbf{e}_\theta = -\sin\theta\, d\theta\,\mathbf{e}_r' + \cos\theta\, d\mathbf{e}_r' - \cos\theta\, d\theta\,\mathbf{e}_z = -\left(\sin\theta\,\mathbf{e}_r' + \cos\theta\,\mathbf{e}_z\right)d\theta + \cos\theta\, d\mathbf{e}_r'.$$

Using Eq. (2.35.3) and Eq. (2.35.5), the preceding equation becomes

$$d\mathbf{e}_\theta = -\mathbf{e}_r d\theta + \cos\theta\, d\phi\,\mathbf{e}_\phi. \tag{2.35.7}$$

From Figure 2.35-1(a) and similar to the polar coordinate, we have

$$d\mathbf{e}_\phi = d\phi\left(-\mathbf{e}_r'\right). \tag{2.35.8}$$

With $\mathbf{e}_r' = \cos\theta\,\mathbf{e}_\theta + \sin\theta\,\mathbf{e}_r$ (see Figure 2.35-1(b)), the preceding equation becomes

$$d\mathbf{e}_\phi = -\sin\theta\, d\phi\,\mathbf{e}_r - \cos\theta\, d\phi\,\mathbf{e}_\theta. \tag{2.35.9}$$

Summarizing the preceding, we have

$$d\mathbf{e}_r = d\theta\,\mathbf{e}_\theta + \sin\theta\, d\phi\,\mathbf{e}_\phi, \quad d\mathbf{e}_\theta = -\mathbf{e}_r d\theta + \cos\theta\, d\phi\,\mathbf{e}_\phi, \quad d\mathbf{e}_\phi = -\sin\theta\, d\phi\,\mathbf{e}_r - \cos\theta\, d\phi\,\mathbf{e}_\theta, \tag{2.35.10}$$

and from Eq. (2.35.2), we have

$$d\mathbf{r} = dr\,\mathbf{e}_r + r\,d\theta\,\mathbf{e}_\theta + r\,\sin\theta\, d\phi\,\mathbf{e}_\phi. \tag{2.35.11}$$

We can now obtain the components of ∇f, $\nabla\mathbf{v}$, div $\mathbf{v}$, curl $\mathbf{v}$, div $\mathbf{T}$, $\nabla^2 f$, and $\nabla^2\mathbf{v}$ for spherical coordinates.

(i) Components of ∇f:

Let $f(r,\theta,\phi)$ be a scalar field. By the definition of ∇f, we have

$$df = \nabla f\cdot d\mathbf{r} = \left[(\nabla f)_r\mathbf{e}_r + (\nabla f)_\theta\mathbf{e}_\theta + (\nabla f)_\phi\mathbf{e}_\phi\right]\cdot\left(dr\,\mathbf{e}_r + r\,d\theta\,\mathbf{e}_\theta + r\sin\theta\, d\phi\,\mathbf{e}_\phi\right), \tag{2.35.12}$$

that is,

$$df = (\nabla f)_r\, dr + (\nabla f)_\theta\, r\,d\theta + (\nabla f)_\phi\, r\sin\theta\, d\phi. \tag{2.35.13}$$

From calculus, the total derivative of df is

$$df = \frac{\partial f}{\partial r}dr + \frac{\partial f}{\partial\theta}d\theta + \frac{\partial f}{\partial\phi}d\phi. \tag{2.35.14}$$

Comparing Eq. (2.35.14) and Eq. (2.35.13), we have

$$(\nabla f)_r = \frac{\partial f}{\partial r}, \quad (\nabla f)_\theta = \frac{1}{r}\frac{\partial f}{\partial\theta}, \quad (\nabla f)_\phi = \frac{1}{r\sin\theta}\frac{\partial f}{\partial\phi}. \tag{2.35.15}$$

(ii) Components of $\nabla\mathbf{v}$:

Let the vector field be represented by

$$\mathbf{v}(r,\theta,\phi) = v_r(r,\theta,\phi)\mathbf{e}_r + v_\theta(r,\theta,\phi)\mathbf{e}_\theta + v_\phi(r,\theta,\phi)\mathbf{e}_\phi. \tag{2.35.16}$$

Letting $\mathbf{T} = \nabla\mathbf{v}$, we have

$$d\mathbf{v} = \mathbf{T}d\mathbf{r} = \mathbf{T}\left(dr\mathbf{e}_r + rd\theta\mathbf{e}_\theta + r\sin\theta d\phi\mathbf{e}_\phi\right) = dr\mathbf{T}\mathbf{e}_r + rd\theta\mathbf{T}\mathbf{e}_\theta + r\sin\theta d\phi\mathbf{T}\mathbf{e}_\phi. \tag{2.35.17}$$

By the definition of components of a tensor T in spherical coordinates, we have

$$\begin{aligned}\mathbf{T}\mathbf{e}_r &= T_{rr}\mathbf{e}_r + T_{\theta r}\mathbf{e}_\theta + T_{\phi r}\mathbf{e}_\phi,\\ \mathbf{T}\mathbf{e}_\theta &= T_{r\theta}\mathbf{e}_r + T_{\theta\theta}\mathbf{e}_\theta + T_{\phi\theta}\mathbf{e}_\phi,\\ \mathbf{T}\mathbf{e}_\phi &= T_{r\phi}\mathbf{e}_r + T_{\theta\phi}\mathbf{e}_\theta + T_{\phi\phi}\mathbf{e}_\phi.\end{aligned} \tag{2.35.18}$$

Substituting these into Eq. (2.35.17), we get

$$\begin{aligned}d\mathbf{v} = \left(T_{rr}dr + T_{r\theta}rd\theta + T_{r\phi}r\ \sin\theta d\phi\right)\mathbf{e}_r + \left(T_{\theta\theta}rd\theta + T_{\theta r}dr + T_{\theta\phi}r\ \sin\theta d\phi\right)\mathbf{e}_\theta\\ +\left(T_{\phi r}dr + T_{\phi\theta}rd\theta + T_{\phi\phi}r\ \sin\theta d\phi\right)\mathbf{e}_\phi.\end{aligned} \tag{2.35.19}$$

We also have, from Eq. (2.35.16),

$$d\mathbf{v} = dv_r\mathbf{e}_r + v_r d\mathbf{e}_r + dv_\theta\mathbf{e}_\theta + v_\theta d\mathbf{e}_\theta + dv_\phi\mathbf{e}_\phi + v_\phi d\mathbf{e}_\phi. \tag{2.35.20}$$

Using the expression for the total derivatives:

$$\begin{aligned}dv_r &= \frac{\partial v_r}{\partial r}dr + \frac{\partial v_r}{\partial\theta}d\theta + \frac{\partial v_r}{\partial\phi}d\phi,\\ dv_\theta &= \frac{\partial v_\theta}{\partial r}dr + \frac{\partial v_\theta}{\partial\theta}d\theta + \frac{\partial v_\theta}{\partial\phi}d\phi,\\ dv_\phi &= \frac{\partial v_\phi}{\partial r}dr + \frac{\partial v_\phi}{\partial\theta}d\theta + \frac{\partial v_\phi}{\partial\phi}d\phi,\end{aligned} \tag{2.35.21}$$

Eq. (2.35.10) and Eq. (2.35.20) become

$$\begin{aligned}d\mathbf{v} = &\left\{\frac{\partial v_r}{\partial r}dr + \left(\frac{\partial v_r}{\partial\theta} - v_\theta\right)d\theta + \left(\frac{\partial v_r}{\partial\phi} - v_\phi\sin\theta\right)d\phi\right\}\mathbf{e}_r\\ &+\left\{\frac{\partial v_\theta}{\partial r}dr + \left(v_r + \frac{\partial v_\theta}{\partial\theta}\right)d\theta + \left(\frac{\partial v_\theta}{\partial\phi} - v_\phi\ \cos\theta\right)d\phi\right\}\mathbf{e}_\theta\\ &+\left\{\frac{\partial v_\phi}{\partial r}dr + \frac{\partial v_\phi}{\partial\theta}d\theta + \left(\frac{\partial v_\phi}{\partial\phi} + v_r\ \sin\theta + v_\theta\ \cos\theta\right)d\phi\right\}\mathbf{e}_\phi.\end{aligned} \tag{2.35.22}$$

Now, comparing Eq. (2.35.22) with Eq. (2.35.19), we have

$$\begin{aligned}\left(T_{rr}dr + T_{r\theta}rd\theta + T_{r\phi}r\ \sin\theta d\phi\right) &= \left\{\frac{\partial v_r}{\partial r}dr + \left(\frac{\partial v_r}{\partial\theta} - v_\theta\right)d\theta + \left(\frac{\partial v_r}{\partial\phi} - v_\phi\ \sin\theta\right)d\phi\right\},\\ \left(T_{\theta r}dr + T_{\theta\theta}rd\theta + T_{\theta\phi}r\ \sin\theta d\phi\right) &= \left\{\frac{\partial v_\theta}{\partial r}dr + \left(v_r + \frac{\partial v_\theta}{\partial\theta}\right)d\theta + \left(\frac{\partial v_\theta}{\partial\phi} - v_\phi\ \cos\theta\right)d\phi\right\},\\ \left(T_{\phi r}dr + T_{\phi\theta}rd\theta + T_{\phi\phi}r\ \sin\theta d\phi\right) &= \left\{\frac{\partial v_\phi}{\partial r}dr + \frac{\partial v_\phi}{\partial\theta}d\theta + \left(\frac{\partial v_\phi}{\partial\phi} + v_r\ \sin\theta + v_\theta\ \cos\theta\right)d\phi\right\}.\end{aligned} \tag{2.35.23}$$

These equations must be valid for arbitrary values of dr, $d\theta$ and $d\phi$, therefore,

$$\begin{aligned}
&T_{rr}=\frac{\partial v_r}{\partial r},\quad T_{r\theta}r=\left(\frac{\partial v_r}{\partial\theta}-v_\theta\right),\quad T_{r\phi}r\sin\theta=\left(\frac{\partial v_r}{\partial\phi}-v_\phi\sin\theta\right),\\
&T_{\theta r}=\frac{\partial v_\theta}{\partial r},\quad T_{\theta\theta}r=\left(v_r+\frac{\partial v_\theta}{\partial\theta}\right),\quad T_{\theta\phi}r\sin\theta=\left(\frac{\partial v_\theta}{\partial\phi}-v_\phi\cos\theta\right),\\
&T_{\phi r}=\frac{\partial v_\phi}{\partial r},\quad T_{\phi\theta}r=\frac{\partial v_\phi}{\partial\theta},\quad T_{\phi\phi}r\sin\theta=\left(\frac{\partial v_\phi}{\partial\phi}+v_r\sin\theta+v_\theta\cos\theta\right).
\end{aligned}\tag{2.35.24}$$

In matrix form, we have

$$[\nabla\mathbf{v}]=\begin{bmatrix}
\dfrac{\partial v_r}{\partial r} & \dfrac{1}{r}\dfrac{\partial v_r}{\partial\theta}-\dfrac{v_\theta}{r} & \dfrac{1}{r\sin\theta}\dfrac{\partial v_r}{\partial\phi}-\dfrac{v_\phi}{r}\\
\dfrac{\partial v_\theta}{\partial r} & \dfrac{1}{r}\dfrac{\partial v_\theta}{\partial\theta}+\dfrac{v_r}{r} & \dfrac{1}{r\sin\theta}\dfrac{\partial v_\theta}{\partial\phi}-\dfrac{v_\phi\cot\theta}{r}\\
\dfrac{\partial v_\phi}{\partial r} & \dfrac{1}{r}\dfrac{\partial v_\phi}{\partial\theta} & \dfrac{1}{r\sin\theta}\dfrac{\partial v_\phi}{\partial\phi}+\dfrac{v_r}{r}+\dfrac{v_\theta\cot\theta}{r}
\end{bmatrix}.\tag{2.35.25}$$

(iii) div **v**:
Using Eq. (2.35.25), we obtain

$$\begin{aligned}
\operatorname{div}\mathbf{v}=\operatorname{tr}(\nabla\mathbf{v})&=\frac{\partial v_r}{\partial r}+\frac{1}{r}\frac{\partial v_\theta}{\partial\theta}+\frac{1}{r\sin\theta}\frac{\partial v_\phi}{\partial\phi}+\frac{2v_r}{r}+\frac{v_\theta\cot\theta}{r}\\
&=\frac{1}{r^2}\frac{\partial(r^2v_r)}{\partial r}+\frac{1}{r\sin\theta}\frac{\partial(v_\theta\sin\theta)}{\partial\theta}+\frac{1}{r\sin\theta}\frac{\partial v_\phi}{\partial\phi}.
\end{aligned}\tag{2.35.26}$$

(iv) Components of curl **v**:
The vector curl **v** = twice the dual vector of $(\nabla\mathbf{v})^{\mathrm{A}}$, therefore

$$\begin{aligned}
\operatorname{curl}\mathbf{v}=&\left\{\frac{v_\phi\cot\theta}{r}+\frac{1}{r}\frac{\partial v_\phi}{\partial\theta}-\frac{1}{r\sin\theta}\frac{\partial v_\theta}{\partial\phi}\right\}\mathbf{e}_r+\left\{\frac{1}{r\sin\theta}\frac{\partial v_r}{\partial\phi}-\frac{1}{r}\frac{\partial(rv_\phi)}{\partial r}\right\}\mathbf{e}_\theta\\
&+\left\{\frac{1}{r}\frac{\partial(rv_\theta)}{\partial r}-\frac{1}{r}\frac{\partial v_r}{\partial\theta}\right\}\mathbf{e}_\phi.
\end{aligned}\tag{2.35.27}$$

(v) Components of div **T**:
Using the definition of div **T** given in Eq. (2.33.27) and take $\mathbf{a}=\mathbf{e}_r$, we have

$$(\operatorname{div}\mathbf{T})_r=\operatorname{div}(\mathbf{T}^{\mathrm{T}}\mathbf{e}_r)-\operatorname{tr}((\nabla\mathbf{e}_r)\mathbf{T}^{\mathrm{T}}).\tag{2.35.28}$$

To evaluate the first term on the right-hand side, we note that

$$\mathbf{T}^{\mathrm{T}}\mathbf{e}_r=T_{rr}\mathbf{e}_r+T_{r\theta}\mathbf{e}_\theta+T_{r\phi}\mathbf{e}_\phi,\tag{2.35.29}$$

so that by using Eq. (2.35.26) for the divergence of a vector in spherical coordinates, we obtain,

$$\operatorname{div}(\mathbf{T}^{\mathrm{T}}\mathbf{e}_r) = \frac{1}{r^2}\frac{\partial(r^2 T_{rr})}{\partial r} + \frac{1}{r\sin\theta}\frac{\partial(T_{r\theta}\sin\theta)}{\partial\theta} + \frac{1}{r\sin\theta}\frac{\partial T_{r\phi}}{\partial\phi}. \tag{2.35.30}$$

To evaluate the second term in Eq. (2.35.28), we first used Eq. (2.35.25) to evaluate $\nabla\mathbf{e}_r$, then calculate $(\nabla\mathbf{e}_r)\mathbf{T}^{\mathrm{T}}$:

$$[\nabla\mathbf{e}_r] = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1/r & 0 \\ 0 & 0 & 1/r \end{bmatrix},\ [(\nabla\mathbf{e}_r)\mathbf{T}^{\mathrm{T}}] = \begin{bmatrix} 0 & 0 & 0 \\ T_{r\theta}/r & T_{\theta\theta}/r & T_{\phi\theta}/r \\ T_{r\phi}/r & T_{\theta\phi}/r & T_{\phi\phi}/r \end{bmatrix} \tag{2.35.31}$$

thus,

$$\operatorname{tr}((\nabla\mathbf{e}_r)\mathbf{T}^{\mathrm{T}}) = \frac{T_{\theta\theta}}{r} + \frac{T_{\phi\phi}}{r}. \tag{2.35.32}$$

Substituting Eq. (2.35.32) and Eq. (2.35.30) into Eq. (2.35.28), we obtain,

$$(\operatorname{div}\mathbf{T})_r = \frac{1}{r^2}\frac{\partial(r^2 T_{rr})}{\partial r} + \frac{1}{r\sin\theta}\frac{\partial(T_{r\theta}\sin\theta)}{\partial\theta} + \frac{1}{r\sin\theta}\frac{\partial T_{r\phi}}{\partial\phi} - \frac{T_{\theta\theta}+T_{\phi\phi}}{r}. \tag{2.35.33}$$

In a similar manner, we can obtain (see Prob. 2.75)

$$(\operatorname{div}\mathbf{T})_\theta = \frac{1}{r^3}\frac{\partial(r^3 T_{\theta r})}{\partial r} + \frac{1}{r\sin\theta}\frac{\partial(T_{\theta\theta}\sin\theta)}{\partial\theta} + \frac{1}{r\sin\theta}\frac{\partial T_{\theta\phi}}{\partial\phi} + \frac{T_{r\theta}-T_{\theta r}-T_{\phi\phi}\cot\theta}{r} \tag{2.35.34}$$

$$(\operatorname{div}\mathbf{T})_\phi = \frac{1}{r^3}\frac{\partial\left(r^3 T_{\phi r}\right)}{\partial r} + \frac{1}{r\sin\theta}\frac{\partial(T_{\phi\theta}\sin\theta)}{\partial\theta} + \frac{1}{r\sin\theta}\frac{\partial T_{\phi\phi}}{\partial\phi} + \frac{T_{r\phi}-T_{\phi r}+T_{\theta\phi}\cot\theta}{r}. \tag{2.35.35}$$

(vi) Laplacian of f:
From

$$\begin{gathered}\operatorname{div}\mathbf{v} = \frac{1}{r^2}\frac{\partial(r^2 v_r)}{\partial r} + \frac{1}{r\sin\theta}\frac{\partial v_\theta\sin\theta}{\partial\theta} + \frac{1}{r\sin\theta}\frac{\partial v_\phi}{\partial\phi}, \\ \nabla f = \frac{\partial f}{\partial r}\mathbf{e}_r + \frac{1}{r}\frac{\partial f}{\partial\theta}\mathbf{e}_\theta + \frac{1}{r\sin\theta}\frac{\partial f}{\partial\phi}\mathbf{e}_\phi,\end{gathered} \tag{2.35.36}$$

we have

$$\begin{aligned}\nabla^2 f = \operatorname{div}(\nabla f) &= \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial f}{\partial r}\right) + \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}\left(\frac{1}{r}\frac{\partial f}{\partial\theta}\sin\theta\right) + \frac{1}{r\sin\theta}\frac{\partial}{\partial\phi}\left(\frac{1}{r\sin\theta}\frac{\partial f}{\partial\phi}\right) \\ &= \frac{\partial^2 f}{\partial r^2} + \frac{2}{r}\frac{\partial f}{\partial r} + \frac{1}{r^2}\left(\frac{\partial^2 f}{\partial\theta^2}\right) + \frac{\cot\theta}{r^2}\frac{\partial f}{\partial\theta} + \frac{1}{r^2\sin^2\theta}\left(\frac{\partial^2 f}{\partial\phi^2}\right).\end{aligned} \tag{2.35.37}$$

(vii) Laplacian of a vector function $\mathbf{v}$:
It can be obtained (see Prob. 2.75)

$$\begin{aligned}\nabla(\text{div } \mathbf{v}) = &\left(\frac{1}{r^2}\frac{\partial^2 r^2 v_r}{\partial r^2} - \frac{2}{r^3}\frac{\partial r^2 v_r}{\partial r} + \frac{1}{r\sin\theta}\left(\frac{\partial^2 v_\theta \sin\theta}{\partial r \partial\theta} + \frac{\partial^2 v_\phi}{\partial r \partial\phi}\right) - \frac{1}{r^2\sin\theta}\left(\frac{\partial v_\theta \sin\theta}{\partial\theta} + \frac{\partial v_\phi}{\partial\phi}\right)\right)\mathbf{e}_r \\ &+\left(\frac{1}{r^3}\frac{\partial^2 r^2 v_r}{\partial\theta\partial r} + \frac{1}{r^2\sin\theta}\left(\frac{\partial^2 v_\theta \sin\theta}{\partial\theta^2} + \frac{\partial^2 v_\theta \sin\theta}{\partial\theta^2}\right) - \frac{1}{r^2}\left(\frac{\cos\theta}{\sin^2\theta}\right)\frac{\partial v_\theta \sin\theta}{\partial\theta} + \frac{1}{r^2}\frac{\partial}{\partial\theta}\frac{1}{\sin\theta}\frac{\partial v_\phi}{\partial\phi}\right)\mathbf{e}_\theta \\ &+\left(\frac{1}{r^3\sin\theta}\frac{\partial}{\partial\phi}\frac{\partial(r^2 v_r)}{\partial r} + \frac{1}{r^2\sin^2\theta}\frac{\partial^2(v_\theta\sin\theta)}{\partial\phi\partial\theta} + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 v_\phi}{\partial^2\phi}\right)\mathbf{e}_\phi,\end{aligned} \tag{2.35.38}$$

and

$$\begin{aligned}\text{curl curl } \mathbf{v} = &\left\{\frac{1}{r^2}\left(\frac{\partial^2 r v_\theta}{\partial\theta\partial r} - \frac{\partial^2 v_r}{\partial\theta^2}\right) + \frac{\cot\theta}{r}\left(\frac{1}{r}\frac{\partial r v_\theta}{\partial r} - \frac{1}{r}\frac{\partial v_r}{\partial\theta}\right) - \left(\frac{1}{r^2\sin^2\theta}\frac{\partial^2 v_r}{\partial\phi^2} - \frac{1}{r^2\sin\theta}\frac{\partial^2 r v_\phi}{\partial\phi\partial r}\right)\right\}\mathbf{e}_r \\ &+\left\{\frac{1}{r^2\sin^2\theta}\left(\frac{\partial^2 v_\phi\sin\theta}{\partial\phi\partial\theta} - \frac{\partial^2 v_\theta}{\partial\phi^2}\right) - \frac{1}{r^2}\left(\frac{\partial r v_\theta}{\partial r} - \frac{\partial v_r}{\partial\theta}\right) - \left(\frac{1}{r}\frac{\partial^2 r v_\theta}{\partial r^2} - \frac{1}{r^2}\frac{\partial r v_\theta}{\partial r} - \frac{1}{r}\frac{\partial^2 v_r}{\partial r\partial\theta} + \frac{1}{r^2}\frac{\partial v_r}{\partial\theta}\right)\right\}\mathbf{e}_\theta \\ &+\left\{\begin{array}{l}\left(\dfrac{1}{r\sin\theta}\dfrac{\partial v_r}{\partial r\partial\phi} - \dfrac{1}{r^2}\dfrac{1}{\sin\theta}\dfrac{\partial v_r}{\partial\phi} - \dfrac{1}{r}\dfrac{\partial^2 r v_\phi}{\partial r^2} + \dfrac{1}{r^2}\dfrac{\partial r v_\phi}{\partial r}\right) + \dfrac{1}{r^2}\left(\dfrac{1}{\sin\theta}\dfrac{\partial v_r}{\partial\phi} - \dfrac{\partial r v_\phi}{\partial r}\right) \\ -\dfrac{1}{r^2}\dfrac{1}{\sin\theta}\left(-v_\phi\sin\theta + \sin\theta\dfrac{\partial^2 v_\phi}{\partial\theta^2} - \dfrac{\partial^2 v_\theta}{\partial\theta\partial\phi}\right) + \dfrac{\cos\theta}{r^2\sin^2\theta}\left(\dfrac{\partial v_\phi\sin\theta}{\partial\theta} - \dfrac{\partial v_\theta}{\partial\phi}\right)\end{array}\right\}\mathbf{e}_\phi.\end{aligned} \tag{2.35.39}$$

Thus, $\nabla^2\mathbf{v} = \nabla(\text{div } \mathbf{v}) - \text{curl curl } \mathbf{v}$ leads to:

$$(\nabla^2\mathbf{v})_r = \left(\begin{array}{l}\dfrac{1}{r^2}\dfrac{\partial^2 r^2 v_r}{\partial r^2} - \dfrac{2}{r^3}\dfrac{\partial r^2 v_r}{\partial r} + \dfrac{1}{r^2}\dfrac{\partial^2 v_r}{\partial\theta^2} + \dfrac{\cot\theta}{r^2}\dfrac{\partial v_r}{\partial\theta} + \dfrac{1}{r^2\sin^2\theta}\dfrac{\partial^2 v_r}{\partial\phi^2} - \dfrac{2}{r^2\sin\theta}\dfrac{\partial v_\theta\sin\theta}{\partial\theta} \\ -\dfrac{2}{r^2\sin\theta}\dfrac{\partial v_\phi}{\partial\phi}\end{array}\right), \tag{2.35.40}$$

$$(\nabla^2\mathbf{v})_\theta = \left(\begin{array}{l}\dfrac{1}{r^2}\dfrac{\partial}{\partial r}\left(r^2\dfrac{\partial v_\theta}{\partial r}\right) + \dfrac{1}{r^2}\dfrac{\partial}{\partial\theta}\left\{\dfrac{1}{\sin\theta}\dfrac{\partial}{\partial\theta}(v_\theta\sin\theta)\right\} + \dfrac{1}{r^2\sin^2\theta}\dfrac{\partial^2 v_\theta}{\partial\phi^2} \\ +\dfrac{2}{r^2}\dfrac{\partial v_r}{\partial\theta} - \dfrac{2}{r^2}\dfrac{\cot\theta}{\sin\theta}\dfrac{\partial v_\phi}{\partial\phi}\end{array}\right), \tag{2.35.41}$$

$$(\nabla^2\mathbf{v})_\phi = \left(\begin{array}{l}\dfrac{1}{r^2}\dfrac{\partial}{\partial r}\left(r^2\dfrac{\partial v_\phi}{\partial r}\right) + \dfrac{1}{r^2}\dfrac{\partial}{\partial\theta}\left\{\dfrac{1}{\sin\theta}\dfrac{\partial}{\partial\theta}(v_\phi\sin\theta)\right\} + \dfrac{1}{r^2\sin^2\theta}\dfrac{\partial^2 v_\phi}{\partial\phi^2} \\ +\dfrac{2}{r^2\sin\theta}\dfrac{\partial v_r}{\partial\phi} + \dfrac{2}{r^2}\dfrac{\cot\theta}{\sin\theta}\dfrac{\partial v_\theta}{\partial\phi}\end{array}\right). \tag{2.35.42}$$

CHAPTER

Stress and Integral Formulations of General Principles

4

In the previous chapter, we considered the purely kinematic description of the motion of a continuum without any consideration of the forces that cause the motion and deformation. In this chapter, we consider a means of describing the forces in the interior of a body idealized as a continuum. It is generally accepted that matter is formed of molecules, which in turn consist of atoms and subatomic particles. Therefore, the internal forces in real matter are those between these particles. In the classical continuum theory where matter is assumed to be continuously distributed, the forces acting at every point inside a body are introduced through the concept of body forces and surface forces. Body forces are those that act throughout a volume (e.g., gravity, electrostatic force) by a long-range interaction with matter or charges at a distance. Surface forces are those that act on a surface (real or imagined), separating parts of the body. We assume that it is adequate to describe the surface forces at a point on a surface through the definition of a *stress vector*, discussed in Section 4.1, which pays no attention to the curvature of the surface at the point. Such an assumption is known as *Cauchy's stress principle* and is one of the basic axioms of classical continuum mechanics.

4.1 STRESS VECTOR

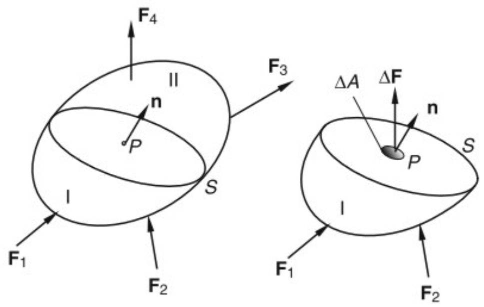

FIGURE 4.1-1

Let us consider a body depicted in Figure 4.1-1. Imagine a plane such as S, which passes through an arbitrary internal point P and which has a unit normal vector $\mathbf{n}$. The plane cuts the body into two portions. One portion

Copyright © 2010, Elsevier Ltd. All rights reserved.

lies on the side of the arrow of **n** (designated by II in the figure) and the other portion on the tail of **n** (designated by I). Considering portion I as a free body, there will be on plane S a resultant force $\Delta\mathbf{F}$ acting on a small area ΔA containing P. We define the stress vector (acting from II to I) at the point P on the plane S as the limit of the ratio $\Delta\mathbf{F}/\Delta A$ as $\Delta A \to 0$. That is, with $\mathbf{t_n}$ denoting the stress vector,

$$\mathbf{t_n} = \lim_{\Delta A \to 0} \frac{\Delta\mathbf{F}}{\Delta A}. \tag{4.1.1}$$

If portion II is considered as a free body, then by Newton's law of action and reaction, we shall have a stress vector (acting from I to II) $\mathbf{t_{-n}}$ at the same point on the same plane equal and opposite to that given by Eq. (4.1.1). That is,

$$\mathbf{t_n} = -\mathbf{t_{-n}}. \tag{4.1.2}$$

The subscript $-\mathbf{n}$ for $\mathbf{t}$ (i.e., $\mathbf{t_{-n}}$) indicates that outward normal for the portion II is in the negative direction of **n**.

Next, let S be a surface (instead of a plane) passing the point P. Let $\Delta\mathbf{F}$ be the resultant force on a small area ΔS on the surface S. The *Cauchy stress vector* at P on S is defined as

$$\mathbf{t} = \lim_{\Delta S \to 0} \frac{\Delta\mathbf{F}}{\Delta S}. \tag{4.1.3}$$

We now state the following principle, known as the *Cauchy's stress principle:* The stress vector at any given place and time has a common value on all parts of material having a common tangent plane at P and lying on the same side of it. In other words, if **n** is the unit outward normal (i.e., a vector of unit length pointing outward, away from the material) to the tangent plane, then

$$\mathbf{t} = \mathbf{t}(\mathbf{x}, t, \mathbf{n}), \tag{4.1.4}$$

where the scalar t denotes time.

In the following section, we show from Newton's second law that this dependence of the Cauchy's stress vector on the outward normal vector **n** can be expressed as

$$\mathbf{t} = \mathbf{T}(\mathbf{x}, t)\mathbf{n}, \tag{4.1.5}$$

where **T** is a linear transformation.

4.2 STRESS TENSOR

According to Eq. (4.1.4), the stress vector on a plane passing through a given spatial point **x** at a given time t depends only on the unit normal vector **n** to the plane. Thus, let **T** be the transformation such that

$$\mathbf{t_n} = \mathbf{Tn}. \tag{4.2.1}$$

We wish to show that this transformation is linear. Let a small tetrahedron be isolated from the body with the point P as one of its vertices (see Figure 4.2-1). The size of the tetrahedron will ultimately be made to approach zero volume so that, in the limit, the inclined plane will pass through the point P. The outward normal to the face PAB is $-\mathbf{e}_1$. Thus, the stress vector on this face is denoted by $\mathbf{t}_{-\mathbf{e}_1}$ and the force on the face is $\mathbf{t}_{-\mathbf{e}_1}\Delta A_1$, where ΔA_1 is the area of PAB. Similarly, the force acting on PBC, PAC and the inclined face ABC are $\mathbf{t}_{-\mathbf{e}_2}\Delta A_2$, $\mathbf{t}_{-\mathbf{e}_3}\Delta A_3$, and $\mathbf{t_n}\,\Delta A_n$, respectively. Thus, from Newton's second law written for the tetrahedron, we have

$$\sum \mathbf{F} = \mathbf{t}_{-\mathbf{e}_1}(\Delta A_1) + \mathbf{t}_{-\mathbf{e}_2}(\Delta A_2) + \mathbf{t}_{-\mathbf{e}_3}(\Delta A_3) + \mathbf{t_n}\Delta A_n = m\mathbf{a}. \tag{4.2.2}$$

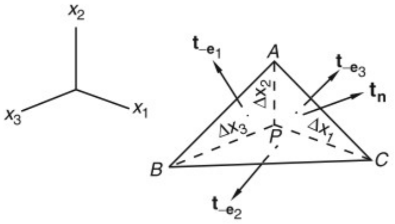
x_2
x_3
x_1
$\mathbf{t}_{-\mathbf{e}_1}$
A
$\mathbf{t}_{-\mathbf{e}_3}$
$\mathbf{t}_{\mathbf{n}}$
Δx_2
Δx_3
P
Δx_1
B
C
$\mathbf{t}_{-\mathbf{e}_2}$

FIGURE 4.2-1

Since the mass m = (density)(volume) and the volume of the tetrahedron is proportional to the product of three infinitesimal lengths (in fact, the volume equals $(1/6)\Delta x_1 \Delta x_2 \Delta x_3$), when the size of the tetrahedron approaches zero, the right-hand side of Eq. (4.2.2) will approach zero faster than the terms on the left, where the stress vectors are multiplied by areas, the product of two infinitesimal lengths. Thus, in the limit, the acceleration term drops out exactly from Eq. (4.2.2). (We note that any body force, e.g., weight, that is acting will be of the same order of magnitude as that of the acceleration term and will also drop out.) Thus,

$$\sum \mathbf{F} = \mathbf{t}_{-\mathbf{e}_1}(\Delta A_1) + \mathbf{t}_{-\mathbf{e}_2}(\Delta A_2) + \mathbf{t}_{-\mathbf{e}_3}(\Delta A_3) + \mathbf{t}_{\mathbf{n}}\Delta A_n = 0. \tag{4.2.3}$$

Let the unit normal vector of the inclined plane *ABC* be

$$\mathbf{n} = n_1\mathbf{e}_1 + n_2\mathbf{e}_2 + n_3\mathbf{e}_3. \tag{4.2.4}$$

The areas ΔA_1, ΔA_2 and ΔA_3, being the projections of ΔA_n on the coordinate planes, are related to ΔA_n by

$$\Delta A_1 = n_1\Delta A_n, \quad \Delta A_2 = n_2\Delta A_n, \quad \Delta A_3 = n_3\Delta A_n. \tag{4.2.5}$$

Using Eq. (4.2.5), Eq. (4.2.3) becomes

$$\mathbf{t}_{-\mathbf{e}_1}n_1 + \mathbf{t}_{-\mathbf{e}_2}n_2 + \mathbf{t}_{-\mathbf{e}_3}n_3 + \mathbf{t}_{\mathbf{n}} = 0. \tag{4.2.6}$$

But from the law of the action and reaction,

$$\mathbf{t}_{-\mathbf{e}_1} = -\mathbf{t}_{\mathbf{e}_1}, \quad \mathbf{t}_{-\mathbf{e}_2} = -\mathbf{t}_{\mathbf{e}_2}, \quad \mathbf{t}_{-\mathbf{e}_3} = -\mathbf{t}_{\mathbf{e}_3}, \tag{4.2.7}$$

therefore, Eq. (4.2.6) becomes

$$\mathbf{t}_{\mathbf{n}} = n_1\mathbf{t}_{\mathbf{e}_1} + n_2\mathbf{t}_{\mathbf{e}_2} + n_3\mathbf{t}_{\mathbf{e}_3}. \tag{4.2.8}$$

Now, using Eq. (4.2.4) and Eq. (4.2.8), Eq. (4.2.1) becomes

$$\mathbf{T}(n_1\mathbf{e}_1 + n_2\mathbf{e}_2 + n_3\mathbf{e}_3) = n_1\mathbf{T}\mathbf{e}_1 + n_2\mathbf{T}\mathbf{e}_2 + n_3\mathbf{T}\mathbf{e}_3. \tag{4.2.9}$$

That is, the transformation **T**, defined by

$$\mathbf{t}_{\mathbf{n}} = \mathbf{T}\mathbf{n}, \tag{4.2.10}$$

is a linear transformation. It is called the *stress tensor* or the *Cauchy stress tensor*.

4.3 COMPONENTS OF STRESS TENSOR

According to Eq. (4.2.10) of the previous section, the stress vectors $\mathbf{t}_{\mathbf{e}_i}$ on the three coordinate planes (the $\mathbf{e}_i$-planes) are related to the stress tensor **T** by

$$\mathbf{t}_{\mathbf{e}_1} = \mathbf{T}\mathbf{e}_1, \quad \mathbf{t}_{\mathbf{e}_2} = \mathbf{T}\mathbf{e}_2, \quad \mathbf{t}_{\mathbf{e}_3} = \mathbf{T}\mathbf{e}_3. \tag{4.3.1}$$

By the definition of the components of a tensor [see Eq. (2.7.2)], we have

$$\mathbf{T}\mathbf{e}_i = \mathbf{T}_{mi}\mathbf{e}_m. \tag{4.3.2}$$

Thus,

$$\begin{aligned}\mathbf{t}_{\mathbf{e}_1} &= T_{11}\mathbf{e}_1 + T_{21}\mathbf{e}_2 + T_{31}\mathbf{e}_3,\\ \mathbf{t}_{\mathbf{e}_2} &= T_{12}\mathbf{e}_1 + T_{22}\mathbf{e}_2 + T_{32}\mathbf{e}_3,\\ \mathbf{t}_{\mathbf{e}_3} &= T_{13}\mathbf{e}_1 + T_{23}\mathbf{e}_2 + T_{33}\mathbf{e}_3.\end{aligned} \tag{4.3.3}$$

Since $\mathbf{t}_{\mathbf{e}_1}$ is the stress vector acting on the plane whose outward normal is $\mathbf{e}_1$, it is clear from the first equation of Eq. (4.3.3) that T_{11} is its normal component and T_{21} and T_{31} are its tangential components. Similarly, T_{22} is the normal component on the $\mathbf{e}_2$-plane and T_{12} and T_{32} are the tangential components on the same plane, and so on.

We note that for each stress component T_{ij}, the second index j indicates the plane on which the stress component acts and the first index indicates the direction of the component; e.g., T_{12} is the stress component in the direction of $\mathbf{e}_1$ acting on the plane whose outward normal is in the direction of $\mathbf{e}_2$. We also note that the positive normal stresses are also known as *tensile stresses,* and negative normal stresses are known as *compressive stresses*. *Tangential stresses* are also known as *shearing stresses*. Both T_{21} and T_{31} are shearing stress components acting on the same plane (the $\mathbf{e}_1$-plane). Thus, the resultant shearing stress on this plane is given by

$$\boldsymbol{\tau}_1 = T_{21}\mathbf{e}_2 + T_{31}\mathbf{e}_3. \tag{4.3.4}$$

The magnitude of this shearing stress is given by

$$|\boldsymbol{\tau}_1| = \sqrt{T_{21}^2 + T_{31}^2}. \tag{4.3.5}$$

Similarly, on $\mathbf{e}_2$-plane,

$$\boldsymbol{\tau}_2 = T_{12}\mathbf{e}_1 + T_{32}\mathbf{e}_3, \tag{4.3.6}$$

and on $\mathbf{e}_3$-plane,

$$\boldsymbol{\tau}_3 = T_{13}\mathbf{e}_1 + T_{23}\mathbf{e}_2. \tag{4.3.7}$$

From $\mathbf{t} = \mathbf{T}\mathbf{n}$, the components of **t** are related to those of **T** and **n** by the equation

$$t_i = T_{ij}n_j, \tag{4.3.8}$$

or, in a form more convenient for computation,

$$[\mathbf{t}] = [\mathbf{T}][\mathbf{n}]. \tag{4.3.9}$$

Thus, it is clear that if the matrix of **T** is known, the stress vector **t** on any inclined plane is uniquely determined from Eq. (4.3.9). In other words, the state of stress at a point is completely characterized by the stress tensor **T**. Also, since **T** is a second-order tensor, any one matrix of **T** determines the other matrices of **T** (see Section 2.18).

We should also note that some authors use the convention $\mathbf{t} = \mathbf{T}^{\mathrm{T}}\mathbf{n}$ so that $\mathbf{t}_{\mathbf{e}_i} = T_{ij}\mathbf{e}_j$. Under that convention, for example, T_{21} and T_{23} are tangential components of the stress vector on the plane whose normal is $\mathbf{e}_2$, and so on. These differences in meaning regarding the nondiagonal elements of $\mathbf{T}$ disappear if the stress tensor is symmetric.

4.4 SYMMETRY OF STRESS TENSOR: PRINCIPLE OF MOMENT OF MOMENTUM

By the use of the moment of momentum equation for a differential element, we shall now show that the stress tensor is generally a symmetric tensor.* Consider the free body diagram of a differential parallelepiped isolated from a body, as shown in Figure 4.4-1. Let us find the moment of all the forces about an axis passing through the center point A and parallel to the x_3-axis:

$$\begin{aligned}\sum(M_A)_3 = {} & T_{21}(\Delta x_2)(\Delta x_3)(\Delta x_1/2) + (T_{21} + \Delta T_{21})(\Delta x_2)(\Delta x_3)(\Delta x_1/2) \\ & - T_{12}(\Delta x_1)(\Delta x_3)(\Delta x_2/2) + (T_{12} + \Delta T_{12})(\Delta x_1)(\Delta x_3)(\Delta x_2/2).\end{aligned} \tag{4.4.1}$$

In writing Eq. (4.4.1), we have assumed the absence of body moments. Dropping the terms containing small quantities of higher order, we obtain

$$\sum(M_A)_3 = (T_{21} - T_{12})(\Delta x_1)(\Delta x_2)(\Delta x_3). \tag{4.4.2}$$

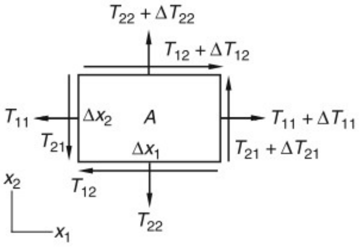

FIGURE 4.4-1

Now, whether the element is in static equilibrium or not,

$$\sum(M_A)_3 = I_{33}\alpha = 0. \tag{4.4.3}$$

This is because the angular acceleration term, $I_{33}\alpha$, is proportional to the moment of inertia I_{33}, which is given by $(1/12)(\text{density})\Delta x_1\Delta x_2\Delta x_3[(\Delta x_1)^2 + (\Delta x_2)^2]$ and is therefore a small quantity of higher order compared with the term $(T_{21}-T_{12})(\Delta x_1)(\Delta x_2)(\Delta x_3)$. Thus,

$$\sum(M_A)_3 = (T_{21} - T_{12})(\Delta x_1)(\Delta x_2)(\Delta x_3) = 0. \tag{4.4.4}$$

---

*See Prob. 4.29 for a case in which the stress tensor is not symmetric.

With similar derivations for the moments about the other two axes, we have

$$T_{12} = T_{21}, \quad T_{13} = T_{31}, \quad T_{23} = T_{32}. \tag{4.4.5}$$

These equations state that the stress tensor is symmetric, i.e., $\mathbf{T} = \mathbf{T}^{\mathrm{T}}$. Therefore, there are only six independent stress components.

Example 4.4.1

The state of stress at a certain point is $\mathbf{T} = -p\mathbf{I}$, where p is a scalar. Show that there is no shearing stress on any plane containing this point.

Solution

The stress vector on any plane passing through the point with normal $\mathbf{n}$ is

$$\mathbf{t_n} = \mathbf{Tn} = -p\mathbf{In} = -p\mathbf{n}.$$

Therefore, it is normal to the plane. This simple stress state is called a *hydrostatic state of stress*.

Example 4.4.2

With reference to a rectangular Cartesian coordinate system, the matrix of a state of stress at a certain point in a body is given by

$$[\mathbf{T}] = \begin{bmatrix} 2 & 4 & 3 \\ 4 & 0 & 0 \\ 3 & 0 & -1 \end{bmatrix} MPa.$$

(a) Find the stress vector and the magnitude of the normal stress on a plane that passes through the point and is parallel to the plane $x_1 + 2x_2 + 2x_3 - 6 = 0$.

(b) If $\mathbf{e}'_1 = \frac{1}{3}(2\mathbf{e}_1 + 2\mathbf{e}_2 + \mathbf{e}_3)$ and $\mathbf{e}'_2 = \frac{1}{\sqrt{2}}(\mathbf{e}_1 - \mathbf{e}_2)$, find T'_{12}.

Solution

(a) The plane $x_1 + 2x_2 + 2x_3 - 6 = 0$ has a unit normal given by

$$\mathbf{n} = \frac{1}{3}(\mathbf{e}_1 + 2\mathbf{e}_2 + 2\mathbf{e}_3).$$

The stress vector is obtained from Eq. (4.3.9) as

$$[\mathbf{t}] = [\mathbf{T}][\mathbf{n}] = \frac{1}{3}\begin{bmatrix} 2 & 4 & 3 \\ 4 & 0 & 0 \\ 3 & 0 & -1 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 16 \\ 4 \\ 1 \end{bmatrix},$$

or

$$\mathbf{t} = \frac{1}{3}(16\mathbf{e}_1 + 4\mathbf{e}_2 + \mathbf{e}_3)\ MPa.$$

The magnitude of the normal stress is, with $T_n \equiv T_{(n)(n)}$,

$$T_n = \mathbf{t} \cdot \mathbf{n} = \frac{1}{9}(16 + 8 + 2) = 2.89\ MPa.$$

(b) To find the primed components of the stress tensor, we have

$$T'_{12} = \mathbf{e}'_1 \cdot \mathbf{T}\mathbf{e}'_2 = \frac{1}{3\sqrt{2}}[2 \quad 2 \quad 1]\begin{bmatrix} 2 & 4 & 3 \\ 4 & 0 & 0 \\ 3 & 0 & -1 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} = \frac{7}{3\sqrt{2}} = 1.65\ MPa.$$

Example 4.4.3

The distribution of stress inside a body is given by the matrix

$$[\mathbf{T}] = \begin{bmatrix} -p+\rho gy & 0 & 0 \\ 0 & -p+\rho gy & 0 \\ 0 & 0 & -p+\rho gy \end{bmatrix},$$

where p, ρ, and g are constants. Figure 4.4-2(a) shows a rectangular block inside the body.

(a) What is the distribution of the stress vector on the six faces of the block?
(b) Find the total resultant force acting on the face $y = 0$ and $x = 0$.

Solution

(a) From $\mathbf{t} = \mathbf{Tn}$, we have

$$\begin{array}{lll}
\text{On } x=0, & [\mathbf{n}] = [-1 \quad 0 \quad 0], & [\mathbf{t}] = [\,p-\rho gy \quad 0 \quad 0\,], \\
\text{On } x=a, & [\mathbf{n}] = [+1 \quad 0 \quad 0], & [\mathbf{t}] = [\,-p+\rho gy \quad 0 \quad 0\,], \\
\text{On } y=0, & [\mathbf{n}] = [0 \quad -1 \quad 0], & [\mathbf{t}] = [\,0 \quad p \quad 0\,], \\
\text{On } y=b, & [\mathbf{n}] = [0 \quad +1 \quad 0], & [\mathbf{t}] = [\,0 \quad -p+\rho gb \quad 0\,], \\
\text{On } z=0, & [\mathbf{n}] = [0 \quad 0 \quad -1], & [\mathbf{t}] = [\,0 \quad 0 \quad p-\rho gy\,], \\
\text{On } z=c, & [\mathbf{n}] = [0 \quad 0 \quad +1], & [\mathbf{t}] = [\,0 \quad 0 \quad -p+\rho gy\,].
\end{array}$$

The distribution of the stress vector on four faces of the cube is shown in Figure 4.4-2(b).

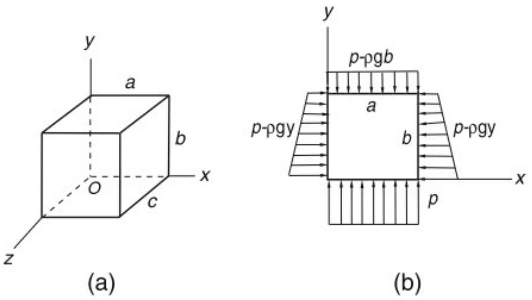

FIGURE 4.4-2

05 05 2020

Bernhard Rieman 1826-1866 ... pdf

→ Turning Points in the Conception of Mathematics.

Obs: Ver todo livro!

165) 20200506

Books_79_2020/
Introduction to Continuum Mechanics.../

A1) Introduction to continuum mechanics ... pdf

Pág. 22 (26)

2.15 Orthogonal Tensors

Obs. Livro todo!

Pág. 45 (49)

2.26 Tensor-Valued Functions of a Scalar

A2) Arithmetic algebraic geometry (2001) ... pdf

Pág. 149 (158)

Introduction to Serre's conjecture

↳ elliptic curve

data 06.05.2020

S T [✓ Q] Q S S D

Pág. 129 (138) Chapter 6

Some other conjectures

6.1. Integral points on elliptic curves

B) Books_79_2020/
Phase Transitions in Materials/
Phase_Transitions_in_Materials.pdf

Pág. 133 (154) Chapter 6 Energy

6.1 Atomic Schrödinger Equations and Formalism

Box 6.1 Dirac Notation

C) Books_79_2020/
Pythagoras' Legacy Mathematics in Ten Great Ideas/
Pythagoras_Legacy.pdf

Obs: Muito bom!

Pág. 82 (93) e: A very special number

Pág. 93 (104) i: Imaginary numbers

Pág. 75 (86) Pascal's Triangle

Obs: Pág. 82 (93) → Gauss

)*+,)))-) ' * 7 A
2 * A> & *!8 ! !A 6!
(%$: 6 8

CHAPTER 1

Introduction to Serre's conjecture

1.1. Introduction

Let's start with an elliptic curve $E/\mathbf{Q}$. Nowadays, it's a familiar activity to consider the Galois representations defined by groups of division points of E. Namely, let n be a positive integer, and let $E[n]$ be the kernel of multiplication by n on $E(\overline{\mathbf{Q}})$. The group $E[n]$ is free of rank two over $\mathbf{Z}/n\mathbf{Z}$. After a choice of basis, the action of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ on $E[n]$ is given by a homomorphism

$$\rho_n : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{Aut}(E[n]) \approx \mathrm{GL}(2, \mathbf{Z}/n\mathbf{Z}).$$

This homomorphism is unramified at each prime p that is prime to the product of n with the conductor of E (see Exercise 15). For each such p, the element $\rho_n(\mathrm{Frob}_p)$ is a 2×2 matrix that is well defined up to conjugation. Its determinant is $p \bmod n$; its trace is $a_p \bmod n$, where a_p is the usual "trace of Frobenius" attached to E and p, i.e., the quantity $1 + p - \#E(\mathbf{F}_p)$. In his 1966 article [**107**], G. Shimura studied these representations and the number fields that they cut out for the curve $E = J_0(11)$. (This curve was also studied by Serre [**91**, pg. 254].) He noticed that for prime values $n = \ell$, the representations ρ_n tended to have large images. In [**93**], J-P. Serre proved that for any fixed elliptic curve E, not having complex multiplication, the indices

$$[\mathrm{GL}(2, \mathbf{Z}/n\mathbf{Z}) : \rho_n(\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}))]$$

are bounded independently of n. In Shimura's example, Serre proved that

$$[\mathrm{GL}(2, \mathbf{Z}/\ell\mathbf{Z}) : \rho_\ell(\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}))] = 1$$

for all $\ell \neq 5$ (see [**93**, §5.5.1]).

In this article, we will be concerned mainly with two-dimensional representations over finite fields. To that end, we restrict attention to the case where $n = \ell$ is prime. The representation ρ_ℓ is "modular" in the familiar sense that it's a representation of a group over a field in positive characteristic. The theme of this course is that it's modular in a different and deeper sense: it comes from a modular form! Indeed, according to a recent preprint of Breuil, Conrad, Diamond and Taylor (see [**7, 19, 114, 117**]), the Shimura-Taniyama conjecture is now a theorem—all elliptic curves over $\mathbf{Q}$ are modular!! Thus if N is the conductor of E, there is a weight-two newform $f = \sum_{n=1}^{\infty} c_n q^n$ ($q = e^{2\pi i z}$) on $\Gamma_0(N)$ with the property that $a_p = c_p$ for all p prime to N. Accordingly, ρ_ℓ is connected up with modular forms via the relation $\mathrm{tr}(\rho_\ell(\mathrm{Frob}_p)) \equiv c_p \pmod{\ell}$, valid for all but finitely many primes p.

The Shimura-Taniyama conjecture asserts that for each positive integer N there is a bijection between isogeny classes of elliptic curves A over $\mathbf{Q}$ of conductor N and rational newforms f on $\Gamma_0(N)$ of weight two. Given A, the Shimura-Taniyama conjecture produces a modular form $f = \sum_{n=1}^{\infty} c_n q^n$ whose Dirichlet series is equal to the L-series of A:

$$\sum_{n=1}^{\infty} \frac{c_n}{n^s} = L(f,s) \xlongequal{\qquad\qquad} L(A,s) = \sum_{n=1}^{\infty} \frac{a_n}{n^s}\ .$$

The integers a_n encode information about the number of points on A over various finite fields $\mathbf{F}_p$. If p is a prime not dividing N, then $a_p = p + 1 - \#A(\mathbf{F}_p)$; if $p \mid N$,

$$a_p = \begin{cases} -1 & \text{if } A \text{ has non-split multiplicative reduction at } p \\ 1 & \text{if } A \text{ has split multiplicative reduction at } p \\ 0 & \text{if } A \text{ has additive reduction at } p. \end{cases}$$

The integers a_n are obtained recursively from the a_p as follows:

- $a_{p^r} = \begin{cases} a_{p^{r-1}} a_p - p a_{p^{r-2}} & \text{if } p \nmid N \\ a_p^r & \text{if } p \mid N \end{cases}$
- $a_{nm} = a_n a_m, \qquad$ if $(n,m) = 1$.

The conjectures made by Serre in [**102**], which are the subject of this paper, concern representations $\rho : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}(2, \overline{\mathbf{F}}_\ell)$. We always require (usually tacitly) that our representations are continuous. The continuity condition just means that the kernel of ρ is open, so that it corresponds to a finite Galois extension K of $\mathbf{Q}$. The representation ρ then embeds $\mathrm{Gal}(K/\mathbf{Q})$ into $\mathrm{GL}(2, \overline{\mathbf{F}}_\ell)$. Since K is a finite extension of $\mathbf{Q}$, the image of ρ is contained in $\mathrm{GL}(2, \mathbf{F})$ for some finite subfield $\mathbf{F}$ of $\overline{\mathbf{F}}_\ell$.

$$\begin{tikzcd} \overline{\mathbf{Q}} \arrow[d, dash] \arrow[dd, bend left, dash, "G_{\mathbf{Q}}"] \\ K \arrow[d, dash] \\ \mathbf{Q} \end{tikzcd} \qquad\qquad \begin{tikzcd} G_{\mathbf{Q}} \arrow[d, two heads] \arrow[dr, "\rho"] & \\ \mathrm{Gal}(K/\mathbf{Q}) \arrow[r, hook] & \mathrm{GL}(2, \overline{\mathbf{F}}_\ell) \end{tikzcd}$$

For various technical reasons, the original conjectures of Serre insist that ρ be irreducible. It is nevertheless fruitful to consider the reducible case as well (see [**111**]).

The conjectures state (in particular) that each continuous irreducible ρ that satisfies a necessary parity condition "arises from" (or is associated with) a suitable modular form mod ℓ. To explain what's going on, let's start with

$$\Delta := \sum_{n=1}^{\infty} \tau(n) q^n = q \prod_{i=1}^{\infty} (1 - q^i)^{24},$$

the unique (normalized) cusp form of weight 12 on $\mathrm{SL}(2, \mathbf{Z})$. In [**92**], Serre conjectured the existence of a "strictly compatible" family of ℓ-adic representations of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ whose L-function is the L-function of Δ, namely

$$L(\Delta, s) = \sum_{n=1}^{\infty} \tau(n) n^{-s} = \prod_p (1 - \tau(p) p^{-s} + p^{11-2s})^{-1},$$

where the product is taken over all prime numbers p. The conjectured ℓ-adic representations were constructed soon after by Deligne [**24**]. Specifically, Deligne constructed, for each prime ℓ, a representation

$$\rho_{\ell^\infty} : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}(2, \mathbf{Z}_\ell),$$

unramified outside ℓ, such that for all primes $p \neq \ell$,

$$\operatorname{tr}(\rho_{\ell^\infty}(\mathrm{Frob}_p)) = \tau(p), \qquad \det(\rho_{\ell^\infty}(\mathrm{Frob}_p)) = p^{11}.$$

On reducing ρ_{ℓ^∞} mod ℓ, we obtain a representation

$$\rho_\ell : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}(2, \mathbf{F}_\ell)$$

with analogous properties. (Equalities are replaced by congruences mod ℓ.) In other words, the ρ_ℓ for Δ are just like the ρ_ℓ for an elliptic curve E, except that the integers a_p are replaced by the corresponding values of the τ-function. The determinant of ρ_ℓ is the 11th power of the mod ℓ cyclotomic character $\chi : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathbf{F}_\ell^*$, i.e., the character giving the action of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ on the group of ℓth roots of unity in $\overline{\mathbf{Q}}$ (see Section 1.5).

More generally, take a weight $k \geq 12$ and suppose that $f = \sum_n c_n q^n$ is a nonzero weight-k cusp form for $\mathrm{SL}(2, \mathbf{Z})$ that satisfies $f|T_n = c_n f$ for all $n \geq 1$, T_n being the nth Hecke operator on the space of cusp forms of weight k for $\mathrm{SL}(2, \mathbf{Z})$ (see Section 1.5). Then the complex numbers c_n $(n \geq 1)$ are algebraic integers. Moreover, the field $E := \mathbf{Q}(\ldots c_n \ldots)$ generated by the c_n is a totally real number field (of finite degree over $\mathbf{Q}$). Thus the c_n all lie in the integer ring $\mathcal{O}_E$ of E. For each ring homomorphism $\varphi : \mathcal{O}_E \to \overline{\mathbf{F}}_\ell$, one finds a representation

$$\rho = \rho_\varphi : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}(2, \overline{\mathbf{F}}_\ell),$$

unramified outside ℓ, such that

$$\operatorname{tr}(\rho(\mathrm{Frob}_p)) = \varphi(c_p), \qquad \det(\rho(\mathrm{Frob}_p)) = p^{k-1}$$

for all $p \neq \ell$. We have $\det \rho = \chi^{k-1}$. Of course, there is no guarantee that ρ is irreducible. We can (and do) suppose that ρ is semisimple by replacing it by its semisimplification. Then ρ is determined up to isomorphism by the displayed trace and determinant conditions; this follows from the Cebotarev density theorem and the Brauer-Nesbitt theorem [**21**], which states that semisimple representations are determined by their characteristic polynomials.

It is important to note that k is necessarily an even integer; otherwise the space $S_k(\mathrm{SL}(2, \mathbf{Z}))$ of weight-k cusp forms on $\mathrm{SL}(2, \mathbf{Z})$ is easily seen to be 0. Thus the determinant χ^{k-1} of ρ is an odd power of χ. In particular, $\det \rho : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathbf{F}_\ell^*$ is unramified outside ℓ and takes the value -1 on complex conjugations $c \in \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$. It's a nice exercise to check that, conversely, all continuous homomorphisms with these properties are odd powers of χ (see Exercise 1).

In the early 1970s, Serre conjectured that all homomorphisms that are "like ρ" come from cusp forms of some weight on $\mathrm{SL}(2, \mathbf{Z})$. Namely, let

$$\rho : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}(2, \overline{\mathbf{F}}_\ell)$$

be a continuous, irreducible representation that is (1) unramified outside ℓ and (2) of odd determinant, in the sense that $\det \rho(c) = -1 \in \overline{\mathbf{F}}_\ell$ for complex conjugations $c \in \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$. In a May, 1973 letter to Tate, Serre conjectured that ρ is of the form ρ_φ. This means that there is a weight $k \geq 12$, an eigenform $f \in S_k(\mathrm{SL}(2, \mathbf{Z}))$, and a homomorphism $\varphi : \mathcal{O}_E \to \overline{\mathbf{F}}_\ell$ (where $\mathcal{O}_E$ is the ring of integers of the field generated by the coefficients of f) so that ρ_φ and ρ are isomorphic.

To investigate Serre's conjecture, it is fruitful to consider the operation $\rho \mapsto \rho \otimes \chi$ on representations. This "twisting" operation preserves the set of representations that come from modular forms. Indeed, let $\theta = q\frac{d}{dq}$ be the classical differential operator $\sum a_n q^n \mapsto \sum n a_n q^n$. According to Serre and Swinnerton-Dyer [**61, 94,**

112], if f is a mod ℓ form of weight k, then θf is a mod ℓ form of weight $k+\ell+1$. Then if ρ is associated to f, $\rho\otimes\chi$ is associated with θf. According to a result of Atkin, Serre and Tate (see [**97**, Th. 3] and Section 2.1), if ρ comes from an eigenform in some space $S_k(\mathrm{SL}(2,\mathbf{Z}))$, then a suitable twist $\rho\otimes\chi^i$ of f comes from a form of weight $\leq \ell+1$.

Serre's conjecture thus has the following consequence: each two-dimensional irreducible odd representation of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ over $\overline{\mathbf{F}}_\ell$ that is unramified outside ℓ has a twist (by a power of χ) coming from an eigenform on $\mathrm{SL}(2,\mathbf{Z})$ of weight at most $\ell+1$. In particular, suppose that $\ell < 11$. Then the spaces $S_k(\mathrm{SL}(2,\mathbf{Z}))$ with $k \leq \ell+1$ are all 0; as a result, they contain no nonzero eigenforms! The conjecture that all ρ are modular (of level 1) thus predicts that there are *no* representations of the type contemplated if ℓ is 2, 3, 5 or 7. In support of the conjecture, the non-existence statement was proved for $\ell = 2$ by J. Tate in a July, 1973 letter to Serre [**113**]. Soon after, Serre treated the case $\ell = 3$ by methods similar to those of Tate. (See [**113**, p. 155] for a discussion and a reference to a note in Serre's Œuvres.) Quite recently, Sharon Brueggeman considered the case $\ell = 5$; she proved that the conjectured result follows from the Generalized Riemann Hypothesis (see [**8**]). In another direction, Hyunsuk Moon generalized Tate's result and proved that there are only finitely many isomorphism classes of continuous semisimple Galois representations $\rho : G_{\mathbf{Q}} \to \mathrm{GL}_4(\overline{\mathbf{F}}_2)$ unramified outside 2 such that field $K/\mathbf{Q}$ corresponding to the kernel of ρ is totally real (see [**76**]). Similar work in this direction has been carried out by Joshi [**58**], under additional local hypothesis.

Serre discussed his conjecture with Deligne, who pointed out a number of surprising consequences. In particular, suppose that one takes a ρ coming from an eigenform f' of some weight and of level $N > 1$. On general grounds, ρ has the right to be ramified at primes p dividing N as well as at the prime ℓ. Suppose that, by accident as it were, ρ turned out to be unramified at all primes $p \mid N$. Then the conjecture would predict the existence of a level-1 form f' (presumably of the same weight as f) whose mod ℓ representation was isomorphic to ρ. For example, if $N = \ell^\alpha$ is a power of ℓ, then the conjecture predicts that ρ arises from a form f' of level 1. How could one manufacture the f'?

The passage $f \rightsquigarrow f'$ comes under the rubric of "level optimization". When you take a representation ρ that comes from high level N, and it seems as though that representation comes from a lower level N', then to "optimize the level" is to cough up a form of level N' that gives ρ.

Deligne pointed out also that Serre's conjecture implies that representations ρ over $\overline{\mathbf{F}}_\ell$ are required to "lift" to λ-adic representations of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$. In the recent articles [**80**] and [**81**], R. Ramakrishna used purely Galois cohomological techniques to prove results in this direction.

1.2. The weak conjecture of Serre

In the mid 1980s, Gerhard Frey began lecturing on a link between Fermat's Last Theorem and elliptic curves (see [**42, 43**]). (Earlier, Hellegouarch had also considered links between Fermat's Last Theorem and elliptic curves; see the MathSciNet review and Appendix of [**48**].) As is now well known, Frey proposed that if $a^\ell + b^\ell$ was a perfect ℓth power, then the elliptic curve $y^2 = x(x-a^\ell)(x+b^\ell)$ could be proved to be non-modular. Soon after, Serre pointed out that the non-modularity contemplated by Frey would follow from suitable level-optimization results concerning modular forms [**101**]. To formulate such optimization results, Serre went back to the tentative conjecture that he had made 15 years earlier and decided to study representations $\rho : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}(2, \overline{\mathbf{F}}_\ell)$ that are not necessarily unramified at ℓ. The results, of course, were the conjectures of [**102**].

An important consequence of these conjectures is the so-called "weak conjecture of Serre." As background, we recall that Hecke eigenforms on congruence subgroups of $\mathrm{SL}(2, \mathbf{Z})$ give rise to two-dimensional representations of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$. If we set things up correctly, we get representations over $\overline{\mathbf{F}}_\ell$. More specifically, take integers $k \geq 2$ and $N \geq 1$; these are the weight and level, respectively. Let $f = \sum a_n q^n$ be a normalized Hecke eigenform in the space $S_k(\Gamma_1(N))$ of complex weight-k cusp forms on the subgroup $\Gamma_1(N)$ of $\mathrm{SL}(2, \mathbf{Z})$. Thus f is nonzero and it satisfies $f|T_n = a_n f$ for all $n \geq 1$. Further, there is a character $\varepsilon : (\mathbf{Z}/N\mathbf{Z})^* \to \mathbf{C}^*$ so that $f|\langle d \rangle = \varepsilon(d) f$ for all $d \in (\mathbf{Z}/N\mathbf{Z})^*$, where $\langle d \rangle$ is the diamond-bracket operator. Again, let $\mathcal{O}$ be the ring of integers of the field $\mathbf{Q}(\dots a_n \dots)$ generated by the a_n; this field is a number field that is either totally real or a CM field. Consider a ring homomorphism $\varphi : \mathcal{O} \to \overline{\mathbf{F}}_\ell$ as before. Associated to this set-up is a representation $\rho : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}(2, \overline{\mathbf{F}}_\ell)$ with properties that connect it up with f (and φ). First, the representation is unramified at all p not dividing ℓN. Next, for all such p, we have

$$\mathrm{tr}(\rho(\mathrm{Frob}_p)) = a_p, \qquad \det(\rho(\mathrm{Frob}_p)) = p^{k-1}\varepsilon(p);$$

the numbers a_p and $p^{k-1}\varepsilon(p)$, literally in $\mathcal{O}$, are mapped tacitly into $\overline{\mathbf{F}}_\ell$ by φ. The representation ρ is determined up to isomorphism by the trace and determinant identities that are displayed, plus the supplemental requirement that it be semisimple. We are interested mainly in the (generic) case in which ρ is irreducible; in that case, it is of course semisimple.

The construction of ρ from f, k and φ was described in [**24**]. In this article, Deligne sketches a method that manufactures for each non-archimedean prime λ of E a representation $\tilde{\rho}_\lambda : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}(2, E_\lambda)$, where E_λ denotes the completion of E at λ. Given φ, we let $\lambda = \ker \varphi$ and find a model of $\tilde{\rho}_\lambda$ over the ring of integers $\mathcal{O}_\lambda$ of E_λ. Reducing $\tilde{\rho}_\lambda$ modulo λ, we obtain a representation over the finite field $\mathcal{O}_\lambda/\lambda\mathcal{O}_\lambda$, and φ embeds this field into $\overline{\mathbf{F}}_\ell$.

In fact, as Shimura has pointed out, the machinery of [**24**] can be avoided if one seeks only the mod λ representation attached to f (as opposed to the full λ-adic representation $\tilde{\rho}_\lambda$). As the first author pointed out in [**87**], one can use congruences among modular forms to find a form of weight two and level $N\ell^2$ that gives rise to ρ. Accordingly, one can find ρ concretely by looking within the group of ℓ-division points of a suitable abelian variety over $\mathbf{Q}$: the variety $J_1(\ell^2 N)$, which is defined in Section 2.3 and in Conrad's Appendix.

Which representations of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ arise in this way (as k, N, f and φ all vary)? As in the case $N = 1$ (i.e., that where $\Gamma_1(N) = \mathrm{SL}(2, \mathbf{Z})$), any ρ that comes from modular forms is an odd representation: we have $\det(\rho(c)) = -1$ when $c \in \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ is a complex conjugation. To see this, we begin with the fact that $\varepsilon(-1) = (-1)^k$, which generalizes (1.4); this follows from the functional equation that relates $f(\frac{az+b}{cz+d})$ to $f(z)$ when $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ is an element of $\Gamma_0(N)$ (see Exercise 7). On the other hand, using the Cebotarev density theorem, we find that $\det \rho = \chi^{k-1}\varepsilon$, where χ is again the mod ℓ cyclotomic character and where ε is regarded now as a map $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \overline{\mathbf{F}}_\ell^*$ in the obvious way, namely by composing $\varepsilon : (\mathbf{Z}/N\mathbf{Z})^* \to \overline{\mathbf{F}}_\ell^*$

with the mod N cyclotomic character. The value on c of the latter incarnation of ε is the number $\varepsilon(-1) = (-1)^k$. Since $\chi(c) = -1$, we deduce that $(\det\rho)(c) = -1$, as was claimed.

Serre's weak conjecture states that, conversely, every irreducible odd representation $\rho : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}(2, \overline{\mathbf{F}}_\ell)$ is modular in the sense that it arises from some f and φ.

A concrete consequence of the conjecture is that all odd irreducible 2-dimensional Galois representations ρ come from abelian varieties over $\mathbf{Q}$. Given ρ, one should be able to find a totally real or CM number field E, an abelian variety A over $\mathbf{Q}$ of dimension $[E : \mathbf{Q}]$ that comes equipped with an action of the ring of integers $\mathcal{O}$ of E, and a ring homomorphism $\varphi : \mathcal{O} \to \overline{\mathbf{F}}_\ell$ with the following property: Let $\lambda = \ker\varphi$. *Then the representation* $A[\lambda] \otimes_{\mathcal{O}/\lambda} \overline{\mathbf{F}}_\ell$ *is isomorphic to* ρ. (In comparing $A[\lambda]$ and ρ, we use $\varphi : \mathcal{O}/\lambda \hookrightarrow \overline{\mathbf{F}}_\ell$ to promote the 2-dimensional $A[\lambda]$ into a representation over $\overline{\mathbf{F}}_\ell$.)

Much of the evidence for the weak conjecture concerns representations taking values in $\mathrm{GL}(2, \mathbf{F}_q)$ where the finite field $\mathbf{F}_q$ has small cardinality. In his original article [**102**, §5], Serre's discusses a large number of examples of such representations. Serre uses theorems of Langlands [**68**] and Tunnell [**115**] to establish his weak conjecture for odd irreducible representations with values in $\mathrm{GL}(2, \mathbf{F}_2)$ and $\mathrm{GL}(2, \mathbf{F}_3)$. Further, he reports on numerical computations of J.-F. Mestre that pertain to representations over $\mathbf{F}_4$ (and trivial determinant). Additionally, Serre remarks [**102**, p. 219] that the weak conjecture is true for those representations with values in $\mathrm{GL}(2, \overline{\mathbf{F}}_p)$ that are dihedral in the sense that their projective images (in $\mathrm{PGL}(2, \overline{\mathbf{F}}_p)$) are dihedral groups. (See also [**29**, §5] for a related argument.) This section of Serre's paper concludes with examples over $\mathbf{F}_9$ and $\mathbf{F}_7$.

More recently, representations over the fields $\mathbf{F}_4$ and $\mathbf{F}_5$ were treated, under somewhat mild hypotheses, by Shepherd-Barron and Taylor [**105**]. For example, Shepherd-Barron and Taylor show that $\rho : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}(2, \mathbf{F}_5)$ is isomorphic to the 5-torsion representaton of an elliptic curve over $\mathbf{Q}$ provided that $\det\rho$ is the mod 5 cyclotomic character. Because elliptic curves over $\mathbf{Q}$ are modular, it follows that ρ is modular.

1.3. The strong conjecture

Fix an odd irreducible Galois representation

$$\rho : G_{\mathbf{Q}} \to \mathrm{GL}(2, \overline{\mathbf{F}}_\ell).$$

As discussed above, the weak conjecture asserts that ρ is modular, in the sense that there exists integers N and k such that ρ comes from some $f \in S_k(\Gamma_1(N))$. The *strong conjecture* goes further and gives a recipe for integers $N(\rho)$ and $k(\rho)$, then asserts that ρ comes from $S_{k(\rho)}(\Gamma_1(N(\rho)))$. In any particular instance, the strong conjecture is, a priori, easier to verify or disprove than the weak conjecture because $S_{k(\rho)}(\Gamma_1(N(\rho)))$ is a finite-dimensional vector space that can be computed (using, e.g., the algorithm in [**73**]). The relation between the weak and strong conjectures is analogous to the relation between the assertion that an elliptic curve is modular of some level and the assertion that an elliptic curve A is modular of a specific level, the conductor of A.

For each prime p, let $I_p \subset G_{\mathbf{Q}}$ denote an inertia group at p. The optimal level is a product

$$N(\rho) = \prod_{p \neq \ell} p^{n(p)},$$

where $n(p)$ depends only on $\rho|_{I_p}$. The optimal weight $k(\rho)$ depends only on $\rho|_{I_\ell}$. The integer $n(p)$ is a conductor in additive notation. In particular, $n(p) = 0$ if and only if ρ is unramified at p.

View ρ as a homomorphism $G_{\mathbf{Q}} \to \operatorname{Aut}(V)$, where V is a two-dimensional vector space over $\overline{\mathbf{F}}_\ell$. It is natural to consider the subspace of inertia invariants:

$$V^{I_p} := \{v \in V : \rho(\sigma)v = v, \text{ all } \sigma \in I_p\}.$$

For example, $V^{I_p} = V$ if and only if ρ is unramified at p. Define

$$n(p) := \dim(V/V^{I_p}) + \operatorname{Swan}(V),$$

where the wild term $\operatorname{Swan}(V)$ is the Swan conductor

$$\operatorname{Swan}(V) := \sum_{i=1}^{\infty} \frac{1}{[G_0 : G_i]} \dim(V/V^{G_i}) \geq 0.$$

Here $G_0 = I_p$ and the $G_i \subset G_0$ are the higher ramification groups.

Suppose that ρ arises from a newform $f \in S_k(\Gamma_1(N))$. A theorem of Carayol [**12**], which was proved independently by Livné [**70**, Prop. 0.1], implies that $N(\rho) \mid N$. It is productive to study the quotient $N/N(\rho)$. Let $\mathcal{O}$ be the ring of integers of the field generated by the Fourier coefficients of f and let $\varphi : \mathcal{O} \to \overline{\mathbf{F}}_\ell$ be the map such that $\varphi(a_p) = \operatorname{tr}(\rho(\operatorname{Frob}_p))$. Let λ be a prime of $\mathcal{O}$ lying over ℓ and E_λ be the completion of $\operatorname{Frac}(\mathcal{O})$ at λ. Deligne [**24**] attached to the pair f, λ a representation

$$\rho_\lambda : G_{\mathbf{Q}} \to \mathrm{GL}(2, E_\lambda) = \operatorname{Aut}(\tilde{V})$$

where $\tilde{V}$ is a two-dimensional vector space over E_λ. The representation ρ_λ can be conjugated so that its images lies inside $\mathrm{GL}(2, \mathcal{O}_\lambda)$; the reduction of ρ_λ modulo λ is then ρ. The following diagram summarizes the set up:

$$\begin{tikzcd}
 & & \mathrm{GL}(2, E_\lambda) = \operatorname{Aut}(\tilde{V}) \\
G_{\mathbf{Q}} \arrow[r] \arrow[rru, bend left, "\rho_\lambda"] \arrow[rrd, bend right, "\rho"'] & \mathrm{GL}(2, \mathcal{O}_\lambda) \arrow[ru, hook] \arrow[rd, "\varphi"] & \\
 & & \mathrm{GL}(2, \overline{\mathbf{F}}_\ell) = \operatorname{Aut}(V)
\end{tikzcd}$$

Let $m(p)$ be the power of p dividing the conductor of ρ_λ. In [**12**], Carayol proves that $m(p) = \operatorname{ord}_p N$. As above, $m(p) = \dim(\tilde{V}/\tilde{V}^{I_p}) + (\text{wild term})$, and the wild term is the same as for ρ. The power of p dividing $N/N(\rho)$ is $\dim(\tilde{V}/\tilde{V}^{I_p}) - \dim(V/V^{I_p}) = \dim V^{I_p} - \dim \tilde{V}^{I_p}$. Though $\tilde{V}$ and V are vector spaces over different fields, we can compare the dimensions of their inertia invariant subspaces. The formula

$$\operatorname{ord}_p(N) = n(p) + (\dim V^{I_p} - \dim \tilde{V}^{I_p}) \tag{1.1}$$

indicates how this difference is the deviation of N from the optimal level locally at p. This is the description of $n(p)$ that is used in proving that if ρ is modular at all, then it is possible to refine N and k to eventually discover that ρ arises from a newform in $S_{k(\rho)}(\Gamma_1(N(\rho)))$. After much work (see [**26, 87**]) it has been shown that *for $\ell > 2$ the weak and strong conjectures are equivalent.* See [**9**] for equivalence in many cases when $\ell = 2$.

Rearranging (1.1) into

$$n(p) = \operatorname{ord}_p(N) - (\dim V^{I_p} - \dim \tilde{V}^{I_p})$$

provides us with a way to read off $N(\rho)$ from $\operatorname{ord}_p(N)$, $\dim V^{I_p}$, and $\dim \tilde{V}^{I_p}$. If $f \in S_k(\Gamma_1(N))$ gives rise to ρ and $\ell \nmid N$, then $k(\rho) \leq k$. In contrast, if we allow powers of ℓ in the level then the weight k can always be made equal to 2.

1.4. Representations arising from an elliptic curve

Equations for elliptic curves can be found in the Antwerp tables [**4**] and the tables of Cremona [**20**]. For example, consider the elliptic curve B given by the equation $y^2+y = x^3+x^2-12x+2$. This is the curve labeled **141A1** in [**20**]; it has conductor $N = 3 \cdot 47$ and discriminant $\Delta = 3^7 \cdot 47$. There is a newform $f \in S_2(\Gamma_0(141))$ attached to B. Because N is square free, the elliptic curve B is *semistable*, in the sense that B has multiplicative reduction at each prime.

The curve B is isolated in its isogeny class; equivalently, for every ℓ the representation

$$\rho_\ell : G_{\mathbf{Q}} \to \operatorname{Aut}(B[\ell]) \approx \mathrm{GL}(2, \mathbf{F}_\ell)$$

is irreducible (see Exercise 4 and Exercise 5). We will frequently consider the representations ρ_ℓ attached to B. The following proposition shows that because B is semistable, each ρ_ℓ is surjective [**93**].

Proposition 1.1. *If A is a semistable elliptic curve over $\mathbf{Q}$ and ℓ is a prime such that ρ_ℓ is irreducible, then ρ_ℓ is surjective.*

Proof. Serre proved this when ℓ is odd; see [**93**, Prop. 21], [**103**, §3.1]. If ρ_2 isn't surjective, then by [**93**, Prop. 21(b)] and Theorem 2.10 it's unramified outside 2. This contradicts [**113**]. □

To give a flavor of Serre's invariants, we describe $N(\rho_\ell)$ and $k(\rho_\ell)$ for the representations ρ_ℓ attached to B. (Note that we still have not defined $k(\rho)$.) At primes p of bad reduction, after a possible unramified quadratic extension of $\mathbf{Q}_p$, the elliptic curve B is a Tate curve. This implies that for $p \neq \ell$, the representation ρ_ℓ is unramified at p if and only if $\operatorname{ord}_p(\Delta) \equiv 0 \pmod{\ell}$; for more details, see Section 2.4.

The optimal level $N(\rho_\ell)$ is a divisor of $3 \cdot 47$; it is divisible only by primes for which ρ_ℓ is ramified, and is not divisible by ℓ. The representation ρ_ℓ is unramified at 3 if and only if $\ell \mid \operatorname{ord}_3(\Delta) = 7$, i.e., when $\ell = 7$. Furthermore, ρ_ℓ is always ramified at 47. First suppose $\ell \notin \{3, 47\}$. If in addition $\ell \neq 7$ then $N(\rho_\ell) = 3 \cdot 47$, and $k(\rho_\ell) = 2$ since $\ell \nmid 3 \cdot 47$. If $\ell = 7$ then $N(\rho_\ell) = 47$, and again $k(\rho_\ell) = 2$. The remaining cases are $\ell = 3$ and $\ell = 47$. If $\ell = 47$ then $N(\rho_\ell) = 3$, and because $\ell - 1$ is the order of the cyclotomic character, $k(\rho_\ell) \equiv 2 \pmod{47-1}$; Serre's recipe then gives $k(\rho_\ell) = 2 + (47 - 1) = 48$. Similarly, when $\ell = 3$, we have $N(\rho_\ell) = 47$ and $k(\rho_\ell) = 2 + (3 - 1) = 4$. The following table summarizes the Serre invariants:

Table 1.4. The Serre invariants of ρ_ℓ

ℓ	$N(\rho_\ell)$	$k(\rho_\ell)$
3	47	4
7	47	2
47	3	48
all other ℓ	141	2

To verify the strong conjecture of Serre for $\ell = 3, 47$, we use a standard tradeoff of level and weight, which relates eigenforms in $S_2(\Gamma_0(141); \mathbf{F}_\ell)$ to eigenforms in $S_{2+\ell-1}(\Gamma_0(141/\ell); \mathbf{F}_\ell)$ (see Section 3.1). The only exceptional prime is $\ell = 7$, for which the minimal weight $k(\rho)$ is 2. The strong conjecture of Serre predicts the existence of an eigenform $g \in S_2(\Gamma_0(47))$ that gives rise to ρ_ℓ. Our initial instinct is to look for an elliptic curve A of conductor 47 such that $A[\ell] \cong B[\ell]$, as $G_\mathbf{Q}$-modules. In fact, there are no elliptic curves of conductor 47. This is because $S_2(\Gamma_0(47))$ is four dimensional, having basis the Galois conjugates of a single eigenform $g = \sum c_n q^n$. The Fourier coefficients c_n of g generate the full ring of integers in the field K obtained from $\mathbf{Q}$ by adjoining a root of $h = x^4 - x^3 - 5x^2 + 5x - 1$. The discriminant $1957 = 19 \cdot 103$ of K equals the discriminant of h, so a root of h generates the full ring of integers. The eigenvalue c_2 satisfies $h(c_2) = 0$. Since $h \equiv (x + 2)(x^3 + 4x^2 + x + 3) \pmod 7$, there is a prime λ lying over 7 such that $\mathcal{O}/\lambda \cong \mathbf{F}_7$; the isomorphism is given by $c_2 \mapsto -2 \bmod 7$. As a check, note that $\#B(\mathbf{F}_2) = 5$ so $a_2 = 3 - 5 = -2 = \varphi(c_2)$. More generally, for $p \nmid 7 \cdot 141$, we have $\varphi(c_p) \equiv a_p \bmod 7$. This equality of traces implies that the representation $\rho_{g,\lambda}$ is isomorphic to $\rho = \rho_{A,7}$, so A is modular of level 47.

1.5. Background material

In this section, we review the cyclotomic character, Frobenius elements, modular forms, and Tate curves. We frequently write $G_\mathbf{Q}$ for $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$. Many of these basics facts are also summarized in [**23**].

1.5.1. The cyclotomic character

The mod ℓ *cyclotomic character* is defined by considering the group $\boldsymbol{\mu}_\ell$ of ℓth roots of unity in $\overline{\mathbf{Q}}$; the action of the Galois group $G_\mathbf{Q}$ on the cyclic group $\boldsymbol{\mu}_\ell$ gives rise to a continuous homomorphism

$$\chi_\ell : G_\mathbf{Q} \to \mathrm{Aut}(\boldsymbol{\mu}_\ell). \tag{1.2}$$

Since $\boldsymbol{\mu}_\ell$ is a cyclic group of order ℓ, its group of automorphisms is canonically the group $(\mathbf{Z}/\ell\mathbf{Z})^* = \mathbf{F}_\ell^*$. We emerge with a map $G_\mathbf{Q} \to \mathbf{F}_\ell^*$, which is the character in question.

Let A be an elliptic curve and ℓ be a prime number. The Weil pairing e_ℓ (see [**109**, III.8]) sets up an isomorphism of $G_\mathbf{Q}$-modules

$$e_\ell : \bigwedge^2 A[\ell] \xrightarrow{\cong} \boldsymbol{\mu}_\ell. \tag{1.3}$$

The determinant of the representation $\rho_{A,\ell}$ is the mod ℓ cyclotomic character χ_ℓ.

Suppose now that $c \in G_\mathbf{Q}$ is the automorphism "complex conjugation." Then the determinant of $\rho_{A,\ell}(c)$ is $\chi_\ell(c)$. Now c operates on roots of unity by the map

$\zeta \mapsto \zeta^{-1}$, since roots of unity have absolute value 1. Accordingly,

$$\det \rho_{A,\ell}(c) = -1; \tag{1.4}$$

one says that $\rho_{A,\ell}$ is *odd*. If $\ell \neq 2$, then $\rho_{A,\ell}(c)$ is conjugate over $\overline{\mathbf{F}}_\ell$ to $\left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$ (Exercise 7). If $\ell = 2$ then the characteristic polynomial of $\rho_{A,\ell}(c)$ is $(x+1)^2$ so $\rho_{A,\ell}(c)$ is conjugate over $\overline{\mathbf{F}}_\ell$ to either the identity matrix or $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$.

1.5.2. Frobenius elements

Let K be a number field. The Galois group $\mathrm{Gal}(K/\mathbf{Q})$ leaves the ring $\mathcal{O}_K$ of integers of K invariant, so that one obtains an induced action on the ideals of $\mathcal{O}_K$. The set of prime ideals $\mathfrak{p}$ of $\mathcal{O}_K$ lying over p (i.e., that contain p) is permuted under this action. For each $\mathfrak{p}$, the subgroup $D_\mathfrak{p}$ of $\mathrm{Gal}(K/\mathbf{Q})$ fixing $\mathfrak{p}$ is called the *decomposition group* of $\mathfrak{p}$. Meanwhile, $\mathbf{F}_\mathfrak{p} := \mathcal{O}_K/\mathfrak{p}$ is a finite extension of $\mathbf{F}_p$. The extension $\mathbf{F}_\mathfrak{p}/\mathbf{F}_p$ is necessarily Galois; its Galois group is cyclic, generated by the Frobenius automorphism $\varphi_p : x \mapsto x^p$ of $\mathbf{F}_\mathfrak{p}$. There is a natural surjective map $D_\mathfrak{p} \to \mathrm{Gal}(\mathbf{F}_\mathfrak{p}/\mathbf{F}_p)$; its injectivity is equivalent to the assertion that p is unramified in $K/\mathbf{Q}$. Therefore, whenever this assertion is true, there is a unique $\sigma_\mathfrak{p} \in D_\mathfrak{p}$ whose image in $\mathrm{Gal}(\mathbf{F}_\mathfrak{p}/\mathbf{F}_p)$ is φ_p. The automorphism $\sigma_\mathfrak{p}$ is then a well-defined element of $\mathrm{Gal}(K/\mathbf{Q})$, the Frobenius automorphism for $\mathfrak{p}$. The various $\mathfrak{p}$ are all conjugate under $\mathrm{Gal}(K/\mathbf{Q})$ and that the Frobenius automorphism for the conjugate of $\mathfrak{p}$ by g is $g\sigma_\mathfrak{p}g^{-1}$. In particular, the various $\sigma_\mathfrak{p}$ are all conjugate; this justifies the practice of writing σ_p for any one of them and stating that σ_p is well defined up to conjugation.

We next introduce the concept of Frobenius elements in $G_\mathbf{Q} = \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$. Let p again be a prime and let $\mathfrak{p}$ now be a prime of the ring of integers of $\overline{\mathbf{Q}}$ lying over p. To $\mathfrak{p}$ we associate: (1) its residue field $\mathbf{F}_\mathfrak{p}$, which is an algebraic closure of $\mathbf{F}_p$, and (2) a decomposition subgroup $D_\mathfrak{p}$ of $G_\mathbf{Q}$. There is again a surjective map $D_\mathfrak{p} \to \mathrm{Gal}(\mathbf{F}_\mathfrak{p}/\mathbf{F}_p)$. The Frobenius automorphism φ_p topologically generates the target group. We shall use the symbol Frob_p to denote any preimage of φ_p in any $D_\mathfrak{p}$ corresponding to a prime lying over p, and refer to Frob_p as a Frobenius element for p in $G_\mathbf{Q}$. This element is doubly ill defined. The ambiguity in Frob_p results from the circumstance that $\mathfrak{p}$ needs to be chosen and from the fact that $D_\mathfrak{p} \to \mathrm{Gal}(\mathbf{F}_\mathfrak{p}/\mathbf{F}_p)$ has a large kernel, the inertia subgroup $I_\mathfrak{p}$ of $D_\mathfrak{p}$. The usefulness of $\mathrm{Frob}_\mathfrak{p}$ stems from the fact that the various $\mathfrak{p}$ are all conjugate, so that likewise the subgroups $D_\mathfrak{p}$ and $I_\mathfrak{p}$ are conjugate. Thus if ρ is a homomorphism mapping $G_\mathbf{Q}$ to some other group, the kernel of ρ contains one $I_\mathfrak{p}$ if and only if it contains all $I_\mathfrak{p}$. In this case, one says that ρ is *unramified* at p; the image of Frob_p is then an element of the target that is well defined up to conjugation.

Consider an elliptic curve A over $\mathbf{Q}$ and let ℓ be a prime number. The fixed field of $\rho_{A,\ell}$ is a finite Galois extension $K_\ell/\mathbf{Q}$ whose Galois group G_ℓ is a subgroup of $\mathrm{GL}(2, \mathbf{F}_\ell)$. A key piece of information about the extension $K_\ell/\mathbf{Q}$ is that its discriminant is divisible at most by ℓ and primes dividing the conductor of A. In other words, if $p \neq \ell$ is a prime number at which A has good reduction, then $K_\ell/\mathbf{Q}$ is unramified at ℓ (see Exercise 15); one says that the representation $\rho_{A,\ell}$ is unramified at p. Whenever this occurs, the construction described above produces a Frobenius element σ_p in G_ℓ that is well defined up to conjugation.

Fix again an elliptic curve A and a prime number ℓ, and let $\rho_{A,\ell} : G_\mathbf{Q} \to \mathrm{GL}(2, \mathbf{F}_\ell)$ be the associated representation. For each prime p not dividing ℓ at which A has good reduction the Frobenius $\sigma_p = \rho_{A,\ell}(\mathrm{Frob}_p)$ is well defined only up

CHAPTER 6

Some other conjectures

We close with some additional comments and conjectures.

6.1. Integral points on elliptic curves

The starting point for any work on integral points on elliptic curves is the following theorem of Siegel.

Theorem 6.1.1 (Siegel)**.** *Every affine model of an elliptic curve over* $\mathbf{Q}$ *has only finitely many points over* $\mathbf{Z}$.

For more recent results on integral points on elliptic curves, see for example [**170**], [**171**], [**48**], [**163**], [**178**], [**169**], [**47**], [**161**], and [**187**].

I learned from [**159**] the statement of the following question:

Question 6.1.2. *Is there is a uniform bound on the number of integral points on minimal (affine) Weierstrass equations for elliptic curves over* $\mathbf{Q}$*?*

The following conjecture of Lang (see Conjecture IX.3.5 of [**156**] and p. 140 of [**78**]) predicts a bound on the number of integral points in terms of the rank of the Mordell-Weil group.

Conjecture 6.1.3 (Lang)**.** *There is an absolute constant* C *such that if* E *is an elliptic curve over* $\mathbf{Q}$ *given by a minimal (affine) Weierstrass equation, then the number of integral points on* E *is at most* $C^{1+\mathrm{rank}(E(\mathbf{Q}))}$.

Silverman [**157**] proved Lang's Conjecture for elliptic curves over $\mathbf{Q}$ with everywhere potentially good reduction (i.e., with integral j-invariant).

See [**137**] for a good discussion of integral points on curves.

6.2. Bounds on numbers of rational points on curves

Faltings [**43**] proved Mordell's Conjecture that if C is a smooth curve of genus $g \geq 2$ defined over a number field K, then $C(K)$ is finite. The following conjectures predict bounds on the number of points.

Caporaso-Harris-Mazur Conjecture ([**22**]). *If K is a number field and $2 \leq g \in \mathbf{Z}$, then there exists a constant $B(K,g)$ such that no smooth curve of genus g defined over K has more than $B(K,g)$ K-rational points.*

Strong Caporaso-Harris-Mazur Conjecture ([**22**]). *If $2 \leq g \in \mathbf{Z}$, then there exists a constant $N(g)$ such that for every number field K there are only finitely many smooth curves of genus g defined over K with more than $N(g)$ K-rational points.*

See also [**23**], where they show that a conjecture of Lang implies the Caporaso-Harris-Mazur Conjecture.

Theorem 6.2.1 (Brumer and Mestre). *For all g,*

$$N(g) \geq 16(g+1).$$

Theorem 6.2.2 (Shioda [**152**]). *For all g,*

$$B(\mathbf{Q},g) \geq 8g+16.$$

Theorem 6.2.3 (Keller-Kulesz [**70**]).

$$B(\mathbf{Q},2) \geq 588, \qquad B(\mathbf{Q},3) \geq 176.$$

Keller and Kulesz showed that the genus 2 curve

$$y^2 = 278,271,081x^2(x^2-9)^2 - 229,833,600(x^2-1)^2$$

has at least 588 rational points, and the genus 3 curve

$$y^2 = 7,920,000(x^2+1)^4 - 136,782,591x^2(x^2-1)^2$$

has at least 176 rational points.

6.3. Conductors of elliptic curves

Let $M(N)$ denote the number of ($\mathbf{Q}$-isomorphism classes of) elliptic curves over $\mathbf{Q}$ of conductor N.

Conjecture 6.3.1 (Brumer-Silverman [**20**]). *With p prime, $M(p)$ is unbounded.*

Brumer and Silverman point out that their conjecture implies the unboundedness of both ranks of elliptic curves over $\mathbf{Q}$ (using Silverman's result bounding the number of integral points in terms of the rank, for elliptic curves with integral j-invariant, and Serre's Theorem 5.3.2 above that if N_E is prime then $\Delta_{\min}(E)$ divides p^5) and 3-ranks of ideal class groups of quadratic fields. Conjecture 6.3.1 would contradict an affirmative answer to Question 6.1.2.

6.4. Abelian varieties, Shimura varieties, and open questions

Abelian varieties are parametrized by Shimura varieties, the way elliptic curves are parametrized by modular curves. As can be seen from the work of Faltings (the Mordell Conjecture) and Wiles (Fermat's Last Theorem), one often cannot remain within the realm of elliptic curves and modular curves, and must use and understand higher dimensional abelian varieties and Shimura varieties. Many results that are known for elliptic curves are still interesting conjectures for abelian varieties. This leaves open wide areas of research for the years to come.

The Mumford-Tate Conjecture (see [**138**], [**39**]) deals with the images of ℓ-adic representations associated to varieties, and is related to the Tate Conjectures [**172**] (which in turn are related to the conjectures of Birch and Swinnerton-Dyer).

As mentioned earlier, see for example [**124**], [**111**], and [**12**] for the conjectures of Beilinson and of Bloch-Kato, which generalize the conjectures of Birch and Swinnerton-Dyer.

There are Modularity Conjectures for elliptic curves over number fields other than $\mathbf{Q}$, and for abelian varieties with real multiplication (or more generally, for abelian varieties of GL_2-type). See for example [**125**], [**58**], and [**59**] (see [**50**] and [**49**] for analogues over function fields). See also Ken Ribet's lectures on Serre's Conjectures [**136**], in this volume.

7)*+,)))-) ' >/ > > 6!
("" % 6 ' F A

6 Energy

The internal energy E is the first term in the free energy, $F = E - TS$. This chapter describes features of E that are important for phase transitions in materials. The internal energy originates from the quantum mechanics of electrons at atoms, specifically the chemical bonds between atoms. The simplest bond is between two atoms in a diatomic molecule, which is developed first to illustrate concepts of bonding, antibonding, electronegativity, covalency, and ionicity.

Crystals have long-range translational symmetry that brings a new quantum number $\vec{k}$ for states of delocalized electrons. The concept of energy bands is presented first by extending the concepts of molecular bonding and antibonding to electron states spread over many atoms. An even simpler model of a gas of free electrons is also developed for electrons in metals.

The strength of bonding depends on the distance between atoms. The interatomic potential of a chemical bond gives rise to elastic constants that characterize how a bulk material responds to small deformations. The chapter ends with a discussion of the elastic energy generated when a particle of a new phase forms inside a parent phase, and the two phases differ in specific volume.

6.1 Atomic Schrödinger Equations and Formalism

Start with two atoms, A and B, initially far apart. The properties of their electrons are calculated independently with two single-atom Schrödinger equations

$$-\frac{\hbar^2}{2m}\nabla^2\psi_{\mathrm{A}}(\vec{r}) + V_{\mathrm{A}}(\vec{r})\,\psi_{\mathrm{A}}(\vec{r}) = \epsilon_{\mathrm{A}}\psi_{\mathrm{A}}(\vec{r}), \tag{6.1}$$

$$-\frac{\hbar^2}{2m}\nabla^2\psi_{\mathrm{B}}(\vec{r}) + V_{\mathrm{B}}(\vec{r})\,\psi_{\mathrm{B}}(\vec{r}) = \epsilon_{\mathrm{B}}\psi_{\mathrm{B}}(\vec{r}). \tag{6.2}$$

At isolated atoms, the single-electron states (denoted $|\mathrm{A}\rangle$ and $|\mathrm{B}\rangle$ in Dirac notation) have wavefunctions ψ_{A} and ψ_{B}, and energies ϵ_{A} and ϵ_{B}. The potentials $V_{\mathrm{A}}(\vec{r})$ and $V_{\mathrm{B}}(\vec{r})$ are centered at the two separated nuclei. Each Laplacian, ∇^2, acts on the coordinates of one electron, wherever it is. To be more explicit, sometimes the Laplacian is given a subscript identifying the particular electron, e.g., the kinetic energy of electron number 1 is written $-(\hbar^2/2m)\,\nabla_1^2$, and electron number 2 has the kinetic energy $-(\hbar^2/2m)\,\nabla_2^2$.

We are interested in how the states for single electrons become something new when the two nuclei are brought together close enough so the wavefunctions of their atomic states overlap slightly. In a short time, an individual electron will have amplitude at both atoms, and a molecule is formed. A single-electron wavefunction for the diatomic molecule is called a "molecular orbital." A state for one electron now needs to include the effects of both nuclei and their core electrons, since the electron spends time about both. The potential is a real challenge for several reasons, primarily because the potential for one electron depends on the presence of other electrons.[1] We begin, however, by ignoring possible changes in potential, and we assume that the total potential for the molecule is simply the sum of potentials of the isolated atoms

$$V_{\mathrm{M}}(\vec{r}) = V_{\mathrm{A}}(\vec{r}) + V_{\mathrm{B}}(\vec{r}). \tag{6.3}$$

Another assumption, a guess really, is that the molecular orbital can be constructed as a weighted sum of the atomic orbitals as

$$|\mathrm{M}\rangle = c_{\mathrm{A}}|\mathrm{A}\rangle + c_{\mathrm{B}}|\mathrm{B}\rangle, \tag{6.4}$$

$$\psi_{\mathrm{M}}(\vec{r}) = c_{\mathrm{A}}\psi_{\mathrm{A}}(\vec{r}) + c_{\mathrm{B}}\psi_{\mathrm{B}}(\vec{r}), \tag{6.5}$$

written in Dirac notation (Eq. 6.4) and with explicit wavefunctions (Eq. 6.5). It is important to remember that the molecular state $|\mathrm{M}\rangle$ accommodates one electron, so the coefficients c_{A} and c_{B} are less than 1 (the atomic states $|\mathrm{A}\rangle$ and $|\mathrm{B}\rangle$ accommodate one electron each). The wavefunctions $\psi_{\mathrm{A}}(\vec{r})$ and $\psi_{\mathrm{B}}(\vec{r})$ are centered at different nuclei, but an electron can spend time in both of them (in states $|\mathrm{A}\rangle$ and $|\mathrm{B}\rangle$) when the two nuclei are brought near each other.

Box 6.1

Dirac Notation

Many calculations in quantum mechanics involve integrations over regions containing electron wavefunctions. Working with conventional integral formalism and its coordinates leads to bulky equations where it can be hard to see the forest through the trees. Paul Dirac dispensed with the coordinates altogether with a clever and compact notation. The quantum state of an electron 1 at atom A, having the wavefunction $\psi_{\mathrm{A}}(\vec{r}_1)$, is denoted $|A\rangle$ in Dirac notation. Without coordinates, this state function is not the wavefunction, but the two are related by projecting the quantum state onto spatial coordinates (i.e., finding the presence of the state $|A\rangle$ at $\vec{r}_1$). This projection and its complex conjugate are written as

$$\psi_{\mathrm{A}}(\vec{r}_1) = \langle \vec{r}_1|\mathrm{A}\rangle, \qquad \psi^*_{\mathrm{A}}(\vec{r}_1) = \langle \mathrm{A}|\vec{r}_1\rangle. \tag{6.6}$$

A typical calculation of orthogonality or normalization takes the form

$$\int_{-\infty}^{\infty} \psi^*_{\mathrm{A}}(\vec{r}_1)\,\psi_{\mathrm{A}}(\vec{r}_1)\,\mathrm{d}^3\vec{r}_1 = \int_{-\infty}^{\infty} \langle A|\vec{r}_1\rangle\langle\vec{r}_1|\mathrm{A}\rangle\,\mathrm{d}^3\vec{r}_1 = \langle A|\int_{-\infty}^{\infty} |\vec{r}_1\rangle\langle\vec{r}_1|\mathrm{A}\rangle\,\mathrm{d}^3\vec{r}_1. \tag{6.7}$$

---

[1] This is, however, a problem already faced when calculating energies of electrons in a multi-electron atom. The effect of a second electron is to push around the first electron, but this also alters the potential and wavefunction of the second electron. Iterative methods are useful for these types of problems.

One interpretation of the integral on the RHS is that it means projecting $|\mathrm{A}\rangle$ onto spatial coordinates for electron 1 as $\langle\vec{r}_1|\mathrm{A}\rangle$, using these projections to weight $|\vec{r}_1\rangle$, and then summing (integrating) over all $\vec{r}_1$. This reconstructs the state $|\mathrm{A}\rangle$ from the $|\vec{r}_1\rangle$, so Eq. 6.7 takes the compact form[a]

$$\int_{-\infty}^{\infty} \psi_{\mathrm{A}}^*(\vec{r}_1)\psi_{\mathrm{A}}(\vec{r}_1)\,\mathrm{d}^3\vec{r}_1 = \langle\mathrm{A}|\mathrm{A}\rangle. \tag{6.8}$$

The form $\langle\ |\ \rangle$ denotes a definite integral, typically with limits over all space. The same is true of expectation values of operators. For example

$$\int_{-\infty}^{\infty} \psi_{\mathrm{A}}^*(\vec{r}_1)\frac{-\mathrm{e}^2}{r_1}\psi_{\mathrm{B}}(\vec{r}_1)\,\mathrm{d}^3\vec{r}_1 = \langle\mathrm{A}|\frac{-\mathrm{e}^2}{r_1}|\mathrm{B}\rangle, \tag{6.9}$$

or for another example:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi_{\mathrm{A}}^*(\vec{r}_1)\psi_{\mathrm{B}}^*(\vec{r}_2)H'(\vec{r}_2)\psi_{\mathrm{B}}(\vec{r}_1)\psi_{\mathrm{A}}(\vec{r}_2)\,\mathrm{d}^3\vec{r}_1\,\mathrm{d}^3\vec{r}_2 = \langle\mathrm{A}|\mathrm{B}\rangle_1\langle\mathrm{B}|H'|\mathrm{A}\rangle_2, \tag{6.10}$$

where the electron subscripts on the RHS of Eqs. 6.9 and 6.10 are usually ignored.

Dirac notation allows rapid and lucid mathematical manipulations. In the end, however, calculating integrals requires thought about the best coordinate system, and typically multiple integrations over three spatial variables.

[a] The coordinates form a complete set for expressing any function of position. For complete sets we can invoke the property of "closure," defined with an integral or a sum as, for example, $\int_{-\infty}^{\infty} |\vec{r}\rangle\langle\vec{r}|\,\mathrm{d}^3\vec{r} = 1$ or $\sum_i |i\rangle\langle i| = 1$.

6.2 Molecular Orbital Theory of Diatomic Molecules

6.2.1 Molecular Schrödinger Equation

The Schrödinger equation for a single electron in the molecule is

$$H|\mathrm{M}\rangle = \epsilon|\mathrm{M}\rangle, \tag{6.11}$$

where the Hamiltonian includes a kinetic energy operator for one electron, and the potential for this electron, the $V_{\mathrm{M}}(\vec{r})$ of Eq. 6.3. Two key operations are now done to convert the RHS of the Schrödinger equation 6.11 into a real energy ϵ without the wavefunction:

1. multiply Eq. 6.11 by $\langle\mathrm{A}|$, and make a second equation by multiplying Eq. 6.11 by $\langle\mathrm{B}|$;
2. for both equations, integrate over the coordinates of the electron.

Doing both steps simultaneously with Dirac notation

$$\langle\mathrm{A}|H|\mathrm{M}\rangle = \epsilon\langle\mathrm{A}|\mathrm{M}\rangle, \tag{6.12}$$

$$\langle\mathrm{B}|H|\mathrm{M}\rangle = \epsilon\langle\mathrm{B}|\mathrm{M}\rangle. \tag{6.13}$$

Using Eq. 6.4 for $|\mathrm{M}\rangle$

$$\langle \mathrm{A}|H|\Big(c_\mathrm{A}|\mathrm{A}\rangle + c_\mathrm{B}|\mathrm{B}\rangle\Big) = \epsilon\langle \mathrm{A}|\Big(c_\mathrm{A}|\mathrm{A}\rangle + c_\mathrm{B}|\mathrm{B}\rangle\Big), \tag{6.14}$$

$$\langle \mathrm{B}|H|\Big(c_\mathrm{A}|\mathrm{A}\rangle + c_\mathrm{B}|\mathrm{B}\rangle\Big) = \epsilon\langle \mathrm{B}|\Big(c_\mathrm{A}|\mathrm{A}\rangle + c_\mathrm{B}|\mathrm{B}\rangle\Big), \tag{6.15}$$

Moving the constants c_A and c_B through the integrals

$$c_\mathrm{A}\,\langle \mathrm{A}|H|\mathrm{A}\rangle + c_\mathrm{B}\,\langle \mathrm{A}|H|\mathrm{B}\rangle = \epsilon\, c_\mathrm{A}\,\langle \mathrm{A}|\mathrm{A}\rangle + \epsilon\, c_\mathrm{B}\,\langle \mathrm{A}|\mathrm{B}\rangle, \tag{6.16}$$

$$c_\mathrm{A}\,\langle \mathrm{B}|H|\mathrm{A}\rangle + c_\mathrm{B}\,\langle \mathrm{B}|H|\mathrm{B}\rangle = \epsilon\, c_\mathrm{A}\,\langle \mathrm{B}|\mathrm{A}\rangle + \epsilon\, c_\mathrm{B}\,\langle \mathrm{B}|\mathrm{B}\rangle. \tag{6.17}$$

Definitions and Approximations

The atomic states $|\mathrm{A}\rangle$ and $|\mathrm{B}\rangle$ are normalized to accommodate one electron, so

$$\langle \mathrm{A}|\mathrm{A}\rangle = \langle \mathrm{B}|\mathrm{B}\rangle = 1. \tag{6.18}$$

Although coordinates are not included in Dirac notation, the wavefunctions $\psi_\mathrm{A}(\vec{r})$ and $\psi_\mathrm{B}(\vec{r})$ are physically centered at different nuclei. The tails of the wavefunction for $|\mathrm{A}\rangle$ overlap the tails of the wavefunction for $|\mathrm{B}\rangle$. We define the overlap integral S

$$S \equiv \langle \mathrm{A}|\mathrm{B}\rangle, \tag{6.19}$$

$$S \equiv \int_{-\infty}^{+\infty} \psi_\mathrm{A}^*(\vec{r})\;\psi_\mathrm{B}(\vec{r})\;\mathrm{d}^3\vec{r}. \tag{6.20}$$

For now, for simplicity we set $S = 0$ (expecting that S is small). A closely related approximation is that the tails of the wavefunction for $|\mathrm{A}\rangle$ overlap only weakly with the potential energy from the second nucleus at B, so we neglect energies like $\langle \mathrm{A}|V_\mathrm{B}|\mathrm{A}\rangle$. This allows the approximations

$$\langle \mathrm{A}|H|\mathrm{A}\rangle \simeq \epsilon_\mathrm{A}, \tag{6.21}$$

$$\langle \mathrm{B}|H|\mathrm{B}\rangle \simeq \epsilon_\mathrm{B}, \tag{6.22}$$

where ϵ_A and ϵ_B are the energies of electrons in states at the isolated atoms A and B. With these simplifications, Eqs. 6.16 and 6.17 become

$$c_\mathrm{A}\,\epsilon_\mathrm{A} + c_\mathrm{B}\,\langle \mathrm{A}|H|\mathrm{B}\rangle = \epsilon\, c_\mathrm{A}, \tag{6.23}$$

$$c_\mathrm{A}\,\langle \mathrm{B}|H|\mathrm{A}\rangle + c_\mathrm{B}\,\epsilon_\mathrm{B} = \epsilon\, c_\mathrm{B}. \tag{6.24}$$

The important physical quantity that remains after our manipulations and approximations is the “bond integral,” h

$$h \equiv \langle \mathrm{A}|H|\mathrm{B}\rangle, \tag{6.25}$$

$$h \equiv \int_{-\infty}^{+\infty} \psi_\mathrm{A}^*(\vec{r})\;H\;\psi_\mathrm{B}(\vec{r})\;\mathrm{d}^3\vec{r}. \tag{6.26}$$

Energies of Molecular States

Some consequences of the bond integral h are explored below, but first we complete the formal effort to obtain the electron energies in the molecule, ϵ. With Eq. 6.25 for h, the pair of Eqs. 6.23 and 6.24 is written as one compact matrix equation

$$\begin{bmatrix} \epsilon_\mathrm{A} - \epsilon & h \\ h & \epsilon_\mathrm{B} - \epsilon \end{bmatrix} \begin{bmatrix} c_\mathrm{A} \\ c_\mathrm{B} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{6.27}$$

It proves convenient to replace ϵ_A and ϵ_B with their average $\overline{\epsilon} = (\epsilon_\mathrm{A} + \epsilon_\mathrm{B})/2$ and the difference $\Delta = (\epsilon_\mathrm{B} - \epsilon_\mathrm{A})/2$ so

$$\epsilon_\mathrm{A} = \overline{\epsilon} - \Delta \quad , \quad \epsilon_\mathrm{B} = \overline{\epsilon} + \Delta, \tag{6.28}$$

$$\begin{bmatrix} \overline{\epsilon} - \Delta - \epsilon & h \\ h & \overline{\epsilon} + \Delta - \epsilon \end{bmatrix} \begin{bmatrix} c_\mathrm{A} \\ c_\mathrm{B} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{6.29}$$

The matrix equation 6.29 has a useful solution when the determinant is zero, giving the secular equation

$$(\overline{\epsilon} - \Delta - \epsilon)(\overline{\epsilon} + \Delta - \epsilon) - h^2 = 0, \tag{6.30}$$

$$(\epsilon - \overline{\epsilon})^2 = \Delta^2 + h^2, \tag{6.31}$$

$$\epsilon^\pm = \overline{\epsilon} \mp \sqrt{\Delta^2 + h^2}, \tag{6.32}$$

where the two values of ϵ are labeled as ϵ^+ and ϵ^-. Compared to the average of the atomic energies, $\overline{\epsilon}$, one molecular state has an energy higher by the amount $\sqrt{\Delta^2 + h^2}$, and the other has an energy lower by the amount $\sqrt{\Delta^2 + h^2}$. These states are naturally called "antibonding" (with higher energy ϵ^-) and "bonding" (with lower energy ϵ^+), respectively.

6.2.2 Molecular Wavefunctions of Bonding and Antibonding States

Homonuclear Molecular Orbitals

We now solve for the c_A and c_B in Eq. 6.4, from which we can obtain the electron density in the bonding and antibonding molecular states, or "molecular orbitals." For homonuclear molecular orbitals involving identical atoms, for which $\epsilon_\mathrm{B} = \epsilon_\mathrm{A}$ so $\Delta = 0$, Eq. 6.32 gives $\epsilon^\pm = \overline{\epsilon} \mp |h|$. Substituting $\epsilon^\pm$ for ϵ in Eq. 6.29

$$\begin{bmatrix} \pm|h| & h \\ h & \pm|h| \end{bmatrix} \begin{bmatrix} c_\mathrm{A}^\pm \\ c_\mathrm{B}^\pm \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{6.33}$$

The first line of Eq. 6.33 gives

$$\pm|h|\, c_\mathrm{A}^\pm + h\, c_\mathrm{B}^\pm = 0, \tag{6.34}$$

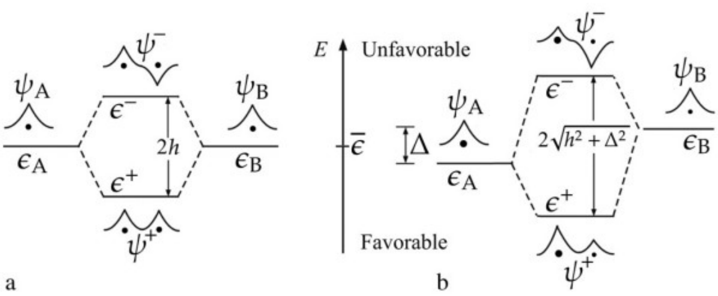

Figure 6.1 Energies and wavefunctions of molecular orbital formation in diatomic molecules with atom A on left and B on right in both panels a and b. Energy scale and definitions are in the middle. (**a**) Homonuclear molecule, such as H_2. With identical atomic energies and wavefunctions, the one-electron molecular states are symmetric. (**b**) Heteronuclear molecule, such as HF. With $\epsilon_B > \epsilon_A$, the bonding level ψ^+ is lower because it puts more of the electron in the favorable atomic wavefunction about atom A. The antibonding orbital ψ^- puts more electron in the unfavorable atomic wavefunction about atom B.

and since h is negative (bond energies are favorable)

$$c_A^{\pm} = \pm c_B^{\pm}. \tag{6.35}$$

For the bonding state with $\{\epsilon^+, c_A^+, c_B^+\}$, we obtain $c_A^+ = c_B^+$.
For the antibonding state with $\{\epsilon^-, c_A^-, c_B^-\}$, we obtain $c_A^- = -c_B^-$.

The two molecular states for single electrons, $|M^+\rangle$ and $|M^-\rangle$, are depicted in Fig. 6.1 as molecular orbital wavefunctions $\psi^+(\vec{r})$ and $\psi^-(\vec{r})$. The two panels a and b show bonding of two identical atoms (homonuclear), and bonding of two different atoms (heteronuclear). When identical atoms are brought together ($\Delta = (\epsilon_B - \epsilon_A)/2 = 0$), the bonding energy ϵ^+ and the antibonding energy ϵ^- are favored or disfavored over the isolated atoms by $\pm h$, where h is the bond integral of Eq. 6.25 (set $\Delta = 0$ in Eq. 6.32).

The $c_A^{\pm}$ and $c_B^{\pm}$ need normalization to 1 because there is one electron in either the bonding or antibonding molecular orbital

$$1 = \langle M^+|M^+\rangle = \left(c_A^+\langle A| + c_B^+\langle B|\right)\left(c_A^+|A\rangle + c_B^+|B\rangle\right), \tag{6.36}$$

$$1 = \langle M^-|M^-\rangle = \left(c_A^-\langle A| + c_B^-\langle B|\right)\left(c_A^-|A\rangle + c_B^-|B\rangle\right), \tag{6.37}$$

and with $c_A^+ = c_B^+$ and $c_A^- = -c_B^-$

$$1 = \langle M^+|M^+\rangle = c_A^{+2}\langle A|A\rangle + c_A^{+2}\langle A|B\rangle + c_A^{+2}\langle B|A\rangle + c_A^{+2}\langle B|B\rangle, \tag{6.38}$$

$$1 = \langle M^-|M^-\rangle = c_A^{-2}\langle A|A\rangle - c_A^{-2}\langle A|B\rangle - c_A^{-2}\langle B|A\rangle + c_A^{-2}\langle B|B\rangle. \tag{6.39}$$

Using the overlap integral S of Eq. 6.19 for $\langle A|B\rangle$ and $\langle B|A\rangle$

$$c_A^+ = +c_B^+ = \frac{1}{\sqrt{2}\,\sqrt{1+S}}, \tag{6.40}$$

$$c_A^- = -c_B^- = \frac{1}{\sqrt{2}\,\sqrt{1-S}}. \tag{6.41}$$

Comparing Eqs. 6.40 and 6.41, it is seen that with increasing overlap S, the bonding wavefunction needs some reduction in amplitude, and the antibonding some increase in amplitude. This can be understood from the electron density in the regions of overlap, as indicated schematically for ψ^+ and ψ^- in Fig. 6.1a. Midway between the two atoms, the wavefunction amplitude is doubled for the bonding wavefunction ψ^+, compared with the wavefunction in the tail of an atom wavefunction. It is zero in the middle of the antibonding wavefunction ψ^-. Obtaining the electron density as $\psi^*\psi$, the density for the bonding ψ^+ is increased by a factor of 2 over that for two independent atoms, whereas the density for the antibonding ψ^- is zero. Normalization of the whole wavefunction to 1 therefore decreases the total amplitude of the bonding wavefunction and increases it for the antibonding.

Heteronuclear Molecular Orbitals

The bonding energy increases with Δ, as given in Eq. 6.32 and shown in Fig. 6.1b. This Δ is a difference of atom properties, ϵ_A and ϵ_B. This difference is specified as the difference in "electronegativity," χ. The larger is χ, the more favorable are the single electron states at the atom. Our results of Eq. 6.32 show that the larger the difference in electronegativity $|\chi_A - \chi_B|$, the stronger the bonding energy. On the other hand, with increasing $|\chi_A - \chi_B|$ the antibonding state becomes even more unfavorable compared to the homonuclear molecule. As discussed further in Section 6.2.4, Fig. 6.1b illustrates what happens – the bonding level puts more electron density on the atom with the larger χ (and lower ϵ_A). The antibonding molecular orbital puts more electron density on the less favorable atom with the smaller χ. The difference between bonding and antibonding energies grows with Δ, or with increasing $|\chi_A - \chi_B|$.

6.2.3 Origin of the Bond Energy

A common argument is that the higher density of electrons in the region between the atoms, where the electron can see both nuclei with their attractive potentials, is why the bonding level is more favorable energetically. Unfortunately, this argument has a fundamental flaw. The electron density is normalized, so putting more electron density between the nuclei must remove electron density from other parts of the molecule. Indeed, Eqs. 6.40 and 6.41 show that $c_A^+ < c_A^-$, for example. Normalization removes electron density from nearest the nuclei where the potential energy is most favorable, and places it where the potential is modestly favorable. The consequence is that the antibonding level sees an increase in electron density at the nuclei, and the bonding level a decrease. The potential energy actually favors the antibonding level over the bonding level for molecular orbital theory with atomic wavefunctions.

The first step to repair this embarrassing situation is to include the electron kinetic energy. Here the disadvantage of the antibonding wavefunction is more obvious. Its node between the nuclei requires that the electron wavefunction undergoes a more rapid oscillation, increasing the kinetic energy because of a larger $\nabla^2\psi$. In contrast, the bonding wavefunction is smoother in this region between the nuclei, so it has the lower kinetic energy. This explanation of bonding is now more correct.

Molecular wavefunctions of the H_2^+ molecular ion have been calculated accurately, since there is only one electron, and the complicated electron–electron interactions are absent. Both the electron potential energy (which decreases when the electron is bunched near the nucleus) and the electron kinetic energy (which increases when the electron is confined) need to be considered in detail, for both the bonding and antibonding states. One should also consider the neglected overlap energies such as $\langle A|V_B|A\rangle$, which destabilize the antibonding state and favor the bonding state. For atomic *s*-electron wavefunctions it is found, perhaps not surprisingly, that atomic wavefunctions are inadequate even for a qualitative explanation of the chemical bond. The antibonding state is still favored slightly over the bonding state for most interatomic separations. What needs to be done is alter the shapes of $\psi_A(\vec{r})$ and $\psi_B(\vec{r})$ to minimize the energies of the bonding and antibonding states. This can be done fairly well by changing the characteristic length for the exponential decay of the atomic *s*-electron wavefunctions. By making a more rapid decay with distance of $\psi_A(\vec{r})$ and $\psi_B(\vec{r})$ for the bonding state, it is possible to place more electron density at the nuclei. The electron kinetic energy is increased a bit because the wavefunction is more confined, but the potential energy becomes much more favorable, and the bonding state does become bonding. The antibonding wavefunctions tend to lower their energy if the atomic *s*-states are expanded a bit, reducing the kinetic energy somewhat.[2] The real situation points out the risk of relying too heavily on analytical models, even for calculating bonding and antibonding states of the simplest molecule, H_2^+.

6.2.4 Ionicity and Electronegativity

For a heteronuclear molecule the two atoms differ in electronegativity, so $\Delta = (\epsilon_B - \epsilon_A)/2 \neq 0$. We expect unequal electron densities on each atom, so $|c_A| \neq |c_B|$. Expressions analogous to Eqs. 6.40 and 6.41 are obtained directly from Eqs. 6.33. Approximately[3]

$$c_A^{\pm} = \frac{1}{\sqrt{2}} \sqrt{1 \pm \frac{\delta - S}{\sqrt{1+\delta^2}}}, \tag{6.42}$$

$$c_B^{\pm} = \pm\frac{1}{\sqrt{2}} \sqrt{1 \mp \frac{\delta + S}{\sqrt{1+\delta^2}}}, \tag{6.43}$$

where the new variable

$$\delta \equiv \frac{\Delta}{|h|} \tag{6.44}$$

---

[2] A further improvement can be obtained by optimizing the atomic wavefunctions at each interatomic separation, but here the result is not changed qualitatively.

[3] When $\delta = 0$, there is a small difference between Eqs. 6.40 and 6.41 and Eqs. 6.42 and 6.43, since in Eqs. 6.42 and 6.43 some terms of order S^2 were neglected, and normalization was not done directly.

scales with the electronegativity difference of the atoms. Assuming the A-atom has the higher electronegativity as in Fig. 6.1b, electron density will transfer from B to A. We seek to write the molecular charge density $\rho = \psi^* \psi$ in the form

$$\rho(\vec{r}) = (1 + \alpha_\mathrm{i})\rho_\mathrm{A}(\vec{r}) + (1 - \alpha_\mathrm{i})\rho_\mathrm{B}(\vec{r}) + \alpha_\mathrm{c}\rho_\mathrm{bond}(\vec{r}), \tag{6.45}$$

where α_i is a fractional charge transfer caused by the ionic character of the bond (the difference in electronegativity), and α_c weights the charge density of covalent bonding that was considered in the previous section for a homonuclear molecule such as H_2.The charge redistribution upon bonding, $\rho_\mathrm{bond}(\vec{r})$, is defined with respect to the charge already present around two overlapping but independent atoms

$$\rho_\mathrm{bond}(\vec{r}) = 2\psi_\mathrm{A}(\vec{r})\psi_\mathrm{B}(\vec{r}) - S\Big[\rho_\mathrm{A}(\vec{r}) + \rho_\mathrm{B}(\vec{r})\Big], \tag{6.46}$$

where $\rho_\mathrm{A} = \psi^*_\mathrm{A}(\vec{r})\psi_\mathrm{A}(\vec{r})$ and $\rho_\mathrm{B} = \psi^*_\mathrm{B}(\vec{r})\psi_\mathrm{B}(\vec{r})$. Both ψ_A and ψ_B were assumed real in Eq. 6.46.

It turns out that

$$\alpha_\mathrm{i} = \frac{\delta}{\sqrt{1+\delta^2}}, \tag{6.47}$$

$$\alpha_\mathrm{c} = \frac{1}{\sqrt{1+\delta^2}}, \tag{6.48}$$

and $\alpha_\mathrm{i}^2 + \alpha_\mathrm{c}^2 = 1$. When $\delta = 0$, so the atoms have the same electronegativity, $\alpha_\mathrm{i} = 0$ and $\alpha_\mathrm{c} = 1$. This is a purely covalent bond.

When $\delta > 1$ and the effects of electronegativity are dominant, $\alpha_\mathrm{i} > \alpha_\mathrm{c}$, and the bond is mostly ionic. Assuming that these bonding tendencies are characteristics of atoms, or differences between atoms, it is appropriate to assign an electronegativity to each chemical element. Measures of electronegativity have been defined by different approaches, but the general trends of different electronegativity scales tend to be in agreement. It is a challenge, however, to account for bonding between different types of outer electrons of different atoms. Furthermore, there is a fundamental problem with all electronegativity scales – they predict that all pairs of chemical elements will form compounds. This is not so, and proper explanations of compound formation require concepts and physical properties that cannot be assigned to individual atoms alone.

6.3 Electronic Bands and the Tight-Binding Model

6.3.1 Translational Symmetry and Phase Factors

Translational symmetry is the fundamental difference between a molecule and a crystal, and translational symmetry has a profound effect on the spectrum of electronic energies in crystals. Neglecting surfaces, the electron density must be the same if position is shifted by a repeat distance of the crystal, i.e., a translation vector of the lattice. This does not apply, of course, to an isolated diatomic molecule.

6

e

A very special number

You have no idea how much poetry there is in the calculation of a table of logarithms.

—Carl Friedrich Gauss (1777–1855)

Prologue

As discussed in chapter 4, π is a transcendental number. Another significant transcendental number is e, which is equal to 2.71828.... The relation and importance of these two numbers is brought out eloquently by Ian Stewart (2008, 101):

> The number e is one of those strange special numbers that appear in mathematics, and have major significance. Another such number is π. These two numbers are the tip of an iceberg—there are many others. They are also arguably the most important of the special numbers because they crop up all over the mathematical landscape.

In correspondence between Gottfried Leibniz and Christian Huygens, there is mention of a constant with value 2.71828, which they represented with the letter b. It was Leonhard Euler who used e in a 1731 letter to Christian Goldbach, and then again in his 1736 book, *Mechanica.* From then on, this has been the symbol used as the standard one in mathematics. The mathematical definition of e is: the limiting value of the series represented by the expression $(1 + 1/n)^n$, as n becomes large without bound. A few values are given in Table 6.1.

Table 6.1 ***Values for (1 + 1/n)ⁿ***

n	$(1 + 1/n)^n$	*Value*
1	$(1 + 1/1)^1$	2.0
2	$(1 + 1/2)^2$	2.25
5	$(1 + 1/5)^5$	2.49
10	$(1 + 1/10)^{10}$	2.59
100	$(1 + 1/100)^{100}$	2.70
…	…	…
10,000	$(1 + 1/10{,}000)^{10{,}000}$	2.71815

Another formula used by Euler to determine the value of *e* (and there are a significant number of them) is the following one. Note that "!" means factorial, the product of an integer and all the integers below it: 4! = 4 × 3 × 2 × 1, 3! = 3 × 2 × 1, etc.

$$e = 1/0! + 1/1! + 1/2! + 1/3! + 1/4! + 1/5! + 1/6! + 1/7! \ldots$$

Equivalent to:

$$e = 1 + 1 + 1/2 + 1/6 + 1/24 + 1/120 + 1/720 + 1/5040 \ldots$$

Equivalent to:

$$e = 1 + 1 + 0.5 + 0.16 + 0.042 + 0.008 + 0.00138 + 0.000198 \ldots$$

Now, one may ask: What possible significance does such a number have? As it turns out, it forms the base of natural logarithms, which have had broad implications and applications; it appears in functions associated with the calculus; it surfaces in formulas for various kinds of curves; it crops up frequently in probability theory; it appears in the calculation of compound interest; and it has been found in natural and social phenomena such as the following:

1. radioactive decay
2. the growth of bacterial colonies
3. epidemiological patterns of all kinds
4. patterns in the accumulation of money
5. rates of change in many physical phenomena

Even though the value of 2.71828 appeared in an appendix in John Napier's 1618 book on logarithms (chapter 5), the first awareness of its mathematical significance is by English mathematician and clergyman William Oughtred (1574–1660), who defined it as the base of the *natural logarithm*. Here's what this means, recalling logarithms from the previous chapter:

> The natural logarithm of a number such as 7.389 is defined as: $\log_e$ (7.389), also written as: ln (7.389), which is approximately 2.

There is little doubt that this number has enormous significance, although the reasons why this is so remain as obscure as they are for π. This chapter deals with *e*. So many books, articles, research papers, websites, etc. have been written on this subject that it would be presumptuous and even pretentious to claim originality and even depth of coverage here (see bibliography for relevant titles). The purpose is simply to link *e* to other great ideas, showing how mathematical discovery forms a chain—a chain constructed with a handful of great ideas that appear across time and space, finding form and explanation in the writings and musings of individual mathematicians.

Mathematical connectivity

The number *e* is connected to many other concepts, structures, and patterns within mathematics, suggesting that it may be yet another kind of code for establishing interrelationships among mathematical entities. A significant property of *e* is that it is associated with various infinite series, thus linking it to the concept of mathematical infinity (to be discussed in chapter 8).

The Swiss mathematician Jacob Bernoulli (1654–1705) came up with the value of *e* by studying compound interest. He noted that a 100% increase when paid out annually results in the sum doubling year after year. However, splitting the 100% over shorter periods results in more than doubling the return at the end of the year. Bernoulli estimated what the payout would be if the interest was calculated continuously over infinitely small periods—which can be represented as an exponential function. The answer he discovered was that it grew at the annual rate of 2.71828…, which is the value of *e*. Soon after, similar findings emerged from diverse fields that involve growth or change, including epidemiology, the study of bacterial infections, and the like. Incidentally, Bernoulli's estimation of *e* involved thinking of *e* as resulting from an infinite series. As one compounds more and more interest, the value of one's investments approach a limit. Bernoulli thus defined *e* in terms of an infinite series that converges to a limit—that is, we can get closer and closer to the limit, but never quite reach it.

Euler was the first mathematician to study e in depth, coming up with the formula for an infinite series mentioned above in his 1748 book, *Introductio in analysin infinitorum* ("Introduction to Infinite Analysis"). Euler also noticed something intriguing, namely that e^x, now called an exponential function, had intriguing properties. One formula for this function is the so-called Taylor Series, formulated by British mathematician Brook Taylor in 1715:

$$e^x = 1 + x/1! + x^2/2! + x^3/3! + x^4/4! \ldots$$

Equivalent to:

$$e^x = 1 + x + x^2/2 + x^3/6 + x^4/24 \ldots$$

One of the most surprising findings regarding this function is that in differentiation none of the terms disappear, which they normally do. Rather, they repeat ad infinitum. Differentiation is an operation in the calculus. It allows us to determine the derivative of a function, which measures the slope of the graph of the function at a particular point on the graph. In other words, the derivative is a measure of change in the value of the function as the independent variable changes. This is yet another exceptional finding, implying that e^x is its own derivative and, thus, that the graph of e^x describes its own rate of change. Figure 6.1 is the graph showing how the slope rises dramatically.

So, more technically, if $f(x) = e^x$, then $dy/dx = e^x$ (where dy/dx is the symbol for differentiation). The negative exponential function, $f(x) = e^{-x}$, shows a slope

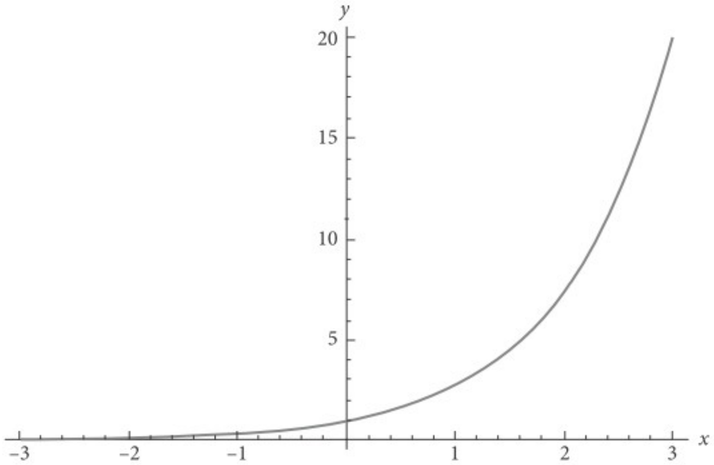

Figure 6.1 *The Exponential Function* e^x (from Wikimedia Commons)

in the opposite direction, thus visually demonstrating the relation between positive and negative exponents in the study of functions—another amazing discovery (see Figure 6.2).

Yet another serendipitous appearance of e^x is in the logarithmic spiral shown in Figure 6.3.

In this spiral, the distance from the center O is $1/e^x$ where x is the angle that results from the spiral turn. As x gets larger, $1/e^x$ becomes smaller. It was Jacob Bernoulli who discovered this remarkable property, and, given that spirals occur throughout nature, this is a rather important one. So surprised was Bernoulli by this finding that he called it a *spira mirabilis*, a miraculous spiral. Research on this figure has revealed a host of other hidden patterns. Nature is replete with logarithmic spirals having a similar mathematical structure.

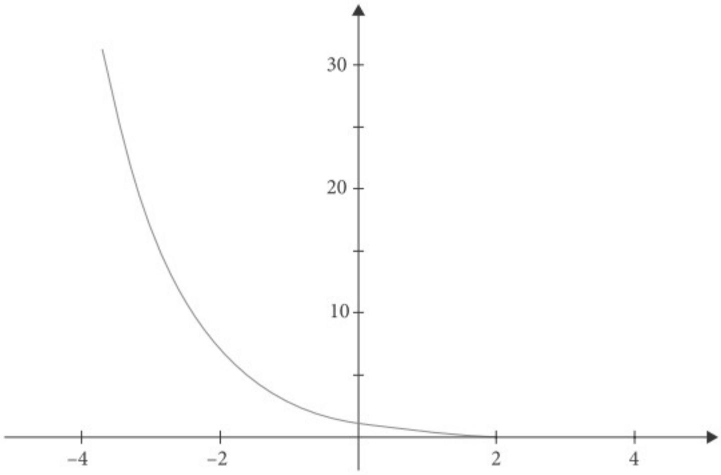

Figure 6.2 *The Exponential Function*, e^{-x} (from Wikimedia Commons)

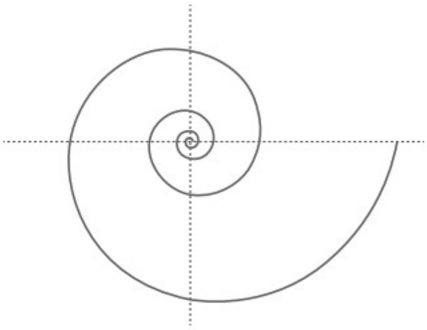

Figure 6.3 *The Logarithmic Spiral* (from Wikimedia Commons)

Bernoulli was so thrilled by his discovery that he wanted to be commemorated in death with it. As it turned out, an error was made, as Freiberger and Thomas (2014, 92) quip:

> Bernoulli was so fascinated by this self-similarity that he planned to have a logarithmic spiral engraved on his tombstone, along with the words (in Latin) 'Although changed, I shall arise the same.' Unfortunately, though, the wrong type of spiral, an Archimedean one whose loops are an equal distance apart, ended up adorning his grave. Poor Bernoulli, he must have turned (logarithmically) in his grave. Perhaps he should have gone for a simpler shape, the trusty old triangle.

The function e^x is a special one—called the *natural exponential function*. In general, an exponential function is one that has the form: $f(x) = a^x$, that is, a function in which the base is constant and the variable (x) is an exponent. Exponential functions have had many implications, not only in pure mathematics, but in in the following fields:

1. The study of radioactivity, since radioactive substances decay at an exponential rate.
2. Population analysis, since if unchecked, a population (organisms, people, cells, etc.) will grow at a rate proportional to itself.
3. In electrical engineering, since the theory of alternating current relies on the exponential function.

The connection between domains of science and exponential functions is something that is truly amazing, showing that mathematics and science are intrinsically intertwined. Indeed, mathematics is rightfully called the language of science. It has also been found that logarithmic functions occur in many areas of physics and biology. For example, the logarithm of measures of living tissue (skin area, weight), lengths of inert appendages (hair, claws), and certain physiological measures such as blood pressure all display a normal distribution. In economic theory, changes in the logarithms of exchange rates, as well as price and stock market indices also display the same type of distribution. The natural logarithm is also used by Google to give every page on the World Wide Web a score (PageRank), which is a rough measure of importance based on a logarithmic scale.

Euler's identity

Working with various exponential functions, Euler discovered an intriguing formula, known appropriately as Euler's formula, where "*i*" is an imaginary number (see chapter 7):

$$e^{i\varphi} = \cos \varphi + i \sin \varphi$$

The sine (shortened to *sin*) of an acute angle is defined in terms of a right triangle: it is the ratio of the length of the side that is opposite a given angle to the length of the longest side of the triangle (the hypotenuse). The cosine (shortened to *cos*) is the ratio of the length of the adjacent side to a given angle divided by the hypotenuse. The above formula can be plotted on the complex number plane as shown in Figure 6.4 below—complex numbers will be examined in the next chapter (Im = imaginary; Re = real):

Now, let's bring π into the picture. It is known in trigonometry that:

$$\sin \pi = 0$$

$$\cos \pi = -1$$

Let's replace φ with π in the formula above with these values:

$$e^{i\varphi} = \cos \varphi + i \sin \varphi$$

$$e^{i\pi} = \cos \pi + i \sin \pi$$

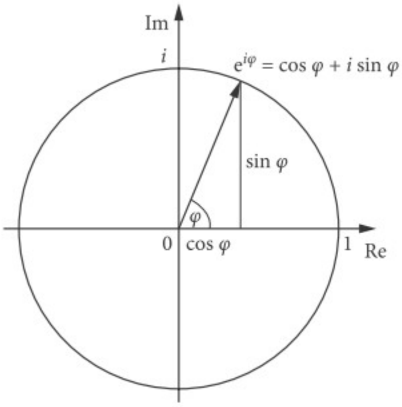

Figure 6.4 *Euler's identity* (from Wikimedia Commons)

$$e^{i\pi} = -1 + (i \times 0)$$

$$e^{i\pi} + 1 = 0$$

This is called Euler's identity. It connects some of the most significant numbers in the history of mathematics—e, i, π, 1, and 0. In 1990, the *Mathematical Intelligencer* periodical took a survey of its readers, conducted by mathematician David Wells, asking them to vote on the "most beautiful theorem in mathematics." Euler's identity came out as the winner by a large margin. It certainly is easy to see why mathematicians perceive it as having an inherent beauty, since it connects some of the most important numbers in mathematics.

There are many interesting and bizarre aspects of this equation, starting with the fact that it is really not an equation in the traditional sense, as Henshaw (2014, 58) aptly observes:

> Generally, the term "equation" implies something that it can be solved. For example, we can solve $x + 2 = 4$ and get the result $x = 2$. But Euler's identity isn't like that. There's nothing to solve for. It's just a statement of fact, like saying that $2 + 2 = 4$. And so, just as $2 + 2 = 4$ is a fact, $e^{i\pi} + 1 = 0$ is also a fact.

Epilogue

The fact that e is a descriptor of physical reality, especially in the domains of growth, decay, change, and probability once again raises the question left by Pythagoras' legacy: Why is mathematics such a powerful and unexpected language of truth? Also, as we have seen in this chapter: Why is change in natural phenomena logarithmic and described by exponential functions?

Even practical matters are governed by such mathematics. Consider the following:

> Bank A offers 100% interest on money. In one year, a sum of \$1000 becomes \$2000. Bank B pays 50% twice a year. In this case, \$1000 becomes \$1500 after six months and this becomes \$2250 after twelve months. This is clearly the better deal. Why is this so?

The formula for compound interest is the following one:

$(1 + r/n)^n$

r = annual interest rate (using 1 rather than 100%)

n = number of yearly periods

In the yearly model, $r = 1$ and $n = 1$:

$(1 + 1/1)^1 = 2.0$

In the half-yearly model, $r = 1$ and $n = 2$ (twice a year):

$(1 + 1/2)^2 = 2.25$

In the monthly model, $r = 1$, $n = 12$ (twelve times a year):

$\left(1 + 1/12\right)^{12} = 2.613$

The higher the value of n the greater the yield, and this shows why Bank B gave the better deal. Recall the formula for e at the start of this chapter:

$$\left(1 + 1/n\right)^n$$

As can be seen, the compound interest formula is based on this very formula. When the interest rate is n, the formulas become the same.

As Euler's identity seems to suggest, there may be some deeper meaning in *e*, as it relates to other numbers, of which we are not yet aware. All this suggests that mathematics is itself a conundrum. Its discoveries, such as the Pythagorean theorem, *e*, π, and so on, cannot be tied down to a specific meaning, even if they emerge in a context, including practical calculations such as those related to compound interest.

Explorations

1. A possible series for e

The number *e* has been determined by a number of series, as illustrated in this chapter. Determine if any of the following approaches *e* (2.71828); that is, if the sum of the terms might approach *e* as a limit.

$$(a)\ 1 + 1/2 + 1/3 + 1/4 + 1/5 + 1/6 + 1/7 + 1/8 \ldots$$

$$(b)\ 1 - 1/2 + 1/3 - 1/4 + 1/5 - 1/6 + 1/7 - 1/8 \ldots$$

$$(c)\ 1 \times 1/2 \times 1/3 \times 1/4 \times 1/5 \times 1/6 \times 1/7 \times 1/8 \ldots$$

2. Another series

Here is another series. It was introduced in this chapter:

$$1/0! + 1/1! + 1/2! + 1/3! + 1/4! + 1/5! + \dots$$

Note that in this case $1/0! = 1$. Does this series allow us to estimate the value of e to any degree of accuracy?

3. Plotting a function

Plot the function, $y = f(x) = x^2$ and then describe its general shape. In Table 6.2 below are some values:

Table 6.2 ***Values for f(x) = x²***

x	**0**	**1**	**−1**	**2**	**−2**	…
$f(x) = x^2$	$0^2 = 0$	$1^2 = 1$	$-1^2 = 1$	$2^2 = 4$	$-2^2 = 4$	…

4. Compound interest

The notions in this chapter are used in calculations of compound interest, as discovered by Jacob Bernoulli (above). Here is a problem that is similar to the one above.

> You are offered a new part-time job as a pizza delivery person, working only on weekends. Your boss gives you a choice in salary options as follows:
>
> *Option A*:
> $4,000 for your first year of work, and a raise of $800 for each year after the first.
>
> *Option B*:
> $2,000 for your first six months of work, and a raise of $200 every six months thereafter.
>
> Which offer is the better one?

5. The exponential function e^x

Below are a few values for e^x:

$$x = 1, e^x = 2.718$$
$$x = 2, e^x = 7.389$$
$$x = 3, e^x = 20.085$$
$$x = 4, e^x = 54.598$$

What power of x would produce the value of 1096.633? What does this imply generally about change (growth, decay, etc.) that is described by this function?

7

i

Imaginary numbers

Imaginary numbers are a fine and wonderful refuge of the divine spirit, almost an amphibian between being and non-being.

—Gottfried Leibniz (1831–1881)

Prologue

Consider the solution of the following quadratic equation, which is an equation with a variable, x in this case, which is squared:

$$x^2 + 1 = 0$$

$$x^2 = -1$$

$$x = \sqrt{-1}\left(\text{or, more precisely, } \pm\sqrt{-1}\right)$$

What kind of number is $\sqrt{-1}$? In a way that parallels the unexpected discovery of $\sqrt{2}$ by the Pythagoreans (chapter 1), when this number surfaced as a solution to a quadratic equation, mathematicians asked themselves what it could possibly mean. Not knowing what to call it, René Descartes named it, logically enough, an *imaginary number*. Euler introduced the letter "*i*" to stand for this number. Like the irrationals, the discovery of i led to new ideas and discoveries. One of these was *complex numbers*—numbers having the form

$(a + bi)$, where a and b are real numbers and i is $\sqrt{-1}$. Incredibly, complex numbers turn out to have many applications. They are used, for instance, to describe electric circuits and electromagnetic radiation and they are fundamental to quantum theory in physics. Returning to the equation above, $x^2 + 1 = 0$, if i is used like any other number, the equation has the following solution: $x = \pm i$.

In his treatise on equations, *Ars magna* (The Great Art), Italian mathematician Gerolamo Cardano (1501–1576) was among the first to come across the square root of negative numbers (Salem, Testard, and Salem 1992, 48–49). In one part, he was examining two numbers, x and y, which add up to 10 and, when multiplied, come to 40:

$$(1)\ x + y = 10$$

$$(2)\ xy = 40$$

The solution is shown below:

From (1): $y = 10 - x$
Substituting in (2): $x(10 - x) = 40$
Expanding and transposing: $-x^2 + 10x - 40 = 0$
Multiplying by -1: $x^2 - 10x + 40 = 0$
$x = 5 \pm \sqrt{-15}$

Cardano decided to abandon exploring the implications of this result calling the solution fictitious. On the other hand, Cardano's contemporary and compatriot Rafael Bombelli (1526–1572) saw such numbers as simply that—numbers that show up as the roots of some equations, formulating the first coherent theory of complex numbers in his *Algebra* of 1572. So, the values of Cardano's equation could be written as follows:

$$x = 5 \pm i\sqrt{15}$$

This chapter will look at the implications and significance of imaginary numbers, which constitute yet another of the great ideas of mathematics that have not only changed the course of mathematics but also of human history. Contemporary science and engineering would be much more limited without them.

Quadratic equations

Quadratic equations were known throughout the ancient world, from Babylon to China and India. The concept of "variable" and the use of symbols to represent numbers is traced to Greek mathematician Diophantus (c. 250 CE), who lived in Alexandria, Egypt, where he developed the techniques that later earned him the title "father of algebra." He also discussed methods of solving quadratic equations—similar methods were elaborated a few centuries later by the Hindu scholar Aryabhata (c. 476–550 CE).

It is useful here to rapidly review quadratic equations, which are defined as equations in which the power of one of the variables is 2. The general form of the quadratic equation is as follows:

$$ax^2 + bx + c = 0$$

The a and b are the numerical coefficients of x^2 and x respectively. The letter c represents the constant in the equation. Thus, for example, if $a = 3$, $b = 2$, and $c = 5$, then: $ax^2 + bx + c = 3x^2 + 2x + 5 = 0$. Consider the simple equation below, noting that in this case, $a = 1$, $b = 0$, and $c = -16$:

$$x^2 - 16 = 0$$

The solution is determined as follows:

$$x^2 - 16 = 0$$

$$x^2 = 16$$

$$x = +4 \text{ or } -4, \text{ shortened to } \pm 4$$

Now, consider the following equation:

$$x^2 + 8x + 15 = 0$$

We know that the expression $(x^2 + 8x + 16)$ represents a perfect square because it can be rewritten as $(x + 4)^2$. So, let's add 16 to both sides of the equation above, in order to produce a perfect square on the left side. This does not alter the values of the equation since the same number has been added to both sides. Note that 15 is moved to the right side, becoming -15:

$$x^2 + 8x + 16 = -15 + 16$$

$$x^2 + 8x + 16 = 1$$

$$(x + 4)^2 = 1$$

$$\surd(x+4)^2 = \surd 1$$

$$(x + 4) = 1$$

$$x = -3$$

Or:

$$-(x + 4) = 1$$

$$x = -5$$

Solution set:

$$x = \{-3, -5\}$$

The most common method for solving a quadratic equation is to use the following formula:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Without going into a step-by-step explanation, we plug these values in the equation above and we end up with the same solution: $x = \{-3, -5\}$.

The Babylonians were among the first to develop an algorithm for solving quadratic equations, as is evidenced by a Babylonian tablet preserved in the British Museum in which the following problem is posed (Solomon 2008, 11–12):

> The area of a square added to the side of a square comes to 0.75. What is the side of the square?

The solution unfolds as shown in Table 7.1, with corresponding contemporary algebraic notation added to show the reasoning involved.

The graphs of quadratic functions show parabolic curvature. Where the parabola touches or crosses the x-axis is where the two solutions lie. For example, the graph in Figure 7.1 shows the solution set of $x^2 - x - 2 = 0$, namely $x = \{-1, +2\}$, which are the points on the x-axis crossed by the curve:

Table 7.1 ***Babylonian Algorithm for Quadratic Equations***

Babylonian Tablet	*Contemporary Notation*
I have added the area and the side of my square: 0.75.	$x^2 + x = 0.75$
You write down 1, the coefficient.	Coefficient of x is 1
You break half of 1.	Half of 1 is 0.5
You multiply 0.5 and 0.5.	$(0.5)^2 = 0.25$
You add 0.25 and 0.75.	$0.25 + 0.75 = 1$
This is the square of 1.	$\sqrt{1} = 1$
Subtract 0.5, which you multiplied.	$1 - 0.5 = 0.5$
0.5 is the side of the square	$x = 0.5$

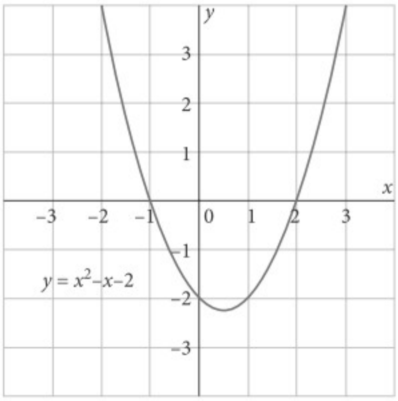

Figure 7.1 *Graph of* $y = x^2 - x - 2$ (from Wikimedia Commons)

Complex numbers

Squaring a negative number produces a positive number (chapter 3). So, what kind of number is $\sqrt{-1}$? All we can say is that it does not behave like any of the real numbers:

$$i = \sqrt{-1}$$
$$i^2 = (\sqrt{-1}) \times (\sqrt{-1}) = -1$$

The first mention of such numbers is traceable, actually, to the great twelfth-century Hindu mathematician Bhāskara (1114–1185). This was followed centuries later by Cardano and Bombelli, as mentioned above. It was Jean-Robert Argand (1768–1822) who showed how imaginary numbers and real numbers could be interconnected, followed by Carl Friedrich Gauss (1777–1855), who introduced the term, *complex number* in 1831. For example, every real number can be represented as a complex number, by simply letting the imaginary part be 0. So, for instance, 5 is:

$$a + bi = 5 + 0i = 5$$

In 1797 Gauss proved something revolutionary—namely, that complex numbers could solve any equation built from real numbers. This means that every equation has a full set of solutions among the complex numbers.

When imaginary numbers were first discovered, it was not clear how they fit into the number system or how they could be represented on the Cartesian plane. This conundrum led to the ingenious invention of a diagram by Argand that made it possible to show the relation of imaginary numbers to real ones. Figure 7.2 is the representation of $(a + bi)$ on the Argand plane (Im = imaginary number line; Re = real number line).

To add two complex numbers, say $(a + bi)$ and $(c + di)$, we add each part (real and imaginary) separately:

$$(a + bi) + (c + di) = (a + c) + (b + d)i$$

Below is an example:

$$(4 + 3i) + (5 + 6i) = (4 + 5) + (3 + 6)i = 9 + 9i$$

To multiply complex numbers we multiply them in the usual algebraic fashion:

$$(a + bi)(c + di) = ac + (ad)i + (bc)i + (bd)i^2$$

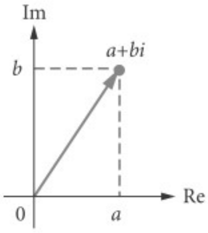

Figure 7.2 *Graph of a + bi* (from Wikimedia Commons)

Below is an example. Note that $i^2 = -1$, so $18i^2 = (18)(-1) = -18$:

$$(4+3i)(5+6i) = 20 + (24)i + (15)i + 18i^2 = 20 + (24)i + (15)i - 18 = (2+39i)$$

The other operations can also be specified in the same way, using the formulas $(a + bi)$ and $(c + di)$. The point is, as Bombelli and others suggested, that complex arithmetic is just like any type of arithmetic. The difference is, of course, the addition of i to it.

Fundamental Theorem of Algebra

The study of complex numbers led to an extraordinary discovery, known as the Fundamental Theorem of Algebra, which states that the field of complex numbers is algebraically closed—that is, there is no polynomial expression that does not have at least one complex root. It stipulates that an equation of degree n has n roots. So, for example a quadratic equation will have two roots, a cubic equation three, a quartic equation four, and so on. Let's look at a few for the sake of illustration:

Linear (one root):
$x - 5 = 0$
$x = \{5\}$

Quadratic (two roots):
$x^2 - 4 = 0$
$x = \{+2, -2\}$

Cubic (three roots):
$2x^3 + 3x^2 - 11x - 6 = 0$
$x = \{2, -1/2, -3)$

Quartic (four roots):
$3x^4 + 6x^3 - 123x^2 - 126x + 1080 = 0$
$x = \{3, 5, -4, -6\}$

This theorem was already implicit in the work of Descartes and Albert Girard (1595–1632). The first attempt to prove it is traced to French mathematician Jean d'Alembert (1717–1783) in 1746. But his proof contained some infelicities, which were corrected by Gauss, who presented his corrections in 1799. Gauss published a complete proof in 1816, building on ideas established by Euler.

The Fundamental Theorem is part of a broader theory of equations. It was Evariste Galois (1811–1832) who took a closer look at equations after the theorem was established, showing that solutions to equations could be carried out with permutations, leading to the emergence of group theory. Solutions to the quadratic, the cubic, and quartic have been found. However, no solution to the quintic (equation with a variable to the power of five) has ever emerged. Galois found that the groups of permutations associated with the quadratic, cubic, and quartic share a property that the quintic does not. This is the reason why no formula can be devised to solve the quintic.

Epilogue

The number *i* came from solving quadratic equations and had no meaning initially—much like √2 had no meaning for the Pythagoreans. After being labeled an imaginary number, it took on a mathematical life all its own, opening up a new way for expanding, classifying, and thus understanding numbers. Where were these numbers before they were discovered? As Ian Stewart (2013, 313) observes, the concept of *existence* in any treatment of mathematics raises several key questions, the most obvious one being the definition of *existence* itself:

> The deep question here is the meaning of "exist" in mathematics. In the real world, something exists if you can observe it, or, failing that, infer its necessary presence from things that can be observed. We know that gravity exists because we can observe its effects, even though no one can see gravity…However, the number two is not like that. It is not a thing, but a conceptual construct.

The irrational and the imaginary numbers did not "exist" until they cropped up in the use of the Pythagorean theorem and in solving quadratic equations respectively. Were they waiting to be discovered? This question is clearly at the core of the nature of mathematics. This "story" can be told over and over within the field—transfinite numbers, graph theory, and so on. These did not "exist" until they crystallized in the conduct of mathematics, through ingenious notational modifications, diagrammatic insights, various explorations with mathematical signs, and so on. Coming up with a theory of discovery in mathematics is an intractable task, since so many unknown factors are involved.

Consider the following question: What would happen if we add *j* and *k* to *i* in the Argand plane? The result is numbers with four parts. These were described by Irish mathematician William Rowan Hamilton (1806–1865) in

1843. He called them *quaternions.* They have led to a whole set of new rules of arithmetic shown in the multiplication table in Table 7.2.

Hamilton showed, in effect, that complex numbers were a subset of quaternions. Just as complex numbers can be shown as points on a plane, so quaternions can be considered to be points in a graph such as the one in Figure 7.3.

As it turns out, quaternions have uncovered the "existence" of many previously unknown number patterns and have had significant applications in physics. Quaternion theory has become especially significant for describing rotation in space. And it is now an essential tool in computer graphics, molecular modeling, and space flight.

Serendipity is likely to be a product of the imagination, as the late writer and semiotician Umberto Eco argued in his book *Serendipities: Language and Lunacy* (1998). The term was coined by English writer Horace Walpole who came across an ancient Persian tale, *The Three Princes of Serendip*, that suggested serendipity as a principle of life. The tale goes somewhat as follows. Three princes from Ceylon were journeying in a strange land when they came upon a man looking for his lost camel. The princes had never seen the animal, but asked the owner: Was it missing a tooth? Was it blind in one eye? Was it

Table 7.2 ***Multiplication of quaternions***

×	1	*i*	*j*	*k*
1	1	*i*	*j*	*k*
i	*i*	−1	*k*	−*j*
j	*j*	−*k*	−1	*i*
k	*k*	*j*	−*i*	−1

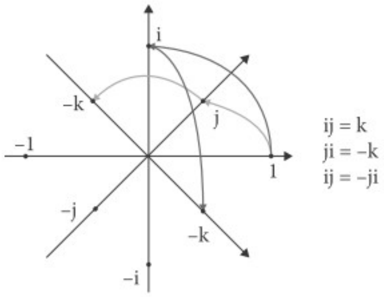

Figure 7.3 *Quaternion graph* (from Wikimedia Commons)

lame? Was it laden with butter on one side and honey on the other? Was a pregnant woman riding it? Incredibly, the answer to all their questions was "yes." The owner accused the princes of having stolen the animal since, clearly, they could not have had such precise knowledge. But the princes merely pointed out that they had observed the road, noticing several patterns in it: for example, the grass on either side was uneven, suggesting a lame gait; there were places where chewed food seemed to emerge from a gap in the animal's mouth; there were uneven patterns of footprints, the signs of awkward mounting and dismounting typical of someone who was pregnant; and there were differing accumulations of ants and flies, which congregate around butter and honey. Their questions were really prompted by inferences based on these observations. Knowing the world in which they lived, they were able to make concrete connections between the observations and what happened. Since Serendip was Ceylon's ancient name, Walpole coined the word *serendipity*, to designate how we come about our discoveries, such as mathematical and scientific ones, in a similar way.

Explorations

1. Imaginary numbers

To get a sense of what i means, figure out the following:

$$(1)\ i^3$$

$$(2)\ i^6$$

$$(3)\ i^0$$

2. A square root

Express $\sqrt{-9}$ as an imaginary number.

3. Conjugates

A conjugate number is a complex number with the opposite middle sign. So, the conjugate of the number $(a + bi)$ is $(a - bi)$, and vice versa. What happens when you multiply conjugates?

$$(a+bi)(a-bi)=?$$

4. Conjugate arithmetic

Now, apply what you discovered in exploration 3 to multiply the following:

$$(\text{a})\ (3+2i)\times(3-2i)=?$$
$$(\text{b})\ (5+3i)\times(5-3i)=?$$

5. Powers of complex numbers

Can you figure out the following?

$$(\text{a})\ (4-5i)^2$$
$$(\text{b})\ (3-3i)^3$$

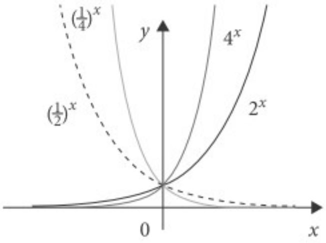

Figure 5.1 *Examples of exponential functions (from Wikimedia Commons)*

1. Equations where the variable or variables are to the power of 1 are called *first degree*: $3x + 1$; $5x + 2y$; etc.
2. Expressions where the variable or variables are to the power of 2 are named accordingly the *second degree*: $3x^2 + 1$; $5x^2 + 2y$; etc.
3. Expressions where the variable or variables are to the power of 3 are called expressions to the *third degree*: $3x^3 + 1$; $5x^3 + 2y$; etc.
4. In general, expressions where the variable is to the power of n are called expressions to the n^{th} degree.

These were known in antiquity, but the lack of an appropriate notation somewhat impeded the development of advanced algebra. The new notation provided mathematicians with a powerful descriptive tool that has had implications throughout mathematics. One of the more fascinating discoveries associated with exponents is the so-called *exponential function*. This is a function of the form $y = a^x$ (for example, $y = 2^x$). All such functions, when plotted onto the Cartesian plane have a parabolic shape as in Figure 5.1.

These functions show up in many places, from the calculation of compound interest to the measurement of anything that grows or decays, such as bacteria. They are used in radiocarbon dating, which has been of great importance to archeology.

Pascal's Triangle

After the spread of exponential arithmetic, mathematicians started looking more closely at certain types of expressions. French mathematician Blaise Pascal (1623–1662) was working with the expression $(a + b)^n$, known as a binomial (consisting of two terms), when he came across a hidden pattern in its expansion, $n = \{0, 1, 2, 3, 4, 5, \ldots\}$. Here are the first eight expansions:

$$\begin{aligned}
(a+b)^0 &= 1 \\
(a+b)^1 &= a+b \\
(a+b)^2 &= a^2+2ab+b^2 \\
(a+b)^3 &= a^3+3a^2b+3ab^2+b^3 \\
(a+b)^4 &= a^4+4a^3b+6a^2b^2+4ab^3+b^4 \\
(a+b)^5 &= a^5+5a^4b+10a^3b^2+10a^2b^3+5ab^4+b^5 \\
(a+b)^6 &= a^6+6a^5b+15a^4b^2+20a^3b^3+15a^2b^4+6ab^5+b^6 \\
(a+b)^7 &= a^7+7a^6b+21a^5b^2+35a^4b^3+35a^3b^4+21a^2b^5+7ab^6+b^7
\end{aligned}$$

In the top row, the numerical coefficient in $(a + b)^0$ is 1; in the next row down the numerical coefficients of $(a + b)^1 = a + b = 1a + 1b$ are both 1; in the next one down the numerical coefficients of $(a + b)^2 = a^2 + 2ab + b^2 = 1a^2 + 2ab + 1b^2$ are 1, 2, and 1; and so on. Extracting these coefficients from the other terms in the expressions, and leaving them in their respective rows, Pascal noted that they could be arranged into the shape of a triangle (Figure 5.2).

Now, surprisingly, the triangle itself harbors patterns that can be broken down reductively as follows:

1. Leaving aside the vertex 1 of the triangle, each row begins and ends in 1.
2. The other numbers in a row are the sum of the two numbers right above it—for example, each of the 10s in the sixth row down equals the sum of 4 and 6 above them.
3. In fact, each number in the triangle is the sum of the two directly above it.
4. It is this structure that makes it an infinite triangle, since we could continue to generate numbers within it with this simple rule ad infinitum.
5. As Sir Isaac Newton noted, the exponent in each expression in the binomial distribution indicates the row in Pascal's triangle where the numerical coefficients occur: for example, in the expansion of $(a + b)^4$,

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1
1 7 21 35 35 21 7 1

Figure 5.2 *Pascal's Triangle*

which is $a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$, the numerical coefficients (1, 4, 6, 4, 1) coincide with the fourth row of numbers in Pascal's triangle (if we do not include the top 1). Similarly, the numerical coefficients in the expansion of $(a + b)^5$, $a^5 + 5a^4b + 10a^3b^2 + 10a^2b^3 + 5ab^4 + b^5$ coincide with the fifth row of numbers in the triangle (1, 5, 10, 10, 5, 1); those in the expansion of $(a + b)^6$ with the sixth row; and so on.

6. The sum of the digits in any row, n, of the triangle is a power of 2, that is, 2^n. For instance, the sum of the digits in the fifth row is: $1 + 5 + 10 + 10 + 5 + 1 = 32$, which is 2^5 (Chamberlain 2015, 164–165).

There are many other hidden patterns in this remarkable triangle. For the sake of historical accuracy, it should be mentioned that the triangle appears in a Sanskrit text named the *Chandas Shastra,* written between 500 and 200 BCE, and in a 1303 Chinese work *Si Yuan Yu Jian* written by Zhu Shijie (1206–1320). There is no evidence to suggest that Pascal was aware of any of these texts—a fact that raises a truly deep enigma with regard to discovery. Why is it that people in different parts of the world and at different times are able to come up with similar ideas? No real answer to this question can be given—it can only be observed. To contextualize Pascal's amazing discovery, Solomon (2008, 80) makes the following relevant remarks:

> To be fair to Pascal, he did not claim to have invented the triangle. He certainly did not name it after himself. He actually referred to it as the "arithmetic triangle" and used it to calculate probabilities.

Logarithms

Exponential notation was the spark for logarithms, which have been used in many areas of mathematics, science, and statistics, leading to many discoveries. A logarithm is the actual power or exponent to which a base, usually 10, must be raised to produce a given number:

> If $n^x = a$, the logarithm of a, with n as the base, is x; symbolically, $\log_n a = x$. For example, $10^3 = 1{,}000$; therefore, $\log_{10} 1{,}000 = 3$.

To get a sense of the usefulness of logarithms, suppose we wanted to calculate the number of ancestors in any previous generation. We have 2 parents, so there are 2 ancestors in the first generation. This can be expressed as $2^1 = 2$. Each of our parents has 2 parents, and so there are $2 \times 2 = 2^2 = 4$ ancestors in the previous generation. Each of the four grandparents has 2 parents, and so there are $4 \times 2 = 2 \times 2 \times 2 = 2^3 = 8$ ancestors in the generation before. The calculation continues according to this pattern. In which generation do we have 1,024 ancestors? This can be rephrased as follows: For which value of n is $2^n = 1{,}024$? We can find the answer by multiplying 2 a number of times until we reach 1,024. But if we know that $\log_2 1{,}024 = 10$, because $2^{10} = 1{,}024$, we can get the answer much more quickly and efficiently. This tells us that in the tenth previous generation we have 1,024 ancestors. Note that in this case the base is not 10, but 2.

The invention of logarithms is traced to Scottish mathematician John Napier (1550–1617), who devised the first ones for carrying out tedious complex arithmetical operations in his 1614 work, *Mirifici logarithmorum canonis descriptio* ("Description of the Marvelous Canons of Logarithms"), coining the term *logarithm* as a combination of Greek *logos* (reason) and *arithmos* (number). In it, Napier showed how simple it is to carry out arithmetical operations with logarithms, since each number can be represented with a logarithmic value and all such values can be included in a table. For example, to multiply any two numbers:

1. Find their logarithms in the appropriate tables.
2. Add the two logarithms.
3. Look the result up in the *antilogarithm* tables (an antilogarithm is the number itself).

Napier's logarithms were based on powers of 0.999999; their translation into tables with 10 as the base was due to English mathematician Henry Briggs (1561–1630) in 1624. Briggs' improvements and refinements of Napier's techniques helped the notion of logarithms gain broad acceptance (Solomon 2008, 67).

The number of discoveries, implications, and ideas that logarithms have spawned, within and outside of mathematics, is mindboggling. One of the most widespread uses of logarithms is in so-called *log scales*. Perhaps the best-known contemporary example is the Richter scale, which is used to measure earthquake sizes, developed by Charles Richter in 1935 (see Figure 5.3). The scale shows how the magnitude of an earthquake is measured in logarithmic terms.

Other log scale phenomena include the decibel scale of sound and the pH scale of acidity. The remarkable thing to consider is that all such discoveries and uses of logarithms came about after the invention of a simple notation—exponents.

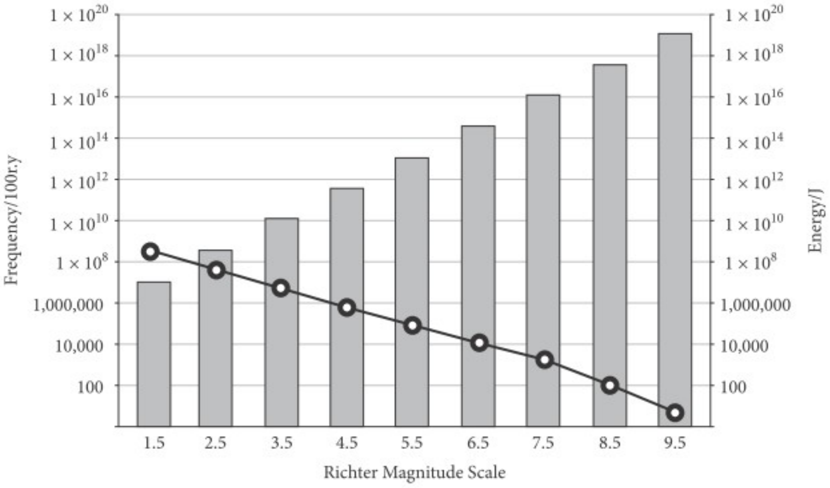

Figure 5.3 *A Richter Scale* (from Wikimedia Commons)

Notations and re-arrangements (such as the layout of Pascal's triangle) contain more hidden information about real phenomena than is initially apparent. It is upon reflecting on the hidden information that we come to extract more meaning from it than we would otherwise be likely to be able to do.

Epilogue

All of mathematics is encoded in symbolic notation. The Pythagorean equation $c^2 = a^2 + b^2$, for example, is essentially a convenient notation of saying the same thing as the sentence "the square on the hypotenuse is equal to the sum of the squares on the other two sides." In so doing, it takes the semantics in the linguistic sentence out, leaving only the structural (symbolic) outline of the information. It is this feature that makes it cognitively powerful, since we can now find many more meanings and applications for it, in addition to the original geometrical one. For example, one can now ask what integers fit the equation and, further, if there are other exponents for which the equation holds in general, $c^n = a^n + b^n$. From this deliberation on the notation itself, detached from its original geometrical meaning, has come much subsequent mathematical contemplation leading to such intriguing ideas as Fermat's Last Theorem, as we saw. In other words, notation suggests meanings that could not otherwise be contemplated.

The ancients did not have exponential notation at their disposal, so they used other devices to accomplish many of the same things as the new notation allowed so efficiently. In his book, *The Sand Reckoner*, Archimedes discussed the notion of powers in a roundabout way. However, it did not allow him to develop such techniques as exponential laws and logarithms.

Einstein's equation $E = mc^2$ has imprinted in it a lot of information about physical reality that could not be expressed in any other way. It says, in a nutshell, that the speed of light is constant and thus that it constrains physical reality. What happens if there is a universe where this formula does not hold? It would be unimaginable, even though it is possible. Above all else, we would need a notation to devise formulas for such a universe. As the philosopher Ludwig Wittgenstein (1922) put it, "Whereof one cannot speak, thereof one must be silent." As the language of Nature, mathematics breaks the silence periodically. $E = mc^2$ speaks volumes, to belabor the point somewhat.

Explorations

1. Square numbers

Here is a tricky problem, involving the exponent 2.

Find three numbers, less than or equal to 10, which when raised to the power of 2 and then added together equal 150.

2. Exponential arithmetic

Here is an age problem that requires exponential arithmetic.

> Alexander loves his grandmother. Here's how to figure out her age. It is between 50 and 100, divisible by 8. Each of the grandmother's sons has as many sons (her grandsons) as brothers. The combined number of her sons and grandsons equals her age. How old is she?

3. Generational logarithm

Above, the generation to which 1,024 ancestors belonged was computed using logarithmic method. Using the same method, to which generation do 256 ancestors belong?

4. A tough nut

This is a tough nut that became a meme on the Internet not too long ago: Can you figure out the value of x in $\sqrt{}(x + 15) + \sqrt{}x = 15$?

5. Laws of exponents

Here is a straightforward exploration what will allow you to review some of the laws of exponents discussed in this chapter. Simply provide the missing exponents in each of the following, as indicated by n in each case.

(a) $(2^n)(2^{12}) = 32{,}768$

(b) $(3^7)(2^n) = 279{,}936$

(c) $(5^n) \div (5^5) = 25$

(d) $(9^2)^n = 81$

data 06.05.2020
D S T [✓ Q] Q S S

Pág. 129 (138) Chapter 6
Some other conjectures
6.1. Integral points on elliptic curves

B) Books_79_2020/
Phase Transitions in Materials/
Phase_Transitions_in_Materials.pdf
Pág. 133 (154) Chapter 6 Energy
6.1 Atomic Schrödinger Equations and Formalism
Box 6.1 Dirac Notation

C) Books_79_2020/
Pythagoras' Legacy Mathematics in Ten Great Ideas/
Pythagoras_Legacy.pdf
Obs: Muito bom!
Pág. 82 (93) e: A very special number
Pág. 93 (104) i: Imaginary numbers
Pág. 75 (86) Pascal's Triangle
Obs: Pág. 82(93) -> Gauss

06 05 2020

You have no idea how much poetry there is in the calculation of a table of logarithms

Carl Friedrich Gauss (1777-1855)

Pag. 93(104)-

Imaginary numbers are a fine and wonderful refuge of the divine spirit, almost an amphibian between being and non-being.

Gottfried Leibniz (1831-1881)
?

D) Books_79_2020/
Superstrings and other Things.../

Superstrings_and_other_things.pdf

Pag. 419(440) Appendix C

Nobel Prize Winners in Physics

E) Books_79_2020/
MAA/

E1) Thinking Geometrically - A Survey of Geometries.pdf

obs: Livro todo!

Pag. 317(342) Chapter 7 Projective Geometry

!)*+,)))-) ' 6 > !>- >/ 6!
(% $ %% *66 ! C 3 7 6 L

? G >1$> # **#

)*+,)))-) ' / G . A> * 8 A
. 6!
-7 5 8 ! 999
(" 1 "% 6 1 H 8 . A

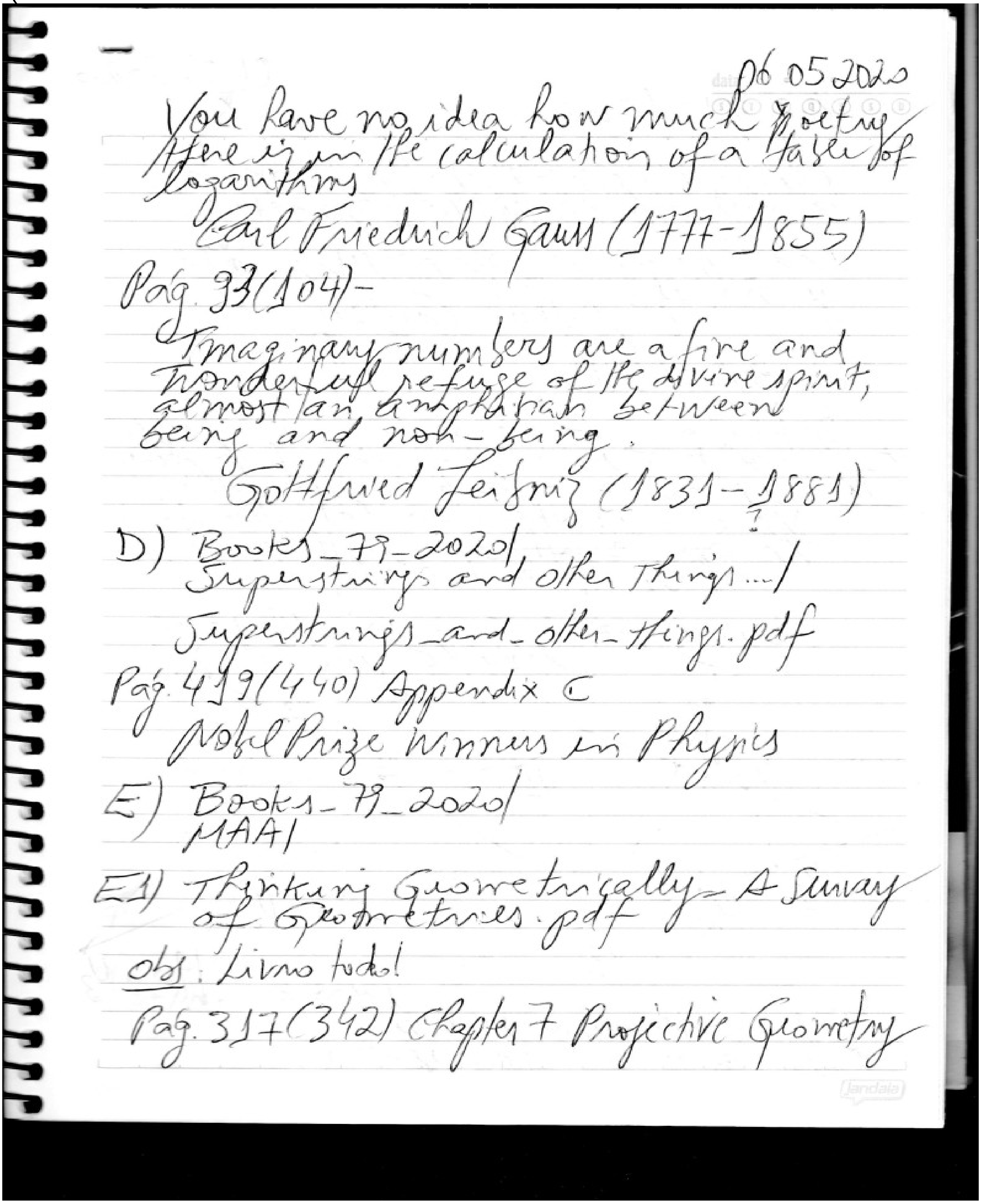

06 05 2020

You have no idea how much poetry there is in the calculation of a table of logarithms

Carl Friedrich Gauss (1777-1855)

Pag. 93(104)-

Imaginary numbers are a fine and wonderful refuge of the divine spirit, almost an amphibian between being and non-being.

Gottfried Leibniz (1831-1881) ?

D) Books_79_2020/
Superstrings and other Things.../

Superstrings_and_other_things.pdf

Pag. 419(440) Appendix C

Nobel Prize Winners in Physics

E) Books_79_2020/
MAA/

E1) Thinking Geometrically - A Survey of Geometries.pdf

obs: Livro todo!

Pag. 317(342) Chapter 7 Projective Geometry

data 06.05.2020

Pág. 318 (342)

Metrical geometry is thus (portanto) a part of [projective] geometry, and [projective] geometry is all geometry. Arthur Cayley.

Pág. 320 (345)

Definition. A perspective with respect to a point T is a mapping of the points (~~U_1~~) U_i on one line to the points V_i on another so that (de modo a) V_i ~~on another a~~ is the image of U_i if and only if

U_i, V_i, and T are on the same line (Figure 7.3)

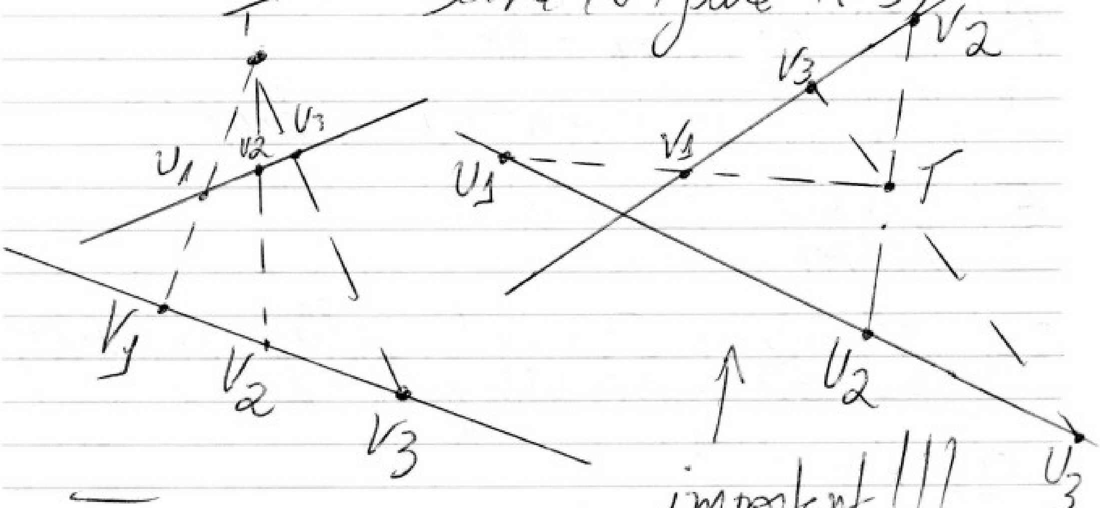

important!!!

$\overline{TU_1}$ → tem um sentido
TU_3 → tem outro sentido.

06.05.2020

Esta perpectiva é muito importante por que podemos descobrir a projeção de um ponto em outro plano quando há uma intersecção de duas retas.

Iremos extrapolar para planos e entre plano e uma superfície curva.

Important:

Pág. 321 (346) Figure 7.4
A Complete quadrangle.

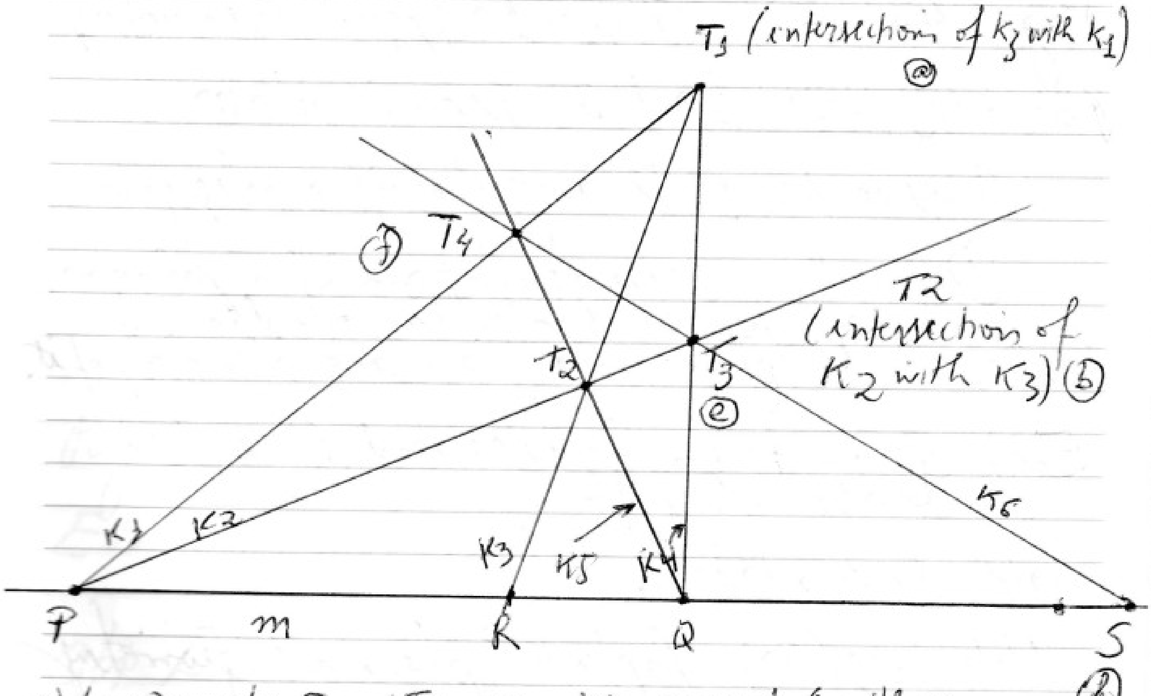
T1 (intersection of K3 with K1) ⓐ
T4 ⓕ
T2 (intersection of K2 with K3) ⓑ
T2
T3 ⓔ
K1
K2
K3
K5
K4
K6
P
m
R
Q
S ⓗ

g) Line → connecting T3 and T4
h) S → line intersections line m

ⓒ K4 → connect Q with T_1
ⓓ K5 → connect Q with T_2
ⓔ K4 → intersects K2, say at T_3
ⓕ K5 → intersects K1, say at T_4.

data 06.05.2020

Pág. 337(362) 7.3 Analytic Projective Geometry

Pág. 338(363) Example 2

We can relate each point of the projective plane to two antipodal points on a sphere.

Fig 7.21

Pág. 361(386) 7.6 Projective Space

Fig 7.32 A two-point perspective view of a cube

Pág. 447(472) Chapter 10 Discrete Geometry

Pág. 449(474) 10.1.2 Triangulations

Figure 10.1 Triangulating a polygon.

Depicts a triangulation of a polygon, a division of it into triangles all whose (de quem) vertices are vertices of it.

E2) Geometry - The Line and the Circle (2018...)...pdf

obs: Livro todo!!!

Pág. 59(80) Chapter 4 Spherical Geometry

Pág. 189(210) 10 Book III: Circles

Pág. 221(242) 11 Book IV: Circles & Polygons

)*+,)))-) ' . A> / 5 !
:2 ** 6!
($: (:$ (%

")*+,)))-) ' ! !> !
A $2 * * 6!
(" % 0

> *And what are these fluxions? The velocities of evanescent increments? And what are these same evanescent increments? They are neither finite quantities, nor quantities infinitely small, nor yet nothing. May we not call them the ghosts of departed quantities?*
>
> George Berkeley, Bishop of Cloyne, *The Analyst* (1734) [52, p. 338]

> *...these increments must be conceived to become continuously smaller, and in this way, their ratio is represented as continuously approaching a certain limit, which is finally attained when the increment becomes absolutely nothing. This limit, which is, as it were, the final ratio of those increments, is the true object of differential calculus.*
>
> Leonhard Euler *Institutiones Calculi Differentialis* (1755) [24]

4

Differentiation

There are two classes of geometric problems treated by calculus: the determination of *tangents* is associated with derivatives, and the determination of *areas* with integrals. The standard practice in calculus texts (which we follow) is to look at derivatives first, but historically area problems were studied far more extensively and far earlier.

In Euclid and Archimedes, the notion of a line tangent to a curve was limited to circles and other *convex* curves; a line was tangent to such a curve if it touched it at exactly one point. (It is a tribute to the care that Archimedes took that he explicitly defined what it means for a curve to be convex, and set forth which properties he assumed followed from this definition [32, pp. 2–4].) A more general idea of tangent lines was given in Fermat's method for finding maxima and minima (Exercise 15 in § 4.7), which was primarily an algebraic algorithm, but supported at least in part by a concept similar to the modern one via secant lines. Descartes had a method for finding the tangent line by first finding the *normal* or perpendicular line (the *circle method* [21, pp. 125–7]) in connection with finding double roots of equations. Isaac Barrow (1630–1677), Newton's teacher and predecessor in the Lucasian Chair at Cambridge, found tangent lines via a method similar to Fermat's, using the "characteristic triangle" which we discuss below.

The approaches of Leibniz and Newton to tangents were quite different. Newton was coming at curves from a kinematic point of view; using the synthetic geometric language of Euclid, many of his tangent/derivative constructions consisted of taking a point moving toward another along a curve, and considering the chord between the two; he would then consider the positions of these chords and conclude that the "ultimate ratio" between these chords and the tangent line at the latter point was "that of equality". In general, his curve was *parametrized*: that is, it was generated by a combination of motions for two quantities, which we would label x (the abscissa) and y (the ordinate); he would consider their "fluxions" $\dot{x}$ and $\dot{y}$ (in our terms, the rates of change of the two quantities) and conclude that the tangent line has slope given by $\frac{\dot{y}}{\dot{x}}$.

Leibniz, on the other hand, thought in terms of a “characteristic triangle” whose sides were “infinitesimal” (*i.e.*, of vanishing length): this was a right triangle whose horizontal leg was the “differential” dx, the vertical leg was dy, and the hypotenuse was ds. These were to represent the infinitesimal increments of the abscissa, ordinate, and of the position along the curve. He would quite happily use the similarity between this triangle and various more conventional ones to deduce results about the tangent and related items. Our modern view combines some of Newton’s kinematic vision with Leibniz’s differential notation, rejecting the idea of actually taking a ratio of the form $\frac{dy}{dx}$ as Leibniz might do, but considering the limiting value of the ratio. However, Leibniz’s notation has survived in all aspects of calculus.

The first quotation at the head of this chapter is a famous passage from a pamphlet published in 1734 by George Berkeley (1685–1753), the Bishop of Cloyne. Berkeley attacked the reasoning behind the calculus of Newton, Leibniz and their colleagues to counter attacks on religious thinking as inexact; his point was that (even though the results *work*) the logical basis for the calculus was no more rigorous (no, less so) than the theological thinking of the time. His criticisms found their target; Newton among others tried to firm up the foundations of their thinking; the quotation from Euler is an indication of the ways this was attempted. But rigorous formulations of the basic ideas of limits, continuity and derivatives really only came in the nineteenth century.

The first careful formulations of the idea of derivative were given by Bernhard Bolzano (1781–1848) in 1817 and Augustin-Louis Cauchy (1789–1857) in 1821; we have already mentioned several times that Bolzano’s work, while somewhat earlier, was not generally known. Cauchy’s definition of the derivative is much the same as ours.

There was a presumption in the seventeenth through mid-nineteenth centuries that all continuous functions can be differentiated, with perhaps a few exceptional points (such as interface points for functions defined in pieces). In fact, Joseph Louis Lagrange (1736–1813) built a whole theory of functions around this assumption (using power series—see Chapter 6) [38]. At the end of this chapter, we will construct an example that shows this assumption can fail spectacularly. The first such example was probably given by Bolzano in 1834, but went unnoticed; however, when Karl Weierstrass (1815–1897) constructed such an example in the 1860s, it had an effect analogous to the discovery of the irrationality of $\sqrt{2}$ in Hellenic times, and was a major push toward the development of a far more careful formulation of the ideas of continuity, limits and derivatives.

4.1 Slope, Speed and Tangents

The **graph** (or **locus**) of an equation in x and y is the collection of points in the plane whose coordinates (x, y) satisfy the equation. The simplest equations involve sums of terms which are constant multiples of x or y, or just constants

$$Ax + By = C; \tag{4.1}$$

these are called **linear equations**, because their graphs are straight lines. When A and B are both nonzero, then the line crosses each coordinate axis at a unique point: it crosses the x-axis (which is the set of points where $y = 0$) at the point $(\frac{C}{A}, 0)$ and the y-axis (where $x = 0$) at $(0, \frac{C}{B})$; $\frac{C}{A}$ (*resp.*, $\frac{C}{B}$) is called the $\boldsymbol{x}$**-intercept** (*resp.*, $\boldsymbol{y}$**-intercept**) of the line.

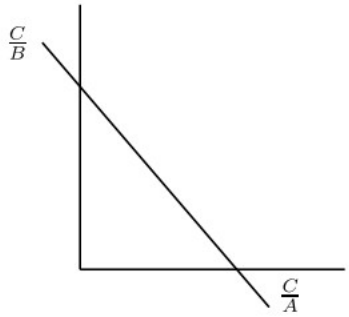

Figure 4.1. $Ax + By = C$

Of course, an equation can be manipulated algebraically, either by multiplying both sides by the same nonzero number or adding the same term to both sides, without changing the graph: in this way (provided $B \neq 0$) we can rewrite equation (4.1) as the definition of a polynomial function of degree 1

$$y = f(x) := mx + b \tag{4.2}$$

where $m = -\frac{A}{B}$ and $b = \frac{C}{B}$. Technically, such a function is not called "linear" (even though its graph is a straight line) unless $b = 0$; in general it is called an **affine function**. The coefficients of this polynomial have geometric interpretations. You can see immediately that b is the y-intercept of the graph. The various lines going through the point $(0, b)$ (or any other given point, for that matter) are distinguished by their "tilt", which is determined by the first-order coefficient m. Geometrically, if we compare the point $(0, b)$ (corresponding to $f(x)$ at $x = 0$) and the point corresponding to another, higher value of x, we see that the value $f(a) = ma + b$ at $x = a$ is precisely ma units higher than that at $x = 0$ (assuming $m > 0$; otherwise we adopt the convention that going "up" (*resp.*, "right") a *negative* number of units is the same as going "down" (*resp.*, "left") by its absolute value). Thus, m represents the ratio between the change, or **increment** $\triangle y$ in the output $y = f(x)$ and the increment $\triangle x$ in the input x as we go from $x = 0$ to any other value of x. In fact, if we start at *any* input value $x = x_1$ (with corresponding output value $y_1 = f(x_1)$) and go to any *other* input $x = x_2$ (with $y = y_2 = f(x_2)$), the increment in x is

$$\triangle x := x_2 - x_1$$

and the corresponding increment in $y = f(x)$ is

$$\triangle y := y_2 - y_1 = f(x_2) - f(x_1)$$

(sometimes also denoted "$\triangle f$") so (when f is affine, given by equation (4.2))

$$\triangle y = (mx_1 + b) - (mx_2 + b) = m(x_2 - x_1) = m\triangle x$$

and thus we have

$$m = \frac{\triangle y}{\triangle x}$$

for *any* pair of *distinct* inputs $x_1 \neq x_2$. The fraction on the right is pronounced "*delta y over delta x*", but another colloquial expression for it is "*rise over run*", reflecting the fact that $\triangle y$ corresponds to vertical displacement and $\triangle x$ to horizontal displacement along the line.

The value $m = \frac{\triangle y}{\triangle x}$ is called the **slope** of the line $y = mx + b$; geometrically,

$$m = \frac{\triangle y}{\triangle x} = \tan\theta \tag{4.3}$$

where θ is the angle between the line and any horizontal line (Figure 4.2): *positive* slope means the line tilts *up* to the right while *negative* slope means it tilts *down* to the right; a horizontal line is the graph of a constant function, with slope zero (Figure 4.3). It also follows that two lines are parallel precisely if they have the same slope (Exercise 7).

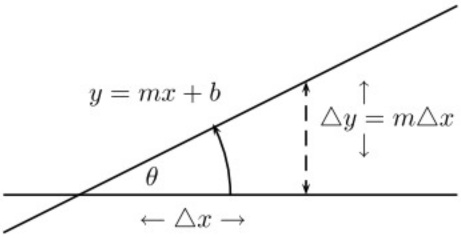
$y = mx + b$
$\triangle y = m\triangle x$
θ
$\leftarrow \triangle x \rightarrow$

Figure 4.2. $m = \tan\theta$

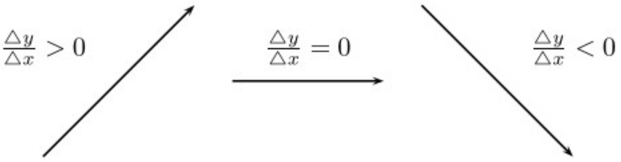
$\frac{\triangle y}{\triangle x} > 0$
$\frac{\triangle y}{\triangle x} = 0$
$\frac{\triangle y}{\triangle x} < 0$

Figure 4.3. Sign of slope

The fact that $\frac{\triangle y}{\triangle x}$ is the same for any pair of points when y is an affine function of x can be used to write down the equation of the line through any pair of distinct points. For example, for $(x_1, y_1) = (-1, 3)$ and $(x_2, y_2) = (2, -3)$, the increments are

$$\begin{aligned} \triangle y &= y_2 - y_1 = (-3) - (3) = -6, \\ \triangle x &= x_2 - x_1 = 2 - (-1) \quad = 3, \end{aligned}$$

so the affine function $f(x) := mx + b$ with $f(-1) = 3$ and $f(2) = -3$ has

$$m = \frac{\triangle y}{\triangle x} = \frac{-6}{3} = -2;$$

then the fact that $f(-1) = 3$ means

$$-2(-1) + b = 3$$

or

$$b = 1$$

so the equation of the given line is

$$y = -2x + 1.$$

(You should check that $(x_2, y_2) = (2, -3)$ also lies on this line.)

The slope $\frac{\triangle y}{\triangle x}$ can also be interpreted as the **rate of change** of $y = f(x)$ between $x = x_1$ and $x = x_2$: in our example, a (horizontal) change in x of 3 units leads to a (vertical) change in $y = -2x + 1$ of -6 units, so y changes -2 units for every unit of change in x.

A little thought, fortified by some calculation (Exercise 8), shows that affine functions (including constants) are the *only* functions with domain $(-\infty, \infty)$ for which $\frac{\triangle y}{\triangle x}$ is the same for any pair of points. For example if we carried out the corresponding calculations for

$$f(x) = 3x^2 - 2$$

with $x_1 = -1$ and $x_2 = 2$, we would get $y_1 = 3(-1)^2 - 2 = 1$, $y_2 = 3(2)^2 - 2 = 10$, so

$$\frac{\triangle y}{\triangle x} = \frac{9}{3} = 3$$

while if we changed x_2 to $x_2 = 1$, we would have $y_2 = 3(1)^2 - 2 = 1$, so

$$\begin{aligned} \triangle y &= y_2 - y_1 = 1 - 1 &&= 0, \\ \triangle x &= x_2 - x_1 = 1 - (-1) &&= 2, \end{aligned}$$

and

$$\frac{\triangle y}{\triangle x} = 0;$$

similarly, if $x_1 = 1$ and $x_2 = 2$, then

$$\begin{aligned} \triangle y &= 10 - 1 = 9, \\ \triangle x &= 2 - 1 \ \ = 1, \end{aligned}$$

so

$$\frac{\triangle y}{\triangle x} = 9.$$

How, then, do we formulate the rate of change of a non-affine function? First, we have to be more specific: clearly $y = 3x^2 - 2$ is *increasing* when x is close to $x = 2$ but *decreasing* when x is near $x = -1$, so we need to specify *where* we want to find the rate of change. Second, the quantity $\frac{\triangle y}{\triangle x}$ represents a kind of *average* rate of change: we can only say that, for example, between $x = -1$ and $x = 2$ the function $y = 3x^2 - 2$ has a net increase of 9 units for a 3-unit increase in x, giving an average rate of change of 3 units increment in y per unit increment in x, while between $x = -1$ and $x = 1$ the function has no *net* change (of course, it does not stay constant in between—it just happens to return to its initial value), so the *average* rate of change between $x = -1$ and $x = 1$ is zero.

If we try to formulate "how fast" a function is changing *at a particular point*, we need to resort to limits. The idea is that if, say, we want to know how fast $y = 3x^2 - 2$ is changing *at* $x = 1$, we work from the assumption that the *average* rate of change over a *small* interval containing $x = 1$ should approximate the "true" or **instantaneous rate of change** at $x = 1$, and the smaller the interval, the better the approximation. This leads to the notion of a derivative.

Definition 4.1.1. *Suppose the function f is defined for all input values near $x = a$. The* ***derivative*** *of $y = f(x)$ at $x = a$ is the limit defined as follows:*

- *for each input x near $x = a$, set*

$$\begin{aligned}\triangle x &= x - a,\\ \triangle y &= f(x) - f(a)\,,\end{aligned}$$

- *the derivative is then the limit*

$$\lim_{x\to a}\frac{\triangle y}{\triangle x} = \lim_{x\to a}\frac{f(x)-f(a)}{x-a}.$$

While this idea was there in Newton's work, in the form of "ultimate ratios", it was not stated so clearly, particularly not in terms of functions (this notion, in fact, came later). The first statement of the derivative of a *function* in the terms above appears to have been by Jean le Rond D'Alembert (1717–1783) in an article in an encyclopedia (1754). Edwards [21, p. 295] has a translated quotation of d'Alembert's statement. Augustin-Louis Cauchy (1789–1857) gave a *rigorous* definition of derivative, more or less as above, in 1823 and used it along with his rigorous formulations of the notions of limit and integral to develop a careful theory of basic calculus.

Note that in this process, $\frac{\triangle y}{\triangle x}$ *does not make sense* when $x = a$. Also, notice that we could equally well have formulated this in terms of $\triangle x$ instead of x, using $x = a + \triangle x$:

$$\frac{\triangle y}{\triangle x} = \frac{f(a+\triangle x) - f(a)}{\triangle x}$$

and taken the limit as $\triangle x \to 0$.[1] Of course, as $\triangle x \to 0$, so[2] does $\triangle y$. A standard notational convention in calculus, which we shall see again in Chapter 5, says that when a quantity is defined in terms of increments of certain variables (like $\frac{\triangle y}{\triangle x}$), then the limiting value as all of these go to zero is denoted by replacing the Greek letter Δ (denoting a concrete small increment, calculated from two distinct but close points) with the corresponding English[3] letter d. Thus, a standard notation for the derivative is

$$\frac{dy}{dx} := \lim_{\triangle x\to 0}\frac{\triangle y}{\triangle x}.$$

Note that the expression $\frac{dy}{dx}$ is *not* the ratio between two quantities called "dx" and "dy"; rather we look at the *whole expression* as the limit of the ratio $\frac{\triangle y}{\triangle x}$. If we want to explicitly indicate *where* the derivative is being taken, we write

$$\left.\frac{dy}{dx}\right|_{x=a} = \lim_{x\to a}\frac{\triangle y}{\triangle x};$$

---

[1] It is common to replace the slightly awkward symbol $\triangle x$ with h in the definition of derivative, so one frequently sees the definition of the derivative written in the (clearly equivalent) form $f'(a) = \lim_{h\to 0}\frac{f(a+h)-f(a)}{h}$.

[2] at least if f is continuous

[3] or French, in the case of Cauchy

also, when $y = f(x)$, we sometimes use "f" in place of "y"

$$\left.\frac{dy}{dx}\right|_{x=a} = \left.\frac{df}{dx}\right|_{x=a} = \frac{df}{dx}.$$

An entirely different notation (due to Lagrange), which indicates in a more graceful way the function being differentiated and the place where the differentiation is being performed, is

$$\left.\frac{df}{dx}\right|_{x=a} = f'(a)\,,$$

and finally, we sometimes write y' in place of $f'(a)$ when $y = f(x)$.

Let us examine this definition in case $f(x) = 3x^2 - 2$ and $a = 1$. We will think in terms of $\triangle x$. We have

$$\begin{aligned}\triangle y = \triangle f &= f(1+\triangle x) - f(1) = [3(1+\triangle x)^2 - 2] - [1] \\ &= 6\triangle x + 3\triangle x^2\end{aligned}$$

so

$$\frac{\triangle y}{\triangle x} = \frac{\triangle f}{\triangle x} = \frac{6\triangle x + 3\triangle x^2}{\triangle x} = \frac{(6+\triangle x)\triangle x}{\triangle x} = 6 + \triangle x$$

and clearly

$$y' = f'(1) = \frac{df}{dx} = \frac{dy}{dx} = \lim_{\triangle x \to 0} \frac{\triangle y}{\triangle x} = \lim_{\triangle x \to 0}(6 + \triangle x) = 6.$$

There is a geometric way of thinking about the process of differentiation. When we have two distinct points (x_1, y_1) and (x_2, y_2) on the graph of $y = f(x)$, the line joining them is sometimes called a **secant line**; the slope of this line is the ratio $\frac{\triangle y}{\triangle x}$ of increments calculated using this pair of points (Figure 4.4). To calculate the derivative $f'(a)$, we nail down the point (x_1, y_1) at $x_1 = a$ (and $y_1 = f(a)$) and consider the secant line joining this to the variable point $x_2 = a + \triangle x$ (and $y_2 = f(a + \triangle x) = y_1 + \triangle y$); now consider the limit as $\triangle x \to 0$ (*i.e.*, $x_2 \to x_1$). All the secant lines we consider in this process go through the point $(x_1, y_1) = (a, f(a))$, but as x_2 changes they rotate around that point: it is geometrically clear that as $\triangle x \to 0$ this rotation brings them ever closer to a limiting position, with slope $\frac{dy}{dx}$: this defines the "tangent line" (to the graph) at the point (Figure 4.5).

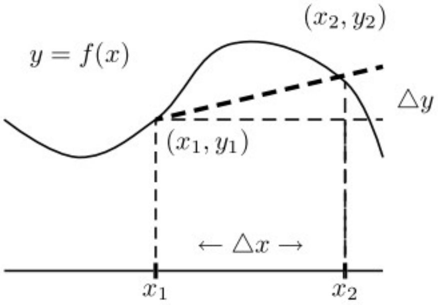

Figure 4.4. Secant Line

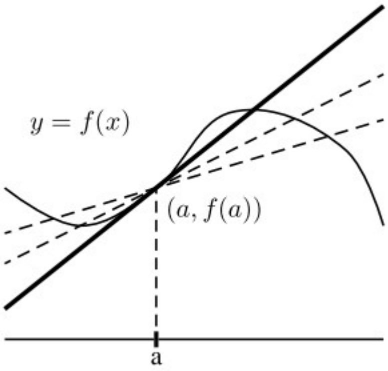

Figure 4.5. Tangent Line

Definition 4.1.2. *Suppose the function f is defined for all x near $x = a$. The line* ***tangent*** *to the graph $y = f(x)$ at $x = a$ is the limit of the secant lines through $(x_1, y_1) = (a, f(a))$ and nearby points (x_2, y_2) on the graph as $\triangle x = x_2 - x_1 \to 0$; the tangent line is the line through $(a, f(a))$ with slope $m = f'(a)$.*

From this description, we can deduce that the equation of the tangent line at $(a, f(a))$ is

$$y - f(a) = f'(a)\,(x - a)$$

or

$$y = f(a) + f'(a)\,(x - a). \tag{4.4}$$

The function defined by equation (4.4) (whose graph is the tangent line) plays a very important role in the application of derivatives to the analysis of f. We will clarify later (Lemma 6.1.1 and Proposition 6.1.2) that this is the affine function which best approximates the behavior of f near $x = a$. We will call this function by several names: it is the **tangent map** or **affine approximation** to $f(x)$ at $x = a$; we denote it by

$$T_a f(x) := f(a) + f'(a)\,(x - a).$$

In all of this discussion, we have ignored a crucial point: does every function *have* a derivative, or equivalently, does every curve have a tangent line? After all, these are defined as limits, and we have already seen that whenever we work with limits we should first make sure they exist. The following shows that not every function can be differentiated.

Remark 4.1.3. *In order for (the limit defining) the derivative $f'(a)$ to exist, the function f must be continuous at $x = a$.*

This is simply a consequence of the definition of the derivative: we are interested in the limit of the fraction $\frac{\triangle y}{\triangle x}$ as $\triangle x \to 0$, and if $\triangle y$ does *not* go to zero, this *must* diverge (why?). But saying $\triangle y \to 0$ is the same as saying $f(a + \triangle x) - f(a) \to 0$, and this is precisely the same as

$$f(a) = \lim_{x \to a} f(x)$$

(Exercise 9).

We say f is **differentiable** at $x = a$ if the limit defining $f'(a)$ exists; the remark tells us that *every differentiable function is continuous*. Thus for example the function

$$f(x) := \begin{cases} \frac{|x|}{x} & \text{for } x \neq 0, \\ 0 & \text{for } x = 0 \end{cases}$$

is *not* differentiable at $x = 0$. We have seen that $f(x) = 3x^2 - 2$ *is* differentiable at $x = 1$, and in the next section we shall establish that most of our "everyday" functions are differentiable wherever they are defined. But there is one glaring exception: the absolute value

$$f(x) = |x|.$$

First, look at the graph of the absolute value (Figure 4.6): for $x \geq 0$, we have $f(x) = x$, a line through the origin with slope $+1$, but for $x \leq 0$, $f(x) = -x$ is a line of slope -1. Certainly near any point $x = a$ *different from* $x = a$, the function looks just like the function $y = x$ (*resp.*, $y = -x$) if $a > 0$ (*resp.*, $a < 0$), so it is differentiable with derivative $y' = \pm 1$. But what is the tangent line at the origin? *All* secants through $(x_1, y_1) = (0, 0)$ and (x_2, y_2) with $x_2 > 0$ (and hence $y_2 = x_2$) have slope $+1$ while *all* secants with $x_2 < 0$ (so $y_2 = -x_2$) have slope -1. Thus, in the definition of $\frac{dy}{dx}$ as the limit of $\frac{\triangle y}{\triangle x}$, we have distinct one-sided limits

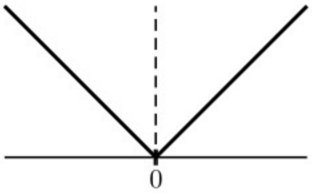

Figure 4.6. $y = |x|$

$$\lim_{\triangle x \to 0^+} \frac{\triangle y}{\triangle x} = \lim_{\triangle x \to 0^+} +1 = +1,$$
$$\lim_{\triangle x \to 0^-} \frac{\triangle y}{\triangle x} = \lim_{\triangle x \to 0^-} -1 = -1.$$

Thus, the (two-sided) *limit* does *not* exist: *the absolute value function $y = |x|$ is continuous, but NOT differentiable at $x = 0$.*

We shall deal with other examples of functions defined in pieces in the next section.

Exercises for § 4.1

Answers to Exercises 1–2ace, 3a, 4aceg, 5acf, 6acd, 10ade, 11c, 12a are given in Appendix B.

Practice problems:

1. Find the slope of each line below:

(a) $y = 3x + 5$
(b) $x = 3y + 5$
(c) $x + y = 0$
(d) $3x + 2y = 5$
(e) $3x = 2y + 5$
(f) $2y = 3x + 5$

> *. . . these were the days of Cavalieri and Roberval, when an area was looked upon as the same thing as an infinite number of line segments, a very helpful if dangerous definition.*
>
> J. L. Coolidge [17]

5

Integration

We have spent a good deal of time and energy studying differentiation; we now turn to the complementary notion of integration. Integration arises in connection with the calculation of areas, volumes, and length of curves, as well as averaging processes and related problems, particularly in probability and statistics.

The study of areas and volumes goes back to the Greeks; in the exercises for the first section of this chapter we will explore the results given by Euclid and Archimedes concerning the area of a circle. Both sources did much more, using versions of the method of exhaustion. For example, Archimedes found the area under a parabola, and volumes of cones, spheres, cylinders, and even paraboloids and hyperboloids. In the eleventh century the Islamic scholar Alhazen (*ca.*965–1039) extended some of these results, for example finding the volume obtained when a parabola is rotated about its base (rather than its axis). In the early seventeenth century Johannes Kepler (1571–1630) and Bonaventura Cavalieri (1598–1647) found formulas for other areas and volumes by heuristic methods. More general methods, closer to ones we now recognize as rigorous, were developed by Pierre de Fermat (1601–1665), Evangelista Torricelli (1608–1647), Gilles Roberval (1602–1675) and Blaise Pascal (1623–1662) in Europe, and John Wallis (1616–1703) and Isaac Barrow (1630–1677) in England.

Then, in the 1660s, the stunning result, known as the *Fundamental Theorem of Calculus*, was discovered: the process of finding areas (integration) is the inverse of the operation of finding tangents (differentiation). This was observed in a special case by Torricelli [52, p. 253] and in more general terms (but in a geometric language that to a modern reader is rather obscure) by James Gregory (1638–1675)—a brilliant Scots mathematician whose early death diminished his historical influence—and Isaac Barrow—Newton's teacher and predecessor in the Lucasian Chair of Mathematics at Cambridge. The full power of this observation was harnessed by Newton and Leibniz (independently) and is at the heart of much of the power of the calculus.

After this breakthrough, calculus developed at a breathtaking pace; by 1700, many of the results and techniques of what we call first-year calculus (and much more) had already been developed by the English and Continental schools. The eighteenth century saw a further blossoming of the field, particularly through power series methods in the hands of the masters Leonard Euler (1707–1783) and Joseph Louis Lagrange (1736–1813). During this period, integration was seen (as it is by many beginning calculus students now) primarily as *antidifferentiation*, and its foundation in the theory of area was somewhat eclipsed.

Then in the early nineteenth century, in connection with problems concerning the wave equation and heat equation, the foundations of integration were re-examined. Augustin-Louis Cauchy (1789–1857) gave a careful formulation of the definite integral (making more rigorous Leibniz's view of a definite integral as a sum of infinitesimals) and used his formulation to discuss integrals in the complex domain, as well as convergence of series. Euler had done some investigating of approximation of integrals by finite sums, but for him this was a property of antiderivatives; for Cauchy it was the definition: in fact, the Fundamental Theorem of Calculus was obtained by him as a consequence of his definition. Cauchy dealt primarily with integrals of continuous functions, but the emerging study of Fourier series required going beyond the assumption of continuity. Cauchy's definition was further refined by Bernhard Riemann (1826–1866) in 1854 to a form close to what we use today, although the recognition that some of his results required the notion of uniform continuity really came later, from Karl Weierstrass (1815–1897). Riemann also gave necessary and sufficient conditions for his integral to exist for a given function; we explore this in the last section of this chapter.

The notion of the integral continued to be extended: **??**Stieltjes18561894 introduced the idea of integrating with respect to a mass distribution in 1894, and a revolutionary new theory of integration was produced by Henri Lebesgue (1875–1941) in his doctoral thesis (1902); the description of the Lebesgue integral is beyond the scope of this book, but in its ability to handle really wild functions and its introduction of new ideas it has had a profound effect on the development of analysis in the last century; even more abstract and general formulations of the integral proliferated throughout that century.

5.1 Area and the Definition of the Integral

The idea of area, like that of velocity, begins with common-sense intuition but ends up requiring sophisticated limit processes to work in any but the simplest situations.

Our basic premise is that area is *additive*: if a region S consists of two pieces S_1 and S_2 which can share boundary points, but not interior points (*i.e.*, , they don't *overlap*) then the area $A(S)$ of the whole region should be the sum of the areas of the pieces:

$$A(S) = A(S_1) + A(S_2)\,. \tag{5.1}$$

We use this in two ways: *subdivision* and *tiling*. Starting with a standard square (sides of length 1) as our unit of area ($A = 1$), we can subdivide it into n^2 subsquares with sides of length $1/n$, so each of these subsquares must have area $1/n^2$. Now, a rectangle whose sides have rational length,[1] say width $w = p/n$ and length $\ell = q/n$, can be tiled by q rows consisting of p squares each with side $1/n$, so we obtain the area formula

---

[1] with common denominator n

$$A = \ell w$$

for (rational) rectangles; it is natural to extrapolate this formula to rectangles with rational or irrational sides.

It is also possible to use these kinds of subdivision and tiling arguments to deduce the area formula for triangles and other polygons. However, this approach collapses when we attempt to calculate the area of regions with curved sides.

To understand this more general situation, we investigate the area of the plane region (see Figure 5.1) S bounded below by the x-axis, on the left and right by two vertical lines ($x = a$ and $x = b$) and above by the graph of a function $y = f(x)$, which we assume is nonnegative and continuous on the interval $x \in [a, b]$:

$$S = \{(x, y) \,|\, a \le x \le b,\ 0 \le y \le f(x)\}\,.$$

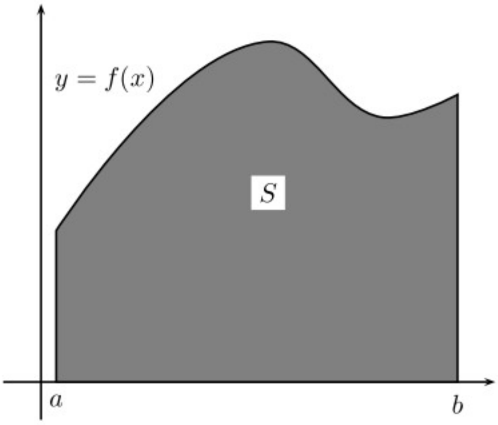

Figure 5.1. Area under a graph

We will illustrate our general discussion using the example

$$f(x) = x^3,\ a = 1,\ b = 2.$$

The idea is to locate the number $A(S)$ (giving the area of S) by a process of approximation analogous to our approximation of $\sqrt{2}$ in § 2.2. Our estimates use the *monotonicity* of area: the area of any subregion is less than the area of the whole:

$$S_1 \subset S_2 \Rightarrow A(S_1) \le A(S_2)\,. \tag{5.2}$$

(This actually follows from additivity of area: see Exercise 3.)

To obtain *lower* bounds for $A(S)$, we fit nonoverlapping rectangles *inside* S (Figure 5.3). Start by subdividing the interval $[a, b]$ into adjacent subintervals: this is specified by a finite list of division points, together with the two endpoints, numbered left-to-right and starting with index 0:

$$a = p_0 < p_1 < \cdots < p_n = b.$$

We refer to the collection

$$\mathcal{P} = \{p_0, p_1, \dots, p_n\}$$

as a **partition** of $[a, b]$. The **component intervals** of $\mathcal{P}$

$$I_j = [p_{j-1}, p_j] \quad j = 1, \dots, n$$

are also numbered left to right; the length of the j^{th} component interval is

$$\triangle x_j = \|I_j\| = p_j - p_{j-1} \quad j = 1, \dots, n.$$

As an example, we consider, for $n = 2, 3, \dots$, the partition $\mathcal{P}_n$ of the interval $[1, 2]$ into n intervals, *all of the same length*—that is, the points p_i, $i = 0, \dots, n$ are equally spaced in $[1, 2]$, so

$$\triangle x_j = \frac{2-1}{n} = \frac{1}{n}$$

for $j = 1, \dots, n$, and we easily see that

$$p_j = 1 + \frac{j}{n}, \quad i = 0, \dots, n$$

or

$$\mathcal{P}_n = \left\{1, 1 + \frac{1}{n}, 1 + \frac{2}{n}, \dots, 1 + \frac{n-1}{n}, 2\right\}.$$

Now, we use each component interval I_j (sitting on the x-axis) as the base of a rectangle of height h_j (and width $w_j = \triangle x_j$). This rectangle is contained inside S precisely if h_j is a lower bound for $f(x)$ on I_j, so the *tallest* rectangle with base I_j that fits inside S (Figure 5.2) has height

$$h_j = h_j^- := \inf_{x \in I_j} f(x)$$

and hence area

$$\triangle A_j^- = h_j^- w_j = \left[\inf_{x \in I_j} f(x)\right] \triangle x_j.$$

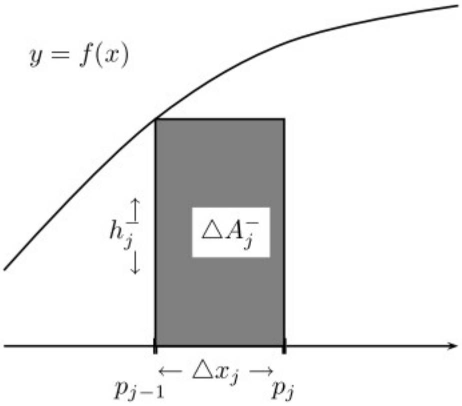

Figure 5.2. Inside Rectangle over $I_j = [p_{j-1}, p_j]$

Thus, starting from the component intervals of the partition $\mathcal{P}$, the largest union of rectangles we can fit inside S (Figure 5.3) has total area

$$\sum_{j=1}^{n} \triangle A_j^- = \sum_{j=1}^{n} \left[\inf_{x\in I_j} f(x)\right] \triangle x_j.$$

The expression on the right is called the **lower sum** for f associated to the partition $\mathcal{P}$ of $[a,b]$; we write

$$\mathcal{L}(\mathcal{P},f) = \sum_{j=1}^{n} \inf_{I_j} f \triangle x_j.$$

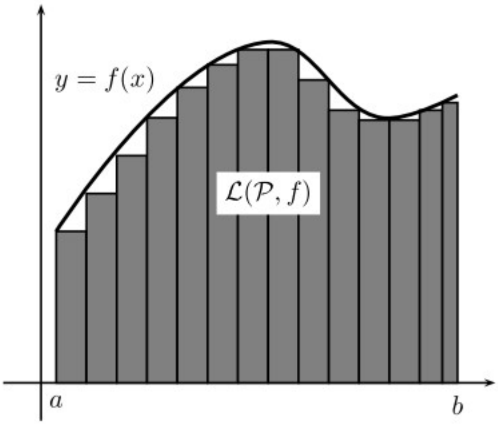

Figure 5.3. $\mathcal{L}(\mathcal{P},f)$

Monotonicity of area then says that, for every partition $\mathcal{P}$,

$$\mathcal{L}(\mathcal{P},f) \leq A(S).$$

In our example, $f(x)=x^3$ is strictly increasing over $[1,2]$, so for each j,

$$\begin{aligned} h_j^- &= \inf_{x\in I_j} x^3 = p_{j-1}^3 \\ &= \left(1+\frac{j-1}{n}\right)^3 \end{aligned}$$

and the lower sum for x^3 associated to the partition $\mathcal{P}_n$ of $[1,2]$ into n equal parts is

$$\begin{aligned} \mathcal{L}(\mathcal{P}_n, x^3) &= \sum_{j=1}^{n} \left[1+\frac{(j-1)}{n}\right]^3 \cdot \frac{1}{n} \\ &= \frac{1}{n} \sum_{i=j-1=0}^{n-1} \left[1+\frac{i}{n}\right]^3 \\ &= \frac{1}{n} \sum_{i=0}^{n-1} \left[1+3\frac{i}{n}+3\frac{i^2}{n^2}+\frac{i^3}{n^3}\right]. \end{aligned}$$

This formidable looking summation can be evaluated using the following summation formulas (Exercise 5):

$$\sum_{k=1}^{n} 1 = n, \tag{5.3}$$

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}, \tag{5.4}$$

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}, \tag{5.5}$$

$$\sum_{k=1}^{n} k^3 = \frac{n^2(n+1)^2}{4} = \left(\sum_{k=1}^{n} k\right)^2. \tag{5.6}$$

Using these (and noting that we are applying them with n replaced by $n-1$ in all but the first sum, since the terms for $i=0$ drop out), we can write

$$\begin{aligned}\mathcal{L}(\mathcal{P}_n, x^3) &= \frac{1}{n}\left[n + \frac{3}{n}\frac{(n-1)(n)}{2} + \frac{3}{n^2}\frac{(n-1)(n)(2n-1)}{6} + \frac{1}{n^3}\frac{(n-1)^2(n)^2}{4}\right] \\ &= 1 + \frac{3}{2}\left(\frac{n-1}{n}\right) + \frac{3}{6}\left(\frac{(n-1)(2n-1)}{n^2}\right) + \frac{1}{4}\left(\frac{(n-1)^2}{n^2}\right). \end{aligned} \tag{5.7}$$

To obtain an *upper* bound for $A(S)$, we use a similar construction to enclose S inside a union of rectangles built on the same bases. Given $I_j = [p_{j-1}, p_j]$ a component of the partition $\mathcal{P} = \{p_0, \ldots, p_n\}$ of $[a, b]$, the rectangle with base I_j and height h_j contains all the points $(x, y) \in S$ with $x \in I_j$ precisely if h_j is an upper bound for $f(x)$ on I_j; the *shortest* such rectangle has height[2]

$$h_j = h_j^+ :== \sup_{I_j} f.$$

This rectangle has area

$$\triangle A_j^+ = \sup_{I_j} f \triangle x_j.$$

The total area of these rectangles (Figure 5.4) gives an upper bound for $A(S)$,

$$A(S) \leq \sum_{j=1}^{n} \triangle A_j^+ = \sum_{j=1}^{n} \sup_{I_j} f \triangle x_j.$$

The sum on the right is called the **upper sum** for f associated to the partition $\mathcal{P}$, and denoted

$$\mathcal{U}(\mathcal{P}, f) = \sum_{j=1}^{n} \sup_{I_j} f \triangle x_j.$$

---

[2] The supremum here exists since f is continuous on I_j, hence bounded by the Extreme Value Theorem (Theorem 3.3.4).

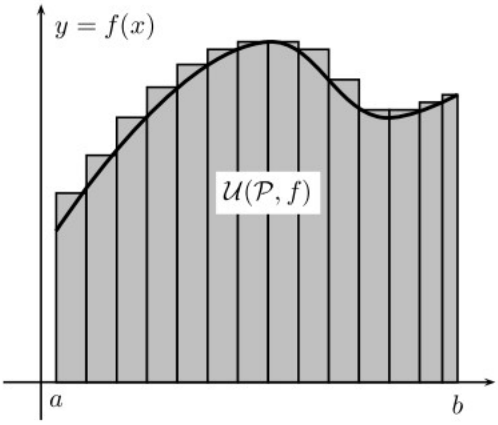

Figure 5.4. $\mathcal{U}(\mathcal{P}, f)$

In our example again, since $f(x) = x^3$ is strictly increasing on $[1, 2]$, we have

$$h_j^+ = \sup_{x \in I_j} x^3 = p_j^3$$

so that the upper sum is

$$\begin{aligned}\mathcal{U}(\mathcal{P}_n, x^3) &= \frac{1}{n} \cdot \sum_{j=1}^{n} \left[1 + 3\frac{j}{n} + 3\frac{j^2}{n^2} + \frac{j^3}{n^3}\right] \\ &= 1 + \frac{3}{2}\left(\frac{n+1}{n}\right) + \frac{3}{6}\left(\frac{(n+1)(2n+1)}{n^2}\right) + \frac{1}{4}\left(\frac{(n+1)^2}{n^2}\right). \quad (5.8)\end{aligned}$$

We see, then, that the area $A(S)$ of the plane region

$$S = \{(x, y) \,|\, a \leq x \leq b,\ 0 \leq y \leq f(x)\}$$

is bounded by the lower and upper sums associated to any partition $\mathcal{P}$ of $[a, b]$:

$$\sum_{j=1}^{n} \inf_{I_j} f \triangle x_j = \mathcal{L}(\mathcal{P}, f) \leq A(S) \leq \mathcal{U}(\mathcal{P}, f) = \sum_{j=1}^{n} \sup_{I_j} f \triangle x_j.$$

We note in particular that the "trivial partition" $\mathcal{P}_0 = \{a, b\}$ has one component interval ($I_1 = [a, b]$) with $\triangle x_1 = b - a$, leading to the estimate

$$\inf_{a \leq x \leq b} f(x)\,(b-a) = \mathcal{L}(\mathcal{P}_0, f) \leq A(S) \leq \mathcal{U}(\mathcal{P}_0, f) = \sup_{a \leq x \leq b} f(x)\,(b-a).$$

Any other partition $\mathcal{P} = \{a = p_0 < p_1 < \cdots < p_n = b\}$ can be obtained from $\mathcal{P}_0$ by adding the new division points $p_1, \ldots, p_{n-1}$ to $\mathcal{P}_0$, or equivalently by subdividing the component $[a, b]$ of $\mathcal{P}_0$ into several adjacent intervals, to form the component intervals of $\mathcal{P}$. For each component interval I_j of $\mathcal{P}$, we clearly have

$$\inf_{I_j} f \geq \inf_{[a,b]} f \text{ and } \sup_{I_j} f \leq \sup_{[a,b]} f$$

and, since

$$\sum_{j=1}^{n} \triangle x_j = b - a,$$

we can conclude that

$$\begin{aligned}
\mathcal{L}(\mathcal{P}, f) &= \sum_{j=1}^{n} \inf_{I_j} f \triangle x_j \\
&\geq \sum_{j=1}^{n} \inf_{[a,b]} f \triangle x_j = \inf_{I_j} f \sum_{j=1}^{n} \triangle x_j = (\inf_{I_j} f)(b-a) \\
&= \mathcal{L}(\mathcal{P}_0, f)
\end{aligned}$$

and, similarly,

$$\mathcal{U}(\mathcal{P}, f) = \sum_{j=1}^{n} \sup_{I_j} f \triangle x_j \leq (\sup_{[a,b]} f)(b-a) = \mathcal{U}(\mathcal{P}_0, f).$$

An analogous argument (Exercise 6) shows that adding new division points to *any* partition $\mathcal{P}$ leads to a new partition $\mathcal{P}'$ whose lower and upper sums are bounded by those of $\mathcal{P}$. We call $\mathcal{P}'$ a **refinement** of $\mathcal{P}$ if, as sets of points,

$$\mathcal{P} \subset \mathcal{P}';$$

that is, every partition point of $\mathcal{P}$ is also a partition point of $\mathcal{P}'$ (or, in different terms, if every component interval of $\mathcal{P}'$ is entirely contained in a single component interval of $\mathcal{P}$). We have

Remark 5.1.1. *If $\mathcal{P}'$ is a refinement of $\mathcal{P}$, then*

$$\mathcal{L}(\mathcal{P}, f) \leq \mathcal{L}(\mathcal{P}', f)$$

and

$$\mathcal{U}(\mathcal{P}, f) \geq \mathcal{U}(\mathcal{P}', f).$$

So far, we have been proceeding as if the area $A(S)$ were a well-defined number satisfying the estimates

$$\mathcal{L}(\mathcal{P}, f) \leq A(S) \leq \mathcal{U}(\mathcal{P}, f) \tag{5.9}$$

for every partition $\mathcal{P}$ of $[a, b]$. But all we *really* know are the lower and upper sums $\mathcal{L}(\mathcal{P}, f)$ and $\mathcal{U}(\mathcal{P}, f)$; we would like to *define* the area $A(S)$ as the number satisfying all the inequalities above. For such a definition to make sense, we need two conditions. First, there must be *at least one* number satisfying all of these inequalities. This boils down to the requirement that *every* lower sum is less than (or equal to) *every* upper sum, even for different partitions. When this holds, we are able to conclude that

$$\sup_{\mathcal{P}} \mathcal{L}(\mathcal{P}, f) \leq \inf_{\mathcal{P}} \mathcal{U}(\mathcal{P}, f),$$

(where the supremum and infimum are taken over all possible finite partitions $\mathcal{P}$ of $[a, b]$) and any number between these two automatically satisfies all the inequalities. But then, for these inequalities to actually *define* a number $A(S)$, we need to insure that *only one* number works—that is, we really need the *equality*

$$\sup_{\mathcal{P}} \mathcal{L}(\mathcal{P}, f) = \inf_{\mathcal{P}} \mathcal{U}(\mathcal{P}, f).$$

The first condition can be established using Remark 5.1.1. Given two partitions $\mathcal{P}_1$ and $\mathcal{P}_2$, we can form their **mutual refinement** $\mathcal{P}$ by taking all division points from either partition; that is, as sets of points,

$$\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2.$$

Now, $\mathcal{P}_i \subset \mathcal{P}$ ($\mathcal{P}$ is a refinement of $\mathcal{P}_i$) for $i = 1, 2$, and hence

$$\mathcal{L}(\mathcal{P}_i, f) \leq \mathcal{L}(\mathcal{P}, f) \leq \mathcal{U}(\mathcal{P}, f) \leq \mathcal{U}(\mathcal{P}_i, f), \quad i = 1, 2. \tag{5.10}$$

In particular, for arbitrary partitions $\mathcal{P}_1$ and $\mathcal{P}_2$,

$$\mathcal{L}(\mathcal{P}_1, f) \leq \mathcal{U}(\mathcal{P}_2, f).$$

But then the upper sum of *any* partition is an upper bound for *all* lower sums, and hence for their least upper bound

$$\sup_{\mathcal{P}} \mathcal{L}(\mathcal{P}, f) \leq \mathcal{U}(\mathcal{P}_2, f),$$

and the *left* side is a lower bound for *all* upper sums, hence for their greatest lower bound:

$$\sup_{\mathcal{P}} \mathcal{L}(\mathcal{P}, f) \leq \inf_{\mathcal{P}} \mathcal{U}(\mathcal{P}, f).$$

We state this formally, for future use:

Remark 5.1.2. *For any two partitions $\mathcal{P}_1$ and $\mathcal{P}_2$,*

$$\mathcal{L}(\mathcal{P}_1, f) \leq \mathcal{U}(\mathcal{P}_2, f)$$

and hence

$$\sup_{\mathcal{P}} \mathcal{L}(\mathcal{P}, f) \leq \inf_{\mathcal{P}} \mathcal{U}(\mathcal{P}, f).$$

So there is always *some* number bounded below by all lower sums and above by all upper sums. Its uniqueness takes further work. Before proceeding, we will enlarge our context. Note that in the construction of lower and upper sums, as well as in the proofs of the various inequalities so far, we have used the assumption that f is continuous *only* to establish that f is bounded on $[a, b]$ (so that the numbers involved in defining $\mathcal{L}(\mathcal{P}, f)$ and $\mathcal{U}(\mathcal{P}, f)$ are all well-defined). Go back over these arguments to convince yourself that everything so far works if we just assume that f is *bounded* (but not necessarily continuous) on $[a, b]$. That is, if there exists $M \in \mathbb{R}$ such that

$$0 \leq f(x) \leq M$$

for all $x \in [a,b]$, then for every partition $\mathcal{P}$ of $[a,b]$ we can define the upper and lower sums

$$\mathcal{L}(\mathcal{P},f) = \sum_{j=1}^{n} \inf_{I_j} f \triangle x_j,$$
$$\mathcal{U}(\mathcal{P},f) = \sum_{j=1}^{n} \sup_{I_j} f \triangle x_j,$$

and these satisfy the inequality

$$\mathcal{L}(\mathcal{P}_1,f) \leq \mathcal{U}(\mathcal{P}_2,f)$$

for any pair of partitions $\mathcal{P}_1$ and $\mathcal{P}_2$. It follows that

$$\sup_{\mathcal{P}} \mathcal{L}(\mathcal{P},f) \leq \inf_{\mathcal{P}} \mathcal{U}(\mathcal{P},f).$$

So for any bounded nonnegative function f, we will say that the set of points in the plane

$$S = \{(x,y) \mid a \leq x \leq b,\ 0 \leq y \leq f(x)\}$$

has a well-defined area if

$$\sup_{\mathcal{P}} \mathcal{L}(\mathcal{P},f) = \inf_{\mathcal{P}} \mathcal{U}(\mathcal{P},f)$$

—that is, only one number can satisfy equation (5.9)—and in that case the **area** $A(S)$ is precisely this (well-defined) number.

In general, it is a daunting task to try to maximize (*resp*., minimize) the lower (*resp*., upper) sums over *all* partitions. In our example, however, we can get at this indirectly. You can verify directly from equation (5.7) and equation (5.8) that

$$\lim_{n\to\infty} \mathcal{L}(\mathcal{P}_n,x^3) = \lim_{n\to\infty} \mathcal{U}(\mathcal{P}_n,x^3) = 1 + \frac{3}{2} + \frac{6}{6} + \frac{1}{4} = 3\frac{3}{4}.$$

But since we know that *every* lower sum is less than *every* upper sum, we see that

$$\begin{aligned} 3\frac{3}{4} &= \lim_{n\to\infty} \mathcal{L}(\mathcal{P}_n,x^3) \\ &\leq \sup_{\mathcal{P}} \mathcal{L}(\mathcal{P},x^3) \leq \inf_{\mathcal{P}} \mathcal{U}(\mathcal{P},x^3) \\ &\leq \lim_{n\to\infty} \mathcal{U}(\mathcal{P}_n,x^3) = 3\frac{3}{4} \end{aligned}$$

and so the area of $\{(x,y) \mid 0 \leq y \leq x^3,\ 1 \leq x \leq 2\}$ is well-defined and equals $3\frac{3}{4}$.

We note briefly that once we drop continuity of f, the area of S *need not* be well-defined. For example, the function

$$f(x) = \begin{cases} 1 & \text{if } x \text{ is rational,} \\ 0 & \text{if } x \text{ is irrational} \end{cases}$$

has, for *every* nontrivial interval I_j,

$$\inf_{I_j} f = 0, \quad \sup_{I_j} f = 1$$

so that for *any* finite partition $\mathcal{P}$ of a nontrivial interval $[a, b]$ ($a < b$),

$$\mathcal{L}(\mathcal{P}, f) = 0 < b - a = \mathcal{U}(\mathcal{P}, f)$$

and hence the area under the graph of f is *not* well defined in the sense above.[3] We shall return to this phenomenon later.

Actually, we can extend the context further. Our calculations were motivated by the problem of finding the area of the region S below the graph of f, and we needed f to be nonnegative on $[a, b]$ to even define S. But the calculations *themselves* make formal sense even when f takes negative values, and lead to a useful object.

So we formulate the more general concept of the **definite integral** of f over $[a, b]$ as follows:

Definition 5.1.3. *Suppose f is bounded on the closed interval $[a, b]$. For every partition of $[a, b]$,*

$$\mathcal{P} = \{a = p_0 < p_1 < \cdots < p_n = b\}$$

with component intervals

$$I_j = [p_{j-1}, p_j] \quad j = 1, \ldots, n$$

*of length $\|I_j\| = p_j - p_{j-1} = \triangle x_j$, define the **lower sum** (resp., **upper sum**) associated to $\mathcal{P}$*

$$\mathcal{L}(\mathcal{P}, f) = \sum_{j=1}^{n} \inf_{I_j} f \triangle x_j,$$

$$\mathcal{U}(\mathcal{P}, f) = \sum_{j=1}^{n} \sup_{I_j} f \triangle x_j.$$

*Then f is **(Riemann) integrable** over $[a, b]$ if*

$$\sup_{\mathcal{P}} \mathcal{L}(\mathcal{P}, f) = \inf_{\mathcal{P}} \mathcal{U}(\mathcal{P}, f)$$

*(the supremum and infimum are taken over all finite partitions $\mathcal{P}$ of $[a, b]$). In this case, we call the number above the **definite (Riemann) integral** of f over $[a, b]$,[4] denoted by*

$$\int_a^b f(x)\, dx = \sup_{\mathcal{P}} \mathcal{L}(\mathcal{P}, f) = \inf_{\mathcal{P}} \mathcal{U}(\mathcal{P}, f).$$

This is, as it stands, a highly abstract definition, and with few exceptions it would be foolhardy to try to calculate the integral (or even verify integrability) directly from the definition. In the next section, we will continue in an abstract vein to deduce some consequences of this definition which will help us better handle the concept. In particular, we will find a large class of functions which are guaranteed to be integrable. Then, in § 5.3, we will establish the Fundamental Theorem of Calculus, which leads to a practical way of calculating the integrals of many functions, a process we will explore at length in §§5.4–5.6.

---

[3] The Lebesgue theory does handle this example.

[4] or "from a to b"

Exercises for § 5.1

Answers to Exercises 1, 2ac are given in Appendix B.

Practice problems:

1. Let $S = \left\{(x, y) \,|\, 1 \leq x \leq 2,\ 0 \leq y \leq x^2\right\}$ be the region bounded below by the x-axis, above by the graph $y = x^2$ of the function $f(x) = x^2$ and on the sides by the lines $x = 1$ and $x = 2$. Let $\mathcal{P}_1 = \left\{1, \frac{3}{2}, 2\right\}$ be the partition of $[1, 2]$ into two equal parts, and $\mathcal{P}_2 = \left\{1, \frac{4}{3}, \frac{5}{3}, 2\right\}$ the partition of $[1, 2]$ into three equal parts.
   (a) Calculate the upper and lower sums $\mathcal{L}(\mathcal{P}_1, f)$ and $\mathcal{U}(\mathcal{P}_1, f)$. (*Hint:* The function f is strictly increasing on $[1, 2]$.)
   (b) Calculate the upper and lower sums $\mathcal{L}(\mathcal{P}_2, f)$ and $\mathcal{U}(\mathcal{P}_2, f)$.
   (c) Let $\mathcal{P}$ be the mutual refinement of $\mathcal{P}_1$ and $\mathcal{P}_2$. Calculate $\mathcal{L}(\mathcal{P}, f)$ and $\mathcal{U}(\mathcal{P}, f)$ and verify equation (5.10) in this case.
   (d) Let $\mathcal{P}' = \{1, 1.2, 1.6, 2\}$. Calculate the upper and lower sums $\mathcal{L}(\mathcal{P}', f)$ and $\mathcal{U}(\mathcal{P}', f)$

2. For each function f and interval $[a, b]$ below, (i) write down a formula for the points x_i of the partition $\mathcal{P}_n$ of $[a, b]$ into n equal pieces; (ii) write an expression for $\mathcal{L}(\mathcal{P}_n, f)$ and $\mathcal{U}(\mathcal{P}_n, f)$ as an explicit sum; (iii) express each sum as a combination of sums of the form $A(n) + B(n) \sum_{i=1}^{n} i + C(n) \sum_{i=1}^{n} i^2$; (iv) use Equations (5.3)–(5.5) to write these as functions of n alone; (v) find the limits of these as $n \to \infty$ (*Hint:* Your answer for lower and upper sums should be the same; then the argument we used in the text for $f(x) = x^3$ shows that this limit equals the definite integral $\int_a^b f(x)\, dx$.)
   (a) $f(x) = x$, $[a, b] = [0, 1]$
   (b) $f(x) = x^2 + 1$, $[a, b] = [0, 1]$
   (c) $f(x) = 2x^2 - 1$, $[a, b] = [-2, 0]$
   (d) $f(x) = 1 - x^2$, $[a, b] = [-2, 2]$ (*Hint:* Consider the sums over $[-2, 0]$ and $[0, 2]$ separately.)

Theory problems:

3. Suppose we know that the area of a union of disjoint regions equals the sum of the areas of the pieces (equation (5.1)). Use this to show that the area of a subregion is at most equal to the area of the whole region (equation (5.2)).

4. Starting with the assumption that the area of a square of side 1 inch is 1 square inch, and the additive property of area (equation (5.1)), derive
   (a) The formula for the area of a rectangle with length and width an *integer* number of inches;

06.05.2020

E3) Calculus deconstructed_a second course in first-year...pdf

Pág. 135(152) 4 Differentiation

Pág. 137(153) Figure 4.1 $Ax+By=C$

Pág. 247(264) 5 Integration.

E4) Geometry Illuminated - An Illustrated Introduction...pdf

Obs. Livro todo!!!

Pág. 105(122) Part II Euclidean Geometry

Pág. 223(240) Part III Euclidean Transformations

↳ muito bom!!!!

Obs.: Aplicar 'quando' (no momento) da projeção da superfície curva sobre um plano.
(simular vice-versa também)

Pág. 395(412) Part IV Hyperbolic Geometry.

E5) Calculus in 3D. Geometry, vectors, and...pdf

Obs. Livro todo!!!

Pág. 115(127) Chapter 2
Curves and Vector-Valued Functions of One Variable.

%)*+,)))-) ' . A & !> *
& ! & ! F ! ! @A6 7 . A 6!
(&& F ! . A

Part II

Euclidean Geometry

The goal of this book is to provide an introduction to both Euclidean and non-Euclidean (hyperbolic) geometry. The two geometries share many features, but they have fundamental and radical differences. Neutral geometry is the part of the path they have in common and that is what we have been studying so far, but we have now come to the fork in the path. That fork comes when you try to answer this question:

> Given a line ℓ and a point P that is not on ℓ, how many lines pass through P and are parallel to ℓ?

Using just the axioms of neutral geometry, we can prove that there is always at least one such parallel. We can also prove that if there is more than one parallel, then there must be infinitely many. But that is the extent of what the neutral axioms can say. The neutral axioms are not enough to determine whether there is one parallel or many. A single axiom is what separates Euclidean and non-Euclidean geometry: the final axiom of Euclidean geometry calls for exactly one parallel, the final axiom of non-Euclidean geometry calls for more than one parallel.

The next several chapters are devoted to Euclidean geometry. Euclidean geometry is very old, and so there is a lot of it. We only cover the fundamentals in this book, but there are many excellent books that do much more, such as *Geometry Revisited* [CG67] by Coxeter and Greitzer.

11

The Axiom on Parallels

Euclidean geometry is distinguished from neutral geometry by one axiom, one that describes the nature of parallel lines. There are many formulations of the parallel axiom for Euclidean geometry, but the one that gets right to the heart of the matter is *Playfair's axiom*, named after the Scottish mathematician John Playfair.

> *(Playfair's axiom) Let ℓ be a line, and let P be a point that is not on ℓ. Then there is exactly one line through P that is parallel to ℓ.*

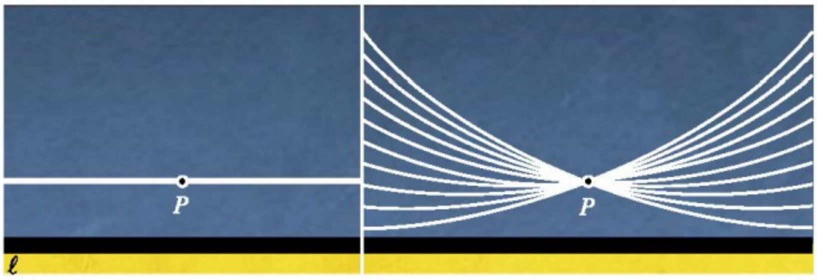

Figure 1. On the left, Euclidean geometry: a single line through P parallel to ℓ. On the right, more than one line through P parallel to ℓ.

In this chapter, we look at a small collection of theorems that are almost immediate consequences of Playfair's axiom, and as such are at the very core of Euclidean geometry. The first of these is Euclid's fifth postulate. This is the controversial postulate in the *Elements*, but also the one that guarantees the same parallel behavior that Playfair's axiom provides. Euclid's fifth is a little unwieldy, particularly when compared to Playfair's axiom, but it is the historical impetus for much of what followed. So let's use Playfair's axiom to prove Euclid's fifth postulate.

Theorem 11.1 *(Euclid's fifth postulate) If lines ℓ_1 and ℓ_2 are crossed by a transversal t, and the sum of adjacent interior angles on one side of t measure less than* 180°, *then ℓ_1 and ℓ_2 intersect on that side of t.*

Proof First some labels.

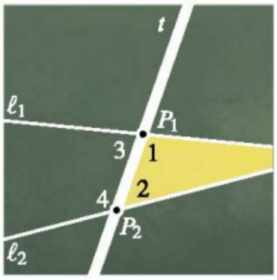

Figure 2. The labels.

Start with lines ℓ_1 and ℓ_2 crossed by transversal t. Label P_1 and P_2, the points of intersection of t with ℓ_1 and ℓ_2 respectively. On one side of t, the two adjacent interior angles should add up to less than 180°. Label the one at P_1 as $\angle 1$ and the one at P_2 at $\angle 2$. Label the supplement of $\angle 1$ as $\angle 3$ and label the supplement of $\angle 2$ as $\angle 4$.

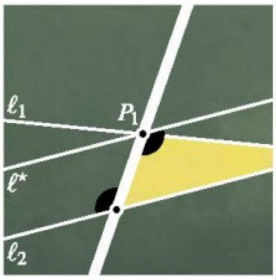

Figure 3. Constructing the unique parallel.

This theorem asserts that the lines intersect, and says where the intersection occurs. We will tackle those two statements in that order. Note that $\angle 1$ and $\angle 4$ are alternate interior angles, but they are not congruent: if they were, their supplements $\angle 2$ and $\angle 3$ would be too, and then

$$(\angle 1) + (\angle 2) = (\angle 1) + (\angle 3) = 180^\circ.$$

There is, however, another line ℓ^* through P_1 that forms an angle congruent to $\angle 4$ (because of the angle construction axiom), and by the alternate interior angle theorem, ℓ^* must be parallel to ℓ_2. Because of Playfair's axiom, ℓ^* is the only parallel to ℓ_2 through P_1. Therefore ℓ_1 intersects ℓ_2.

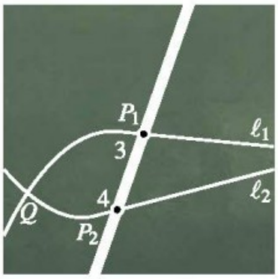

Figure 4. An impossible triangle on the wrong side of t.

The second part of the proof is to figure out on which side of t that ℓ_1 and ℓ_2 cross. Let's see what would happen if they intersected at a point Q on the wrong side of t, the side with $\angle 3$ and $\angle 4$. Then ΔP_1P_2Q would have two interior angles, $\angle 3$ and $\angle 4$, whose measures add up to more than 180°. This would violate the Saccheri-Legendre theorem. So ℓ_1 and ℓ_2 cannot intersect on the side of t with $\angle 3$ and $\angle 4$ and so they must intersect on the side with $\angle 1$ and $\angle 2$.

□

One of the truly useful theorems of neutral geometry is the alternate interior angle theorem. In fact, we just used it in the last proof. You may recall from high school geometry that the converse of the theorem is often even more useful. The problem is that the converse of the alternate interior angle theorem can't be proved using just the axioms of neutral geometry. It depends on the Euclidean behavior of parallel lines.

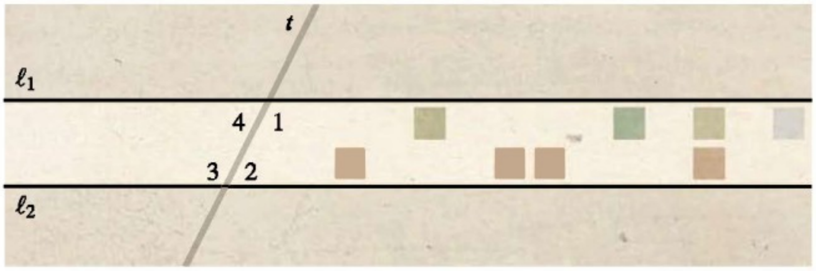

Figure 5. The converse of the alternate interior angle theorem: if $\ell_1 \parallel \ell_2$, then $\angle 1 \simeq \angle 3$ and $\angle 2 \simeq \angle 4$.

Theorem 11.2 *(Converse of the alternate interior angle theorem) If ℓ_1 and ℓ_2 are parallel, then the pairs of alternate interior angles formed by a transversal t are congruent.*

Proof Consider two parallel lines crossed by a transversal. Label adjacent interior angles: $\angle 1$ and $\angle 2$, and $\angle 3$ and $\angle 4$, so that $\angle 1$ and $\angle 4$ are supplementary and $\angle 2$ and $\angle 3$ are supplementary. That means that the pairs of alternate interior angles are $\angle 1$ and $\angle 3$, and $\angle 2$ and $\angle 4$. Now, we just have to do a little arithmetic. From the two pairs of supplementary angles:

(*i*): $(\angle 1) + (\angle 4) = 180°$
(*ii*): $(\angle 2) + (\angle 3) = 180°$.

Adding all four angles together,

$$(\angle 1) + (\angle 2) + (\angle 3) + (\angle 4) = 360°.$$

Now we turn to Euclid's fifth postulate, or rather its contrapositive. The lines ℓ_1 and ℓ_2 are parallel, which means that they do not intersect on either side of t. Therefore Euclid's fifth says that on neither side of t may the sum of adjacent interior angles be less than 180°:

- $(\angle 1) + (\angle 2) \geq 180°$
- $(\angle 3) + (\angle 4) \geq 180°$.

If either sum was greater than 180°, though, the sum of all four angles would have to be more than 360°. We saw above that is not the case, so the inequalities are actually equalities:

(*iii*): $(\angle 1) + (\angle 2) = 180°$
(*iv*): $(\angle 3) + (\angle 4) = 180°$.

Now we have two systems of equations with four unknowns, and it is basic algebra from here. Subtract *(iv)* from *(i)* to get

$$(\angle 1) - (\angle 3) = 0 \implies (\angle 1) = (\angle 3).$$

Subtract *(i)* from *(iii)* to get

$$(\angle 2) - (\angle 4) = 0 \implies (\angle 2) = (\angle 4).$$

Therefore the alternate interior angles are congruent. □

One of the key theorems we proved in the neutral geometry section was the Saccheri-Legendre theorem: that the angle sum of a triangle is at most 180°. That is all you can say

with the axioms of neutral geometry, but in a world with Playfair's axiom and the converse of the alternate interior angle theorem, we can make a more bold statement about angle sums of triangles.

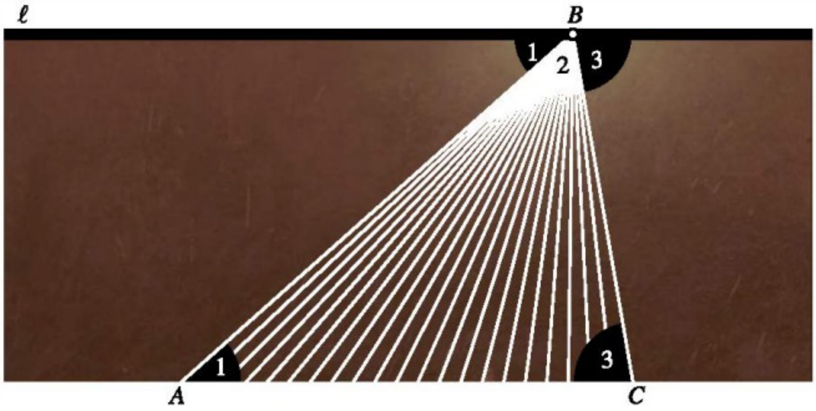

Figure 6. Use the converse of the alternate interior angle theorem to find the angle sum of a triangle.

Theorem 11.3 *The angle sum of a triangle is* 180°.

Proof Given $\triangle ABC$, by Playfair's axiom, there is a unique line ℓ through B that is parallel to $\overline{AC}$. That line and the rays $\overrightarrow{BA}$ and $\overrightarrow{BC}$ form three angles, $\angle 1$, $\angle 2$, and $\angle 3$ as shown in figure 6. By the converse of the alternate interior angle theorem, two pairs of alternate interior angles are congruent: $\angle 1 \simeq \angle A$ and $\angle 3 \simeq \angle C$. Therefore, the angle sum of $\triangle ABC$ is

$$s(\triangle ABC) = (\angle A) + (\angle B) + (\angle C) = (\angle 1) + (\angle 2) + (\angle 3) = 180^\circ. \qquad \square$$

So, the three angles in an equilateral triangle all measure 60° and the measures of the two acute angles in a right triangle must add up to 90°.

In the last chapter on quadrilaterals we talked a little bit about the uncertain status of rectangles in neutral geometry, saying that it is easy to make a convex quadrilateral with three right angles, but that once you have done that, there is no guarantee that the fourth angle will be a right angle. Here it is now in the Euclidean context:

Theorem 11.4 *Let $\angle ABC$ be a right angle. Let r_A and r_C be rays so that r_A has endpoint A, is on the same side of $\overline{AB}$ as C, and is perpendicular to $\overline{AB}$; r_C has endpoint C, is on the same side of $\overline{BC}$ as A, and is perpendicular to $\overline{BC}$. Then r_A and r_C intersect at a point D, and the angle formed at this intersection, $\angle ADC$, is a right angle. Therefore $\square ABCD$ is a rectangle.*

Proof First we need to make sure that r_A and r_C intersect. Let ℓ_A and ℓ_C be the lines containing r_A and r_C respectively. By the alternate interior angle theorem, the right angles at A and B imply that ℓ_A and $\overline{BC}$ are parallel. So $\overline{BC}$ is the one line parallel to ℓ_A through C, and that implies that ℓ_C cannot be parallel to ℓ_A: it has to intersect ℓ_A. Let's call that point of intersection D. This

theorem claims that it is the rays, not the lines, that intersect. Therefore we need to rule out the possibility that the intersection of ℓ_A and ℓ_C might happen on one (or both) of the opposite rays. Since ℓ_A is parallel to $\overline{BC}$, all the points of ℓ_A are on the same side of $\overline{BC}$ as A. None of the points of r_C^{op} is on that side of BC, so D cannot be on r_C^{op}. Likewise, all the points of ℓ_C are on the same side of $\overline{AB}$ as C. None of the points of r_A^{op} is on that side of AB, so D cannot be on r_A^{op}.

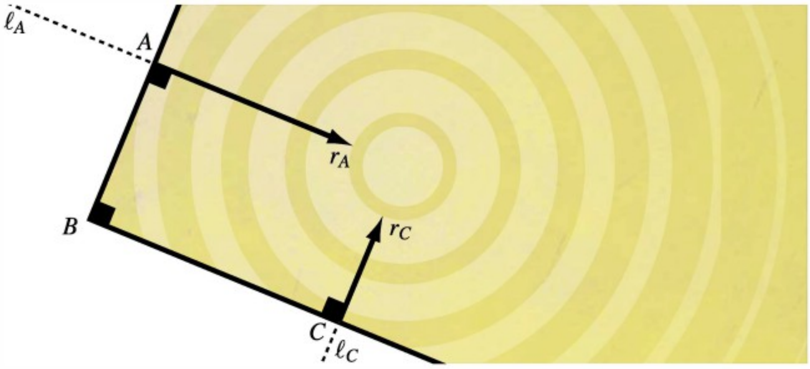

Figure 7. Constructing a rectangle.

So now $\square ABCD$ has three right angles, $\angle A$, $\angle B$, and $\angle C$. It is actually a convex quadrilateral too (verifying this is left as an exercise), so the diagonal AC divides $\square ABCD$ into two triangles $\triangle ABC$ and $\triangle ADC$ (see figure 8). Then, since the angle sum of a triangle is 180°,

$$
\begin{gathered}
s(\triangle ABC) + s(\triangle ADC) = 180^\circ + 180^\circ \\
(\angle CAB) + (\angle B) + (\angle ACB) + (\angle CAD) + (\angle D) + (\angle ACD) = 360^\circ \\
(\angle A) + (\angle B) + (\angle C) + (\angle D) = 360^\circ \\
90^\circ + 90^\circ + 90^\circ + (\angle D) = 360^\circ \\
(\angle D) = 90^\circ.
\end{gathered}
$$

Therefore $\angle D$ is a right angle. □

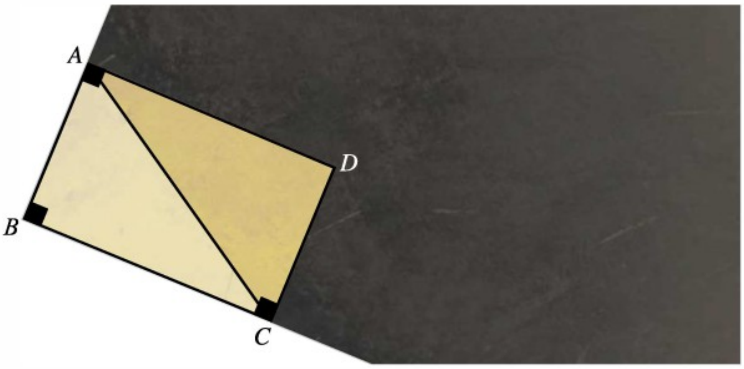

Figure 8. Verifying that the fourth angle is right.

Exercise 11.1 Fill in the gap in the proof by explaining why $\square ABCD$ must be convex.

We see then that, yes, rectangles exist in Euclidean geometry. Two essential properties of rectangles are provided in the following (easy) exercises.

Exercise 11.2 Let $\square ABCD$ be a rectangle. Prove that $\overline{AB}$ is parallel to $\overline{CD}$ and $\overline{AD}$ is parallel to $\overline{BC}$.

Exercise 11.3 Let $\square ABCD$ be a rectangle. Prove that $AB \simeq CD$ and $AD \simeq BC$ and $AC \simeq BD$.

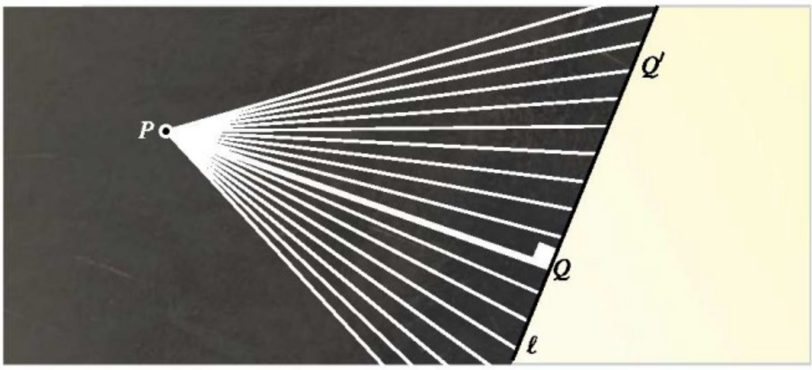

Figure 9. The distance from P to ℓ is measured along the segment from P to ℓ that is perpendicular to ℓ.

For the last result of this section, we return to parallel lines. One of the things that we will see when we study non-Euclidean geometry is that parallel lines tend to diverge from each other. That doesn't happen in Euclidean geometry. This is one of the key differences between the two geometries. Let's make this more precise. Suppose that P is a point that is not on a line ℓ. Define the distance from P to ℓ to be the minimum distance from P to a point on ℓ:

$$d(P,\ell) = \min\left\{ |PQ| \;\middle|\; Q \text{ is on } \ell \right\}.$$

Exercise 11.4 Prove that the minimum actually occurs when Q is the foot of the perpendicular to ℓ through P (this can be proved using solely the axioms of neutral geometry).

For a pair of parallel lines, that distance as measured along perpendiculars does not change.

Theorem 11.5 *If ℓ and ℓ' are parallel lines, then the distance from a point on ℓ to ℓ' is constant. In other words, if P and Q are points on ℓ, then*

$$d(P,\ell') = d(Q,\ell').$$

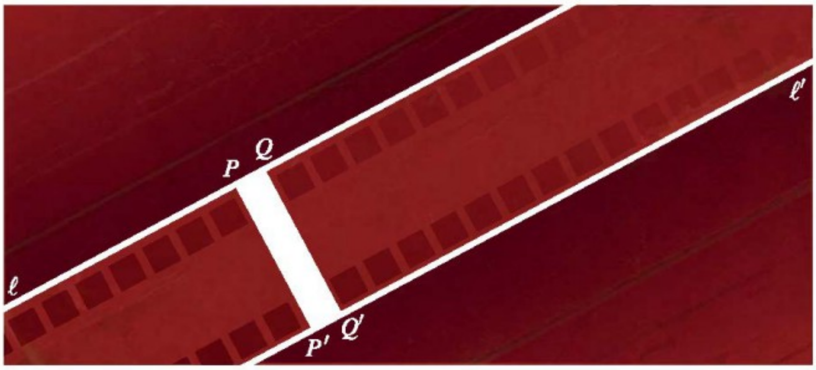

Figure 10. Perpendicular segments PP' and QQ' between parallel lines have the same length. That is, parallel lines are equidistant.

Proof Let P' and Q' be the feet of the perpendiculars on ℓ' from P and Q respectively. That way, by the previous exercise,

$$d(P,\ell') = |PP'| \text{ and } d(Q,\ell') = |QQ'|.$$

Then $\angle PP'Q'$ and $\angle QQ'P'$ are right angles. By the converse of the alternate interior angle theorem, $\angle P$ and $\angle Q$ are right angles too, so $\square PQQ'P'$ is a rectangle. According to exercise 11.2, PP' and QQ', which are opposite sides of a rectangle, are congruent. □

11.1 Exercises

11.5. Let ℓ be a line and P a point not on ℓ. Using only the results of neutral geometry, prove that there is at least one line through P that is parallel to ℓ.

11.6. Suppose that ℓ_1, ℓ_2, and ℓ_3 are three distinct lines such that ℓ_1 and ℓ_2 are parallel, and ℓ_2 and ℓ_3 are parallel. Prove that ℓ_1 and ℓ_3 are parallel.

11.7. Suppose that ℓ_1 and ℓ_2 are parallel lines, and that t is a line that intersects (but is distinct from) ℓ_1. Prove that t intersects ℓ_2.

11.8. Given three distinct parallel lines, prove that there is a labeling of them as ℓ_1, ℓ_2, and ℓ_3 so that

- no point of ℓ_1 lies between a point of ℓ_2 and a point of ℓ_3
- every point of ℓ_2 lies between a point of ℓ_1 and of ℓ_3
- no point of ℓ_3 lies between a point of ℓ_1 and a point of ℓ_2.

In this case, we can say that ℓ_2 is *between* ℓ_1 and ℓ_3.

11.9. This exercise is an extension of the previous one. Given n distinct parallel lines, prove that there is an ordering of them as $\ell_1, \ell_2, \ldots, \ell_n$ so that if $i < j < k$ then ℓ_j is between ℓ_i and ℓ_k.

11.10. Find the angle sum of a convex n-gon as a function of n.

11.11. What is the measure of an interior angle of a regular n-gon (as a function of n)?

11.12. Consider a convex quadrilateral $\square ABCD$. Prove that the two diagonals of $\square ABCD$ bisect each other if and only if $\square ABCD$ is a parallelogram.

11.13. Show that a parallelogram $\square ABCD$ is a rectangle if and only if $AC \simeq BD$.

11.14. Suppose that the diagonals of a convex quadrilateral $\square ABCD$ intersect one another at a point P and that

$$AP \simeq BP \simeq CP \simeq DP.$$

Prove that $\square ABCD$ is a rectangle.

11.15. Suppose that the diagonals of a convex quadrilateral bisect one another at right angles. Prove that the quadrilateral must be a rhombus.

11.16. Let $\triangle ABC$ be isosceles with $AB \simeq AC$, and let a be the midpoint of BC. Let b and c be points on the line through A that is parallel to BC so that $b * A * c$ and so that $bA \simeq cA \simeq Ba$. Prove that $\triangle ABC \simeq \triangle abc$.

11.17. Verify that the Cartesian model (as developed through the exercises in chapters 1–3) satisfies Playfair's axiom.

12

Parallel Projection

In the course of this chapter, we will need to use a few properties of parallelograms. It is rather common to confuse the properties of a parallelogram with its definition, so first, recall the definition of a parallelogram.

Definition 12.1 A *parallelogram* is a simple quadrilateral whose opposite sides are parallel.

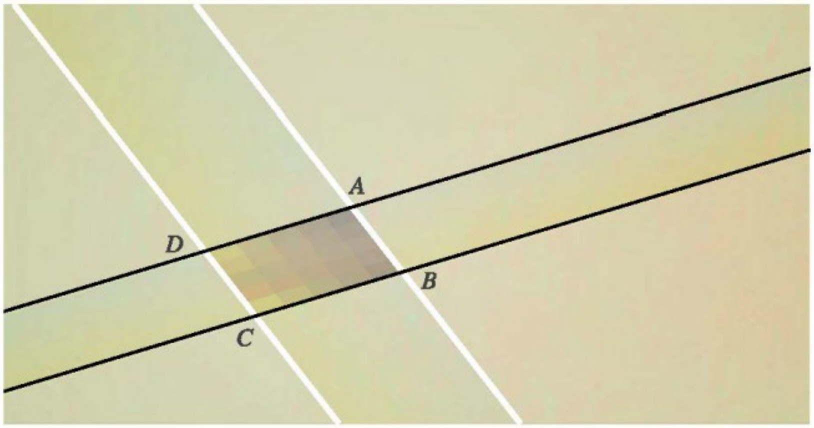

Figure 1. A parallelogram.

Now on to the properties of parallelograms that we will need for this chapter.

Exercise 12.1 Prove that in a parallelogram, the two pairs of opposite sides are congruent and the two pairs of opposite angles are congruent.

Exercise 12.2 Prove that if a convex quadrilateral has one pair of opposite sides that are both parallel and congruent, then it is a parallelogram.

Exercise 12.3 Let $\square ABB'A'$ be a simple quadrilateral. Verify that if AA' and BB' are parallel, but AB and $A'B'$ are not, then AA' and BB' cannot be congruent.

12.1 Parallel projection

The purpose of this chapter is to introduce a mechanism called *parallel projection*, a particular kind of mapping from the points on one line to the points on another. Parallel projection is the piece of machinery that you have to have in place to understand similarity, which is in turn essential for much of what we will be doing in the next chapters. The primary goal at this point is to understand how distances between points may be distorted by the parallel projection mapping. Once that is figured out, we will be able to turn our attention to the geometry of similarity.

Definition 12.2 A *parallel projection* from one line ℓ to another ℓ' is a map $\Phi : \ell \to \ell'$ that assigns to each point P on ℓ a point $\Phi(P)$ on ℓ' so that all the lines connecting a point and its image are parallel to one another.

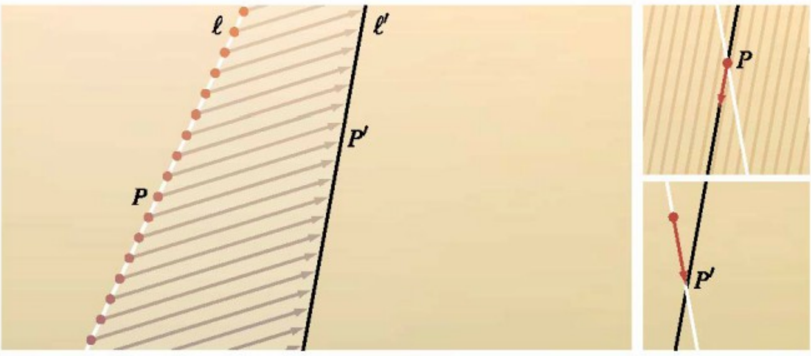

Figure 2. The path from a point P on ℓ to a point P' on ℓ' defines a parallel projection as long as neither P nor P' is the intersection of ℓ and ℓ' (as shown at right).

It is easy to construct parallel projections. Any point P on ℓ and its image $\Phi(P)$ on ℓ' completely determines the projection: for any other point Q on ℓ there is a unique line that passes through Q and is parallel to $\overline{P\,\Phi(P)}$. Wherever the line intersects ℓ' will have to be $\Phi(Q)$. There are only two cases where this construction will not work: if P is the intersection of ℓ and ℓ', then the lines of projection run parallel to ℓ' and so fail to provide a point of intersection; and if $\Phi(P)$ is the intersection of ℓ and ℓ', then the lines of projection actually coincide rather than being parallel.

Theorem 12.3 *A parallel projection is both one-to-one and onto.*

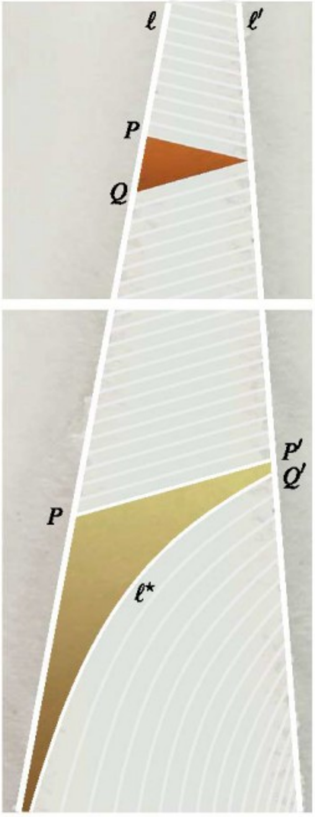

Figure 3. Using proofs by contradiction, we can show that parallel projection is both one-to-one (top) and onto (bottom).

Proof Let $\Phi : \ell \to \ell'$ be a parallel projection. First let's see why Φ is one-to-one. Suppose that it is not. That is, suppose that P and Q are two distinct points on ℓ but that $\Phi(P) = \Phi(Q)$. Then the two projecting lines $\overline{P\,\Phi(P)}$ and $\overline{Q\,\Phi(Q)}$, which ought to be parallel, actually share a point. This can't happen.

Now let's see why Φ is onto. Suppose that Q' is a point on ℓ' that is not in the image of Φ. To get a sense of how Φ is casting points from ℓ to ℓ', consider a point P on ℓ and its image $P' = \Phi(P)$ on ℓ'. By Playfair's axiom, there is a unique line ℓ^* that passes through Q' and is parallel to $\overline{PP'}$. If ℓ^* intersects ℓ at a point Q, then $\Phi(Q) = Q'$, contradicting the initial supposition that Q' is not in the image of Φ. If ℓ^* does not intersect ℓ, then both ℓ and $\overline{PP'}$ are lines that pass through P and are parallel to ℓ^*, in violation of Playfair's axiom. In either case, we arrive at a contradiction. Therefore, an arbitrary point Q' of ℓ must be in the image of Φ, and so Φ is onto. $\square$

Now that we have established that parallel projection is a bijection, it helps to use a naming convention to makes things more readable. I will use a prime mark $'$ to indicate the parallel projection of a point, so $\Phi(P) = P'$, $\Phi(Q) = Q'$, and so on.

12.2 Parallel projection, order, and congruence

So far we have seen that parallel projection establishes a correspondence between the points of one line and the points of another. What about the order of the points? Can points get shuffled up in the process of a parallel projection? No.

Theorem 12.4 *Let $\Phi : \ell \to \ell'$ be a parallel projection. If A, B, and C are points on ℓ and $A * B * C$, then $A' * B' * C'$.*

Proof Because B is between A and C, A and C must be on opposite sides of $\overline{BB'}$. But, $\overline{AA'}$ does not intersect $\overline{BB'}$, so A' has to be on the same side of $\overline{BB'}$ as A; and $\overline{CC'}$ does not

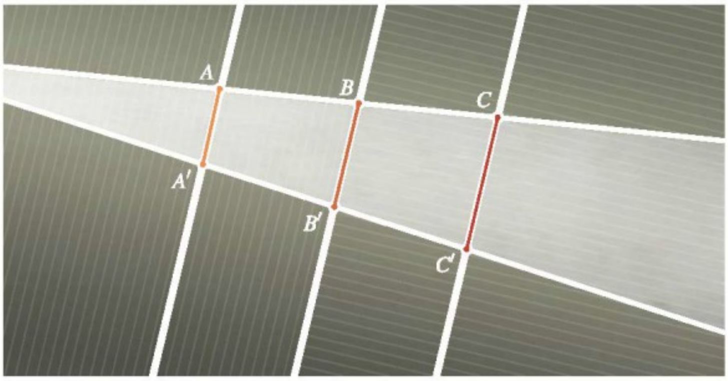

Figure 4. Parallel projection keeps points in order.

intersect $\overline{BB'}$, so C' has to be on the same side of $\overline{BB'}$ as C. Therefore A' and C' have to be on opposite sides of $\overline{BB'}$, and so the intersection of $\overline{BB'}$ and $A'C'$, which is B', must be between A' and C'. □

Now what about congruence?

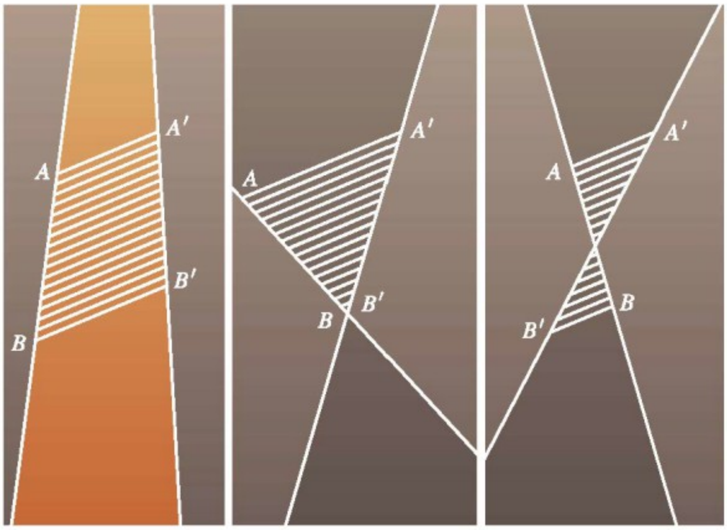

Figure 5. There are three positions for A and B relative to the image line: both on the same side, one on the image line, or one on each side. Likewise, there are three positions for a and b. There are nine cases in all.

Theorem 12.5 *Let $\Phi : \ell \to \ell'$ be a parallel projection. If a, b, A, and B are points on ℓ and if $ab \simeq AB$, then $a'b' \simeq A'B'$.*

Proof There are several cases, depending on the positions of ab and AB relative to ℓ'. They could lie on the same side of ℓ', they could lie on opposite sides of ℓ', one or both could straddle

ℓ', or one or both could have an endpoint on ℓ'. Each case must be handled differently. We will look at the case where both segments are on the same side of ℓ'. The others are left as exercises. There are still two cases to consider, depending on whether or not ℓ and ℓ' are parallel.

Case 1: when ℓ and ℓ' are parallel (see figure 6). This is the simpler case, but it helps illuminate the more general case. Because aa' is parallel to bb' and ab is parallel to $a'b'$, $\square aa'b'b$ is a parallelogram; AA' is parallel to BB' and AB is parallel to $A'B'$ so $\square AA'B'B$ is also a parallelogram. Because the opposite sides of a parallelogram are congruent (exercise 12.1), $a'b' \simeq ab$ and $AB \simeq A'B'$. Since $ab \simeq AB$, $a'b' \simeq A'B'$.

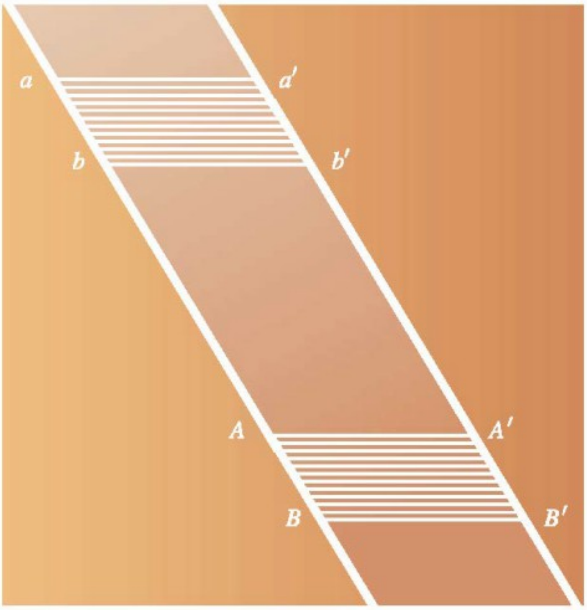

Figure 6. Case 1, when the two lines are parallel.

Case 2: when ℓ and ℓ' are not parallel (see figure 7). This is the more likely case, in which $\square aa'b'b$ and $\square AA'B'B$ will not be parallelograms. We can takes a similar approach to case 1, though, by building some parallelograms into the problem. Because ℓ and ℓ' are not parallel, the segments aa' and bb' cannot have the same length (exercise 12.3), and the segments AA' and BB' cannot have the same length. Let's assume that aa' is shorter than bb' and that AA' is shorter than BB'. If this is not the case, then it is just a matter of switching some labels to make it so. Then

- there is a point c between b and b' so that $bc \simeq aa'$, and
- there is a point C between B and B' so that $BC \simeq AA'$.

This creates four shapes of interest: the quadrilaterals $\square a'abc$ and $\square A'ABC$ that are parallelograms (exercise 12.2), and the triangles $\Delta a'b'c$ and $\Delta A'B'C$. The key here is to prove that $\Delta a'b'c \simeq \Delta A'B'C$.

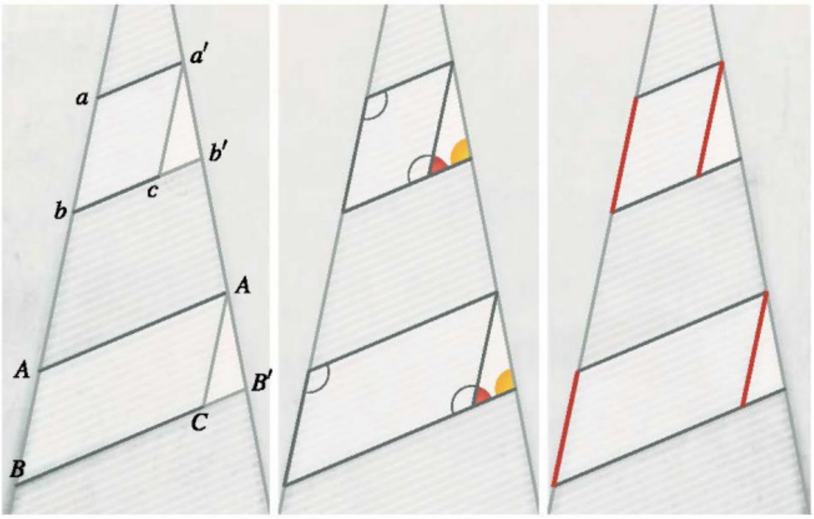

Figure 7. The non-parallel case. (l) First construct parallelograms. (c) The two angles and (r) the sides needed for AAS to show that $\Delta a'b'c' \simeq \Delta A'B'C'$.

- **A:** $\angle b' \simeq \angle B'$. The lines $\overline{cb'}$ and $\overline{CB'}$ are parallel (they are two of the projecting lines) and they are crossed by the transversal ℓ'. By the converse of the alternate interior angle theorem, $\angle a'b'c$ and $\angle A'B'C$ are congruent.
- **A:** $\angle c \simeq \angle C$. The opposite angles of the two parallelograms are congruent. Therefore $\angle a'cb \simeq \angle a'ab$ and $\angle A'AB \simeq \angle A'CB$. But $\overline{aa'}$ and $\overline{AA'}$ are parallel lines cut by the transversal ℓ, so $\angle a'ab \simeq \angle A'AB$, so $\angle a'cb \simeq \angle A'CB$, and so their supplements $\angle a'cb'$ and $\angle A'CB'$ are also congruent.
- **S:** $a'c \simeq A'C$. The opposite sides of the two parallelograms are congruent too. Therefore $a'c \simeq ab$ and $AB \simeq A'C$, and since $ab \simeq AB$, we get $a'c \simeq A'C$.

By AAS, then, $\Delta a'b'c \simeq \Delta A'B'C$. The corresponding sides $a'b'$ and $A'B'$ have to be congruent.

□

Exercise 12.4 Fill in some gaps:

(1) Prove the theorem in the case when both a and A lie on ℓ' (at the intersection of ℓ and ℓ').
(2) Prove the theorem in the case when a and A lie on opposite sides of ℓ' from b and B.

12.3 Parallel projection and distance

That brings us to the question at the very heart of parallel projection. If Φ is a parallel projection and A and B are two points on ℓ, how do the lengths $|AB|$ and $|A'B'|$ compare? In case 1 of the last proof, AB and $A'B'$ ended up being congruent, but that was because ℓ and ℓ' were parallel. In general, AB and $A'B'$ do not have to be congruent. But it turns out that in the process of parallel projecting from one line to another, all distances are scaled by a constant multiple.

Theorem 12.6 *If $\Phi : \ell \to \ell'$ is a parallel projection, then there is a constant k such that*

$$|A'B'| = k|AB|$$

for all points A and B on ℓ.

It is helpful to get a little perspective on this before the proof. First note that the previous theorem gives us a way to narrow the scope of the problem. Fix a point O on ℓ and let r be one of the two rays along ℓ with O as its endpoint. The segment construction axiom says that every segment AB on ℓ is congruent to a segment OP where P is some point on r. We have just seen that parallel projection maps congruent segments to congruent segments. So if Φ scales all segments of the form OP by a factor of k, then it must scale all the segments of ℓ by the same factor.

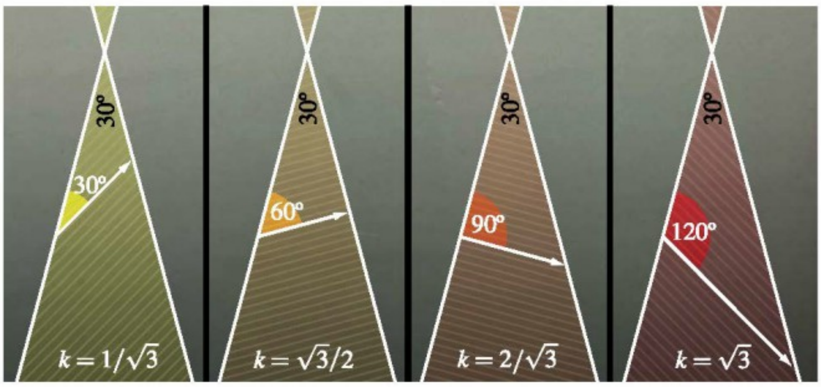

Figure 8. Some parallel projections and their scaling constants.

Second, we can extend that by considering what happens when we parallel project end-to-end congruent copies of a segment. For this, let me introduce another convenient notation convention: for the rest of this argument, let P_d denote the point on r that is a distance d from O. Now, pick a positive real value x, and let

$$k = |O'P'_x|/|OP_x|,$$

so that Φ scales the segment OP_x by a factor of k. Eventually we will have to show that Φ scales all segments by that factor, but for now let's restrict our attention to the segments OP_{nx}, where n is a positive integer. Between O and P_{nx} are P_x, P_{2x}, … $P_{(n-1)x}$ in order:

$$O * P_x * P_{2x} * \cdots * P_{(n-1)x} * P_{nx}.$$

We have seen that parallel projection preserves the order of points, so

$$O' * P'_x * P'_{2x} * \cdots * P'_{(n-1)x} * P'_{nx}.$$

Each segment $P_{ix}P_{(i+1)x}$ is congruent to OP_x and consequently each parallel projection $P'_{ix}P'_{(i+1)x}$ is congruent to $O'P'_x$. Add their lengths:

$$\begin{aligned}|O'P'_{nx}| &= |O'P'_x| + |P'_xP'_{2x}| + |P'_{2x}P'_{3x}| + \cdots + |P'_{(n-1)x}P'_{nx}| \\ &= kx + kx + kx + \cdots + kx \quad (n \text{ times}) \\ &= k \cdot nx\end{aligned}$$

and so Φ scales OP_{nx} by a factor of k.

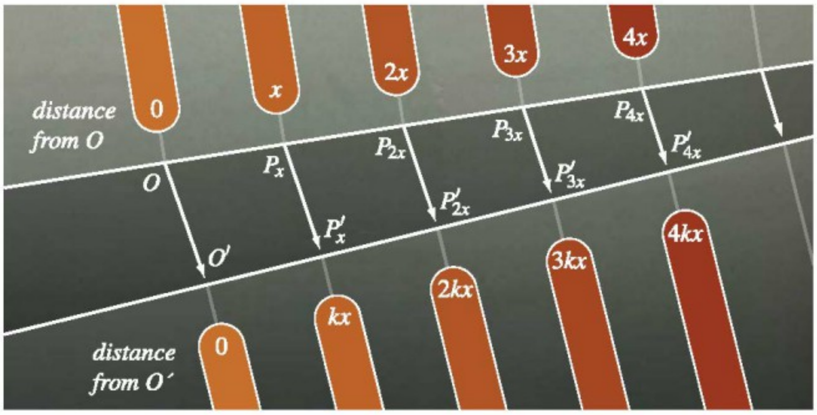

Figure 9. Scaling of segments of length kx, where k is an integer.

Sadly, no matter what x is, the points P_{nx} account for an inconsequential portion of the set of all points of r. However, if OP_x and OP_y were to have two different scaling factors we could use this end-to-end copying to magnify the difference between them. Third, then, consider an example that demonstrates how this works, and which shows how this ultimately prevents there being two different scaling factors. In this example, let's suppose that $|O'P'_1| = 2$, so all integer length segments on ℓ are scaled by a factor of 2. Let's take a look at what this means for $P_{3.45}$. Let k be the scaling factor for $OP_{3.45}$. Figure 10 shows how the first few end-to-end copies of $OP_{3.45}$ begin to limit the possible values of k.

Fourth, note that the floor function, $f(x) = \lfloor x \rfloor$, assigns to each real number x the largest integer that is less than or equal to it. The ceiling function, $f(x) = \lceil x \rceil$, assigns to each real number x the smallest integer that is greater than or equal to it.

It is finally time to prove that parallel projection scales distance.

Proof Let $k = |O'P'_1|$ so that k is the scaling factor for the segment of length 1 (and consequently for all integer length segments). Now take a point P_x on ℓ and let k' be the scaling factor for the segment OP_x. We want to show that $k' = k$ and we can do that by following the same strategy as in the example above, by capturing k' in an increasingly narrow band around k by looking at the parallel projection of P_{nx} as n increases. Start by working back and forth between

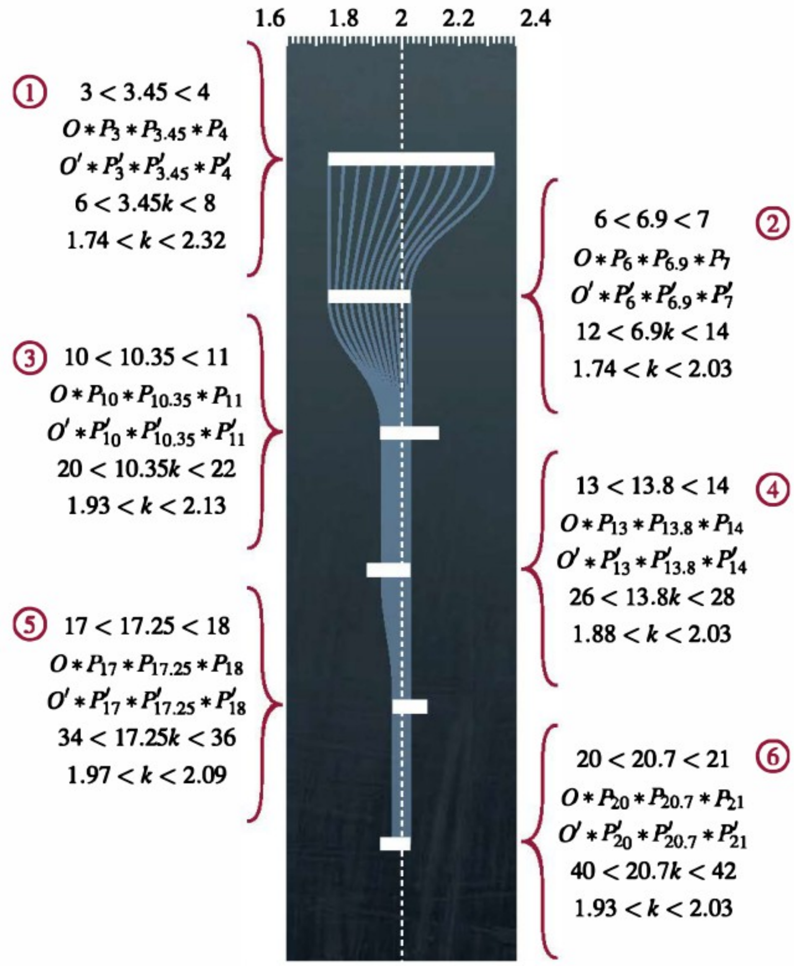

① $3 < 3.45 < 4$
$O * P_3 * P_{3.45} * P_4$
$O' * P'_3 * P'_{3.45} * P'_4$
$6 < 3.45k < 8$
$1.74 < k < 2.32$

$6 < 6.9 < 7$ ②
$O * P_6 * P_{6.9} * P_7$
$O' * P'_6 * P'_{6.9} * P'_7$
$12 < 6.9k < 14$
$1.74 < k < 2.03$

③ $10 < 10.35 < 11$
$O * P_{10} * P_{10.35} * P_{11}$
$O' * P'_{10} * P'_{10.35} * P'_{11}$
$20 < 10.35k < 22$
$1.93 < k < 2.13$

$13 < 13.8 < 14$ ④
$O * P_{13} * P_{13.8} * P_{14}$
$O' * P'_{13} * P'_{13.8} * P'_{14}$
$26 < 13.8k < 28$
$1.88 < k < 2.03$

⑤ $17 < 17.25 < 18$
$O * P_{17} * P_{17.25} * P_{18}$
$O' * P'_{17} * P'_{17.25} * P'_{18}$
$34 < 17.25k < 36$
$1.97 < k < 2.09$

$20 < 20.7 < 21$ ⑥
$O * P_{20} * P_{20.7} * P_{21}$
$O' * P'_{20} * P'_{20.7} * P'_{21}$
$40 < 20.7k < 42$
$1.93 < k < 2.03$

Figure 10. An example. Projecting end-to-end congruent copies of a segment of length 3.45 by a parallel projection that scales all integer length segments by a factor of two.

inequalities of real numbers and orders of points:

$$\lfloor nx \rfloor \leq nx \leq \lceil nx \rceil$$
$$O * P_{\lfloor nx \rfloor} * P_{nx} * P_{\lceil nx \rceil}$$
$$O' * P'_{\lfloor nx \rfloor} * P'_{nx} * P'_{\lceil nx \rceil}$$
$$k\lfloor nx \rfloor \leq k' nx \leq k\lceil nx \rceil.$$

(The notation here is a bit casual—if nx is an integer, then $\lfloor nx \rfloor = nx$, so $P_{\lfloor nx \rfloor}$ is not actually between O and P_{nx}. This does not cause any problems for the proof.) Because $(nx - 1) \leq \lfloor nx \rfloor$

and $\lceil nx \rceil \leq (nx + 1)$ we can move to a less precise but more manageable set of inequalities:

$$k(nx - 1) \leq k'nx \leq k(nx + 1).$$

Divide by nx to get

$$k\left(\frac{nx - 1}{nx}\right) \leq k' \leq k\left(\frac{nx + 1}{nx}\right).$$

As n increases, the ratios $(nx - 1)/(nx)$ and $(nx + 1)/(nx)$ both approach 1. In the limit as n goes to infinity, they are 1. Since the inequalities have to be true for all n, the only possible value for k', then, is k. □

12.4 Exercises

12.5. Suppose that Φ is a parallel projection from ℓ to ℓ'. If ℓ and ℓ' intersect at P, prove that $\Phi(P) = P$.

12.6. Prove that if ℓ and ℓ' are parallel, then the scaling factor of a parallel projection between them must be one, but that if ℓ and ℓ' are not parallel, then there is a parallel projection with every possible scaling factor k where $0 < k < \infty$.

12.7. Let $\Phi : \ell \to \ell'$ be a parallel projection. Show that the inverse map $\Phi^{-1} : \ell' \to \ell$ is also a parallel projection.

12.8. Let $\Phi_1 : \ell \to \ell'$ and $\Phi_2 : \ell' \to \ell''$ be parallel projections. Is the composition $\Phi_2 \circ \Phi_1 : \ell \to \ell''$ necessarily a parallel projection?

12.9. In chapter 6, we constructed a distance function, and one of the keys to the construction was locating the points on a ray that were a distance of $m/2^n$ from its endpoint. In Euclidean geometry, there is a construction that locates all the points on a ray that are any rational distance m/n from its endpoint. Take two non-opposite rays r and r' with a common endpoint O. Along r, lay out m congruent copies of a segment of length 1, ending at the point P_m. Along r', lay out n congruent copies of a segment of length 1, ending at the point Q_n. Mark the point Q_1 on r' that is a distance 1 from O. Verify that the line that passes through Q_1 and is parallel to $P_m Q_n$ intersects r a distance of m/n from O.

Part III

Euclidean Transformations

In the third part of this book, we will look at Euclidean geometry from a different perspective, that of Euclidean transformations. It is a point of view that has been most closely associated with Felix Klein, that the way to study some property (such as congruence) is to study the maps that preserve it. The first chapter sets the scene with a quick development of analytic geometry. Then it is on to Euclidean isometries, bijections of the Euclidean plane that preserve distance. Over several chapters we will study the isometries, and ultimately we will classify all Euclidean isometries into four types: reflections, rotations, translations, and glide reflections. Then it is time to loosen the restriction a bit to consider bijections that preserve congruence, but not necessarily distance. Finally, we will look at inversion, a type of bijection of the punctured plane (the Euclidean plane minus a point). As luck would have it, inversion provides a convenient bridge into non-Euclidean geometry.

21

Analytic Geometry

This chapter is a quick development of analytic geometry and trigonometry in the language of Euclidean geometry. I feel an obligation to provide this bridge between traditional Euclidean geometry (as we have developed it) and more contemporary analytic geometry, but you should already be comfortable with this material, so feel free to skim through it.

21.1 Analytic geometry

At the heart of analytic geometry, there is a correspondence between points and coordinates, ordered pairs of real numbers. The Cartesian approach to the correspondence is a familiar one that goes something like this. Begin with two perpendicular lines (the choice is arbitrary). These are the x- and y-axes. Their intersection is the origin O. We will want to measure signed distances from O along the axes, which means we need to assign a positive direction to them. Once directions have been chosen, each axis will be divided into two rays that share O as their common vertex: a positive axis consisting of points whose signed distance from O is positive, and a negative axis consisting of points whose signed distance from O is negative. From a geometric point of view, the choice of directions is arbitrary, but there is an established convention: the axes are assigned positive directions so that the positive y-axis is a 90° counterclockwise turn from the positive x-axis. Geometry provides no way to distinguish which direction is the counterclockwise direction, so it is a convention that must be communicated by way of illustrations (and clocks).

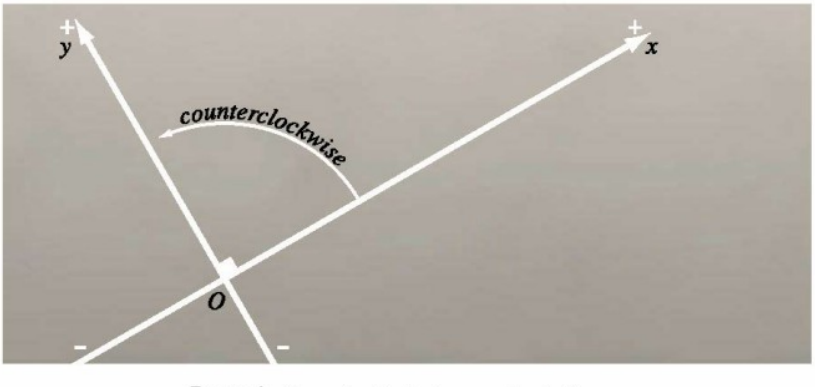

Figure 1. A counterclockwise coordinate frame.

A point P on the x-axis is assigned the coordinates $(p, 0)$, where p is the signed distance from O to P. A point Q on the y-axis is assigned the coordinates $(0, q)$ where q is the signed distance from O to Q. Most points will not lie on either axis. For their coordinates, we must consider their projections onto the axes. If R is such a point, then we draw the lines that pass through R and are perpendicular to the axes. If the points where the perpendiculars cross the axes have coordinates $(a, 0)$ and $(0, b)$, then the coordinates of R are (a, b). With this correspondence, every point corresponds to a unique coordinate pair, and every coordinate pair corresponds to a unique point.

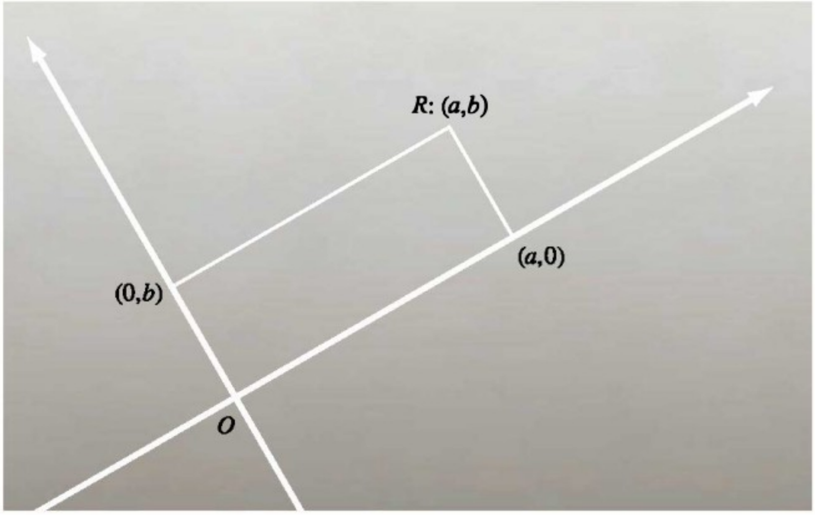

Figure 2. Assigning coordinates to a point.

The next step is to figure out how to calculate the distance between points in terms of their coordinates. This is essential for everything that we are going to do. Let's begin with two special cases.

Lemma 21.1 *(Horizontal and vertical distance) For points that share an x-coordinate, $P_1 = (x, y_1)$ and $P_2 = (x, y_2)$, $|P_1P_2| = |y_1 - y_2|$. For points that share a y-coordinate, $P_3 = (x_3, y)$ and $P_4 = (x_4, y)$, $|P_3P_4| = |x_3 - x_4|$.*

Proof We will just prove the second statement. Label two more points, $Q_1 = (x_1, 0)$ and $Q_2 = (x_2, 0)$. Then $\square P_1P_2Q_2Q_1$ is a rectangle, so its opposite sides P_1P_2 and Q_1Q_2 have to be the same length. This is where we make the direct connection between coordinates and distance: the coordinates along each axis were chosen to reflect their signed distance from the origin O. To be thorough, there are several cases to consider:

$$\begin{aligned}
O * Q_1 * Q_2 &: \quad |Q_1Q_2| = |OQ_2| - |OQ_1| = x_2 - x_1 = |x_1 - x_2| \\
O * Q_2 * Q_1 &: \quad |Q_1Q_2| = |OQ_1| - |OQ_2| = x_1 - x_2 = |x_1 - x_2| \\
Q_1 * O * Q_2 &: \quad |Q_1Q_2| = |OQ_1| + |OQ_2| = -x_1 + x_2 = |x_1 - x_2|
\end{aligned}$$

$$\begin{aligned}
Q_2 * O * Q_1: &\quad |Q_1Q_2| = |OQ_2| + |OQ_1| = -x_2 + x_1 = |x_1 - x_2| \\
Q_1 * Q_2 * O: &\quad |Q_1Q_2| = |OQ_1| - |OQ_2| = -x_1 - (-x_2) = |x_1 - x_2| \\
Q_2 * Q_1 * O: &\quad |Q_1Q_2| = |OQ_2| - |OQ_1| = -x_2 - (-x_1) = |x_1 - x_2|.
\end{aligned}$$

In all cases, $|P_1P_2| = |Q_1Q_2| = |x_1 - x_2|$. □

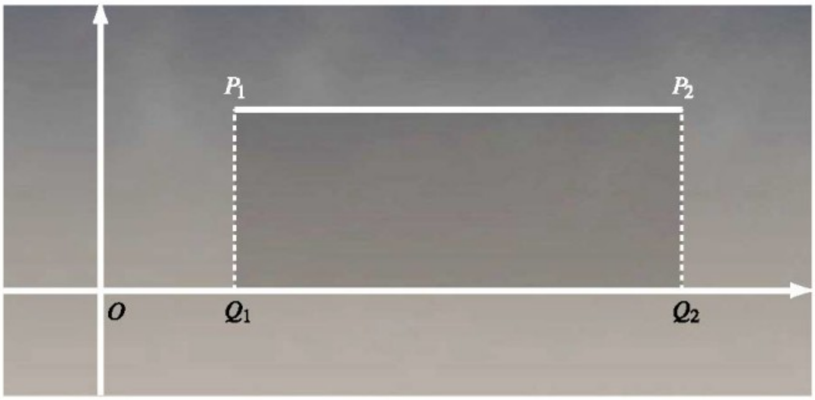

Figure 3. Measuring the distance between points on a horizontal line.

The general distance formula is now an easy consequence of the Pythagorean theorem.

Theorem 21.2 *(The distance formula) For two points $P = (x_1, y_1)$ and $Q = (x_2, y_2)$,*

$$|PQ| = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}.$$

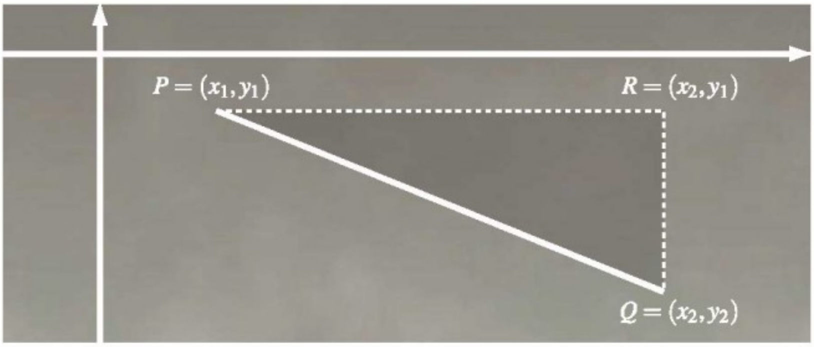

Figure 4. The distance formula is a consequence of the Pythagorean theorem.

Proof If P and Q share either x-coordinates or y-coordinates, then the formula reduces to the special case in the previous lemma (because $\sqrt{a^2} = |a|$). If not, mark one more point: $R = (x_2, y_1)$, the intersection of the horizontal line through P and the vertical line through Q.

Then $|PR| = |x_1 - x_2|$, $|RQ| = |y_1 - y_2|$, and $\triangle PRQ$ is a right triangle. By the Pythagorean theorem,

$$|PQ|^2 = |PR|^2 + |QR|^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2.$$

Take the square root to get the formula. ☐

Corollary 21.3 *(Equation of a circle) The equation of a circle $\mathcal{C}$ with center at $P = (h, k)$ and radius r is*

$$(x - h)^2 + (y - k)^2 = r^2.$$

Proof By definition, the points of $\mathcal{C}$ are those points that are a distance of r from P. Therefore (x, y) is on $\mathcal{C}$ if and only if

$$\sqrt{(x - h)^2 + (y - k)^2} = r.$$

Square both sides of the equation to get the standard form. ☐

Now let's try to find equations that will describe lines. The key is the idea that a line describes the shortest path between points. That is captured more formally in the triangle inequality, which you should recall states that $|AB| + |BC| \geq |AC|$, but that the equality only happens when $A * B * C$.

Theorem 21.4 *Given two distinct points $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ on a line ℓ, a third point $P = (x, y)$ lies on ℓ if and only if its coordinates can be written in the form*

$$x = x_1 + t(x_2 - x_1) \quad \textit{and} \quad y = y_1 + t(y_2 - y_1)$$

for some $t \in \mathbb{R}$.

Proof The orderings of P, P_1, and P_2 on the line create several cases. Let's consider only the case where t is between 0 and 1 and P is between P_1 and P_2. It is representative of the other two cases.

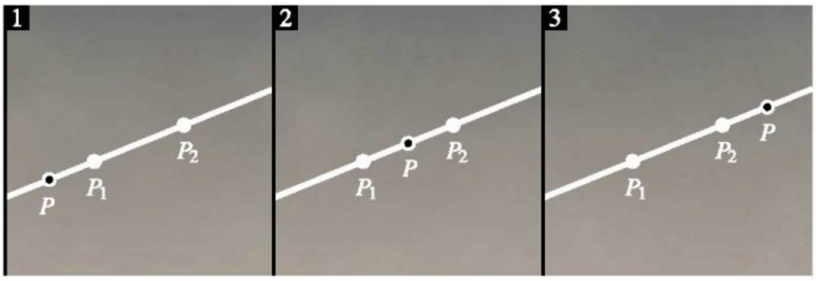

Figure 5. Finding the equation for an arbitrary point P on a line through P_1 and P_2. There are three cases depending upon the order of the points P, P_1, and P_2 on that line.

$\Longrightarrow$ First, let us show that if $P = (x_1 + t(x_2 - x_1), y_1 + t(y_2 - y_1))$ for some value of t between 0 and 1, then P is between P_1 and P_2. We can directly calculate $|P_1P|$ and $|PP_2|$:

$$\begin{aligned}
|P_1P| &= [(x - x_1)^2 + (y - y_1)^2]^{1/2} \\
&= [(x_1 + t(x_2 - x_1) - x_1)^2 + (y_1 + t(y_2 - y_1) - y_1)^2]^{1/2} \\
&= [(tx_2 - tx_1)^2 + (ty_2 - ty_1)^2]^{1/2} \\
&= t[(x_2 - x_1)^2 + (y_2 - y_1)^2]^{1/2} \\
&= t|P_1P_2|. \\
|PP_2| &= [(x_2 - x)^2 + (y_2 - y)^2]^{1/2} \\
&= [(x_2 - (x_1 + t(x_2 - x_1)))^2 + (y_2 - (y_1 + t(y_2 - y_1)))^2]^{1/2} \\
&= [((1 - t)x_2 - (1 - t)x_1)^2 + ((1 - t)y_2 - (1 - t)y_1)^2]^{1/2} \\
&= (1 - t)[(x_2 - x_1)^2 + (y_2 - y_1)^2]^{1/2} \\
&= (1 - t)|P_1P_2|.
\end{aligned}$$

According to the triangle inequality, P is between P_1 and P_2, since

$$|P_1P| + |PP_2| = t|P_1P_2| + (1 - t)|P_1P_2| = |P_1P_2|.$$

$\Longleftarrow$ Conversely, let us show that if P is between P_1 and P_2, then the coordinates of P can be written in the parametric form $(x_1 + t(x_2 - x_1), y_1 + t(y_2 - y_1))$ for some value of t between 0 and 1. P is the only point in the plane that is a distance $d_1 = |P_1P|$ from P_1 and a distance $d_2 = |PP_2|$ from P_2. Because of the uniqueness, we just need to find a point in parametric form that is also those distances from P_1 and P_2. The point that we are looking for is the one where $t = d_1/(d_1 + d_2)$. The two calculations, that the distance from this point to P_1 is d_1, and that the distance from this point to P_2 is d_2, are both straightforward, so are left as exercises.

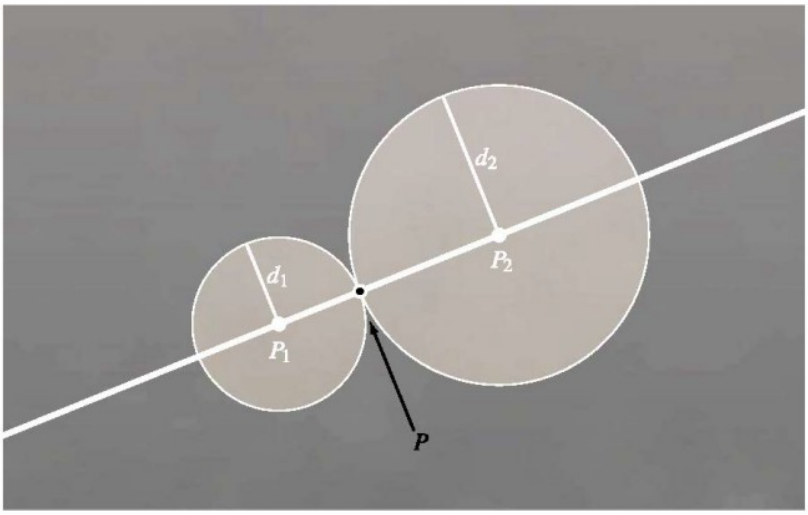

Figure 6. A point P is between P_1 and P_2 if the triangle inequality is an equality. Furthermore, P is the only point that is a distance d_1 from P_1 and a distance d_2 from P_2. □

Exercise 21.1 Complete the calculation in the last proof: using the parametrized form for $r(t)$ given there, show that $r(d_1/(d_1+d_2))$ gives the coordinates of a point that is a distance d_1 from P_1 and d_2 from P_2.

From the parametric form it is easy to get to standard form, and from there to point-slope form, slope-intercept form, and so on. The latter steps are usually done in a pre-calculus course, so I will only go one step further.

Theorem 21.5 *The coordinates (x, y) of the points of a line satisfy an equation of the form $Ax + By = C$ where A, B, and C are real numbers.*

Proof Suppose that (x_1, y_1) and (x_2, y_2) are distinct points on the line. As we saw in the last theorem, the other points on the line have coordinates (x, y) that satisfy the equations

$$\begin{aligned} x &= x_1 + t(x_2 - x_1) \\ y &= y_1 + t(y_2 - y_1). \end{aligned}$$

Now it is just a matter of combining the equations to eliminate the parameter t.

$$\begin{aligned} x - x_1 &= t(x_2 - x_1) \\ y - y_1 &= t(y_2 - y_1). \end{aligned}$$

It is tempting to divide the second equation by the first. That eliminates t and also serves as a definition of the slope of a line (in particular, it shows that the slope is constant). But we cannot divide by zero. Instead, multiply:

$$\begin{aligned} (x - x_1)(y_2 - y_1) &= t(x_2 - x_1)(y_2 - y_1) \\ (y - y_1)(x_2 - x_1) &= t(y_2 - y_1)(x_2 - x_1). \end{aligned}$$

Set the equations equal and simplify:

$$\begin{aligned} (x - x_1)(y_2 - y_1) &= (y - y_1)(x_2 - x_1) \\ x(y_2 - y_1) - x_1(y_2 - y_1) &= y(x_2 - x_1) - y_1(x_2 - x_1) \\ x(y_2 - y_1) - y(x_2 - x_1) &= x_1(y_2 - y_1) - y_1(x_2 - x_1). \end{aligned}$$

This equation has the proper form, with

$$A = (y_2 - y_1), \quad B = -(x_2 - x_1), \quad \text{and } C = x_1(y_2 - y_1) - y_1(x_2 - x_1). \qquad \square$$

Exercise 21.2 Suppose that A, B, and C are real numbers. Show that if A and B are not both zero, then the equation $Ax + By = C$ describes a line.

21.2 The unit circle approach to trigonometry

At the end of chapter 13 on similarity, we defined the six trigonometric functions in terms of the angles of a right triangle, which means that they were restricted to values in the interval $(0, \pi/2)$. There is also a unit circle approach that extends the definitions beyond that narrow window. This is standard fare, so what follows is a brief recap. A point with two positive coordinates (x, y) on the unit circle corresponds to a right triangle whose vertices are $(0, 0)$, $(x, 0)$, and (x, y). If θ is the measure of the angle at the origin, then $\cos\theta = x$ and $\sin\theta = y$ (because the hypotenuse has length one). Now just continue that: any ray from the origin forms an angle θ measured in the counterclockwise direction from the x-axis. The ray intersects the unit circle at a point (x, y) and we define

$$\cos(\theta) = x, \quad \sin(\theta) = y.$$

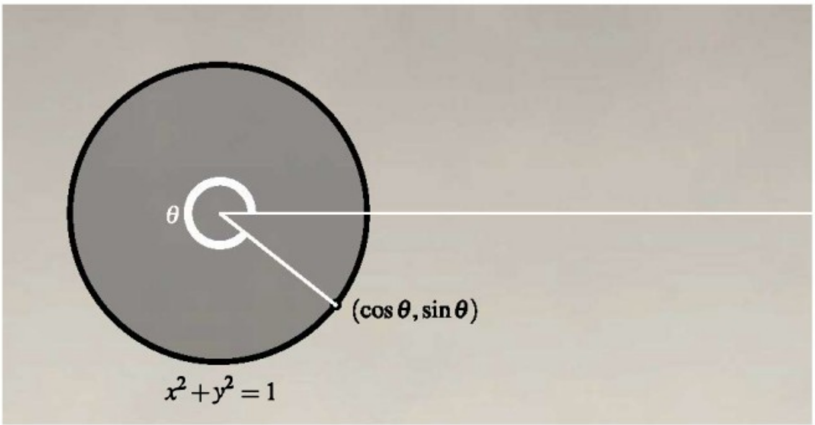

Figure 7. Coordinates on the unit circle correspond to the cosine and sine functions.

Allowing for proper and reflex angles, that extends the domains of sine and cosine to $[0, 2\pi)$, but we can go farther. We can allow the ray to spin around the circle more than once (for θ values greater than 2π) or in the clockwise direction (for negative θ). By imposing periodicity:

$$\cos(\theta + 2n\pi) = \cos(\theta), \quad \sin(\theta + 2n\pi) = \sin(\theta), \quad n \text{ an integer.}$$

The other four trigonometric functions (tangent, cotangent, secant, cosecant) are defined similarly as the ratios

$$\tan(\theta) = y/x, \quad \cot(\theta) = x/y, \quad \sec(\theta) = 1/x, \quad \csc(\theta) = 1/y.$$

There are many relationships among the trigonometric functions, some obvious and some subtle. Let's get the obvious ones out of the way. From the definitions of the functions, we get the reciprocal identities

$$\sec\theta = \frac{1}{\cos\theta}, \quad \csc\theta = \frac{1}{\sin\theta}, \quad \cot\theta = \frac{1}{\tan\theta},$$

and identities that relate tangent and cotangent to sine and cosine:

$$\tan\theta = \frac{\sin\theta}{\cos\theta}, \quad \cot\theta = \frac{\cos\theta}{\sin\theta}.$$

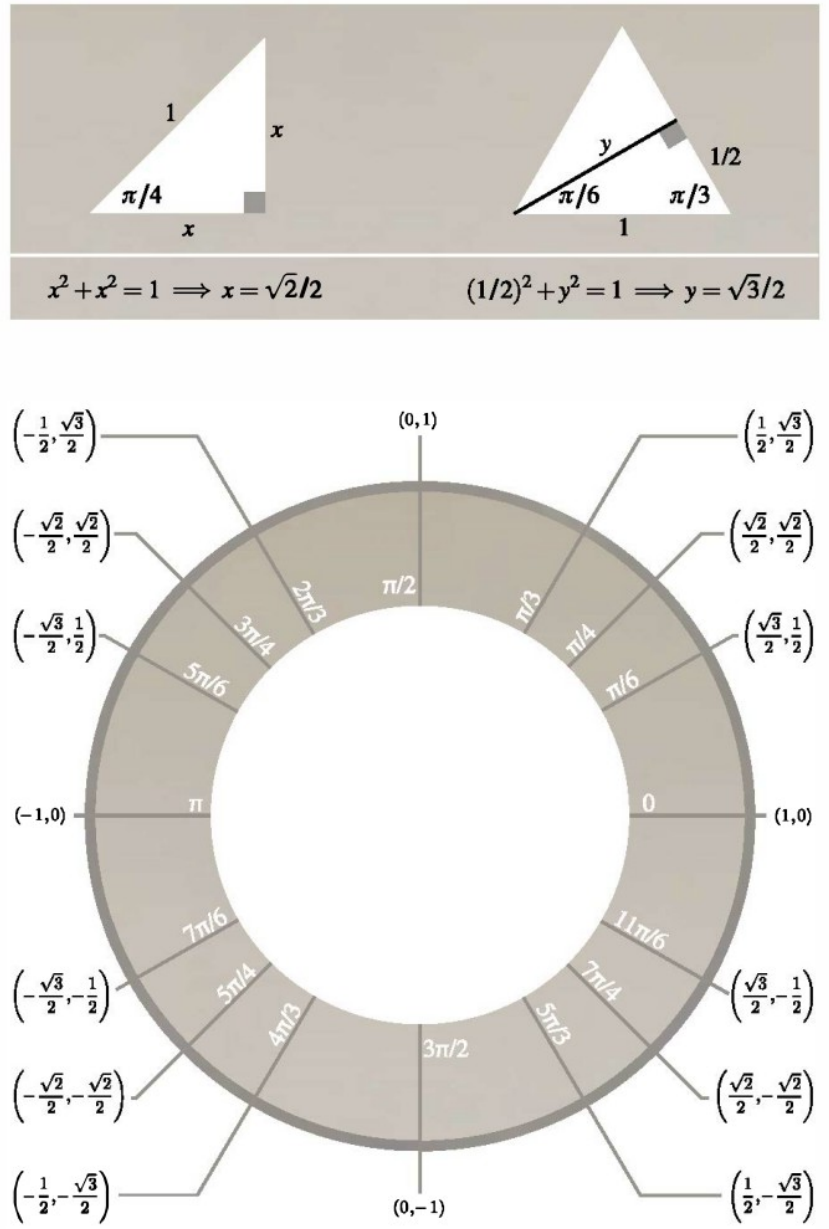
1
x
$\pi/4$
x
y
1/2
$\pi/6$
$\pi/3$
1
$x^2+x^2=1 \implies x=\sqrt{2}/2$
$(1/2)^2+y^2=1 \implies y=\sqrt{3}/2$
$(0,1)$
$\left(-\frac{1}{2},\frac{\sqrt{3}}{2}\right)$
$\left(\frac{1}{2},\frac{\sqrt{3}}{2}\right)$
$\left(-\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}\right)$
$\left(\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}\right)$
$\left(-\frac{\sqrt{3}}{2},\frac{1}{2}\right)$
$\left(\frac{\sqrt{3}}{2},\frac{1}{2}\right)$
$\pi/2$
$2\pi/3$
$\pi/3$
$3\pi/4$
$\pi/4$
$5\pi/6$
$\pi/6$
$(-1,0)$
π
0
$(1,0)$
$7\pi/6$
$11\pi/6$
$5\pi/4$
$7\pi/4$
$4\pi/3$
$3\pi/2$
$5\pi/3$
$\left(-\frac{\sqrt{3}}{2},-\frac{1}{2}\right)$
$\left(\frac{\sqrt{3}}{2},-\frac{1}{2}\right)$
$\left(-\frac{\sqrt{2}}{2},-\frac{\sqrt{2}}{2}\right)$
$\left(\frac{\sqrt{2}}{2},-\frac{\sqrt{2}}{2}\right)$
$\left(-\frac{1}{2},-\frac{\sqrt{3}}{2}\right)$
$(0,-1)$
$\left(\frac{1}{2},-\frac{\sqrt{3}}{2}\right)$

Figure 8. We can use an equilateral triangle and an isosceles right triangle to find the sine and cosine values for the angles $\pi/6$, $\pi/3$, and $\pi/4$. Then use the symmetry of the circle to extend beyond the first quadrant.

From the equation of the circle $x^2 + y^2 = 1$, we get the Pythagorean identities

$$\sin^2\theta + \cos^2\theta = 1, \quad \tan^2\theta + 1 = \sec^2\theta, \quad 1 + \cot^2\theta = \csc^2\theta.$$

By comparing angles taken in the counterclockwise and clockwise directions, we see that cosine and secant are even functions (where $f(-x) = f(x)$) and that the other four are odd functions (where $f(-x) = -f(x)$).

Beyond these, there is a second tier of identities—double angle, half angle, power reduction, etc.—that are not so immediately clear. They can be derived from two big identities, the addition formulas for sine and cosine, but the proofs of those formulas require a more careful look at the geometry of the unit circle. To close this chapter, we will prove them.

Theorem 21.6 *(Addition rule for cosine) For angles measures α and β,*

$$\cos(\alpha + \beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta.$$

The key to the proof is to compare two distances that we know to be the same, one distance expressed in terms of the angle $\alpha + \beta$, the other in terms of the individual angles α and β. The trick is to make the right choice of distances.

Proof On the unit circle, label:

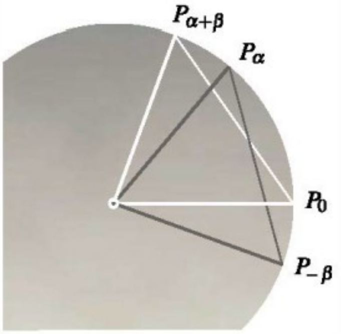

Figure 9. Marking points to derive the cosine addition formula.

$$\begin{aligned} P_0 &= (1, 0) \\ P_\alpha &= (\cos\alpha, \sin\alpha) \\ P_{-\beta} &= (\cos(-\beta), \sin(-\beta)) = (\cos\beta, -\sin\beta) \\ P_{\alpha+\beta} &= (\cos(\alpha+\beta), \sin(\alpha+\beta)). \end{aligned}$$

If O is the origin, then $\triangle OP_0P_{\alpha+\beta}$ and $\triangle OP_{-\beta}P_\alpha$ are congruent (SAS: in each triangle, two of the sides are radii, and the angle between them measures $\alpha + \beta$). So the two segments $P_0P_{\alpha+\beta}$ and $P_{-\beta}P_\alpha$ are congruent, and we can compare their lengths (it is actually easier to work with the squares of those lengths). Throughout these calculations, we make repeated use of the

Pythagorean identity $\sin^2 x + \cos^2 x = 1$.

$$\begin{aligned}
|P_0P_{\alpha+\beta}|^2 &= (\cos(\alpha+\beta)-1)^2 + (\sin(\alpha+\beta)-0)^2\\
&= \cos(\alpha+\beta)^2 - 2\cos(\alpha+\beta) + 1 + \sin^2(\alpha+\beta)\\
&= 2 - 2\cos(\alpha+\beta).\\
|P_{-\beta}P_\alpha|^2 &= (\cos\alpha - \cos\beta)^2 + (\sin\alpha + \sin\beta)^2\\
&= \cos^2\alpha - 2\cos\alpha\cos\beta + \cos^2\beta + \sin^2\alpha + 2\sin\alpha\sin\beta + \sin^2\beta\\
&= 2 - 2\cos\alpha\cos\beta + 2\sin\alpha\sin\beta.
\end{aligned}$$

Set the expressions equal and simplify:

$$\begin{aligned}
2 - 2\cos(\alpha+\beta) &= 2 - 2\cos\alpha\cos\beta + 2\sin\alpha\sin\beta\\
-2\cos(\alpha+\beta) &= -2\cos\alpha\cos\beta + 2\sin\alpha\sin\beta\\
\cos(\alpha+\beta) &= \cos\alpha\cos\beta - \sin\alpha\sin\beta.
\end{aligned}$$ ☐

Now let's look at the addition rule for the sine function. For this proof, one approach would be to use the cofunction identity $\sin(x) = \cos(\pi/2 - x)$ followed by the addition rule for cosine that we just derived (although we have not yet derived this cofunction identity). Instead, let's contrive something in the spirit of the last proof by comparing some distances and then doing a little algebra.

Theorem 21.7 *(Addition rule for sine)*

$$\sin(\alpha+\beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta.$$

Proof On the unit circle, label:

$$\begin{aligned}
P_0 &= (1,0)\\
P_\alpha &= (\cos\alpha, \sin\alpha)\\
P_\beta &= (\cos\beta, \sin\beta)\\
P_{\alpha+\beta} &= (\cos(\alpha+\beta), \sin(\alpha+\beta)).
\end{aligned}$$

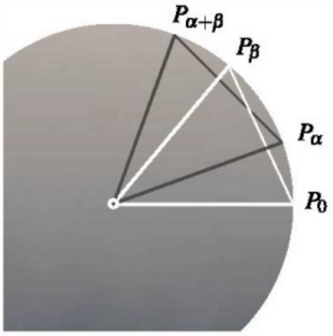

Figure 10. Marking points to derive the sine addition formula.

By SAS, $P_\alpha P_{\alpha+\beta}$ and $P_0 P_\beta$ are congruent. Let's compare the lengths (we use the addition rule for cosine midway through the first distance calculation).

$$\begin{aligned}|P_\alpha P_{\alpha+\beta}|^2 &= (\cos(\alpha+\beta)-\cos(\alpha))^2 + (\sin(\alpha+\beta)-\sin(\alpha))^2\\ &= \cos^2(\alpha+\beta) - 2\cos\alpha\cos(\alpha+\beta) + \cos^2\alpha\\ &\quad + \sin^2(\alpha+\beta) - 2\sin\alpha\sin(\alpha+\beta) + \sin^2\alpha\\ &= 2 - 2\cos\alpha\cos(\alpha+\beta) - 2\sin\alpha\sin(\alpha+\beta)\\ &= 2 - 2\cos\alpha(\cos\alpha\cos\beta - \sin\alpha\sin\beta) - 2\sin\alpha\sin(\alpha+\beta)\\ &= 2 - 2\cos^2\alpha\cos\beta + 2\sin\alpha\cos\alpha\sin\beta - 2\sin\alpha\sin(\alpha+\beta)\end{aligned}$$

and

$$\begin{aligned}|P_0 P_\beta|^2 &= (\cos\beta - 1)^2 + (\sin\beta - 0)^2\\ &= \cos^2\beta - 2\cos\beta + 1 + \sin^2\beta\\ &= 2 - 2\cos\beta.\end{aligned}$$

Set the expressions equal and simplify:

$$\begin{aligned}2 - 2\cos^2\alpha\cos\beta + 2\sin\alpha\cos\alpha\sin\beta - 2\sin\alpha\sin(\alpha+\beta) &= 2 - 2\cos\beta\\ -2\cos^2\alpha\cos\beta + 2\sin\alpha\cos\alpha\sin\beta - 2\sin\alpha\sin(\alpha+\beta) &= -2\cos\beta\\ \cos^2\alpha\cos\beta - \sin\alpha\cos\alpha\sin\beta + \sin\alpha\sin(\alpha+\beta) &= \cos\beta.\end{aligned}$$

Solve for $\sin(\alpha+\beta)$:

$$\begin{aligned}\sin\alpha\sin(\alpha+\beta) &= \cos\beta - \cos^2\alpha\cos\beta + \sin\alpha\cos\alpha\sin\beta\\ &= \cos\beta(1-\cos^2\alpha) + \sin\alpha\cos\alpha\sin\beta\\ &= \cos\beta\sin^2\alpha + \sin\alpha\cos\alpha\sin\beta\\ &= \sin\alpha(\sin\alpha\cos\beta + \cos\alpha\sin\beta).\end{aligned}$$

As long as $\sin\alpha$ is not zero, we can divide both sides by it, and what's left is what we want. That leaves one loose thread: what if $\sin\alpha$ is zero? That is left as an exercise. ☐

Exercise 21.3 Suppose that $\sin(\alpha) = 0$, and hence that α is an integer multiple of π. Prove that the addition formula for sine holds in this case as well.

21.3 Exercises

21.4. (The midpoint formula) Let $P = (a, b)$ and $Q = (c, d)$. Verify that the coordinates of the midpoint of PQ are

$$\left(\frac{a+c}{2}, \frac{b+d}{2}\right).$$

06.05.2020

E3) Calculus deconstructed_a second course in first-year...pdf

Pág. 135(152) 4 Differentiation

Pág. 137(153) Figure 4.1 $Ax+By=C$

Pág. 247(264) 5 Integration.

E4) Geometry Illuminated - An Illustrated Introduction...pdf

Obs. Livro todo!!!

Pág. 105(122) Part II Euclidean Geometry

Pág. 223(240) Part III Euclidean Transformations

→ muito bom!!!!

Obs: Aplicar 'quando' (no momento) da projeção da superfície curva sobre um plano.
(simular vice-versa também)!

Pág. 395(412) Part IV Hyperbolic Geometry.

E5) Calculus in 3D, Geometry, vectors, and...pdf

Obs. Livro todo!!!

Pág. 115(127) Chapter 2
Curves and Vector-Valued Functions of One Variable.

Part IV

Hyperbolic Geometry

The last part of this book is an introduction to hyperbolic geometry, the non-Euclidean neutral geometry. There are three major components.

- In chapters 36 and 37 we will try to get some sense of what to expect of hyperbolic geometry by looking at what would happen if Playfair's axiom were false.
- In chapters 38–41 we will build several models for hyperbolic geometry, starting with a partial model called the Beltrami pseudosphere, then moving to the upper half plane model, before ultimately settling on the Poincaré disk model. The calculations to get to the Beltrami pseudosphere model are difficult. They require multivariable calculus, linear algebra, and even some differential equations. You may want to skip straight ahead to the Poincaré disk model in chapter 41. In that chapter we verify that the Poincaré model satisfies all the axioms of neutral geometry, and so you can see how the Poincaré disk model works. I have included the chapters on the pseudosphere and the upper half plane to provide a better idea of why the Poincaré disk model works.
- The remaining chapters look at some elementary hyperbolic geometry in the Poincaré disk, deriving, for instance, hyperbolic analogs of the Pythagorean theorem and the circumference formula.

36

The Search for a Rectangle

The geometers who sought to prove Euclid's parallel axiom focused a lot of attention on rectangles. They are ubiquitous in Euclidean geometry, but without Euclid's parallel axiom, the situation is different. Let us see what happens when you let a rectangle into your world.

36.1 If there were a rectangle . . .

By definition, a rectangle is a quadrilateral with four right angles. When we talk of a rectangle, we often talk of its *base* and *height*. There is ambiguity about which side to call the base and which to call the height, but the conversation would not make any sense if it weren't for this:

Lemma 36.1 *The opposite sides of a rectangle are congruent.*

Exercise 36.1 Prove the lemma (assuming that rectangles exist). Hint: the result is easy to prove if you use the converse of the alternate interior angle theorem, and the existence of a single rectangle implies the converse of the theorem. However, we have not yet established that. Consider the following approach: let $\square ABCD$ be a rectangle and suppose that $|BC| > |AD|$. Then there is a point E between B and C so that $AD \simeq BE$. Look for a contradiction involving $\triangle DCE$.

Now let's get to work.

Theorem 36.2 *Suppose that a rectangle R exists, with base b and height h. Then for an integer $n \geq 1$, there is a rectangle with base nb and height h.*

Proof By induction. The base case, when $n = 1$, is trivial because R meets the requirements. Now for the inductive step. Suppose that a rectangle with base nb and height h exists. Label it $\square ABCD$ where AB is the base. We need to build a rectangle with base $(n+1)b$ and height h. Two of its vertices are going to be A and D. The other two are E and F where

- $A * B * E$ and $|BE| = b$, and
- $D * C * F$ and $|CF| = b$.

-7 & 6 F !X 6 C / +

By SASAS, $\square BEFC$ is congruent to R, and therefore $\angle E$ and $\angle F$ are both right angles. That is, the interior angles of $\square AEFD$ are right angles so $\square AEFD$ is a rectangle. Its height is $|AD| = h$. Its base is

$$|AE| = |AB| + |BE| = nb + b = (n+1)b.$$

By induction, then, we can build nb-by-h rectangles.

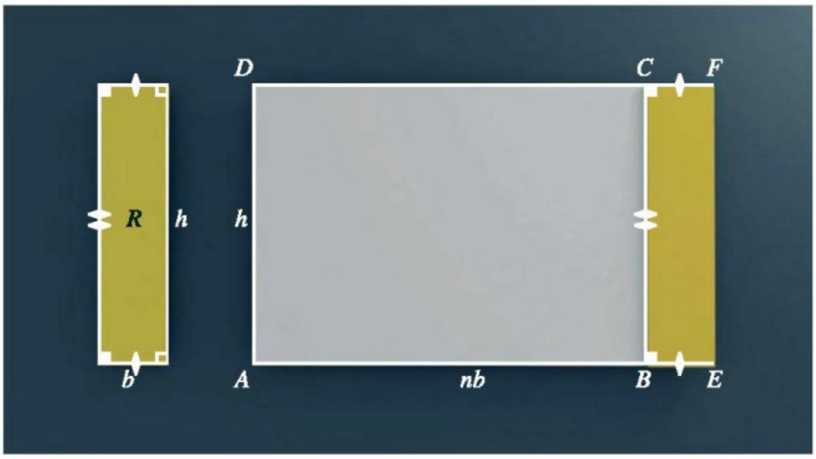

Figure 1. If there is a rectangle with base b and height h, then there is a rectangle with base nb and height h (a proof by induction).

$\square$

Theorem 36.3 *Suppose that a rectangle R exists, with base b and height h. Then for a positive real number x, there is a rectangle with base x and height h.*

Proof There is an integer n so that $nb > x$. By the previous result, there is a rectangle $\square ABCD$ with base $|AB| = nb$ and height $|AD| = h$. Then there are points E and F so that

- $A * E * B$ and $|AE| = x$, and
- $D * F * C$ and $|DF| = x$.

The segment EF cuts $\square ABCD$ into two pieces. We can show that both of those pieces are also rectangles. To see this, start by looking at their angle sums. As a corollary to the Saccheri-Legendre theorem, the angle sum of a quadrilateral is at most 2π. Therefore

$$(\angle A) + (\angle D) + (\angle AEF) + (\angle DFE) \le 2\pi$$
$$(\angle B) + (\angle C) + (\angle BEF) + (\angle CFE) \le 2\pi.$$

Since $(\angle A) = (\angle B) = (\angle C) = (\angle D) = \pi/2$, we get

$$(\angle AEF) + (\angle DFE) \le \pi$$
$$(\angle BEF) + (\angle CFE) \le \pi.$$

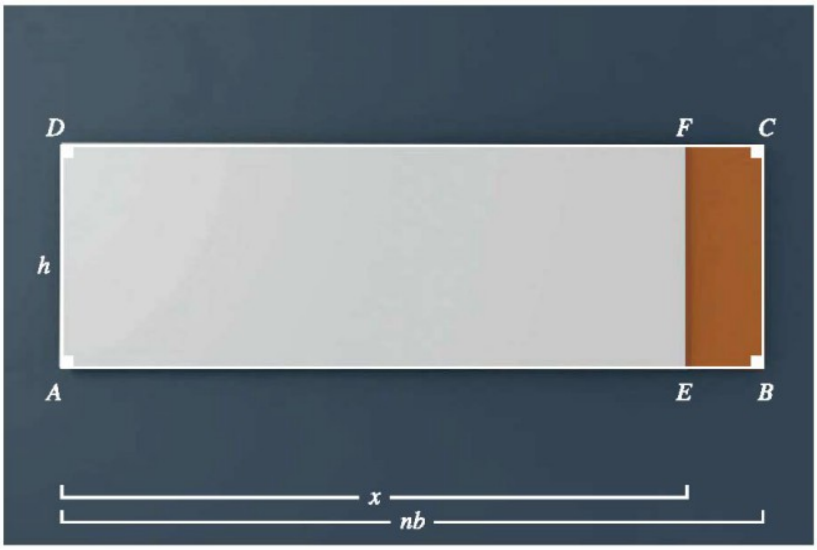

Figure 2. If there is a rectangle with base b and height h, then there is a rectangle with base x and height h. Extend the given rectangle to a new rectangle with base $nb > x$. Then divide that into two pieces, one with a base of x.

But

$$EA \simeq FD, \quad \angle A \simeq \angle D, \quad AD = DA, \quad \angle D \simeq \angle A, \text{ and } DF \simeq AE.$$

By SASAS, $\square AEFD$ is congruent to itself in a non-trivial way, so matching corresponding pieces, $\angle DFE \simeq \angle AEF$. Their supplements, $\angle CFE$ and $\angle BEF$, are also congruent.

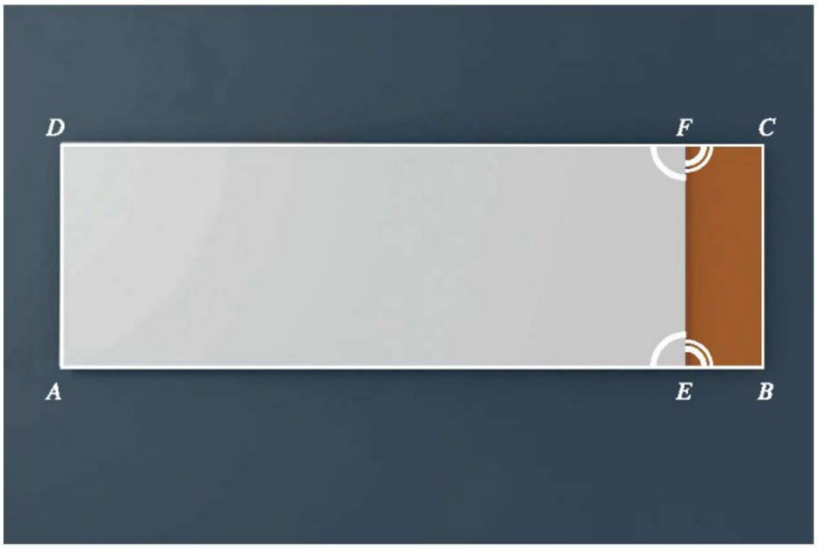

Figure 3. The segment EF cuts the rectangle $\square ABCD$ into two smaller pieces, both of which must also be rectangles.

Substitute this information into the pair of inequalities to get

$$2(\angle AEF) \leq \pi \implies (\angle AEF) \leq \pi/2$$
$$2(\angle BEF) \leq \pi \implies (\angle BEF) \leq \pi/2.$$

But $\angle AEF$ and $\angle BEF$ are supplementary, and so together sum to π. The only way this can happen is if $(\angle AEF) = (\angle BEF) = \pi/2$. That is, they are right angles, as are $\angle DFE$ and $\angle CFE$. Hence $\square AEFD$ and $\square BEFC$ are rectangles. In particular, $\square AEFD$ is a rectangle with base x and height h. □

Theorem 36.4 *Suppose that a rectangle R exists, with base b and height h. Then for positive real numbers x and y, there is a rectangle with base x and height y.*

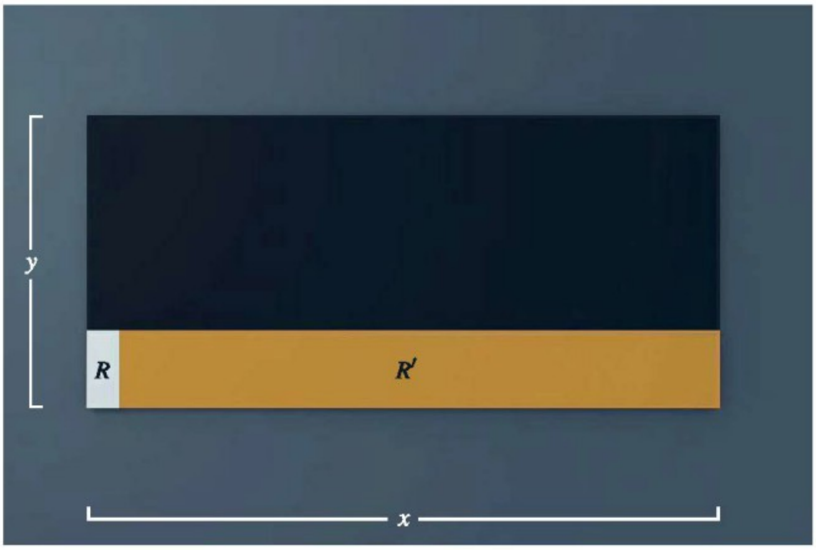

Figure 4. Use the previous result in both directions to build a rectangle with base x and height y.

Proof By the previous result, there is a rectangle R' with base x and height h. The roles of base and height are interchangeable, so that we can think of R' as a rectangle with base h and height x. Using the previous result, there is a rectangle R'' with base y and height x. Switch the roles of base and height in R'' back, so that R'' has a base x and a height y. □

The last three results tell us that if there is a rectangle in our geometry, then they are abundant. What are the implications of all of these rectangles?

Theorem 36.5 *If rectangles of all sizes exist, then the angle sum of a triangle is π.*

Proof There are two cases: first when the triangle is a right triangle, and second when it is not.

Case 1: Suppose that ΔABC is a right triangle with right angle at C. There is a rectangle with base $|AC|$ and height $|BC|$. Label it $\square acbd$ so that ac is the base and cb is the height. The diagonal ab cuts the rectangle into two right triangles. By SAS, both are congruent to ΔABC. Therefore the angle sum of the two copies of ΔABC equals the angle sum of a rectangle:

$$2((\angle A) + (\angle B) + (\angle C)) = 2\pi \implies (\angle A) + (\angle B) + (\angle C) = \pi.$$

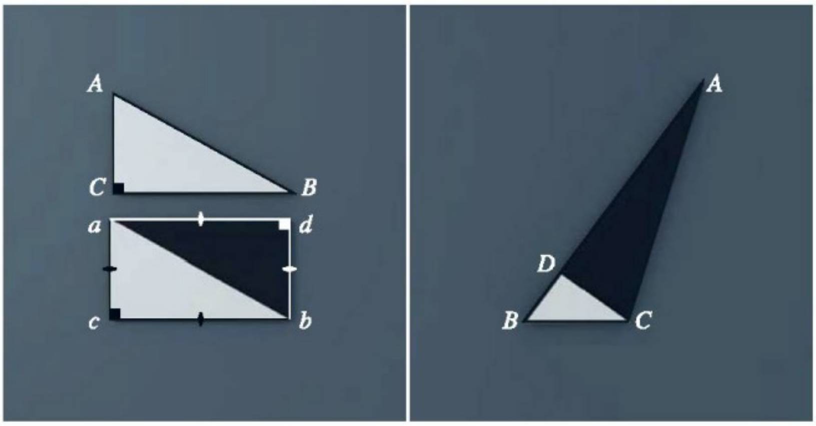

Figure 5. Given rectangles, the sum of the interior angles of the triangles must be π: (l) case 1, for right triangles. (r) case 2, for non-right triangles.

Case 2: Now suppose that ΔABC is not a right triangle, and label it so that $\angle C$ is the largest angle. In that case, the altitude from C intersects AB between A and B. Label the intersection point as D. Then CD divides ΔABC into two smaller right triangles: ΔACD and ΔBCD. From the previous part, both have an angle sum of π. They have right angles at D, so the other two angles sum to $\pi/2$. So

$$(\angle A) + (\angle ACD) = \pi/2 \quad \text{and} \quad (\angle BCD) + (\angle B) = \pi/2.$$

Add the equations to get

$$(\angle A) + (\angle ACD) + (\angle BCD) + (\angle B) = \pi/2 + \pi/2$$
$$(\angle A) + (\angle C) + (\angle B) = \pi.$$

□

Theorem 36.6 *If the angle sum of a triangle is π, then for a line ℓ and a point P not on ℓ, there is a unique line through P parallel to ℓ. That is, Playfair's axiom is true (and does not need to be taken as an axiom).*

Proof Begin by dropping the perpendicular from P to ℓ, and label its foot as Q_0. Let ℓ' be the line through P that is perpendicular to PQ_0. By the alternate interior angle theorem, it is parallel to ℓ. We need to show it is the only parallel. Suppose that ℓ'' is some other line through P. It forms an acute angle with PQ_0 on one of the sides of PQ_0 (and not on the other). On the acute side of PQ_0 we are going to build a sequence of isosceles triangles $\{\Delta PQ_{i-1}Q_i\}$.

(1) Locate Q_1 on ℓ so that $PQ_0 \simeq Q_0Q_1$. Then ΔPQ_0Q_1 is isosceles. The angle at Q_0 is right, so if the angle sum of a triangle is π, then

$$\pi/2 + (\angle P) + (\angle Q_1) = \pi \quad \Longrightarrow \quad (\angle P) + (\angle Q_1) = \pi/2.$$

By the isosceles triangle theorem, $\angle P$ and $\angle Q_1$ are congruent, so they both measure $\pi/4$.

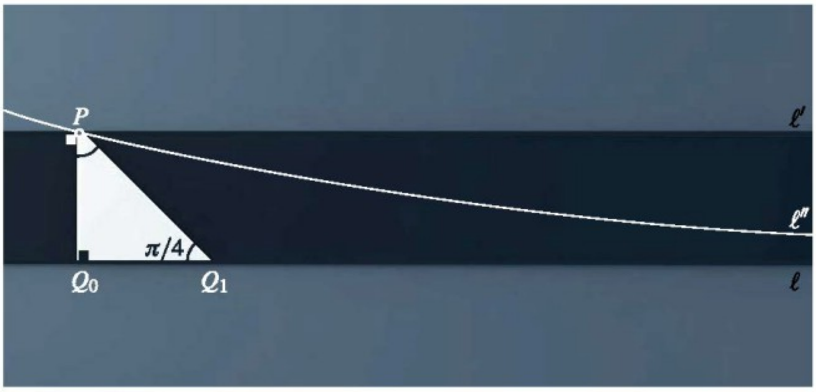

Figure 6. Constructing longer and longer isosceles triangles between ℓ and ℓ'. The first step.

(2) Now move farther out on ℓ, and locate Q_2 so that $Q_1Q_2 \simeq PQ_1$. The angle at Q_1 is $\pi - \pi/4 = 3\pi/4$, so if the angle sum of a triangle is π, then in ΔPQ_1Q_2,

$$3\pi/4 + (\angle P) + (\angle Q_2) = \pi \quad \Longrightarrow \quad (\angle P) + (\angle Q_2) = \pi/4.$$

Because $\angle P$ and $\angle Q_2$ are congruent in this triangle, they each measure $\pi/8$. Therefore

$$(\angle Q_0PQ_2) = (\angle Q_0PQ_1) + (\angle Q_1PQ_2) = \pi/4 + \pi/8 = 3\pi/8.$$

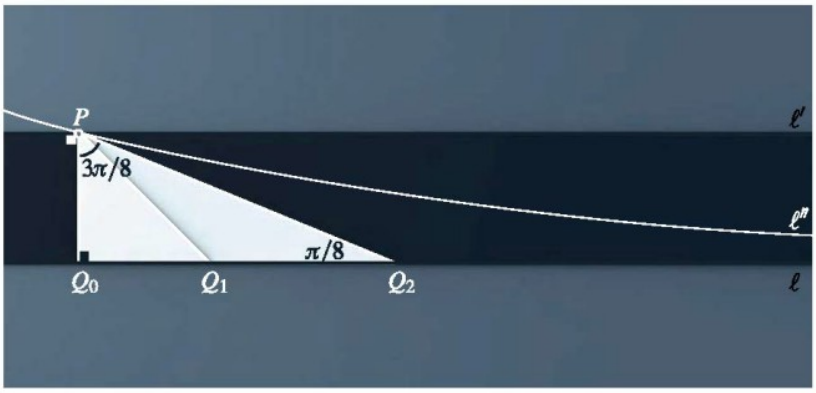

Figure 7. The second step.

(3) We can continue in this way, building isosceles triangles $\triangle PQ_iQ_{i+1}$ with congruent sides PQ_i and Q_iQ_{i+1}. Measuring angles as we go:

i	$(\angle PQ_{i-1}Q_i)$	$(\angle Q_0PQ_i)$
1	$\pi/4$	$\pi/4$
2	$\pi/8$	$3\pi/8$
3	$\pi/16$	$7\pi/16$
4	$\pi/32$	$15\pi/32$
5	$\pi/64$	$31\pi/64$

In general,

$$(\angle Q_0PQ_i) = \frac{2^i - 1}{2^{i+1}}\pi.$$

As i approaches infinity, the angle approaches $\pi/2$. Now ℓ'' forms an acute angle with PQ_0. Therefore, for a large enough i, ℓ'' will lie between $\overrightarrow{PQ_0}$ and $\overrightarrow{PQ_i}$. By the crossbar theorem, ℓ'' will then intersect Q_0Q_i, and so ℓ'' will not be parallel to ℓ. That is, a line through P other than ℓ' will intersect ℓ. This means ℓ' is the unique parallel to ℓ through P. □

36.2 The search for a rectangle

Euclidean geometry is neutral geometry with one more axiom, Playfair's axiom, added. Before the discovery of non-Euclidean geometry, geometers were trying to prove that neutral geometry *is* Euclidean geometry by trying to prove Playfair's axiom from the axioms of neutral geometry. One way to do that would be to prove that there are rectangles in neutral geometry. Hence, much effort was spent in attempts to build a rectangle using only the axioms of neutral geometry. There isn't a construction that is going to just give you all four right angles right away, so the idea was to construct a quadrilateral that has some of the properties of a rectangle, and then to work from there. One such quasi-rectangle is called a Saccheri quadrilateral.

Definition 36.7 A convex quadrilateral $\square ABCD$ is called a *Saccheri quadrilateral* with *base* AB, *summit* CD, and *legs* BC and AD if:

- $\angle A$ and $\angle B$ are right angles, and
- segments BC and AD are congruent.

The angles $\angle C$ and $\angle D$ are called the *summit angles* of the Saccheri quadrilateral.

Saccheri quadrilaterals can be constructed in neutral geometry: start with a segment for the base, draw two rays from the endpoints of the base to form the two right angles, and then mark off congruent segments along them. The segments are the legs. Join their endpoints to form the summit and you have a Saccheri quadrilateral. A Saccheri quadrilateral seems to be about half the way to a rectangle because two of its angles are right, and one pair of opposite sides are congruent. The only Saccheri quadrilaterals in Euclidean geometry are rectangles because by the converse of the alternate interior angle theorem, the two summit angles must be right angles. With only the axioms of neutral geometry, we can't prove that the summit angles are right angles, but we can prove weaker statements. The first statement exploits the internal symmetry of a Saccheri quadrilateral (much like the proof of the isosceles triangle theorem).

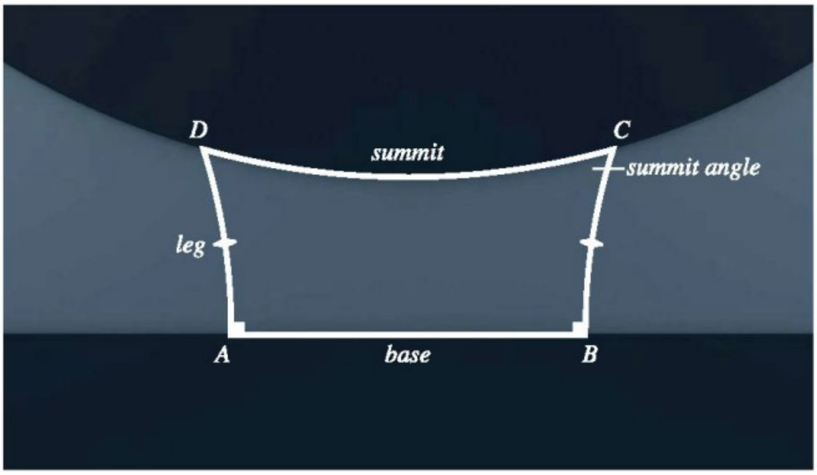

Figure 8. The parts of a Saccheri quadrilateral.

Theorem 36.8 *Let $\square ABCD$ be a Saccheri quadrilateral with base AB. Then the summit angles $\angle C$ and $\angle D$ are congruent.*

Proof Observe that

$$DA \simeq CB, \quad \angle A \simeq \angle B, \quad AB = BA, \quad \angle B \simeq \angle A, \quad \text{and} \quad BC \simeq AD.$$

By SASAS, $\square ABCD$ and $\square BADC$ are congruent. Matching the corresponding parts, $\angle C \simeq \angle D$.

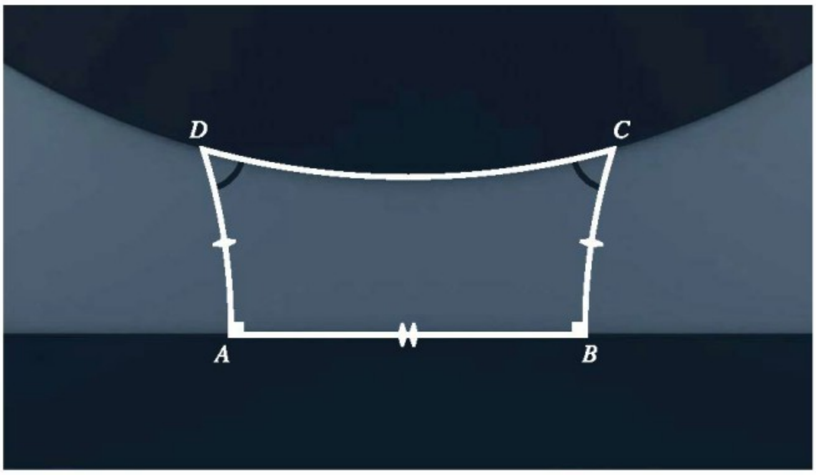

Figure 9. By SASAS, the summit angles of a Saccheri quadrilateral are congruent. □

Theorem 36.9 *Let $\square ABCD$ be a Saccheri quadrilateral with base AB. Then the summit angles $\angle C$ and $\angle D$ cannot be obtuse.*

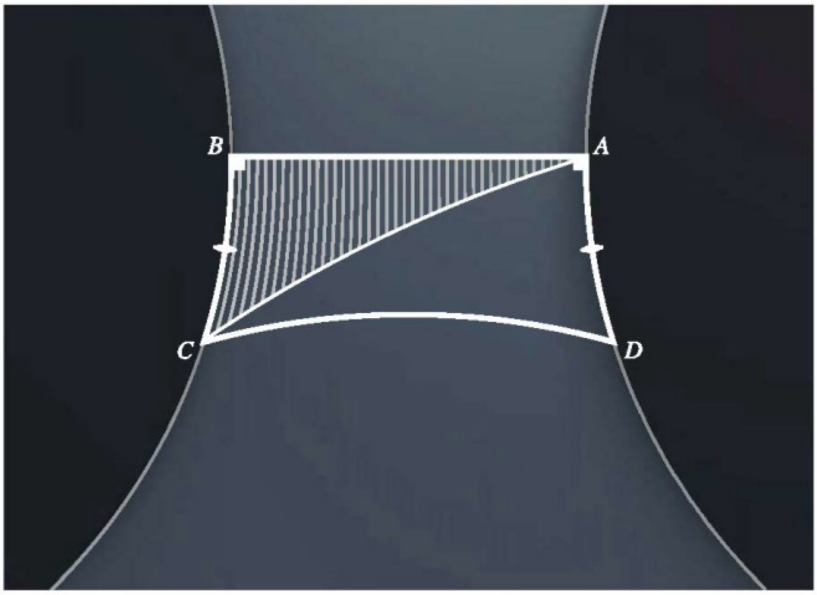

Figure 10. The summit angles cannot be obtuse.

Proof The diagonal AC divides the Saccheri quadrilateral into two triangles, and by the Saccheri-Legendre theorem the angle sum of each is at most π. Therefore the angle sum of $\square ABCD$ is at most 2π. So

$$\begin{aligned}(\angle A) + (\angle B) + (\angle C) + (\angle D) &\le 2\pi\\ (\pi/2) + (\pi/2) + (\angle C) + (\angle D) &\le 2\pi\\ (\angle C) + (\angle D) &\le \pi.\end{aligned}$$

We just proved that $\angle C$ and $\angle D$ are congruent, though, so they must measure at most $\pi/2$. □

If the summit angles are right, then the Saccheri quadrilateral is a rectangle. So in non-Euclidean geometry, the summit angles must be acute.

Definition 36.10 The *altitude* of a Saccheri quadrilateral is the line segment from the midpoint of its base to the midpoint of its summit.

Theorem 36.11 *Let $\square ABCD$ be a Saccheri quadrilateral with base AB and altitude EF (with E on the base and F on the summit). Then EF is perpendicular to the base AB and the summit CD.*

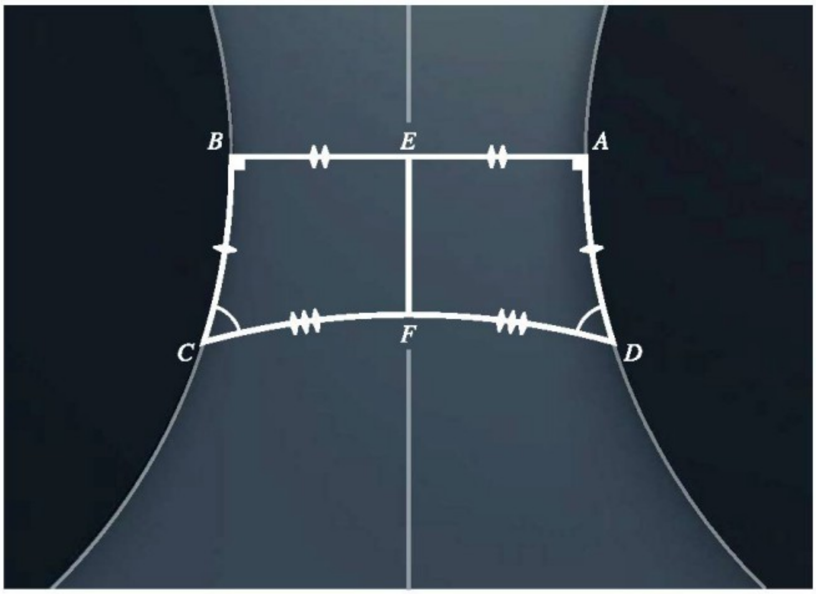

Figure 11. The segment EF is the altitude of $\square ABCD$. It is perpendicular to AB and CD.

Proof Compare the two quadrilaterals $\square AEFD$ and $\square BEFC$. Since $\square ABCD$ is Saccheri, $AD \simeq BC$ and $\angle A \simeq \angle B$, and since the summit angles of a Saccheri quadrilateral are congruent, $\angle D \simeq \angle C$. By the definition of an altitude, $AE \simeq BE$ and $DF \simeq CF$. By SASAS, then, $\square AEFD \simeq \square BEFC$. In particular, $\angle AEF$ is congruent to its own supplement, $\angle BEF$, and $\angle DFE$ is congruent to its own supplement, $\angle CFE$. They are both right angles, then, and so the altitude is perpendicular to both base and summit. □

Theorem 36.12 *Let $\square ABCD$ be a Saccheri quadrilateral with base AB and summit CD. Then $|CD| \geq |AB|$.*

Proof Assume instead that AB is greater than CD. Label E and F the midpoints of AB and CD, so that EF is the altitude. Then $|EB| > |FC|$ so we can locate a point G between E and B so that $EG \simeq FC$. Since $\square EFCG$ has right angles at E and F and a pair of congruent opposite sides, it is a Saccheri quadrilateral with base EF. As we saw above, the summit angle $\angle EGC$ cannot be obtuse, so its supplement cannot be acute. But ΔBGC has a right angle at B and an angle at G that isn't acute, so the angle sum of ΔBGC will exceed π, and this violates the Saccheri-Legendre theorem. □

Theorem 36.13 *Let $\square ABCD$ be a Saccheri quadrilateral with base AB. Let EF be the altitude, with E on AB and F on CD. Then the altitude EF cannot be longer than the legs AD and BC.*

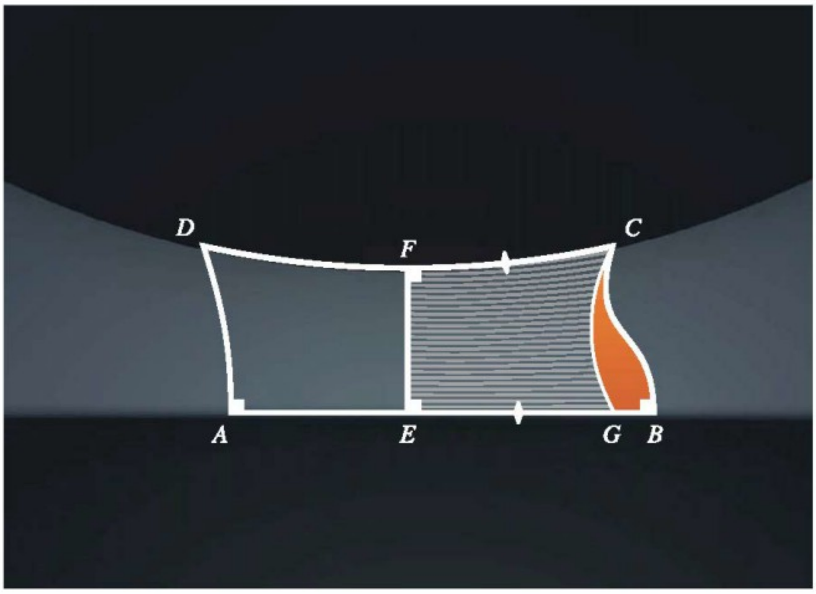

Figure 12. The base cannot be longer than the summit.

Proof Let's build a different Saccheri quadrilateral (see figure 13). Label:

F' : on $\overrightarrow{FE}$ so that $F * E * F'$ and $EF' \simeq EF$, and
C' : on $\overrightarrow{CB}$ so that $C * B * C'$ and $BC' \simeq BC$.

Now we have the congruences:

$$CB \simeq C'B, \quad BE = BE, \quad EF \simeq EF', \quad \angle CBE \simeq \angle C'BE, \text{ and } \angle BEF \simeq \angle BEF'.$$

By SASAS, $\square BEFC$ and $\square BEF'C'$ are congruent. Matching their corresponding pieces, $\angle F'$ is congruent to $\angle F$ and so is a right angle, and $F'C' \simeq FC$. Thus $\square CFF'C'$ is a Saccheri quadrilateral with base FF' and summit CC'. By the previous result, its base cannot be longer than the summit, and so

$$|EF| = \tfrac{1}{2}|FF'| \leq \tfrac{1}{2}|CC'| = |CB|.$$

We see, then, that EF, the altitude of $\square ABCD$ cannot be longer than the leg BC. □

Thinking more generally, what can we say about other perpendiculars that fall between the altitude and a leg?

Theorem 36.14 *Let $\square ABCD$ be a Saccheri quadrilateral with base AB and altitude EF. Let GH be another segment that connects the base to the summit, so that $E * G * B$, $F * H * C$, and GH is perpendicular to the base. Then*

$$|EF| \leq |GH| \leq |BC|.$$

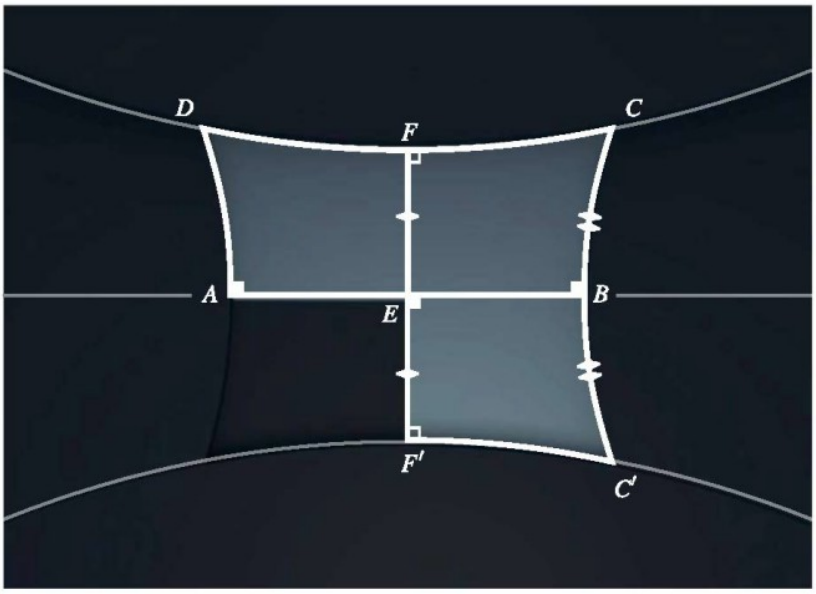

Figure 13. The altitude cannot be longer than a leg.

Proof We will prove the first inequality, and leave the second as an exercise. Suppose that $|EF| > |GH|$. Then there is a point K between E and F so that $EK \simeq GH$. Then $\square EGHK$ is a Saccheri quadrilateral with base EG. Now let's look at $\triangle KFH$. In it, $\angle F$ is right because the altitude is perpendicular to the summit; $\angle K$ is right or obtuse, since it is supplementary to a

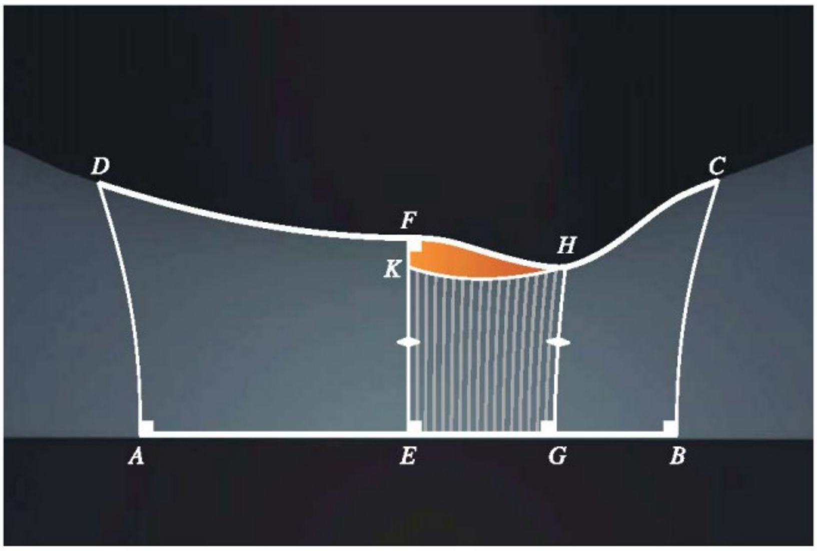

Figure 14. Given a segment GH that is between the leg BC and the altitude EF, $|GH| \geq |EF|$ (a proof by contradiction).

summit angle (remember that the summit angles cannot be obtuse). Then the sum of the angles in $\triangle KFH$ is greater than π, which contradicts the Saccheri-Legendre theorem. Therefore $|GH|$ cannot be less than $|EF|$. □

Exercise 36.2 Complete the proof of the theorem by showing that $|GH| \leq |BC|$.

If the first half of this chapter was a little jarring—it is a little crazy to think about a geometry without rectangles—I hope the second half gave some comfort. On the one hand, yes, things really are different in non-Euclidean geometry; but on the other, there is still order and structure. The next chapter is a more direct approach to understanding the order and structure: we will look at the behavior of parallel lines.

36.3 References

The general theme of this lesson, to start from a rectangle and to use it to establish Playfair's axiom, is taken from Wallace and West's *Roads to Geometry* [WW04].

36.4 Exercises

36.3. Prove that the diagonals of a Saccheri quadrilateral are congruent.

36.4. Prove that the diagonals of a Saccheri quadrilateral intersect one another at a point on its altitude.

36.5. Demonstrate that in neutral geometry it is possible to construct a convex quadrilateral whose four sides are all congruent.

36.6. Prove that if there is a triangle with an angle sum of π, then there must be a rectangle.

36.7. Let $\square ABCD$ be a convex quadrilateral so that $\angle A$ and $\angle B$ are right angles, and so that $\angle C \simeq \angle D$. Prove that $\square ABCD$ is a Saccheri quadrilateral.

36.8. Let $\square ABCD$ be a Saccheri quadrilateral with base AB and altitude EF (so that E is on the base and F is on the altitude). Prove that $|EC| \geq |FB|$.

36.9. Given $\triangle ABC$, let ℓ be the line that passes through the midpoints of AB and AC. Let F_B be the foot of the perpendicular to ℓ that passes through B, and let F_C be the foot of the perpendicular to ℓ that passes through C. Prove that $\square BCF_CF_B$ is a Saccheri quadrilateral with base F_BF_C.

36.10. Show that in non-Euclidean geometry, not every triangle has a circumcenter. The strategy is as follows. The circumcenter is the intersection of the perpendicular bisectors. Let P be a point not on the line ℓ. Let Q be the foot of the perpendicular to ℓ that passes through P. Let ℓ' be a line through Q that is parallel to ℓ but not perpendicular to PQ. Construct a triangle that has ℓ and ℓ' as perpendicular bisectors. Since the perpendicular bisectors don't even intersect, there cannot be a circumcenter.

36.11. Prove that in non-Euclidean geometry, a triangle has an incenter. (We shall see later, though, that some circles cannot be contained in any triangle. This is most easily seen in terms of area.)

37

Non-Euclidean Parallels

From here on out, let's assume that Playfair's axiom is false. That means that our geometry is not Euclidean. It means our triangle angle sums are strictly less than π. It means that Saccheri quadrilaterals are not rectangles, because rectangles do not exist, and so the summit angles are acute (not right). And other absurdities.

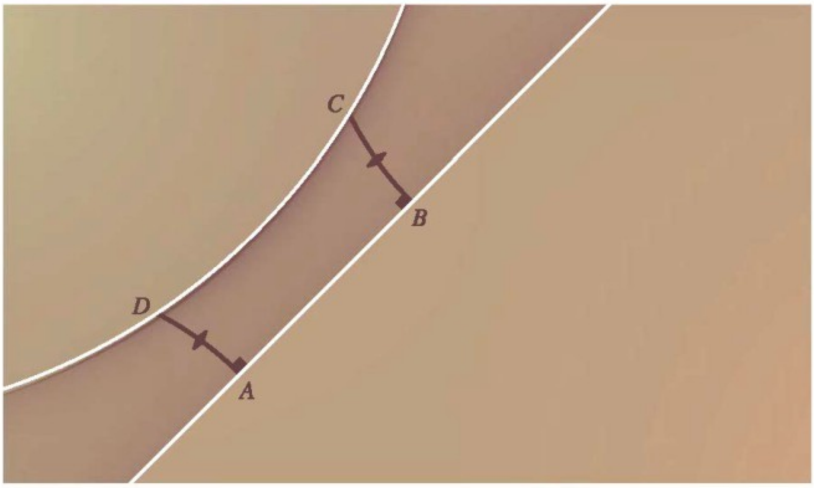

Figure 1. The base and summit of a Saccheri quadrilateral are on parallel lines.

It may have been implied, but in the last chapter we never proved that the base and summit of a Saccheri quadrilateral are parallel.

Exercise 37.1 Let $\square ABCD$ be a Saccheri quadrilateral with base AB and summit CD. Prove that $\overline{AB}$ and $\overline{CD}$ are parallel.

In this section, we will explore parallels more systematically. By the end of it, we should have a better sense of what to look for in a non-Euclidean geometry. To start, let's see that non-Euclidean behavior is pervasive, and not just an anomaly that occurs at a few isolated points.

Theorem 37.1 *Let ℓ be a line and let P be a point that is not on ℓ. Then there are infinitely many lines through P that are parallel to ℓ.*

Proof There are two pieces to this proof. First that there are always at least two parallels. Then that there are infinitely many.

Part 1, that there are (at least) two parallels. Drop the perpendicular from P to ℓ, and mark its intersection with ℓ as Q. Let ℓ' be the line through P that is perpendicular to PQ. According to the alternate interior angle theorem, ℓ' is parallel to ℓ. Now suppose that it is the only parallel. To get a contradiction, we are going to build a Saccheri quadrilateral and see that it would have to be a rectangle. Pick a point on ℓ other than Q and label it R. On the line through R that is perpendicular to ℓ, mark the point S so that $RS \simeq PQ$ (choose S on the same side of ℓ as P). Then $\square PQRS$ is a Saccheri quadrilateral with base QR and summit PS. According to exercise 37.1, $\overline{PS}$ and $\overline{QR} = \ell$ are parallel. But there is only one line through P that is parallel to ℓ, namely ℓ'. Therefore $\overline{PS} = \ell'$. In $\square PQRS$, the summit angle at P is a right angle; and the other summit angle must be congruent to it. Therefore all four angles of $\square PQRS$ are right angles and it is a rectangle. There are no rectangles in non-Euclidean geometry, so there must be more than one parallel to ℓ that passes through P.

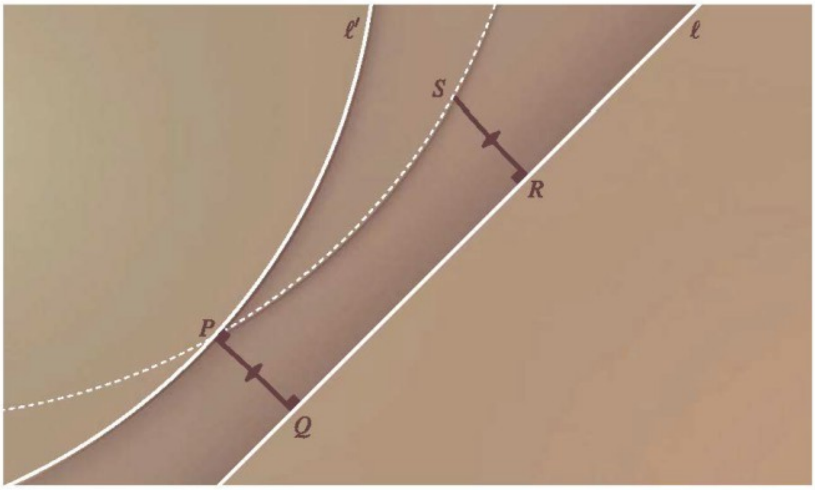

Figure 2. If there are no rectangles, then there must be at least two lines through P that are parallel to $\overline{QR}$.

Part 2, that there are infinitely many parallels. Let ℓ_1 and ℓ_2 be two lines through P that are parallel to ℓ. We can box-in P by surrounding it with more labeled points:

A, A' : on ℓ_1 so that $A * P * A'$,
B : on ℓ_2 and on the same side of PQ as A,
B' : on ℓ_2 and on the same side of PQ as A' (so $B * P * B'$).

Let C be a point on the segment AB. We will argue that $\overline{PC}$ is also parallel to ℓ. Suppose $\overline{PC}$ intersects ℓ at a point D. Then one of $\overrightarrow{PA}$, $\overrightarrow{PA'}$, $\overrightarrow{PB}$, or $\overrightarrow{PB'}$ would have to lie in the interior of

$\angle QPD$. According to the crossbar theorem, the ray would have to intersect QD. This cannot happen since ℓ_1 and ℓ_2 (which contain all four rays) are parallel to ℓ.

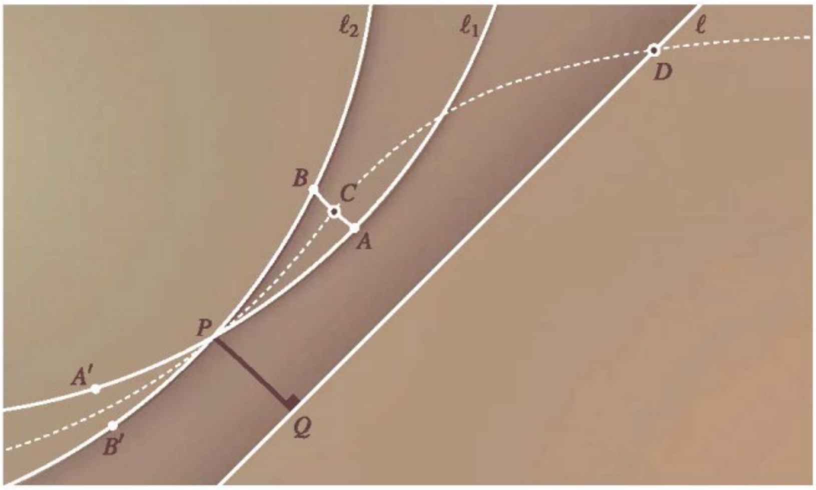

Figure 3. Given two parallels through P, we can find infinitely more between them. □

In Euclidean geometry, parallel lines are everywhere equidistant, so we can measure the distance between them with a single number. Non-Euclidean geometry is not so simple. Given parallel lines ℓ_1 and ℓ_2 in non-Euclidean geometry, the distance from a point P on ℓ_2 to the line ℓ_1 is measured as usual: drop a perpendicular from P to ℓ_1 and mark its foot (the intersection with ℓ_1) as F_P. Then the distance from P to ℓ_1 is $|PF_P|$.

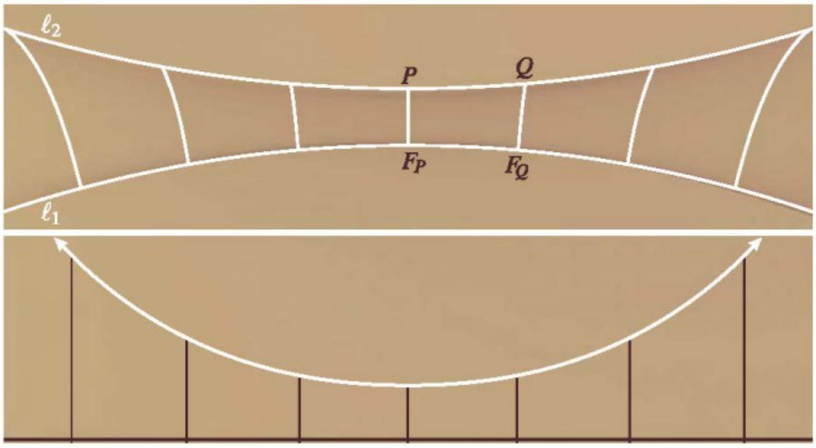

Figure 4. We can view the distance between parallel lines ℓ_1 and ℓ_2 as a function of the points on ℓ_2.

We can define a function $f : \ell_2 \dashrightarrow \mathbb{R}$ by setting $f(P)$ to be the distance from P to ℓ_1. The rest of this chapter investigates it. Before we go any further, we need the following lemma (note its similarity to the theorem in the last chapter that compared the length of the summit to the length of the base).

Lemma 37.2 *Let ℓ_1 and ℓ_2 be parallel lines, let P and Q be points on ℓ_2, and let F_P and F_Q be the respective feet of the perpendiculars on ℓ_1. Then $|F_P F_Q| < |PQ|$.*

Exercise 37.2 Prove the lemma.

Theorem 37.3 *The function f measuring the distance from points of ℓ_2 to the line ℓ_1 is continuous.*

Proof It is always possible to drop a perpendicular from a point P on ℓ_2 to the line ℓ_1, and it is always possible to measure the distance from P to its foot F_P, so the function f is always defined. For f to be continuous at the point P, we need a bit more: as a point Q on ℓ_2 approaches P, the value of $f(Q)$ must approach $f(P)$. To be more precise,

> For every $\epsilon > 0$, there must be a $\delta > 0$ so that if $|PQ| < \delta$ then $|f(P) - f(Q)| < \epsilon$.

The verification of this hinges on a few simple geometric observations that follow from the triangle inequality.

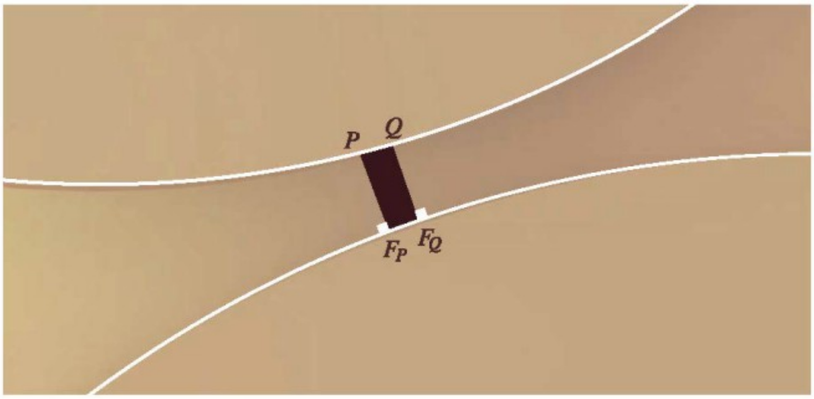

Figure 5. To show that f is continuous, consider two points P and Q that are close to one another.

- $|PF_P| \le |PQ| + |QF_Q| + |F_P F_Q|$.
- $|QF_Q| \le |PQ| + |PF_P| + |F_P F_Q| \implies |PF_P| \ge |QF_Q| - |PQ| - |F_P F_Q|$.
- Therefore

$$|QF_Q| - |PQ| - |F_P F_Q| \le |PF_P| \le |PQ| + |QF_Q| + |F_P F_Q|.$$

- Subtract $|QF_Q|$:

$$-|PQ| - |F_P F_Q| \le |PF_P| - |QF_Q| \le |PQ| + |F_P F_Q|.$$

- We don't know whether $|PF_P| > |QF_Q|$ or if it is the other way around. Either way, we can say that

$$\big||PF_P| - |QF_Q|\big| \le |PQ| + |F_PF_Q|.$$

With that information, verifying continuity is straightforward. Fix $\epsilon > 0$ and put $\delta = \epsilon/2$. If $|PQ| < \delta = \epsilon/2$, then $|F_PF_Q| < \epsilon/2$ as well (according to lemma 37.2), and so

$$|f(P) - f(Q)| = \big||PF_P| - |QF_Q|\big| \le |PQ| + |F_PF_Q| < \epsilon/2 + \epsilon/2 = \epsilon.$$

Therefore f is continuous at P. Since P was an arbitrary point on ℓ_2, f is continuous on all of ℓ_2. □

Theorem 37.4 *For a real number r, $|\{f^{-1}(r)\}| \le 2$. That is, there are at most two points on ℓ_2 that are at a distance r from ℓ_1.*

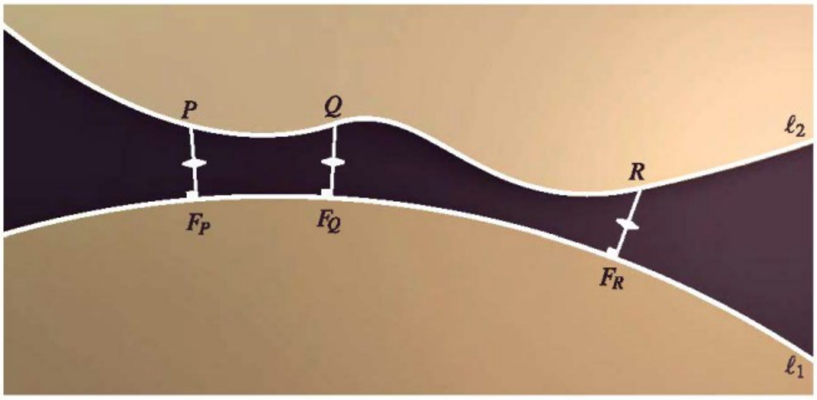

Figure 6. It is not possible for all of P, Q, and R to be the same distance from ℓ_1.

Proof Suppose that there is a real number r so that $|\{f^{-1}(r)\}| > 2$. Then there are (at least) three distinct points P, Q, and R on ℓ_2 with $P * Q * R$ so that $f(P) = f(Q) = f(R) = r$. If we label the feet of their respective perpendiculars as F_P, F_Q, and F_R, then

$$|PF_P| = |QF_Q| = |RF_R| = r.$$

Thus $\square F_PF_QQP$ and $\square F_QF_RRQ$ are Saccheri quadrilaterals. They cannot be rectangles, though, so their summit angles $\angle PQF_Q$ and $\angle RQF_Q$ must be acute. But the angles are supplementary so they cannot both be acute. □

Theorem 37.5 *The function f has no local maxima and at most one local minimum.*

Proof Suppose that f has a local maximum at a point B. Then, since f is continuous, for a sufficiently small value positive value ϵ, there are points A and C on either side of B so that

$$f(A) = f(C) = f(B) - \epsilon.$$

Let F_A, F_B, and F_C be the feet of the perpendiculars to ℓ_1 from A, B, and C respectively. Then

$$|AF_A| = f(A) = f(C) = |CF_C|$$

so $\square F_AF_CCA$ is a Saccheri quadrilateral with base F_AF_C. The segment BF_B connects the summit to the base and is perpendicular to the base. We saw at the end of the last chapter that such segments cannot be longer than the legs. But the legs of this Saccheri quadrilateral are AF_A and CF_C, and they are shorter (by ϵ) than BF_B. This is a contradiction, and so f cannot have a local maximum.

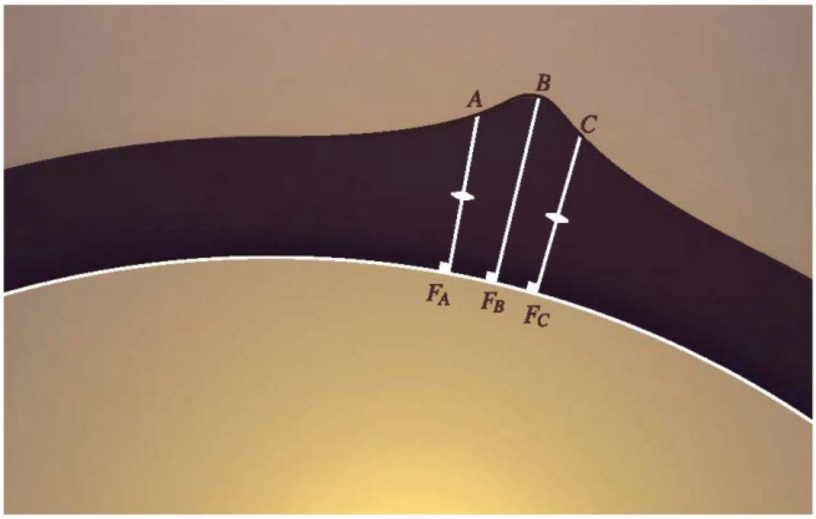

Figure 7. It is not possible for f to have a local maximum (as is suggested here at the point B).

The second part of the theorem follows immediately: if f had more than one minimum, then, because it is continuous, it would have a local maximum in between them. □

Theorem 37.6 *Let P and Q be points on ℓ_2 and suppose that $f(Q) > f(P)$. Then f approaches infinity along $\overrightarrow{PQ}$.*

It should be clear that the value of f continues to increase past Q. If at some point it began to decrease, then we would have found a local maximum, and we have just seen that f has none. But could it be that f continues to grow, but is bounded above? Is there an upper bound that f approaches in the limit but never surpasses? This is a more delicate question. The strategy of the proof is to build a wall of congruent Saccheri quadrilaterals that appears to contain $\overrightarrow{PQ}$, and then see why that cannot happen.

Proof Suppose that f does have an upper bound so that it approaches but never exceeds a number k. Starting from the point F_Q, the foot of the perpendicular from Q, mark off congruent segments along ℓ_1 on the same side of PF_P as $\overrightarrow{PQ}$. That is, label points F_i so that

$$F_Q = F_0 * F_1 * F_2 * F_3 * \cdots$$

and so that the segments F_iF_{i+1} are congruent. Above each point F_i (on the same side of ℓ_1 as ℓ_2), erect a perpendicular segment F_iP_i of length k. Then each $\square F_iF_{i+1}P_{i+1}P_i$ is a Saccheri quadrilateral, and by SASAS they are all congruent.

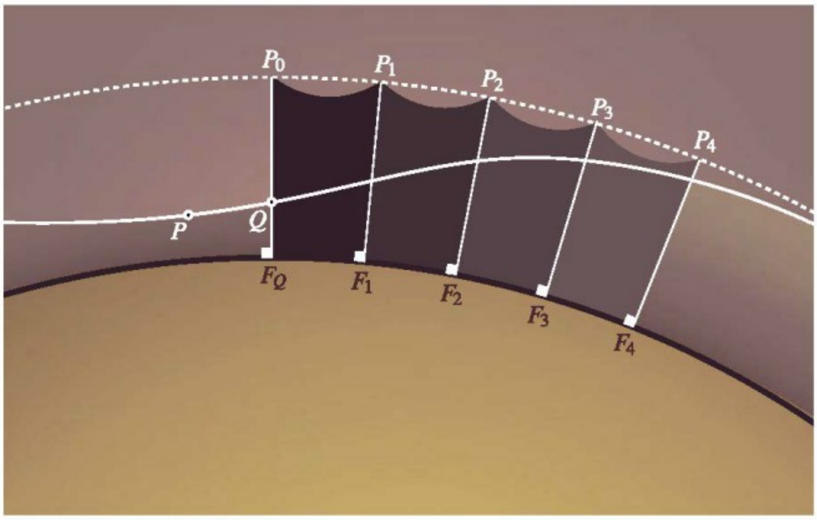

Figure 8. It is not possible for f to level off at some asymptotic value, as suggested here.

Their altitudes are shorter than their legs, by an amount that we can call ϵ (since the quadrilaterals are congruent, the discrepancy will be the same for all of them). Now let's get back to $\overrightarrow{PQ}$. For a while, it may run below the summits of the quadrilaterals we have assembled. But f has a limiting value of k along it, so eventually there is a point beyond which f is greater than $k - \epsilon$. Look beyond that point, at the interaction of $\overrightarrow{PQ}$ with one of the summits P_nP_{n+1}. Points P_n and P_{n+1} are on the opposite side of ℓ_2 from ℓ_1. But the altitude between them is on the same side of ℓ_2 as ℓ_1. This would imply that ℓ_2 intersects P_nP_{n+1} twice. Two distinct lines can intersect only once, so this is a contradiction. □

Looking in the other direction, if $f(Q) > f(P)$, then $f(P) < f(Q)$, and it is natural to wonder what happens as we progress along $\overrightarrow{QP}$. Does the distance to ℓ_2 continue to decrease, or does it at some point begin to increase again?

Definition 37.7 If f achieves a minimum, then ℓ_1 and ℓ_2 are called *divergent* parallels. If instead f approaches zero, then ℓ_1 and ℓ_2 are called *asymptotic* parallels.

If ℓ_1 and ℓ_2 are divergent parallels, then once $\overrightarrow{QP}$ has crossed over the minimum value, the value of f will begin to go up. Once that happens, by the previous result, f will go to infinity. Both types of parallels exist in non-Euclidean geometry. The base and summit of a Saccheri quadrilateral are divergent parallels. Asymptotic parallels are more elusive, but one of the exercises outlines a proof of their existence. Beyond that, there is one more obvious question: are there any parallels that are neither divergent nor asymptotic? Suppose ℓ_1 and ℓ_2 are not divergent. Then f does not achieve its minimum. It is easy then to pick points P and Q

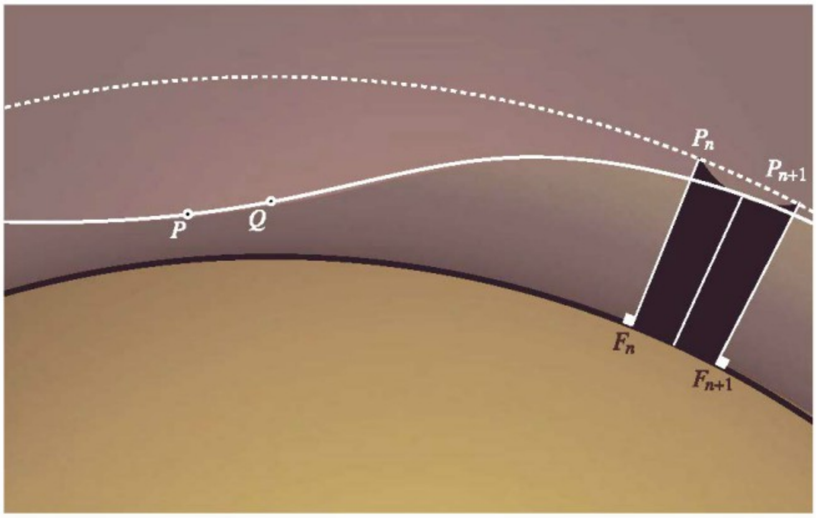

Figure 9. If $\overrightarrow{PQ}$ approached an asymptote, it would eventually intersect the summit of a constructed Saccheri quadrilateral more than once.

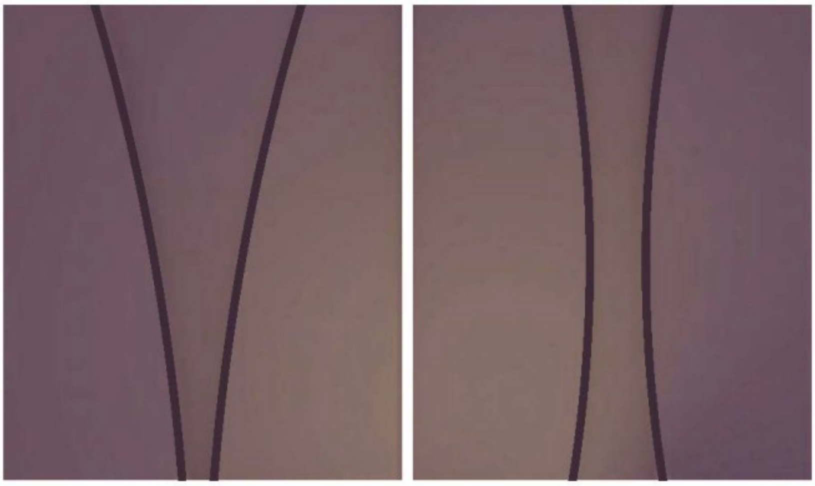

Figure 10. (l) asymptotic parallels; (r) divergent parallels.

so that $f(P) < f(Q)$, so f must continue to decrease past P as we move out along $\overrightarrow{QP}$. By definition, the lines are asymptotic if the value of f decreases to zero. Is it possible that the value of f is always decreasing, but never drops below some positive number k? No. The argument against a nonzero lower bound is essentially the same as the argument against an upper bound in the proof above.

Theorem 37.8 *If ℓ_1 and ℓ_2 are parallel, then they are either divergent or asymptotic.*

Exercise 37.3 Prove theorem 37.8.

The results of this chapter should give you some sense of the look of non-Euclidean geometry. But can such a geometry exist? If we want to answer YES, we need to find a model that meets all the requirements of neutral geometry, and has the non-Euclidean characteristics. In the next chapters, we will look at such models.

37.1 Exercises

37.4. Let r_1 and r_2 be two non-opposite rays with endpoint A. Label point B and C on r_1 and B' and C' on r_2 so that $A * B * C$ and $A * B' * C'$. Prove that the angle sum of $\triangle ACC'$ is less than the angle sum of $\triangle ABB'$.

37.5. Suppose that ℓ_1 and ℓ_2 are divergent parallels. Let P be the point on ℓ_2 that is the minimum distance from ℓ_1, and let Q be the foot of the perpendicular from P. Prove that QP intersects ℓ_2 at a right angle.

37.6. Let ℓ be a line and let P be a point that is not on ℓ. Prove that there is a line through P which is asymptotic to ℓ. Hint: let r_1 and r_2 be rays from P, one that intersects ℓ and one that is along a divergent parallel to ℓ. Let A be a point on r_1 and B be a point on r_2. A point Q on AB falls into one of two sets: either PQ is parallel to ℓ or it isn't. By Dedekind's axiom, there is a point Q_0 that separates those two sets. Show that PQ_0 is an asymptotic parallel.

37.7. Suppose that lines ℓ_1 and ℓ_2 intersect at a point O. Suppose that P and Q are on ℓ_2 so that $O * P * Q$, and that F_P and F_Q are the feet of the perpendiculars, from P and Q respectively, to ℓ_1. Prove that $|PF_P| < |QF_Q|$.

37.8. Let r_1 and r_2 be two rays with a common endpoint O. Prove that for a positive real number x, there is a point P on r_1 so that $d(P, r_2) > x$.

06.05.2020

E3) Calculus deconstructed _ a second course in first-year... pdf

Pág. 135 (152) 4 Differentiation

Pág. 137 (153) Figure 4.1 $Ax + By = C$

Pág. 247 (264) 5 Integration.

E4) Geometry Illuminated - An Illustrated Introduction... pdf

Obs. Livro todo!!!

Pág. 105 (122) Part II Euclidean Geometry

Pág. 223 (240) Part III Euclidean Transformations

↳ muito bom!!!!

Obs. Aplicar 'quando' (no momento) da projeção da superfície curva sobre um plano.
(simular vice-versa também)!

Pág. 395 (412) Part IV Hyperbolic Geometry.

E5) Calculus in 3D. Geometry, vectors, and... pdf

Obs. Livro todo!!!

Pág. 115 (127) Chapter 2
Curves and Vector-Valued Functions of One Variable.

F)*+,)))-) ' "0 . A2 8 2
! 8 2 ** 6!
(1 6 8 ! ! < 7

2

Curves and Vector-Valued Functions of One Variable

2.1 Conic Sections

We begin this chapter by looking at the *conic sections*, which were regarded by the Greeks as the simplest curves after the straight line and circle.

A major source of information about classical Greek mathematics is Pappus of Alexandria (*ca.* 300 AD), a formidable geometer of the late third century AD.[1] In his *Mathematical Collection*[2] he surveyed the work of his predecessors; many of these works have been lost. He classified mathematical problems according to the kinds of loci (curves) required for their solution:

- *planar* problems can be solved using circles and straight lines, or **planar loci**. These are often called **compass and straightedge constructions**;
- *solid* problems involve the intersection of a plane with a cone (**solid loci**, or **conic sections**);

---

[1] The work of Pappus is sometimes taken to mark the end of the classical Greek tradition in mathematics.

[2] Parts of this survive in a twelfth-century copy.

A

Conic Sections: Apollonius' approach

Here we give some more details of the argument for certain assertions in § 2.1. Our exposition loosely follows [27, pp. 355-9].
Recall the setup (Figure A.1):

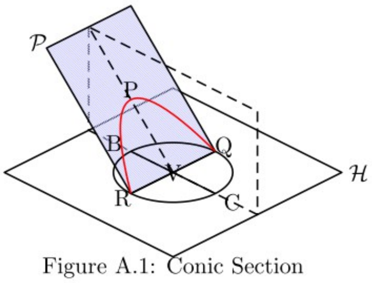

Figure A.1: Conic Section

We wish to investigate the conic section $\gamma = \mathcal{P} \cap \mathcal{K}$, where the plane $\mathcal{P}$ does not contain the origin, and intersects any horizontal plane in a line parallel to the x-axis. The yz-plane (which is perpendicular to any such line) intersects γ in a point P, and possibly in a second point P'—the vertices of γ. Given a point Q on γ distinct from the vertices (*i.e.*, not in the yz-plane), the horizontal plane $\mathcal{H}$ through Q intersects $\mathcal{K}$ in a circle containing Q. The intersection of $\mathcal{H}$ with the yz-plane is a line through the center of this circle, and so contains a diameter BC of the circle. We draw the chord of the circle through Q perpendicular to this diameter, denoting by R the other end of the chord, and by V its intersection with BC. The

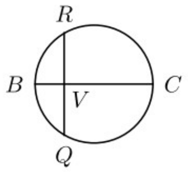

Figure A.2: Lines in $\mathcal{H}$

line segments QV and PV are, respectively, the **ordinate** and **abcissa**. We know that V bisects QR, and also, by Prop. 13, Book VI of the *Elements*, that

$$|QV|^2 = |QV| \cdot |VR| = |BV| \cdot |VC| . \tag{A.1}$$

Parabolas: Suppose first that PV is parallel to AC, so that P is the only vertex of γ. Consider the triangle $\triangle ABC$ in the yz-plane, noting that P lies on AB and V lies on BC (Figure A.3). Since AC is parallel to PV,

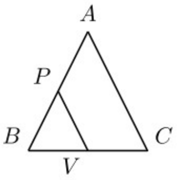

Figure A.3: Equation (A.2)

the triangles $\triangle ABC$ and $\triangle PBV$ are similar and (since $AD = AC$) isosceles. In particular,

$$\frac{|BV|}{|BP|} = \frac{|BC|}{|BA|}$$

and (again since AC and PV are parallel)

$$\frac{|VC|}{|BC|} = \frac{|PA|}{|BA|}$$

or equivalently

$$\frac{|VC|}{|PA|} = \frac{|BC|}{|BA|}.$$

Multiplication of these two equations yields

$$\frac{|BV|\,|VC|}{|BP|\,|PA|} = \left(\frac{|BC|}{|BA|}\right)^2.$$

Since $|BP| = |PV|$, we conclude that

$$|BV|\,|VC| = \left[\left(\frac{|BC|}{|BA|}\right)^2 |PA|\right] |PV|\,. \tag{A.2}$$

Note that replacing Q with another point Q' on γ replaces $\mathcal{H}$ with a parallel plane $\mathcal{H}'$, and gives a picture similar to Figure A.3 (Figure A.4). In particular, the quantity in brackets in Equation (A.2) depends only on γ: it is called the **parameter of ordinates** for γ—we will denote it by p. Apollonius represents it by a line segment PL perpendicular to the abcissa PV (indicated in Figure A.4), called the **orthia**,[1] or in Latin, the **latus rectum** .[2]

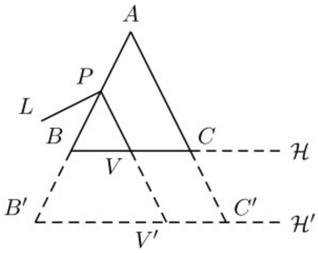

Figure A.4: Independence of Q

---

[1] Fried and Unguru argue that for Apollonius, the orthia is a specific line segment, not the representation of an independent quantity.

[2] The term "latus rectum" is also used to refer to the **parameter of ordinates** (the number p in Equation (A.3)).

If we introduce rectangular coordinates in $\mathcal{P}$ with the origin at P and axes parallel to QV and PV, setting

$$y = |QV|$$
$$x = |PV|$$

then Equation (A.1) says that

$$y^2 = |BV|\,|VC|\,,$$

and substituting this information into Equation (A.2) leads to the equation

$$y^2 = px \tag{A.3}$$

where p is the parameter of ordinates defined above.

For the other two cases, when PV is not parallel to AC, then the line PV (extended) meets the line AB (extended) at the second vertex P'. If ϕ denotes the (acute) angle between $\mathcal{P}$ and a horizontal plane $\mathcal{H}$, then V lies between P and P' if $0 \leq \phi < \frac{\pi}{2} - \alpha$ and P lies between V and P' if $\frac{\pi}{2} - \alpha < \phi \leq \frac{\pi}{2}$.

Ellipses: In the first case (see Figure A.5), let J be the point at which the line through A parallel to PV (and hence to PP') meets BC (extended). As in the case of the parabola (but with C replaced by J), the triangles $\triangle ABJ$ and $\triangle PBV$ are similar (but no longer isosceles), so

$$\frac{|BV|}{|PV|} = \frac{|BJ|}{|AJ|}.$$

Also, since the lines AP' and VJ are transversals to the two parallels AJ and PP' meeting at C, the triangles $\triangle AJC$ and $\triangle P'VC$ are similar, so

$$\frac{|VC|}{|VP'|} = \frac{|JC|}{|AJ|}.$$

Multiplying these equalities and invoking Equation (2.2), we have

$$\frac{|QV|^2}{|PV|\,|VP'|} = \frac{|BV|\,|VC|}{|PV|\,|VP'|} = \frac{|BJ|\,|JC|}{|AJ|^2}$$

or, as the analogue of Equation (A.2),

$$|QV|^2 = \left[\left(\frac{|BJ|\,|JC|}{|AJ|^2}\right)|VP'|\right]|PV|\,. \tag{A.4}$$

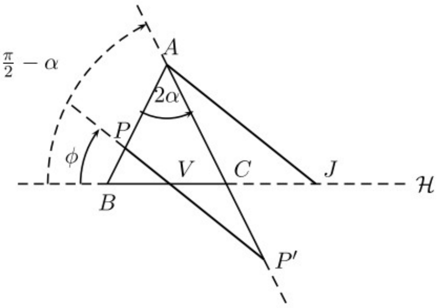

Figure A.5: $0 \leq \phi \leq \frac{\pi}{2} - \alpha$

Again as in the case of the parabola, the fraction in parentheses in Equation (A.4) depends only on the curve γ. We again form the "orthia" of γ, a line segment PL perpendicular to PV with length[3]

$$p = |PL| = \left(\frac{|BJ|\,|JC|}{|AJ|^2}\right)|PP'|\,. \tag{A.5}$$

Now let S be the intersection of LP' with the line through V parallel to PL (Figure A.6). Note that the triangles $\triangle LP'P$ and $\triangle SP'V$ are similar, so

$$\frac{|PL|}{|PP'|} = \frac{|VS|}{|VP'|}$$

---

[3] This defintion of the orthia is on the face of it quite different from the definition in the case of the parabola. I have not been able to find a reasonable explanation of why the two definitions yield analogous line segments. It can be shown (M. N. Fried, private correspondence) that if one considers the orthia of hyperbolic or elliptic sections whose diameter has inclination approaching that of a generator (*i.e.*, approaching a parabolic section), then these orthia tend toward the parabolic orthia of the limit. This, however, is clearly an anachronistic point of view, and Fried has pointed out that Apollonius never discusses varying the section. In fact, in his view, Apollonius did not intend the properties of the conic sections with respect to application of areas—known as the **symptomata**—to unify the different types; he viewed them as separate objects, and only noted these somewhat analogous properties as incidental observations. Fried's point of view (but not a specific commentary on this issue) is given at length in [16, pp. 74-90].

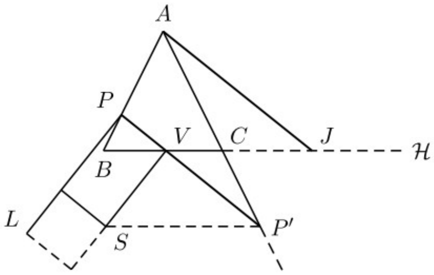

Figure A.6: Definition of S

and substituting this (via Equation (A.5)) into Equation (A.4), we have

$$\begin{aligned}|QV|^2 &= \left[\left(\frac{|PL|}{|PP'|}\right)|VP'|\right]|PV| \\ &= \left[\left(\frac{|VS|}{|VP'|}\right)|VP'|\right]|PV|\end{aligned}$$

or

$$|QV|^2 = |VS| \cdot |PV|\,. \tag{A.6}$$

This is like Equation (A.3), but $|PL|$ is replaced by the shorter length $|VS|$. To obtain the rectangular equation of the ellipse, we set

$$d = |PP'|$$

(the **diameter**): by similarity of $\triangle LP'P$ and $\triangle SP'V$,

$$\frac{|VS|}{|PL|} = \frac{|VP'|}{|PP'|} = 1 - \frac{|PV|}{|PP'|}$$

so (again setting $x = |PV|$ and $y = |QV|$) we have as the equation of the ellipse

$$y^2 = |VS|\,x = p\left(1 - \frac{x}{d}\right)x = px - \frac{p}{d}x^2. \tag{A.7}$$

Hyperbolas: In the final case, when $\frac{\pi}{2} - \alpha < \phi \leq \frac{\pi}{2}$, P lies between V and P' (Figure A.7). Formally, our constructions in this case are the same as in the case of the ellipse: as before, the two similarities $\triangle ABJ \sim \triangle PBV$ and $\triangle AJC \sim \triangle P'VC$ lead to Equation (A.4); we form

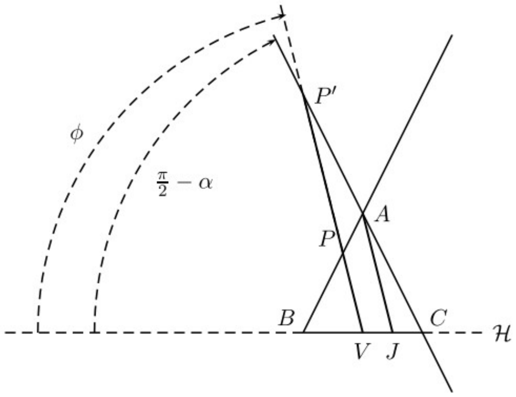

Figure A.7: $\frac{\pi}{2} - \alpha < \phi \leq \frac{\pi}{2}$

the orthia PL perpendicular to PV satisfying Equation (A.5) and let S be the intersection of $P'L$ (extended) with the line through V parallel to PL (Figure A.8). The same arguments as in the ellipse case yield Equation (A.6), but this time the segment VS *exceeds* PL.

A verbatim repetition of the calculation leading to Equation (A.7) leads to its hyperbolic analogue,

$$y^2 = px + \frac{p}{d}x^2. \tag{A.8}$$

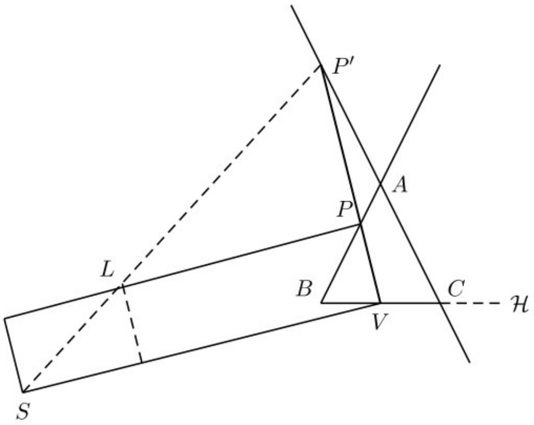
P'
A
P
L
B
C
$\mathcal{H}$
V
S

Figure A.8: Definition of S

H

Surface Area: the Counterexample of Schwarz and Peano

We present here an example, due to H. A. Schwarz [48] and G. Peano [44][1] which shows that the analogue for surfaces of the usual definition of the length of a curve cannot work.

Recall that if a curve $\mathcal{C}$ was given as the path of a moving point $\overrightarrow{p}(t)$, $a \leq t \leq b$, we partitioned $[a, b]$ via $\mathcal{P} = \{a = t_0 < t_1 < \cdots < t_n\}$ and approximated $\mathcal{C}$ by a piecewise-linear path consisting of the line segments $[\overrightarrow{p}(t_{i-1}), \overrightarrow{p}(t_i)]$, $i = 1, \ldots, n$. Since a straight line is the shortest distance between two points, the distance travelled by $\overrightarrow{p}(t)$ between $t = t_{i-1}$ and $t = t_i$ is at least the length of this segment, which is $\|\overrightarrow{p}(t_i - \overrightarrow{p}(t_{i-1}))\|$. Thus, the total length of the piecewise-linear approximation is a lower bound for the length of the actual path: we say $\mathcal{C}$ is *rectifiable* if the supremum of the lengths of all the piecewise-linear paths arising from different partitions of $[a, b]$ is finite, and in that case define the (arc)length of the curve to be this supremum. We saw in § 2.5 that every regular arc (that is, a curve given by a one-to-one differentiable parametrization with non-vanishing velocity) the length can be calculated from any regular parametrization as

---

[1] Schwarz tells the story of this example in a note [49, pp. 369-370]. Schwarz initially wrote down his example in a letter to one Angelo Gnocchi in December 1880, with a further clarification in January 1881. In May, 1882 Gnocchi wrote to Schwarz that in a conversation with Peano, the latter had explained that Serret's definition of surface area (to which Schwarz's example is a counterexample) could not be correct, giving detailed reasons why it was wrong; Gnocchi had then told Peano of Schwarz's letters. Gnocchi reported the example in the Notices of the Turin Academy, at which point it came to the attention of Charles Hermite (1822-1901), who publihsed the correspondence in his *Cours d'analyse*. Meanwhile, Peano published his example. After seeing Peano's article, Schwarz contacted him and learned that Peano had independently come up with the same example in 1882.

$$\mathfrak{s}\left(\mathcal{C}\right) = \int_a^b \left\| \dot{\overrightarrow{p}}(t) \right\| \, dt.$$

The analogue of this procedure could be formulated for surface area as follows. [2] Let us suppose for simplicity that a surface $\mathfrak{S}$ in $\mathbb{R}^3$ is given by the parametrization $\overrightarrow{p}(s, t)$, with domain a rectangle $[a, b] \times [c, d]$. If we partition $[a, b] \times [c, d]$ as we did in § 5.1, we would like to approximate $\mathfrak{S}$ by rectangles in space whose vertices are the images of "vertices" $p_{i,j} = \overrightarrow{p}(x_i, y_j)$ of the subrectangles R_{ij}. This presents a difficulty, since *four* points in $\mathbb{R}^3$ need not be coplanar. However, we can easily finesse this problem if we note that *three* points in $\mathbb{R}^3$ are *always* contained in some plane. Using diagonals (see Figure H.1)[3] we can divide each subrectangle R_{ij} into two triangles, say

$$\begin{aligned} T_{ij1} &= \triangle p_{i-1,j-1} p_{i,j-1} p_{i,j} \\ T_{ij2} &= \triangle p_{i-1,j-1} p_{i-1,j} p_{i,j}. \end{aligned}$$

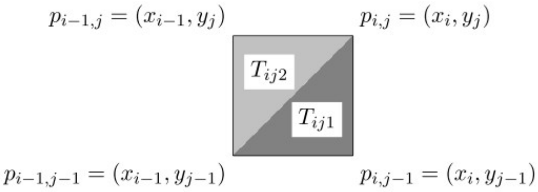
$p_{i-1,j} = (x_{i-1}, y_j)$
$p_{i,j} = (x_i, y_j)$
T_{ij2}
T_{ij1}
$p_{i-1,j-1} = (x_{i-1}, y_{j-1})$
$p_{i,j-1} = (x_i, y_{j-1})$

Figure H.1: Dividing R_{ij} into triangles

This tiling $\{T_{ijk} \,|\, i = 1, \ldots, m, \quad j = 1, \ldots, n, \quad k = 1, 2\}$ of the rectangle $[a, b] \times [c, d]$ is called a **triangulation**. Now, it would be natural to try to look at the total of the areas of the triangles whose vertices are the points $\overrightarrow{{p^*}_{i,j}} = \overrightarrow{p}(p_{i,j})$ on the surface

---

[2] This approach was in fact followed by J. M. Serret

[3] There are two ways to do divide each rectangle, but as we shall see, this will not prove to be an issue.

$$\begin{aligned} T^*_{ij1} &= \triangle \overrightarrow{{p^*}_{i-1,j-1}}\,\overrightarrow{{p^*}_{i,j-1}}\,\overrightarrow{{p^*}_{i,j}} \\ T^*_{ij2} &= \triangle \overrightarrow{{p^*}_{i-1,j-1}}\,\overrightarrow{{p^*}_{i-1,j}}\,\overrightarrow{{p^*}_{i,j}}. \end{aligned}$$

and define the area of $\mathfrak{S}$ to be the supremum of these over all triangulations of $[a, b] \times [c, d]$.

Unfortunately, this approach doesn't work; an example found (simultaneously) in 1892 by H. A. Schwartz and G. Peano shows

Proposition H.0.1. *There exist triangulations for the standard parametrization of the cylinder such that the total area of the triangles is arbitrarily high.*

Proof. Consider the finite cylinder surface

$$\begin{aligned} x^2 + y^2 &= 1 \\ 0 \leq z &\leq 1. \end{aligned}$$

We partition the interval $[0, 1]$ of z-values into m equal parts using the $m + 1$ horizontal circles $z = \frac{k}{m}$, $k = 0, \ldots, m$. Then we divide each circle into n equal arcs, but in such a way that the endpoints of arcs on any particular circle are directly above or below the midpoints of the arcs on the neighboring circles. One way to do this is to define the angles

$$\theta_{jk} = \begin{cases} \frac{2\pi j}{n} & \text{for } k \quad \textit{even}, \\ \frac{2\pi j}{n} - \frac{\pi}{n} & \text{for } k \ \textit{odd} \end{cases}$$

and then the points

$$p_{jk} = (\cos \theta_{jk}, \sin \theta_{jk}, \frac{k}{m}).$$

That is, the points $\{p_{jk}\}$ for k fixed and $j = 1, \ldots, n$ divide the k^{th} circle into n equal arcs. Now consider the triangles whose vertices are the endpoints of an arc and the point on a neighboring circle directly above or below the midpoint of that arc (Figure H.2).

The resulting triangulation of the cylinder is illustrated in Figure H.3.

To calculate the area of a typical triangle, we note first (Figure H.4) that

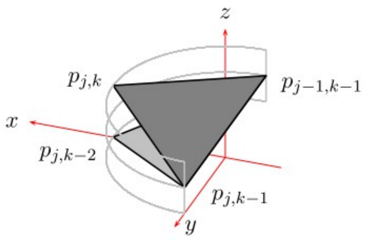

z
$p_{j,k}$
$p_{j-1,k-1}$
x
$p_{j,k-2}$
$p_{j,k-1}$
y

Figure H.2: Typical Triangles

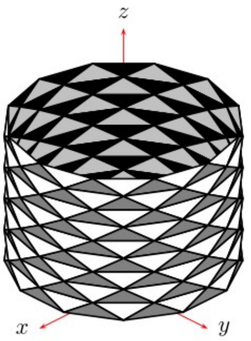

Figure H.3: Triangulation of the Cylinder

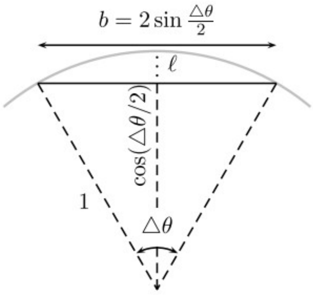
$b = 2\sin\frac{\triangle\theta}{2}$
ℓ
$\cos(\triangle\theta/2)$
1
$\triangle\theta$

Figure H.4: The Base of a Typical Triangle

its base is a chord of the unit circle subtending an arc of $\triangle\theta = \frac{2\pi}{n}$ radians; it follows from general principles that the length of the chord is

$$b_n = 2\sin\frac{\triangle\theta}{2} = 2\sin\frac{\pi}{n}.$$

We also note for future reference that the part of the perpendicular bisector of the chord from the chord to the circle has length

$$\ell = 1 - \cos\frac{\triangle\theta}{2} = 1 - \cos\frac{\pi}{n}.$$

To calculate the height of a typical triangle, we note that the vertical distance between the plane containing the base of the triangle and the other vertex is $\frac{1}{m}$, while the distance (in the plane containing the base) from the base to the point below the vertex (the dotted line in Figure H.5) is $\ell = 1 - \cos\frac{\pi}{n}$; it follows that the height of the triangle (the dashed line in Figure H.5) is itself the hypotenuse of a right triangle with sides ℓ and $\frac{1}{m}$, so

$$h_{m,n} = \sqrt{\left(\frac{1}{m}\right)^2 + \ell^2} = \sqrt{\left(\frac{1}{m}\right)^2 + \left(1 - \cos\frac{\pi}{n}\right)^2}.$$

Thus the area of a single triangle of our triangulation (for a given choice of

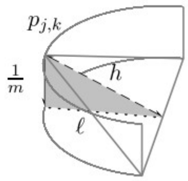

Figure H.5: Calculating the Height of a Typical Triangle

m and n) is

$$\begin{aligned}\triangle A_{m,n} &= \frac{1}{2} b_n h_{m,n} \\ &= \frac{1}{2}\left[2\sin\frac{\pi}{n}\right]\sqrt{\left(\frac{1}{m}\right)^2 + \left(1-\cos\frac{\pi}{n}\right)^2} \\ &= \left[\sin\frac{\pi}{n}\right]\sqrt{\left(\frac{1}{m}\right)^2 + \left(1-\cos\frac{\pi}{n}\right)^2}.\end{aligned}$$

Now let us count the number of triangles. There are $m+1$ horizontal circles, each cut into n arcs, and the chord subtending each arc is the base of exactly two triangles of our triangulation, except for the two "end" circles, for which each chord is the base of *one* triangle. This means there are $2mn$ triangles, giving a total area of

$$\begin{aligned}A_{m,n} &= 2mn\triangle Am,n \\ &= 2mn\left[\sin\frac{\pi}{n}\right]\sqrt{\left(\frac{1}{m}\right)^2 + \left(1-\cos\frac{\pi}{n}\right)^2} \\ &= 2\left[n\sin\frac{\pi}{n}\right]\sqrt{1+m^2\left(1-\cos\frac{\pi}{n}\right)^2}.\end{aligned}$$

Now, the quantity in brackets converges to π as $n \to \infty$ and, for n fixed, the square root goes to ∞ as $m \to \infty$; it follows that we can fix a sequence of pairs of values $\{(m_k, n_k)\}$ (for example, as Schwarz suggests, $m = n^3$) such that the quantity A_{m_k,n_k} diverges to infinity, establishing that the supremum of the total area of piecewise-linear approximations of a cylinder is infinite, and hence gives a bad definition for the area of the cylinder itself. □

To see what is going on here, we might note that at the vertex opposite the chord of each triangle, the plane tangent to the cylinder is vertical, while the plane of the triangle makes an angle with it of size $\theta(m,n)$, where

$$\tan\theta(m,n) = \frac{\ell}{1/m} = \frac{1-\cos\frac{\pi}{n}}{1/m} = m\left(1-\cos\frac{\pi}{n}\right).$$

If for example $m = n^3$, one can check (using, *e.g.*, L'Hôpital's rule) that the tangent goes to infinity with n, so in the limit the triangles approach being *perpendicular* to the cylinder.

It turns out that the approach of Serret (using these triangulations) can be made to work, provided we replace the *supremum* of all such total areas with the *limit*, as $\varepsilon \to 0$, of the *infimum* of all such total areas for triangulations with mesh size less than ε. In effect, this means we are looking at triangulations in which the triangles are as close as possible to being tangent to the cylinder.

F')*+,)))-) ' 5 * 7 ! . A
$2 * A 6!
-7 5 8 ! 999

1 ? G >1$> # **#

?--J " % $ 1L)?--J > > ")? G >1$>) **)

1)*+,)))-) ' & , !
&2 && ** " 2 / * * !H8
B
@

F1)*+,)))-) ' 5 $$ 2
* * !H8
($' % & '1

)*+,)))-) ' 5 $$ 2
* * > % 7 6

7)*+,)))-) ' 5 $$ 2
* * > 7 6

data 06.05.2020

Pág. 707(719) A. Conic Sections: Apollonius' approach

Pág. 791(803) H. Surface Area: the Counterexample of Schwarz and Peano.

E6) Linear Algebra and Geometry (2019...) ... pdf

Obs. Livro todo!!!

E7) ~~Maa Dictionary Maasai...pdf~~

Books_79_2020/ .djvu
MAA/

E7.1) In mathematical circles. Quadrants I, II ... djvu

obs.: Historic

E7.2) Lure of the integers.. djvu

Pág. 196(214) Integer 67

An Inequality on $\pi(x)$

It is not known whether or not

$$\pi(x+y) \leq \pi(x) + \pi(y)$$

for all sufficiently large x, y.

data 06.05.2020

Extending work of Rosser and Schoenfeld, various authors have proved related results.

Krawczyk proved the following, $\pi(kx) < k\pi(x)$ for $k \geqslant e, x \geqslant 67$.

For $\epsilon > 0, x \geqslant 17, y \geqslant 17, x+y \geqslant 1 + e^{4(1+1/\epsilon)}$, it has been shown that

$$\pi(x+y) < (1+\epsilon)(\pi(x)+\pi(y)).$$

There is an interesting connection between Conjecture A ($\pi(x+y) \leq \pi(x)+\pi(y)$, $x, y \geqslant 2$) and the generalization of the twin prime conjecture that follows it.

E7.3) Symmetry ... djvu

Pag. 11(24) Chapter 2 Inside and Outside

2.1 Reflecting in a Circle'. (inversion)

Important:

Pag. 14(27) 2.2 Composition of two inde-

1 ")*+,)))-) ' A A ** 6
!H8
(% 6 & ! ! - !
)*+,)))-) ' A A ** 6
% %" 6!

CHAPTER 2

Inside and Outside

2.1 REFLECTING IN A CIRCLE[1]

If, on a square grid, a gridline h is chosen as axis of a straight-line reflection, the reflection of lattice points can be done by "counting squares" (Figure 18). In this way one half-plane is mapped onto the other half-plane.

Now we try, with a comparable procedure, to map the interior of a circle on the exterior, and the other way round. This seems at first sight self-contradictory, since the outside of a circle is much "bigger" than its interior. However, Figure 19 shows a "square structure", in which the "squares" become smaller as one moves inwards. Thus, indeed, can a mapping of A onto A' be achieved, square by square.

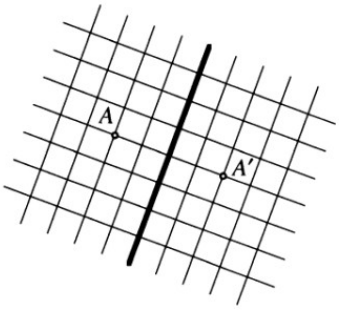

FIGURE 18
Reflection in a gridline on a square grid

---

[1] Here, and subsequently, the author has "reflection" where many standard English-language texts would speak of "inversion."

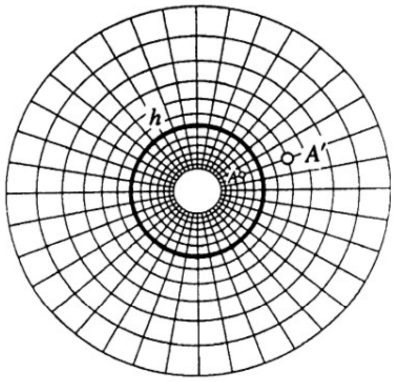

FIGURE 19
Reflection in a circle

This mapping we call a ***circle-reflection*** (or ***inversion***). The highlighted circle h is called the ***reference circle;*** it corresponds to the axis of a straightline-reflection. In constructing such a "circular square" pattern we have choices, so we must clarify how the radii of the concentric circles for the scaling $t = \cdots, -2, -1, 0, 1, 2, \cdots$ are related to each other (Figure 20). For that purpose, we imagine the "square pattern" further subdivided, so that the quadrilaterals are almost squares.

We now study such a "near-square", lying in a circular sector of angle α radians at a distance r from the center of the circle (Figure 21). This near-square has a side of length Δr in the radial direction and a side of length

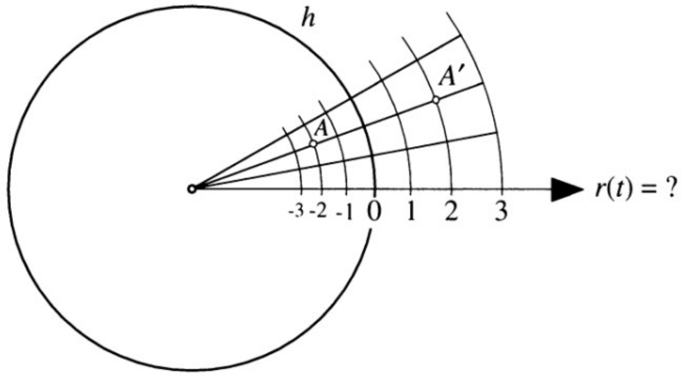

FIGURE 20
$r(t) = ?$

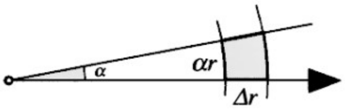

FIGURE 21
Near-squares

αr in the tangential direction. We may reason heuristically as follows. These two lengths should be equal, so that $\Delta r = \alpha r$. In terms of differentials we obtain $dr = \alpha r$, so, through separation of variables,

$$\frac{dr}{r} = \alpha,$$

and thus, through integration, $\ln r = \alpha t + C_1$, so that, finally, $r(t) = Ce^{\alpha t}$. If we denote the radius of the reference-circle h by r_0, we get, from the boundary condition $r(0) = r_0$, the function

$$r(t) = r_0 e^{\alpha t}.$$

The radii depend exponentially on t. As $t \to -\infty$, $r \to 0$; thus the interior of the circle has room for infinitely many "squares". For the further investigation of the mapping of A onto A' polar coordinates are appropriate (Figure 22).

The point A and its image A' have the same polar angle ϕ, and their polar distances belong to equal and opposite t-values: thus

$$r = r_0 e^{\alpha t}, \quad r' = r_0 e^{-\alpha t}.$$

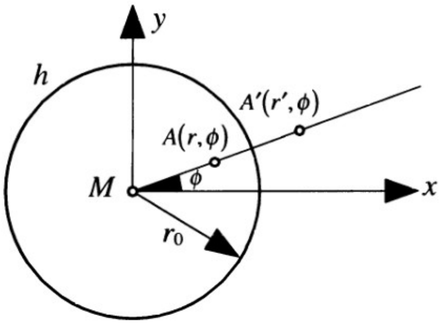

FIGURE 22
Polar coordinates

Thus we obtain

$$rr' = r_0^2.$$

Remark If we choose a unit circle for the reference circle h, so that $r_0 = 1$, then the polar distances r and r' of a point and its image are reciprocals of each other. In this case the mapping can be described in the complex plane by $w = f(z) = \frac{1}{z}$.

2.2 COMPOSITION OF TWO CIRCLE-REFLECTIONS

We now study the case of two circle-reflections (or inversions) whose reference-circles h and H are concentric, with radii r_0 and R_0. Under reflection in h the point A is mapped to the point A'; by reflection in H the point A' is mapped to the point A'' (Figure 23). In the passage from A to A' we have $rr' = r_0^2$; in going from A' to A'' we have $r'r'' = R_0^2$. For the passage from A to A'' we thus have $\frac{r''}{r} = \frac{R_0^2}{r_0^2}$; so the composition of two circle-reflections with concentric reference-circles is a central dilatation with magnification ratio $\frac{R_0^2}{r_0^2}$; the center of dilatation is the common center of the two reference-circles.

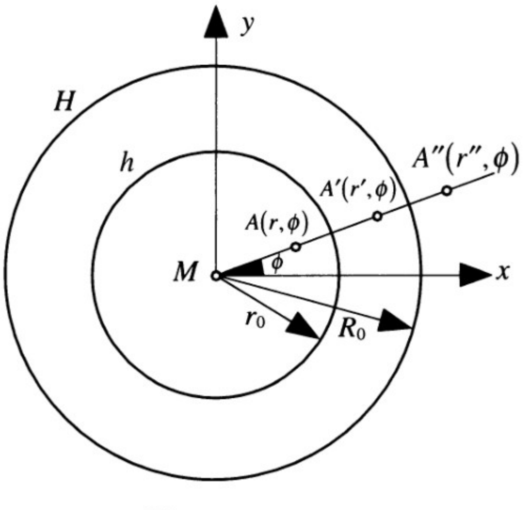

FIGURE 23
The composition of two circle-reflections

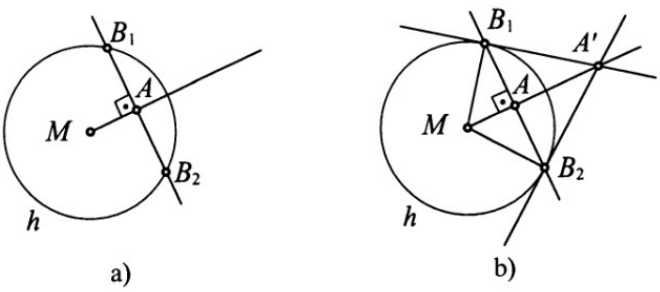

FIGURE 24
Direct construction of the image point

2.3 DIRECT CONSTRUCTION OF THE IMAGE POINT

Figure 24 shows in two steps how, for a given point A in the interior of the reference-circle h, the image point A' may be constructed.

To prove the correctness of this construction we consider the right-angled triangle $MA'B_1$ (Figure 24b) with its right angle at B_1. In this triangle r_0 is the length of MB_1, r is the length of the part MA of the hypotenuse, and r' is the length of the hypotenuse. By the similarity of the triangles MAB_1 and MB_1A' we have $rr' = r_0^2$.

Question 2.1 How may the point A' be constructed, when its pre-image A lies outside the reference-circle h?

Question 2.2 Which are the fixed points of a circle-reflection?

2.4 CIRCLE-REFLECTION INVARIANTS

The usual straight-line reflection is a congruence mapping: thus the images of straight lines, circles, squares, etc. are again straight lines, circles, squares, etc. The circle-reflection is not a congruence mapping, since the bounded interior of the reference-circle is mapped onto the unbounded exterior. Nevertheless some configurations are such that their images are of the same kind.

A bishop on a chessboard follows along the diagonals of neighboring ("by corners") chessboard squares. A sequence of such diagonals lies naturally on a straight line. The situation is very different in the case of the square grid of Figure 19. Here a sequence of square-diagonals gives rise to a curved

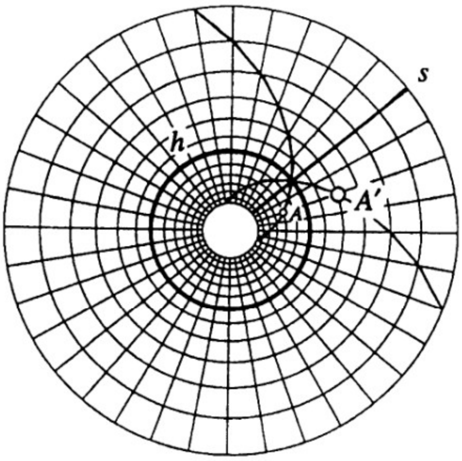

FIGURE 25
"Diagonal curves"

line of spiral shape (Figure 25). Precisely, a sequence of square-diagonals gives rise to a broken line with vertices at the grid points of the square grid. But if we subdivide the square grid ever finer and think of it provided with square-diagonals, there arises a beautifully curved spiral. Reflection of this spiral in the reference-circle h, which maps A to A', gives rise to a spiral of exactly the same kind. The two spirals are thus mirror images with respect to circle-reflection in h. Moreover, they are obviously also mirror images with respect to straight line reflection in the radial grid line s. For such a diagonal spiral the polar angle ϕ increases, for each grid step in the radial direction, by the same value α; α is the angle of intersection of two neighboring radial grid lines. If we denote by t the number of grid steps from h towards the exterior, then the relation $\phi = \alpha t$ holds, so that $t = \frac{\phi}{\alpha}$. Since $r(t) = r_0 e^{\alpha t}$, the diagonal spiral is represented in polar coordinates by the equation

$$r\left(\frac{\phi}{\alpha}\right) = r_0 e^{\phi}.$$

The diagonal spiral is thus the graph of an exponential function drawn in polar coordinates. In general a spiral with the polar representation

$$r(\phi) = r_0 e^{m\phi}$$

is called a ***logarithmic spiral,*** since the polar angle ϕ depends logarithmically on the polar distance r. Figure 26 shows a logarithmic spiral with $m = \frac{1}{5}$.

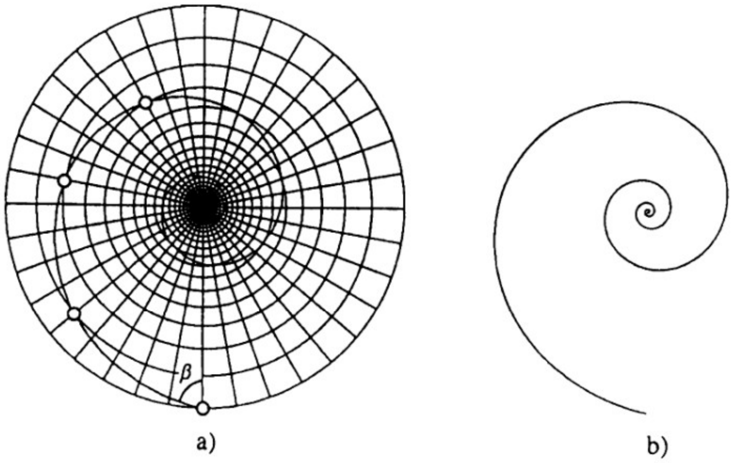

a) b)

FIGURE 26
A logarithmic spiral with $m = \frac{1}{5}$

With this spiral a unit-square in the radial direction arises from 5 unit-squares in the tangential direction; the "gradient" is thus $m = \frac{1}{5}$. In general, a logarithmic spiral with gradient m is transformed, under a circle-reflection in a reference-circle concentric with the spiral center, into a logarithmic spiral with a gradient $(-m)$. In a logarithmic spiral the angle to the radial gridlines is always the same, namely $\beta = \operatorname{arc\,cot} m$.

A logarithmic spiral thus follows a constant course β with respect to the grid-center. Such curves appear in the real world. For a few insects, for example, for bees, the lateral placing of the immovable eyes has the consequence that the insects fly towards an object in view, for example a flower, with a systematic angular error β; the flight curve is thus a logarithmic spiral. Also, in aviation and navigation, constant courses are adopted; the curves thus arising, the so-called ***loxodromes,*** are, in view of the sphericity of the earth, not logarithmic spirals, but they can be approximated by such curves near the poles.

Figure 27a shows a loxodrome, in which, for two unit-squares in the easterly direction, there occurs a unit-square in the northerly direction. It cuts each meridian at an angle $\arctan 2 \approx 63.43°$, and so follows a constant course of 63.43°. Figure 27b shows the same loxodrome as viewed from the north. In Figure 28 (from [32, p. 318]) we recognize the same loxodrome as the basis of the design.

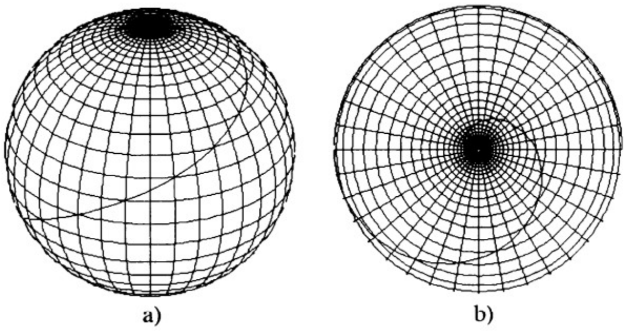
a) b)

FIGURE 27
Loxodrome with constant course

Question 2.3 How long is a logarithmic spiral from reference-circle to its innermost point?

Question 2.4 Which curves are given by the polar equations $r(\phi) = a\phi + b,\ r(\phi) = \frac{a}{\phi}$?

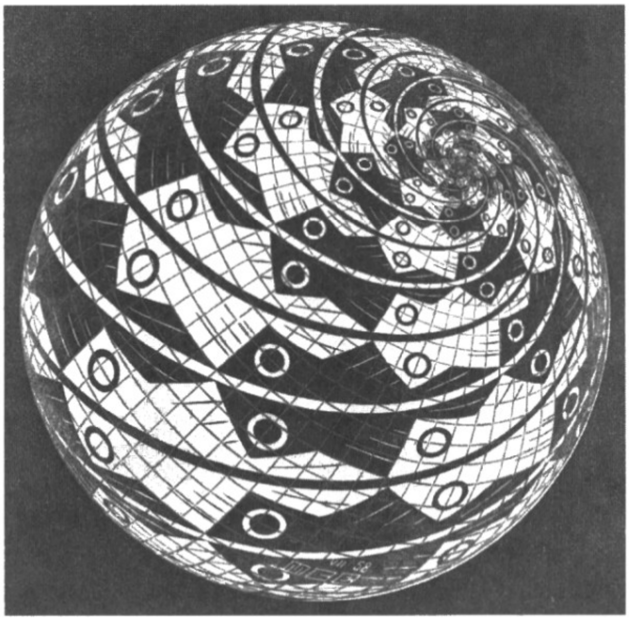

FIGURE 28
M. C. Escher: *Spherical surface with fish,* July 1958. ⓒ 1997 Cordon Art–Baarn–Holland. All rights reserved.

Question 2.5 How long is the loxodrome of Figure 27 from South Pole to North Pole?

2.5 IMAGE OF A STRAIGHT LINE

We study the image of a straight line g under a circle-reflection in a reference-circle h. To this end we choose a Cartesian coordinate system such that h is given by the equation $x^2 + y^2 = r_0^2$ and g is given by the equation $x = c$ (Figure 29a).

The line g is then given in polar coordinates by the equation $c = r\cos\phi$. Since $rr' = r_0^2$, the equation for the image figure, in polar coordinates, is

$$r' = \frac{r_0^2}{c}\cos\phi\,.$$

The points which satisfy this equation lie on the circle having the line-segment joining $M(0,0)$ and $B'(\frac{r_0^2}{c}, 0)$ as diameter (see Figure 29a). The image g' of g is thus a circle through M itself. The center M of h can, however, be interpreted as the image of the ***point at infinity*** on g. The tangent to g' at M is parallel to g.

Remark The image of a straight line g is, of course, also a circle if G intersects the reference-circle (Figure 29b). In this case g' may simply be described as the circle through M and the two fixed points F_1 and F_2.

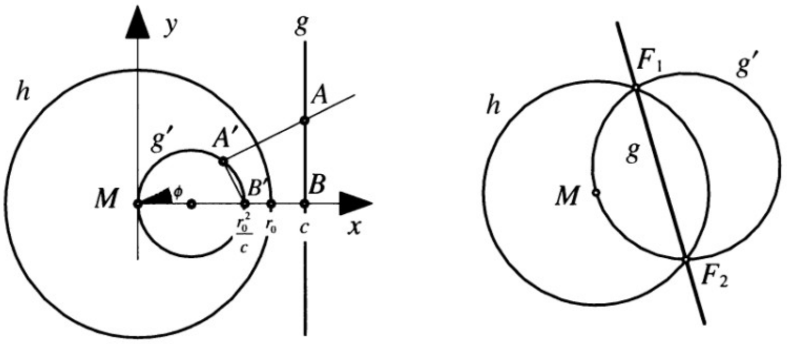

a) b)

FIGURE 29
Image of a straight line

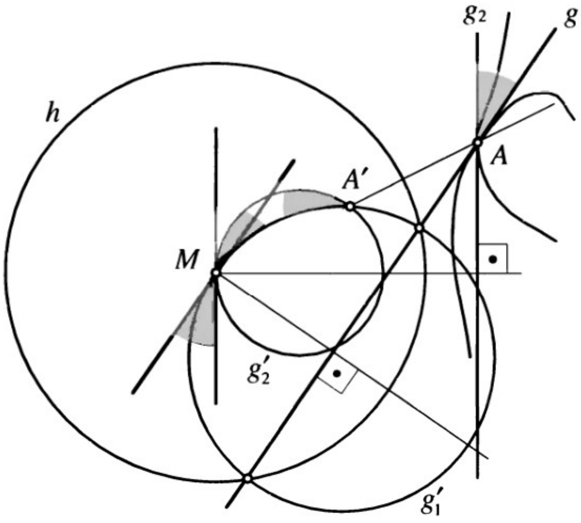

FIGURE 30
The angle remains fixed

Question 2.6 Are there exceptional straight lines whose image under a circle-reflection is not a circle?

We can now show that the angle is preserved under a circle-reflection in h. The angle of intersection of two curves is defined as the angle of intersection of their tangents at the point of intersection A (Figure 30). But this angle is equal to the angle between the image circles of the two tangents to the curves at M and is thus equal to the angle at the image point A'.

A mapping under which angles are preserved is called ***angle-preserving*** or ***conformal.*** Conformal mappings play an important role in cartography in the construction of angle-preserving maps.

Circle-reflections are conformal mappings.

In particular, right angles are preserved under conformal mappings. The image of a Cartesian square lattice is thus itself an orthogonal network, consisting of orthogonal circles through M (Figure 31).

Question 2.7 How did the gothic quatrefoil window in the center of Figure 31 arise?

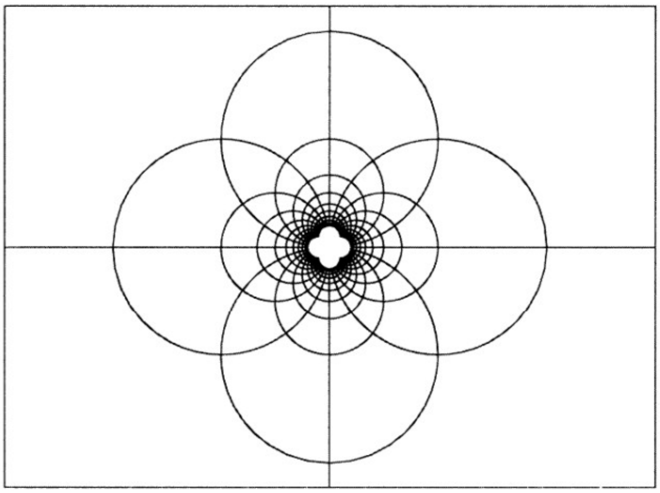

FIGURE 31
Image of a square lattice

2.6 REPRESENTATION IN CARTESIAN COORDINATES

We consider a Cartesian coordinate system with the center M of the reference-circle as origin (Figure 32). With the notations $A(x, y)$ and $A'(x', y')$, we can write $r = \sqrt{x^2 + y^2}$, $r' = \sqrt{x'^2 + y'^2}$ and $\frac{x'}{x} = \frac{r'}{r}$.

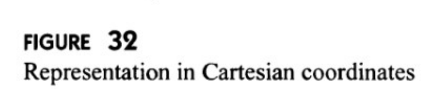

FIGURE 32
Representation in Cartesian coordinates

Since $rr' = r_0^2$, it further follows that

$$\frac{x'}{x} = \frac{r'}{r} = \frac{r_0^2}{r^2} = \frac{r_0^2}{x^2+y^2}.$$

This yields the mapping equations

$$x' = r_0^2\,\frac{x}{x^2+y^2}, \quad y' = r_0^2\,\frac{y}{x^2+y^2}.$$

Remark For the inverse mapping we have the symmetric mapping equations

$$x = r_0^2\,\frac{x'}{x'^2+y'^2}, \quad y = r_0^2\,\frac{y'}{x'^2+y'^2}.$$

2.7 IMAGE OF A CIRCLE

The image of a circle k concentric with the reference-circle h is again a circle k' concentric with h; for the radii ρ, ρ' of the circles k, k' we have $\rho\rho' = r_0^2$. A circle k through the center M of h has as image a straight line. We study now the image of a circle k which is neither concentric with h nor passes through M (Figure 33). In an appropriate Cartesian coordinate system such a circle k with center at $(u, 0)$ and radius ρ is given by the equation $(x-u)^2 + y^2 = \rho^2$. By substituting from the mapping equations for the

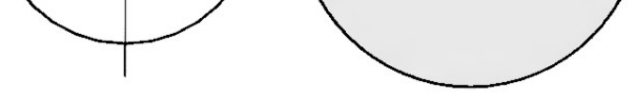

FIGURE 33
Image of a circle

inverse mapping, we obtain from this for the image figure k' the equation

$$\left(r_0^2\,\frac{x'}{x'^2+y'^2}-u\right)^2+\left(r_0^2\,\frac{y'}{x'^2+y'^2}\right)^2=\rho^2.$$

We manipulate this into the form

$$r_0^4-2ur_0^2x'+u^2(x'^2+y'^2)=\rho^2(x'^2+y'^2).$$

This is again the equation of a circle; it can be brought into the form

$$\left(x'-\frac{ur_0^2}{u^2-\rho^2}\right)^2+y'^2=\left(\frac{r_0^2\rho}{u^2-\rho^2}\right)^2$$

The image k' of the circle k is thus a circle with center $(u',0)$ and radius ρ', where

$$u'=\frac{ur_0^2}{u^2-\rho^2},\quad \rho'=\frac{r_0^2\rho}{u^2-\rho^2}.$$

Remark We have $uu'=\frac{u^2r_0^2}{u^2-\rho^2}\neq r_0^2$; this means that the center of one circle is *not* mapped onto the center of the other.

Under a straightline reflection, a circle which cuts the axis of reflection orthogonally is fixed (Figure 34a).

The same is true with a circle-reflection; a circle k which cuts the reference-circle h orthogonally is a fixed circle (Figure 34b). In this situation we have, from Pythagoras' Theorem, $u^2-\rho^2=r_0^2$. Hence

$$u'=\frac{ur_0^2}{u^2-\rho^2}=u;\quad \rho'=\frac{r_0^2\rho}{u^2-\rho^2}=\rho.$$

Remark Thus, with circle-reflections, circles and straight lines are transformed, according to the situation, into circles and/or straight lines. It is

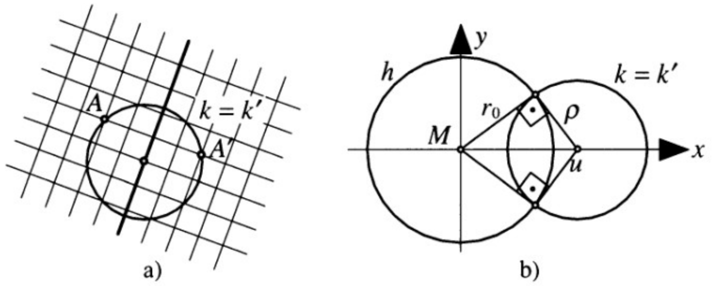

a) b)

FIGURE 34
Fixed circles

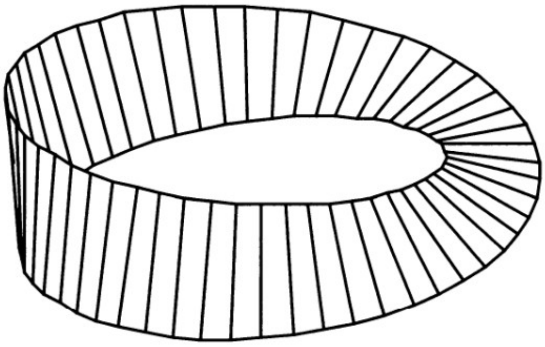

FIGURE 35
The Möbius band

therefore meaningful to subsume both ideas under one all-embracing idea. The concept of ***Möbius circle*** brings together circles and straight lines; a Möbius circle is one or the other. Under a circle-reflection the image of a Möbius circle is always a Möbius circle. The name Möbius circle reminds us of A. F. Möbius.

August Ferdinand Möbius (born 17/11/1790 in Schulpforta, died 26/9/1868 in Leipzig) was, after studies at Leipzig University and study leave in Göttingen and Halle, called in 1816 on the recommendation of Gauss to Leipzig. In 1820 he became Director of the Observatory, and in 1844 full professor for astronomy and mechanics. He contributed decisively to the education of Gymnasium (academic high school) teachers in the kingdom of Saxony and was heavily involved in the new orientation of geometry in the first half of the 19th century. He is also the discoverer of the Möbius band (Figure 35). A Möbius band is easily constructed from a strip of paper, into which one introduces a single twist before sticking the ends together. The Möbius band has some interesting topological properties. It is a "one-sided surface", that is, it has no "inside" and "outside". If we start anywhere on the apparent outside to color the band we find ourselves, as we continue the work, suddenly on the apparent inside. The ants of Figure 36 (from [32, p. 324]) pass without problem from the apparent outside to the apparent inside, without having to climb over the edge of the strip.

Question 2.8 What happens if we try to cut a Möbius band in two along the center line?

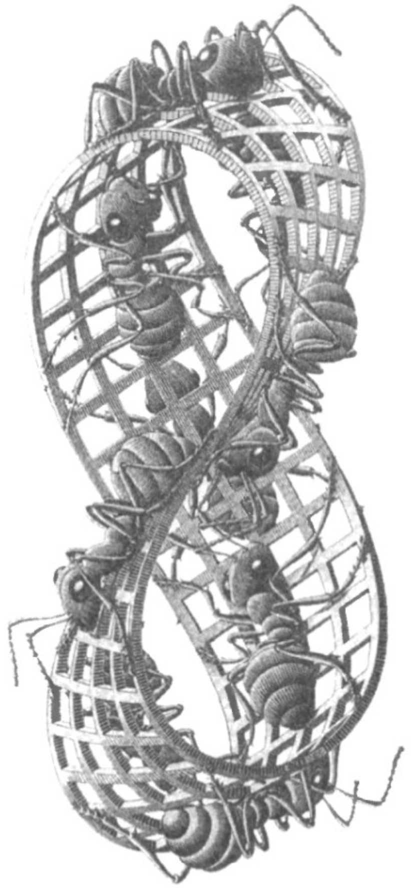

FIGURE 36
M. C. Escher: Möbius strips II (Red Forest Ants), 1963. © 1997 Cordon Art–Baarn–Holland. All rights reserved.

The concept of circle-reflection can be extended to sphere-reflection in 3-dimensional space. Then the inside of the sphere is mapped onto the outside and conversely.

Question 2.9 How does a mathematician catch a lion?

2.8 SQUARE-REFLECTION

The idea of reflecting from the inside to the outside and vice versa can be carried over to other closed figures. For the inspiration of this idea I must thank Georg Schierscher. In Figure 37 a reflection in a square of side length 2 is represented. The "reflection in a square" is defined as follows: the point A' lies on the same ray emanating from the center M of the square as the

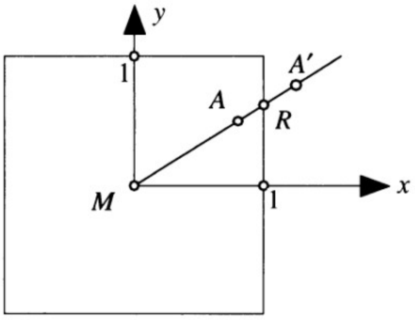

FIGURE 37
Reflection in a square

original point A. The point R, with now a variable polar distance, which we still denote by r_0, is the intersection of the ray with the square. For the polar distances r and r' of A and A', respectively, we have the relationship $rr' = r_0^2$.

Question 2.10 Express the mapping equations of this reflection in a square in Cartesian coordinates. How does the image of the square grid parallel to the axes look? What is the image of the square grid turned through 45°?

2.9 OTHER REFLECTIONS

Question 2.11 What problem presents itself if one reflects in a parabola p of the parabolic network of Figure 38 by "counting the unit squares"?
Further "near-reflections" are considered in [4].

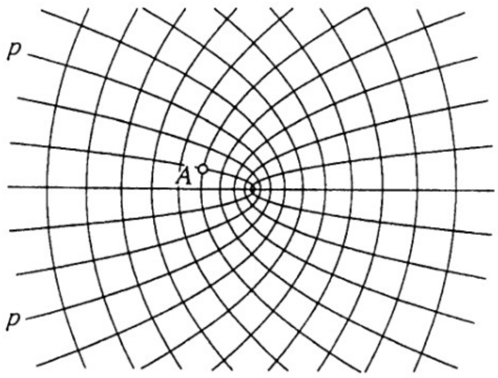

FIGURE 38
Reflection in a parabola?

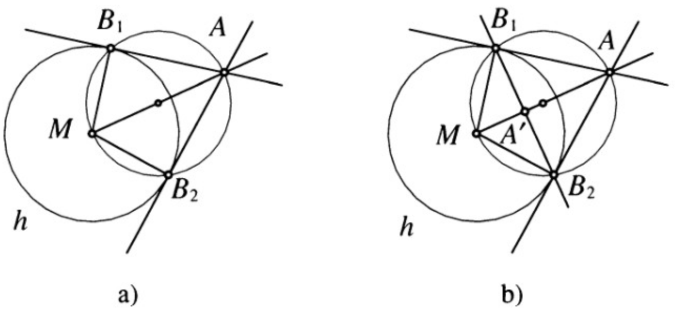

FIGURE 39
The point A lies outside h

Answers to Questions

Answer 2.1 The construction proceeds in the reverse sequence according to Figure 39.

Answer 2.2 Precisely the points of the reference-circle h are the fixed points.

Answer 2.3 A path-length ds on a logarithmic spiral has, in view of the constant course, a radial variation given by $dr = ds \cos \beta$. Thus the length of the logarithmic spiral from the reference-circle h with radius r_0 to its innermost point is given by $s_0 = \frac{r_0}{\cos \beta}$.

Answer 2.4 A linear function $r(\phi) = a\phi + b$ in polar coordinates yields an ***archimedean spiral*** (Figure 40a). For an archimedean spiral the radial difference between two revolutions is always the same, namely, $r(\phi+2\pi)-r(\phi) = 2a\pi$. Such spirals arise from the rolling up of material of constant thickness (rolls of carpet, rolls of film, rolls of toilet paper). The mirror-image of an archimedean spiral in a circle-reflection is no longer an archimedean spiral.

The function $r(\phi) = \frac{a}{\phi}$ yields a ***hyperbolic spiral,*** which, in view of its appearance, is also known as a ***bishop's staff*** or ***crosier*** (Figure 40b). We see such a hyperbolic spiral when we look at a screw in the direction of its axis, for example, by looking at the central shaft of a winding staircase.

Answer 2.5 According to the Theorem of Pythagoras, the arc length for a unit length in the northerly direction is $\sqrt{5}$ times this unit length (Figure 41). The entire loxodrome length is thus $\sqrt{5}$ times the meridian length. For the loxodrome with constant course β the length, from Pole to Pole, is $\frac{1}{\cos \beta}$ times the meridian length. Thus it is also clear that, in general, the loxodrome does not constitute the shortest path on the spherical

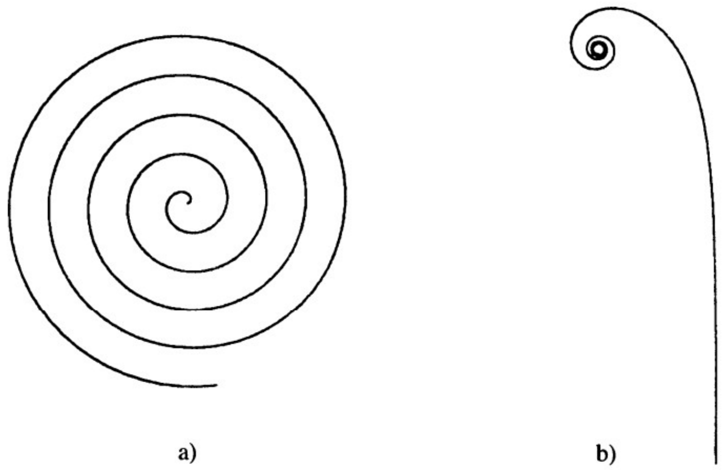

FIGURE 40
Archimedean and hyperbolic spirals

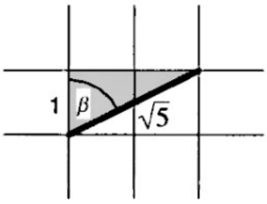
1
β
$\sqrt{5}$

FIGURE 41
Arc length of the loxodrome

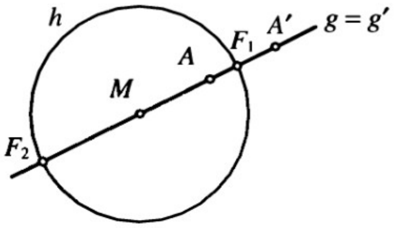
h
A'
$g = g'$
F_1
A
M
F_2

FIGURE 42
Image of a straight line through the center

surface. The shortest path is given by arcs of great circles. Nevertheless on sea journeys the loxodrome is preferred on practical grounds (constant course) when the difference in length is not too great.

Answer 2.6 A straight line g through the center M is a fixed line, though not a line of fixed points. It has only the two diametrical fixed points F_1 and F_2. As to the other points, the inner and outer points are exchanged; the point M is mapped to the point at infinity, and conversely (Figure 42). The image straight line g' can be viewed as a circle of infinite radius.

Answer 2.7 Figure 31 is the image of a bounded Cartesian square network; the region runs in both dimensions from -12 to $+12$. The gothic quatrefoil window is the image of the boundary square with $x = \pm 12$ and $y = \pm 12$ under reflection in a circle.

Answer 2.8 The Möbius band cannot literally be cut in two along its center line. If attempted, there results a *single* band, twice as long and with a double twist. Try it!

Answer 2.9 He (or she) sits in the center of an empty lion cage and comes out through a reflection in a sphere. Then he (or she) is outside and the lion is inside. Indeed, all lions are then in the cage. Moreover, the mathematician, and the rest of humanity, no longer have their hearts in the right place. (Translator's note: the lion cage is assumed spherical.)

Answer 2.10 For the mapping equations we need a separation of cases. If $|y| \leq |x|$ we have

$$x' = \frac{x}{x^2}; \ y' = \frac{y}{x^2},$$

and if $|y| > |x|$ we have

$$x' = \frac{x}{y^2}; \ y' = \frac{y}{y^2}.$$

Figure 43a shows the image of the gridline $y = \text{const}$. Instead of the circles of

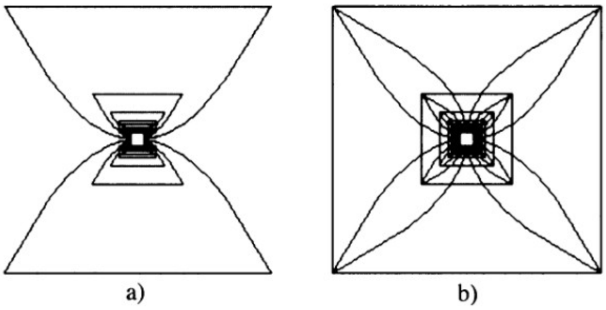
a) b)

FIGURE 43
Image of the square network parallel to the axes

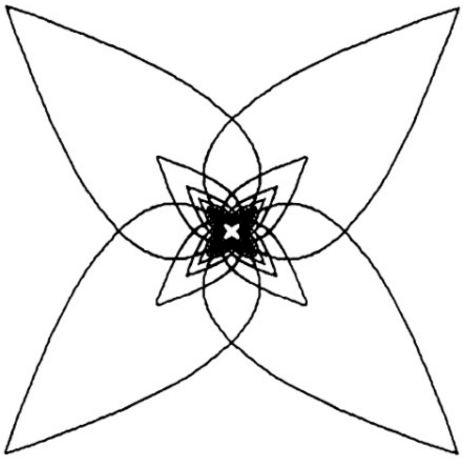

FIGURE 44
Image of the square network turned through 45°

Figure 31 we obtain curves which are put together from quadratic parabolae and line segments. Figure 43b shows the image of the entire square grid parallel to the axes. Figure 44 shows the image of the square grid turned through 45°.

Answer 2.11 The mirror image is not uniquely defined; there is an "upper" and a "lower" solution (Figure 45).

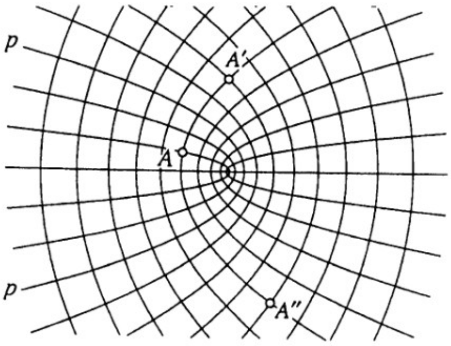

FIGURE 45
The mirror image is not uniquely defined.

data 06.05.2020

Extending work of Rosser and Schoenfeld, various authors have proved related results.

Krawczyk proved the following, $\pi(kx) < k\pi(x)$ for $k \geq e, x \geq 67$.

For $\epsilon > 0, x \geq 17, y \geq 17, x+y \geq 1 + e^{4(1+1/\epsilon)}$, it has been show that

$$\pi(x+y) < (1+\epsilon)(\pi(x)+\pi(y)).$$

There is an interesting connection between Conjecture A ($\pi(x+y) \leq \pi(x)+\pi(y)$, $x, y \geq 2$) and the generalization of the twin prime conjecture that follows it.

E7.3) Symmetry ... djvu

Pag. 11(24) Chapter 2 Inside and Outside

2.1 Reflecting in a Circle". (inversion)

Important:

Pag. 14(27) 2.2 Composition of two inve-

data 06.05.2020

S T [✓ Q] Q S S D

reflections. Figure 23

Fig. 24 (Pág. 15(28)).

Direct construction of the image point.

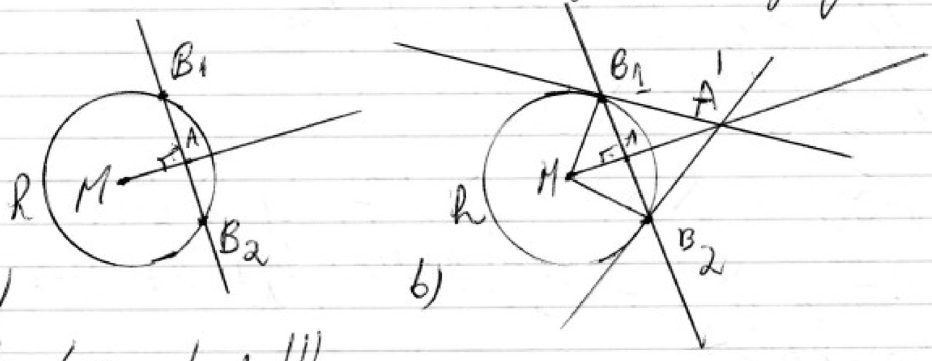

Obs: Livro todo!!!

E7.4) Geometric inequalities (1975...)..djvu

Pág. 65 (71) Chapter Three
The Reflection Principle

Figure 3.1.

Obs. Livro todo!!!

E7.5) Proof and other Dilemmas_ Mathematics and Philosophy... djvu

E7.6) Carl Friedrich Gauss — Titan of Science..! djvu

Pág. 163 (211) Chapter 14 Surface Theory

1)*+,)))-) ' ! - 0 >
! 6 A ** 6 : !H8

1 ')*+,)))-) ' < ! . > /
%2 ** !H8
("' 6 " / A

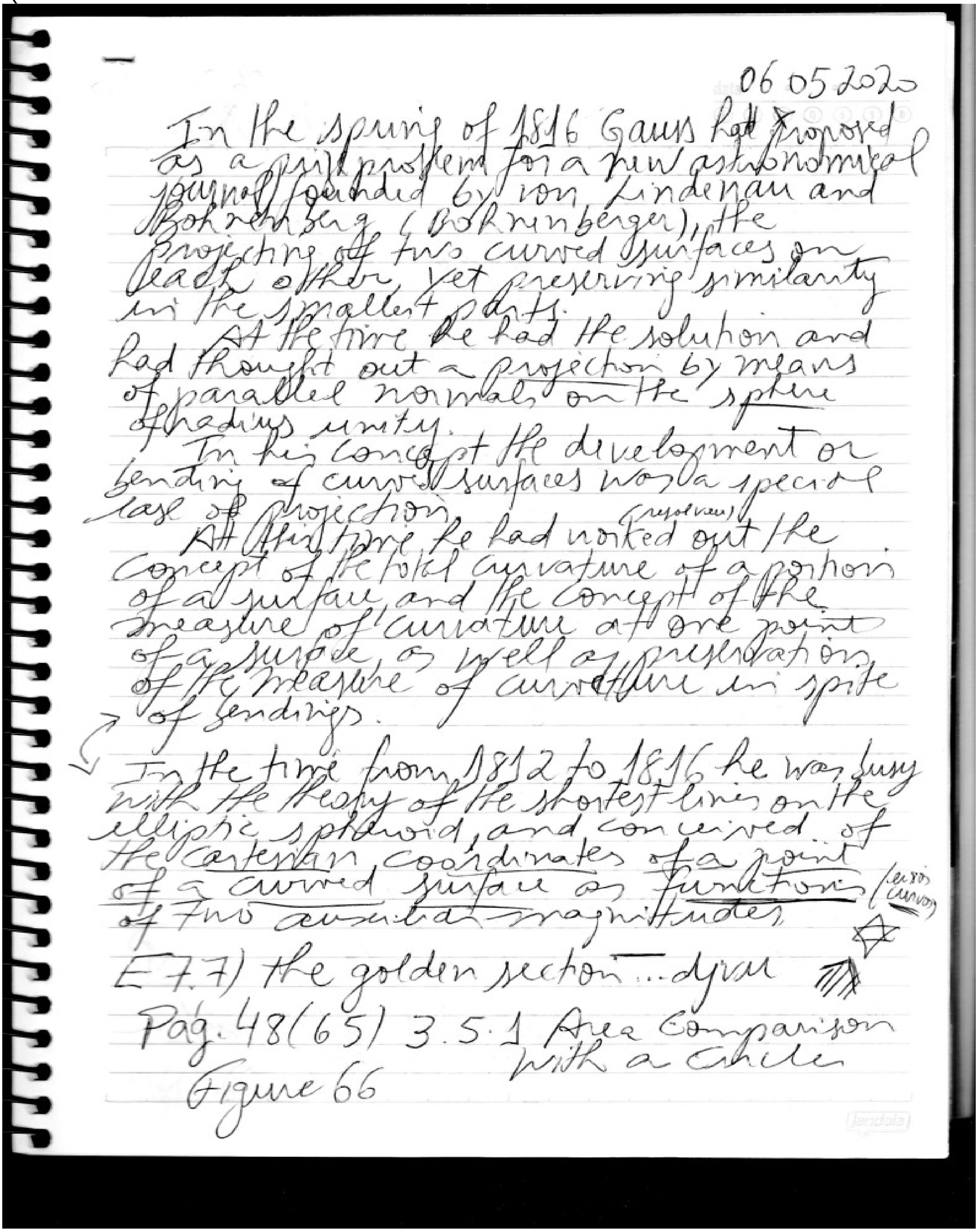

06 05 2020

In the spring of 1816 Gauss had proposed as a prize problem for a new astronomical journal founded by von Lindenau and Bohnenberg (Bohnenberger), the projecting of two curved surfaces on each other, yet preserving similarity in the smallest parts.

At the time he had the solution and had thought out a projection by means of parallel normals on the sphere of radius unity.

In his concept the development or bending of curved surfaces was a special case of projection.

(resolveu)
At this time he had worked out the concept of the total curvature of a portion of a surface and the concept of the measure of curvature at one point of a surface, as well as preservation of the measure of curvature in spite of bendings.

In the time from 1812 to 1816 he was busy with the theory of the shortest lines on the elliptic spheroid, and conceived of the cartesian coordinates of a point of a curved surface as functions of two auxiliar magnitudes. (eixos curvos)

E7.7) the golden section... dyvu

Pag. 48(65) 3.5.1 Area Comparison with a Circle

Figure 66

06.05.2020

The ellipse and the circle should have the same area.

Pág. 49 (66) 3.5.2 Geometry in the Music Cassette.

Pág. 52 (69) 3.5.3 the ellipse in a Square Lattice.

Pág. 52 (69) 3.6 Golden Trigonometry.

$$\sin 18^\circ = \frac{\rho}{2} = \frac{\tau - 1}{2}$$

$$\cos 18^\circ = \sqrt{1 - \frac{\rho^2}{4}} = \frac{\sqrt{2+\tau}}{2}$$

$$\tan 18^\circ = \frac{\tau - 1}{\sqrt{2+\tau}}$$

Pág. 75 (92) 5.3 Power of $1+\sqrt{2}$

E 7.8) Circles, a mathematical view... dyvu

Obs.: Livro todo !!!

Pág. 01 (39) Chapter 1
The nine-point circle.

Pág. 35 (73) 7. Circles cutting at a given angle.

1 :)*+,)))-) ' > 8
$$ 2 * * !H8
-7 5 8 ! 999
("$ 6 / 6

CHAPTER I

The properties of circles discussed in this chapter are those which have the habit of appearing in other branches of mathematics.

1. The nine-point circle

This circle is the first really exciting one to appear in any course on elementary geometry. It is a circle which is found to pass through nine points intimately connected with a given triangle *ABC*. Three proofs are given. Each has its own peculiar merits.

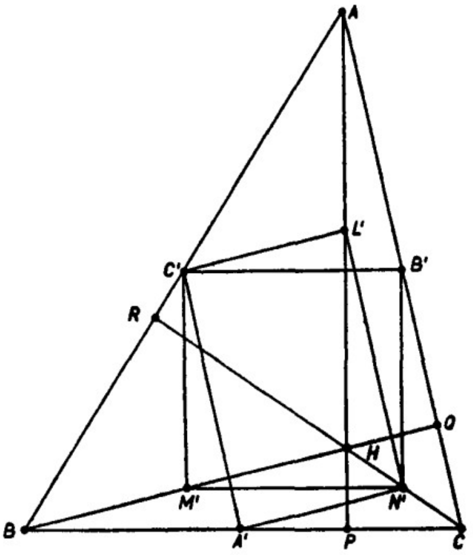

FIG. 1

(i) *H* is the orthocentre of triangle *ABC*, the intersection of the altitudes *AP*, *BQ*, *CR*. The midpoints of *BC*, *CA* and *AB* are *A'*, *B'* and *C'* respectively. *L'*, *M'*, *N'* are the midpoints of *HA*, *HB*, *HC*. We show that a circle passes through the nine points *A'*, *B'*, *C'*, *L'*, *M'*, *N'*, *P*, *Q*, *R*. We use the theorem that the line joining the midpoints of two sides of a triangle is parallel to the third side.

Both *B'C'* and *N'M'* are ‖ to *BC*, and both *B'N'* and *C'M'* are ‖ to *AH*. Hence *B'C'M'N'* is a rectangle. Similarly *C'A'N'L'* is a rectangle.

Hence $A'L'$, $B'M'$ and $C'N'$ are three diameters of one circle. Since $A'P$ and $L'P$ are $\perp$, this circle passes through P. Similarly it passes through Q and R.

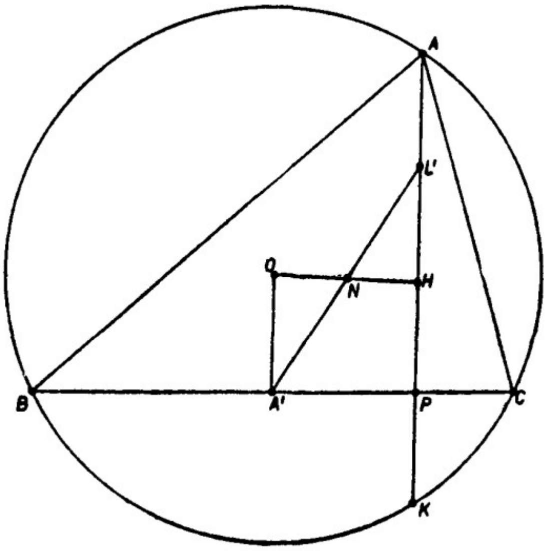

FIG. 2

(ii) Let O be the circumcentre of ABC and R the circumradius. We know that $AH = 2OA'$. Since L' bisects AH, $L'H = OA'$, and therefore OH and $L'A'$ bisect each other, in N, say.

Now $NL' = NA' = NP$ (since $\angle APB = 90°$). We also know that $HP = PK$, and since $ON = NH$, we have

$$NP = \tfrac{1}{2}OK = \tfrac{1}{2}R.$$

Therefore the circle centre N and radius $\frac{1}{2}R$ passes through P, A', L'. Similarly it passes through Q, B', M' and R, C', N'.

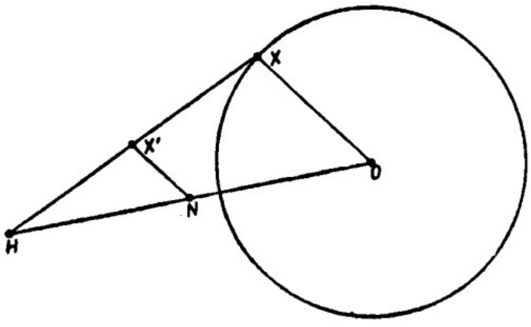

FIG. 3

(iii) The third proof uses a lemma from the theory of similar figures. If X is a point on a given circle of centre O and radius R, and H is any given fixed point, the locus of X', the midpoint of HX, is a circle centre N, where N bisects HO, and radius $\frac{1}{2}R$. The proof is clear.

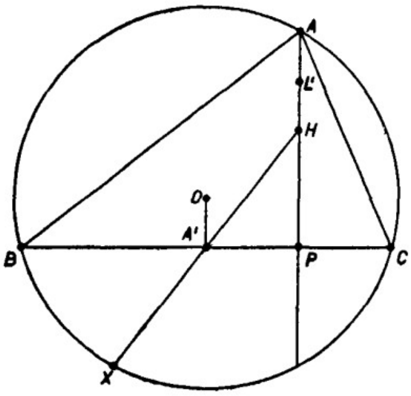

Fig. 4

Applying this lemma to the case of the circumcircle of ABC, H being the orthocentre, we see that a circle centre N and radius $\frac{1}{2}R$ passes through P, Q, R, L', M', N'. We must now show that this circle also passes through A', B', C'. It is sufficient to show that if HX contains A', then HX is bisected at A'.

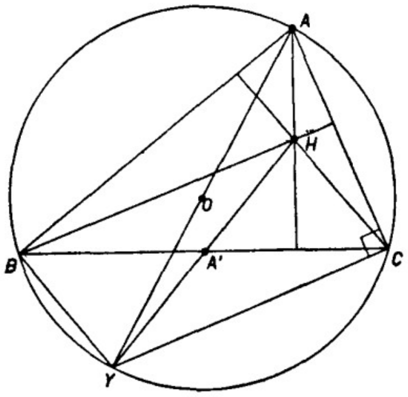

Fig. 5

Let AO meet the circumcircle in Y. Then $\angle ACY = 90°$, so that $YC \parallel BH$. Since also $\angle ABY = 90°$, $YB \parallel CH$. Therefore $YBHC$ is a parallelogram, and the diagonals YH and BC

bisect each other. Hence Y is the same point as X in the previous diagram, and the theorem is proved.

The methods used above are elementary, and at least one of the three proofs will be familiar to most readers of this book. In chapter II we shall need to assume some theorems in inversion and coaxal systems of circles. As fashions in mathematical teaching change, it cannot be assumed that all readers will be familiar with these theories, and a brief (but adequate) outline of each will now be given.

2. Inversion

This is a one-to-one transformation of the points of the plane by means of a given circle, which we shall call Σ, of radius k and centre O. To obtain the transform X' of any given point X, or the *inverse* of X in the circle Σ, we join X to O, and find X' on OX such that $OX.OX' = k^2$. The point O itself is excluded from the points of the plane which may be transformed. It is clear that

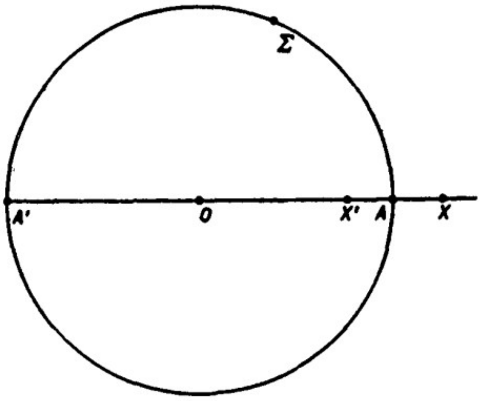

FIG. 6

(i) X is the inverse of X';
(ii) points on Σ transform into themselves;
(iii) if A, A' are the ends of the diameter of Σ through X, then X, X' are harmonic conjugates with respect to A, A'.

All circles through X and X' cut Σ orthogonally, since the square of the tangent from O to such a circle $= OX.OX' = k^2$.

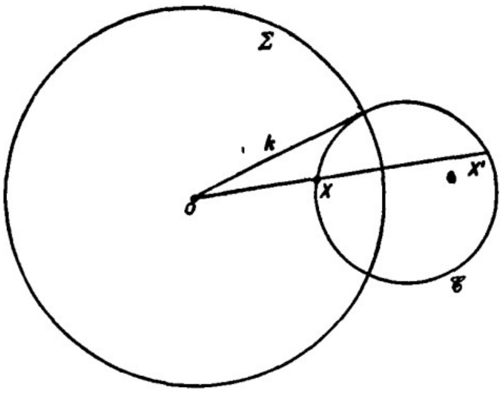

FIG. 7

As a corollary to this remark, we see that if $\mathscr{C}$ is any circle orthogonal to Σ, all points on $\mathscr{C}$ invert into points on $\mathscr{C}$, for $OX.OX' = k^2$. It is sometimes necessary to consider the process of inversion when Σ, the circle of inversion, is a line. If we keep A fixed, and let A' move off to infinity, we see that X' moves towards the geometrical image of X in the resulting line.

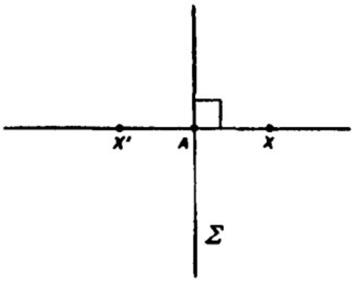

FIG. 8

For future application we now give a construction, using only a

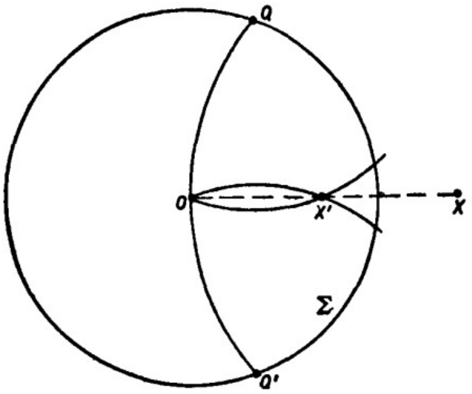

FIG. 9

pair of compasses, by which we can find the inverse of a point X in a circle Σ of given centre O. Let the circle centre X and radius XO cut Σ in Q and Q', and let O and X' be the intersections of the two circles with centres Q and Q' and radii QO and $Q'O$ respectively. Then X' is the inverse of X in Σ.

For QOX and $X'OQ$ are isosceles triangles with a common base angle. They are therefore similar, and $OX : OQ = OQ : OX'$. Hence $OX . OX' = OQ^2$.

This construction lies at the basis of the discussion on *compass geometry* in §11.

We are interested in the locus of inverse points X' as X describes a given curve $\mathscr{C}$. This locus is called the *inverse of the curve* $\mathscr{C}$ with respect to Σ. When $\mathscr{C}$ is a circle orthogonal to Σ, we have seen that the inverse is $\mathscr{C}$ itself. The general theorem is:

The inverse of a circle is a straight line or a circle.

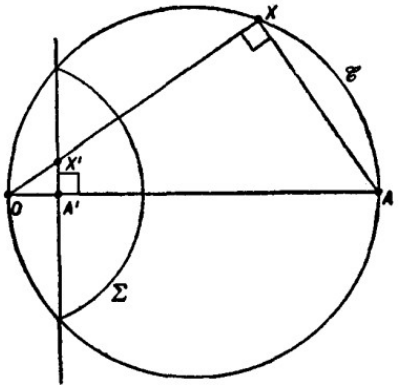

FIG. 10

(i) The centre of inversion O is on $\mathscr{C}$:

Let OA be the diameter of $\mathscr{C}$ through O, and let A' be the inverse of A. Then $OA . OA' = k^2$. If X is any point on $\mathscr{C}$, and X' its inverse,

$$OX . OX' = k^2 = OA . OA'.$$

Hence A, X, X', A' are concyclic points, and since $\angle AXO = 90°$, $\angle X'A'A = 90°$. Therefore X' describes the line through $A' \perp$ to OA.

We note that this line is $\parallel$ to the tangent to $\mathscr{C}$ at O. It naturally passes through the intersections of $\mathscr{C}$ and Σ, since points on Σ invert into themselves.

Corollary. The inverse of a straight line is a circle through the centre of inversion.

(ii) The centre of inversion O is not on $\mathscr{C}$. Let C be the centre of $\mathscr{C}$. Since

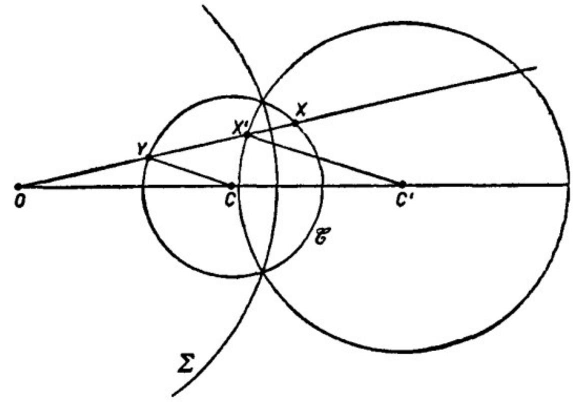

FIG. 11

$$OX.OX' = k^2,$$

and $$OX.OY = t^2,$$

where t is the tangent[†] from O to $\mathscr{C}$, by division we have

$$\frac{OX'}{OY} = \frac{k^2}{t^2}.$$

Since this ratio is constant, it follows that if C' is a point on OC such that $C'X' \parallel CY$, then C' is a fixed point, and $C'X'$ is constant. Hence X' moves on a circle centre C'.

Note that C' is not necessarily the inverse of C. We shall soon find what point the centre of $\mathscr{C}$ does invert into.

We need only one further theorem from the theory of inversion:

The angle of intersection of two circles is unaltered by inversion.

In fact a more general theorem is true, but this one will suffice for our subsequent needs.

[†] We shall see later that t^2 may be negative. Then O is inside $\mathscr{C}$.

(i) If both circles $\mathscr{C}$ and $\mathscr{D}$ pass through the centre O of inversion, they transform into straight lines parallel to the tangents to $\mathscr{C}$ and $\mathscr{D}$ at O, so that the theorem is evident.

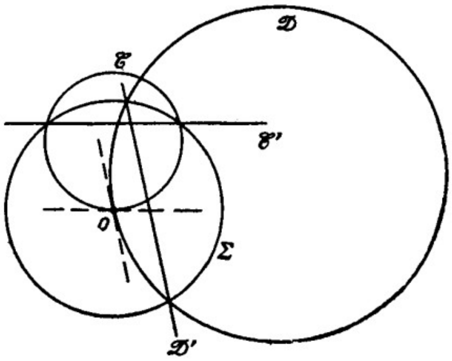

FIG. 12

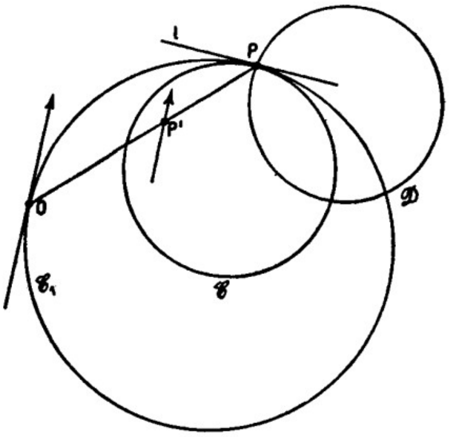

FIG. 13

(ii) If at least one of the circles $\mathscr{C}$, $\mathscr{D}$ does not pass through O, let P be a point of intersection of the two circles, and let l, m be the tangents at P to $\mathscr{C}$, $\mathscr{D}$ respectively. We may draw a circle $\mathscr{C}_1$ through O which touches l at P, and a circle $\mathscr{D}_1$ through O which touches m at P. Then the angle between $\mathscr{C}_1$ and $\mathscr{D}_1$ is the same as that between $\mathscr{C}$ and $\mathscr{D}$.

Now if P' is the inverse of P, the tangent at P' to the inverse of $\mathscr{C}$ is the inverse of $\mathscr{C}_1$. Hence the angle between the circles inverse to $\mathscr{C}$ and to $\mathscr{D}$ is equal to the angle between $\mathscr{C}_1$ and $\mathscr{D}_1$ (by (i) above), and this is equal to the angle between $\mathscr{C}$ and $\mathscr{D}$.

With the help of this theorem we can now find the inverse of the centre of a circle $\mathscr{C}$. All lines through the centre C cut $\mathscr{C}$ orthogonally. These lines invert into circles through the centre of inversion O and the inverse of C. Since all circles through O and the inverse of C cut the inverse of $\mathscr{C}$ orthogonally, the inverse of C must be *the inverse of* O *in the inverse of* $\mathscr{C}$.

If O lies on $\mathscr{C}$, so that the inverse of $\mathscr{C}$ is a straight line, the centre of $\mathscr{C}$ inverts into the geometrical image of O in this line.

We now give some applications of the theory of inversion.

3. Feuerbach's theorem

The nine-point circle of a triangle touches the incircle and excircles of the triangle.

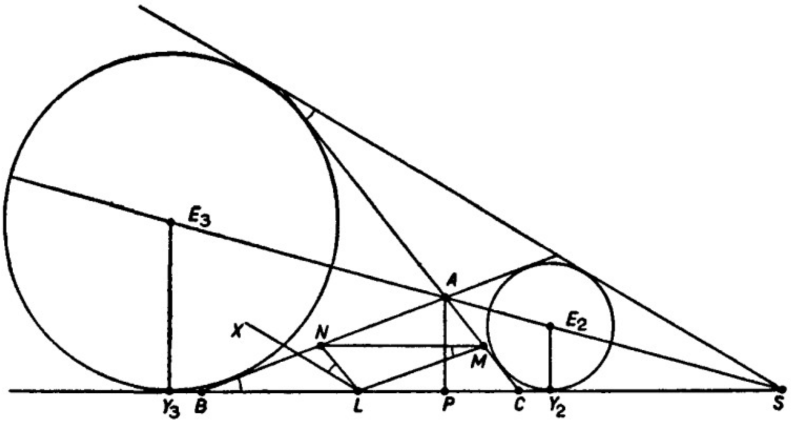

FIG. 14

We consider the excircles opposite B and C. Let their centres be E_2 and E_3, and their points of contact with BC Y_2 and Y_3. The points A, S divide E_2E_3 internally and externally in the same ratio, that of the radii of the two excircles. Dropping perpendiculars on BC, it follows that P and S also divide Y_2Y_3 internally and externally in the same ratio. That is, P and S are harmonic conjugates with respect to Y_2 and Y_3.

Let L, M, N be midpoints of BC, CA, AB. We note that $BY_2 = \frac{1}{2}(BC + CA + AB) = CY_3$, so that L is also the midpoint of Y_2Y_3. From the harmonic relation just proved we see that

$$LP.LS = LY_2{}^2 = LY_3{}^2.$$

Now let LX be the tangent at L to the circle LMN, where X is on the same side of BC as N. Then

$$\angle XLN = \angle NML = \angle ABC.$$

Hence the angle between LX and $CA = \angle ABC$. It follows that LX is $\parallel$ to the fourth common tangent of the two excircles, the common tangent through S other than BC.

Now since $LP.LS = LY_2^2 = LY_3^2$, the nine-point circle LMN is the inverse of this fourth common tangent with respect to the circle centre L and radius LY_2. For this nine-point circle contains P, and inverts into the line through S parallel to the tangent to the circle at L.

Also since the circle of inversion cuts both excircles orthogonally, each excircle inverts into itself. We have seen that the inverse of the nine-point circle touches these excircles. It follows that the nine-point circle touches the excircles centres E_2, E_3. In precisely the same way we can show that the nine-point circle touches the incircle and the excircle opposite A.

4. Extension of Ptolemy's theorem

If A, B, C, D are any four points in a plane, then

$$AB.CD + AD.BC > AC.BD,$$

unless A, B, C, D lie, in the order $ABCD$, on a circle or a straight line. In the latter case, the inequality becomes an equality.

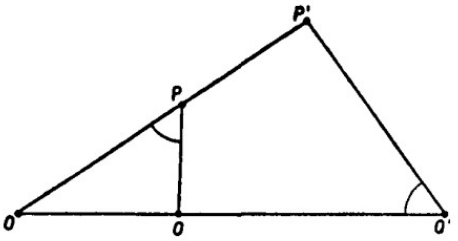

FIG. 15

This is proved by noting the influence of inversion on the length of a segment. If the points P, Q invert into the points P', Q', with O as centre of inversion, then OPQ and $OQ'P'$ are similar triangles. Hence

$$\frac{P'Q'}{PQ} = \frac{OP'}{OQ} = \frac{OP.OP'}{OP.OQ} = \frac{k^2}{OP.OQ},$$

so that

$$P'Q' = \frac{k^2\,PQ}{OP.OQ}.$$

Now, given the four points A, B, C, D, we invert with respect to A. Let B', C', D' be the respective inverses of B, C, D. Then $B'C' + C'D' > B'D'$ unless C' lies on the line $B'D'$ between B' and D′. In the latter case we have

$$B'C' + C'D' = B'D'.$$

The first inequality becomes

$$\frac{BC}{AB.AC} + \frac{CD}{AC.AD} > \frac{BD}{AB.AD},$$

or

$$AB.CD + AD.BC > AC.BD,$$

unless C' lies on the line $B'D'$ between B' and D', in which case A, B, C, D lie on a circle in the order $ABCD$, or on a straight line.

5. Fermat's problem

A, B, C are any three points in a plane. To find a point P such that $PA + PB + PC$ shall be least.

Let B and C be the acute angles of the triangle ABC. On BC, and away from A, describe an equilateral triangle BCD. Then by the extension of Ptolemy's theorem, unless P lies on the circle BCD so that the order of the points is $BPCD$, we have

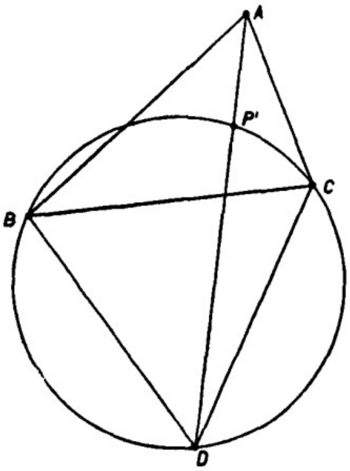

FIG. 16

$$BP.CD + PC.DB > PD.BC,$$

or

$$PB + PC > PD,$$

since $CD = DB = BC$. Therefore

$$PA + PB + PC > PA + PD.$$

Now, unless P lies on AD, we have $PA + PD > AD$. Hence, unless P is at P' (the other intersection of AD with the circle BCD), we have $PA + PB + PC > AD$.

But if P is at P', both the above inequalities become equalities; so that $P'A + P'B + P'C = AD$.

Hence $P'A + P'B + P'C < PA + PB + PC$.

Therefore P' is the required point.

If $\angle BAC = 120°$, $A = P'$, and A is the required point.

If $\angle BAC > 120°$, A is still the required point.

6. The centres of similitude of two circles

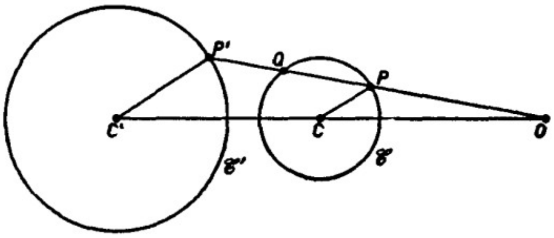

FIG. 17

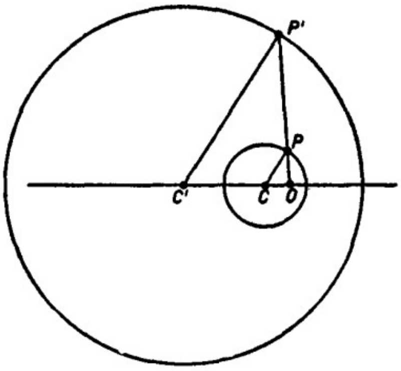

FIG. 18

These points have already appeared in some of our proofs. We now define them.

External centre of similitude. Let $\mathscr{C}$ and $\mathscr{C}'$ be two circles of unequal radii, with centres C and C'. Let CP and $C'P'$ be parallel radii drawn in the *same* sense. Then it is clear that $P'P$ cuts $C'C$ in a fixed point O, which divides $C'C$ externally in the ratio of the radii of the two circles. This is the external centre of similitude. If the ratio of the radii $= k$, then $OP' : OP = k$. We also have $OP \, . \, OQ =$ constant. Therefore

$$OP' \, . \, OQ = \text{constant}.$$

It follows that with centre of inversion O, and a suitable radius of inversion, $\mathscr{C}$ can be inverted into $\mathscr{C}'$.

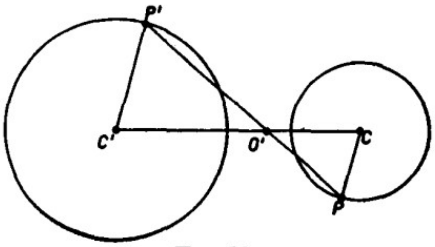

FIG. 19

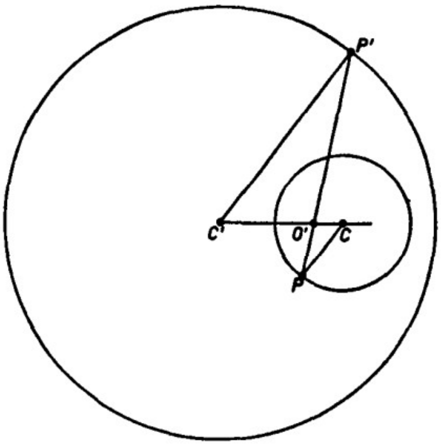

FIG. 20

Internal centre of similitude. Here the radii of the circles $\mathscr{C}$ and $\mathscr{C}'$ may be equal. CP and $C'P'$ are parallel radii in opposite senses. $P'P$ cuts $C'C$ in a fixed point O', the internal centre of similitude, which divides $C'C$ internally in the ratio of the radii.

Direct common tangents of $\mathscr{C}$ and $\mathscr{C}'$, if they exist, cut the line of centres in O, and transverse common tangents, if they exist, cut the line of centres in O'.

If $\mathscr{C}$ and $\mathscr{C}'$ are concentric, the common centre is both the internal and external centre of similitude.

If $\mathscr{C}$ and $\mathscr{C}'$ touch externally, that is, if they lie on opposite sides of the tangent at the point of contact, this point is the *internal* centre of similitude. If the circles touch internally, the point of contact is the *external* centre.

If we have three circles, no two of which have equal radii, and the centres of which form a triangle, the six centres of similitude lie in threes on four straight lines. This is easily proved, using the theorem of Menelaus.

7. Coaxal systems of circles

These are best introduced by using the methods of coordinate geometry.

The equation of any line passing through the origin of coordinates in the (x,y)-plane may be written as

$$\lambda x + \mu y = 0.$$

These lines are said to form a *pencil* of lines, and the equation of the system depends *linearly* on two distinct lines of the pencil, given by $x = 0$ and $y = 0$.

Similarly, if $u \equiv ax + by + c = 0$ and $v \equiv a'x + b'y + c' = 0$ are any two distinct lines, the line

$$\lambda u + \mu v = 0$$

represents, for suitable λ, μ, any line through the intersection of $u = 0$ and $v = 0$.

When we consider two distinct circles $\mathscr{C}$ and $\mathscr{C}'$, given by the equations

$$C \equiv x^2 + y^2 + 2gx + 2fy + c = 0,$$
$$C' \equiv x^2 + y^2 + 2g'x + 2f'y + c' = 0,$$

the system $\lambda C + \mu C' = 0$, derived linearly from them, is a system of circles, since it may also be written as

$$x^2 + y^2 + \frac{2x(\lambda g + \mu g')}{\lambda + \mu} + \frac{2y(\lambda f + \mu f')}{\lambda + \mu} + \frac{\lambda c + \mu c'}{\lambda + \mu} = 0.$$

If (x',y') is a point of intersection of $\mathscr{C}$ and $\mathscr{C}'$,

$$C(x',y') = C'(x',y') = 0,$$

that is, the point satisfies the equations of both circles. Since also, for all values of λ, μ,

$$\lambda C(x',y') + \mu C'(x',y') = 0,$$

this same point lies on all circles of the system $\lambda C + \mu C' = 0$. Such a system of circles therefore passes through the points of intersection (if any) of $\mathscr{C}$ and $\mathscr{C}'$. The system may be described as a *pencil* of circles, but for reasons which will soon become clear, we give it the more normal title of a *coaxal system* of circles.

A geometrical meaning may be given to the ratio of the parameters $\lambda : \mu$. Let $P\ (x',y')$ be any point in the plane, and let any line through P cut $\mathscr{C}$ in Q and R. Then it is elementary that the product

$$PQ.PR = x'^2 + y'^2 + 2gx' + 2fy' + c.$$

This constant, which is > 0 if P is outside the circle and equal to the square of the tangent from P to $\mathscr{C}$, is called, in all cases, the *power* of P with respect to $\mathscr{C}$. If P is inside $\mathscr{C}$, the power is negative, and it is zero, of course, if P is on $\mathscr{C}$. Since

$$\lambda C + \mu C' = 0$$

may be written as

$$\frac{\lambda}{\mu} = \frac{-C'}{C},$$

we see that *the locus of a point which moves so that the ratio of its powers with respect to two given circles $\mathscr{C}$ and $\mathscr{C}'$ is a constant is a circle of the coaxal system determined by $\mathscr{C}$ and $\mathscr{C}'$.*

A particular and important member of the system is obtained by taking the ratio of the powers to be unity. We then obtain

$$C - C' \equiv 2(g - g')x + 2(f - f')y + c - c' = 0,$$

a straight line, called the *radical axis* of the system. If $\mathscr{C}$ and $\mathscr{C}'$ intersect, this line is the common chord of the two circles, and it is evident that points on the common chord of two circles do have equal powers with respect to each circle. But the radical axis exists even if the circles do not intersect. It is easily verified that the radical axis of two circles is perpendicular to the line of centres.

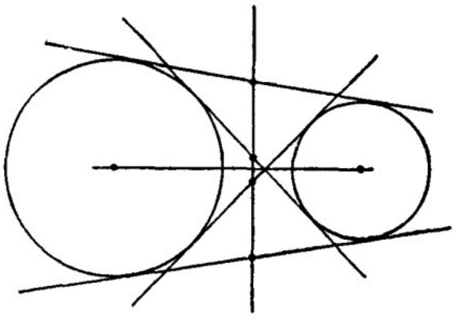

FIG. 21

We deduce that the midpoints of the common tangents of two circles lie on a line which is perpendicular to the line of centres.

We obtain the same coaxal system of circles if we take any two distinct members of the system $\lambda C + \mu C' = 0$, say

$$C_1 \equiv \lambda_1 C + \mu_1 C' = 0,$$

and

$$C_2 \equiv \lambda_2 C + \mu_2 C' = 0,$$

where $\lambda_1/\mu_1 \neq \lambda_2/\mu_2$, and write down the system of circles which depend linearly on C_1 and C_2. This should be verified. Another theorem of importance is that any two circles of a coaxal system have the same radical axis. This explains the term ***coaxal system.*** We can verify this directly. The radical axis of C_1 and C_2 is

$$2x\left[\frac{\lambda_1 g + \mu_1 g'}{\lambda_1 + \mu_1} - \frac{\lambda_2 g + \mu_2 g'}{\lambda_2 + \mu_2}\right] + 2y\left[\frac{\lambda_1 f + \mu_1 f'}{\lambda_1 + \mu_1} - \frac{\lambda_2 f + \mu_2 f'}{\lambda_2 + \mu_2}\right] + \left[\frac{\lambda_1 c + \mu_1 c'}{\lambda_1 + \mu_1} - \frac{\lambda_2 c + \mu_2 c'}{\lambda_2 + \mu_2}\right] = 0.$$

This easily reduces to

$$(\lambda_1\mu_2 - \lambda_2\mu_1)[2x(g - g') + 2y(f - f') + c - c'] = 0.$$

It follows that the centres of all circles of a coaxal system lie on a line perpendicular to the radical axis. This is also clear from the fact that the centre of $\lambda C + \mu C' = 0$ is the point

$$\left[\frac{\lambda(-g) + \mu(-g')}{\lambda + \mu},\ \frac{\lambda(-f) + \mu(-f')}{\lambda + \mu}\right].$$

In the next chapter we shall need to be familiar with a few more properties of coaxal systems, and we derive them here.

8. Canonical form for coaxal system

The line of centres and the radical axis of a coaxal system offer, in a sense, a natural system of coordinate axes for the system. We take the line of centres as the x-axis, and the radical axis as the y-axis, and see what form the equation of the system takes.

Since their centres are on the x-axis, any two circles $\mathscr{C}$ and $\mathscr{C}'$ have equations of the form,

$$C \equiv x^2 + y^2 + 2gx + c = 0,$$

and

$$C' \equiv x^2 + y^2 + 2g'x + c' = 0.$$

The radical axis of these two circles is

$$C - C' \equiv 2(g - g')x + c - c' = 0,$$

and since this is to be $x = 0$, we must have $c = c'$. The circles of the system may therefore be written in the form

$$x^2 + y^2 + 2\lambda x + c = 0,$$

where c is constant, and λ a parameter. This is a canonical, or standard form for a coaxal system.

We note that the system now depends linearly on the circle $x^2 + y^2 + c = 0$, and the radical axis $x = 0$.

We can now easily distinguish between the intersecting and non-intersecting type of coaxal system. Intersections, if any, must lie on the radical axis $x = 0$. This meets circles of the system where $y^2 + c = 0$. If c is positive, there are no intersections, and we have the non-intersecting type of coaxal system. If c is negative, all circles of the system cut the radical axis in the same two points.

(i) $c = k^2 > 0$. Any circle of the system may be written in the form

$$(x + \lambda)^2 + y^2 = \lambda^2 - k^2.$$

For real circles, we must have $\lambda^2 \geqslant k^2$. There are two circles of zero radius in the system, with centres at $(-k, 0)$ and $(k, 0)$. These points, which we denote by L and L', are called the *limiting points* of the system. No circles of the system have their centres between L and L'. If a circle of the system cuts the line of centres in A, A', then $OA . OA' = k^2 = OL^2$, and it follows that L and L' are harmonic conjugates with respect to A and A'. Hence any circle through L and L' cuts all circles of the coaxal system orthogonally.

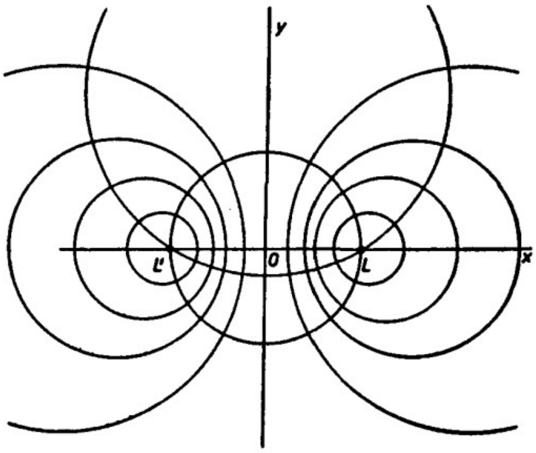

FIG. 22

(ii) $c = -k^2 < 0$. This case presents no features which are not evident geometrically.

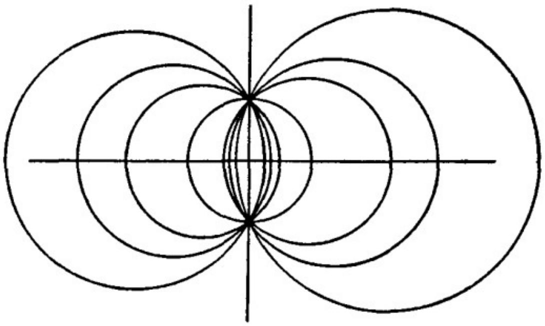

FIG. 23

(iii) $c = 0$. This can be regarded as an intermediate case.

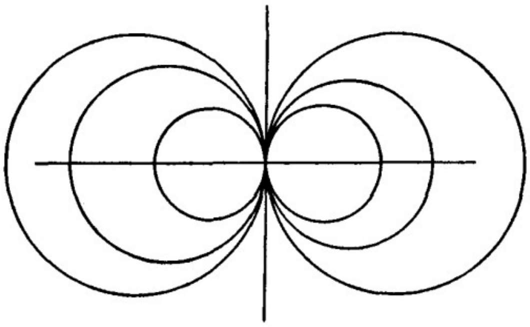

FIG. 24

All circles of the system touch the radical axis, and the point of contact may be regarded as consisting of two coincident limiting points.

We have remarked that in (i) any circle through L and L' is orthogonal to every circle of the given system. The equation of a circle through L and L' is

$$x^2 + y^2 + 2\mu y - k^2 = 0,$$

and this, of course, is a coaxal system of the intersecting type. The orthogonality relation is evident algebraically, since the condition to be satisfied, $2gg' + 2ff' - c - c' = 0$, becomes here $2\lambda . 0 + 2 . 0 . \mu - k^2 + k^2 = 0$, which is the case.

Hence, associated with every coaxal system of a given type there is a coaxal system of the opposite type, and the circles of one cut the circles of the other orthogonally. Such systems are called *conjugate systems*, for reasons which will become clear in chapter II.

9. Further properties

If we try to find the circles of the coaxal system

$$x^2 + y^2 + 2\lambda x + c = 0$$

which pass through $P(x',y')$, we see that

$$\lambda = -\frac{(x'^2 + y'^2 + c)}{2x'};$$

that is, there is a unique value of the parameter which determines the circle. Hence one circle of a coaxal system passes through a given point P. The construction of this circle is evident for intersecting systems, since a circle is determined by three points, but not so evident for a non-intersecting system. In this case we invoke the *conjugate system*, which is of the intersecting type, and draw the unique circle of this system which passes through P. We can then construct the required circle by drawing the unique circle through P which is orthogonal to the constructed circle and has its centre on the line of centres of the given coaxal system.

If we invert a coaxal system of circles, it is fairly clear that we obtain an intersecting coaxal system in the intersecting case, but this is not so clear in the non-intersecting case. However we can bring in the *conjugate system*, which is of the intersecting type, and remembering that circles which are orthogonal invert into orthogonal circles, we see that the circles of the non-intersecting system invert into circles which are orthogonal to all the circles of a coaxal system. Such a system of circles must form the conjugate coaxal system.

However, inversion is a transformation which sometimes does more than we demand of it, and the reader may feel reluctant to accept the proof indicated when we show, as we do now, that *any non-intersecting system of coaxal circles may be inverted into a system of concentric circles.* The question he will ask is: if a system of concentric circles is a coaxal system, why have we not come across it in the above treatment?

We leave the reader to ponder on this apparent paradox, and prove the theorem. Let L and L' be the limiting points of the non-intersecting coaxal system, and invert with respect to a circle centre L. Then the circles of the system invert into circles which are orthogonal to the inverses of all circles through L and L'. But these inverses are straight lines through a fixed point, the inverse of L'. Hence the coaxal system inverts into a system of concentric circles, with centre at the inverse of L'.

This theorem provides a solution of the attractive problem: given two circles $\mathscr{C}$ and $\mathscr{C}'$, one lying inside the other, can we find a chain of circles, $\mathscr{C}_1, \mathscr{C}_2, \mathscr{C}_3, \ldots, \mathscr{C}_n$, all of which touch $\mathscr{C}$ and $\mathscr{C}'$, such that $\mathscr{C}_1$ touches $\mathscr{C}_2$ and $\mathscr{C}_n$, $\mathscr{C}_2$ touches $\mathscr{C}_3$ and $\mathscr{C}_1, \ldots,$ $\mathscr{C}_n$ touches $\mathscr{C}_{n-1}$ and $\mathscr{C}_1$?

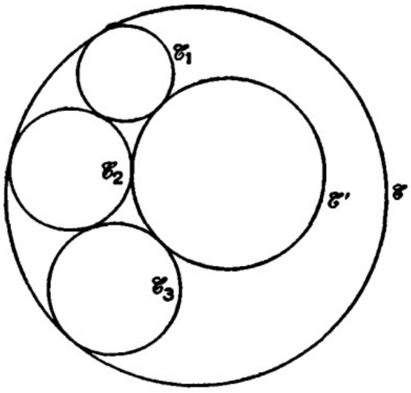

FIG. 25

Inverting $\mathscr{C}$ and $\mathscr{C}'$ into concentric circles, we see that the problem does not always have a solution, but that when there is a solution, we can begin the chain by drawing any circle which touches $\mathscr{C}$ and $\mathscr{C}'$, and can then fit in the other members of the chain in an obvious manner.

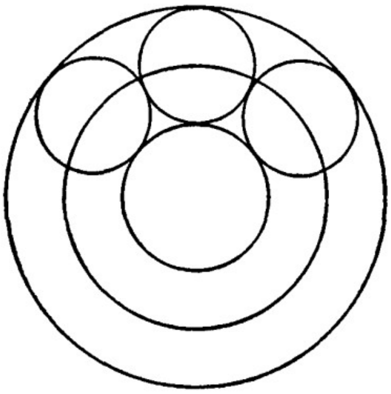

FIG. 26

We conclude this discussion of coaxal circles by considering the radical axes, in pairs, of three circles $\mathscr{C}$, $\mathscr{C}'$, $\mathscr{C}''$. The intersection of two of these radical axes is a point which has equal powers with respect to all three circles. This point is therefore on the third radical axis. Hence, if they are not parallel, the radical axes of the three pairs of circles which can be formed from three given circles are concurrent. The point of concurrence is called the *radical centre* of the three circles. The circle which has its centre at the radical centre, and radius equal to the length of the tangent from the centre to any one of the three circles, if it lies outside one, and therefore all three circles, cuts each of the three circles orthogonally and is the only circle which does this. With respect to this circle as circle of inversion, each of the three circles inverts into itself.

10. Problem of Apollonius

Given three circles $\mathscr{C}_1$, $\mathscr{C}_2$ and $\mathscr{C}_3$, to draw a circle touching all three.

We need two preliminary results:

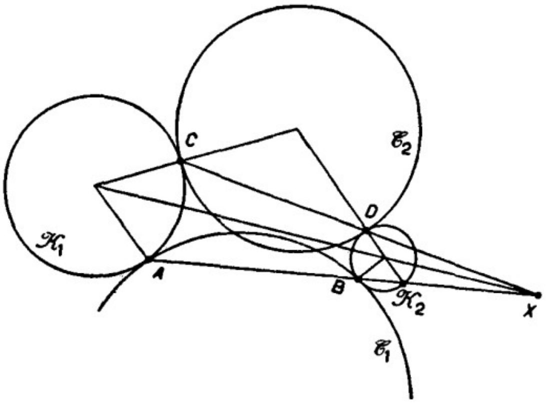

FIG. 27

(i) If two non-concentric circles $\mathscr{K}_1$, $\mathscr{K}_2$ touch two other non-concentric circles $\mathscr{C}_1$, $\mathscr{C}_2$ at A, B and C, D respectively, and if the contacts at C and D are of like[†] or unlike type according as those at A and B are, then AB and CD meet in a centre of similitude of the circles $\mathscr{K}_1$, $\mathscr{K}_2$, or they are parallel to the line of centres of these circles.

To prove this result we join the centres of $\mathscr{K}_1$, $\mathscr{K}_2$ and $\mathscr{C}_1$ and find, using the theorem of Menelaus, where AB cuts the line joining the centres of $\mathscr{K}_1$, $\mathscr{K}_2$. If the given conditions are satisfied, and we consider the triangle formed by the centres of $\mathscr{K}_1$,

---

[†] Contacts are of " like type " when both are external or both are internal.

$\mathscr{K}_2$ and $\mathscr{C}_2$, we see that CD passes through the same point X, and that this is a centre of similitude of $\mathscr{K}_1$ and $\mathscr{K}_2$.

(ii) We remarked earlier (see page 12) that $XA.XB = XC.XD$, so that X is on the radical axis of $\mathscr{C}_1$ and $\mathscr{C}_2$.

We also note that if the contacts at A and B are of like or unlike type according as those at C and D are, then the contacts at A and C are of like or unlike type according as those at B and D are. Therefore AC and BD meet at a centre of similitude of $\mathscr{C}_1$ and $\mathscr{C}_2$, and this centre is on the radical axis of $\mathscr{K}_1, \mathscr{K}_2$.

We have therefore shown that the radical axis of $\mathscr{C}_1$ and $\mathscr{C}_2$ contains a centre of similitude of $\mathscr{K}_1$ and $\mathscr{K}_2$, and the radical axis of $\mathscr{K}_1$ and $\mathscr{K}_2$ contains a centre of similitude of $\mathscr{C}_1$ and $\mathscr{C}_2$.

We now return to our problem, and assume, for simplicity, that the radii of $\mathscr{C}_1$, $\mathscr{C}_2$, $\mathscr{C}_3$ are three distinct numbers, so that all the centres of similitude exist.

Let $\mathscr{K}_0$ be a circle touching the three circles externally. Inversion from the radical centre of $\mathscr{C}_1$, $\mathscr{C}_2$, $\mathscr{C}_3$ changes these circles into themselves, and $\mathscr{K}_0$ is transformed into a circle $\mathscr{K}_1$ which is touched internally by the three circles. The joins of the points of contact of each circle $\mathscr{C}_i$ $(i = 1, 2, 3)$ with $\mathscr{K}_0$ and with $\mathscr{K}_1$ pass through the radical centre.

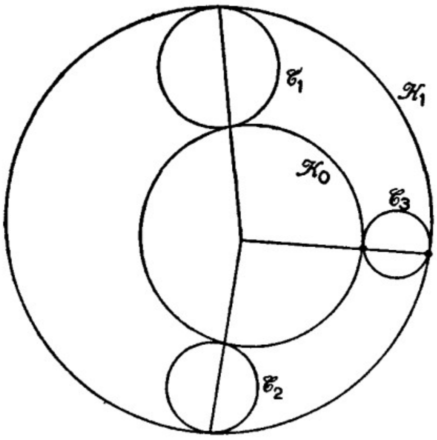

FIG. 28

The pole, for $\mathscr{C}_1$, of the line joining the points of contact of $\mathscr{K}_0$ and $\mathscr{C}_1$ and of $\mathscr{K}_1$ and $\mathscr{C}_1$ is the intersection of the tangents at these points of contact. These tangents are the radical axes of $\mathscr{K}_0$ and $\mathscr{C}_1$ and of $\mathscr{K}_1$ and $\mathscr{C}_1$, and since the radical axes of three circles, taken in pairs, are concurrent, this pole lies on the radical axis of $\mathscr{K}_0$ and $\mathscr{K}_1$.

By the reciprocal property of pole and polar, the polar of this point contains the pole, for $\mathscr{C}_1$, of the radical axis. Now the polar is the line joining the points of contact of $\mathscr{K}_0$ and $\mathscr{C}_1$ and of $\mathscr{K}_1$ and $\mathscr{C}_1$ and passes through the radical centre of $\mathscr{C}_1$, $\mathscr{C}_2$, $\mathscr{C}_3$. Since also the radical axis of $\mathscr{K}_0$ and $\mathscr{K}_1$ contains the external centres of similitude of the circles $\mathscr{C}_1$, $\mathscr{C}_2$, $\mathscr{C}_3$ taken in pairs (this was proved in (ii) above), we have the following construction, due to Gergonne:

Draw the line through the external centres of similitude of $\mathscr{C}_1$, $\mathscr{C}_2$ and $\mathscr{C}_3$, taken in pairs, and join its poles for each of these circles to the radical centre of the three circles. These joins cut the given circles in points of contact of two of the required circles.

The centres of similitude of $\mathscr{C}_1$, $\mathscr{C}_2$, $\mathscr{C}_3$ lie by threes on four lines, and each line of centres of similitude gives two tangent circles. In the most favourable case we therefore expect to find *eight* solutions of the problem of Apollonius.

We shall verify this from another point of view in chapter II.

11. Compass geometry

We conclude this chapter by showing how the straight edge may be dispensed with in the elementary constructions of Euclidean geometry.

A straight edge is used to connect two given points by a straight line. No pair of compasses will do this. At the same time the straight edge and compasses are primarily used to determine certain points. The fundamental problems in Euclidean geometry are:

(i) to find the intersections of two circles, the centre and radius of each being given;

(ii) to find the intersection of a line given by two points with a circle given by its centre and radius;

(iii) to find the intersections of two lines, each given by two points.

The solution of (i) is naturally possible using only the compasses. Problems (ii) and (iii) reduce to (i) after inversion.

Therefore, to show that all Euclidean constructions are possible using only a pair of compasses, we must show:

(a) how to find the inverse of a given point in a circle of given centre and radius;

(b) how to find the centre of a circle which is to pass through three given points.

We gave the solution of (a) in §2, page 5.

Before dealing with (b), we show how to find, using compasses only, the midpoint of the segment determined by two given points. In the first place we show how to extend a given segment AO to twice its length. Draw the circle centre O and radius OA, and mark off the chords $AP = PQ = QA' = OA$. Then AOA' is a straight line, and $AA' = 2OA$.

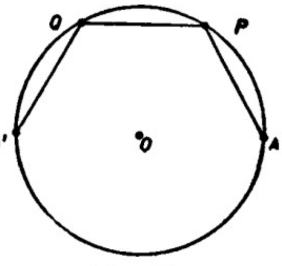

FIG. 29

We can now find the midpoint of the segment OA. Extend AO to A', where $AO = OA'$, and find the inverse of A' in the circle centre A and radius AO. If A'' is the inverse,

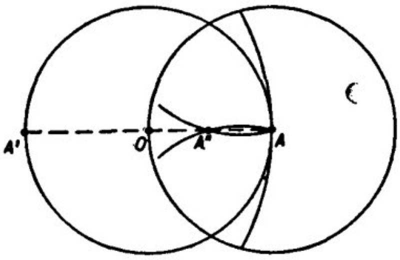

FIG. 30

$$AA''.AA' = AO^2,$$

and since $AA' = 2AO$, the point A'' is the midpoint of OA.

We can extend OA to A', where $OA' = nOA$, n being any positive integer, and the inverse of A' in the circle centre A and radius AO gives OA/n.

(b) Let A, B, C be the three given points. Invert with respect to the circle centre A and radius AB. The circle ABC inverts

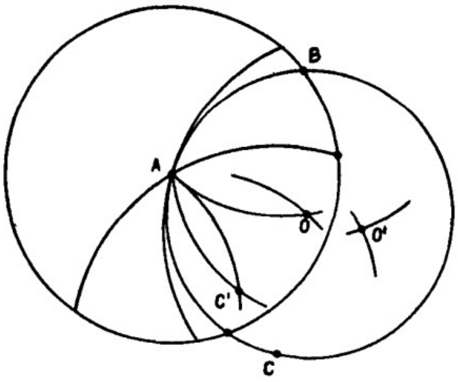

FIG. 31

into the line BC', where C' is the inverse of C in the circle of inversion. This point can be constructed. The centre of circle ABC inverts into the inverse of the centre of inversion, which is A, in the inverse of circle ABC, namely, the line BC' (see §2, page 9). The inverse of A in BC' is the geometrical image of A in BC'. This is easily constructed, as indicated in Fig. 31. If this geometrical image be O', we finally find the inverse of O' in the circle of inversion. This final point O is the centre of circle ABC.

The reader should carry out this construction for himself.

To find where two lines determined by points A,B and C,D respectively meet, we invert with respect to any suitable circle, centre O, say. If the respective inverse points are A',B' and C',D', the intersection of the circles $OA'B'$ and $OC'D'$, other than O, is the inverse of the point required. We can draw these circles, by the above construction, and find this point.

Bright students often rediscover this compass geometry for themselves, and are dashed to find that it was fully investigated, without the benefit of inversion, by Mascheroni in *La Geometria del Compasso*, Pavia, 1797.

7. Circles cutting at a given angle

We now return to the general relation (1) of the previous section, and see that the points representing circles which cut a fixed circle $\mathscr{C}'$ at an angle α or $\pi - \alpha$ satisfy the equation

$$4(x^2 + y^2 - z)(\xi'^2 + \eta'^2 - \zeta')\cos^2\alpha = (2x\xi' + 2y\eta' - z - \zeta')^2. \tag{1}$$

This represents a quadric surface, and the form of its equation shows that it has *ring-contact* with Ω along the intersection of Ω with the plane

$$2x\xi' + 2y\eta' - z - \zeta' = 0. \tag{2}$$

If the reader is unfamiliar with the idea of ring-contact he need not be disturbed, since it is only the algebraic form of (1) which will concern us. We know already that the curve in which the plane given by (2) meets Ω projects into the circle $\mathscr{C}'$ itself.

We wish to investigate the circles which cut three given circles $\mathscr{C}_1$, $\mathscr{C}_2$ and $\mathscr{C}_3$ at angles α or $\pi - \alpha$. To do this, we must examine the common intersections of the three quadric surfaces:

$$4(x^2 + y^2 - z)(\xi_r^2 + \eta_r^2 - \zeta_r)\cos^2\alpha = (2x\xi_r + 2y\eta_r - z - \zeta_r)^2 \qquad (r = 1, 2, 3).$$

For the sake of brevity we write these equations as

$$4(x^2 + y^2 - z)k_r\cos^2\alpha = X_r^2 \qquad (r = 1, 2, 3), \tag{3}$$

where k_r is a constant, and $X_r = 0$ is the polar plane of (ξ_r, η_r, ζ_r). If we equate the common value of $x^2 + y^2 - z$ in the three equations, we obtain the equations

$$\frac{X_1^2}{k_1} = \frac{X_2^2}{k_2} = \frac{X_3^2}{k_3}.$$

We note that these equations do not depend on $\cos\alpha$. If we take them in pairs, the system is equivalent to

$$\begin{cases} \sqrt{k_2}\,X_1 \pm \sqrt{k_1}\,X_2 = 0, \\ \sqrt{k_3}\,X_2 \pm \sqrt{k_2}\,X_3 = 0. \end{cases} \tag{4}$$

Taking all the possible combinations of signs, we have four sets of two linear equations in x, y, z. Since two planes determine a line, the solutions of (4) lie on four straight lines. Since $X_1 = X_2 = X_3 = 0$ automatically satisfies (4), these four lines all pass through the point determined by the intersection of the polar planes of (ξ_r, η_r, ζ_r) $(r = 1, 2, 3)$. This point represents the circle orthogonal to the three given circles.

Each one of the lines determined by (4) meets each one of the three quadrics given by (3) in the same two points. Hence there are *eight* solutions of the equations, for a given value of $\cos\alpha$, and we have proved the following theorem:

There are eight circles which intersect three given circles at a given angle α *or* $\pi - \alpha$. *As the angle* α *varies the circles vary in four coaxal systems with a common circle, the circle orthogonal to the three given circles.*

The case $\alpha = 0$ gives the solution of the problem of Apollonius (see chapter I, §10). Since each one of the lines determined by (4) gives two solutions of (3), each one of the four coaxal systems contains two of the circles which satisfy the given conditions.

8. Representation of inversion

Let A, A' be inverse points in a circle $\mathscr{C}$. Then $\mathscr{C}$ is orthogonal to any circle through A and A', and $\mathscr{C}$ is a member of the coaxal system which is determined by the limiting points A, A'. Let P represent the circle $\mathscr{C}$. The zero circles centres A, A' are represented by points on Ω. By the remark above, P is collinear with these two points on Ω.

Since P is regarded as fixed, and A, A' as variable, we have the following method for finding the inverse, in a given circle $\mathscr{C}$, of a given curve, $\mathscr{D}$, say:

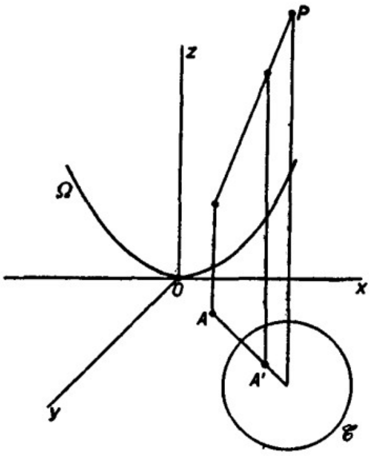

FIG. 40

Erect perpendiculars to the plane Oxy at the points of $\mathscr{D}$, and let $\mathscr{E}$ be the curve in which these perpendiculars meet Ω. The cone of lines joining P to the points of $\mathscr{E}$ meets Ω in a further curve $\mathscr{E}'$. Project $\mathscr{E}'$ down on to the plane Oxy. The resulting curve $\mathscr{D}'$ is the inverse of $\mathscr{D}$ in $\mathscr{C}$.

We remark that this construction shows at once that the eight circles which were discussed in the preceding section are inverse in pairs in the common orthogonal circle of the three circles. Since each of the three given circles inverts into itself with respect to the common orthogonal circle, this result is to be expected.

If the curve $\mathscr{D}$ to be inverted is a circle, we know that the curve $\mathscr{E}$ obtained by projecting it on to Ω is a plane section. The cone of lines joining P to the points of $\mathscr{E}$ is a quadric cone, and since the inverse of a circle is another circle, this cone meets Ω again in a further plane section $\mathscr{E}'$. Again, a circle, the circle of inversion and the inverse circle are coaxal, so that we obtain the theorem:

A quadric cone which meets Ω in one plane section will meet it again in another plane section, and the poles of the two plane sections are collinear with the vertex of the cone.

The more advanced reader will realize that this is a *projective* theorem, and is therefore true for any quadric Ω.

9. The envelope of a system

We may measure the complexity of a system of circles by the complexity of the curve in E_3 which represents the system. From this point of view a coaxal system of circles is the simplest kind of system, since a line is the simplest curve in E_3. A coaxal system of circles has no *envelope*: that is, there is no curve touched by all circles of the system. For intersecting systems we may, if we wish, say that the envelope consists of the two points common to all circles of the system. But interesting envelopes only appear when we consider *conic systems of circles*.

A conic system of circles is one represented by a conic in E_3. A conic is a plane curve, and this plane will have a pole with respect to Ω. This pole represents a circle which is orthogonal to all the circles of the conic system. Again, the projection of the conic on to the plane Oxy is, except in one special case, a conic, and gives the locus of centres of circles of the conic system.

Hence the centres of circles of a conic system lie on a conic, and all circles of the system are orthogonal to a fixed circle.

data 06.05.2020

The ellipse and the circle should have the same area.

Pag. 49 (66) 3.5.2 Geometry in the Music Cassette.

Pag. 52 (69) 3.5.3 The Ellipse in a Square Lattice.

Pag. 52 (69) 3.6 Golden Trigonometry.

$$\sin 18^\circ = \frac{\rho}{2} = \frac{\tau - 1}{2}$$

$$\cos 18^\circ = \sqrt{1 - \frac{\rho^2}{4}} = \frac{\sqrt{2+\tau}}{2}$$

$$\tan 18^\circ = \frac{\tau - 1}{\sqrt{2+\tau}}$$

Pag. 75 (92) 5.3 Power of $1+\sqrt{2}$

E 7.8) Circles, a mathematical view... djvu

Obs.: Livro todo !!!

Pag. 01 (39) Chapter 1
The nine-point circle.

Pag. 35 (73) 7. Circles cutting at a given angle.

F1 $. L ! + &
)*+,)))-) ' . L ! + $: 2
** !H8

1)*+,)))-) ' . A & ! *
& ! & ! F ! ! @A6 7 . A 6!

1)*+,)))-) ' / F
! G 12 !H8
(;

(1 F ! 0
)*+,)))-) ' / F
! G 1 1 : 6!

Euler and Differentials[1]

Anthony P. Ferzola

Two recent articles by Dunham [5] and Flusser [10] have presented examples of Leonhard Euler's work in algebra. Both papers are a joy to read; watching Euler manipulate and calculate with incredible facility is a pleasure. A modern mathematician can see the logical flaws in some of the arguments, yet at the same time be aware that the mind behind it all is that of a unique master.

These two articles reminded me how much fun it is to read Euler. In researching the evolution of the differential a few years ago, I found the work of Euler refreshingly different from that of other seventeenth- and eighteenth-century mathematicians. One can read about Euler's use and misuse of infinite series in most histories of mathematics (e.g., [2, pp. 486–490]). This paper offers a glimpse at how Euler used infinitesimals and infinite series to compute differentials for the elementary functions encountered in a typical undergraduate calculus sequence. I hope the reader of this brief survey of Euler's work with differentials will seek out original sources such as [8] and [9]. As Harold Edwards [7] has cogently argued, we have much to learn from reading the masters.

Euler and the 18th Century

Euler (1707–1783) was the most prolific and one of the most influential mathematicians who ever lived. He made major contributions to both pure and applied mathematics and his collected works amount to over 70 volumes. So strong was his influence that historians like Boyer [2] and Edwards [6] refer to the eighteenth century as the Age of Euler.

Euler made the function concept fundamental in analysis. He saw a function as both any quantity depending on variables and also as any algebraic combination of constants and variables (including infinite sums or products). This is obviously not a modern definition of a function. Still, Euler used his function concept to maximal advantage. As we examine some of Euler's computations, keep in mind the immense insight and unity he achieved with the function approach—a point of view we now take for granted.

---

[1] Reprinted from the *College Mathematics Journal*, Vol. 25, March, 1994, pp. 102–111. This article received the George Pólya Award in 1995.

In his *Introducio in analysin infinitorum* (1748), one sees the first systematic interpretation of logarithms as exponents. Prior to Euler, logarithms were typically viewed as terms of an arithmetic series in one-to-one correspondence with terms of a geometric series [3]. Euler viewed trigonometric functions as numerical ratios rather than as ratios of line segments. He also studied properties of the elementary transcendental functions by the frequent use of their infinite series expansions [6, p. 270]. Euler often used infinite series indiscriminately, without regard to questions of convergence.

Euler's understanding and use of differentials within the framework of functions is the focus of this paper. Before presenting his work, a word about the differential before Euler.

For Leibniz (1646–1716) the differentials dx and dy were, as the name suggests, (infinitesimal) differences in the abscissa x and the ordinate y, respectively [4, pp. 70–76]. The infinitesimal was considered to be a number smaller than any positive number. The omission of the "even smaller" higher-order infinitesimals such as $(dx)^2$ or $dx\,dy$, which were deemed negligible relative to dx and dy, was basic to his methods. So powerful were the notation and methods that the differential calculus was truly a *differential* calculus for nearly one and a half centuries: The differential (and not the derivative) was the main object of study.

Leibniz gave other interpretations of the differential, but the mathematicians working in the early eighteenth century tended to favor Leibniz's formulation of a differential as an infinitesimal. It appears in the work of Johann Bernoulli (1667–1748) and in the first calculus textbook, *Analyse des infiniment petits pour l'intelligence des lignes courbes* (1696), which was written by L'Hôpital and which made free use of Bernoulli's ideas (see [18, p. 315]). Euler was one of Bernoulli's pupils.

Many of Euler's results and infinite series discussed below were known to Newton, Leibniz, Bernoulli, and others. Euler's work with differentials is unique, however, in his definition of infinitesimals as absolute zeros and in his heavy reliance on infinite series to *develop* his differential calculus.

Differentials as Absolute Zeros

In his *Institutiones calculi differentialis* (1755), Euler stated: "To those who ask what the infinitely small quantity in mathematics is, we answer it is actually equal to zero" [18, p. 384]. Euler felt that the view of the infinitesimal as zero adequately removed the mystery and ambiguity of statements such as "The infinitesimal is smaller than any given quantity" or the postulate of Johann Bernoulli that "Adding an infinitesimal to a quantity leaves the quantity unchanged."

Euler then said that the quotient $0/0$ can actually take on any value because

$$n \cdot 0 = 0$$

for all real n and therefore, he concluded,

$$\frac{n}{1} = \frac{0}{0}. \tag{1}$$

He noted that if two zeros can have an arbitrary ratio, then different symbols should be used for the zero in the numerator and the zero in the denominator of the fraction on the right-hand side of equation (1). It is here that Euler introduced the Leibnizian notation of differentials.

Euler denoted an infinitely small quantity by dx. Here $dx = 0$ and $a\,dx = 0$ for any finite quantity a. But for Euler these two zeros are different zeros that cannot be confused when the ratio $a\,dx/dx = a$ is investigated [18, p. 385]. In a similar way dy/dx can denote a finite ratio even though dx and dy are zero. "Thus for Euler the calculus was simply the determination of the ratios of evanescent increments—a heuristic procedure for finding the value of the expression 0/0" [1].

The neglect of higher-order infinitesimals was also explained employing quotients. Noting that $dx = 0$ and $(dx)^2 = 0$, where $(dx)^2$ is a zero (or infinitesimal) of second order, Euler reasoned that

$$dx + (dx)^2 = dx$$

because

$$\frac{dx + (dx)^2}{dx} = 1 + dx = 1.$$

By the same reasoning, Euler established that

$$dx + (dx)^{n+1} = dx$$

for all $n > 0$. The omission of higher-order differentials was frequently utilized by Euler in finding the differential dy, where y is a function of x.

Computations with Elementary Functions

The computations discussed in this section are all found in Euler's *Institutiones calculi differentialis*. Their most noteworthy feature is the use of power series expressions for functions from the outset, with no mention of questions of convergence. Thus, whereas in modern textbooks the justification of such infinite series expansions is an advanced topic in differential calculus, for Euler they were the foundation for the calculation of derivatives.

To find dy if $y = x^n$ (n any real number), Euler used the binomial expansion [9, p. 99]. If x is increased by an infinitesimal amount dx, then y experiences a change of dy where

$$\begin{aligned} dy &= (x + dx)^n - x^n \\ &= nx^{n-1}\,dx + \frac{n(n-1)}{1}\,x^{n-2}\,(dx)^2 + \cdots \\ &= nx^{n-1}\,dx \end{aligned}$$

upon the omission of the higher-order infinitesimals $(dx)^2$, etc. Newton and Leibniz did similar computations for finding the derivative of $y = x^n$, Leibniz using a comparable differential argument while Newton worked with fluxions [6, p. 192]. Within the rigorous context of

$$\lim_{\Delta x \to 0} \frac{(x + \Delta x)^n - x^n}{\Delta x}$$

we all use the essence of this computation (for positive integer powers of x) in our first semester calculus courses.

Euler derived the product rule as follows:

$$\begin{aligned} d(pq) &= (p+dp)(q+dq)-pq \\ &= p\,dq + q\,dp + dp\,dq \\ &= p\,dq + q\,dp \end{aligned}$$

where the last step is due to the omission of the higher-order infinitesimal $dp\,dq$. Similar computations were done by Leibniz [4, p. 143]. This argument is analogous to the proof of the product rule still found in a few present-day textbooks (e.g., [12]).

Euler's derivation of the quotient rule is unique in its use of a geometric series [9, p. 103]:

$$\begin{aligned} \frac{1}{q+dq} &= \frac{1}{q}\left(\frac{1}{1+dq/q}\right) \\ &= \frac{1}{q}\left(1-\frac{dq}{q}+\frac{dq^2}{q^2}-\cdots\right) \\ &= \frac{1}{q}-\frac{dq}{q^2}. \end{aligned}$$

Then

$$\begin{aligned} d\left(\frac{p}{q}\right) &= \frac{p+dp}{q+dq}-\frac{p}{q} \\ &= (p+dp)\frac{1}{q+dq}-\frac{p}{q} \\ &= (p+dp)\left(\frac{1}{q}-\frac{dq}{q^2}\right)-\frac{p}{q} \\ &= \frac{dp}{q}-\frac{p\,dq}{q^2} \\ &= \frac{q\,dp-p\,dq}{q^2}. \end{aligned}$$

In chapter 6, Euler found the differentials of transcendental functions. For computing the differential of the natural logarithm (which he denoted by the single letter "ℓ" but which we will denote by the usual "log"), Euler used Mercator's series [9, p. 122]:

$$\log(1+z) = z-\frac{z^2}{2}+\frac{z^3}{3}-\cdots.$$

Given $y = \log(x)$ then

$$\begin{aligned} dy &= \log(x+dx)-\log(x) \\ &= \log\left(1+\frac{dx}{x}\right) \\ &= \frac{dx}{x}-\frac{(dx)^2}{2x^2}+\frac{(dx)^3}{3x^3}-\cdots \\ &= \frac{dx}{x}. \end{aligned}$$

To illustrate the chain rule, Euler did many examples. For instance, if $y = \log(x^n)$ then letting $p = x^n$ yields $y = \log(p)$, which implies that $dy = dp/p$ where $dp = nx^{n-1}\,dx$. Thus $dy = n\,dx/x$.

Euler's computation of dy for $y = \log(x)$ can be found in a modern nonstandard analysis text [15, p. 65]. This may seem unremarkable since nonstandard analysis was developed by Abraham Robinson in the mid-twentieth century to place the notion of infinitesimals and their manipulation on solid logical ground. In fact, it is rare to find nonstandard analysis arguments that are exactly like Euler's, because nonstandard analysis arguments are rarely done in the context of infinite series (see [11] and [16]).

As an example of Euler's work with trigonometric functions, consider the computation of dy for $y = \sin x$ [9, p. 132]. For this purpose he explicitly used the sine and cosine series

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots \tag{2}$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots \tag{3}$$

to show that $\sin(dx) = dx$ and $\cos(dx) = 1$. He obtained these results by substituting dx into (2) and (3) and ignoring higher-order differentials. He also employed the trigonometric identity

$$\sin(a + b) = \sin a \cos b + \sin b \cos a. \tag{4}$$

Thus, using (4):

$$\begin{aligned} dy &= \sin(x + dx) - \sin x \\ &= \sin x \cos dx + \sin dx \cos x - \sin x \\ &= \sin x + \cos x\, dx - \sin x \\ &= \cos x\, dx. \end{aligned}$$

This is the most beautifully efficient computation of all those presented, especially when compared to the usual limit computation of the derivative of $y = \sin x$. There one needs to work as follows:

$$\begin{aligned} \lim_{\Delta x \to 0} \frac{\sin(x + \Delta x) - \sin x}{\Delta x} &= \lim_{\Delta x \to 0} \frac{\sin x \cos(\Delta x) + \sin(\Delta x) \cos x - \sin x}{\Delta x} \\ &= \cos x \lim_{\Delta x \to 0} \frac{\sin(\Delta x)}{\Delta x} + \sin x \lim_{\Delta x \to 0} \frac{\cos(\Delta x) - 1}{\Delta x} \end{aligned}$$

Then $y' = \cos x$ is obtained using two limits (which must be proven):

$$\lim_{x \to 0} \frac{\sin x}{x} = 1 \tag{5}$$

and

$$\lim_{x \to 0} \frac{\cos x - 1}{x} = 0.$$

The first of these limits is captured in Euler's equation $\sin(dx) = dx$. The second limit is comparable to Euler's equation $\cos(dx) = 1$ or $\cos(dx) - 1 = 0$. Although

Euler's derivation is computationally more compact than the standard modern approach, the latter is logically sound. Any method for differentiating the sine function must deal in particular with (5). This is proven geometrically, since in the standard modern approach one defines at the outset the geometric meaning of the trigonometric functions (i.e., cosine and sine parametrize the unit circle). The proof of (5) is relatively easy when compared to the difficulty involved in showing the geometric meaning of the functions Euler defined (without regard to questions of convergence) as the sums of the power series (2) and (3).

In Euler's three-volume *Institutiones calculi integralis* (1768–1770), he defined integration, like Leibniz and Johann Bernoulli, as the formal inverse of the differential. He used the integral symbol and wrote, for example,

$$\int nx^{n-1}\,dx = x^n$$

$$\int \frac{dx}{x} = \log x$$

$$\int \cos x\,dx = \sin x$$

all plus or minus an appropriate constant. The first volume of this work reads like a modern calculus textbook chapter on techniques of integration. Integration by substitution, by parts, by partial fractions, and by trigonometric substitution are all illustrated in a logical and systematic way. Undoubtedly, Euler's well-organized and all-encompassing use of differentials in a function context did much to solidify the popularity of the differential and integral notations on the continent.

The Total Differential

Euler's *Institutiones calculi differentialis* was the first systematic exposition of the calculus of functions of several variables. He understood a function of n variables to be any finite or infinite expression involving these variables. As soon as he introduced these functions, Euler addressed the question of the relationship among the differentials of all the variables involved.

He obtained the result that if

$$V = f(x, y, z)$$

then

$$dV = p\,dx + q\,dy + r\,dz,$$

where p, q, and r are all functions of x, y, and z [9, pp. 144–145]. He arrived at this formula in an interesting way. If X is a function of x alone and is increased by an infinitesimal amount dx, then

$$dX = P\,dx$$

by the usual one-variable argument. Similarly, if Y and Z are functions of y alone and z alone respectively, then

$$dY = Q\,dy$$

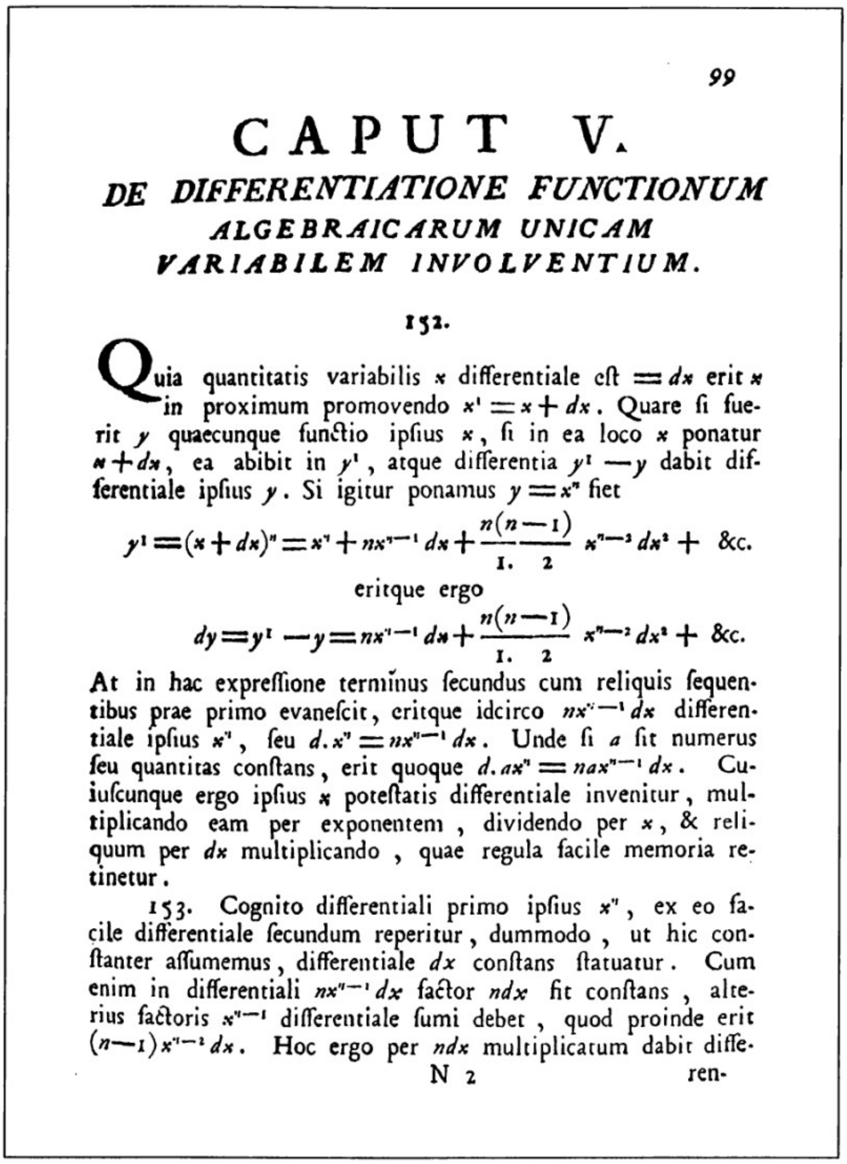

99

CAPUT V.

DE DIFFERENTIATIONE FUNCTIONUM ALGEBRAICARUM UNICAM VARIABILEM INVOLVENTIUM.

152.

Quia quantitatis variabilis x differentiale eſt $= dx$ erit x in proximum promovendo $x^{I} = x + dx$. Quare ſi fuerit y quaecunque functio ipſius x, ſi in ea loco x ponatur $x + dx$, ea abibit in y^{I}, atque differentia $y^{I} - y$ dabit differentiale ipſius y. Si igitur ponamus $y = x^{n}$ fiet

$$y^{I} = (x + dx)^{n} = x^{n} + nx^{n-1}dx + \frac{n(n-1)}{1.\ 2}x^{n-2}dx^{2} + \&c.$$

eritque ergo

$$dy = y^{I} - y = nx^{n-1}dx + \frac{n(n-1)}{1.\ 2}x^{n-2}dx^{2} + \&c.$$

At in hac expreſſione terminus ſecundus cum reliquis ſequentibus prae primo evaneſcit, eritque idcirco $nx^{n-1}dx$ differentiale ipſius x^{n}, ſeu $d.x^{n} = nx^{n-1}dx$. Unde ſi a ſit numerus ſeu quantitas conſtans, erit quoque $d.ax^{n} = nax^{n-1}dx$. Cuiuſcunque ergo ipſius x poteſtatis differentiale invenitur, multiplicando eam per exponentem, dividendo per x, & reliquum per dx multiplicando, quae regula facile memoria retinetur.

153. Cognito differentiali primo ipſius x^{n}, ex eo facile differentiale ſecundum reperitur, dummodo, ut hic conſtanter aſſumemus, differentiale dx conſtans ſtatuatur. Cum enim in differentiali $nx^{n-1}dx$ factor ndx ſit conſtans, alterius factoris x^{n-1} differentiale ſumi debet, quod proinde erit $(n-1)x^{n-2}dx$. Hoc ergo per ndx multiplicatum dabit diffe-

N 2 ren-

Title page, Chapter 5 of *Institutiones calculi differentialis*

and

$$dZ = R\,dz.$$

If $V = X + Y + Z$ (i.e., a special function of three variables) then

$$\begin{aligned} dV &= dX + dY + dZ \\ &= P\,dx + Q\,dy + R\,dz. \end{aligned}$$

If $V = XYZ$, then

$$dV = (X + P\,dx)(Y + Q\,dy)(Z + R\,dz) - XYZ.$$

This simplifies (upon omission of higher-order differential terms such as $ZPQ\,dx\,dy$) to

$$dV = YZP\,dx + XZQ\,dy + XYR\,dz.$$

From these two examples, Euler expected that any algebraic expression of x, y, and z has differential

$$dV = p\,dx + q\,dy + r\,dz \tag{6}$$

because a function of three variables can be thought of as a sum of products of these variables. He generalized the result for any number of variables [9, p. 146].

Later in the same work, he addressed the concept of partial differentiation [9, pp. 156–157]. If y and z are held constant, then by equation (6)

$$dV = p\,dx$$

as there is no change in y or z. (Notice how, for Euler, no change in y is not the same as saying dy is the infinitesimal change in y, even though he defined infinitesimals as being zero.) He then wrote

$$p = (dV/dx),$$

where the parentheses about the quotient remind one that p equals the differential of V (with only the x being variable) divided by dx. Similar meanings apply to $q = (dV/dy)$ and $r = (dV/dz)$. This was Euler's notation and understanding of the concept of partial derivatives. The current symbol ∂ dates from the 1840's [14]. Obviously, (6) becomes

$$dV = (dV/dx)\,dx + (dV/dy)\,dy + (dV/dz)\,dz,$$

although Euler did not explicitly write this.

It is worth noting that Euler's exposition of differentials for functions of several variables immediately followed his work with differentials for functions of one variable. Exploring the differential calculus for both single and multivariable functions before passing on to integration is an old idea which I think has merit. It gives the calculus sequence a stronger focus and unity, by concentrating effort on one basic concept (the derivative) in various settings before moving on to its inverse. A recent textbook by Small and Hosack [17] takes this approach. Perhaps we will see more of this, especially since computer algebra systems such as *Derive*, *Maple*, and *Mathematica* have taken the pain out of such tasks as surface sketching.

Differentials in Multiple Integrals

Euler frequently let his readers in on his thought processes, even when the procedures seemed fruitless. This was mathematics being done for all to see, not a slick modern textbook treatment. There was no taking down the scaffolding *à la* Gauss.

Euler, in *De formulis integralibus duplicatis* (1769), gave one of the first clear discussions of double integrals. In the first half of the eighteenth century, $\iint f(x, y)\,dx\,dy$ denoted the solution of $\partial^2 z/\partial x\,\partial y = f(x, y)$ obtained by antidifferentiation. Euler supplemented this by providing a (thoroughly modern) procedure for evaluating definite double integrals over a bounded domain **R** enclosed by arcs in the xy plane. Euler used iterated integrals:

$$\iint_{\mathbf{R}} f(x, y)\, dx\, dy = \int_a^b dx \int_{f_1(x)}^{f_2(x)} z\, dy,$$

where $z = f(x, y)$. For $z > 0$, Euler saw this as a volume, since $\int z\, dy$ gives the area of a "slice" (parallel to the y-axis) of the three-dimensional region above **R** and under $z = f(x, y)$, and the following integration with respect to x "adds up the slices" to yield the volume [8, p. 293]. This is perhaps the first time Leibniz's powerful differential notation was used in tandem with a volume argument employing Cavalieri's method of indivisibles [2, p. 361].

Euler also interpreted $dx\, dy$ as an "area element" of **R**. That is, **R** is made up of an infinite set of infinitesimal area elements $dx\, dy$. This is most clearly seen when Euler attempted to change variables [8, pp. 302–303]. And it was here that Euler ran into difficulties.

He reasoned that if $dx\, dy$ is an area element and we change variables via the transformation

$$\begin{aligned} x &= x(t, v) = a + mt + v\sqrt{1 - m^2} \\ y &= y(t, v) = b + t\sqrt{1 - m^2} - mv \end{aligned}$$

(a translation by the vector (a, b), a clockwise rotation through the angle α, where $\cos\alpha = m$, and a reflection through the x-axis), then $dx\, dy$ should equal $dt\, dv$. But

$$\begin{aligned} dx &= m\, dt + dv\sqrt{1 - m^2} \\ dy &= dt\sqrt{1 - m^2} - m\, dv \end{aligned}$$

and multiplication gives

$$dx\, dy = m\sqrt{1 - m^2}\,(dt)^2 + (1 - 2m^2)\, dt\, dv - m\sqrt{1 - m^2}\,(dv)^2.$$

Euler rejected this as wrong and meaningless. (How many calculus students wonder, explicitly or implicitly, why we cannot just multiply the differential forms for dx and dy?) Euler decided to attack the problem in a formal non-geometric way, not using area elements but rather by changing variables one at a time (for details, see [13]). In this way he arrived at the correct general result:

$$\iint f(x, y)\, dx\, dy = \iint f\big(x(t, v), y(t, v)\big) \left|\frac{\partial(x, y)}{\partial(t, v)}\right| dt\, dv.$$

In 1899, another great mathematician with a computational flair, Élie Cartan, arrived at the straightforward multiplicative result Euler sought, by using Grassmann's exterior product with differential forms. This is a formal product where the usual distributive laws hold but with the conditions that

$$dx\, dx = dy\, dy = 0$$

and

$$dx\, dy = -dy\, dx$$

(see [17, p. 514], and [13]). Thus, for Euler's differentials

$$\begin{aligned} dx\,dy &= \left(m\,dt + dv\sqrt{1-m^2}\right)\left(dt\sqrt{l-m^2} - m\,dv\right) \\ &= dt\,dt\,m\sqrt{1-m^2} - m^2\,dt\,dv + (1-m^2)\,dv\,dt - dv\,dv\,m\sqrt{1-m^2} \\ &= -m^2\,dt\,dv + dv\,dt(1-m^2) \\ &= -m^2\,dt\,dv - dt\,dv(1-m^2) \\ &= -dt\,dv. \end{aligned}$$

The minus sign appears because the transformation (involving a reflection) does not preserve orientation. In general, given any transformation from the tv-plane to the xy-plane, the exterior product yields

$$dx\,dy = \frac{\partial(x,y)}{\partial(t,v)}\,dt\,dv.$$

Conclusion

Even in this rudimentary survey of Euler's work with differentials in calculus, it is fascinating to watch a genius grapple with an ambiguous concept (infinitesimal) and attempt to clarify it (absolute zero)—however flawed the attempt. Reading Euler has enriched my teaching of the calculus by keeping me mindful that my students are tackling a subject whose foundations humbled the greatest minds of the past. Even the seemingly fruitless paths can be instructive, as we have seen. It took mathematicians about 150 years to come up with the exterior product for differential forms that Euler needed for the change of variables formula in multiple integrals. How many other Eulerian dead ends may be worth pursuing? Again, the advice of Harold Edwards [7] points the way for the teacher and the researcher: "Read the masters!"

References

1. C. Boyer, *The History of Calculus*, Dover, New York, 1959, p. 244.
2. ——, *A History of Mathematics*, Wiley, New York, 1968.
3. F. Cajori, *A History of Mathematics*, Macmillan, London, 1931, p. 235.
4. J. Child, *The Early Mathematical Manuscripts of Leibniz*, Open Court, Chicago, 1920.
5. W. Dunham, Euler and the fundamental theorem of algebra, *College Mathematics Journal* 22 (1991) 282–293.
6. C. H. Edwards, Jr., *The Historical Development of the Calculus*, Springer-Verlag, New York, 1979.
7. H. M. Edwards, Read the masters! in L. A. Steen, ed., *Mathematics Tomorrow*, Springer-Verlag, New York, 1981.
8. L. Euler, *De formulis integralibus duplicatis*, in A. Gutzmer, ed., *Opera omnia: L. Euler*, Vol. 17, Series 1, B. G. Teubner, Leipzig, 1915.
9. ——, *Institutiones calculi differentialis*, in C. Kowalewski, ed., *Opera Omnia: L. Euler*, Vol. X, Series 1, B. G. Teubner, Leipzig, 1913.
10. P. Flusser, Euler's amazing way to solve equations, *Mathematics Teacher* 85 (1992) 224–227.

11. J. M. Henle and E. M. Kleinberg, *Infinitesimal Calculus*, MIT Press, Cambridge, 1979.

12. D. Hughes-Hallett, A. M. Gleason, et. al., *Calculus: Preliminary Edition*, Wiley, New York, 1992, p. 239.

13. V. Katz, Change of variables in multiple integrals: Euler to Cartan, *Mathematics Magazine* 55 (1982) 3–11.

14. ——, The history of differential forms from Clairaut to Poincaré, *Historia Mathematica* 8 (1981) 161–188, p. 161.

15. A. Robert, *Nonstandard Analysis*, Wiley, New York, 1988, p. 65.

16. A. Robinson, *Non-Standard Analysis*, North-Holland, Amsterdam, 1970.

17. D. B. Small and J. M. Hosack, *Calculus: An Integrated Approach*, McGraw-Hill, New York, 1990.

18. D. J. Struik, *A Source Book in Mathematics 1200–1800*, Harvard University Press, Cambridge, 1969.

06.05.2020

E7.9) Generalized Riemann_Integral ☆
(1980...)...djvu.

E7.10) Geometry Illuminated...djvu

Obs: Tenho em pdf.

E7.11) The genius of Euler...djvu

Pág. X (10)

All celebrated mathematicians now alive are his disciples: there is no one who... is not guided and sustained by the genius of Euler.

From the Marquis de Condorcet's Eulogy to Euler, 1783

Pág. 155 (172) Euler and Differentials[1]

F) Books_79_2020/
Fractional_Trigonometry/ ✡

The Fractional Trigonometry...pdf

Obs: Livro todo!!! Muito bom!!!

G) Books_79_2020/
Trisectors Dudley/

A_Budget_of_Trisectors.pdf

Obs. Livro todo!!!

? G > $> #< >/ A#
)*+,)))-) ' / < / A I
*66 < 0 F= ! 12 D I A Y 6!

)*+,)))-) ' / A>
66 ! = ! 6!
-7
5 8 ! 999

)*+,)))-) ' *>? ! > >/ 6!
-7 5 8 ! 999

)Z&)?--J " % $ 1L)?--J > > ")? G >1$>)
F8)
-7
5 8 ! 999

*+,>& -+/*3/> %#

*+, & -+/*3/ " % 1 1L)*+,>& -+/*3/> %).M M)

)*+,)))-) ' D = @ ! ! 5 V !
M M M M & M M 6 :$:2 * ! 6!

7)*+,)))-) ' D = @ ! ! 5 V !
M M M M && M M ! X 6 $ 2 * ! 6!

D *+,>& -+/*3/> #.M M #

*+, & -+/*3/ " % 1 1L)*+,>& -+/*3/>).M M)
-7
)*+,)))-) ' 8 C !
7 6! 6 999

D 8 6 H 8 ! 6 6!
-7
F 8 (6 ! Q

?--J " %
$ 1L)?--J > > ")? G >: >)5 8 >.M M > H 8 >0 > > 6
)

data 06.05.2020
S T [✓ Q] Q S S D

H) Circles Eves/
Howard W. Eves
Obs.: Diretório todo!!!

I) ARQ_IMPORTANT_04/
Géométrie/

a) Jacques Hadamard - Leçons de
géométrie élémentaire I...pdf

b) Jacques Hadamard - Leçons de
géométrie élémentaire II...pdf

J) ARQ_IMPORTANT_05/
Géométrie/
J1) livre-exercices-de-geometrie-
fgm-bnf.pdf

Obs.: 1176 pages.

J2) livre-geometrie-projective-
daniel-perrin...pdf

Obs. 1328 pags, mas este tem 526 páginas.

Material Completo:

Books_80_2020/
Livre_Géométrie_Projective_Daniel_
Perrin_important/

→ todo o diretório!!!

G)*+,)))-) ' "" $ %:"$ 6!
/ 7 ? !
-7 5 8 ! 999

?--J " % $ 1L)?--J > > ")? G >: >).M M > H 8)

)*+,)))-) ' [. A>
@ A ! 2 ? G W 6!

)*+,)))-) '
* > !>. A>>*> !A> > 6 6!

")*+,)))-) ' ? A ! . A> * 3
6 ! < / @ A 2 < < 6!

%)*+,)))-) ' ? ! 6!
-7 / 6 A ! . A

)*+,)))-) ' A . A2 :" $" >
F6 A2 @ A2 ! $$ 2 6 6!

')*+,)))-) ' F ! > !>3 F ! 6!

1)*+,)))-) ' . A / @ A>
F ! 2 @A6 7 2 ! H 8 . :2 6 6!

:)*+,)))-) ' 8 M M \ ! !6
' 6!

$)*+,)))-) ' 6 > 6!
-7 5 8 7 999

)*+,)))-) ' ! M M 6 H 8 6!

)*+,)))-) ' 7 6 8
A 6 1 6!

)*+,)))-) ' C L ! 8
6!

< %# # 1 ": %%
QC & " (":1 ! ! QC = 8
' !

! " " #$ $ % & !! ' ()%
" " " #$ $ * #%
+,-.+ + +/ + !!
0 12" 2 1 1 2 31 2224 5 2 263 !7*!)8 3"

+,-.+ + +/ + !!
0 12" 2 1 1 2 31 2224 5 2 263 !7*) * 3"

6 2/ $& 3 9 2: $
/6
0 " " ; <<<

2/ $0 =- > ? 3 " $
/6 0 " " ; <<<

@//A (# * %B+@//A 2 2 (+@ 7 2) 2 +0 =-
> ? 3 " +

+,-.+ + +/ + !! 0 =- > ? 3 "
3 C *% 3"

+,-.+ + +/ + !! 0 =- > ? 3 "
3 = 3"
"

" ,-.2@, & 2 $
/6 0 " " ; <<<

,-.2@, & $ #$
:1 1 2 D 8 2 3 $
/6 E 0 1 F " 18 $

,-.2@, & $ #$
/-0G/:/=,CH2I,-@/CH2 $
/6
+,-.+ + +/ + !!
/ J2I 6 J2 ? " 6 3"

,-.2@, & $ #$
- : ? " : , ? $

+,-.+ + +/ + !! 4C 8 J 5 KL K
- : ? " : , ? M 3 @ G " 6 3"

6 +,-.+ + +/ + !! - : ? "
: , ? %M 3 & 6 3"

+,-.+ + +/ + !! " 3 3 3 3"
+,-.+ + +/ + !! " 3 3 3 3"

/6 ,-.2@, & $ # N
0 " " ; <<<

,-.2@, & $! N : 8 $
/6 0 " " ; <<<

D ,-.2@, & $! N & $
/6 0 " " ; <<<

D #2 2: ?$ N ?

+,-.+ + +/ + !! J " ? 3"

+,-.+ + +/ + !! J " ? 3"

+,-.+ + +/ + !! " ? (3"

+,-.+ + +/ + !! J " ? # 3"

D6 : 2 2 $

+,-.+ + +/ + !! =!(%)#H 3"

+,-.+ + +/ + !! =!(%#(# 3"

D : 2 2 $

+,-.+ + +/ + !! 2 = (**((3"

+,-.+ + +/ + !! 2 2 = (**%# # 3"

+,-.+ + +/ + !! (2 = (***!) 3"

+,-.+ + +/ + !! #2 = (!) 3"

+,-.+ + +/ + !! 2 = (! !((* 3"

+,-.+ + +/ + !! !2 = (! %#!% 3"

+,-.+ + +/ + !! %2 = (!)%(3"

+,-.+ + +/ + !!)2 = (! % 3"

+,-.+ + +/ + !! *2 = (! H 3"

+,-.+ + +/ + !! 2 2 = (!)!#% 3"

+,-.+ + +/ + !! 2 2 = (! *) ! 3"

+,-.+ + +/ + !! 2 2 = (! * 3"

+,-.+ + +/ + !! 2 2 = (! * 3"

+,-.+ + +/ + !! 2 = (! ! 3"

" G & # $

+,-.+ + +/ + !! " 3"

+,-.+ + +/ + !! " 3"

+,-.+ + +/ + !! " (3"

+,-.+ + +/ + !! " # 3"

7 ,-.2@, & $ %$
0 , : ? $

+,-.+ + +/ + !! 0 @ 1 M = 3 0
, : ? , " 2 & *!%M - " &
6 "D8

+,-.+ + +/ + !! 0 @ 1 M = 3 0
, : ? , " 2 && *!%M - " &
6 "D8

,-.2@, & $)$
@ 7 2 2 $

,-.2@, & $)$
- $

+,-.+ + +/ + !! I M : M 3 " 3 M
, : ? G ? *M 3 3"

6 +,-.+ + +/ + !! > > ? = > "
: ? *M > 3 ? 3"

+,-.+ + +/ + !! O - " ?
" 8 ? %M 3 3"

+,-.+ + +/ + !! O - I : ?
" - 8 ? %M 3 3"

+,-.+ + +/ + !! O - " ?
" 8 ?2 %M 3 3"

" +,-.+ + +/ + !! : ? " 3 3
- M @ 7 P @ 3"

+,-.+ + +/ + !! : ? -
M 6 " C 8 ? 3"

+,-.+ + +/ + !! : ? - "
" 3 *M & 3"

QRSTUVWXYZW[\]^_`abWc^d_Qb\eef\\\geg[[[[[Wh`iRijkdW[ll\\\m^nkaobpqristTXuokjkqbT
3 " 3 ? M > 3 ? 3"

+,-.+ + +/ + !! 3 M 8 J ? "
= 3 8 8 (M > 3 ? 3"

+,-.+ + +/ + !! / G?3 v E
@ : ? !M @ 7 P @ 3"

D +,-.+ + +/ + !! @ ? " : ?2
3 - > !M I 8 6 3"

7 +,-.+ + +/ + !! - 6? v ?
3 J , ? : ? M , ? "D8

+,-.+ + +/ + !! K : I : ?
? " 2O > " - 3"

,-.2@, & $)$
3 20 ?$
/6
0 " " ; <<<

,-.2@, & $)$

+,-.+ + +/ + !! 6 B 3"

+,-.+ + +/ + !! , 3"
/6 & 3 <<<

,-.2@, & $ $
33 $
33 2 $

. ,-.2@, & $ $
3 : ? " & ,33 $

- ,-.2@, & $ $
0 " v 7 $

data 06.05.2020
S T [✓ Q] Q S S D

166) 20200506

A) Books_80_2020/
Géométrie-Chasles_2_Volumes/

a) Traité_de_géométrie_supérieure...
508r.pdf

b) Traité_de_géométrie_supérieure...
529k.pdf

B) PC_Oueto/
Importantíssimo_Gauss/

Obs: Todo diretório !!!

C) PC_oueto/
The VNR Concise Encyclopedia of Mathematics/

Obs: Todo diretório

Copy: Books_80_2020/
The VNR Concise Encyclopedia of Mathematics/
The VNR Concise Encyclopedia of Mathematics-Springer Netherl.pdf.

D) ARQ_BASICS_02/
Pappus

Obs. Todo diretório !!!

06.05.2020

E) ARQ_BASICS/04/
Géométrie_Projective_important/

Obs: gx le théorème de Céva/

F) ARQ_BASICS/04/
ORTHOGONAUX_DARBOUX_images/

G) ARQ_BASICS/04/
Riemannian Geometry and
Geometric Analysis/

a) [Universitext] Jurgen ... Riemannian.pdf
(6th Ed)

b) Riemannian Geometry and Geometric
Analysis (2017...)..pdf
(7th Ed)

H) ARQ_BASICS/04/
Traité des propriétés projectives des
figures_Poncelet/

Obs.: Todo diretório!!!

I) ARQ_BASICS/06/
Gravitation/

Obs.: Todo diretório!!!

J) ARQ_BASICS/07/
Infinitesimal/

Obs.: Todo diretório!!!

data 06 05 2020
S T [✓ Q] Q S S D

a) 04-Volumes_Germany/
→ Cauchy

b) Gauss_2_Volumes/

c) Gauss_12_Volumes/ ✡

d) Houel Infinitesimal 4 Volumes/

K) ARQ_BASICS/07/
The Secrets of Ancient Geometry-
2-Volumes//

L) ARQ_BASICS/08/
Books_25_2020/

M) ARQ_BASICS/08/
Riemann/
a) Dani... Geometry in History...pdf
b) Eighteen Essays in Non-Euclidean
Geometry (2019...)...pdf.
c) From Riemann to differential geometry
and relativity (2017, Springer).pdf
d) Geometry and Spectra of Compact
Riemann Surfaces...pdf
e) Geometry of Riemann Surfaces...pdf
f) Geometry of Riemann surfaces and their
moduli spaces...pdf

06 05 2020

g) Lobachevskii_Pangeometry...pdf

h) Metric Spaces, Convexity and Nonpositive Curvature...pdf

i) On the Hypotheses which Lie at the Bases of Geometry...pdf
→ Riemann

j) Peter Pesic_Beyond Geometry_Classic Papers from Riemann to Einstein ...pdf

k) Riemann Surfaces by Way of Complex Analytic Geometry...djvu

L) Vargas J. G.-Differential Geometry for Physicists and Mathematics...pdf

N) ARQ-BASICS/08/
Spherical_Trigonometry/

obs.: Todo diretório!!!

O) ARQ-BASICS/08/

01) claimofleibnitzt00slomrich.pdf

02) Perfect Ali Pi.pdf

Obs:

P) ARQ-BASICS/10/
Pappus/ and Pappus_1/

data 06.05.2020

Q) ARQ-BASICS/10/
Spherical Geometry and Its Applications/

R) ARQ-BASICS/10/
The Cinderella.2 Manual
Working.../

167) 20200507

Deus é oniexistente (eterno)
onipotente
onisciente
onipresente.

O Universo é parte de Deus, e portanto, o Universo sempre existiu.

O Big Bang é um momento cíclico do Universo.

O Universo é ativo em decorrência das Forças Fundamentais de Interação

Weak
Strong
Electromagnetic

A gravidade não é uma força fundamental, mas é decorrência da interação das 3 forças fundamentais do Universo.

06.05.2020

Q) ARQ_BASICS/10/
Spherical Geometry and Its
Applications/

R) ARQ_BASICS/10/
The Cinderella.2 Manual
Working.../

167) 20200507

Deus é omniexistente (eterno)
omnipotente
omnisciente
omnipresente.

O Universo é parte de Deus,
e portanto, o Universo sempre existiu.

O Big Bang é um momento cíclico
do Universo.

O Universo é ativo em decorrência
dos Forças Fundamentais de Interação

Weak
Strong
Electromagnetic

A gravidade não é uma força
fundamental, mas é decorrência da
interação dos 3 forças fundamentais
do Universo.

data 07.05.2020

Gravitation is by far weakest × of the three interactions at atomic scale, where electromagnetic interactions dominate.

Albert Einstein não conseguiu fazer a unificação das forças fundamentais pois considerou a gravidade como uma força fundamental e não uma consequência das 3 forças fundamentais de interação do Universo.

General relativity generalizes special relativity and refines Newton's law of universal gravitation

↳ descreveu o efeito das 3 forças fundamentais do Universo e não uma força em si mesma (independente), mas decorrentes daquelas.

providing a unified description of gravity as a geometric property of space and time (spacetime).

In particular the curvature of spacetime is directly related to the energy and momentum of whatever (tanto faz) matter and radiation are present.

Light Cone:

Geometric Newton → classical mechanics is a limiting case of special relativity

data 07.05.2020
8

mechanics. In the language of symmetry, where gravity can be neglected, physics is Lorentz invariant (Hendrik Lorentz) as in special relativity rather than Galilei invariant as in classical mechanics.

The defining symmetry of special relativity ~~th~~ is the Poincaré group, which includes translations, rotations and boosts.

The differences between the two become significant when dealing with speeds approaching the speed of light and with high-energy phenomena.

A velocidade da luz é definida como uma constante universal, mas esta não absolutamente constante, pois a sua velocidade depende do meio em que trafega e com base nas 3 leis fundamentais de interações, aplicáveis a curvatura do espaço-tempo.

Voltando (Light cone)

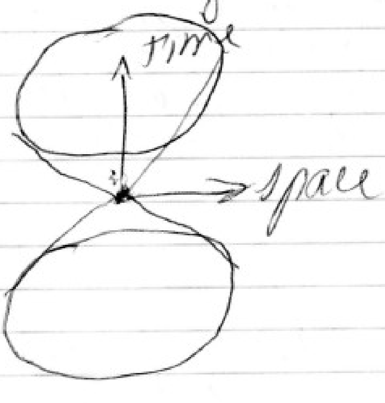

Meu: light cone

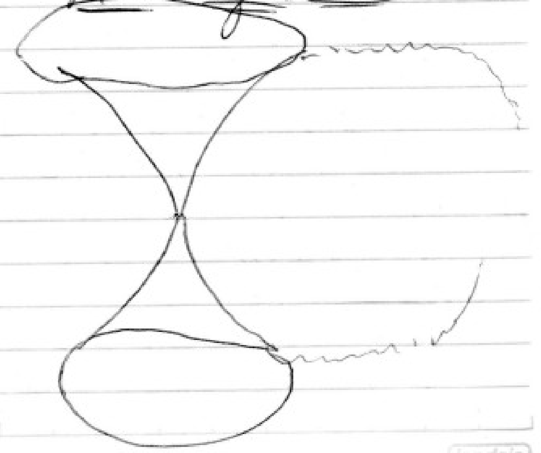

07 05 2020

With Lorentz symmetry additional* structures are defined by the set of light cone. These sets are observer-independent.

In conjunction with the world lines of freely falling particles, the light-cones can be used to reconstruct the space-time's semi-Riemannian metric, at least up to a positive scalar factor.

In mathematical terms, this defines a conformal structure or conformal geometry.

↳ The conformal isometries of a domain therefore (portanto) consist of holomorphic self-maps. In particular, on the conformal compactification - the Riemann Sphere - the conformal transformation are given by the Möbius transformations

(o meu é em 3D
2 cone invertido [illegible]

reta, mas na curvatura tempo-espaço não há uma reta perpendicular ao plano em que houve a transformação.

$$z \rightarrow \frac{az+b}{cz+d}$$

where ad-bc is nonzero.

↳ can be visualized as the complex number plane wrapped around a sphere (by some form of stereographic projection).

Important:

Stereographic projection is conformal, meaning that it preserves the angles at which curves cross each other. Stereographic projection does not preserve area. In general, the area of a region of sphere does not equal the area of its projection onto the plane.

$dA = \frac{4}{(1+X^2+Y^2)^2}\, dX\,dY$. Along the unit circle $X^2+Y^2=1$, there is no inflation of area in the limit, giving a scale factor of 1. Near (0,0) areas are inflated by a factor of 4, and near infinity areas are inflated by arbitrarily small factors. The metric space is given in (X,Y) coordinates by $\frac{4}{(1+X^2+Y^2)^2}(dX^2+dY^2)$, and is the unique formula found in Bernhard Riemann's Habilitationsschrift on the foundations of geometry, delivered at Göttingen in 1854, and entitled
Über die Hypothesen welche der Geometrie zu Grunde liegen.
Copy: ARQ_BASICS_04/DEUS/Geom.pdf
Source: www.maths.tcd.ie/pub/HistMath/People/Riemann/Geom/Geom.pdf.
Source: https://www.emis.de/classics/Riemann/
→ the Mathematical Papers of Georg Friedrich Bernhard Riemann.

No map from the sphere to the plane can be both conformal and area-preserving. If it were, then it would be a local isometry and would preserve Gaussian curvature.
→ is an intrinsic measure of curvature, depending only on ~~th~~ distances that are measured on the surface, not on the way it is isometrically embedded in Euclidean space. This is content of the Theorema egregium (1827).
Copy: ARQ_BASICS_04/DEUS/Gaussian_curvature/
a) Andrew Pressley - Elementary Differential Geometry (2010...)...pdf Pag.247(248).

07 05 2020

b) ARQ-BASICS_04/DEUS/Gaussian-curvature/
cs6.pdf
c) L6_GaussBonnet.pdf
d) CDG.pdf
e) Gauss_Egregium.pdf
f) rutgers-lib-26114_PDF-1.pdf ⇒ Introduction to the Curvature of Surfaces.
g) Palais_Fundamental Theorem.pdf

Total Curvature (Gaussian Curvature) The surface integral of the Gaussian curvature over some region of a surface is called the total curvature. The total curvature of a geodesic triangle equals the deviation of the sum of its angles from π. The sum of the angles of a triangle on a surface of positive curvature will exceed π, while the sum of the angles of a triangle on a surface of negative curvature will be less than π. On a surface of zero curvature, such as the Euclidean plane, the angles will sum to precisely π radians. $\sum_{i=1}^{3} \theta_i = \pi + \iint_T K \, dA$

o). Gauss-Bonnet theorem
cont. ... Gaussian curvature. The sphere and the plane have different Gaussian curvatures, so this is impossible.

Circles on the sphere that do not pass through the point of projection are projected to circles on the plane. Circles on the sphere that do not pass through the point of projection are projected to straight lines on the plane. These lines are sometimes thought of as circles through the point at infinity, or circles of infinite radius. All lines in the plane, when transformed to circles on the sphere by the inverse of stereographic projection, meet at the projection point. Parallel lines, which do not intersect in the plane, are transformed to circles tangent at projection point. Intersecting lines are transformed to circles that

data 07.05.2020

intersect transversally at two points in the sphere one of which is the projection point.

Cross Section: A graph of $z=x^2+xy+y^2$. For the partial derivative at (1,1,3) that leaves y constant, the corresponding tangent line is parallel to the xz plane. Observamos uma curvatura da superfície, compare esta curvatura com as equações dA da Stereographic Projection. ✡.

Voltando: Light cone. e complementando com o Penrose Carter diagram of an infinite MINKOWSKI UNIVERSE / wikipedia.org/wiki/General-relativity

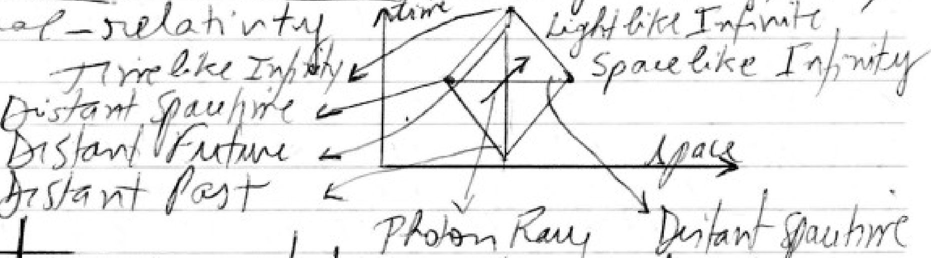
time
Lightlike Infinite
Spacelike Infinity
Timelike Infinity
Distant Spacetime
Distant Future
Distant Past
space
Photon Ray
Distant Spacetime

obs: 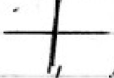 A Cruz neste diagrama é um ângulo reto no seu ponto central no meu caso, já há uma curvatura intrínseca. Todas as linhas são curvas no meu modelo.

Neste diagrama, devemos 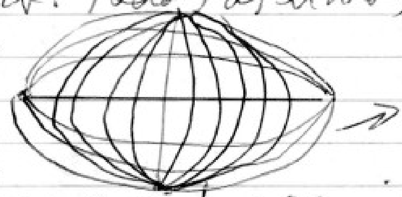 dentro do losângulo.

observar que o ponto inicial de referência deverá ser alterado para o ponto inicial do Big Bang, ou seja para o momento em que ocorreu o Big Bang mas havendo a alteração do modelo utilizado para o desenvolvimento do Universo.

As 3 forças fundamentais de interação do Universo geram como efeito a gravidade, como uma consequência "sine qua non" da curvatura do Universo, gerada em decorrência destas 3 forças fundamentais universais e não em decorrência direta da gravidade, mas indiretamente,

07.05.2020

Speed of light: In 1865, James Clerk Maxwell proposed that light was an electromagnetic wave, and therefore travelled at the speed c appearing in his theory of electromagnetism. In 1905, Albert Einstein postulated that the speed of light c with respect to any inertial frame is a constant and is independent of the motion of the light source. c = celeritas

☆ In 1856, Wilhelm Eduard Weber and Rudolf Kohlrausch had used c for a different constant later shown to equal √2 times the speed of light in vacuum.

Poincaré group: first defined by Hermann Minkowski (1908) as the group of Minkowski spacetime isometries

↳ ignoring the effects of gravity (i.e) Minkowski space → the presence of gravity spacetime is described by a curved 4-dimensional manifold for which the tangent space to any point is a 4-dimensional Minkowski space. Thus, the structure of Minkowski space is still essential in the description of general relativity.

Hyperbolic stereographic projection

" Red circular arc is geodesic in Poincaré disk model; it projects to the brown geodesic on the green hyperboloid.

Curvature: Gaussian Curvature

Formally, Gaussian curvature only depends on the Riemann metric of the surface. This is Gauss's celebrated Theorema Egregium, which he found while concerned with geographic surveys and mapmaking.

Gaussian curvature K at the point P can be computed by the Bertrand–Diguet–Puiseux Theorem as

$$K = \lim_{r \to 0^+} 3\left(\frac{2\pi r - C(r)}{\pi r^3}\right).$$

!) !*

data 07.05.2020
S T Q [✓ Q] S S D

168) Books_80-2020/
  a) Deus/ → referência ao item 167.
  b) Gauss/ → referência ao item 167.
  c) Géométrie_Projective → também se refere 167.
  (contém arquivos básicos com fundamentações),
  d) Livre_Géométrie_Projective_Daniel_Perrin.../
  e) Physics and Mathematics of Quantum Many-Body Systems/
    e1) 3030412644.pdf
    e2) Bohmian mechanics - ~~an~~ the physics and mathematics of quantum theory...pdf
    e3) Einstein's space-time_an introduction to special and general relativity...pdf
    e4) Geometry, topology...pdf
    e5) the Quantum Mechanics Conundrum...pdf
  f) The Problem of the Motion of Bodies/
    3319048392 pdf —

169) 20200507

Arquivos importantes para a fundamentação da pesquisa está no diretório:
ARQ_BASICS_20200504
→ contém 5 arquivos.
Outros: HD_1.36_01
HD_1.36_02
HD-610
HD-620
HD-Geoflex
Pendrive_Geral
Pendrives
QBex (PC)
Terra (1TB-Backup)

202005072146

@ 7 2) 2 $

+x& +@//A (# * %B+@//A 2 2 (+@ 7 2) 2 +

!%

+,-.+ + +/ + !% , " ? > ?
I : ? M 3 3"
' 8 3 " 3

' #% #) : y 0 >

10

Gauss' Theorema Egregium

One of Gauss' most important discoveries about surfaces is that the Gaussian curvature is unchanged when the surface is bent without stretching. Gauss called this result 'egregium', and the Latin word for 'remarkable' has remained attached to his theorem ever since. We shall deduce the Theorema Egregium from two results which relate the first and second fundamental forms of a surface, and which have other important consequences.

10.1 The Gauss and Codazzi–Mainardi equations

It is natural to ask if there are any relations between the first and second fundamental forms of a surface. Note that, by Examples 6.1.2, 6.1.4, 7.1.1 and 7.1.2, the plane and the unit cylinder, when suitably parametrized, have the same first fundamental form but different second fundamental forms, and so the second fundamental form certainly cannot be 'deduced' from the first. Nevertheless, there are some nontrivial relations between the two forms.

Proposition 10.1.1 (Codazzi–Mainardi Equations)

Let

$$Edu^2 + 2Fdudv + Gdv^2 \quad \text{and} \quad Ldu^2 + 2Mdudv + Ndv^2$$

Andrew Pressley, *Elementary Differential Geometry: Second Edition*,
Springer Undergraduate Mathematics Series, DOI 10.1007/978-1-84882-891-9_10,
© Springer-Verlag London Limited 2010

be the first and second fundamental forms of a surface patch $\boldsymbol{\sigma}(u, v)$, and define the Christoffel symbols as in Proposition 7.4.4. Then,

$$\begin{aligned} L_v - M_u &= L\Gamma_{12}^1 + M(\Gamma_{12}^2 - \Gamma_{11}^1) - N\Gamma_{11}^2, \\ M_v - N_u &= L\Gamma_{22}^1 + M(\Gamma_{22}^2 - \Gamma_{12}^1) - N\Gamma_{12}^2. \end{aligned} \tag{10.1}$$

Proposition 10.1.2 (Gauss Equations)

If K is the Gaussian curvature of the surface patch $\boldsymbol{\sigma}(u, v)$ in the preceding proposition, then

$$\begin{aligned} EK &= (\Gamma_{11}^2)_v - (\Gamma_{12}^2)_u + \Gamma_{11}^1\Gamma_{12}^2 + \Gamma_{11}^2\Gamma_{22}^2 - \Gamma_{12}^1\Gamma_{11}^2 - (\Gamma_{12}^2)^2 \\ FK &= (\Gamma_{12}^1)_u - (\Gamma_{11}^1)_v + \Gamma_{12}^2\Gamma_{12}^1 - \Gamma_{11}^2\Gamma_{22}^1 \\ FK &= (\Gamma_{12}^2)_v - (\Gamma_{22}^2)_u + \Gamma_{12}^1\Gamma_{12}^2 - \Gamma_{22}^1\Gamma_{11}^2 \\ GK &= (\Gamma_{22}^1)_u - (\Gamma_{12}^1)_v + \Gamma_{22}^1\Gamma_{11}^1 + \Gamma_{22}^2\Gamma_{12}^1 - (\Gamma_{12}^1)^2 - \Gamma_{12}^2\Gamma_{22}^1. \end{aligned}$$

Proof

We prove both propositions simultaneously. Write down the equation $(\boldsymbol{\sigma}_{uu})_v = (\boldsymbol{\sigma}_{uv})_u$, using Proposition 7.4.4 for $\boldsymbol{\sigma}_{uu}$ and $\boldsymbol{\sigma}_{uv}$:

$$\begin{aligned} &(\Gamma_{11}^1\boldsymbol{\sigma}_u + \Gamma_{11}^2\boldsymbol{\sigma}_v + L\mathbf{N})_v = (\Gamma_{12}^1\boldsymbol{\sigma}_u + \Gamma_{12}^2\boldsymbol{\sigma}_v + M\mathbf{N})_u, \\ \therefore \quad &\left(\frac{\partial\Gamma_{11}^1}{\partial v} - \frac{\partial\Gamma_{12}^1}{\partial u}\right)\boldsymbol{\sigma}_u + \left(\frac{\partial\Gamma_{11}^2}{\partial v} - \frac{\partial\Gamma_{12}^2}{\partial u}\right)\boldsymbol{\sigma}_v + (L_v - M_u)\mathbf{N} \\ &= \Gamma_{12}^1\boldsymbol{\sigma}_{uu} + (\Gamma_{12}^2 - \Gamma_{11}^1)\boldsymbol{\sigma}_{uv} - \Gamma_{11}^2\boldsymbol{\sigma}_{vv} - L\mathbf{N}_v + M\mathbf{N}_u \qquad (10.2) \\ &= \Gamma_{12}^1(\Gamma_{11}^1\boldsymbol{\sigma}_u + \Gamma_{11}^2\boldsymbol{\sigma}_v + L\mathbf{N}) + (\Gamma_{12}^2 - \Gamma_{11}^1)(\Gamma_{12}^1\boldsymbol{\sigma}_u + \Gamma_{12}^2\boldsymbol{\sigma}_v + M\mathbf{N}) \\ &\quad -\Gamma_{11}^2(\Gamma_{22}^1\boldsymbol{\sigma}_u + \Gamma_{22}^2\boldsymbol{\sigma}_v + N\mathbf{N}) - L\mathbf{N}_v + M\mathbf{N}_u, \end{aligned}$$

using Proposition 7.4.4 again. Now, $\mathbf{N}_u$ and $\mathbf{N}_v$ are perpendicular to $\mathbf{N}$, and so are linear combinations of $\boldsymbol{\sigma}_u$ and $\boldsymbol{\sigma}_v$. Hence, equating $\mathbf{N}$ components on both sides of the last equation gives

$$L_v - M_u = L\Gamma_{12}^1 + M(\Gamma_{12}^2 - \Gamma_{11}^1) - N\Gamma_{11}^2,$$

which is the first of the Codazzi–Mainardi equations (10.1). The other equation follows in a similar way by equating coefficients of $\mathbf{N}$ in the equation $(\boldsymbol{\sigma}_{uv})_v = (\boldsymbol{\sigma}_{vv})_u$.

Now we use the formulas in the proof of Proposition 8.1.2 to express $\mathbf{N}_u$ and $\mathbf{N}_v$ in terms of $\boldsymbol{\sigma}_u$ and $\boldsymbol{\sigma}_v$. Equating coefficients of $\boldsymbol{\sigma}_u$ in Eq. 10.2 then gives

$$(\Gamma_{11}^1)_v - (\Gamma_{12}^1)_u = \Gamma_{12}^2\Gamma_{12}^1 - \Gamma_{11}^2\Gamma_{22}^1 + Lc - Ma \tag{10.3}$$

in the notation of the proof of Proposition 8.1.2, from which we find that

$$a = \frac{LG - MF}{EG - F^2}, \quad c = \frac{MG - NF}{EG - F^2}$$

and hence

$$Lc - Ma = \frac{L(MG - NF) - M(LG - MF)}{EG - F^2} = -\frac{F(LN - M^2)}{EG - F^2} = -FK$$

by Corollary 8.1.3. Substituting in Eq. 10.3 and rearranging gives the second of the Gauss equations. The other three are proved in the same way, equating coefficients of $\boldsymbol{\sigma}_v$ in $(\boldsymbol{\sigma}_{uu})_v = (\boldsymbol{\sigma}_{uv})_u$ and those of $\boldsymbol{\sigma}_u$ and $\boldsymbol{\sigma}_v$ in $(\boldsymbol{\sigma}_{uv})_v = (\boldsymbol{\sigma}_{vv})_u$. □

The following theorem tells us that there are no other relations between the first and second fundamental forms other than those in Propositions 10.1.1 and 10.1.2.

Theorem 10.1.3

Let $\boldsymbol{\sigma} : U \to \mathbb{R}^3$ and $\tilde{\boldsymbol{\sigma}} : U \to \mathbb{R}^3$ be surface patches with the same first and second fundamental forms. Then, there is a direct isometry M of $\mathbb{R}^3$ such that $\tilde{\boldsymbol{\sigma}} = M(\boldsymbol{\sigma})$.

Moreover, let V be an open subset of $\mathbb{R}^3$ and let E, F, G, L, M and N be smooth functions on V. Assume that $E > 0, G > 0, EG - F^2 > 0$ and that the equations in Propositions 10.1.1 and 10.1.2 hold, with $K = \frac{LN - M^2}{EG - F^2}$ and the Christoffel symbols defined as in Proposition 7.4.4. Then, if $(u_0, v_0) \in V$, there is an open set U contained in V and containing (u_0, v_0), and a surface patch $\boldsymbol{\sigma} : U \to \mathbb{R}^3$, such that $Edu^2 + 2Fdudv + Gdv^2$ and $Ldu^2 + 2Mdudv + Ndv^2$ are the first and second fundamental forms of $\boldsymbol{\sigma}$, respectively.

This theorem is an analogue for surfaces of Theorem 2.2.5, which shows that unit-speed plane curves are determined up to a direct isometry of $\mathbb{R}^3$ by their signed curvature. We shall not prove Theorem 10.1.3 here. The first part depends on uniqueness theorems for the solution of systems of ordinary differential equations, and is not particularly difficult. The second part is more sophisticated and depends on existence theorems for the solution of certain *partial* differential equations. The following example illustrates what is involved.

Example 10.1.4

Consider the first and second fundamental forms $du^2 + dv^2$ and $-du^2$, respectively. Let us first see whether a surface patch with these first and second fundamental forms exists. Since all the coefficients of these forms are constant, all the Christoffel symbols are zero and the Codazzi–Mainardi equations are obviously satisfied. The first formula in Proposition 10.1.2 gives $K = 0$, so the only other condition to be checked is $LN - M^2 = 0$, and this clearly holds since $M = N = 0$. Theorem 10.1.3 therefore tells us that a surface patch with the given first and second fundamental forms exists.

To find it, we note that the Gauss equations give

$$\boldsymbol{\sigma}_{uu} = -\mathbf{N}, \quad \boldsymbol{\sigma}_{uv} = \mathbf{0}, \quad \boldsymbol{\sigma}_{vv} = \mathbf{0}.$$

The last two equations tell us that $\boldsymbol{\sigma}_v$ is a constant vector, say $\mathbf{a}$, so

$$\boldsymbol{\sigma}(u, v) = \mathbf{b}(u) + \mathbf{a}v, \tag{10.4}$$

where $\mathbf{b}$ is a function of u only. The first equation then gives $\mathbf{N} = -\mathbf{b}''$ (a dash denoting d/du). We now need to use the expressions for $\mathbf{N}_u$ and $\mathbf{N}_v$ in terms of $\boldsymbol{\sigma}_u$ and $\boldsymbol{\sigma}_v$ in the proof of Proposition 8.1.2. The matrix of the Weingarten map with respect to the basis $\{\boldsymbol{\sigma}_u, \boldsymbol{\sigma}_v\}$ of the tangent plane is

$$\mathcal{F}_I^{-1}\mathcal{F}_{II} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix},$$

so Proposition 8.1.2 gives

$$\mathbf{N}_u = \boldsymbol{\sigma}_u, \quad \mathbf{N}_v = \mathbf{0}.$$

The second equation tells us nothing new, since we already know that $\mathbf{N} = -\mathbf{b}''$ depends only on u. The first equation gives

$$\mathbf{b}''' + \mathbf{b}' = \mathbf{0}.$$

Hence, $\mathbf{b}'' + \mathbf{b}$ is a constant vector, which we can take to be zero by applying a translation to $\boldsymbol{\sigma}$ (see Eq. 10.4). Then,

$$\mathbf{b}(u) = \mathbf{c}\cos u + \mathbf{d}\sin u,$$

where $\mathbf{c}$ and $\mathbf{d}$ are constant vectors, and $\mathbf{N} = -\mathbf{b}'' = \mathbf{b}$. This must be a unit vector for all values of u. It is easy to see that this is possible only if $\mathbf{c}$ and $\mathbf{d}$ are perpendicular unit vectors, in which case we can arrange that $\mathbf{c} = (1, 0, 0)$ and $\mathbf{d} = (0, 1, 0)$ by applying an isometry of $\mathbb{R}^3$, giving $\mathbf{b}(u) = (\cos u, \sin u, 0)$. Finally, $\boldsymbol{\sigma}_u \times \boldsymbol{\sigma}_v = \lambda\mathbf{N}$ for some non-zero scalar λ, so $\mathbf{b}' \times \mathbf{a} = \lambda\mathbf{b}$. This forces $\mathbf{a} = (0, 0, \lambda)$, and the patch is given by

$$\boldsymbol{\sigma}(u, v) = (\cos u, \sin u, \lambda v),$$

a parametrization of the unit cylinder (which the reader had probably guessed some time ago).

EXERCISES

10.1.1 A surface patch has first and second fundamental forms

$$\cos^2 v\,du^2 + dv^2 \quad \text{and} \quad -\cos^2 v\,du^2 - dv^2,$$

respectively. Show that the surface is an open subset of a sphere of radius one. Write down a parametrization of S^2 with these first and second fundamental forms.

10.1.2 Show that there is no surface patch whose first and second fundamental forms are

$$du^2 + \cos^2 u\,dv^2 \quad \text{and} \quad \cos^2 u\,du^2 + dv^2,$$

respectively.

10.1.3 Suppose that a surface patch $\boldsymbol{\sigma}(v, w)$ has first and second fundamental forms

$$\frac{dv^2 + dw^2}{w^2} \quad \text{and} \quad L dv^2 + N dw^2,$$

respectively, where $w > 0$. Prove that L and N do not depend on v, that $LN = -1/w^4$ and that

$$Lw^5 \frac{dL}{dw} = 1 - L^2 w^4.$$

Solve this equation for L and deduce that $\boldsymbol{\sigma}$ cannot be defined in the whole of the half-plane $w > 0$. Compare the discussion of the pseudosphere in Example 9.3.3.

10.1.4 Suppose that the first and second fundamental forms of a surface patch are $Edu^2 + Gdv^2$ and $Ldu^2 + Ndv^2$, respectively. Show that the Codazzi–Mainardi equations reduce to

$$L_v = \frac{1}{2}E_v\left(\frac{L}{E} + \frac{N}{G}\right), \quad N_u = \frac{1}{2}G_u\left(\frac{L}{E} + \frac{N}{G}\right).$$

Deduce that the principal curvatures $\kappa_1 = L/E$ and $\kappa_2 = N/G$ satisfy the equations

$$(\kappa_1)_v = \frac{E_v}{2E}(\kappa_2 - \kappa_1), \quad (\kappa_2)_u = \frac{G_u}{2G}(\kappa_1 - \kappa_2).$$

10.2 Gauss' remarkable theorem

We noted after Proposition 7.4.4 that the Christoffel symbols depend only on the coefficients of the *first* fundamental form. It follows from the formulas in Proposition 10.1.2 that this is also true of the Gaussian curvature K. In view of Corollary 6.2.3, we obtain:

Theorem 10.2.1 (Gauss' Theorema Egregium)

The Gaussian curvature of a surface is preserved by local isometries.

This means, more explicitly, that if $\mathcal{S}_1$ and $\mathcal{S}_2$ are two surfaces and $f : \mathcal{S}_1 \to \mathcal{S}_2$ is a local isometry, then for any point $\mathbf{p} \in \mathcal{S}_1$ the Gaussian curvature of $\mathcal{S}_1$ at $\mathbf{p}$ is equal to that of $\mathcal{S}_2$ at $f(\mathbf{p})$. The theorem is sometimes expressed by saying that the Gaussian curvature is an *intrinsic* property of a surface, for it implies that the Gaussian curvature could be measured by a bug living in the surface.

By substituting into the equations in Proposition 10.1.2 the expressions for the Christoffel symbols in Proposition 7.4.4, we can of course find an explicit expression for K in terms of E, F and G. (It appears that we could get four such formulas, one from each of the equations in Proposition 10.1.2, but these turn out to be the same.) Here is the result.

Corollary 10.2.2

The Gaussian curvature is given by

$$K = \frac{\begin{vmatrix} -\frac{1}{2}E_{vv} + F_{uv} - \frac{1}{2}G_{uu} & \frac{1}{2}E_u & F_u - \frac{1}{2}E_v \\ F_v - \frac{1}{2}G_u & E & F \\ \frac{1}{2}G_v & F & G \end{vmatrix} - \begin{vmatrix} 0 & \frac{1}{2}E_v & \frac{1}{2}G_u \\ \frac{1}{2}E_v & E & F \\ \frac{1}{2}G_u & F & G \end{vmatrix}}{(EG - F^2)^2}.$$

We shall not go through the details of this calculation, partly because the algebra is tedious, and partly because the following special cases are often all that is needed.

Corollary 10.2.3

(i) If $F = 0$, we have

$$K = -\frac{1}{2\sqrt{EG}} \left\{ \frac{\partial}{\partial u} \left(\frac{G_u}{\sqrt{EG}} \right) + \frac{\partial}{\partial v} \left(\frac{E_v}{\sqrt{EG}} \right) \right\}.$$

(ii) If $E = 1$ and $F = 0$, we have

$$K = -\frac{1}{\sqrt{G}}\frac{\partial^2\sqrt{G}}{\partial u^2}.$$

Proof

If $F = 0$ Proposition 7.4.4 gives

$$\Gamma_{11}^1 = \frac{E_u}{2E},\ \Gamma_{11}^2 = -\frac{E_v}{2G},\ \Gamma_{12}^1 = \frac{E_v}{2E},\ \Gamma_{12}^2 = \frac{G_u}{2G},\ \Gamma_{22}^1 = -\frac{G_u}{2E},\ \Gamma_{22}^2 = \frac{G_v}{2G}.$$

Substituting into the first formula in Proposition 10.1.2 gives

$$\begin{aligned} EK &= -\left(\frac{E_v}{2G}\right)_v - \left(\frac{G_u}{2G}\right)_u + \frac{E_uG_u}{4EG} - \frac{E_vG_v}{4G^2} + \frac{E_v^2}{4EG} - \frac{G_u^2}{4G^2} \\ \text{i.e.,}\quad -2K\sqrt{EG} &= \frac{E_{vv}+G_{uu}}{(EG)^{1/2}} - \frac{E_v(EG_v+E_vG)}{2(EG)^{3/2}} - \frac{G_u(E_uG+EG_u)}{2(EG)^{3/2}} \\ &= \frac{E_{vv}}{(EG)^{1/2}} - \frac{1}{2}\frac{E_v(EG)_v}{(EG)^{3/2}} + \frac{G_{uu}}{(EG)^{1/2}} - \frac{1}{2}\frac{G_u(EG)_u}{(EG)^{3/2}} \\ &= \left(\frac{E_v}{(EG)^{1/2}}\right)_v + \left(\frac{G_u}{(EG)^{1/2}}\right)_u, \end{aligned}$$

proving the equation in (i). If, in addition, $E = 1$, the second term on the right-hand side of the formula in (i) vanishes, so

$$K = -\frac{1}{2\sqrt{G}}\frac{\partial}{\partial u}\left(\frac{G_u}{\sqrt{G}}\right) = -\frac{1}{\sqrt{G}}\frac{\partial^2\sqrt{G}}{\partial u^2}. \qquad \square$$

Example 10.2.4

For the surface of revolution

$$\boldsymbol{\sigma}(u,v) = (f(u)\cos v, f(u)\sin v, g(u)),$$

where $f > 0$ and $\dot{f}^2 + \dot{g}^2 = 1$ (a dot denoting d/du), we found in Example 6.1.3 that $E = 1$, $F = 0$, and $G = f(u)^2$. Hence, Corollary 10.2.3(ii) applies and gives

$$K = -\frac{1}{\sqrt{G}}\frac{\partial^2\sqrt{G}}{\partial u^2} = -\frac{\ddot{f}}{f},$$

in agreement with Example 8.1.4.

The Theorema Egregium provides a necessary condition for the existence of a local isometry between surfaces: if such a local isometry exists, the Gaussian curvature must be the same at corresponding points of the two surfaces. We give two examples of this idea; others can be found in the exercises.

Our first result shows that it is impossible to draw a 'perfect' map of the Earth (which is why cartography is an interesting subject).

Proposition 10.2.5

Any map of any region of the earth's surface must distort distances.

Proof

A map of a region of the earth's surface which did *not* distort distances would be a diffeomorphism from this region of a sphere to a region in a plane (the map) which multiplied all distances by the same constant factor, say C. We might as well assume that the plane passes through the origin. Then, by composing this map with the map $\mathbf{r} \mapsto C^{-1}\mathbf{r}$ from the plane to itself, we would get an *isometry* between this region of the sphere and some region of a plane. This would imply, by the Theorema Egregium, that these regions of the sphere and the plane have the same Gaussian curvature. But we know that a plane has Gaussian curvature zero everywhere, and a sphere has constant positive Gaussian curvature everywhere (if the sphere has radius R, the Gaussian curvature is $1/R^2$). So no such isometry can exist. □

Note, on the other hand, that it *is* possible to draw a map of the Earth that correctly represents angles, and a map that correctly represents areas, for we saw in Example 6.3.5 that the stereographic projection is conformal, and Archimedes' Theorem 6.4.6 shows that there is a map that correctly represents areas. However, it is not possible to represent both angles and areas correctly with the same map (Exercise 6.4.5(ii)).

Our next example shows how the Theorema Egregium can sometimes be used to determine all the isometries of a surface.

Proposition 10.2.6

The only isometries of a helicoid (Exercise 4.2.6)

$$\boldsymbol{\sigma}(u, v) = (u \cos v, u \sin v, v)$$

are S_λ, $R_x \circ S_\lambda$, $R_y \circ S_\lambda$ and $R_z \circ S_\lambda$ for some value of λ, where S_λ is the screwing motion $\boldsymbol{\sigma}(u, v) \mapsto \boldsymbol{\sigma}(u, v + \lambda)$, and R_x, R_y and R_z are rotations by π around the x-, y- and z-axes.

Proof

Suppose that an isometry of the helicoid takes $\boldsymbol{\sigma}(u,v)$ to $\boldsymbol{\sigma}(\tilde{u},\tilde{v})$, where $\tilde{u}$ and $\tilde{v}$ are smooth functions of u and v. Since the Gaussian curvature at $\boldsymbol{\sigma}(u,v)$ is $-1/(1+u^2)^2$ (see Exercise 8.1.2), the Theorema Egregium tells us that

$$\frac{-1}{(1+u^2)^2} = \frac{-1}{(1+\tilde{u}^2)^2},$$

so $\tilde{u} = \pm u$. Applying a rotation R_z by π around the z-axis changes u to $-u$ (and fixes v), so we assume that $\tilde{u} = u$. Let $\tilde{v} = f(u,v)$. By Corollary 6.2.3, the patches $\boldsymbol{\sigma}(u,v)$ and $\tilde{\boldsymbol{\sigma}}(u,v) = \boldsymbol{\sigma}(u, f(u,v))$ have the same first fundamental form. That of $\boldsymbol{\sigma}$ is $du^2 + (1+u^2)dv^2$, and that of $\tilde{\boldsymbol{\sigma}}$ is found to be

$$(1+(1+u^2)f_u^2)du^2 + 2(1+u^2)f_u f_v du dv + (1+u^2)f_v^2 dv^2.$$

Equating these, we find that $f_u = 0$ and $f_v = \pm 1$. Hence,

$$\tilde{v} = f(u,v) = \pm v + \lambda,$$

where λ is a constant. A rotation R_x by π around the x-axis changes v to $-v$ (and fixes u), so we take the $+$ sign. This gives the isometry

$$S_\lambda : \boldsymbol{\sigma}(u,v) \mapsto \boldsymbol{\sigma}(u, v+\lambda).$$

This proves the proposition (the isometry $R_y \circ S_\lambda$ arises because $R_y = R_x \circ R_z$). □

EXERCISES

10.2.1 Show that if a surface patch has first fundamental form $e^\lambda(du^2+dv^2)$, where λ is a smooth function of u and v, its Gaussian curvature K satisfies

$$\Delta\lambda + 2Ke^\lambda = 0,$$

where Δ denotes the Laplacian $\partial^2/\partial u^2 + \partial^2/\partial v^2$.

10.2.2 With the notation of Exercise 9.5.1, define $u = r\cos\theta$, $v = r\sin\theta$, and let $\tilde{\boldsymbol{\sigma}}(u,v)$ be the corresponding reparametrization of $\boldsymbol{\sigma}$. It can be shown that $\tilde{\boldsymbol{\sigma}}$ is an allowable surface patch for $\mathcal{S}$ defined on the open set $u^2+v^2 < \epsilon^2$. (Note that this is not quite obvious because $\boldsymbol{\sigma}$ is not allowable when $r = 0$.)

(i) Show that the first fundamental form of $\tilde{\boldsymbol{\sigma}}$ is $\tilde{E}du^2 + 2\tilde{F}dudv + \tilde{G}dv^2$, where

$$\tilde{E} = \frac{u^2}{r^2} + \frac{Gv^2}{r^4}, \quad \tilde{F} = \left(1 - \frac{G}{r^2}\right)\frac{uv}{r^2}, \quad \tilde{G} = \frac{v^2}{r^2} + \frac{Gu^2}{r^4}.$$

(ii) Show that $u^2(\tilde{E}-1) = v^2(\tilde{G}-1)$, and by considering the Taylor expansions of $\tilde{E}$ and $\tilde{G}$ about $u = v = 0$, deduce that

$$G(r, \theta) = r^2 + kr^4 + \text{remainder}$$

for some constant k, where remainder$/r^4$ tends to zero as r tends to zero.

(iii) Show that $k = -K(P)/3$, where $K(P)$ is the value of the Gaussian curvature of $\mathcal{S}$ at P.

10.2.3 With the notation of Exercises 9.5.1 and 10.2.2, show that:

(i) The circumference of the geodesic circle with centre P and radius R is

$$C_R = 2\pi R \left(1 - \frac{K(P)}{6} R^2 + \text{remainder}\right),$$

where remainder$/R^2$ tends to zero as R tends to zero.

(ii) The area inside the geodesic circle in (i) is

$$A_R = \pi R^2 \left(1 - \frac{K(P)}{12} R^2 + \text{remainder}\right),$$

where the remainder satisfies the same condition as in (i).

Verify that these formulas are consistent with those in spherical geometry obtained in Exercise 6.5.3.

10.2.4 Let A, B and C be the vertices of a triangle $\mathcal{T}$ on a surface $\mathcal{S}$ whose sides are arcs of geodesics, and let α, β and γ be its internal angles (so that α is the angle at A, etc.). Assume that the triangle is contained in a geodesic patch $\boldsymbol{\sigma}$ as in Exercise 9.5.1 with $P = A$. Thus, with the notation in that exercise, if we take $\mathbf{v}$ to be tangent at A to the side passing through A and B, then the sides meeting at A are the parameter curves $\theta = 0$ and $\theta = \alpha$, and the remaining side can be parametrized by $\boldsymbol{\gamma}(\theta) = \boldsymbol{\sigma}(f(\theta), \theta)$ for some smooth function f and $0 \leq \theta \leq \alpha$.

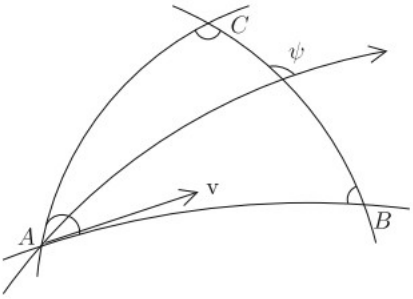

(i) Use the geodesic equations (9.2) to show that

$$f'' - \frac{f'\lambda'}{\lambda^2} = \frac{1}{2}\frac{\partial G}{\partial r},$$

where a dash denotes $d/d\theta$ and $\lambda = \parallel \boldsymbol{\gamma}' \parallel$.

(ii) Show that, if $\psi(\theta)$ is the angle between $\boldsymbol{\sigma}_r$ and the tangent vector to the side opposite A at $\boldsymbol{\gamma}(\theta)$, then

$$\psi'(\theta) = -\frac{\partial\sqrt{G}}{\partial r}(f(\theta), \theta).$$

(iii) Show that, if K is the Gaussian curvature of $\mathcal{S}$,

$$\int_{\mathcal{T}} K \, d\mathcal{A}_{\boldsymbol{\sigma}} = \alpha + \beta + \gamma - \pi.$$

This result will be generalized in Corollary 13.2.3.

10.2.5 Show that the Gaussian curvature of the Möbius band in Example 4.5.3 is equal to $-1/4$ everywhere along its median circle. Deduce that this Möbius band *cannot* be constructed by taking a strip of paper and joining the ends together with a half-twist. (The analytic description of the 'cut and paste' Möbius band is more complicated than the version in Example 4.5.3.)

10.2.6 Show that the only isometries from the catenoid to itself are products of rotations around its axis, reflections in planes containing the axis, and reflection in the plane containing the waist of the catenoid.

10.3 Surfaces of constant Gaussian curvature

The results of the preceding two sections enable us to gain a good understanding of surfaces of constant Gaussian curvature. We begin with the following local description.

Theorem 10.3.1

Any point of a surface of constant Gaussian curvature is contained in a patch that is isometric to an open subset of a plane, a sphere or a pseudosphere.

Proof

Let $\mathbf{p}$ be a point on a surface $\mathcal{S}$ with constant Gaussian curvature K. By applying a dilation of $\mathbb{R}^3$ (see Exercise 8.1.5), we need only consider the cases $K = 0, 1$ and -1.

We take a geodesic patch $\boldsymbol{\sigma}(u, v)$ with $\boldsymbol{\sigma}(0, 0) = \mathbf{p}$. Writing $g = \sqrt{G}$, the first fundamental form is

$$du^2 + g(u, v)^2 dv^2.$$

By Corollary 10.2.3(ii),

$$\frac{\partial^2 g}{\partial u^2} + Kg = 0. \tag{10.5}$$

Note that

$$g(0, v) = 1, \quad g_u(0, v) = 0, \tag{10.6}$$

by Proposition 9.5.1.

If $K = 0$, the solution of Eq. 10.5 is $g(u, v) = \alpha u + \beta$, where α and β are smooth functions of v only. The initial conditions (10.6) give $\alpha = 0$, $\beta = 1$, so $g = 1$ and the first fundamental form of $\boldsymbol{\sigma}$ is $du^2 + dv^2$. This is the same as the first fundamental form of the usual parametrization of the plane (see Example 6.1.2), and Corollary 6.2.3 now shows that $\boldsymbol{\sigma}$ is isometric to an open subset of the plane.

If $K = 1$, the general solution of Eq. 10.5 is $g = \alpha \cos u + \beta \sin u$, where α and β only depend on v. The boundary conditions (10.6) give $\alpha = 1$, $\beta = 0$, and the first fundamental form of $\boldsymbol{\sigma}$ is $du^2 + \cos^2 u \, dv^2$. This is the first fundamental form of S^2, with u and v being latitude and longitude, respectively (see Example 6.1.3). Hence, $\boldsymbol{\sigma}$ is isometric to an open subset of S^2.

Finally, if $K = -1$, we find in the same way that the first fundamental form of $\boldsymbol{\sigma}$ is

$$du^2 + \cosh^2 u \, dv^2.$$

We have not encountered this first fundamental form before. However, let us reparametrize $\boldsymbol{\sigma}$ by defining

$$V = e^v \tanh u, \quad W = e^v \operatorname{sech} u.$$

We then find, using the formulas in Exercise 6.1.4, that the first fundamental form becomes

$$\frac{dV^2 + dW^2}{W^2}.$$

Comparing with Exercise 8.3.1, we see that this is the first fundamental form of a parametrization of the pseudosphere. $\square$

The next result gives a much more precise, though still local, description of surfaces of constant negative Gaussian curvature.

Proposition 10.3.2

Let $\mathcal{S}$ be a surface of constant Gaussian curvature -1. If $\mathbf{p} \in \mathcal{S}$, there is a surface patch of $\mathcal{S}$ containing $\mathbf{p}$ whose first and second fundamental forms are

$$du^2 + 2\cos\theta\, du dv + dv^2 \quad \text{and} \quad 2\sin\theta\, du dv, \tag{10.7}$$

respectively, where $\theta(u, v)$ is a smooth function such that $0 < \theta < \pi$ for all u, v.

Thus, the parameter curves of this parametrization form a *Chebyshev net* (see Exercises 6.1.9 and 7.4.6).

Proof

Since K is negative the surface $\mathcal{S}$ has no umbilics. Hence, there is a surface patch $\boldsymbol{\sigma}(U, V)$ containing $\mathbf{p}$ whose first and second fundamental forms are

$$EdU^2 + GdV^2 \quad \text{and} \quad LdU^2 + NdV^2,$$

respectively (Proposition 8.4.1). The principal curvatures are

$$\kappa_1 = \frac{L}{E}, \quad \kappa_2 = \frac{N}{G}. \tag{10.8}$$

Since $\kappa_1\kappa_2 = -1$, there are two cases:

(i) $\kappa_1 > 0$ and $\kappa_2 < 0$ for all U, V

(ii) $\kappa_1 < 0$ and $\kappa_2 > 0$ for all U, V

We shall concentrate on case (i), indicating briefly the modifications necessary for case (ii).

In case (i), there is a smooth function $\omega(U, V)$ such that $0 < \omega < \pi/2$ and

$$\kappa_1 = \tan\omega, \quad \kappa_2 = -\cot\omega. \tag{10.9}$$

(In case (ii), we put $\kappa_1 = -\cot\omega$, $\kappa_2 = \tan\omega$.) The first Codazzi-Mainardi equation is

$$L_V = \frac{1}{2}E_V\left(\frac{L}{E} + \frac{N}{G}\right)$$

(cf. Exercise 10.1.4). From Eqs. 10.8 and 10.9, we get

$$\begin{aligned}(E\tan\omega)_V &= \frac{1}{2}E_V(\tan\omega - \cot\omega),\\ \text{i.e.,}\quad \omega_V &= -\frac{E_V}{2E}\cot\omega.\end{aligned} \tag{10.10}$$

Then,

$$\left(\frac{\cos\omega}{\sqrt{E}}\right)_V = -\frac{\omega_V \sin\omega}{\sqrt{E}} - \frac{E_V \cos\omega}{2E^{3/2}} = -\frac{\sin\omega}{\sqrt{E}}\left(\omega_V + \frac{E_V}{2E}\cot\omega\right) = 0$$

by Eq. 10.10. Hence, there is a positive smooth function $e(U)$ such that

$$E = e(U)\cos^2\omega.$$

Similarly, the second Codazzi-Mainardi equation implies that there is a positive smooth function $g(V)$ such that

$$G = g(V)\sin^2\omega.$$

(In case (ii), we get $E = e(U)\sin^2\omega$, $G = g(V)\cos^2\omega$.)

Let $\tilde{U}(U)$ and $\tilde{V}(V)$ be smooth functions such that

$$\frac{d\tilde{U}}{dU} = \sqrt{e(U)}, \quad \frac{d\tilde{V}}{dV} = \sqrt{g(V)},$$

and let $\tilde{\boldsymbol{\sigma}}(\tilde{U},\tilde{V})$ be the corresponding reparametrization of $\boldsymbol{\sigma}(U,V)$. Then, the first fundamental form of $\tilde{\boldsymbol{\sigma}}$ is

$$\cos^2\omega\, d\tilde{U}^2 + \sin^2\omega\, d\tilde{V}^2$$

and its second fundamental form is

$$\frac{L}{e}d\tilde{U}^2 + \frac{N}{g}d\tilde{V}^2 = \sin\omega\cos\omega(d\tilde{U}^2 - d\tilde{V}^2).$$

Setting $u = (\tilde{U}+\tilde{V})/2$, $v = (\tilde{U}-\tilde{V})/2$ gives a reparametrization of $\tilde{\boldsymbol{\sigma}}$ whose first and second fundamental forms are as stated in the proposition, with $\theta = 2\omega$. (In case (ii) we set $u = (\tilde{U}+\tilde{V})/2$, $v = (\tilde{V}-\tilde{U})/2$.) □

The function $\theta(u,v)$ appearing in (10.7) is not arbitrary, as the Gauss equations (Proposition 10.1.2) must still be satisfied. As we remarked earlier, these four equations are equivalent to each other, so we need only consider one of them. The Christoffel symbols are found to be

$$\begin{aligned} \Gamma_{11}^1 = \theta_u\cot\theta,\ \ \Gamma_{12}^1 &= 0,\ \ \Gamma_{22}^1 = -\theta_v\operatorname{cosec}\theta, \\ \Gamma_{11}^2 = -\theta_u\operatorname{cosec}\theta,\ \ \Gamma_{12}^2 &= 0,\ \ \Gamma_{22}^2 = \theta_v\cot\theta. \end{aligned}$$

Substituting in the first of the Gauss equations gives

$$\begin{aligned} -1 &= (-\theta_u\operatorname{cosec}\theta)_v - \theta_u\theta_v\operatorname{cosec}\theta\cot\theta, \\ \text{i.e.,}\qquad \theta_{uv} &= \sin\theta. \end{aligned} \tag{10.11}$$

This is called the *sine-Gordon equation*; it arises in fluid mechanics and various other areas of physics.

In view of Theorem 10.1.3, we can summarize our discussion as follows.

Theorem 10.3.3

Let $\theta(u,v)$ be a smooth solution of the sine-Gordon equation (10.11) such that $0 < \theta < \pi$. Then there exists a surface, unique up to a direct isometry of $\mathbb{R}^3$, with constant Gaussian curvature -1 and first and second fundamental forms given by (10.7). Conversely, any surface of constant Gaussian curvature -1 has a parametrization with first and second fundamental forms (10.7) in which θ is a solution of the sine-Gordon equation.

We conclude this section with a beautiful global result that characterizes the *compact* surfaces of constant Gaussian curvature. Note that, by Proposition 8.6.1, the value of the constant Gaussian curvature in this case must be > 0.

Theorem 10.3.4

Every connected compact surface whose Gaussian curvature is constant is a sphere.

The proof of this theorem depends on the following lemma.

Lemma 10.3.5

Let $\boldsymbol{\sigma} : U \to \mathbb{R}^3$ be a surface patch containing a point $\mathbf{p}$ that is not an umbilic. Let $\kappa_1 \geq \kappa_2$ be the principal curvatures of $\boldsymbol{\sigma}$ and suppose that κ_1 has a local maximum at $\mathbf{p}$ and κ_2 has a local minimum there. Then, the Gaussian curvature of $\boldsymbol{\sigma}$ at $\mathbf{p}$ is ≤ 0.

Proof 10.3.5 Since $\mathbf{p}$ is not an umbilic, $\kappa_1 > \kappa_2$ at $\mathbf{p}$, so by shrinking U if necessary, we may assume that $\kappa_1 > \kappa_2$ everywhere.

By Proposition 8.4.1, we can assume that the first and second fundamental forms of $\boldsymbol{\sigma}$ are

$$Edu^2 + Gdv^2 \quad \text{and} \quad Ldu^2 + Ndv^2,$$

respectively. By Exercise 10.1.4,

$$E_v = -\frac{2E}{\kappa_1 - \kappa_2}(\kappa_1)_v, \quad G_u = \frac{2G}{\kappa_1 - \kappa_2}(\kappa_2)_u,$$

and by Corollary 10.2.3(i), the Gaussian curvature

$$K = -\frac{1}{2\sqrt{EG}}\left(\frac{\partial}{\partial u}\left(\frac{G_u}{\sqrt{EG}}\right) + \frac{\partial}{\partial v}\left(\frac{E_v}{\sqrt{EG}}\right)\right).$$

Since $\mathbf{p}$ is a stationary point of κ_1 and κ_2, we have $(\kappa_1)_v = (\kappa_2)_u = 0$, and hence $E_v = G_u = 0$, at $\mathbf{p}$. Hence, at $\mathbf{p}$,

$$K = -\frac{1}{2EG}(G_{uu} + E_{vv}) = -\frac{1}{2EG}\left(\frac{2G}{\kappa_1 - \kappa_2}(\kappa_2)_{uu} - \frac{2E}{\kappa_1 - \kappa_2}(\kappa_1)_{vv}\right)$$

(again dropping terms involving E_v, G_u and the first derivatives of κ_1 and κ_2). Since κ_1 has a local maximum at $\mathbf{p}$, $(\kappa_1)_{vv} \leq 0$ there, and since κ_2 has a local minimum at $\mathbf{p}$, $(\kappa_2)_{uu} \geq 0$ there. Hence, the last equation shows that $K \leq 0$ at $\mathbf{p}$. □

Proof 10.3.4 The proof of Theorem 10.3.4 uses a little point set topology. We consider the continuous function on the surface $\mathcal{S}$ given by $J = (\kappa_1 - \kappa_2)^2$, where κ_1 and κ_2 are the principal curvatures. Note that this function is well defined even though κ_1 and κ_2 are not, partly because we do not know which principal curvature is to be called κ_1 and which κ_2, and partly because the principal curvatures are only well defined up to sign unless $\mathcal{S}$ has a definite orientation. We shall prove that this function is identically zero on $\mathcal{S}$, so that every point of $\mathcal{S}$ is an umbilic. Since the Gaussian curvature $K > 0$, it follows from Proposition 8.2.9 that $\mathcal{S}$ is an open subset of a sphere, say $\mathbf{S}$. In fact, $\mathcal{S}$ must be the whole of $\mathbf{S}$. For, since $\mathcal{S}$ is compact, it is necessarily a closed subset of $\mathbb{R}^3$, and hence a *closed* subset of $\mathbf{S}$. But since $\mathbf{S}$ is connected, the only non-empty subset of $\mathbf{S}$ that is both open and closed is $\mathbf{S}$ itself.

Suppose then, to get a contradiction, that J is not identically zero on $\mathcal{S}$. Since $\mathcal{S}$ is compact, J must attain its maximum value at some point $\mathbf{p} \in \mathcal{S}$, and this maximum value is > 0. Choose a patch $\boldsymbol{\sigma} : U \to \mathbb{R}^3$ of $\mathcal{S}$ containing $\mathbf{p}$, and let κ_1 and κ_2 be its principal curvatures. Since $\kappa_1\kappa_2 > 0$, by reparametrizing if necessary, we can assume that κ_1 and κ_2 are both > 0 (see Exercise 8.2.8). Suppose that $\kappa_1 > \kappa_2$ at $\mathbf{p}$; since κ_1 and κ_2 are continuous (in fact, smooth) functions on U (Exercise 8.2.8 again), by shrinking U if necessary, we can assume that $\kappa_1 > \kappa_2$ everywhere on U. Since K is a constant > 0, the function $\left(x - \frac{K}{x}\right)^2$ increases with x provided that $x > K/x > 0$. Since $\kappa_1 > K/\kappa_1 = \kappa_2 > 0$, this function is increasing at $x = \kappa_1$, so κ_1 must have a local maximum at $\mathbf{p}$, and then $\kappa_2 = K/\kappa_1$ must have a local minimum there. By Lemma 10.3.5, $K \leq 0$ at $\mathbf{p}$. This contradicts the assumption that $K > 0$. □

EXERCISES

10.3.1 Show that a compact surface with Gaussian curvature > 0 everywhere and constant *mean* curvature is a sphere.

10.3.2 Show that the solution of the sine-Gordon equation corresponding to the pseudosphere constructed in Section 8.3 is

$$\theta(u,v) = 2\tan^{-1}(\sinh(u-v+c)),$$

where c is a constant.

10.4 Geodesic mappings

We considered in Chapter 6 local diffeomorphisms between surfaces that preserve lengths, angles or areas. In the same way, it is natural to consider local diffeomorphisms that preserve geodesics. In fact, to obtain interesting results we have to consider local diffeomorphisms $F : \mathcal{S} \to \tilde{\mathcal{S}}$ from a surface $\mathcal{S}$ to a surface $\tilde{\mathcal{S}}$ such that, if $\boldsymbol{\gamma}$ is a geodesic on $\mathcal{S}$, then $F \circ \boldsymbol{\gamma}$ is a *reparametrization* of a geodesic on $\tilde{\mathcal{S}}$ (cf. Exercise 10.4.1). We recall from Exercise 9.1.2 that a curve that is a reparametrization of a geodesic is called a pre-geodesic. Thus, we are led to

Definition 10.4.1

If $\mathcal{S}$ and $\tilde{\mathcal{S}}$ are surfaces, a local diffeomorphism $F : \mathcal{S} \to \tilde{\mathcal{S}}$ is said to be *geodesic* if f takes every pre-geodesic on $\mathcal{S}$ to a pre-geodesic on $\tilde{\mathcal{S}}$.

The following result provides some examples of geodesic local diffeomorphisms.

Proposition 10.4.2

The following are geodesic local diffeomorphisms:

(i) Every local isometry.

(ii) Every dilation of $\mathbb{R}^3$.

(iii) Every composite of geodesic local diffeomorphisms.

Proof

(i) If $\boldsymbol{\gamma}$ is a pre-geodesic on a surface patch $\boldsymbol{\sigma}$ of a surface $\mathcal{S}$, then $\boldsymbol{\gamma}(t) = \boldsymbol{\gamma}_1(\varphi(t))$ for some geodesic $\boldsymbol{\gamma}_1$ and some reparametrization map φ (Exercise 9.1.2). If $F : \mathcal{S} \to \tilde{\mathcal{S}}$ is an isometry, then $\tilde{\boldsymbol{\gamma}}_1 = F \circ \boldsymbol{\gamma}_1$ is a geodesic on the surface patch $\tilde{\boldsymbol{\sigma}} = f \circ \boldsymbol{\sigma}$ of $\tilde{\mathcal{S}}$ by Corollary 9.2.7. Hence, $F \circ \boldsymbol{\gamma} = F \circ \boldsymbol{\gamma}_1 \circ \varphi = \tilde{\boldsymbol{\gamma}}_1 \circ \varphi$ is a reparametrization of the geodesic $\tilde{\boldsymbol{\gamma}}_1$, and so is a pre-geodesic.

(ii) The dilation $(x, y, z) \mapsto a(x, y, z)$, where a is a non-zero constant, multiplies the first fundamental form by a^2 and hence leaves the Christoffel symbols unchanged (see Proposition 7.4.4). By Proposition 9.2.3, applying the dilation leaves the equations determining the geodesics unchanged. It follows that the dilation takes geodesics to geodesics. By the argument in part (i), it also takes pre-geodesics to pre-geodesics.

(iii) This is obvious. □

As the preceding proof made clear, local isometries and dilations take geodesics to geodesics. The following is an example of a geodesic local diffeomorphism that *does not take geodesics to geodesics*. By the preceding proposition, it cannot be a composite of local isometries and dilations.

Example 10.4.3

We consider the map F from the lower hemisphere of the unit sphere S^2 to the plane $z = -1$ given by projection from the centre of the sphere.

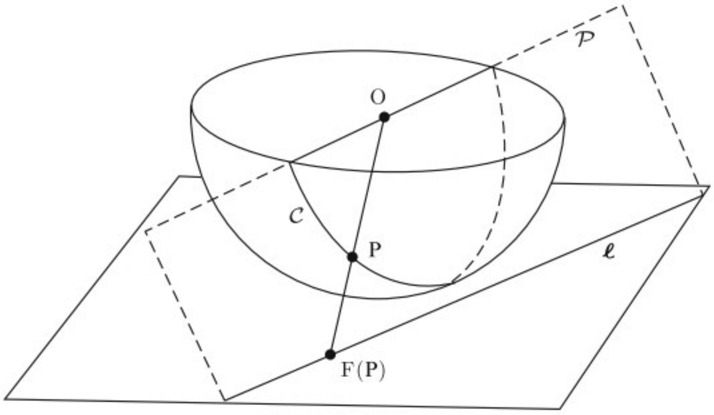

Every pre-geodesic $\mathcal{C}$ on S^2 is a great circle formed by the intersection of S^2 with a plane $\mathcal{P}$ passing through the origin. Central projection takes the part of $\mathcal{C}$ in the lower hemisphere to the straight line l that is the intersection of $\mathcal{P}$ with the plane $z = -1$. Hence, F is a geodesic diffeomorphism.

As a particular case, let $\mathcal{P}$ be the xz-plane, so that $\mathcal{C}$ can be parametrized by $\boldsymbol{\gamma}(t) = (\cos t, 0, \sin t)$; note that $\boldsymbol{\gamma}$ is unit-speed and so is a geodesic. It is easy to show that $F(\boldsymbol{\gamma}(t)) = (-\cot t, 0, -1)$: this curve is a parametrization of the straight line $y = 0$, $z = -1$ (hence it is a pre-geodesic), but it is *not* a geodesic as it does not have constant speed (Proposition 9.1.2).

The inverse of the diffeomorphism F in Example 10.4.3 provides a parametrization of (an open subset of) S^2 with the property that geodesics on S^2 correspond to straight lines in the parameter plane. Beltrami proved the following beautiful theorem which determines exactly which surfaces admit parametrizations with this property. It motivated him to carry out a detailed investigation of surfaces of constant Gaussian curvature, which in turn led to his work on hyperbolic geometry that we describe in Chapter 11.

Theorem 10.4.4

Let $\mathcal{S}$ be a connected surface. If there is a geodesic local diffeomorphism from $\mathcal{S}$ to a plane, then $\mathcal{S}$ has constant Gaussian curvature. Conversely, if $\mathcal{S}$ has constant Gaussian curvature, then for any point $\mathbf{p} \in \mathcal{S}$ there is a surface patch $\boldsymbol{\sigma} : U \to \mathcal{S}$ of $\mathcal{S}$ such that $\mathbf{p} \in \boldsymbol{\sigma}(U)$ and $\boldsymbol{\sigma}^{-1} : \boldsymbol{\sigma}(U) \to U$ is a geodesic diffeomorphism.

Proof

If F is a geodesic local diffeomorphism from $\mathcal{S}$ to the plane, $\boldsymbol{\sigma} = F^{-1}$ is a parametrization of an open subset of $\mathcal{S}$ with the property that $\boldsymbol{\sigma}$ takes every straight line in the plane to a pre-geodesic on $\mathcal{S}$. Denoting the parameters by u, v, we can take the line to have equation

$$v = \lambda u + \mu, \tag{10.12}$$

where λ and μ are constants. By Exercise 9.1.2, there is a parametrization $t \mapsto (u(t), v(t))$ of this line such that $\boldsymbol{\gamma}(t) = \boldsymbol{\sigma}(u(t), v(t))$ is a unit-speed geodesic on $\mathcal{S}$. By Proposition 9.2.3, we have

$$\begin{aligned} \ddot{u} + (\Gamma_{11}^1 + 2\lambda\Gamma_{12}^1 + \lambda^2\Gamma_{22}^1)\dot{u}^2 &= 0, \\ \lambda\ddot{u} + (\Gamma_{11}^2 + 2\lambda\Gamma_{12}^2 + \lambda^2\Gamma_{22}^2)\dot{u}^2 &= 0. \end{aligned}$$

Eliminating $\ddot{u}$ and noting that $\dot{u}$ cannot be zero (for then we would have $\dot{v} = 0$ too, since $\dot{v} = \lambda\dot{u}$ by Eq. 10.12, contradicting the fact that $\boldsymbol{\gamma}$ is unit-speed), we get

$$\lambda(\Gamma_{11}^1 + 2\lambda\Gamma_{12}^1 + \lambda^2\Gamma_{22}^1) = \Gamma_{11}^2 + 2\lambda\Gamma_{12}^2 + \lambda^2\Gamma_{22}^2.$$

Since *every* straight line in the uv-plane corresponds to a geodesic $\boldsymbol{\gamma}$ in this way, the preceding equation must hold for all values of the constant λ. This implies that

$$\Gamma_{11}^2 = \Gamma_{22}^1 = 0, \quad \Gamma_{11}^1 = 2\Gamma_{12}^2, \quad \Gamma_{22}^2 = 2\Gamma_{12}^1.$$

The Gauss equations (Proposition 10.1.2) now become

$$\begin{aligned}
KE &= (\Gamma_{12}^2)^2 - (\Gamma_{12}^2)_u,\\
KF &= (\Gamma_{12}^1)_u - 2(\Gamma_{12}^2)_v + \Gamma_{12}^2\Gamma_{12}^1,\\
KF &= (\Gamma_{12}^2)_v - 2(\Gamma_{12}^1)_u + \Gamma_{12}^1\Gamma_{12}^2,\\
KG &= (\Gamma_{12}^1)^2 - (\Gamma_{12}^1)_v.
\end{aligned}$$

Subtracting the middle two equations gives $(\Gamma_{12}^1)_u = (\Gamma_{12}^2)_v$, and using this the four equations become

$$KE = (\Gamma_{12}^2)^2 - (\Gamma_{12}^2)_u, \tag{10.13}$$
$$KF = -(\Gamma_{12}^2)_v + \Gamma_{12}^1\Gamma_{12}^2, \tag{10.14}$$
$$KF = -(\Gamma_{12}^1)_u + \Gamma_{12}^1\Gamma_{12}^2, \tag{10.15}$$
$$KG = (\Gamma_{12}^1)^2 - (\Gamma_{12}^1)_v. \tag{10.16}$$

Differentiating Eq. 10.13 with respect to v and Eq. 10.14 with respect to u, and then equating the expressions for $(\Gamma_{12}^2)_{uv} = (\Gamma_{12}^2)_{vu}$ gives

$$\begin{aligned}
EK_v - FK_u &= K(F_u - E_v) + 2\Gamma_{12}^2(\Gamma_{12}^2)_v - \Gamma_{12}^2(\Gamma_{12}^1)_u - \Gamma_{12}^1(\Gamma_{12}^2)_u\\
&= K(F_u - E_v) + 2\Gamma_{12}^2(\Gamma_{12}^1\Gamma_{12}^2 - KF) - \Gamma_{12}^2(\Gamma_{12}^1\Gamma_{12}^2 - KF)\\
&\quad - \Gamma_{12}^1((\Gamma_{12}^2)^2 - KE)\\
&= K(F_u - E_v) + K(E\Gamma_{12}^1 - F\Gamma_{12}^2),
\end{aligned} \tag{10.17}$$

where we used Eqs. 10.13–10.15 in passing from the first line to the second.

Using the definition of the Christoffel symbols (Proposition 7.4.4), we find that

$$E\Gamma_{12}^1 - F\Gamma_{12}^2 = \frac{(EG + F^2)E_v - 2EFG_u}{2(EG - F^2)}. \tag{10.18}$$

Now, the equations $\Gamma_{11}^2 = 0$ and $\Gamma_{11}^1 = 2\Gamma_{12}^2$ give

$$EE_v = 2EF_u - FE_u, \quad 3FE_v = 2EG_u + 2FF_u - GE_u.$$

Hence,

$$\begin{aligned}
(EG + F^2)E_v - 2EFG_u &= 2(EG - F^2)E_v + (3F^2 - EG)E_v - 2EFG_u\\
&= 2(EG - F^2)E_v + F(2EG_u + 2FF_u - GE_u)\\
&\quad - G(2EF_u - FE_u) - 2EFG_u\\
&= 2(EG - F^2)(E_v - F_u).
\end{aligned}$$

Inserting this result into Eq. 10.18 gives

$$E\Gamma_{12}^1 - F\Gamma_{12}^2 = E_v - F_u,$$

and then Eq. 10.17 gives

$$EK_v - FK_u = 0.$$

A similar calculation starting with Eqs. 10.15 and 10.16 leads to

$$FK_v - GK_u = 0.$$

Hence,

$$(EG - F^2)K_v = GFK_u - FGK_u = 0,$$

so $K_v = 0$ (since $EG - F^2 > 0$). Similarly $K_u = 0$. Hence, K is constant.

For the converse, assume that $\mathcal{S}$ has constant Gaussian curvature. By Theorem 10.3.1, each point of $\mathcal{S}$ is contained in a surface patch that is isometric to an open subset of a plane, a sphere or a pseudosphere. Since local isometries and dilations are geodesic local diffeomorphisms (Proposition 10.4.2), we have only to show that there are geodesic local diffeomorphisms from $\mathcal{S}$ to the plane in three cases:

(i) $\mathcal{S}$ is a plane, when there is nothing to prove.

(ii) $\mathcal{S}$ is the unit sphere, which is Example 10.4.3.

(iii) $\mathcal{S}$ is the pseudosphere in Section 8.3, which is Exercise 9.3.4.

The proof of Beltrami's theorem is complete. □

EXERCISES

10.4.1 Show that a local diffeomorphism between surfaces that takes unit-speed geodesics to unit-speed geodesics must be a local isometry.

10.4.2 Show that a local diffeomorphism between surfaces that is the composite of a dilation and a local isometry takes geodesics to geodesics. Is the converse of this statement true?

10.4.3 This exercise shows that a geodesic local diffeomorphism F from a surface $\mathcal{S}$ to a surface $\tilde{\mathcal{S}}$ that is also conformal is the composite of a dilation and a local isometry.

(i) Let $\mathbf{p} \in \mathcal{S}$ and let $\boldsymbol{\sigma}$ be a geodesic patch containing $\mathbf{p}$ as in Proposition 9.5.1, with first fundamental form du^2+Gdv^2. Show that $\tilde{\boldsymbol{\sigma}} = F \circ \boldsymbol{\sigma}$ is a patch of $\tilde{\boldsymbol{\sigma}}$ containing $F(\mathbf{p})$ with first fundamental form $\lambda(du^2 + Gdv^2)$ for some smooth function $\lambda(u, v)$.

(ii) Show that the parameter curves $v =$ constant are pre-geodesics on $\tilde{\boldsymbol{\sigma}}$ and deduce that λ is independent of v.

(iii) Show that if $\boldsymbol{\gamma}$ is a geodesic on $\boldsymbol{\sigma}$ and θ is the oriented angle between $\boldsymbol{\gamma}$ and the parameter curves $v =$ constant,

$$\frac{d\theta}{dv} + \frac{G_u}{2\sqrt{G}} = 0. \tag{10.19}$$

(iv) Show that

$$\frac{d\theta}{dv} + \frac{(\lambda G)_u}{2\lambda\sqrt{G}} = 0. \tag{10.20}$$

(v) Deduce from Eqs. 10.19 and 10.20 that λ is constant.

(vi) Show that $F : \boldsymbol{\sigma} \to \tilde{\boldsymbol{\sigma}}$ is the composite of a dilation and a local isometry.

DIFFERENTIAL GEOMETRY OF CURVES AND SURFACES

6. Gauss' Theorema Egregium

Question. *How can we decide if two given surfaces can be obtained from each other by "bending without stretching"?*

The simplest example is a flat strip, say of paper, which can be rolled (without stretching!) to give a cylinder; a bit more precisely, a piece of the plane can be deformed to a piece of a cylinder. On the other hand, it is intuitively clear that we cannot deform without stretching a piece of a plane to a piece of a sphere (or of an ellipse, of a hyperboloid, of a torus etc. — just try to wrap such an object!) In general, the question is hard. Let's try to translate the question into a more formal language:

- bending = *map between surfaces*
- without stretching = *preserving the distances between two points*, which is the same as *local isometry* (see the last section in chapter 4 of the notes)

Gauss came up with the idea that an answer to our question can be given by measuring curvature of curves on a surface, more precisely, the principal curvatures, but in a very ingenious way. Because if we compare the plane and the cylinder we see that their principal curvatures are $k_1^{\text{plane}} = k_2^{\text{plane}} = 0$ whereas $k_1^{\text{cylinder}} = 1$, $k_2^{\text{cylinder}} = 0$. Even though they are locally isometric, the plane and the cylinder have different principal curvatures. Nevertheless, their product is the same! Recall that the latter product is what we called the Gauss curvature. In fact, Gauss' observation was that what we noticed in the case of the deformation plane→cylinder is a very general phenomenon: if there is a deformation without stretching of the surface S_1 to the surface S_2, then the product of the principal curvatures at any point on S_1 is preserved by the deformation. Another instructive (and easy to visualize, thanks to the computer generated animation available online) example is the continuous deformation of the helicoid in the catenoid. If you follow a point on the helicoid during this process, the Gauss curvature doesn't change, even though the surface changes its shape radically during the process. Gauss' theorem can be stated as follows:

Theorema Egregium.[1] *If $f : S_1 \to S_2$ is a local isometry, then the Gauss curvature of S_1 at P equals the Gauss curvature of S_2 at $f(P)$.*

Remark. 1. The theorem can only be used to rule out (local) isometries between surfaces. By this we mean that the converse of the theorem is not true: one can find a diffeomorphism (that is, a differentiable bijective map) between surfaces which preserves the Gauss curvature at any point, but is not an isometry. You will see such an example in Homework no. 5. In the same spirit: we cannot deduce that the plane and the cylinder (or the helicoid and the catenoid) are locally isometric just because they have the same Gauss curvature!

2. In fact the importance of the theorem goes beyond the application mentioned above (that is, rule out locally isometric surfaces). Many authors consider it "the most important single theorem in differential geometry". But it's not easy to explain why, so we refer the reader to Spivak's monograph [Sp], the end of chapter 3, part B (page 143 - 144, "What does Theorema Egregium really mean?").

The rest of the chapter is devoted to the proof of this theorem. The key point is to consider a local parametrization (U, φ) and produce a formula for $K(\varphi(Q))$ (where Q is in U) which depends only on the functions E, F, and G (the coefficients of the first fundamental form).

---

[1] Gauss' original paper where this theorem was stated and proven was written in Latin. He says that this is a *remarkable theorem*, which in Latin means *theorema egregium* (egregium=remarkable).

First we note that for any Q in U, the vectors φ'_u, φ'_v, and N furnish a basis of $\mathbb{R}^3$. Thus we can consider the expansions

$$\begin{aligned}
\varphi''_{uu} &= \Gamma^1_{11}\varphi'_u + \Gamma^2_{11}\varphi'_v + L_1 N\\
\varphi''_{uv} &= \Gamma^1_{12}\varphi'_u + \Gamma^2_{12}\varphi'_v + L_2 N\\
\varphi''_{vu} &= \Gamma^1_{21}\varphi'_u + \Gamma^2_{21}\varphi'_v + L_2 N\\
\varphi''_{vv} &= \Gamma^1_{22}\varphi'_u + \Gamma^2_{22}\varphi'_v + L_3 N\\
N'_u &= a_{11}\varphi'_u + a_{21}\varphi'_v\\
N'_v &= a_{12}\varphi'_u + a_{22}\varphi'_v.
\end{aligned} \tag{1}$$

The numbers Γ^k_{ij} are called the *Christoffel symbols* of the parametrization (U, φ). Altogether, they are in number of $2 \times 2 \times 2 = 8$, but in fact, since $\varphi'_{uv} = \varphi'_{vu}$, we have

$$\Gamma^1_{12} = \Gamma^1_{21} \text{ and } \Gamma^2_{12} = \Gamma^2_{21}.$$

We take dot product of the first four equations with N and use equations (3), page 12, chapter 5 of the notes. We obtain

$$L_1 = e,\ L_2 = f,\ L_3 = g,$$

where e, f, g are the coefficients of the second fundamental form. The Christoffel symbols can be determined by taking again the first four equations and making dot products with φ'_u and φ'_v. We obtain:

$$\begin{aligned}
\Gamma^1_{11}E + \Gamma^2_{11}F &= \varphi'_{uu} \cdot \varphi'_u = \frac{1}{2}E_u\\
\Gamma^1_{11}F + \Gamma^2_{11}G &= \varphi'_{uu} \cdot \varphi'_v = F_u - \frac{1}{2}E_u
\end{aligned} \tag{2}$$

$$\begin{aligned}
\Gamma^1_{12}E + \Gamma^2_{12}F &= \varphi'_{uv} \cdot \varphi'_u = \frac{1}{2}E_v\\
\Gamma^1_{12}F + \Gamma^2_{12}G &= \varphi'_{uv} \cdot \varphi'_v = \frac{1}{2}G_u
\end{aligned} \tag{3}$$

$$\begin{aligned}
\Gamma^1_{22}E + \Gamma^2_{22}F &= \varphi'_{vv} \cdot \varphi'_u = F_v - \frac{1}{2}G_u\\
\Gamma^1_{22}F + \Gamma^2_{22}G &= \varphi'_{vv} \cdot \varphi'_v = \frac{1}{2}G_v
\end{aligned} \tag{4}$$

Each of the three groups of equations can be considered as a linear system of two equations with two unknowns. The determinant of each system is $EG - F^2$ which is strictly positive (why?). We can solve each system and obtain concrete expressions for Γ^k_{ij} in terms of E, F, G and their derivatives.

Now we consider the obvious equation

$$(\varphi''_{uu})'_v - (\varphi''_{uv})'_u = 0.$$

If we use (1), we deduce

$$\begin{aligned}
&\Gamma^1_{11}\varphi''_{uv} + \Gamma^2_{11}\varphi''_{vv} + eN'_v + (\Gamma^1_{11})'_v\varphi'_u + (\Gamma^2_{11})'_v\varphi'_v + e'_v N\\
&= \Gamma^1_{12}\varphi''_{uu} + \Gamma^2_{12}\varphi''_{vu} + fN'_u + (\Gamma^1_{12})'_u\varphi'_u + (\Gamma^2_{12})'_u\varphi'_v + f'_u N
\end{aligned}$$

We use again (1) to substitute all second order derivatives. Then we equate the coefficient of φ'_v. We obtain:

$$\Gamma^1_{11}\Gamma^2_{12} + \Gamma^2_{11}\Gamma^2_{22} + ea_{22} + (\Gamma^2_{11})'_v$$
$$= \Gamma^1_{12}\Gamma^2_{11} + \Gamma^2_{12}\Gamma^2_{12} + fa_{21} + (\Gamma^2_{12})'_u.$$

But we actually know a_{21} and a_{22}, by the Weingarten equations (see chapter 5, section 4 of the notes). By making the replacements we obtain:

$$(\Gamma^2_{12})'_u - (\Gamma^2_{11})'_v + \Gamma^1_{12}\Gamma^2_{11} + \Gamma^2_{12}\Gamma^2_{12} - \Gamma^2_{11}\Gamma^2_{22} - \Gamma^1_{11}\Gamma^2_{12}$$
$$= -E\frac{eg - f^2}{EG - F^2} = -EK$$

and this gives the desired

formula for K in terms of E, F, G and their partial derivatives of first and second order with respect to u and v.

To finish the proof, we consider the local isometry $f : S_1 \to S_2$. Take P a point of S_1. Because $d(f)_P : T_PS_1 \to T_{f(P)}S_2$ is a linear isomorphism, f is a local diffeomorphism at P, which implies that there exists a local parametrization (U, φ) of S_1 around P such that $(U, f \circ \varphi)$ is a local parametrization of S_2 around $f(P)$. For any Q in U, the canonical basis of $T_{f(\varphi(Q))}S_2$ consists of

$$(f \circ \varphi)'_u(Q) = d(f)_{\varphi(Q)}(\varphi'_u(Q)),\ (f \circ \varphi)'_v(Q) = d(f)_{\varphi(Q)}(\varphi'_v(Q)).$$

Moreover, if we denote $P = \varphi(Q)$, we have

$$\mathrm{I}_{f(P)}(d(f)_P(w)) = \mathrm{I}_P(w),$$

for all $w = w_1\varphi'_u(Q) + w_2\varphi'_v(Q)$ in T_PS_1. We also have

$$d(f)_P(w) = w_1(f \circ \varphi)'_u(Q) + w_2(f \circ \varphi)'_v(Q),$$

and

$$\mathrm{I}_P(w) = E(Q)w_1^2 + 2F(Q)w_1w_2 + G(Q)w_2^2$$

as well as

$$\mathrm{I}_{f(P)}(d(f)_P(w)) = \tilde{E}(Q)w_1^2 + \tilde{F}(Q)w_1w_2 + \tilde{G}(Q)w_2^2.$$

We deduce $E(Q) = \tilde{E}(Q)$, $F(Q) = \tilde{F}(Q)$, and $G(Q) = \tilde{G}(Q)$ for any $Q = (u, v)$ in U. Combined with the formula for K from above, these imply that the curvature of S_1 at $\varphi(Q)$ equals the curvature of S_2 at $f(\varphi(Q))$, for any Q in U. □

REFERENCES

[dC] M. P. do Carmo, *Differential Geometry of Curves and Surfaces*, Prentice Hall 1976

[Gr-Abb-Sa] A. Gray, E. Abbena, and S. Salamon, *Modern Differential Geometry with MATHEMATICA*, 3rd edition

[Sp] M. Spivak, *A Comprehensive Introduction to Differential Geometry*, vol. II

+,-.+ + +/ + !% E!2: @ 3"
/6

" +,-.+ + +/ + !% I: 3"

+,-.+ + +/ + !% : 2> 3"

CURVATURE AND THE THEOREMA EGREGIUM OF GAUSS

DEANE YANG

In this note, we describe a simple way to define the second fundamental form of a hypersurface in $\mathbb{R}^n$ and use it to prove Gauss's Theorema Egregium, as well as its analogue in higher dimensions.

The basic idea is to approximate a hypersurface, near a given point, by the graph of a quadratic polynomial near its critical point. In particular, given any point p in a hypersurface $S \subset \mathbb{R}^n$, the hypersurface can be positioned so that p is at the origin and S is tangent to the hyperplane $T = \{x^n = 0\}$. It follows that S is locally the graph of a function $f(x^1, \dots, x^{n-1})$ satisfying $f(0) = 0$ and $df(0) = 0$. The second order Taylor expansion of f is therefore

$$f(v) = H(v, v) + o(|v|^2), \ v \in T,$$

where H is a symmetric tensor. The second fundamental form at p is defined to be H. The uniqueness of H implies that it is an extrinsic geometric invariant tensor of S.

The statement of the Theorem Egregium then falls out of a straightforward calculation that seeks to identify a tensor invariant, defined in terms of the second fundamental form, that is independent of the extrinsic isometric embedding. It therefore is an intrinsic geometric invariant of the hypersurface. Another straightforward tensor calculation leads to an intrinsic definition of Gauss curvature for a surface in $\mathbb{R}^3$ and, more generally, the Riemann curvature tensor.

Note that neither the Levi-Civita connection nor the Gauss map is used explicitly below.

The discussion below has a straightforward extension to submanifolds of higher codimension.

1. Rigid motions

Let $e_1, \dots, e_n$ denote the standard basis of $\mathbb{R}^n$. Given a nonzero vector $\nu \in \mathbb{R}^n$, let $\nu^\perp$ denote the linear hyperplane normal to ν.

A *rigid motion* is a map $R\colon \mathbb{R}^n \to \mathbb{R}^n$ given by

$$R(x) = A(x - p),$$

where $A \in \mathrm{SO}(n)$ and $p \in \mathbb{R}^n$.

2. Definition of a hypersurface

Define a *hypersurface* to be a subset $S \subset \mathbb{R}^n$ such that for each $x \in S$, there exists a linear subspace $T \subset \mathbb{R}^n$ of codimension 1 such that S near x is graph of a smooth function of the plane $x + T$. In other words, there exists neighborhood $N \subset \mathbb{R}^n$ of 0 and a smooth function $f : N \cap T \to \mathbb{R}$ such that

$$S \cap (x + N) = \{x + v + f(v)u \ : \ v \in T \cap N\}, \tag{1}$$

where u is a nonzero vector normal to T and $f(0) = 0$.

Date: December 31, 2015 (Revised November 13, 2019).

3. Tangent space

For each $x \in S$, there is a unique choice of T, which will be denoted T_xS and called the *tangent plane to* S *at* x, such that the corresponding function f satisfies $f(0) = 0$ and $df(0) = 0$. If S is orientable, then there exists a smooth map $u : S \to S^{n-1}$, where $u(x)$ is normal to T_xS. This is called the *Gauss map*. We will not use the Gauss map below.

4. The second fundamental form

For each $x \in S$, let f be the function be the one in (1), with $T = T_xS$. The *second fundamental form* of S at x is defined to be the Hessian $H(x)$ of f at 0. The Hessian $H(x)$ is a symmetric tensor on T_xS, where, for any $v \in T_xS$,

$$H(x)(v, v) = \left.\frac{d^2}{dt^2}\right|_{t=0} f(tv).$$

Since f is uniquely determined by S and $x \in S$, so is H. It is therefore a geometric invariant of the surface S. Since it depends on the embedding of S, it is an extrinsic invariant.

Given $x \in S$, we can apply a rigid motion to S so that $x = 0$, $u(x) = e_n$, and therefore, in a neighborhood of $0 \in \mathbb{R}^n$,

$$S = \{(x', f(x'))\},$$

where $f(0) = 0$ and $df(0) = 0$. For each $v = v^1e_1 + \cdots + v^{n-1}e_{n-1} \in T_0S (= e_n^\perp)$,

$$H(0)(v, v) = v^iv^jH(0)(e_i, e_j) = H_{ij}v^iv^j, \text{ where } H_{ij} = H_{ji} = H(0)(e_i, e_j).$$

5. Isometric hypersurfaces

Let T denote $e_n^\perp$.

Two hypersurfaces S and $\widehat{S}$ are *isometric* if there exists a smooth diffeomorphism

$$\Phi : S \to \widehat{S}$$

such that the length of any smooth curve $C \subset S$ is equal to the length of the curve $\Phi(C) \subset \widehat{S}$.

Given $x \in S$, let $\hat{x} = \Phi(x) \in \widehat{S}$. Applying rigid motions to both S and $\widehat{S}$, we can assume that that $\hat{x} = x = 0$ and the respective tangent hyperplanes are $T_{\hat{x}}\widehat{S} = T_xS = T$. This in turn defines uniquely the functions $f : T \to \mathbb{R}$, whose graph near 0 is S, and $\hat{f} : T \to \mathbb{R}$, whose graph near 0 is $\widehat{S}$.

Moreover, there is a diffeomorphism $\phi : T \to T$ such that, for each $v \in T$,

$$\Phi(v + f(v)e_n) = \phi(v) + \hat{f}(\phi(v))e_n.$$

The diffeomorphism $\Phi : S \to \widehat{S}$ preserves the lengths of all curves if and only if

$$\partial_i\hat{y} \cdot \partial_j\hat{y} = \partial_iy \cdot \partial_jy,$$

where, for each $v \in T$,

$$\begin{aligned} \hat{y}(v) &= \phi^k(v) + \hat{f}(\phi(v))e_n \\ y(v) &= v + f(v)e_n, \end{aligned}$$

for each $v \in T$. In other words,

$$\partial_i\phi \cdot \partial_j\phi + \partial_p\hat{f}\partial_i\phi^p\partial_q\hat{f}\partial_j\phi^q = \delta_{ij} + \partial_if\partial_jf. \tag{2}$$

Note that

$$\begin{aligned} \partial_i f(0) &= \partial_i \hat{f}(0) = 0 \\ \partial_i \phi^j(0) &= \delta_i^j, \end{aligned}$$

for all $1 \leq i, j \leq n-1$. Therefore, if we differentiate (2) and evaluate at $v = 0$, we get

$$\partial^2_{ik}\phi^j + \partial^2_{jk}\phi^i = 0.$$

This implies that, at $v = 0$,

$$\partial^2_{ik}\phi^j = -\partial^2_{jk}\phi^i = \partial^2_{ji}\phi^k = -\partial^2_{ik}\phi^j, \tag{3}$$

and therefore

$$\partial^2_{jk}\phi^i(0) = 0, \text{ for all } 1 \leq i, j, k \leq n-1.$$

If we differentiate (2) again, evaluate at $x = 0$, and cycle through the indices i, j, k, l, we get

$$\partial^3_{ikl}\phi^j + \partial^3_{jkl}\phi^i + \partial^2_{ik}\hat{f}\partial^2_{jl}\hat{f} + \partial^2_{il}\hat{f}\partial^2_{jk}\hat{f} = \partial^2_{ik}f\partial^2_{jl}f + \partial^2_{il}f\partial^2_{jk}f \tag{4}$$

$$\partial^3_{jli}\phi^k + \partial^3_{kli}\phi^j + \partial^2_{jl}\hat{f}\partial^2_{ki}\hat{f} + \partial^2_{ji}\hat{f}\partial^2_{kl}\hat{f} = \partial^2_{jl}f\partial^2_{ki}f + \partial^2_{ji}f\partial^2_{kl}f \tag{5}$$

$$\partial^3_{kij}\phi^l + \partial^3_{lij}\phi^k + \partial^2_{ki}\hat{f}\partial^2_{lj}\hat{f} + \partial^2_{kj}\hat{f}\partial^2_{li}\hat{f} = \partial^2_{ki}f\partial^2_{lj}f + \partial^2_{kj}f\partial^2_{li}f \tag{6}$$

$$\partial^3_{ljk}\phi^i + \partial^3_{ijk}\phi^l + \partial^2_{lj}\hat{f}\partial^2_{ik}\hat{f} + \partial^2_{lk}\hat{f}\partial^2_{ij}\hat{f} = \partial^2_{lj}f\partial^2_{ik}f + \partial^2_{lk}f\partial^2_{ij}f. \tag{7}$$

Therefore, the equation

$$(4) - (5) + (6) - (7) \tag{8}$$

eliminates ϕ and gives

$$\partial^2_{ik}\hat{f}\partial^2_{jl}\hat{f} - \partial^2_{il}\hat{f}\partial^2_{ik}\hat{f} = \partial^2_{ik}f\partial^2_{jl}f - \partial^2_{il}f\partial^2_{ik}f, \tag{9}$$

Ig follows that, if H and $\widehat{H}$ are the second fundamental forms of S and $\widehat{S}$ at 0, then

$$\widehat{R}_{ijkl} = R_{ijkl}, \tag{10}$$

where

$$R_{ijkl} = H_{ik}H_{jl} - H_{il}H_{jk} \tag{11}$$

$$\widehat{R}_{ijkl} = \widehat{H}_{ik}\widehat{H}_{jl} - \widehat{H}_{il}\widehat{H}_{jk}. \tag{12}$$

This leads to the definition of the tensor R, where, for any $v_1, v_2, v_3, v_4 \in T_xS$,

$$R(v_1, v_2, v_3, v_4) = H(v_1, v_3)H(v_2, v_4) - H(v_1, v_4)H(v_2, v_3).$$

What (10) shows is that the tensor R for a surface S depends only on its intrinsic geometry and not on the isometric embedding. The tensor R is, of course, the Riemann curvature tensor, and the equations (11) are the Gauss equations.

If $n = 2$, then the only nontrivial component of the Riemann curvature tensor is

$$K = R(e_1, e_2, e_1, e_2) = H_{11}H_{22} - H_{12}^2,$$

which is known as the Gauss curvature, and equation (10) is the Theorem Egregium of Gauss.

Example 1. We can show that the standard sphere in $\mathbb{R}^3$ is not locally isometric to the plane $x^3 = 0$ using this. Using the definition of the second fundamental form, we can verify that, for the plane, $H_{ij} = 0$ and therefore $R_{1212} = 0$. On the other hand, for the sphere, $H_{ij} = \delta_{ij}$ and therefore $R_{1212} = 1$. Therefore, the two surfaces are not locally isometric.

6. Tensor identities

Most of the proof above involves only differentiation and straightforward calculations. The only significant steps are the following:

(1) Definition of a hypersurface
(2) If a map between two hypersurfaces preserves lengths of curves, then the map satisfies equation (2).
(3) Most importantly, the tensor calculations done in (3) and (8). These are equivalent to the following tensor identities:

$$\begin{aligned}(T \otimes S^2T) \cap (\Lambda^2 T \otimes T) &= \{0\}\\ (T \otimes S^3T) \cap (\Lambda^2 T \otimes \Lambda^2 T) &= \{0\}.\end{aligned}$$

7. Intrinsic formula for Riemann curvature

The embedding of S as a graph over T is given by $v \in T \mapsto v + f(v)e_n$ and therefore, the Riemannian metric is given by

$$\begin{aligned}g_{ij}(v) &= \partial_i(v + f(v)e_n) \cdot \partial_j(v + f(v)e_n)\\ &= (e_i + \partial_i f(v)e_n) \cdot (e_k + \partial_j f(v)e_n)\\ &= \delta_{ij} + \partial_i f \partial_j f.\end{aligned}$$

Therefore, since $df(0) = 0$, it follows that at $v = 0$,

$$\begin{aligned}\partial_k g_{ij} &= \partial_k(\delta_{ij} + \partial_i f \partial_j f)\\ &= \partial^2_{ki} f \partial_j f + \partial_i f \partial^2_{kj} f\\ &= 0.\\ \partial^2_{ij} g_{kl} &= \partial^2_{ij}(\partial_k f \partial_l f)\\ &= \partial^2_{ik} f \partial^2_{jl} f + \partial^2_{il} f \partial^2_{jk} f.\end{aligned}$$

Permuting the indices, we get

$$\partial^2_{ik} g_{jl} = \partial^2_{ij} f \partial^2_{kl} f + \partial^2_{il} f \partial^2_{jk} f \tag{13}$$

$$\partial^2_{jl} g_{ik} = \partial^2_{ij} f \partial^2_{kl} f + \partial^2_{jk} f \partial^2_{il} f \tag{14}$$

$$\partial^2_{il} g_{jk} = \partial^2_{ij} f \partial^2_{kl} f + \partial^2_{ik} f \partial^2_{jl} f \tag{15}$$

$$\partial^2_{jk} g_{il} = \partial^2_{ij} f \partial^2_{kl} f + \partial^2_{jl} f \partial^2_{ik} f. \tag{16}$$

Therefore, the equation

$$-(13) - (14) + (15) + (16) \tag{17}$$

implies that, at $x = 0$,

$$\begin{aligned}-\partial^2_{ik} g_{jl} - \partial^2_{jl} g_{ik} + \partial^2_{il} g_{jk} + \partial^2_{jk} g_{il} &= -\partial^2_{il} f \partial^2_{jk} f - \partial^2_{jk} f \partial^2_{jl} f + \partial^2_{ik} f \partial^2_{jl} f + \partial^2_{jl} f \partial^2_{ik} f\\ &= 2R_{ijkl}.\end{aligned}$$

This proves, at least for a Riemannian manifold that can be isometrically embedded as a hypersurface in Euclidean space, that the Riemann curvature tensor is an intrinsic geometric invariant.

8. The Riemann curvature of an abstract Riemannian manifold

Let M be a n-manifold with Riemannian metric g. Given local coordinates $x^1, \ldots, x^n$, the metric is

$$g = g_{ij}(x)\, dx^i\, dx^j,$$

where $g_{ij}(x) = g(\partial_i, \partial_j)$. The calculations in the previous section suggest the following: Given $p \in M$, let $x = (x^1, \ldots, x^n)$ be coordinates such that $x(p) = 0$, $g_{ij}(0) = \delta_{ij}$ and $\partial_k g_{ij}(0) = 0$, for all $1 \le i, j, k \le n$. Then define the Riemann curvature tensor at p to be given by

$$R(v_1, v_2, v_3, v_4) = R_{ijkl} v_1^i v_2^j v_3^k v_4^l,$$

where, at $x = 0$,

$$R_{ijkl} = -\frac{1}{2}(\partial^2_{ik} g_{jl} + \partial^2_{jl} g_{ik} - \partial^2_{il} g_{jk} - \partial^2_{jk} g_{il}).$$

To prove that this is a well defined tensor, it suffices to prove the following lemmas, which have straightforward proofs.

Lemma 1. *Let g be a smooth Riemannian metric on an n-manifold M. For each $p \in M$, there exist local coordinates $x = (x^1, \ldots, x^n)$ such that $x(p) = 0$ and $g = g_{ij}\, dx^i\, dx^j$, where, for every $1 \le i, j, k \le n$,*

$$\begin{aligned} g_{ij}(0) &= \delta_{ij} \\ \partial_k g_{ij}(0) &= 0. \end{aligned}$$

Remark. Exponential or normal coordinates satisfy Lemma 1. However, a direct proof of Lemma 1 is a lot simpler than the construction of exponential coordinates.

Lemma 2. *Let M and N be smooth n-manifolds, h be a Riemannian metric on N, $\Phi : M \to N$ a smooth map, and $g = \Phi^* h$. Let $x = (x^1, \ldots, x^n)$ be local coordinates on a neighborhood of $p \in M$ and $y = (y^1, \ldots, y^n)$ be local coordinates on a neighborhood of $\Phi(p) \in N$ such that*

$$\begin{aligned} x(p) &= 0 \\ y(\Phi(p)) &= 0 \\ \frac{\partial y^i}{\partial x^j}(0) &= \delta^i_j \\ h_{ij}(0) &= \delta_{ij} \\ \partial_k h_{ij}(0) &= 0. \end{aligned}$$

Then at $x = y = 0$,

$$\begin{aligned} g_{ij}(0) &= \delta_{ij} \\ \partial_k g_{ij}(0) &= 0 \\ \partial^2_{ik} g_{jl} + \partial^2_{jl} g_{ik} - \partial^2_{il} g_{jk} - \partial^2_{jk} g_{il} &= \partial^2_{ik} h_{jl} + \partial^2_{jl} h_{ik} - \partial^2_{il} h_{jk} - \partial^2_{jk} h_{il}. \end{aligned}$$

9. Gauss curvature

Department of Mathematics, New York University

+,-.+ + +/ + !% 6 ! #2 IO 3"

+,-.+ + +/ + !% 2O " 0 3"

!)

I>C $ z !%

6 : $ z !%

:1 1 2 D 8 N !%

" E 8 :1 1 2 D 8 2I 2 $

? " . ? @ "? ? $

@//A (# * %B+@//A 2 2 (+@ 7 2) 2 + ? "
. ? @ "? ? +

+,-.+ + +/ + !) ((# !## 3"

+,-.+ + +/ + !) @ 0 3 ?
" w ? *M 3 @ G " 6 3"

(+,-.+ + +/ + !) > y 3 2
" 3 " 8 ? %M 3 3"

# +,-.+ + +/ + !) : ?M 3 ?M " 3 ? 6?
7 = 7 3"

+,-.+ + +/ + !) ((* #)(* 3"

& 3

+,-.+ + +/ + !% : 3"

O #$ $ *

Chapter 5

Chapter 5 - Part 1

Volume-06---20211206

! ! " " # $ % & '
! ! ! " " (

') * + , - . !

' / 0 1 ! , 2 3 4 - 5 ! 3 6 7 !
!

8 ! 9 8 ! :

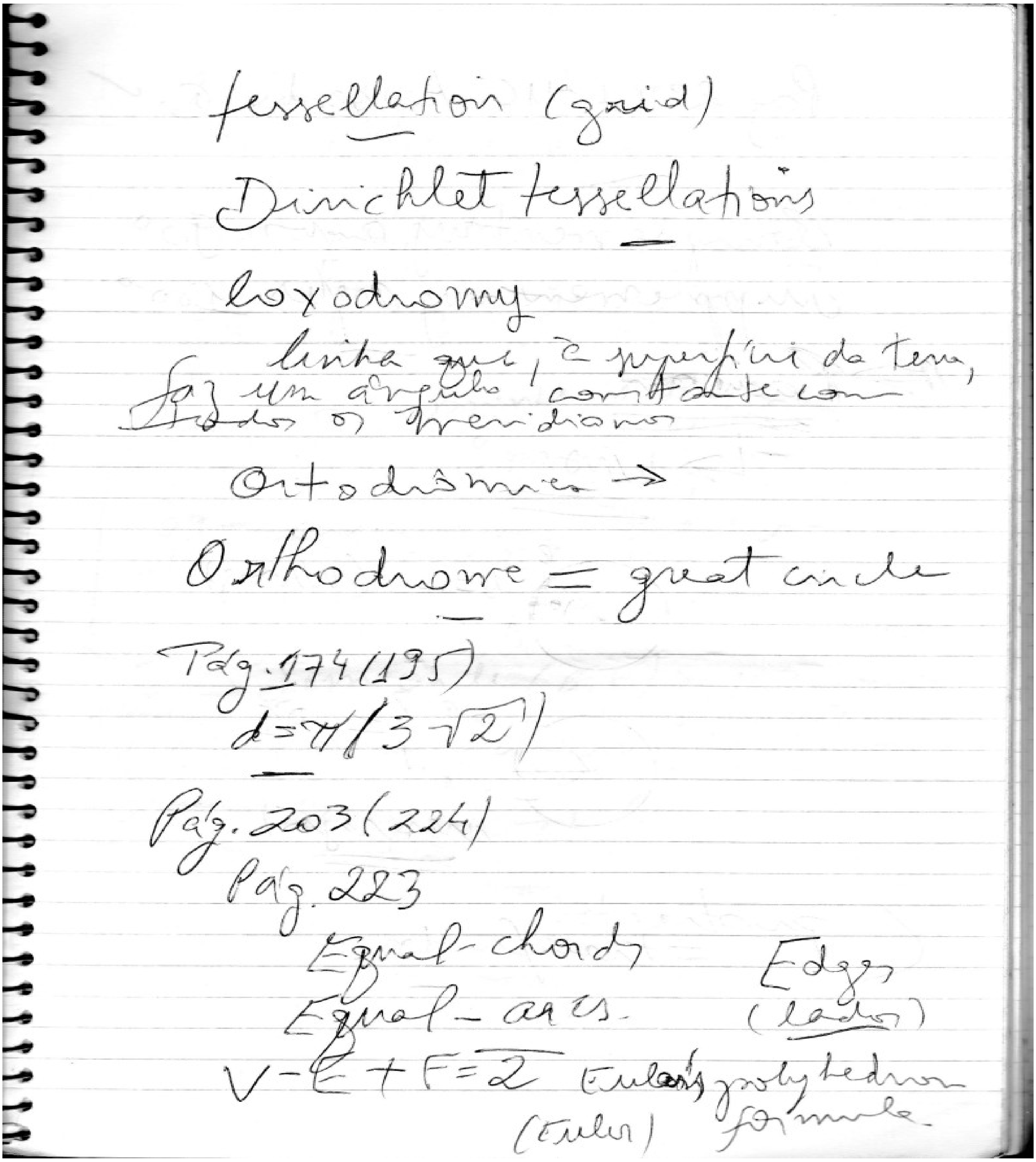

tessellation (grid)

Dirichlet tessellations

loxodromy

linha que, à superfície da Terra, faz um ângulo constante com todos os meridianos

Ortodrômica →

Orthodrome = great circle

Pág. 174 (195)

$d = \pi/(3\sqrt{2})$

Pág. 203 (224)

Pág. 223

Equal-chords

Equal-arcs.

Edges (lados)

$V - E + F = 2$ Euler's polyhedron formula

(Euler)

Pag 224 + 416 Appendix ①.

Complementary angles 90°
Supplementary angles 180°

⇒ **Radian**

↳ **wrong:**

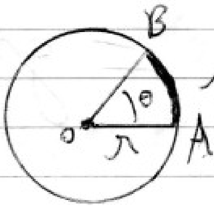

s=r (wrong)

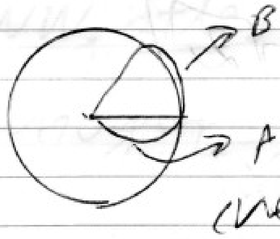

= ok

(ver razão)

(protractor
= transferidor

& /

20200812

01 - Circle (A, B) = $x^2 + y^2 = 1$

02 - Semicircle (B, C)

↳ Select two end points

03 - Perpendicular Line (D, x Axis)

04 - Semicircle (B, F)

05 - Circular Arc (H, G, C)

↳ Select center and two points

06 - Semicircle (B, I)

07 - Semicircle (I, D)

08 - Semicircle (B, A)

09 - Circular Arc (I, C, B)

10 - Circle (I, D, J)

Select through 3 Points

Vejamos:

Semicircle (B, F) ao ser comparado com o Semicircle (I, D), observamos uma não coincidência das linhas na parte de baixo dos referidos semicircles.

7a e 7b

—

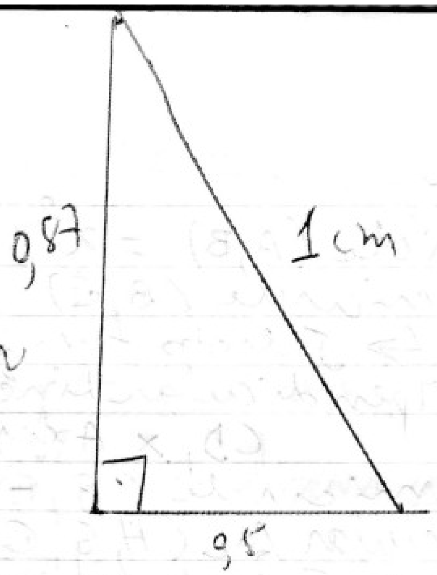

Calcular para checar

Wrong!!!

$a^2 = b^2 + c^2$

$\underline{1} = (0,5)^2 + c^2$

$1 - (0,5)^2 = c^2$

$1 - 0,25 = c^2$

$\underline{1^2} = 2$ e não $\underline{1}$ (usei Pitágoras)

$1 = \sqrt{a^2 + b^2}$

$b = \sqrt{c^2 - a^2}$

$b = \sqrt{1^2 - (0,5)^2}$

$\sqrt{1 - 0,25}$

$b = \sqrt{0,75}$

0,866....

$c = \sqrt{a^2 + b^2}$

$a = \sqrt{c^2 - b^2}$

$b = \sqrt{c^2 - a^2}$

%
;<)=; ; ;8 ; % ! ! + 3 3
0 > ? + ! - , @@ !.A

;<)=; ; ;8 ; % ! ! + 3 3
0 > ? + ! - , @@ -! ' (

PREFACE

> *Symmetry, as wide or as narrow as you may define its meaning, is one idea by which man through the ages has tried to comprehend and create order, beauty, and perfection.*
>
> —HERMANN WEYL

The unifying concept underlying fractals, chaos, and power laws is self-similarity. Self-similarity, or invariance against changes in scale or size, is an attribute of many laws of nature and innumerable phenomena in the world around us. Self-similarity is, in fact, one of the decisive symmetries that shape our universe and our efforts to comprehend it.

Symmetry itself is one of the most fundamental and fruitful concepts of human thought [Wey 81]. By symmetry we mean an *invariance* against change: something stays the same, in spite of some potentially consequential alteration. Mirror symmetry, that is, invariance against "flipping sides," is perhaps the most widely noticed symmetry. Nature built many of her organisms in nearly symmetric ways, and most fundamental laws of physics, such as Newton's law of gravitation, have an *exact* mirror symmetry: there is no difference between left and right in the attraction of heavenly (and most earthbound) bodies. However, the nonconservation of parity in radioactive decay—that is, the *violation* of point symmetry in the "weak" interactions—has finally taught even the physicists to take the distinction between right and left seriously.

Another important symmetry is invariance with respect to geometric translation. Our trust in invariance under transpositions in space and time is, in fact, so unlimited that we believe that the laws of nature are the same all over the cosmos—and that they have been, and will remain so, for all time.

An equally momentous symmetry is invariance with respect to rotation. A circle is invariant under rotation around its center by any angle. A square can be rotated only through angles that are multiples of $360°/4 = 90°$; it is said to have a fourfold symmetry axis. A regular hexagon has a sixfold symmetry. While the rotational symmetry of a flower or a starfish may be imperfect, the exact isotropy found in the fundamental laws of nature is one of the most powerful principles in elucidating the structure of individual atoms, complicated molecules, and entire crystals. Transposition, rotation, and mirror symmetries, acting together, shape crystals from diamonds to snowflakes. And the same three symmetries govern much of what we find pleasing in ornamental designs.

An even more astounding symmetry is the exact identity of like elementary particles. There simply is no difference between an electron here and an electron there—on a distant star, for example. In fact, the perfect identity of photons, the particles of light, has disqualified them from being counted as so many identifiable individuals, resulting in a new kind of particle statistics, discovered by S. N. Bose and rendered palatable by Einstein—a way of counting not heretofore encountered in a world filled with tangible objects.

It was one of the greatest mathematicians of our century, Emmy Noether, who first pointed out the connection between the symmetries of the fundamental laws of physics with respect to displacements in space and time and rotations, on the one hand, and the conservation of linear momentum, energy, and angular momentum on the other. (Noether taught at Göttingen, where David Hilbert, overcoming obstinate prejùdice, had finally secured a faculty position for her. In the dismantling of German science in 1933, she was forced to leave Gottingen. She died at Bryn Mawr in 1935.)

Other symmetries have had equally profound consequences in our understanding of the universe we inhabit. Invariance against uniform motion has given us special relativity, a fusion of space and time into space-time and, as its best-known consequence, the equation $E = mc^2$. The equivalence of acceleration and gravity postulated by Einstein is the basis of his general theory of relativity, which further revolutionized our appreciation of space, time, and matter.

Yet, among all these symmetries flowering in the Garden of Invariance, there sprouts one that, until recently, has not been sufficiently cherished: the ubiquitous invariance against changes in size, called *self-similarity* or, if more than one scale factor is involved, *self-affinity*. The enormously fruitful concepts of self-similarity and self-affinity pervade nature from the distribution of atoms in matter to that of the galaxies in the universe. And in mathematics, too, self-similarity is deeply entrenched. Some 300 years ago the German philosopher and polymath Gottfried Wilhelm Leibniz used the scaling invariance of the infinitely long straight line for its definition. Cantor sets and Weierstrass functions are other early examples—albeit less smooth—of self-similar structures in mathematics, later joined by Julia sets and other marvels of set theory.

It is perhaps symptomatic that with *set theory* still another abstract branch of mathematics has penetrated the real world. There simply seems to be no limit to Eugene Wigner's "unreasonable effectiveness" of mathematics. Indeed, who would have thought that such utterly mathematical constructions as *Cantor sets*, invented solely to reassure the skeptics that sets could both have zero measure and still be uncountable, would make a real difference in any practical realm, let alone become a pivotal concept? Yet this is precisely what happened for many natural phenomena from gelation, polymerization, and coagulation in colloidal physics and chemistry to nonlinear systems in innumerable branches of science. Percolation, dendritic growth, fracture surfaces, electrical discharges (lightnings and Lichtenberg figures), and the composition of quasicrystals are best described by set-theoretic constructs.

Or take the weird functions Karl Weierstrass invented a hundred years ago purely to prove that a function could be both everywhere continuous and yet nowhere differentiable. The fact that such an analytic pathology describes something in the real world—nay, is *elemental* to understanding the strange attractors of nonlinear dynamic systems (such as the double swing and the three-body problem)—gives one pause.

The word *symmetric* is of ancient Greek parentage and means well-proportioned, well-ordered—certainly nothing even remotely chaotic. Yet, paradoxically, self-similarity, the topic of this tome, alone among all the symmetries gives birth to its very antithesis: *chaos*, a state of utter confusion and disorder. As we shall endeavor to show, the genesis of chaos is, in fact, closely related to self-similarity and its inherent lack of "smoothness."

Perhaps not surprisingly, self-similarity entails numerous paradoxes in measurements of time, length, and even musical pitch. Think of Zeno's tardy turtle, pursued—but never overtaken—by swift Achilles. Why do certain lengths increase without bound when we measure them with ever smaller yardsticks? How would Euclid have explained plane geometric figures whose areas scale not as the squares of their apparent perimeters but as some lesser power, such as 1.77 and other fractional exponents? What should we think of musical sounds that, when scaled *up* in frequency, sound—incredibly—*lower* in pitch? How are such monstrosities possible? And how can we describe them in a consistent, meaningful manner?

Here a particularly felicitous thought by Felix Hausdorff comes to the rescue. His and Abram Besicovitch's new ways of looking at dimension dethrones it from its integer position and propels it into the realm of real numbers, giving us one of the sharpest tools—the *Hausdorff dimension* and its ramifications—with which to attack the strange sets that self-similarity breeds.

And while recalling some of the glorious names of the past, we should never forget our great contemporary, the inimitable Benoit B. Mandelbrot, who, single-handed, rescued set theory's most brittle functions and "dustiest" sets from near-oblivion and planted them right in the middle of our daily experience and consciousness. Yes, for all these years, we *have* been living with fractal arteries, not far from fractal river systems draining fractal mountainscapes under fractal clouds, toward fractal coastlines. But, kin to Molière's would-be gentleman, we lacked the proper prose—*fractal*, noun and adjective—that Benoit B. begot.

But our story also has a silent and immobile hero: the digital computer. There can be little doubt that computers have acted as the most forceful forceps in extracting fractals from the dark recesses of abstract mathematics and delivering their geometric intricacies into bright daylight. In fact, the impact of fractal images, often of unimagined beauty and appeal, has given computer graphics a surprising new dimension.

Synopsis

We open our treatise with one of the most charming uses that similarity was ever put too: the young Einstein's proof of Pythagoras's theorem. By adding just a single straight line, in the right place, to a right triangle and applying plenty of similarity, the popular theorem is proved without further prodding.

We then invade the unlimited domain of *self*-similarity as manifest in fractals, multifractals, and the scaling laws of physics, psychophysics, and boundless other fields.

In phase spaces, we encounter deterministic chaos and strange attractors. Percolation and other phase transitions lead us to critical exponents and a hierarchy of different dimensions. Following Poincaré, we immerse ourselves in the self-similarities of iterated mappings, from baker transformations and Bernoulli shifts to logistic parabolas and circle maps. Neither tori, *cantori*, nor Arnold tongues will faze us as we (sur)mount devil's staircases to unwind among the rational winding numbers festooning Farey trees.

And when we talk about nonlinear dynamics we must remember some of the great contributors of recent vintage: Siegel, Moser, Lorenz, Wilson, Feigenbaum, and—last but not least—the great Russian "school" exemplified by such names as Lyapunov, Arnol'd, Sinai, Chirikov, Alexeev, Anosov, Pesin, and the recently deceased master mathematician Kolmogorov.

Cayley trees, also known as Bethe lattices, will provide us with a fitting point of departure for many a practical fractal, such as our bronchial and vascular systems. Cellular automata concern us as models of both biological growth and chemical reactions.

We are also strangely attracted to symbolic dynamics, kneading (and needing) the Morse-Thue sequence and, especially, the Fibonacci rabbit sequence and their discrete self-similarities that, indiscreetly, tell us so much about period doubling, mode locking, frustrated Ising spins, and fivefold symmetric quasicrystals. Many of these subjects were shrouded in mystery and beset by paradoxes before the sharp scalpels, fashioned by scaling and renormalization theories, revealed the underlying tissue and made them tractable. In fact, it is no accident that viable fundamental field theories in physics are renormalizable, as they must be if they are to shun sham scales.

And, of course, we will not hesitate to run down random fractals, from Brownian motion to diffusion-limited aggregation and stock market hiccups (some hiccups of late!). The poor gambler's ruin and the St. Petersburg paradox will provide further food for fractal reflections.

These, then, are some of the exciting, and sobering, themes sounded in the present volume. The aim of this exposition is to enhance the reader's understanding of self-similarity, perhaps the most pregnant of all of nature's symmetries, and to illustrate the wide-ranging applications of scaling invariance in physics, chemistry, biology, music, and—particularly—the visual arts, as manifested in the recent renaissance of computer graphics through fractal images and their iterative beauty.

Acknowledgments

This book owes its existence to many sources. Apart from a brief encounter, in my dissertation, with chaos among the normal modes of concert halls, a "nonintegrable" system if there ever was one, my main stimulus came from the early demonstrations by Heinz-Otto Peitgen and Peter Richter of fractal Julia sets. Their beautiful images, and the intriguing mathematics which underlies them, as epitomized in their book *The Beauty of Fractals*, have made a lasting impression on me.

My first meeting with Mandelbrot's work was his analysis of word frequencies in natural and artificial languages, which touched upon my own interests in computer speech synthesis and recognition. Mandelbrot's monumental monograph *The Fractal Geometry of Nature* influenced me immeasurably, as it did so many other people.

I have also greatly benefited from Robert Devaney's books and lectures on chaos, Jens Feder's *Fractals*, Michael Barnsley's *Fractals Everywhere*, and Dietrich Stauffer's charming *Introduction to Percolation Theory*.

I learned a lot about new developments during the 1988 Gordon Conference on Fractals, organized by Richard Voss and Paul Meakin, which brought together many of the world's outstanding experts in the field.

At the University of Göttingen, it was mainly through the work of Werner Lauterborn that I saw chaos in action in nonlinear dynamic systems from cavitation bubbles to Toda chains. My own students at the Drittes Physikalisches Institut provided both stimulus and rectification during a series of lecture courses on self-similarity, fractals, and chaos. Heinrich Henze and Karl Lautscham made skillful additions to the demonstrations that accompanied these lectures. Holger Behme, Wolf Dieter Brandt, and Tino Gramss contributed many of the computer graphics in this book.

The late Walter Kaufmann-Bühler of Springer Verlag, New York, and Ronald Graham, a longtime colleague at Bell Laboratories, provided early encouragement.

Hildegard Franks and Esperanza Plata in Murray Hill, New Jersey, supported by Maija Sutter and Irena Schönke in Göttingen, converted nearly illegible

1
CHAPTER

Introduction

> *I want to know how God created this world. I am not interested in this or that phenomenon, in the spectrum of this or that element. I want to know His thoughts; the rest are details.*
>
> —ALBERT EINSTEIN

Nature abounds with periodic phenomena: from the motion of a swing to the oscillations of atoms, from the chirping of a grasshopper to the orbits of the heavenly bodies. And our terrestrial bodies, too, participate in this universal minuet—from the heart beat and circadian rhythms to monthly and even longer cycles.

Of course, nothing in nature is *exactly* periodic. All motion has a beginning and an end, so that, in the mathematical sense, strict periodicity does not exist in the real world. Nevertheless, periodicity has proved to be a supremely useful concept in elucidating underlying laws and mechanisms in many fields.

One reason for the universality of simple harmonic motion is the linearity—or near-linearity—of many physical systems and the invariance with displacement in space and time of the laws governing their behavior.

But there are numerous other phenomena in which linearity breaks down and, instead of periodicity, we get aperiodic or even chaotic motion: the smooth waves on a well-behaved lake turn to violent turbulence in the mountain brook, and the daily sunrise, the paradigm of predictability, is overshadowed by cloud formations, a haven for *chaos*—albeit *deterministic* chaos.

But no matter how chaotic life gets, with all regularity gone to bits, another fundamental bulwark often remains unshaken, rising above the turbulent chaos: *self-similarity*, an invariance with respect to *scaling*; in other words, invariance not with *additive* translations, but invariance with *multiplicative* changes of scale. In short, a self-similar object appears unchanged after increasing or shrinking its size. Indeed, in turbulent flows, large eddies beget smaller ones, and these spawn smaller ones still—and so on *ad infinitum* (almost). In general, one of the conspicuous consequences of self-similarity is the appearance of exceedingly fine-grained structures, now generally called *fractals* after Benoit B. Mandelbrot, the father of fractals [Man 83].[1]

Many laws of nature are independent, or nearly so, of a scaling factor. The fact that scaling usually has a limit (Planck's constant, when things get too small, or the speed of light, when objects fly too fast) does no harm to the usefulness of "thinking self-similar," just as the lack (outside mathematics) of strict periodicity is no great impediment in the real world. In a sense, self-similarity is akin to periodicity on a *logarithmic* scale.

Self-similarity, strict or otherwise, reigns in many fields in many guises, and in this book we shall explore some of the many manifestations of self-similarity in the world around us. Among the topics treated are the following:

- Scaling laws and their exponents in physics, psychophysics, and physiology
- Random walks in the stock market and under the microscope; floods, forest fires, the distribution of galaxies, and other "accidents" with statistical self-similarity
- Scaling invariance, self-similarity, and some of their mathematical models, such as Cantor sets and Julia sets
- Fractals and their characterization by Hausdorff, and other noninteger dimensions; fractal paradoxes and their resolution; Weierstrass functions and Hilbert curves; Koch flakes, Sierpinski gaskets, and other non-Euclidean constructions in two and more dimensions; fat fractals and multifractals
- Iterated mappings and a selection of the ensuing self-similarities
- The logistic parabola and other unimodal maps with universal scaling laws; period-doubling bifurcation to chaos; the Feigenbaum constant, symbolic dynamics and the Morse-Thue sequence; Sharkovskii's universal ordering of orbits
- Complexification of the quadratic map and the Mandelbrot set
- Devil's staircases, Farey trees, Arnold tongues, and modelocking; the "rabbit sequence" and the quasi-periodic route to temporal and spatial chaos; Ising spins and quasicrystals
- Laplace's triangle and cellular automata

---

[1] References in brackets are listed alphabetically at the end of the book. The numbers refer to the year of publication.

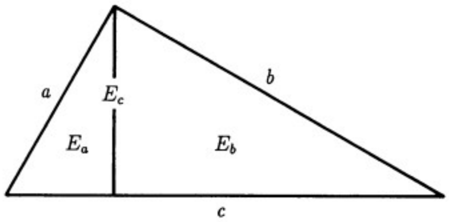

Figure 1 Pythagoras's theorem: sketch for proof by the 11-year-old Einstein based on similarity.

In this chapter some of these topics are introduced informally, together with the leading dramatis personae.

Einstein, Pythagoras, and Simple Similarity

I will a little think.
—ALBERT EINSTEIN, in America

When Jacob Einstein taught (Euclidean) geometry to his 11-year-old nephew Albert, the young Einstein—even then striving for utmost parsimony—felt that some of Euclid's proofs were unnecessarily complicated.[2] For example, in a typical proof of Pythagoras's theorem $a^2 + b^2 = c^2$, was it really mandatory to have all those extra lines, angles, and squares in addition to the basic right triangle with hypotenuse, c and sides a and b?

After "a little thinking," the sharp youngster came up with a proof that required only *one* additional line, the altitude above the hypotenuse (see Figure 1). This height divides the large triangle into two smaller triangles that are *similar* to each other and similar to the large triangle. (Triangles are similar if their angles are the same, which is easily seen to be the case in Figure 1.)

Now, in Euclidean geometry, the area ratio of two similar (closed) figures is equal to the *square* of the ratio of corresponding *linear* dimensions. Thus, the areas E_a, E_b, and E_c (E as in German *Ebene*) of the three triangles in Figure 1 are related to their hypotenuses a, b, and c by the following equations:

$$E_a = ma^2 \tag{1}$$

---

[2] I have the story from Schneior Lifson of the Weizmann Institute in Tel Aviv, who has it from Einstein's assistant Ernst Strauss, to whom it was told by old Albert himself.

$$E_b = mb^2 \tag{2}$$

$$E_c = mc^2 \tag{3}$$

where m is a dimensionless nonzero multiplier that is the *same in all three equations.*

Now a second look at Figure 1 will reveal that the area of the large triangle is, of course, the sum of the areas of the two smaller triangles,

$$E_a + E_b = E_c$$

or, with equations 1 to 3,

$$ma^2 + mb^2 = mc^2$$

Dividing this identity by the common measure m promptly produces Pythagoras's renowned result

$$a^2 + b^2 = c^2$$

proved here by an 11-year old person[3] by combining two fertile scientific principles that were going to stand the grown-up Einstein in good stead: simplicity and symmetry, of which self-similarity is a special case. Yet the true beauty of Einstein's proof is not that it is so simple, but that it exposes the true essence of Pythagoras's theorem: *similarity* and *scaling*.

The resemblance of equation 3 to Einstein's later discovery, his famous $E = mc^2$, is of course entirely fortuitous. The equivalence of mass m and energy E, which is at the basis of nuclear power in all its guises, is a consequence of Lorentz invariance. This invariance, which underlies special relativity, was predicted by Einstein in 1905 after, it seems, several false starts and a "little more thinking" (see Figure 2).

A Self-Similar Array of Self-Preserving Queens

One of numerous chess problems is the placement of as many queens as possible on a chessboard of a given size so that no queen "attacks" (shares a row, column, or $\pm 45°$ diagonal with) any other queen. For a $k \times k$ square board, there can be at most k nonattacking queens. But are k peacefully coexisting queens always possible? What if k is very large? Doesn't the complexity of placing the queens grow exponentially with the size of the board? As we shall see, the placement is actually very simple, even for arbitrarily large boards, if we focus our attention

---

[3] Really a "nonperson" at that stage, considering the neglect he suffered in his Munich high school [Pye 85].

' $! ,
;B<C< D A A / ;* E / ; ; !
- , ; * ! , ! F)
-!
' / %$ D8D <)CF &<D 88)B&D<CF G CF

and given by

$$\int_V \partial_i f\, dV = \oint_S f\, dS_i, \tag{1.139a}$$

$$\int_S \varepsilon_{ijk}\partial_k f\, dS_j = \oint_\Gamma f\, dx^i, \tag{1.139b}$$

where, as in the case of the divergence and curl theorems, S is the boundary of V for the first equation, while Γ is the boundary of S in the second. The proofs of these relations are completely analogous to those of the earlier theorems. In the case of the divergence theorem, the only modification is that we only need to consider one term, and we integrate in the x^i direction first. In fact, we could prove the divergence theorem by first proving the more general form and afterwards letting f be the jth component of a vector $\vec{v}$. We would obtain

$$I_i^j = \int_V \partial_i v^j dV = \oint_S v^j dS_i \tag{1.140}$$

and upon letting $i = j$ and summing

$$I_i^i = \int_V \partial_i v^i dV = \int_V \nabla \cdot \vec{v}\, dV = \oint_S v^i dS_i = \oint_S \vec{v} \cdot d\vec{S}. \tag{1.141}$$

Proving the generalised curl theorem starting from the regular curl theorem is left as Problem 1.44.

1.6 NON-CARTESIAN COORDINATE SYSTEMS

While Cartesian coordinate systems provide a solid basis for dealing with scalars, vectors, and later on tensors, in affine spaces, we will encounter problems that are easier to describe and solve using a different set of coordinates that are not necessarily linear or even orthogonal. In several cases, we will also be able to apply symmetry arguments in a particular set of coordinates that are not as easily applicable in other coordinate systems.

1.6.1 General theory

A coordinate system of an N dimensional affine space is a labelling of all the points in the space by N numbers, which we call *coordinates*. Each combination of possible coordinates should uniquely define a point in the space. Each y^a is a continuous function called a *coordinate function* from the N dimensional space to the real numbers. From this point on, we will adopt the convention of using early letters in the alphabet (a, b, c, ...) as indices for general coordinate systems and reserve letters from the middle of the alphabet (i, j, k, ...) for indices in Cartesian coordinates.

Example 1.25 Given an affine N dimensional space, we can always define a Cartesian coordinate system as we have done previously. We do so by selecting one point in the affine space as the origin O and selecting a set of orthonormal basis vectors $\vec{e}_i$. The coordinates are then defined by mapping the coordinates x^i onto the point P such that

$$\overrightarrow{OP} = \vec{x} = \vec{e}_i x^i, \tag{1.142}$$

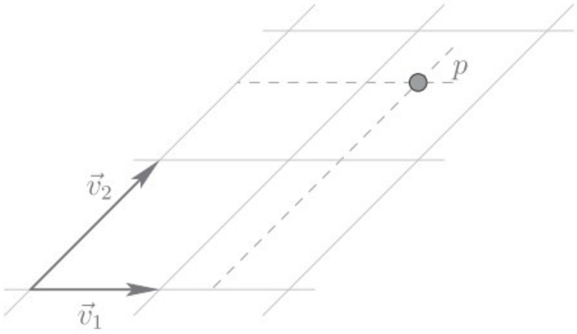

Figure 1.16 A coordinate system for a two-dimensional affine space may be based on the linearly independent, but not orthogonal, set of vectors $\vec{v}_1$ and $\vec{v}_2$. In this case, the coordinates of the point p would be $y^1 = 1.4$, $y^2 = 1.6$.

where $\overrightarrow{OP}$ is the difference vector between P and O. The coordinate functions are given by

$$x^i(P) = \vec{e}_i \cdot \overrightarrow{OP}. \tag{1.143}$$

Since we can always start by writing down a Cartesian coordinate system for an affine space, the coordinates of a more general coordinate system may be written in terms of these Cartesian coordinates for simplicity.

Example 1.26 We can create a coordinate system that is not based on orthogonal basis vectors in a similar way to how we created the Cartesian coordinate system. Instead of a set of orthonormal basis vectors, it suffices to pick a set of linearly independent vectors $\vec{v}_a$. The mapping of the coordinates y^a to the point P is given by

$$\overrightarrow{OP} = \vec{x} = \vec{v}_a y^a. \tag{1.144}$$

However, note that the coordinate functions are not given by $y^a = \vec{v}_a \cdot \overrightarrow{OP}$ in this case as the vectors $\vec{v}_a$ are not necessarily orthogonal or even normalised. For concreteness, let us take a two-dimensional space parametrised using the basis $\vec{v}_1 = \vec{e}_1$ and $\vec{v}_2 = \vec{e}_1 + \vec{e}_2$, where $\vec{e}_1$ and $\vec{e}_2$ form an orthonormal basis, see Fig. 1.16. We obtain

$$\vec{x} = y^1\vec{e}_1 + y^2(\vec{e}_1 + \vec{e}_2) = \vec{e}_1(y^1 + y^2) + \vec{e}_2 y^2. \tag{1.145}$$

From this we can deduce that the coordinate functions are given by

$$y^1 = (\vec{e}_1 - \vec{e}_2) \cdot \vec{x} = x^1 - x^2 \neq \vec{v}_1 \cdot \vec{x}, \tag{1.146a}$$

$$y^2 = \vec{e}_2 \cdot \vec{x} = x^2 \neq \vec{v}_2 \cdot \vec{x}. \tag{1.146b}$$

Equivalently, we could have used

$$y^1 = x^1 - x^2, \tag{1.147a}$$

$$y^2 = x^2, \tag{1.147b}$$

to *define* the coordinate functions y^1 and y^2. In what follows we will discuss how this applies to the general setting with an arbitrary coordinate system.

1.6.1.1 Tangent vector basis

As indicated above, we will now assume that the coordinate functions y^a are known functions of the Cartesian coordinates x^i. Given these functions, we define the *coordinate lines* for the y^a coordinate as the curves for which y^b is constant for all $b \neq a$. The *tangent vector* of this coordinate line is given by

$$\vec{E}_a = \frac{\partial \vec{x}}{\partial y^a} = \frac{\partial x^i}{\partial y^a}\vec{e}_i, \tag{1.148}$$

where $\vec{e}_i$ are the Cartesian basis vectors, which are constant. The partial derivative in this expression is the partial derivative of the Cartesian coordinate functions expressed as functions of the y coordinates. In order for the coordinate system to uniquely define a point, the tangent vectors $\vec{E}_a$ must be linearly independent. We can use this fact in order to use the $\vec{E}_a$ vectors as a set of basis vectors and express any vector as a linear combination of these basis vectors. This presents us with a conundrum as the vectors $\vec{E}_a$ may generally depend on the coordinates. The way forward here is to realise that most vectors are defined only at a given point in space. In particular, for vector fields, it is natural to use the basis defined at a particular spatial point as the basis for the fields at that point. In terms of components, we would write

$$\vec{v}(\vec{x}) = v^a(\vec{x})\vec{E}_a(\vec{x}). \tag{1.149}$$

In what follows, we will often suppress the explicit mention of the spatial dependence of both vector components and basis vectors, i.e., we will assume that $v^a \equiv v^a(\vec{x})$ and $\vec{E}_a \equiv \vec{E}_a(\vec{x})$.

The set of vectors $\vec{E}_a$ is called a *tangent vector basis* and the vector components v^a are called *contravariant vector components*. The reason for this nomenclature is that a vector is not dependent upon the basis used to represent it and in order to ensure that this is the case, the vector components must transform in the opposite way as compared to the basis vectors under a coordinate change. In particular for a vector $\vec{v}$

$$\vec{v} = v^a\vec{E}_a = v^a\frac{\partial x^i}{\partial y^a}\vec{e}_i = v^i\vec{e}_i. \tag{1.150}$$

Scalar multiplication with $\vec{e}_j$ now leads to

$$v^j = \frac{\partial x^j}{\partial y^a}v^a \quad \Longrightarrow \quad v^b = \frac{\partial y^b}{\partial x^j}\frac{\partial x^j}{\partial y^a}v^a = \frac{\partial y^b}{\partial x^j}v^j. \tag{1.151}$$

Here we have used the chain rule and the fact that

$$\frac{\partial y^b}{\partial x^j}\frac{\partial x^j}{\partial y^a} = \delta^b_a. \tag{1.152}$$

Note that we have here introduced a slightly new notation for the δ symbol, namely that it has one index up and one index down. As we shall see, the position of the indices (up or down), that we have so far been quite flexible with, plays a crucial role for non-Cartesian coordinate systems. In Eq. (1.151), we have found that the vector components v^a transform with the partial derivative of the y coordinates with respect to the x coordinates instead of vice versa, as the tangent vector basis does.

We also note that the form of Eq. (1.151) is not restricted to when one of the coordinate systems is Cartesian. If we have two coordinate systems with coordinates y^a and $y'^{a'}$, respectively, we obtain

$$v^{a'} = \frac{\partial y'^{a'}}{\partial x^j} v^j = \frac{\partial y'^{a'}}{\partial x^j} \frac{\partial x^j}{\partial y^a} v^a = \frac{\partial y'^{a'}}{\partial y^a} v^a, \tag{1.153}$$

where we have again applied the chain rule for the derivatives.

Example 1.27 In Example 1.26, we used the vectors $\vec{v}_1$ and $\vec{v}_2$ to define a new coordinate system on a two-dimensional affine space. Towards the end of the example, we noted that the coordinate system could also be defined by specifying the coordinate functions as in Eqs. (1.147). Inverting these equations leads to

$$x^1 = y^1 + y^2, \quad x^2 = y^2. \tag{1.154}$$

It follows directly that the tangent vector basis is given by

$$\vec{E}_1 = \frac{\partial \vec{x}}{\partial y^1} = \vec{e}_1 = \vec{v}_1, \quad \vec{E}_2 = \frac{\partial \vec{x}}{\partial y^2} = \vec{e}_1 + \vec{e}_2 = \vec{v}_2. \tag{1.155}$$

Thus, the tangent vector basis is the basis we used to define the coordinate system in Example 1.26. Note that this is a particular result for the situation when the coordinate functions are linear, we call this type of coordinates *affine coordinates*.

1.6.1.2 Dual basis

Apart from the tangent vector basis, there is a different set of basis vectors which we can define using a given coordinate system. As we move on to more advanced topics, the need shall arise to separate the following concept further from the concept of tangent vectors, but for now let us define a set of vector fields by taking the gradient of the coordinate functions

$$\vec{E}^a = \nabla y^a. \tag{1.156}$$

These vector fields also form a basis at each point in the base space and their relation to the Cartesian basis is

$$\vec{e}_i \cdot \vec{E}^a = \frac{\partial y^a}{\partial x^i} \quad \Longrightarrow \quad \vec{E}^a = \frac{\partial y^a}{\partial x^i} \vec{e}_i. \tag{1.157}$$

Note that the vectors $\vec{E}^a$ follow the very same transformation rule as the contravariant vector components and thus transforms in the opposite fashion as compared to the tangent vector basis. This new set of basis vectors is called the *dual basis* and has a set of interesting properties. The first one that we note is that the vectors in the dual basis are not necessarily parallel to the corresponding tangent vectors.

Example 1.28 Returning again to Example 1.26, Eqs. (1.147) directly gives us

$$\vec{E}^1 = \vec{e}_1 - \vec{e}_2, \quad \vec{E}^2 = \vec{e}_2, \tag{1.158}$$

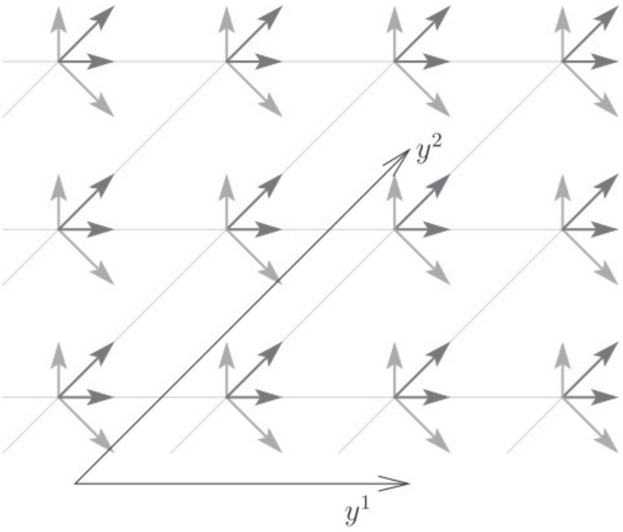

Figure 1.17 The two-dimensional coordinate system $y^1 = x^1 - x^2$, $y^2 = x^2$ along with the associated tangent vector basis (dark) and dual basis (light). The coordinate lines of y^1 are at the same time the level surfaces of y^2. By definition, the tangent vector field $\vec{E}_1$ is tangent to the y^1 coordinate lines while the dual field $\vec{E}^2$ is normal to the y^2 level surfaces. Thus, we must have $\vec{E}_1 \cdot \vec{E}^2 = 0$. The corresponding argument can be made for $\vec{E}_2$ and $\vec{E}^1$.

which are clearly not parallel to $\vec{E}_1$ and $\vec{E}_2$, respectively. The two different bases are illustrated in Fig. 1.17. However, as shown in the figure, the tangent vector field $\vec{E}_1$ is orthogonal to the dual vector field $\vec{E}^2$.

The fact that $\vec{E}_1 \cdot \vec{E}^2 = 0$ in the above example is by no means a coincidence. From the properties of the tangent vector and dual fields, we generally obtain

$$\vec{E}_a \cdot \vec{E}^b = \vec{e}_i \cdot \vec{e}_j \frac{\partial x^i}{\partial y^a} \frac{\partial y^b}{\partial x^j} = \frac{\partial x^i}{\partial y^a} \frac{\partial y^b}{\partial x^i} = \frac{\partial y^b}{\partial y^a} = \delta^b_a. \tag{1.159}$$

Thus, although $\vec{E}^a$ is not generally parallel with $\vec{E}_a$, it is the only dual basis vector that has a non-zero scalar product with $\vec{E}_a$ and this scalar product is equal to one.

Example 1.29 While the tangent vector basis and dual basis are not necessarily parallel to each other, there is an important class of coordinate systems in which they are. An example of this is the coordinate system

$$s = \sinh(x^1), \quad t = \sinh(x^2), \tag{1.160}$$

where we have called the new coordinates s and t rather than y^1 and y^2 in order to avoid confusion in later examples. This coordinate system has the same coordinate lines and coordinate level surfaces as the original Cartesian system as it only rescales the coordinates.

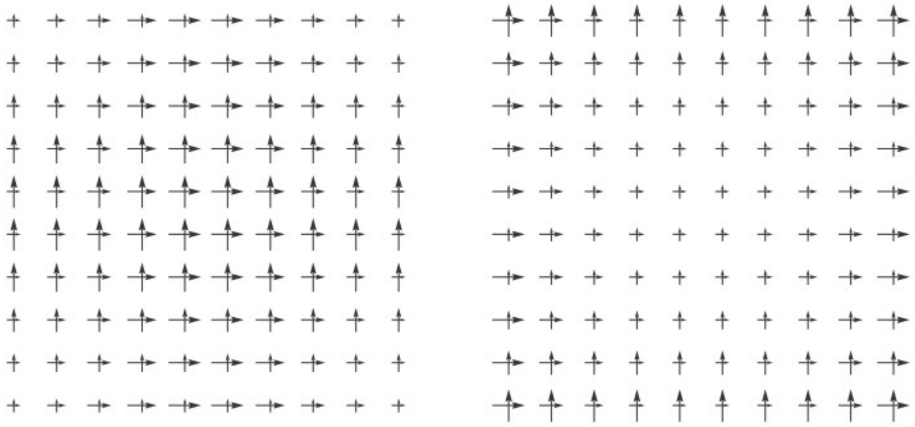

Figure 1.18 The tangent vector basis (left) and dual basis (right) for the coordinate system $s = \sinh(x^1)$, $t = \sinh(x^2)$ with $s = t = 0$ in the middle of the figures (not to scale between the figures). When the magnitude of the tangent vector basis is large, then the magnitude of the dual basis is small and vice versa.

We find that

$$\vec{E}_s = \vec{e}_1 \frac{1}{\sqrt{1+s^2}}, \quad \vec{E}_t = \vec{e}_2 \frac{1}{\sqrt{1+t^2}} \tag{1.161a}$$

$$\vec{E}^s = \vec{e}_1 \sqrt{1+s^2}, \quad \vec{E}^t = \vec{e}_2 \sqrt{1+t^2}, \tag{1.161b}$$

where we have taken the habit of using the coordinate names rather than numbers when speaking about specific components. In Fig. 1.18, we show the tangent vector and dual basis fields for this coordinate system. Another difference from our previous example is that the basis vectors depend on the coordinates, something which is often the case for non-Cartesian coordinate systems.

When it comes to decomposing an arbitrary vector $\vec{v}$ into a linear combination of $\vec{E}^a$ we find that, with $\vec{v} = v_a \vec{E}^a$,

$$\vec{E}_a \cdot \vec{v} = \vec{E}_a \cdot v_b \vec{E}^b = \delta_a^b v_b = v_a. \tag{1.162}$$

Using the expression for $\vec{E}_a$ in Cartesian coordinates, we find that

$$v_a = \frac{\partial x^i}{\partial y^a} \vec{e}_i \cdot \vec{v} = \frac{\partial x^i}{\partial y^a} v^i. \tag{1.163}$$

The components v_a thus transform in the same way as the tangent vector basis under coordinate changes and are therefore called *covariant vector components*.

Before specialising to coordinate systems that fulfil an additional orthogonality criterion, let us examine how to express the scalar product of two general vectors $\vec{v}$ and $\vec{w}$ in a general coordinate system. Writing $\vec{v}$ in terms of the tangent vector basis and $\vec{w}$ in the dual basis, we find that

$$\vec{v} \cdot \vec{w} = v^a w_b \vec{E}_a \cdot \vec{E}^b = v^a w_b \delta^a_b = v^a w_a. \tag{1.164}$$

The scalar product of the two vectors is therefore given by the sum of the products of the covariant and contravariant components of the vectors, respectively. Naturally, there is nothing stopping us from instead expressing $\vec{v}$ in the dual basis and $\vec{w}$ in the tangent vector basis and arriving at $\vec{v} = \vec{w} = v_a w^a$.

1.6.2 Orthogonal coordinates

As we have seen, the tangent vector basis $\vec{E}_a$ need not be orthogonal and, depending on the coordinates, we may end up with $\vec{E}_a \cdot \vec{E}_b \neq 0$ for $a \neq b$. However, there is an important class of coordinate systems where the tangent vector basis is orthogonal as we saw in Example 1.29. A coordinate system where this is the case is called an *orthogonal coordinate system* and in such systems

$$\vec{E}_a \cdot \vec{E}_b = h_a^2 \delta_{ab}, \qquad \text{(no sum)} \tag{1.165}$$

where we have introduced the *scale factors* $h_a = |\vec{E}_a|$, which are the moduli of the tangent vector basis. Since the vectors $\vec{E}_a$ are orthogonal, we can define an *orthonormal basis* $\vec{e}_a$ according to

$$\vec{e}_a = \frac{1}{h_a}\vec{E}_a = \frac{1}{h_a}\frac{\partial \vec{x}}{\partial y^a}. \qquad \text{(no sum)} \tag{1.166}$$

The normalisation condition gives

$$h_a = \sqrt{\sum_i \left(\frac{\partial x^i}{\partial y^a}\right)^2}. \tag{1.167}$$

For our purposes, it is not relevant to work with covariant and contravariant indices once we have established the orthonormal basis $\vec{e}_a$. For the remainder of this chapter, we will therefore use a notation similar to that used for Cartesian coordinates, with the only difference being in the indices used. Naturally, any vector can be written as a linear combination of the $\vec{e}_a$ basis

$$\vec{v} = \tilde{v}_a \vec{e}_a = \sum_a \tilde{v}_a \frac{1}{h_a}\vec{E}_a, \tag{1.168}$$

which leads to the relation $\tilde{v}_a = h_a v^a$ (no sum), where $\tilde{v}_a$ are the *physical components* of the vector $\vec{v}$ in the orthonormal basis, since they are directly related to the projection of the vector onto the basis unit vectors.

Example 1.30 In Example 1.29 we introduced an orthogonal coordinate system with

$$\vec{E}_s = \frac{1}{\sqrt{1+s^2}}\vec{e}_1, \quad \vec{E}_t = \frac{1}{\sqrt{1+t^2}}\vec{e}_2. \tag{1.169}$$

Taking the modulus of these two tangent vectors gives us

$$h_s = \frac{1}{\sqrt{1+s^2}}, \quad h_t = \frac{1}{\sqrt{1+t^2}} \tag{1.170}$$

and thus

$$\vec{e}_s = \vec{e}_1, \quad \vec{e}_t = \vec{e}_2. \tag{1.171}$$

In the above example, the orthonormal basis turned out to be the same as the Cartesian one, but the $\vec{e}_a$ generally depend on the point in the base space, just as $\vec{E}_a$ does, due to the possibility of $\vec{E}_a$ changing direction from point to point.

Let us now find the relation between the dual basis $\vec{E}^a$ and the orthonormal basis $\vec{e}_a$. In order to do this, we expand $\vec{E}^a$ in the orthonormal basis as $\vec{E}^a = \tilde{E}^a_b \vec{e}_b$ and obtain

$$\tilde{E}^a_b = \vec{e}_b \cdot \vec{E}^a = \frac{1}{h_b}\vec{E}_b \cdot \vec{E}^a = \frac{1}{h_b}\delta^a_b. \qquad \text{(no sum)} \tag{1.172}$$

It follows directly that

$$\vec{e}_a = \frac{1}{h_a}\vec{E}_a = h_a \vec{E}^a. \qquad \text{(no sum)} \tag{1.173}$$

Thus, not only is the tangent vector basis orthogonal, it is also parallel to the dual basis, with the only difference between the two being a factor of h_a^2. This also means that we could just as well construct the orthonormal basis starting from the dual basis.

1.6.2.1 Integration in orthogonal coordinates

It is worth spending some time and effort on discussing line, surface and volume integrals in orthogonal coordinate systems. In all cases, the relevant quantity will be

$$\frac{\partial \vec{x}}{\partial y^a} = \vec{E}_a = h_a \vec{e}_a. \qquad \text{(no sum)} \tag{1.174}$$

The form of a line integral does not change significantly, we only need to note that

$$\int_\Gamma \vec{v} \cdot d\vec{x} = \int_\Gamma \vec{v} \cdot \vec{E}_a \dot{y}^a d\tau, \tag{1.175}$$

where $\dot{y}^a = dy^a/d\tau$ and τ is a curve parameter. This lets us express the integral as

$$\int_\Gamma \vec{v} \cdot d\vec{x} = \int_\Gamma \sum_a \tilde{v}_a h_a \dot{y}^a d\tau. \tag{1.176}$$

In particular, if integrating along the y^b coordinate line, we would use y^b as a parameter and thus have

$$\int_\Gamma \vec{v} \cdot d\vec{x} = \int_\Gamma \sum_a \tilde{v}_a h_a \frac{\partial y^a}{\partial y^b} dy^b = \int_\Gamma \tilde{v}_b h_b dy^b. \qquad \text{(no sum)} \tag{1.177}$$

Similar to the above, if we have a surface integral over a coordinate level surface S_c with $y^c =$ constant and the other coordinates being y^a and y^b, which we can use for parametrising the surface, then the surface element will be given by

$$\begin{aligned} d\vec{S} &= \frac{\partial \vec{x}}{\partial y^a} \times \frac{\partial \vec{x}}{\partial y^b} dy^a dy^b = \vec{E}_a \times \vec{E}_b \, dy^a dy^b \\ &= h_a h_b (\vec{e}_a \times \vec{e}_b) dy^a dy^b = h_a h_b \vec{e}_c dy^a dy^b, \qquad \text{(no sum)} \end{aligned} \tag{1.178}$$

assuming that $\vec{e}_a$, $\vec{e}_b$, and $\vec{e}_c$ form a right-handed triplet. Since $\vec{e}_c$ is a unit vector, it is the unit normal and the surface area is $dS = h_a h_b dy^a dy^b$ (no sum). A flux integral over this surface is therefore given by

$$\int_{S_c} \vec{v} \cdot d\vec{S} = \int_{S_c} \tilde{v}_c h_a h_b dy^a dy^b. \qquad \text{(no sum)} \tag{1.179}$$

Finally, parametrising a volume with the coordinates, the volume element dV may be expressed as

$$\begin{aligned} dV &= \frac{\partial \vec{x}}{\partial y^1} \cdot \left(\frac{\partial \vec{x}}{\partial y^2} \times \frac{\partial \vec{x}}{\partial y^3} \right) dy^1 dy^2 dy^3 = \vec{E}_1 \cdot (\vec{E}_2 \times \vec{E}_3) dy^1 dy^2 dy^3 \\ &= h_1 h_2 h_3 dy^1 dy^2 dy^3, \end{aligned} \tag{1.180}$$

with a similar generalisation of $dV = \prod_a h_a dy^a$ in an arbitrary number of dimensions. For ease of notation, we introduce the *Jacobian determinant*

$$\mathcal{J} = \prod_a h_a = h_1 h_2 h_3, \tag{1.181}$$

where the last equality holds in three dimensions, which is the product of all of the scale factors. However, it should be noted that the Jacobian determinant also appears in non-orthogonal coordinates, although it does not take the simple form shown above. In general coordinates, we would have

$$\mathcal{J} = \vec{E}_1 \cdot (\vec{E}_2 \times \vec{E}_3), \tag{1.182}$$

again with a straightforward generalisation to an arbitrary number of dimensions. As a result, a volume integral expressed in an arbitrary coordinate system takes the form

$$\int_\Omega \ldots dV = \int_{\Omega_p} \ldots \mathcal{J}\, dy^1 dy^2 dy^3, \tag{1.183}$$

where Ω_p is the set of coordinates corresponding to the volume Ω.

Example 1.31 There is an intuitive interpretation of the integral expressions derived above. The tangent vector $\vec{E}_a = h_a \vec{e}_a$ (no sum) is the change in the position vector associated to an infinitesimal change in the coordinate y^a. A line element therefore becomes directly proportional to $\vec{E}_a dy^a$. Due to the orthogonality of these changes, a surface element spanned by two of the $\vec{E}_a$ will be equal to the product of the scale factors in magnitude and directed in the third direction and the volume spanned by all three of them will be the product of all of the side lengths of the spanned rectangular box.

As an important final comment on integration in orthogonal coordinate systems, it must be noted that if the integrand is a vector quantity, then the result cannot be obtained by integrating the components of the orthonormal basis $\vec{e}_a$ one by one. The reason for this is that the basis vectors themselves are not constant and may therefore not be taken outside of the integral as in the case of Cartesian coordinates.

1.6.2.2 Differentiation in orthogonal coordinates

Just as integration takes a slightly different form in general coordinate systems as compared to Cartesian coordinates, the differential operators acting on scalar and vector fields take on a slightly modified form, with the appearance of scale factors slightly simplifying the results. The operator that is easiest to transform to orthogonal coordinates is the gradient of a scalar field φ, which will be given by

$$\nabla \varphi = \vec{e}_i \partial_i \varphi = \vec{e}_i \frac{\partial y^a}{\partial x^i} \partial_a \varphi = \vec{E}^a \partial_a \varphi, \tag{1.184}$$

which holds in any coordinate system. It follows that the partial derivatives $\partial_a \varphi$ are naturally covariant vector components as they will follow the covariant transformation rules. Writing this in terms of the orthonormal basis, the *gradient in orthogonal coordinates* is

$$\nabla \varphi = \sum_a \frac{1}{h_a} \vec{e}_a \partial_a \varphi. \tag{1.185}$$

Due to the appearance of scale factors, expressing the divergence and curl in terms of the orthonormal basis $\vec{e}_a$ and its vector components poses a bit more of a challenge and we will be greatly helped by the following argumentation: Let us consider the dual basis $\vec{E}^a = \nabla y^a = \vec{e}_a / h_a$ (no sum). Taking the cross product between two of these vectors, we obtain

$$\vec{E}^a \times \vec{E}^b = \frac{1}{h_a h_b} \vec{e}_a \times \vec{e}_b = \frac{1}{h_a h_b} \sum_c \varepsilon_{abc} \vec{e}_c = \sum_c \frac{h_c}{\mathcal{J}} \varepsilon_{abc} \vec{e}_c, \tag{1.186}$$

where ε_{abc} has the same properties as the permutation symbol in Cartesian coordinates, i.e., it is completely anti-symmetric and equal to one when $\vec{e}_a$, $\vec{e}_b$, and $\vec{e}_c$ form a right-handed system. Taking the divergence of the original expression leads to

$$\nabla \cdot (\vec{E}^a \times \vec{E}^b) = \nabla \cdot (\nabla y^a \times \nabla y^b) = \nabla y^b \cdot (\nabla \times \nabla y^a) - \nabla y^a \cdot (\nabla \times \nabla y^b) = 0, \tag{1.187}$$

due to $\nabla \times \nabla \varphi = 0$ for all functions φ, and thus also for the coordinate functions. Inserting the expression for $\vec{E}^a \times \vec{E}^b$ in terms of the orthonormal basis into this gives the result

$$\sum_c \nabla \cdot \frac{h_c}{\mathcal{J}} \varepsilon_{abc} \vec{e}_c = 0. \tag{1.188}$$

Multiplying this with ε_{dab}, summing over a and b, and making use of the ε-δ-relation we finally arrive at

$$\sum_{abc} \varepsilon_{dab} \varepsilon_{abc} \nabla \cdot \frac{h_c}{\mathcal{J}} \vec{e}_c = 2 \nabla \cdot \frac{h_d}{\mathcal{J}} \vec{e}_d = 0. \tag{1.189}$$

Equipped with this relation, we move on to expressing the *divergence in orthogonal coordinates*. By expanding the vector $\vec{v}$ as $\vec{v} = \tilde{v}_a \vec{e}_a$, we obtain

$$\begin{aligned} \nabla \cdot \vec{v} = \nabla \cdot \tilde{v}_a \vec{e}_a &= \sum_a \nabla \cdot \left(\frac{\tilde{v}_a \mathcal{J}}{h_a} \frac{h_a}{\mathcal{J}} \vec{e}_a \right) = \sum_a \left[\frac{h_a}{\mathcal{J}} \vec{e}_a \cdot \nabla \left(\frac{\tilde{v}_a \mathcal{J}}{h_a} \right) + \frac{\tilde{v}_a \mathcal{J}}{h_a} \nabla \cdot \left(\frac{h_a}{\mathcal{J}} \vec{e}_a \right) \right] \\ &= \sum_a \frac{h_a}{\mathcal{J}} \vec{e}_a \cdot \nabla \left(\frac{\tilde{v}_a \mathcal{J}}{h_a} \right) = \frac{1}{\mathcal{J}} \sum_a \partial_a \left(\frac{\tilde{v}_a \mathcal{J}}{h_a} \right), \end{aligned} \tag{1.190}$$

where in the last step we have used Eq. (1.185) to express the gradient in terms of the orthonormal basis.

Before moving on with the curl, we note that the *Laplace operator in orthogonal coordinates* can be expressed by combining Eqs. (1.185) and (1.190) into

$$\nabla^2 \varphi = \nabla \cdot \nabla \varphi = \frac{1}{\mathcal{J}} \sum_a \partial_a \left(\frac{\mathcal{J}}{h_a^2} \partial_a \varphi \right). \tag{1.191}$$

Finally, the *curl in orthogonal coordinates* can be expressed using the scale factors and vector components $\tilde{v}_a$ by noting that

$$\nabla \times \vec{v} = \nabla \times v_a \vec{E}^a = -\vec{E}^a \times \nabla v_a + v_a \nabla \times \vec{E}^a = -\vec{E}^a \times \nabla v_a. \tag{1.192}$$

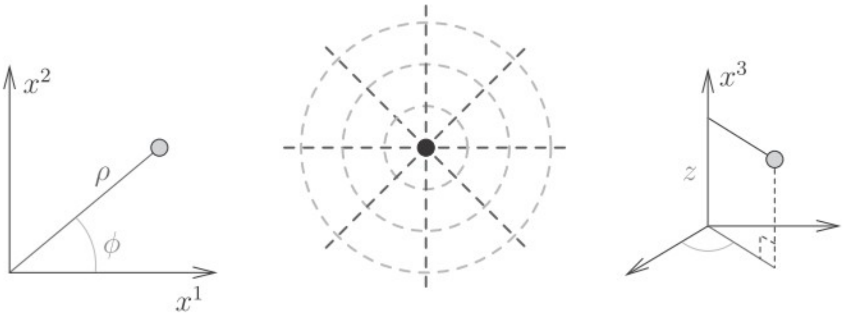

Figure 1.19 The left figure shows the relation between polar coordinates ρ and ϕ and a Cartesian coordinate system in two dimensions. The middle figure shows coordinate lines of ρ (dark) and ϕ (light) based on a polar coordinate system with the black point as origin. A cylinder coordinate system can be created by adding a third direction perpendicular to the plane as shown in the right figure.

Expressing $\vec{E}^a$ in terms of $\vec{e}_a$ and using Eq. (1.185), we find

$$\nabla \times \vec{v} = \sum_{abc} \vec{e}_a \varepsilon_{abc} \frac{1}{h_b h_c} \partial_b \left(h_c \tilde{v}_c \right) = \frac{1}{\mathcal{J}} \sum_{abc} \vec{e}_a \varepsilon_{abc} h_a \partial_b (h_c \tilde{v}_c). \tag{1.193}$$

In matrix determinant notation, this can schematically be written as

$$\nabla \times \vec{v} = \frac{1}{\mathcal{J}} \begin{vmatrix} h_1\vec{e}_1 & \partial_1 & h_1\tilde{v}_1 \\ h_2\vec{e}_2 & \partial_2 & h_2\tilde{v}_2 \\ h_3\vec{e}_3 & \partial_3 & h_3\tilde{v}_3 \end{vmatrix}, \tag{1.194}$$

where the partial derivatives are taken to act on the quantities $h_a\tilde{v}_a$ (no sum) in the third column.

1.6.3 Polar and cylinder coordinates

With the general theory in place, we will now turn our attention to some specific coordinate systems that are highly relevant in several physical applications. The first such system is the set of *polar coordinates* in two dimensions, which are defined by

$$x^1 = \rho \cos(\phi), \tag{1.195a}$$

$$x^2 = \rho \sin(\phi), \tag{1.195b}$$

see Fig. 1.19. This set of coordinates is uniquely defines any point in a Cartesian coordinate system based on the coordinates ρ and ϕ with $0 \leq \rho < \infty$ and ϕ being 2π periodic. We may also extend this to a three dimensional space by mapping the third Cartesian coordinate x^3 to a third coordinate z according to

$$x^3 = z, \tag{1.195c}$$

which along with the earlier definitions of ρ and ϕ defines *cylinder coordinates*. Below we will assume that we are working in three dimensions with cylinder coordinates, remembering that the polar coordinates may be recovered by omitting the z direction, with the obvious exception of the curl, where three dimensions are necessary in order for it to be defined. Using the definition of the tangent vector basis in Eq. (1.148), we find that

$$\vec{E}_\rho = \vec{e}_1 \cos(\phi) + \vec{e}_2 \sin(\phi), \tag{1.196a}$$

$$\vec{E}_\phi = -\vec{e}_1 \rho \sin(\phi) + \vec{e}_2 \rho \cos(\phi), \tag{1.196b}$$

$$\vec{E}_z = \vec{e}_3. \tag{1.196c}$$

It is straightforward to verify that these vectors are orthogonal and therefore we can apply the full framework we have developed for orthogonal coordinate systems. In particular, taking the modulus of these vectors, we find the scale factors for cylinder coordinates

$$h_\rho = 1, \quad h_\phi = \rho, \quad h_z = 1 \tag{1.197}$$

and the orthonormal basis is given by

$$\vec{e}_\rho = \vec{E}_\rho = \vec{e}_1 \cos(\phi) + \vec{e}_2 \sin(\phi), \tag{1.198a}$$

$$\vec{e}_\phi = \frac{1}{\rho}\vec{E}_\phi = -\vec{e}_1 \sin(\phi) + \vec{e}_2 \cos(\phi), \tag{1.198b}$$

$$\vec{e}_z = \vec{E}_z = \vec{e}_3. \tag{1.198c}$$

It should be noted that the position vector $\vec{x}$ can be expressed in cylinder coordinates as

$$\vec{x} = \rho \vec{e}_\rho + z \vec{e}_z. \tag{1.199}$$

Applying the theory developed in Section 1.6.2, we can immediately express the gradient, divergence, curl, and Laplace operator in cylinder coordinates

$$\nabla f = \vec{e}_\rho \partial_\rho f + \vec{e}_\phi \frac{1}{\rho} \partial_\phi f + \vec{e}_z \partial_z f, \tag{1.200a}$$

$$\nabla \cdot \vec{v} = \frac{1}{\rho}(\partial_\rho \rho \tilde{v}_\rho + \partial_\phi \tilde{v}_\phi) + \partial_z \tilde{v}_z, \tag{1.200b}$$

$$\nabla \times \vec{v} = \vec{e}_\rho \left(\frac{1}{\rho} \partial_\phi \tilde{v}_z - \partial_z \tilde{v}_\phi \right) + \vec{e}_\phi \left(\partial_z \tilde{v}_\rho - \partial_\rho \tilde{v}_z \right) + \frac{1}{\rho} \vec{e}_z \left(\partial_\rho \rho \tilde{v}_\phi - \partial_\phi \tilde{v}_\rho \right), \tag{1.200c}$$

$$\nabla^2 f = \frac{1}{\rho} \partial_\rho (\rho \partial_\rho f) + \frac{1}{\rho^2} \partial_\phi^2 f + \partial_z^2 f. \tag{1.200d}$$

Example 1.32 An important vector field which is most easily described in cylinder coordinates is

$$\vec{v} = \frac{1}{\rho} \vec{e}_\phi. \tag{1.201}$$

It can be used to describe the magnetic field strength around an infinite conductor carrying a current in magnetostatics, but for our current purposes it is sufficient to notice that its divergence and curl both vanish as

$$\nabla \cdot \vec{v} = \frac{1}{\rho} \partial_\phi \frac{1}{\rho} = 0, \tag{1.202}$$

$$\nabla \times \vec{v} = -\vec{e}_\rho \partial_z \frac{1}{\rho} + \vec{e}_z \frac{1}{\rho} \partial_\rho \left(\frac{\rho}{\rho} \right) = 0. \tag{1.203}$$

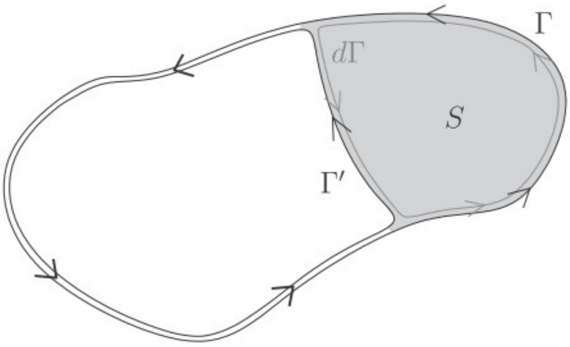

Figure 1.20 The change of contour from Γ to Γ′ does not change the value of the circulation integral of $\vec{v}$ as long as $\nabla \times \vec{v} = 0$ on the area S.

These relations are true for all $\rho > 0$, but may not be generally true at the line $\rho = 0$, where the cylinder coordinate system is singular and this vector field will require special treatment. Taking the circulation integral of $\vec{v}$ around a closed curve Γ, we may rewrite it as a circulation integral around Γ′ according to Fig. 1.20 as long as $\nabla \times \vec{v} = 0$ on an area with the difference contour as a border, in this case as long as the z-axis does not pass through the loop $d\Gamma$. We may therefore change the shape of the contour as long as we do not change the number of times it winds around the z-axis. For a curve that winds around the z-axis once, it may be deformed into a circle of radius R in the plane $z = 0$. We parametrise this circle by $\rho = R$, $\phi = t$, and $z = 0$, with $0 < t < 2\pi$ and the circulation integral becomes

$$\oint_\Gamma \vec{v} \cdot d\vec{x} = \int_0^{2\pi} \frac{R}{R} dt = 2\pi. \tag{1.204}$$

In the above example, we might have been tempted to apply the curl theorem and ended up with

$$\oint_\Gamma \vec{v} \cdot d\vec{x} = \int_S (\nabla \times \vec{v}) \cdot d\vec{S} = 0, \tag{1.205}$$

where S is the disc with the circle Γ as border. Obviously, $0 \neq 2\pi$ and it would seem that something is amiss. However, this is precisely an effect of the vector field $\vec{v}$ being singular at $\rho = 0$, implying that $\nabla \times \vec{v}$ is in some way related to a δ function along $\rho = 0$, since everywhere else the relation $\nabla \times \vec{v} = 0$ holds true.

1.6.4 Spherical coordinates

The second coordinate system we will have a closer look at is that of *spherical coordinates*. It is most useful whenever a problem exhibits complete rotational symmetry in three dimensions, as is the case for a point charge or a spherically symmetric mass distribution, as we will see in the examples of this section. While the cylinder coordinates in three dimensions introduced polar coordinates in a plane and left the third coordinate untouched, spherical coordinates introduces two angles θ and φ, which uniquely define the direction to a point from the origin, and a radius r, which uniquely defines the distance. The spherical

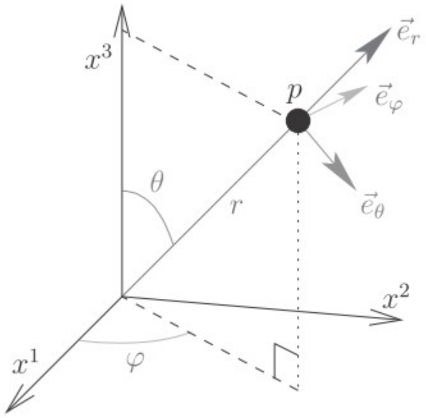

Figure 1.21 The definition of the spherical coordinates in relation to a Cartesian coordinate system. Two angles and a distance uniquely identifies the point p. The unit vectors $\vec{e}_r$, $\vec{e}_\theta$, and $\vec{e}_\varphi$ represent the directions in which p would move when the corresponding coordinates change.

coordinates are defined by

$$x^1 = r\sin(\theta)\cos(\varphi), \tag{1.206a}$$

$$x^2 = r\sin(\theta)\sin(\varphi), \tag{1.206b}$$

$$x^3 = r\cos(\theta), \tag{1.206c}$$

and they are graphically represented in Fig. 1.21. The angle θ is the angle the position vector $\vec{x}$ makes with the x^3-axis, while φ measures the angle between the projection of $\vec{x}$ onto the x^1-x^2-plane and the x^1-axis, much like the polar coordinate ϕ.

From the definition of spherical coordinates, we can find the tangent vector basis in terms of the original Cartesian basis vectors

$$\vec{E}_r = \vec{e}_1\sin(\theta)\cos(\varphi) + \vec{e}_2\sin(\theta)\sin(\varphi) + \vec{e}_3\cos(\theta), \tag{1.207a}$$

$$\vec{E}_\theta = r[\vec{e}_1\cos(\theta)\cos(\varphi) + \vec{e}_2\cos(\theta)\sin(\varphi) - \vec{e}_3\sin(\theta)], \tag{1.207b}$$

$$\vec{E}_\varphi = r\sin(\theta)[-\vec{e}_1\sin(\varphi) + \vec{e}_2\cos(\varphi)]. \tag{1.207c}$$

As in the case of cylinder coordinates, it is straightforward to show that this set of basis vectors is orthogonal and the scale factors

$$h_r = 1, \quad h_\theta = r, \quad \text{and} \quad h_\varphi = r\sin(\theta) \tag{1.208}$$

are found by computing their moduli. Dividing the tangent vector basis by the scale factors provides us with the orthonormal basis

$$\vec{e}_r = \vec{E}_r = \vec{e}_1\sin(\theta)\cos(\varphi) + \vec{e}_2\sin(\theta)\sin(\varphi) + \vec{e}_3\cos(\theta), \tag{1.209a}$$

$$\vec{e}_\theta = \frac{1}{r}\vec{E}_\theta = \vec{e}_1\cos(\theta)\cos(\varphi) + \vec{e}_2\cos(\theta)\sin(\varphi) - \vec{e}_3\sin(\theta), \tag{1.209b}$$

$$\vec{e}_\varphi = \frac{1}{r\sin(\theta)}\vec{E}_\varphi = -\vec{e}_1\sin(\varphi) + \vec{e}_2\cos(\varphi). \tag{1.209c}$$

This set of basis vectors is also illustrated in Fig. 1.21. Furthermore, we obtain the relation

$$\vec{x} = r\vec{e}_r \tag{1.210}$$

for the position vector at each point in space. This is a natural consequence of the selection of the angular parameters as defining the direction to a given point and the coordinate r giving the distance, just as $\vec{x} = \rho\vec{e}_\rho$ in polar coordinates for two dimensions. In fact, polar and spherical coordinates are the two and three dimensional versions of a more general set of spherical coordinates in N dimensions.

The gradient, divergence, curl, and Laplace operators are expressed in spherical coordinates according to

$$\nabla f = \vec{e}_r\partial_r f + \frac{1}{r}\vec{e}_\theta\partial_\theta f + \frac{1}{r\sin(\theta)}\vec{e}_\varphi\partial_\varphi f, \tag{1.211a}$$

$$\nabla\cdot\vec{v} = \frac{1}{r^2}\partial_r(r^2\tilde{v}_r) + \frac{1}{r\sin(\theta)}\partial_\theta(\sin(\theta)\tilde{v}_\theta) + \frac{1}{r\sin(\theta)}\partial_\varphi\tilde{v}_\varphi, \tag{1.211b}$$

$$\begin{aligned}\nabla\times\vec{v} = {} & \frac{1}{r\sin(\theta)}\vec{e}_r(\partial_\theta\sin(\theta)\tilde{v}_\varphi - \partial_\varphi\tilde{v}_\theta) + \frac{1}{r}\vec{e}_\theta\left(\frac{1}{\sin(\theta)}\partial_\varphi\tilde{v}_r - \partial_r r\tilde{v}_\varphi\right)\\ & + \frac{1}{r}\vec{e}_\varphi(\partial_r r\tilde{v}_\theta - \partial_\theta\tilde{v}_r),\end{aligned} \tag{1.211c}$$

$$\nabla^2 f = \frac{1}{r^2}\left[\partial_r\left(r^2\partial_r f\right) + \frac{1}{\sin(\theta)}\partial_\theta\left(\sin(\theta)\partial_\theta f\right) + \frac{1}{\sin^2(\theta)}\partial_\varphi^2 f\right]. \tag{1.211d}$$

Example 1.33 An important vector field in physics is the field of a *point source* in three dimensions, which is most easily expressed in spherical coordinates as

$$\vec{v} = \frac{1}{r^2}\vec{e}_r. \tag{1.212}$$

Inserting $\tilde{v}_r = 1/r^2$ into the expressions for the divergence and curl, respectively, we obtain

$$\nabla\cdot\vec{v} = \frac{1}{r^2}\partial_r\frac{r^2}{r^2} = 0, \tag{1.213}$$

$$\nabla\times\vec{v} = \frac{1}{r}\vec{e}_\theta\left(\frac{1}{\sin(\theta)}\partial_\varphi\frac{1}{r^2}\right) - \frac{1}{r}\vec{e}_\varphi\partial_\theta\frac{1}{r^2} = 0. \tag{1.214}$$

As for the field in Example 1.32, this field is both divergence and curl free everywhere where the coordinate system is not singular, in this case at $r = 0$, where the field has a singularity.

If we perform the flux integral over a closed surface S with outward pointing normal vector, the divergence theorem now tells us that the flux Φ through S is given by

$$\Phi = \oint_S \vec{v}\cdot d\vec{S} = \int_V \nabla\cdot\vec{v}\,dV = 0, \tag{1.215}$$

where V is the volume enclosed by S as long as the point $r = 0$ is not part of this volume. If $r = 0$ is enclosed by S, then we can add and subtract the flux integral over a small sphere S_ε of radius ε to the original integral without changing its value, since we are adding and subtracting the same thing, see Fig. 1.22. As a result, the flux through S is given by

$$\Phi = \left(\oint_S \vec{v}\cdot d\vec{S} - \oint_{S_\varepsilon}\vec{v}\cdot d\vec{S}\right) + \oint_{S_\varepsilon}\vec{v}\cdot d\vec{S} = \int_{V_\varepsilon}\nabla\cdot\vec{v}\,dV + \oint_{S_\varepsilon}\vec{v}\cdot d\vec{S}, \tag{1.216}$$

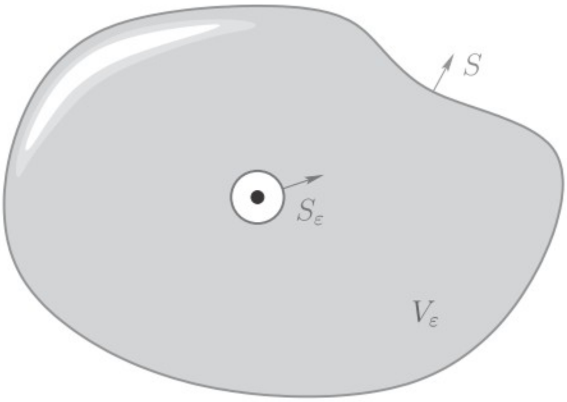

Figure 1.22 By adding and subtracting the flux through the small sphere S_ε, we can rewrite the flux integral over S as the flux integral over S_ε and a volume integral over the volume V_ε in between the S and S_ε.

where V_ε is the volume in between S and S_ε. We have here used the fact that the negative of the integral over S_ε is equal to the integral over S_ε with reversed normal vector, i.e., a normal vector pointing out of the volume V_ε. Since $r = 0$ is not in V_ε, the field $\vec{v}$ is divergence free in this volume and we therefore obtain

$$\Phi = \oint_{S_\varepsilon} \vec{v} \cdot d\vec{S}. \tag{1.217}$$

The surface element $d\vec{S}$ is given by Eq. (1.178) as

$$d\vec{S} = \vec{e}_r h_\theta h_\varphi d\theta\, d\varphi = \vec{e}_r \varepsilon^2 \sin(\theta)\, d\theta\, d\varphi, \tag{1.218}$$

where we have evaluated the scale factors at $r = \varepsilon$ in the last step. We find that

$$\Phi = \oint_{S_\varepsilon} \vec{v} \cdot d\vec{S} = \int_0^\pi \sin(\theta) d\theta \int_0^{2\pi} d\varphi = 4\pi. \tag{1.219}$$

Thus, the flux of the field $\vec{v}$ through a closed surface S is zero if it does not contain the point $r = 0$ and 4π if it does. In order for the divergence theorem to hold, this implies that

$$\nabla \cdot \vec{v} = 4\pi \delta^{(3)}(\vec{x}), \tag{1.220}$$

where $\delta^{(3)}(\vec{x})$ is the three dimensional *delta function* defined by

$$\int_V \delta^{(3)}(\vec{x})\, dV = \begin{cases} 1, & (\vec{x} = 0 \text{ is in } V) \\ 0, & (\text{otherwise}) \end{cases}. \tag{1.221}$$

For the time being, we will just accept that this is the case, but we shall return to this relation when we discuss Green's functions in Chapter 7.

Example 1.34 You may have realised that the vector field in the previous example has a form that reminds us of the gravitational field outside a spherically symmetric mass distribution

$$\vec{g} = -\vec{e}_r \frac{GM}{r^2} \tag{1.222}$$

or the electric field outside a spherically symmetric charge distribution

$$\vec{E} = \vec{e}_r \frac{kQ}{r^2}, \tag{1.223}$$

where G is Newton's gravitational constant, M the total mass of the mass distribution, k is Coulomb's constant, and Q the total charge of the charge distribution. From Gauss's law on differential form, we have

$$\nabla \cdot \vec{E} = \frac{\rho}{\varepsilon_0}. \tag{1.224}$$

Taking the electric field to be given by Eq. (1.223) in all of space, the resulting charge density must be

$$\rho = \varepsilon_0 \nabla \cdot \vec{E} = 4\pi\varepsilon_0 kQ\delta^{(3)}(\vec{x}) = Q\delta^{(3)}(\vec{x}), \tag{1.225}$$

where the last step applies the fact that $k = 1/4\pi\varepsilon_0$. Thus, the charge density is zero everywhere except for in the origin and the total charge, obtained by integrating the charge density over all of space, is Q. The charge distribution therefore corresponds to a single point charge Q in the origin. Similar arguments can be made for the gravitational field.

Even if the charge distribution is not confined to the origin, but still spherically symmetric, the fields above are still solutions to the field equations outside the charge distribution. This can be derived by using the rotational symmetry of the problem. If the charge distribution is rotated around any axis through the center, then the new charge distribution is equivalent to the original one. However, if we take the rotation axis to be the x^3-axis, then the rotation would have rotated the components of the field orthogonal to $\vec{e}_3$, leading to a different solution for the same charge distribution unless those components are zero, see Fig. 1.23. From this we can draw the conclusion that the field will be of the form $\vec{E} = E(r)\vec{e}_r$, since the solution must be unique and the rotational symmetry also implies that the magnitude of the field cannot depend on the angles θ and φ. If we now take a sphere S_r of radius r such that the entire charge distribution is contained within it, then the integral form of Gauss's law states

$$\oint_{S_r} \vec{E} \cdot d\vec{S} = E(r) \oint_{S_r} dS = E(r) 4\pi r^2 = \frac{Q}{\varepsilon_0}. \tag{1.226}$$

Solving for $E(r)$ results in

$$E(r) = \frac{1}{4\pi\varepsilon_0} \frac{Q}{r^2} = \frac{kQ}{r^2} \tag{1.227}$$

as expected. Solving vector equations by using symmetry arguments in this fashion is not always straightforward and will generally involve special cases or complicated integrals. This will be somewhat simplified by the use of potential theory, which is introduced in the next section.

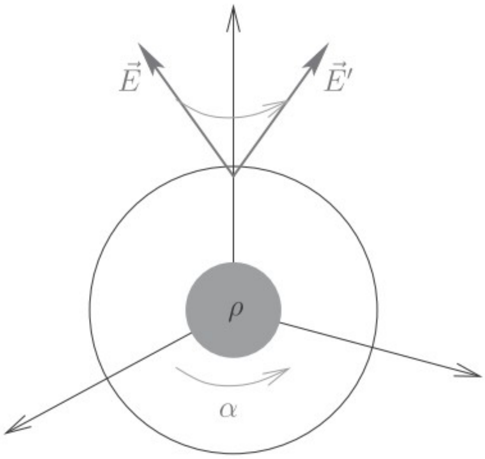

Figure 1.23 Under a rotation by an angle α, if there are components of the electric field $\vec{E}$ orthogonal to the axis of rotation, they will be changed by the rotation to $\vec{E}'$, even if the charge distribution remains the same. This would lead to a different solution unless the orthogonal components are identically zero.

1.7 POTENTIALS

It is sometimes very useful to consider vector fields that can be constructed as derivatives of other fields, both in order to solve differential equations involving fields and in order to make statements of physical importance without having to perform longer calculations. In this section, we will discuss such vector fields, the underlying theory, and some of the physical applications. There are essentially two ways of writing a vector field $\vec{v}$ as a derivative of a different field, either

$$\vec{v} = -\nabla\phi, \tag{1.228a}$$

where ϕ is a scalar field, or

$$\vec{v} = \nabla \times \vec{A}, \tag{1.228b}$$

where $\vec{A}$ is a vector field. The fields ϕ and $\vec{A}$ are known as the *scalar potential* and *vector potential*, respectively. The minus sign in the equation for the scalar potential is purely conventional, but it will appear in this fashion in many physics applications, which is why we adhere to this convention.

1.7.1 Scalar potentials

Vector fields with scalar potentials are typically encountered at very early stages of physics education although it will generally not be stated in the language of vector analysis.

Example 1.35 The gravitational potential Φ is the scalar potential of the gravitational field $\vec{g}$

$$\vec{g} = -\nabla\Phi. \tag{1.229}$$

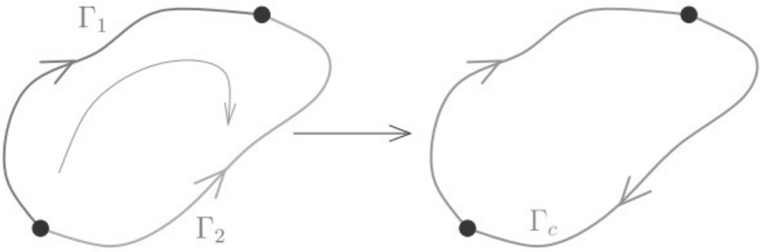

Figure 1.24 The two curves Γ_1 and Γ_2 have the same endpoints. The closed curve Γ_c can be constructed by first going along Γ_1 in the original direction and then along Γ_2 in the opposite direction. This results in a relation between the line integrals along Γ_1 and Γ_2 to the circulation integral along Γ_c.

From Eq. (1.222), we find that, outside a spherical mass distribution,

$$\vec{e}_r \cdot \vec{g} = -\frac{GM}{r^2} = -\partial_r \Phi, \tag{1.230}$$

leading to

$$\Phi = -\frac{GM}{r} \tag{1.231}$$

up to an integration constant that is chosen to be zero by convention, such that the gravitational potential approaches zero as $r \to \infty$.

Coming back to the results of Section 1.4.3, we immediately find that any vector field $\vec{v}$ with a scalar potential ϕ is curl free, as follows from

$$\nabla \times \vec{v} = -\nabla \times \nabla\phi = 0, \tag{1.232}$$

where we have used that $\nabla \times \nabla$ is zero for all scalar fields. It turns out that being curl free is also a sufficient condition for having a scalar potential. Before showing this by explicitly constructing a scalar potential for a curl free vector field, let us examine the properties and physical applications of these fields.

A *conservative vector field* is a vector field $\vec{v}$ for which the line integral

$$L = \int_\Gamma \vec{v} \cdot d\vec{x} \tag{1.233}$$

only depends on the endpoints of the curve Γ. In particular, if we take two different curves Γ_1 and Γ_2 with the same endpoints (see Fig. 1.24), then we can construct a closed curve Γ_c by first going along Γ_1 in the original direction and then along Γ_2 in the opposite direction. This gives us the relation

$$\oint_{\Gamma_c} \vec{v} \cdot d\vec{x} = \int_{\Gamma_1} \vec{v} \cdot d\vec{x} - \int_{\Gamma_2} \vec{v} \cdot d\vec{x} \equiv L_1 - L_2, \tag{1.234}$$

for any vector field $\vec{v}$. Using the curl theorem, we obtain the relation

$$L_1 - L_2 = \int_S (\nabla \times \vec{v}) \cdot d\vec{S}, \tag{1.235}$$

where S is a surface with Γ_c as boundary. If $\nabla \times \vec{v} = 0$, it therefore holds that $L_1 = L_2$, meaning that $\vec{v}$ is conservative if it is curl free. As we have seen that all vector fields with a scalar potential are curl free, they are therefore also conservative. This is also reflected in Eq. (1.106), where we saw that the line integral of a gradient was only dependent on the endpoints of the curve.

Example 1.36 A homogeneous electric field $\vec{E}$ is independent of the position. Naturally this implies that

$$\nabla \times \vec{E} = 0 \tag{1.236}$$

and the field is therefore conservative. The line integral between the two points $\vec{x}_A$ and $\vec{x}_B$ is therefore independent of the path taken and we can take a straight path parametrised by $\vec{x}(t) = \vec{x}_A + t(\vec{x}_B - \vec{x}_A)$, where $t = 0$ represents the start of the curve and $t = 1$ represents the end. Along this curve, $d\vec{x}/dt = \vec{x}_B - \vec{x}_A$, leading to

$$\int_{\vec{x}_A}^{\vec{x}_B} \vec{E} \cdot d\vec{x} = \int_0^1 \vec{E} \cdot (\vec{x}_B - \vec{x}_A)\, dt = \vec{E} \cdot (\vec{x}_B - \vec{x}_A). \tag{1.237}$$

Since $\vec{E}$ is conservative, this will be the result for the line integral of any curve starting in $\vec{x}_A$ and ending in $\vec{x}_B$. We also note that if we define $V = -\vec{E} \cdot (\vec{x} - \vec{x}_A)$, then

$$-\nabla V = \nabla E^i(x^i - x^i_A) = \vec{e}_j E^i \partial_j (x^i - x^i_A) = \vec{e}_j E^i \delta_{ij} = \vec{E}. \tag{1.238}$$

Therefore, the constant field $\vec{E}$ has the potential V.

The above example provides a hint to how we can construct a potential for any conservative field. Assuming a conservative vector field $\vec{v}$, we can construct the scalar field

$$\varphi(\vec{x}) = -\int_{\vec{x}_0}^{\vec{x}} \vec{v}(\vec{\xi}) \cdot d\vec{\xi}, \tag{1.239}$$

which is a line integral along any curve connecting $\vec{x}_0$ and $\vec{x}$. This is well defined due to $\vec{v}$ being conservative, meaning that the integral does not depend on the particular curve chosen. In order to compute the gradient of this scalar field, we need to express the partial derivatives $\partial_i \varphi$ in terms of $\vec{v}$. By definition, the partial derivatives are given by

$$\partial_i \varphi = \lim_{\varepsilon \to 0} \left(\frac{\varphi(\vec{x} + \varepsilon \vec{e}_i) - \varphi(\vec{x})}{\varepsilon} \right). \tag{1.240}$$

We now note that

$$\varphi(\vec{x} + \varepsilon \vec{e}_i) - \varphi(\vec{x}) = -\int_{\vec{x}_0}^{\vec{x}+\varepsilon\vec{e}_i} \vec{v} \cdot d\vec{x} + \int_{\vec{x}_0}^{\vec{x}} \vec{v} \cdot d\vec{x} = -\int_{\vec{x}}^{\vec{x}+\varepsilon\vec{e}_i} \vec{v} \cdot d\vec{x}, \tag{1.241}$$

where we have used that we can select a curve first passing through $\vec{x}$ and then continuing to $\vec{x} + \varepsilon \vec{e}_i$ when computing $\varphi(\vec{x} + \varepsilon \vec{e}_i)$. Again, due to $\vec{v}$ being conservative, we can select the curve $\vec{x}(t) = \vec{x} + t\varepsilon \vec{e}_i$, with t going from 0 to 1, when evaluating the remaining integral. We find that

$$-\int_{\vec{x}}^{\vec{x}+\varepsilon\vec{e}_i} \vec{v} \cdot d\vec{x} = -\varepsilon \int_0^1 \vec{v} \cdot \vec{e}_i dt = -\varepsilon \int_0^1 v^i dt = -\varepsilon v^i(\vec{x} + \varepsilon t_* \vec{e}_i), \tag{1.242}$$

where $0 \leq t_* \leq 1$, and we have used the mean value theorem of integration in the last step. Taking the limit, we obtain

$$\partial_i \varphi = -\lim_{\varepsilon \to 0} \left(v^i(\vec{x} + \varepsilon t_* \vec{e}_i) \right) = -v^i(\vec{x}). \tag{1.243}$$

It follows that

$$-\nabla\varphi = -\vec{e}_i \partial_i \varphi = \vec{e}_i v^i = \vec{v} \tag{1.244}$$

and thus φ is a scalar potential of $\vec{v}$, meaning that a scalar potential of $\vec{v}$ exists as long as $\vec{v}$ is conservative.

Collecting our results, we find that the following three statements are equivalent:

1. The vector field $\vec{v}$ is curl free, $\nabla \times \vec{v} = 0$.
2. The vector field $\vec{v}$ is conservative.
3. The vector field $\vec{v}$ has a scalar potential φ such that $\vec{v} = -\nabla\varphi$.

Example 1.37 A *conservative force field* is a force field $\vec{F}$ for which the work W done by moving an object from $\vec{x}_A$ to $\vec{x}_B$, given by the line integral

$$W = \int_{\vec{x}_A}^{\vec{x}_B} \vec{F} \cdot d\vec{x}, \tag{1.245}$$

does not depend on the path taken. By definition, this force field is a conservative vector field, meaning that it has a scalar potential V and is curl free. The scalar potential is the *potential energy* of the force field and we have

$$\vec{F} = -\nabla V. \tag{1.246}$$

Conservative force fields play a central role in classical mechanics.

In the construction of the scalar potential of a conservative vector field, we introduced the arbitrary point $\vec{x}_0$. As any choice of $\vec{x}_0$ results in a valid scalar potential, the resulting scalar potential is not unique. For the conservative vector field $\vec{v}$, we can define two scalar potentials

$$\varphi_1 = -\int_{\vec{x}_1}^{\vec{x}} \vec{v} \cdot d\vec{x} \quad \text{and} \quad \varphi_2 = -\int_{\vec{x}_2}^{\vec{x}} \vec{v} \cdot d\vec{x}. \tag{1.247}$$

Both of these scalar fields have the property $-\nabla\varphi_i = \vec{v}$ and are therefore scalar potentials of $\vec{v}$. However, they differ by

$$\varphi_2(\vec{x}) - \varphi_1(\vec{x}) = \int_{\vec{x}_1}^{\vec{x}_2} \vec{v} \cdot d\vec{x} \tag{1.248}$$

which is a constant independent of $\vec{x}$. In general, if $\vec{v}$ has the scalar potential φ, then $\phi = \varphi + k$, where k is a constant, will fulfil

$$-\nabla\phi = -\nabla(\varphi + k) = -\nabla\varphi - \nabla k = \vec{v} \tag{1.249}$$

and also be a scalar potential of $\vec{v}$. Any scalar potential is therefore only defined up to a constant and scalar potentials differing by a constant will give the same vector field.

It is also worth noting that the construction of the scalar potential as an integral of the vector field may not always be the easiest way of computing the potential. In general, finding the potential amounts to solving a set of differential equations

$$\partial_i \varphi = -v^i, \tag{1.250}$$

which can often be done in a straightforward fashion, as in Example 1.35.

Example 1.38 In Example 1.36 we found the potential of a constant electric field $\vec{E}$ by integration along a selected curve. We could also have integrated the equation $\partial_i V = -E^i$ directly. Restricting ourselves to two dimensions and starting with $i = 1$, we find that

$$\partial_1 V = -E^1 \quad \Longrightarrow \quad V = -E^1 x^1 + f(x^2), \tag{1.251}$$

where f is only a function of x^2 and thus $\partial_1 f(x^2) = 0$. Inserting this into $\partial_2 V = -E^2$ gives us

$$\partial_2(-E^1 x^1 + f(x^2)) = f'(x^2) = -E^2 \quad \Longrightarrow \quad f(x^2) = -E^2 x^2 + V_0, \tag{1.252}$$

where V_0 is a constant. It follows that

$$V = -E^1 x^1 - E^2 x^2 + V_0 = -\vec{E} \cdot \vec{x} + V_0, \tag{1.253}$$

where V_0 is an arbitrary integration constant. This example has a straightforward generalisation to any number of dimensions.

1.7.2 Vector potentials

Unlike scalar potentials, vector potentials are generally not encountered in physics education until the study of magnetostatics.

Example 1.39 A static magnetic field $\vec{B}$ has a vector potential $\vec{A}$ such that

$$\vec{B} = \nabla \times \vec{A}. \tag{1.254}$$

As we shall see, this follows from *Gauss's law of magnetism*, which on differential form is given by

$$\nabla \cdot \vec{B} = 0. \tag{1.255}$$

In analogy to curl free vector fields having scalar potentials, we shall see that divergence free vector fields have vector potentials.

For any vector field $\vec{v}$ with a vector potential $\vec{A}$, we find that

$$\nabla \cdot \vec{v} = \nabla \cdot (\nabla \times A) = 0. \tag{1.256}$$

All fields with vector potentials are therefore divergence free. This is analogous to what we found in the scalar potential case, where all fields with scalar potentials were found to be curl free. Furthermore, just as vector fields with scalar potentials having line integrals depending only on the curve endpoints, the flux of a field with a vector potential through a surface S only depends on the boundary curve, which follows directly from the curl theorem (see Eq. (1.136)).

In order to show that $\nabla \cdot \vec{v} = 0$ implies the existence of a vector potential, we will proceed with an explicit construction of the potential $\vec{A}$. Unfortunately, the construction cannot proceed in a way similar to the construction of a scalar potential based on a curl free field, mainly due to the loss of a path independent line integral. Instead, we start by considering the scalar field

$$\phi(\vec{x}) = \frac{1}{r}, \tag{1.257}$$

where r is the spherical coordinate. By taking the gradient of this field, we obtain

$$-\nabla\phi = -\vec{e}_r \partial_r \frac{1}{r} = \frac{1}{r^2}\vec{e}_r, \tag{1.258}$$

which is exactly the point source vector field considered in Example 1.33. Using the results from this example, we find that

$$-\nabla^2\phi = 4\pi\delta^{(3)}(\vec{x}). \tag{1.259}$$

By making the translation $\vec{x} \to \vec{x} - \vec{x}_0$, we obtain the relation

$$-\nabla^2 \frac{1}{|\vec{x} - \vec{x}_0|} = 4\pi\delta^{(3)}(\vec{x} - \vec{x}_0). \tag{1.260}$$

Furthermore, for any vector field $\vec{v}$, we can use the defining property of $\delta^{(3)}$ in order to obtain

$$\vec{v}(\vec{x}) = \int \delta^{(3)}(\vec{x} - \vec{x}_0)\vec{v}(\vec{x}_0)dV_0 = -\frac{1}{4\pi}\int \vec{v}(\vec{x}_0)\left(\nabla^2 \frac{1}{|\vec{x} - \vec{x}_0|}\right)dV_0, \tag{1.261}$$

where the integral is taken over all of space and the fields will be assumed to go to zero sufficiently fast as $r \to \infty$. By moving the Laplace operator out of the integral, we obtain

$$\vec{v}(\vec{x}) = -\frac{1}{4\pi}\nabla^2 \int \frac{\vec{v}(\vec{x}_0)}{|\vec{x} - \vec{x}_0|}dV_0 = -\frac{1}{4\pi}\nabla^2\vec{w}, \tag{1.262}$$

where

$$\vec{w} = \int \frac{\vec{v}(\vec{x}_0)}{|\vec{x} - \vec{x}_0|}dV_0. \tag{1.263}$$

From Eq. (1.97), it follows that

$$\vec{v}(\vec{x}) = \frac{1}{4\pi}\nabla \times (\nabla \times \vec{w}) - \frac{1}{4\pi}\nabla(\nabla \cdot \vec{w}). \tag{1.264}$$

We can now rewrite $\nabla \cdot \vec{w}$ according to

$$\nabla \cdot \vec{w} = \int \nabla \cdot \frac{\vec{v}(\vec{x}_0)}{|\vec{x} - \vec{x}_0|}dV_0 = \int \vec{v}(\vec{x}_0) \cdot \nabla \frac{1}{|\vec{x} - \vec{x}_0|}dV_0. \tag{1.265}$$

We now note that $\nabla f(\vec{x} - \vec{x}_0) = -\nabla_0 f(\vec{x} - \vec{x}_0)$, where ∇_0 is taken to be the gradient with respect to $\vec{x}_0$, which leads to

$$\nabla \cdot \vec{w} = -\int \vec{v}(\vec{x}_0) \cdot \nabla_0 \frac{1}{|\vec{x} - \vec{x}_0|}dV_0 = -\int \nabla_0 \cdot \left(\frac{\vec{v}(\vec{x}_0)}{|\vec{x} - \vec{x}_0|}\right)dV_0 = 0, \tag{1.266}$$

where we have used the fact that $\vec{v}$ was taken to be divergence free and again assumed that the $\vec{v}$ goes to zero sufficiently fast as $r \to \infty$ for the surface integral in the divergence theorem to vanish as we integrate over the entire space. It follows that

$$\vec{v} = \nabla \times \vec{A}, \tag{1.267}$$

where

$$\vec{A} = \frac{1}{4\pi}\nabla \times \vec{w} = \frac{1}{4\pi}\nabla \times \int \frac{\vec{v}(\vec{x}_0)}{|\vec{x} - \vec{x}_0|} dV_0, \tag{1.268}$$

and thus $\vec{v}$ has $\vec{A}$ as a vector potential.

While the above tells us of the existence of a vector potential, it is not necessarily the most convenient way of computing it. Again in analogy to the scalar potential case, it is often more advantageous to solve the set of differential equations obtained when demanding that $\vec{v} = \nabla \times \vec{A}$.

Example 1.40 Following Example 1.18, we know that the constant vector field $\vec{\omega} = 2\omega\vec{e}_3$ has the vector potential $\vec{A} = \omega(-x^2\vec{e}_1 + x^1\vec{e}_2)$ (note that $\vec{A}$ was called $\vec{v}$ in Example 1.18) and we now set out to reproduce this result. Starting from $\nabla \cdot \vec{\omega} = 0$, which follows directly from $\vec{\omega}$ being constant, we know that the vector potential exists. We therefore write

$$\begin{aligned}\nabla \times \vec{A} &= \left(\partial_2 A^3 - \partial_3 A^2\right)\vec{e}_1 + \left(\partial_3 A^1 - \partial_1 A^3\right)\vec{e}_2 + \left(\partial_1 A^2 - \partial_2 A^1\right)\vec{e}_3 \\ &= 2\omega\vec{e}_3.\end{aligned} \tag{1.269}$$

The vector field components A^i are generally too many to solve for in a unique way and we therefore try to find a solution such that $A^3 = 0$. It follows that

$$\partial_3 A^1 = 0 \quad \Longrightarrow \quad A^1 = f^1(x^1, x^2), \tag{1.270a}$$

$$\partial_3 A^2 = 0 \quad \Longrightarrow \quad A^2 = f^2(x^1, x^2), \tag{1.270b}$$

$$\partial_1 f^2 - \partial_2 f^1 = 2\omega. \tag{1.270c}$$

The last of these equations also has several possible solutions, one of which is

$$f^1 = -2\omega x^2, \quad f^2 = 0, \tag{1.271a}$$

which is different from the vector potential we already saw. However, another solution is clearly

$$f^1 = -\omega x^2, \quad f^2 = \omega x^1, \tag{1.271b}$$

which leads to precisely the potential already obtained in Example 1.18. Both of these solutions are perfectly valid vector potentials for $\vec{\omega}$ and it follows that the vector potential is not unique. This is a general result that will hold for any divergence free vector field.

The procedure in the above example can be generally applied to any divergence free vector field $\vec{v}$. We can start by making the ansatz that the vector potential $\vec{A}$ is of the form

$$\vec{A} = A^1\vec{e}_1 + A^2\vec{e}_2, \tag{1.272}$$

which is assuming that the $\vec{e}_3$ component is identically zero. The first component of the relation $\vec{v} = \nabla \times \vec{A}$ is now given by

$$-\partial_3 A^2 = v^1 \quad \Longrightarrow \quad A^2 = -\int_{x_0^3}^{x^3} v^1(x^1, x^2, z)\, dz + f^2(x^1, x^2), \tag{1.273}$$

for each set of x^1 and x^2. Similarly, we obtain

$$A^1 = \int_{x_0^3}^{x^3} v^2(x^1, x^2, z)\, dz + f^1(x^1, x^2). \tag{1.274}$$

The third component will now be of the form

$$v^3 = \partial_1 A^2 - \partial_2 A^1 = -\int_{x_0^3}^{x^3} [\partial_1 v^1(\vec{x}_z) + \partial_2 v^2(\vec{x}_z)] dz + \partial_1 f^2 - \partial_2 f^1, \tag{1.275}$$

where $\vec{x}_z = x^1\vec{e}_1 + x^2\vec{e}_2 + z\vec{e}_3$ and f^1 and f^2 are functions of x^1 and x^2. However, since $\vec{v}$ is divergence free, we can replace $\partial_1 v^1 + \partial_2 v^2$ with $-\partial_3 v^3$. We now obtain

$$v^3(\vec{x}) = \int_{x_0^3}^{x^3} \partial_3 v^3(\vec{x}_z) dz + \partial_1 f^2 - \partial_2 f^1 = v^3(\vec{x}) - v^3(x^1, x^2, x_0^3) + \partial_1 f^2 - \partial_2 f^1 \tag{1.276}$$

or, cancelling $v^3(\vec{x})$ from both sides,

$$v^3(x^1, x^2, x_0^3) = \partial_1 f^2 - \partial_2 f^1. \tag{1.277}$$

As in the example, this differential equation can be solved by

$$f^1 = -\int_{x_0^2}^{x^2} v^3(x^1, y, x_0^3)\, dy, \quad f^2 = 0 \tag{1.278}$$

and we obtain a vector potential

$$\vec{A} = \vec{e}_1 \left[\int_{x_0^3}^{x^3} v^2(x^1, x^2, z)\, dz - \int_{x_0^2}^{x^2} v^3(x^1, y, x_0^3)\, dy \right] + \vec{e}_2 \int_{x_0^3}^{x^3} v^1(x^1, x^2, z)\, dz. \tag{1.279}$$

Apart from giving us an alternative way of showing that the vector potential exists, this explicit construction gives us a straightforward recipe for computing a vector potential for any divergence free vector field without applying our original construction.

Note that this explicit construction of a vector potential only depends on the local properties of the vector field $\vec{v}$ and does not require it to go to zero sufficiently fast as $r \to \infty$.

Example 1.41 Another example of a divergence free vector field can be found in fluid dynamics. For any fluid, the local velocity field $\vec{v}$ tells us how fast the fluid is moving at each point in space. For an *incompressible fluid*, i.e., a fluid for which the density is constant regardless of external pressure or forces, we must have $\nabla \cdot \vec{v} = 0$, implying the existence of a vector potential for the fluid flow. While this may not be exactly true for any fluid, it is often a good approximation and we will return to the rationale behind this condition for an incompressible fluid in Section 3.9.3.

As we have seen, the vector potential, like the scalar potential for curl free fields, is not unique. In the scalar potential case, we found that any scalar fields that differed by a constant had the same gradient. In the case of a field with a vector potential, we can clearly also add a constant vector field to the potential without affecting the field itself. However, this is not the most general way of obtaining equivalent vector potentials. Let us assume that the vector field $\vec{v}$ is divergence free and that it has two different vector potentials $\vec{A}_1$ and $\vec{A}_2$. By taking the curl of the difference of these vector potentials, we can deduce that

$$\nabla \times (\vec{A}_1 - \vec{A}_2) = \nabla \times \vec{A}_1 - \nabla \times \vec{A}_2 = \vec{v} - \vec{v} = 0. \tag{1.280}$$

Thus, the difference vector field $\delta\vec{A} = \vec{A}_1 - \vec{A}_2$ is curl free. From the previous section, it directly follows that $\delta\vec{A}$ has a scalar potential φ such that $\delta\vec{A} = -\nabla\varphi$. Therefore, for any two vector potentials of the same field, we must have

$$\vec{A}_2 = \vec{A}_1 + \nabla\varphi, \tag{1.281}$$

for some φ. In addition, if we know one vector potential $\vec{A}$, then also

$$\nabla \times (\vec{A} + \nabla\varphi) = \nabla \times \vec{A} + \nabla \times \nabla\varphi = \vec{v} \tag{1.282}$$

for any scalar field φ. It follows that any difference between two vector potentials of the same divergence free field is a gradient and that adding a gradient to the vector potential of a vector field will give a new vector potential of the same field.

1.7.3 Scalar and vector potentials

So far we have seen that curl free vector fields have scalar potentials and that divergence free vector fields have vector potentials. The natural question arises whether or not anything can be said for fields that are neither curl nor divergence free. We shall find that any vector field, regardless of its curl and divergence, may be written using a scalar and a vector potential. In fact, we will be able to decompose any vector field into two parts, where one is curl free and the other divergence free according to

$$\vec{v} = \vec{v}_s + \vec{v}_v, \quad \text{where} \quad \vec{v}_s = -\nabla\varphi, \quad \vec{v}_v = \nabla \times \vec{A}. \tag{1.283}$$

The existence of such a decomposition is known as *Helmholtz's theorem* (or the *fundamental theorem of vector analysis*) and the decomposition itself is called a *Helmholtz decomposition*.

Helmholtz's theorem can be stated in the following way: If V is a volume in three dimensional space, a vector field $\vec{v}$ may be decomposed in a curl free and a divergence free component. We will show that this is true by explicitly constructing such a decomposition in a similar fashion to when we showed that any divergence free field has a vector potential. Starting from $\phi = 1/\left|\vec{x} - \vec{x}_0\right|$, we follow the same steps as in Eqs. (1.258) to (1.264), but with the integration domain being V rather than all of space. The scalar and vector potentials are given by $\varphi = \nabla \cdot \vec{w}/4\pi$ and $\vec{A} = \nabla \times \vec{w}/4\pi$, respectively. We find that

$$\begin{aligned}
\varphi &= \frac{1}{4\pi}\int_V \vec{v}(\vec{x}_0) \cdot \nabla \frac{1}{|\vec{x} - \vec{x}_0|} dV_0 = -\frac{1}{4\pi}\int_V \vec{v}(\vec{x}_0) \nabla_0 \frac{1}{|\vec{x} - \vec{x}_0|} dV_0 \\
&= -\frac{1}{4\pi}\int_V \nabla_0 \left(\vec{v}(\vec{x}_0) \frac{1}{|\vec{x} - \vec{x}_0|} \right) dV_0 + \frac{1}{4\pi}\int_V \frac{1}{|\vec{x} - \vec{x}_0|} \nabla_0 \cdot \vec{v}(\vec{x}_0) dV_0 \\
&= \frac{1}{4\pi}\int_V \frac{1}{|\vec{x} - \vec{x}_0|} \nabla_0 \cdot \vec{v}(\vec{x}_0) dV_0 - \frac{1}{4\pi}\oint_S \frac{\vec{v}(\vec{x}_0)}{|\vec{x} - \vec{x}_0|} \cdot d\vec{S}_0, \qquad (1.284) \\
\vec{A} &= -\frac{1}{4\pi}\int_V \vec{v}(\vec{x}_0) \times \nabla \frac{1}{|\vec{x} - \vec{x}_0|} dV_0 = \frac{1}{4\pi}\int_V \vec{v}(\vec{x}_0) \times \nabla_0 \frac{1}{|\vec{x} - \vec{x}_0|} dV_0 \\
&= -\frac{1}{4\pi}\int_V \nabla_0 \times \left(\frac{\vec{v}(\vec{x}_0)}{|\vec{x} - \vec{x}_0|} \right) dV_0 + \frac{1}{4\pi}\int_V \frac{1}{|\vec{x} - \vec{x}_0|} \nabla_0 \times \vec{v}(\vec{x}_0) dV_0
\end{aligned}$$

$$= \frac{1}{4\pi}\int_V \frac{1}{|\vec{x}-\vec{x}_0|}\nabla_0 \times \vec{v}(\vec{x}_0)dV_0 + \frac{1}{4\pi}\oint_S \frac{\vec{v}(\vec{x}_0)}{|\vec{x}-\vec{x}_0|} \times d\vec{S}_0. \tag{1.285}$$

Note that in order to compute these potentials, we only need to know the divergence and curl of $\vec{v}$ in V as well as the value of $\vec{v}$ on the surface S, which is the boundary surface of V.

Just as for scalar and vector potentials, the Helmholtz decomposition is generally not unique. Given a decomposition

$$\vec{v} = -\nabla\varphi_1 + \nabla \times \vec{A}_1, \tag{1.286}$$

we can try to find another pair of potentials φ_2 and $\vec{A}_2$ satisfying the same relation. Such a pair of potentials must fulfil

$$\vec{v} - \vec{v} = -\nabla(\varphi_1 - \varphi_2) + \nabla \times (\vec{A}_1 - \vec{A}_2) = 0. \tag{1.287}$$

Taking the divergence of this relation gives

$$-\nabla^2(\varphi_1 - \varphi_2) = 0, \tag{1.288}$$

implying that the difference of the scalar potentials $f = \varphi_1 - \varphi_2$ must fulfil *Laplace's equation*

$$\nabla^2 f = 0. \tag{1.289}$$

For the difference of the vector potentials, we find that

$$\nabla \times (\vec{A}_1 - \vec{A}_2) = \nabla f. \tag{1.290}$$

Since ∇f is divergence free, it has a vector potential and this equation therefore has a solution. We can therefore add any function satisfying Laplace's equation to the scalar potential as long as we modify the vector potential accordingly.

Example 1.42 In the Helmholtz decomposition above, we will end up with a scalar potential

$$\varphi = -\frac{1}{4\pi}\oint_S \frac{\vec{v}(\vec{x}_0)}{|\vec{x}-\vec{x}_0|} \cdot d\vec{S}_0 \tag{1.291}$$

even in the case when $\nabla \cdot \vec{v} = 0$. However, by the arguments of the previous section, we should be able to write the divergence free vector field $\vec{v}$ as $\vec{v} = \nabla \times \vec{A}$ for some vector potential $\vec{A}$. We note that

$$\nabla^2\varphi = -\frac{1}{4\pi}\oint_S \left(\nabla^2 \frac{1}{|\vec{x}-\vec{x}_0|}\right)\vec{v}(\vec{x}_0)\cdot d\vec{S}_0 = \oint_S \delta^{(3)}(\vec{x}-\vec{x}_0)\vec{v}(\vec{x}_0)\cdot d\vec{S}_0 = 0, \tag{1.292}$$

where we have assumed $\vec{x}$ is an interior point of V so that $\delta^{(3)}(\vec{x}-\vec{x}_0)$ is zero on the boundary S. Since φ fulfils $\nabla^2\varphi = 0$, we can choose a new scalar potential $\varphi_2 = 0$ which fulfils $\nabla^2(\varphi_2 - \varphi) = 0$ and find a corresponding vector potential $\vec{A}_2$ such that

$$\nabla \times (\vec{A}_2 - \vec{A}) = -\nabla\varphi, \quad \nabla \times \vec{A}_2 = -\nabla\varphi + \nabla \times \vec{A} = \vec{v}. \tag{1.293}$$

Thus, we can still construct a vector potential of $\vec{v}$ as per the previous section.

Example 1.43 We have seen that a constant electric field $\vec{E} = E^i \vec{e}_i$ has a scalar potential $V = -\vec{E} \cdot \vec{x}$, the existence of which was implied by $\nabla \cdot \vec{E} = 0$. However, since $\vec{E}$ is constant, we also have

$$\nabla \cdot \vec{E} = \partial_i E^i = 0, \tag{1.294}$$

implying that $\vec{E} = \nabla \times \vec{A}$ for some vector potential $\vec{A}$. Such a vector potential can be constructed by letting

$$\vec{A} = \frac{1}{2}\vec{E} \times \vec{x}, \tag{1.295}$$

which gives us

$$\nabla \times \vec{A} = \frac{1}{2}\nabla \times (\vec{E} \times \vec{x}) = \frac{1}{2}[\vec{E}(\nabla \cdot \vec{x}) - (\vec{E} \cdot \nabla)\vec{x}] = \frac{1}{2}(3\vec{E} - \vec{E}) = \vec{E}. \tag{1.296}$$

This is also consistent with $\nabla^2 \varphi = 0$ implying that the scalar potential can be removed by a redefinition of the vector potential.

Often, the potentials are not physically observable quantities, but the vector fields they describe are. In these situations the potentials can be changed freely as long as they give rise to the same fields. This freedom also means that, when solving physical problems, we may impose additional constraints on the potentials in such a way that the problem is as easy to solve as possible. This can be done as long as our imposed conditions are not so stringent that they imply that the problem has no solution.

1.8 PROBLEMS

Problem 1.1. Write down the following vector expressions in component form using the Einstein summation convention:

a) $\vec{v} \cdot (k\vec{w} + \vec{u})$

b) $[(k\vec{v}) \times \vec{w}](\ell + m)$

c) $\vec{v} \times (\vec{w} \times \vec{u})$

d) $(\vec{v} \times \vec{w}) \times \vec{u}$

Here, $\vec{u}$, $\vec{v}$, and $\vec{w}$ are vectors while k, ℓ, and m are scalars. Your answers may contain at most one permutation symbol ε_{ijk}.

Problem 1.2. Verify that the cross product satisfies Eq. (1.7) when written on component form.

Problem 1.3. Check that the components of a vector $\vec{v}$ can be found through the use of Eq. (1.10).

Problem 1.4. Consider the three vectors

$$\vec{v}_1 = 3\vec{e}_1 - \vec{e}_2, \quad \vec{v}_2 = 2\vec{e}_2, \quad \vec{e}_3 = -\vec{e}_1 + \vec{e}_2 + 5\vec{e}_3. \tag{1.297}$$

Compute the following:

a) The magnitude of each of the vectors.

b) The inner product between any pair of the vectors.

c) The cross product between any pair of the vectors.

d) The angles between the vectors.

e) The volume of the parallelepiped spanned by the vectors.

Problem 1.5. Show, using only Eq. (1.6), that the cross product $\vec{v} \times \vec{w}$ is orthogonal to both $\vec{v}$ and $\vec{w}$, i.e., show that

$$\vec{v} \cdot (\vec{v} \times \vec{w}) = \vec{w} \cdot (\vec{v} \times \vec{w}) = 0. \tag{1.298}$$

Problem 1.6. Consider the three vectors

$$\vec{e}_{1'}' = \frac{\vec{e}_1}{\sqrt{2}} - \frac{\vec{e}_3}{\sqrt{2}}, \quad \vec{e}_{2'}' = \frac{\vec{e}_1}{\sqrt{3}} + \frac{\vec{e}_2}{\sqrt{3}} + \frac{\vec{e}_3}{\sqrt{3}}, \quad \vec{e}_{3'}' = \frac{\vec{e}_1}{\sqrt{6}} - \sqrt{\frac{2}{3}}\vec{e}_2 + \frac{\vec{e}_3}{\sqrt{6}}. \tag{1.299}$$

Show that this set of vectors is orthogonal and normalised. Determine whether they form a right- or left-handed set and compute the corresponding transformation coefficients.

Problem 1.7. Show that the Kronecker delta fulfils the relation

$$\delta_{ii} = N, \tag{1.300}$$

where N is the dimensionality of the space.

Problem 1.8. Show that Eq. (1.26) turns into Eq. (1.6) when writing out the sums explicitly by making use of Eq. (1.24).

Problem 1.9. Prove the ε-δ-relation Eq. (1.41) and use it to derive the bac-cab rule Eq. (1.40). *Hint:* In order to prove the ε-δ-relation, you do not need to perform the sum explicitly for all possible indices. It is sufficient to do it for some combinations of free indices and then use the symmetries and anti-symmetries of the expression.

Problem 1.10. Use the ε-δ-relation to express the quantity $\varepsilon_{ijk}\varepsilon_{jk\ell}$ in terms of the Kronecker delta.

Problem 1.11. Consider the vector fields $\vec{v} = x^2\vec{e}_1 - x^1\vec{e}_2$ and $\vec{w} = x^3\vec{e}_1 - x^1\vec{e}_3$ in three dimensions. Compute the following quantities:

a) $\vec{v} \cdot \vec{w}$

b) $\vec{v} \times \vec{w}$

c) $\vec{v} \cdot (\vec{w} \times \vec{x})$

where $\vec{x}$ is the position vector $\vec{x} = x^i\vec{e}_i$.

Problem 1.12. Consider the parallelogram spanned by the vectors $\vec{v}$ and $\vec{w}$, see Fig. 1.25. Show that the diagonals $\vec{v} + \vec{w}$ and $\vec{v} - \vec{w}$ of this parallelogram are orthogonal if and only if $\vec{v}$ and $\vec{w}$ have the same magnitude.

Problem 1.13. A cube of side length ℓ is spanned by the three vectors $\vec{v}_1 = \ell\vec{e}_1$, $\vec{v}_2 = \ell\vec{e}_2$, and $\vec{v}_3 = \ell\vec{e}_3$. Compute the angle between two of the cube diagonals by using the inner product of the displacement vectors between opposite corners.

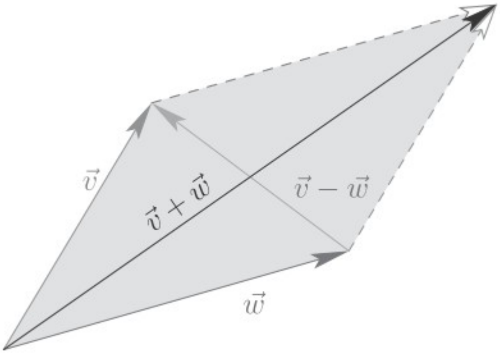

Figure 1.25 A parallelogram spanned by the vectors $\vec{v}$ and $\vec{w}$. The aim of Problem 1.12 is to show that the diagonal vectors $\vec{v}+\vec{w}$ and $\vec{v}-\vec{w}$ are orthogonal only if $\vec{v}$ and $\vec{w}$ have the same magnitude.

Problem 1.14. Find an expression for the squared magnitude of the vector $\vec{v}\times\vec{w}$ in terms of the components of $\vec{v}$ and $\vec{w}$. Also express the magnitudes of $\vec{v}$ and $\vec{w}$ and use your results to find an expression for the sine of the angle between $\vec{v}$ and $\vec{w}$.

Problem 1.15. Consider the two lines given by

$$\vec{x}_1(t) = t\vec{e}_1 + (2-t)\vec{e}_2 \quad \text{and} \quad \vec{x}_2(s) = 2s\vec{e}_1 - \vec{e}_2 - s\vec{e}_3, \tag{1.301}$$

respectively. Define the difference vector between two points on the lines $\vec{d}(t,s) = \vec{x}_1(t) - \vec{x}_2(s)$ and use it to find the shortest distance between the lines. Show that for the points defining the shortest distance, $\vec{d}$ is orthogonal to the tangent vectors of both lines.

Problem 1.16. An object moves in such a fashion that its position as a function of time t may be described by

$$\vec{x}(t) = r_0\cos(\omega t)\vec{e}_1 + r_0\sin(\omega t)\vec{e}_2 + v_0 t\vec{e}_3. \tag{1.302}$$

a) What is the distance from the origin at an arbitrary time t?

b) What are the velocity and acceleration vectors associated to this motion?

Problem 1.17. Many physical situations involving precession, such as the precession of a spinning top or magnetic spin precession, can be described by the vectorial differential equation

$$\frac{d\vec{L}}{dt} = \vec{v}\times\vec{L}, \tag{1.303}$$

where $\vec{L}$ is an angular momentum and $\vec{v}$ is a constant vector. Show that the magnitude of the angular momentum $\vec{L}$ and its inner product with $\vec{v}$ are constants in time.

Problem 1.18. Compute the divergence and curl of the following vector fields:

a) $\vec{x} = x^i\vec{e}_i$

b) $\vec{v}_1 = \vec{a}\times\vec{x}$

c) $\vec{v}_2 = x^2\vec{e}_1 - x^1\vec{e}_2$

where $\vec{a}$ is a constant vector $\vec{a} = a^i \vec{e}_i$.

Problem 1.19. Rewrite the following expressions in terms of the scalar field ϕ and its derivatives only, i.e., leave no derivatives of the position vector $\vec{x}$ or derivatives of products:

a) $\nabla \cdot (\phi \vec{x})$

b) $\nabla \cdot (\vec{x} \times \nabla\phi)$

c) $\nabla \cdot (\phi \nabla \phi)$

d) $\nabla \times (\vec{x} \times \nabla\phi)$

Problem 1.20. Confirm that if the partial derivatives of a scalar field ϕ commute, i.e., $\partial_i \partial_j \phi = \partial_j \partial_i \phi$, and if the partial derivatives of a vector field $\vec{v}$ commute, then Eqs. (1.91) are satisfied.

Problem 1.21. Express $\nabla \times (\nabla \times \vec{v})$ in terms of the permutation symbol and use the ε-δ-relation to prove Eq. (1.97).

Problem 1.22. Starting from the definition in Eq. (1.36), show that the vector $\vec{S}$ defined in this equation is orthogonal to all $\vec{v}_i$.

Problem 1.23. Consider the following surfaces:

a) $\phi_1(\vec{x}) = x^1 + x^2 + x^3 = 5$

b) $\phi_2(\vec{x}) = (x^1)^2 + (x^2)^2 - x^3 = 0$

c) $\phi_3(\vec{x}) = x^3 - r_0 \cos(kx^1) = -4$

For each surface, parametrise it using the parameters $t = x^1$ and $s = x^2$ and find an expression for the directed area element $d\vec{S}$. Use the fact that the area element is orthogonal to the surface to find a normal vector $\vec{n}$ with magnitude one. Verify that the surface normal in each case is parallel to $\nabla\phi_k$.

Problem 1.24. Show that for a general scalar field $\phi(\vec{x})$, the relation

$$[(\vec{x} \times \nabla) \times (\vec{x} \times \nabla)]\phi = -\vec{x} \times \nabla\phi \tag{1.304}$$

is satisfied.

Problem 1.25. Let $\phi(\vec{x})$ be a scalar field and find the conditions that it has to fulfil in order to satisfy the relation

$$\nabla \times (\nabla\phi \times \vec{a}) = \nabla(\nabla\phi \cdot \vec{a}), \tag{1.305}$$

where $\vec{a}$ is a constant non-zero vector.

Problem 1.26. For a vector field $\vec{v}(\vec{x})$ that fulfils the relation

$$\vec{v}(k\vec{x}) = k^n \vec{v}(\vec{x}), \tag{1.306}$$

where k and n are constants, show that

$$(\vec{x} \cdot \nabla)\vec{v}(\vec{x}) = n\vec{v}(\vec{x}) \tag{1.307a}$$

and compute

$$\nabla \cdot \{\vec{x}[\vec{x} \cdot \vec{v}(\vec{x})]\}. \tag{1.307b}$$

Problem 1.27. For the motion of a rigid body rotating around the point $\vec{x}_0$, the velocity field of the body may be written as

$$\vec{v}(\vec{x}, t) = \vec{\omega}(t) \times (\vec{x} - \vec{x}_0), \tag{1.308}$$

where $\vec{\omega}(t)$ describes the angular velocity of the rotation. Determine an expression for the acceleration $\vec{a}(\vec{x}, t) = d\vec{v}/dt$ and compute the quantities $\nabla \cdot \vec{v}$, $\nabla \times \vec{v}$, $\nabla \cdot \vec{a}$, and $\nabla \times \vec{a}$. *Hint:* When determining the acceleration field, consider the motion $\vec{x}(t)$ of a point in the body that satisfies $d\vec{x}(t)/dt = \vec{v}$.

Problem 1.28. A particle is moving in a force field $\vec{F} = k(x^1\vec{e}_2 - x^2\vec{e}_1)$ from the position $\vec{x} = r_0\vec{e}_1$ to the position $\vec{x} = r_0\vec{e}_2$. Compute the work done by the force when the particle takes the path

a) along a circle of constant radius r_0 and

b) along the straight line between the endpoints.

Determine whether or not the force field is conservative.

Problem 1.29. The *Lorentz force* on a particle with electric charge q in a magnetic field $\vec{B}(\vec{x})$ is given by

$$\vec{F} = q\vec{v} \times \vec{B}, \tag{1.309}$$

where $\vec{v}$ is the particle's velocity $\vec{v} = d\vec{x}/dt$. Show that the total work

$$W = \int \vec{F} \cdot d\vec{x} \tag{1.310}$$

done by the magnetic field on the particle is equal to zero.

Problem 1.30. The half-sphere of radius R defined by $x^3 > 0$ and $\vec{x}^{\,2} = R^2$ can be parametrised by the cylinder coordinates ρ and ϕ as

$$x^1 = \rho\cos(\phi), \quad x^2 = \rho\sin(\phi), \quad x^3 = \sqrt{R^2 - \rho^2}, \tag{1.311}$$

for $\rho < R$. Use this parametrisation to compute the total area of the half-sphere according to

$$A = \int dS. \tag{1.312}$$

Problem 1.31. Consider the vector field

$$\vec{v} = (x^1)^k\vec{e}_1 + (x^2)^k\vec{e}_2 + (x^3)^k\vec{e}_3, \tag{1.313}$$

where k is a positive integer. Compute the flux integral

$$\Phi = \oint_S \vec{v} \cdot d\vec{S}, \tag{1.314}$$

where S is the sphere of radius R with the surface normal pointing away from the origin.

Problem 1.32. In a region where the matter density is given by

$$\rho(\vec{x}) = \frac{\rho_0}{L^2}\vec{x}^{\,2}, \tag{1.315}$$

compute the total mass contained in the volumes given by

a) the cube $0 < x^i < L$ $(i = 1, 2, 3)$ and

b) the sphere $r < L$ in spherical coordinates.

Problem 1.33. Consider a fluid with constant density ρ_0 and the velocity field

$$\vec{v} = \frac{v_0}{L}(x^1\vec{e}_2 - x^2\vec{e}_1 + L\vec{e}_3). \tag{1.316}$$

a) Verify that the velocity field is divergence free $\nabla \cdot \vec{v} = 0$.

b) Compute the total momentum inside the cube $0 < x^i < L$ $(i = 1, 2, 3)$.

Problem 1.34. For the fluid in Problem 1.33, compute the mass flux of the fluid

$$\Phi = \int_S \rho_0 \vec{v} \cdot d\vec{S} \tag{1.317}$$

when the surface S in cylinder coordinates is given by

a) the disc $z = z_0$, $\rho < r_0$, $0 < \phi < 2\pi$, with surface normal $\vec{n} = \vec{e}_3$,

b) the cylinder $0 < z < z_0$, $\rho = r_0$, $0 < \phi < 2\pi$, with surface normal $\vec{n} = \vec{e}_\rho$,

c) the ϕ coordinate surface $0 < z < z_0$, $0 < \rho < r_0$, $\phi = \phi_0$, with surface normal $\vec{n} = \vec{e}_\phi$.

Note: We have here used ρ_0 to denote a density, be careful not to confuse it with the cylinder coordinate ρ.

Problem 1.35. A small element $d\vec{x}$ of a conductor carrying a current I in a magnetic field $\vec{B}$ will be subjected to a force

$$d\vec{F} = -I\vec{B} \times d\vec{x}. \tag{1.318}$$

In a constant magnetic field $\vec{B} = \vec{B}_0$

a) compute the total force $\vec{F}$ acting on the closed loop given by

$$x^1 = r_0\cos(t), \quad x^2 = r_0\sin(t), \quad x^3 = 0, \qquad (0 < t < 2\pi) \tag{1.319}$$

b) compute the total torque $\vec{M}$ relative to the origin for the same loop. The torque due to a force $d\vec{F}$ is given by $d\vec{M} = \vec{x} \times d\vec{F}$.

Problem 1.36. For a static electromagnetic field, the electric field $\vec{E}$ has a scalar potential ϕ and the magnetic field $\vec{B}$ is divergence free. Assuming that S is a closed equipotential surface, i.e., that ϕ takes a constant value ϕ_0 on S, compute the volume integral

$$I = \int_V \vec{E} \cdot \vec{B}\, dV, \tag{1.320}$$

where V is the volume enclosed by S, by using the divergence theorem.

Problem 1.37. Compute the closed surface integral

$$\vec{I} = \oint_S \vec{x} \times d\vec{S} \tag{1.321}$$

by rewriting it as an integral over the volume V enclosed by S and simplifying the resulting expression.

Problem 1.38. Use the generalised version of the curl theorem to rewrite the circulation integral

$$\vec{I} = \oint_\Gamma \vec{x} \times d\vec{x} \tag{1.322}$$

around the closed loop Γ as an integral over a surface S bounded by Γ. Simplify your result as far as possible without knowing the loop Γ and then apply your result to the loop given by

$$x^1(t) = r_0 \cos(t), \quad x^2(t) = r_0 \sin(t), \quad x^3(t) = 0. \tag{1.323}$$

Verify that you obtain the same result from the circulation integral as you get from the surface integral.

Problem 1.39. Given that the scalar field $\phi(\vec{x})$ is constant on the boundary curve of a surface S, show that

$$\int_S [(\nabla\phi) \times (\nabla\psi)] \cdot d\vec{S} = 0 \tag{1.324}$$

for any scalar field $\psi(\vec{x})$.

Problem 1.40. Consider again the fluid treated in Problems 1.33 and 1.34. Rewrite the mass flux out of the closed surface S composed of the cylinder area $-z_0 < z < z_0$, $\rho = r_0$, $0 < \phi < 2\pi$ and the two discs $z = \pm z_0$, $\rho < r_0$, $0 < \phi < 2\pi$ as an integral over the enclosed volume and compute the volume integral. Verify that the flux is equivalent to the flux resulting from adding up the fluxes through S (use your results from Problem 1.34).

Problem 1.41. Compute the integral expression

$$I = \int_V \vec{v} \cdot \vec{w}\, dV, \tag{1.325}$$

where $\vec{v}$ is curl free in the volume V and $\vec{w}$ is orthogonal to the surface normal at the boundary of V and divergence free inside V.

Problem 1.42. The surface integral

$$\vec{\Phi} = \oint_S \frac{\vec{x}}{r^5} \vec{x} \cdot d\vec{S} \tag{1.326}$$

can be rewritten as a volume integral

$$\vec{\Phi} = \int_V \vec{x}\phi(\vec{x})\, dV. \tag{1.327}$$

Determine an expression for the scalar field $\phi(\vec{x})$ assuming that $\vec{x} = 0$ is not within the volume V.

Problem 1.43. Let $\vec{A}$ be a vector field satisfying the relation $\nabla \times (\nabla \times \vec{A}) = 0$ in the volume V and $\vec{n} \cdot [\vec{A} \times (\nabla \times \vec{A})] = 0$ on the boundary surface S of V, where $\vec{n}$ is the surface normal. Show that

$$\int_V (\nabla \times \vec{A})^2 dV = 0. \tag{1.328}$$

Problem 1.44. Use the regular curl theorem to prove the generalised curl theorem, given by Eq. (1.139b).

Problem 1.45. The vector potential $\vec{A}$ of a magnetic dipole with the constant dipole moment $\vec{m}$ is given by

$$\vec{A} = \frac{\mu_0}{4\pi} \frac{\vec{m} \times \vec{x}}{r^3}. \tag{1.329}$$

Compute the corresponding magnetic field $\vec{B} = \nabla \times \vec{A}$ for $r > 0$. Also compute the curl of this magnetic field.

Problem 1.46. With the dual basis at hand, show that the contravariant vector components v^a of a vector $\vec{v}$ may be found as

$$v^a = \vec{E}^a \cdot \vec{v}, \tag{1.330}$$

similar to how the covariant vector components were computed using the tangent vector basis in Eq. (1.162).

Problem 1.47. Verify that the tangent vector bases in cylinder coordinates (see Eqs. (1.196)) and in spherical coordinates (see Eqs. (1.207)) are orthogonal bases.

Problem 1.48. Compute the divergence and curl of the vector field $\vec{v} = \frac{1}{\rho}\vec{e}_\phi$, given in cylinder coordinates, for $\rho > 0$. Use your result to compute the circulation integral

$$I = \oint_\Gamma \vec{v} \cdot d\vec{x}, \tag{1.331}$$

where Γ is the closed curve given by

$$\rho(t) = \rho_0[2 + \sin(t)], \quad \phi(t) = t, \quad z(t) = t(t - 4\pi), \tag{1.332}$$

where $0 < t < 4\pi$.

Problem 1.49. *Hyperbolic coordinates* u and v in two dimensions may be introduced to parametrise the region $x^1, x^2 > 0$ according to

$$x^1 = ve^u, \quad x^2 = ve^{-u}. \tag{1.333}$$

a) Sketch the coordinate lines corresponding to u and v.

b) Find the inverse transformation that expresses u and v as functions of x^1 and x^2.

c) Compute the tangent vector and dual bases.

d) Determine whether or not hyperbolic coordinates form an orthogonal coordinate system. If they do, compute the scale factors. If they do not, compute the inner products $\vec{E}_u \cdot \vec{E}_v$ and $\vec{E}^u \cdot \vec{E}^v$.

Problem 1.50. The *parabolic coordinates* t and s in two dimensions are defined through the relations

$$x^1 = ts, \quad x^2 = \frac{1}{2}(t^2 - s^2), \tag{1.334a}$$

where $t, s > 0$.

a) Sketch the coordinate lines corresponding to t and s.

b) Find the inverse transformation that expresses t and s as functions of x^1 and x^2.

c) Compute the tangent vector and dual bases.

d) Determine whether or not hyperbolic coordinates form an orthogonal coordinate system. If they do, compute the scale factors. If they do not, compute the inner products $\vec{E}_t \cdot \vec{E}_s$ and $\vec{E}^t \cdot \vec{E}^s$.

e) Find the expressions for the gradient, divergence, curl, and Laplace operators in three dimensions when also parametrising the third direction with an additional coordinate z as

$$x^3 = z. \tag{1.334b}$$

f) Express the position vector $\vec{x}$ in parabolic coordinates.

Problem 1.51. A screened point charge at the origin may be described by the Yukawa potential

$$\phi(\vec{x}) = \frac{q}{4\pi r} e^{-kr}, \tag{1.335}$$

where k is a constant and r is the radial spherical coordinate. Compute the flux of the corresponding vector field $\vec{v} = -\nabla\phi$ out of the sphere of radius R both by directly performing the surface integral and by applying the divergence theorem. What can be said about $\nabla^2\phi(\vec{x})$ at the origin?

Problem 1.52. Compute $\nabla \cdot \vec{x}$ and $\nabla \times \vec{x}$ using the expressions for the divergence and curl in cylinder and spherical coordinates. Verify that the results are the same as the result you would find doing the same computation in Cartesian coordinates.

Problem 1.53. Verify that the vector field

$$\vec{v} = \frac{1}{r^2}\vec{e}_r \tag{1.336}$$

is divergence free for $r > 0$. According to our discussion on potentials, this implies that $\vec{v}$ has a vector potential $\vec{A}$ such that

$$\vec{v} = \nabla \times \vec{A}. \tag{1.337}$$

Find a vector potential $\vec{A}$ of $\vec{v}$ that has the form $\vec{A} = A_\varphi \vec{e}_\varphi$ and is divergence free.

Problem 1.54. Assume that the vector fields $\vec{v}(\vec{x})$ and $\vec{w}(\vec{x})$ satisfy

$$\nabla \cdot \vec{v} = \nabla \cdot \vec{w} \quad \text{and} \quad \nabla \times \vec{v} = \nabla \times \vec{w} \tag{1.338a}$$

within a volume V and that

$$\vec{n} \cdot \vec{v} = \vec{n} \cdot \vec{w} \tag{1.338b}$$

on the boundary surface of V. Show that $\vec{v}(\vec{x}) = \vec{w}(\vec{x})$ within the volume V.

Problem 1.55. The force on a charge q at position $\vec{x} \neq 0$ from an electric dipole of dipole moment $\vec{p} = p\vec{e}_3$ at the origin is given by

$$\vec{F} = pq\left(\frac{2\cos(\theta)}{r^3}\vec{e}_r + \frac{\sin(\theta)}{r^3}\vec{e}_\theta\right) \tag{1.339}$$

in spherical coordinates. Compute $\nabla \cdot \vec{F}$ and $\nabla \times \vec{F}$. Does this force field have a corresponding scalar and vector potential?

2.3.5 Orthogonal coordinates

Let us briefly examine some of the above ideas applied to an orthogonal coordinate system. The first observation we can make is that, due to the orthogonality, the metric tensor is diagonal by definition

$$(g_{ab}) = \mathrm{diag}(h_1^2, h_2^2, \ldots, h_N^2), \tag{2.130}$$

where the h_a are the scale factors. This follows directly from the relation $g_{ab} = \vec{E}_a \cdot \vec{E}_b$ while keeping in mind that the coordinate system is orthogonal. Inserting this into the expression for the Christoffel symbols in Eq. (2.86), we find that

$$\Gamma^c_{ab} = \delta_{cb}\partial_a \ln(h_b) + \delta_{ca}\partial_b \ln(h_a) - \frac{\delta_{ab}}{2h_c^2}\partial_c h_a^2, \qquad \text{(no sum)} \tag{2.131}$$

where we have used that $(\partial_a h_b^2)/h_b^2 = 2\partial_a \ln(h_b)$. As can be seen from this expression, only Christoffel symbols that contain at least two of the same index can be non-zero. Furthermore, the scale factor belonging to that index must depend on the remaining index. In particular, we obtain

$$\Gamma^c_{ab} = \begin{cases} \partial_b \ln(h_a), & (c = a) \\ -\frac{1}{2h_c^2}\partial_c h_a^2, & (a = b \neq c) \end{cases}. \tag{2.132}$$

Due to the metric being diagonal, the metric determinant is given by the product of the diagonal elements

$$g = \prod_{a=1}^{N} h_a^2 \quad \Longrightarrow \quad \sqrt{g} = \prod_{a=1}^{N} h_a. \tag{2.133}$$

We note that the expression for $\sqrt{g}$ is exactly the same as that for the Jacobian determinant $\mathcal{J}$ in Eq. (1.181) for transformations to an orthogonal coordinate system from a Cartesian one. Naturally, this is no coincidence since $\sqrt{g} = 1$ in any Cartesian system and $\sqrt{g}$ is a scalar density of weight one.

Example 2.23 The polar coordinates ρ and ϕ in $\mathbb{R}^2$ constitute an orthogonal coordinate system as discussed in Section 1.6.3, where it was also found that

$$h_\rho = 1, \quad h_\phi = \rho. \tag{2.134}$$

From this follows that

$$g_{\rho\rho} = 1, \quad g_{\phi\phi} = \rho^2, \quad g_{\rho\phi} = g_{\phi\rho} = 0. \tag{2.135}$$

As we only have two coordinates, the argumentation above does not reduce the possible number of non-zero Christoffel symbols. However, since none of the scale factors depend on ϕ, the Christoffel symbols containing derivatives with respect to ϕ are all identically equal to zero. This includes

$$\Gamma^\phi_{\phi\phi} = \partial_\phi \ln(h_\phi) = 0, \tag{2.136a}$$

$$\Gamma^\phi_{\rho\rho} = -\frac{1}{2h_\phi}\partial_\phi h_\rho^2 = 0, \tag{2.136b}$$

$$\Gamma^\rho_{\rho\phi} = \Gamma^\rho_{\phi\rho} = \partial_\phi \ln(h_\rho) = 0. \tag{2.136c}$$

At the same time, the scale factor h_ρ is constant, which gives us

$$\Gamma^\rho_{\rho\rho} = \partial_\rho \ln(h_\rho) = 0. \tag{2.136d}$$

Finally, the non-zero Christoffel symbols are

$$\Gamma^{\rho}_{\phi\phi} = -\frac{1}{2h_\rho}\partial_\rho h_\phi^2 = -\rho, \tag{2.136e}$$

$$\Gamma^{\phi}_{\phi\rho} = \Gamma^{\phi}_{\rho\phi} = \partial_\rho \ln(h_\phi) = \frac{1}{\rho}. \tag{2.136f}$$

These results are in accordance with the derivatives of the tangent vector basis

$$\partial_\rho \vec{E}_\rho = 0, \quad \partial_\phi \vec{E}_\rho = \frac{1}{\rho}\vec{E}_\phi, \quad \partial_\rho \vec{E}_\phi = \frac{1}{\rho}\vec{E}_\phi, \quad \partial_\phi \vec{E}_\phi = -\rho\vec{E}_\rho, \tag{2.137}$$

which can be derived starting from Eq. (1.196). Note that it will often be easier to compute the Christoffel symbols directly from their definition in terms of the partial derivatives of the tangent vector basis.

2.4 TENSORS IN CARTESIAN COORDINATES

As mentioned in the beginning of this chapter, many texts will start the discussion on tensors by considering the special case of tensors in Cartesian coordinates, much in the same way as we first discussed scalars and vectors in Cartesian coordinates in Chapter 1. Instead, we choose to start by discussing the full framework of tensors in arbitrary coordinate systems, using the definition in terms of the outer product in order to build a deeper foundation for our upcoming discussion of more general spaces in Chapter 9, where the discussion on arbitrary coordinate systems will provide us with some additional intuition. However, before jumping into the more general framework, let us take a step back and look at the special case of Cartesian coordinate systems.

We start this endeavour by taking a set of constant orthonormal unit vectors $\vec{e}_i$ and fixing an origin. Any point in space is then described by the coordinates x^i, where the position vector $\vec{x} = x^i\vec{e}_i$ describes the displacement from the origin. The fact that the basis vectors are constant and orthonormal quickly leads us to the conclusion that

$$\vec{e}_i = \vec{E}^i = \vec{E}_i, \tag{2.138}$$

meaning that there is no difference between the tangent vector basis and the dual basis. Thus, in Cartesian coordinates, there is no need to distinguish covariant and contravariant indices and we can write all indices up or down at our leisure. Naturally, there is an argument for keeping the indices where they would belong in a more general setting.

Even when keeping to Cartesian coordinates, we may change to a different Cartesian coordinate system. Selecting the new Cartesian coordinate system $x'^{i'}$, which has the basis vectors $\vec{e}\,'_{i'}$ and whose origin is displaced with respect to the origin of the x^i coordinates by a vector $\vec{A}$, we find that

$$\vec{x} = \vec{x}\,' + \vec{A} \quad\Longrightarrow\quad x'^{i'} = (x^i - A^i)\vec{e}\,'_{i'}\cdot\vec{e}_i = a^{i'}_i(x^i - A^i), \tag{2.139}$$

where $a^{i'}_i = \vec{e}\,'_{i'}\cdot\vec{e}_i$ are the transformation coefficients defined in connection to Eq. (1.11). As expected, we find that the more general tensor transformations using partial derivatives reduce to these transformation coefficients as

$$\frac{\partial x'^{i'}}{\partial x^i} = a^{i'}_i = a^i_{i'} = \frac{\partial x^i}{\partial x'^{i'}}. \tag{2.140}$$

Again, it becomes apparent that there is no difference between covariant and contravariant indices in Cartesian coordinates as they transform in the same fashion due to the $i' \leftrightarrow i$ symmetry of the transformation coefficients. With the basis being independent of the position, we may also have tensors that are defined without reference to a particular point in space. In fact, the very first example, Example 2.1 of this chapter, involved such a tensor, the moment of inertia, which is a property of an extended object rather than a single point, although it is expressed with respect to a point (usually the center of mass).

Example 2.24 The Kronecker delta δ^a_b as discussed earlier is generally a tensor. When restricted to Cartesian coordinate systems, we may exchange it for δ_{ij}, which we discussed already in Chapter 1. We have here been able to move the contravariant index down as there is no difference between covariant and contravariant indices in Cartesian coordinates. We also note that, due to the fact that $\vec{e}_i \cdot \vec{e}_j = \delta_{ij}$ and $\vec{E}_i = \vec{e}_i$ in Cartesian coordinates, the metric tensor in Cartesian coordinates becomes

$$g_{ij} = \vec{e}_i \cdot \vec{e}_j = \delta_{ij}. \tag{2.141}$$

In fact, an equivalent definition of transformations between Cartesian coordinate systems is affine transformations that preserve the metric tensor as δ_{ij}.

Naturally, since all basis vectors and components of the metric components are constant, the Christoffel symbols are all identically equal to zero in Cartesian coordinates.

Since the metric tensor for a Cartesian coordinate system is just the Kronecker delta δ_{ij}, it follows directly that the metric determinant is given by $g = 1$ in all such systems. At the same time, we are already aware that the metric determinant is a scalar density with weight two and therefore should transform from frame to frame along with a factor $\mathcal{J}^2$. There is only one way these two can be compatible, which is that any transition between Cartesian coordinate systems must have a Jacobian $\mathcal{J} = \pm 1$, where the positive sign holds for transitions between coordinate systems with the same handedness and the negative sign for transitions between systems with opposite handedness. This also follows from the fact that Cartesian coordinate transformations preserve the δ_{ij} form of the metric tensor, which imposes constraints on the transformation coefficients $a^{i'}_i$. From this insight follows that there is no distinction between tensor densities of different weights as long as we restrict ourselves to right-handed Cartesian coordinates. In particular, all tensor densities can be considered as tensors in right-handed Cartesian coordinates due to the tensor density transformation of Eq. (2.110) reducing to the normal tensor transformation when $\mathcal{J} = 1$, regardless of the weight w.

Example 2.25 Our prime example of a tensor density was the permutation symbol $\varepsilon_{a_1 \ldots a_N}$. When viewed in right-handed Cartesian coordinates, it can be taken as $\varepsilon_{i_1 \ldots i_N} = \varepsilon^{i_1 \ldots i_N} = \eta_{i_1 \ldots i_N}$. Again we note that the positioning of the indices does not matter in this case. The permutation symbol tensor densities, regardless of whether they are viewed as covariant or contravariant, are equivalent to each other and so is the totally anti-symmetric tensor $\eta_{i_1 \ldots i_N}$.

2.5 TENSOR INTEGRALS

Tensor integrals in general coordinates will not necessarily be well defined. This will be particularly true when we discuss curved spaces, where there is no unique way of relating vectors at different points in the base space to each other. In this section, we will therefore start by restricting ourselves to the integration of tensors in Cartesian coordinates and later examine which concepts we can carry with us to integrals in more general spaces.

2.5.1 Integration of tensors in Cartesian coordinates

The integration of vectors discussed in Chapter 1 can be readily extended to the case of tensors in Cartesian coordinates. As was the case for the integration of vectors, the volume, surface, and line elements can be parametrised and computed and the volume element will be an infinitesimal number, while the surface and line elements will be infinitesimal vector quantities. Since the Cartesian bases $e_{i_1 \ldots i_n}$ for any type of tensors are constant, they can be taken out of the integral and each component can be integrated as a separate function.

2.5.1.1 *Volume integration*

For volume integrals, we first note that the volume element takes on the same expression regardless of the coordinates as long as we use the Cartesian coordinates as integration parameters. In particular, we find that

$$dV = \vec{e}_1 \cdot (\vec{e}_2 \times \vec{e}_3)\, dx^1 dx^2 dx^3 = \varepsilon_{123} dx^1 dx^2 dx^3 = dx^1 dx^2 dx^3. \tag{2.142}$$

This also follows from the fact that $\mathcal{J} = 1$ for all Cartesian coordinate transformations and we will see the generalisation to arbitrary coordinates when we discuss more general integrations.

Since the volume element is an infinitesimal number, the volume integral of any tensor $T_{i_1 \ldots i_n}$ is a tensor $I_{i_1 \ldots i_n}$ of the same rank and is given by

$$I_{i_1 \ldots i_n} = \int_V T_{i_1 \ldots i_n} dV = \int_V T_{i_1 \ldots i_n} dx^1 dx^2 dx^3, \tag{2.143}$$

where V is the volume over which the integral is taken.

Example 2.26 The moment of inertia tensor of a solid object, which we have discussed earlier, may be expressed using a volume integral. The velocity $\vec{v}$ of the object at a point $\vec{x}$ is given by

$$\vec{v} = \vec{\omega} \times (\vec{x} - \vec{x}_0), \tag{2.144}$$

where $\vec{x}_0$ is the fixed point of the rotation. The mass within a small volume dV around this point is $dm = \rho\, dV$, where ρ is the mass density of the object. The angular momentum of this mass with respect to $\vec{x}_0$ is therefore given by $d\vec{L} = (\vec{x} - \vec{x}_0) \times dm\, \vec{v}$ and the total angular momentum of the solid can therefore also be written

$$L_i = \vec{e}_i \cdot \int_V \rho \vec{x} \times (\vec{\omega} \times \vec{x}) dV = \omega_j \int_V \rho (x_k x_k \delta_{ij} - x_i x_j) dV, \tag{2.145}$$

where we have set $\vec{x}_0 = 0$ for simplicity and V is the full extension of the solid. Identification of this expression with Eq. (2.2) shows that the moment of inertia tensor is given by the volume integral of the tensor $\rho(x_k x_k \delta_{ij} - x_i x_j)$

$$I_{ij} = \int_V \rho(x_k x_k \delta_{ij} - x_i x_j) dV. \tag{2.146}$$

In particular, we note that if the volume mass is purely along the rotational axis, i.e., $\vec{\omega} \propto \vec{x}$, then the angular momentum is zero as it should be.

2.5.1.2 *Surface integrals*

Since the surface element is an infinitesimal vector quantity, we can use it to either increase the rank of the tensor being integrated by one by performing an outer product or decrease it by one by contraction. There is also the possibility of using the scalar surface element, in which case the integral results in a tensor of the same type as the integrand, just as we saw for the volume integral. Naturally, which of these options should be used depends on the physical situation that we wish to describe. Just as for the integrals of vector quantities, the surface element will be given by Eq. (1.59).

Example 2.27 The stress tensor σ_{ij} relates a surface element $d\vec{S}$ to a force acting across the surface as described in Example 2.7. Integrating over the surface S of an object, we can find the total force acting on the object as

$$F_i = \oint_S df_i = \oint_S \sigma_{ij} dS_j. \tag{2.147}$$

It should be noted that this is the external force on the object resulting from contact forces at the surface. There is still the possibility of a force, such as gravity, acting directly on the constituents within the volume. For example, the gravitational force on an object of mass density ρ that takes up the volume V would be given by the volume integral

$$\vec{F}_g = \int_V \rho \vec{g}\, dV, \tag{2.148}$$

where $\vec{g}$ is the external gravitational field.

Example 2.28 Coming back to the moment of inertia tensor, which we saw can be written as a volume integral, some mass distributions will allow us to disregard one dimension of the distribution. In the case we started with, a disc of mass M, we neglected the thickness of the disc. This can be done as long as the thickness is much smaller than the radius R of the disc and the volume integral with the volume density ρ can be replaced by a surface integral with the surface density ρ_s

$$I_{ij} = \int_V \rho(\delta_{ij} x_k x_k - x_i x_j) dV \to \int_S \rho_s(\delta_{ij} x_k x_k - x_i x_j) dS. \tag{2.149}$$

In our case with the disc, we can compute the moment of inertia tensor about its center by the following consideration. We start by introducing a coordinate system such that the disc is in the x_1-x_2-plane. First of all, the term with $x_i x_j$ will be equal to zero whenever i or j is equal to three as $x_3 = 0$ everywhere on the disc. Furthermore, when $(i, j) = (1, 2)$ the integrand for the term $x_1 x_2$ is anti-symmetric with respect to $x_1 \to -x_1$ (and $x_2 \to -x_2$), while the integration domain is symmetric. Thus, the second term will only give a contribution when $i = j = 1$ or 2. Due to the symmetry of the disc, the value will be the same for both these cases and we find

$$\int_S \rho x_1^2 dS = \int_0^R r^3\, dr \int_0^{2\pi} \rho_s \cos^2(\varphi) d\varphi = \pi \rho_s \frac{R^4}{4}. \tag{2.150}$$

For the first term including the δ_{ij}, we find that

$$\int_S \delta_{ij} \rho_s x_k x_k dS = \delta_{ij} \rho_s \int_0^R r^3 dr \int_0^{2\pi} d\varphi = \pi \rho_s \delta_{ij} \frac{R^4}{2}. \tag{2.151}$$

We now take into account that the mass of the disc is

$$M = \int_S \rho_s dS = \pi R^2 \rho_s, \tag{2.152}$$

resulting in

$$I_{ij} = \frac{MR^2}{4} (\delta_{ij} + \delta_{i3}\delta_{j3}). \tag{2.153}$$

Thus, the non-zero components of the moment of inertia tensor in this coordinate system are $I_{11} = I_{22} = MR^2/4$ and $I_{33} = MR^2/2$. These are precisely the values quoted in the beginning of this chapter. Equivalently, including the tensor basis, this may be expressed as

$$I = \frac{MR^2}{4} (\delta + e_{33}), \tag{2.154}$$

where δ is the rank two Kronecker delta and $e_{33} = \vec{e}_3 \otimes \vec{e}_3$.

2.5.1.3 Line integrals

As was the case with the surface integral, the line element will generally come with a direction. Thus, also for a line integral, the resulting tensor may be of one order lower or higher than the tensor being integrated, depending on whether or not there is a contraction between the integrated tensor and the line element. We may also find situations when the direction of the line element is unimportant, resulting in a tensor of the same order as the integrand.

Example 2.29 The magnetic force on an infinitesimal part $d\vec{x}$ of a conductor carrying a current I is given by $d\vec{F} = I d\vec{x} \times \vec{B}$, where $\vec{B}$ is the external magnetic field. The total force on the conductor may therefore be written as a line integral

$$\vec{F} = I \int_\Gamma d\vec{x} \times \vec{B}, \tag{2.155}$$

where Γ is the curve in space occupied by the conductor. In terms of the magnetic field tensor $F_{ij} = \varepsilon_{ijk}B^k$ introduced earlier, the force can be rewritten as

$$F_i = I\int_\Gamma F_{ij}dx^j, \tag{2.156}$$

i.e., in terms of the line integral of F_{ij} contracted with the line element. Note that we have here used F_i and F_{ij} to denote different tensors, the first is the total force vector and the second the magnetic field tensor.

2.5.1.4 *Integral theorems*

As already mentioned, the integral theorems discussed in Section 1.5 may be generalised to the forms

$$\int_V \partial_i T_{j_1\ldots j_n}\,dV = \oint_S T_{j_1\ldots j_n}\,dS_i, \tag{2.157a}$$

$$\int_S \varepsilon_{ijk}\partial_k T_{\ell_1\ldots\ell_n}\,dS_j = \oint_\Gamma T_{\ell_1\ldots\ell_n}\,dx^i, \tag{2.157b}$$

where we have replaced the f of Eqs. (1.139) with the general tensor component $T_{j_1\ldots j_n}$ as already alluded to when we first discussed these theorems. Contractions of any of the indices of $T_{j_1\ldots j_n}$ with the free index i are also possible and the integral theorems will still hold since it does not matter whether the contraction is done before or after the integration. This may also be seen by writing the contraction in terms of the constant Kronecker delta, which may be moved out of the integral due to its components being constant, for example

$$\int_V \partial_i T_{ijk}dV = \delta_{i\ell}\int_V \partial_i T_{\ell jk}dV = \delta_{i\ell}\oint_S T_{\ell jk}dS_i = \oint_S T_{ijk}dS_i. \tag{2.158}$$

Just as the divergence and curl theorems, these generalised integral theorems are powerful tools that may be used in order to simplify several calculations that might otherwise be hard to perform.

Example 2.30 The surface force acting on an object can be found by taking the surface integral

$$F_i = \oint_S \sigma_{ij}dS_j \tag{2.159}$$

over its entire surface, where σ_{ij} is the stress tensor. By the generalisation of the divergence theorem, we find that

$$F_i = \int_V \partial_j\sigma_{ij}dV. \tag{2.160}$$

A special application of this formula is *Archimedes' principle*. For an object submerged in a fluid, the contact force from the fluid on the object will be the same as that which the fluid would exert on an equivalent volume of fluid that replaced the object, see Fig. 2.2. Within a stationary fluid of density ρ_0 in a homogeneous gravitational field $\vec{g}$, the stress tensor is given by $\sigma_{ij} = -p\delta_{ij}$, where $p = \rho_0\vec{g}\cdot\vec{x}$ is the pressure in the fluid at $\vec{x}$. It follows that

$$F_i = -\int_V \delta_{ij}\partial_j p\,dV = -\int_V \partial_i\rho_0 g_j x^j\,dV = -g_i\rho_0 V = -g_i M_V, \tag{2.161}$$

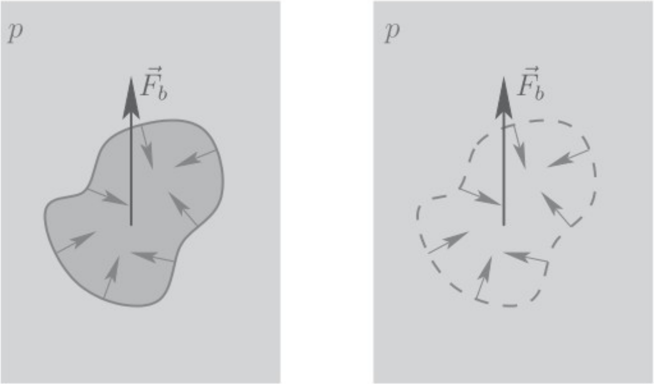

Figure 2.2 According to Archimedes' principle, the buoyant force $\vec{F}_b$ on an object immersed in a fluid with the pressure field p is equal to the force on the corresponding volume if the object was replaced by fluid. The buoyant force can be expressed as an integral over the object's surface using the stress tensor.

where $M_V = \rho_0 V$ is the mass of the displaced fluid. The force is therefore in the opposite direction of the gravitational field and equal in magnitude to the gravitational force on a mass M_V. Including the gravitational force, the total force on the object will therefore be given by

$$\vec{F}_{\text{total}} = \vec{g}(\rho - \rho_0)V, \tag{2.162}$$

where ρ is the density of the object itself. If the object is denser than the fluid, the net force is in the same direction as the gravitational field and this situation is referred to as the object having negative buoyancy. If the fluid is denser than the object, the total force is directed in the opposite direction from the gravitational field and the object has positive buoyancy. In the situation where $\rho = \rho_0$, the net force is zero and the buoyancy is neutral.

2.5.2 The volume element and general coordinates

When discussing volume integrals in general, we have so far only expressed the volume element in Cartesian or orthogonal coordinates, where it takes the form

$$dV = \prod_{a=1}^{N} h_a dx^a. \qquad \text{(no sum)} \tag{2.163}$$

This expression was derived from the volume spanned by the tangent vector basis $\vec{E}_a$ and the approach remains valid for a general coordinate system. Wanting to express the volume described by $y_0^a < y^a < y_0^a + dy^a$, the volume is spanned by the vectors $\vec{E}_a dy^a$ (no sum), see Fig. 2.3. In three dimensions, this results in

$$dV = \vec{E}_1 \cdot (\vec{E}_2 \times \vec{E}_3) dy^1 dy^2 dy^3. \tag{2.164}$$

Expressing this triple product in Cartesian coordinates, we find that

$$dV = \varepsilon_{ijk} \frac{\partial x^i}{\partial y^1} \frac{\partial x^j}{\partial y^2} \frac{\partial x^k}{\partial y^3} dy^1 dy^2 dy^3. \tag{2.165}$$

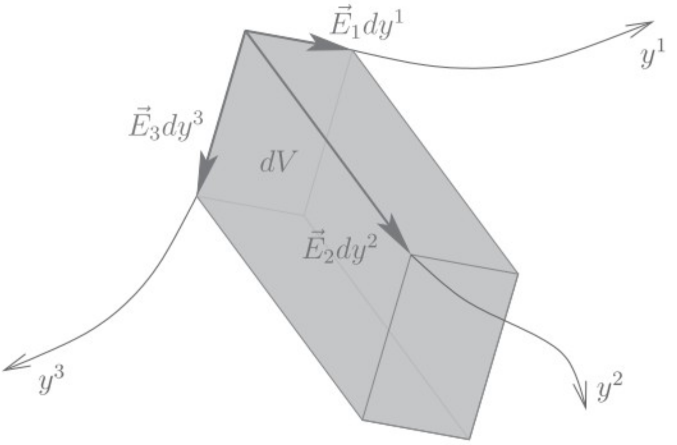

Figure 2.3 In a general coordinate system, the volume dV described by the coordinates $y_0^a < y^a < y_0^a + dy^a$ is spanned by the vectors $\vec{E}_a dy^a$ (no sum). The volume is therefore given by the triple product $dV = \vec{E}_1 \cdot (\vec{E}_2 \times \vec{E}_3) dy^1 dy^2 dy^3$. This is a good approximation as long as the dy^a are small and exact when they are infinitesimal.

The quantity multiplying the differentials dy^a now looks stunningly familiar as the 123 component of a completely anti-symmetric tensor with Cartesian components ε_{ijk}. Fortunately, we already know that this tensor is $\eta_{abc} = \sqrt{g}\,\varepsilon_{abc}$ and it directly follows that

$$dV = \sqrt{g}\,\varepsilon_{123} dy^1 dy^2 dy^3 = \sqrt{g}\,dy^1 dy^2 dy^3. \tag{2.166}$$

The generalisation of this argumentation to an N dimensional space is straightforward and we find that

$$dV = \sqrt{g} \prod_{k=1}^{N} dy^{a_k} \equiv \sqrt{g}\,d^N y. \tag{2.167}$$

This form of the volume element should not come as a surprise as we have already seen that the volume element expressed in a general orthogonal coordinate system is $dV = \mathcal{J} d^N y$, where $\mathcal{J}$ is the product of the scale factors, which, as noted earlier, is the square root of the metric determinant in such a coordinate system.

Despite our success in expressing the volume element, integration in a general coordinate system faces another obstacle that is not present in Cartesian coordinates. Unlike the Cartesian basis vectors, the tangent vector basis $\vec{E}_a$ and the dual basis $\vec{E}^a$ are not constant. This results in the complication that we cannot perform integrals of tensor quantities in general coordinates by just integrating the components one by one. For integrals of tensor quantities, we face the problem of which basis to use to express the result. Being an integral over a volume, surface, or curve, the resulting tensor quantity does not belong to a particular point in space, so there can be no preference for any particular basis. In an affine space, we may brush this aside and express the result in a Cartesian coordinate system, but in more general curved spaces, there will not exist a natural way of comparing, much less defining the sum of, tensors at different points in space. However, scalar quantities do not require a basis and integrals of quantities with all indices contracted will therefore continue to make sense even in general spaces.

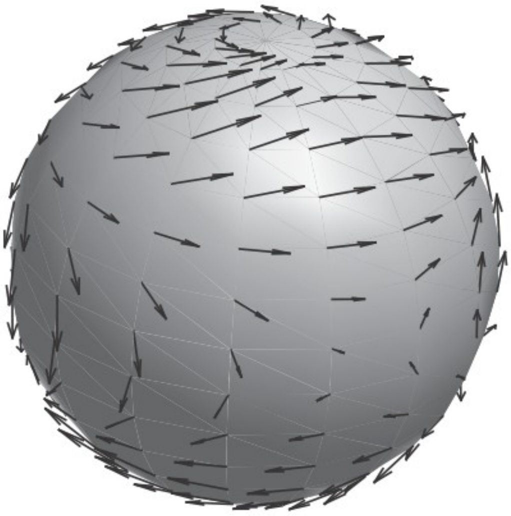

Figure 2.4 While the flow of a substance on a sphere may be defined as a vector at each point of the sphere, there is no meaningful definition of the vector integral of this flow. However, it is possible to define integrals on the sphere that describe, for example, the total flux into or out of a given region.

Example 2.31 Consider the concentration of a substance on the surface of a sphere. At each point on the sphere, there is a tangent vector J^a describing the flow of the substance, see Fig. 2.4, but the tangent vector spaces at different points are different. There is therefore no relevant way of integrating over the sphere to obtain the average flow of the substance as a vector in any of the tangent spaces of the sphere. Of course, there may be some notion of such an average flow if we are looking at the sphere as embedded in our three dimensional affine space, but with a description that is restricted to the sphere only, this is an impossibility. However, the flux integral

$$\Phi = \int_\Gamma J^a dS_a \tag{2.168}$$

of the vector field J^a over some curve Γ, where dS_a is orthogonal to the curve, is a well defined quantity that describes the amount of substance flowing across Γ per time. These issues will be discussed in more detail in Chapter 9.

2.6 TENSOR EXAMPLES

The last example of the preceding section gives us an appetiser of things to come when we will consider more general spaces. We postpone this discussion for later as the discussion will benefit from material that will be covered in the following chapters, in particular Chapter 8 on variational calculus will be of importance. Instead, we conclude this chapter by giving some more structured examples of the application of tensor analysis within different fields of physics. It should be stressed that these examples are only written for the purpose of

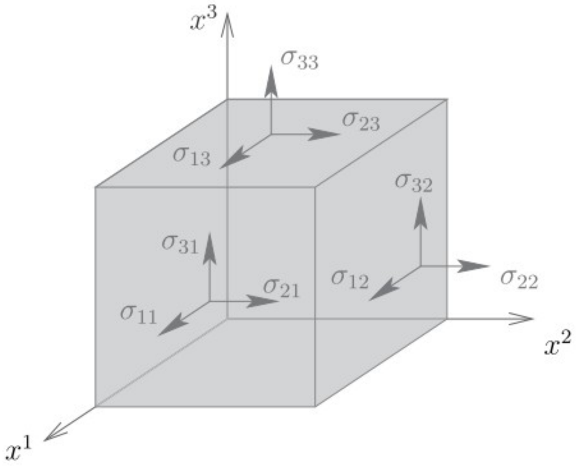

Figure 2.5 The nine components of the stress tensor represented as forces acting on the different faces of a cube. The stress σ_{ij} represents a force in the x^i direction on a surface with a normal $\vec{e}_j$.

discussing the application of tensor analysis in different fields and not as guides or learning material for the subject matters. Each subject is much broader than what can be presented here and will require dedicated texts to understand in full. However, it will provide us with some basics and a context in which we can apply the knowledge acquired throughout this chapter.

2.6.1 Solid mechanics

Solid mechanics is the study of solids and the forces that act within them. In the examples given in this section, we will restrict ourselves to static situations, where any object is in force (and torque) equilibrium. While rank two tensors are relatively common in many areas of physics, as they represent the most general linear relationship between two vectors, solid mechanics presents us with two of the most regularly encountered rank four tensors, the stiffness and compliance tensors.

2.6.1.1 *The stress tensor*

We are already familiar with one of the main actors of solid mechanics from our earlier examples. The *stress tensor* is a rank two tensor σ_{ij} that relates the directed surface element dS_i to the contact force dF_i acting on that surface element

$$dF_i = \sigma_{ij} dS_j. \tag{2.169}$$

For example, the component σ_{12} describes the force per unit area in the x^1 direction when the area element has a surface normal in the x^2 direction, see Fig. 2.5. For a solid in static equilibrium, we can find a set of differential equations that describe how the stress tensor depends on any volume forces, most commonly a gravitational force. We do this by considering an arbitrary volume V with a boundary surface S. Assuming that the body is subject to a volume force $f_i dV$, where f_i is a vector field, the total force on the volume is

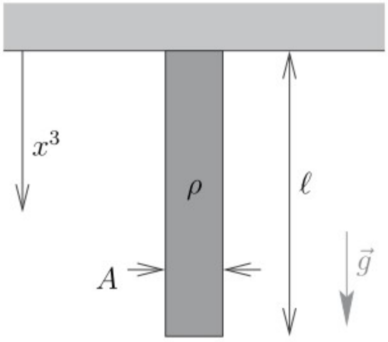

Figure 2.6 A rod of density ρ, cross sectional area A, and length ℓ hanging from the ceiling will be subject to internal stresses. The x^3 direction is taken to point in the same direction as the homogeneous gravitational field $\vec{g}$.

given by

$$F_i = \oint_S \sigma_{ij} dS_j + \int_V f_i dV, \tag{2.170}$$

where the first term represents the surface forces from the rest of the solid and the second term represents the volume force. Applying the generalised divergence theorem to this relation, we find that

$$F_i = \int_V (\partial_j \sigma_{ij} + f_i) dV = 0, \tag{2.171}$$

where we have used that in order for the solid to be in static equilibrium, its acceleration, and therefore the total force acting on it, must be equal to zero. Since the volume V was arbitrary, this relation must hold for any volume, which is true only if the integrand is identically equal to zero, i.e.,

$$\partial_j \sigma_{ij} = -f_i. \tag{2.172}$$

Given appropriate boundary conditions, this differential equation will allow us to solve for the stress tensor for an object subjected to a force density f_i.

Example 2.32 Consider a homogeneous rod of density ρ and square cross section of area A that is hanging from the ceiling as shown in Fig. 2.6. The volume force density acting on this rod due to gravity is given by $f_i = \rho g \delta_{i3}$, where the x^3 direction is defined to be in the same direction as the gravitational field. The resulting differential equation for the stress tensor is

$$\partial_j \sigma_{ij} = -\rho g \delta_{i3}. \tag{2.173}$$

With no forces acting perpendicular to the x^3 direction, the resulting differential equations for $i = 1, 2$ are solved by $\sigma_{1j} = \sigma_{2j} = 0$. Assuming no forces acting on the free surfaces of the rod, we also find $\sigma_{31} = \sigma_{32} = 0$ as well as the only non-zero component

$$\sigma_{33} = \rho g(\ell - x^3). \tag{2.174}$$

Note that this gives a force

$$\vec{F} = -\vec{e}_3 \rho g A \ell = -\vec{e}_3 M g, \tag{2.175}$$

where M is the mass of the rod, from the ceiling acting on the top of the rod. Not surprisingly, this is the exact force needed to cancel the gravitational force of the rod.

Force equilibrium is not sufficient for an object to be in static equilibrium, we also need to impose torque equilibrium in order not to induce rotations. Again looking at a volume V, the torque $\vec{\tau}$ relative to the origin is given by

$$\tau_i = \varepsilon_{ijk}\left(\oint_S x^j \sigma_{k\ell} dS_\ell + \int_V x^j f_k dV\right). \tag{2.176}$$

Note that, due to the force equilibrium that we have already imposed, torque equilibrium around the origin is equivalent to torque equilibrium about any other point. Just as for the force equilibrium, we can apply the generalised divergence theorem to the surface integral, resulting in

$$\begin{aligned}\tau_i &= \int_V \varepsilon_{ijk}(\partial_\ell x^j \sigma_{k\ell} + x^j f_k) dV = \int_V \varepsilon_{ijk}[\sigma_{kj} + x^j(\partial_\ell \sigma_{k\ell} + f_k)]dV \\ &= \int_V \varepsilon_{ijk}\sigma_{kj} dV,\end{aligned} \tag{2.177}$$

where we have applied the requirement of force equilibrium in the last step. Demanding torque equilibrium requires that $\vec{\tau} = 0$ and since this has to hold for any volume V, the integrand must be equal to zero. Multiplying by $\varepsilon_{i\ell m}$, we find that

$$\varepsilon_{i\ell m}\varepsilon_{ijk}\sigma_{jk} = 2\sigma_{[\ell m]} = 0, \tag{2.178}$$

meaning that the anti-symmetric part of the stress tensor σ_{ij} must vanish, i.e., the stress tensor is symmetric.

2.6.1.2 The strain tensor

We have so far only described a situation where we have been concerned with the internal forces inside a solid. If the solid is very stiff, the forces will not cause noticeable deformations of the solid. However, for some applications, it is necessary to also know how the material deforms based on the internal stresses. In order to handle this, we first need a good description of the deformations of the solid. We begin this endeavour by defining the *displacement field* $\vec{u}(\vec{x})$, which describes the displacement of a point in the solid that would be at position $\vec{x}$ if the solid was not strained. It follows that this point is located at

$$\vec{x}\,' = \vec{x} + \vec{u}(\vec{x}). \tag{2.179}$$

Although this field describes how each point of the solid has been displaced, a non-zero displacement field does not necessarily mean that the solid is strained. Possibilities for such situations include translations and rotations

$$x'^i = R_{ij}x^j + a^i \tag{2.180}$$

of the solid, where R_{ij} are the components of a rotation matrix and $\vec{a}$ is a constant vector. We thus need to relate the displacement field to the strain that is present in the solid.

In the following, we will assume that the solid is relatively stiff, meaning that the partial derivatives of the displacement field are much smaller than one $|\partial_i u_j| \ll 1$. In other words, the difference in the displacement between nearby points in the solid is much smaller than the distance between the points. We define the *linear strain* between two nearby points as

$$\epsilon = \frac{ds' - ds}{ds}, \tag{2.181}$$

where ds is the distance between the points in the unstrained solid and ds' the distance between the points after being displaced. Note that we have here used the symbol ϵ rather than ε for the strain in order to avoid confusion with the permutation symbol when introducing the strain tensor. For the two points $\vec{x}$ and $\vec{x} + \vec{n}\,ds$, where $\vec{n}$ is a unit vector describing the direction of the original spatial separation between the points, we find that the new displacement is given by

$$\vec{x}\,'(\vec{x} + \vec{n}\,ds) - \vec{x}\,'(\vec{x}) \simeq [\vec{n} + (\vec{n} \cdot \nabla)\vec{u}]ds. \tag{2.182}$$

Squaring this relation we find that, to linear order in u,

$$ds'^2 \simeq ds^2[1 + n_i n_j(\partial_j u_i + \partial_i u_j)] \tag{2.183a}$$

and taking the square root of this expression and expanding in the small derivatives gives us

$$ds' \simeq ds\left[1 + n_i n_j \frac{1}{2}(\partial_j u_i + \partial_i u_j)\right]. \tag{2.183b}$$

It therefore follows that the linear strain is given by

$$\epsilon = n_i n_j \frac{1}{2}(\partial_j u_i + \partial_i u_j) = n_i n_j \epsilon_{ij}, \tag{2.184}$$

where we have defined the symmetric *strain tensor*

$$\epsilon_{ij} = \frac{1}{2}(\partial_j u_i + \partial_i u_j). \tag{2.185}$$

The strain tensor can be used to describe the deformation of the solid in many different manners, but we will leave most of this for dedicated texts on solid mechanics and only provide the following example.

Example 2.33 The *volumetric strain* of a small element of a solid with unstrained volume dV can be quantified as

$$\delta = \frac{dV' - dV}{dV}, \tag{2.186}$$

where dV' is the volume of the element when the solid is strained. Taking an original volume spanned by the coordinate basis vectors $d\vec{x}_i = \vec{e}_i ds_i$ (no sum), we find that

$$d\vec{x}\,'_1 = (\vec{e}_1 + \partial_1 \vec{u})ds_1, \tag{2.187}$$

with the equivalent expressions for $d\vec{x}\,'_2$ and $d\vec{x}\,'_3$. The volume dV' is therefore given by

$$dV' = d\vec{x}\,'_1 \cdot (d\vec{x}\,'_2 \times d\vec{x}\,'_3) \simeq (1 + \partial_i u_i)ds_1 ds_2 ds_3 = (1 + \epsilon_{ii})dV. \tag{2.188}$$

Thus, the volumetric strain is given by the trace ϵ_{ii} of the strain tensor.

With the above example in mind, we note that it is always possible to rewrite the strain tensor as

$$\epsilon_{ij} = \frac{1}{3}\theta\delta_{ij} + \underbrace{\left(\epsilon_{ij} - \frac{1}{3}\theta\delta_{ij}\right)}_{\equiv\kappa_{ij}}, \tag{2.189}$$

where θ is an arbitrary scalar field. With the choice of θ equal to the volumetric strain ϵ_{kk}, the tensor κ_{ij} becomes traceless and is known as the *shear strain tensor* (or *deviatoric strain tensor*).

2.6.1.3 *The stiffness and compliance tensors*

For elastic materials, the internal stresses are linearly related to the strains in the material. For our purposes, we will not consider more complicated situations with large strains and inelastic materials and the most general way of linearly relating two rank two tensors is by a rank four tensor. We define this rank four tensor in the relation between the stress and strain tensors to be the *stiffness tensor* $c_{ijk\ell}$. The corresponding relation, *Hooke's law*, is given by

$$\sigma_{ij} = c_{ijk\ell}\epsilon_{k\ell} \tag{2.190}$$

and the symmetry of the stress tensor implies that the stiffness tensor must be symmetric in the first two indices $c_{ijk\ell} = c_{jik\ell} = c_{\{ij\}k\ell}$. In addition, the fact that the strain tensor is symmetric also implies that any anti-symmetric part with respect to the last two indices will be irrelevant for the relation and we might as well take $c_{ijk\ell} = c_{ij\ell k} = c_{ij\{k\ell\}}$. A general rank four tensor would have $3^4 = 81$ independent components. However, the symmetries of the stiffness tensor brings this number down to $6^2 = 36$.

Example 2.34 If a material is well described only by the deformations in one direction and the deformation only depends on that direction, it can be regarded within a one-dimensional setting with a single coordinate x. As a result, the stress and displacement only have one component and can be written as σ and ε, respectively. Hooke's law now takes the form

$$\sigma = c\varepsilon, \tag{2.191}$$

where c is the single component c_{1111} of the stiffness tensor.

By the assumption of linearity, the stiffness tensor itself is not dependent on the actual strains or stresses in the material, only on the properties of the material itself. Depending on the structure of the material, further simplifications may be performed to reduce the number of independent components. The most constraining case is obtained if we assume that the material is isotropic, meaning that the stiffness tensor must have the same components in all frames. We already know that the rank two Kronecker delta tensor is isotropic and thus any outer product of Kronecker deltas will also be an isotropic rank four tensor. The most general linear combination $C_{ijk\ell}$ of such products is

$$C_{ijk\ell} = c_1\delta_{ij}\delta_{k\ell} + c_2\delta_{ik}\delta_{j\ell} + c_3\delta_{i\ell}\delta_{jk}, \tag{2.192}$$

which has three independent parameters. However, only the first of these terms has the symmetry required from the stiffness tensor and the last two terms display this symmetry only if $c_2 = c_3$. It is possible to show that this is the only isotropic tensor of rank four with the required symmetries and we can therefore write the stiffness tensor of an isotropic material on the form

$$c_{ijk\ell} = \lambda\delta_{ij}\delta_{k\ell} + \mu(\delta_{ik}\delta_{j\ell} + \delta_{i\ell}\delta_{jk}). \tag{2.193}$$

The parameters λ and μ are material constants and we note that the stress is now related to the strain according to

$$\sigma_{ij} = \lambda\delta_{ij}\delta_{k\ell}\varepsilon_{k\ell} + \mu(\varepsilon_{ij} + \varepsilon_{ji}) = \lambda\varepsilon_{kk}\delta_{ij} + 2\mu\varepsilon_{ij}. \tag{2.194}$$

Decomposing the strain tensor into volumetric and shear strains, we find that

$$\sigma_{ij} = \left(\lambda + \frac{2\mu}{3}\right)\varepsilon_{kk}\delta_{ij} + 2\mu\kappa_{ij} \equiv K\varepsilon_{kk}\delta_{ij} + 2G\kappa_{ij}. \tag{2.195}$$

In this last expression we have introduced the *bulk modulus* K and the *shear modulus* G. These are both material constants and can be found tabulated in standard reference texts.

The relation in Eq. (2.190) can also be inverted. We can do this by finding a tensor $s_{ijk\ell}$ such that

$$\epsilon_{ij} = s_{ijk\ell}\sigma_{k\ell}. \tag{2.196}$$

This tensor is known as the *compliance tensor*, which tells us how easy it is to deform the material, with larger components indicating larger deformation for the same stress. By the same argumentation as for the stiffness tensor, the symmetries of the stress and strain tensors imply the very same symmetries for the compliance tensor. The usual form of writing the compliance tensor for an isotropic material is

$$s_{ijk\ell} = \frac{1}{E}\left[\frac{1+\nu}{2}(\delta_{ik}\delta_{j\ell} + \delta_{i\ell}\delta_{jk}) - \nu\delta_{ij}\delta_{k\ell}\right] \tag{2.197a}$$

implying that

$$\epsilon_{ij} = \frac{1}{E}\left[(1+\nu)\sigma_{ij} - \nu\sigma_{kk}\delta_{ij}\right]. \tag{2.197b}$$

The material constants involved here are *Young's modulus* E and *Poisson's ratio* ν. Being the inverse relation of that given by the stiffness tensor, these constants are related to the bulk and shear moduli and are also found in standard references. Relating E and ν to K and G is left as Problem 2.41.

2.6.2 Electromagnetism

Electromagnetism is the theory describing the interactions of electric and magnetic fields with matter based upon charges and currents. For the simplest ideas in the theory, it is sufficient to have a description based on vector analysis using the electric field $\vec{E}$ and magnetic field $\vec{B}$. In free space, these fields satisfy *Maxwell's equations*

$$\nabla \cdot \vec{E} = \frac{\rho}{\varepsilon_0}, \tag{2.198a}$$

$$\nabla \cdot \vec{B} = 0, \tag{2.198b}$$

$$\nabla \times \vec{E} + \frac{\partial \vec{B}}{\partial t} = 0, \tag{2.198c}$$

$$\nabla \times \vec{B} - \frac{1}{c^2}\frac{\partial \vec{E}}{\partial t} = \mu_0\vec{J}, \tag{2.198d}$$

;B<C< D A A / ;* E / ; ; !
- , ; ! ? ! + ! , !) ! + !
- # -! ' / / -

Chapter 1

Coordinate Systems and Vectors

Coordinates and vectors—in one form or another—are two of the most fundamental concepts in any discussion of mathematics as applied to physical problems. So, it is beneficial to start our study with these two concepts. Both vectors and coordinates have generalizations that cover a wide variety of physical situations including not only ordinary three-dimensional space with its ordinary vectors, but also the four-dimensional spacetime of relativity with its so-called *four vectors*, and even the infinite-dimensional spaces used in quantum physics with their vectors of infinite components. Our aim in this chapter is to review the *ordinary space* and how it is used to describe physical phenomena. To facilitate this discussion, we first give an outline of some of the properties of vectors.

1.1 Vectors in a Plane and in Space

We start with the most common definition of a vector as a directed line segment without regard to where the vector is located. In other words, a vector is a directed line segment whose only important attributes are its direction and its length. As long as we do not change these two attributes, the vector is not affected. Thus, we are allowed to move a vector parallel to itself without changing the vector. Examples of vectors[1] are position $\mathbf{r}$, displacement $\Delta\mathbf{r}$, velocity $\mathbf{v}$, momentum $\mathbf{p}$, electric field $\mathbf{E}$, and magnetic field $\mathbf{B}$. The vector that has no length is called the **zero vector** and is denoted by $\mathbf{0}$.

general properties of vectors

Vectors would be useless unless we could perform some kind of operation on them. The most basic operation is changing the length of a vector. This is accomplished by multiplying the vector by a real positive number. For example, $3.2\mathbf{r}$ is a vector in the same direction as $\mathbf{r}$ but 3.2 times longer. We

---

[1] Vectors will be denoted by Roman letters printed in boldface type.

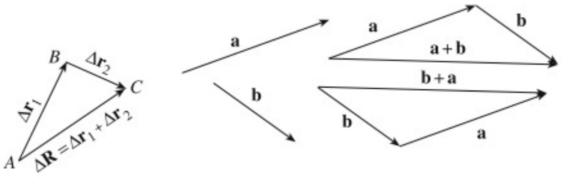

(a) (b)

Figure 1.1: Illustration of the commutative law of addition of vectors.

operations on vectors

can flip the direction of a vector by multiplying it by -1. That is, $(-1)\times\mathbf{r} = -\mathbf{r}$ is a vector having the same length as $\mathbf{r}$ but pointing in the opposite direction. We can combine these two operations and think of multiplying a vector by any real (positive or negative) number. The result is another vector *lying along the same line* as the original vector. Thus, $-0.732\mathbf{r}$ is a vector that is 0.732 times as long as $\mathbf{r}$ and points in the opposite direction. The zero *vector* is obtained every time one multiplies any vector by the *number* zero.

Another operation is the addition of two vectors. This operation, with which we assume the reader to have some familiarity, is inspired by the obvious addition law for displacements. In Figure 1.1(a), a displacement, $\Delta\mathbf{r}_1$ from A to B is added to the displacement $\Delta\mathbf{r}_2$ from B to C to give $\Delta\mathbf{R}$ their resultant, or their sum, i.e., the displacement from A to C: $\Delta\mathbf{r}_1+\Delta\mathbf{r}_2=\Delta\mathbf{R}$. Figure 1.1(b) shows that addition of vectors is commutative: $\mathbf{a}+\mathbf{b}=\mathbf{b}+\mathbf{a}$. It is also associative, $\mathbf{a}+(\mathbf{b}+\mathbf{c})=(\mathbf{a}+\mathbf{b})+\mathbf{c}$, i.e., the order in which you add vectors is irrelevant. It is clear that $\mathbf{a}+\mathbf{0}=\mathbf{0}+\mathbf{a}=\mathbf{a}$ for any vector $\mathbf{a}$.

Example 1.1.1. The parametric equation of a line through two given points can be obtained in vector form by noting that any point in space defines a vector whose components are the coordinates of the given point.[2] If the components of the points P and Q in Figure 1.2 are, respectively, (p_x, p_y, p_z) and (q_x, q_y, q_z), then we can define vectors $\mathbf{p}$ and $\mathbf{q}$ with those components. An arbitrary point X with components (x, y, z) will lie on the line PQ if and only if the vector $\mathbf{x}=(x,y,z)$ has its tip on that line. This will happen if and only if the vector joining P and X, namely $\mathbf{x}-\mathbf{p}$, is proportional to the vector joining P and Q, namely $\mathbf{q}-\mathbf{p}$. Thus, for some real number t, we must have

vector form of the parametric equation of a line

$$\mathbf{x}-\mathbf{p}=t(\mathbf{q}-\mathbf{p}) \qquad \text{or} \qquad \mathbf{x}=t(\mathbf{q}-\mathbf{p})+\mathbf{p}.$$

This is the vector form of the equation of a line. We can write it in component form by noting that the equality of vectors implies the equality of corresponding components. Thus,

$$\begin{aligned} x &= (q_x - p_x)t + p_x, \\ y &= (q_y - p_y)t + p_y, \\ z &= (q_z - p_z)t + p_z, \end{aligned}$$

which is the usual parametric equation for a line. ■

---

[2] We shall discuss components and coordinates in greater detail later in this chapter. For now, the knowledge gained in calculus is sufficient for our discussion.

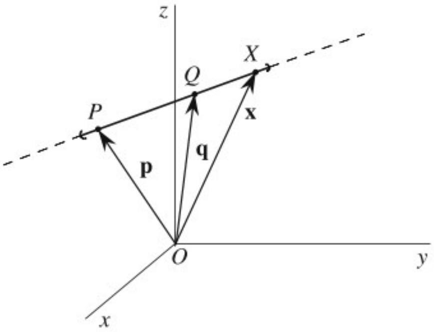

Figure 1.2: The parametric equation of a line in space can be obtained easily using vectors.

There are some special vectors that are extremely useful in describing physical quantities. These are the **unit vectors**. If one divides a vector by its length, one gets a unit vector in the direction of the original vector. Unit vectors are generally denoted by the symbol $\hat{\mathbf{e}}$ with a subscript which designates its direction. Thus, if we divided the vector $\mathbf{a}$ by its length $|\mathbf{a}|$ we get the unit vector $\hat{\mathbf{e}}_a$ in the direction of $\mathbf{a}$. Turning this definition around, we have

use of unit vectors

Box 1.1.1. *If we know the magnitude* $|\mathbf{a}|$ *of a vector quantity as well as its direction* $\hat{\mathbf{e}}_a$*, we can construct the vector:* $\mathbf{a} = |\mathbf{a}|\hat{\mathbf{e}}_a$.

This construction will be used often in the sequel.

The most commonly used unit vectors are those in the direction of coordinate axes. Thus $\hat{\mathbf{e}}_x$, $\hat{\mathbf{e}}_y$, and $\hat{\mathbf{e}}_z$ are the unit vectors pointing in the positive directions of the x-, y-, and z-axes, respectively.[3] We shall introduce unit vectors in other coordinate systems when we discuss those coordinate systems later in this chapter.

unit vectors along the x-, y-, and z-axes

1.1.1 Dot Product

The reader is no doubt familiar with the concept of **dot product** whereby two vectors are "multiplied" and the result is a number. The dot product of $\mathbf{a}$ and $\mathbf{b}$ is defined by

$$\mathbf{a} \cdot \mathbf{b} \equiv |\mathbf{a}|\,|\mathbf{b}|\cos\theta, \tag{1.1}$$

dot product defined

where $|\mathbf{a}|$ is the length of $\mathbf{a}$, $|\mathbf{b}|$ is the length of $\mathbf{b}$, and θ is the angle between the two vectors. This definition is motivated by many physical situations.

---

[3] These unit vectors are usually denoted by **i**, **j**, and **k**, a notation that can be confusing when other non-Cartesian coordinates are used. We shall not use this notation, but adhere to the more suggestive notation introduced above.

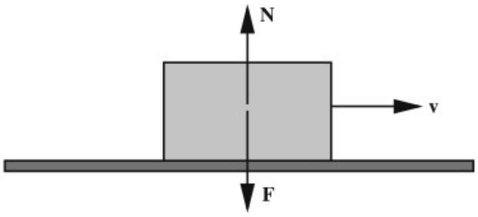

Figure 1.3: No work is done by a force orthogonal to displacement. If such a work were not zero, it would have to be positive or negative; but no consistent rule exists to assign a sign to the work.

The prime example is work which is defined as the scalar product of force and displacement. The presence of $\cos\theta$ ensures the requirement that the work done by a force perpendicular to the displacement is zero. If this requirement were not met, we would have the precarious situation of Figure 1.3 in which the two vertical forces add up to zero but the total work done by them is not zero! This is because it would be impossible to assign a "sign" to the work done by forces being displaced perpendicular to themselves, and make the rule of such an assignment in such a way that the work of **F** in the figure cancels that of **N**. (The reader is urged to try to come up with a rule—e.g., assigning a positive sign to the work if the velocity points to the right of the observer and a negative sign if it points to the observer's left—and see that it will not work, no matter how elaborate it may be!) The only logical definition of work is that which includes a $\cos\theta$ factor.

properties of dot product

The dot product is clearly commutative, $\mathbf{a}\cdot\mathbf{b} = \mathbf{b}\cdot\mathbf{a}$. Moreover, it distributes over vector addition

$$(\mathbf{a}+\mathbf{b})\cdot\mathbf{c} = \mathbf{a}\cdot\mathbf{c} + \mathbf{b}\cdot\mathbf{c}.$$

To see this, note that Equation (1.1) can be interpreted as the product of the length of **a** with the projection of **b** along **a**. Now Figure 1.4 demonstrates[4] that the projection of $\mathbf{a}+\mathbf{b}$ along **c** is the sum of the projections of **a** and **b** along **c** (see Problem 1.2 for details). The third property of the inner product is that $\mathbf{a}\cdot\mathbf{a}$ is always a *positive* number unless **a** is the zero vector in which case $\mathbf{a}\cdot\mathbf{a} = 0$. In mathematics, the collection of these three properties—commutativity, positivity, and distribution over addition—*defines* a dot (or inner) product on a vector space.

properties defining the dot (inner) product

The definition of the dot product leads directly to $\mathbf{a}\cdot\mathbf{a} = |\mathbf{a}|^2$ or

$$|\mathbf{a}| = \sqrt{\mathbf{a}\cdot\mathbf{a}}, \tag{1.2}$$

which is useful in calculating the length of sums or differences of vectors.

---

[4] Figure 1.4 appears to prove the distributive property only for vectors lying in the same plane. However, the argument will be valid even if the three vectors are not coplanar. Instead of dropping perpendicular *lines* from the tips of **a** and **b**, one drops perpendicular *planes*.

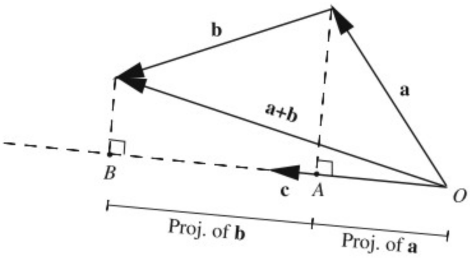

Figure 1.4: The distributive property of the dot product is clearly demonstrated if we interpret the dot product as the length of one vector times the projection of the other vector on the first.

One can use the distributive property of the dot product to show that if (a_x, a_y, a_z) and (b_x, b_y, b_z) represent the components of $\mathbf{a}$ and $\mathbf{b}$ along the axes x, y, and z, then

dot product in terms of components

$$\mathbf{a} \cdot \mathbf{b} = a_x b_x + a_y b_y + a_z b_z. \tag{1.3}$$

From the definition of the dot product, we can draw an important conclusion. If we divide both sides of $\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}|\,|\mathbf{b}| \cos\theta$ by $|\mathbf{a}|$, we get

$$\frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}|} = |\mathbf{b}| \cos\theta \quad \text{or} \quad \left(\frac{\mathbf{a}}{|\mathbf{a}|}\right) \cdot \mathbf{b} = |\mathbf{b}| \cos\theta \;\Rightarrow\; \hat{\mathbf{e}}_a \cdot \mathbf{b} = |\mathbf{b}| \cos\theta.$$

a useful relation to be used frequently in the sequel

Noting that $|\mathbf{b}| \cos\theta$ is simply the projection of $\mathbf{b}$ along $\mathbf{a}$, we conclude

> **Box 1.1.2.** *To find the perpendicular projection of a vector* **b** *along another vector* **a**, *take the dot product of* **b** *with* $\hat{\mathbf{e}}_a$, *the unit vector along* **a**.

Sometimes "component" is used for perpendicular projection. This is not entirely correct. For any set of three *mutually perpendicular* unit vectors in space, Box 1.1.2 can be used to find the *components* of a vector along the three unit vectors. Only if the unit vectors are mutually perpendicular do components and projections coincide.

1.1.2 Vector or Cross Product

Given two space vectors, $\mathbf{a}$ and $\mathbf{b}$, we can find a third *space* vector $\mathbf{c}$, called the **cross product** of $\mathbf{a}$ and $\mathbf{b}$, and denoted by $\mathbf{c} = \mathbf{a} \times \mathbf{b}$. The magnitude of $\mathbf{c}$ is defined by $|\mathbf{c}| = |\mathbf{a}|\,|\mathbf{b}| \sin\theta$ where θ is the angle between $\mathbf{a}$ and $\mathbf{b}$. The direction of $\mathbf{c}$ is given by the **right-hand rule**: If $\mathbf{a}$ is turned to $\mathbf{b}$ (note the order in which $\mathbf{a}$ and $\mathbf{b}$ appear here) through the angle between $\mathbf{a}$ and $\mathbf{b}$,

cross product of two space vectors

right-hand rule explained

a (right-handed) screw that is perpendicular to **a** and **b** will advance in the direction of $\mathbf{a}\times\mathbf{b}$. This definition implies that

$$\mathbf{a}\times\mathbf{b} = -\mathbf{b}\times\mathbf{a}.$$

cross product is antisymmetric

This property is described by saying that the cross product is **antisymmetric**. The definition also implies that

$$\mathbf{a}\cdot(\mathbf{a}\times\mathbf{b}) = \mathbf{b}\cdot(\mathbf{a}\times\mathbf{b}) = 0.$$

That is, $\mathbf{a}\times\mathbf{b}$ is perpendicular to both **a** and **b**.[5]

The vector product has the following properties:

$$\begin{aligned}
&\mathbf{a}\times(\alpha\mathbf{b}) = (\alpha\mathbf{a})\times\mathbf{b} = \alpha(\mathbf{a}\times\mathbf{b}), \qquad && \mathbf{a}\times\mathbf{b} = -\mathbf{b}\times\mathbf{a},\\
&\mathbf{a}\times(\mathbf{b}+\mathbf{c}) = \mathbf{a}\times\mathbf{b}+\mathbf{a}\times\mathbf{c}, && \mathbf{a}\times\mathbf{a} = 0.
\end{aligned} \tag{1.4}$$

cross product in terms of components

Using these properties, we can write the vector product of two vectors in terms of their components. We are interested in a more general result valid in other coordinate systems as well. So, rather than using x, y, and z as subscripts for unit vectors, we use the numbers 1, 2, and 3. In that case, our results can also be used for spherical and cylindrical coordinates which we shall discuss shortly.

$$\begin{aligned}
\mathbf{a}\times\mathbf{b} &= (\alpha_1\hat{\mathbf{e}}_1+\alpha_2\hat{\mathbf{e}}_2+\alpha_3\hat{\mathbf{e}}_3)\times(\beta_1\hat{\mathbf{e}}_1+\beta_2\hat{\mathbf{e}}_2+\beta_3\hat{\mathbf{e}}_3)\\
&= \alpha_1\beta_1\hat{\mathbf{e}}_1\times\hat{\mathbf{e}}_1+\alpha_1\beta_2\hat{\mathbf{e}}_1\times\hat{\mathbf{e}}_2+\alpha_1\beta_3\hat{\mathbf{e}}_1\times\hat{\mathbf{e}}_3\\
&\quad +\alpha_2\beta_1\hat{\mathbf{e}}_2\times\hat{\mathbf{e}}_1+\alpha_2\beta_2\hat{\mathbf{e}}_2\times\hat{\mathbf{e}}_2+\alpha_2\beta_3\hat{\mathbf{e}}_2\times\hat{\mathbf{e}}_3\\
&\quad +\alpha_3\beta_1\hat{\mathbf{e}}_3\times\hat{\mathbf{e}}_1+\alpha_3\beta_2\hat{\mathbf{e}}_3\times\hat{\mathbf{e}}_2+\alpha_3\beta_3\hat{\mathbf{e}}_3\times\hat{\mathbf{e}}_3.
\end{aligned}$$

But, by the last property of Equation (1.4), we have

$$\hat{\mathbf{e}}_1\times\hat{\mathbf{e}}_1 = \hat{\mathbf{e}}_2\times\hat{\mathbf{e}}_2 = \hat{\mathbf{e}}_3\times\hat{\mathbf{e}}_3 = 0.$$

right-handed set of unit vectors

Also, if we assume that $\hat{\mathbf{e}}_1$, $\hat{\mathbf{e}}_2$, and $\hat{\mathbf{e}}_3$ form a so-called **right-handed set**, i.e., if

$$\begin{aligned}
\hat{\mathbf{e}}_1\times\hat{\mathbf{e}}_2 &= -\hat{\mathbf{e}}_2\times\hat{\mathbf{e}}_1 = \hat{\mathbf{e}}_3,\\
\hat{\mathbf{e}}_1\times\hat{\mathbf{e}}_3 &= -\hat{\mathbf{e}}_3\times\hat{\mathbf{e}}_1 = -\hat{\mathbf{e}}_2,\\
\hat{\mathbf{e}}_2\times\hat{\mathbf{e}}_3 &= -\hat{\mathbf{e}}_3\times\hat{\mathbf{e}}_2 = \hat{\mathbf{e}}_1,
\end{aligned} \tag{1.5}$$

then we obtain

$$\mathbf{a}\times\mathbf{b} = (\alpha_2\beta_3-\alpha_3\beta_2)\hat{\mathbf{e}}_1+(\alpha_3\beta_1-\alpha_1\beta_3)\hat{\mathbf{e}}_2+(\alpha_1\beta_2-\alpha_2\beta_1)\hat{\mathbf{e}}_3$$

---

[5] This fact makes it clear why $\mathbf{a}\times\mathbf{b}$ is not defined in the plane. Although it is possible to define $\mathbf{a}\times\mathbf{b}$ for vectors **a** and **b** lying in a plane, $\mathbf{a}\times\mathbf{b}$ *will not lie in that plane* (it will be perpendicular to that plane). For the vector product, **a** and **b** (although lying in a plane) must be considered as space vectors.

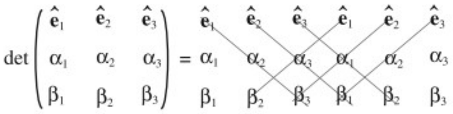

Figure 1.5: A 3×3 determinant is obtained by writing the entries twice as shown, multiplying all terms on each slanted line and adding the results. The lines from upper left to lower right bear a positive sign, and those from upper right to lower left a negative sign.

which can be nicely written in a determinant form[6]

cross product in terms of the determinant of components

$$\mathbf{a} \times \mathbf{b} = \det \begin{pmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3 \\ \alpha_1 & \alpha_2 & \alpha_3 \\ \beta_1 & \beta_2 & \beta_3 \end{pmatrix}. \tag{1.6}$$

Figure 1.5 explains the rule for "expanding" a determinant.

Example 1.1.2. From the definition of the vector product and Figure 1.6(a), we note that

area of a parallelogram in terms of cross product of its two sides

$$|\mathbf{a} \times \mathbf{b}| = \text{area of the parallelogram defined by } \mathbf{a} \text{ and } \mathbf{b}.$$

So we can use Equation (1.6) to find the area of a parallelogram defined by two vectors directly in terms of their components. For instance, the area defined by $\mathbf{a} = (1, 1, -2)$ and $\mathbf{b} = (2, 0, 3)$ can be found by calculating their vector product

$$\mathbf{a} \times \mathbf{b} = \det \begin{pmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3 \\ 1 & 1 & -2 \\ 2 & 0 & 3 \end{pmatrix} = 3\hat{\mathbf{e}}_1 - 7\hat{\mathbf{e}}_2 - 2\hat{\mathbf{e}}_3,$$

and then computing its length

$$|\mathbf{a} \times \mathbf{b}| = \sqrt{3^2 + (-7)^2 + (-2)^2} = \sqrt{62}. \qquad \blacksquare$$

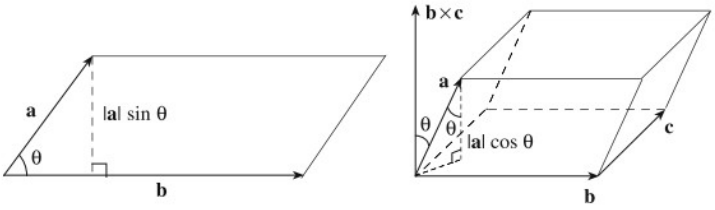

(a) (b)

Figure 1.6: (a) The area of a parallelogram is the absolute value of the cross product of the two vectors describing its sides. (b) The volume of a parallelepiped can be obtained by mixing the dot and the cross products.

---

[6] No knowledge of determinants is necessary at this point. The reader may consider (1.6) to be a mnemonic device useful for remembering the components of $\mathbf{a} \times \mathbf{b}$.

Example 1.1.3. The volume of a parallelepiped defined by three non-coplanar vectors, **a**, **b**, and **c**, is given by $|\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})|$. This can be seen from Figure 1.6(b), where it is clear that

volume of a parallelepiped as a combination of dot and cross products

$$\text{volume} = (\text{area of base})(\text{altitude}) = |\mathbf{b} \times \mathbf{c}|(|\mathbf{a}|\cos\theta) = |(\mathbf{b} \times \mathbf{c}) \cdot \mathbf{a}|.$$

The absolute value is taken to ensure the positivity of the area. In terms of components we have

$$\begin{aligned}\text{volume} &= |(\mathbf{b} \times \mathbf{c})_1\alpha_1 + (\mathbf{b} \times \mathbf{c})_2\alpha_2 + (\mathbf{b} \times \mathbf{c})_3\alpha_3| \\ &= |(\beta_2\gamma_3 - \beta_3\gamma_2)\alpha_1 + (\beta_3\gamma_1 - \beta_1\gamma_3)\alpha_2 + (\beta_1\gamma_2 - \beta_2\gamma_1)\alpha_3|,\end{aligned}$$

volume of a parallelepiped as the determinant of the components of its side vectors

which can be written in determinant form as

$$\text{volume} = |\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})| = \left|\det\begin{pmatrix}\alpha_1 & \alpha_2 & \alpha_3 \\ \beta_1 & \beta_2 & \beta_3 \\ \gamma_1 & \gamma_2 & \gamma_3\end{pmatrix}\right|.$$

Note how we have put the absolute value sign around the determinant of the matrix, so that the area comes out positive. ■

Historical Notes

The concept of vectors as directed line segments that could represent velocities, forces, or accelerations has a very long history. Aristotle knew that the effect of two forces acting on an object could be described by a single force using what is now called *the parallelogram law.* However, the real development of the concept took an unexpected turn in the nineteenth century.

With the advent of *complex numbers* and the realization by Gauss, Wessel, and especially Argand, that they could be represented by points in a plane, mathematicians discovered that complex numbers could be used to study vectors in a plane. A complex number is represented by a pair[7] of real numbers—called the real and imaginary parts of the complex number—which could be considered as the two components of a planar vector.

This connection between vectors in a plane and complex numbers was well established by 1830. Vectors are, however, useful only if they are treated as objects in *space.* After all, velocities, forces, and accelerations are mostly three-dimensional objects. So, the two-dimensional complex numbers had to be generalized to three dimensions. This meant inventing ways of adding, subtracting, multiplying, and dividing objects such as (x, y, z).

William R. Hamilton 1805–1865

The invention of a spatial analogue of the planar complex numbers is due to **William R. Hamilton**. Next to Newton, Hamilton is the greatest of all English mathematicians, and like Newton he was even greater as a physicist than as a mathematician. At the age of five Hamilton could read Latin, Greek, and Hebrew. At eight he added Italian and French; at ten he could read Arabic and Sanskrit, and at fourteen, Persian. A contact with a lightning calculator inspired him to study mathematics. In 1822 at the age of seventeen and a year before he entered Trinity College in Dublin, he prepared a paper on caustics which was read before the Royal Irish Academy in 1824 but not published. Hamilton was advised to rework and expand it. In 1827 he submitted to the Academy a revision which initiated the science of geometrical optics and introduced new techniques in analytical mechanics.

---

[7] See Chapter 18.

In 1827, while still an undergraduate, he was appointed Professor of Astronomy at Trinity College in which capacity he had to manage the astronomical observations and teach science. He did not do much of the former, but he was a fine lecturer.

Hamilton had very good intuition, and knew how to use analogy to reason from the known to the unknown. Although he lacked great flashes of insight, he worked very hard and very long on special problems to see what generalizations they would lead to. He was patient and systematic in working on specific problems and was willing to go through detailed and laborious calculations to check or prove a point.

After mastering and clarifying the concept of complex numbers and their relation to planar vectors (see Problem 18.11 for the connection between complex multiplication on the one hand, and dot and cross products on the other), Hamilton was able to think more clearly about the three-dimensional generalization. His efforts led unfortunately to frustration because the vectors (a) required four components, and (b) defied commutativity! Both features were revolutionary and set the standard for algebra. He called these new numbers **quaternions**.

In retrospect, one can see that the new three-dimensional complex numbers *had* to contain four components. Each "number," when acting on a vector, rotates the latter about an axis and stretches (or contracts) it. Two angles are required to specify the axis of rotation, one angle to specify the amount of rotation, and a fourth number to specify the amount of stretch (or contraction).

Hamilton announced the invention of quaternions in 1843 at a meeting of the Royal Irish Academy, and spent the rest of his life developing the subject.

1.2 Coordinate Systems

Coordinates are "functions" that specify **points** of a space. The smallest number of these functions necessary to specify a point is called the **dimension** of that space. For instance, a point of a plane is specified by two numbers, and as the point moves in the plane the two numbers change, i.e., the coordinates are functions of the position of the point. If we designate the point as P, we may write the coordinate functions of P as $(f(P), g(P))$.[8] Each pair of such functions is called a **coordinate system**.

coordinate systems as functions.

There are two coordinate systems used for a plane, **Cartesian**, denoted by $(x(P), y(P))$, and **polar**, denoted by $(r(P), \theta(P))$. As shown in Figure 1.7,

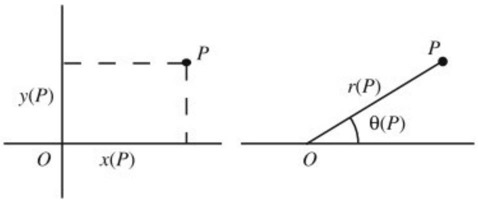

Figure 1.7: Cartesian and polar coordinates of a point P in two dimensions.

---

[8] Think of f (or g) as a rule by which a unique number is assigned to each point P.

the "function" x is defined as giving the distance from P to the vertical axis, while θ is the function which gives the angle that the line OP makes with a given fiducial (usually horizontal) line. The origin O and the fiducial line are completely arbitrary. Similarly, the functions r and y give distances from the origin and to the horizontal axis, respectively.

Box 1.2.1. *In practice, one drops the argument P and writes (x, y) and (r, θ).*

the three common coordinate systems: Cartesian, cylindrical and spherical

We can generalize the above concepts to three dimensions. There are three coordinate functions now. So for a point P in space we write

$$(f(P), g(P), h(P)),$$

where f, g, and h are functions on the three-dimensional space. There are three widely used coordinate systems, **Cartesian** $(x(P), y(P), z(P))$, **cylindrical** $(\rho(P), \varphi(P), z(P))$, and **spherical** $(r(P), \theta(P), \varphi(P))$. $\varphi(P)$ is called the **azimuth** or the **azimuthal angle** of P, while $\theta(P)$ is called its **polar angle**. To find the spherical coordinates of P, one chooses an arbitrary point as the origin O and an arbitrary line through O called the **polar axis**. One measures $\overline{OP}$ and calls it $r(P)$; $\theta(P)$ is the angle between $\overline{OP}$ and the polar axis. To find the third coordinate, we construct the plane through O and perpendicular to the polar axis, drop a projection from P to the plane meeting the latter at H, draw an arbitrary fiducial line through O in this plane, and measure the angle between this line and $\overline{OH}$. This angle is $\varphi(P)$. Cartesian and cylindrical coordinate systems can be described similarly. The three coordinate systems are shown in Figure 1.8. As indicated in the figure, the polar axis is usually taken to be the z-axis, and the fiducial line from which $\varphi(P)$ is measured is chosen to be the x-axis. Although there are other coordinate systems, the three mentioned above are by far the most widely used.

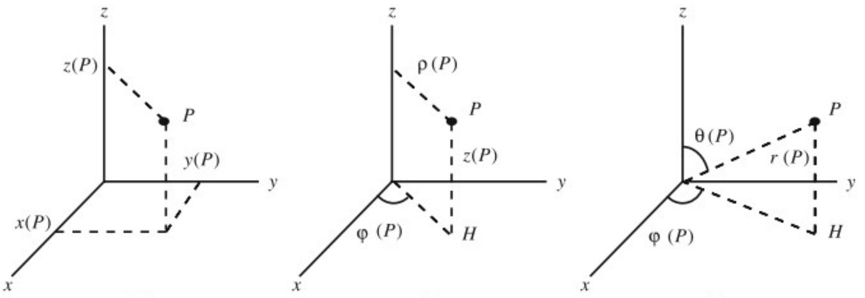

(a) (b) (c)

Figure 1.8: (a) Cartesian, (b) cylindrical, and (c) spherical coordinates of a point P in three dimensions.

Which one of the three systems of coordinates to use in a given physical problem is dictated mainly by the geometry of that problem. As a rule, spherical coordinates are best suited for spheres and spherically symmetric problems. Spherical symmetry describes situations in which quantities of interest are functions only of the distance from a fixed point and not on the orientation of that distance. Similarly, cylindrical coordinates ease calculations when cylinders or cylindrical symmetries are involved. Finally, Cartesian coordinates are used in rectangular geometries.

limitations of non-Cartesian coordinates

Of the three coordinate systems, Cartesian is the most complete in the following sense: *A point in space can have only* one triplet *as its coordinates.* This property is not shared by the other two systems. For example, a point P located on the z-axis of a cylindrical coordinate system does not have a well-defined $\varphi(P)$. In practice, such imperfections are not of dire consequence and we shall ignore them.

transformation from spherical to cylindrical coordinates

Once we have three coordinate systems to work with, we need to know how to translate from one to another. First we give the transformation rule from spherical to cylindrical. It is clear from Figure 1.9 that

$$\rho = r\sin\theta, \quad \varphi_{\text{cyl}} = \varphi_{\text{sph}}, \quad z = r\cos\theta. \tag{1.7}$$

Thus, given (r, θ, φ) of a point P, we can obtain (ρ, φ, z) of the same point by substituting in the RHS.

transformation from cylindrical to Cartesian coordinates

Next we give the transformation rule from cylindrical to Cartesian. Again Figure 1.9 gives the result:

$$x = \rho\cos\varphi, \quad y = \rho\sin\varphi, \quad z_{\text{car}} = z_{\text{cyl}}. \tag{1.8}$$

transformation from spherical to Cartesian coordinates

We can combine (1.7) and (1.8) to connect Cartesian and spherical coordinates:

$$x = r\sin\theta\cos\varphi, \quad y = r\sin\theta\sin\varphi, \quad z = r\cos\theta. \tag{1.9}$$

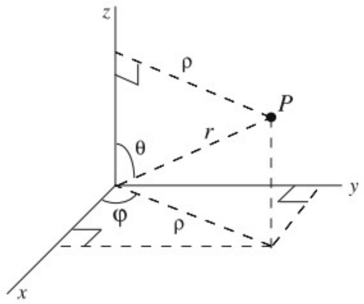

Figure 1.9: The relation between the cylindrical and spherical coordinates of a point P can be obtained using this diagram.

Box 1.2.2. *Equations (1.7)–(1.9) are extremely important and worth being committed to memory. The reader is advised to study Figure 1.9 carefully and learn to reproduce (1.7)–(1.9) from the figure!*

The transformations given are in their standard form. We can turn them around and give the inverse transformations. For instance, squaring the first and third equations of (1.7) and adding gives $\rho^2 + z^2 = r^2$ or $r = \sqrt{\rho^2 + z^2}$. Similarly, dividing the first and third equation yields $\tan\theta = \rho/z$, which implies that $\theta = \tan^{-1}(\rho/z)$, or equivalently,

$$\frac{z}{r} = \cos\theta \;\Rightarrow\; \theta = \cos^{-1}\left(\frac{z}{r}\right) = \cos^{-1}\left(\frac{z}{\sqrt{\rho^2 + z^2}}\right).$$

transformation from cylindrical to spherical coordinates

Thus, the inverse of (1.7) is

$$r = \sqrt{\rho^2 + z^2}, \quad \theta = \tan^{-1}\left(\frac{\rho}{z}\right) = \cos^{-1}\left(\frac{z}{\sqrt{\rho^2 + z^2}}\right), \quad \varphi_{\text{sph}} = \varphi_{\text{cyl}}. \tag{1.10}$$

Similarly, the inverse of (1.8) is

$$\begin{aligned} \rho &= \sqrt{x^2 + y^2}, \\ \varphi &= \tan^{-1}\left(\frac{y}{x}\right) = \cos^{-1}\left(\frac{x}{\sqrt{x^2 + y^2}}\right) = \sin^{-1}\left(\frac{y}{\sqrt{x^2 + y^2}}\right), \\ z_{\text{cyl}} &= z_{\text{car}}, \end{aligned} \tag{1.11}$$

transformation from Cartesian to spherical coordinates

and that of (1.9) is

$$\begin{aligned} r &= \sqrt{x^2 + y^2 + z^2}, \\ \theta &= \tan^{-1}\left(\frac{\sqrt{x^2 + y^2}}{z}\right) = \cos^{-1}\left(\frac{z}{\sqrt{x^2 + y^2 + z^2}}\right) \\ &= \sin^{-1}\left(\frac{\sqrt{x^2 + y^2}}{\sqrt{x^2 + y^2 + z^2}}\right), \\ \varphi &= \tan^{-1}\left(\frac{y}{x}\right) = \cos^{-1}\left(\frac{x}{\sqrt{x^2 + y^2}}\right) = \sin^{-1}\left(\frac{y}{\sqrt{x^2 + y^2}}\right). \end{aligned} \tag{1.12}$$

An important question concerns the *range* of these quantities. In other words: *In what range should we allow these quantities to vary in order to cover the whole space?* For Cartesian coordinates all three variables vary between $-\infty$ and $+\infty$. Thus,

range of coordinate variables

$$-\infty < x < +\infty, \quad -\infty < y < +\infty, \quad -\infty < z < +\infty.$$

The ranges of cylindrical coordinates are

$$0 \le \rho < \infty, \quad 0 \le \varphi \le 2\pi, \quad -\infty < z < \infty.$$

Note that ρ, being a distance, cannot have negative values.[9] Similarly, the ranges of spherical coordinates are

$$0 \leq r < \infty, \quad 0 \leq \theta \leq \pi, \quad 0 \leq \varphi \leq 2\pi.$$

Again, r is never negative for similar reasons as above. Also note that the range of θ excludes values larger than π. This is because the range of φ takes care of points where θ "appears" to have been increased by π.

Historical Notes

One of the greatest achievements in the development of mathematics since Euclid was the introduction of coordinates. Two men take credit for this development: Fermat and Descartes. These two great French mathematicians were interested in the unification of geometry and algebra, which resulted in the creation of a most fruitful branch of mathematics now called **analytic geometry**. Fermat and Descartes who were heavily involved in physics, were keenly aware of both the need for quantitative methods and the capacity of algebra to deliver that method.

Pierre de Fermat
1601–1665

Fermat's interest in the unification of geometry and algebra arose because of his involvement in optics. His interest in the attainment of maxima and minima—thus his contribution to calculus—stemmed from the investigation of the passage of light rays through media of different indices of refraction, which resulted in *Fermat's principle* in optics and the law of refraction. With the introduction of coordinates, Fermat was able to quantify the study of optics and set a trend to which all physicists of posterity would adhere. It is safe to say that without analytic geometry the progress of science, and in particular physics, would have been next to impossible.

Born into a family of tradespeople, **Pierre de Fermat** was trained as a lawyer and made his living in this profession becoming a councillor of the parliament of the city of Toulouse. Although mathematics was but a hobby for him and he could devote only spare time to it, he made great contributions to number theory, to calculus, and, together with Pascal, initiated work on probability theory.

The coordinate system introduced by Fermat was not a convenient one. For one thing, the coordinate axes were not at right angles to one another. Furthermore, the use of negative coordinates was not considered. Nevertheless, he was able to translate geometric curves into algebraic equations.

René Descartes was a great philosopher, a founder of modern biology, and a superb physicist and mathematician. His interest in mathematics stemmed from his desire to understand nature. He wrote:

> ...I have resolved to quit only abstract geometry, that is to say, the consideration of questions which *serve only to exercise the mind*, and this, in order to study another kind of geometry, which has for its object the explanation of the phenomena of nature.

René Descartes
1596–1650

His father, a relatively wealthy lawyer, sent him to a Jesuit school at the age of eight where, due to his delicate health, he was allowed to spend the mornings in bed, during which time he worked. He followed this habit during his entire life. At twenty he graduated from the University of Poitier as a lawyer and went to Paris where he studied mathematics with a Jesuit priest. After one year he decided to join the army of Prince Maurice of Orange in 1617. During the next nine years he vacillated between various armies while studying mathematics.

---

[9] In some calculus books ρ is allowed to have negative values to account for points on the opposite side of the origin. However, in physics literature ρ is assumed to be positive.To go to "the other side" of the origin along ρ, we change φ by π, keeping ρ positive at all times.

He eventually returned to Paris, where he devoted his efforts to the study of optical instruments motivated by the newly discovered power of the telescope. In 1628 he moved to Holland to a quieter and freer intellectual environment. There he lived for the next twenty years and wrote his famous works. In 1649 Queen Christina of Sweden persuaded Descartes to go to Stockholm as her private tutor. However the Queen had an uncompromising desire to draw curves and tangents at 5 a.m., causing Descartes to break the lifelong habit of getting up at 11 o'clock! After only a few months in the cold northern climate, walking to the palace for the 5 o'clock appointment with the queen, he died of pneumonia in 1650.

Descartes described his algebraic approach to geometry in his monumental work *La Géométrie*. It is in this work that he solves geometrical problems using algebra by introducing coordinates. These coordinates, as in Fermat's case, were not lengths along perpendicular axes. Nevertheless they paved the way for the later generations of scientists such as Newton to build on Descartes' and Fermat's ideas and improve on them.

Newton uses polar coordinates for the first time

Throughout the seventeenth century, mathematicians used one axis with the y values drawn at an oblique or right angle onto that axis. Newton, however, in a book called *The Method of Fluxions and Infinite Series* written in 1671, and translated much later into English in 1736, describes a coordinate system in which points are located in reference to a fixed point and a fixed line through that point. This was the first introduction of essentially the *polar coordinates* we use today.

1.3 Vectors in Different Coordinate Systems

Many physical situations require the study of vectors in different coordinate systems. For example, the study of the solar system is best done in spherical coordinates because of the nature of the gravitational force. Similarly calculation of electromagnetic fields in a cylindrical cavity will be easier if we use cylindrical coordinates. This requires not only writing functions in terms of these coordinate variables, but also expressing vectors in terms of unit vectors suitable for these coordinate systems. It turns out that, for the three coordinate systems described above, the most natural construction of such vectors renders them mutually perpendicular.

orthonormal basis

Any set of three (two) mutually perpendicular unit vectors in space (in the plane) is called an **orthonormal basis**.[10] Basis vectors have the property that any vector can be written in terms of them.

Let us start with the plane in which the coordinate system could be Cartesian or polar. In general, we construct an orthonormal basis *at a point* and note that

---

[10] The word "orthonormal" comes from *orthogonal* meaning "perpendicular," and *normal* meaning "of unit length."

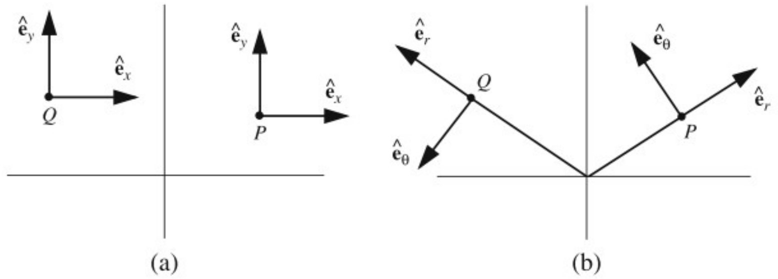

Figure 1.10: The unit vectors in (a) Cartesian coordinates and (b) polar coordinates. The unit vectors at P and Q are the same for Cartesian coordinates, but different in polar coordinates.

Box 1.3.1. *The orthonormal basis, generally speaking, depends on the point at which it is constructed.*

The vectors of a basis are constructed as follows. To find the unit vector corresponding to a coordinate at a point P, hold the other coordinate fixed and *increase* the coordinate in question. The initial direction of motion of P is the direction of the unit vector sought. Thus, we obtain the Cartesian unit vectors at point P of Figure 1.10(a): $\hat{\mathbf{e}}_x$ is obtained by holding y fixed and letting x vary in the increasing direction; and $\hat{\mathbf{e}}_y$ is obtained by holding x fixed at P and letting y increase. In each case, the unit vectors show the initial direction of the motion of P. It should be clear that one obtains the same set of unit vectors regardless of the location of P. However, the reader should take note that this is true only for coordinates that are defined in terms of axes whose directions are fixed, such as Cartesian coordinates.

general rule for constructing a basis at a point

If we use polar coordinates for P, then holding θ fixed at P gives the direction of $\hat{\mathbf{e}}_r$ as shown in Figure 1.10(b), because for fixed θ, that is the direction of increase for r. Similarly, if r is fixed at P, the initial direction of motion of P when θ is increased is that of $\hat{\mathbf{e}}_\theta$ shown in the figure. If we choose another point such as Q shown in the figure, then a new set of unit vectors will be obtained which are different form those of P. This is because polar coordinates are not defined in terms of any fixed axes.

Since $\{\hat{\mathbf{e}}_x, \hat{\mathbf{e}}_y\}$ and $\{\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta\}$ form a basis in the plane, any vector $\mathbf{a}$ in the plane can be expressed in terms of either basis as shown in Figure 1.11. Thus, we can write

$$\mathbf{a} = a_{xP}\hat{\mathbf{e}}_{xP} + a_{yP}\hat{\mathbf{e}}_{yP} = a_{rP}\hat{\mathbf{e}}_{rP} + a_{\theta P}\hat{\mathbf{e}}_{\theta P} = a_{rQ}\hat{\mathbf{e}}_{rQ} + a_{\theta Q}\hat{\mathbf{e}}_{\theta Q}, \qquad (1.13)$$

where the coordinates are subscripted to emphasize their dependence on the points at which the unit vectors are erected. In the case of Cartesian coordinates, this, of course, is not necessary because the unit vectors happen to be independent of the point. In the case of polar coordinates, although this

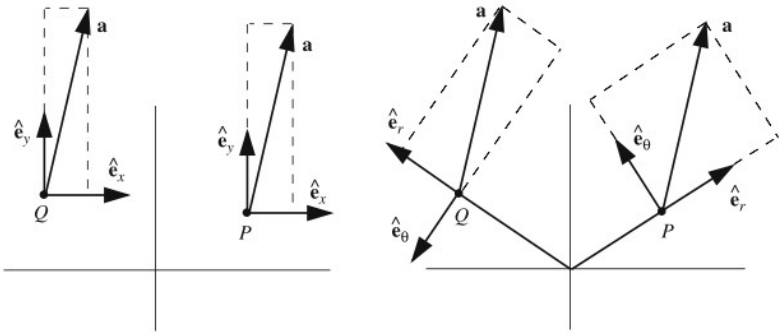

(a) (b)

Figure 1.11: (a) The vector **a** has the same components along unit vectors at P and Q in Cartesian coordinates. (b) The vector **a** has different components along unit vectors at different points for a polar coordinate system.

dependence exists, we normally do not write the points as subscripts, being aware of this dependence every time we use polar coordinates.

angle brackets denote vector components

So far we have used parentheses to designate the (components of) a vector. Since, parentheses—as a universal notation—are used for coordinates of points, we shall write components of a vector in angle brackets. So Equation (1.13) can also be written as

$$\mathbf{a} = \langle a_x, a_y \rangle_P = \langle a_r, a_\theta \rangle_P = \langle a_r, a_\theta \rangle_Q,$$

where again the subscript indicating the point at which the unit vectors are defined is normally deleted. However, we need to keep in mind that although $\langle a_x, a_y \rangle$ is independent of the point in question, $\langle a_r, a_\theta \rangle$ is very much point-dependent. Caution should be exercised when using this notation as to the location of the unit vectors.

The unit vectors in the coordinate systems of space are defined the same way. We follow the rule given before:

Box 1.3.2. (***Rule for Finding Coordinate Unit Vectors***). *To find the unit vector corresponding to a coordinate at a point P, hold the other coordinates fixed and* increase *the coordinate in question. The initial direction of motion of P is the direction of the unit vector sought.*

It should be clear that the Cartesian basis $\{\hat{\mathbf{e}}_x, \hat{\mathbf{e}}_y, \hat{\mathbf{e}}_z\}$ is the same for all points, and usually they are drawn at the origin along the three axes. An arbitrary vector **a** can be written as

$$\mathbf{a} = a_x\hat{\mathbf{e}}_x + a_y\hat{\mathbf{e}}_y + a_z\hat{\mathbf{e}}_z \qquad \text{or} \qquad \mathbf{a} = \langle a_x, a_y, a_z \rangle, \tag{1.14}$$

where we used angle brackets to denote *components* of the vector, reserving the parentheses for *coordinates* of points in space.

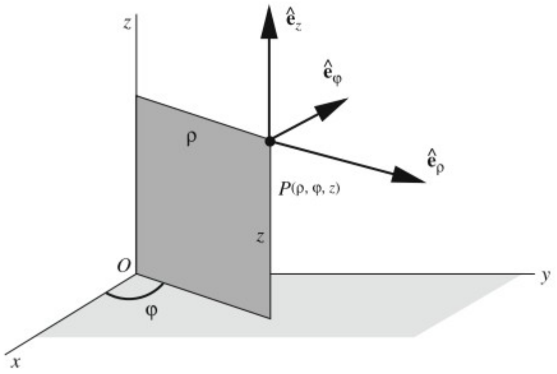

Figure 1.12: Unit vectors of cylindrical coordinates.

The unit vectors at a point P in the other coordinate systems are obtained similarly. In cylindrical coordinates, $\hat{\mathbf{e}}_\rho$ lies along and points in the direction of increasing ρ at P; $\hat{\mathbf{e}}_\varphi$ is perpendicular to the plane formed by P and the z-axis and points in the direction of increasing φ; $\hat{\mathbf{e}}_z$ points in the direction of positive z (see Figure 1.12). We note that only $\hat{\mathbf{e}}_z$ is independent of the point at which the unit vectors are defined because z is a fixed axis in cylindrical coordinates. Given any vector $\mathbf{a}$, we can write it as

$$\mathbf{a} = a_\rho \hat{\mathbf{e}}_\rho + a_\varphi \hat{\mathbf{e}}_\varphi + a_z \hat{\mathbf{e}}_z \qquad \text{or} \qquad \mathbf{a} = \langle a_\rho, a_\varphi, a_z \rangle. \tag{1.15}$$

radial direction

The unit vectors in spherical coordinates are defined similarly: $\hat{\mathbf{e}}_r$ is taken along r and points in the direction of increasing r; this direction is called **radial**; $\hat{\mathbf{e}}_\theta$ is taken to lie in the plane formed by P and the z-axis, is perpendicular to r, and points in the direction of increasing θ; $\hat{\mathbf{e}}_\varphi$ is as in the cylindrical case (Figure 1.13). An arbitrary vector in space can be expressed in terms of the spherical unit vectors at P:

$$\mathbf{a} = a_r \hat{\mathbf{e}}_r + a_\theta \hat{\mathbf{e}}_\theta + a_\varphi \hat{\mathbf{e}}_\varphi \qquad \text{or} \qquad \mathbf{a} = \langle a_r, a_\theta, a_\varphi \rangle. \tag{1.16}$$

It should be emphasized that

Box 1.3.3. *The cylindrical and spherical unit vectors* $\hat{\mathbf{e}}_\rho$, $\hat{\mathbf{e}}_r$, $\hat{\mathbf{e}}_\theta$, *and* $\hat{\mathbf{e}}_\varphi$ *are dependent on the position of* P.

position vector

Once an origin O is designated, every point P in space will define a vector, called a **position vector** and denoted by $\mathbf{r}$. This is simply the vector drawn from O to P. In Cartesian coordinates this vector has *components* $\langle x, y, z \rangle$, thus one can write

$$\mathbf{r} = x\hat{\mathbf{e}}_x + y\hat{\mathbf{e}}_y + z\hat{\mathbf{e}}_z. \tag{1.17}$$

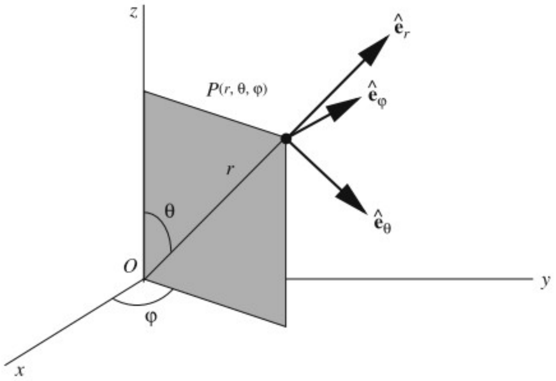

Figure 1.13: Unit vectors of spherical coordinates. Note that the intersection of the shaded plane with the xy-plane is a line along the cylindrical coordinate ρ.

difference between coordinates and components explained

But (x, y, z) are also the *coordinates* of the point P. This can be a source of confusion when other coordinate systems are used. For example, in spherical coordinates, the *components* of the vector $\mathbf{r}$ at P are $\langle r, 0, 0\rangle$ because $\mathbf{r}$ has only a component along $\hat{\mathbf{e}}_r$ and none along $\hat{\mathbf{e}}_\theta$ or $\hat{\mathbf{e}}_\varphi$. One writes[11]

$$\mathbf{r} = r\hat{\mathbf{e}}_r. \tag{1.18}$$

However, the *coordinates* of P are still (r, θ, φ)! Similarly, the coordinates of P are (ρ, φ, z) in a cylindrical system, while

$$\mathbf{r} = \rho\,\hat{\mathbf{e}}_\rho + z\hat{\mathbf{e}}_z, \tag{1.19}$$

because $\mathbf{r}$ lies in the ρz-plane and has no component along $\hat{\mathbf{e}}_\varphi$. Therefore,

Box 1.3.4. *Make a clear distinction between the **components** of the vector* $\mathbf{r}$ *and the **coordinates** of the point* P.

A common symptom of confusing components with coordinates is as follows. Point P_1 has position vector $\mathbf{r}_1$ with spherical components $\langle r_1, 0, 0\rangle$ at P_1. The position vector of a second point P_2 is $\mathbf{r}_2$ with spherical components $\langle r_2, 0, 0\rangle$ at P_2. It is easy to fall into the trap of thinking that $\mathbf{r}_1 - \mathbf{r}_2$ has spherical components $\langle r_1 - r_2, 0, 0\rangle$! This is, of course, not true, because the spherical unit vectors at P_1 are completely different from those at P_2, and, therefore, contrary to the Cartesian case, we cannot simply subtract components.

---

[11] We should really label everything with P. But, as usual, we assume this labeling to be implied.

One of the great advantages of vectors is their ability to express results independent of any specific coordinate systems. Physical laws are always coordinate-independent. For example, when we write $\mathbf{F} = m\mathbf{a}$ both $\mathbf{F}$ and $\mathbf{a}$ could be expressed in terms of Cartesian, spherical, cylindrical, or any other convenient coordinate system. This independence allows us the freedom to choose the coordinate systems most convenient for the problem at hand. For example, it is extremely difficult to solve the planetary motions in Cartesian coordinates, while the use of spherical coordinates facilitates the solution of the problem tremendously.

Physical laws ought to be coordinate independent!

Example 1.3.1. We can express the *coordinates* of the **center of mass** (CM) of a collection of particles in terms of their position vectors.[12] Thus, if $\overline{\mathbf{r}}$ denotes the position vector of the CM of the collection of N mass points, $m_1, m_2, \ldots, m_N$ with respective position vectors $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N$ relative to an origin O, then[13]

center of mass

$$\overline{\mathbf{r}} = \frac{m_1\mathbf{r}_1 + m_2\mathbf{r}_2 + \cdots + m_N\mathbf{r}_N}{m_1 + m_2 + \cdots + m_N} = \frac{\sum_{k=1}^{N} m_k\mathbf{r}_k}{M}, \tag{1.20}$$

where $M = \sum_{k=1}^{N} m_k$ is the total mass of the system. One can also think of Equation (1.20) as a *vector* equation. To find the component equations in a coordinate system, one needs to pick a *fixed* point (say the origin), a set of unit vectors at that point (usually the unit vectors along the axes of some coordinate system), and substitute the *components* of $\mathbf{r}_k$ along those unit vectors to find the components of $\overline{\mathbf{r}}$ along the unit vectors. ■

1.3.1 Fields and Potentials

The distributive property of the dot product and the fact that the unit vectors of the bases in all coordinate systems are mutually perpendicular can be used to derive the following:

dot product in terms of components in the three coordinate systems

$$\begin{aligned}
\mathbf{a} \cdot \mathbf{b} &= a_x b_x + a_y b_y + a_z b_z \quad \text{(Cartesian)},\\
\mathbf{a} \cdot \mathbf{b} &= a_\rho b_\rho + a_\varphi b_\varphi + a_z b_z \quad \text{(cylindrical)},\\
\mathbf{a} \cdot \mathbf{b} &= a_r b_r + a_\theta b_\theta + a_\varphi b_\varphi \quad \text{(spherical)}.
\end{aligned} \tag{1.21}$$

The first of these equations is the same as (1.3).

It is important to keep in mind that the components are to be expressed in the same set of unit vectors. This typically means setting up mutually perpendicular unit vectors (an orthonormal basis) *at a single point* and resolving all vectors along those unit vectors.

The dot product, in various forms and guises, has many applications in physics. As pointed out earlier, it was introduced in the definition of work, but soon spread to many other concepts of physics. One of the simplest—and most important—applications is its use in writing the laws of physics in a coordinate-independent way.

---

[12] This implies that the equation is most useful only when Cartesian coordinates are used, because only for these coordinates do the components of the position vector of a point coincide with the coordinates of that point.

[13] We assume that the reader is familiar with the symbol $\sum$ simply as a summation symbol. We shall discuss its properties and ways of manipulating it in Chapter 9.

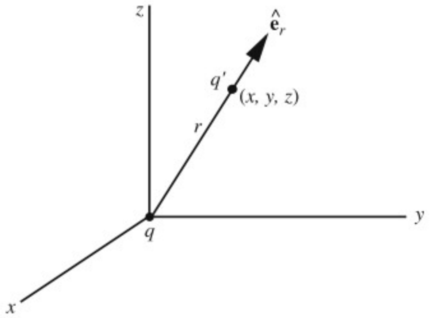

Figure 1.14: The diagram illustrating the electrical force when one charge is at the origin.

Example 1.3.2. A point charge q is situated at the origin. A second charge q' is located at (x, y, z) as shown in Figure 1.14. We want to express the electric force on q' in Cartesian, spherical, and cylindrical coordinate systems.

Coulomb's law

We know that the electric force, as given by **Coulomb's law**, lies along the line joining the two charges and is either attractive or repulsive according to the signs of q and q'. All of this information can be summarized in the formula

$$\mathbf{F}_{q'} = \frac{k_e q q'}{r^2}\hat{\mathbf{e}}_r \tag{1.22}$$

where $k_e = 1/(4\pi\epsilon_0) \approx 9 \times 10^9$ in SI units. Note that if q and q' are unlike, $qq' < 0$ and $\mathbf{F}_{q'}$ is opposite to $\hat{\mathbf{e}}_r$, i.e., it is *attractive*. On the other hand, if q and q' are of the same sign, $qq' > 0$ and $\mathbf{F}_{q'}$ is in the same direction as $\hat{\mathbf{e}}_r$, i.e., *repulsive*.

Equation (1.22) expresses $\mathbf{F}_{q'}$ in spherical coordinates. Thus, its components in terms of unit vectors at q' are $\langle k_e qq'/r^2, 0, 0\rangle$. To get the components in the other coordinate systems, we rewrite (1.22). Noting that $\hat{\mathbf{e}}_r = \mathbf{r}/r$, we write

$$\mathbf{F}_{q'} = \frac{k_e q q'}{r^2}\frac{\mathbf{r}}{r} = \frac{k_e q q'}{r^3}\mathbf{r}. \tag{1.23}$$

For Cartesian coordinates we use (1.12) to obtain $r^3 = (x^2+y^2+z^2)^{3/2}$. Substituting this and (1.17) in (1.23) yields

$$\mathbf{F}_{q'} = \frac{k_e q q'}{(x^2+y^2+z^2)^{3/2}}(x\hat{\mathbf{e}}_x + y\hat{\mathbf{e}}_y + z\hat{\mathbf{e}}_z).$$

Therefore, the components of $\mathbf{F}_{q'}$ in Cartesian coordinates are

$$\left\langle \frac{k_e q q' x}{(x^2+y^2+z^2)^{3/2}}, \frac{k_e q q' y}{(x^2+y^2+z^2)^{3/2}}, \frac{k_e q q' z}{(x^2+y^2+z^2)^{3/2}} \right\rangle.$$

Finally, using (1.10) and (1.19) in (1.23), we obtain

$$\mathbf{F}_{q'} = \frac{k_e q q'}{(\rho^2+z^2)^{3/2}}(\rho\,\hat{\mathbf{e}}_\rho + z\hat{\mathbf{e}}_z).$$

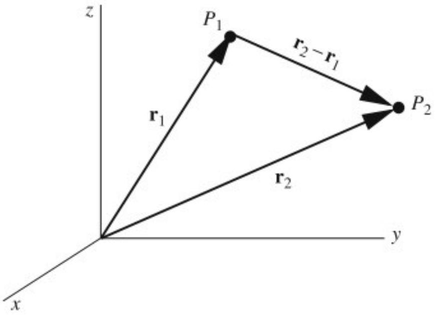

Figure 1.15: The displacement vector between P_1 and P_2 is the difference between their position vectors.

Thus the components of $\mathbf{F}_{q'}$ along the cylindrical unit vectors constructed at the location of q' are

$$\left\langle \frac{k_e q q' \rho}{(\rho^2 + z^2)^{3/2}}, 0, \frac{k_e q q' z}{(\rho^2 + z^2)^{3/2}} \right\rangle . \qquad \blacksquare$$

Since $\mathbf{r}$ gives the position of a point in space, one can use it to write the distance between two points P_1 and P_2 with position vectors $\mathbf{r}_1$ and $\mathbf{r}_2$. Figure 1.15 shows that $\mathbf{r}_2 - \mathbf{r}_1$ is the displacement vector from P_1 to P_2. The importance of this vector stems from the fact that many physical quantities are functions of distances between point particles, and $\mathbf{r}_2 - \mathbf{r}_1$ is a concise way of expressing this distance. The following example illustrates this.

Historical Notes

During the second half of the eighteenth century many physicists were engaged in a quantitative study of electricity and magnetism. **Charles Augustin de Coulomb**, who developed the so-called *torsion balance* for measuring weak forces, is credited with the discovery of the law governing the force between electrical charges.

Charles Coulomb
1736–1806

Coulomb was an army engineer in the West Indies. After spending nine years there, due to his poor health, he returned to France about the same time that the French Revolution began, at which time he retired to the country to do scientific research.

Beside his experiments on electricity, Coulomb worked on applied mechanics, structural analysis, the fracture of beams and columns, the thrust of arches, and the thrust of the soil.

Henry Cavendish
1731–1810

At about the same time that Coulomb discovered the law of electricity, there lived in England a very reclusive character named **Henry Cavendish**. He was born into the nobility, had no close friends, was afraid of women, and disinterested in music or arts of any kind. His life revolved around experiments in physics and chemistry that he carried out in a private laboratory located in his large mansion.

During his long life he published only a handful of relatively unimportant papers. But after his death about one million pounds sterling were found in his bank account and twenty bundles of notes in his laboratory. These notes remained in the possession of his relatives for a long time, but when they were published one hundred years later, it became clear that Henry Cavendish was one of the greatest experimental physicists ever. He discovered all the laws of electric and magnetic interactions at the same time as Coulomb, and his work in chemistry matches that of Lavoisier. Furthermore, he used a torsion balance to measure the universal gravitational constant for the first time, and as a result was able to arrive at the exact mass of the Earth.

Example 1.3.3. COULOMB'S LAW FOR TWO ARBITRARY CHARGES
Suppose there are point charges q_1 at P_1 and q_2 at P_2. Let us write the force exerted on q_2 by q_1. The magnitude of the force is

$$F_{21} = \frac{k_e q_1 q_2}{d^2},$$

where $d = \overline{P_1 P_2}$ is the distance between the two charges. We use d because the usual notation r has special meaning for us: it is one of the coordinates in spherical systems. If we multiply this magnitude by the unit vector describing the direction of the force, we obtain the full force vector (see Box 1.1.1). But, assuming repulsion for the moment, this unit vector is

$$\frac{\mathbf{r}_2 - \mathbf{r}_1}{|\mathbf{r}_2 - \mathbf{r}_1|} \equiv \hat{\mathbf{e}}_{21}.$$

Also, since $d = |\mathbf{r}_2 - \mathbf{r}_1|$, we have

$$\mathbf{F}_{21} = \frac{k_e q_1 q_2}{d^2}\hat{\mathbf{e}}_{21} = \frac{k_e q_1 q_2}{|\mathbf{r}_2 - \mathbf{r}_1|^2}\frac{\mathbf{r}_2 - \mathbf{r}_1}{|\mathbf{r}_2 - \mathbf{r}_1|}$$

Coulomb's law when charges are arbitrarily located

or

$$\mathbf{F}_{21} = \frac{k_e q_1 q_2}{|\mathbf{r}_2 - \mathbf{r}_1|^3}(\mathbf{r}_2 - \mathbf{r}_1). \tag{1.24}$$

Although we assumed repulsion, we see that (1.24) includes attraction as well. Indeed, if $q_1 q_2 < 0$, $\mathbf{F}_{21}$ is opposite to $\mathbf{r}_2 - \mathbf{r}_1$, i.e., $\mathbf{F}_{21}$ is directed from P_2 to P_1. Since $\mathbf{F}_{21}$ is the force *on* q_2 by q_1, this is an *attraction*. We also note that Newton's third law is included in (1.24):

$$\mathbf{F}_{12} = \frac{k_e q_2 q_1}{|\mathbf{r}_1 - \mathbf{r}_2|^3}(\mathbf{r}_1 - \mathbf{r}_2) = -\mathbf{F}_{21}$$

because $\mathbf{r}_2 - \mathbf{r}_1 = -(\mathbf{r}_1 - \mathbf{r}_2)$ and $|\mathbf{r}_2 - \mathbf{r}_1| = |\mathbf{r}_1 - \mathbf{r}_2|$.

vector form of gravitational force

We can also write the gravitational force immediately

$$\mathbf{F}_{21} = -\frac{G m_1 m_2}{|\mathbf{r}_2 - \mathbf{r}_1|^3}(\mathbf{r}_2 - \mathbf{r}_1), \tag{1.25}$$

where m_1 and m_2 are point masses and the minus sign is introduced to ensure attraction. ■

Now that we have expressions for electric and gravitational forces, we can obtain the electric field of a point charge and the gravitational field of a point mass. First recall that the **electric field** at a point P is defined to be the force on a test charge q located at P divided by q. Thus if we have a charge q_1, at P_1 with position vector $\mathbf{r}_1$ and we are interested in its fields at P with position vector $\mathbf{r}$, we introduce a charge q at $\mathbf{r}$ and calculate the force on q from Equation (1.24):

$$\mathbf{F}_q = \frac{k_e q_1 q}{|\mathbf{r} - \mathbf{r}_1|^3}(\mathbf{r} - \mathbf{r}_1).$$

electric field of a point charge

Dividing by q gives

$$\mathbf{E}_1 = \frac{k_e q_1}{|\mathbf{r} - \mathbf{r}_1|^3}(\mathbf{r} - \mathbf{r}_1), \tag{1.26}$$

where we have given the field the same index as the charge producing it.

The calculation of the gravitational field follows similarly. The result is

$$\mathbf{g}_1 = -\frac{Gm_1}{|\mathbf{r} - \mathbf{r}_1|^3}(\mathbf{r} - \mathbf{r}_1). \tag{1.27}$$

field point and source point

In (1.26) and (1.27), P is called the **field point** and P_1 the **source point**. Note that in both expressions, the field position vector comes first.

superposition principle explained

If there are several point charges (or masses) producing an electric (gravitational) field, we simply add the contributions from each source. The principle behind this procedure is called the **superposition principle**. It is a principle that "seems" intuitively obvious, but upon further reflection its validity becomes surprising. Suppose a charge q_1 produces a field $\mathbf{E}_1$ around itself. Now we introduce a second charge q_2 which, far away and isolated from any other charges, produced a field $\mathbf{E}_2$ around itself. It is not at all obvious that once we move these charges together, the individual fields should not change. After all, this is not what happens to human beings! We act completely differently when we are alone than when we are in the company of others. The presence of others drastically changes our individual behaviors. Nevertheless, charges and masses, unfettered by any social chains, retain their individuality and produce fields as if no other charges were present.

It is important to keep in mind that the superposition principle applies only to point sources. For example, a charged conducting sphere will not produce the same field when another charge is introduced nearby, because the presence of the new charge alters the charge distribution of the sphere and indeed does change the sphere's field. However each individual point charge (electron) on the sphere, whatever location on the sphere it happens to end up in, will retain its individual electric field.[14]

Going back to the electric field, we can write

$$\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2 + \cdots + \mathbf{E}_n$$

for n point charges $q_1, q_2, \ldots, q_n$ (see Figure 1.16). Substituting from (1.26), with appropriate indices, we obtain

$$\mathbf{E} = \frac{k_e q_1}{|\mathbf{r} - \mathbf{r}_1|^3}(\mathbf{r} - \mathbf{r}_1) + \frac{k_e q_2}{|\mathbf{r} - \mathbf{r}_2|^3}(\mathbf{r} - \mathbf{r}_2) + \cdots + \frac{k_e q_n}{|\mathbf{r} - \mathbf{r}_n|^3}(\mathbf{r} - \mathbf{r}_n)$$

or, using the summation symbol, we obtain

---

[14] The superposition principle, which in the case of electrostatics and gravity is needed to calculate the fields of large sources consisting of many point sources, becomes a vital pillar upon which quantum theory is built and by which many of the strange phenomena of quantum physics are explained.

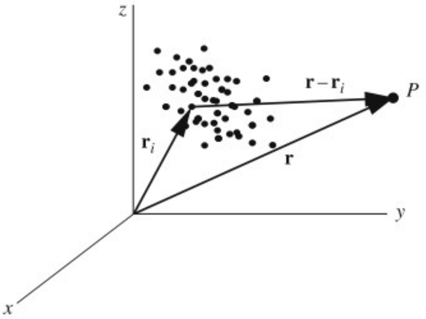

Figure 1.16: The electrostatic field of N point charges is the sum of the electric fields of the individual charges.

Box 1.3.5. *The electric field of n point charges $q_1, q_2, \ldots, q_n$, located at position vectors $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n$ is $\mathbf{E} = \sum_{i=1}^{n} \frac{k_e q_i}{|\mathbf{r}-\mathbf{r}_i|^3}(\mathbf{r}-\mathbf{r}_i)$, and the analogous expression for the gravitational field of n point masses $m_1, m_2, \ldots, m_n$ is $\mathbf{g} = -\sum_{i=1}^{n} \frac{G m_i}{|\mathbf{r}-\mathbf{r}_i|^3}(\mathbf{r}-\mathbf{r}_i)$.*

Historical Notes

The concept of force has a fascinating history which started in the works of Galileo around the beginning of the seventeenth century, mathematically formulated and precisely defined by Sir Isaac Newton in the second half of the seventeenth century, revised and redefined in the form of **fields** by Michael Faraday and James Maxwell in the mid nineteenth century, and finally brought to its modern quantum field theoretical form by Dirac, Heisenberg, Feynman, Schwinger, and others by the mid twentieth century.

notion of field elaborated

Newton, in his theory of gravity, thought of gravitational force as "action-at-a-distance," an agent which affects something that is "there" because of the influence of something that is "here." This kind of interpretation of force had both philosophical and physical drawbacks. It is hard to accept a ghostlike influence on a distant object. Is there an agent that "carries" this influence? What is this agent, if any? Does the influence travel infinitely fast? If we remove the Sun from the Solar System would the Earth and other planets "feel" the absence of the Sun immediately?

These questions, plus others, prompted physicists to come up with the idea of a field. According to this interpretation, the Sun, by its mere presence, creates around itself an invisible three dimensional "sheet" such that, if any object is placed in this sheet, it feels the gravitational force. The reason that planets feel the force of gravity of the Sun is because they happen to be located in the gravitational field of the Sun. The reason that an apple falls to the Earth is because it is in the gravitational field of the Earth and not due to some kind of action-at-a-distance ghost.

Therefore, according to this concept, the force acts on an object *here*, because there exists a field right *here*. And force becomes a local concept. The field concept removes the difficulties associated with action-at-a-distance. The "agent" that transmits the influence from the source to the object, is the field. If the Sun is stolen from the solar system, the Earth will not feel the absence of the Sun immediately. It will receive the information of such cosmic burglary after a certain time-lapse corresponding to the time required for the disturbance to travel from the Sun to the Earth. We can liken such a disturbance (disappearance of the Sun) to a disturbance in the smooth water of a quiet pond (by dropping a stone into it). Clearly, the disturbance travels from the source (where the stone was dropped) to any other point with a finite speed, the speed of the water waves.

The concept of a field was actually introduced first in the context of electricity and magnetism by Michael Faraday as a means of "visualizing" electromagnetic effects to replace certain mathematical ideas for which he had little talent. However, in the hands of James Maxwell, fields were molded into a physical entity having an existence of their own in the form of electromagnetic waves to be produced in 1887 by Hertz and used in 1901 by Marconi in the development of radio.

potential

A concept related to that of fields is **potential** which is closely tied to the work done by the fields on a charge (in the case of electrostatics) or a mass (in the case of gravity). It can be shown[15] that the gravitational potential $\Phi(\mathbf{r})$ at $\mathbf{r}$, of n point masses, is given by

$$\Phi(\mathbf{r}) = -\sum_{i=1}^{n} \frac{Gm_i}{|\mathbf{r} - \mathbf{r}_i|} \tag{1.28}$$

and that of n point charges by

$$\Phi(\mathbf{r}) = \sum_{i=1}^{n} \frac{k_e q_i}{|\mathbf{r} - \mathbf{r}_i|}. \tag{1.29}$$

Note that in both cases, the potential goes to zero as $\mathbf{r}$ goes to infinity. This has to do with the choice of the location of the zero of potential, which we have chosen to be the point at infinity in Equations (1.28) and (1.29).

Example 1.3.4. The electric charges q_1, q_2, q_3, and q_4 are located at Cartesian $(a, 0, 0)$, $(0, a, 0)$, $(-a, 0, 0)$, and $(0, -a, 0)$, respectively. We want to find the electric field and the electrostatic potential at an arbitrary point on the z-axis. We note that

$$\mathbf{r}_1 = a\hat{\mathbf{e}}_x, \quad \mathbf{r}_2 = a\hat{\mathbf{e}}_y, \quad \mathbf{r}_3 = -a\hat{\mathbf{e}}_x, \quad \mathbf{r}_4 = -a\hat{\mathbf{e}}_y, \quad \mathbf{r} = z\hat{\mathbf{e}}_z,$$

so that

$$\begin{aligned} \mathbf{r} - \mathbf{r}_1 &= -a\hat{\mathbf{e}}_x + z\hat{\mathbf{e}}_z, & \mathbf{r} - \mathbf{r}_2 &= -a\hat{\mathbf{e}}_y + z\hat{\mathbf{e}}_z, \\ \mathbf{r} - \mathbf{r}_3 &= a\hat{\mathbf{e}}_x + z\hat{\mathbf{e}}_z, & \mathbf{r} - \mathbf{r}_4 &= a\hat{\mathbf{e}}_y + z\hat{\mathbf{e}}_z, \end{aligned}$$

---

[15] See Chapter 14 for details.

and $|\mathbf{r}-\mathbf{r}_i|^3 = (a^2+z^2)^{3/2}$ for all i. The electric field can now be calculated using Box 1.3.5:

$$\begin{aligned}\mathbf{E} &= \frac{k_e q_1}{(a^2+z^2)^{3/2}}(-a\hat{\mathbf{e}}_x + z\hat{\mathbf{e}}_z) + \frac{k_e q_2}{(a^2+z^2)^{3/2}}(-a\hat{\mathbf{e}}_y + z\hat{\mathbf{e}}_z) \\ &+ \frac{k_e q_3}{(a^2+z^2)^{3/2}}(a\hat{\mathbf{e}}_x + z\hat{\mathbf{e}}_z) + \frac{k_e q_4}{(a^2+z^2)^{3/2}}(a\hat{\mathbf{e}}_y + z\hat{\mathbf{e}}_z) \\ &= \frac{k_e}{(a^2+z^2)^{3/2}}\left[(-aq_1+aq_3)\hat{\mathbf{e}}_x + (-aq_2+aq_4)\hat{\mathbf{e}}_y + (q_1+q_2+q_3+q_4)z\hat{\mathbf{e}}_z\right].\end{aligned}$$

It is interesting to note that if the sum of all charges is zero, the z-component of the electric field vanishes at all points on the z-axis. Furthermore, if, in addition, $q_1 = q_3$ and $q_2 = q_4$, there will be no electric field at any point on the z-axis.

The potential is obtained similarly:

$$\begin{aligned}\Phi &= \frac{k_e q_1}{(a^2+z^2)^{1/2}} + \frac{k_e q_2}{(a^2+z^2)^{1/2}} + \frac{k_e q_3}{(a^2+z^2)^{1/2}} + \frac{k_e q_4}{(a^2+z^2)^{1/2}} \\ &= \frac{k_e(q_1+q_2+q_3+q_4)}{\sqrt{a^2+z^2}}.\end{aligned}$$

So, the potential is zero at all points of the z-axis, as long as the total charge is zero. ■

1.3.2 Cross Product

The unit vectors in the three coordinate systems are not only mutually perpendicular, but in the order in which they are given, they also form a right-handed set [see Equation (1.5)]. Therefore, we can use Equation (1.6) and write

$$\mathbf{a}\times\mathbf{b} = \underbrace{\det\begin{pmatrix}\hat{\mathbf{e}}_x & \hat{\mathbf{e}}_y & \hat{\mathbf{e}}_z \\ a_x & a_y & a_z \\ b_x & b_y & b_z\end{pmatrix}}_{\text{in Cartesian CS}} = \underbrace{\det\begin{pmatrix}\hat{\mathbf{e}}_\rho & \hat{\mathbf{e}}_\varphi & \hat{\mathbf{e}}_z \\ a_\rho & a_\varphi & a_z \\ b_\rho & b_\varphi & b_z\end{pmatrix}}_{\text{in cylindrical CS}} = \underbrace{\det\begin{pmatrix}\hat{\mathbf{e}}_r & \hat{\mathbf{e}}_\theta & \hat{\mathbf{e}}_\varphi \\ a_r & a_\theta & a_\varphi \\ b_r & b_\theta & b_\varphi\end{pmatrix}}_{\text{in spherical CS}} \tag{1.30}$$

Two important prototypes of the concept of cross product are **angular momentum** and **torque**. A particle moving with instantaneous linear momentum $\mathbf{p}$ relative to an origin O has *instantaneous* angular momentum $\mathbf{L} = \mathbf{r}\times\mathbf{p}$ if its instantaneous position vector with respect to O is $\mathbf{r}$. In Figure 1.17 we have shown $\mathbf{r}$, $\mathbf{p}$, and $\mathbf{r}\times\mathbf{p}$. Similarly, if the instantaneous force on the above particle is $\mathbf{F}$, then the instantaneous torque acting on it is $\mathbf{T} = \mathbf{r}\times\mathbf{F}$.

angular momentum and torque as examples of cross products

If there are more than one particle we simply add the contribution of individual particles. Thus, the total angular momentum $\mathbf{L}$ of N particles and the total torque $\mathbf{T}$ acting on them are

$$\mathbf{L} = \sum_{k=1}^{N}\mathbf{r}_k\times\mathbf{p}_k \qquad \text{and} \qquad \mathbf{T} = \sum_{k=1}^{N}\mathbf{r}_k\times\mathbf{F}_k, \tag{1.31}$$

where $\mathbf{r}_k$ is the position of the kth particle, $\mathbf{p}_k$ its instantaneous momentum, and $\mathbf{F}_k$ the instantaneous force acting on it.

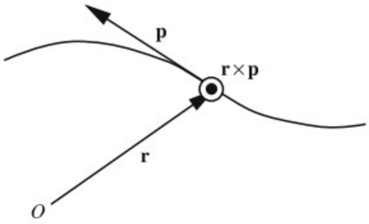

Figure 1.17: Angular momentum of a moving particle with respect to the origin O. The circle with a dot in its middle represents a vector pointing out of the page. It is assumed that $\mathbf{r}$ and $\mathbf{p}$ lie in the page.

Example 1.3.5. In this example, we show that the torque on a collection of three particles is caused by *external* forces only. The torques due to the internal forces add up to zero. The generalization to an arbitrary number of particles will be done in Example 9.2.1 when we learn how to manipulate summation symbols.

For $N = 3$, the second formula in Equation (1.31) reduces to

$$\mathbf{T} = \mathbf{r}_1 \times \mathbf{F}_1 + \mathbf{r}_2 \times \mathbf{F}_2 + \mathbf{r}_3 \times \mathbf{F}_3.$$

Each force can be divided into an external part and an internal part, the latter being the force caused by the presence of the other particles. So, we have

$$\begin{aligned}
\mathbf{F}_1 &= \mathbf{F}_1^{(\text{ext})} + \mathbf{F}_{12} + \mathbf{F}_{13}, \\
\mathbf{F}_2 &= \mathbf{F}_2^{(\text{ext})} + \mathbf{F}_{21} + \mathbf{F}_{23}, \\
\mathbf{F}_3 &= \mathbf{F}_3^{(\text{ext})} + \mathbf{F}_{31} + \mathbf{F}_{32},
\end{aligned}$$

where $\mathbf{F}_{12}$ is the force on particle 1 exerted by particle 2, etc. Substituting in the above expression for the torque, we get

$$\begin{aligned}
\mathbf{T} &= \mathbf{r}_1 \times \mathbf{F}_1^{(\text{ext})} + \mathbf{r}_2 \times \mathbf{F}_2^{(\text{ext})} + \mathbf{r}_3 \times \mathbf{F}_3^{(\text{ext})} \\
&\quad + \mathbf{r}_1 \times \mathbf{F}_{12} + \mathbf{r}_1 \times \mathbf{F}_{13} + \mathbf{r}_2 \times \mathbf{F}_{21} + \mathbf{r}_2 \times \mathbf{F}_{23} + \mathbf{r}_3 \times \mathbf{F}_{31} + \mathbf{r}_3 \times \mathbf{F}_{32} \\
&= \mathbf{T}^{(\text{ext})} + (\mathbf{r}_1 - \mathbf{r}_2) \times \mathbf{F}_{12} + (\mathbf{r}_1 - \mathbf{r}_3) \times \mathbf{F}_{13} + (\mathbf{r}_2 - \mathbf{r}_3) \times \mathbf{F}_{23},
\end{aligned}$$

where we used the third law of motion: $\mathbf{F}_{12} = -\mathbf{F}_{21}$, etc. Now we note that the internal force between two particles, 1 and 2 say, is along the line joining them, i.e., along $\mathbf{r}_1 - \mathbf{r}_2$. It follows that all the cross products in the last line of the equation above vanish and we get $\mathbf{T} = \mathbf{T}^{(\text{ext})}$. ■

We have already seen that multiplying *a vector by a number* gives another vector. A physical example of this is electric force which is obtained by multiplying electric field by electric charge. In fact we divided the electric force by charge to get the electric field. Historically, it was the law of the force which was discovered first and *then* the concept of electric field was defined. We have also seen that one can get a new vector by cross-multiplying *two vectors*. The rule of this kind of multiplication is, however, more complicated. It turns out

from electric field to electric force

8H <

that the magnetic force is related to the magnetic field via such a cross multiplication. What is worse is that the magnetic field is also related to its source (electric charges *in motion*) via such a product. Little wonder that magnetic phenomena are mathematically so much more complicated than their electric counterparts. That is why in the study of magnetism, one first introduces the concept of magnetic field and how it is related to the motion of charges producing it, and *then* the force of this field on moving charges.

magnetic field of a moving charge or Biot–Savart law

Example 1.3.6. A charge q, located instantaneously at the origin, is moving with velocity $\mathbf{v}$ relative to P [see Figure 1.18(a)]. Assuming that $|\mathbf{v}|$ is much smaller than the speed of light, the *instantaneous* magnetic field at P due to q is given by $\mathbf{B} = \frac{k_m q\,\mathbf{v} \times \hat{\mathbf{e}}_r}{r^2}$, or, using $\hat{\mathbf{e}}_r = \mathbf{r}/r$, by $\mathbf{B} = \frac{k_m q\,\mathbf{v} \times \mathbf{r}}{r^3}$. This is a simple version of a more general formula known as the **Biot–Savart law**. In the above relations, k_m is the analog of k_e in the electric case.

If we are interested in the magnetic field when q is located at a point other than the origin, we replace $\mathbf{r}$ with the vector from the instantaneous location of the moving charge to P. This is shown in Figure 1.18(b), where the vector from q_1 to P is to replace $\mathbf{r}$ in the above equation. More specifically, we have

$$\mathbf{B}_1 = \frac{k_m q_1 \mathbf{v}_1 \times (\mathbf{r} - \mathbf{r}_1)}{|\mathbf{r} - \mathbf{r}_1|^3}. \tag{1.32}$$

If there are N charges, the total magnetic field will be

$$\mathbf{B} = \sum_{k=1}^{N} \frac{k_m q_k \mathbf{v}_k \times (\mathbf{r} - \mathbf{r}_k)}{|\mathbf{r} - \mathbf{r}_k|^3}, \tag{1.33}$$

where we have used the superposition principle. ■

magnetic force on a moving charge.

When a charge q moves with velocity $\mathbf{v}$ in a magnetic field $\mathbf{B}$, it experiences a force given by

$$\mathbf{F} = q\mathbf{v} \times \mathbf{B}. \tag{1.34}$$

It is instructive to write the magnetic force exerted by a charge q_1 moving with velocity $\mathbf{v}_1$ on a second charge q_2 moving with velocity $\mathbf{v}_2$. We leave this as an exercise for the reader.

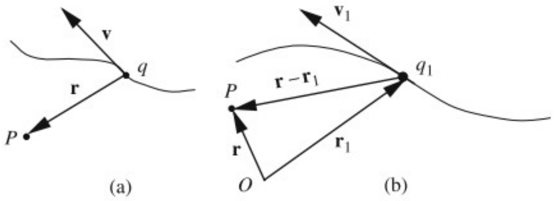

Figure 1.18: The (instantaneous) magnetic field at P of a moving point charge (a) when P is at the origin, and (b) when P is different from the origin. The field points out of the page for the configuration shown.

Example 1.3.7. A charge q moves with constant speed v (assumed to be small compared to the speed of light) on a straight line. We want to calculate the magnetic field produced by the charge at a point P located at a distance ρ from the line as a function of time. Cylindrical coordinates are most suitable for this problem because of the existence of a natural axis. Choose the path of the charge to be the z-axis. Also assume that P lies in the xy-plane, and that q was at the origin at $t = 0$. Then $\mathbf{v} = v\hat{\mathbf{e}}_z$, $\mathbf{r} = \rho\hat{\mathbf{e}}_\rho$, $\mathbf{r}_1 = vt\hat{\mathbf{e}}_z$, $\mathbf{r} - \mathbf{r}_1 = \rho\hat{\mathbf{e}}_\rho - vt\hat{\mathbf{e}}_z$. So

$$|\mathbf{r} - \mathbf{r}_1| = \sqrt{(\rho\hat{\mathbf{e}}_\rho - vt\hat{\mathbf{e}}_z) \cdot (\rho\hat{\mathbf{e}}_\rho - vt\hat{\mathbf{e}}_z)} = \sqrt{\rho^2 + v^2t^2}$$

and $\mathbf{v} \times (\mathbf{r} - \mathbf{r}_1) = v\hat{\mathbf{e}}_z \times (\rho\hat{\mathbf{e}}_\rho - vt\hat{\mathbf{e}}_z) = \rho v\hat{\mathbf{e}}_\varphi$. Therefore, the magnetic field is

$$\mathbf{B} = \frac{k_m q\mathbf{v} \times (\mathbf{r} - \mathbf{r}_1)}{|\mathbf{r} - \mathbf{r}_1|^3} = \frac{k_m q\rho v}{(\rho^2 + v^2t^2)^{3/2}}\hat{\mathbf{e}}_\varphi.$$

Readers familiar with the relation between magnetic fields and currents in long wires will note that the magnetic field above obeys the right-hand rule. ■

1.4 Relations Among Unit Vectors

We have seen that, depending on the nature of problems encountered in physics, one coordinate system may be more useful than others. We have also seen that the *coordinates* can be transformed back and forth using functional relations that connect them. Since many physical quantities are vectors, transformation and expression of components in bases of various coordinate systems also become important. The key to this transformation is writing one set of unit vectors in terms of others. In the derivation of these relations, we shall make heavy use of Box 1.1.2.

First we write the cylindrical unit vectors in terms of Cartesian unit vectors. Since $\{\hat{\mathbf{e}}_x, \hat{\mathbf{e}}_y, \hat{\mathbf{e}}_z\}$ form a basis, any vector can be written in terms of them. In particular, $\hat{\mathbf{e}}_\rho$ can be expressed as

$$\hat{\mathbf{e}}_\rho = a_1\hat{\mathbf{e}}_x + b_1\hat{\mathbf{e}}_y + c_1\hat{\mathbf{e}}_z \tag{1.35}$$

with a_1, b_1, and c_1 to be determined. Next we recall that

Box 1.4.1. *The dot product of two unit vectors is the cosine of the angle between them.*

Furthermore, Figure 1.12 shows that the angle between $\hat{\mathbf{e}}_\rho$ and $\hat{\mathbf{e}}_x$ is φ, and that between $\hat{\mathbf{e}}_\rho$ and $\hat{\mathbf{e}}_y$ is $\pi/2 - \varphi$. So, by dotting both sides of Equation (1.35) by $\hat{\mathbf{e}}_x$, $\hat{\mathbf{e}}_y$, and $\hat{\mathbf{e}}_z$ in succession, we obtain

$$\underbrace{\hat{\mathbf{e}}_x \cdot \hat{\mathbf{e}}_\rho}_{=\cos\varphi} = a_1 + 0 + 0 = a_1 \;\Rightarrow\; a_1 = \cos\varphi,$$

$$\underbrace{\hat{\mathbf{e}}_y \cdot \hat{\mathbf{e}}_\rho}_{=\sin\varphi} = 0 + b_1 + 0 = b_1 \;\Rightarrow\; b_1 = \sin\varphi,$$

$$\underbrace{\hat{\mathbf{e}}_z \cdot \hat{\mathbf{e}}_\rho}_{=0} = 0 + 0 + c_1 = c_1 \;\Rightarrow\; c_1 = 0.$$

Therefore,

$$\hat{\mathbf{e}}_\rho = \hat{\mathbf{e}}_x \cos\varphi + \hat{\mathbf{e}}_y \sin\varphi.$$

With the first and third cylindrical unit vectors $\hat{\mathbf{e}}_\rho$ and $\hat{\mathbf{e}}_z$ at our disposal,[16] we can determine the second, using Equation (1.5):

$$\hat{\mathbf{e}}_\varphi = \hat{\mathbf{e}}_z \times \hat{\mathbf{e}}_\rho = \det\begin{pmatrix} \hat{\mathbf{e}}_x & \hat{\mathbf{e}}_y & \hat{\mathbf{e}}_z \\ 0 & 0 & 1 \\ \cos\varphi & \sin\varphi & 0 \end{pmatrix} = -\hat{\mathbf{e}}_x \sin\varphi + \hat{\mathbf{e}}_y \cos\varphi.$$

Thus,

cylindrical unit vectors in terms of Cartesian unit vectors

$$\begin{aligned} \hat{\mathbf{e}}_\rho &= \hat{\mathbf{e}}_x \cos\varphi + \hat{\mathbf{e}}_y \sin\varphi, \\ \hat{\mathbf{e}}_\varphi &= -\hat{\mathbf{e}}_x \sin\varphi + \hat{\mathbf{e}}_y \cos\varphi, \\ \hat{\mathbf{e}}_z &= \hat{\mathbf{e}}_z. \end{aligned} \tag{1.36}$$

This equation can easily be inverted to find the Cartesian unit vectors in terms of the cylindrical unit vectors. For example, the coefficients in

$$\hat{\mathbf{e}}_x = a_2\hat{\mathbf{e}}_\rho + b_2\hat{\mathbf{e}}_\varphi + c_2\hat{\mathbf{e}}_z$$

can be obtained by dotting both sides of it with $\hat{\mathbf{e}}_\rho$, $\hat{\mathbf{e}}_\varphi$, and $\hat{\mathbf{e}}_z$, respectively,

$$\begin{aligned} \hat{\mathbf{e}}_\rho \cdot \hat{\mathbf{e}}_x &= a_2 + 0 + 0 \;\Rightarrow\; \cos\varphi = a_2, \\ \hat{\mathbf{e}}_\varphi \cdot \hat{\mathbf{e}}_x &= 0 + b_2 + 0 \;\Rightarrow\; -\sin\varphi = b_2, \\ \hat{\mathbf{e}}_z \cdot \hat{\mathbf{e}}_x &= 0 + 0 + c_2 \;\Rightarrow\; 0 = c_2, \end{aligned}$$

where we have used $\hat{\mathbf{e}}_\rho \cdot \hat{\mathbf{e}}_x = \cos\varphi$, and $\hat{\mathbf{e}}_\varphi \cdot \hat{\mathbf{e}}_x = -\sin\varphi$—obtained by dotting the first and second equations of (1.36) with $\hat{\mathbf{e}}_x$—as well as $\hat{\mathbf{e}}_z \cdot \hat{\mathbf{e}}_x = 0$. Similarly, one can obtain $\hat{\mathbf{e}}_y$ in terms of the cylindrical unit vectors. The entire result is

Cartesian unit vectors in terms of cylindrical unit vectors

$$\begin{aligned} \hat{\mathbf{e}}_x &= \hat{\mathbf{e}}_\rho \cos\varphi - \hat{\mathbf{e}}_\varphi \sin\varphi \\ \hat{\mathbf{e}}_y &= \hat{\mathbf{e}}_\rho \sin\varphi + \hat{\mathbf{e}}_\varphi \cos\varphi \\ \hat{\mathbf{e}}_z &= \hat{\mathbf{e}}_z \end{aligned} \tag{1.37}$$

Now we express the spherical unit vectors in terms of the cylindrical ones. This is easily done for $\hat{\mathbf{e}}_r$, because it has only $\hat{\mathbf{e}}_\rho$ and $\hat{\mathbf{e}}_z$ components (why?). Thus, with

$$\hat{\mathbf{e}}_r = a_3\hat{\mathbf{e}}_\rho + b_3\hat{\mathbf{e}}_z,$$

we obtain

$$\begin{aligned} \hat{\mathbf{e}}_\rho \cdot \hat{\mathbf{e}}_r &= a_3 + 0 \;\Rightarrow\; a_3 = \sin\theta, \\ \hat{\mathbf{e}}_z \cdot \hat{\mathbf{e}}_r &= 0 + b_3 \;\Rightarrow\; b_3 = \cos\theta, \end{aligned}$$

---

[16] Remember that $\hat{\mathbf{e}}_z$ is a unit vector in both coordinate systems. So, one can say that the *cylindrical* $\hat{\mathbf{e}}_z$ has components $\langle 0, 0, 1\rangle$ in the *Cartesian* basis $\{\hat{\mathbf{e}}_x, \hat{\mathbf{e}}_y, \hat{\mathbf{e}}_z\}$.

where in the last step of each line, we used the fact that the angle between $\hat{\mathbf{e}}_r$ and $\hat{\mathbf{e}}_z$ is θ and that between $\hat{\mathbf{e}}_r$ and $\hat{\mathbf{e}}_\rho$ is $\pi/2-\theta$ (see Figure 1.13). With a_3 and b_3 so determined, we can write

$$\hat{\mathbf{e}}_r = \hat{\mathbf{e}}_\rho \sin\theta + \hat{\mathbf{e}}_z \cos\theta.$$

Having two spherical unit vectors $\hat{\mathbf{e}}_r$ and $\hat{\mathbf{e}}_\varphi$ at our disposal,[17] we can determine the third one, using (1.5) and (1.30):

$$\hat{\mathbf{e}}_\theta = \hat{\mathbf{e}}_\varphi \times \hat{\mathbf{e}}_r = \det\begin{pmatrix} \hat{\mathbf{e}}_\rho & \hat{\mathbf{e}}_\varphi & \hat{\mathbf{e}}_z \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{pmatrix} = \hat{\mathbf{e}}_\rho \cos\theta - \hat{\mathbf{e}}_z \sin\theta.$$

Thus,

$$\begin{aligned} \hat{\mathbf{e}}_r &= \hat{\mathbf{e}}_\rho \sin\theta + \hat{\mathbf{e}}_z \cos\theta, \\ \hat{\mathbf{e}}_\theta &= \hat{\mathbf{e}}_\rho \cos\theta - \hat{\mathbf{e}}_z \sin\theta, \\ \hat{\mathbf{e}}_\varphi &= \hat{\mathbf{e}}_\varphi. \end{aligned} \tag{1.38}$$

spherical unit vectors in terms of cylindrical unit vectors

The inverse relations can be obtained as before. We leave the details of the calculation as an exercise for the reader.

Combining Equations (1.36) and (1.38), we can express spherical unit vectors in terms of the Cartesian unit vectors:

spherical unit vectors in terms of Cartesian unit vectors

$$\begin{aligned} \hat{\mathbf{e}}_r &= \hat{\mathbf{e}}_x \sin\theta\cos\varphi + \hat{\mathbf{e}}_y \sin\theta\sin\varphi + \hat{\mathbf{e}}_z \cos\theta, \\ \hat{\mathbf{e}}_\theta &= \hat{\mathbf{e}}_x \cos\theta\cos\varphi + \hat{\mathbf{e}}_y \cos\theta\sin\varphi - \hat{\mathbf{e}}_z \sin\theta, \\ \hat{\mathbf{e}}_\varphi &= -\hat{\mathbf{e}}_x \sin\varphi + \hat{\mathbf{e}}_y \cos\varphi. \end{aligned} \tag{1.39}$$

Equations (1.39) and (1.36) are very useful when calculating vector quantities in spherical and cylindrical coordinates as we shall see in many examples to follow. These equations also allow us to express a unit vector in one of the three coordinate systems in terms of the unit vectors of any other coordinate system.

Example 1.4.1. P_1 and P_2 have Cartesian coordinates $(1,1,1)$ and $(-1,2,-1)$, respectively. A vector $\mathbf{a}$ has spherical components $\langle 0,2,0\rangle$ at P_1. We want to find the spherical components of $\mathbf{a}$ at P_2. These are given by $\mathbf{a}\cdot\hat{\mathbf{e}}_{r_2}$, $\mathbf{a}\cdot\hat{\mathbf{e}}_{\theta_2}$, and $\mathbf{a}\cdot\hat{\mathbf{e}}_{\varphi_2}$. In order to calculate these dot products, it is most convenient to express all vectors in Cartesian form. So, using Equation (1.39), we have

$$\mathbf{a} = 2\hat{\mathbf{e}}_{\theta_1} = 2\left(\hat{\mathbf{e}}_x \cos\theta_1\cos\varphi_1 + \hat{\mathbf{e}}_y \cos\theta_1\sin\varphi_1 - \hat{\mathbf{e}}_z \sin\theta_1\right),$$

where (r_1,θ_1,φ_1) are coordinates of P_1. We can calculate these from the Cartesian coordinates of P_1:

$$r_1 = \sqrt{1^2+1^2+1^2} = \sqrt{3}, \quad \cos\theta_1 = \frac{z_1}{r_1} = \frac{1}{\sqrt{3}}, \quad \tan\varphi_1 = \frac{y_1}{x_1} = 1 \;\Rightarrow\; \varphi_1 = \frac{\pi}{4}.$$

---

[17] Recall that $\hat{\mathbf{e}}_\varphi$ is both a cylindrical and a spherical unit vector.

Therefore,

$$\mathbf{a} = 2\left(\hat{\mathbf{e}}_x \frac{1}{\sqrt{3}}\frac{1}{\sqrt{2}} + \hat{\mathbf{e}}_y \frac{1}{\sqrt{3}}\frac{1}{\sqrt{2}} - \hat{\mathbf{e}}_z \sqrt{\frac{2}{3}}\right) = \frac{2}{\sqrt{6}}\hat{\mathbf{e}}_x + \frac{2}{\sqrt{6}}\hat{\mathbf{e}}_y - \frac{4}{\sqrt{6}}\hat{\mathbf{e}}_z.$$

Now we need to express $\hat{\mathbf{e}}_{r_2}$, $\hat{\mathbf{e}}_{\theta_2}$, and $\hat{\mathbf{e}}_{\varphi_2}$ in terms of Cartesian unit vectors. Once again we use Equation (1.39) for which we need the spherical coordinates of P_2:

$$r_2 = \sqrt{(-1)^2 + 2^2 + (-1)^2} = \sqrt{6}, \quad \cos\theta_2 = \frac{z_2}{r_2} = -\frac{1}{\sqrt{6}}, \quad \tan\varphi_2 = \frac{y_2}{x_2} = -2.$$

Similarly, Equations (1.11) and (1.12) yield

$$\sin\theta_2 = +\sqrt{\frac{5}{6}}, \qquad \cos\varphi_2 = -\frac{1}{\sqrt{5}}, \qquad \sin\varphi_2 = +\frac{2}{\sqrt{5}}.$$

Then

$$\begin{aligned}
\hat{\mathbf{e}}_{r_2} &= \hat{\mathbf{e}}_x \sin\theta_2\cos\varphi_2 + \hat{\mathbf{e}}_y \sin\theta_2 \sin\varphi_2 + \hat{\mathbf{e}}_z \cos\theta_2 \\
&= \hat{\mathbf{e}}_x \sqrt{\frac{5}{6}}\left(-\frac{1}{\sqrt{5}}\right) + \hat{\mathbf{e}}_y \sqrt{\frac{5}{6}}\frac{2}{\sqrt{5}} - \hat{\mathbf{e}}_z \frac{1}{\sqrt{6}} \\
&= -\frac{1}{\sqrt{6}}\hat{\mathbf{e}}_x + \frac{2}{\sqrt{6}}\hat{\mathbf{e}}_y - \frac{1}{\sqrt{6}}\hat{\mathbf{e}}_z, \\
\hat{\mathbf{e}}_{\theta_2} &= \hat{\mathbf{e}}_x \cos\theta_2\cos\varphi_2 + \hat{\mathbf{e}}_y \cos\theta_2 \sin\varphi_2 - \hat{\mathbf{e}}_z \sin\theta_2 \\
&= \hat{\mathbf{e}}_x\left(-\frac{1}{\sqrt{6}}\right)\left(-\frac{1}{\sqrt{5}}\right) + \hat{\mathbf{e}}_y\left(-\frac{1}{\sqrt{6}}\right)\frac{2}{\sqrt{5}} - \hat{\mathbf{e}}_z\sqrt{\frac{5}{6}} \\
&= \frac{1}{\sqrt{30}}\hat{\mathbf{e}}_x - \frac{2}{\sqrt{30}}\hat{\mathbf{e}}_y - \frac{5}{\sqrt{30}}\hat{\mathbf{e}}_z, \\
\hat{\mathbf{e}}_{\varphi_2} &= -\hat{\mathbf{e}}_x \sin\varphi_2 + \hat{\mathbf{e}}_y \cos\varphi_2 = -\frac{2}{\sqrt{5}}\hat{\mathbf{e}}_x - \frac{1}{\sqrt{5}}\hat{\mathbf{e}}_y.
\end{aligned}$$

We can now take the dot products required for the components:

$$\begin{aligned}
r \text{ comp} = \mathbf{a}\cdot\hat{\mathbf{e}}_{r_2} &= \left(\frac{2}{\sqrt{6}}\hat{\mathbf{e}}_x + \frac{2}{\sqrt{6}}\hat{\mathbf{e}}_y - \frac{4}{\sqrt{6}}\hat{\mathbf{e}}_z\right)\cdot\left(-\frac{1}{\sqrt{6}}\hat{\mathbf{e}}_x + \frac{2}{\sqrt{6}}\hat{\mathbf{e}}_y - \frac{1}{\sqrt{6}}\hat{\mathbf{e}}_z\right) \\
&= -\frac{2}{6} + \frac{4}{6} + \frac{4}{6} = 1,
\end{aligned}$$

$$\begin{aligned}
\theta \text{ comp} = \mathbf{a}\cdot\hat{\mathbf{e}}_{\theta_2} &= \left(\frac{2}{\sqrt{6}}\hat{\mathbf{e}}_x + \frac{2}{\sqrt{6}}\hat{\mathbf{e}}_y - \frac{4}{\sqrt{6}}\hat{\mathbf{e}}_z\right)\cdot\left(\frac{1}{\sqrt{30}}\hat{\mathbf{e}}_x - \frac{2}{\sqrt{30}}\hat{\mathbf{e}}_y - \frac{5}{\sqrt{30}}\hat{\mathbf{e}}_z\right) \\
&= \frac{2}{6\sqrt{5}} - \frac{4}{6\sqrt{5}} + \frac{20}{6\sqrt{5}} = \frac{3}{\sqrt{5}}, \\
\varphi \text{ comp} = \mathbf{a}\cdot\hat{\mathbf{e}}_{\varphi_2} &= \left(\frac{2}{\sqrt{6}}\hat{\mathbf{e}}_x + \frac{2}{\sqrt{6}}\hat{\mathbf{e}}_y - \frac{4}{\sqrt{6}}\hat{\mathbf{e}}_z\right)\cdot\left(-\frac{2}{\sqrt{5}}\hat{\mathbf{e}}_x - \frac{1}{\sqrt{5}}\hat{\mathbf{e}}_y\right) \\
&= -\frac{4}{\sqrt{30}} - \frac{2}{\sqrt{30}} = -\frac{6}{\sqrt{30}} = -\sqrt{\frac{6}{5}}.
\end{aligned}$$

It now follows that

$$\mathbf{a} = \hat{\mathbf{e}}_{r_2} + \frac{3}{\sqrt{5}}\hat{\mathbf{e}}_{\theta_2} - \sqrt{\frac{6}{5}}\hat{\mathbf{e}}_{\varphi_2}.$$

As a check, we note that

$$|\mathbf{a}| = \sqrt{1^2 + \left(\frac{3}{\sqrt{5}}\right)^2 + \left(-\sqrt{\frac{6}{5}}\right)^2} = \sqrt{\frac{5+9+6}{5}} = \sqrt{4} = 2,$$

which agrees with the length of $\mathbf{a}$. ■

Example 1.4.2. Points P_1 and P_2 have spherical coordinates $(r_1, \theta_1, \varphi_1)$ and $(r_2, \theta_2, \varphi_2)$, respectively. We want to find: (a) the angle between their position vectors $\mathbf{r}_1$ and $\mathbf{r}_2$ in terms of their coordinates; (b) the spherical components of $\mathbf{r}_2$ at P_1; and (c) the spherical components of $\mathbf{r}_1$ at P_2. Once again, we shall express all vectors in terms of Cartesian unit vectors when evaluating dot products.

(a) The cosine of the angle—call it γ_{12}—between the position vectors is simply $\hat{\mathbf{e}}_{r_1} \cdot \hat{\mathbf{e}}_{r_2}$. We can readily find this by using Equation (1.39):

$$\begin{aligned}
\cos\gamma_{12} = \hat{\mathbf{e}}_{r_1} \cdot \hat{\mathbf{e}}_{r_2} &= (\hat{\mathbf{e}}_x \sin\theta_1 \cos\varphi_1 + \hat{\mathbf{e}}_y \sin\theta_1 \sin\varphi_1 + \hat{\mathbf{e}}_z \cos\theta_1) \\
&\quad \cdot (\hat{\mathbf{e}}_x \sin\theta_2 \cos\varphi_2 + \hat{\mathbf{e}}_y \sin\theta_2 \sin\varphi_2 + \hat{\mathbf{e}}_z \cos\theta_2) \\
&= \sin\theta_1 \cos\varphi_1 \sin\theta_2 \cos\varphi_2 + \sin\theta_1 \sin\varphi_1 \sin\theta_2 \sin\varphi_2 + \cos\theta_1 \cos\theta_2 \\
&= \sin\theta_1 \sin\theta_2 (\cos\varphi_1 \cos\varphi_2 + \sin\varphi_1 \sin\varphi_2) + \cos\theta_1 \cos\theta_2 \\
&= \sin\theta_1 \sin\theta_2 \cos(\varphi_1 - \varphi_2) + \cos\theta_1 \cos\theta_2.
\end{aligned}$$

(b) To find the spherical components of $\mathbf{r}_2$ at P_1, we need to take the dot product of $\mathbf{r}_2$ with the spherical unit vectors at P_1:

$$\begin{aligned}
r \text{ comp} &= \mathbf{r}_2 \cdot \hat{\mathbf{e}}_{r_1} = r_2 \hat{\mathbf{e}}_{r_2} \cdot \hat{\mathbf{e}}_{r_1} \\
&= r_2 \left[\sin\theta_1 \sin\theta_2 \cos(\varphi_1 - \varphi_2) + \cos\theta_1 \cos\theta_2\right], \\
\theta \text{ comp} &= \mathbf{r}_2 \cdot \hat{\mathbf{e}}_{\theta_1} = r_2 \hat{\mathbf{e}}_{r_2} \cdot \hat{\mathbf{e}}_{\theta_1} \\
&= r_2 \left(\hat{\mathbf{e}}_x \sin\theta_2 \cos\varphi_2 + \hat{\mathbf{e}}_y \sin\theta_2 \sin\varphi_2 + \hat{\mathbf{e}}_z \cos\theta_2\right) \\
&\quad \cdot \left(\hat{\mathbf{e}}_x \cos\theta_1 \cos\varphi_1 + \hat{\mathbf{e}}_y \cos\theta_1 \sin\varphi_1 - \hat{\mathbf{e}}_z \sin\theta_1\right) \\
&= r_2 (\sin\theta_2 \cos\varphi_2 \cos\theta_1 \cos\varphi_1 + \sin\theta_2 \sin\varphi_2 \cos\theta_1 \sin\varphi_1 - \cos\theta_2 \sin\theta_1) \\
&= r_2 [\sin\theta_2 \cos\theta_1 \cos(\varphi_1 - \varphi_2) - \cos\theta_2 \sin\theta_1], \\
\varphi \text{ comp} &= \mathbf{r}_2 \cdot \hat{\mathbf{e}}_{\varphi_1} = r_2 \hat{\mathbf{e}}_{r_2} \cdot \hat{\mathbf{e}}_{\varphi_1} \\
&= r_2 \left(\hat{\mathbf{e}}_x \sin\theta_2 \cos\varphi_2 + \hat{\mathbf{e}}_y \sin\theta_2 \sin\varphi_2 + \hat{\mathbf{e}}_z \cos\theta_2\right) \cdot \left(-\hat{\mathbf{e}}_x \sin\varphi_1 + \hat{\mathbf{e}}_y \cos\varphi_1\right) \\
&= r_2 \left(-\sin\theta_2 \cos\varphi_2 \sin\varphi_1 + \sin\theta_2 \sin\varphi_2 \cos\varphi_1\right) = r_2 \sin\theta_2 \sin(\varphi_2 - \varphi_1).
\end{aligned}$$

(c) The spherical components of $\mathbf{r}_1$ at P_2 can be found similarly. In fact, switching the indices "1" and "2" in the expressions of part (b) gives the desired formulas. ■

Example 1.4.3. To illustrate further the conversion of vectors from one coordinate system to another, consider a charge q that is located at the *cylindrical* coordinates $(a, \pi/3, -a)$. We want to find the *spherical* components of the electrostatic field $\mathbf{E}$ of this charge at a point P with *Cartesian* coordinates (a, a, a).

The most straightforward way of doing this is to convert all coordinates to Cartesian, find the field, and then take the dot products with appropriate unit vectors. The Cartesian coordinates of the charge are

$$\begin{aligned}
x_q &= \rho_q \cos\varphi_q = a \cos\left(\frac{\pi}{3}\right) = \tfrac{1}{2}a, \\
y_q &= \rho_q \sin\varphi_q = a \sin\left(\frac{\pi}{3}\right) = \frac{\sqrt{3}a}{2} = 0.866a, \\
z_q &= -a.
\end{aligned}$$

Thus,

$$\mathbf{r}-\mathbf{r}_q=(a-\tfrac{1}{2}a)\hat{\mathbf{e}}_x+(a-0.866a)\hat{\mathbf{e}}_y+(a-(-a))\hat{\mathbf{e}}_z=0.5a\hat{\mathbf{e}}_x+0.134a\hat{\mathbf{e}}_y+2a\hat{\mathbf{e}}_z$$

and

$$|\mathbf{r}-\mathbf{r}_q|=\sqrt{(0.5a)^2+(0.134a)^2+(2a)^2}=2.066a,$$

and the electric field at P can be written in terms of Cartesian unit vectors at P:

$$\begin{aligned}\mathbf{E}&=\frac{k_eq}{|\mathbf{r}-\mathbf{r}_q|^3}(\mathbf{r}-\mathbf{r}_q)=k_eq\frac{0.5a\hat{\mathbf{e}}_x+0.134a\hat{\mathbf{e}}_y+2a\hat{\mathbf{e}}_z}{(2.066a)^3}\\&=k_eq\frac{0.5\hat{\mathbf{e}}_x+0.134\hat{\mathbf{e}}_y+2\hat{\mathbf{e}}_z}{8.818a^2}=\frac{k_eq}{a^2}(0.0567\hat{\mathbf{e}}_x+0.0152\hat{\mathbf{e}}_y+0.2268\hat{\mathbf{e}}_z).\end{aligned}$$

To find the *spherical* components of the field at P, we first express the spherical unit vectors at P in terms of Cartesian unit vectors. For this, we need the spherical coordinates of P:

$$\begin{aligned}r&=\sqrt{a^2+a^2+a^2}=\sqrt{3}\,a=1.732a,\\\cos\theta&=\frac{z}{r}=\frac{a}{\sqrt{3}\,a}=\frac{1}{\sqrt{3}}=0.577\ \Rightarrow\ \theta=0.955,\\\tan\varphi&=\frac{y}{x}=\frac{a}{a}=1\ \Rightarrow\ \varphi=\frac{\pi}{4}=0.785.\end{aligned}$$

It now follows that

$$\begin{aligned}\hat{\mathbf{e}}_r&=\hat{\mathbf{e}}_x\sin\theta\cos\varphi+\hat{\mathbf{e}}_y\sin\theta\sin\varphi+\hat{\mathbf{e}}_z\cos\theta=0.577\hat{\mathbf{e}}_x+0.577\hat{\mathbf{e}}_y+0.577\hat{\mathbf{e}}_z,\\\hat{\mathbf{e}}_\theta&=\hat{\mathbf{e}}_x\cos\theta\cos\varphi+\hat{\mathbf{e}}_y\cos\theta\sin\varphi-\hat{\mathbf{e}}_z\sin\theta=0.408\hat{\mathbf{e}}_x+0.408\hat{\mathbf{e}}_y-0.816\hat{\mathbf{e}}_z,\\\hat{\mathbf{e}}_\varphi&=-\hat{\mathbf{e}}_x\sin\varphi+\hat{\mathbf{e}}_y\cos\varphi=-0.707\hat{\mathbf{e}}_x+0.707\hat{\mathbf{e}}_y.\end{aligned}$$

Now we take the dot product of $\mathbf{E}$ with these unit vectors to find its spherical components at P. The reader may first easily check that

$$\begin{aligned}&\hat{\mathbf{e}}_r\cdot\hat{\mathbf{e}}_x=0.577, &&\hat{\mathbf{e}}_r\cdot\hat{\mathbf{e}}_y=0.577, &&\hat{\mathbf{e}}_r\cdot\hat{\mathbf{e}}_z=0.577,\\&\hat{\mathbf{e}}_\theta\cdot\hat{\mathbf{e}}_x=0.408, &&\hat{\mathbf{e}}_\theta\cdot\hat{\mathbf{e}}_y=0.408, &&\hat{\mathbf{e}}_\theta\cdot\hat{\mathbf{e}}_z=-0.816,\\&\hat{\mathbf{e}}_\varphi\cdot\hat{\mathbf{e}}_x=-0.707, &&\hat{\mathbf{e}}_\varphi\cdot\hat{\mathbf{e}}_y=0.707, &&\hat{\mathbf{e}}_\varphi\cdot\hat{\mathbf{e}}_z=0.\end{aligned}$$

We can now finally calculate the field components:

$$\begin{aligned}E_r&=\mathbf{E}\cdot\hat{\mathbf{e}}_r=\frac{k_eq}{a^2}(0.0567\hat{\mathbf{e}}_r\cdot\hat{\mathbf{e}}_x+0.0152\hat{\mathbf{e}}_r\cdot\hat{\mathbf{e}}_y+0.2268\hat{\mathbf{e}}_r\cdot\hat{\mathbf{e}}_z)\\&=\frac{k_eq}{a^2}(0.0567\times0.577+0.0152\times0.577+0.2268\times0.577)=0.1724\frac{k_eq}{a^2},\\E_\theta&=\mathbf{E}\cdot\hat{\mathbf{e}}_\theta=\frac{k_eq}{a^2}(0.0567\hat{\mathbf{e}}_\theta\cdot\hat{\mathbf{e}}_x+0.0152\hat{\mathbf{e}}_\theta\cdot\hat{\mathbf{e}}_y+0.2268\hat{\mathbf{e}}_\theta\cdot\hat{\mathbf{e}}_z)\\&=\frac{k_eq}{a^2}(0.0567\times0.408+0.0152\times0.408-0.2268\times0.816)=-0.1558\frac{k_eq}{a^2},\\E_\varphi&=\mathbf{E}\cdot\hat{\mathbf{e}}_\varphi=\frac{k_eq}{a^2}(0.0567\hat{\mathbf{e}}_\varphi\cdot\hat{\mathbf{e}}_x+0.0152\hat{\mathbf{e}}_\varphi\cdot\hat{\mathbf{e}}_y+0.2268\hat{\mathbf{e}}_\varphi\cdot\hat{\mathbf{e}}_z)\\&=\frac{k_eq}{a^2}(-0.0567\times0.707+0.0152\times0.707)=-0.0294\frac{k_eq}{a^2}.\end{aligned}$$

The choice of Cartesian coordinates was the most straightforward one, but one can choose any other coordinate system to calculate the field and find the components in any other set of unit vectors. The reader is urged to try the other choices. ■

;B<C< D A A / ;* E / ; ; !
- , ;* ! , ! F < - A I !
C ! F! H ! J A , -!
8H 0 A ! KKK
' #/ / - 1 D H ! ,- H

3

Complex numbers and hyperbolic functions

This chapter is concerned with the representation and manipulation of complex numbers. Complex numbers pervade this book, underscoring their wide application in the mathematics of the physical sciences. The application of complex numbers to the description of physical systems is left until later chapters and only the basic tools are presented here.

3.1 The need for complex numbers

Although complex numbers occur in many branches of mathematics, they arise most directly out of solving polynomial equations. We examine a specific quadratic equation as an example.

Consider the quadratic equation

$$z^2 - 4z + 5 = 0. \tag{3.1}$$

Equation (3.1) has two solutions, z_1 and z_2, such that

$$(z - z_1)(z - z_2) = 0. \tag{3.2}$$

Using the familiar formula for the roots of a quadratic equation, (1.4), the solutions z_1 and z_2, written in brief as $z_{1,2}$, are

$$\begin{aligned} z_{1,2} &= \frac{4 \pm \sqrt{(-4)^2 - 4(1 \times 5)}}{2} \\ &= 2 \pm \frac{\sqrt{-4}}{2}. \end{aligned} \tag{3.3}$$

Both solutions contain the square root of a negative number. However, it is not true to say that there are no solutions to the quadratic equation. The *fundamental theorem of algebra* states that a quadratic equation will always have two solutions and these are in fact given by (3.3). The second term on the RHS of (3.3) is called an *imaginary* term since it contains the square root of a negative number;

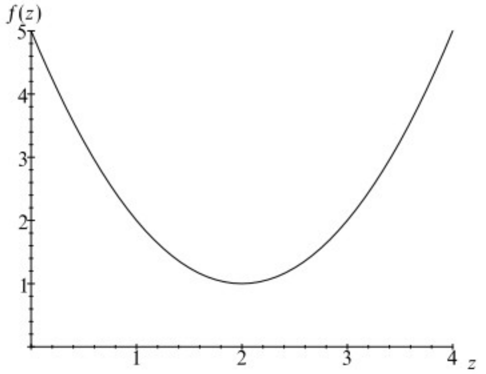

Figure 3.1 The function $f(z) = z^2 - 4z + 5$.

the first term is called a *real* term. The full solution is the sum of a real term and an imaginary term and is called a *complex number*. A plot of the function $f(z) = z^2 - 4z + 5$ is shown in figure 3.1. It will be seen that the plot does not intersect the z-axis, corresponding to the fact that the equation $f(z) = 0$ has no purely real solutions.

The choice of the symbol z for the quadratic variable was not arbitrary; the conventional representation of a complex number is z, where z is the sum of a real part x and i times an imaginary part y, i.e.

$$z = x + iy,$$

where i is used to denote the square root of -1. The real part x and the imaginary part y are usually denoted by $\operatorname{Re} z$ and $\operatorname{Im} z$ respectively. We note at this point that some physical scientists, engineers in particular, use j instead of i. However, for consistency, we will use i throughout this book.

In our particular example, $\sqrt{-4} = 2\sqrt{-1} = 2i$, and hence the two solutions of (3.1) are

$$z_{1,2} = 2 \pm \frac{2i}{2} = 2 \pm i.$$

Thus, here $x = 2$ and $y = \pm 1$.

For compactness a complex number is sometimes written in the form

$$z = (x, y),$$

where the components of z may be thought of as coordinates in an xy-plot. Such a plot is called an *Argand diagram* and is a common representation of complex numbers; an example is shown in figure 3.2.

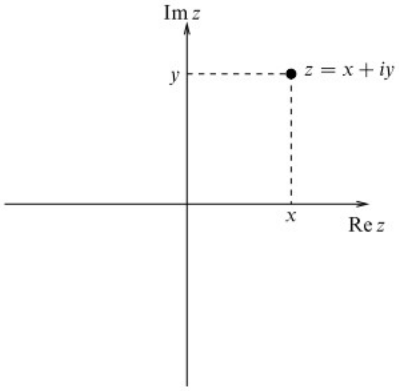

Figure 3.2 The Argand diagram.

Our particular example of a quadratic equation may be generalised readily to polynomials whose highest power (degree) is greater than 2, e.g. cubic equations (degree 3), quartic equations (degree 4) and so on. For a general polynomial $f(z)$, of degree n, the fundamental theorem of algebra states that the equation $f(z) = 0$ will have exactly n solutions. We will examine cases of higher-degree equations in subsection 3.4.3.

The remainder of this chapter deals with: the algebra and manipulation of complex numbers; their polar representation, which has advantages in many circumstances; complex exponentials and logarithms; the use of complex numbers in finding the roots of polynomial equations; and hyperbolic functions.

3.2 Manipulation of complex numbers

This section considers basic complex number manipulation. Some analogy may be drawn with vector manipulation (see chapter 7) but this section stands alone as an introduction.

3.2.1 Addition and subtraction

The addition of two complex numbers, z_1 and z_2, in general gives another complex number. The real components and the imaginary components are added separately and in a like manner to the familiar addition of real numbers:

$$z_1 + z_2 = (x_1 + iy_1) + (x_2 + iy_2) = (x_1 + x_2) + i(y_1 + y_2),$$

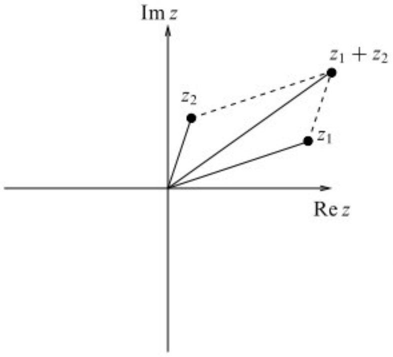

Figure 3.3 The addition of two complex numbers.

or in component notation

$$z_1 + z_2 = (x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2).$$

The Argand representation of the addition of two complex numbers is shown in figure 3.3.

By straightforward application of the commutativity and associativity of the real and imaginary parts separately, we can show that the addition of complex numbers is itself commutative and associative, i.e.

$$z_1 + z_2 = z_2 + z_1,$$
$$z_1 + (z_2 + z_3) = (z_1 + z_2) + z_3.$$

Thus it is immaterial in what order complex numbers are added.

►*Sum the complex numbers* $1 + 2i$, $3 - 4i$, $-2 + i$.

Summing the real terms we obtain

$$1 + 3 - 2 = 2,$$

and summing the imaginary terms we obtain

$$2i - 4i + i = -i.$$

Hence

$$(1 + 2i) + (3 - 4i) + (-2 + i) = 2 - i. \blacktriangleleft$$

The subtraction of complex numbers is very similar to their addition. As in the case of real numbers, if two identical complex numbers are subtracted then the result is zero.

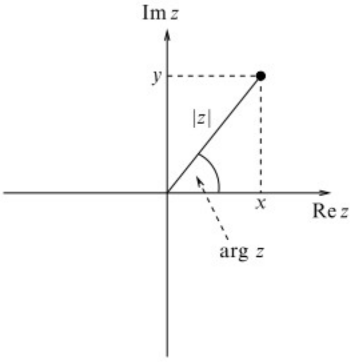

Figure 3.4 The modulus and argument of a complex number.

3.2.2 Modulus and argument

The modulus of the complex number z is denoted by $|z|$ and is defined as

$$|z| = \sqrt{x^2 + y^2}. \tag{3.4}$$

Hence the modulus of the complex number is the distance of the corresponding point from the origin in the Argand diagram, as may be seen in figure 3.4.

The argument of the complex number z is denoted by $\arg z$ and is defined as

$$\arg z = \tan^{-1}\left(\frac{y}{x}\right). \tag{3.5}$$

It can be seen that $\arg z$ is the angle that the line joining the origin to z on the Argand diagram makes with the positive x-axis. The anticlockwise direction is taken to be positive by convention. The angle $\arg z$ is shown in figure 3.4. Account must be taken of the signs of x and y individually in determining in which quadrant $\arg z$ lies. Thus, for example, if x and y are both negative then $\arg z$ lies in the range $-\pi < \arg z < -\pi/2$ rather than in the first quadrant ($0 < \arg z < \pi/2$), though both cases give the same value for the ratio of y to x.

►*Find the modulus and the argument of the complex number $z = 2 - 3i$.*

Using (3.4), the modulus is given by

$$|z| = \sqrt{2^2 + (-3)^2} = \sqrt{13}.$$

Using (3.5), the argument is given by

$$\arg z = \tan^{-1}\left(-\tfrac{3}{2}\right).$$

The two angles whose tangents equal -1.5 are -0.9828 rad and 2.1588 rad. Since $x = 2$ and $y = -3$, z clearly lies in the fourth quadrant; therefore $\arg z = -0.9828$ is the appropriate answer. ◄

3.2.3 Multiplication

Complex numbers may be multiplied together and in general give a complex number as the result. The product of two complex numbers z_1 and z_2 is found by multiplying them out in full and remembering that $i^2 = -1$, i.e.

$$\begin{aligned} z_1 z_2 &= (x_1 + iy_1)(x_2 + iy_2) \\ &= x_1 x_2 + i x_1 y_2 + i y_1 x_2 + i^2 y_1 y_2 \\ &= (x_1 x_2 - y_1 y_2) + i(x_1 y_2 + y_1 x_2). \end{aligned} \tag{3.6}$$

►*Multiply the complex numbers* $z_1 = 3 + 2i$ *and* $z_2 = -1 - 4i$.

By direct multiplication we find

$$\begin{aligned} z_1 z_2 &= (3 + 2i)(-1 - 4i) \\ &= -3 - 2i - 12i - 8i^2 \\ &= 5 - 14i. \blacktriangleleft \end{aligned} \tag{3.7}$$

The multiplication of complex numbers is both commutative and associative, i.e.

$$z_1 z_2 = z_2 z_1, \tag{3.8}$$

$$(z_1 z_2) z_3 = z_1 (z_2 z_3). \tag{3.9}$$

The product of two complex numbers also has the simple properties

$$|z_1 z_2| = |z_1||z_2|, \tag{3.10}$$

$$\arg(z_1 z_2) = \arg z_1 + \arg z_2. \tag{3.11}$$

These relations are derived in subsection 3.3.1.

►*Verify that (3.10) holds for the product of* $z_1 = 3 + 2i$ *and* $z_2 = -1 - 4i$.

From (3.7)

$$|z_1 z_2| = |5 - 14i| = \sqrt{5^2 + (-14)^2} = \sqrt{221}.$$

We also find

$$\begin{aligned} |z_1| &= \sqrt{3^2 + 2^2} = \sqrt{13}, \\ |z_2| &= \sqrt{(-1)^2 + (-4)^2} = \sqrt{17}, \end{aligned}$$

and hence

$$|z_1||z_2| = \sqrt{13}\sqrt{17} = \sqrt{221} = |z_1 z_2|. \blacktriangleleft$$

We now examine the effect on a complex number z of multiplying it by ± 1 and $\pm i$. These four multipliers have modulus unity and we can see immediately from (3.10) that multiplying z by another complex number of unit modulus gives a product with the same modulus as z. We can also see from (3.11) that if we

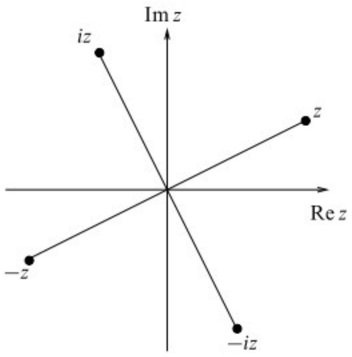

Figure 3.5 Multiplication of a complex number by ± 1 and $\pm i$.

multiply z by a complex number then the argument of the product is the sum of the argument of z and the argument of the multiplier. Hence multiplying z by unity (which has argument zero) leaves z unchanged in both modulus and argument, i.e. z is completely unaltered by the operation. Multiplying by -1 (which has argument π) leads to rotation, through an angle π, of the line joining the origin to z in the Argand diagram. Similarly, multiplication by i or $-i$ leads to corresponding rotations of $\pi/2$ or $-\pi/2$ respectively. This geometrical interpretation of multiplication is shown in figure 3.5.

►*Using the geometrical interpretation of multiplication by i, find the product* $i(1-i)$.

The complex number $1-i$ has argument $-\pi/4$ and modulus $\sqrt{2}$. Thus, using (3.10) and (3.11), its product with i has argument $+\pi/4$ and unchanged modulus $\sqrt{2}$. The complex number with modulus $\sqrt{2}$ and argument $+\pi/4$ is $1+i$ and so

$$i(1-i) = 1+i,$$

as is easily verified by direct multiplication. ◄

The division of two complex numbers is similar to their multiplication but requires the notion of the complex conjugate (see the following subsection) and so discussion is postponed until subsection 3.2.5.

3.2.4 Complex conjugate

If z has the convenient form $x+iy$ then the complex conjugate, denoted by z^*, may be found simply by changing the sign of the imaginary part, i.e. if $z = x+iy$ then $z^* = x - iy$. More generally, we may define the complex conjugate of z as the (complex) number having the same magnitude as z that when multiplied by z leaves a real result, i.e. there is no imaginary component in the product.

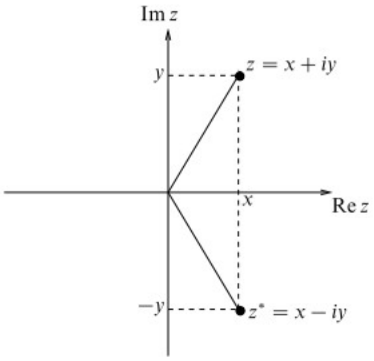

Figure 3.6 The complex conjugate as a mirror image in the real axis.

In the case where z can be written in the form $x + iy$ it is easily verified, by direct multiplication of the components, that the product zz^* gives a real result:

$$zz^* = (x + iy)(x - iy) = x^2 - ixy + ixy - i^2y^2 = x^2 + y^2 = |z|^2.$$

Complex conjugation corresponds to a reflection of z in the real axis of the Argand diagram, as may be seen in figure 3.6.

►*Find the complex conjugate of* $z = a + 2i + 3ib$.

The complex number is written in the standard form

$$z = a + i(2 + 3b);$$

then, replacing i by $-i$, we obtain

$$z^* = a - i(2 + 3b). ◄$$

In some cases, however, it may not be simple to rearrange the expression for z into the standard form $x + iy$. Nevertheless, given two complex numbers, z_1 and z_2, it is straightforward to show that the complex conjugate of their sum (or difference) is equal to the sum (or difference) of their complex conjugates, i.e. $(z_1 \pm z_2)^* = z_1^* \pm z_2^*$. Similarly, it may be shown that the complex conjugate of the product (or quotient) of z_1 and z_2 is equal to the product (or quotient) of their complex conjugates, i.e. $(z_1z_2)^* = z_1^*z_2^*$ and $(z_1/z_2)^* = z_1^*/z_2^*$.

Using these results, it can be deduced that, no matter how complicated the expression, its complex conjugate may *always* be found by replacing every i by $-i$. To apply this rule, however, we must always ensure that all complex parts are first written out in full, so that no i's are hidden.

►*Find the complex conjugate of the complex number $z = w^{(3y+2ix)}$, where $w = x + 5i$.*

Although we do not discuss complex powers until section 3.5, the simple rule given above still enables us to find the complex conjugate of z.

In this case w itself contains real and imaginary components and so must be written out in full, i.e.

$$z = w^{3y+2ix} = (x + 5i)^{3y+2ix}.$$

Now we can replace each i by $-i$ to obtain

$$z^* = (x - 5i)^{(3y-2ix)}.$$

It can be shown that the product zz^* is real, as required. ◄

The following properties of the complex conjugate are easily proved and others may be derived from them. If $z = x + iy$ then

$$(z^*)^* = z, \tag{3.12}$$

$$z + z^* = 2\,\mathrm{Re}\,z = 2x, \tag{3.13}$$

$$z - z^* = 2i\,\mathrm{Im}\,z = 2iy, \tag{3.14}$$

$$\frac{z}{z^*} = \left(\frac{x^2 - y^2}{x^2 + y^2}\right) + i\left(\frac{2xy}{x^2 + y^2}\right). \tag{3.15}$$

The derivation of this last relation relies on the results of the following subsection.

3.2.5 Division

The division of two complex numbers z_1 and z_2 bears some similarity to their multiplication. Writing the quotient in component form we obtain

$$\frac{z_1}{z_2} = \frac{x_1 + iy_1}{x_2 + iy_2}. \tag{3.16}$$

In order to separate the real and imaginary components of the quotient, we multiply both numerator and denominator by the complex conjugate of the denominator. By definition, this process will leave the denominator as a real quantity. Equation (3.16) gives

$$\begin{aligned}\frac{z_1}{z_2} &= \frac{(x_1 + iy_1)(x_2 - iy_2)}{(x_2 + iy_2)(x_2 - iy_2)} = \frac{(x_1x_2 + y_1y_2) + i(x_2y_1 - x_1y_2)}{x_2^2 + y_2^2}\\ &= \frac{x_1x_2 + y_1y_2}{x_2^2 + y_2^2} + i\frac{x_2y_1 - x_1y_2}{x_2^2 + y_2^2}.\end{aligned}$$

Hence we have separated the quotient into real and imaginary components, as required.

In the special case where $z_2 = z_1^*$, so that $x_2 = x_1$ and $y_2 = -y_1$, the general result reduces to (3.15).

►*Express z in the form $x + iy$, when*

$$z = \frac{3 - 2i}{-1 + 4i}.$$

Multiplying numerator and denominator by the complex conjugate of the denominator we obtain

$$\begin{aligned} z &= \frac{(3-2i)(-1-4i)}{(-1+4i)(-1-4i)} = \frac{-11-10i}{17} \\ &= -\frac{11}{17} - \frac{10}{17}i. \blacktriangleleft \end{aligned}$$

In analogy to (3.10) and (3.11), which describe the multiplication of two complex numbers, the following relations apply to division:

$$\left|\frac{z_1}{z_2}\right| = \frac{|z_1|}{|z_2|}, \tag{3.17}$$

$$\arg\left(\frac{z_1}{z_2}\right) = \arg z_1 - \arg z_2. \tag{3.18}$$

The proof of these relations is left until subsection 3.3.1.

3.3 Polar representation of complex numbers

Although considering a complex number as the sum of a real and an imaginary part is often useful, sometimes the *polar representation* proves easier to manipulate. This makes use of the complex exponential function, which is defined by

$$e^z = \exp z \equiv 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots. \tag{3.19}$$

Strictly speaking it is the function $\exp z$ that is defined by (3.19). The number e is the value of $\exp(1)$, i.e. it is just a number. However, it may be shown that e^z and $\exp z$ are equivalent when z is real and rational and mathematicians then *define* their equivalence for irrational and complex z. For the purposes of this book we will not concern ourselves further with this mathematical nicety but, rather, assume that (3.19) is valid for all z. We also note that, using (3.19), by multiplying together the appropriate series we may show that (see chapter 24)

$$e^{z_1} e^{z_2} = e^{z_1 + z_2}, \tag{3.20}$$

which is analogous to the familiar result for exponentials of real numbers.

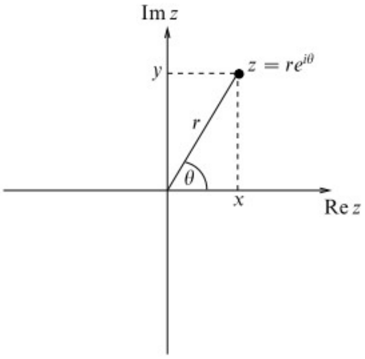

Figure 3.7 The polar representation of a complex number.

From (3.19), it immediately follows that for $z = i\theta$, θ real,

$$e^{i\theta} = 1 + i\theta - \frac{\theta^2}{2!} - \frac{i\theta^3}{3!} + \cdots \tag{3.21}$$

$$= 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \cdots + i\left(\theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \cdots\right) \tag{3.22}$$

and hence that

$$e^{i\theta} = \cos\theta + i\sin\theta, \tag{3.23}$$

where the last equality follows from the series expansions of the sine and cosine functions (see subsection 4.6.3). This last relationship is called *Euler's equation.* It also follows from (3.23) that

$$e^{in\theta} = \cos n\theta + i\sin n\theta$$

for all n. From Euler's equation (3.23) and figure 3.7 we deduce that

$$\begin{aligned} re^{i\theta} &= r(\cos\theta + i\sin\theta) \\ &= x + iy. \end{aligned}$$

Thus a complex number may be represented in the polar form

$$z = re^{i\theta}. \tag{3.24}$$

Referring again to figure 3.7, we can identify r with $|z|$ and θ with $\arg z$. The simplicity of the representation of the modulus and argument is one of the main reasons for using the polar representation. The angle θ lies conventionally in the range $-\pi < \theta \le \pi$, but, since rotation by θ is the same as rotation by $2n\pi + \theta$, where n is any integer,

$$re^{i\theta} \equiv re^{i(\theta+2n\pi)}.$$

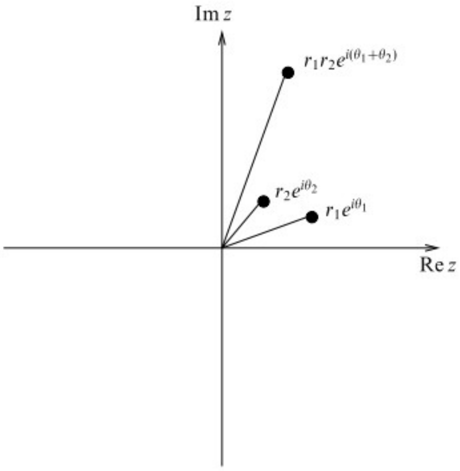

Figure 3.8 The multiplication of two complex numbers. In this case r_1 and r_2 are both greater than unity.

The algebra of the polar representation is different from that of the real and imaginary component representation, though, of course, the results are identical. Some operations prove much easier in the polar representation, others much more complicated. The best representation for a particular problem must be determined by the manipulation required.

3.3.1 Multiplication and division in polar form

Multiplication and division in polar form are particularly simple. The product of $z_1 = r_1 e^{i\theta_1}$ and $z_2 = r_2 e^{i\theta_2}$ is given by

$$\begin{aligned} z_1 z_2 &= r_1 e^{i\theta_1} r_2 e^{i\theta_2} \\ &= r_1 r_2 e^{i(\theta_1+\theta_2)}. \end{aligned} \tag{3.25}$$

The relations $|z_1 z_2| = |z_1||z_2|$ and $\arg(z_1 z_2) = \arg z_1 + \arg z_2$ follow immediately. An example of the multiplication of two complex numbers is shown in figure 3.8.

Division is equally simple in polar form; the quotient of z_1 and z_2 is given by

$$\frac{z_1}{z_2} = \frac{r_1 e^{i\theta_1}}{r_2 e^{i\theta_2}} = \frac{r_1}{r_2} e^{i(\theta_1-\theta_2)}. \tag{3.26}$$

The relations $|z_1/z_2| = |z_1|/|z_2|$ and $\arg(z_1/z_2) = \arg z_1 - \arg z_2$ are again

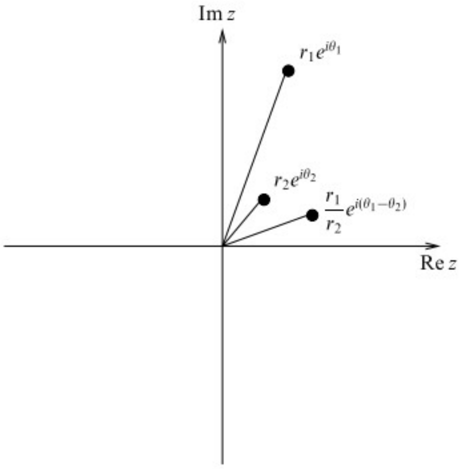

Figure 3.9 The division of two complex numbers. As in the previous figure, r_1 and r_2 are both greater than unity.

immediately apparent. The division of two complex numbers in polar form is shown in figure 3.9.

3.4 de Moivre's theorem

We now derive an extremely important theorem. Since $\left(e^{i\theta}\right)^n = e^{in\theta}$, we have

$$(\cos\theta + i\sin\theta)^n = \cos n\theta + i\sin n\theta, \tag{3.27}$$

where the identity $e^{in\theta} = \cos n\theta + i\sin n\theta$ follows from the series definition of $e^{in\theta}$ (see (3.21)). This result is called *de Moivre's theorem* and is often used in the manipulation of complex numbers. The theorem is valid for all n whether real, imaginary or complex.

There are numerous applications of de Moivre's theorem but this section examines just three: proofs of trigonometric identities; finding the nth roots of unity; and solving polynomial equations with complex roots.

3.4.1 Trigonometric identities

The use of de Moivre's theorem in finding trigonometric identities is best illustrated by example. We consider the expression of a multiple-angle function in terms of a polynomial in the single-angle function, and its converse.

►*Express* $\sin 3\theta$ *and* $\cos 3\theta$ *in terms of powers of* $\cos\theta$ *and* $\sin\theta$.

Using de Moivre's theorem,

$$\begin{aligned}\cos 3\theta + i\sin 3\theta &= (\cos\theta + i\sin\theta)^3\\ &= (\cos^3\theta - 3\cos\theta\sin^2\theta) + i(3\sin\theta\cos^2\theta - \sin^3\theta).\end{aligned} \tag{3.28}$$

We can equate the real and imaginary coefficients separately, i.e.

$$\begin{aligned}\cos 3\theta &= \cos^3\theta - 3\cos\theta\sin^2\theta\\ &= 4\cos^3\theta - 3\cos\theta\end{aligned} \tag{3.29}$$

and

$$\begin{aligned}\sin 3\theta &= 3\sin\theta\cos^2\theta - \sin^3\theta\\ &= 3\sin\theta - 4\sin^3\theta. \blacktriangleleft\end{aligned}$$

This method can clearly be applied to finding power expansions of $\cos n\theta$ and $\sin n\theta$ for any positive integer n.

The converse process uses the following properties of $z = e^{i\theta}$,

$$z^n + \frac{1}{z^n} = 2\cos n\theta, \tag{3.30}$$

$$z^n - \frac{1}{z^n} = 2i\sin n\theta. \tag{3.31}$$

These equalities follow from simple applications of de Moivre's theorem, i.e.

$$\begin{aligned}z^n + \frac{1}{z^n} &= (\cos\theta + i\sin\theta)^n + (\cos\theta + i\sin\theta)^{-n}\\ &= \cos n\theta + i\sin n\theta + \cos(-n\theta) + i\sin(-n\theta)\\ &= \cos n\theta + i\sin n\theta + \cos n\theta - i\sin n\theta\\ &= 2\cos n\theta\end{aligned}$$

and

$$\begin{aligned}z^n - \frac{1}{z^n} &= (\cos\theta + i\sin\theta)^n - (\cos\theta + i\sin\theta)^{-n}\\ &= \cos n\theta + i\sin n\theta - \cos n\theta + i\sin n\theta\\ &= 2i\sin n\theta.\end{aligned}$$

In the particular case where $n = 1$,

$$z + \frac{1}{z} = e^{i\theta} + e^{-i\theta} = 2\cos\theta, \tag{3.32}$$

$$z - \frac{1}{z} = e^{i\theta} - e^{-i\theta} = 2i\sin\theta. \tag{3.33}$$

►*Find an expression for* $\cos^3\theta$ *in terms of* $\cos 3\theta$ *and* $\cos\theta$.

Using (3.32),

$$\begin{aligned}\cos^3\theta &= \frac{1}{2^3}\left(z+\frac{1}{z}\right)^3\\ &= \frac{1}{8}\left(z^3+3z+\frac{3}{z}+\frac{1}{z^3}\right)\\ &= \frac{1}{8}\left(z^3+\frac{1}{z^3}\right)+\frac{3}{8}\left(z+\frac{1}{z}\right).\end{aligned}$$

Now using (3.30) and (3.32), we find

$$\cos^3\theta = \tfrac{1}{4}\cos 3\theta + \tfrac{3}{4}\cos\theta. \blacktriangleleft$$

This result happens to be a simple rearrangement of (3.29), but cases involving larger values of n are better handled using this direct method than by rearranging polynomial expansions of multiple-angle functions.

3.4.2 Finding the nth roots of unity

The equation $z^2 = 1$ has the familiar solutions $z = \pm 1$. However, now that we have introduced the concept of complex numbers we can solve the general equation $z^n = 1$. Recalling the fundamental theorem of algebra, we know that the equation has n solutions. In order to proceed we rewrite the equation as

$$z^n = e^{2ik\pi},$$

where k is any integer. Now taking the nth root of each side of the equation we find

$$z = e^{2ik\pi/n}.$$

Hence, the solutions of $z^n = 1$ are

$$z_{1,2,\ldots,n} = 1,\ \ e^{2i\pi/n},\ \ \ldots,\ \ e^{2i(n-1)\pi/n},$$

corresponding to the values $0, 1, 2, \ldots, n-1$ for k. Larger integer values of k do not give new solutions, since the roots already listed are simply cyclically repeated for $k = n, n+1, n+2$, etc.

►*Find the solutions to the equation* $z^3 = 1$.

By applying the above method we find

$$z = e^{2ik\pi/3}.$$

Hence the three solutions are $z_1 = e^{0i} = 1$, $z_2 = e^{2i\pi/3}$, $z_3 = e^{4i\pi/3}$. We note that, as expected, the next solution, for which $k = 3$, gives $z_4 = e^{6i\pi/3} = 1 = z_1$, so that there are only three separate solutions. ◄

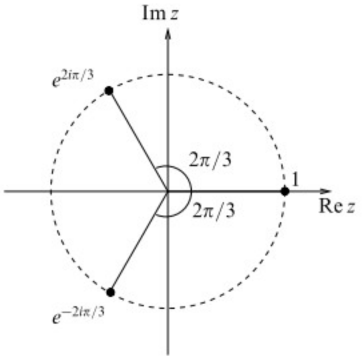

Figure 3.10 The solutions of $z^3 = 1$.

Not surprisingly, given that $|z^3| = |z|^3$ from (3.10), all the roots of unity have unit modulus, i.e. they all lie on a circle in the Argand diagram of unit radius. The three roots are shown in figure 3.10.

The cube roots of unity are often written 1, ω and ω^2. The properties $\omega^3 = 1$ and $1 + \omega + \omega^2 = 0$ are easily proved.

3.4.3 Solving polynomial equations

A third application of de Moivre's theorem is to the solution of polynomial equations. Complex equations in the form of a polynomial relationship must first be solved for z in a similar fashion to the method for finding the roots of real polynomial equations. Then the complex roots of z may be found.

►*Solve the equation* $z^6 - z^5 + 4z^4 - 6z^3 + 2z^2 - 8z + 8 = 0$.

We first factorise to give

$$(z^3 - 2)(z^2 + 4)(z - 1) = 0.$$

Hence $z^3 = 2$ or $z^2 = -4$ or $z = 1$. The solutions to the quadratic equation are $z = \pm 2i$; to find the complex cube roots, we first write the equation in the form

$$z^3 = 2 = 2e^{2ik\pi},$$

where k is any integer. If we now take the cube root, we get

$$z = 2^{1/3}e^{2ik\pi/3}.$$

To avoid the duplication of solutions, we use the fact that $-\pi < \arg z \leq \pi$ and find

$$
\begin{aligned}
z_1 &= 2^{1/3}, \\
z_2 &= 2^{1/3}e^{2\pi i/3} = 2^{1/3}\left(-\frac{1}{2} + \frac{\sqrt{3}}{2}i\right), \\
z_3 &= 2^{1/3}e^{-2\pi i/3} = 2^{1/3}\left(-\frac{1}{2} - \frac{\sqrt{3}}{2}i\right).
\end{aligned}
$$

The complex numbers z_1, z_2 and z_3, together with $z_4 = 2i$, $z_5 = -2i$ and $z_6 = 1$ are the solutions to the original polynomial equation.

As expected from the fundamental theorem of algebra, we find that the total number of complex roots (six, in this case) is equal to the largest power of z in the polynomial. ◀

A useful result is that the roots of a polynomial with real coefficients occur in conjugate pairs (i.e. if z_1 is a root, then z_1^* is a second distinct root, unless z_1 is real). This may be proved as follows. Let the polynomial equation of which z is a root be

$$a_n z^n + a_{n-1}z^{n-1} + \cdots + a_1 z + a_0 = 0.$$

Taking the complex conjugate of this equation,

$$a_n^*(z^*)^n + a_{n-1}^*(z^*)^{n-1} + \cdots + a_1^* z^* + a_0^* = 0.$$

But the a_n are real, and so z^* satisfies

$$a_n(z^*)^n + a_{n-1}(z^*)^{n-1} + \cdots + a_1 z^* + a_0 = 0,$$

and is also a root of the original equation.

3.5 Complex logarithms and complex powers

The concept of a complex exponential has already been introduced in section 3.3, where it was assumed that the definition of an exponential as a series was valid for complex numbers as well as for real numbers. Similarly we can define the logarithm of a complex number and we can use complex numbers as exponents.

Let us denote the natural logarithm of a complex number z by $w = \mathrm{Ln}\, z$, where the notation Ln will be explained shortly. Thus, w must satisfy

$$z = e^w.$$

Using (3.20), we see that

$$z_1 z_2 = e^{w_1}e^{w_2} = e^{w_1+w_2},$$

and taking logarithms of both sides we find

$$\mathrm{Ln}\,(z_1 z_2) = w_1 + w_2 = \mathrm{Ln}\, z_1 + \mathrm{Ln}\, z_2, \tag{3.34}$$

which shows that the familiar rule for the logarithm of the product of two real numbers also holds for complex numbers.

We may use (3.34) to investigate further the properties of $\mathrm{Ln}\, z$. We have already noted that the argument of a complex number is multivalued, i.e. $\arg z = \theta + 2n\pi$, where n is any integer. Thus, in polar form, the complex number z should strictly be written as

$$z = re^{i(\theta+2n\pi)}.$$

Taking the logarithm of both sides, and using (3.34), we find

$$\mathrm{Ln}\, z = \ln r + i(\theta + 2n\pi), \tag{3.35}$$

where $\ln r$ is the natural logarithm of the real positive quantity r and so is written normally. Thus from (3.35) we see that $\mathrm{Ln}\, z$ is itself multivalued. To avoid this multivalued behaviour it is conventional to define another function $\ln z$, the *principal value* of $\mathrm{Ln}\, z$, which is obtained from $\mathrm{Ln}\, z$ by restricting the argument of z to lie in the range $-\pi < \theta \leq \pi$.

►*Evaluate* $\mathrm{Ln}\,(-i)$.

By rewriting $-i$ as a complex exponential, we find

$$\mathrm{Ln}\,(-i) = \mathrm{Ln}\,\left[e^{i(-\pi/2+2n\pi)}\right] = i(-\pi/2 + 2n\pi),$$

where n is any integer. Hence $\mathrm{Ln}\,(-i) = -i\pi/2,\ 3i\pi/2,\ \ldots$. We note that $\ln(-i)$, the principal value of $\mathrm{Ln}\,(-i)$, is given by $\ln(-i) = -i\pi/2$. ◄

If z and t are both complex numbers then the zth power of t is defined by

$$t^z = e^{z\,\mathrm{Ln}\, t}.$$

Since $\mathrm{Ln}\, t$ is multivalued, so too is this definition.

►*Simplify the expression* $z = i^{-2i}$.

Firstly we take the logarithm of both sides of the equation to give

$$\mathrm{Ln}\, z = -2i\,\mathrm{Ln}\, i.$$

Now inverting the process we find

$$e^{\mathrm{Ln}\, z} = z = e^{-2i\,\mathrm{Ln}\, i}.$$

We can write $i = e^{i(\pi/2+2n\pi)}$, where n is any integer, and hence

$$\begin{aligned} \mathrm{Ln}\, i &= \mathrm{Ln}\,\left[e^{i\left(\pi/2+2n\pi\right)}\right] \\ &= i\left(\pi/2 + 2n\pi\right). \end{aligned}$$

We can now simplify z to give

$$\begin{aligned} i^{-2i} &= e^{-2i\times i\left(\pi/2+2n\pi\right)} \\ &= e^{(\pi+4n\pi)}, \end{aligned}$$

which, perhaps surprisingly, is a real quantity rather than a complex one. ◄

Complex powers and the logarithms of complex numbers are discussed further in chapter 24.

3.6 Applications to differentiation and integration

We can use the exponential form of a complex number together with de Moivre's theorem (see section 3.4) to simplify the differentiation of trigonometric functions.

►*Find the derivative with respect to x of* $e^{3x}\cos 4x$.

We could differentiate this function straightforwardly using the product rule (see subsection 2.1.2). However, an alternative method in this case is to use a complex exponential. Let us consider the complex number

$$z = e^{3x}(\cos 4x + i\sin 4x) = e^{3x}e^{4ix} = e^{(3+4i)x},$$

where we have used de Moivre's theorem to rewrite the trigonometric functions as a complex exponential. This complex number has $e^{3x}\cos 4x$ as its real part. Now, differentiating z with respect to x we obtain

$$\frac{dz}{dx} = (3+4i)e^{(3+4i)x} = (3+4i)e^{3x}(\cos 4x + i\sin 4x), \qquad (3.36)$$

where we have again used de Moivre's theorem. Equating real parts we then find

$$\frac{d}{dx}\left(e^{3x}\cos 4x\right) = e^{3x}(3\cos 4x - 4\sin 4x).$$

By equating the imaginary parts of (3.36), we also obtain, as a bonus,

$$\frac{d}{dx}\left(e^{3x}\sin 4x\right) = e^{3x}(4\cos 4x + 3\sin 4x). \blacktriangleleft$$

In a similar way the complex exponential can be used to evaluate integrals containing trigonometric and exponential functions.

►*Evaluate the integral* $I = \int e^{ax}\cos bx\, dx$.

Let us consider the integrand as the real part of the complex number

$$e^{ax}(\cos bx + i\sin bx) = e^{ax}e^{ibx} = e^{(a+ib)x},$$

where we use de Moivre's theorem to rewrite the trigonometric functions as a complex exponential. Integrating we find

$$\begin{aligned}\int e^{(a+ib)x}\,dx &= \frac{e^{(a+ib)x}}{a+ib} + c\\ &= \frac{(a-ib)e^{(a+ib)x}}{(a-ib)(a+ib)} + c\\ &= \frac{e^{ax}}{a^2+b^2}\left(ae^{ibx} - ibe^{ibx}\right) + c, \qquad (3.37)\end{aligned}$$

where the constant of integration c is in general complex. Denoting this constant by $c = c_1 + ic_2$ and equating real parts in (3.37) we obtain

$$I = \int e^{ax}\cos bx\,dx = \frac{e^{ax}}{a^2+b^2}(a\cos bx + b\sin bx) + c_1,$$

which agrees with result (2.37) found using integration by parts. Equating imaginary parts in (3.37) we obtain, as a bonus,

$$J = \int e^{ax}\sin bx\,dx = \frac{e^{ax}}{a^2+b^2}(a\sin bx - b\cos bx) + c_2. \blacktriangleleft$$

3.7 Hyperbolic functions

The *hyperbolic functions* are the complex analogues of the trigonometric functions. The analogy may not be immediately apparent and their definitions may appear at first to be somewhat arbitrary. However, careful examination of their properties reveals the purpose of the definitions. For instance, their close relationship with the trigonometric functions, both in their identities and in their calculus, means that many of the familiar properties of trigonometric functions can also be applied to the hyperbolic functions. Further, hyperbolic functions occur regularly, and so giving them special names is a notational convenience.

3.7.1 Definitions

The two fundamental hyperbolic functions are $\cosh x$ and $\sinh x$, which, as their names suggest, are the hyperbolic equivalents of $\cos x$ and $\sin x$. They are defined by the following relations:

$$\cosh x = \tfrac{1}{2}(e^x + e^{-x}), \tag{3.38}$$

$$\sinh x = \tfrac{1}{2}(e^x - e^{-x}). \tag{3.39}$$

Note that $\cosh x$ is an even function and $\sinh x$ is an odd function. By analogy with the trigonometric functions, the remaining hyperbolic functions are

$$\tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \tag{3.40}$$

$$\operatorname{sech} x = \frac{1}{\cosh x} = \frac{2}{e^x + e^{-x}}, \tag{3.41}$$

$$\operatorname{cosech} x = \frac{1}{\sinh x} = \frac{2}{e^x - e^{-x}}, \tag{3.42}$$

$$\coth x = \frac{1}{\tanh x} = \frac{e^x + e^{-x}}{e^x - e^{-x}}. \tag{3.43}$$

All the hyperbolic functions above have been defined in terms of the real variable x. However, this was simply so that they may be plotted (see figures 3.11–3.13); the definitions are equally valid for any complex number z.

3.7.2 Hyperbolic–trigonometric analogies

In the previous subsections we have alluded to the analogy between trigonometric and hyperbolic functions. Here, we discuss the close relationship between the two groups of functions.

Recalling (3.32) and (3.33) we find

$$\cos ix = \tfrac{1}{2}(e^x + e^{-x}),$$

$$\sin ix = \tfrac{1}{2}i(e^x - e^{-x}).$$

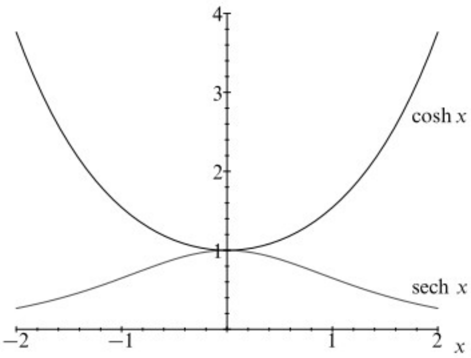

Figure 3.11 Graphs of cosh x and sech$x$.

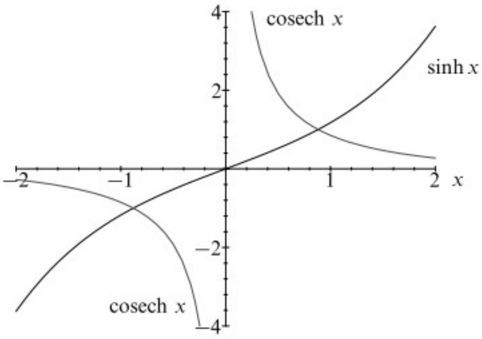

Figure 3.12 Graphs of sinh x and cosech$x$.

Hence, by the definitions given in the previous subsection,

$$\cosh x = \cos ix, \tag{3.44}$$

$$i \sinh x = \sin ix, \tag{3.45}$$

$$\cos x = \cosh ix, \tag{3.46}$$

$$i \sin x = \sinh ix. \tag{3.47}$$

These useful equations make the relationship between hyperbolic and trigono-

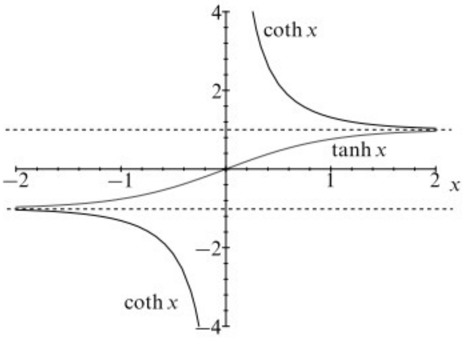

Figure 3.13 Graphs of tanh x and coth x.

metric functions transparent. The similarity in their calculus is discussed further in subsection 3.7.6.

3.7.3 Identities of hyperbolic functions

The analogies between trigonometric functions and hyperbolic functions having been established, we should not be surprised that all the trigonometric identities also hold for hyperbolic functions, with the following modification. Wherever $\sin^2 x$ occurs it must be replaced by $-\sinh^2 x$, and vice versa. Note that this replacement is necessary even if the $\sin^2 x$ is hidden, e.g. $\tan^2 x = \sin^2 x/\cos^2 x$ and so must be replaced by $(-\sinh^2 x/\cosh^2 x) = -\tanh^2 x$.

►*Find the hyperbolic identity analogous to* $\cos^2 x + \sin^2 x = 1$.

Using the rules stated above $\cos^2 x$ is replaced by $\cosh^2 x$, and $\sin^2 x$ by $-\sinh^2 x$, and so the identity becomes

$$\cosh^2 x - \sinh^2 x = 1.$$

This can be verified by direct substitution, using the definitions of $\cosh x$ and $\sinh x$; see (3.38) and (3.39). ◄

Some other identities that can be proved in a similar way are

$$\text{sech}^2 x = 1 - \tanh^2 x, \tag{3.48}$$

$$\text{cosech}^2 x = \coth^2 x - 1, \tag{3.49}$$

$$\sinh 2x = 2 \sinh x \cosh x, \tag{3.50}$$

$$\cosh 2x = \cosh^2 x + \sinh^2 x. \tag{3.51}$$

3.7.4 Solving hyperbolic equations

When we are presented with a hyperbolic equation to solve, we may proceed by analogy with the solution of trigonometric equations. However, it is almost always easier to express the equation directly in terms of exponentials.

►*Solve the hyperbolic equation* $\cosh x - 5\sinh x - 5 = 0$.

Substituting the definitions of the hyperbolic functions we obtain

$$\tfrac{1}{2}(e^x + e^{-x}) - \tfrac{5}{2}(e^x - e^{-x}) - 5 = 0.$$

Rearranging, and then multiplying through by $-e^x$, gives in turn

$$-2e^x + 3e^{-x} - 5 = 0$$

and

$$2e^{2x} + 5e^x - 3 = 0.$$

Now we can factorise and solve:

$$(2e^x - 1)(e^x + 3) = 0.$$

Thus $e^x = 1/2$ or $e^x = -3$. Hence $x = -\ln 2$ or $x = \ln(-3)$. The interpretation of the logarithm of a negative number has been discussed in section 3.5. ◄

3.7.5 Inverses of hyperbolic functions

Just like trigonometric functions, hyperbolic functions have inverses. If $y = \cosh x$ then $x = \cosh^{-1} y$, which serves as a definition of the inverse. By using the fundamental definitions of hyperbolic functions, we can find closed-form expressions for their inverses. This is best illustrated by example.

►*Find a closed-form expression for the inverse hyperbolic function* $y = \sinh^{-1} x$.

First we write x as a function of y, i.e.

$$y = \sinh^{-1} x \quad \Rightarrow \quad x = \sinh y.$$

Now, since $\cosh y = \frac{1}{2}(e^y + e^{-y})$ and $\sinh y = \frac{1}{2}(e^y - e^{-y})$,

$$\begin{aligned} e^y &= \cosh y + \sinh y \\ &= \sqrt{1 + \sinh^2 y} + \sinh y \\ e^y &= \sqrt{1 + x^2} + x, \end{aligned}$$

and hence

$$y = \ln(\sqrt{1 + x^2} + x). \text{ ◄}$$

In a similar fashion it can be shown that

$$\cosh^{-1} x = \ln(\sqrt{x^2 - 1} + x).$$

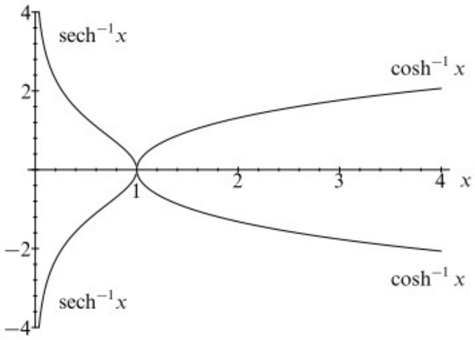

Figure 3.14 Graphs of $\cosh^{-1}x$ and $\mathrm{sech}^{-1}x$.

►*Find a closed-form expression for the inverse hyperbolic function* $y = \tanh^{-1}x$.

First we write x as a function of y, i.e.

$$y = \tanh^{-1}x \qquad \Rightarrow \qquad x = \tanh y.$$

Now, using the definition of $\tanh y$ and rearranging, we find

$$x = \frac{e^y - e^{-y}}{e^y + e^{-y}} \qquad \Rightarrow \qquad (x+1)e^{-y} = (1-x)e^y.$$

Thus, it follows that

$$\begin{aligned} e^{2y} = \frac{1+x}{1-x} \quad \Rightarrow \quad e^y &= \sqrt{\frac{1+x}{1-x}}, \\ y &= \ln\sqrt{\frac{1+x}{1-x}}, \\ \tanh^{-1}x &= \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right). \end{aligned}$$ ◄

Graphs of the inverse hyperbolic functions are given in figures 3.14–3.16.

3.7.6 Calculus of hyperbolic functions

Just as the identities of hyperbolic functions closely follow those of their trigonometric counterparts, so their calculus is similar. The derivatives of the two basic

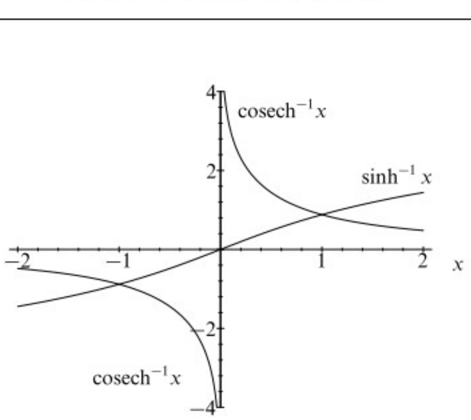

Figure 3.15 Graphs of $\sinh^{-1} x$ and $\operatorname{cosech}^{-1} x$.

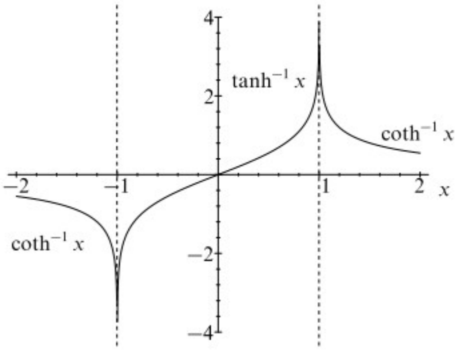

Figure 3.16 Graphs of $\tanh^{-1} x$ and $\coth^{-1} x$.

hyperbolic functions are given by

$$\frac{d}{dx}(\cosh x) = \sinh x, \tag{3.52}$$

$$\frac{d}{dx}(\sinh x) = \cosh x. \tag{3.53}$$

They may be deduced by considering the definitions (3.38), (3.39) as follows.

►*Verify the relation* $(d/dx)\cosh x = \sinh x$.

Using the definition of $\cosh x$,

$$\cosh x = \tfrac{1}{2}(e^x + e^{-x}),$$

and differentiating directly, we find

$$\begin{aligned}\frac{d}{dx}(\cosh x) &= \tfrac{1}{2}(e^x - e^{-x})\\ &= \sinh x. \blacktriangleleft\end{aligned}$$

Clearly the integrals of the fundamental hyperbolic functions are also defined by these relations. The derivatives of the remaining hyperbolic functions can be derived by product differentiation and are presented below only for completeness.

$$\frac{d}{dx}(\tanh x) = \text{sech}^2 x, \tag{3.54}$$

$$\frac{d}{dx}(\text{sech } x) = -\text{sech } x \tanh x, \tag{3.55}$$

$$\frac{d}{dx}(\text{cosech } x) = -\text{cosech } x \coth x, \tag{3.56}$$

$$\frac{d}{dx}(\coth x) = -\text{cosech}^2 x. \tag{3.57}$$

The inverse hyperbolic functions also have derivatives, which are given by the following:

$$\frac{d}{dx}\left(\cosh^{-1}\frac{x}{a}\right) = \frac{1}{\sqrt{x^2 - a^2}}, \tag{3.58}$$

$$\frac{d}{dx}\left(\sinh^{-1}\frac{x}{a}\right) = \frac{1}{\sqrt{x^2 + a^2}}, \tag{3.59}$$

$$\frac{d}{dx}\left(\tanh^{-1}\frac{x}{a}\right) = \frac{a}{a^2 - x^2}, \quad \text{for } x^2 < a^2, \tag{3.60}$$

$$\frac{d}{dx}\left(\coth^{-1}\frac{x}{a}\right) = \frac{-a}{x^2 - a^2}, \quad \text{for } x^2 > a^2. \tag{3.61}$$

These may be derived from the logarithmic form of the inverse (see subsection 3.7.5).

►*Evaluate* $(d/dx)\sinh^{-1}x$ *using the logarithmic form of the inverse.*

From the results of section 3.7.5,

$$\begin{aligned}\frac{d}{dx}\left(\sinh^{-1}x\right) &= \frac{d}{dx}\left[\ln\left(x+\sqrt{x^2+1}\right)\right]\\ &= \frac{1}{x+\sqrt{x^2+1}}\left(1+\frac{x}{\sqrt{x^2+1}}\right)\\ &= \frac{1}{x+\sqrt{x^2+1}}\left(\frac{\sqrt{x^2+1}+x}{\sqrt{x^2+1}}\right)\\ &= \frac{1}{\sqrt{x^2+1}}.\end{aligned}$$ ◄

3.8 Exercises

3.1 Two complex numbers z and w are given by $z = 3+4i$ and $w = 2-i$. On an Argand diagram, plot

(a) $z+w$, (b) $w-z$, (c) wz, (d) z/w,
(e) z^*w+w^*z, (f) w^2, (g) $\ln z$, (h) $(1+z+w)^{1/2}$.

3.2 By considering the real and imaginary parts of the product $e^{i\theta}e^{i\phi}$ prove the standard formulae for $\cos(\theta+\phi)$ and $\sin(\theta+\phi)$.

3.3 By writing $\pi/12 = (\pi/3)-(\pi/4)$ and considering $e^{i\pi/12}$, evaluate $\cot(\pi/12)$.

3.4 Find the locus in the complex z-plane of points that satisfy the following equations.

(a) $z-c=\rho\left(\dfrac{1+it}{1-it}\right)$, where c is complex, ρ is real and t is a real parameter that varies in the range $-\infty<t<\infty$.
(b) $z=a+bt+ct^2$, in which t is a real parameter and a, b, and c are complex numbers with b/c real.

3.5 Evaluate

(a) $\mathrm{Re}(\exp 2iz)$, (b) $\mathrm{Im}(\cosh^2 z)$, (c) $(-1+\sqrt{3}i)^{1/2}$,
(d) $|\exp(i^{1/2})|$, (e) $\exp(i^3)$, (f) $\mathrm{Im}(2^{i+3})$, (g) i^i, (h) $\ln[(\sqrt{3}+i)^3]$.

3.6 Find the equations in terms of x and y of the sets of points in the Argand diagram that satisfy the following:

(a) $\mathrm{Re}\,z^2=\mathrm{Im}\,z^2$;
(b) $(\mathrm{Im}\,z^2)/z^2=-i$;
(c) $\arg[z/(z-1)]=\pi/2$.

3.7 Show that the locus of all points $z=x+iy$ in the complex plane that satisfy

$$|z-ia|=\lambda|z+ia|,\quad \lambda>0,$$

is a circle of radius $|2\lambda a/(1-\lambda^2)|$ centred on the point $z=ia[(1+\lambda^2)/(1-\lambda^2)]$. Sketch the circles for a few typical values of λ, including $\lambda<1$, $\lambda>1$ and $\lambda=1$.

3.8 The two sets of points $z=a$, $z=b$, $z=c$, and $z=A$, $z=B$, $z=C$ are the corners of two similar triangles in the Argand diagram. Express in terms of $a,b,\ldots,C$

11

Line, surface and volume integrals

In the previous chapter we encountered continuously varying scalar and vector fields and discussed the action of various differential operators on them. In addition to these differential operations, the need often arises to consider the integration of field quantities along lines, over surfaces and throughout volumes. In general the integrand may be scalar or vector in nature, but the evaluation of such integrals involves their reduction to one or more scalar integrals, which are then evaluated. In the case of surface and volume integrals this requires the evaluation of double and triple integrals (see chapter 6).

11.1 Line integrals

In this section we discuss *line* or *path integrals*, in which some quantity related to the field is integrated between two given points in space, A and B, along a prescribed curve C that joins them. In general, we may encounter line integrals of the forms

$$\int_C \phi \, d\mathbf{r}, \qquad \int_C \mathbf{a} \cdot d\mathbf{r}, \qquad \int_C \mathbf{a} \times d\mathbf{r}, \tag{11.1}$$

where ϕ is a scalar field and $\mathbf{a}$ is a vector field. The three integrals themselves are respectively vector, scalar and vector in nature. As we will see below, in physical applications line integrals of the second type are by far the most common.

The formal definition of a line integral closely follows that of ordinary integrals and can be considered as the limit of a sum. We may divide the path C joining the points A and B into N small line elements $\Delta\mathbf{r}_p$, $p = 1, \ldots, N$. If (x_p, y_p, z_p) is any point on the line element $\Delta\mathbf{r}_p$ then the second type of line integral in (11.1), for example, is defined as

$$\int_C \mathbf{a} \cdot d\mathbf{r} = \lim_{N\to\infty} \sum_{p=1}^{N} \mathbf{a}(x_p, y_p, z_p) \cdot \Delta\mathbf{r}_p,$$

where it is assumed that all $|\Delta\mathbf{r}_p| \to 0$ as $N \to \infty$.

Each of the line integrals in (11.1) is evaluated over some curve C that may be either open (A and B being distinct points) or closed (the curve C forms a loop, so that A and B are coincident). In the case where C is closed, the line integral is written $\oint_C$ to indicate this. The curve may be given either parametrically by $\mathbf{r}(u) = x(u)\mathbf{i} + y(u)\mathbf{j} + z(u)\mathbf{k}$ or by means of simultaneous equations relating x, y, z for the given path (in Cartesian coordinates). A full discussion of the different representations of space curves was given in section 10.3.

In general, the value of the line integral depends not only on the end-points A and B but also on the path C joining them. For a closed curve we must also specify the direction around the loop in which the integral is taken. It is usually taken to be such that a person walking around the loop C in this direction always has the region R on his/her left; this is equivalent to traversing C in the anticlockwise direction (as viewed from above).

11.1.1 Evaluating line integrals

The method of evaluating a line integral is to reduce it to a set of scalar integrals. It is usual to work in Cartesian coordinates, in which case $d\mathbf{r} = dx\,\mathbf{i} + dy\,\mathbf{j} + dz\,\mathbf{k}$. The first type of line integral in (11.1) then becomes simply

$$\int_C \phi\, d\mathbf{r} = \mathbf{i}\int_C \phi(x,y,z)\,dx + \mathbf{j}\int_C \phi(x,y,z)\,dy + \mathbf{k}\int_C \phi(x,y,z)\,dz.$$

The three integrals on the RHS are ordinary scalar integrals that can be evaluated in the usual way once the path of integration C has been specified. Note that in the above we have used relations of the form

$$\int \phi\,\mathbf{i}\,dx = \mathbf{i}\int \phi\,dx,$$

which is allowable since the Cartesian unit vectors are of constant magnitude and direction and hence may be taken out of the integral. If we had been using a different coordinate system, such as spherical polars, then, as we saw in the previous chapter, the unit basis vectors would not be constant. In that case the basis vectors could not be factorised out of the integral.

The second and third line integrals in (11.1) can also be reduced to a set of scalar integrals by writing the vector field $\mathbf{a}$ in terms of its Cartesian components as $\mathbf{a} = a_x\mathbf{i} + a_y\mathbf{j} + a_z\mathbf{k}$, where a_x, a_y, a_z are each (in general) functions of x, y, z. The second line integral in (11.1), for example, can then be written as

$$\begin{aligned}\int_C \mathbf{a}\cdot d\mathbf{r} &= \int_C (a_x\mathbf{i} + a_y\mathbf{j} + a_z\mathbf{k})\cdot(dx\,\mathbf{i} + dy\,\mathbf{j} + dz\,\mathbf{k})\\ &= \int_C (a_x\,dx + a_y\,dy + a_z\,dz)\\ &= \int_C a_x\,dx + \int_C a_y\,dy + \int_C a_z\,dz. \qquad (11.2)\end{aligned}$$

A similar procedure may be followed for the third type of line integral in (11.1), which involves a cross product.

Line integrals have properties that are analogous to those of ordinary integrals. In particular, the following are useful properties (which we illustrate using the second form of line integral in (11.1) but which are valid for all three types).

(i) Reversing the path of integration changes the sign of the integral. If the path C along which the line integrals are evaluated has A and B as its end-points then

$$\int_A^B \mathbf{a}\cdot d\mathbf{r} = -\int_B^A \mathbf{a}\cdot d\mathbf{r}.$$

This implies that if the path C is a loop then integrating around the loop in the opposite direction changes the sign of the integral.

(ii) If the path of integration is subdivided into smaller segments then the sum of the separate line integrals along each segment is equal to the line integral along the whole path. So, if P is any point on the path of integration that lies between the path's end-points A and B then

$$\int_A^B \mathbf{a}\cdot d\mathbf{r} = \int_A^P \mathbf{a}\cdot d\mathbf{r} + \int_P^B \mathbf{a}\cdot d\mathbf{r}.$$

►*Evaluate the line integral* $I = \int_C \mathbf{a}\cdot d\mathbf{r}$, *where* $\mathbf{a} = (x+y)\mathbf{i} + (y-x)\mathbf{j}$, *along each of the paths in the xy-plane shown in figure 11.1, namely*

(i) *the parabola* $y^2 = x$ *from* $(1,1)$ *to* $(4,2)$,
(ii) *the curve* $x = 2u^2 + u + 1$, $y = 1 + u^2$ *from* $(1,1)$ *to* $(4,2)$,
(iii) *the line* $y = 1$ *from* $(1,1)$ *to* $(4,1)$, *followed by the line* $x = 4$ *from* $(4,1)$ *to* $(4,2)$.

Since each of the paths lies entirely in the xy-plane, we have $d\mathbf{r} = dx\,\mathbf{i} + dy\,\mathbf{j}$. We can therefore write the line integral as

$$I = \int_C \mathbf{a}\cdot d\mathbf{r} = \int_C [(x+y)\,dx + (y-x)\,dy]. \tag{11.3}$$

We must now evaluate this line integral along each of the prescribed paths.

Case (i). Along the parabola $y^2 = x$ we have $2y\,dy = dx$. Substituting for x in (11.3) and using just the limits on y, we obtain

$$I = \int_{(1,1)}^{(4,2)} [(x+y)\,dx + (y-x)\,dy] = \int_1^2 [(y^2+y)2y + (y-y^2)]\,dy = 11\tfrac{1}{3}.$$

Note that we could just as easily have substituted for y and obtained an integral in x, which would have given the same result.

Case (ii). The second path is given in terms of a parameter u. We could eliminate u between the two equations to obtain a relationship between x and y directly and proceed as above, but it is usually quicker to write the line integral in terms of the parameter u. Along the curve $x = 2u^2 + u + 1$, $y = 1 + u^2$ we have $dx = (4u+1)\,du$ and $dy = 2u\,du$.

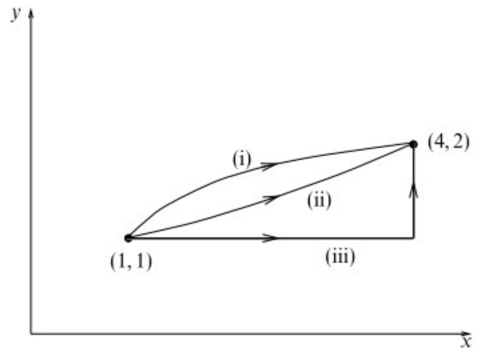

Figure 11.1 Different possible paths between the points (1, 1) and (4, 2).

Substituting for x and y in (11.3) and writing the correct limits on u, we obtain

$$\begin{aligned} I &= \int_{(1,1)}^{(4,2)} [(x+y)\,dx + (y-x)\,dy] \\ &= \int_0^1 [(3u^2+u+2)(4u+1) - (u^2+u)2u]\,du = 10\tfrac{2}{3}. \end{aligned}$$

Case (iii). For the third path the line integral must be evaluated along the two line segments separately and the results added together. First, along the line $y = 1$ we have $dy = 0$. Substituting this into (11.3) and using just the limits on x for this segment, we obtain

$$\int_{(1,1)}^{(4,1)} [(x+y)\,dx + (y-x)\,dy] = \int_1^4 (x+1)\,dx = 10\tfrac{1}{2}.$$

Next, along the line $x = 4$ we have $dx = 0$. Substituting this into (11.3) and using just the limits on y for this segment, we obtain

$$\int_{(4,1)}^{(4,2)} [(x+y)\,dx + (y-x)\,dy] = \int_1^2 (y-4)\,dy = -2\tfrac{1}{2}.$$

The value of the line integral along the whole path is just the sum of the values of the line integrals along each segment, and is given by $I = 10\frac{1}{2} - 2\frac{1}{2} = 8$. ◄

When calculating a line integral along some curve C, which is given in terms of x, y and z, we are sometimes faced with the problem that the curve C is such that x, y and z are not single-valued functions of one another over the entire length of the curve. This is a particular problem for closed loops in the xy-plane (and also for some open curves). In such cases the path may be subdivided into shorter line segments along which one coordinate is a single-valued function of the other two. The sum of the line integrals along these segments is then equal to the line integral along the entire curve C. A better solution, however, is to represent the curve in a parametric form $\mathbf{r}(u)$ that is valid for its entire length.

►*Evaluate the line integral $I = \oint_C x\,dy$, where C is the circle in the xy-plane defined by $x^2 + y^2 = a^2$, $z = 0$.*

Adopting the usual convention mentioned above, the circle C is to be traversed in the anticlockwise direction. Taking the circle as a whole means x is not a single-valued function of y. We must therefore divide the path into two parts with $x = +\sqrt{a^2 - y^2}$ for the semicircle lying to the right of $x = 0$, and $x = -\sqrt{a^2 - y^2}$ for the semicircle lying to the left of $x = 0$. The required line integral is then the sum of the integrals along the two semicircles. Substituting for x, it is given by

$$\begin{aligned} I = \oint_C x\,dy &= \int_{-a}^{a} \sqrt{a^2 - y^2}\,dy + \int_{a}^{-a} \left(-\sqrt{a^2 - y^2}\right)\,dy \\ &= 4\int_0^a \sqrt{a^2 - y^2}\,dy = \pi a^2. \end{aligned}$$

Alternatively, we can represent the entire circle parametrically, in terms of the azimuthal angle ϕ, so that $x = a\cos\phi$ and $y = a\sin\phi$ with ϕ running from 0 to 2π. The integral can therefore be evaluated over the whole circle at once. Noting that $dy = a\cos\phi\,d\phi$, we can rewrite the line integral completely in terms of the parameter ϕ and obtain

$$I = \oint_C x\,dy = a^2 \int_0^{2\pi} \cos^2\phi\,d\phi = \pi a^2. \blacktriangleleft$$

11.1.2 Physical examples of line integrals

There are many physical examples of line integrals, but perhaps the most common is the expression for the total work done by a force $\mathbf{F}$ when it moves its point of application from a point A to a point B along a given curve C. We allow the magnitude and direction of $\mathbf{F}$ to vary along the curve. Let the force act at a point $\mathbf{r}$ and consider a small displacement $d\mathbf{r}$ along the curve; then the small amount of work done is $dW = \mathbf{F}\cdot d\mathbf{r}$, as discussed in subsection 7.6.1 (note that dW can be either positive or negative). Therefore, the total work done in traversing the path C is

$$W_C = \int_C \mathbf{F}\cdot d\mathbf{r}.$$

Naturally, other physical quantities can be expressed in such a way. For example, the electrostatic potential energy gained by moving a charge q along a path C in an electric field $\mathbf{E}$ is $-q\int_C \mathbf{E}\cdot d\mathbf{r}$. We may also note that Ampère's law concerning the magnetic field $\mathbf{B}$ associated with a current-carrying wire can be written as

$$\oint_C \mathbf{B}\cdot d\mathbf{r} = \mu_0 I,$$

where I is the current enclosed by a closed path C traversed in a right-handed sense with respect to the current direction.

Magnetostatics also provides a physical example of the third type of line integral in (11.1). If a loop of wire C carrying a current I is placed in a magnetic field $\mathbf{B}$ then the force $d\mathbf{F}$ on a small length $d\mathbf{r}$ of the wire is given by $d\mathbf{F} = I\, d\mathbf{r} \times \mathbf{B}$, and so the total (vector) force on the loop is

$$\mathbf{F} = I \oint_C d\mathbf{r} \times \mathbf{B}.$$

11.1.3 Line integrals with respect to a scalar

In addition to those listed in (11.1), we can form other types of line integral, which depend on a particular curve C but for which we integrate with respect to a scalar du, rather than the vector differential $d\mathbf{r}$. This distinction is somewhat arbitrary, however, since we can always rewrite line integrals containing the vector differential $d\mathbf{r}$ as a line integral with respect to some scalar parameter. If the path C along which the integral is taken is described parametrically by $\mathbf{r}(u)$ then

$$d\mathbf{r} = \frac{d\mathbf{r}}{du}\, du,$$

and the second type of line integral in (11.1), for example, can be written as

$$\int_C \mathbf{a} \cdot d\mathbf{r} = \int_C \mathbf{a} \cdot \frac{d\mathbf{r}}{du}\, du.$$

A similar procedure can be followed for the other types of line integral in (11.1).

Commonly occurring special cases of line integrals with respect to a scalar are

$$\int_C \phi\, ds, \qquad \int_C \mathbf{a}\, ds,$$

where s is the arc length along the curve C. We can always represent C parametrically by $\mathbf{r}(u)$, and from section 10.3 we have

$$ds = \sqrt{\frac{d\mathbf{r}}{du} \cdot \frac{d\mathbf{r}}{du}}\, du.$$

The line integrals can therefore be expressed entirely in terms of the parameter u and thence evaluated.

►*Evaluate the line integral $I = \int_C (x-y)^2\, ds$, where C is the semicircle of radius a running from $A = (a, 0)$ to $B = (-a, 0)$ and for which $y \geq 0$.*

The semicircular path from A to B can be described in terms of the azimuthal angle ϕ (measured from the x-axis) by

$$\mathbf{r}(\phi) = a\cos\phi\, \mathbf{i} + a\sin\phi\, \mathbf{j},$$

where ϕ runs from 0 to π. Therefore the element of arc length is given, from section 10.3, by

$$ds = \sqrt{\frac{d\mathbf{r}}{d\phi} \cdot \frac{d\mathbf{r}}{d\phi}}\, d\phi = a(\cos^2\phi + \sin^2\phi)\, d\phi = a\, d\phi.$$

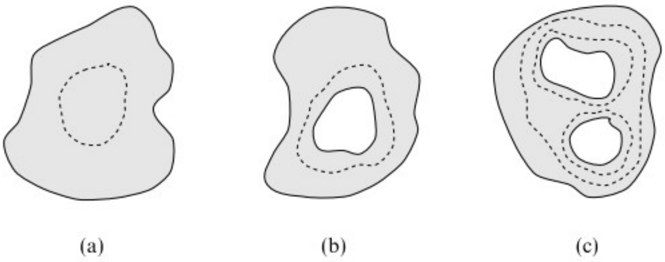

Figure 11.2 (a) A simply connected region; (b) a doubly connected region; (c) a triply connected region.

Since $(x-y)^2 = a^2(1-\sin 2\phi)$, the line integral becomes

$$I = \int_C (x-y)^2\,ds = \int_0^{\pi} a^3(1-\sin 2\phi)\,d\phi = \pi a^3. \blacktriangleleft$$

As discussed in the previous chapter, the expression (10.58) for the square of the element of arc length in three-dimensional orthogonal curvilinear coordinates u_1, u_2, u_3 is

$$(ds)^2 = h_1^2\,(du_1)^2 + h_2^2\,(du_2)^2 + h_3^2\,(du_3)^2,$$

where h_1, h_2, h_3 are the scale factors of the coordinate system. If a curve C in three dimensions is given parametrically by the equations $u_i = u_i(\lambda)$ for $i = 1, 2, 3$ then the element of arc length along the curve is

$$ds = \sqrt{h_1^2\left(\frac{du_1}{d\lambda}\right)^2 + h_2^2\left(\frac{du_2}{d\lambda}\right)^2 + h_3^2\left(\frac{du_3}{d\lambda}\right)^2}\,d\lambda.$$

11.2 Connectivity of regions

In physical systems it is usual to define a scalar or vector field in some region R. In the next and some later sections we will need the concept of the *connectivity* of such a region in both two and three dimensions.

We begin by discussing planar regions. A plane region R is said to be *simply connected* if every simple closed curve within R can be continuously shrunk to a point without leaving the region (see figure 11.2(a)). If, however, the region R contains a hole then there exist simple closed curves that cannot be shrunk to a point without leaving R (see figure 11.2(b)). Such a region is said to be doubly connected, since its boundary has two distinct parts. Similarly, a region with $n-1$ holes is said to be *n-fold connected*, or *multiply connected* (the region in figure 11.2(c) is triply connected).

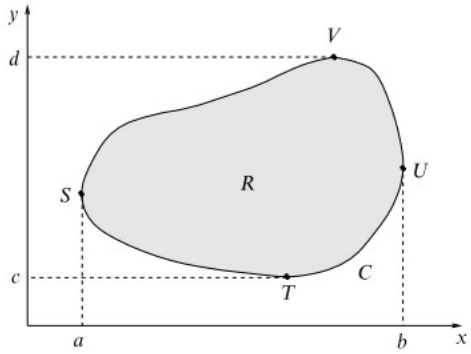

Figure 11.3 A simply connected region R bounded by the curve C.

These ideas can be extended to regions that are not planar, such as general three-dimensional surfaces and volumes. The same criteria concerning the shrinking of closed curves to a point also apply when deciding the connectivity of such regions. In these cases, however, the curves must lie in the surface or volume in question. For example, the interior of a torus is not simply connected, since there exist closed curves in the interior that cannot be shrunk to a point without leaving the torus. The region between two concentric spheres of different radii is simply connected.

11.3 Green's theorem in a plane

In subsection 11.1.1 we considered (amongst other things) the evaluation of line integrals for which the path C is closed and lies entirely in the xy-plane. Since the path is closed it will enclose a region R of the plane. We now discuss how to express the line integral around the loop as a double integral over the enclosed region R.

Suppose the functions $P(x,y)$, $Q(x,y)$ and their partial derivatives are single-valued, finite and continuous inside and on the boundary C of some simply connected region R in the xy-plane. *Green's theorem in a plane* (sometimes called the divergence theorem in two dimensions) then states

$$\oint_C (P\,dx + Q\,dy) = \iint_R \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dx\,dy, \tag{11.4}$$

and so relates the line integral around C to a double integral over the enclosed region R. This theorem may be proved straightforwardly in the following way. Consider the simply connected region R in figure 11.3, and let $y = y_1(x)$ and

$y = y_2(x)$ be the equations of the curves STU and SVU respectively. We then write

$$\begin{aligned}\iint_R \frac{\partial P}{\partial y}\,dx\,dy &= \int_a^b dx \int_{y_1(x)}^{y_2(x)} dy\,\frac{\partial P}{\partial y} = \int_a^b dx\Big[P(x,y)\Big]_{y=y_1(x)}^{y=y_2(x)} \\ &= \int_a^b \Big[P(x,y_2(x)) - P(x,y_1(x))\Big]\,dx \\ &= -\int_a^b P(x,y_1(x))\,dx - \int_b^a P(x,y_2(x))\,dx = -\oint_C P\,dx.\end{aligned}$$

If we now let $x = x_1(y)$ and $x = x_2(y)$ be the equations of the curves TSV and TUV respectively, we can similarly show that

$$\begin{aligned}\iint_R \frac{\partial Q}{\partial x}\,dx\,dy &= \int_c^d dy \int_{x_1(y)}^{x_2(y)} dx\,\frac{\partial Q}{\partial x} = \int_c^d dy\Big[Q(x,y)\Big]_{x=x_1(y)}^{x=x_2(y)} \\ &= \int_c^d \Big[Q(x_2(y),y) - Q(x_1(y),y)\Big]\,dy \\ &= \int_d^c Q(x_1,y)\,dy + \int_c^d Q(x_2,y)\,dy = \oint_C Q\,dy.\end{aligned}$$

Subtracting these two results gives Green's theorem in a plane.

►*Show that the area of a region R enclosed by a simple closed curve C is given by* $A = \frac{1}{2}\oint_C(x\,dy - y\,dx) = \oint_C x\,dy = -\oint_C y\,dx$. *Hence calculate the area of the ellipse* $x = a\cos\phi$, $y = b\sin\phi$.

In Green's theorem (11.4) put $P = -y$ and $Q = x$; then

$$\oint_C (x\,dy - y\,dx) = \iint_R (1+1)\,dx\,dy = 2\iint_R dx\,dy = 2A.$$

Therefore the area of the region is $A = \frac{1}{2}\oint_C(x\,dy - y\,dx)$. Alternatively, we could put $P = 0$ and $Q = x$ and obtain $A = \oint_C x\,dy$, or put $P = -y$ and $Q = 0$, which gives $A = -\oint_C y\,dx$.

The area of the ellipse $x = a\cos\phi$, $y = b\sin\phi$ is given by

$$\begin{aligned}A = \frac{1}{2}\oint_C (x\,dy - y\,dx) &= \frac{1}{2}\int_0^{2\pi} ab(\cos^2\phi + \sin^2\phi)\,d\phi \\ &= \frac{ab}{2}\int_0^{2\pi} d\phi = \pi ab. \blacktriangleleft\end{aligned}$$

It may further be shown that Green's theorem in a plane is also valid for multiply connected regions. In this case, the line integral must be taken over all the distinct boundaries of the region. Furthermore, each boundary must be traversed in the positive direction, so that a person travelling along it in this direction always has the region R on their left. In order to apply Green's theorem

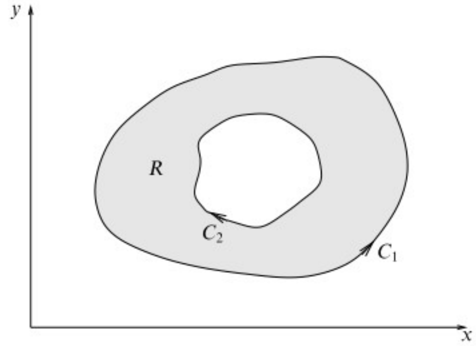

Figure 11.4 A doubly connected region R bounded by the curves C_1 and C_2.

to the region R shown in figure 11.4, the line integrals must be taken over both boundaries, C_1 and C_2, in the directions indicated, and the results added together.

We may also use Green's theorem in a plane to investigate the path independence (or not) of line integrals when the paths lie in the xy-plane. Let us consider the line integral

$$I = \int_A^B (P\,dx + Q\,dy).$$

For the line integral from A to B to be independent of the path taken, it must have the same value along any two arbitrary paths C_1 and C_2 joining the points. Moreover, if we consider as the path the closed loop C formed by $C_1 - C_2$ then the line integral around this loop must be zero. From Green's theorem in a plane, (11.4), we see that a *sufficient* condition for $I = 0$ is that

$$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}, \tag{11.5}$$

throughout some simply connected region R containing the loop, where we assume that these partial derivatives are continuous in R.

It may be shown that (11.5) is also a *necessary* condition for $I = 0$ and is equivalent to requiring $P\,dx + Q\,dy$ to be an exact differential of some function $\phi(x, y)$ such that $P\,dx + Q\,dy = d\phi$. It follows that $\int_A^B (P\,dx + Q\,dy) = \phi(B) - \phi(A)$ and that $\oint_C (P\,dx + Q\,dy)$ around any closed loop C in the region R is identically zero. These results are special cases of the general results for paths in three dimensions, which are discussed in the next section.

►*Evaluate the line integral*

$$I = \oint_C \left[(e^x y + \cos x \sin y)\, dx + (e^x + \sin x \cos y)\, dy\right],$$

around the ellipse $x^2/a^2 + y^2/b^2 = 1$.

Clearly, it is not straightforward to calculate this line integral directly. However, if we let

$$P = e^x y + \cos x \sin y \qquad \text{and} \qquad Q = e^x + \sin x \cos y,$$

then $\partial P/\partial y = e^x + \cos x \cos y = \partial Q/\partial x$, and so $P\,dx + Q\,dy$ is an exact differential (it is actually the differential of the function $f(x, y) = e^x y + \sin x \sin y$). From the above discussion, we can conclude immediately that $I = 0$. ◄

11.4 Conservative fields and potentials

So far we have made the point that, in general, the value of a line integral between two points A and B depends on the path C taken from A to B. In the previous section, however, we saw that, for paths in the xy-plane, line integrals whose integrands have certain properties are independent of the path taken. We now extend that discussion to the full three-dimensional case.

For line integrals of the form $\int_C \mathbf{a} \cdot d\mathbf{r}$, there exists a class of vector fields for which the line integral between two points is *independent* of the path taken. Such vector fields are called *conservative*. A vector field $\mathbf{a}$ that has continuous partial derivatives in a simply connected region R is conservative if, and only if, any of the following is true.

- (i) The integral $\int_A^B \mathbf{a} \cdot d\mathbf{r}$, where A and B lie in the region R, is independent of the path from A to B. Hence the integral $\oint_C \mathbf{a} \cdot d\mathbf{r}$ around any closed loop in R is zero.
- (ii) There exists a single-valued function ϕ of position such that $\mathbf{a} = \nabla\phi$.
- (iii) $\nabla \times \mathbf{a} = \mathbf{0}$.
- (iv) $\mathbf{a} \cdot d\mathbf{r}$ is an exact differential.

The validity or otherwise of any of these statements implies the same for the other three, as we will now show.

First, let us assume that (i) above is true. If the line integral from A to B is independent of the path taken between the points then its value must be a function only of the positions of A and B. We may therefore write

$$\int_A^B \mathbf{a} \cdot d\mathbf{r} = \phi(B) - \phi(A), \tag{11.6}$$

which defines a single-valued scalar function of position ϕ. If the points A and B are separated by an infinitesimal displacement $d\mathbf{r}$ then (11.6) becomes

$$\mathbf{a} \cdot d\mathbf{r} = d\phi,$$

which shows that we require $\mathbf{a}\cdot d\mathbf{r}$ to be an exact differential: condition (iv). From (10.27) we can write $d\phi = \nabla\phi \cdot d\mathbf{r}$, and so we have

$$(\mathbf{a} - \nabla\phi) \cdot d\mathbf{r} = 0.$$

Since $d\mathbf{r}$ is arbitrary, we find that $\mathbf{a} = \nabla\phi$; this immediately implies $\nabla \times \mathbf{a} = \mathbf{0}$, condition (iii) (see (10.37)).

Alternatively, if we suppose that there exists a single-valued function of position ϕ such that $\mathbf{a} = \nabla\phi$ then $\nabla \times \mathbf{a} = \mathbf{0}$ follows as before. The line integral around a closed loop then becomes

$$\oint_C \mathbf{a} \cdot d\mathbf{r} = \oint_C \nabla\phi \cdot d\mathbf{r} = \oint d\phi.$$

Since we defined ϕ to be single-valued, this integral is zero as required.

Now suppose $\nabla \times \mathbf{a} = \mathbf{0}$. From Stoke's theorem, which is discussed in section 11.9, we immediately obtain $\oint_C \mathbf{a}\cdot d\mathbf{r} = 0$; then $\mathbf{a} = \nabla\phi$ and $\mathbf{a}\cdot d\mathbf{r} = d\phi$ follow as above.

Finally, let us suppose $\mathbf{a} \cdot d\mathbf{r} = d\phi$. Then immediately we have $\mathbf{a} = \nabla\phi$, and the other results follow as above.

►*Evaluate the line integral* $I = \int_A^B \mathbf{a} \cdot d\mathbf{r}$, *where* $\mathbf{a} = (xy^2 + z)\mathbf{i} + (x^2y + 2)\mathbf{j} + x\mathbf{k}$, *A is the point* (c, c, h) *and B is the point* $(2c, c/2, h)$, *along the different paths*

(i) *C_1, given by* $x = cu$, $y = c/u$, $z = h$,
(ii) *C_2, given by* $2y = 3c - x$, $z = h$.

Show that the vector field $\mathbf{a}$ *is in fact conservative, and find* ϕ *such that* $\mathbf{a} = \nabla\phi$.

Expanding out the integrand, we have

$$I = \int_{(c,c,h)}^{(2c,c/2,h)} \left[(xy^2 + z)\,dx + (x^2y + 2)\,dy + x\,dz\right], \tag{11.7}$$

which we must evaluate along each of the paths C_1 and C_2.

(i) Along C_1 we have $dx = c\,du$, $dy = -(c/u^2)\,du$, $dz = 0$, and on substituting in (11.7) and finding the limits on u, we obtain

$$I = \int_1^2 c\left(h - \frac{2}{u^2}\right)\,du = c(h-1).$$

(ii) Along C_2 we have $2\,dy = -dx$, $dz = 0$ and, on substituting in (11.7) and using the limits on x, we obtain

$$I = \int_c^{2c} \left(\tfrac{1}{2}x^3 - \tfrac{9}{4}cx^2 + \tfrac{9}{4}c^2x + h - 1\right)\,dx = c(h-1).$$

Hence the line integral has the same value along paths C_1 and C_2. Taking the curl of $\mathbf{a}$, we have

$$\nabla \times \mathbf{a} = (0-0)\mathbf{i} + (1-1)\mathbf{j} + (2xy - 2xy)\mathbf{k} = \mathbf{0},$$

so $\mathbf{a}$ is a conservative vector field, and the line integral between two points must be independent of the path taken. Since **a** is conservative, we can write $\mathbf{a} = \nabla\phi$. Therefore, ϕ must satisfy

$$\frac{\partial \phi}{\partial x} = xy^2 + z,$$

which implies that $\phi = \frac{1}{2}x^2y^2 + zx + f(y, z)$ for some function f. Secondly, we require

$$\frac{\partial \phi}{\partial y} = x^2y + \frac{\partial f}{\partial y} = x^2y + 2,$$

which implies $f = 2y + g(z)$. Finally, since

$$\frac{\partial \phi}{\partial z} = x + \frac{\partial g}{\partial z} = x,$$

we have g = constant = k. It can be seen that we have explicitly constructed the function $\phi = \frac{1}{2}x^2y^2 + zx + 2y + k$. ◄

The quantity ϕ that figures so prominently in this section is called the *scalar potential function* of the conservative vector field **a** (which satisfies $\nabla \times \mathbf{a} = \mathbf{0}$), and is unique up to an arbitrary additive constant. Scalar potentials that are multivalued functions of position (but in simple ways) are also of value in describing some physical situations, the most obvious example being the scalar magnetic potential associated with a current-carrying wire. When the integral of a field quantity around a closed loop is considered, provided the loop does not enclose a net current, the potential is single-valued and all the above results still hold. If the loop does enclose a net current, however, our analysis is no longer valid and extra care must be taken.

If, instead of being conservative, a vector field **b** satisfies $\nabla \cdot \mathbf{b} = 0$ (i.e. **b** is solenoidal) then it is both possible and useful, for example in the theory of electromagnetism, to define a *vector field* **a** such that $\mathbf{b} = \nabla \times \mathbf{a}$. It may be shown that such a vector field **a** always exists. Further, if **a** is one such vector field then $\mathbf{a}' = \mathbf{a} + \nabla\psi + \mathbf{c}$, where ψ is any scalar function and **c** is any constant vector, also satisfies the above relationship, i.e. $\mathbf{b} = \nabla \times \mathbf{a}'$. This was discussed more fully in subsection 10.8.2.

11.5 Surface integrals

As with line integrals, integrals over surfaces can involve vector and scalar fields and, equally, can result in either a vector or a scalar. The simplest case involves entirely scalars and is of the form

$$\int_S \phi \, dS. \tag{11.8}$$

As analogues of the line integrals listed in (11.1), we may also encounter surface integrals involving vectors, namely

$$\int_S \phi \, d\mathbf{S}, \qquad \int_S \mathbf{a} \cdot d\mathbf{S}, \qquad \int_S \mathbf{a} \times d\mathbf{S}. \tag{11.9}$$

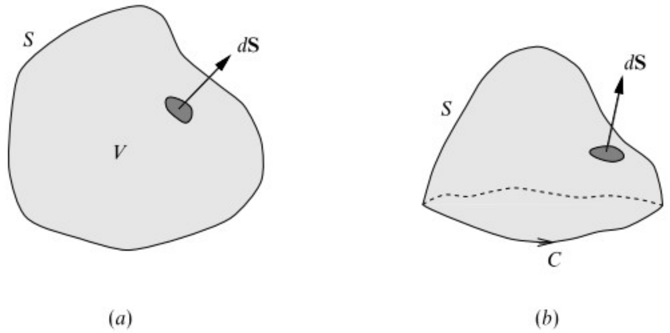

Figure 11.5 (a) A closed surface and (b) an open surface. In each case a normal to the surface is shown: $d\mathbf{S} = \hat{\mathbf{n}}\, dS$.

All the above integrals are taken over some surface S, which may be either open or closed, and are therefore, in general, double integrals. Following the notation for line integrals, for surface integrals over a closed surface $\int_S$ is replaced by $\oint_S$.

The vector differential $d\mathbf{S}$ in (11.9) represents a vector area element of the surface S. It may also be written $d\mathbf{S} = \hat{\mathbf{n}}\, dS$, where $\hat{\mathbf{n}}$ is a unit normal to the surface at the position of the element and dS is the scalar area of the element used in (11.8). The convention for the direction of the normal $\hat{\mathbf{n}}$ to a surface depends on whether the surface is open or closed. A closed surface, see figure 11.5(a), does not have to be simply connected (for example, the surface of a torus is not), but it does have to enclose a volume V, which may be of infinite extent. The direction of $\hat{\mathbf{n}}$ is taken to point outwards from the enclosed volume as shown. An open surface, see figure 11.5(b), spans some perimeter curve C. The direction of $\hat{\mathbf{n}}$ is then given by the right-hand sense with respect to the direction in which the perimeter is traversed, i.e. follows the right-hand screw rule discussed in subsection 7.6.2. An open surface does not have to be simply connected but for our purposes it must be two-sided (a Möbius strip is an example of a one-sided surface).

The formal definition of a surface integral is very similar to that of a line integral. We divide the surface S into N elements of area ΔS_p, $p = 1, 2, \ldots, N$, each with a unit normal $\hat{\mathbf{n}}_p$. If (x_p, y_p, z_p) is any point in ΔS_p then the second type of surface integral in (11.9), for example, is defined as

$$\int_S \mathbf{a} \cdot d\mathbf{S} = \lim_{N\to\infty} \sum_{p=1}^{N} \mathbf{a}(x_p, y_p, z_p) \cdot \hat{\mathbf{n}}_p \Delta S_p,$$

where it is required that all $\Delta S_p \to 0$ as $N \to \infty$.

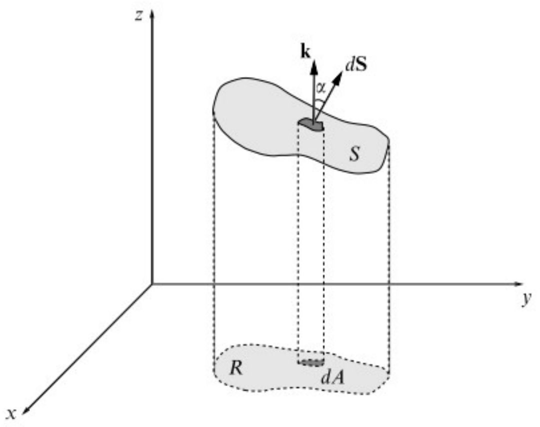

Figure 11.6 A surface S (or part thereof) projected onto a region R in the xy-plane; $d\mathbf{S}$ is a surface element.

11.5.1 Evaluating surface integrals

We now consider how to evaluate surface integrals over some general surface. This involves writing the scalar area element dS in terms of the coordinate differentials of our chosen coordinate system. In some particularly simple cases this is very straightforward. For example, if S is the surface of a sphere of radius a (or some part thereof) then using spherical polar coordinates θ, ϕ on the sphere we have $dS = a^2 \sin\theta \, d\theta \, d\phi$. For a general surface, however, it is not usually possible to represent the surface in a simple way in any particular coordinate system. In such cases, it is usual to work in Cartesian coordinates and consider the projections of the surface onto the coordinate planes.

Consider a surface (or part of a surface) S as in figure 11.6. The surface S is projected onto a region R of the xy-plane, so that an element of surface area dS projects onto the area element dA. From the figure, we see that $dA = |\cos\alpha| \, dS$, where α is the angle between the unit vector $\mathbf{k}$ in the z-direction and the unit normal $\hat{\mathbf{n}}$ to the surface at P. So, at any given point of S, we have simply

$$dS = \frac{dA}{|\cos\alpha|} = \frac{dA}{|\hat{\mathbf{n}} \cdot \mathbf{k}|}.$$

Now, if the surface S is given by the equation $f(x, y, z) = 0$ then, as shown in subsection 10.7.1, the unit normal at any point of the surface is given by $\hat{\mathbf{n}} = \nabla f / |\nabla f|$ evaluated at that point, cf. (10.32). The scalar element of surface area then becomes

$$dS = \frac{dA}{|\hat{\mathbf{n}} \cdot \mathbf{k}|} = \frac{|\nabla f| \, dA}{\nabla f \cdot \mathbf{k}} = \frac{|\nabla f| \, dA}{\partial f / \partial z}, \tag{11.10}$$

where $|\nabla f|$ and $\partial f/\partial z$ are evaluated on the surface S. We can therefore express any surface integral over S as a double integral over the region R in the xy-plane.

►*Evaluate the surface integral* $I = \int_S \mathbf{a}\cdot d\mathbf{S}$, *where* $\mathbf{a} = x\mathbf{i}$ *and* S *is the surface of the hemisphere* $x^2+y^2+z^2=a^2$ *with* $z\geq 0$.

The surface of the hemisphere is shown in figure 11.7. In this case dS may be easily expressed in spherical polar coordinates as $dS = a^2\sin\theta\, d\theta\, d\phi$, and the unit normal to the surface at any point is simply $\hat{\mathbf{r}}$. On the surface of the hemisphere we have $x = a\sin\theta\cos\phi$ and so

$$\mathbf{a}\cdot d\mathbf{S} = x\,(\mathbf{i}\cdot\hat{\mathbf{r}})\,dS = (a\sin\theta\cos\phi)(\sin\theta\cos\phi)(a^2\sin\theta\,d\theta\,d\phi).$$

Therefore, inserting the correct limits on θ and ϕ, we have

$$I = \int_S \mathbf{a}\cdot d\mathbf{S} = a^3\int_0^{\pi/2} d\theta\,\sin^3\theta\int_0^{2\pi} d\phi\,\cos^2\phi = \frac{2\pi a^3}{3}.$$

We could, however, follow the general prescription above and project the hemisphere S onto the region R in the xy-plane that is a circle of radius a centred at the origin. Writing the equation of the surface of the hemisphere as $f(x,y) = x^2+y^2+z^2-a^2 = 0$ and using (11.10), we have

$$I = \int_S \mathbf{a}\cdot d\mathbf{S} = \int_S x\,(\mathbf{i}\cdot\hat{\mathbf{r}})\,dS = \int_R x\,(\mathbf{i}\cdot\hat{\mathbf{r}})\,\frac{|\nabla f|\,dA}{\partial f/\partial z}.$$

Now $\nabla f = 2x\mathbf{i}+2y\mathbf{j}+2z\mathbf{k} = 2\mathbf{r}$, so on the surface S we have $|\nabla f| = 2|\mathbf{r}| = 2a$. On S we also have $\partial f/\partial z = 2z = 2\sqrt{a^2-x^2-y^2}$ and $\mathbf{i}\cdot\hat{\mathbf{r}} = x/a$. Therefore, the integral becomes

$$I = \iint_R \frac{x^2}{\sqrt{a^2-x^2-y^2}}\,dx\,dy.$$

Although this integral may be evaluated directly, it is quicker to transform to plane polar coordinates:

$$\begin{aligned} I &= \iint_{R'} \frac{\rho^2\cos^2\phi}{\sqrt{a^2-\rho^2}}\,\rho\,d\rho\,d\phi \\ &= \int_0^{2\pi}\cos^2\phi\,d\phi\int_0^a \frac{\rho^3\,d\rho}{\sqrt{a^2-\rho^2}}. \end{aligned}$$

Making the substitution $\rho = a\sin u$, we finally obtain

$$I = \int_0^{2\pi}\cos^2\phi\,d\phi\int_0^{\pi/2} a^3\sin^3 u\,du = \frac{2\pi a^3}{3}. \blacktriangleleft$$

In the above discussion we assumed that any line parallel to the z-axis intersects S only once. If this is not the case, we must split up the surface into smaller surfaces S_1, S_2 etc. that are of this type. The surface integral over S is then the sum of the surface integrals over S_1, S_2 and so on. This is always necessary for closed surfaces.

Sometimes we may need to project a surface S (or some part of it) onto the zx- or yz-plane, rather than the xy-plane; for such cases, the above analysis is easily modified.

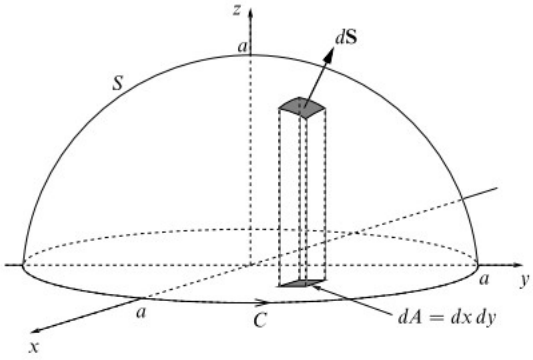

Figure 11.7 The surface of the hemisphere $x^2 + y^2 + z^2 = a^2$, $z \geq 0$.

11.5.2 Vector areas of surfaces

The vector area of a surface S is defined as

$$\mathbf{S} = \int_S d\mathbf{S},$$

where the surface integral may be evaluated as above.

►*Find the vector area of the surface of the hemisphere* $x^2 + y^2 + z^2 = a^2$ *with* $z \geq 0$.

As in the previous example, $d\mathbf{S} = a^2 \sin\theta \, d\theta \, d\phi \, \hat{\mathbf{r}}$ in spherical polar coordinates. Therefore the vector area is given by

$$\mathbf{S} = \iint_S a^2 \sin\theta \; \hat{\mathbf{r}} \, d\theta \, d\phi.$$

Now, since $\hat{\mathbf{r}}$ varies over the surface S, it also must be integrated. This is most easily achieved by writing $\hat{\mathbf{r}}$ in terms of the constant Cartesian basis vectors. On S we have

$$\hat{\mathbf{r}} = \sin\theta\cos\phi\,\mathbf{i} + \sin\theta\sin\phi\,\mathbf{j} + \cos\theta\,\mathbf{k},$$

so the expression for the vector area becomes

$$\begin{aligned}\mathbf{S} &= \mathbf{i}\left(a^2\int_0^{2\pi}\cos\phi\,d\phi\int_0^{\pi/2}\sin^2\theta\,d\theta\right) + \mathbf{j}\left(a^2\int_0^{2\pi}\sin\phi\,d\phi\int_0^{\pi/2}\sin^2\theta\,d\theta\right)\\ &\quad + \mathbf{k}\left(a^2\int_0^{2\pi}d\phi\int_0^{\pi/2}\sin\theta\cos\theta\,d\theta\right)\\ &= \mathbf{0} + \mathbf{0} + \pi a^2\mathbf{k} = \pi a^2\mathbf{k}.\end{aligned}$$

Note that the magnitude of $\mathbf{S}$ is the projected area, of the hemisphere onto the xy-plane, and not the surface area of the hemisphere. ◄

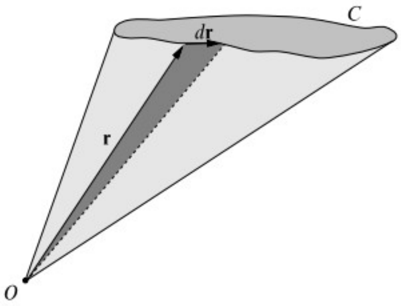

Figure 11.8 The conical surface spanning the perimeter C and having its vertex at the origin.

The hemispherical shell discussed above is an example of an open surface. For a closed surface, however, the vector area is always zero. This may be seen by projecting the surface down onto each Cartesian coordinate plane in turn. For each projection, every positive element of area on the upper surface is cancelled by the corresponding negative element on the lower surface. Therefore, each component of $\mathbf{S} = \oint_S d\mathbf{S}$ vanishes.

An important corollary of this result is that the vector area of an open surface depends only on its perimeter, or boundary curve, C. This may be proved as follows. If surfaces S_1 and S_2 have the same perimeter then $S_1 - S_2$ is a closed surface, for which

$$\oint d\mathbf{S} = \int_{S_1} d\mathbf{S} - \int_{S_2} d\mathbf{S} = \mathbf{0}.$$

Hence $\mathbf{S}_1 = \mathbf{S}_2$. Moreover, we may derive an expression for the vector area of an open surface S solely in terms of a line integral around its perimeter C. Since we may choose any surface with perimeter C, we will consider a cone with its vertex at the origin (see figure 11.8). The vector area of the elementary triangular region shown in the figure is $d\mathbf{S} = \frac{1}{2}\mathbf{r} \times d\mathbf{r}$. Therefore, the vector area of the cone, and hence of *any* open surface with perimeter C, is given by the line integral

$$\mathbf{S} = \frac{1}{2}\oint_C \mathbf{r} \times d\mathbf{r}.$$

For a surface confined to the xy-plane, $\mathbf{r} = x\mathbf{i} + y\mathbf{j}$ and $d\mathbf{r} = dx\,\mathbf{i} + dy\,\mathbf{j}$, and we obtain for this special case that the area of the surface is given by $A = \frac{1}{2}\oint_C (x\,dy - y\,dx)$, as we found in section 11.3.

►*Find the vector area of the surface of the hemisphere* $x^2 + y^2 + z^2 = a^2$, $z \geq 0$, *by evaluating the line integral* $\mathbf{S} = \frac{1}{2}\oint_C \mathbf{r} \times d\mathbf{r}$ *around its perimeter.*

The perimeter C of the hemisphere is the circle $x^2 + y^2 = a^2$, on which we have

$$\mathbf{r} = a\cos\phi\,\mathbf{i} + a\sin\phi\,\mathbf{j}, \qquad d\mathbf{r} = -a\sin\phi\,d\phi\,\mathbf{i} + a\cos\phi\,d\phi\,\mathbf{j}.$$

Therefore the cross product $\mathbf{r} \times d\mathbf{r}$ is given by

$$\mathbf{r} \times d\mathbf{r} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a\cos\phi & a\sin\phi & 0 \\ -a\sin\phi\,d\phi & a\cos\phi\,d\phi & 0 \end{vmatrix} = a^2(\cos^2\phi + \sin^2\phi)\,d\phi\,\mathbf{k} = a^2\,d\phi\,\mathbf{k},$$

and the vector area becomes

$$\mathbf{S} = \tfrac{1}{2}a^2\mathbf{k}\int_0^{2\pi} d\phi = \pi a^2\,\mathbf{k}. ◄$$

11.5.3 Physical examples of surface integrals

There are many examples of surface integrals in the physical sciences. Surface integrals of the form (11.8) occur in computing the total electric charge on a surface or the mass of a shell, $\int_S \rho(\mathbf{r})\,dS$, given the charge or mass density $\rho(\mathbf{r})$. For surface integrals involving vectors, the second form in (11.9) is the most common. For a vector field $\mathbf{a}$, the surface integral $\int_S \mathbf{a}\cdot d\mathbf{S}$ is called the *flux* of $\mathbf{a}$ through S. Examples of physically important flux integrals are numerous. For example, let us consider a surface S in a fluid with density $\rho(\mathbf{r})$ that has a velocity field $\mathbf{v}(\mathbf{r})$. The mass of fluid crossing an element of surface area $d\mathbf{S}$ in time dt is $dM = \rho\mathbf{v}\cdot d\mathbf{S}\,dt$. Therefore the *net* total mass flux of fluid crossing S is $M = \int_S \rho(\mathbf{r})\mathbf{v}(\mathbf{r})\cdot d\mathbf{S}$. As a another example, the electromagnetic flux of energy out of a given volume V bounded by a surface S is $\oint_S(\mathbf{E}\times\mathbf{H})\cdot d\mathbf{S}$.

The solid angle, to be defined below, subtended at a point O by a surface (closed or otherwise) can also be represented by an integral of this form, although it is not strictly a flux integral (unless we imagine isotropic rays radiating from O). The integral

$$\Omega = \int_S \frac{\mathbf{r}\cdot d\mathbf{S}}{r^3} = \int_S \frac{\hat{\mathbf{r}}\cdot d\mathbf{S}}{r^2}, \tag{11.11}$$

gives the *solid angle* Ω *subtended at* O *by a surface* S if $\mathbf{r}$ is the position vector measured from O of an element of the surface. A little thought will show that (11.11) takes account of all three relevant factors: the size of the element of surface, its inclination to the line joining the element to O and the distance from O. Such a general expression is often useful for computing solid angles when the three-dimensional geometry is complicated. Note that (11.11) remains valid when the surface S is not convex and when a single ray from O in certain directions would cut S in more than one place (but we exclude multiply connected regions).

In particular, when the surface is closed $\Omega = 0$ if O is outside S and $\Omega = 4\pi$ if O is an interior point.

Surface integrals resulting in vectors occur less frequently. An example is afforded, however, by the total resultant force experienced by a body immersed in a stationary fluid in which the hydrostatic pressure is given by $p(\mathbf{r})$. The pressure is everywhere inwardly directed and the resultant force is $\mathbf{F} = -\oint_S p\,d\mathbf{S}$, taken over the whole surface.

11.6 Volume integrals

Volume integrals are defined in an obvious way and are generally simpler than line or surface integrals since the element of volume dV is a scalar quantity. We may encounter volume integrals of the forms

$$\int_V \phi\,dV, \qquad \int_V \mathbf{a}\,dV. \tag{11.12}$$

Clearly, the first form results in a scalar, whereas the second form yields a vector. Two closely related physical examples, one of each kind, are provided by the total mass of a fluid contained in a volume V, given by $\int_V \rho(\mathbf{r})\,dV$, and the total linear momentum of that same fluid, given by $\int_V \rho(\mathbf{r})\mathbf{v}(\mathbf{r})\,dV$ where $\mathbf{v}(\mathbf{r})$ is the velocity field in the fluid. As a slightly more complicated example of a volume integral we may consider the following.

►*Find an expression for the angular momentum of a solid body rotating with angular velocity* $\boldsymbol{\omega}$ *about an axis through the origin.*

Consider a small volume element dV situated at position $\mathbf{r}$; its linear momentum is $\rho\,dV\dot{\mathbf{r}}$, where $\rho = \rho(\mathbf{r})$ is the density distribution, and its angular momentum about O is $\mathbf{r}\times\rho\dot{\mathbf{r}}\,dV$. Thus for the whole body the angular momentum $\mathbf{L}$ is

$$\mathbf{L} = \int_V (\mathbf{r}\times\dot{\mathbf{r}})\rho\,dV.$$

Putting $\dot{\mathbf{r}} = \boldsymbol{\omega}\times\mathbf{r}$ yields

$$\mathbf{L} = \int_V [\mathbf{r}\times(\boldsymbol{\omega}\times\mathbf{r})]\,\rho\,dV = \int_V \boldsymbol{\omega} r^2\rho\,dV - \int_V (\mathbf{r}\cdot\boldsymbol{\omega})\mathbf{r}\rho\,dV. \blacktriangleleft$$

The evaluation of the first type of volume integral in (11.12) has already been considered in our discussion of multiple integrals in chapter 6. The evaluation of the second type of volume integral follows directly since we can write

$$\int_V \mathbf{a}\,dV = \mathbf{i}\int_V a_x\,dV + \mathbf{j}\int_V a_y\,dV + \mathbf{k}\int_V a_z\,dV, \tag{11.13}$$

where a_x, a_y, a_z are the Cartesian components of $\mathbf{a}$. Of course, we could have written $\mathbf{a}$ in terms of the basis vectors of some other coordinate system (e.g. spherical polars) but, since such basis vectors are not, in general, constant, they

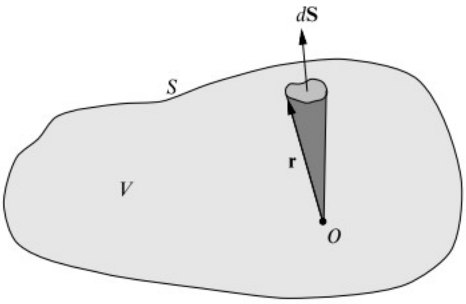

Figure 11.9 A general volume V containing the origin and bounded by the closed surface S.

cannot be taken out of the integral sign as in (11.13) and must be included as part of the integrand.

11.6.1 Volumes of three-dimensional regions

As discussed in chapter 6, the volume of a three-dimensional region V is simply $V = \int_V dV$, which may be evaluated directly once the limits of integration have been found. However, the volume of the region obviously depends only on the surface S that bounds it. We should therefore be able to express the volume V in terms of a surface integral over S. This is indeed possible, and the appropriate expression may derived as follows. Referring to figure 11.9, let us suppose that the origin O is contained within V. The volume of the small shaded cone is $dV = \frac{1}{3}\mathbf{r} \cdot d\mathbf{S}$; the total volume of the region is thus given by

$$V = \frac{1}{3} \oint_S \mathbf{r} \cdot d\mathbf{S}.$$

It may be shown that this expression is valid even when O is not contained in V. Although this surface integral form is available, in practice it is usually simpler to evaluate the volume integral directly.

►*Find the volume enclosed between a sphere of radius a centred on the origin and a circular cone of half-angle α with its vertex at the origin.*

The element of vector area $d\mathbf{S}$ on the surface of the sphere is given in spherical polar coordinates by $a^2 \sin\theta \, d\theta \, d\phi \, \hat{\mathbf{r}}$. Now taking the axis of the cone to lie along the z-axis (from which θ is measured) the required volume is given by

$$\begin{aligned} V = \frac{1}{3} \oint_S \mathbf{r} \cdot d\mathbf{S} &= \frac{1}{3} \int_0^{2\pi} d\phi \int_0^{\alpha} a^2 \sin\theta \, \mathbf{r} \cdot \hat{\mathbf{r}} \, d\theta \\ &= \frac{1}{3} \int_0^{2\pi} d\phi \int_0^{\alpha} a^3 \sin\theta \, d\theta = \frac{2\pi a^3}{3}(1 - \cos\alpha). \end{aligned}$$ ◄

11.7 Integral forms for grad, div and curl

In the previous chapter we defined the vector operators grad, div and curl in purely mathematical terms, which depended on the coordinate system in which they were expressed. An interesting application of line, surface and volume integrals is the expression of grad, div and curl in coordinate-free, geometrical terms. If ϕ is a scalar field and $\mathbf{a}$ is a vector field then it may be shown that at any point P

$$\nabla\phi = \lim_{V\to 0}\left(\frac{1}{V}\oint_S \phi\, d\mathbf{S}\right) \tag{11.14}$$

$$\nabla\cdot\mathbf{a} = \lim_{V\to 0}\left(\frac{1}{V}\oint_S \mathbf{a}\cdot d\mathbf{S}\right) \tag{11.15}$$

$$\nabla\times\mathbf{a} = \lim_{V\to 0}\left(\frac{1}{V}\oint_S d\mathbf{S}\times\mathbf{a}\right) \tag{11.16}$$

where V is a small volume enclosing P and S is its bounding surface. Indeed, we may consider these equations as the (geometrical) *definitions* of grad, div and curl. An alternative, but equivalent, geometrical definition of $\nabla\times\mathbf{a}$ at a point P, which is often easier to use than (11.16), is given by

$$(\nabla\times\mathbf{a})\cdot\hat{\mathbf{n}} = \lim_{A\to 0}\left(\frac{1}{A}\oint_C \mathbf{a}\cdot d\mathbf{r}\right), \tag{11.17}$$

where C is a plane contour of area A enclosing the point P and $\hat{\mathbf{n}}$ is the unit normal to the enclosed planar area.

It may be shown, *in any coordinate system*, that all the above equations are consistent with our definitions in the previous chapter, although the difficulty of proof depends on the chosen coordinate system. The most general coordinate system encountered in that chapter was one with orthogonal curvilinear coordinates u_1, u_2, u_3, of which Cartesians, cylindrical polars and spherical polars are all special cases. Although it may be shown that (11.14) leads to the usual expression for grad in curvilinear coordinates, the proof requires complicated manipulations of the derivatives of the basis vectors with respect to the coordinates and is not presented here. In Cartesian coordinates, however, the proof is quite simple.

►*Show that the geometrical definition of* grad *leads to the usual expression for* $\nabla\phi$ *in Cartesian coordinates.*

Consider the surface S of a small rectangular volume element $\Delta V = \Delta x\,\Delta y\,\Delta z$ that has its faces parallel to the x, y, and z coordinate surfaces; the point P (see above) is at one corner. We must calculate the surface integral (11.14) over each of its six faces. Remembering that the normal to the surface points outwards from the volume on each face, the two faces with $x =$ constant have areas $\Delta\mathbf{S} = -\mathbf{i}\,\Delta y\,\Delta z$ and $\Delta\mathbf{S} = \mathbf{i}\,\Delta y\,\Delta z$ respectively. Furthermore, over each small surface element, we may take ϕ to be constant, so that the net contribution to the surface integral from these two faces is then

$$[(\phi + \Delta\phi) - \phi]\,\Delta y\,\Delta z\,\mathbf{i} = \left(\phi + \frac{\partial \phi}{\partial x}\,\Delta x - \phi\right)\,\Delta y\,\Delta z\,\mathbf{i}$$
$$= \frac{\partial \phi}{\partial x}\,\Delta x\,\Delta y\,\Delta z\,\mathbf{i}.$$

The surface integral over the pairs of faces with $y =$ constant and $z =$ constant respectively may be found in a similar way, and we obtain

$$\oint_S \phi\, d\mathbf{S} = \left(\frac{\partial \phi}{\partial x}\,\mathbf{i} + \frac{\partial \phi}{\partial y}\,\mathbf{j} + \frac{\partial \phi}{\partial z}\,\mathbf{k}\right)\,\Delta x\,\Delta y\,\Delta z.$$

Therefore $\nabla\phi$ at the point P is given by

$$\nabla\phi = \lim_{\Delta x, \Delta y, \Delta z \to 0}\left[\frac{1}{\Delta x\,\Delta y\,\Delta z}\left(\frac{\partial \phi}{\partial x}\,\mathbf{i} + \frac{\partial \phi}{\partial y}\,\mathbf{j} + \frac{\partial \phi}{\partial z}\,\mathbf{k}\right)\,\Delta x\,\Delta y\,\Delta z\right]$$
$$= \frac{\partial \phi}{\partial x}\,\mathbf{i} + \frac{\partial \phi}{\partial y}\,\mathbf{j} + \frac{\partial \phi}{\partial z}\,\mathbf{k}. \blacktriangleleft$$

We now turn to (11.15) and (11.17). These geometrical definitions may be shown straightforwardly to lead to the usual expressions for div and curl in orthogonal curvilinear coordinates.

►*By considering the infinitesimal volume element $dV = h_1h_2h_3\,\Delta u_1\,\Delta u_2\,\Delta u_3$ shown in figure 11.10, show that (11.15) leads to the usual expression for $\nabla\cdot\mathbf{a}$ in orthogonal curvilinear coordinates.*

Let us write the vector field in terms of its components with respect to the basis vectors of the curvilinear coordinate system as $\mathbf{a} = a_1\hat{\mathbf{e}}_1 + a_2\hat{\mathbf{e}}_2 + a_3\hat{\mathbf{e}}_3$. We consider first the contribution to the RHS of (11.15) from the two faces with $u_1 =$ constant, i.e. $PQRS$ and the face opposite it (see figure 11.10). Now, the volume element is formed from the orthogonal vectors $h_1\,\Delta u_1\,\hat{\mathbf{e}}_1$, $h_2\,\Delta u_2\,\hat{\mathbf{e}}_2$ and $h_3\,\Delta u_3\,\hat{\mathbf{e}}_3$ at the point P and so for $PQRS$ we have

$$\Delta\mathbf{S} = h_2h_3\,\Delta u_2\,\Delta u_3\,\hat{\mathbf{e}}_3\times\hat{\mathbf{e}}_2 = -h_2h_3\,\Delta u_2\,\Delta u_3\,\hat{\mathbf{e}}_1.$$

Reasoning along the same lines as in the previous example, we conclude that the contribution to the surface integral of $\mathbf{a}\cdot d\mathbf{S}$ over $PQRS$ and its opposite face taken together is given by

$$\frac{\partial}{\partial u_1}(\mathbf{a}\cdot\Delta\mathbf{S})\,\Delta u_1 = \frac{\partial}{\partial u_1}(a_1h_2h_3)\,\Delta u_1\,\Delta u_2\,\Delta u_3.$$

The surface integrals over the pairs of faces with $u_2 =$ constant and $u_3 =$ constant respectively may be found in a similar way, and we obtain

$$\oint_S \mathbf{a}\cdot d\mathbf{S} = \left[\frac{\partial}{\partial u_1}(a_1h_2h_3) + \frac{\partial}{\partial u_2}(a_2h_3h_1) + \frac{\partial}{\partial u_3}(a_3h_1h_2)\right]\,\Delta u_1\,\Delta u_2\,\Delta u_3.$$

Therefore $\nabla\cdot\mathbf{a}$ at the point P is given by

$$\nabla\cdot\mathbf{a} = \lim_{\Delta u_1, \Delta u_2, \Delta u_3 \to 0}\left[\frac{1}{h_1h_2h_3\,\Delta u_1\,\Delta u_2\,\Delta u_3}\oint_S \mathbf{a}\cdot d\mathbf{S}\right]$$
$$= \frac{1}{h_1h_2h_3}\left[\frac{\partial}{\partial u_1}(a_1h_2h_3) + \frac{\partial}{\partial u_2}(a_2h_3h_1) + \frac{\partial}{\partial u_3}(a_3h_1h_2)\right]. \blacktriangleleft$$

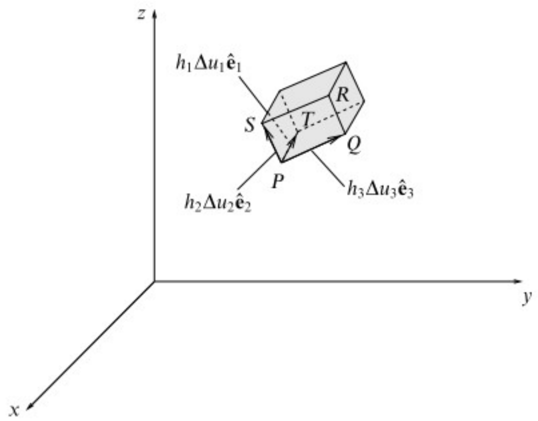

z
y
x
$h_1\Delta u_1\hat{\mathbf{e}}_1$
$h_2\Delta u_2\hat{\mathbf{e}}_2$
$h_3\Delta u_3\hat{\mathbf{e}}_3$
S
T
R
Q
P

Figure 11.10 A general volume ΔV in orthogonal curvilinear coordinates u_1, u_2, u_3. PT gives the vector $h_1\,\Delta u_1\,\hat{\mathbf{e}}_1$, PS gives $h_2\,\Delta u_2\,\hat{\mathbf{e}}_2$ and PQ gives $h_3\,\Delta u_3\,\hat{\mathbf{e}}_3$.

►*By considering the infinitesimal planar surface element $PQRS$ in figure 11.10, show that (11.17) leads to the usual expression for $\nabla\times\mathbf{a}$ in orthogonal curvilinear coordinates.*

The planar surface $PQRS$ is defined by the orthogonal vectors $h_2\,\Delta u_2\,\hat{\mathbf{e}}_2$ and $h_3\,\Delta u_3\,\hat{\mathbf{e}}_3$ at the point P. If we traverse the loop in the direction $PSRQ$ then, by the right-hand convention, the unit normal to the plane is $\hat{\mathbf{e}}_1$. Writing $\mathbf{a} = a_1\hat{\mathbf{e}}_1 + a_2\hat{\mathbf{e}}_2 + a_3\hat{\mathbf{e}}_3$, the line integral around the loop in this direction is given by

$$\begin{aligned}\oint_{PSRQ}\mathbf{a}\cdot d\mathbf{r} &= a_2h_2\,\Delta u_2 + \left[a_3h_3 + \frac{\partial}{\partial u_2}(a_3h_3)\,\Delta u_2\right]\,\Delta u_3\\ &\quad - \left[a_2h_2 + \frac{\partial}{\partial u_3}(a_2h_2)\,\Delta u_3\right]\,\Delta u_2 - a_3h_3\,\Delta u_3\\ &= \left[\frac{\partial}{\partial u_2}(a_3h_3) - \frac{\partial}{\partial u_3}(a_2h_2)\right]\,\Delta u_2\,\Delta u_3.\end{aligned}$$

Therefore from (11.17) the component of $\nabla\times\mathbf{a}$ in the direction $\hat{\mathbf{e}}_1$ at P is given by

$$\begin{aligned}(\nabla\times\mathbf{a})_1 &= \lim_{\Delta u_2,\Delta u_3\to 0}\left[\frac{1}{h_2h_3\,\Delta u_2\,\Delta u_3}\oint_{PSRQ}\mathbf{a}\cdot d\mathbf{r}\right]\\ &= \frac{1}{h_2h_3}\left[\frac{\partial}{\partial u_2}(h_3a_3) - \frac{\partial}{\partial u_3}(h_2a_2)\right].\end{aligned}$$

The other two components are found by cyclically permuting the subscripts 1, 2, 3. ◄

Finally, we note that we can also write the ∇^2 operator as a surface integral by setting $\mathbf{a} = \nabla\phi$ in (11.15), to obtain

$$\nabla^2\phi = \nabla\cdot\nabla\phi = \lim_{V\to 0}\left(\frac{1}{V}\oint_S \nabla\phi\cdot d\mathbf{S}\right).$$

11.8 Divergence theorem and related theorems

The divergence theorem relates the total flux of a vector field out of a closed surface S to the integral of the divergence of the vector field over the enclosed volume V; it follows almost immediately from our geometrical definition of divergence (11.15).

Imagine a volume V, in which a vector field $\mathbf{a}$ is continuous and differentiable, to be divided up into a large number of small volumes V_i. Using (11.15), we have for each small volume

$$(\nabla \cdot \mathbf{a})V_i \approx \oint_{S_i} \mathbf{a} \cdot d\mathbf{S},$$

where S_i is the surface of the small volume V_i. Summing over i we find that contributions from surface elements interior to S cancel since each surface element appears in two terms with opposite signs, the outward normals in the two terms being equal and opposite. Only contributions from surface elements that are also parts of S survive. If each V_i is allowed to tend to zero then we obtain the *divergence theorem*,

$$\int_V \nabla \cdot \mathbf{a}\, dV = \oint_S \mathbf{a} \cdot d\mathbf{S}. \tag{11.18}$$

We note that the divergence theorem holds for both simply and multiply connected surfaces, provided that they are closed and enclose some non-zero volume V. The divergence theorem may also be extended to tensor fields (see chapter 26).

The theorem finds most use as a tool in formal manipulations, but sometimes it is of value in transforming surface integrals of the form $\int_S \mathbf{a} \cdot d\mathbf{S}$ into volume integrals or vice versa. For example, setting $\mathbf{a} = \mathbf{r}$ we immediately obtain

$$\int_V \nabla \cdot \mathbf{r}\, dV = \int_V 3\, dV = 3V = \oint_S \mathbf{r} \cdot d\mathbf{S},$$

which gives the expression for the volume of a region found in subsection 11.6.1. The use of the divergence theorem is further illustrated in the following example.

►*Evaluate the surface integral* $I = \int_S \mathbf{a} \cdot d\mathbf{S}$, *where* $\mathbf{a} = (y - x)\,\mathbf{i} + x^2 z\,\mathbf{j} + (z + x^2)\,\mathbf{k}$ *and* S *is the open surface of the hemisphere* $x^2 + y^2 + z^2 = a^2$, $z \geq 0$.

We could evaluate this surface integral directly, but the algebra is somewhat lengthy. We will therefore evaluate it by use of the divergence theorem. Since the latter only holds for closed surfaces enclosing a non-zero volume V, let us first consider the closed surface $S' = S + S_1$, where S_1 is the circular area in the xy-plane given by $x^2 + y^2 \leq a^2$, $z = 0$; S' then encloses a hemispherical volume V. By the divergence theorem we have

$$\int_V \nabla \cdot \mathbf{a}\, dV = \oint_{S'} \mathbf{a} \cdot d\mathbf{S} = \int_S \mathbf{a} \cdot d\mathbf{S} + \int_{S_1} \mathbf{a} \cdot d\mathbf{S}.$$

Now $\nabla \cdot \mathbf{a} = -1 + 0 + 1 = 0$, so we can write

$$\int_S \mathbf{a} \cdot d\mathbf{S} = -\int_{S_1} \mathbf{a} \cdot d\mathbf{S}.$$

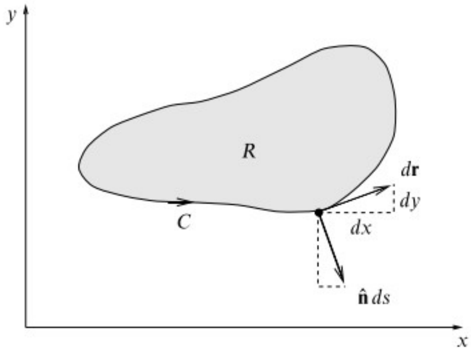

Figure 11.11 A closed curve C in the xy-plane bounding a region R. Vectors tangent and normal to the curve at a given point are also shown.

The surface integral over S_1 is easily evaluated. Remembering that the normal to the surface points outward from the volume, a surface element on S_1 is simply $d\mathbf{S} = -\mathbf{k}\,dx\,dy$. On S_1 we also have $\mathbf{a} = (y - x)\,\mathbf{i} + x^2\,\mathbf{k}$, so that

$$I = -\int_{S_1} \mathbf{a}\cdot d\mathbf{S} = \iint_R x^2\,dx\,dy,$$

where R is the circular region in the xy-plane given by $x^2 + y^2 \le a^2$. Transforming to plane polar coordinates we have

$$I = \iint_{R'} \rho^2\cos^2\phi\;\rho\,d\rho\,d\phi = \int_0^{2\pi}\cos^2\phi\,d\phi\int_0^a \rho^3\,d\rho = \frac{\pi a^4}{4}. \blacktriangleleft$$

It is also interesting to consider the two-dimensional version of the divergence theorem. As an example, let us consider a two-dimensional planar region R in the xy-plane bounded by some closed curve C (see figure 11.11). At any point on the curve the vector $d\mathbf{r} = dx\,\mathbf{i} + dy\,\mathbf{j}$ is a tangent to the curve and the vector $\hat{\mathbf{n}}\,ds = dy\,\mathbf{i} - dx\,\mathbf{j}$ is a normal pointing out of the region R. If the vector field $\mathbf{a}$ is continuous and differentiable in R then the two-dimensional divergence theorem in Cartesian coordinates gives

$$\iint_R \left(\frac{\partial a_x}{\partial x} + \frac{\partial a_y}{\partial y}\right) dx\,dy = \oint \mathbf{a}\cdot\hat{\mathbf{n}}\,ds = \oint_C (a_x\,dy - a_y\,dx).$$

Letting $P = -a_y$ and $Q = a_x$, we recover Green's theorem in a plane, which was discussed in section 11.3.

11.8.1 Green's theorems

Consider two scalar functions ϕ and ψ that are continuous and differentiable in some volume V bounded by a surface S. Applying the divergence theorem to the vector field $\phi\nabla\psi$ we obtain

$$\oint_S \phi\nabla\psi \cdot d\mathbf{S} = \int_V \nabla\cdot(\phi\nabla\psi)\,dV$$
$$= \int_V \left[\phi\nabla^2\psi + (\nabla\phi)\cdot(\nabla\psi)\right]\,dV. \qquad (11.19)$$

Reversing the roles of ϕ and ψ in (11.19) and subtracting the two equations gives

$$\oint_S (\phi\nabla\psi - \psi\nabla\phi)\cdot d\mathbf{S} = \int_V (\phi\nabla^2\psi - \psi\nabla^2\phi)\,dV. \qquad (11.20)$$

Equation (11.19) is usually known as Green's first theorem and (11.20) as his second. Green's second theorem is useful in the development of the Green's functions used in the solution of partial differential equations (see chapter 21).

11.8.2 Other related integral theorems

There exist two other integral theorems which are closely related to the divergence theorem and which are of some use in physical applications. If ϕ is a scalar field and $\mathbf{b}$ is a vector field and both ϕ and $\mathbf{b}$ satisfy our usual differentiability conditions in some volume V bounded by a closed surface S then

$$\int_V \nabla\phi\,dV = \oint_S \phi\,d\mathbf{S}, \qquad (11.21)$$
$$\int_V \nabla\times\mathbf{b}\,dV = \oint_S d\mathbf{S}\times\mathbf{b}. \qquad (11.22)$$

►*Use the divergence theorem to prove (11.21).*

In the divergence theorem (11.18) let $\mathbf{a} = \phi\mathbf{c}$, where $\mathbf{c}$ is a constant vector. We then have

$$\int_V \nabla\cdot(\phi\mathbf{c})\,dV = \oint_S \phi\mathbf{c}\cdot d\mathbf{S}.$$

Expanding out the integrand on the LHS we have

$$\nabla\cdot(\phi\mathbf{c}) = \phi\nabla\cdot\mathbf{c} + \mathbf{c}\cdot\nabla\phi = \mathbf{c}\cdot\nabla\phi,$$

since $\mathbf{c}$ is constant. Also, $\phi\mathbf{c}\cdot d\mathbf{S} = \mathbf{c}\cdot\phi d\mathbf{S}$, so we obtain

$$\int_V \mathbf{c}\cdot(\nabla\phi)\,dV = \oint_S \mathbf{c}\cdot\phi\,d\mathbf{S}.$$

Since $\mathbf{c}$ is constant we may take it out of both integrals to give

$$\mathbf{c}\cdot\int_V \nabla\phi\,dV = \mathbf{c}\cdot\oint_S \phi\,d\mathbf{S},$$

and since $\mathbf{c}$ is arbitrary we obtain the stated result (11.21). ◄

Equation (11.22) may be proved in a similar way by letting $\mathbf{a} = \mathbf{b}\times\mathbf{c}$ in the divergence theorem, where $\mathbf{c}$ is again a constant vector.

11.8.3 Physical applications of the divergence theorem

The divergence theorem is useful in deriving many of the most important partial differential equations in physics (see chapter 20). The basic idea is to use the divergence theorem to convert an integral form, often derived from observation, into an equivalent differential form (used in theoretical statements).

►*For a compressible fluid with time-varying position-dependent density $\rho(\mathbf{r}, t)$ and velocity field $v(\mathbf{r}, t)$, in which fluid is neither being created nor destroyed, show that*

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0.$$

For an arbitrary volume V in the fluid, the conservation of mass tells us that the rate of increase or decrease of the mass M of fluid in the volume must equal the net rate at which fluid is entering or leaving the volume, i.e.

$$\frac{dM}{dt} = -\oint_S \rho \mathbf{v} \cdot d\mathbf{S},$$

where S is the surface bounding V. But the mass of fluid in V is simply $M = \int_V \rho\, dV$, so we have

$$\frac{d}{dt}\int_V \rho\, dV + \oint_S \rho \mathbf{v} \cdot d\mathbf{S} = 0.$$

Taking the derivative inside the first integral on the RHS and using the divergence theorem to rewrite the second integral, we obtain

$$\int_V \frac{\partial \rho}{\partial t}\, dV + \int_V \nabla \cdot (\rho \mathbf{v})\, dV = \int_V \left[\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v})\right] dV = 0.$$

Since the volume V is arbitrary, the integrand (which is assumed continuous) must be identically zero, so we obtain

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0.$$

This is known as the *continuity equation*. It can also be applied to other systems, for example those in which ρ is the density of electric charge or the heat content, etc. For the flow of an incompressible fluid, $\rho =$ constant and the continuity equation becomes simply $\nabla \cdot \mathbf{v} = 0$. ◄

In the previous example, we assumed that there were no sources or sinks in the volume V, i.e. that there was no part of V in which fluid was being created or destroyed. We now consider the case where a finite number of *point* sources and/or sinks are present in an incompressible fluid. Let us first consider the simple case where a single source is located at the origin, out of which a quantity of fluid flows radially at a rate Q ($\mathrm{m}^3\ \mathrm{s}^{-1}$). The velocity field is given by

$$\mathbf{v} = \frac{Q\mathbf{r}}{4\pi r^3} = \frac{Q\hat{\mathbf{r}}}{4\pi r^2}.$$

Now, for a sphere S_1 of radius r centred on the source, the flux across S_1 is

$$\oint_{S_1} \mathbf{v} \cdot d\mathbf{S} = |\mathbf{v}| 4\pi r^2 = Q.$$

Since $\mathbf{v}$ has a singularity at the origin it is not differentiable there, i.e. $\nabla \cdot \mathbf{v}$ is not defined there, but at all other points $\nabla \cdot \mathbf{v} = 0$, as required for an incompressible fluid. Therefore, from the divergence theorem, for any closed surface S_2 that does not enclose the origin we have

$$\oint_{S_2} \mathbf{v} \cdot d\mathbf{S} = \int_V \nabla \cdot \mathbf{v}\, dV = 0.$$

Thus we see that the surface integral $\oint_S \mathbf{v} \cdot d\mathbf{S}$ has value Q or zero depending on whether or not S encloses the source. In order that the divergence theorem is valid for *all* surfaces S, irrespective of whether they enclose the source, we write

$$\nabla \cdot \mathbf{v} = Q\delta(\mathbf{r}),$$

where $\delta(\mathbf{r})$ is the three-dimensional Dirac delta function. The properties of this function are discussed fully in chapter 13, but for the moment we note that it is defined in such a way that

$$\delta(\mathbf{r} - \mathbf{a}) = 0 \qquad \text{for } \mathbf{r} \neq \mathbf{a},$$

$$\int_V f(\mathbf{r})\delta(\mathbf{r} - \mathbf{a})\, dV = \begin{cases} f(\mathbf{a}) & \text{if } \mathbf{a} \text{ lies in } V \\ 0 & \text{otherwise} \end{cases}$$

for any well-behaved function $f(\mathbf{r})$. Therefore, for any volume V containing the source at the origin, we have

$$\int_V \nabla \cdot \mathbf{v}\, dV = Q \int_V \delta(\mathbf{r})\, dV = Q,$$

which is consistent with $\oint_S \mathbf{v} \cdot d\mathbf{S} = Q$ for a closed surface enclosing the source. Hence, by introducing the Dirac delta function the divergence theorem can be made valid even for non-differentiable point sources.

The generalisation to several sources and sinks is straightforward. For example, if a source is located at $\mathbf{r} = \mathbf{a}$ and a sink at $\mathbf{r} = \mathbf{b}$ then the velocity field is

$$\mathbf{v} = \frac{(\mathbf{r} - \mathbf{a})Q}{4\pi|\mathbf{r} - \mathbf{a}|^3} - \frac{(\mathbf{r} - \mathbf{b})Q}{4\pi|\mathbf{r} - \mathbf{b}|^3}$$

and its divergence is given by

$$\nabla \cdot \mathbf{v} = Q\delta(\mathbf{r} - \mathbf{a}) - Q\delta(\mathbf{r} - \mathbf{b}).$$

Therefore, the integral $\oint_S \mathbf{v} \cdot d\mathbf{S}$ has the value Q if S encloses the source, $-Q$ if S encloses the sink and 0 if S encloses neither the source nor sink or encloses them both. This analysis also applies to other physical systems – for example, in electrostatics we can regard the sources and sinks as positive and negative point charges respectively and replace $\mathbf{v}$ by the electric field $\mathbf{E}$.

11.9 Stokes' theorem and related theorems

Stokes' theorem is the 'curl analogue' of the divergence theorem and relates the integral of the curl of a vector field over an open surface S to the line integral of the vector field around the perimeter C bounding the surface.

Following the same lines as for the derivation of the divergence theorem, we can divide the surface S into many small areas S_i with boundaries C_i and unit normals $\hat{\mathbf{n}}_i$. Using (11.17), we have for each small area

$$(\nabla \times \mathbf{a}) \cdot \hat{\mathbf{n}}_i \; S_i \approx \oint_{C_i} \mathbf{a} \cdot d\mathbf{r}.$$

Summing over i we find that on the RHS all parts of all interior boundaries that are not part of C are included twice, being traversed in opposite directions on each occasion and thus contributing nothing. Only contributions from line elements that are also parts of C survive. If each S_i is allowed to tend to zero then we obtain Stokes' theorem,

$$\int_S (\nabla \times \mathbf{a}) \cdot d\mathbf{S} = \oint_C \mathbf{a} \cdot d\mathbf{r}. \qquad (11.23)$$

We note that Stokes' theorem holds for both simply and multiply connected open surfaces, provided that they are two-sided. Stokes' theorem may also be extended to tensor fields (see chapter 26).

Just as the divergence theorem (11.18) can be used to relate volume and surface integrals for certain types of integrand, Stokes' theorem can be used in evaluating surface integrals of the form $\oint_S (\nabla \times \mathbf{a}) \cdot d\mathbf{S}$ as line integrals or vice versa.

►*Given the vector field* $\mathbf{a} = y\,\mathbf{i} - x\,\mathbf{j} + z\,\mathbf{k}$, *verify Stokes' theorem for the hemispherical surface* $x^2 + y^2 + z^2 = a^2$, $z \geq 0$.

Let us first evaluate the surface integral

$$\int_S (\nabla \times \mathbf{a}) \cdot d\mathbf{S}$$

over the hemisphere. It is easily shown that $\nabla \times \mathbf{a} = -2\,\mathbf{k}$, and the surface element is $d\mathbf{S} = a^2 \sin\theta \, d\theta \, d\phi \; \hat{\mathbf{r}}$ in spherical polar coordinates. Therefore

$$\begin{aligned}\int_S (\nabla \times \mathbf{a}) \cdot d\mathbf{S} &= \int_0^{2\pi} d\phi \int_0^{\pi/2} d\theta \, \left(-2a^2 \sin\theta\right) \hat{\mathbf{r}} \cdot \mathbf{k} \\ &= -2a^2 \int_0^{2\pi} d\phi \int_0^{\pi/2} \sin\theta \left(\frac{z}{a}\right) d\theta \\ &= -2a^2 \int_0^{2\pi} d\phi \int_0^{\pi/2} \sin\theta \cos\theta \, d\theta = -2\pi a^2.\end{aligned}$$

We now evaluate the line integral around the perimeter curve C of the surface, which is the circle $x^2 + y^2 = a^2$ in the xy-plane. This is given by

$$\oint_C \mathbf{a}\cdot d\mathbf{r} = \oint_C (y\,\mathbf{i} - x\,\mathbf{j} + z\,\mathbf{k})\cdot(dx\,\mathbf{i} + dy\,\mathbf{j} + dz\,\mathbf{k})$$
$$= \oint_C (y\,dx - x\,dy).$$

Using plane polar coordinates, on C we have $x = a\cos\phi$, $y = a\sin\phi$ so that $dx = -a\sin\phi\,d\phi$, $dy = a\cos\phi\,d\phi$, and the line integral becomes

$$\oint_C (y\,dx - x\,dy) = -a^2\int_0^{2\pi}(\sin^2\phi + \cos^2\phi)\,d\phi = -a^2\int_0^{2\pi} d\phi = -2\pi a^2.$$

Since the surface and line integrals have the same value, we have verified Stokes' theorem in this case. ◄

The two-dimensional version of Stokes' theorem also yields Green's theorem in a plane. Consider the region R in the xy-plane shown in figure 11.11, in which a vector field $\mathbf{a}$ is defined. Since $\mathbf{a} = a_x\,\mathbf{i} + a_y\,\mathbf{j}$, we have $\nabla\times\mathbf{a} = (\partial a_y/\partial x - \partial a_x/\partial y)\,\mathbf{k}$, and Stokes' theorem becomes

$$\iint_R \left(\frac{\partial a_y}{\partial x} - \frac{\partial a_x}{\partial y}\right) dx\,dy = \oint_C (a_x\,dx + a_y\,dy).$$

Letting $P = a_x$ and $Q = a_y$ we recover Green's theorem in a plane, (11.4).

11.9.1 Related integral theorems

As for the divergence theorem, there exist two other integral theorems that are closely related to Stokes' theorem. If ϕ is a scalar field and $\mathbf{b}$ is a vector field, and both ϕ and $\mathbf{b}$ satisfy our usual differentiability conditions on some two-sided open surface S bounded by a closed perimeter curve C, then

$$\int_S d\mathbf{S}\times\nabla\phi = \oint_C \phi\,d\mathbf{r}, \tag{11.24}$$

$$\int_S (d\mathbf{S}\times\nabla)\times\mathbf{b} = \oint_C d\mathbf{r}\times\mathbf{b}. \tag{11.25}$$

►*Use Stokes' theorem to prove (11.24).*

In Stokes' theorem, (11.23), let $\mathbf{a} = \phi\mathbf{c}$, where $\mathbf{c}$ is a constant vector. We then have

$$\int_S [\nabla\times(\phi\mathbf{c})]\cdot d\mathbf{S} = \oint_C \phi\mathbf{c}\cdot d\mathbf{r}. \tag{11.26}$$

Expanding out the integrand on the LHS we have

$$\nabla\times(\phi\mathbf{c}) = \nabla\phi\times\mathbf{c} + \phi\nabla\times\mathbf{c} = \nabla\phi\times\mathbf{c},$$

since $\mathbf{c}$ is constant, and the scalar triple product on the LHS of (11.26) can therefore be written

$$[\nabla\times(\phi\mathbf{c})]\cdot d\mathbf{S} = (\nabla\phi\times\mathbf{c})\cdot d\mathbf{S} = \mathbf{c}\cdot(d\mathbf{S}\times\nabla\phi).$$

Substituting this into (11.26) and taking **c** out of both integrals because it is constant, we find

$$\mathbf{c}\cdot\int_S d\mathbf{S}\times\nabla\phi=\mathbf{c}\cdot\oint_C \phi\, d\mathbf{r}.$$

Since **c** is an arbitrary constant vector we therefore obtain the stated result (11.24). ◄

Equation (11.25) may be proved in a similar way, by letting $\mathbf{a}=\mathbf{b}\times\mathbf{c}$ in Stokes' theorem, where **c** is again a constant vector. We also note that by setting $\mathbf{b}=\mathbf{r}$ in (11.25) we find

$$\int_S (d\mathbf{S}\times\nabla)\times\mathbf{r}=\oint_C d\mathbf{r}\times\mathbf{r}.$$

Expanding out the integrand on the LHS gives

$$(d\mathbf{S}\times\nabla)\times\mathbf{r}=d\mathbf{S}-d\mathbf{S}(\nabla\cdot\mathbf{r})=d\mathbf{S}-3\,d\mathbf{S}=-2\,d\mathbf{S}.$$

Therefore, as we found in subsection 11.5.2, the vector area of an open surface S is given by

$$\mathbf{S}=\int_S d\mathbf{S}=\frac{1}{2}\oint_C \mathbf{r}\times d\mathbf{r}.$$

11.9.2 Physical applications of Stokes' theorem

Like the divergence theorem, Stokes' theorem is useful in converting integral equations into differential equations.

►*From Ampère's law, derive Maxwell's equation in the case where the currents are steady, i.e.* $\nabla\times\mathbf{B}-\mu_0\mathbf{J}=\mathbf{0}$.

Ampère's rule for a distributed current with current density **J** is

$$\oint_C \mathbf{B}\cdot d\mathbf{r}=\mu_0\int_S \mathbf{J}\cdot d\mathbf{S},$$

for any circuit C bounding a surface S. Using Stokes' theorem, the LHS can be transformed into $\int_S(\nabla\times\mathbf{B})\cdot d\mathbf{S}$; hence

$$\int_S(\nabla\times\mathbf{B}-\mu_0\mathbf{J})\cdot d\mathbf{S}=0$$

for *any* surface S. This can only be so if $\nabla\times\mathbf{B}-\mu_0\mathbf{J}=\mathbf{0}$, which is the required relation. Similarly, from Faraday's law of electromagnetic induction we can derive Maxwell's equation $\nabla\times\mathbf{E}=-\partial\mathbf{B}/\partial t$. ◄

In subsection 11.8.3 we discussed the flow of an incompressible fluid in the presence of several sources and sinks. Let us now consider *vortex* flow in an incompressible fluid with a velocity field

$$\mathbf{v}=\frac{1}{\rho}\hat{\mathbf{e}}_\phi,$$

in cylindrical polar coordinates ρ, ϕ, z. For this velocity field $\nabla\times\mathbf{v}$ equals zero everywhere except on the axis $\rho = 0$, where $\mathbf{v}$ has a singularity. Therefore $\oint_C \mathbf{v} \cdot d\mathbf{r}$ equals zero for any path C that does not enclose the vortex line on the axis and 2π if C does enclose the axis. In order for Stokes' theorem to be valid for all paths C, we therefore set

$$\nabla \times \mathbf{v} = 2\pi\delta(\rho),$$

where $\delta(\rho)$ is the Dirac delta function, to be discussed in subsection 13.1.3. Now, since $\nabla \times \mathbf{v} = \mathbf{0}$, except on the axis $\rho = 0$, there exists a scalar potential ψ such that $\mathbf{v} = \nabla\psi$. It may easily be shown that $\psi = \phi$, the polar angle. Therefore, if C does not enclose the axis then

$$\oint_C \mathbf{v} \cdot d\mathbf{r} = \oint d\phi = 0,$$

and if C does enclose the axis,

$$\oint_C \mathbf{v} \cdot d\mathbf{r} = \Delta\phi = 2\pi n,$$

where n is the number of times we traverse C. Thus ϕ is a multivalued potential.

Similar analyses are valid for other physical systems – for example, in magnetostatics we may replace the vortex lines by current-carrying wires and the velocity field $\mathbf{v}$ by the magnetic field $\mathbf{B}$.

11.10 Exercises

11.1 The vector field $\mathbf{F}$ is defined by

$$\mathbf{F} = 2xz\mathbf{i} + 2yz^2\mathbf{j} + (x^2 + 2y^2z - 1)\mathbf{k}.$$

Calculate $\nabla \times \mathbf{F}$ and deduce that $\mathbf{F}$ can be written $F = \nabla\phi$. Determine the form of ϕ.

11.2 The vector field $\mathbf{Q}$ is defined by

$$\mathbf{Q} = \left[3x^2(y+z) + y^3 + z^3\right]\mathbf{i} + \left[3y^2(z+x) + z^3 + x^3\right]\mathbf{j} + \left[3z^2(x+y) + x^3 + y^3\right]\mathbf{k}.$$

Show that $\mathbf{Q}$ is a conservative field, construct its potential function and hence evaluate the integral $J = \int \mathbf{Q} \cdot d\mathbf{r}$ along any line connecting the point A at $(1, -1, 1)$ to B at $(2, 1, 2)$.

11.3 $\mathbf{F}$ is a vector field $xy^2\mathbf{i} + 2\mathbf{j} + x\mathbf{k}$, and L is a path parameterised by $x = ct$, $y = c/t$, $z = d$ for the range $1 \le t \le 2$. Evaluate (a) $\int_L \mathbf{F}\,dt$, (b) $\int_L \mathbf{F}\,dy$ and (c) $\int_L \mathbf{F} \cdot d\mathbf{r}$.

11.4 By making an appropriate choice for the functions $P(x, y)$ and $Q(x, y)$ that appear in Green's theorem in a plane, show that the integral of $x - y$ over the upper half of the unit circle centred on the origin has the value $-\frac{2}{3}$. Show the same result by direct integration in Cartesian coordinates.

11.5 Determine the point of intersection P, in the first quadrant, of the two ellipses

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \quad \text{and} \quad \frac{x^2}{b^2} + \frac{y^2}{a^2} = 1.$$

Taking $b < a$, consider the contour L that bounds the area in the first quadrant that is common to the two ellipses. Show that the parts of L that lie along the coordinate axes contribute nothing to the line integral around L of $x\,dy - y\,dx$. Using a parameterisation of each ellipse similar to that employed in the example

' $ - @)- C ,
< ! ! -A -3 L H ! H
-, , -, !, , ,- -

29

Representation theory

As indicated at the start of the previous chapter, significant conclusions can often be drawn about a physical system simply from the study of its symmetry properties. That chapter was devoted to setting up a formal mathematical basis, group theory, with which to describe and classify such properties; the current chapter shows how to implement the consequences of the resulting classifications and obtain concrete physical conclusions about the system under study. The connection between the two chapters is akin to that between working with coordinate-free vectors, each denoted by a single symbol, and working with a coordinate system in which the same vectors are expressed in terms of components.

The 'coordinate systems' that we will choose will be ones that are expressed in terms of matrices; it will be clear that ordinary numbers would not be sufficient, as they make no provision for any non-commutation amongst the elements of a group. Thus, in this chapter the group elements will be *represented* by matrices that have the same commutation relations as the members of the group, whatever the group's original nature (symmetry operations, functional forms, matrices, permutations, etc.). For some abstract groups it is difficult to give a written description of the elements and their properties without recourse to such representations. Most of our applications will be concerned with representations of the groups that consist of the symmetry operations on molecules containing two or more identical atoms.

Firstly, in section 29.1, we use an elementary example to demonstrate the kind of conclusions that can be reached by arguing purely on symmetry grounds. Then in sections 29.2–29.10 we develop the formal side of representation theory and establish general procedures and results. Finally, these are used in section 29.11 to tackle a variety of problems drawn from across the physical sciences.

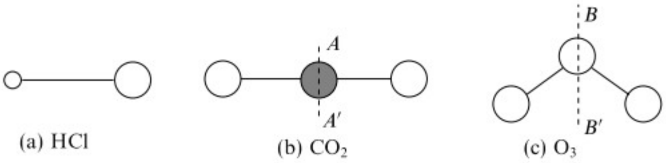

Figure 29.1 Three molecules, (a) hydrogen chloride, (b) carbon dioxide and (c) ozone, for which symmetry considerations impose varying degrees of constraint on their possible electric dipole moments.

29.1 Dipole moments of molecules

Some simple consequences of *symmetry* can be demonstrated by considering whether a permanent electric dipole moment can exist in any particular molecule; three simple molecules, hydrogen chloride, carbon dioxide and ozone, are illustrated in figure 29.1. Even if a molecule is electrically neutral, an electric dipole moment will exist in it if the centres of gravity of the positive charges (due to protons in the atomic nuclei) and of the negative charges (due to the electrons) do not coincide.

For hydrogen chloride there is no reason why they should coincide; indeed, the normal picture of the binding mechanism in this molecule is that the electron from the hydrogen atom moves its average position from that of its proton nucleus to somewhere between the hydrogen and chlorine nuclei. There is no compensating movement of positive charge, and a net dipole moment is to be expected – and is found experimentally.

For the linear molecule carbon dioxide it seems obvious that it cannot have a dipole moment, because of its symmetry. Putting this rather more rigorously, we note that any rotation about the long axis of the molecule leaves it totally unchanged; consequently, any component of a permanent electric dipole perpendicular to that axis must be zero (a non-zero component would rotate although no physical change had taken place in the molecule). That only leaves the possibility of a component parallel to the axis. However, a rotation of π radians about the axis AA' shown in figure 29.1(b) carries the molecule into itself, as does a reflection in a plane through the carbon atom and perpendicular to the molecular axis (i.e. one with its normal parallel to the axis). In both cases the two oxygen atoms change places but, as they are identical, the molecule is indistinguishable from the original. Either 'symmetry operation' would reverse the sign of any dipole component directed parallel to the molecular axis; this can only be compatible with the indistinguishability of the original and final systems if the parallel component is zero. Thus on symmetry grounds carbon dioxide cannot have a permanent electric dipole moment.

Finally, for ozone, which is angular rather than linear, symmetry does not place such tight constraints. A dipole-moment component parallel to the axis BB' (figure 29.1(c)) is possible, since there is no symmetry operation that reverses the component in that direction and at the same time carries the molecule into an indistinguishable copy of itself. However, a dipole moment perpendicular to BB' is not possible, since a rotation of π about BB' would both reverse any such component and carry the ozone molecule into itself – two contradictory conclusions unless the component is zero.

In summary, symmetry requirements appear in the form that some or all components of permanent electric dipoles in molecules are forbidden; they do not show that the other components do exist, only that they may. The greater the symmetry of the molecule, the tighter the restrictions on potentially non-zero components of its dipole moment.

In section 23.11 other, more complicated, physical situations will be analysed using results derived from representation theory. In anticipation of these results, and since it may help the reader to understand where the developments in the next nine sections are leading, we make here a broad, powerful, but rather formal, statement as follows.

If a physical system is such that after the application of particular rotations or reflections (or a combination of the two) the final system is indistinguishable from the original system then its behaviour, and hence the functions that describe its behaviour, must have the corresponding property of invariance when subjected to the same rotations and reflections.

29.2 Choosing an appropriate formalism

As mentioned in the introduction to this chapter, the elements of a finite group $\mathcal{G}$ can be *represented* by matrices; this is done in the following way. A suitable column matrix u, known as a *basis vector*,[§] is chosen and is written in terms of its components u_i, the *basis functions*, as $\mathsf{u} = (u_1\ u_2\ \cdots\ u_n)^{\mathrm{T}}$. The u_i may be of a variety of natures, e.g. numbers, coordinates, functions or even a set of labels, though for any one basis vector they will all be of the same kind.

Once chosen, the basis vector can be used to generate an n-dimensional *representation* of the group as follows. An element X of the group is selected and its effect on each basis function u_i is determined. If the action of X on u_1 is to produce u_1', etc. then the set of equations

$$u_i' = Xu_i \tag{29.1}$$

[§] This usage of the term *basis vector* is not exactly the same as that introduced in subsection 8.1.1.

of simple pair interchanges to which a permutation is equivalent. That these assignments are in accord with the group multiplication table 28.8 should be checked.

Thus the three irreps of the group $\mathcal{G}$ (i.e. the group $3m$ or C_{3v} or S_3), are, using the conventional notation A_1, A_2, E (see section 29.8), as follows:

		Element					
		I	A	B	C	D	E
Irrep	A_1	1	1	1	1	1	1
	A_2	1	1	1	-1	-1	-1
	E	M_I	M_A	M_B	M_C	M_D	M_E

(29.12)

where

$$\mathsf{M}_I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathsf{M}_A = \begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}, \qquad \mathsf{M}_B = \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix},$$

$$\mathsf{M}_C = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathsf{M}_D = \begin{pmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}, \quad \mathsf{M}_E = \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}.$$

29.5 The orthogonality theorem for irreducible representations

We come now to the central theorem of representation theory, a theorem that justifies the relatively routine application of certain procedures to determine the restrictions that are inherent in physical systems that have some degree of rotational or reflection symmetry. The development of the theorem is long and quite complex when presented in its entirety, and the reader will have to refer elsewhere for the proof.[§]

The theorem states that, in a certain sense, the irreps of a group $\mathcal{G}$ are as orthogonal as possible, as follows. If, for each irrep, the elements in any one position in each of the g matrices are used to make up g-component column matrices then

(i) any two such column matrices coming from different irreps are orthogonal;
(ii) any two such column matrices coming from different positions in the matrices of the same irrep are orthogonal.

This orthogonality is in addition to the irreps' being in the form of orthogonal (unitary) matrices and thus each comprising mutually orthogonal rows and columns.

[§] See, e.g., H. F. Jones, *Groups, Representations and Physics* (Bristol: Institute of Physics, 1998); J. F. Cornwell, *Group Theory in Physics*, vol 2 (London: Academic Press, 1984); J-P. Serre, *Linear Representations of Finite Groups* (New York: Springer, 1977).

More mathematically, if we denote the entry in the ith row and jth column of a matrix $\mathsf{D}(X)$ by $[\mathsf{D}(X)]_{ij}$, and $\hat{\mathsf{D}}^{(\lambda)}$ and $\hat{\mathsf{D}}^{(\mu)}$ are two irreps of $\mathcal{G}$ having dimensions n_λ and n_μ respectively, then

$$\sum_X \left[\hat{\mathsf{D}}^{(\lambda)}(X)\right]^*_{ij} \left[\hat{\mathsf{D}}^{(\mu)}(X)\right]_{kl} = \frac{g}{n_\lambda}\delta_{ik}\delta_{jl}\delta_{\lambda\mu}. \tag{29.13}$$

This rather forbidding-looking equation needs some further explanation.

Firstly, the asterisk indicates that the complex conjugate should be taken if necessary, though all our representations so far have involved only real matrix elements. Each Kronecker delta function on the right-hand side has the value 1 if its two subscripts are equal and has the value 0 otherwise. Thus the right-hand side is only non-zero if $i = k$, $j = l$ and $\lambda = \mu$, all at the same time.

Secondly, the summation over the group elements X means that g contributions have to be added together, each contribution being a product of entries drawn from the representative matrices in the two irreps $\hat{\mathsf{D}}^{(\lambda)} = \{\hat{\mathsf{D}}^{(\lambda)}(X)\}$ and $\hat{\mathsf{D}}^{(\mu)} = \{\hat{\mathsf{D}}^{(\mu)}(X)\}$. The g contributions arise as X runs over the g elements of $\mathcal{G}$.

Thus, putting these remarks together, the summation will produce zero if either

(i) the matrix elements are not taken from exactly the same position in every matrix, including cases in which it is not possible to do so because the irreps $\hat{\mathsf{D}}^{(\lambda)}$ and $\hat{\mathsf{D}}^{(\mu)}$ have different dimensions, or

(ii) even if $\hat{\mathsf{D}}^{(\lambda)}$ and $\hat{\mathsf{D}}^{(\mu)}$ do have the same dimensions and the matrix elements are from the same positions in every matrix, they are different irreps, i.e. $\lambda \neq \mu$.

Some numerical illustrations based on the irreps A_1, A_2 and E of the group $3m$ (or C_{3v} or S_3) will probably provide the clearest explanation (see (29.12)).

(a) Take $i = j = k = l = 1$, with $\hat{\mathsf{D}}^{(\lambda)} = A_1$ and $\hat{\mathsf{D}}^{(\mu)} = A_2$. Equation (29.13) then reads

$$1(1) + 1(1) + 1(1) + 1(-1) + 1(-1) + 1(-1) = 0,$$

as expected, since $\lambda \neq \mu$.

(b) Take (i, j) as $(1, 2)$ and (k, l) as $(2, 2)$, corresponding to different matrix positions within the same irrep $\hat{\mathsf{D}}^{(\lambda)} = \hat{\mathsf{D}}^{(\mu)} = $ E. Substituting in (29.13) gives

$$0(1) + \left(-\tfrac{\sqrt{3}}{2}\right)\left(-\tfrac{1}{2}\right) + \left(\tfrac{\sqrt{3}}{2}\right)\left(-\tfrac{1}{2}\right) + 0(1) + \left(-\tfrac{\sqrt{3}}{2}\right)\left(-\tfrac{1}{2}\right) + \left(\tfrac{\sqrt{3}}{2}\right)\left(-\tfrac{1}{2}\right) = 0.$$

(c) Take (i, j) as $(1, 2)$, and (k, l) as $(1, 2)$, corresponding to the same matrix positions within the same irrep $\hat{\mathsf{D}}^{(\lambda)} = \hat{\mathsf{D}}^{(\mu)} = $ E. Substituting in (29.13) gives

$$0(0) + \left(-\tfrac{\sqrt{3}}{2}\right)\left(-\tfrac{\sqrt{3}}{2}\right) + \left(\tfrac{\sqrt{3}}{2}\right)\left(\tfrac{\sqrt{3}}{2}\right) + 0(0) + \left(-\tfrac{\sqrt{3}}{2}\right)\left(-\tfrac{\sqrt{3}}{2}\right) + \left(\tfrac{\sqrt{3}}{2}\right)\left(\tfrac{\sqrt{3}}{2}\right) = \tfrac{6}{2}.$$

(d) No explicit calculation is needed to see that if $i = j = k = l = 1$, with $\hat{\mathsf{D}}^{(\lambda)} = \hat{\mathsf{D}}^{(\mu)} = \mathrm{A}_1$ (or A_2), then each term in the sum is either 1^2 or $(-1)^2$ and the total is 6, as predicted by the right-hand side of (29.13) since $g = 6$ and $n_\lambda = 1$.

29.6 Characters

The actual matrices of general representations and irreps are cumbersome to work with, and they are not unique since there is always the freedom to change the coordinate system, i.e. the components of the basis vector (see section 29.3), and hence the entries in the matrices. However, one thing that does not change for a matrix under such an equivalence (similarity) transformation – i.e. under a change of basis – is the trace of the matrix. This was shown in chapter 8, but is repeated here. The trace of a matrix A is the sum of its diagonal elements,

$$\mathrm{Tr}\,\mathsf{A} = \sum_{i=1}^{n} \mathsf{A}_{ii}$$

or, using the summation convention (section 26.1), simply A_{ii}. Under a similarity transformation, again using the summation convention,

$$\begin{aligned}
[\mathsf{D}_{\mathsf{Q}}(X)]_{ii} &= [\mathsf{Q}^{-1}]_{ij}[\mathsf{D}(X)]_{jk}[\mathsf{Q}]_{ki} \\
&= [\mathsf{D}(X)]_{jk}[\mathsf{Q}]_{ki}[\mathsf{Q}^{-1}]_{ij} \\
&= [\mathsf{D}(X)]_{jk}[\mathsf{I}]_{kj} \\
&= [\mathsf{D}(X)]_{jj},
\end{aligned}$$

showing that the traces of equivalent matrices are equal.

This fact can be used to greatly simplify work with representations, though with some partial loss of the information content of the full matrices. For example, using trace values alone it is not possible to distinguish between the two groups known as $4mm$ and $\bar{4}2m$, or as C_{4v} and D_{2d} respectively, even though the two groups are not isomorphic. To make use of these simplifications we now define the characters of a representation.

Definition. *The* characters $\chi(\mathsf{D})$ *of a representation* D *of a group* $\mathcal{G}$ *are defined as the traces of the matrices* $\mathsf{D}(X)$*, one for each element* X *of* $\mathcal{G}$.

At this stage there will be g characters, but, as we noted in subsection 28.7.3, elements A, B of $\mathcal{G}$ in the same conjugacy class are connected by equations of the form $B = X^{-1}AX$. It follows that their matrix representations are connected by corresponding equations of the form $\mathsf{D}(B) = \mathsf{D}(X^{-1})\mathsf{D}(A)\mathsf{D}(X)$, and so by the argument just given their representations will have equal traces and hence equal characters. Thus *elements in the same conjugacy class have the same characters,*

$3m$	I	A, B	C, D, E	
A_1	1	1	1	z; z^2; $x^2 + y^2$
A_2	1	1	-1	R_z
E	2	-1	0	(x, y); (xz, yz); (R_x, R_y); $(x^2 - y^2, 2xy)$

Table 29.1 The character table for the irreps of group $3m$ (C_{3v} or S_3). The right-hand column lists some common functions that transform according to the irrep against which each is shown (see text).

though, in general, these will vary from one representation to another. However, it might also happen that two or more conjugacy classes have the same characters in a representation – indeed, in the trivial irrep A_1, see (29.12), every element inevitably has the character 1.

For the irrep A_2 of the group $3m$, the classes $\{I\}$, $\{A, B\}$ and $\{C, D, E\}$ have characters 1, 1 and -1, respectively, whilst they have characters 2, -1 and 0 respectively in irrep E.

We are thus able to draw up a *character table* for the group $3m$ as shown in table 29.1. This table holds in compact form most of the important information on the behaviour of functions under the two-dimensional rotational and reflection symmetries of an equilateral triangle, i.e. under the elements of group $3m$. The entry under I for any irrep gives the dimension of the irrep, since it is equal to the trace of the unit matrix whose dimension is equal to that of the irrep. In other words, for the λth irrep $\chi^{(\lambda)}(I) = n_\lambda$, where n_λ is its dimension.

In the extreme right-hand column we list some common functions of Cartesian coordinates that transform, under the group $3m$, according to the irrep on whose line they are listed. Thus, as we have seen, z, z^2, and $x^2 + y^2$ are all unchanged by the group operations (though x and y individually are affected) and so are listed against the one-dimensional irrep A_1. Each of the pairs (x, y), (xz, yz), and $(x^2 - y^2, 2xy)$, however, is mixed as a pair by some of the operations, and so these pairs are listed against the two-dimensional irrep E: each pair forms a basis set for this irrep.

The quantities R_x, R_y and R_z refer to rotations about the indicated axes; they transform in the same way as the corresponding components of angular momentum **J**, and their behaviour can be established by examining how the components of $\mathbf{J} = \mathbf{r} \times \mathbf{p}$ transform under the operations of the group. To do this explicitly is beyond the scope of this book. However, it can be noted that R_z, being listed opposite the one-dimensional A_2, is unchanged by I and by the rotations A and B but changes sign under the mirror reflections C, D, and E, as would be expected.

29.6.1 Orthogonality property of characters

Some of the most important properties of characters can be deduced from the orthogonality theorem (29.13),

$$\sum_X \left[\hat{\mathsf{D}}^{(\lambda)}(X)\right]^*_{ij} \left[\hat{\mathsf{D}}^{(\mu)}(X)\right]_{kl} = \frac{g}{n_\lambda}\delta_{ik}\delta_{jl}\delta_{\lambda\mu}.$$

If we set $j = i$ and $l = k$, so that both factors in any particular term in the summation refer to diagonal elements of the representative matrices, and then sum both sides over i and k, we obtain

$$\sum_X \sum_{i=1}^{n_\lambda} \sum_{k=1}^{n_\mu} \left[\hat{\mathsf{D}}^{(\lambda)}(X)\right]^*_{ii} \left[\hat{\mathsf{D}}^{(\mu)}(X)\right]_{kk} = \frac{g}{n_\lambda} \sum_{i=1}^{n_\lambda} \sum_{k=1}^{n_\mu} \delta_{ik}\delta_{ik}\delta_{\lambda\mu}.$$

Expressed in term of characters, this reads

$$\sum_X \left[\chi^{(\lambda)}(X)\right]^* \chi^{(\mu)}(X) = \frac{g}{n_\lambda} \sum_{i=1}^{n_\lambda} \delta_{ii}^2 \delta_{\lambda\mu} = \frac{g}{n_\lambda} \sum_{i=1}^{n_\lambda} 1 \times \delta_{\lambda\mu} = g\delta_{\lambda\mu}. \tag{29.14}$$

In words, the (g-component) 'vectors' formed from the characters of the various irreps of a group are mutually orthogonal, but each one has a squared magnitude (the sum of the squares of its components) equal to the order of the group.

Since, as noted in the previous subsection, group elements in the same class have the same characters, (29.14) can be written as a sum over classes rather than elements. If c_i denotes the number of elements in class $\mathcal{C}_i$ and X_i any element of $\mathcal{C}_i$, then

$$\sum_i c_i \left[\chi^{(\lambda)}(X_i)\right]^* \chi^{(\mu)}(X_i) = g\delta_{\lambda\mu}. \tag{29.15}$$

Although we do not prove it here, there also exists a 'completeness' relation for characters. It makes a statement about the products of characters for a fixed pair of group elements, X_1 and X_2, when the products are summed over all possible irreps of the group. This is the converse of the summation process defined by (29.14). The completeness relation states that

$$\sum_\lambda \left[\chi^{(\lambda)}(X_1)\right]^* \chi^{(\lambda)}(X_2) = \frac{g}{c_1}\delta_{\mathcal{C}_1\mathcal{C}_2}, \tag{29.16}$$

where element X_1 belongs to conjugacy class $\mathcal{C}_1$ and X_2 belongs to $\mathcal{C}_2$. Thus the sum is zero unless X_1 and X_2 belong to the same class. For table 29.1 we can verify that these results are valid.

(i) For $\hat{\mathsf{D}}^{(\lambda)} = \hat{\mathsf{D}}^{(\mu)} = \mathrm{A}_1$ or A_2, (29.15) reads

$$1(1) + 2(1) + 3(1) = 6,$$

whilst for $\hat{\mathsf{D}}^{(\lambda)} = \hat{\mathsf{D}}^{(\mu)} = \mathrm{E}$, it gives

$$1(2^2) + 2(1) + 3(0) = 6.$$

(ii) For $\hat{\mathsf{D}}^{(\lambda)} = \mathrm{A}_2$ and $\hat{\mathsf{D}}^{(\mu)} = \mathrm{E}$, say, (29.15) reads

$$1(1)(2) + 2(1)(-1) + 3(-1)(0) = 0.$$

(iii) For $X_1 = A$ and $X_2 = D$, say, (29.16) reads

$$1(1) + 1(-1) + (-1)(0) = 0,$$

whilst for $X_1 = C$ and $X_2 = E$, both of which belong to class $\mathcal{C}_3$ for which $c_3 = 3$,

$$1(1) + (-1)(-1) + (0)(0) = 2 = \frac{6}{3}.$$

29.7 Counting irreps using characters

The expression of a general representation $\mathsf{D} = \{\mathsf{D}(X)\}$ in terms of irreps, as given in (29.11), can be simplified by going from the full matrix form to that of characters. Thus

$$\mathsf{D}(X) = m_1\hat{\mathsf{D}}^{(1)}(X) \oplus m_2\hat{\mathsf{D}}^{(2)}(X) \oplus \cdots \oplus m_N\hat{\mathsf{D}}^{(N)}(X)$$

becomes, on taking the trace of both sides,

$$\chi(X) = \sum_{\lambda=1}^{N} m_\lambda \chi^{(\lambda)}(X). \qquad (29.17)$$

Given the characters of the irreps of the group $\mathcal{G}$ to which the elements X belong, and the characters of the representation $\mathsf{D} = \{\mathsf{D}(X)\}$, the g equations (29.17) can be solved as simultaneous equations in the m_λ, either by inspection or by multiplying both sides by $\left[\chi^{(\mu)}(X)\right]^*$ and summing over X, making use of (29.14) and (29.15), to obtain

$$m_\mu = \frac{1}{g}\sum_X \left[\chi^{(\mu)}(X)\right]^* \chi(X) = \frac{1}{g}\sum_i c_i \left[\chi^{(\mu)}(X_i)\right]^* \chi(X_i). \qquad (29.18)$$

That an unambiguous formula can be given for each m_λ, once the *character set* (the set of characters of each of the group elements or, equivalently, of each of the conjugacy classes) of D is known, shows that, for any particular group, two representations with the same characters are equivalent. This strongly suggests something that can be shown, namely, *the number of irreps = the number of conjugacy classes*. The argument is as follows. Equation (29.17) is a set of simultaneous equations for N unknowns, the m_λ, some of which may be zero. The value of N is equal to the number of irreps of $\mathcal{G}$. There are g different values of X, but the number of *different* equations is only equal to the number of distinct conjugacy classes, since any two elements of $\mathcal{G}$ in the same class have the same character set and therefore generate the same equation. For a unique solution to simultaneous equations in N unknowns, exactly N independent equations are needed. Thus N is also the number of classes, establishing the stated result.

►*Determine the irreps contained in the representation of the group 3m in the vector space spanned by the functions x^2, y^2, xy.*

We first note that although these functions are not orthogonal they form a basis set for a representation, since they are linearly independent quadratic forms in x and y and any other quadratic form can be written (uniquely) in terms of them. We must establish how they transform under the symmetry operations of group $3m$. We need to do so only for a representative element of each conjugacy class, and naturally we take the simplest in each case.

The first class contains only I (as always) and clearly $\mathsf{D}(I)$ is the 3×3 unit matrix.

The second class contains the rotations, A and B, and we choose to find $\mathsf{D}(A)$. Since, under A,

$$x \;\to\; -\frac{1}{2}x + \frac{\sqrt{3}}{2}y \qquad \text{and} \qquad y \;\to\; -\frac{\sqrt{3}}{2}x - \frac{1}{2}y,$$

it follows that

$$x^2 \;\to\; \tfrac{1}{4}x^2 - \tfrac{\sqrt{3}}{2}xy + \tfrac{3}{4}y^2, \qquad\qquad y^2 \;\to\; \tfrac{3}{4}x^2 + \tfrac{\sqrt{3}}{2}xy + \tfrac{1}{4}y^2 \tag{29.19}$$

and

$$xy \;\to\; \tfrac{\sqrt{3}}{4}x^2 - \tfrac{1}{2}xy - \tfrac{\sqrt{3}}{4}y^2. \tag{29.20}$$

Hence $\mathsf{D}(A)$ can be deduced and is given below.

The third and final class contains the reflections, C, D and E; of these C is much the easiest to deal with. Under C, $x \to -x$ and $y \to y$, causing xy to change sign but leaving x^2 and y^2 unaltered. The three matrices needed are thus

$$\mathsf{D}(I) = \mathsf{I}_3, \quad \mathsf{D}(C) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \quad \mathsf{D}(A) = \begin{pmatrix} \frac{1}{4} & \frac{3}{4} & -\frac{\sqrt{3}}{2} \\ \frac{3}{4} & \frac{1}{4} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{4} & -\frac{\sqrt{3}}{4} & -\frac{1}{2} \end{pmatrix};$$

their traces are respectively 3, 1 and 0.

It should be noticed that much more work has been done here than is necessary, since the traces can be computed immediately from the effects of the symmetry operations on the basis functions. All that is needed is the weight of each basis function in the transformed expression for that function; these are clearly 1, 1, 1 for I, and $\frac{1}{4}$, $\frac{1}{4}$, $-\frac{1}{2}$ for A, from (29.19) and (29.20), and 1, 1, -1 for C, from the observations made just above the displayed matrices. The traces are then the sums of these weights. The off-diagonal elements of the matrices need not be found, nor need the matrices be written out.

From (29.17) we now need to find a superposition of the characters of the irreps that gives representation D in the bottom line of table 29.2.

By inspection it is obvious that $\mathsf{D} = \mathrm{A}_1 \oplus \mathrm{E}$, but we can use (29.18) formally:

$$\begin{aligned} m_{A_1} &= \tfrac{1}{6}[1(1)(3) + 2(1)(0) + 3(1)(1)] = 1, \\ m_{A_2} &= \tfrac{1}{6}[1(1)(3) + 2(1)(0) + 3(-1)(1)] = 0, \\ m_E &= \tfrac{1}{6}[1(2)(3) + 2(-1)(0) + 3(0)(1)] = 1. \end{aligned}$$

Thus A_1 and E appear once each in the reduction of D, and A_2 not at all. Table 29.1 gives the further information, not needed here, that it is the combination $x^2 + y^2$ that transforms as a one-dimensional irrep and the pair $(x^2 - y^2,\ 2xy)$ that forms a basis of the two-dimensional irrep, E. ◄

	Classes		
Irrep	I	AB	CDE
A_1	1	1	1
A_2	1	1	-1
E	2	-1	0
D	3	0	1

Table 29.2 The characters of the irreps of the group $3m$ and of the representation D, which must be a superposition of some of them.

29.7.1 *Summation rules for irreps*

The first summation rule for irreps is a simple restatement of (29.14), with μ set equal to λ; it then reads

$$\sum_X \left[\chi^{(\lambda)}(X)\right]^* \chi^{(\lambda)}(X) = g.$$

In words, the sum of the squares (modulus squared if necessary) of the characters of an irrep taken over all elements of the group adds up to the order of the group. For group $3m$ (table 29.1), this takes the following explicit forms:

for A_1, $\quad 1(1^2) + 2(1^2) + 3(1^2) = 6;$
for A_2, $\quad 1(1^2) + 2(1^2) + 3(-1)^2 = 6;$
for E, $\quad 1(2^2) + 2(-1)^2 + 3(0^2) = 6.$

We next prove a theorem that is concerned not with a summation within an irrep but with a summation over irreps.

Theorem. *If n_μ is the dimension of the μth irrep of a group $\mathcal{G}$ then*

$$\sum_\mu n_\mu^2 = g,$$

where g is the order of the group.

Proof. Define a representation of the group in the following way. Rearrange the rows of the multiplication table of the group so that whilst the elements in a particular order head the columns, their inverses in the same order head the rows. In this arrangement of the $g \times g$ table, the leading diagonal is entirely occupied by the identity element. Then, for each element X of the group, take as representative matrix the multiplication-table array obtained by replacing X by 1 and all other element symbols by 0. The matrices $\mathsf{D}^{\text{reg}}(X)$ so obtained form the *regular representation* of $\mathcal{G}$; they are each $g \times g$, have a single non-zero entry '1' in each row and column and (as will be verified by a little experimentation) have

(a)

	I	A	B
I	I	A	B
A	A	B	I
B	B	I	A

(b)

	I	A	B
I	I	A	B
B	B	I	A
A	A	B	I

Table 29.3 (a) The multiplication table of the cyclic group of order 3, and (b) its reordering used to generate the regular representation of the group.

the same multiplication structure as the group $\mathcal{G}$ itself, i.e. they form a faithful representation of $\mathcal{G}$.

Although not part of the proof, a simple example may help to make these ideas more transparent. Consider the cyclic group of order 3. Its multiplication table is shown in table 29.3(a) (a repeat of table 28.10(a) of the previous chapter), whilst table 29.3(b) shows the same table reordered so that the columns are still labelled in the order I, A, B but the rows are now labelled in the order $I^{-1} = I$, $A^{-1} = B$, $B^{-1} = A$. The three matrices of the regular representation are then

$$\mathsf{D}^{\rm reg}(I) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathsf{D}^{\rm reg}(A) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \quad \mathsf{D}^{\rm reg}(B) = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

An alternative, more mathematical, definition of the regular representation of a group is

$$\left[\mathsf{D}^{\rm reg}(G_k)\right]_{ij} = \begin{cases} 1 & \text{if } G_k G_j = G_i, \\ 0 & \text{otherwise.} \end{cases}$$

We now return to the proof. With the construction given, the regular representation has characters as follows:

$$\chi^{\rm reg}(I) = g, \qquad \chi^{\rm reg}(X) = 0 \quad \text{if } X \neq I.$$

We now apply (29.18) to $\mathsf{D}^{\rm reg}$ to obtain for the number m_μ of times that the irrep $\hat{\mathsf{D}}^{(\mu)}$ appears in $\mathsf{D}^{\rm reg}$ (see 29.11))

$$m_\mu = \frac{1}{g} \sum_X \left[\chi^{(\mu)}(X)\right]^* \chi^{\rm reg}(X) = \frac{1}{g} \left[\chi^{(\mu)}(I)\right]^* \chi^{\rm reg}(I) = \frac{1}{g} n_\mu g = n_\mu.$$

Thus an irrep $\hat{\mathsf{D}}^{(\mu)}$ of dimension n_μ appears n_μ times in $\mathsf{D}^{\rm reg}$, and so by counting the total number of basis functions, or by considering $\chi^{\rm reg}(I)$, we can conclude that

$$\sum_{\mu} n_{\mu}^2 = g. \tag{29.21}$$

This completes the proof.

As before, our standard demonstration group $3m$ provides an illustration. In this case we have seen already that there are two one-dimensional irreps and one two-dimensional irrep. This is in accord with (29.21) since

$$1^2 + 1^2 + 2^2 = 6, \qquad \text{which is the order } g \text{ of the group.}$$

Another straightforward application of the relation (29.21), to the group with multiplication table 29.3(a), yields immediate results. Since $g = 3$, none of its irreps can have dimension 2 or more, as $2^2 = 4$ is too large for (29.21) to be satisfied. Thus all irreps must be one-dimensional and there must be three of them (consistent with the fact that each element is in a class of its own, and that there are therefore three classes). The three irreps are the sets of 1×1 matrices (numbers)

$$\mathrm{A}_1 = \{1, 1, 1\} \qquad \mathrm{A}_2 = \{1, \omega, \omega^2\} \qquad \mathrm{A}_2^* = \{1, \omega^2, \omega\},$$

where $\omega = \exp(2\pi i/3)$; since the matrices are 1×1, the same set of nine numbers would be, of course, the entries in the character table for the irreps of the group. The fact that the numbers in each irrep are all cube roots of unity is discussed below. As will be noticed, two of these irreps are complex – an unusual occurrence in most applications – and form a complex conjugate pair of one-dimensional irreps. In practice, they function much as a two-dimensional irrep, but this is to be ignored for formal purposes such as theorems.

A further property of characters can be derived from the fact that all elements in a conjugacy class have the same order. Suppose that the element X has order m, i.e. $X^m = I$. This implies for a representation D of dimension n that

$$[\mathsf{D}(X)]^m = I_n. \tag{29.22}$$

Representations equivalent to D are generated as before by using similarity transformations of the form

$$\mathsf{D}_{\mathsf{Q}}(X) = \mathsf{Q}^{-1}\mathsf{D}(X)\mathsf{Q}.$$

In particular, if we choose the columns of Q to be the eigenvectors of $\mathsf{D}(X)$ then, as discussed in chapter 8,

$$\mathsf{D}_{\mathsf{Q}}(X) = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix}$$

where the λ_i are the eigenvalues of $\mathsf{D}(X)$. Therefore, from (29.22), we have that

$$\begin{pmatrix} \lambda_1^m & 0 & \cdots & 0 \\ 0 & \lambda_2^m & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n^m \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Hence all the eigenvalues λ_i are mth roots of unity, and so $\chi(X)$, the trace of $\mathsf{D}(X)$, is the sum of n of these. In view of the implications of Lagrange's theorem (section 28.6 and subsection 28.7.2), the only values of m allowed are the divisors of the order g of the group.

29.8 Construction of a character table

In order to decompose representations into irreps on a routine basis using characters, it is necessary to have available a character table for the group in question. Such a table gives, for each irrep μ of the group, the character $\chi^{(\mu)}(X)$ of the class to which group element X belongs. To construct such a table the following properties of a group, established earlier in this chapter, may be used:

(i) the number of classes equals the number of irreps;
(ii) the 'vector' formed by the characters from a given irrep is orthogonal to the 'vector' formed by the characters from a different irrep;
(iii) $\sum_\mu n_\mu^2 = g$, where n_μ is the dimension of the μth irrep and g is the order of the group;
(iv) the identity irrep (one-dimensional with all characters equal to 1) is present for every group;
(v) $\sum_X \left|\chi^{(\mu)}(X)\right|^2 = g$.
(vi) $\chi^{(\mu)}(X)$ is the sum of n_μ mth roots of unity, where m is the order of X.

►*Construct the character table for the group* 4*mm (or C_{4v}) using the properties of classes, irreps and characters so far established.*

The group 4*mm* is the group of two-dimensional symmetries of a square, namely rotations of 0, $\pi/2$, π and $3\pi/2$ and reflections in the mirror planes parallel to the coordinate axes and along the main diagonals. These are illustrated in figure 29.3. For this group there are eight elements:

- the identity, I;
- rotations by $\pi/2$ and $3\pi/2$, R and R';
- a rotation by π, Q ;
- four mirror reflections m_x, m_y, m_d and $m_{d'}$.

Requirements (i) to (iv) at the start of this section put tight constraints on the possible character sets, as the following argument shows.

The group is non-Abelian (clearly $Rm_x \neq m_xR$), and so there are fewer than eight classes, and hence fewer than eight irreps. But requirement (iii), with $g = 8$, then implies

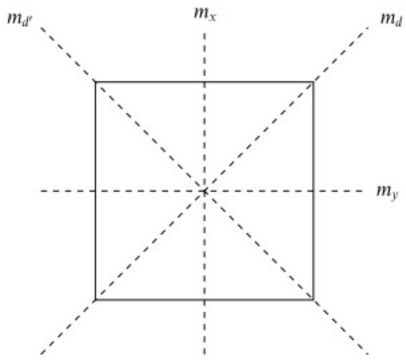

Figure 29.3 The mirror planes associated with $4mm$, the group of two-dimensional symmetries of a square.

that at least one irrep has dimension 2 or greater. However, there can be no irrep with dimension 3 or greater, since $3^2 > 8$, nor can there be more than one two-dimensional irrep, since $2^2 + 2^2 = 8$ would rule out a contribution to the sum in (iii) of 1^2 from the identity irrep, and this must be present. Thus the only possibility is one two-dimensional irrep and, to make the sum in (iii) correct, four one-dimensional irreps.

Therefore using (i) we can now deduce that there are five classes. This same conclusion can be reached by evaluating $X^{-1}YX$ for every pair of elements in $\mathcal{G}$, as in the description of conjugacy classes given in the previous chapter. However, it is tedious to do so and certainly much longer than the above. The five classes are I, Q, $\{R, R'\}$, $\{m_x, m_y\}$, $\{m_d, m_{d'}\}$.

It is straightforward to show that only I and Q commute with every element of the group, so they are the only elements in classes of their own. Each other class must have at least 2 members, but, as there are three classes to accommodate $8 - 2 = 6$ elements, there must be exactly 2 in each class. This does not pair up the remaining 6 elements, but does say that the five classes have 1, 1, 2, 2, and 2 elements. Of course, if we had started by dividing the group into classes, we would know the number of elements in each class directly.

We cannot entirely ignore the group structure (though it sometimes happens that the results are independent of the group structure – for example, all non-Abelian groups of order 8 have the same character table!); thus we need to note in the present case that $m_i^2 = I$ for $i = x,\ y,\ d$ or d' and, as can be proved directly, $Rm_i = m_iR'$ for the same four values of label i. We also recall that for any pair of elements X and Y, $\mathsf{D}(XY) = \mathsf{D}(X)\mathsf{D}(Y)$. We may conclude the following for the one-dimensional irreps.

(a) In view of result (vi), $\chi(m_i) = \mathsf{D}(m_i) = \pm 1$.

(b) Since $R^4 = I$, result (vi) requires that $\chi(R)$ is one of $1,\ i,\ -1,\ -i$. But, since $\mathsf{D}(R)\mathsf{D}(m_i) = \mathsf{D}(m_i)\mathsf{D}(R')$, and the $\mathsf{D}(m_i)$ are just numbers, $\mathsf{D}(R) = \mathsf{D}(R')$. Further

$$\mathsf{D}(R)\mathsf{D}(R) = \mathsf{D}(R)\mathsf{D}(R') = \mathsf{D}(RR') = \mathsf{D}(I) = 1,$$

and so $\mathsf{D}(R) = \pm 1 = \mathsf{D}(R')$.

(c) $\mathsf{D}(Q) = \mathsf{D}(RR) = \mathsf{D}(R)\mathsf{D}(R) = 1$.

If we add this to the fact that the characters of the identity irrep A_1 are all unity then we can fill in those entries in character table 29.4 shown in bold.

Suppose now that the three missing entries in a one-dimensional irrep are p, q and r, where each can only be ± 1. Then, allowing for the numbers in each class, orthogonality

8H & - KKK
+ @ / C - ! (3 -
! , 2

$4mm$	I	Q	R, R'	m_x, m_y	$m_d, m_{d'}$
A_1	**1**	**1**	**1**	**1**	**1**
A_2	**1**	**1**	1	-1	-1
B_1	**1**	**1**	-1	1	-1
B_2	**1**	**1**	-1	-1	1
E	**2**	-2	0	0	0

Table 29.4 The character table deduced for the group $4mm$. For an explanation of the entries in bold see the text.

with the characters of A_1 requires that

$$1(1)(1) + 1(1)(1) + 2(1)(p) + 2(1)(q) + 2(1)(r) = 0.$$

The only possibility is that two of p, q, and r equal -1 and the other equals $+1$. This can be achieved in three different ways, corresponding to the need to find three further different one-dimensional irreps. Thus the first four lines of entries in character table 29.4 can be completed. The final line can be completed by requiring it to be orthogonal to the other four. Property (v) has not been used here though it could have replaced part of the argument given. ◀

29.9 Group nomenclature

The nomenclature of published character tables, as we have said before, is erratic and sometimes unfortunate; for example, often E is used to represent, not only a two-dimensional irrep, but also the identity operation, where we have used I. Thus the symbol E might appear in both the column and row headings of a table, though with quite different meanings in the two cases. In this book we use roman capitals to denote irreps.

One-dimensional irreps are regularly denoted by A and B, B being used if a rotation about the principal axis of $2\pi/n$ has character -1. Here n is the highest integer such that a rotation of $2\pi/n$ is a symmetry operation of the system, and the principal axis is the one about which this occurs. For the group of operations on a square, $n = 4$, the axis is the perpendicular to the square and the rotation in question is R. The names for the group, $4mm$ and C_{4v}, derive from the fact that here n is equal to 4. Similarly, for the operations on an equilateral triangle, $n = 3$ and the group names are $3m$ and C_{3v}, but because the rotation by $2\pi/3$ has character $+1$ in all its one-dimensional irreps (see table 29.1), only A appears in the irrep list.

Two-dimensional irreps are denoted by E, as we have already noted, and three-dimensional irreps by T, although in many cases the symbols are modified by primes and other alphabetic labels to denote variations in behaviour from one irrep to another in respect of mirror reflections and parity inversions. In the study of molecules, alternative names based on molecular angular momentum properties are common. It is beyond the scope of this book to list all these variations, or to give a large selection of character tables; our aim is to demonstrate and justify the use of those found in the literature specifically dedicated to crystal physics or molecular chemistry.

Variations in notation are not restricted to the naming of groups and their irreps, but extend to the symbols used to identify a typical element, and hence all members, of a conjugacy class in a group. In physics these are usually of the types n_z, $\bar{n}_z$ or m_x. The first of these denotes a rotation of $2\pi/n$ about the z-axis, and the second the same thing followed by parity inversion (all vectors $\mathbf{r}$ go to $-\mathbf{r}$), whilst the third indicates a mirror reflection in a plane, in this case the plane $x = 0$.

Typical chemistry symbols for classes are NC_n, NC_n^2, NC_n^x, NS_n, σ_v, σ^{xy}. Here the first symbol N, where it appears, shows that there are N elements in the class (a useful feature). The subscript n has the same meaning as in the physics notation, but σ rather than m is used for a mirror reflection, subscripts v, d or h or superscripts xy, xz or yz denoting the various orientations of the relevant mirror planes. Symmetries involving parity inversions are denoted by S; thus S_n is the chemistry analogue of $\bar{n}$. None of what is said in this and the previous paragraph should be taken as definitive, but merely as a warning of common variations in nomenclature and as an initial guide to corresponding entities. Before using any set of group character tables, the reader should ensure that he or she understands the precise notation being employed.

29.10 Product representations

In quantum mechanical investigations we are often faced with the calculation of what are called matrix elements. These normally take the form of integrals over all space of the product of two or more functions whose analytic forms depend on the microscopic properties (usually angular momentum and its components) of the electrons or nuclei involved. For 'bonding' calculations involving 'overlap integrals' there are usually two functions involved, whilst for transition probabilities a third function, giving the spatial variation of the interaction Hamiltonian, also appears under the integral sign.

If the environment of the microscopic system under investigation has some symmetry properties, then sometimes these can be used to establish, without detailed evaluation, that the multiple integral must have zero value. We now express the essential content of these ideas in group theoretical language.

Suppose we are given an integral of the form

$$J = \int \Psi\phi \, d\tau \quad \text{or} \quad J = \int \Psi\xi\phi \, d\tau$$

to be evaluated over all space in a situation in which the physical system is invariant under a particular group $\mathcal{G}$ of symmetry operations. For the integral to be non-zero the integrand must be invariant under each of these operations. In group theoretical language, *the integrand must transform as the identity, the one-dimensional representation* A_1 *of* $\mathcal{G}$; more accurately, some non-vanishing part of the integrand must do so.

An alternative way of saying this is that if under the symmetry operations of $\mathcal{G}$ the integrand transforms according to a representation D and D does not contain A_1 amongst its irreps then the integral J is necessarily zero. It should be noted that the converse is not true; J may be zero even if A_1 is present, since the integral, whilst showing the required invariance, may still have the value zero.

It is evident that we need to establish how to find the irreps that go to make up a representation of a double or triple product when we already know the irreps according to which the factors in the product transform. The method is established by the following theorem.

Theorem. *For each element of a group the character in a product representation is the product of the corresponding characters in the separate representations.*

Proof. Suppose that $\{u_i\}$ and $\{v_j\}$ are two sets of basis functions, that transform under the operations of a group $\mathcal{G}$ according to representations $\mathsf{D}^{(\lambda)}$ and $\mathsf{D}^{(\mu)}$ respectively. Denote by u and v the corresponding basis vectors and let X be an element of the group. Then the functions generated from u_i and v_j by the action of X are calculated as follows, using (29.1) and (29.4):

$$u_i' = Xu_i = \left[\left(\mathsf{D}^{(\lambda)}(X)\right)^{\mathrm{T}} \mathsf{u}\right]_i = \left[\mathsf{D}^{(\lambda)}(X)\right]_{ii} u_i + \sum_{l\neq i} \left[\left(\mathsf{D}^{(\lambda)}(X)\right)^{\mathrm{T}}\right]_{il} u_l,$$

$$v_j' = Xv_j = \left[\left(\mathsf{D}^{(\mu)}(X)\right)^{\mathrm{T}} \mathsf{v}\right]_j = \left[\mathsf{D}^{(\mu)}(X)\right]_{jj} v_j + \sum_{m\neq j} \left[\left(\mathsf{D}^{(\mu)}(X)\right)^{\mathrm{T}}\right]_{jm} v_m.$$

Here $[\mathsf{D}(X)]_{ij}$ is just a single element of the matrix $\mathsf{D}(X)$ and $[\mathsf{D}(X)]_{kk} = [\mathsf{D}^{\mathrm{T}}(X)]_{kk}$ is simply a diagonal element from the matrix – the repeated subscript does not indicate summation. Now, if we take as basis functions for a product representation $\mathsf{D}^{\mathrm{prod}}(X)$ the products $w_k = u_i v_j$ (where the $n_\lambda n_\mu$ various possible pairs of values i, j are labelled by k), we have also that

$$\begin{aligned} w_k' &= Xw_k = Xu_iv_j = (Xu_i)(Xv_j) \\ &= \left[\mathsf{D}^{(\lambda)}(X)\right]_{ii} \left[\mathsf{D}^{(\mu)}(X)\right]_{jj} u_iv_j + \text{terms not involving the product } u_iv_j. \end{aligned}$$

This is to be compared with

$$w_k' = Xw_k = \left[\left(\mathsf{D}^{\mathrm{prod}}(X)\right)^{\mathrm{T}} \mathsf{w}\right]_k = \left[\mathsf{D}^{\mathrm{prod}}(X)\right]_{kk} w_k + \sum_{n\neq k} \left[\left(\mathsf{D}^{\mathrm{prod}}(X)\right)^{\mathrm{T}}\right]_{kn} w_n,$$

where $\mathsf{D}^{\mathrm{prod}}(X)$ is the product representation matrix for element X of the group. The comparison shows that

$$\left[\mathsf{D}^{\mathrm{prod}}(X)\right]_{kk} = \left[\mathsf{D}^{(\lambda)}(X)\right]_{ii} \left[\mathsf{D}^{(\mu)}(X)\right]_{jj}.$$

It follows that

$$\begin{aligned}
\chi^{\text{prod}}(X) &= \sum_{k=1}^{n_\lambda n_\mu} \left[\mathsf{D}^{\text{prod}}(X)\right]_{kk} \\
&= \sum_{i=1}^{n_\lambda} \sum_{j=1}^{n_\mu} \left[\mathsf{D}^{(\lambda)}(X)\right]_{ii} \left[\mathsf{D}^{(\mu)}(X)\right]_{jj} \\
&= \left\{ \sum_{i=1}^{n_\lambda} \left[\mathsf{D}^{(\lambda)}(X)\right]_{ii} \right\} \left\{ \sum_{j=1}^{n_\mu} \left[\mathsf{D}^{(\mu)}(X)\right]_{jj} \right\} \\
&= \chi^{(\lambda)}(X)\ \chi^{(\mu)}(X). \qquad (29.23)
\end{aligned}$$

This proves the theorem, and a similar argument leads to the corresponding result for integrands in the form of a product of three or more factors.

An immediate corollary is that *an integral whose integrand is the product of two functions transforming according to two different irreps is necessarily zero.* To see this, we use (29.18) to determine whether irrep A_1 appears in the product character set $\chi^{\text{prod}}(X)$:

$$m_{A_1} = \frac{1}{g} \sum_X \left[\chi^{(A_1)}(X)\right]^* \chi^{\text{prod}}(X) = \frac{1}{g} \sum_X \chi^{\text{prod}}(X) = \frac{1}{g} \sum_X \chi^{(\lambda)}(X)\chi^{(\mu)}(X).$$

We have used the fact that $\chi^{(A_1)}(X) = 1$ for all X but now note that, by virtue of (29.14), the expression on the right of this equation is equal to zero unless $\lambda = \mu$.

Any complications due to non-real characters have been ignored – in practice, they are handled automatically as it is usually $\Psi^*\phi$, rather than $\Psi\phi$, that appears in integrands, though many functions are real in any case, and nearly all characters are.

Equation (29.23) is a general result for integrands but, specifically in the context of chemical bonding, it implies that for the possibility of bonding to exist, the two quantum wavefunctions must transform according to the same irrep. This is discussed further in the next section.

29.11 Physical applications of group theory

As we indicated at the start of chapter 28 and discussed in a little more detail at the beginning of the present chapter, some physical systems possess symmetries that allow the results of the present chapter to be used in their analysis. We consider now some of the more common sorts of problem in which these results find ready application.

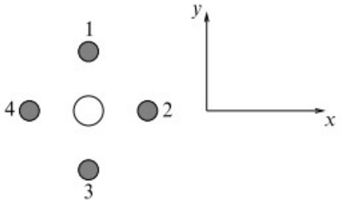

Figure 29.4 A molecule consisting of four atoms of iodine and one of manganese.

29.11.1 Bonding in molecules

We have just seen that whether chemical bonding can take place in a molecule is strongly dependent upon whether the wavefunctions of the two atoms forming a bond transform according to the same irrep. Thus it is sometimes useful to be able to find a wavefunction that does transform according to a particular irrep of a group of transformations. This can be done if the characters of the irrep are known and a sensible starting point can be guessed. We state without proof that starting from any n-dimensional basis vector $\Psi \equiv (\Psi_1 \; \Psi_2 \; \cdots \; \Psi_n)^{\mathrm{T}}$, where $\{\Psi_i\}$ is a set of wavefunctions, the new vector $\Psi^{(\lambda)} \equiv (\Psi_1^{(\lambda)} \; \Psi_2^{(\lambda)} \; \cdots \; \Psi_n^{(\lambda)})^{\mathrm{T}}$ generated by

$$\Psi_i^{(\lambda)} = \sum_X \chi^{(\lambda)^*}(X) X \Psi_i \tag{29.24}$$

will transform according to the λth irrep. If the randomly chosen Ψ happens not to contain any component that transforms in the desired way then the $\Psi^{(\lambda)}$ so generated is found to be a zero vector and it is necessary to select a new starting vector. An illustration of the use of this 'projection operator' is given in the next example.

►*Consider a molecule made up of four iodine atoms lying at the corners of a square in the xy-plane, with a manganese atom at its centre, as shown in figure 29.4. Investigate whether the molecular orbital given by the superposition of p-state (angular momentum* $l = 1$*) atomic orbitals*

$$\Psi_1 = \Psi_y(\mathbf{r} - \mathbf{R}_1) + \Psi_x(\mathbf{r} - \mathbf{R}_2) - \Psi_y(\mathbf{r} - \mathbf{R}_3) - \Psi_x(\mathbf{r} - \mathbf{R}_4)$$

can bond to the d-state atomic orbitals of the manganese atom described by either (i) $\phi_1 = (3z^2 - r^2)f(r)$ *or* (ii) $\phi_2 = (x^2 - y^2)f(r)$, *where* $f(r)$ *is a function of r and so is unchanged by any of the symmetry operations of the molecule. Such linear combinations of atomic orbitals are known as ring orbitals.*

We have eight basis functions, the atomic orbitals $\Psi_x(N)$ and $\Psi_y(N)$, where $N = 1, 2, 3, 4$ and indicates the position of an iodine atom. Since the wavefunctions are those of p-states they have the forms $xf(r)$ or $yf(r)$ and lie in the directions of the x- and y-axes shown in the figure. Since r is not changed by any of the symmetry operations, $f(r)$ can be treated as a constant. The symmetry group of the system is $4mm$, whose character table is table 29.4.

Case (i). The manganese atomic orbital $\phi_1 = (3z^2 - r^2)f(r)$, lying at the centre of the molecule, is not affected by any of the symmetry operations since z and r are unchanged by them. It clearly transforms according to the identity irrep A_1. We therefore need to know which combination of the iodine orbitals $\Psi_x(N)$ and $\Psi_y(N)$, if any, also transforms according to A_1.

We use the projection operator (29.24). If we choose $\Psi_x(1)$ as the arbitrary one-dimensional starting vector, we unfortunately obtain zero (as the reader may wish to verify), but $\Psi_y(1)$ is found to generate a new non-zero one-dimensional vector transforming according to A_1. The results of acting on $\Psi_y(1)$ with the various symmetry elements X can be written down by inspection (see the discussion in section 29.2). So, for example, the $\Psi_y(1)$ orbital centred on iodine atom 1 and aligned along the positive y-axis is changed by the anticlockwise rotation of $\pi/2$ produced by R' into an orbital centred on atom 4 and aligned along the negative x-axis; thus $R'\Psi_y(1) = -\Psi_x(4)$. The complete set of group actions on $\Psi_y(1)$ is:

$$I,\ \Psi_y(1); \quad Q,\ -\Psi_y(3); \quad R,\ \Psi_x(2); \quad R',\ -\Psi_x(4);$$

$$m_x,\ \Psi_y(1); \quad m_y,\ -\Psi_y(3); \quad m_d,\ \Psi_x(2); \quad m_{d'},\ -\Psi_x(4).$$

Now $\chi^{(\mathrm{A}_1)}(X) = 1$ for all X, so (29.24) states that the sum of the above results for $X\Psi_y(1)$, all with weight 1, gives a vector (here, since the irrep is one-dimensional, just a wavefunction) that transforms according to A_1 and is therefore capable of forming a chemical bond with the manganese wavefunction ϕ_1. It is

$$\Psi^{(\mathrm{A}_1)} = 2[\Psi_y(1) - \Psi_y(3) + \Psi_x(2) - \Psi_x(4)],$$

though, of course, the factor 2 is irrelevant. This is precisely the ring orbital Ψ_1 given in the problem, but here it is generated rather than guessed beforehand.

Case (ii). The atomic orbital $\phi_2 = (x^2 - y^2)f(r)$ behaves as follows under the action of typical conjugacy class members:

$$I,\ \phi_2; \quad Q,\ \phi_2; \quad R,\ (y^2 - x^2)f(r) = -\phi_2; \quad m_x,\ \phi_2; \quad m_d,\ -\phi_2.$$

From this we see that ϕ_2 transforms as a one-dimensional irrep, but, from table 29.4, that irrep is B_1 not A_1 (the irrep according to which Ψ_1 transforms, as already shown). Thus ϕ_2 and Ψ_1 cannot form a bond. ◄

The original question did not ask for the the ring orbital to which ϕ_2 may bond, but it can be generated easily by using the values of $X\Psi_y(1)$ calculated in case (i) and now weighting them according to the characters of B_1:

$$\begin{aligned}\Psi^{(\mathrm{B}_1)} &= \Psi_y(1) - \Psi_y(3) + (-1)\Psi_x(2) - (-1)\Psi_x(4) \\ &\quad + \Psi_y(1) - \Psi_y(3) + (-1)\Psi_x(2) - (-1)\Psi_x(4) \\ &= 2[\Psi_y(1) - \Psi_x(2) - \Psi_y(3) + \Psi_x(4)].\end{aligned}$$

Now we will find the other irreps of $4mm$ present in the space spanned by the basis functions $\Psi_x(N)$ and $\Psi_y(N)$; at the same time this will illustrate the important point that since we are working with characters we are only interested in the diagonal elements of the representative matrices. This means (section 29.2) that if we work in the natural representation $\mathsf{D}^{\mathrm{nat}}$ we need consider only those functions that transform, wholly or partially, into themselves. Since we have no need to write out the matrices explicitly, their size (8×8) is no drawback. All the irreps spanned by the basis functions $\Psi_x(N)$ and $\Psi_y(N)$ can be determined by considering the actions of the group elements upon them, as follows.

(i) Under I all eight basis functions are unchanged, and $\chi(I) = 8$.

(ii) The rotations R, R' and Q change the value of N in every case and so all diagonal elements of the natural representation are zero and $\chi(R) = \chi(Q) = 0$.

(iii) m_x takes x into $-x$ and y into y and, for $N = 1$ and 3, leaves N unchanged, with the consequences (remember the forms of $\Psi_x(N)$ and $\Psi_y(N)$) that

$$\begin{aligned} &\Psi_x(1) \rightarrow -\Psi_x(1), \quad \Psi_x(3) \rightarrow -\Psi_x(3), \\ &\Psi_y(1) \rightarrow \Psi_y(1), \qquad \Psi_y(3) \rightarrow \Psi_y(3). \end{aligned}$$

Thus $\chi(m_x)$ has four non-zero contributions, -1, -1, 1 and 1, together with four zero contributions. The total is thus zero.

(iv) m_d and $m_{d'}$ leave no atom unchanged and so $\chi(m_d) = 0$.

The character set of the natural representation is thus 8, 0, 0, 0, 0, which, either by inspection or by applying formula (29.18), shows that

$$\mathsf{D}^{\text{nat}} = \text{A}_1 \oplus \text{A}_2 \oplus \text{B}_1 \oplus \text{B}_2 \oplus 2\text{E},$$

i.e. that all possible irreps are present. We have constructed previously the combinations of $\Psi_x(N)$ and $\Psi_y(N)$ that transform according to A_1 and B_1. The others can be found in the same way.

29.11.2 Matrix elements in quantum mechanics

In section 29.10 we outlined the procedure for determining whether a matrix element that involves the product of three factors as an integrand is necessarily zero. We now illustrate this with a specific worked example.

►*Determine whether a 'dipole' matrix element of the form*

$$J = \int \Psi_{d_1} x \Psi_{d_2}\, d\tau,$$

where Ψ_{d_1} and Ψ_{d_2} are d-state wavefunctions of the forms $xyf(r)$ and $(x^2 - y^2)g(r)$ respectively, can be non-zero (i) *in a molecule with symmetry C_{3v} (or 3m), such as ammonia, and* (ii) *in a molecule with symmetry C_{4v} (or 4mm), such as the MnI$_4$ molecule considered in the previous example.*

We will need to make reference to the character tables of the two groups. The table for C_{3v} is table 29.1 (section 29.6); that for C_{4v} is reproduced as table 29.5 from table 29.4 but with the addition of another column showing how some common functions transform.

We make use of (29.23), extended to the product of three functions. No attention need be paid to $f(r)$ and $g(r)$ as they are unaffected by the group operations.

Case (i). From the character table 29.1 for C_{3v}, we see that each of xy, x and $x^2 - y^2$ forms part of a basis set transforming according to the two-dimensional irrep E. Thus we may fill in the array of characters (using chemical notation for the classes, except that we continue to use I rather than E) as shown in table 29.6. The last line is obtained by

$4mm$	I	Q	R, R'	m_x, m_y	$m_d, m_{d'}$	
A_1	1	1	1	1	1	z; z^2; $x^2 + y^2$
A_2	1	1	1	−1	−1	R_z
B_1	1	1	−1	1	−1	$x^2 - y^2$
B_2	1	1	−1	−1	1	xy
E	2	−2	0	0	0	(x, y); (xz, yz); (R_x, R_y)

Table 29.5 The character table for the irreps of group $4mm$ (or C_{4v}). The right-hand column lists some common functions, or, for the two-dimensional irrep E, pairs of functions, that transform according to the irrep against which they are shown.

Function	Irrep	Classes		
		I	$2C_3$	$3\sigma_v$
xy	E	2	−1	0
x	E	2	−1	0
$x^2 - y^2$	E	2	−1	0
product		8	−1	0

Table 29.6 The character sets, for the group C_{3v} (or $3mm$), of three functions and of their product $x^2y(x^2 - y^2)$.

Function	Irrep	Classes				
		I	C_2	$2C_6$	$2\sigma_v$	$2\sigma_d$
xy	B_2	1	1	−1	−1	1
x	E	2	−2	0	0	0
$x^2 - y^2$	B_1	1	1	−1	1	−1
product		2	−2	0	0	0

Table 29.7 The character sets, for the group C_{4v} (or $4mm$), of three functions, and of their product $x^2y(x^2 - y^2)$.

multiplying together the corresponding characters for each of the three elements. Now, by inspection, or by applying (29.18), i.e.

$$m_{A_1} = \tfrac{1}{6}[1(1)(8) + 2(1)(-1) + 3(1)(0)] = 1,$$

we see that irrep A_1 does appear in the reduced representation of the product, and so J is not necessarily zero.

Case (ii). From table 29.5 we find that, under the group C_{4v}, xy and $x^2 - y^2$ transform as irreps B_2 and B_1 respectively and that x is part of a basis set transforming as E. Thus the calculation table takes the form of table 29.7 (again, chemical notation for the classes has been used).

Here inspection is sufficient, as the product is exactly that of irrep E and irrep A_1 is certainly not present. Thus J is necessarily zero and the dipole matrix element vanishes. ◄

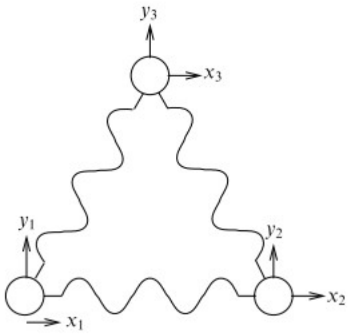

Figure 29.5 An equilateral array of masses and springs.

29.11.3 Degeneracy of normal modes

As our final area for illustrating the usefulness of group theoretical results we consider the normal modes of a vibrating system (see chapter 9). This analysis has far-reaching applications in physics, chemistry and engineering. For a given system, normal modes that are related by some symmetry operation have the same frequency of vibration; the modes are said to be *degenerate*. It can be shown that such modes span a vector space that transforms according to some irrep of the group $\mathcal{G}$ of symmetry operations of the system. Moreover, the degeneracy of the modes equals the dimension of the irrep. As an illustration, we consider the following example.

►*Investigate the possible vibrational modes of the equilateral triangular arrangement of equal masses and springs shown in figure 29.5. Demonstrate that two are degenerate.*

Clearly the symmetry group is that of the symmetry operations on an equilateral triangle, namely $3m$ (or C_{3v}), whose character table is table 29.1. As on a previous occasion, it is most convenient to use the natural representation D^{nat} of this group (it almost always saves having to write out matrices explicitly) acting on the six-dimensional vector space $(x_1, y_1, x_2, y_2, x_3, y_3)$. In this example the natural and regular representations coincide, but this is not usually the case.

We note that in table 29.1 the second class contains the rotations A (by $\pi/3$) and B (by $2\pi/3$), also known as R and R'. This class is known as 3_z in crystallographic notation, or C_3 in chemical notation, as explained in section 29.9. The third class contains C, D, E, the three mirror reflections.

Clearly $\chi(I) = 6$. Since all position labels are changed by a rotation, $\chi(3_z) = 0$. For the mirror reflections the simplest representative class member to choose is the reflection m_y in the plane containing the y_3-axis, since then only label 3 is unchanged; under m_y, $x_3 \to -x_3$ and $y_3 \to y_3$, leading to the conclusion that $\chi(m_y) = 0$. Thus the character set is 6, 0, 0.

Using (29.18) and the character table 29.1 shows that

$$\mathsf{D}^{\text{nat}} = \mathrm{A}_1 \oplus \mathrm{A}_2 \oplus 2\mathrm{E}.$$

However, we have so far allowed x_i, y_i to be completely general, and we must now identify and remove those irreps that do not correspond to vibrations. These will be the irreps corresponding to bodily translations of the triangle and to its rotation without relative motion of the three masses.

Bodily translations are linear motions of the centre of mass, which has coordinates

$$x = (x_1 + x_2 + x_3)/3 \quad \text{and} \quad y = (y_1 + y_2 + y_3)/3).$$

Table 29.1 shows that such a coordinate pair (x, y) transforms according to the two-dimensional irrep E; this accounts for one of the two such irreps found in the natural representation.

It can be shown that, as stated in table 29.1, planar bodily rotations of the triangle – rotations about the z-axis, denoted by R_z – transform as irrep A_2. Thus, when the linear motions of the centre of mass, and pure rotation about it, are removed from our reduced representation, we are left with $E \oplus A_1$. So, E and A_1 must be the irreps corresponding to the internal vibrations of the triangle – one doubly degenerate mode and one non-degenerate mode.

The physical interpretation of this is that two of the normal modes of the system have the same frequency and one normal mode has a different frequency (barring accidental coincidences for other reasons). It may be noted that in quantum mechanics the energy quantum of a normal mode is proportional to its frequency. ◀

In general, group theory does not tell us what the frequencies are, since it is entirely concerned with the symmetry of the system and not with the values of masses and spring constants. However, using this type of reasoning, the results from representation theory can be used to predict the degeneracies of atomic energy levels and, given a perturbation whose Hamiltonian (energy operator) has some degree of symmetry, the extent to which the perturbation will resolve the degeneracy. Some of these ideas are explored a little further in the next section and in the exercises.

29.11.4 Breaking of degeneracies

If a physical system has a high degree of symmetry, invariant under a group $\mathcal{G}$ of reflections and rotations, say, then, as implied above, it will normally be the case that some of its eigenvalues (of energy, frequency, angular momentum etc.) are degenerate. However, if a perturbation that is invariant only under the operations of the elements of a smaller symmetry group (a subgroup of $\mathcal{G}$) is added, some of the original degeneracies may be broken. The results derived from representation theory can be used to decide the extent of the degeneracy-breaking.

The normal procedure is to use an N-dimensional basis vector, consisting of the N degenerate eigenfunctions, to generate an N-dimensional representation of the symmetry group of the perturbation. This representation is then decomposed into irreps. In general, eigenfunctions that transform according to different irreps no longer share the same frequency of vibration. We illustrate this with the following example.

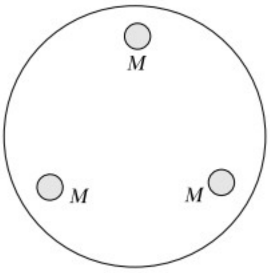

Figure 29.6 A circular drumskin loaded with three symmetrically placed masses.

►*A circular drumskin has three equal masses placed on it at the vertices of an equilateral triangle, as shown in figure 29.6. Determine which degenerate normal modes of the drumskin can be split in frequency by this perturbation.*

When no masses are present the normal modes of the drum-skin are either non-degenerate or two-fold degenerate (see chapter 21). The degenerate eigenfunctions Ψ of the nth normal mode have the forms

$$J_n(kr)(\cos n\theta)e^{\pm i\omega t} \qquad \text{or} \qquad J_n(kr)(\sin n\theta)e^{\pm i\omega t}.$$

Therefore, as explained above, we need to consider the two-dimensional vector space spanned by $\Psi_1 = \sin n\theta$ and $\Psi_2 = \cos n\theta$. This will generate a two-dimensional representation of the group $3m$ (or C_{3v}), the symmetry group of the perturbation. Taking the easiest element from each of the three classes (identity, rotations, and reflections) of group $3m$, we have

$$\begin{aligned}
&I\Psi_1 = \Psi_1, \qquad I\Psi_2 = \Psi_2,\\
&A\Psi_1 = \sin\left[n\left(\theta - \tfrac{2}{3}\pi\right)\right] = \left(\cos\tfrac{2}{3}n\pi\right)\Psi_1 - \left(\sin\tfrac{2}{3}n\pi\right)\Psi_2,\\
&A\Psi_2 = \cos\left[n\left(\theta - \tfrac{2}{3}\pi\right)\right] = \left(\cos\tfrac{2}{3}n\pi\right)\Psi_2 + \left(\sin\tfrac{2}{3}n\pi\right)\Psi_1,\\
&C\Psi_1 = \sin[n(\pi - \theta)] = -(\cos n\pi)\Psi_1,\\
&C\Psi_2 = \cos[n(\pi - \theta)] = (\cos n\pi)\Psi_2.
\end{aligned}$$

The three representative matrices are therefore

$$\mathsf{D}(I) = \mathsf{I}_2, \quad \mathsf{D}(A) = \begin{pmatrix} \cos\frac{2}{3}n\pi & -\sin\frac{2}{3}n\pi \\ \sin\frac{2}{3}n\pi & \cos\frac{2}{3}n\pi \end{pmatrix}, \quad \mathsf{D}(C) = \begin{pmatrix} -\cos n\pi & 0 \\ 0 & \cos n\pi \end{pmatrix}.$$

The characters of this representation are $\chi(I) = 2$, $\chi(A) = 2\cos(2n\pi/3)$ and $\chi(C) = 0$. Using (29.18) and table 29.1, we find that

$$\begin{aligned}
m_{\mathrm{A}_1} &= \tfrac{1}{6}\left(2 + 4\cos\tfrac{2}{3}n\pi\right) = m_{\mathrm{A}_2}\\
m_{\mathrm{E}} &= \tfrac{1}{6}\left(4 - 4\cos\tfrac{2}{3}n\pi\right).
\end{aligned}$$

Thus

$$\mathsf{D} = \begin{cases} \mathrm{A}_1 \oplus \mathrm{A}_2 & \text{if } n = 3,\ 6,\ 9,\ \ldots,\\ \mathrm{E} & \text{otherwise.} \end{cases}$$

Hence the normal modes $n = 3,\ 6,\ 9,\ \ldots$ each transform under the operations of $3m$

as the sum of two one-dimensional irreps and, using the reasoning given in the previous example, are therefore split in frequency by the perturbation. For other values of n the representation is irreducible and so the degeneracy cannot be split. ◄

29.12 Exercises

29.1 A group $\mathcal{G}$ has four elements I, X, Y and Z, which satisfy $X^2 = Y^2 = Z^2 = XYZ = I$. Show that $\mathcal{G}$ is Abelian and hence deduce the form of its character table.

Show that the matrices

$$\mathsf{D}(I) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathsf{D}(X) = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix},$$
$$\mathsf{D}(Y) = \begin{pmatrix} -1 & -p \\ 0 & 1 \end{pmatrix}, \qquad \mathsf{D}(Z) = \begin{pmatrix} 1 & p \\ 0 & -1 \end{pmatrix},$$

where p is a real number, form a representation D of $\mathcal{G}$. Find its characters and decompose it into irreps.

29.2 Using a square whose corners lie at coordinates $(\pm 1, \pm 1)$, form a natural representation of the dihedral group $\mathcal{D}_4$. Find the characters of the representation, and, using the information (and class order) in table 29.4 (p. 1102), express the representation in terms of irreps.

Now form a representation in terms of eight 2×2 orthogonal matrices, by considering the effect of each of the elements of $\mathcal{D}_4$ on a general vector (x, y). Confirm that this representation is one of the irreps found using the natural representation.

29.3 The quaternion group $\mathcal{Q}$ (see exercise 28.20) has eight elements $\{\pm 1, \pm i, \pm j, \pm k\}$ obeying the relations

$$i^2 = j^2 = k^2 = -1, \quad ij = k = -ji.$$

Determine the conjugacy classes of $\mathcal{Q}$ and deduce the dimensions of its irreps. Show that $\mathcal{Q}$ is homomorphic to the four-element group $\mathcal{V}$, which is generated by two distinct elements a and b with $a^2 = b^2 = (ab)^2 = I$. Find the one-dimensional irreps of $\mathcal{V}$ and use these to help determine the full character table for $\mathcal{Q}$.

29.4 Construct the character table for the irreps of the permutation group S_4 as follows.

(a) By considering the possible forms of its cycle notation, determine the number of elements in each conjugacy class of the permutation group S_4, and show that S_4 has five irreps. Give the logical reasoning that shows they must consist of two three-dimensional, one two-dimensional, and two one-dimensional irreps.

(b) By considering the odd and even permutations in the group S_4, establish the characters for one of the one-dimensional irreps.

(c) Form a natural matrix representation of 4×4 matrices based on a set of objects $\{a, b, c, d\}$, which may or may not be equal to each other, and, by selecting one example from each conjugacy class, show that this natural representation has characters 4, 2, 1, 0, 0. In the four-dimensional vector space in which each of the four coordinates takes on one of the four values a, b, c or d, the one-dimensional subspace consisting of the four points with coordinates of the form $\{a, a, a, a\}$ is invariant under the permutation group and hence transforms according to the invariant irrep A_1. The remaining three-dimensional subspace is irreducible; use this and the characters deduced above to establish the characters for one of the three-dimensional irreps, T_1.

' $

Fractals, Chaos, Power Laws
Schroeder.

Mathematical methods for physicists
Chow

$$\sqrt{r^2 d\theta^2 + dr^2}$$

↓

Distance between infinitesimally separated points in Cartesian coordinates (top) and polar coordinates (bottom), as given by Pythagoras's theorem.

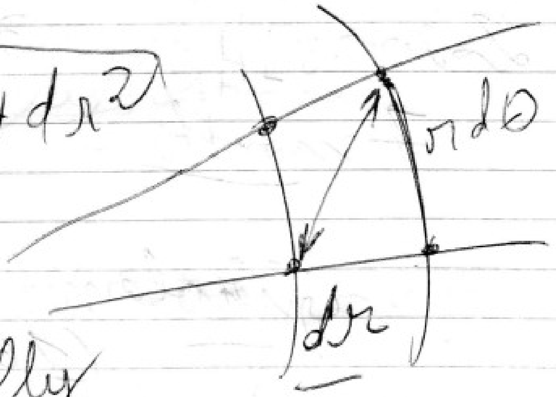

$$\sum_{i,j}^{n}$$

12
110

$\tau o \psi o \upsilon \lambda o \gamma O \varsigma$

Como calcular o comprimento ?
do arco de uma circunferência?

44' → 66

44-51
167-172
80-85

20200814　319d9c1f1

7 months

Girolamo Saccheri
(1667-1733)

1733

↳ Euclides absoluto de toda falha.

Lobachevski 1826

Bolyai

⇒ Riemann　ungarav@gmail.com

hiperespacio

(el electron es zurdo y otros ensayos científicos!

5º Postulado Euclides

Si una recta corta a otras (otras) dos formando ángulos internos por el mismo lado que suman menos de dos ángulos rectos, esas dos rectas prolongadas indefinidamente se cortarán por ese lado en que los ángulos suman menos de dos rectos.

Source,

youtube.com/watch?=v=
3_OTz3m2W_g

Mathematical Methods for Physicists
Herman

---

user/perezlaser/

gaussianos.com

Prof. Ezequiel (Profesor)

A Rainha das Ciências Garbi

luz dodrode
→ "anel de Einstein"!

20200818

Euclid's Elements,
Completing Oliver Byrne's
work
200 €

Pág. 258 – the Delian Quest
2020

Source:
www.archive.org/details/
AUL070416

Delian Solution
www.archive.org/details/
The Delian Solution

Pappi Alexandrini Collectionis
Volume 1, 2, 3

Quadrature du Cercle
(Gallico)

Itamaratuiá, 77

Acesso Soluções de Pagamentos S.A
CNPJ 13.140.088/0001-99
Av. Rebouças 1368 Pinheiros
CEP 05402-100

---

www.meuacesso.com.br

Port Malone – Circles

- "" H " E " (#"H- E@$/($$/ " M
- "" H " E " (#"H- E@ # // M 0 N

;<)=; ; ;8 ; O!P , O N O
O- OQ RI H O -- 0 OH- E/ $#%%$ -!
H ;<)=; ; ;8 ;
O!P , O N O! O P& OQ R* 2OS - OH- E#%$% ! -!
;<)=; ; ;8 ;
0 T O O O ! N O A OQ R0 NOI H OH- E@ # // -!

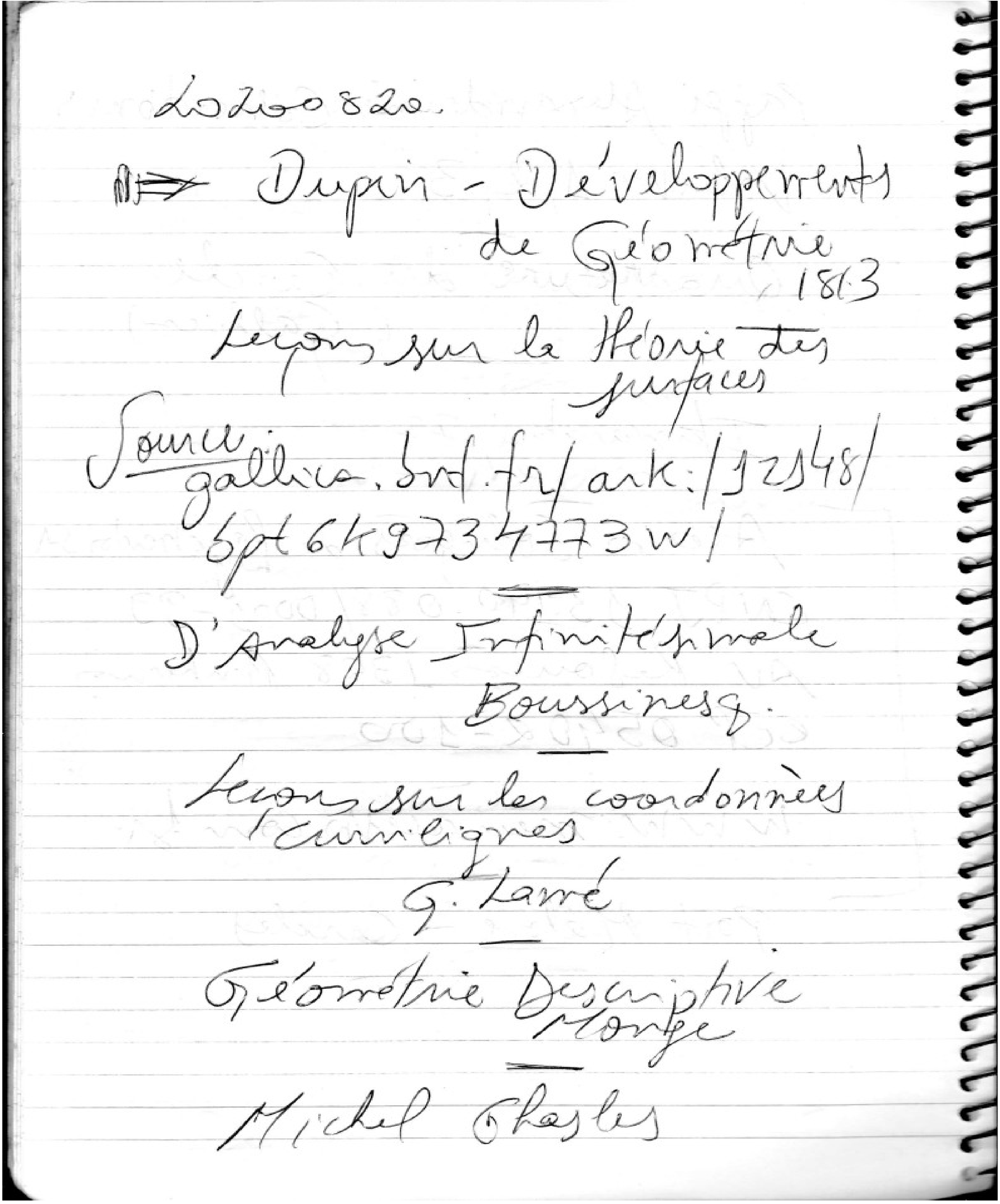

Lo2oo820.

⟹ Dupin - Développements de Géométrie 1813

Leçons sur la théorie des surfaces

Source: gallica.bnf.fr/ark:/12148/bpt6k9734773w/

D'Analyse Infinitésimale Boussinesq.

Leçons sur les coordonnées curvilignes G. Lamé

Géométrie Descriptive Monge

Michel Chasles

Prof.
Stenio Musich

www.steniomusich.com
www.matemática.org
www.mathematics.pictures
www.mathematics.page

20211112.

Obs: Entrou no lugar da página 33, pois o sistema pulou e eu refiz o escaneamento da página em questão.

20200821.

https://publikationsserver.tu-braunschweig.de/receive/dbbs_mods_00049589

Important:

https://echo.mpiwg-berlin.mpg.de/content/mathematics/mathematics/

Adrien Marie Legendre

20200822

$x^2+y^2=\underline{1}$ (Wrong!!!)

$x^2+y^2=\underline{\underline{\underline{4}}}$

$\underline{1}^2=\underline{\underline{2}}$

$a^2=b^2+c^2$

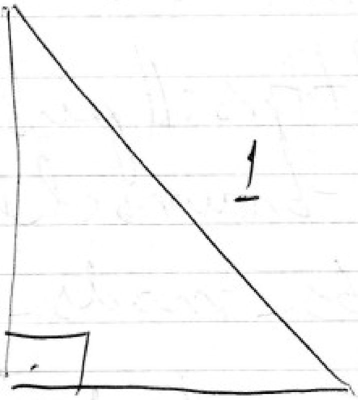

$1=\sqrt{a^2+b^2}$

$\Downarrow$

a^2+b^2

$1^2+1^2=\sqrt{2}$

$\neq$

$1=\sqrt{2}$ (wrong!!!)

Vejamos:

$? = \sqrt{a^2+b^2}$

$\sqrt{1^2+1^2}$

$\sqrt{2+2}$

$2=\sqrt{4}$

e não 1!!! fera!!!???

Correto!!!

$1^2 = 2$ pois $2^2 = 4$

(seguindo-se o mesmo princípio de cálculo)

Vejamos Radian!

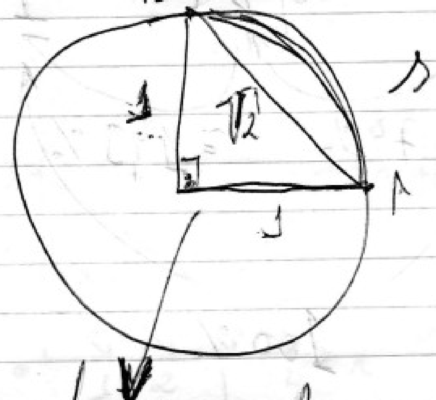

s=r (wrong!!!)

É um triângulo retângulo

Triângulo pitagórico:

3, 4, 5

Meu:

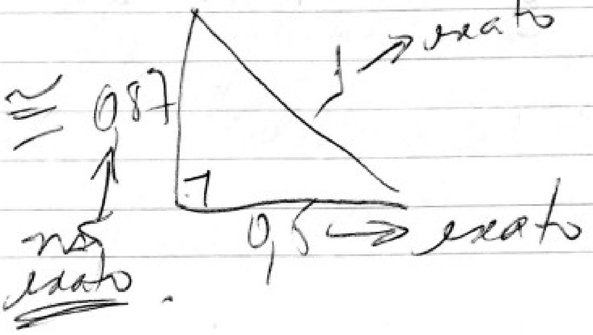
≅ 0,87
não exato
1 → exato
0,5 → exato

Vejamos a proporção:

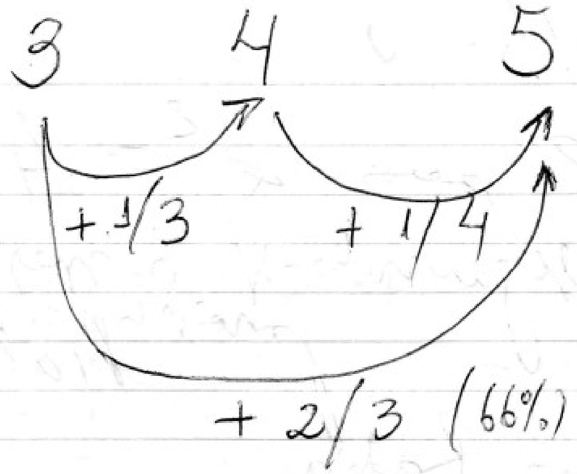
3
4
5
+ 1/3
+ 1/4
+ 2/3 (66%)

$\frac{1}{3} + \frac{1}{4} = \frac{4+3=7}{12} \frac{}{12}$

70 |12
100 0,58$\overline{3}$
040
04

$0,87 - 100$
$1 - x$

$0,87x = 100$
$x = \frac{100}{0,87}$

10000 |0,87
130 114,9
430
800
17

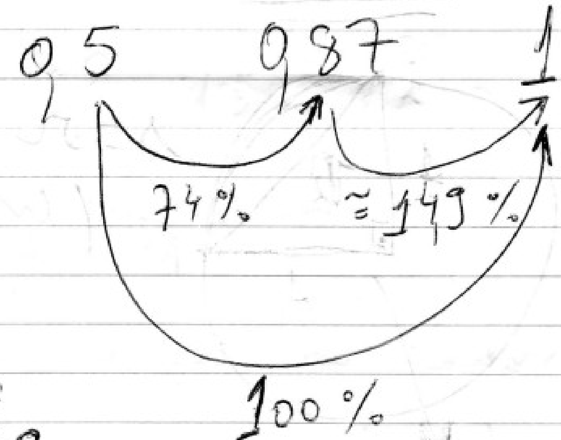
0,5
0,87
1
74%
≈ 14,9 %
100 %

$0,5 - 100$
$0,87 - x$

$0,5x = 87$
$x = \frac{87}{0,5}$

870 |0,5
37 174 %
20

74% a mai

Outras observações:

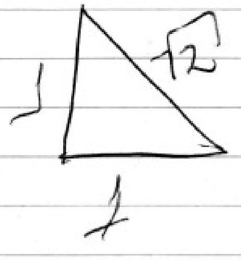

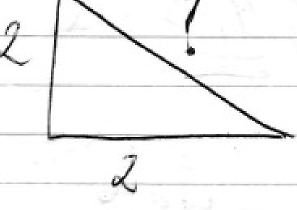

$? = \sqrt{a^2 + b^2}$
$? = \sqrt{2^2 + 2^2}$
$? = \sqrt{4 + 4}$
$? = \sqrt{8}$

Ficar!!! no mesmo triângulo.

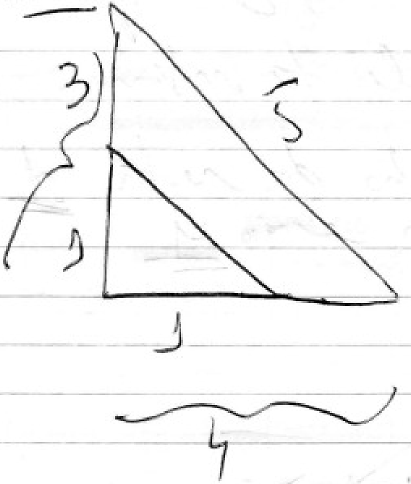

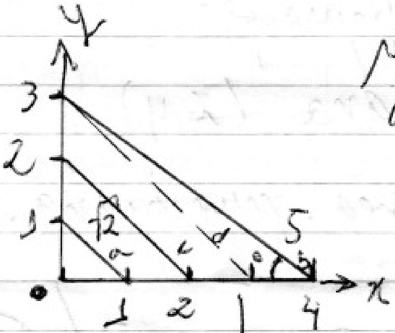

retas
a ∦ b → não são paralelos

a ∥ c = são paralelos.

$$3^2 + 4^2 = 5^2$$
$$9 + 16 = 25$$
$$25 = 25$$

Calcular o ângulo da Representação (triângulo) pitagórico sobre o eixo x (calcule).
θ

curvando-se este triângulo pitagórico sobre uma superfície curva até que a reta b fique paralela às retas a, c.

Calcule a curvatura do eixo x (função) (eixo curvilíneo x) no exato ponto em que as 3 retas fiquem paralelas.

Calcule também:

o ângulo da reta d em relação ao eixo y

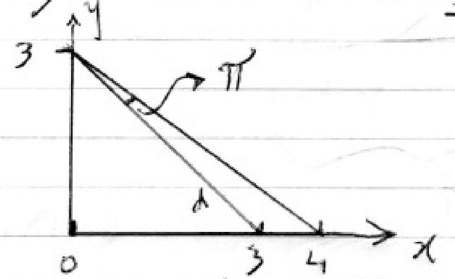

nas 2 configurações:

a) cartesiana (x, y)

b) curvilínea murichiana. (x_m, y_m)

20200822 0245

Simule a função curvilínea até o ponto em que esta se iguala à equação $x^2 + y^2 = 1$ ou 4

o eixo x faz o mesmo traçado da equação $x^2 + y^2 = 1$ ou 4 (circunferência)

veja se as quatro retas estão em paralelo.

$x^2 + y^2 = 1$ ou $\underline{4}$

$a^2 + b^2 = r^2$

$\underbrace{1^2 + 1^2}_{4} = r^2 = r^2 = 4$

$r = \sqrt{4}$

$r = 2$

Analizando

$(0,5)^{\wedge 2} + 0,5^{\wedge 2} + 0,5^{\wedge 2} + 0,5^{\wedge 2} = 1$

$0,25 + 0,25 + 0,25 + 0,25 = 1$

$x^2 + y^2 = 4$

$1 + 1 + 1 + 1 = 4$

$2^1 + 2^1 = 4$

$x^2 + y^2 = 4$

$(\sqrt{2})^2 + (\sqrt{2})^2 = 4$ (4x)

obs $\frac{1}{\sqrt{2}} = 0,70710678 = \underline{\underline{\sqrt{0,5}}}$

$\sqrt{0,5} =$

0,70710678...

vejamos

0,25 x 2 = 0,50

0,25 x 0,25 =

0,0625

$0,5^{\wedge 2} = 0,25$

vejamos

0,5 x 0,5 = 0,25

0,5 x 0,5 = 0,25

0,5 x 0,5 = 0,25

0,5 x 0,5 = 25

$0,5 + \frac{1}{\sqrt{2}} = 1,20710678$

diferença $\sqrt{0,5}$

$0,5 + 1\sqrt{2} = 1,91421356$

$1 + 1 + 1 + 1 = 4$

$0,5 + 0,5$

$(0,5)^2 + (0,5)^2 \quad (0,5)^2 + (0,5)^2$

$0,25 + 0,25 \quad 0,25 + 0,25$

$\Sigma 4$

16 partes (0,25)

$a^2 + b^2$

8 partes (0,5)

$(0,5)^2 \times 8$

Recalculando:

vejamos

$(0,5)^2 + (0,5)^2 + (0,5)^2 + (0,5)^2 \Rightarrow$ Ⓐ

$0,25 \quad 0,25 \quad 0,25 \quad 0,25 = 1$ $+$

$(0,5)^2 + (0,5)^2 + (0,5)^2 + (0,5)^2 \Rightarrow$ Ⓑ

$+$

$(0,5)^2 + (0,5)^2 + (0,5)^2 + (0,5)^2 \Rightarrow$ Ⓒ

$+$

$(0,5)^2 + (0,5)^2 + (0,5)^2 + (0,5)^2 \Rightarrow D$

$=$

④

$a^2 + b^2 + c^2 + d^2 = 1$

$\times 4 = 4$

ou

$1^4 = 4$

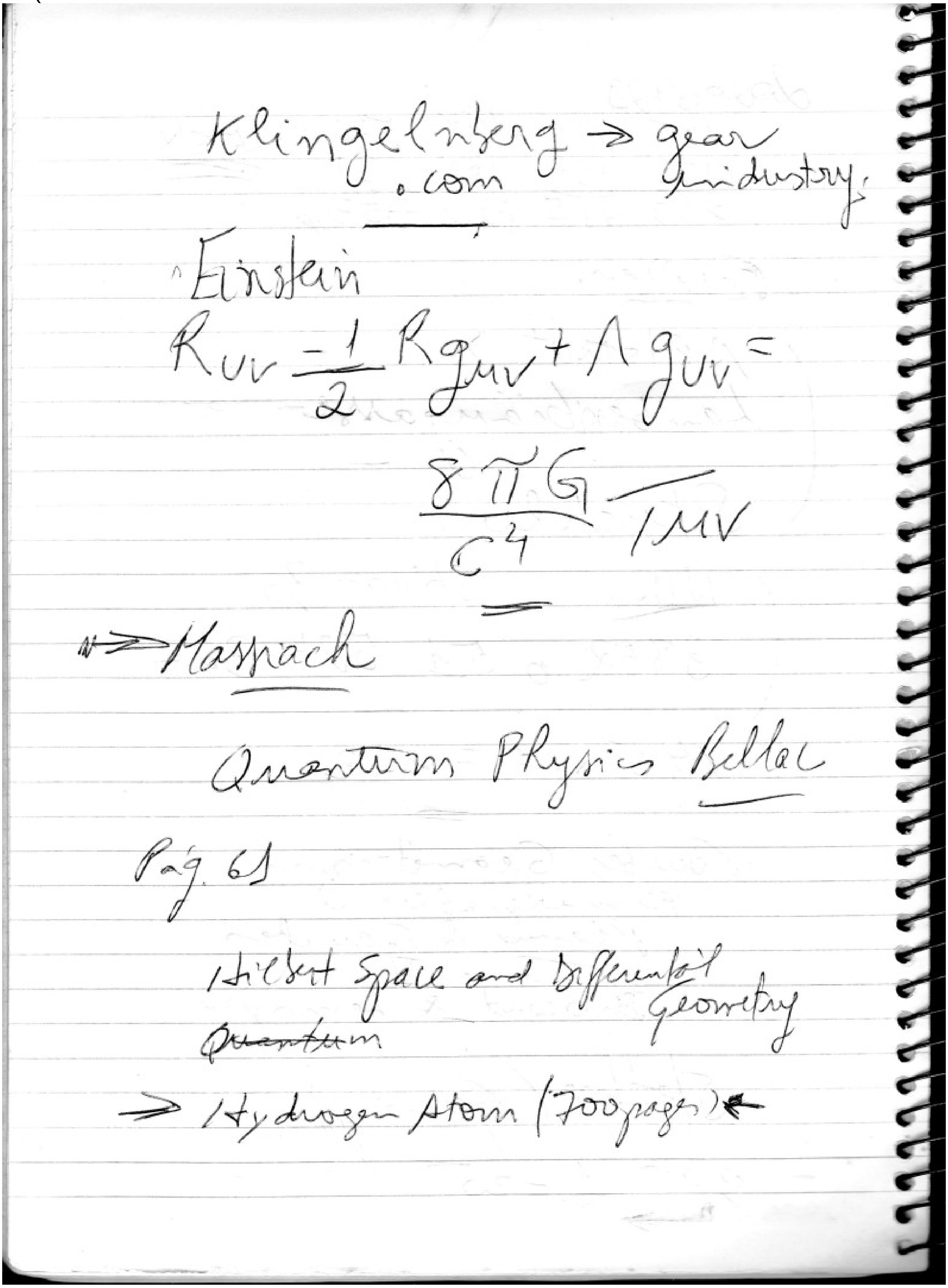

Klingelnberg → gear
.com industry.

Einstein

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

→ Massach

Quantum Physics Bellac

Pag. 61

Hilbert Space and Differential Geometry
~~Quantum~~

→ Hydrogen Atom (700 pages) ←

20200831.

→ Siebenter (7)

2030 - 1
5559 - 2
5560 - 3

5561 - 5

5558 - 1

58 - 1

11112033 → 7
11112031 → 4
11112032 - 6

11173197 - 6

10049140
↳ Theoria Motus

10082225
Trigonometry

Opere Di Galileo Galilei
10525557 - 3 tomo
56 - 2 tomo
55 - 1 tomo

https://reader.digitale-sammlungen.de/de/fs1/object/display/bsb10525555_00007.html

) "" ! ! " "
) "" !* + , ! "! + + -.$(-
%) /) ! 0 0
) ""! ! ! ! "1//2 "! !) 3 !4+ + - ----
) ""! ! ! ! "1//2 "! !) 3 !4+ + - ---
) ""! ! ! ! "1//2 "! !) 3 !4+ + - ---&

) ""! ! ! ! "1//2 "! !) 3 !4+ + -.$(-
) ""! ! ! ! "1//2 "! !) 3 !4+ + -.$(
) ""! ! ! ! "1//2 "! !) 3 !4+ + -.$(&

' #.

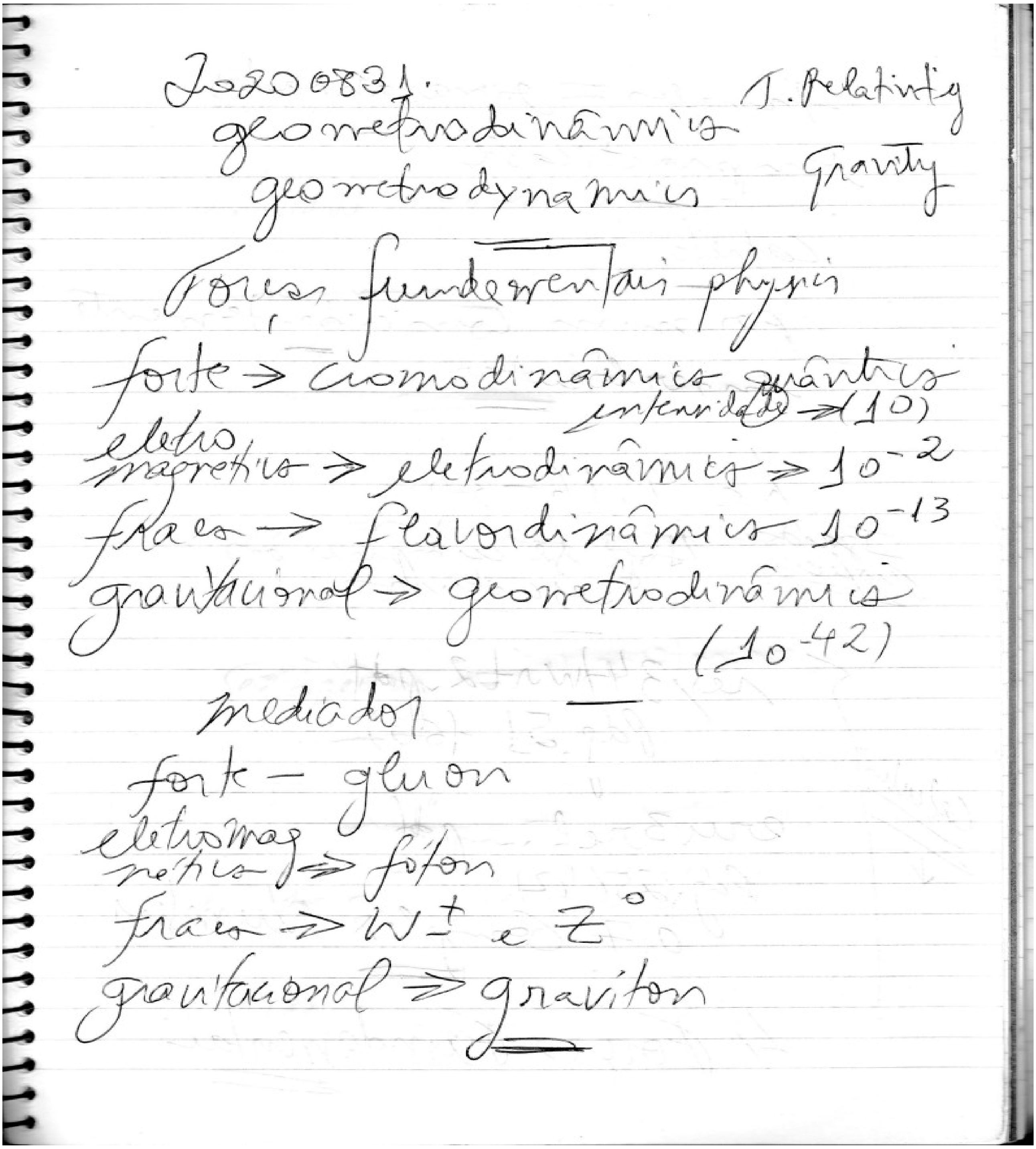

20200831.
T. Relativity
geometrodinâmica
geometrodynamics
Gravity

Forças fundamentais physics

forte → cromodinâmica quântica
intensidade → (10)
eletromagnética → eletrodinâmica → 10^{-2}
fraca → flavordinâmica 10^{-13}
gravitacional → geometrodinâmica
(10^{-42})

mediador

forte – gluon
eletromagnética → fóton
fraca → $W^{\pm}$ e Z^0
gravitacional → graviton

20200911.

Traité de géométrie supérieure
Chasles

Pag. 732 y 742

Fig. 184 p. 564.

Pag. 564 (660)

—

Kolmogorov

—

$1 + \sqrt{5}/2 = 1.618034.$

—

Ramanujan

$\frac{9801\sqrt{2}}{4412}$ for π

Namagiri Thayar (Mother Lakshmi)
(Goddess Mahalakshmi)
of Namakkal

—

(

' $ 1 7 8 , ! !)

9:5;9 9 9/ 9 (8 , ! !) 0 ! !
/ ! < +! ,) : 2 5)! = - -)

9:5;9 9 9/ 9 (8 , ! !) 0 ! !
/ ! < +! ,) : 2 5)! = #(-)

' $ (- &

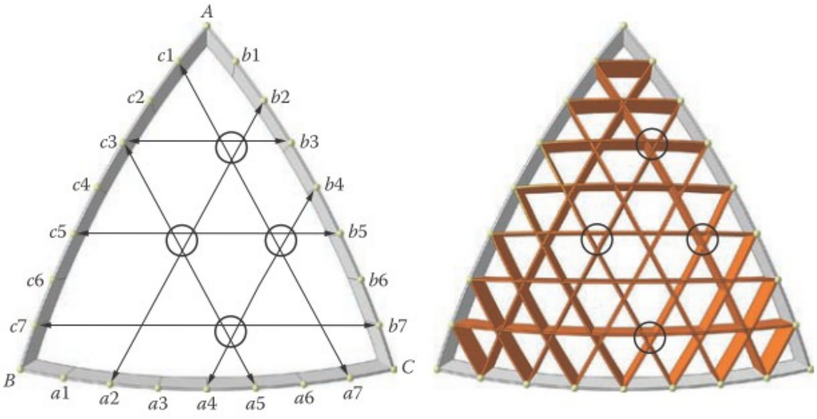
A
c1
b1
c2
b2
c3
b3
c4
b4
c5
b5
c6
b6
c7
b7
B
C
a1
a2
a3
a4
a5
a6
a7

(a) (b)

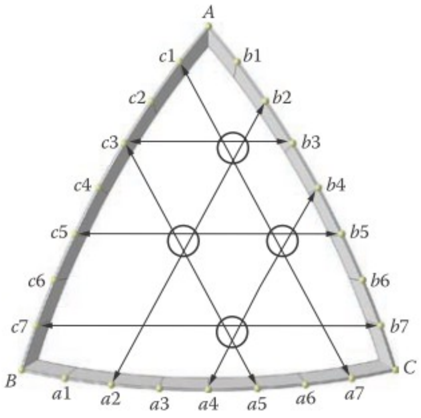
A
c1
b1
c2
b2
c3
b3
c4
b4
c5
b5
c6
b6
c7
b7
B
C
a1
a2
a3
a4
a5
a6
a7

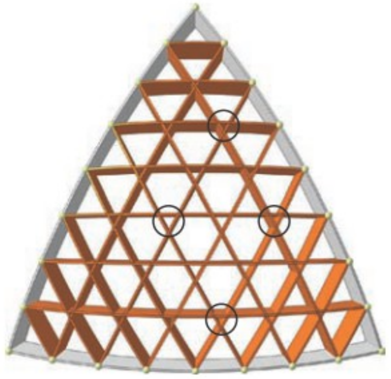

8.5.7 Equal-Arcs (Three Great Circles)

The previous example cast some doubts on the advantages of using equal-arc reference points to define great circle intersections. However, when three intersecting great circles are used to define grid points instead of two, the results are outstanding, yielding the most uniform tessellation gird of the three schemas thus far. Like the previous Equal-arcs (two great circles) schema, this one is not without its surprises, but they are easy to deal with.

In this schema, equal-arc reference points along the PPT's edges are selected so that pairs of them define three great circle arcs approximately parallel to their respective PPT edge. Unlike Equal-chords schemas where great circles reference equal-chord points, those based on equal-arc references do not intersect at a single line. Each pair of great circles intersects at a different line and defines its own pair of antipodal points. With three intersecting great circles, six antipodal points are created and three of them (one from each antipodal pair) fall on the surface of the PPT. Figure 8.17(a) illustrates the result.

In Figure 8.17(a), examples of four three-way intersections are circled. Taking just one of these examples, the great circle arcs between *c*3-*b*3, *c*1-*a*7 and *b*2-*a*2 intersect at three different places, creating three different points within the PPT. These points are on the surface of the sphere but they form a small window triangle rather than the single tessellation vertex we need. Figure 8.17(b) shows the 21 window triangles that result when the entire PPT is subdivided. Notice that the windows become larger and larger towards the center of the PPT; the smallest ones are the ones nearest the PPT's three apices.

Each window is actually a small triangle whose planar surface is just inside the sphere; only the apices of the window triangle are actually on the surface of the sphere. Figure 8.18(a) shows a detail of the window and its vertices. The red arcs represent the great circle arcs, and the small balls represent their intersection points. What is needed is a single point that summarizes the three window vertices. The triangle's centroid is the most appropriate choice because if the triangle were continuously scaled down while maintaining the proportions of its sides, its geometry would converge at a single point at its centroid. The centroid of a triangle is defined as the intersection of its medians. The average of its three

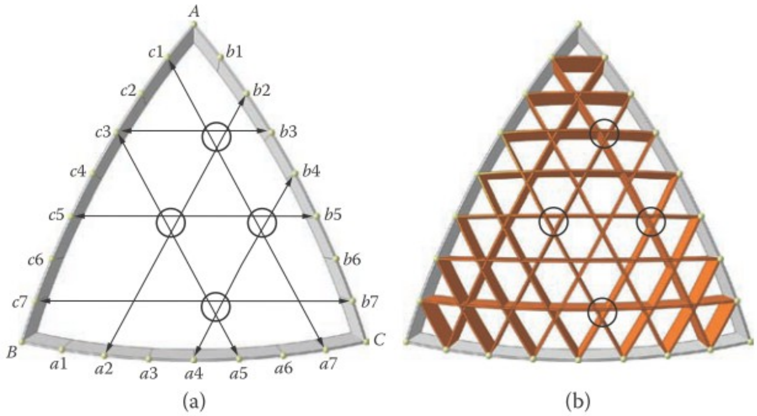

(a) (b)

Figure 8.17. Equal-arcs (three great circles) schema.

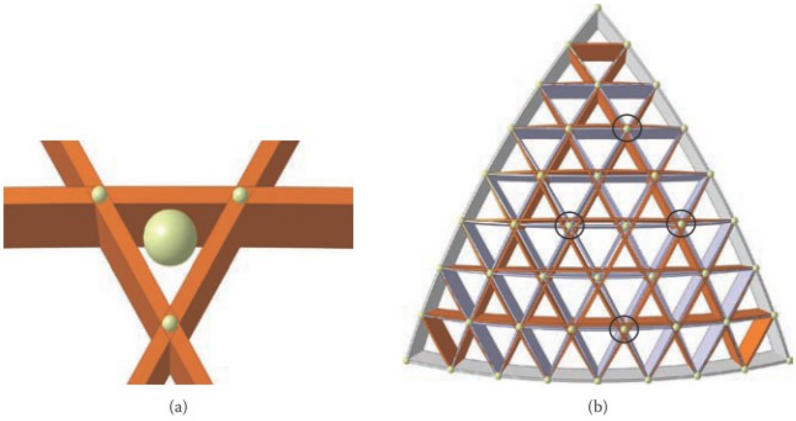
(a) (b)

Figure 8.18. Equal-arcs windows and tessellation vertices.

vertex coordinates produces the same point we call C. To find the window's centroid point:

$$Cx = \frac{(x_1 + x_2 + x_3)}{3},$$
$$Cy = \frac{(y_1 + y_2 + y_3)}{3},$$
$$Cz = \frac{(z_1 + z_2 + z_3)}{3}.$$

We are almost done; only one small detail remains. The window's centroid is on the plane of the triangle, which is inside the sphere. The point must be projected from the origin to the surface of the sphere to create the final tessellation vertex we need. This is a simple matter of normalizing—creating a unit vector length—for the point. Appendix C explains this simple procedure. Figure 8.18(b) shows the tessellation vertices that result when all the centroids of the window triangles are found and projected to the sphere's surface. The original great circle arcs are shown in red, the final grid in blue-green. This figure highlights intersections that were used as examples in previous figures.

Figure 8.19 shows the completed PPT with the surface and panel treatments we have been using. The results are highly uniform, and are the best yet. The final grid resembles Equal-chords to some extent. The largest and smallest triangles are in the same positions as Equal-chords. The mid-PPT face triangle is also equilateral, and just as in Equal-chords, triangle areas increase as you move from the PPT apex to mid-edge, and from mid-edge to midface. The main differences between the Equal-chords and Equal-arcs (three great circles) schema are chord lengths and face angles. We compare the two schemas in the next chapter.

Figure 8.20 shows a complete sphere based on the PPT just created. Notice that its part counts are the same as the other schemas. The subdivision method did not change this fact,

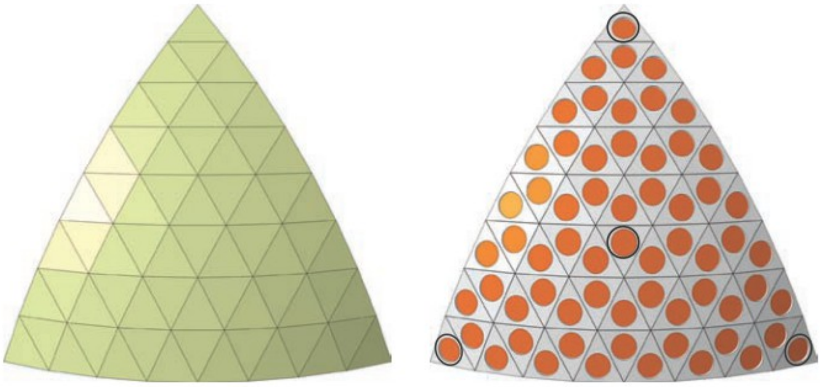

Figure 8.19. Equal-arcs (three great circles) tessellated PPT.

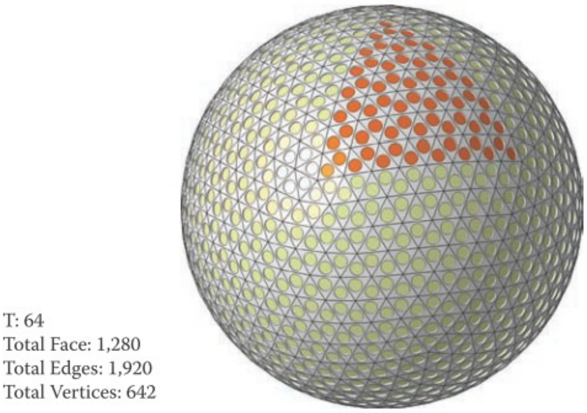

Figure 8.20. Equal-arcs (three great circles) tessellated sphere.

since the schemas all have the same *b*,*c* specification. Certainly, Equal-arcs (three great circles) involves more computation than either of the previous two schemas, but the results are worthwhile. Analysis will show that it has the best distribution of points and the least variation in chord lengths, angles, and face angles of any Class I schema.

8.5.8 Mid-Arcs

The earliest work with Mid-arcs was done by Jeffrey Lindsay in the early 1950s. At that time, the technique was referred to as the alternate method. We call the technique Mid-arcs because it is more descriptive of how the technique works. Deltahedra spherical faces are all equilateral. Mid-arcs takes advantage of this fact and subdivides a PPT by creating one mid-arc reference point on each of its edges. Each pair of mid-arc points defines a great

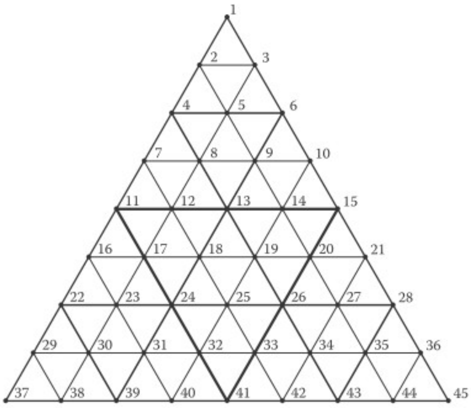
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45

Figure 8.21. Mid-arcs vertex grid reference numbers.

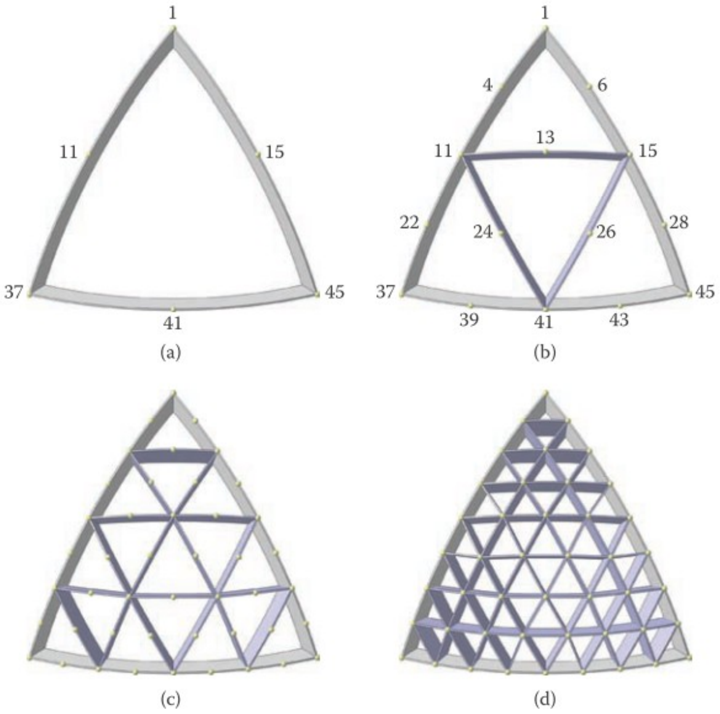
1
11
15
37
45
41
1
4
6
11
13
15
22
24
26
28
37
39
41
43
45

(a)

(b)

(c)

(d)

Figure 8.22. Mid-arcs tessellation schema.

circle arc across the triangle and these three arcs subdivide the triangle into four smaller ones. There is no need to intersect great circles. We use the same method of defining four new triangles within an existing one until we have subdivided the entire PPT. Each time we subdivide all the triangles in a PPT this way, we double the PPT's frequency. Mid-arcs is very easy to understand and to calculate. It is popular with cartographers and astronomers[15] who have devised ways to reference individual triangles on the sphere by taking advantage of the schema's hierarchical nesting of triangles within triangles.[16] Often, they subdivide a spherical octahedron rather than an icosahedron because the octahedron results in natural great circles that can be aligned with the earth's equator and Greenwich Meridian.

Figure 8.21 shows how Mid-arcs work. The pattern of triangles within triangles is evident. Each successive layer of triangles is based on the mid-arc points of an outer triangle. This recursive nesting of triangles within triangles is the essence of Mid-arcs.[17]

Figure 8.22 applies the schema to the same icosahedral PPT we have been using all along. We number the vertices from top to bottom, left to right so that Figure 8.21 and Figure 8.22 can be compared. As usual, subdivision begins at the PPT edges, where each has a single mid-arc reference point.[18] In Figure 21(a), they are numbered 11, 15, and 41. These points can be found by projecting the mid-chord point between 1 and 37 to the surface of the sphere or finding the middle of the arc between the PPT's apices. Either way, great circle arcs between these points define four new subtriangles (1-11-15, 11-37-41, 11-41-15, and 15-41-45), each of which has mid-arc points (4, 6, 13, 22, 24, 26, 28, 39, and 43); see Figure 8.22(b). These new mid-arc points define 12 more edges and 30 new vertex points when all edges are considered (c). This process repeats itself once again until a total of 42 tessellation vertices and 108 edges are defined. Figure 8.22(d) shows the complete tessellation. Keep in mind that all of these vertices and edges were derived from the apices of

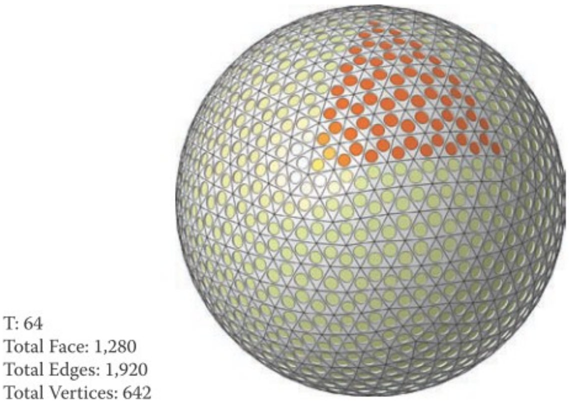

Figure 8.23. Mid-arcs tessellated sphere.

---

[15] (Dutton 1996, 1991) and (Szalay et al. 2005).

[16] (Fekete 1990)

[17] Cartographers sometimes use quaternary subdivision meshing schemas that hierarchically subdivide equilateral triangles similar to Mid-arcs (Dutton 1996). See also Section 3.3.

[18] The edge midpoints of the icosahedron (or of the dodecahedron) are the vertices of the icosidodecahedron (20 triangular faces and 12 pentagonal faces). See Chapter 6 for examples.

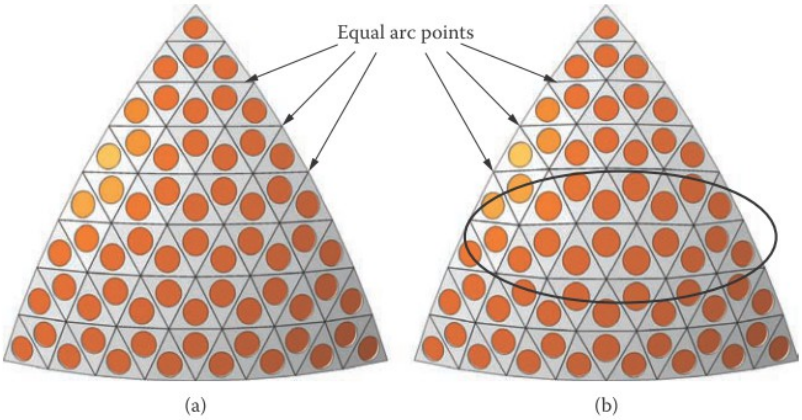

Figure 8.24. Equal-arcs versus Mid-arcs PPT tessellations.

the original PPT and the mid-arc reference point along each edge. Calculations have to be precise; otherwise, errors will accumulate. Our example stops at 8^v, but we could keep on subdividing each triangle indefinitely, doubling the frequency each time.

The Mid-arcs schema is straightforward and easy to calculate. However, there are some limitations. Unlike the other Class I schemas, Mid-arcs can only tessellate frequencies that are powers of 2 (i.e. 2^v, 4^v, 8^v, 16^v, 32^v, etc.). Any even or odd frequency is possible with the other Class I schemas, but only even ones are possible with Mid-arcs. The second drawback concerns the tessellation faces themselves. Despite the fact that we start with an equilateral PPT face and place reference points only in the middle of each of its edges, the vast majority of subdivision triangles created in the final tessellation are not equilateral. Only those faces whose centroids are tangent to the PPT's own centroid are equilateral; all others are slightly isosceles or acute. In this example, only triangles 1-37-45, 11-15-41, 13-24-26, and 18-25-19 are equilateral.

A few numbers illustrate this last point. The icosahedron's PPT edges are each 63.43495º. The first new mid-arc triangle (11-15-41) is also equilateral with 36º edges (its centroid is tangent to the PPT's centroid), but the other three triangles created (1-11-15, 11-27-41, and 15-41-45) are slightly isosceles. Two of their edges are 31.71747º, and the other edge shared with the mid-PPT triangle is 36º. Although other Class I schemas have this same characteristic, Mid-arcs produce more variation in chord length and triangle area than others do. We will see this later in the next chapter. Figure 8.23 shows a full sphere subdivided by the Mid-arcs technique along with its part counts. The PPT we subdivided is highlighted.

Figure 8.24 illustrates another disadvantage of Mid-arcs tessellation (left) by comparing its grid side by side with an Equal-arcs (three great circles) tessellation for the same PPT. Although the PPT edge reference points are the same in both schemas, Mid-arcs faces increase in size more quickly as they approach the center of the PPT, and the chords change length faster as they approach the middle of the PPT. As a result, sequences of chords that cross the PPT are not as smooth in transition as they are with Equal-arcs. An example run of chords showing this difference is circled. Figure 8.25 makes the same comparison, but

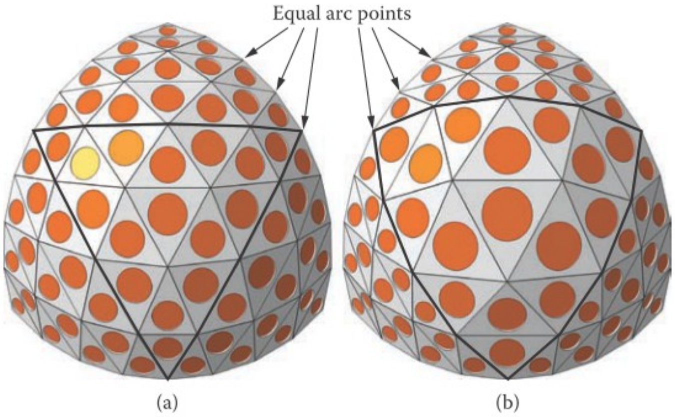

Figure 8.25. Tetrahedral Equal-arcs versus Mid-arcs PPT tessellations.

uses tetrahedral PPTs instead. Tetrahedron PPTs are four times larger than an icosahedron's face (not to scale in the figure), and this makes it easier to see the differences between the two schemas. The first-level subdivision of the PPT is highlighted in each illustration, and the differences are pronounced. Mid-arcs central PPT face is noticeably larger and the faces near the PPT apices are smaller than they are in Equal-arcs tessellations. In Figure 8.25(b) the tetrahedron's central equilateral triangle (highlighted) has three times as much area as the ones at each apex. The tetrahedron's equilateral faces have significantly more spherical excess than the icosahedron's and highlight these differences.

Mid-arcs is the simplest subdivision technique presented thus far. The fact that frequencies are always powers of 2 and the repetitive definition of triangles within triangles offers a natural hierarchy that can be exploited to reference individual triangles on the sphere.[19] Mid-arcs also demonstrates a basic spherical principle—there are no similar triangles on a sphere. You cannot subdivide a spherical equilateral triangle around its mid-edge points and define four other equilaterals as you can in plane geometry.

8.5.9 Subdividing Other Deltahedra

We have developed four different Class I schemas in this chapter. Although each example subdivides a spherical icosahedron, the same methodology applies to tetrahedra and octahedra. The only difference is the initial PPT's characteristics (vertex angle, edge arc length, and vertex coordinates) used to start the process. Use the values in Table 8.2 to subdivide the others.

Figure 8.26 shows examples of 8*v* equal-arc schemas for the tetrahedron and octahedron (geodesic notation $\{3,3+\}_{0,8}$ and $\{3,4+\}_{0,8}$, respectively). They are shown in a partially disassembled state and with mixed surface treatments. Both of these polyhedra have larger PPT areas than the icosahedron for the same radius sphere. As a result, their tessellation triangles show much more variation in tessellation triangle area and shape, as the incenter circles clearly show. The tetrahedron shows the greatest variation of the two. These examples clearly demonstrate how important the choice of foundation polyhedra is when the

---

[19] (Fekete 1990)

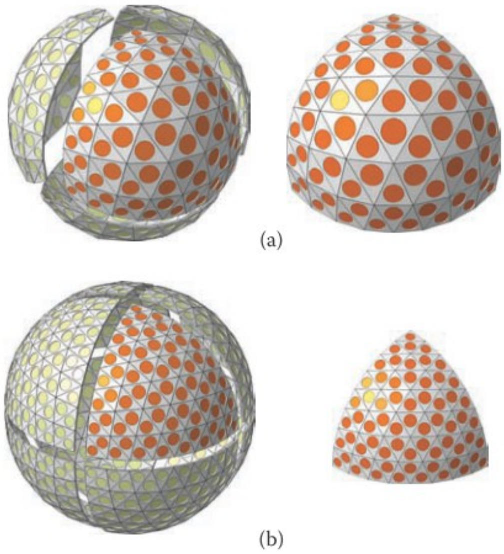

(a)

(b)

Figure 8.26. Equal-arc (three great circles) tessellations of tetrahedra and octahedra.

goal is to create the most uniform subdivision possible. It is no wonder the icosahedron is the most frequently used polyhedral base.

8.5.10 Summary

Class I schemas were among the first spherical polyhedra subdivision methods developed. They apply to any deltahedra. Except for Mid-arcs, which is limited to frequencies that are powers of 2, any even or odd frequency is possible and, depending on prime factors, may contain subgrids at harmonic frequencies. Class I subdivisions maintain the polyhedron's edge as part of the final tessellation grid, though the edge is segmented. The geodesic notation for Class I is distinctive: either the *b* or *c* coefficient is always 0.

Class I tessellations begin with Equal-chords or Equal-arcs reference points along the PPT's edges. Pairs of points (and the origin) define intersecting great circles that are approximately parallel to the PPT's edges. Grid points result where their line of intersection pierces the sphere's surface:

- *Equal-chords*. Based on equal-chord reference points, great circles parallel to the PPT's edges are intersected and their intersection points define points on the tessellation grid within the PPT's surface. Two intersect at the same line as three.

- *Equal-arcs (two great circles)*. Based on equal-arc reference points, two great circles are intersected. They intersect at a single line and define individual tessellation points on the PPT's surface. This procedure is computationally simple, but the resulting tessellation vertices are not rotationally symmetric and only mirror symmetric around one of the PPT's altitude axis.

- *Analysis, visualization, and data extractions*. Demonstrates what can be done with the completed sphere.

- *Summary*. Recaps the highlights of the design example.

Each of our design examples start with a reference model that is used to define the spherical design's theoretical framework (points, lines, planes, and surfaces). The reference model does not contain any geometric information regarding a real connector, panel, or strut. For example, the reference model might contain a line indicating where the center line of a pipe will be, but in the final design, it does not contain the geometry of the pipe itself. The reference model might also contain planes. They might indicate the limits of a surface or pad in the final design.

In the next step, the reference model is used to build up the actual spherical design parts, such as struts, connectors, surfaces, or panels. For example, in this step, a pipe of a particular diameter, length, and wall thickness is created using the center line in the reference model. Most of the time, spherical designers develop parts for only a single LCD or PPT. When the final design of this small area is complete, it is instantiated many times. When a part is instantiated, it is not copied. Instead, a reference to the part's original geometry is created and new location information is added, which describes how the original part (the one referred to) is to be transformed (moved, rotated, scaled, etc.) to position it somewhere else. Instancing allows us to cover the entire sphere with a design based on just one LCD or PPT. Each instance of the small part is positioned side by side and aligned properly on the sphere by applying constraints between points and edges of adjacent parts. A group of constrained parts defines an assembly (of LCD or PPT parts). Due to the symmetry in most spherical designs, the assembly itself can be instanced many times and constraints applied to make still larger assemblies. For instance, there are 120 LCDs in a spherical icosahedron. A single spherical design for one LCD could be instanced six times to make a single assembly for one icosahedral PPT. This assembly, in turn, could be instanced 20 more times. Constraints between each instance positions them to cover the entire sphere. This scenario is only one possibility. We will demonstrate some designs where there are several levels of assemblies and subassemblies.

In a sense, what we have just described is using CAD such as a digital brick layer. We design one brick—a very precise and intricate one—and then lay instances of this brick (references to the original brick, plus geometric transformation information describing its new location) side by side to make a wall. Our bricks and wall, however, are spherical triangles, hexagons, and hyperbolic surfaces. Mathematical constraints between assemblies act as mortar. The reference methodology we use here is applicable to any spherical polyhedron, subdivision method, design layout, or frequency. It exploits the high degree of symmetry and repetition found in spherical designs. Our method begins with reference models.

10.4.1 Start with a Few Good Points

Reference models greatly facilitate spherical design. Every example in this chapter uses one. The first three examples reuse the same reference model, while the last two use a different one. A reference model is made with CATIA and it contains only points, lines, surfaces, and planes. There are no physical design elements in the reference model; that is, it does not contain elements such as connectors, panels, or struts.

The reference model is highly precise, and its layout is carefully chosen to include only the minimum number of elements that will define the reference geometry needed

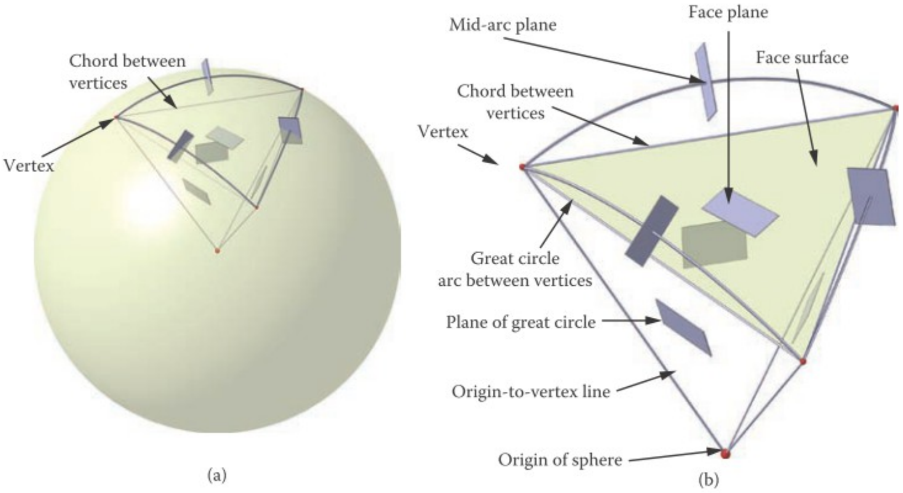

(a) (b)

Figure 10.16. Simple spherical reference model.

to build our parts. Reference models might seem an unnecessary step, but they are well worth the effort. If they are accurate, every spherical design, based on them, will be accurate as well.

Reference models can be reused repeatedly to make a surprising variety of spherical designs. Our first three examples all use the same reference model, but their physical designs could not be more different. Figure 10.16 shows an overall (a) and detailed (b) view of a very simple spherical triangle reference model. None of this reference geometry is visible, but we have used small spheres, cylinders, pads, and coloring to indicate where points, lines, planes, and surfaces are located.

To build this model, we calculated four precise points. Three points lay on the surface of the sphere and define the triangle's vertices. They have the exact angular relationship to one another we want, and they are perfectly tangent to the surface of the sphere. The forth point defines the sphere's origin. The points used in this figure are the vertices of a standard spherical icosahedron face; thus, this model also describes a spherical equilateral triangle.

Using just these four points, CATIA part design and surfacing tools will define everything else. A line tool defines vertex-to-vertex lines (the triangle's edges or chords) and vertex-to-origin reference lines. A plane tool defines planes at the triangle's face and at the three great circles, which make the sides of the spherical triangle. Three more planes are defined normal to both the triangle's mid-chord and mid-arc sides. A circle tool defines great circle arcs on the planes between vertex points.

The reference model is now complete and includes one spherical surface (only shown in Figure 10.16(a)), three points on the spherical surface (which define the icosahedron's face triangle), one origin point, three vertex-to-vertex chord lines, three vertex-to-origin lines, one triangle face plane, one triangular surface, three great circle planes, and three mid-chord/arc normal planes. CATIA defined all of these elements from the initial four input points.

Most of the time, the number of starter points is quite modest, and you can calculate them with a scientific calculator, using the equations provided in previous chapters. In other cases, you may find it easier to use the CAD program itself to calculate the points. We use both techniques in this chapter. Regardless of the input method you use, you may be surprised at how few points are actually needed to build the reference model of a complex spherical design.

The spherical designs in this chapter are based on equal-arc (thee great circles),triacon, and skew subdivisions. A triacon reference model of a single-edge diamond, or even half-diamond, is usually enough to describe a design covering the entire sphere. For equal-arc subdivisions, a single LCD or PPT is used.

Most CAD systems have reflection or mirroring tools. Any geometry can be copied and reflected around a specified plane. Thus, the reference geometry for a left-handed LCD can be reflected to make a right-handed LCD with minimum work. If the design was completely symmetrical around one great circle, the design model for one hemisphere can be reflected to complete the other hemisphere. Reflection is just one of many labor-saving CAD tools.

Some subdivisions result in enantiomorphic shapes. For example, triangular, pentagonal, or hexagonal shapes that do not have equal-length sides often reoccur in other parts of the subdivision as left- or right-handed versions. Copy-reflecting can solve most of these cases. The Ford hex-triangle example, developed later in this chapter, demonstrates enantiomorphic conditions, which are easily solved by mirroring the proper reference model geometry.

10.4.2 Spherical Assembly

The octet truss example introduced constraints and their role in assembling 3D models. In that example, individual octet truss connectors were assembled, along with their rivet fasteners, into a large, nine-way connector.

Spherical assembly makes use of the same techniques, but because the number of parts is so much bigger, even greater productivity gains are possible with careful assembly techniques. A complex and detailed section of the sphere can automatically define every other location on the entire sphere. This is usually accomplished by assembling basic design units into a larger unit and then assembling those units into an even larger assembly.

There are many benefits to using assemblies. Any change to the basic design unit instantly propagates wherever the unit is instanced. This benefit alone makes spherical assembly worth using. There are no restrictions. Automatic drawings can be made, analysis performed, and all the usual visualizations and space analysis can be performed.

Spherical assemblies result in surprisingly small CAD models. This is because they contain just the geometry of the basic design unit, its reference model (which could be deleted after the design unit is created), and the mathematical constraints, which define how the design units are instanced and constrained in larger assemblies. In other words, the design model is not copied repeatedly; it is instanced. This is a key concept, because without it, CAD geometric models would be huge and keeping them consistent would be a major task. There is only one copy of the basic design part, plus some administrative overhead (mostly constraints and visualization options, such as colors or materials), which apply to each instance. The CAD user will certainly notice when spherical assemblies are used; system response does not degrade much, even as he or she works with increasingly large frequencies and more complex part detail. Any spherist anticipating using CAD would be well advised to understand spherical assemblies and to use them wherever there is a high degree of symmetry in the application.

10.5 Three Class II Triacon Designs

Next we will develop three Class II triacon designs. All of them are based on icosahedral 4^v triacon breakdowns and are easily visualized. We will make one comprehensive reference model and reuse it to make very different spherical designs.

The standard unit of symmetry in triacon breakdowns is the polyhedral edge-symmetric diamond. Two of the diamond's vertices are tangent to adjacent polyhedra vertices. The other two diamond vertices are tangent to the polyhedron's face midpoint (see Section 8.6). Figure 10.17 (a) shows a 4^v triacon edge-zenith diamond in the context of a complete spherical icosahedron reference model. Again, the small planes, arcs, and colored surfaces, shown in the figure, simply serve to make these reference elements visible. Only one diamond reference diamond model is needed because the 29 others around the sphere (one diamond straddles each of the icosahedron's 30 edges) are identical except for their absolute vertex coordinate values. Thus, we can create a single design unit based on this one diamond and then copy it to all the other positions on the sphere. Figure 10.17 (b) is a close-up view of just the edge-zenith triacon reference model. The surface of the sphere is shown in gold, spotlighting the diamond's key elements. The *xyz*-axis is shown again to help associate illustrations (a) and (b).

This model contains the same reference elements as the spherical triangle example, shown in Figure 10.16, except that there are more of them: 9 spherical surface points, 1 origin point, 8 triangular surfaces, 16 vertex-to-vertex edges and arcs, and 16 great circle planes (one of every vertex-to-vertex arc). Small red spheres indicate the two diamond points that are tangent to the icosahedron's vertices. Notice that the triacon diamond is symmetrical around both the *zx*- and *zy*-planes. This means you could start building this reference model with as few as five or six points and use the CAD system to complete the

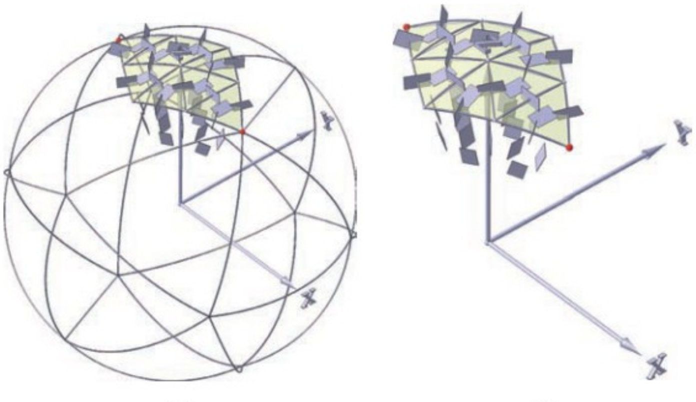

(a) (b)

Figure 10.17. Icosahedron 4^v triacon reference model.

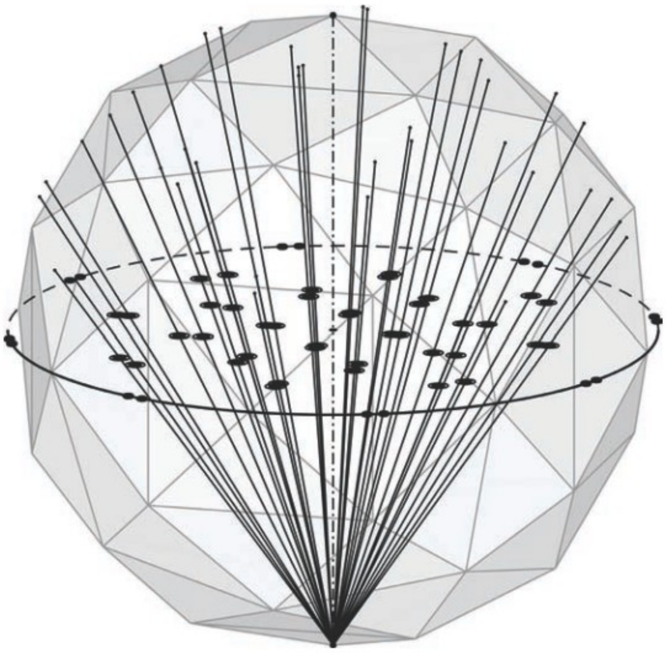

B

Stereographic Projection

In many spherical designs, it is helpful to visualize points and great and lesser circles in two dimensions or as a flat drawing. As the illustrations in this book demonstrate, there are many ways to project spherical geometry onto a 2D surface. However, there is one projection method in particular that offers both visualization and direct analysis capabilities. *Stereographic projection* is a way to show the 3D relationship of points and arcs on a sphere in a 2D diagram called a stereogram. Geodesic subdivisions distribute many points and arcs on the surface of a sphere, and stereographic projection is the best way to represent them in a 2D way that preserves angular relationships.

Stereograms are like maps or graphs of the points and arcs on a sphere. We use stereographics to visualize how well a particular subdivision schema performs and to make comparisons among schemas. With experience, stereograms are easy to read. They immediately show symmetry, if it exists; they also show whether points are well distributed or if the subdivision faces are oriented uniformly. Often, you can identify the subdivision schema and the base spherical polyhedra being used simply by inspecting the stereogram.

Stereographics were studied by ancient Greek mathematicians and used to solve spherical trigonometry problems. Later, they were used to make astronomical instruments, such as the astrolabe. Today, the technique is used by cartographers, who make maps of the earth or maps from satellite images of moons and planets. Geologists and civil engineers use stereographics to analyze large-scale rock formations, while crystallographers use the technique to study small crystals and their chemical composition.

B.1 Points on a Sphere

Stereographic projections depict points and arcs on the surface of the sphere. Points on the surface of the sphere are positions in 3D space, and we identify each of these points with the vector that points from the center of the sphere to this point. A vector has direction and magnitude. A line established by the point and the origin of the sphere define the direction. The magnitude is the length or distance between the point and the origin. For almost every case example in this book, the magnitude is one unit because we choose to work with a generalized sphere of one radius unit. That unit can be one millimeter or one mile; it does not matter.

In the geodesic applications developed here, points on the sphere can be the vertices of triangular grids or vectors that are normal to each face. Face normals are very useful vectors because their direction passes through the origin and is perpendicular to the face, or more precisely, to the plane of the face. The face normal has a magnitude as well. This magnitude is directly proportional to the distance of the face from the origin. However, we can project (or scale) the normal point and define a new point that *does* lie on the surface of the unit sphere. The new point is called a unit normal and maintains the original direction of the normal; thus, it remains on a line from the origin and remains perpendicular to the plane of the face. The stereographic projection of either geodesic tessellation vertices, or the unit normals of their faces, is the best way of representing the inter-relationship of their directions and orientations on a plane.

B.2 Stereographic Properties

Many techniques exist for projecting shapes on a sphere onto 2D surfaces. Cartographers have invented scores. But unlike other projection methods, stereographic projection is particularly suitable when analyzing and comparing spherical subdivisions.

First, stereographic projections are conformal. That is, the stereogram preserves the angular relationship between points or intersecting arcs on the surface of the sphere. If lines of latitude and longitude, which meet at right angles on a globe, are projected, they will meet at right angles on the stereoprojection. This also means that the angles between points or between crossing arcs can be directly measured on the stereogram. *Conformality* is an important property because it preserves the general shape of projected objects.

Second, circles on the sphere project as circles on the projection plane. If a great circle passes through the center of a projection, it projects as a straight line (a segment of a circle with an infinite radius). Other spherical projections can display angular relationships between poles on the sphere, but the stereographic projection is the only one that projects any circle on the surface of the sphere into a circle on the projection.[1] Geodesics distribute many great circle arcs and spherical caps, lesser circles around points on the sphere, and are often used to analyze how uniformly points are distributed. All of them project as circles on the stereogram.

---

[1] Some circles project as straight lines or circles with an infinite radius.

Projections can be made from either the north or south pole of the sphere, and the projection plane, called the primitive, can be the equator (actually any great circle) or tangent to the sphere at either pole (or any plane passing through the sphere). Each combination changes the appearance of the stereogram, but the information it contains and how it is used remain the same. These choices make stereographic projection flexible, although some combinations are more useful to geodesic work than others. We discuss some of these choices in this section, and we adopt some standard uses that we use to analyze subdivisions.

B.3 A History of Diverse Uses

The stereographic projection has its origins no later than the third century BC. Appollonius of Perga included theorems with direct application to proving circle preservation in his *Conics*, which dates from the late third century BC. Applications of the stereographic projection are certain to have been made by the Greeks no later than about 225 BC in the form of the anaphoric clock, which has a disk with stereographically projected stars rotating behind a grid including a stereographically projected horizon and tropics. Hipparchus certainly used the stereographic projection to make star maps and is incorrectly (in my opinion) credited by Ptolemy as being the originator of the projection method. Ptolemy‘s *Planisphaerium* demonstrates that a solid theoretical foundation for the projection was in place before 150 AD.

The Greeks called the process of representing a sphere on a plane “unfolding the sphere.” The stereographic projection was named by François d’Aguilon (Franciscus Aguilonius) in *Opticorum libri sex* (Antwerp, 1613) from στερεοζ (solid). Hence, stereographic means “drawing solids (on a plane).” It is clear that the Greeks did not prove the conformal properties of the projection, but may have assumed it. It is equally clear that the Greeks did know and use the preservation of circles property. Preservation of angles (conformality) under the stereographic projection was first proved by Thomas Harriot (1560–1621).

Islamic astronomers used the projection and proved its basic properties no later than the ninth century. A notable entirely European contribution was *De plana spera*, an early thirteenth-century work by Jordanus de Nemore that presented the theoretical foundation for the stereographic projection. The stereographic projection was applied to astrolabes, quadrants, and other astronomical instruments in both Islamic countries and in Christian Europe in medieval times. Claudius Ptolemy (ca. 150 AD) wrote extensively on stereographics in his work entitled *Planisphaerium*, and in 1613, François d’Aguilon gave the technique its name, “stereographic.”

B.4 The Astrolabe

Stereoprojection is a natural choice for star charts and reference grids for familiar objects or patterns in the night sky. Stereographics was the fundamental projection in the most important astronomy instrument ever devised: the astrolabe. An astrolabe is used to solve problems related to time and the positions of the sun and stars. The first astrolabes date to before 400 AD, and they remained very popular in Europe and the Islamic world into the seventeenth century. No other instrument enjoyed such a long life. The earliest astrolabes

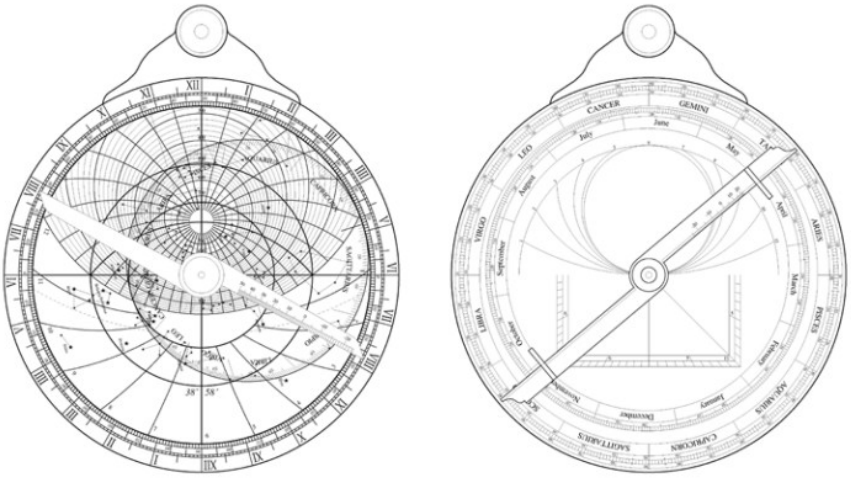

Figure B.1. Astrolabe front and back.

that evolved in Greek antiquity were refined in the Islamic world and were virtually perfected by the tenth century AD. They were introduced into Europe in the eleventh century from Moslem Spain (al-Andalus). Figure B.1 shows the front and back sides of a classic astrolabe.[2]

The earliest known astrolabes, brass disks 6 to 12 inches in diameter, were made in Alexandria. Muslims first encountered the astrolabe in the eighth century in Harran, currently in Turkey near the Syrian border. Even the earliest designs were highly sophisticated instruments. Every part was carefully designed and engraved with scales for specific functions. The main body, the *mater* (Latin for "mother"), was a round plate with a large, hollow circular pocket that accommodated a set of thin, circular disks called *climates*, *plates*, or *tympans* (Latin for "plate"). On the climate disk was engraved a stereoprojection celestial grid of the sky above the observer's local horizon, at the latitude from which the user would make his observations. The center of the climate disk corresponds to the north celestial pole. Well equipped astrolabes would also include a set of climate plates, stored in the mater, each engraved with a sky reference grid for a latitude where the instrument might be used.

Centered over the climate disk is a pierced disk called the *rete* (Latin for "net") which is free to rotate over the climate disk to simulate the daily motion of the stars. It is a frame with openings so the climate disk underneath can be seen. Pointers mark the positions of bright stars. The location of the sun is shown by the projection of the ecliptic as an offset circle on the rete. Scales on the back side of the astrolabe are used to determine the position of the sun on the ecliptic (i.e., the sun's longitude).

Early astronomers believed the earth to be fixed in the heavens and that the stars revolved around it. Although we now know this is not true, the concept is still effectively used today by ocean voyagers who use sextants to navigate. The star's position over the stereographic reference lines indicates where in the sky (azimuth and altitude) the sun or a star can be found.

---

[2] Astrolabe illustrations courtesy of James Morrison.

Many European astrolabes included a rotating rule over the rete. It served many purposes, acting like the hand of a clock, providing an altitude scale, and aiding in the interpretation of the rete's position over the climate. It also was an index to scales engraved on the outer rim, or *limb*, of the astrolabe.

Progressive design refinements and advanced craftsmanship, plus the widespread dissemination of astrolabes in Europe in the fifteenth and sixteenth centuries, led to instrument refinements and contributed to the astrolabe being applied to surveying, orienteering, and timekeeping along with astronomical uses. Many of these additional features required engraved scales to be added to the back of the instrument. The astrolabe's popularity for casting horoscopes guaranteed its continued use even to the present day, and astrolabes were even available in paper and other affordable forms. The astrolabe was truly the most elegant and widely used scientific instrument ever made. Despite its many design variations, the fundamental use of stereoprojection remained constant.

B.5 Crystallography and Geology

New uses for stereographics have been found in the fields of crystallography, civil engineering, and structural geology. Crystallographers study the atomic structure of solids. Before the use of X-ray diffraction techniques, stereographic techniques were used to graph and analyze the geometry of crystals. Stereograms are made of the orientation of a crystal's planes and the angles between them. In order to make the stereogram, crystal is located inside an imaginary sphere. Points on the sphere are located directly above each face (on an axis normal to the face). These points are then stereoprojected from a pole in the hemisphere opposite the point. Figure B.2 shows the basic arrangement. The result is the dot-like pattern on the primitive, the equatorial plane where points are projected. Each dot is the projection of a point on the sphere directly above a crystal's face. The angles between dots indicate the angles between crystal faces. Other dot patterns indicate symmetrical relationships between vertices, edges, and crystal faces.

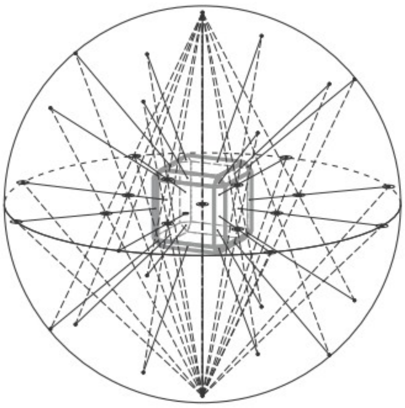

Figure B.2. Stereoprojection in crystallography.

20200917

Claude Shannon 1949

limite teórico de transmissão de dados.

⇒ Book: Divided Spheres —

Pag. 211 (232)

8.5.7 Equal-Area (Three Great Circles)

Pag. 292 (313)

Figure 10.16. Spherical

Pag. 373 (394)

Stereographic Projection

Shanah tovah Umetukah!

(Feliz Ano Novo e Doce)

20200918 (5781)

. ' #

Modelo de combate de Lei de Lanchester Quadrático

1 x 1,5 (superioridade)

—

inadimplida a obrigação

—

e-rara.ch

25/Maio

⇒ 5781

⇒ São Paulo Brasil

Calcumus

Dez/2010 US$ 1,70

Você é escravo daquilo que você deseja.
Indian Proverb.

—

ARQ_BASICS_02\12\
Uncertain Projective Geometry/
Uncertain Projective Geometry.pdf

Pag. 26 (43)

$d(1,0) = |c| / \sqrt{a^2+b^2}$

---

pc_oweto\MHAMY11.pdf

Pag. 196/197 (100)

184. The Equations of a Line Perpendicular to Two Given Straight Lines.

→ nonparallel. straight lines

---

e-rara.ch

Leibniz →

---

pstenio45@gmail.com

---

pdfdrive.com

→ aumento

Tolermonita luminosa (Roma)

Yang Chan Chivé Pedreira.

9:5;9 9 9/ 9 . ? 6 , 0 <
5 < ! /+6 5 #@) 1
A ! +)!
' #-

in 2D and 3D are covered. They are represented using homogeneous vectors, but we will give a Euclidean interpretation for the vectors of every entity. At the end of this section, we will see that the chosen representations are special cases of the Plücker coordinates, a representation of subspaces within a projective space.

2.2.1 Points and Lines in 2D

Given a *2D point* $\boldsymbol{x} = (x, y)^\mathsf{T} \in \mathbb{R}^2$ on the Euclidean plane, one may obtain a homogeneous representation using equation (2.3):

$$\mathbf{x} = \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \cong \begin{pmatrix} wx \\ wy \\ w \end{pmatrix} \cong \begin{pmatrix} u \\ v \\ w \end{pmatrix}$$

see also figure 2.1(d) on page 21.

A Euclidean representation of a 2D point is straight forward by specifying two coordinates, but the Euclidean representation of a *2D line* can be manifold:

slope-intercept form	$y = m\,x + b$
axis-intercept form	$\frac{x}{a} + \frac{y}{b} = 1$
angle-distance form	$x\,cos(\varphi) + y\,sin(\varphi) = d$
homogeneous form	$a\,x + b\,y + c = 0$
determinant form	$\begin{vmatrix} x_1 & x_2 & x \\ y_1 & y_2 & y \\ 1 & 1 & 1 \end{vmatrix} = 0$
point-angle form	$(x - x_0)\,cos(\varphi) + (y - y_0)\,sin(\varphi) = 0$

Some of these representations do not cover all possible cases (e.g. vertical lines for the slope-intercept form) and some require more parameters than necessary: the determinant form requires two points $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ hence four parameters opposed to the fact that only two parameters are necessary, for example angle and distance.

We will start with the homogeneous form $a\,x+b\,y+c = 0$, then the parameters can be written in a homogeneous vector

$$\mathbf{l} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

If we multiply the homogeneous form with a scalar w and use the homogeneous vector $\mathbf{x} = (wx, wy, w)^\mathsf{T} \cong (u, v, w)^\mathsf{T}$ for a point, we obtain

$$a\,u + b\,v + c\,w = \mathbf{l}^\mathsf{T}\mathbf{x} = \mathbf{x}^\mathsf{T}\mathbf{l} = 0 \qquad (2.12)$$

which is the 2D case for equation (2.10) for hyperplanes.

A *line segment* $\mathbf{s}_l$ is a finite segment of an (infinite) line. It can be represented by a triple $(\mathbf{l}_{fs}, \mathbf{x}_{fs}, l)$, the infinite line $\mathbf{l}_{fs}$, a mid-point $\mathbf{x}_{fs}$ of the line segment and a length l. This is equivalent compared to defining two enclosing points on the line, cf. definition 1 on page 6.

2.2.1.1 Homogeneous and Euclidean Parts From the homogeneous vectors for points it is easy to determine the Euclidean distance from the point to the origin $\boldsymbol{O}$: for a projective point $\mathbf{x} = (u, v, w)^\mathsf{T}$, the Euclidean distance to the origin is given by $d(\mathbf{x}, \boldsymbol{O}) = \sqrt{(u/w)^2 + (v/w)^2} = \sqrt{u^2 + v^2}/|w|$.

The coordinates of the 2D line $\mathbf{l} = (a, b, c)^\mathsf{T}$ may be interpreted by comparing equation (2.12) with the angle-distance form: they are essentially the same up to a multiplication factor $\lambda_l = 1/\sqrt{a^2 + b^2}$:

$$\mathbf{l} = \begin{pmatrix} a \\ b \\ c \end{pmatrix} \cong 1/\sqrt{a^2 + b^2} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} cos(\varphi) \\ sin(\varphi) \\ -d \end{pmatrix} \tag{2.13}$$

As the scalar d refers to the distance of the line to the origin, we obtain $d(\mathbf{l}, \boldsymbol{O}) = |c|/\sqrt{a^2 + b^2}$. Note that the angle φ is the angle of the normal of the line $\mathbf{l}$.

Motivated by the above observation of computing the distances to the origin and as suggested by BRAND 1966 p. 51ff., we will divide the homogeneous vectors for 2D points and lines into two parts:

- the *Euclidean part* depends upon the position of the Euclidean origin $\boldsymbol{O}$. This part of the homogeneous vector will be indexed with O and is the enumerator of the fractions $d(\mathbf{x}, \boldsymbol{O})$ and $d(\mathbf{l}, \boldsymbol{O})$ as above.
- the *homogeneous part* is independent of the position of the origin $\boldsymbol{O}$. It refers to the denominator of the fractions $d(\mathbf{x}, \boldsymbol{O})$ and $d(\mathbf{l}, \boldsymbol{O})$.

For a 2D point $\mathbf{x}$ and for a 2D line $\mathbf{l}$ the partitioning yields

$$\mathbf{x} = \begin{pmatrix} u \\ v \\ \hline w \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_O \\ \hline x_h \end{pmatrix} \quad ; \quad \mathbf{l} = \begin{pmatrix} a \\ b \\ \hline c \end{pmatrix} = \begin{pmatrix} \boldsymbol{l}_h \\ \hline l_O \end{pmatrix}$$

2.2.1.2 Normalization to Euclidean Coordinates Dividing the homogeneous vectors by the norm of the homogeneous parts, $\mathbf{x}/|x_h|$ and $\mathbf{l}/|\boldsymbol{l}_h|$, we obtain the Euclidean representation of the 2D entities: the Euclidean coordinates of a 2D point and the angle-distance form of a 2D line. We call this division *Euclidean normalization:* for any homogeneous vector $\mathbf{x}$ representing a geometric entity, the Euclidean normalization $\mathsf{N}_O(\cdot)$ is defined as

$$\mathsf{N}_O(\mathbf{x}) := \mathbf{x}/|\boldsymbol{x}_h| \tag{2.14}$$

where $\boldsymbol{x}_h$ is the homogeneous part of the entity $\mathbf{x}$.

Note that the proposed normalization $\mathsf{N}_O(\cdot)$ is ambiguous with respect to its sign: a Euclidean point $\boldsymbol{x} = (2,3)^\mathsf{T}$ may be represented by $\mathbf{x}_1 = (4,6,2)^\mathsf{T}$ and $\mathrm{x}_2 = (-4,-6,-2)^\mathsf{T}$. This leads to $\mathsf{N}_O(\mathbf{x}_1) = -\mathsf{N}_O(\mathbf{x}_2)$. The same applies to a 2D line with $\varphi = 90°$ and $d = 3$, where two representations $\mathbf{m}_1 = (2,0,6)^\mathsf{T}$ and $\mathbf{m}_2 = (-2,0,-6)^\mathsf{T}$ yield $\mathsf{N}_O(\mathbf{m}_1) = -\mathsf{N}_O(\mathbf{m}_1)$, see figure 2.3.

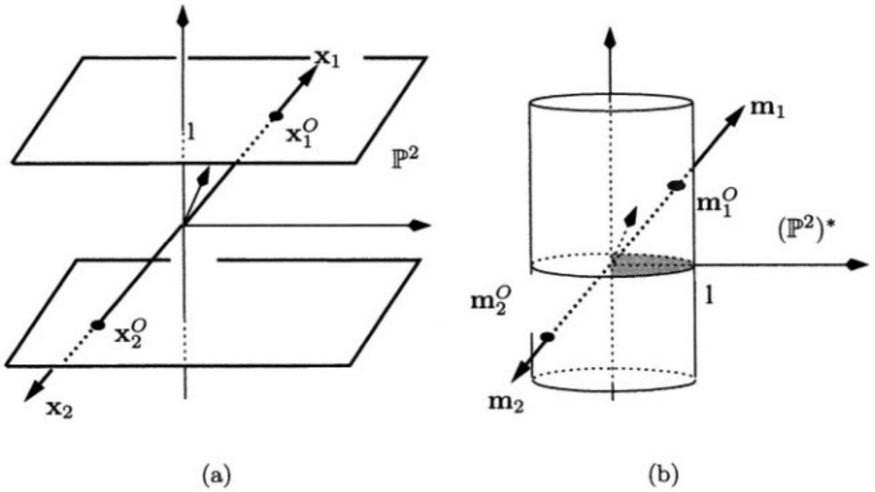

Fig. 2.3. Ambiguous Euclidean normalization of a 2D point $\boldsymbol{x}$ and a 2D line $\boldsymbol{m}$, which can be represented by $\mathbf{x}_1, \mathbf{x}_2$ resp. $\mathbf{m}_1, \mathbf{m}_2$: the normalizations $\mathsf{N}_O(\mathbf{x}_1)$ and $\mathsf{N}_O(\mathbf{x}_2)$ resp. $\mathsf{N}_O(\mathbf{m}_1)$ and $\mathsf{N}_O(\mathbf{m}_2)$ yield different results. We may choose one representation where the homogeneous part is positive oriented, see text for details.

To reach a unique representation, it is possible to demand that the homogeneous part $(\boldsymbol{x}_e)_h$ of any *normalized* homogeneous vector $\mathbf{x}_e = \mathsf{N}_O(\mathbf{x})$ should be *positive oriented:* for points, this means that the homogeneous scale factor x_h is positive, thus for a 2D point $\mathbf{x}$, we skip the norm by $\mathsf{N}_O(\mathbf{x}) = \mathbf{x}/x_h$. For lines, the direction $\boldsymbol{l}_h$ should be on the upper half of the circle S^2, thus $\langle (0,1)^\mathsf{T}, \boldsymbol{l}_h \rangle > 0$, see figure 2.3(b) As for the spherical normalization, the ambiguity may be resolved using oriented projective geometry, STOLFI 1991, but the ambiguity only affects the Euclidean output of the reasoning process: as long as only homogeneous coordinates are used, the sign of the vectors is not important. We will not use Euclidean coordinates directly during the complete reasoning, see chapter 3, thus the ambiguity can be disregarded in our case.

To summarize, we are now able to transfer back and forth between Euclidean and homogeneous representation as in figure 2.2 on page 23. Furthermore, we can easily obtain the distances to the origin:

$$d(\mathbf{x}, \boldsymbol{O}) = \frac{|\boldsymbol{x}_O|}{|x_h|} \quad ; \quad d(\mathbf{l}, \boldsymbol{O}) = \frac{|l_O|}{|\boldsymbol{l}_h|} \tag{2.15}$$

In the next sections, we will define Euclidean and homogeneous parts for 3D entities, too, and we will see that there both parts of the vector play the same role as in the 2D case.

2.2.1.3 Infinite Points and Line As stated earlier, projective geometry allows us to represent entities at infinity. This means that the distance from the origin is not defined in the Euclidean interpretation, thus the formulas in equation (2.15) are not valid anymore. This can only happen when the norm of the enumerator, i.e. the norm of the homogeneous part is zero.

Property 1 *If a homogeneous vector is non-zero, then it represents an entity at infinity if and only if the norm of the homogeneous part is zero.*

Infinite points are of the form $\mathbf{x}_\infty = (u, v, 0)^\mathsf{T}$. Since $\mathbf{x}_\infty$ is a homogeneous vector, this is a one-parameter family of vectors; the parameter is the direction of the infinite point: taking the angle-distance form from equation (2.13), an infinite point $\mathbf{x}_\infty$ lies on a 2D line $\mathbf{l} = \lambda(cos(\varphi), sin(\varphi), -d)$ if and only if the direction angle φ is perpendicular to the directional vector $(u, v)^\mathsf{T}$.
The (unique) 2D line at infinity is of the form $\mathbf{l}_\infty = (0, 0, c)^\mathsf{T}$, and it is obvious that all 2D points $\mathbf{x}_\infty$ at infinity are on the line at infinity.

Table 2.1. Interpretation of the canonical basis $\mathbf{e}_i$ as 2D points and 2D lines.

homogeneous vector	2D point interpretation	2D line interpretation
$\mathbf{e}_1 = (1,0,0)^\mathsf{T}$	infinite point in x-direction	y-axis
$\mathbf{e}_2 = (0,1,0)^\mathsf{T}$	infinite point in y-direction	x-axis
$\mathbf{e}_3 = (0,0,1)^\mathsf{T}$	origin	line at infinity

Canonical Examples As examples for points and lines, we look at the canonical basis vectors $\mathbf{e}_i = (0, \ldots, \underset{i}{1}, \ldots, 0)$ and interpret them as 2D points or lines, see table 2.1. This will be useful when analyzing homogeneous transformations, section 2.3 on page 34 and construction of new entities, section 3.1.1 on page 49.

2.2.2 Points and Planes in 3D

Points in 3D are represented analogously to points in 2D: given a *3D point* $\boldsymbol{X} = (X, Y, Z)^\mathsf{T} \in \mathbb{R}^3$, the homogeneous representation is according to equation (2.3):

$$\mathbf{X} = \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \cong \begin{pmatrix} TX \\ TY \\ TZ \\ T \end{pmatrix} \cong \begin{pmatrix} U \\ V \\ W \\ T \end{pmatrix}$$

A *plane* in 3D can be represented in several ways, similar to 2D line representations, see table 2.2.1 on page 25. Since 3D planes are hyperplanes for $\mathbb{P}^3$ as 2D lines are hyperplanes for $\mathbb{P}^2$, a plane is represented as a homogeneous vector $\mathbf{A} = (A, B, C, D)^\mathsf{T}$, resembling the homogeneous form $AX + BY + CZ + D = 0$ with a Euclidean point $X = (X,Y,Z)$. Again the homogeneous form can be rewritten with a homogeneous point vector **X:**

$$AU + BV + CW + D = \mathbf{A}^\mathsf{T}\mathbf{X} = \mathbf{X}^\mathsf{T}\mathbf{A}$$

Since we deal with polyhedral objects, a *plane segment* or a *planar patch* can be defined by its polygonal outline: a planar patch $\mathbf{P} = (\mathbf{A}_P, \mathbf{l}_{P1}, \ldots, \mathbf{l}_{Pn})$ is a planar region enclosed by n 2D lines $\mathbf{l}_{Pi}, i = 0 \ldots n$ which lie on the plane $\mathbf{A}_P$, cf. definition 1 on page 6.

2.2.2.1 Homogeneous and Euclidean Parts We can transfer the definition from section 2.2.1 and define the homogeneous and Euclidean parts of a point **X** and a plane **A**

$$\mathbf{X} = \begin{pmatrix} U \\ V \\ W \\ \hline T \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_O \\ \hline X_h \end{pmatrix} \quad ; \quad \mathbf{A} = \begin{pmatrix} A \\ B \\ C \\ \hline D \end{pmatrix} = \begin{pmatrix} \boldsymbol{A}_h \\ \hline A_O \end{pmatrix}$$

The Euclidean normalization of **X** is then $\mathsf{N}_O(\mathbf{X}) = \mathbf{X}/X_h$ and $\mathsf{N}_O(\mathbf{A}) = \mathbf{A}/|\boldsymbol{A}_h|$. Similar to the 2D case, the ambiguity in the sign of $\mathbf{A}_e = \mathsf{N}_O(\mathbf{A})$ is resolved by testing whether the normalized direction $(\boldsymbol{A}_e)_h$ is on the upper half of the sphere S^3, thus $\langle (0,0,1)^\mathsf{T}, (\boldsymbol{A}_e)_h \rangle > 0$. The Euclidean distance $d(\cdot, \boldsymbol{O})$ to the origin is

$$d(\mathbf{X}, \boldsymbol{O}) = \frac{|\boldsymbol{X}_O|}{|X_h|} = |\mathsf{N}_O(\mathbf{X})| \quad ; \quad d(\mathbf{A}, \boldsymbol{O}) = \frac{|A_O|}{|\boldsymbol{A}_h|} = |\mathsf{N}_O(\mathbf{A})| \qquad (2.16)$$

2.2.2.2 Ideal Points and Plane and Canonical Examples With property 1 on page 28 it is clear that the infinite point $\mathbf{X}_\infty = (U, V, W, 0)^\mathsf{T}$ lies on the infinite plane $\mathbf{A}_\infty = (0, 0, 0, 1)$. The direction of an infinite point $\mathbf{X}_\infty$ is determined by Euclidean part $(\boldsymbol{X}_\infty)_O$.

The interpretation for the canonical vectors $\mathbf{E}_i, i = 1, \ldots, 4$ are listed in table 2.2, it is a direct generalization of table 2.1 on the preceding page.

Table 2.2. Interpretation of the canonical basis $\mathbf{E}_i$ as 3D points and 3D planes.

homogeneous vector	3D point interpretation	3D line interpretation
$\mathbf{E}_1 = (1,0,0,0)^\mathsf{T}$	infinite point in x-direction	yz-plane
$\mathbf{E}_2 = (0,1,0,0)^\mathsf{T}$	infinite point in y-direction	xz-plane
$\mathbf{E}_3 = (0,0,1,0)^\mathsf{T}$	infinite point in z-direction	xy-plane
$\mathbf{E}_4 = (0,0,0,1)^\mathsf{T}$	origin	plane at infinity

2.2.3 Lines in 3D

A *3D line* is different from the previous entities since it is neither a point nor a hyperplane. It is rather a one-dimensional subspace in 3D which can be represented by two (Euclidean) points $\boldsymbol{X}_0$ and $\boldsymbol{Y}_0$, cf. figure 2.4(a). One needs four parameters to represent a line in 3D, as shown in figure 2.4(b): assume two planes have been chosen in advance such as the coordinate planes on the *XY* respectively *XZ* axes. Then the two intersection points of the 3D line and the planes can be described by two parameters each.

There exists a number of other representations for 3D lines, but in the following we will derive a homogeneous vector representation based on two given points, see figure 2.4(a): a 3D line refers to a homogeneous 6-vector

$$\mathbf{L}(\boldsymbol{X}_0, \boldsymbol{Y}_0) = \begin{pmatrix} \boldsymbol{Y}_0 - \boldsymbol{X}_0 \\ \boldsymbol{X}_0 \times \boldsymbol{Y}_0 \end{pmatrix} = \begin{pmatrix} \boldsymbol{L}_h \\ \boldsymbol{L}_O \end{pmatrix}$$

which consists of two parts:

- the first part $\boldsymbol{L}_h$ contains the direction $\boldsymbol{L}_h = \boldsymbol{Y}_0 - \boldsymbol{X}_0$ of the line $\mathbf{L}$ and has 2 degrees of freedom. Note that this part is the *homogeneous part* of the 3D line vector as it is independent from the origin,
- the two points $\boldsymbol{X}_0, \boldsymbol{Y}_0$ and the origin $\boldsymbol{O}$ define a plane, whose normal is given by the cross-product $\boldsymbol{L}_O = \boldsymbol{X}_0 \times \boldsymbol{Y}_0$. With the direction of $\boldsymbol{L}_O$ we have fixed the line $\mathbf{L}(\boldsymbol{X}_0, \boldsymbol{Y}_0)$ in a specific plane. Fixing the plane adds only one additional independent parameter to the existing two of the first part $\boldsymbol{L}_h$, since $\boldsymbol{L}_O$ and $\boldsymbol{L}_h$ are constrained to be perpendicular to each other,

  $$\boldsymbol{L}_O^\mathsf{T} \boldsymbol{L}_h = 0 \qquad (2.17)$$

  This constraint is called the *Plücker constraint* of a 3D line, see below for more details.

  In order to uniquely fix the 3D line we need the minimal distance $d(\mathbf{L}, \boldsymbol{O})$ of the line to the origin as the fourth parameter of the 3D-line. $d(\mathbf{L}, \boldsymbol{O})$ is equivalent to the altitude h of the triangle $(\boldsymbol{O}, \boldsymbol{X}_0, \boldsymbol{Y}_0)$. Because the norm $|\boldsymbol{L}_O|$ of the cross-product $\boldsymbol{X}_0 \times \boldsymbol{Y}_0$ yields twice the area of the triangle and $|\boldsymbol{L}_h|$ is the length of the base of the triangle, we can write

$$d(\mathbf{L}, \boldsymbol{O}) = \frac{|\boldsymbol{L}_O|}{|\boldsymbol{L}_h|} \tag{2.18}$$

Thus the distance as the fourth parameter the 3D line is implicitly given by $\boldsymbol{L}_O$. The 3-vector $\boldsymbol{L}_O = \boldsymbol{X}_0 \times \boldsymbol{Y}_0$ completely depends on the distance of the origin and thus $\boldsymbol{L}_O$ is the *Euclidean part* of the 6-vector $\mathbf{L} = (\boldsymbol{L}_h, \boldsymbol{L}_O)^\mathsf{T}$. The Euclidean part vanishes, i.e. $\boldsymbol{L}_O = \mathbf{0}$, if and only if the line intersects the origin.

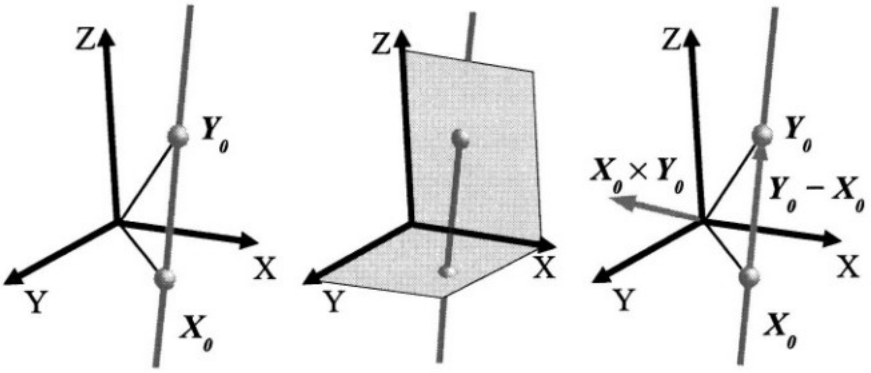

(a) (b) (c)

Fig. 2.4. A 3D line may be determined by two points which intersect with two planes of the coordinate system; the two points have 2 degrees of freedom since they are fixed to planes (left). Two Euclidean points $\boldsymbol{X}_0$ and $\boldsymbol{Y}_0$ define a 3D line (middle). In this work we choose two 3-vectors for the representation of $\mathbf{L}$: the direction of the line $\boldsymbol{L}_h = \boldsymbol{X}_0 - \boldsymbol{Y}_0$ and the 3-vector $\boldsymbol{L}_O = \boldsymbol{X}_0 \times \boldsymbol{Y}_0$. We assume $\boldsymbol{L}_O$ to have the length equal to the distance of the 3D line $\mathbf{L}$ to the origin (right).

To summarize the above analysis, a *3D line* is represented by a homogeneous 6-vector $\mathbf{L} = (L_1, L_2, L_3, L_4, L_5, L_6)^\mathsf{T}$, where the homogeneous part $\boldsymbol{L}_h = (L_1, L_2, L_3)^\mathsf{T}$ is constrained to be orthogonal to the Euclidean part $\boldsymbol{L}_O = (L_4, L_5, L_6)^\mathsf{T}$, i.e. $\boldsymbol{L}_h^\mathsf{T} \boldsymbol{L}_O = 0$. Thus the 6-vector $\mathbf{L}$ has 4 degrees of freedom, considering both the orthogonality and homogeneous constraint. As in 2D, a line segment $\mathbf{S}_L$ is a finite fragment of $\mathbf{L}$ and is represented by a triple $(\mathbf{L}, \mathbf{X}_{fs}, L)$, where $\mathbf{X}_{fs}$ is the mid-point of the line segment and L is the length of the segment.

The Euclidean normalization of a 3D line $\mathsf{N}_O(\mathbf{L}) = \mathbf{L}/|\boldsymbol{L}_h|$ uses the formula already known from points and hyperplanes. The ambiguity in the sign of the homogeneous part $(\boldsymbol{L}_e)_h$ of $\mathbf{L}_e = \mathsf{N}_O(\mathbf{L})$ is resolved by deciding whether the normalized direction $(\boldsymbol{L}_e)_h$ is on the upper half of the sphere S^3: $\langle (0,0,1)^\mathsf{T}, (L_e)_h \rangle > 0$.

Some Examples Again, we look at some canonical examples. A line $\mathbf{L} = (\boldsymbol{L}_h, \boldsymbol{L}_O)^\mathsf{T}$ lies at infinity if the homogeneous part vanishes, $\boldsymbol{L}_h = \mathbf{0}$. Thus a line at infinity is defined by the normal direction $\boldsymbol{L}_O$ and has 2 degrees of freedom. Note that a line at infinity can only be computed using equation(2.19).

Taking the canonical 6-vectors $\mathbf{E}_i$, we obtain the coordinate axis and three distinct lines at infinity, namely the lines lying on the canonical planes of the coordinate system, cf. table 2.3.

Table 2.3. Interpretation of the canonical basis $\mathbf{E}_i$ as 3D lines.

homogeneous 6-vector	3D line interpretation
$\mathbf{E}_1 = (1,0,0,0,0,0)^\mathsf{T}$	x-axis
$\mathbf{E}_2 = (0,1,0,0,0,0)^\mathsf{T}$	y-axis
$\mathbf{E}_3 = (0,0,1,0,0,0)^\mathsf{T}$	z-axis
$\mathbf{E}_4 = (0,0,0,1,0,0)^\mathsf{T}$	line at infinity on yz-plane
$\mathbf{E}_6 = (0,0,0,0,0,1)^\mathsf{T}$	line at infinity on xy-plane
$\mathbf{E}_5 = (0,0,0,0,1,0)^\mathsf{T}$	line at infinity on xz-plane

We summarize the representations of points, lines and planes in 2D and 3D in table 2.4

Table 2.4. Homogeneous representation of points, lines and planes in 2D and 3D. A line in 3D is represented with Plücker coordinates, for which the Plücker constraint $\boldsymbol{L}^\mathsf{T}\boldsymbol{L}_O = 0$ holds.

	2D	3D
point	$\mathbf{x}^\mathsf{T} = (u, v; w) = (\boldsymbol{x}_O^\mathsf{T}, x_h)$	$\mathbf{X}^\mathsf{T} = (U, V, W; T) = (\boldsymbol{X}_O^\mathsf{T}, X_h)$
line	$\mathbf{l}^\mathsf{T} = (a, b; c) = (\boldsymbol{l}_h^\mathsf{T}, l_O)$	$\mathbf{L}^\mathsf{T} = (L_1, L_2, L_3; L_4, L_5, L_6) = (\boldsymbol{L}_h^\mathsf{T}, \boldsymbol{L}_O^\mathsf{T})$
plane	-	$\mathbf{A}^\mathsf{T} = (A, B, C; D) = (\boldsymbol{A}_h^\mathsf{T}, A_O)$

2.2.4 Plücker Coordinates

The previous representation of the 3D line is compatible with the representation using the so-called Plücker coordinates. The geometrician Julius Plücker[2] introduced a representation of k-dimensional subspaces within an

---

[2] Julius Plücker, born 1801 in Elberfeld, Germany, died 22 May 1868 in Bonn, Germany. Professor of mathematics at Halle in 1834, at Bonn since 1836.

n-**dimensional** projective space based on $\binom{n+1}{k+1}$ minors of a $(n+1) \times (k+1)$ matrix with given coordinates, cf. FAUGERAS AND LUONG 2001.

An example for Plücker coordinates was already introduced in section 2.1.2, where three points $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ form a 2-dimensional subspace, a hyperplane. From the 4×3 matrix $\mathsf{M}_A = (\mathbf{X}\,\mathbf{Y}\,\mathbf{Z})$ one obtains 4 minors or 4 Plücker coordinates.

Assume now a 3D line $\mathbf{L}$ which was formed by two points $\mathbf{X}$, $\mathbf{Y}$. It is a one-dimensional subspace with $\mathbb{P}^3$ and there are $\binom{n+1}{k+1} = 6$ coordinates, which are derived from 2×2 minors l_{ij} of a 4×2 matrix $\mathsf{M}_L = (\mathbf{X}, \mathbf{Y})$. This matrix consists of two homogeneous vectors of 3D points $\mathbf{X}, \mathbf{Y}$ - we can also use 3D planes, as we will see in section 2.5 on page 40. For M_L we obtain

$$\mathsf{M}_L = \begin{pmatrix} X_1 & Y_1 \\ X_2 & Y_2 \\ X_3 & Y_3 \\ X_4 & Y_4 \end{pmatrix} ; \quad \begin{matrix} l_{12} = \begin{vmatrix} X_1 & Y_1 \\ X_2 & Y_2 \end{vmatrix} & l_{23} = \begin{vmatrix} X_2 & Y_2 \\ X_3 & Y_3 \end{vmatrix} & l_{34} = \begin{vmatrix} X_3 & Y_3 \\ X_4 & Y_4 \end{vmatrix} \\ l_{14} = \begin{vmatrix} X_1 & Y_1 \\ X_4 & Y_4 \end{vmatrix} & l_{24} = \begin{vmatrix} X_2 & Y_2 \\ X_4 & Y_4 \end{vmatrix} & l_{13} = \begin{vmatrix} X_1 & Y_1 \\ X_3 & Y_3 \end{vmatrix} \end{matrix} \tag{2.19}$$

If we assume the homogeneous part of $\mathbf{X}, \mathbf{Y}$ to be fixed as $X_4 = Y_4 = 1$, it is easy to see that the Plücker coordinates l_{ij} directly relate to our 6-vector $\mathbf{L} = (\boldsymbol{L}_h, \boldsymbol{L}_O)^\mathsf{T}$: $\boldsymbol{L}_h = -(l_{14}, l_{24}, l_{34})^\mathsf{T} = (l_{41}, l_{42}, l_{43})^\mathsf{T}$ and $\boldsymbol{L}_O = (l_{23}, -l_{13}, l_{12})^\mathsf{T} = (l_{23}, l_{31}, l_{12})^\mathsf{T}$. Scaling $\mathbf{X}, \mathbf{Y}$ with homogeneous factor λ resp. μ scales the Plücker coordinates by $\lambda \cdot \mu$.

Note that the Plücker coordinates l_{ij} have been expressed according to a particular order of the index pair i, j. We are free though to choose other index pairs, that means any other combination of 2×2 minors. For our choice of $\mathbf{L} = (\boldsymbol{L}_h, \boldsymbol{L}_O)$ we have $(i, j) \in \{(4,1), (4,2), (4,3), (2,3), (3,1), (1,2)\}$. This has only the effect on the ordering and the signs of the Plücker coordinates within a vector, but makes it possible to distinguish clearly between a homogeneous and a Euclidean part. Note that other authors choose different Plücker coordinates for lines, cf. HARTLEY AND ZISSERMAN 2000, STOLFI 1991.

Plücker Constraint In general, the Plücker coordinates are not independent from each other, in fact there is a set of quadratic relations between them. For the cases in 2D and 3D space, a quadratic relation of Plücker coordinates only occurs for line in 3D, we obtain only one condition:

$$l_{41}l_{23} + l_{42}l_{31} + l_{43}l_{12} = 0 \tag{2.20}$$

All 6-vectors $\mathbf{L} = (l_{41}, l_{42}, l_{43}, l_{23}, l_{31}, l_{12})$ satisfying equation (2.20) correspond to valid 3D lines. As mentioned above the Plücker constraint (or Plücker condition) is obviously satisfied with the proposed construction as $\boldsymbol{L}_h^\mathsf{T} \boldsymbol{L}_O = 0$. Note that the constraint (2.20) defines a manifold in $\mathbb{R}^6$, containing all valid 3D-lines. If the vector $\mathbf{L}$ is known to be spherically normalized, the constraint defines a manifold on the sphere S^5.

Thus the vector notation for lines appears to have a disadvantage compared to the vector representation for points and planes: in addition to the homogeneous factor it has another degree of freedom which is constraint by the Plücker condition. This situation can be compared to the fundamental matrix, see section 2.3.3 on page 36, which is a homogeneous 3 × 3 matrix with 7 degrees of freedom. In chapter 3 and 4 we will see that it is easy to cope with the over-parameterization in the context of geometric reasoning.

2.3 Basic Geometric Transformations

After introducing the basic geometric entities, we now review possible geometric transformations on them. One property of these transformations is that they at least preserve straight lines and incidence relations. An in-depth treatment of such geometric transformations on projective entities can be found in HARTLEY AND ZISSERMAN 2000.

Transformations may be applied to points, lines or planes, though usually one refers to point transformations. Reformulating point transformations to transformations of 2D lines and 3D planes is mentioned in section 2.5.2 on page 44.

In the following, we concentrate on the following types of transformations:

- Linear mappings within image space resp. object space, called homographies in 2D and 3D with affine transformations and similarity transformations as special cases
- Linear mapping from object space to image space, called projective camera transformations with affine camera transformations as a special case
- Linear mapping between two different images, called fundamental matrix.

A more detailed analysis and the application of the transformations are introduced in section 3.1.2 on page 56.

Matrix and Vector Representation The transformations will generally be represented by a $n \times m$ matrix M, such that $\boldsymbol{x}' = M\boldsymbol{x}$. If a matrix is a transformation of homogeneous vectors, then it is a homogeneous transformation $\mathsf{M} \cong \lambda\mathsf{M}$ as it is defined up to scale, denoted be upright sans-serif letters.

Alternatively, we will sometimes represent the matrix as a vector $\boldsymbol{m} = \mathrm{vec}(M^{\mathsf{T}})$ where the transposed rows of M are stacked up in a mn-vector, see definition B.2 on page 184.

2.3.1 Homography

Homographies are projective transformations, that linearly map points within a projective space $\mathbb{P}^n$. Thus it can be written as a simple matrix-vector multiplication

$$\mathbf{x}' = \mathsf{H}\mathbf{x}$$

with $\mathbf{x}', \mathbf{x} \in \mathbb{R}^{n+1}$ and the $(n+1) \times (n+1)$ regular matrix H.

An affine transformation H_a is a special case of a homography and combines translation T, rotation R, the scaling of both axes M and skew W of the original space $\mathbb{P}^2$ resp. $\mathbb{P}^3$. For 2D, we obtain

$$\begin{aligned}
\mathsf{H}_a &= \mathsf{W}\,\mathsf{M}\,\mathsf{R}\,\mathsf{T} \\
&= \begin{pmatrix} 1 & s/2 & 0 \\ s/2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos(\alpha) & \sin(\alpha) & 0 \\ -\sin(\alpha) & \cos(\alpha) & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & -t_x \\ 0 & 1 & -t_y \\ 0 & 0 & 1 \end{pmatrix} \\
&= \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ 0 & 0 & 1 \end{pmatrix}
\end{aligned} \tag{2.21}$$

As an affine transformation in 2D has six degrees of freedom, all six entries in H_a are not constrained. We may constrain the upper left 2×2 matrix $H_a' = \mathsf{H}_a^{(1,2),(1,2)}$ to obtain a similarity transformation by ${H_a'}^{\mathsf{T}} H_a' = m^2 I$. This leads to two independent quadratic constraints in 2D. In a similar way, we obtain five independent quadratic constraints in 3D.

In general, the entries of a homography H can be interpreted using the canonical examples from section 2.2.1.3 on page 28: the multiplication $\mathsf{H}\,\mathbf{e}_i$ with the canonical vectors $\mathbf{e}_i$ yields the i-th column of the homography H. Thus for example the first column $(h_{11}, h_{21}, h_{31})^\mathsf{T}$ is the image of the infinite point in the direction of the x-axis. It is obvious that for an affine transformations H_a infinite points remain at infinity.

The first row $(h_{12}, h_{12}, h_{13})^\mathsf{T}$ of a homography H can be interpreted using lines: as $\mathsf{H}^\mathsf{T}\mathbf{l}' = \mathbf{l}$, cf. equation (2.11), the first row is the line which is mapped to $\mathbf{e}_1$. Thus it is easy to see that for affine transformations, the line at infinity remains at infinity.

2.3.2 Projective Camera

The most common model for cameras is the pinhole camera model: a point $\mathbf{X}$ in 3D space is projected into an image by computing a viewing ray $\mathbf{L}'(\mathbf{x}')$ from the unique projection center $\mathbf{X}_0$ to $\mathbf{X}$ and intersecting this viewing ray with a unique image plane, see figure 2.5. The image formation process described above can also be modeled as a general linear transformation of $\mathbb{P}^3$ to $\mathbb{P}^2$, which again can be written as a matrix-vector multiplication $\mathbf{x}' = \mathsf{P}\mathbf{X}$, where P is a 3×4 matrix.

This time we interpret the *rows* $\mathbf{A}^\mathsf{T}$, $\mathbf{B}^\mathsf{T}$ and $\mathbf{C}^\mathsf{T}$ of the matrix $\mathsf{P} = \begin{pmatrix} \mathbf{A}^\mathsf{T} \\ \mathbf{B}^\mathsf{T} \\ \mathbf{C}^\mathsf{T} \end{pmatrix}$ as planes $\mathbf{A}, \mathbf{B}, \mathbf{C}$ and the multiplication $\mathbf{x}' = \mathsf{P}\,\mathbf{X}$ contains three equations

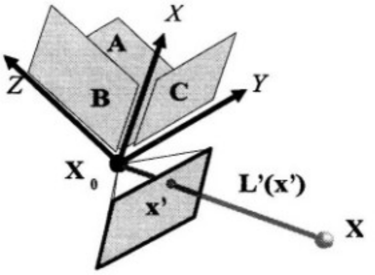

Fig. 2.5. Pinhole camera model and the three camera planes **A**, **B** and **C**.

of the form (2.10) on page 24. If $\mathbf{x}' = (u, v, w)^\mathsf{T}$, then $u = 0$ if and only if $\mathbf{X} \in \mathbf{A}$. Similarly, $v = 0 \Leftrightarrow \mathbf{X} \in \mathbf{B}$ and $w = 0 \Leftrightarrow \mathbf{X} \in \mathbf{C}$, thus a point $\mathbf{x}'$ lies at infinity if $\mathbf{X} \in \mathbf{C}$. The three planes intersect in the projection center $\boldsymbol{X}_0$, as the image point is undefined if $\mathbf{X} = (\boldsymbol{X}_0, 1)^\mathsf{T}$, already observed by DAS 1949.

Given the exterior orientation of a camera with rotation R and projection center $\boldsymbol{X}_0$ and the inner orientation with the calibration matrix K, a Euclidean interpretation of the projection matrix can be formulated as follows:

$$\mathsf{P} = \mathsf{K}R[I| - \boldsymbol{X}_0] \quad \text{with } \mathsf{K} = \begin{pmatrix} c & cs & x_H \\ 0 & c(1+m) & y_h \\ 0 & 0 & 1 \end{pmatrix} \tag{2.22}$$

with skew s, focal length c, principle point $(x_H, y_H)^\mathsf{T}$ and scale difference m between the first and second coordinate axis in the image.

Finally, as in the last section we can fix the last row of the transformation matrix with $\mathbf{C}^\mathsf{T} = (0, 0, 0, 1)^\mathsf{T}$ to obtain a special subclass of projective cameras: the plane $\mathbf{C}$ then lies at infinity and therefore the projection center $\boldsymbol{X}_0$, too. This means that all viewing rays are parallel to each other and we obtain the affine camera

$$\mathsf{P}^a = \begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

A more detailed description of various camera models can be found in HARTLEY AND ZISSERMAN 2000 or in FÖRSTNER 2001B.

2.3.3 Fundamental Matrix

Assume there are two images given, taken with different cameras P_1 and P_2. Suppose now a point $\mathbf{X}$ in object space is projected in the first and second image, $\mathbf{x}' = \mathsf{P}_1\mathbf{X}$ and $\mathbf{x}'' = \mathsf{P}_2\mathbf{X}$, cf. figure 2.6. It is clear that the projection center $\boldsymbol{X}_{01}$, $\boldsymbol{X}_{02}$ and the point $\mathbf{X}$ define a plane, the so-called epipolar plane.

Now suppose we take two arbitrarily chosen image points $\mathbf{x}'$ and $\mathbf{x}''$ from the first resp. second image. If we want to test whether they are caused by the same object point, one has to test whether the two viewing rays $\mathbf{L}'$ and $\mathbf{L}''$ are coplanar.

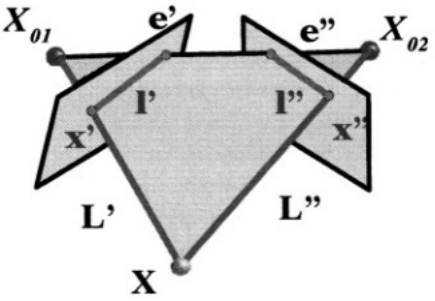

Fig. 2.6. Epipolar geometry: two projection centers $\boldsymbol{X}_{01}$ and $\boldsymbol{X}_{02}$ are connected by an epipolar line, which intersects the image planes at $\mathbf{e}'$ and $\mathbf{e}''$. The two viewing rays $\mathbf{L}'$ and $\mathbf{L}''$ point to the object point $\mathbf{X}$.

But one can also apply this test in 2D by projecting the viewing-ray $\mathbf{L}'$ into the second image, thus obtaining an image line $\mathbf{l}''$, the so-called epipolar line of $\mathbf{x}'$ which in turn has to be incident to the second image point $\mathbf{x}''$. This 2D relationship between the first point $\mathbf{x}'$ and the second point $\mathbf{x}''$ can be expressed as a coplanarity condition with the 3×3 so-called fundamental matrix F:

$$\mathbf{x}'^{\mathsf{T}} \mathsf{F} \mathbf{x}'' = 0 \tag{2.23}$$

Thus the fundamental matrix relates one image to another image. There are many ways to derive an algebraic expression for the fundamental matrix, for a review see HARTLEY AND ZISSERMAN 2000 or FAUGERAS AND LUONG 2001[3]. In section 3.1.3 we will give an explicit formula for computing F given two cameras $\mathsf{P}_1, \mathsf{P}_2$.

The vector defined by $\mathsf{F}\mathbf{x}''$ is actually the homogeneous vector for the epipolar line $\mathbf{l}''$ and by transposing equation 2.23, one also obtains the epipolar line $\mathbf{l}'$, in summary:

$$\mathbf{l}' = \mathbf{l}'(\mathbf{x}'') = \mathsf{F}\mathbf{x}'' \qquad \mathbf{l}'' = \mathbf{l}''(\mathbf{x}') = \mathsf{F}^{\mathsf{T}}\mathbf{x}'$$

Note that the projection of camera center $\boldsymbol{X}_{01}$ into the second image results in a point $\mathbf{e}'$, which is called epipole; likewise we can obtain the second epipole $\mathbf{e}''$. For epipoles, there is no epipolar line defined, thus e.g. $\mathsf{F}\mathbf{e}'' = \mathbf{0}$ and $\mathbf{e}''$ being the nullvector of F.

The fundamental matrix has 7 degrees of freedom: this is due to the fact that the 3×3 matrix is homogeneous and singular, i.e. $\det(\mathsf{F}) = 0$. Thus there are two constraints for the 9 matrix elements.

---

[3] Note that they use the transpose F^{T} as a definition for the fundamental matrix.

If the internal camera parameters K_1 and K_2 are known, only 5 parameters are unknown and the matrix is called *essential matrix,* LONGUET-HIGGINS 1981 and has 5 degrees of freedom.

2.4 Conditioning of Homogeneous Entities

Up to now we have not imposed any restrictions on the homogeneous coordinates of geometric entities nor on the parameters of geometric transformations. As the homogeneous entities are created by Euclidean objects, cf. figure 2.2, their coordinates are expressed with respect to an existing reference coordinate system. In close-range applications, we usually have the reference system nearby the entities. But in aerial imagery, the reference system is usually a geodetic system, like the Gauß-Krüger coordinate system or its international variant, the UTM system (universal-transversal-mercator system), cf. WITTE AND SCHMIDT 1995. For example, the tip of the dome of Aachen has the coordinates $x = 2505940.53[m], y = 5626590.37[m]$. Embedding this point in homogeneous coordinates as in equation (2.3) would yield a difference of seven orders of magnitudes between the Euclidean part $\boldsymbol{X}_O$ and the homogeneous part X_h. When considering subsequent computations with these coordinates, such a coordinate representation is undesirable in terms of numerical stability.

A solution is to define a local coordinate system that is close to the set of given object points, which is advisable to do prior to all methods that are discussed in this work. But if the distances between different points are large with respect to the units, an additional scaling of the coordinates is needed. For homogeneous vectors of geometric entities this means to either choose a larger X_h instead of 1 or equivalently to scale down the Euclidean coordinates by some factor. Both solutions can be achieved by scaling one part appropriately without changing the other part. We call this scaling operation *conditioning* of a geometric entity.

One may choose the second option for conditioning, i.e. scaling the Euclidean part of a geometric entity without scaling its homogeneous part. This has the effect, that the geometric entities of the vectors move closer to the Euclidean origin, cf. section 2.2. The conditioning operation can be represented by a matrix multiplication $\mathbf{x}' = \mathsf{W}(f)\mathbf{x}$, where f is the conditioning factor and W is the identity matrix except of the diagonal elements, which correspond to the Homogeneous part of the vector. The following matrices for 2D and 3D entities are obtained:

$$\begin{aligned} \mathsf{W}_x(f) &= \begin{pmatrix} f\mathsf{I}_2 & \mathbf{0} \\ \mathbf{0}^\mathsf{T} & 1 \end{pmatrix} ; \mathsf{W}_l(f) = \begin{pmatrix} \mathsf{I}_2 & \mathbf{0} \\ \mathbf{0}^\mathsf{T} & f \end{pmatrix} ; \\ \mathsf{W}_X(f) &= \begin{pmatrix} f\mathsf{I}_3 & \mathbf{0} \\ \mathbf{0}^\mathsf{T} & 1 \end{pmatrix} ; \mathsf{W}_L(f) = \begin{pmatrix} \mathsf{I} & \mathsf{0}_3 \\ \mathsf{0}_3 & f\mathsf{I}_3 \end{pmatrix} ; \mathsf{W}_A(f) = \begin{pmatrix} f\mathsf{I}_3 & \mathbf{0} \\ \mathbf{0}^\mathsf{T} & 1 \end{pmatrix} \end{aligned} \tag{2.24}$$

For scaling down the Euclidean part, one has to Choose $f < 1$. To go back to the original coordinates, one simply applies the inverse of the conditioning factor, $\mathbf{x} = \mathsf{W}(\frac{1}{f})$. Given a set of geometric entities, one has to find the entity $\mathbf{x}$ that is furthest away from the origin, and choose a conditioning factor f_{min}, such that $\mathbf{x}' = \mathsf{W}(f_{min})\mathbf{x}$ is sufficiently close to the origin. Then apply the conditioning factor to all other entities.

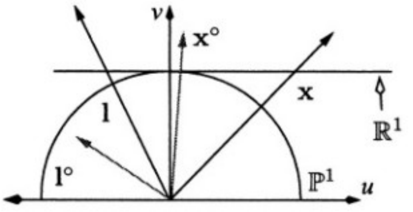

Fig. 2.7. Effect of conditioning of homogeneous vectors: a point $\mathbf{x}$ moves closer to the u-**axis**, a line $\mathbf{l}$ moves closer to the v-**axis**, but the angle between the two new vectors remains invariant.

For the projective axis, the effect of conditioning to the homogeneous vectors is depicted in figure 2.7: for a point $\mathbf{x} = (u, v)^\mathsf{T}$, the first coordinate is scaled by a factor f, the second coordinate remains invariant. The vector is now closer to the v-**axis**, a Euclidean normalization onto $\mathbb{R}^1$ would yield a scalar closer to zero. For a hyperplane $\mathbf{l} = (a, b)^\mathsf{T}$, the second coordinate is assumed to be the Euclidean part and thus the first coordinate is scaled by f, yielding a vector closer to the u-**axis**. Note that the dot-product between $\mathbf{x}$ and $\mathbf{l}$ remains invariant when scaling with the same conditioning factor.

For a transformation, we generally have two different sets of entities $\mathbf{x}_i$ and $\mathbf{x}'_i$, for which two different conditioning factors f and f' have to be computed. A conditioned transformation matrix H° can be obtained by

$$\mathsf{H}^\circ = \mathsf{W}(f')\mathsf{H}\mathsf{W}^{-1}(f) \tag{2.25}$$

The reconditioning of the estimated transformation H° is then given by

$$\mathsf{H} = \mathsf{W}^{-1}(f')\mathsf{H}^\circ\mathsf{W}(f) \tag{2.26}$$

The conditioning operation is not only useful to obtain small coordinates: as we are going to see in the next two chapters, conditioning becomes important when

- computing optimal estimations from a set of observations, cf. section 3.4
- reducing the bias of the statistical representation of uncertain geometric entities, cf. section 4.3.

Algorithm 2.1 shows a method for conditioning of a set of geometric entities. The algorithm is used in section 3.4 on page 80 when constructing an unknown entity by a set of observations and will turn out especially useful

Conditioning a set of geometric entities
Objective: Transform a set of lines, points and planes in 3D such that they move closer to the Euclidean origin.

Given are $n = n_X + n_L + n_A$ entities with n_X 3D points $\mathbf{X}_k$, n_L lines $\mathbf{L}_l$ and n_A planes $\mathbf{A}_m$.

Additionally, the required conditioning factor $f_{min} < 1$ is given, i.e. the minimal ratio between the homogeneous part and the Euclidean part of a geometric entity.

```
1. find the maximum ratio $max_{hO}$ of the norm of the Euclidean
   and homogeneous parts for points, lines and planes
   $max_{hO} := \max_{i,j,k} \frac{|\mathbf{v}_h|}{|\mathbf{v}_O|}$ for $\mathbf{v} \in \{\mathbf{X}_i, \mathbf{L}_j, \mathbf{A}_k\}$
2. determine whether it is necessary to scale the Euclidean parts.
   if $\max_{hO} < f_{min}$ then
      $f := f_{min} \cdot max_{hO}$
      if $f < \epsilon_M$ then $f := f_{min} / max_O$
            with $max_O = \max_{i,j,k} |\mathbf{v}_O|$ for $\mathbf{v} \in \{\mathbf{X}_i, \mathbf{L}_j, \mathbf{A}_k\}$
      $\mathbf{v}^\circ := \mathsf{W}(f)\mathbf{v}$ for $\mathbf{v} \in \{\mathbf{X}_i, \mathbf{L}_j, \mathbf{A}_k\}$
```

Algorithm 2.1: Algorithm for conditioning a set of geometric entities. The conditioning changes the ratio between the norms of the Euclidean and homogeneous part of a vector, which should be smaller than f_{min}. The machine accuracy ϵ_M is the smallest number such that $1.0 + \epsilon_M > 1.0$. The conditioned entity is denoted by $\mathbf{v}^\circ$, for a definition of the matrices $\mathsf{W}(f)$, see equation (2.24).

when discussing uncertainty representations of the homogeneous vectors, see section 4.3.3 on page 117.

Note that because we have subdivided all homogeneous vectors of geometric entities into a homogeneous and a Euclidean part, the conditioning of geometric entities is very simple and works consistently for homogeneous vectors for points, lines and planes within a coordinate frame.

2.5 Duality Principle

The duality principle is an important feature of projective geometry as all propositions occur twice: taking the dual of a true proposition yields a new proposition, possibly with a different meaning. As an example for duality, consider the property from section 2.1.2: three 2D points $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \in \mathbb{P}^2$ are collinear iff $|\mathbf{x}_1\ \mathbf{x}_2\ \mathbf{x}_3| = 0$. It is possible to replace the points with lines, which are elements of the dual space $(\mathbb{P}^2)^*$. Then we obtain the dual property: three 2D lines $\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3$ intersect in one point iff $|\mathbf{l}_1\ \mathbf{l}_2\ \mathbf{l}_3| = 0$, see figure 2.8.

In the following we will discuss dual entities and their transformations. They become important for the geometric constructions described in chapter 3, where we introduce the duality between the join and intersection of entities.

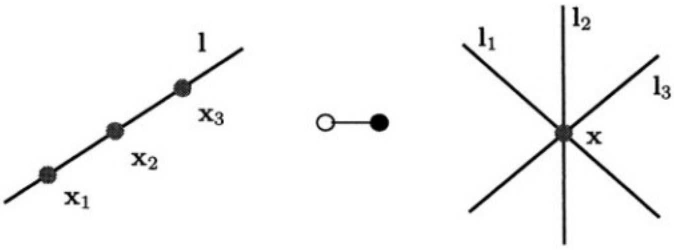
$\mathbf{l}$
$\mathbf{x}_3$
$\mathbf{x}_2$
$\mathbf{x}_1$
$\mathbf{l}_1$
$\mathbf{l}_2$
$\mathbf{l}_3$
$\mathbf{x}$

Fig. 2.8. Checking whether three points are collinear *(left)* is dual to checking whether three lines intersect in one point *(right),* see text for details. The sign ○—● is used here to express the duality between the two geometric situations.

2.5.1 Dual Entities

Dual Points and Hyperplanes As demonstrated by the last example and mentioned in section 2.1.2, points and hyperplanes in 2D and 3D are called dual to each other. As they use the same kind of homogeneous vectors, one can interpret a homogeneous point vector $\mathbf{x}$ as a hyperplane: we denote this change of interpretation by $\overline{\mathbf{x}}$ and call $\overline{\mathbf{x}}$ the dual of $\mathbf{x}$. The vector $\overline{\mathbf{x}}$ has the same coordinates, but now represents a hyperplane instead of a point.

For example, given a 2D point $\mathbf{x} = (x, y, 1)^\mathsf{T}$ its dual $\overline{\mathbf{x}}$ is a 2D line. Setting $\mathbf{l} = \overline{\mathbf{x}}$, we obtain the normalized line $\mathsf{N}_O(\mathbf{l}) = \frac{1}{x^2+y^2}(x, y, 1)^\mathsf{T}$. Thus the distance of the line $\mathbf{l} = \overline{\mathbf{x}}$ is the inverse distance of the point $\mathbf{x}$ to the origin. Additionally, the point and its dual line lie on opposite sides of the origin, see figure 2.9(left). The same reasoning applies to 3D points $\mathbf{X}$, where its dual is a plane $\overline{\mathbf{X}}$, see figure 2.9(middle).

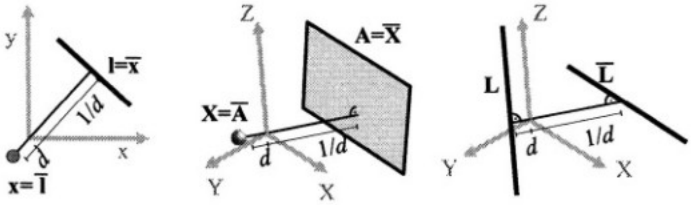
y
$\mathbf{l}=\overline{\mathbf{x}}$
$1/d$
x
d
$\mathbf{x}=\overline{\mathbf{l}}$
Z
$\mathbf{A}=\overline{\mathbf{X}}$
$\mathbf{X}=\overline{\mathbf{A}}$
d
$1/d$
Y
X
Z
$\mathbf{L}$
$\overline{\mathbf{L}}$
d
$1/d$
Y
X

Fig. 2.9. Visualization of dual entities. *Left:* the dual $\overline{\mathbf{x}}$ of a 2D point $\mathbf{x}$ is a 2D line: *middle:* the dual $\overline{\mathbf{X}}$ of a 3D point $\mathbf{X}$ is a plane; *right:* the dual $\overline{\mathbf{L}}$ of a 3D line $\mathbf{L}$ is again a 3D line. the distance d of the dual entities is inverse to the distance of the original entities.

Dual Lines One question has not been answered yet: what is the dual of a 3D line $\mathbf{L}$? To answer this question, we consider again the relationship that was derived between a point and a hyperplane, see equation (2.10). In 3D, a point $\mathbf{U}$ is incident to a plane $\mathbf{A}$ if $\mathbf{U}^\mathsf{T}\mathbf{A} = 0$. If the plane $\mathbf{A}$ was derived by three points $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$, we obtain

$$\mathbf{U}^\mathsf{T}\mathbf{A} = 0 \quad \Leftrightarrow \quad \mathbf{A}^\mathsf{T}\mathbf{U} = 0 \quad \Leftrightarrow \quad |\mathbf{X}, \mathbf{Y}, \mathbf{Z}\ ;\ \mathbf{U}| = 0 \tag{2.27}$$

see equation (2.6). The dot-product $\mathbf{U}^\mathsf{T}\mathbf{A}$ is actually an instance of the so-called cap-product, denoted by $< \mathbf{U}, \mathbf{A} >$: in general the cap-product is defined as a product of elements of two dual projective subspaces onto $\mathbb{R}$. More precisely, the cap-product is defined for the two corresponding *vector* spaces where the sum of the dimensions of the underlying *vector* subspaces is equal to the dimension of the vector space. For example, a 3D point with projective dimension 0 has a vector dimension of 1 as the projective point in $\mathbb{P}^3$ is represented by a line in $\mathbb{R}^4$, cf. figure 2.1(d) for the 2D case. The dual of a 3D point is a 3D plane with projective dimension 2 resp. vector dimension 3. Summing up the vector dimensions yields 1 + 3 = 4 the dimension of $\mathbb{R}^4$. To derive a cap-product involving 3D lines, we can re-interpret the determinant $|\mathbf{X}, \mathbf{Y}, \mathbf{Z}\ ;\ \mathbf{U}|$ by defining two lines $\mathbf{L}(\mathbf{X}, \mathbf{Y})$ and $\mathbf{M}(\mathbf{Z},\mathbf{U})$:

$$|\underbrace{\mathbf{X}\,\mathbf{Y}}_{\mathbf{L}}\ ;\ \underbrace{\mathbf{Z}\,\mathbf{U}}_{\mathbf{M}}| = 0 \tag{2.28}$$

This equation holds if the four points are coplanar or equivalently, if the lines $\mathbf{L}$ and $\mathbf{M}$ intersect in one point. We can expand the determinant in (2.28) to hold the Plücker coordinates of the lines $\mathbf{L}$, $\mathbf{M}$:

$$\begin{aligned}
\begin{vmatrix} X_1 & Y_1 & Z_1 & U_1 \\ X_2 & Y_2 & Z_2 & U_2 \\ X_3 & Y_3 & Z_3 & U_3 \\ X_4 & Y_4 & Z_4 & U_4 \end{vmatrix} &= \begin{vmatrix} X_1 & Y_1 \\ X_4 & Y_4 \end{vmatrix} \begin{vmatrix} Z_2 & U_2 \\ Z_3 & U_3 \end{vmatrix} - \begin{vmatrix} X_2 & Y_2 \\ X_4 & Y_4 \end{vmatrix} \begin{vmatrix} Z_1 & U_1 \\ Z_3 & U_3 \end{vmatrix} + \begin{vmatrix} X_3 & Y_3 \\ X_4 & Y_4 \end{vmatrix} \begin{vmatrix} Z_1 & U_1 \\ Z_2 & U_2 \end{vmatrix} \\
&\quad + \begin{vmatrix} X_2 & Y_2 \\ X_3 & Y_3 \end{vmatrix} \begin{vmatrix} Z_1 & U_1 \\ Z_4 & U_4 \end{vmatrix} - \begin{vmatrix} X_1 & Y_1 \\ X_3 & Y_3 \end{vmatrix} \begin{vmatrix} Z_2 & U_2 \\ Z_4 & U_4 \end{vmatrix} + \begin{vmatrix} X_1 & Y_1 \\ X_2 & Y_2 \end{vmatrix} \begin{vmatrix} Z_3 & U_3 \\ Z_4 & U_4 \end{vmatrix} \qquad (2.29) \\
&= +l_{14}m_{23} - l_{24}m_{13} + l_{34}m_{12} + l_{23}m_{14} - l_{13}m_{24} + l_{12}m_{34} \\
&= -l_{41}m_{23} - l_{42}m_{31} - l_{43}m_{12} - l_{23}m_{41} - l_{31}m_{42} - l_{12}m_{43}
\end{aligned}$$

The 2 × 2 minors are the Plücker coordinates for the lines $\mathbf{L}$ and $\mathbf{M}$, see also section 2.2.4 on page 32. One can check that equation (2.28) is equivalent to

$$-\boldsymbol{L}_h^\mathsf{T}\boldsymbol{M}_O - \boldsymbol{L}_O^\mathsf{T}\boldsymbol{M}_h = 0 \tag{2.30}$$

Thus we have derived a condition for the incidence of two lines $\mathbf{L}$ and $\mathbf{M}$ from the incidence of a plane $\mathbf{A}$ and a point $\mathbf{U}$.

Equation (2.30) defines the cap-product $< \mathbf{L}, \mathbf{M} >$ of two lines in a similar way as $< \mathbf{A}, \mathbf{U} >$ with equation (2.10); the dual of a 3D line is again a 3D line. Interpreting the cap-product by a dot-product of dual entities, we obtain

$$< \mathbf{L}, \mathbf{M} > := -\boldsymbol{L}_h^{\mathsf{T}} \boldsymbol{M}_O - \boldsymbol{L}_O^{\mathsf{T}} \boldsymbol{M}_h = -\mathbf{L}^{\mathsf{T}} \overline{\mathbf{M}} \quad \text{with } \overline{\mathbf{M}} := \begin{pmatrix} \boldsymbol{M}_O \\ \boldsymbol{M}_h \end{pmatrix}$$

where $\overline{\mathbf{M}}$ is the dual of the line $\mathbf{M}$ and is obtained by switching the homogeneous and Euclidean part of the vector $\mathbf{M}$. The visualization of a line and its dual is shown in figure 2.9(left). Note that again the point and its dual line are on opposite sides of the origin and the dual has an inverse distance to the origin with respect to the original entity.

To summarize, the possible cap-products in 2D and 3D are:

entity 1	dim. of vector subspace	dual entity 2	dim. of vector subspace	dim. of vector space	cap-product
point $\mathbf{x}$	1	line $\mathbf{l}$	2	1+2=3	$< \mathbf{x}, \mathbf{l} > \; := \mathbf{x}^{\mathsf{T}} \mathbf{l}$
point $\mathbf{X}$	1	plane $\mathbf{A}$	3	1+3=4	$< \mathbf{X}, \mathbf{A} > \; := \mathbf{X}^{\mathsf{T}} \mathbf{A}$
line $\mathbf{L}$	2	line $\mathbf{M}$	2	2+2=4	$< \mathbf{L}, \mathbf{M} > \; := \mathbf{L}^{\mathsf{T}} \overline{\mathbf{M}}$

Usually the cap-product is defined in the context of the so-called double or Grassmann-Cayley algebra, where operations such as join and intersection form an algebra on projective subspaces. In the next chapter we introduce these operations on the geometric entities defined. Again, for a rigorous treatment of this subject, see FAUGERAS AND LUONG 2001.

An alternative proof that two lines are incident if equation (2.30) does not explicitly use determinants but only analytical geometry and follows BRAND 1966, p. 54: assume a point $(\boldsymbol{X}^{\mathsf{T}}, 1) \in \mathbf{L}^{\mathsf{T}} = (\boldsymbol{L}_h^{\mathsf{T}}, \boldsymbol{L}_O^{\mathsf{T}})$ and $(\boldsymbol{Y}^{\mathsf{T}}, 1) \in \mathbf{M}^{\mathsf{T}} = (\boldsymbol{M}_h^{\mathsf{T}}, \boldsymbol{M}_O^{\mathsf{T}})$. The lines are coplanar if and only if the vectors $(X — Y)$, $\boldsymbol{L}_h$ and $\boldsymbol{M}_h$ are coplanar, or algebraically $(\boldsymbol{X} - \boldsymbol{Y})^{\mathsf{T}} (\boldsymbol{L}_h \times \boldsymbol{M}_h) = 0$. The assertion follows now from the fact that $\boldsymbol{L}_O = \boldsymbol{X} \times \boldsymbol{L}_h$ and $\boldsymbol{M}_O = \boldsymbol{Y} \times \boldsymbol{M}_h$.

Algebraic Expression for Dual Entities As we want to express all operations on entities within linear algebra, we define the dual operator as a linear transformation from one projective subspace to another. For points and lines in 2D and for points and planes in 3D, the dual operator is the identity matrix I_3 resp. I_4:

$$\overline{\mathbf{x}} := I_3 \, \mathbf{x} \quad , \; \overline{\mathbf{l}} := I_3 \, \mathbf{l} \qquad ; \qquad \overline{\mathbf{X}} := I_4 \, \mathbf{X} \quad , \; \overline{\mathbf{A}} := I_4 \, \mathbf{A}$$

For lines in 3D, we define the dual operator as follows, considering our choice for the Plücker-coordinates:

$$\overline{\mathbf{L}} := C\mathbf{L} \quad \text{with } C := \begin{pmatrix} 0&0&0&1&0&0\\ 0&0&0&0&1&0\\ 0&0&0&0&0&1\\ 1&0&0&0&0&0\\ 0&1&0&0&0&0\\ 0&0&1&0&0&0 \end{pmatrix} = \begin{pmatrix} 0 & I_3 \\ I_3 & 0 \end{pmatrix} \tag{2.31}$$

Note that applying the dual operators I_3, I_4 and C changes the interpretation of a homogeneous vector from a projective entity to its dual.

2.5.2 Dual of Point Transformations

The existence of dual entities also has an effect on the transformations mentioned in section 2.3 on page 34, namely homographies, projective transformations and fundamental matrices. There exists a theorem to derive a duality property for linear mappings of all geometric entities. In section 2.1.2 it was already mentioned that given any linear map H from $\mathbb{P}^n$ to $\mathbb{P}^m$, the dual mapping is the transpose H^T that maps the dual spaces $(\mathbb{P}^m)^*$ to $(\mathbb{P}^n)^*$. This behavior can be transferred to any of the proposed geometric entities, see FAUGERAS AND LUONG 2001, p. 154.

For example assume a projective transformation P of a 3D-point $\mathbf{X}$ to 2D point $\mathbf{x}'$, the duality relation for $\mathbf{x}' = \mathsf{P}\mathbf{X}$ can be written as

$$\mathbf{x}' = \mathsf{P}\mathbf{X} \quad \circ\!\!-\!\!\bullet \quad \mathsf{P}^\mathsf{T}\mathbf{l}' = \mathbf{A} \tag{2.32}$$

where again the simplified notation $\overline{\mathbf{x}'} = \mathbf{l}'$ and $\overline{\mathbf{X}} = \mathbf{A}$ holds. The sign $\circ\!\!-\!\!\bullet$ is used here to express the duality between the two expressions.

Since we know how to interpret the elements of the dual spaces $(\mathbb{P}^m)^*$ and $(\mathbb{P}^n)^*$ we can directly apply this to 2D and 3D. In equation (2.32), a 2D image line may be back-projected in 3D using a projective camera matrix P and results in a projective plane $\mathbf{A}$, which intersects the focal point and the 2D line on the image plane.

Similarly, the rules of duality for homographies H, projective transformations P and fundamental matrices F are

$$\begin{aligned} \mathbf{x}' = \mathsf{H}\mathbf{x} \quad &\circ\!\!-\!\!\bullet \quad \mathsf{H}^\mathsf{T}\mathbf{l}' = \mathbf{l} \\ \mathbf{X}' = \mathsf{H}\mathbf{X} \quad &\circ\!\!-\!\!\bullet \quad \mathsf{H}^\mathsf{T}\mathbf{A}' = \mathbf{A} \\ \mathbf{L}' = \mathsf{H}_L\mathbf{L} \quad &\circ\!\!-\!\!\bullet \quad \mathsf{H}_L^\mathsf{T}\overline{\mathbf{L}'} = \overline{\mathbf{L}} \\ \mathbf{x}' = \mathsf{P}\mathbf{X} \quad &\circ\!\!-\!\!\bullet \quad \mathsf{P}^\mathsf{T}\mathbf{l}' = \mathbf{A} \\ \mathbf{l}' = \mathsf{F}\mathbf{x}'' \quad &\circ\!\!-\!\!\bullet \quad \mathsf{F}^\mathsf{T}\mathbf{x}' = \mathbf{l}'' \\ \mathbf{l}' = \mathsf{Q}\mathbf{L} \quad &\circ\!\!-\!\!\bullet \quad \mathsf{Q}^\mathsf{T}\mathbf{x}' = \overline{\mathbf{L}} \end{aligned}$$

The last duality relation involves the projective camera transformation Q for 2D lines, which will be covered in section 3.1.2.2 on page 57.

Right now, the representation of the line transformation $\mathbf{L}' = \mathsf{H}_L\mathbf{L}$ is not obvious as not all homographies $\mathsf{H}_L : \mathbb{P}^5 \rightarrow \mathbb{P}^5$ within the five-dimensional projective space correspond to valid homographies in $\mathbb{P}^3$. We will deal with line transformations H_L in section 3.1.2 on page 56: as soon as we have introduced algebraic expressions for constructions join and intersection, we are able to derive a construction of H_L given a 3D homography H. The same argument applies to projective camera transformations Q for 3D lines, see section 3.1.2.2.

6 +B 9:5;9 9 9/ 9 .)!) +C)!

Pag. 26 (43)

$d(1,0)=|c|/\sqrt{a^2+b^2}$

—

pc_oweto\MHAMYA1.pdf
Pag. 196/197 (100)
184. The Equations of a Line
Perpendicular to Two
Given Straight Lines.
→ nonparallel.
straight lines

—

e-rara.ch

—

Leibniz →

—

pstenio45@gmail.com

—

pdfdrive.com

—

→ aivento
Tobermonita luminosa (Roma)

Yang Chan chiné Pedrina

Now apply formula (1). According to (1b), Sec. 161, we have $x_1=2$, $y_1=0$, $z_1=0$, $l_1=3$, $m_1=2$, $n_1=1$, so that

$$\begin{vmatrix} y_0-y_1 & z_0-z_1 \\ m_1 & n_1 \end{vmatrix}=\begin{vmatrix} 0 & 1 \\ 2 & 1 \end{vmatrix}=-2,$$

$$\begin{vmatrix} z_0-z_1 & x_0-x_1 \\ n_1 & l_1 \end{vmatrix}=\begin{vmatrix} 1 & -1 \\ 1 & 3 \end{vmatrix}=4,$$

$$\begin{vmatrix} x_0-x_1 & y_0-y_1 \\ l_1 & m_1 \end{vmatrix}=\begin{vmatrix} -1 & 0 \\ 3 & 2 \end{vmatrix}=-2$$

We obtain

$$d=\frac{\sqrt{(-2)^2+4^2+(-2)^2}}{\sqrt{3^2+2^2+1^2}}=2\sqrt{\frac{3}{7}}$$

163. The Condition for Two Straight Lines Intersecting or Lying in a Single Plane

If the straight lines

$$\frac{x-x_1}{l_1}=\frac{y-y_1}{m_1}=\frac{z-z_1}{n_1}, \tag{1}$$

$$\frac{x-x_2}{l_2}=\frac{y-y_2}{m_2}=\frac{z-z_2}{n_2} \tag{2}$$

lie in a single plane, then

$$\begin{vmatrix} x_2-x_1 & y_2-y_1 & z_2-z_1 \\ l_1 & m_1 & n_1 \\ l_2 & m_2 & n_2 \end{vmatrix}=0 \tag{3}$$

or, in vector form,

$$(\boldsymbol{r}_2-\boldsymbol{r}_1)\,\boldsymbol{a}_1\boldsymbol{a}_2=0 \tag{3a}$$

Conversely, if Condition (3) is fulfilled, then the straight lines lie in a single plane.

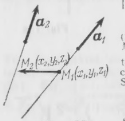

Fig. 177

Explanation. If the straight lines (1) and (2) lie in a single plane, then the straight line M_1M_2 (Fig. 177) also lies in that plane, i. e. the vectors $\overrightarrow{M_1M_2}$, $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ are coplanar (and conversely). This is what Eq. (3) expresses (see Sec. 120).

Note. If $\frac{l_1}{l_2}=\frac{m_1}{m_2}=\frac{n_1}{n_2}$ [here, (3) will definitely be satisfied], then the straight lines are parallel, otherwise the straight lines satisfying Condition (3) intersect.

Example. Determine whether the straight lines

$$\frac{x}{1}=\frac{y}{2}=\frac{z}{3}, \tag{1}$$

$$\frac{x+1}{2}=\frac{y-1}{1}=\frac{z+1}{4} \tag{2}$$

intersect, and if they do, at what point.

Solution. The straight lines (1) and (2) lie in one plane since the determinant (3) equal to $\begin{vmatrix} -1 & 1 & -1 \\ 1 & 2 & 3 \\ 2 & 1 & 4 \end{vmatrix}$ vanishes. These lines are not parallel (the direction numbers are not proportional). In order to find the point of intersection, we have to solve a system of four equations (1), (2) in three unknowns. As a rule, such a system does not have any solutions, but in the given case [because Condition (3) is fulfilled] there is a solution. Solving a system of any three equations, we get $x=1$, $y=2$, $z=3$. The fourth equation is satisfied. The point of intersection is (1, 2, 3).

164. The Equations of a Line Perpendicular to Two Given Straight Lines

The straight line UV intersecting two nonparallel straight lines (L_1 and L_2 in Fig. 178)

$$\frac{x-x_1}{l_1}=\frac{y-y_1}{m_1}=\frac{z-z_1}{n_1},$$

$$\frac{x-x_2}{l_2}=\frac{y-y_2}{m_2}=\frac{z-z_2}{n_2}$$

and perpendicular to them is represented (in vector form) by the equations

$$(\boldsymbol{r}-\boldsymbol{r}_1)\,\boldsymbol{a}_1\boldsymbol{a}=0, \tag{1}$$

$$(\boldsymbol{r}-\boldsymbol{r}_2)\,\boldsymbol{a}_2\boldsymbol{a}=0 \tag{2}$$

where $\boldsymbol{a}_1=\{l_1,\ m_1,\ n_1\}$, $\boldsymbol{a}_2=\{l_2,\ m_2,\ n_2\}$ and $\boldsymbol{a}=\boldsymbol{a}_1\times\boldsymbol{a}_2$.

Taken separately, Eq. (1) is the plane P_1 drawn through the straight line L_1 parallel to the vector $\boldsymbol{a}=\boldsymbol{a}_1\times\boldsymbol{a}_2$ (Sec. 159). Similarly, (2) is plane P_2 drawn through L_2 parallel to $\boldsymbol{a}$.

Point K_1, where UV intersects L_1, is found at the intersection of L_1 with the plane P_2. Similarly, we find, the point K_2 and then the length of the common perpendicular K_1K_2.

Note. If L_1 and L_2 are parallel [then $\boldsymbol{a}=0$ and Eqs. (1), (2) become identities], there is an infinity of straight lines UV. To obtain the equation of one of them, take on L_1 (Fig. 179) an arbitrary point K_1 and form the equation of the straight line passing through K_1 in the direction of the vector $\boldsymbol{a}_1\times\boldsymbol{b}$, where $\boldsymbol{b}=\boldsymbol{a}_1\times(\boldsymbol{r}_2-\boldsymbol{r}_1)$.

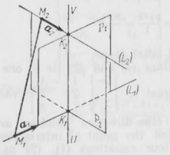

Fig. 178

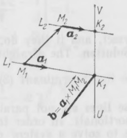

Fig. 179

Example 1. Find the equations of the perpendicular to the straight lines

$$x=2+2t,\qquad y=1+4t,\qquad z=-1-t, \tag{3}$$

$$x=-31+3t',\qquad y=6+2t',\qquad z=3+6t' \tag{4}$$

Solution. We have $\boldsymbol{a}_1=\{2,\ 4,\ -1\}$, $\boldsymbol{a}_2=\{3,\ 2,\ 6\}$, $\boldsymbol{a}=\boldsymbol{a}_1\times\boldsymbol{a}_2=\{26,\ -15,\ -8\}$.

The desired perpendicular is given by the equations

$$\begin{vmatrix} x-2 & y-1 & z+1 \\ 2 & 4 & -1 \\ 26 & -15 & -8 \end{vmatrix}=0,$$

$$\begin{vmatrix} x+31 & y-6 & z-3 \\ 3 & 2 & 6 \\ 26 & -15 & -8 \end{vmatrix}=0$$

or, after simplifying,

$$\begin{cases} 47x+10y+134z+30=0, & (5)\\ 74x+180y-97z+1505=0 & (6)\end{cases}$$

The point K_1 of intersection of the common perpendicular with the straight line (3) is found from the system (3)-(6). We get $K_1(-2,\ -7,\ 1)$. Similarly, we find $K_2(-28,\ 8,\ 9)$. The length d of the common perpendicular is

$$d=\sqrt{(-2+28)^2+(-7-8)^2+(1-9)^2}=\sqrt{965}$$

Example 2. Find the equations of the line perpendicular to the straight lines

$$x=2+2t,\qquad y=3+2t,\qquad z=t, \tag{7}$$

$$x=5+2t',\qquad y=4+2t',\qquad z=1+t' \tag{8}$$

The lines are parallel: $\boldsymbol{a}_1=\boldsymbol{a}_2=\{2,\ 2,\ 1\}$, $\boldsymbol{r}_2-\boldsymbol{r}_1=\{3,\ 1,\ 1\}$, $\boldsymbol{b}=\boldsymbol{a}_1\times(\boldsymbol{r}_2-\boldsymbol{r}_1)=\{1,\ 1,\ -4\}$. The direction vector of the common perpendicular $\boldsymbol{a}_1\times\boldsymbol{b}=\{-9,\ 9,\ 0\}$ or, multiplying by $\frac{1}{9}$, $\{-1,\ 1,\ 0\}$. For the initial point we take an arbitrary point $K_1(2+2t;\ 3+2t;\ t)$ of the line (7). We obtain the equation of the common perpendicular

$$\frac{x-(2+2t)}{-1}=\frac{y-(3+2t)}{1}=\frac{z-t}{0} \tag{9}$$

where t is an arbitrary number. To find the point K_2 of intersection of the common perpendicular (9) with the straight line (8), substitute expression (8) into Eq. (9). This yields

$$\frac{3+2(t'-t)}{-1}=\frac{1+2(t'-t)}{1}=\frac{1+(t'-t)}{0}$$

Any one of these equations yields $t'=t-1$; substituting into (8), we get $K_2(3+2t,\ 2+2t,\ t)$, so that

$$d=|K_1K_2|=$$

$$=\sqrt{[(3+2t)-(2+2t)]^2+[(2+2t)-(3+2t)]^2+[t-t]^2}=\sqrt{2}$$

165. The Shortest Distance Between Two Straight Lines

The shortest distance between the straight lines L_1 and L_2 is the length d of their common perpendicular. It can be found by forming the equations of the common perpendicular (Sec. 164, Examples 1 and 2). A simpler way, however, is to find d directly.

(1) If L_1 and L_2 are not parallel (Fig. 180), then

$$d=\frac{|(\boldsymbol{r}_2-\boldsymbol{r}_1)\,\boldsymbol{a}_1\boldsymbol{a}_2|}{|\boldsymbol{a}_1\times\boldsymbol{a}_2|}=\frac{|\boldsymbol{r}_2-\boldsymbol{r}_1)\,\boldsymbol{a}_1\boldsymbol{a}_2|}{\sqrt{(\boldsymbol{a}_1\times\boldsymbol{a}_2)^2}} \tag{1}$$

20200922

geometrias.eu/depositum/
Tratado0508.pdf

—

ridicando comida → negando

—

Traité des Courbes Spéciales
Remarquables planes
et gauches
F. Gomes Teixeira

—

3 Tomos

416, 504 e 440 pags

ISBN 978-2-87647-162-7

mathcurve.com

—

Brocard

ForgottenBooks.com

Calculus
Shockley

20201015

FDM 29091753

Odontologia 6207

27 Out 14h.

The Shelby Electric Plant

Shelby, Ohio

Adolphe Chaillet

lâmpada elétrica bombeiros USA.

Lorentz (gamma)

↳ drdonlincoln.com

espaço anti-de Sitter ou hiperbólico, que possui uma curvatura geométrica negativa constante.

buracos negros supercondutores
universo anti-de Sitter ⤴

≠

do nosso universo

—

Test Velocidade NET

beta.simet.nic.br

—

Gauss → unidade de densidade de fluxo magnético.

MRI → 10.000 gauss = 1 Tesla

lorelei.math.uni-potsdam.de
lorelei

—

VIKI VAPORUBE MENTOLADO
ATRÁS DA ORELHA
→ mosquito vai embora

—

mensóregia

—

2020 1017
+275
378
→ 653

Hello 1984

1009 pages

⇒ Pappus - La collection Mathématiques

Paul Ver Eecke

São Domingos Sávio → patrono dos estudantes

— — —

A) Oeuvres complètes de Christiaan Huygens

gallica.bnf.fr/ark:/12148/
bpt6k77856x

B) opticks - Newton

gallica.bnf.fr/ark:/12148/
bpt6k3362k

Wolfgang Kubicki FDP

Jogo perigoso com a democracia.

⇒ Pappus .jp2 ⇒ OCR

gallica.bnf.fr/ark:/12148/
bpt6k9721942t.r

=

Journal de mathématiques
pures et appliquées
Tome XII
Joseph Liouville.

=

Traité de géométrie
supérieure par
M. Chasles

↓

—

gallica.bnf.fr/ark:/12148/
bpt6k9605508v
bpt6k996370.r
bpt6k9805529h ← ✡
bpt6k5790435r ←

=

bpt6k1410388s

=

0B B)B

) "" + " 7 " #("+) 7. -- (,

9:5;9 9 9/ 9
E BC! C B B C)B CCCK L C C+) 7. -- (,)!

+) "" + " 7 " #("+) 7.. $&

9:5;9 9 9/ 9
E BC! C B B C)B CCCK L C C+) 7.. $&)!

) "" + " 7 " #("+) 7.(-- .

9:5;9 9 9/ 9
E BC! C B B C)B CCCK L C C+) 7.(-- .)!

!) "" + " 7 " #("+) 7 # $((

9:5;9 9 9/ 9
5 C C C C B B CK LH C0 C+) 7 # $(( )!

) "" + " 7 " #("+) 7-&. #$-

9:5;9 9 9/ 9
5)) C C C) M C! CK L C C+) 7-&. #$-)!

) "" + " 7 " #("+) 7 - &#-$

9:5;9 9 9/ 9 E BC! C B B C
C) CK LH <C > N : C+) 7 - &#-$)!

/ , F ! $

) "" + " 7 " #("+) 7 - $ (+

9:5;9 9 9/ 9
H C , C!OF ! C C CK LF ! C $ $ (-C+) 7 - $ (+)!

+) "" + " 7 " #("+) 7 - -.&.$

9:5;9 9 9/ 9
H C , C!OF ! C C CK LF ! C $ $ (-C+) 7 - -.&.$)!

) "" + " 7 " #("+) 7 - #. &

9:5;9 9 9/ 9
H C , C!OF ! C C CK LF ! C $ $ (-C+) 7 - #. &)!

!) "" + " 7 " #("+) 7 - .$&
/+
H G * , ! B B B B G !OF !

9:5;9 9 9/ 9
H CG * C , C! CB B CK LF ! C $ $ (-C+) 7 - .$&)!

0B B
) "" + " 7 " #("+) 7$$& ##*

9:5;9 9 9/ 9
0B B C C , BC K5) ! C CK L CP C+) 7$$& ##*)!

+) "" + " 7 " #("+) 7 -#$.&

9:5;9 9 9/ 9
:) N C G C C O C CK L C C+) 7 -#$.&)!

$ 6 , ! >

) "" + " 7 " #("+) 7. &

9:5;9 9 9/ 9 $
:)) C!O < C C! C B B CK L CP C+) 7. &)!

+) "" + " 7 " #("+) 7.& ## (

9:5;9 9 9/ 9 $
:)) C!O < C C! C B B CK L CP C+) 7.& ## ( )!

)

bpt6k290445

bpt6k6567453r - Stéréotomie

20201020

$E = E_F + E_{Fraco} + [\text{Eletrons}] + E_{c(\text{cinetica})}$

↳ velocidade

= força e gera a pseudo-grandeza

→ Principio Maupertuis

↳ ação no intervalo de tempo

$$\delta \int_{t_1}^{t_2} L \, dt = 0$$

Arvin Ash

youtube.com/c/ArvinAsh/

' $

NOVOVINO - 157 - 2020 / Books - 157 - 2020 /

A). 1830 - 1930 A Century of Geometry /
Century of Geometry... pdf

Pag. 201 (208)

What is space?

Hermes Trismegistos

"Deus est sphaera infinita
cuius centrum est ubique,
circumferencia vero nusquam"

(God is the infinite sphere
whose centre is everywhere
yet whose surface is nowhere).

B) Importantissimo /
Quantum Triangulations... pdf

✡ Annalisa Marzuoli

annalisa.marzuoli@
unipv.it

Pag. 29 (47) Fig. 1.14

The image of the curve
(1.77) is characterized
by the intersection in
($\mathbb{R}^3$; O, x, y, z) ...

#
9:5;9 9 9/ 9 # ; E !
) @ ;) @ Q H ! ! 5 >) &)!

+ 9:5;9 9 9/ 9 # ; E !
) @ @ ! ;)) 1 A ! +)!

9:5;9 9 9/ 9 # E 0 < 8<
E) 1 A ! + ..&)!

/+

9:5;9 9 9/ 9 # ; E !
) @ ;) @ Q H ! ! 5 >) &)!
' . #& > #

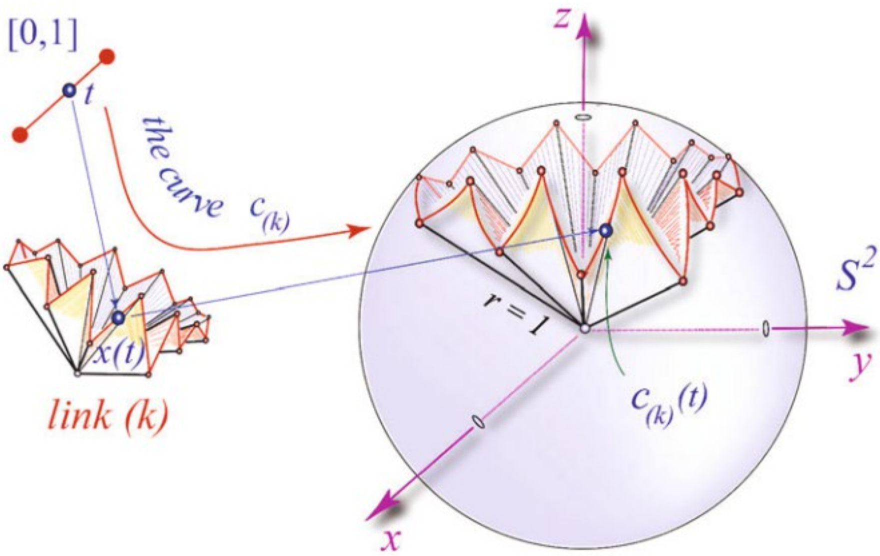
[0,1]
t
the curve $c_{(k)}$
$x(t)$
link (k)
z
S^2
$r = 1$
y
$c_{(k)}(t)$
x

which is smooth in the $\mathbb{R}^{N_1}$ topology, the two local coordinate charts $(\mathcal{U}_{T(1)}, \pi_{T(1)})$ and $(\mathcal{U}_{T(2)}, \pi_{T(2)})$ have diffeomorphic intersection, and consequently $POL_{g,N_0}(M)$ is a differentiable orbifold locally modelled on $(\mathcal{U}_T, \pi_T) \subset \mathcal{T}^{met}_{g,N_0}(M)/Aut(T)$ as stated.

A direct consequence of Proposition 1.4 is the following

Theorem 1.3 *The space of Polyhedral structures $POL_{g,N_0}(M)$ generated by closed polyhedral surfaces (T_l, M) of genus g with N_0 vertices has*

$$\mathcal{C}ard\ \left\{\mathcal{T}_{g,N_0}(M)\right\}, \tag{1.59}$$

distinct orbifold strata, where $\mathcal{C}ard$ denotes the cardinality of the set of distinct triangulations considered.

Proof Since, according to Proposition 1.4, the strata $\frac{\mathbb{R}^{N_1}_+}{Aut(T)}$ of $POL_{g,N_0}(M)$ are labelled by the finite set of distinct triangulations $[T] \in \mathcal{T}_{g,N_0}(M)$, the theorem immediately follows. ■

In order to explore the basic properties of the space of polyhedral structures $POL_{g,N_0}(M)$ we need to discuss the structure of $\mathcal{T}^{met}_{g,N_0}(M)$ in the neighborhood of a given polyhedral surface (T_l, M). This will provide the appropriate notion of local deformation of polyhedral surfaces. A thorough analysis of the deformation theory of polyhedral surface, (in the general case of Euclidean, Hyperbolic, and Spherical triangulated surfaces), is presented in a series of remarkable papers by Feng Luo [7, 12, 13], who, building on the influential work of Thurston [22], Rivin [19], and Leibon [11], provides a very detailed and illustrative geometrical framework for the theory. Here we take a slightly different point of view, more apt to the physical applications, emphasizing the natural connections with the Witten–Kontsevich theory.

1.8 Cotangent Cones and Circle Bundles $\mathcal{Q}_{(k)}$ Over $\mathcal{T}^{met}_{g,N_0}(M)$

A basic question in discussing the deformation theory of $\mathcal{T}^{met}_{g,N_0}(M)$ is to characterize the space of the infinitesimal deformations of a given polyhedral surface (T_l, M), i.e. a suitable notion of tangent space to $\mathcal{T}^{met}_{g,N_0}(M)$ at (T_l, M). The idea is to formalize the intuitive picture of what a deformation of a neighborhood of a vertex $\sigma^0(k)$ in a polyhedral surface looks like. To this end, we inject the given vertex $\sigma^0(k)$ in the origin O of $(\mathbb{R}^3; O, x, y, z)$, where we identify $\mathbb{R}^3$, endowed with the vector product ×, with the Lie algebra $\mathfrak{so}(3)$ of the rotation group $SO(3)$. Since $\sigma^0(k)$ is a conical point whose generators are edges of Euclidean triangles, we can associate with $\sigma^0(k)$ the polyhedral cone with apex O, whose generators have lengths normalized to 1 and whose directrix is a piecewise–geodesic curve on the surface of a unit sphere $\mathbb{S}^2$ centered at the origin. This curve is a spherical polygon whose side–lengths are provided by the original vertex angles $\{\theta_\alpha\}$ generating the conical geometry around $\sigma^0(k)$. This construction has a natural $SO(3)$ symmetry which, when reduced, generates a set of spherical polygons which parametrizes all possible distinct polyhedral cones in $(\mathbb{R}^3; O, x, y, z)$ having the same intrinsic metric geometry of the local Euclidean structure $\mathscr{E}\left(Star[\sigma^0(k)]\right)$ near $\sigma^0(k)$. This set characterizes the *polyhedral cotangent cone* to $\mathscr{E}\left(Star[\sigma^0(k)]\right)$ at $\sigma^0(k)$. There is a natural bundle associated with this construction which, to some extent, provides a suitable notion of (co)tangent space to $\mathscr{T}^{met}_{g,N_0}(M)$ at (T_l, M).

We begin with a few notational remarks. Recall that if $q(k) \geq 2$ triangles $\sigma^2(h_\alpha, k, h_{\alpha+1})$ are incident on $\sigma^0(k)$, then

$$Star[\sigma^0(k)] \doteq \cup_{\alpha=1}^{q(k)} \sigma^2(h_\alpha, k, h_{\alpha+1}), \quad h_{q(k)+1} \equiv h_1, \tag{1.60}$$

where the index α is taken modulo $q(k)$. We denote by

$$\mathscr{E}\left(Star[\sigma^0(k)]\right) \doteq \left\{(\theta_{\alpha,k,\alpha+1},\ \theta_{k,\alpha+1,\alpha},\ \theta_{\alpha+1,\alpha,k})\right\}, \tag{1.61}$$

the generic conically complete locally Euclidean structure on $Star[\sigma^0(k)]$, and consider the ordered sequence of vertex angles at $\sigma^0(k)$

$$\left\{\theta_{\alpha,k,\alpha+1}\right\}_{\alpha=1}^{\alpha=q(k)}, \quad \theta_{q(k),k,q(k)+1} \equiv \theta_{q(k),k,1}, \tag{1.62}$$

as a point of $[0, \pi]^{q(k)}$, modulo cyclic permutations in $\mathbb{Z}_{q(k)}$. The boundary points 0 and π corresponds to the degeneration of those triangles $\sigma^2(h_\alpha, k, h_{\alpha+1})$ whose corresponding vertex angle $\theta_{\alpha,k,\alpha+1}$ either goes to 0, (a *thinning* triangle), or to π, (a *fattening* triangle). Note that in the former case we have that the edge length $l(h_\alpha, h_{\alpha+1}) \to 0$ with $l(l(k, h_\alpha)$ and $l(k, h_{\alpha+1})$ both > 0; whereas, in the latter case we have $l(h_\alpha, h_{\alpha+1}) = l(k, h_\alpha) + l(k, h_{\alpha+1})$. According to Proposition 1.1 to any $\mathscr{E}\left(Star[\sigma^0(k)]\right)$ we can associate a conical defect

$$\Theta(k) \doteq \sum_{\alpha=1}^{q(k)} \theta_{\alpha,k,\alpha+1}, \tag{1.63}$$

supported at $\sigma^0(k)$, which induces a conical geometry on $Star[\sigma^0(k)]$. Since (1.63) depends only on the $\sigma^0(k)$ angles $\{\theta_{\alpha,k,\alpha+1}\} \in \mathscr{E}\left(Star[\sigma^0(k)]\right)$, it is natural to define the set of *vertex angle structures* at $\sigma^0(k)$ associated with $\mathscr{E}\left(Star[\sigma^0(k)]\right)$ according to

$$\mathscr{V}ert_{q(k)} := \left\{\{\theta_\alpha\}_k \in [0, \pi]^{q(k)}/\mathbb{Z}^{q(k)} \,\middle|\, \exists \mathscr{E}\left(Star[\sigma^0(k)]\right) \ \ s.t. \{\theta_\alpha\}_k = \{\theta_{\alpha,k,\alpha+1}\}\right\}, \tag{1.64}$$

(we shall often drop the index k in $\{\theta_\alpha\}_k$ if it is clear from the geometrical context which vertex we are considering). Similarly, for a given $\Theta(k) \in \mathbb{R}^+$, we define the set of *vertex angle structures* associated with the conical angle $\Theta(k)$ as

$$\mathcal{V}ert_{q(k)}[\Theta(k)] \;:=\; \left\{ \{\theta_\alpha\}_k \in \mathcal{V}ert_{q(k)} \Big| \sum_{\alpha=1}^{q(k)} \theta_\alpha \,=\, \Theta(k) \right\}. \tag{1.65}$$

Remark 1.5 Since the automorphism group $Aut(T)$ of a given a triangulation (T_l, M) in $\mathcal{T}^{met}_{g,N_0}(M)$ fixes every vertex, we can consistently label the vertices of (T_l, M) by the index set $V := (1, \ldots, k, \ldots, N_0)$. This also induces a corresponding labeling on the vertex angles $\{\theta(h, k, j)\}$ of (T_l, M) which is coherent with the vertex star–grouping (1.62). Explicitly, if we define the ordering

$$\ldots \prec \{\theta_\alpha\}_k \prec \{\theta_\beta\}_{k+1} \prec \ldots, \tag{1.66}$$

i.e.,

$$\ldots \prec \{\theta_{\alpha,k,\alpha+1}\}_{\alpha=1}^{\alpha=q(k)} \prec \{\theta_{\beta,k+1,\beta+1}\}_{\beta=1}^{\alpha=q(k+1)} \prec \ldots, \tag{1.67}$$

we can label the vertex angles $\{\theta(h, k, j)\}$ by an index set Φ with cardinality $3N_2$. In particular we can associate to the polyhedral surface (T_l, M) the vertex angles vector

$$\begin{aligned} \mathcal{T}^{met}_{g,N_0}(M) &\longrightarrow (0, \pi)^{3N_2} \,=\, \oplus_{k=1}^{N_0}\; \mathcal{V}ert_{q(k)} \\ (T_l, M) &\longmapsto \overrightarrow{\vartheta}(T_l) := \left(\{\theta_\alpha\}_1, \ldots, \{\theta_\beta\}_{N_0}\right), \end{aligned} \tag{1.68}$$

and the *angle–vertex* incidence matrix

$$A_{T_l} \;:=\; [a_{kJ}]_{k \in V,\, J \in \Phi}\,, \tag{1.69}$$

where $a_{kJ} \equiv 1$ if the vertex angle $\angle_J$ is incident on the vertex $\sigma^0(k)$, $a_{kJ} \equiv 0$ otherwise. Namely, $a_{kJ} \equiv 1$ iff $\angle_J = \theta_{\alpha,k,\alpha+1}$ for some $\alpha \in (1, \ldots, q(k))$. Note in particular that the subset of triangulations $\mathcal{T}^{met}_{g,N_0}(M;\, \{\Theta(k)\})$ with a prescribed set of conical angles $\{\Theta(k)\}$ can be equivalently characterized by the condition

$$\mathcal{T}^{met}_{g,N_0}(M;\, \{\Theta(k)\}) = \left\{ (T_l, M) \in \mathcal{T}^{met}_{g,N_0}(M) \Big|\; A_{T_l}\, \overrightarrow{\vartheta}(T_l)^t = \overrightarrow{\Theta}(T_l)^t \right\}, \tag{1.70}$$

where $A^{T_l}\overrightarrow{\vartheta}(T_l)^t$ denotes the matrix product between A_{T_l} and the transposed (column) vector $\overrightarrow{\vartheta}(T_l)$, and where $\overrightarrow{\Theta}(T_l)$ denotes the given conical angles vector associated with (T_l, M), i.e.

$$\overrightarrow{\Theta}(T_l) := \;(\Theta(1), \ldots, \Theta(k), \ldots, \Theta(N_0))\,. \tag{1.71}$$

□

The general idea behind the introduction of $\mathcal{V}ert_{q(k)}$ and $\mathcal{V}ert_{q(k)}[\Theta(k)]$ is that these sets may serve as coordinate spaces for the conical angles $\{\Theta(k)\}$ associated with polyhedral surfaces. However this fails short of being a significant parametrization since the incidence order $q(k)$ may vary with the vertex $\sigma^0(k)$ and with the polyhedral surface $(T_l, M) \in \mathscr{T}^{met}_{g,N_0}(M)$ considered. To bypass this difficulty, it is convenient to *telescope* [4] the set $\{\theta_\alpha\}_{\alpha=1}^{q(k)}$ in $\mathscr{V}ert_{q(k)}$ as $q(k) \to \infty$, i.e. to inject $\{\theta_\alpha\}_{\alpha=1}^{q(k)}$ not in a particular $\mathscr{V}ert_{q(k)}$ but rather in the inductive limit of $\mathscr{V}ert_{q(k)}$ as $q(k) \to \infty$. This is easily realized by observing that there is a natural sequence of inclusions

$$\left(\theta_1, \dots, \theta_{q(k)}\right) \hookrightarrow \mathscr{V}ert_{q(k)} \hookrightarrow \dots \mathscr{V}ert_{q(k)+h} \hookrightarrow \mathscr{V}ert_{q(k)+h+1} \hookrightarrow \dots, \tag{1.72}$$

defined, (e.g., in the $\mathscr{V}ert_{q(k)} \hookrightarrow \mathscr{V}ert_{q(k)+1}$ case), by the $q(k)+1$ distinct injections

$$\left(\theta_1, \dots, \theta_{q(k)}\right) \xrightarrow{j_\beta^{q,q+1}} \left(\theta_1, \dots, \theta_{\beta-1}, 0, \theta_\beta, \dots, \theta_{q(k)}\right), \tag{1.73}$$

where $\beta = 1, \dots, q(k)+1$. Thus, independently from the particular vertex $\sigma^0(k)$ and triangulation (T_l, M) considered, we introduce the following characterization

Definition 1.17 The inductive limit of the family $\left\{\mathscr{V}ert_q, j_\beta^{q,q+1}\right\}$, as $q \to \infty$, defines the space of *vertex angle structures*

$$\mathscr{V}ert := \lim_{q \to \infty} \left\{\mathscr{V}ert_q \hookrightarrow \mathscr{V}ert_{q+1}, \; j_\beta^{q,q+1}\right\} \simeq [0, \pi]^\infty, \tag{1.74}$$

associated with the set of Euclidean vertex stars $\left\{\mathscr{E}\left(Star[\sigma^0(k)]\right)\right\}$ of polyhedral surfaces in $\mathscr{T}^{met}_{g,N_0}(M)$.

Explicitly, we may think of a sequences $\{\theta_\alpha\}_{\alpha=1}^{q(k)} \in \mathscr{V}ert_{q(k)}$, associated with $\mathscr{E}\left(Star[\sigma^0(k)]\right)$, as the coordinates of the corresponding conical structure in the local chart $\mathscr{V}ert_{q(k)}$ of $\mathscr{V}ert$, i.e.

$$\oplus_{k=1}^{N_0} \; \mathscr{V}ert_{q(k)} \hookrightarrow \mathscr{V}ert. \tag{1.75}$$

We can locally model the geometrical realizations of the local chart $\mathscr{V}ert_{q(k)}$ in terms of spherical polygons (Fig. 1.14). This will characterize a natural polygonal bundle over $\mathscr{V}ert$ which, as described in the introductory remarks, will play the role of the cotangent bundle to $\mathscr{T}^{met}_{g,N_0}(M)$. To carry out such a construction explicitly, let us consider the link of $\sigma^0(k)$ in $Star[\sigma^0(k)] \in (T_l, M) \in \mathscr{T}^{met}_{g,N_0}(M)$,

$$link(k) := link_{(T_l,M)}\left(\sigma^0(k)\right) \simeq \mathbb{S}^1, \tag{1.76}$$

viz. the union of all edges $\sigma^1(h_\alpha, h_{\alpha+1})$ in $Star[\sigma^0(k)]$ satisfying $\sigma^1(h_\alpha, h_{\alpha+1}) \cap \sigma^0(k) = \emptyset$, (see Theorem 1.1). We can naturally map $link(k)$ to a closed rectifiable piecewise–smooth Jordan curve $c_{(k)}$ on the unit sphere

$$\begin{aligned} c_{(k)} \; : \; [0,1] &\longrightarrow link(k) \hookrightarrow \mathbb{S}^2, \\ t \;\; &\longmapsto \;\; x(t) \hookrightarrow c_{(k)}(t), \end{aligned} \tag{1.77}$$

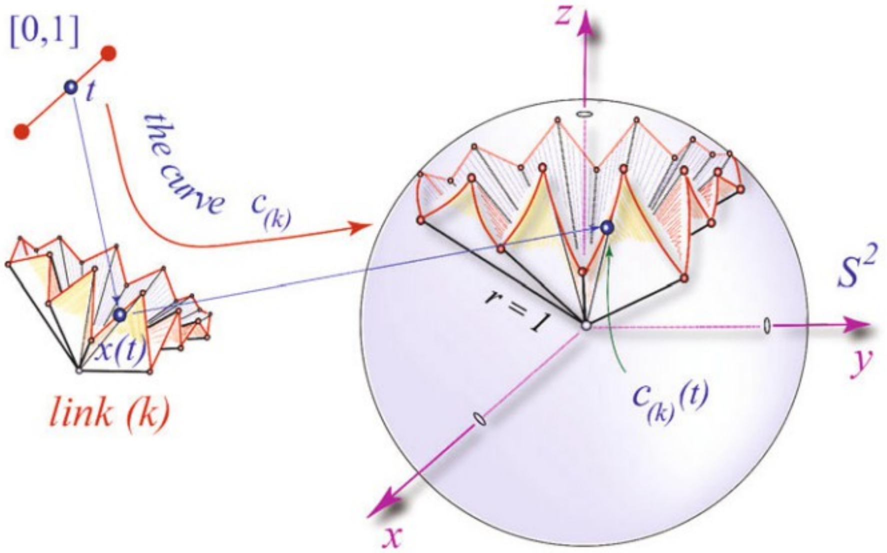

Fig. 1.14 The image of the curve (1.77) is characterized by the intersection, in ($\mathbb{R}^3$; O, x, y, z), of a geometric realization of $\mathcal{E}\left(Star[\sigma^0(k)]\right)$ with a unit sphere $\mathbb{S}^2$ centered in the origin $O \equiv \sigma^0(k)$. This intersection is well-defined up to a natural $SO(3)$ action

where $x(t) \in link(k)$, and where $c_{(k)}(0) = c_{(k)}(1)$. The image $c_{(k)}([0,1])$ of the curve $c_{(k)}$ can be explicitly constructed by intersecting, in $\left((\mathbb{R}^3, \times) \cong \mathfrak{so}(3);\, O, x, y, z\right)$, a geometrical realization of the given Euclidean structure $\mathcal{E}\left(Star[\sigma^0(k)]\right)$ with a unit sphere $\mathbb{S}^2$ centered in the origin O, (see Fig. 1.14).This generates a cone whose intrinsic metric depends only on the length $|c_{(k)}|$ of the curve $c_{(k)}$, (see Fig. 1.15), and we have the characterization, (see also [5])

Lemma 1.1 *The conical angle $\Theta(k)$ at the vertex $\sigma^0(k)$ is the length $|c_{(k)}|$ of the Jordan curve* (1.77). ☐

The mapping (1.77) also induces on the curve $c_{(k)}$ the decoration defined by the $q(k)$ points $\{p_\alpha\}$ images of the vertices $\{\sigma^0(h_\alpha)\} \in link(k)$. In particular the vertex angle $\theta_\alpha(k)$ associated with the triangle $\sigma^2(h_\alpha, k, h_\alpha)$ provides the length

$$\left|c_{(k)}(\alpha, \alpha+1)\right| := \theta_\alpha(k), \tag{1.78}$$

of the arc of geodesic $c_{(k)}(\alpha, \alpha+1)$ between the point p_α and $p_{\alpha+1}$.

Remark 1.6 The $\{c_{(k)}(\alpha, \alpha+1)\}$ are geodesic arcs since they are defined by the intersection between $\mathbb{S}^2$ and planes passing through the center of $\mathbb{S}^2$.

M. Kapovich and J. Millson have studied in depth the properties of the set of spherical polygons and of the associated moduli spaces, (see [9] and references therein). Here we exploit a few elementary aspects of their analysis playing an important role in our framework. We start by recalling the

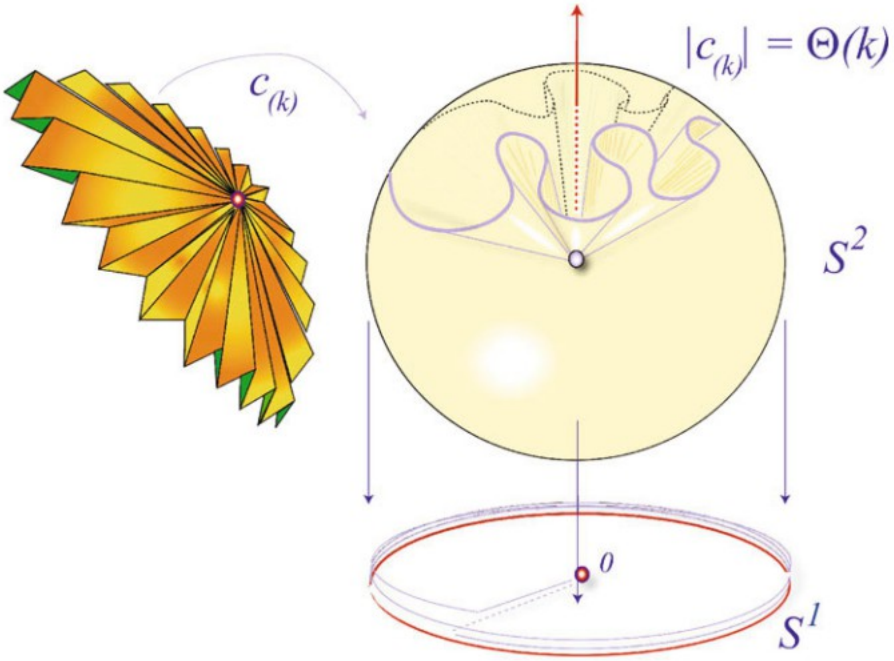

Fig. 1.15 The cone associated with the curve $c_{(k)}$ on the unit sphere S^2. The intrinsic metric of such a cone depends only on the length $|c_{(k)}|$ of the curve $c_{(k)}$, (to emphasize this we have rounded the image of the curve). This length is the conical angle $\Theta(k)$ at the vertex. If the cone generated by $c_{(k)}$ is cut open along a generator and unfolded into the (complex) plane, the curve $c_{(k)}$ covers the circle S^1 and $\Theta(k)$ gives the length (with multiplicity) of the circuitation around S^1

Definition 1.18 (Labeled Spherical Polygons [9]) A $q(k)$–tuple of oriented geodesic arcs $c_{(k)}(\alpha, \alpha+1) \in \mathbb{S}^2$ of lengths

$$0 < \left|c_{(k)}(\alpha, \alpha+1)\right| < \pi, \tag{1.79}$$

such that the end–point of the geodesic segment $c_{(k)}(\alpha, \alpha+1)$ coincides with the initial point of the geodesic segment $c_{(k)}(\alpha+1, \alpha+2)$ defines a closed spherical $q(k)$–gon

$$P_{(k)}(c_{(k)}) := \{c_{(k)}(\alpha, \alpha+1)\}. \tag{1.80}$$

□

Let us remark that the labeling here refers to the fact that the geodesic segments $c_{(k)}(\alpha, \alpha+1)$ defining the edges of the polygon are path–ordered. Moreover, as long as $0 < \left|c_{(k)}(\alpha, \alpha+1)\right| < \pi$, the geodesic arcs $c_{(k)}(\alpha, \alpha+1)$ are equivalently characterized by the ordered sequence of $q(k)$ points $\{p_\alpha\}$, and we can identify $P_{(k)}(c_{(k)}) = \{p_\alpha\}$. The more general case $\left|c_{(k)}(\alpha, \alpha+1)\right| \in [0, \pi]$, which corresponds to spherical polygons associated with degenerating triangles in $\mathscr{E}(Star[\sigma^0(k)])$, can be handled with a detailed analysis of the possible singular polygonal configurations associated with the triangle degenerations. In order to avoid the complications due to these singular configurations, we shall consider uniquely the generic case under the restriction (1.79).

There is a natural $SO(3)$ diagonal action on the spherical polygons $P_{(k)}(c_{(k)})$ defined by [9]

$$g \cdot P_{(k)}(c_{(k)}) := \left(g \cdot \overrightarrow{Op}_1, \ldots, g \cdot \overrightarrow{Op}_{q(k)} \right), \tag{1.81}$$

where $g \in SO(3)$ and where $\overrightarrow{Op}_\alpha$ is the vector in $\mathbb{R}^3$ connecting the center O of $\mathbb{S}^2$ with the point $p_\alpha \in \mathbb{S}^2$. Denote by $\mathscr{P}_{q(k)}(\mathbb{S}^2)$ the space of spherical polygons [9]. Since, according to the above remarks, there is a one–to–one correspondence between $SO(3)$ equivalence classes of spherical polygons and polyhedral cones modelling a neighborhood of the vertex $\sigma^0(k)$ it is natural to introduce the

Definition 1.19 The quotient space of $\mathscr{P}_{q(k)}(\mathbb{S}^2)$ by the $SO(3)$ action described above,

$$\mathscr{Q}_{q(k)}(\mathbb{S}^2) := \mathscr{P}_{q(k)}(\mathbb{S}^2) \,/\, SO(3), \tag{1.82}$$

defines the space of *polyhedral cones* over the vertex $\sigma^0(k) \in (T_l, M)$.

In order to fix the $SO(3)$ gauge freedom in $\mathscr{P}_{q(k)}(\mathbb{S}^2)$ in (1.82) we start by remarking that the configuration space of spherical polygons with given side-lengths $\{\left|c_{(k)}(\alpha, \alpha+1)\right|\}$ can be characterized as the set

$$\mathscr{P}_{(k)}(\mathbb{S}^2; \{\left|c_{(k)}(\alpha, \alpha+1)\right|\}) := \tag{1.83}$$

$$\left\{ \{p_\alpha\} \in \mathbb{S}^{2\,q(k)} \mid \overrightarrow{Op}_\alpha \cdot \overrightarrow{Op}_{\alpha+1} = \cos\, \left|c_{(k)}(\alpha, \alpha+1)\right| \right\},$$

where $\cdot$ denotes the scalar product in $\mathbb{R}^3$. In particular, we can always choose the geodesic arc $c_{(k)}(q(k), 1)$ of the spherical polygon $P_{(k)}(c_{(k)})$ to lie in the (O, x, y)–plane $\subset \mathbb{R}^3$. Let $(\overrightarrow{\varepsilon}_1, \overrightarrow{\varepsilon}_2)$ be orthonormal basis vectors in (O, x, y), and let us denote by p_0 the midpoint of the geodesic arc $c_{(k)}(q(k), 1)$. We take p_0 to be the origin of the piecewise geodesic path $c_{(k)}$. Rotate the polygon $P_{(k)}(c_{(k)})$ around the origin in the (O, x, y)–plane so that $\overrightarrow{Op}_0 \equiv \overrightarrow{\varepsilon}_1$. This fixes the position of the vector $\overrightarrow{Op}_1$ according to (Fig. 1.16)

$$\overrightarrow{Op}_1 = \cos \frac{\left|c_{(k)}(q(k), 1)\right|}{2} \overrightarrow{\varepsilon}_1 + \sin \frac{\left|c_{(k)}(q(k), 1)\right|}{2} \overrightarrow{\varepsilon}_2. \tag{1.84}$$

In such a representation there is a natural map defined by

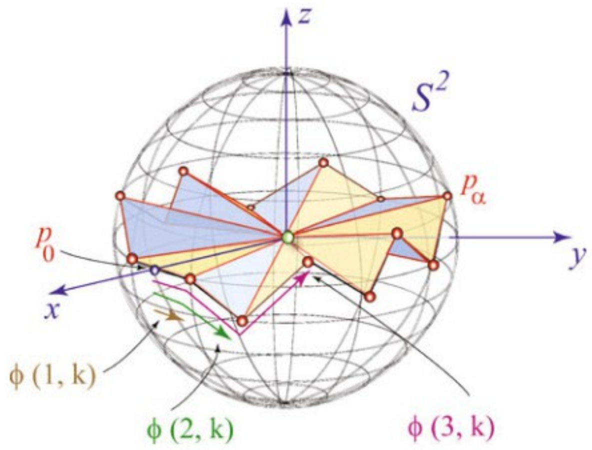

Fig. 1.16 The decoration defined by the $q(k)$ points $\{p_\alpha\}$ images of the vertices $\{\sigma^0(h_\alpha)\} \in link(k)$. The arc–length computed from the origin p_0, defined by choosing the geodesic arc $c_{(k)}(q(k),1)$ of the spherical polygon $P_{(k)}(c_{(k)})$ to lie in the (O, x, y)–plane $\subset \mathbb{R}^3$, characterizes the map $\phi(\alpha,k)$ associated with the point p_α

$$\begin{aligned}\pi_{\phi(k)} \,:\, \mathscr{P}_{q(k)}(\mathbb{S}^2) \;&\longrightarrow\; \left[\mathbb{S}^1/2\pi\right]^{q(k)} \\ \{p_\alpha\} \;&\longmapsto\; \{\phi(\alpha,k)\}_{\alpha=1}^{q(k)} \;:=\; \left\{\frac{\left|c_{(k)}(0,\alpha)\right|}{\Theta(k)}\right\}_{\alpha=1}^{q(k)},\end{aligned} \tag{1.85}$$

where $\left|c_{(k)}(0,\alpha)\right|$ denotes the length of the piecewise geodesic arc of curve $c_{(k)}(0,\alpha)$ between the midpoint $p_0 \in c_{(k)}(q(k),1)$, (the origin of the piecewise geodesic path $c_{(k)}$), and the polygon vertex p_α. Note that the normalization to the total conical angle $\Theta(k) = \sum_{\alpha=1}^{q(k)} \theta_\alpha$ is well–defined since, according to Lemma 1.1,

$$\Theta(k) = \sum_{\alpha=1}^{q(k)} \left|c_{(k)}(\alpha,\alpha+1)\right|. \tag{1.86}$$

Moreover, we have

$$\phi(\alpha+1,k) \;=\; \phi(\alpha,k) \;+\; \frac{\left|c_{(k)}(\alpha,\alpha+1)\right|}{\Theta(k)}\,, \quad \alpha = 1,\ldots,q(k)-1, \tag{1.87}$$

which directly follows from the definition of $\phi(\alpha,k)$, and

$$\phi(1,k) \;=\; \phi(q(k),k) + \frac{\left|c_{(k)}(q(k),1)\right|}{\Theta(k)} \;-\; 1\,, \quad \alpha = q(k), \tag{1.88}$$

which is a consequence of the identity

$$\frac{\left|c_{(k)}(q(k),1)\right|}{\Theta(k)} \;=\; 1 \;-\; \sum_{\alpha=1}^{q(k)-1} \left(\phi(\alpha+1,k) - \phi(\alpha,k)\right). \tag{1.89}$$

Conversely,

$$\left|c_{(k)}(\alpha,\alpha+1)\right| \;=\; \Theta(k)\;[\phi(\alpha+1,k)-\phi(\alpha,k)]\;, \quad \alpha=1,\ldots,q(k)-1, \tag{1.90}$$

and

$$\left|c_{(k)}(q(k),1)\right| \;=\; \Theta(k)\;[\phi(1,k)-\phi(q(k),k)+1]\,. \tag{1.91}$$

It follows that, relative to the framing (1.84), we can uniquely represent any spherical polygon, with given edge lengths $\{\left|c_{(k)}(\alpha,\alpha+1)\right| \;=\; \theta_\alpha(k)\}$, in terms of the set

$$\Xi_{\{\phi(\alpha,k)\}} \;:=\; \left\{\{\phi(\alpha+1,k)-\phi(\alpha,k)\} \mid \{\phi(\alpha,k)\}\in\left[\mathbb{S}^1/2\pi\right]^{q(k)}\right\}. \tag{1.92}$$

The quotient map

$$\mathscr{P}_{(k)}(\mathbb{S}^2;\{\left|c_{(k)}(\alpha,\alpha+1)\right|\}) \longrightarrow \frac{\mathscr{P}_{(k)}(\mathbb{S}^2;\{\left|c_{(k)}(\alpha,\alpha+1)\right|\})}{SO(3)}, \tag{1.93}$$

associating to a polygon $P_{(k)}(c_{(k)})$ its representative in the framing (1.84), shows that (1.90) and (1.91) provide a homeomorphism

$$\Xi_{q(k)}(\{\phi(\alpha,k)\}) \;\longrightarrow\; \frac{\mathscr{P}_{(k)}(\mathbb{S}^2;\{\left|c_{(k)}(\alpha,\alpha+1)\right|\})}{SO(3)}, \tag{1.94}$$

which characterizes $\Xi_{q(k)}(\{\phi(\alpha,k)\})$ as a slice for the $SO(3)$ action on the spaces of spherical polygons with given edge–lengths. It follows that we can represent the space of polyhedral cones over the vertex $\sigma^0(k)$, $\mathscr{Q}_{q(k)}(\mathbb{S}^2)$, according to

$$\mathscr{Q}_{q(k)}(\mathbb{S}^2) \simeq \Xi_{q(k)}(\{\phi(\alpha,k)\}). \tag{1.95}$$

In particular, if $V(T)$ denotes the vertex set of (T_l, M), then the map

$$\begin{aligned} \pi_k \;:\; V(T)\,\times\,\mathscr{V}ert &\longrightarrow \sqcup_{k=1}^{N_0}\mathscr{Q}_{q(k)}(\mathbb{S}^2) \qquad (1.96)\\ (\sigma^0(k),\,\{\theta_\alpha\}_k) &\longmapsto (\{\phi(\alpha,k)\},\,\{\phi(\alpha+1,k)-\phi(\alpha,k)\}=\{\theta_\alpha\}_k)\,, \end{aligned}$$

is a choice of a polyhedral cone (Fig. 1.17) having, near the given vertex $\sigma^0(k)$, the same intrinsic Euclidean structure $\mathscr{E}\left(Star[\sigma^0(k)]\right)$ defined by the vertex angles $\{\theta_\alpha\}_k$. Thus, it is natural to introduce the following

Definition 1.20 The mapping

$$(T_l, M)|_{\sigma^0(k)} \longmapsto \mathscr{Q}_{q(k)}(\mathbb{S}^2), \tag{1.97}$$

defines the polyhedral cotangent cone to (T_l, M) at the vertex $\sigma^0(k)$.

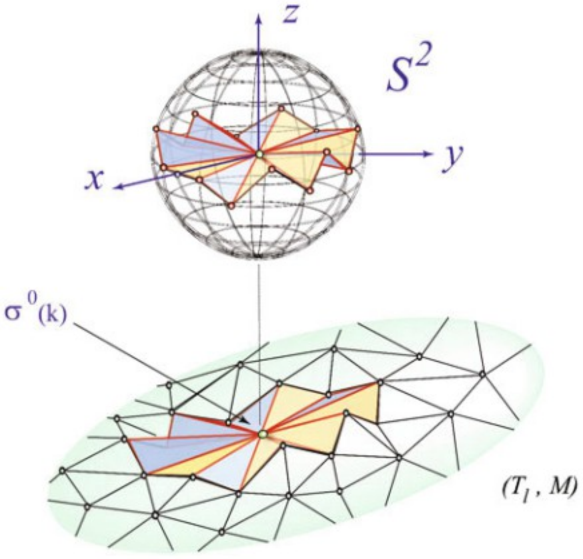

Fig. 1.17 $\mathscr{Q}_{q(k)}(\mathbb{S}^2)$ characterizes the polyhedral cotangent cone to (T_l, M) at the vertex $\sigma^0(k)$

Roughly speaking, the identification of $\mathscr{Q}_{q(k)}(\mathbb{S}^2)$ with the polyhedral cotangent space at $\sigma^0(k)$ is also justified by noticing that the polyhedra in $\mathscr{Q}_{q(k)}(\mathbb{S}^2)$ are side–wise dual to the tangent space [21] over the corresponding open triangle $\breve{\sigma}^2(\alpha, k, \alpha+1)$.

In a well–defined sense, the pair $\left(\sigma^0(k),\ \mathscr{Q}_{q(k)}(\mathbb{S}^2)\right)$ comes decorated with *coordinates*: $(\{\phi(\alpha, k)\})$ associated with the vertex $\sigma^0(k)$, and $(\{\phi(\alpha+1, k) - \phi(\alpha, k)\})$ associated with the choice of a polyhedral cone in $\mathscr{Q}_{q(k)}(\mathbb{S}^2)$ over $\sigma^0(k)$. For later use it is useful to formalize this remark into the

Definition 1.21 The map

$$(\sigma^0(k),\ \mathscr{Q}_{q(k)}(\mathbb{S}^2)) \longmapsto (\{\phi(\alpha, k)\},\ \{\phi(\alpha+1, k) - \phi(\alpha, k)\}), \tag{1.98}$$

defines the polyhedral cotangent coordinates associated with the vertex $\sigma^0(k) \in (T_l, M)$. In particular the assignment of a 1-form in $(\sigma^0(k),\ \mathscr{Q}_{q(k)}(\mathbb{S}^2))$,

$$\nu_k := \sum_{\alpha=1}^{q(k)} (\phi(\alpha, k) - \phi(\alpha+1, k))\ d\left(\phi(\alpha, k)\right), \tag{1.99}$$

(recall that $\phi(\alpha+1, k) := \phi(1, k)$ for $\alpha = q(k)$), characterizes the choice of a polyhedral cone in $(\sigma^0(k),\ \mathscr{Q}_{q(k)}(\mathbb{S}^2))$.

Note that the 1-form ν_k is intrinsically defined over $\mathscr{Q}_{(k)}(\mathbb{S}^2)$, in the sense that it does not depend on the chosen origin point p_0 used for characterizing the polygonal arc-length map $\phi(\alpha, k)$. The above characterization of the polyhedral cotangent cone to (T_l, M) at the vertex $\sigma^0(k)$ can be extended (Fig. 1.18) to $\mathscr{T}^{met}_{g,N_0}(M)$ according to the

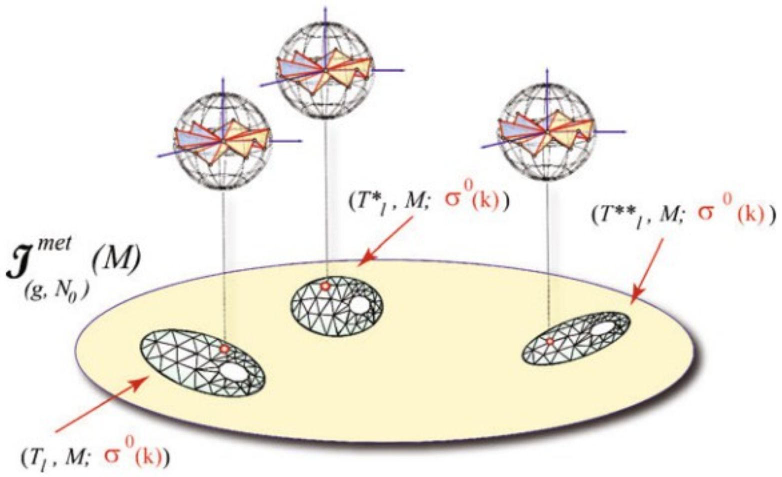
$(T^*_l, M;\ \sigma^0(k))$
$(T^{**}_l, M;\ \sigma^0(k))$
$\mathcal{J}^{met}_{(g,N_0)}(M)$
$(T_l, M;\ \sigma^0(k))$

Fig. 1.18 The polyhedral cotangent cones $\mathcal{Q}_{q(k)}(\mathbb{S}^2)$ to the polyhedral surface (T_l, M) at the vertex $\sigma^0(k)$, characterizes as $(T_l, M; \sigma^0(k))$ varies in $\mathcal{T}^{met}_{g,N_0}(M)$, a (polygonal) circle bundle over $\mathcal{T}^{met}_{g,N_0}(M)$. Since there are N_0 labeled vertices for each (T_l, M), there are N_0 such a bundle over $\mathcal{T}^{met}_{g,N_0}(M)$

Lemma 1.2 *Let* $C_k := (\{\phi(\alpha, k)\}, \{\phi(\alpha + 1, k) - \phi(\alpha, k)\})$ *denote the generic polyhedral cone in* $\mathcal{Q}_{q(k)}(\mathbb{S}^2)$*, then the mapping*

$$\begin{aligned} \mathcal{T}^{met}_{g,N_0}(M) &\longrightarrow \otimes_{k=1}^{N_0} \mathcal{Q}_{q(k)}(\mathbb{S}^2) \\ (T_l, M) &\longmapsto (C_1 \ldots, C_k, \ldots, C_{N_0}) \end{aligned} \tag{1.100}$$

defined by the N_0 *polyhedral cotangent cones at the (labelled vertices of the) the polyhedral surfaces* $(T_l, M) \in \mathcal{T}^{met}_{g,N_0}(M)$*, defines* N_0 *circle bundles* $\mathcal{Q}_{(k)}$ *over* $\mathcal{T}^{met}_{g,N_0}(M)$*.*

Proof The $\mathbb{S}^1$ fiber of $\mathcal{Q}_{(k)}$ is naturally associated with the spherical polygon defining the polyhedral cones in $\mathcal{Q}_{(k)}$. Moreover, the definition of $\phi(k, \alpha)$ easily implies that the bundle projection

$$\begin{aligned} pr_k : \mathcal{Q}_{(k)} &\longrightarrow \mathcal{T}^{met}_{g,N_0}(M) \\ C_k &\longmapsto C_k|_{\sigma^0(k) \in (T_l, M)}\,, \end{aligned} \tag{1.101}$$

is continuous when restricted to open sets of (T_l, M) in $\mathcal{T}^{met}_{g,N_0}(M)$. By considering the inductive limit Ξ_∞ of the family of polyhedral spaces $\left\{\Xi_{\{\phi(\alpha,k)\}}\right\}$ as $q(k) \to \infty$, and by considering the polyhedral cotangent cone $\mathcal{Q}_{(k)}$ as a point in Ξ_∞, it is also easy to see that pr_k is well–defined under the *flip move* (1.51), (we need to use Ξ_∞ because the incidence number $q(k)$ changes under flip moves). ■

CNAME

WWW ghs.google.com

pdfdrive.com

→ Eberhard Zeidler

ezeidler@mis.mpg.de

ironia

J.S. Andrew Hamilton

~~www~~ jila.colorado.edu/~ajsh/

Andrew.Hamilton@colorado.edu

Réciprocité de Legendre

a) archive.numdam.org/item/NAM_1872_2-11__354_0/

b) rzuser.uni-heidelberg.de/~hb3/fchrono.html.

Algebraic Geometry Gortz.

Problems in plane and solid geometry
Prasolov

365
[illegible]
→ 613

The Fractal Geometry of Nature.

Eberhard Knobloch

---

sanet.st
sanet.ws

Cyberduck Programe → Google Drive

cyberduck.io

---

ironia.

03, 27, 39, 46, 47, 60

20201027.

backblaze.com

$0,87 \times 0,87 = 7,569$

$\sqrt{8} = 2,828...$

$\sqrt{2} = 1,414...$

$5/15 = 0,\overline{3}$

landandseacollection.com

Nietzsche

Steinofen   forno de pedra

✓ Fazer TXT com o nome de todos os arquivos 08, 09 e 10.

sort

dir /s /w /p /o:n>____.txt

/h - hidden files

dir /s/w/p/o:n>____.txt

dir >prn = print

dir > \relatorio\dir.doc

dir c:\*.*/s/w/p/o:>____.txt

eugene381@comcast.net

String Theory Polchinski
2 Volumes

silêncio obsequioso.

herético.

papolatria.

Baruch de Espinoza
(1632-1677)

Leibniz conheceu Nicolaus Steno.

1684 Nova Methodus pro
Maximis et Minimis

(Calculo diferencial)
Sec XVIII disputa Leibniz x Newton.

20201029 - Laplace

Imp:

C:\PC_Outlo\Importantíssimo_Books_38_2020\Distance Geometry.pdf

Google Drive

Shift + Z

Latria → idolatrar (Deus)

Dulia – venerar

Munich Institute of Mathematics of Zurich

~~Romanos~~ 8,31

Tessalonicense Cap 2, 6-7 Catecom
~~Taquetom~~.

Excentricidades.

Satisfazer Pessoas → devedor.

Conics and Cubics Bix

Georg Glaeser

8 unidades astronômicas da lente (pouco menos que a distância do sol a Saturno).

Hellmuth Stachel

Steiner's Curvature Centroid

Flexible quadrangular mesh after Hellmuth Stachel.

Frank Hettlich

007 - Duran Duran
Aha
Tina Turner
Adele Skyfall

Lc 14, 1-6
hidrópico

' #

20201109

Schuldig = Culpado

Leibniz $x^{-2} + x^{-\frac{3}{2}} = 4$

Leibniz Biografia
Autograzzs

google chrome.

math.odu.edu/~jhh/Volume-1.pdf
2

Selected Papers Feller

Esquema Ponzi USA 1910

Books/ Books_2020_05_a_07

172.217.28.131
p. 80.

08/11 — Morreu Vanuza
— Aniversário 75 anos de Antonio Marcos
— Mário Zan (morte) 14 anos

09/11 Fundação Congregação dos Redentoristas
09/11/1732 (Ligório) Pe

Dia da Catedral de Latrão
→ catedral do papa dedicada
ao Sr. Jesus Cristo (Salvador)
09/11 e São João Batista.

Recordar data de fundação. → 278 anos

09/11 Hoje Jesus expulsou os vendilhões do templo.
3 dias → reconstruiria o templo. (seu corpo)
ressurreição

' ##

Torno VDF.

Misturador rotativo e horizontal
1952

→ Books-2020-22-Updated-20200713

Google 4,3 milhões Gb/dia

Sean M. Carroll

Fundamentals of Calculus
Morris

→ 10321 ⇒

Archimedes

Volume da Esfera = 4 x volume do cone com raio e altura iguais ao raio da esfera.

/+)

% - ' #.
9:5;9 9 9/ 9 - E ? , >
< 7) % + C))!
' $ > & -

at the quantum level that manifests in two states, like the *spin* of an electron which comes as either 'spin up' or 'spin down'[24]. These two quantum states comprise a *qubit,* designated by, [0] and [1], respectively, and may be combined into what is called a *superposition* of states:

$$\psi = c_1[0] + c_2[1]. \tag{27}$$

Here, c_1 and c_2 are *complex* numbers, such that $|c_1|^2$ represents the probability of the electron being in state [0], and $|c_2|^2$ represents the probability of the electron being in state [1]. This is another example of where complex numbers are vital in dealing with problems of the real world.

As we are dealing with probabilities, it is required that: $|c_1|^2 + |c_2|^2 = 1$. Under this proviso, taking all possible complex values c_1 and c_2, we have just created a two-dimensional (complex) Hilbert space. All the vectors, ψ, in this space are represented by the infinitely many different allowable values of c_1 and c_2 in eq. (27).

These vector states can essentially be geometrically represented by what is known as the *Bloch sphere*, as in Fig. 7.5 (left)[25]. When a measurement of the qubit is made, its state will be determined to be either [0] or [1] (due to the collapse of the wave function), and whose value effectively depends on the 'latitude' of the state vector's position on the Bloch sphere.

A quantum computer with 300 qubits can in some sense exist in 2^{300} states *at the same time*. This is more than the number of atoms in the observable Universe[26]. Quantum computers are still in the research and development stage and are suited mainly to particular types of problems (Fig. 7.5 (R)). According to computer scientist Scott Aaronson, the biggest advantages are for "(1) simulating Quantum Physics and Chemistry themselves, which has tremendous potential applications to medicine, materials science, and elsewhere, and (2) breaking the public-key cryptography that's currently used to secure the Internet."

We have just had a peek into the rather mysterious future of computing. Furthermore, quantum computer scientist Seth Lloyd proposed in his book,

---

[24] The notion of spin is derived from the angular momentum of a rotating object, but at the quantum level for a charged particle (like an electron), spin relates to the magnetic field it generates *as if it were* a spinning sphere of charge.

[25] On the Bloch sphere, for θ and ϕ as depicted in Fig. 7.5 (L), there is the geometric representation for the qubit state ψ as:

$$\psi = \sin\frac{\theta}{2}[0] + e^{i\phi}\cos\frac{\theta}{2}[1].$$

The sphere is named after Swiss-American physicist Felix Bloch (1905–1983). Note our familiar friend, the complex exponential $e^{i\phi}$ from Chapter 3.

[26] Note that $2^{300} \approx 10^{90}$, and as mentioned in Chapter 2, the number of atoms in the observable Universe is around 10^{80}.

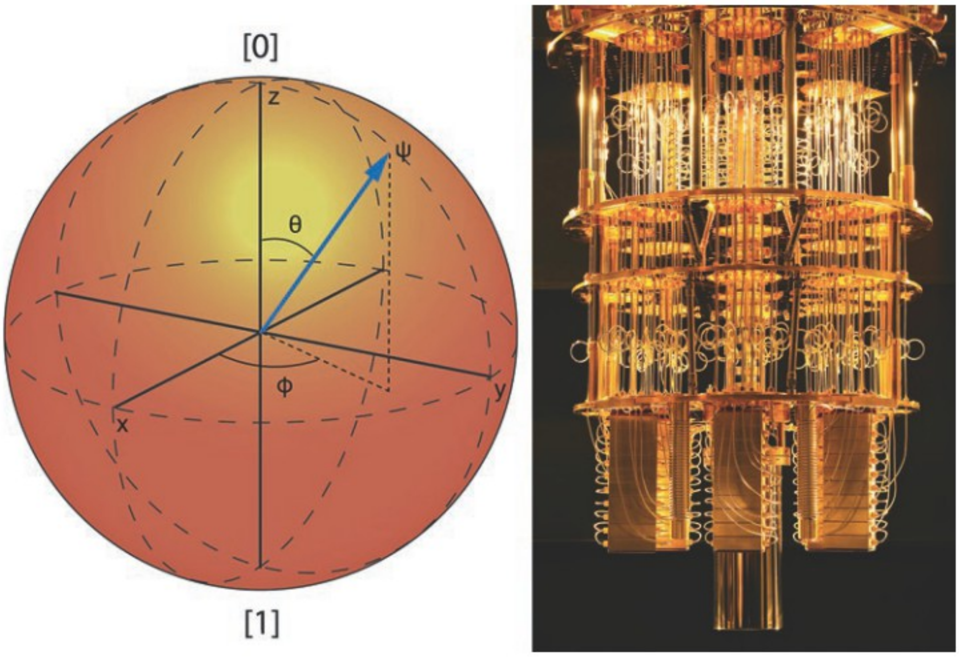

Figure 7.5: (L) The Bloch sphere geometric depiction of a state vector of a qubit. (R) Part of the cooling system and cabling of the rather beautiful IBM Q quantum computer. The chip is located in the canister at the bottom, which must be kept near absolute zero. (Images courtesy of (L) Katy Metcalf, (R) IBM Research *Flickr* CC BY-ND 2.0.)

Programming the Universe, that the entire Universe is a quantum computer that "computes its own behavior." This idea is similar to the digital Physics of Edward Fredkin mentioned in Chapter 5, in the context of the Universe being a cellular automaton.

Related to this latter idea is the notion that in order to simulate behavior in the microscopic quantum world effectively, it will take a quantum computer to do so. In the words of theoretical physicist John Preskill, "…might we already have persuasive evidence that Nature performs tasks going beyond what can be simulated efficiently by classical computers?" This is the quantum computing version of the expression, "It takes a thief to catch a thief."

Preskill proposed the term 'quantum supremacy' in 2012 to denote the point in time when a quantum computer will be able to perform a computing task that would not be possible on a classical computer *in a reasonable period of time*, irrespective of whether or not the computation was in any way useful.

The achievement of this goal of quantum supremacy has been claimed by an AI team at Google, published in an article in *Nature* (October 23, 2019), who used their

53-qubit quantum computer chip called Sycamore. They have performed a calculation in roughly 200 seconds that they argued would have taken the most powerful supercomputer around 10,000 years to perform. It must be added, however, that the problem selected was designed very specifically to favor a quantum computer and is known to be difficult for a classical computer. The authors claimed[27]:

> "Quantum processors based on superconducting qubits can now perform computations in a Hilbert space of dimension $2^{53} \approx 9 \times 10^{15}$, beyond the reach of the fastest classical supercomputers available today. To our knowledge, this experiment marks the first computation that can be performed only on a quantum processor."

On the other hand, in an IBM research blog[28], three of their scientists asserted that the problem Google solved could have been done in 2.5 days on one of its classical supercomputers, which is rather quicker than Google's estimate of 10,000 years. Therefore, "… [Google's assertion] should not be viewed as proof that quantum computers are 'supreme' over classical computers." Needless to say, IBM and Google are fierce rivals in the field of quantum computing.

Part of the problem is the very notion of quantum supremacy and its interpretation. Moreover, whether the computer time is 10,000 years or 2.5 days is not really the point. The IBM authors go on to state that, "quantum computers will never reign 'supreme' over classical computers, but will rather work in concert with them, since each have their unique strengths." Indeed, some problems are more amenable to classical computers and some are more suitable for a quantum computer, though the line between the two types of problems is often a gray area.

What the Google result does demonstrate is control over 53 qubits that are inherently unstable and were held in a quantum superposition state long enough to do a complex calculation. So, achieving this sort of result is an impressive technical achievement and a scientific milestone pointing the way to the future.

For those interested, there are now facilities for anyone from the general public who wishes to learn how to create a program on a quantum computer and to execute it[29]. A comprehensive account of quantum computing, together with much related subject matter, can be found in the fine monograph by Scott Aaronson in the Bibliography.

---

[27] F. Arute *et al.*, Quantum supremacy using a programmable superconducting processor, *Nature*, 574, 505-510, 2019.

[28] E. Pednault, J. Gunnels, J. Gambetta, On Quantum Supremacy, Oct. 21, 2019, IBM Research blog. This announcement pre-dates the Google publication in *Nature*, as the Google announcement was leaked publicly prior to publication.

[29] D-Wave Leap: https://www.dwavesys.com/take-leap
IBM Q Experience: https://www.research.ibm.com/ibm-q/Microsoft Quantum Development Kit: https://www.microsoft.com/en-us/quantum/development-kit

/ + 7 9:5;9 9 9/ 9 - E H) E
E < ! :))) ...)!

' -

Kathryn Mc Cullers

Christy Orshak.

— 7 Combinações:

Par ←

14 + 17 + 19 = 50 P I I

15 + 16 + 19 = 50 I P I

Par & Par = 16 + 16 + 18 = 50 ✡ P P P

16 + 17 + 17 = 50 P I I

—

13 + 18 + 19 = 50 (I P I

12 + 19 + 19 = 50 P I I

~~9~~ pares (8 pares) 3 = P I I

10 ímpares. 2 = I P I

1 = P P P ⇐

outros:

⇒ 14 + 18 + 18 = 50 P P P

+

↳ 2 = P P P

/+ + NR) S -

20201121.

Radian-20201121.ggb.

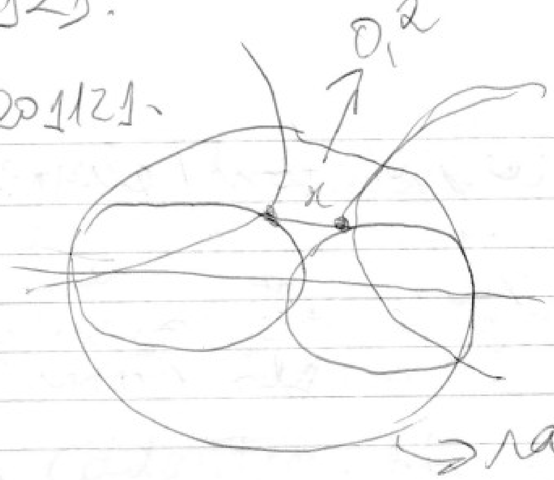

→ raio 0,5

$$\begin{array}{r} \overset{3}{0,28} \\ \times \; 4 \\ \hline 1,12 \end{array} \quad + \quad \begin{array}{r} 0,2 \times \\ 1,4 \\ \hline 0,8 \end{array} = 1.92$$

perímetro

0,28 → outro 0,31 = 0,03

$0,03 \,|\, 2$ → $\underline{0,015}$

Pontos B e E

✓ ✓

(0,5, 0) (0,5, 0,5)

→ equations

Circle (B, E)

$(x-0.5)^2+y^2=\underline{0,25}$

$$\sqrt{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2}$$

R (0.3, 0.4)

B_1 (0,26, 0.43)

$$\sqrt{(0.3-0.26)^2+(0.4-0.43)^2}$$

$(0.04)^2 \quad (-0.03)^2$

$0.0016 + 0.0009 =$

$\sqrt{0.0025} = 0.05$

B_1 (0.26, 0.43)

R (0.3, 0.4)

$$\sqrt{(0.26-0.3)^2+(0.43-0.4)^2}$$

$(-0.04)^2 \quad (0.03)^2$

$0.0016 + 0.0009$

$\sqrt{0.0025} = 0.05$

f_1 r // m (não é paralelo)

R (0,3, 0.4)

A_1 (0.3, 0.46)

Distância:

$$\sqrt{(0{,}3-0.3)^2 - (0{,}4-0.46)^2}$$

$$0 \quad - \quad (-0.06)^2$$

$$0 \quad - \quad 0.0036$$

$$\underline{\underline{-0{,}0036}}$$

$$\left\lfloor \sqrt[2017]{2015^{2015}} + \sqrt[2015]{2017^{2017}} \right] = 2016$$

Math Garden

Jürgen Richter-Gebert
Ulrich Kortenkamp

__Book:__
Perspectives on
Projective
Geometry.

$$\boxed{3{,}21^\circ}$$

$0.21^\circ \rightarrow \qquad \underline{1^\circ} \rightarrow 60' = 7' + 0{,}2'$

$\qquad\qquad 0{,}12 - x$

$\downarrow$

$\underline{\underline{3^\circ 7' 12''}} \qquad \underline{1'} \rightarrow 60'' \qquad 12''$

$\qquad\qquad 0{,}2 \rightarrow$

$3 + 0,21^\circ = 3^\circ 7' 12''$

(3,21)

$$\begin{array}{r} 360\times \\ 60 \\ \hline 21600\times \\ 60 \\ \hline 1296000 \text{ seg} \end{array}$$

$3 \times 60 = 180 \times 60 = 10800$

$7 \times 60 = 420$

$$\begin{array}{r} 10800 \\ 420 \\ 12 \quad 12 \\ \hline 11232 \end{array}$$

$1296000 \div 11232 =$

$115,\overline{384615}$

Obs:

$\pi = 180$

$180 \times 60 \times 60 = 648000$

$648000 \div 11232 =$

$57,\overline{692307}$

' #

Escada de Loreto - New Mexico.

33 degraus

Capela de Loreto → Santa Fé.

1873

Irmãs de Loreto

[Tau manifesto
Michael Hartl's

$$e^{\tau i} = 1$$

↓

Tau

$$C_n = \frac{1}{T}\int_0^T f(x)\, e^{inx}\, dx$$

$$n! \approx \sqrt{\tau n}\left(\frac{n}{e}\right)^{n}$$

Euler

"Let π be the circumference of a circle whose radius = 1"

6.2832...

Tentamen Explicationis Phaenomenorum Aeris

Leonh Eulero

Pag. 347

6.28 → tau

π = perimeter

Introductio Infinitorum

Caput VIII (8)

describing the semi-circumference of a circle with radius 1, and after expanding out 128 digits of the thing number, the author writes:

"which for the sake of brevity I may write π"

$\overline{1.5708}\ldots \begin{cases}\pi/2\\ \pi/4\end{cases}$

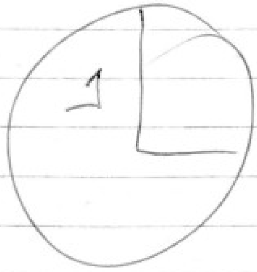

Commentarii Academiae Scientiarum Imperialis

Tomus II

$$\pi = \frac{\text{semicircumference}}{\text{radius}}$$

Important: 20201128

C:\ZIP\PI\Euler\
introductionla01eule.pdf
Pág. 92(116)
Chapitre VIII
Des Quantités transcendantes
qui naissent du Cercle

PIX

20201128
BB - Zap (61) 40040001.

Secrets of the Universe

20201128
36.75.98.34.bc.googleusercontent.com
plattelnotius2@gmail.com

Santa Esmeralda -
Don't let me be
misunderstood.

C:\Data-NovoVIVO-178-2020\
Books-178-2020\
Lectures in Geometric
Combinatorics\
0821841408.pdf

Pag. 52(62)

Fig. 2 - Affine Gale
diagram for
Theorem 6.11

Pág. 57(67) - Triangulations

of Point Configurations.

80 moedas   4 amigos
1/8 viúva
resto p/ 3 filhos

$$80 = \frac{80}{8} + \frac{80}{4} + 3x$$

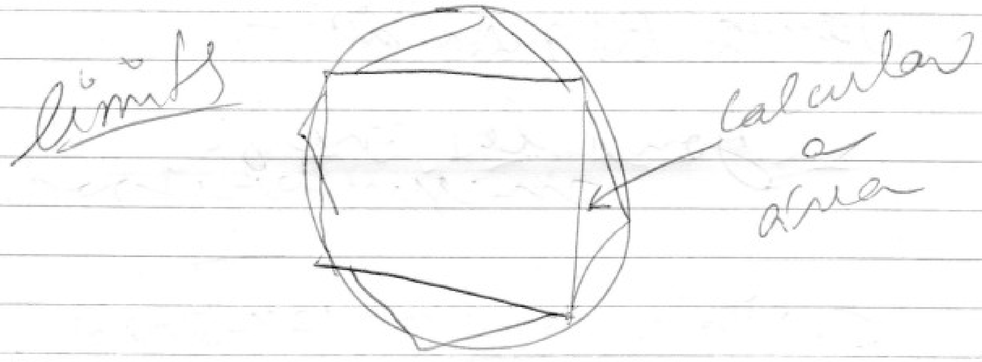
limits
calcular a área

quadrado e octágono > mudança
que gera um aumento de área.

mpg.de

/+

H !F O ! : * 8 * 5

F

) ""! ! ! ! "1//2 "! !) 3 !4+ + (&-

9:5;9 9 9/ 9 $((. $#((((+ + (&-)!

+) ""! ! ! ! "1//2 "! !) 3 !4+ + (&-

9:5;9 9 9/ 9 $((.#&-((((+ + (&-)!

) ""! ! ! ! "1//2 "! !) 3 !4+ + (&-$

9:5;9 9 9/ 9 $((.#&-((((+ + (&-)!

!) ""! ! ! ! "1//2 "! !) 3 !4+ + (&-#

9:5;9 9 9/ 9 $((. $ ((((+ + (&-#)!

&

) 8 B G

) ""! ! ! ! "1//2 "! !) 3 !4+ + -.# #-

9:5;9 9 9/ 9 & $(($ ((((+ + -.# #-)!

+) ""! ! ! ! "1//2 "! !) 3 !4+ + --&

9:5;9 9 9/ 9 & $(($ #. ((((+ + --&)!

) ""! ! ! ! "1//2 "! !) 3 !4+ + --&&

9:5;9 9 9/ 9 & $(($ - -((((+ + --&&)!

!) ""! ! ! ! "1//2 "! !) 3 !4+ + --&(

9:5;9 9 9/ 9 & $(($ - (((((+ + --&( )!

) ""! ! ! ! "1//2 "! !) 3 !4+ + --(

9:5;9 9 9/ 9 & $(($ - (((((+ + --( )!

pedi a noiva Eugênia em casamento.

—

www. research-collection.
ethz.ch

—

www. digitale-sammlungen.de

—

https://download. digitale-
sammulungen. de/
BOOKS/
download.pl?
id=bsb10081751
52
53
54

Cours de Mathématiques
https://download. digitale-
sammlungen. de//BOOKS/
download.pl?id=
bsb10594245 (vol.1)
bsb10625576 (Vol.2)
5577 (Vol.3)
5578 (Vol.4)
5580 (vol.5)

═

digilib.mpiwg-berlin.
mpg.de

Euler: digitale-sammlungen.de

bsb10052831 (Mechanics)
32 (Mechanics)
33 (Solidorum)

bsb10053431 (Differentialis)
432 (Diff-Vol 1)
433 (Diff-Vol 2)
434 (Diff-Vol 3)
435 (Vol. 1)
436 (Vol. 2)
437 (Vol. 3)
438 (Vol. 4)

>   " "   . -- $.

Chapter 5 - Part 2

Volume-07---20211206

! " !" # # $ % & !
" !" ' " ! # # "
& '
(!) (! '*' +

& ,
-./0- - -1 ! - . 234 " 5 6 7 " !
8 9 . 6 , 4"

-./0- - -1 ! - . 234 " 5 6 7 " !
8 9 . 6 , ":;

& < $' = '$!44 ! ! >
1?
& $ $ '< 4 @ ! 6

Since F leaves the horizontal and vertical lines invariant, we see that

$$\begin{aligned} F(a,b) &= F(\{(x,b) : x \in \mathbb{R}\}) \cap F(\{(a,y) : y \in \mathbb{R}\}) \\ &= \{(x,b) : x \in \mathbb{R}\} \cap \{(a,y) : y \in \mathbb{R}\} \\ &= (a,b). \end{aligned}$$

To show that $f = I$, we need to see the effect of the map $f\colon \mathbb{P}^2_{\mathbb{R}} \to \mathbb{P}^2_{\mathbb{R}}$ on the line $\{z = 0\}$.

Let $[x,y,0]$ be a point on the line $\{z = 0\}$. We may consider $[x,y,0]$ as the equivalence class $[\ell]$ of parallel lines ℓ in $\mathbb{A}^2_{\mathbb{R}}$. Each of these lines ℓ in $\mathbb{A}^2_{\mathbb{R}}$ is pointwise fixed by the affine transformation $F\colon \mathbb{A}^2_{\mathbb{R}} \to \mathbb{A}^2_{\mathbb{R}}$. Hence $f([\ell]) = [\ell]$, i.e., $f([x,y,0]) = [x,y,0]$. This completes the proof. ☐

3.8 Principle of Duality

One of the most beautiful properties of the projective planes is the principle of duality, which says that *in a projective plane, every definition remains the same, every theorem remains true when we interchange the words "point" and "line" and consequently interchange "lie on" and "passing through", "join" and "intersection", "collinear" and "concurrent" etc.* If in a statement, we make these changes, the resulting statement is said to be dual to the original one.

To establish this principle, it is enough to show that the axioms imply their own dual. Once we establish this, then given a theorem and its proof, we can dualize the theorem and the proof can be mechanically written down. Now we will prove that the axioms are self-dual.

1. Axiom 1 says that any two distinct points lie on a unique line. The dual statement of this is that *any two lines meet at a unique point.* This statement is nothing but axiom 2. Thus we see that axioms 1 and 2 are dual to each other.

2. Axiom 3 is the existence of four points no three of which are collinear. The dual statement is the existence of a set of four lines no three of which are concurrent. This is equivalent to saying that a projective plane contains a quadrilateral. This was proved in Proposition 3.2.8.

This establishes the principle of duality.

3.9 Classical Theorems in $\mathbb{P}^2_{\mathbb{K}}$

Theorem 3.9.1 (Desargues theorem). *Let* $\triangle(ABC)$ *and* $\triangle(PQR)$ *be two triangles in* $\mathbb{P}^2_{\mathbb{K}}$ *such that the lines* AP, BQ *and* CR *are concurrent. Let* $X = AB \cap PQ$, $Y = BC \cap QR$ *and* $Z = AC \cap PR$. *Then the three points* X, Y *and* Z *are collinear.*

Proof. The proof will utilize a trick which will be employed repeatedly.

Without loss of generality we may assume that the line XY is the line at ∞. Then the rest of the configuration lies in $\mathbb{A}^2_{\mathbb{K}}$, the affine part of $\mathbb{P}^2_{\mathbb{K}}$, i.e., $\mathbb{A}^2_{\mathbb{K}} = \mathbb{P}^2_{\mathbb{K}} \setminus \ell_{XY}$. By assumption the affine lines $AB \setminus \{X\}$ and $PQ \setminus \{X\}$ are parallel in $\mathbb{A}^2_{\mathbb{K}}$; similarly the affine lines $BC \setminus \{Y\}$ and $QR \setminus \{Y\}$ are parallel in $\mathbb{A}^2_{\mathbb{K}}$. Therefore by the affine Desargues theorem, the affine lines $CA \setminus \ell_{XY}$ and $PR \setminus \ell_{XY}$ are parallel. This means that, in $\mathbb{P}^2_{\mathbb{K}}$, the two lines CA and PR meet at a point on the line ℓ_{XY}. Hence the three points X, Y and Z are collinear in $\mathbb{P}^2_{\mathbb{K}}$.

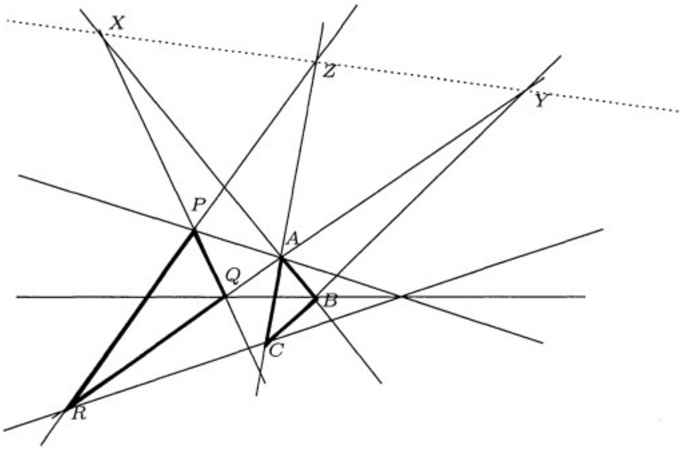

Figure 3.9.1 Desargues Theorem

Second Proof of Desargues theorem. This proof illustrates the use of homogeneous coordinates.

By the fundamental theorem of projective geometry, we may assume that ABC is the standard triangle $\triangle([1,0,0][0,1,0][0,0,1])$. Then the point of concurrence is $[1,1,1]$. The details are left to the reader. (See Figure 3.9.1.)

Since the lines AP, BQ and CR are concurrent at the point $[1,1,1]$, there exist scalars $\lambda_i, \mu_i \in \mathbb{K}^*$ for $1 \le i \le 3$ such that $P = [\lambda_1 + \mu_1, \mu_1, \mu_1]$, $Q = [\mu_2, \lambda_2 + \mu_2, \mu_2]$ and $R = [\mu_3, \mu_3, \lambda_3 + \mu_3]$. Observe that the line $PQ = \{[s(\lambda_1+\mu_1, \mu_1, \mu_1)+t(\mu_2\lambda_2+\mu_2, \mu_2)] : (s,t) \in \mathbb{K}2 \setminus (0,0)\}$ and the line $AB = \{[\alpha, \beta, 0] : (\alpha, \beta) \in \mathbb{K}2 \setminus (0,0)\}$. Since $X = PQ \cap AB$ there exist scalars $s_1, t_1 \in \mathbb{K}$ such that $[s_1(\lambda_1 + \mu_1, \mu_1, \mu_1) + (\mu_2, \lambda_2 + \mu_2, \mu_2)]$ is in AB. Hence $s_1\mu_1 + t_1\mu_2 = 0$ and $X = [s_1\lambda_1, s_2\lambda_2, 0]$. Similarly there exist scalars s_2, t_2, s_3, t_3 such that

- $s_2\mu_2 + t_2\mu_3 = 0$ and $Y = [0, s_2\lambda_2, t_2\lambda_3]$.
- $s_3\mu_3 + t_3\mu_1 = 0$ and $Z = [t_3\lambda_1, 0, s_3\lambda_3]$.

Hence the points X, Y and Z are collinear iff

$$\begin{vmatrix} s_1\lambda_1 & 0 & t_3\lambda_1 \\ t_1\lambda_2 & s_2\lambda_2 & 0 \\ 0 & t_2\lambda_3 & s_3\lambda_3 \end{vmatrix} = (s_1s_2s_3 + t_1t_2t_3)\lambda_1\lambda_2\lambda_3 = 0.$$

Since $\lambda_1\lambda_2\lambda_3 \neq 0$, it is enough to show that $(s_1s_2s_3 + t_1t_2t_3) = 0$. Now the three equations $s_1\mu_1 + t_1\mu_2 = 0$, $s_2\mu_2 + t_2\mu_3 = 0$ and $s_3\mu_1 + t_3\mu_1 = 0$ together imply that $(s_1s_2s_3 + t_1t_2t_3)\mu_1\mu_2\mu_3 = 0$. Again, since $\mu_1\mu_2\mu_3 \neq 0$, it follows that $(s_1s_2s_3 + t_1t_2t_3) = 0$ and the proof of the theorem is complete.

□

We would like to know whether the converse of Desargues theorem is true. We leave this as a set of exercises below.

Exercise 3.9.2.
(i) State the converse of Desargues theorem.
(ii) State the dual of Desargues theorem.
(iii) What is the relation between the converse and the dual of Desargues theorem.
(iv) Can you prove them?

Theorem 3.9.3 (Pappus-Pascal theorem)**.** *Let ABC and PQR be two collinear triples in $\mathbb{P}^2_{\mathbb{K}}$ such that $\ell_{ABC} \neq \ell_{PQR}$. Then the three points $X = AQ \cap BR$, $Y = BP \cap CQ$ and $Z = AP \cap CR$ are collinear.*

Proof. See Figure 3.9.2.

Let ℓ_{XY} be the line joining the points X and Y in $\mathbb{P}^2_{\mathbb{K}}$. To prove the theorem, we need to show that the point Z lies on the line ℓ_{XY}. We

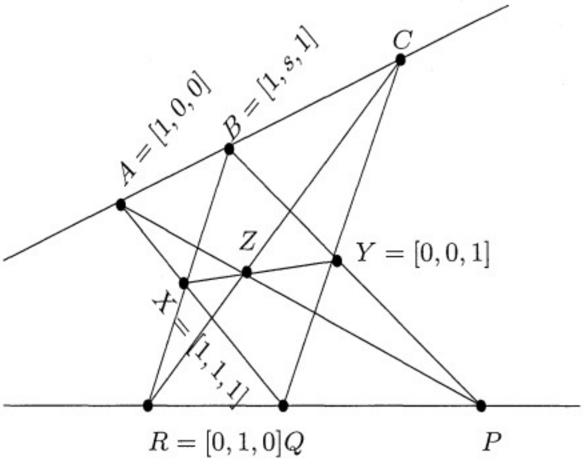

Figure 3.9.2 Pappus-Pascal Theorem

again use the same trick as we did in the proof of Desargues theorem, and take the line ℓ_{XY} as the line at ∞ and look at the configuration in the affine plane $\mathbb{A}^2_{\mathbb{K}} = \mathbb{P}^2_{\mathbb{K}} \setminus \ell_{XY}$. Here we have two lines ℓ_{ABC} and ℓ_{PQR} such that $AQ \parallel BR$ and $BP \parallel CQ$. (Why?) Therefore by the affine form of Pascal's theorem, it follows that $AP \parallel CR$. $\square$

The dual of Pappus-Pascal theorem is known as Pappus-Brianchon theorem. To state the theorem, we introduce some notations.

Definition 3.9.4. Given two lines ℓ_1 and ℓ_2 in a projective plane Π, we denote by $\ell_1\ell_2$, the point of intersection $\ell_1 \cap \ell_2$ of the lines ℓ_1 and ℓ_2.

Given four lines ℓ_1, ℓ_2, ℓ_3 and ℓ_4 in a projective plane Π, the line joining $\ell_1\ell_2$ and $\ell_3\ell_4$ is denoted by $(\ell_1\ell_2)(\ell_3\ell_4)$.

Theorem 3.9.5 (Pappus-Brianchon theorem). *Let ℓ_i, $1 \leq i \leq 6$, be six distinct lines in $\mathbb{P}^2_{\mathbb{K}}$ such that P is the point of concurrence of the lines ℓ_1, ℓ_3 and ℓ_5 and Q the point of concurrence of the lines ℓ_2, ℓ_4 and ℓ_6. Then the lines $(\ell_1\ell_4)(\ell_2\ell_3)$, $(\ell_4\ell_5)(\ell_3\ell_6)$ and $(\ell_5\ell_2)(\ell_4\ell_3)$ are concurrent.*

Proof. The proof is left as an exercise. We just need to use the duality principle. $\square$

Let us recall that a complete quadrangle is a set of four points joined in pairs by six distinct lines; the four points are called the vertices and the lines the edges. Two sides are said to be opposite if they have no common vertex.

Theorem 3.9.6 (Fano's Theorem)**.** *Let $\mathbb{K}$ be a field of characteristic not equal to* 2*. Then the intersections of the three pairs of opposite sides in a quadrangle are never collinear in $\mathbb{P}^2_{\mathbb{K}}$.*

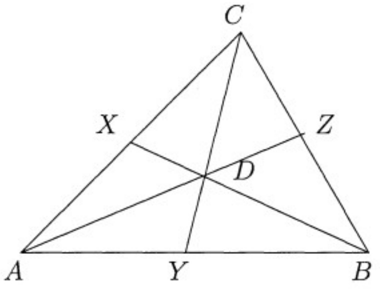

Figure 3.9.3 Fano's Theorem

Proof. Let $ABCD$ be a complete quadrangle. (See Figure 3.9.3 for the configuration). The three pairs of opposite sides are (AC, BD), (AB, CD) and (AD, BC). Let X be the point of intersection of the lines AC and BD, Y the point of intersection of the lines AB and CD and Z the point of intersection of the lines AD and BC. Now our claim is that the three points X, Y and Z are never collinear. We may assume that the line XY is the line at ∞. Let us remove the line XY and look at the remaining configuration in $\mathbb{A}^2_{\mathbb{K}} = \mathbb{P}^2_{\mathbb{K}} \setminus \ell_{XY}$.

In $\mathbb{A}^2_{\mathbb{K}}$, we see that the lines $AC \setminus \{X\}$ and $BD \setminus \{X\}$ are parallel; similarly, the lines $AB \setminus \{Y\}$ and $CD \setminus \{Y\}$ are parallel. This means that the configuration $ABCD$ in $\mathbb{A}^2_{\mathbb{K}}$ is a parallelogram in $\mathbb{A}^2_{\mathbb{K}}$. We know that the diagonals AC and BD of the parallelogram $ABCD$ bisect each other in $\mathbb{A}^2_{\mathbb{K}}$ if the characteristic of $\mathbb{K}$ is not equal to 2 (Theorem 2.9.10). But the lines AC and BD meet at the point Z. Therefore the point $Z = \frac{1}{2}(A+C) = \frac{1}{2}(B+D)$. This shows that the point $Z \in \mathbb{A}^2_{\mathbb{K}} = \mathbb{P}^2_{\mathbb{K}} \setminus \ell_{XY}$, i.e., Z does not lie on the line ℓ_{XY}. Hence the points X, Y and Z are not collinear.

□

The dual of Fano's Theorem is

Theorem 3.9.7. *The diagonals of a complete quadrilateral are never concurrent in* $\mathbb{P}^2_{\mathbb{K}}$*, where* $\text{char}\,\mathbb{K} \neq 2$. □

Remark 3.9.8. If the characteristic of the field $\mathbb{K}$ is 2, then the points X, Y and Z of Fano's configuration are collinear.

Thus, Fano's theorem in $\mathbb{P}^2_{\mathbb{K}}$ is true if and only if the characteristic of the field $\mathbb{K}$ is not equal to 2. (See Theorem 6.11 in [6].) On a general projective plane, Fano's theorem is taken as an axiom. (See [3], p.231)

3.10 Projective Line

Even though our primary concern is plane geometries, we shall define projective spaces of dimension n for each $n \in \mathbb{N}$ over an arbitrary field $\mathbb{K}$ and later specialize to the case $n = 1$.

We fix a field $\mathbb{K}$. The n dimensional *projective space* over $\mathbb{K}$ is defined to be the set of all one dimensional vector subspaces of $\mathbb{K}^{n+1}$. It is denoted by $\mathbb{P}^n(\mathbb{K})$. In particular, we call $\mathbb{P}^1(\mathbb{K})$ as the *projective line* over $\mathbb{K}$.

As in the case of the projective plane, we have a system of homogeneous coordinates in $\mathbb{P}^n(\mathbb{K})$. A point P has homogeneous coordinates $[x_0, x_1, \ldots, x_n]$ if $(x_0, \ldots, x_n)$ is a nonzero vector in the one dimensional vector subspace P. Again as in the projective plane, the group $\mathrm{GL}(n+1, \mathbb{K})$ of nonsingular linear maps over $\mathbb{K}^{n+1}$ acts on $\mathbb{P}^n(\mathbb{K})$ via the map $[x_0, x_1, \ldots, x_n] \mapsto [A(x_0, x_1, \ldots, x_n)]$. We then get the projective linear group $\mathrm{PGL}(n+1, \mathbb{K}) := \mathrm{GL}(n+1, \mathbb{K})/\mathbb{K}^*$, the quotient of $\mathrm{GL}(n+1, \mathbb{K})$ modulo its centre.

Henceforth, we shall concentrate on the projective line over $\mathbb{K}$.

Lemma 3.10.1. *Given three distinct points* P, Q, R *in* $\mathbb{P}^1(\mathbb{K})$*, we can find vectors* $p, q, r \in \mathbb{K}^2$ *such that (i)* $P = \mathbb{K}p$, $Q = \mathbb{K}q$ *and* $R = \mathbb{K}r$ *and (ii)* $p = q + r$.

Proof. Easy. Choose nonzero elements q_1, r_1 in Q, R respectively. Then $\{q_1, r_1\}$ is a basis of $\mathbb{K}^2$. Write $p = aq_1 + br_1$. Note that $a \neq 0$ and $b \neq 0$. (Why?) Let $q = aq_1$ and $r = br_1$. □

The following result should be compared with Theorem 3.7.1.

Theorem 3.10.2. *Given two ordered sets* $(P_1Q_1R_1)$ *and* $(P_2Q_2R_2)$ *of three points in* $\mathbb{P}^1(\mathbb{K})$*, there exists a unique* $A \in PGL(2, \mathbb{K})$ *which takes* P_1 *to* P_2 *etc.*

Proof. By the previous lemma, we can find vectors p_i, q_i, r_i such that $P_i = \mathbb{K}p_i$ etc. Also, we have $p_i = q_i + r_i$. Now $\{q_1, r_1\}$ and $\{q_2, r_2\}$ are bases of $\mathbb{K}^2$ so that there exists a unique linear map $A\colon \mathbb{K}^2 \to \mathbb{K}^2$ such that $Aq_1 = q_2$ etc. Clearly, $Ap_1 = p_2$.

To prove uniqueness, it suffices to show that if $A \in PGL(2, \mathbb{K})$ has two distinct fixed points, then it is the identity. Choose vectors p, q, r as in the lemma. Then $Ap = ap$, $Aq = bq$ and $Ar = cr$. Since

$$a(q + r) = ap = Ap = A(q + r) = Aq + Ar = bq + cr,$$

by linear independence of $\{q, r\}$, it follows that $a = b = c$. As a consequence, $A = aI$. □

Lemma 3.10.3. *Let A, B, C, D be four distinct points of $\mathbb{P}^1(\mathbb{K})$. We can find vectors $a, b, c, d \in \mathbb{K}^2$ such that*

(i) $A = \mathbb{K}a$, $B = \mathbb{K}b$, $C = \mathbb{K}c$ *and* $D = \mathbb{K}d$.
(ii) $c = a + b$.
(iii) $d = \lambda a + b$.

Furthermore, the scalar λ in (iii) is independent of the choices made.

Proof. We have already achieved (ii) in Lemma 3.10.1. Let $0 \neq d' \in D$. Then $d' = \alpha a + \beta b$. Note that $\alpha \neq 0 \neq \beta$ so that

$$0 \neq \frac{d'}{\beta} = \frac{\alpha}{\beta} a + b \in D.$$

Choose $d = d'/\beta$ so that $D = \mathbb{K}d$, $d = \lambda a + b$, where $\lambda = \alpha/\beta$.

To prove the uniqueness of λ: Let a_1, b_1, c_1, d_1 do the job. There exist scalars r, s such that $a_1 = ra$ and $b_1 = sb$. Then $c_1 = a_1 + b_1 = ra + sb$ must be a scalar multiple of $c = a + b$. Hence we conclude that $r = s$. (Why?) If $d_1 = \mu a_1 + b_1$, we have

$$d_1 = \mu ra + rb = r(\mu a + b)$$

is a scalar multiple of

$$d = \lambda a + b$$

iff $\mu = \lambda$. □

Definition 3.10.4. Let (A, B, C, D) be an ordered set of four distinct points in $\mathbb{P}^1(\mathbb{K})$. Then its *cross-ratio* is defined to be the scalar λ of the last theorem.

In the rest of the section, we shall approach this concept from a different direction and arrive at a formula for the cross-ratio.

The action of $\mathrm{GL}(2,\mathbb{K})$ on $\mathbb{K}^2$ induces an action on $\mathbb{P}^1_{\mathbb{K}}$ as follows: Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ be an element in $\mathrm{GL}(2,\mathbb{K})$. Then, for $(x,y) \in \mathbb{K}^2$,

$$A(x,y) = (ax+by, cx+dy).$$

Since A is nonsingular, $A(x,y)$ is non-zero if the vector (x,y) is non-zero. Using this observation, we define a map $f_A \colon \mathbb{P}^1_{\mathbb{K}} \to \mathbb{P}^1_{\mathbb{K}}$ by setting

$$f_A[x,y] := [ax+by, cx+dy], \text{ for } [x,y] \in \mathbb{P}^1_{\mathbb{K}}.$$

We need to show that this is well-defined. That is, if $[x'y'] = [x,y]$, then $[A(x',y')] = [A(x,y)]$. This is easy and left to the reader. Note that if $B = \lambda A$ for some non-zero $\lambda \in \mathbb{K}$, then $f_A = f_B$ on $\mathbb{P}^1_{\mathbb{K}}$.

Exercise 3.10.5. Show that if $A, B \in \mathrm{GL}(2,\mathbb{K})$ are such that $f_A = f_B$ on $\mathbb{P}^1_{\mathbb{K}}$, then $B = \lambda A$ for some $\lambda \in \mathbb{K}$. *Hint:* Refer to Lemma 3.6.1.

Given a field $\mathbb{K}$, we add an "ideal point" ∞, not in $\mathbb{K}$, to get $\mathbb{K}_\infty := \mathbb{K} \cup \{\infty\}$.

The proof of the following proposition is left to the reader.

Proposition 3.10.6. *The map* $\Psi \colon \mathbb{P}^1_{\mathbb{K}} \to \mathbb{K}_\infty$ *given by*

$$\Psi([z,w]) = \begin{cases} z/w & \text{if} \quad w \neq 0 \\ \infty & \text{if} \quad w = 0 \end{cases}$$

is a bijection.

The inverse $\Psi^{-1} \colon \mathbb{K}_\infty \to \mathbb{P}^1_{\mathbb{K}}$ *is given by* $z \mapsto [z,1]$ *for* $z \neq \infty$ *and* $\infty \mapsto [1,0]$. ☐

Exercise 3.10.7. Let X and Y be arbitrary sets. If $f \colon X \to X$ is a bijection and if $\varphi \colon X \to Y$ is a bijection, the map $\varphi \circ f \circ \varphi^{-1} \colon Y \to Y$ is a bijection.

We restrict ourselves to the complex number field $\mathbb{C}$.

Any bijection of $\mathbb{P}^1_{\mathbb{C}}$ induces a bijection of $\mathbb{C}_\infty$ by the above exercise. In particular, for $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in PGL(2,\mathbb{C})$, the map $f_A \colon \mathbb{P}^1_{\mathbb{C}} \to \mathbb{P}^1_{\mathbb{C}}$ induces a bijection $F_A : \mathbb{C}_\infty \to \mathbb{C}_\infty$. Now we would like to write down the map F_A explicitly.

By Exercise 3.10.7, the map $F_A = \Psi \circ f_A \circ \Psi^{-1}$.

Let $z \neq \infty$. Then $\Psi^{-1}(z) = [z, 1]$. This point $[z, 1]$ is mapped by f_A to $[A(z, 1)] = [az + b, cz + d]$. The map Ψ sends the point $[az + b, cz + d]$ to $(az + b)/(cz + d)$, if $cz + d \neq 0$.

Similarly, one shows that $F_A(\infty) = a/c$ and $F_A(-d/c) = \infty$. To summarize, for $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in GL(2, \mathbb{C})$, the map F_A is given as follows:

$$F_A(z) = \begin{cases} \frac{az+b}{cz+d} & \text{if} \quad z \in \mathbb{C},\, z \neq -\frac{d}{c} \\ \infty & \text{if} \quad z = \frac{-d}{c} \\ a/c & \text{if} \quad z = \infty. \end{cases} \tag{3.10.1}$$

The map $F_A \colon \mathbb{C}_\infty \to \mathbb{C}_\infty$ is called the *fractional linear transformation*, (bilinear transformation, Möbius transformation) associated with A. Note that given complex numbers a, b, c and d such that $ad - bc \neq 0$, we can define a map $\varphi \colon \mathbb{C}_\infty \to \mathbb{C}_\infty$ just by (3.10.1) without introducing $\mathbb{P}^1_{\mathbb{C}}$.

Exercise 3.10.8. For what $A \in$ PGL $(2, \mathbb{C})$, do we have $f_A(\infty) = \infty$?

Remark 3.10.9. All the analysis above goes through if we replace $\mathbb{C}$ by any field $\mathbb{K}$.

Remark 3.10.10. There are three classes of elementary fractional linear transformations: The translation $z \mapsto z + b$, the multiplication by a nonzero scalar $z \mapsto az$ (which is geometrically magnification by $|a|$ and a rotation by $e^{i\theta}$ if $a = |a|e^{i\theta}$) and the inversion $z \mapsto 1/z$. In fact, any fractional linear transformation can be written as a composite of these three. To see this observe that

$$\varphi(z) = \frac{az + b}{cz + d} = \frac{a}{c} + \frac{bc - ad}{c^2} \frac{1}{z + (d/c)}.$$

(Hint: Division of polynomials!) Thus φ is given by

$$\begin{aligned} z &\mapsto w_1 = z + (d/c) \\ w_1 &\mapsto w_2 = \frac{1}{w_1} = \frac{1}{z + (d/c)} \\ w_2 &\mapsto w_3 = \frac{bc - ad}{c^2} w_2 = \frac{bc - ad}{c^2} \frac{1}{z + (d/c)} \\ w_3 &\mapsto w_4 = \frac{a}{c} + w_3 = \frac{a}{c} + \frac{bc - ad}{c^2} \frac{1}{z + (d/c)}. \end{aligned}$$

We claim that given three distinct points x_1, x_2 and x_3 in $\mathbb{A}^1_{\mathbb{K}} \subseteq \mathbb{P}^1_{\mathbb{K}}$, there exists a unique A in $PGL(2, \mathbb{K})$ such that $\Psi_A(x_1) = 0$, $\Psi_A(x_2) = 1$ and $\Psi_A(x_3) = \infty$. If such an A exists, then

1. $\Psi_A(x_1) = 0$ iff $ax_1 + b = 0$ iff $b = -ax_1$.
2. $\Psi_A(x_3) = \infty$ iff $cx_3 + d = 0$ iff $d = -cx_3$.
3. $\Psi_A(x_2) = 1$ iff $ax_2 + b = cx_2 + d$. Now by substituting the values of b and d in terms of x_1 and x_3, we see that $ax_2 + b = cx_2 + d$ iff $a(x_2 - x_1) = c(x_2 - x_3)$. Hence $c = \dfrac{a(x_2 - x_1)}{(x_2 - x_3)}$.

Therefore

$$\begin{aligned}\Psi_A(x) &= \frac{ax+b}{cx+d} \\ &= \frac{ax - ax_1}{\frac{a(x_2-x_1)}{x_2-x_3}x - \frac{a(x_2-x_1)}{x_2-x_3}x_3} \\ &= \frac{(x-x_1)(x_2-x_3)}{(x_2-x_1)(x-x_3)} \\ &= \frac{(x-x_1)(x_2-x_3)}{(x-x_3)(x_2-x_1)}.\end{aligned}$$

This shows that the required matrix A is given by

$$A = \begin{pmatrix} x_2 - x_3 & -(x_2 - x_3)x_1 \\ x_2 - x_1 & -x_3(x_2 - x_1) \end{pmatrix}.$$

As a mnemonic short-cut, we arrive at the above formula as follows. Since x_1 should go to zero, we need a term of the form $(x - x_1)$. Since x_2 should go to 1, this expression is modified as $(x-x_1)/(x_2-x_1)$. Now, x_3 should go to ∞. To achieve this, we need the reciprocal of $x - x_3$. Thus we get an expression of the form

$$\left(\frac{x - x_1}{x_2 - x_1}\right)\left(\frac{1}{x - x_3}\right).$$

Under this $x_1 \mapsto 0$ and $x_3 \mapsto \infty$. To ensure $x_2 \mapsto 1$, we modify this as

$$\left(\frac{x - x_1}{x_2 - x_1}\right)\left(\frac{x_2 - x_3}{x - x_3}\right).$$

Since the points x_1, x_2 and x_3 are all distinct,

$$\det(A) = (x_2 - x_3)(x_2 - x_1)(x_1 - x_3)$$

is non-zero. Hence A is nonsingular. We denote $\Psi_A(x)$ by $[x_1, x_2, x_3; x]$ and call this element in $\mathbb{K}$ as the *cross-ratio* of the point x with respect to (x_1, x_2, x_3).

Theorem 3.10.11. *The two definitions of cross-ratio are equivalent.*

Proof. Let P_i for $1 \leq i \leq 3$ be three distinct points in $\mathbb{P}^1_{\mathbb{K}}$. Let P be an arbitrary point in $\mathbb{P}^1_{\mathbb{K}}$. We want to show that $(P_1, P_2, P_3, P) = [P_1, P_2, P_3; P]$.

Let us first compute the cross-ratio of P with respect to the triple (P_1, P_2, P_3) according to the first definition. Let $P_1 = \mathbb{K}u_1$, $P_2 = \mathbb{K}u_2$ and $P_3 = \mathbb{K}(u_1 + u_2)$. Then, if $P = \mathbb{K}(xu_1 + yu_2)$, according to the first definition, the cross-ratio $(P_1, P_2, P_3; P) = x/y$.

Let us now consider the standard frame $Q_1 = \mathbb{K}e_1$, $Q_2 = \mathbb{K}e_2$ and $Q_3 = \mathbb{K}(e_1 + e_2)$. Let us now compute the cross-ratio $(Q_1, Q_2, Q_3; P_i)$ and $(Q_1, Q_2, Q_3; P)$ according to the first definition.

Let $P_1 = \mathbb{K}(t_1e_1 + e_2) = \mathbb{K}u_1$, $P_2 = \mathbb{K}(t_2e_1 + e_2) = \mathbb{K}u_2$ and $P_3 = \mathbb{K}(t_3e_1 + e_2) = \mathbb{K}(u_1 + u_2)$. Let us also write $P = \mathbb{K}(te_1 + e_2)$.

If we now show that $t = x/y$, then we are through. First notice that

$$u_1 = a(t_1e_1 + e_2) \tag{3.10.2}$$

$$u_2 = b(t_2e_1 + e_2) \tag{3.10.3}$$

for some non-zero scalars a and b in $\mathbb{K}$. Now, since $P = \mathbb{K}(te_1 + e_2) = \mathbb{K}(xu_1 + yu_2)$, we see that

$$\begin{aligned} te_1 + e_2 &= \alpha(xu_1 + yu_2) && (3.10.4)\\ &= \alpha(x(a(t_1e_1 + e_2)) + y(b(t_2e_1 + e_2))) && (3.10.5)\\ &= \alpha(xat_1 + ybt_2)e_1 + \alpha(xa + yb)e_2. && (3.10.6)\end{aligned}$$

Since e_1 and e_2 form a basis of $\mathbb{K}^2$, it follows that

$$\begin{aligned} t &= xat_1 + ybt_2 && (3.10.7)\\ 1 &= xa + yb. && (3.10.8)\end{aligned}$$

Hence, using (3.10.8) we can write (3.10.7) as

$$\begin{aligned} \alpha(xat_1 + ybt_2) &= \alpha t(xa + yb)\\ &= \alpha txa + \alpha tyb. && (3.10.9)\end{aligned}$$

We can re-write (3.10.9) as

$$\frac{x}{y}a(t_1 - t) + b(t_2 - t) = 0. \tag{3.10.10}$$

It can be further simplified to

$$\frac{x}{y} = \frac{b}{a}\frac{t - t_2}{t_1 - t}. \tag{3.10.11}$$

Therefore, if we can compute the ratio b/a in terms of the t_i's and t, then we will have an expression for x/y involving only the t_i's and we hope that it will yield a proof.

Since $t_3e_1 + e_2 \in \mathbb{K}(u_1 + u_2) = P_3$, there exists a non-zero scalar β such that $t_3e_1 + e_2 = \beta(u_1 + u_2)$. Therefore

$$\begin{aligned} t_3e_1 + e_2 &= \beta(u_1 + u_2) \\ &= \beta[a(t_1e_1 + e_2) + b(t_2e_1 + e_2)] \\ &= \beta[(at_1 + bt_2)e_1 + (a + b)e_2]. \end{aligned} \tag{3.10.12}$$

Again using the fact that the vectors e_1 and e_2 form a basis of $\mathbb{K}^2$, we get from (3.10.12) that $\beta(a + b) = 1$ and $\beta(at_1 + bt_2) = t_3$. Now

$$\begin{aligned} \beta\left((at_1 + bt_2) + (a + b)\right) &= t_3 \\ &= t_3\beta(a + b) \quad \text{and} \\ \beta\left(a(t_1 - t_3) + b(t_2 - t_3)\right) &= 0. \end{aligned} \tag{3.10.13}$$

Since $\beta \neq 0$, we get from (3.10.13) that $a(t_1 - t_3) + b(t_2 - t_3) = 0$. Hence

$$\frac{b}{a} = \frac{t_1 - t_3}{t_3 - t_2}. \tag{3.10.14}$$

Substituting this ratio in (3.10.11) we get

$$\frac{x}{y} = \frac{(t - t_2)(t_1 - t_3)}{(t - t_1)(t_2 - t_3)}. \tag{3.10.15}$$

This completes the proof. □

Exercise 3.10.12. Let $A = \mathbb{K}a$, $B = \mathbb{K}b$ and $C = \mathbb{K}c$, where $c = a + b$, be three distinct points in $\mathbb{P}^1_{\mathbb{K}}$. Let D be any other point in $\mathbb{P}^1_{\mathbb{K}}$. Then we know that the cross-ratio $(A, B, C; D) = \lambda$ where $D = \mathbb{K}(\lambda a + b)$. Show that

1. $(B, A, C; D) = 1/\lambda$,
2. $(C, B, A; D) = \lambda/(1 - \lambda)$,
3. $(A, C, B; D) = \lambda - 1$,
4. $(B, C, A; D) = (1 - \lambda)/\lambda$ and
5. $(C, A, B; D) = 1/(\lambda - 1)$.

Lemma 3.10.13. *If* φ *is a fractional linear transformation and if* z_k*,* $1 \leq k \leq 4$ *are distinct points of* $\mathbb{K}_\infty$*, then*

$$(z_1, z_2, z_3; z_4) = (\varphi(z_1), \varphi(z_2), \varphi(z_3); \varphi(z_4)).$$

That is, the cross-ratio is invariant under fractional linear transformations.

Proof. By using Remark 3.10.10, it suffices to show that the cross-ratio is invariant under elementary fractional linear transformations. This is easy to verify and left to the reader. $\square$

3.11 Projective Completion of Conics

Given a conic C in $\mathbb{A}^2_{\mathbb{R}}$, we look at its image in the projective completion $\mathbb{P}^2_{\mathbb{R}}$ of $\mathbb{A}^2_{\mathbb{R}}$. We denote this conic by C^*. The image will again be a conic, which might have some extra points. So first we set up an algebraic correspondence and then understand the geometric meaning of this completion. We show that the extra points in $\mathbb{P}^2_{\mathbb{R}}$ of the completed conic are the limiting directions of the tangent lines at points on the affine part of the conic as we move along the conic towards infinity.

Let us start with an example to illustrate what we are going to do.

Example 3.11.1. Let ℓ be a line through origin in $\mathbb{A}^2_{\mathbb{R}}$. Then the image of this curve has only one extra point in $\mathbb{P}^2_{\mathbb{R}}$; namely the point at ∞, which as we saw in the completion of $\mathbb{A}^2_{\mathbb{R}}$, can be viewed as the direction vector of the line ℓ. This fact is also geometrically obvious as a line ℓ has the same direction at any of its points.

As a concrete case let us look at the x-axis which we denote by ℓ. Its homogeneous equation is $y = 0$. In $\mathbb{P}^2_{\mathbb{R}}$, the projective completion of $\mathbb{A}^2_{\mathbb{R}}$, the image of all the points that lie on the line $\{(x, 0)\}$ will have homogeneous coordinates $[x, 0, 1]$. They lie on the line ℓ^* and the extra point that lies on ℓ^* is given by $\{y = 0\} \cap \{z = 0\}$ in $\mathbb{P}^2_{\mathbb{R}}$. We easily see that $[1, 0, 0]$ is the only extra point.

Definition 3.11.2. A *conic* C in $\mathbb{A}^2_{\mathbb{R}}$ is given by the solution of an equation of the form $\langle Ax, x\rangle + \langle b, x\rangle + c = 0$, where A is a 2×2 symmetric matrix and b is a point in $\mathbb{A}^2_{\mathbb{R}}$ and c is a real number; and $\langle .,. \rangle$ is the usual dot product in $\mathbb{R}^2$.

The projective completion of the conic C in $\mathbb{P}^2_{\mathbb{R}}$ is got by "homogenising" the equation $\langle Ax, x\rangle + \langle b, x\rangle + c = 0$. It is also denoted by C^*. We will explain this further: Let $P = [x, y, z]$ with $z \neq 0$ be a point in $\mathbb{P}^2_{\mathbb{R}}$.

&

avxhm.se/ebook/3030613976.html

Pi → não é TRANSCENDENTAL NUMBER
perhaps RATIONAL

Lindemann–Weierstrass Theorem.
is a result that is very useful in establishing the transcendence of numbers.

about: support

⇛ Cersei

"campo externo de empacotamento" ao fluido.

31,684.17 GB.

vacina por decurso de prazo

&

Jason became fascinated with number Pi, irrational number that is

"a fundamental building block of everything around me, a signifier of nature's perfect symmetry".

Struck by Genius

A geometria do pi é tecida no espaço-tempo

Sempre achei que a geometria de Pi descrevia anéis Einstein (também conhecidos como lentes gravitacionais).

the geometry of Pi

A geometria do pi é tecida no espaço-tempo e é impressionantemente bonita e é encontrada às vezes em lugares surpreendentes.

Sempre achei que a geometria de Pi descrevia anéis Einstein (também conhecidos como lentes gravitacionais), mas parece estar relacionado com estes novos objetos também.

The geometry of Pi is woven into spacetime and it is stunningly beautiful and it is found sometimes in surprising places.

I've always thought the geometry of Pi described Einstein rings

(aka gravitational lensing), but it seems to be related to these new objects also.

—

jason-padgett.pixels.com

—

jason.s.padgett@durham.ac.uk

—

Geometrie

—

Comic Kendig

Complex Variables Krantz

Steven G. Krantz

⇒ sk@wustl.edu
math

—

& '

Boster - 193 - 2020 /
Introduction to Modern
Dynamics Chaos... \

019884462x.pdf

Pág. 136(150)
4.42 Feigenbaum
number and
universality

$$\delta = \lim_{n \to \infty} \frac{S_{n+2} - S_{n+1}}{S_{n+1} - S_n} =$$

4.66920160910299067185 3
20382

This Feigenbaum ratio is the
same for all quadratic maps

Pág. 168(182)
5.4 KAM theory
continue fraction

A quadratic irrational is the solution to a quadratic equation with integer coefficients and has periodic continued fractions, such as

$$\sqrt{2} = [1; 2, 2, 2, \ldots] \quad (5.24)$$

Nonquadratic irrational have nonrepeating continued fractions, such as

$$\pi = [3; 7, 15, 1, 292, 1, 1, 1, 2, 1, 3, 1, 14, 2, 1, 1, 2, 2, 2, 2, 1, 84, 2, \ldots] \quad (5.25)$$

$$\pi - 3 \approx [0; 7, 15, 1, 0, 0, 0, \ldots]$$

$$= \cfrac{1}{7 + \cfrac{1}{15 + \cfrac{1}{1}}} = 0{,}14160 = \frac{177}{1250} \quad (5.26)$$

' & '
-./0- - -1 ! - ' @!; " @ A % " "
@6 ! ! B A 9 B 4! ! " > " 2" 13 " $ 4"

-./0- - -1 ! - ' @!; " @ A % " "
@6 ! C ! B A 9 B 4! B ! " > 13 " D ; 6 $ 4"

& ' + -./0- - -1 ! - ' @!; " @ A
% " " @6 ! ! B A 9 B 4! ! " > " 2" 13 " $ 4"

+ + = ?! ? ! " ; ! 6

& ' , -./0- - -1 ! - ' @!; " @ A
% " " @6 ! C ! B A 9 B 4! B ! " > 13 " D ; 6
$ 4"

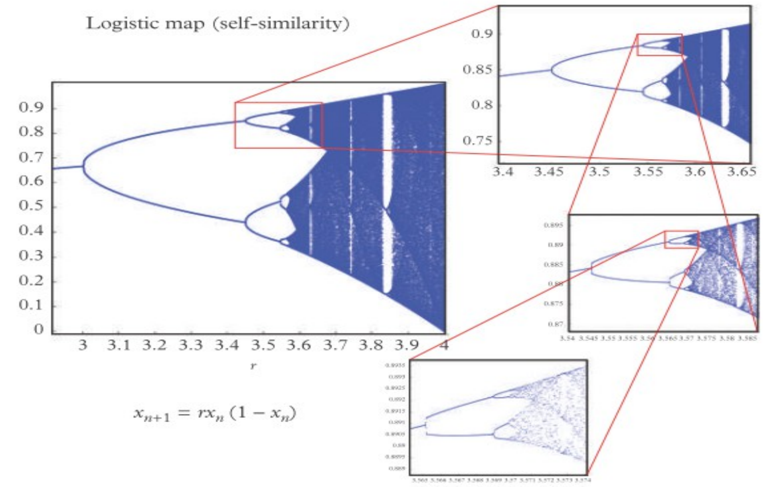

Figure 4.15 *Bifurcation diagram of the logistic map. The sub-areas are self-similar, with the same features appearing as for the larger scale.*

4.4.2 Feigenbaum number and universality

The logistic map is only one of a virtually limitless number of iterated maps. Each different mapping function produces a different graphical solution with different bifurcation cascades. But if every nonlinear system is different, then what type of over-arching principles can be applied to chaos? The answer is that there are common themes among wide classes of nonlinear systems, including quantitative similarities that in some cases become identities. For instance, the logistic map has an inverted quadratic dependence represented by the parabola in Fig. 4.14. One of the discoveries in the field of chaos was the universal behavior of the bifurcation cascade for *every* iterated map that has a lowest-order quadratic dependence (Fig. 4.16). The bifurcation thresholds of r have a fixed ratio in the limit as the bifurcations approach chaos,

$$\delta = \lim_{n\to\infty} \frac{S_{n+2} - S_{n+1}}{S_{n+1} - S_n} = 4.669201609102990671853203 82$$

This Feigenbaum ratio is the same for all quadratic maps (which tend to be the most common). Therefore, in the face of the apparent randomness of chaotic systems, universals do emerge.

4.5 Three-dimensional state space and chaos

The character of nonlinear dynamical systems changes dramatically when going from 2D to 3D state space. Dynamical systems are subject to the no-crossing

& '

Boster - 193 - 2020 / Introduction to Modern Dynamics chaos... \

019884462x.pdf

Pág. 136(150)

4.42 Feigenbaum number and universality

$$\delta = \lim_{n \to \infty} \frac{S_{n+2} - S_{n+1}}{S_{n+1} - S_n} = 4.669201609102990671853\,20382$$

This Feigenbaum ratio is the same for all quadratic maps

Pág. 168(182)

5.4 KAM theory

continue fraction

A quadratic irrational is the solution to a quadratic equation with integer coefficients and has periodic continued fractions, such as

$$\sqrt{2} = [1; 2, 2, 2, \dots] \quad (5.24)$$

Nonquadratic irrational have nonrepeating continued fractions, such as

$$\pi = [3; 7, 15, 1, 292, 1, 1, 1, 2, 1, 3, 1, 14, 2, 1, 1, 2, 2, 2, 2, 1, 84, 2, \dots] \quad (5.25)$$

$$\pi - 3 \approx [0; 7, 15, 1, 0, 0, 0, \dots]$$

$$= \cfrac{1}{7 + \cfrac{1}{15 + \cfrac{1}{1}}} = 0{,}14160 = \frac{177}{1250} \quad (5.26)$$

+ & "

20201225 M

Quéops - 3341 - 3094 A.C.
(radiocarbono → madeira)

you belong to the city
Glenn Frey.

All round the world
Lisa Stansfield.

Being with you Smokey Robinson

⇒ Books - 193 - 2020 / Mathematics for Machine Technology /
Fujio Yamaguchi - Computer Aided Geometric... pdf
Pág. 11 (46)
Fig 1.9 three types of line segments

Obs.: Root Cause of All Evil

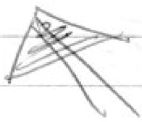

Pag. 300 (326)
Conventional geometric
Newton-Raphson
method, ...
Fig. 15.2
Determining intersection
point of curve and
line segment.

Circles Pedoe
Analytical Conics Spain
Non-Euclidean Geometry
Kulczycki
Introduction to Projective
Geometry Pedoe

Books - 183 - 2020 / Geometry
Revisited / Compact Projective

+

! -./0- - -1 ! - + ; > E6 B E! 6 E
!9 ! ! ! " 4 . " " 5 @ 2 ; % B . !"
$<$ 4"

? -./0- - -1 ! - + ; > E6 B E! 6 E
!9 ! ! ! " 4 . " " 5 @ %% 2 ; % B
. !" $$ 4"

-./0- - -1 ! - + =! 7! "? 9 4 . " "
5 @ A 7 ! " 4"

" -./0- - -1 ! - + . ? ! FF 5
4 ! 4 ! " " " G H 6 ' ":;

-./0- - -1 ! - + = : I! ! ; ! "
! 4 ! " " " 4 $<< ":;

-./0- - -1 ! - + = : I! ! 4 . " "
5 @ . > ! 6 = @ ! .44 ! 4 > 96 4"
1? % 4 ! JJJ

& + = $ > 64

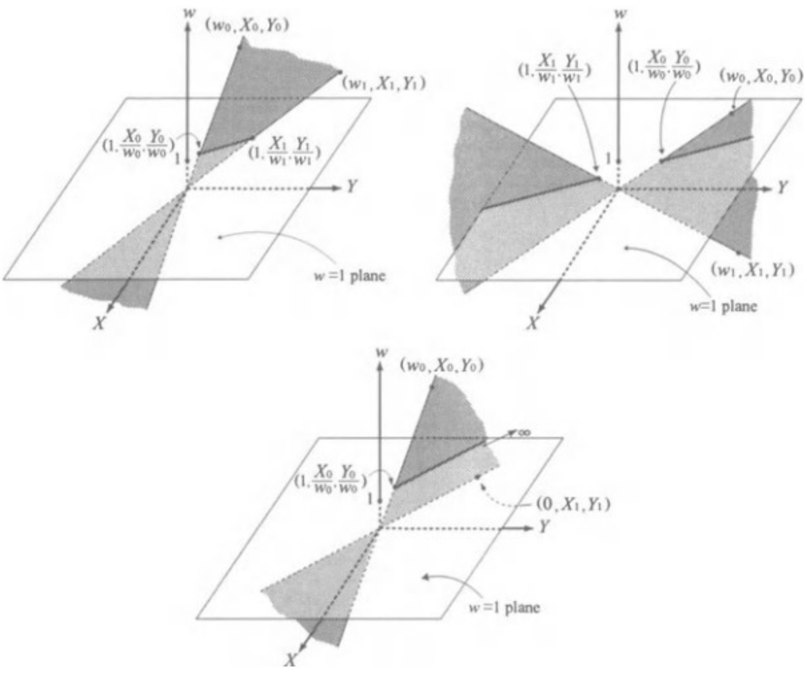
w
(w_0, X_0, Y_0)
(w_1, X_1, Y_1)
$(1, \frac{X_0}{w_0}, \frac{Y_0}{w_0})$
1
$(1, \frac{X_1}{w_1}, \frac{Y_1}{w_1})$
Y
X
w =1 plane
∞
$(0, X_1, Y_1)$

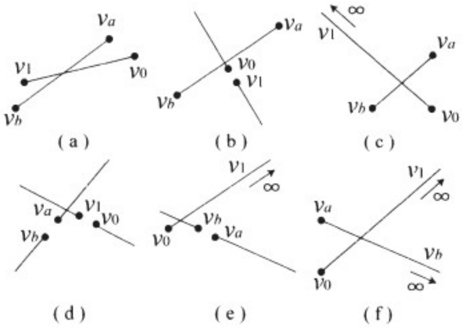

Figure 1.7: Six cases of line segment intersections.

Since a line segment could be transformed into three shapes (Fig. 1.5), and a triangle into five shapes (Fig. 1.6), by a general projective transformation, we have to make $3\times 2=6$ (Fig. 1.7) or $3\times 5=15$ separate routines respectively to cover all the possible cases.

In actual processing, the three types of line segments must each be treated with separate mathematical expressions. For instance, if we denote two endpoint vectors by $\boldsymbol{v}_0$ and $\boldsymbol{v}_1$, we can write as follows:

$$\text{Internal line segment}: \boldsymbol{v} = (1-t)\boldsymbol{v}_0 + t\boldsymbol{v}_1, \qquad (0 \le t \le 1), \quad (1.11)$$

$$\text{Half line}: \boldsymbol{v} = \boldsymbol{v}_0 + \boldsymbol{\alpha}t, \qquad (t \ge 0), \quad (1.12)$$

$$\text{External line segment}: \boldsymbol{v} = \boldsymbol{v}_0 + \boldsymbol{\alpha}t,\ \boldsymbol{v} = \boldsymbol{v}_1 - \boldsymbol{\alpha}t, \quad (t \ge 0). \quad (1.13)$$

In the package, the program, first of all, has to test among all the possible cases which the present case is and pass the control to a corresponding routine. This is not an easy task again. Thus, a seemingly simple problem becomes highly complex by the time a complete program is written.

1.4 Problems of Rational Curves and Surfaces

Because of the high representational capabilities, rational polynomial curves and surfaces are becoming standard descriptions in computer aided geometric design systems.

In dealing with a rational polynomial curve or surface, however, we come across various difficulties.

First, a rational polynomial curve does not allow us to generate an exact point on the curve corresponding to a parameter value because it forces to perform division operations in computing the rational expression itself. Therefore in reality an exact point can not generally be obtained, for example, from a quadratic rational polynomial curve although it is an exact expression of conics.

Second, as mentioned in Section 1.2, the geometric Newton-Raphson method sometimes fails when it is applied to a rational polynomial curve while it does not when applied to an ordinary one.

Third, formulas and algorithms tend to be more complex than those for an ordinary polynomial curve. The projective transformation of the curve is this complex and is not linear with respect to the position vector $\boldsymbol{p}(t)$:

$$\boldsymbol{p}^*(t) = \frac{\boldsymbol{p}(t)\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} + [t_x\ t_y\ t_z]}{px + qy + rz + s}. \tag{1.14}$$

There are many other formulas and algorithms for rational polynomial curves which take much more complex forms than those for ordinary ones.

Fourth, there is a strict sign restriction on weights w_i, i.e., the signs of w_i must not be mixed in the processing of a rational polynomial curve because various inconveniences, such as non-convex hull property, zero denominator of tangent vectors and so on, may arise if the restriction is not observed. This restriction drastically reduces the high representational capability of rational polynomial curves and surfaces.

Suppose $\boldsymbol{q}_0$, $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ define an ellipse (Fig. 1.8). The negation of the weight of $\boldsymbol{q}_1$ generates its complementary segment. This is a simple way of generating a whole curve. This curve segment is, however, not allowed for processing because the signs of the weights of the control vertices are mixed.

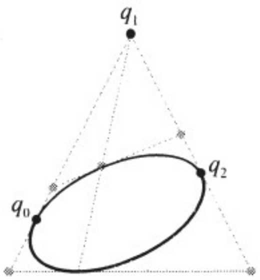

Figure 1.8: Complementary segment.

1.5 Root Cause of All Evil

We have so far pointed out various types of problems inherent in the conventional processing. Let us consider the cause of each problem in the order presented above.

As for the example of non-exactness, it is evident that the cause of non-exactness in the pentagon problem lies in performing the division operations to generate Euclidean coordinates.

In the case of the non-robustness problem of the Newton-Raphson method applied to rational curves and surfaces, it is also self-evident that the cause of the non-robustness lies in the divisional expression of the rational curves and surfaces because this type of non-robustness does not arise when it is applied to ordinary polynomial curves and surfaces.

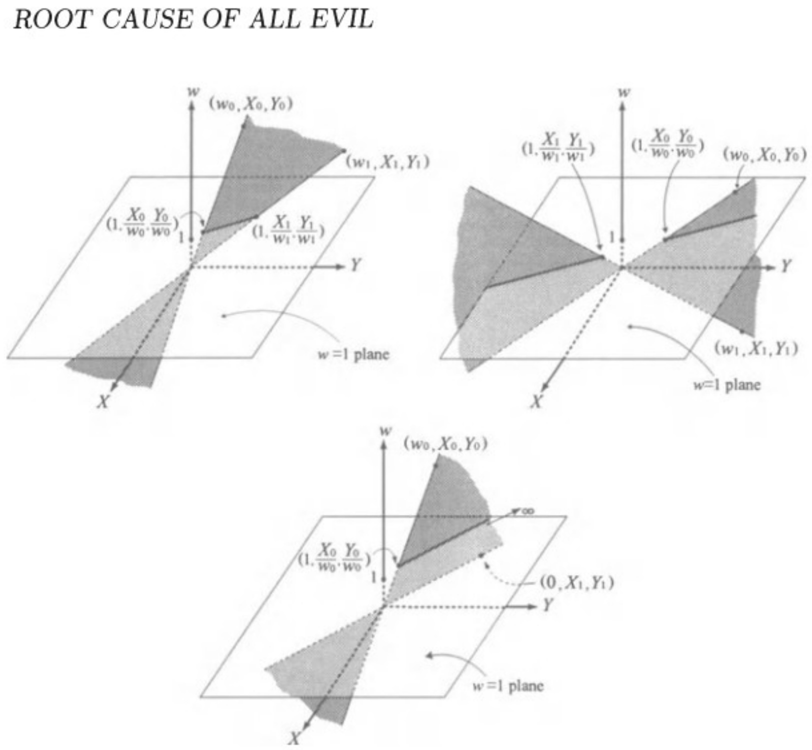
w
(w_0, X_0, Y_0)
(w_1, X_1, Y_1)
$(1, \frac{X_0}{w_0}, \frac{Y_0}{w_0})$
1
$(1, \frac{X_1}{w_1}, \frac{Y_1}{w_1})$
Y
X
w=1 plane
w
$(1, \frac{X_1}{w_1}, \frac{Y_1}{w_1})$
$(1, \frac{X_0}{w_0}, \frac{Y_0}{w_0})$
(w_0, X_0, Y_0)
1
Y
(w_1, X_1, Y_1)
w=1 plane
X
w
(w_0, X_0, Y_0)
∞
$(1, \frac{X_0}{w_0}, \frac{Y_0}{w_0})$
1
$(0, X_1, Y_1)$
Y
w=1 plane
X

Figure 1.9: Three types of line segments.

As for the example of non-compactness, the cause may be a little difficult to find out. Let us consider this way; the three types of line segments are the cuts of wedge-shaped planes with the apexes at the origin of three-space O-wXY (Fig. 1.9). Then we understand that the three types of line segments are the results of division operations, and that division operations could cause to increase the number of shape types of a geometric element.

In Section 1.4, we listed up various problems inherent in the processing of rational curves and surfaces. Because these do not appear in the ordinary curves and surfaces, it is evident that the divisional forms of rational curves and surfaces are the causes of these problems.

Considering the discussion above, it does not seem to be difficult for us to conclude that these problems are caused by division operations. Thus we say that *"the division operation is the root cause of all evil."* (This will be borne out by various test results shown in Chapter 25) This observation invites us to consider the data set immediately before division operations, that is, the denominator w and the numerators X, Y, Z as coordinates of a point in place of the coordinates $(x, y, z) = (X/w, Y/w, Z/w)$. While the post-division data set is called *ordinary coordinates* or *Euclidean coordinates*, the pre-division data set (w, X, Y, Z) is called *homogeneous coordinates.*

It is now the time to study homogeneous coordinates.

of $f(x_{i+1})$:

$$f(x_{i+1}) = f(x_i + \delta x) = f(x_i) + \frac{\dot{f}(x_i)}{1!}\delta x + \frac{\ddot{f}(x_i)}{2!}(\delta x)^2 + \dots. \tag{15.2}$$

Setting $f(x_{i+1})=0$ yields:

$$\delta x = -\frac{f(x_i)}{\dot{f}(x_i)}. \tag{15.3}$$

Thus,

$$x_{i+1} = x_i - \frac{f(x_i)}{\dot{f}(x_i)}. \tag{15.4}$$

Determination of the roots of an equation using eq.(15.4) is called the *Newton-Raphson method* (Fig. 15.1). Highly accurate solutions can be obtained efficiently with this method.

The application of the Newton-Raphson method to determination of intersections between geometric elements is often called the *geometric Newton-Raphson method*, which we are dealing with in this chapter.

15.2 Conventional Euclidean Approach

To illustrate the conventional geometric Newton-Raphson method, where processing takes place in Euclidean space, we consider the problem of finding the intersection point between a plane curve and a line segment. We shall consider the cases for ordinary and rational polynomial curves separately.

Ordinary Polynomial Curve Case

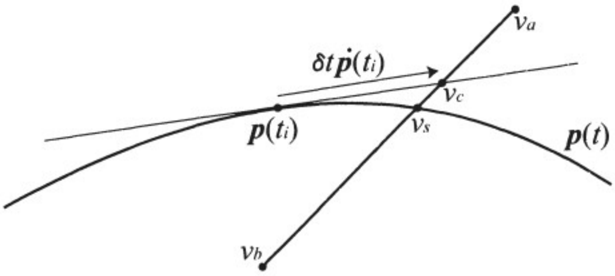

Figure 15.2: Determining intersection point of curve and line segment.

We use the Newton-Raphson method to find the intersection of an ordinary parametric polynomial plane curve (Fig. 15.2):

$$\boldsymbol{p}(t) = [x(t) \;\; y(t)]$$

and a line segment $v_a v_b$, where

$$\boldsymbol{v}_a = [x_a \;\; y_a], \quad \boldsymbol{v}_b = [x_b \;\; y_b].$$

The basic algorithm is as follows:

1. Specify $i = 0$ and an approximate parameter $t = t_0$ to the intersection point.
2. Determine the tangent at $p(t_i)$ on the curve and then find its intersection point with the other geometric element.
3. An increment δt is determined from the intersection point.
4. Increase the parameter t_i by δt, and return to step 2.

The tangent line of the curve at $p(t_i)$ is given by:

$$-\dot{y}(t_i)x + \dot{x}(t_i)y + x(t_i)\dot{y}(t_i) - \dot{x}(t_i)y(t_i) = 0. \tag{15.5}$$

Meanwhile, the line equation for the line segment $v_a v_b$ is given by:

$$(y_a - y_b)x - (x_a - x_b)y + x_a y_b - x_b y_a = 0. \tag{15.6}$$

The intersection point v_c of these two lines is then obtained as:

$$\begin{aligned} x_c &= \frac{\begin{vmatrix} \dot{x}(t_i) & x(t_i)\dot{y}(t_i) - \dot{x}(t_i)y(t_i) \\ x_b - x_a & x_a y_b - x_b y_a \end{vmatrix}}{\begin{vmatrix} \dot{x}(t_i) & \dot{y}(t_i) \\ x_b - x_a & y_b - y_a \end{vmatrix}}, \\ y_c &= \frac{\begin{vmatrix} \dot{y}(t_i) & x(t_i)\dot{y}(t_i) - \dot{x}(t_i)y(t_i) \\ y_b - y_a & x_a y_b - x_b y_a \end{vmatrix}}{\begin{vmatrix} \dot{x}(t_i) & \dot{y}(t_i) \\ x_b - x_a & y_b - y_a \end{vmatrix}}. \end{aligned} \tag{15.7}$$

Now, $\boldsymbol{v}_c$, $\boldsymbol{p}(t_i)$, $\dot{\boldsymbol{p}}(t_i)$ and the parameter increment δt have the following relation:

$$\boldsymbol{v}_c - \boldsymbol{p}(t_i) = \delta t \dot{\boldsymbol{p}}(t_i). \tag{15.8}$$

So we can determine the increment δt from either the expression for x- or y-coordinate above as:

$$\delta t = -\frac{\begin{vmatrix} x(t_i) - x_a & y(t_i) - y_a \\ x_b - x_a & y_b - y_a \end{vmatrix}}{\begin{vmatrix} \dot{x}(t_i) & \dot{y}(t_i) \\ x_b - x_a & y_b - y_a \end{vmatrix}}. \tag{15.9}$$

There is another more general, simpler method to determine δt, which will be useful in later discussions.

First we express the line segment in a parametric form as:

$$\boldsymbol{v}(u) = (1 - u)\boldsymbol{v}_a + u\boldsymbol{v}_b, \quad 0 \le u \le 1. \tag{15.10}$$

If an intersection exists between the curve and line segment, then the following is satisfied at that point:

$$\boldsymbol{p}(t) = \boldsymbol{v}(u).$$

Now, if t_i and u_i are sufficiently close to the intersection point, then there exist δt and δu such that the following equation holds:

$$\boldsymbol{p}(t_i + \delta t) = \boldsymbol{v}(u_i + \delta u). \tag{15.11}$$

Applying the Taylor expansion to both sides and extracting up to the first-order term, we obtain:

$$\begin{aligned} \boldsymbol{p}(t_i) + \delta t \dot{\boldsymbol{p}}(t_i) &\doteqdot \boldsymbol{v}(u_i) + \delta u \dot{\boldsymbol{v}}(u_i) \\ &= (1 - u_i - \delta u)\boldsymbol{v}_a + (u_i + \delta u)\boldsymbol{v}_b. \end{aligned} \tag{15.12}$$

This shows that $\boldsymbol{p}(t_i)+\delta t \dot{\boldsymbol{p}}(t_i)$, $\boldsymbol{v}_a$ and $\boldsymbol{v}_b$ are linearly dependent, so:

$$\begin{vmatrix} 1 & x(t_i) + \delta t \dot{x}(t_i) & y(t_i) + \delta t \dot{y}(t_i) \\ 1 & x_a & y_a \\ 1 & x_b & y_b \end{vmatrix} = 0. \tag{15.13}$$

Solving this for δt yields the same expression as eq.(15.9).

Rational Polynomial Curve Case

Consider now a rational curve given by:

$$\boldsymbol{p}(t) = [x(t) \;\; y(t)] = [\frac{X(t)}{w(t)} \;\; \frac{Y(t)}{w(t)}]. \tag{15.14}$$

From this, the expression for δt can be obtained in the same manner as for the ordinary polynomial curve above, or:

$$\delta t = -\frac{\begin{vmatrix} \frac{X(t_i)}{w(t_i)} - x_a & \frac{Y(t_i)}{w(t_i)} - y_a \\ x_b - x_a & y_b - y_a \end{vmatrix}}{\begin{vmatrix} \frac{\dot{X}(t_i)w(t_i) - X(t_i)\dot{w}(t_i)}{w(t_i)^2} & \frac{\dot{Y}(t_i)w(t_i) - Y(t_i)\dot{w}(t_i)}{w(t_i)^2} \\ x_b - x_a & y_b - y_a \end{vmatrix}}. \tag{15.15}$$

If we let $w(t_i) = 1$ above, we obtain the same expression as eq.(15.9).

We now rewrite eq.(15.15) using 3×3 determinants as follows:

$$\delta t = -\frac{\begin{vmatrix} 1 & x_a & y_a \\ 0 & \frac{X(t_i)}{w(t_i)} - x_a & \frac{Y(t_i)}{w(t_i)} - y_a \\ 0 & x_b - x_a & y_b - y_a \end{vmatrix}}{\begin{vmatrix} 1 & x_a & y_a \\ 0 & \frac{\dot{X}(t_i)w(t_i) - X(t_i)\dot{w}(t_i)}{w(t_i)^2} & \frac{\dot{Y}(t_i)w(t_i) - Y(t_i)\dot{w}(t_i)}{w(t_i)^2} \\ 0 & x_b - x_a & y_b - y_a \end{vmatrix}}$$

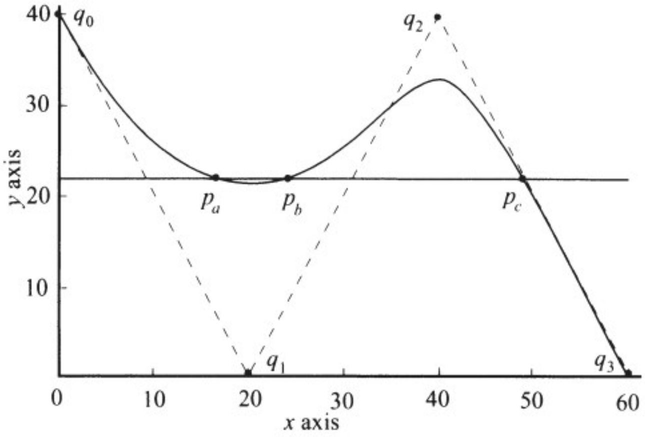

$$\boldsymbol{q}_0 = [0 \;\; 40], \quad \boldsymbol{q}_1 = [20 \;\; 0], \quad \boldsymbol{q}_2 = [40 \;\; 40], \quad \boldsymbol{q}_3 = [60 \;\; 0],$$
$$w_0 = 1, \qquad w_1 = 3, \qquad w_2 = 9, \qquad w_3 = 1.$$

Figure 15.3: Intersection of curve and line segment.

$$= -\frac{\begin{vmatrix} 1 & \frac{X(t_i)}{w(t_i)} & \frac{Y(t_i)}{w(t_i)} \\ w_a & X_a & Y_a \\ w_b & X_b & Y_b \end{vmatrix}}{\begin{vmatrix} 0 & \frac{\dot{X}(t_i)w(t_i) - X(t_i)\dot{w}(t_i)}{w(t_i)^2} & \frac{\dot{Y}(t_i)w(t_i) - Y(t_i)\dot{w}(t_i)}{w(t_i)^2} \\ w_a & X_a & Y_a \\ w_b & X_b & Y_b \end{vmatrix}}. \tag{15.16}$$

15.3 Problems with Euclidean Approach

Problems arise when the conventional method, based on eq.(15.16), is used to process rational polynomial curves. These are now discussed.

A non-robust example was given in Section 1.2 (Fig. 15.3). As Table 15.1 shows for this case, the parameter converges when given the initial value of 0.5, but diverges out to infinity from an initial value of, say, 0.7. That is, the robustness of the algorithm is lost. Convergence to the roots depends on initial values.

Note that this non-robust phenomenon does not arise when the geometric Newton-Raphson method is applied to an ordinary polynomial curve, and that the non-robustness occurs even when the w-values of the control points of the

Pag. 300 (326)

Conventional geometric Newton-Raphson method, ...

Fig. 15.2

Determining intersection point of curve and line segment.

Circles Pedoe

Analytical Conics Spain

Non-Euclidean Geometry Kulczycki

Introduction to Projective Geometry Pedoe

Books - 183 - 2020 / Geometry Revisited / Compact Projective

, & '+
! -./0- - -1 ! - , 5 6 ; " > ! ! !
. ! . ! $ ":;

? -./0- - -1 ! - , 5 6 / ; " >
! ! ! . ! . ! $ 4" K $

-./0- - -1 ! - , 5 6 ; " ! ! !

" -./0- - -1 ! - , 4! : ; ! H !
% " 1 5 6 4"

& + > : ; 4

12.0 General assumption. Throughout this section, $\mathbb{F}$ shall denote one of the Cayley–Dickson division algebras $\mathbb{R}$, $\mathbb{C}$, $\mathbb{H}$, or $\mathbb{O}$.

12.1 The affine plane $\mathscr{A}_2\mathbb{F}$ over $\mathbb{F}$ is constructed as follows. The set of points is $\mathbb{F} \times \mathbb{F}$, and the following subsets of $\mathbb{F} \times \mathbb{F}$ are called lines:

$$\begin{aligned} [s,t] &= \{\,(x, sx+t) \mid x \in \mathbb{F}\,\} \quad \text{for } s,t \in \mathbb{F}\,, \\ [c] &= \{c\} \times \mathbb{F} \quad \text{for } c \in \mathbb{F}\,. \end{aligned}$$

We say that the line $[s,t]$ has *slope* s, and the line $[c]$ has *slope* ∞.

Because $\mathbb{F}$ is a division algebra, it is easily verified that this structure has the following properties:

1) For two points $(x_1, y_1) \neq (x_2, y_2)$ there is a unique line joining them (i.e., containing both of them), namely the line $[s, y_1 - sx_1]$, where s is uniquely determined by the equation

$$s(x_2 - x_1) = y_2 - y_1 \tag{1}$$

if $x_1 \neq x_2$, and the line $[x_1]$ if $x_1 = x_2$.

2) Two lines of different slopes have a unique point of intersection. In fact, we have $[c] \cap [s,t] = \{\,(c, sc+t)\,\}$, and for $s_1 \neq s_2$ we have $[s_1, t_1] \cap [s_2, t_2] = \{\,(x, s_1x + t_1)\,\}$, where x is uniquely determined by the equation

$$(s_1 - s_2)x = t_2 - t_1\,. \tag{2}$$

3) Two different lines are disjoint if they have the same slope, and by property 2) this condition is also necessary. Two lines will be called *parallel* if they have the same slope, i.e., if they are disjoint or equal.

4) The so-called *parallel axiom* holds: For each line L and each point (x, y), there is a unique line which passes through (x, y) and is parallel to L. Depending on whether the slope of L is ∞ or $s \in \mathbb{F}$, this parallel is given by $[x]$ or by $[s, y - sx]$, respectively.

Properties 1) – 4) say that $\mathscr{A}_2\mathbb{F}$ satisfies the axioms of an *affine plane*, see (21.8).

Since $\mathbb{F}$ is biassociative (11.11 and 13), the solutions of equations (1) and (2) may be written down explicitly as

$$s = (y_2 - y_1)(x_2 - x_1)^{-1} \tag{1'}$$

$$x = (s_1 - s_2)^{-1}(t_2 - t_1)\,. \tag{2'}$$

12.2 The projective completion $\overline{\mathscr{A}_2}\mathbb{F}$ of the affine plane $\mathscr{A}_2\mathbb{F}$ is obtained by adjoining 'points at infinity', one for each parallel class. All the lines of the parallel class are thought to intersect in the respective point at infinity. Furthermore, one adds a 'line at infinity' passing through just these points at infinity.

For $s \in \mathbb{F} \cup \{\infty\}$, the point at infinity on the lines of slope s (which form a parallel class) will be denoted by (s), and the line at infinity is

$$[\infty] := \{\, (s) \mid s \in \mathbb{F} \cup \{\infty\} \,\} \ .$$

By this construction, we obtain a geometry in which any two lines $L_1 \neq L_2$ always have a unique point of intersection $L_1 \wedge L_2$. (If L_1, L_2 are lines of the affine plane $\mathscr{A}_2\mathbb{F}$, their intersection point in the projective completion $\overline{\mathscr{A}_2}\mathbb{F}$ is a point at infinity if, and only if, in $\mathscr{A}_2\mathbb{F}$ the lines L_1, L_2 are parallel.) The property that for any two points $p_1 \neq p_2$ there is a unique line $p_1 p_2$ joining them is preserved in this extension process. Thus, the projective completion $\overline{\mathscr{A}_2}\mathbb{F}$ is a *projective* plane in the sense of definition (21.1).

Of course, this construction applies quite generally to any affine plane, see Hughes–Piper [73] Theorem 3.10, p. 83.

For certain points and lines of the projective completion $\overline{\mathscr{A}_2}\mathbb{F}$, we shall systematically use special names as indicated in the diagram of Figure 12a.

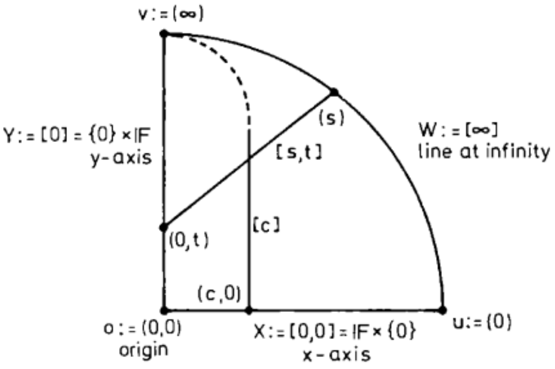

Figure 12a

12.3 Collineations. A *collineation* of an affine or projective plane is a bijection of the set of points onto itself mapping lines onto lines.

As for the latter condition, it suffices to postulate that lines are mapped *into* lines, see (23.2).

A collineation of an affine plane maps every parallel class of lines to a parallel class of lines; one therefore obtains a unique extension to a collineation of the projective completion by permuting the points at infinity accordingly. Conversely, if a collineation of the projective completion leaves the line at infinity invariant, then it arises in this way from an affine collineation. Henceforth, we shall not distinguish between an affine collineation and its projective extension. We remark that, besides these collineations, the projective completion may have collineations which move the line at infinity, cf. (13.5) and (17.6).

It is straightforward that for a collineation φ, for points $p_1 \neq p_2$ and for lines $L_1 \neq L_2$ one has

$$(p_1 p_2)^\varphi = p_1^\varphi p_2^\varphi \quad \text{and} \quad (L_1 \wedge L_2)^\varphi = L_1^\varphi \wedge L_2^\varphi \,.$$

The collineations of a given plane form a group under composition. For a subgroup Δ of this group, for a point p and a line L, the stabilizers Δ_p and Δ_L are the subgroups consisting of all collineations in Δ which fix p or leave L invariant, respectively.

12.4 Axial collineations. We now present a notion which is basic for the study of collineation groups of projective planes, together with some standard results, cf. also (23.7 ff).

(a) Consider a point p and a line L. We say that a collineation φ has *axis* L if φ fixes every point on L; dually, φ has *center* p if φ leaves every line through p invariant. A collineation of a *projective* plane has an axis if, and only if, it has a center (see Hughes–Piper [73] Theorem 4.9, p. 94; in the present chapter, we shall not use this fact). The center may be on the axis or not.

(b) For a group Δ of collineations, $\Delta_{[p,L]}$ denotes the subset of all collineations in Δ with center p and axis L; it is obviously a subgroup. For $\gamma \in \Delta$ one easily verifies the following

Conjugation formula: $\gamma^{-1}\Delta_{[p,L]}\,\gamma = \Delta_{[p^\gamma, L^\gamma]}$,

cf. also (91.1a).

(c) *Uniqueness properties.* A collineation $\delta \in \Delta_{[p,L]}$ having a fixed point q outside $L \cup \{p\}$ must be the identity. Indeed, by joining q to the points of L, one sees that q is a center of δ, too; therefore, every point not on pq is a fixed point, being the intersection of two different lines through the centers p and q, and the same argument applied to such a fixed point instead of q finally shows that $\delta = \mathrm{id}$.

Dually, the only collineation in $\Delta_{[p,L]}$ leaving a line $M \neq L$ with $p \notin M$ invariant is the identity, because then M is an axis as well, its points being the points of intersection of the invariant line M and the invariant lines through the center p.

In particular, if a non-identical collineation has a center and an axis, these are uniquely determined.

We now give concrete descriptions of collineations of $\overline{\mathscr{A}_2\mathbb{F}}$ having a center and an axis. We first deal with such collineations for which center and axis are incident, so-called *elations*. The other case, in which center and axis are not incident, will be taken up later (12.13).

12.5 Proposition: Elations of $\overline{\mathscr{A}_2}\mathbb{F}$.

(a) *The collineations of* $\overline{\mathscr{A}_2}\mathbb{F}$ *having the line at infinity* $W = [\infty]$ *as axis and a center on* W *are precisely the translations*

$$\tau_{a,b} : (x, y) \mapsto (x + a, y + b);\ \ \tau_{a,b}|_W = \mathrm{id}$$

with $a, b \in \mathbb{F}$. *The translation group*

$$\mathsf{T} = \left\{ \tau_{a,b} \;\middle|\; a, b \in \mathbb{F} \right\}$$

is commutative and sharply transitive on the affine point set $\mathbb{F} \times \mathbb{F}$.

(b) *The collineations of* $\overline{\mathscr{A}_2}\mathbb{F}$ *with axis* $Y = [0]$ *and center* $v = (\infty)$ *are precisely the shears*

$$\sigma_a : (x, y) \mapsto (x, y + ax),\ \ (s) \mapsto (s + a),\ \ (\infty) \mapsto (\infty)$$

for $a \in \mathbb{F}$; *they form a commutative group which is sharply transitive on* $W \backslash \{v\}$.

(c) *The collineations of* $\overline{\mathscr{A}_2}\mathbb{F}$ *with axis* $X = [0, 0]$ *and center* $u = (0)$ *are precisely the collineations*

$$\sigma'_a : (x, y) \mapsto (x + ay, y),\ (s) \mapsto ((s^{-1} + a)^{-1}) \quad \textit{for } s \neq 0, -a^{-1},$$
$$(0) \mapsto (0),\ \ (-a^{-1}) \mapsto (\infty),\ \ (\infty) \mapsto (a^{-1})$$

for $a \in \mathbb{F} \setminus \{0\}$ *(together with* $\sigma'_0 = \mathrm{id}$).

Note that (c) is obtained from (b) by conjugation with the following collineation, see the conjugation formula in (12.4b).

12.6. *The reflection*

$$(x, y) \mapsto (y, x),\ (s) \mapsto (s^{-1}) \quad \textit{for } s \neq 0,$$
$$(0) \mapsto (\infty),\ \ (\infty) \mapsto (0)$$

is a collineation of $\overline{\mathscr{A}_2}\mathbb{F}$ *with axis* $[1, 0]$ *and center* (-1).

Proof of (12.5 and 6). 1) We first consider the restrictions of the maps in (12.5a and b) and in (12.6) to the affine plane with point set $\mathbb{F} \times \mathbb{F}$. One may easily verify by direct computation that these restrictions transform affine lines into lines and therefore are collineations of the affine plane $\mathscr{A}_2\mathbb{F}$. For (12.5a and b), this only requires the distributive laws; for (12.6), one also uses biassociativity (11.9). The necessary computations show that lines are mapped as follows:

(1) $$[s, t]^{\tau_{a,b}} = [s, t + b - sa];\quad [c]^{\tau_{a,b}} = [c + a]$$

(2) $$[s, t]^{\sigma_a} = [s + a, t];\quad [c]^{\sigma_a} = [c]\,,$$

and that the reflection of (12.6) interchanges

$$(3)\qquad \begin{array}{lll} [s,t] & \text{with} & [s^{-1}, -s^{-1}t] \quad \text{for } s \neq 0\,, \\ [0,t] & \text{with} & [t]\,. \end{array}$$

The effect of these collineations on the slopes of lines shows that their projective extensions act on the line at infinity W in the specified way.

From (3), one infers that the collineation defined in (12.6) leaves all the lines of slope -1 invariant, so that it has center (-1). It is clear that this collineation has axis $[1,0] = \{\,(x,x) \mid x \in \mathbb{F}\,\}$.

Thus, (12.6) is proved. Concerning (12.5), it now suffices to prove (a) and (b), because (c) then follows from (b), as was noted above.

2) The collineation $\tau_{a,b}$ has axis W. If $a \neq 0$ and if $s \in \mathbb{F}$ is the solution of $sa = b$, then by (1) all the lines $[s,t]$, $t \in \mathbb{F}$, are invariant under $\tau_{a,b}$. Now these are precisely the affine lines through the point (s), so that $\tau_{a,b}$ has center (s). For $a = 0$, the center is (∞), since then all the lines $[c]$ are invariant according to equation (1).

Commutativity of T and transitivity on $\mathbb{F} \times \mathbb{F}$ are obvious. We use a standard transitivity argument in order to show that T contains every collineation with axis W and center p on W. For fixed $p \in W$, the group $\mathsf{A}_{[p,W]}$ of all these collineations contains $\mathsf{T}_{[p,W]}$. Now, for any line $M \neq W$ through p, the latter group is transitive on $M \setminus \{p\}$; indeed, for two different points $q_1, q_2 \in M \setminus \{p\}$, the translation $\tau \in \mathsf{T}$ mapping q_1 to q_2 leaves $M = pq_1 = pq_2$ invariant, so that, by the uniqueness properties (12.4c), the center of τ belongs to M and therefore equals $p = M \wedge W$. Again by the uniqueness properties (12.4c), the group $\mathsf{A}_{[p,W]}$ is now seen to be sharply transitive on $M \setminus \{p\}$, so that $\mathsf{A}_{[p,W]} = \mathsf{T}_{[p,W]}$; see also (23.9). This completes the proof of (12.5a).

3) It is obvious that σ_a has axis $Y = \{0\} \times \mathbb{F}$, and from (2) one sees that all the lines $[c]$ of slope ∞, in other words the lines through $v = (\infty)$, are invariant, so that σ_a has center v. Transitivity of the group $\{\,\sigma_a \mid a \in \mathbb{F}\,\}$ on $W \setminus \{v\}$ is obvious. By the same argument as above, it follows that this group coincides with the group $\mathsf{A}_{[v,Y]}$ of all collineations with axis Y and center v, since the latter group is seen to be sharply transitive on $W \setminus \{v\}$ by virtue of the uniqueness properties (12.4c). Thus, (12.5b) is verified. □

Next, we prove some general statements about collineations having non-collinear fixed points.

12.7 Lemma: Stabilizer of a triangle. *The collineations of* $\overline{\mathcal{A}_2\mathbb{F}}$ *fixing the points* $o = (0,0)$, $u = (0)$, *and* $v = (\infty)$ *are precisely the transformations of the form*

$$(x,y) \mapsto (x^\alpha, y^\beta),\ \ (s) \mapsto (s^\gamma),\ \ (\infty) \mapsto (\infty)$$

with automorphisms α, β, γ *of the additive group of* $\mathbb{F}$ *satisfying*

$$(*) \qquad (sx)^\beta = s^\gamma x^\alpha$$

for all $s, x \in \mathbb{F}$. *In particular, these collineations are completely determined by their actions on the coordinate axes* $\mathbb{F} \times \{0\}$ *and* $\{0\} \times \mathbb{F}$.

Proof. A collineation of this kind leaves the coordinate axes $X = ou$ and $Y = ov$ invariant and maps parallels of X and Y to parallels of X and Y, respectively; this means that in the image of (x, y) the first coordinate is independent of y and the second one is independent of x, so that the collineation is of the stated form with bijections α, β, γ of $\mathbb{F}$ fixing 0. We now find the algebraic conditions for such a transformation to be a collineation. Consider the point $(x, sx+t)$ on the line $[s, t]$ joining $(0, t)$ and (s). The corresponding condition for the image points, requiring that the point $(x^\alpha, (sx+t)^\beta)$ lies on the line joining $(0, t^\beta)$ and (s^γ), that is, on the line $[s^\gamma, t^\beta]$, is expressed algebraically by

$$(1) \qquad (sx+t)^\beta = s^\gamma x^\alpha + t^\beta .$$

The special case $t = 0$, which is just $(*)$, allows us to write (1) as $(sx+t)^\beta = (sx)^\beta + t^\beta$, thus exhibiting the additivity of β, which in turn, once more via $(*)$, implies the additivity of α and γ.

Conversely, if $(*)$ holds and if β is additive, then (1) is valid, and we have a collineation. □

12.8 Corollary: Stabilizer of a quadrangle. *The collineations of* $\overline{\mathscr{A}_2}\mathbb{F}$ *fixing the points* $o = (0, 0)$, $u = (0)$, $v = (\infty)$, *and* $e = (1, 1)$ *are precisely the transformations*

$$(x, y) \mapsto (x^\alpha, y^\alpha), \ (s) \mapsto (s^\alpha), \ (\infty) \mapsto (\infty) \quad \textit{with } \alpha \in \operatorname{Aut}\mathbb{F} .$$

Proof. These collineations are the transformations given in (12.7) which, in addition, satisfy $1^\alpha = 1 = 1^\beta$. From (12.7(∗)) it then follows that also $1^\gamma = 1$, that $\alpha = \beta = \gamma$, and that, consequently, α is an automorphism of $\mathbb{F}$. □

12.9 Remark. *The collineations in* (12.8) *coincide with the affine collineations fixing the points* o, $e_1 = (1, 0)$, *and* $e_2 = (0, 1)$.

Indeed, if we consider them as collineations of the projective completion $\overline{\mathscr{A}_2}\mathbb{F}$, the line at infinity W remains invariant; now $u = (oe_1) \wedge W$, $v = (oe_2) \wedge W$, $e = (e_1 v) \wedge (e_2 u)$, and conversely $e_1 = (ve) \wedge (ou)$, $e_2 = (ue) \wedge (ov)$. □

In the field case $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}, \mathbb{H}\}$, the affine collineation group can now be determined completely.

12.10 Fundamental theorem of affine geometry. *Let* $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}, \mathbb{H}\}$. *The collineations of the affine plane* $\mathscr{A}_2\mathbb{F}$ *are precisely the mappings*

$$(x, y) \mapsto \varphi(x^\alpha, y^\alpha) + (a, b)$$

for $a, b \in \mathbb{F}$, $\alpha \in \operatorname{Aut}\mathbb{F}$, *and* $\varphi \in \mathrm{GL}_2\mathbb{F}$ *(the group of linear transformations of* $\mathbb{F} \times \mathbb{F}$ *considered as a* right $\mathbb{F}$*-vector space). They form the group* $\mathsf{A} = \mathrm{A\Gamma L}_2\mathbb{F}$, *which is the semidirect product* $\mathsf{A} = \mathsf{T} \rtimes \mathsf{A}_o$ *of the translation group* T *described in* (12.5) *by the stabilizer of the origin*

$$\mathsf{A}_o = \Gamma\mathrm{L}_2\mathbb{F} = \operatorname{Aut}\mathbb{F} \ltimes \mathrm{GL}_2\mathbb{F}\,.$$

Remark. This theorem and the proof given below are in fact valid over any (not necessarily commutative) field $\mathbb{F}$, cf. (23.6) and the references given there.

Proof of (12.10). Let A denote the group of all affine collineations. As T is the group of all collineations with axis W and center on W, see (12.5), the conjugation formula (12.4b) implies that T is a normal subgroup of A . By transitivity of T on the affine point set, the Frattini argument (91.2a) shows that $\mathsf{A} = \mathsf{A}_o \cdot \mathsf{T} = \mathsf{T} \cdot \mathsf{A}_o$.

The group $\mathrm{GL}_2\mathbb{F}$ consists of collineations because in the field case the lines are just the one-dimensional affine subspaces of the *right* $\mathbb{F}$-vector space $\mathbb{F} \times \mathbb{F}$, as is immediate from their definition (12.1), so that linear transformations map lines to lines. Now $\mathrm{GL}_2\mathbb{F}$ is sharply transitive on the set of bases of $\mathbb{F} \times \mathbb{F}$, whence $\mathsf{A}_o = \mathsf{A}_{o,(1,0),(0,1)} \cdot \mathrm{GL}_2\mathbb{F}$ by the Frattini argument. Finally, the stabilizer $\mathsf{A}_{o,(1,0),(0,1)}$ is described by $\operatorname{Aut}\mathbb{F}$ according to (12.9 and 8). □

We now determine how these affine collineations act on the line at infinity W. Since $\mathsf{A} = \mathsf{T} \cdot \mathsf{A}_o$, and since the translation group T acts trivially on W, only the collineations fixing o need to be considered.

12.11. *Let* $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}, \mathbb{H}\}$. *The stabilizer* A_o *consists of all the mappings*

$$\gamma_{\alpha,\varphi} : (x, y) \mapsto \varphi(x^\alpha, y^\alpha)$$

for $\alpha \in \operatorname{Aut}\mathbb{F}$ *and* $\varphi \in \mathrm{GL}_2\mathbb{F}$. *If* φ *is described by a matrix as*

$$\varphi : \mathbb{F}^2 \to \mathbb{F}^2 : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix},$$

then the projective extension of the collineation $\gamma_{\alpha,\varphi}$ *acts on the line at infinity as the fractional semilinear transformation*

$$\gamma_{\alpha,\varphi}|_W : W \to W : (s) \mapsto \left((c + d \cdot s^\alpha)(a + b \cdot s^\alpha)^{-1}\right)\,.$$

This should be understood also for $s = 0$ and $s = \infty$, with the usual conventions about the rôle of 0 and ∞ in such expressions, e.g. with $0^{-1} = \infty$, $\infty^{-1} = 0$.

In order to obtain the asserted description of $\gamma_{\alpha,\varphi}|_W$, we decompose the given collineation as $\gamma_{\alpha,\varphi} = \gamma_{\alpha,\mathrm{id}}\,\gamma_{\mathrm{id},\varphi}$. The projective extension of $\gamma_{\alpha,\mathrm{id}}$ is the collineation described in (12.8) mapping (s) to (s^α). The projective extension of $\gamma_{\mathrm{id},\varphi}$ maps (s) to (s'), where s' is the slope of the image of the line $[s,0] = \{\,(x,sx) \mid x \in \mathbb{F}\,\} = (1,s)\mathbb{F}$. The image line is $((a+bs),(c+ds))\mathbb{F} = (1,s')\mathbb{F}$, so that $s' = (c+ds)(a+bs)^{-1}$. The description of $\gamma_{\alpha,\varphi}|_W$ now follows by composition. □

We thus have obtained the following result.

Fundamental theorem for the projective line. *Let* $\mathbb{F} \in \{\mathbb{R},\mathbb{C},\mathbb{H}\}$. *The group* $\mathsf{A}|_W$ *of transformations induced on the line at infinity* W *by the group* A *of collineations of* $\mathscr{A}_2\mathbb{F}$ *(via projective extension) is the product*

$$\mathsf{A}|_W = \operatorname{Aut}\mathbb{F} \cdot \mathrm{PGL}_2\mathbb{F}$$

of the group

$$\mathrm{PGL}_2\mathbb{F} = \left\{\, (s) \mapsto \left((c+ds)(a+bs)^{-1}\right) \;\middle|\; \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{GL}_2\mathbb{F} \,\right\}$$

of fractional linear transformations by the group

$$\{\,(s) \mapsto (s^\alpha),\ (\infty) \mapsto (\infty) \mid \alpha \in \operatorname{Aut}\mathbb{F}\,\} \cong \operatorname{Aut}\mathbb{F}\,.$$

Note. In (15.6), we shall present a version of the preceding theorem which is valid for $\mathbb{F} = \mathbb{O}$, as well.

Addenda.

(a) *The last-mentioned group is the stabilizer of* (0), (1), *and* (∞) *in* $\mathsf{A}|_W$.

(b) *The stabilizer of* (0) *and* (∞) *in* $\mathrm{PGL}_2\mathbb{F}$ *consists of the transformations* $(s) \mapsto (dsa^{-1})$ *for* $a,d \in \mathbb{F}^\times$. *These are precisely the maps* $(s) \mapsto (Bs)$ *where* B *belongs to the group* $\mathrm{GO}_n^+\mathbb{R}$ *defined in* (11.21).

(c) *The fractional linear group* $\mathrm{PGL}_2\mathbb{F}$ *is triply transitive on* W. *It is sharply triply transitive if, and only if* $\mathbb{F}$ *is commutative.*

(d) $\operatorname{Aut}\mathbb{F}$ *is trivial for* $\mathbb{F} = \mathbb{R}$. *In the case* $\mathbb{F} = \mathbb{H}$, *the transformation* $(s) \mapsto (s^\alpha)$ *for* $\alpha \in \operatorname{Aut}\mathbb{F}$ *belongs to* $\mathrm{PGL}_2\mathbb{H}$. *Thus,*

$$\mathsf{A}|_W = \mathrm{PGL}_2\mathbb{F} \quad \textit{for } \mathbb{F} = \mathbb{R},\mathbb{H}\,.$$

Proof of Addenda. We begin with assertion (b). The first part of the assertion is immediate from the explicit description of $\mathrm{PGL}_2\mathbb{F}$. The second part has been proved in (11.23).

We now turn to (c). The stabilizer of (∞) in $\mathrm{PGL}_2\mathbb{F}$ contains the transformations $(s) \mapsto (s+c)$ for $c \in \mathbb{F}$ and hence is transitive on $W \setminus (\infty)$. Under $(s) \mapsto (s^{-1})$, the points (0) and (∞) are interchanged. Hence, the group $\mathrm{PGL}_2\mathbb{F}$ is doubly transitive on W, and (b) shows that it is even triply transitive. Also from (b), we infer that the stabilizer Λ of the triple (0), (1), (∞) in $\mathrm{PGL}_2\mathbb{F}$ is induced by the inner automorphisms of $\mathbb{F}$, so that this stabilizer is trivial precisely if $\mathbb{F}$ is commutative. Thus, (c) is proved. At the same time, we have obtained that this stabilizer is contained in the subgroup of $\mathsf{A}|_W$ corresponding to $\operatorname{Aut}\mathbb{F}$, which proves (a). Moreover, the latter subgroup reduces to $\Lambda \subseteq \mathrm{PGL}_2\mathbb{F}$ whenever all automorphisms of $\mathbb{F}$ are inner, so that then $\mathsf{A}|_W = \mathrm{PGL}_2\mathbb{F}$. This is the case for $\mathbb{F} = \mathbb{H}$ by (11.29), whence (d). $\square$

From now on, we include the case $\mathbb{F} = \mathbb{O}$ in our discussion again. Our next topic is a concrete description of collineations with *non*-incident center and axis. Such collineations are called *homologies*.

12.13 Proposition: Homologies of $\overline{\mathscr{A}_2}\mathbb{F}$.

(a) *The collineations of* $\overline{\mathscr{A}_2}\mathbb{F}$ *having the line at infinity* $W = [\infty]$ *as axis and fixing the origin* $o = (0,0)$ *are precisely the mappings*

$$\mu_a : (x, y) \mapsto (xa, ya);\ \ \mu_a|_W = \mathrm{id}$$

$$\text{for } 0 \neq a \in \begin{cases} \mathbb{F} & \text{if } \mathbb{F} \in \{\mathbb{R}, \mathbb{C}, \mathbb{H}\} \\ \operatorname{Ker}\mathbb{O} = \mathbb{R} & \text{if } \mathbb{F} = \mathbb{O}. \end{cases}$$

They have center o. *These collineations are called* homotheties.

In the field case $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}, \mathbb{H}\}$, *the group of these homotheties is transitive on* $X \setminus \{o, X \wedge W\}$. *For* $\mathbb{F} = \mathbb{O}$, *this is not so.*

(b) *The collineations of* $\overline{\mathscr{A}_2}\mathbb{F}$ *with axis* $X = [0,0]$ *and center* $v = (\infty)$ *are precisely the mappings*

$$(x, y) \mapsto (x, ay),\ \ (s) \mapsto (as)$$

for $0 \neq a \in \mathbb{F}$ *if* $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}, \mathbb{H}\}$, *and for* $0 \neq a \in \mathbb{R}$ *if* $\mathbb{F} = \mathbb{O}$.

(c) *The collineations of* $\overline{\mathscr{A}_2}\mathbb{F}$ *with axis* $Y = [0]$ *and center* $u = (0)$ *are precisely the mappings*

$$(x, y) \mapsto (ax, y),\ \ (s) \mapsto (sa^{-1})$$

for $0 \neq a \in \mathbb{F}$ *if* $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}, \mathbb{H}\}$, *and for* $0 \neq a \in \mathbb{R}$ *if* $\mathbb{F} = \mathbb{O}$.

Remark. The difference between $\mathbb{O}$ and the fields $\mathbb{R}, \mathbb{C}, \mathbb{H}$ exhibited here reflects the non-associativity of $\mathbb{O}$, as will become clear in the proof. Synthetically, this difference is expressed by the fact that Desargues' theorem for non-incident centers and axes, which holds in planes over (not necessarily commutative) fields, is not valid in the octonion plane. The fundamental interrelations between transitivity properties of groups of homologies, the validity of Desargues' theorem, and associativity properties of the coordinate domain are due to Baer [42] Theorem 5.1, p. 146 and Theorem 6.2, p. 151, cf. also (23.22) and the books mentioned there.

Proof of (12.13). (a) The collineations in question are precisely the transformations of the form stated in (12.7) with $\gamma = \mathrm{id}$, i.e., the collineations

$$\delta : (x, y) \mapsto (x^\alpha, y^\beta);\ \ \delta|_W = \mathrm{id}$$

with bijections α, β of $\mathbb{F}$ satisfying

$$(*) \qquad (sx)^\beta = s \cdot x^\alpha$$

for all $s, x \in \mathbb{F}$. With $x = 1$ and $a := 1^\alpha$ this gives $s^\beta = sa$; and putting $s = 1$ we obtain $\beta = \alpha$. Thus, δ is of the form

$$(1) \qquad \delta : (x, y) \mapsto (xa, ya)\ ,$$

and $(*)$ requires that

$$(2) \qquad (sx)a = s(xa)$$

for all $s, x \in \mathbb{F}$.

Now, if $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}, \mathbb{H}\}$, then every element a satisfies (2) because of associativity; if $\mathbb{F} = \mathbb{O}$, then (2) holds precisely for the elements a of the kernel $\operatorname{Ker}\mathbb{O} = \mathbb{R}$, see (11.20). Thus, the first part of assertion (a) is proved.

For $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}, \mathbb{H}\}$, the easy direction of the preceding argument would suffice, namely the verification that the homotheties μ_a are indeed collineations for all $a \in \mathbb{F} \setminus \{0\}$. The proof can then be completed by the following standard transitivity argument. Obviously, in these cases, the set $\{\, \mu_a \mid a \in \mathbb{F} \setminus \{0\}\,\}$ of homotheties is transitive on $X \setminus \{o, X \wedge W\} = \mathbb{F} \times \{0\} \setminus \{(0, 0)\}$; one now observes that, by the uniqueness properties (12.4c), the group $\mathsf{A}_{[o,W]}$ of all homologies with center o and axis W is sharply transitive there and thus consists precisely of the homotheties as stated. See also (23.9).

A further proof for $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}, \mathbb{H}\}$ can be obtained from the fundamental theorem of affine geometry (12.10). The collineations fixing the origin are given in (12.11) together with their actions on W; the collineation $\gamma_{\alpha,\varphi}$ described there has axis W if, and only if $c + d \cdot s^\alpha = s(a + b \cdot s^\alpha)$ for all $s \in \mathbb{F}$. By putting in turn $s = 0, s = 1$, and $s = -1$ one sees that this is equivalent to $c = 0 = b$, $d = a$ and $s^\alpha = a^{-1}sa$ for all $s \in \mathbb{F}$. In this situation, $\gamma_{\alpha,\varphi}$ maps (x, y) to $\varphi(x^\alpha, y^\alpha) = (a \cdot x^\alpha, a \cdot y^\alpha) = (a \cdot (a^{-1}xa), a \cdot (a^{-1}xa)) = (xa, ya)$; this proves our proposition anew. Note that, in the non-commutative case $\mathbb{F} = \mathbb{H}$ at hand, $\gamma_{\alpha,\varphi}$ is not a linear transformation (of the affine point set $\mathbb{H}^2$ considered as a *right* $\mathbb{H}$-vector space)!

(b) may be proved analogously. The collineations with axis X and center v are the mappings

$$(x, y) \mapsto (x, y^{\beta}), \ \ (s) \mapsto (s^{\gamma})$$

with bijections β, γ of $\mathbb{F}$ satisfying $(sx)^{\beta} = s^{\gamma} x$ for all $s, x \in \mathbb{F}$. With $s = 1$ and $c := 1^{\gamma}$, we obtain $x^{\beta} = cx$. Putting $x = 1$, we infer $\gamma = \beta$. Thus, the collineations in question are the transformations $(x, y) \mapsto (x, cy)$, where $c \neq 0$ satisfies

$$(3) \qquad c(sx) = (cs)x \quad \text{for all } s, x \in \mathbb{F} .$$

For $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}, \mathbb{H}\}$, the latter condition is trivially fulfilled because of associativity. For $\mathbb{F} = \mathbb{O}$, by applying conjugation to (3), we obtain the equivalent condition $(\bar{x}\,\bar{s})\bar{c} = \bar{x}(\bar{s}\bar{c})$ for all $s, x \in \mathbb{F}$, which says that $\bar{c}$, and hence c, belongs to the kernel $\mathbb{R}$ of $\mathbb{O}$.

In the field cases $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}, \mathbb{H}\}$, analogous variations of the proof as in (a) are possible, of course.

(c) is obtained from (b) by conjugation; one applies the conjugation formula (12.4b) to the reflection $(x, y) \mapsto (y, x)$, $(s) \mapsto (s^{-1})$ of (12.6), which interchanges the axis X with Y and the center u with v. □

Collineations of $\mathscr{A}_2\mathbb{O}$, Moufang identities, and triality

Finally, we proceed to a more detailed study of the stabilizer ∇ of the triangle $o = (0, 0)$, $u = (0)$, and $v = (\infty)$ in the collineation group of the projective octonion plane $\overline{\mathscr{A}_2\mathbb{O}}$. The group ∇ can also be viewed as the stabilizer of the coordinate axes $X = ou$ and $Y = ov$ in the group of collineations of the affine plane $\mathscr{A}_2\mathbb{O}$. This stabilizer is an easily accessible part of the collineation group. A study of the whole collineation group of the projective plane $\overline{\mathscr{A}_2\mathbb{O}}$ will be the subject of separate sections (17 and 18).

First, we exhibit special collineations which generate ∇ and which are closely related to the Moufang identities (12.15). By an idea of Salzmann [59c], these collineations can be constructed very easily as compositions of elations which are known from (12.5). When translating the geometric fact that these transformations are collineations into algebraic language, one immediately obtains the Moufang identities. They express other facets of the weak associativity properties of $\mathbb{O}$ besides alternativity (11.13). Actually, they are inherent in alternativity, although this is not entirely obvious. Here, this fact may be obtained by inspection of our argument. For purely algebraic proofs, see Pickert [75] 6.1 no. 2, p. 160, Schafer [66] III.1, p. 28, Zhevlakov et al. [82] Lemma 2.7, p. 35.

With the special collineations mentioned above at our disposal, it is then easy to determine ∇. The structure of ∇ is closely related to the triality principle (12.18).

As a special trait of our presentation, triality does not appear as an algebraic fact having geometric applications, but is intertwined with geometric phenomena right from the start.

Our arguments would be valid more generally for any alternative field instead of $\mathbb{O}$. In particular, they work over (not necessarily commutative) fields, as well; but then, the collineations in question may also be obtained directly from the fundamental theorem of affine geometry (12.10), see (12.19), and the Moufang identities are trivially satisfied because of associativity.

12.14 Proposition. *The following mappings are collineations of* $\overline{\mathcal{A}_2\mathbb{O}}$*, for every* $a \in \mathbb{O} \setminus \{0\}$:

$$\begin{aligned} \gamma_a &: (x, y) \mapsto (a^{-1}x, ay), \ (s) \mapsto (asa) \\ \gamma'_a &: (x, y) \mapsto (axa, ya), \ \ (s) \mapsto (sa^{-1}) \\ \gamma''_a &: (x, y) \mapsto (xa, aya), \ \ (s) \mapsto (as)\,. \end{aligned}$$

Proof. We use the elations $\sigma_a : (x, y) \mapsto (x, y + ax)$, $\sigma'_a : (x, y) \mapsto (x + ay, y)$ from (12.5b and c) and check that γ_a is the following product of such elations:

$$(*) \qquad \gamma_a = \sigma'_1\, \sigma_{a-1}\, \sigma'_{-a^{-1}}\, \sigma_{a-a^2}\,.$$

The biassociativity of $\mathbb{O}$ (11.9) will be essential for the necessary calculations. Under the given product of elations, an affine point is mapped as follows:

$$\begin{aligned} (x, y) &\mapsto (x + y, y) \\ &\mapsto (x + y, y + (a - 1)(x + y)) = (x + y, a(x + y) - x) \\ &\mapsto \big(x + y - a^{-1}(a(x + y)) + a^{-1}x, a(x + y) - x\big) = \big(a^{-1}x, a(x + y) - x\big) \\ &\mapsto \big(a^{-1}x, a(x + y) - x + (a - a^2)(a^{-1}x)\big) = \big(a^{-1}x, a(x + y) - x + x - ax\big) \\ &= (a^{-1}x, ay)\,. \end{aligned}$$

The line $[s, 0]$ joining $(0, 0)$ and $(1, s)$ is mapped onto the line joining the image points $(0, 0)$ and (a^{-1}, as); this line is $[asa, 0]$, since $(asa)a^{-1} = as$ by biassociativity, so that indeed $(a^{-1}, as) \in [asa, 0]$. The projective extension thus maps (s) to (asa), and $(*)$ is proved.

According to (12.7), the fact that γ_a is a collineation is expressed algebraically by the identity $a(sx) = (asa)(a^{-1}x)$. Substituting $z = a^{-1}x$, $az = x$ yields the first *Moufang identity*

$$(1) \qquad a(s(az)) = (asa)z\,.$$

Applying conjugation on both sides of (1) and renaming, one obtains the second Moufang identity

$$(2) \qquad ((ba)x)a = b(axa)\,.$$

With $s = ba$, $sa^{-1} = b$ this is transformed into the identity $(sx)a = (sa^{-1})(axa)$, from which in turn, by applying (12.7) backwards, it follows that γ'_a is a collineation. An immediate verification with the help of biassociativity shows that $\gamma''_a = \gamma_a\gamma'_a$, so that this is a collineation as well. By (12.7) once again, the latter fact is equivalent to the third Moufang identity

$$a(sx)a = (as)(xa)\ . \tag{3}$$

We collect these identities as a corollary to the proof.

12.15 The Moufang identities. *For all $a, b, c \in \mathbb{O}$, the following hold*:

$$\begin{aligned} a(b(ac)) &= (aba)c \\ ((ab)c)b &= a(bcb) \\ (ab)(ca) &= a(bc)a\ . \end{aligned}$$

□

12.16 Lemma. *The group of collineations generated by the collineations γ_a and γ'_a of* (12.14) *for $a \in \mathbb{O} \setminus \{0\}$ fixes the vertices of the triangle o, u, v and is transitive on the set $(\mathbb{O}\setminus\{0\}) \times (\mathbb{O}\setminus\{0\})$ of affine points not incident with the sides of this triangle.*

Proof. For $a, b \in \mathbb{O}\setminus\{0\}$, we show that the point $(1, 1)$ may be mapped to the point (a, b) by a product of collineations of the specified kind. Indeed, since ba belongs to a subfield of $\mathbb{O}$ isomorphic to $\mathbb{C}$, see (11.7), there is $c \in \mathbb{O}$ with $ba = c^3$. Now, using alternativity (11.8), one easily verifies that $\gamma_{c^{-1}}\gamma'_c\gamma_b$ maps $(1, 1)$ to (a, b). □

In the following proposition, we use the notation for orthogonal groups introduced in (11.21).

12.17 Proposition: The stabilizer of the coordinate axes.

(a) *The group ∇ of collineations of $\overline{\mathcal{A}_2\mathbb{O}}$ leaving the points o, u, and v fixed consists of the transformations*

$$(A, B \mid C) : (x, y) \mapsto (Ax, By),\ \ (s) \mapsto (Cs)\ ,$$

where A,B,C belong to $\mathrm{GO}_8^+\mathbb{R}$ and satisfy the triality condition

$$B(s \cdot x) = Cs \cdot Ax \quad \textit{for all } s, x \in \mathbb{O}\ . \tag{$*$}$$

(b) *The group ∇ is the direct product of the subgroup consisting of the transformations*

$$\mu_{r,t} : (x, y) \mapsto (rx, ty),\ \ (s) \mapsto (tr^{-1} \cdot s) \quad \textit{for } 0 < r, t \in \mathbb{R}$$

! "
-./0- - -1 ! - , 5 6 ; " ! ! !
. . ! < 4"
& + !4 4 4

CHAPTER 2

Some Properties of Circles

> **Although the Greeks worked fruitfully, not only in geometry but also in the most varied fields of mathematics, nevertheless we today have gone beyond them everywhere and certainly also in geometry.**
>
> ***F. Klein***

The circle has been held in highest esteem through the ages. Its perfect form has affected philosophers and astronomers alike. Until Kepler derived his laws, the thought that planets might move in anything but circular paths was unthinkable. Nowadays, the words "square", "line", and the like sometimes have derogatory connotations, but the circle—never. Cleared of superstitious nonsense and pseudo-science, it still stands out, as estimable as ever.

Limitations of space make it impossible for us to present more than a few of the most interesting properties developed since Euclid of the circle and its relation to triangles and other polygons.

2.1 The power of a point with respect to a circle

We begin our investigations by recalling two of Euclid's theorems: III.35, about the product of the parts into which two chords of a circle divide each other (that is, in the notation of Figure 2.1A, $PA \times PA' = PB \times PB'$), and III.36, comparing a secant and a tangent drawn from the same point P outside the circle (in Figure 2.1B, $PA \times PA' = PT^2$). If we agree to regard a tangent as the limiting form of a secant, we can combine these results as follows:

THEOREM 2.11. *If two lines through a point* P *meet a circle at points* A, A' *(possibly coincident) and* B, B' *(possibly coincident), respectively, then* $PA \times PA' = PB \times PB'$.

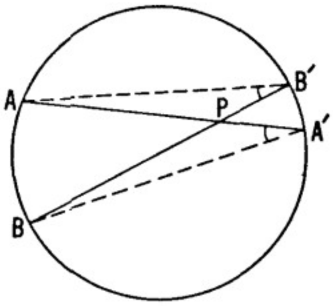

Figure 2.1A

For a proof we merely have to observe that the similar triangles PAB' and PBA' (with a common angle at P) yield

$$\frac{PA}{PB'} = \frac{PB}{PA'}.$$

In Figure 2.1B, we can equally well use the similar triangles PAT and PTA' to obtain

$$\frac{PA}{PT} = \frac{PT}{PA'},$$

and then say $PA \times PA' = PT^2 = PB \times PB'$.

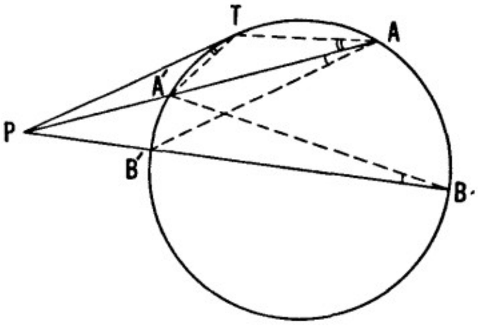

Figure 2.1B

Let R denote the radius of the circle, and d the distance from P to the center. By taking BB' to be the diameter through P (with B farther from P than B'), we see that, if P is inside the circle (as in Figure 2.1A),

$$AP \times PA' = BP \times PB' = (R+d)(R-d) = R^2 - d^2,$$

and if P is outside (as in Figure 2.1B),

$$PA \times PA' = PB \times PB' = (d+R)(d-R) = d^2 - R^2.$$

The equation

$$AP \times PA' = R^2 - d^2$$

provides a quick proof of a formula due to Euler:

THEOREM 2.12. *Let O and I be the circumcenter and incenter, respectively, of a triangle with circumradius R and inradius r; let d be the distance OI. Then*

$$d^2 = R^2 - 2rR. \tag{2.12}$$

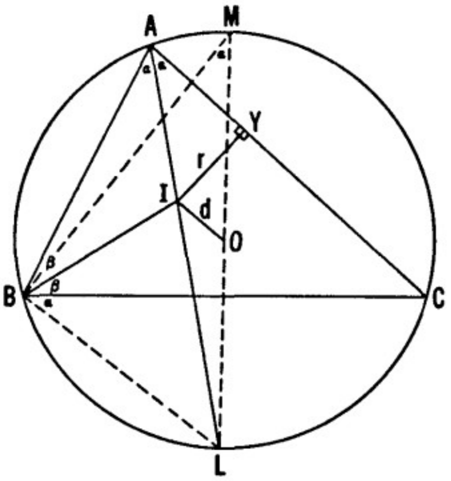

Figure 2.1C

Figure 2.1C shows the internal bisector of $\angle A$ extended to meet the circumcircle at L, the midpoint of the arc BC not containing A. LM is the diameter perpendicular to BC. Writing, for convenience, $\alpha = \frac{1}{2}A$ and $\beta = \frac{1}{2}B$, we notice that

$$\angle BML = \angle BAL = \alpha, \quad \text{and} \quad \angle LBC = \angle LAC = \alpha.$$

Since the exterior angle of $\triangle ABI$ at I is

$$\angle BIL = \alpha + \beta = \angle LBI,$$

$\triangle LBI$ is isosceles: $LI = LB$. Thus

$$\begin{aligned} R^2 - d^2 &= LI \times IA = LB \times IA \\ &= LM \frac{LB/LM}{IY/IA} IY = LM \frac{\sin\alpha}{\sin\alpha} IY \\ &= LM \times IY = 2Rr, \end{aligned}$$

that is, $d^2 = R^2 - 2rR$, as we wished to prove.

For any circle of radius R and any point P distant d from the center, we call

$$d^2 - R^2$$

the *power* of P with respect to the circle. It is clearly positive when P is outside, zero when P lies on the circumference, and negative when P is inside. For the first of these cases we have already obtained the alternative expression

$$PA \times PA',$$

where A and A' are any two points on the circle, collinear with P (as in Theorem 2.11). This expression for the power of a point P remains valid for all positions of P if we agree to adopt Newton's idea of *directed* line segments: a kind of one-dimensional vector algebra in which

$$AP = -PA.$$

The product (or quotient) of two directed segments on one line is regarded as being positive or negative according as the directions agree or disagree. With this convention, the equation

$$d^2 - R^2 = PA \times PA'$$

holds universally. If P is inside the circle,

$$d^2 - R^2 = -(R^2 - d^2) = -AP \times PA' = PA \times PA';$$

and if P is on the circumference, either A or A' coincides with P, so that one of the segments has length zero. In fact, after observing that the product $PA \times PA'$ has the same value for every secant (or chord) through P, we could have used this value as a *definition* for the power of P with respect to the circle.

The word *power* was first used in this sense by Jacob Steiner, whose name has already appeared in Chapter 1.

EXERCISES

1. What is the (algebraically) smallest possible value that the power of a point can have with respect to a circle of given radius R? Which point has this critical power?

2. What is the locus of points of constant power (greater than $-R^2$) with respect to a given circle?

3. If the power of a point has the positive value t^2, interpret the length t geometrically.

4. If PT and PU are tangents from P to two concentric circles, with T

on the smaller, and if the segment PT meets the larger circle at Q, then $PT^2 - PU^2 = QT^2$.

5. The circumradius of a triangle is at least twice the inradius.

6. Express (in terms of r and R) the power of the incenter with respect to the circumcircle.

7. The notation of directed segments enables us to express Stewart's theorem (Exercise 4 of Section 1.2) in the following symmetrical form [5, p. 152]: *If P, A, B, C are four points of which the last three are collinear, then* $PA^2 \times BC + PB^2 \times CA + PC^2 \times AB + BC \times CA \times AB = 0$.

8. A line through the centroid G of ΔABC intersects the sides of the triangle at points X, Y, Z. Using the concept of directed line segments, prove that
$$\frac{1}{GX} + \frac{1}{GY} + \frac{1}{GZ} = 0.$$

9. How far away is the horizon as seen from the top of a mountain one mile high? (Assume the earth to be a sphere of diameter 7920 miles.)

2.2 The radical axis of two circles

The following anecdote was related by E. T. Bell [3, p. 48]. Young Princess Elisabeth, exiled from Bohemia, had successfully attacked a problem in elementary geometry by using coordinates. As Bell states it, "The problem is a fine specimen of the sort that are *not* adapted to the crude brute force of elementary Cartesian geometry." Her teacher was René Descartes (after whom Cartesian coordinates were named†). His reaction was that "he would not undertake to carry out her solution . . . in a month."

The lesson is clear: a solution that is possible in a certain manner may still not be the best or most economical one. At any rate, here is one theorem for which an analytic proof, without being any more difficult than the usual synthetic proof [6, p. 86], has some interesting repercussions:

THEOREM 2.21. *The locus of all points whose powers with respect to two nonconcentric circles are equal is a line perpendicular to the line of centers of the two circles.*

---

† There are some who claim that it was Pierre Fermat (1601–1665) who actually invented analytic geometry. Their contention is that he gave the essential idea to Descartes in a letter.

In terms of rectangular Cartesian coordinates, the square of the distance d between any two points (x, y) and (a, b) is

$$(x - a)^2 + (y - b)^2.$$

Therefore the power of (x, y) with respect to the circle with center (a, b) and radius r is

$$d^2 - r^2 = (x - a)^2 + (y - b)^2 - r^2.$$

In particular, the circle itself, being the locus of points (x, y) of power zero, has the equation

$$(2.22) \qquad (x - a)^2 + (y - b)^2 - r^2 = 0.$$

The same equation, in the form $(x - a)^2 + (y - b)^2 = r^2$, expresses the circle as the locus of points whose distances from (a, b) have the constant value r.

When this circle is expressed in the form

$$(2.23) \qquad x^2 + y^2 - 2ax - 2by + c = 0$$

(where $c = a^2 + b^2 - r^2$), the power of an arbitrary point (x, y) is again expressed by the left side of the equation, namely

$$x^2 + y^2 - 2ax - 2by + c.$$

Another circle having the same center (a, b) but a different radius has an equation of the same form with a different c, and any circle having a different center has an equation of the form

$$(2.24) \qquad x^2 + y^2 - 2a'x - 2b'y + c' = 0,$$

where either $a' \neq a$ or $b' \neq b$ or both. We are thus free to use the equations (2.23) and (2.24) for the two non-concentric circles mentioned in Theorem 2.21. The locus of all points (x, y) whose powers with respect to these two circles are equal is

$$x^2 + y^2 - 2ax - 2by + c = x^2 + y^2 - 2a'x - 2b'y + c'.$$

Since $x^2 + y^2$ cancels, this locus is the *line*

$$(a' - a)x + (b' - b)y = \tfrac{1}{2}(c' - c).$$

By choosing our frame of reference so that the x-axis joins the two centers, we may express the two circles in the simpler form

$$(2.25) \quad x^2 + y^2 - 2ax + c = 0, \qquad x^2 + y^2 - 2a'x + c' = 0,$$

where $a' \neq a$. Then the locus becomes

$$x = \frac{c' - c}{2(a' - a)}.$$

This line, being parallel to the y-axis, is *perpendicular* to the x-axis, which is the join of centers. Since the line can be defined geometrically in terms of the circles (as containing all points of equal power), we could have taken it to be the y-axis itself, as in Figure 2.2A. Thus any two non-concentric circles can be expressed in the still simpler form

$$(2.26) \qquad x^2 + y^2 - 2ax + c = 0, \qquad x^2 + y^2 - 2a'x + c = 0.$$

Now the locus is $x = 0$. Conversely, every point $(0, y)$ on the line $x = 0$ has the *same* power $y^2 + c$ with respect to both circles.

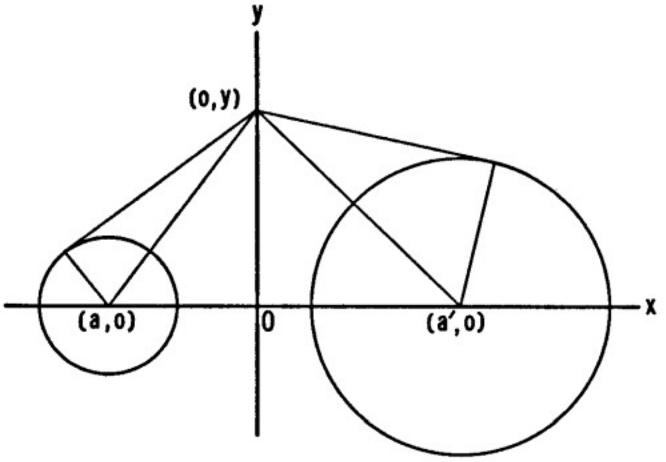

Figure 2.2A

This remark completes the proof. Of course, we could have shortened it by expressing the two circles immediately in the form (2.25); but then we would have missed the beautiful lemma that, for any circle expressed in the standard form (2.23), *the power of the general point* (x, y) *is equal to the expression on the left side of the equation.*

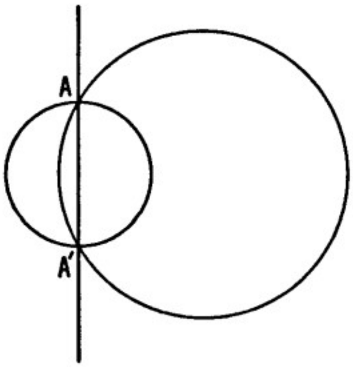

Figure 2.2B

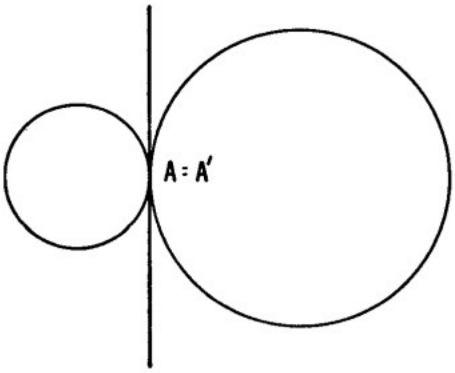

Figure 2.2C

The locus of points of equal power with respect to two non-concentric circles is called their *radical axis*. In the special case when the two circles intersect at two points A and A' (Figure 2.2B), each of these points has zero power for both circles, and therefore the radical axis is simply the line AA'. Similarly, when the two circles touch each other (Figure 2.2C), their radical axis is their common tangent at their point of contact.

EXERCISES

1. What is the locus of all points from which the tangents to two given circles have equal lengths?

2. When the distance between the centers of two circles is greater than the sum of the radii, the circles have four common tangents. The midpoints of these four line segments are collinear.

3. Let PAB, AQB, ABR, $P'BA$, $BQ'A$, BAR' be six *similar* triangles all on the same side of their common side AB. (Three of them are shown in Figure 2.2D; the rest can be derived by reflection† in the perpendicular bisector of the segment AB.) Those vertices of the triangles that do not lie on AB (namely, P, Q, R, P', Q', R') all lie on one circle. *Hint*: Compare the powers of A and B with respect to the circle PQR.

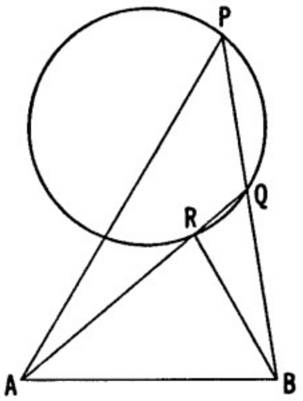

Figure 2.2D

4. Given a and b, for what values of c does the equation 2.23 represent a circle?

5. Describe a construction for the radical axis of two given non-concentric circles: a construction that remains valid when one circle encloses the other.

---

† The operation of reflection is useful in solving many geometric problems. See, for instance, Yaglom [29].

2.3 Coaxal circles

The two circles (2.26) (which may be any two non-concentric circles) are members of an infinite family, represented by the equation

$$x^2 + y^2 - 2ax + c = 0,$$

where c is fixed while a varies over the whole range of real values (except, if c is positive, the values between $\pm\sqrt{c}$). This family is called a *pencil of coaxal circles*, because every two of its members have the same line of centers and the same radical axis. If c is negative, every member of the family meets the y-axis at the same two points $(0, \pm\sqrt{-c})$, and the pencil consists simply of *all the circles through these two points*. Similarly, if $c = 0$, the pencil consists of all the circles that touch the y-axis at the origin. The case when c is positive is illustrated in Figure 2.3A.

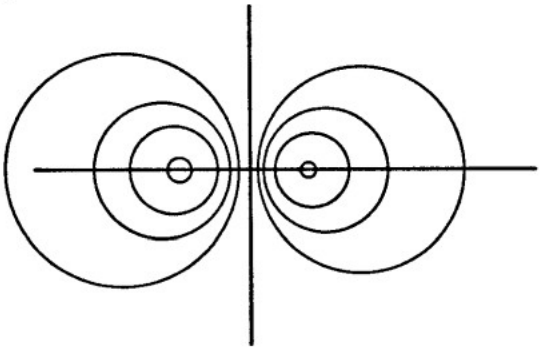

Figure 2.3A

If *three* non-coaxal circles are such that no two are concentric, we can take them in pairs and thus find three radical axes. Any point that has the same power for all three circles must lie on all three of these lines. Conversely, any point of intersection of two of the three radical axes, having the same power for all three circles, must lie on the third line as well. If two of the axes are parallel, then all three must be parallel. In particular:

THEOREM 2.31. *If the centers of three circles form a triangle, there is just one point whose powers with respect to the three circles are all equal.*

This common point of the three radical axes is called the *radical center* of the three circles.

EXERCISES

1. Two circles are in contact internally at a point T. Let the chord AB of

the larger circle be tangent to the smaller circle at a point P. Then the line TP bisects $\angle ATB$.

2. If three non-intersecting circles have radical center O, the points of contact of the six tangents from O to the circles all lie on one circle.

2.4 More on the altitudes and orthocenter of a triangle

The circumcircle of a triangle, already encountered in the previous chapters, deserves further examination. Figure 2.4A shows the circumcircle ABC with center O, diameter AA_0 through A, and radius $OL = R$ perpendicular to BC. We see also the altitude $AD = h_a$. The equal angles at B and A_0 make $\Delta ABD \sim \Delta AA_0C$, so that

$$\frac{h_a}{c} = \frac{b}{AA_0}$$

and

$$h_a = \frac{bc}{2R}. \tag{2.41}$$

Subtracting from $\angle BAC$ the two equal angles

$$\angle A_0AC = \angle BAD = 90° - B,$$

we are left with

$$\begin{aligned}\angle DAA_0 &= A - 2(90° - B) = A + 2B - (A + B + C)\\ &= B - C.\end{aligned}$$

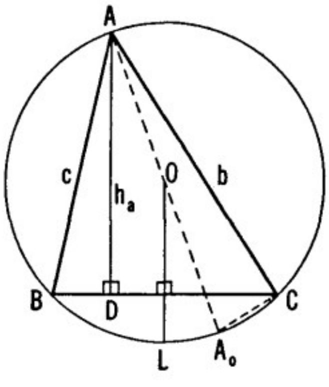

Figure 2.4A

This expression for $\angle DAA_0 = \angle DAO$ has been made with reference to the figure in which $B > C$. If instead we had taken $B < C$, the equal angles A_0AC and BAD would have overlapped, with the result that $\angle DAO = C - B$. We can include both cases by writing

$$\angle DAO = |B - C|. \tag{2.42}$$

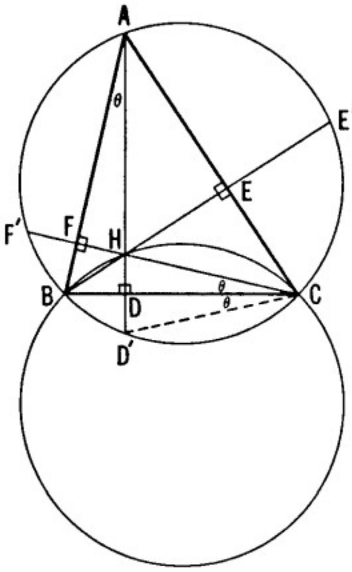

Figure 2.4B

Figure 2.4B shows the three altitudes AD, BE, CF extended to meet the circumcircle at D', E', F'. Of course, H is the orthocenter. Now $\angle DAB = \angle FCB$, both being complements of the angle B. This explains our use of the same symbol θ for both. Also $\angle BCD' = \angle BAD'$, so we have labeled $\angle BCD'$ accordingly. The congruent right-angled triangles CDH and CDD' show us that

$$HD = DD'. \tag{2.43}$$

Similarly, $HE = EE'$ and $HF = FF'$.

Since the circle with diameter AB passes through D and E, Theorem 2.11 tells us that $HA \times HD = HB \times HE$. Similarly $HB \times HE = HC \times HF$. Hence

$$HA \times HD = HB \times HE = HC \times HF. \tag{2.44}$$

If X, Y, Z are any points on the respective sides BC, CA, AB, circles constructed on the cevians AX, BY, CZ as diameters will pass through the feet of the altitudes: D, E, F, respectively. (The second and third circles are shown in Figure 2.4C.) The three expressions equated in (2.44) are the powers of H with respect to these three circles. Hence H is the radical center of the circles, and we have proved two interesting theorems that have appeared at various times as puzzlers:

THEOREM 2.45. *If circles are constructed on two cevians as diameters, their radical axis passes through the orthocenter H of the triangle.*

THEOREM 2.46. *For any three non-coaxal circles having cevians for diameters, H is their radical center.*

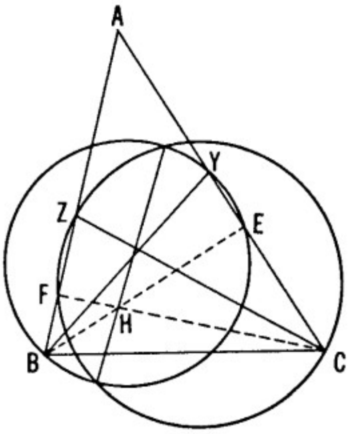

Figure 2.4C

Alternatively, the same results can be obtained by means of the following simple considerations. If AD is the altitude from A, the pencil of coaxal circles·through A and D may be described as the circles having cevians through A as diameters. Two of these cevians are the sides AB and AC. Thus the circles on BC, CA, AB as diameters have the altitudes for their radical axes in pairs, and H for their radical center. (In this manner, the concurrence of the altitudes is seen to be a special case of Theorem 2.31). It follows that H has the same power for *all* circles having cevians for diameters.

Notice the word "non-coaxal" in the statement of Theorem 2.46. This implies that the three cevians are not all drawn from the same vertex of $\triangle ABC$. We shall see in our next theorem that it implies slightly more!

Several amusing problems can be derived from Theorem 2.46 (as applied to cevians AX, BY, CZ) by introducing non-essentials. Although the three cevians need not be concurrent, it makes for greater confusion to let them be concurrent. Thus, we might ask: if circles are constructed on the medians (or altitudes, or angle bisectors) as diameters, prove that their radical center is the orthocenter of the triangle.

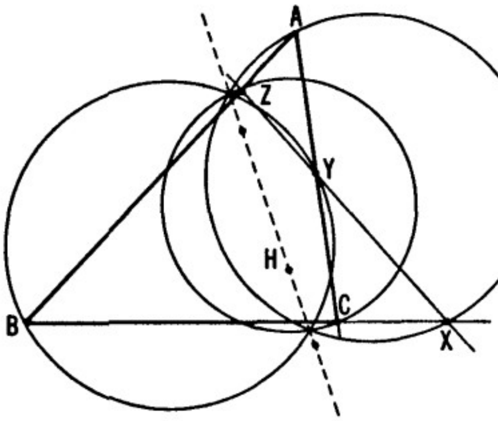

Figure 2.4D

The most interesting case of non-concurrent cevians arises when X, Y, Z are *collinear* points on the lines BC, CA, AB (extended if necessary), as in Figure 2.4D; for then we can equally well say that X, B, C are collinear points on the sides of $\triangle AYZ$, or that Y, C, A are collinear points on the sides of $\triangle BZX$, or that Z, A, B are collinear points on the sides of $\triangle CXY$. Hence the circles on AX, BY, CZ as diameters are so situated that their radical axes pass through H and also (for the same reason) through the orthocenters of the other three triangles. Since these four orthocenters are obviously distinct, the radical axes must coincide, and we have proved

THEOREM 2.47. *If four lines meet one another at six points* A, B, C, X, Y, Z, *so that the sets of collinear points are* XBC, YCA, ZAB, XYZ, *then the circles on* AX, BY, CZ *as diameters are coaxal, and the orthocenters of the four triangles* AYZ, BZX, CXY, ABC *are collinear.*

Another property of a triangle and its altitudes is illustrated in Figure 1.3C. If we inspect the diagram carefully, we reach the conclusion that, just as H is the orthocenter of $\triangle ABC$, A *is the orthocenter of* $\triangle HBC$, and, for the same reason, B is the orthocenter of $\triangle HAC$, and C is the orthocenter of $\triangle HAB$. This configuration $ABCH$ is known as an *orthocentric quadrangle*, and has a number of interesting properties. We merely examine one of them, namely: *If* $ABCH$ *is an orthocentric quadrangle, the circumcircles of the four triangles formed by taking any three of the vertices have equal radii.*

The simplest proof makes use of the equation (2.43) and Figure 2.4B. In this figure, $\triangle HBC$ and $\triangle D'BC$ are congruent, so they must have congruent circumcircles. Hence the circumcircle of $\triangle D'BC$ (or $\triangle ABC$) is congruent to the circumcircle of $\triangle HBC$, and similarly for the other triangles.

EXERCISES

1. The points where the extended altitudes meet the circumcircle form a triangle similar to the orthic triangle.

2. The internal angle bisectors of $\triangle ABC$ are extended to meet the circumcircle at points L, M, N, respectively. Find the angles of $\triangle LMN$ in terms of the angles A, B and C.

2.5 Simson lines

If perpendiculars are dropped onto the sides of a triangle ABC from a point P, the feet of these perpendiculars usually form the vertices of a triangle $A_1B_1C_1$ (the *pedal triangle* discussed in Section 1.9). Let us now examine the exceptional case where the point P lies on the circumcircle, as in Figure 2.5A. To be definite, we have taken P to lie on the arc CA that does not contain B, between A and the point diametrically opposite to B. All other cases can be derived by re-naming A, B, C. Because of the right angles at A_1, B_1 and C_1, P lies also on the circumcircles of triangles A_1BC_1, A_1B_1C and AB_1C_1. Therefore

$$\angle APC = 180° - B = \angle C_1PA_1$$

and, subtracting $\angle APA_1$, we deduce

$$\angle A_1PC = \angle C_1PA.$$

But since points A_1, C, P, B_1 lie on a circle,

$$\angle A_1PC = \angle A_1B_1C,$$

and since points A, B_1, P, C_1 lie on a circle,

$$\angle C_1PA = \angle C_1B_1A.$$

Thus

$$\angle A_1B_1C = \angle C_1B_1A,$$

so that the points A_1, B_1, C_1 are collinear; the pedal triangle is "degenerate.".

Conversely, if a point P is so situated that the pedal triangle of $\triangle ABC$ is degenerate, P must evidently lie in the region of the plane that is inside one angle of $\triangle ABC$ and beyond the opposite side. By re-naming the vertices if necessary, we can assume that this "one angle" is B, and that C_1 lies on the extension of the side BA beyond A, as in Figure 2.5A. We can then reverse the steps in the above discussion of angles and conclude that P lies on the circumcircle. Hence

THEOREM 2.51. *The feet of the perpendiculars from a point to the sides of a triangle are collinear if and only if the point lies on the circumcircle.*

The line containing the feet is known as the *Simson line* (or sometimes just the *simson*) of the point with respect to the triangle. Robert Simson (1687–1768) made several contributions to both geometry and arithmetic. For instance, it was he who discovered that, if f_n is the nth term of the Fibonacci sequence 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, ⋯ , then $f_{n-1}f_{n+1} - f_n^2 = (-1)^n$ [6, pp. 165–168]. The "simson" was attributed to him because it seemed to be typical of his geometrical ideas. However, historians have searched through his works for it in vain. Actually it was discovered in 1797 by William Wallace.

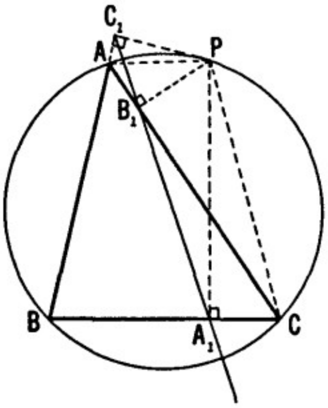

Figure 2.5A

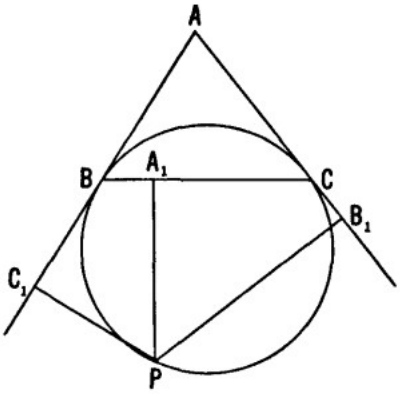

Figure 2.5B

EXERCISES

1. Does our proof of Theorem 2.51 require any modification when $\triangle ABC$ has an obtuse angle?

2. What point on the circle has CA as its Simson line?

3. Are there any points that lie on their own Simson lines? What lines are these?

4. The tangents at two points B and C on a circle meet at A. Let $A_1B_1C_1$ be the pedal triangle of the isosceles triangle ABC for an arbitrary point P on the circle, as in Figure 2.5B. Then

$$PA_1^2 = PB_1 \times PC_1.$$

2.6 Ptolemy's theorem and its extension

The concept of the Simson line can be used to derive a very useful theorem, as follows. Let us examine Figure 2.5A again. Although the "pedal triangle" $A_1B_1C_1$ is degenerate, the lengths of its "sides" are still given by Theorem 1.91:

$$B_1C_1 = \frac{aAP}{2R}, \qquad A_1C_1 = \frac{bBP}{2R}, \qquad A_1B_1 = \frac{cCP}{2R}.$$

Since $A_1B_1 + B_1C_1 = A_1C_1$, we deduce $cCP + aAP = bBP$, that is

$$AB \times CP + BC \times AP = AC \times BP.$$

Since $ABCP$ is a cyclic quadrilateral, we have thus proved Ptolemy's theorem:

THEOREM 2.61. *If a quadrilateral is inscribed in a circle, the sum of the products of the two pairs of opposite sides is equal to the product of the diagonals.*

Ptolemy's theorem has a converse that can be strengthened by observing that, for any location of B_1 other than on the segment A_1C_1, the equation $A_1B_1 + B_1C_1 = A_1C_1$ has to be replaced by the "triangle inequality"

$$A_1B_1 + B_1C_1 > A_1C_1,$$

which yields

$$AB \times CP + BC \times AP > AC \times BP.$$

Hence

THEOREM 2.62. *If ABC is a triangle and P is not on the arc CA of its circumcircle, then*

$$AB \times CP + BC \times AP > AC \times BP.$$

EXERCISES

1. Let P be any point in the plane of an equilateral triangle ABC. Then $PC + PA = PB$ or $PC + PA > PB$ according as P does or does not lie on the arc CA of the circumcircle. [For an interesting application of this result, see **23**, pp. 11–12.]

2. If a point P lies on the arc CD of the circumcircle of a square $ABCD$, then $PA(PA + PC) = PB(PB + PD)$.

3. If a circle cuts two sides and a diagonal of a parallelogram $ABCD$ at points P, R, Q as shown in Figure 2.6A, then

$$AP \times AB + AR \times AD = AQ \times AC.$$

Hint: Apply Theorem 2.61 to the quadrilateral $PQRA$ and then replace the sides of $\triangle PQR$ by the corresponding sides of the similar triangle CBA.

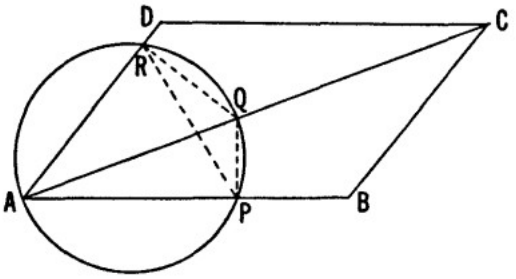

Figure 2.6A

2.7 More on Simson lines

The Simson line has many interesting properties, and it may be worthwhile to investigate some of them. Let us begin by examining Figure 2.7A, which is the same as Figure 2.5A except that the perpendicular PA_1 has been extended to meet the circumcircle at U, and the line AU has been drawn.

The cyclic quadrilaterals $PAUC$ and PB_1A_1C tell us that

$$\angle PUA = \angle PCA = \angle PCB_1 = \angle PA_1B_1.$$

Therefore *the line AU is parallel to the Simson line A_1B_1.*

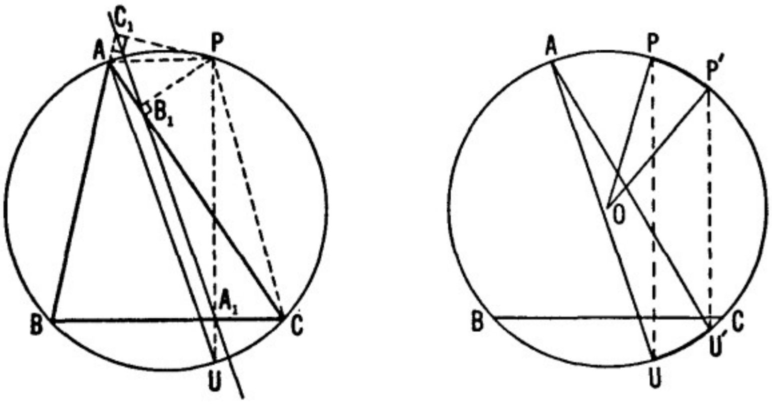

Figure 2.7A Figure 2.7B

Let us now compare the Simson line of P with the Simson line of another point P' (also, of course, on the circumcircle). The angle between these two Simson lines is simply the angle UAU' between the lines AU and AU' which are parallel to them (Figure 2.7B). The two chords PU and $P'U'$, both perpendicular to BC, are parallel to each other, and cut off equal arcs PP' and UU'. Thus

$$\angle UAU' = \tfrac{1}{2}\angle UOU' = \tfrac{1}{2}\angle POP'$$

or, if we distinguish between positive and negative angles,

$$\angle UAU' = \tfrac{1}{2}\angle UOU' = -\tfrac{1}{2}\angle POP'.$$

We have thus proved:

THEOREM 2.71. *The angle between the Simson lines of two points P and P' on the circumcircle is half the angular measure of the arc $P'P$.*

If we imagine P to run steadily round the circumcircle, the line AU will rotate steadily about A at half the angular velocity in the opposite sense, so as to reverse its direction by the time P has described the whole circumference. Meanwhile, the Simson line will turn in a corresponding manner about a continuously changing center of rotation. In fact, the Simson line envelops a beautifully symmetrical curve called a deltoid or "Steiner's hypocycloid" [**20**]. The motion is demonstrated very clearly in the film *Simson Line* by T. J. Fletcher.

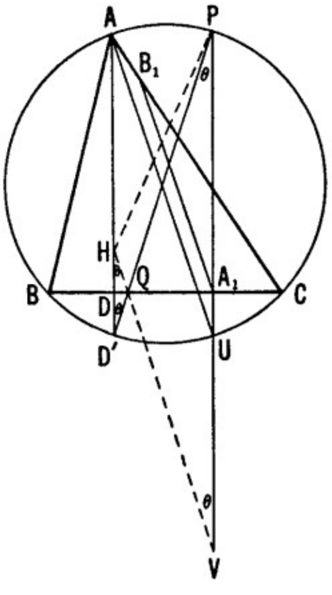

Figure 2.7C

To continue our investigation, let us now examine Figure 2.7C, which is a combination of Figures 2.4B and 2.7A with the extra lines HP, $D'P$ (meeting BC at Q) and HQ (extended to meet PU at V). Since both HD' and PV are perpendicular to BC, equation (2.43) shows that the triangles QHD' and QPV are isosceles. In other words, HV is the image of $D'P$ by reflection in BC. Since

$$\angle D'HV = \angle PVH = \angle D'PU = \angle D'AU,$$

the line HV is parallel to AU, which we have already shown to be parallel to the Simson line of P.

Finally, we observe that, in $\triangle PHV$, the Simson line A_1B_1 is parallel to the side HV and bisects the side PV (at A_1). Hence it must also bisect the remaining side PH:

THEOREM 2.72. *The Simson line of a point (on the circumcircle) bisects the segment joining that point to the orthocenter.*

This has been merely an introduction to the topic of Simson lines. They have many other properties which we must regretfully leave to other sources.

EXERCISES

1. The Simson lines of diametrically opposite points on the circumcircle are perpendicular to each other and meet on the nine-point circle.

2. Let ABC be an equilateral triangle inscribed in a circle with center O, and let P be any point on the circle. Then the Simson line of P bisects the radius OP.

2.8 The Butterfly

The Butterfly theorem has been around for quite a while. We state it as follows (see Figure 2.8A):

THEOREM 2.81. *Through the midpoint M of a chord PQ of a circle, any other chords AB and CD are drawn; chords AD and BC meet PQ at points X and Y. Then M is the midpoint of XY.*

For this theorem numerous proofs have been developed, varying in length and difficulty. Three have been received from Dr. Zoll of Newark State College. He mentioned that one of these was submitted in 1815 by W. G. Horner, discoverer of Horner's Method for approximating the roots of a polynomial equation. (According to E. T. Bell, Horner's Method was anticipated by a Chinese.) For another proof, see R. Johnson [17, p. 78]. The shortest proof employs projective geometry [7, pp. 78, 144]. The one presented here, though not very short, is simple and easy to remember.

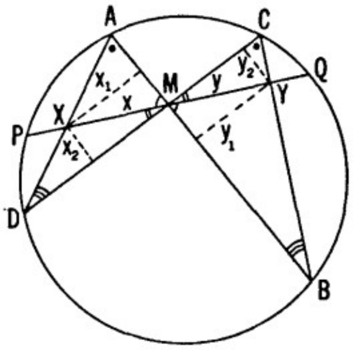

Figure 2.8A

We begin by dropping perpendiculars x_1 and y_1 from X and Y to AB, x_2 and y_2 from X and Y to CD. Writing for convenience $a = PM = MQ$, $x = XM$, $y = MY$, we observe that the pairs of similar triangles Mx_1 and My_1, Mx_2 and My_2, Ax_1 and Cy_2, Dx_2 and By_1 yield

$$\frac{x}{y} = \frac{x_1}{y_1}, \quad \frac{x}{y} = \frac{x_2}{y_2}, \quad \frac{x_1}{y_2} = \frac{AX}{CY}, \quad \frac{x_2}{y_1} = \frac{XD}{YB},$$

whence

$$\frac{x^2}{y^2} = \frac{x_1}{y_1}\frac{x_2}{y_2} = \frac{x_1}{y_2}\frac{x_2}{y_1} = \frac{AX \times XD}{CY \times YB} = \frac{PX \times XQ}{PY \times YQ}$$

$$= \frac{(a-x)(a+x)}{(a+y)(a-y)} = \frac{a^2-x^2}{a^2-y^2} = \frac{a^2}{a^2} = 1,$$

and $x = y$, as we wished to prove.

EXERCISES

1. In Figure 2.8A the lines AC and BD (extended) intersect PQ (extended) at two points which, like X and Y, are equidistant from M.

2. Let PT and PB be two tangents to a circle, AB the diameter through B, and TH the perpendicular from T to AB. Then AP bisects TH.

3. Let the incircle (with center I) of $\triangle ABC$ touch the side BC at X, and let A' be the midpoint of this side. Then the line $A'I$ (extended) bisects AX.

2.9 Morley's theorem

One of the most surprising theorems in elementary geometry was discovered about 1904 by Frank Morley (the father of Christopher Morley, whose novel, *Thunder on the Left*, has a kink in its time sequence that appeals particularly to geometers). He mentioned it to friends in Cambridge, England, and published it twenty years later in Japan. Meanwhile it was rediscovered and presented as a problem in the Educational Times. Two solutions were sent in, one of which, by M. T. Naraniengar,† is as neat as any of the dozens that have been devised since then. The theorem states:

THEOREM 2.91. *The points of intersection of the adjacent trisectors of the angles of any triangle are the vertices of an equilateral triangle.*

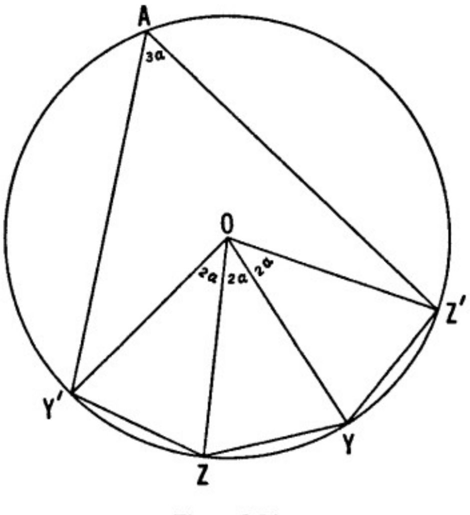

Figure 2.9A

Naraniengar's proof requires a preparatory theorem or lemma (illustrated in Figure 2.9A):

LEMMA. *If four points Y', Z, Y, Z' satisfy the conditions*

---

† Mathematical Questions and Their Solutions from the Educational Times (New Series), 15 (1909), p. 47.

$$Y'Z = ZY = YZ'$$

and

$$\angle YZY' = \angle Z'YZ = 180° - 2\alpha > 60°$$

then they lie on a circle. Moreover, if a point A, *on the side of the line* $Y'Z'$ *away from* Y, *is so situated that* $\angle Y'AZ' = 3\alpha$, *then this fifth point* A *also lies on the same circle.*

To prove the lemma, let the internal bisectors of the equal angles YZY' and $Z'YZ$ meet at O. Then $OY'Z$, OZY, OYZ' are three congruent isosceles triangles having base angles $90° - \alpha$. Their equal sides OY', OZ, OY, OZ' are radii of a circle with center O, and their angles at this common vertex are 2α. In other words, each of the equal chords $Y'Z$, ZY, YZ' subtends an angle 2α at the center O and consequently subtends an angle α at any point on the arc $Y'Z'$ not containing Y. This arc may be described as the locus of points (on the side of the line $Y'Z'$ away from Y) from which the chord $Y'Z'$ subtends an angle 3α. One such point is A; therefore A lies on the circle.

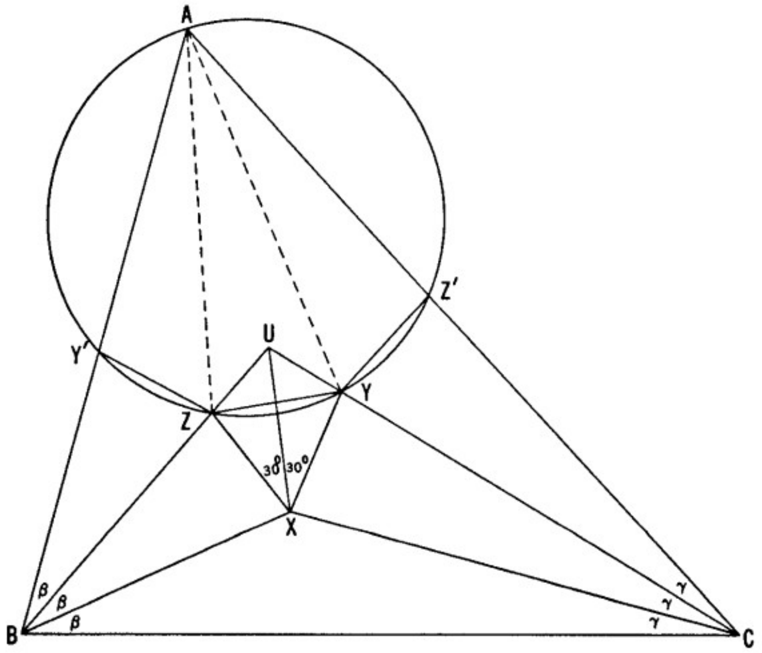

Figure 2.9B

Now we are ready to attack Theorem 2.91 itself. In Figure 2.9B, the trisectors of the angles $B = 3\beta$ and $C = 3\gamma$ meet as shown at the points U and X. In $\triangle BCU$, the angles at B and C are bisected by BX and CX; hence X is the incenter, and the angle at U is bisected by UX. If we construct points Y and Z on the lines CU and BU so that XY and XZ make equal angles 30° with XU on opposite sides, then $\Delta UXY \cong \Delta UXZ$, $XY = XZ$, and since the angle at X is 60° it follows that ΔXYZ is equilateral.

Also ΔUZY is isosceles. Its angle at U is the same as that of ΔUBC, whose other angles are 2β and 2γ; therefore the equal angles of ΔUYZ at Y and Z are each $\beta + \gamma$.

Writing $\alpha = A/3$, we deduce from $A + B + C = 180°$ that

$$\alpha + \beta + \gamma = 60°, \qquad \text{whence} \qquad \beta + \gamma = 60° - \alpha.$$

Thus

$$\angle YZU = 60° - \alpha \qquad \text{and} \qquad \angle XZU = 120° - \alpha.$$

Our next step is to mark off $BY' = BX$ on BA, and $CZ' = CX$ on CA. We now have

$$\Delta BZX \cong \Delta BZY' \qquad \text{and} \qquad \Delta CYX \cong \Delta CYZ',$$

so that

$$Y'Z = ZX = ZY = YX = YZ'.$$

Before we can apply the lemma, we still have to evaluate $\angle YZY'$ and $\angle Z'YZ$. However, this is a simple matter. Since the equal angles BZY' and BZX have equal supplements,

$$\angle UZY' = \angle XZU = 120° - \alpha$$

and

$$\begin{aligned} \angle YZY' &= \angle YZU + \angle UZY' = (60° - \alpha) + (120° - \alpha) \\ &= 180° - 2\alpha. \end{aligned}$$

Similarly, $\angle Z'YZ = 180° - 2\alpha$; and of course $\alpha = \frac{1}{3}A < 60°$.

Applying the lemma, we deduce that the five points Y', Z, Y, Z', A all lie on a circle. Since the equal chords $Y'Z$, ZY, YZ' subtend equal angles α at A, the lines AZ and AY trisect the angle A of ΔABC. In other words, the points X, Y, Z, which were artificially constructed so as to form an equilateral triangle, are in fact the points described in Morley's theorem. The proof is now complete.

EXERCISES

1. Let the angle trisectors AZ and CX (extended) meet at V, BX and AY at W. Then the three lines UX, VY, WZ are concurrent. (That is, in the language of projective geometry, UVW and XYZ are *perspective* triangles. In general, UVW is *not* equilateral.)

2. For what kind of triangle ABC will the pentagon $AY'ZYZ'$ be regular?

3. When $\triangle ABC$ is equilateral, the four points Y', Z, Y, Z' occur among the vertices of a regular enneagon (9-gon) in which A is the vertex opposite to the side ZY.

4. For a triangle with angles 3α, 3β, 3γ and circumradius R, Morley's triangle has sides $8R \sin \alpha \sin \beta \sin \gamma$.

5. If $Z'Y = YZ = ZY'$ on the side $Z'Y'$ of a rectangle $BCZ'Y'$ whose center X forms an equilateral triangle with Y and Z, then BX and BZ trisect the right angle at B.

Cálculo Integral Avançado
Bouchara

Aline Szewkies
alinetourism@gmail.com

Cemitério no Monte Sião.

+972-52-7027085

Books-001-2021\Riemann Geometry\
David G. Blair - The Detection of
Gravitational Waves (2005).pdf
Pág. 6(31)
fractionally larger than π.

John Wheeler introduced the
slogan "matter tells space
how to curve, space tells
matter how to move" to
describe general relativity.

This hides the complexity
because in reality:

matter, the motion of matter,
and radiation density,
including propagating
waves of curvature,
tell spacetime how to
curve.

Curvature creates curvature,
and influences its propagation.

—

Einstein realised the
connection between
gravitational dynamics
and Riemann geometry.

—

201.181.244.35.bc.googleusercontent.com

—

20210102

Books-001-2021\Math-
Distances\
Computer-Aided Geometric
Design...pdf

-./0- - -1 ! - @!; " 5 8 ! > @
5 !; ! ! H!; , 4"
& '

1.3 Non-linearity and wave phenomena

Non-linearity means that in general wavefronts can interact with each other and with regions of static curvature. This can allow a variety of interesting phenomena, some of which we discuss below. There are probably plenty more such phenomena which will be among the surprises that await the birth of gravitational wave astronomy.

(a) Waves can be scattered by strong background curvature. For example, the gravitational waves emanating from the coalescence of a binary neutron star system, and its subsequent collapse to form a black hole, will be scattered in the intense curvature of the forming black hole (Thorne, 1987, and references therein).

(b) Waves can be focussed by background curvature. On a cosmological scale galaxies can focus gravitational waves from distant sources as they are already observed to do optically with quasars. On a more local scale, the Sun can act as a lens to focus plane gravitational waves to a focal point near to the orbit of Jupiter (about 20 AU from the Sun). There is a catch here, however, and it is probably not worth sending a gravitational wave detector out there. The problem is that the solar lens is not a lot larger than a gravitational wavelength. Thorne (1983) shows that strong focussing only occurs if the gravitational diameter of the Sun, $4GM/c^2$, is large compared to the wavelength; otherwise diffraction spreads the beam sufficiently that the focussed amplitude is not much increased. Thus only waves of frequency $> 10^4$ Hz will be strongly focussed, and, as we shall see later, these are not of great interest.

(c) Waves can be parametrically amplified in regions of time varying curvature. If the curvature is varying dynamically, for example in the early universe after the big bang, gravitational waves can be amplified as they interact with the background. This can increase the amplitude of gravitational waves emanating from fluctuations in the early universe. Similarly, the background curvature may be modulated by one source of waves (which an engineer might call the pump). A second incident wave (the signal) may be amplified and frequency shifted by this interaction exactly as a microwave or optical signal is amplified by a pumped non-linear medium. Thus a coalescing binary with orbital frequency ω_b will upconvert incident radiation to frequency $\omega_b + \omega_i$ while simultaneously generating radiation at twice the binary frequency. Similarly, one can imagine (probably very unlikely) processes whereby a massive black hole focusses intense gravitational waves causing non-linear harmonic generation.

(d) Waves interact with self generated curvature. This is epitomised by the geon, proposed by Wheeler (1962). A geon is a bundle of gravitational waves held together by its own self generated curvature. While most doubt that such objects could exist, one can conjecture about beams or pulses of self focussing gravitational waves, confining themselves in their own self generated valley of background curvature.

(e) Colliding gravitational waves can focus each other so strongly that the wavefront may collapse to form a singularity. If the singularity forms only around one wavefront, a high velocity black hole will fly out from the collision, but if it forms about both wavefronts the resulting black hole will be nearly stationary (Yurtsever, 1988a–c).

These are some of the exotic aspects of the non-linearity. At a more immediately practical level, the problem of the non-linearity of general relativity affects above all the modelling of sources of gravitational waves. Thus the majority of source calculations can only be approximations. Even with the use of powerful numerical codes, the behaviour of matter under conditions of intense gravity and high densities is so uncertain that our confidence in their results cannot be high. (See, for example, Nakamura, 1987.) In spite of this, however, we will see in chapter 2 that good order of magnitude estimates exist for a range of wave generation processes.

1.4 Introduction to gravitational waves

In the case of the analysis for detection of gravitational waves it is a different matter. For even the strongest imaginable gravitational waves from astrophysical sources we need only consider plane, linear waves on an essentially flat space background. This is because the amplitude of waves crossing the solar system, even from relatively nearby strong sources, must always be exceedingly low, despite the fact that the energy flux may be enormous. We will come back to this point in chapter 2. As a result, in spite of non-linearity, we go ahead and make a separation, which is not fully precise. We assume superposition, which is essential to our concept of gravitational waves, which is not strictly rigorous, which is bothersome to some theorists, but which need bother experimentalists not at all. We separate spacetime curvature into a background term and a wave term by assuming that the background curvature varies slowly while the wave varies rapidly. That is

$$R_{\alpha\beta\mu\nu} = R^{\mathrm{B}}_{\alpha\beta\mu\nu} + R^{\mathrm{W}}_{\alpha\beta\mu\nu}. \tag{1.3}$$

It is possible to go on to further simplify things by assuming plane waves on a background of flat Minkowski spacetime, or alternatively to simplify Einstein's equations for weak fields and explicitly solve for a wave solution. We shall only sketch the barest details of the derivation here, as it involves considerable mathematical detail readily found elsewhere, and anyhow is outside the scope of this book.

! -./0- - -1 ! - 4"
? -./0- - -1 ! - ; ! " ! 4
. " " 5 @ 4 ! 8 7 " ? $<< 4"
1? E ; " JJJ

1. Basic Theory of Curves and Surfaces

1.1 General

1.1.1 Properties of Object Shapes and Their Mathematical Representation

Before a computer can perform processing relating to a shape, a mathematical description of that shape must be provided on the computer's memory. Such a description should preserve as many of the properties of the actual object shape as possible. From the point of view of computer processing the following properties are particularly important.

(1) Spatial Uniqueness

Two objects cannot occupy the same space at the same time. This property is called *spatial uniqueness*. Object shapes are expressed in a computer by data structures using numerical shape data and functions, but spatial uniqueness is very difficult to express. Therefore a program is needed to test for overlap between positions of two bodies.

(2) Boundedness and Continuity

An object shape occupies a bounded space defined by a number of mutually adjacent surfaces of finite extent. In this book, this property is called the *boundedness* of the object shape. In performing computer-aided design it is extremely important for the range over which the shape being designed extends to be conveniently expressed. Let us consider the case of a curve. The curve in Fig. 1.1 is multivalued: it takes on several y-values for some values of x, and also several x-values for some values of y. If this curve is described by

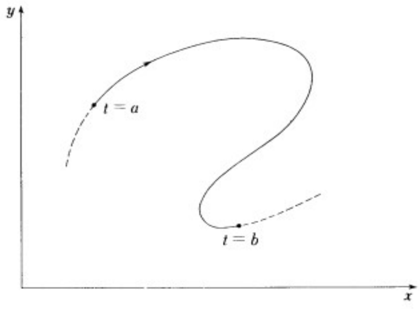

Fig. 1.1. A curve segment for which y is multi-valued with respect to x and x is multi-valued with respect to y

a function $y=F(x)$, trying to describe its range can be very difficult. On the other hand, if it is described by a parametric vector function $\boldsymbol{P}(t)=[x(t)\ y(t)]$, then the range can be described simply as $a \leqq t \leqq b$ where $t=a$ and $t=b$ at the two ends of the curve.

Curved (or plane) object surfaces are in continuous connection, without a break, along curved boundary lines or ridge lines. This property of an object is called *continuity*. Many types of processing on shapes make effective use of this continuity property. Consequently, in expressing an object shape on a computer, it is important to express how the surfaces connect to each other along boundary curves. This description is performed by a *data structure*.

(3) Independence of Coordinate Axes

An object shape is independent of the coordinate system in which it is described. This seems natural enough, but it is an important condition in formulating a mathematical model of the object. As an example, consider the problem of selecting points on a clay model, measuring their coordinates and then fitting them to a mathematical representation of a curve. This is normally done by letting the measured points define the curve and *interpolating* between them. The final curve must be the same regardless of the coordinate system in which the coordinates of the points were measured. This property is called *shape invariance under transformation*. If the method of arriving at a mathematical representation is inappropriate, the condition of shape invariance under transformation will not be satisfied.

For example, consider the following parametric curve.

$$\boldsymbol{P}(t)=A_0(t)\,\boldsymbol{P}_0+A_1(t)\,\boldsymbol{P}_1 \quad (0 \leqq t \leqq 1) \tag{1.1}$$

where $A_0(t)=(1-t)^3$ and $A_1(t)=t^3$. In this case, the relations $A_0(0)=1$, $A_0(1)=0$, $A_1(0)=0$, $A_1(1)=1$ hold, so we see that $\boldsymbol{P}(t)$ is a curve which passes through the two points $\boldsymbol{P}_0$ and $\boldsymbol{P}_1$. Taking $\boldsymbol{P}_0=[0\ 0]$ and $\boldsymbol{P}_1=[10\ 10]$, $\boldsymbol{P}(t)$ describes the line segment connecting $\boldsymbol{P}_0$ and $\boldsymbol{P}_1$ (refer to Fig. 1.2). Expressing $\boldsymbol{P}_0$ and $\boldsymbol{P}_1$ in the x^*y^* coordinate system, $\boldsymbol{P}_0^*=[-3\ -2]$ and $\boldsymbol{P}_1^*=[7\ 8]$; then if we let

$$\boldsymbol{P}^*(t)=A_0(t)\,\boldsymbol{P}_0^*+A_1(t)\,\boldsymbol{P}_1^*$$

$\boldsymbol{P}^*(t)$ passes through the same points as $\boldsymbol{P}(t)$ at $t=0$ and $t=1$, but for $0<t<1$ $\boldsymbol{P}^*(t)$ describes a curve different from the straight line of $\boldsymbol{P}(t)$. That is, the straight line representation of Eq. (1.1) is not desirable.

In Eq. (1.1) a curve is described as a vector using the parameter t. There is an alternate way to describe curves, the non-parametric method in the form $y=F(x)$. In the latter representation, at a point at which the curve becomes parallel to the y-axis the slope dy/dx becomes infinite, making it difficult to handle on a computer. In practice it becomes necessary to use another suitable coordinate system in which the slope does not become infinite. However, depending on the shape of the curve being represented, there are

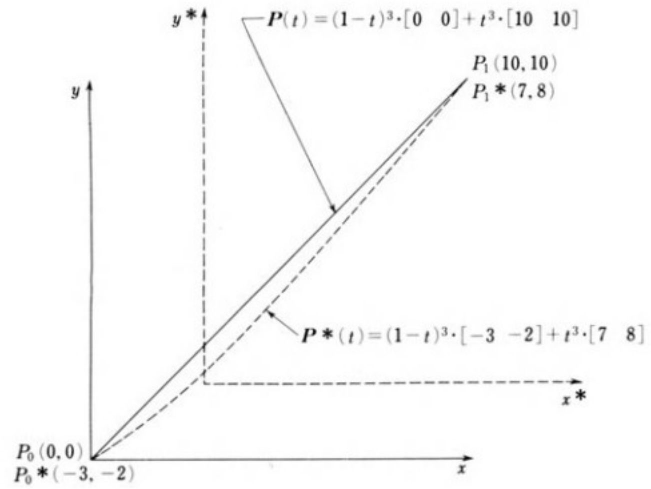
y^*
$P(t)=(1-t)^3\cdot[0\ \ 0]+t^3\cdot[10\ \ 10]$
$P_1\,(10,10)$
y
$P_1{*}\,(7,8)$
$P{*}(t)=(1-t)^3\cdot[-3\ \ -2]+t^3\cdot[7\ \ 8]$
x^*
$P_0\,(0,0)$
$P_0{*}\,(-3,-2)$
x

Fig. 1.2. An example in which a shape does not remain invariant under coordinate transformation

some cases in which the slope will become infinite no matter what coordinate system is chosen. This happens when the curve is closed. On the other hand, with a parametric representation the slope is:

$$\frac{dy}{dx}=\frac{dy}{dt}\bigg/\frac{dx}{dt}=\frac{\dot{y}}{\dot{x}}$$

so the direction parallel to the y-axis can be simply represented by:

$$\dot{x}=0.$$

The shape of an object is independent, unaffected by the coordinate system in which it is described. This property of an object is called *axis independence*. Axis independence must exist in order for a shape to be expressed on a computer.

In the above we have studied a relation between two types of mathematical representation of a curve and the properties of a shape. We have seen that an expression for a curve in the form $y=F(x)$ is inadequate to represent the properties possessed by shapes. A parametric representation overcomes this difficulty.

The advantages which a parametric representation of a curve has over a non-parametric representation can be summarized as follows.

1) Axis independence can be expressed (refer to Sect. 1.1.3).
2) It is easier to specify the regions of a multivalued curve.
3) Since it is an independent expression with respect to x, y and z (x and y in the case of a plane curve), affine transformations and projective

transformations can easily be performed by 4 × 4 matrices (3 × 3 matrices for plane curves).

4) Representation of space curves is easy.

In order to express a space curve by a non-parametric representation, it must be defined as the intersection of two curved surfaces: $y = F(x)$ and $z = G(x)$, or $\varphi_1(x, y, z) = 0$ and $\varphi_2(x, y, z) = 0$. In a parametric representation it can be represented in the form $\boldsymbol{P}(t) = [x(t)\ y(t)\ z(t)]$.

5) Since x, y and z are expressed as explicit functions of the parameter t, a point on the curve can be easily computed.

Since curved surfaces are basically regarded as being represented by groups of curves, the above arguments with respect to curves can be thought of as also applying to curved surfaces.

1.1.2 Design and Mathematical Representations

A mathematical representation of a curve or curved surface must not only express the properties possessed by a shape, they must also satisfy certain necessary conditions for design.

(1) Abundant Expression Power Having Generality

The designer's demands with regard to shapes are not limited; they change frequently, and a mathematical representation of a shape must be able to respond to them flexibly. For a mathematical representation of a curve, in some cases a straight line is appropriate, while in other cases a conic section such as a circle or ellipse might be appropriate. In addition, in some cases it is desirable not to be limited to fixed shapes, but rather for the representation to have the capability to respond flexibly to any arbitrary shape that the designer draws freely with a stylus.

Ordinarily, a curve or curved surface is thought of as consisting of a number of curve segments or surface patches which are connected together. When designing a smooth curve or curved surface, it is necessary to make sure that the curve segments or surface patches are connected smoothly. Mathematically, continuity to a sufficiently high order derivative (usually up to the slope or the curvature) is required. The smoothness of a curve involves not only this kind of local continuity but also the overall large-scale smoothness of the curve. Figure 1.3 shows examples of curves which are connected in such a manner that there is (a) continuity of position only (class C^0), (b) continuous to the slope (class C^1), (c) continuous to the curvature (class C^2). Figure 1.4 shows a corner R of a curve. In (a), the connection is done with straight line segments and a circular arc, so the slope is continuous at the connection point, but the curvature is discontinuous (class C^1). When the discontinuity of curvature reaches a certain level, it becomes noticeable to the naked eye. (b) shows a corner R with the connection performed with a curve in such a manner that the curvature is continuous (class C^2).

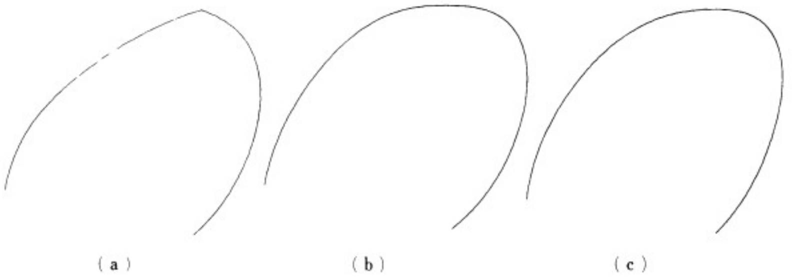
(a) (b) (c)

Fig. 1.3. (a) A class C^0 curve connection, (b) a class C^1 curve connection, (c) a class C^2 curve connection

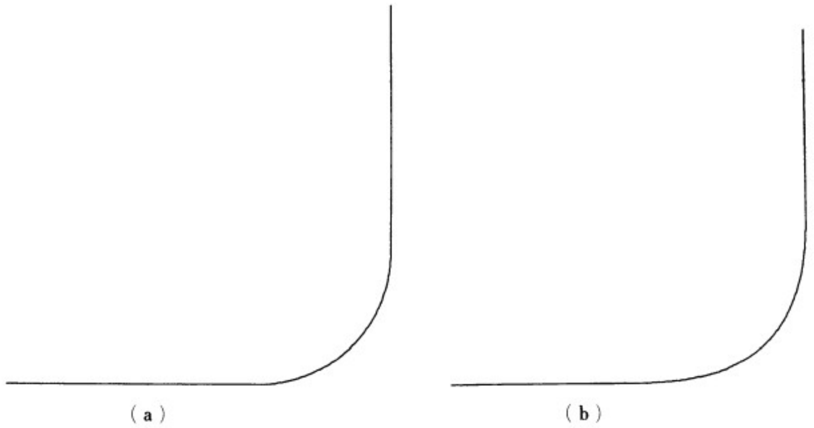
(a) (b)

Fig. 1.4. Expression of a corner R (a) Class C^1, (b) class C^2

In designing a curve or curved surface, it sometimes happens that it is desired to have the slope be discontinuous at a certain point. In fact, the discontinuity of slope is sometimes necessary. A point where the slope is discontinuous is called a *cusp*. Figure 1.3(a) shows a cusp. A line of discontinuity in a curved surface is called a *crease* (refer to Fig. 1.5).

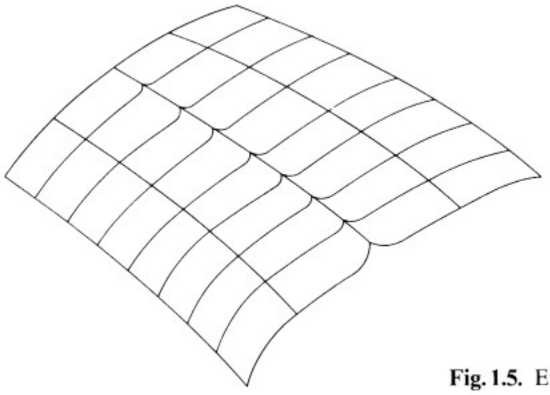
Fig. 1.5. Example of a crease in a curved surface

(2) Ease of Shape Control

Shapes of curves and curved surfaces which appear on the display could ideally be manipulated at will by the designer. At the start of the design work, the designer focuses on the overall shape that is displayed and forms his judgments. Consequently, it is desirable for the designer to be able to exercise *global* shape control (Fig. 1.6). Along the curve there are a number of *control points* which control the shape. Usually the control points are the same points that were used to define the curve. It would be very troublesome if it were necessary to change the parameters of all of the control points in order to change the shape of the curve. Ideally one would be able to change the shape of the entire curve by controlling just one suitable control point. As the design work progresses, the designer's attention shifts from the overall curve to localized sections of it. Then the designer wants to be able to change the shapes of only those sections while leaving the rest of the curve unchanged. Consequently, it is desirable to have the capability for not only global control of the shape of a curve, but also *local* control (Fig. 1.7). A local control capability is a necessary condition for being able to converge the trial-and-error design process.

There are cases in which it is impossible to control the locations of control points in such a way as to obtain the desired shape with only one

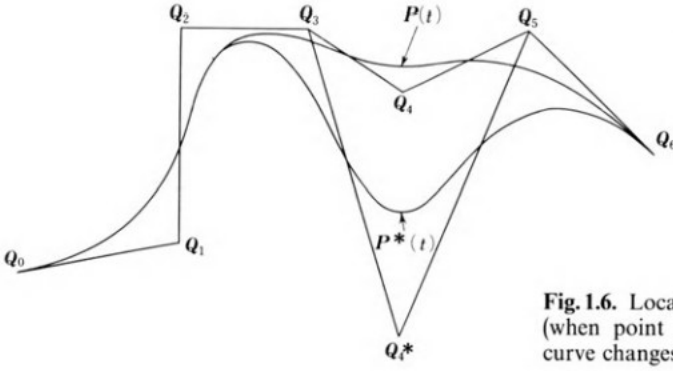

Fig. 1.6. Local control of a curve (when point Q_4 changes to Q_4^*, the curve changes from $P(t)$ to $P^*(t)$)

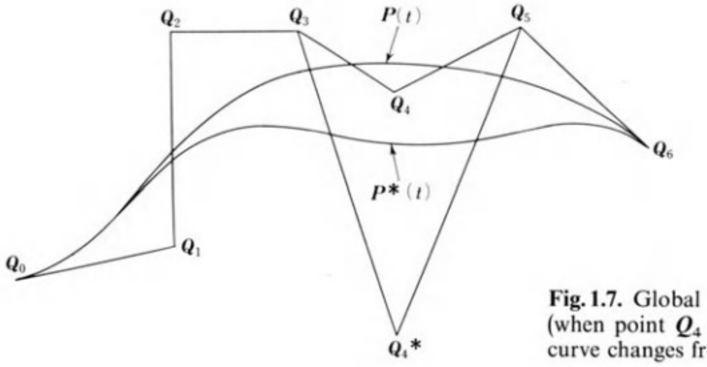

Fig. 1.7. Global control of a curve (when point Q_4 is changed to Q_4^* the curve changes from $P(t)$ to $P^*(t)$)

mathematical expression. In such cases, it is necessary to have some method to improve the capability of the mathematical expression to express a shape. Figures 1.8 and 1.9 show examples of ways in which the ability to express shapes can be increased. In Fig. 1.8 the shape is preserved while one curve segment is split into two segments. In Fig. 1.9 the shape is preserved while the degree of the curve segments is increased by one. What is important here is that the capability to express shapes is increased without changing the shapes themselves.

Some curves have special properties such as the *convex hull property* or the *variation diminishing property*. In such cases shape control is made easier because the changes that a shape will undergo when the control point parameters are varied can be predicted (refer to Chaps. 5 and 6).

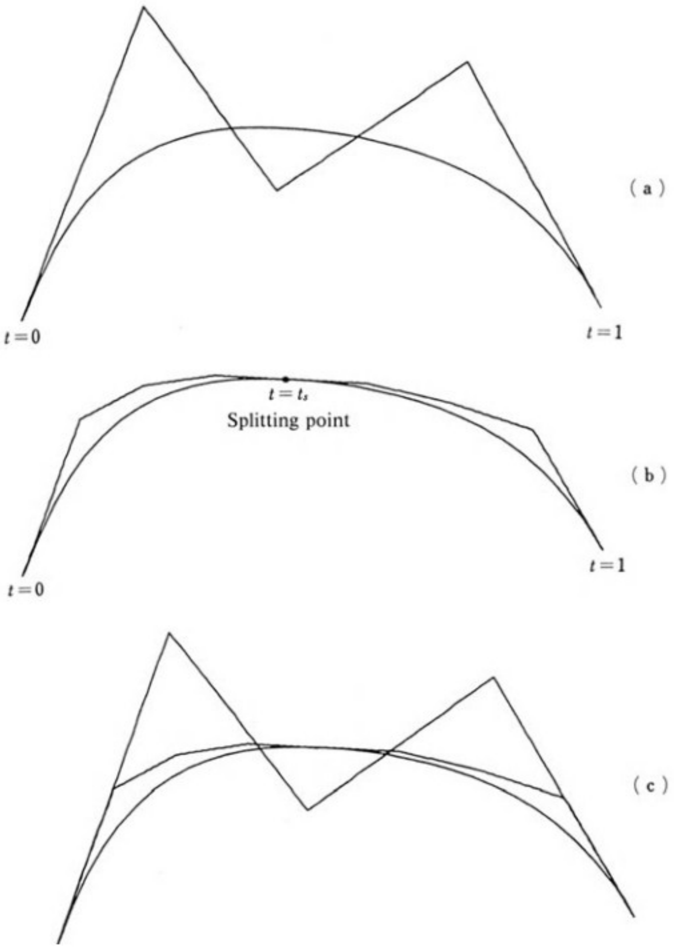

Fig. 1.8. An example of how a curve can be split into two segments while preserving its shape (a Bézier curve) (a) Before splitting, (b) after splitting, (c) after splitting

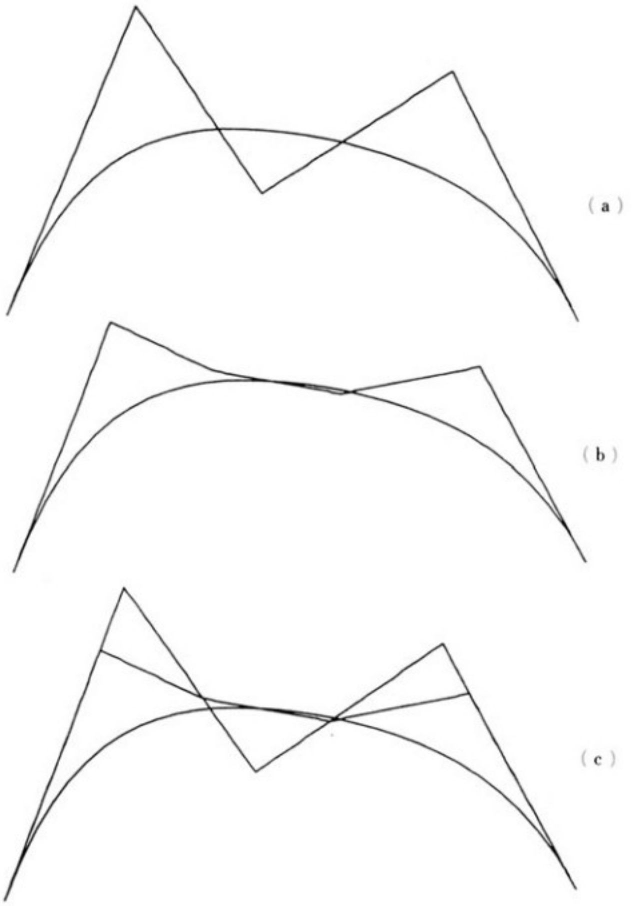

Fig. 1.9. An example in which the shape of the curve is preserved and (in form) the degree of the curve is increased (Bézier curve) (a) Before degree is increased, (b) after degree is increased; (c) before and after degree is increased

1.1.3 Invariance of a Shape Under Coordinate Transformation

In Sect. 1.1.1 we referred to the invariance of a shape under coordinate transformation. Now let us look at the conditions for such invariance to be satisfied.

Let a curve be represented by

$$\begin{aligned} \boldsymbol{P}(t) &= [x(t)\ y(t)\ z(t)] \\ &= [\phi_0(t)\ \phi_1(t) \ldots \phi_n(t)]\ [\boldsymbol{Q}_0\ \boldsymbol{Q}_1 \ldots \boldsymbol{Q}_n]^T. \end{aligned} \tag{1.2}$$

$\phi_0(t)$, $\phi_1(t)$, …, $\phi_n(t)$ are linearly independent polynomial functions of t; $\boldsymbol{Q}_0$, $\boldsymbol{Q}_1$, …, $\boldsymbol{Q}_n$ are the position vectors that define the curve.

Let us separately consider the 3-dimensional rotation matrix $\boldsymbol{T}$ (a 3×3 matrix) and the translation $\boldsymbol{m}=[m_x\ m_y\ m_z]$. In order for the shape to be invariant, it is necessary that the curve expressed in terms of $\boldsymbol{Q}_0^*, \boldsymbol{Q}_1^*, \ldots, \boldsymbol{Q}_n^*$ in the transformed coordinate system transform by the same amount as $\boldsymbol{Q}_0, \boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_n$ do.

For the rotation matrix $\boldsymbol{T}$ we have:

$$\begin{aligned}\boldsymbol{P}^*(t)&=[\phi_0(t)\ \ \phi_1(t)\ \ldots\ \phi_n(t)]\begin{bmatrix}\boldsymbol{Q}_0^*\\ \boldsymbol{Q}_1^*\\ \vdots\\ \boldsymbol{Q}_n^*\end{bmatrix}=[\phi_0(t)\ \ \phi_1(t)\ \ldots\ \phi_n(t)]\begin{bmatrix}\boldsymbol{Q}_0\,\boldsymbol{T}\\ \boldsymbol{Q}_1\,\boldsymbol{T}\\ \vdots\\ \boldsymbol{Q}_n\,\boldsymbol{T}\end{bmatrix}\\ &=[\phi_0(t)\ \ \phi_1(t)\ \ldots\ \phi_n(t)]\begin{bmatrix}\boldsymbol{Q}_0\\ \boldsymbol{Q}_1\\ \vdots\\ \boldsymbol{Q}_n\end{bmatrix}\boldsymbol{T}=\boldsymbol{P}(t)\,\boldsymbol{T}\end{aligned}$$

so that in the case of rotation the condition of invariance is automatically satisfied.

Next let us consider the translation.

$$\begin{aligned}\boldsymbol{P}^*(t)&=[\phi_0(t)\ \ \phi_1(t)\ \ldots\ \phi_n(t)]\begin{bmatrix}\boldsymbol{Q}_0^*\\ \boldsymbol{Q}_1^*\\ \vdots\\ \boldsymbol{Q}_n^*\end{bmatrix}\\ &=[\phi_0(t)\ \ \phi_1(t)\ \ldots\ \phi_n(t)]\begin{bmatrix}\boldsymbol{Q}_0+\boldsymbol{m}\\ \boldsymbol{Q}_1+\boldsymbol{m}\\ \vdots\\ \boldsymbol{Q}_n+\boldsymbol{m}\end{bmatrix}\\ &=[\phi_0(t)\ \ \phi_1(t)\ \ldots\ \phi_n(t)]\begin{bmatrix}\boldsymbol{Q}_0\\ \boldsymbol{Q}_1\\ \vdots\\ \boldsymbol{Q}_n\end{bmatrix}+(\phi_0(t)+\phi_1(t)+\ldots+\phi_n(t))\,\boldsymbol{m}\\ &=\boldsymbol{P}(t)+(\phi_0(t)+\phi_1(t)+\ldots+\phi_n(t))\,\boldsymbol{m}.\end{aligned}\tag{1.3}$$

From Eq. (1.3), in order for the shape to be invariant the relation

$$\phi_0(t)+\phi_1(t)+\ldots+\phi_n(t)\equiv 1\tag{1.4}$$

must be satisfied. Equation (1.4) is sometimes called Cauchy's relation.

1.2 Curve Theory

1.2.1 Parametric Representation of Curves; Tangent Lines and Osculating Planes

Let us express a curve as a vector using the parameter t:

$$\boldsymbol{P}(t) = [x(t) \;\; y(t) \;\; z(t)] .$$

Next consider the derivative vector evaluated at $t = t_0$:[*]

$$\dot{\boldsymbol{P}}(t_0) = [\dot{x}(t_0) \;\; \dot{y}(t_0) \;\; \dot{z}(t_0)] .$$

If the condition

$$\dot{\boldsymbol{P}}(t_0) \neq \boldsymbol{0} \tag{1.5}$$

is satisfied, then the curve is said to be *regular* at $t = t_0$. A point at which a curve is regular is called an *ordinary point*, a point at which it is not regular is called a *singular point*. For condition (1.5) for regularity to be satisfied, $\dot{x}(t_0)$, $\dot{y}(t_0)$ and $\dot{z}(t_0)$ must not all be zero at once. This condition is expressed mathematically by the following inequality:

$$\dot{x}(t_0)^2 + \dot{y}(t_0)^2 + \dot{z}(t_0)^2 > 0 . \tag{1.6}$$

If the derivatives of $x(t)$, $y(t)$ and $z(t)$ of a curve $\boldsymbol{P}(t)$ $(a \leqq t \leqq b)$ to order r exist and are continuous, and if the curve is regular throughout the interval, then that curve is said to be of class C^r.

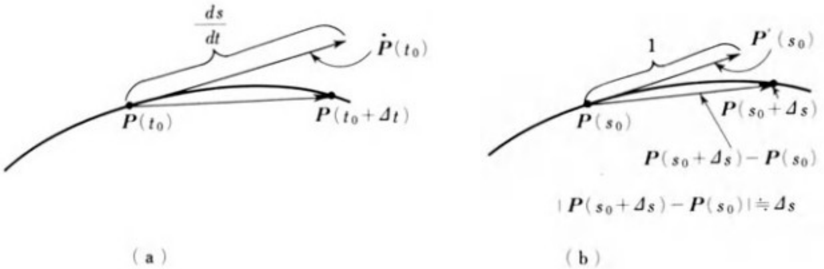
$\frac{ds}{dt}$
$\dot{\boldsymbol{P}}(t_0)$
$\boldsymbol{P}(t_0)$
$\boldsymbol{P}(t_0 + \Delta t)$
1
$\boldsymbol{P}'(s_0)$
$\boldsymbol{P}(s_0)$
$\boldsymbol{P}(s_0 + \Delta s)$
$\boldsymbol{P}(s_0 + \Delta s) - \boldsymbol{P}(s_0)$
$|\boldsymbol{P}(s_0 + \Delta s) - \boldsymbol{P}(s_0)| \fallingdotseq \Delta s$
(a)
(b)

Fig. 1.10. Diagrams explaining the tangent vector $\dot{\boldsymbol{P}}(t_0)$ and the unit tangent vector $\boldsymbol{P}'(s_0)$ (a) The tangent vector, (b) the unit tangent vector

---

[*] If t is thought of as time, then the first derivative with respect to t is the velocity vector and the second derivative with respect to t is the acceleration vector

In Fig. 1.10(a), consider the vector $\boldsymbol{P}(t_0+\Delta t)-\boldsymbol{P}(t_0)$, from $\boldsymbol{P}(t_0)$ to $\boldsymbol{P}(t_0+\Delta t)$ on the curve $\boldsymbol{P}(t)$, divided by Δt:

$$\frac{\boldsymbol{P}(t_0+\Delta t)-\boldsymbol{P}(t_0)}{\Delta t}.$$

This vector indicates the direction from $\boldsymbol{P}(t_0)$ to $\boldsymbol{P}(t_0+\Delta t)$. If the curve is regular at $t=t_0$, then this vector converges to the finite magnitude vector $\dot{\boldsymbol{P}}(t_0)$ in the limit as $\Delta t \to 0$. $\dot{\boldsymbol{P}}(t_0)$ is called the tangent vector at the point $t=t_0$.

Let us now use the length s of the curve as a parameter. We will indicate the derivative with respect to s by the prime symbol $'$ to distinguish it from the derivative with respect to t.

Since

$$\frac{d\boldsymbol{P}}{ds} \equiv \boldsymbol{P}' = \frac{d\boldsymbol{P}}{dt}\,\frac{dt}{ds} = \frac{\dot{\boldsymbol{P}}}{\dot{s}} \tag{1.7}$$

$$\frac{ds}{dt} \equiv \dot{s} = \sqrt{\left(\frac{dx}{dt}\right)^2+\left(\frac{dy}{dt}\right)^2+\left(\frac{dz}{dt}\right)^2} = \sqrt{\dot{\boldsymbol{P}}^2} \tag{1.8}$$

then if $\dot{\boldsymbol{P}} \neq \boldsymbol{0}$, we have:

$$\boldsymbol{P}' = \frac{\dot{\boldsymbol{P}}}{\dot{s}} = \frac{\dot{\boldsymbol{P}}}{\sqrt{\dot{\boldsymbol{P}}^2}} \tag{1.9}$$

$$\therefore \ |\boldsymbol{P}'| = 1. \tag{1.10}$$

That is, $\boldsymbol{P}'$ is in the same direction as $\dot{\boldsymbol{P}}$, and its magnitude is 1. $\boldsymbol{P}'$ is called the *unit tangent vector* (refer to Fig. 1.10(b)). It can be seen from Eq. (1.8) that $\dot{s}$ is the magnitude of the tangent vector. In this book, the magnitude of the tangent vector will sometimes be denoted by α, and the unit tangent vector by $\boldsymbol{t}$. Hence we have:

$$\dot{s} \equiv \alpha \tag{1.11}$$

$$\boldsymbol{P}' \equiv \boldsymbol{t}. \tag{1.12}$$

The equation of a tangent line can be expressed in terms of a parameter u as:

$$\boldsymbol{R}(u) = \boldsymbol{P}(t_0) + u\boldsymbol{t}(t_0). \tag{1.13}$$

Here the parameter u is the distance from the tangent point.

At a point where a curve is regular, the tangent line is unique. However, at a singular point there occur various anomalous cases. Examples of singular points are shown in Figs. 1.11 and 1.12. At a cusp, as shown in Fig. 1.11, $\dot{\boldsymbol{P}}$

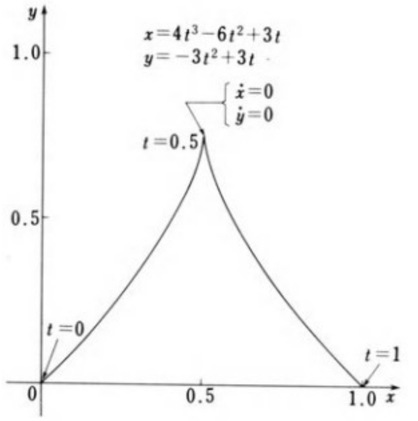

Fig. 1.11. Example of a singular point (1)

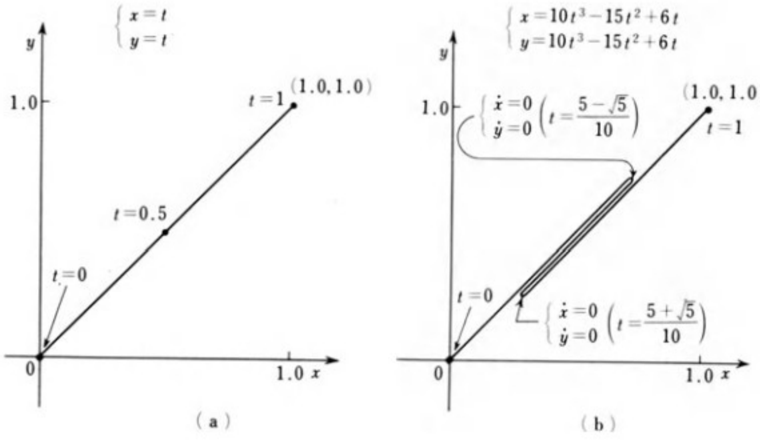

Fig. 1.12. Example of a singular point (2)

can become discontinuous or a **0** vector. Figure 1.12 shows examples in which line segments are expressed parametrically. With the expression used in Fig. 1.12(a), the line segments are regular over the entire interval. In contrast, in Fig. 1.12(b), although the line being expressed is the same two singular points occur within the interval. In this case, as t increases from the starting point $t=0$ to $t=(5-\sqrt{5})/10$, the point on the line progresses steadily toward the end point. At $t=(5-\sqrt{5})/10$ $\dot{x}=\dot{y}=0$, so this point is a singular point. From $t=(5-\sqrt{5})/10$ to $t=(5+\sqrt{5})/10$, as t increases the point on the line progresses back toward the starting point. $t=(5+\sqrt{5})/10$ is another singular point; there the direction reverses again and the point on the line progresses toward the end point $t=1$. The possibilities for such *parameterization,* the assigning of a relation between points on a line and a parameter, are numerous. It is desirable to choose a parameterization such that the curve is normal at as many points as possible.

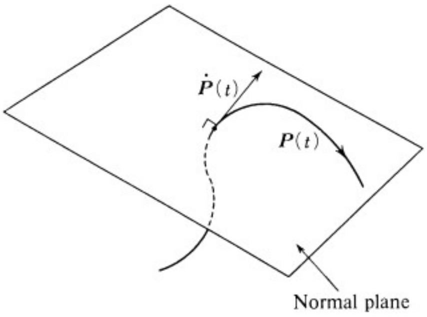

Fig. 1.13. The normal plane

Another type of singular point is one at which the curve *degenerates* to a single point.

A plane which passes through point $\boldsymbol{P}(t_0)$ on a curve and is perpendicular to the tangent line to the curve at that point is called *normal plane* (Fig. 1.13). The equation of a normal plane is:

$$(\boldsymbol{R}-\boldsymbol{P}(t_0))\cdot\dot{\boldsymbol{P}}(t_0)=0. \tag{1.14}$$

Consider a point $\boldsymbol{P}(t_0)$ on a curve and a plane which passes through it and two points very close to it, $\boldsymbol{P}(t_0+\Delta_1 t)$ and $\boldsymbol{P}(t_0+\Delta_2 t)$. The plane that is approached in the limit as $\Delta_1 t$ and $\Delta_2 t$ approach 0 independently is called the *osculating plane* at point $\boldsymbol{P}(t_0)$. The osculating plane is given by the following equation.

$$[\boldsymbol{R}-\boldsymbol{P}(t_0), \dot{\boldsymbol{P}}(t_0), \ddot{\boldsymbol{P}}(t_0)]=0. \tag{1.15}$$

The brackets indicate a triple scalar product; $\dot{\boldsymbol{P}}(t_0)\times\ddot{\boldsymbol{P}}(t_0)\neq 0$. Equation (1.15) is the condition for the three vectors $\boldsymbol{R}-\boldsymbol{P}(t_0)$, $\dot{\boldsymbol{P}}(t_0)$ and $\ddot{\boldsymbol{P}}(t_0)$ to lie in the same plane.

A general plane passing through the point $\boldsymbol{P}(t_0)$ can be written in terms of its unit normal vector $\boldsymbol{a}$ as $\boldsymbol{a}\cdot(\boldsymbol{R}-\boldsymbol{P}(t_0))=0$. The length of the perpendicular distance from the point $\boldsymbol{P}(t_0+\Delta t)$ on the curve $\boldsymbol{P}(t)$ to this plane is:

$$\boldsymbol{a}\cdot(\boldsymbol{P}(t_0+\Delta t)-\boldsymbol{P}(t_0))=\boldsymbol{a}\cdot\left(\dot{\boldsymbol{P}}(t_0)\,\Delta t+\frac{\ddot{\boldsymbol{P}}(t_0)}{2!}\,\Delta t^2+\dots\right)$$

(Fig. 1.14). In the special case when $\boldsymbol{a}\cdot\dot{\boldsymbol{P}}(t_0)=0$, $\boldsymbol{a}\cdot\ddot{\boldsymbol{P}}(t_0)=0$, that is, when the plane is the osculating plane, this length is a 3rd-order infinitesimal with respect to Δt. Consequently, the osculating plane is the plane that best fits the curve at the point $\boldsymbol{P}(t_0)$.

The line that lies in the osculating plane, passes through the point $\boldsymbol{P}(t_0)$ and is perpendicular to the tangent vector $\dot{\boldsymbol{P}}(t_0)$ at that point is called the *principal normal*. The line that passes through point $\boldsymbol{P}(t_0)$ and is perpendicular to the osculating plane is called the *binormal*, and the plane determined by the

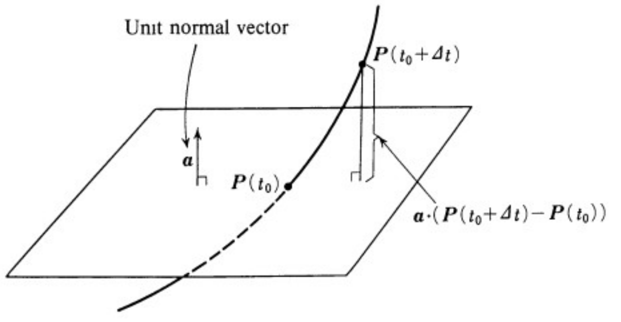

Fig. 1.14. Perpendicular distance from a point on a curve to a plane

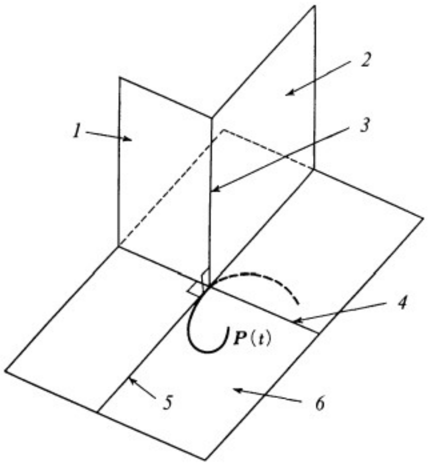

Fig. 1.15. Relation between the normal plane, osculating plane and rectifying plane. *1* normal plane; *2* rectifying plane; *3* binormal; *4* principal normal; *5* tangent line, *6* osculating plane

tangent and the binormal is called the *rectifying plane*. These relations are shown in Fig. 1.15.

1.2.2 Curvature and Torsion

Let us consider the meaning of the second derivative $\boldsymbol{P}''(s) \equiv d^2\boldsymbol{P}/ds^2$. From the definition of a derivative, we have:

$$\boldsymbol{P}''(s_0) = \lim_{\Delta s \to 0} \frac{\boldsymbol{P}'(s_0+\Delta s) - \boldsymbol{P}'(s_0)}{\Delta s}. \tag{1.16}$$

As shown in Fig. 1.16(a) and (b), in the limit as $\Delta s \to 0$ the numerator $\boldsymbol{P}'(s_0+\Delta s) - \boldsymbol{P}'(s_0)$ is perpendicular to the tangent vector at the point $\boldsymbol{P}(s_0)$ and points toward the center of curvature of the curve. It can be seen from

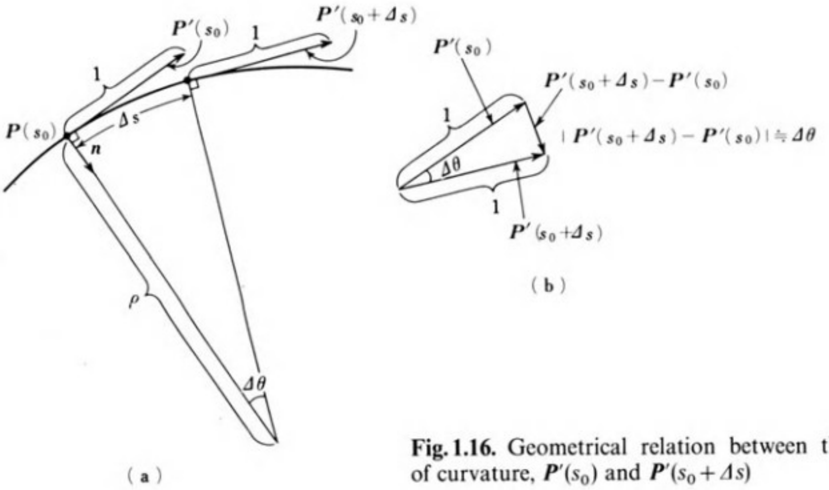

Fig. 1.16. Geometrical relation between the center of curvature, $\boldsymbol{P}'(s_0)$ and $\boldsymbol{P}'(s_0+\Delta s)$

Fig. 1.16(b) that its magnitude is $\Delta\theta$. Consequently, the magnitude of $\boldsymbol{P}''(s_0)$ is:

$$|\boldsymbol{P}''(s_0)| = \lim_{\Delta s\to 0}\frac{\Delta\theta}{\Delta s} = \lim_{\Delta s\to 0}\frac{\frac{1}{\varrho}\Delta s}{\Delta s} = \frac{1}{\varrho} \equiv \kappa .$$

Here ϱ is the radius of curvature and κ is the curvature. $\boldsymbol{P}''$ can be expressed in terms of them as:

$$\boldsymbol{P}'' = \frac{1}{\varrho}\,\boldsymbol{n} \equiv \kappa\boldsymbol{n} . \tag{1.17}$$

Here $\boldsymbol{n}$ is the unit vector pointing toward the center of curvature. Since $\boldsymbol{P}''(s)$ is a vector that has a magnitude equal to the curvature at point s and points toward the center of curvature, it is sometimes called the curvature vector. Curvature is the rate of turning of the unit tangent vector $\boldsymbol{t}$ with respect to the length of the curve s, in other words, a quantity that indicates how rapidly or slowly the curve is turning.

Next, let us find the relationship between the curvature vector $\boldsymbol{P}''(s)$ and the derivative vectors $\dot{\boldsymbol{P}}$ and $\ddot{\boldsymbol{P}}$ with respect to the parameter t. From Eq. (1.9):

$$\boldsymbol{P}'(s) = \frac{\dot{\boldsymbol{P}}}{\sqrt{\dot{\boldsymbol{P}}^2}} .$$

Differentiating the above equation with respect to s gives:

$$\boldsymbol{P}''(s)=\frac{d}{dt}\left(\frac{\dot{\boldsymbol{P}}}{\sqrt{\dot{\boldsymbol{P}}^2}}\right)\frac{dt}{ds}=\frac{\ddot{\boldsymbol{P}}\sqrt{\dot{\boldsymbol{P}}^2}-\dot{\boldsymbol{P}}\left(\dfrac{2\dot{\boldsymbol{P}}}{2\sqrt{\dot{\boldsymbol{P}}^2}}\cdot\ddot{\boldsymbol{P}}\right)}{\dot{\boldsymbol{P}}^2}\;\frac{1}{\sqrt{\dot{\boldsymbol{P}}^2}}$$

$$=\frac{1}{\dot{\boldsymbol{P}}^2}\left[\ddot{\boldsymbol{P}}-\frac{\dot{\boldsymbol{P}}}{\sqrt{\dot{\boldsymbol{P}}^2}}\left(\frac{\dot{\boldsymbol{P}}}{\sqrt{\dot{\boldsymbol{P}}^2}}\cdot\ddot{\boldsymbol{P}}\right)\right]^{4)} \tag{1.18}$$

$$\equiv\frac{\boldsymbol{B}}{\dot{\boldsymbol{P}}^2} \tag{1.19}$$

where:

$$\boldsymbol{B}\equiv\ddot{\boldsymbol{P}}-\frac{\dot{\boldsymbol{P}}}{\sqrt{\dot{\boldsymbol{P}}^2}}\left(\frac{\dot{\boldsymbol{P}}}{\sqrt{\dot{\boldsymbol{P}}^2}}\cdot\ddot{\boldsymbol{P}}\right)\equiv|\boldsymbol{B}|\,\boldsymbol{n}. \tag{1.20}$$

Equation (1.18) gives the relation between the curvature vector and the derivative vectors $\dot{\boldsymbol{P}}(t)$ and $\ddot{\boldsymbol{P}}(t)$ with respect to the parameter t.

Taking the absolute value of Eq. (1.19) gives:

$$|\boldsymbol{P}''(s)|=\frac{1}{\varrho}=\frac{|\boldsymbol{B}|}{|\dot{\boldsymbol{P}}|^2}. \tag{1.21}$$

By using Eqs. (1.18) and (1.21), the center of curvature can be determined graphically as shown in Fig. 1.17.

If we use the relation:

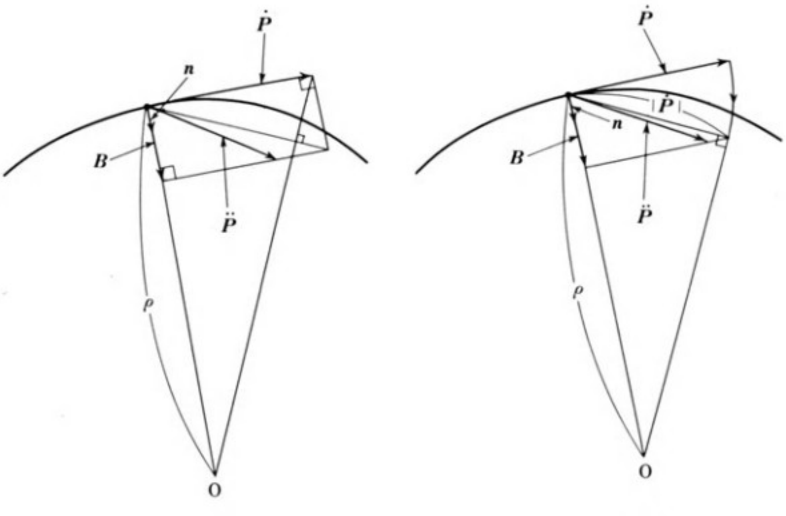

(a) (b)

Fig. 1.17. How to find the center of curvature graphically

$$\dot{s} = \frac{\dot{\boldsymbol{P}}}{\sqrt{\dot{\boldsymbol{P}}^2}} \cdot \ddot{\boldsymbol{P}} \tag{1.22}$$

then Eq. (1.18) becomes:

$$\boldsymbol{P}'' = \frac{1}{\dot{s}^2} (\ddot{\boldsymbol{P}} - \ddot{s}\boldsymbol{t}).$$

Combining this equation with Eq. (1.17) gives:

$$\ddot{\boldsymbol{P}} = \frac{\dot{s}^2}{\varrho} \boldsymbol{n} + \dot{s}\boldsymbol{t}. \tag{1.23}$$

If the parameter t is regarded as time, Eq. (1.23) expresses the well-known dynamical relation that the acceleration $\ddot{\boldsymbol{P}}$ acting on a mass point equals the sum of the tangential acceleration $\ddot{s}\boldsymbol{t}$ and the normal acceleration $(\dot{s}^2/\varrho)\,\boldsymbol{n}$. This relation is shown in Fig. 1.18.

The form of Eq. (1.18) for the curvature vector can be changed as follows:

$$\begin{aligned} \boldsymbol{P}''(s) \equiv \boldsymbol{t}' &= \frac{1}{\dot{\boldsymbol{P}}^4} [\dot{\boldsymbol{P}}^2 \ddot{\boldsymbol{P}} - \dot{\boldsymbol{P}}(\dot{\boldsymbol{P}} \cdot \ddot{\boldsymbol{P}})] \\ &= \frac{(\dot{\boldsymbol{P}} \times \ddot{\boldsymbol{P}}) \times \dot{\boldsymbol{P}}}{\dot{\boldsymbol{P}}^4} && (1.24)^{*)} \\ &= \frac{|\dot{\boldsymbol{P}} \times \ddot{\boldsymbol{P}}|}{|\dot{\boldsymbol{P}}|^3} \boldsymbol{n} && (1.25)^{**)} \\ &\equiv \kappa \boldsymbol{n}. && (1.25') \end{aligned}$$

Consequently, the curvature κ becomes:

$$\kappa = \frac{|\dot{\boldsymbol{P}} \times \ddot{\boldsymbol{P}}|}{|\dot{\boldsymbol{P}}|^3} = \frac{|\dot{\boldsymbol{P}} \times \ddot{\boldsymbol{P}}|}{\dot{s}^3}. \tag{1.26}$$

A point on a curve at which $\kappa = 0$ is called a *point of inflection.*

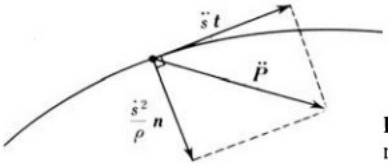

Fig. 1.18. $\ddot{\boldsymbol{P}}$ is the sum of the tangent acceleration and normal acceleration

---

*) Using the identity $\boldsymbol{A} \times (\boldsymbol{B} \times \boldsymbol{C}) = (\boldsymbol{A} \cdot \boldsymbol{C})\boldsymbol{B} - (\boldsymbol{A} \cdot \boldsymbol{B})\boldsymbol{C}$ for the triple vector product

**) $(\boldsymbol{P} \times \boldsymbol{P})$ is perpendicular to $\boldsymbol{P}$, this vector product is itself a vector which points toward the center of curvature.

The vector $\boldsymbol{b}=\boldsymbol{t}\times\boldsymbol{n}$ is called the *unit binormal vector*. The unit binormal vector can be expressed as follows:

$$
\begin{aligned}
\boldsymbol{b} &= \boldsymbol{t}\times\boldsymbol{n}\\
&= \frac{\dot{\boldsymbol{P}}}{|\dot{\boldsymbol{P}}|}\times\left[\frac{1}{\kappa}\,\frac{(\dot{\boldsymbol{P}}\times\ddot{\boldsymbol{P}})\times\dot{\boldsymbol{P}}}{|\dot{\boldsymbol{P}}|^4}\right]=\frac{1}{\kappa}\,\frac{1}{|\dot{\boldsymbol{P}}|^5}\,[\dot{\boldsymbol{P}}\times((\dot{\boldsymbol{P}}\times\ddot{\boldsymbol{P}})\times\dot{\boldsymbol{P}})]\\
&= \frac{1}{\kappa}\,\frac{1}{|\dot{\boldsymbol{P}}|^5}\,[|\dot{\boldsymbol{P}}|^2\,(\dot{\boldsymbol{P}}\times\ddot{\boldsymbol{P}})-(\dot{\boldsymbol{P}}\cdot(\dot{\boldsymbol{P}}\times\ddot{\boldsymbol{P}}))\,\dot{\boldsymbol{P}}]
\end{aligned}
$$

(using the identity for the triple vector product)

$$
=\frac{1}{\kappa}\,\frac{\dot{\boldsymbol{P}}\times\ddot{\boldsymbol{P}}}{|\dot{\boldsymbol{P}}|^3}\quad(\because\ \dot{\boldsymbol{P}}\cdot(\dot{\boldsymbol{P}}\times\ddot{\boldsymbol{P}})=0) \tag{1.27}
$$

$$
=\frac{\dot{\boldsymbol{P}}\times\ddot{\boldsymbol{P}}}{|\dot{\boldsymbol{P}}\times\ddot{\boldsymbol{P}}|}. \tag{1.28}
$$

Figure 1.19 shows the relation between the unit tangent vector $\boldsymbol{t}$, the unit principal normal vector $\boldsymbol{n}$ and the unit binormal vector $\boldsymbol{b}$.

Letting $\Delta\phi$ be the angle between the osculating planes at two slightly separated points $\boldsymbol{P}(s_0)$ and $\boldsymbol{P}(s_0+\Delta s)$, the *torsion* of the curve at point $\boldsymbol{P}(s_0)$ is defined as:

$$
\tau=\lim_{\Delta s\to 0}\frac{\Delta\phi}{\Delta s}. \tag{1.29}
$$

Torsion is a measure of the amount of rotation of the osculating plane with respect to the length of the curve s. In other words, torsion is a quantity that indicates whether the curve is twisting rapidly or slowly. Torsion can be expressed mathematically in terms of the derivatives $\dot{\boldsymbol{P}}$, $\ddot{\boldsymbol{P}}$ and $\dddot{\boldsymbol{P}}$ with respect to the parameter t as follows:

$$
\tau=\frac{(\dot{\boldsymbol{P}}\times\ddot{\boldsymbol{P}})\cdot\dddot{\boldsymbol{P}}}{(\dot{\boldsymbol{P}}\times\ddot{\boldsymbol{P}})^2}=\frac{[\dot{\boldsymbol{P}},\ddot{\boldsymbol{P}},\dddot{\boldsymbol{P}}]}{(\dot{\boldsymbol{P}}\times\ddot{\boldsymbol{P}})^2}=\frac{[\dot{\boldsymbol{P}},\ddot{\boldsymbol{P}},\dddot{\boldsymbol{P}}]}{\dot{s}^6\,\kappa^2}. \tag{1.30}
$$

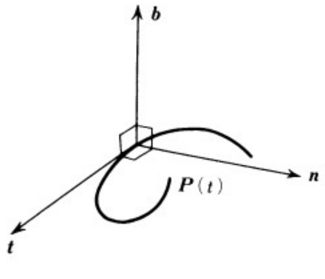

Fig. 1.19. Relation between the unit tangent vector, unit principal normal vector, and unit binormal vector $\boldsymbol{b}$, unit binormal vector; $\boldsymbol{n}$, unit principal normal vector; $\boldsymbol{t}$, unit tangent vector

The condition for a curve to be a plane curve, in other words a curve with zero torsion, is:

$$(\dot{\boldsymbol{P}} \times \ddot{\boldsymbol{P}}) \cdot \dddot{\boldsymbol{P}} = \begin{vmatrix} \dot{x} & \dot{y} & \dot{z} \\ \ddot{x} & \ddot{y} & \ddot{z} \\ x & y & z \end{vmatrix} = 0. \tag{1.31}$$

1.2.3 Frenet Frames and the Frenet-Serret Equations

If we take the directions of the unit tangent vector $\boldsymbol{t}$, the unit principal normal vector $\boldsymbol{n}$ and the unit binormal vector $\boldsymbol{b}$ to be the positive directions along the tangent, the principal normal and the binormal, respectively, then the coordinate system formed by the tangent, the principal normal and the binormal is a right-handed system. These coordinate axes move as the parameter t varies (refer to Fig. 1.19). The set $(\boldsymbol{P}; \boldsymbol{t}, \boldsymbol{n}, \boldsymbol{b})$ is called a *moving frame* or a *Frenet frame*. When properties of a space curve are investigated in the vicinity of various points on it, the Frenet frame of each point is used.

Let us now look at the changes in the Frenet frame as the curve parameter is varied. First from the relation for the unit tangent vector, we have:

$$\frac{d\boldsymbol{P}}{ds} \equiv \boldsymbol{P}' \equiv \boldsymbol{t}.$$

Then, from another of the relations which we derived (refer to Eq. (1.17)):

$$\frac{d\boldsymbol{t}}{ds} \equiv \boldsymbol{t}' = \kappa \boldsymbol{n}.$$

Next, let us find $d\boldsymbol{n}/ds \equiv \boldsymbol{n}'$. Since $\boldsymbol{n}$ is a unit vector,

$$\boldsymbol{n}^2 = 1.$$

Differentiating both sides of this equation with respect to s gives:

$$\boldsymbol{n} \cdot \frac{d\boldsymbol{n}}{ds} = 0.$$

We see from this that $d\boldsymbol{n}/ds$ is perpendicular to $\boldsymbol{n}$. Therefore $d\boldsymbol{n}/ds$ can be written in the following form:

$$\frac{d\boldsymbol{n}}{ds} = a_1 \boldsymbol{t} + a_2 \boldsymbol{b}. \tag{1.32}$$

Taking the scalar product of both sides of this equation with $\boldsymbol{t}$ gives:

$$\boldsymbol{t} \cdot \frac{d\boldsymbol{n}}{ds} = a_1. \tag{1.33}$$

Since we also have the relation:

$$\boldsymbol{t} \cdot \boldsymbol{n} = 0$$

differentiating by s gives:

$$\frac{d\boldsymbol{t}}{ds} \cdot \boldsymbol{n} + \boldsymbol{t} \cdot \frac{d\boldsymbol{n}}{ds} = 0.$$

Consequently we have:

$$\boldsymbol{t} \cdot \frac{d\boldsymbol{n}}{ds} = -\kappa. \tag{1.34}$$

Comparing Eqs. (1.33) and (1.34), we see that $a_1 = -\kappa$. Next, take the scalar product of $\boldsymbol{b}$ with both sides of Eq. (1.32):

$$\boldsymbol{b} \cdot \frac{d\boldsymbol{n}}{ds} = a_2. \tag{1.35}$$

Since we also have the relation:

$$\boldsymbol{b} \cdot \boldsymbol{n} = 0$$

differentiating by s gives:

$$\frac{d\boldsymbol{b}}{ds} \cdot \boldsymbol{n} + \boldsymbol{b} \cdot \frac{d\boldsymbol{n}}{ds} = 0. \tag{1.36}$$

The unit binormal vector $\boldsymbol{b}$ is perpendicular to the osculating plane. Since the torsion is the rate of rotation of the osculating plane with respect to s,

$$\left| \frac{d\boldsymbol{b}}{ds} \right| = \tau. \tag{1.37}$$

Differentiating the relation

$$\boldsymbol{b} \cdot \boldsymbol{t} = 0$$

by s gives:

$$\frac{d\boldsymbol{b}}{ds} \cdot \boldsymbol{t} + \boldsymbol{b} \cdot \frac{d\boldsymbol{t}}{ds} = 0.$$

The second term on the left-hand side of this equation is:

$$\boldsymbol{b} \cdot \frac{d\boldsymbol{t}}{ds} = \boldsymbol{b} \cdot \kappa \boldsymbol{n} = 0,$$

so that:

$$\frac{d\boldsymbol{b}}{ds} \cdot \boldsymbol{t} = 0.$$

From this relation, we see that $d\boldsymbol{b}/ds$ is a vector that is perpendicular to both $\boldsymbol{b}$ and $\boldsymbol{t}$, that is, a vector that is directed along the principal normal. Its magnitude is given by Eq. (1.37) as τ. If we adopt the convention:

$$\frac{d\boldsymbol{b}}{ds} = -\tau \boldsymbol{n}$$

then Eq. (1.36) becomes:

$$-\tau + \boldsymbol{b} \cdot \frac{d\boldsymbol{n}}{ds} = 0.$$

Comparing this equation with Eq. (1.35) we see that $a_2 = \tau$, so that Eq. (1.32) becomes:

$$\frac{d\boldsymbol{n}}{ds} \equiv \boldsymbol{n}' = -\kappa \boldsymbol{t} + \tau \boldsymbol{b}. \tag{1.38}$$

The above results can be summarized as follows:

$$\boldsymbol{P}' = \boldsymbol{t} \tag{1.39}$$

$$\boldsymbol{t}' = \kappa \boldsymbol{n} \tag{1.40}$$

$$\boldsymbol{n}' = -\kappa \boldsymbol{t} + \tau \boldsymbol{b} \tag{1.41}$$

$$\boldsymbol{b}' = -\tau \boldsymbol{n}. \tag{1.42}$$

These equations are called the *Frenet-Serret* equations. The differential geometric properties of space curves are all derived from these equations.

1.2.4 Calculation of a Point on a Curve[6)]

Consider a curve that is expressed as a polynomial in the parameter t, such as the following cubic expression:

$$\boldsymbol{P}(t) = \boldsymbol{A}t^3 + \boldsymbol{B}t^2 + \boldsymbol{C}t + \boldsymbol{D} \quad (0 \leqq t \leqq 1). \tag{1.43}$$

A point on the curve can be calculated by dividing the t-interval up into n segments as follows:

$$\left.\begin{aligned} t &= 0,\ \delta,\ 2\delta, \ldots, n\delta \\ \delta &= 1/n \end{aligned}\right\} \tag{1.44}$$

and substituting into Eq. (1.43) n sequence. In this case it is convenient to change Eq. (1.43) into the following form:

$$\boldsymbol{P}(t) = ((\boldsymbol{A}t + \boldsymbol{B})t + \boldsymbol{C})t + \boldsymbol{D}. \tag{1.45}$$

Equation (1.43) requires 6 multiplications per coordinate per point, but Eq. (1.45) requires only 3, so computations can be performed more rapidly.

Next, let us consider a method of calculating points on a curve at high speed using no multiplications at all, by using finite difference formulas.

$$\begin{aligned}
\boldsymbol{P}(t) &= \boldsymbol{A}t^3 + \boldsymbol{B}t^2 + \boldsymbol{C}t + \boldsymbol{D} \\
\Delta \boldsymbol{P}(t) &= \boldsymbol{P}(t+\delta) - \boldsymbol{P}(t) \\
&= \boldsymbol{A}(t+\delta)^3 + \boldsymbol{B}(t+\delta)^2 + \boldsymbol{C}(t+\delta) + \boldsymbol{D} - \boldsymbol{A}t^3 - \boldsymbol{B}t^2 - \boldsymbol{C}t - \boldsymbol{D} \\
&= 3\boldsymbol{A}\delta t^2 + (3\boldsymbol{A}\delta^2 + 2\boldsymbol{B}\delta)t + \boldsymbol{A}\delta^3 + \boldsymbol{B}\delta^2 + \boldsymbol{C}\delta \\
\Delta^2 \boldsymbol{P}(t) &= \Delta \boldsymbol{P}(t+\delta) - \Delta \boldsymbol{P}(t) \\
&= 3\boldsymbol{A}\delta(t+\delta)^2 + (3\boldsymbol{A}\delta^2 + 2\boldsymbol{B}\delta)\,(t+\delta) + \boldsymbol{A}\delta^3 + \boldsymbol{B}\delta^2 + \boldsymbol{C}\delta \\
&\quad - 3\boldsymbol{A}\delta t^2 - (3\boldsymbol{A}\delta^2 + 2\boldsymbol{B}\delta)t - \boldsymbol{A}\delta^3 - \boldsymbol{B}\delta^2 - \boldsymbol{C}\delta \\
&= 6\boldsymbol{A}\delta^2 t + 6\boldsymbol{A}\delta^3 + 2\boldsymbol{B}\delta^2 \\
\Delta^3 \boldsymbol{P}(t) &= \Delta^2 \boldsymbol{P}(t+\delta) - \Delta^2 \boldsymbol{P}(t) \\
&= 6\boldsymbol{A}\delta^2(t+\delta) + 6\boldsymbol{A}\delta^3 + 2\boldsymbol{B}\delta^2 - 6\boldsymbol{A}\delta^2 t - 6\boldsymbol{A}\delta^3 - 2\boldsymbol{B}\delta^2 \\
&= 6\boldsymbol{A}\delta^3.
\end{aligned}$$

The 3rd-order finite difference is a constant vector that is independent of the parameter. In general, for an nth-order polynomial espression, the nth-order finite difference is a constant vector.

Let us use the notation $\boldsymbol{P}_0$, $\Delta\boldsymbol{P}_0$, $\Delta^2\boldsymbol{P}_0$, $\Delta^3\boldsymbol{P}_0$ and $\boldsymbol{P}_i$, $\Delta\boldsymbol{P}_i$, $\Delta^2\boldsymbol{P}_i$, $\Delta^3\boldsymbol{P}_i$ to denote the various order finite differences at $t=0$ and $t=i\delta$ respectively. Then we have the relations:

$$\left.\begin{aligned}
\boldsymbol{P}_1 &= \boldsymbol{P}_0 + \Delta \boldsymbol{P}_0 \\
\Delta \boldsymbol{P}_1 &= \Delta \boldsymbol{P}_0 + \Delta^2 \boldsymbol{P}_0 \\
\Delta^2 \boldsymbol{P}_1 &= \Delta^2 \boldsymbol{P}_0 + \Delta^3 \boldsymbol{P}_0 \\
\Delta^3 \boldsymbol{P}_1 &= \Delta^3 \boldsymbol{P}_0
\end{aligned}\right\} \tag{1.46}$$

so that if the various order finite differences at $t=0$, that is, $\boldsymbol{P}_0$, $\Delta\boldsymbol{P}_0$, $\Delta^2\boldsymbol{P}_0$, $\Delta^3\boldsymbol{P}_0$ are known, then the points on the curve at $t=\delta$, 2δ, ..., $i\delta$, ... can be found in sequence solely by repeated additions. However, if, for example,

$\delta = 1/2^5$, finite differences such as $\varDelta^3 \boldsymbol{P}_0$ become vectors with very small magnitudes, so that in some cases the length of a computer word is not adequate to maintain sufficient accuracy. In such cases this problem can be solved by storing $\varDelta \boldsymbol{P}_0$, $\varDelta^2 \boldsymbol{P}_0$, $\varDelta^3 \boldsymbol{P}_0$ as $1/\delta$ precision, $1/\delta^2$ precision, $1/\delta^3$ precision numbers, respectively, and applying Eqs. (1.46) in the form:

$$\left.\begin{aligned}
\boldsymbol{P}_1 &= \boldsymbol{P}_0 + \delta\left(\frac{1}{\delta}\,\varDelta \boldsymbol{P}_0\right)\\
\frac{1}{\delta}\,\varDelta \boldsymbol{P}_1 &= \frac{1}{\delta}\,\varDelta \boldsymbol{P}_0 + \delta\left(\frac{1}{\delta^2}\,\varDelta^2 \boldsymbol{P}_0\right)\\
\frac{1}{\delta^2}\,\varDelta^2 \boldsymbol{P}_1 &= \frac{1}{\delta^2}\,\varDelta^2 \boldsymbol{P}_0 + \delta\left(\frac{1}{\delta^3}\,\varDelta^3 \boldsymbol{P}_0\right)\\
\frac{1}{\delta^3}\,\varDelta^3 \boldsymbol{P}_1 &= \frac{1}{\delta^3}\,\varDelta^3 \boldsymbol{P}_0
\end{aligned}\right\} \tag{1.47}$$

If we take $\delta = 1/2^m$ (where m is a positive integer) in the above equations, then the necessary computations can be performed at high speed using exclusively shift operations and additions, without time-consuming multiplications.

In calculations with Eqs. (1.47) it is convenient to express the various order finite differences at $t=0$ as a 4×3 matrix:

$$\begin{bmatrix} \boldsymbol{P}_0 \\ \dfrac{1}{\delta}\,\varDelta \boldsymbol{P}_0 \\ \dfrac{1}{\delta^2}\,\varDelta^2 \boldsymbol{P}_0 \\ \dfrac{1}{\delta^3}\,\varDelta^3 \boldsymbol{P}_0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{D} \\ \boldsymbol{A}\delta^2 + \boldsymbol{B}\delta + \boldsymbol{C} \\ 6\boldsymbol{A}\delta + 2\boldsymbol{B} \\ 6\boldsymbol{A} \end{bmatrix} \tag{1.48}$$

In this book, the matrix in Eq. (1.48) will be referred to as the finite difference matrix.

1.2.5 Connection of Curve Segments

Let us now consider the problem of connecting a curve segment $\boldsymbol{P}_{\text{II}}(t)$ $(0 \leqq t \leqq 1)$ to the curve segment $\boldsymbol{P}_{\text{I}}(t)$ $(0 \leqq t \leqq 1)$ (Fig. 1.20).

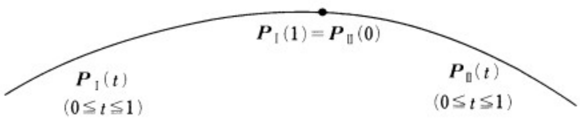
$\boldsymbol{P}_{\text{I}}(1)=\boldsymbol{P}_{\text{II}}(0)$
$\boldsymbol{P}_{\text{I}}(t)$
$(0\leq t\leq 1)$
$\boldsymbol{P}_{\text{II}}(t)$
$(0\leq t\leq 1)$

Fig. 1.20. Continuity conditions for curves Conditions for continuity up to curvature
① $\boldsymbol{P}_{\text{II}}(0)=\boldsymbol{P}_{\text{I}}(1)$; ② $\dot{\boldsymbol{P}}_{\text{II}}(0)=\dfrac{\alpha_2}{\alpha_1}\,\dot{\boldsymbol{P}}_{\text{I}}(1)$, ③ $\ddot{\boldsymbol{P}}_{\text{II}}(0)=\left(\dfrac{\alpha_2}{\alpha_1}\right)^2\ddot{\boldsymbol{P}}_{\text{I}}(1)+\beta\,\dot{\boldsymbol{P}}_{\text{I}}(1)$

First, we have the condition of continuity of position:

$$\boldsymbol{P}_{\mathrm{II}}(0)=\boldsymbol{P}_{\mathrm{I}}(1). \tag{1.49}$$

If the magnitudes of the tangent vectors of $\boldsymbol{P}_{\mathrm{I}}(t)$ and $\boldsymbol{P}_{\mathrm{II}}(t)$ at the connection point are α_1 and α_2 respectively, then the condition for the slope to be continuous is:

$$\frac{1}{\alpha_2}\dot{\boldsymbol{P}}_{\mathrm{II}}(0)=\frac{1}{\alpha_1}\dot{\boldsymbol{P}}_{\mathrm{I}}(1)=\boldsymbol{t} \tag{1.50}$$

or:

$$\dot{\boldsymbol{P}}_{\mathrm{II}}(0)=\frac{\alpha_2}{\alpha_1}\dot{\boldsymbol{P}}_{\mathrm{I}}(1)=\lambda\dot{\boldsymbol{P}}_{\mathrm{I}}(1)\quad\left(\lambda=\frac{\alpha_2}{\alpha_1}\right). \tag{1.51}$$

Next, let us find the condition for the centers of curvature of the two curve segments at the connection point to vary continuously. Equation (1.17) for the curvature is:

$$\boldsymbol{P}''=\kappa\boldsymbol{n}.$$

In this equation κ is the magnitude of the curvature and $\boldsymbol{n}$ is the unit principal normal vector, pointing toward the center of curvature. From the above equation, for the center of curvature of the curve to vary smoothly at the connection point, it is necessary for both κ and $\boldsymbol{n}$ to be continuous.

If the binormal $\boldsymbol{b}=\boldsymbol{t}\times\boldsymbol{n}$ and $\boldsymbol{t}$ are continuous then $\boldsymbol{n}$ is also continuous. Consequently, to find the condition κ and $\boldsymbol{n}$ to be continuous, we require a condition for both κ and $\boldsymbol{b}$ to be continuous. From Eq. (1.27):

$$\kappa\boldsymbol{b}=\frac{\dot{\boldsymbol{P}}\times\ddot{\boldsymbol{P}}}{|\dot{\boldsymbol{P}}|^3}.$$

Consequently, in order for the center of curvature of a curve to vary continuously at a connection point, we must have:

$$\frac{\dot{\boldsymbol{P}}_{\mathrm{II}}(0)\times\ddot{\boldsymbol{P}}_{\mathrm{II}}(0)}{|\dot{\boldsymbol{P}}_{\mathrm{II}}(0)|^3}=\frac{\dot{\boldsymbol{P}}_{\mathrm{I}}(1)\times\ddot{\boldsymbol{P}}_{\mathrm{I}}(1)}{|\dot{\boldsymbol{P}}_{\mathrm{I}}(1)|^3}.$$

Substituting the relation in Eq. (1.51) into this equation and rearranging gives:

$$\dot{\boldsymbol{P}}_{\mathrm{I}}(1)\times\left[\ddot{\boldsymbol{P}}_{\mathrm{II}}(0)-\left(\frac{\alpha_2}{\alpha_1}\right)^2\ddot{\boldsymbol{P}}_{\mathrm{I}}(1)\right]=\boldsymbol{0}.$$

From this equation, the following equation is obtained, with β an arbitrary scalar[7].

$$\begin{aligned}\ddot{\boldsymbol{P}}_{\mathrm{II}}(0)&=\left(\frac{\alpha_2}{\alpha_1}\right)^2\ddot{\boldsymbol{P}}_{\mathrm{I}}(1)+\beta\dot{\boldsymbol{P}}_{\mathrm{I}}(1)\\&=\lambda^2\ddot{\boldsymbol{P}}_{\mathrm{I}}(1)+\beta\dot{\boldsymbol{P}}_{\mathrm{I}}(1)\quad\left(\lambda=\frac{\alpha_2}{\alpha_1}\right).\end{aligned}\tag{1.52}$$

1.2.6 Parameter Transformation

Normally, a curve segment $\boldsymbol{P}(t)$ is normalized so that $t=0$ at one end and $t=1$ at the other end. If we are to use only one part of this curve, from $t=t_0$ to $t=t_1$, or perform some processing on this part, it is convenient to normalize the parameter over this part of the curve (refer to Fig. 1.21). In order to keep the degree of the curve unchanged when performing this transformation, the following linear relation must hold between the old parameter t and the parameter after transformation t^*.

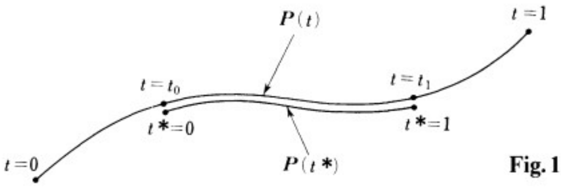

Fig. 1.21. Curve parameter transformation

$$t=(1-t^*)\,t_0+t^*t_1=t_0+(t_1-t_0)\,t^*.\tag{1.53}$$

Substituting this relation for t into the original equation of the curve gives the normalized curve equation in terms of t^*, with t^* varying from 0 to 1 over the interval of interest. For example, in the case of the cubic curve:

$$\boldsymbol{P}(t)=[t^3\ \ t^2\ \ t\ \ 1]\ [\boldsymbol{A}\ \ \boldsymbol{B}\ \ \boldsymbol{C}\ \ \boldsymbol{D}]^T\tag{1.54}$$

the normalized curve equation over the interval $t_0\leqq t\leqq t_1$ is:

$$\begin{aligned}\boldsymbol{P}(t^*)&=[\{t_0+(t_1-t_0)\,t^*\}^3\quad\{t_0+(t_1-t_0)\,t^*\}^2\quad\{t_0+(t_1-t_0)\,t^*\}\quad 1]\begin{bmatrix}\boldsymbol{A}\\\boldsymbol{B}\\\boldsymbol{C}\\\boldsymbol{D}\end{bmatrix}\\&=\begin{matrix}[(t_1-t_0)^3t^{*3}+3\,(t_1-t_0)^2\,t_0\,t^{*2}+3\,(t_1-t_0)\,t_0^2\,t^*+t_0^3\\(t_1-t_0)^2\,t^{*2}+2\,(t_1-t_0)\,t_0\,t^*+t_0^2\quad(t_1-t_0)\,t^*+t_0\quad 1]\end{matrix}\begin{bmatrix}\boldsymbol{A}\\\boldsymbol{B}\\\boldsymbol{C}\\\boldsymbol{D}\end{bmatrix}=\end{aligned}$$

$$= [t^{*3}\ t^{*2}\ t^*\ 1] \begin{bmatrix} (t_1-t_0)^3 & 0 & 0 & 0 \\ 3(t_1-t_0)^2 t_0 & (t_1-t_0)^2 & 0 & 0 \\ 3(t_1-t_0)t_0^2 & 2(t_1-t_0)t_0 & t_1-t_0 & 0 \\ t_0^3 & t_0^2 & t_0 & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{B} \\ \boldsymbol{C} \\ \boldsymbol{D} \end{bmatrix} \tag{1.55}$$

$$= [t^{*3}\ t^{*2}\ t^*\ 1] \begin{bmatrix} (t_1-t_0)^3 \boldsymbol{A} \\ 3(t_1-t_0)^2 t_0 \boldsymbol{A} + (t_1-t_0)^2 \boldsymbol{B} \\ 3(t_1-t_0)t_0^2 \boldsymbol{A} + 2(t_1-t_0)t_0 \boldsymbol{B} + (t_1-t_0)\boldsymbol{C} \\ t_0^3 \boldsymbol{A} + t_0^2 \boldsymbol{B} + t_0 \boldsymbol{C} + \boldsymbol{D} \end{bmatrix} \tag{1.56}$$

$$\equiv [t^{*3}\ t^{*2}\ t^*\ 1] \begin{bmatrix} \boldsymbol{A}^* \\ \boldsymbol{B}^* \\ \boldsymbol{C}^* \\ \boldsymbol{D}^* \end{bmatrix}.$$

Let us now see how the tangent vector of the curve changes under the parameter transformation of Eq. (1.53). The tangent vector after transformation becomes:

$$\frac{d\boldsymbol{P}}{dt^*} = \frac{d\boldsymbol{P}}{dt}\frac{dt}{dt^*} = (t_1-t_0)\frac{d\boldsymbol{P}}{dt}. \tag{1.57}$$

That is, the magnitude of the tangent vector is reduced in proportion to the reduction of the range through which the original parameter varies ($0 \leqq t_1 - t_0 \leqq 1$). For the second derivative vector we have:

$$\frac{d^2\boldsymbol{P}}{dt^{*2}} = (t_1-t_0)^2 \frac{d^2\boldsymbol{P}}{dt^2} \tag{1.58}$$

so that it is reduced in proportion to the square of the reduction of the parameter range.

Let us now consider a more general parameter transformation, the rational transformation given by the following equation:

$$t = \frac{at^* + b}{ct^* + 1}.$$

This transformation is sometimes called *bilinear transformation* or a *homographic transformation.* With $t^*=0$ and $t^*=1$ at $t=t_0$ and $t=t_1$, respectively, this transformation takes the following form, with η an arbitrary scalar:

$$t = \frac{(\eta t_1 - t_0)t^* + t_0}{(\eta - 1)t^* + 1}. \tag{1.59}$$

The shape within the old parameter range $t_0 \leqq t \leqq t_1$ is now expressed without shape change in terms of the new parameter which varies in the domain $0 \leqq t^* \leqq 1$. Compared to the transformation of Eq. (1.53), the transformation of Eq. (1.59) has an increased degree of freedom because η can be chosen freely. When $\eta = 1$, the transformation of Eq. (1.59) reduces to the transformation of Eq. (1.53). As η varies, what changes is the relation between a point on the curve and the parameter value; the shape does not change.

In the special case $t_0 = 0$, $t_1 = 1$ we have:

$$t = \frac{\eta t^*}{(\eta - 1)t^* + 1}. \tag{1.60}$$

Equation (1.59) is frequently used in rational polynomial parameter transformations, because only the parameter range changes, while the shape and the degree of the equation remain unchanged. Equation (1.60) is used when one wishes to change only the relation between points on the curve and parameter values, without changing the parameter range.

When the following curve:

$$\left.\begin{aligned} x(t) &= \frac{a_{11}t^2 + a_{21}t + a_{31}}{a_{13}t^2 + a_{23}t + a_{33}} \\ y(t) &= \frac{a_{12}t^2 + a_{22}t + a_{32}}{a_{13}t^2 + a_{23}t + a_{33}} \end{aligned}\right\} \tag{1.61}$$

expressed in terms of rational polynomials in which both the numerator and denominators are quadratic expressions is transformed using Eq. (1.60), the result is similar in which both the numerators and denominators are quadratic expressions. Let us demonstrate this.

The homogeneous vector representation $\boldsymbol{P}(t) = [x(t)\ y(t)\ 1]$ for the rational polynomial curve expressed by Eqs. (1.61) is*):

$$w(t)\boldsymbol{P}(t) = [t^2\ \ t\ \ 1] \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}. \tag{1.62}$$

Substituting the parameter relation of Eq. (1.60) into this equation gives:

$$w(t^*)\boldsymbol{P}(t^*) = \left[\left\{\frac{\eta t^*}{(\eta-1)t^*+1}\right\}^2 \ \ \frac{\eta t^*}{(\eta-1)t^*+1} \ \ 1\right] \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \doteq$$

---

*) As for homogeneous coordinates, refer to Newman & Sproull Principles of Interactive Computer Graphics, Second edition, McGraw-Hill

$$\doteq [\eta^2 t^{*2} \quad \eta t^* \{(\eta-1)t^*+1\} \quad \{(\eta-1)t^*+1\}^2] \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}^{**)}$$

$$= [t^{*2} \; t^* \; 1] \begin{bmatrix} \eta^2 & \eta(\eta-1) & (\eta-1)^2 \\ 0 & \eta & 2(\eta-1) \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}. \tag{1.63}$$

Consequently, in terms of ordinary coordinates we have:

$$x(t^*) = \frac{\{\eta^2 a_{11} + \eta(\eta-1)a_{21} + (\eta-1)^2 a_{31}\} t^{*2} + \{\eta a_{21} + 2(\eta-1)a_{31}\} t^* + a_{31}}{\{\eta^2 a_{13} + \eta(\eta-1)a_{23} + (\eta-1)^2 a_{33}\} t^{*2} + \{\eta a_{23} + 2(\eta-1)a_{33}\} t^* + a_{33}}$$

$$y(t^*) = \frac{\{\eta^2 a_{12} + \eta(\eta-1)a_{22} + (\eta-1)^2 a_{32}\} t^{*2} + \{\eta a_{22} + 2(\eta-1)a_{32}\} t^* + a_{32}}{\{\eta^2 a_{13} + \eta(\eta-1)a_{23} + (\eta-1)^2 a_{33}\} t^{*2} + \{\eta a_{23} + 2(\eta-1)a_{33}\} t^* + a_{33}}.$$

Next, let us look at the change in the tangent vector under this parameter transformation. The tangent vector after the transformation is:

$$\frac{d\boldsymbol{P}}{dt^*} = \frac{d\boldsymbol{P}}{dt}\frac{dt}{dt^*}.$$

From Eq. (1.60), we have:

$$\frac{dt}{dt^*} = \frac{\eta}{\{(\eta-1)t^*+1\}^2}$$

$$\therefore \frac{d\boldsymbol{P}}{dt^*} = \frac{\eta}{\{(\eta-1)t^*+1\}^2}\frac{d\boldsymbol{P}}{dt}. \tag{1.64}$$

Therefore, the tangent vectors at points $t^*=0$ and $t^*=1$ are, respectively:

$$\left.\frac{d\boldsymbol{P}}{dt^*}\right|_{t^*=0} = \eta \left.\frac{d\boldsymbol{P}}{dt}\right|_{t=0}, \qquad \left.\frac{d\boldsymbol{P}}{dt^*}\right|_{t^*=1} = \frac{1}{\eta}\left.\frac{d\boldsymbol{P}}{dt}\right|_{t=1}. \tag{1.65}$$

As a result of using transformation (1.60), the shape is unchanged, the range within which the parameter varies remains the same, and the variation of a point on the curve for a given change in the parameter is changed. As a result, the magnitudes of the tangent vectors at both ends are changed, but the product of the magnitudes of the two tangent vectors remains the same. That is:

**) Multiplying a homogeneous vector by a nonzero scalar constant does not alter its ordinary coordinate vector We shall define the equality in ordinary coordinate vector with the symbol =

$$\left.\frac{d\boldsymbol{P}}{dt^*}\right|_{t^*=0}\cdot\left.\frac{d\boldsymbol{P}}{dt^*}\right|_{t^*=1}=\left.\frac{d\boldsymbol{P}}{dt}\right|_{t=0}\cdot\left.\frac{d\boldsymbol{P}}{dt}\right|_{t=1}. \tag{1.66}$$

1.2.7 Partitioning of a Curve Segment

Let us consider a case in which a curve $\boldsymbol{P}(t)$ is split into two curves $\boldsymbol{P}_1(u)$ $(0 \leqq u \leqq 1)$ and $\boldsymbol{P}_2(u)$ $(0 \leqq u \leqq 1)$ at a point corresponding to $t=t_s$. We will demonstrate this using the cubic curve (1.43) (Fig. 1.22).

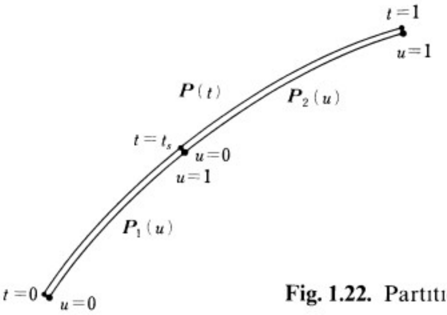

Fig. 1.22. Partitioning of a curve segment

The first curve $\boldsymbol{P}_1(u)$ is obtained by setting $t_0=0$, $t_1=t_s$ and $t^*=u$ in Eq. (1.56). This gives the following expression:

$$\boldsymbol{P}_1(u)=[u^3\ u^2\ u\ 1]\begin{bmatrix} t_s^3\boldsymbol{A} \\ t_s^2\boldsymbol{B} \\ t_s\boldsymbol{C} \\ \boldsymbol{D}\end{bmatrix}. \tag{1.67}$$

The second curve $\boldsymbol{P}_2(u)$ is obtained by setting $t_0=t_s$, $t_1=1$ and $t^*=u$ in Eq. (1.56) to give the expression:

$$\boldsymbol{P}_2(u)=[u^3\ u^2\ u\ 1]\begin{bmatrix} (1-t_s)^3\boldsymbol{A} \\ 3(1-t_s)^2t_s\boldsymbol{A}+(1-t_s)^2\boldsymbol{B} \\ 3(1-t_s)t_s^2\boldsymbol{A}+2(1-t_s)t_s\boldsymbol{B}+(1-t_s)\boldsymbol{C} \\ t_s^3\boldsymbol{A}+t_s^2\boldsymbol{B}+t_s\boldsymbol{C}+\boldsymbol{D}\end{bmatrix}. \tag{1.68}$$

Partitioning the curve in this way makes it possible to apply local control to its shape.

1.2.8 Parametric Cubic Curves

The parametric cubic curve segment (1.43) is completely determined by setting the parameter equal to 0, the position vector to $\boldsymbol{Q}_0$ and the tangent vector to $\dot{\boldsymbol{Q}}_0$ at one end, and the parameter to 1, the position vector to $\boldsymbol{Q}_1$ and the tangent vector to $\dot{\boldsymbol{Q}}_1$ at the other end (for further details refer to Sect. 3.2.1). This gives:

$$\begin{aligned}
&\boldsymbol{A}=2(\boldsymbol{Q}_0-\boldsymbol{Q}_1)+\dot{\boldsymbol{Q}}_0+\dot{\boldsymbol{Q}}_1\\
&\boldsymbol{B}=3(\boldsymbol{Q}_1-\boldsymbol{Q}_0)-2\,\dot{\boldsymbol{Q}}_0-\dot{\boldsymbol{Q}}_1\\
&\boldsymbol{C}=\dot{\boldsymbol{Q}}_0\\
&\boldsymbol{D}=\boldsymbol{Q}_0\,.
\end{aligned}$$

If we have two curve segments such that the two end points of one curve and the directions of the tangent vectors are the same as those of the other,

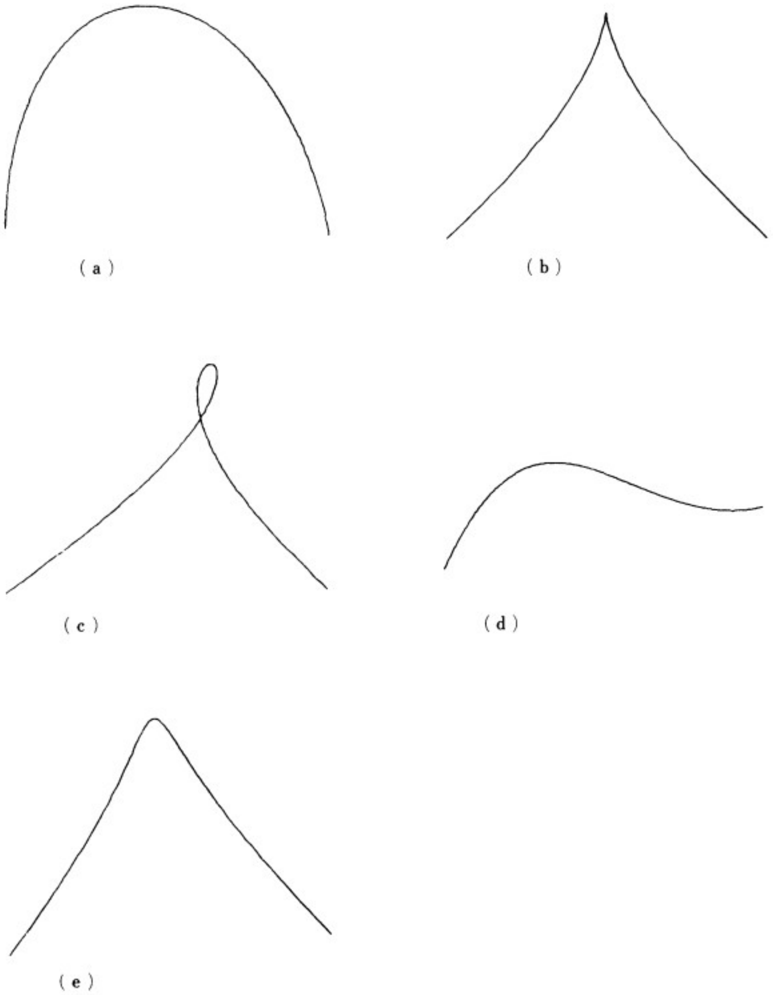

Fig. 1.23. Shapes of parametric cubic curves. (a) case with no abnormalities, (b) cusp; (c) loop, (d) one inflection point; (e) two inflection points

then, if the magnitudes of the tangent vectors at these end points are different, the shapes will be different. The magnitudes as well as directions of the tangent vectors affect the shapes of the curves.

Let us focus attention on a point of inflection of a curve segment. From Eq. (1.26), a point of inflection is a point at which:

$$\dot{\boldsymbol{P}} \times \ddot{\boldsymbol{P}} = \boldsymbol{0}.$$

If we exclude points at which $\dot{\boldsymbol{P}} = \boldsymbol{0}$, then a point of inflection is a point at which:

$$\ddot{\boldsymbol{P}} = g\,\dot{\boldsymbol{P}} \tag{1.69}$$

where g is a scalar. Since this is a second-degree equation with respect to t, we see that as the parameter varies from 0 to 1, a parametric cubic curve can have a maximum of 2 inflection points. A typical shape of a parametric cubic curve is shown in Fig. 1.23.

Next, let us look at the torsion of a parametric cubic curve.

$$\left.\begin{aligned} \dot{\boldsymbol{P}}(t) &= [\dot{x}\ \dot{y}\ \dot{z}] = 3\boldsymbol{A}t^2 + 2\boldsymbol{B}t + \boldsymbol{C} \\ \ddot{\boldsymbol{P}}(t) &= [\ddot{x}\ \ddot{y}\ \ddot{z}] = 6\boldsymbol{A}t + 2\boldsymbol{B} \\ \dddot{\boldsymbol{P}}(t) &= [x\ \ y\ \ z] = 6\boldsymbol{A} \end{aligned}\right\} . \tag{1.70}$$

Substituting these relations into Eq. (1.31), we obtain:

$$\begin{vmatrix} \dot{x} & \dot{y} & \dot{z} \\ \ddot{x} & \ddot{y} & \ddot{z} \\ x & y & z \end{vmatrix} = \begin{vmatrix} 3A_x t^2 + 2B_x t + C_x & 3A_y t^2 + 2B_y t + C_y & 3A_z t^2 + 2B_z t + C_z \\ 6A_x t + 2B_x & 6A_y t + 2B_y & 6A_z t + 2B_z \\ 6A_x & 6A_y & 6A_z \end{vmatrix}$$

$$= -12 \begin{vmatrix} A_x & A_y & A_z \\ B_x & B_y & B_z \\ C_x & C_y & C_z \end{vmatrix}$$

Therefore, the condition for a parametric cubic curve to be a plane curve is:

$$\begin{vmatrix} A_x & A_y & A_z \\ B_x & B_y & B_z \\ C_x & C_y & C_z \end{vmatrix} = 0. \tag{1.71}$$

From this equation a curve for which $A_x = A_y = A_z = 0$, that is a quadratic parametric curve, has zero torsion and is a plane curve. Therefore, we see that the minimum degree that a polynomial curve must have in order to be a space curve is 3. In addition, even for a cubic curve, if the 2nd-degree coefficients (B_x, B_y, B_z) or the 1st-degree coefficients (C_x, C_y, C_z) are 0, the curve is a plane curve.

1.2.9 Length and Area of a Curve

A very small segment of a curve has approximately the same length as the chord which subtends that segment. The chord vector obtained by varying the parameter by Δt is approximately $\dot{\boldsymbol{P}}(t)\Delta t$, so the length l of the curve from the parameter value t_0 to t_1 can be calculated as:

$$l=\int_{t_0}^{t_1} |\dot{\boldsymbol{P}}(t)|\, dt. \tag{1.72}$$

If we assume that the curve is a plane curve that can be represented by Eq. (1.43), then:

$$\begin{aligned} l&=\int_{t_0}^{t_1} \sqrt{A_4t^4+A_3t^3+A_2t^2+A_1t+A_0}\, dt\\ &\equiv\int_{t_0}^{t_1} f(t)\, dt. \end{aligned} \tag{1.73}$$

Here:

$$\begin{aligned} &f(t)=\sqrt{A_4t^4+A_3t^3+A_2t^2+A_1t+A_0}\\ &A_4=9(A_x^2+A_y^2)\\ &A_3=12(A_xB_x+A_yB_y)\\ &A_2=6(A_xC_x+A_yC_y)+4(B_x^2+B_y^2)\\ &A_1=4(B_xC_x+B_yC_y)\\ &A_0=C_x^2+C_y^2. \end{aligned}$$

We apply the following parameter transformation:

$$t=t_0+(t_1-t_0)u$$

in order to be able to use the Gauss quadrature method. The transformation is applied to Eq. (1.73) to give:

$$\begin{aligned} l&=\int_{t_0}^{t_1} f(t)\, dt=(t_1-t_0)\int_0^1 f(t_0+(t_1-t_0)u)\, du=(t_1-t_0)\int_0^1 g(u)\, du\\ &\fallingdotseq(t_1-t_0)\sum_{i=1}^{m} w_i g(u_i). \end{aligned} \tag{1.74}$$

In this equation, m is the number of knots in Gauss' formula, the u_i are the knots and the w_i are weight factors.

Next, in the case of a plane curve, let us find the area of the "pie slice" formed by the origin and a curve segment (Fig. 1.24). The pie slice defined by

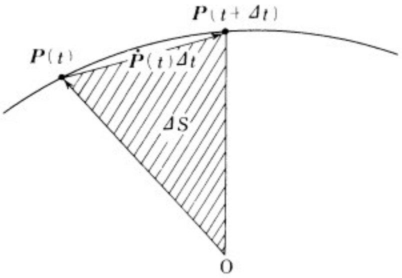

Fig. 1.24. Area enclosed by a curve

the origin and the two points on the curve $\boldsymbol{P}(t)$ and $\boldsymbol{P}(t+\Delta t)$ is approximated by the triangle $O-\boldsymbol{P}(t)-\boldsymbol{P}(t+\Delta t)$. Since this area ΔS is $1/2\,|\boldsymbol{P}(t)\times\dot{\boldsymbol{P}}(t)|\,\Delta t$, the area S swept out by the curve as the parameter varies from t_0 to t_1 is:

$$S=\frac{1}{2}\int_{t_0}^{t_1}|\boldsymbol{P}(t)\times\dot{\boldsymbol{P}}(t)|\,dt. \tag{1.75}$$

1.2.10 Intersection of a Curve with a Plane

A plane is represented by the equation:

$$ax+by+cz+d=0.$$

Since the curve $\boldsymbol{P}(t)=[x(t)\ \ y(t)\ \ z(t)]$ satisfies this equation, we have:

$$ax(t)+by(t)+cz(t)+d=0. \tag{1.76}$$

The roots of this equation can be found by the Newton-Raphson method. If the curve $\boldsymbol{P}(t)$ has been normalized in the range $0\leqq t\leqq 1$, the roots will be confined to this interval. The calculation can be made very efficient by pointing at the intersection points on the display with a stylus to obtain the first iterations.

1.2.11 Intersection of Two Curves

For simplicity let us restrict ourselves to the problem of finding intersections of two plane curves

$$\boldsymbol{P}_1(t)=[x_1(t)\ y_1(t)\ z_1(t)] \quad \text{and} \quad \boldsymbol{P}_2(u)=[x_2(u)\ y_2(u)\ z_2(u)].$$

In this case, we have:

$$\left.\begin{aligned} x_1(t)&=x_2(u)\\ y_1(t)&=y_2(u)\end{aligned}\right\}\qquad (0\leqq t\leqq 1,\ 0\leqq u\leqq 1). \tag{1.77}$$

Let us seek solutions of the simultaneous equations for the two unknowns t and u ($0 \leqq t \leqq 1$, $0 \leqq u \leqq 1$) by the Newton-Raphson method. The calculation can be made very efficient by pointing at the intersection points on the display with a stylus to obtain the first iterations.

If both curves are given by cubic polynomials of the form (1.43), then Eq. (1.77) becomes the following:

$$\left.\begin{aligned} A_{1x}t^3+B_{1x}t^2+C_{1x}t+D_{1x}-(A_{2x}u^3+B_{2x}u^2+C_{2x}u+D_{2x})=0 \\ A_{1y}t^3+B_{1y}t^2+C_{1y}t+D_{1y}-(A_{2y}u^3+B_{2y}u^2+C_{2y}u+D_{2y})=0 \end{aligned}\right\}. \tag{1.78}$$

1.3 Theory of Surfaces

1.3.1 Parametric Representation of Surfaces

Suppose that a curved surface is described by a polynomial in terms of the two parameters u and w which vary on its surface:

$$\boldsymbol{P}(u, w)=[x(u, w) \;\; y(u, w) \;\; z(u, w)]. \tag{1.79}$$

If a curved surface satisfies the conditions that the three Jacobians:

$$J_x \equiv \frac{\partial(y, z)}{\partial(u, w)}{}^{*)}, \qquad J_y \equiv \frac{\partial(z, x)}{\partial(u, w)}, \qquad J_z \equiv \frac{\partial(x, y)}{\partial(u, w)}$$

are not all zero at the same time, in other words, that:

$$J_x^2+J_y^2+J_z^2 \neq 0 \tag{1.80}$$

that the curved surface $\boldsymbol{P}(u, w)$ has all derivatives with respect to x, y and z up to r-th order, and, moreover, that all such derivatives are continuous, then the curved surface defined by Eq. (1.79) is said to be of class C^r.

Condition (1.80) is the condition for the curved surface to not degenerate to a point or a curve, and to not contain any singular points such as spikes. This condition requires constraints on both the curve itself and the parameters. Even for the same curve, it is possible that for one parameter expression condition (1.80) will be satisfied, and for another parameter expression it will not.**)

---

*) This notation is defined by

$$\frac{\partial(y, z)}{\partial(u, w)}=\begin{vmatrix} \dfrac{\partial y}{\partial u} & \dfrac{\partial z}{\partial u} \\[2ex] \dfrac{\partial y}{\partial w} & \dfrac{\partial z}{\partial w} \end{vmatrix}$$

**) For example, if the xy plane is expressed as $x=u$, $y=w$, $z=0$ then $J_z=1$, but if the same plane is expressed as $x=u^3$, $y=w^3$, $z=0$ then at the origin $J_x=J_y=J_z=0$

In the discussion that follows, unless otherwise stated it is assumed that condition (1.80) holds.

If $w = w_0$ (fixed), the curved surface representation describes a curve $\boldsymbol{P}(u, w_0)$ on the curved surface. This curve is called the u-curve. Similarly, the curve $\boldsymbol{P}(u_0, w)$ on the curved surface corresponding to $u = u_0$ (fixed) is called the w-curve.

At a point $\boldsymbol{P}(u_0, w_0)$ on the curved surface, the tangent vector in the direction of the u-curve is:

$$\boldsymbol{P}_u = \left[\frac{\partial x(u, w_0)}{\partial u} \quad \frac{\partial y(u, w_0)}{\partial u} \quad \frac{\partial z(u, w_0)}{\partial u}\right]_{u=u_0}.$$

Similarly, the tangent vector in the direction of the w-curve is:

$$\boldsymbol{P}_w = \left[\frac{\partial x(u_0, w)}{\partial w} \quad \frac{\partial y(u_0, w)}{\partial w} \quad \frac{\partial z(u_0, w)}{\partial w}\right]_{w=w_0}.$$

If these two vectors are drawn starting from the point $\boldsymbol{P}(u_0, w_0)$, if condition (1.80) is satisfied the angle between them is non-zero (Fig. 1.25). That is, condition (1.80) is equivalent to saying that the vectors $\boldsymbol{P}_u$ and $\boldsymbol{P}_w$ are linearly independent at each point on the curved surface. Condition (1.80) is also equivalent to $\boldsymbol{P}_u \times \boldsymbol{P}_w \neq \boldsymbol{0}$ *).

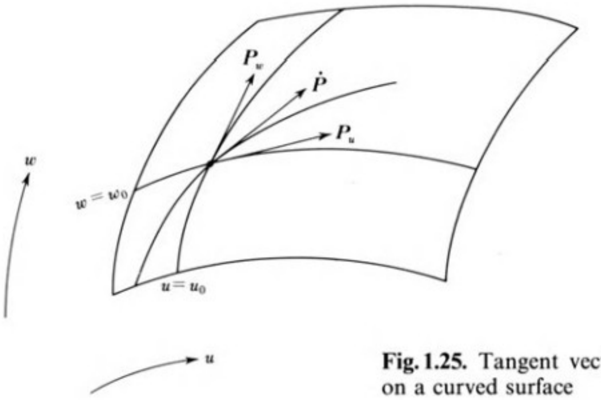

Fig. 1.25. Tangent vectors in two parameter directions on a curved surface

A curve in an arbitrary direction on a curved surface can be expressed in terms of a parameter t as $\boldsymbol{P}(u(t), w(t))$ $(a \leqq t \leqq b)$. Therefore, at point $\boldsymbol{P}(u_0, w_0)$, the tangent vector in the direction of that curve becomes:

$$\left.\frac{d\boldsymbol{P}}{dt}\right|_{t=t_0} = \boldsymbol{P}_u \left.\frac{du}{dt}\right|_{t=t_0} + \boldsymbol{P}_w \left.\frac{dw}{dt}\right|_{t=t_0}. \tag{1.81}$$

---

*) Note that $(\boldsymbol{P}_u \times \boldsymbol{P}_w)^2 = J_x^2 + J_y^2 + J_z^2$

Note that $u_0 = u(t_0)$ and $w_0 = w(t_0)$. The above formula expresses the fact that a tangent vector in an arbitrary direction on the curved surface lies on the plane formed by $\boldsymbol{P}_u$ and $\boldsymbol{P}_w$. This plane is called the tangent plane at the point $\boldsymbol{P}(u_0, w_0)$. Letting $\boldsymbol{R}$ be a position vector on the tangent plane, the tangent plane is given by the following equation.

$$[\boldsymbol{R} - \boldsymbol{P}(u_0, w_0),\ \boldsymbol{P}_u,\ \boldsymbol{P}_w] = 0. \tag{1.82}$$

The straight line that passes through point $\boldsymbol{P}(u_0, w_0)$ and is perpendicular to the tangent plane at that point is called the normal line. The normal vector $\boldsymbol{e}$ can be expressed in terms of $\boldsymbol{P}_u$ and $\boldsymbol{P}_w$ in a right-hand systems as follows:

$$\boldsymbol{e} = \frac{\boldsymbol{P}_u \times \boldsymbol{P}_w}{\sqrt{(\boldsymbol{P}_u \times \boldsymbol{P}_w)^2}} = \frac{\boldsymbol{P}_u \times \boldsymbol{P}_w}{\sqrt{J_x^2 + J_y^2 + J_z^2}}. \tag{1.83}$$

$\boldsymbol{e}$ is called the unit normal vector.

1.3.2 The First Fundamental Matrix of a Surface

As was stated in the preceding section, a curve on the curved surface $\boldsymbol{P}(u, w)$ can be given by $u = u(t)$, $w = w(t)$, $(a \leqq t \leqq b)$. The following notation is used for this.

$$\boldsymbol{u}(t) = [u(t)\ w(t)]. \tag{1.84}$$

From Eq. (1.81), the tangent vector to this curve is given by:

$$\begin{aligned}\dot{\boldsymbol{P}} &= \boldsymbol{P}_u \dot{u} + \boldsymbol{P}_w \dot{w} = [\dot{u}\ \ \dot{w}] \begin{bmatrix} \boldsymbol{P}_u \\ \boldsymbol{P}_w \end{bmatrix} \\ &\equiv \dot{\boldsymbol{u}} \boldsymbol{A}. \end{aligned} \tag{1.85}$$

Here $\boldsymbol{A}$ is defined by:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{P}_u \\ \boldsymbol{P}_w \end{bmatrix}. \tag{1.86}$$

The square of the magnitude $\dot{s}$ of the tangent vector is given by:

$$\begin{aligned} \dot{s}^2 &= \dot{\boldsymbol{P}}(t)^2 = \dot{\boldsymbol{P}}(t) \cdot \dot{\boldsymbol{P}}(t)^T \\ &= [\dot{u}\ \ \dot{w}] \begin{bmatrix} \boldsymbol{P}_u \\ \boldsymbol{P}_w \end{bmatrix} [\boldsymbol{P}_u\ \ \boldsymbol{P}_w] \begin{bmatrix} \dot{u} \\ \dot{w} \end{bmatrix} = [\dot{u}\ \ \dot{w}] \begin{bmatrix} \boldsymbol{P}_u^2 & \boldsymbol{P}_u \cdot \boldsymbol{P}_w \\ \boldsymbol{P}_u \cdot \boldsymbol{P}_w & \boldsymbol{P}_w^2 \end{bmatrix} \begin{bmatrix} \dot{u} \\ \dot{w} \end{bmatrix} \\ &\equiv \dot{\boldsymbol{u}} \boldsymbol{F} \dot{\boldsymbol{u}}^T. \end{aligned} \tag{1.87}$$

Here $\boldsymbol{F}$ is the matrix defined by:

$$\boldsymbol{F}=\begin{bmatrix} \boldsymbol{P}_u^2 & \boldsymbol{P}_u\cdot\boldsymbol{P}_w \\ \boldsymbol{P}_u\cdot\boldsymbol{P}_w & \boldsymbol{P}_w^2 \end{bmatrix}. \tag{1.88}$$

This matrix $\boldsymbol{F}$ is called the first fundamental matrix of the curved surface.

Using the first fundamental matrix, the unit tangent vector $\boldsymbol{t}$ of a curve on a curved surface is:

$$\boldsymbol{t}=\frac{\dot{\boldsymbol{P}}(t)}{|\dot{\boldsymbol{P}}(t)|}=\frac{\dot{\boldsymbol{u}}\boldsymbol{A}}{(\dot{\boldsymbol{u}}\boldsymbol{F}\dot{\boldsymbol{u}}^T)^{\frac{1}{2}}}. \tag{1.89}$$

Therefore, the cosine of the angle θ between the two curves $\boldsymbol{u}_1=\boldsymbol{u}_1(t)$ and $\boldsymbol{u}_2=\boldsymbol{u}_2(t)$ on the curved surface is, using (1.89):

$$\cos\theta=\boldsymbol{t}_1\cdot\boldsymbol{t}_2=\frac{\dot{\boldsymbol{u}}_1\boldsymbol{A}\cdot\boldsymbol{A}^T\dot{\boldsymbol{u}}_2^T}{(\dot{\boldsymbol{u}}_1\boldsymbol{F}\dot{\boldsymbol{u}}_1^T)^{\frac{1}{2}}(\dot{\boldsymbol{u}}_2\boldsymbol{F}\dot{\boldsymbol{u}}_2^T)^{\frac{1}{2}}}=\frac{\dot{\boldsymbol{u}}_1\boldsymbol{F}\dot{\boldsymbol{u}}_2^T}{(\dot{\boldsymbol{u}}_1\boldsymbol{F}\dot{\boldsymbol{u}}_1^T)^{\frac{1}{2}}(\dot{\boldsymbol{u}}_2\boldsymbol{F}\dot{\boldsymbol{u}}_2^T)^{\frac{1}{2}}}. \tag{1.90}$$

The length of a curve on the curved surface is given by:

$$l=\int_{t_0}^{t_1}\dot{s}\,dt=\int_{t_0}^{t_1}(\dot{\boldsymbol{u}}\boldsymbol{F}\dot{\boldsymbol{u}}^T)^{\frac{1}{2}}\,dt. \tag{1.91}$$

Next, let us consider a small area of a curved surface surrounded by the parametric curves $u=u_0$, $u=u_0+\Delta u$, $w=w_0$, $w=w_0+\Delta w$. Referring to Fig. 1.26, we see that the area ΔS of this small segment of surface is:

$$\Delta S \fallingdotseq |\boldsymbol{P}_u\times\boldsymbol{P}_w|\,\Delta u\,\Delta w.$$

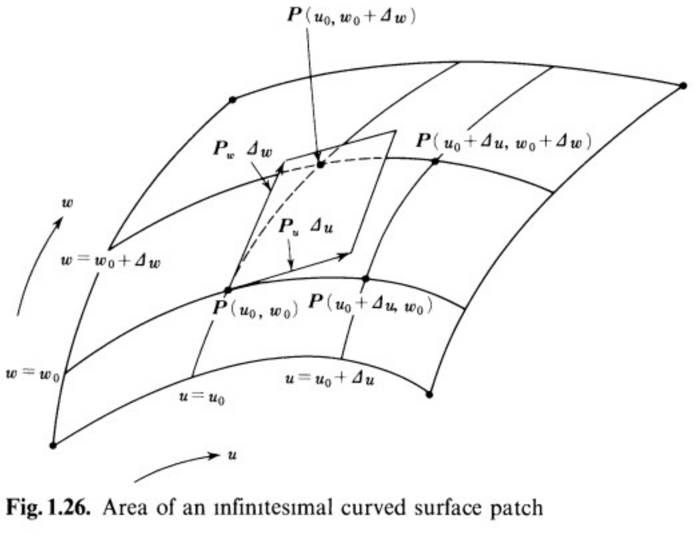

Fig. 1.26. Area of an infinitesimal curved surface patch

We also have:

$$|\boldsymbol{P}_u \times \boldsymbol{P}_w|^2 = \boldsymbol{P}_u^2 \boldsymbol{P}_w^2 - (\boldsymbol{P}_u \cdot \boldsymbol{P}_w)^{2\,*)} = |\boldsymbol{F}|. \tag{1.92}$$

Therefore, the curved surface area corresponding to the region R on the uw plane is:

$$S = \iint_R |\boldsymbol{F}|^{\frac{1}{2}}\, du dw. \tag{1.93}$$

1.3.3 Determining Conditions for a Tangent Vector to a Curve on a Surface

As shown in the preceding section, since the unit tangent vector of a curve on a curved surface is given by Eq. (1.89), the condition that determines the unit tangent vector $\boldsymbol{t}$ is:

$$\dot{\boldsymbol{u}} \boldsymbol{F} \dot{\boldsymbol{u}}^T > 0. \tag{1.94}$$

In the case $\boldsymbol{P}_u^2 \neq 0$, changing the form of the equation slightly gives:

$$\begin{aligned} \dot{\boldsymbol{u}} \boldsymbol{F} \dot{\boldsymbol{u}}^T &= [\dot{u}\ \dot{w}] \begin{bmatrix} \boldsymbol{P}_u^2 & \boldsymbol{P}_u \cdot \boldsymbol{P}_w \\ \boldsymbol{P}_u \cdot \boldsymbol{P}_w & \boldsymbol{P}_w^2 \end{bmatrix} \begin{bmatrix} \dot{u} \\ \dot{w} \end{bmatrix} = \boldsymbol{P}_u^2 \dot{u}^2 + 2\boldsymbol{P}_u \cdot \boldsymbol{P}_w \dot{u}\dot{w} + \boldsymbol{P}_w^2 \dot{w}^2 \\ &= \boldsymbol{P}_u^2 \left(\dot{u} + \frac{\boldsymbol{P}_u \cdot \boldsymbol{P}_w}{\boldsymbol{P}_u^2} \dot{w} \right)^2 + \frac{\boldsymbol{P}_u^2 \boldsymbol{P}_w^2 - (\boldsymbol{P}_u \cdot \boldsymbol{P}_w)^2}{\boldsymbol{P}_u^2} \dot{w}^2 \\ &= \boldsymbol{P}_u^2 \left(\dot{u} + \frac{\boldsymbol{P}_u \cdot \boldsymbol{P}_w}{\boldsymbol{P}_u^2} \dot{w} \right)^2 + \frac{|\boldsymbol{P}_u \times \boldsymbol{P}_w|^2}{\boldsymbol{P}_u^2} \dot{w}^2 > 0 \quad (\boldsymbol{P}_u^2 \neq 0). \end{aligned} \tag{1.95}$$

This equation implies that if the conditions

$$\dot{\boldsymbol{u}} \neq \boldsymbol{0} \quad \text{and} \quad \boldsymbol{P}_u \times \boldsymbol{P}_w \neq \boldsymbol{0} \tag{1.96}$$

are satisfied, the unit tangent vectors are determined for all curves $\boldsymbol{u} = \boldsymbol{u}(t)$. If the second condition is satisfied but the first is not, then, as can be seen from Eq. (1.81) and (1.83), the tangent plane and normal line are determined, but the unit tangent vector $\boldsymbol{t}$ is not determined due to twisting of the curve $\boldsymbol{u} = \boldsymbol{u}(t)$. If the first condition is satisfied but the second is not, there can be a cusp or the u-curve and the w-curve can be parallel, with the result that the tangent plane is undetermined[7].

---

*) Using the vector identity $(\boldsymbol{A} \times \boldsymbol{B})(\boldsymbol{C} \times \boldsymbol{D}) = (\boldsymbol{AC})(\boldsymbol{BD}) - (\boldsymbol{AD})(\boldsymbol{BC})$

1.3.4 Curvature of a Surface

From Eq. (1.9) we have:

$$\dot{\boldsymbol{P}} = \dot{s}\boldsymbol{P}' = \dot{s}\boldsymbol{t}.$$

Differentiating by t we obtain the equation:

$$\begin{aligned}\ddot{\boldsymbol{P}} &= \ddot{s}\boldsymbol{t} + \dot{s}\dot{\boldsymbol{t}} = \ddot{s}\boldsymbol{t} + \dot{s}^2\boldsymbol{t}' \\ &= \ddot{s}\boldsymbol{t} + \dot{s}^2\kappa\boldsymbol{n}.\end{aligned} \tag{1.97}$$

Now let us find $\ddot{\boldsymbol{P}}$ for the curve $\boldsymbol{u}(t) = [u(t)\ w(t)]$ on the curved surface $\boldsymbol{P}(u, w)$. From Eq. (1.81) we have:

$$\dot{\boldsymbol{P}} = \boldsymbol{P}_u\dot{u} + \boldsymbol{P}_w\dot{w}.$$

Therefore:

$$\begin{aligned}\ddot{\boldsymbol{P}} &= \frac{d\boldsymbol{P}_u}{dt}\dot{u} + \boldsymbol{P}_u\ddot{u} + \frac{d\boldsymbol{P}_w}{dt}\dot{w} + \boldsymbol{P}_w\ddot{w} \\ &= \boldsymbol{P}_{uu}\dot{u}^2 + \boldsymbol{P}_{uw}\dot{u}\dot{w} + \boldsymbol{P}_u\ddot{u} + \boldsymbol{P}_{ww}\dot{w}^2 + \boldsymbol{P}_{wu}\dot{w}\dot{u} + \boldsymbol{P}_w\ddot{w} \\ &= \boldsymbol{P}_{uu}\dot{u}^2 + \boldsymbol{P}_{uw}\dot{u}\dot{w} + \boldsymbol{P}_{wu}\dot{u}\dot{w} + \boldsymbol{P}_{ww}\dot{w}^2 + \boldsymbol{P}_u\ddot{u} + \boldsymbol{P}_w\ddot{w}.\end{aligned} \tag{1.98}$$

Taking the inner product of the unit normal vector $\boldsymbol{e}$ to the curved surface and $\ddot{\boldsymbol{P}}$ from Eq. (1.97), since $\boldsymbol{e}$ and $\boldsymbol{t}$ are perpendicular with each other:

$$\boldsymbol{e} \cdot \ddot{\boldsymbol{P}} = \dot{s}^2\kappa\boldsymbol{n} \cdot \boldsymbol{e}. \tag{1.99}$$

Also taking the inner product of the unit normal vector $\boldsymbol{e}$ and $\ddot{\boldsymbol{P}}$ from Eq. (1.98), since $\boldsymbol{e}$ is perpendicular to both $\boldsymbol{P}_u$ and $\boldsymbol{P}_w$:

$$\begin{aligned}\boldsymbol{e} \cdot \ddot{\boldsymbol{P}} &= \boldsymbol{e} \cdot \boldsymbol{P}_{uu}\dot{u}^2 + \boldsymbol{e} \cdot \boldsymbol{P}_{uw}\dot{u}\dot{w} + \boldsymbol{e} \cdot \boldsymbol{P}_{wu}\dot{u}\dot{w} + \boldsymbol{e} \cdot \boldsymbol{P}_{ww}\dot{w}^2 \\ &= [\dot{u}\ \dot{w}]\begin{bmatrix}\boldsymbol{e} \cdot \boldsymbol{P}_{uu} & \boldsymbol{e} \cdot \boldsymbol{P}_{uw} \\ \boldsymbol{e} \cdot \boldsymbol{P}_{wu} & \boldsymbol{e} \cdot \boldsymbol{P}_{ww}\end{bmatrix}\begin{bmatrix}\dot{u} \\ \dot{w}\end{bmatrix} \\ &\equiv \dot{\boldsymbol{u}}\boldsymbol{G}\dot{\boldsymbol{u}}^T\end{aligned} \tag{1.100}$$

where:

$$\boldsymbol{G} = \begin{bmatrix}\boldsymbol{e} \cdot \boldsymbol{P}_{uu} & \boldsymbol{e} \cdot \boldsymbol{P}_{uw} \\ \boldsymbol{e} \cdot \boldsymbol{P}_{wu} & \boldsymbol{e} \cdot \boldsymbol{P}_{ww}\end{bmatrix}. \tag{1.101}$$

From Eq. (1.99) and (1.100) we obtain:

$$\dot{s}^2\kappa\boldsymbol{n} \cdot \boldsymbol{e} = \dot{\boldsymbol{u}}\boldsymbol{G}\dot{\boldsymbol{u}}^T. \tag{1.102}$$

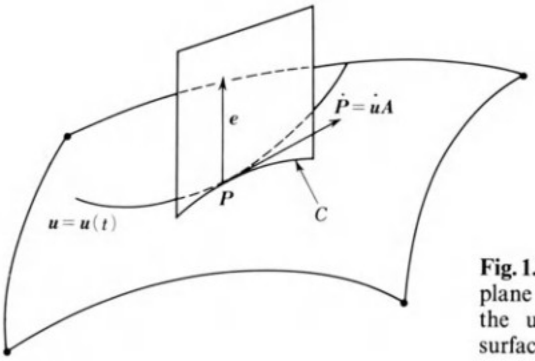

Fig. 1.27. The intersection curve C of the plane containing the tangent vector $\dot{\boldsymbol{P}}$ and the unit normal vector $\boldsymbol{e}$ with a curved surface

The matrix $\boldsymbol{G}$ is called the second fundamental matrix of the curved surface. Normally, in the curved surface representation encountered in CAD, it often happens that $\boldsymbol{P}_{uw}=\boldsymbol{P}_{wu}$. In such a case, $\boldsymbol{G}$ is a symmetrical matrix.

Let a space curve $\boldsymbol{u}=\boldsymbol{u}(t)$ be given on the curved surface $\boldsymbol{P}(u, w)$ and let C be the curve formed by the intersection of the curved surface with the plane which includes the tangent vector $\dot{\boldsymbol{P}}=\dot{\boldsymbol{u}}\boldsymbol{A}$ at a point P on the curve and the unit normal vector $\boldsymbol{e}$ (refer to Fig. 1.27). The curvature of the curve C is called the normal curvature relative to the direction $\dot{\boldsymbol{u}}\boldsymbol{A}$ at point P. The normal curvature is the projected length of the curvature vector of the curve $\boldsymbol{u}$ to $\boldsymbol{e}$. Letting κ_n be the normal curvature, from Eq. (1.102) we have:

$$\dot{s}^2\kappa_n=\dot{\boldsymbol{u}}\boldsymbol{G}\dot{\boldsymbol{u}}^T$$

$$\therefore\ \kappa_n=\frac{\dot{\boldsymbol{u}}\boldsymbol{G}\dot{\boldsymbol{u}}^T}{\dot{s}^2}=\frac{\dot{\boldsymbol{u}}\boldsymbol{G}\dot{\boldsymbol{u}}^T}{\dot{\boldsymbol{u}}\boldsymbol{F}\dot{\boldsymbol{u}}^T}. \tag{1.103}$$

The sign of κ_n is plus if the curve C is concave in the direction of $\boldsymbol{e}$, negative if it is convex. As shown in Fig. 1.27, $\kappa_n<0$ at point P.

In general, if the direction of the intersection curve at point P changes, the curvature κ_n will also change. The direction in which κ_n takes an extreme value is called the principal direction of the normal curvature. In Eq. (1.103), if we take:

$$\xi=\dot{u},\quad \eta=\dot{w},\quad L=\boldsymbol{e}\cdot\boldsymbol{P}_{uu},\quad M=\boldsymbol{e}\cdot\boldsymbol{P}_{uw}=\boldsymbol{e}\cdot\boldsymbol{P}_{wu},$$
$$N=\boldsymbol{e}\cdot\boldsymbol{P}_{ww},\quad E=\boldsymbol{P}_u^2,\quad F=\boldsymbol{P}_u\cdot\boldsymbol{P}_w,\quad G=\boldsymbol{P}_w^2$$

then:

$$L\xi^2+2M\xi\eta+N\eta^2-\kappa_n(E\xi^2+2F\xi\eta+G\eta^2)=0. \tag{1.104}$$

To find the extreme value of κ_n, take the partial derivatives with respect to ξ and η and set $\partial\kappa_n/\partial\xi=0$ and $\partial\kappa_n/\partial\eta=0$ to obtain:

$$\left.\begin{array}{l}(L-\kappa_nE)\xi+(M-\kappa_nF)\eta=0\\(M-\kappa_nF)\xi+(N-\kappa_nG)\eta=0\end{array}\right\}. \tag{1.105}$$

Eliminating ξ and η from these equations gives:

$$(EG-F^2)\kappa_n^2-(EN+GL-2FM)\kappa_n+LN-M^2=0. \tag{1.106}$$

This quadratic equation always has two real solutions $\kappa_{n\max}$ and $\kappa_{n\min}$. $\kappa_{n\max}$ and $\kappa_{n\min}$ are the maximum and minimum values of the normal curvature, called the principal curvatures.

From the relation between the roots and the coefficients we have:

$$\begin{aligned} K \equiv \kappa_{n\max}\kappa_{n\min} &= \frac{LN-M^2}{EG-F^2} = \frac{(\boldsymbol{e}\cdot\boldsymbol{P}_{uu})(\boldsymbol{e}\cdot\boldsymbol{P}_{ww})-(\boldsymbol{e}\cdot\boldsymbol{P}_{uw})^2}{\boldsymbol{P}_u^2\boldsymbol{P}_w^2-(\boldsymbol{P}_u\cdot\boldsymbol{P}_w)} \\ &= \frac{|\boldsymbol{G}|}{|\boldsymbol{F}|} \end{aligned} \tag{1.107}$$

$$\begin{aligned} 2H \equiv \kappa_{n\min}+\kappa_{n\max} &= \frac{EN+GL-2FM}{EG-F^2} \\ &= \frac{\boldsymbol{P}_u^2(\boldsymbol{e}\cdot\boldsymbol{P}_{ww})-2(\boldsymbol{P}_u\cdot\boldsymbol{P}_w)(\boldsymbol{e}\cdot\boldsymbol{P}_{uw})+\boldsymbol{P}_w^2(\boldsymbol{e}\cdot\boldsymbol{P}_{uu})}{|\boldsymbol{F}|}. \end{aligned} \tag{1.108}$$

K is called the total curvature or the Gauss curvature, while H is called the mean curvature. It is easy to show that the two directions in which the principal curvatures are obtained are perpendicular to each other.

1.3.5 Calculation of a Point on a Surface

Many of the curved surface patch equations which are normally used are of the bi-cubic surface type:

$$\boldsymbol{P}(u,w)=\sum_{i=0}^{3}\sum_{j=0}^{3}\boldsymbol{A}_{3-i,3-j}u^i w^j \tag{1.109}$$

$$\begin{aligned} &=(\boldsymbol{A}_{00}w^3+\boldsymbol{A}_{01}w^2+\boldsymbol{A}_{02}w+\boldsymbol{A}_{03})u^3+(\boldsymbol{A}_{10}w^3+\boldsymbol{A}_{11}w^2+\boldsymbol{A}_{12}w+\boldsymbol{A}_{13})u^2 \\ &\quad+(\boldsymbol{A}_{20}w^3+\boldsymbol{A}_{21}w^2+\boldsymbol{A}_{22}w+\boldsymbol{A}_{23})u+\boldsymbol{A}_{30}w^3+\boldsymbol{A}_{31}w^2+\boldsymbol{A}_{32}w+\boldsymbol{A}_{33} \end{aligned} \tag{1.110}$$

$$=[u^3\ u^2\ u\ 1]\begin{bmatrix}\boldsymbol{A}_{00} & \boldsymbol{A}_{01} & \boldsymbol{A}_{02} & \boldsymbol{A}_{03}\\ \boldsymbol{A}_{10} & \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \boldsymbol{A}_{13}\\ \boldsymbol{A}_{20} & \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \boldsymbol{A}_{23}\\ \boldsymbol{A}_{30} & \boldsymbol{A}_{31} & \boldsymbol{A}_{32} & \boldsymbol{A}_{33}\end{bmatrix}\begin{bmatrix}w^3\\ w^2\\ w\\ 1\end{bmatrix}. \tag{1.111}$$

For a curved surface of this type, just as in the case of a curve, a point on a curved surface can be calculated rapidly by finite difference calculation.

Taking w to be fixed in Eq. (1.110), we have:

$$\boldsymbol{P}(u, w) = \boldsymbol{A}u^3 + \boldsymbol{B}u^2 + \boldsymbol{C}u + \boldsymbol{D}$$

where:

$$\begin{aligned}
\boldsymbol{A} &= \boldsymbol{A}_{00}w^3 + \boldsymbol{A}_{01}w^2 + \boldsymbol{A}_{02}w + \boldsymbol{A}_{03}\\
\boldsymbol{B} &= \boldsymbol{A}_{10}w^3 + \boldsymbol{A}_{11}w^2 + \boldsymbol{A}_{12}w + \boldsymbol{A}_{13}\\
\boldsymbol{C} &= \boldsymbol{A}_{20}w^3 + \boldsymbol{A}_{21}w^2 + \boldsymbol{A}_{22}w + \boldsymbol{A}_{23}\\
\boldsymbol{D} &= \boldsymbol{A}_{30}w^3 + \boldsymbol{A}_{31}w^2 + \boldsymbol{A}_{32}w + \boldsymbol{A}_{33}
\end{aligned}$$

which is the equation of a cubic curve. Then, using the finite difference matrix in Eq. (1.48), a point on a curve on the curved surface for which w is fixed can be found.

Letting δ be the finite difference increment in the u-direction and denoting a forward finite difference by Δ, the finite difference matrix for $u=0$ is, from Eq. (1.48):

$$\begin{bmatrix} \boldsymbol{P}_{0,w} \\ \frac{1}{\delta}\Delta \boldsymbol{P}_{0,w} \\ \frac{1}{\delta^2}\Delta^2 \boldsymbol{P}_{0,w} \\ \frac{1}{\delta^3}\Delta^3 \boldsymbol{P}_{0,w} \end{bmatrix} = \begin{bmatrix} \boldsymbol{D} \\ \boldsymbol{A}\delta^2 + \boldsymbol{B}\delta + \boldsymbol{C} \\ 6\boldsymbol{A}\delta + 2\boldsymbol{B} \\ 6\boldsymbol{A} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{30}w^3 + \boldsymbol{A}_{31}w^2 + \boldsymbol{A}_{32}w + \boldsymbol{A}_{33} \\ (\delta^2\boldsymbol{A}_{00} + \delta\boldsymbol{A}_{10} + \boldsymbol{A}_{20})w^3 + (\delta^2\boldsymbol{A}_{01} + \delta\boldsymbol{A}_{11} + \boldsymbol{A}_{21})w^2 \\ \quad + (\delta^2\boldsymbol{A}_{02} + \delta\boldsymbol{A}_{12} + \boldsymbol{A}_{22})w + \delta^2\boldsymbol{A}_{03} + \delta\boldsymbol{A}_{13} + \boldsymbol{A}_{23} \\ (6\delta\boldsymbol{A}_{00} + 2\boldsymbol{A}_{10})w^3 + (6\delta\boldsymbol{A}_{01} + 2\boldsymbol{A}_{11})w^2 \\ \quad + (6\delta\boldsymbol{A}_{02} + 2\boldsymbol{A}_{12})w + 6\delta\boldsymbol{A}_{03} + 2\boldsymbol{A}_{13} \\ 6\boldsymbol{A}_{00}w^3 + 6\boldsymbol{A}_{01}w^2 + 6\boldsymbol{A}_{02}w + 6\boldsymbol{A}_{03} \end{bmatrix} \tag{1.112}$$

Therefore, the finite difference matrix for $w=0$ is:

$$\begin{bmatrix} \boldsymbol{P}_{0,0} \\ \frac{1}{\delta}\Delta \boldsymbol{P}_{0,0} \\ \frac{1}{\delta^2}\Delta^2 \boldsymbol{P}_{0,0} \\ \frac{1}{\delta^3}\Delta^3 \boldsymbol{P}_{0,0} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{33} \\ \delta^2\boldsymbol{A}_{03} + \delta\boldsymbol{A}_{13} + \boldsymbol{A}_{23} \\ 6\delta\boldsymbol{A}_{03} + 2\boldsymbol{A}_{13} \\ 6\boldsymbol{A}_{03} \end{bmatrix}. \tag{1.113}$$

Next, in Eq. (1.112), let us take the finite difference in the w-direction at $w=0$. Letting δ also be the finite difference increment in this direction and denoting the forward finite difference in this direction by ∇, then the first order finite differences are:

$$\begin{bmatrix} \frac{1}{\delta}\nabla \boldsymbol{P}_{0,0} \\ \frac{1}{\delta^2}\nabla(\Delta \boldsymbol{P}_{0,0}) \\ \frac{1}{\delta^3}\nabla(\Delta^2 \boldsymbol{P}_{0,0}) \\ \frac{1}{\delta^4}\nabla(\Delta^3 \boldsymbol{P}_{0,0}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{30}\delta^2+\boldsymbol{A}_{31}\delta+\boldsymbol{A}_{32} \\ \begin{matrix}(\delta^2\boldsymbol{A}_{00}+\delta\boldsymbol{A}_{10}+\boldsymbol{A}_{20})\delta^2+(\delta^2\boldsymbol{A}_{01}+\delta\boldsymbol{A}_{11}+\boldsymbol{A}_{21})\delta \\ +\delta^2\boldsymbol{A}_{02}+\delta\boldsymbol{A}_{12}+\boldsymbol{A}_{22}\end{matrix} \\ \begin{matrix}(6\,\delta\boldsymbol{A}_{00}+2\boldsymbol{A}_{10})\delta^2+(6\,\delta\boldsymbol{A}_{01}+2\boldsymbol{A}_{11})\delta \\ +6\,\delta\boldsymbol{A}_{02}+2\boldsymbol{A}_{12}\end{matrix} \\ 6\,\boldsymbol{A}_{00}\delta^2+6\,\boldsymbol{A}_{01}\delta+6\,\boldsymbol{A}_{02} \end{bmatrix}. \tag{1.114}$$

The second and third order finite differences are:

$$\begin{bmatrix} \frac{1}{\delta^2}\nabla^2 \boldsymbol{P}_{0,0} \\ \frac{1}{\delta^3}\nabla^2(\Delta \boldsymbol{P}_{0,0}) \\ \frac{1}{\delta^4}\nabla^2(\Delta^2 \boldsymbol{P}_{0,0}) \\ \frac{1}{\delta^5}\nabla^2(\Delta^3 \boldsymbol{P}_{0,0}) \end{bmatrix} = \begin{bmatrix} 6\,\boldsymbol{A}_{30}\delta+2\boldsymbol{A}_{31} \\ 6(\delta^2\boldsymbol{A}_{00}+\delta\boldsymbol{A}_{10}+\boldsymbol{A}_{20})\delta+2(\delta^2\boldsymbol{A}_{01}+\delta\boldsymbol{A}_{11}+\boldsymbol{A}_{21}) \\ 6(6\delta\boldsymbol{A}_{00}+2\boldsymbol{A}_{10})\delta+2(6\delta\boldsymbol{A}_{01}+2\boldsymbol{A}_{11}) \\ 36\,\boldsymbol{A}_{00}\delta+12\boldsymbol{A}_{01} \end{bmatrix} \tag{1.115}$$

$$\begin{bmatrix} \frac{1}{\delta^3}\nabla^3 \boldsymbol{P}_{0,0} \\ \frac{1}{\delta^4}\nabla^3(\Delta \boldsymbol{P}_{0,0}) \\ \frac{1}{\delta^5}\nabla^3(\Delta^2 \boldsymbol{P}_{0,0}) \\ \frac{1}{\delta^6}\nabla^3(\Delta^3 \boldsymbol{P}_{0,0}) \end{bmatrix} = \begin{bmatrix} 6\boldsymbol{A}_{30} \\ 6(\delta^2\boldsymbol{A}_{00}+\delta\boldsymbol{A}_{10}+\boldsymbol{A}_{20}) \\ 6(6\delta\boldsymbol{A}_{00}+2\boldsymbol{A}_{10}) \\ 36\,\boldsymbol{A}_{00} \end{bmatrix}. \tag{1.116}$$

Let us name the 4×3 matrices on the right-hand sides of Eqs. (1.113), (1.114), (1.115) and (1.116), respectively as finite difference matrix A, finite difference matrix B, finite difference matrix C and finite difference matrix D (Fig. 1.28).

Calculations on the curve $\boldsymbol{P}(u, 0)$ in the u-direction at $w=0$ can be carried out by finite differences using matrix A, that is, Eq. (1.113). Next, to find the finite difference matrix to be used to calculate the curve $\boldsymbol{P}(u, \delta)$, as shown in Fig. (1.28) perform finite difference calculations for each element corresponding to matrices A, B, C and D. Therefore, to generate a family of curves extending along the u-direction separated by a parametric distance γ ($\delta=1/2^m$,

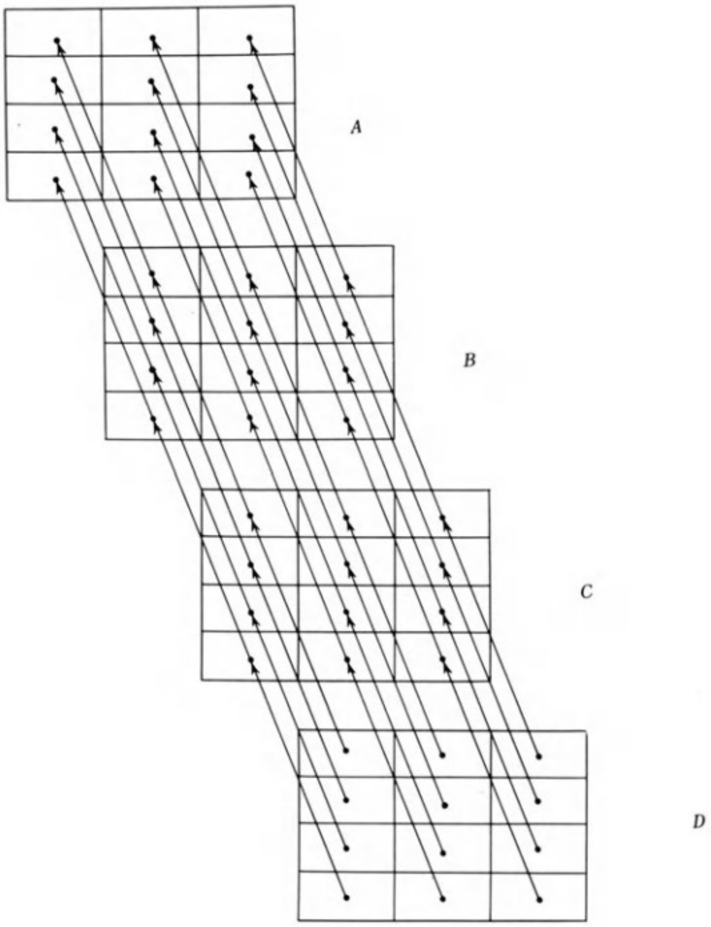

Fig. 1.28. Finite difference computations between elements of finite difference matrices A, B, C, D

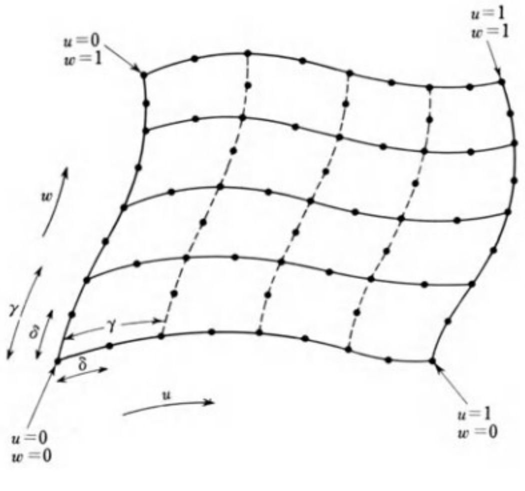
u=0
w=1
u=1
w=1
w
γ
δ
γ
δ
u
u=1
w=0
u=0
w=0

Fig. 1.29. Generation of points on a curved surface by finite difference computations

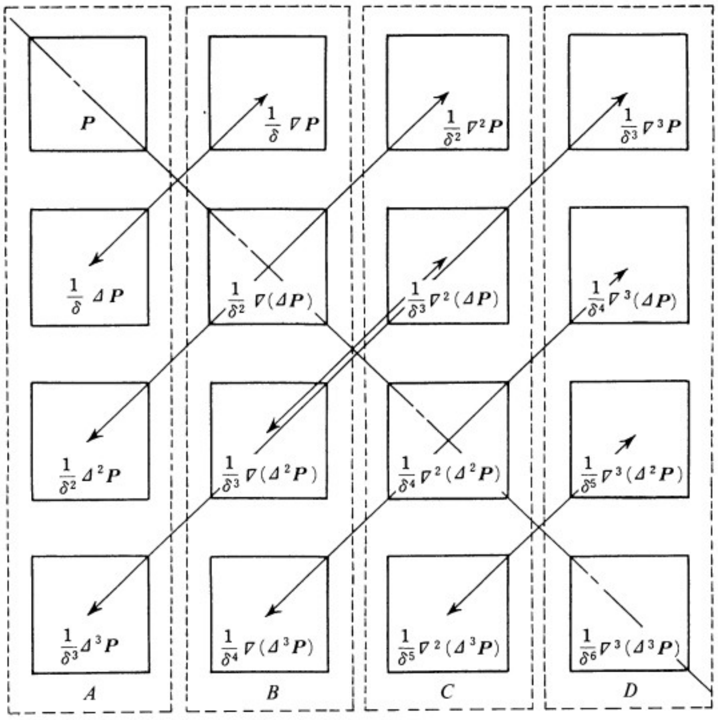
P
$\frac{1}{\delta}\nabla P$
$\frac{1}{\delta^2}\nabla^2 P$
$\frac{1}{\delta^3}\nabla^3 P$
$\frac{1}{\delta}\Delta P$
$\frac{1}{\delta^2}\nabla(\Delta P)$
$\frac{1}{\delta^3}\nabla^2(\Delta P)$
$\frac{1}{\delta^4}\nabla^3(\Delta P)$
$\frac{1}{\delta^2}\Delta^2 P$
$\frac{1}{\delta^3}\nabla(\Delta^2 P)$
$\frac{1}{\delta^4}\nabla^2(\Delta^2 P)$
$\frac{1}{\delta^5}\nabla^3(\Delta^2 P)$
$\frac{1}{\delta^3}\Delta^3 P$
$\frac{1}{\delta^4}\nabla(\Delta^3 P)$
$\frac{1}{\delta^5}\nabla^2(\Delta^3 P)$
$\frac{1}{\delta^6}\nabla^3(\Delta^3 P)$
A
B
C
D

Fig. 1.30. Vector matrix transposition of finite difference matrices A, B, C, D

$\gamma = 1/2^n$, $m \geqq n$) in the w-direction, it is sufficient to perform 1 curve-generating finite difference operation in the u-direction for every 2^{m-n} finite difference operations between matrices (Fig. 1.29).

Next, to find a family of curves extending in the w-direction, as shown in Fig. 1.30 matrix transpositions are performed using the vectors of finite difference matrices A, B, C and D. In other words, the process described above can be performed with the order of u and w reversed; this corresponds to the transformations $A_{ij} \leftrightarrow A_{ji}$ in the finite difference matrices.

1.3.6 Subdivision of Surface Patches

Let us now divide up a curved surface patch, using a bi-cubic surface described by Eq. (1.111) (Fig. 1.31).

First, let us find equations to describe the surface patch $\boldsymbol{P}(u^*, w^*)$ $(0 \leqq u^* \leqq 1;\ 0 \leqq w^* \leqq 1)$ within the bi-cubic surface patch $\boldsymbol{P}(u, w)$ surrounded by $u = u_0$, $u = u_1$, $w = w_0$ and $w = w_1$. To do this the following parameter transformation must be performed in Eq. (1.111).

$$u = u_0 + u^*(u_1 - u_0) \tag{1.117}$$

$$w = w_0 + w^*(w_1 - w_0). \tag{1.118}$$

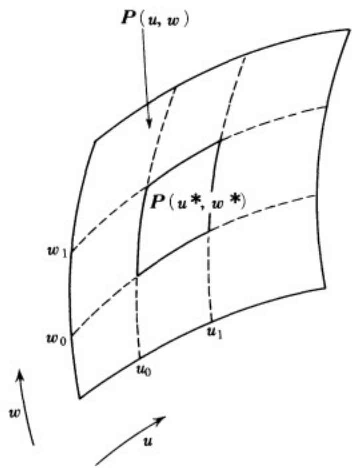

Fig. 1.31. Division of a surface patch (1)

Substituting (1.117) and (1.118) into (1.111), and using the result for parametric transformation of a curve (Eq. (1.55)), gives:

$$\boldsymbol{P}(u^*, w^*) = [u^{*3}\ u^{*2}\ u^*\ 1]$$

$$\times\begin{bmatrix}(u_1-u_0)^3 & 0 & 0 & 0\\ 3(u_1-u_0)^2u_0 & (u_1-u_0)^2 & 0 & 0\\ 3(u_1-u_0)u_0^2 & 2(u_1-u_0)u_0 & u_1-u_0 & 0\\ u_0^3 & u_0^2 & u_0 & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{A}_{00} & \boldsymbol{A}_{01} & \boldsymbol{A}_{02} & \boldsymbol{A}_{03}\\ \boldsymbol{A}_{10} & \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \boldsymbol{A}_{13}\\ \boldsymbol{A}_{20} & \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \boldsymbol{A}_{23}\\ \boldsymbol{A}_{30} & \boldsymbol{A}_{31} & \boldsymbol{A}_{32} & \boldsymbol{A}_{33}\end{bmatrix}$$

$$\times\begin{bmatrix}(w_1-w_0)^3 & 3(w_1-w_0)^2w_0 & 3(w_1-w_0)w_0^2 & w_0^3\\ 0 & (w_1-w_0)^2 & 2(w_1-w_0)w_0 & w_0^2\\ 0 & 0 & w_1-w_0 & w_0\\ 0 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix}w^{*3}\\ w^{*2}\\ w^*\\ 1\end{bmatrix}$$

$$(0 \leqq u^* \leqq 1;\quad 0 \leqq w^* \leqq 1). \tag{1.119}$$

Using Eq. (1.119), the curved surface formula for the case in which the surface patch $\boldsymbol{P}(u, w)$ in Fig. 1.32 is divided into 3 sections is obtained as follows:

- Formula for curved surface section A:
  in Eq. (1.119), set $u_0 = u_s$, $u_1 = 1$, $w_0 = 0$, $w_1 = 1$.
- Formula for curved surface section B:
  in Eq. (1.119), set $u_0 = 0$, $u_1 = u_s$, $w_0 = 0$, $w_1 = w_s$.
- Formula for curved surface section C:
  in Eq. (1.119), set $u_0 = 0$, $u_1 = u_s$, $w_0 = w_s$, $w_1 = 1$.

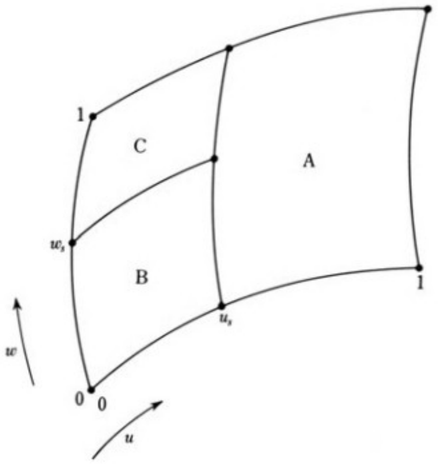
1
C
A
w_s
B
1
u_s
w
0
0
u

Fig. 1.32. Division of a surface patch (2)

1.3.7 Connection of Surface Patches

Let us consider the problem of connecting surface patch $\boldsymbol{P}_{\text{II}}(u, w)$ along its curve $u=0$ continuously to surface patch $\boldsymbol{P}_{\text{I}}(u, w)$ along its curve $u=1$ (Fig. 1.33).

First, the condition for the boundary curves of both patches to coincide is:

$$\boldsymbol{P}_{\text{II}}(0, w)=\boldsymbol{P}_{\text{I}}(1, w) \quad (0 \leqq w \leqq 1). \tag{1.120}$$

Next, on the boundary curve between both patches, in order for the slope to be continuous in the direction across the boundary curve it is necessary for

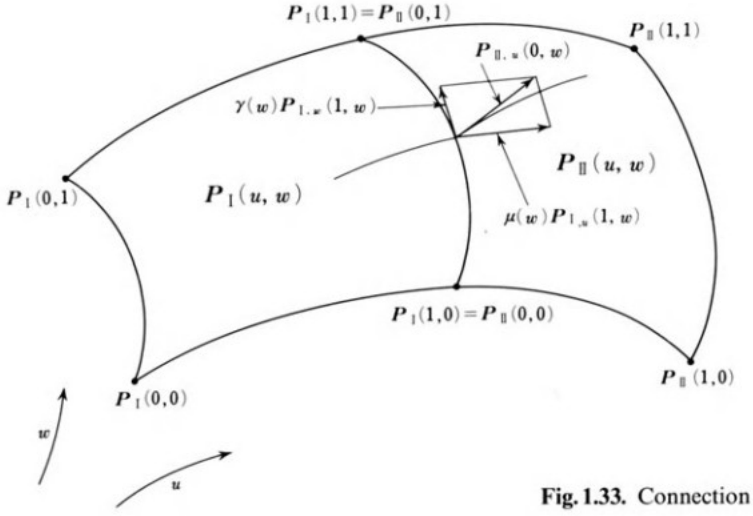
$\boldsymbol{P}_{\text{I}}(1,1)=\boldsymbol{P}_{\text{II}}(0,1)$
$\boldsymbol{P}_{\text{II}}(1,1)$
$\boldsymbol{P}_{\text{II},u}(0, w)$
$\gamma(w)\boldsymbol{P}_{\text{I},w}(1, w)$
$\boldsymbol{P}_{\text{II}}(u, w)$
$\boldsymbol{P}_{\text{I}}(0,1)$
$\boldsymbol{P}_{\text{I}}(u, w)$
$\mu(w)\boldsymbol{P}_{\text{I},u}(1, w)$
$\boldsymbol{P}_{\text{I}}(1,0)=\boldsymbol{P}_{\text{II}}(0,0)$
$\boldsymbol{P}_{\text{II}}(1,0)$
$\boldsymbol{P}_{\text{I}}(0,0)$
w
u

Fig. 1.33. Connection of surface patches

the tangent plane $\boldsymbol{P}_{\mathrm{II}}(0, w)$ $(0 \leqq w \leqq 1)$ to coincide with the tangent plane $\boldsymbol{P}_{\mathrm{I}}(1, w)$ $(0 \leqq w \leqq 1)$. In this case, the direction of the normal vector is continuous; that is, taking $\mu(w)$ to be a positive scalar function, we have:

$$\boldsymbol{P}_{\mathrm{II},u}(0, w) \times \boldsymbol{P}_{\mathrm{II},w}(0, w) = \mu(w)\, \boldsymbol{P}_{\mathrm{I},u}(1, w) \times \boldsymbol{P}_{\mathrm{I},w}(1, w) \; (0 \leqq w \leqq 1). \tag{1.121}$$

Since $\boldsymbol{P}_{\mathrm{II},w}(0, w) = \boldsymbol{P}_{\mathrm{I},w}(1, w)$, if $\gamma(w)$ is an arbitrary scalar function we have:

$$\boldsymbol{P}_{\mathrm{II},u}(0, w) = \mu(w)\, \boldsymbol{P}_{\mathrm{I},u}(1, w) + \gamma(w)\, \boldsymbol{P}_{\mathrm{I},w}(1, w) \;\; (0 \leqq w \leqq 1). \tag{1.122}$$

The implication of Eq. (1.122) is that $\boldsymbol{P}_{\mathrm{II},u}(0, w)$ of the second surface patch lies on the tangent plane of the first surface patch along the boundary curve.

Taking $\gamma(w) = 0$ as a special case, we obtain:

$$\boldsymbol{P}_{\mathrm{II},u}(0, w) = \mu(w)\, \boldsymbol{P}_{\mathrm{I},u}(1, w) \;\; (0 \leqq w \leqq 1). \tag{1.123}$$

In this case, the u-curve is smoothly connected across the boundary curve between the two surface patches (Fig. 1.34). This is a slope continuity condition and is a more severe condition than in the case of Eq. (1.122).

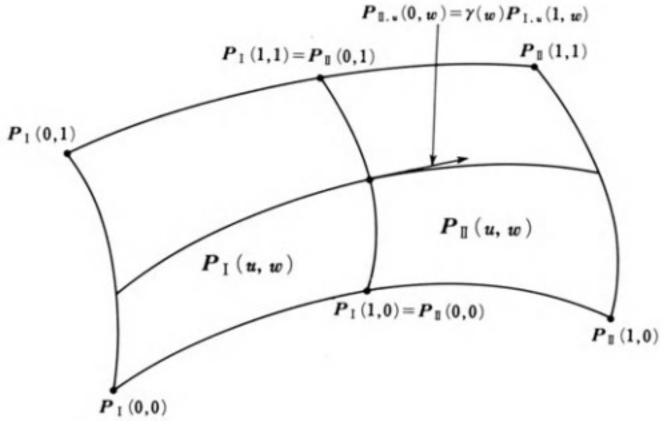
$P_{\mathrm{II},u}(0, w) = \gamma(w) P_{\mathrm{I},u}(1, w)$
$P_{\mathrm{I}}(1,1) = P_{\mathrm{II}}(0,1)$
$P_{\mathrm{II}}(1,1)$
$P_{\mathrm{I}}(0,1)$
$P_{\mathrm{II}}(u, w)$
$P_{\mathrm{I}}(u, w)$
$P_{\mathrm{I}}(1,0) = P_{\mathrm{II}}(0,0)$
$P_{\mathrm{II}}(1,0)$
$P_{\mathrm{I}}(0,0)$

Fig. 1.34. Special case of connection by slope continuity (w-curves are continuous)

The condition expressed by Eq. (1.120) requires that on the boundary curve, the degree of w on $\boldsymbol{P}_{\mathrm{II}}(u, w)$ must agree with the degree of w on $\boldsymbol{P}_{\mathrm{I}}(u, w)$. This does not constrain the degree of w over all of the patch $\boldsymbol{P}_{\mathrm{II}}(u, w)$. The degree of w on $\boldsymbol{P}_{\mathrm{II}}(u, w)$ can be specified to have an arbitrary value n not lower than the degree of w on $\boldsymbol{P}_{\mathrm{I}}(u, w)$ as long as it agrees with the degree on $\boldsymbol{P}_{\mathrm{I}}(u, w)$ on the boundary curve.

In addition, surface patches can be connected as shown in Fig. 1.35 by means of a suitable parametric transformation.

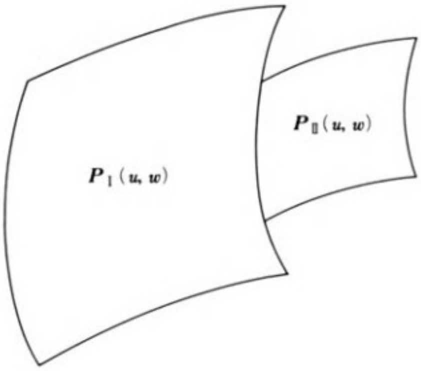

Fig. 1.35. Special case of patch connection

1.3.8 Degeneration of a Surface Patch

It is possible for 2 coincident vertices of the 4 corners of a surface patch to form a triangular surface patch (Fig. 1.36). Curve $\boldsymbol{P}(0, w)$ of this surface patch becomes a single point $\boldsymbol{D}$ regardless of the value which w takes from 0 to 1. In other words, the boundary curve degenerates to a point. When performing shape processing, for example in NC processing, when, for example, a tool path is calculated it is necessary to find the normal vector at every point on the surface. Normally, a normal vector on a surface is found as the vector product of tangent vectors in two parametric directions. However, at point $\boldsymbol{D}$ the vector product becomes:

$$\boldsymbol{P}_u(0, w) \times \boldsymbol{P}_w(0, w). \tag{1.124}$$

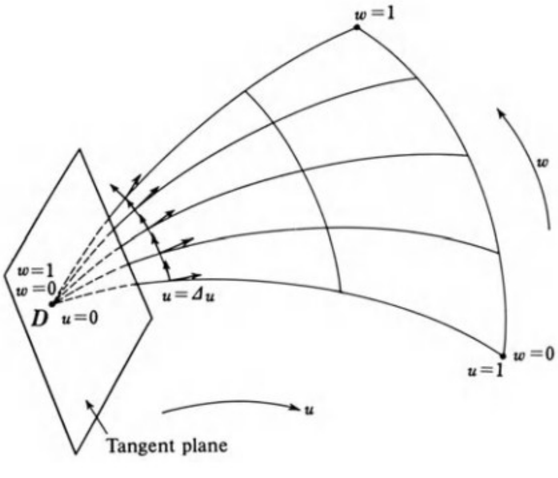

Fig. 1.36. A surface patch that has degenerated to a triangular shape

Since $\boldsymbol{P}_w(0, w) = \boldsymbol{0}$, the normal vector at point $\boldsymbol{D}$ cannot be determined by this method. In such a case, the normal vector is determined as a limit, as follows.

The unit normal vector $\boldsymbol{e}$ on the curve $u = \Delta u$ is:

$$\boldsymbol{e} = \frac{\boldsymbol{P}_u(\Delta u, w) \times \boldsymbol{P}_w(\Delta u, w)}{|\boldsymbol{P}_u(\Delta u, w) \times \boldsymbol{P}_w(\Delta u, w)|}. \tag{1.125}$$

We also have:

$$\boldsymbol{P}_u(\Delta u, w) \fallingdotseq \boldsymbol{P}_u(0, w) + \Delta u\, \boldsymbol{P}_{uu}(0, w) \tag{1.126}$$

$$\begin{aligned} \boldsymbol{P}_w(\Delta u, w) &\fallingdotseq \boldsymbol{P}_w(0, w) + \Delta u\, \boldsymbol{P}_{uw}(0, w) \\ &= \Delta u\, \boldsymbol{P}_{uw}(0, w). \end{aligned} \tag{1.127}$$

Substituting the relations in Eqs. (1.126) and (1.127) into (1.125) and taking the limit gives, for the unit normal vector at point $\boldsymbol{D}$:

$$\begin{aligned} \boldsymbol{e}_D &= \lim_{\Delta u \to 0} \boldsymbol{e} = \lim_{\Delta u \to 0} \frac{\boldsymbol{P}_u(0, w) \times \boldsymbol{P}_{uw}(0, w) + \Delta u\, \boldsymbol{P}_{uu}(0, w) \times \boldsymbol{P}_{uw}(0, w)}{|\boldsymbol{P}_u(0, w) \times \boldsymbol{P}_{uw}(0, w) + \Delta u\, \boldsymbol{P}_{uu}(0, w) \times \boldsymbol{P}_{uw}(0, w)|} \\ &= \frac{\boldsymbol{P}_u(0, w) \times \boldsymbol{P}_{uw}(0, w)}{|\boldsymbol{P}_u(0, w) \times \boldsymbol{P}_{uw}(0, w)|}. \end{aligned} \tag{1.128}$$

In order for the unit normal vector to be determined uniquely, formula (1.128) must have a unique value for any arbitrary w from 0 to 1.

In other words, when a triangular patch is expressed as an ordinary quadrangular patch, the unit normal vector at the degenerate point $\boldsymbol{D}$ cannot be found by normal means; it must be found using (1.128). And in order for formula (1.128) to be uniquely determined, special care must be taken (refer to Sects. 3.39 and 5.2.4).

1.3.9 Calculation of a Normal Vector on a Surface

A normal vector $\boldsymbol{N}$ at a point on a curved surface is calculated as follows:

$$\begin{aligned} \boldsymbol{N} &= \frac{\partial \boldsymbol{P}(u, w)}{\partial u} \times \frac{\partial \boldsymbol{P}(u, w)}{\partial w} \\ &= \left[\frac{\partial(y, z)}{\partial(u, w)} \; \frac{\partial(z, x)}{\partial(u, w)} \; \frac{\partial(x, y)}{\partial(u, w)} \right] \\ &\equiv [J_x \;\; J_y \;\; J_z]. \end{aligned} \tag{1.129}$$

The Jacobians such as $\dfrac{\partial(y, z)}{\partial(u, w)}$ are defined as follows:

$$\frac{\partial(y,z)}{\partial(u,w)} \equiv J_x = \begin{vmatrix} \dfrac{\partial y}{\partial u} & \dfrac{\partial z}{\partial u} \\ \dfrac{\partial y}{\partial w} & \dfrac{\partial z}{\partial w} \end{vmatrix}.$$

In a special case, for example, when a triangular patch is represented as a quadrangular patch, one boundary line degenerates to a point. At this point the Jacobian becomes singular and the normal vector cannot be found by the above method (refer to Sect. 1.3.8).

The unit normal vector $\boldsymbol{e}$ is:

$$\boldsymbol{e} = \left[\frac{J_x}{\sqrt{J_x^2+J_y^2+J_z^2}} \quad \frac{J_y}{\sqrt{J_x^2+J_y^2+J_z^2}} \quad \frac{J_z}{\sqrt{J_x^2+J_y^2+J_z^2}} \right]. \tag{1.130}$$

1.3.10 Calculation of Surface Area and Volume of a Surface

To find the surface area of a curved surface, the first fundamental matrix can be used and the calculation performed with Eq. (1.93), but Jacobians can also be used, as follows.

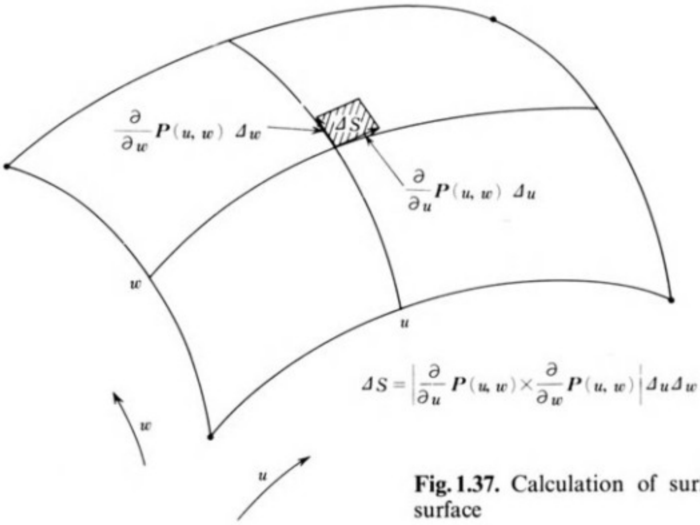
$\frac{\partial}{\partial w}\boldsymbol{P}(u,w)\ \Delta w$
ΔS
$\frac{\partial}{\partial u}\boldsymbol{P}(u,w)\ \Delta u$
w
u
$\Delta S = \left|\frac{\partial}{\partial u}\boldsymbol{P}(u,w) \times \frac{\partial}{\partial w}\boldsymbol{P}(u,w)\right| \Delta u \Delta w$
w
u

Fig. 1.37. Calculation of surface area of a curved surface

The surface area ΔS of an infinitesimal parallel quadrangle formed by the infinitesimal tangent vectors $\dfrac{\partial}{\partial u}\boldsymbol{P}(u,w)\,\Delta u$ and $\dfrac{\partial}{\partial w}\boldsymbol{P}(u,w)\,\Delta w$ in the u- and w-directions is (Fig. 1.37):

$$\Delta S = \left| \frac{\partial}{\partial u} \boldsymbol{P}(u, w) \times \frac{\partial}{\partial w} \boldsymbol{P}(u, w) \right| \Delta u \Delta w = |\boldsymbol{N}| \Delta u \Delta w$$
$$= \sqrt{J_x^2 + J_y^2 + J_z^2}\ \Delta u \Delta w. \tag{1.131}$$

The surface area S over the entire surface patch is then obtained as follows:

$$S = \int dS$$
$$= \int_0^1 \int_0^1 \sqrt{J_x^2 + J_y^2 + J_z^2}\ du dw. \tag{1.132}$$

Next, let us find the volume V of the space enclosed when a curved surface patch is projected onto the xy plane (Fig. 1.38). The projected area ΔS_z obtained by projecting an infinitesimal area on the xy plane is obtained by setting $J_x = J_y = 0$ in Eq. (1.131):

$$\Delta S_z = J_z \Delta u \Delta w.$$

This gives the following expression for the volume V enclosed by the column formed when ΔS is projected onto the xy plane:

$$\Delta V = z(u, w)\, \Delta S_z = z(u, w)\, J_z \Delta u \Delta w.$$

Accordingly, the total volume V becomes:

$$V = \int_0^1 \int_0^1 z(u, w)\, J_z du dw. \tag{1.133}$$

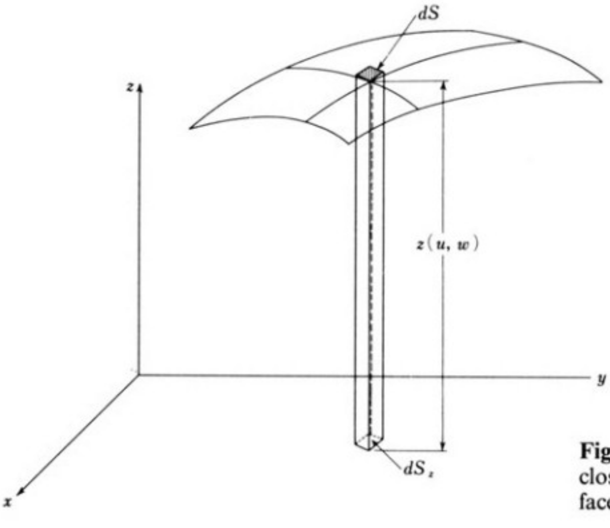

Fig. 1.38. Volume of the space enclosed by projecting a curved surface onto the xy plane

1.3.11 Offset Surfaces

The following equation can be used to generate a point on the curved surface $\boldsymbol{P}_0(u, w)$ that is offset a distance d with respect to the curved surface $\boldsymbol{P}(u, w)$.

$$\boldsymbol{P}_0(u, w) = \boldsymbol{P}(u, w) + \boldsymbol{e}(u, w)\, d. \tag{1.134}$$

In this equation, $\boldsymbol{e}(u, w)$ is the unit normal vector in the offset direction at point $\boldsymbol{P}(u, w)$.

To express $\boldsymbol{P}_0(u, w)$ as a bi-cubic surface in the format of Eq. (1.111), point $\boldsymbol{P}_0(u, w)$ can be calculated for the points corresponding to points $u, w = 0$, 1/3, 2/3, 1 on the curved surface $\boldsymbol{P}(u, w)$; then the bi-cubic surface passing through these 16 points can be found by solving simultaneous equations (Fig. 1.39). This curved surface is an approximation to the offset curved surface.

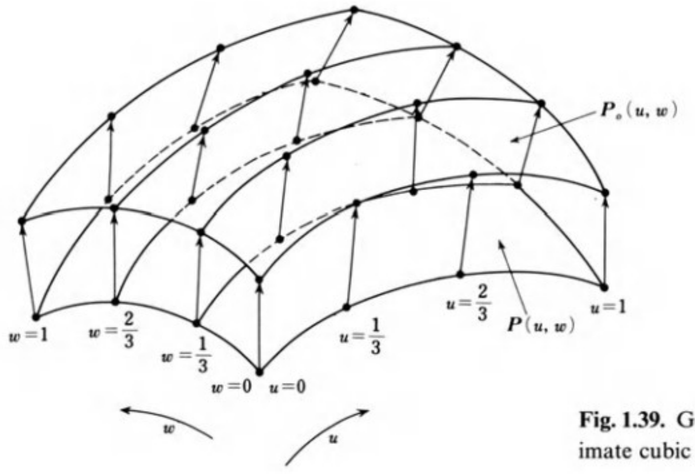

Fig. 1.39. Generation of an approximate cubic offset surface

References (Chap. 1)

4) Hosaka, Mamoru and Mitsuru Kuroda: On generation of curves and surfaces in CAD (in Japanese) *Information Processing* 17 (12) December, 1976.
5) Sasaki, Shigeo: *Differential Geometry* (in Japanese), Kyoritsu Shuppan, Basic Mathematics Textbooks 16, pp 13—16
6) Yamaguchi, Fujio· A method of designing free form surfaces by computer display (2nd report) (in Japanese), *Precision Machinery 43* (9).
7) Faux, I D. and M J. Pratt *Computational Geometry for Design and Manufacture,* Ellis Horwood, Mathematics and its Applications.

-./0- - -1 ! - @ ! ! " 4 ! ! 4"

" -./0- - -1 ! - 5 6 / ; " 4"

-./0- - -1 ! - % " ! 4"

-./0- - -1 ! - L ;M 5 6B 8 9 %%
6 4"

-./0- - -1 ! - > ! 4"

-./0- - -1 ! - = : I! ! ; ! "
! 4 ! " " " $<<B 4 ":;

-./0- - -1 ! - L ;M 5 6 8 9 %
! 6 F"! 4"

: -./0- - -1 ! - L ;M 5 6 8 9 %%
6 %% F"! < 4"

1? " ; " L ;M 5 6 4 ! JJJ

9 -./0- - -1 ! - 4 . " " 5 @
. > ! 6 = @ ! .44 ! B 4 > 96 4"

& + +$, = +

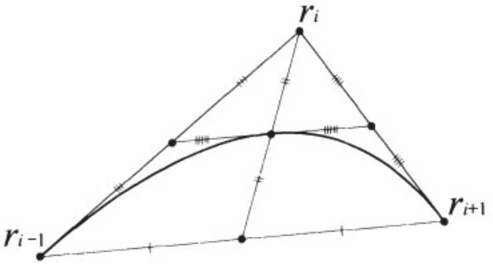

= + / ! ? ! 9 M ; ! " 4 6

Chapter 21

Polyhedron Subdivisions

This chapter is included to show a simple application of the Four-dimensional Processing to surface generation by polyhedron subdivision, and to show an example of how generality of representational capability is increased by a simple homogenization.

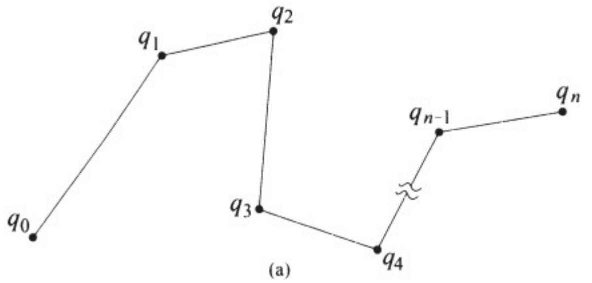

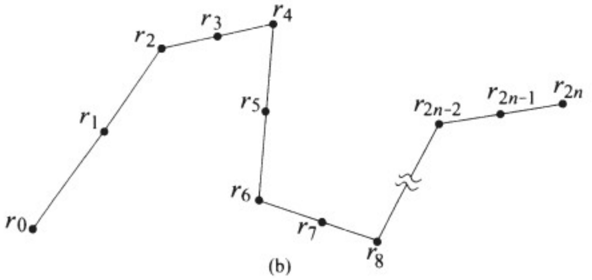

Figure 21.1: Chaikin's algorithm (1).

21.1 Chaikin's Algorithm

G.M. Chaikin proposed a method of generating curves procedurally, using a very simple geometrical algorithm, without depending on mathematical repre-

F. Yamaguchi, *Computer-Aided Geometric Design*
© Springer-Verlag Tokyo 2002

sentations [9]. Suppose that a polygon $q_0q_1\dots q_n$ is given, as in Fig. 21.1(a). Find the midpoints of all of the sides, and relabel the initially given vertices and these points as $r_0, r_1, \dots, r_{2n}$ (Fig. 21.1(b)).

Fig. 21.2(a) shows just the vertex $b(r_i)$ of the initial polygon and the midpoints $a(r_{i-1})$ and $c(r_{i+1})$ of the two sides attached to it.

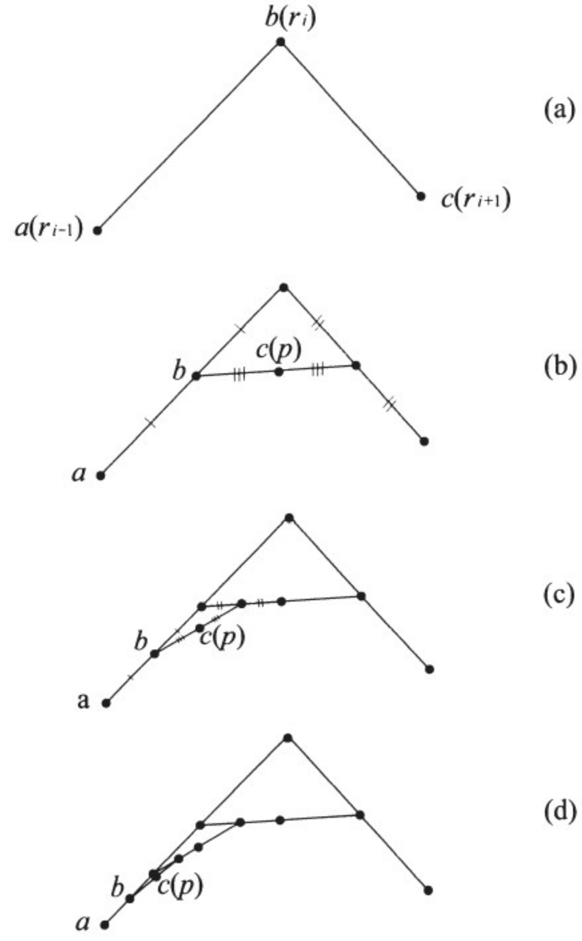

Figure 21.2: Chaikin's algorithm (2).

Letting p be the midpoint of the line segment formed by connecting the midpoints of sides ab and bc, we have:

$$\boldsymbol{p} = \frac{\frac{\boldsymbol{a}+\boldsymbol{b}}{2} + \frac{\boldsymbol{b}+\boldsymbol{c}}{2}}{2} = \frac{\boldsymbol{a} + 2\boldsymbol{b} + \boldsymbol{c}}{4}. \tag{21.1}$$

This point p becomes one point on the curve being generated. This division produces a set of 2 sides on either side of the point p. Eq.(21.1) can be applied to both of these sides. We push the set of 2 sides on the right side, for example, into stack and then apply eq.(21.1) again to the set of 2 sides on the left. This operation is repeated, pushing the set of 2 sides on the right into stack each time, until the length of the set of 2 sides on the left reaches a certain limit. At this point the contents of stack are poped up and similar operations are performed on them until stack is empty. This generates a sequence of points on the curve corresponding to $r_{i-1}r_ir_{i+1}$. If the processing above is performed for all groups of $r_{i-1}r_ir_{i+1} (i = 2, 4, \ldots, 2n-2)$, a curve corresponding to the given polygon $q_0q_1 \ldots q_n$ is generated. An example of such a curve is shown in Fig. 21.3 .

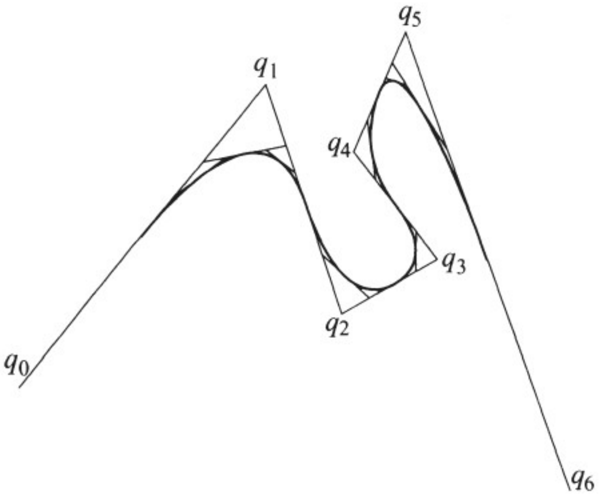

Figure 21.3: An example of a curve generated by Chaikin's method.

As can be understood from the generation process above, the curve segment is completely determined by only 3 points r_{i-1}, r_i, r_{i+1}. It has one end at r_{i-1} and the other at r_{i+1}. Its slope at r_{i-1} is in the direction of the line segment $\overrightarrow{r_{i-1}r_i}$, and its slope at r_{i+1} is in the direction of the line segment $\overrightarrow{r_ir_{i+1}}$. Since the curve passes through the point p calculated with eq.(21.1), it is evident that the curve is a parabola [1].

Let us find the equation of this parabola [48]. Writing the curve function in the form

$$\boldsymbol{p}_i(t) = [t^2 \; t \; 1] \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \\ \boldsymbol{c} \end{bmatrix}, \tag{21.2}$$

we obtain:

$$\dot{\boldsymbol{p}}_i(t) = [2t \; 1 \; 0] \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \\ \boldsymbol{c} \end{bmatrix}. \tag{21.3}$$

---

[1] See page 340 of [58].

Since we have the relations $\boldsymbol{p}_i(0)=\boldsymbol{r}_{i-1}$, $\boldsymbol{p}_i(1)=\boldsymbol{r}_{i+1}$, $\dot{\boldsymbol{p}}_i(0)=h(\boldsymbol{r}_i-\boldsymbol{r}_{i-1})$, $\dot{\boldsymbol{p}}_i(1)=k(\boldsymbol{r}_{i+1}-\boldsymbol{r}_i)$, the following equation is obtained:

$$\begin{bmatrix} \boldsymbol{r}_{i-1} \\ \boldsymbol{r}_{i+1} \\ h(\boldsymbol{r}_i-\boldsymbol{r}_{i-1}) \\ k(\boldsymbol{r}_{i+1}-\boldsymbol{r}_i) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \\ 2 & 1 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \\ \boldsymbol{c} \end{bmatrix}. \tag{21.4}$$

Converting the first matrix on the right-hand side into a square matrix gives:

$$\begin{bmatrix} \boldsymbol{r}_{i-1} \\ \boldsymbol{r}_{i+1} \\ h(\boldsymbol{r}_i-\boldsymbol{r}_{i-1}) \\ k(\boldsymbol{r}_{i+1}-\boldsymbol{r}_i) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 2 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \\ \boldsymbol{c} \\ \boldsymbol{0} \end{bmatrix}. \tag{21.5}$$

This implies:

$$\begin{aligned} \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \\ \boldsymbol{c} \\ \boldsymbol{0} \end{bmatrix} &= \begin{bmatrix} 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 2 & 1 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{r}_{i-1} \\ \boldsymbol{r}_{i+1} \\ h(\boldsymbol{r}_i-\boldsymbol{r}_{i-1}) \\ k(\boldsymbol{r}_{i+1}-\boldsymbol{r}_i) \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & -1/2 & 1/2 \\ -1 & 1 & 1/2 & -1/2 \\ 0 & 1 & -1/2 & -1/2 \\ 1 & -1 & 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} \boldsymbol{r}_{i-1} \\ \boldsymbol{r}_{i+1} \\ h(\boldsymbol{r}_i-\boldsymbol{r}_{i-1}) \\ k(\boldsymbol{r}_{i+1}-\boldsymbol{r}_i) \end{bmatrix} \\ &= \begin{bmatrix} \frac{h}{2}\boldsymbol{r}_{i-1} - \frac{1}{2}(h+k)\boldsymbol{r}_i + \frac{k}{2}\boldsymbol{r}_{i+1} \\ -(1+\frac{h}{2})\boldsymbol{r}_{i-1} + \frac{1}{2}(h+k)\boldsymbol{r}_i + (1-\frac{k}{2})\boldsymbol{r}_{i+1} \\ \frac{h}{2}\boldsymbol{r}_{i-1} + \frac{1}{2}(k-h)\boldsymbol{r}_i + (1-\frac{k}{2})\boldsymbol{r}_{i+1} \\ (1-\frac{h}{2})\boldsymbol{r}_{i-1} + \frac{1}{2}(h-k)\boldsymbol{r}_i - (1-\frac{k}{2})\boldsymbol{r}_{i+1} \end{bmatrix}. \end{aligned} \tag{21.6}$$

Equating the 4-th rows of the matrices on both sides gives:

$$(1-\frac{h}{2})\boldsymbol{r}_{i-1} + \frac{1}{2}(h-k)\boldsymbol{r}_i - (1-\frac{k}{2})\boldsymbol{r}_{i+1} = \boldsymbol{0}, \tag{21.7}$$

which yields:

$$h=k=2. \tag{21.8}$$

Therefore:

$$\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \\ \boldsymbol{c} \end{bmatrix} = \begin{bmatrix} \boldsymbol{r}_{i-1} - 2\boldsymbol{r}_i + \boldsymbol{r}_{i+1} \\ 2\boldsymbol{r}_i - 2\boldsymbol{r}_{i-1} \\ \boldsymbol{r}_{i-1} \end{bmatrix}. \tag{21.9}$$

Thus, the parabola function $\boldsymbol{p}_i(t)$ becomes:

$$\boldsymbol{p}_i(t) = [t^2 \ t \ 1] \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \\ \boldsymbol{c} \end{bmatrix}$$

$$
\begin{aligned}
&= [t^2\ t\ 1] \begin{bmatrix} \boldsymbol{r}_{i-1} - 2\boldsymbol{r}_i + \boldsymbol{r}_{i+1} \\ 2\boldsymbol{r}_i - 2\boldsymbol{r}_{i-1} \\ \boldsymbol{r}_{i-1} \end{bmatrix} \\
&= [t^2\ t\ 1] \begin{bmatrix} 1 & -2 & 1 \\ -2 & 2 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{r}_{i-1} \\ \boldsymbol{r}_i \\ \boldsymbol{r}_{i+1} \end{bmatrix} \\
&= [(1-t)^2\ 2t(1-t)\ t^2] \begin{bmatrix} \boldsymbol{r}_{i-1} \\ \boldsymbol{r}_i \\ \boldsymbol{r}_{i+1} \end{bmatrix} \\
&\equiv [B_{0,2}(t)\ B_{1,2}(t)\ B_{2,2}(t)] \begin{bmatrix} \boldsymbol{r}_{i-1} \\ \boldsymbol{r}_i \\ \boldsymbol{r}_{i+1} \end{bmatrix}, (i=2,4,\ldots,2n-2),
\end{aligned}
\tag{21.10}
$$

where $B_{0,2}(t), B_{1,2}(t), B_{2,2}(t)$ are Bernstein polynomials of degree 2. It is clear that $p_i(t)$ is a quadratic Bézier curve segment corresponding to the sequence of vertices r_{i-1}, r_i, r_{i+1}. It is easy to confirm that the following relations hold in this case (see Fig. 21.4):

$$
\boldsymbol{p}_i(1/2) = \frac{\boldsymbol{r}_{i-1} + 2\boldsymbol{r}_i + \boldsymbol{r}_{i+1}}{4}, \quad \dot{\boldsymbol{p}}_i(1/2) = \boldsymbol{r}_{i+1} - \boldsymbol{r}_{i-1}. \tag{21.11}
$$

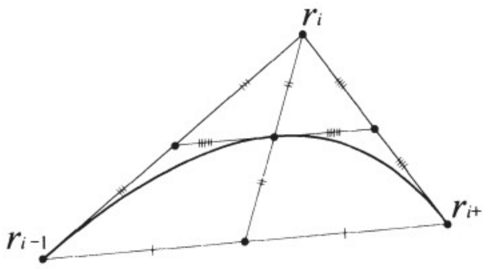

Figure 21.4: Relation between Chaikin's curve and polygon.

Let us now convert this into an expression in terms of the initially given sequence of vertices r_{i-2}, r_i, r_{i+2} rather than r_{i-1}, r_i, r_{i+1}. Since we have the relations:

$$
\begin{aligned}
\boldsymbol{r}_{i-1} &= \frac{\boldsymbol{r}_{i-2} + \boldsymbol{r}_i}{2}, \\
\boldsymbol{r}_{i+1} &= \frac{\boldsymbol{r}_i + \boldsymbol{r}_{i+2}}{2},
\end{aligned}
$$

the curve becomes:

$$
\boldsymbol{p}_i(t) = [(1-t)^2\ 2t(1-t)\ t^2] \begin{bmatrix} \boldsymbol{r}_{i-1} \\ \boldsymbol{r}_i \\ \boldsymbol{r}_{i+1} \end{bmatrix}
$$

$$= [(1-t)^2 \; 2t(1-t) \; t^2] \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1 & 0 \\ 0 & 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} \boldsymbol{r}_{i-2} \\ \boldsymbol{r}_i \\ \boldsymbol{r}_{i+2} \end{bmatrix}$$

$$= [\frac{1}{2}(1-t)^2 \quad -t^2+t+\frac{1}{2} \quad \frac{1}{2}t^2] \begin{bmatrix} \boldsymbol{r}_{i-2} \\ \boldsymbol{r}_i \\ \boldsymbol{r}_{i+2} \end{bmatrix}$$

$$\equiv [N_{0,3}(t) \; N_{1,3}(t) \; N_{2,3}(t)] \begin{bmatrix} \boldsymbol{r}_{i-2} \\ \boldsymbol{r}_i \\ \boldsymbol{r}_{i+2} \end{bmatrix}, (i=2,4,\ldots,2n-2).$$

If we let $i/2$ be a new subscript and call this i instead, then

$$\boldsymbol{p}_i(t) = [N_{0,3}(t) \; N_{1,3}(t) \; N_{2,3}(t)] \begin{bmatrix} \boldsymbol{q}_{i-1} \\ \boldsymbol{q}_i \\ \boldsymbol{q}_{i+1} \end{bmatrix}, (i=1,2,\ldots,n-1). \qquad (21.12)$$

This means that the function expresses a uniform B-spline curve of order 3 with respect to polygon $q_0q_1q_2\ldots q_{n-1}q_n$.

21.2 Homogenization of Chaikin's Algorithm

We shall homogenize the Chaikin's algorithm. Suppose a sequence of vectors $\boldsymbol{Q}_0, \boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_n$ is given in homogeneous coordinate vectors $[w_i \; X_i \; Y_i]$, and in vector space $\mathbf{R}^3$ we shall perform exactly the same operations as we did in the preceding section.

For $r_0, r_1, \ldots, r_{2n}$ in the preceding section we use capital bold letters $\boldsymbol{R}_0, \boldsymbol{R}_1, \ldots, \boldsymbol{R}_{2n}$ this time (see Fig. 21.5(a)). By using an operation corresponding to eq.(21.1) we get a new sequence of vectors.

This subdivision generates the following homogeneous Bézier curve of degree 2, which is equivalent to a rational quadratic Bézier curve:

$$\overline{\boldsymbol{P}_i(t)} = \overline{[B_{0,2}(t) \; B_{1,2}(t) \; B_{2,2}(t)] \begin{bmatrix} \boldsymbol{R}_{i-1} \\ \boldsymbol{R}_i \\ \boldsymbol{R}_{i+1} \end{bmatrix}}, \quad (i=2,4,\ldots,2n-2), \qquad (21.13)$$

where $\boldsymbol{R}_i = [w_i \; X_i \; Y_i]$.

Or, the curve form is also represented as below (see Fig. 21.5(b)):

$$\overline{\boldsymbol{P}_i(t)} = \overline{[N_{0,3}(t) \; N_{1,3}(t) \; N_{2,3}(t)] \begin{bmatrix} \boldsymbol{Q}_{i-1} \\ \boldsymbol{Q}_i \\ \boldsymbol{Q}_{i+1} \end{bmatrix}}, \quad (i=1,2,\ldots,n-1). \qquad (21.14)$$

Eq.(21.14) is the homogeneous representation of a rational B-spline curve of order 3 (degree 2).

% <
! -./0- - -1 ! - < / ! ; . 4 6 ! "
6 . 4"
& ' =

2

Geometry of Riemann Spaces

The fact that the geometry of the space in which we live is Euclidean is a very basic daily experience. This may explain why it took so long before it was realised that this may actually not be correct, and that the question of the geometry of the space around us is a matter of empirical assessment. Early in the 19th century Gauss studied the geometry of curved surfaces, and showed that all references to a flat embedding space could be eliminated. In the same way Riemann formulated in 1854 the geometry of 3D spaces. He found that Euclidean geometry is merely one possibility out of many. Riemann's method could be generalized to spaces of arbitrary dimension. The geometry of these curved Riemann spaces is wholly described within the space itself, by the use of co-ordinates and the metric tensor. No embedding is required. These geometrical concepts gradually spread beyond the mathematical incrowd, and in the last quarter of the 19th century the idea that a fourth (spatial) dimension might exist had mesmerized the public's imagination, perhaps even more so than black holes did a century later. One of the products of that period was Abbott's famous Flatland.[1] The flatland analogy is nowadays a standard technique of teachers to explain some of the intricacies of curved spaces.

The theoretical framework of Riemann spaces is also the starting point for the mathematical formulation of GR. In this chapter we discuss the tools that any student should master in order to be able to deal with GR beyond the level of handwaving. In doing so we have deliberately chosen to stay close to intuition as that outweighs the merits of rigour, certainly on first acquaintance.

2.1 Definition

A Riemann space has the following properties:

---

[1] Abbott, E.A.: 1884, *Flatland: A Romance of many Dimensions, by a Square*, Seeley & Co. (London).

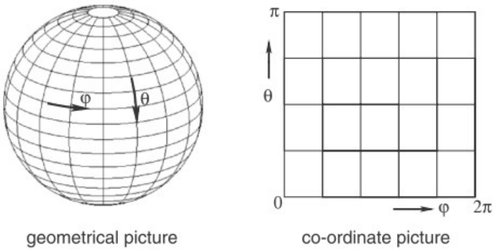

Fig. 2.1. A geometrical picture and the corresponding co-ordinate picture of the space defined by (2.2). Co-ordinate pictures will be frequently used.

1. Any point can be identified by a set of co-ordinates $\{x^\mu\}$; the number of independent x^μ is called the dimension.
2. It is possible to define continuously differentiable functions of $\{x^\mu\}$, in particular one-to-one co-ordinate transformations $\{x^\mu\} \leftrightarrow \{\overline{x}^\nu\}$.
3. There is a metric that specifies the distance $\mathrm{d}s^2$ between two nearby points x^μ and $x^\mu + \mathrm{d}x^\mu$:

$$\mathrm{d}s^2 = g_{\alpha\beta}\,\mathrm{d}x^\alpha \mathrm{d}x^\beta \quad ; \qquad g_{\alpha\beta} = g_{\beta\alpha} \; . \tag{2.1}$$

An antisymmetric part of $g_{\alpha\beta}$ does not contribute to $\mathrm{d}s^2$. Example: a spherical surface with radius 1 and co-ordinates θ, φ:

$$\mathrm{d}s^2 = \mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\varphi^2 \; . \tag{2.2}$$

Notation: $\mathrm{d}\theta^2 \equiv (\mathrm{d}\theta)^2$, $\mathrm{d}\varphi^2 \equiv (\mathrm{d}\varphi)^2$, but $\mathrm{d}s^2 = (\mathrm{d}s)^2$ only if $\mathrm{d}s^2 > 0$ as in (2.2). But the metric is in general not positive definite! In this simple case the geometrical structure may be visualised through embedding in an Euclidean space of one higher dimension, but for Riemann spaces of higher dimension this is no longer possible. Moreover, a Riemann space of dimension D cannot always be embedded in a flat space of dimension $D+1$. It is often useful to draw a *co-ordinate picture* of a suitably chosen subspace, even though it contains no information on the geometry, see Fig. 2.1.

An important point is that the metric determines the local structure of the space, but reveals nothing about its global (topological) structure. A plane, a cone and a cylinder all have the same metric $\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2$, but entirely different global structures.

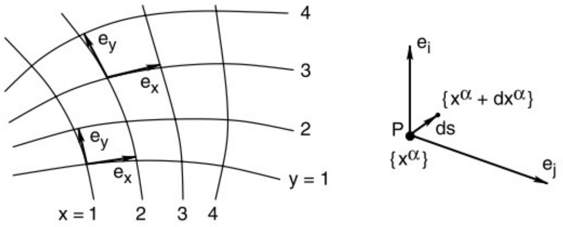

Fig. 2.2. Co-ordinate lines and base vectors spanning the tangent space. The choice of the co-ordinates is entirely free, and in practice dictated by the question which co-ordinates are the most expedient to use.

2.2 The tangent space

In each point we construct a set of base vectors tangent to the co-ordinate lines, as in Fig. 2.2. The arrow points towards increasing x^i. The base vectors span the flat tangent space, which has the same dimension as Riemann space. This construction evidently requires the existence of a flat embedding space, but that can be avoided as follows. Consider the curves $\{x^\alpha(p)\}$ through a point P in Riemann space (p = curve parameter), and construct $A^\sigma = [\mathrm{d}x^\sigma/\mathrm{d}p]_P$. These vectors A^σ span the abstract tangent space of P, which exists independent of any embedding. Usually, however, the abstract tangent space may be identified with the tangent space constructed in Fig. 2.2. For our discussion there is no real advantage in making the distinction and we shall work with the intuitive picture of Fig. 2.2.

We may use any metric we like in the tangent space, but there exists a preferred metric. Consider an infinitesimal section of Riemann space. This section is flat and virtually coincides with the tangent space. To an infinitesimal vector $\mathrm{d}\boldsymbol{s} = \mathrm{d}x^\alpha \boldsymbol{e}_\alpha$ *in the tangent space* we may therefore assign the length of the line element $\mathrm{d}s$ *in Riemann space*, i.e. we require $\mathrm{d}\boldsymbol{s} \cdot \mathrm{d}\boldsymbol{s} = \mathrm{d}s^2$:

$$\begin{aligned} \mathrm{d}\boldsymbol{s} \cdot \mathrm{d}\boldsymbol{s} \;&=\; (\mathrm{d}x^\alpha \boldsymbol{e}_\alpha) \cdot (\mathrm{d}x^\beta \boldsymbol{e}_\beta) \;=\; \boldsymbol{e}_\alpha \cdot \boldsymbol{e}_\beta \; \mathrm{d}x^\alpha \mathrm{d}x^\beta \\ &=\; g_{\alpha\beta} \; \mathrm{d}x^\alpha \mathrm{d}x^\beta \ , \end{aligned} \tag{2.3}$$

and it follows that

$$g_{\alpha\beta} \;\equiv\; \boldsymbol{e}_\alpha \cdot \boldsymbol{e}_\beta \ . \tag{2.4}$$

Here $\cdot$ represents the vector inner product. This may be the usual inner product, for example when we deal with 2D surfaces embedded in a flat R_3. But in case of the Minkowski spacetime of SR, and in GR, the inner product is not positive definite, and we may have that $\boldsymbol{A}\cdot\boldsymbol{A}<0$ (for spacelike vectors). By taking $\mathrm{d}x^\alpha = 1$ in (2.3) and all other $\mathrm{d}x^\beta = 0$ we see that $\boldsymbol{e}_\alpha\cdot\boldsymbol{e}_\alpha = \mathrm{d}\boldsymbol{s}\cdot\mathrm{d}\boldsymbol{s}$ (no summation). It follows that the 'length' of $\boldsymbol{e}_\alpha$ corresponds to a jump $\Delta x^\alpha = 1$, at constant value of the other co-ordinates. Due to the curvature this is of course only approximately correct. These base vectors are called a co-ordinate basis because they are defined entirely by the co-ordinates and the metric. The length of the base vectors depends on the choice of the co-ordinates, and is in general a function of position. Consider for example polar co-ordinates in a plane, Fig. 2.3. The length of $\boldsymbol{e}_\mathrm{r}$ is constant, while $|\boldsymbol{e}_\varphi| \propto r$:

$$\mathrm{d}s^2 \;=\; \underset{\substack{\uparrow \\ \boldsymbol{e}_\mathrm{r}\cdot\boldsymbol{e}_\mathrm{r}}}{1}\cdot\mathrm{d}r^2 \;+\; \underset{\substack{\uparrow \\ \boldsymbol{e}_\varphi\cdot\boldsymbol{e}_\varphi}}{r^2}\mathrm{d}\varphi^2\ . \tag{2.5}$$

Now that we have defined the basis we may construct finite vectors $\boldsymbol{A} = A^\alpha\boldsymbol{e}_\alpha$ in the tangent space through the usual parallelogram construction. These so called contravariant components A^α are the components of $\boldsymbol{A}$ along the basis.

The next step is to define another (covariant) representation A_α of $\boldsymbol{A}$ by demanding that $\boldsymbol{A}\cdot\boldsymbol{A} = A_\alpha A^\alpha$, for every $\boldsymbol{A}$:

$$\boldsymbol{A}\cdot\boldsymbol{A} \;=\; (A^\alpha\boldsymbol{e}_\alpha)\cdot(A^\beta\boldsymbol{e}_\beta) \;=\; g_{\alpha\beta}\,A^\beta A^\alpha \;\equiv\; A_\alpha A^\alpha\ , \tag{2.6}$$

which leads to:

$$A_\alpha = g_{\alpha\beta}A^\beta\ . \tag{2.7}$$

In a more advanced treatment a distinction is made between tensors as geometrical objects, their contravariant representation located in an abstract tangent space, and the dual tangent space, in which the covariant representations reside. In the current, more primitive context the following interpretation suggests itself. Since $A_\gamma = g_{\gamma\beta}A^\beta = \boldsymbol{e}_\gamma\cdot\boldsymbol{e}_\beta A^\beta = (A^\beta\boldsymbol{e}_\beta)\cdot\boldsymbol{e}_\gamma = \boldsymbol{A}\cdot\boldsymbol{e}_\gamma$, it follows that A_γ is the projection of $\boldsymbol{A}$ on $\boldsymbol{e}_\gamma$. Hence, the contravariant components A^β are the components of $\boldsymbol{A}$ along the base vectors $\boldsymbol{e}_\beta$ (parallelogram construction), while the covariant component A_α is the projection of $\boldsymbol{A}$ on the base vector $\boldsymbol{e}_\alpha$, Fig. 2.3, right:

$$\text{contravariant }(A^\beta):\ \ \boldsymbol{A}\ = A^\beta\boldsymbol{e}_\beta\ , \tag{2.8}$$

$$\text{covariant }\ (A_\alpha):\ \ A_\alpha = \boldsymbol{A}\cdot\boldsymbol{e}_\alpha\ . \tag{2.9}$$

Finally, the concept of *index raising and lowering.* We can lower an index with the help of (2.7). The inverse operation of raising is defined as:

$$A^\gamma \;=\; g^{\gamma\alpha}A_\alpha\ . \tag{2.10}$$

The meaning of $g^{\gamma\alpha}$ can be gleaned from:

$$A^\gamma \;=\; g^{\gamma\alpha}A_\alpha \;=\; g^{\gamma\alpha}g_{\alpha\nu}A^\nu\ , \tag{2.11}$$

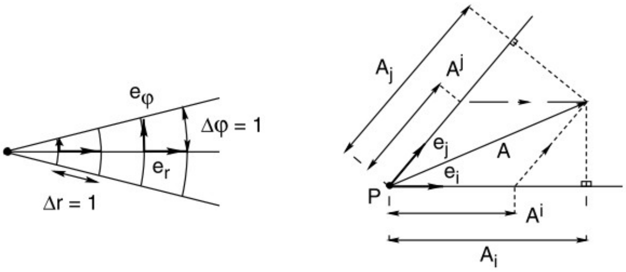

Fig. 2.3. Left: polar co-ordinates and the base vectors $\boldsymbol{e}_{\mathrm{r}}$ and $\boldsymbol{e}_{\varphi}$. Right: interpretation of the contravariant and covariant representation of a vector $\boldsymbol{A}$.

so that $g^{\gamma\alpha}g_{\alpha\nu} = \delta^{\gamma}_{\nu}$, i.e. $\{g^{\gamma\alpha}\}$ is the inverse of $\{g_{\alpha\nu}\}$. In summary:

$$\left.\begin{array}{lrcl} \text{index lowering :} & A_\alpha &=& g_{\alpha\beta}A^\beta \ , \\ \text{index raising :} & A^\gamma &=& g^{\gamma\nu}A_\nu \ , \\ & \{g^{\gamma\nu}\} &=& \{g_{\alpha\beta}\}^{-1} \ . \end{array}\right\} \tag{2.12}$$

We have silently adopted the *summation convention*: if an index occurs twice, once as a lower and once as an upper index, summation over that index is implied. Note that the rules for index raising and lowering are always valid, and have nothing to do with the question whether one is dealing with a tensor or not. The tensor concept is related to behaviour under co-ordinate transformations, which was not an issue above, and to which we turn our attention now.

2.3 Tensors

We are now in a position to do linear algebra in the tangent space, but we leave that aside and study the effect of co-ordinate transformations. Consider two overlapping sets of co-ordinates $\{x^\mu\}$ and $\{x^{\mu'}\}$. The notation is sloppy – it would be more appropriate to write $\{\overline{x}^\mu\}$ instead of $\{x^{\mu'}\}$, but $\{x^{\mu'}\}$ is much more expedient if used with care. A displacement $\delta x^{\mu'}$ is related to a displacement δx^ν through:

$$\delta x^{\mu'} = \frac{\partial x^{\mu'}}{\partial x^\nu}\,\delta x^\nu \equiv x^{\mu'}_{\ ,\nu}\,\delta x^\nu \ . \tag{2.13}$$

Notation:

$$X_{,\nu} \equiv \frac{\partial X}{\partial x^{\nu}} \; ; \qquad X_{,\nu\rho} \equiv \frac{\partial^2 X}{\partial x^{\nu} \partial x^{\rho}} \; , \quad \text{etc.} \tag{2.14}$$

where X can be anything (A_{α}, $g^{\alpha\beta}, ...$). We may freely interchange indices behind the comma: $X_{,\alpha\beta\gamma} = X_{,\alpha\gamma\beta} = X_{,\gamma\alpha\beta}$ etc.

Any set A^{μ} transforming according to (2.13) is called a contravariant tensor of rank 1:

$$A^{\mu'} = x^{\mu'}_{,\nu}\, A^{\nu} \quad \leftrightarrow \quad A^{\nu} \;\; \text{contravariant.} \tag{2.15}$$

Hence δx^{ν} is a contravariant tensor. Tensors of rank 1 are often referred to as vectors, and henceforth we shall use the word vector in this sense only. A function such as the temperature distribution $T(x)$ is called a *scalar*, a tensor of rank zero. Its value in a point is independent of the co-ordinate system, i.e. invariant for co-ordinate transformations: $T'(x') = T(x)$, where T' is the new function prescription. The derivative of a scalar Q,

$$B_{\mu} = \frac{\partial Q}{\partial x^{\mu}} \equiv Q_{,\mu} \tag{2.16}$$

transforms like $B_{\mu'} = Q'_{,\mu'} = Q_{,\nu}\, x^{\nu}_{,\mu'} = x^{\nu}_{,\mu'}\, B_{\nu}$. Every B_{ν} that transforms in this way is called a covariant vector or tensor of rank 1:

$$B_{\mu'} = x^{\nu}_{,\mu'}\, B_{\nu} \quad \leftrightarrow \quad B_{\nu} \;\; \text{covariant.} \tag{2.17}$$

From two covariant vectors we can form $T_{\mu\nu} = A_{\mu} B_{\nu}$, a covariant tensor of rank 2. More general tensors can be constructed through summation, $T_{\mu\nu} = A_{\mu} B_{\nu} + C_{\mu} D_{\nu} + ...$ This process may be continued: $T_{\alpha\beta} C^{\gamma}$ and $A_{\mu} C^{\nu} B_{\rho}$ are mixed tensors of rank 3 (provided T, A, B and C are tensors themselves). The indices of tensors of higher rank transform according to (2.15) resp. (2.17), for example:

$$T^{\alpha'}{}_{\beta'\gamma'}{}^{\delta'} = x^{\alpha'}_{,\mu}\, x^{\nu}_{,\beta'}\, x^{\sigma}_{,\gamma'}\, x^{\delta'}_{,\tau}\, T^{\mu}{}_{\nu\sigma}{}^{\tau} \; . \tag{2.18}$$

There is no other choice because (2.18) must hold for the special tensor $T^{\alpha}{}_{\beta\gamma}{}^{\delta} = P^{\alpha} Q_{\beta} R_{\gamma} S^{\delta}$, and the transformation rules for vectors have already been fixed! Note that we get a glimpse here of how the Lorentz transformations of SR will be generalised in GR: relation (1.7) of SR will be replaced by (2.15). This transformation is still locally linear, but different in each point of Riemann space as the $\{x^{\mu'}{}_{,\nu}\}$ are functions of position. The single global Lorentz transformation will be replaced by a mesh of local Lorentz transformations.

The horizontal position of the indices is important: $T^{\mu}{}_{\nu}$ is different from $T_{\nu}{}^{\mu}$! The summation over double indices is called *contraction*. It lowers the rank by two. For example $T^{\mu}{}_{\mu}$, $T^{\alpha}{}_{\beta\alpha}{}^{\gamma}$, $P_{\alpha\beta}\, Q^{\beta\gamma}$, $T^{\alpha}{}_{\beta\alpha}{}^{\beta}$ (double contraction). Double indices are *dummies*: $T^{\alpha}{}_{\alpha} = T^{\mu}{}_{\mu}$, dummies may occur only twice, once as an upper and once as a lower index. If you encounter expressions like $C^{\mu\mu}$, $P^{\alpha\beta}Q^{\alpha}_{\;\gamma}$ or $P_{\alpha\beta}\,Q^{\alpha\gamma}\,R_{\delta\alpha}$ then you have made a mistake somewhere!

Index raising and lowering, finally, is done by factors $g_{\alpha\beta}$ or $g^{\mu\nu}$ for each upper/lower index, e.g.:

$$\left.\begin{aligned} T^{\mu\nu} &= g^{\mu\alpha}\,T_{\alpha}^{\ \ \nu}\ , \\ T^{\alpha}{}_{\beta}{}^{\gamma\delta} &= g^{\alpha\mu}\,g_{\beta\nu}\,g^{\delta\sigma}\;T_{\mu}{}^{\nu\gamma}{}_{\sigma}\ , \quad \text{etc.} \end{aligned}\right\} \tag{2.19}$$

Again, like in (2.18), we have hardly any other choice here, because (2.19) must hold for the special tensors $T^{\mu\nu} = P^{\mu}Q^{\nu}$ and $T^{\alpha}{}_{\beta}{}^{\gamma\delta} = P^{\alpha}Q_{\beta}R^{\gamma}S^{\delta}$, and the rules for index raising and lowering for vectors have already been fixed. We are now in a position that we can raise and lower indices at liberty. We emphasise once more that the rules (2.12) and (2.19) for index gymnastics are generally valid, also for non-tensors. For example, $Q_{\mu\nu} = A_{\mu,\nu}$ is not a tensor (exercise 2.4), and yet $Q^{\mu}{}_{\nu} = g^{\mu\alpha}Q_{\alpha\nu}$.

Exercise 2.1: The unit tensor is defined as $\delta^{\alpha}{}_{\beta} = 1$ for $\alpha = \beta$, otherwise 0. Prove that $\delta^{\alpha}{}_{\beta}$ is a tensor, and that $\delta^{\alpha}{}_{\beta} = \delta_{\beta}{}^{\alpha}$, so that we may write δ^{α}_{β} without risk of confusion. Show that $\delta_{\alpha\beta} = g_{\alpha\beta}$. Is $\eta_{\alpha\beta}$ a tensor? And $g_{\alpha\beta}$? One could define $\delta_{\alpha\beta} = 1$ for $\alpha = \beta$, and 0 otherwise, but then $\delta_{\alpha\beta}$ is not a tensor.

Hint: $\delta^{\alpha'}{}_{\beta'}$ must be equal to $x^{\alpha'}{}_{,\nu}\,x^{\mu}{}_{,\beta'}\,\delta^{\nu}{}_{\mu}$, or $\delta^{\alpha'}{}_{\beta'} = x^{\alpha'}{}_{,\nu}\,x^{\nu}{}_{,\beta'} = x^{\alpha'}_{\ ,\beta'}$ (chain rule) $= 1$ for $\alpha' = \beta'$ otherwise 0. Hence $\delta^{\alpha}{}_{\beta}$ is tensor. And $\delta_{\beta}{}^{\alpha} = g_{\beta\mu}\,g^{\alpha\nu}\delta^{\mu}{}_{\nu} = g_{\beta\mu}\,g^{\alpha\mu} = g^{\alpha\mu}g_{\mu\beta} = 1$ for $\alpha = \beta$, otherwise 0, i.e. identical to $\delta^{\alpha}{}_{\beta}$; $\delta_{\alpha\beta} = g_{\alpha\nu}\delta^{\nu}_{\beta} = g_{\alpha\beta}$; $\eta_{\alpha\beta}$ is a tensor in SR only, i.e. under Lorentz transformations; $g_{\alpha\beta}$ tensor: use (2.1), require that $\mathrm{d}s^2$ is also tensor in GR (invariant scalar), and $\mathrm{d}x^{\alpha}$ is tensor, then exercise 2.3. Other definition $\delta_{\alpha\beta}$: $\delta_{\alpha'\beta'} = x^{\nu}{}_{,\alpha'}\,x^{\mu}{}_{,\beta'}\,\delta_{\nu\mu}$? No, because the chain rule can no longer be used.

Exercise 2.2: If $T^{\alpha\beta}$ and $P^{\mu}{}_{\nu}$ are tensors then $P^{\mu}{}_{\mu}$ is a scalar, but $T^{\alpha\alpha}$ is not. The inner product $A_{\nu}B^{\nu}$ of two vectors is a scalar.

Hint: $P^{\mu'}{}_{\mu'} = x^{\mu'}{}_{,\alpha}\,x^{\beta}{}_{,\mu'}\,P^{\alpha}{}_{\beta}$, then the chain rule.

Exercise 2.3: Quotient theorem: If $A^{\lambda}P_{\lambda\mu\nu}$ is a tensor for arbitrary vector A^{λ}, then $P_{\lambda\mu\nu}$ is a tensor; $\mu\nu$ may be replaced with an arbitrary sequence of upper / lower indices.

Hint: $A^\lambda P_{\lambda\mu\nu}$ is a tensor, i.e. $A^{\lambda'} P_{\lambda'\mu'\nu'} = x^\alpha_{,\mu'}\, x^\beta_{,\nu'}\, A^\sigma P_{\sigma\alpha\beta}$ (λ' and σ are dummies!), then substitute $A^\sigma = x^\sigma_{,\lambda'}\, A^{\lambda'}$, etc.

Exercise 2.4: The derivative $A_{\mu,\nu}$ of a covariant vector A_μ is not a tensor, as it transforms according to:

$$A_{\mu',\nu'} \;=\; A_{\alpha,\beta}\, x^\alpha_{,\mu'}\, x^\beta_{,\nu'} \;+\; A_\alpha\, x^\alpha_{,\mu'\nu'}\;. \tag{2.20}$$

The problem is in the second term of (2.20). In SR only linear (Lorentz) transformations are allowed. In that case the second term is zero and $A_{\mu,\nu}$ is a tensor.

Hint: Start from $A_{\mu',\nu'} = (x^\alpha_{,\mu'}\, A_\alpha)_{,\nu'}$, then use the product rule.

Exercise 2.5: Prove $T_\alpha{}^\nu A_\nu \;=\; T_{\alpha\nu}A^\nu\,;\; T_\alpha{}^\alpha \;=\; T^\alpha{}_\alpha\,;\; g^\nu{}_\nu \;=\; 4\,;\;\; \eta^\nu{}_\nu \;=\; g^{00} - g^{11} - g^{22} - g^{33}$.

Hint: We know that $g^\nu{}_\nu = g^{\nu\alpha}g_{\alpha\nu} = \delta^\nu_\nu = 4$. The following may be illuminating: the scalar $g^\nu{}_\nu$ is invariant, compute in a freely falling frame: $g^\nu{}_\nu = \eta^\nu{}_\nu$, SR holds in that frame: $\eta^\nu{}_\nu = \eta^{\nu\alpha}\eta_{\alpha\nu} = 4$. But in GR: $\eta^\nu{}_\nu = g^{\nu\alpha}\eta_{\alpha\nu} =$ etc.

2.4 Parallel transport and Christoffel symbols

Consider a particle at position P in Riemann space, Fig. 2.4. The vectors associated with it (velocity, spin, ..) reside in the tangent space of P. At some later time the particle has moved to position Q, but the tangent space of Q does not coincide with that of P. To be able to do dynamics, we must develop a way to compare vectors in the different tangent spaces along the worldline of the particle. In other words, we need something against which to gauge the concept of ‘change’. This is what parallel transport in GR is about.

Fig. 2.4 shows the curve $x^\sigma(p)$ in Riemann space. The vector $\boldsymbol{A}$ is always in the tangent space, but the tangent spaces of P, Q, R, .. are disjunct, and comparison of $\boldsymbol{A}(P)$ with $\boldsymbol{A}(Q)$ or $\boldsymbol{A}(R)$ is not possible. To this end we define a *connection* between tangent spaces, that is, a mathematical prescription telling us how a vector $\boldsymbol{A}(P)$ lies in the tangent space of Q if we ‘transport’ it along a given path from P to Q. This can be done in a variety of ways, but much of the mathematical freedom that we have is eliminated by the physical

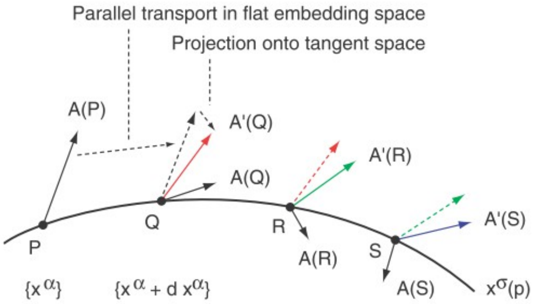
Parallel transport in flat embedding space
Projection onto tangent space
A(P)
A'(Q)
A(Q)
A'(R)
A'(S)
Q
R
P
A(R)
S
A(S)
{x^α}
{x^α + d x^α}
x^σ(p)

Fig. 2.4. Conceptual definition of parallel displacement of a vector along a curve $x^\sigma(p)$ in Riemann space: first an ordinary parallel displacement in the flat embedding space (resulting in the dashed arrows) followed by projection on the local tangent space. The process is repeated in infinitesimal steps.

requirement that we recover what we ordinarily do when we transport a vector parallel to itself in a flat space. Imagine the Riemann space embedded in a flat space of higher dimension. We know how to move $\boldsymbol{A}(P)$ around parallel to itself in this embedding space, because it is flat. Having arrived in Q, the result is projected onto the local tangent space. To order $O(\mathrm{d}x^\alpha)$ projection does not change the length of the vector: the projection angle γ is $O(\mathrm{d}x^\alpha)$, but $\cos\gamma = 1$ up to $O(\mathrm{d}x^\alpha)$. This process is now repeated with infinitesimal steps, and generates the coloured vector field $\boldsymbol{A}'$ in Fig. 2.4, starting from $\boldsymbol{A}(P)$. In this way we have generalized the concept of parallel transport to curved spaces, in such a way that it reduces to normal parallel transport for flat spaces. Not surprisingly, it is also the definition that turns out to work in GR. The result of the transport operation depends on the path, see Fig. 2.5. However, when $\boldsymbol{e}$ in Fig. 2.5 is parallel-transported along a small curve on the sphere there is virtually no change, because there is hardly any curvature felt (exercise 2.17).

We now formalise our intuitive approach. The difference $\mathrm{d}\boldsymbol{A} = \boldsymbol{A}(Q) - \boldsymbol{A}(P)$ is not defined, but up to order $O(\mathrm{d}x^\alpha)$ we have that $\mathrm{d}\boldsymbol{A} \simeq \boldsymbol{A}(Q) - \boldsymbol{A}'(Q)$, and this is useful as both vectors lie in the same tangent space. The vector $\mathrm{d}\boldsymbol{A}$ may be interpreted as the intrinsic change of $\boldsymbol{A}$, after correction for the 'irrelevant' change in the orientation of the tangent space:

$$\mathrm{d}\boldsymbol{A} \simeq \boldsymbol{A}(Q) - \boldsymbol{A}'(Q) \tag{2.21}$$

$$= \mathrm{d}(A^\mu \boldsymbol{e}_\mu) = (\mathrm{d}A^\mu)\boldsymbol{e}_\mu + A^\mu(\mathrm{d}\boldsymbol{e}_\mu) \,. \tag{2.22}$$

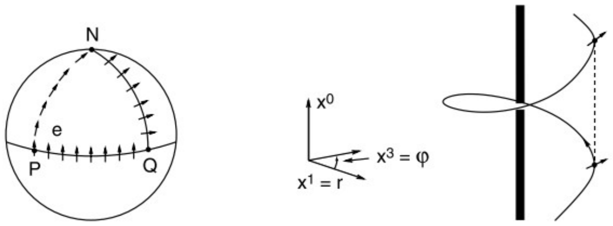

Fig. 2.5. Left: Parallel displacement of the vector $\boldsymbol{e}$ along PNQ and along PQ produces entirely different results. To the right, the geodesic precession of a top in orbit around a central mass, see text.

Here, $\mathrm{d}\boldsymbol{A}$ has been split into two contributions: the change $\mathrm{d}A^\mu \equiv A^\mu(Q) - A^\mu(P)$ of the contravariant components of $\boldsymbol{A}$, and a contribution from the change of the base vectors. On general grounds we anticipate $\mathrm{d}\boldsymbol{e}_\mu$ to be proportional to both $\{\mathrm{d}x^\beta\}$ and $\{\boldsymbol{e}_\alpha\}$:

$$\mathrm{d}\boldsymbol{e}_\mu = \Gamma^\alpha_{\ \mu\beta}\, \mathrm{d}x^\beta\, \boldsymbol{e}_\alpha \ . \tag{2.23}$$

$\Gamma^\alpha_{\ \mu\beta}$ is called the *Christoffel symbol of the second kind*, and as may be expected it is intimately related to the metric tensor:

$$\Gamma^\mu_{\ \nu\sigma} = \tfrac{1}{2} g^{\mu\lambda} \left(g_{\lambda\nu,\sigma} + g_{\lambda\sigma,\nu} - g_{\nu\sigma,\lambda} \right) \equiv g^{\mu\lambda}\, \Gamma_{\lambda\nu\sigma} \ . \tag{2.24}$$

The = sign is proved in § 2.5. The ≡ sign defines the *Christoffel symbol of the first kind*, simply by raising one index with $g^{\mu\lambda}$. According to (2.23) the Christoffel symbols define the connection between the base vectors of the tangent spaces at different positions. As pointed out above, there exist more general connection coefficients than (2.24), but these play no role in GR.

Insert (2.23) in (2.22) and rename the dummy-indices:

$$\mathrm{d}\boldsymbol{A} = \left(\mathrm{d}A^\mu + \Gamma^\mu_{\ \nu\sigma} A^\nu\, \mathrm{d}x^\sigma \right) \boldsymbol{e}_\mu \equiv (\mathrm{D}A^\mu)\, \boldsymbol{e}_\mu \ . \tag{2.25}$$

The right hand side defines the intrisic change $\mathrm{D}A^\mu$, which apparently obeys the following equation:

$$\frac{\mathrm{D}A^\mu}{\mathrm{D}p} = \frac{\mathrm{d}A^\mu}{\mathrm{d}p} + \Gamma^\mu_{\ \nu\sigma} A^\nu \frac{\mathrm{d}x^\sigma}{\mathrm{d}p} \qquad \text{(contravariant);} \tag{2.26}$$

$$\frac{\mathrm{D}A_\mu}{\mathrm{D}p} = \frac{\mathrm{d}A_\mu}{\mathrm{d}p} - \Gamma^\nu_{\ \mu\sigma} A_\nu \frac{\mathrm{d}x^\sigma}{\mathrm{d}p} \qquad \text{(covariant).} \tag{2.27}$$

For the second relation (2.27) see exercise 2.8. We may apply these equations in two ways. For a *given* vector field we may compute $\mathrm{D}A^{\mu}$ or $\mathrm{D}A_{\mu}$ for a displacement $\mathrm{d}p$ along $x^{\sigma}(p)$. On the other hand, one may solve $\mathrm{D}A^{\mu}/\mathrm{D}p = 0$ or $\mathrm{D}A_{\mu}/\mathrm{D}p = 0$ starting from an initial value $A^{\mu}(P)$ or $A_{\mu}(P)$, and construct a vector field along $x^{\sigma}(p)$ for which $\mathrm{d}\boldsymbol{A} = \boldsymbol{A} - \boldsymbol{A}' = 0$. Parallel transport of a vector along $x^{\sigma}(p)$ is therefore described by the differential equation

$$\frac{\mathrm{D}A_{\mu}}{\mathrm{D}p} = 0 \qquad \text{or} \qquad \frac{\mathrm{D}A^{\mu}}{\mathrm{D}p} = 0 \ . \tag{2.28}$$

We mention a few properties of the Christoffel symbols. They are symmetrical in the last two indices:

$$\Gamma^{\mu}{}_{\nu\sigma} = \Gamma^{\mu}{}_{\sigma\nu} \ ; \qquad \Gamma_{\lambda\nu\sigma} = \Gamma_{\lambda\sigma\nu} \ . \tag{2.29}$$

By interchanging the indices in (2.24) we may infer $\Gamma_{\nu\lambda\sigma}$, and on adding that to $\Gamma_{\lambda\nu\sigma}$ one obtains

$$\Gamma_{\lambda\nu\sigma} + \Gamma_{\nu\lambda\sigma} = g_{\lambda\nu,\sigma} \ . \tag{2.30}$$

The Christoffel symbol transforms according to

$$\Gamma^{\mu'}{}_{\nu'\sigma'} = \Gamma^{\rho}{}_{\alpha\beta}\, x^{\mu'}{}_{,\rho}\, x^{\alpha}{}_{,\nu'}\, x^{\beta}{}_{,\sigma'} + x^{\mu'}{}_{,\rho}\, x^{\rho}{}_{,\nu'\sigma'} \ . \tag{2.31}$$

The proof is for diehards (see literature). The first term is what we would expect if the Christoffel symbol were a tensor, but the second term makes that it is actually not a tensor. The concept of parallel transport will be used in § 2.5 to define geodesics.

In SR the velocity and spin vector of a particle on which no forces are exerted are constant. They are transported parallel along the 'straight' orbit of the particle. The idea of GR is that a particle under the influence of gravity moves freely in a *curved* spacetime. A natural generalisation is that velocity and spin vector of the particle can be found by parallel transport along the orbit in spacetime. In this way we are able to understand the geodesic precession of a top. Fig. 2.5 shows a co-ordinate picture, with $x^0 = ct$ on the vertical axis and polar co-ordinates $x^1 = r$ and $x^3 = \varphi$ in the horizontal plane. The worldline of the top orbiting the central object (vertical bar) is a spiral. The spin 4-vector (whose spatial part is directed along the spin axis) is parallel-transported along the worldline. After one revolution the top has returned to same spatial position, but because spacetime is not flat – not visible in a co-ordinate picture – the spin vector has changed its direction. At this point one may wonder how the effect is related to the Thomas precession. We refer to Ch. 8 for a more general treatment, from which both Thomas precession and geodesic precession emerge in the appropriate limit.

Exercise 2.6: The length of a vector remains constant under parallel transport:

$$\mathrm{d}\,A^\nu A_\nu \;=\; \mathrm{d}\,(g_{\mu\nu}A^\mu A^\nu) = 0 \ .$$

Hint: First attempt: d = D = intrinsic change: $\mathrm{D}A^\nu A_\nu = (\mathrm{D}A^\nu)A_\nu + A^\nu(\mathrm{D}A_\nu) = 0$, because $\mathrm{D}A^\nu = (\mathrm{D}A^\nu/\mathrm{D}p)\,\mathrm{d}p = 0$, etc. But (2.27) must still be proven, and for that we need $\mathrm{d}\,A^\nu A_\nu = 0$. Second attempt: d = total change: $\mathrm{d}\,g_{\mu\nu}A^\mu A^\nu = 2A_\nu\mathrm{d}A^\nu + A^\mu A^\nu\, g_{\mu\nu,\sigma}\,\mathrm{d}x^\sigma$; (2.26): $\mathrm{d}A^\nu = -\Gamma^\nu_{\ \mu\sigma}\,A^\mu\mathrm{d}x^\sigma$; exercise 2.5 and (2.30): $2A_\nu\mathrm{d}A^\nu = -2\Gamma_{\nu\mu\sigma}\,A^\nu A^\mu\mathrm{d}x^\sigma = -g_{\nu\mu,\sigma}\,A^\nu A^\mu\mathrm{d}x^\sigma$.

Exercise 2.7: Prove that $\mathrm{d}\,A^\nu B_\nu \;=\; 0$ under parallel transport.

Hint: The length of $A^\nu + B^\nu$ is constant.

Exercise 2.8: For parallel transport of a covariant vector:

$$\mathrm{d}B_\mu \;=\; \Gamma^\nu_{\ \mu\sigma}\,B_\nu\,\mathrm{d}x^\sigma \ . \tag{2.32}$$

Hint: $0 \;=\; \mathrm{d}\,A^\mu B_\mu \;=\; A^\mu\,\mathrm{d}B_\mu \;+\; B_\mu\,\mathrm{d}A^\mu$, and $\mathrm{d}A^\mu$ is known.

Exercise 2.9: Prove that

$$\Gamma^\mu_{\ \nu\mu} \;=\; g_{,\nu}/2g \;=\; \tfrac{1}{2}\big(\log|g|\big)_{,\nu} \quad ; \qquad g \;=\; \det\{g_{\alpha\beta}\} \ . \tag{2.33}$$

Hint: (2.24): $\Gamma^\mu_{\ \nu\mu} = \frac{1}{2}g^{\lambda\mu}g_{\lambda\mu,\nu}$. For a matrix M we have that $\mathrm{Tr}\,(M^{-1}M_{,\nu}) = (\mathrm{Tr}\,\log M)_{,\nu} = (\log\det M)_{,\nu}$. Take $M = \{g_{\alpha\beta}\}$.

2.5 Geodesics

Intuitively, a geodesic is a line that is 'as straight as possible' on a curved surface. We say that a curve $x^\mu(p)$ is a geodesic when the tangent vector $\mathrm{d}x^\mu/\mathrm{d}p$ remains a tangent vector under parallel transport along $x^\mu(p)$. Therefore $\dot{x}^\mu \equiv \mathrm{d}x^\mu/\mathrm{d}p$ must satisfy (2.28), and we arrive at the geodesic equation:

$$\frac{\mathrm{D}}{\mathrm{D}p}\left(\frac{\mathrm{d}x^\mu}{\mathrm{d}p}\right) \;=\; 0 \quad\rightarrow\quad \ddot{x}^\mu \;+\; \Gamma^\mu_{\ \nu\sigma}\,\dot{x}^\nu\,\dot{x}^\sigma \;=\; 0\ , \tag{2.34}$$

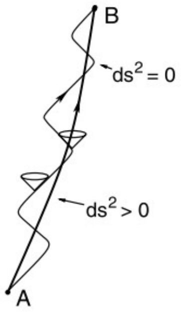

Fig. 2.6. A timelike geodesic connecting events A and B is the curve with the *maximum* possible interval length between A and B, see text.

with $\dot{} = \mathrm{d}/\mathrm{d}p$. For timelike geodesics[2] the parameter p in (2.34) is proportional to the interval length s. Proof: according to exercise 2.6 the length of $\dot{x}^\alpha = \mathrm{d}x^\alpha/\mathrm{d}p$ is constant along $x^\mu(p)$, i.e. $\dot{x}^\alpha \dot{x}_\alpha = g_{\alpha\beta}\, \dot{x}^\alpha\, \dot{x}^\beta \equiv (\mathrm{d}s/\mathrm{d}p)^2$ is constant. For timelike geodesics $\mathrm{d}s^2 > 0$, and we may take the square root to conclude that $\mathrm{d}s = \mathrm{const} \cdot \mathrm{d}p$. Later, when GR is cast into the geometrical framework developed here, this result will be connected to proper time (a physical concept that does not yet exist here): $\mathrm{d}s = c\mathrm{d}\tau$, so that

$$\mathrm{d}p \ \propto\ \mathrm{d}s \ \propto\ \mathrm{d}\tau \qquad \text{for timelike geodesics.} \tag{2.35}$$

This is important as it implies that we may, for timelike geodesics, replace the curve parameter p in (2.34) by the interval length s or the proper time τ.

Intuitively, a geodesic is also the shortest possible route between two points. For a positive definite metric this is indeed the case, but $\mathrm{d}s^2$ can be positive as well as negative in GR. Assuming that the interval $\int \mathrm{d}s = \int \dot{s}\, \mathrm{d}p$ of a timelike geodesic is an extremum (see below), it is easy to see that it should be a maximum: there always exists an arbitrarily nearby worldline that has a smaller $\int \mathrm{d}s$, by letting it jump more or less from light-cone to light-cone, as in Fig. 2.6 (see e.g. Wald (1984) § 9.3). The construction of Fig. 2.6 fails for spacelike geodesics.

---

[2] In an analogy with (1.4) we speak of a timelike (spacelike) worldline or geodesic when $\mathrm{d}s^2 > 0$ ($\mathrm{d}s^2 < 0$). A null worldline or null geodesic has $\mathrm{d}s^2 = 0$. For spacelike and null geodesics p can no longer be interpreted as an interval length.

Eq. (2.34) may also be derived from a variational principle.[3] The simplest is $\delta\int \dot{s}\,\mathrm{d}p = 0$, and this is equivalent to $\delta\int F(\dot{s})\,\mathrm{d}p = 0$ provided F is monotonous, $F' \neq 0$. We choose $\delta\int \dot{s}^2\,\mathrm{d}p = 0$, or

$$\delta\int L\,\mathrm{d}p = 0\ ; \quad L(x^\alpha, \dot{x}^\beta) = (\mathrm{d}s/\mathrm{d}p)^2 = g_{\alpha\beta}\,\dot{x}^\alpha\,\dot{x}^\beta\ . \tag{2.36}$$

The solution is determined by the Euler-Lagrange equations (Appendix C)

$$\frac{\partial L}{\partial x^\lambda} = \frac{\mathrm{d}}{\mathrm{d}p}\left(\frac{\partial L}{\partial \dot{x}^\lambda}\right)\ . \tag{2.37}$$

Now, $\partial L/\partial x^\lambda = g_{\alpha\beta,\lambda}\,\dot{x}^\alpha\dot{x}^\beta$ because only $g_{\alpha\beta}$ depends on $\{x^\mu\}$. By using $\partial\dot{x}^\alpha/\partial\dot{x}^\lambda = \delta^\alpha_\lambda$ one gets $\partial L/\partial\dot{x}^\lambda = 2g_{\alpha\lambda}\,\dot{x}^\alpha$. Substitute this in (2.37):

$$\begin{aligned} g_{\alpha\beta,\lambda}\,\dot{x}^\alpha\,\dot{x}^\beta &= 2(g_{\alpha\lambda}\,\dot{x}^\alpha)^{\cdot} \\ &= 2(g_{\alpha\lambda,\beta}\,\dot{x}^\beta\,\dot{x}^\alpha + g_{\alpha\lambda}\,\ddot{x}^\alpha)\ , \end{aligned}$$

or

$$g_{\alpha\lambda}\,\ddot{x}^\alpha + \tfrac{1}{2}(2g_{\lambda\alpha,\beta} - g_{\alpha\beta,\lambda})\,\dot{x}^\alpha\,\dot{x}^\beta = 0\ . \tag{2.38}$$

Now comes a frequently used trick: renaming of dummy indices: $2g_{\lambda\alpha,\beta}\cdot\dot{x}^\alpha\,\dot{x}^\beta = g_{\lambda\alpha,\beta}\,\dot{x}^\alpha\,\dot{x}^\beta + g_{\lambda\beta,\alpha}\,\dot{x}^\beta\,\dot{x}^\alpha = (g_{\lambda\alpha,\beta} + g_{\lambda\beta,\alpha})\,\dot{x}^\alpha\,\dot{x}^\beta$. Substitution in (2.38) and multiplication with $g^{\mu\lambda}$ gives:

$$\ddot{x}^\mu + \tfrac{1}{2}g^{\mu\lambda}\,(g_{\lambda\alpha,\beta} + g_{\lambda\beta,\alpha} - g_{\alpha\beta,\lambda})\,\dot{x}^\alpha\,\dot{x}^\beta = 0\ . \tag{2.39}$$

This is of the form of (2.34) and the factor multiplying $\dot{x}^\alpha\,\dot{x}^\beta$ must be equal to $\Gamma^\mu_{\ \alpha\beta}$, which proves (2.24). Variational calculus is a very efficient tool for this type of problem. Without much difficulty, it permits us to find the geodesic equation directly from the metric, and from this equation one may just read the Christoffel symbols $\Gamma^\mu_{\ \nu\sigma}$. This is usually a lot faster than calculating them from (2.24), and this method is therefore highly recommended.

The following result is very helpful when analysing the dynamics of a test particle in GR (assuming that its orbit is a geodesic), because it allows us to find *constants of the motion.* From the text below (2.37) we see that $\partial L/\partial x^\lambda$ vanishes if $g_{\alpha\beta,\lambda} = 0$. And then eq. (2.37) says that $\partial L/\partial\dot{x}^\lambda = 2g_{\alpha\lambda}\dot{x}^\alpha$ is constant. In terms of the *4-velocity* $u^\mu = \mathrm{d}x^\mu/\mathrm{d}p$ we have found that the covariant 4-velocity $u_\lambda = g_{\lambda\alpha}u^\alpha$ is constant:

$$g_{\alpha\beta,\lambda} = 0 \quad\rightarrow\quad u_\lambda = g_{\lambda\nu}\,\dot{x}^\nu = \text{constant} \tag{2.40}$$

with $^{\cdot} = \mathrm{d}/\mathrm{d}p$. The fact that u_λ is a constant along a geodesic if the metric is independent of x^λ – doesn't that ring a bell?

---

[3] Here we switch to another definition of geodesics without proving its equivalence with (2.34).

Exercise 2.10: Show that the geodesics of the Lorentz metric ($g_{\alpha\beta} = \eta_{\alpha\beta}$) are straight lines.

Exercise 2.11: Show that the variational problem (2.36) is equivalent to $\delta \int F(L)\, \mathrm{d}p = 0$ if F is monotonous, $F' \neq 0$.

Hint: Write down (2.37) with $L \to F(L)$; use $\partial F(L)/\partial x^\lambda = F'\, \partial L/\partial x^\lambda$, and $(F'\, \partial L/\partial \dot{x}^\lambda)^\cdot = (F')^\cdot\, \partial L/\partial \dot{x}^\lambda + F'(\partial L/\partial \dot{x}^\lambda)^\cdot$. But $(F')^\cdot = F''\mathrm{d}L/\mathrm{d}p = 0$ (L is constant on $x^\mu(p)$ because $\dot{x}^\alpha \dot{x}_\alpha$ is).

2.6 The covariant derivative

For a given vector field A^μ that is not restricted to the curve $x^\sigma(p)$ we can elaborate $\mathrm{d}A^\mu/\mathrm{d}p$ in (2.26) as $\mathrm{d}A^\mu/\mathrm{d}p = A^\mu{}_{,\sigma}\, \dot{x}^\sigma$, because we are able to compute derivatives in other directions than along the curve. This leads to the introduction of the covariant derivative

$$\frac{\mathrm{D}A^\mu}{\mathrm{D}p} = \left(A^\mu{}_{,\sigma} + \Gamma^\mu_{\nu\sigma}\, A^\nu\right) \dot{x}^\sigma \equiv A^\mu{}_{:\sigma}\, u^\sigma \ , \tag{2.41}$$

where $u^\sigma = \dot{x}^\sigma = \mathrm{d}x^\sigma/\mathrm{d}p$ and

$$A^\mu{}_{:\sigma} \equiv A^\mu{}_{,\sigma} + \Gamma^\mu_{\nu\sigma}\, A^\nu \tag{2.42}$$

is the *covariant derivative* of A^μ. It may be regarded as the 'intrinsic derivative', the derivative after correction for the meaningless change in orientation of the base vectors. In a similar way we may obtain the covariant derivative of a covariant vector from (2.27):

$$A_{\mu:\sigma} = A_{\mu,\sigma} - \Gamma^\nu_{\mu\sigma}\, A_\nu \ . \tag{2.43}$$

Important is that both $A^\mu{}_{:\sigma}$ and $A_{\mu:\sigma}$ are tensors if A^μ is a vector, even though neither of the two terms on the right hand sides of (2.42) and (2.43) are tensors themselves. The proof is a matter of combining relations (2.20) and (2.31), and is left to the reader.

Next follow a few definitions. The covariant derivative of a product XY of two tensors is:

$$(XY)_{:\sigma} = X_{:\sigma}\, Y + X\, Y_{:\sigma} \ . \tag{2.44}$$

For example:

$$
\begin{aligned}
(A_\mu B_\nu)_{:\sigma} &= (A_{\mu,\sigma} - \Gamma^\alpha_{\mu\sigma} A_\alpha) B_\nu + A_\mu (B_{\nu,\sigma} - \Gamma^\alpha_{\nu\sigma} B_\alpha) \\
&= (A_\mu B_\nu)_{,\sigma} - \Gamma^\alpha_{\mu\sigma} A_\alpha B_\nu - \Gamma^\alpha_{\nu\sigma} A_\mu B_\alpha \,. \qquad (2.45)
\end{aligned}
$$

Accordingly, we define the covariant derivative of a covariant second rank tensor as:

$$T_{\mu\nu:\sigma} = T_{\mu\nu,\sigma} - \Gamma^\alpha_{\mu\sigma} T_{\alpha\nu} - \Gamma^\alpha_{\nu\sigma} T_{\mu\alpha} \,. \qquad (2.46)$$

The recipe for tensors of higher rank should be clear by now. For example, if we need an expression for $T_\alpha{}^\beta{}_{\gamma:\sigma}$, then we merely have to work out $(P_\alpha Q^\beta R_\gamma)_{:\sigma}$ as in (2.44) and (2.45). The general pattern is $T^*_{*:\sigma} = T^*_{*,\sigma} \pm \Gamma$-term for every index. For a scalar:

$$Q_{:\sigma} = Q_{,\sigma} \,. \qquad (2.47)$$

Covariant derivatives do not commute, unlike normal derivatives ($X_{,\alpha\beta} = X_{,\beta\alpha}$ for every X). We calculate $B_{\mu:\nu:\sigma}$ by substituting $T_{\mu\nu} = B_{\mu:\nu}$ in (2.46):

$$B_{\mu:\nu:\sigma} = B_{\mu:\nu,\sigma} - \Gamma^\alpha_{\mu\sigma} B_{\alpha:\nu} - \Gamma^\alpha_{\nu\sigma} B_{\mu:\alpha} \,, \qquad (2.48)$$

which should be elaborated further with (2.43). After that, interchange ν and σ and subtract. The result of a somewhat lengthy calculation is:

$$B_{\mu:\nu:\sigma} - B_{\mu:\sigma:\nu} = B_\alpha R^\alpha{}_{\mu\nu\sigma} \qquad (2.49)$$

with

$$R^\alpha{}_{\mu\nu\sigma} = \Gamma^\alpha_{\mu\sigma,\nu} - \Gamma^\alpha_{\mu\nu,\sigma} + \Gamma^\tau_{\mu\sigma} \Gamma^\alpha_{\tau\nu} - \Gamma^\tau_{\mu\nu} \Gamma^\alpha_{\tau\sigma} \,. \qquad (2.50)$$

$R^\alpha{}_{\mu\nu\sigma}$ is called the RIEMANN tensor. It is a tensor because (2.49) is valid for every vector B_α and because the left hand side is a tensor. Then apply the quotient theorem. Apparently, covariant derivatives commute only if $R^\alpha{}_{\mu\nu\sigma} = 0$. The Riemann tensor plays a crucial role in GR because it contains all information about the curvature of space. Note the remarkable fact that according to (2.49) the difference of two consecutive covariant differentiations is proportional to the vector itself. The explanation is given in the next section.

Exercise 2.12: Show that

$$T^{\mu\nu}{}_{:\sigma} = T^{\mu\nu}{}_{,\sigma} + \Gamma^\mu_{\alpha\sigma} T^{\alpha\nu} + \Gamma^\nu_{\alpha\sigma} T^{\mu\alpha} \,. \qquad (2.51)$$

Great care is needed in using these relations. For example, let $T^{\mu\nu}$ be diagonal. Then it seems evident that $T^{1\mu}{}_{:\mu} = T^{11}{}_{:1}$, but that is not the case. Why not?

Hint: Write out $(A^\mu B^\nu)_{:\sigma}$ as in (2.45). It is due to the action of the invisible dummy index α.

Exercise 2.13: An important property is that the metric tensor behaves as a constant under covariant differentiation:

$$g_{\mu\nu:\sigma} = 0 \ . \tag{2.52}$$

Hint: Use (2.46) and (2.30).

Exercise 2.14: Prove the following compact form of the geodesic equation:

$$u^\sigma u_{\mu:\sigma} = 0 \qquad \text{or} \qquad u^\sigma u^\mu{}_{:\sigma} = 0 \ . \tag{2.53}$$

Hint: The last relation is just $0 = \mathrm{D}u^\mu/\mathrm{D}p = (2.41)$; the first relation with (2.52): $0 = g_{\lambda\mu} u^\mu{}_{:\sigma} u^\sigma = (g_{\lambda\mu} u^\mu)_{:\sigma} u^\sigma =$ etc.

Exercise 2.15: A reminder of the linear algebra aspects of tensor calculus. Given a 2D Riemann space with co-ordinates x, y, a metric and two vectors in the tangent space of the point (x, y):

$$\mathrm{d}s^2 = \mathrm{d}x^2 + 4\mathrm{d}x\mathrm{d}y + \mathrm{d}y^2 \ ; \quad A^\alpha = \begin{pmatrix} 1 \\ 4 \end{pmatrix} \ ; \quad B_\alpha = \begin{pmatrix} y \\ x \end{pmatrix} \ .$$

Write down $g_{\mu\nu}$ and $g^{\mu\nu}$ and show that all Christoffel symbols are zero. Compute A_ν and $B^\nu{}_{:\nu}$.

Hint: $g_{11} = g_{22} = 1$; $g_{12} = g_{21} = 2$, use (2.24) for the Christoffel symbols;

$$g^{\mu\nu} = \frac{1}{3} \begin{pmatrix} -1 & 2 \\ 2 & -1 \end{pmatrix} \ ; \quad A_\mu = \begin{pmatrix} 9 \\ 6 \end{pmatrix} \ ; \quad B^\nu{}_{:\nu} = \frac{4}{3} \ .$$

The Γ's being zero we have $B^\nu{}_{:\nu} = B^\nu{}_{,\nu}$.

2.7 Riemann tensor and curvature

The metric tensor does not tell us whether a space is flat, because the use of 'strange' co-ordinates is not prohibited. For example $\mathrm{d}s^2 = \mathrm{d}r^2 + r^2\mathrm{d}\varphi^2$

(planar polar co-ordinates) defines a flat space, but (2.2) defines a curved space. The metric tensor contains apparently a mix of information on co-ordinates and curvature. The intrinsic curvature properties are determined by the Riemann tensor. We shall illustrate this by transporting a vector A^μ parallel to itself along two different paths to the same final position, see Fig. 2.7. According to (2.26), $\mathrm{d}A^\mu = -f_\sigma(x)\mathrm{d}x^\sigma$ with $f_\sigma(x) = \Gamma^\mu{}_{\nu\sigma}A^\nu$ (the upper index μ is omitted for brevity as it does not change). The difference of the two final vectors is:

$$\begin{aligned}
\mathrm{d}A^\mu &= A_1^\mu - A_2^\mu \\
&= -f_\sigma(x)\mathrm{d}\xi^\sigma - f_\sigma(x+\mathrm{d}\xi)\mathrm{d}\eta^\sigma + f_\sigma(x)\mathrm{d}\eta^\sigma + f_\sigma(x+\mathrm{d}\eta)\mathrm{d}\xi^\sigma \\
&\simeq -f_\sigma\mathrm{d}\xi^\sigma - f_\sigma\mathrm{d}\eta^\sigma - f_{\sigma,\lambda}\,\mathrm{d}\xi^\lambda\mathrm{d}\eta^\sigma + f_\sigma\mathrm{d}\eta^\sigma + f_\sigma\mathrm{d}\xi^\sigma + f_{\sigma,\lambda}\,\mathrm{d}\eta^\lambda\mathrm{d}\xi^\sigma \\
&= (f_{\sigma,\lambda} - f_{\lambda,\sigma})\,\mathrm{d}\xi^\sigma\mathrm{d}\eta^\lambda\ . \qquad (2.54)
\end{aligned}$$

Now substitute $f_\sigma = \Gamma^\mu{}_{\nu\sigma}A^\nu = A^\mu{}_{:\sigma} - A^\mu{}_{,\sigma}$. The terms $A^\mu{}_{,\sigma}$ cancel, and after some index gymnastics we arrive at (exercise 2.16):

$$\begin{aligned}
\mathrm{d}A^\mu &= (A^\mu{}_{:\sigma,\lambda} - A^\mu{}_{:\lambda,\sigma})\,\mathrm{d}\xi^\sigma\mathrm{d}\eta^\lambda \\
&= g^{\mu\nu}(A_{\nu:\sigma,\lambda} - A_{\nu:\lambda,\sigma})\,\mathrm{d}\xi^\sigma\mathrm{d}\eta^\lambda \\
&= g^{\mu\nu}(A_{\nu:\sigma:\lambda} - A_{\nu:\lambda:\sigma})\,\mathrm{d}\xi^\sigma\mathrm{d}\eta^\lambda \\
&= g^{\mu\nu}R_{\alpha\nu\sigma\lambda}\,A^\alpha\,\mathrm{d}\xi^\sigma\mathrm{d}\eta^\lambda \\
&= g^{\mu\nu}R_{\nu\alpha\lambda\sigma}\,A^\alpha\,\mathrm{d}\xi^\sigma\mathrm{d}\eta^\lambda \\
&= R^\mu{}_{\alpha\lambda\sigma}\,A^\alpha\,\mathrm{d}\xi^\sigma\mathrm{d}\eta^\lambda\ . \qquad (2.55)
\end{aligned}$$

On account of (2.24) the Christoffel symbols vanish identically in a flat space with rectangular co-ordinates, since $g_{\mu\nu}$ has only constant elements. Therefore the Riemann tensor (2.50) is zero as well. The transformation properties of a tensor then ensure that $R^\alpha{}_{\mu\nu\sigma}$ is zero in a flat space for any choice of the co-ordinates.[4] In that case parallel transport along a closed path leaves a vector unchanged.[5] But in a curved space the orientation of the vector will have

---

[4] Contrary to the Christoffel symbols, which are not tensors. For example, the Christoffel symbols vanish in rectangular co-ordinates in a plane, but *not* in polar co-ordinates.

[5] Conversely, if the Riemann tensor is zero, it can be proven that there exist co-ordinates so that $g_{\mu\nu}$ is constant which implies that the space is flat, see e.g. Dirac (1975) § 12.

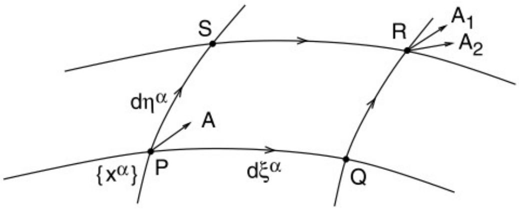

Fig. 2.7. Parallel transport of the vector ***A*** from P to R along path 1 (PQR) and path 2 (PSR) produces a different result.

changed. Once this is accepted intuitively, it is clear that the difference $\mathrm{d}A^{\mu}$ must be proportional to the length of the vector, which explains the factor A^{α} in (2.55). The derivation in (2.55) shows that the difference $\mathrm{d}A^{\mu}$ is also proportional to the difference of two consecutive covariant differentiations, and this explains why this difference is proportional to the vector itself, as in (2.49).

There are several other ways to illustrate the relation between the Riemann tensor and curvature. One is the equation for the *geodesic deviation*, see exercise 2.18. Another is the relation between *Gaussian curvature* and the Riemann tensor. Gaussian curvature refers to surfaces embedded in a flat 3D space. The curvature κ in a point P of a curve on the surface is defined as the inverse radius of the osculating circle at P. Each point has two principal curvatures κ_1 and κ_2, and the Gaussian curvature $K \equiv \kappa_1\kappa_2$ is an invariant determined by the geometry of the surface, which has several interesting properties.[6] Turning now to Riemann spaces, take two orthogonal unit vectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ in the tangent space of a point P which are not null. Now consider those geodesics in Riemann space that are tangent in P to the plane spanned by $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$. These geodesics subtend, locally around P, a 2D curved subspace of Riemann space. The Gaussian curvature of this 2D space at P is $R_{\alpha\mu\nu\sigma}e_1^{\alpha}e_2^{\mu}e_1^{\nu}e_2^{\sigma}$, apart from the sign.[7]

The Riemann tensor obeys several symmetry relations that reduce the number of independent components from n^4 to $n^2(n^2-1)/12$ (see literature). In 4 dimensions $R^{\alpha}{}_{\nu\rho\sigma}$ has only 20 independent components, and all contractions of $R^{\alpha}{}_{\nu\rho\sigma}$ are either zero or equal, apart from the sign. We choose the *Ricci tensor*:[8]

---

[6] E.g. Gauss's theorem on integral curvature: the sum of the three interior angles of a geodesic triangle (bounded by 3 geodesics) equals π plus the surface integral of K.

[7] For a proof of these statements see e.g. Robertson and Noonan (1969) p. 216.

$$R_{\mu\nu} \equiv R^{\alpha}{}_{\mu\nu\alpha} \qquad \text{(RICCI)}. \tag{2.56}$$

The explicit expression follows from (2.50):

$$R_{\mu\nu} \;=\; \Gamma^{\alpha}_{\mu\alpha,\nu} \;-\; \Gamma^{\alpha}_{\mu\nu,\alpha} \;-\; \Gamma^{\alpha}_{\mu\nu}\,\Gamma^{\beta}_{\alpha\beta} \;+\; \Gamma^{\alpha}_{\mu\beta}\,\Gamma^{\beta}_{\nu\alpha}\;. \tag{2.57}$$

We infer from (2.33) that $\Gamma^{\alpha}_{\mu\alpha,\nu} = \frac{1}{2}\big(\log|g|\big)_{,\mu\nu}$ so that all terms in (2.57) are symmetric in μ and in ν. Hence $R_{\mu\nu}$ is symmetric:

$$R_{\mu\nu} = R_{\nu\mu}\;. \tag{2.58}$$

We may contract once more:

$$R \;\equiv\; R^{\nu}{}_{\nu} = g^{\nu\mu}R_{\mu\nu} = R^{\alpha\beta}{}_{\beta\alpha}\;. \tag{2.59}$$

R is called the *total curvature*. Finally we introduce the *Einstein tensor* $G_{\mu\nu}$:

$$G_{\mu\nu} \;=\; R_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}R \qquad \text{(EINSTEIN)}. \tag{2.60}$$

The Einstein tensor will be useful later because its divergence is zero:

$$G^{\mu\nu}{}_{:\nu} \;=\; (R^{\mu\nu} - \tfrac{1}{2}g^{\mu\nu}R)_{:\nu} \;=\; 0\;. \tag{2.61}$$

Riemann, Ricci en Einstein tensor contain at most second derivatives of $g_{\alpha\beta}$. By substituting (2.24) in (2.50) we get:

$$\begin{aligned} R^{\alpha}{}_{\mu\nu\sigma} \;=\; &\tfrac{1}{2}g^{\alpha\beta}\left(g_{\beta\sigma,\mu\nu} \;-\; g_{\mu\sigma,\beta\nu} \;-\; g_{\beta\nu,\mu\sigma} \;+\; g_{\mu\nu,\beta\sigma}\right) \\ &+\; g^{\alpha\beta}\left(\Gamma_{\tau\beta\sigma}\,\Gamma^{\tau}_{\mu\nu} \;-\; \Gamma_{\tau\beta\nu}\,\Gamma^{\tau}_{\mu\sigma}\right)\;. \end{aligned} \tag{2.62}$$

The corresponding expressions for $R_{\mu\nu}$ and for $G_{\mu\nu}$ can be found from this by contraction. The first term contains all second-order derivatives. The first-order derivatives are in the second term. The proofs of (2.61) and (2.62) can be found in the literature, but are not important here.

Exercise 2.16: Provide the missing details of the derivation of (2.55).

Hint: Second = sign: $A^{\mu}{}_{:\sigma,\lambda} \;=\; (g^{\mu\nu}A_{\nu})_{:\sigma,\lambda} \;=\; (g^{\mu\nu}A_{\nu:\sigma})_{,\lambda} \;=\; g^{\mu\nu}{}_{,\lambda}\,A_{\nu:\sigma} \;+\; g^{\mu\nu}A_{\nu:\sigma,\lambda}$, but A^{μ} is parallel transported, hence $A_{\nu:\sigma} = 0$, etc. Third = sign: $A_{\nu:\sigma:\lambda} \;=\; A_{\nu:\sigma,\lambda}$ from (2.48). Fifth = sign: $R_{\alpha\nu\sigma\lambda} \;=\; R_{\nu\alpha\lambda\sigma}$ is a symmetry relation of the Riemann tensor.

---

[8] Other authors define $R_{\mu\nu} \;=\; R^{\alpha}{}_{\mu\alpha\nu}$, another source of sign differences. For a complete classification of all sign conventions see the red pages in Misner et al. (1971). In terms of this classification we follow the $-\;+\;-$ convention.

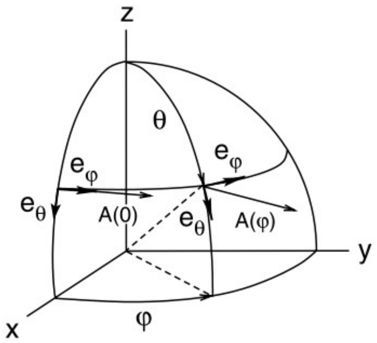

Fig. 2.8. Parallel transport of a vector $\boldsymbol{A}$ over the surface of a sphere with radius $r = 1$, see exercise 2.17.

Exercise 2.17: Consider a 2D spherical surface with radius $r = 1$, see Fig. 2.8. Calculate the Christoffel symbols and the total curvature R. Convince yourself that $R \propto r^{-2}$. Show that a vector $\boldsymbol{A}$ will rotate in the tangent space as it is parallel-transported along a circle $\theta = \theta_0$. Try to understand this with the intuitive definition of parallel transport in § 2.4. Start in $(\theta,\, \varphi) = (\theta_0,\, 0)$ with $(A^\theta,\, A^\varphi) = (0, 1/\sin\theta_0)$. Show that $A^i A_i$ is always 1, i.e. $|\boldsymbol{A}| \equiv 1$, and that after one full revolution $\boldsymbol{A}$ has rotated over an angle $2\pi\cos\theta_0$. Discuss the limiting cases $\theta_0 = \pi/2$ (geodesic!) and $\theta_0 \ll 1$.

Hint: (2.2): $g_{11} = 1,\ g_{22} = \sin^2\theta\ (\theta = 1,\ \varphi = 2)$. Do *not* use (2.24), but rather (2.37) with $L(\theta, \dot\theta, \dot\varphi) = \dot\theta^2 + \sin^2\theta\,\dot\varphi^2$:

$$\begin{aligned} \frac{\partial L}{\partial \theta} &= \Big(\frac{\partial L}{\partial \dot\theta}\Big)^{\cdot} \ \rightarrow\ \ddot\theta \,-\, \sin\theta\cos\theta\,\dot\varphi^2 \,=\, 0\ ; \\ \frac{\partial L}{\partial \varphi} &= \Big(\frac{\partial L}{\partial \dot\varphi}\Big)^{\cdot} \ \rightarrow\ \ddot\varphi \,+\, 2\,\cot\theta\,\dot\theta\dot\varphi \,=\, 0\ . \end{aligned}$$

By comparing with (2.34) we may just read the Γ's: ${\Gamma^1}_{22} = -\sin\theta\cos\theta$; ${\Gamma^2}_{12} = \cot\theta$ (double product!). All other Γ's are zero. (2.33) → ${\Gamma^\alpha}_{\mu\alpha,\nu} = (\log\sin\theta)_{,\mu\nu} \ \rightarrow\ {\Gamma^\alpha}_{1\alpha,1} = -1/\sin^2\theta$. And ${\Gamma^\alpha}_{11,\alpha} = 0$; ${\Gamma^\alpha}_{22,\alpha} = -(\sin\theta \cdot \cos\theta)_{,\theta} = \sin^2\theta - \cos^2\theta$. Algebra: $R_{11} = -1$ and $R_{22} = -\sin^2\theta$. Finally $R = g^{\mu\nu}R_{\mu\nu} = g^{11}R_{11} + g^{22}R_{22} = R_{11} + (1/\sin^2\theta)\,R_{22} = -2$. For a sphere with radius r: $R = -2/r^2$ (minus sign due to sign convention).
Parallel transport: p is proportional to the arc length (why?), so choose $p = \varphi$; (2.28)+(2.26): ${A^\mu}_{,\varphi} + {\Gamma^\mu}_{\nu\sigma}\,A^\nu {x^\sigma}_{,\varphi} = 0$ with $x^1_{,\varphi} = \mathrm{d}\theta/\mathrm{d}\varphi = 0$ and $x^2_{,\varphi} = \mathrm{d}\varphi/\mathrm{d}\varphi = 1$:

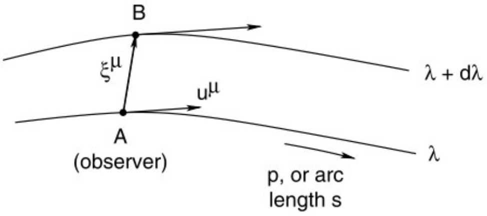

Fig. 2.9. The geodesic deviation.

$$A^{\theta}_{\ ,\varphi} = \sin\theta_0\cos\theta_0\,A^{\varphi}\ ; \qquad A^{\varphi}_{\ ,\varphi} = -\cot\theta_0\,A^{\theta}\ .$$

Eliminate A^{φ}: $A^{\theta}_{\ ,\varphi\varphi} + \cos^2\theta_0\,A^{\theta} = 0$, same equation holds for A^{φ}. Harmonic oscillator with frequency $\cos\theta_0$. Solution for given initial value:

$$A^{\theta} = \sin(\varphi\cos\theta_0)\ ; \qquad A^{\varphi} = \cos(\varphi\cos\theta_0)/\sin\theta_0\ .$$

$\boldsymbol{A}$ rotates clockwise when looking down on the tangent space from outside; $\theta_0 = \pi/2$: $A^{\theta} \equiv 0$ and $A^{\varphi} \equiv 1/\sin\theta_0 = 1$, therefore $\boldsymbol{A}$ remains a tangent vector; $\theta_0 \ll 1$ (small circle around the north pole): in that case the tangent space is always almost parallel to the equatorial plane, with base vectors $\boldsymbol{x}$ en $\boldsymbol{y}$, and $\boldsymbol{e}_{\theta} \simeq \boldsymbol{x}\cos\varphi + \boldsymbol{y}\sin\varphi$ and $\boldsymbol{e}_{\varphi} \simeq (\boldsymbol{y}\cos\varphi - \boldsymbol{x}\sin\varphi)\sin\theta_0$. For $\theta_0 \ll 1$ it follows that $\boldsymbol{A} = A^{\theta}\boldsymbol{e}_{\theta} + A^{\varphi}\boldsymbol{e}_{\varphi} \simeq \boldsymbol{y}$, so that $\boldsymbol{A}$ remains virtually unchanged with respect to a fixed frame.

Exercise 2.18: Given a set of geodesics $x^{\mu}(p,\lambda)$ where p is the curve parameter and λ labels different geodesics (λ is constant along one geodesic). Consider two neighbouring geodesics λ and $\lambda + \delta\lambda$. The points A and B are connected by the vector $\xi^{\mu} = x^{\mu}(p,\lambda+\delta\lambda) - x^{\mu}(p,\lambda) \simeq (\partial x^{\mu}/\partial\lambda)\delta\lambda \equiv e^{\mu}\delta\lambda$. Prove that:

$$\frac{\mathrm{D}^2\xi^{\mu}}{\mathrm{D}p^2} = R^{\mu}_{\ \alpha\beta\nu}\,u^{\alpha}u^{\beta}\xi^{\nu}\ ; \qquad u^{\alpha} = \dot{x}^{\alpha} = \frac{\partial x^{\alpha}}{\partial p}\ . \tag{2.63}$$

This is the equation for the geodesic deviation, that will play an important role later. In a flat space the Riemann tensor is zero, and then ξ^{μ} is a linear function of p, and for timelike geodesics also a linear function of the arc length s, as expected. In a curved space however this is no longer the case. For example, on a sphere $\xi^{\mu}(s)$ will be something like a sine-function.

Hint: The proof comes in three steps:

$$\text{(a)}\qquad \frac{\partial e^\mu}{\partial p} = \frac{\partial^2 x^\mu}{\partial p\,\partial\lambda} = \frac{\partial u^\mu}{\partial \lambda} = u^\mu{}_{,\alpha}\,\frac{\partial x^\alpha}{\partial\lambda} = u^\mu{}_{,\alpha}\, e^\alpha\ ;$$

$$\begin{aligned}
\text{(b)}\ \ e^\mu{}_{:\alpha}\, u^\alpha &\equiv \frac{\mathrm{D}e^\mu}{\mathrm{D}p} = \frac{\partial e^\mu}{\partial p} + \Gamma^\mu_{\ \alpha\beta}\, e^\alpha u^\beta \\
&= u^\mu{}_{,\alpha}\, e^\alpha + \Gamma^\mu_{\ \alpha\beta}\, e^\alpha u^\beta = u^\mu{}_{:\alpha}\, e^\alpha\ ;
\end{aligned}$$

$$\begin{aligned}
\text{(c)}\qquad \frac{\mathrm{D}^2 e^\mu}{\mathrm{D}p^2} &\equiv (e^\mu{}_{:\alpha}\, u^\alpha)_{:\beta}\, u^\beta = (u^\mu{}_{:\alpha}\, e^\alpha)_{:\beta}\, u^\beta \\
&= u^\mu{}_{:\alpha}\, e^\alpha{}_{:\beta}\, u^\beta + u^\mu{}_{:\alpha:\beta}\, e^\alpha u^\beta \\
&= u^\mu{}_{:\alpha}\, u^\alpha{}_{:\beta}\, e^\beta + u^\mu{}_{:\alpha:\beta}\, u^\alpha e^\beta + \left(u^\mu{}_{:\alpha:\beta} - u^\mu{}_{:\beta:\alpha}\right) u^\beta e^\alpha \\
&= \left(u^\mu{}_{:\alpha}\, u^\alpha\right)_{:\beta} e^\beta + g^{\mu\nu}\left(u_{\nu:\alpha:\beta} - u_{\nu:\beta:\alpha}\right) u^\beta e^\alpha \\
&= g^{\mu\nu}\, u_\sigma\, R^\sigma{}_{\nu\alpha\beta}\, u^\beta e^\alpha \\
&= R_\sigma{}^\mu{}_{\alpha\beta}\, u^\sigma u^\beta e^\alpha \\
&= R^\mu{}_{\sigma\beta\alpha}\, u^\sigma u^\beta e^\alpha\ .
\end{aligned}$$

In (c) we have twice used (b), next $u^\mu{}_{:\beta:\alpha}\, e^\alpha u^\beta = u^\mu{}_{:\alpha:\beta}\, e^\beta u^\alpha$ is added and substracted again, and then (2.53) and (2.49). The last = sign is a symmetry relation of the Riemann tensor. Because $\delta\lambda$ is constant, the equation also holds for $\xi^\mu = e^\mu \delta\lambda$.

Exercise 2.19: Be aware of some inconsistencies in the notation. We encountered one in exercise 2.12. Meet two more here. In § 2.2 and § 2.3 it was stressed that the rules for index raising and lowering are always valid. Does that mean that

$$g^{\mu\alpha} g_{\alpha\lambda,\nu} \overset{?}{=} g^\mu_{\ \lambda,\nu}\ ; \tag{2.64}$$

$$g_{\mu\alpha}\dot{u}^\alpha \overset{?}{=} \dot{u}_\mu\ . \tag{2.65}$$

Hint: In exercise 2.12 the trouble was caused by a hidden index; here we discover that the symbols without derivative had already been defined; one way to see that (2.64) cannot be correct is to note that $g^\mu_{\ \lambda,\nu} \equiv \delta^\mu_{\ \lambda,\nu} = 0$, and since $\det\{g^{\mu\alpha}\} \neq 0 \rightarrow g_{\alpha\lambda,\nu} = 0 \ \rightarrow g_{\alpha\lambda} = \text{const}$. Instead, $0 = (g^{\mu\alpha} g_{\alpha\lambda})_{,\nu} = g^{\mu\alpha}{}_{,\nu}\, g_{\alpha\lambda} + g^{\mu\alpha} g_{\alpha\lambda,\nu}$, etc. Likewise, u_μ is *defined* as $g_{\mu\alpha} u^\alpha$ so that $\dot{u}_\mu = (g_{\mu\alpha} u^\alpha)^{\cdot} = g_{\mu\alpha,\sigma}\, u^\sigma u^\alpha + g_{\mu\alpha}\dot{u}^\alpha$. Also correct is $\dot{u}_\mu = u_{\mu,\alpha}\dot{x}^\alpha = u_{\mu,\alpha} u^\alpha$.

2

Geometry of Riemann Spaces

The fact that the geometry of the space in which we live is Euclidean is a very basic daily experience. This may explain why it took so long before it was realised that this may actually not be correct, and that the question of the geometry of the space around us is a matter of empirical assessment. Early in the 19th century Gauss studied the geometry of curved surfaces, and showed that all references to a flat embedding space could be eliminated. In the same way Riemann formulated in 1854 the geometry of 3D spaces. He found that Euclidean geometry is merely one possibility out of many. Riemann's method could be generalized to spaces of arbitrary dimension. The geometry of these curved Riemann spaces is wholly described within the space itself, by the use of co-ordinates and the metric tensor. No embedding is required. These geometrical concepts gradually spread beyond the mathematical incrowd, and in the last quarter of the 19th century the idea that a fourth (spatial) dimension might exist had mesmerized the public's imagination, perhaps even more so than black holes did a century later. One of the products of that period was Abbott's famous Flatland.[1] The flatland analogy is nowadays a standard technique of teachers to explain some of the intricacies of curved spaces.

The theoretical framework of Riemann spaces is also the starting point for the mathematical formulation of GR. In this chapter we discuss the tools that any student should master in order to be able to deal with GR beyond the level of handwaving. In doing so we have deliberately chosen to stay close to intuition as that outweighs the merits of rigour, certainly on first acquaintance.

2.1 Definition

A Riemann space has the following properties:

---

[1] Abbott, E.A.: 1884, *Flatland: A Romance of many Dimensions, by a Square*, Seeley & Co. (London).

! -./0- - -1 ! - < 2" .?? .?? = ! ! "
D ; 6 $$ 4"

! -./0- - -1 ! - < 2" .?? B E ! ! F 7! 4
= ! ! " 8 !"; $ 4"

!' -./0- - -1 ! - < %! ! > . ! " = ! ! "
. / ! ! 6 @ < 4"

? -./0- - -1 ! - < > . ! !
4" K + 4!
1? -./0- - -1 ! - < > . ! !
4 ! 8 7 " ? < 4" K +' 4!

& < +

In 1947 Nafe, Nelson and Rabi [24] reported an anomalous value by about 0.26% in the hyperfine splitting of hydrogen and deuterium, which was quickly confirmed by Nagle et al. [25], and Breit [26] suggested a possible anomaly $g \neq 2$ of the magnetic moment of the electron. Soon after, Kusch and Foley [27], by a study of the hyperfine–structure of atomic spectra in a constant magnetic field, presented the first precision determination of the magnetic moment of the electron $g_e = 2.00238(10)$ in 1948, just before the theoretical result had been settled. Together with Schwinger's result $a_e^{(2)} = \alpha/(2\pi) \simeq 0.00116$ (which accounts for 99% of the anomaly) this provided one of the first tests of the virtual quantum corrections, usually called radiative corrections, predicted by a relativistic quantum field theory. The discovery of the fine structure of the hydrogen spectrum (Lamb–shift) by Lamb and Retherford [28] and the corresponding calculations by Bethe, Kroll & Lamb and Weisskopf & French [29] was the other triumph of testing the new level of theoretical understanding with precision experiments. These successes had a dramatic impact in establishing quantum field theory as a general framework for the theory of elementary particles and for our understanding of the fundamental interactions. It stimulated the development of QED[7] in particular and the concepts of quantum field theory in general. With the

In 1947 Nafe, Nelson and Rabi [24] reported an anomalous value by about 0.26% in the hyperfine splitting of hydrogen and deuterium, which was quickly confirmed by Nagle et al. [25], and Breit [26] suggested a possible anomaly $g \neq 2$ of the magnetic moment of the electron. Soon after, Kusch and Foley [27], by a study of the hyperfine–structure of atomic spectra in a constant magnetic field, presented the first precision determination of the magnetic moment of the electron $g_e = 2.00238(10)$ in 1948, just before the theoretical result had been settled. Together with Schwinger's result $a_e^{(2)} = \alpha/(2\pi) \simeq 0.00116$ (which accounts for 99% of the anomaly) this provided one of the first tests of the virtual quantum corrections, usually called radiative corrections, predicted by a relativistic quantum field theory. The discovery of the fine structure of the hydrogen spectrum (Lamb–shift) by Lamb and Retherford [28] and the corresponding calculations by Bethe, Kroll & Lamb and Weisskopf & French [29] was the other triumph of testing the new level of theoretical understanding with precision experiments. These successes had a dramatic impact in establishing quantum field theory as a general framework for the theory of elementary particles and for our understanding of the fundamental interactions. It stimulated the development of QED[7] in particular and the concepts of quantum field theory in general. With the advent of non–Abelian gauge theories, proposed by Yang and Mills (YM) [31] in 1954, and after 't Hooft and Veltman [32] found the missing clues to understanding and handling them on the quantum level, many years later in 1971, the SM [33] (Glashow, Weinberg, Salam 1981/1987) finally emerged as a comprehensive theory of weak, electromagnetic and strong interactions. The strong interactions had emerged as Quantum Chromodynamics (QCD) [34] (Fritzsch, Gell-Mann, Leutwyler 1973), exhibiting the property of Asymptotic Freedom (AF) [35] (Gross, Politzer and Wilczek 1973). All this structure today is crucial for obtaining sufficiently precise predictions for the anomalous magnetic moment of the muon as we will see.

The most important condition for the anomalous magnetic moment to be a useful monitor for testing a theory is its unambiguous predictability within that theory. The predictability crucially depends on the following properties of the theory:

(1) it must be a local relativistic quantum field theory and
(2) it must be renormalizable.

As a consequence $g - 2$ vanishes at tree level. This means that g cannot be an independently adjustable parameter in any renormalizable QFT, which in turn implies that $g - 2$ is a calculable quantity and the predicted value can be confronted with experiments. As we will see $g - 2$ can in fact be both predicted as well as experimentally measured with very high accuracy. By confronting precise theoretical predictions with precisely measured experimental data it is possible to subject the theory to very stringent tests and to find its possible limitation.

The particle–antiparticle duality [11], also called crossing or charge conjugation property, which is a basic consequence of any relativistic local QFT, implies first and foremost that particles and antiparticles have identical masses and spins. In

---

[7] Today we understand QED as an Abelian gauge theory. This important structural property was discovered by Weyl [30] in 1929.

Anomalous Magnetic Moment
of the Muon (2014).pdf

Pag. 08(24) ->

"a possible anomaly
$g \neq 2$ of the magnetic
moment of the
electron"

Obs.:

Books-001-2021\Organic
Chemistry (Mastering Chemistry),
9th Edition\
rtu3463ryu54etr635.pdf
↳ Organic Chemistry

Pag. 05(39)
Fig 1-4
nucleus — node
(region of
zero electron
density)

substances such as diamond, graphite, carbon dioxide, ammonium cyanate, and sodium carbonate are derived from minerals and have typical inorganic properties. Most of the millions of carbon compounds are classified as organic, however.

We humans are composed largely of organic molecules, and we are nourished by the organic compounds in our food. The proteins in our skin, the lipids in our cell membranes, the glycogen in our livers, and the DNA in the nuclei of our cells are all organic compounds. Our bodies are also regulated and defended by complex organic compounds.

Chemists have learned to synthesize or simulate many of these complex molecules. The synthetic products serve as drugs, medicines, plastics, pesticides, paints, and fibers. Many of the most important advances in medicine are actually advances in organic chemistry. New synthetic drugs are developed to combat disease, and new polymers are molded to replace failing organs. Organic chemistry has gone full circle. It began as the study of compounds derived from "organs," and now it gives us the drugs and materials we need to save or replace those organs.

The AbioCor® self-contained artificial heart, which is used to sustain patients who are waiting for a heart transplant. The outer shell is polycarbonate, and the valves and inner bladder are polyurethane. Both of these durable substances are synthetic organic compounds.

1-2 Principles of Atomic Structure

Before we begin our study of organic chemistry, we must review some basic principles. These concepts of atomic and molecular structure are crucial to your understanding of the structure and bonding of organic compounds.

1-2A Structure of the Atom

Atoms are made up of protons, neutrons, and electrons. Protons are positively charged and are found together with (uncharged) neutrons in the nucleus. Electrons, which have a negative charge that is equal in magnitude to the positive charge on the proton, occupy the space surrounding the nucleus (Figure 1-2). Protons and neutrons have similar masses, about 1800 times the mass of an electron. Almost all the atom's mass is in the nucleus, but it is the electrons that take part in chemical bonding and reactions.

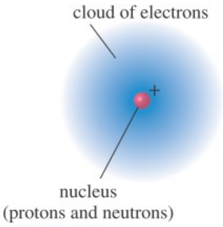

FIGURE 1-2
Basic atomic structure. An atom has a dense, positively charged nucleus surrounded by a cloud of electrons.

Each element is distinguished by the number of protons in the nucleus (the atomic number). The number of neutrons is usually similar to the number of protons, although the number of neutrons may vary. Atoms with the same number of protons but different numbers of neutrons are called **isotopes.** For example, the most common kind of carbon atom has six protons and six neutrons in its nucleus. Its mass number (the sum of the protons and neutrons) is 12, and we write its symbol as ^{12}C. About 1% of carbon atoms have seven neutrons; the mass number is 13, written ^{13}C. A very small fraction of carbon atoms have eight neutrons and a mass number of 14. The ^{14}C isotope is radioactive, with a half-life (the time it takes for half of the nuclei to decay) of 5730 years. The predictable decay of ^{14}C is used to determine the age of organic materials up to about 50,000 years old.

1-2B Electron Shells and Orbitals

An element's chemical properties are determined by the number of protons in the nucleus and the corresponding number of electrons around the nucleus. The electrons form bonds and determine the structure of the resulting molecules. Because they are small and light, electrons show properties of both particles and waves; in many ways, the electrons in atoms and molecules behave more like waves than like particles.

Electrons that are bound to nuclei are found in **orbitals.** Orbitals are mathematical descriptions that chemists use to explain and predict the properties of atoms and molecules. The *Heisenberg uncertainty principle* states that we can never determine exactly where the electron is; nevertheless, we can determine the **electron density,** the probability of finding the electron in a particular part of the orbital. An orbital, then, is an allowed energy state for an electron, with an associated probability function that defines the distribution of electron density in space.

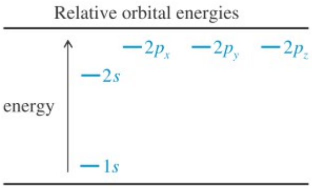

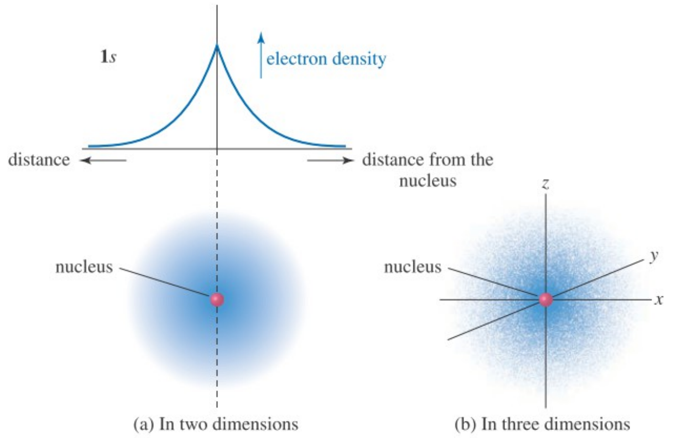

(a) In two dimensions (b) In three dimensions

FIGURE 1-3
Graph and diagram of the 1*s* atomic orbital. The electron density is highest at the nucleus and drops off exponentially with increasing distance from the nucleus in any direction.

Atomic orbitals are grouped into different "shells" at different distances from the nucleus. Each shell is identified by a principal quantum number n, with $n = 1$ for the lowest-energy shell closest to the nucleus. As n increases, the shells are farther from the nucleus, are higher in energy, and can hold more electrons. Most of the common elements in organic compounds are found in the first two rows of the periodic table, indicating that their electrons are found in the first two electron shells. The first shell ($n = 1$) can hold two electrons, and the second shell ($n = 2$) can hold eight.

The first electron shell contains just the 1*s* orbital. All *s* orbitals are spherically symmetrical, meaning that they are nondirectional. The electron density is only a function of the distance from the nucleus. The electron density of the 1*s* orbital is graphed in Figure 1-3. Notice how the electron density is highest *at* the nucleus and falls off exponentially with increasing distance from the nucleus. The 1*s* orbital might be imagined as a cotton boll, with the cottonseed at the middle representing the nucleus. The density of the cotton is highest nearest the seed, and it becomes less dense at greater distances from this "nucleus."

The second electron shell consists of the 2*s* and 2*p* orbitals. The 2*s* orbital is spherically symmetrical like the 1*s* orbital, but its electron density is not a simple exponential function. The 2*s* orbital has a smaller amount of electron density close to the nucleus. Most of the electron density is farther away, beyond a region of zero electron density called a **node.** Because most of the 2*s* electron density is farther from the nucleus than that of the 1*s*, the 2*s* orbital is higher in energy. Figure 1-4 shows a graph of the 2*s* orbital.

In addition to the 2*s* orbital, the second shell also contains three 2*p* atomic orbitals, one oriented in each of the three spatial directions. These orbitals are called the $2p_x$, the $2p_y$, and the $2p_z$, according to their direction along the x, y, or z axis. The 2*p* orbitals are slightly higher in energy than the 2*s*, because the average location of the electron in a 2*p* orbital is farther from the nucleus. Each *p* orbital consists of two lobes, one on either side of the nucleus, with a **nodal plane** at the nucleus. The nodal plane is a flat (planar) region of space, including the nucleus, with zero electron density. The three 2*p* orbitals differ only in their spatial orientation, so they have identical energies. Orbitals with identical energies are called **degenerate orbitals.** Figure 1-5 shows the shapes of the three degenerate 2*p* atomic orbitals.

The *Pauli exclusion principle* tells us that each orbital can hold a maximum of two electrons, provided that their spins are paired. The first shell (one 1*s* orbital) can accommodate two electrons. The second shell (one 2*s* orbital and three 2*p* orbitals) can accommodate eight electrons, and the third shell (one 3*s* orbital, three 3*p* orbitals, and five 3*d* orbitals) can accommodate 18 electrons.

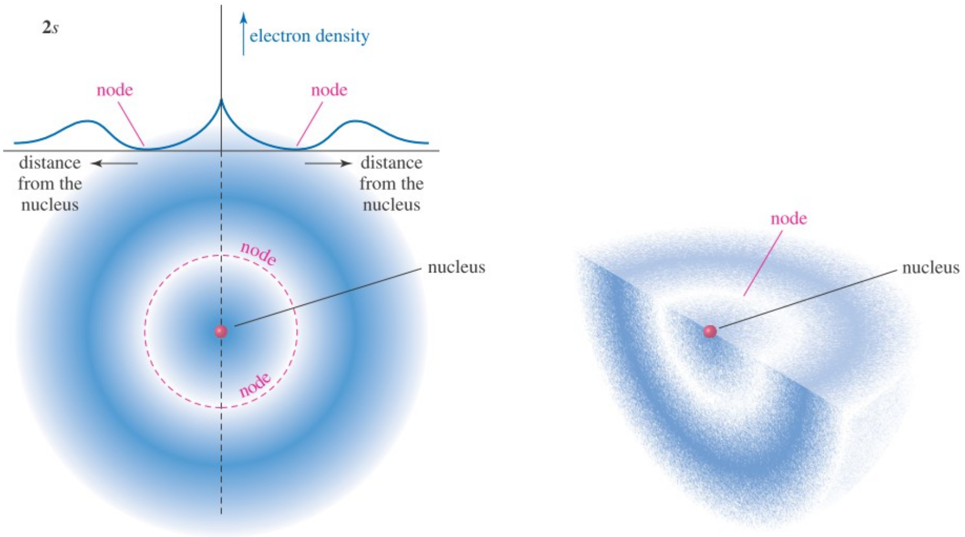
2s
electron density
node
node
distance from the nucleus
distance from the nucleus
node
node
nucleus
node
nucleus

(a) In two dimensions

(b) In three dimensions (cutaway)

FIGURE 1-4
Graph and diagram of the 2*s* atomic orbital. The 2*s* orbital has a small region of high electron density close to the nucleus, but most of the electron density is farther from the nucleus, beyond a node, or region of zero electron density.

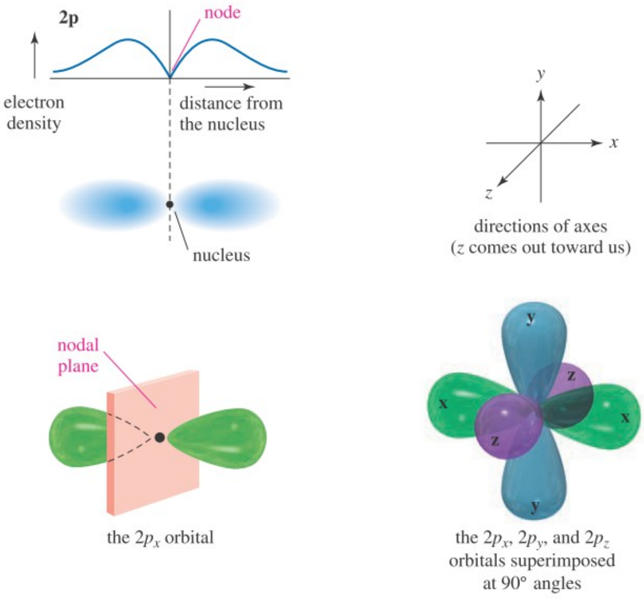
2p
node
electron density
distance from the nucleus
nucleus
y
x
z
directions of axes
(z comes out toward us)
nodal plane
the $2p_x$ orbital
y
z
x
x
z
y
the $2p_x$, $2p_y$, and $2p_z$ orbitals superimposed at 90° angles

FIGURE 1-5
The 2*p* orbitals. Three 2*p* orbitals are oriented at right angles to each other. Each is labeled according to its orientation along the *x*, *y*, or *z* axis.

" -./0- - -1 ! - < 2 ;! B 2 ? "" ! "
5 ! " > = B 4 ! 8 7 " ? 4"
& ' ! " D ; 3 6

1.2 Strict and Uniform Convexity

Let X be a Banach space with dual X^*; the value of $x^* \in X^*$ at $x \in X$ is denoted by $\langle x, x^* \rangle_X$ or $\langle x, x^* \rangle$. We recall that X is said to be *strictly convex* if whenever $x, y \in X$ are such that $x \neq y$ and $\|x\| = \|y\| = 1$, and $\lambda \in (0,1)$, then $\|\lambda x + (1-\lambda)y\| < 1$. This simply means that the unit sphere in X does not contain any line segment. Note that an equivalent condition on X is that no sphere of any radius and any centre contains a line segment. For example, if $\mathbb{R}^2$ is equipped with the l_p norm it is strictly convex if $1 < p < \infty$ but not if $p = 1$ or ∞.

Proposition 1.1. *The space X is strictly convex if and only if it is the case that whenever $x, y \in X$ are such that $\|x + y\| = \|x\| + \|y\|$ then either $y = 0$ or $x = \lambda y$ for some $\lambda \geq 0$.*

Proof. Suppose that X is strictly convex and $x, y \in X \backslash \{0\}$ are such that $x \neq y$ and $\|x + y\| = \|x\| + \|y\|$. Then $\|x\| \neq \|y\|$, for otherwise $\left\| \frac{x+y}{2} \right\| = \|x\|$, contradicting the strict convexity of X. However, if $\|y\| < \|x\|$, we put $\lambda = \|y\| / \|x\|$ and observe that

$$1 \geq \left\| \frac{x}{\|x\|} + \lambda \left(\frac{y}{\|y\|} - \frac{x}{\|x\|} \right) \right\| \geq \frac{\|x+y\| - \lambda \|x\|}{\|x\|} = \frac{\|x\| + \|y\| - \lambda \|x\|}{\|x\|} = 1,$$

which means that $x = y/\lambda$. The converse is obvious. □

It is plain that every linear subspace of a strictly convex space is itself strictly convex, with the inherited norm. Here are some of the less trivial properties of such spaces.

Proposition 1.2. *Let X be a Banach space with dual X^*. Then X is strictly convex if and only if given any $x^* \in X^* \backslash \{0\}$, there is at most one $x \in X$ such that $\|x\| = 1$ and $\langle x, x^* \rangle = \|x^*\|$; if X is reflexive, such an x exists.*

Proof. Let X be strictly convex and $x^* \in X^* \backslash \{0\}$. Suppose there are two distinct such points x, say x_1 and x_2. Then if $0 < \lambda < 1$,

$$\begin{aligned} \|x^*\| &= \lambda \langle x_1, x^* \rangle + (1-\lambda) \langle x_2, x^* \rangle = \langle \lambda x_1 + (1-\lambda) x_2, x^* \rangle \\ &\leq \|x^*\| \, \|\lambda x_1 + (1-\lambda) x_2\| < \|x^*\|, \end{aligned}$$

which is absurd. Conversely, suppose that $\|x + \lambda(y - x)\| = 1$ for some $x, y \in X$ with $\|x\| = \|y\| = 1$ and some $\lambda \in (0,1)$. By the Hahn–Banach theorem, there exists $x^* \in X^*$ such that $\langle x + \lambda(y - x), x^* \rangle = 1$ and $\|x^*\| = 1$. Then $(1-\lambda) \langle x, x^* \rangle + \lambda \langle y, x^* \rangle = 1$, and since $|\langle x, x^* \rangle|, |\langle y, x^* \rangle| \leq 1$ we must have $\langle x, x^* \rangle = \langle y, x^* \rangle = 1$. By hypothesis, this implies that $x = y$, and so X is strictly convex.

Now suppose that X is reflexive and let (x_k) be a sequence in X such that $\|x_k\| = 1$ for all $k \in \mathbb{N}$ and $\|x^*\| = \lim_{k \to \infty} \langle x_k, x^* \rangle$. Since X is reflexive, there is a weakly convergent subsequence of (x_k), still denoted by (x_k) for simplicity, with weak limit x, say. Then $\|x\| \leq 1$ and $\langle x, x^* \rangle = \lim_{k \to \infty} \langle x_k, x^* \rangle = \|x^*\|$. □

Proposition 1.3. *Let X be a Banach space. Then X^* is strictly convex if and only if given any $x \in X\backslash\{0\}$, there is a unique $x^* \in X^*$ such that $\|x^*\| = 1$ and $\langle x, x^*\rangle = \|x\|$.*

Proof. Suppose that X^* is strictly convex. By the Hahn–Banach theorem, there exists $x^* \in X^*$ with $\|x^*\| = 1$ and $\langle x, x^*\rangle = \|x\|$. Suppose there exists $y^* \in X^*$, $y^* \neq x^*$, with $\|y^*\| = 1$ and $\langle x, y^*\rangle = \|x\|$. Then if $0 < \lambda < 1$,

$$\|x\| = \lambda \langle x, x^*\rangle + (1-\lambda) \langle x, y^*\rangle \leq \|\lambda x^* + (1-\lambda)y^*\| \, \|x\| < \|x\|,$$

and we have a contradiction. The proof of the converse is similar to that of the corresponding statement in Proposition 1.2. □

Proposition 1.4. *Let K be a convex subset of a strictly convex Banach space X. Then there is at most one element $x \in K$ such that*

$$\|x\| = \inf\{\|y\| : y \in K\}.$$

If, in addition, X is reflexive and K is closed and non-empty, then such a point x exists.

Proof. Suppose there exist $x, y \in K$ with $\|x\| = \|y\| = \inf\{\|z\| : z \in K\}$, $x \neq y$. Let $0 < \lambda < 1$: then $\lambda x + (1-\lambda)y \in K$, $\|\lambda x + (1-\lambda)y\| < \|x\|$ and we have a contradiction.

For the second assertion, let (x_k) be a sequence in K such that $\lim_{k\to\infty} \|x_k\| = l := \inf\{\|y\| : y \in K\}$. By reflexivity, this sequence has a subsequence, still denoted by (x_k) for convenience, such that $x_k \rightharpoonup x$ for some $x \in X$, by which we mean that (x_k) converges weakly to x in X; in fact, $x \in K$ since K is convex and closed, and hence weakly closed. Moreover, $\|x\| \leq \lim_{k\to\infty} \|x_k\| = l$. □

To measure the degree of strict convexity of a Banach space X, we define its *modulus of convexity* $\delta_X : [0,2] \to [0,1]$ by

$$\delta_X(\varepsilon) = \inf\left\{1 - \frac{1}{2}\|x+y\| : x, y \in X, \|x\| \leq 1, \|y\| \leq 1, \|x-y\| \geq \varepsilon\right\}. \tag{1.5}$$

This function is introduced so that, given any two distinct points x and y in the closed unit ball B in X, we shall have an idea of how far inside B is the midpoint of the line segment joining x to y. Note that the same function is obtained if the infimum is taken over all $x, y \in X$ with $\|x\| = \|y\| = 1$ and $\|x-y\| = \varepsilon$: see [88], Vol. II, p. 60. The Banach space X is called *uniformly convex* if for all $\varepsilon \in (0,2]$, $\delta_X(\varepsilon) > 0$. This means that X is uniformly convex if, given any $\varepsilon \in (0,2]$, $\|x+y\| \leq 2(1 - \delta_X(\varepsilon))$ whenever $x, y \in X$ are such that $\|x\| = \|y\| = 1$ and $\|x-y\| \geq \varepsilon$. It is easy to see that an equivalent condition is that whenever (x_k) and (y_k) are sequences in X such that $\|x_k\| \leq 1, \|y_k\| \leq 1$ $(k \in \mathbb{N})$ and $\lim_{k\to\infty} \|x_k + y_k\| = 2$, then $\lim_{k\to\infty} \|x_k - y_k\| = 0$. Note that δ_X is increasing on $[0,2]$, continuous on $[0,2)$ (not necessarily at 2) and $\delta_X(0) = 0$. Moreover, $\delta_X(2) = 1$ if and only if X is strictly convex.

Plainly every closed linear subspace of a uniformly convex space is uniformly convex when given the inherited norm. Moreover, it is clear that every uniformly convex space is strictly convex: as we shall see later, the converse is false. However, in finite-dimensional spaces strict convexity does imply uniform convexity, for then the fact that $1-\frac{1}{2}\|x+y\|>0$ whenever $\|x\|=\|y\|=1$ and $\|x-y\|=\varepsilon>0$ implies that $\delta_X(\varepsilon)>0$ since closed bounded sets are compact in finite-dimensional spaces. The simplest example of a uniformly convex space is any Hilbert space H, for the parallelogram law enables us to see that

$$\delta_H(\varepsilon)=1-(1-\varepsilon^2/4)^{1/2},$$

which is clearly positive for all $\varepsilon\in(0,2]$; in fact, since

$$(1-x^q)^{1/q}\le 1-x^q/q \quad (0\le x\le 1, 1<q<\infty),$$

we have

$$\delta_H(\varepsilon)\ge\varepsilon^2/8.$$

Note that in the opposite direction, for every Banach space X of dimension at least 2, it is known that

$$\delta_X(\varepsilon)\le\delta_H(\varepsilon)=1-(1-\varepsilon^2/4)^{1/2}\le C\varepsilon^2;$$

see [88], Vol. II, p. 63. In this sense, Hilbert spaces are the 'most' uniformly convex spaces.

If $1<p<\infty$, the sequence space l_p and the Lebesgue space L_p (over any set and with any measure) are uniformly convex. This was established by Clarkson [25], who introduced the notion of uniform convexity and proved the following inequalities (now known as the Clarkson inequalities) that hold for arbitrary elements $f,g\in L_p$:

$$\left(\|(f+g)/2\|_p^{p'}+\|(f-g)/2\|_p^{p'}\right)^{1/p'}\le\left(\frac{\|f\|_p^p+\|g\|_p^p}{2}\right)^{1/p} \text{ if } 1\le p\le 2, \quad (1.6)$$

and

$$\left(\|(f+g)/2\|_p^{p}+\|(f-g)/2\|_p^{p}\right)^{1/p}\le\left(\frac{\|f\|_p^p+\|g\|_p^p}{2}\right)^{1/p} \text{ if } 2\le p\le \infty. \quad (1.7)$$

Here $\|\cdot\|_p$ denotes the norm on L_p. In the cases $2\le p\le\infty$ and $1\le p\le 2$ the inequalities in (1.6) and (1.7) respectively hold in the reversed sense. He used these to obtain lower bounds for $\delta_p:=\delta_{L_p}$ of the form

$$\delta_p(\varepsilon)\ge(\varepsilon/C)^r, \quad (1.8)$$

where $r = p'$ if $1 < p \leq 2$, $r = p$ if $2 \leq p < \infty$, and C is a positive constant that depends on p. These lower bounds are not optimal when $1 < p \leq 2$; however, from them the uniform convexity of L_p when $1 < p < \infty$ follows immediately.

Estimates from below of the modulus of convexity of a Banach space of the form (1.8) are very useful, particularly if the best value of r is known, and so we present below results leading to the optimal value of r for L_p, following the approach of Ball et al. [3]. We begin with a technical lemma.

Lemma 1.1. *Let* $1 < p < \infty$ *and define* $\lambda_p : [0,\infty) \to [0,\infty)$ *by*

$$\lambda_p(t) = (1+t)^{p-1} + |1-t|^{p-1} \ sgn\ (1-t).$$

Then for all $x, y \in \mathbb{R}$,

$$|x+y|^p + |x-y|^p = \sup\{\lambda_p(t)\,|x|^p + \lambda_p(1/t)\,|y|^p : 0 < t < \infty\}$$

if $1 < p \leq 2$; *if* $2 \leq p < \infty$, *the same holds with* sup *replaced by* inf.

Proof. Suppose that $1 < p \leq 2$. By homogeneity and symmetry it is enough to deal with the case $0 < y \leq x = 1$. Put

$$f(t) = \lambda_p(t) + \lambda_p(1/t)y^p,\ 0 < t < \infty,$$

and note that $f(y) = (1+y)^p + (1-y)^p$. Moreover, if $t \neq y$, a routine calculation shows that

$$f'(t) = (p-1)\{1-(y/t)^p\}\left\{(1+t)^{p-2} - |1-t|^{p-2}\right\}.$$

Thus $f'(t) \geq 0$ if $0 < t < y$, $f'(t) \leq 0$ if $y < t < \infty$. It follows that the maximum value of f on $(0,\infty)$ occurs when $t = y$, and the desired inequality follows. The proof when $2 \leq p < \infty$ is similar. □

In [3] it is shown that Clarkson's inequalities may be obtained from this result, which also gives rise to the following theorem, due originally to Hanner [71] and from which the modulus of convexity δ_p can be estimated.

Theorem 1.1. *Suppose that* $1 < p \leq 2$ *and let* $f, g \in L_p$. *Then*

$$\|f+g\|_p^p + \|f-g\|_p^p \geq \left(\|f\|_p + \|g\|_p\right)^p + \left|\|f\|_p - \|g\|_p\right|^p.$$

If $2 \leq p < \infty$ *the inequality is reversed.*

Proof. Suppose that $1 < p \leq 2$. Then by Lemma 1.1,

$$\begin{aligned}
\|f+g\|_p^p+\|f-g\|_p^p &= \int \{|f+g|^p+|f-g|^p\} \\
&= \int \sup_{0<t<\infty} \{\lambda_p(t)\,|f|^p+\lambda_p(1/t)\,|g|^p\} \\
&\geq \sup_{0<t<\infty} \int \{\lambda_p(t)\,|f|^p+\lambda_p(1/t)\,|g|^p\} \\
&= \sup_{0<t<\infty} \{\lambda_p(t)\,\|f\|^p+\lambda_p(1/t)\,\|g\|^p\} \\
&= \left(\|f\|_p+\|g\|_p\right)^p+\left|\|f\|_p-\|g\|_p\right|^p,
\end{aligned}$$

the final step following from another application of Lemma 1.1. The case $2 \leq p < \infty$ is similar. □

Finally we give a theorem of Ball and Pisier (see [3]).

Theorem 1.2. *Suppose that* $1 < p \leq 2$ *and let* $f, g \in L_p$. *Then*

$$\|f+g\|_p^2+\|f-g\|_p^2 \geq 2\,\|f\|_p^2+2(p-1)\,\|g\|_p^2.$$

If $2 \leq p < \infty$ *the inequality is reversed.*

Proof. Suppose that $1 < p \leq 2$. We use the inequality

$$|a+b|^p+|a-b|^p \geq 2\{a^2+(p-1)b^2\}^{p/2}\ (a,b \in \mathbb{R}) \tag{1.9}$$

which is ascribed to Gross in [3]. Accepting this for the moment, note that by Hölder's inequality,

$$\|f+g\|_p^p+\|f-g\|_p^p \leq 2^{(2-p)/2}\left(\|f+g\|_p^2+\|f-g\|_p^2\right)^{p/2},$$

from which we have, by Theorem 1.1,

$$\begin{aligned}
\left(\|f+g\|_p^2+\|f-g\|_p^2\right)^{p/2} &\geq 2^{-(2-p)/2}\left\{\left(\|f\|_p+\|g\|_p\right)^p+\left|\|f\|_p-\|g\|_p\right|^p\right\} \\
&\geq 2^{p/2}\left\{\|f\|_p^2+(p-1)\,\|g\|_p^2\right\}^{p/2},
\end{aligned}$$

the final inequality following from (1.9). This gives the theorem when $1 < p \leq 2$; similar arguments work when $2 \leq p < \infty$.

To prove (1.9) it is enough to show that

$$h(t) := (1+t)^p+(1-t)^p-2\{1+(p-1)t^2\}^{p/2} \geq 0,\ 0 \leq t \leq 1. \tag{1.10}$$

Note that if $0 \le t < 1$,

$$(1+t)^p + (1-t)^p = 2 + p(p-1)t^2 + 2p(p-1)\sum_{n=2}^{\infty} \frac{(3-p)...(2n-1-p)}{(2n)!} t^{2n}$$

and

$$\{1+(p-1)t^2\}^{p/2} = 1 + \frac{p}{2}(p-1)t^2 + \sum_{n=2}^{\infty} \frac{p(2-p)...(2n-2-p)}{n!2^n}(p-1)^n t^{2n}(-1)^{n+1},$$

so that

$$h(t) = 2p(p-1)\sum_{n=2}^{\infty} \frac{(3-p)...(2n-1-p)}{(2n)!} t^{2n} + I,$$

where

$$I = 2\sum_{m=1}^{\infty} \frac{p(2-p)...(4m-2-p)}{(2m+1)!2^{2m+1}} \left\{2(2m+1) - (4m-p)(p-1)t^2\right\}(p-1)^{2m}t^{4m}.$$

Since $2(2m+1) - (4m-p)(p-1)t^2 > 0$ when $0 \le t < 1$, $h(t) \ge 0$ and (1.10) follows. □

Now these results are applied to give optimal estimates for δ_p, the modulus of convexity of L_p. First suppose that $1 < p \le 2$. Then from Theorem 1.2 with $u = f+g$ and $v = f-g$ we see that

$$\|u\|_p^2 + \|v\|_p^2 \ge 2\left\|\frac{u+v}{2}\right\|_p^2 + 2(p-1)\left\|\frac{u-v}{2}\right\|_p^2,$$

from which we have

$$\delta_p(\varepsilon) \ge 1 - (1-(p-1)(\varepsilon/2)^2)^{1/2} \ge (p-1)\varepsilon^2/8.$$

In the terminology of [110], according to which given any $r \in [2,\infty)$ a Banach space X is said to be r-uniformly convex if $\delta_X(\varepsilon) \ge (\varepsilon/C)^r$ for some positive constant C, this means that L_p is 2-uniformly convex. When $2 \le p < \infty$, direct use of Theorem 1.1 shows that $\delta_p(\varepsilon) \ge 1-(1-(\varepsilon/2)^p)^{1/p} \ge 1-(1-(\varepsilon/2)^p/p) = (\varepsilon/2)^p/p$, so that L_p is p-uniformly convex. The asymptotic behaviour of δ_p for $1 < p < \infty$ is stated in [88], Vol. II, p. 63; see also [107], p. 237.

To provide other interesting examples of uniformly convex spaces it is convenient to introduce certain spaces of l_p type. Let $I \subset \mathbb{N}$ be non-empty, and for each $i \in I$ let A_i be a Banach space; put $A = \{A_i : i \in I\}$ and let $1 < p < \infty$. Define

$$l_p(A) = \left\{ a = (a_i)_{i\in I} : a_i \in A_i, \left\|a \mid l_p(A)\right\| = \left(\sum_{i\in I} \|a_i \mid A_i\|^p\right)^{1/p} < \infty \right\}.$$

This is a linear space when given the natural definitions of addition and multiplication by scalars; it is routine to verify that it becomes a Banach space when endowed with the norm $\left\|\cdot \mid l_p(A)\right\|$. The following result is due to Day [30].

Theorem 1.3. *The space $l_p(A)$ is strictly convex if and only if each A_i is strictly convex; it is uniformly convex if and only if every A_i is uniformly convex with modulus of convexity δ_i and $\delta(\varepsilon) := \inf_{i\in I}\delta_i(\varepsilon) > 0$ for all $\varepsilon \in (0,2)$.*

Proof. First suppose that every A_i is strictly convex, and let $a,b \in l_p(A)\backslash\{0\}$ be such that $\left\|a+b \mid l_p(A)\right\| = \left\|a \mid l_p(A)\right\| + \left\|b \mid l_p(A)\right\|$. Then

$$\left(\sum_{i\in I}\|a_i+b_i \mid A_i\|^p\right)^{1/p} = \left(\sum_{i\in I}\|a_i \mid A_i\|^p\right)^{1/p} + \left(\sum_{i\in I}\|b_i \mid A_i\|^p\right)^{1/p},$$

and so, by a natural adaptation of the classical method of dealing with the case of equality in Minkowski's inequality, we see that $a = \lambda b$ for some $\lambda > 0$. By Proposition 1.1 it follows that $l_p(A)$ is strictly convex. That strict convexity of $l_p(A)$ implies that of each A_i is obvious.

Now assume that each A_i is uniformly convex and has modulus of convexity δ_i, with $\delta(\varepsilon) := \inf_{i\in I}\delta_i(\varepsilon) > 0$ for all $\varepsilon \in (0,2)$. Let $\varepsilon > 0$, ε small, and suppose initially that $a,b \in l_p(A)$ are such that $\left\|a \mid l_p(A)\right\| = \left\|b \mid l_p(A)\right\| = 1$, $\left\|a-b \mid l_p(A)\right\| > \varepsilon$ and $\|a_i \mid A_i\| = \|b_i \mid A_i\| = \beta_i$ for all $i \in I$; put $\|a_i - b_i \mid A_i\| = \gamma_i$, and note that $\gamma_i \le 2\beta_i$. Then $\|a_i + b_i \mid A_i\| < 2(1-\delta(\gamma_i/\beta_i))\beta_i$ for all $i \in I$, and so

$$\left\|a+b \mid l_p(A)\right\| \le 2\left\{\sum_{i\in I}((1-\delta(\gamma_i/\beta_i))\beta_i)^p\right\}^{1/p}.$$

Let $E = \{i \in I : \gamma_i/\beta_i > \varepsilon/4\}$ and $F = I\backslash E$. Then

$$1 = \left(\sum_{i\in I}\beta_i^p\right)^{1/p} \ge \left(\sum_{i\in F}\beta_i^p\right)^{1/p} \ge (4/\varepsilon)\left(\sum_{i\in F}\gamma_i^p\right)^{1/p}.$$

Hence

$$\left(\sum_{i\in E}\gamma_i^p\right)^{1/p} = \left(\sum_{i\in I}\gamma_i^p - \sum_{i\in F}\gamma_i^p\right)^{1/p} \ge (\varepsilon^p - (\varepsilon/4)^p)^{1/p} \ge 3\varepsilon/4,$$

so that

$$\alpha := \left(\sum_{i\in E}\beta_i^p\right)^{1/p} \ge 3\varepsilon/8.$$

It follows that

$$\begin{aligned}\left\|a+b \mid l_p(A)\right\| &\leq 2\left\{(1-\delta(\varepsilon/4))^p \sum_{i\in E}\beta_i^p + \sum_{i\in F}\beta_i^p\right\}^{1/p}\\ &\leq 2\left\{(1-\delta(\varepsilon/4))^p\alpha^p + 1-\alpha^p\right\}^{1/p}\\ &\leq 2\left\{1-(1-(1-\delta(\varepsilon/4))^p)\alpha^p\right\}^{1/p}\\ &\leq 2\left\{1-(1-(1-\delta(\varepsilon/4))^p)(3\varepsilon/8)^p\right\}^{1/p} = 2(1-\delta_0(\varepsilon)),\ \text{say,}\end{aligned}$$

which is an inequality of the required form.

Next suppose that $\left\|a \mid l_p(A)\right\| = \left\|b \mid l_p(A)\right\| = 1$ and that $\left\|a+b \mid l_p(A)\right\| > 2(1-\delta_p(\eta))$, where δ_p is the modulus of convexity of $l_p(A)$ and $\eta \in (0,2]$. Then

$$2(1-\delta_p(\eta)) \leq \left(\sum_{i\in I}(\|a_i \mid A_i\| + \|b_i \mid A_i\|)^p\right)^{1/p} \leq 2,$$

and since the sequences $(\|a_i \mid A_i\|)$ and $(\|b_i \mid A_i\|)$ are in l_p, we see that

$$\left(\sum_{i\in I}|\|a_i \mid A_i\| - \|b_i \mid A_i\||^p\right)^{1/p} < \eta.$$

Let $c=(c_i)=(b_i\|a_i \mid A_i\|/\|b_i \mid A_i\|)$, so that $\|c_i \mid A_i\| = \|a_i \mid A_i\|$ for all $i \in I$ and $\left\|b-c \mid l_p(A)\right\| = \left(\sum_{i\in I}|\|b_i \mid A_i\| - \|a_i \mid A_i\||^p\right)^{1/p} < \eta$. Thus by the earlier part of our discussion, $\left\|a-c \mid l_p(A)\right\| \leq \varepsilon/2$ if $\left\|a+c \mid l_p(A)\right\| \geq 2(1-\delta_0(\varepsilon/2))$; moreover, $\left\|a+c \mid l_p(A)\right\| \geq \left\|a+b \mid l_p(A)\right\| - \left\|c-b \mid l_p(A)\right\| > 2(1-\delta_p(\eta)-\eta/2)$. Choose $\eta \in (0,\varepsilon/2)$ so that $\delta_p(\eta)+\eta/2 < \delta_0(\varepsilon/2)$. Then $\left\|a+b \mid l_p(A)\right\| > 2(1-\delta_p(\eta))$ implies that $\left\|a+c \mid l_p(A)\right\| > 2(1-\delta_0(\varepsilon/2))$, which in turn implies that $\left\|a-c \mid l_p(A)\right\| < \varepsilon/2$. Thus $\left\|a-b \mid l_p(A)\right\| \leq \left\|a-c \mid l_p(A)\right\| + \left\|b-c \mid l_p(A)\right\| < \varepsilon$. Set $\delta_p^{'}(\varepsilon) = \delta_p(\eta) > 0$: then $\left\|a-b \mid l_p(A)\right\| < \varepsilon$ if $\left\|a \mid l_p(A)\right\| = \left\|b \mid l_p(A)\right\| = 1$ and $\left\|a+b \mid l_p(A)\right\| > 2(1-\delta_p^{'}(\varepsilon))$. This establishes the uniform convexity of $l_p(A)$.

Finally, suppose that $l_p(A)$ is uniformly convex. Consideration of the points $a = (\delta_{ij}) \in A$, where δ_{ij} is the Kronecker delta, shows that each A_j is uniformly convex. □

If each $A_i = L_{p_i}$, then since by our earlier results the modulus of convexity of L_{p_i} satisfies

$$\delta_{L_{p_i}}(\varepsilon) \geq \left(\frac{p_i-1}{8}\right)\varepsilon^2 \text{ if } 1<p_i\leq 2, \text{ and } \delta_{L_{p_i}}(\varepsilon) \geq (\varepsilon/2)^{p_i}/p_i \text{ if } 2<p_i<\infty,$$

it follows that $l_p(A)$ is uniformly convex if and only if the sequence $\{p_i\}$ is bounded away from 1 and ∞.

As a further immediate application of this result we deal with Sobolev spaces. Let Ω be an open subset of $\mathbb{R}^n$, let $p \in (1,\infty)$ and suppose that $k \in \mathbb{N}$. The Sobolev space $W_p^k(\Omega)$ is the linear space of all (equivalence classes of) functions $u \in D^\alpha u$ such that for all $\alpha = (\alpha_j) \in \mathbb{N}_0^n$ with $|\alpha| := \alpha_1 + ... + \alpha_n \leq k$, the distributional derivative $D^\alpha u := \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} ... \partial x_n^{\alpha_n}}$ belongs to $L_p(\Omega)$. Endowed with the norm

$$\left\| u \mid W_p^k(\Omega) \right\| := \left(\sum_{|\alpha| \leq k} \left\| D^\alpha u \mid L_p(\Omega) \right\|^p \right)^{1/p}$$

it becomes a Banach space, which is isometrically isomorphic to a closed linear subspace V of the product of $N := \#\{\alpha \in \mathbb{N}_0^n : |\alpha| \leq k\}$ copies of $L_p(\Omega)$, the isomorphism being established by the map $u \longmapsto (D^\alpha u)_{|\alpha| \leq k}$. With $I = \{1, 2, ..., N\}$ and $A_j = L_p(\Omega)$ for all $j \in I$, it is plain that this product is just $l_p(A)$, and in view of Theorem 1.3, $l_p(A)$ is uniformly convex. Thus V is uniformly convex, and therefore so is $W_p^k(\Omega)$. We have thus proved

Corollary 1.1. *For all $p \in (1,\infty)$ and all $k \in \mathbb{N}$, the Sobolev space $W_p^k(\Omega)$ is uniformly convex.*

Remark 1.1. (i) In [29] and [30] the spaces $l_p(A)$ are used to show that strict convexity does not imply uniform convexity, and indeed that there are separable, reflexive, strictly convex Banach spaces that are not isomorphic to any uniformly convex space. When $A = \{l_{p_j} : j \in \mathbb{N}\}$, $l_p(A)$ is isomorphic to a uniformly convex space if and only if the sequence (p_j) is bounded away from 1 and ∞. A result more general than Theorem 1.3 is given in [31], the idea being to replace the l_p structure of the space $l_p(A)$ by means of a more general norm.

(ii) In [30] an analogue of Theorem 1.3 is given for spaces of Lebesgue type. Let $1 < p < \infty$, let X be a Banach space, let Ω be an open subset of $\mathbb{R}^n$ and denote by $L_p(X)$ the space of all functions $f : \Omega \to X$ such that

$$\left\| f \mid L_p(X) \right\| := \left(\int_\Omega \| f(x) \mid X \|^p \, dx \right)^{1/p} < \infty.$$

Then [30] shows that $L_p(X)$, endowed with the norm $\left\| \cdot \mid L_p(X) \right\|$, is uniformly convex if and only if X is uniformly convex. The same holds with any measure space instead of Ω and Lebesgue measure. An application of this result will be made in Chap. 3.

We now give some useful properties of uniformly convex spaces.

Proposition 1.5. *Let X be uniformly convex and let (x_k) be a sequence in X that converges weakly to $x \in X$, with $\|x_k\| \to \|x\|$. Then $\|x_k - x\| \to 0$.*

Proof. As the result is trivial if $x = 0$ we assume that $x \neq 0$; we may plainly also assume that for all $k \in \mathbb{N}$, $x_k \neq 0$. Put $\xi_k = 1 - \|x\| / \|x_k\|$; $\xi_k \to 0$ as $k \to \infty$. Set

$y_k = x_k / \|x_k\|$, $y = x/\|x\|$ and note that $y_k = (1-\xi_k)x_k/\|x\|$ converges weakly to y. Moreover, $\|y_k\| = \|y\| = 1$. By the Hahn–Banach theorem, there exists $y^* \in X^*$ such that $\langle y, y^* \rangle_X = 1 = \|y^*\|$. Hence

$$2 \geq \|y_k + y\| \geq \langle y_k + y, y^* \rangle_X \to 2\langle y, y^* \rangle_X = 2,$$

so that $\lim_{k\to\infty} \|y_k + y\| = 2$. Since X is uniformly convex, $\lim_{k\to\infty}\|y_k - y\| = 0$, and as $x_k - x = \|x_k\| y_k - \|x\| y$, it follows easily that $\|x_k - x\| \to 0$. ☐

Theorem 1.4. *Every uniformly convex space is reflexive.*

Proof. This classical result is due to D.P. Milman [98]; here we give the short proof contained in [88], Vol. II, Prop. 1.e.3. Suppose that X is uniformly convex, let $J : X \to X^{**}$ be the canonical map, let B, B^{**} be the closed unit balls in X, X^{**} respectively and denote by σ the weak topology $\sigma(X^{**}, X^*)$. Let $x^{**} \in X^{**}$, $\|x^{**}\| = 1$. Then since $J(B)$ is σ-dense in B^{**} (see [15], Chap. IV, Sect. 5), there is a generalised sequence $\{x_\alpha\}_{\alpha\in A}$ $(x_\alpha \in B)$ such that $Jx_\alpha \overset{\sigma}{\to} x^{**}$. Since $Jx_\alpha + Jx_\beta \overset{\sigma}{\to} 2x^{**}$, the weak lower-semicontinuity of the norm shows that $\lim_{\alpha,\beta} \left\|Jx_\alpha + Jx_\beta\right\| = 2$, and hence

$$\lim_{\alpha,\beta} \left\|x_\alpha + x_\beta\right\| = 2. \tag{1.11}$$

As X is uniformly convex, it follows from (1.11) that $\lim_{\alpha,\beta}\left\|x_\alpha - x_\beta\right\| = 0$, so that for some $x \in B$, $\lim_\alpha \|x_\alpha - x\| = 0$. Thus $\lim_\alpha \|Jx_\alpha - Jx\| = 0$ and hence $x^{**} = Jx$: X must be reflexive. ☐

Together with Proposition 1.4 this gives

Proposition 1.6. *Let K be a closed, convex, non-empty subset of a uniformly convex space X. Then K has a unique element of minimal norm.*

Now let X be uniformly convex and let $K \subset X$ be closed, convex and non-empty. Given any $x \in X$, the set $K - \{x\} := \{y - x : y \in K\}$ is closed and convex and so has a unique element $w(x)$ of minimal norm, by Proposition 1.6. Clearly $w(x) = z(x) - x$ for some unique $z(x) \in K$ and

$$\|z(x) - x\| = \inf\{\|y - x\| : y \in K\} = d(x, K).$$

In other words, there is a unique point $P_K x := z(x)$ of K such that $d(x,K) = \|x - P_K x\|$. The mapping $P_K : X \to K$ is called the *projection map of X onto K*; plainly $P_K x = x$ if and only if $x \in K$.

Proposition 1.7. *Let K be a closed, convex, non-empty subset of a uniformly convex space X. Then the projection map P_K is continuous.*

Proof. Suppose the result is false. Then there exist $x \in X$, a sequence (x_n) in X and $\varepsilon > 0$ such that $\lim_{n\to\infty} x_n = x$ and $\|P_K x_n - P_K x\| \geq \varepsilon$ for all $n \in \mathbb{N}$. Since $|d(x,K) - d(x_n,K)| \leq \|x_n - x\|$, it follows that

$$|\|x_n - P_K x_n\| - \|x - P_K x\|| \leq \|x_n - x\| \to 0 \text{ as } n \to \infty.$$

As $(P_K x_n)$ is bounded, it has a weakly convergent subsequence, still denoted by $(P_K x_n)$ for convenience, with weak limit z, say. Since K is closed and convex, it is weakly closed and so $z \in K$. Moreover, $x_n - P_K x_n \rightharpoonup x - z$ and

$$\|x - z\| \leq \liminf_{n\to\infty} \|x_n - P_K x_n\| = \|x - P_K x\|,$$

which implies that $z = P_K x$. Hence

$$x_n - P_K x_n \rightharpoonup x - P_K x \text{ and } \|x_n - P_K x_n\| \to \|x - P_K x\|.$$

By Proposition 1.5, the uniform convexity of X now implies that $x_n - P_K x_n \to x - P_K x$, so that $P_K x_n \to P_K x$: contradiction. □

If X is a Hilbert space H with inner product $(\cdot,\cdot)$, this last result can be sharpened, for then

$$\|P_K x - P_K y\| \leq \|x - y\| \text{ for all } x, y \in H.$$

To establish this, note that for all $w \in K$ and all $x \in H$,

$$\|x - P_K x\| \leq \|x - w\|.$$

Thus if $\lambda \in (0,1)$,

$$\|x - P_K x\|^2 \leq \|x - (1-\lambda) P_K x - \lambda P_K y\|^2 = \|x - P_K x + \lambda (P_K x - P_K y)\|^2.$$

Hence

$$0 \leq 2 \text{ re } (x - P_K x, P_K x - P_K y) + \lambda \|P_K x - P_K y\|^2.$$

Similarly,

$$0 \leq 2 \text{ re } (y - P_K y, P_K y - P_K x) + \lambda \|P_K x - P_K y\|^2.$$

Addition of these inequalities gives

$$\text{re } (x - y + P_K y - P_K x, P_K x - P_K y) \geq -\lambda \|P_K x - P_K y\|^2.$$

Letting $\lambda \to 0$ we see that

$$\text{re } (x - y, P_K y - P_K x) \geq \|P_K x - P_K y\|^2,$$

from which the result follows easily.

A map $\mu : [0,\infty) \to [0,\infty)$ that is continuous, strictly increasing and satisfies $\mu(0) = 0$, $\lim_{t\to\infty} \mu(t) = \infty$, is called a *gauge function*. A map J from a Banach space X to 2^{X^*}, the set of all subsets of X^*, is said to be a *duality map on X with gauge function* μ if

$$J(x) = \{x^* \in X^* : \langle x, x^* \rangle = \|x^*\| \|x\|, \|x^*\| = \mu(\|x\|)\}.$$

Note that by the Hahn–Banach theorem, for each $x \in X$ the set $J(x)$ is non-empty. It is also convex. To justify this, let $x^*, y^* \in J(x)$ and $\lambda \in (0,1)$. Put $z^* = \lambda x^* + (1-\lambda) y^*$ and observe that

$$\langle x, z^* \rangle = \lambda \langle x, x^* \rangle + (1-\lambda) \langle x, y^* \rangle = \mu(\|x\|) \|x\|.$$

Hence $\|z^*\| \geq \mu(\|x\|)$. However,

$$\|z^*\| \leq \lambda \|x^*\| + (1-\lambda) \|y^*\| = \mu(\|x\|)$$

and so $\|z^*\| = \mu(\|x\|)$. Thus $z^* \in J(x)$ and the proof is complete.

Let X be a Banach space with strictly convex dual X^* and let J be a duality map on X with gauge function μ. Then for each $x \in X$, the set $J(x)$ consists of precisely one point. In fact, for each $x \in X$, the points in $J(x)$ lie on the sphere in X^* with centre 0 and radius $\mu(\|x\|)$. If $J(x)$ contained two distinct points, the midpoint of the line segment joining them would be in the convex set $J(x)$, which is impossible as X^* is strictly convex.

In view of this result, we shall regard a duality map J on X as a map from X to X^* when X^* is strictly convex. It is known (see [19], pp. 42–43 and [94], p. 176) that if X and X^* are strictly convex, then J is an injective map of X onto X^* that is strictly monotone (that is, $\langle x-y, Jx-Jy \rangle > 0$ if $x, y \in X, x \neq y$); it is weakly continuous in the sense that if $x_k \to x_0$ in X, then Jx_k converges weak* in X^* to Jx_0 (the convergence is strong if, in addition, X^* is uniformly convex); the map $J^{-1} : X^* \to X$ is a duality map on X^* with gauge function μ^{-1} if, in addition, X is reflexive, X^{**} being identified with X. When X is a Hilbert space, so that X^* may be identified with X, the most natural duality map on X is the identity map, corresponding to the gauge function μ with $\mu(t) = t$.

When $1 < p < \infty$ and $\mu(t) = t^{p-1}$, it may be checked that the duality map J on $L_p(\Omega)$ (where Ω is, say, a domain in $\mathbb{R}^n$) with gauge function μ is given by

$$J(u) = |u|^{p-2} u. \tag{1.12}$$

The duality map J on l_p with the same gauge function μ is defined by $J((x_k)) = (|x_k|^{p-2} x_k)$. Duality maps on Sobolev spaces will be discussed later.

Next we turn to differentiability. Let F be a real-valued function on a Banach space X. We say that F is *Gâteaux-differentiable* at $x_0 \in X$ if there exists $x^* \in X^*$ such that

$$\langle h, x^* \rangle = \lim_{t \to 0} \frac{F(x_0 + th) - F(x_0)}{t} \text{ for all } h \in X. \tag{1.13}$$

The limit above is called the *derivative of F in the direction h*; the functional x^* is denoted by grad $F(x_0)$ and will be referred to as the *gradient* or *Gâteaux derivative* of F at x_0. Another notion of differentiability of F is that of Fréchet: F is said to be *Fréchet-differentiable* at x_0 if there exists a functional $F'(x_0) \in X^*$, called the (Fréchet) derivative of F at x_0, such that

$$\frac{|F(x_0+h)-F(x_0)-\langle h, F'(x_0)\rangle|}{\|h\|} \to 0 \text{ as } \|h\| \to 0. \tag{1.14}$$

Clearly Fréchet-differentiability implies Gâteaux-differentiability, with equality of grad $F(x_0)$ and $F'(x_0)$; in the reverse direction, it can be shown that if grad $F(x)$ exists throughout some neighbourhood of x_0 and is continuous at x_0, then F is Fréchet-differentiable at x_0 and grad $F(x_0) = F'(x_0)$. As Gâteaux derivatives are often easier to calculate than Fréchet derivatives, this result is quite useful.

The case when $F(x) = \|x\|$ $(x \in X)$ is of particular interest.

Proposition 1.8. *Let X be a Banach space with norm $\|\cdot\|$. Then $\|\cdot\|$ is Gâteaux-differentiable on $X\backslash\{0\}$ if and only if X^* is strictly convex.*

Proof. For the moment we suppose that X is real. Let $x, h \in X$. If $0 < s < t$, then the convexity of $y \mapsto \|x+y\|$ implies that

$$\|x+sh\|-\|x\| = \left\|x+\frac{s}{t}th+(1-\frac{s}{t})\cdot 0\right\| - \|x\| \leq \frac{s}{t}\|x+th\| - \|x\|,$$

which shows that $t \mapsto t^{-1}(\|x+th\|-\|x\|)$ is monotone increasing in $t>0$. Moreover, $t^{-1}(\|x+th\|-\|x\|)$ is bounded below, for use of the triangle inequality shows that

$$t^{-1}(\|x+th\|-\|x\|) \geq -\|h\|.$$

Hence $\phi_+(h) := \lim_{t\to 0+} t^{-1}(\|x+th\|-\|x\|)$ exists for all $x, h \in X$, as does $\phi_-(h) := \lim_{t\to 0+} t^{-1}(\|x\|-\|x-th\|)$. Since

$$\|x+th\|-\|x\| \geq \|x\|-\|x-th\|,$$

we see that $\phi_-(h) \leq \phi_+(h)$.

If $x^* \in X^*$ is such that $\|x^*\| = 1$ and $\langle x, x^*\rangle = \|x\|$, then

$$\langle h, x^*\rangle = \frac{\langle x-x+th, x^*\rangle}{t} \geq \frac{\|x\|-\|x-th\|}{t},$$

and so $\phi_-(h) \leq \langle h, x^*\rangle$; similarly, $\langle h, x^*\rangle \leq \phi_+(h)$. Put

$$M^*(x) = \{x^* \in X^* : \phi_-(h) \leq \langle h, x^*\rangle \leq \phi_+(h) \text{ for all } h \in X\}.$$

We claim that

$$M^*(x) = \{x^* \in X^* : \|x^*\| = 1 \text{ and } \langle x, x^*\rangle = \|x\|\}.$$

To justify this, let $x^* \in M^*(x)$ and observe that $\phi_-(x) \leq \langle x, x^* \rangle \leq \phi_+(x)$, while clearly $\phi_-(x) = \phi_+(x) = \|x\|$. Hence $\langle x, x^* \rangle = \|x\|$ and also $\|x^*\| = 1$. The claim is justified.

Now suppose that $\|\cdot\|$ is Gâteaux-differentiable on $X \backslash \{0\}$ and let $x \in X \backslash \{0\}$. Then the functionals ϕ_- and ϕ_+ introduced above coincide, which implies that $\langle h, x^* \rangle = \phi_-(h)$ for all $h \in X$ and hence that $M^*(x)$ has exactly one element. By Proposition 1.3, X^* is strictly convex.

Conversely, suppose that X^* is strictly convex and let $x \in X \backslash \{0\}$. Then $M^*(x)$ consists of a single element, x^* say. We claim that this implies that $x^* = \phi_- = \phi_+$. For fixed $y_0 \in X$ define a functional x_1^* on the space spanned by y_0 according to the prescription $\langle \lambda y_0, x_1^* \rangle = \lambda \phi_+(y_0)$. If $\lambda \geq 0$, $\phi_+(\lambda y_0) = \lambda \phi_+(y_0)$, while if $\lambda < 0$,

$$\lambda \phi_+(y_0) = -|\lambda| \phi_+(y_0) = |\lambda| \phi_-(-y_0) \leq |\lambda| \phi_+(-y_0) = \phi_+(\lambda y_0).$$

Hence $\langle \lambda y_0, x_1^* \rangle \leq \phi_+(\lambda y_0)$ for all $\lambda \in \mathbb{R}$. Moreover, ϕ_+ is subadditive, since for $0 < t \leq s$,

$$\begin{aligned}\phi_+(y_1 + y_2) &\leq (t/2)^{-1} \left\{ \left\| x + \frac{t}{2} y_1 + \frac{t}{2} y_2 \right\| - \|x\| \right\} \\ &= t^{-1} \{ \|2x + t y_1 + t y_2\| - 2\|x\| \} \\ &\leq t^{-1} \{ \|x + t y_1\| - \|x\| \} + s^{-1} \{ \|x + s y_2\| - \|x\| \}.\end{aligned}$$

It follows from the Hahn–Banach theorem that there is a linear extension $x_0^* \in X^*$ of x_1^* such that $\langle y, x_0^* \rangle \leq \phi_+(y)$ on X. Thus $|\langle y, x_0^* \rangle| \leq \|y\|$ for all $y \in X$ and so $x_0^* \in X^*$. Moreover, $\langle -y, x_0^* \rangle \geq -\phi_+(y) = \phi_-(y)$ on X : we conclude that $x_0^* = x^*$. Repetition of this argument with ϕ_- instead of ϕ_+ shows that $\phi_-(y) = \langle y, x^* \rangle = \phi_+(y)$ for all $y \in X$, which means that $\|\cdot\|$ is Gâteaux-differentiable on $X \backslash \{0\}$.

When X is complex the arguments are similar, with re x^* instead of x^*. □

Remark 1.2. (i) Proposition 1.8 leads to the description as *smooth* those spaces X for which X^* is strictly convex; the arguments above show that for such spaces X, if $x \in X \backslash \{0\}$, then grad $\|x\|$ is the unique element $x^* \in X^*$ such that $\langle x, x^* \rangle = \|x\|$ and $\|x^*\| = 1$. This implies that the map $J : X \to X^*$ given by $Jx = \mu(\|x\|)\text{grad}\,\|x\|$ $(x \neq 0)$, $J0 = 0$, is the duality map on X with gauge function μ.

(ii) If X^* is uniformly convex and X is real, it can be shown (see, for example, [96], Lemma II.5.7) that grad $\|x\|$ is the Fréchet derivative of $\|\cdot\|$ at x for each $x \in X \backslash \{0\}$.

The Gâteaux derivative of the norm is quite easy to determine directly in some important cases. For example, suppose that $1 < p < \infty$ and consider the sequence space l_p, with norm $\|x\| = \left(\sum_1^\infty |x_k|^p\right)^{1/p}$. It is easy to see that as $t \to 0$, then for all $x, h \in X$ with $x \neq 0$,

$$t^{-1} \sum_1^\infty \{ |x_k + t h_k|^p - |x_k|^p \} \to p \sum_1^\infty |x_k|^{p-1} \text{ sgn } x_k \cdot h_k.$$

From this we have

$$\text{grad } \|x\| = p^{-1}\|x\|^{1-p} \text{ grad } \|x\|^p = \left(\|x\|^{1-p}|x_k|^{p-1} \text{ sgn } x_k\right), x \neq 0. \tag{1.15}$$

In a similar way it can be shown that the Gâteaux derivative of the norm $\|x\| = \left(\int_\Omega |x(\xi)|^p d\xi\right)^{1/p}$ on $L_p(\Omega)$, where Ω is, for example, an open subset of $\mathbb{R}^n$, is given by

$$(\text{grad } \|x\|)(\xi) = \|x\|^{1-p}|x(\xi)|^{p-1}\text{sgn } x(\xi), \xi \in \Omega. \tag{1.16}$$

Of course, (1.15) and (1.16) also follow immediately from our earlier calculations of duality maps for l_p and $L_p(\Omega)$.

Further illustrations of the usefulness of Proposition 1.8 can be given by means of polar sets. Given any closed linear subspace M of X, we write

$$M^0 = \{x^* \in X^* : \langle x, x^*\rangle_X = 0 \text{ for all } x \in M\},$$

and call this the polar set of M; similarly, if N is a closed linear subspace of X^*, we set

$${}^0N = \{x \in X : \langle x, x^*\rangle_X = 0 \text{ for all } x^* \in N\}.$$

Corresponding notation is used with respect to linear subspaces of other Banach spaces and their duals. The polar set M^0 of any closed linear subspace M of X is isometrically isomorphic to $(X/M)^*$: in fact, with the canonical map of X onto X/M denoted by ϕ, the adjoint ϕ^* of ϕ is an isometric isomorphism of $(X/M)^*$ onto M^0 (see, for example, [15], Chap. IV, Section 5, Prop. 9). Note also that if X is reflexive, then $(M^0)^*$ is isometrically isomorphic to X/M and X/M is reflexive.

Proposition 1.9. *Let M be a closed linear subspace of a reflexive Banach space X. If X is strictly convex, so are M and X/M; if X^* is strictly convex, so are $(X/M)^*$ and M^0.*

Proof. Suppose first that X is strictly convex. Then clearly so is M. Since X is reflexive, the norm on X^*, and hence that on M^0 and thus on $(X/M)^*$ (which is isometrically isomorphic to M^0) is Gâteaux-differentiable at all non-zero points. Hence by Proposition 1.8, the reflexive space X/M is strictly convex. If X^* is strictly convex, then so is its closed subspace M^0, and as $(X/M)^*$ is isometrically isomorphic to M^0, the result follows. □

Remark 1.3. In [32] the relationship between uniform convexity of a Banach space X and that of its quotient spaces is analysed. Let $\mathscr{M}$ be the family of all closed linear subspaces $M \neq \{0\}$ of X. It turns out that X is uniformly convex if and only if all the quotient spaces X/M $(M \in \mathscr{M})$ are uniformly convex, with a common modulus of convexity; the same holds if $\mathscr{M}$ is replaced by $\{M \in \mathscr{M} : \dim(X/M) = 2\}$. This paper also discusses the interaction between uniform convexity and dual spaces. To explain this, we introduce the modulus of smoothness of X : this is the function $\rho_X : (0,\infty) \to [0,\infty)$ defined by

$$\rho_X(\varepsilon) = \sup\left\{ \frac{\|x+y\|+\|x-y\|}{2} - 1 : \|x\| = 1, \|y\| = \varepsilon \right\}.$$

Since

$$2\|x\| = \|(x+y)+(x-y)\| \leq \|x+y\| + \|x-y\|$$

it follows that $\rho_X(\varepsilon) \geq 0$.

The space X is called *uniformly smooth* if

$$\lim_{\varepsilon \to 0} \rho_X(\varepsilon)/\varepsilon = 0.$$

Note that in the definition of $\rho_X(\varepsilon)$ the supremum may be taken over all $x, y \in X$ with $\|x\| \leq 1$ and $\|y\| \leq \varepsilon$ without affecting the outcome. Uniform smoothness means that if x and y are on the unit sphere and lie in a narrow cone, then $\|x+y\|$ is close to 2; loosely speaking, the unit ball of X does not have any corners. Hilbert spaces H are the 'most' uniformly smooth spaces in the sense that

$$\rho_X(\varepsilon) \geq \rho_H(\varepsilon) \text{ for all } \varepsilon > 0.$$

Observe also that X is uniformly smooth if and only if

$$\lim_{t \to 0} \frac{\|x+ty\| - \|x\|}{t} \text{ exists, uniformly on } \{(x,y) : \|x\| = \|y\| = 1\}.$$

The main result of [32] is that X is uniformly convex (respectively, uniformly smooth) if and only if X^* is uniformly smooth (respectively, uniformly convex). For a discussion of this and related notions we refer to [33] (Chap. VII, Sect. 2), [3] and [107], 5.5.2. Further insight may be obtained by means of the concept of super-reflexivity. To explain this we need some terminology: given Banach spaces X and Y, the space X is said to be *finitely representable in* Y if, for each finite-dimensional subspace X_n of X and each $\lambda > 1$, there is an isomorphism T_n of X_n into Y for which

$$\lambda^{-1}\|x\|_X \leq \|T_n x\|_Y \leq \lambda\|x\|_X \text{ for all } x \in X_n.$$

A Banach space X is called *super-reflexive* if no nonreflexive Banach space is finitely representable in X. Each of the following properties is equivalent to super-reflexivity of X (see, for example, [107], 5.5.2.4) :

(a) X^* is super-reflexive.
(b) X has an equivalent norm with respect to which it is uniformly convex.
(c) X^* has an equivalent norm with respect to which it is uniformly convex.
(d) X has an equivalent norm with respect to which it is uniformly smooth.

Given a super-reflexive Banach space X with a Schauder basis $(x_n)_{n\in\mathbb{N}}$, where $\|x_n\|_X = 1$ for all $n \in \mathbb{N}$, it can be shown (see [76] and [70]; see also [61], Chap. 9, Theorem 9.25) that there exist $p, q \in (1, \infty)$ and $K > 0$ such that for every $x = \sum_{n=1}^{\infty} \alpha_n x_n \in X$,

$$K^{-1}\left(\sum_{n=1}^{\infty}|\alpha_n|^q\right)^{1/q} \le \|x\|_X \le K\left(\sum_{n=1}^{\infty}|\alpha_n|^p\right)^{1/p}.$$

In particular, if X is uniformly convex, then there exists $q \in (1,\infty)$ such that for every $x \in X$ the biorthogonal functionals x_n^* corresponding to the x_n satisfy $(x_n^*(x)) \in l_q$.

We see from the proof of Proposition 1.8 that given any Banach space X and any $x,h \in X$, then

$$(x,h)_+ := \|x\| \lim_{t\to 0+} t^{-1}(\|x+th\|-\|x\|) \text{ and } (x,h)_- := \lim_{t\to 0+} t^{-1}(\|x\|-\|x-th\|) \tag{1.17}$$

both exist. Now suppose that $\|\cdot\|$ is Gâteaux-differentiable on $X\backslash\{0\}$. Then $(x,h)_+ = (x,h)_-$ for all $x,h \in X$; we define $(x,h)_X$ to be this common value and call it the *semi-inner product* of x and h. Thus

$$(x,h)_X = \|x\| \langle h, \text{grad } \|x\|\rangle_X, \text{ for } x,h \in X,\ x \neq 0. \tag{1.18}$$

We define $(0,h)_X = 0$ for all $h \in X$. Note that $(x,h)_X$ depends linearly on h and that $(x,x)_X = \|x\|^2$; however, in general, $(x,h)_X \neq (h,x)_X$. By way of illustration we note that when $X = L_p(\Omega)$ and $1 < p < \infty$,

$$(x,h)_X = \|x\|^{2-p}\int_{\Omega}|x(\xi)|^{p-2}\overline{x(\xi)}h(\xi)d\xi. \tag{1.19}$$

This leads us to the notion of orthogonality of elements of a Banach space. Following James [75], we give

Definition 1.1. Let X be a Banach space and let $x,h \in X$. Then x is said to be orthogonal to h, written $x \perp h$, if

$$\|x+\lambda h\| \ge \|x\| \text{ for all } \lambda \in \mathbb{R}; \tag{1.20}$$

x is orthogonal to a subset H of X if $x \perp h$ for all $h \in H$. Orthogonality of subsets M_1, M_2 of X, written $M_1 \perp M_2$, is defined in the obvious way.

Note that in general orthogonality is not symmetric, that is, $x \perp h$ does not imply $h \perp x$. Indeed (see [75]), if X is strictly convex with $\dim X \ge 3$, then if orthogonality is symmetric, X must be an inner-product space.

The linkage between orthogonality and semi-inner products is given by the next result, contained in [75].

Theorem 1.5. *Let X be a Banach space with norm $\|\cdot\|$ that is Gâteaux-differentiable at every non-zero point, and let $x,h \in X$. Then $x \perp h$ if and only if $(x,h)_X = 0$. Moreover, given $x \in X\backslash\{0\}$ and $y \in X$, there is precisely one $\lambda \in \mathbb{R}$ such that $x \perp \lambda x + y$ and this is given by*

$$\langle y, grad\ \|x\|\rangle = -\lambda\,\|x\|. \tag{1.21}$$

Finally we mention decompositions of X and X^*, involving this notion of orthogonality, given by Alber [2]. We say that X is the James orthogonal sum of closed linear subspaces M_1, M_2, written $X = M_1 \uplus M_2$, if (a) each $x \in X$ has a unique decomposition $x = m_1 + m_2$, where $m_1 \in M_1$, $m_2 \in M_2$; (b) $M_1 \cap M_2 = \{0\}$; (c) $M_1 \perp M_2$. His result is the following

Theorem 1.6. *Let M be a closed linear subspace of a uniformly convex and uniformly smooth Banach space X; denote by $J_X : X \to X^*$ the duality map with gauge function μ given by $\mu(t) = t$ $(t \geq 0)$. Then*

$$X = M \uplus J_X^{-1} M^0 \text{ and } X^* = M^0 \uplus J_X M.$$

Further decomposition results will be given in the next section.

1.3 Representations of Compact Linear Operators

Let H be a Hilbert space and suppose that $A : H \to H$ is compact and self-adjoint. Then for all $x \in H$,

$$Ax = \sum_n \lambda_n (x, \phi_n) \phi_n,$$

where the λ_n are eigenvalues of A, each repeated according to multiplicity and ordered by decreasing modulus, while the ϕ_n are orthonormal eigenvectors of A corresponding to the eigenvalues λ_n; here $(\cdot,\cdot)$ denotes the inner product in H. Moreover, if the kernel of A is trivial, then the ϕ_n form a complete orthonormal set in H. All this is classical, as is its extension (by Erhard Schmidt) to arbitrary compact linear operators acting between (possibly different) Hilbert spaces; for details of the proofs see [41], Chap. II, Sect. 5, for example. It turns out (see [44]) that results of a broadly similar character to these can be obtained in a Banach space setting, under some restrictions on the spaces; we describe this work below and make applications in later chapters.

Throughout this section it will be supposed that X and Y are real, reflexive Banach spaces with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$, respectively, and that $T : X \to Y$ is compact and linear. The results to be given also hold for complex spaces, but with minor technical complications in the proofs that we prefer to avoid here. We begin with a simple result.

Proposition 1.10. *There exists $x_1 \in X$, with $\|x_1\|_X = 1$, such that $\|T\| = \|Tx_1\|_Y$.*

Proof. We may assume that $T \neq 0$ as otherwise the result is obvious. Let $\{w_k\}$ be a sequence of elements of X with $\|w_k\|_X = 1$ for all $k \in \mathbb{N}$ and $\lim_{k\to\infty} \|Tw_k\|_Y = \|T\|$. Since $\{w_k\}$ is bounded and X is reflexive, there is a weakly convergent subsequence of $\{w_k\}$, still denoted by $\{w_k\}$ for convenience, with weak limit $w \in X$. As T is compact, $Tw_k \to Tw$. Thus $\|w\|_X \leq \liminf_{k\to\infty} \|w_k\|_X = 1$ and $\|Tw\|_Y = \|T\|$, from which it is immediate that $\|w\|_X = 1$. Now take $x_1 = w$. $\square$

Book-001-2021\Math-Distance\
Curves and Surfaces in
Computer Aided Geometric
Design.... pdf

Pág. 10(20)

Fig. 0.5.

Example of a surface in
which a pentagonal surface
patch occurs.

→

Book-001-2021\Eigenvalues....\
Eigenvalues.... pdf

Pág. 3(16)

1.2 Strict and Uniform
Convexity

Banach Space

$x, y \in X$ are such that
$\|x+y\| = \|x\| + \|y\|$ then

either $y=0$ or $x=\lambda y$, for some $\lambda > 0$.

obj.: Proposition 1.1

20240103: Cosmolect1-11.pdf

pag. 12 (gravitational instability picture)

⇒ pag. 18 Curved space-time

Photons propagate along geodesic curves

...

Pag. 31

— Pág. 46 apparent angular size of that physical size is determined by spatial curvature.

...

either $y=0$ ou $x=\lambda y$, for
some $\lambda \geq 0$.

obs.: Proposition 1.1

20240103:
Cosmolect 1-11.pdf
pag. 12 (gravitational instability picture)

⟹ pag. 18 Curved space-time

Photons propagate along
geodesic curves ...

Pag. 31

— Pag. 46 apparent angular size
of that physical size
is determined by
spatial curvature. ..

4"
-./0- - -1 ! - < 4"
& !; ! ! !? 6 4

- voids have typical diameters of ~ 50 Mpc
- density contrast of structures: $\delta \equiv \frac{\rho - \bar\rho}{\bar\rho} > 1$
- structures (should have) formed via gravitational collapse from seed fluctuations

✲⋙

2. Growth Factor

- gravitational instability picture: density contrast evolves according to

$$\boxed{\ddot\delta + 2\,\frac{\dot S}{S}\,\dot\delta = 4\pi\,G\,\rho\,\delta}$$

- two solutions; one growing, one decaying; growth factor:

$$\boxed{\delta(S) = \delta_0\,D_+(S)}$$

- simple case (Einstein-de Sitter universe):

$$\boxed{D_+(S) = S = (1+z)^{-1}}$$

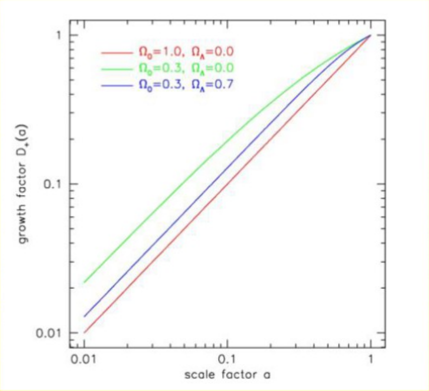

$$\boxed{\frac{\partial f}{\partial t} + \vec{v}\,\frac{\partial f}{\partial \vec{x}} - \nabla\Phi\,\frac{\partial f}{\partial \vec{v}} = C[f]}$$

- integration over momentum space: eq. for temperature fluctuations

2. Curved Space-time

- photons propagate along geodesic curves:

$$\boxed{\frac{\mathrm{d}^2 x^\mu}{\mathrm{d}\lambda^2} + \Gamma^\mu{}_{\nu\rho}\,\frac{\mathrm{d}x^\nu}{\mathrm{d}\lambda}\,\frac{\mathrm{d}x^\rho}{\mathrm{d}\lambda} = 0}$$

($\Gamma^\mu{}_{\nu\rho}$: Christoffel symbols; λ: affine parameter)

- photon momentum can be written:

$$\boxed{p^\mu = \frac{\mathrm{d}x^\mu}{\mathrm{d}\lambda}}$$

- weakly perturbed metric ($c^2\,\Phi$: Newtonian potential):

$$\boxed{\mathrm{d}s^2 = (1 + 2\,\Phi)\;\mathrm{d}t^2 - S^2(t)\;(1 - 2\,\Phi)\;\mathrm{d}\vec{x}^{\,2}}$$

- change of photon momentum:

$$\boxed{\frac{1}{p}\frac{\mathrm{d}p}{\mathrm{d}t} = -\left(\frac{1}{S(t)}\frac{\mathrm{d}S}{\mathrm{d}t} - \frac{\partial\Phi}{\partial t} + \frac{c}{S(t)}\,\vec{\gamma}\,\frac{\partial\Phi}{\partial\vec{x}}\right)}$$

with $\gamma^i = p^i/|\vec{p}\,|$

3. Collision Terms

- collisions between photons and electrons: Compton scattering
- photon energy at recombination: $k_{\mathrm{B}}T \sim 0.3\ \mathrm{eV} \ll m_e c^2$; Compton scattering can be approximated by Thomson scattering:

$$\boxed{\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \frac{3}{16\pi}\,\sigma_{\mathscr{T}}\left(1+\cos^2\theta\right)}$$

- collision term:

$$\boxed{C[f] = \frac{3}{16\pi}\,\sigma_{\mathscr{T}}\,c\,n_e\int(f'-f)\left(1+\cos^2\theta\right)\mathrm{d}^2\Omega}$$

- additional equations describing matter flow: hydrodynamic equations (continuity, Euler equation)

& '

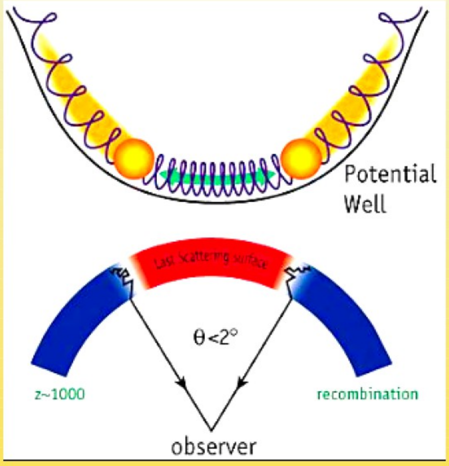
Potential
Well
Last scattering surface
θ<2°
z~1000
recombination
observer

Gravity tries to make matter fall into potential wells

Radiation pressure pushed back...

Oscillations results

& +

$$\boxed{\lambda_{\max} = \frac{c}{\sqrt{3}} t_{\rm rec} \approx 240\,000 \text{ Lj}}$$

- apparent angular size of that physical size is determined by spatial curvature
- curvature can be read off location of first acoustic peak

either $y=0$ ou $x=\lambda y$, for some $\lambda \geq 0$.

obj.: Proposition 1.1

20240103: Cosmolect1-11.pdf

Pag. 12 (gravitational instability picture)

⟹ Pág 18 Curved space-time

Photons propagate along geodesic curves

...

Pag. 31

— Pág. 46 apparent angular size of that physical size is determined by spatial curvature.

...

& ,

Antenokeeffe

geometrias.blogspot.com/2020/
05/antenokeeffe-reconstrucao.html

/2020/05/dual-do-problema-
de-malfatti-da-entrada.html

/2020/04/o-original-problema-
de-malfatti.html

2020/02/entre-triangulos-
porismo-e.html

2020/01/circumcentros-
o-o-o-b-e-o-c-de-um.
html

polarprof-001-site1.htempurl.com/
geometriagon

Thomas Andrews (1873 - 1912)
projetista naval irlandês do Titanic
→ morreu no navio (inafundável)

William Bullock (1813 - 1867)
Em 1863 automatizou a impressora rotativa.

Em 1867 quando instalava uma de suas máquinas rotativas automáticas, ele chutou uma correia que se soltou sozinha e acabou prendendo a sua perna. Morreu do procedimento de amputação.

Otto Lilienthal (1848 - 1896)
engenheiro mecânico alemão, o "pai dos voos de planador"

2000 testes e 5 horas nos céus, morreu um dia após despencar 17 metros com seu planador Normal Glider e fraturar a coluna.

"Sacrifícios precisam ser feitos", teria dito

Israel - Muro das Lamentações

SMALL WAILING WALL
(pequeno lamento)

rosenhebrewschool.com

tau.ac.il

⇒ tetraon.dk/book-inside/

"Aliá"

Lei do Retorno.

$$\operatorname{sen}\theta^2 + \cos\theta^2 = 1$$

$$\begin{cases} -1 \leq \operatorname{sen}\theta \leq 1 \\ -1 \leq \cos\theta \leq 1 \end{cases}$$

⇒ Regra dos ângulos complementares
(de um triângulo retângulo)

Ângulos:

& ,

caterva $\longrightarrow$

1) $sen(\alpha) = sen(\beta)$

$\alpha = \beta + 2k\pi \text{ , } k \in \mathbb{Z}$

2) $\cos(\alpha) = \cos(\beta)$

mesma obsuitta,

3) $tg(\alpha) = tg(\beta)$

$\alpha = (\beta + \pi) + 2k\pi =$

$\beta + (2k+1)\pi$, para $k \in \mathbb{Z}$

Logo

$\alpha = \beta + k\pi$

20210105

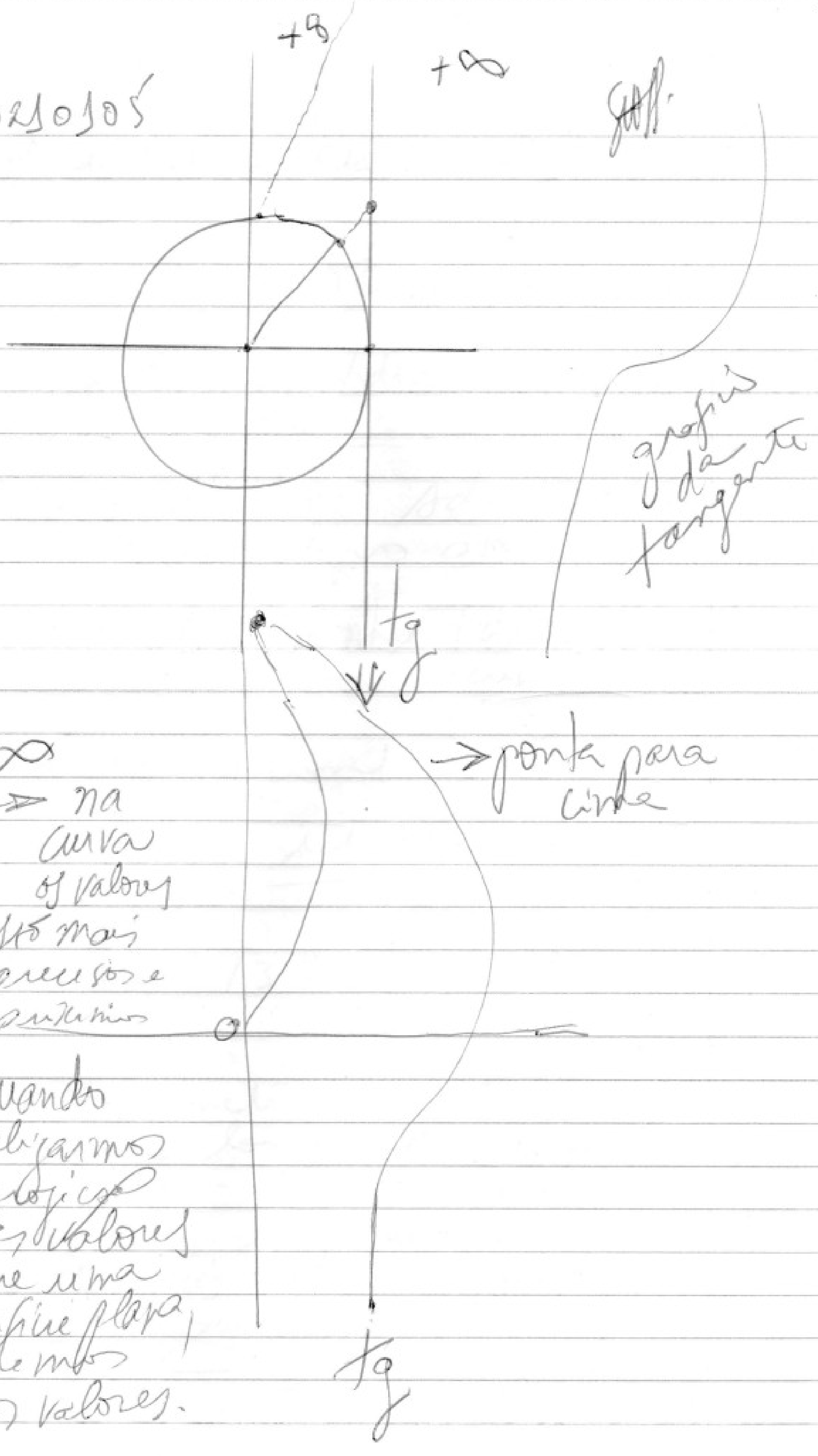
+8
+∞
grafico da tangente
tg
→ ponta para cima
tg

+∞
↳ na curva os valores são mais precisos e próximos

Quando realizarmos a projeção destes valores sobre uma superfície plana, obtemos outros valores.

1? 2 "! JJJ

& ,<

20210106 (20210106)
Upload Books-2021-02.7z

—

Terra = 1,4602 seg + rápido.

86400 seg /p/ dia   19 Julho 2020

1970 → até hoje adicionado 27 seg
no tempo atómico internacional.

31 Dez ou 30 Junho (datas que são
acrescentados).

"leap second"

1 seg será deletado.

Universo 13,77 b anos ± 40 m

—

interferometria de linha de
base muito longa. (medição de origem
de uma fonte)
de rádio no Universo.

(usa-se várias antenas separadas por
milhares de quilómetros.

—

&

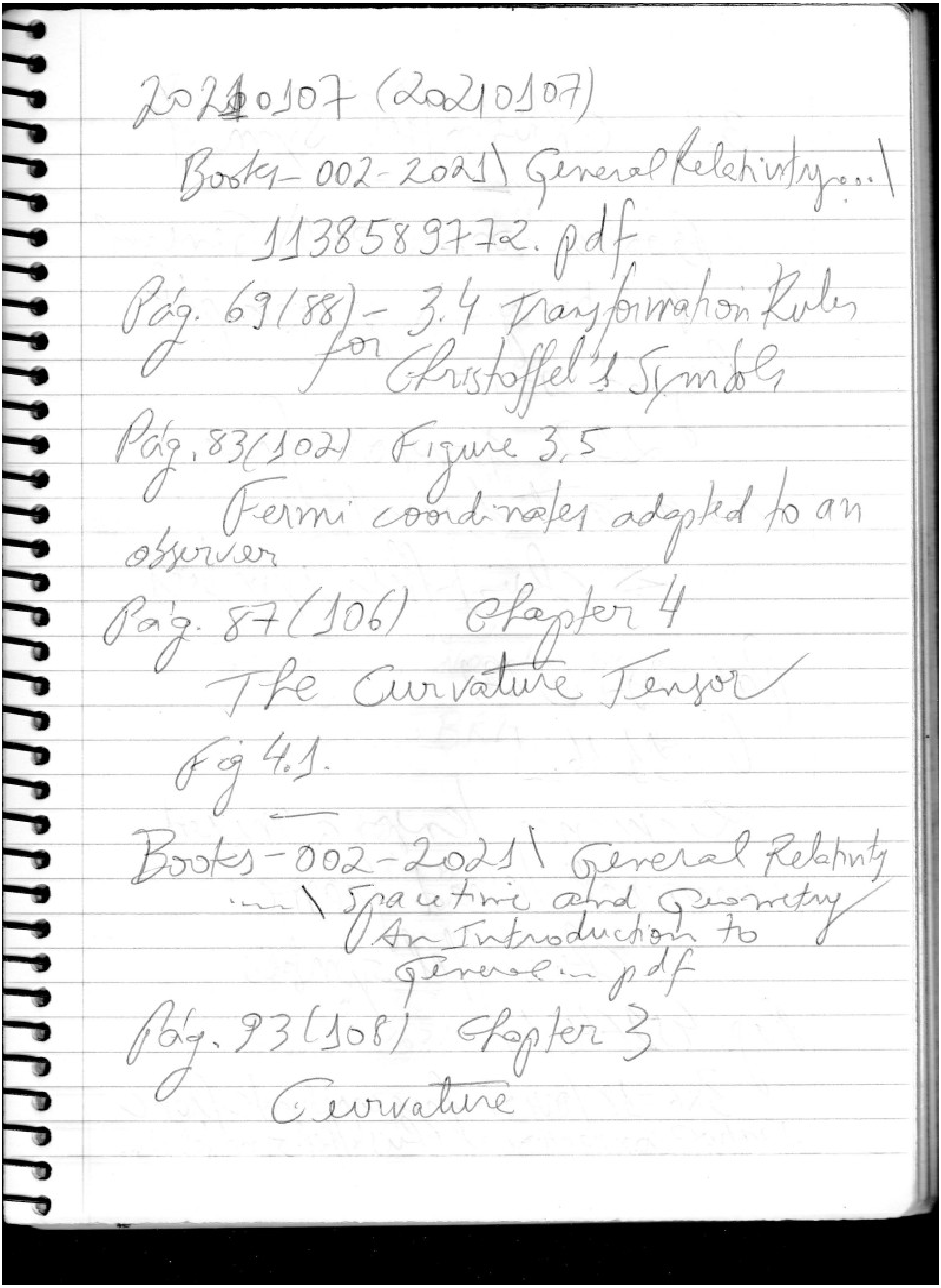

20210107 (20210107)

Books-002-2021\ General Relativity...\
1138589772.pdf

Pág. 69(88) - 3.4 Transformation Rules
for Christoffel's Symbols

Pág. 83(102) Figure 3.5
Fermi coordinates adapted to an
observer

Pág. 87(106) Chapter 4
The Curvature Tensor

Fig 4.1.

Books-002-2021\ General Relativity
...\ Spacetime and Geometry
An Introduction to
General... pdf

Pág. 93(108) Chapter 3
Curvature

20210107 (20210107)

Books-002-2021\ General Relativity...\
1138589772.pdf

Pág. 69(88) - 3.4 Transformation Rules
for Christoffel's Symbols

Pág. 83(102) Figure 3.5
Fermi coordinates adapted to an
observer

Pág. 87(106) Chapter 4
The Curvature Tensor

Fig 4.1.

Books-002-2021\ General Relativity
...\ Spacetime and Geometry
An Introduction to
General... pdf

Pág. 93(108) Chapter 3
Curvature

The second covariant derivative $\nabla\nabla\Phi$ is a $\binom{0}{2}$ tensor, the components of which can be computed using Eq. 3.35

$$\nabla_\beta\nabla_\alpha\Phi \equiv \Phi_{,\beta;\alpha} = \Phi_{,\beta,\alpha} - \Phi_{,\mu}\Gamma^{\mu}_{\beta\alpha}\,. \tag{3.37}$$

In a LIF, since $\Gamma^{\mu}_{\beta\alpha} = 0$, covariant and ordinary derivative coincide, and Eq. 3.37 becomes

$$\nabla_\alpha\nabla_\beta\Phi = \Phi_{,\beta;\alpha} = \Phi_{,\beta,\alpha}\,, \tag{3.38}$$

and since ordinary partial derivatives commute, i.e. $\Phi_{,\beta,\alpha} = \Phi_{,\alpha,\beta}$, it follows that in a LIF

$$\nabla_\alpha\nabla_\beta\Phi = \nabla_\beta\nabla_\alpha\Phi\,. \tag{3.39}$$

This equation shows that the tensor $\nabla\nabla\Phi$ is symmetric, and since symmetry is a tensorial property, it must hold in any frame; therefore from Eqs. 3.39 and 3.37 we find

$$\Phi_{,\beta,\alpha} - \Phi_{,\mu}\Gamma^{\mu}_{\beta\alpha} = \Phi_{,\alpha,\beta} - \Phi_{,\mu}\Gamma^{\mu}_{\alpha\beta} \tag{3.40}$$

in any coordinate system. It follows that for any Φ

$$\Phi_{,\mu}\Gamma^{\mu}_{\beta\alpha} = \Phi_{,\mu}\Gamma^{\mu}_{\alpha\beta}\,, \tag{3.41}$$

and consequently

$$\Gamma^{\mu}_{\beta\alpha} = \Gamma^{\mu}_{\alpha\beta}\,, \tag{3.42}$$

i.e. *Christoffel's symbols are symmetric in the lower indices.*

3.4 TRANSFORMATION RULES FOR CHRISTOFFEL'S SYMBOLS

According to Eq. 1.38 (or to Eq. 3.9), Christoffel's symbols are

$$\Gamma^{\rho}_{\tau\sigma} = \frac{\partial x^\rho}{\partial \xi^\mu}\frac{\partial^2 \xi^\mu}{\partial x^\tau \partial x^\sigma}\,.$$

Let us consider a general coordinate transformation $x^\mu = x^\mu(x^{\alpha'})$. In the new frame Christoffel's symbols are

$$\begin{aligned}
\Gamma^{\lambda'}_{\mu'\nu'} &= \frac{\partial x^{\lambda'}}{\partial \xi^\alpha}\frac{\partial^2 \xi^\alpha}{\partial x^{\mu'}\partial x^{\nu'}} = \\
&= \frac{\partial x^{\lambda'}}{\partial x^\rho}\frac{\partial x^\rho}{\partial \xi^\alpha}\frac{\partial}{\partial x^{\mu'}}\left(\frac{\partial \xi^\alpha}{\partial x^\sigma}\frac{\partial x^\sigma}{\partial x^{\nu'}}\right) = \\
&= \frac{\partial x^{\lambda'}}{\partial x^\rho}\frac{\partial x^\rho}{\partial \xi^\alpha}\left[\frac{\partial^2 \xi^\alpha}{\partial x^\tau \partial x^\sigma}\frac{\partial x^\tau}{\partial x^{\mu'}}\frac{\partial x^\sigma}{\partial x^{\nu'}} + \frac{\partial \xi^\alpha}{\partial x^\sigma}\frac{\partial^2 x^\sigma}{\partial x^{\mu'}\partial x^{\nu'}}\right] = \\
&= \frac{\partial x^{\lambda'}}{\partial x^\rho}\frac{\partial x^\tau}{\partial x^{\mu'}}\frac{\partial x^\sigma}{\partial x^{\nu'}}\Gamma^{\rho}_{\tau\sigma} + \frac{\partial x^{\lambda'}}{\partial x^\rho}\frac{\partial^2 x^\rho}{\partial x^{\mu'}\partial x^{\nu'}}\,.
\end{aligned} \tag{3.43}$$

It should be noted that if Christoffel's symbols were the components of a tensor, only the first term in the last row of Eq. 3.43 would be present (see Eq. 2.158). Therefore, as pointed out in Sec. 3.1, Christoffel's symbols are not the components of a tensor.

However, the $\Gamma^{\gamma}_{\alpha\beta}$'s transform as tensors for *linear transformations*, i.e. for coordinate transformations of the form $x'^\mu = \Lambda^{\mu'}{}_\alpha x^\alpha$ where $\Lambda^{\mu'}{}_\alpha$ are constants. Indeed in this case $\frac{\partial^2 x^\rho}{\partial x^{\mu'}\partial x^{\nu'}} = 0$.

Box 3-B

Exercise

Compute Christoffel's symbols $\Gamma^{\mu}_{\alpha\beta}$ using the affine connection

$$\frac{\partial \vec{e}_{(\alpha)}}{\partial x^{\beta}} = \Gamma^{\mu}_{\alpha\beta}\vec{e}_{(\mu)}\,. \tag{3.44}$$

Let us consider for example a two-dimensional flat space in polar coordinates, i.e. $(x^{1'}, x^{2'}) \equiv (r, \theta)$; the basis vectors are related to the coordinate basis associated to Cartesian coordinates by Eqs. 2.73

$$\begin{aligned} \vec{e}_{(1')} &= \vec{e}_{(r)} = \cos\theta\vec{e}_{(1)} + \sin\theta\vec{e}_{(2)}\,, \\ \vec{e}_{(2')} &= \vec{e}_{(\theta)} = -r\sin\theta\vec{e}_{(1)} + r\cos\theta\vec{e}_{(2)}\,. \end{aligned} \tag{3.45}$$

We shall indicate $(\vec{e}_{(1)}, \vec{e}_{(2)})$ with $(\vec{e}_{(x)}, \vec{e}_{(y)})$, and $(\vec{e}_{(1')}, \vec{e}_{(2')})$ with $(\vec{e}_{(r)}, \vec{e}_{(\theta)})$. From Eqs. 3.45 we find

$$\frac{\partial \vec{e}_{(r)}}{\partial r} = \frac{\partial}{\partial r}(\cos\theta\vec{e}_{(x)} + \sin\theta\vec{e}_{(y)}) = 0\,, \tag{3.46}$$

and consequently, from Eq. 3.44

$$\Gamma^{\mu}_{rr}\vec{e}_{(\mu)} = \Gamma^{r}_{rr}\vec{e}_{(r)} + \Gamma^{\theta}_{rr}\vec{e}_{(\theta)} = 0 \longrightarrow \Gamma^{r}_{rr} = \Gamma^{\theta}_{rr} = 0\,. \tag{3.47}$$

Moreover

$$\begin{aligned} \frac{\partial \vec{e}_{(r)}}{\partial \theta} &= \frac{\partial}{\partial \theta}(\cos\theta\vec{e}_{(x)} + \sin\theta\vec{e}_{(y)}) = \\ &= -\sin\theta\vec{e}_{(x)} + \cos\theta\vec{e}_{(y)} = \frac{1}{r}\vec{e}_{(\theta)}\,; \end{aligned}$$

therefore

$$\frac{1}{r}\vec{e}_{(\theta)} = \Gamma^{\mu}_{r\theta}\vec{e}_{(\mu)} = \Gamma^{r}_{r\theta}\vec{e}_{(r)} + \Gamma^{\theta}_{r\theta}\vec{e}_{(\theta)} \longrightarrow \Gamma^{r}_{r\theta} = 0 \;\;,\;\; \Gamma^{\theta}_{r\theta} = \frac{1}{r}\,. \tag{3.48}$$

Proceeding along these lines it is easy to show that

$$\Gamma^{r}_{\theta r} = 0\,,\; \Gamma^{\theta}_{\theta r} = \frac{1}{r}\,,\; \Gamma^{r}_{\theta\theta} = -r\,,\; \Gamma^{\theta}_{\theta\theta} = 0\,. \tag{3.49}$$

3.5 THE COVARIANT DERIVATIVE OF TENSORS

Following the approach adopted in Sec. 3.2 we shall now define the covariant derivative of tensors of any rank. Let us consider, as an example, the covariant derivative of a $\binom{0}{2}$ tensor. Given the tensor $\mathbf{T} = T_{\alpha\beta}\,\tilde{\omega}^{(\alpha)} \otimes \tilde{\omega}^{(\beta)}$ and a vector $\vec{V}$, let $\tilde{q}$ be the one-form obtained by contracting $\mathbf{T}$ with $\vec{V}$, i.e., in components

$$q_{\alpha} = T_{\alpha\beta}V^{\beta}\,. \tag{3.50}$$

According to Eq. 3.35 the covariant derivative of $\tilde{q}$ is a $\binom{0}{2}$ tensor with components

$$\nabla_{\mu}q_{\alpha} = q_{\alpha,\mu} - \Gamma^{\sigma}_{\mu\alpha}q_{\sigma}\,, \tag{3.51}$$

This is a tensor equation, and must hold in any coordinate frame; therefore, in a generic frame the equation of motion of the free particle is

$$u^\alpha u^\mu{}_{;\alpha} = 0\,. \tag{3.121}$$

If we expand the covariant derivative

$$u^\alpha u^\mu{}_{;\alpha} = u^\alpha u^\mu{}_{,\alpha} + u^\alpha \Gamma^\mu{}_{\alpha\beta} u^\beta\,, \tag{3.122}$$

and replace $u^\mu = dx^\mu/d\lambda$, Eq. 3.121 becomes

$$\frac{d^2x^\mu}{d\lambda^2} + \Gamma^\mu_{\alpha\beta}\frac{dx^\alpha}{d\lambda}\frac{dx^\beta}{d\lambda} = 0\,, \tag{3.123}$$

which is the geodesic equation, Eq. 3.114. Thus, the geodesic equation can also be written in the form 3.121, which describes the parallel transport of the tangent vector $\vec{u}$ along the geodesic. This means that if we parallely transport the tangent vector from a point $\mathbf{p}$ to a point $\mathbf{p}'$ along the geodesic line, the transported vector in $\mathbf{p}'$ will be tangent to the curve. Consequently, a curve $\mathcal{C}$ with tangent vector $\vec{u}$ is a geodesic if

$$\nabla_{\vec{u}}\ \vec{u} = 0\,. \tag{3.124}$$

Therefore, *geodesics are those curves which parallel-transport their own tangent vectors.*

The parameter along the geodesic is not unique. Let s be a new parameter; since $\frac{d}{d\lambda} = \frac{d}{ds}\frac{ds}{d\lambda}$, Eq. 3.123 becomes

$$\frac{d^2x^\alpha}{ds^2} + \Gamma^\alpha_{\mu\nu}\left[\frac{dx^\mu}{ds}\frac{dx^\nu}{ds}\right] = -\left[\frac{d^2s}{d\lambda^2}\Big/\left(\frac{ds}{d\lambda}\right)^2\right]\frac{dx^\alpha}{ds}\,. \tag{3.125}$$

This equation reduces to Eq. 3.123 only if s is related to λ by a linear transformation

$$s = a\lambda + b, \qquad a, b = \text{const}\,, \tag{3.126}$$

in which case the right-hand side of Eq. 3.125 vanishes. The parameters for which the geodesic equation takes the form 3.123 are called **affine parameters** and are all related by a linear transformation.

Eq. 3.121 describes timelike, spacelike, and null geodesics. If the geodesic is timelike, i.e. $\vec{u}\cdot\vec{u} < 0$, it represents the wordline of a massive particle; in this case, by performing the linear transformation in Eq. 3.126 it is possible to change the affine parameter in such a way that the new parameter is the particle proper time (see Box 2-F) multiplied by c, i.e. the parameter τ appearing in Eq. 3.114. With this choice, $\vec{u}\cdot\vec{u} = -1$. If the geodesic is a null curve, i.e. $\vec{u}\cdot\vec{u} = 0$, it represents the wordline of a massless particle; in this case it is not possible to choose the proper time as affine parameter, because it cannot be defined; the geodesic can be parametrized with a different affine parameter, for instance with the arc length. If the geodesic is spacelike, i.e. $\vec{u}\cdot\vec{u} > 0$, it does not represent the worldline of a physical particle.

3.9 FERMI COORDINATES

In this section we shall show how to construct a coordinate frame adapted to an observer. This frame is especially useful to describe physical experiments.

An observer in General Relativity is characterized by a timelike curve, $\tau \mapsto \gamma(\tau)$, where τ is the proper time (times c). The tangent vector to the curve is the four-velocity of the

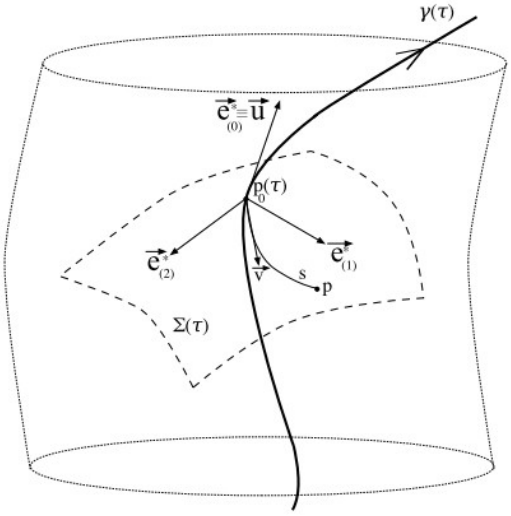

Figure 3.5: Fermi coordinates adapted to an observer. For simplicity, we do not show one of the three space dimensions.

observer $\vec{u}$, and $\vec{u} \cdot \vec{u} = -1$. The observer can either be freely-falling (i.e. with vanishing acceleration $a^\mu = u^\nu u^\mu{}_{;\nu} = 0$) or accelerated ($a^\mu \neq 0$).

It is possible to define, in a region close to the curve (a "worldtube" surrounding γ), coordinates $\{x^{*\mu}\}$ adapted to the observer, called the *Fermi coordinates*, as follows (see Fig. 3.5):

- At any point of the curve $\mathbf{p}_0(\tau) \in \gamma$, we can define a basis $\{\vec{e}^{\,*}_{(\mu)}\}$ of the vector space $\mathbf{T}_{\mathbf{p}_0}$ such that $\vec{e}^{\,*}_{(0)} \equiv \vec{u}$, and $\vec{e}^{\,*}_{(i)}$ $(i = 1, \ldots, 3)$ are orthogonal to $\vec{e}^{\,*}_{(0)}$ and to each other. In addition, we impose that $\vec{e}^{\,*}_{(i)}$ have norm one. Thus, the basis is orthonormal ($\vec{e}^{\,*}_{(\mu)} \cdot \vec{e}^{\,*}_{(\nu)} = \eta_{\mu\nu}$), and the timelike basis vector coincides with the four-velocity of the observer.

- At each value of τ, the three spacelike vectors $\vec{e}^{\,*}_{(i)}$ locally define a three-dimensional spacelike surface Σ, to which they are tangent in $\mathbf{p}_0(\tau)$.

- Given a point $\mathbf{p}$ in the worldtube, such that $\mathbf{p} \in \Sigma(\tau)$, let us consider a spacelike geodesic from $\mathbf{p}_0(\tau)$ to $\mathbf{p}$ on the surface $\Sigma(\tau)$. Let $\vec{v}$ be the tangent vector in $\mathbf{p}_0(\tau)$ to this geodesic with unit length. Since the geodesic belongs to $\Sigma(\tau)$, $\vec{v}$ is a linear combination of the three spacelike vectors $\vec{e}^{\,*}_{(i)}$ $(i = 1, \ldots, 3)$, i.e.

$$\vec{v} = v^i \vec{e}^{\,*}_{(i)} \,. \tag{3.127}$$

- Let s be the affine parameter of the geodesic from $\mathbf{p}_0(\tau)$ to $\mathbf{p}$. We define the Fermi coordinates of the point $\mathbf{p}$ as follows:

$$x^{*0} = \tau \,, \qquad x^{*i} = s v^i \,. \tag{3.128}$$

  It can be shown that s is also the proper length of the path (as defined in Eq. 2.211) of this spacelike geodesic between $\mathbf{p}_0$ and $\mathbf{p}$.

- Repeating the same procedure for all points of the observer worldline, we can define, in the entire worldtube, the Fermi coordinate system $\{x^{*\mu}\}$ adapted to the observer worldline.

It can be shown that (with an appropriate choice of the three-dimensional basis $\{\vec{e}^{*}_{(i)}\}$ for each surface $\Sigma(\tau)$) the spacetime metric in Fermi coordinates is

$$ds^2 = -(1 + 2a_\mu x^{*\mu})(dx^{*0})^2 + \delta_{ij} dx^{*i} dx^{*j} + O(x^{*2})\,, \tag{3.129}$$

where a^ν are the components of the observer's four-acceleration. If the observer is freely-falling ($\vec{a} = 0$), $ds^2 = \eta_{\mu\nu} + O(x^{*2})$, and this coordinate system coincides with the LIF centered in $\mathbf{p}_0$. Indeed, in a LIF the metric coincides with Minkowski's metric modulo *quadratic* terms in the coordinates. If, instead, the observer is accelerated, the deviation of the metric 3.129 from Minkowski's metric is *linear* in the coordinates $x^{*\mu}$: $g_{00} - \eta_{00} = -2a_\mu x^{*\mu}$. The linear term is precisely due to the acceleration of the observer.

Remarkably, the metric 3.129 does not depend on the spacetime curvature (up to $O(x^{*2})$). Indeed, introducing the Fermi coordinates for an accelerated observer in Special Relativity, i.e. $\xi^\alpha \to x^{*\mu}(\xi^\alpha)$, leads to the metric $ds^2 = -(1+2a_\mu x^{*\mu})(dx^{*0})^2 + \delta_{ij} dx^{*i} dx^{*j}$, which differs from Eq. 3.129 by the $O(x^{*2})$ terms only. We conclude that due to the Equivalence Principle, according to which the laws of physics in a LIF take, locally, the form prescribed by Special Relativity for inertial frames, the laws of physics in a Fermi frame take, locally, the form prescribed by Special Relativity for *accelerated* frames.

3.10 NON-COORDINATE BASES

In Boxes 2-G and 2-P we showed that changing from Minkowskian coordinates $\{\xi^{\alpha'}\} \equiv (ct, x, y)$ to polar coordinates $\{x^\alpha\} \equiv (ct, r, \theta)$, the coordinate basis for vectors[3]

$$\{\vec{e}_{(\alpha')}\} = \begin{cases} \vec{e}_{(0')} & \to & (1,0,0) \\ \vec{e}_{(1')} & \to & (0,1,0) \\ \vec{e}_{(2')} & \to & (0,0,1) \end{cases} \tag{3.130}$$

transforms to $\{\vec{e}_{(\alpha)}\}$, where (see Eq. 2.73)

$$\begin{cases} \vec{e}_{(0)} = \vec{e}_{(0')} \\ \vec{e}_{(1)} = \vec{e}_{(r)} = \cos\theta \vec{e}_{(1')} + \sin\theta \vec{e}_{(2')} \\ \vec{e}_{(2)} = \vec{e}_{(\theta)} = -r\sin\theta \vec{e}_{(1')} + r\cos\theta \vec{e}_{(2')}\,. \end{cases} \tag{3.131}$$

These expressions follow from the transformation laws

$$\vec{e}_{(\alpha)} = \Lambda^{\mu'}{}_{\alpha}\; \vec{e}_{(\mu')}\,, \tag{3.132}$$

where $\Lambda^{\mu'}{}_{\alpha} = \frac{\partial x^{\mu'}}{\partial x^\alpha}$ and $\Lambda^{\alpha}{}_{\mu'} = \frac{\partial x^\alpha}{\partial x^{\mu'}}$. We now want to choose a different basis for vectors. For example, while the vectors $\{\vec{e}_{(\alpha)}\}$ are not normalized because

$$\vec{e}_{(\alpha)} \cdot \vec{e}_{(\beta)} = g_{\alpha\beta} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & r^2 \end{pmatrix} \neq \eta_{\alpha\beta}\,, \tag{3.133}$$

---

[3] Note that, at variance with Boxes 2-G and 2-P, we indicate with a prime the Minkowskian coordinates.

we may choose a basis of normalized vectors given by

$$\begin{cases} \vec{e}_{(\hat{0})} = \vec{e}_{(0)} \\ \vec{e}_{(\hat{r})} = \vec{e}_{(r)} \\ \vec{e}_{(\hat{\theta})} = \frac{1}{r}\vec{e}_{(\theta)}\,, \end{cases} \tag{3.134}$$

such that

$$\vec{e}_{(\hat{\alpha})} \cdot \vec{e}_{(\hat{\beta})} = \eta_{\hat{\alpha}\hat{\beta}}\,. \tag{3.135}$$

In this basis the transformation 3.131 becomes

$$\begin{aligned} \vec{e}_{(\hat{r})} &= \cos\theta\vec{e}_{(1')} + \sin\theta\vec{e}_{(2')} \\ \vec{e}_{(\hat{\theta})} &= -\sin\theta\vec{e}_{(1')} + \cos\theta\vec{e}_{(2')}\,. \end{aligned} \tag{3.136}$$

In the previous sections we have always introduced the basis vectors as the coordinate basis associated to a coordinate frame, but they can also be defined independently of the choice of the coordinate frame, as in Eq. 3.134. In this case, it is natural to ask whether there exists a set of coordinates $\{x^{\hat{\alpha}}\}$ such that

$$e_{(\hat{\alpha})} = \Lambda^{\beta}{}_{\hat{\alpha}}\ \vec{e}_{(\beta)} = \frac{\partial x^{\beta}}{\partial x^{\hat{\alpha}}}\ \vec{e}_{(\beta)}\,, \tag{3.137}$$

so that the basis $\{\vec{e}_{(\hat{\alpha})}\}$ is a coordinate basis. The same question can be formulated for the basis one-forms, i.e. whether there exist coordinates $\{x^{\hat{\alpha}}\}$ such that

$$\tilde{\omega}^{(\hat{\alpha})} = \Lambda^{\hat{\alpha}}{}_{\beta}\ \tilde{\omega}^{(\beta)} = \frac{\partial x^{\hat{\alpha}}}{\partial x^{\beta}}\ \tilde{\omega}^{(\beta)} \tag{3.138}$$

and

$$\Lambda^{\hat{\alpha}}{}_{\beta} = \frac{\partial x^{\hat{\alpha}}}{\partial x^{\beta}}\,. \tag{3.139}$$

For instance, in the considered example

$$\Lambda^{\beta}{}_{\hat{\alpha}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{r} \end{pmatrix} \qquad \text{and} \qquad \Lambda^{\hat{\alpha}}{}_{\beta} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & r \end{pmatrix}\,. \tag{3.140}$$

If Eq. 3.139 is true, the following condition must be satisfied

$$\frac{\partial}{\partial x^{\gamma}}\Lambda^{\hat{\alpha}}{}_{\beta} = \frac{\partial^2 x^{\hat{\alpha}}}{\partial x^{\gamma}\partial x^{\beta}} = \frac{\partial^2 x^{\hat{\alpha}}}{\partial x^{\beta}\partial x^{\gamma}} = \frac{\partial}{\partial x^{\beta}}\Lambda^{\hat{\alpha}}{}_{\gamma}\,. \tag{3.141}$$

This is an “integrability condition” that all the components of $\Lambda^{\hat{\alpha}}{}_{\gamma}$ must satisfy in order for the coordinates $\{x^{\hat{\alpha}}\}$ to exist. In the considered example Eq. 3.141 gives

$$\frac{\partial}{\partial\theta}\Lambda^{\hat{2}}{}_{1} = \frac{\partial}{\partial r}\Lambda^{\hat{2}}{}_{2} \qquad \to \qquad 0 = 1 \tag{3.142}$$

which is certainly not true. We conclude that the normalized vector basis in Eq. 3.134 and the dual basis for one forms $\{\tilde{\omega}^{(\hat{\alpha})}\}$ *are not* coordinate bases, since we cannot associate a coordinate transformation to them.

CHAPTER 4

The curvature tensor

In this chapter we shall introduce the curvature tensor, showing that it describes the curvature of the spacetime. This derivation is based on the parallel transport of a vector along a closed loop. We shall discuss the main properties of the curvature tensor and show how the latter is related to the equation of geodesic deviation.

4.1 PARALLEL TRANSPORT ALONG A LOOP

Let us consider an infinitesimal closed loop, having as sides the coordinate lines $x^1 = a$, $x^1 = a + \delta a$, $x^2 = b$, $x^2 = b + \delta b$ (see Fig. 4.1).

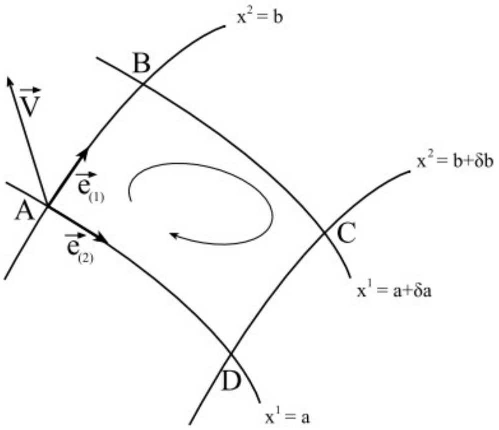

Figure 4.1: A generic vector $\vec{V}$ is parallely transported along an infinitesimal, closed loop whose boundaries are portions of coordinate lines.

We wish to compute how the components of a generic vector $\vec{V}$ change when the vector is parallely transported along the loop. The loop can be divided into four parts, so that the total variation of the vector components is

$$\delta V^\alpha = \delta V^\alpha_{AB} + \delta V^\alpha_{BC} + \delta V^\alpha_{CD} + \delta V^\alpha_{DA} \,. \tag{4.1}$$

The vector tangent to the curves AB and CD and that tangent to the curves BC and DA are, respectively, the basis vectors $\vec{e}_{(1)}$ and $\vec{e}_{(2)}$; since their only non-vanishing components are $e^1_{(1)} = 1$ and $e^2_{(2)} = 1$, the equation of parallel transport, Eq. 3.92, gives

$$\text{on AB and CD :} \qquad \nabla_{\vec{e}_{(1)}}\vec{V} = e^{\mu}_{(1)} V^{\alpha}{}_{;\mu} = 0 \quad \rightarrow \quad \frac{\partial V^{\alpha}}{\partial x^1} + \Gamma^{\alpha}_{\beta 1} V^{\beta} = 0\,, \tag{4.2}$$

$$\text{on BC and DA :} \qquad \nabla_{\vec{e}_{(2)}}\vec{V} = e^{\mu}_{(2)} V^{\alpha}{}_{;\mu} = 0 \quad \rightarrow \quad \frac{\partial V^{\alpha}}{\partial x^2} + \Gamma^{\alpha}_{\beta 2} V^{\beta} = 0\,. \tag{4.3}$$

Eq. 4.2 has to be integrated along the line $x^2 = b$ from A to B, and $x^2 = b + \delta b$ from C to D; Eq. 4.3 has to be integrated along $x^1 = a + \delta a$ from B to C, and on $x^1 = a$ from D to A. The integrals to evaluate thus are

$$\delta V^{\alpha}_{AB} = -\textstyle\int_{A(x^2=b)}^{B} \Gamma^{\alpha}_{\beta 1} V^{\beta} dx^1\,, \qquad \delta V^{\alpha}_{BC} = -\displaystyle\int_{B(x^1=a+\delta a)}^{C} \Gamma^{\alpha}_{\beta 2} V^{\beta} dx^2\,, \tag{4.4}$$

$$\delta V^{\alpha}_{CD} = -\textstyle\int_{C(x^2=b+\delta b)}^{D} \Gamma^{\alpha}_{\beta 1} V^{\beta} dx^1\,, \quad \delta V^{\alpha}_{DA} = -\displaystyle\int_{D(x^1=a)}^{A} \Gamma^{\alpha}_{\beta 2} V^{\beta} dx^2\,. \tag{4.5}$$

The integral in δV^{α}_{BC} is a function of $x^1 = a + \delta a$, with $\delta a \ll a$; therefore it can be Taylor-expanded as

$$\delta V^{\alpha}_{BC} = -\left[\int_{B(x^1=a)}^{C} \Gamma^{\alpha}_{\beta 2} V^{\beta} dx^2 + \frac{\partial}{\partial x^1}\left(\int_{B}^{C} \Gamma^{\alpha}_{\beta 2} V^{\beta} dx^2\right)\bigg|_{(x^1=a)} \delta a\right]. \tag{4.6}$$

The integral in δV^{α}_{CD} can be expanded in a similar way

$$\delta V^{\alpha}_{CD} = -\left[\int_{C(x^2=b)}^{D} \Gamma^{\alpha}_{\beta 1} V^{\beta} dx^1 + \frac{\partial}{\partial x^2}\left(\int_{C}^{D} \Gamma^{\alpha}_{\beta 1} V^{\beta} dx^1\right)\bigg|_{(x^2=b)} \delta b\right]. \tag{4.7}$$

Thus, finally

$$\begin{aligned}\delta V^{\alpha} \simeq & -\int_{A(x^2=b)}^{B} \Gamma^{\alpha}_{\beta 1} V^{\beta} dx^1 - \int_{D(x^1=a)}^{A} \Gamma^{\alpha}_{\beta 2} V^{\beta} dx^2 \\ & -\left[\int_{B(x^1=a)}^{C} \Gamma^{\alpha}_{\beta 2} V^{\beta} dx^2 + \frac{\partial}{\partial x^1}\left(\int_{B}^{C} \Gamma^{\alpha}_{\beta 2} V^{\beta} dx^2\right)\bigg|_{(x^1=a)} \delta a\right] \\ & -\left[\int_{C(x^2=b)}^{D} \Gamma^{\alpha}_{\beta 1} V^{\beta} dx^1 + \frac{\partial}{\partial x^2}\left(\int_{C}^{D} \Gamma^{\alpha}_{\beta 1} V^{\beta} dx^1\right)\bigg|_{(x^2=b)} \delta b\right].\end{aligned} \tag{4.8}$$

If we now replace the coordinates of the points A, B, C, D in the previous equation, i.e.

$$A = (a, b), \quad C = (a + \delta a, b + \delta b), \quad B = (a + \delta a, b), \quad \text{and} \quad D = (a, b + \delta b)\,, \tag{4.9}$$

only two terms survive in Eq. 4.8, and since the operators of integration and differentiation commute, the final expression of δV^{α} can be written as

$$\begin{aligned}\delta V^{\alpha} &\simeq -\delta a \int_{b}^{b+\delta b} \frac{\partial}{\partial x^1}\left(\Gamma^{\alpha}_{\beta 2} V^{\beta}\right) dx^2 + \delta b \int_{a}^{a+\delta a} \frac{\partial}{\partial x^2}\left(\Gamma^{\alpha}_{\beta 1} V^{\beta}\right) dx^1 \\ &\simeq \delta a \delta b \left[-\frac{\partial}{\partial x^1}\left(\Gamma^{\alpha}_{\beta 2} V^{\beta}\right) + \frac{\partial}{\partial x^2}\left(\Gamma^{\alpha}_{\beta 1} V^{\beta}\right)\right].\end{aligned} \tag{4.10}$$

By expanding the derivatives and making use of Eqs. 4.2 and 4.3, Eq. 4.10 can be further simplified to give

$$\begin{aligned}\delta V^\alpha &= \delta a\delta b\left[\frac{\partial\Gamma^\alpha_{\beta 1}}{\partial x^2}V^\beta+\Gamma^\alpha_{\sigma 1}\frac{\partial V^\sigma}{\partial x^2}-\frac{\partial\Gamma^\alpha_{\beta 2}}{\partial x^1}V^\beta-\Gamma^\alpha_{\sigma 2}\frac{\partial V^\sigma}{\partial x^1}\right] \\ &= \delta a\delta b\left[\frac{\partial\Gamma^\alpha_{\beta 1}}{\partial x^2}-\frac{\partial\Gamma^\alpha_{\beta 2}}{\partial x^1}-\Gamma^\alpha_{\sigma 1}\Gamma^\sigma_{\beta 2}+\Gamma^\alpha_{\sigma 2}\Gamma^\sigma_{\beta 1}\right]V^\beta\,.\end{aligned} \tag{4.11}$$

Note that δa and δb are the non-vanishing components of the displacement vectors $\vec{\delta x}_{(1)}$ and $\vec{\delta x}_{(2)}$ along the direction of the basis vectors $\vec{e}_{(1)}$ and $\vec{e}_{(2)}$, i.e.

$$\delta x^\mu_{(1)} = (0,\delta a,0,0)=\delta a\ \delta^\mu_1\ , \tag{4.12}$$

$$\delta x^\mu_{(2)} = (0,0,\delta b,0)=\delta b\ \delta^\mu_2\ ; \tag{4.13}$$

therefore Eq. 4.11 can be written as follows

$$\delta V^\alpha=\delta x^\nu_{(1)}\ \delta x^\mu_{(2)}\ \left[\frac{\partial\Gamma^\alpha_{\beta\nu}}{\partial x^\mu}-\frac{\partial\Gamma^\alpha_{\beta\mu}}{\partial x^\nu}-\Gamma^\alpha_{\sigma\nu}\Gamma^\sigma_{\beta\mu}+\Gamma^\alpha_{\sigma\mu}\Gamma^\sigma_{\beta\nu}\right]V^\beta\,. \tag{4.14}$$

If the loop is defined through different coordinate lines, the same proof leads to the same final expression, Eq. 4.14. A similar, but more involved, calculation shows that Eq. 4.14 also holds when the sides of the loop are not coordinate lines, i.e. when $\vec{\delta x}_{(1)}$ and $\vec{\delta x}_{(2)}$ are general vectors.

We shall now show that the terms in square brackets in Eq. 4.14 are the components of a tensor. According to the definition given in Chapter 2, a $\binom{1}{3}$ tensor **T** is a linear function of one one-form and three vectors which, when applied to these arguments, produces a scalar, i.e.

$$T(\tilde q,\vec A,\vec B,\vec C)=T^\alpha{}_{\beta\rho\delta}\ q_\alpha A^\beta B^\rho C^\delta\,, \tag{4.15}$$

where $T^\alpha{}_{\beta\rho\delta}$ are the components of **T**. To begin with, we note that V^α, δV^α, $\delta x^\mu_{(1)}$, and $\delta x^\mu_{(2)}$ are the components of generic vectors. Let us consider a generic one-form field with components q_α. The quantity $\delta V^\alpha q_\alpha$ is a scalar, because it is the contraction of a vector and a one-form, and

$$\delta V^\alpha q_\alpha=\left[\frac{\partial\Gamma^\alpha_{\beta\nu}}{\partial x^\mu}-\frac{\partial\Gamma^\alpha_{\beta\mu}}{\partial x^\nu}-\Gamma^\alpha_{\sigma\nu}\Gamma^\sigma_{\beta\mu}+\Gamma^\alpha_{\sigma\mu}\Gamma^\sigma_{\beta\nu}\right]q_\alpha V^\beta\delta x^\nu_{(1)}\ \delta x^\mu_{(2)}\,. \tag{4.16}$$

Therefore, the contraction of the components of the vectors $\delta x^\nu_{(1)}$, $\delta x^\mu_{(2)}$, and V^β, and of the one-form q_α with the quantity in brackets is a scalar. In addition, Eq. 4.16 is linear in $\vec V$, $\tilde q$, $\vec{\delta x}_{(1)}$, and $\vec{\delta x}_{(2)}$. Indeed, if we consider for example a displacement $\vec{\delta x}_{(1a)}+\vec{\delta x}_{(1b)}$ along $\vec e_{(1)}$, it is immediate to check that

$$\delta V^\alpha q_\alpha=\delta x^\nu_{(1a)}\delta x^\mu_{(2)}\,[...]\,q_\alpha V^\beta+\delta x^\nu_{(1b)}\delta x^\mu_{(2)}\,[...]\,q_\alpha V^\beta\,, \tag{4.17}$$

and similarly for the other quantities. Thus, we can conclude that the terms in square brackets in Eq. 4.16 are the components of a $\binom{1}{3}$ tensor, the **Riemann tensor**

$$R^\alpha{}_{\beta\mu\nu}\equiv\Gamma^\alpha_{\beta\nu,\mu}-\Gamma^\alpha_{\beta\mu,\nu}-\Gamma^\alpha_{\sigma\nu}\Gamma^\sigma_{\beta\mu}+\Gamma^\alpha_{\sigma\mu}\Gamma^\sigma_{\beta\nu}\,. \tag{4.18}$$

Note that the Riemann tensor is antisymmetric in ν and μ; indeed, if we interchange $\vec{\delta x}_{(1)}$ and $\vec{\delta x}_{(2)}$ in Eq. 4.14, δV^α changes sign, because the loop goes in the opposite direction.

This shows that the sign of Eq. 4.18 can be chosen arbitrarily, and for this reason the definitions of the Riemann tensor given in textbooks may differ for an overall sign.

The Riemann tensor depends on the affine connection and on its first derivatives, i.e. on the first and second derivatives of the metric tensor. If we choose a LIF the $\Gamma^{\alpha}_{\beta\mu}$'s vanish, but their derivatives do not; the Riemann tensor is generically nonzero in a LIF.

We shall call a spacetime *flat* if the parallel transport of any vector field around any loop vanishes, otherwise we shall call it *curved*. Therefore, in a flat spacetime the Riemann tensor vanishes everywhere and (since it is a tensor) in any frame; in a curved spacetime, instead, the Riemann tensor is non-zero at some point of the manifold, and this is also true in any frame. It can be shown (but the rigorous derivation goes beyond the scope of this book) that a global, Minkowskian coordinate system can be defined if and only if the Riemann tensor vanishes on the entire spacetime[1].

In conclusion, *if the spacetime is flat the Riemann tensor is zero, if it is curved the Riemann tensor is non-zero, and this is true in any coordinate frame.* This justifies why the Riemann tensor is also called the **curvature tensor**.

Furthermore, Eq. 4.14 shows that if $R^{\alpha}{}_{\beta\gamma\delta} = 0$, i.e. the spacetime is flat, a vector $\vec{V}$ parallely transported along *any* closed loop does not change, i.e. $\delta V^{\alpha} = 0$. Conversely, if the Riemann tensor does not vanish, i.e. the spacetime is curved, there will exist at least one loop such that $\delta V^{\alpha} \neq 0$.

It should be stressed that

- The Riemann tensor is linear in the second derivatives of $g_{\mu\nu}$, and non-linear in the first derivatives. Note that it has the same differential structure of the Gaussian curvature introduced in Chapter 1, Eq. 1.12.
- In a LIF $\Gamma^{\alpha}_{\nu\sigma} = 0$, therefore the non-linear part of the Riemann tensor vanishes as well.

By contracting the Riemann tensor with the metric we can construct a $\binom{0}{2}$ tensor, called the **Ricci tensor**

$$R_{\mu\nu} = g^{\alpha\beta} R_{\alpha\mu\beta\nu} = R^{\alpha}{}_{\mu\alpha\nu} \,. \tag{4.19}$$

By further contracting the Ricci tensor with the metric, we can also define the **scalar (or Ricci) curvature**

$$R = g^{\alpha\beta} R_{\alpha\beta} = R^{\alpha}{}_{\alpha} \,. \tag{4.20}$$

As the Riemann tensor, both the Ricci tensor and the Ricci scalar are linear in the second derivatives of $g_{\mu\nu}$, and non-linear in the first derivatives.

In a LIF, the components of the Riemann tensor have a very simple form since the non-linear part of the tensor vanishes, i.e.

$$R^{\mu}{}_{\nu\alpha\beta} = \Gamma^{\mu}_{\nu\beta,\alpha} - \Gamma^{\mu}_{\nu\alpha,\beta} \,, \tag{4.21}$$

and by replacing the expression of the Christoffel symbols given in Eq. 3.68, we find

$$R^{\alpha}{}_{\beta\mu\nu} = \frac{1}{2} g^{\alpha\sigma} \left[g_{\sigma\nu,\beta\mu} - g_{\sigma\mu,\beta\nu} + g_{\beta\mu,\sigma\nu} - g_{\beta\nu,\sigma\mu} \right] \,, \tag{4.22}$$

or, lowering the index α,

$$R_{\alpha\beta\mu\nu} = g_{\alpha\lambda} R^{\lambda}{}_{\beta\mu\nu} = \frac{1}{2} \left[g_{\alpha\nu,\beta\mu} - g_{\alpha\mu,\beta\nu} + g_{\beta\mu,\alpha\nu} - g_{\beta\nu,\alpha\mu} \right] \,. \tag{4.23}$$

---

[1] This can be understood by noting that in a LIF $g_{\mu\nu} = \eta_{\mu\nu} + O(x^2)$, and the $O(x^2)$ terms depend on the second derivatives of the metric; it can be shown that the $O(x^2)$ terms can be expressed in terms of the Riemann tensor components. If these terms vanish everywhere, the LIF can be extended to the entire spacetime.

Consequently, in a LIF the components of the Ricci tensor are

$$R_{\beta\nu} = R^{\alpha}{}_{\beta\alpha\nu} = \frac{1}{2} g^{\alpha\sigma} \left[g_{\sigma\nu,\beta\alpha} - g_{\sigma\alpha,\beta\nu} + g_{\beta\alpha,\sigma\nu} - g_{\beta\nu,\sigma\alpha} \right] . \tag{4.24}$$

4.2 SYMMETRIES OF THE RIEMANN TENSOR

Using Eq. 4.23 it is easy to show that the curvature tensor is antisymmetric in the first and in the second pair of indices, i.e.

$$R_{\alpha\beta\mu\nu} = -R_{\beta\alpha\mu\nu} \,, \quad \text{and} \quad R_{\alpha\beta\mu\nu} = -R_{\alpha\beta\nu\mu} \,, \tag{4.25}$$

whereas it is symmetric under exchange of the two pairs of indices

$$R_{\alpha\beta\mu\nu} = R_{\mu\nu\alpha\beta} \,. \tag{4.26}$$

In addition it satisfies the *Ricci identities*

$$R_{\alpha\beta\mu\nu} + R_{\alpha\nu\beta\mu} + R_{\alpha\mu\nu\beta} = 0 \,. \tag{4.27}$$

Since $R_{\alpha\beta\mu\nu}$ is a tensor, these symmetry properties, which can easily be derived in a LIF using Eq. 4.23, hold in any reference frame. The symmetries of the Riemann tensor reduce the number of its independent components from $4^4 = 256$ to only 20.

As a consequence of Eq. 4.26, the Ricci tensor $R_{\mu\nu} = g^{\alpha\beta} R_{\mu\alpha\nu\beta}$ is symmetric under exchange of its indices:

$$R_{\mu\nu} = R_{\nu\mu} \,, \tag{4.28}$$

and therefore it has only 10 independent components.

4.3 THE RIEMANN TENSOR GIVES THE COMMUTATOR OF COVARIANT DERIVATIVES

Let us consider the second covariant derivatives of a vector field $\vec{V}$

$$\nabla_\alpha \nabla_\beta V^\mu = \nabla_\alpha (V^\mu{}_{;\beta}) = (V^\mu{}_{;\beta})_{,\alpha} + \Gamma^\mu_{\sigma\alpha} V^\sigma{}_{;\beta} - \Gamma^\sigma_{\beta\alpha} V^\mu{}_{;\sigma} \,. \tag{4.29}$$

In a LIF $\Gamma^\mu_{\sigma\alpha} = 0$, and Eq. 4.29 becomes

$$\nabla_\alpha \nabla_\beta V^\mu = (V^\mu{}_{;\beta})_{,\alpha} = V^\mu{}_{,\beta,\alpha} + \Gamma^\mu_{\nu\beta,\alpha} V^\nu \,. \tag{4.30}$$

By interchanging α and β

$$\nabla_\beta \nabla_\alpha V^\mu = (V^\mu{}_{;\alpha})_{,\beta} = V^\mu{}_{,\alpha,\beta} + \Gamma^\mu_{\nu\alpha,\beta} V^\nu \,. \tag{4.31}$$

The commutator of the covariant derivatives is then

$$[\nabla_\alpha, \nabla_\beta] \, V^\mu \equiv \nabla_\alpha \nabla_\beta V^\mu - \nabla_\beta \nabla_\alpha V^\mu = \left(\Gamma^\mu_{\nu\beta,\alpha} - \Gamma^\mu_{\nu\alpha,\beta} \right) V^\nu \,. \tag{4.32}$$

The term in round brackets in this equation coincides with the expression of the Riemann tensor in the LIF (see Eq. 4.21), therefore

$$[\nabla_\alpha, \nabla_\beta] \, V^\mu = R^\mu{}_{\nu\alpha\beta} V^\nu \,. \tag{4.33}$$

This is a tensor equation and since it is valid in a given reference frame, it will be valid in *any* frame. Eq. 4.33 shows that *in curved spacetime the covariant derivatives do not commute* and therefore the order in which they appear is important.

4.4 THE BIANCHI IDENTITIES

We shall now derive the *Bianchi identities*, a set of differential equations satisfied by the Riemann tensor. Since in a LIF the expression of the tensor is much simpler, we shall use Eq. 4.23 to prove the identities in this frame, and then generalize them to any frame.

Let us differentiate Eq. 4.23 with respect to x^λ

$$R_{\alpha\beta\mu\nu,\lambda} = \frac{1}{2}\left[g_{\alpha\nu,\beta\mu\lambda} - g_{\alpha\mu,\beta\nu\lambda} + g_{\beta\mu,\alpha\nu\lambda} - g_{\beta\nu,\alpha\mu\lambda}\right] . \tag{4.34}$$

Using Eq. 4.34 and the symmetry of $g_{\alpha\beta}$, it is easy to show that

$$R_{\alpha\beta\mu\nu,\lambda} + R_{\alpha\beta\lambda\mu,\nu} + R_{\alpha\beta\nu\lambda,\mu} = 0 . \tag{4.35}$$

In a LIF ordinary and covariant derivatives coincide, therefore Eq. 4.35 can also be written as

$$R_{\alpha\beta\mu\nu;\lambda} + R_{\alpha\beta\lambda\mu;\nu} + R_{\alpha\beta\nu\lambda;\mu} = 0 . \tag{4.36}$$

These equations have been derived in a LIF, and since they are tensor equations, they will be valid in any frame.

Eqs. 4.36 are the Bianchi identities which, as we shall later see, play an important role in the derivation of Einstein's equations.

4.5 THE EQUATION OF GEODESIC DEVIATION

If the spacetime is curved, the Equivalence Principle establishes that we can always choose a LIF where the affine connection vanishes. In this frame a free particle (i.e. one subjected to no other force than gravity) follows the geodesic equation with zero acceleration.

Conversely, if the spacetime is flat we can always choose a coordinate frame in which the affine connection does not vanish as, for instance, in Minkowski's spacetime when spherical coordinates are chosen. In this case a free particle would follow a geodesic with non-zero acceleration, and we would not be able to establish whether this acceleration is due to a true gravitational field or to the fact that the reference frame is not inertial. In other words, the local curvature of spacetime cannot be measured by observing the motion of a single particle.

We shall now show that the local curvature of spacetime can be measured by comparing the motion of two close particles, i.e. by comparing the behaviour of close geodesics.

Let us consider a two-parameter family of geodesics $x^\mu(\tau, p)$, where τ is the affine parameter and p labels different curves of the family. Along each geodesic $p = const$. We also consider the family of non-geodesic curves which join points of the family $x^\mu(\tau, p)$ having the same τ; these curves are parametrized by p and are indicated with $\tau = const$ in Fig. 4.2.

Two particles move along two infinitely close geodesics $x^\mu(\tau, p)$ and $x^\mu(\tau, p) + \delta x^\mu = x^\mu(\tau, p + \delta p)$. Here $\vec{\delta x}$ is the infinitesimal displacement vector between points having the same τ on the two geodesics. Let $\vec{t}$ be the tangent vector to the geodesics, and $\vec{b}$ the tangent vector to the $p = const$ curves, as indicated in Fig. 4.2, i.e.

$$t^\alpha = \frac{\partial x^\alpha}{\partial \tau} \qquad \text{and} \qquad b^\alpha = \frac{\partial x^\alpha}{\partial p} . \tag{4.37}$$

The displacement vector is $\delta x^\alpha = b^\alpha \delta p$. For simplicity, we choose the parameter p such that the tangent vector b^α and the displacement vector δx^α coincide. With this definition, $\delta x^\alpha = \frac{\partial x^\alpha}{\partial p}$, and

$$\frac{\partial t^\alpha}{\partial p} = \frac{\partial \delta x^\alpha}{\partial \tau} . \tag{4.38}$$

3)*+,)))- !) & /! ! " 4 5 *
% " 4 ! + ! 2 5 /"
' &$!/ $

CHAPTER 3

Curvature

3.1 ■ OVERVIEW

We all know what curvature means, at least informally, and in the first two chapters of this book we have felt free to refer on occasion to the concept of curvature without giving it a careful definition. Clearly curvature depends somehow on the metric, which defines the geometry of our manifold; but it is not immediately clear how we should attribute curvature to any given metric (since, as we have seen, even the metric of a flat space can look arbitrarily complicated in a sufficiently extravagant coordinate system). As is often the case in mathematics, we require quite a bit of care to formalize our intuition about a concept into a usable mathematical structure; formalizing what we think of as "curvature" is the subject of this chapter.

The techniques we are about to develop are absolutely crucial to the subject; it is safe to say that there is a higher density of useful formulas per page in this chapter than in any of the others. Let's quickly summarize the most important ones, to provide a roadmap for the formalism to come.

All the ways in which curvature manifests itself rely on something called a "connection," which gives us a way of relating vectors in the tangent spaces of nearby points. There is a unique connection that we can construct from the metric, and it is encapsulated in an object called the *Christoffel symbol*, given by

$$\Gamma^{\lambda}_{\mu\nu} = \tfrac{1}{2} g^{\lambda\sigma}(\partial_\mu g_{\nu\sigma} + \partial_\nu g_{\sigma\mu} - \partial_\sigma g_{\mu\nu}). \tag{3.1}$$

The notation makes $\Gamma^{\lambda}_{\mu\nu}$ look like a tensor, but in fact it is not; this is why we call it an "object" or "symbol." The fundamental use of a connection is to take a *covariant derivative* ∇_μ (a generalization of the partial derivative); the covariant derivative of a vector field V^ν is given by

$$\nabla_\mu V^\nu = \partial_\mu V^\nu + \Gamma^{\nu}_{\mu\sigma} V^\sigma, \tag{3.2}$$

and covariant derivatives of other sorts of tensors are given by similar expressions. The connection also appears in the definition of *geodesics* (a generalization of the notion of a straight line). A parameterized curve $x^\mu(\lambda)$ is a geodesic if it obeys

$$\frac{d^2 x^\mu}{d\lambda^2} + \Gamma^{\mu}_{\rho\sigma} \frac{dx^\rho}{d\lambda} \frac{dx^\sigma}{d\lambda} = 0, \tag{3.3}$$

known as the geodesic equation.

Finally, the technical expression of curvature is contained in the Riemann tensor, a $(1, 3)$ tensor obtained from the connection by

$$R^{\rho}{}_{\sigma\mu\nu} = \partial_{\mu}\Gamma^{\rho}_{\nu\sigma} - \partial_{\nu}\Gamma^{\rho}_{\mu\sigma} + \Gamma^{\rho}_{\mu\lambda}\Gamma^{\lambda}_{\nu\sigma} - \Gamma^{\rho}_{\nu\lambda}\Gamma^{\lambda}_{\mu\sigma}. \tag{3.4}$$

Everything we want to know about the curvature of a manifold is given to us by the Riemann tensor; it will vanish if and only if the metric is perfectly flat. Einstein's equation of general relativity relates certain components of this tensor to the energy-momentum tensor.

These four equations are all of primary importance in the study of curved manifolds. We will now see how they arise from a careful consideration of how familiar notions of geometry in flat space adapt to this more general context.

3.2 ■ COVARIANT DERIVATIVES

In our discussion of manifolds, it became clear that there were various notions we could talk about as soon as the manifold was defined: we could define functions, take their derivatives, consider parameterized paths, set up tensors, and so on. Other concepts, such as the volume of a region or the length of a path, required some additional piece of structure, namely the introduction of a metric. It would be natural to think of the notion of curvature as something that depends exclusively on the metric. In a more careful treatment, however, we find that curvature depends on a connection, and connections may or may not depend on the metric. Nevertheless, we will also show how the existence of a metric implies a certain unique connection, whose curvature may be thought of as that of the metric. This is the connection used in general relativity, so in this particular context it is legitimate to think of curvature as characterizing the metric, without introducing any additional structures.

The connection becomes necessary when we attempt to address the problem of the partial derivative not being a good tensor operator. What we would like is a covariant derivative, that is, an operator that reduces to the partial derivative in flat space with inertial coordinates, but transforms as a tensor on an arbitrary manifold. It is conventional to spend a certain amount of time motivating the introduction of a covariant derivative, but in fact the need is obvious; equations such as $\partial_{\mu}T^{\mu\nu} = 0$ must be generalized to curved space somehow. So let's agree that a covariant derivative would be a good thing to have, and go about setting it up.

In flat space in inertial coordinates, the partial derivative operator ∂_{μ} is a map from (k, l) tensor fields to $(k, l + 1)$ tensor fields, which acts linearly on its arguments and obeys the Leibniz rule on tensor products. All of this continues to be true in the more general situation we would now like to consider, but the map provided by the partial derivative depends on the coordinate system used. We would therefore like to define a **covariant derivative** operator ∇ to perform the functions of the partial derivative, but in a way independent of coordinates. Rather than simply postulating the answer (which would be perfectly acceptable), let's motivate it by thinking carefully about what properties a covariant generalization of the partial derivative *should* have—mathematical structures are, after all, invented by human beings, not found lying on sidewalks. We begin by requiring that ∇ be a map from (k, l) tensor fields, to $(k, l+1)$ tensor fields which has these two properties:

1. linearity: $\nabla(T + S) = \nabla T + \nabla S$;
2. Leibniz (product) rule: $\nabla(T \otimes S) = (\nabla T) \otimes S + T \otimes (\nabla S)$.

If ∇ is going to obey the Leibniz rule, it can always be written as the partial derivative plus some linear transformation. That is, to take the covariant derivative we first take the partial derivative, and then apply a correction to make the result covariant. [We aren't going to prove this reasonable-sounding statement; see Wald (1984) if you are interested.] Let's consider what this means for the covariant derivative of a vector V^ν. It means that, for each direction μ, the covariant derivative ∇_μ will be given by the partial derivative ∂_μ plus a correction specified by a set of n matrices $(\Gamma_\mu)^\rho{}_\sigma$ (one $n \times n$ matrix, where n is the dimensionality of the manifold, for each μ). In fact the parentheses are usually dropped and we write these matrices, known as the **connection coefficients**, with haphazard index placement as $\Gamma^\rho_{\mu\sigma}$. We therefore have

$$\boxed{\nabla_\mu V^\nu = \partial_\mu V^\nu + \Gamma^\nu_{\mu\lambda} V^\lambda.} \tag{3.5}$$

Notice that in the second term the index originally on V has moved to the Γ, and a new index is summed over. If this is the expression for the covariant derivative of a vector in terms of the partial derivative, we should be able to determine the transformation properties of $\Gamma^\nu_{\mu\lambda}$ by demanding that the left-hand side be a (1, 1) tensor. That is, we want the transformation law to be

$$\nabla_{\mu'} V^{\nu'} = \frac{\partial x^\mu}{\partial x^{\mu'}} \frac{\partial x^{\nu'}}{\partial x^\nu} \nabla_\mu V^\nu. \tag{3.6}$$

Let's look at the left side first; we can expand it using (3.5) and then transform the parts that we understand (which is everything except $\Gamma^{\nu'}_{\mu'\lambda'}$):

$$\begin{aligned}\nabla_{\mu'} V^{\nu'} &= \partial_{\mu'} V^{\nu'} + \Gamma^{\nu'}_{\mu'\lambda'} V^{\lambda'} \\ &= \frac{\partial x^\mu}{\partial x^{\mu'}} \frac{\partial x^{\nu'}}{\partial x^\nu} \partial_\mu V^\nu + \frac{\partial x^\mu}{\partial x^{\mu'}} V^\nu \frac{\partial}{\partial x^\mu} \frac{\partial x^{\nu'}}{\partial x^\nu} + \Gamma^{\nu'}_{\mu'\lambda'} \frac{\partial x^{\lambda'}}{\partial x^\lambda} V^\lambda. \end{aligned} \tag{3.7}$$

On the right-hand side we can also expand $\nabla_\mu V^\nu$:

$$\frac{\partial x^\mu}{\partial x^{\mu'}} \frac{\partial x^{\nu'}}{\partial x^\nu} \nabla_\mu V^\nu = \frac{\partial x^\mu}{\partial x^{\mu'}} \frac{\partial x^{\nu'}}{\partial x^\nu} \partial_\mu V^\nu + \frac{\partial x^\mu}{\partial x^{\mu'}} \frac{\partial x^{\nu'}}{\partial x^\nu} \Gamma^\nu_{\mu\lambda} V^\lambda. \tag{3.8}$$

These last two expressions are to be equated; the first terms in each are identical and therefore cancel, so we have

$$\Gamma^{\nu'}_{\mu'\lambda'}\frac{\partial x^{\lambda'}}{\partial x^{\lambda}}V^{\lambda}+\frac{\partial x^{\mu}}{\partial x^{\mu'}}V^{\lambda}\frac{\partial}{\partial x^{\mu}}\frac{\partial x^{\nu'}}{\partial x^{\lambda}}=\frac{\partial x^{\mu}}{\partial x^{\mu'}}\frac{\partial x^{\nu'}}{\partial x^{\nu}}\Gamma^{\nu}_{\mu\lambda}V^{\lambda}, \tag{3.9}$$

where we have changed a dummy index from ν to λ. This equation must be true for any vector V^{λ}, so we can eliminate that on both sides. Then the connection coefficients in the primed coordinates may be isolated by multiplying by $\partial x^{\lambda}/\partial x^{\sigma'}$ and relabeling $\sigma' \to \lambda'$. The result is

$$\Gamma^{\nu'}_{\mu'\lambda'}=\frac{\partial x^{\mu}}{\partial x^{\mu'}}\frac{\partial x^{\lambda}}{\partial x^{\lambda'}}\frac{\partial x^{\nu'}}{\partial x^{\nu}}\Gamma^{\nu}_{\mu\lambda}-\frac{\partial x^{\mu}}{\partial x^{\mu'}}\frac{\partial x^{\lambda}}{\partial x^{\lambda'}}\frac{\partial^2 x^{\nu'}}{\partial x^{\mu}\partial x^{\lambda}}. \tag{3.10}$$

This is not, of course, the tensor transformation law; the second term on the right spoils it. That's okay, because *the connection coefficients are not the components of a tensor*. They are purposefully constructed to be nontensorial, but in such a way that the combination (3.5) transforms as a tensor—the extra terms in the transformation of the partials and the Γ's exactly cancel. This is why we are not so careful about index placement on the connection coefficients; they are not a tensor, and therefore you should try not to raise and lower their indices.

What about the covariant derivatives of other sorts of tensors? By similar reasoning to that used for vectors, the covariant derivative of a one-form can also be expressed as a partial derivative plus some linear transformation. But there is no reason as yet that the matrices representing this transformation should be related to the coefficients $\Gamma^{\nu}_{\mu\lambda}$. In general we could write something like

$$\nabla_{\mu}\omega_{\nu}=\partial_{\mu}\omega_{\nu}+\widetilde{\Gamma}^{\lambda}_{\mu\nu}\omega_{\lambda}, \tag{3.11}$$

where $\widetilde{\Gamma}^{\lambda}_{\mu\nu}$ is a new set of matrices for each μ. Pay attention to where all of the various indices go. It is straightforward to derive that the transformation properties of $\widetilde{\Gamma}$ must be the same as those of Γ, since otherwise $\nabla_{\mu}\omega_{\nu}$ wouldn't transform as a tensor, but otherwise no relationship has been established. To do so, we need to introduce two new properties that we would like our covariant derivative to have, in addition to the two above:

3. commutes with contractions: $\nabla_{\mu}(T^{\lambda}{}_{\lambda\rho})=(\nabla T)_{\mu}{}^{\lambda}{}_{\lambda\rho}$,
4. reduces to the partial derivative on scalars: $\nabla_{\mu}\phi=\partial_{\mu}\phi$.

There is no way to "derive" these properties; we are simply demanding that they be true as part of the definition of a covariant derivative. Note that property 3 is equivalent to saying that the Kronecker delta (the identity map) is covariantly constant, $\nabla_{\mu}\delta^{\lambda}_{\sigma}=0$; this is certainly a reasonable thing to ask.

Let's see what these new properties imply. Given some one-form field ω_{μ} and vector field V^{μ}, we can take the covariant derivative of the scalar defined by $\omega_{\lambda}V^{\lambda}$ to get

$$\begin{aligned}\nabla_\mu(\omega_\lambda V^\lambda) &= (\nabla_\mu\omega_\lambda)V^\lambda + \omega_\lambda(\nabla_\mu V^\lambda)\\ &= (\partial_\mu\omega_\lambda)V^\lambda + \widetilde{\Gamma}^\sigma_{\mu\lambda}\omega_\sigma V^\lambda + \omega_\lambda(\partial_\mu V^\lambda) + \omega_\lambda\Gamma^\lambda_{\mu\rho}V^\rho.\end{aligned} \tag{3.12}$$

But since $\omega_\lambda V^\lambda$ is a scalar, this must also be given by the partial derivative:

$$\begin{aligned}\nabla_\mu(\omega_\lambda V^\lambda) &= \partial_\mu(\omega_\lambda V^\lambda)\\ &= (\partial_\mu\omega_\lambda)V^\lambda + \omega_\lambda(\partial_\mu V^\lambda).\end{aligned} \tag{3.13}$$

This can only be true if the terms in (3.12) with connection coefficients cancel each other; that is, rearranging dummy indices, we must have

$$0 = \widetilde{\Gamma}^\sigma_{\mu\lambda}\omega_\sigma V^\lambda + \Gamma^\sigma_{\mu\lambda}\omega_\sigma V^\lambda. \tag{3.14}$$

But both ω_σ and V^λ are completely arbitrary, so

$$\widetilde{\Gamma}^\sigma_{\mu\lambda} = -\Gamma^\sigma_{\mu\lambda}. \tag{3.15}$$

The two extra conditions we have imposed therefore allow us to express the covariant derivative of a one-form using the same connection coefficients as were used for the vector, but now with a minus sign (and indices matched up somewhat differently):

$$\boxed{\nabla_\mu\omega_\nu = \partial_\mu\omega_\nu - \Gamma^\lambda_{\mu\nu}\omega_\lambda.} \tag{3.16}$$

It should come as no surprise that the connection coefficients encode all the information necessary to take the covariant derivative of a tensor of arbitrary rank. The formula is quite straightforward; for each upper index you introduce a term with a single $+\Gamma$, and for each lower index a term with a single $-\Gamma$:

$$\begin{aligned}\nabla_\sigma T^{\mu_1\mu_2\cdots\mu_k}{}_{\nu_1\nu_2\cdots\nu_l} &= \partial_\sigma T^{\mu_1\mu_2\cdots\mu_k}{}_{\nu_1\nu_2\cdots\nu_l}\\ &\quad + \Gamma^{\mu_1}_{\sigma\lambda}T^{\lambda\mu_2\cdots\mu_k}{}_{\nu_1\nu_2\cdots\nu_l} + \Gamma^{\mu_2}_{\sigma\lambda}T^{\mu_1\lambda\cdots\mu_k}{}_{\nu_1\nu_2\cdots\nu_l} + \cdots\\ &\quad - \Gamma^\lambda_{\sigma\nu_1}T^{\mu_1\mu_2\cdots\mu_k}{}_{\lambda\nu_2\cdots\nu_l} - \Gamma^\lambda_{\sigma\nu_2}T^{\mu_1\mu_2\cdots\mu_k}{}_{\nu_1\lambda\cdots\nu_l} - \cdots.\end{aligned} \tag{3.17}$$

This is the general expression for the covariant derivative. You can check it yourself; it comes from the set of axioms we have established, and the usual requirements that tensors of various sorts be coordinate-independent entities. Sometimes an alternative notation is used; just as commas are used for partial derivatives, semicolons are used for covariant ones:

$$\nabla_\sigma T^{\mu_1\mu_2\cdots\mu_k}{}_{\nu_1\nu_2\cdots\nu_l} \equiv T^{\mu_1\mu_2\cdots\mu_k}{}_{\nu_1\nu_2\cdots\nu_l;\sigma}. \tag{3.18}$$

Once again, in this book we will stick to "∇_σ."

To define a covariant derivative, then, we need to put a connection on our manifold, which is specified in some coordinate system by a set of coefficients $\Gamma^{\lambda}_{\mu\nu}$ ($n^3 = 64$ independent components in $n = 4$ dimensions) that transform according to (3.10). (The name *connection* comes from the fact that it is used to transport vectors from one tangent space to another, as we will soon see; it is sometimes used to refer to the operator ∇, and sometimes to the coefficients $\Gamma^{\lambda}_{\mu\nu}$.) Evidently, we could define a large number of connections on any manifold, and each of them implies a distinct notion of covariant differentiation. In general relativity this freedom is not a big concern, because it turns out that every metric defines a unique connection, which is the one used in GR. Let's see how that works.

The first thing to notice is that the difference of two connections is a tensor. Imagine we have defined two different kinds of covariant derivative, ∇_μ and $\widehat{\nabla}_\mu$, with associated connection coefficients $\Gamma^{\lambda}_{\mu\nu}$ and $\widehat{\Gamma}^{\lambda}_{\mu\nu}$. Then the difference

$$S^{\lambda}{}_{\mu\nu} = \Gamma^{\lambda}_{\mu\nu} - \widehat{\Gamma}^{\lambda}_{\mu\nu} \tag{3.19}$$

is a $(1, 2)$ tensor. (Notice that we had to choose a convention for index placement.) We could show this by brute force, plugging in the transformation laws for the connection coefficients, but let's be a little more slick. Given an arbitrary vector field V^λ, we know that both $\nabla_\mu V^\lambda$ and $\widehat{\nabla}_\mu V^\lambda$ are tensors, so their difference must also be. This difference is simply

$$\begin{aligned}\nabla_\mu V^\lambda - \widehat{\nabla}_\mu V^\lambda &= \partial_\mu V^\lambda + \Gamma^{\lambda}_{\mu\nu} V^\lambda - \partial_\mu V^\lambda - \widehat{\Gamma}^{\lambda}_{\mu\nu} V^\lambda \\ &= S^{\lambda}{}_{\mu\nu} V^\nu.\end{aligned} \tag{3.20}$$

Since V^λ was arbitrary, and the left-hand side is a tensor, $S^{\lambda}{}_{\mu\nu}$ must be a tensor. As a trivial consequence, we learn that any set of connection coefficients can be expressed as some fiducial connection plus a tensorial correction,

$$\Gamma^{\lambda}_{\mu\nu} = \widehat{\Gamma}^{\lambda}_{\mu\nu} + S^{\lambda}{}_{\mu\nu}. \tag{3.21}$$

Next notice that, given a connection specified by $\Gamma^{\lambda}_{\mu\nu}$, we can immediately form another connection simply by permuting the lower indices. That is, the set of coefficients $\Gamma^{\lambda}_{\nu\mu}$ will also transform according to (3.10) (since the partial derivatives appearing in the last term can be commuted), so they determine a distinct connection. There is thus a tensor we can associate with any given connection, known as the **torsion tensor**, defined by

$$T^{\lambda}{}_{\mu\nu} = \Gamma^{\lambda}_{\mu\nu} - \Gamma^{\lambda}_{\nu\mu} = 2\Gamma^{\lambda}_{[\mu\nu]}. \tag{3.22}$$

It is clear that the torsion is antisymmetric in its lower indices, and a connection that is symmetric in its lower indices is known as "torsion-free."

We can now define a unique connection on a manifold with a metric $g_{\mu\nu}$ by introducing two additional properties:

- torsion-free: $\Gamma^{\lambda}_{\mu\nu} = \Gamma^{\lambda}_{(\mu\nu)}$.
- metric compatibility: $\nabla_\rho g_{\mu\nu} = 0$.

A connection is **metric compatible** if the covariant derivative of the metric with respect to that connection is everywhere zero. This implies a couple of nice properties. First, it's easy to show that both the Levi–Civita tensor and the inverse metric also have zero covariant derivative,

$$\begin{aligned} \nabla_\lambda \epsilon_{\mu\nu\rho\sigma} &= 0 \\ \nabla_\rho g^{\mu\nu} &= 0. \end{aligned} \tag{3.23}$$

Second, a metric-compatible covariant derivative commutes with raising and lowering of indices. Thus, for some vector field V^λ,

$$g_{\mu\lambda}\nabla_\rho V^\lambda = \nabla_\rho(g_{\mu\lambda}V^\lambda) = \nabla_\rho V_\mu. \tag{3.24}$$

With non-metric-compatible connections we would have to be very careful about index placement when taking a covariant derivative.

Our claim is therefore that there is exactly one torsion-free connection on a given manifold that is compatible with some given metric on that manifold. We do not want to make these two requirements part of the definition of a covariant derivative; they simply single out one of the many possible ones.

We can demonstrate both existence and uniqueness by deriving a manifestly unique expression for the connection coefficients in terms of the metric. To accomplish this, we expand out the equation of metric compatibility for three different permutations of the indices:

$$\begin{aligned} \nabla_\rho g_{\mu\nu} &= \partial_\rho g_{\mu\nu} - \Gamma^{\lambda}_{\rho\mu} g_{\lambda\nu} - \Gamma^{\lambda}_{\rho\nu} g_{\mu\lambda} = 0 \\ \nabla_\mu g_{\nu\rho} &= \partial_\mu g_{\nu\rho} - \Gamma^{\lambda}_{\mu\nu} g_{\lambda\rho} - \Gamma^{\lambda}_{\mu\rho} g_{\nu\lambda} = 0 \\ \nabla_\nu g_{\rho\mu} &= \partial_\nu g_{\rho\mu} - \Gamma^{\lambda}_{\nu\rho} g_{\lambda\mu} - \Gamma^{\lambda}_{\nu\mu} g_{\rho\lambda} = 0. \end{aligned} \tag{3.25}$$

We subtract the second and third of these from the first, and use the symmetry of the connection to obtain

$$\partial_\rho g_{\mu\nu} - \partial_\mu g_{\nu\rho} - \partial_\nu g_{\rho\mu} + 2\Gamma^{\lambda}_{\mu\nu} g_{\lambda\rho} = 0. \tag{3.26}$$

It is straightforward to solve this for the connection by multiplying by $g^{\sigma\rho}$. The result is

$$\boxed{\Gamma^{\sigma}_{\mu\nu} = \tfrac{1}{2} g^{\sigma\rho}(\partial_\mu g_{\nu\rho} + \partial_\nu g_{\rho\mu} - \partial_\rho g_{\mu\nu}).} \tag{3.27}$$

This formula is one of the most important in this subject; commit it to memory. Of course, we have only proved that if a metric-compatible and torsion-free connection exists, it must be of the form (3.27); you can check for yourself that the right-hand side of (3.27) transforms like a connection.

This connection we have derived from the metric is the one on which conventional general relativity is based. It is known by different names: sometimes the **Christoffel** connection, sometimes the **Levi–Civita** connection, sometimes the **Riemannian** connection. The associated connection coefficients are sometimes called **Christoffel symbols** and written as $\left\{ {\sigma \atop \mu\nu} \right\}$; we will sometimes call them Christoffel symbols, but we won't use the funny notation. The study of manifolds with metrics and their associated connections is called Riemannian geometry, or sometimes pseudo-Riemannian when the metric has a Lorentzian signature.

Before putting our covariant derivatives to work, we should mention some miscellaneous properties. First, note that in ordinary flat space there is an implicit connection we use all the time—the Christoffel connection constructed from the flat metric. The coefficients of the Christoffel connection in flat space vanish in Cartesian coordinates, but not in curvilinear coordinate systems. Consider for example the plane in polar coordinates, with metric

$$ds^2 = \mathrm{d}r^2 + r^2 \mathrm{d}\theta^2. \tag{3.28}$$

The nonzero components of the inverse metric are readily found to be $g^{rr} = 1$ and $g^{\theta\theta} = r^{-2}$. Notice that we use r and θ as indices in an obvious notation. We can compute a typical connection coefficient:

$$\begin{aligned}
\Gamma^r_{rr} &= \tfrac{1}{2} g^{r\rho}(\partial_r g_{r\rho} + \partial_r g_{\rho r} - \partial_\rho g_{rr}) \\
&= \tfrac{1}{2} g^{rr}(\partial_r g_{rr} + \partial_r g_{rr} - \partial_r g_{rr}) \\
&\quad + \tfrac{1}{2} g^{r\theta}(\partial_r g_{r\theta} + \partial_r g_{\theta r} - \partial_\theta g_{rr}) \\
&= \tfrac{1}{2}(1)(0+0-0) + \tfrac{1}{2}(0)(0+0-0) \\
&= 0.
\end{aligned} \tag{3.29}$$

Sadly, it vanishes. But not all of them do:

$$\begin{aligned}
\Gamma^r_{\theta\theta} &= \tfrac{1}{2} g^{r\rho}(\partial_\theta g_{\theta\rho} + \partial_\theta g_{\rho\theta} - \partial_\rho g_{\theta\theta}) \\
&= \tfrac{1}{2} g^{rr}(\partial_\theta g_{\theta r} + \partial_\theta g_{r\theta} - \partial_r g_{\theta\theta}) \\
&= \tfrac{1}{2}(1)(0+0-2r) \\
&= -r.
\end{aligned} \tag{3.30}$$

Continuing to turn the crank, we eventually find

$$\begin{aligned}
\Gamma^r_{\theta r} &= \Gamma^r_{r\theta} = 0 \\
\Gamma^\theta_{rr} &= 0 \\
\Gamma^\theta_{r\theta} &= \Gamma^\theta_{\theta r} = \frac{1}{r} \\
\Gamma^\theta_{\theta\theta} &= 0.
\end{aligned} \tag{3.31}$$

From these and similar expressions, we can derive formulas for the divergence, gradient, and curl in curvilinear coordinate systems.

Contrariwise, even in a curved space it is still possible to make the Christoffel symbols vanish at any one point. This is because, as we argued in the last chapter, we can always make the first derivative of the metric vanish at a point; so by (3.27) the connection coefficients derived from this metric will also vanish. Of course this can only be established at a point, not in some neighborhood of the point. We will discuss this more fully in Section 3.4.

Another useful property is that the formula for the divergence of a vector (with respect to the Christoffel connection) has a simplified form. The covariant divergence of V^μ is given by

$$\nabla_\mu V^\mu = \partial_\mu V^\mu + \Gamma^\mu_{\mu\lambda} V^\lambda. \tag{3.32}$$

It is straightforward to show that the Christoffel connection satisfies

$$\Gamma^\mu_{\mu\lambda} = \frac{1}{\sqrt{|g|}} \partial_\lambda \sqrt{|g|}, \tag{3.33}$$

and we therefore obtain

$$\nabla_\mu V^\mu = \frac{1}{\sqrt{|g|}} \partial_\mu (\sqrt{|g|} V^\mu). \tag{3.34}$$

There are also formulas for the divergences of higher-rank tensors, but they are generally not such a great simplification.

We use the Christoffel covariant derivative in the curved-space version of Stokes's theorem (see Appendix E). If V^μ is a vector field over a region Σ with boundary $\partial\Sigma$, Stokes's theorem is

$$\boxed{\int_\Sigma \nabla_\mu V^\mu \sqrt{|g|}\, d^n x = \int_{\partial\Sigma} n_\mu V^\mu \sqrt{|\gamma|}\, d^{n-1} x,} \tag{3.35}$$

where n_μ is normal to $\partial\Sigma$, and γ_{ij} is the induced metric on $\partial\Sigma$. If the connection weren't metric-compatible or torsion-free, there would be additional terms in this equation.

The last thing we need to mention is that converting partial derivatives into covariant derivatives is not always necessary in order to construct well-defined tensors; in particular, the exterior derivative and the vector-field commutator are both well-defined in terms of partials, essentially because both involve an antisymmetrization that cancels the nontensorial piece of the partial derivative transformation law. The same feature implies that they could, on the other hand, be equally well-defined in terms of (torsion-free) covariant derivatives; the antisymmetrization causes the connection coefficient terms to vanish. Thus, if ∇ is the Christoffel connection, ω_μ is a one-form, and X^μ and Y^μ are vector fields, we can write

$$(\mathrm{d}\omega)_{\mu\nu} = \partial_{[\mu}\omega_{\nu]} = \nabla_{[\mu}\omega_{\nu]} \tag{3.36}$$

and

$$[X, Y]^{\mu} = X^{\lambda}\partial_{\lambda}Y^{\mu} - Y^{\lambda}\partial_{\lambda}X^{\mu} = X^{\lambda}\nabla_{\lambda}Y^{\mu} - Y^{\lambda}\nabla_{\lambda}X^{\mu}. \tag{3.37}$$

If the connection is not torsion-free, the last equalities in these expressions are no longer true; the more fundamental definitions of the exterior derivative and the commutator are those in terms of the partial derivative.

Before moving on, let's review the process by which we have been adding structures to our mathematical constructs. We started with the basic notion of a set, which you were presumed to be familiar with (informally, if not rigorously). We introduced the concept of open subsets of our set; this is equivalent to introducing a topology, and promoted the set to a topological space. Then by demanding that each open set look like a region of $\mathbf{R}^n$ (with n the same for each set) and that the coordinate charts be smoothly sewn together, the topological space became a manifold. A manifold is simultaneously a very flexible and powerful structure, and comes equipped naturally with a tangent bundle, tensor bundles of various ranks, the ability to take exterior derivatives, and so forth. We then proceeded to put a metric on the manifold, resulting in a manifold with metric (sometimes Riemannian manifold). Independently of the metric we found we could introduce a connection, allowing us to take covariant derivatives. Once we have a metric, however, there is automatically a unique torsion-free metric-compatible connection. Likewise we could introduce an independent volume form, although one is automatically determined by the metric. In principle there is nothing to stop us from introducing more than one connection, or volume form, or metric, on any given manifold. In general relativity we do have a physical metric, which determines volumes and the covariant derivative, and the independence of these notions is not a crucial feature.

3.3 ■ PARALLEL TRANSPORT AND GEODESICS

Now that we know how to take covariant derivatives, let's step back and put this in the context of differentiation more generally. We think of a derivative as a way of quantifying how fast something is changing. In the case of tensors, the crucial issue is "changing with respect to what?" An ordinary function defines a number at each point in spacetime, and it is straightforward to compare two different numbers, so we shouldn't be surprised that the partial derivative of functions remained valid on arbitrary manifolds. But a tensor is a map from vectors and dual vectors to the real numbers, and it's not clear how to compare such maps at different points in spacetime. Since we have successfully constructed a covariant derivative, can we think of it as somehow measuring the rate of change of tensors? The answer turns out to be yes: the covariant derivative quantifies the instantaneous rate of

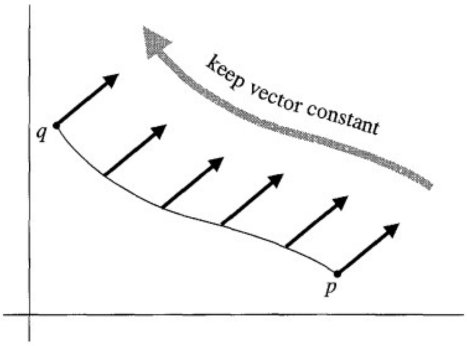

FIGURE 3.1 In flat space, we can parallel transport a vector by simply keeping its Cartesian components constant.

change of a tensor field in comparison to what the tensor would be if it were "parallel transported." In other words, a connection defines a specific way of keeping a tensor constant (along some path), on the basis of which we can compare nearby tensors.

It turns out that the concept of parallel transport is interesting in its own right, and worth spending some time thinking about. Recall that in flat space it is unnecessary to be very careful about the fact that vectors are elements of tangent spaces defined at individual points; it is actually very natural to compare vectors at different points (where by "compare" we mean add, subtract, take the dot product, and so on). The reason why it is natural is because it makes sense, in flat space, to move a vector from one point to another while keeping it constant, as illustrated in Figure 3.1. Then, once we get the vector from one point to another, we can do the usual operations allowed in a vector space.

This concept of moving a vector along a path, keeping constant all the while, is known as *parallel transport*. Parallel transport requires a connection to be well-defined; the intuitive manipulation of vectors in flat space makes implicit use of the Christoffel connection on this space. The crucial difference between flat and curved spaces is that, in a curved space, *the result of parallel transporting a vector from one point to another will depend on the path taken between the points.* Without yet assembling the complete mechanism of parallel transport, we can use our intuition about the two-sphere to see that this is the case. Start with a vector on the equator, pointing along a line of constant longitude. Parallel transport it up to the north pole along a line of longitude in the obvious way. Then take the original vector, parallel transport it along the equator by an angle θ, and then move it up to the north pole as before. As shown in Figure 3.2, it should be clear that the vector, parallel transported along two paths, arrived at the same destination with two different values (rotated by θ).

It therefore appears as if there is no natural way to uniquely move a vector from one tangent space to another; we can always parallel-transport it, but the result depends on the path, and there is no natural choice of which path to take.

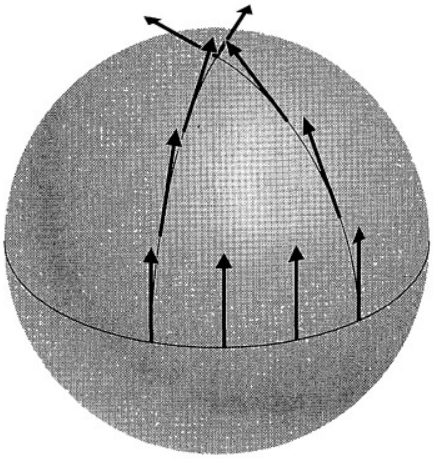

FIGURE 3.2 Parallel transport on a two-sphere. On a curved manifold, the result of parallel transport can depend on the path taken.

Unlike some of the problems we have encountered, *there is no solution to this one*—we simply must learn to live with the fact that two vectors can only be compared in a natural way if they are elements of the same tangent space. For example, two particles passing by each other have a well-defined relative velocity, which cannot be greater than the speed of light. But two particles at different points on a curved manifold do not have any well-defined notion of relative velocity—the concept simply makes no sense. Of course, in certain special situations it is still useful to talk as if it did make sense, but occasional usefulness is not a substitute for rigorous definition. In cosmology, for example, the light from distant galaxies is redshifted with respect to the frequencies we would observe from a nearby stationary source. Since this phenomenon bears such a close resemblance to the conventional Doppler effect due to relative motion, we are very tempted to say that the galaxies are "receding away from us" at a speed defined by their redshift. At a rigorous level this is nonsense, what Wittgenstein would call a "grammatical mistake"—the galaxies are not receding, since the notion of their velocity with respect to us is not well-defined. What is actually happening is that the metric of spacetime between us and the galaxies has changed (the universe has expanded) along the path of the photon from here to there, leading to an increase in the wavelength of the light. As an example of how you can go wrong, naive application of the Doppler formula to the redshift of galaxies implies that some of them are receding faster than light, in apparent contradiction with relativity. The resolution of this apparent paradox is simply that the very notion of their recession should not be taken literally.

Enough about what we cannot do; let's see what we can. Parallel transport is supposed to be the curved-space generalization of the concept of "keeping the vector constant" as we move it along a path; similarly for a tensor of arbitrary rank.

Given a curve $x^\mu(\lambda)$, the requirement of constancy of a tensor $T^{\mu_1\mu_2\cdots\mu_k}{}_{\nu_1\nu_2\cdots\nu_l}$ along this curve in flat space is simply that the components be constant:

$$\frac{d}{d\lambda}T^{\mu_1\mu_2\cdots\mu_k}{}_{\nu_1\nu_2\cdots\nu_l} = \frac{dx^\mu}{d\lambda}\frac{\partial}{\partial x^\mu}T^{\mu_1\mu_2\cdots\mu_k}{}_{\nu_1\nu_2\cdots\nu_l} = 0.$$

To make this properly tensorial we simply replace this partial derivative by a covariant one, and define the **directional covariant derivative** to be

$$\frac{D}{d\lambda} = \frac{dx^\mu}{d\lambda}\nabla_\mu. \tag{3.38}$$

This is a map, defined only along the path, from (k, l) tensors to (k, l) tensors. We then define **parallel transport** of the tensor T along the path $x^\mu(\lambda)$ to be the requirement that the covariant derivative of T along the path vanishes:

$$\left(\frac{D}{d\lambda}T\right)^{\mu_1\mu_2\cdots\mu_k}{}_{\nu_1\nu_2\cdots\nu_l} \equiv \frac{dx^\sigma}{d\lambda}\nabla_\sigma T^{\mu_1\mu_2\cdots\mu_k}{}_{\nu_1\nu_2\cdots\nu_l} = 0. \tag{3.39}$$

This is a well-defined tensor equation (since both the tangent vector $dx^\mu/d\lambda$ and the covariant derivative ∇T are tensors), known as the **equation of parallel transport**. For a vector it takes the form

$$\frac{d}{d\lambda}V^\mu + \Gamma^\mu_{\sigma\rho}\frac{dx^\sigma}{d\lambda}V^\rho = 0. \tag{3.40}$$

We can look at the parallel transport equation as a first-order differential equation defining an initial-value problem: given a tensor at some point along the path, there will be a unique continuation of the tensor to other points along the path such that the continuation solves (3.39). We say that such a tensor is parallel-transported.

The notion of parallel transport is obviously dependent on the connection, and different connections lead to different answers. If the connection is metric-compatible, the metric is always parallel transported with respect to it:

$$\frac{D}{d\lambda}g_{\mu\nu} = \frac{dx^\sigma}{d\lambda}\nabla_\sigma g_{\mu\nu} = 0. \tag{3.41}$$

It follows that the inner product of two parallel-transported vectors is preserved. That is, if V^μ and W^ν are parallel-transported along a curve $x^\sigma(\lambda)$, we have

$$\begin{aligned}\frac{D}{d\lambda}(g_{\mu\nu}V^\mu W^\nu) &= \left(\frac{D}{d\lambda}g_{\mu\nu}\right)V^\mu W^\nu + g_{\mu\nu}\left(\frac{D}{d\lambda}V^\mu\right)W^\nu + g_{\mu\nu}V^\mu\left(\frac{D}{d\lambda}W^\nu\right)\\ &= 0.\end{aligned} \tag{3.42}$$

This means that parallel transport with respect to a metric-compatible connection preserves the norm of vectors, the sense of orthogonality, and so on.

With parallel transport defined, the next logical step is to discuss geodesics. A geodesic is the curved-space generalization of the notion of a straight line in Euclidean space. We all know what a straight line is: it's the path of shortest distance between two points. But there is an equally good definition—a straight line is a path that parallel-transports its own tangent vector. It will turn out that these two concepts coincide if and only if the connection is the Christoffel connection.

We'll start with the second definition (a geodesic is a curve along which the tangent vector is parallel-transported), since it is computationally much more straightforward. The tangent vector to a path $x^\mu(\lambda)$ is $dx^\mu/d\lambda$. The condition that it be parallel transported is thus

$$\frac{D}{d\lambda}\frac{dx^\mu}{d\lambda} = 0, \tag{3.43}$$

or alternatively

$$\boxed{\frac{d^2x^\mu}{d\lambda^2} + \Gamma^\mu_{\rho\sigma}\frac{dx^\rho}{d\lambda}\frac{dx^\sigma}{d\lambda} = 0.} \tag{3.44}$$

This is the **geodesic equation**, another one you should memorize. We can easily see that it reproduces the usual notion of straight lines if the connection coefficients are the Christoffel symbols in Euclidean space; in that case we can choose Cartesian coordinates in which $\Gamma^\mu_{\rho\sigma} = 0$, and the geodesic equation is just $d^2x^\mu/d\lambda^2 = 0$, which is the equation for a straight line.

That was embarrassingly simple; let's turn to the more nontrivial case of the shortest-distance definition. As we know, various subtleties are involved in the definition of distance in a Lorentzian spacetime; for null paths the distance is zero, for timelike paths it's more convenient to use the proper time. So in the name of simplicity let's do the calculation just for a timelike path—the resulting equation will turn out to be good for any path, so we are not losing any generality. We therefore consider the proper time functional,

$$\tau = \int \left(-g_{\mu\nu}\frac{dx^\mu}{d\lambda}\frac{dx^\nu}{d\lambda}\right)^{1/2} d\lambda, \tag{3.45}$$

where the integral is over the path. To search for shortest-distance paths, we could do the usual calculus-of-variations treatment to seek critical points of this functional. They will turn out to be curves of *maximum* proper time, consistent with our discussion of the twin paradox in Chapter 1. However, we can simplify the algebra by means of a trick. The integral (3.45) is of the form $\tau = \int \sqrt{-f}\, d\lambda$, where $f = g_{\mu\nu}(dx^\mu/d\lambda)(dx^\nu/d\lambda)$. The variation looks like

$$\begin{aligned}\delta\tau &= \int \delta\sqrt{-f}\, d\lambda \\ &= -\int \frac{1}{2}(-f)^{-1/2}\,\delta f\, d\lambda.\end{aligned} \tag{3.46}$$

It makes things easier if we now specify that our parameter is the proper time τ itself, rather than the arbitrary parameter λ, so that the tangent vector is the four-velocity U^μ. This fixes the value of f,

$$f = g_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau} = g_{\mu\nu}U^\mu U^\nu = -1. \tag{3.47}$$

From (3.46) we then have

$$\delta\tau = -\frac{1}{2}\int \delta f\, d\tau. \tag{3.48}$$

Stationary points of (3.45)—paths for which $\delta\tau = 0$—are therefore equivalent to stationary points (with fixed parameterization) of the simpler integral

$$I = \frac{1}{2}\int f\, d\tau = \frac{1}{2}\int g_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau}\, d\tau. \tag{3.49}$$

(The $\frac{1}{2}$ is by no means necessary, but will make things nicer later on.) Taking variations of this expression is thus a shortcut to finding shortest-distance paths, one that we will wisely follow.

Stationary points of I will of course obey the Euler–Lagrange equations (1.128), but evaluating them involves repeated application of the chain rule, and it is just as simple to directly consider the change in the integral under infinitesimal variations of the path,

$$\begin{aligned} x^\mu &\rightarrow x^\mu + \delta x^\mu \\ g_{\mu\nu} &\rightarrow g_{\mu\nu} + (\partial_\sigma g_{\mu\nu})\delta x^\sigma. \end{aligned} \tag{3.50}$$

The second line comes from Taylor expansion in curved spacetime, which as you can see, uses the partial derivative, not the covariant derivative. This is because we are simply thinking of the components $g_{\mu\nu}$ as functions on spacetime in some specific coordinate system. Plugging this into (3.49) and keeping only terms first-order in δx^μ, we get

$$\delta I = \frac{1}{2}\int\left[\partial_\sigma g_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau}\delta x^\sigma + g_{\mu\nu}\frac{d(\delta x^\mu)}{d\tau}\frac{dx^\nu}{d\tau} + g_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{d(\delta x^\nu)}{d\tau}\right] d\tau. \tag{3.51}$$

The last two terms can be integrated by parts; for example,

$$\begin{aligned} \frac{1}{2}\int\left[g_{\mu\nu}\frac{dx^\mu}{d\tau}\frac{d(\delta x^\nu)}{d\tau}\right] d\tau &= -\frac{1}{2}\int\left[g_{\mu\nu}\frac{d^2x^\mu}{d\tau^2} + \frac{dg_{\mu\nu}}{d\tau}\frac{dx^\mu}{d\tau}\right]\delta x^\nu\, d\tau \\ &= -\frac{1}{2}\int\left[g_{\mu\nu}\frac{d^2x^\mu}{d\tau^2} + \partial_\sigma g_{\mu\nu}\frac{dx^\sigma}{d\tau}\frac{dx^\mu}{d\tau}\right]\delta x^\nu\, d\tau, \end{aligned} \tag{3.52}$$

where we have neglected boundary terms, which vanish because we take our variation δx^μ to vanish at the endpoints of the path. In the second line we have used the chain rule on the derivative of $g_{\mu\nu}$. The variation (3.51) then becomes, after rearranging some dummy indices,

$$\delta I = -\int \left[g_{\mu\sigma} \frac{d^2x^\mu}{d\tau^2} + \frac{1}{2}\left(\partial_\mu g_{\nu\sigma} + \partial_\nu g_{\sigma\mu} - \partial_\sigma g_{\mu\nu}\right) \frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau} \right] \delta x^\sigma \, d\tau. \tag{3.53}$$

Since we are searching for stationary points, we want δI to vanish for any variation δx^σ; this implies

$$g_{\mu\sigma} \frac{d^2x^\mu}{d\tau^2} + \frac{1}{2}\left(\partial_\mu g_{\nu\sigma} + \partial_\nu g_{\sigma\mu} - \partial_\sigma g_{\mu\nu}\right) \frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau} = 0, \tag{3.54}$$

and multiplying by the inverse metric $g^{\rho\sigma}$ finally leads to

$$\frac{d^2x^\rho}{d\tau^2} + \frac{1}{2} g^{\rho\sigma}\left(\partial_\mu g_{\nu\sigma} + \partial_\nu g_{\sigma\mu} - \partial_\sigma g_{\mu\nu}\right) \frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau} = 0. \tag{3.55}$$

We see that this is precisely the geodesic equation (3.40), but with the specific choice of Christoffel connection (3.27). Thus, on a manifold with metric, extremals of the length functional are curves that parallel transport their tangent vector with respect to the Christoffel connection associated with that metric. It doesn't matter if any other connection is defined on the same manifold. Of course, in GR the Christoffel connection is the only one used, so the two notions are the same.

The variational principle provides a convenient way to actually calculate the Christoffel symbols for a given metric. Rather than simply plugging into (3.27), it is often less work to explicitly vary the integral (3.49), with the metric of interest substituted in for $g_{\mu\nu}$. An example of this procedure is shown in Section 3.5.

3.4 ■ PROPERTIES OF GEODESICS

The primary usefulness of geodesics in general relativity is that they are the paths followed by unaccelerated test particles. A **test particle** is a body that does not itself influence the geometry through which it moves—never perfectly true, but often an excellent approximation. This concept allows us to explore, for example, the properties of the gravitational field around the Sun, without worrying about the field of the planet whose motion we are considering. The geodesic equation can be thought of as the generalization of Newton's law $\mathbf{f} = m\mathbf{a}$, for the case $\mathbf{f} = 0$, to curved spacetime. It is also possible to introduce forces by adding terms to the right-hand side; in fact, looking back to the expression (1.105)

$$\frac{d^2x^\mu}{d\tau^2} + \Gamma^\mu_{\rho\sigma} \frac{dx^\rho}{d\tau}\frac{dx^\sigma}{d\tau} = \frac{q}{m} F^\mu{}_\nu \frac{dx^\nu}{d\tau}. \tag{3.56}$$

We will talk about this more later, but in fact your guess would be correct.

We should say some more careful words about the parameterization of a geodesic path. When we presented the geodesic equation as the requirement that the tangent vector be parallel-transported, (3.44), we parameterized our path with some parameter λ, whereas when we found the formula (3.55) for the extremal of the spacetime interval, we wound up with a very specific parameterization, the proper time. Of course from the form of (3.55) it is clear that a transformation,

$$\tau \to \lambda = a\tau + b, \tag{3.57}$$

for some constants a and b, leaves the equation invariant. Any parameter related to the proper time in this way is called an **affine parameter**, and is just as good as the proper time for parameterizing a geodesic. What was hidden in our derivation of (3.44) was that *the demand that the tangent vector be parallel-transported actually constrains the parameterization of the curve*, specifically to one related to the proper time by (3.57). In other words, if you start at some point and with some initial direction, and then construct a curve by beginning to walk in that direction and keeping your tangent vector parallel transported, you will not only define a path in the manifold but also (up to linear transformations) define the parameter along the path.

Of course, there is nothing to stop you from using any other parameterization you like, but then (3.44) will not be satisfied. More generally you will satisfy an equation of the form

$$\frac{d^2x^\mu}{d\alpha^2} + \Gamma^\mu_{\rho\sigma}\frac{dx^\rho}{d\alpha}\frac{dx^\sigma}{d\alpha} = f(\alpha)\frac{dx^\mu}{d\alpha}, \tag{3.58}$$

for some parameter $\alpha(\lambda)$, where $f(\alpha)$ is related to the affine parameter by

$$f(\alpha) = -\left(\frac{d^2\alpha}{d\lambda^2}\right)\left(\frac{d\alpha}{d\lambda}\right)^{-2}. \tag{3.59}$$

Conversely, if (3.58) is satisfied along a curve you can always find an affine parameter $\lambda(\alpha)$ for which the geodesic equation (3.44) will be satisfied.

For timelike paths, we can write the geodesic equation in terms of the four-velocity $U^\mu = dx^\mu/d\tau$ as

$$U^\lambda \nabla_\lambda U^\mu = 0. \tag{3.60}$$

Similarly, in terms of the four-momentum $p^\mu = mU^\mu$, the geodesic equation is simply

$$p^\lambda \nabla_\lambda p^\mu = 0. \tag{3.61}$$

This relation expresses the idea that freely-falling particles keep moving in the direction in which their momenta are pointing.

For null paths, the proper time vanishes and τ is not an appropriate affine parameter. Nevertheless, it is still perfectly well-defined to ask whether a parameterized path $x^\mu(\lambda)$ satisfies the geodesic equation (3.44). If a null path is a geodesic for some parameter λ, it will also be a geodesic for any other affine parameter of the form $a\lambda + b$. However, there is no preferred choice among these parameters like the proper time is for timelike paths. Once we choose a parameter at some point along the path, of course, there is a unique continuation to the rest of the path if we want to solve the geodesic equation. It is often convenient to choose the normalization of the affine parameter λ along a null geodesic such that $dx^\mu/d\lambda$ is equal to the momentum four-vector:

$$p^\mu = \frac{dx^\mu}{d\lambda}. \tag{3.62}$$

This is in contrast to timelike paths, where $dx^\mu/d\tau$ is the momentum per unit mass. Then an observer with four-velocity U^μ measures the energy of the particle (or equivalently the frequency, since we are setting $\hbar = 1$) to be

$$E = -p_\mu U^\mu. \tag{3.63}$$

This expression always tells us the energy of a particle with momentum p^μ as measured by an observer with four-velocity U^μ, whether p^μ is null or timelike; you can check it by going to locally inertial coordinates. (A caveat: this expression for E does not include potential energy, only the intrinsic energy from motion and inertia. In a general spacetime there will not be a well-defined notion of gravitational potential energy, although in special cases it does exist.)

An important property of geodesics in a spacetime with Lorentzian metric is that the character (timelike/null/spacelike) of the geodesic, relative to a metric-compatible connection, never changes. This is simply because parallel transport preserves inner products, and the character is determined by the inner product of the tangent vector with itself. This is why we were consistent to consider purely timelike paths when we derived (3.55); for spacelike paths we would have derived the same equation, since the only difference is an overall minus sign in the final answer.

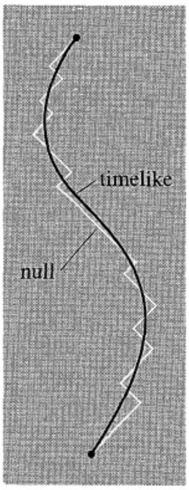

FIGURE 3.3 We can always approximate a timelike path by a sequence of null paths with a total path length of zero. Hence, timelike geodesics must be maxima of the proper time rather than minima.

Let's now explain the earlier remark that timelike geodesics are maxima of the proper time. The reason we know this is true is that, given any timelike curve (geodesic or not), we can approximate it to arbitrary accuracy by a null curve. To do this all we have to do is to consider "jagged" null curves that follow the timelike one, as portrayed in Figure 3.3. As we increase the number of sharp corners, the null curve comes closer and closer to the timelike curve while still having zero path length. Timelike geodesics cannot therefore be curves of minimum proper time, since they are always infinitesimally close to curves of less proper time (zero, in fact); actually they maximize the proper time. This is how you can remember which twin in the twin paradox ages more—the one who stays home is basically on a geodesic, and therefore experiences more proper time. Of course even this is being a little cavalier; actually every time we say "maximize" or "minimize" we should add the modifier "locally." Often the case is that between two points on a manifold there is more than one geodesic. For instance, on S^2 we can draw a great circle through any two points, and imagine traveling between them either the short way or the long way around. One of these is obviously longer than the other, although both are stationary points of the length functional.

Geodesics provide a convenient a way of mapping the tangent space T_p of a point p to a region of the manifold that contains p, called the **exponential map**. This map in turn defines a set of coordinates for this region that are automatically the locally inertial coordinates discussed in Section 2.5 [coordinates $x^{\hat{\mu}}$ around a point p such that $g_{\hat{\mu}\hat{\nu}}(p) = \eta_{\hat{\mu}\hat{\nu}}$ and $\partial_{\hat{\sigma}} g_{\hat{\mu}\hat{\nu}}(p) = 0$]. We begin by noticing that any vector $k \in T_p$ defines a unique geodesic passing through it, for which k is the tangent vector at p, and $\lambda(p) = 0$:

$$\frac{dx^\mu}{d\lambda}(\lambda = 0) = k^\mu. \tag{3.64}$$

Uniqueness follows from the fact that the geodesic equation is a second-order differential equation, and specifying initial data in the form $x^\mu(p)$ and $k^\mu = (dx^\mu/d\lambda)(p)$ completely determines a solution. On this geodesic there will be a unique point in M for which $\lambda = 1$. The exponential map at p, $\exp_p : T_p \to M$, is then defined as

$$\exp_p(k) = x^\nu(\lambda = 1), \tag{3.65}$$

where $x^\nu(\lambda)$ solves the geodesic equation subject to (3.64), as shown in Figure 3.4.

For some set of tangent vectors k^μ near the zero vector, this map will be well-defined, and in fact invertible. Depending on the geometry, however, different geodesics emanating from a single point may eventually cross, at which point $\exp_p : T_p \to M$ is no longer one-to-one. Furthermore, the range of the exponential map is not necessarily the whole manifold, and the domain is not necessarily the whole tangent space. The range can fail to be all of M simply because there can be two points that are not connected by any geodesic. An example is given by anti-de Sitter space, discussed in Chapter 8. The domain can fail to be all of T_p because a geodesic may run into a singularity, which we think of as "the edge of the manifold." Manifolds that have such singularities are known as **geodesically incomplete**. In a more careful discussion it would actually be the

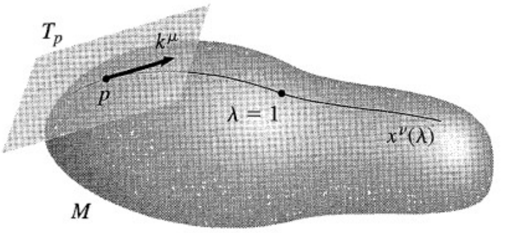

FIGURE 3.4 The exponential map takes a vector in T_p to a point in M that lies at unit affine parameter along the geodesic to which the vector is tangent.

other way around: the best way we have of defining a singularity is as a place where geodesics appear to "end," after we remove trivial cases in which a part of the manifold is artificially excluded by hand. See Wald (1984) or Hawking and Ellis (1973). This problem is not merely technical; the singularity theorems of Hawking and Penrose state that, for certain matter content, spacetimes in general relativity are almost guaranteed to be geodesically incomplete. As examples, two of the most useful spacetimes in GR—the Schwarzschild solution describing black holes and the Friedmann–Robertson–Walker solutions describing homogeneous, isotropic cosmologies—both feature important singularities; these will be discussed in later chapters.

We now use the exponential map to construct locally inertial coordinates. The easy part is to find basis vectors $\{\hat{e}_{(\hat{\mu})}\}$ for T_p such that the components of the metric are those of the canonical form:

$$g_{\hat{\mu}\hat{\nu}} = g(\hat{e}_{(\hat{\mu})}, \hat{e}_{(\hat{\nu})}) = \eta_{\hat{\mu}\hat{\nu}}. \tag{3.66}$$

Here $g(\ ,\ )$ denotes the metric, thought of as a multilinear map from $T_p \times T_p$ to **R**. And the hats have different meanings: over e they remind us that it's a basis vector, and over the indices they remind us that we are in locally inertial coordinates (as we shall see). This is easy because it's just linear algebra, not yet referring to coordinates; starting with any set of components for $g_{\mu\nu}$, we can always diagonalize this matrix and then rescale our basis vectors to satisfy (3.66). The hard part, we would expect, is finding a coordinate system $x^{\hat{\mu}}$ for which the basis vectors $\{\hat{e}_{(\hat{\mu})}\}$ comprise the coordinate basis, $\hat{e}_{(\hat{\mu})} = \partial_{\hat{\mu}}$, and such that the first partial derivatives of $g_{\hat{\mu}\hat{\nu}}$ vanish. In fact, however, the exponential map achieves this automatically. For any point q sufficiently close to p, there is a unique geodesic path connecting p to q, and a unique parameterization λ such that $\lambda(p) = 0$ and $\lambda(q) = 1$. At p the tangent vector k to this geodesic can be written as a linear combination of our basis vectors, $k = k^{\hat{\mu}}\hat{e}_{(\hat{\mu})}$. We define the sought-after coordinates x^{μ} simply to be these components: $x^{\hat{\mu}}(q) = k^{\hat{\mu}}$. In other words, we have defined the coordinates $x^{\hat{\mu}}(q)$ to be the components (with respect to our normalized basis $\{\hat{e}_{(\hat{\mu})}\}$) of the tangent vector k that gets mapped to q by $\exp_p$. Coordinates constructed in this way are known as **Riemann normal coordinates** at p.

We still need to verify that these Riemann normal coordinates satisfy $\partial_{\hat{\sigma}} g_{\hat{\mu}\hat{\nu}}(p) = 0$. Note that a ray in the tangent space (a parameterized set of vectors of the form $\lambda k^{\hat{\mu}}$, for some fixed vector $k^{\hat{\mu}}$) gets mapped to a geodesic by the exponential map. Therefore, in Riemann normal coordinates, a curve $x^{\hat{\mu}}(\lambda)$ of the form

$$x^{\hat{\mu}}(\lambda) = \lambda k^{\hat{\mu}} \tag{3.67}$$

will solve the geodesic equation. Indeed, *any* geodesic through p may be expressed this way, for some appropriate vector $k^{\hat{\mu}}$. We therefore have

$$\frac{d^2 x^{\hat{\mu}}}{d\lambda^2} = 0 \tag{3.68}$$

along any geodesic through p in this coordinate system. But, by the geodesic equation, we also have

$$\frac{d^2x^{\hat{\mu}}}{d\lambda^2}(p) = -\Gamma^{\hat{\mu}}_{\hat{\rho}\hat{\sigma}}(p)k^{\hat{\rho}}k^{\hat{\sigma}}, \tag{3.69}$$

where $k^{\hat{\rho}} = (dx^{\hat{\rho}}/d\lambda)(p)$. Since this holds for arbitrary $k^{\hat{\rho}}$, we conclude that

$$\Gamma^{\hat{\mu}}_{\hat{\rho}\hat{\sigma}}(p) = 0. \tag{3.70}$$

Now apply metric compatibility:

$$\begin{aligned} 0 &= \nabla_{\hat{\sigma}} g_{\hat{\mu}\hat{\nu}} \\ &= \partial_{\hat{\sigma}} g_{\hat{\mu}\hat{\nu}} - \Gamma^{\hat{\lambda}}_{\hat{\sigma}\hat{\mu}} g_{\hat{\lambda}\hat{\nu}} - \Gamma^{\hat{\lambda}}_{\hat{\sigma}\hat{\nu}} g_{\hat{\mu}\hat{\lambda}} \\ &= \partial_{\hat{\sigma}} g_{\hat{\mu}\hat{\nu}}, \end{aligned} \tag{3.71}$$

where all quantities are evaluated at p. We see that Riemann normal coordinates provide a realization of the locally inertial coordinates discussed in Section 2.5. They are not unique; there are an infinite number of non-Riemann-normal coordinate systems in which $g_{\hat{\mu}\hat{\nu}}(p) = \eta_{\hat{\mu}\hat{\nu}}$ and $\partial_{\hat{\sigma}} g_{\hat{\mu}\hat{\nu}}(p) = 0$, but in an expansion around p they will differ from the Riemann normal coordinates only at third order in $x^{\hat{\mu}}$.

3.5 ■ THE EXPANDING UNIVERSE REVISITED

Let's put some of the technology we have developed to work in understanding a simple metric. Recall the expanding-universe metric we studied in Chapter 2,

$$\begin{aligned} ds^2 &= -\mathrm{d}t^2 + a^2(t)[\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2] \\ &= -\mathrm{d}t^2 + a^2(t)\delta_{ij}\mathrm{d}x^i\mathrm{d}x^j. \end{aligned} \tag{3.72}$$

This metric describes a universe consisting of flat spatial sections expanding as a function of time, with the relative distance between particles at fixed spatial coordinates growing proportionally to the scale factor $a(t)$.

Faced with a metric, the first thing we do is to calculate the Christoffel symbols. As mentioned at the end of Section 3.3, the easiest technique for doing so is actually to vary explicitly an integral of the form (3.49). Plugging in the metric under consideration, we have

$$I = \frac{1}{2}\int \left[-\left(\frac{dt}{d\tau}\right)^2 + a^2(t)\delta_{ij}\frac{dx^i}{d\tau}\frac{dx^j}{d\tau} \right] d\tau. \tag{3.73}$$

The technique is to consider variations $x^\mu \to x^\mu + \delta x^\mu$ and demand that δI vanish. We get n equations on an n-dimensional manifold (in this case $n = 4$), one for each μ; each equation corresponds to a component of the geodesic equation

(3.44). In the equation derived by varying with respect to x^μ, the coefficient of $(dx^\rho/d\tau)(dx^\sigma/d\tau)$ will be $\Gamma^\mu_{\rho\sigma}$.

For the metric (3.72), we need to consider separately variations with respect to $x^0 = t$ and one of the x^i's (it doesn't matter which one, since the results for each spacelike direction will be equivalent). Let's start with $t \to t + \delta t$. The nontrivial time dependence comes from the scale factor, for which, to first order,

$$a(t + \delta t) = a(t) + \dot{a}\delta t, \tag{3.74}$$

where $\dot{a} = da/dt$. We therefore have

$$\begin{aligned}\delta I &= \frac{1}{2}\int\left[-2\frac{dt}{d\tau}\frac{d(\delta t)}{d\tau} + 2a\dot{a}\delta_{ij}\frac{dx^i}{d\tau}\frac{dx^j}{d\tau}\delta t\right]d\tau \\ &= \int\left[\frac{d^2t}{d\tau^2} + a\dot{a}\delta_{ij}\frac{dx^i}{d\tau}\frac{dx^j}{d\tau}\right]\delta t\, d\tau,\end{aligned} \tag{3.75}$$

where we have dropped a boundary term after integrating by parts (as always). Setting the coefficient of δt equal to zero implies

$$\frac{d^2t}{d\tau^2} + a\dot{a}\delta_{ij}\frac{dx^i}{d\tau}\frac{dx^j}{d\tau} = 0, \tag{3.76}$$

which is supposed to be equivalent to the geodesic equation (with $\mu = 0$)

$$\frac{d^2x^0}{d\tau^2} + \Gamma^0_{\rho\sigma}\frac{dx^\rho}{d\tau}\frac{dx^\sigma}{d\tau} = 0. \tag{3.77}$$

Comparison of these two equations implies

$$\begin{aligned}\Gamma^0_{00} &= 0, \\ \Gamma^0_{i0} = \Gamma^0_{0i} &= 0, \\ \Gamma^0_{ij} &= a\dot{a}\,\delta_{ij}.\end{aligned} \tag{3.78}$$

We can repeat this procedure for a spatial coordinate, $x^i \to x^i + \delta x^i$. The variation is then

$$\begin{aligned}\delta I &= \frac{1}{2}\int a^2\left(2\delta_{ij}\frac{dx^i}{d\tau}\frac{d(\delta x^j)}{d\tau}\right)d\tau \\ &= -\int\left(a^2\frac{d^2x^i}{d\tau^2} + 2a\frac{da}{d\tau}\frac{dx^i}{d\tau}\right)\delta_{ij}\delta x^j\, d\tau.\end{aligned} \tag{3.79}$$

We can express $da/d\tau$ in terms of $\dot{a}$ by using the chain rule,

$$\frac{da}{d\tau} = \dot{a}\frac{dt}{d\tau}. \tag{3.80}$$

Then setting the coefficient of δx^j equal to zero in (3.79) implies

$$\frac{d^2x^i}{d\tau^2}+2\frac{\dot{a}}{a}\frac{dt}{d\tau}\frac{dx^i}{d\tau}=0. \tag{3.81}$$

Comparing to the geodesic equation, we find that the Christoffel symbols must satisfy

$$\Gamma^i_{\rho\sigma}\frac{dx^\rho}{d\tau}\frac{dx^\sigma}{d\tau}=2\frac{\dot{a}}{a}\frac{dt}{d\tau}\frac{dx^i}{d\tau}. \tag{3.82}$$

The Christoffel symbols are therefore given by

$$\begin{aligned}\Gamma^i_{00}&=0\\ \Gamma^i_{j0}=\Gamma^i_{0j}&=\frac{\dot{a}}{a}\delta^i_j\\ \Gamma^i_{jk}&=0.\end{aligned} \tag{3.83}$$

Together, (3.78) and (3.83) are all of the connection coefficients for the metric (3.72). These are, of course, necessary both for studying geodesics of the spacetime and for taking covariant derivatives; in fact, (3.76) and (3.81) together *are* the geodesic equation. Let's put this to work by solving for null geodesics, those followed by massless particles such as photons, for which we have to use λ rather than τ as a parameter. Without loss of generality we can consider paths along the x-direction, for which $x^\mu(\lambda)=\{t(\lambda),x(\lambda),0,0\}$. It is trivial to solve for null paths of this sort, using $ds^2=0$. We have

$$0=-\mathrm{d}t^2+a^2(t)\mathrm{d}x^2, \tag{3.84}$$

which implies

$$\frac{dx}{d\lambda}=\frac{1}{a}\frac{dt}{d\lambda}. \tag{3.85}$$

In Section 2.6 we solved this for $a=t^q$, but here we will remain more general. Also, we have chosen to consider paths moving in the positive x-direction, which determines the sign of $dx/d\lambda$. We must distinguish, however, between "null paths" and "null geodesics": the latter are a much more restrictive class, and to show that these paths are geodesics, we need to solve for the coordinates t and x in terms of the parameter λ.

Let's solve for $dt/d\lambda$, which will turn out to be the quantity in which we are most interested. Plugging the null condition (3.85) into the $\mu=0$ component of the geodesic equation (3.76), and remembering to replace $\tau\to\lambda$, we get

$$\frac{d^2t}{d\lambda^2}+\frac{\dot{a}}{a}\left(\frac{dt}{d\lambda}\right)^2=0. \tag{3.86}$$

It is straightforward to verify that this is solved by

$$\frac{dt}{d\lambda} = \frac{\omega_0}{a}, \tag{3.87}$$

where ω_0 is a constant. For a given $a(t)$, this could be instantly integrated to yield $t(\lambda)$. But more interesting is to consider the energy E of the photon as it would be measured by a comoving observer (one at fixed spatial coordinates), who would have four-velocity

$$U^\mu = (1, 0, 0, 0). \tag{3.88}$$

Don't get tricked into thinking that the timelike component of the four-velocity of a particle at rest will always equal unity; we need to satisfy the normalization condition $g_{\mu\nu}U^\mu U^\nu = -1$, which in the rest frame ($U^i = 0$) implies $U^0 = \sqrt{-g_{00}}$. According to (3.63), and using $p^\mu = dx^\mu/d\lambda$, we have

$$\begin{aligned} E &= -p_\mu U^\mu \\ &= -g_{00}\frac{dx^0}{d\lambda}U^0 \\ &= \frac{\omega_0}{a}. \end{aligned} \tag{3.89}$$

We see why the notation ω_0 was chosen for the constant of proportionality in (3.87): ω_0 is simply the frequency of the photon when $a = 1$. Recall that $E = \hbar\omega$, and we are using units in which $\hbar = 1$.

We have uncovered a profound phenomenon: the **cosmological redshift**. A photon emitted with energy E_1 at scale factor a_1 and observed with energy E_2 at scale factor a_2 will have

$$\frac{E_2}{E_1} = \frac{a_1}{a_2}. \tag{3.90}$$

This is called a "redshift" because the wavelength of the photon is inversely proportional to the frequency, and in an expanding universe the wavelength therefore grows with time. As a practical matter this provides an easy way to measure the change in the scale factor between us and distant galaxies, and also serves as a proxy for the distance: since the universe has been monotonically expanding, a greater redshift implies a greater distance. In conventional notation, the amount of redshift is denoted by

$$z = \frac{\omega_1 - \omega_2}{\omega_2} = \frac{a_2}{a_1} - 1, \tag{3.91}$$

so that z vanishes if there has been no expansion, for instance, if the emitter and observer are so close that there hasn't been enough time for the universe to appreciably expand.

As mentioned in Section 3.3, the cosmological redshift is *not* a Doppler shift (despite the understandable temptation to refer to the "velocity" of receding galaxies). Now we can understand this statement quantitatively. You might imagine that, as far as the behavior of emitted photons is concerned, there is little difference between two galaxies physically moving apart in a flat spacetime and two galaxies at fixed comoving coordinates in an expanding spacetime. But let's consider a specific (unrealistic, but educational) example. Start with flat spacetime, and imagine that our two galaxies are initially not moving apart, but are at rest in some globally inertial coordinate system. One emits a photon toward the other; while the photon is traveling, we quickly move the two galaxies apart until they are twice their original separation, then leave them stationary at that distance; and then the photon is absorbed by the second galaxy. Clearly there will be no Doppler shift, since the galaxies were at rest both at emission and absorption. Now consider the analogous phenomenon in an expanding spacetime, with the galaxies stuck at fixed comoving coordinates. We begin with the scale factor constant (the universe is not expanding). One galaxy emits a photon, and we imagine that during the photon's journey the universe starts expanding until the scale factor is twice its original size, and then stops expanding before the photon is absorbed. In this case there certainly will be a redshift, despite the fact that there was no "relative motion" (an ill-defined concept in any case) at either absorption or emission; the photon's wavelength will have doubled as the scale factor doubled, so we observe a redshift $z = 1$. This demonstrates the conceptual distinction between the cosmological redshift and the conventional Doppler effect.

Beyond the geodesic equation, covariant derivatives will play a role in generalizing laws of physics from the flat spacetime of special relativity to the curved geometry of general relativity. As we will discuss in more detail in the next chapter, a simple rule of thumb is simply to replace all partial derivatives by covariant derivatives, and all appearances of the flat spacetime metric $\eta_{\mu\nu}$ by the curved metric $g_{\mu\nu}$. For example, the energy-momentum conservation equation of special relativity, $\partial_\mu T^{\mu\nu} = 0$, where $T^{\mu\nu}$ is the energy-momentum tensor, becomes

$$\nabla_\mu T^{\mu\nu} = 0. \tag{3.92}$$

In cosmology, we typically model the matter filling the universe as a perfect fluid; the corresponding energy-momentum tensor comes from generalizing (1.114) to curved spacetime,

$$T^{\mu\nu} = (\rho + p)U^\mu U^\nu + p g^{\mu\nu}. \tag{3.93}$$

Recall that ρ is the energy density, p is the pressure, and U^μ is the four-velocity of the fluid. For the metric (3.72) the components of the inverse metric are

$$g^{\mu\nu} = \begin{pmatrix} -1 & & & \\ & a^{-2} & & \\ & & a^{-2} & \\ & & & a^{-2} \end{pmatrix}. \tag{3.94}$$

We can take the fluid to be in its rest frame in these coordinates, so that the components of the four-velocity are $U^\mu = (1, 0, 0, 0)$. In fact the fluid would have to be in its rest frame for this particular metric to solve Einstein's equation, as we will later see. The energy-momentum tensor therefore takes the form

$$T^{\mu\nu} = \begin{pmatrix} \rho & & & \\ & a^{-2}p & & \\ & & a^{-2}p & \\ & & & a^{-2}p \end{pmatrix}. \tag{3.95}$$

Note that these components are specific to the metric (3.72), and will generally look different for other metrics.

Let's see what the energy-momentum conservation equation $\nabla_\mu T^{\mu\nu} = 0$ implies for a perfect fluid in an expanding universe. The rule for covariant derivatives (3.17) implies

$$\nabla_\mu T^{\mu\nu} = \partial_\mu T^{\mu\nu} + \Gamma^\mu_{\mu\lambda} T^{\lambda\nu} + \Gamma^\nu_{\mu\lambda} T^{\mu\lambda} = 0. \tag{3.96}$$

This equation has four components, one for each μ, although the three $\mu = i \in \{1, 2, 3\}$ are equivalent. Let's first look at the $\nu = 0$ component, piece by piece. The first term is straightforward,

$$\partial_\mu T^{\mu 0} = \partial_0 T^{00} = \dot{\rho}. \tag{3.97}$$

The second term is

$$\Gamma^\mu_{\mu\lambda} T^{\lambda 0} = \Gamma^\mu_{\mu 0} T^{00} = 3\frac{\dot{a}}{a}\rho, \tag{3.98}$$

and the third term is

$$\Gamma^0_{\mu\lambda} T^{\mu\lambda} = \Gamma^0_{00} T^{00} + \Gamma^0_{11} T^{11} + \Gamma^0_{22} T^{22} + \Gamma^0_{33} T^{33} = 3\frac{\dot{a}}{a}p. \tag{3.99}$$

In each of these sets of equations, we have first invoked the fact that $T^{\mu\nu}$ is diagonal, and then used the explicit formulae for the energy-momentum tensor and the connection coefficients in this metric. All together, then, we find

$$\dot{\rho} = -3\frac{\dot{a}}{a}(\rho + p). \tag{3.100}$$

Now let's look at one of the spatial components, choosing $\nu = 1$ for definiteness. Once again working piece by piece, we have for the first term in (3.96),

$$\partial_\mu T^{\mu 1} = \partial_1 T^{11} = a^{-2}\partial_x p. \tag{3.101}$$

The second and third terms are

$$\Gamma^\mu_{\mu\lambda} T^{\lambda 1} = \Gamma^\mu_{\mu 1} T^{11} = 0, \tag{3.102}$$

and

$$\Gamma^1_{\mu\lambda}T^{\mu\lambda} = \Gamma^1_{00}T^{00} + \Gamma^1_{11}T^{11} + \Gamma^1_{22}T^{22} + \Gamma^1_{33}T^{33} = 0. \tag{3.103}$$

Equivalent results will hold for $\nu = 2$ and $\nu = 3$. So the spatial components of the energy-momentum conservation equation simply amount to

$$\partial_i p = 0. \tag{3.104}$$

It is illuminating to compare these results to those we would obtain in Minkowski spacetime, which can be found by simply setting $a = 1$, $\dot{a} = 0$. The pressure-gradient equation (3.104) is unaffected, so there is no effect of curvature on the spatial components: for a fluid that is motionless as measured by a comoving observer, the pressure must be constant throughout space. For the timelike component, on the other hand, the expansion of the universe introduces a nonzero right-hand side to (3.100). To understand the consequences of this new feature, let us consider equations of state of the form

$$p = w\rho, \tag{3.105}$$

where w is some constant. Then (3.100) becomes

$$\frac{\dot{\rho}}{\rho} = -3(1+w)\frac{\dot{a}}{a}, \tag{3.106}$$

which can be solved to yield

$$\rho \propto a^{-3(1+w)}. \tag{3.107}$$

In Chapter 1 we mentioned three kinds of perfect fluid with equations of state of the form (3.105): dust, with $w = 0$; radiation, with $w = \frac{1}{3}$; and vacuum, with $w = -1$. A set of nonrelativistic, noninteracting particles behaves like dust; a set of photons or other massless particles behaves like radiation; and a nonzero constant energy density throughout spacetime acts like vacuum. From (3.107) we see that the equation of state determines how the energy density evolves as the universe expands:

$$\begin{array}{rll} \text{matter} & p = 0 & \rho \propto a^{-3} \\ \text{radiation} & p = \frac{1}{3}\rho & \rho \propto a^{-4} \\ \text{vacuum} & p = -\rho & \rho = \text{constant}. \end{array} \tag{3.108}$$

We will explore these behaviors more thoroughly in Chapter 8; for right now let's simply note that they make sense. For dust, the energy density comes from the rest mass of each particle; if all the particles have mass m, the energy density is simply $\rho = nm$, where n is the number density. Since the number density goes down as a^{-3} (the physical volume of a comoving region goes up, while the total number of particles stays constant), while the masses remain unchanged, we expect that the energy density obeys $\rho \propto a^{-3}$. For radiation, meanwhile, the energy of each particle (such as a photon) redshifts away as a^{-1} as the universe expands; since the number density still obeys $n \propto a^{-3}$, we expect that $\rho \propto a^{-4}$. Finally, the vacuum energy density is an intrinsic and unchanging amount of energy in any physical volume; it doesn't redshift at all as the universe expands, so we get $\rho =$ constant.

This example brings to life the differences between flat and curved spacetimes. For example, consider what we might be tempted to call the "energy," the integral over space of the energy density: $E = \int \rho a^3\, d^3x$, where the boundaries are at fixed comoving coordinates, so the region expands along with the universe, and the factor of a^3 comes from the square root of the determinant of the spatial metric $a^2\delta_{ij}$. This number is clearly not conserved in general. For dust, since $\rho \propto a^{-3}$, E remains constant as the universe expands; but for radiation it decreases, and for vacuum energy it increases. This is upsetting, since conservation of energy is one of the more cherished principles of physics. What has happened? One way of thinking about this is from the viewpoint of Noether's theorem, which states that every symmetry implies a conserved quantity. Energy is the conserved quantity that derives from invariance under time translations. Clearly, in an expanding universe, the energy-momentum tensor is defined on a background that is changing with time; therefore there is no reason to believe that the energy should be conserved. ("There is no timelike Killing vector," in the language to be introduced in Section 3.8.) Nevertheless, we continue to refer to $\nabla_\mu T^{\mu\nu} = 0$ as the energy-momentum conservation equation. It conveys the idea that there is a definite law obeyed by the energy-momentum tensor, even if there is no integral corresponding to a conserved energy. The transition from flat to curved spacetime induces the additional Christoffel-symbol terms in (3.96); these terms serve, roughly speaking, to allow transfer of energy between the matter fields (comprising $T^{\mu\nu}$) and the gravitational field. This notion is not very formal, however, and you shouldn't push it too far—it turns out to be difficult to associate a local energy density to the gravitational field, although it is possible in certain circumstances.

Of course there is also a notion of time-translation invariance that refers not to the background spacetime, but to the theory itself (that is, to the equations that define the theory rather than a specific solution to them). We haven't yet developed the dynamical equations of general relativity, but they will turn out to be invariant under time translations, as well as under any other sort of coordinate transformations, as indeed they must be. This general coordinate invariance leads to a set of constraints on allowed configurations of the theory, and generally requires a more subtle analysis.

In the end, you should come to accept that there is a profound difference between flat and curved spacetimes, and some of our favorite notions from flat-spacetime physics will be seriously modified in this more general context. This is not a sign of any flaw in general relativity, but a natural consequence of discarding the rigid spacetime geometry we learn to take for granted.

3.6 ■ THE RIEMANN CURVATURE TENSOR

Having set up the machinery of covariant derivatives and parallel transport, we are at last prepared to discuss curvature proper. The curvature is quantified by the Riemann tensor, which is derived from the connection. The idea behind this measure of curvature is that we know what we mean by "flatness" of a connection—the conventional (and usually implicit) Christoffel connection associated with a Euclidean or Minkowskian metric has a number of properties that can be thought of as different manifestations of flatness. These include the fact that parallel transport around a closed loop leaves a vector unchanged, that covariant derivatives of tensors commute, and that initially parallel geodesics remain parallel. As we shall see, the Riemann tensor arises when we study how any of these properties are altered in more general contexts.

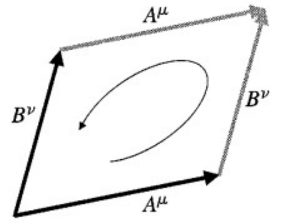

FIGURE 3.5 An infinitesimal loop defined by two vectors A^μ and B^ν.

We have already argued, using the two-sphere as an example, that parallel transport of a vector around a closed loop in a curved space will lead to a transformation of the vector. The resulting transformation depends on the total curvature enclosed by the loop; it would be more useful to have a local description of the curvature at each point, which is what the Riemann tensor is supposed to provide. One conventional way to introduce the Riemann tensor, therefore, is to consider parallel transport around an infinitesimal loop. We are not going to do that here, but take a more direct route. Nevertheless, even without working through the details, it is possible to see what form the answer should take. Since spacetime looks flat in sufficiently small regions, our loop will be specified by two (infinitesimal) vectors A^μ and B^ν. We imagine parallel transporting a vector V^μ by first moving it in the direction of A^μ, then along B^ν, then backward along A^μ and B^ν to return to the starting point, as shown in Figure 3.5. We know the action of parallel transport is independent of coordinates, so there should be some tensor that tells us how the vector changes when it comes back to its starting point; it will be a linear transformation on a vector, and therefore involve one upper and one lower index. But it will also depend on the two vectors A and B that define the loop; therefore there should be two additional lower indices to contract with A^μ and B^ν. Furthermore, the tensor should be antisymmetric in these two indices, since interchanging the vectors corresponds to traversing the loop in the opposite direction, and should give the inverse of the original answer. This is consistent with the fact that the transformation should vanish if A and B are the same vector. We therefore expect that the expression for the change δV^ρ experienced by this vector when parallel transported around the loop should be of the form

$$\delta V^\rho = R^\rho{}_{\sigma\mu\nu} V^\sigma A^\mu B^\nu, \tag{3.109}$$

where $R^\rho{}_{\sigma\mu\nu}$ is a $(1, 3)$ tensor known as the **Riemann tensor** (or simply curvature tensor). It is antisymmetric in the last two indices:

$$R^\rho{}_{\sigma\mu\nu} = -R^\rho{}_{\sigma\nu\mu}. \tag{3.110}$$

Of course, if (3.109) is taken as a definition of the Riemann tensor, a convention needs to be chosen for the ordering of the indices. There is no agreement at all on what this convention should be, so be careful.

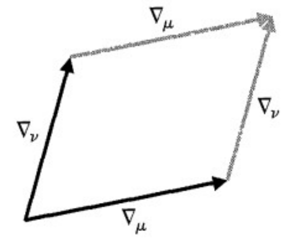

FIGURE 3.6 The commutator of two covariant derivatives.

Knowing what we do about parallel transport, we could very carefully perform the necessary manipulations to see what happens to the vector under this operation, and the result would be a formula for the curvature tensor in terms of the connection coefficients. It is much quicker, however, to consider a related operation, the commutator of two covariant derivatives. The relationship between this and parallel transport around a loop should be evident; the covariant derivative of a tensor in a certain direction measures how much the tensor changes relative to what it would have been if it had been parallel transported, since the covariant derivative of a tensor in a direction along which it is parallel transported is zero. The commutator of two covariant derivatives, then, measures the difference between parallel transporting the tensor first one way and then the other, versus the opposite ordering, as shown in Figure 3.6.

The actual computation is very straightforward. Considering a vector field V^ρ, we take

$$\begin{aligned}
[\nabla_\mu, \nabla_\nu]V^\rho &= \nabla_\mu\nabla_\nu V^\rho - \nabla_\nu\nabla_\mu V^\rho \\
&= \partial_\mu(\nabla_\nu V^\rho) - \Gamma^\lambda_{\mu\nu}\nabla_\lambda V^\rho + \Gamma^\rho_{\mu\sigma}\nabla_\nu V^\sigma - (\mu \leftrightarrow \nu) \\
&= \partial_\mu\partial_\nu V^\rho + (\partial_\mu\Gamma^\rho_{\nu\sigma})V^\sigma + \Gamma^\rho_{\nu\sigma}\partial_\mu V^\sigma - \Gamma^\lambda_{\mu\nu}\partial_\lambda V^\rho - \Gamma^\lambda_{\mu\nu}\Gamma^\rho_{\lambda\sigma}V^\sigma \\
&\quad + \Gamma^\rho_{\mu\sigma}\partial_\nu V^\sigma + \Gamma^\rho_{\mu\sigma}\Gamma^\sigma_{\nu\lambda}V^\lambda - (\mu \leftrightarrow \nu) \\
&= (\partial_\mu\Gamma^\rho_{\nu\sigma} - \partial_\nu\Gamma^\rho_{\mu\sigma} + \Gamma^\rho_{\mu\lambda}\Gamma^\lambda_{\nu\sigma} - \Gamma^\rho_{\nu\lambda}\Gamma^\lambda_{\mu\sigma})V^\sigma - 2\Gamma^\lambda_{[\mu\nu]}\nabla_\lambda V^\rho.
\end{aligned} \tag{3.111}$$

In the last step we have relabeled some dummy indices and eliminated some terms that cancel when antisymmetrized. We recognize that the last term is simply the torsion tensor, and that the left hand side is manifestly a tensor; therefore the expression in parentheses must be a tensor itself. We write

$$[\nabla_\mu, \nabla_\nu]V^\rho = R^\rho{}_{\sigma\mu\nu}V^\sigma - T^\lambda{}_{\mu\nu}\nabla_\lambda V^\rho, \tag{3.112}$$

where the Riemann tensor is identified as

$$\boxed{R^\rho{}_{\sigma\mu\nu} = \partial_\mu\Gamma^\rho_{\nu\sigma} - \partial_\nu\Gamma^\rho_{\mu\sigma} + \Gamma^\rho_{\mu\lambda}\Gamma^\lambda_{\nu\sigma} - \Gamma^\rho_{\nu\lambda}\Gamma^\lambda_{\mu\sigma}.} \tag{3.113}$$

Notice a number of things about the derivation of this expression:

- Of course we have not demonstrated that (3.113) is actually the same tensor that appeared in (3.109), but in fact it's true. You are asked to show this in the Exercises.
- It is perhaps surprising that the commutator $[\nabla_\mu, \nabla_\nu]$, which appears to be a differential operator, has an action on vector fields that (in the absence of

torsion, at any rate) is a simple multiplicative transformation. The Riemann tensor measures that part of the commutator of covariant derivatives that is proportional to the vector field, while the torsion tensor measures the part that is proportional to the covariant derivative of the vector field; the second derivative doesn't enter at all.

- Notice that the expression (3.113) is constructed from nontensorial elements; you can check that the transformation laws all work out to make this particular combination a legitimate tensor.
- The antisymmetry of $R^\rho{}_{\sigma\mu\nu}$ in its last two indices is immediate from this formula and its derivation.
- We constructed the curvature tensor completely from the connection (no mention of the metric was made). We were sufficiently careful that the above expression is true for any connection, whether or not it is metric compatible or torsion free.
- Using what are by now our usual methods, the action of $[\nabla_\rho, \nabla_\sigma]$ can be computed on a tensor of arbitrary rank. The answer is

$$\begin{aligned}
[\nabla_\rho, \nabla_\sigma] & X^{\mu_1\cdots\mu_k}{}_{\nu_1\cdots\nu_l} \\
&= -T^\lambda{}_{\rho\sigma}\nabla_\lambda X^{\mu_1\cdots\mu_k}{}_{\nu_1\cdots\nu_l} \\
&\quad + R^{\mu_1}{}_{\lambda\rho\sigma} X^{\lambda\mu_2\cdots\mu_k}{}_{\nu_1\cdots\nu_l} + R^{\mu_2}{}_{\lambda\rho\sigma} X^{\mu_1\lambda\cdots\mu_k}{}_{\nu_1\cdots\nu_l} + \cdots \\
&\quad - R^\lambda{}_{\nu_1\rho\sigma} X^{\mu_1\cdots\mu_k}{}_{\lambda\nu_2\cdots\nu_l} - R^\lambda{}_{\nu_2\rho\sigma} X^{\mu_1\cdots\mu_k}{}_{\nu_1\lambda\cdots\nu_l} - \cdots. \qquad (3.114)
\end{aligned}$$

Both the torsion tensor and the Riemann tensor, thought of as multilinear maps, have elegant expressions in terms of the vector-field commutator. Thinking of the torsion as a map from two vector fields to a third vector field, we have

$$T(X, Y) = \nabla_X Y - \nabla_Y X - [X, Y], \qquad (3.115)$$

and thinking of the Riemann tensor as a map from three vector fields to a fourth one, we have (in funny-looking but standard notation)

$$R(X, Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z. \qquad (3.116)$$

In these expressions, the notation ∇_X refers to the covariant derivative along the vector field X; in components, $\nabla_X = X^\mu \nabla_\mu$. So, for example, (3.116) is equivalent to

$$\begin{aligned}
R^\rho{}_{\sigma\mu\nu} X^\mu Y^\nu Z^\sigma &= X^\lambda \nabla_\lambda (Y^\eta \nabla_\eta Z^\rho) - Y^\lambda \nabla_\lambda (X^\eta \nabla_\eta Z^\rho) \\
&\quad - (X^\lambda \partial_\lambda Y^\eta - Y^\lambda \partial_\lambda X^\eta) \nabla_\eta Z^\rho, \qquad (3.117)
\end{aligned}$$

which you can check is equivalent to (3.113). Note that the two vectors X and Y in (3.116) correspond to the last two indices in the component form of the Riemann tensor. The last term in (3.116), involving the commutator $[X, Y]$, vanishes when X and Y are taken to be the coordinate basis vector fields (since $[\partial_\mu, \partial_\nu] = 0$), which is why this term did not arise when we originally took the commutator of two covariant derivatives. We will not use this notation extensively, but you might see it in the literature, so you should be able to decode it.

Having defined the curvature tensor as something that characterizes the connection, let us now admit that in GR we are most concerned with the Christoffel connection. In this case the connection is derived from the metric, and the associated curvature may be thought of as that of the metric itself. This identification allows us to finally make sense of our informal notion that spaces for which the metric looks Euclidean or Minkowskian are flat. In fact it works both ways:

- If a coordinate system exists in which the components of the metric are constant, the Riemann tensor will vanish.
- If the Riemann tensor vanishes, we can always construct a coordinate system in which the metric components are constant.

Technically, these statements should be restricted to a region of the manifold that is simply-connected (all loops in the region can be smoothly deformed to a point without leaving the region); we will implicitly assume this condition below.

The first of these is easy to show. If we are in some coordinate system such that $\partial_\sigma g_{\mu\nu} = 0$ everywhere, not just at a point, then $\Gamma^\rho_{\mu\nu} = 0$ and $\partial_\sigma \Gamma^\rho_{\mu\nu} = 0$; thus $R^\rho{}_{\sigma\mu\nu} = 0$ by (3.113). But this is a tensor equation, and if it is true in one coordinate system it must be true in any coordinate system. Therefore, the statement that the Riemann tensor vanishes is a necessary condition for it to be possible to find coordinates in which the components of $g_{\mu\nu}$ are constant everywhere.

The second claim, that $R^\rho{}_{\sigma\mu\nu} = 0$ everywhere implies we can find a coordinate system in which the metric components are constant everywhere, is harder to prove (but not very hard). Consider as a warm-up a one-form $\omega = \omega_\mu \mathrm{d}x^\mu$, defined at some point p. For any path $x^\mu(\lambda)$ that includes p, we can construct a unique one-form field along the path by demanding that ω_μ be parallel-transported:

$$\frac{dx^\mu}{d\lambda}\nabla_\mu \omega_\nu = 0. \tag{3.118}$$

In general, if we performed this procedure for distinct paths that started at p and passed through some other point q, the value of ω_μ at q would depend on the path. However, if the Riemann tensor vanishes everywhere, the parallel-transport will be path-independent, and we can define a unique one-form field throughout the manifold. Therefore (3.118) must be true for arbitrary $dx^\mu/d\lambda$; this can only be true if ω_μ is covariantly constant:

$$\nabla_\mu \omega_\nu = 0. \tag{3.119}$$

On an arbitrary manifold there will be no solutions to this equation; it is only possible here because we are assuming that the curvature vanishes. We can take the antisymmetric part of (3.119), and from (3.36) we know that this is just the exterior derivative:

$$\nabla_{[\mu}\omega_{\nu]} = \partial_{[\mu}\omega_{\nu]} = 0, \tag{3.120}$$

or, in index-free notation,

$$\mathrm{d}\omega = 0. \tag{3.121}$$

In other words, ω is closed. It is also exact (there exists a scalar function α such that $\omega = \mathrm{d}\alpha$), since we have restricted the topology of the region in which we are working. In components we have

$$\omega_\mu = \partial_\mu \alpha. \tag{3.122}$$

There is nothing special about the one-form ω, so we can repeat this procedure with a set of one-forms $\hat{\theta}^{(a)}$, where $a \in \{1 \dots n\}$ on an n-dimensional manifold. We may choose our one-forms to comprise a normalized basis for the dual space T^*_p, such that the components of the metric with respect to this basis are those of the canonical form; in other words,

$$ds^2(p) = \eta_{ab}\, \hat{\theta}^{(a)} \otimes \hat{\theta}^{(b)}. \tag{3.123}$$

Here we use η_{ab} in a generalized sense, as a matrix with either $+1$ or -1 for each diagonal element and zeroes elsewhere. The actual arrangement of the $+1$'s and -1's depends on the canonical form of the metric, but is irrelevant for the present argument. Now let us parallel transport the entire set of basis forms all over the manifold; the vanishing of the Riemann tensor ensures that the result will be independent of the path taken. Since the metric is always automatically parallel-transported with respect to a metric-compatible connection, the metric components will remain unchanged,

$$ds^2(\text{anywhere}) = \eta_{ab}\, \hat{\theta}^{(a)} \otimes \hat{\theta}^{(b)}. \tag{3.124}$$

We therefore have specified a set of one-form fields, which everywhere define a basis in which the metric components are constant. This is completely unimpressive; it can be done on any manifold, regardless of what the curvature is. What we would like to show is that this is a *coordinate* basis (which will only be possible if the curvature vanishes). However, by the same arguments that led to (3.122), we know that all of the $\hat{\theta}^{(a)}$'s are exact forms, so that there exists a set of functions y^a such that the one-form fields are their gradients,

$$\hat{\theta}^{(a)} = \mathrm{d}y^a. \tag{3.125}$$

These n functions are precisely the sought-after coordinates; all over the manifold the metric takes the form

$$ds^2 = \eta_{ab}\, \mathrm{d}y^a \mathrm{d}y^b. \tag{3.126}$$

At this point you are welcome to switch from using a, b as indices to μ, ν if you prefer.

We have thus verified that the Riemann tensor provides us with an answer to the question of whether some horrible-looking metric is secretly that of flat space in a perverse coordinate system. If we calculate the Riemann tensor of such a metric and find that it vanishes, we know that the metric is flat; if it doesn't vanish, there is curvature.

3.7 ■ PROPERTIES OF THE RIEMANN TENSOR

The Riemann tensor, with four indices, naively has n^4 independent components in an n-dimensional space. In fact the antisymmetry property (3.110) means that there are only $n(n-1)/2$ independent values these last two indices can take on, leaving us with $n^3(n-1)/2$ independent components. When we consider the Christoffel connection, however, a number of other symmetries reduce the independent components further. Let's consider these now.

The simplest way to derive these additional symmetries is to examine the Riemann tensor with all lower indices,

$$R_{\rho\sigma\mu\nu} = g_{\rho\lambda} R^{\lambda}{}_{\sigma\mu\nu}. \tag{3.127}$$

Let us further consider the components of this tensor in locally inertial coordinates $x^{\hat{\mu}}$ established at a point p. Then the Christoffel symbols themselves will vanish, although their derivatives will not. We therefore have

$$\begin{aligned} R_{\hat{\rho}\hat{\sigma}\hat{\mu}\hat{\nu}}(p) &= g_{\hat{\rho}\hat{\lambda}}(\partial_{\hat{\mu}}\Gamma^{\hat{\lambda}}_{\hat{\nu}\hat{\sigma}} - \partial_{\hat{\nu}}\Gamma^{\hat{\lambda}}_{\hat{\mu}\hat{\sigma}}) \\ &= \tfrac{1}{2} g_{\hat{\rho}\hat{\lambda}} g^{\hat{\lambda}\hat{\tau}}(\partial_{\hat{\mu}}\partial_{\hat{\nu}} g_{\hat{\sigma}\hat{\tau}} + \partial_{\hat{\mu}}\partial_{\hat{\sigma}} g_{\hat{\tau}\hat{\nu}} - \partial_{\hat{\mu}}\partial_{\hat{\tau}} g_{\hat{\nu}\hat{\sigma}} - \partial_{\hat{\nu}}\partial_{\hat{\mu}} g_{\hat{\sigma}\hat{\tau}} \\ &\qquad - \partial_{\hat{\nu}}\partial_{\hat{\sigma}} g_{\hat{\tau}\hat{\mu}} + \partial_{\hat{\nu}}\partial_{\hat{\tau}} g_{\hat{\mu}\hat{\sigma}}) \\ &= \tfrac{1}{2}(\partial_{\hat{\mu}}\partial_{\hat{\sigma}} g_{\hat{\rho}\hat{\nu}} - \partial_{\hat{\mu}}\partial_{\hat{\rho}} g_{\hat{\nu}\hat{\sigma}} - \partial_{\hat{\nu}}\partial_{\hat{\sigma}} g_{\hat{\rho}\hat{\mu}} + \partial_{\hat{\nu}}\partial_{\hat{\rho}} g_{\hat{\mu}\hat{\sigma}}). \end{aligned} \tag{3.128}$$

In the first line we have used $\Gamma^{\hat{\tau}}_{\hat{\mu}\hat{\nu}}(p) = 0$, in the second line we have used $\partial_{\hat{\mu}} g^{\hat{\lambda}\hat{\tau}} = 0$ in Riemann normal coordinates, and the fact that partials commute in the third line. From this expression we can notice immediately three properties of $R_{\rho\sigma\mu\nu}$: it is antisymmetric in its first two indices,

$$\boxed{R_{\rho\sigma\mu\nu} = -R_{\sigma\rho\mu\nu},} \tag{3.129}$$

it is antisymmetric in its last two indices [which we already knew from (3.110)],

$$\boxed{R_{\rho\sigma\mu\nu} = -R_{\rho\sigma\nu\mu},} \tag{3.130}$$

and it is invariant under interchange of the first pair of indices with the second:

$$\boxed{R_{\rho\sigma\mu\nu} = R_{\mu\nu\rho\sigma}.} \tag{3.131}$$

With a little more work, which is left to your imagination, we can see that the sum of cyclic permutations of the last three indices vanishes:

$$\boxed{R_{\rho\sigma\mu\nu} + R_{\rho\mu\nu\sigma} + R_{\rho\nu\sigma\mu} = 0.} \tag{3.132}$$

Given (3.130), it's easy to see that this last property is equivalent to the vanishing of the antisymmetric part of the last three indices:

$$\boxed{R_{\rho[\sigma\mu\nu]} = 0.} \tag{3.133}$$

All of these properties have been derived in a special coordinate system, but they are all tensor equations; therefore they will be true in any coordinates (so we haven't bothered with hats on the indices). Not all of them are independent; with some effort, you can show that (3.129), (3.130), and (3.133) together imply (3.131). The logical interdependence of the equations is usually less important than the fact that they are true.

Given these relationships between the different components of the Riemann tensor, how many independent quantities remain? Let's begin with the facts that $R_{\rho\sigma\mu\nu}$ is antisymmetric in the first two indices, antisymmetric in the last two indices, and symmetric under interchange of these two pairs. This means that we can think of it as a symmetric matrix $R_{[\rho\sigma][\mu\nu]}$, where the pairs $\rho\sigma$ and $\mu\nu$ are thought of as individual indices. An $m \times m$ symmetric matrix has $m(m+1)/2$ independent components, while an $n \times n$ antisymmetric matrix has $n(n-1)/2$ independent components. We therefore have

$$\frac{1}{2}\left[\frac{1}{2}n(n-1)\right]\left[\frac{1}{2}n(n-1)+1\right] = \frac{1}{8}(n^4 - 2n^3 + 3n^2 - 2n) \tag{3.134}$$

independent components. We still have to deal with the additional symmetry (3.133). An immediate consequence of (3.133) is that the totally antisymmetric part of the Riemann tensor vanishes,

$$R_{[\rho\sigma\mu\nu]} = 0. \tag{3.135}$$

In fact, this equation plus the other symmetries (3.129), (3.130), and (3.131), are enough to imply (3.133), as can be easily shown by expanding (3.135) and manipulating the resulting terms. Therefore imposing the additional constraint of (3.135) is equivalent to imposing (3.133), once the other symmetries have been accounted for. How many independent restrictions does this represent? Let us imagine decomposing

$$R_{\rho\sigma\mu\nu} = X_{\rho\sigma\mu\nu} + R_{[\rho\sigma\mu\nu]}. \tag{3.136}$$

It is easy to see that any totally antisymmetric 4-index tensor is automatically antisymmetric in its first and last indices, and symmetric under interchange of the two pairs. Therefore these properties are independent restrictions on $X_{\rho\sigma\mu\nu}$, unrelated to the requirement (3.135). Now a totally antisymmetric 4-index tensor has $n(n-1)(n-2)(n-3)/4!$ terms, and therefore (3.135) reduces the number of independent components by this amount. We are left with

$$\tfrac{1}{8}(n^4 - 2n^3 + 3n^2 - 2n) - \tfrac{1}{24}n(n-1)(n-2)(n-3) = \tfrac{1}{12}n^2(n^2-1) \tag{3.137}$$

independent components of the Riemann tensor.

In four dimensions, therefore, the Riemann tensor has 20 independent components. (In one dimension it has none.) These twenty functions are precisely the 20 degrees of freedom in the second derivatives of the metric that we could not set to zero by a clever choice of coordinates when we first discussed locally inertial coordinates in Chapter 2. This should reinforce your confidence that the Riemann tensor is an appropriate measure of curvature.

In addition to the algebraic symmetries of the Riemann tensor (which constrain the number of independent components at any point), it obeys a differential identity, which constrains its relative values at different points. Consider the covariant derivative of the Riemann tensor, evaluated in locally inertial coordinates:

$$\begin{aligned}\nabla_{\hat\lambda} R_{\hat\rho\hat\sigma\hat\mu\hat\nu} &= \partial_{\hat\lambda} R_{\hat\rho\hat\sigma\hat\mu\hat\nu} \\ &= \tfrac{1}{2}\partial_{\hat\lambda}(\partial_{\hat\mu}\partial_{\hat\sigma} g_{\hat\rho\hat\nu} - \partial_{\hat\mu}\partial_{\hat\rho} g_{\hat\nu\hat\sigma} - \partial_{\hat\nu}\partial_{\hat\sigma} g_{\hat\rho\hat\mu} + \partial_{\hat\nu}\partial_{\hat\rho} g_{\hat\mu\hat\sigma}).\end{aligned} \tag{3.138}$$

It may seem illegitimate to take the derivative of an expression that is only true at a point, but the terms we are neglecting are all proportional to $\partial_{\hat\sigma} g_{\hat\mu\hat\nu}$, and therefore vanish. We would like to consider the sum of cyclic permutations of the first three indices:

$$\begin{aligned}&\nabla_{\hat\lambda} R_{\hat\rho\hat\sigma\hat\mu\hat\nu} + \nabla_{\hat\rho} R_{\hat\sigma\hat\lambda\hat\mu\hat\nu} + \nabla_{\hat\sigma} R_{\hat\lambda\hat\rho\hat\mu\hat\nu} \\ &\quad = \tfrac{1}{2}(\partial_{\hat\lambda}\partial_{\hat\mu}\partial_{\hat\sigma} g_{\hat\rho\hat\nu} - \partial_{\hat\lambda}\partial_{\hat\mu}\partial_{\hat\rho} g_{\hat\nu\hat\sigma} - \partial_{\hat\lambda}\partial_{\hat\nu}\partial_{\sigma} g_{\hat\rho\hat\mu} + \partial_{\hat\lambda}\partial_{\hat\nu}\partial_{\hat\rho} g_{\hat\mu\hat\sigma} \\ &\qquad + \partial_{\hat\rho}\partial_{\hat\mu}\partial_{\hat\lambda} g_{\hat\sigma\hat\nu} - \partial_{\hat\rho}\partial_{\hat\mu}\partial_{\hat\sigma} g_{\hat\nu\hat\lambda} - \partial_{\hat\rho}\partial_{\hat\nu}\partial_{\hat\lambda} g_{\hat\sigma\hat\mu} + \partial_{\hat\rho}\partial_{\hat\nu}\partial_{\hat\sigma} g_{\hat\mu\hat\lambda} \\ &\qquad + \partial_{\hat\sigma}\partial_{\hat\mu}\partial_{\hat\rho} g_{\hat\lambda\hat\nu} - \partial_{\hat\sigma}\partial_{\hat\mu}\partial_{\hat\lambda} g_{\hat\nu\hat\rho} - \partial_{\hat\sigma}\partial_{\hat\nu}\partial_{\hat\rho} g_{\hat\lambda\hat\mu} + \partial_{\hat\sigma}\partial_{\hat\nu}\partial_{\hat\lambda} g_{\hat\mu\hat\rho}) \\ &\quad = 0.\end{aligned} \tag{3.139}$$

Once again, since this is an equation between tensors it is true in any coordinate system, even though we derived it in a particular one. We recognize by now that the antisymmetry $R_{\rho\sigma\mu\nu} = -R_{\sigma\rho\mu\nu}$ allows us to write this result as

$$\nabla_{[\lambda} R_{\rho\sigma]\mu\nu} = 0. \tag{3.140}$$

This is known as the **Bianchi identity**. For a general connection there would be additional terms involving the torsion tensor. It is closely related to the Jacobi identity, since (recalling the definition of the Riemann tensor in terms of the commutator of covariant derivatives) it expresses

$$[[\nabla_\lambda, \nabla_\rho], \nabla_\sigma] + [[\nabla_\rho, \nabla_\sigma], \nabla_\lambda] + [[\nabla_\sigma, \nabla_\lambda], \nabla_\rho] = 0. \tag{3.141}$$

The Riemann tensor has four indices. At times it is useful to express a tensor as a sum of various pieces that are individually easier to handle and may have direct physical interpretations. The trick is to do this in a coordinate-invariant way. For example, we could decompose the Riemann tensor into $R^\rho{}_{\sigma ij}$ and $R^\rho{}_{\sigma i0}$, from which we could reconstruct the entire tensor (since $R^\rho{}_{\sigma 00}$ vanishes). But clearly this decomposition is not invariant under change of basis; we want to find a decomposition that is preserved when we change coordinates. What we are really doing is considering representations of the Lorentz group. We have two fundamental tricks at our disposal: taking contractions, and taking symmetric/antisymmetric parts. For example, given an arbitrary $(0, 2)$ tensor $X_{\mu\nu}$, we can decompose it into its symmetric and antisymmetric pieces,

$$X_{\mu\nu} = X_{(\mu\nu)} + X_{[\mu\nu]}, \tag{3.142}$$

and the symmetric part can be further decomposed into its trace $X = g^{\mu\nu} X_{(\mu\nu)}$ and a trace-free part $\widehat{X}_{\mu\nu} = X_{(\mu\nu)} - \frac{1}{n} X g_{\mu\nu}$, so that

$$X_{\mu\nu} = \frac{1}{n} X g_{\mu\nu} + \widehat{X}_{\mu\nu} + X_{[\mu\nu]}. \tag{3.143}$$

(Note that $X_{[\mu\nu]}$ is automatically traceless.) When we change coordinates, the different pieces $X g_{\mu\nu}$, $\widehat{X}_{\mu\nu}$, and $X_{[\mu\nu]}$ are rotated into themselves, not into each other; we say that they define "invariant subspaces" of the space of $(0, 2)$ tensors. For more complicated tensors the equivalent decomposition might not be so simple.

For the Riemann tensor, our first step is to take a contraction to form the **Ricci tensor**:

$$\boxed{R_{\mu\nu} = R^\lambda{}_{\mu\lambda\nu}.} \tag{3.144}$$

For the curvature tensor formed from an arbitrary (not necessarily Christoffel) connection, there are a number of independent contractions to take. Our primary concern is with the Christoffel connection, for which (3.144) is the only independent contraction; all others either vanish, or are related to this one. The Ricci tensor associated with the Christoffel connection is automatically symmetric,

$$\boxed{R_{\mu\nu} = R_{\nu\mu},} \tag{3.145}$$

as a consequence of the symmetries of the Riemann tensor. The trace of the Ricci tensor is the **Ricci scalar** (or **curvature scalar**):

$$\boxed{R = R^\mu{}_\mu = g^{\mu\nu} R_{\mu\nu}.} \tag{3.146}$$

We could also form the trace-free part $\widehat{R}_{\mu\nu} = R_{\mu\nu} - \frac{1}{n} R g_{\mu\nu}$, but this turns out not to be especially useful; it is more common to express things in terms of $R_{\mu\nu}$ and R.

The Ricci tensor and scalar contain all of the information about traces of the Riemann tensor, leaving us the trace-free parts. These are captured by the **Weyl tensor**, which is basically the Riemann tensor with all of its contractions removed. It is given in n dimensions by

$$\begin{aligned} C_{\rho\sigma\mu\nu} &= R_{\rho\sigma\mu\nu} - \frac{2}{(n-2)} \left(g_{\rho[\mu} R_{\nu]\sigma} - g_{\sigma[\mu} R_{\nu]\rho} \right) \\ &\quad + \frac{2}{(n-1)(n-2)} g_{\rho[\mu} g_{\nu]\sigma} R \end{aligned} \tag{3.147}$$

This messy formula is designed so that all possible contractions of $C_{\rho\sigma\mu\nu}$ vanish, while it retains the symmetries of the Riemann tensor:

$$\begin{aligned} C_{\rho\sigma\mu\nu} &= C_{[\rho\sigma][\mu\nu]}, \\ C_{\rho\sigma\mu\nu} &= C_{\mu\nu\rho\sigma}, \\ C_{\rho[\sigma\mu\nu]} &= 0. \end{aligned} \tag{3.148}$$

The Weyl tensor is only defined in three or more dimensions, and in three dimensions it vanishes identically. One of the most important properties of the Weyl tensor is that it is invariant under conformal transformations (discussed in the Appendices). This means that if you compute $C^\rho{}_{\sigma\mu\nu}$ (note that the first index is upstairs) for some metric $g_{\mu\nu}$, and then compute it again for a metric given by $\omega^2(x) g_{\mu\nu}$, where $\omega(x)$ is an arbitrary nonvanishing function of spacetime, you get the same answer. For this reason it is often known as the *conformal tensor*.

An especially useful form of the Bianchi identity comes from contracting twice on (3.139):

$$\begin{aligned} 0 &= g^{\nu\sigma} g^{\mu\lambda} (\nabla_\lambda R_{\rho\sigma\mu\nu} + \nabla_\rho R_{\sigma\lambda\mu\nu} + \nabla_\sigma R_{\lambda\rho\mu\nu}) \\ &= \nabla^\mu R_{\rho\mu} - \nabla_\rho R + \nabla^\nu R_{\rho\nu}, \end{aligned} \tag{3.149}$$

or

$$\nabla^\mu R_{\rho\mu} = \tfrac{1}{2} \nabla_\rho R. \tag{3.150}$$

Notice that, unlike the partial derivative, it makes sense to raise an index on the covariant derivative, due to metric compatibility. We define the **Einstein tensor** as

$$\boxed{G_{\mu\nu} = R_{\mu\nu} - \tfrac{1}{2} R g_{\mu\nu}.} \tag{3.151}$$

In four dimensions the Einstein tensor can be thought of as a trace-reversed version of the Ricci tensor. We then see that the twice-contracted Bianchi identity (3.150) is equivalent to

$$\boxed{\nabla^\mu G_{\mu\nu} = 0.} \tag{3.152}$$

The Einstein tensor, which is symmetric due to the symmetry of the Ricci tensor and the metric, will be of great importance in general relativity.

We should pause at this point to contrast the formalism we have developed with our intuitive notion of curvature. Our intuition is unfortunately contaminated by the fact that we are used to thinking about one- and two-dimensional spaces embedded in the (almost) Euclidean space in which we live. We think, for example, of a straight line as having no curvature, while a circle (S^1) is curved. However, according to (3.137), in one, two, three, and four dimensions there are 0, 1, 6 and 20 independent components of the Riemann tensor, respectively. (Everything we say about the curvature in these examples refers to the curvature associated with the Christoffel connection, and therefore the metric.) Therefore it is impossible for a one-dimensional space such as S^1 to have any curvature as we have defined it. The apparent contradiction stems from the fact that our intuitive notion of curvature depends on the extrinsic geometry of the manifold, which characterizes how a space is embedded in some larger space, while the Riemann curvature is a property of the intrinsic geometry of a space, which could be measured by observers confined to the manifold. Beings that lived on a circle and had no access to the larger world would necessarily think that they lived in a flat geometry—for example, there is no possibility of a nondegenerate infinitesimal loop around which we could parallel-transport a vector and have it come back rotated from its original position. Extrinsic curvature, discussed in Appendix D, is occasionally useful in GR when we wish to describe submanifolds of spacetime; but most often we are interested in the intrinsic geometry of spacetime itself, which does not rely on any embeddings.

We can illustrate the intrinsic/extrinsic difference further with an example from two dimensions, where the curvature has one independent component. In fact, all of the information about the curvature is contained in the single component of the Ricci scalar. Consider a torus, portrayed in Figure 3.7, which can be thought of as a square region of the plane with opposite sides identified (topologically, $S^1 \times S^1$). Although a torus embedded in three dimensions looks curved from our point of view, it should be clear that we can put a metric on the torus whose components are constant in an appropriate coordinate system—simply unroll it and use the Euclidean metric of the plane, $ds^2 = dx^2 + dy^2$. In this metric, the torus is flat.

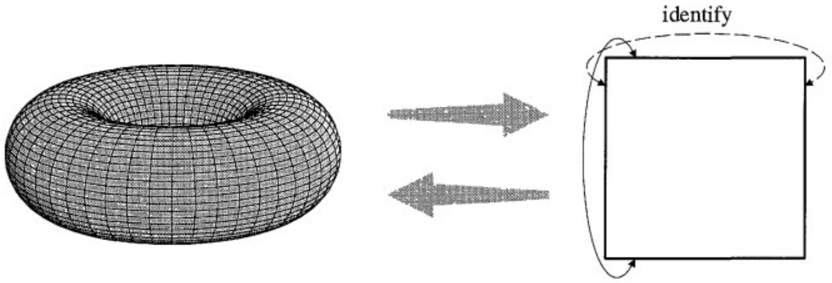

FIGURE 3.7 A torus thought of as a square in flat space with opposite sides identified.

There is also nothing to stop us from introducing a different metric in which the torus is not flat, but the point we are trying to emphasize is that it can be made flat in some metric. Every time we embed a manifold in a larger space, the manifold inherits an "induced metric" from the background in which it is embedded, as discussed in the Appendix A. Our point here is that a torus embedded in a flat three-dimensional Euclidean space will have an induced metric that is curved, but we can nevertheless choose to put a different metric on it so that the intrinsic geometry is flat.

Let's turn to a simple example where the curvature does not vanish. We have already talked about the two-sphere S^2, with metric

$$ds^2 = a^2(d\theta^2 + \sin^2\theta\, d\phi^2), \tag{3.153}$$

where a is the radius of the sphere. It will actually be the radius if our sphere is embedded in $\mathbf{R}^3$, but we can call it the radius even in the absence of any embedding. Two-dimensional people living on the sphere could calculate a by measuring the area of the sphere, dividing by 4π, and taking the square root; using the word "radius" to refer to this quantity is merely a convenience. We should also point out that the notion of a sphere is sometimes used in the weaker topological sense, without any particular metric being assumed; the metric we are using is called the *round metric*. Without going through the details, the nonzero connection coefficients for (3.153) are

$$\begin{aligned} \Gamma^\theta_{\phi\phi} &= -\sin\theta\cos\theta \\ \Gamma^\phi_{\theta\phi} = \Gamma^\phi_{\phi\theta} &= \cot\theta. \end{aligned} \tag{3.154}$$

Let's compute a promising component of the Riemann tensor:

$$\begin{aligned} R^\theta{}_{\phi\theta\phi} &= \partial_\theta\Gamma^\theta_{\phi\phi} - \partial_\phi\Gamma^\theta_{\theta\phi} + \Gamma^\theta_{\theta\lambda}\Gamma^\lambda_{\phi\phi} - \Gamma^\theta_{\phi\lambda}\Gamma^\lambda_{\theta\phi} \\ &= (\sin^2\theta - \cos^2\theta) - (0) + (0) - (-\sin\theta\cos\theta)(\cot\theta) \\ &= \sin^2\theta. \end{aligned} \tag{3.155}$$

The notation is obviously imperfect, since the Greek letter λ is a dummy index that is summed over, while the Greek letters θ and ϕ represent specific coordinates. Lowering an index, we have

$$\begin{aligned} R_{\theta\phi\theta\phi} &= g_{\theta\lambda} R^{\lambda}{}_{\phi\theta\phi} \\ &= g_{\theta\theta} R^{\theta}{}_{\phi\theta\phi} \\ &= a^2 \sin^2\theta . \end{aligned} \tag{3.156}$$

It is easy to check that all of the components of the Riemann tensor either vanish or are related to this one by symmetry. We can go on to compute the Ricci tensor via $R_{\mu\nu} = g^{\alpha\beta} R_{\alpha\mu\beta\nu}$. We obtain

$$\begin{aligned} R_{\theta\theta} &= g^{\phi\phi} R_{\phi\theta\phi\theta} = 1 \\ R_{\theta\phi} &= R_{\phi\theta} = 0 \\ R_{\phi\phi} &= g^{\theta\theta} R_{\theta\phi\theta\phi} = \sin^2\theta . \end{aligned} \tag{3.157}$$

The Ricci scalar is similarly straightforward:

$$R = g^{\theta\theta} R_{\theta\theta} + g^{\phi\phi} R_{\phi\phi} = \frac{2}{a^2}. \tag{3.158}$$

Therefore the Ricci scalar, which for a two-dimensional manifold completely characterizes the curvature, is a constant over the two-sphere. If we had perturbed the metric (corresponding physically to bumps on the sphere), this would no longer have been the case. Note that the scalar curvature decreases as the radius of the sphere increases. Even in more general contexts, we will sometimes refer to the "radius of curvature" of a manifold as providing a length scale over which the curvature varies; the larger the radius of curvature, the smaller the curvature itself.

3.8 ■ SYMMETRIES AND KILLING VECTORS

The real world is a messy place, and we have no hope of finding a metric that describes our actual universe, or even any small part thereof, with perfect precision. Instead, we model spacetime via various approximations appropriate to the physical situation being studied. For example, the geometry outside a star or planet may be approximated, to some order of precision, as being spherically symmetric, even if the real situation includes small deviations from this symmetry—these may be added in later as perturbations.

General relativity is no different from other fields of physics, then, in being especially interested in solutions with symmetry. In fact, such properties may be even more crucial in GR than in, say, electromagnetism, since the nonlinear nature of Einstein's equation (discussed in the next chapter) makes it hard to find any exact solutions at all. In the context of curved spacetime, however, we need to be more careful than usual about what exactly is meant by "symmetry." In this section we develop some useful tools for studying symmetry; a deeper investigation can be found in Appendix B.

We think of a manifold M as possessing a symmetry if the geometry is invariant under a certain transformation that maps M to itself; that is, if the metric is the same, in some sense, from one point to another. In fact different tensor fields may possess different symmetries; symmetries of the metric are called **isometries**. Sometimes the existence of isometries is obvious; consider, for example, four-dimensional Minkowski space,

$$ds^2 = \eta_{\mu\nu} dx^\mu dx^\nu = -dt^2 + dx^2 + dy^2 + dz^2. \tag{3.159}$$

We know of several isometries of this space; these include translations ($x^\mu \to x^\mu + a^\mu$, with a^μ fixed) and Lorentz transformations ($x^\mu \to \Lambda^\mu{}_\nu x^\nu$, with $\Lambda^\mu{}_\nu$ a Lorentz-transformation matrix). The fact that the metric is invariant under translations is made immediately apparent by the simple fact that the metric coefficients $\eta_{\mu\nu}$ are independent of the individual coordinate functions x^μ. Indeed, whenever $\partial_{\sigma_*} g_{\mu\nu} = 0$ for some fixed σ_* (but for all μ and ν), there will be a symmetry under translations along x^{σ_*}:

$$\partial_{\sigma_*} g_{\mu\nu} = 0 \quad \Rightarrow \quad x^{\sigma_*} \to x^{\sigma_*} + a^{\sigma_*} \text{ is a symmetry.} \tag{3.160}$$

The careful reader will have noticed that we still haven't precisely defined what we mean by symmetry; roughly we imagine that the metric is invariant under some transformation, but the precise meaning is only developed in Appendix B. Also, the implication arrow in (3.160) only goes one way, and it would be nice to have a clean criterion for deciding when a given transformation counts as a symmetry; this will come soon.

Isometries of the form (3.160) have immediate consequences for the motion of test particles as described by the geodesic equation. Recall from (3.61) that the geodesic equation can be written in terms of the four-momentum $p^\mu = mU^\mu$ (valid for timelike paths, at least) as

$$p^\lambda \nabla_\lambda p^\mu = 0. \tag{3.161}$$

By metric compatibility we are free to lower the index μ, and then we may expand the covariant derivative to obtain

$$p^\lambda \partial_\lambda p_\mu - \Gamma^\sigma_{\lambda\mu} p^\lambda p_\sigma = 0. \tag{3.162}$$

The first term tells us how the momentum components change along the path,

$$p^\lambda \partial_\lambda p_\mu = m \frac{dx^\lambda}{d\tau} \partial_\lambda p_\mu = m \frac{dp_\mu}{d\tau}, \tag{3.163}$$

while the second term is

$$\Gamma^{\sigma}_{\lambda\mu} p^{\lambda} p_{\sigma} = \tfrac{1}{2} g^{\sigma\nu} (\partial_{\lambda} g_{\mu\nu} + \partial_{\mu} g_{\nu\lambda} - \partial_{\nu} g_{\lambda\mu}) p^{\lambda} p_{\sigma} \tag{3.164}$$

$$= \tfrac{1}{2} (\partial_{\lambda} g_{\mu\nu} + \partial_{\mu} g_{\nu\lambda} - \partial_{\nu} g_{\lambda\mu}) p^{\lambda} p^{\nu} \tag{3.165}$$

$$= \tfrac{1}{2} (\partial_{\mu} g_{\nu\lambda}) p^{\lambda} p^{\nu}, \tag{3.166}$$

where we have used the symmetry of $p^{\lambda} p^{\nu}$ to go from the second line to the third. So, without yet making any assumptions about symmetry, we see that the geodesic equation can be written as

$$m \frac{dp_{\mu}}{d\tau} = \frac{1}{2} (\partial_{\mu} g_{\nu\lambda}) p^{\lambda} p^{\nu}. \tag{3.167}$$

Therefore, if all of the metric coefficients are independent of the coordinate x^{σ_*}, we find that this isometry implies that the momentum component p_{σ_*} is a conserved quantity of the motion:

$$\partial_{\sigma_*} g_{\mu\nu} = 0 \quad \Rightarrow \quad \frac{dp_{\sigma_*}}{d\tau} = 0. \tag{3.168}$$

This will hold along any geodesic, even though we only derived it for timelike ones. The conserved quantities implied by isometries are extremely useful in studying the motion of test particles in curved backgrounds.

Of course, even though independence of the metric components on one or more coordinates implies the existence of isometries, the converse does not necessarily hold. Symmetry under Lorentz transformations, for example, is not manifest as independence of $\eta_{\mu\nu}$ on any coordinates; indeed, in four dimensions, there are four types of translations and six types of Lorentz transformations, for a total of ten, which is obviously larger than the number of dimensions the metric could possibly be independent of. What is more, it would be simple enough to transform to a complicated coordinate system where not even the translational symmetries were obvious. Such a coordinate transformation would change the metric components, but not the underlying geometry, which is what the symmetry is really characterizing. Clearly a more systematic procedure is called for.

We can develop such a procedure by casting the right-hand equation of (3.168), expressing constancy of one of the components of the momentum, in a more manifestly covariant form. If x^{σ_*} is the coordinate which $g_{\mu\nu}$ is independent of, let us consider the vector ∂_{σ_*}, which we label as K:

$$K = \partial_{\sigma_*}, \tag{3.169}$$

which is equivalent in component notation to

$$K^{\mu} = (\partial_{\sigma_*})^{\mu} = \delta^{\mu}_{\sigma_*}. \tag{3.170}$$

We say that the vector K^{μ} generates the isometry; this means that the transformation under which the geometry is invariant is expressed infinitesimally as a motion in the direction of K^μ. Again, the notion is developed more fully in Appendix B. In terms of this vector, the noncovariant-looking quantity p_{σ_*} is simply

$$p_{\sigma_*} = K^\nu p_\nu = K_\nu p^\nu. \tag{3.171}$$

Meanwhile, the constancy of this (scalar) quantity along the path is equivalent to the statement that its directional derivative along the geodesic vanishes:

$$\frac{dp_{\sigma_*}}{d\tau} = 0 \quad \leftrightarrow \quad p^\mu \nabla_\mu (K_\nu p^\nu) = 0. \tag{3.172}$$

Expanding the expression on the right, we obtain

$$\begin{aligned} p^\mu \nabla_\mu (K_\nu p^\nu) &= p^\mu K_\nu \nabla_\mu p^\nu + p^\mu p^\nu \nabla_\mu K_\nu \\ &= p^\mu p^\nu \nabla_\mu K_\nu \\ &= p^\mu p^\nu \nabla_{(\mu} K_{\nu)}, \end{aligned} \tag{3.173}$$

where in the second line we have invoked the geodesic equation ($p^\mu \nabla_\mu p^\nu = 0$). In the third line we have used the fact that $p^\mu p^\nu$ is automatically symmetric in μ and ν, so only the symmetric part of $\nabla_\mu K_\nu$ could possibly contribute. We therefore conclude that any vector K_μ that satisfies $\nabla_{(\mu} K_{\nu)} = 0$ implies that $K_\nu p^\nu$ is conserved along a geodesic trajectory:

$$\boxed{\nabla_{(\mu} K_{\nu)} = 0 \quad \Rightarrow \quad p^\mu \nabla_\mu (K_\nu p^\nu) = 0.} \tag{3.174}$$

The equation on the left is known as **Killing's equation**, and vector fields that satisfy it are known as **Killing vector fields** (or simply Killing vectors). You can verify for yourself that, if the metric is independent of some coordinate x^{σ_*}, the vector ∂_{σ_*} will satisfy Killing's equation. In fact, if a vector K^μ satisfies Killing's equation, it will always be possible to find a coordinate system in which $K = \partial_{\sigma_*}$; but in general we cannot find coordinates in which all the Killing vectors are simultaneously of this form, nor is this form necessary for the vector to satisfy Killing's equation.

As we investigate in Appendix B, Killing vector fields on a manifold are in one-to-one correspondence with continuous symmetries of the metric on that manifold. Every Killing vector implies the existence of conserved quantities associated with geodesic motion. This can be understood physically: by definition the metric is unchanging along the direction of the Killing vector. Loosely speaking, therefore, a free particle will not feel any forces in this direction, and the component of its momentum in that direction will consequently be conserved. In fact, using the same kind of logic by which we showed that $K_\nu p^\nu$ is conserved along a geodesic if $\nabla_{(\mu} K_{\nu)} = 0$ generalizes to additional indices, a **Killing tensor** is a symmetric l-index tensor $K_{\nu_1 \cdots \nu_l}$ that satisfies the obvious generalization of Killing's equation, and correspondingly leads to conserved quantities by contracting with l copies of the momentum:

$$\nabla_{(\mu} K_{\nu_1 \cdots \nu_l)} = 0 \quad \Rightarrow \quad p^\mu \nabla_\mu (K_{\nu_1 \cdots \nu_l} p^{\nu_1} \cdots p^{\nu_l}) = 0. \tag{3.175}$$

Simple examples of Killing tensors are the metric itself, and symmetrized tensor products of Killing vectors. Killing tensors are not related in a simple way to symmetries of the spacetime, but they will simplify our analysis of rotating black holes and expanding universes.

Derivatives of Killing vectors can be related to the Riemann tensor by

$$\nabla_\mu \nabla_\sigma K^\rho = R^\rho{}_{\sigma\mu\nu} K^\nu, \tag{3.176}$$

as you are asked to prove in the exercises. Contracting this expression yields

$$\nabla_\mu \nabla_\sigma K^\mu = R_{\sigma\nu} K^\nu. \tag{3.177}$$

These relations, along with the Bianchi identity and Killing's equation, suffice to show that the directional derivative of the Ricci scalar along a Killing vector field will vanish,

$$K^\lambda \nabla_\lambda R = 0. \tag{3.178}$$

This last fact is another reflection of the idea that the geometry is not changing along a Killing vector field.

Besides leading to conserved quantities for the motion of individual particles, the existence of a timelike Killing vector allows us to define a conserved energy for the entire spacetime. Given a Killing vector K_ν and a conserved energy-momentum tensor $T_{\mu\nu}$, we can construct a current

$$J_T^\mu = K_\nu T^{\mu\nu} \tag{3.179}$$

that is automatically conserved,

$$\begin{aligned} \nabla_\mu J_T^\mu &= (\nabla_\mu K_\nu) T^{\mu\nu} + K_\nu (\nabla_\mu T^{\mu\nu}) \\ &= 0. \end{aligned} \tag{3.180}$$

The first term vanishes by virtue of Killing's equation (since the symmetry of the upper indices serves to automatically symmetrize the lower indices), and the second term vanishes by conservation of $T_{\mu\nu}$. If K_ν is timelike, we can integrate over a spacelike hypersurface Σ to define the total energy,

$$E_T = \int_\Sigma J_T^\mu n_\mu \sqrt{\gamma}\, d^3x, \tag{3.181}$$

where γ_{ij} is the induced metric on Σ and n_μ is the normal vector to Σ. In Appendix E we discuss integration over hypersurfaces, and in particular Stokes's theorem; as explained there, E_T will be the same when integrated over any spacelike hypersurface, and is therefore conserved. This result fits nicely with our discussion in Section 3.5, where we found that the total energy is not typically conserved in an expanding universe; expansion means that the metric is changing with time, so there is no isometry in this direction. When there is a timelike Killing vector, we can write the metric in a form where it is independent of the timelike coordinate, and Noether's theorem implies a conserved energy. Similarly, spacelike Killing vectors may be used to construct conserved momenta (or angular momenta).

Although it may or may not be simple to actually solve Killing's equation in any given spacetime, it is frequently possible to write down some Killing vectors by inspection. (Of course a generic metric has no Killing vectors at all, but to keep things simple we often deal with metrics with high degrees of symmetry.) For example, in $\mathbf{R}^3$ with metric $ds^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2$, independence of the metric components with respect to x, y, and z immediately yields three Killing vectors:

$$\begin{aligned} X^\mu &= (1, 0, 0) \\ Y^\mu &= (0, 1, 0) \\ Z^\mu &= (0, 0, 1). \end{aligned} \tag{3.182}$$

These clearly represent the three translations. There are also three rotational symmetries in $\mathbf{R}^3$, which are not quite as simple. To find them, imagine first going to polar coordinates,

$$\begin{aligned} x &= r \sin\theta \cos\phi \\ y &= r \sin\theta \sin\phi \\ z &= r \cos\theta, \end{aligned} \tag{3.183}$$

where the metric takes the form

$$ds^2 = \mathrm{d}r^2 + r^2 \mathrm{d}\theta^2 + r^2 \sin^2\theta \, \mathrm{d}\phi^2. \tag{3.184}$$

Now the metric (the *same* metric, just in a different coordinate system) is manifestly independent of ϕ. We therefore know that $R = \partial_\phi$ is a Killing vector. Transforming back to Cartesian coordinates, this becomes

$$R = -y\partial_x + x\partial_y. \tag{3.185}$$

The Cartesian components R^μ are therefore $(-y, x, 0)$. Since this represents a rotation about the z-axis, it is straightforward to guess the components of all three rotational Killing vectors:

$$\begin{aligned} R^\mu &= (-y,\ \ x,\ \ 0) \\ S^\mu &= (\ \ z,\ \ 0, -x) \\ T^\mu &= (\ \ 0, -z,\ \ y), \end{aligned} \tag{3.186}$$

representing rotations about the z, y, and x-axes, respectively. You can check for yourself that these actually do solve Killing's equation. The overall signs don't matter, since minus a Killing vector is still a Killing vector.

This exercise leads directly to the Killing vectors for the two-sphere S^2 with metric

$$ds^2 = d\theta^2 + \sin^2\theta\, d\phi^2. \tag{3.187}$$

Since this sphere can be thought of as the locus of points at unit distance from the origin in $\mathbf{R}^3$, and the rotational Killing vectors all rotate such a sphere into itself, they also represent symmetries of S^2. To get explicit coordinate-basis representations for these vectors, we first transform the three-dimensional vectors (3.186) to polar coordinates $x^{\mu'} = (r, \theta, \phi)$. A straightforward calculation reveals

$$\begin{aligned} R &= \partial_\phi \\ S &= \cos\phi\, \partial_\theta - \cot\theta \sin\phi\, \partial_\phi \\ T &= -\sin\phi\, \partial_\theta - \cot\theta \cos\phi\, \partial_\phi . \end{aligned} \tag{3.188}$$

Notice that there are no components along ∂_r, which makes sense for a rotational isometry. Therefore the expressions (3.188) for the three rotational Killing vectors in $\mathbf{R}^3$ are exactly the same as those of S^2 in spherical polar coordinates.

In $n \geq 2$ dimensions, there can be more Killing vectors than dimensions. This is because a set of Killing vector fields can be linearly independent, even though at any one point on the manifold the vectors at that point are linearly dependent. It is trivial to show (so you should do it yourself) that a linear combination of Killing vectors with *constant* coefficients is still a Killing vector (in which case the linear combination does not count as an independent Killing vector), but this is not generally true with coefficients that vary over the manifold. You can also show that the commutator of two Killing vector fields is a Killing vector field; this is very useful to know, but it may be the case that the commutator gives you a vector field that is not linearly independent (or it may simply vanish). The problem of finding all of the Killing vectors of a metric is therefore somewhat tricky, as it is not always clear when to stop looking.

3.9 ■ MAXIMALLY SYMMETRIC SPACES

How symmetric can a space possibly be? An example of a space with the highest possible degree of symmetry is $\mathbf{R}^n$ with the flat Euclidean metric. Consider the isometries of this space, which we know to be translations and rotations in n dimensions, from the perspective of what they do in the neighborhood of some fixed point p. The translations are those transformations that move the point; there are n independent axes along which it can be moved, and hence n total translations. The rotations, centered at p, are those transformations that leave p invariant; they can be thought of as moving one of the axes through p into one of the others. There are n axes, and for each axis there are $n-1$ other axes into which it can be rotated, but we shouldn't count a rotation of y into x as separate from a rotation of x into y, so the total number of independent rotations is $\frac{1}{2}n(n-1)$. We therefore have

$$n + \tfrac{1}{2}n(n-1) = \tfrac{1}{2}n(n+1) \tag{3.189}$$

independent symmetries of $\mathbf{R}^n$. But our counting argument only referred to the behavior of the symmetry in a neighborhood of p, not globally all over the manifold; so even in the presence of curvature the counting should be the same. If the metric signature is not Euclidean, some of the rotations will actually be boosts, but again the counting will be the same. The number of isometries is, of course, the number of linearly independent Killing vector fields. We therefore refer to an n-dimensional manifold with $\frac{1}{2}n(n+1)$ Killing vectors as a **maximally symmetric space**. The most familiar examples of maximally symmetric spaces are n-dimensional Euclidean spaces $\mathbf{R}^n$ and the n-dimensional spheres S^n. For an n-dimensional sphere we usually think of the isometries as consisting of $\frac{1}{2}n(n+1)$ independent rotations, rather than as some collection of both rotations and translations. However, if we consider the action of these rotations on some fixed point p, a moment's thought convinces us that the entire set can be decomposed into $\frac{1}{2}n(n-1)$ rotations around the point (keeping p fixed), and another n that move p along each direction, just as in $\mathbf{R}^n$.

If a manifold is maximally symmetric, the curvature is the same everywhere (as expressed by translation-like isometries) and the same in every direction (as expressed by rotation-like isometries). Hence, if we know the curvature of a maximally symmetric space at one point, we know it everywhere. Indeed, there are only a small number of possible maximally symmetric spaces; they are classified by the curvature scalar R (which will be constant everywhere), the dimensionality n, the metric signature, and perhaps some discrete pieces of information relating to the global topology (distinguishing, for example, an n-torus from $\mathbf{R}^n$, and tori of different sizes from each other). It follows that we should be able to reconstruct the entire Riemann tensor of such a space from the Ricci scalar R; let's see how this works.

The basic idea is simply that, since the geometry looks the same in all directions, the curvature tensor should look the same in all directions. What might this mean? First choose locally inertial coordinates at some point p, so that $g_{\hat{\mu}\hat{\nu}} = \eta_{\hat{\mu}\hat{\nu}}$. Of course, locally inertial coordinates are not unique; for example, we can perform a Lorentz transformation at p and the metric components will remain those of $\eta_{\hat{\mu}\hat{\nu}}$. (By "doing a Lorentz transformation" we really are referring to a change of basis vectors in T_p; in a curved spacetime, this only makes sense at a single point, not over a region.) Since the geometry is maximally symmetric, we want the same to be true of the Riemann tensor; that is, the components of $R_{\hat{\rho}\hat{\sigma}\hat{\mu}\hat{\nu}}$ should not change under a Lorentz transformation either, since there is no preferred direction in spacetime. But there are unique tensors that do not change their components under Lorentz transformations—the metric, the Kronecker delta, and the Levi–Civita tensor. This means that, in these coordinates and at this point, the components of $R_{\hat{\rho}\hat{\sigma}\hat{\mu}\hat{\nu}}$ will be proportional to a tensor constructed from these invariant tensors. Attempting to match the symmetries of the Riemann tensor reveals that there is a unique possibility:

$$R_{\hat{\rho}\hat{\sigma}\hat{\mu}\hat{\nu}} \propto g_{\hat{\rho}\hat{\mu}}g_{\hat{\sigma}\hat{\nu}} - g_{\hat{\rho}\hat{\nu}}g_{\hat{\sigma}\hat{\mu}}. \tag{3.190}$$

But this is a completely tensorial relation, so it must be true in any coordinate system. We have argued in favor of this relation at a single point p, but in a maximally symmetric space all points are created equal, so it must also be true at any other point as well. The proportionality constant is easily fixed by contracting both sides twice [the left-hand side becomes R, and the right-hand side is $n(n-1)$]. We end up with an equation true in any maximally symmetric space, at any point, in any coordinate system:

$$R_{\rho\sigma\mu\nu} = \frac{R}{n(n-1)}(g_{\rho\mu}g_{\sigma\nu} - g_{\rho\nu}g_{\sigma\mu}). \tag{3.191}$$

Likewise, if the Riemann tensor satisfies this condition (with R a constant over the manifold), the metric will be maximally symmetric. In two dimensions, finding that R is a constant is sufficient to prove that a space is maximally symmetric, since there is only one independent component of the curvature. In higher dimensions you have to work harder.

Locally, then (ignoring questions of global topology), a maximally symmetric space of given dimension and signature is fully specified by R. The basic classification of such spaces is simply whether R is positive, zero, or negative, since the magnitude of R represents an overall scaling of the size of the space. For Euclidean signatures, the flat maximally symmetric spaces are planes or appropriate higher-dimensional generalizations, while the positively curved ones are spheres. Maximally symmetric Euclidean spaces of negative curvature are hyperboloids, denoted H^n. These are less familiar because even a two-dimensional hyperboloid cannot be isometrically embedded in $\mathbf{R}^3$. Let's examine this two-dimensional hyperboloid briefly.

There are a number of ways of representing H^2, which has the same topology as $\mathbf{R}^2$. One simple way is as the **Poincaré half-plane**, which is the region $y > 0$ of a two-dimensional region with coordinates $\{x, y\}$ and metric

$$ds^2 = \frac{a^2}{y^2}(dx^2 + dy^2). \tag{3.192}$$

The geometry of the Poincaré half-plane is of course different from that of the upper half of $\mathbf{R}^2$, despite the use of similar coordinates. For example, we can compute the length of a line segment stretching vertically ($x =$ constant) from y_1 to y_2:

$$\begin{aligned}\Delta s &= \int_{y_1}^{y_2} \sqrt{g_{\mu\nu}\frac{dx^\mu}{dy}\frac{dx^\nu}{dy}}\,dy \\ &= a\int_{y_1}^{y_2}\frac{dy}{y} \\ &= a\ln\left(\frac{y_2}{y_1}\right).\end{aligned} \tag{3.193}$$

This is not at all the result $\Delta s = y_2 - y_1$ we would expect in Euclidean space. In particular, notice that the path length becomes infinite for paths that approach the boundary $y = 0$. In other words, it's not really a boundary at all; it's infinitely far away, as far as anyone living on the hyperboloid is concerned.

The nonvanishing Christoffel symbols for (3.192) are

$$\begin{aligned}\Gamma^x_{xy} = \Gamma^x_{yx} &= -y^{-1} \\ \Gamma^y_{xx} &= y^{-1} \\ \Gamma^y_{yy} &= -y^{-1}.\end{aligned} \tag{3.194}$$

From these it is straightforward to show that geodesics satisfy

$$(x - x_0)^2 + y^2 = l^2, \tag{3.195}$$

for some constants x_0 and l. Curves of this form are semicircles with centers located on the x-axis, as shown in Figure 3.8. In the limit as $x_0 \to \infty$ and $l \to \infty$ with $l - x_0$ fixed, we get a straight vertical line. Following our discussion of S^2 at the end of Section 3.7, we calculate a representative component of the Riemann tensor to be

$$R^x{}_{yxy} = -y^{-2}. \tag{3.196}$$

As with the two-sphere, all other components are either vanishing or related to this by symmetries. This is simply a reflection of the fact that we are in two di-

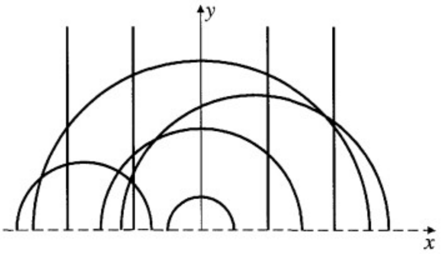

FIGURE 3.8 The upper half plane with a negatively curved metric. Geodesics are semicircles and straight lines that intersect the x-axis vertically.

mensions, with only one independent component of curvature. Turning the crank yields the Ricci tensor,

$$\begin{aligned} R_{xx} &= -y^{-2} \\ R_{xy} &= 0 \\ R_{yy} &= -y^{-2}, \end{aligned} \tag{3.197}$$

and the curvature scalar,

$$R = -\frac{2}{a^2}. \tag{3.198}$$

We see that it matches that of S^2 with the opposite sign, and in particular that it is a constant. Since we are in two dimensions, this is enough to ensure that our metric really is maximally symmetric. Of course there are coordinates in which H^2 looks very different; one is introduced in the Exercises.

Locally, then, a maximally symmetric space of Euclidean signature is either a plane, a sphere, or a hyperboloid, depending on the sign of R. Globally, any maximally symmetric space (of Euclidean signature) can be constructed by taking a carefully chosen region of one of these three spaces and identifying different sides, as the flat torus can be constructed from $\mathbf{R}^2$. As an aside, let's briefly mention a connection between local geometry and global topology, encompassed by the Gauss–Bonnet theorem. For a two-dimensional compact boundaryless orientable manifold, this reads

$$\chi(M) = \frac{1}{4\pi}\int_M R\sqrt{|g|}\, d^n x, \tag{3.199}$$

where $\chi(M)$ is a topological invariant of the space, known as the Euler characteristic. In general it can be calculated from the cohomology spaces mentioned in Chapter 2; in two dimensions, however, it is simply given by

$$\chi(M) = 2(1-g), \tag{3.200}$$

where g is the genus of the surface (zero for a sphere, and equal to the number of handles of a torus or Riemann surface). The Gauss–Bonnet theorem holds whether or not the curvature R is a constant; when it is, however, we see that all Riemann surfaces of genus $g \geq 2$ must have negative curvature, just as a sphere must be positively curved and a torus must ~~be~~ flat.

Continuing our aside, think for the moment about string theory, which claims that the fundamental objects comprising the universe are small one-dimensional loops of string. Such strings have two-dimensional "world-sheets" rather than one-dimensional worldlines. Doing perturbation theory in string theory (the equivalent of calculating Feynman diagrams in quantum field theory) involves summing over all world-sheet geometries (generally, for technical reasons, Euclidean geometries). This sounds like a lot of geometries, but in two dimensions any metric can be written as some fiducial metric times a conformal factor. This should be plausible, since there is only one curvature component; you are asked to prove it in the Exercises. The fiducial metric can be chosen differently for each world-sheet topology, and we can make our lives easier by choosing it to be (locally) a metric of maximal symmetry—the round sphere for genus zero, the plane for genus one, and the hyperboloid for higher genera. Even more fortunately, the string theories of greatest physical interest are the so-called critical string theories, for which the conformal factor itself doesn't matter. This is one of the things that makes doing calculations in perturbative string theory possible; we only have to sum over a discrete set of topologies, with a finite number of modular parameters for each topology (such as the parameters telling us the sizes of the different directions in a torus).

We close this section with one last point. We have explored the maximally symmetric spaces of Euclidean signatures; there are, of course, corresponding spacetimes with Lorentzian signatures. We know that the maximally symmetric spacetime with $R = 0$ is simply Minkowski space. The positively curved maximally symmetric spacetime is called de Sitter space, while that with negative curvature is imaginatively labeled anti-de Sitter space. These spacetimes will be more thoroughly discussed in Chapter 8.

It should be clear by now that the Appendices flesh out these ideas in important ways. Impatient readers may skip over them, but it would be a shame to do so.

3.10 ■ GEODESIC DEVIATION

The Riemann tensor shows up as a consequence of curvature in one more way: geodesic deviation. You have undoubtedly heard that the defining property of Euclidean (flat) geometry is the parallel postulate: initially parallel lines remain parallel forever. Of course in a curved space this is not true; on a sphere, certainly, initially parallel geodesics will eventually cross. We would like to quantify this behavior for an arbitrary curved space.

The problem is that the notion of "parallel" does not extend naturally from flat to curved spaces. The best we can do is to consider geodesic curves that might be initially parallel, and see how they behave as we travel down the geodesics. To this end we consider a one-parameter family of geodesics, $\gamma_s(t)$. That is, for each $s \in \mathbf{R}$, γ_s is a geodesic parameterized by the affine parameter t. The collection of these curves defines a smooth two-dimensional surface (embedded in a manifold M of arbitrary dimensionality). The coordinates on this surface may be chosen to be s and t, provided we have chosen a family of geodesics that do not cross. The entire surface is the set of points $x^\mu(s,t) \in M$. We have two natural vector fields: the tangent vectors to the geodesics,

$$T^\mu = \frac{\partial x^\mu}{\partial t}, \tag{3.201}$$

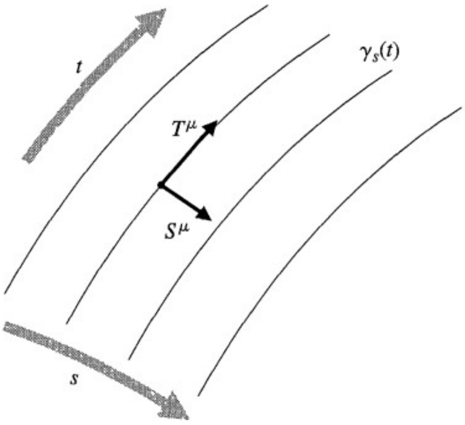

FIGURE 3.9 A set of geodesics $\gamma_s(t)$, with tangent vectors T^μ. The vector field S^μ measures the deviation between nearby geodesics.

and the deviation vectors

$$S^\mu = \frac{\partial x^\mu}{\partial s}. \tag{3.202}$$

This name derives from the informal notion that S^μ points from one geodesic toward the neighboring ones.

The idea that S^μ points from one geodesic to the next inspires us to define the "relative velocity of geodesics,"

$$V^\mu = (\nabla_T S)^\mu = T^\rho \nabla_\rho S^\mu, \tag{3.203}$$

and the "relative acceleration of geodesics,"

$$A^\mu = (\nabla_T V)^\mu = T^\rho \nabla_\rho V^\mu. \tag{3.204}$$

You should take the names with a grain of salt, but these vectors are certainly well-defined. This notion of *relative* acceleration between geodesics should be distinguished from the acceleration of a path away from being a geodesic, which would be given by $a^\mu = T^\sigma \nabla_\sigma T^\mu$.

Since S and T are basis vectors adapted to a coordinate system, their commutator vanishes:

$$[S, T] = 0. \tag{3.205}$$

From (3.37) we then have

$$S^\rho \nabla_\rho T^\mu = T^\rho \nabla_\rho S^\mu. \tag{3.206}$$

With this in mind, let's compute the acceleration:

$$\begin{aligned}
A^\mu &= T^\rho \nabla_\rho (T^\sigma \nabla_\sigma S^\mu) \\
&= T^\rho \nabla_\rho (S^\sigma \nabla_\sigma T^\mu) \\
&= (T^\rho \nabla_\rho S^\sigma)(\nabla_\sigma T^\mu) + T^\rho S^\sigma \nabla_\rho \nabla_\sigma T^\mu \\
&= (S^\rho \nabla_\rho T^\sigma)(\nabla_\sigma T^\mu) + T^\rho S^\sigma (\nabla_\sigma \nabla_\rho T^\mu + R^\mu{}_{\nu\rho\sigma} T^\nu) \\
&= (S^\rho \nabla_\rho T^\sigma)(\nabla_\sigma T^\mu) + S^\sigma \nabla_\sigma (T^\rho \nabla_\rho T^\mu) - (S^\sigma \nabla_\sigma T^\rho)\nabla_\rho T^\mu \\
&\quad + R^\mu{}_{\nu\rho\sigma} T^\nu T^\rho S^\sigma \\
&= R^\mu{}_{\nu\rho\sigma} T^\nu T^\rho S^\sigma .
\end{aligned} \tag{3.207}$$

Let's think about this line by line. The first line is the definition of A^μ, and the second line comes directly from (3.206). The third line is simply the Leibniz rule. The fourth line replaces a double covariant derivative by the derivatives in the opposite order plus the Riemann tensor. In the fifth line we use Leibniz again (in the opposite order from usual), and then we cancel two identical terms and notice that the term involving $T^\rho \nabla_\rho T^\mu$ vanishes because T^μ is the tangent vector to a geodesic. The result,

$$\boxed{A^\mu = \frac{D^2}{dt^2} S^\mu = R^\mu{}_{\nu\rho\sigma} T^\nu T^\rho S^\sigma ,} \tag{3.208}$$

is the **geodesic deviation equation**. It expresses something that we might have expected: the relative acceleration between two neighboring geodesics is proportional to the curvature.

The geodesic deviation equation characterizes the behavior of a one-parameter family of neighboring geodesics. We will sometimes be interested in keeping track of the behavior of a multi-dimensional set of neighboring geodesics, perhaps representing a bundle of photons or a distribution of massive test particles. Such a set of geodesics forms a congruence; in Appendix F we derive equations that describe the evolution of such congruences.

Physically, of course, the acceleration of neighboring geodesics is interpreted as a manifestation of gravitational tidal forces. In the next chapter we explore in more detail how properties of curved spacetime are reflected by physics in a gravitational field.

3.11 ■ EXERCISES

1. Verify these consequences of metric compatibility ($\nabla_\sigma g_{\mu\nu} = 0$):

$$\begin{aligned}
\nabla_\sigma g^{\mu\nu} &= 0 \\
\nabla_\lambda \epsilon_{\mu\nu\rho\sigma} &= 0.
\end{aligned} \tag{3.209}$$

2. You are familiar with the operations of gradient ($\nabla\phi$), divergence ($\nabla\cdot\mathbf{V}$) and curl ($\nabla\times\mathbf{V}$) in ordinary vector analysis in three-dimensional Euclidean space. Using covariant derivatives, derive formulae for these operations in spherical polar coordinates $\{r, \theta, \phi\}$ defined by

$$x = r\sin\theta\cos\phi \tag{3.210}$$

$$y = r\sin\theta\sin\phi \tag{3.211}$$

$$z = r\cos\theta. \tag{3.212}$$

Compare your results to those in Jackson (1999) or an equivalent text. Are they identical? Should they be?

3. Imagine we have a *diagonal* metric $g_{\mu\nu}$. Show that the Christoffel symbols are given by

$$\Gamma^{\lambda}_{\mu\nu} = 0 \tag{3.213}$$

$$\Gamma^{\lambda}_{\mu\mu} = -\tfrac{1}{2}(g_{\lambda\lambda})^{-1}\partial_\lambda g_{\mu\mu} \tag{3.214}$$

$$\Gamma^{\lambda}_{\mu\lambda} = \partial_\mu\left(\ln\sqrt{|g_{\lambda\lambda}|}\right) \tag{3.215}$$

$$\Gamma^{\lambda}_{\lambda\lambda} = \partial_\lambda\left(\ln\sqrt{|g_{\lambda\lambda}|}\right) \tag{3.216}$$

In these expressions, $\mu \neq \nu \neq \lambda$, and repeated indices are *not* summed over.

4. In Euclidean three-space, we can define paraboloidal coordinates (u, v, ϕ) via

$$x = uv\cos\phi \quad y = uv\sin\phi \quad z = \tfrac{1}{2}(u^2 - v^2).$$

- **(a)** Find the coordinate transformation matrix between paraboloidal and Cartesian coordinates $\partial x^{\alpha}/\partial x^{\beta'}$ and the inverse transformation. Are there any singular points in the map?
- **(b)** Find the basis vectors and basis one-forms in terms of Cartesian basis vectors and forms.
- **(c)** Find the metric and inverse metric in paraboloidal coordinates.
- **(d)** Calculate the Christoffel symbols.
- **(e)** Calculate the divergence $\nabla_\mu V^\mu$ and Laplacian $\nabla_\mu\nabla^\mu f$.

5. Consider a 2-sphere with coordinates (θ, ϕ) and metric

$$ds^2 = d\theta^2 + \sin^2\theta\, d\phi^2. \tag{3.217}$$

- **(a)** Show that lines of constant longitude (ϕ = constant) are geodesics, and that the only line of constant latitude (θ = constant) that is a geodesic is the equator ($\theta = \pi/2$).
- **(b)** Take a vector with components $V^\mu = (1, 0)$ and parallel-transport it once around a circle of constant latitude. What are the components of the resulting vector, as a function of θ?

6. A good approximation to the metric outside the surface of the Earth is provided by

$$ds^2 = -(1 + 2\Phi)dt^2 + (1 - 2\Phi)dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2), \tag{3.218}$$

where

$$\Phi = -\frac{GM}{r} \tag{3.219}$$

may be thought of as the familiar Newtonian gravitational potential. Here G is Newton's constant and M is the mass of the earth. For this problem Φ may be assumed to be small.

(a) Imagine a clock on the surface of the Earth at distance R_1 from the Earth's center, and another clock on a tall building at distance R_2 from the Earth's center. Calculate the time elapsed on each clock as a function of the coordinate time t. Which clock moves faster?

(b) Solve for a geodesic corresponding to a circular orbit around the equator of the Earth ($\theta = \pi/2$). What is $d\phi/dt$?

(c) How much proper time elapses while a satellite at radius R_1 (skimming along the surface of the earth, neglecting air resistance) completes one orbit? You can work to first order in Φ if you like. Plug in the actual numbers for the radius of the Earth and so on (don't forget to restore the speed of light) to get an answer in seconds. How does this number compare to the proper time elapsed on the clock stationary on the surface?

7. For this problem you will use the parallel propagator introduced in Appendix I to see how the Riemann tensor arises from parallel transport around an infinitesimal loop. Consider the following loop:

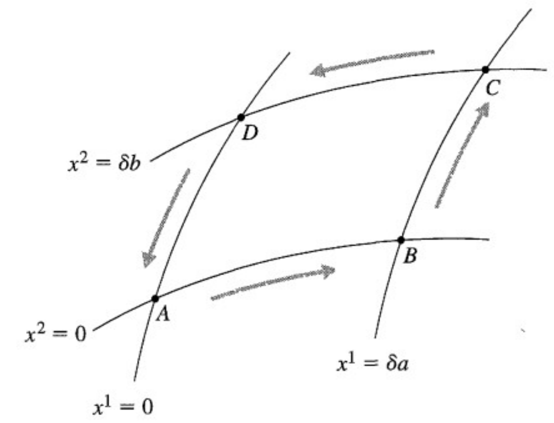

Using the infinite series expression for the parallel propagator, compute to lowest nontrivial order in δa and δb the transformation induced on a vector that is parallel transported around this loop from A to B to C to D and back to A, and show it is proportional to the appropriate components of the Riemann tensor. To make things easy, you can use x^1 and x^2 as parameters on the appropriate legs of the journey.

8. The metric for the three-sphere in coordinates $x^\mu = (\psi, \theta, \phi)$ can be written

$$ds^2 = d\psi^2 + \sin^2\psi\,(d\theta^2 + \sin^2\theta\, d\phi^2). \tag{3.220}$$

(a) Calculate the Christoffel connection coefficients. Use whatever method you like, but it is good practice to get the connection coefficients by varying the integral (3.49).
(b) Calculate the Riemann tensor, Ricci tensor, and Ricci scalar.
(c) Show that (3.191) is obeyed by this metric, confirming that the three-sphere is a maximally symmetric space (as you would expect).

9. Show that the Weyl tensor $C^{\mu}{}_{\nu\rho\sigma}$ is left invariant by a conformal transformation.

10. Show that, for $n \geq 4$, the Weyl tensor satisfies a version of the Bianchi identity,

$$\nabla_\rho C^{\rho}{}_{\sigma\mu\nu} = 2\frac{(n-3)}{(n-2)}\left(\nabla_{[\mu}R_{\nu]\sigma} + \frac{1}{2(n-1)}g_{\sigma[\mu}\nabla_{\nu]}R\right). \tag{3.221}$$

11. Since the Poincare half-plane with metric (3.192) is maximally symmetric, we might expect that it is rotationally symmetric around any point, although this certainly isn't evident in the $\{x, y\}$ coordinates. If that is so, it should be possible to put the metric in a form where the rotational symmetry is manifest, such as

$$ds^2 = f^2(r)[dr^2 + r^2 d\theta^2]. \tag{3.222}$$

To show that this works, calculate the curvature scalar for this metric and solve for the function $f(r)$ subject to the condition $R = -2/a^2$ everywhere. What is the range of the coordinate r?

12. Show that any Killing vector K^μ satisfies the relations mentioned in the text:

$$\begin{aligned} \nabla_\mu \nabla_\sigma K^\rho &= R^{\rho}{}_{\sigma\mu\nu}K^\nu \\ K^\lambda \nabla_\lambda R &= 0. \end{aligned} \tag{3.223}$$

13. Find explicit expressions for a complete set of Killing vector fields for the following spaces:
(a) Minkowski space, with metric $ds^2 = -dt^2 + dx^2 + dy^2 + dz^2$.
(b) A spacetime with coordinates $\{u, v, x, y\}$ and metric

$$ds^2 = -(du dv + dv du) + a^2(u)dx^2 + b^2(u)dy^2, \tag{3.224}$$

where a and b are unspecified functions of u. This represents a gravitational wave spacetime. (*Hints*, which you need not show: there are five Killing vectors in all, and all of them have a vanishing u component K^u.)

Be careful, in all of these cases, about the distinction between upper and lower indices.

14. Consider the three Killing vectors of the the two-sphere, (3.188). Show that their commutators satisfy the following algebra:

$$\begin{aligned} [R, S] &= T \\ [S, T] &= R \\ [T, R] &= S. \end{aligned} \tag{3.225}$$

15. Use Raychaudhuri's equation, discussed in Appendix F, to show that, if a fluid is flowing on geodesics through spacetime with zero shear and expansion, then spacetime must have a timelike Killing vector.

16. Consider again the metric on a three-sphere,

$$ds^2 = \mathrm{d}\psi^2 + \sin^2\psi\,(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2). \tag{3.226}$$

In this problem we make use of noncoordinate bases, discussed in Appendix J. In an orthonormal frame of one-forms $\hat{\theta}^{(a)}$ the metric would become

$$ds^2 = \hat{\theta}^{(1)} \otimes \hat{\theta}^{(1)} + \hat{\theta}^{(2)} \otimes \hat{\theta}^{(2)} + \hat{\theta}^{(3)} \otimes \hat{\theta}^{(3)}. \tag{3.227}$$

(a) Find such an orthonormal frame of one-forms, such that the matrix e_a^μ is diagonal. Don't worry about covering the entire manifold.

(b) Compute the components of the spin connection by solving $\mathrm{d}e^a + \omega^a{}_b \wedge e^b = 0$.

(c) Compute the components of the Riemann tensor $R^\rho{}_{\sigma\mu\nu}$ in the coordinate basis adapted to x^μ by computing the components of the curvature two-form $R^a{}_{b\mu\nu}$ and then converting.

-3 7 2 " 888

in empty space, Einstein-Cartan theory is indistinguishable from general relativity in experiments carried out in vacuum.

2.11 Connection coefficients in terms of the metric

The connection coefficients have been defined, equation (2.46), as derivatives of the tangent basis vectors $\boldsymbol{e}_\mu$. However, the connection coefficients can be expressed purely in terms of the (first derivatives of the) metric, without reference to the individual basis vectors. The partial derivatives of the metric are

$$\begin{aligned}
\frac{\partial g_{\lambda\mu}}{\partial x^\nu} &= \frac{\partial \boldsymbol{e}_\lambda \cdot \boldsymbol{e}_\mu}{\partial x^\nu} \\
&= \boldsymbol{e}_\lambda \cdot \frac{\partial \boldsymbol{e}_\mu}{\partial x^\nu} + \boldsymbol{e}_\mu \cdot \frac{\partial \boldsymbol{e}_\lambda}{\partial x^\nu} \\
&= \boldsymbol{e}_\lambda \cdot \boldsymbol{e}_\kappa \, \Gamma^\kappa_{\mu\nu} + \boldsymbol{e}_\mu \cdot \boldsymbol{e}_\kappa \, \Gamma^\kappa_{\lambda\nu} \\
&= g_{\lambda\kappa} \, \Gamma^\kappa_{\mu\nu} + g_{\mu\kappa} \, \Gamma^\kappa_{\lambda\nu} \\
&= \Gamma_{\lambda\mu\nu} + \Gamma_{\mu\lambda\nu} \ , \qquad (2.59)
\end{aligned}$$

which is a sum of two connection coefficients. Here $\Gamma_{\lambda\mu\nu}$ with all indices lowered is defined to be $\Gamma^\kappa_{\mu\nu}$ with the first index lowered by the metric,

$$\Gamma_{\lambda\mu\nu} \equiv g_{\lambda\kappa} \Gamma^\kappa_{\mu\nu} \ . \qquad (2.60)$$

Combining the metric derivatives in the following fashion yields an expression for a single connection,

$$\begin{aligned}
\frac{\partial g_{\lambda\mu}}{\partial x^\nu} + \frac{\partial g_{\lambda\nu}}{\partial x^\mu} - \frac{\partial g_{\mu\nu}}{\partial x^\lambda} &= \Gamma_{\lambda\mu\nu} + \Gamma_{\mu\lambda\nu} + \Gamma_{\lambda\nu\mu} + \Gamma_{\nu\lambda\mu} - \Gamma_{\mu\nu\lambda} - \Gamma_{\nu\mu\lambda} \\
&= 2\,\Gamma_{\lambda\mu\nu} - S_{\lambda\mu\nu} - S_{\mu\nu\lambda} - S_{\nu\mu\lambda} \ , \qquad (2.61)
\end{aligned}$$

with $S_{\lambda\mu\nu} \equiv g_{\lambda\kappa} S^\kappa_{\mu\nu}$, which shows that, in the presence of torsion,

$$\Gamma_{\lambda\mu\nu} = \frac{1}{2} \left(\frac{\partial g_{\lambda\mu}}{\partial x^\nu} + \frac{\partial g_{\lambda\nu}}{\partial x^\mu} - \frac{\partial g_{\mu\nu}}{\partial x^\lambda} + S_{\lambda\mu\nu} + S_{\mu\nu\lambda} + S_{\nu\mu\lambda} \right) \quad \textit{not} \text{ a coordinate tensor} \ . \qquad (2.62)$$

If torsion vanishes, as general relativity assumes, then

$$\boxed{\Gamma_{\lambda\mu\nu} = \frac{1}{2} \left(\frac{\partial g_{\lambda\mu}}{\partial x^\nu} + \frac{\partial g_{\lambda\nu}}{\partial x^\mu} - \frac{\partial g_{\mu\nu}}{\partial x^\lambda} \right)} \quad \textit{not} \text{ a coordinate tensor} \ . \qquad (2.63)$$

This is the formula that allows connection coefficients to be calculated from the metric.

2.12 Torsion-free covariant derivative

Einstein's principle of equivalence postulates that a locally inertial frame exists at each point of spacetime, and this implies that torsion vanishes. However, torsion is of special interest as a generalization of general relativity because, as discussed in §2.19.2, the torsion tensor and the Riemann curvature tensor can be regarded as fields associated with local gauge groups of respectively displacements and Lorentz transformations. Together displacements and Lorentz transformations form the Poincaré group of symmetries of spacetime. Torsion arises naturally in extensions of general relativity such as supergravity (Nieuwenhuizen, 1981), which unify the Poincaré group by adjoining supersymmetry transformations.

The torsion-free part of the covariant derivative is a covariant derivative even when torsion is present (that is, it yields a tensor when acting on a tensor). The torsion-free covariant derivative is important, even when torsion is present, for several reasons. Firstly, as will be discovered from an action principle in Chapter 4, the covariant derivative that goes in the geodesic equation (2.88) is the torsion-free covariant derivative, equation (2.90). Secondly, the torsion-free covariant curl defines the exterior derivative in the theory of differential forms, §15.6. The exterior derivative has the property that it is inverse to integration over curved hypersurfaces. Integration is central to various aspects of general relativity, such as the development of Lagrangian and Hamiltonian mechanics. Thirdly, the Lie derivative, §7.34, is a covariant derivative defined in terms of torsion-free covariant derivatives. Finally, Yang-Mills gauge symmetries, such as the $U(1)$ gauge symmetry of electromagnetism, require the gauge field to be defined in terms of the torsion-free covariant derivative, in order to preserve the gauge symmetry.

When torsion is present and it is desirable to make the torsion part explicit, it is convenient to distinguish torsion-free quantities with a $^\circ$ overscript. The torsion-free part $\mathring{\Gamma}_{\lambda\mu\nu}$ of the connection, also called the **Levi-Civita connection**, is given by the right hand side of equation (2.63). When expressed in a coordinate frame (as opposed to a tetrad frame, §11.15), the components of the torsion-free connections $\mathring{\Gamma}_{\lambda\mu\nu}$ are also called **Christoffel symbols**. Sometimes, the components $\mathring{\Gamma}_{\lambda\mu\nu}$ with all indices lowered are called Christoffel symbols of the first kind, while components $\mathring{\Gamma}^{\lambda}_{\mu\nu}$ with first index raised are called Christoffel symbols of the second kind. There is no need to remember the jargon, but it is useful to know what it means if you meet it.

The torsion-full connection $\Gamma_{\lambda\mu\nu}$ is a sum of the torsion-free connection $\mathring{\Gamma}_{\lambda\mu\nu}$ and a tensor called the **contortion tensor** (not contorsion!) $K_{\lambda\mu\nu}$,

$$\Gamma_{\lambda\mu\nu} = \mathring{\Gamma}_{\lambda\mu\nu} + K_{\lambda\mu\nu} \quad \textit{not} \text{ a coordinate tensor} . \tag{2.64}$$

From equation (2.62), the contortion tensor $K_{\lambda\mu\nu}$ is related to the torsion tensor $S_{\lambda\mu\nu}$ by

$$K_{\lambda\mu\nu} = \tfrac{1}{2}\left(S_{\lambda\mu\nu} + S_{\mu\nu\lambda} + S_{\nu\mu\lambda}\right) = -\,S_{\nu\lambda\mu} + \tfrac{3}{2}S_{[\lambda\mu\nu]} \quad \text{a coordinate tensor} . \tag{2.65}$$

The contortion $K_{\lambda\mu\nu}$ is antisymmetric in its first two indices,

$$K_{\lambda\mu\nu} = -K_{\mu\lambda\nu} , \tag{2.66}$$

and thus like the torsion tensor $S_{\lambda\mu\nu}$ has $6 \times 4 = 24$ degrees of freedom. The torsion tensor $S_{\lambda\mu\nu}$ can be expressed in terms of the contortion tensor $K_{\lambda\mu\nu}$,

$$S_{\lambda\mu\nu} = K_{\lambda\mu\nu} - K_{\lambda\nu\mu} = -\,K_{\mu\nu\lambda} + 3\,K_{[\lambda\mu\nu]} \quad \text{a coordinate tensor} . \tag{2.67}$$

The torsion-full covariant derivative D_ν differs from the torsion-free covariant derivative $\mathring{D}_\nu$ by the contortion,

$$D_\nu A^\kappa \equiv \mathring{D}_\nu A^\kappa + K^{\kappa}_{\mu\nu} A^\mu \quad \text{a coordinate tensor} . \tag{2.68}$$

In this book torsion will not be assumed automatically to vanish, and thus by default the symbol D_ν will denote the torsion-full covariant derivative. When torsion is assumed to vanish, or when D_ν denotes the torsion-free covariant derivative, it will be explicitly stated so.

Concept question 2.5 Can the metric be Minkowski in the presence of torsion? In §2.10.1 it was argued that the postulate of the existence of locally inertial frames implies that torsion vanishes. The basis of the argument was the proposition that derivatives of the tangent axes vanish, equation (2.54b). Impose instead the weaker condition that the derivatives of the metric (i.e. scalar products of tangent axes) vanish,

$$\frac{\partial g_{\lambda\mu}}{\partial x^\nu} = 0 \ . \tag{2.69}$$

Can torsion be non-vanishing under this weaker condition? **Answer.** Yes. In fact torsion may exist even in flat (Minkowski) space, where the metric is everywhere Minkowski, $g_{\lambda\mu} = \eta_{\lambda\mu}$. The condition (2.69) of vanishing metric derivatives is equivalent to the vanishing of the torsion-free connections,

$$\frac{1}{2}\frac{\partial g_{\lambda\mu}}{\partial x^\nu} = \Gamma_{(\lambda\mu)\nu} = \mathring{\Gamma}_{(\lambda\mu)\nu} + K_{(\lambda\mu)\nu} = \mathring{\Gamma}_{(\lambda\mu)\nu} = 0 \ . \tag{2.70}$$

Thus the condition (2.69) of vanishing metric derivatives imposes no condition on torsion.

Exercise 2.6 Covariant curl and coordinate curl. Show that the covariant curl of a covariant vector A_λ is

$$D_\kappa A_\lambda - D_\lambda A_\kappa = \frac{\partial A_\lambda}{\partial x^\kappa} - \frac{\partial A_\kappa}{\partial x^\lambda} + S^\mu_{\kappa\lambda} A_\mu \ . \tag{2.71}$$

Conclude that the coordinate curl of a vector equals its torsion-free covariant curl,

$$\mathring{D}_\kappa A_\lambda - \mathring{D}_\lambda A_\kappa = \frac{\partial A_\lambda}{\partial x^\kappa} - \frac{\partial A_\kappa}{\partial x^\lambda} \ . \tag{2.72}$$

Of course, if torsion vanishes as general relativity assumes, then the covariant curl is the torsion-free covariant curl. Note that since both $D_\kappa A_\lambda - D_\lambda A_\kappa$ on the left hand side and $S^\mu_{\kappa\lambda} A_\mu$ on the right hand side of equation (2.71) are both tensors, it follows that the coordinate curl $\partial A_\lambda / \partial x^\kappa - \partial A_\kappa / \partial x^\lambda$ is a tensor even in the presence of torsion.

Exercise 2.7 Covariant divergence and coordinate divergence. Show that the covariant divergence of a contravariant vector A^μ is

$$D_\mu A^\mu = \frac{1}{\sqrt{-g}} \frac{\partial(\sqrt{-g} A^\mu)}{\partial x^\mu} + S^\nu_{\mu\nu} A^\mu \ , \tag{2.73}$$

where $g \equiv |g_{\mu\nu}|$ is the determinant of the metric matrix. Conclude that the torsion-free covariant divergence is

$$\mathring{D}_\mu A^\mu = \frac{1}{\sqrt{-g}} \frac{\partial(\sqrt{-g} A^\mu)}{\partial x^\mu} \ . \tag{2.74}$$

Of course, if torsion vanishes as general relativity assumes, then the covariant divergence is the torsion-free covariant divergence. Note that since both the covariant divergence on the left hand side of equation (2.73) and the torsion term on the right hand side of equation (2.73) are both tensors, the torsion-free covariant divergence (2.74) is a tensor even in the presence of torsion.

Solution. The covariant divergence is

$$D_\mu A^\mu = \frac{\partial A^\mu}{\partial x^\mu} + \Gamma^\nu_{\mu\nu} A^\mu \ . \tag{2.75}$$

From equation (2.62),

$$\begin{aligned} \Gamma^\nu_{\mu\nu} &= \frac{1}{2} g^{\lambda\nu} \frac{\partial g_{\lambda\nu}}{\partial x^\mu} + S^\nu_{\mu\nu} \\ &= \frac{\partial \ln |\sqrt{-g}|}{\partial x^\mu} + S^\nu_{\mu\nu} \ . \end{aligned} \tag{2.76}$$

The second line of equations (2.76) follows because for any matrix M,

$$\begin{aligned} \delta \ln |M| &= \ln |M + \delta M| - \ln |M| \\ &= \ln |M^{-1}(M + \delta M)| \\ &= \ln |1 + M^{-1} \delta M| \\ &= \ln (1 + \mathrm{Tr}\ M^{-1} \delta M) \\ &= \mathrm{Tr}\ M^{-1} \delta M \ . \end{aligned} \tag{2.77}$$

The torsion-free covariant divergence is

$$\mathring{D}_\mu A^\mu = \frac{\partial A^\mu}{\partial x^\mu} + \mathring{\Gamma}^\nu_{\mu\nu} A^\mu \ , \tag{2.78}$$

where the torsion-free coordinate connection is

$$\mathring{\Gamma}^\nu_{\mu\nu} = \frac{1}{2} g^{\lambda\nu} \frac{\partial g_{\lambda\nu}}{\partial x^\mu} = \frac{\partial \ln |\sqrt{-g}|}{\partial x^\mu} \ . \tag{2.79}$$

Concept question 2.8 If torsion does not vanish, does torsion-free covariant differentiation commute with the metric? Answer. Yes. Unlike the torsion-full covariant derivative, Concept Question 2.3, the torsion-free covariant derivative of the tangent basis vectors $\boldsymbol{e}_\kappa$ does *not* vanish, but rather depends on the contortion $K^\nu_{\kappa\mu} \boldsymbol{e}_\nu$,

$$\mathring{D}_\mu \boldsymbol{e}_\kappa = D_\mu \boldsymbol{e}_\kappa + K^\nu_{\kappa\mu} \boldsymbol{e}_\nu = K^\nu_{\kappa\mu} \boldsymbol{e}_\nu \ . \tag{2.80}$$

However, the torsion-free covariant derivative of the metric, that is, of scalar products of the tangent basis vectors, does vanish,

$$\mathring{D}_\mu g_{\kappa\lambda} = \mathring{D}_\mu (\boldsymbol{e}_\kappa \cdot \boldsymbol{e}_\lambda) = K^\nu_{\kappa\mu} \boldsymbol{e}_\nu \cdot \boldsymbol{e}_\lambda + K^\nu_{\lambda\mu} \boldsymbol{e}_\kappa \cdot \boldsymbol{e}_\nu = K_{\lambda\kappa\mu} + K_{\kappa\lambda\mu} = 0 \ , \tag{2.81}$$

thanks to the antisymmetry of the contortion tensor in its first two indices. As a corollary, torsion-free covariant differentiation commutes with the operations of raising and lowering indices, and of contraction.

2.13 Mathematical aside: What if there is no metric?

General relativity is a metric theory. Many of the structures introduced above can be defined mathematically without a metric. For example, it is possible to define the tangent space of vectors with basis $\boldsymbol{e}_\mu$, and to define a dual vector space with basis $\boldsymbol{e}^\mu$ such that $\boldsymbol{e}^\mu \cdot \boldsymbol{e}_\nu = \delta^\mu_\nu$, equation (2.36). Elements of the dual vector space are called covectors. Similarly it is possible to define connections and covariant derivatives without a metric. However, this book follows general relativity in assuming that spacetime has a metric.

2.14 Coordinate 4-velocity

Consider a particle following a worldline

$$x^\mu(\tau) \ , \tag{2.82}$$

where τ is the particle's proper time. The proper time along any interval of the worldline is $d\tau \equiv \sqrt{-ds^2}$.

Define the **coordinate 4-velocity** u^μ by

$$\boxed{u^\mu \equiv \frac{dx^\mu}{d\tau}} \quad \text{a coordinate 4-vector} \ . \tag{2.83}$$

The magnitude squared of the 4-velocity is constant

$$u_\mu u^\mu = g_{\mu\nu} \frac{dx^\mu}{d\tau} \frac{dx^\nu}{d\tau} = \frac{ds^2}{d\tau^2} = -1 \ . \tag{2.84}$$

The negative sign arises from the choice of metric signature: with the signature $-{+}{+}{+}$ adopted here, there is a $-$ sign between ds^2 and $d\tau^2$. Equation (2.84) can be regarded as an integral of motion associated with conservation of particle rest mass.

2.15 Geodesic equation

Let $\boldsymbol{u} \equiv \boldsymbol{e}_\mu u^\mu$ be the 4-velocity in coordinate-independent notation. The principle of equivalence (which imposes vanishing torsion) implies that the **geodesic equation**, the equation of motion of a freely-falling particle, is

$$\boxed{\frac{d\boldsymbol{u}}{d\tau} = 0} \ . \tag{2.85}$$

Why? Because $d\boldsymbol{u}/d\tau = 0$ in the particle's own free-fall frame, and the equation is coordinate-independent. In the particle's own free-fall frame, the particle's 4-velocity is $u^\mu = \{1, 0, 0, 0\}$, and the particle's locally inertial axes $\boldsymbol{e}_\mu = \{\boldsymbol{e}_0, \boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3\}$ are constant.

' & $

Exercise 11.6 Symmetry of the Riemann tensor. Show that the cyclic symmetry (11.63) implies the symmetry $kl \leftrightarrow mn$, given the antisymmetries $k \leftrightarrow l$ and $m \leftrightarrow n$. Given Exercise 11.5, this shows that the symmetry $kl \leftrightarrow mn$ is, like the cyclic symmetry, a consequence of vanishing torsion.
Solution. Show that

$$2(R_{klmn} - R_{mnkl}) = 3\left(R_{k[lmn]} - R_{l[kmn]} - R_{m[nkl]} + R_{n[mkl]}\right) \ , \tag{11.71}$$

or alternatively,

$$2(R_{klmn} - R_{mnkl}) = 3\left(R_{[klm]n} - R_{[kln]m} - R_{[mnk]l} + R_{[mnl]k}\right) \ . \tag{11.72}$$

Exercise 11.7 Number of components of the Riemann tensor. How many independent components does the Riemann tensor have, in 4-dimensional spacetime?
Solution. If torsion vanishes, 20. If torsion does not vanish, 36. The extra 16 components come from $R_{[klm]n}$, which is related to torsion by equation (11.69), and which has $4 \times 4 = 16$ components if torsion does not vanish.

Concept question 11.8 Must connections vanish if Riemann vanishes? Must the tetrad connections Γ_{lmn} vanish if the Riemann tensor vanishes identically, $R_{klmn} = 0$? **Answer.** No. For a counterexample, take flat (Minkowski) space expressed in spherical polar coordinates $\{t, r, \theta, \phi\}$. The non-vanishing tetrad-frame connections are $\Gamma_{212} = \Gamma_{313} = 1/r$ and $\Gamma_{323} = \cot\theta/r$ (compare equations (20.21)).

11.16.1 Riemann tensor in a mixed coordinate-tetrad frame

In Chapter 16, Einstein's equations will be obtained from an action principle, as first done by Hilbert (1915). The Hilbert Lagrangian takes a particularly insightful form if the Riemann tensor is expressed in a mixed coordinate-tetrad basis.

The coordinate-frame covariant derivative D_κ of a tetrad-frame vector a_n is

$$D_\kappa a_n = e^k{}_\kappa D_k a_n = \frac{\partial a_n}{\partial x^\kappa} - \Gamma^m_{n\kappa} a_m \quad \text{a coordinate-tetrad tensor} \ , \tag{11.73}$$

where $\Gamma^m_{n\kappa}$ is the tetrad-frame connection with its last index converted into the coordinate frame with the vierbein,

$$\Gamma^m_{n\kappa} \equiv e^k{}_\kappa \Gamma^m_{nk} \quad \text{a coordinate vector, but } \textit{not} \text{ a tetrad tensor} \ . \tag{11.74}$$

As usual, the connection with all indices lowered is defined by $\Gamma_{ln\kappa} \equiv \gamma_{lm}\Gamma^m_{n\kappa}$. The connections $\Gamma_{mn\kappa}$ should not be confused with the coordinate-frame connections (Christoffel symbols) $\Gamma_{\mu\nu\kappa}$. The relation between the two is, from equation (11.44),

$$\Gamma_{mn\kappa} = -\, e^k{}_\kappa \, d_{mnk} + e_m{}^\mu e_n{}^\nu \, \Gamma_{\mu\nu\kappa} \ . \tag{11.75}$$

In 4 dimensions there are $6 \times 4 = 24$ distinct connections $\Gamma_{mn\kappa}$ (with or without torsion), whereas there are $4 \times 10 = 40$ distinct coordinate-frame connections $\Gamma_{\mu\nu\kappa}$ (without torsion, or $4 \times 4 \times 4 = 64$ with torsion).

The last term on the right hand side of equation (11.60) for the Riemann tensor can be written, in view of equations (11.49) and (11.32),

$$(\Gamma^p_{kl} - \Gamma^p_{lk} - S^p_{kl})\Gamma_{mnp} = (\partial_l e_k{}^\kappa - \partial_k e_l{}^\kappa)\Gamma_{mn\kappa} \ . \tag{11.76}$$

The Riemann tensor $R_{\kappa\lambda mn}$ in the mixed coordinate-tetrad basis is then

$$\boxed{R_{\kappa\lambda mn} = \frac{\partial \Gamma_{mn\lambda}}{\partial x^\kappa} - \frac{\partial \Gamma_{mn\kappa}}{\partial x^\lambda} + \Gamma^p_{m\lambda}\Gamma_{pn\kappa} - \Gamma^p_{m\kappa}\Gamma_{pn\lambda}} \quad \text{a coordinate-tetrad tensor} \ , \tag{11.77}$$

which is valid with or without torsion. Equation (11.77) resembles superficially the coordinate-frame expression (2.112) for the Riemann tensor, but it is more economical in that there are only 24 connections $\Gamma_{mn\kappa}$ instead of the 40 (or 64, with torsion) coordinate-frame connections $\Gamma_{\mu\nu\kappa}$.

The torsion $S^m_{\kappa\lambda}$ in the mixed coordinate-tetrad basis is

$$\boxed{S^m_{\kappa\lambda} = -\frac{\partial e^m{}_\lambda}{\partial x^\kappa} + \frac{\partial e^m{}_\kappa}{\partial x^\lambda} - \Gamma^m_{l\kappa} e^l{}_\lambda + \Gamma^m_{k\lambda} e^k{}_\kappa} \quad \text{a coordinate-tetrad tensor} \ . \tag{11.78}$$

Equations (11.77) and (11.78) constitute Cartan's equations of structure (Cartan, 1904) (see §16.14.2).

11.17 Ricci, Einstein, Bianchi

The usual suite of formulae leading to Einstein's equations apply. Since all the quantities are tensors, and all the equations are tensor equations, their form follows immediately from their coordinate counterparts.

Ricci tensor:

$$\boxed{R_{km} \equiv \gamma^{ln} R_{klmn}} \ . \tag{11.79}$$

Ricci scalar:

$$\boxed{R \equiv \gamma^{km} R_{km}} \ . \tag{11.80}$$

Einstein tensor:

$$\boxed{G_{km} \equiv R_{km} - \tfrac{1}{2}\gamma_{km} R} \ . \tag{11.81}$$

Einstein's equations:

$$\boxed{G_{km} = 8\pi G T_{km}} \ . \tag{11.82}$$

The trace of the Einstein equations implies that $R = -8\pi G T$, so the Einstein equations (11.82) can equally well be written with the trace terms transferred from the left to the right hand side,

$$R_{km} = 8\pi G \left(T_{km} - \tfrac{1}{2}\gamma_{km} T\right) \ . \tag{11.83}$$

Bianchi identities in the absence of torsion:

$$\boxed{D_k R_{lmnp} + D_l R_{mknp} + D_m R_{klnp} = 0} \ , \tag{11.84}$$

which most importantly imply covariant conservation of the Einstein tensor, hence conservation of energy-momentum

$$\boxed{D^k T_{km} = 0} \ . \tag{11.85}$$

11.18 Expressions with torsion

If torsion does not vanish, then the Riemann tensor, and consequently also the Ricci and Einstein tensors, can be split into torsion-free (distinguished by a $^\circ$ overscript) and torsion parts (e.g. Hehl, Heyde, and Kerlick 1976). A similar split occurs in the ADM formalism where a certain gauge choice (fixing the time component $\boldsymbol{\gamma}_0$ of the tetrad to be orthogonal to hypersurfaces of constant time) splits the tetrad connection into a tensor part, the extrinsic curvature, and a remainder, equation (17.27).

The contortion tensor K_{lmn} was defined previously as the torsion part of the connection Γ_{lmn}, equation (11.55). The unique non-vanishing contraction of the contortion tensor defines the contortion vector K_m,

$$K_m \equiv K^n_{mn} = S^n_{mn} \ . \tag{11.86}$$

The torsion-full Riemann tensor R_{klmn} is a sum of the torsion-free Riemann tensor $\mathring{R}_{klmn}$ and a torsion part (note that $K_{pkl} - K_{plk} - S_{pkl} = 0$, so the "extra" term in R_{klmn}, equation (11.60), vanishes when K_{pkl} is the contortion),

$$R_{klmn} = \mathring{R}_{klmn} + \mathring{D}_k K_{mnl} - \mathring{D}_l K_{mnk} + K^p_{ml} K_{pnk} - K^p_{mk} K_{pnl} \ . \tag{11.87}$$

The Ricci tensor is

$$R_{km} = \mathring{R}_{km} - \mathring{D}_k K_m - \mathring{D}^n K_{mnk} + K_{mpn} K^{np}{}_k - K_{mpk} K^p \ , \tag{11.88}$$

and the Ricci scalar is

$$R = \mathring{R} - 2\mathring{D}_n K^n + K_{mpn} K^{npm} - K_n K^n \ . \tag{11.89}$$

The antisymmetric part of the Einstein tensor is, from contracting equation (11.69),

$$G_{[km]} = \tfrac{3}{2} R_{[klm]}{}^l = \tfrac{3}{2} \left(D_{[k} S^l_{lm]} + S^p_{[kl} S^l_{m]p} \right) \ , \tag{11.90}$$

which vanishes for vanishing torsion.

The Jacobi identity (2.126) implies, in addition to the 16 conditions (11.69), the 24 Bianchi identities

$$D_{[k} R_{lm]np} + S^q_{[kl} R_{m]qnp} = 0 \ . \tag{11.91}$$

The doubly-contracted Bianchi identities are

$$-\tfrac{3}{2}\gamma^{kn}\gamma^{lp} \left(D_{[k} R_{lm]np} + S^q_{[kl} R_{m]qnp} \right) = D^k G_{mk} - \tfrac{1}{2} S^q_{kl} R_{mq}{}^{kl} - S^q_{km} R_q{}^k = 0 \ . \tag{11.92}$$

11.19 General relativity in 2 spacetime dimensions

General relativity in 2 spacetime dimensions is weird. There are zero Bianchi identities (2.128) in 2 spacetime dimensions, so the Bianchi identities do not identify any covariantly conserved tensor. The Einstein tensor itself vanishes identically in 2 spacetime dimensions.

There are consistent extensions of general relativity in 2 spacetime dimensions, such as string-inspired dilaton gravity (Grumiller, Kummer, and Vassilevich, 2002). However, those will not be considered here.

Historically, the main application of 2-dimensional relativity has been to explore quantum field theory in curved spacetime, since in 2 spacetime dimensions the quantum energy-momentum tensor induced by any prescribed geometry can be calculated exactly (even though the classical energy-momentum tensor is indeterminate).

The closest thing to a consistent realization of classical general relativity in 2 spacetime dimensions is as follows.

In 2 spacetime dimensions, the Riemann tensor has just one distinct component, R_{0101}, and that component is determined entirely by the Ricci scalar R. The tetrad-frame Riemann and Ricci tensors are related to the Ricci scalar R by

$$R_{klmn} = \tfrac{1}{2} \left(\gamma_{km} \gamma_{ln} - \gamma_{kn} \gamma_{lm} \right) R \ , \quad R_{km} = \tfrac{1}{2} \gamma_{km} R \ . \tag{11.93}$$

In an arbitrary number of N spacetime dimensions, contracting the Einstein equations implies that the Ricci scalar R is proportional to the trace T of the energy-momentum tensor,

$$(1 - \tfrac{1}{2} N) R = \kappa_N T \ , \tag{11.94}$$

where κ_N is Newton's gravitational constant in N spacetime dimensions, suitably normalized. For $N = 2$, the factor on the left of equation (11.94) vanishes; but one can imagine absorbing the zero factor into a redefinition of the gravitational constant κ_N, so that

$$R = \kappa_2' T \tag{11.95}$$

for some κ_2'. Now impose that the energy-momentum tensor T_{km} is covariantly conserved,

$$D^k T_{km} = 0 \ . \tag{11.96}$$

In $N = 2$ spacetime dimensions, the trace relation (11.95) together with the covariant conservation condition (11.96) imply almost uniquely the form of the energy-momentum tensor.

The conserved energy-momentum tensor T_{km} takes its simplest expression when the metric is expressed in conformally flat form. The metric in $N = 2$ spacetime dimensions is a symmetric 2×2 matrix. By a suitable coordinate transformation of the 2 coordinates, the metric can be brought to the conformally flat form

$$ds^2 = e^{2\xi} (- dt^2 + dx^2) = - e^{2\xi} dv du \ , \tag{11.97}$$

where $v \equiv t + x$ and $u \equiv t - x$ are null coordinates, and ξ is a function of the two coordinates. The

Newman-Penrose tetrad-frame components of the conserved energy-momentum tensor T_{km} are then

$$-\frac{R}{2} = -4e^{-2\xi}\frac{\partial^2\xi}{\partial v\partial u} = -\kappa_2'\frac{T}{2} = \kappa_2' T_{vu} \ , \tag{11.98a}$$

$$4e^{-2\xi}\left[\frac{\partial^2\xi}{\partial v^2} - \left(\frac{\partial\xi}{\partial v}\right)^2 + f^+(v)\right] = \kappa_2' T_{vv} \ , \tag{11.98b}$$

$$4e^{-2\xi}\left[\frac{\partial^2\xi}{\partial u^2} - \left(\frac{\partial\xi}{\partial u}\right)^2 + f^-(u)\right] = \kappa_2' T_{uu} \ , \tag{11.98c}$$

where $f^+(v)$ and $f^-(u)$ are arbitrary functions of respectively v and u. There is a residual gauge freedom $v \to V(v)$ and $u \to U(u)$ in the choice of null coordinates that allows the conformal function to be adjusted $\xi \to \xi + \xi^+(v) + \xi^-(u)$ by arbitrary additive functions of v and u. This residual gauge freedom allows the functions $f^+(v)$ and $f^-(u)$ in equations (11.98b) and (11.98c) to be adjusted arbitrarily. If desired, $f^+(v)$ and $f^-(u)$ can be set to zero.

The classical 2-dimensional general relativity described by equations (11.98) is not very interesting; for example there is no 2-dimensional analogue of the Schwarzschild black hole, Exercise 11.9.

Where equations (11.98) prove more interesting is that they also describe the expectation value $\langle T_{kl}\rangle$ of the renormalized quantum energy-momentum induced by a given geometry in 2 spacetime dimensions (Birrell and Davies, 1982). That is, the expectation value $\langle T\rangle$ of the quantum trace in 2 spacetime dimensions is proportional to the Ricci scalar (Birrell and Davies, 1982, eq. (6.121)), and the quantum energy-momentum tensor $\langle T_{kl}\rangle$ is covariantly conserved, therefore equations (11.98) are satisfied by $\langle T_{kl}\rangle$. In 4 spacetime dimensions the quantum energy-momentum tensor $\langle T_{kl}\rangle$ is extremely difficult to calculate in a general spacetime, so clues from 2 spacetime dimensions can be illuminating.

Exercise 11.9 Black holes in 2 spacetime dimensions? Does the analogue of a Schwarzschild black hole exist in 2 spacetime dimensions?

Solution. No. Require the spacetime to be empty outside some radius. The vanishing of the Ricci scalar (11.98a) implies that

$$\xi = \xi^+(v) + \xi^-(u) \tag{11.99}$$

for some functions ξ^+ and ξ^- of the null coordinates v and u. But then the coordinate tranformations of the null coordinates

$$dV = e^{2\xi^+(v)}dv \ , \quad dU = e^{2\xi^-(u)}du \tag{11.100}$$

bring the line element to

$$ds^2 = -dV dU \ , \tag{11.101}$$

which is just flat (Minkowski) space in $N = 2$ dimensions.

Exercise 11.10 Tidal forces falling into a Schwarzschild black hole. In the Schwarzschild or that the energy-momentum tensor is symmetric, equation (16.279). This motivates the ADM strategy of simply discarding the 6 antisymmetric components of the Einstein equations. These 6 components comprise the 3 antisymmetric components of the 9 equations of motion (16.331b), corresponding to the 3 antisymmetric space-space Einstein equations, and the 3 equations of motion (16.332), corresponding to the 3 antisymmetric time-space Einstein equations. Discarding the antisymmetric Einstein's equations seems innocent enough, until one realises that the antisymmetry of the energy-momentum is a consequence of the law of conservation of spin angular-momentum, equation (16.278), which law is responsible for the Gaussian constraints (16.334b) and (16.334c). Thus discarding the 6 antisymmetric Einstein equations is equivalent to using up the 6 Gaussian constraints. As a corollary, the 6 Gaussian constraints can no longer be treated as constraints; rather, they must be treated as identities. This should raise a red flag: it may not be a good numerical strategy to trade hyperbolic equations of motion for elliptic identities.

Finally, the usual ADM strategy (though not a necessary one — see Chapter 17), is to work entirely with coordinate-frame quantities. An advantage of this approach is that all quantities are spatially Lorentz gauge-invariant (the ADM gauge choice (16.325) removes the gauge freedom of Lorentz boosts). In particular, the 9 spatial vierbein $e_{a\alpha}$ reduce to the 6 components of the Lorentz gauge-invariant spatial metric $g_{\alpha\beta}$, and the 24 Lorentz connections are replaced by the 6 components of the symmetric (for vanishing torsion) extrinsic curvatures $\Gamma_{\alpha 0\beta}$ together with the $3 \times 6 = 18$ torsion-free coordinate-frame spatial connections (Christoffel symbols) $\Gamma_{\alpha\beta\gamma}$. In summary, in the ADM formalism there are $6 + 6 = 12$ equations of motion for $g_{\alpha\beta}$ and $\Gamma_{\alpha 0\beta}$, 18 identities for $\Gamma_{\alpha\beta\gamma}$, and 4 Hamiltonian constraints.

16.15.7 BSSN formalism

The BSSN formalism, discussed further in Chapter 17, §17.7, has gained popularity because it proves to have better numerical stability when applied to problems such as the merger of two black holes (Shibata and Nakamura, 1995; Baumgarte and Shapiro, 1998; Shinkai, 2009; Baumgarte and Shapiro, 2010; Brown et al., 2012).

The BSSN formalism follows ADM for the most part, in particular imposing the ADM gauge choice (16.325). However, BSSN retains the 3 momentum Einstein equations (16.332) as equations of motion, and restores the 3 Gaussian constraints (16.334c). Like the Hamiltonian constraints, the Gaussian constraints must be arranged to be satisfied in the initial conditions, but may be discarded thereafter.

16.15.8 An alternative strategy for solving the gravitational equations numerically

A drawback of the ADM and BSSN approaches is that expending 3 of the Lorentz gauge conditions on the vierbein, equation (16.325), preempts being able to use those Lorentz gauge conditions in potentially advantageous ways on the Lorentz connections, as is done for example in proofs of singularity theorems (conditions (18.26), for example). Moreover, as discussed in §16.15.6, transferring Lorentz gauge conditions from the Lorentz connections to the vierbein has the side-effect of replacing "nice" hyperbolic equations of motion with "nasty" elliptic identities.

This suggests that better numerical behaviour could result from using the Hamiltonian system gravitational equations in the native form derived in §16.15.3, where they consist of $12 + 12$ equations of motion, $6 + 4$ constraints, and 6 identities, equations (16.317) and (16.318) (as opposed to the $6 + 6$ equations of motion, 4 constraints, and 18 identities of the ADM formalism). As remarked previously, the system of gravitational equations derived in §16.15.3 resembles the Hamiltonian system of electromagnetic equations, which consist of $3 + 3$ equations of motion, 1 constraint, and 3 identities, equations (16.76) and (16.77).

Consider briefly the problem of solving the system of gravitational equations (16.317) and (16.318) numerically. First, all 40 equations must be arranged to be satisfied on an initial hypersurface of constant time. Subsequently, the 10 constraint equations may be discarded (or kept as a check of numerical accuracy). In integrating the remaining system of 30 equations from one hypersurface of constant time to the next, the first step is to use the $12 + 12$ equations of motion to find the 12 gravitational coordinates $\boldsymbol{e}$ and their 12 conjugate momenta $\boldsymbol{\pi}$ on the next (upper) hypersurface. As in ADM, coordinate gauge freedom allows the time components $\boldsymbol{e}_{\bar{t}}$ of the vierbein (the lapse and shift) to be chosen arbitrarily. The 12 components of $\boldsymbol{e}$ and the 4 components of $\boldsymbol{e}_{\bar{t}}$ account for all 16 components of the vierbein. Lorentz gauge freedom allows the time components $\boldsymbol{\Gamma}_{\bar{t}}$ of the Lorentz connection to be chosen arbitrarily. The 12 components of $\boldsymbol{\pi}$ and the 6 components of $\boldsymbol{\Gamma}_{\bar{t}}$ account for 18 of the 24 degrees of freedom of the Lorentz connection. The remaining 6 degrees of freedom of the Lorentz connection are determined by the 6 identities.

The 6 identities are packaged together with the 6 Gaussian constraints in the 12 equations (16.321a). Now the 12 equations (16.321a) comprise an expression for the 12 components of the spatial vector 2-form $\frac{1}{2}[\boldsymbol{\Gamma}, \boldsymbol{e}]$ in terms of spatial exterior derivatives $\mathrm{d}\boldsymbol{e}$ of the spatial vierbein and the spin angular-momentum (the latter vanishing if torsion is assumed to vanish). Since $\boldsymbol{e}$ over the upper spatial hypersurface of constant time is already in hand from integrating the equations of motion, taking the spatial exterior derivative $\mathrm{d}\boldsymbol{e}$ is a matter of differentiating numerically over the spatial hypersurface. The 12 components of $\frac{1}{2}[\boldsymbol{\Gamma}, \boldsymbol{e}]$ are

$$\tfrac{1}{2}[\boldsymbol{\Gamma}, \boldsymbol{e}] = \boldsymbol{\Gamma} \cdot \boldsymbol{e} = -2\, \Gamma_{k[\alpha\beta]}\, \boldsymbol{\gamma}^k\, d^2 x^{\alpha\beta} \ . \tag{16.336}$$

The Gaussian constraints involve the 6 components of $\frac{1}{2}[\boldsymbol{\Gamma}, \boldsymbol{e}]$ comprising

$$\boldsymbol{e} \wedge \tfrac{1}{2}[\boldsymbol{\Gamma}, \boldsymbol{e}] = -12\, e_{k\alpha}\, \Gamma_{l[\beta\gamma]}\, \boldsymbol{\gamma}^k \wedge \boldsymbol{\gamma}^l\, d^3 x^{\alpha\beta\gamma} \ , \tag{16.337}$$

summed over distinct antisymmetric sequences of indices kl and $\alpha\beta\gamma$. The components of equation (16.337) may be regarded as a 6×12 matrix acting on a 12-component vector consisting of the $\Gamma_{k[\alpha\beta]}$. The null space of the 6×12 matrix is another 6×12 matrix which, when acting on the 12-component vector $\Gamma_{k[\alpha\beta]}$, gives the 6 linear combinations of $\frac{1}{2}[\boldsymbol{\Gamma}, \boldsymbol{e}]$ whose values are determined by the 6 identities distinct from the 6 Gaussian constraints. Actually, it is not essential to choose the 6 identities to lie in the 6-dimensional null space; it suffices that the 6 identities span the null space. That is, the 6 identities and 6 Gaussian constraints should together span the full 12-dimensional space of $\frac{1}{2}[\boldsymbol{\Gamma}, \boldsymbol{e}]$. Equivalently, no linear combination of the 6 identities should be a constraint.

-3 4! ! !

3.1 Christoffel Symbol

Books - 002 - 2021\ General Relativity...\ grbook.pdf

Pag. 67 (97) -

2.12 Torsion-free covariant derivative

⇒ Christoffels symbols

Pag. 297 (327)

11.16.1

Riemann tensor a mixed coordinate-tetrad frame

Christoffel symbols

Pag. 451 (481)

3×6=18 torsion-free coordinate-frame spatial connections (Christoffel symbols)...

%   ' $

⇒ 06 Jan 2021 – 1.22 TB (20210106)

CVGA2019-Videos

⇒ { 184.123.201.35.bc.

99.185.102.34.bc. ⇐

{ 173.193.186.35.bc.

Book1-002-2021 | Cálculo |

Cálculo-vectorial-claudio-pita.pdf

Pág. 514 (523)

Figura 4

Curvas paralelas a la elipse $\alpha(t) =$

$(a \cos t, b \operatorname{sen} t)$,

$0 \leq t \leq 2\pi$

)*+,)))- !) ! 2 ! ! " / ! < /"
' (0 ($ 1 ! 0
' ($ (2! ! ! !

En los ejercicios 16–20 determine el círculo osculador de la curva dada en el (los) punto(s) indicado(s).

16. $y = x^2$, en el origen de coordenadas.

17. $y = \cos x$, en el punto $(0, 1)$.

18. $y = x^3$, en el punto $(1, 1)$ y en el punto $(-1, -1)$.

19. $y = \dfrac{1}{1+x^2}$, en el punto $(0, 1)$.

20. $y = e^{-x^2}$, en el punto $(0, 1)$. (Compare con los resultados de los ejercicios 16 y 19).

21. Sea $F\colon U \subseteq \mathbb{R}^2 \to \mathbb{R}$ una función de clase $\mathscr{C}^2$ tal que grad $F(x, y) \neq 0$ $\forall (x, y) \in U$. Si $c \in \mathbb{R}$ pertenece al rango de F, la expresión $F(x, y) = c$ define una curva en el plano cuya curvatura en un punto (x, y) de ella (demuestre que) es igual a

$$k = \frac{-\left(\dfrac{\partial F}{\partial y}\right)^2 \dfrac{\partial^2 F}{\partial x^2} + 2\dfrac{\partial F}{\partial x}\dfrac{\partial F}{\partial y}\dfrac{\partial^2 F}{\partial x \partial y} - \left(\dfrac{\partial F}{\partial x}\right)^2 \dfrac{\partial^2 F}{\partial y^2}}{\|\operatorname{grad} F\|^3}$$

22. Se llama *evoluta* de una curva $\mathbf{f}\colon I \subseteq \mathbb{R} \to \mathbb{R}^2$ a la curva que describen los centros de curvatura de $\mathbf{f}$.

a. Determine la evoluta de la curva $\mathbf{f}\colon \mathbb{R} \to \mathbb{R}^2$, $\mathbf{f}(t) = (t, t^2)$.

b. Determine la evoluta de la curva $\mathbf{f}\colon \mathbb{R} \to \mathbb{R}^2$, $\mathbf{f}(t) = (t, t^3)$.

(*) 5.8 Curvas paralelas

El contenido de esta sección es (con algunas modificaciones de estilo) un artículo escrito independientemente del libro (referencia [PiIII]), y motivado por una pregunta "inocente" de un estudiante de ingeniería, cuya respuesta resultó ser bastante poco inocente (este es el tipo de preguntas interesantes en matemáticas: de contenido profundo y de formulación sencilla). La pregunta fue: "¿es la curva paralela a una parábola también una parábola?". Más bien, supongamos que en cada punto de la parábola $y = x^2$ nos alejamos de ella sobre la recta normal una distancia r constante. Formamos así una nueva curva "paralela" a $y = x^2$. La figura que queda tiene al menos todo el aspecto de una nueva parábola. ¿Esta nueva curva es efectivamente una parábola? Al intentar contestar esta pregunta surgieron una gran cantidad de ideas interesantes que culminaron en lo que ahora se presenta en esta sección. La parte de la matemática que más intervino en la respuesta es la que hemos desarrollado en el presente capítulo (geometría diferencial de curvas en el plano).

En realidad, el problema que abordaremos es más ambicioso que el planteado inicialmente. Estudiaremos curvas que tienen la propiedad de mantenerse a una distancia constante de una curva dada (medida sobre la normal de esta curva en cada punto), llamadas "curvas paralelas" (a la curva dada). Exploraremos las relaciones que existen entre algunas propiedades (como la regularidad) y aspectos geométricos (como curvatura, longitud de la curva y área encerrada) de la curva dada y las correspondientes de sus curvas paralelas. Se presentan —a manera de ejemplos ilustrativos—, los estudios correspondientes a las curvas paralelas a elipses y a parábolas.

Las curvas que aparecen en esta sección se suponen infinitamente derivables (de clase $\mathscr{C}^\infty$), de modo que "curva regular" significa una curva de clase $\mathscr{C}^\infty$ cuyo vector velocidad nunca es nulo.

Consideremos la curva regular $\boldsymbol{\alpha}: I \to \mathbb{R}^2$, $\boldsymbol{\alpha}(t) = (x(t), y(t))$. Sea $\mathbf{n}(t)$ el vector normal unitario obtenido al girar $\boldsymbol{\alpha}'(t)$ (normalizado) un ángulo de $\pi/2$ en dirección antihoraria. Sea $\boldsymbol{\beta}: I \to \mathbb{R}^2$ la curva definida como $\boldsymbol{\beta}(t) = \boldsymbol{\alpha}(t) + r\mathbf{n}(t)$, donde r es un número real no nulo dado. Obsérvese que para cada $t \in I$ se tiene que $\|\boldsymbol{\beta}(t) - \boldsymbol{\alpha}(t)\| = |r|$, es decir $\boldsymbol{\beta}$, es una curva que se mantiene a una distancia constante —igual a $|r|$— de $\boldsymbol{\alpha}$. LLamaremos a $\boldsymbol{\beta}$ "la curva r-paralela a $\boldsymbol{\alpha}$". Se entiende que r puede ser positivo o negativo, (por razones obvias se descarta el caso $r = 0$).

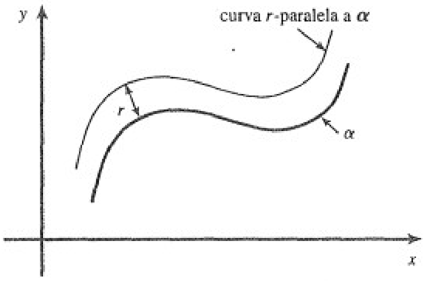

Figura 1. Curva paralela a $\boldsymbol{\alpha}$.

Comencemos por hacer explícitas las funciones coordenadas de la curva $\boldsymbol{\beta}$. Como $\boldsymbol{\alpha}'(t) = (x'(t), y'(t))$, se tiene que el vector $\mathbf{n}(t)$ es

$$\mathbf{n}(t) = \frac{1}{\|\boldsymbol{\alpha}'(t)\|}(-y'(t), x'(t))$$

de modo que

$$\boldsymbol{\beta}(t) = \boldsymbol{\alpha}(t) + r\mathbf{n}(t) = \left(x(t) - \frac{ry'(t)}{\|\boldsymbol{\alpha}'(t)\|}, y(t) + \frac{rx'(t)}{\|\boldsymbol{\alpha}'(t)\|}\right) \tag{1}$$

Obtengamos una expresión para el vector velocidad $\boldsymbol{\beta}'(t)$. Se tiene, derivando, (haciendo uso de que $(\|\boldsymbol{\rho}(t)\|)' = (\boldsymbol{\rho}(t) \cdot \boldsymbol{\rho}'(t))/\|\boldsymbol{\rho}(t)\|$ en donde $'$ denota derivación respecto de t de la norma del vector $\boldsymbol{\rho}(t)$) lo siguiente

$$\begin{aligned}\boldsymbol{\beta}'(t) = \Bigg(&\|x'(t)\| - r\frac{\|\boldsymbol{\alpha}'(t)\|y''(t) - y'(t)(\boldsymbol{\alpha}'(t) \cdot \boldsymbol{\alpha}''(t))/\|\boldsymbol{\alpha}'(t)\|}{\|\boldsymbol{\alpha}'(t)\|^2}, \\ &y'(t) + r\frac{\|\boldsymbol{\alpha}'(t)\|x''(t) - x'(t)(\boldsymbol{\alpha}'(t) \cdot \boldsymbol{\alpha}''(t))/\|\boldsymbol{\alpha}'(t)\|}{\|\boldsymbol{\alpha}'(t)\|^2}\Bigg)\end{aligned}$$

Simplificando nos queda

$$\boldsymbol{\beta}'(t) = \left(x'(t) - r\frac{x'(t)(x'(t)y''(t) - x''(t)y'(t))}{\|\boldsymbol{\alpha}'(t)\|^3}, y'(t) + r\frac{y'(t)(x''(t)y'(t) - x'(t)y''(t))}{\|\boldsymbol{\alpha}'(t)\|^3}\right)$$

o sea

$$\boldsymbol{\beta}'(t) = \left(1 - r\frac{x'(t)y''(t) - x''(t)y'(t)}{\|\boldsymbol{\alpha}'(t)\|^3}\right)(x'(t), y'(t))$$

Recordando que la curvatura con signo de la curva $\boldsymbol{\alpha}$, que denotaremos por $k_{\alpha}(t)$, está dada por

$$k_{\alpha}(t) = \frac{x'(t)y''(t) - x''(t)y'(t)}{\|\boldsymbol{\alpha}'(t)\|^3}$$

nos queda finalmente la fórmula para $\boldsymbol{\beta}'(t)$ como

$$\boxed{\boldsymbol{\beta}'(t) = (1 - rk_{\alpha}(t))\boldsymbol{\alpha}'(t)} \tag{2}$$

Siendo la curva $\boldsymbol{\alpha}$ regular, nos preguntamos si la curva r-paralela $\boldsymbol{\beta}$ también lo es. Con la fórmula que obtuvimos para $\boldsymbol{\beta}'(t)$ es fácil contestar esta pregunta.

> **Teorema 5.8.1** Sea $\boldsymbol{\alpha}: I \to \mathbb{R}^2$ una curva regular, y sea $\boldsymbol{\beta}: I \to \mathbb{R}^2$ la curva r-paralela a $\boldsymbol{\alpha}$. La curva $\boldsymbol{\beta}$ es regular si y sólo si r^{-1} no pertenece al rango de $k_{\alpha}(t)$ $(t \in I)$.

Demostración Puesto que $\|\boldsymbol{\beta}'(t)\| = |1 - rk_{\alpha}(t)|\|\boldsymbol{\alpha}'(t)\|$, y por hipótesis $\|\boldsymbol{\alpha}'(t)\| \neq 0$, se tiene que $\boldsymbol{\beta}'(t) \neq 0$ si y sólo si $1 - rk_{\alpha}(t) \neq 0$ $\forall t \in I$, lo cual significa que $r^{-1} \neq k_{\alpha}(t)$ $\forall t \in I$, es decir que $r^{-1} \notin$ rango de $k_{\alpha}(t)$. Q.E.D.

Por ejemplo, si consideramos la curva regular $\boldsymbol{\alpha}: [a, b] \to \mathbb{R}^2$, definida en el intervalo compacto $[a, b]$ de $\mathbb{R}$, la función curvatura $k_{\alpha}(t)$, $t \in [a, b]$ alcanzará su máximo absoluto dentro del intervalo $[a, b]$, de modo que si tomamos $r = \pm\epsilon$, $(\epsilon > 0)$, con

$$\epsilon^{-1} > \text{máx}_{t\in[a,b]} |k_{\alpha}(t)|$$

las curvas $\boldsymbol{\beta}_1(t) = \boldsymbol{\alpha}(t) + \epsilon\mathbf{n}(t)$ y $\boldsymbol{\beta}_2(t) = \boldsymbol{\alpha}(t) - \epsilon\mathbf{n}(t)$, que son curvas regulares ϵ-paralelas a $\boldsymbol{\alpha}$, (y que cada una de ellas se encuentra en "lados distintos de $\boldsymbol{\alpha}(t)$", es decir, $\boldsymbol{\beta}_1$ se encuentra en el lado marcado por la dirección del vector $\mathbf{n}(t)$, mientras que $\boldsymbol{\beta}_2$ se encuentra en el lado opuesto) constituyen la frontera de la vecindad abierta

$$V_{\epsilon}(\boldsymbol{\alpha}) = \{(x, y) \in \mathbb{R}^2 \mid \|(x, y) - \boldsymbol{\beta}_i(t)\| < \epsilon,\ i = 1, 2, t \in I\}$$

la cual es un ejemplo de un tipo de vecindades abiertas de curvas llamadas "vecindades tubulares".

A manera de ejemplo, consideremos la curva $\boldsymbol{\alpha}: [0, 2\pi] \to \mathbb{R}^2$, dada por $\boldsymbol{\alpha}(t) = (t, \operatorname{sen} t)$. La curva r-paralela a $\boldsymbol{\alpha}$ es $\boldsymbol{\beta}: [0, 2\pi] \to \mathbb{R}^2$

$$\boldsymbol{\beta}(t) = \left(t - (r\cos t)/(1 + \cos^2 t)^{1/2}, \operatorname{sen} t + r/(1 + \cos^2 t)^{1/2}\right)$$

La curvatura de $\boldsymbol{\alpha}$ en t es

$$k_{\alpha}(t) = -(\operatorname{sen} t)/(1 + \cos^2 t)^{3/2}$$

que alcanza su máximo en $t = 3\pi/2$, el cual vale $k_{\alpha}(3\pi/2) = 1$. Entonces, tomando $\epsilon > 0$ de modo que $\epsilon^{-1} > 1$, es decir que $\epsilon < 1$, tenemos que las curvas regulares ϵ-paralelas $\boldsymbol{\beta}(t)$ dadas en

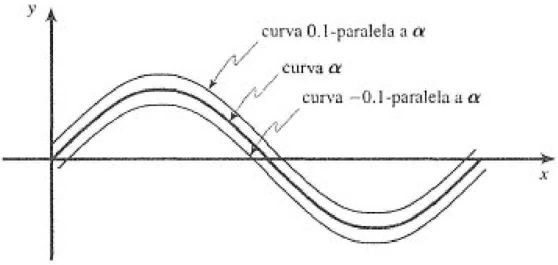

Figura 2. Curvas ±0.1-paralelas a la gráfica de $y = \operatorname{sen} x$, $0 \leq x \leq 2\pi$.

la fórmula anterior con $r = \pm\epsilon$, constituyen la frontera de una vecindad tubular de $\boldsymbol{\alpha}$. La siguiente figura muestra el caso en que $\epsilon = 0.1$.

Veamos ahora cómo están relacionadas las curvaturas k_α, de la curva $\boldsymbol{\alpha}$, con la de su paralela $\boldsymbol{\beta}$, k_β. Tomemos $r^{-1} \notin$ rango de $k_\alpha(t)$, de modo que la curva r-paralela $\boldsymbol{\beta}$ es regular. Se tiene, derivando la expresión (2), que

$$\boldsymbol{\beta}''(t) = (1 - rk_\alpha(t))\boldsymbol{\alpha}''(t) - rk'_\alpha(t)\boldsymbol{\alpha}'(t)$$

de modo que

$$\boldsymbol{\beta}'(t) \times \boldsymbol{\beta}''(t) = \big((1 - rk_\alpha(t))\boldsymbol{\alpha}'(t)\big) \times \big((1 - rk_\alpha(t))\boldsymbol{\alpha}''(t) - rk'_\alpha(t)\boldsymbol{\alpha}'(t)\big) = (1 - rk_\alpha(t))^2\boldsymbol{\alpha}'(t) \times \boldsymbol{\alpha}''(t)$$

y entonces

$$k_\beta(t) = \frac{\|\boldsymbol{\beta}'(t) \times \boldsymbol{\beta}''(t)\|}{\|\boldsymbol{\beta}'(t)\|^3} = \frac{(1 - rk_\alpha(t))^2\|\boldsymbol{\alpha}'(t) \times \boldsymbol{\alpha}''(t)\|}{|1 - rk_\alpha(t)|^3\|\boldsymbol{\alpha}'(t)\|^3}$$

Por las hipótesis hechas sobre r se tiene que $1 - rk_\alpha(t) > 0\ \forall t \in I$, o bien $1 - rk_\alpha(t) < 0\ \forall t \in I$, y entonces la fórmula anterior queda como

$$\boxed{k_\beta(t) = \pm\frac{|k_\alpha(t)|}{1 - rk_\alpha(t)}}$$

con los signos correspondientes a cada uno de los casos mencionados. Con esta fórmula es fácil ver que si la curva $\boldsymbol{\alpha}$ tiene un vértice en t_0 (es decir que $k'_\alpha(t_0) = 0$) entonces su curva r-paralela $\boldsymbol{\beta}$ (regular) tendrá también un vértice en ese punto. En efecto, se tiene que

$$k_\beta(t) = \pm(1 - rk_\alpha(t))^{-2}k'_\beta$$

de donde se ve, de hecho, que los vértices de $\boldsymbol{\alpha}$ y de $\boldsymbol{\beta}$ coinciden.

Consideremos, a manera de ejemplo la elipse $\boldsymbol{\alpha}\colon [0, 2\pi] \to \mathbb{R}^2$ dada por $\boldsymbol{\alpha}(t) = (a\cos t, b\operatorname{sen} t)$ donde $a > b > 0$. La curvatura $k_\alpha(t)$ es

$$k_\alpha(t) = \frac{ab}{(a^2\operatorname{sen}^2 t + b^2\cos^2 t)^{3/2}}$$

Los extremos absolutos de esta función en el intervalo $[0, 2\pi]$ se alcanzan en $t = 0, \pi$ (máximo absoluto) y en $t = \pi/2, 3\pi/2$ (mínimo absoluto), los cuales valen $k_{\boldsymbol{\alpha}}(0) = k_{\boldsymbol{\alpha}}(\pi) = a/b^2$ y $k_{\boldsymbol{\alpha}}(\pi/2) = k_{\boldsymbol{\alpha}}(3\pi/2) = b/a^2$. Tomando entonces r tal de modo $r^{-1} \notin [b/a^2, a/b^2]$, la curva r-paralela a $\boldsymbol{\alpha}$, que según (1) es

$$\boldsymbol{\beta}(t) = \left(\cos t\left(a - \frac{rb}{(a^2 \operatorname{sen}^2 t + b^2 \cos^2 t)^{1/2}}\right), \operatorname{sen} t\left(b - \frac{ra}{(a^2 \operatorname{sen}^2 t + b^2 \cos^2 t)^{1/2}}\right)\right)$$

será una curva regular con curvatura

$$k_{\boldsymbol{\beta}}(t) = \pm\frac{k_{\boldsymbol{\alpha}}(t)}{1 - rk_{\boldsymbol{\alpha}}(t)} = \pm\frac{ab}{(a^2 \operatorname{sen}^2 t + b^2 \cos^2 t)^{3/2} - rab}$$

donde el signo más corresponde al caso en el que $r < b^2/a$ (pues en este caso $r^{-1} > a/b^2 =$ máximo absoluto de $k_{\boldsymbol{\alpha}}(t)$ en $[0, 2\pi]$, y por lo tanto se tiene que $1 - rk_{\boldsymbol{\alpha}}(t) > 0\ \forall t \in [0, 2\pi]$), y el signo menos corresponde al caso en que $r > a^2/b$ (pues en este caso $r^{-1} < b/a^2 =$ mínimo absoluto de $k_{\boldsymbol{\alpha}}(t)$ en $[0, 2\pi]$, y por lo tanto se tiene que $1 - rk_{\boldsymbol{\alpha}}(t) < 0\ \forall t \in [0, 2\pi]$). La siguiente figura muestra el caso en que $a = 2$, $b = 1$, con $\boldsymbol{\beta}_1$ una curva r-paralela a $\boldsymbol{\alpha}$ siendo $r = 0.25 < 0.5 = b^2/a$, y $\boldsymbol{\beta}_2$ una curva r-paralela a $\boldsymbol{\alpha}$ con $r = 5 > 4 = a^2/b$. Obsérvese que la curva $\boldsymbol{\alpha}$ tiene curvatura positiva, al igual que las curvas $\boldsymbol{\beta}_1$ y $\boldsymbol{\beta}_2$. Sin embargo, el vector velocidad de $\boldsymbol{\beta}_1$ siempre tiene la misma dirección que el vector velocidad de $\boldsymbol{\alpha}$, en tanto que el de $\boldsymbol{\beta}_2$ siempre tiene la dirección contraria al correspondiente de $\boldsymbol{\alpha}$ (hecho que se ve fácilmente de la fórmula (2)). Nótese también que cualquier curva r-paralela a $\boldsymbol{\alpha}$ con $r < 0$, es automáticamente regular, pues siendo $k_{\boldsymbol{\alpha}}(t) > 0\ \forall t \in [0, 2\pi]$, se tiene $1 - rk_{\boldsymbol{\alpha}}(t) > 0\ \forall t \in [0, 2\pi]$ (figura 3).

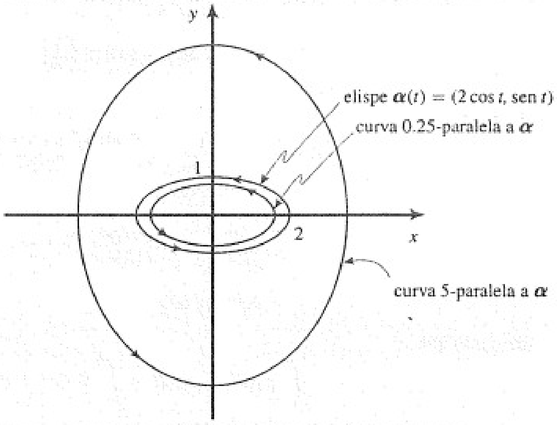
y
elispe $\boldsymbol{\alpha}(t) = (2\cos t, \operatorname{sen} t)$
curva 0.25-paralela a $\boldsymbol{\alpha}$
1
2
x
curva 5-paralela a $\boldsymbol{\alpha}$

Figura 3. Curvas 0.25 y 5-paralelas a la elipse $\boldsymbol{\alpha}_t = (2\cos t, \operatorname{sen} t)$ $0 \le t \le 2\pi$.

Ahora estudiaremos la relación entre la longitud de la curva regular dada $\boldsymbol{\alpha}\colon [a, b] \to \mathbb{R}^2$ con la correspondiente de su curva r-paralela $\boldsymbol{\beta}\colon [a, b] \to \mathbb{R}^2$. Solamente se considerará el caso en que la curva $\boldsymbol{\beta}$ sea regular, es decir, cuando $\boldsymbol{\beta}$ es una curva r-parelela a $\boldsymbol{\alpha}$, con r^{-1} fuera del rango de $k_{\alpha}(t)$ $(t \in [a, b])$.

Recordemos que para la curva regular $\boldsymbol{\alpha}$ se tiene bien definida su longitud (denotada como $L_{[a,b]}(\boldsymbol{\alpha})$) como

$$L_{[a,b]}(\boldsymbol{\alpha}) = \int_a^b \|\boldsymbol{\alpha}'(t)\|\, dt$$

Siendo la curva $\boldsymbol{\beta}$ también regular, su longitud se calcula como

$$L_{[a,b]}(\boldsymbol{\beta}) = \int_a^b \|\boldsymbol{\beta}'(t)\|\, dt$$

o bien, puesto que $\boldsymbol{\beta}'(t) = (1 - rk_{\alpha}(t))\boldsymbol{\alpha}'(t)$, se tiene

$$\begin{aligned}
L_{[a,b]}(\boldsymbol{\beta}) &= \int_a^b |1 - rk_{\alpha}(t)|\|\boldsymbol{\alpha}'(t)\|\, dt \\
&= \int_a^b \big(\operatorname{sgn}(1 - rk_{\alpha}(t))\big)(1 - rk_{\alpha}(t))\|\boldsymbol{\alpha}'(t)\|\, dt \\
&= \operatorname{sgn}(1 - rk_{\alpha}(t))\left[\int_a^b \|\boldsymbol{\alpha}'(t)\|\, dt - r\int_a^b k_{\alpha}(t)\|\boldsymbol{\alpha}'(t)\|\, dt\right] \\
&= \operatorname{sgn}(1 - rk_{\alpha}(t))\left[L_{[a,b]} - r\int_a^b k_{\alpha}(t)\|\boldsymbol{\alpha}'(t)\|\, dt\right]
\end{aligned}$$

Por otra parte, si llamamos $\theta(t)$ al ángulo formado por el vector velocidad $\boldsymbol{\alpha}'(t)$ (que por hipótesis es no nulo para toda $t \in [a, b]$) con la parte positiva del eje de las abscisas $(0 \leq \theta(t) < \pi)$, se tiene

$$\theta(t) = \arctan \frac{y'(t)}{x'(t)}$$

de donde

$$\begin{aligned}
\theta'(t) &= \frac{1}{1 + \left(\dfrac{y'(t)}{x'(t)}\right)^2}\left(\frac{x'(t)y''(t) - y'(t)x''(t)}{(x'(t))^2}\right) \\
&= \frac{x'(t)y''(t) - y'(t)x''(t)}{((x'(t))^2 + (y'(t))^2)^{3/2}}\big(x'(t)^2 + (y'(t))^2\big)^{1/2} \\
&= k_{\alpha}(t)\|\boldsymbol{\alpha}'(t)\|
\end{aligned}$$

de modo que

$$\int_a^b k_{\alpha}(t)\|\boldsymbol{\alpha}'(t)\|\, dt = \int_a^b \theta'(t)dt = \theta(b) - \theta(a)$$

y por lo tanto finalmente podemos escribir la fórmula de $L_{[a,b]}(\boldsymbol{\beta})$ como

$$\boxed{L_{[a,b]}(\boldsymbol{\beta}) = \operatorname{sgn}(1 - rk_{\alpha}(t))\big[L_{[a,b]}(\boldsymbol{\alpha}) - r(\theta(b) - \theta(a))\big]} \tag{3}$$

Más aún, si la curva $\boldsymbol{\alpha}$ es cerrada, la diferencia $\theta(b) - \theta(a)$ es un múltiplo de 2π. En forma más precisa se tiene que $\theta(b) - \theta(a) = 2\pi I$, en I se llama índice de rotación de la curva (que de manera intuitiva lo pensamos como la cantidad de vueltas completas que da el vector tangente sobre la curva). Así pues, en este caso la fórmula (3) se ve como

$$L_{[a,b]}(\boldsymbol{\beta}) = \operatorname{sgn}(1 - rk_{\boldsymbol{\alpha}}(t))\left[L_{[a,b]}(\boldsymbol{\alpha}) - 2\pi r|I|\right]$$

de modo que si $1 - rk_{\boldsymbol{\alpha}}(t) > 0\ \forall t \in [a, b]$ (situación que se presenta si $r^{-1} > \text{máx}_{t\in[a,b]} |k_{\boldsymbol{\alpha}}(t)|$, o bien, en particular, si la curva $\boldsymbol{\alpha}$ es una curva simple convexa positivamente orientada y $r < 0$), la longitud de la curva $\boldsymbol{\beta}$, r-paralela a $\boldsymbol{\alpha}$ es

$$L_{[a,b]}(\boldsymbol{\beta}) = L_{[a,b]}(\boldsymbol{\alpha}) - 2\pi r|I|$$

en tanto que si $1 - rk_{\boldsymbol{\alpha}}(t) < 0\ \forall t \in [a, b]$ (por ejemplo, si $r^{-1} < \text{mín}_{t\in[a,b]} |k_{\boldsymbol{\alpha}}(t)|$), la curva regular $\boldsymbol{\beta}$, r-paralela a $\boldsymbol{\alpha}$ tiene por longitud

$$L_{[a,b]}(\boldsymbol{\beta}) = -L_{[a,b]}(\boldsymbol{\alpha}) + 2\pi r|I|$$

Retomando el ejemplo de la elipse $\boldsymbol{\alpha}: [0, 2\pi] \to \mathbb{R}^2$, $\boldsymbol{\alpha}(t) = (2\cos t, \operatorname{sen} t)$, habíamos ya considerado curvas r-paralelas a ella con $r = 0.25$ (caso en el que $1 - rk_{\boldsymbol{\alpha}}(t) > 0\ \forall t \in [0, 2\pi]$), que llamamos $\boldsymbol{\beta}_1$, y con $r = 5$ (caso en el que $1 - rk_{\boldsymbol{\alpha}}(t) < 0\ \forall t \in [0, 2\pi]$), que llamamos $\boldsymbol{\beta}_2$ (ver figura 3). Tanto $\boldsymbol{\beta}_1$ como $\boldsymbol{\beta}_2$ son regulares y para calcular su longitud podemos usar las fórmulas recientemente obtenidas. Se tiene

$$\begin{aligned} L_{[0,2\pi]}(\boldsymbol{\beta}_1) &= L_{[0,2\pi]}(\boldsymbol{\alpha}) - 2\pi r = 9.6885 - 0.5\pi = 8.1177 \\ L_{[0,2\pi]}(\boldsymbol{\beta}_2) &= -L_{[0,2\pi]}(\boldsymbol{\alpha}) + 2\pi r = -9.6885 + 10\pi = 21.7275 \end{aligned}$$

Consideremos ahora una curva regular cerrada simple $\boldsymbol{\alpha}: [a, b] \to \mathbb{R}^2$ (es decir, se tiene $\boldsymbol{\alpha}^{(j)}(a) = \boldsymbol{\alpha}^{(j)}(b)$, $j = 0, 1, 2, \ldots$, y $\boldsymbol{\alpha}(t_1) \neq \boldsymbol{\alpha}(t_2)$ para toda pareja $t_1, t_2 \in [a, b)$, $t_1 \neq t_2$). Así si tiene sentido hablar del "área encerrada por la curva $\boldsymbol{\alpha}$". De hecho, con ayuda del Teorema de Green que estudiaremos en el capítulo 7, podemos calcular dicha área (que denotaremos por $A(\boldsymbol{\alpha})$) como (con la curva $\boldsymbol{\alpha}(t) = (x(t), y(t))$ positivamente orientada)

$$A(\boldsymbol{\alpha}) = \frac{1}{2}\int_a^b [x(t)y'(t) - x'(t)y(t)]\, dt$$

(en este momento, aceptemos que esta fórmula proporciona efectivamente el área mencionada). Vamos ahora a estudiar cómo está relacionada esta área con la correspondiente de su curva r-paralela (cuando tenga sentido hablar del área de esta curva). Primero obsérvese que si la curva $\boldsymbol{\alpha}$ es cerrada entonces la curva $\boldsymbol{\beta}$ será también cerrada (hecho que se ve fácilmente por la manera como está definida $\boldsymbol{\beta}$). La propiedad de regularidad de la curva $\boldsymbol{\alpha}$, como ya se ha visto, se puede perder en la curva $\boldsymbol{\beta}$; esto depende de la existencia de raíces de la ecuación $1 - rk_{\boldsymbol{\alpha}}(t) = 0$ en $[a, b]$. Este hecho, sin embargo, no nos impide usar la fórmula del Teorema de Green para calcular el área encerrada por la curva, pues ésta sigue siendo válida para curvas regulares por secciones en $[a, b]$. Entonces la propiedad importante que debe conservar la curva $\boldsymbol{\beta}$ para poder seguir hablando del área que ésta encierra es la de ser una curva simple (sin autointersecciones). Notemos también que la orientación positiva de $\boldsymbol{\alpha}$ puede perderse en $\boldsymbol{\beta}$, aun conservando esta última la propiedad de ser simple (ver el ejemplo de la curva $\boldsymbol{\beta}$, 1.5-paralela a la elipse $\boldsymbol{\alpha}(t) = (2\cos t, \operatorname{sen} t)$, que tiene orientación negativa). De cualquier modo, este conflicto se arregla fácilmente con una reparametrización de la curva que invierta su orientación.

Partamos entonces de una curva $\boldsymbol{\alpha}\colon [a, b] \to \mathbb{R}^2$, $\boldsymbol{\alpha}(t) = (x(t), y(t))$ regular, cerrada, simple y positivamente orientada y consideremos la curva $\boldsymbol{\beta}\colon [a, b] \to \mathbb{R}^2$, r-paralela a $\boldsymbol{\alpha}$, con r de modo que $\boldsymbol{\beta}$ sea simple. Supongamos también, por el momento, que $\boldsymbol{\beta}$ queda orientada positivamente. Escribiendo

$$\boldsymbol{\beta}(t) = (x_{\boldsymbol{\beta}}(t), y_{\boldsymbol{\beta}}(t)), \quad \boldsymbol{\beta}'(t) = (x'_{\boldsymbol{\beta}}(t), y'_{\boldsymbol{\beta}}(t))$$

donde entonces

$$\begin{aligned} x_{\boldsymbol{\beta}}(t) &= x(t) - \frac{ry'(t)}{\|\boldsymbol{\alpha}'(t)\|}, \qquad & y_{\boldsymbol{\beta}}(t) &= y(t) + \frac{rx'(t)}{\|\boldsymbol{\alpha}'(t)\|}, \\ x'_{\boldsymbol{\beta}}(t) &= (1 - rk_{\boldsymbol{\alpha}}(t))x'(t) & y'_{\boldsymbol{\beta}}(t) &= (1 - rk_{\boldsymbol{\alpha}}(t))y'(t) \end{aligned}$$

tenemos que el área encerrada por $\boldsymbol{\beta}$, que denotaremos por $A(\boldsymbol{\beta})$ es

$$\begin{aligned} A(\boldsymbol{\beta}) &= \frac{1}{2}\int_a^b \Big[x_{\boldsymbol{\beta}}(t)y'_{\boldsymbol{\beta}}(t) - x'_{\boldsymbol{\beta}}(t)y_{\boldsymbol{\beta}}(t)\Big]\,dt \\ &= \frac{1}{2}\int_a^b \left[\left(x(t) - \frac{ry'(t)}{\|\boldsymbol{\alpha}'(t)\|}\right)((1 - rk_{\boldsymbol{\alpha}}(t))y'(t)) - \left(y(t) - \frac{rx'(t)}{\|\boldsymbol{\alpha}'(t)\|}\right)((1 - rk_{\boldsymbol{\alpha}}(t))x'(t))\right] dt \\ &= \frac{1}{2}\int_a^b (1 - rk_{\boldsymbol{\alpha}}(t))\big(x(t)y'(t) - y(t)x'(t) - r\|\boldsymbol{\alpha}'(t)\|\big)\,dt \\ &= \frac{1}{2}\int_a^b \big(x(t)y'(t) - y(t)x'(t)\big)\,dt - \frac{1}{2}r\int_a^b \|\boldsymbol{\alpha}'(t)\|dt \\ &\quad - \frac{1}{2}r\int_a^b \big(x(t)y'(t) - y(t)x'(t)\big)k_{\boldsymbol{\alpha}}(t)\,dt + \frac{1}{2}r^2\int_a^b k_{\boldsymbol{\alpha}}(t)\|\boldsymbol{\alpha}'(t)\|\,dt \end{aligned}$$

El primero, segundo y cuarto sumandos son expresiones fácilmente reconocibles. En efecto, el primero no es más que el área encerrada por la curva $\boldsymbol{\alpha}$. La integral que aparece en el segundo sumando es la longitud de la curva $\boldsymbol{\alpha}$ (que denotaremos por $L(\boldsymbol{\alpha})$), y, por último, la integral que aparece en el cuarto sumando es (como ya habíamos observado en los cálculos correspondientes a la longitud de la curva paralela $\boldsymbol{\beta}$)

$$\int_a^b k_{\boldsymbol{\alpha}}(t)\|\boldsymbol{\alpha}'(t)\|\,dt = \theta(b) - \theta(a) = 2\pi$$

pues la curva $\boldsymbol{\alpha}$ es simple y positivamente orientada (su índice de rotación es $+1$). Entonces la última expresión nos queda como

$$A(\boldsymbol{\beta}) = A(\boldsymbol{\alpha}) - 0.5rL(\boldsymbol{\alpha}) + \pi r^2 - 0.5r\int_a^b \big(x(t)y'(t) - y(t)x'(t)\big)k_{\boldsymbol{\alpha}}(t)dt$$

Trabajemos con la integral que aparece en esta fórmula. Llamémosla ξ. Se tiene que ejemplo de la curva $\boldsymbol{\beta}$, 1.5-paralela a la elipse $\boldsymbol{\alpha}(t) = (2\cos t, \operatorname{sen} t)$, que tiene orientación negativa). De cualquier modo, este conflicto se arregla fácilmente con una reparametrización de la curva que invierta su orientación.

Partamos entonces de una curva $\boldsymbol{\alpha}\colon [a, b] \to \mathbb{R}^2$, $\boldsymbol{\alpha}(t) = (x(t), y(t))$ regular, cerrada, simple y positivamente orientada y consideremos la curva $\boldsymbol{\beta}\colon [a, b] \to \mathbb{R}^2$, r-paralela a $\boldsymbol{\alpha}$, con r de modo que $\boldsymbol{\beta}$ sea simple. Supongamos también, por el momento, que $\boldsymbol{\beta}$ queda orientada positivamente. Escribiendo

$$\boldsymbol{\beta}(t) = (x_{\beta}(t), y_{\beta}(t)), \quad \boldsymbol{\beta}'(t) = (x'_{\beta}(t), y'_{\beta}(t))$$

donde entonces

$$x_{\beta}(t) = x(t) - \frac{ry'(t)}{\|\boldsymbol{\alpha}'(t)\|}, \qquad y_{\beta}(t) = y(t) + \frac{rx'(t)}{\|\boldsymbol{\alpha}'(t)\|},$$
$$x'_{\beta}(t) = (1 - rk_{\alpha}(t))x'(t) \qquad y'_{\beta}(t) = (1 - rk_{\alpha}(t))y'(t)$$

tenemos que el área encerrada por $\boldsymbol{\beta}$, que denotaremos por $A(\boldsymbol{\beta})$ es

$$\begin{aligned}
A(\boldsymbol{\beta}) &= \frac{1}{2}\int_a^b \left[x_{\beta}(t)y'_{\beta}(t) - x'_{\beta}(t)y_{\beta}(t)\right] dt \\
&= \frac{1}{2}\int_a^b \left[\left(x(t) - \frac{ry'(t)}{\|\boldsymbol{\alpha}'(t)\|}\right)\left((1 - rk_{\alpha}(t))y'(t)\right) - \left(y(t) - \frac{rx'(t)}{\|\boldsymbol{\alpha}'(t)\|}\right)\left((1 - rk_{\alpha}(t))x'(t)\right)\right] dt \\
&= \frac{1}{2}\int_a^b (1 - rk_{\alpha}(t))\big(x(t)y'(t) - y(t)x'(t) - r\|\alpha'(t)\|\big)\, dt \\
&= \frac{1}{2}\int_a^b \big(x(t)y'(t) - y(t)x'(t)\big)\, dt - \frac{1}{2}r\int_a^b \|\alpha'(t)\| dt \\
&\quad - \frac{1}{2}r\int_a^b \big(x(t)y'(t) - y(t)x'(t)\big)k_{\alpha}(t)\, dt + \frac{1}{2}r^2\int_a^b k_{\alpha}(t)\|\alpha'(t)\|\, dt
\end{aligned}$$

El primero, segundo y cuarto sumandos son expresiones fácilmente reconocibles. En efecto, el primero no es más que el área encerrada por la curva α. La integral que aparece en el segundo sumando es la longitud de la curva α (que denotaremos por $L(\alpha)$), y, por último, la integral que aparece en el cuarto sumando es (como ya habíamos observado en los cálculos correspondientes a la longitud de la curva paralela $\boldsymbol{\beta}$)

$$\int_a^b k_{\alpha}(t)\|\boldsymbol{\alpha}'(t)\|\, dt = \theta(b) - \theta(a) = 2\pi$$

pues la curva $\boldsymbol{\alpha}$ es simple y positivamente orientada (su índice de rotación es $+1$). Entonces la última expresión nos queda como

$$A(\boldsymbol{\beta}) = A(\boldsymbol{\alpha}) - 0.5rL(\boldsymbol{\alpha}) + \pi r^2 - 0.5r\int_a^b \big(x(t)y'(t) - y(t)x'(t)\big)k_{\alpha}(t)dt$$

Trabajemos con la integral que aparece en esta fórmula. Llamémosla ξ. Se tiene que

Para la curva 5-paralela a la elipse tenemos

$$A(\boldsymbol{\beta}) = 2\pi - 5(9.6885) + \pi(5)^2 = 36.3807$$

y para la curva 1.5-paralela a la elipse se tiene

$$A(\boldsymbol{\beta}) = 2\pi - 1.5(9.6885) + \pi(1.5)^2 = -1.18095$$

Nótese que el valor del área encerrada por esta curva 1.5-paralela a la elipse es negativo, pues, como ya habíamos comentado, la orientación que tiene esta curva es negativa. Es claro, sin embargo, que en este caso el valor absoluto de nuestro resultado es el área procurada.

En el caso particular de la elipse $\boldsymbol{\alpha}\colon [0, 2\pi] \to \mathbb{R}^2$, dada por $\boldsymbol{\alpha}(t) = (a\cos t, b\,\mathrm{sen}\,t)$, con $a > b > 0$, podemos ser más explícitos en cuanto a la utilización de la fórmula del área encerrada por sus curvas r-paralelas $\boldsymbol{\beta}\colon [0, 2\pi] \to \mathbb{R}^2$

$$\boldsymbol{\beta}(t) = \left(\cos t\left(a - \frac{rb}{(a^2\,\mathrm{sen}^2\,t + b^2\cos^2 t)^{1/2}}\right), \mathrm{sen}\,t\left(b - \frac{ra}{(a^2\,\mathrm{sen}^2\,t + b^2\cos^2 t)^{1/2}}\right)\right)$$

Estudiemos las intersecciones de esta curva con el eje x. Haciendo

$$y(t) = \mathrm{sen}\,t\left(b - \frac{ra}{(a^2\,\mathrm{sen}^2\,t + b^2\cos^2 t)^{1/2}}\right) = 0$$

Nótese que siempre hay dos raíces de esta ecuación en $[0, 2\pi)$, a saber $t = 0, \pi$. Es decir, la curva $\boldsymbol{\beta}$ corta al eje x al menos en estos dos puntos. Puede, sin embargo, haber otras raíces de la ecuación, provenientes de

$$b - \frac{ra}{(a^2\,\mathrm{sen}^2\,t + b^2\cos^2 t)^{1/2}} = 0$$

Obsérvese que si $r < 0$, estas raíces no existen (las curvas paralelas hacia el exterior de la elipse cortan solamente en $t = 0$ y $t = \pi$ al eje de las x). Tomemos entonces $r > 0$. En este caso se tiene que

$$\mathrm{sen}^2\,t = \frac{r^2a^2 - b^4}{b^2a^2 - b^4}$$

Estos valores de t existen (y son diferentes de 0 y π) si y sólo si

$$0 < \frac{r^2a^2 - b^4}{b^2a^2 - b^4} < 1$$

es decir, si $b^2/a < r < b$. En tal caso, obsérvese que existen dos raíces de la ecuación $y(t) = 0$ (además de las ya mencionadas 0 y π) entre 0 y 2π, a saber

$$t_1 = \mathrm{arcsen}\left(\frac{r^2a^2 - b^4}{b^2a^2 - b^4}\right)^{1/2}, \qquad t_2 = t_1 + \pi$$

y que además se tiene $x_{\boldsymbol{\beta}}(t_1) = x_{\boldsymbol{\beta}}(t_2)$. Esto significa que bajo la condición $b^2/a < r < b$, existe una autointersección de la curva $\boldsymbol{\beta}$, la cual ocurre sobre el eje de las x (en el punto $\boldsymbol{\beta}(t_1) = \boldsymbol{\beta}(t_2)$).

De manera análoga se ve que la curva $\boldsymbol{\beta}$ siempre tiene dos intersecciones con el eje y, en $t = \pi/2$ y en $t = 3\pi/2$, y que puede tener más, provenientes de la ecuación

$$a - \frac{rb}{(a^2 \operatorname{sen}^2 t + b^2 \cos^2 t)^{1/2}} = 0$$

de donde se ve también que de ser $r < 0$, tales nuevas intersecciones (las raíces de esta ecuación) no existen (las curvas paralelas hacia el exterior de la elipse cortan al eje de las y solamente en $t = \pi/2$ y $t = 3\pi/2$). Escribiendo la última ecuación como

$$\cos^2 t = \frac{a^4 - r^2b^2}{a^4 - a^2b^2}$$

vemos que las nuevas raíces existen si y sólo si

$$0 < \frac{a^4 - r^2b^2}{a^4 - a^2b^2} < 1$$

o sea si $a < r < a^2/b$. En tal caso, existen dos raíces, a saber

$$t_1 = \arccos\left(\frac{a^4 - r^2b^2}{a^4 - a^2b^2}\right), \qquad t_2 = t_1 + \pi$$

teniéndose además que $y_{\boldsymbol{\beta}}(t_1) = y_{\boldsymbol{\beta}}(t_2)$. Esto significa entonces que con la condición $a < r < a^2/b$, hay dos intersecciones en el eje y (además de las correspondientes a $t = \pi/2$ y $t = 3\pi/2$) las cuales son autointersecciones de la curva $\boldsymbol{\beta}$.

En conclusión: si $b^2/a < r < b$, la curva $\boldsymbol{\beta}$ tiene autointersecciones (en el eje x); si $a < r < a^2/b$, la curva $\boldsymbol{\beta}$ tiene autointersecciones (en el eje y). Como $a > b$, observamos que en el intervalo $b \leq r \leq a$, la curva $\boldsymbol{\beta}$ no tiene autointersecciones (nótese que en este intervalo la curva $\boldsymbol{\beta}$ no es regular pues, como habíamos ya visto, la región en la cual la curva $\boldsymbol{\beta}$ pierde su regularidad corresponde a valores de r tales que $b^2/a \leq r \leq a^2/b$). Más aún, con el valor de r en el intervalo $[b, a]$, la curva r-paralela $\boldsymbol{\beta}$, siendo cerrada, simple y regular por partes, tiene orientación negativa (un ejemplo de esta situación es la curva 1.5-paralela a la elipse $\boldsymbol{\alpha}(t) = (2\cos t, \operatorname{sen} t)$ que ya ha sido estudiada, pues en este caso $r = 1.5 \in [1, 2] = [b, a]$). A manera de resumen de las variaciones que tienen las propiedades de regularidad, simplicidad y positividad en la orientación de la curva r-paralela $\boldsymbol{\beta}$, al variar r, presentamos el esquema siguiente

$$\text{———}\overset{0}{*}\text{———}\overset{b^2/a}{*}\text{——————}\overset{b}{*}\text{——————}\overset{a}{*}\text{——————}\overset{a^2/b}{*}\text{——————}\!\rightarrow r$$

<table>
<tr><td rowspan="2">La curva es regular (,) y simple (,]. Tiene orientación positiva (,].</td><td colspan="3">La curva no es regular [,]</td><td rowspan="2">La curva es regular (,) y simple [,). Tiene orientación positiva [,).</td></tr>
<tr><td>La curva no es simple (autointersecciones en el eje x) (,)</td><td>La curva es simple. Tiene orientación negativa [,]</td><td>La curva no es simple (autointersecciones en el eje y) (,)</td></tr>
</table>

Las siguientes figuras ilustran cada una de estas situaciones

En consecuencia, la fórmula para calcular el área encerrada por la curva r-paralela $\boldsymbol{\beta}$ a la elipse $\boldsymbol{\alpha}$, tiene sentido cuando $r \notin (b^2/a, b) \cup (a, a^2/b)$. En todos los demás casos la curva $\boldsymbol{\beta}$ es simple,

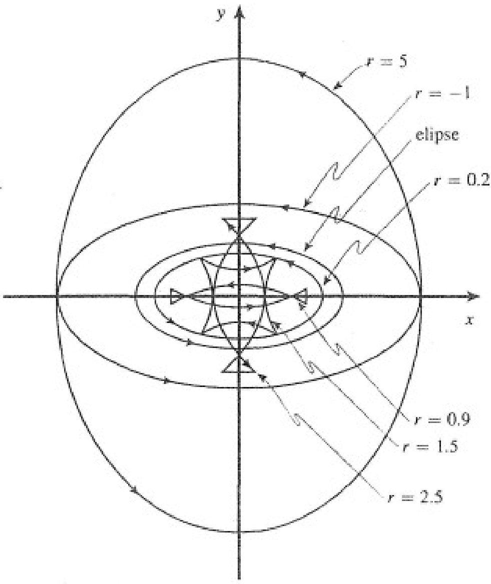

Figura 4. Curvas paralelas a la elipse $\alpha(t) = (a \cos t, b \operatorname{sen} t)$, $0 \leq t \leq 2\pi$.

propiedad fundamental para que tenga sentido hablar del área encerrada por la curva. Más aún, cuando $r \leq b^2/a$, o $r \geq a^2/b$, siendo $\boldsymbol{\beta}$ una curva positivamente orientada, el área que ésta encierra es

$$A(\boldsymbol{\beta}) = A(\boldsymbol{\alpha}) - rL(\boldsymbol{\alpha}) + \pi r^2 = \pi ab - rL(\boldsymbol{\alpha}) + \pi r^2$$

mientras que si $b \leq r \leq a$, la curva simple $\boldsymbol{\beta}$ tiene orientación negativa, y entonces el área encerrada por ella debe ser

$$A(\boldsymbol{\beta}) = -\pi ab + rL(\boldsymbol{\alpha}) - \pi r^2$$

En el teorema siguiente se recopilan los resultados obtenidos respecto a longitud y área encerradas por las curvas paralelas a una elipse.

Teorema 5.8.2 Sea $\boldsymbol{\alpha}: [0, 2\pi] \to \mathbb{R}^2$ la elipse

$$\boldsymbol{\alpha}(t) = (a \cos t, b \operatorname{sen} t)$$

en donde $a > b > 0$. La curva $\boldsymbol{\beta}: [0, 2\pi] \to \mathbb{R}^2$

$$\boldsymbol{\beta}(t) = \left(\cos t \left(a - \frac{rb}{(a^2 \operatorname{sen}^2 t + b^2 \cos^2 t)^{1/2}} \right), \operatorname{sen} t \left(b - \frac{ra}{(a^2 \operatorname{sen}^2 t + b^2 \cos^2 t)^{1/2}} \right) \right)$$

es una curva r-paralela a $\boldsymbol{\alpha}$. Sean $L(\boldsymbol{\alpha})$ y $L(\boldsymbol{\beta})$ la longitud de las curvas $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}$, y $A(\boldsymbol{\alpha})$ y $A(\boldsymbol{\beta})$ el área que éstas encierran, respectivamente.

a. Si $r \leq b^2/a$, la longitud de la curva $\boldsymbol{\beta}$ se calcula como

$$L(\boldsymbol{\beta}) = L(\boldsymbol{\alpha}) - 2\pi r$$

b. Si $r \geq a^2/b$, la longitud de la curva $\boldsymbol{\beta}$ se calcula como

$$L(\boldsymbol{\beta}) = -L(\boldsymbol{\alpha}) + 2\pi r$$

c. Si $r \in (-\infty, b^2/a] \cup [a^2/b, +\infty)$, el área encerrada por la

$$A(\boldsymbol{\beta}) = \pi ab - rL(\boldsymbol{\alpha}) + \pi r^2$$

d. Si $r \in [b, a]$, el área encerrada por la curva $\boldsymbol{\beta}$ es

$$A(\boldsymbol{\beta}) = -\pi ab + rL(\boldsymbol{\alpha}) - \pi r^2 \qquad \blacksquare$$

Para finalizar esta sección, estudiaremos las curvas paralelas a una parábola. Consideremos la parábola $\boldsymbol{\alpha}\colon I \to \mathbb{R}^2$, $\boldsymbol{\alpha}(t) = (t, pt^2)$, donde I es un intervalo de $\mathbb{R}$ simétrico respecto del origen, y p es un real positivo dado. Según la fórmula (1), la curva $\boldsymbol{\beta}$, r-paralela a $\boldsymbol{\alpha}$ es

$$\boldsymbol{\beta}(t) = \left(t - \frac{2rpt}{(1 + 4p^2t^2)^{1/2}}, t^2 + \frac{r}{(1 + 4p^2t^2)^{1/2}} \right)$$

Ciertamente ésta es una curva simétrica respecto del eje de ordenadas. Una pregunta natural sobre ella es: ¿esta curva —paralela a la parábola $y = px^2$—, también es una parábola? (Como se dijo al inicio, esta pregunta fue la que motivó en principio el presente estudio). No es difícil imaginar que la respuesta a la pregunta será negativa, pues, como hemos visto ya en varias ocasiones, la curva $\boldsymbol{\beta}$ puede llegar a perder su regularidad (propiedad que debe poseer cualquier curva parabólica). En efecto, la curvatura de $\boldsymbol{\alpha}$ en t viene dada por

$$k_{\alpha}(t) = \frac{2p}{(1 + 4p^2t^2)^{3/2}}$$

El rango de esta función es (pensando que $\boldsymbol{\alpha}$ está definida en todo $\mathbb{R}$) el intervalo $(0, 2p]$, de modo que, según el teorema 1, tomando $r^{-1} \in (0, 2p]$, la curva $\boldsymbol{\beta}$, r-paralela a $\boldsymbol{\alpha}$, no es regular, y no se puede esperar que ésta sea una parábola. Sin embargo, la inquietud puede persistir al imaginar una curva r-paralela a $\boldsymbol{\alpha}$ con $r < 0$ muy pequeño en valor absoluto (con $\boldsymbol{\beta}$ "muy cerca de la parte exterior de $\boldsymbol{\alpha}$"). En este caso, sin embargo, la respuesta sigue siendo negativa. En efecto, si la curva $\boldsymbol{\beta}$ fuera una parábola, ésta tendría que ser del tipo $y = \mu x^2 + \nu$, para ciertas constantes μ y ν. (En efecto, la curva $\boldsymbol{\beta}$ es simétrica respecto del eje de ordenadas y posee un extremo local en $\boldsymbol{\beta}(0) = (0, r)$, el cual corresponde al punto que está a r unidades del vértice de la parábola —que es, además, el único vértice de la curva, i.e. el punto en que $k'_{\alpha}(t) = 0$. Así, la única curva parabólica con estas características es $y = \mu x^2 + \nu$). Para probar nuestra afirmación ("la curva $\boldsymbol{\beta}$ no es una parábola") usaremos la siguiente caracterización de este tipo de curvas.

Lema La (traza de la) curva regular $\boldsymbol{\alpha}\colon \mathbb{R} \to \mathbb{R}^2$, $\boldsymbol{\alpha}(t) = (x(t), y(t))$, simétrica respecto del eje de ordenadas, es una parábola del tipo $y = \mu x^2 + \nu$, si y sólo si

$$y'(t)(x'(t))^2 - x(t)(x'(t)y''(t) - x''(t)y'(t)) = 0$$

Demostración. Primeramente veremos que la expresión anterior es invariante por reparametrizaciones de $\boldsymbol{\alpha}$. Sea $\tilde{\boldsymbol{\alpha}}: \mathbb{R} \to \mathbb{R}^2$ una reparametrización de $\boldsymbol{\alpha}$. Entonces $\tilde{\boldsymbol{\alpha}}(t) = \boldsymbol{\alpha}(\phi(t))$ para alguna función $\phi: \mathbb{R} \to \mathbb{R}$ diferenciable con derivada nunca nula en $\mathbb{R}$. Se tiene que

$$\tilde{\boldsymbol{\alpha}}'(t) = \phi'(t)\boldsymbol{\alpha}'(\phi(t)), \quad \tilde{\boldsymbol{\alpha}}''(t) = \phi''(t)\boldsymbol{\alpha}'(\phi(t)) + (\phi'(t))^2\boldsymbol{\alpha}''(\phi(t))$$

de modo que, escribiendo $\tilde{\boldsymbol{\alpha}}(t) = (\tilde{x}(t), \tilde{y}(t))$, se tiene

$$\begin{aligned}\tilde{y}'(t)\left(\tilde{x}'(t)\right)^2 - \tilde{x}(t)\left(\tilde{x}'(t)\tilde{y}''(t) - \tilde{x}''(t)\tilde{y}'(t)\right) &= \phi'(t)y'(\phi(t))[\phi'(t)x'(\phi(t))]^2 \\ &\quad - x(\phi(t))\left\{\phi'(t)x'(\phi(t))\left[\phi''(t)y'(\phi(t)) + (\phi'(t))^2 y''(\phi(t))\right]\right. \\ &\quad \left.- \phi'(t)y'(\phi(t))\left[\phi''(t)x'(\phi(t)) + (\phi'(t))^2 x''(\phi(t))\right]\right\} \\ &\quad (\phi'(t))^3\left[y'(t)(x'(t))^2 - x(t)(x'(t)y''(t) - x''(t)y'(t))\right] = 0\end{aligned}$$

Supongamos entonces que la (traza de la) curva α es de la forma mencionada en el lema. Es decir, supongamos que $y(t) = \mu(x(t))^2 + \nu$, $\forall t \in \mathbb{R}$. Tomemos $x(t) = t$, y entonces $y(t) = \mu t^2 + \nu$. Se tiene

$$y'(t)(x'(t))^2 - x(t)(x'(t)y''(t) - x''(t)y'(t)) = (2\mu t) - t(2\mu) = 0$$

En forma recíproca, suponga que se cumple la relación dada. Poniendo $x(t) = t$, nos queda que $y(t)$ satisface

$$y'(t) - t(y''(t)) = 0$$

de donde, integrando nos queda

$$y(t) = \mu t^2 + \nu$$

como queríamos. Q.E.D.

Ahora es fácil ver que nuestra curva $\boldsymbol{\beta}$, r-paralela a la parábola $\boldsymbol{\alpha}$, no es una parábola del tipo $y = \mu x^2 + \nu$. De hecho se tiene (escribiendo las funciones coordenadas de $\boldsymbol{\beta}$ como $x_{\beta}(t)$ y $y_{\beta}(t)$)

$$y'_{\beta}(t)(x'_{\beta}(t))^2 - x_{\beta}(t)(x'_{\beta}(t)y''_{\beta}(t) - x''_{\beta}(t)y'_{\beta}(t)) = 8rp^3t^3k_{\alpha}(t)(1 - rk_{\alpha}(t))^2 \neq 0$$

Si consideramos la parábola $y = px^2$ en todo su dominio, es decir, la (traza de la) curva $\boldsymbol{\alpha}: \mathbb{R} \to \mathbb{R}^2$, $\boldsymbol{\alpha}(t) = (t, pt^2)$, como dijimos anteriormente, la curva $\boldsymbol{\beta}$, r-paralela a $\boldsymbol{\alpha}$, será regular si y sólo si $r^{-1} \notin (0, 2p] =$ rango de $k_{\alpha}(t)$. Es decir, $\boldsymbol{\beta}$ es regular si y sólo si $r < 1/2p$. Esto significa que, globalmente, sólo se pueden tener curvas regulares paralelas a la parábola $y = px^2$ por afuera de ella ($r < 0$), o bien, por adentro hasta $r = 1/2p$ (no incluso, o sea $0 < r < 1/2p$). Para $r \geq 1/2p$, la curva $\boldsymbol{\beta}$ no es regular, teniéndose de hecho que si $r > 1/2p$, esta curva presentará una autointersección sobre el eje de las ordenadas. En efecto, haciendo $x_{\beta}(t) = 0$, es decir

$$t - \frac{2rpt}{(1 + 4p^2t^2)^{1/2}} = 0$$

vemos que una raíz de esta ecuación es siempre $t = 0$ (la curva $\boldsymbol{\beta}$ siempre pasa por el eje de ordenadas en $\boldsymbol{\beta}(0) = (0, r)$), teniéndose además la posibilidad de que

$$1 - \frac{2rpt}{(1 + 4p^2t^2)^{1/2}} = 0$$

(Nótese que esta ecuación no tiene solución si $r < 0$, es decir, las curvas r-paralelas hacia el exterior de la parábola solamente cruzan el eje de ordenadas en el punto $(0, r)$), de donde

$$\bar{t}_{1,2} = \pm(1/2p)(4r^2p^2 - 1)^{1/2}$$

(nótese también que si $r < 1/2p$ estas raíces no existen). Para $r > 1/2p$, éstas son dos raíces reales distintas en las que (puesto que $y_{\boldsymbol{\beta}}(t)$ es una función par) $y_{\boldsymbol{\beta}}(\bar{t}_1) = y_{\boldsymbol{\beta}}(\bar{t}_2)$. Estas raíces corresponden entonces a una autointersección de la curva $\boldsymbol{\beta}$ sobre el eje de ordenadas. Para $r = 1/2p$ (las raíces $\bar{t}_{1,2}$ son iguales —a cero), la curva $\boldsymbol{\beta}$, aunque no regular (de hecho $\boldsymbol{\beta}'(0) = 0$), es simple.

Veamos un poco más de cerca qué ocurre cuando $r \geq 1/2p$ (es decir, cuando la curva $\boldsymbol{\beta}$ no es regular). En este caso la ecuación

$$1 - rk_{\alpha}(t) = 1 - \frac{2pr}{(1 + 4p^2t^2)^{3/2}} = 0$$

posee dos raíces en

$$t^*_{1,2} = \pm(1/2p)\left((2pr)^{2/3} - 1\right)^{1/2}$$

Nuevamente observamos que si $r = 1/2p$, estas raíces son iguales, y la curva $\boldsymbol{\beta}$, aunque no regular (pierde su regularidad únicamente en $t^*_1 = t^*_2$), es una curva simple (en este caso todas las raíces $\bar{t}_1$, $\bar{t}_2$, t^*_1, t^*_2 son iguales a cero). Cuando $r > 1/2p$, las raíces t^*_1 y t^*_2 son distintas, y se tiene además que $\bar{t}_1 < t^*_1 < t^*_2 < \bar{t}_2$. Por la simetría de la curva $\boldsymbol{\beta}$ respecto del eje de ordenadas, podemos concluír que la parte de la curva $\boldsymbol{\beta}$ entre $\bar{t}_1$ y $\bar{t}_2$, corresponde a un "loop" (simétrico respecto del eje de ordenadas) dentro del cual la curva pierde su regularidad en los puntos $\boldsymbol{\beta}(t^*_1)$ y $\boldsymbol{\beta}(t^*_2)$.

La figura 5 (tomando $p = 1$) ilustra los casos en los que $r = -1$ (curva regular paralela hacia el exterior de la parábola $y = x^2$), $r = 0.25 < 0.5 = 1/2p$ (curva regular paralela hacia el interior de la parábola) y $r = 1.8 > 1/2p$ (curva no regular paralela hacia el interior de la parábola).

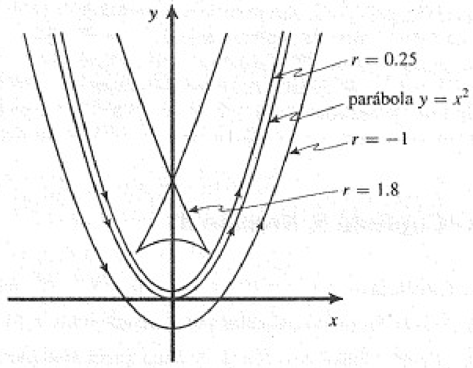

Figura 5. La parábola $y = x^2$ y sus curvas r-paralelas con $r = -1, 0.25, 1.8$.

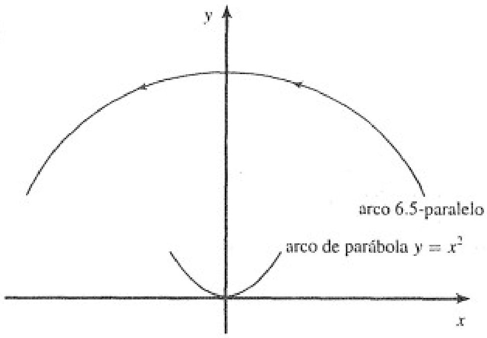

Figura 6. El arco de parábola $y = x^2$, $-1 \leq x \leq 1$ y su curva 6.5-paralela.

Si consideramos solamente un arco de parábola (simétrico respecto del eje de ordenadas), digamos $\boldsymbol{\alpha}: [-a, a] \to \mathbb{R}^2$, $\boldsymbol{\alpha}(t) = (t, pt^2)$, es posible hacer algunas observaciones adicionales a las anteriormente hechas en el caso general. En este caso la función curvatura $k_{\boldsymbol{\alpha}}: [-a, a] \to \mathbb{R}$

$$k_{\alpha}(t) = \frac{2p}{(1 + 4p^2t^2)^{3/2}}$$

alcanza su máximo absoluto en $t = 0$, que vale $k_{\alpha}(0) = 2p$, y alcanza también su mínimo absoluto en $t = \pm a$, que vale $k_{\alpha}(\pm a) = 2p/(1 + 4p^2a^2)^{3/2}$. LLamemos q a este valor. Se tiene entonces que la curva $\boldsymbol{\beta}$ será regular si y sólo si $r^{-1} \notin [q, 2p]$, es decir, si $r < 1/2p$ (como en el caso general) o si $r > 1/q = (1 + 4p^2a^2)^{3/2}/2p$. Este valor de q es positivo, lo indica la curva r-paralela $\boldsymbol{\beta}$ que está en la parte "interna" de la parábola. Es posible entonces tener una "imagen paralela al arco de parábola $y = px^2$, que sea regular". Observe sin embargo que el valor $1/q$ es en general muy grande (del orden de $4a^3$), lo que significa que "hay que despegarse mucho —hacia adentro, paralelamente— de la parábola para poder recuperar la regularidad de la curva r-paralela $\boldsymbol{\beta}$". Consideremos a manera de ejemplo el arco de parábola $\boldsymbol{\alpha}: [-1, 1] \to \mathbb{R}^2$, $\boldsymbol{\alpha}(t) = (t, t^2)$. En este caso $1/2p = 1/2$ y $q = 2/(5)^{3/2} = 0.1789$. Cuando $r < 1/2$, la curva $\boldsymbol{\beta}: [-1, 1] \to \mathbb{R}^2$, r-paralela a $\boldsymbol{\alpha}$, es regular, estando ésta en el exterior de la parábola cuando $r < 0$ y en su interior cuando $0 < r < 1/2$. Mientras r se encuentre entre $1/2$ y $1/q = 5.59016$, la curva $\boldsymbol{\beta}$ (en el interior de la parábola) no será regular y ésta vuelve a ser regular para $r > 5.59016$. La figura 6 ilustra esta situación.

Ejercicios (Capítulo 5, Sección 8)

1. Estudie las curvas paralelas a una recta $\boldsymbol{\lambda}: \mathbb{R} \to \mathbb{R}^2$, $\boldsymbol{\lambda}(t) = (t, mt + b)$.

2. Estudie las curvas paralelas a una circunferencia $\boldsymbol{\lambda}: [0, 2\pi] \to \mathbb{R}^2$, $\boldsymbol{\lambda}(t) = (a \cos t, a \operatorname{sen} t)$.

(***) **3.** ¿Cierto o falso? Sea $F(x, y) = 0$ una curva en el plano. Considere sus curvas ϵ-paralelas con ϵ pequeño. Estas pueden escribirse como $G(x, y) = 0$ si y sólo si la curva $F(x, y) = 0$ es una recta o una circunferencia.

⇒ 06 Jan 2021 – 1.22 TB (20210106)

CVGA2019-Videos

⇒ { 184.123.201.35.bc.
99.185.102.34.bc. ⇐
{ 173.193.186.35.bc.

Book1-002-2021 \ Cálculo \

Cálculo-vectorial-claudi-pita....pdf
Pág. 514 (523)
Figura 4
Curvas paralelas a la elipse $\alpha(t) =$
$(a \cos t, b \operatorname{sen} t),$
$0 \leq t \leq 2\pi$

'
)*+,)))- !) % " 7 < 4 5
2 ! " ! + /"
)*+,)))- !) % " 7 < 4 5 +
"=2

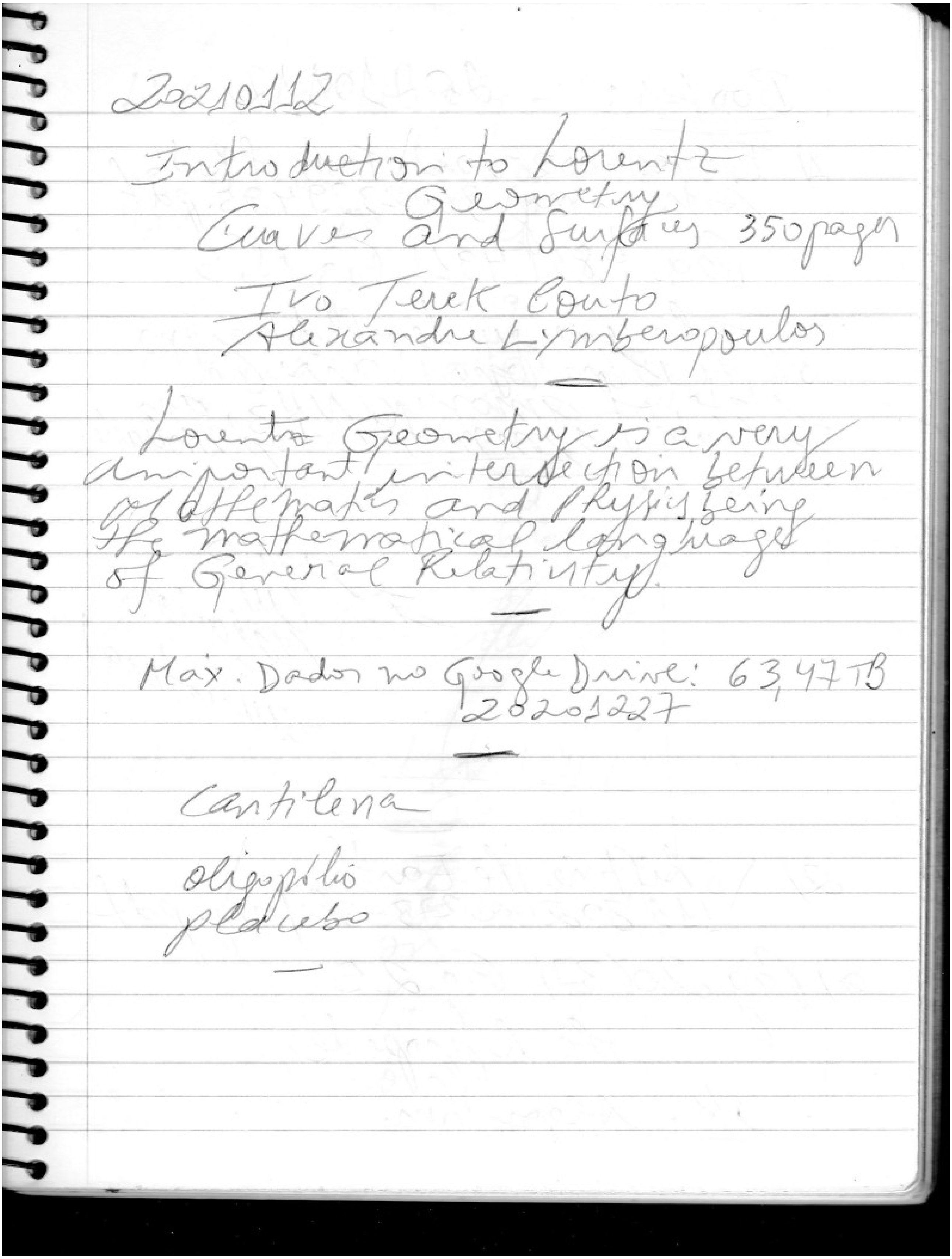

20210112

Introduction to Lorentz
Geometry
Curves and Surfaces 350 pages

Ivo Terek Couto
Alexandre Lymberopoulos

—

Lorentz Geometry is a very an important intersection between of Mathematics and Physics, being the mathematical language of General Relativity.

—

Max. Dados no Google Drive: 63,47 TB
20201227

—

cantilena

oligopólio
placebo

—

' .
!)*+,)))- !) $ 0 $&&0&(/"
' 0.$ 0& 1 ((

Table 15.1 Intensities of satellite groups relative to the Main Component (After Mauersberger 1983)

(J,K)	(1,1)	(2,2)	(3,3)	(4,4)	(5,5)	(6,6)	(2,1)
I_{inner}	0.295	0.0651	0.0300	0.0174	0.0117	0.0081	0.0651
I_{outer}	0.238	0.0628	0.0296	0.0173	0.0114	0.0081	0.0628

of $> 10^6$ years. This could occur much more quickly on grain surfaces, or with charged particles. If the exchange is slow, the ortho and para modifications of a particular species behave like different molecules; a comparison of ortho and para populations might give an estimate of temperatures in the distant past, perhaps at the time of molecular formation.

For the H_2 molecule, the symmetry of the rotational wavefunction depends on the total angular momentum J as $(-1)^J$. In the $J = 0$ state the rotational wavefunction is symmetric. However, the total wavefunction must be antisymmetric since protons are fermions. Thus, the $J = 0, 2, 4$, etc., rotational levels are para-H_2, while the $J = 1, 3, 5$, etc., are ortho-H_2. Spectral lines can connect only one modification. In the case of H_2, dipole rotational transitions are not allowed, but quadrupole rotational transitions ($\Delta J = \pm 2$) are. Thus, the 28 μm line of H_2 connects the $J = 2$ and $J = 0$ levels of para-H_2. Transitions between the ground and vibrational states are also possible.

Finally, as a more complex example of the relation of identical nuclei, we consider the case of three identical nuclei. This is the case for NH_3, CH_3CN and CH_3C_2H. Exchanging two of the nuclei is equivalent to a rotation by 120°. An exchange as was used for the case of two nuclei would not, in general, lead to a suitable symmetry. Instead combinations of spin states must be used (see Townes and Schawlow 1975). These lead to the result that the ortho to para ratio is two to one if the identical nuclei are protons. That is, NH_3, CH_3CN or CH_3C_2H the ortho form has $S(J, K) = 2$, while the para form has $S(J, K) = 1$. In summary, the division of molecules with identical nuclei into ortho and para species determines selection rules for radiative transitions and also rules the for collisions.

15.5.3 Hyperfine Structure

For symmetric top molecules, examples of hyperfine spectra are found for the inversion doublet transitions of NH_3. Since both the upper and lower levels have the same quantum numbers (J, K), there will be five groups of hyperfine components separated by a few MHz. Because of interactions between the spins of H nuclei there will be an additional splitting, within each group, of order a few kHz. In Table 15.1 we give the relative intensities of the NH_3 satellites for the case of low optical depth and LTE. For a molecule such as OH, one of the electrons is unpaired. The interaction of the nuclear magnetic moment with the magnetic moment of an unpaired electron is described as magnetic hyperfine structure. This splits a specific line into a number of components. In the case of the OH molecule, this interaction gives rise to a hyperfine splitting of the energy levels, in addition to the much larger Λ doubling. Together with the Λ doublet splitting, this gives rise to a quartet of energy levels in the OH ground state. Transitions between these energy levels produce the four ground state lines of OH at 18 cm wavelength (see Fig. 15.9).

NH_3 is an example of an oblate symmetric top molecule commonly found in the ISM. A diagram of the lower energy levels of NH_3 are shown in Fig. 15.5. A prolate top molecule has a cigar-like shape. Then the value A replaces C, and A>B. The energy-level diagrams for prolate symmetric top molecules found in the ISM, such as CH_3CCH and CH_3CN, follow this rule. However, since these molecules are much heavier than NH_3, the rotational transitions give rise to lines in the millimeter wavelength range, whereas the rotational lines of ammonia are in the sub-mm range (see Churchwell and Hollis 1983).

Differences in the orientation of the nuclei can be of importance. If a reflection of all particles about the center of mass leads to a configuration which cannot be obtained by a rotation of the molecule, these reflections represent two different states. For NH_3, we show this situation in the upper part of Fig. 15.5. Then there are two separate, degenerate states which exist for each value of (J, K) for $J \geq 1$. The $K = 0$ ladder is an exception because of the Pauli principle. These states are doubly degenerate as long as the potential barrier separating the two configurations is infinitely high. However, in molecules such as NH_3 the two configurations are separated only by a smaller potential barrier. This gives rise to a measurable splitting of the degenerate energy levels, which is referred to as inversion doubling. For NH_3, transitions between these inversion doublet levels are caused by the quantum mechanical tunneling of the nitrogen nucleus through the plane of the three protons. The wavefunctions of the two inversion doublet states have opposite parities, so that dipole transitions are possible. Thus dipole transitions occur between states with the same (J, K) quantum numbers. The splitting of the (J, K) levels for NH_3 shown in Fig. 15.5 is exaggerated; the inversion transitions give rise to spectral lines in the wavelength range near 1 cm. For CH_3CCH or CH_3CN, the splitting caused by inversion doubling is very small since the barrier is much higher than for NH_3.

The direction of the dipole moment of symmetric top molecules is parallel to the K axis. Spectral line radiation can be emitted only by a changing dipole moment. Since radiation will be emitted perpendicular to the direction of the dipole moment, there can be no radiation along the symmetry axis. Thus the K quantum number *cannot* change in dipole radiation, so allowed dipole transitions cannot connect different K ladders. The different K ladders are connected by octopole radiative transitions which require $\Delta K = \pm 3$. These are very slow, however, and collisions are far more likely to cause an exchange of population between different K ladders. We will discuss this in Sect. 15.6.3 in connection with the determination of T_K using the ratio of populations of different (J, K) states in symmetric top molecules.

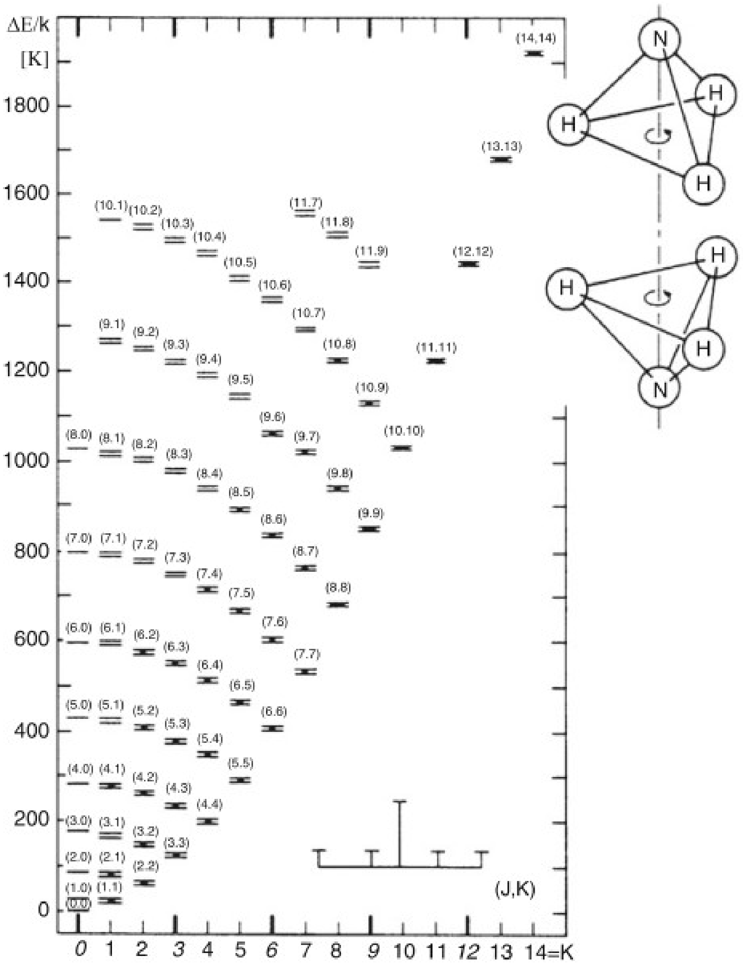
ΔE/k
[K]
1800
1600
1400
1200
1000
800
600
400
200
0
(14,14)
(13,13)
(10.1)
(10.2)
(10.3)
(10.4)
(10.5)
(10.6)
(11.7)
(11.8)
(11.9)
(12.12)
(10.7)
(10.8)
(11.11)
(10.9)
(10.10)
(9.1)
(9.2)
(9.3)
(9.4)
(9.5)
(9.6)
(9.7)
(9.8)
(9.9)
(8.0)
(8.1)
(8.2)
(8.3)
(8.4)
(8.5)
(8.6)
(8.7)
(8.8)
(7.0)
(7.1)
(7.2)
(7.3)
(7.4)
(7.5)
(7.6)
(7.7)
(6.0)
(6.1)
(6.2)
(6.3)
(6.4)
(6.5)
(6.6)
(5.0)
(5.1)
(5.2)
(5.3)
(5.4)
(5.5)
(4.0)
(4.1)
(4.2)
(4.3)
(4.4)
(3.0)
(3.1)
(3.2)
(3.3)
(2.0)
(2.1)
(2.2)
(1.0)
(1.1)
(0.0)
(J,K)
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14=K
N
H
H
H
H
H
H
N

Fig. 15.5 The energy-level diagram of the vibrational ground state of ammonia, NH_3, an oblate symmetric top molecule. Ortho-NH_3 has $K = 0, 3, 6, 9, \cdots$, while para-NH_3 has all other K values (see text). Rotational transitions with $\Delta J = 1$, $\Delta K = 0$, give rise to lines in the far IR. This molecule also has transitions with $\Delta J = 0$, $\Delta K = 0$ between inversion doublet levels. The interaction of the nuclear spin of ^{14}N with the electrons causes quadrupole hyperfine structure. In the $\Delta J = 0$, $\Delta K = 0$ transitions, the line is split into five groups of components. A sketch of the structure of the groups of hyperfine components of the $(J, K) = (1,1)$ inversion doublet line is indicated in the *lower right*; the separation is of order of MHz. In the *upper right* is a sketch of the molecule before and after an inversion transition, which gives rise to a 1.3 cm photon (Adapted from Wilson et al. 2006)

Books: 20210112

1) Books-002-2021\Tools of Radio Astronomy\3642399495.pdf

Pág. 483(491) Fig 15.5

The energy-level diagram of the vibrational ground state of amonia NH_3, an oblate symmetric top molecule.

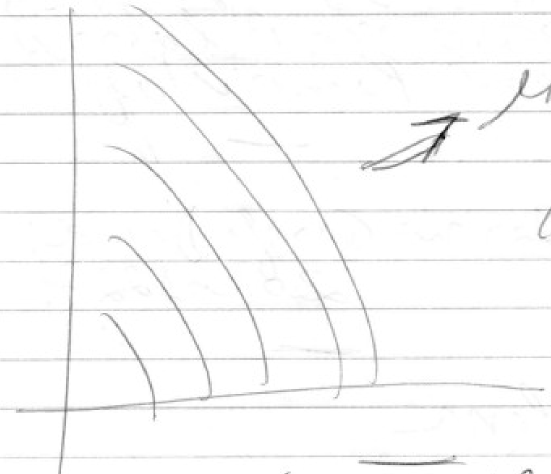

erpayo curve (calcular a curvatura)

2) \Shifting the Earth...\
11180242273-Shifting.pdf

a) Pág.20(37) Fig.2.5

The hippopede

Obs: Algorithm.

3)*+,)))- !) . 0 $ /"
' $ 1 (

By generating retrograde motion through this final construction, Eudoxus demonstrates his mastery of geometry. Below, we discuss some details of the construction. The presentation fully exploits mathematics that were not available to Eudoxus, including trigonometry, Cartesian coordinates, analytic geometry, linear algebra, and Euler angles. With these tools, not fully developed until two millennia after Eudoxus, the solution is manageable. Before continuing, take a moment to imagine the orbit of a point along the equator of a spinning sphere fixed on another spinning sphere that is then again fixed on another spinning sphere, which is finally tethered to an outermost spinning sphere. Imagine, without our modern mathematical tools, structuring the correct alignment of the axes and the speed of the spins to create a desired motion, such as retrograde motion. As you imagine this, Eudoxus earns your respect.[1]

The Whole

Eudoxus' heavens, the universal sphere and all other spheres perfectly centered around the earth, fit together as a whole through a system ordered according to distance from the earth. The center of the universe is Earth. On the closest sphere sits the moon. The next occupied sphere hosts Mercury, followed by Venus, the sun, Mars, Jupiter, Saturn, and finally the universal sphere. Aristotle, memorialized this universe in his writings. Later in time, the Church would deify Aristotle and in the process, canonize Eudoxus' universe.

THE MATH

Below we present the mathematics of Eudoxus' solar orbit and the orbit of a planet with retrograde motion. To begin, let us describe the construction of the orbits as an algorithm. Once this verbal description is available, we can associate mathematical operations with each step. The algorithm applies to each body separately.

Preliminary Step. Suspend the spheres concentrically so that they are all centered on the earth and so that their axes of rotation are all aligned along the north–south axis of the earth. Seat the planet or sun upon the equator of the innermost sphere.

Step 1. Among the stationary spheres, spin the innermost one at its appropriate speed and direction.

Step 2. Tilt the axis of the sphere from step 1 at an appropriate angle, and affix the axis to the immediate sphere that envelopes it.

Step 3. Return to step 1 until all the spheres are spinning.

---

[1] As an aid, a nice visual that displays the retrograde motion of a planet along three concentric circles is available on the website http://hsci.cas.ou.edu/images/applets/hippopede.html.

Two operations are of interest: (1) spinning a sphere with its axis aligned along the north–south direction and (2) tilting and affixing the sphere. These operations proceed from the innermost sphere outward toward the universal sphere.

To describe these operations mathematically, some notation is necessary. We consider three-dimensional Euclidean space in which (x,y,z) form the standard orthogonal coordinate system. Center the earth at the origin of the system and point the z axis directly north. The z axis is the axis of rotation for the universal sphere. Alongside Euclidean space, we append a time variable t, with the solar year as its unit measurement. A vector $(x(t), y(t), z(t))^T$ gives the (x,y,z) position of the sun or planet at time t.

Each sphere has a rate and direction of rotation designated by the real number ω_j. A positive value of ω_j indicates spin in a clockwise direction as viewed by an individual on the z axis looking down on the universal sphere from the outside. The subscript j identifies the associated sphere from the innermost to the universal sphere. With this notation, ω_1 is the angular velocity of the innermost sphere, ω_2 is the angular velocity of its immediate outer neighbor, ω_3 is the angular velocity of the third most inner sphere, and so forth. If there are m spheres, then ω_m is the angular velocity of the universal sphere.

Aside from the angular velocity, each sphere, excluding the universal sphere, has a tilt angle with respect to the axis of its immediate outer neighbor. We designate the tilt angle by θ_j, whereas with the angular velocity, the subscript j identifies the associated sphere.

With this notation, we are ready to mathematically describe the algorithm. The preliminary step sets the position of the body prior to performing any operations. The body maintains a constant distance from the earth that we arbitrarily set at one. This requires that the innermost sphere's radius be one. The preliminary step initially seats the body at the position $(0, 0, 1)^T$ on the innermost sphere; note that this position lies on the sphere's equator.

Step 1 imputes a rotation about the z axis through the angle $\omega_j t$, the angle is the product of the angular velocity and time. A vector executes the desired rotation when operated on by the following matrix.

$$R_{\omega_j}(t) = \begin{bmatrix} \cos(\omega_j t) & -\sin(\omega_j t) & 0 \\ \sin(\omega_j t) & \cos(\omega_j t) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

For example, applying the matrix to the vector $(1, 0, 3)^T$ yields the following result:

$$R_{\omega_j}(t)\begin{bmatrix}1\\0\\3\end{bmatrix} = \begin{bmatrix} \cos(\omega_j t) & -\sin(\omega_j t) & 0 \\ \sin(\omega_j t) & \cos(\omega_j t) & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix}1\\0\\3\end{bmatrix} = \begin{bmatrix}\cos(\omega_j t)\\ \sin(\omega_j t)\\ 3\end{bmatrix}$$

The (x,y) components of the vector $(1, 0, 3)^T$ are rotated to the values $(\cos(\omega_j t), \sin(\omega_j t))$, while the z component remains fixed at the value 3.

Step 2 tilts the sphere at a prescribed angle θ_j. A tilt is also a rotation that we perform in the (y, z) plane leaving the x axis fixed. The rotation matrix is given as follows.

$$T_{\theta_j} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta_j) & -\sin(\theta_j) \\ 0 & \sin(\theta_j) & \cos(\theta_j) \end{bmatrix}$$

For example, applying the matrix to the vector $(3, 0, 1)^T$ yields the following result:

$$T_{\theta_j}\begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta_j) & -\sin(\theta_j) \\ 0 & \sin(\theta_j) & \cos(\theta_j) \end{bmatrix}\begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ -\sin(\theta_j) \\ \cos(\theta_j) \end{bmatrix}$$

The x value of the vector remains fixed, while the (y,z) components are tilted through the angle θ_j.

We are now in a position to mathematically describe the motion of the sun as follows:

$$\begin{bmatrix} x(t) \\ y(t) \\ z(t) \end{bmatrix} = R_{\omega_2}(t)T_{\theta_1}R_{\omega_1}(t)\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \cos(\omega_2 t) & -\sin(\omega_2 t) & 0 \\ \sin(\omega_2 t) & \cos(\omega_2 t) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$
$$\times \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta_1) & -\sin(\theta_1) \\ 0 & \sin(\theta_1) & \cos(\theta_1) \end{bmatrix}\begin{bmatrix} \cos(\omega_1 t) & \sin(\omega_1 t) & 0 \\ \sin(\omega_1 t) & \cos(\omega_1 t) & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

We can march through the algorithm by reading the expression from right to left, reflecting the fact that the first operation on the initial vector is the far-right matrix, the second operation is the next matrix in the center, and the final operation is the far-left matrix.

Preliminary Step. The sun's preliminary position is set at (1,0,0).

Step 1. Since the solar ring completes one revolution annually, the value ω_1 is -2π. The matrix,

$$\begin{bmatrix} \cos(-2\pi t) & -\sin(-2\pi t) & 0 \\ \sin(-2\pi t) & \cos(-2\pi t) & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

spins the initial vector in a counterclockwise direction.

Step 2. The value θ_1 is equal to the angle of the ecliptic with the equatorial plane. The matrix,

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\left(2\pi\dfrac{23.5}{360}\right) & -\sin\left(2\pi\dfrac{23.5}{360}\right) \\ 0 & \sin\left(2\pi\dfrac{23.5}{360}\right) & \cos\left(2\pi\dfrac{23.5}{360}\right) \end{bmatrix},$$

tilts the solar ring through an angle of 23.5 degrees aligning the sun's orbit with the ecliptic.

Return to Step 1. Since the stars circle about the universal axis 366 times per year, the value ω_2 is 732π. The matrix,

$$\begin{bmatrix} \cos(732\pi t) & -\sin(732\pi t) & 0 \\ \sin(732\pi t) & \cos(732\pi t) & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

rotates the universal sphere clockwise. We are done.

Performing the matrix multiplications and simplification gives the following result for the sun's trajectory:

$$\begin{bmatrix} x(t) \\ y(t) \\ z(t) \end{bmatrix} = \begin{bmatrix} \cos(\omega_2 t)\cos(\omega_1 t) - \cos(\theta_1)\sin(\omega_2 t)\sin(\omega_1 t) \\ \sin(\omega_2 t)\cos(\omega_1 t) - \cos(\theta_1)\cos(\omega_2 t)\sin(\omega_1 t) \\ -\sin(\theta_1)\sin(\omega_2 t) \end{bmatrix}$$

There are a few notable features in this solution. As expressed by the z value, the sun oscillates up to a maximum value of $\sin(\theta_1)$ and a minimum value of $-\sin(\theta_1)$. The way everything has been set, the initial time represents the fall equinox, where the ecliptic and the equatorial plane intersect. The sun first attains its minimum and maximum values at $t = \frac{1}{4}$, the winter solstice, and $t = \frac{3}{4}$, the summer solstice, and annually after that. The maximum angle of latitude is θ_1, which, as previously noted, is the angle between the ecliptic and the equatorial plane.

Because the rates of rotation, ω_1 and ω_2, are both integral multiples of 2π, the values $(x(t),y(t),z(t))$ are cyclic with an annual period. Indeed, the sun returns to its initial position at (1,0,0) after one year and then repeats its orbit.

After this solar warmup exercise, we are ready to chase the planets and capture retrograde motion. The rotate, tilt, rotate, tilt ... algorithm applied to four spheres expresses itself through the following mathematical equation:

$$\begin{bmatrix} x(t) \\ y(t) \\ z(t) \end{bmatrix} = R_{\omega_4}(t)T_{\theta_3}R_{\omega_3}(t)T_{\theta_2}R_{\omega_2}(t)T_{\theta_1}R_{\omega_1}(t)\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

Here we are concerned primarily with the first and second spheres as these spheres are most responsible for the retrograde motion. Set the angular velocities of the two innermost spheres equal in magnitude but opposite in direction to one another: $\omega_2 = -\omega_1$. We examine this case for an arbitrary tilt angle θ_1.

Evaluating the expression

$$R_{-\omega_1}(t)T_{\theta_1}R_{\omega_1}(t)\begin{bmatrix}1\\0\\0\end{bmatrix}$$

gives the following result:

$$\begin{aligned}\begin{bmatrix}\tilde{x}(t)\\\tilde{y}(t)\\\tilde{z}(t)\end{bmatrix} &= R_{-\omega_1}(t)T_{\theta_1}R_{\omega_1}(t)\begin{bmatrix}1\\0\\0\end{bmatrix}\\ &= \begin{bmatrix}\cos^2(\omega_1 t)+\cos(\theta_1)\sin^2(\omega_1 t)\\(\cos(\theta_1)-1)\sin(\omega_1 t)\cos(\omega_1 t)\\\sin(\theta_1)\sin(\omega_1 t)\end{bmatrix}\end{aligned}\tag{2.1}$$

Our first problem is to describe the path of the planet absent the influence of the third sphere and universal sphere. (The tildas emphasize that this is not a complete description of the path since not all spheres are considered.) First let us describe the relation between $\tilde{x}(t)$ and $\tilde{y}(t)$. Using trigonometric identities, it is possible to solve for both $\sin(\omega_1 t)$ and $\cos(\omega_1 t)$ in terms of $\tilde{x}$. On substituting the solutions into the equation for $\tilde{y}(t)$, the relation between $\tilde{x}$ and $\tilde{y}$ emerges. The steps are as follows:

$$\begin{aligned}\tilde{x} &= \cos^2(\omega_1 t)+\cos(\theta_1)\sin^2(\omega_1 t)\\ &= 1-\sin^2(\omega_1 t)+\cos(\theta_1)\sin^2(\omega_1 t)\\ &= 1+(\cos(\theta_1)-1)\sin^2(\omega_1 t)\end{aligned}$$

Solving for $\sin^2(\omega_1 t)$ yields the following equality:

$$\sin^2(\omega_1 t) = \frac{\tilde{x}-1}{\cos(\theta_1)-1}\tag{2.2}$$

A similar calculation in which one uses the Pythagorean identity to substitute for $\sin^2(\omega_1 t)$ yields the following result:

$$\cos^2(\omega_1 t) = \frac{\tilde{x}-\cos(\theta_1)}{1-\cos(\theta_1)}\tag{2.3}$$

Substituting expressions (2.2) and (2.3) into the equation for $\tilde{y}(t)$ [i.e., (2.1)] and simplifying gives the following relation between $\tilde{x}$ and $\tilde{y}$:

$$\begin{aligned}
\tilde{y} &= (\cos(\theta_1) - 1)\sin(\omega_1 t)\cos(\omega_1 t) \\
\tilde{y}^2 &= (\cos(\theta_1) - 1)^2 \sin^2(\omega_1 t)\cos^2(\omega_1 t) \\
&= (\cos(\theta_1) - 1)^2 \left(\frac{\tilde{x} - 1}{\cos(\theta_1) - 1}\right)\left(\frac{\tilde{x} - \cos(\theta_1)}{1 - \cos(\theta_1)}\right) \\
&= (1 - \tilde{x})(\tilde{x} - \cos(\theta_1)) \\
&= -\tilde{x}^2 + (1 + \cos(\theta_1))\,\tilde{x} - \cos(\theta_1) \\
&= -\left(\tilde{x} - \frac{1 + \cos(\theta_1)}{2}\right)^2 + \left(\frac{1 + \cos(\theta_1)}{2}\right)^2 - \cos(\theta_1) \\
&= -\left(\tilde{x} - \frac{1 + \cos(\theta_1)}{2}\right)^2 + \left(\frac{1 - \cos(\theta_1)}{2}\right)^2
\end{aligned}$$

In two dimensions the path of $\tilde{x}(t)$ and $\tilde{y}(t)$ is the circle centered at $(\tilde{x}, \tilde{y}) = \left(\frac{1+\cos(\theta_1)}{2}, 0\right)$ with radius $\frac{1-\cos(\theta_1)}{2}$:

$$\left(\tilde{x} - \frac{1 + \cos(\theta_1)}{2}\right)^2 + \tilde{y}^2 = \left(\frac{1 - \cos(\theta_1)}{2}\right)^2 \tag{2.4}$$

Returning to three-dimensional space, the result obtained above necessitates that the planet's orbit confine itself to the surface of the cylinder whose cross section at any constant value of z is the circle of equation (2.4). Also note that in three dimensions, because the inner sphere's center always remains fixed and the inner sphere's radius is one, the planet meanders about the sphere $\tilde{x}^2 + \tilde{y}^2 + \tilde{z}^2 = 1$. The shape of the planet's path, absent the influence of the third sphere and the universal sphere, is then the intersection of these two surfaces (see Figure 2.5).

The resulting intersection appears like a figure eight that has been stamped on a sphere. A further analysis of the path given by equation (2.1) reveals that the planet begins at the cross of the figure eight and skates about the upper loop (assuming that θ_1 ranges between 0 and 180 degrees). After returning to the cross, the planet skates about the lower loop. Eudoxus calls the figure a hippopede, meaning horse fetter.

Before continuing, note the influence of the tilt angle θ_1 as it ranges from 0 to 180 degrees. When the tilt angle is 0 degree, the two spheres are aligned and spinning in opposite directions. The planet never moves but remains at its initial point. One can see that the circle of equation (2.4) shrinks to a point at which the corresponding cylinder shrinks to its central axis and intersects the sphere tangentially at the point (1,0,0). As we increase the tilt angle toward 90 degrees, the radius of the cylinder increases, while its surface remains tangent to the sphere at the point (1,0,0). This increases the size of the figure eight, and

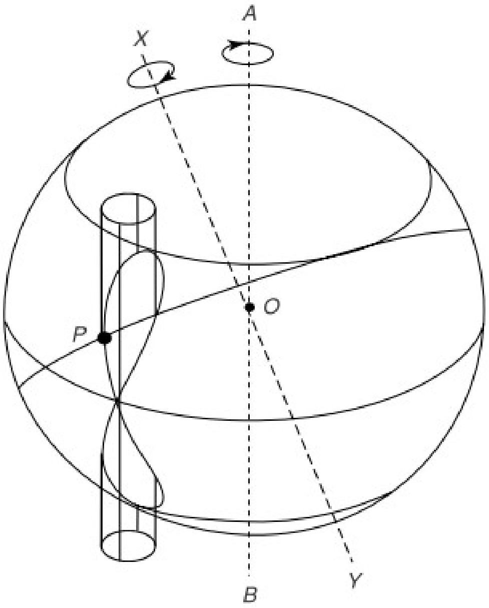

Figure 2.5. The hippopede.

the figure reaches its greatest vertical extent when the angle is 90 degrees and the loops touch the poles. Beyond 90 degrees the loops begin to approach one another, sliding toward the equator around the poles until at 180 degrees they meet up with one another and the planet spins around the equatorial plane at the rate $2\omega_1$.

Proceeding to step 2 of the algorithm, we tilt the figure eight and affix it to the third sphere. Let the tilt angle θ_2 be 90 degrees so that the figure eight lies on its side, becoming an infinity sign. Applying the tilt yields the following result:

$$\begin{aligned}
\begin{bmatrix} \tilde{x}(t) \\ \tilde{y}(t) \\ \tilde{z}(t) \end{bmatrix} &= T_{\theta_2} R_{-\omega_1}(t) T_{\theta_1} R_{\omega_1}(t) \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \\
&= T_{\theta_2} \begin{bmatrix} \cos^2(\omega_1 t) + \cos(\theta_1)\sin^2(\omega_1 t) \\ (\cos(\theta_1) - 1)\sin(\omega_1 t)\cos(\omega_1 t) \\ \sin(\theta_1)\sin(\omega_1 t) \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \cos^2(\omega_1 t) + \cos(\theta_1)\sin^2(\omega_1 t) \\ (\cos(\theta_1) - 1)\sin(\omega_1 t)\cos(\omega_1 t) \\ \sin(\theta_1)\sin(\omega_1 t) \end{bmatrix} \\
&= \begin{bmatrix} \cos^2(\omega_1 t) + \cos(\theta_1)\sin^2(\omega_1 t) \\ -\sin(\theta_1)\sin(\omega_1 t) \\ (\cos(\theta_1) - 1)\sin(\omega_1 t)\cos(\omega_1 t) \end{bmatrix}
\end{aligned}$$

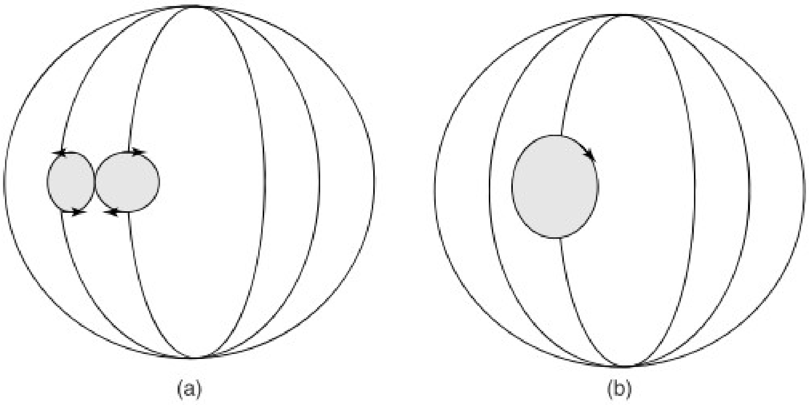

Figure 2.6. Comparison of motion along hippopede (a) and circle (b).

After this tilt, the planet skates around the resting figure eight traversing back and forth somewhat along the universal equator (see Figure 2.6a). The next step is to impute a rotation that governs the drift of the planet through the zodiac. It is the combination of the uniform drift along with the skate around the figure eight that creates retrograde motion. As an aid toward visualizing what occurs, let's consider a simpler construction with a similar feature.

For the simplified construct, assume that the planet lies at the intersection of a cylinder whose central axis is the x axis and a sphere. The intersection is a circle as displayed in Figure 2.6b, which contrasts this construct with that of Eudoxus. As displayed in the figure, the planet rotates uniformly about the intersecting circle, causing it to change direction, just as Eudoxus' skating planet. Next let's rotate this planet about the heavens. To make a mechanical device that replicates the motion, build a track around a great circle of the universal sphere and place a wheel connected to Earth with an axle on the track. Somewhat closer to Earth, on the axle, attach a pointer that rotates with the wheel. As the wheel rotates uniformly about its track, the tip of the pointer gives the location of the planet. As Figure 2.7 illustrates, while following a trend through the constellations, the orbit loops about itself in retrograde motion. Eudoxus' skating planet, skating back and forth along its figure eight as it trends through the zodiac, produces retrograde motion in a similar manner.

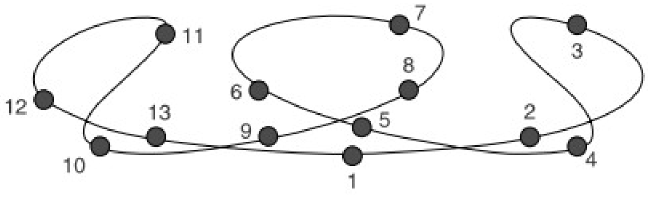

Figure 2.7. Simplified model yielding three retrograde loops every trip around the zodiac.

Returning to Eudoxus' construction, we next spin the third sphere. The third sphere governs the rate at which the planet floats through the constellations, requiring that the angular velocity ω_3 give the radian measurement of the planet's cycle through the universal sphere in one year's time. (Recall that the sun cycles through the zodiac one time per year, so the angular velocity of its sphere is 2π.) Mars skates through the zodiac around 1.06 times every 2 years or gets a little over halfway through in one year. The associated angular velocity ω_3 is about -1.06π. Applying the spin of the third sphere to the planet results in the following trajectory:

$$\begin{aligned}
\begin{bmatrix} \tilde{x}(t) \\ \tilde{y}(t) \\ \tilde{z}(t) \end{bmatrix} &= R_{\omega_3} T_{\theta_2} R_{-\omega_1}(t) T_{\theta_1} R_{\omega_1}(t) \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \\
&= R_{\omega_3} \begin{bmatrix} \cos^2(\omega_1 t) + \cos(\theta_1)\sin^2(\omega_1 t) \\ -\sin(\theta_1)\sin(\omega_1 t) \\ (\cos(\theta_1) - 1)\sin(\omega_1 t)\cos(\omega_1 t) \end{bmatrix} \\
&= \begin{bmatrix} \cos(\omega_3 t) & \sin(\omega_3 t) & 0 \\ -\sin(\omega_3 t) & \cos(\omega_3 t) & 0 \\ 0 & 0 & 1 \end{bmatrix} \\
&\quad \times \begin{bmatrix} \cos^2(\omega_1 t) + \cos(\theta_1)\sin^2(\omega_1 t) \\ -\sin(\theta_1)\sin(\omega_1 t) \\ (\cos(\theta_1) - 1)\sin(\omega_1 t)\cos(\omega_1 t) \end{bmatrix} \\
&= \begin{bmatrix} \left(\cos^2(\omega_1 t) + \cos(\theta_1)\sin^2(\omega_1 t)\right)\cos(\omega_3 t) \\ -\sin(\theta_1)\sin(\omega_1 t)\sin(\omega_3 t) \\ -\left(\cos^2(\omega_1 t) + \cos(\theta_1)\sin^2(\omega_1 t)\right)\sin(\omega_3 t) \\ +\sin(\theta_1)\sin(\omega_1 t)\cos(\omega_3 t) \\ (\cos(\theta_1) - 1)\sin(\omega_1 t)\cos(\omega_1 t) \end{bmatrix}
\end{aligned}$$

As the $\tilde{z}(t)$ component of the trajectory indicates, the planet oscillates about the equatorial plane $\tilde{z} = 0$. In reality, the planets hug the ecliptic. A tilt of the third sphere toward the ecliptic ($\theta_3 = 23.5$ degrees) reorients the planet:

$$\begin{aligned}
\begin{bmatrix} \tilde{x}(t) \\ \tilde{y}(t) \\ \tilde{z}(t) \end{bmatrix} &= T_{\theta_3} R_{\omega_3} T_{\theta_2} R_{-\omega_1}(t) T_{\theta_1} R_{\omega_1}(t) \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \\
&= T_{\theta_3} \begin{bmatrix} \left(\cos^2(\omega_1 t) + \cos(\theta_1)\sin^2(\omega_1 t)\right)\cos(\omega_3 t) \\ -\sin(\theta_1)\sin(\omega_1 t)\sin(\omega_3 t) \\ -\left(\cos^2(\omega_1 t) + \cos(\theta_1)\sin^2(\omega_1 t)\right)\sin(\omega_3 t) \\ +\sin(\theta_1)\sin(\omega_1 t)\cos(\omega_3 t) \\ (\cos(\theta_1) - 1)\sin(\omega_1 t)\cos(\omega_1 t) \end{bmatrix}
\end{aligned}$$

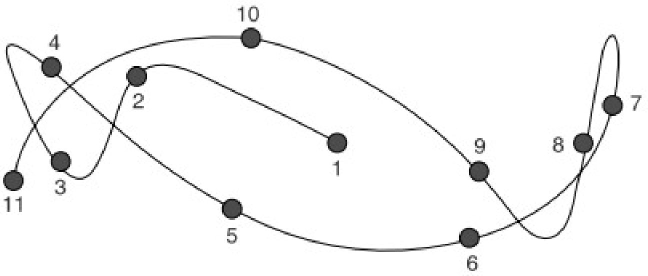

Figure 2.8. Eudoxus' planet with retrograde orbit.

$$T_{\theta_3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta_3) & -\sin(\theta_3) \\ 0 & \sin(\theta_3) & \cos(\theta_3) \end{bmatrix}$$

The results for $(\tilde{x}(t), \tilde{y}(t), \tilde{z}(t))$ are as follows:

$$\begin{aligned} \tilde{x}(t) &= \left(\cos^2(\omega_1 t) + \cos(\theta_1)\sin^2(\omega_1 t)\right)\cos(\omega_3 t) \\ &\quad - \sin(\theta_1)\sin(\omega_1 t)\sin(\omega_3 t) \\ \tilde{y}(t) &= \left[-\left(\cos^2(\omega_1 t) + \cos(\theta_1)\sin^2(\omega_1 t)\right)\sin(\omega_3 t)\right. \\ &\quad \left. + \sin(\theta_1)\sin(\omega_1 t)\cos(\omega_3 t)\right]\cos(\theta_3) \\ &\quad - (\cos(\theta_1) - 1)\sin(\omega_1 t)\cos(\omega_1 t)\sin(\theta_3) \\ \tilde{z}(t) &= \left[-\left(\cos^2(\omega_1 t) + \cos(\theta_1)\sin^2(\omega_1 t)\right)\sin(\omega_3 t)\right. \\ &\quad \left. + \sin(\theta_1)\sin(\omega_1 t)\cos(\omega_3 t)\right]\sin(\theta_3) \\ &\quad + (\cos(\theta_1) - 1)\sin(\omega_1 t)\cos(\omega_1 t)\cos(\theta_3) \end{aligned}$$

These equations encode retrograde motion through the zodiac. Figure 2.8 is a graph of the orbit. The animation at http://hsci.cas.ou.edu/images/applets/hippopede.html uses these equations. There is only one more operation left to complete the description: to spin the universal sphere to endow the planet with its daily motion. We leave Eudoxus' model incomplete and allow the reader seeking perfection to remove the tildas and complete it.

APOLLONIUS’ OSCULATING CIRCLE AND THE ELLIPSE

We restrict our investigation to the case of the ellipse. Begin with the equation for an ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \tag{4.1}$$

where a is the major axis and b is the minor axis of the ellipse.

Let (x_0, y_0) be an arbitrary point in the plane of the ellipse. We seek an equation for a normal line from the ellipse to the point. The equation for a general line passing through the point (x_0, y_0) is $y - y_0 = m(x - x_0)$, where m is the slope of the line. It is necessary to determine the slope for a normal line. As Apollonius neatly demonstrates, it is possible to solve this without resorting to calculus.[4] Here we fully employ the power of calculus. With the knowledge that the gradient of a function points in the normal direction of the function’s constant level surfaces, we find the gradient of the function $f(x, y) = \frac{x^2}{a^2} + \frac{y^2}{b^2}$, from which it is possible to determine the slope of the normal line:

$$\left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right) = \left(\frac{2x}{a^2}, \frac{2y}{b^2}\right)$$

---

[4] Select an arbitrary point on the ellipse, (x_e, y_e) and write a general equation for all lines passing through that point:

$$\begin{aligned} y - y_e &= m_t(x - x_e) \\ y &= m_t(x - x_e) + y_e \end{aligned}$$

Next, substitute for the y variable in (4.1) using the right-hand side (RHS) of this equation. The result is a quadratic equation.

$$\begin{aligned} \frac{x^2}{a^2} + \frac{(m_t(x - x_e) + y_e)^2}{b^2} &= 1 \\ \left(\frac{1}{a^2} + \frac{m_t^2}{b^2}\right)x^2 + \left(\frac{2m_t(y_e - m_t x_e)}{b^2}\right)x + \frac{(y_e - m_t x_e)^2}{b^2} - 1 &= 0 \end{aligned}$$

As the tangent line intersects this equation in one point, we look for a single solution to the quadratic equation:

$$x = \frac{-\left(\frac{2m_t(y_e - m_t x_e)}{b^2}\right) \pm \sqrt{\left(\frac{2m_t(y_e - m_t x_e)}{b^2}\right)^2 - 4\left(\frac{1}{a^2} + \frac{m_t^2}{b^2}\right)\left(\frac{(y_e - m x_e)^2}{b^2} - 1\right)}}{2\left(\frac{1}{a^2} + \frac{m_t^2}{b^2}\right)}$$

The solution reduces to a single point when the expression under the radical vanishes. The reader can check that this occurs when $m_t = -\frac{b^2 x_e}{a^2 y_e}$. The normal line is at right angles to the tangent line and accordingly has slope $m = \frac{a^2 y_e}{b^2 x_e}$.

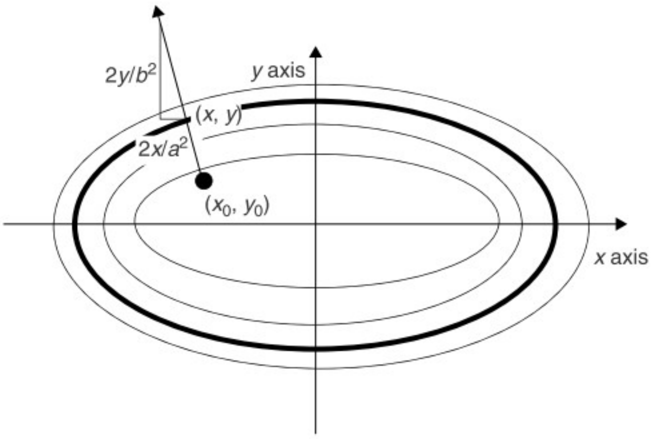

Figure 4.2. Gradient vector gives slope of normal line.

Then the slope of the normal line is given by the following equation (see Figure 4.2):

$$m = \frac{a^2 y}{b^2 x} \tag{4.2}$$

Placing the value of m into the equation of the normal line passing through the point (x_0, y_0) and simplifying results in an equation that describes a double branched hyperbola:

$$\begin{aligned} y - y_0 &= m(x - x_0) \\ y - y_0 &= \frac{a^2 y}{b^2 x}(x - x_0) \\ b^2 x(y - y_0) &= a^2 y(x - x_0) \\ (a^2 - b^2)xy + b^2 y_0 x - a^2 x_0 y &= 0 \end{aligned} \tag{4.3}$$

We can conclude that when the normal line to the point (x_0, y_0) intersects the ellipse of equation (4.1) at coordinates (x, y), it also intersects the double-branched hyperbola of equation (4.3) at the same coordinates (x, y). Several more conclusions are possible:

1. The point (x_0, y_0) determines the configuration of the hyperbola and hence the number of intersections with the ellipse.
2. We conclude that the number of intersections between the ellipse and the hyperbola gives the number of normal lines, which is the answer to Apollonius' problem. Figure 4.3 illustrates the possible ways in which the figures can intersect. We can see that at most there can be four normal lines to

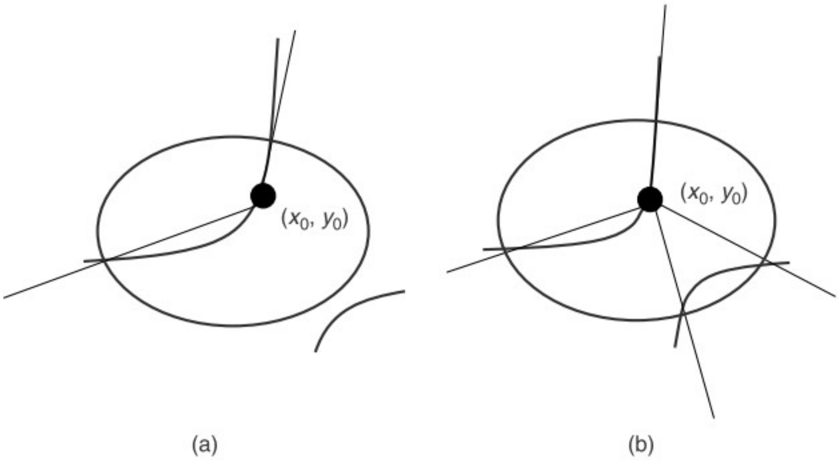

Figure 4.3. Intersection of hyperbola with ellipse: (a) two intersections, two normal lines; (b) four intersections, four normal lines.

a point as given by four points of intersection (Figure 4.3b). Noting that the minimum and maximum distances from the ellipse to a point are given by the length of the normal section, we can also conclude that there are at least two normals to a point, one associated with the minimum distance and one associated with the maximum distance. In fact, since one branch of the hyperbola passes through the origin and extends to infinity in two directions, there must be at least two points of intersection (Figure 4.3a).

3. Finally, we note that three normals to a point occur when one branch of the hyperbola is tangent to the ellipse, yielding a single point of intersection (Figure 4.4). This final conclusion is critical in our further analysis.

An algebraic analysis also yields some of the same conclusions. Let us find an equation for the intersections between the ellipse of equation (4.1) and the hyperbola of equation (4.3). Rearranging equation (4.3) and squaring results in the following equation:

$$(a^2 - b^2)xy + b^2xy_0 - a^2yx_0 = 0$$

$$(a^2 - b^2)xy - a^2yx_0 = -b^2xy_0$$

$$\left((a^2 - b^2)x - a^2x_0\right) y = -b^2xy_0$$

$$\left((a^2 - b^2)x - a^2x_0\right)^2 y^2 = b^4x^2y_0^2$$

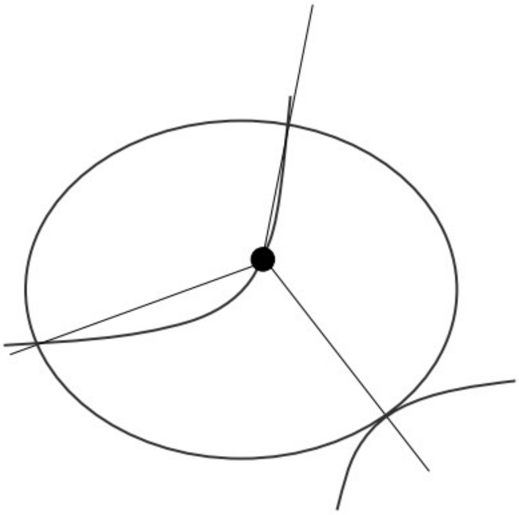

Figure 4.4. Intersection of hyperbola with ellipse: three intersections, three normal lines.

Next, solving for the value y^2 in equation (4.1) and substituting the result into this equation yields a quartic polynomial in x:

$$\begin{aligned}\frac{x^2}{a^2}+\frac{y^2}{b^2}&=1\\ y^2&=b^2\left(1-\frac{x^2}{a^2}\right)\\ \left((a^2-b^2)x-a^2x_0\right)^2y^2&=b^4x^2y_0^2\\ \left((a^2-b^2)x-a^2x_0\right)^2\left(1-\frac{x^2}{a^2}\right)&=b^2x^2y_0^2\\ \left((a^2-b^2)x-a^2x_0\right)^2(a^2-x^2)-a^2b^2x^2y_0^2&=0 \qquad (4.4)\end{aligned}$$

Before continuing, we make the following observations. At most, there are four solutions to the quartic polynomial, indicating that any point has at most four normals. Evaluating the quartic polynomial at the points $x=\pm a$ and $x=0$ yields the following:

$$\begin{aligned}\left((a^2-b^2)a^2-a^2x_0\right)^2(a^2-a^2)-b^2a^4y_0^2&=-b^2a^4y_0^2\\ \left((a^2-b^2)0-a^2x_0\right)^2(a^2-0^2)-a^2b^20^2y_0^2&=a^6x_0^2\end{aligned}$$

Since the sign of the quartic changes between $x=-a$ and $x=0$, there is at least one root between these values. By the same argument, there is at least one root between $x = 0$ and $x = a$. So there are at least two roots, equating to two normal lines to any point.[5] The algebraic analysis is consistent with the geometric analysis.

The first individual to solve the quartic polynomial was Lodovico Ferarri in the sixteenth century.[6] This solution was not available to Apollonius. Nevertheless, Apollonius went on to discover further relations that allow one to specify the number of normal lines to any given point. Below, we find the solution without explicitly solving the quartic polynomial.

It is of interest to find the set of points where there are three normals. It turns out that this set defines a closed curve; within the curve there are four normals, on the curve there are three normals, and outside the curve there are two normals. To determine the curve, let x_0 and y_0 be variables and seek a curve of the form $g(x_0, y_0) = 0$.

Note that equation (4.3) is linear in both x_0 and y_0. Using the observation that a branch of the hyperbola associated with a point having three normals is tangent to the ellipse (see Figure 4.4), one can obtain another relation between x_0 and y_0. At the tangent point the normal to the ellipse is aligned with that of the hyperbola. Using equation (4.3), we set $h(x, y) = (a^2 - b^2)xy + b^2xy_0 - a^2yx_0$, giving a family of hyperbolas. The gradient of $h(x, y)$ is aligned with the normal vector:

$$\left(\frac{\partial h}{\partial x}, \frac{\partial h}{\partial y}\right) = \left((a^2 - b^2)y + b^2y_0, (a^2 - b^2)x - a^2x_0\right)$$

The slope of the normal line is $m = \frac{(a^2-b^2)x-a^2x_0}{(a^2-b^2)y+b^2y_0}$, which we can equate with the slope from the previous calculation, equation (4.2):

$$\begin{aligned} m &= \frac{(a^2 - b^2)x - a^2x_0}{(a^2 - b^2)y + b^2y_0} \\ &= \frac{a^2y}{b^2x} \end{aligned} \tag{4.5}$$

Simplifying the expression yields another linear equation in x_0 and y_0:

$$\begin{aligned} \frac{(a^2 - b^2)x - a^2x_0}{(a^2 - b^2)y + b^2y_0} &= \frac{a^2y}{b^2x} \\ \left((a^2 - b^2)x - a^2x_0\right)b^2x &= \left((a^2 - b^2)y + b^2y_0\right)a^2y \end{aligned}$$

---

[5] Because of the symmetry of the ellipse of equation (4.1), no spurious roots were introduced by squaring equation (4.2).

[6] An interesting story surrounds the discovery of the solution to the quartic polynomial. Ferarri and a compatriot, Tartaglia, went at one another over priority rights for discovery of the cubic. Their public argument culminated in a public mathematical duel. For details, see *The Ellipse* by the author [(Mazer 2010); see Bibliography at end of this book].

$$-a^2b^2(xx_0 + yy_0) = \frac{a^2 - b^2}{a^2b^2}(b^2x^2 - a^2y^2)$$

$$xx_0 + yy_0 = \frac{a^2 - b^2}{a^2b^2}(b^2x^2 - a^2y^2) \tag{4.6}$$

Equations (4.3) and (4.6) form a pair of linear equations in the variables x_0 and y_0. We next solve this system of equations for the variable x_0:

$$a^2yx_0 - b^2xy_0 = -(a^2 - b^2)xy$$

$$xx_0 + yy_0 = \frac{a^2 - b^2}{a^2b^2}(b^2x^2 - a^2y^2)$$

Multiplying the top equation by y and the bottom equation by b^2x and adding the two outcomes eliminates the y_0 variable and results in an equation in x_0:

$$(a^2y^2 + b^2x^2)x_0 = -(a^2 - b^2)xy^2 + \frac{a^2 - b^2}{a^2}(b^2x^3 - a^2xy^2)$$

$$(a^2y^2 + b^2x^2)x_0 = \frac{a^2 - b^2}{a^2}b^2x^3$$

Next, using the equation for the ellipse [eq. (4.1)], note that $a^2y^2 + b^2x^2 = a^2b^2$. Substituting this identity into the expression above results in a relationship between x and x_0:

$$(a^2y^2 + b^2x^2)x_0 = \frac{a^2 - b^2}{a^2}b^2x^3$$

$$a^2b^2x_0 = \frac{a^2 - b^2}{a^2}b^2x^3 \tag{4.7}$$

$$x_0 = \frac{a^2 - b^2}{a^4}x^3$$

Equation (4.7) has a nice geometric interpretation as follows. Assume that we start off with a point on the ellipse, (x, y) and ask ourselves at what value x_0 does the normal line to (x, y) intersect exactly two other normal lines. Equation (4.7) answers this question. With this relationship it is possible to find the curve of interest: $g(x_0, y_0)$.

One can make a similar calculation, or assert by symmetry the following equation between y and y_0:

$$y_0 = \frac{b^2 - a^2}{b^4}y^3 \tag{4.8}$$

From equations (4.7) and (4.8) one obtains the following:

$$\left(\frac{ax_0}{a^2-b^2}\right)^{2/3} = \frac{x^2}{a^2}$$

$$\left(\frac{by_0}{a^2-b^2}\right)^{2/3} = \frac{y^2}{b^2}$$

Adding these equations, using the equation for the ellipse, and simplifying yields the desired curve:

$$\left(\frac{ax_0}{a^2-b^2}\right)^{2/3} + \left(\frac{by_0}{a^2-b^2}\right)^{2/3} = \frac{x^2}{a^2} + \frac{y^2}{b^2}$$

$$\left(\frac{ax_0}{a^2-b^2}\right)^{2/3} + \left(\frac{by_0}{a^2-b^2}\right)^{2/3} = 1$$

$$(ax_0)^{2/3} + (by_0)^{2/3} = \left(a^2-b^2\right)^{2/3} \tag{4.9}$$

With one caveat that we address below, the points on the curve given by equation (4.9) are intersections of three normal lines from the ellipse. Figure 4.5 illustrates this curve. Apollonius demonstrates that within this curve, each point stands on four normal lines from the ellipse while outside this curve, each point stands on only two normal lines from the ellipse. The curve is known as the *evolute of the ellipse*. After addressing the caveat, we will return to the evolute.[7]

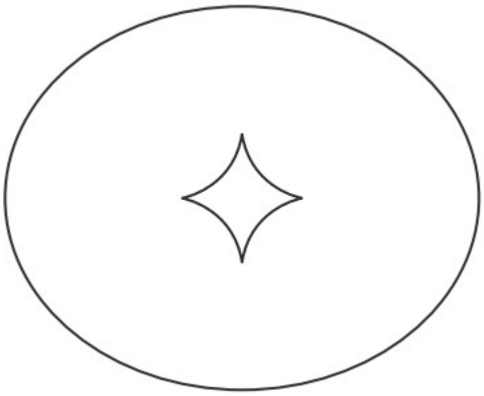

Figure 4.5. The evolute of the ellipse. The evolute consists of all points having three normal lines to the ellipse.

---

[7] Because ancient Greek mathematicians, who called themselves *geometers*, did not express functional relationships as we do today, Apollonius did not find an expression for the evolute that is equivalent

-3 / ## # # # > ! !/ ! 99 0
/! ! !" ! ? / ' @ !

At this point, a judge would most certainly uphold an objection that the reader may wish to place forward. Twice in the derivation the author was less than careful concerning possible divisions by zero. This occurs when determining the slope value m once using the ellipse and once using the tangent hyperbola. The reader can verify that all the divisions by zero occur when either x and x_0 are zero or y and y_0 are zero; that is, the location of all illegal operations is on the x axis or the y axis. Using symmetry, one can deduce that any point on either axis stands on an even number of normals, precluding the possibility of standing on only three normals. One can see this for a point on the x axis as follows. If (x_0, y_0) lies on the x axis, then the point stands on two normals emanating from the points $(a, 0)$ and $(-a, 0)$. Suppose that there is a third normal emanating from the ellipse at the point (x_n, y_n). Then, by symmetry, another normal emanates from the point $(x_n, -y_n)$, precluding the possibility of only three normals. There can be only two or four normals. Figure 4.6 illustrates the argument; the point (x_0, y_0) lies at the corner of the evolute. On the corner point there are only two normals. A similar argument applies to points on the y axis. The outcome is that points on both the evolute and the coordinate axes stand on only two normal lines, not three. With this statement the caveat has been put to rest.

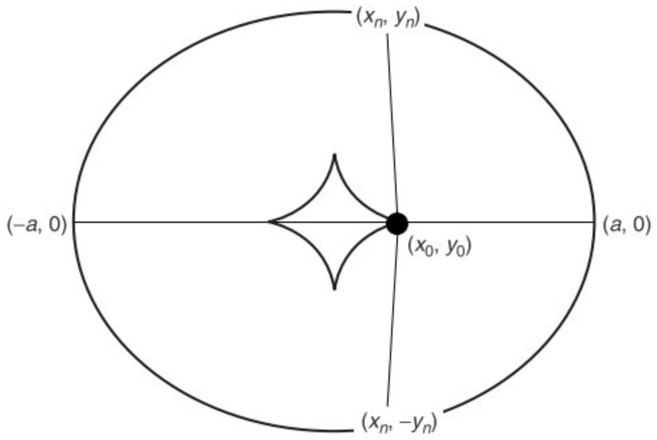

Figure 4.6. A point on the x axis cannot be on the evolute as well. By symmetry, there must be an even number of normals to a point on the x axis.

to equation (4.8). Nevertheless, he did find all the relationships necessary to derive the evolute. In particular, Apollonius writes relations similar to that of the ellipse [eq. (4.1)], the hyperbola [eq. (4.3)], and the slope of the tangent hyperbola [eq. (4.5)]. In his book, *A History of Greek Mathematics*, Heath uses Apollonius' relations to derive the evolute in its modern form (Heath 1921).

The Radius of Curvature

In the spirit of this chapter, just for fun, we now take an alternative path to the evolute. The path leads to a generalized evolute for any given curve, but the solution for the ellipse indicates a direction. For the ellipse, one may consider the point (x, y) at which the hyperbola of equation (4.3) becomes tangent as a limiting point at which two intersections of the receding hyperbola collide. The corresponding point on the evolute, namely, (x_0, y_0), is then the limiting point of intersection for two colliding normal lines (see Figure 4.7). Let us calculate the limiting point of intersection for a general curve given by $f(x, y) = 0$.

Before proceeding, some notation is necessary. Let $\mathbf{n}(x, y)$ be the unit normal vector to the curve having components $\mathbf{n} = (n_1, n_2)$. Note that both components are also functions: $n_1 = n_1(x, y)$ and $n_2 = n_2(x, y)$. Also since $\mathbf{n}$ is the unit normal, its length is one: $n_1^2 + n_2^2 = 1$. Below, we will indicate the direction in which $\mathbf{n}(x, y)$ points. Let (x, y) be on the curve $f(x, y) = 0$ and (x_a, y_a) also be on the curve with x_a close to x so that $x_a = x + \Delta$. Associated with both points (x, y) and (x_a, y_a) are normal lines that intersect at some point denoted by (x_0, y_0). With this notation the problem is to determine $\lim_{\Delta \to 0}(x_0, y_0)$ (see Figure 4.8).

The expression $(x + n_1(x, y)r, y + n_2(x, y)r)$ yields the normal line to (x, y) as the value r ranges over all real numbers. Similarly, the expression $(x_a + n_1(x_a, y_a)s, y_a + n_2(x_a, y_a)s)$ yields the normal line to (x_a, y_a) as the value s ranges over all real numbers. Because it is at the intersection of these two lines, the point (x_0, y_0) must satisfy the following set of equations for specified values of r and s:

$$\begin{aligned} x_0 &= x + n_1(x, y)r = x_a + n_1(x_a, y_a)s \\ y_0 &= y + n_2(x, y)r = y_a + n_2(x_a, y_a)s \end{aligned} \tag{4.10}$$

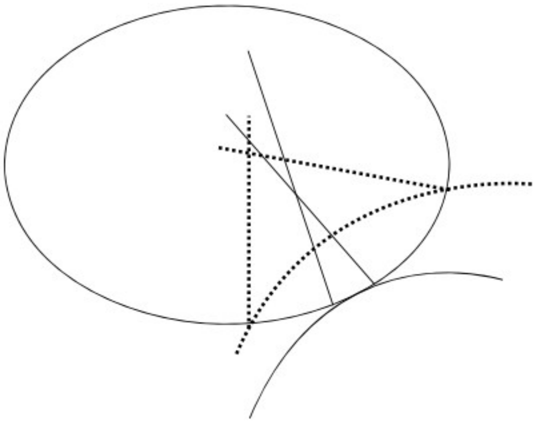

Figure 4.7. Intersections of normals associated with hyperbolas. As the hyperbolas recede and become tangent to the ellipse, the intersections approach a point on the evolute.

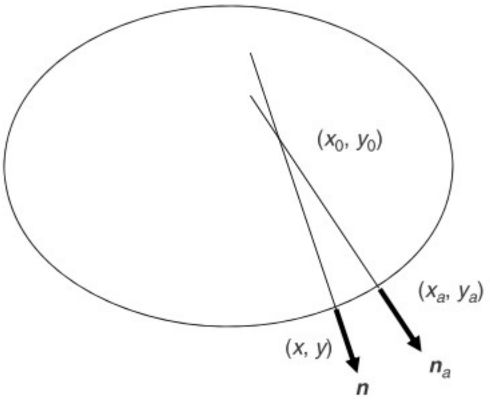

Figure 4.8. The limit of (x_0, y_0) as (x_a, y_a) approaches (x, y) is a point on the evolute.

These equations allow one to solve for the values of r and s that specify the point of intersection. The values r and s converge in the limit as $\Delta \to 0$. Accordingly, we focus on r:

$$\begin{aligned} n_1(x, y)r - n_1(x_a, y_a)s &= x_a - x \\ n_2(x, y)r - n_2(x_a, y_a)s &= y_a - y \end{aligned}$$

Eliminating the s variable and solving for r yields the following equation:

$$\begin{aligned} (n_1(x, y)n_2(x_a, y_a) - n_2(x, y)n_1(x_a, y_a))\, r = n_2(x_a, y_a)(x_a - x) \\ -n_1(x_a, y_a)(y_a - y) \end{aligned}$$

From this equation one obtains the value of r:

$$r = \frac{n_2(x_a, y_a)(x_a - x) - n_1(x_a, y_a)(y_a - y)}{n_1(x, y)n_2(x_a, y_a) - n_2(x, y)n_1(x_a, y_a)} \tag{4.11}$$

The objective is to determine $\lim_{\Delta \to 0} r$. Toward this end, the expressions are expanded to first order as a series in Δ. By assumption, the term $x_a - x = \Delta$. Next, let's determine the series expansion for the term y_a in terms of Δ, $y_a = y(x) + y'(x)\Delta +$ higher-order terms. To accomplish this, we assume that the equation $f(x, y) = 0$ allows one to express y as a differentiable function of x.[8] With this assumption, we note the following:

$$f(x_a, y_a) = f(x, y) + \frac{\partial f}{\partial x}\Delta + \frac{\partial f}{\partial y}\Delta_y + \text{higher-order terms}$$

---

[8] The necessary conditions are known from the implicit function theorem.

Both (x, y) and (x_a, y_a) lie on the curve yielding the following expression:

$$\frac{\partial f}{\partial x}\Delta + \frac{\partial f}{\partial y}\Delta_y + \text{higher-order terms} = 0$$
$$\frac{\partial f}{\partial x}\Delta + \frac{\partial f}{\partial y}\frac{dy}{dx}\Delta + \text{higher-order terms} = 0$$
$$\left(\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{dy}{dx}\right)\Delta + \text{higher-order terms} = 0$$

It follows that the coefficient $\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{dy}{dx}$ is identically zero, from which we calculate the derivative, $y'(x)$:

$$\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{dy}{dx} = 0$$
$$y'(x) = \frac{dy}{dx} = -\frac{\frac{\partial f}{\partial x}}{\frac{\partial f}{\partial y}}$$

Since the gradient and normal vectors are aligned, $\frac{\frac{\partial f}{\partial x}}{\frac{\partial f}{\partial y}} = \frac{n_1}{n_2}$ and we obtain the result for $y'(x)$.

$$y'(x) = -\frac{n_1}{n_2} \tag{4.12}$$

Finally, we determine the expansion for the normal components, $n_1(x_a, y_a)$ and $n_2(x_a, y_a)$, in which the higher-order terms are ignored:

$$\begin{aligned} n_1(x_a, y_a) &= n_1(x, y) + \frac{\partial n_1}{\partial x}\Delta + \frac{\partial n_1}{\partial y}\Delta_y \\ &= n_1(x, y) + \frac{\partial n_1}{\partial x}\Delta + \frac{\partial n_1}{\partial y}\frac{dy}{dx}\Delta \\ &= n_1(x, y) + \frac{\partial n_1}{\partial x}\Delta - \frac{\partial n_1}{\partial y}\frac{n_1}{n_2}\Delta \\ &= n_1(x, y) + \left(\frac{\partial n_1}{\partial x} - \frac{\partial n_1}{\partial y}\frac{n_1}{n_2}\right)\Delta \end{aligned} \tag{4.13}$$

Similarly, the expansion for $n_2(x_a, y_a)$ is the following:

$$n_2(x_a, y_a) = n_2(x, y) + \left(\frac{\partial n_2}{\partial x} - \frac{\partial n_2}{\partial y}\frac{n_1}{n_2}\right)\Delta \tag{4.14}$$

Substituting the first-order expansions [eqs. (4.13) and (4.14)] into equation (4.11) results in the following expression for r:

$$r = \frac{\left[n_2 + \left(\frac{\partial n_2}{\partial x} - \frac{\partial n_2}{\partial y}\frac{n_1}{n_2}\right)\Delta\right]\Delta + \left[n_1 + \left(\frac{\partial n_1}{\partial x} - \frac{\partial n_1}{\partial y}\frac{n_1}{n_2}\right)\Delta\right]\frac{n_1}{n_2}\Delta}{n_1\left[n_2 + \left(\frac{\partial n_2}{\partial x} - \frac{\partial n_2}{\partial y}\frac{n_1}{n_2}\right)\Delta\right] - n_2\left[n_1 + \left(\frac{\partial n_1}{\partial x} - \frac{\partial n_1}{\partial y}\frac{n_1}{n_2}\right)\Delta\right]}$$

Eliminating all higher-order terms, simplifying, and taking the limit as $\Delta \to 0$ results in the desired expression:

$$\lim_{\Delta\to 0} r = \lim_{\Delta\to 0} \frac{\left[n_2 + \left(\frac{\partial n_2}{\partial x} - \frac{\partial n_2}{\partial y}\frac{n_1}{n_2}\right)\Delta\right]\Delta + \left[n_1 + \left(\frac{\partial n_1}{\partial x} - \frac{\partial n_1}{\partial y}\frac{n_1}{n_2}\right)\Delta\right]\frac{n_1}{n_2}\Delta}{n_1\left[n_2 + \left(\frac{\partial n_2}{\partial x} - \frac{\partial n_2}{\partial y}\frac{n_1}{n_2}\right)\Delta\right] - n_2\left[n_1 + \left(\frac{\partial n_1}{\partial x} - \frac{\partial n_1}{\partial y}\frac{n_1}{n_2}\right)\Delta\right]}$$

$$\lim_{\Delta\to 0} r = \lim_{\Delta\to 0} \frac{n_2\Delta + n_1\frac{n_1}{n_2}\Delta}{n_1\left(\frac{\partial n_2}{\partial x} - \frac{\partial n_2}{\partial y}\frac{n_1}{n_2}\right)\Delta - n_2\left(\frac{\partial n_1}{\partial x} - \frac{\partial n_1}{\partial y}\frac{n_1}{n_2}\right)\Delta}$$

$$\lim_{\Delta\to 0} r = \frac{n_2 + n_1\frac{n_1}{n_2}}{n_1\left(\frac{\partial n_2}{\partial x} - \frac{\partial n_2}{\partial y}\frac{n_1}{n_2}\right) - n_2\left(\frac{\partial n_1}{\partial x} - \frac{\partial n_1}{\partial y}\frac{n_1}{n_2}\right)}$$

$$\lim_{\Delta\to 0} r = \frac{n_2^2 + n_1^2}{n_1 n_2\left(\frac{\partial n_2}{\partial x} - \frac{\partial n_2}{\partial y}\frac{n_1}{n_2}\right) - n_2^2\left(\frac{\partial n_1}{\partial x} - \frac{\partial n_1}{\partial y}\frac{n_1}{n_2}\right)}$$

$$\lim_{\Delta\to 0} r = \frac{1}{n_1 n_2\left(\frac{\partial n_2}{\partial x} - \frac{\partial n_2}{\partial y}\frac{n_1}{n_2}\right) - n_2^2\left(\frac{\partial n_1}{\partial x} - \frac{\partial n_1}{\partial y}\frac{n_1}{n_2}\right)}$$

One final simplification provides the result:

$$\lim_{\Delta\to 0} r = \frac{1}{n_1 n_2\left(\frac{\partial n_2}{\partial x} + \frac{\partial n_1}{\partial y}\right) - \left(n_1^2\frac{\partial n_2}{\partial y} + n_2^2\frac{\partial n_1}{\partial x}\right)} \tag{4.15}$$

To complete the description of the limiting value, we take the direction of **n** in such a manner to ensure that the limit of equation (4.15) is positive. Once the limiting value of r is available, the limiting expressions for x_0 and y_0 are also available. Below we use the following notation for the limiting values:

$$\lim_{\Delta\to 0} r = r^* \quad \text{and} \quad \lim_{\Delta\to 0} (x_0, y_0) = (x_0^*, y_0^*)$$

By definition, the point (x_0^*, y_0^*) is on the evolute. With this notation, the following identities hold:

$$x_0^* = x + n_1 r^* \quad \text{and} \quad y_0^* = y + n_2 r^* \tag{4.16}$$

The Osculating Circle

We could return to the ellipse and use equations (4.16) to establish a relationship between x_0^* and y_0^*. But there is something else even more fun.

Note that r^* is a function of the point (x, y), $r^* = r^*(x, y)$, and the value r^* is the distance from the point (x, y) to the point (x_0^*, y_0^*). The distance r^* is known as the radius of curvature[9] at the point (x, y), and the point (x_0^*, y_0^*) is the center of curvature. Assume that the original curve, $f(x, y) = 0$, is a circle of radius R. Then, because all normal lines intersect at the center, r^* is the same value as R. Now return to the general case, $f(x, y) = 0$, any arbitrary curve. Because, at a given point (x, y), the general curve and a circle tangent to the curve at (x, y) with radius r^* both yield the same value for the radius of curvature, the curves are similar around the point (x, y). In fact, among all tangent circles, that of radius r^* is the circle that best approximates the original curve, as we next demonstrate.

First consider what is meant by the "best approximating circle to the original curve at a point (x, y)." Three choices specify any circle: two to designate its center, (x_c, y_c), and one for the radius, R. Let (x_a, y_a) be a point on the circle in the vicinity of (x, y), and express y_a as a function of x_a.[10] As above, it is possible to expand the functions in a Taylor series about the value x. A similar Taylor series about the value x is available using the original curve, $f(x, y) = 0$. As there are three parameter choices for the circle, it is possible to choose them so that the first three terms of each Taylor series formed from the circle match the first three terms from the Taylor series formed from the original curve. The circle that matches the first three terms is the best approximating circle at the point (x,y). Let us find this circle.

The following expression is the Taylor series of a function expanded about the value x:

$$y_a = y(x) + y'(x)\Delta + \tfrac{1}{2}y''(x)\Delta^2 + \text{higher-order terms} \tag{4.17}$$

The first three terms of two power series are equal when the value of y and its first two derivatives are the same at the given value x. Equation (4.16) gives the value of the first term, $y = y_0^* - r^* n_1$. Equation (4.12) provides the first-order term, $y'(x)\Delta = -\frac{n_1}{n_2}\Delta$; equivalently, $y'(x) = -\frac{n_1}{n_2}$.

Differentiating $y'(x)$ gives the second derivative:

$$y''(x) = -\frac{n_2\left(\dfrac{dn_1}{dx}\right) - n_1\left(\dfrac{dn_2}{dx}\right)}{n_2^2}$$

---

[9] The value $\kappa = \frac{1}{r^*}$ is known as the *curvature*.

[10] We assume that the conditions of the implicit function theorem hold so that the expression holds.

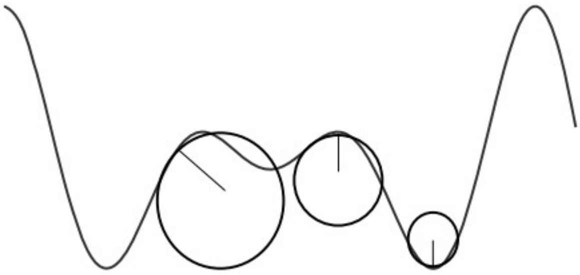

Figure 4.9. The osculating circle at different points on a curve.

Since y_a is considered as a function of x_a, the normal component n_1 can be considered as a function of x_a with expansion, $n_1(x_a) = n_1(x) + \frac{d}{dx}n_1(x)\Delta$ + higher-order terms. Equation (4.13) provides the first-order term of the expansion yielding the expression $\frac{d}{dx}n_1(x) = \left(\frac{\partial n_1}{\partial x} - \frac{\partial n_1}{\partial y}\frac{n_1}{n_2}\right)$. Similarly, from equation (4.14), $\frac{d}{dx}n_2(x) = \left(\frac{\partial n_2}{\partial x} - \frac{\partial n_2}{\partial y}\frac{n_1}{n_2}\right)$. Using these expressions and equation (4.15) results in the following expression for the second derivative of y:[11]

$$\begin{aligned} y''(x) &= -\frac{n_2\left(\dfrac{\partial n_1}{\partial x} - \dfrac{\partial n_1}{\partial y}\dfrac{n_1}{n_2}\right) - n_1\left(\dfrac{\partial n_2}{\partial x} - \dfrac{\partial n_2}{\partial y}\dfrac{n_1}{n_2}\right)}{n_2^2} \\ &= -\frac{n_2^2\left(\dfrac{\partial n_1}{\partial x} - \dfrac{\partial n_1}{\partial y}\dfrac{n_1}{n_2}\right) - n_2 n_1\left(\dfrac{\partial n_2}{\partial x} - \dfrac{\partial n_2}{\partial y}\dfrac{n_1}{n_2}\right)}{n_2^3} \\ &= \frac{1}{r^* n_2^3} \end{aligned} \tag{4.18}$$

For any curve, the three values $y(x)$, $y'(x)$, and $y''(x)$ depend only on the curve's normal components n_1 and n_2, the curve's radius of curvature r^*, and the curve's center of curvature (x_0^*, y_0^*). As such, the first three terms in a Taylor series of two curves sharing common normal components, radius of curvature, and center of curvature are identical. Figure 4.9 illustrates three circles tangent to three different points along a curve. The circles share a common normal with the curves at their points of intersection. Also, the center of each circle is the same as the curve's center of curvature: $(x_c, y_c) = (x_0^*, y_0^*)$. The circle is called the *osculating circle*.

The Evolute and the Osculating Circle

Associated with each point along a curve is an osculating circle and the osculating circle's center. The collection of all such centers forms another curve and this new curve defines the evolute (see Figure 4.10). The preceding sections demonstrate that the points sitting on three normals from an ellipse are on the evolute of the ellipse. Our initial evaluation of the evolute didn't even apply the evolute's definition. It was an accidental outcome, and we weren't aware that the curve of equation (4.9) satisfies the properties of the evolute. Below, we calculate the evolute of the ellipse from its definition and show that the curve is the same curve resulting from the initial approach. The calculation simplifies equations (4.13) and (4.14) to obtain the expressions (4.7) and (4.8).

---

[11] The result for the second derivative is frequently written as $y''(x) = \kappa/n_2^3$, where κ is the curvature.

We begin by calculating the unit normals to the ellipse. We have already found normal vectors by taking the gradient of the function $f(x, y) = \frac{x^2}{a^2} + \frac{y^2}{b^2}$. The gradient is $\left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right) = \left(\frac{2x}{a^2}, \frac{2y}{b^2}\right)$. Normalizing the gradient and pointing the vectors inward yields the components of the unit normals:

$$n_1 = -\frac{x}{a^2\sqrt{\dfrac{x^2}{a^4} + \dfrac{y^2}{b^4}}}$$

$$n_2 = -\frac{y}{b^2\sqrt{\dfrac{x^2}{a^4} + \dfrac{y^2}{b^4}}}$$

Next, using equation (4.18), we obtain the following expression for the radius of curvature r^*:

$$r^* = \frac{1}{y''(x)n_2^3}$$

The derivative $y'(x) = -\frac{n_1}{n_2} = -\frac{b^2x}{a^2y}$. Differentiating this value results in the second derivative.

$$\begin{aligned}
y''(x) &= -\frac{b^2}{a^2}\left(\frac{y + \dfrac{b^2x^2}{a^2y}}{y^2}\right) \\
&= -\frac{b^2}{a^4y^3}\left(a^2y^2 + b^2x^2\right) \\
&= -\frac{b^2}{a^4y^3}\left(a^2b^2\right) \qquad \text{from the equation of an ellipse} \\
&= -\frac{b^4}{a^2y^3}
\end{aligned}$$

Placing these expressions into equations (4.16) yields the desired expression:

$$\begin{aligned}
x_0^* &= x + r^* n_1 \\
&= x + \frac{n_1}{y''(x) n_2^3} \\
&= x - \frac{x}{a^2\sqrt{\dfrac{x^2}{a^4} + \dfrac{y^2}{b^4}}} \frac{a^2 y^3}{b^4} \frac{\left(b^2\sqrt{\dfrac{x^2}{a^4} + \dfrac{y^2}{b^4}}\right)^3}{y^3} \\
&= x - b^2 x \left(\frac{x^2}{a^4} + \frac{y^2}{b^4}\right) \\
&= x - b^2 x \left(\frac{x^2}{a^4} + \frac{1}{b^2}\left(1 - \frac{x^2}{a^2}\right)\right) \\
&= x - \frac{b^2 x^3}{a^4} - x + \frac{x^3}{a^4} \\
&= \frac{a^2 - b^2}{a^4} x^3
\end{aligned}$$

A similar calculation beginning with $y_0^* = y + r^* n_2$ results in the equation $y_0^* = \frac{b^2 - a^2}{a^4} y^3$. Comparing this with equations (4.7) and (4.8), we find all the

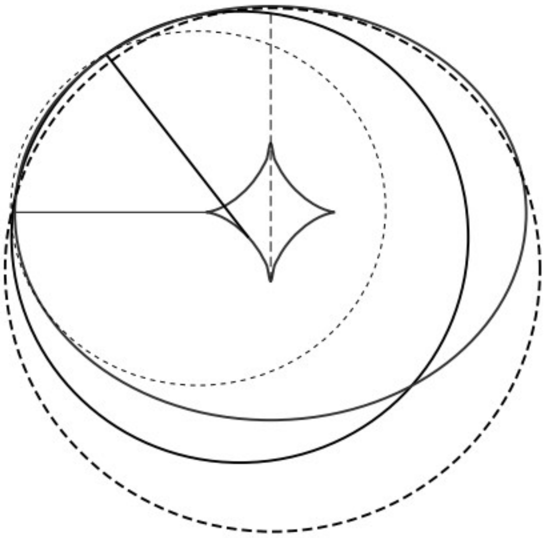

Figure 4.10. The evolute constructed from the centers of the osculating circles.

points (x_0, y_0) coincide with (x_0^*, y_0^*), affirming that the curve of expression 4.9, $(ax_0)^{2/3} + (by_0)^{2/3} = \left(a^2 - b^2\right)^{2/3}$ is, indeed, the evolute.

Figure 4.10 illustrates the construction of an ellipse's evolute using the family of osculating circles.

FUN'S HARVEST

An unanswered question is Apollonius' path to the evolute for the ellipse. Did he utilize the tangent hyperbola as in our initial approach, unaware of the osculating circle; discover and utilize the osculating as in our latter approach; or come to it through some other process? If Apollonius was aware of the osculating circle and its radius of curvature, he did not present its properties in his manuscript *On Conics*, but he most certainly motivated others to pursue it.

Christiaan Huygens (1629–1695) explicitly defines the radius of curvature and the osculating circle for for general curves. Leibniz applies his methods of calculus to determine the radius of curvature. Gauss (1777–1855), Riemann (1826–1866), and Ricci-Curbastro (1853–1925) create a general setting for non-Euclidean geometry in which the curvature is central. Then Einstein (1879–1955) incorporates curvature into his field equations of general relativity. Mathematics is fun, but not frivolous.

b) Pag. 49(66) Fig 4.4

Intersection of hyperbola with ellipse: three intersections, the normal lines.

Pag. 53(70) Fig. 4.6

Pag 54(71) The radius of curvature

Fig. 4.7 Intersections of normals associated with hyperbola. As the hyperbolas recede and become tangent to the ellipse, the intersections approach a point on the evolute.

Import

Pag. 56(75) The Osculating Circle

Fig. 4.9 The osculating circle at different points on a curve

c) Pag. 85(102) - Figure 5.12

Three consecutive oppositions.

Note that the equations yield real values as long as $d < 1$, meaning that the distance of the deferent's center from Earth is smaller than the deferent's radius.[15]

The equation for Mars' actual coordinates [eq. (5.2)] is correct if one uses the values m_1 and m_2 of equations (5.6) and (5.7).

Before we continue, one feature of the model is noteworthy. As the longitudinal angle θ is not centered at the center of Mars' deferent, the speed of Mars' mean about the deferent is not uniform. Figure 5.9 illustrates the manner in which the speed varies. The equant-centered angles A and B are equal, so Mars' mean travels along each segment embraced by the angles' arms in an equal time. Because the segment embraced by B is longer than that embraced by A, the speed of Mars' mean is faster along the segment embraced by B. In general, the speed of Mars' mean decreases as Mars' mean approaches the equant and increases as Mars' mean recedes from the equant. As we will see in the next chapter, Copernicus found this feature most distasteful, desiring instead uniform circular motion. Kepler, on the other hand, viewed this feature favorably.

The Martian Year. Ptolemy's Martian year is the time that Mars' mean requires to trot once about the deferent. This is the same as the requirement for the longitude θ to advance by 360 degrees. Ptolemy uses an ingenious relationship between his choice of angle to represent the anomaly (referencing the apogee) and the occurrence of opposition. Figure 5.10 illustrates this relationship. The

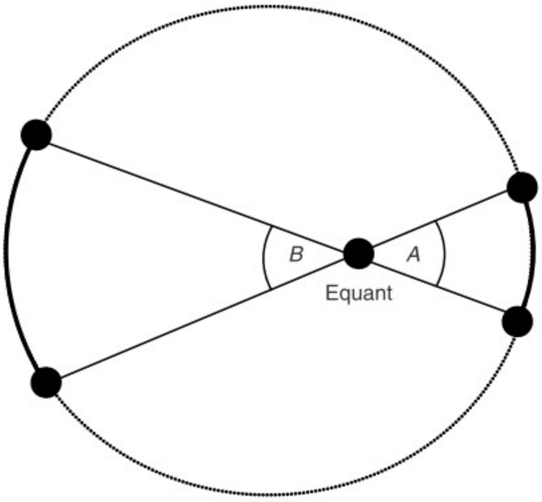

Figure 5.9. Equal angles about the equant indicate equal travel times for Mars' mean about the deferent. The size of the arc increases with distance from the equant corresponding to a higher speed.

---

[15] A quadratic formula yielding two solutions arises. We choose the solution lying on the relevant ray.

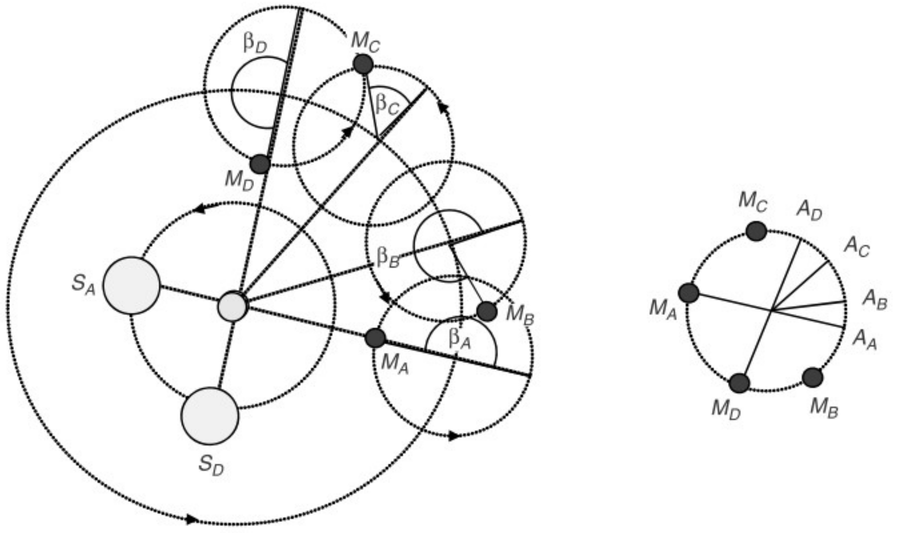

(a) (b)

Figure 5.10. The change in Ptolemy's anomaly between consecutive oppositions; the anomaly completes one full circuit as both the initial and final anomalies are 180 degrees: (a) four observations; (b) corresponding positions of apogee and Mars on epicycle.

angle β is Ptolemy's measurement of the anomaly. As the figure illustrates, the angle executes one cycle between occurrences of opposition.

Because the speed of Mars' mean is not uniform about the deferent, the angle β does not vary with uniform speed. Indeed, if it did so, then the time between occurrences of central retrograde would be constant. However, Ptolemy reasoned that the rate of change in β has a long-run average. He also correctly reasoned that the long-run average of the rate of change in the angle β would be the same as the constant rate of change given by the angle β_Q of Figure 5.11:

$$\beta_Q = \alpha - \theta \tag{5.8}$$

By using a long-term historical record that provides the long-run changes in both α and β, the long-run change in θ is available. Dividing this change by the timeframe of the historical record yields the rate of change in θ. Consistent with previous notation, we indicate the rate of change in θ by ω. Once ω is known, one can calculate the Martian year. The following equations express the relevant relations:

$$\theta_0 = \alpha_0 - \beta_{Q0} \tag{5.9}$$

$$\theta_1 = \alpha_1 - \beta_{Q1} \tag{5.10}$$

$$\Delta\theta = \theta_1 - \theta_0 \tag{5.11}$$

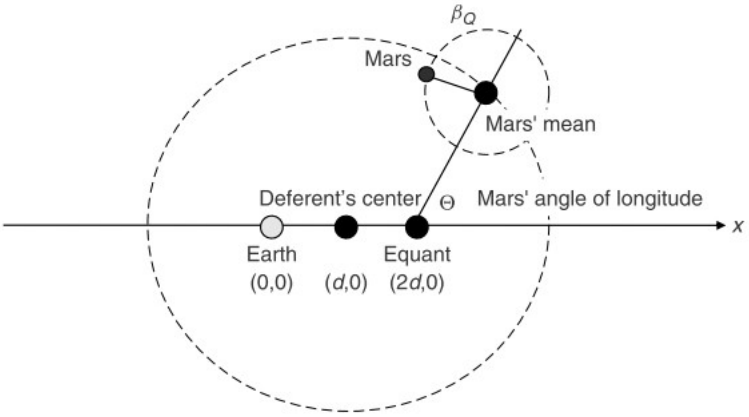

Figure 5.11. The angle β_Q.

$$= (\alpha_1 - \alpha_0) - (\beta_{Q1} - \beta_{Q0}) \tag{5.12}$$

$$\omega = \frac{\Delta\theta}{T} \tag{5.13}$$

in which the subscript 0 denotes observations at the beginning of the historical record, the subscript 1 denotes observations at the end of the historical record, and T denotes the time difference between the initial and final observations.

Ptolemy had a historical record available. Hipparchus had proposed this calculation and initiated the necessary observations. The record indicates 38 occurrences of opposition in 79.0095615[16] solar years or equivalently 37 cycles of the angle β during the timeframe $T = 79.0095615$. The angle β then changed by $37 \times 360 = 13{,}320$; that is, $\beta_{Q1} - \beta_{Q0} = 13{,}320$. We know that the angle α changes by 360 degrees in one solar year's time, so over the time range of observations, α changed by $360 \times 79.0095615 = 28{,}443.44214$ degrees; specifically, $\alpha_1 - \alpha_0$. Using equation (5.11), we find that the angle θ changed by 15,123.44214 degrees, from which one obtains the rate $\omega = \frac{15123.44214}{79.0095615} = 191.4128094$ degrees per solar year [eq. (5.12)]. With this rate of change in the equant-centered longitude, one obtains the Martian year at $\frac{360}{191.4128094} = 1.880752$ solar years.[17]

Ptolemy's Juggernaut. Ptolemy set the objective of determining all of the parameters of his model in a manner that is consistent with the observations of

---

[16] Ptolemy presents the historical record in Chapter IX of *The Almagest*. I backed into the number of years in order to match the rate of change in the Martian longitude that my translated version of *The Almagest* presents. The number of solar years differs slightly from that in the numbers I read where it sets the time frame at 79 solar years, 3 and $\frac{13}{60}$ days or equivalently, 79.008807 solar years. Confusing matters further, the translation of *The Almagest* states that the equivalent number of days is $28857\frac{53}{60}$, which is inconsistent with the number of solar years.

[17] The preceding calculation is made with a solar year equaling 365.25 days. Below, calculations are made by the year in accordance with the ancient Egyptian calendar that fixes each year at 365 days. The rate of rotation with respect to the Egyptian year is $\omega = 191.281794599$.

Table 5.1. For convenience, we once again write the full model and emphasize the variables that are time-dependent:

$$M_1(t) = m_1(t) + r_e \cos\alpha(t) \tag{5.14}$$

$$M_2(t) = m_2(t) + r_e \sin\alpha(t) \tag{5.15}$$

$$m_1(t) = \left(-d\cos\theta(t) + \sqrt{1 - d^2\sin^2\theta(t)}\right)\cos\theta(t) + 2d \tag{5.16}$$

$$m_2(t) = \left(-d\cos\theta(t) + \sqrt{1 - d^2\sin^2\theta(t)}\right)\sin\theta(t) \tag{5.17}$$

$$S_1(t) = r_S \cos\lambda(t) \tag{5.18}$$

$$S_2(t) = r_S \sin\lambda(t) \tag{5.19}$$

$$\lambda(t) = \lambda_A + \omega_\lambda t \tag{5.20}$$

$$\alpha(t) = \lambda(t) \tag{5.21}$$

$$\theta(t) = \theta_A + \omega_\theta t \tag{5.22}$$

The model variables are

t	Time measured in solar years
$(M_1(t), M_2(t))$	Coordinates of Mars
$(m_1(t), m_2(t))$	Coordinates of Mars' mean (center of Mars' epicycle)
$(S_1(t), S_2(t))$	Coordinates of the sun
$\lambda(t)$	Solar longitude
$\alpha(t)$	Mars' anomaly
$\theta(t)$	Mars' equant-centered longitude

Note that the model centers the sun's deferent on Earth.

There are eight model parameters:

r_e	Radius of Mars' epicycle
d	Distance from center of Mars' deferent to Earth, which is also half the distance from Mars' equant to Earth
r_S	Radius of the sun's deferent
λ_A	Initial solar longitude, which is also Mars' initial anomaly
θ_A	Initial longitude of Mars' mean
ω_λ	Angular speed of the sun's longitude
ω_θ	Angular speed of Mars' equant-centered longitude

An additional parameter has already been set; the radius of Mars' deferent is 1.

Ptolemy must find values for all of the parameters that ensure coincidence between the model and the observations. Aside from the radius of the deferent of Mars', two others have been set: the angular speed of the sun's longitude,

$\omega_\lambda = 360$ degrees per solar year and the angular speed of Mars' equant-centered longitude, which was found above to be 191.281794599 degrees per year. The values of the solar radius and the radius of Mars' epicycle do not impact the observations of Table 5.1. Table 5.1 gives observations of opposition, and only the sun's longitude along with Mars' mean and Mars' anomaly matter; the observations would occur with the correct solar longitude, location of Mars' mean, and anomaly regardless of the solar and epicyclic radii[18]. Accordingly, we will not consider the solar radius or the radius of Mars' epicycle. There are only three remaining parameters to specify: d, the distance from the center of Mars' deferent to Earth; λ_A, the initial solar longitude, which is also Mars' initial anomaly; and θ_A, the initial longitude of Mars' mean. Following Ptolemy, we enforce consistency between the observations and the model by selecting appropriate values for the undetermined parameters.

Figure 5.12 illustrates the model parameters and notation. The figure places Mars' mean at several observations of central retrograde; the mean lies on the line that passes through the sun, Earth, and Mars with the sun and Mars in opposition from Earth. The figure illustrates the notation that is used below. A variable with a letter as a subscript indicates the observation associated with that variable. For example, the figure illustrates Mars' mean longitude at the second observation θ_B and the sun's longitude at the third observation λ_C. With this notation, the parameters that we must determine are θ_A, Mars' initial longitude λ_A, and d.

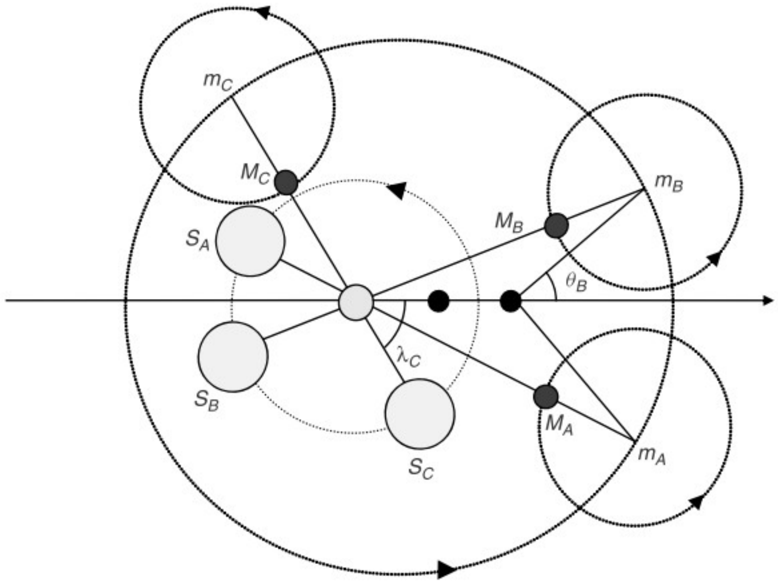

Figure 5.12. Three consecutive oppositions. Angles θ_B and λ_C are displayed.

---

[18] The radii could not be too big. The sun always remains the closest body to Earth.

Coincidence of the model and observations requires that the model's time lapse between occurrences of opposition is in agreement with the observed elapsed times of Table 5.1. Ptolemy requires a solution method to determine the parameters that ensure coincidence. Before continuing we remark on the number of free parameters. While there appear to be three free parameters, there are in fact only two, either d and λ_A or d and θ_A. To see this, let us assume that the two parameters d and λ_A are determined. Then the remaining angle θ_A is not free, but must be set by the condition that the observations are in opposition and only one value for θ_A modulo 360 degrees meets the requirement. Using a similar argument, one notes that a determination of the parameters d and θ_A sets the parameter λ_A.

As noted below, only two data points are used to determine two free parameters, the time interval between the first and second observations, and then the time interval between the first and third observations. Three observations are necessary to gather these two data points. Note that once the position of the sun and Mars' mean are correctly determined, one reconciles Mars' actual position by setting the initial value of Mars' anomaly equal to the initial value of the solar longitude, $\alpha_A = \lambda_A$. Accordingly, in the calculations below we need not further consider Mars' actual position, only its mean position.[19]

Table 5.2 gives the difference in the equant-centered Martian longitude $\Delta\theta$, where the difference is with respect to the first observation. This value is determined by multiplying the rate of change in θ (i.e., ω_θ) by the elapsed time between the observations. An equivalent calculation is made for the difference in solar longitude: $\Delta\lambda = \Delta\alpha$. Table 5.2 also displays the reduced angles, whereby the corresponding angle is reset to zero whenever it surpasses a multiple of 360 degrees. As noted above, the table gives two data points: the time interval between the first and second observations, and then the time interval between the first and third observations.

We begin by establishing a relation between the distance from Earth to the equant $2d$ and the remaining parameters. At central retrograde, when Mars' mean is in opposition to the sun, the coordinates of Mars' mean with reference to the

TABLE 5.2. Differences in Critical Angles

Observation	Year	Years Elapsed	$\Delta\lambda = \Delta\alpha$	$\Delta\theta$
A	15 of Hadrian	—	—	—
B	19 of Hadrian	4.191324	1507.83 → 67.83	801.73 → 81.73
C	2 of Antonine	8.454452	3041.57 → 161.57	1617.20 → 177.20

---

[19] The reader should note that none of the calculations that follow utilize the radius of the epicycle r_e. As described above, this means that coincidence between the observations of the table and the model is independent of the radius. Ptolemy then has one extra parameter that he may freely choose to match additional observations, separate from Table 5.1. The reader may consult Ptolemy's *Almagest* for details.

solar longitude are

$$m_1 = -h\cos\lambda$$
$$m_2 = -h\sin\lambda$$

where h is the distance between Earth and Mars' mean, $h = \sqrt{m_1^2 + m_2^2}$.

Equations (5.15) and (5.16) are used to determine the following expression for h:

$$h = \left\{\left[\left(-d\cos\theta + \sqrt{1 - d^2\sin^2\theta}\right)\cos\theta + 2d\right]^2 + \left(-d\cos\theta + \sqrt{1 - d^2\sin^2\theta}\right)^2\sin^2\theta\right\}^{1/2} \tag{5.23}$$

Mars' mean resides on the line $m_1 = m_2\cot\theta + 2d$, where $2d$ is the distance from Earth to the equant.

Substituting the expressions for Mars' mean coordinates into the line and simplifying yields the relation for $2d$ that we are seeking:

$$\begin{aligned}
-h\cos\lambda &= -h\sin\lambda\cot\theta + 2d \\
2d &= h(\sin\lambda\cot\theta - \cos\lambda) \\
2d\sin\theta &= h(\sin\lambda\cos\theta - \cos\lambda\sin\theta) \\
2d\sin\theta &= h\sin(\lambda - \theta) \\
2d &= \frac{h\sin(\lambda - \theta)}{\sin\theta}
\end{aligned} \tag{5.24}$$

Equation (5.24) is valid at all three observation points. Equating the result at the first observation A with the result at either of the other observations and simplifying yields an expression relating the angle $\lambda_A - \theta_A$ with remaining parameters:

$$\begin{aligned}
\frac{h_A\sin(\lambda_A - \theta_A)}{\sin\theta_A} &= \frac{h\sin(\lambda - \theta)}{\sin\theta} \\
\frac{h_A\sin(\lambda_A - \theta_A)}{\sin\theta_A} &= \frac{h\sin(\lambda_A - \theta_A + \Delta\lambda - \Delta\theta)}{\sin(\theta_A + \Delta\theta)} \\
\frac{h_A\sin(\lambda_A - \theta_A)\sin(\theta_A + \Delta\theta)}{h\sin(\theta_A)} &= \sin(\lambda_A - \theta_A + \Delta\lambda - \Delta\theta) \\
\frac{h_A\sin(\lambda_A - \theta_A)\sin(\theta_A + \Delta\theta)}{h\sin(\theta_A)} &= \sin(\lambda_A - \theta_A)\cos(\Delta\lambda - \Delta\theta) \\
&\quad + \cos(\lambda_A - \theta_A)\sin(\Delta\lambda - \Delta\theta)
\end{aligned}$$

$$\begin{aligned}
\frac{h_A \sin(\theta_A + \Delta\theta)}{h \sin\theta_A} &= \cos(\Delta\lambda - \Delta\theta) + \cot(\lambda_A - \theta_A)\sin(\Delta\lambda - \Delta\theta) \\
\cot(\lambda_A - \theta_A)\sin(\Delta\lambda - \Delta\theta) &= \frac{h_A \sin(\theta_A + \Delta\theta)}{h \sin\theta_A} - \cos(\Delta\lambda - \Delta\theta) \\
\cot(\lambda_A - \theta_A) &= \frac{h_A \sin(\theta_A + \Delta\theta)}{h \sin\theta_A \sin(\Delta\lambda - \Delta\theta)} - \cot(\Delta\lambda - \Delta\theta)
\end{aligned} \tag{5.25}$$

As noted above, the final expression holds with the right-hand side (RHS) evaluated at either the second or third observation. Equating the RHS at the second and third observations and simplifying results in an expression for θ_A:

$$\begin{aligned}
&\frac{h_A \sin(\theta_A + \Delta\theta_B)}{h_B \sin\theta_A \sin(\Delta\lambda_B - \Delta\theta_B)} - \cot(\Delta\lambda_B - \Delta\theta_B) = \frac{h_A \sin(\theta_A + \Delta\theta_C)}{h_C \sin\theta_A \sin(\Delta\lambda_C - \Delta\theta_C)} \\
&\qquad - \cot(\Delta\lambda_C - \Delta\theta_C) \\
&\frac{h_A [\sin\theta_A \cos\Delta\theta_B + \cos\theta_A \sin\Delta\theta_B)]}{h_B \sin\theta_A \sin(\Delta\lambda_B - \Delta\theta_B)} - \cot(\Delta\lambda_B - \Delta\theta_B) \\
&\quad = \frac{h_A [\sin\theta_A \cos\Delta\theta_C + \cos\theta_A \sin\Delta\theta_C)]}{h_C \sin\theta_A \sin(\Delta\lambda_C - \Delta\theta_C)} - \cot(\Delta\lambda_C - \Delta\theta_C) \\
&\frac{h_A [\cos\Delta\theta_B + \cot\theta_A \sin\Delta\theta_B)]}{h_B \sin(\Delta\lambda_B - \Delta\theta_B)} - \cot(\Delta\lambda_B - \Delta\theta_B) \\
&\quad = \frac{h_A [\cos\Delta\theta_C + \cot\theta_A \sin\Delta\theta_C)]}{h_C \sin(\Delta\lambda_C - \Delta\theta_C)} - \cot(\Delta\lambda_C - \Delta\theta_C) \\
&\left[\frac{h_A \sin\Delta\theta_B}{h_B \sin(\Delta\lambda_B - \Delta\theta_B)} - \frac{h_A \sin\Delta\theta_C}{h_C \sin(\Delta\lambda_C - \Delta\theta_C)}\right]\cot\theta_A \\
&\quad = \frac{h_A \cos\Delta\theta_C}{h_C \sin(\Delta\lambda_C - \Delta\theta_C)} - \frac{h_A \cos\Delta\theta_B}{h_B \sin(\Delta\lambda_B - \Delta\theta_B)} \\
&\qquad + \cot(\Delta\lambda_B - \Delta\theta_B) - \cot(\Delta\lambda_C - \Delta\theta_C)
\end{aligned}$$

The result for $\cot\theta_A$ is the following.

$$\begin{aligned}
\cot\theta_A = &\left[\frac{h_A \cos\Delta\theta_C}{h_C \sin(\Delta\lambda_C - \Delta\theta_C)} - \frac{h_A \cos\Delta\theta_B}{h_B \sin(\Delta\lambda_B - \Delta\theta_B)}\right. \\
&\left. + \cot(\Delta\lambda_B - \Delta\theta_B) - \cot(\Delta\lambda_C - \Delta\theta_C)\right] \\
&\div \left[\frac{h_A \sin\Delta\theta_B}{h_B \sin(\Delta\lambda_B - \Delta\theta_B)} - \frac{h_A \sin\Delta\theta_C}{h_C \sin(\Delta\lambda_C - \Delta\theta_C)}\right]
\end{aligned} \tag{5.26}$$

If only we knew the factors h, we would have a solution. Placing the values from Table 5.2 into the RHS of equation (5.26) yields the value of θ_A. With θ_A

available, equation (5.25) yields the value of λ_A. Finally using both θ_A and λ_A in equation (5.24), one obtains the value of $2d$. However, we don't know a priori the values of h.

In *The Almagest*, Ptolemy, using far more geometric methods, confronts a similar problem. He then reveals a solution. Using an initial guess for the unknown factors that are akin to our values of h, Ptolemy establishes an initial estimate of the undetermined parameters. Then, using the estimate of the undetermined parameters, he updates the unknown factors. With the updated factors, Ptolemy once again refines his estimate of the undetermined parameters and continues in iterative fashion. We formalize and execute this iterative process.

Step 0. Choose initial values for h_A, h_B, and h_C.

Step 1. Place the latest value of h_A, h_B, and h_C in equations (5.26), (5.25), and (5.24), respectively, to determine estimates for θ_A, λ_A, and d.

Step 2. Using the latest values of θ_A, λ_A, and d, determine the longitude of Mars' mean and solar longitude at each observed position. Then use these values in equation (5.23), to update the current estimates of h_A, h_B, and h_C. If the updated estimates change materially from the previous estimates, return to step 1. Otherwise accept the estimates as the model parameters.

We perform one iterate of the algorithm and then present results in a table.

Step 0. Choose an initial values for h_A, h_B, and h_C.

If the value of d is small, then the values for h are all nearly one. Hoping that this is the case, we take $h = 1$ at each observation as the initial choice.

Step 1. Use the latest value of h_A, h_B, and h_C in equations (5.26), (5.25), and (5.24), respectively, to determine estimates for θ_A, λ_A, and d.
Equation (5.26) yields the estimate for θ_A:

$$\begin{aligned}
\cot\theta_A &= \left[\frac{h_A\cos\Delta\theta_C}{h_C\sin(\Delta\lambda_C-\Delta\theta_C)} - \frac{h_A\cos\Delta\theta_B}{h_B\sin(\Delta\lambda_B-\Delta\theta_B)}\right.\\
&\qquad \left. + \cot(\Delta\lambda_B-\Delta\theta_B) - \cot(\Delta\lambda_C-\Delta\theta_C)\right]\\
&\qquad \div\left[\frac{h_A\sin\Delta\theta_B}{h_B\sin(\Delta\lambda_B-\Delta\theta_B)} - \frac{h_A\sin\Delta\theta_C}{h_C\sin(\Delta\lambda_C-\Delta\theta_C)}\right]\\
\cot\theta_A &= \left[-\frac{\cos 177.2}{\sin 15.63} + \frac{\cos 81.73}{\sin 13.9} - \cot 13.9 + \cot 15.63\right]\\
&\qquad \div\left[-\frac{\sin 81.73}{\sin 13.9} + \frac{\sin 177.2}{\sin 15.63}\right]\\
&= -0.97448\\
\theta_A &= -45.740429
\end{aligned}$$

Equation (5.25) yields the estimate for λ_A:

$$\begin{aligned}
\cot(\lambda_A - \theta_A) &= \frac{h_A \sin(\theta_A + \Delta\theta_B)}{h_B \sin\theta_A \sin(\Delta\lambda_B - \Delta\theta_B)} - \cot(\Delta\lambda_B - \Delta\theta_B) \\
&= \frac{\sin 36.0}{\sin 45.7 \sin 13.9} + \cot 13.9 \\
&= 0.133313 \\
\lambda_A - \theta_A &= 7.638289 + 180 \\
\lambda_A &= 7.638289 - 45.740429 + 180 \\
&= 141.897860
\end{aligned}$$

Equation (5.24) yields the estimate for d:

$$\begin{aligned}
d &= \frac{h_A \sin(\lambda_A - \theta_A)}{2 \sin\theta_A} \\
&= -\frac{\sin 187.6}{2 \sin 45.7} \\
&= 0.092796
\end{aligned}$$

Step 2. Using the latest values of θ_A, λ_A, and d, determine the longitude of Mars' mean and solar longitude at each observed position. Then use these values in equations (5.23) to update the current estimates of h_A, h_B, and h_C. If the updated estimates change materially from the previous estimates, return to step 1. Otherwise accept the estimates as the model parameters.

We provide the calculation for h_A only. The other factors h_B, and h_C, follow in an identical manner:

$$\begin{aligned}
h_A &= \sqrt{m_1^2 + m_2^2} \\
m_1 &= \left(-d \cos\theta_A + \sqrt{1 - d^2 \sin^2\theta_A}\right) \cos\theta_A + 2d \\
&= \left(-0.09 \cos 45.7 + \sqrt{1 - 0.09^2 \sin^2 45.7}\right) \cos 45.7 + .18 \\
m_2 &= \left(-d \cos\theta_A + \sqrt{1 - d^2 \sin^2\theta_A}\right) \sin\theta_A \\
&= \left(-0.09 \cos 45.7 + \sqrt{1 - 0.09^2 \sin^2 45.7}\right) \sin 45.7 \\
h_A &= 1.070834 \\
h_B &= 1.079119 \\
h_C &= 0.946415
\end{aligned}$$

TABLE 5.3. Results

Iteration	θ_A	λ_A	d
1	−45.74042905	141.8978603	0.092796342
2	−41.94622205	145.1827914	0.099407705
3	−41.57671571	145.4795767	0.099954443
4	−41.54442592	145.5051929	0.099998729
5	−41.54174508	145.5073096	0.10000231
6	−41.54152639	145.507482	0.100002599
7	−41.54150865	145.507496	0.100002623
8	−41.54150722	145.5074971	0.100002625
9	−41.5415071	145.5074972	0.100002625
10	−41.54150709	145.5074972	0.100002625
11	−41.54150709	145.5074972	0.100002625

Using a computer, the iteration method has been programmed to generate the results of Table 5.3.

INSPIRATION

Using his more geometric method, Ptolemy passes through three iterations and finds the values $\theta_A = -41.55$, $\lambda_A = 145.5$, and $d = 0.1$. With these values, the model and the observations coincide to a remarkable degree of accuracy. The discrepancies between Ptolemy's values and the final result of Table 5.3 are immaterial and well within the tolerance of rounding error. For all intents and purposes Ptolemy, hits the mark.

The difficulties that Ptolemy had to overcome stymied Ptolemy's predecessors. Ptolemy derived his method prior to the advent of elementary algebra and Cartesian coordinates. He carried out exceedingly difficult calculations without the use of a computer, calculator, or even pencil and paper. After my own battles using a computer, I am dumbfounded and cannot imagine performing this feat under the conditions of Ptolemy's times. Not only did Ptolemy carry out his method for Mars; he also carried out similar calculations for Jupiter and Saturn. In addition, he used different but equally, if not more, complex methods for Mercury and Venus.

Let us jump briefly ahead to Kepler, whose remark concerning a similar calculation of his own is a testimony to Ptolemy. Initially, Kepler didn't buy into Ptolemy's midway compromise that fixes Mars' deferent's center at half the distance to the equant. He believed that data would show that the Ptolemaic system requires seating the equant at a different location. Kepler considered the position of the equant as another parameter and set about finding this parameter in addition to the other undetermined parameters. As he adds an additional free parameter, he requires an additional data point from an additional observation at opposition. After devising an iteration scheme inspired by, but more complicated than, Ptolemy's, Kepler calculates. More than 70 iterations later, Kepler arrives at the answer. In his book *New Astronomy*, Kepler describes the toll it has taken on him:

> If this wearisome method has filled you with loathing, it should more properly fill you with compassion for me, as I have gone through it at least seventy times at the expense of a great deal of time, and you will cease to wonder that the fifth year has gone by since I took up Mars . . . .

In performing his own calculation, Kepler found sympathy for Ptolemy, who, as noted above, carried out his calculations for all the planets. What did Ptolemy receive in return for his efforts? *The Almagest* is foundational; Ptolemy leaves a legacy that both Copernicus and Kepler require for their works. One imagines that Ptolemy believed he had crossed the divide between mathematics and theology and found perfection.

$ '
!)*+,)))- !) $ 4 5 * / 2
/"

3)*+,)))- !) $ A " ! 4 5
! ! %%6 BCC ! " C C ! CD! /"

)*+,)))- !) $ 0(0$5 0$(" C ! 5
4 5 * 3 ! 2 /"

' . (!/ &

9

Projective Algebraic Curves

One of the main functions of this book is to place algebraic curves in a natural setting (the complex projective plane) where they can be studied easily. For some readers, particularly those whose background is not in mathematics, this may prove to be a psychological barrier. I can only assure them that the reward is much greater than the mental effort involved. History has shown that placing algebraic curves in a natural setting provides a flood of illumination, enabling one much better to comprehend the features one meets in everyday applications.

Let me motivate this by looking briefly at the important question of understanding how two curves intersect each other. The simplest situation is provided by two lines. In general, you expect two distinct lines to meet in a single point. But there are lines ($x = 0$ and $x = 1$ for instance) which do not intersect. This really is a nuisance. It would be much nicer to say that these lines intersect at a single point, namely a 'point at infinity'. In this chapter, we will extend the ordinary affine plane to a 'projective plane' by adding points at infinity, and gain much as a result.

9.1 The Projective Plane

The construction is based on the following intuition. Think of the affine plane $\mathbb{K}^2$ as a plane L in $\mathbb{K}^3$ which does not pass through the origin O, and let L_0 be the parallel plane through O. Then we have a natural one-to-one correspondence between points in $\mathbb{K}^2$, and lines through the origin O in $\mathbb{K}^3$ not in L_0. To a point in L we assign the unique line through O and that point, and conversely, to a line through O not in L_0 we assign the point where it meets L. (Figure 9.1.) In this way, we conceive of lines through O in $\mathbb{K}^3$ in L_0 as in some sense 'points at infinity' in $\mathbb{K}^2$. That motivates the following formal definitions.

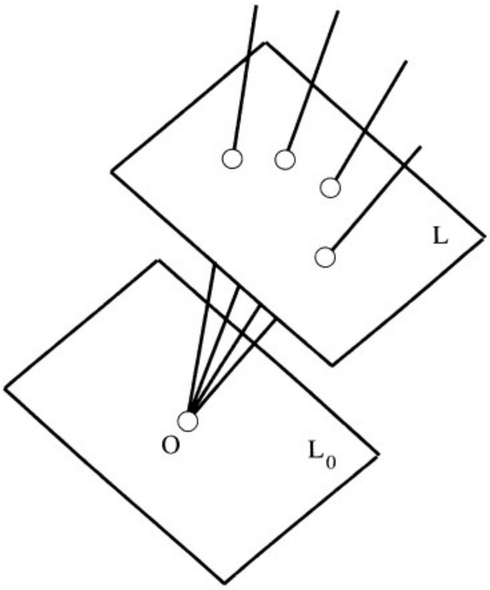

Fig. 9.1. Affine plane in 3-space

The *projective plane* $\mathbb{PK}^2$ is the set of lines through the origin O in $\mathbb{K}^3$. These lines are called the *points* of $\mathbb{PK}^2$. Given a non-zero vector $X = (x_1, x_2, x_3)$ in $\mathbb{K}^3$ there is a unique line through X, O determining a unique point in $\mathbb{PK}^2$, denoted by $(x_1 : x_2 : x_3)$. Beware, there is no point $(0 : 0 : 0)$! We call X a *representative*, and x_1, x_2, x_3 *homogeneous coordinates* for the point. Homogeneous coordinates are not unique; for any non-zero scalar λ, the vector λX represents the *same* point in $\mathbb{PK}^2$. Thus, given two points $A = (a_1 : a_2 : a_3)$, $B = (b_1 : b_2 : b_3)$ in $\mathbb{PK}^2$ the relation $A = B$ means that there exists a scalar $\lambda \neq 0$ for which $b_1 = \lambda a_1$, $b_2 = \lambda a_2$, $b_3 = \lambda a_3$. For many practical purposes we will use homogeneous coordinates labelled x, y and z.

Example 9.1 The *sphere model* of $\mathbb{PR}^2$. Each line through the origin in $\mathbb{R}^3$ meets the unit sphere $x^2 + y^2 + z^2 = 1$ in two diametrically opposite points: so $\mathbb{PR}^2$ can be thought of as a sphere, with diametrically opposite points identified. We will expand on this model of the real projective plane as we proceed.

Exercises

9.1.1 The *complex conjugate* point associated to a point $P = (a : b : c)$ in $\mathrm{P}\mathbb{C}^2$ is defined to be the point $\bar{P} = (\bar{a} : \bar{b} : \bar{c})$ where $\bar{a}$, $\bar{b}$, $\bar{c}$ are the complex conjugates of a, b, c. Show that this concept is well-defined, i.e. that the point $\bar{P}$ does not depend on the representative chosen for the point P. The point P is said to be *real* when $P = \bar{P}$. Show that P is real if and only if P has a representative with real homogeneous coordinates.

9.2 Projective Lines

Now consider any plane U through the origin in $\mathbb{K}^3$, so necessarily having an equation $ax + by + cz = 0$, where a, b, c are scalars, not all zero. Every line through the origin in U represents a point in $\mathrm{P}\mathbb{K}^2$. The resulting set of points in $\mathrm{P}\mathbb{K}^2$ is called the *line* in $\mathrm{P}\mathbb{K}^2$ with equation $ax + by + cz = 0$. Examples are the *coordinate lines* $x = 0$, $y = 0$ and $z = 0$. In the spherical model of the real projective plane, lines correspond to intersections of the unit sphere $x^2 + y^2 + z^2 = 1$ with planes through the origin, hence to great circles.

Example 9.2 A set of points in $\mathrm{P}\mathbb{K}^2$ is *collinear* when they all lie on one line. The condition for three points $P_1 = (x_1 : y_1 : z_1)$, $P_2 = (x_2 : y_2 : z_2)$, $P_3 = (x_3 : y_3 : z_3)$ to be collinear is easily obtained. It is the condition for the non-zero vectors P_1, P_2, P_3 in $\mathbb{K}^3$ to lie on a single plane through the origin, i.e. for those vectors to be linearly dependent, i.e. for the determinant

$$\begin{vmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{vmatrix} = 0.$$

Example 9.3 Through any two distinct points in $\mathrm{P}\mathbb{K}^2$ there is a unique line. That is simply the statement that through two distinct lines through the origin in $\mathbb{K}^3$ there is a unique plane through the origin. To find the equation of the line through two distinct points with representatives $P_1 = (x_1, y_1, z_1)$, $P_2 = (x_2, y_2, z_2)$ we proceed as follows. We seek all representatives $P = (x, y, z)$ for which P_1, P_2, P are coplanar; by Example 9.2 the condition is

$$\begin{vmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x & y & z \end{vmatrix} = 0.$$

For instance, if $P_1 = (0 : 0 : 1)$, $P_2 = (1 : 0 : 1)$ then the line joining P_1, P_2 is $y = 0$.

The next example describes the crucial property which distinguishes the affine from the projective plane.

Example 9.4 Any two distinct lines in $\mathbb{PK}^2$ intersect in a unique point. Indeed, two distinct lines in $\mathbb{PK}^2$ correspond to two distinct planes through the origin in $\mathbb{K}^3$, so intersect in a unique line through the origin in $\mathbb{K}^3$, i.e. in a unique point in $\mathbb{PK}^2$. It is this property which marks the fundamental difference between the affine plane and the projective plane. A practical question is to find the point of intersection of two distinct lines $ax+by+cz = 0$ and $a'x+b'y+c'z = 0$. By linear algebra, a solution of these equations is (X, Y, Z) where

$$X = \begin{vmatrix} b & c \\ b' & c' \end{vmatrix}, \quad Y = -\begin{vmatrix} a & c \\ a' & c' \end{vmatrix}, \quad Z = \begin{vmatrix} a & b \\ a' & b' \end{vmatrix}.$$

Note that at least one of X, Y, Z is $\neq 0$, else (by linear algebra) the vectors (a, b, c), (a', b', c') are linearly dependent, and the lines coincide.

At this point, it is worth pointing out that $\mathbb{PK}^2$ is a 'projective plane' in the sense of Section 2.3. The three points $(1 : 0 : 0)$, $(0 : 1 : 0)$, $(0 : 0 : 1)$ are not collinear; Example 9.3 tells us that through any two distinct points there is a unique line, and Example 9.4 tells us that any two lines intersect in at least one point.

Example 9.5 Any projective line $ax + by + cz = 0$ in $\mathbb{PK}^2$ can be 'rationally parametrized'. (A formal treatment of these ideas will be given in Chapter 18.) Recall that a line in $\mathbb{PK}^2$ corresponds to a plane through the origin in $\mathbb{K}^3$, and that any two *distinct* points A, B on the line correspond to basis vectors for that plane. It follows that any point on the plane has the form $X = sA + tB$ for some scalars s, t. Writing $A = (a_1, a_2, a_3)$, $B = (b_1, b_2, b_3)$, $X = (x, y, z)$ that means we can write $x = sa_1 + tb_1$, $y = sa_2 + tb_2$, $z = sa_3 + tb_3$ for some ratio $(s : t)$. For instance, the line $z = 0$ passes through the points $A = (1 : 0 : 0)$, $B = (0 : 1 : 0)$ yielding the parametrization $x = s$, $y = t$, $z = 0$. As in the affine case, the parametrization is not unique, it depends on the choice of A, B.

A few words are in order on the question of 'visualizing' the zero set of a projective line. Recall first the affine case. In the real case, we have the familiar lines in the $\mathbb{R}^2$; and in the complex case, we saw in Example 2.13

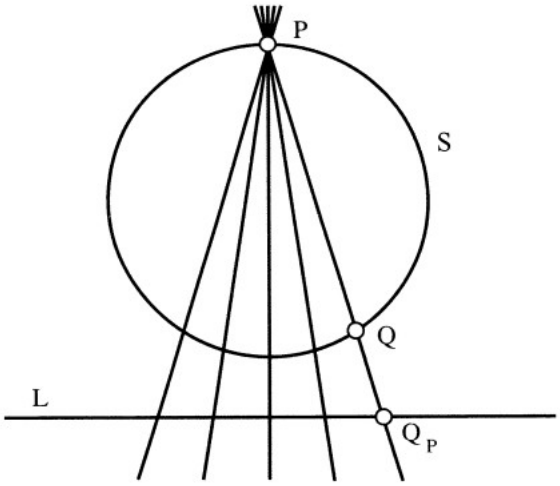

Fig. 9.2. Real projective line as a circle

that a line in $\mathbb{C}^2$ could be thought of as a real plane. Consider now the projective case, where a line is defined by a plane through the origin in $\mathbb{K}^3$. It will be no restriction to take the plane $z = 0$, so that the points of the (projective) line correspond to the (affine) lines through the origin in the (x, y)-plane, and are determined by the ratio $(x : y)$. Provided $x \neq 0$ we can identify the ratio with the quotient y/x in $\mathbb{K}$. Thus we can view the points of the projective line as the elements of $\mathbb{K}$ plus an extra element ∞. In the real and complex cases, this gives rise to simple visualizations.

Example 9.6 In the real affine plane, let S be a circle, let P be a point on S, and let L be a line not passing through P. We think of L as a copy of $\mathbb{R}$. Define *stereographic projection* to be the natural mapping $S - \{P\} \to L$, which takes a point $Q \neq P$ on S to the point on L where it meets the line through P, Q. (Figure 9.2.) Under this mapping the points of $S - \{P\}$ correspond one-to-one with the elements of $\mathbb{R}$, and we can extend the bijection to the whole of S by mapping P itself to the element ∞. Thus *we can think of a real projective line as a circle.*

Example 9.7 In real affine space, let S be a sphere, let P be a point on S, and let L be a plane not passing through P. We think of L as a copy of $\mathbb{C}$. Here too we can define *stereographic projection* to be the natural mapping $S - \{P\} \to L$, which takes a point $Q \neq P$ on S to the point on

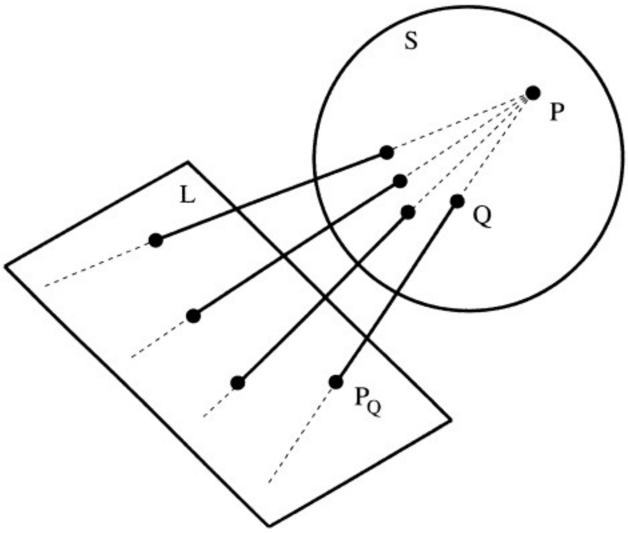

Fig. 9.3. Complex projective line as a sphere

L where it meets the line through P, Q. (Figure 9.3.) Under this mapping the points of $S - \{P\}$ correspond one-to-one with the elements of $\mathbb{C}$, and we can extend the bijection to the whole of S by mapping P itself to the element ∞. Thus *we can think of a complex projective line as a sphere.* This continues the train of thought begun in Example 2.13, namely that *'curves' in the complex projective plane can be thought of as real surfaces.*

Example 9.8 Just for the fun of it, consider the projective plane $P\mathbb{Z}_2^2$ over the field $\mathbb{Z}_2$ with two elements 0, 1. There are seven points: $(0:0:1)$, $(1:0:1)$, $(0:1:1)$, $(1:1:1)$, $(1:0:0)$, $(0:1:0)$, $(1:1:0)$. And there are seven lines: $x = 0$, $y = 0$, $z = 0$, $x = y$, $y = z$, $z = x$, $z = x + y$. Think of it as in Figure 9.4.

Exercises

9.2.1 In each of the following cases, find the equation of the line joining the points A, B in $P\mathbb{C}^2$.

(i) $A = (1:0:-1)$ $B = (1:1:-2)$
(ii) $A = (1:0:0)$ $B = (1:2:3)$
(iii) $A = (-1:1:1)$ $B = (2:3:3)$
(iv) $A = (1:i:0)$ $B = (1:-i:0)$.

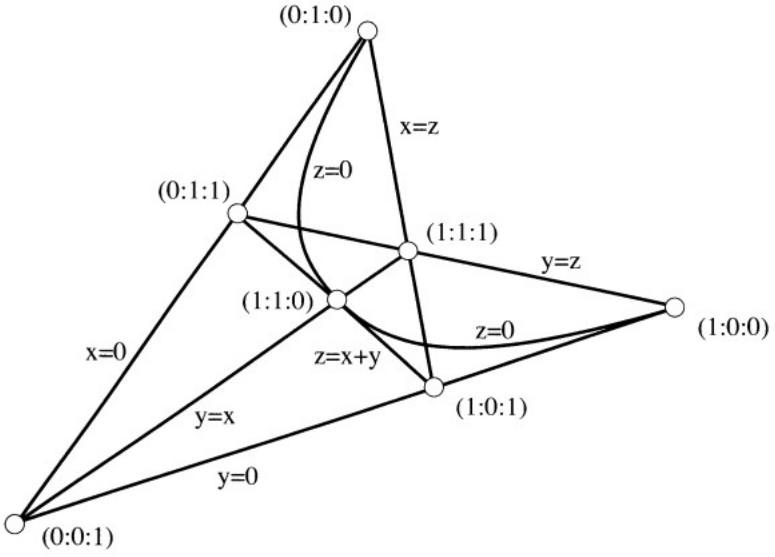

Fig. 9.4. The projective plane over $\mathbb{Z}_2$

9.2.2 In each of the following cases find the point of intersection of the two given lines in $\mathrm{P}\mathbb{C}^2$.

(i) $2x + 3y + z = 0, \quad x - y + 2z = 0$
(ii) $x + y + z = 0, \quad x + y + 2z = 0$
(iii) $x + 3y - 2z = 0, \quad y + z = 0.$

9.2.3 The *complex conjugate* line associated to a line $L = ax + by + cz$ in $\mathrm{P}\mathbb{C}^2$ is the line $\bar{L} = \bar{a}x + \bar{b}y + \bar{c}z$. Show that this concept is well-defined, i.e. that the line $\bar{L}$ does not depend on the equation of L. L is said to be *real* when $L = \bar{L}$. Show that L is real if and only if it has a defining equation with real coefficients.

9.2.4 Let P be a non-real point in $\mathrm{P}\mathbb{C}^2$. Show that the line joining P, $\bar{P}$ is real. Let L be a non-real line in $\mathrm{P}\mathbb{C}^2$. Show that the point of intersection of L, $\bar{L}$ is real.

9.2.5 Nine points P_{ij} in $\mathrm{P}\mathbb{C}^2$ are defined as follows, where ω is a primitive complex cube root of unity. (These points will arise naturally in a geometric context in Example 13.6.)

$$\begin{array}{lll} P_{00} = (0 : -1 : 1) & P_{01} = (0 : -\omega : 1) & P_{02} = (0 : -\omega^2 : 1) \\ P_{10} = (1 : 0 : -1) & P_{11} = (1 : 0 : -\omega) & P_{12} = (1 : 0 : -\omega^2) \\ P_{20} = (-1 : 1 : 0) & P_{21} = (-\omega : 1 : 0) & P_{22} = (-\omega^2 : 1 : 0). \end{array}$$

Any two of these points determine a line. Show that in total there are twelve such lines, and that each one passes through exactly three of the points P_{ij}. You might find it interesting to form a diagram of these nine points and twelve lines, and then compare it with that in Figure 2.2.

9.2.6 Show that the projective plane $\mathrm{P}\mathbb{Z}_3^2$ has precisely 13 points, and precisely 13 lines. How many points lie on a line? And how many lines pass through a point?

9.3 Affine Planes in the Projective Plane

We need to say something about the fundamental matter of how affine planes 'sit' in projective planes. To this end, we return to the basic intuition of Section 9.1. We start from an arbitrary affine plane L in $\mathbb{K}^3$, not through the origin O, and denote by L_0 the parallel plane through O. That gave us a natural one-to-one correspondence between points in L, and lines in $\mathbb{K}^3$ through O not lying in the plane L_0. In the language of the preceding section, L_0 defines a line in $\mathrm{P}\mathbb{K}^2$, the *line at infinity* associated to L, whose points are the *points at infinity*; explicitly, if L has equation $ax + by + cz = d$ then L_0 has equation $ax + by + cz = 0$. Thus we obtain a natural one-to-one correspondence between points in L and points in $\mathrm{P}\mathbb{K}^2$ not lying on the line L_0. In other words, we can think of the projective plane $\mathrm{P}\mathbb{K}^2$ as the 'affine' plane L, plus an associated 'line at infinity' L_0. That is the underlying geometric picture. The reader should bear in mind that the 'line at infinity' depends on a choice of plane L, there is nothing unique about it.

For practical purposes, we need to be more explicit about exactly how we identify L with the affine plane $\mathbb{K}^2$. The idea is that, provided L is not parallel to a given coordinate axis, we can identify L with the corresponding coordinate plane (obtained by setting the coordinate equal to zero) by projection onto that plane, in the direction of the axis. The correspondence could be written down quite explicitly, by solving the equation $ax + by + cz = d$ for one of x, y, z in terms of the other two. For instance the affine plane L defined by $x + y + z = 1$ is not parallel to the z-axis, and can be identified with the coordinate plane $z = 0$ via $(x, y) \sim (x, y, 1 - x - y)$. For most practical purposes, it will be sufficient to take L to be one of the planes $x = 1$, $y = 1$, $z = 1$, with associated lines at infinity $x = 0$, $y = 0$, $z = 0$. These choices give rise to three basic identifications:

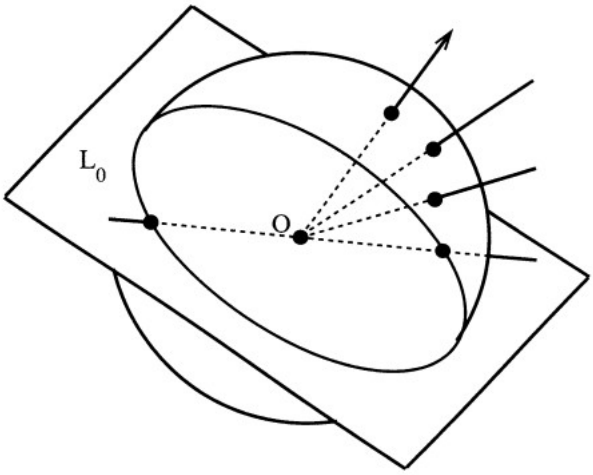

Fig. 9.5. Hemisphere model of $P\mathbb{R}^2$

choice	identification
$x = 1$	$(y, z) \sim (1 : y : z)$
$y = 1$	$(z, x) \sim (x : 1 : z)$
$z = 1$	$(x, y) \sim (x : y : 1)$

Example 9.9 The hemisphere model of $P\mathbb{R}^2$. Suppose we are given a plane L_0 through the origin O in $\mathbb{R}^3$. Each line through O meets the closed unit hemisphere $x^2 + y^2 + z^2 = 1$, $L_0 \geq 0$, with boundary a great circle, at least once. More precisely, every line through O meets the open hemisphere $x^2 + y^2 + z^2 = 1$, $L_0 > 0$ just *once*, whilst lines which meet the boundary great circle $x^2 + y^2 + z^2 = 1$, $L_0 = 0$ do so *twice*. Thus, in contrast to the sphere model, we have only to identify diametrically opposite points *on the boundary*. Perhaps this represents the best visualization of $P\mathbb{R}^2$, with the open hemisphere corresponding to $\mathbb{R}^2$, and the boundary circle to a line. Looking at the hemisphere from a distant point on the line perpendicular to L_0 through O one 'sees' the projective plane as a disc. (Figure 9.5.) Points in the interior correspond one-to-one with points in the real affine plane $\mathbb{R}^2$; the boundary circle corresponds to a line, but 'opposite' points have to be identified. It is conceptually important to realize that, whereas the sphere model provides

us with a *single* representation of the real projective plane, the hemisphere model provides a whole *family* of representations, one for each choice of the plane L_0.

9.4 Projective Curves

In this section we will work in the projective plane $\mathrm{P}\mathbb{K}^2$, with homogeneous coordinates x, y, z. By a *projective algebraic curve* of degree d in $\mathrm{P}\mathbb{K}^2$ we mean a non-zero form $F(x, y, z)$ of degree d, *up to multiplication by a non-zero scalar*. As in the affine case, we will abbreviate the term 'algebraic curve' to 'curve'. Projective curves of degree 1, 2, 3, … are called *lines*, *conics*, *cubics*, and so on. Over the real field we refer to *real* projective curves, and over the complex field to *complex* projective curves. Given a curve F we can associate to it the set

$$C(F) = \{(x, y, z) \in \mathbb{K}^3 : F(x, y, z) = 0\}.$$

This set has the crucial property of being a *cone*, meaning that if (x, y, z) is a point in $C(F)$ then so too is (tx, ty, tz), for any choice of scalar t; that is immediate from the defining property $F(tx, ty, tz) = t^d F(x, y, z)$ of forms. (Lemma 3.12.) In view of this fact, it makes sense to define the *zero set* of F to be the set $V(F)$ defined below; the key fact is that whether a point belongs to $V(F)$ does not depend on the choice of representative for that point.

$$V(F) = \{(x : y : z) \in \mathrm{P}\mathbb{K}^2 : F(x, y, z) = 0\}.$$

Recall from Section 3.4 that any form F over the field $\mathbb{K}$ admits a unique factorization $F = cF_1^{r_1} \dots F_s^{r_s}$ with $c \neq 0$ in $\mathbb{K}$, and the F_i irreducible forms such that F_j is not a factor of F_k for $j \neq k$. When F is a projective curve, we refer to $F_1, \dots, F_s$ as the *irreducible components* of F. And, as in the affine case, the zero set of F is the union of the zero sets of $F_1, \dots, F_s$.

Example 9.10 Let F be a curve in the real projective plane $\mathrm{P}\mathbb{R}^2$. We can think of the zero set in terms of the sphere model of the projective plane. Since the set $C(F)$ is a cone, the zero set is completely determined by the 'curve' of intersection with the sphere $x^2 + y^2 + z^2 = 1$. Thus we can think of the zero set as a 'curve' on the sphere, with diametrically opposite points identified. (Figure 9.6.) For instance, when F is a line the cone is a plane through the origin, intersecting the sphere in a great circle. Likewise, we can think of the zero set in terms of the hemisphere

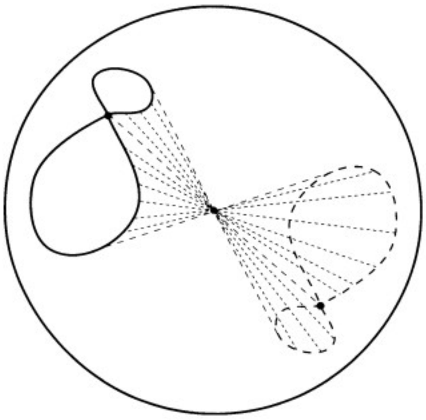

Fig. 9.6. Projective curve in the spherical model

model, where it is completely determined by the 'curve' of intersection with the hemisphere $x^2 + y^2 + z^2 = 1$, $z \geq 0$. Thus we can think of the zero set as a 'curve' on the hemisphere, with diametrically opposite points on the boundary great circle identified. For instance, when F is a line the resulting cone intersects the hemisphere in a great semicircle, and we identify the two end points.

9.5 Affine Views of Projective Curves

The object of this section is to establish the basic relations between affine and projective curves. Suppose we have a projective curve F, and make a choice of affine plane L with equation $ax + by + cz = d$. Then by eliminating one of the variables x, y, z between the relations $F(x, y, z) = 0$, $ax + by + cz = d$ we obtain an affine curve f in the (y, z)-plane, the (x, z)-plane or the (x, y)-plane. We assume here that the coefficient of the variable in the equation of L is non-zero. We refer to f as an *affine view* of F. Thus, to any projective curve F, we associate a family of affine views f, one for each choice of affine plane L and each choice of variable.

Example 9.11 Consider the real projective conic $F = x^2 + y^2 - z^2$. The affine views of F corresponding to the planes $z = 1$, $x + z = 1$ and $x + y + z = 1$ are the real ellipse $x^2 + y^2 = 1$, the parabola $y^2 = 1 - 2x$, and the hyperbola $2xy = 2x + 2y - 1$. This example makes the point

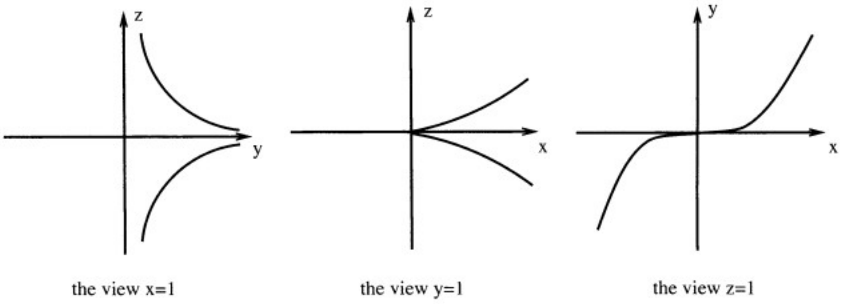

Fig. 9.7. Three affine views of a cubic curve

that different affine views of a projective curve may well fail to be affine equivalent.

Perhaps the best mental picture is to start from the spherical model of the real projective plane, with F defining a 'curve' on the surface of the unit sphere. An affine view is then determined by an affine plane L, with L_0 the parallel plane through O. The plane L_0 then determines a hemispherical model of the real projective plane, with the curve on the sphere giving a curve on the open hemisphere, and opposite points of intersection with the great circle $L_0 = 0$ identified. For instance, when F is a line, the intersection with the sphere is a great circle, intersecting the hemisphere in 'half' a great circle, with its end points identified. As L_0 changes, so the view of the 'curve' will change. In practice, we will work with the three *principal* affine views, meaning those associated to the affine planes $x = 1$, $y = 1$, $z = 1$. (Note that *any* point in the projective plane has a representative with either $x = 1$, $y = 1$ or $z = 1$ so these views cover the whole plane.) The process of obtaining the equation for f by substituting $x = 1$, $y = 1$ or $z = 1$ in $F(x, y, z)$ is known as *dehomogenization*, since it destroys the homogeneity of F.

Example 9.12 Consider, for instance, the projective cubic $F = yz^2 - x^3$ in $P\mathbb{R}^2$. Dehomogenizing with respect to x, y, z we obtain the three affine views of Figure 9.7, namely $f = y - x^3$ in the (x, y)-plane, $g = yz^2 - 1$ in the (y, z)-plane, and $h = z^2 - x^3$ in the (z, x)-plane. No two of these affine views are affine equivalent.

It is illuminating to think of Example 9.12 in terms of hemispherical models. If we take the hemispheres resulting from the planes $x = 0$,

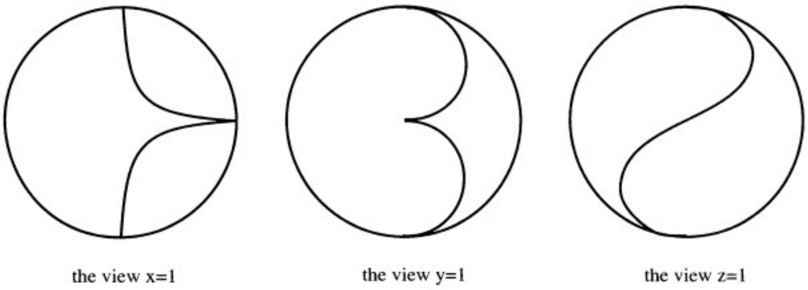

Fig. 9.8. Three hemispherical views of a cubic curve

$y = 0$, $z = 0$ in $\mathbb{R}^3$, we obtain the three pictures in Figure 9.8 which make it more plausible that they are different 'views' of the same curve F.

That brings us to one of the key constructions in this book, namely that of reversing the above process. More precisely, given any affine curve $f(x, y)$ of degree d in $\mathbb{K}^2$, and any affine plane L in $\mathbb{K}^3$ not through the origin, and not parallel to the z-axis, there exists a projective curve $F(x, y, z)$ of degree d in $P\mathbb{K}^2$ with the property that $F(x, y, z) = f(x, y)$ for every point (x, y, z) in L. Indeed, if L is defined by an equation $l = 1$, where $l = ax + by + cz$ with $c \neq 0$, and $f(x, y) = \sum a_{ij}x^i y^j$, we have only to observe that we can take

$$F(x, y, z) = l^d f\left(\frac{x}{l}, \frac{y}{l}\right) = \sum a_{ij}x^i y^j l^{d-i-j}.$$

Note that the condition $c \neq 0$ ensures that F has a term of degree d with non-zero coefficient. This is the process of *homogenization* with respect to the form l, you homogenize f by multiplying each monomial by that power of l necessary to bring the degree up to d. In practice, the only case of this construction we will meet is when L is the plane $z = 1$, so

$$F(x, y, z) = z^d f\left(\frac{x}{z}, \frac{y}{z}\right) = \sum a_{ij}x^i y^j z^{d-i-j}.$$

For instance, to the affine line $f = ax + by + c$ we associate the projective line $F = ax + by + cz$. Note that f itself is one of the family of affine views provided by F; indeed, if we homogenize f to obtain F, and then dehomogenize using z again, we are back to f. However, beware! That can fail when the order of the processes is reversed. For instance, if we dehomogenize the projective curve $F = z^2(x^2 + z^2)$, we obtain the affine curve $f = x^2 + 1$, and then homogenization with respect to z yields a *different* projective curve $G = x^2 + z^2$. Of course the process of homogenization applies equally well when the variables in the affine

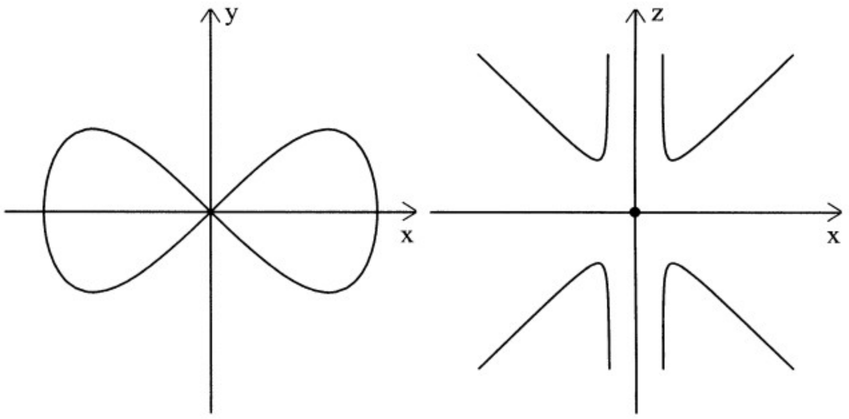

Fig. 9.9. Two affine views of the eight-curve

plane are labelled as x, z (when we homogenize with respect to y) or as y, z (when we homogenize with respect to x).

Example 9.13 Consider the cubic $f = y - x^3$ in $\mathbb{PR}^2$. The associated projective curve is $F = yz^2 - x^3$. But, as we saw in Example 9.12, the zero sets of the principal affine views $f = y - x^3$, $g = z^2 - x^3$ and $h = yz^2 - 1$ are visually quite different. At first sight it is difficult to imagine that the views g, h have any connexion with the original curve f. In particular, the view g shows that there is a 'cusp' at infinity on f, something that one might not have expected.

This example makes a very fundamental conceptual point. An affine curve f is simply one out of a very large family of affine views of the associated projective curve F. By working solely in the affine case, we deny ourselves the rich geometry provided by the projective situation. It is virtually impossible to get a genuine understanding of the geometry of an affine curve without moving into the projective situation. Here is another illuminating illustration.

Example 9.14 The projective curve in $\mathbb{PR}^2$ associated to the eight-curve $f = y^2 - x^2 + x^4$ is $F = y^2z^2 - x^2z^2 + x^4$. This meets the line at infinity $z = 0$ when $x^4 = 0$, yielding the point at infinity $Y = (0 : 1 : 0)$ on the y-axis. The affine view given by $y = 1$ is $g = z^2 - x^2z^2 + x^4$, and in that view Y corresponds to the origin in the (z, x)-plane. The zero set of g is easily sketched, by writing $g = 0$ in the form $z^2 = x^4/(x^2-1)$, and reveals that the origin is an isolated point on the zero set of g. (Figure 9.9.) Thus Y is an 'isolated' point on the projective curve – another unexpected revelation.

There is a very simple relationship between the zero sets of an affine curve f and its homogenization F with respect to z. Since $F(x, y, 1) = f(x, y)$ we have $f(x, y) = 0$ if and only if $F(x, y, 1) = 0$. Identifying the point (x, y) in $\mathbb{K}^2$ with the point $(x : y : 1)$ in $P\mathbb{K}^2$ (as above), we see that the zero set of F can be thought of as the zero set of f, plus the points at infinity $(x : y : 0)$ satisfying $F(x, y, 0) = 0$. Note that $F(x, y, 0)$ is a binary form in x, y, so finding the points at infinity on the curve is tantamount to finding the linear factors of that form.

Example 9.15 Any projective curve F in $P\mathbb{C}^2$ has an infinite zero set. Indeed the zero set of F 'contains' the zero set of any affine curve f in $\mathbb{C}^2$ obtained by dehomogenization, so is infinite by Lemma 2.1.

Example 9.16 To any line $ax + by + c$ in $\mathbb{K}^2$ we can associate a line $ax + by + cz$ in $P\mathbb{K}^2$. The zero set of the projective line comprises that of the affine line together with a unique point at infinity $(x : y : 0)$ for which $ax + by = 0$, i.e. the point $(-b : a : 0)$. Note that the point at infinity on the line does not depend on the coefficient c. All lines parallel to $ax+by+c = 0$ in $\mathbb{K}^2$ have the same point at infinity. In particular, the point at infinity on the x-axis is $X = (1 : 0 : 0)$ and the point at infinity on the y-axis is $Y = (0 : 1 : 0)$.

Example 9.17 The parabola $f = y^2 - x$ in $\mathbb{R}^2$ gives rise to the conic $F = y^2-xz$ in $P\mathbb{R}^2$ which meets the line at infinity $z = 0$ when $F(x, y, 0) = y^2 = 0$, i.e. at the single point $X = (1 : 0 : 0)$.

Example 9.18 The Maltese cross (Figure 6.6) is the quartic in $\mathbb{R}^2$ defined by $f = xy(x^2 - y^2) - (x^2 + y^2)$. The associated projective curve in $P\mathbb{R}^2$ is $F = xy(x^2 - y^2) - z^2(x^2 + y^2)$, which meets the line at infinity $z = 0$ when $xy(x^2 - y^2) = 0$. Thus we obtain the points at infinity on the lines $x = 0$, $y = 0$, $y = x$ and $y = -x$, namely $(0 : 1 : 0)$, $(1 : 0 : 0)$, $(1 : 1 : 0)$ and $(1 : -1 : 0)$.

Example 9.19 The general circle in $\mathbb{R}^2$ with centre (a, b) and radius $r > 0$ is $f = (x - a)^2 + (y - b)^2 - r^2$. The associated projective curve is $F = (x-az)^2+(y-bz)^2-r^2z^2$ in $P\mathbb{R}^2$. This meets the line at infinity $z = 0$

when $x^2+y^2=0$, giving no points in $\mathrm{P}\mathbb{R}^2$. However, we can also think of F as a projective curve in $\mathrm{P}\mathbb{C}^2$, in that case the relation $x^2+y^2=0$ yields $y=\pm ix$, so there are two points at infinity in $\mathrm{P}\mathbb{C}^2$, namely the complex conjugate points $I=(1:i:0)$ and $J=(1:-i:0)$.

The points I, J in Example 9.19 play a special role in the history of curve theory; they are called the *circular points at infinity*, since any circle (thought of as a complex projective curve) passes through them. The notation has its origins in the biblical names Isaak and Jakob. Real algebraic curves for which the associated complex projective curve passes through I, J are said to be *circular*, and play an important role in mechanical engineering problems. We have already met a number of circular curves, for instance, the pedal curves of the parabola and the unit circle, the three- and four-leaved clovers, and the astroid.

Exercises

9.5.1 Find the points at infinity on the projective curves in $\mathrm{P}\mathbb{R}^2$ corresponding to the following curves f in $\mathbb{R}^2$.

(i) $f=x^3-xy^2-y$
(ii) $f=(y-x^2)^2-xy^3$
(iii) $f=x^4+4xy-y^4$
(iv) $f=x^5+y^5-2xy^2$
(v) $f=x^2y^2+x^2-y^2$.

9.5.2 Find the points at infinity on the projective curves in $\mathrm{P}\mathbb{R}^2$ corresponding to the following curves f in $\mathbb{R}^2$.

(i) $f=x^5+y^5-2x^2-5xy+2y^2$
(ii) $f=x^5+y^5-5x^2y^2+3x-3y$
(iii) $f=(2x+y)(x+y)^2+(2x-y)(x+y)+(x-y)$
(iv) $f=x^4y+x^2y^3+x^4-y^3+1$.

9.5.3 Let $f=cf_1^{r_1}\dots f_s^{r_s}$ be the unique factorization of the polynomial f into irreducible factors, and let $F, F_1, \dots, F_s$ be the forms obtained from $f, f_1, \dots, f_s$ by homogenization. Show that $F=cF_1^{r_1}\dots F_s^{r_s}$ is the unique factorization of the form F into irreducible forms.

9.5.4 Let h be any irreducible conic in $\mathbb{R}^2$ having at least one real point and let H be the corresponding projective conic in $\mathrm{P}\mathbb{C}^2$. Show that if H passes through the circular points at infinity I, J then

h is a circle. (Point out *precisely* where you use the hypothesis that h is irreducible, and where you use the hypothesis that h has a real point.)

9.5.5 Let $f(x, y)$ be a curve of degree d in $\mathbb{R}^2$, and let $F(x, y, z)$ be the corresponding projective curve in $P\mathbb{C}^2$. Show that $f(x, y)$ is circular if and only if its HOT are divisible by $x^2 + y^2$. Use this characterization to verify that the pedal curves of the parabola and the unit circle, the three- and four-leaved clovers, and the astroid are all circular curves.

9.5.6 Let F, G be two curves in $P\mathbb{R}^2$ intersecting in finitely many points. Show that the intersections are either real, or occur in complex conjugate pairs.

10

Singularities of Projective Curves

The next step in our development is to extend the concept of 'singularity' from the affine to the projective situation. We will follow the same pattern as in the affine case. The starting point is to extend the idea of intersection number (of a line with a curve at some point) to the projective situation; we can then define the 'multiplicity' of a point on a projective curve, and use this to introduce 'singular' points. As an application of these ideas, we return to conics, and show that by thinking of conics in the projective (rather than the affine) plane we can gain geometric insight into the delta invariants of Chapter 5.

10.1 Intersection Numbers

For projective curves we can mimic what we did in Chapter 6 for affine curves. Let $F(x, y, z)$ be a curve of degree d in $\mathbb{PK}^2$. And let L be a line. L is determined by any two distinct points, say $A = (a_1 : a_2 : a_3)$ and $B = (b_1 : b_2 : b_3)$: then, as in Example 9.5, the line L can be parametrized as $x_1 = sa_1 + tb_1$, $x_2 = sa_2 + tb_2$, $x_3 = sa_3 + tb_3$. The intersections of F with L are given by $\Phi(s, t) = 0$, where Φ is defined by

$$\Phi(s,t) = F(sa_1 + tb_1, sa_2 + tb_2, sa_3 + tb_3).$$

Since $F(x, y, z)$ is a form of degree d we have $\Phi(\lambda s, \lambda t) = \lambda^d \Phi(s, t)$ for any scalar λ, so $\Phi(s, t)$ is a binary form of degree d, the *intersection form*. Of course, Φ may be identically zero. It is worth comparing this with the situation in the affine case where an intersection polynomial $\phi(t)$, associated to an affine curve f of degree d and a line l, is only known to be of degree $\leq d$. Concerning the generalities of how lines intersect projective curves, we have the following analogue of Lemma 4.4.

Lemma 10.1 *Let F be a curve of degree d in $\mathrm{P}\mathbb{K}^2$, and let L be a line. Then either L meets F in $\leq d$ points, or L is a component of F. Thus, for a curve F with no line components, the degree d is an upper bound for the number of intersections of F with any line L.*

Proof As above, we can associate an intersection form Φ of degree d to the (parametrized) line L and the curve F. Provided Φ is not identically zero, the equation $\Phi(s,t) = 0$ has $\leq d$ roots, and there are $\leq d$ points of intersection. Suppose now that Φ is identically zero, so F vanishes at every point on L. Let L have equation $ax + by + cz = 0$. It is no restriction to suppose that $c \neq 0$, so $X = x$, $Y = y$, $Z = ax + by + cz$ is an invertible linear mapping. Then $F(X, Y, Z)$ vanishes at every point of $Z = 0$, so the binary form $F(X, Y, 0)$ is identically zero, and Z must be a factor of $F(X, Y, Z)$. In terms of the variables x, y, z that means that L is a factor of $F(x, y, z)$, as required. □

Next, let us look in more detail at the way in which a line L intersects a projective curve F of degree d. Assume L is not a component of F. Let P be a point on L corresponding to the ratio $(s_0 : t_0)$, in some parametrization of L. We define the *intersection number* $I(P, F, L)$ to be the multiplicity of $(s_0 : t_0)$ as a root of the equation $\Phi(s,t) = 0$; that means that $\Phi(s,t)$ has a factor $(t_0 s - s_0 t)^m$ for some integer $m \geq 0$, and that $I(P, F, L) = m$. Clearly $I(P, F, L)$ is an integer ≥ 0, and vanishes if and only if P is not an intersection of F, L. It is a convention that $I(P, F, L) = \infty$ when L is a component of F.

Example 10.1 The projective curve corresponding to the eight-curve $f = y^2 - x^2 + x^4$ in $\mathbb{R}^2$ is $F = y^2z^2 - x^2z^2 + x^4$ in $\mathrm{P}\mathbb{R}^2$. Let L be the line $x = 0$. Then L is the line through $P = (0 : 0 : 1)$, $Q = (0 : 1 : 0)$ and parametrized as $x = 0$, $y = t$, $z = s$. The resulting intersection form is $\Phi(s,t) = s^2t^2$, having roots $(1 : 0)$ and $(0 : 1)$ both of multiplicity 2; the root $(1 : 0)$ corresponds to P, whilst $(0 : 1)$ corresponds to Q. Since both roots have multiplicity 2 the intersection numbers are $I(P, F, L) = 2$, $I(Q, F, L) = 2$.

The next step in our development is to establish that $I(P, F, L)$ does not depend on the chosen parametrization of the line L. We follow the paradigm of the affine case, by establishing a preliminary lemma. By a projective *change of parameter* we mean an invertible linear map $U(s,t) = (u, v)$ where $u = \alpha s + \gamma t$, $v = \beta s + \delta t$ and α, β, γ, δ are scalars with $\alpha\delta - \beta\gamma \neq 0$. We now have

Lemma 10.2 *Suppose that two forms* $\Phi(s,t)$, $\Psi(u,v)$ *are related by a projective change of parameter, i.e.* $\Phi(s,t) = \Psi(U(s,t))$ *for some projective change of parameter* $U(s,t)$. *Then* (s_0, t_0) *is a zero of multiplicity* m *of* $\Phi(s,t)$ *if and only if* $U(s_0,t_0)$ *is a zero of multiplicity* m *of* $\Psi(u,v)$.

Proof Write $U(s_0,t_0) = (u_0, v_0)$. Observe (using the Factor Theorem) that $-t_0 s + s_0 t$ is a factor of $\Phi(t)$ if and only if (s_0, t_0) is a zero of $\Phi(t)$, if and only if $U(s_0,t_0)$ is a zero of $\Psi(U(s,t))$, if and only if $-v_0 u + u_0 v$ is a factor of $\Psi(u,v)$. However, by Lemma 4.6 corresponding factors $t - t_0$, $u - u(t_0)$ have the same multiplicity, as required. □

Lemma 10.3 $I(P,F,L)$ *depends only on* P, F, L, *it does not depend on the choice of parametrization of* L.

Proof This follows exactly the same lines as the corresponding result (Lemma 6.3) in the affine case. Consider two parametrizations of L, the first via distinct points A, B with parameter the ratio $(s:t)$, yielding the intersection form $\Phi(s,t) = F(sA + tB)$, and the second via distinct points C, D with parameter the ratio $(u:v)$, yielding the intersection form $\Psi(u,v) = F(uC + vD)$. The connexion between $\Phi(s,t)$, $\Psi(u,v)$ is easily established. Let $A = \alpha C + \beta D$, $B = \gamma C + \delta D$. Then, by direct calculation

$$\Phi(s,t) = F(sA + tB) = F(u(s,t)C + v(s,t)D) = \Psi(u(s,t), v(s,t))$$

where $u(s,t) = \alpha s + \gamma t$, $v(s,t) = \beta s + \delta t$. Thus $\Phi(s,t)$ is obtained from $\Psi(u,v)$ by the projective change of parameter $U(s,t) = (u(s,t), v(s,t))$. (Note that $\alpha\delta - \beta\gamma \neq 0$ else A, B coincide.) Then $(s_0 : t_0)$ is a zero of $\Phi(s,t)$ if and only if $U(s_0,t_0)$ is a zero of $\Psi(u,v)$, and it follows from Lemma 10.2 that the multiplicities of the zeros coincide. The result follows from the definition of $I(P,F,L)$. □

The *total number* of intersections of F with L (counted properly) is the sum of the intersection numbers $I(P,F,L)$ with P on F. As the sum of the multiplicities of the roots of $\Phi(s,t)$ is $\leq d$, we deduce that the total number of intersections of a curve F of degree d with a line L is $\leq d$. Over the complex field the intersection form Φ has exactly d zeros, counted with multiplicities, so the total number of intersections (counted properly) is *precisely* d, in particular Φ has at least one zero. Thus we have established

Lemma 10.4 *Let F be a curve of degree d in* $\mathbb{PC}^2$, *and let L be a line which is not a component of F. Then L meets F in exactly d points, counted properly. In particular, L meets F in at least one point.*

Lemma 10.4 is a special case of a celebrated result in the theory of algebraic curves, namely Bézout's Theorem. Roughly speaking, this says the following. Any two curves F, G in $\mathbb{PC}^2$ with no common component intersect in only finitely many points P. Moreover, it is possible to associate an 'intersection number' $I(P, F, G)$ to any such curves F, G, and any point P, in such a way that the sum of the intersection numbers at all points of intersection is de, where $d = \deg F$, $e = \deg G$. And in the special case when G is a line L, this 'intersection number' agrees with the number $I(P, F, L)$ defined above. The proof of Bézout's Theorem is presented in Chapter 14, together with the algebraic ideas on which it depends.

It is worth remarking at this stage that the definition of $I(P, F, L)$ has a fatal limitation. We used the fact that L can be parametrized globally by parameters s, t. In principle, one could extend the idea by replacing L by any curve G which can be globally parametrized. However, this approach founders on the fundamental fact that rather few algebraic curves have this property. (We will have more to say about this matter in the final chapter.) In practice you can reduce the computation of intersection numbers in the projective plane to one in an affine plane, via the following result.

Lemma 10.5 *Let P be an intersection point of a curve F of degree d with a line L in the projective plane* $\mathbb{PK}^2$. *Dehomogenize with respect to one of the variables x, y, z to obtain a point p, a curve f, and a line l in the affine plane* $\mathbb{K}^2$. *Then $I(P, F, L) = I(p, f, l)$.*

Proof At least one of the homogeneous coordinates of P is non-zero. Without loss of generality, assume it is z. Let $B = (b_1 : b_2 : 0)$ be the unique point where $z = 0$ meets L, and let $A = (a_1 : a_2 : 1)$ be any point on L distinct from B. Then the projective line L is parametrized as $sA + tB$, with $s = 0$ corresponding to B, and gives rise to an intersection form $\Phi(s, t) = F(sA + tB)$. Likewise, the affine line l is parametrized as $a + tb$, where $a = (a_1, a_2)$, $b = (b_1, b_2)$, and gives rise to an intersection polynomial $\phi(t) = f(a + tb)$. Moreover, under the identification $(x, y) \sim (x : y : 1)$ the point on l with parameter t_0 corresponds to the point on

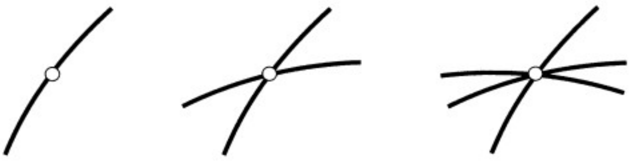

Fig. 10.1. Simple, double and triple points

L with parameter $s = 1$, $t = t_0$. For $s \neq 0$ we have

$$\Phi(s,t) = F(sA + tB) = s^d f(a + ub) = s^d \phi(u)$$

where $u = t/s$. Thus the zeros (s_0, t_0) of $\Phi(s,t)$ with $s_0 \neq 0$ correspond to zeros t_0 of $\phi(t)$, and have the same multiplicity. The result follows from the definitions of $I(P,F,L)$, $I(p,f,l)$. □

Example 10.2 Consider the curve $F = y^2z^2 - x^2z^2 + x^4$ in $\mathbb{PR}^2$, the line L given by $x = 0$, and the point $P = (0:0:1)$. Dehomogenizing with respect to z, we obtain the affine curve $f = y^2 - x^2 + x^4$, the line l given by $x = 0$, and the point $p = (0,0)$. We can parametrize l as $x = 0$, $y = t$ with p corresponding to the value $t = 0$ of the parameter. Then the intersection polynomial is $\phi(t) = t^2$, and $I(p,f,l) = 2$, agreeing with the result of Example 10.1.

10.2 Multiplicity of a Point on a Curve

Let P be a point on a curve F in $\mathbb{PK}^2$. Consider the pencil of all lines L through P. By the *multiplicity* of P on F we mean the minimal value m of the intersection number $I(P,F,L)$ as L varies over the pencil, i.e. $I(P,F,L) \geq m$ for *every* line L through P, and there exist lines L for which $I(P,F,L) = m$. Points of multiplicity 1, 2, 3, 4, … are said to be *simple*, *double*, *triple*, *quadruple*, … points of F. A crude mental picture for a point P of multiplicity m is that there are m 'branches' of the curve self-intersecting at P. (Figure 10.1.) But, as in the affine case, that is not the whole story.

In view of Lemma 10.5, we have the following result, which allows us to compute multiplicities via the technique of Lemma 6.4.

Lemma 10.6 *Let P be a point on a curve F in $\mathbb{PK}^2$. Dehomogenize with* *respect to one of the variables x, y, z to obtain a point p on a curve f in $\mathbb{K}^2$. Then the multiplicity of P on F coincides with that of p on f.*

Example 10.3 The eight-curve gives rise to the projective curve $F = y^2z^2 - x^2(z^2 - x^2)$ in $P\mathbb{R}^2$, intersecting the line at infinity at the unique point $Q = (0 : 1 : 0)$. To find the multiplicity of Q we dehomogenize with respect to y to obtain the affine curve $f = z^2 - x^2(z^2 - x^2)$ and the point $q = (0, 0)$ in the (x, z)-plane. Here the LOT are z^2, so q has multiplicity 2 on f, and hence Q has multiplicity 2 on F.

Exercises

10.2.1 Let F be an irreducible quartic curve in $P\mathbb{C}^2$. Show that F cannot have a point P of multiplicity ≥ 4. Show also that if F has a triple point P then P is the only point on F of multiplicity ≥ 2.

10.2.2 Let P be a point on a projective curve F in $P\mathbb{K}^2$ of multiplicity $\geq m$. Show that P is a point on the projective curve F_z of multiplicity $\geq (m - 1)$. And show, by means of an example, that P could have multiplicity $> m - 1$.

10.3 Singular Points

A point P on a projective curve F is *singular* when it has multiplicity ≥ 2. The curve F itself is *singular* when it has at least one singular point, otherwise it is *non-singular*. We will see in Chapter 18 that an irreducible curve F has at most finitely many singular points.

Example 10.4 A cubic curve F in $P\mathbb{K}^2$ with two singular points P, Q necessarily reduces to a conic, and the line L joining P, Q. Indeed, since P, Q are singular, we have $I(P, F, L) \geq 2$, $I(Q, F, L) \geq 2$ so the total number of intersections of L with F is ≥ 4, contradicting Lemma 10.1 unless L is a component of F.

Here is a simple criterion for a point to be singular on a curve F, analogous to Lemma 6.6 in the affine case.

Lemma 10.7 *A point P is a singular point on a curve F if and only if $F(P) = 0$ and $F_x(P) = F_y(P) = F_z(P) = 0$.*

Proof The condition for $P = (p_1 : p_2 : p_3)$ to be singular is that $I(P, F, L) \geq 2$ for *every* line L through P. We need to interpret this in detail. Let L be any line through P, and let $Q = (q_1 : q_2 : q_3)$ be another point on L. Recall that the intersection form $\Phi(s,t) = F(sP + tQ)$ is a form of degree d in s, t so has the form

$$\Phi(s,t) = s^d F_0(P,Q) + s^{d-1}tF_1(P,Q) + \cdots + t^d F_d(P,Q)$$

where each $F_k(P,Q)$ is a polynomial in the entries of P, Q. The point P corresponds to the ratio $(1 : 0)$ and the condition $I(P,F,L) \geq 2$ is that $(1 : 0)$ should have multiplicity ≥ 2 as a root of $\Phi(s,t) = 0$, i.e. that $\Phi(s,t)$ has t^2 as a factor, i.e. that $F_0(P,Q) = 0$, $F_1(P,Q) = 0$. The condition $F_0(P,Q) = 0$ holds automatically, as P lies on F, so it remains to interpret the condition $F_1(P,Q) = 0$. Write

$$\phi(t) = F(P + tQ) = tF_1(P,Q) + \cdots + t^d F_d(P,Q)$$

so $F_1(P,Q) = \phi'(0)$. The Chain Rule gives $\phi'(t) = q_1F_x + q_2F_y + q_3F_z$, where the partials are evaluated at $P + tQ$; setting $t = 0$ yields

$$F_1(P,Q) = q_1F_x(P) + q_2F_y(P) + q_3F_z(P).$$

This expression is zero for all choices of q_1, q_2, q_3 if and only if $F_x(P) = F_y(P) = F_z(P) = 0$, as required. □

Example 10.5 A point R of intersection of two projective curves F, G is necessarily a singular point of $H = FG$. The point R satisfies the relations $F(R) = 0$, $G(R) = 0$ and hence $H(R) = 0$. Moreover, the rule for differentiating a product gives

$$H_x = F_xG + FG_x, \quad H_y = F_yG + FG_y, \quad H_z = F_zG + FG_z$$

so R also satisfies the relations $H_x(R) = 0$, $H_y(R) = 0$, $H_z(R) = 0$ and is therefore singular on H.

The labour involved in finding singular points in practice can be reduced via the following useful little result about forms, known as the Euler Lemma.

Lemma 10.8 *Let $F(x_1,\ldots,x_n)$ be a form of degree d over a field $\mathbb{K}$, then we have the Euler relation*

$$dF = \sum x_kF_{x_k}.$$

Proof Recall first (Lemma 3.12) that a form $F(x_1,\ldots,x_n)$ of degree d satisfies the following identity, for all $t \in \mathbb{K}$

$$F(tx_1,\ldots,tx_n) = t^dF(x_1,\ldots,x_n).$$

Differentiating both sides of this relation with respect to t, using the Chain Rule, we get the following relation. The required result then follows on setting $t = 1$.

$$\sum x_k F_{x_k}(tx_1, \ldots, tx_n) = dt^{d-1} F(tx_1, \ldots, tx_n).$$

□

The point of the Euler Lemma is that if all the partials F_x, F_y, F_z of a projective curve F vanish at a point P then *automatically* F itself vanishes at P. Thus *the singular points of a projective curve F are obtained by solving the equations*

$$0 = F_x = F_y = F_z.$$

Many of the classical affine curves turn out to have singularities 'at infinity', helping to explain their geometry.

Example 10.6 Agnesi's versiera (Example 2.8) is the real affine cubic $f = x^2y - 4a^2(2a - y)$ with $a > 0$. Although it has no singularities in $\mathbb{R}^2$, the corresponding projective curve $F = x^2y - 4a^2z^2(2az - yz)$ does have a singularity at infinity. Singularities are given by $0 = F_x = 2xy$, $0 = F_y = x^2 + 4a^2z^2$, $0 = F_z = -8a^2z(3az - y)$. The first relation gives $x = 0$ or $y = 0$; $x = 0$ leads to the singularity $P = (0 : 1 : 0)$, the point at infinity on the y-axis, whilst $y = 0$ produces only the trivial solution. The nature of the singularity at P is revealed by setting $y = 1$ in F, yielding the affine curve

$$g = x^2 - 4a^2z^2(2az - 1) = x^2 + 4a^2z^2 + \text{HOT}$$

having a singularity at the origin in the (x, z)-plane, namely an acnode. In particular, this explains why all the 'vertical' lines in the (x, y)-plane only meet the curve once – they meet the curve twice more at the point P.

Example 10.7 The eight-curve gives rise to the projective curve $F = y^2z^2 - x^2(z^2 - x^2)$ in $\mathbb{PR}^2$. Singular points require $0 = F_x = 2x(2x^2 - z^2)$, $0 = F_y = 2yz^2$, $0 = F_z = 2z(y^2 - x^2)$. Clearly, $y = 0$ or $z = 0$: $y = 0$ gives $P = (0 : 0 : 1)$ and $z = 0$ gives $Q = (0 : 1 : 0)$. This throws light on Example 10.1; there we saw that the line $x = 0$ meets the curve solely at P, Q with intersection number 2 in each case. Since Q is a singular point of the curve, that is no longer too surprising.

Example 10.8 Consider the family of *Steiner* cubics in $\mathbb{PC}^2$ given by the formulas

$$F = \mu(x^3 + y^3 + z^3) + 3\lambda xyz.$$

We claim that F is singular if and only if $\mu = 0$ or $\lambda^3 = -1$, and that in each of these four cases F reduces to a triangle. (By a *triangle* we mean a projective cubic which reduces to three non-concurrent lines, known as the *sides*; the three points of intersection of the sides are the *vertices* of the triangle.) When $\mu = 0$, the cubic is $F = xyz$, a triangle with singular points at the vertices $(1 : 0 : 0)$, $(0 : 1 : 0)$, $(0 : 0 : 1)$. When $\mu \neq 0$ we can assume that $\mu = 1$, by multiplying the equation through by a scalar. Then singular points are given by $F_x = 0$, $F_y = 0$, $F_z = 0$, i.e. by

$$x^2 = -\lambda yz, \qquad y^2 = -\lambda zx, \qquad z^2 = -\lambda xy.$$

Note that if one of x, y, z is $= 0$ then all are $= 0$, so we can assume x, y, z are all $\neq 0$. Equating the product of the three left-hand sides, with the product of the three right-hand sides, we get $(1 + \lambda^3)x^2y^2z^2 = 0$. When $\lambda^3 \neq -1$, one of x, y, z must be $= 0$, a contradiction, so in that case F is non-singular. Suppose $\lambda^3 = -1$. Multiplying the three relations by x, y, z respectively yields $x^3 = y^3 = z^3 = -\lambda xyz$. Let 1, ω, ω^2 be the three complex cube roots of 1. Then λ is one of -1, $-\omega$, $-\omega^2$; moreover it is no restriction to suppose that $x = 1$ so y, z must each be one of 1, ω, ω^2. That yields the following singular points.

$$\begin{cases} \lambda = -1 & (1 : 1 : 1) \quad (1 : \omega : \omega^2) \quad (1 : \omega^2 : \omega) \\ \lambda = -\omega & (1 : 1 : \omega^2) \quad (1 : \omega : \omega) \quad (1 : \omega^2 : 1) \\ \lambda = -\omega^2 & (1 : 1 : \omega) \quad (1 : \omega : 1) \quad (1 : \omega^2 : \omega^2). \end{cases}$$

In each case, we have three non-collinear points; the lines joining any two of these points meet F in ≥ 4 points, so must be components, and the cubic reduces to a triangle, with vertices the given singular points.

Exercises

10.3.1 Find the singular points of the quartic curve in $\mathbb{PC}^2$ given by $F = x^2y^2 + y^2z^2 + z^2x^2$, and determine their multiplicities.

10.3.2 Find the singular points of the quartic curve in $\mathbb{PC}^2$ given by $F = (x^2 - z^2)^2 - y^2z(2y + 3z)$, and determine their multiplicities.

10.3.3 Show that the quartic curve $F = (x^2 + y^2 - zx)^2 - z^2(x^2 + y^2)$ in $\mathbb{PC}^2$ has exactly three singular points.

10.3.4 Let λ be a complex number and let F_λ be the quartic curve in $P\mathbb{C}^2$ given by $F_\lambda = x^4 + xy^3 + y^4 - \lambda zy^3 - 2x^2yz - xy^2z + y^2z^2$. Show that $O = (0 : 0 : 1)$ is singular on F_λ. Further, show that there are exactly two values of λ for which F_λ has a second singular point.

10.3.5 Show that the quintic $F = (x^2 - a^2z^2)^2y - (y^2 - b^2z^2)^2x$ in $P\mathbb{C}^2$ has exactly four singular points. ($a \neq 0$, $b \neq 0$.)

10.3.6 Show that the sextic curve $F = (z^2 - x^2 - y^2)^3 - 27x^2y^2z^2$ in $P\mathbb{C}^2$ has exactly ten singular points. (This is the complex projective version of the astroid.)

10.3.7 Show that the sextic curve $F = (x^2 + y^2)^3 - 4x^2y^2z^2$ in $P\mathbb{C}^2$ has exactly three singular points, namely a quadruple point P, and two double points at the circular points at infinity I, J.

10.3.8 The Euler Lemma expresses a form F of degree d in terms of the first order partials of F. Extend this to second order partials by showing that

$$d(d-1)F = \sum x_i x_j F_{x_i x_j}.$$

10.3.9 Let $f(x, y)$ be a polynomial over the field $\mathbb{K}$ which can be written in the form $f = F + G$ where F, G are forms of different degrees $d, e \geq 2$. Use the Euler Lemma to show that the singular points of f in $\mathbb{K}^2$ satisfy the relations $F = 0$, $G = 0$. If, in addition, F, G have no common factors, deduce that $(0, 0)$ is the only singularity of f. Deduce that the Maltese cross curve in $\mathbb{R}^2$ defined by $f = xy(x^2 - y^2) - (x^2 + y^2)$ has just one singularity, namely a double point at $(0, 0)$.

10.3.10 Let $p(x)$ be a complex polynomial of degree $d \geq 3$, let f be the affine curve in $\mathbb{C}^2$ defined by $f = y^2 - p(x)$, and let F be the corresponding projective curve in $P\mathbb{C}^2$. Show that F meets the line at infinity at a unique point P, of multiplicity $(d - 2)$. (The singular points of f were determined in Exercise 6.3.9.)

10.4 Delta Invariants viewed Projectively

In this section we will consolidate our account of affine conics in Chapter 5 by showing that the projective viewpoint gives considerable geometric insight into the delta invariants δ, Δ of that chapter. Recall that a general conic in $\mathbb{K}^2$ has the shape

$$q = ax^2 + 2hxy + by^2 + 2gx + 2fy + c \qquad (\star)$$

where the coefficients a, b, c, f, g, h lie in $\mathbb{K}$, and at least one of a, b, h is non-zero. The corresponding projective curve in $P\mathbb{K}^2$ is

$$Q = ax^2 + by^2 + cz^2 + 2hxy + 2fyz + 2gzx \qquad (\star\star)$$

The first thing to ask is how the conic meets the line at infinity. Setting $z = 0$ in $(\star\star)$ gives a binary quadratic equation $ax^2 + 2hxy + by^2 = 0$. Over the complex field this has two roots (giving two points at infinity on the conic) with coincidence if and only if the discriminant vanishes. The discriminant is $-4(ab - h^2) = -4\delta$, where δ is the affine invariant we met in Chapter 5, thus the geometric significance of the relation $\delta = 0$ is that the associated projective conic meets the line at infinity at a single point with intersection number 2. Over the real field we can say more. When $\delta < 0$, the binary quadratic has two real roots (and the conic has two real points at infinity), and when $\delta > 0$ the binary quadratic has no real roots (and the conic has no real points at infinity). Thus in the real case, the conditions $\delta < 0$, $\delta = 0$, $\delta > 0$ correspond to the projective conic Q having two, one or no points at infinity. Looking at Table 5.4 we see that for an *irreducible* real conic these conditions characterize respectively the hyperbola, the parabola and the two types of ellipse.

A geometric understanding of the invariant Δ is based on the following elementary result.

Lemma 10.9 *The conic $(\star\star)$ is singular if and only the matrix A below is singular.*

$$A = \begin{pmatrix} a & h & g \\ h & b & f \\ g & f & c \end{pmatrix}$$

Proof Q is singular if and only if there exists a non-trivial solution (x, y, z) for the following system of linear equations: and by linear algebra that is the case if and only if the matrix A of coefficients is singular.

$$\begin{aligned} Q_x &= 2(ax + hy + gz) = 0 \\ Q_y &= 2(hx + by + fz) = 0 \\ Q_z &= 2(gx + fy + cz) = 0. \end{aligned}$$

□

Recall from linear algebra that a matrix A is singular if and only if its determinant is zero. However the determinant of A is precisely the invariant Δ of Chapter 5, so the geometric interpretation of the condition $\Delta = 0$ is that the associated projective conic Q has a singular point.

Example 10.9 The conic $Q(x, y, z) = xy + yz + zx$ in $P\mathbb{K}^2$ is non-singular because its matrix A (below) has non-zero determinant.

$$A = \frac{1}{2}\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

Lemma 10.10 *A reducible conic Q is necessarily singular. And a singular conic Q with at least two points is reducible. Over the complex field a conic is singular if and only if it is reducible.*

Proof Suppose Q reduces, necessarily, to two lines L, M. These lines intersect in at least one point R which is singular on Q, by Example 10.5. Conversely, suppose Q is singular, with at least two points. Let R be a singular point of Q, and let S be a point on Q distinct from R. Then the line joining R, S meets Q in ≥ 3 points, so by Lemma 10.1 is a component, and Q is reducible. The final statement is immediate from the fact that any curve in $P\mathbb{C}^2$ has an infinite zero set. □

Over the real field, a singular conic may fail to be reducible; for instance, although the conic $Q = x^2 + y^2$ in $P\mathbb{R}^2$ is irreducible, the zero set contains only one point $P = (0 : 0 : 1)$, which is singular on Q. Finally, it is instructive to look again at the lists of normal forms for real affine conics in Table 5.4. All the normal forms with $\Delta = 0$ do indeed exhibit a singular point in the projective plane, though in the 'parallel lines' cases that point is 'at infinity'.

3) Books-002-2021\ Pappus\
Geometry A Comprehensive
Course.pdf

Pag. 363 - The Projective Plane
and Projective Space.

4) \ Pappus\ 2D Coordinate
Geometry.pdf

5) \ Geometry of Curves\
et45643yrt435....pdf

→ Elementary Geometry of
Algebraic Bcurves
(curves)

Pag. 108(125) - Chapter 9
Projective Algebraic Curves

a) Pag. 112(129) Fig 9.2
Real projective line as
a circle

Fig 9.3 Complex projective
line as a sphere

Fig. 9.4 the projective plane over $\mathbb{Z}_2$

b/ Pag. 116(133) Fig. 9.5
Hemisphere model of $PR^2(R)$

c/ Pag. 125(142) - Chapter 10
Singularities of Projective Curves.
d/ Pag. 137(154) - Chapter 11
Projective Equivalence
e/ Pag. 148(165) - Chapter 12
Projective Tangents
6/ Geometry of Curves|
try2y'r643ytr346-
geometry-of-curves.pdf
a/ Pag. 133(152) - Chapter 7
Curvature
b/ Pag. 151(170) - Chapter 8
Curvature: applications.
c/ Pag. 167(186) - Chapter 9

")*+,)))- !) $ 5 =5 0$5 $0 4 5
2 /"
' $$ (!/

7

Curvature

The (signed) curvature of a curve parametrised by its arc-length is the rate of change of direction of the tangent vector. The absolute value of the curvature is a measure of how sharply the curve bends. Curves which bend sharply will have a large absolute curvature. Curves which bend slowly, which are almost straight lines, will have a small absolute curvature. Straight lines themselves have zero curvature. Curves which swing to the left have positive curvature and curves which swing to the right have negative curvature. The curvature of the direction of a road will affect the maximum speed at which vehicles can travel without skidding, and the curvature in the trajectory of an aeroplane will affect whether the pilot will suffer "blackout" as a result of the g–forces involved. In this chapter we give the basic definitions and properties of curvature. In Chapter 8 we consider some applications of the curvature of parametrised curves, and in Chapter 9 we consider the circle of curvature.

7.1 Cartesian coordinates

7.1.1 General parameters

Let $t \mapsto \mathbf{r}(t)$ be a regular smooth parametrised curve[†]. The parameter t will not in general be arc-length. Since the curve is regular and smooth, the parametrisation will determine a direction along the curve as t increases. An *oriented curve* is a curve with such a direction along it. We wish to measure the rate of change of direction of the tangent vector in a way which is intrinsic to curve, that is, in a way which is not dependent on a particular parametrisation of the curve, but only on the orientation of the curve (direction along the curve) determined by the parametrisation. We measure this rate of change of direction with respect to the arc-length s. Let ψ be the angle measured counter-clockwise from the direction of the initial half-line (the positive x-axis) to the tangent vector (see Figure 7.1). Notice that[†]

[†] Often we use t as parameter. The reader should however become familiar with the use of other parameters such as s, usually the arc-length, and φ, often used in case the parameter is an angle.

$$\frac{d\mathbf{r}}{ds} = (\cos\psi, \sin\psi).$$

We assume in general that the direction of the positive y–axis is given by a counter-clockwise rotation through an angle $\frac{\pi}{2}$ from the direction of the positive x–axis.

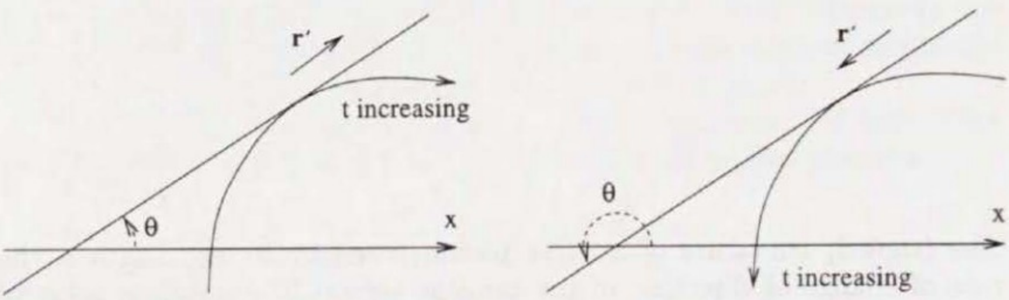

Figure 7.1 *Curvature.*

Definition 7.1. The *curvature* or *signed curvature* κ of a smooth curve $t \mapsto \mathbf{r}(t)$ at a regular point is the rate of change of direction at that point of the tangent line with respect to arc-length, that is

$$\boxed{\kappa = \frac{d\psi}{ds}\ ,}$$

where arc-length $t \mapsto s(t)$ is measured along the curve so that s increases as t increases. The *absolute curvature* of the curve at a point is the absolute value $|\kappa|$ of the curvature at that point.

In general the arc-length parametrisation of the curves we consider is not given by a simple formula and, in order to calculate the curvature of a specific parametrised curve, we need to determine a formula for curvature in terms of the given parameter. As a first step to this we give in the following lemma a formula for curvature in terms of the rate of change of direction of the tangent vector with respect to the parameter t by which the curve is described.

[†] We use ψ for the polar angle of $\mathbf{r}'$, and generally use θ or φ when considering the polar angle of $\mathbf{r}$.

Lemma 7.2. *The curvature of a curve $t \mapsto \mathbf{r}(t)$ at a regular point is*

$$\kappa = \kappa(t) = \frac{d\psi}{dt}\Big/\left|\frac{d\mathbf{r}}{dt}\right| = \frac{\dfrac{d\psi}{dt}}{\sqrt{\left(\dfrac{dx}{dt}\right)^2 + \left(\dfrac{dy}{dt}\right)^2}} = \frac{d\psi}{dt}\Big/\frac{ds}{dt} = \frac{d\psi}{ds}\,,$$

where arc-length $t \mapsto s(t)$ is measured in the direction given by t increasing.

*Solution**. We need to prove the last of the equalities. Since the arc-length is measured in terms of t increasing, we have at regular points $\dfrac{ds}{dt} > 0$ and therefore $\dfrac{ds}{dt} = |\mathbf{r}'|$. For a regular (smooth) curve $t \mapsto \mathbf{r}(t)$, we also have that $t \mapsto s(t)$ is smooth (see §4.6.2). By the inverse function theorem (see §4.7), in this case of a regular smooth curve, the function $t \mapsto s(t)$ has an inverse $s \mapsto t(s)$ which is also smooth. Therefore ψ can be regarded as a function of s by $s \mapsto \psi(t(s))$. The last of the equalities in the definition now follows from the chain rule $\dfrac{d\psi}{dt} = \dfrac{d\psi}{ds}\dfrac{ds}{dt}$, on regarding ψ as a function of s on the one hand and as a function of t on the other, where $s = s(t)$. □

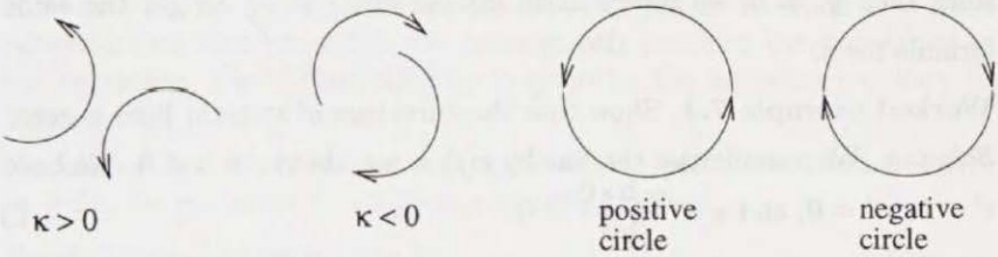

Figure 7.2 *Sign of the curvature.*

The curvature can be positive, negative or zero. The angle ψ is measured counterclockwise. Therefore the curvature of a parametrised curve is *positive* $\left(\kappa = \dfrac{d\psi}{ds} > 0\right)$ if the curve swings to the left as t increases, that is ψ increases as s increases. Similarly the curvature is *negative* $\left(\kappa = \dfrac{d\psi}{ds} < 0\right)$ if the curve swings to the right as t increases, that is ψ decreases as s increases (see Figure 7.2). Thus the *sign* of the curvature at a point depends on the direction along which the curve is parametrised at the point. If the curve is reparametrised in the opposite direction then the sign of the curvature becomes the opposite of what it had been.

---

* For the meaning of * see the preface.

Mnemonic

$$\boxed{\text{A curve with positive curvature swings to the left.}}$$

We next give a formula for curvature which is the one normally used in calculations.

Theorem 7.3. *At a regular point of a smooth curve* $t \mapsto \mathbf{r}(t)$, *the curvature satisfies*

$$\boxed{\kappa = \frac{x'\,y'' - y'\,x''}{((x')^2 + (y')^2)^{\frac{3}{2}}} = \frac{x'\,y'' - y'\,x''}{|\mathbf{r}'|^3} = -\,\frac{\mathbf{r}'\,.\,(-y'', x'')}{|\mathbf{r}'|^3}\,.}$$

Proof. Differentiating $\tan\psi = \dfrac{y'}{x'}$ with respect to t, we have, for $x' \neq 0$,

$$\frac{x'\,y'' - y'\,x''}{(x')^2} = \sec^2\psi \cdot \psi' = (1 + \tan^2\psi) \cdot \psi' = \left(1 + \left(\frac{y'}{x'}\right)^2\right) \cdot \psi',$$

and the result follows easily on multiplying each side by $(x')^2$. At any (regular) point at which $x' = 0$, we cannot use the above argument, but, since then $y' \neq 0$, we differentiate instead $\cot\psi = \dfrac{x'}{y'}$ to get the same formula for κ. □

Worked example 7.4. Show that the curvature of straight lines is zero.

Solution. We parametrise the line by $\mathbf{r}(t) = \mathbf{a} + t\mathbf{b}$ where $\mathbf{b} \neq \mathbf{0}$. We have $\mathbf{r}' = \mathbf{b}$, $\mathbf{r}'' = \mathbf{0}$, and $\kappa = \dfrac{-\mathbf{b}\cdot\mathbf{0}}{|\mathbf{b}|^3} = 0$. □

Worked example 7.5. Show that the curvature of a positively described (counter-clockwise) circle of radius R is $\dfrac{1}{R}$, and that the curvature of a negatively described (clockwise) circle of radius R is $-\dfrac{1}{R}$.

Solution. We parametrise the circle having centre $\mathbf{a}$ and radius R in the counter-clockwise direction by $\mathbf{r}(t) = \mathbf{a} + R(\cos t, \sin t)$. Therefore we have $\mathbf{r}' = R(-\sin t, \cos t)$, and $\mathbf{r}'' = -R(\cos t, \sin t)$. Therefore

$$\kappa = \frac{x'\,y'' - y'\,x''}{((x')^2 + (y')^2)^{\frac{3}{2}}} = \frac{R^2(\sin^2 t + \cos^2 t)}{(R^2\sin^2 t + R^2\cos^2 t)^{\frac{3}{2}}} = \frac{1}{R}\,.$$

If the circle is oriented in the clockwise direction, by

$$\mathbf{r}(t) = \mathbf{a} + R(\cos t, -\sin t)$$

for example, we obtain in a similar way that the curvature is $-\dfrac{1}{R}$. □

By Example 7.5, we have the following memory-aid for the sign of the curvature of a curve.

Mnemonic

A positively (counter-clockwise) described circle has positive curvature.

Worked example 7.6. Find the curvature of the curve $x = t^3 - t, y = t^2$.

Solution. We have $\mathbf{r}' = (3t^2 - 1, 2t)$, and $\mathbf{r}'' = (6t, 2)$. Therefore

$$\kappa = \frac{x'\,y'' - y'\,x''}{((x')^2 + (y')^2)^{\frac{3}{2}}} = \frac{(3t^2 - 1).2 - 2t.6t}{((3t^2 - 1)^2 + (2t)^2)^{\frac{3}{2}}} = -\frac{6t^2 + 2}{(9t^4 - 2t^2 + 1)^{\frac{3}{2}}}\,.$$

□

Notice that Theorem 7.3 tells us that

$$\frac{x'\,y'' - y'\,x''}{((x')^2 + (y')^2)^{\frac{3}{2}}}$$

is invariant under change of rectangular axes (rotation and/or translation), since the curvature $\dfrac{d\psi}{ds}$ is clearly invariant under such a change. If the axes are changed by a reflexion, the curvature simply changes sign. The curvature is also independent of the choice of parametrisation in a fixed direction (see Exercise 7.20). We subsequently consider the derivatives of the curvature. The differentiability is given by the following corollary to Theorem 7.3.

Corollary 7.7. *At a regular point of a smooth parametrisation* $t \mapsto \mathbf{r}(t)$ $(n \geq 2)$, *the curvature* $t \mapsto \kappa(t)$ *is a smooth function.*

Proof. The curvature is given by

$$\kappa = \frac{x'\,y'' - y'\,x''}{((x')^2 + (y')^2)^{\frac{3}{2}}}\,.$$

The numerator is a polynomial in the smooth functions x', y', x'', and y'', and therefore is a smooth function; similarly $(x')^2 + (y')^2$ is smooth. Also $X \to X^{\frac{3}{2}}$ is a smooth function of X at points for which $X > 0$. The result follows since the composite of two smooth functions is smooth, and the quotient of a smooth function by a non-zero smooth function is also smooth. □

7.1.2 Coordinates as parameter

An important special case of Theorem 7.3 is the formula for the curvature of a graph. The graph of a function $y = g(x)$ is usually assumed to be parametrised using x as parameter, that is by $\mathbf{r}(x) = (x, g(x))$. With this parametrisation the graph of $y = g(x)$ has the orientation given by x increasing. Similarly the curve given by $x = h(y)$ is usually assumed to be parametrised using y as parameter, that is by $\mathbf{r}(y) = (h(y), y)$, and has then the orientation given by y increasing.

Corollary 7.8. *The graph, parametrised by the coordinate x, of a smooth function $y = g(x)$ has curvature*

$$\boxed{\kappa(x) = \frac{\dfrac{d^2y}{dx^2}}{\left(1 + \left(\dfrac{dy}{dx}\right)^2\right)^{\frac{3}{2}}} = \frac{g''(x)}{(1 + (g'(x))^2)^{\frac{3}{2}}} \,.}$$

Proof. We have $\mathbf{r}' = (1, g'(x))$ and $\mathbf{r}'' = (0, g''(x))$. Thus

$$\kappa = \frac{x'\,y'' - y'\,x''}{|\mathbf{r}'|^3} = \frac{1\,.\,g''(x) - g'(x)\,.\,0}{(1 + (g'(x))^2)^{\frac{3}{2}}} \,.$$

☐

For a curve parametrised by the x–coordinate, the curvature is positive where the concavity of the curve is "upward", and is negative where the concavity is "downward"; moreover, in this case of a curve parametrised by the x–coordinate, the curvature has the same sign as $\dfrac{d^2y}{dx^2}$ by Corollary 7.8.

Compare the formula of Corollary 7.8 carefully with the formula for the curvature for general parametrisations. Of course, by Theorem 4.22, any algebraic curve may be parametrised locally, near a non-singular point, by either the x or the y coordinate, or by both. Each such local parametrisation will give a local orientation along the curve, determined by x increasing or y increasing respectively, and hence will determine a sign for the curvature. For some curves the curvature will have the opposite sign and the curve will have the opposite local orientation for increasing x to those given by increasing y.

Worked example 7.9. Find the curvature of the graph of $y = x^3$ parametrised by x.

Solution. We have $\dfrac{dy}{dx} = 3x^2$, and $\dfrac{d^2y}{dx^2} = 6x$. Therefore

$$\kappa = \frac{6x}{(1 + 9x^4)^{\frac{3}{2}}} \,.$$

☐

Worked example 7.10. Find the point of greatest absolute curvature of the graph of $y = \log x$.

Solution. The domain of parametrisation is $x > 0$. We have

$$\frac{dy}{dx} = \frac{1}{x}, \quad \text{and}$$
$$\frac{d^2y}{dx^2} = -\frac{1}{x^2}.$$

Therefore

$$\kappa = -\frac{x}{(1+x^2)^{\frac{3}{2}}}, \quad \text{and}$$

$$\frac{d\kappa}{dx} = \frac{2x^2-1}{(1+x^2)^{\frac{5}{2}}} = \frac{2\left(x-\frac{1}{\sqrt{2}}\right)\left(x+\frac{1}{\sqrt{2}}\right)}{(1+x^2)^{\frac{5}{2}}}.$$

The critical point is at $x = \frac{1}{\sqrt{2}}$, and the required point, which gives a minimum value of the signed curvature corresponding to the parametrisation by x and a maximum value of the absolute curvature, is $\left(\frac{1}{\sqrt{2}}, -\frac{1}{2}\log 2\right)$. The curve is not defined at $x = -\frac{1}{\sqrt{2}}$. □

7.1.3 A short cut

We note the following technique which is sometimes useful in shortening the calculation of curvature and in determining whether or not a cusp is ordinary. Recall that $(-v', u')$ is obtained by rotating (u', v') counter-clockwise through the angle $\frac{\pi}{2}$, in the case where $(u', v') \neq 0$. Let $\mathbf{Z}(t) = (X(t), Y(t))$ satisfy $\mathbf{Z}' = f(t).(u(t), v(t))$. Suppose we wish to calculate

$$\mathbf{Z}' \cdot (-Y'', X'').$$

We have

$$\mathbf{Z}'' = f'(t).(u(t), v(t)) + f(t).(u'(t), v'(t)).$$

Therefore

$$\mathbf{Z}' \cdot (-Y'', X'') = \mathbf{Z}' \cdot (f(t).(-v'(t), u'(t))),$$

since $(u'(t), v'(t)) \cdot (-v'(t), u'(t)) = 0$.

Notice that we have differentiated

$$f(t).(u(t), v(t))$$

as a product of a scalar function and a vector function, rather than writing it in the form $(f(t).u(t), f(t).v(t))$.

Exercises

7.1. Find the curvature at regular points of the following curves:

a) The rectangular hyperbola $t \mapsto \mathbf{r}(t) = \left(at, \frac{a}{t}\right)$.

$$\left[\frac{2a^2t^4}{a^3(1+t^4)^{\frac{3}{2}}};\ \text{none, none}\right]$$

b) The cuspidal cubic $t \mapsto \mathbf{r}(t) = (t^2, t^3)$.

c) The crunodal cubic $t \mapsto \mathbf{r}(t) = (t^2-1, t(t^2-1))$.

$$\left[\frac{6t^2+2}{(9t^4-2t^2+1)^{\frac{3}{2}}};\ \text{none, none}\right]$$

d) The acnodal cubic $t \mapsto \mathbf{r}(t) = (1+t^2, t(1+t^2))$.

In each case find the non-regular points and the points at which the curvature is zero.

7.2. Prove that the ellipse $x = a\cos\varphi, y = b\sin\varphi$ $(0 < b < a)$ has curvature

$$\kappa = \frac{ab}{(a^2\sin^2\varphi + b^2\cos^2\varphi)^{\frac{3}{2}}}.$$

7.3. Show that the right-hand branch $x = a\sec\varphi, y = b\tan\varphi$ of the hyperbola $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$ has curvature

$$\kappa = -\frac{ab}{(a^2\tan^2\varphi + b^2\sec^2\varphi)^{\frac{3}{2}}}.$$

Find the curvature on the left-hand branch $x = -a\sec\varphi, y = b\tan\varphi$.

7.4. Show that the parabola $x = at^2, y = 2at$ has curvature

$$\kappa = -\frac{1}{2a\,(1+t^2)^{\frac{3}{2}}}.$$

7.5. Show that the curvature of the cycloid $x = a(t - \sin t)$, $y = a(1-\cos t)$ is

$$-\frac{1}{4a\left|\sin\frac{t}{2}\right|}.$$

7.6. Show that the curvature of the curve

$$t \mapsto \mathbf{r}(t) = (a\sin 2t\,(1+\cos 2t), a\cos 2t\,(1-\cos 2t))$$

is $4a\cos 3t$.

7.7. The curvature of the ellipse given by $x = a\cos\varphi, y = b\sin\varphi$ $(0 < b < a)$ is

$$\kappa = \frac{ab}{(a^2\sin^2\varphi + b^2\cos^2\varphi)^{\frac{3}{2}}}.$$

Show that

$$a^2 \sin^2 \varphi + b^2 \cos^2 \varphi = (a^2 - b^2) \sin^2 \varphi + b^2$$

is strictly monotone increasing in the interval $0 \le \varphi \le \frac{\pi}{2}$ and hence deduce that there are precisely four points on the ellipse at which the curvature of the ellipse is $\frac{1}{a}$. (This is the curvature of the circle of radius a which has, as a diameter, the major axis of the ellipse.)

7.8. Show that the curve $\mathbf{r}(t) = a(\sin 2bt, \sin bt)$ is regular and that its tangent is parallel to the y–axis if, and only if, $bt = \pm\frac{\pi}{4}$ or $\pm\frac{3\pi}{4}$. Find the point where the tangent is parallel to the x–axis. Determine the curvature at those points where the tangent is parallel to either axis.

7.9. Find the curvature at the origin of the curve $y = \sin ax^2$ parametrised by the x coordinate. $[\kappa = a]$

7.10. Find the curvature at the origin of the curve $y = x^2 - x^3$ parametrised by the x coordinate.

7.11. Show that the curvature of the catenary $y = a \cosh \frac{x}{a}$ parametrised by the x coordinate is $\frac{a}{y^2}$.

7.12. Find the curvature of a curve given by $x = h(y)$ and parametrised by the y–coordinate.

7.13. A curve which can be parametrised near the origin by a smooth function $y = f(x)$ has the x–axis as tangent at the origin. Prove that the curvature at the origin is given by Newton's formula:

$$\kappa(0) = \lim_{x \to 0} \frac{2y}{x^2}.$$

Obtain a similar formula in case the curve is parametrised near the origin by a smooth function $x = g(y)$ and has the y–axis as tangent at the origin. (*Hint:* Use L'Hôpital's theorem – see §4.7.5.)

$$\left[\kappa(0) = -\lim_{y \to 0} \frac{2x}{y^2}\right]$$

7.14. Find the curvature at the origin of the quartic curve

$$3x^4 + 2y^4 + 4xy^2 + xy - x^2 + 2y = 0.$$

(*Hint:* Use the local parametrisation theorem (Theorem 4.22) to show that there is a local parametrisation $y = f(x)$ near the origin, divide the algebraic equation by x^2, and use Exercise 7.13.) $[\kappa = 1]$

7.15. Show that the curvature of the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ parametrised locally by x is $\kappa = \pm \frac{ab}{(a^2 - e^2x^2)^{\frac{3}{2}}}$, where e is the eccentricity. Indicate on a diagram which parts of the ellipse have positive curvature in this case and which have negative curvature.

7.16. Find the maximum distance from the origin of points on the curve

$$t \mapsto \mathbf{r}(t) = \left(a\sin t - b\sin\frac{at}{b}\, , a\cos t - b\cos\frac{at}{b}\right) .$$

Show that the curvature at such a point of maximum distance is $\dfrac{a+b}{4ab}$.

7.17. Show that the curvature of the hyperbola $\dfrac{x^2}{a^2} - \dfrac{y^2}{b^2} = 1$ parametrised by x is $\kappa = \pm\dfrac{ab}{(e^2x^2 - a^2)^{\frac{3}{2}}}$, where e is the eccentricity. Indicate on a diagram which parts of the hyperbola have positive curvature in this case and which have negative curvature.

7.18. Show that the curvature of a general curve satisfies

$$\kappa = \frac{\mathbf{r}''\,.\,\mathbf{n}}{|\mathbf{r}'|^3} ,$$

where $\mathbf{n} = (-y', x')$ is the (usual) normal.

7.19. Let $s \mapsto \mathbf{r}(s)$ be a curve parametrised by arc-length. Prove that $\mathbf{r}'$ and $(-y'', x'')$ are parallel. Deduce that $|\kappa(s)| = |\mathbf{r}''(s)|$. Prove further that $\mathbf{r}'' \cdot \mathbf{r}'' + \mathbf{r}' \cdot \mathbf{r}''' = 0$ and deduce that $\kappa(s)^3 = x''\,y''' - y''\,x'''$.

7.20. Show that the curvature of a curve $t \mapsto \mathbf{r}(t)$ is unchanged by rotation or translation of the coordinate axes. Prove that it is also left unchanged by a change of parameter given by $t = g(\tau)$ where g is a smooth function and $g'(\tau) > 0$ for all relevant values of τ.

7.21. Let $t \to \mathbf{r}(t)$ be a regular parametrised curve and let $t \to \mathbf{n}(t)$ be a smooth function such that $\mathbf{n}(t)$ is one of the two unit normals to the curve at $\mathbf{r}(t)$. Let the curvature, for this choice of normal, be defined by

$$\mathbf{n}'(t) + \kappa(t)\mathbf{r}'(t) = \mathbf{0}.$$

Deduce that $\mathbf{n} \cdot \mathbf{r}'' = \kappa\mathbf{r}' \cdot \mathbf{r}'$ for each value of t, and that $\mathbf{n} \cdot \mathbf{r}''' = 3\kappa\mathbf{r}' \cdot \mathbf{r}''$ at t if, and only if, $\kappa'(t) = 0$. Prove also that if $\mathbf{n}' \cdot \mathbf{n}' = 1$ for all t, then $\kappa'(t) = 0$ if, and only if, $\mathbf{r}' \cdot \mathbf{r}'' = 0$.

7.2 Curves given by polar equation

We now develop a technique which we can use to find the curvature of curves given by a polar equation. Let

$$\mathbf{w} = (\cos\varphi, \sin\varphi),$$

be the parametrisation of the unit circle $x^2 + y^2 = 1$ where the parameter φ in this case is the polar angle θ, then we have

$$\begin{aligned}\mathbf{w}' &= (-\sin\varphi, \cos\varphi), \text{ and}\\ \mathbf{w}'' &= -(\cos\varphi, \sin\varphi).\end{aligned}$$

The vectors $(\cos\varphi, \sin\varphi)$ and $(-\sin\varphi, \cos\varphi)$ satisfy

$$\begin{aligned}(\cos\varphi, \sin\varphi)\cdot(\cos\varphi, \sin\varphi) &= 1,\\ (-\sin\varphi, \cos\varphi)\cdot(-\sin\varphi, \cos\varphi) &= 1, \text{ and}\\ (\cos\varphi, \sin\varphi)\cdot(-\sin\varphi, \cos\varphi) &= 0,\end{aligned} \tag{7.1}$$

that is they are orthogonal unit vectors or orthonormal vectors.

A curve given by a polar equation $r = r(\theta)$, has the parametric equation

$$\mathbf{r} = r(\varphi)(\cos\varphi, \sin\varphi),$$

according to our convention of replacing θ, the polar angle, by φ. Then we have by the chain rule

$$\begin{aligned}\mathbf{r}' &= r'(\varphi)(\cos\varphi, \sin\varphi) + r(\varphi)(-\sin\varphi, \cos\varphi),\\ \mathbf{r}'' &= r''(\varphi)(\cos\varphi, \sin\varphi) + 2r'(\varphi)(-\sin\varphi, \cos\varphi) - r(\varphi)(\cos\varphi, \sin\varphi),\end{aligned}$$

and

$$\begin{aligned}(-y'', x'') = r''(\varphi)(-\sin\varphi, \cos\varphi) &- \\ &2r'(\varphi)(\cos\varphi, \sin\varphi) - r(\varphi)(-\sin\varphi, \cos\varphi).\end{aligned}$$

Notice how $\mathbf{r}$, $\mathbf{r}'$, $\mathbf{r}''$, and $(-y'', x'')$ are expressed as linear combinations of the two orthogonal unit vectors $(\cos\varphi, \sin\varphi)$ and $(-\sin\varphi, \cos\varphi)$. These two vectors vary by rotating as φ increases. The above technique of writing the successive derivatives of $\mathbf{r}$ as linear combinations of these two rotating orthogonal vectors considerably shortens and simplifies the calculation needed to determine the curvature of curves given by their polar equation. We now have easily, on using Equations 7.1,

$$\begin{aligned}|\mathbf{r}'|^2 &= \{r'(\varphi)\}^2 + \{r(\varphi)\}^2, \text{ and}\\ -\mathbf{r}'\cdot(-y'', x'') &= r^2 + 2(r')^2 - r\,r''.\end{aligned}$$

The following theorem is obtained by substituting these into the formula for curvature given in Theorem 7.3.

Theorem 7.11. *The curvature of a curve given by a polar equation $r = r(\theta)$ is*

$$\kappa(\theta) = \frac{r^2 + 2\left(\dfrac{dr}{d\theta}\right)^2 - r\dfrac{d^2r}{d\theta^2}}{\left(r^2 + \left(\dfrac{dr}{d\theta}\right)^2\right)^{\frac{3}{2}}}.$$

The parameter used for this result is θ and therefore the curve has the orientation given by increasing θ. As usual the sign of the curvature is positive if the curve turns to the left as θ increases. Notice that the formula for $\kappa(\theta)$ is unchanged when r is replaced by $-r$ and θ is replaced by $\theta + \pi$. Therefore the formula is valid where $r = r(\theta)$ is given by the polar equation, and is equally valid where r is the polar coordinate ($r \geq 0$) of the point given by the polar equation; recall that the polar coordinates are either $(r(\theta), \theta)$ or $(-r(\theta), \theta + \pi)$, where r is given by the polar equation.

We now use the above technique as a model for calculating the curvature in a specific example. It is suggested that the reader calculate the curvature of curves given by their polar equation directly in this way rather than by learning and using the formula of Theorem 7.11. We stress that direct calculations are far more complicated if the above technique involving orthogonal vectors is not used.

Worked example 7.12. Find the curvature of the equi-angular spiral

$$\mathbf{r}(\varphi) = e^{a\varphi}(\cos\varphi, \sin\varphi),$$

where a is a real number.

Solution. We have

$$\begin{aligned}\mathbf{r}' &= ae^{a\varphi}(\cos\varphi, \sin\varphi) + e^{a\varphi}(-\sin\varphi, \cos\varphi),\\ \mathbf{r}'' &= a^2e^{a\varphi}(\cos\varphi, \sin\varphi) + 2ae^{a\varphi}(-\sin\varphi, \cos\varphi) - e^{a\varphi}(\cos\varphi, \sin\varphi),\end{aligned}$$

and

$$(-y'', x'') = a^2e^{a\varphi}(-\sin\varphi, \cos\varphi) - 2ae^{a\varphi}(\cos\varphi, \sin\varphi) - e^{a\varphi}(-\sin\varphi, \cos\varphi).$$

Therefore $\mathbf{r}' \cdot (-y'', x'') = -2a^2e^{2a\varphi} + a^2e^{2a\varphi} - e^{2a\varphi}$ (deduce this directly using the fact that $(\cos\varphi, \sin\varphi)$ and $(-\sin\varphi, \cos\varphi)$ are orthogonal unit vectors). Hence the curvature is

$$\kappa = \frac{(1+a^2)e^{2a\varphi}}{(1+a^2)^{\frac{3}{2}}e^{3a\varphi}} = \frac{e^{-a\varphi}}{\sqrt{1+a^2}}.$$

□

Exercises

7.22. For the following curves use the technique described in this section to show that the curvature is as given:

a) The circle $r = a\cos\theta$, $\kappa = \dfrac{2}{a}$.

b) The Archimedean spiral $r = a\theta$, $\kappa = \dfrac{2a^2 + r^2}{(a^2 + r^2)^{\frac{3}{2}}}$.

c) The cardioid $r = a(1 + \cos\theta)$, $\kappa = \dfrac{3}{4a}\left|\sec\dfrac{\theta}{2}\right|$.

7.23. Calculate the curvature of the curves of Exercise 7.22 by using the formula of Theorem 7.11.

7.24. Show that the astroid $\varphi \mapsto \mathbf{r}(\varphi) = (a\cos^3\varphi, a\sin^3\varphi)$ has curvature at regular points given by

$$\kappa = -\frac{2}{3a|\sin 2\varphi|}.$$

7.25. Show that the curvature of $r = a\cos 2\theta$ is

$$\frac{3\sin^2 2\theta + 5}{a(3\sin^2 2\theta + 1)^{\frac{3}{2}}}.$$

7.26. Show that the curvature of the curve

$$x(t) = \int_0^t \frac{\cos u}{\sqrt{u}}\,du\,, \quad y(t) = \int_0^t \frac{\sin u}{\sqrt{u}}\,du$$

is $\kappa = -\sqrt{t}$. Sketch the curve.

7.27. Show that the curvature of $r\cos 2\theta = a$ at $r = a$ is

$$\kappa = -\frac{3}{a}.$$

7.28. Show that the curvature of $r^2 = a^2(1 + \cos\theta)$ at $\theta = 0$ is

$$\kappa = \frac{5}{4\sqrt{2}\,a}.$$

7.29. For the sinusoidal spirals $r^m = a^m \cos m\theta$, prove that

$$\frac{dr}{d\theta} = -r\tan m\theta, \text{ and that } \frac{d^2r}{d\theta^2} = r(\tan^2 m\theta - m\sec^2 m\theta).$$

Deduce that

$$\kappa = \frac{(m+1)|r|^{m-1}}{a^m}.$$

(Special cases of this general curve for various values of rational m include the line ($m = -1$), the circle ($m = 1$), the parabola $\left(m = -\frac{1}{2}\right)$, the cardioid $\left(m = \frac{1}{2}\right)$, the rectangular hyperbola ($m = -2$), and Bernoulli's lemniscate ($m = 2$).)

7.3 Curves in the Argand diagram

For certain curves, especially ones given by a polar equation, the representation of points $\mathbf{r} = (x, y)$ of the plane by the complex number $z = x + iy$ can very considerably shorten and simplify the calculation of the curvature of the curve.

Let the parametrised curve $J \mapsto \mathbf{C}$ in the Argand diagram be given by $t \mapsto z(t) = x(t) + iy(t)$. Recall that multiplying by i in the Argand diagram corresponds to rotating counter-clockwise through an angle $\frac{\pi}{2}$. The next theorem, which is is equivalent to Theorem 7.3, follows on using the bijection between the plane $\mathbf{R}^2$ and the Argand diagram $\mathbf{C}$ which associates the vector $\mathbf{r} = (x, y)$ in the plane to the vector $z = x + iy$ in the Argand diagram: we simply replace the vector $\mathbf{r}$ in the formula for curvature in Theorem 7.3 by the vector z, and recall that the scalar product in $\mathbf{C}$ of z and $w = u + iv$ is given by $(z, w) \mapsto \operatorname{re}(z\overline{w})$, and thus the scalar product of z and iw is $-\operatorname{im}(z\overline{w})$.

Theorem 7.13. *At a regular point of a smooth parametrisation the curvature satisfies*

$$\boxed{\kappa = -\frac{\operatorname{im}\left(z'\,\overline{z''}\right)}{|z'|^3}\,.}$$

Note especially the negative signs which occur in the last formulae of Theorem 7.3 and in Theorem 7.13. Also when calculating curvature we use the formula $|z'|^2 = z' \,.\, \overline{z'}$, which again can shorten the calculations for curves related to circles.

The technique we use in calculations is similar to that given in §7.2. Let

$$w(\varphi) = e^{i\varphi},$$

be the parametrisation of the unit circle $|z| = 1$ where the parameter φ is the polar angle θ, then we have

$$w' = ie^{i\varphi}, \text{ and}$$
$$w'' = -e^{i\varphi}.$$

The vectors $e^{i\varphi}$ and $ie^{i\varphi}$ are orthogonal unit vectors, or orthonormal vectors, in the Argand diagram, that is they satisfy $|e^{i\varphi}| = 1$, $|ie^{i\varphi}| = 1$, and

$$\operatorname{re}\left(e^{i\varphi}.\overline{ie^{i\varphi}}\right) = \operatorname{re}(-i) = 0.$$

The condition for complex numbers z and w to be orthogonal is that $\operatorname{re}(z.\overline{w}) = 0$, or, equivalently, that multiplying one by i gives a scalar (i.e., real) multiple of the second. A curve given by a polar equation $r = r(\theta)$, has the parametric equation

$$z(\varphi) = r(\varphi)e^{i\varphi}$$

according to our convention of replacing θ by φ. Then we have

$$z' = r'e^{i\varphi} + rie^{i\varphi},$$
$$z'' = (r'' - r)e^{i\varphi} + 2r'ie^{i\varphi}.$$

Notice how z, z', and z'' are expressed as linear combinations of the two orthogonal unit vectors $e^{i\varphi}$ and $ie^{i\varphi}$, which rotate as φ increases. We now have easily

$$|z'|^2 = \{r'(\varphi)\}^2 + \{r(\varphi)\}^2, \text{ and}$$
$$-\text{im}\,(z'\,\overline{z''}) = r^2 + 2(r')^2 - r\,r''.$$

Theorem 7.11 again follows from Theorem 7.13 and these last two equations. However, we use the above technique, which is similar to that of §7.2, for calculating the curvature in specific examples.

Worked example 7.14. Show that the equi-angular spiral $z = e^{(a+i)\varphi}$, where a is real, has curvature $\kappa = \dfrac{e^{-a\varphi}}{\sqrt{1+a^2}}$.

Solution. We have

$$\begin{aligned} z' &= (a+i)e^{(a+i)\varphi}, \\ z'' &= (a+i)^2 e^{(a+i)\varphi}, \\ |z'| &= |(a+i)e^{i\varphi}e^{a\varphi}| = \sqrt{1+a^2}e^{a\varphi}, \\ z'\,\overline{z''} &= (a+i)e^{(a+i)\varphi}\,\overline{(a+i)^2 e^{(a+i)\varphi}} \\ &= (a+i)e^{(a+i)\varphi}\,(a-i)^2 e^{(a-i)\varphi} = (a-i)(1+a^2)e^{2a\varphi}, \text{ and} \\ -\text{im}\,(z'\,\overline{z''}) &= (1+a^2)e^{2a\varphi}. \end{aligned}$$

Hence the curvature $\kappa = \dfrac{(1+a^2)e^{2a\varphi}}{(1+a^2)^{\frac{3}{2}}e^{3a\varphi}} = \dfrac{e^{-a\varphi}}{\sqrt{1+a^2}}$. ☐

Exercises

7.30. Use the technique of this section to show that the curvature is as given for the following curves (in each case a is a real number):

a) The circle $z = (a\cos\varphi)\,e^{i\varphi}$, $\kappa = \dfrac{2}{a}$.

b) The Archimedean spiral $z = a\varphi e^{i\varphi}$, $\kappa = \dfrac{2+\varphi^2}{a(1+\varphi^2)^{\frac{3}{2}}}$.

c) The cardioid $z = a(1+\cos\varphi)e^{i\varphi}$, $\kappa = \dfrac{3}{4a}\left|\sec\dfrac{\varphi}{2}\right|$.

7.4 An alternative formula

As well as being an important theoretical result which we shall use later, the following theorem gives a useful alternative way of calculating the curvature, especially where points of the plane are represented by complex numbers in the Argand diagram.

Theorem 7.15. *At a regular point of smooth parametrised curve in the Argand diagram we have*

$$\boxed{\frac{d}{dt}\left(\frac{z'(t)}{|z'(t)|}\right) = i\kappa z'(t).}$$

In particular, in the case where the parametrisation is by the arc-length s, we have

$$\boxed{z''(s) = i\kappa z'(s).}$$

Proof. Since $z' = |z'|e^{i\psi}$, where $\psi = \psi(t)$ as in §7.1.1 is the angle measured counter-clockwise from the initial half-line to the tangent vector $z' = z'(t)$ or equivalently is the polar angle of z', we have, on using Lemma 7.2,

$$\frac{d}{dt}\left(\frac{z'}{|z'|}\right) = \frac{d}{dt}\left(e^{i\psi}\right) = ie^{i\psi}\psi' = ie^{i\psi}|z'|\kappa = i\kappa z',$$

as required. □

Notice that $\dfrac{z'(t)}{|z'(t)|}$ is the unit (length) tangent vector, and that iz' is a normal vector. The derivative of the unit tangent vector is therefore orthogonal to the tangent vector provided that the curvature is not zero. In the case of arc-length parametrisation, $z'(s)$ is already a unit tangent vector for each s and $iz'(s)$ is the corresponding unit normal vector.

Worked example 7.16. Use Theorem 7.15 to show that the curvature at all points of a circle of radius R is either $\dfrac{1}{R}$ or $-\dfrac{1}{R}$, depending on the orientation.

Solution. Consider the circle oriented counter-clockwise by the parametrisation $z = z_0 + Re^{it}$ $(0 \le t \le 2\pi)$. This has centre $z_0 = x_0 + iy_0$ and radius R. We have

$$z' = iRe^{it}, \quad \frac{z'}{|z'|} = ie^{it}, \quad \text{and} \quad \frac{d}{dt}\left(\frac{z'}{|z'|}\right) = -e^{it} = \frac{1}{R}iz'.$$

Thus, by Theorem 7.15, the curvature at all points of this circle of radius R (which is oriented counter-clockwise) is $\dfrac{1}{R}$. Similarly for circles

$$z = z_0 + Re^{-it} \quad (0 \le t \le 2\pi)$$

oriented clockwise the curvature is $-\dfrac{1}{R}$ at all points. □

The equivalent of Theorem 7.15 for curves in $\mathbf{R}^2$ is the following corollary.

Corollary 7.17. *At a regular point of a smooth parametrised curve in* $\mathbf{R}^2$ *we have*

$$\boxed{\frac{d}{dt}\left(\frac{1}{|\mathbf{r}'(t)|}\,\mathbf{r}'(t)\right) = \kappa\,.\,(-y'(t), x'(t)).}$$

In particular, in the case where the parametrisation is by the arc-length s, *we have*

$$\boxed{\mathbf{r}''(s) = \kappa\,(-y'(s), x'(s)) \quad \textit{and} \quad (-y''(s), x''(s)) = -\kappa\mathbf{r}'(s).}$$

Note that in the latter case, $\mathbf{r}'(s)$ is a unit tangent vector for each s and $(-y'(s), x'(s))$ is the corresponding unit normal vector.

It is not necessary to calculate the curvature itself to find points at which the curvature is zero. The following corollary can save unnecessary calculation.

Corollary 7.18. *At a regular point of a smooth parametrisation the curvature is zero if, and only if,*

$$x'y'' - y'x'' = \mathbf{r}' \cdot (-y'', x'') = -\mathrm{im}\,\left(z'\,\overline{z''}\right) = 0.$$

Equivalently the curvature is zero if, and only if,

$$\frac{d}{dt}\left(\frac{1}{|\mathbf{r}'|}\,\mathbf{r}'\right) = \frac{d}{dt}\left(\frac{z'}{|z'|}\right) = 0\,.$$

In order to use this result to find points at which the curvature is zero, we must check that the points we find are regular points. The curvature is not normally defined at non-regular points.

Exercises

7.31. Use Theorem 7.13 to show that the Archimedean spiral $z = a\varphi e^{i\varphi}$ has curvature $\kappa = \dfrac{\varphi^2 + 2}{a(\varphi^2 + 1)^{\frac{3}{2}}}$.

7.32. Use Theorem 7.15 to show that the equi-angular spiral $z = e^{(a+i)t}$ has curvature $\dfrac{e^{-at}}{\sqrt{1 + a^2}}$.

7.33. Let $\mathbf{n} = \dfrac{1}{|\mathbf{r}'|}(-y', x')$ be the standard unit normal vector. Show that

$$\mathbf{n}' = -\kappa\mathbf{r}'.$$

7.34. Show that the line $z = z_0 + tw$ $(w \neq 0)$ has curvature $\kappa = 0$.

' (!/ . 2! !// !

8

Curvature: applications

In this chapter we give applications of the curvature of parametrised curves. First we show how the curvature and its derivatives can be used to determine inflexions, vertices, and undulations. Then we show how the curvature of algebraic curves can be determined. Finally we show how cusps can be classified by considering the limiting curvature at the cusp, together with the sign of the curvature at points on each side of the cusp.

8.1 Inflexions of parametric curves at regular points

We defined a point of simple inflexion at a regular point of a parametrised curve to be a point at which the tangent has three-point contact, or, equivalently, a point at which $\mathbf{r}''$ is parallel to $\mathbf{r}'$ and $\mathbf{r}'''$ is not parallel to $\mathbf{r}'$ (see §5.3.1). Recall that $\mathbf{r}''$ is 'parallel to $\mathbf{r}'$ includes the possibility that $\mathbf{r}'' = \mathbf{0}$. We now give a criterion for determining inflexions using the curvature and its derivatives.

Criterion 8.1. *A regular point of a twice differentiable parametrisation is a point of simple inflexion if, and only if,* $\kappa = 0$ *and* $\kappa' \neq 0$.

Proof. This result follows easily from Theorem 7.3. First $(-y'', x'')$ is obtained from $\mathbf{r}''$ by a counter-clockwise rotation through the angle $\frac{\pi}{2}$. We have $\kappa = 0$ if, and only if, $\mathbf{r}' \cdot (-y'', x'') = 0$, that is if, and only if, $\mathbf{r}''$ is a parallel to $\mathbf{r}'$. Also

$$\begin{aligned}\kappa' &= \frac{d}{dt}\left(-\frac{\mathbf{r}' \cdot (-y'', x'')}{|\mathbf{r}'|^3}\right)\\ &= -\frac{\mathbf{r}'' \cdot (-y'', x'')}{|\mathbf{r}'|^3} - \frac{\mathbf{r}' \cdot (-y''', x''')}{|\mathbf{r}'|^3} - \mathbf{r}' \cdot (-y'', x'')\frac{d}{dt}\left(\frac{1}{|\mathbf{r}'|^3}\right).\end{aligned}$$

The first term on the right is always 0, the third term is 0 whenever κ is 0, and the second term is 0, if, and only if, $\mathbf{r}'''$ is a parallel to $\mathbf{r}'$. □

More generally a point of inflexion is a point (a, b) where $\kappa = 0$ and where κ has opposite signs on the curve close to (a, b) on either side of (a, b), that is it is positive along the curve on one side of (a, b) and negative along the curve on the other side of (a, b).

Example 8.2. On the line $\mathbf{r} = (a, b) + t(p, q)$ we have $\mathbf{r}' = (p, q)$ and $\mathbf{r}'' = \mathbf{0}$. Thus $\kappa = \kappa' = 0$ at all points on the line. Therefore at every point on a line we have $\kappa = 0$; but no point on a line is a point of simple inflexion, since κ' is zero at all such points.

Worked example 8.3. Prove that the acnodal cubic $y^2 = x^3 - x^2$ has precisely two points of simple inflexion and find them.

Solution. We parametrise the curve excluding its isolated point, by

$$\mathbf{r} = (x, y) = (t^2 + 1,\ t(t^2 + 1))$$

for real values of t. We have

$$\begin{aligned} \mathbf{r}' &= (2t,\ 3t^2 + 1), \\ \mathbf{r}'' &= (2,\ 6t), \\ (-y'', x'') &= (-6t,\ 2), \\ (\mathbf{r}')^2 &= 9t^4 + 10t^2 + 1, \text{ and} \\ -\mathbf{r}' \cdot (-y'', x'') &= 6t^2 - 2. \end{aligned}$$

Thus the curvature is given by

$$\kappa = \frac{6t^2 - 2}{(9t^4 + 10t^2 + 1)^{\frac{3}{2}}},$$

which is zero if, and only if, $t = \pm\frac{1}{\sqrt{3}}$. Writing

$$t^2 - \tfrac{1}{3} = \left(t - \tfrac{1}{\sqrt{3}}\right)\left(t + \tfrac{1}{\sqrt{3}}\right),$$

we have, by Lemma 8.4 below,

$$\kappa'\left(\tfrac{1}{\sqrt{3}}\right) = \left.\frac{6\left(t + \frac{1}{\sqrt{3}}\right)}{(9t^4 + 10t^2 + 1)^{\frac{3}{2}}}\right|_{t=\frac{1}{\sqrt{3}}} \neq 0,$$

Similarly $\kappa'\left(-\frac{1}{\sqrt{3}}\right) \neq 0$. Therefore the acnodal cubic has precisely two points of simple inflexion, at $t = \pm\frac{1}{\sqrt{3}}$. □

8.1.1 *A useful lemma*

The following lemma can save a great deal of work in calculating the derivative of a function at point at which the function is zero, and in determining the sign of such a derivative. One use is to show easily (in suitable cases) which inflexions are simple. Another is to determine whether a critical point is a maximum or a minimum.

Lemma 8.4. *Let $f(x) = g(x)h(x)$ where g and h are differentiable at a and $g(a) = 0$. Then $f'(a) = g'(a)h(a)$.*

By this lemma, in order to determine $f'(a)$, or simply the sign of $f'(a)$, it is not necessary to differentiate the whole of the formula for f, but simply that 'factor' which is zero at $x = a$.

Proof. We have

$$f'(a) = g'(a)h(a) + g(a)h'(a)$$

by the rule for differentiating a product, and the result is immediate. □

Example 8.5. Let $f(x) = (x-2)\tan^{-1}\left(\frac{x^2+x+1}{x^4+1}\right)$, then (immediately) we have

$$f'(2) = \tan^{-1}\left(\frac{4+2+1}{16+1}\right) = \tan^{-1}\left(\frac{7}{17}\right).$$

Exercises

8.1. Find the curvature of the following curves and determine their inflexions

a) $(\frac{1}{2}t^2, t + \frac{1}{3}t^3)$ $\left[\frac{t^2-1}{(t^4+3t^2+1)^{\frac{3}{2}}}; \pm 1, \text{ both simple}\right]$

b) $(t + \frac{1}{2}t^2, \frac{1}{3}t^3)$

8.2. Let $f(x) = g(x)h(x)$ where g and h are twice differentiable at a and $g(a) = g'(a) = 0$. Show that $f''(a) = g''(a)h(a)$.

8.3. Let $f(x) = g(x)h(x)$ where g and h are n–times differentiable at a and $g(a) = g'(a) = \cdots = g^{(n-1)}(a) = 0$. Show that $f^n(a) = g^n(a)h(a)$.

8.2 Vertices and undulations at regular points

We defined a point of simple undulation at a regular point of a parametrised curve to be a point at which the tangent has four-point contact, or, equivalently a point at which $\mathbf{r}'$, $\mathbf{r}''$, and $\mathbf{r}'''$, are parallel and $\mathbf{r}^{(4)}$ is not parallel to $\mathbf{r}'$ (see §5.3.1). We now give an alternative criterion for undulation in terms of the curvature.

Criterion 8.6. *A regular point of a three times differentiable parametrisation is a point of simple undulation if, and only if, $\kappa = \kappa' = 0$ and $\kappa'' \neq 0$.*

Proof. By Criterion 8.1 we know that $\kappa = 0$ if, and only if, $\mathbf{r}'$ and $\mathbf{r}''$ are parallel, that $\kappa = \kappa' = 0$ if, and only if, $\mathbf{r}'$, $\mathbf{r}''$, and $\mathbf{r}'''$ are parallel, and that in general

$$\kappa' = -\frac{\mathbf{r}' \cdot (-y''', x''')}{|\mathbf{r}'|^3} - (\mathbf{r}' \cdot (-y'', x''))\,\frac{d}{dt}\left(\frac{1}{|\mathbf{r}'|^3}\right).$$

At a point where $\mathbf{r}'$, $\mathbf{r}''$ and $\mathbf{r}'''$ are parallel, we have, on using Lemma 8.4,

$$\kappa'' = -\frac{\mathbf{r}' \cdot (-y^{(4)}, x^{(4)})}{|\mathbf{r}'|^3}.$$

Thus given that $\kappa = \kappa' = 0$, we have that $\kappa'' \neq 0$ if, and only if, $\mathbf{r}^{(4)}$ is not parallel to $\mathbf{r}'$. □

More generally a point of undulation is a point (a, b) where $\kappa = 0$ and where κ has the same sign on the curve on either side of (a, b), that is either it is positive along the curve on both sides of (a, b) or negative along the curve on both sides of (a, b).

Definition 8.7. A *vertex* is a (regular) point for which $\kappa' = 0$.

Thus points, at which the curvature $\kappa(t)$ has a local maximum, a local minimum or a stationary point of inflexion, are vertices.

Note carefully the difference between simple inflexion, vertex, and simple undulation using the curvature criteria. Note also the difference between vertex and corner; a vertex is a regular point and a corner is not.

Worked example 8.8. Show that, on circles and lines, all points are vertices.

Solution. For a line we have $\kappa = 0$ and for a circle we have that κ is constant. In each case $\kappa' = 0$. □

Worked example 8.9. Find the curvature of $y = x^4$. Deduce that the curve has a simple undulation at the origin, and, additionally, two vertices close to the origin. Show that there is a strict local maximum of the curvature at each of these two vertices.

Solution. We parametrise the curve by $x \mapsto (x, x^4) = \mathbf{r}(x)$. We have

$$\kappa = \frac{x'\,y'' - y'\,x''}{|\mathbf{r}'|^3} = \frac{12x^2}{(16x^6+1)^{3/2}}, \text{ and}$$
$$\kappa' = \frac{24x(1-56x^6)}{(16x^6+1)^{5/2}}.$$

Also at the point where $x = 0$, we have by Lemma 8.4,

$$\kappa''(0) = \left.\frac{24(1-56x^6)}{(16x^6+1)^{5/2}}\right|_{x=0} = 24.$$

Therefore $\kappa(0) = \kappa'(0) = 0$ and $\kappa''(0) \neq 0$. Thus there is a simple undulation at $x = 0$. Also there are two vertices and the values taken by x at these are the two real sixth roots of $\frac{1}{56}$, which are $\approx \pm .51$; thus the vertices are $\approx (\pm.51, .07)$. Using Lemma 8.4, we readily see that a strict local maximum of the curvature occurs at each of the vertex points, since, at the vertex points,

$$\kappa'' = \left.\frac{24x(-56.6x^5)}{(16x^6+1)^{5/2}}\right|_{x=\pm\frac{1}{\sqrt[6]{56}}} < 0.$$

$\square$

Worked example 8.10. Let $\mathbf{r}(t) = (3t^2-1,\, t\,(3t^2-1)\,)$. Show that the curve $t \mapsto \mathbf{r}(t)$ is regular, and find its curvature. Show that the curve has no inflexions and only one vertex. Find the points at which the tangents are horizontal and the points at which the tangents are vertical. Find the slopes of the tangent lines to the curve at the point $\mathbf{r} = \mathbf{0}$. Find the local maxima and minima of y regarded as a function of x. Write down the algebraic equation of the curve.

Solution. We have $\mathbf{r}' = (6t, 9t^2-1)$, $\mathbf{r}'' = (6, 18t)$, and $|\mathbf{r}'|^2 = (9t^2+1)^2$. The parametrisation is regular, since $\mathbf{r}' \neq \mathbf{0}$. The curvature is

$$\kappa = \frac{x'\,y'' - y'\,x''}{|\mathbf{r}'|^3} = \frac{6(9t^2+1)}{(9t^2+1)^3} = \frac{6}{(9t^2+1)^2}\ , \text{ and}$$

$$\kappa' = \frac{6.18t.(-2)}{(9t^2+1)^3}\,.$$

Thus there are no inflexions and a vertex only at $t=0$: at $t=0$ we have (using Lemma 8.4)

$$\kappa'(0) = 6.18.(-2) < 0,$$

and thus the curvature has a local maximum where $t=0$, that is at the point $(-1,0)$. The tangent is horizontal (i.e., $y'=0$ and $x'\neq 0$) if, and only if, $t = \pm\frac{1}{3}$, and is vertical (i.e., $x'=0$ and $y'\neq 0$) if, and only if, $t=0$. Also $\mathbf{r} = \mathbf{0}$ if, and only if, $t = \pm\frac{1}{\sqrt{3}}$. At $t = \pm\frac{1}{\sqrt{3}}$ we have

$$\frac{dy}{dx} = \frac{y'}{x'} = \pm\frac{1}{\sqrt{3}}.$$

Now

$$\frac{dy}{dx} = \frac{y'}{x'} = \frac{9t^2-1}{6t} = 0$$

if, and only if, $t = \pm\frac{1}{3}$. Also

$$\frac{d^2y}{dx^2}\,x' = \frac{d}{dt}\left(\frac{dy}{dx}\right),$$

and again using Lemma 8.4, we see that $\dfrac{d^2y}{dx^2} \neq 0$ at $t = \pm\frac{1}{3}$, and therefore the local maxima and minima of y, regarded as a function of x, are at $(-\frac{2}{3}, \pm\frac{2}{9})$. For large values of t, the curve is 'close' to the semi-cubical parabola $x^3 = 3y^2$ parametrised by $t \mapsto (3t^2, 3t^3)$. The algebraic equation of the curve is given by

$$y^2 = x^2t^2 = \frac{x^2(x+1)}{3},$$

that is by $3y^2 = x^2(x+1)$: so it is a cubic curve. □

Worked example 8.11. Show that the ellipse $\mathbf{r} = (a\cos t, b\sin t)$ has no inflexions and has precisely four vertices.

Solution. We have

$$\kappa = \frac{ab}{(a^2\sin^2 t + b^2\cos^2 t)^{\frac{3}{2}}}, \text{ and}$$

$$\kappa' = -\frac{3ab(a^2-b^2)\sin t\,\cos t}{(a^2\sin^2 t + b^2\cos^2 t)^{5/2}}.$$

Thus the ellipse has no inflexions and has precisely four vertices. The parameter values at which these vertices occur are $t = 0, \dfrac{\pi}{2}, \pi, \dfrac{3\pi}{2}$. Using Lemma 8.4, we readily see that $\kappa'' \neq 0$ at each of these vertices. Therefore the curvature has a local maximum or a local minimum at each vertex.

□

This example illustrates a general property of simple closed curves, that is that such curves have at least four vertices.

Theorem 8.12. *(Four-vertex theorem) A simple closed curve has at least four vertices; at least two correspond to local maximum values of κ and at least two correspond to local minimum values of κ.*

A *simple* curve here means a regular curve with no multiple points except the end points, that is $\mathbf{r}(t_0) = \mathbf{r}(T)$. The proof of this theorem is hard and beyond the scope of this book. The proof can be found in texts on algebraic curves, or algebraic geometry.

Exercises

8.4. Show that the curve $t \mapsto \mathbf{r}(t) = (t^2-1,\ t\,(t^2-1))$ is regular, and find its curvature. Show that $\kappa' = 0$ if, and only if, $9t^5 + 4t^3 - 1 = 0$, and deduce that the curve has no inflexions and precisely three vertices. Find the points at which the tangents are horizontal and the points at which the tangents are vertical. Find the slopes of the tangent lines to the curve at the point $\mathbf{r} = \mathbf{0}$. Find the local maxima and minima of y regarded as a function of x. Write down the algebraic equation of the curve.

$$\left[\frac{6t^2+2}{(9t^4-2t^2+1)^{\frac{3}{2}}}; 0, t^2 = \tfrac{1}{18}\left(\sqrt{52}-4\right); 0, \pm\tfrac{1}{\sqrt{3}}; \pm 1; \pm\tfrac{1}{\sqrt{3}}; y^2 = x^2(1+x)\right]$$

8.5. Find the curvature of $\left(t + \frac{1}{2}t^2, \frac{1}{2}t^2\right)$. Show that the curve has no inflexions and has precisely one vertex. Write down the algebraic equation of the curve.

8.6. Prove that the curvature of $r = 2 + \cos\theta$ is

$$\kappa = \frac{6(1+\cos\theta)}{(5+4\cos\theta)^{\frac{3}{2}}}.$$

Show that the curve has a simple undulation at $\theta = -\pi$, and find the vertices.

(*Hint:* Use Exercise 8.2.) $\left[0, \frac{2\pi}{3}, \pi, \frac{4\pi}{3}\right]$

8.3 Curvature of algebraic curves*

In general algebraic curves cannot be parametrised in a simple way, that is by using a simple formula which gives a parametrisation of the whole curve. Therefore the formulae for the curvature which we have considered previously cannot be used. However algebraic curves can be parametrised locally near a non-singular point, though not in general using a simple formula. We now use this result to give a parameter-free formula for the curvature of algebraic curves.

Given a polynomial $f(x,y)$ let $J = \begin{pmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{pmatrix}$ be the Hessian matrix.

Theorem 8.13. *The curvature of the algebraic curve $f(x,y) = 0$ at a non-singular point (x,y) on the curve is*

$$\kappa = \pm \frac{(f_y, -f_x)\, J \begin{pmatrix} f_y \\ -f_x \end{pmatrix}}{(f_x^2 + f_y^2)^{\frac{3}{2}}} = \pm \frac{f_y^2 f_{xx} - 2f_x f_y f_{xy} + f_x^2 f_{yy}}{(f_x^2 + f_y^2)^{\frac{3}{2}}}$$
$$= \pm \frac{f_y^2 f_{xx} - 2f_x f_y f_{xy} + f_x^2 f_{yy}}{|\mathbf{grad}\, f|^3},$$

where the sign $-$ is taken in case the motion along the curve is in the direction of the vector $(f_y, -f_x)$ and where the sign $+$ is taken in case the motion along the curve is in the direction of the vector $-(f_y, -f_x)$.

Proof. By Theorem 4.4, the algebraic curve can be given a regular local parametrisation near a non-singular point by $t \mapsto \mathbf{r}(t)$. Differentiating

$$f(x(t), y(t)) = 0,$$

using the chain rule, we obtain

$$\mathbf{r}' \cdot (f_x, f_y) = 0.$$

Therefore $\mathbf{r}'$ is orthogonal to (f_x, f_y), and we have $\mathbf{r}' = \lambda(f_y, -f_x)$ where $\lambda \neq 0$ is a function of t, and $(-y', x') = \lambda(f_x, f_y)$. On differentiating once more we obtain

---

* For the meaning of * see the preface.

$$\mathbf{r}'' \cdot (f_x, f_y) + (x', y')\, J \begin{pmatrix} x' \\ y' \end{pmatrix} = 0.$$

Therefore we have

$$x'\, y'' - y'\, x'' = \mathbf{r}'' \cdot (-y', x') = \lambda\, \mathbf{r}'' \cdot (f_x, f_y) = -\lambda^3 (f_y, -f_x)\, J \begin{pmatrix} f_y \\ -f_x \end{pmatrix}.$$

Since $|\mathbf{r}'| = |\lambda|\, |(f_y, -f_x)|$, the result follows. ☐

Corollary 8.14. *The algebraic curve has a point of inflexion at a non-singular point (a,b) if, and only if,*

$$f_y^2 f_{xx} - 2 f_x f_y f_{xy} + f_x^2 f_{yy}$$

is zero at (a,b) and changes sign as (x,y) moves through (a,b) along the curve.

Thus in order to find all points of inflexion of an algebraic curve we first determine the points where the curvature is zero by finding the simultaneous solutions of the equations

$$f(x,y) = 0, \text{ and}$$
$$f_y^2 f_{xx} - 2 f_x f_y f_{xy} + f_x^2 f_{yy} = 0,$$

and then find those solutions which are non-singular points. Next we determine whether the curvature changes sign as we move along the curve past the point of zero curvature. If the curvature changes sign, the point is an inflexion. If the sign of the curvature is the same immediately before and after the point, the point is an undulation. In practice we can check on the change of sign if, for example, the curve can be parametrised near the point by x, and

$$f_y^2 f_{xx} - 2 f_x f_y f_{xy} + f_x^2 f_{yy}$$

becomes a function of x after substituting from $f(x,y) = 0$. Such a local parametrisation by x exists by Theorem 4.4 provided that the tangent at the point is not parallel to the y–axis.

Worked example 8.15. Calculate the curvature of the hyperbola

$$x^2 - 3y^2 = 1$$

at $(2,1)$.

Solution. Let $f = x^2 - 3y^2 - 1$. We have $f_x = 2x$, $f_y = -6y$, $f_{xx} = 2$, $f_{xy} = 0$, and $f_{yy} = -6$. Therefore

$$\kappa = \pm \frac{(-6y)^2 . 2 + (4x)^2 . (-6)}{(4x^2 + 36y^2)^{\frac{3}{2}}} = \pm \frac{9y^2 - 12x^2}{(x^2 + 9y^2)^{\frac{3}{2}}}.$$

In particular

$$\kappa(2,1) = \pm \frac{-3}{\sqrt{13}}.$$

$\square$

Worked example 8.16. Find the points of inflexion of the acnodal cubic

$$f(x,y) = y^2 + x^2 - x^3 = 0.$$

Solution. We have $f_x = 2x - 3x^2$, $f_y = 2y$, $f_{xx} = 2 - 6x$, $f_{xy} = 0$, and $f_{yy} = 2$. Eliminating y from

$$f(x,y) = y^2 + x^2 - x^3 = 0, \text{ and}$$

$$\begin{aligned} f_y^2 f_{xx} - 2f_x f_y f_{xy} + f_x^2 f_{yy} &= 4y^2(2-6x) + 2(2x-3x^2)^2 \\ &= 8y^2 - 24xy^2 + 8x^2 - 24x^3 + 18x^4, \end{aligned}$$

we have, on the curve, that

$$f_y^2 f_{xx} - 2f_x f_y f_{xy} + f_x^2 f_{yy} = 8x^3 - 6x^4.$$

The origin is a singular point. Therefore there are precisely two points of zero curvature. They are $\left(\frac{4}{3}, \pm\frac{4}{3\sqrt{3}}\right)$. At these points

$$\mathbf{grad}\, f = (2x - 3x^2, 2y)$$

is not parallel to the y–axis, and therefore the curve can be parametrised locally by x. Also

$$8x^3 - 6x^4 = 6x^3\left(\tfrac{4}{3} - x\right)$$

changes sign as we move along the curve past each of the points and therefore the curvature also changes sign. Thus the two points are points of inflexion. (Near points of undulation the curvature does not change sign.)

$\square$

Exercises

8.7. Calculate the curvature of the ellipse $\dfrac{x^2}{4} + y^2 = 1$ at $(2,0)$ using Theorem 8.13. Compare your result with that given by using the parametrisation $t \mapsto (2\cos t, \sin t)$. $\left[\mp\frac{3}{8}\right]$

8.8. Show that the curvature of $x^4 + y^2 = 2a(x+y)$ at the origin is

$$\pm\frac{1}{2\sqrt{2a}}.$$

8.9. Show that the curvature of $y^2(a-x) = x^2(a+x)$ at $(-a,0)$ is $\pm\dfrac{4}{a}$.

8.10. Show that the curvature of $x^2y = x^2 + y^2$ at non-singular points is

$$\pm\frac{2x^3y^4(y^2-3x^2)}{(4y^6 + x^2(y^2-x^2)^2)^{\frac{3}{2}}}.$$

8.11. Show that the curvature of $x^3 - xy^2 - y = 0$ at points where y has a local maximum or minimum value is $\pm 2^{-\frac{1}{2}}3^{\frac{5}{4}}$.

8.4 Limiting curvature of algebraic curves at cusps

As a further application of curvature, we show how curvature can be used to classify the cusps of algebraic curves. We divide cusps into three types, ceratoid cusps, cusps with zero limiting curvature, and rhamphoid cusps (see Figure 8.1). Although the curvature is not defined at a cusp, in that the formula for curvature is not meaningful at a cusp, it is nevertheless useful to consider the limit of the curvature at a point as the point approaches the cusp. The following analysis of types of cusps does not extend to non-algebraic curves in general.

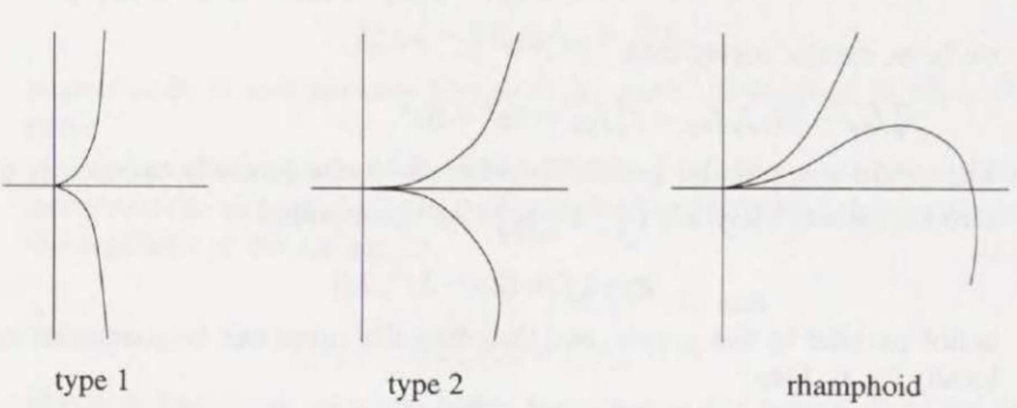

Figure 8.1 *Limiting curvature at cusps.*

8.4.1 Type 1: Ceratoid cusps (limiting radius of curvature zero).

A point $z(t_0)$ is a *ceratoid cusp* if it is a cusp for which $|\kappa(t)|$ tends to ∞ along both branches as t tends to t_0, or equivalently for which the radius of curvature tends to zero as t tends to t_0. For ceratoid cusps the curve appears to move away from the cuspidal tangent line 'fairly rapidly' as it leaves the cusp, since near the cusp the curvature is very large and therefore the curve bends sharply.

Example 8.17. For the semi-cubical parabola $x^3 = y^2$ parametrised by $t \mapsto \mathbf{r}(t) = (t^2, t^3)$ for all real t, we have

$$\begin{aligned} \mathbf{r}' &= (2t, 3t^2), \\ \mathbf{r}'' &= (2, 6t), \text{ and} \\ \mathbf{r}''' &= (0, 6). \end{aligned}$$

The curvature of the curve is given by

$$\kappa = \frac{6t^2}{(4t^2+9t^4)^{\frac{3}{2}}} = \frac{6}{|t|(4+9t^2)^{\frac{3}{2}}}.$$

In this calculation we have used the formula $\sqrt{t^2} = |t|$. We have that κ tends to ∞ as t tends to 0, and therefore the limiting radius of curvature is zero. The semi-cubical parabola has an ordinary cusp at $\mathbf{r}'(0) = \mathbf{0}$, since $\mathbf{r}''(0) = (2, 0)$ and $\mathbf{r}'''(0) = (0, 6)$ are linearly independent.

8.4.2 Type 2: limiting radius of curvature infinite

The second type of cusp is a cusp at a point $z(t_0)$ for which $\kappa(t)$ tends to 0 along both branches as t tends to t_0, or equivalently for which the limiting radius of curvature is infinite. In this case the curve appears to linger near the cuspidal tangent line as it leaves the cusp, since the curvature is close to zero near the cusp.

Example 8.18. For the curve $(2\,x)^5 = (5\,y)^2$, parametrised by

$$t \mapsto \mathbf{r}(t) = \left(\tfrac{1}{2}t^2, \tfrac{1}{5}t^5\right)$$

for all real t, we have

$$\begin{aligned} \mathbf{r}' &= (t, t^4), \\ \mathbf{r}'' &= (1, 4t^3), \text{ and} \\ \mathbf{r}''' &= (0, 12t^2). \end{aligned}$$

The curvature of the curve is given by

$$\kappa = \frac{3t^4}{(t^2 + t^8)^{\frac{3}{2}}} = \frac{3|t|}{(1 + t^6)^{\frac{3}{2}}} \; .$$

We have that κ tends to 0 as t tends to 0, and the limiting radius of curvature is infinite. This cusp is not ordinary since $\mathbf{r}'' = (1, 0)$ and $\mathbf{r}''' = (0, 0)$ are not linearly independent at it. In this calculation we have used the formula $\sqrt{t^2} = |t|$. Recall that the curvature is positive as the curve swings to the left; this is indicated in the figure.

Cusps ($\mathbf{r}'' \neq 0$), for which, in moving through the cusp, the curve crosses the cuspidal tangent line from one side to the other, are necessarily of Type 1 or of Type 2 (ie. $|\kappa| = \infty$ or $\kappa = 0$ in the limit). For Type 1 and Type 2 cusps the curvature has the same sign, close to the cusp, on each side of the cusp.

8.4.3 Type 3: Rhamphoid cusps

Rhamphoid cusps are cusps for which the curve, close to the cusp, lies on one side of the cuspidal tangent line. For such cusps the curvature has different signs, close to the cusp, on each side of the cusp. These cusps can also have limiting radii of curvature other than 0 or ∞.

Example 8.19. The curve $(y - x^2)^2 = x^5$, or $y = x^2 \pm x^{\frac{5}{2}}$, parametrised by $t \mapsto \mathbf{r}(t) = (t^2, t^4 + t^5)$ for all real t, satisfies

$$\begin{aligned} \mathbf{r}' &= (2t, 4t^3 + 5t^4), \\ \mathbf{r}'' &= (2, 12t^2 + 20t^3), \text{ and} \\ \mathbf{r}''' &= (0, 24t + 60t^2). \end{aligned}$$

The curvature of the curve is given by

$$\kappa = \frac{(16t^3 + 30t^4)}{(4t^2 + (4t^3 + 5t^4)^2)^{\frac{3}{2}}} = \frac{t}{|t|}\frac{(16 + 30t)}{(4 + (4t^2 + 5t^3)^2)^{\frac{3}{2}}} .$$

In this calculation we have used the formula $\sqrt{t^2} = |t|$. We have $|\kappa|$ tends to 2 as t tends to 0, κ tends to $+2 > 0$ as $t > 0$ tends to 0, and κ tends to $-2 < 0$ as $t < 0$ tends to 0. Also we have the slope

$$\frac{dy}{dx} = \frac{y'}{x'} = \frac{t^3(4 + 5t)}{2} .$$

This curve has a point of simple inflexion where $\kappa = 0$, that is, where $t = -\frac{8}{15}$. Also, regarded as the graph of a function of x, the curve has a horizontal tangent at $t = -\frac{4}{5}$. Again the cusp is not ordinary since $\mathbf{r}'' = (2, 0)$ and $\mathbf{r}''' = (0, 0)$ are not linearly independent at it.

Drawing cusps. Note the different shapes of these cusps depending on the different limiting radii of curvature, and on whether or not the cusp is rhamphoid.

Exercises

8.12. For each of the following curves, show that there is a cusp at the origin, and classify the cusp by finding the limiting radius of curvature and determining whether or not it is rhamphoid.

i) $(\frac{1}{2}t^2, \frac{1}{3}t^3)$, [$\infty$; type 1, ordinary]

ii) $(\frac{1}{2}t^2, \frac{1}{4}t^4 + \frac{1}{5}t^5)$,

iii) (t^2, t^5), [0; type 2]

vi) $(\frac{1}{2}t^2, \frac{1}{6}t^6 + \frac{1}{6}t^6)$.

8.13. Show that $\mathbf{r}(t) = (t^2, t^n)$ $(n \geq 7)$ has, at the origin, a type 2 cusp if n is odd.

8.14. Show that $\mathbf{r}(t) = (t^2, t^n + t^{n+1})$ $(n \geq 6)$ has, at the origin, a rhamphoid cusp if n is even.

8.15. Use L'Hôpital's theorem (see §4.7.5) to show that, at a cusp $\mathbf{r}(t_0)$ of a parametrised curve $t \mapsto \mathbf{r}(t)$,

$$\lim_{t \to t_0} \frac{x'\,y'' - y'\,x''}{(x')^2 + (y')^2} = \frac{x''\,y''' - y''\,x'''}{2(x'')^2 + 2(y'')^2} .$$

Deduce that $\lim\limits_{t \to t_0} |\kappa(t)| = \infty$ at a simple cusp.

We give here some more general exercises which can include curve sketching or plotting if desired.

8.16. Let $\mathbf{r}(t) = (\sqrt{3}(t^2 - 1),\ t\,(t^2 - 1))$. Show that the curve $t \mapsto \mathbf{r}(t)$ is regular. Prove that the curvature of $t \mapsto \mathbf{r}(t)$ at $\mathbf{r}(t)$, is given by

$$\kappa = \frac{2\sqrt{3}}{(3t^2 + 1)^2}\,,$$

and show that the curve has no inflexions and only one vertex. Show further that the curvature κ does not have an inflexion at the vertex of the curve. Find the points at which the tangents are horizontal and the points at which the tangents are vertical. Hence sketch the curve, noting the points with horizontal tangents and the points with vertical tangents, and the local maxima and minima of y. Write down the algebraic equation of the curve.

$$\left[t = 0, \kappa''(0) \neq 0 : \pm\tfrac{1}{\sqrt{3}}; 0 : 9y^2 - 3x^2 - \sqrt{3}x^3 = 0\right]$$

8.17. Show that the curve $z(t) = (\cos t, \cos t \sin t)$ is regular and has curvature

$$\kappa = \frac{\cos t\,(1 + 2\sin^2 t)}{(1 - 3\sin^2 t + 4\sin^4 t)^{\frac{3}{2}}}\,.$$

Hence find the inflexions of the curve. Show that every point of $z(t)$ belongs to the algebraic curve

$$y^2 = x^2 - x^4.$$

By writing $x^2 - x^4 = x^2(1 - x^2)$, show conversely that every point of the algebraic curve has the form $z(t)$ for some t. Show that the algebraic curve has a singular point at $(0, 0)$ and no other singular point. Sketch the curve, indicating the positions of the inflexions. $\left[n\pi + \dfrac{\pi}{2}; |x| \leq 1, \text{ thus } x = \cos t\right]$

8.18. Show that the limaçon with polar equation

$$r = \sqrt{2} + \cos\theta$$

has inflexions where $\cos\theta = -\frac{2\sqrt{2}}{3}$ and vertices where

$$\sin\theta = 0 \text{ or } \cos\theta = -\tfrac{1}{\sqrt{2}}\,.$$

Find the tangent directions at the vertices. Sketch the curve.

8.19. Show that the limaçon with polar equation

$$r = \sqrt{2} - \cos\theta$$

has inflexions where $\cos\theta = \frac{2\sqrt{2}}{3}$. By checking that the appropriate determinant is zero, or otherwise, show also that the limaçon has vertices where $\sin\theta = 0$ or $\cos\theta = \frac{1}{\sqrt{2}}$. Find the tangent directions at the vertices. Sketch the curve.

$$\left[\kappa = \frac{-3\sqrt{2}\cos\theta + 4}{(2\sqrt{2}\cos\theta - 3)^{\frac{3}{2}}} : i, i \pm 1\right]$$

8.20. State the four vertex theorem. Find the vertices of the limaçon with polar equation

$$r = 1 + 2\cos\theta\,,$$

and make a careful drawing of this curve using polar graph paper. Why does this curve not provide a counterexample to the four vertex theorem?

$$[0, \pi;\ \text{not simple}]$$

8.21. Show that the parametric curve $t \mapsto \mathbf{r}(t) = (t^2 + t^4, t^2 + t^5)$ has exactly one non-regular point and that this is a cusp but not an ordinary cusp. Write down the limiting tangent direction at the cusp. Show that $z(t)$ has zero curvature where $10t^3 + 15t - 8 = 0$. Use the intermediate value theorem and Rolle's theorem to deduce that $z(t)$ has one inflexion between $t = 0$ and $t = \frac{1}{2}$ and has no other inflexions. Show that $|\kappa(t)|$ tends to $\frac{1}{\sqrt{2}}$ as t tends to 0. Draw the curve for $-1.1 \leq t \leq 1.1$ using 5 cm as 1 unit.

$$\left[\mathbf{r}'''(0) = \mathbf{0}; (1,1);\ \text{inflexion since } \frac{d}{dt}(10t^3 + 15t - 8) \neq 0\right]$$

8.22. Show that the parametric curve $t \mapsto \mathbf{r}(t) = (\frac{1}{4}t^4, t^2 + t^3)$ has exactly one non-regular point and that this is a cusp but not an ordinary cusp. Determine the limiting tangent direction at the cusp. Find the curvature $\kappa(t)$ at a general regular point, and show that $|\kappa(t)|$ tends to $\frac{1}{2}$ at the cusp. Show that there is exactly one point P at which the curve has an inflexion. Determine the tangent direction at P. Draw the curve taking 10 cm as 1 unit. Indicate the point P and plot about 9 other points. (Use the portrait mode with the origin near the left.)

8.23. Show that the parametric curve $t \mapsto \mathbf{r}(t) = (t^2, -t^4 + t^5)$ has exactly one non-regular point and that this is a cusp but not an ordinary cusp. Determine the limiting tangent direction at the cusp. Also determine the point P on the curve, not at the cusp, at which the tangent is parallel to the x-axis. Find the curvature $\kappa(t)$ at a general regular point, and show that $|\kappa(t)|$ tends to 2 at the cusp. Show that there is exactly one point Q at which the curve has an inflexion. Draw the curve for $-1.1 \leq t \leq 1.1$ using 5 cm as 1 unit. Indicate the points P and Q and plot about 9 other points. (Use the portrait mode with the origin near the top left: first calculate $\mathbf{r}(t)$ for $t = \pm 1.1$.)

$$[\mathbf{r}'''(0) = \mathbf{0}; (1,0); \tfrac{4}{5}, \tfrac{8}{15}]$$

8.24. Show that the parametric curve $t \mapsto \mathbf{r}(t) = (t^2 - t^3, t^5)$ has exactly one non-regular point and that this is a cusp but not an ordinary cusp. Determine the cuspidal tangent line. Also determine the point P on the curve, not at the cusp, at which the tangent is parallel to the y-axis. Find the curvature $\kappa(t)$ at a general regular point. Show that there is exactly one point Q at which the curve has an inflexion. Draw the curve for $-1.1 \leq t \leq 1.1$ using 5 cm as 1 unit. Indicate the points P and Q and plot about 9 other points. (Use the landscape mode with the origin at the centre.)

' . !/ & 2!

9

Circle of curvature

We consider now the radius of curvature, the circle of curvature, and the osculating circle at regular points of parametrised curves. The *radius of curvature* is the absolute value of the reciprocal of the curvature. The *circle of curvature* at a point on the curve is the circle passing through the point and having the same curvature and same tangent direction as the curve at the point. The radius of the circle of curvature is equal to the radius of curvature. The *osculating circle* at a point on the curve is the unique circle having $\geq$ 3–point contact with the curve at the point. We show that the circle of curvature coincides with the osculating circle.

9.1 Centre of curvature and circle of curvature

In general there is a unique circle having the same curvature and same tangent direction as the curve at a common point which is regular for the curve. The centre of this circle is called the centre of curvature at the point and the radius of the circle is called the radius of curvature at the point. The centre of curvature can also be described as the limiting position of the point of intersection of 'consecutive' normals. We now define the centre of curvature and the circle of curvature.

Definition 9.1. The *standard parametrisation* of a circle having curvature $\kappa \quad (\kappa \neq 0)$, and centre z_0, is

$$\boxed{z(t) = z_0 + \frac{1}{\kappa} e^{i\kappa t} \quad \text{or} \quad \mathbf{r}(t) = \mathbf{r}_0 + \frac{1}{\kappa}(\cos \kappa t, \sin \kappa t)} \quad \left(0 \leq t \leq \frac{2\pi}{|\kappa|}\right).$$

Notice that the standard parametrisation of a circle is a unit speed (arc-length) parametrisation (i.e., $|z'| = 1$), and that t is not the argument (polar angle) of $z - z_0$. This circle is positively oriented (counter-clockwise) if $\kappa > 0$, and is negatively oriented (clockwise) if $\kappa < 0$.

Definition 9.2. At a regular point $\mathbf{r}(t_0)$, respectively $z(t_0)$, of the curve $t \mapsto \mathbf{r}(t)$, respectively $t \mapsto z(t)$, the *centre of curvature* is

$$\boxed{\begin{aligned} &z_* = z_*(t_0) = z(t_0) + \frac{1}{\kappa}\frac{iz'(t_0)}{|z'(t_0)|}, \text{ or, equivalently,} \\ &\mathbf{r}_* = \mathbf{r}_*(t_0) = \mathbf{r}(t_0) + \frac{1}{\kappa|\mathbf{r}'(t_0)|}(-y'(t_0), x'(t_0)) . \end{aligned}}$$

Thus the centre of curvature lies along the normal to the curve at $z(t_0)$, at an 'oriented distance' $\frac{1}{\kappa}$ to the 'left' of the curve. This means that the centre of curvature lies on the 'left' if $\kappa > 0$, and lies on the 'right' if $\kappa < 0$. 'Left' and 'right' here are determined by the direction of the parametrisation. We shall use the notations $t \mapsto \mathbf{r}(t)$ and $t \mapsto z(t)$ interchangeably for a curve. Many statements, proofs, and solutions are simpler using the curve $t \mapsto z(t)$ in the Argand diagram.

Definition 9.3. At a regular point $\mathbf{r}(t_0)$ of the curve $t \mapsto \mathbf{r}(t)$ the *radius of curvature* is $\frac{1}{|\kappa|}$ for $\kappa \neq 0$, and is ∞ for $\kappa = 0$.

The centre of curvature (see Figure 9.1) can be considered as the limiting position of the point of intersection of 'consecutive' normals to the curve. Notice that the radius of curvature of a circle of radius R, whatever its orientation, is R.

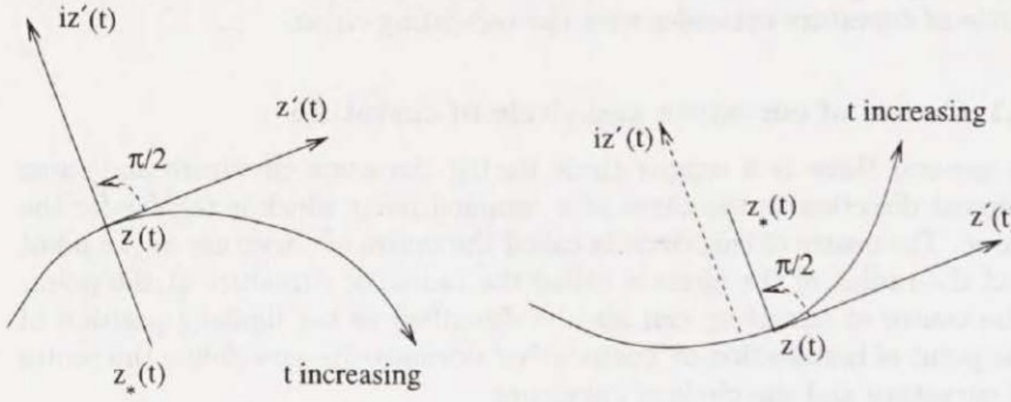

negative curvature (κ < 0) positive curvature (κ > 0)

Figure 9.1 *Centre of curvature.*

Definition 9.4. The *circle of curvature* at a regular point $\mathbf{r}(t_0)$, at which the curvature is non-zero, is the circle centre $\mathbf{r}_*(t_0)$ and radius $\frac{1}{|\kappa|}$.

The equation of the circle of curvature in standard complex parametric form is

$$\boxed{Z(s) = z_*(t_0) + \frac{1}{\kappa}e^{i\kappa s}} \qquad \left(0 \le s \le \frac{2\pi}{|\kappa|}\right),$$

and the equation in standard real parametric form is

$$\boxed{\mathbf{R}(s) = \mathbf{r}_*(t_0) + \frac{1}{\kappa}(\cos\kappa s, \sin\kappa s)\,.}$$

The algebraic equation is

$$\boxed{(X - x_*)^2 + (Y - y_*)^2 = \frac{1}{\kappa^2}\,.}$$

Other forms of the equation are

$$|\, Z - z_* \,| = \frac{1}{|\kappa|} \quad \text{or} \quad (Z - z_*)(\overline{Z} - \overline{z_*}) = \frac{1}{\kappa^2}\,.$$

Here we use $Z = X + iY$ and $\mathbf{R} = (X, Y)$ rather than $z = x + iy$ and $\mathbf{r} = (x, y)$ for the variable.

Theorem 9.5. *Let $\mathbf{r}(t_0)$ be a regular point of the curve $t \mapsto \mathbf{r}(t)$. Then there is a unique standardly parametrised circle, namely the circle of curvature, which passes through $\mathbf{r}(t_0)$, and which at $\mathbf{r}(t_0)$ has the same tangent line, the same direction of tangent vector, and the same curvature as the curve. Furthermore the sign of the orientation of this circle (counterclockwise is positive) coincides with the sign of the curvature.*

Proof. From the definition of the centre of curvature, we deduce that

$$|z_* - z(t_0)| = \frac{1}{|\kappa|}.$$

Thus the circle of curvature

$$|Z - z_*| = \frac{1}{|\kappa|}$$

passes through $z(t_0)$. Let

$$Z(s) = a + \frac{1}{\kappa}e^{i\kappa s}$$

be a standardly parametrised circle which passes through $z(t_0)$, where κ is the curvature of the curve at $t = t_0$ (see Definition 9.1). Differentiating this equation gives $Z' = ie^{i\kappa s}$. This circle meets the curve at the point where $z(t_0) = Z(s)$. At this point we have

$$Z(s) - a = \frac{1}{\kappa}e^{i\kappa s} = -\frac{1}{\kappa}iZ'(s)\,, \quad \text{and,}$$

$$z(t_0) - z_*(t_0) = -\frac{1}{\kappa}i\frac{z'(t_0)}{|z'(t_0)|}\,.$$

Therefore $a = z_*(t_0)$ if, and only if,

$$Z'(s) = \frac{z'(t_0)}{|z'(t_0)|},$$

that is if, and only if, the tangent vector $z'(t_0)$ to the curve and the tangent vector $Z'(s)$ to the circle are parallel and have the same direction (recall that $s \mapsto Z(s)$ is a unit speed curve). Since $a = z(t_0)$ for the circle of curvature, the existence, and uniqueness of the required circle follows. ☐

The centre of curvature and the circle of curvature at a point lie on the 'concave side' of the curve at that point. Notice that the 'concave side' of the curve does not depend on the orientation of the curve, that is the direction of the parametrisation. The circle of curvature coincides with the osculating circle, and is often called the osculating circle. The osculating circle is the unique circle having $\geq$ 3–point contact with the curve at $\mathbf{r}(t_0)$ (see §9.2).

Mnemonic

> The circle of curvature has the same direction as the curve at their point of tangency.

Worked example 9.6. Find the centre of curvature and the circle of curvature of the curve $\mathbf{r} = (t^3 - t, t^2)$ at the origin.

Solution. The origin is given by the parameter value $t = 0$. By Example 7.6, we have that, at $t = 0$, $\kappa = -2$, and $\mathbf{r}' = (-1, 0)$. Thus the centre of curvature is

$$\mathbf{r}_* = \mathbf{r} + \frac{1}{\kappa|\mathbf{r}'|}(-y', x') = \mathbf{0} - \frac{1}{2}(0, -1) = \frac{1}{2}(0, 1).$$

The circle of curvature is

$$(X - 0)^2 + (Y - \frac{1}{2})^2 = \left(\frac{1}{2}\right)^2,$$

that is

$$X^2 + Y^2 - Y = 0.$$

☐

Worked example 9.7. Show that the circle of curvature of the curve

$$y = 2x + 3x^2 - 2xy + y^2 + 2y^3$$

at the origin is $3x^2 + 3y^2 - 5y + 10x = 0$.

Solution. Let $f = -y + 2x + 3x^2 - 2xy + y^2 + 2y^3$. We have

$$\frac{\partial f}{\partial y}(0, 0) = 1 \neq 0.$$

Therefore by Theorem 4.22, we can assume that y is given as a function of x, $y = y(x)$, near the origin. Differentiating the defining equation implicitly with respect to x we have

$$\begin{aligned} y' &= 2 + 6x - 2xy' - 2y + 2yy' + 6y^2y', \text{ and} \\ y'' &= 6 - 4y' - 2xy'' + 2(y')^2 + 2yy'' + 6y^2y'' + 12yy'y''. \end{aligned}$$

Putting $x = y = 0$ in the first gives $y' = 2$. Substituting into the second then gives $y'' = 6$. Therefore

$$\kappa = \frac{6}{(1+4)^{\frac{3}{2}}} = \frac{6\sqrt{5}}{25},$$

and $(x_*, y_*) = \frac{5}{6}(-2, 1)$. The equation of the circle of curvature is therefore

$$(x + \tfrac{5}{3})^2 + (y - \tfrac{5}{6})^2 = \tfrac{125}{36},$$

which simplifies to give

$$3x^2 + 3y^2 - 5y + 10x = 0.$$

□

Exercises

9.1. Find the circle of curvature at $t = 0$ of the following curves.

a) the crunodal cubic $\mathbf{r}(t) = (t^2 - 1, t(t^2 - 1))$.

$$\left[\left(x + \tfrac{1}{2}\right)^2 + y^2 = \tfrac{1}{4}\right]$$

b) the acnodal cubic $\mathbf{r}(t) = (t^2 + 1, t(t^2 + 1))$.

9.2. Find the coordinates of the centre of curvature of the curve $y = \cosh x$ at the point $(0, 1)$. Find the circle of curvature.
(*Hint:* Recall that the derivatives of sinh and cosh are cosh and sinh respectively.)

9.3. Find a point on the curve $y = x^2 + x - 1$ for which the centre of curvature lies on the y–axis. $[x = -1]$

9.4. Write down the formula for the centre of curvature for a curve $y = f(x)$ parametrised by its x–coordinate.

$$\left[\mathbf{r}(x) = (x, f(x)) + \frac{1 + (f'(x))^2}{f''(x)}(-f'(x), 1)\right]$$

9.5. A curve is parametrised by

$$t \mapsto \mathbf{r}(t) = \left(\frac{t^2}{1+t^2}, \frac{t^3}{1+t^2}\right).$$

Prove that the centre of curvature is $(-t^2 - \frac{1}{6}t^4, \frac{4}{3}t)$.

9.6. Show that the circles of curvature of the parabola $y^2 = 4ax$ for the ends of the latus rectum are $x^2 + y^2 - 10ax \pm 4ay - 3a^2 = 0$, and that they cut the curve again in the points $(9a, \mp 6a)$.

9.7. Prove that the curvature of the curve $y = \frac{1}{4}x^2 - \frac{1}{2}\log x$ $(x > 0)$ is

$$\frac{4x}{(1+x^2)^2}\,.$$

Find the point at which the curve is parallel to the x–axis, and prove that the circle of curvature at this point touches the y–axis.

$[(1, \frac{3}{4})]$

9.8. For the equi-angular spiral $r = ae^{m\theta}$, show that the centre of curvature is at the point where the line through the pole $r = 0$ and orthogonal to the radius vector meets the normal.

9.9. For what value of λ does the parabola $y^2 = 4\lambda x$ have the same circle of curvature as the ellipse

$$\left(\frac{x-a}{a}\right)^2 + \frac{y^2}{b^2} = 1$$

at the origin? $\left[\dfrac{b^2}{16a}\right]$

9.10. Show that the circle of curvature at the origin for the curve

$$x + y = ax^2 + by^2 + cx^3$$

is

$$(a+b)(x^2+y^2) = 2x + 2y.$$

(*Hint:* Use Theorem 8.13. Recall that

$$\mathbf{r}_* = \mathbf{r} + \frac{1}{\kappa}\mathbf{n},$$

where $\mathbf{n}$ is the unit normal vector appropriate to the choice of the sign of the curvature. Alternatively use the method of Exercise 9.7.)

9.11. Calculate the curvature of the ellipse $\dfrac{x^2}{a^2} + \dfrac{y^2}{b^2} = 1$ at the point $(a\cot t, b\sin t)$. On an ellipse with $a = 8$ cm and $b = 4$ cm, draw accurately the circle of curvature at the point $\left(\dfrac{\sqrt{3}a}{2}, \dfrac{b}{2}\right)$. On a separate copy of the ellipse draw the circles of curvature at $(a, 0)$ and at $(0, b)$. Describe the relation of the points $z(t)$ of a curve to a circle tangent to the curve at $z(t_0)$ for t sufficiently close to t_0, where the circle is not the circle of curvature at $z(t_0)$. Illustrate this by drawing two further circles tangent to the ellipse at $\left(-\dfrac{\sqrt{3}a}{2}, \dfrac{b}{2}\right)$ and noting the relevant behaviour on your diagram.

9.12. Let $y = f(x)$ be a curve whose centres of curvature all lie in the x–axis. Prove that $yy'' + (1 + (y')^2) = 0$. Hence show, by integrating this equation, that the points on the curve must satisfy an equation of the form $(x-a)^2 + y^2 = c^2$.

9.13. Show that the centre of curvature at a non-singular point (a, b) of an algebraic curve $f(x, y) = 0$ is

$$\mathbf{r}_* = (a, b) - \frac{|\mathbf{grad}\, f|^2}{f_y^2 f_{xx} - 2f_x f_y f_{xy} + f_x^2 f_{yy}}\, \mathbf{grad}\, f\ ,$$

where all partial derivatives are evaluated at (a, b).

9.2 Contact between curves and circles

We considered contact between an algebraic curve and a parametrised curve in Chapter 5. We consider here the special case where the algebraic curve is a circle that intersects a given parametric curve. We describe the significance of 1–point contact, 2–point contact and 3–point contact in this case. In general there is a unique circle having ≥ 3–point contact with a curve at a regular point. This circle is called the *osculating circle*. We show also that the osculating circle coincides with the circle of curvature at the point. This gives us an alternative way of defining the circle of curvature.

Let

$$|\mathbf{r} - (a, b)|^2 = k^2 \quad (k > 0)$$

be a fixed circle and let

$$t \mapsto \mathbf{r}(t)$$

be a parametrised curve. Define the function γ by

$$\begin{aligned}\gamma(t) = |\mathbf{r}(t) - (a, b)|^2 - k^2 &= (\mathbf{r}(t) - (a, b)) \,.\, (\mathbf{r}(t) - (a, b)) - k^2\\ &= (x(t) - a)^2 + (y(t) - b)^2 - k^2 .\end{aligned}$$

The two curves intersect where $\gamma(t) = 0$. We assume that $t \mapsto \mathbf{r}(t)$ is smooth (or, more generally, suitably differentiable). In the special case where $x(t)$ and $y(t)$ are polynomials, $\gamma(t)$ is a polynomial in t.

We now consider a number of special cases.

9.2.1 At least 1–point contact: $\gamma(t_0) = 0$

This means that the curve intersects the circle at $\mathbf{r}(t_0)$. In the case of exactly 1–point contact, the circle and the curve will not have the same tangent line at the point of intersection.

9.2.2 At least 2–point contact: $\gamma(t_0) = \gamma'(t_0) = 0$

We assume that t_0 is a regular value. The second condition is

$$\gamma'(t_0) = 2\mathbf{r}' \,.\, (\mathbf{r} - (a, b)) = 0\,. \tag{9.1}$$

Thus the curve intersects the circle at $\mathbf{r}(t_0)$ and the tangent to the curve there is orthogonal to that radius of the circle which passes through $\mathbf{r}(t_0)$,

9.13. Show that the centre of curvature at a non-singular point (a, b) of an algebraic curve $f(x, y) = 0$ is

$$\mathbf{r}_* = (a, b) - \frac{|\mathbf{grad}\, f|^2}{f_y^2 f_{xx} - 2f_x f_y f_{xy} + f_x^2 f_{yy}}\, \mathbf{grad}\, f\ ,$$

where all partial derivatives are evaluated at (a, b).

9.2 Contact between curves and circles

We considered contact between an algebraic curve and a parametrised curve in Chapter 5. We consider here the special case where the algebraic curve is a circle that intersects a given parametric curve. We describe the significance of 1–point contact, 2–point contact and 3–point contact in this case. In general there is a unique circle having $\geq$ 3–point contact with a curve at a regular point. This circle is called the *osculating circle*. We show also that the osculating circle coincides with the circle of curvature at the point. This gives us an alternative way of defining the circle of curvature.

Let

$$|\mathbf{r} - (a, b)|^2 = k^2 \quad (k > 0)$$

be a fixed circle and let

$$t \mapsto \mathbf{r}(t)$$

be a parametrised curve. Define the function γ by

$$\begin{aligned}\gamma(t) = |\mathbf{r}(t) - (a, b)|^2 - k^2 &= (\mathbf{r}(t) - (a, b)) \,.\, (\mathbf{r}(t) - (a, b)) - k^2 \\ &= (x(t) - a)^2 + (y(t) - b)^2 - k^2.\end{aligned}$$

The two curves intersect where $\gamma(t) = 0$. We assume that $t \mapsto \mathbf{r}(t)$ is smooth (or, more generally, suitably differentiable). In the special case where $x(t)$ and $y(t)$ are polynomials, $\gamma(t)$ is a polynomial in t.

We now consider a number of special cases.

9.2.1 At least 1–point contact: $\gamma(t_0) = 0$

This means that the curve intersects the circle at $\mathbf{r}(t_0)$. In the case of exactly 1–point contact, the circle and the curve will not have the same tangent line at the point of intersection.

9.2.2 At least 2–point contact: $\gamma(t_0) = \gamma'(t_0) = 0$

We assume that t_0 is a regular value. The second condition is

$$\gamma'(t_0) = 2\mathbf{r}' \,.\, (\mathbf{r} - (a, b)) = 0\,. \tag{9.1}$$

Thus the curve intersects the circle at $\mathbf{r}(t_0)$ and the tangent to the curve there is orthogonal to that radius of the circle which passes through $\mathbf{r}(t_0)$,

Ω in the family $\mathbf{F}$. For example, if $\mathbf{F}$ is the family of all lines and P is a regular/non-singular point, the osculating line at P is the tangent line to Γ at P; and if $\mathbf{F}$ is the family of all circles and P is a regular/non-singular point, the osculating circle at P is the circle of curvature of Γ at P. In each of these two cases there is a unique osculating curve.

Exercises

9.14. Show that the curve $t \mapsto \mathbf{r}(t)$ has $\geq$ 4-point contact with its osculating circle at a regular point $\mathbf{r}(t)$ if, and only if,

$$\frac{x'y'' - y'x''}{(x')^2 + (y')^2}\,(3\mathbf{r}'' \cdot \mathbf{r}') = \mathbf{r}''' \cdot (-y', x').$$

9.15. Find the points on the following curves at which the osculating circle has $\geq$ 4-point contact:

a) the ellipse $\mathbf{r}(\varphi) = (a\cos\varphi, b\sin\varphi)$ $(a \neq b)$, $[0, \tfrac{1}{2}\pi, \pi, \tfrac{3}{2}\pi]$

b) the cycloid $\mathbf{r}(\varphi) = (a\varphi - b\sin\varphi, a - b\cos\varphi)$ $(a \neq b)$.

(*Hint:* Use Exercise 9.14.)

9.16. Show that, if two curves have $\geq$ 3-point contact at a common point, then they have the same osculating circle.

(*Hint:* Use Exercise 5.6.)

'   & $!/   =   2   2

16

Projective curves

Algebraic curves in the plane are often 'incomplete', in that certain 'points at infinity' are missing from them. We show here how plane algebraic curves can be extended to projective curves in the projective plane; any additional points obtained in the extension are the points at infinity for the original plane curve. The plane curve becomes an affine view of the projective curve. The asymptotes to the plane curve are tangents to the projective curve at the points at infinity. Different affine views of the projective curve can give further information about the original plane curve. A projective curve has many different affine views; for example, ellipses, hyperbolae, and parabolae are all affine views of the same projective curve. We give a classical definition of elements of a projective space as lines through the origin of the appropriate coordinate space. We show that points of the projective plane have a 'coordinate' representation, and use coordinates to obtain subsequent results and for solving examples. The asymptotes of the plane curve can be determined by studying the projective curve. Another application is to determine the boundedness of the plane curve.

16.1 The projective line

The projective curves we consider lie in the projective plane. The projective plane is given by adding certain points at infinity to the plane. We first consider the projective line, which is given by adding a point at infinity to the line.

Definition 16.1. The *projective line* $\mathbf{RP}^1$ is the set of lines in $\mathbf{R}^2$ which pass through the origin $(0, 0)$. A *point* in $\mathbf{RP}^1$ is a line through the origin of $\mathbf{R}^2$.

We now show that the projective line identifies with (is bijective with) any line in $\mathbf{R}^2$, not through the origin, together with the point at infinity for that line. Consider the line $\Lambda : ax + by = d$ in $\mathbf{R}^2$ which does not pass through the origin, that is $d \neq 0$; for any line we have $(a, b) \neq \mathbf{0}$. Then any point P on Λ determines a unique element L of $\mathbf{RP}^1$, where L is the line through P and $(0,0)$. Conversely any element L of $\mathbf{RP}^1$, other than the line $\Lambda_0 : ax+by=0$, determines a point P of Λ. The line $\Lambda_0 : ax+by=0$ is parallel to Λ. We thus have a bijection

$$\mathbf{RP}^1 \leftrightarrow \Lambda \ + \ \{\text{one extra point}\},$$

between $\mathbf{RP}^1$ and the union of the line Λ, which does not pass through the origin, and an extra point. The extra point corresponds to the line Λ_0 under the bijection. The extra point is called the *point at infinity* of Λ (see Figure 16.1). The line Λ is an *affine line*; it lies in the plane $\mathbf{R}^2$. Two lines $\Lambda : ax+by=d$ and $\Omega : px+qy=r$, which do not pass through the origin, will have different points at infinity (that is $ax+by=0$ and $px+qy=0$ are different lines) if they are not parallel, and will have the same point at infinity (that is $ax+by=0$ and $px+qy=0$ are the same line) if they are parallel.

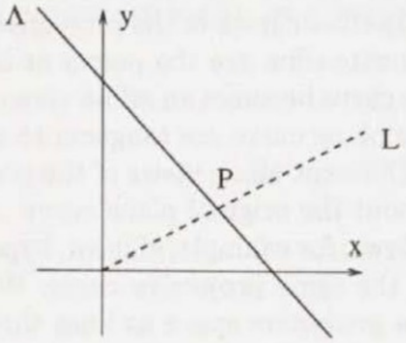

Figure 16.1 *The point at infinity of a line.*

Remark 16.2. Each element of $\mathbf{RP}^1$ determines two (opposite) points on the unit circle $x^2+y^2=1$. Indeed $\mathbf{RP}^1$ is bijective with the set[†] obtained from the upper semi-circle $\{x^2+y^2=1, y \geq 0\}$ after 'gluing together' the points $(1,0)$ and $(-1,0)$. See Figure 16.2.

16.2 The projective plane

The definition given above for the 'one-dimensional' projective line can be used as a model for defining projective spaces of 'higher dimensions'. We consider here the projective plane, which is two-dimensional projective space.

Definition 16.3. The *projective plane* $\mathbf{RP}^2$ is the set of lines in $\mathbf{R}^3$ which

---

[†] The projective line is topologically a circle, that is its shape can be deformed or stretched out to form a circle.

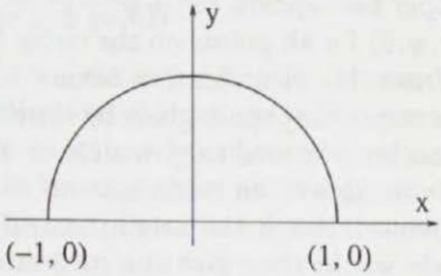

Figure 16.2 *The projective line as a circle.*

pass through the origin $(0,0,0)$. A point of $\mathbf{R}\mathrm{P}^2$ is a line passing through the origin of $\mathbf{R}^3$.

We denote by $(x:y:z)$ the element of $\mathbf{R}\mathrm{P}^2$ determined by the unique line through (x,y,z) and $(0,0,0)$ where $(x,y,z) \neq 0$. Note that

$$(\lambda x:\lambda y:\lambda z) = (x:y:z)$$

for $(x,y,z) \neq \mathbf{0}$ and $\lambda \neq 0$. The algebraic equation of the line in $\mathbf{R}^3$ corresponding to the point $(a:b:c)$ of $\mathbf{R}\mathrm{P}^2$ is

$$\frac{x}{a} = \frac{y}{b} = \frac{z}{c}.$$

This form of the equation allows up to two of a, b, c to be zero; for example if $a = 0$ the line is given by $x = 0$ and $cy = bz$. The parametric equation of the line is $(x,y,z) = \lambda(a,b,c)$. There is no point of $\mathbf{R}\mathrm{P}^2$ which corresponds to the symbol $(0:0:0)$.

Mnemonic

The projective plane is the set of points $(x:y:z)$ with $(\lambda x:\lambda y:\lambda z) = (x:y:z)$ and $(x,y,z) \neq (0,0,0)$

The coordinates x, y, z are called *homogeneous coordinates.* The numbers x, y, z are not themselves fixed for a point of the projective plane. The ratios between them are fixed for a given point.

16.2.1 Geometrical interpretation in terms of the unit sphere

Each element of $\mathbf{R}\mathrm{P}^2$, that is each line through the origin, meets the unit sphere $x^2 + y^2 + z^2 = 1$ in $\mathbf{R}^3$ in two (opposite) points. Indeed $\mathbf{R}\mathrm{P}^2$ can be considered as the space obtained from the unit sphere by identifying such pairs (x,y,z) and $-(x,y,z)$ of opposite points to give the point

$$(x:y:z).$$

(Formally $\mathbf{R}\mathrm{P}^2$ is the set of equivalence classes under the equivalence relation $(x,y,z) \sim -(x,y,z)$.) Similarly $\mathbf{R}\mathrm{P}^2$ can be considered as the space obtained from the upper hemisphere $\{x^2+y^2+z^2=1, z \geq 0\}$ by identifying $(x,y,0)$ with $-(x,y,0)$ for all points on the circle $\{x^2+y^2=1,\ z=0\}$. The space obtained from this identification cannot be deformed to give a sphere[†], unlike the corresponding space given by identifying opposite points of the circle, which can be deformed to give a circle. Rather than choosing the upper hemisphere as above, we could instead choose the hemisphere on one side of any plane through the origin; identifying opposite points on the boundary circle would then give the projective plane. A different choice of axes for $\mathbf{R}^3$ will result in different homogeneous coordinates for the projective plane.

16.2.2 Points at infinity for an affine plane in $\mathbf{R}^3$

Consider now a plane $\Pi : ax+by+cz=d$ which does not pass through the origin of $\mathbf{R}^3$ (that is $d \neq 0$). For any plane we have $(a,b,c) \neq \mathbf{0}$. Each element P of Π determines a unique element L of $\mathbf{RP}^2$ (L is the line through $(0,0,0)$ and P). Conversely each element L of $\mathbf{RP}^2$, other than those elements representing lines in the plane $\Pi_0 : ax+by+cz=0$, determines a unique element P of Π. Under this bijection, the affine plane Π can be regarded as a subset of $\mathbf{RP}^2$. The points of $\mathbf{RP}^2$ not lying in this subset are the *points at infinity* for the affine plane Π. This set of points at infinity for the affine plane $\Pi : ax+by+cz=d$ is the projective line obtained from the great circle[‡]

$$\{x^2+y^2+z^2=1,\ ax+by+cz=0\}$$

by identifying opposite points. We thus have a bijection

$$\boxed{\mathbf{RP}^2 \leftrightarrow \Pi + \text{ the projective line of points at infinity}}$$

between $\mathbf{RP}^2$ and the union of Π, which is a plane not through the origin, and the projective line of points of Π at infinity.

Mnemonic

$$\boxed{\text{The set of points at infinity for an affine plane is a projective line.}}$$

[†] Topologically, the projective plane is a non-orientable closed surface and therefore cannot be represented in three-dimensional Euclidean space without self-intersections; models can be constructed with self-intersections. In four-dimensional Euclidean space, models of the projective plane can be defined without such self-intersections.

[‡] A great circle on a sphere is the circle of intersection of the sphere and any plane through its centre.

16.2.3 Duality of lines and points

In the plane $\mathbf{R}^2$, any two distinct points determine a unique line which passes through them. However it is not generally true that any two distinct lines determine a unique common point, since parallel lines in general have no point in common. In the projective plane any two distinct lines do have a unique common point, and therefore a duality exists in the 'geometry' of the projective plane between lines and points.

Theorem 16.4. *In the projective plane*

i) two distinct lines meet in precisely one point, and

ii) there is a unique line through any two distinct points.

Proof. We give a geometric proof. Projective lines are obtained from great circles on the sphere by identifying opposite points. Two distinct great circles intersect at two opposite points, which become identified to one point of the projective plane. Two distinct points of the sphere, which are not opposite, determine a unique great circle which passes through both, and this circle becomes identified as a projective line. The result follows. ☐

16.2.4 Diagrammatic representation of the projective plane

It will sometimes be convenient to represent the projective plane diagrammatically by a portion of the plane. We represent the axes $x = 0$, $y = 0$ and $z = 0$ by 'lines' in the plane (see Figure 16.3).

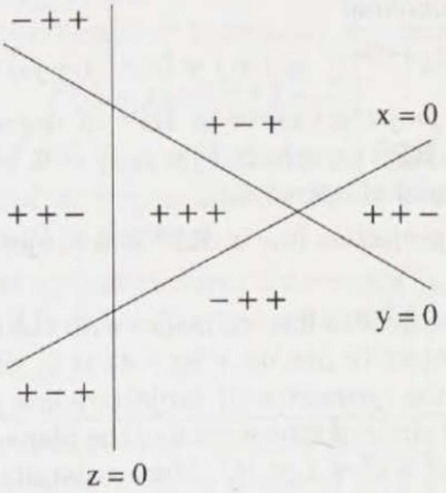

Figure 16.3 *The axes of the projective plane.*

The notation '$- + +$' for example means that the points in this region can be written $(x:y:z)$ where $x < 0$ and $y, z > 0$. The axes divide $\mathbf{RP}^2$ into four such regions. Notice that the regions determined by '$- + +$' and by '+ − −' are identical for example. The axes come from great circles, on the sphere $x^2 + y^2 + z^2 = 1$ in $\mathbf{R}^3$, given by intersection with the planes $x = 0$, $y = 0$ and $z = 0$: these great circles bound the 'first octant'

$$\{x \geq 0, y \geq 0, z \geq 0\}$$

of the sphere. The diagram should be thought of as lying within a closed curve, where 'roughly opposite' points of the closed curve are identified to give $\mathbf{RP}^2$. Care should be taken in drawing conclusions from this diagrammatic representation, particularly those relating to lines. For example there is no sense of 'parallel lines' in the projective plane; indeed, any pair of lines meet. In the affine plane given by $z = 1$, parallel lines meet on the line at infinity $z = 0$ in the projective plane. Notice in the diagram that crossing over the x–axis for example changes the sign of the x–coordinate.

16.3 Projective curves

Recall that a *homogeneous function of degree n* is a function $F : \mathbf{R}^3 \to \mathbf{R}$ which satisfies

$$F(\lambda x, \lambda y, \lambda z) = \lambda^n F(x, y, z)$$

for each real number λ and for each (x, y, z). We are especially interested in homogeneous polynomials of degree n. Such a homogeneous polynomial will have the form

$$F(x, y, z) = \sum a_{ij} x^i y^j z^{n-(i+j)}.$$

The *degree* of each monomial

$$a_{ij} x^i y^j z^{n-(i+j)} \text{ is } i + j + (n - (i + j)) = n.$$

Definition 16.5. A *projective curve* in $\mathbf{RP}^2$ of degree $n \geq 1$ is the set of points $(x : y : z)$ in $\mathbf{RP}^2$ for which $F(x, y, z) = 0$, where $F(x, y, z)$ is a homogeneous polynomial of degree n.

Definition 16.6. A *projective line* in $\mathbf{RP}^2$ is a projective curve of degree one.

This definition of 'projective line' coincides with the one defined geometrically above. The projective line $ax + by + cz = 0$, where $(x : y : z)$ lies in $\mathbf{RP}^2$, corresponds to the (geometrical) projective line given by identifying opposite points on the circle of intersection of the plane $ax + by + cz = 0$ and the unit sphere $x^2 + y^2 + z^2 = 1$ in $\mathbf{R}^3$. More generally the set of points in $\mathbf{R}^3$ satisfying $F(x, y, z) = 0$, for a homogeneous polynomial F, form a surface in $\mathbf{R}^3$. This surface is the union of the lines in a pencil of lines through the origin. Each of the lines in this pencil determines a point of $\mathbf{RP}^2$, and the set of such points in $\mathbf{RP}^2$ is the projective curve $F(x, y, z) = 0$ where $(x : y : z)$ lies in $\mathbf{RP}^2$.

Note the special case above of the projective line $ax + by + cz = 0$ of points at infinity for the affine plane $ax + by + cz = d$, considered as a subset of $\mathbf{RP}^2$.

In general a line meets a curve of degree $n \geq 2$ in the plane in $\leq n$ points. We now prove that a similar result holds in the projective plane.

Theorem 16.7. *A projective line meets a projective curve of degree n, which does not contain it, in at most n distinct points.*

Proof. We give a proof in the case where $z = 0$ is the line; the proof in the general case then follows by choosing a rotation of $\mathbf{R}^3$ which maps the plane determined by the given line to the plane $z = 0$. Let the projective curve be $F(x, y, z) = 0$. Now $F(x, y, 0)$ is the sum of the terms of $f(x, y) = 0$ of highest degree. Therefore $F(x, y, 0)$ is not identically zero. Thus $F(x, y, 0)$ is a homogeneous polynomial of degree n which factorises as

$$Ky^m \prod_{i=1}^{n-m} (x - c_i y)$$

where some c_i may not be real and some real c_i may be repeated (we are interested here only in real c_i). The only possible points of intersection are therefore the points $(c_i\!:\!1\!:\!0)$ and $(1\!:\!0\!:\!0)$. The result follows. $\square$

16.4 The projective curve determined by a plane curve

Given a plane algebraic curve of degree n with equation $f(x, y) = 0$, we define a homogeneous polynomial $F(x, y, z)$ of degree n which determines a projective curve $F(x, y, z) = 0$.

The homogeneous polynomial F is formally determined by

$$F(x, y, z) = z^n f\left(\frac{x}{z}, \frac{y}{z}\right).$$

The easiest way to find $F(x, y, z)$ is to determine the degree n of f, then to multiply each monomial in f by a suitable power of z in order to obtain, in each case, a monomial of degree n.

Definition 16.8. The projective curve *determined by* or *associated to* the plane curve $f(x, y) = 0$ in $\mathbf{R}^2$ is $F(x, y, z) = 0$ in $\mathbf{RP}^2$.

Mnemonic

A projective curve has elements $(x\!:\!y\!:\!z)$ in $\mathbf{RP}^2$.
Affine curves in $\mathbf{R}^3$ have elements (x, y, z) in $\mathbf{R}^3$.
A plane curve has elements (x, y) in $\mathbf{R}^2$.

Worked example 16.9. Find the equation of the projective curve determined by the plane curve

$$y^2 = x^3 + x.$$

Solution. The plane curve has degree 3, therefore we must multiply each monomial by a suitable power of z to give a homogeneous polynomial of degree 3. The projective curve therefore is $y^2z = x^3 + xz^2$. □

16.5 Affine views of a projective curve

Let the projective curve Γ in $\mathbf{R}\mathrm{P}^2$ be given by $F(x, y, z) = 0$; thus, the points of Γ are

$$\{(x{:}y{:}z) : F(x, y, z) = 0\}.$$

We consider the affine plane in $\mathbf{R}^3$ given by $\{(x, y, z) : ax + by + cz = d\}$ where $d \neq 0$. In this plane in $\mathbf{R}^3$, the projective curve Γ determines the affine view Π given by

$$\{(x, y, z) : F(x, y, z) = 0,\ ax + by + cz = d\} :$$

the affine curve Π is the intersection in $\mathbf{R}^3$ of the affine plane $ax+by+cz = d$ and the surface $\{(x, y, z) : F(x, y, z) = 0\}$. The projective curve can be regarded as the union of the affine curve and the set of points at infinity for the affine curve. The set of points at infinity for the affine curve in the plane $ax + by + cz = d$ is

$$\{(x{:}y{:}z) : ax + by + cz = 0,\ F(x, y, z) = 0\}.$$

Definition 16.10. The *affine view* (or *affine section*) of the projective curve

$$\{(x{:}y{:}z) : F(x, y, z) = 0\}$$

for the plane $ax + by + cz = d$ $(d \neq 0)$ is the affine curve in $\mathbf{R}^3$,

$$\{(x, y, z) : F(x, y, z) = 0,\ ax + by + cz = d\}.$$

Notice that the projective curve lies in $\mathbf{R}\mathrm{P}^2$ and the affine view lies in the plane $ax + by + cz = d$ in $\mathbf{R}^3$. This plane does not pass through the origin of $\mathbf{R}^3$.

Mnemonic

Projective curve ↔ affine view + points at infinity for affine view.

Worked example 16.11. Find the affine view for $y + z = 1$ of the projective curve

$$x^3 - y^3 - y^2z = 0.$$

Solution. Substituting $z = 1-y$ gives $x^3 = y^2$. The equation $x^3 = y^2$ in $\mathbf{R}^3$ gives the cylinder passing through the curve $x^3 = y^2$ in the xy–plane and with generators parallel to the z–axis. The affine view is the intersection of this cylinder with the plane $y + z = 1$. It has a cusp at $(0, 0, 1)$. □

16.6 Plane curves as views of a projective curve

An important case of affine views is where the affine plane is taken to be $z = 1$. The affine view in $\mathbf{R}^3$ is then

$$\{(x, y, z) : z = 1, F(x, y, 1) = 0\}$$

in the plane $z = 1$. We may identify this with the plane curve

$$F(x, y, 1) = 0$$

in $\mathbf{R}^2$. Notice that this plane curve is not given by putting $z = 0$ in $F(x, y, z) = 0$. The points at infinity for the plane curve are then

$$\{(x{:}y{:}z) : z = 0, F(x, y, 0) = 0\}$$

in $\mathbf{RP}^2$.

Theorem 16.12. *Let $F(x, y, z) = 0$ be the projective curve determined by the plane curve $f(x, y) = 0$, then the affine view for $z = 1$ of the projective curve $F(x, y, z) = 0$ is $f(x, y) = 0$. Conversely let $f(x, y) = 0$ be the affine view for $z = 1$ of a projective curve $F(x, y, z) = 0$, for which z is not a factor of $F(x, y, z)$, then the projective curve determined by $f(x, y) = 0$ is $F(x, y, z) = 0$.*

Proof. We assume that the curves are of degree $n \geq 1$.
i) We have formally

$$F(x, y, z) = z^n f\left(\frac{x}{z}, \frac{y}{z}\right).$$

The affine view for $z = 1$ is

$$F(x, y, 1) = f(x, y).$$

ii) We have $f(x, y) = F(x, y, 1)$. This determines the projective curve

$$z^n f\left(\frac{x}{z}, \frac{y}{z}\right) = z^n F(\frac{x}{z}, \frac{y}{z}, 1) = F(x, y, z).$$

□

The affine views for the planes $x = 1$ and $y = 1$ can similarly be identified with plane curves in $\mathbf{R}^2 = \{(y, z)\}$ and $\mathbf{R}^2 = \{(x, z)\}$ respectively, and a result similar to Theorem 16.12 holds.

Worked example 16.13. Determine the affine views and their points at infinity of the projective curve $y^2 z = x^3$ for the planes $x = 1$, $y = 1$, and $z = 1$.

Solution. To find the affine view for the plane $y = 1$, we put $y = 1$ in the equation $y^2 z = x^3$. The affine view is $z = x^3$. To find the point at infinity we put $y = 0$ in the equation $y^2 z = x^3$; the plane curve therefore has just one point $(0{:}0{:}1)$ at infinity. The affine view for $z = 1$ is the semi-cubical parabola $y^2 = x^3$, which also has just one point $(0{:}1{:}0)$ at infinity. The affine view for $x = 1$ is $y^2z = 1$, which has two points $(0{:}1{:}0)$ and $(0{:}0{:}1)$ at infinity. Recall that these points at infinity are in $\mathbf{RP}^2$. $\square$

We now give an example involving conic sections.

Worked example 16.14. Show that ellipses, hyperbolae, and parabolae can occur as affine views of the projective curve given by $F = x^2 + y^2 - z^2$.

Solution. The homogeneous polynomial $F = x^2 + y^2 - z^2$ determines the right circular cone

$$\{(x, y, z) \in \mathbf{R}^3 : x^2 + y^2 - z^2 = 0\},$$

which is a surface in $\mathbf{R}^3$ (see Figure 16.4), and determines the projective curve

$$\Gamma = \{(x{:}y{:}z) : x^2 + y^2 - z^2 = 0\}$$

in $\mathbf{RP}^2$. The affine view of this projective curve depends on the choice of the affine plane $ax + by + cz = d$. We note three of the views which can occur.

i) The affine view for $z = 1$. In this case the line at infinity is $z = 0$, and the affine view is $x^2 + y^2 = 1$, which is an ellipse. There are no points at infinity.

ii) The affine view for $x = 1$. In this case the line at infinity is $x = 0$, and the affine view is $z^2 - y^2 = 1$, which is a hyperbola. There are two points $(0{:}1{:}\pm1)$ at infinity.

iii) The affine view for $y + z = 1$. In this case the line at infinity is $y + z = 0$, and the affine view is given by

$$x^2 + y^2 - (1 - y)^2 = 0 \quad \text{and} \quad y + z = 1,$$

that is

$$x^2 = 1 - 2y = -2\left(y - \tfrac{1}{2}\right) \quad \text{and} \quad y + z = 1.$$

This is a parabola in the plane $y + z = 1$. It is the intersection of the parabolic cylinder $x^2 = -2\left(y - \frac{1}{2}\right)$ in $\mathbf{R}^3$ with the plane $y + z = 1$. To find the points at infinity, we put $y + z = 0$. This gives $x = 0, y + z = 0$; therefore there is precisely one point $(0{:}1{:}-1)$ at infinity. (Compare Theorem 16.34 below.) $\square$

Mnemonic

Ellipses, hyperbolae, and parabolae are affine views of the same projective curve.

The set of points at infinity for a projective curve depends on the view chosen. Different views will give different sets of points at infinity, as illustrated in the above examples.

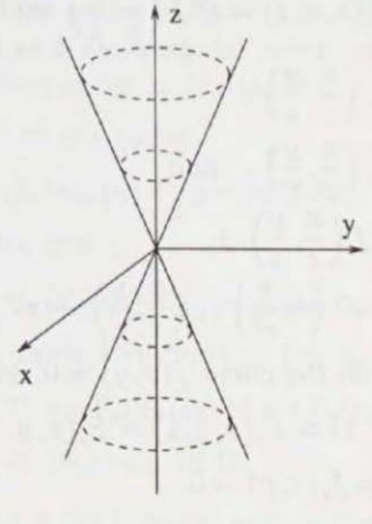

Figure 16.4 *A right circular cone in 3-space.*

16.7 Tangent lines to projective curves

The tangent lines to a plane algebraic curve $f(x, y) = 0$ are most readily defined at non-singular points, that is at points at which $\mathbf{grad}\, f \neq 0$. We consider the analogous situation for projective curves.

Definition 16.15. A *non-singular point* of the projective curve

$$F(x, y, z) = 0$$

is a point on the curve at which

$$\mathbf{grad}\, F = \left(\frac{\partial F}{\partial x}, \frac{\partial F}{\partial x}, \frac{\partial F}{\partial x}\right) \neq 0.$$

A *singular point* is a point at which

$$\mathbf{grad}\, F = \left(\frac{\partial F}{\partial x}, \frac{\partial F}{\partial x}, \frac{\partial F}{\partial x}\right) = 0.$$

The singular points of a plane curve correspond to singular points of the associated projective curve as in the following theorem.

Theorem 16.16. *(x, y) is a singular point of the plane curve $f(x, y) = 0$ if, and only if, $(x : y : 1)$ is a singular point of the associated projective curve $F(x, y, z) = 0$.*

Proof. Differentiating $F(x,y,z) = z^n f\left(\frac{x}{z}, \frac{y}{z}\right)$ we have

$$\begin{aligned} F_x(x,y,z) &= z^n \frac{1}{z} f_x\left(\frac{x}{z}, \frac{y}{z}\right), \\ F_y(x,y,z) &= z^n \frac{1}{z} f_y\left(\frac{x}{z}, \frac{y}{z}\right), \quad \text{and} \\ F_z(x,y,z) &= nz^{n-1} f\left(\frac{x}{z}, \frac{y}{z}\right) + \\ &\qquad z^n\left(-\frac{x}{z^2}\right) f_x\left(\frac{x}{z}, \frac{y}{z}\right) + z^n\left(-\frac{y}{z^2}\right) f_y\left(\frac{x}{z}, \frac{y}{z}\right). \end{aligned}$$

Thus, for a point (x,y) on the curve $f(x,y) = 0$, we have

$$F_x(x,y,1) = F_y(x,y,1) = F_z(x,y,1) = 0$$

if, and only if, $f_x(x,y) = f_y(x,y) = 0$. □

Similar results apply for the affine views given by $x = 1$ and $y = 1$, and more generally for the affine views given by $ax + by + cz = 1$.

We now determine the tangent lines at non-singular points of a projective curve. The tangent line of a projective curve is a projective line.

Theorem 16.17. *Let (x_0, y_0, z_0) be a non-singular point of the projective curve $F(x,y,z) = 0$, then the tangent line (a projective line in $\mathbf{R}P^2$) at (x_0, y_0, z_0) is*

$$\boxed{\mathbf{r}\cdot\mathbf{grad}\, F = xF_x(x_0,y_0,z_0) + yF_y(x_0,y_0,z_0) + zF_z(x_0,y_0,z_0) = 0.}$$

Proof.* We give a geometrical proof. The tangent plane to the surface $F(x,y,z) = 0$ in $\mathbf{R}^3$ at (x_0, y_0, z_0) is

$$(x - x_0)F_x + (y - y_0)F_y + (z - z_0)F_z = 0$$

where

$$F_x = \frac{\partial F}{\partial x}, \quad F_y = \frac{\partial F}{\partial y} \quad \text{and} \quad F_z = \frac{\partial F}{\partial z}$$

at (x_0, y_0, z_0). This equation is of the form

$$xF_x + yF_y + zF_z = \text{constant}.$$

Since this plane must pass through the origin (the surface is given by a pencil of lines through the origin), it has the form

$$xF_x + yF_y + zF_z = 0.$$

□

The partial derivatives F_x, F_y, and F_z must be evaluated at the fixed point (x_0, y_0, z_0).

---

* For the meaning of * see the preface.

Theorem 16.18. *The associated projective line of the tangent line to the plane curve $f(x,y)=0$ at a non-singular point (x_0,y_0) is the tangent line to the associated projective curve at $(x_0:y_0:1)$.*

Proof. The tangent line at (x_0,y_0) is

$$(x-x_0)f_x(x_0,y_0)+(y-y_0)f_y(x_0,y_0)=0.$$

The associated projective line is

$$(x-x_0z)f_x(x_0,y_0)+(y-y_0z)f_y(x_0,y_0)=0.$$

The tangent line to the projective curve at $(x_0:y_0:1)$ is

$$xF_x(x_0:y_0:1)+yF_y(x_0:y_0:1)+zF_z(x_0:y_0:1)=0,$$

that is, as in the proof of Theorem 16.16,

$$xf_x(x_0,y_0)+yf_y(x_0,y_0)+z(nf(x_0,y_0)-x_0f_x(x_0,y_0)-y_0f_y(x_0,y_0))=0.$$

The result follows since (x_0,y_0) lies on the curve $f(x,y)=0$. □

The tangent line $xF_x+yF_y+zF_z=0$ to the projective curve

$$F(x,y,z)=0$$

at $(x_0:y_0:1)$ determines the tangent line $xF_x+yF_y+F_z=0$ at $(x_0,y_0,1)$ to the affine view given by $z=1$; that is the tangent line $xF_x+yF_y+F_z=0$ at (x_0,y_0) in the case of the corresponding plane curve. We consider later the tangent lines to the projective curve at the points at infinity for the plane curve.

Worked example 16.19. Find the singular points of the projective curve

$$x^2z=y^3$$

and determine the tangent lines at the non-singular points $(x_0:y_0:z_0)$.

Solution. $F_x=2xz, F_y=-3y^2, F_z=x^2$, thus $\mathbf{grad}\,F=0$ if, and only if, $x=y=0$. There is exactly one singular point at $(0:0:1)$. The tangent at a non-singular point $(x_0:y_0:z_0)$ is

$$2x_0z_0x-3y_0^2y+x_0^2z=0.$$

□

16.8 Boundedness of the associated affine curve

One of the important applications of projective geometry to the study of plane curves is to determine boundedness.

Definition 16.20. A curve in $\mathbf{R}^2$ or $\mathbf{R}^3$ is *bounded* if there is a number K such that $|\mathbf{r}|\le K$ for each point $\mathbf{r}$ of the curve.

Theorem 16.21. *Let a given affine or plane curve have no points at infinity, then the plane curve is bounded (as a curve in $\mathbf{R}^3$ or $\mathbf{R}^2$).*

*Sketch of proof**. Our proof uses similar techniques to the proof of Theorem 15.18. Given a polynomial $f(x, y)$ of degree N, we may write it in the form $f(x, y) = g(x, y) + h(x, y)$ where $g(x, y)$ is the homogeneous polynomial consisting of all monomials of $f(x, y)$ of degree N, and $h(x, y)$ has degree $\leq N - 1$. We construct the homogeneous polynomial $F(x, y, z)$ associated with $f(x, y)$ as above. Clearly $F(x, y, 0) = g(x, y)$, and, since there are no points at infinity, the equation $F(x, y, 0) = 0$ has no non-trivial solution, that is $g(x, y)$ has no real linear factors. Therefore $g(x, y)$ is a product of irreducible quadratic homogeneous polynomials and consequently has even degree $N = 2n$. As in the proof of Theorem 15.18, there is a $K > 0$ such that

$$| g(x, y) | \geq K(x^2 + y^2)^n$$

for all real x and y. On the other hand if $\phi(x, y)$ is a homogeneous polynomial of degree r, there is an $L > 0$ such that

$$| \phi(x, y) | \leq L(x^2 + y^2)^{r/2}$$

for all real x and y (we use the inequalities

$$|x| \leq (x^2 + y^2)^{1/2} \quad \text{and} \quad |y| \leq (x^2 + y^2)^{1/2}).$$

Now $h(x, y)$ is a sum of finitely many homogeneous polynomials each of degree $\leq 2n - 1$, therefore there is an $M > 0$ such that

$$| h(x, y) | \leq M(x^2 + y^2)^{(2n-1)/2}$$

for all real x and y satisfying $(x^2 + y^2) \geq 1$. Thus we have

$$|f(x, y)| = |g(x, y) + h(x, y)| \geq (x^2 + y^2)^{(2n-1)/2}(K(x^2 + y^2)^{1/2} - M)) > 0$$

for all real x and y with

$$x^2 + y^2 > \frac{M^2}{K^2}.$$

Therefore $f(x, y) = 0$ has no solutions for

$$x^2 + y^2 > \max \left\{1, \frac{M^2}{K^2}\right\},$$

which proves that the curve $f(x, y) = 0$ is bounded. □

Worked example 16.22. Show that the affine view for $z = 1$ of the projective curve $(x^2 + y^2)^2 - (x^2 - y^2)z^2 = 0$ is bounded.

Solution. Substituting $z = 0$ into $(x^2 + y^2)^2 - (x^2 - y^2)z^2 = 0$, we obtain $x = y = z = 0$. Therefore the affine view for $z = 1$ has no points at infinity ((0:0:0) does not represent a point in $\mathbf{R}P^2$). Thus the affine curve

$$(x^2 + y^2)^2 = x^2 - y^2$$

is bounded. □

---

* For the meaning of * see the preface.

The converse of Theorem 16.21 is not true. The existence of points at infinity does not imply that the affine curve is not bounded. The points at infinity could be isolated points of the projective curve. More generally it can be shown that an affine or plane curve is bounded if, and only if, the only points at infinity are isolated.

Worked example 16.23. Find the singular points of the projective curve

$$y^2z^2 = x^2z^2 - x^4.$$

Show that the affine view for $z = 1$ has a point at infinity but is bounded.

Solution. The affine view for $z = 1$ is the figure–8 curve

$$y^2 = x^2(1 - x^2).$$

It has singular points at $(0{:}0{:}1)$ and $(0{:}1{:}0)$. Now $(1 - x^2) \geq 0$ and

$$\frac{d}{dx}(x^2(1 - x^2)) = 2x(1 - 2x^2).$$

Thus y is zero at $x = 0$ and $x = \pm 1$ and has critical points at

$$x = \pm\frac{1}{\sqrt{2}}.$$

It follows that the affine view lies in the rectangle

$$\{-1 \leq x \leq 1, -\tfrac{1}{2} \leq y \leq \tfrac{1}{2}\}.$$

However this affine view has an isolated singular point $(0{:}1{:}0)$ at infinity. Notice that for the affine view given by $y = 1$ this singular point of the projective curve becomes the origin of the (x, z)–plane, and the quadratic terms of the equation have an iterated linear factor which does not correspond to a (real) tangent at the origin. □

16.9 Summary of the analytic viewpoint

1. The projective plane $\mathbf{RP}^2$ is the set of triples $(x{:}y{:}z)$ where
   i) $(x, y, z) \neq 0$ that is x, y, z are not all zero, and
   ii) $(\lambda x{:}\lambda y{:}\lambda z) = (x{:}y{:}z)$ for each $\lambda \neq 0$.

For $(x, y, z) \neq 0$, the set $\{\lambda(x, y, z)\}$ for varying λ, including $(0, 0, 0)$ given by $\lambda = 0$, is a line in $\mathbf{R}^3$ through (x, y, z). Excluding $(0, 0, 0)$ from the line determines a point of $\mathbf{RP}^2$ in this analytic sense by ii.

2. A curve of degree n in $\mathbf{RP}^2$ is the set of points of $\mathbf{RP}^2$ satisfying $F(x, y, z) = 0$, where F is a homogeneous polynomial of degree $n \geq 1$.

3. A plane $ax+by+cz = d$ in $\mathbf{R}^3$, for which $d \neq 0$, determines (bijectively) an affine plane regarded as a subset of $\mathbf{RP}^2$; and the set of points of $\mathbf{RP}^2$, which are not on this plane, constitute the (projective) line at infinity of the affine plane. The projective line given by the equation $ax + by + cz = 0$

in $\mathbf{RP}^2$ is the line is the line at infinity of the affine plane $ax + by + cz = d$.

4. The affine plane $ax + by + cz = d$ ($d \neq 0$) intersects a projective curve of degree n in an algebraic curve of degree $\leq n$.

5. Often the affine plane considered is $z = 1$ and the projective line of points at infinity is then precisely the set $\{(x : y : 0)\}$, where x, y are not both zero. Each point in this latter set can be written $(x : 1 : 0)$ except the point $(1 : 0 : 0)$.

16.10 Asymptotes

The tangent lines to a projective curve, excluding tangent lines at points at infinity, correspond bijectively with tangent lines to the affine or plane curve. We now consider tangent lines to the projective curve at the points at infinity (see Figure 16.5).

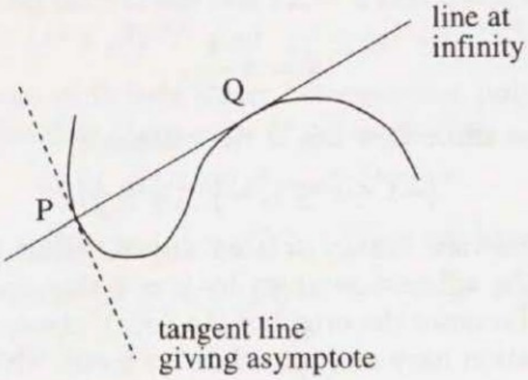

Figure 16.5 *Tangents at points at infinity.*

Definition 16.24. An *asymptote* of a plane curve Γ is a tangent to the projective curve determined by it at a point at infinity, which tangent is not the line at infinity.

Theorem 16.25. *A plane or affine curve of degree n has at most n distinct points at infinity.*

Proof. The points at infinity are the points of intersection of the associated projective curve, which has the same degree, and the line at infinity. The result follows from Theorem 16.7. ☐

Example 16.26. $\prod_{r=1}^{n}(x - ry) + 1 = 0$ has n distinct points $(r : 1 : 0)$ at infinity. For $n = 2$ this is a hyperbola.

For points at infinity on a projective curve there are two cases to consider as follows.

1. P is a point at infinity and at least one tangent at P is not the line at infinity. In this case each tangent, at P, which is not the line at infinity will determine an asymptote.

2. P is a point at infinity and the unique tangent at P is the line at infinity. In this case the point at infinity does not give rise to an asymptote.

16.10.1 Procedure for determining the asymptotes

We demonstrate the method for finding the asymptotes to a plane curve

$$f(x, y) = 0.$$

First find the points at infinity on the related projective curve

$$F(x, y, z) = z^n f\left(\frac{x}{z}, \frac{y}{z}\right) = 0.$$

Find the tangents at these points: these tangents will be projective lines. For those tangents which are not the line at infinity, determine the related affine line in the plane $z = 1$. The asymptotes to the plane curve will have the same equations as these affine lines.

Worked example 16.27. Find the points at infinity and the asymptotes of the hyperbola

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1.$$

Solution. The plane curve is the affine view for $z = 1$ of the projective curve

$$F(x, y, z) = \frac{x^2}{a^2} - \frac{y^2}{b^2} - z^2 = 0.$$

The points at infinity for the plane curve are given by $z = 0$ and

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 0:$$

the points at infinity are therefore $(a{:}b{:}0)$ and $(a{:}-b{:}0)$. The tangent line to the projective curve at (p, q, r) is

$$\frac{2p}{a^2}x - \frac{2q}{b^2}y - 2rz = 0.$$

For the two points at infinity the tangents are the projective lines

$$\frac{x}{a} = \pm\frac{y}{b}.$$

Putting $z = 1$, we now obtain the plane lines

$$\frac{x}{a} = \pm\frac{y}{b}$$

in $\mathbf{R}^2$; these are the asymptotes. □

In the above example the equations for the projective lines and plane lines corresponding to the asymptotes are the same. This will not happen in general since substituting $z = 1$ will usually give a constant term.

Example 16.28. In Example 16.14 we considered the affine views of

$$x^2 + y^2 - z^2 = 0.$$

We now investigate the asymptotes of these views.

i) The affine view for $z = 1$. In this case the line at infinity is $z = 0$. There are no points at infinity and therefore no asymptotes.

ii) The affine view for $x = 1$. In this case the line at infinity is $x = 0$. There are two points $(0\colon 1\colon \pm 1)$ at infinity. The tangents to the projective curve at these two points are $y \mp z = 0$. The asymptotes to the plane curve, given by substituting $x = 1$, are $y \mp z = 0$.

iii) The affine view for $y + z = 1$. In this case the line at infinity is $y + z = 0$. There is precisely one point $(0\colon 1\colon -1)$ at infinity. The tangent to the projective curve at this point is $y + z = 0$, which is the line at infinity. Therefore the affine curve has no asymptote.

Recall that the three homogeneous-coordinate 'axes' $x = 0$, $y = 0$ and $z = 0$ are determined by great-circles on the sphere. The diagrams in Figure 16.6 demonstrate what is happening.

16.11 Singular points and inflexions of projective curves

A singular point of a projective curve $F(x, y, z) = 0$ of degree n is a point at which

$$\mathbf{grad}\, F = (F_x, F_y, F_z) = 0.$$

By Euler's theorem (see §4.7.6),

$$nF = xF_x + yF_y + zF_z$$

and therefore any point $(x\colon y\colon z)$, which satisfies $F_x = F_y = F_z = 0$, lies on the curve. By Theorem 16.16 above, a point in an affine view is a singular point if, and only if, the corresponding point of the projective curve is a singular point.

Definition 16.29. A singular point of a projective curve is *isolated*, *nodal*, or a *cusp* according as the associated point (not at infinity) in an affine view is isolated, nodal, or a cusp.

It can be shown that the above definition is independent of the view chosen.

Definition 16.30. A *singular point at infinity* of a plane algebraic curve $f(x, y) = 0$, is a point at infinity which is also a singular point of the associated projective curve $F(x, y, z) = 0$ defined by

$$F(x, y, z) = z^n f\left(\frac{x}{z}, \frac{y}{z}\right).$$

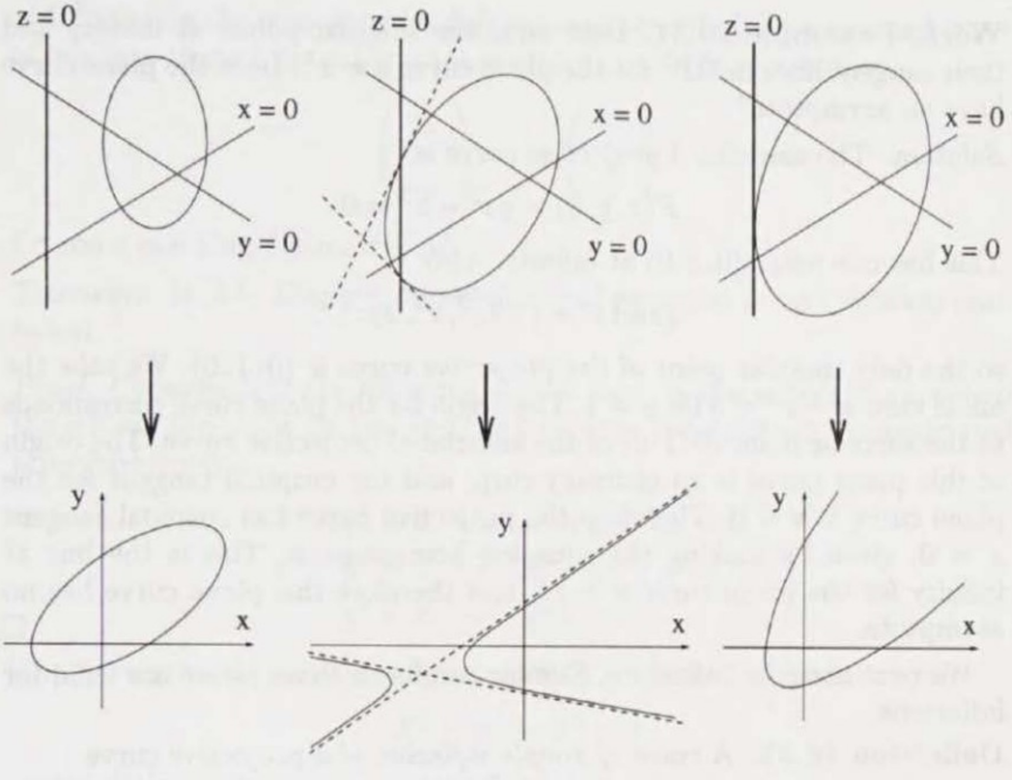

Ellipse Hyperbola Parabola

Figure 16.6 *Asymptotes of second order projective curves.*

In order to classify a singular point at infinity for a plane or affine curve, we can take a different affine view of the associated projective curve for which the singular point which we are considering is not at infinity, and classify it (cusp, nodal, isolated) in the usual way. It can be shown that if a singular point is a particular type in one affine view (for example ordinary cusp, 3–node, etc.), then it is of the same type in other affine views.

The tangent lines at singular points of a projective curve can be found as follows. First we take an affine view (usually given by one of $x = 1$, $y = 1$ or $z = 1$), for which the singular point is not the point at infinity, and then find the tangents to the corresponding plane curve at this singular point in one of the usual ways (see for example Chapter 15). On making homogeneous the equations of these tangent lines to the plane curve, then finding the projective lines determined by them, we obtain the equations of the required tangent lines to the projective curve. It can be shown that the tangent lines found in this way do not depend on the affine view used in determining them.

Worked example 16.31. Determine the singular points at infinity and their tangent lines in $\mathbf{RP}^2$ for the plane curve $y = x^3$. Does the plane curve have an asymptote?

Solution. The associated projective curve is

$$F(x, y, z) = yz^2 - x^3 = 0.$$

This has one point $(0:1:0)$ at infinity. Also

$$\mathbf{grad}\, F = (-3x^2, z^2, 2yz),$$

so the only singular point of the projective curve is $(0:1:0)$. We take the affine view $z^2 - x^3 = 0$ for $y = 1$. The origin for the plane curve corresponds to the singular point $(0:1:0)$ of the associated projective curve. The origin of this plane curve is an ordinary cusp, and the cuspidal tangent for the plane curve is $z = 0$. Therefore the projective curve has cuspidal tangent $z = 0$, given by making the equation homogeneous. This is the line at infinity for the plane curve $y = x^3$, and therefore this plane curve has no asymptote. □

We next consider inflexions. Similar results to those above are valid for inflexions.

Definition 16.32. A *point of simple inflexion* of a projective curve

$$F(x, y, z) = 0$$

is a point which corresponds to a point of simple inflexion in some affine view.

Again it can be shown that the above definition is independent of the view chosen. Points of higher inflexion and points of simple and higher undulation can be defined in a similar way, and again it can be shown that the definitions are independent of the view chosen.

16.12 Equivalence of curves*

Equivalence of curves is dependent on the type of transformations allowed. Two curves in the plane are congruent if there is a rigid map sending one to the other; congruence is one type of equivalence. In Chapter 2 we showed how conics can be moved to canonical position using a rigid map. We now describe another form of equivalence of plane curves, and a form of equivalence for projective curves.

16.12.1 Projective equivalence of plane curves

Definition 16.33. Two plane curves given by $f(x, y) = 0$ and $g(x, y) = 0$ are *projectively equivalent* if there is a non-singular linear transformation of $\mathbf{R}^3$ sending the homogeneous defining polynomial of the associated curve of one into that of the other, that is $G(x,y,z) = F(u,v,w)$ where

---

* For the meaning of * see the preface.

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = M \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

for some non-singular matrix M.

Theorem 16.34. *Ellipses, hyperbolae, and parabolae are projectively equivalent.*

Proof. Consider first the following special case: in the array the curves on the left are the plane curves, and the curves on the right are the associated projective curves.

$$\begin{aligned} x^2 + y^2 = 1 &\quad\mapsto\quad x^2 + y^2 - z^2 = 0, \\ x^2 - y^2 = 1 &\quad\mapsto\quad x^2 - y^2 - z^2 = 0, \\ y^2 = x &\quad\mapsto\quad y^2 = xz. \end{aligned}$$

The homogeneous polynomial representing the last of these becomes

$$X^2 - Y^2 - Z^2 = 0$$

under the non-singular linear transformation

$$x = X + Z,\ y = Y \text{ and } z = X - Z.$$

It follows readily that these three curves are projectively equivalent: the only further transformation required is a change of coordinate axes in each case. By elementary geometrical considerations any ellipse, hyperbola, or parabola is projectively equivalent to one of these three. To prove this, first move a given (plane) conic to canonical position by a rotation and a translation of $\mathbf{R}^2$. Each of these transformations determines a non-singular linear transformation of $\mathbf{R}^3$. The rotation determines the obvious rotation which has the same effect on the xy-plane as the given rotation and which leaves the z-axis fixed. The translation of $\mathbf{R}^2$, given by

$$x \mapsto x + a \text{ and } y \mapsto y + b,$$

determines the non-singular linear transformation of $\mathbf{R}^3$ given by

$$x \mapsto x + az,\ y \mapsto y + bz \text{ and } z \mapsto z.$$

A further non-singular linear transformation (consisting of compressions along the axes) of $\mathbf{R}^2$ then reduces the (Euclidean) canonical form (see Chapter 1) of the ellipse, hyperbola, or parabola to the appropriate one of the forms given above. The three affine equations on the left of the above array are, respectively, the affine canonical forms of the ellipse, hyperbola, and parabola. □

Worked example 16.35. Prove that the acnodal cubic $x^3 - x^2 - y^2 = 0$ is projectively equivalent to the witch of Agnesi $1 - x(1 + y^2) = 0$.

Solution. The two associated projective curves are respectively

$$x^3 - x^2 z - y^2 z = 0$$

and $z^3 - xz^2 - xy^2 = 0$. The non-singular linear transformation

$$x \mapsto z, z \mapsto x, y \mapsto y$$

transforms one into the other. $\square$

16.12.2 Equivalence of projective curves

Definition 16.36. Two projective curves Φ and Γ given respectively by $F(x, y, z) = 0$ and $G(x, y, z) = 0$ are *equivalent* if $G(x, y, z) = F(u, v, w)$ where

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = M \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

for some non-singular matrix M.

'Equivalence of projective curves' is clearly an equivalence relation. Under this equivalence, singular points are mapped to singular points, tangents to tangents etc. Two affine curves are *projectively equivalent* if they are affine views of two equivalent projective curves (see §16.12.1). Two plane curves are *affinely equivalent* if one can be obtained from the other by a translation together with a non-singular (i.e., invertible) linear map (the latter need not be a rotation). Also shearing maps, such as $x \mapsto x$, $y \mapsto x + y$, are non-singular linear maps but do not preserve angles or distance. Such affine maps do preserve 'tangents', 'two lines being parallel', 'cusps', and 'singular points', but do not preserve 'normals'.

16.13 Examples of asymptotic behaviour

We give examples of asymptotic behaviour in Figure 16.7 as follows.

i) The projective curve does not cross its tangent at a point at infinity and the tangent is not the line at infinity $\Rightarrow$ two distant parts of the affine curve typically approach the asymptote on opposite sides of the asymptote.

ii) The projective curve has, as its tangent at a point at infinity, the line at infinity $\Rightarrow$ two distant parts of the affine curve tend to infinity, though not along an asymptote.

iii) The projective curve has a simple inflexion at a point at infinity and its tangent there is not the line at infinity $\Rightarrow$ two distant parts of the affine curve approach the asymptote on the same side of the asymptote.

iv) The projective curve has a simple double point at infinity and neither

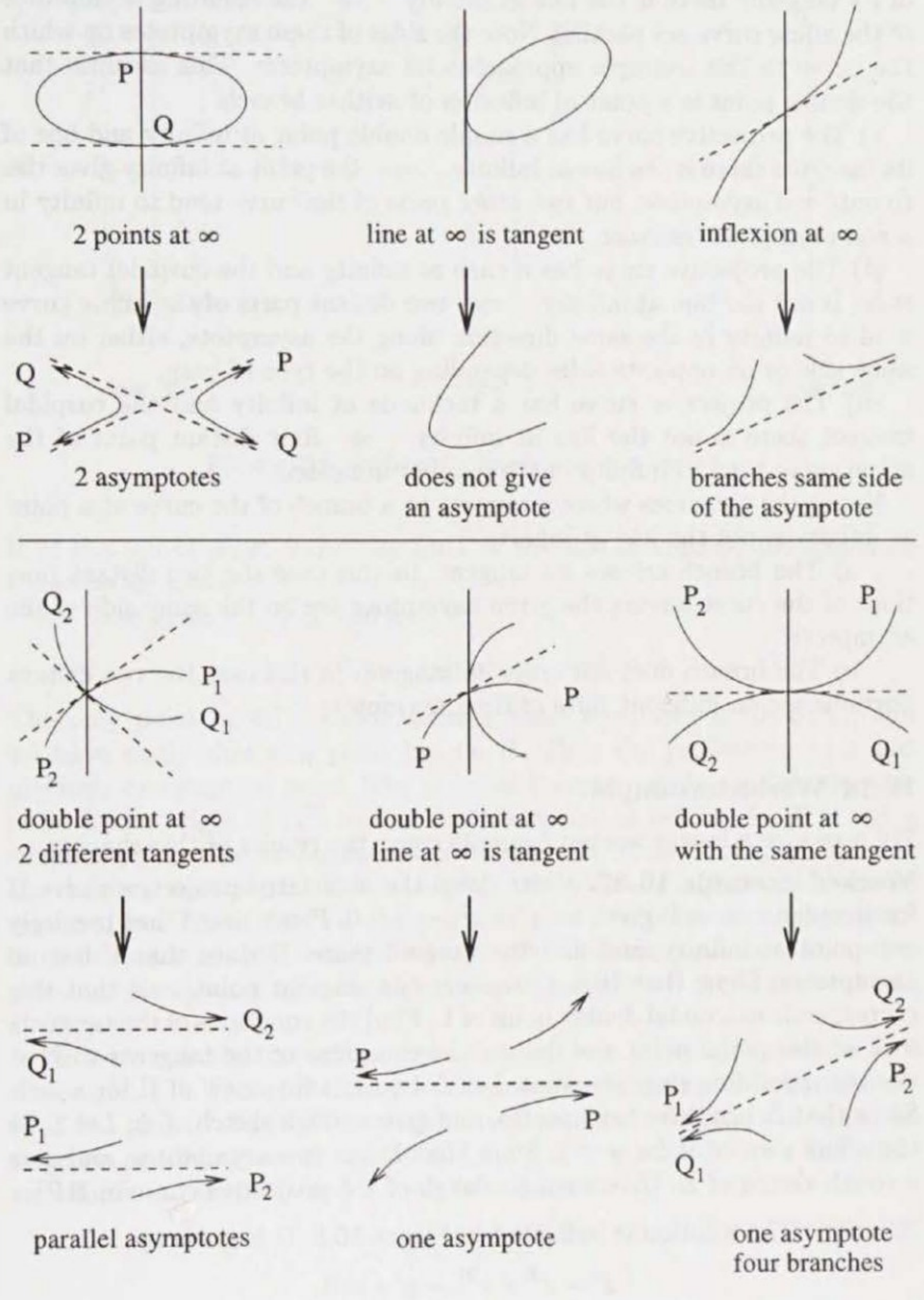
P
Q
2 points at ∞
line at ∞ is tangent
inflexion at ∞
Q
P
P
Q
2 asymptotes
does not give
an asymptote
branches same side
of the asymptote
Q_2
P_1
Q_1
P_2
P
P
P_2
P_1
Q_2
Q_1
double point at ∞
2 different tangents
double point at ∞
line at ∞ is tangent
double point at ∞
with the same tangent
Q_2
Q_1
P_1
P_2
P
P
Q_2
P_2
P_1
Q_1
parallel asymptotes
one asymptote
one asymptote
four branches

Figure 16.7 *More examples of asymptotic behaviour.*

of its tangents there is the line at infinity ⇒ the resulting asymptotes of the affine curve are parallel. Note the sides of these asymptotes on which the curve in this example approaches its asymptotes. This assumes that the double point is a point of inflexion of neither branch.

v) The projective curve has a simple double point at infinity and one of its tangents there is the line at infinity ⇒ the point at infinity gives rise to only one asymptote, but two other parts of the curve tend to infinity in a non-asymptotic manner.

vi) The projective curve has a cusp at infinity and the cuspidal tangent there is not the line at infinity ⇒ two distant parts of the affine curve tend to infinity in the same direction along the asymptote, either on the same side or on opposite sides depending on the type of cusp.

vii) The projective curve has a tachnode at infinity and the cuspidal tangent there is not the line at infinity ⇒ four distant parts of the affine curve tend to infinity in the manner indicated.

Notice the two cases where a tangent to a branch of the curve at a point at infinity is not the line at infinity.

a) The branch crosses its tangent. In this case the two distant portions of the curve having the given asymptote are on the same side of the asymptote.

b) The branch does not cross its tangent. In this case the two distant portions are on different sides of their asymptote.

16.14 Worked example

We now give a longer worked example using the results of this chapter.

Worked example 16.37. Write down the associated projective curve Π for the plane curve Γ given by $x^3 + x^2 - y^2 = 0$. Prove that Γ has precisely one point at infinity, and find the tangent there. Deduce that Γ has no asymptotes. Show that Π has precisely one singular point, and that this corresponds to a nodal double point of Γ. Find the equations of the tangents to Γ at this nodal point and deduce the equations of the tangents to Π at the corresponding singular point. Let Δ be the affine view of Π for $x = 1$. Show that Δ has three asymptotes, and give a rough sketch of Δ. Let Σ be the affine view of Π for $y = 1$. Show that Σ has two asymptotes, and give a rough sketch of Σ. Give a rough sketch of the projective curve in $\mathbf{RP}^2$.

Solution. The solution is indicated in Figure 16.8. Π is given by

$$F = x^3 + x^2z - y^2z = 0.$$

We have

$$\mathbf{grad}\, F = (3x^2 + 2xz, -2yz, x^2 - y^2).$$

Now Π meets $z = 0$ where $x^3 = 0$, that is only at $(0\!:\!1\!:\!0)$. The tangent to Π at $(0\!:\!1\!:\!0)$ is $z = 0$ (given by $(x, y, z) \cdot \mathbf{grad}\, F = 0$). Thus the tangent to

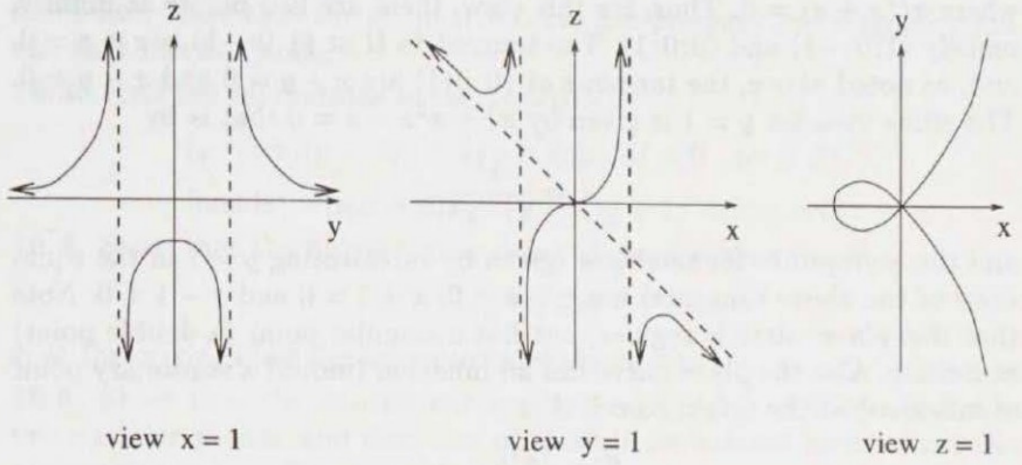

Figure 16.8 *Different affine views of a curve.*

Π at the (only) point at infinity for Γ is the line at infinity, and therefore Γ has no asymptotes.

We have **grad** $F = 0$ if, and only if,

$$3x^2 + 2xz = -2yz = x^2 - y^2 = 0.$$

The only point of $\mathbf{R}P^2$ which satisfies these equations is $(0:0:1)$, and we have easily that this point lies on Π. Thus the projective curve has precisely one singular point. The point of Γ corresponding to this singular point is the origin of $\mathbf{R}^2$. Notice that the point at infinity for Γ is not a singular point. The homogeneous polynomial of terms of lowest degree in the defining equation for Γ is $x^2 - y^2 = (x+y)(x-y)$. Since there are no repeated real linear factors, the origin is a double point and the tangent lines there are $x + y = 0$ and $x - y = 0$. The tangents to Π at the singular point (given by making these equations homogeneous) are $x + y = 0$ and $x - y = 0$.

We now consider the affine view of Π for $x = 1$. Now Π meets $x = 0$ where $y^2 z = 0$. Thus, for this view, there are two points at infinity, namely $(0:1:0)$ and $(0:0:1)$. The tangent to Π at $(0:1:0)$ is $z = 0$, and the tangents at $(0:0:1)$ are $x+y = 0$ and $x-y = 0$, as noted above. The affine view for $x = 1$ is given by $1 + z - y^2 z = 0$ that is by

$$z = \frac{1}{y^2 - 1}$$

and the asymptotes for this view (given by substituting $x = 1$ in the equations of the above tangents) are $z = 0$, $1 + y = 0$ and $1 - y = 0$. Note that this plane curve has no singular point in the plane, but has a singular point (a double point) at infinity.

We now consider the affine view of Π for $y = 1$. Now Π meets $y = 0$

where $x^2(x+z)=0$. Thus, for this view, there are two points at infinity, namely $(1{:}0{:}-1)$ and $(0{:}0{:}1)$. The tangent to Π at $(1{:}0{:}-1)$ is $x+z=0$, and, as noted above, the tangents at $(0{:}0{:}1)$ are $x+y=0$ and $x-y=0$. The affine view for $y=1$ is given by $x^3+x^2z-z=0$ that is by

$$z=\frac{x^3}{1-x^2}$$

and the asymptotes for this view (given by substituting $y=1$ in the equations of the above tangents) are $x+z=0$, $x+1=0$ and $x-1=0$. Note that this plane curve is regular, but has a singular point (a double point) at infinity. Also the plane curve has an inflexion (indeed a stationary point of inflexion) at the origin (check that

$$\frac{dz}{dx}=\frac{d^2z}{dx^2}=0).$$

Note in the sketch of the projective curve that no affine view of the projective curve is bounded, since any projective line (as a line infinity) meets the projective curve in at least one point. (Compare this with the situation for the projective curve associated with a hyperbola.) ☐

In the two views considered above the affine curves are graphs of functions, that is their equations take on the form $z=\varphi(y)$ and $z=\psi(x)$ respectively. Thus the curve is parametrised by the y or x variable respectively. The crunodal cubic is not the graph of a function even after a rotation.

The projective curve given by a quadratic homogeneous polynomial such as $x^2+y^2-z^2=0$ can be drawn in such a way that it 'resembles' a closed curve in the projective plane. This is not true for all projective curves as the above example shows. If it cannot be drawn to resemble a closed curve in the plane, we can usually note in our sketch in the projective plane that certain points must be identified in order to close the curve – recall that projective space is obtained from the sphere by identifying each point with its opposite. The projective curve Π sketched here, regarded as a curve on the sphere, will have as its 'opposite' on the sphere a similar looking curve, and when the two are joined together, they will form one continuous curve. On the other hand the quadratic cone projective curve determines on the sphere two disjoint closed curves, one of which is 'opposite' to the other.

Exercises

16.1. Find the singular points of $(x^2+y^2)^2-(x^2-y^2)z^2=0$.

$$[(1{:}1{:}\pm 2),(1{:}-1{:}\pm 2),(0{:}0{:}1)]$$

16.2. Show that the cubic

$$(x-y)(x^2+y^2)+2xyz=0$$

intersects the line $z=0$ in precisely one point and find the tangent line at this point. Show that the projective curve has precisely one singular point and that this is a node.

16.3. Find the asymptotes to the hyperbola

$$(y - \alpha x)(y - \beta x) + 2fx + 2gy + d = 0 \quad (\alpha \neq \beta).$$

$$[\text{one is} \quad -\alpha(\alpha - \beta)x + (\alpha - \beta)y + 2f + 2\alpha g = 0]$$

16.4. Show that the only singular point of the algebraic curve

$$x^4 - y^4 = 4xy^2$$

is at the origin. Find the asymptotes of the curve.

16.5. Show that the quartic curve $y^2(x^2 + z^2) - x^4 = 0$ has precisely two singular points, and that one of these is an isolated point. Find the asymptotes to the affine view given by $z = 1$.

$$[(0{:}0{:}1), (0{:}1{:}0); x^2 + z^2 - x^4 = 0; x - y = 0, x + y = 0]$$

16.6. Let $f(x, y) = g(x, y) + h(x, y) + k(x, y)$ be a polynomial, where g is non-trivial and homogeneous of degree $n \geq 2$, h is homogeneous of degree $n - 1$, and the degree of k is $\leq n - 2$. Show that the points at infinity $(x_0 : y_0 : 0)$ of $f(x, y) = 0$ are given by the real solutions of $g(x, y) = 0$. Show further that, where $g(x, y)$ has no repeated real linear factors, the asymptotes to the plane curve are the lines

$$x g_x(x_0, y_0) + y g_y(x_0, y_0) + h(x_0, y_0) = 0$$

corresponding to these solutions.

16.7. The projective curve Π given by

$$x^4 + x^2yz - y^3z + y^4 = 0$$

has as affine view for $z = 1$ the curve Σ, given by

$$x^4 + x^2y - y^3 + y^4 = 0\,.$$

Prove that the origin is a triple point of the affine curve Σ, and find the tangent lines to the curve there. Show that Π meets the axes $x = 0$, $y = 0$ and $z = 0$ altogether in precisely two points, and find the equations of the four tangents to Π at these two points. Deduce also that Σ is a bounded curve.

Show that the affine view Γ, for $x = 1$, of the projective curve Π is a graph. Write down the coordinates of the points of infinity for Γ, and determine the equations of its asymptotes. Make a rough sketch of Γ indicating the asymptotes and carefully indicating any intersection of the curve with the axes.

[$x = \pm y, y = 0 : (0{:}0{:}1), (0{:}1{:}1); x = \pm y, y = 0, y = z : (0{:}0{:}1), (0{:}1{:}1);$
$y = \pm 1, y = 0, y = z$]

16.8. The algebraic curve Σ with equation

$$y^2(1 - x) - x^2(1 + x) = 0$$

has a double point at the origin.

a) Let Π be the associated projective curve of Σ. Show that Σ has precisely one point at infinity, namely $(x{:}y{:}z) = (0{:}1{:}0)$ on Π. Let Γ be the affine view of Π for $y = 1$. Determine the tangent line to Γ at the origin, and show that it has 3–point contact with Γ at the origin. Write down the equation of the asymptote to Σ.

$[z = x, x = 1]$

b) Each line $y - m = mx$ through $V = (-1, 0)$ meets Σ in points Q and R and meets the y–axis in a point $P = (0, m)$. Show that $OP = QP = PR$, where O is the origin. The curve Σ is the right strophoid.

Draw this strophoid, in the usual way, using a drawn line and marked points V and O. For a line L through V which meets the drawn line in a point P, mark on L the points Q and R which satisfy $OP = QP = PR$. Repeat this for a suitable collection of lines through the origin, and draw the curve in the usual way.

16.9. The projective curve Π given by

$$y^4 + x^2y^2 + xy^2z - x^3z = 0$$

has as affine view for $z = 1$ the curve Σ, given by

$$y^4 + x^2y^2 + xy^2 - x^3 = 0.$$

Prove that the origin is a triple point of the affine curve Σ, and find the tangent lines to the curve there. Show that Π meets the axes $x = 0$, $y = 0$ and $z = 0$ altogether in precisely two points, and find the equations of the four tangents to Π at these two points. Deduce that Σ has no asymptotes.

Show that the affine view Γ, for $x = 1$, of the projective curve Π is a graph. Write down the coordinates of the points of infinity for Γ, and determine the equations of its asymptotes. Make a rough sketch of Γ indicating the asymptotes and carefully indicating any intersection of the curve with the axes.

$[x = \pm y, x = 0 : (0{:}0{:}1), (1{:}0{:}0); x = \pm y, x = 0, z = 0 : (0{:}0{:}1); y = \pm 1]$

16.10. The algebraic curve Σ with equation

$$(x^2 + y^2)y^2 - x^2 = 0$$

has a tachnode at the origin.

a) Let Π be the associated projective curve of Σ. Show that Σ has precisely one point at infinity, namely $(x{:}y{:}z) = (1{:}0{:}0)$ on Π. Let Γ be the affine view of Π for $x = 1$. Show that Γ has a node at the origin and find the tangent lines there. Write down the equations of the asymptotes to Σ.

$[y = \pm z, y = \pm 1]$

b) Each line $y = mx$ through the origin $O = (0, 0)$ meets Σ in points P and P', and meets the line $y = 1$ in a point Q. Prove that $OP = OP' = CQ$, where $C = (0, 1)$.

Use the following method to draw the curve Σ (Gutschoven's curve) using drawn rectangular axes and line $y = 1$. For a line L through O which meets $y = 1$ in point Q, mark on L the points P and P' which satisfy $OP = OP' = CQ$. Repeat this for a suitable collection of lines through the origin, and draw the curve in the usual way.

16.11. The algebraic curve given by

$$y^2(1-x) - x^2(m+x) = 0 \quad (m \neq 0, -1)$$

has a singular point at the origin and no other singular points. Show that the origin is an isolated point in case $m < 0$, and is a nodal double point in case $m > 0$. In the latter case write down the equations of the tangent lines.

The projective curve Π given by

$$y^2(z-x) - x^2(4z+x) = 0$$

has as affine view for $z = 1$ the curve Σ, given by

$$y^2(1-x) - x^2(4+x) = 0.$$

Show that Π meets the axes $x = 0$, $y = 0$ and $z = 0$ altogether in precisely three points, and find the equations of the four tangents to Π at these three points, including the two at the double point. Show that Σ has an asymptote and write down its equation.

Show that the affine view Γ, for $y = 1$, of the projective curve Π is a graph. Write down the coordinates of the points at infinity for Γ, and determine the equations of its asymptotes. Make a rough sketch of Γ indicating the asymptotes and carefully indicating any intersection of the curve with the axes.

$[y = \pm\sqrt{m}x; y = \pm 2x, z = x, x + 4z = 0; x = 1; 2x = \pm 1, x + 4z = 0]$

16.12. The projective curve Π given by

$$(x^2+y^2)(x-2y) + z(x^2-y^2) = 0$$

has, as affine view for $z = 1$, the affine curve Σ given by

$$(x^2+y^2)(x-2y) + x^2 - y^2 = 0.$$

Show that the origin is a singular point of Σ.

Find the other points of intersection of the projective curve with the three lines $x = 0, y = 0, z = 0$, and find the equations of the tangents at all four points.

Hence write down the equations of the asymptotes to the related affine curve Ω with equation

$$(x^2+1)(x-2) + z(x^2-1) = 0,$$

given that the tangents to the projective curve at $(0{:}0{:}1)$ have equations

$$x = \pm y.$$

Give a rough sketch of Ω, including on your sketch the asymptotes, and the tangent lines at the points where the curve crosses the axes.
(A scale of 1 unit = 2 cm is suggested.)

16.13. Show that the algebraic curve Σ with equation

$$(x-1)^2(x^2+y^2)-9x^2=0$$

has a double point at the origin.

a) Let Π be the associated projective curve of Σ. Show that Σ has precisely one point at infinity, namely $(x:y:z)=(0:1:0)$ on Π. Let Γ be the affine view of Π for $y=1$. The tangent line to Γ at the origin is $x=z$. Show that it has 4-point contact with Γ at the origin. Write down the equation of the asymptote to Σ. $[x=1]$

b) Each line $y=mx$ meets the line $x=1$ in a point P and the curve Σ in precisely 2 points Q and R. Show that Q and R lie on opposite sides of the line $x=1$, and that $QP=PR=3$. The curve Σ is a conchoid of Nicomedes.

Draw this conchoid, in the usual way, using a drawn line and origin. For a line L through the origin which meets the given line at P, mark the two points on L distant 6 cm from P (use the scale 1 unit = 2 cm). Repeat this for a suitable collection of lines through the origin, and draw the curve in the usual way.

16.14. The projective curve Π, given by

$$x\left(x^2+y^2\right)+\left(3x^2-y^2\right)z=0,$$

has as affine view for $z=1$ the trisectrix of MacLaurin Σ, given by

$$x\left(x^2+y^2\right)+\left(3x^2-y^2\right)=0.$$

Prove that the origin is a double point of the affine curve Σ and find the tangent lines to the curve there. Show that Π meets the axes $x=0$, $y=0$ and $z=0$ altogether in three points, and find the equations of the four tangents to Π at these three points, including the two at the double point. Show that Σ has an asymptote and write down its equation.

Let Γ be the affine view, for $y=1$, of the projective curve Π. Determine the equations of the asymptotes of Γ. Make a rough sketch of Γ indicating the asymptotes and carefully indicating the tangent line where Γ crosses the axis $z=0$. (You may assume that the projective curve Π has an inflexion at the point where it meets the axis $z=0$.)

16.15. The algebraic curve Σ with equation

$$x^3+xy^2-2y^2=0$$

has a cusp at the origin.

a) Let Π be the associated projective curve of Σ. Show that Σ has precisely one point at infinity, namely $(x:y:z)=(0:1:0)$ on Π. Let Γ be the affine view of Π for $y = 1$. Determine the tangent line to Γ at the origin, and show that it has 3–point contact with Γ at the origin. Write down the equation of the asymptote to Σ. [$2z = x, x = 2$]

b) Each line $y = mx$ meets Σ in a point P and meets the circle

$$(x-1)^2 + y^2 = 1$$

and the line $x = 2$ in points Q and R. By considering the x–coordinates of P, Q and R, or otherwise, show that $OP = QR$, where O is the origin. The curve Σ is the cissoid of Diocles.

Draw this cissoid, in the usual way, using a drawn line, drawn circle, and marked origin. For a line L through the origin which meets the given line at R and the given circle at Q, mark on L the point P which satisfies $OP = QR$. Repeat this for a suitable collection of lines through the origin, and draw the curve in the usual way.

$$\left[x \text{ coordinate of } P \text{ is } \frac{2m^2}{1+m^2}\right]$$

16.16. The projective curve Π, given by

$$y^2(z-x) - x^2(z+x) = 0,$$

has as affine view for $z = 1$, the strophoid of Barrow Σ, given by

$$y^2(1-x) - x^2(1+x) = 0.$$

Prove that the origin is a double point of the affine curve Σ and find the tangent lines to the curve there. Show that Π meets the axes $x = 0$, $y = 0$ and $z = 0$ altogether in three points, and find the equations of the four tangents to Π at these three points, including the two at the double point. Show that Σ has an asymptote and write down its equation.

Show that the affine view Γ, for $y = 1$, of the projective curve Π is a graph. Find its asymptotes. Make a rough sketch of Γ indicating the asymptotes and marking the tangent where it crosses the axis $z = 0$. (You may assume that the projective curve Π has an inflexion at the point where it meets the axis $z = 0$.)

16.17. The projective curve Π, given by

$$x^3 + 3xy^2 + x^2z - y^2z = 0,$$

has as affine view for $z = 1$, the folium of Descartes Σ, given by

$$x^3 + 3xy^2 + x^2 - y^2 = 0.$$

Prove that the origin is a double point of the affine curve Σ and find the tangent lines to the curve there. Show that Π meets the axes $x = 0$, $y = 0$ and $z = 0$ altogether in three points, and find the equations of the four tangents to Π at these three points, including the two at the double point. Show that Σ has an asymptote and write down its equation.

Make a rough sketch of the affine view of the projective curve Π for $y = 1$, marking the tangent where it crosses the axis $z = 0$, and the asymptotes for this view. (You may assume that the projective curve Π has an inflexion at the point where it meets the axis $z = 0$.)

$[y = \pm x; y = \pm x, 3x = z, x + z = 0; 3x = 1; x = \pm 1, x + z = 0]$

16.18. Show that the projective curve P with equation

$$x^3 - x^2y - 3x^2z + 7xz^2 - 3yz^2 - 9z^3 = 0$$

has only one singular point and find it. The cubic curve Γ is obtained from P by putting $z = 1$.

Find the equation of the tangent line to Γ at $(0, -3)$ and, by considering its intersection multiplicity with the curve, deduce that the point is a point of inflexion. Similarly show that $(3, 1)$ is a point of inflexion. Deduce that $(-3, -7)$ is the third point of inflexion and determine the slope of the tangent line at this point. Find the asymptote to the curve.
(You may assume that the singular point of P is isolated.)

3) Books-002-2021\ Pappus\
Geometry A Comprehensive
Course.pdf

Pag 363 - The Projective Plane
and Projective Space.

4) \ Pappus\ 2D Coordinate
Geometry.pdf

5) \ Geometry of Curves\
et45643yrt435....pdf

→ Elementary Geometry of
Algebraic Curves
(curves)

Pág. 108(125) - Chapter 9
Projective Algebraic Curves

a) Pág. 112(129) Fig 9.2
Real projective line as
a circle

Fig 9.3 Complex projective
line as a sphere

Fig. 9.4 The projective plane over $\mathbb{Z}_2$

b) Pag. 116(133) Fig. 9.5
Hemisphere model of $P\mathbb{R}^2(\mathbb{R})$

c) Pag. 125(142) - Chapter 10
Singularities of Projective Curves.
d) Pag. 137(154) - Chapter 11
Projective Equivalence
e) Pag. 148(165) - Chapter 12
Projective Tangents
6) Geometry of Curves|
try2y'y643ytr346 -
geometry-of-curves.pdf
a) Pag. 133(152) - Chapter 7
Curvature
b) Pag. 151(170) - Chapter 8
Curvature: applications.
c) Pag. 167(186) - Chapter 9

Circle of Curvature
d) Pág. 297 (316) Chapter 16
Projective Curves
(important)

"Se a geometria do espaço for euclidiana, o volume do universo é infinito; mas se a geometria for esférica, então o volume é finito!"

—

riemannian curvature tensor

—

youtube.com/user/eigenchris/

—

→ BOLOGNA
COSULICH.

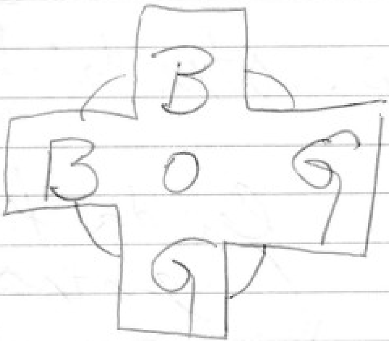

—

embargo manifesto.

—

mp@unibo.it

—

20210126

ressonância orbital TOI-178
18:9:6:4:3

( )

Guataporanga
"Veado Vermelho ou Fera Vermelha"

15 000 ÷ 5,5 = 2,700.000,00

2727272,72

⇒ MoneyGram ($)

Relp. studentionline@unibo.it

Lei da termodinâmica e entropia.
↳ desordem

a entropia só pode aumentar, nunca diminuir.

O tempo só avança na direção em que a entropia aumenta.

Barbour → o que determina a passagem do tempo não é o aumento da entropia,

mas o aumento da complexidade
sem limites de tempo
e espaço.

No Universo, a entropia não
aumenta, mas a complexidade
sim.

Lobachevsky

The God Problem: how a godless
Cosmos creates

Bloom, Howard K 1943-

2012 711 pag.

⇒ Holocausto - Shoá

I don't care that they stole my idea
I care that they don't have any
of their own
Tesla

ALYAH — MITZVA.

Rabini Rony Gurwicz   OLEH

' ( )"
! *+,-* * *. ! * ' ! / 0! 1 2 3
4 / + ! / ! # 5 6 ! 7! + ! ! ! ! 3
7#

8 *+,-* * *. ! * ' ! / 0! 1 5 7 3 !
9 # 2 3 5 6 ! 7! / + ! ! ! ! 3 7#

+,- * *. ! * ' ! / 0! 1 5 # !
7! / " : # ! 2 3 ! # ;! + ! ! ! ! 3
7#

.8 < #! # ; ====

. 7#! #

+,- * *. ! * ' ! 7 ! ! !
7 ; ! # 9 ! ! 2 3 + > ; 3 7#

+,- * *. ! * ' 7 ; ! # ? ; 2 3
> ;) 7#

" *+,-* * *. ! * ' 7 ; ! # 7 ? ; 3
> ; 3) 7#

( )'

20210130 Número.

- N Perfeito: 33550336 (1456 anos).
- N Perfeito: 8589869056
  137438691328

Em 1588 - Pietro Antonio Cataldi.

- Leonard Euler (1772)
  19 dígitos
- 51 números perfeitos
  49.724.095 dígitos

A conjectura sobre os números perfeitos ímpares.

off:

"todos os números perfeitos são pares".

Existe números perfeitos ímpares?

universe expand
unevenly → desigualmente.

20210205

jean@cis.upenn.edu.

Condado Travis, cidade Austin
↳ Samsung. (Travis)! Texas

Data-NovoVivo-004-2021\
Books-004-2021\
Geometrical and visual optics...\
Geometrical and visual optics...pdf

Pág. 106 (119)

Figure 7-8 A e B

$n' = 1.333$

estudar!!!

(veja curvature)

Obs: $F = \frac{n'}{f'}$ Lens Effectivity

20210205

jean@cis.upenn.edu.

Condado Travis, cidade Austin
(Travis)! Texas
↳ samsung.

Data-NovoVivo-004-2021\
Books-004-2021\
Geometrical and visual optics...\
Geometrical and visual optics...pdf

Pag. 106 (119)

Figure 7-8 A e B

$n' = 1.333$

estudar!!!

(veja curvature)

Obs: $F = \frac{n'}{f'}$ Lens Effectivity

@ (
+,- * *. ! * @ 2 ! ! # ! .7 / " #
<# 7# *+,-* * *. ! * @ 2 ! ! # !
.7 / " # <# 7#
( ) 9 % + A

In summary, to correct hyperopia, the secondary focal point of a positive correcting lens must be coincident with the eye's far point. When this is the case, an infinitely distant object is imaged at the far point. Since the far point is conjugate with the retina, an image is focused on the retina.

LENS EFFECTIVITY

Up to now, we've assumed the lens used to correct ametropia is positioned at the eye's anterior surface. This would be the case for a contact lens or corneal refractive surgery, but not for a spectacle lens, which is positioned at a significant distance in front of the eye. The distance between the cornea and the ocular (back) surface of a spectacle lens is referred to as the **vertex distance**.

Let's consider a myopic eye that can be corrected with a −6.68 D lens placed in the plane of the cornea. If we wish to correct the refractive error with a spectacle lens located 15.00 mm in front of the eye, what power spectacle lens is required?

The key to answering this question is to realize the far point of the myopic eye is at a fixed distance from the eye. This is illustrated in Figure 7-8. To correct the refractive error, the secondary focal point of the correcting lens must be coincident with this far point. In the case at hand, the far point is 14.97 cm anterior to the cornea. If the correcting lens is 15.00 mm anterior to the cornea, its secondary focal length must be −13.47 cm [(−14.97 cm) − (−1.50 cm) = −13.47 cm]. The power of a minus lens with a secondary focal length of −13.47 cm is

$$F = \frac{n'}{f'}$$

$$F = \frac{(100)\,(1.00)}{-\,13.47\ \text{cm}}$$

$$F = -\,7.42\ \text{D}$$

This myopic eye can be corrected with either a −6.68 D contact lens or a −7.42 D spectacle lens at a vertex distance of 15.00 mm. Both correct the patient's myopia and are said to have the same *effective* power. Note when we correct the patient's myopia with a spectacle lens, its absolute power is greater than that of the contact lens. Think of it this way: since the spectacle lens is closer to the far point, it must have a shorter focal length.

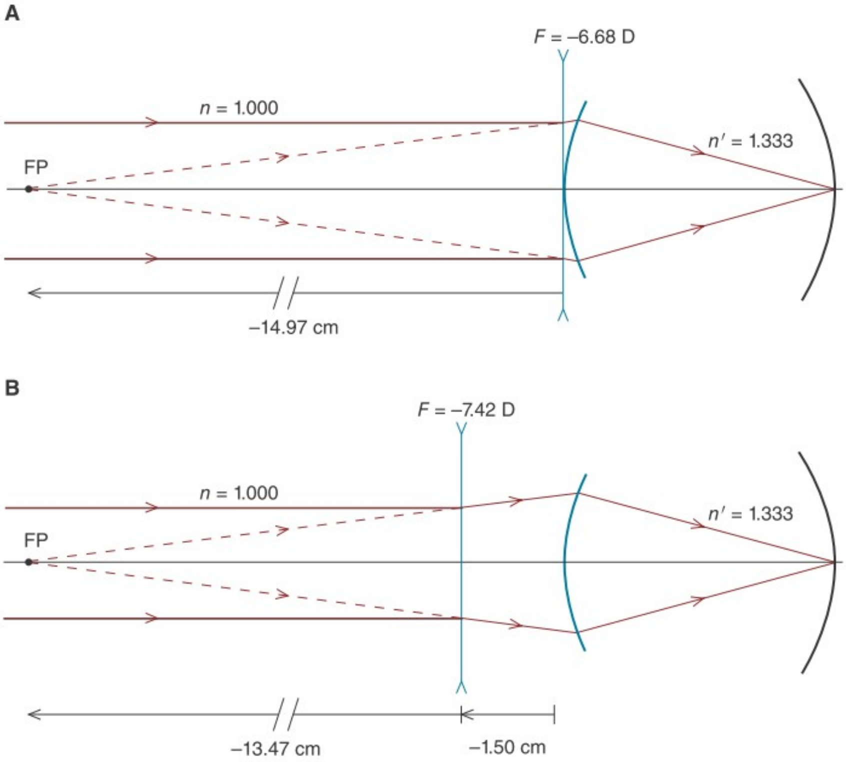

Figure 7-8. This refractive error can be corrected with (**A**) a −6.68 D contact lens or (**B**) a −7.42 D spectacle lens. In both cases, an infinitely distant object is focused at the eye's far point.

Clinical Highlight

Understanding lens effectivity is critical to understanding and solving common clinical symptoms. For instance, patients with myopia sometimes report blurred distant objects become clearer when they push their spectacles very close to their eyes. How can we explain this? Most likely, the corrective lenses are not sufficiently strong. As illustrated in Figure 7-9, by pushing the lenses closer to the eyes, the secondary focal points of the corrective lenses are made coincident with the far points, thereby increasing lens effectivity.

The principles of lens effectivity also apply to hyperopia. *Consider the 2.82 D hyperope we discussed earlier in the chapter. This hyperopic patient requires a corrective lens of +2.82 D in the plane of the cornea (i.e., a contact lens whose power is +2.82D). What power spectacle lens is required to correct this refractive error? Assume a vertex distance of 15.00 mm.*

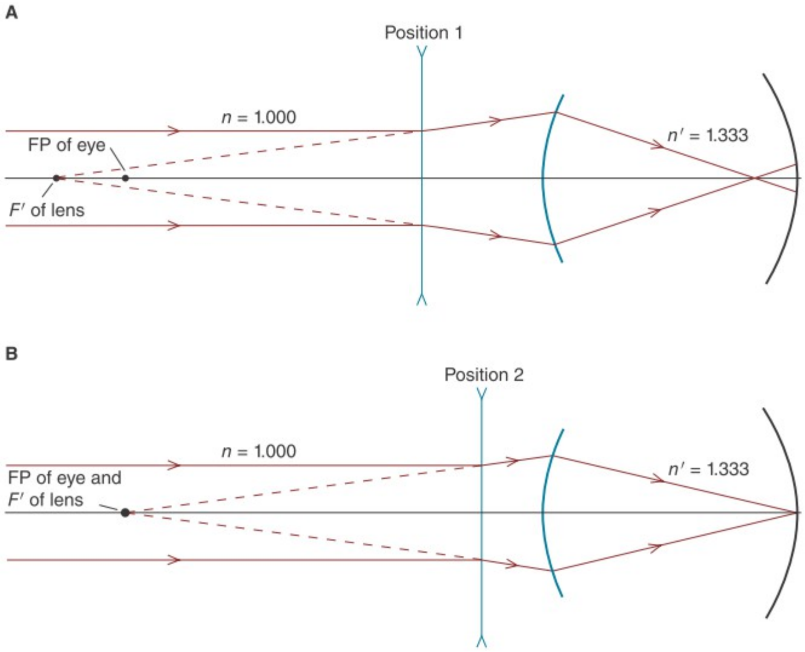

Figure 7-9. The effective power of a spectacle lens used to correct myopia is increased by moving the lens toward the eye [from position 1 in (**A**) to position 2 in (**B**)].

The secondary focal point of the corrective lens, whether it is a contact lens or a spectacle lens, must be coincident with the far point of the eye (which is 35.46 cm to the right of the cornea). From Figure 7-10, we see the secondary focal length of the spectacle lens must be 36.96 cm (35.46 + 1.50 cm = 36.96 cm). The power of this spectacle lens is

$$F = \frac{n'}{f'}$$

$$F = \frac{(100)\,(1.00)}{36.96\ \text{cm}}$$

$$F = +\,2.71\ \text{D}$$

This hyperope can be corrected with either a +2.82 D contact lens or a +2.71 D spectacle lens. The contact lens has a higher power because it must have a shorter focal length if its secondary focal point is to be coincident with the eye's far point. In this example, the difference between the power of the contact lens and the spectacle lens is small. For higher powers, however, lens effectivity is an important clinical factor. For instance, a patient who is corrected with a +10.00 D contact lens requires a spectacle lens whose power is +8.70 D.

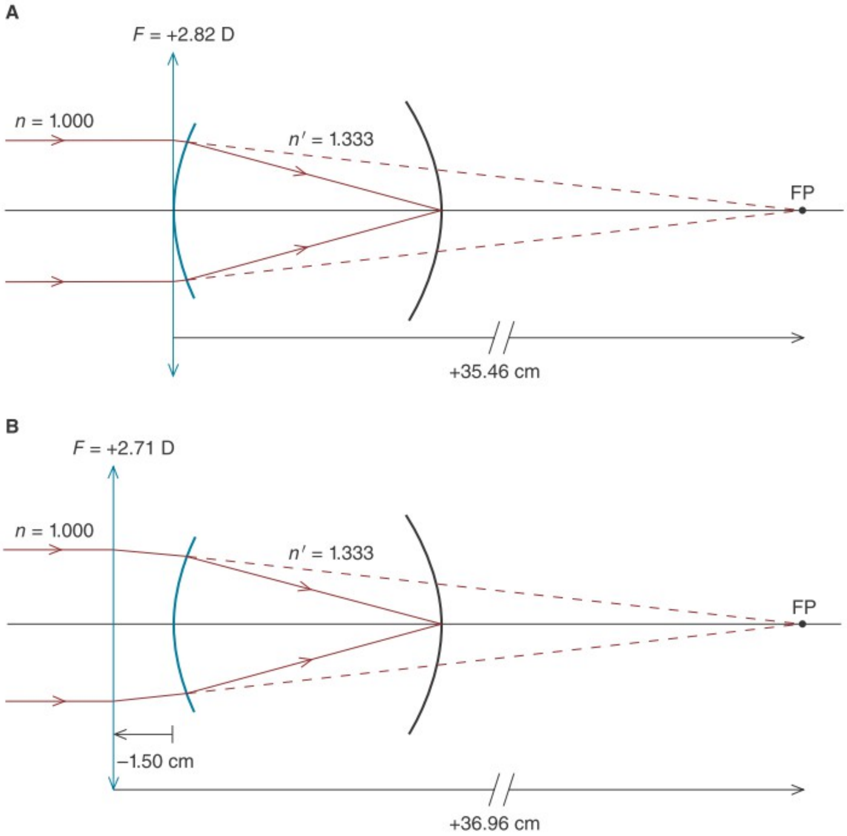

Figure 7-10. This refractive error can be corrected with (**A**) a +2.82 D contact lens or (**B**) a +2.71 D spectacle lens. In both cases, an infinitely distant object is focused at the eye's far point. Although lens effectivity is relatively small with this low power prescription, it can be significant with higher power lenses.

Clinical Highlight

Hyperopic patients sometimes notice their distance visual acuity improves when they slide their spectacles down their nose so they are farther from their eyes. What is happening here? Most likely, the lenses—when worn at the normal distance—are not sufficiently strong. By moving the lenses away from his eyes, the hyperopic patient is moving the secondary focal points of the spectacle lenses closer to his far points (Fig. 7-11). Rather than using a stronger lens, the same effect can be obtained by moving a lens away from the cornea, thereby increasing its effectivity.

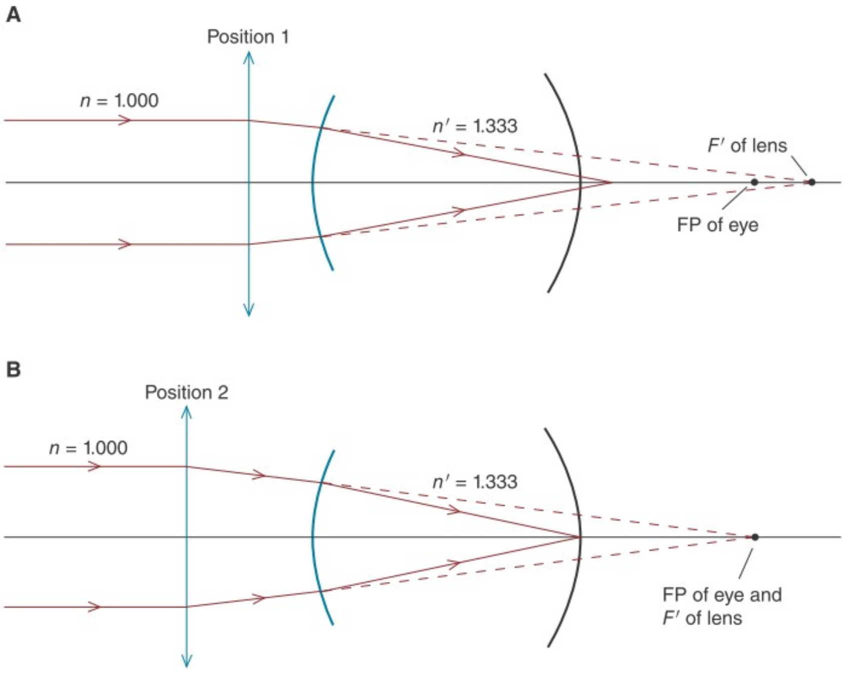

Figure 7-11. The effective power of a plus spectacle lens that is used to correct hyperopia is increased by moving the lens away from the eye [from position 1 in (**A**) to position 2 in (**B**)].

Up to now, we've discussed examples that involve replacing a contact lens with a spectacle lens. The same principles apply to replacing a spectacle lens with a contact lens. As a practical matter, refractions are almost always performed at the spectacle plane, about 15.00 mm in front of the cornea. Therefore, it is more common to calculate the contact lens power from the spectacle power than vice versa.

Let's look at an example. *Your refraction at a vertex distance of 15.00 mm results in a spectacle prescription of −9.00 D in each eye. If the patient wishes to be fit with contact lenses, what power should they be?*

First, without doing any calculations, you should know that the power of the contact lenses will be less than the spectacle lenses. Since the contact lenses are closer to the eyes, they are farther from the far point and consequently have a longer focal length and lower power.[8] To be specific, whereas the secondary focal length for the spectacles is −11.11 cm (e.g., 100/−9.00 D = −11.11 cm), the secondary focal length for the contact lenses is −12.61 cm [e.g., (−11.11 cm) + (−1.5 cm) = −12.61 cm]. A −7.93 D contact lens has this focal length (e.g., 100/−12.61 cm = −7.93 D).

Just to be sure we have these important concepts down, let's do one more example. *If we perform a trial-lens refraction at a distance of 12 mm and find +12.00 D, what power contact lens should be prescribed?*

Before we do any calculations, we know that the contact lens must have a shorter focal length than the spectacle lens, which means it has more power. Rather than a focal length of 8.33 cm (for the spectacle lens), the contact lens focal length is 1.2 cm shorter, or 7.13 cm. Such a contact lens has a dioptric power of +14.02 D.

Lens effectivity can also be calculated using equations. Although some people prefer to use equations, I think it's better to solve these problems without them. It's all too easy to forget equations, but once you understand the basic concepts of lens effectivity, they should stay with you over the course of your professional career.

CORRECTION OF AMETROPIA WITH LASER AND SURGICAL PROCEDURES

Ametropia is most often corrected with an ophthalmic lens positioned in the spectacle plane or a contact lens in the corneal plane. In recent years, it has become commonplace to compensate (or partially compensate) for refractive errors with surgical and laser procedures.

The first widely utilized surgical procedure was *radial keratotomy* (RK). This procedure was used primarily to compensate for myopia. Radial incisions are made in the cornea, leading to the flattening of this tissue and a resultant decrease in the eye's refractive power (Fig. 7-12).

Photorefractive procedures that utilize excimer laser technology have supplanted radial keratotomy. In *photorefractive keratectomy* (PRK), the corneal epithelium is removed prior to using an excimer laser to sculpt the underlying stroma. With *laser-assisted in situ keratomileusis* (LASIK), a microtome or laser is used to create a corneal flap (Fig. 7-13). The exposed stroma is sculpted with the goal of adjusting the refractive power of the eye and reducing the amount of

---

[8] When we say that one corrective lens is stronger (or more powerful) than another, we are typically referring to the absolute power of the lenses. That is, a −10.00 D lens is said to be stronger than a −5.00 D lens.

Pág 102 (115)

Hyperopia

$L' = \frac{n'}{l'}$

$L' = \frac{(1000)(1.333)}{2122\ \text{mm}}$

constant

por qué?

$L' = +62.82\ D$

JEE Maths - Playlist - JEE Main 2021

96 USA
1 HK.

103 USA
1 HK

VILA Antonina

630

Anhangabaú

20210215

23h 56min 4,09053 seg

40.075 km (circunferência)

460 m/s = 1.656 km/h (rotação)

365 d 6 h 9 min 10 seg (translação).

940 milhões km/ano

30 km/s = 108000 km/h.

Sistema Solar = 220 km/s = 792.000 km/h

Ano Galáctico = 225.000.000 anos.

Via Láctea cruzando o Universo
em direção das constelações de
Leão e da Virgem

navegamos 630 Km/s

2.200.000 Km/h

---

460 m/s — 0,46
30 Km/s — 30
220 Km/s — 220
630 Km/s — 630

880,46 Km/s.

The Mission-Severine

⇒ https://admin.google.com/u/1/

Cidade Jardim 19.973
Parque Ibirapuera 22.500
Ibirapuera 23.000 preço

---

gsuite.google.com/signup/essentials/welcome
/businessstarter/
welcome

Av Cambucha

Remark: One frequently encounters the expression for $\tilde{E}_v$ with the first term missing. This is appropriate if the energy taken with respect to the ground state $n_j = 0$, in which case the numerator in the expression for q_v is absent.

(d) Last, we consider electronic degrees of freedom. Here no closed formulation is available; one simply begins with the listing of the energies of the molecular electronic states. The electronic ground state is set at energy zero; relative to this, the electronic energy levels are specified as

$$\varepsilon_e = g_{e0} + g_{e1}\varepsilon_{e1} + g_{e2}\varepsilon_{e2} + \ldots. \tag{1.12.48}$$

Normally, the excited electronic energy levels are very large compared to k_BT, whence it is usually sufficient to set the electronic molecular partition function as

$$q_e = g_{e0} + g_{e1}e^{-\varepsilon_1/k_BT} + g_{e2}e^{-\varepsilon_2/k_BT}, \tag{1.12.49}$$

and then use Eq. (1.12.49), as is, to determine how these electronic degrees of freedom contribute to the various thermodynamic functions of state as specified in Chapter 10.

Intervening Remark

A note of caution: The above exposition on the application of statistical mechanics is very rudimentary. While conveying the essence of the approach, the derivations apply only to highly idealized cases. We have not considered interactions among gaseous systems, rotation–vibration effects, anharmonicities in vibrations, electron interactions, and the like, which are needed for a realistic description of physical properties of actual systems. Such topics are handled in more advanced and specialized treatments in numerous textbooks on statistical mechanics, and fall outside the limited scope of the present discussion. Just be aware of the limitations inherent in the above treatment.

Statistical Treatment of Nonmetallic Solids

Nonmetallic solids will be considered as consisting of a great number of monatomic constituents held rigidly together as a lattice, and executing harmonic vibrations within their confined spaces. We neglect any internal degrees of freedom and ignore the six translational and rotational degrees of freedom of the entire rigid body, leaving us with $3N - 6 \approx 3N$ vibrational coordinates that are inserted in the partition function for the solid.

As an introduction, rewrite the expression for the molar energy of harmonic oscillators, cited above, to specify the average molar energy of a set of oscillators vibrating at a single frequency ν:

$$\tilde{E} = \tilde{E}_0 + 3R\frac{\Theta_v}{e^{\Theta_v/T} - 1}; \quad \tilde{E}_0 = \frac{3}{2}N_A h\nu, \quad \Theta_v \equiv h\nu/k_B, \tag{1.12.50}$$

where N_A is the Avogadro's number. Then the heat capacity is given by

$$\tilde{C}_V = 3R\frac{(\Theta_v/T)^2 e^{\Theta_v/T}}{\left(e^{\Theta_v/T} - 1\right)^2}. \tag{1.12.51}$$

This relation, commonly known as Einstein's (1905) law of heat capacities, correctly corresponds to the limiting cases $\tilde{C}_V = 0$ at $T = 0$ and to $\tilde{C}_V = 3R$ at infinite temperature—the law of Dulong and Petit. It immediately rationalizes the deviations of the observed $\tilde{C}_V$ values from this law at lower temperatures that had mystified scientists before the advent of quantum mechanics. However, the above equation generally deviates to some extent from experiment; in particular, at low T, the above relation shows that $\tilde{C}_V \sim T^{-2}e^{-\Theta_v/T}$, while experiments conform to a $\tilde{C}_V \sim T^3$ variation.

To accommodate these findings, Debye (1912) eliminated the unrealistic assumption that all atoms vibrate at the same fixed frequency. Since the N atoms on the crystal form a coupled system, one anticipates a spread of oscillator frequencies that needs to be introduced. For this purpose, Debye adopted the theory of a homogeneous elastic continuum, in which c_l and c_t are the velocities of one longitudinal and two transverse waves propagating in a volume V. As is well established (a euphemism for: "Everyone knows that…"), the number of frequencies in the range ν to $\nu + d\nu$ as given by the theory of elasticity is specified by

$$dZ = 4\pi V\left(\frac{1}{c_l^3} + \frac{2}{c_t^3}\right)\nu^2 d\nu. \tag{1.12.52}$$

The total number of available frequencies is $3N$, corresponding to a range of frequencies $0 \le \nu \le \nu_0$. Then

$$Z = 3N = \frac{4}{3}\pi V\left(\frac{1}{c_l^3} + \frac{2}{c_t^3}\right)\nu_0^3. \tag{1.12.53}$$

On combining Eqs. (1.12.52) and (1.12.53), we find that

$$dZ = \frac{9N}{\nu_0^3}\nu^2 d\nu, \tag{1.12.54}$$

which will now be used as the applicable distribution function for available frequencies. Then the energy of 1 mol of solid is specified by

$$\tilde{E} - \tilde{E}_0 = \frac{9R}{\nu_0^3 k_B}\int_0^{\nu_0} \frac{h\nu^3 d\nu}{e^{h\nu/k_BT} - 1} = \frac{9RT^4}{\Theta_v^3}\int_0^{\Theta_v/T} \frac{y^3 dy}{e^y - 1} = \frac{9RT}{x_D^3}\int_0^{x_D} \frac{y^3 dy}{e^y - 1}, \tag{1.12.55}$$

where $y \equiv h\nu/k_BT, x_D \equiv h\nu_0/k_BT$. Hence, the heat capacity is found to be

$$\frac{\tilde{C}_V}{3R} = \frac{12}{x_D^3}\int_0^{x_D} \frac{y^3 dy}{e^y - 1} - \frac{3x_D}{e^{x_D} - 1}. \tag{1.12.56}$$

The fraction in the integral can be expanded in an infinite series before the integration is attempted. The detailed execution, which involves some mathematical acrobatics, leads to the approximate result

$$\frac{\tilde{C}_V}{3R} = \frac{4}{5}\frac{\pi^4}{x_D^3} - \frac{3x_D}{e^{x_D} - 1} + \text{Correction Terms}. \tag{1.12.57}$$

At low temperatures, hence, large x_D, only the first term is of importance, which shows that the heat capacity varies as $\tilde{C}_V \sim T^3$, in good agreement with experiment.

Despite the cobbling together of elements of classical and quantum statistics, the above theory is consistently used in the analysis of data, as manifested by the empirical determination of Θ_D values for a large variety of materials. One of the saving graces is the fact that the distribution function appears under the integral sign. Hence, even considerable deviations of the real from the posited distribution function can be tolerated, without greatly changing the value of the integral in Eq. (1.12.56).

Electronic Heat Capacity of Metals

We consider the heat capacity attributed to N electrons circulating freely among N atoms that constitute a regular array of a solid. They are distributed among the N energy states of a partially filled energy band in accord with the Fermi–Dirac theory (Section 10.1) which is characterized by the Fermi energy ε_F. For a brief review of essentials, see the appendix below. Thus, at low temperatures, $k_BT \ll \varepsilon_F$, the energy required to heat N electrons from absolute zero to temperature T is given by (almost all the action involves electrons with energy at or close to ε_F)

$$E_{el}(T) = \int_{\varepsilon_F}^{\infty} d\varepsilon(\varepsilon - \varepsilon_F)f(\varepsilon;T)\rho(\varepsilon) + \int_{0}^{\varepsilon_F} d\varepsilon(\varepsilon_F - \varepsilon)[1 - f(\varepsilon;T)]\rho(\varepsilon). \tag{1.12.58}$$

Here $\rho(\varepsilon)$ is the density of states (DOS) in the range $d\varepsilon$ about ε; f is the Fermi function derived in Section 10.1. The first integral represents the energy required to elevate electrons at the Fermi level to the $f(\varepsilon;T)\rho(\varepsilon)$ available energy states per unit volume at an energy $\varepsilon > \varepsilon_F$; the upper band edge energy at $W/2$ has been replaced by infinity because $\rho(\varepsilon)$ vanishes over that range. The second integral similarly shows the energy required to remove $[1 - f(\varepsilon;T)]\rho(\varepsilon)$ electrons per unit volume from energy $\varepsilon < \varepsilon_F$; to the Fermi level; the lower band edge energy at $-W/2$ has been replaced by zero, because $\rho(\varepsilon)$ vanishes in that range.

From the above, we obtain the electronic heat capacity as

$$C_{el} = (\partial E_{el}/\partial T) = \int_{0}^{\infty} d\varepsilon(\varepsilon - \varepsilon_F)(\partial f/\partial T)(\rho(\varepsilon)). \tag{1.12.59}$$

At low temperatures, $(\partial f/\partial T)$ differs from zero only for energies very close to ε_F; we may thus introduce the approximation

$$C_{el} = \rho(\varepsilon_F)\int_{0}^{\infty} d\varepsilon(\varepsilon - \varepsilon_F)(\partial f/\partial T), \tag{1.12.60}$$

where

$$\frac{\partial f}{\partial T} = \frac{\varepsilon - \varepsilon_F}{k_BT^2}\,\frac{e^{(\varepsilon-\varepsilon_F)/k_BT}}{\left[e^{(\varepsilon-\varepsilon_F)/k_BT} + 1\right]^2}. \tag{1.12.61}$$

Gielis curves Source: the Geometrical Beauty of Plants.

—

01 → 23, 24, 25, 26 e 27
02 → 02, 03, 04, 05
14, 15, 16, ...

=

Books-004-2021/Thermodynamics Principles/0128219408.pdf

Pág. 86 (96) -

Statistical Treatment of Nonmetallic Solids

... This relation, commonly known as Einstein's (1905) law of heat capacities, correctly corresponds to the limiting cases

$C_v = 0$ at $T = 0$ and to

$C_v = 3R$ at infinite temperature – the law of Dulong and Petit.

(%
+,- * *. ! * B!C#! 8 ;! !
, ! ! ! # + ! ! ! !)%@ 7#
( )

Notation

All manifolds are assumed to be smooth and connected. The notation we use is generally standard. $C^{k+\alpha}$, $0 < \alpha < 1$, is the space of function whose kth derivatives satisfy a Hölder condition with exponent α, C^{ω} is the space of real analytic functions, and the Sobolev space $H_0^1(\Omega)$ is the completion of smooth functions with compact support in Ω in the norm $(\int_{\Omega}|\nabla u|^2)^{1/2}$. Our Laplacian, Δ, on real-valued functions has the sign so that $\Delta u = +u''$ on $\mathbf{R}^1$. The Ricci curvature of the standard unit sphere S^n with metric g in $\mathbf{R}^{n+1}$ is $(n-1)g$, and its scalar curvature is $n(n-1)$. If α is a one-form on a Riemannian manifold, then $\alpha^{\#}$ is the dual vector field. In classical tensor notation we use a semicolon to denote covariant differentiation. ω_n is the volume of the unit sphere S^n in $\mathbf{R}^{n+1}$.

I. Gaussian Curvature

1. Surfaces in $\mathbf{R}^3$. As sometimes occurs, one of the most interesting unresolved problems involving Gaussian curvature is also one of the easiest to understand. Let $z = u(x, y)$ be a surface in $\mathbf{R}^3$. Its Gaussian curvature $K(x, y)$ is given by the formula

$$\frac{u_{xx}u_{yy} - u_{xy}^2}{\left(1 + |\nabla u|^2\right)^2} = K(x, y). \tag{1.1}$$

Turn this around and let K be a given C^∞ function of x and y. Is there a surface $z = u(x, y)$ with curvature K? Thus one wants to solve (1.1) for u. The most modest question is to at least find this surface locally, say in some neighborhood of the origin. If K is a real analytic function, then a local solution exists by the Cauchy–Kovalevsky theorem. If $K(0,0) > 0$, then (1.1) is elliptic at the origin, while if $K(0,0) < 0$, then it is hyperbolic there. In either case standard machinery allows one to conclude that there is a solution in some neighborhood of the origin. The case $K(0,0) = 0$ is not yet fully understood. If $K(x, y) \geqslant 0$ or $\nabla K(0,0) \neq 0$, then Lin **[Li]** has recently shown that (1.1) is locally solvable, using the Nash–Moser implicit function theorem. Nothing more is known. It is even quite possible that there are some functions $K \in C^\infty$ for which (1.1) has no solution in any neighborhood of the origin—just as for the classical Lewy example. Equations such as (1.1) that involve the determinant of the hessian of a function are usually called *Monge–Ampère* equations.

The problem of solving (1.1) is closely related to the problem of realizing an abstract Riemannian metric $ds^2 = E\,d\xi^2 + 2F\,d\xi\,d\eta + G\,d\eta^2$ on $\mathbf{R}^2$ locally as the graph of a smooth surface $z = z(x, y)$ in $\mathbf{R}^3$, so then

$$ds^2 = dx^2 + dy^2 + dz^2. \tag{1.2}$$

Rewrite this as $ds^2 - dz^2 = dx^2 + dy^2$. Since $dx^2 + dy^2$ is the standard flat metric on $\mathbf{R}^2$, one seeks $z(\xi, \eta)$ with $z(0,0) = 0$, $\nabla z(0,0) = 0$ so that the metric $g = ds^2 - dz^2$ is flat, i.e., has zero Gaussian curvature in some neighborhood of the origin. This leads to a Monge–Ampère equation similar to (1.1), with similar difficulty if the Gaussian curvature of ds^2 is zero. If one can find $z(\xi, \eta)$, then since g is flat an easy geometric argument shows that in geodesic coordinates we can write $g = dx^2 + dy^2$. In these new coordinates $z(x, y)$ satisfies (1.2). It is likely that a good understanding of (1.1) will at the same time resolve the question of whether one can always locally isometrically embed a 2-manifold in $\mathbf{R}^3$.

Once one understands the local solvability of (1.1), it is natural to seek a global solution in a bounded open set in $\mathbf{R}^2$, possibly with boundary conditions. We will discuss this later in Chapter 4 under the special assumption that $K > 0$, so one can use techniques for elliptic equations. There seems to be little information if $K \leqslant 0$ (see [**Az**]).

In the early years of this century, Minkowski and Weyl posed the following problems:

Minkowski problem. Prove the existence of a compact convex surface in $\mathbf{R}^3$ whose Gaussian curvature is prescribed as a given positive function of the unit normal.

Weyl problem. Let g be a given metric on the sphere S^2 with positive Gaussian curvature K. Prove the existence of a compact convex surface in $\mathbf{R}^3$ with g as its metric, i.e., isometrically embed (S^2, g) in $\mathbf{R}^3$ assuming $K > 0$.

Both of these problems lead to nonlinear elliptic partial differential equations. They were eventually solved by Nirenberg [**N**] and Pogorelov [**P1**]. More recently, a higher-dimensional version of the Minkowski problem has also been solved (see [**P2**]), but there is no information on any such version of the Weyl problem; for instance, if (S^n, g) has positive sectional curvature, can it be embedded in $\mathbf{R}^N$ for some "good" N depending on n?

A related question was asked by Oliker [**O**]. Let $\Sigma^n \hookrightarrow \mathbf{R}^{n+1}$ be a smooth compact graph over the standard unit sphere $S^n \hookrightarrow \mathbf{R}^{n+1}$, so if $x \in S^n$ then the points y on Σ^n can be written as a radial graph $y = u(x)x$, where u is some positive function defined on S^n (and can be extended to $\mathbf{R}^{n+1} - \{0\}$ to be constant along radii). The Gauss–Kronecker curvature K of Σ^n is given by a Monge–Ampère equation involving the determinant of the hessian of u on S^n. Conversely, given a positive function $K(|x|)$ on $\mathbf{R}^{n+1} - \{0\}$, Oliker asked if there is a closed convex surface $\Sigma^n \hookrightarrow \mathbf{R}^{n+1}$ over S^n whose Gauss–Kronecker curvature is K. Combining Oliker's [**O**] work with a significant improvement by Delanoe [**Del**], the result is that there is at least one such hypersurface if K satisfies the additional assumption that there are numbers $0 < R_1 \leqslant R_2$ such that for all x on S^n

$$\inf K(R_1 x) \geqslant R_1^{-n} \quad \text{and} \quad \sup K(R_2 x) \leqslant R_2^{-n}. \tag{1.3}$$

Moreover, the proof gives a hypersurface Σ^n in the shell between the spheres of radii R_1 and R_2. If assumption (1.3) is satisfied for some other numbers $R_2 < R_1' \leqslant R_2'$, then of course one finds another solution, so the solution is not unique. Oliker gives some additional assumption guaranteeing the uniqueness—up to homothety—of the hypersurface. Oliker's existence proof of a solution of the Monge–Ampère equation uses the continuity method, while Delanoe's uses the Schauder fixed point theorem.

Treibergs and Wei **[TW, T]** discuss similar problems for the mean curvature, so they are led to a quasi-linear elliptic equation.

2. Prescribing the curvature form on a surface. If (M^2, g) is a compact two-dimensional Riemannian manifold without boundary, then the Gauss–Bonnet theorem asserts that

$$\int_M K_g \, dA_g = 2\pi\chi(M). \tag{1.4}$$

If one thinks of $\Omega = K\,dA$ as the curvature two-form, then one type of converse to Gauss–Bonnet is, given any two-form Ω satisfying $\int_M \Omega = 2\pi\chi(M)$, find a metric g so that $\Omega = K\,dA$. This was solved by Wallach and Warner **[WW]**, who gave the following simple proof. Fix a metric g_0 on M and seek g *pointwise conformal* to g_0, that is,

$$g = e^{2u} g_0 \tag{1.5}$$

for some function u to be determined. Using classical formulas one finds that

$$\Omega = K\,dA = (K_0 - \Delta_0 u)\,dA_0 = \Omega_0 - (\Delta_0 u)\,dA_0.$$

Thus one wants to solve the linear equation $(\Delta_0 u)\,dA_0 = \Omega_0 - \Omega$, which we can rewrite as

$$\Delta_0 u = *(\Omega_0 - \Omega), \tag{1.6}$$

where $*$ is the Hodge operator in the g_0 metric. Since

$$\int_M (\Omega_0 - \Omega) = 0, \tag{1.7}$$

by linear elliptic theory, equation (1.6) has a solution u which is unique up to an additive constant. This solves the problem.

We do not believe anyone has solved the analogous problem on compact surfaces with boundary, or the more difficult question in higher dimensions (except in the case of Kähler manifolds, in which case this is the Calabi problem to be discussed in Chapter 3).

3. Prescribing the Gaussian curvature on a surface. For a surface M^2 there are two types of questions:

1. Find a metric g with constant Gaussian curvature. Here—and in similar higher-dimensional problems—the point is to obtain metrics with special properties which often help clarify the geometry of M.

2 (Inverse problem). Given a function K, find a metric g whose Gaussian curvature is K.

If the manifold M is not compact, one may wish to require that the metric be complete, while if M is compact with boundary, then we may impose some boundary conditions.

In practice, one often seeks the unknown metric g by picking some metric g_0 and then somehow deforming it to the desired metric g. The simplest type of deformation is the pointwise conformal deformation (1.5). From a geometric viewpoint, this does not give much flexibility since there is only one unknown function u while on an n-dimensional manifold the metric is locally an $n \times n$ symmetric matrix and hence has $n(n+1)/2$ components. Nonetheless, these deformations are surprisingly useful—and their geometric interpretation is often significant for geometric and physical problems.

Throughout these lectures we shall see several other effective ways to deform a metric, but much work remains to be done on this issue.

(a) *Compact surfaces.* One aspect of the two above problems, which may be viewed as proving the existence of a solution of a certain partial differential equation such as (1.1) or (1.6), is first finding the obstructions there may be to the existence of a solution. For instance, the Gauss–Bonnet theorem (1.4) gives some topological obstructions to the existence of certain metrics. We used this in (1.7). For the "inverse problem" on the torus T^2, since $\chi(T^2) = 0$ there is no metric with $K > 0$ (or $K < 0$) everywhere, since (1.4) implies that either $K \equiv 0$ or else K changes sign. There is a similar obvious sign condition on K for other Euler characteristics.

THEOREM 1.8 (KAZDAN–WARNER). *On a compact M^2, a function $K \in C^\infty(M^2)$ is the Gaussian curvature of some metric g if and only if K satisfies the obvious Gauss–Bonnet sign condition.*

There are two proofs of this. To begin with, for any metric g there is a complicated formula for the Gaussian curvature K in terms of the first two derivatives of g. We write this briefly as

$$F(g) = K. \tag{1.9}$$

Thus, given K, (1.9) is a quasilinear partial differential equation for g. The shortest proof **[KW5]** begins with a known metric g_0 with constant curvature K_0, so $F(g_0) = K_0$, and then appeals to the inverse function theorem to solve (1.9) for K near K_0 in an appropriate topology. Next one observes that if "K near K_0" means in the L_p norm, then given any function K with $K(x_0) = K_0$ for some x_0, there is a diffeomorphism ϕ of M so that $|K \circ \phi - K_0| < \varepsilon$ in L_p (the diffeomorphism just spreads a neighborhood of x_0 over most of M). Thus one can solve

$$F(g_1) = K \circ \phi. \tag{1.10}$$

But now the metric $g = \phi^{*-1}(g_1)$ satisfies (1.9) (i.e. from the viewpoint of M, both K and $K \circ \phi$ are the same functions). The only difficulty in carrying out this program is that the inverse function theorem does not immediately apply in some standard cases, namely on S^2 and T^2; but this can be circumvented by additional technical devices.

The second proof is geometrically richer. To solve (1.9) one picks some metric g_0 on M and seeks the desired metric g pointwise conformal to g_0, so $g = e^{2u}g_0$ for some function u to be found in order to satisfy (1.9), that is, $F(e^{2u}g_0) = K$.

This last equation is exactly

$$\Delta_0 u = K_0 - Ke^{2u}, \tag{1.11}$$

where Δ_0 and K_0 are the Laplacian and Gaussian curvature of g_0. If one wishes, one can ignore the geometric origins of this equation and ask for which functions K_0 and K can one solve it on any compact manifold M^n of any dimension. It turns out that the theory of this equation is surprisingly intricate—with many cases still not understood. We shall only summarize briefly (see **[KW1]** for details).

The simplest case is if both $K_0 < 0$ and $K < 0$; then (1.11) has a unique solution on M^n for any n. This is easily proved using sub and super solutions.

If $K_0 = 0$, then there are two necessary conditions on K ($\not\equiv 0$) for the solvability:

(i) K changes sign,

(ii) $\int K \, dA_0 < 0$.

If $\dim M \leqslant 2$, these conditions are also sufficient. The existence proof uses the calculus of variations and breaks down if $\dim M \geqslant 3$, where no sufficient conditions are known, despite the fact that the equation $\Delta_0 u = -Ke^{2u}$ looks so simple.

Finally, if $K_0 > 0$, even less is known. Only the case where $\dim M = 1$ is clearly understood (then one can solve (1.11) if and only if K is positive somewhere). If $\dim M = 2$, then one can solve (1.10) if K_0 is sufficiently small—by the calculus of variations. Only for the special case of $M = S^2$ or P^2 with their canonical metrics is a little more known. Say $K_0 = c > 0$ is a constant. Then Moser **[M]** used the calculus of variations to prove that on S^2 (resp. P^2) if $0 < c < 1$ (resp. $0 < c < 2$), then there is a solution if and only if K is positive somewhere. Since $K_0 = c = 1$ on S^2 and P^2 with their standard metrics, this proves Theorem 1.8 on P^2, but just misses on S^2. Indeed, if $c \geqslant 1$, there is a *nonexistence* result on S^2 which is implied by the following identity that every solution of $\Delta_0 u = c - Ke^{2u}$ must satisfy:

$$\int_{S^2} X(K) e^{2u} \, dA_0 = (c-1) \int_{S^2} (\operatorname{div} X) K e^{2u} \, dA_0, \tag{1.12}$$

where X is in the six-dimensional space of conformal vector fields (i.e. the Lie algebra of the group of conformal maps) on S^2. The special case where X is in the three-dimensional space of gradients of first-order spherical harmonics was proved by Kazdan and Warner **[KW1]**. Later, Bourguignon and Ezin **[BE]** observed that one should add the other conformal vector fields, i.e., the three-dimensional orthogonal group $O(3)$. In particular, if $c = 1$ and $K = \psi$ is a first-order spherical harmonic (plus a constant if one wishes), then (1.12) is not satisfied with $X = \nabla\psi$ because then $X(K) = |\nabla\psi|^2$ so the left side is positive while the right side is zero.

It is not known if the existence of a function u satisfying (1.12) is sufficient for the solvability of (1.11) on S^2. In fact, this is not even known in the special case where K is rotationally symmetric and one seeks a rotationally symmetric solution (in this special case (1.11) reduces to an ordinary differential equation). As we shall see later, similar issues arise in other problems.

Berger [**B**] has suggested using (1.11) to give a proof of the uniformization theorem for Riemann surfaces: one is given g_0 on M and seeks g with $K \equiv$ constant. If $\chi(M) \leqslant 0$, then the existence of g is a consequence of the above discussion, while if $M = S^2$ there is still no direct proof using (1.11). In the S^2 case, the obstruction (1.12) causes difficulty. One would also like to know that if g_0 (and K) are invariant under some group action, then there is an invariant solution u; this is true for (1.11) if both $K_0 < 0$ and $K < 0$ since the solution is unique in this case.

Despite these difficulties in solving (1.11), one can still use it to prove Theorem 1.8, the idea being to solve

$$\Delta_0 u = K_0 - (K \circ \phi) e^{2u} \tag{1.13}$$

for some function u *and* some diffeomorphism ϕ. The procedure here [**KW3**] is similar to the first proof of Theorem 1.8 outlined above.

(b) *Noncompact surfaces.* If M is not compact, but has a boundary, one can impose boundary conditions and ask the same questions—and run into similar difficulties (see Cherrier [**Ch**]).

Little is known if (M, g_0) is complete but not compact and one seeks a complete metric g with curvature K. The appropriate Gauss–Bonnet result is an inequality found by Cohn–Vossen (see [**Bl**]):

$$\int_M K \, dA \leqslant 2\pi\chi(M). \tag{1.14}$$

Except for [**KW2**], where all possible complete metrics on $\mathbf{R}^2$ are determined, essentially all work has centered on the case of pointwise conformal deformations on $\mathbf{R}^2$ with its standard flat metric g_0, so one wants to solve (1.11) in the very special case

$$\Delta_0 u = -K e^{2u}. \tag{1.15}$$

The results depend on the sign of $K(x)$ as $|x| \to \infty$. We give some of these results. Again, the case $K \leqslant 0$ is simpler to understand.

If $K \leqslant 0$ and $K(x) \leqslant -c/|x|^2 < 0$ as $|x| \to \infty$, then Sattinger [**Sa**] (sharpening earlier results by Ahlfors, Osserman, and others) showed that there is *no* solution. On the other hand, the work of Ni [**Ni1**] (see also [**KN**]) and McOwen [**Mc1**] proves that if $-c/|x|^l \leqslant K(x) \leqslant 0$ for some $l > 2$, then for every b satisfying $-(l-2)/2 < b < 0$ there is a solution u with the asymptotic behavior

$$u(x) = -b \log|x| + u_\infty + O(|x|^\gamma) \tag{1.16}$$

as $|x| \to \infty$; here u_∞ is a constant and this holds for any

$$\gamma > max(-1, 2 - l - 2b).$$

Note $\gamma < 0$ since $2 - l - 2b < 0$. Thus the metric $g = e^{2u}g_0$ is also complete. Moreover, the total curvature satisfies

$$\int_{\mathbf{R}^2} K\, dA_g = \int_{\mathbf{R}^2} Ke^{2u}\, dA_0 = 2\pi b. \tag{1.17}$$

Thus, the choice of b determines the total curvature.

If K is positive somewhere, it is useful to recall a classical result of Bonnet which asserts that if a complete Riemannian manifold has $K > \text{const} > 0$, then the manifold is compact. Thus, on a complete noncompact M one needs some decay of $K(x)$ as $|x| \to \infty$. Aviles **[Av]** and McOwen **[Mc2]** show that if K is positive somewhere and $K(x) = O(|x|^{-l})$ as $|x| \to \infty$, where $l > 0$, then for every b satisfying $\max(0, 2 - l) < b < 2$ there is a solution with the asymptotic behavior (1.16) and with total curvature (1.17). From (1.16) the resulting metric $g = e^{2u}g_0$ is complete as long as $c \leqslant 1$, which is consistent with the Cohn–Vossen's Gauss–Bonnet inequality (1.14).

In addition, Aviles shows that under certain growth restrictions on $K(x)$ the inequality (1.14) is both necessary and sufficient for the metric $g = e^{2u}g_0$ to be complete.

Even in this case of $\mathbf{R}^2$ with the standard flat metric, our understanding of (1.15) requires more work. Little is known for other 2-manifolds. The next-simplest case is probably the hyperbolic disc H^2 with its metric g_0 having curvature $K_0 = -1$. Then (1.11) is

$$\Delta_0 u = -1 - Ke^{2u}. \tag{1.18}$$

Now the Laplacian is *not* uniformly elliptic, and there are many bounded harmonic functions. There should be interesting new phenomena waiting to be found (see **[AvM]**).

Hyperbolic manifolds Marden

Prescribing the curvature
of a riemannian manifold
Jerry L. Kazdan

Gaussian Curvature
Pag. 01 (09)

Source: Books-004-2021/
Prescribing the curvature..../
Jerry L. Kazdan-Prescribing
the curvature... pdf /

Obs: surface in $\mathbb{R}^3$

As sometimes occurs, one of the most intriguing unresolved problems involving Gaussian curvature is also one of the easiest to understand.

Let $z = u(x,y)$ be a surface in $\mathbb{R}^3$.
Its Gaussian curvature
$K(x,y)$ is given by the formula

$$\frac{u_{xx}\,u_{yy} - u_{xy}^2}{(1 + |\nabla u|^2)^2} = K(x,y).$$

Solucionar!!!

(@

Obs: 20210221

πr^2 (Área Círculo)

$r=2$

$3.14 . 2^2 = 3.14 \times 4 = 12{,}56$

1/4 do círculo = Área 3,14

(raio 2)

Faça cálculos com $r=1$ e
$r=2$

Simule!!

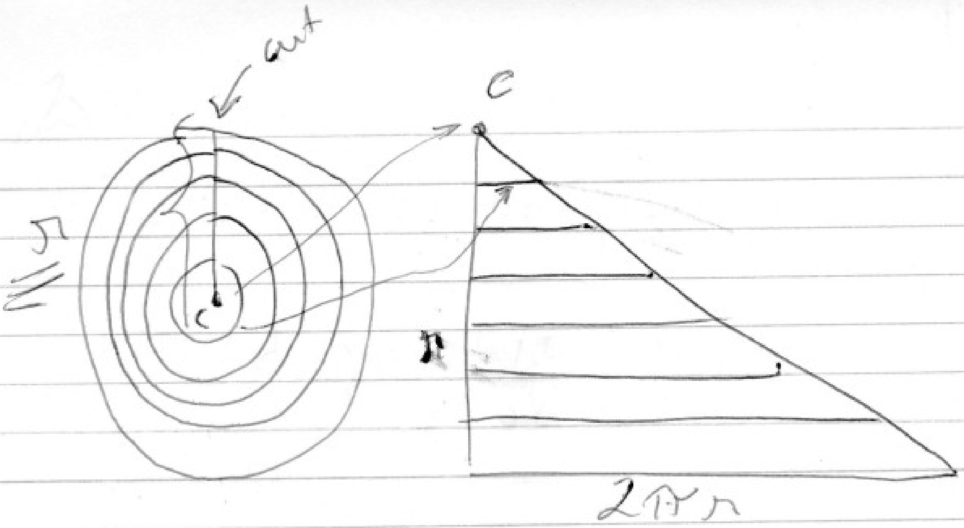

$$\frac{\cancel{2}\pi r \cdot r}{\cancel{2}} = \pi r^2$$

(Area)

(%

20210223.

828,57 Gb.

—

Aplicativo "Meu gov.br"

—

Gêmeos → combina { Áries
Leão
Sagitário

—

I Am mine Pearl Jam.

AKAMAI. —

Proporção Áurea

—

1,618 $\varphi = \frac{(1+\sqrt{5})}{2}$

1,618033988...

$a/b = (a+b)/a = \varphi$ "phi" $= 1.61803...$

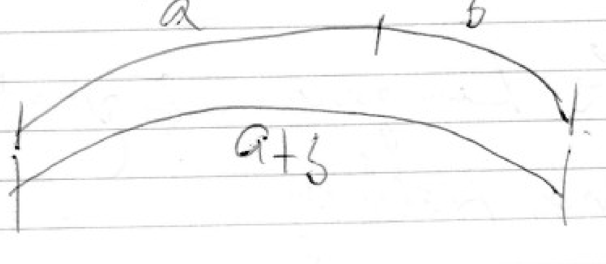

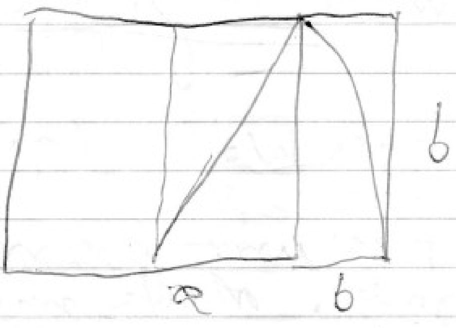

Phídeas
5º A C.
séc.

Fibonacci.

0 1 1 2 3 5 8 13 21 34 55 89 144

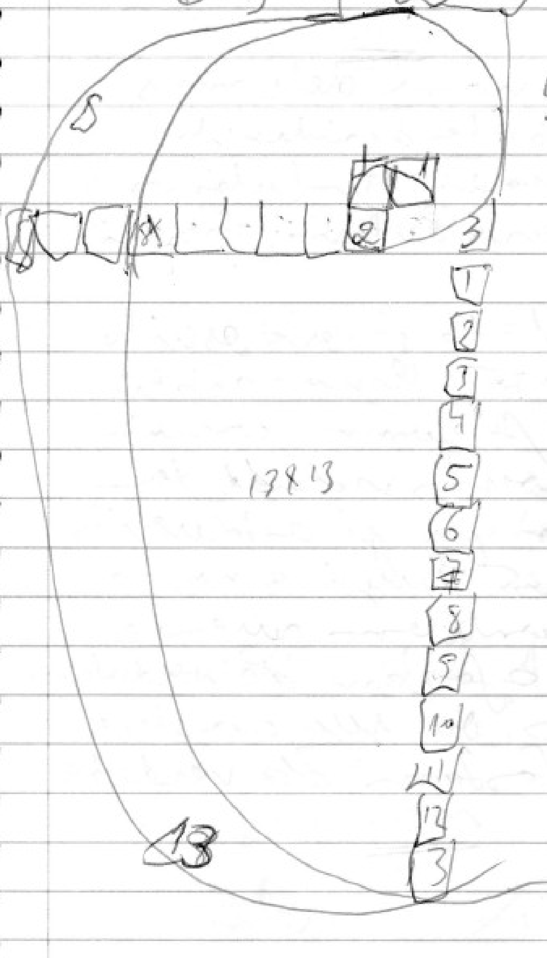
8
5
1
2
3
1
2
3
4
5
6
7
8
9
10
11
12
13
13x13
13

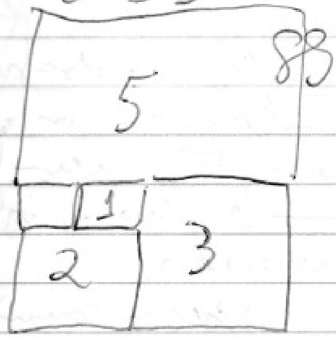

Sequência de
Fibonacci
+
Retângulo de ouro
=
espiral de
ouro.

Códice Fiscale.
MSCSTN69H01Z602Q

stenio.musich@studio.unibo.it

> accesso:
universitaly.it/
Obs:
universitaly.it/index.php/
students/stranieri

studenti.unibo.it

SAT
English Tolc-E
Tolc-I ←
Prova ammissione [illegible]
e751

(@

Condritos carbonáceos

6 nº que definem todo o universo

1. matéria bariônica
2. matéria escura
3. energia escura
4. profundidade óptica
(mais astrofísico de todos os parâmetros do modelo padrão da cosmologia)

~~5.~~ Para determinar a profundidade óptica do universo é feita uma medição da polarização da CMB

CMB – radiação cósmica de fundo em micro-ondas.

Polarização }
intensidade } 3 características
comprimento de onda } de uma onda de luz.

5. } Flutuações quânticas
6. } (diminutas)

Chamado de espectro de potência primordial e descreve as flutuações na densidade do universo no espaço tridimensional.

Parâmetro restante, chamado
índice espectral escalar.

As pequenas variações de
energia que estavam presentes
no Universo primitivo, dependem
da escala angular.

O ruído cósmico primordial
é ligeiramente rosado.

Universo: 13,8 bilhões de
anos
(40 milhões de anos
variação ±)

A observação mais restritiva é a da
anisotropia da CMB:
as diminutas
flutuações da temperatura.

=

$360 \div 13 = 27,\overline{692307}$

$360 \div 11 = 32,\overline{72}$

$\boxed{360 \div 15 = 24}$ hours

$360 \div 16 = 22,5$

$360 \div 18 = 20$

wrinkling = enrugamento

ver Magazines (Universo)!!!

[Google Workspace Enterprise
0800 724 8154

Enterprise Plus (Ilimitado)

https://support.google.com/a/answer/10077880

Standard => TB unlimited

(@@

A direct method approach for the existence of convex hypersurfaces with prescribed Gauss-Kronecker curvature.

Source: www.digizeitschriften.de/dms/img/?PID=GDZPPN002439964

Google Takeout

⇒ Filme: Good Will Hunting

04 min 56 – solucionar problema.

obs: 03:57 – Sistema Avançado de Fourier

Solution: 08:26

%   (@@
+,-   *   *.   ! * %
: %"" D)DDD : %"" D)DDD " 7#
(@

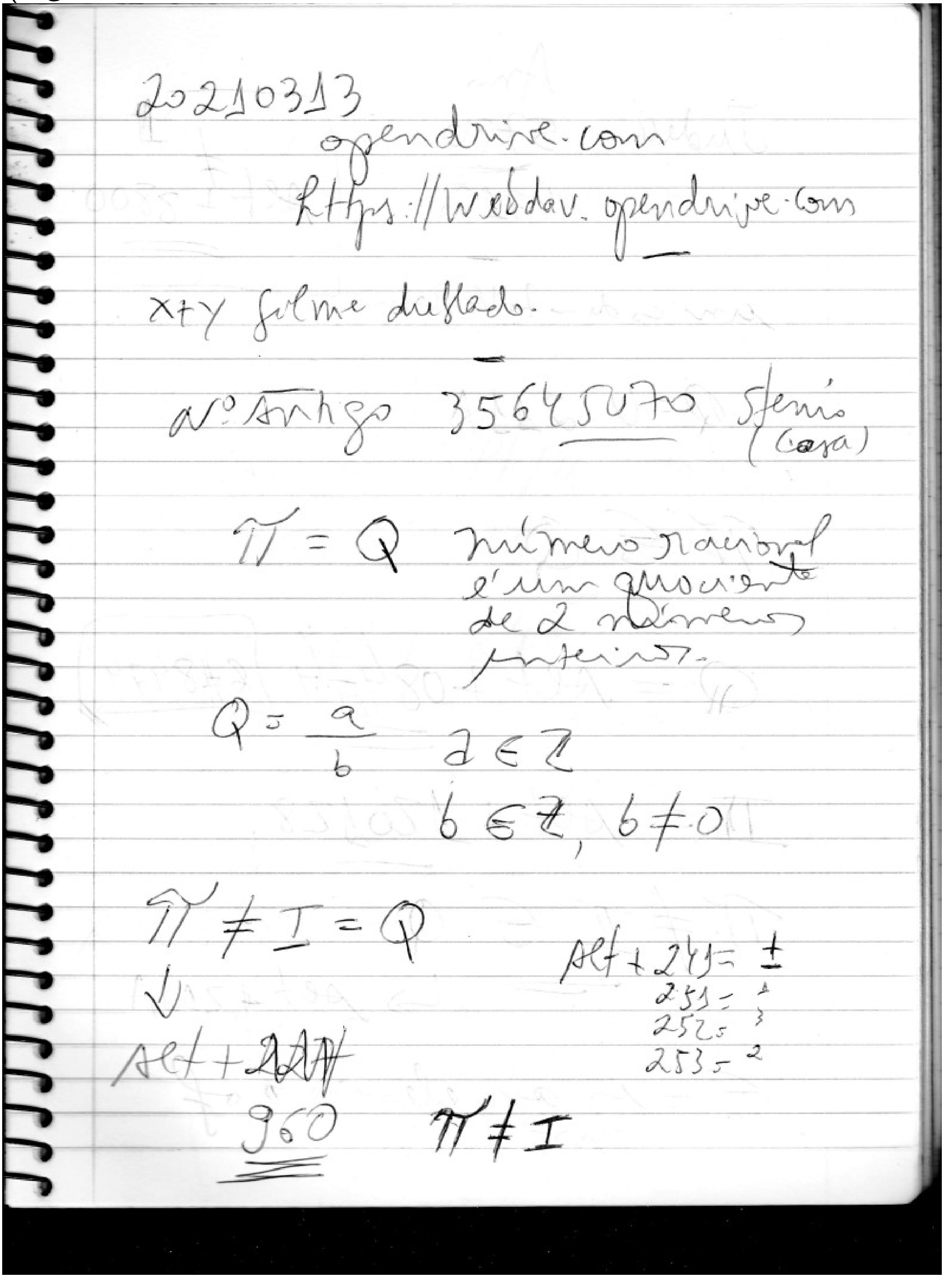

20210313

opendrive.com

https://webdav.opendrive.com

x+y filme dublado.

nº Antigo 35645070 Stenio (casa)

$\pi = Q$ número racional é um quociente de 2 números inteiros.

$Q = \frac{a}{b}$ $a \in \mathbb{Z}$

$b \in \mathbb{Z}$, $b \neq 0$

$\pi \neq I = Q$

↓

Alt + ~~227~~

960

$\pi \neq I$

Alt + 241 = ±

251 = ^{1}

252 = ^{3}

253 = ^{2}

Ano

Judeus = 5780

4719

≠

Alt $\overset{\neq}{+}$ 8800

unicode-table.com

$\mathbb{Q}$ = U+211A

$\pi \in \mathbb{Q}$

$\mathbb{Q}$ = Alt + 08474 $\boxed{\&\#8474}$

$\mathbb{I}$ = Alt + 120128

$\pi \neq \mathbb{I} \in \mathbb{Q}$

↳ Alt + 211A

$\in$ = is an element of
not =

$\pi = \textcircled{X}$

↓

$\pi =$ 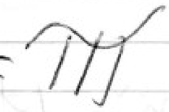

$\pi = \text{Alt} + \underline{960}$

Font MUS:

Local/Microsoft/Windows/EUDC/

mus-20210314.TTE

Q = Alt + 8474

Font: Microsoft Sans Serif.

Symbol font

Alt + 8474 = →

=

Editor de Caractere Particular

→

Execute: eudcedit

Use no Mapa de Caracteres:
Todas as fontes

MUS
mathematical
Uniforme
~~Specification~~ Space

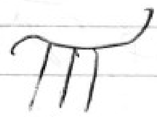

Coloquei no Website em
20210314

202103150357

Vocês estão querendo construir um túnel para o futuro usando uma colher de chá. Vocês estão equivocados, vocês devem usar a dinamite. Vocês devem implodir a forma como fazem matemática, física, química etc.

David de Michelangelo foi construído com uma colher de chá.

Precisamos ir além e unir a colher de chá e a dinamite.

A constante de expansão do Universo (Hubble) está sendo calculada de forma imprecisa pois vocês estão usando conceitos e cálculos ultrapassados.

É chegada a hora de quebrarmos os grilhões que nos prendem ao passado como uma bola de ferro presa aos nossos pés. Para voar precisamos deixar o chão que achamos ser seguro e que muitas vezes poderá se abrir bem embaixo dos nossos pés.

20210315

Die Vermessung der Welt
(Gauss)
Medindo o Mundo

Bernhard Rieman

David Hilbert – queria lançar o plano de reconstruir os fundamentos da matemática, para torná-los consistentes e livres de paradoxos.

Gödel, com sua tese de doutorado, pôs um ponto final a essa pretensão, demonstrando que havia problemas na matemática que eram impossíveis de serem resolvidos.

A deslumbrante clareza e exatidão da matemática era, na realidade, um labirinto repleto de paradoxos.

Gödel provou: 1) Em um sistema formal axiomático consistente que possa expressar fatos sobre aritmética básica, há enunciados verdadeiros que não podem ser provados; 2) Que a consistência do sistema não pode ser provada dentro do mesmo sistema.
São os teoremas da incompletude.

Zero:

Fortaleza de Gwalior na
Índia Central.
(registro mais antigo do
zero)

India – Shunya
(idéia filosófica do
vazio para
representar o novo
termo matemático).

—

Collected mathematical works.
Bohr, Harald August 1887-1951
3 volumes 1952

---

nbarchive.dk/publications/
bcw/
(Niels Bohr Collected
Works)
13 vol.
ISBN 978-0444532862

8504 pages
51.8 pounds
Obs: Amazon
Niels Bohr - Collected
Works

AbeBooks.co.uk (co.uk)
bv=12677084031 £ 4,880.36

(@

MICROLINO

500 elitmulo €$=35.000

Geometria Analitica —

Lezione di Geometria Analitica

1961 Guido Castelnuovo 18,00

Collected Mathematical Papers

George David Birkhoff 824 pg

3 vol.

$40 \div 6{,}5 = 6{,}\overline{153846}$

$40 \div 5{,}5 = 7{,}\overline{27}$ ⫽ > 1,53

$40 \div 4{,}5 = 8{,}\overline{8}$

$40 \div 3{,}5 = 11{,}\overline{428571}$ > 2,628571

⇛ $40 \div 2{,}5 = 16$ (par) > 4,57[illegible]

$40 \div 1{,}5 = 26{,}\overline{6}$

$40 \div 2{,}5 = \boxed{16}$ $360°/16 =$ 22,5

×

$40 \times 22{,}5$[illegible] = [illegible] = $\boxed{900}$

360°:

(1 → 22,5
2 → 45
(3 → 67,5
4 → 90
(5 → 112,5
6 → 135,0
(7 → 157,5
8 → 180
(9 → 202,5
10 → 225

(11 → 247,5
12 → 270
(13 → 292,5
14 → 315
(15 → 337,5
16 → 360

22,5 → ÷ 40 =
0,5625

$\boxed{22{,}5 \div 50 = 0{,}45}$

1 impar

$50 \div 2{,}5 = 20$ $\quad 360 \div 20 = 18$

$50 \times 18 = \boxed{900}$ partes

$18 \div 20 = 0{,}36$

18	108	198	288
36	126	216	306
54	144	234	324
72	162	252	342
90	180	270	360

$360 \div 0{,}36 = 1000$

$\pi = \frac{C}{D} = \frac{360°}{100\,\text{mm}} = 3{,}6$

$R = 0{,}50$

$(50 \div 0{,}50 = 100)$

$\pi =$

$3{,}14 - 3{,}6$

$0{,}46$

$3{,}1415 - 3{,}6 =$

$0{,}4585$

$3{,}14159 - 3{,}6 =$

$0{,}45841$

error

π π

$3,6 \div \pi = 3,6 - \pi =$

$\underbrace{0,45}\,84\,\underbrace{07\ldots}_{\text{erro}}$

$3,6 - \pi = 0,45 \Rightarrow 0,4584$

$\pi = 3,15 \Rightarrow 3,1416$

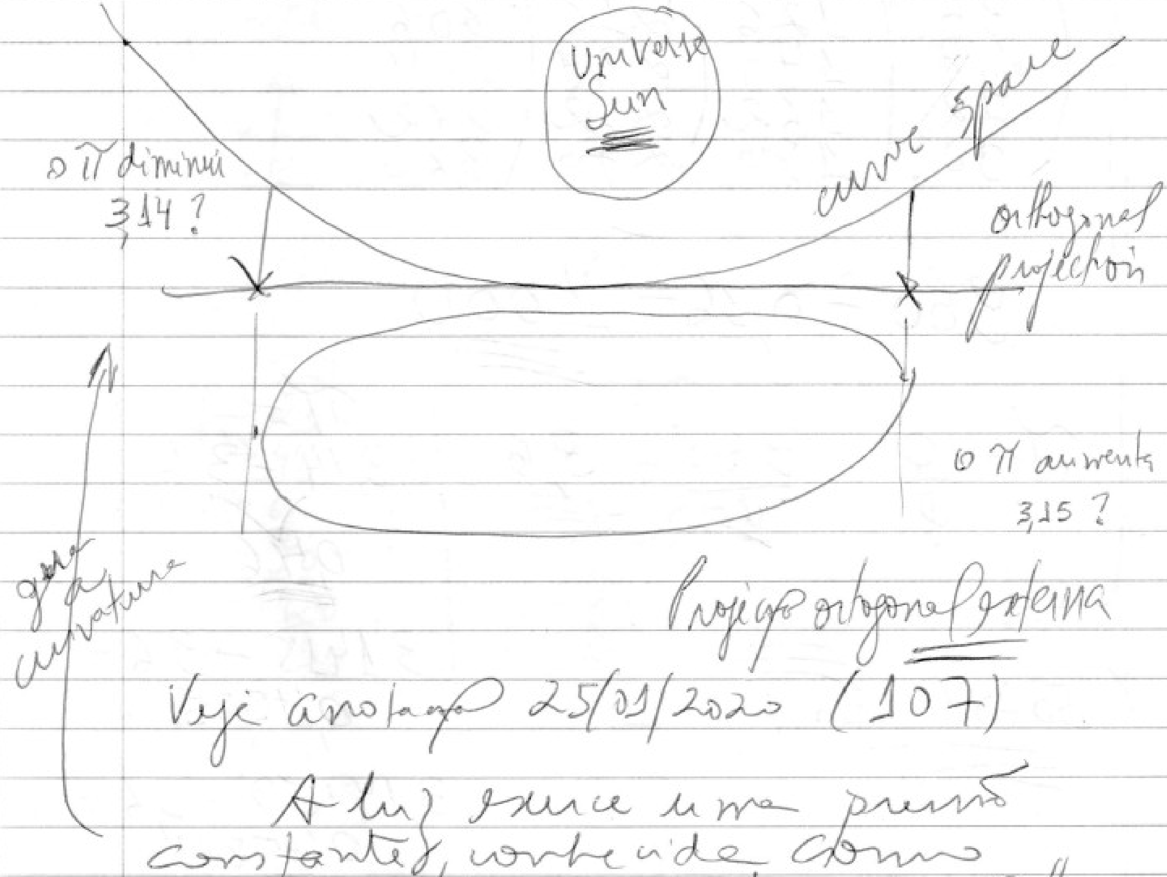
Universe Sun
o π diminui 3,14?
curve space
orthogonal projection
o π aumenta 3,15?
gera curvatura

Projeção ortogonal externa

Veja anotação 25/01/2020 (107)

A luz exerce uma pressão constante, conhecida como "pressão de radiação"

Veja:

Anotação 108 (110)

110 (02/02/2020)

Gauss foi o primeiro a testar a curvatura do espaço físico:

179° 53' 59.32''

obs: Não há geodésicas paralelas

A History of non-Euclidean geometry

B. A. Rosenfeld

Pág 280

Osculating circle at that point.

etc.

)   (")
+,-   *   *.   ! *)   +   3   # !   3
;   7   !   7!   7   )%% #?;
+,-   *   *.   ! *)   +   3   # !   3
;   7   !   7!   7   )%% 7#
(%   )   !7   %

Chapter 8

The Curvature of Space

Curvature and Intrinsic Geometry of a Surface in the Works of Euler

By the *curvature* of a curve at a point we mean the limit of the ratio of the angle $\Delta\alpha$ between the tangents at the endpoints of an arc to the length Δs of that arc as the latter contracts to the point; that is, the limit

$$k = \frac{d\alpha}{ds} = \lim_{\Delta s \to 0} \frac{\Delta\alpha}{\Delta s}. \tag{8.1}$$

k is also the reciprocal of the *radius of curvature* at the point in question, that is, the radius of the osculating circle at that point. (The osculating circle at a point P of a curve is defined as the limit of circles determined by three points on the curve as they tend to P.) The concepts of the curvature of a curve and of the osculating (literally "kissing," from the Latin *osculans*) circle were already known to Leibniz.[1] Leibniz also suggested the possibility of characterizing the curvature of a surface by means of an osculating sphere.[2]

Euler characterized the curvature of a surface in his *Investigations of the curvature of surfaces* (Recherches sur la courbure des surfaces. Berlin, 1767) *[176, vol. 28, pp. 1–22]*. Euler called a normal section of a surface $z = f(x, y)$ perpendicular to the xOy plane a *principal section*. By varying the angle φ between a normal section and the principal section Euler found that at each point of the surface there is a maximal radius of curvature f and a minimal radius of curvature g, that their planes are perpendicular to one another, and

---

[1] Leibniz first considered osculation of curves—in particular, osculation of curves and circles—in his *New reflections on the nature of the angles of tangency and osculation and their use in mathematical practice to replace complex figures with simpler ones* (Meditatio nova de natura Anguli contactus et osculi, horumque usu in practica Mathesi ad figuras faciliores succedaneas difficilioribus substituendas. Leipzig 1686) *[312, v. 7, pp. 326–329]*.

[2] In a letter to Johann Bernoulli, dated July 29, 1698.

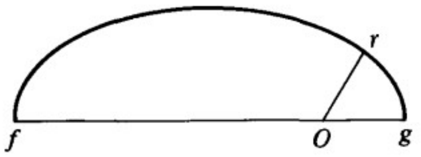

Figure 101

that the connection between an arbitrary radius of curvature $r = r(\varphi)$ and the radii of curvature f and g is given by

$$r = \frac{2fg}{f + g - (f - g)\cos 2\varphi}.$$

Euler writes:

> For a simple derivation of this formula we join the maximal and minimal osculating radii by putting $Of = f$ and $Og = g$ and describe on fg a semi-ellipse with one focus at O (Figure 101). Then for the section MN one must take the angle fOr, doubled with respect to the angle EZM, and the line Oz will be equal to the osculating radius for the section MN. Thus drawing conclusions about the curvature of surfaces, at first glance a matter of great complexity, reduces for each element to the determination of two osculating radii one of which is maximal and the other minimal for that element; these two things entirely determine the nature of the curvature, and we obtain the curvature of all possible sections perpendicular to the given element *[176, vol. 28, p. 21]*.

In 1837 Charles Dupin (1784 1873) *[157, p. 109]* modified Euler's formula to read

$$\frac{1}{r} = \frac{1}{f}\cos^2 \varphi + \frac{1}{g}\sin^2 \varphi.$$

In this paper *On solids whose surfaces can be developed onto a plane* (De solidis quarum superficiem in planum explicare licet. Petersburg, 1772) *[176, vol. 28, pp. 161–186]* Euler introduced the concept of a *developable surface*, that is, a surface that can be applied to a plane without folding or tearing, and proved the fundamental theorem that such a surface is a cylinder, or a cone, or a surface formed by the tangents to a space curve. His point of departure was that on such a surface an infinitesimal triangle must be congruent to the corresponding triangle in the plane to which the surface is applied. Further, he introduced on the surface curvilinear coordinates t, u equal to the rectangular coordinates of the corresponding points of the plane and, denoting the partial derivatives $\partial x/\partial t$, $\partial y/\partial t$, $\partial z/\partial t$ by l, m, n, and $\partial x/\partial u$, $\partial y/\partial u$, $\partial z/\partial u$ by λ, μ, ν, set down the developability conditions in the form $l^2 + m^2 + n^2 = 1, \lambda^2 + \mu^2 + \nu^2 = 1, l\lambda + m\mu + n\nu = 0$, that is, as the requirements of unicity and perpendicularity of the partial derivatives with respect to t and u of the radius vector of a point of the surface. An equivalent form of this condition is that the square of an arc element of a curve on a developable surface equals the sum of the squares of the differentials of the coordinates,

$$ds^2 = dt^2 + du^2,$$

that is, the arc element of a developable surface coincides with the arc element of a plane.

In a note published only in 1862 *[176, vol. 29, pp. 437–440]* Euler also established the general conditions of applicability of a surface to other surfaces. These results of Euler pertain to the so-called *intrinsic geometry of a surface*, that is, the study of properties of a surface that are unchanged when the surface is bent.

Euler's work on *geodesics* also pertains to the intrinsic geometry of surfaces. A *geodesic* is the shortest of all curves on a surface joining two of its points (in the plane the geodesics are straight lines and on a sphere they are great circles).[3]

The problem of finding geodesics on a surface was first posed in 1697 by Johann Bernoulli (1667–1748). Shortly after that, Johann Bernoulli wrote a letter to l'Hospital to the effect that he had found the general differential equation of geodesics. This result was published only in 1742 *[50, p. 364]*. Jacob Bernoulli (1654–1705) showed in 1698 that when cylindrical and conical surfaces are applied to planes, their geodesics go over into straight lines.

Euler was the first to publish the differential equation of geodesics. He did this in the paper *On the shortest curve on an arbitrary surface joining two arbitrary points* (De linea brevissima in superficie quacumque duo quaelibet puncta jungente. Petersburg, 1732) *[176, v. 25, pp. 1–12]*. Euler's equation was

$$\frac{Qddx + Pddy}{Qdx + Pdx} = \frac{dxddx + dyddy}{dt^2 + dx^2 + dy^2},$$

where t, x, y are rectangular coordinates in three-dimensional space, and the functions P and Q are determined from the differential equation of the surface $Pdx = Qdy + Rdt$. Euler returned to the problem of geodesics many times. In the second volume of his *Mechanics* (Mechanica. Petersburg) *[177, v. 2, p. 426]* he showed that, "in the absence of forces," the path of a point on a surface is a geodesic (an assumption used earlier by Bernoulli). In this connection, Euler showed that the *principal normal of a geodesic* (a straight line in the osculating plane of that curve passing through the point of contact and perpendicular to the tangent) always coincides with the normal to the surface.

---

[3] Many of Euler's papers deal with the determination of geodesics. The first was published in 1728 *[176, vol. 25, pp. 1–12]*, and the last was published posthumously *[176, vol. 25, pp. 269–279]*. Euler called geodesics *shortest curves*. The term *geodesic* first appeared in Laplace's *Celestial Mechanics* (Paris, 1799). Later it was applied to all quadrics. After the appearance of Liouville's paper (1814) it was applied to all surfaces.

We note that in the cartography paper *On the representation of the surface of a sphere in the plane* Euler compared the arc length elements of the plane and the sphere and proved the impossibility of an isometric mapping of a sphere to a plane. Also, he posed the question about the three mappings of the sphere to the plane that are of greatest importance in cartography, namely, the mapping in which the meridians and parallels are mapped onto an orthogonal system of lines, the mapping that preserves angles between curves (conformal mapping), and the mapping that preserves areas (equiareal mapping).

Independently of Euler, Monge considered developable surfaces in his *Memoir on evolutes, radii of curvature and various kinds of inflection of curves with twofold curvature* (Mémoire sur les développées, des rayons de courbure et les differents genres d'inflexion des courbes à double courbure. Paris, 1785) *[374]*. We note that Monge's book *Applications of analysis to geometry* (Applications de l'analyse à la géométrie. Paris, 1807) *[373]* contained the first systematic exposition of the theory of surfaces.

Intrinsic Geometry of Surfaces in the Work of Gauss

The general theory of the intrinsic geometry of a surface was formulated by Carl Friedrich Gauss in his *General investigations of curved surfaces* (Disquisitiones generales circa superficies curvas. Göttingen, 1828) *[196, vol. 4, pp. 217–258]*. Gauss's paper was the result of the assignment, given to him in 1820, to produce a cartographic survey of the kingdom of Hanover. This assignment made him reflect on problems of geodesy. In turn, this brought him to the theory of surfaces.

Gauss's theory unified the surface theories of Euler and Monge and the theory of quadratic forms developed by him in his *Arithmetical investigations* (Disquisitiones arithmeticae. Göttingen, 1801) *[196, vol. 1]*. In *Investigations of curved surfaces* Gauss introduced curvilinear coordinates p, q of points of the surface and defined the functions

$$a = \frac{dx}{dp}, \qquad b = \frac{dy}{dp}, \qquad c = \frac{dz}{dp}, \qquad a' = \frac{dx}{dq}, \qquad b' = \frac{dy}{dq},$$

$$c' = \frac{dz}{dq}; \qquad bc' - cb' = A, \qquad ca' - ac' = B', \qquad ab' - ba' = C;$$

$$\frac{ddx}{dp^2} = \alpha, \qquad \frac{ddx}{dpdq} = \beta, \qquad \frac{ddx}{dq^2} = \gamma, \qquad \frac{ddy}{dp^2} = \alpha',$$

$$\frac{ddy}{dpdq} = \beta', \qquad \frac{ddy}{dq^2} = \gamma', \qquad \frac{ddz}{dp^2} = \alpha'', \qquad \frac{ddz}{dpdq} = \beta'', \qquad \frac{ddz}{dq^2} = \gamma''.$$

In modern terms, the vectors $\{a, b, c\}$ and $\{a', b', c'\}$ are the partial derivatives of the radius vector of a point on the surface with respect to p and q

(Gauss denoted ordinary and partial derivatives the same way); the vector $\{A, B, C\}$ is the cross product of $\{a, b, c\}$ and $\{a', b', c'\}$ and its direction is that of the normal to the surface; and $\{\alpha, \beta, \gamma\}$, $\{\alpha', \beta', \gamma'\}$, and $\{\alpha'', \beta'', \gamma''\}$ are the second partial derivatives of the radius vector of a point on the surface with respect to p and q.

Further, Gauss introduced the quadratic forms

$$ds^2 = Edp^2 + 2Fdpdq + Gdq^2$$

and

$$Ddp^2 + D'dpdq + D''dq^2,$$

where

$$aa + bb + cc = E, \quad aa' + bb' + cc' = F, \quad a'a' + b'b' + c'c' = G,$$

$$A\alpha + B\beta + C\gamma = D, \quad A\alpha' + B\beta' + C\gamma' = D', \quad A\alpha'' + B\beta'' + C\gamma'' = D''.$$

The first of these forms, the so-called first fundamental form, gives the square of an arc length element of the surface, that is, of an element of length of an arc of a curve on the surface, and the second differs from the modern second fundamental form only by the multiplier $AA + BB + CC = EG - FF$.

Gauss introduced the all-important concept of "measure of curvature" now known as the *Gaussian curvature* of a surface. Making use of the spherical image of the surface, that is, the mapping that associates to a point P on the surface, a point P' on the unit sphere, such that the radius OP' is parallel to the normal to the surface at P (in modern terms, the end of the unit normal to the surface with beginning at a fixed point), Gauss defined the concepts of total curvature and "measure of curvature" as follows:

> We shall say that a part of a curved surface bounded by a given contour has *total curvature* given by the area of the corresponding figure on the surface of the sphere. One must draw a clear distinction between the total curvature and a kind of specific curvature that we shall call the *measure of curvature*. The latter pertains to a point on the surface and is the fraction obtained by dividing the total curvature of an element of the surface adjacent to the point by the area of that element, and thus gives the ratio of corresponding infinitesimal areas on the sphere and on the curved surface. We hope that our subsequent exposition will fully explain the utility of these new concepts" *[196, vol. 4, p. 226]*.

Gauss goes on to show that

> the "measure of curvature" at a point of the surface is equal to a fraction whose numerator is 1 and whose denominator is the product of the two principal curvatures of the normal sections.
>
> It is clear that the measure of curvature is positive for convex-convex or concave-concave surfaces (this is a trivial distinction) and negative for convex-concave ones. If the surface is made up of parts of both

kinds, then on their boundary the measure of curvature must vanish *[196, vol. 4, pp. 231]*.

Then Gauss obtains further expressions for the "measure of curvature" of which the most important are the expression of the "measure of curvature" k in terms of the coefficients of his two forms,

$$k = \frac{DD'' - D'D'}{EG - FF},$$

and the expression in terms of the coefficients of the first form alone and their derivatives with respect to p and q:

$$\begin{aligned} 4(EG - FF)^2 k = {} & E\left(\frac{dE}{dq}\frac{dG}{dq} - 2\frac{dF}{dp}\frac{dG}{dq} + \left(\frac{dG}{dp}\right)^2\right) + \\ & + F\left(\frac{dE}{dp}\frac{dG}{dq} - \frac{dE}{dq}\frac{dG}{dp} - 2\frac{dE}{dq}\frac{dF}{dq} + \frac{dF}{dp}\frac{dF}{dq} - \right. \\ & \left. - 2\frac{dF}{dp}\frac{dG}{dp}\right) + G\left(\frac{dE}{dp}\frac{dG}{dp} - \frac{dE}{dp}\frac{dF}{dq} + \left(\frac{dE}{dq}\right)^2\right) - \\ & - 2(EG - FF)\left(\frac{ddE}{dq^2} - 2\frac{ddF}{dpdq} + \frac{ddG}{dp^2}\right). \end{aligned}$$

In connection with the last formula Gauss makes the following comment:

> The formula in the last section leads, of itself, to the following remarkable
>
> *Theorem.* If a curved surface is applied to any other surface then the measure of curvature at each of its points remains unchanged.
>
> Also, it is clear that every finite part of a curved surface will, after application to another surface, retain its total curvature.
>
> The special case to which geometers have until now limited their investigations is the case of surfaces applicable to a plane. Our theory readily shows that the measure of curvature of such surfaces at any point is zero. . . .
>
> What we have set forth in the previous section involves a special approach to the investigation of surfaces worthy of the close attention of geometers. Namely, when a surface is regarded not as the boundary of a solid but as a solid with one vanishing dimension, flexible but not stretchable, then the properties of the surface are partly dependent on the form to which it has been reduced and in which it is being studied, and partly independent [of it in the sense that] they remain invariant regardless of its form under bending. The latter properties, whose investigation opens a new and fruitful area of geometry, include the measure of curvature and total curvature as defined by us. They also include the study of geodesic curves, as well as many other matters to be discussed in the sequel. When this approach is adopted, then the plane, and surfaces applicable to a plane, such as cylindrical and conical surfaces,

> are regarded as essentially the same, and the general method of characterizing a surface studied in this manner is based on the formula $\sqrt{Edp^2 + 2Fdpdq + Gdq^2}$ that expresses the connection of the line element of the surface with two variables p and q *[196, vol. 4, pp. 237–238]*.

Gauss is defining the *intrinsic geometry* of a surface invariant under bending and application to another surface. Some of the properties studied in intrinsic geometry are the lengths of curves on the surface, the angles between curves, and what Gauss calls *geodesic curves*, that is, geodesics. Gauss's theorem, called by him a *remarkable theorem* (Theorema Egregium), asserts that the *measure of curvature* is a property that belongs to intrinsic geometry.

Further, Gauss considers the total curvature of a geodesic triangle (that is, a triangle whose sides are arcs of geodesics). The total curvature of such a triangle is given by $\int k d\sigma$, where $d\sigma$ is an element of the surface of the triangle. He finds that

> the total curvature of a triangle is equal to the part of the spherical surface that corresponds to the triangle taken with a plus or minus sign according as the surface on which the triangle lies is concave-concave or concave-convex; as the unit of area we take a square with side one (the radius of the sphere); then the area of the sphere is 4π. It follows that the [surface of the] part of the sphere corresponding to the triangle is to the surface of the sphere as $\pm(A + B + C - \pi)$ to 4π. This theorem, that undoubtedly belongs with the most beautiful theorems of the theory of surfaces, can be stated as follows:
>
> *The excess above* 180° *of the sum of the angles of a triangle formed by geodesic curves on a curved concave-concave surface and the defect below* 180° *of the sum of the angles of a triangle formed by geodesic curves on a concave-convex surface are each measured by the area of the part of the spherical surface that corresponds to the given triangle via normal directions, provided that the whole surface* (*of the sphere*) *is taken as* 720 *degrees [196, vol. 4, pp. 245–246]*.

This theorem of Gauss implies that the Gaussian curvature at a point is equal to the limit of the ratio of the angular excess $A + B + C - \pi$ of a geodesic triangle ABC on the surface to its area as the triangle shrinks to the given point.

Minding's Theory of Surfaces of Constant Curvature

Gauss's work on the intrinsic geometry of surfaces was continued by Ferdinand Minding (1806–1885), who worked in Dorpat (now Tartu). In his *Remark on the development of curved lines on surfaces* (Bemerkung über die Abwickelung krummer Linien auf Flächen. Berlin, 1830) *[366]*, Minding introduced the fundamental concept of *geodesic curvature* of a curve on a surface, defined as the limit of the ratio of the angle between tangent geodesics at the endpoints of an arc of the curve to the arc length as the arc shrinks to a point, and showed that this limit is invariant under bending. In the paper *How to decide whether two surfaces are mutually applicable; including remarks on surfaces of constant measure of curvature* (Wie sich entscheiden lässt ob zwei gegebene krumme Flächen auf einander abwickelbar sind, nebst Bemerkungen über die Flächen von unveränderlicher Krümmungsweise. Berlin, 1839) *[367]* Minding expressed the line element of a surface with constant "measure of curvature" as

$$ds^2 = dp^2 + \left(\frac{1}{\sqrt{k}}\sin p\sqrt{k}\right)^2 dq^2$$

and concluded that

> two surfaces of equal constant measure of curvature can be applied to each other in infinitely many ways, since any two points of one can be associated to two arbitrary points of the other provided that the lengths of the shortest curves between the points of each pair are equal. This implies the following consequences:
>
> Every surface whose measure of curvature is zero [results from] the bending of a plane—this much is well known.
>
> Every surface whose measure of curvature (k) is constant and positive can be applied to a sphere of radius $1/\sqrt{k}$ *[367, pp. 375–376]*.

Further, Minding discovered helical surfaces and, in particular, surfaces of constant negative curvature. One such surface is given by the parametric equations

$$r = \sqrt{x^2 + y^2} = \frac{1}{\cosh\varphi}, \qquad z = \varphi - \tanh\varphi.$$

Minding comments that

> This surface is the result of rotating a curve like ... *CBE* ... (Figure 102) about the zz-axis which it approaches asymptotically *[367, p. 380]*.

The curve *CBE*, which is characterized by the constancy of the length of the segment of the tangent between the point of tangency and the zz-axis, is called a *tractrix*, and the surface of revolution referred to by Minding is called a *pseudosphere*.

In *Contributions to the theory of shortest curves on curved surfaces* (Beiträge zur Theorie der kürzesten Linien auf krummen Flächen. Berlin, 1840) *[368]* Minding established the trigonometric relations in geodesic triangles on such surfaces and noted that these formulas could be obtained from the corresponding formulas of spherical geometry on a sphere of radius r by multiplying r by the complex number i.

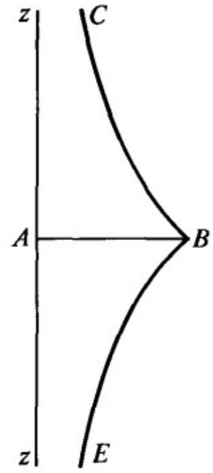

Figure 102

Interpretation of the Lobačevskian Plane on a Pseudosphere

In spite of the fact that Minding's paper appeared in the same journal (*Crelle's*) as Lobačevskiĭ's *Imaginary geometry*, neither of them noticed that the trigonometric formulas of the hyperbolic plane coincide with the trigonometric formulas of a surface of constant negative curvature.[4]

The person who *did* notice the sameness of these formulas was E. Beltrami, who considered the matter in *An attempt at an interpretation of non-Euclidean geometry [42, pp. 374 405]*. Beltrami constructed a model of the hyperbolic plane in a circle and associated to the points of the hyperbolic plane of curvature $-1/R^2$ the coordinates u, v of the corresponding points of the circle. He found that in this coordinate system the line element is given by

$$ds^2 = R^2\frac{(a^2 - u^2)du^2 + 2uvdudv + (a^2 - v^2)dv^2}{a^2 - u^2 - v^2}.$$

Beltrami computed the total curvature of a surface with this line element and noticed that the Gaussian curvature of the hyperbolic plane is everywhere equal to the same number $-1/R^2$; that is, the hyperbolic plane can be viewed as a surface of constant negative curvature.

Thus, speaking of surfaces of constant negative curvature, Beltrami concluded that

---

[4] Lobačevskiĭ's paper appeared in vol. 17 (1837) of *Crelle's Journal* and Minding's paper in vol. 20 (1840) of that journal. B. L. Laptev *[265, p. 20]* explains that Lobačevskiĭ did not borrow the latter volume from the library of Kazan University.

> the theorems of non-Euclidean planimetry apply to them. Further, most of these theorems can be thought of concretely only if we refer them to such surfaces and not to the plane *[42, p. 381]*.

That is why Beltrami suggested that all surfaces of constant negative curvature be called *pseudospherical* surfaces.

Beltrami also showed that each of the sheets of Minding's pseudosphere is isometric to the part of the hyperbolic plane enclosed between two parallel lines and a horocycle perpendicular to them. He also found parts of the hyperbolic plane isometric to other surfaces of revolution of constant negative curvature determined by Minding.

Riemannian Geometry

In the paper *On the hypotheses which lie at the foundations of geometry [122, pp. 55–71]* Riemann introduced a notion of an n-*ply extended manifold* broader than either of the similar notions introduced by Grassmann and Schläfli. After defining this manifold Riemann posed the question of the "measure-relations of which such a manifold is capable" and of the possibility "to express geometrically the calculated results." After noting that the foundations of these two parts of the question are established in Gauss's celebrated memoir *Disquisitiones generales circa superficies curvas [122, pp. 59–60]*, Riemann writes:

> Position-fixing being reduced to quantity-fixings, and the position of a point in the *n*-ply extended manifold being consequently expressed by means of *n* variables $x_1, x_2, x_3, \ldots x_n$, the determination of a line comes to the giving of these quantities as functions of one variable. The problem consists then in establishing a mathematical expression for the length of a line, and to this end we must consider the quantities x as expressible in terms of certain units. I shall treat this problem only under certain restrictions, and I shall confine myself in the first place to lines in which the ratios of the increments dx of the respective variables vary continuously. We may then conceive these lines broken up into elements, within which the ratios of the quantities dx may be regarded as constant; and the problem is then reduced to establishing for each point a general expression for the linear element ds starting from that point, an expression which will thus contain the quantities x and the quantities dx. I shall suppose, secondly, that the length of the linear element, to the first order, is unaltered when all the points of this element undergo the same infinitesimal displacement, which implies at the same time that if all the quantities dx are increased in the same ratio, the linear element will vary also in the same ratio. On these suppositions, the linear element may be any homogeneous function of the first degree of the quantities

> dx, which is unchanged when we change the signs of all the dx, and in which the arbitrary constants are continuous functions of the quantities x. To find the simplest cases, I shall seek first an expression for manifoldness of $n - 1$ dimensions which are everywhere equidistant from the origin of the linear element; that is, I shall seek a continuous function of position whose values distinguish them from one another. In going outwords from the origin, this must either increase in all directions or decrease in all directions; I assume that it increases in all directions, and therefore has a minimum at that point. If, then, the first and second differential coefficients of this function are finite, its first differential must vanish, and the second differential cannot become negative; I assume that it is always positive. This differential expression, then, of the second order remains constant when ds remains constant, and increases in the duplicate ratio when the dx, and therefore also ds, increase in the same ratio; it must therefore be ds^2 multiplied by a constant, and consequently ds is the square root of an always positive integral homogeneous function of the second order of the quantities dx, in which the coefficients are continuous functions of the quantities x. For Space, when the position of points is expressed by rectilinear co-ordinates, $ds = \sqrt{\sum (dx)^2}$ *[122, pp. 60–61]*.

(In Clifford's translation we changed his term n-*dimensional manifoldedness* to Riemann's n-ply extended manifold). By *Space* Riemann means three-dimensional Euclidean space. After considering various possible cases of the dependence of ds on the differentials dx, Riemann restricts himself to "manifolds for which the line element is given by the square root of a differential expression of the second degree," that is,

$$ds^2 = \sum_i \sum_j g_{ij} dx_i dx_j, \tag{8.2}$$

where $g_{ij} = g_{ji}$ are functions of the variables x_i.

Riemann's requirement guarantees that his n-dimensional space is locally Euclidean. The quadratic form (8.2) is assumed to be positive definite, that is, for all dx_i not simultaneously zero $ds^2 > 0$. Hence each pair of points with infinitesimally different coordinates x_i and $x_i + dx_i$ are a definite distance ds apart. By integrating this distance along different lines we determine their length. Among the different lines we can find the shortest (geodesic) lines $x_i = x_i(t)$ that are solutions of the differential equations

$$\frac{d^2 x_i}{ds^2} + \sum_j \sum_k \Gamma^i_{jk} \frac{dx_j}{ds} \frac{dx_k}{ds} = 0, \tag{8.3}$$

where the Γ^i_{jk} are point functions expressed in terms of the coefficients g_{ij} by means of the relations

$$\sum_i g_{il} \Gamma^l_{jk} = \frac{1}{2}\left(\frac{\partial g_{lk}}{\partial x_j} + \frac{\partial g_{il}}{\partial x_k} - \frac{\partial g_{jk}}{\partial x_l}\right). \tag{8.4}$$

The spaces defined in this way are called *Riemannian spaces*. Surfaces in ordinary space are particular Riemannian spaces for $n = 2$; in this case, an instance of a geometry of the kind defined by Riemann is the so-called intrinsic geometry of a surface, elaborated by Gauss in the work *General investigations of curved surfaces* alluded to by Riemann. Clearly, Euclidean space is a special case of a Riemannian space whose geodesics are Euclidean straight lines.

A more general example of a Riemannian space is a hypersurface in $(n + 1)$-dimensional Euclidean space and, in particular, a hypersphere in this space. The geodesics of a hypersphere are its great circles—sections by two-dimensional planes passing through its center. Elliptic space, obtained from a sphere by identifying antipodal points, and Lobačevskian space are also examples of Riemannian spaces.

One of the most important concepts of Riemannian geometry is the *curvature of a space*. The curvature is defined at every point in every two-dimensional direction passing through this point. To determine the curvature of a Riemannian space at a point one considers a geodesic triangle, that is, a curvilinear triangle bounded by arcs of three geodesics, one of whose vertices is at the given point and two of whose sides issuing from this vertex are tangent to the given two-dimensional direction. For this triangle ABC one computes the angular excess $A + B + C - \pi$ (as defined, the angular excess can be positive, zero, or negative), as well as the area, and forms the ratio of the angular excess over the area of the triangle. Now one lets the triangle shrink to the given vertex but insists that the sides issuing from it continue to be tangent to the given two-dimensional direction. It is obvious that in this limiting process the area of the geodesic triangle tends to zero. Since the geometry of the two-dimensional surface in which the passage to the limit is taking place is locally Euclidean and the angular excess of a Euclidean plane is zero, it follows that the angular excess of the geodesic triangle also tends to zero. But the ratio of the angular excess over the area of the triangle tends to a definite limit called the *curvature of the Riemannian space* at the given point in the given two-dimensional direction.

The curvature of n-dimensional Euclidean space, regarded as a special case of a Riemannian space, is equal to zero at all of its points; that is why a Euclidean space is called Riemannian space of *zero curvature* and Riemannian spaces different from Euclidean space are called *curved spaces*.

Riemann defined the curvature of his space in the following manner; in a neighborhood of the point at which he wanted to define curvature he introduced coordinates x_i equal to zero at the point and such that distance s of points in the neighborhood of this point was given in terms of the coordinates by the relation $s^2 = \sum_i x_i^2$. Riemann writes:

> When we introduce these quantities, the square of the line-element is Σdx^2 for infinitesimal values of the x, but the term of next order in it is equal to a homogeneous function of the second order of the $\frac{1}{2}n(n - 1)$ quantities $(x_1 dx_2 - x_2 dx_1)$, $(x_1 dx_3 - x_3 dx_1)$ ... an infinitesimal, there-

> fore, of the fourth order; so that we obtain a finite quantity on dividing this by the square of the infinitesimal triangle, whose vertices are $(0,0,0,\dots)$, $(x_1, x_2, x_3, \dots)$, $(dx_1, dx_2, dx_3, \dots)$. This quantity retains the same value so long as the x and the dx are included in the same binary linear form, or so long as the two geodesics from 0 to x and from 0 to dx remain in the same surface-element; it depends therefore only on place and direction. It is obviously zero when the manifold represented is flat, i.e., when the squared line-element is reducible to $\Sigma\, dr^2$, and may therefore be regarded as the measure of the deviation of the manifold from flatness at the given point in the given surface-direction. Multiplied by $-\frac{3}{4}$ it becomes equal to the quantity which Privy Councillor Gauss has called the total curvature of a surface *[122, pp. 62–63]*.

In view of Riemann's choice of coordinates, the expansion of the square of the line element ds^2 in the neighborhood of the point he considered is given by

$$ds^2 = \sum_i dx_i^2 + \sum_i \sum_j \sum_k c_{ij,k} x_k dx_i dx_j + \sum_i \sum_j \sum_k \sum_l c_{ij,kl} x_k x_l dx_i dx_j + \dots,$$

where $c_{ij,k}$ and $2c_{ij,kl}$ are the values of the derivatives $\dfrac{\partial g_{ij}}{\partial x_k}$ and $\dfrac{\partial^2 g_{ij}}{\partial x_k \partial x_l}$ at the point under consideration. In the first place, Riemann notes that there is no linear term in this expansion, that is, that $c_{ij,k} = 0$. The reason for this is that his coordinate lines are geodesics satisfying equation (8.3), where the coefficients Γ^i_{jk} are expressed in terms of g_{ij} by formulas (8.4). Then Riemann claims that the second-order terms in the expansion of ds^2 form a quadratic form in the $x_i dx_j - x_j dx_i$. If for the sake of uniformity we denote the infinitesimals x_i by δx_i then the latter magnitudes can be written as $\Delta x_{ij} = \delta x_i dx_j - \delta x_j dx_i$. These quantities determine a parallelogram on the vectors $\{dx_i\}$ and $\{\delta x_i\}$, so that the quadratic terms in the expansion can be written as

$$\Delta\sigma^2 = \sum_i \sum_j \sum_k \sum_l R_{ij,kl} \Delta x_{ij} \Delta x_{kl}.$$

The quantity obtained by Riemann, which he calls the fraction from the division of $\Delta\sigma^2$ by the area of the triangle on the vectors $\{dx_i\}$ and $\{\delta x_i\}$, is actually the limit of this fraction as the triangle in question shrinks to the point. Nowadays, a quantity proportional to that defined by Riemann (with a proportionality factor such that in the case of a surface it coincides with the Gaussian curvature) is called the *Riemannian curvature* of the space at a given point in a given two-dimensional direction. Riemann goes on to give our earlier geometric interpretation of the curvature he defined. He notes that on surfaces in ordinary space "the difference between the angle sum of an infinitesimal triangle and two right angles is proportional to the area of the triangle" and writes:

> To give an intelligible meaning to the curvature of an n-fold manifold at a given point and in a given surface-direction through it, we must start from the fact that a geodesic proceeding from a point is entirely determined when its initial direction is given. According to this we obtain a determinate surface if we prolong all the geodesics proceeding from the given point and lying initially in the given surface-direction; this surface has at the given point a definite curvature, which is also the curvature of the n-fold continuum at the given point in the given surface-direction *[122, p. 64]*.

Riemannian geometry was applied by Riemann to the theory of differential equations in *A mathematical work containing an attempt to answer the question proposed by the most illustrious Paris Academy: To determine the heat state of a homogeneous solid so that a system of isothermal curves given at a certain moment in time remains a system of isothermal curves at an arbitrary moment in time and so that the temperature at a point is expressed as a function of time and two more independent variables* (Commentatio mathematica qua respondere tentatur questioni ab Ill$^{\text{ma}}$ Academia Parisensi proposita: Trouver quel doit être l'état calorifique d'un corps solide homogène pour qu'une système de courbes isothermes, à un instant donné, restent isothermes après un temps quelconque, de telle sorte que la température d'un point puisse s'exprimer en fonction du temps et de deux autres variables indépendentes), written in 1861 and published in his collected works (Leipzig, 1876) *[454, pp. 391–404]*. In this paper Riemann solved the problem of reducing the differential equation of heat conduction

$$\sum_i \frac{\partial}{\partial s_i}\left(\sum_j b_{ij}\frac{\partial u}{\partial s_j}\right) = h\frac{\partial u}{\partial t}$$

to simplest form— a problem equivalent to that of transforming the quadratic form $\sum_i \sum_j b_{ij} ds_i ds_j$ to a sum of squares. Riemann found that this can be done if and only if the expression

$$K = \frac{1}{2}\frac{\sum\limits_i \sum\limits_j \sum\limits_k \sum\limits_l (ij,kl)(ds_i\delta s_j - ds_j\delta s_i)(ds_k\delta s_l - ds_l\delta s_k)}{\left(\sum\limits_i \sum\limits_j b_{ij} ds_i ds_j\right)\left(\sum\limits_i \sum\limits_j b_{ij}\delta s_i \delta s_j\right) - \left(\sum\limits_i \sum\limits_j b_{ij} ds_i \delta s_j\right)^2},$$

that is invariant under a change of variables, vanishes. Of this expression, he denotes by (111), Riemann says:

> The expression $\sqrt{\sum_i \sum_j b_{ij} ds_i ds_j}$ can be regarded as a line element of an n-tuply extended space that is outside the bounds of our intuition. If we lead from point $(s_1, s_2, \ldots, s_n)$ in this space all possible geodesics whose initial directions are characterized by the ratios $\alpha ds_1 + \beta\delta s_1 : \alpha ds_2 + \beta\delta s_2 : \ldots : \alpha ds_n + \beta\delta s_n$ (where α and β are arbitrary quantities) then these lines form a certain surface which we can think of as located in the

> ordinary space of our intuition. Then expression (111) is the curvature of this surface at the point $(s_1, s_2, \ldots, s_n)$ *[454, p. 403]*.

The quantities (ij, kl), the so-called four-index Riemann symbols, are the same quantities that we denoted earlier by $R_{ij,kl}$. We note that the quantities $R_{ij,kl}$ can be expressed in terms of the coefficients g_{ij} and their derivatives, using the quantities Γ^i_{jk} (expressed in terms of the g_{ij} and their derivatives by formulas (8.4)), by means of the formula

$$R_{ij,kl} = \left(\frac{\partial \Gamma^h_{jk}}{\partial x_i} - \frac{\partial \Gamma^h_{ik}}{\partial x_j} + \Gamma^h_{ir}\Gamma^r_{jk} - \Gamma^h_{jr}\Gamma^r_{ik}\right) g_{hl}.$$

In the paper mentioned previously Riemann also used the quantities Γ^i_{jk} but denoted them as p_{ijk}.

Riemann's investigations were continued by Elwin Bruno Christoffel (1829–1900) in the paper *On the transformation of homogeneous differential expression of second degree* (Über die Transformation der homogenen Differentialausdrücke zweiten Grades. Berlin, 1869) *[114, vol. 1, pp. 368–377]*, in which he posed the question of the conditions under which the geometry determined by a form $\sum_i \sum_j g_{ij} dx_i dx_j$ coincides with the geometry determined by a form $\sum_i \sum_j h_{ij} dy_i dy_j$. Just as Riemann's intention was to develop Gauss's theory of surfaces, so Christoffel's intention was to generalize the problem of superposition of surfaces. Christoffel's necessary condition for the coincidence of geometries turned out to be the coincidence of the differential forms $\sum_i \sum_j \sum_k \sum_l R_{ij,kl} dx_i dx_k \delta x_j \delta x_l$, computed for the two given forms. Christoffel denoted the quantities Γ^i_{jk} by $\left\{{jk \atop i}\right\}$ and the quantities $\sum_l g_{il} \Gamma^l_{jk}$ by $\left[{jk \atop i}\right]$. That is why these quantities are often called *Christoffel symbols of the first and second kind*, respectively.

Riemannian Spaces of Constant Curvature

After defining Riemannian spaces of variable curvature and noting that Euclidean spaces are spaces of zero curvature Riemann dwelled at length on spaces of constant nonzero curvature:

> Manifolds whose curvature is constantly zero may be treated as a special case of those whose curvature is constant. The common character of these continua whose curvature is constant may be also expressed thus, that figures may be moved in them without stretching. For clearly figures could not be arbitrarily shifted and turned round in them if the curvature at each point were not the same in all directions. On the other hand, however, the measure-relations of the manifold are entirely determined by the curvature; they are therefore exactly the same in all directions at one point as at another, and consequently the same constructions can be made from it: whence it follows that in aggregates with

constant curvature figures may have any arbitrary position given them. The measure-relations of these manifolds depend only on the value of the curvature, and in relation to the analytic expression it may be remarked that if this value is denoted by α, the expression for the line-element may be written

$$\frac{1}{1 + \frac{1}{4}\alpha \sum x^2}\sqrt{\sum dx^2}$$

[122, p. 65].

The latter expression generalizes the previously mentioned expression, obtained by Minding, of the line element of a surface of constant curvature in terms of its Gaussian curvature.

The simplest example of a Riemannian space of constant positive curvature is a sphere in $(n + 1)$-dimensional Euclidean space; the Riemannian curvature of a sphere of radius r, of any dimension, in all two-dimensional directions is equal to $1/r^2$, for the area of any spherical triangle is the product of r^2 and the angular excess of the triangle.

An example of an n-dimensional Riemannian space of constant negative curvature is n-dimensional Lobačevskian space, first defined by Beltrami in the paper *Fundamental theory of spaces of constant curvature* (Teoria fondamentale degli spazî di curvatura costante. Milan, 1869) *[42, pp. 406–429]* published in the same year as his previously discussed paper. This space can also be viewed as one of the sheets of a sphere of radius qi in $(n + 1)$-dimensional pseudo-Euclidean space. The Riemannian curvature of such a sphere in all two-dimensional directions is $-1/q^2$ since the area of an arbitrary spherical triangle on this sphere is the product of its angular defect by $-q^2$.

We wish to mention here a remarkable theorem of F. Schur established in the paper *On the connection between spaces of constant curvature and projective spaces* (Über dem Zusammenhang der Räume constanten Krümmungmasses mit den projectiven Räumen. Leipzig, 1886) *[517]* which asserts that, on a Riemannian manifold, the constancy of the Riemannian curvature in all two-dimensional directions at each point implies the constancy of the Riemannian curvature at all points. Schur's proof is purely geometric and is based on a projective mapping of the bundle of linear elements at one point onto a similar bundle at another point. An analytic proof of this result was given by Luigi Bianchi (1856–1928) in the paper *On four-index symbols and on Riemannian curvature* (Sui simboli a quatro indici e sulla curvatura di Riemann. Rome, 1902) *[56]*.

Elliptic Geometry

Another form of n-dimensional Riemannian space of constant positive curvature is a sphere in $(n + 1)$-dimensional Euclidean space with identified antipodal points called n-dimensional *elliptic space* or *non-Euclidean space of Riemann*. When $(n + 1)$-dimensional Euclidean space is completed to $(n + 1)$-dimensional projective space, pairs of antipodal points of a sphere in $(n + 1)$-dimensional Euclidean space are projected from its center onto the points of its plane at infinity. Since this plane is an n-dimensional projective space, elliptic space can be viewed as a metrized projective space. The details are as follows.

An arbitrary sphere of $(n + 1)$-dimensional Euclidean space is given by an equation of the form

$$A \sum_i x_i^2 + 2 \sum_i b_i x_i + c = 0,$$

which can be written in homogeneous coordinates as

$$A \sum_i x_i^2 + 2 \sum_i b_i x_i x_0 + c(x_0)^2 = 0.$$

The latter equation shows that our sphere intersects the plane at infinity, $x_0 = 0$, in the imaginary quadric $\sum_i x_i^2 = 0$. Hence an elliptic space can be described as a projective space with a given imaginary conic. The distance ω between two points X and Y of an elliptic space of curvature $1/r^2$ is connected with the angle φ between the corresponding diameters of the sphere by the relation $\omega = \varphi r$. On the other hand, Edmond Laguerre (1834–1886) showed in the paper *On the theory of foci* (Sur la theorie des foyers. Paris, 1853) *[298, vol. 2, pp. 6–15]* that the angle φ between two straight lines in Euclidean space can be expressed in terms of the cross ratio (ij, xy) of the straight lines x, y and two *isotropic straight lines* i, j, that is, imaginary lines of zero length that lie in the plane of x and y and pass through the point of intersection of x and y, by means of the relation

$$\varphi = \frac{1}{2i} \ln(ij, xy). \tag{8.5}$$

Since the imaginary lines i, j that join the center of the sphere to the points I, J of the imaginary quadric $\sum_i x_i^2 = 0$ in the plane at infinity are, in view of the equation of the quadric, isotropic, Laguerre's formula (8.5) yields an expression for the distance ω between points X, Y of elliptic space in terms of the cross ratio (IJ, XY) of these points and the two points of intersection of the straight line XY and the imaginary quadric:

$$\omega = \frac{r}{2i} \ln(IJ, XY). \tag{8.6}$$

An elliptic metric in the projective plane was defined in 1859 by Cayley in the *Sixth memoir upon quantics [103, vol. 2, pp. 561–592]*. An elliptic metric in space, and the term *elliptic geometry*, were introduced by Klein in 1871 in the paper *On the so-called non-Euclidean geometry [282, vol. 1, pp. 254–305]*, in which he wrote that

> The basis of a general projective metric in space is provided by an arbitrary *fundamental surface of the second order.*
>
> To define the distance between two points one joins them by a straight line. It intersects the fundamental surface in two new points that are in a definite cross ratio with the two given points. *The logarithm of this cross ratio multiplied by an arbitrary constant c yields what one should call the distance between the two given points [282, vol. 1, p. 300].*

Klein gave a similar definition of the angle between two planes (see p. 238). Further Klein wrote:

> By a *motion* is meant the totality of linear transformations that leave the fundamental surface invariant.
>
> By *spheres* one means quadric surfaces that meet the fundamental surface along a plane curve. The center of the sphere is the pole of the plane that contains the osculating curve....
>
> If the fundamental surface is *imaginary* then all straight lines have finite length and each pencil of planes has finite angle sum. *Elliptic* geometry comes under this case (provided that the constant c' in the definition of angles is taken as $\sqrt{-1}/2$, so that the sum of the angles in a pencil of planes is π).
>
> We do not investigate the case when the fundamental surface is *real* and *ruled* (like a hyperboloid of one sheet) for this case is no way connected with the three geometries (elliptic, hyperbolic, parabolic) considered here.
>
> Finally, if the fundamental surface is *real* and *nonruled* then for the interior points of the surface we obtain a metric that includes the metric of hyperbolic geometry provided that we again take the constant c' to be $\sqrt{-1}/2$.
>
> *Parabolic* geometry is included as a special case of the general metric; this case arises if the fundamental surface specializes (degenerates) into an imaginary conic section. The fundamental conic section of parabolic geometry is the so-called imaginary circle at infinity *[282, vol. 1, p. 301].*

Klein's *fundamental surface* is now called by Cayley's term *the absolute.* We note that Klein's *linear transformations* are collineations, his *hyperbolic geometry* is Lobačevskian geometry, and his *parabolic geometry* is Euclidean geometry; the *imaginary circle at infinity* is the imaginary circle $\sum_i x_i^2 = 0$ in the plane at infinity of Euclidean space that plays the role of the absolute of this space. Klein did not investigate the geometry of the exterior of an oval (nonruled) quadric and the geometry of a ruled quadric. The reason for this is that, unlike elliptic geometry and Lobačevskian geometry, the geometry of these spaces is locally not Euclidean but pseudo-Euclidean.

We also note that the only difference between formula (8.6) and formula (6.27) for Lobačevskian space is that in one case we have the pure imaginary constant r/i and in the other the real constant q.

Clifford Parallels and Surfaces

The geometry of elliptic space was significantly developed by William Kingdon Clifford (1845–1879) in his *Preliminary sketch of biquaternions* (London, 1879) *[122, pp. 181–200]*.

After defining elliptic, hyperbolic, and parabolic geometries whose absolutes are, respectively, an imaginary quadric, a real (oval) quadric, and an imaginary conic in a real plane, Clifford investigates *elliptic geometry*. After defining poles and polar planes and mutually polar straight lines with respect to an absolute and pointing out that two points that are polar conjugates with respect to the absolute "are a *quadrant* apart," that is, are $(\pi/2)r$ apart, or, if $r = 1$, are $\pi/2$ apart, Clifford notes that

> Through an arbitrary point can in general be drawn *one* line perpendicular to a given plane; namely, the line joining the point to the pole of the plane. If, however, the point *is* the pole of the plane, every line through it is perpendicular to the plane. Similarly, from a point not on the polar of a given line can be drawn one and only one perpendicular to the line; namely, the line through the point which meets the given line and its polar *[122, p. 192]*.

and further:

> *In general, two lines can be drawn so that each meets two given lines at right angles, and these are polars of one another*. One line may therefore be converted into another by rotation about two polar axes. These axes are determined as the lines which meet the two given lines and their polars. If we travel continuously along one of these lines and draw perpendiculars on the other, one of these axes determines the shortest distance between the lines, and the other the longest. If then these two are equal, the lines are equidistant along their whole length. Thus *there is a case of exception in which two lines and their polars belong to the same set of generators of a hyperboloid; the lines are then equidistant along their whole length, and meet the same two generators of one system of the absolute*. I shall use the word *parallel* to denote two lines so situated; and they shall be called *right* parallel or *left* parallel according as one is converted into the other by a right-handed or left-handed twist. Through an arbitrary point can be drawn one right parallel and one left parallel to a given line; the angle between them is twice the distance of the point from the line. There are many points of analogy between the *parallels* here defined and those of parabolic geometry. Thus, if a line meets two parallel lines, it makes equal angles with them; and a series of parallel lines meeting a given line constitute a ruled surface of zero curvature. The geometry of this surface is the same as that of a finite parallelogram whose opposite sides are regarded as identical *[122, pp. 192–193]*.

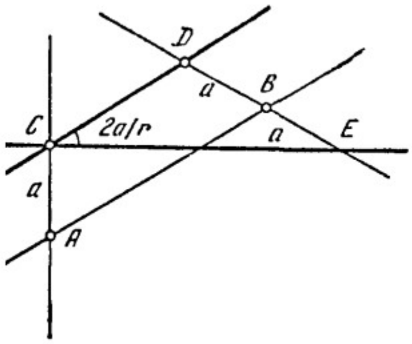

Figure 103

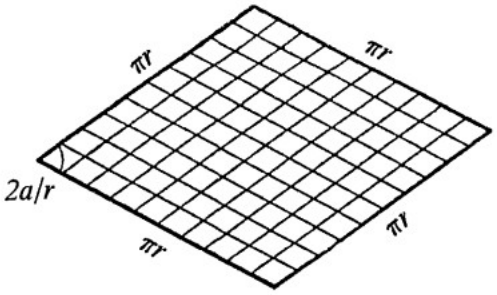

Figure 104

The parallels defined by Clifford are now called *paratactic straight lines*;[5] one must bear in mind that, unlike parallels in Euclidean (*parabolic*) geometry, Clifford parallels are skew straight lines. Figure 103 illustrates the construction of a right, CD, and left, CE, Clifford parallel through a point C to a straight line AB. From the point C one drops the perpendicular CA to the straight line AB and constructs its polar DBE. On the latter one lays off on either side of the point B segments BD and BE equal to the segment $CA = a$. The required parallels are the straight lines CD and CE; the angle DCE is measured in terms of the segment $DE = 2a$ and is equal to $2a/r$. The surface constructed by Clifford is now known as a *Clifford surface*. It is a ruled quadric obtained by rotating one of two paratactic straight lines about the other; its rectilinear generators in both families are paratactic to the axes; at the same time it is also a surface of revolution about the polar of the first axis. The geometry of a Clifford surface is Euclidean. Its area is finite and equal to $\pi^2 r^2 \sin 2a/r$, for it is isometric to a rhombus with sides πr and acute angle $2a/r$ (Figure 104) with sides in each pair of opposite sides glued together. To prove that the geometry of this surface is Euclidean it suffices to write down its parametric equations in homogeneous coordinates whose basis points are located on the axes of the surface,

$$x_0 = \cos\frac{a}{r}\cos u, \qquad x_1 = \cos\frac{a}{r}\sin u,$$

$$x_2 = \sin\frac{a}{r}\cos v, \qquad x_3 = \sin\frac{a}{r}\sin v,$$

and to compute the line element

$$ds^2 = \sum_i dx_i^2 = r^2\left(\cos^2\frac{a}{r}\,du^2 + \sin^2\frac{a}{r}\,dv^2\right).$$

---

[5] This term was introduced by E. Study in the paper *[568]*.

.8 5 7! ! # # 83 # ! ! 7! ! ! !

If in the latter we put

$$U = r\cos\frac{a}{r}\cdot u, \qquad V = r\sin\frac{a}{r}\cdot v,$$

then we have $ds^2 = dU^2 + dV^2$.

If one eliminates from the parametric equations of the Clifford surface the parameters u and v then one obtains its equation in the form

$$\sin^2\frac{a}{r}(x_0^2 + x_1^2) - \cos^2\frac{a}{r}(x_2^2 + x_3^2) = 0.$$

A Clifford surface is the simplest solution of the Clifford-Klein problem of finding spaces with a Euclidean metric that are not isometric to Euclidean space in the large.

Clifford's Idea of the Geometrization of Physics

W. K. Clifford was also interested in questions of the philosophy of space. In his philosophical work *Philosophy of pure science* (London, 1873) there are his famous words on Lobačevskian geometry!

> What Vesalius was to Galen, what Copernicus was to Ptolemy, that was Lobatchewsky to Euclid. There is, indeed, a somewhat instructive parallel between the last two cases. Copernicus and Lobatchewsky were both of Slavic origin. Each of them has brought about a revolution in scientific ideas so great that it can only be compared with that wrought by the other. And the reason of the transcendent importance of these two changes is that they are changes in the conception of the Cosmos *[123, vol. 1, p. 356]*.

What Clifford is saying here is that Lobačevskiĭ's discovery that Euclidean geometry is not the only conceivable geometry is just as revolutionary as Vesalius' shattering of the myth of the exclusive position of man in the animal kingdom and Copernicus' discovery that the Earth is just one of the planets.

Clifford expressed his interesting thoughts on space in his posthumously published *The Common sense of the exact sciences* (London, 1885) *[121]*. Clifford begins the chapter on space with the words

> Geometry is a physical science *[121, p. 43]*,

and at the end of the chapter on position he writes:

> We may conceive our space to have everywhere a nearly uniform curvature, but that slight variations of the curvature may occur from point to point, and themselves vary with the time. These variations of the curvature with the time may produce effects which we not unnaturally attribute to physical causes independent of the geometry of our space.

> We might even go so far as to assign to this variation of the curvature of space "what really happens" in that phenomenon which we term the motion of matter *[121, pp. 202–203]*.

Clifford developed these ideas in greater detail in his report *On the space theory of matter* (Cambridge, 1876). First he paraphrases Riemann:

> Riemann has shewn that as there are different kinds of lines and surfaces, so there are different kinds of space of three dimensions; and that we can only find out by experience to which of these kinds the space in which we live belongs. In particular, the axioms of plane geometry are true within the limits of experiment on the surface of a sheet of paper, and yet we know that the sheet is really covered with a number of small ridges and furrows, upon which (the total curvature not being zero) these axioms are not true. Similarly, he says although the axioms of solid geometry are true within the limits of experiment for finite portions of our space, yet we have no reason to conclude that they are true for very small portions; and if any help can be got thereby for the explanation of physical phenomena, we may have reason to conclude that they are not true for very small portions of space *[122, p. 21]*.

Then he states the following four principles:

> (1) That small portions of space *are* in fact of a nature analogous to little hills on a surface which is on the average flat; namely, that the ordinary laws of geometry are not valid in them.
> (2) That this property of being curved or distorted is continually being passed on from one portion of space to another after the manner of a wave.
> (3) That this variation of the curvature of space is what really happens in that phenomenon which we call the *motion of matter*, whether ponderable or etherial.
> (4) That in the physical world nothing else takes place but this variation, subject (possibly) to the law of continuity *[122, pp. 21–22]*

These principles form a program of geometrization of physics.

Riemann's Topology

In addition to the theory of Riemannian spaces Riemann also founded topology—one of the most important disciplines that significantly broadened our notions of space.

We have already pointed out that Euler interpreted Leibniz's term *geometry of position* in a topological sense, whereas Carnot and Grassmann gave it narrower interpretations. Euler's interpretation of the term was developed by the German physicist Johann Benedict Listing (1808–1862), who suggested the now generally accepted term *topology* in his *Preliminary studies on topology* (Vorstudien zur Topologie. Göttingen, 1847) *[329]*. Listing's term comes from replacing Leibniz's Latin *situs* (place) with its Greek counterpart *topos*. Listing studied

> *linear complections, that is, lines or curves or sets of such, located on a surface, such as a plane or a sphere, or arbitrarily disposed in space [329, p. 867].*

He investigated knots, chains, plaits, and other forms of mutual disposition of linear complections—later he called them *linear complexes*—with numerous examples drawn from biology and technology. Listing's work *The census of spatial complexes or a generalization of Euler's therorem on polyhedra* Der Census räumlicher Komplexe oder Verallgemeinerung des Euler'schen Satzes von den Polyedern. Göttingen, 1862) *[330]* dealt with more general *spatial complexes*.

Gauss interpreted Leibniz's ideas in the same topological sense. Gauss devoted to topological investigations, as to non-Euclidean geometry, a number of rough notes and letters (see *[196, vol. 8, pp. 407–410]* and the study of Jean Claude Pont *[435a, pp. 31–38]*). When he founded the topology of two-dimensional manifolds and laid the foundations of multidimensional manifolds Riemann also interpreted Leibniz's ideas in the same way. Riemann laid the foundations of two-dimensional topology in his *Theory of Abelian functions* (1857) *[454, pp. 82–142]*, mentioned earlier. In this work he wrote, with reference to Leibniz, that

> For the study of functions which arise as integrals of exact differentials, some theorems belonging to *analysis situs* are nearly indispensable *[454, p. 91]*.

Riemann associates to algebraic functions $f(x, y)$ in the complex variables x and y multisheeted surfaces now called *Riemann surfaces*. He divides these surfaces into

> simply connected ones, in which every closed curve bounds a region of the surface—as, for example, a disk—and multiply connected ones, for which this does not happen—as, for example, an annulus bounded by two concentric circles *[454, p. 22]*.

and points out that by means of a system of cuts it is possible to make a multiply connected surface into a simply connected one. Riemann introduces the characteristic p of a plane algebraic curve (later called by Clebsch *[119]* the genus of the curve) which he defines as half the number of cuts needed to make the corresponding multiply connected Riemann surface into a simply connected one. The number p is now called the *genus of the surface*. We note that, as S. L'Huillier showed *[323]*, for polyhedra the genus p is connected with the Euler characteristic $\chi = N_0 - N_1 + N_2$ of a polyhedron by means of the relation $\chi = 2 - 2p$. For a sphere $p = 0$, for a torus $p = 1$, and for a sphere with p handles it is equal to p. Riemann showed that for a Riemann surface the genus p, the number n of its sheets, and the number w of its branch points are connected by the relation $w - 2n = 2p - 2$.

In a survey that is a supplement to the article *On the hypotheses that lie at the foundations of geometry* Riemann adds a remark in connection with the first chapter of the article *Notion of an n-ply extended magnitude*:

> Chapter 1 is at once an introduction to investigations on *analysis situs* *[454, p. 286]*.

Fragments dealing with *analysis situs* were published in 1876 in an edition of Riemann's collected works *[454, pp. 479–482]*. In these fragments Riemann generalized the topological properties of a two-dimensional surface to an n-dimensional manifold, here called an *n-stretch* (*n-Streck*). Here Riemann defines what are now called *homologous* n-stretches:

> an n-stretch A is said to be transformable into an n-stretch B if A and parts of B together form the complete boundary of an interior $(n + 1)$-stretch *[454, p. 479]*.

Then he gives the extremely important definition:

> If in the interior of a continuously extended manifold it is possible to make every unbounded n-stretch bounding by means of m definite parts of nonbounding n-stretches then this manifold has $(m + 1)$-tuple connection of the n-th dimension.
>
> A continuously extended manifold is called *simply connected* if the connection of every dimension is simple *[454, p. 479]*.

It is easy to verify that for a two-dimensional manifold of genus p Riemann's *connection* is $(2p + 1)$-ple. Then Riemann explains the dependence of the connection of the boundary of a manifold on the connection of the manifold itself.

These ideas of Riemann were set forth by his friend Enrico Betti in the previously mentioned paper *On spaces of an arbitrary number of dimensions* *[55, vol. 2, pp. 273–290]*. By "spaces" Betti meant manifolds in multidimensional Euclidean spaces. He wrote:

> If in an n-dimensional space R, bounded by one or more $(n - 1)$-dimensional spaces, every closed m-dimensional space, $m < n$, is the boundary of a part of a linear connected $(m + 1)$-dimensional space, entirely contained in R, then we shall have a connection in $m + 1$ dimensions and we shall say that R has a *simple* connection of the m-th species. If a space R has only simple connections, then we shall say that it is simply connected. If, however, one can imagine in R a number p_m of closed m-dimensional spaces which cannot form the boundary of a linear connected part of an $(m + 1)$-dimensional space, entirely contained in R, and such that every other closed m-dimensional space forms

> by itself or with some or all of these spaces the boundary of a linear connected part of an $(m + 1)$-dimensional space, entirely contained in R, then we shall say that R has a connection of $(p_m + 1)$-th order of the m-th species *[55, vol. 2, p. 278]*.

These ideas were further developed by Poincaré. Already in the memoir *On curves defined by differential equations* (Sur les courbes definies par les équations differentielles. Paris, 1881–1885) *[431, vol. 1, pp. 90–161]* he made extensive use of topological properties of curves to give qualitative descriptions of solutions of differential equations. Poincaré devoted to the topology of multidimensional manifolds the large memoir *Analysis situs* (the title is taken over from Riemann) (Paris, 1895) *[431, vol. 6, pp. 193–288]* and five supplements to it (Palermo-London-Paris, 1899 1904).

Poincaré determined $(n - p)$-dimensional manifold in n-dimensional space by means of p equations and q inequalities between the n coordinates. Then he defined the boundary of a manifold, homomorphism between manifolds, and homology—the fundamental concept of combinatorial topology. Using this concept he defined the *Betti numbers* of a manifold, that is, Betti's *orders of connection* (in the twenties of this century Solomon Lefschetz (1884–1972) and James Alexander (1888–1971) proposed to call *Betti numbers* numbers b_i one less than Poincaré's *Betti numbers*; cf. *[308]*). Poincaré extended Euler's theorem to polyhedra with arbitrary Betti numbers by showing that

$$N_0 - N_1 + N_2 - \cdots + (-1)^n N_{n-1} = b_0 - b_1 + b_2 - \cdots - (-1)^n b_{n-1}.$$

The polynomial $\sum b_i t^i$ is called the *Poincaré polynomial*. In the same paper Poincaré also laid the foundation for the homotopic theory of manifolds in which a leading role is played by the noncommutative *Poincaré group* introduced by him.

Topological Spaces

At the beginning of the 20th century, in connection with the spread of the group-theoretic viewpoint, mathematicians began to study *abstract spaces* together with the study of topological invariants of manifolds that are submanifolds of Euclidean spaces or are obtained from such manifolds by identification of points (as in the case of obtaining the projective plane from a sphere) or by splitting them into points of various sheets (as in the case of constructing the Riemann surface of an algebraic function out of its domain of definition in the complex plane). One of the first works in this direction was the paper *On certain points of the functional calculus* (Sur quelques points du calcul fonctionelles. Palermo, 1906) *[184]* of the French mathematician Maurice Fréchet (1878–1973). In this paper Fréchet defines for the first time an *abstract metric space* exemplified by a space of functions or curves. To construct an abstract space Fréchet found it necessary

> to generalize, first of all, the theory of linear sets that has brought about so much progress in the theory of functions of a single variable,

by which theory he meant the, by that time, extensively developed theory of sets of points of the real line. Fréchet goes on to say:

> If we assume this preliminary investigation of sets then there arises a difficulty. The first generalization that seems natural is that of the notion of a continuous function. But if one wishes to consider operations where the variable is an element of arbitrary nature, then we must first know what is to be meant by neighboring elements or by the limit of a sequence of elements. This seems impossible: usually one gives a special definition of limit for each category of elements under study— points, curves, and so on. I circumvented this difficulty by a method similar to that which allows one to reason in the theory of abstract groups about a composition of explicitly indeterminate form.
>
> Now I note that almost all (but not all) classical definitions of limit can be formulated as follows: in the given category of elements one can associate to each pair of elements a number $\rho(A, B)$ whose properties are very close to the properties of distance of two points, namely that A coincides with B if $\rho(A, B) = 0$ and A tends to B if $\rho(A, B)$ tends to zero. If we accept this hypothesis, less general but nevertheless very broad, then we obtain numerous more specific results.
>
> The approach just outlined leads to the generalizing of almost all theorems about linear sets and about continuous functions (at least those that can be formulated independently of the nature of the investigated sets) *[184, pp. 1–2]*.

When mentioning abstract groups Fréchet emphasizes that he is applying the same methods of creating abstract mathematical concepts that have hitherto been used only in algebra, where, together with the abstract theory of groups, founded in the seventies of the 19th century, there also appeared Dedekind's abstract ring theory.

Developing his program, Fréchet defines a general metric space, which he calls *class* (V), as

> a set of elements of arbitrary nature such that we know how to determine whether two given elements are identical or not and, in addition, such that to any two of them, A, B, we can associate a number $(A, B) = (B, A)$ with the following two properties:
>
> 1′ The necessary and sufficient condition that (A, B) is zero is the identity of A and B.
>
> 2′ There exists a completely determined positive function $f(\varepsilon)$ such that the inequalities $(A, B) \le \varepsilon$ and $(B, C) \le \varepsilon$ imply $(A, C) \le f(\varepsilon)$ for all elements A, B, C. In other words, for (A, C) to be small it suffices that (A, B) and (B, C) are small *[184, p. 18]*.

At present, metric spaces are defined by means of three axioms: the *identity axiom* that coincides with Frechet's axiom 1′; the *symmetry axiom* $(A, B) = (B, A)$ that Frechet regards as part of the definition of distance; and the *triangle inequality axiom* $(A, B) + (B, C) \geq (A, C)$ that replaces Fréchet's axiom 2′.

Fréchet goes on to consider sets in metric spaces, limits of sequences, and limit points of sets. All limit points of a set form its *derived set*. Sets that contain all their limit points are called *closed*, and complements of closed sets with respect to the whole space are called *open* sets. The intersection of all closed sets containing a given set M is called the *closure* $\bar{\bar{M}}$ of the set M. Then, after defining the Cauchy criterion for a sequence of elements $A_1, A_2, \ldots$ of the metric space as the possibility of associating to every $\varepsilon > 0$ an integer n with the property that the inequality $(A_n, A_{n+p}) > \varepsilon$ is satisfied for every p, Fréchet defines a *complete* metric space as a metric space in which every sequence satisfying the Cauchy criterion has a, necessarily unique, limit.

Fréchet restricts himself to spaces that can be regarded in at least one way as the derived sets of countable sets of their elements. He considers metric spaces of functions with different definitions of distance.

The definition of limit points and closed sets made it possible to introduce continuous mappings and homeomorphisms of the metric spaces considered by Fréchet. Fréchet's paper became the starting point for the development of a general theory of abstract spaces and for the development of the basic concepts of functional analysis.

Since distance is not needed for the study of topological properties of abstract spaces, just as it is not needed for the study of topological properties of lines and surfaces, soon after the appearance of abstract spaces Felix Hausdorff (1868–1942) in his *Foundations of set theory* (Grundzüge der Mengenlehre. Leipzig, 1914) *[217; 218]* defined an *abstract topological space*. Hausdorff defined a topological space as any collection of elements, called *points*, with a distinguished collection of subsets $\{U\}$, such that to each point x there are associated some, and, at the very least, one of the sets of the system $\{U\}$, called *neighborhoods* $U(x)$ of x; the intersection of any two neighborhoods $U_1(x)$ and $U_2(x)$ contains a neighborhood $U_3(x)$ of this point, and if y is a point of a neighborhood $U(x)$ then there exists a neighborhood $U(y)$ of y contained in $U(x)$. Prescribing neighborhoods makes it possible to define limit points of any set M as points x such that every neighborhood of x contains at least one point of M other than x. Just as in Fréchet spaces so too in a topological (Hausdorff) space one defines derived, closed, and open sets. It is easy to see that all neighborhoods are open sets.

Every Fréchet metric space is a topological space. In fact, one can take as neighborhoods the sets of points whose distances from the points of a (fixed) countable set, whose derived set is the whole space, are less than some rational numbers. It is easy to see that the number of such neighborhoods is countable, or, as one says, the space has a *countable basis*.

A further development of the Hausdorff definition is the definition of a topological space proposed by the Polish mathematician Kazimierz Kuratowski (1896–1980) in the paper *The operation $\bar{A}$ of analysis situs* (L'operation $\bar{A}$ de l'analysis situs. Warsaw, 1922) *[296]*. By *operation $\bar{A}$* Kuratowski means the transition from a set A to its closure $\bar{A}$. According to Kuratowski, a topological space is a collection of elements of arbitrary nature, called *points*, on whose subsets A there is defined an operation of *closure* satisfying the following axioms:

1′ for the union $A + B$ of two sets A and B we have $\overline{A + B} = \bar{A} + \bar{B}$;
2′ a set A is contained in its closure, $A \subset \bar{A}$;
3′ the closure of the empty set $\varnothing$ coincides with it, $\overline{\varnothing} = \varnothing$;
4′ the closure of the closure coincides with the closure, $\bar{\bar{A}} = \bar{A}$.

If for every set there is defined its closure $\bar{A}$ then the points of $\bar{A}$ not in A are called the *limit points of* A; the sets A that coincide with their closures A are called the *closed sets*, and their complements relative to the whole space are called the *open sets*.

The Soviet mathematician Pavel Sergeevič Aleksandrov [Alexandroff] (1896–1983) proposed in the paper *On the foundation of* n*-dimensional topology* (Zur Begründung der *n*-dimensionalen Topologie. Leipzig, 1925) *[14]* a more symmetric form of Kuratowski's definition:

1. A topological space is a set of elements of arbitrary nature, called *points*, in which certain subsets, called *open sets*, have been singled out such that:
   1′ the whole space is an open set;
   2′ the null set is an open set;
   3′ the intersection of finitely many open sets is an open set;
   4′ every union of open sets is an open set.
2. A topological space is a set of elements of arbitrary nature, called points, in which certain subsets, called *closed sets*, have been singled out such that:
   1′ the whole space is a closed set;
   2′ the null set is a closed set;
   3′ every union of closed sets is a closed set;
   4′ the intersection of finitely many closed sets is a closed set.

In the first case a closed set is defined as the complement of an open set and in the second case an open set is defined as the complement of a closed set. Neighborhoods can be defined as subsystems of open sets such that an arbitrary open set can be represented as the union of sets of this subsystem (see *[218, p. 258]*).

To different systems of closed or open subsets in the same set of points there correspond different topologies on this set. For example, our axiom systems are satisfied if

(A) the only closed (and open) sets are the whole space and the null set;
(B) the closed (and open) sets are all sets of points of the space.

In case (A) the closures of all nonempty sets, including single points, coincide with the whole space. In case (B) the closure of each set coincides with this set, and a minimal system of neighborhoods is a system in which the neighborhood of each point is that point alone. Then the space has no limit points and is called *discrete*; the space of case (A) is called *trivial*.

Hausdorff ruled out case (A) by means of the following axiom: for any two points of the space there are disjoint neighborhoods *[218, p. 260]*. Sometimes Hausdorff's axiom is replaced with a weaker one due to Frigyes Riesz, that the closures of all single points of the space are the points themselves *[456, vol. 1, pp. 155 169]*—or by an even weaker axiom introduced by the Soviet mathematician Andreĭ Nikolaevič Kolmogorov (1903 1987), that any two points of the space have different closures *[15, p. 58]*.

A very important class of topological spaces is the class of *compact spaces* such that every covering of the space by means of open sets contains a finite subcovering. The Soviet mathematicians P. S. Aleksandrov and Pavel Samuilovič Uryson [Urysohn] (1898–1924) studied such spaces in the case when the minimal cardinality of a system of neighborhoods is greater than the cardinality of a countable set and called them *bicompact [16]*. For these spaces P. S. Aleksandrov developed a homology theory analogous to the homology theory of manifolds in Euclidean spaces. By now mathematicians have found a great many topological invariants of the most varied topological spaces.

One of the most obvious topological invariants is the dimension of a topological space. Whereas it is possible to have a one-to-one correspondence between manifolds of different dimensions, a one-to-one bicontinuous correspondence can only be established between manifolds of the same dimension. If the points of a topological space have neighborhoods homeomorphic to n-dimensional Euclidean space then it is natural to say that the dimension of the space is n. In more complicated cases dimension was defined by the Dutch mathematician Luitzen Egbertus Jan Brouwer (1882–1966) in the paper *Proof of the invariance of the dimension number* (Beweis der Invarianz der Dimensionzahl. Leipzig, 1911) *[80]*. The ideas of Brouwer were developed by Urysohn in *A memoir on Cantor manifolds* (Memuar o kantorovyh mnogoobraziyah. 1928) *[599]* and by the German mathematician Karl Menger (b. 1902) in *Dimension theory* (Dimensionstheorie. Leipzig, 1932) *[362]*. They defined dimension by induction on the dimension number, beginning with the dimension -1 which they assigned to the empty set.

Urysohn and Menger also proposed another definition of dimension, equivalent to the preceding for the most important topological spaces: a nonempty space is n-dimensional if every finite covering of the space has a finite subdivision of order $\leq n$ and there exists a finite covering of the space without finite subdivisions of order $< n$. A nonempty space is infinite-dimensional if it is not n-dimensional for any nonnegative n. The dimension of the null set is again taken to be -1.

In the previously mentioned paper Brouwer proved his famous *fixed-point* *theorem*: every continuous mapping of an n-dimensional simplex into itself has at least one fixed point. This theorem was first proved by the Latvian mathematician Piers Bohl (1865–1921), who worked in Riga, in the paper *On the motion of a mechanical system near an equilibrium position* (Über die Bewegung eines mechanischen Systems in der Nähe einer Gleichgewichtlage. Berlin, 1904) *[64, pp. 79–125]*. The Bohl-Brouwer theorem was extended to infinite-dimensional analogues of an n-dimensional simplex by the Soviet mathematician Andrei Nikolaevič Tihonov [Tychonoff] (b. 1906) *[597]*. Bohl came to the fixed-point theorem in connection with the problem of proving the existence of a solution of a system of differential equations connected with the notion of certain mechanical systems he investigated. The Bohl-Brouwer theorem was later used to prove the existence of solutions of finite systems of ordinary differential equations, and the Tihonov theorem was used to prove the existence of solutions of infinite systems of differential equations.

The Influence of the General Theory of Relativity

We have already pointed out the importance of Einstein's discovery of the special theory of relativity for the elaboration of the concept of pseudo-Euclidean geometry. For example, the appearance of Weyl's book *Space, time, matter [626]* in which, among other matters, he set forth his axiomatization of n-dimensional Euclidean space, was connected with the general theory of relativity. But the importance of the general theory of relativity for the development of geometry is far broader and deeper. Whereas in the special theory of relativity spacetime was viewed as a pseudo-Euclidean space, in the general theory it is viewed as an analogue of a Riemannian space that stands in the same relation to such a space as pseudo-Euclidean space to Euclidean space. Such a space is now called a *pseudo-Riemannian* or *general Riemannian space*. Just as in a Riemannian space, so too at each point of this space there is given the square of a line element

$$ds^2 = \sum_i \sum_j g_{ij} dx^i dx^j,$$

where the g_{ij} are point functions and, although the quadratic form is no longer positive definite, it can be reduced at each point to the form

$$ds^2 = -(dx^0)^2 + (dx^1)^2 + (dx^2)^2 + (dx^3)^2.$$

The first sketch of the general theory of relativity was set forth by Einstein together with the German mathematician Marcel Grossmann (1878–1936) in *Outline of a generalized theory of relativity and theory of gravitation* (Entwurf einer verallgemeinerten Relativitätstheorie und Theorie der Gravitation. Leipzig, 1913), and the final theory was presented in *Foundations of the general theory of relativity* (Grundlagen der allgemeinen Relativitätstheorie. Leipzig, 1916) *[340, pp. 109–164]*.

Already in the 1913 paper Einstein and Grossmann used the *tensor calculus* created by the Italian geometer Gregorio Ricci-Curbastro (1853–1925) in *Principles of a theory of differential quadratic forms* (Principii di una theoria delle forme differenziale quadratiche. Milan, 1884) *[453, vol. 1, pp. 138–171]* and subsequently developed together with Tullio Levi-Civita (1873–1941) in *Methods of the absolute differential calculus and their applications* (Méthodes du calcul différentiel absolu et leurs applications. Leipzig, 1901) *[453, vol. 2, pp. 185–271]*. Ricci gave the name *covariant system of the first order* to the functions a_i of the coordinates x^i that change as a result of a coordinate transformation $x^{i'} = x^{i'}(x^1, \ldots, x^n)$ in accordance with the rule

$$a_{i'} = \sum_i a_i \frac{\partial x^i}{\partial x^{i'}}, \tag{8.7}$$

exemplified by the law of transformation of partial derivatives $\frac{\partial \varphi}{\partial x^i}$, and the name *contravariant system of the first order* to functions a^i of the coordinates x^i that transform in accordance with the opposite rule

$$a^{i'} = \sum_i a^i \frac{\partial x^{i'}}{\partial x^i}, \tag{8.8}$$

exemplified by the law of transformation of the differentials dx^i of the coordinates. Ricci considered scalars as systems of zero order.

Ricci called the functions a_{ij}, a^{ij}, and a^i_j that transform in accordance with the respective rules

$$a_{i'j'} = \sum_i \sum_j a_{ij} \frac{\partial x^i}{\partial x^{i'}} \frac{\partial x^j}{\partial x^{j'}}, \quad a^{i'j'} = \sum_i \sum_j a^{ij} \frac{\partial x^{i'}}{\partial x^i} \frac{\partial x^{j'}}{\partial x^j}, \quad a^{i'}_{j'} = \sum_i \sum_j a^i_j \frac{\partial x^{i'}}{\partial x^i} \frac{\partial x^j}{\partial x^{j'}}, \tag{8.9}$$

twice covariant, *twice contravariant*, and *mixed* systems of the second order.

Ricci defined in a similar way systems of higher order. Contravariant systems of the first order at a point can be regarded as coordinates of vectors; covariant systems of the first order at the same point can be regarded as coefficients of linear forms defined on these vectors; twice covariant systems of the second order can be regarded as coefficients of bilinear forms defined on these vectors, and mixed systems of the second order can be regarded as matrices of linear transformations defined on these vectors. Einstein and Grossmann proposed that Ricci's "systems" be called *tensors* (and thus extended to them the term *elastic tensor* and that the "orders" of the systems be called *ranks*. The Dutch geometer Jan Arnoldus Schouten (1883–1971) in his *Ricci calculus* (Der Ricci-Kalkül. Berlin, 1924) *[510]* called Ricci's systems *affinors* and thereby extended to them a term used by F. Jung for linear operators; he called "orders" of the systems *valences*, borrowing this term from chemistry.

Most mathematicians now use the term *tensor* in the sense of Einstein and *valence* in the sense of Schouten. Following Einstein's suggestion, in the tensor calculus one omits the summation sign when summing with respect to covariant and contravariant indices. We will follow this rule in the following.

The coefficients of the quadratic form that defines the metric of a Riemannian space form a twice covariant tensor g_{ij}; the quadratic form is written in tensor notation as

$$ds^2 = g_{ij}dx^i dx^j, \tag{8.10}$$

and the equation of the geodesics is written as

$$\frac{d^2x^i}{dt^2} + \Gamma^i_{jk}\frac{dx^j}{dt}\frac{dx^k}{dt} = 0. \tag{8.11}$$

The Christoffel symbols Γ^i_{jk} do not form a tensor and transform under coordinate transformations in accordance with a more involved law. But the four-index Riemann symbols, now written as $R_{ij,k}^{\cdot\cdot\cdot l}$, do form a tensor of valence four called the *curvature tensor*.

Covariant differentiation as defined by Ricci associates to every tensor field in a Riemannian space—that is, to every function that defines a tensor at every point of a region in a Riemannian space—the field of a new tensor that has one additional covariant valence. The covariant derivative $\nabla_i\varphi$ of a scalar is its partial derivative $\dfrac{\partial\varphi}{\partial x^i}$ with respect to the coordinate x^i. The covariant derivatives $\nabla_j a^i$ and $\nabla_j a_i$ of vectors a^i and a_i are given by

$$\nabla_j a^i = \frac{\partial a^i}{\partial x^j} + \Gamma^i_{jk}a^k, \qquad \nabla_j a_i = \frac{\partial a_i}{\partial x^j} - \Gamma^k_{ji}a_k. \tag{8.12}$$

The covariant derivatives $\nabla_k a_{ij}$ and $\nabla_k a^i_j$ of the tensors a_{ij} and a^i_j have the form

$$\nabla_k a_{ij} = \frac{\partial a_{ij}}{\partial x^k} - \Gamma^l_{ki}a_{lj} - \Gamma^l_{kj}a_{il}, \qquad \nabla_k a^i_j = \frac{\partial a^i_j}{\partial x^k} + \Gamma^i_{kl}a^l_j - \Gamma^l_{kj}a^i_l. \tag{8.13}$$

The Geometry of the General Theory of Relativity

Einstein formulated the principle of general relativity in *The foundations of the general theory of relativity* as follows:

> The general laws of nature are to be expressed by equations which hold good for all systems of co-ordinates, that is, are covariant with respect to any substitutions whatever (generally covariant) *[340, p. 117]*.

The most important element of the general relativity theory is the geometric interpretation of gravity. In this theory gravitation is linked to the curvature of space, whose definition in a pseudo-Riemannian space is the same as that in a Riemannian space; the greater the density of matter in a certain region, and thus the intensity of the gravitational field, the greater the curvature of pseudo-Riemannian spacetime. A mass point subject to gravitational forces alone moves along a geodesic which satisfies the same equations as in Riemannian space.

Light rays, too, propagate along geodesics of pseudo-Riemannian spacetime. The deflection of the trajectories of material particles and of light rays from straight lines represents the attraction of particles of matter and of light by the heavy masses that gives rise to the gravitational field. In this connection we quote from Einstein's paper *On the ether* (Über den Aether. Zürich, 1924) *[166]*:

> General relativity theory sets aside one other shortcoming of classical dynamics: in the latter, inertia and weight appear as phenomena that are completely independent from one another in spite of the fact that both are conditioned by the same material constant -- mass. Relativity theory surmounts this shortcoming by establishing for the dynamical behavior of an electrically neutral particle the law of geodesics in which the interaction of inertia and gravitation is already inseparable. This interaction imparts to the ether a metric that varies from point to point and properties that determine the dynamical behavior of material points. In turn, these are determined by physical factors, namely the distribution of mass or energy. Thus the ether of general relativity differs from the ether of classical mechanics in that it is no longer "absolute" but is determined in the sense of its properties, that vary in space, by the distribution of weighted matter. This definition is complete provided that the world is spatially finite and closed *[166, p. 89]*.

Here the word *ether* is used in the sense of curved space whose geometric properties are determined by matter; in other papers Einstein replaces this term by the term *space-time continuum*. Einstein's reference to the finite and closed character of the universe reflects his original view that the curvature of space implies that space is finite and closed. (Cf. the paper *[649]* of Yakov Borisovič Zel'dovič (1914–1988).)

General relativity theory has given a physical interpretation to the curvature of pseudo-Riemannian spacetime. Light rays propagate along the isotropic geodesics of this spacetime. The deviation of the trajectories of material points and light rays from geodesics represents the attraction of material points and light by the heavy masses that give rise to the gravitational field.

The primary connection between non-Euclidean geometry and general relativity is that the discovery of non-Euclidean geometry extended the range of conceivable spaces and thus prepared the ground for the discovery of spaces of variable curvature, both Riemannian and pseudo-Riemannian; the latter represents the mathematical apparatus of general relativity. However, there is also a direct link between general relativity and non-Euclidean

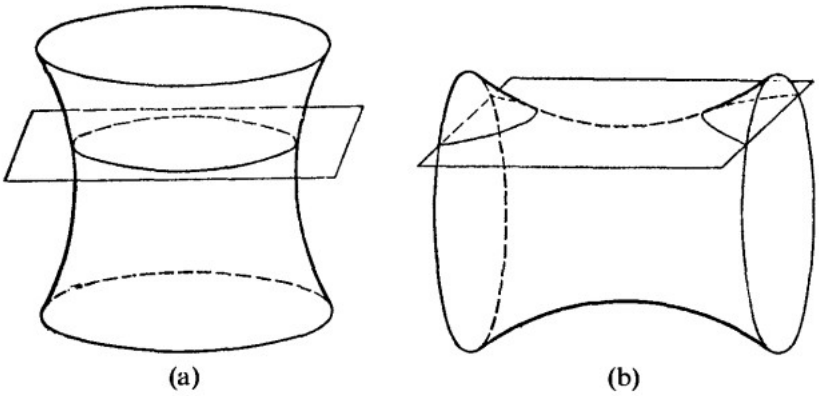

Figure 105

geometry, but what is involved is a four-dimensional rather than a three-dimensional space as in the case of special relativity. The geometry in question describes the spacetime of general relativity "in the mean," under the assumption of the uniform distribution of matter. This view was presented by Einstein already in the paper *Cosmological considerations linked to the general relativity theory* (Kosmologische Betrachtungen zur allgemeinen Relativitätstheorie. Berlin, 1917) *[340, pp. 175–189]*. Here Einstein considered a uniform distribution of matter in space alone, regarding space itself as a sphere in four-dimensional Euclidean space and spacetime as a cylinder in five-dimensional space. However, if we proceed from the assumption that we are dealing with a uniform distribution of matter throughout the spacetime continuum then spacetime is a sphere of real or imaginary radius in five-dimensional pseudo-Euclidean spaces of indices 4 and 3, respectively. If we assume, for simplicity, that the hypersurfaces $t =$ const. are parallel hyperplanes, then, in time, the "space section" of the world decreases or increases, depending on the position of the cutting hyperplane. In the first case the curvature of the "space section" is constant and positive, whereas in the second case it is constant and negative (Figures 105a and 105b), in full agreement with astronomical observations. This confirmation shows that real spacetime, that is, a pseudo-Riemannian space of variable curvature, corresponds in the mean to the picture just outlined; the theory of an expanding universe of constant negative curvature was proposed by A. A. Friedmann in the paper *On the possibility of a world of constant negative curvature* (Über die Möglichkeit einer Welt mit konstanter negativer Krümmung des Raumes. Leipzig, 1924) *[188]*.

The view of spacetime as a four-dimensional pseudo-Riemannian space of index 3 with positive curvature was proposed by the Belgian astronomer Willem de Sitter (1872 1934) in the paper *On Einstein's theory of gravitation and its astronomical consequences* (Brussels, 1917) *[541]*. In the paper *Geometrical note on de Sitter's world* (London, 1924) *[158]* Patrick du Val showed that "de Sitter's world" is a Lobačevskian space of positive curvature, that is, a sphere of real radius in five-dimensional pseudo-Euclidean space of index 4 with identified antipodal points (see also *[129a]*).

Einstein's Philosophy of Space

The discovery of the general theory of relativity was a brilliant concretization of the connection between matter and space and time, which are an existence form of matter— an issue often addressed by the classics of Marxism.

Einstein dealt with the philosophical questions of space in many of his papers. One relevant passage is the preceding quotation from his paper *On the ether*. The following quotation is from the paper *Geometry and experience* (Geometrie und Erfahrung. Berlin, 1921) *[165]*:

> Of all sciences mathematics is held in especially high esteem, for its theorems are absolutely true and indisputable while the areas of other sciences are to some extent debatable and there is always the danger of their results being overturned by new discoveries. But it is not fitting for a researcher working in some other area of science to envy mathematics, for the propositions of mathematics rest not on real objects but exclusively on objects of our imagination. In fact, it is little wonder that one attains logical agreement of deductions if one earlier reached agreement concerning the principal propositions (axioms) as well as the methods to be employed for deriving from the principal propositions other propositions. At the same time, the deep esteem for mathematics has another basis, which is that mathematics is that which gives to the exact sciences a measure of confidence which they could not otherwise attain.
>
> In this connection there arises a riddle that has worried researchers of all times. Whence the remarkable correspondence between mathematics and real things if mathematics is just a product of human thought unrelated to any experience? Can the human mind understand the properties of real things without any experience, just by way of reflection?
>
> To my mind, a concise answer to this question is this: to the extent to which the theorems of mathematics can be applied to reflect the real world they are not exact; they are exact to the extent to which they do not refer to reality. It seems to me that complete clarity in this matter can be achieved only by following the kind of mathematics known as axiomatics. The progress achieved through axiomatics consists in the fact that it sharply separated its formal-logical content from its objective and intuitive content. By point, straight line, and so on, in axiomatic geometry one should mean only contentless ideas. That which gives them content lies outside mathematics.
>
> On the other hand, however, it is true that mathematics in general and geometry in particular owe to their origins the necessity of learning

> a little about the behavior of objects existing in reality. This is shown by the very word "geometry" which means "measurement of the earth" *[165, pp. 3–5]*.

In this paper Einstein arrives

> at the viewpoint held by as original and deep a thinker as Henri Poincaré: Euclidean geometry differs from all conceivably possible axiomatic geometries by its simplicity. And since axiomatic geometry contains no statements about actual reality and can make such statements only in conjunction with physical laws, it seems possible and reasonable to adhere to Euclidean geometry regardless of the properties of reality. On the other hand, in case of an observed disagreement between theory and experience, it is easier to agree to change physical laws than axiomatic Euclidean geometry. If one forgets all about the connection between a practically rigid body and geometry, then it is not easy to refuse to admit that Euclidean geometry must be viewed as simplest.
>
> Why did Poincaré and other researchers reject the equivalence that thrusts itself upon one between a practically rigid body of real experience and a geometric body? Simply because, when one considers them more closely, real rigid bodies in nature turn out not to be rigid at all, since their geometric behavior, that is, their possible mutual disposition, depends on temperature, external forces, and so on. Thus the initial direct connection between geometry and physical reality is destroyed and we are forced to shift to the following more general position characteristic of Poincaré's viewpoint. Geometry (G) tells us nothing about the behavior of real objects; this behavior is described only by geometry together with the totality of physical laws (Ph). In symbolic terms we can say that only the sum (G) + (Ph) lends itself to experimental verification. Thus one can choose arbitrarily (G) as well as different parts of (Ph); all of these laws represent agreements. To avoid contradictions it is necessary to choose the rest of (Ph) so that (G) and all (Ph) are experimentally verifiable *[165, pp. 7–8]*.

We see that Einstein refers here to the philosophical works of Poincaré mentioned earlier but, unlike Poincaré, who treated "objective reality" and "laws of nature" in the spirit of Mach, Einstein proceeds from "physical reality" and the possibility of "choosing arbitrarily" (G), that is, geometric axioms which, together with (Ph), that is, physical laws, must confront experience. Thus Einstein interprets Poincaré's position in materialistic terms and, like Lobačevskiĭ, believes that the question of which of the conceivable geometric systems corresponds to the geometry of the real world must be decided experimentally.

It should be pointed out that Einstein frequently refers to Mach, whom he regards as one of his predecessors. In the obituary *Ernst Mach* (Ernst Mach. Leipzig, 1916) Einstein considered Mach's critique of Newton's mechanics and wrote:

> The quoted lines show that Mach clearly understood the weak points of classical mechanics and came close to discovering the general relativity theory. And that half a century before its creation! *[167]*.

Einstein regarded Mach's idea that the inertia of a body is explainable by its interaction with other bodies—what Einstein called *Mach's principle*—as one of the factors in the preparation of the general relativity theory. Mach's idea that "space and time cannot be completely separated from one another" *[347, p. 350]* is very close to the ideas of special relativity theory.

Undoubtedly, Einstein's famous *thought experiments* were influenced by Mach. At the same time Einstein did not agree with Mach's doctrine of sensations as elements of the world. He not only interpreted the terms *sensations* and *experience* in a materialist spirit but also ascribed his viewpoint to Mach when he said (apropos Mach's use of the term *sensations*) that

> frequently, as a result of insufficient familiarity with his, that is, Mach's, works, some persons tend to confuse them with the terminology of philosophical idealism and solipsism *[167]*.

Here Einstein's opinion was based on the fact that he took Mach's materialist terminology literally and throught it inconceivable that a serious physicist could be a subjective idealist. Later Einstein changed his opinion about Mach's philosophy and during a meeting of the French philosophical society, while answering a question from Emile Meyerson, he referred to Mach as a "deplorable philosopher" (*un déplorable philosophe*) and accused him of a "myopic view of science that led him to reject the existence of atoms" *[364, p. 62]*.

Parallel Displacement

The heightened interest in Riemannian geometry brought about by general relativity theory resulted in an extremely important discovery in this geometry, a concept known as *parallel displacement of vectors*. The discovery was made by Levi-Civita in the paper *The notion of parallelism in an arbitrary manifold and the geometric characterization of Riemannian curvature implied by it* (Nozione di parallelismo in una varieta qualunque e consequente specificazione geometrica della curvatura Riemannina. Palermo, 1917) *[320]*.

If a vector a^i is defined at a point A of a Riemannian space then the result of its displacement to a point B of the line $x^i(t)$ is a solution of the differential equation

$$\frac{da^i}{dt} + \Gamma^i_{jk} a^j \frac{dx^k}{dt} = 0, \tag{8.14}$$

that can be written with the help of the covariant derivative as

$$\nabla_k a^i \frac{dx^k}{dt} = 0. \tag{8.15}$$

It follows that if one defines along a coordinate line a vector field of vectors obtained by parallel displacement of a given vector then at all points of this line the covariant derivative with respect to the corresponding coordinate is equal to zero. Parallel displacement of vectors admits of a visual geometric definition. Parallel displacement of vectors from a point A of the space to an infinitely near point B is generated by mapping a neighborhood of the point A onto a neighborhood of the point B that is the result of applying first a reflection (along geodesics) in the point A in the neighborhood of A, and then a reflection in C, the midpoint of the geodesic arc from A to B, in a neighborhood of C that contains the neighborhood of A. Under this neighborhood mapping the increments of the coordinates of the point A map to the increments of the coordinates of the point B and, consequently, the differentials of the coordinates of A, that is, the vectors at A, map to the differentials of the coordinates of B, that is, the vectors at B. The name *parallel displacement* is due to the fact that the product of two reflections in two points in ordinary space is a translation of this space. Levi-Civita showed that the curvature of a Riemannian space, at a given point, in a two-dimensional direction is the limit of the ratio of the angle by which a vector is rotated as a result of parallel displacement along a closed contour passing through the given point to the area of the surface bounded by this contour when the contour shrinks to the given point (this is Levi-Civita's *geometric characterization of Riemannian curvature*).

The Problem of a Unified Field Theory

In general relativity theory the geometric properties of space are determined by the gravitational field alone, without any effect of the electromagnetic field on the geometry of space. After creating the general theory of relativity Einstein posed the problem of constructing a geometry of four-dimensional spacetime determined by the electromagnetic as well as the gravitational field. The first attempt of construction of a unified field theory was made by Hermann Weyl in his paper *Gravitation and electricity* (Gravitation und Elektrizität. Leipzig, 1918) *[340, pp. 200–216]*. Besides the metric quadratic form

$$ds^2 = \sum_{\mu,\nu} g_{\mu\nu} dx^\mu dx_\nu \tag{2}$$

Weyl considers also the linear form

$$d\phi = \sum_{\mu} \phi_\mu dx_\mu \tag{7}$$

and writes that

> *The internal metrical connexion of space thus depends on a linear form* (7) *besides the quadratic form* (2)—which is determined except as to an arbitrary factor of proportionality *[340, p. 206]*.

The vector ϕ_μ is the *vector-potential* of the electromagnetic field $F_{\mu\nu}$ given by

$$F_{\mu\nu} = \partial\phi_\mu/\partial x_\nu - \partial\phi_\nu/\partial x_\mu.$$

The geometry constructed by Weyl is called the *Weyl connection*. Weyl expounded this theory also in his *Space, Time and Matter* (1918) *[626]*.

Einstein himself devoted the last 35 years of his life to the search for a unified field theory. Of other attempts at constructing a unified field theory we note, first of all, the paper of Theodor Kaluza (1885–1954) *On the unicity problem of physics* (Zum Unitätsproblem der Physik. Berlin, 1921) *[259]*. To increase the number of relevant parameters Kaluza went in the direction of increasing the number of dimensions of his space and assumed that in addition to the four dimensions of physical spacetime there is a fifth dimension without a direct physical sense. Proceeding from Kaluza's idea the mathematician Veblen and the physicist Pauli proposed in 1930–1933 a unified field theory in which five coordinates are viewed as the projective coordinates of particles in a four-dimensional space.

A detailed history of unified field theories in the first third of the 20th century is given by Vladimir Pavlovič Vizgin in the book *Unified field theories in the first third of the 20th century* (Edinye teorii polya v pervoĭ treti XX veka. Moscow, 1985) *[615]*.

We note that when Einstein posed the question of a unified field theory he assumed that the theories of interactions of the gravitational and electromagnetic fields exhaust all forms of physical interactions. The discovery, in the last few decades, of "strong" and "weak" interactions that do not coincide with these interactions deprived the problem posed by Einstein of the importance he attached to it. In his "Remarks to the Einsteinian sketch of a unified field theory" the famous German physicist Werner Heisenberg (1901–1976) wrote:

> This attempt, grandiose in its conception, failed from the very beginning. At the time when Einstein concerned himself with the unified field theory new elementary particles were continuously discovered and with them their associated new fields. This meant that there existed no hard, solid, empirical basis for the implementation of the Einsteinian program and Einstein's attempts produced no conclusive results whatever. However, the failure that dogged the Einsteinian program also had deeper foundations than the uncertainty of empirical facts; these foundations have to do with the relations of Einstein's field-theoretic view of quantum theory. Einstein proceeded from the classical nonlinear field theory of matter for the metric tensor of the field that defined his geometry.

> Einstein hoped that in the end it will be possible to think of atoms and elementary particles in such a theory as singular solutions of the nonlinear equations of the field. In reality, however, the element of discreteness that expresses the existence of elementary particles has a far wider character. It becomes dominant as soon as we go over to the domain of atoms and elementary particles; which is why their description by means of a field theory of the classical type is out of the question. Conversely, we now know that there operate here quantum-mechanical laws whose structure was understood in the twenties. Since Einstein could not come to terms with such a structure, he did not make an attempt to approach the unified field theory using quantum laws.
>
> Nevertheless, experiments with elementary particles carried out in connection with quantum theory contain very many arguments in favor of Einstein's program.
>
> In the last ten years many elementary particles have been discovered and thus also many new fields. Also, it became clear that elementary particles can change into one another. If two elementary particles with very high kinetic energy collide in some way then there can arise new particles, and the laws that govern the coming into existence and the disappearance of elementary particles can, apparently, be formulated by means of relatively simple selection rules and the corresponding quantum numbers. It follows that all elementary particles "consist," so to say, of one and the same substance that can be simply called "energy" or "matter"; their structure and their ability to go over into one another should follow from a simple law for matter.
>
> Thus, by the nature of things, a satisfactory theory of elementary particles must, at the same time, be a unified field theory of matter *[221a, pp. 120 - 125]*.

An attempt to create such a unified theory of elementary particles was made by Heisenberg himself.

Spaces with Affine Connection

Although the problem of a unified field theory has not yielded any substantial results for physics, it has been remarkably fruitful for geometry. Weyl's theory was further developed by Schouten, who introduced the general notion of a space with an affine connection in the paper *On different forms of connection that can be laid at the foundation of a differential geometry* (Über die verschiedenen Arten der Übertragung die einer Differentialgeometrie zugrunde gelegt werden können. Berlin, 1922) *[511]*. A detailed exposition of Schouten's theory is found in his *Ricci calculus*. A space with an affine connection is defined as a manifold of points with coordinates x^i at each of whose points there is given a point function $\Gamma^i_{jk} = \Gamma^i_{kj}$ that is not expressible in the general case in terms of the point functions g_{ij}. Because of this last condition it is not possible, in the general case, to define in a space with an affine connection the length of lines, but it *is* possible to define parallel displacement of vectors by formula (8.14). In a space with an affine connection geodesics are defined. They are solutions of the equations (8.11). For each arc of a geodesic there is defined its midpoint. In such a space one can define reflections in points along geodesics. Just as in a Riemannian space, the geometric sense of parallel displacement of vectors is that it is determined by a mapping of a neighborhood of a point of the space onto a neighborhood of an infinitely near point that is the result of two reflections in two points along geodesics. The name *affine connection* is due to the fact that the mapping of vectors at a point of space onto vectors at an infinitely near point of this space is an affine transformation of vector spaces.

It is clear that Riemannian and pseudo-Riemannian spaces are special cases of spaces with an affine connection. In such spaces the mapping of vectors at one point onto vectors at an infinitely close point under parallel displacement is an isometry of Euclidean and pseudo-Euclidean spaces, respectively.

The space introduced by Weyl in his unified field theory is also a space with an affine connection. In this space the mapping of vectors at one point onto vectors at an infinitely close point under parallel displacement is a similarity transformation.

Even more general spaces were introduced by the French mathematician Élie Cartan (1869–1951). In the papers *On manifolds with an affine connection and the generalized relativity theory* (Sur les variétés à connexion affine et la théorie de la relativité généralisée. Paris, 1923) *[96, part 3, pp. 659–746, 798–825, 921–992]*, *Spaces with a conformal connection* (Les espaces à connexion conforme. Paris, 1923) *[96, part 3, pp. 747–797]* and *On manifolds with a projective connection* (Sur les variétés à connexion projective. Paris, 1924) *[96, part 3, pp. 825–861]* Cartan introduced the notion of a *space with a homogeneous connection*. Whereas in the case of a space with an affine connection there is associated to every point a vector space that can be regarded as an affine space, and under parallel displacement of vectors the space at one point is mapped by an affine transformation onto the space at an infinitely near point, in the case of a space with a homogeneous connection there corresponds to every point of the space a homogeneous space with a definite group of transformations. Also, for every pair of infinitely near points there is specified a transformation of the spaces associated with these points that preserves their geometry. In addition to spaces with an affine connection and special cases of such spaces noted above, the most important spaces with a homogeneous connection are *spaces with a projective* and *conformal connection*, respectively. To each point of these spaces there is associated, respectively, a projective or conformal space, and for each pair of infinitely near points of a space with a connection there is specified, respectively, a projective or conformal mapping of one of the spaces associated with these points onto the other.

Differentiable Manifolds

The spaces introduced by Riemann and the groups introduced by Lie, bearing the names of their respective inventors, as well as pseudo-Riemannian spaces, Weyl spaces, and the more general spaces with an affine or other homogeneous connection were defined only locally—in the domain of a certain coordinate system. The introduction of topological manifolds made possible the study of differential-geometric spaces in the large. In every n-dimensional topological manifold one can define a homeomorphic mapping of a neighborhood U of every point onto some domain of n-dimensional Euclidean space, and the coordinate system in the Euclidean space provides coordinates for the neighborhood U. This system of coordinates in the neighborhood U is now called a *system of local coordinates* or a *local map* of the manifold, and the totality of local maps is called the *atlas* of the manifold. A manifold is called *differentiable of class* v if (1) the domains of the maps of the atlas cover the whole manifold; and (2) if U and U' are two domains of maps of the atlas with nonempty intersection and a point x in the intersection has coordinates x^i and $x^{i'}$ in the two maps then the functions $x^{i'} = f(x^1, \dots, x^n)$ have continuous partial derivatives up to and including order v, and the jacobian $\det\left(\frac{\partial x^{i'}}{\partial x^j}\right)$ is different from zero throughout the intersection of the domains. The manifold is called *analytic* if the functions in (2) are analytic.

The definitions of a differentiable manifold and of an analytic manifold were formulated by Oswald Veblen (1880–1960) and by John Henry Constantine Whitehead (1904–1960) in *The foundations of differential geometry* (Cambridge, 1932) *[612]*.

With every point of a differentiable manifold there is associated a vector space with vectors $\{dx^i\}$ whose coordinates are the differentials of the coordinates x^i; these spaces are called the *tangent spaces* of the differentiable manifold. Under coordinate transformations $x^i \to x^{i'}$ the coordinates of the vectors and tensors of the tangent spaces transform according to the rules (8.7)–(8.9), and so on. If in every tangent space of a differentiable manifold there is given a metric tensor g_{ij} that determines an inner product in the Euclidean or pseudo-Euclidean space then the differentiable manifold is called, respectively, a *Riemannian* or *pseudo-Riemannian space in the large*. If in each of the tangent spaces of a differentiable manifold there is given an object Γ^i_{jk} that transforms under coordinate transformations according to the same rule as the coefficients Γ^i_{jk} of Riemannian spaces then the differentiable space is called a *space with an affine connection in the large*.

Fibrations

Spaces with connections are special cases of fiber spaces. Unlike spaces with Schouten-Cartan connections, fiber spaces are considered in the large. The initial form of the theory of fiber spaces in the large was the theory of "sphere-spaces," that is, fiber spaces whose "fibers" are spheres. This theory arose in connection with the solution of the problem of continuous mapping of an n-dimensional sphere onto an m-dimensional one for, if $n > m$, then to every point of the m-dimensional sphere there corresponds an $(n - m)$-dimensional manifold of the n-dimensional sphere, so that the latter "stratifies" or "fibrates" into these manifolds; the resulting fibres are also called *spheres*. The theory of such spaces was constructed by the American mathematician Hassler Whitney (b. 1907) in the paper *Sphere-spaces* (Boston, 1935) *[632]*. In the paper *On the theory of sphere bundles* (Boston, 1940) *[633]* Whitney calls these spaces *sphere-bundles*. The most complete account of Whitney's theory is contained in his lectures *On the topology of differentiable manifolds* (Ann Arbor, 1941) *[634]*. Together with *sphere-bundles* Whitney also considered *plane-bundles*. Whitney associated to every differentiable manifold a *tangent plane-bundle* consisting of the tangent spaces of the manifold and *tangent sphere-bundle* consisting of spheres in these tangent spaces with centers at the points of tangency, with one sphere in each tangent space. In the first case, the points of the *fibers* characterize the vectors of the differentiable manifold; in the second, they characterize the directions issuing from a point of the manifold.

Soon Whitney's *sphere-bundles* were replaced by arbitrary manifolds. Then there arose the more general concept of a *fiber bundle* or a *fiber space*. The general theory of such spaces was set forth by the American mathematician Norman Steenrod (1910–1971) in the monograph *The topology of fiber bundles* (Princeton, 1951) *[560]*, which completed the development of this theory.

The simplest fiber space is the *direct product* of two topological spaces S and T, that is, the totality of pairs (s, t) of elements s in S and t in T where the closed sets of the topological product are sets of pairs (s, t) in which both elements s and t vary over closed sets in the spaces S and T, respectively. In this case one may view the set of pairs (s, t_0) with fixed t_0 as a *fiber* and the space T of all possible elements t_0 as a basis of the fiber space. One can also view the set of pairs (s_0, t) with fixed s_0 as a *fiber* and the space S as a basis of the same fiber space. In this case the fibration is called *trivial*.

In more complicated cases, a fibre space is not simply the direct product of a *fiber* by a *basis* but also consists of homeomorphic fibers, and, if the latter are viewed as points of some new space, then this space can be regarded as a basis of the fiber space. The simplest fiber space of this kind is a *Möbius strip* obtained from the lateral surface of a right circular cylinder by identification of pairs of points that are symmetric with respect to its center, or from a rectangle (which may be thought of as the application of half of the preceding cylinder to a plane) by gluing the points of a pair of its opposite sides that are symmetric with respect to its center. (Hence the name *skew products* for general fibrations of a space used in the Russian translation of N. Steenrod's book *[560]*). One usually considers also the group of topological mappings of a fiber onto itself that preserves a certain structure of these fibers and mappings of fibers, one onto another, that preserve this structure.

The most important fiber spaces are the *tangent plane-bundles* defined by Whitney for differentiable manifolds. Now they are called simply *tangent bundles* of differentiable manifolds. The bases of these fiber spaces are the differentiable manifolds, and the fibres are their tangent spaces. To the class of fiber spaces there belong Riemannian spaces, pseudo-Riemannian spaces, and spaces with an affine connection in the large together with their tangent spaces that are, respectively, Euclidean, pseudo-Euclidean, and affine, as well as spaces with a projective, conformal, and arbitrary homogeneous connection together with the corresponding homogeneous spaces that also form *fibers*. The groups of transformations of the *fibers* of these spaces are, respectively, the groups of Euclidean and pseudo-Euclidean motions, of affine, projective, conformal, and other transformations.

Exterior Forms, Curvature and Betti Numbers

The field of a skew-symmetric covariant tensor $a_{i_1 i_2 \ldots i_k}$ in a domain of a differentiable manifold defines at every point of this domain an exterior differential form

$$\omega_k = \sum_{i_1} \cdots \sum_{i_k} a_{i_1 i_2 \ldots i_k} dx^{i_1} \wedge dx^{i_2} \wedge \cdots \wedge dx^{i_k},$$

where $\wedge$ denotes exterior multiplication. The form is invariant under even permutations of the dx^{i_α} and changes sign under odd permutations of the dx^{i_α} (in virtue of the skew symmetry of $a_{i_1 \ldots i_k}$ the ordinary differential form $\sum_{i_1} \ldots \sum_{i_k} a_{i_1 i_2 \ldots i_k} dx^{i_1} dx^{i_2} \ldots dx^{i_k}$ is equal to zero). For exterior forms we define the operations of exterior multiplication

$$\omega_{k+l} = \omega_k \wedge \omega_l$$

and exterior differentiation

$$\omega_{k+1} = \mathscr{D}\omega_k = \sum_{i_1} \cdots \sum_{i_k} da_{i_1 \ldots i_k} \wedge dx^{i_1} \wedge \cdots \wedge dx^{i_k}.$$

The differential forms in multiple integrals defining areas and volumes are exterior forms, and the Ostrogradskiĭ formula in the beginning of chapter 7 can be written as

$$\int_{\partial M} \omega = \int_M \mathscr{D}\omega,$$

where M is a bounded domain and ∂M is its boundary.

The theory of exterior forms was founded by É. Cartan in the paper *On certain differential expressions and the Pfaff problem* (Sur certaines expressions différentielles et le problème de Pfaff. Paris, 1899) *[96, part 2, pp. 303–396]*. It was extensively developed by É. Cartan (see his book *Exterior differential systems and their geometrical applications* (Les systèmes différentiels extérieurs et leurs applications géométriques. Paris, 1945) *[97]*, by the Soviet geometer Sergeĭ Pavlovič Finikov (1883–1964), and by their students.

If a moving orthonormal frame $\{\mathbf{e}_i\}$ is associated with every point of a curve or surface in the space R_n or with every figure of a family of figures in this space, then the differentials of the radius vector $\mathbf{x}$ and of the vectors $\mathbf{e}_i$ of the frame can be written as

$$d\mathbf{x} = \mathbf{e}_i \omega^i, \qquad d\mathbf{e}_i = \mathbf{e}_j \omega_i^j, \tag{8.16}$$

where

$$\omega_j^i = -\omega_i^j, \tag{8.17}$$

and the exterior differentials of the forms ω^i and ω_j^i satisfy the *equations of structure*

$$\mathscr{D}\omega^i = \sum_k \omega^k \wedge \omega_k^i, \qquad \mathscr{D}\omega_j^i = \sum_k \omega_j^k \wedge \omega_k^i. \tag{8.18}$$

A curve, surface or family of figures is defined by a system of Pfaff equations $\theta^1 = \theta^2 = \cdots = \theta^r = 0$, where the forms θ^A are linear combinations of the forms ω^i, ω_j^i. For a curve with arc length s and tangent vector $\mathbf{e}_1$ these equations are $\omega^2 = \cdots = \omega^n = 0$ and for a hypersurface with normal vector $\mathbf{e}_n$ the equation is $\omega^n = 0$. If for the curve we put $d\mathbf{e}_1/ds = k_1\mathbf{e}_2$ and choose vectors $\mathbf{e}_3, \ldots, \mathbf{e}_n$ so that $d\mathbf{e}_i/ds$ is a linear combination of $\mathbf{e}_1, \ldots, \mathbf{e}_{i+1}$, then $\omega_i^{i+1}/ds = -\omega_{i+1}^i/ds = k_i$ are curvatures of the curve and the remaining ω_j^i are zeros. In this way we obtain the *Frenet formulas*

$$\begin{aligned} &d\mathbf{e}_1/ds = k_1\mathbf{e}_2, \\ &d\mathbf{e}_2/ds = -k_1\mathbf{e}_1 + k_2\mathbf{e}_3, \ldots, \\ &d\mathbf{e}_i/ds = -k_{i-1}\mathbf{e}_{i-1} + k_i\mathbf{e}_{i+1}, \ldots, d\mathbf{e}_n/ds = -k_{n-1}\mathbf{e}_{n-1} \end{aligned}$$

discovered for R_3 in 1847 by Frederic Frenet (1816–1900) and for R_n in 1874 by C. Jordan. The curvatures k_i define the curve up to a motion of R_n.

For a hypersurface we direct the vectors $\mathbf{e}_1, \ldots, \mathbf{e}_{n-1}$ along curvature lines of the hypersurface and the vector $\mathbf{e}_n$ along its normal. Exterior differentiation of the Pfaff equation $\omega^n = 0$ gives, in view of (8.18), $\mathscr{D}\omega^n = \sum_k \omega^k \wedge \omega_k^n = 0$. Hence $\omega_\alpha^n = \sum_\beta b_{\alpha\beta}\omega^\beta$ $(\alpha, \beta = 1, 2, \ldots, n-1)$. We note that the analogues of Gauss's quadratic forms for surfaces in R_3 and for hypersurfaces in R_n are

$$d\mathbf{x}^2 = \sum_\alpha (\omega^\alpha)^2,$$

$$\mathbf{e}_n d^2\mathbf{x} = -d\mathbf{e}_n d\mathbf{x} = -\left(-\sum_\alpha \omega_\alpha^n \mathbf{e}_\alpha\right)\left(\sum_\beta \omega^\beta \mathbf{e}_\beta\right) = \sum_\alpha \sum_\beta b_{\alpha\beta}\omega^\alpha\omega^\beta.$$

In our basis $b_{\alpha\beta} = b_\alpha\delta_{\alpha\beta}$, $-b_\alpha = k_\alpha$ are main curvatures of the hypersurface; they are differential invariants of the hypersurface. Exterior differentiation of other forms yields a complete system of differential invariants of the hypersurface which defines it up to a motion of the space R_n.

Formulas (8.16) and (8.18) also hold in the affine space E_n and in the pseudo-Euclidean space ${}^lR_n$. In the case of E_n the forms ω_j^i are not connected by any condition. In the case of ${}^lR_n$ with scalar square of a vector $\mathbf{x}^2 = \sum_i \varepsilon_i (x^i)^2$ the conditions (8.17) become

$$\varepsilon_i \omega_j^i = -\varepsilon_j \omega_i^j. \tag{8.19}$$

In the spaces P_n, S_n and ${}^lS_n$ the points are represented by vectors of the spaces E_{n+1}, R_{n+1} and ${}^lR_{n+1}$, respectively. In these spaces the role of the radius vector $\mathbf{x}$ is played by e_0. In these spaces hold the second formulas (8.16) and (8.18). In P_n the forms ω_j^i are connected by the condition $\sum_i \omega_i^i = 0$, in S_n by the condition (8.17), and in ${}^lR_n$ by the condition (8.19).

Analogous methods can be applied in all homogeneous spaces.

These methods can also be applied in spaces with affine connection and in Riemannian and pseudo-Riemannian spaces. In these spaces the vectors $\mathbf{e}_i$ are vectors of the tangent spaces of E_n, R_n and ${}^lR_n$ respectively. In these spaces hold formulas (8.16) and in Riemannian and pseudo-Riemannian spaces formulas (8.17) and (8.19), respectively. The equations of structure there have the form

$$\mathscr{D}\omega^i = \sum_k \omega^k \wedge \omega_k^i + \sum_j \sum_k S_{jk}^i \omega^j \wedge \omega^k,$$
$$\mathscr{D}\omega_j^i = \sum \omega_j^k \wedge \omega_k^i + \sum_k \sum_h R_{jkh}{}^{\cdot\cdot\cdot i} \omega^k \wedge \omega^h$$

where S_{jk}^i is the *torsion tensor* (equal to zero in spaces without torsion, in particular, in Riemannian and pseudo-Riemannian spaces), $R_{jkh}{}^{\cdot\cdot\cdot i}$ is the *curvature tensor* (in Riemannian spaces $R_{ijkl} = g_{hl} R^h{}_{ijk}$ is the Riemann tensor).

An exterior differential form ω is called *closed* if $\mathscr{D}\omega = 0$ and *exact* if $\omega = \mathscr{D}\omega'$. Every exact form is closed since $\mathscr{D}(\mathscr{D}\omega) = 0$. Exterior differentiation is similar to the taking of a boundary, except that in the latter case the dimension of a manifold decreases and in the former case the "order" of a form increases. This analogy was made use of in the paper *On analysis situs* [i.e. topology] *of n-dimensional manifolds* (Sur l'analysis situs des variétés à n-dimensions. Paris, 1931) *[452]* by the Belgian mathematician Georges de Rham (1903 1969). In this paper de Rham defined the *cohomology groups*, in a certain sense dual to Betti's *homology groups*, and applied exterior differential forms for the calculation of Betti numbers. The integral of a closed form is said to be homologous to zero, and h integrals of closed forms on a p-dimensional submanifold which are independent in the usual topological sense are said to be homologically independent if no linear combination of these integrals with constant coefficients not all zero is homologous to zero. De Rham has shown that the Betti number b_p is equal to the number of homologically independent integrals of exact differential forms of order p. This method of calculation of Betti numbers was applied by Cartan in his *Topology of compact homogeneous spaces* (La topologie des espaces homogènes clos. Moscow, 1937) *[96, part 1, pp. 1331 -1338]* and by his student Charles Ehresmann in the paper *On the topology of certain homogeneous spaces* (Sur la topologie de certains espaces homogènes. Princeton, 1934) *[164]*.

The Betti numbers are connected with the Gaussian curvature of a surface: if the Gaussian integral curvature $\int K d\sigma$ is taken over a closed compact surface in R_3, then this integral is equal to $4\pi\left(1-\frac{b_1}{2}\right)$, where b_1 is the first Betti number of the surface. This number is equal to double the number of handles of the surface ($b_1 = 0$ for a sphere and $= 2$ for a torus). About this problem see the book of Kentaro Yano and Salomon Bochner *Curvature and Betti numbers* (Princeton, 1953) *[645]*.

Veja:

Anotação 108 (110)

110 (02/02/2020)

Gauss foi o primeiro a testar a curvatura do espaço físico:

179° 53' 59.32"

obs: Não há geodésicas paralelas

A History of non-Euclidean geometry

B. A. Rosenfeld

Pág. 280

Osculating circle at that point.

etc.

$(180 = \pi)$ $179^\circ 59' 59.32''$

$$\begin{array}{r} 180\times \\ 60\times \\ 60 \\ \hline 648000'' \end{array}$$

$0{,}68''$ ↙

$$\begin{array}{c} 648000 - 100 \\ 0{,}68 - x \\ \hline \end{array}$$

$- 1{,}049382716 \; ^{\boxed{-4}} \%$

$\div 101{,}049382716 = 3{,}10896$

$\div 1{,}049382716 = 2{,}99375299$

$\pi - 1{,}049382716\% =$

$(\div 1{,}[illegible]01049382716) = 3{,}10896768$

↳

Correto: $1{,}049382716^{-4}$

$0{,}0001049382716\%$

$\div 1{,}0001\ldots$

$\pi \div 1{,}0001\ldots =$

$\pi = 3{,}14126301487837089713\ldots$

Fundamentals of Mathematics
Behnke

Vetores — Tighten

{ módulo - tamanho — Grow
direção - ângulo
sentido - para onde

Grandezas:
{ Escalares
Vetoriais

Conics and Cubics Bix

20210322
www.digitale-sammlungen.de
bsb-muenchen.de

"Porque a sabedoria serve de defesa, do mesmo modo que o dinheiro!
Eclesiastes 7.12

Quem toma emprestado é escravo de quem empresta.
Provérbios 22.7

(@

"Vocês foram comprados por alto preço;
não se tornem escravos de homens."
1 Coríntios 7:23

"Tudo tem o seu tempo determinado,
e há tempo para todo o propósito
debaixo do céu."
Eclesiastes 3:1

7 Segredos da Prosperidade Judaica:

1) Ser perito Provérbios 22:29
faça, seja o melhor.
2) Ter uma cultura empreendedora
"Ninguém fará por você aquilo
que deve ser feito por você".
3) Ter mentores
"Quando não há conselhos os planos
se dispersam,
mas havendo muitos conselheiros
eles se firmam."
Provérbios 15:22
4) Aceitar a repreensão
"Pobreza e afronta virão ao
que rejeita a instrução, mas o
que guarda a repreensão será
honrado".
Provérbios 13:18

4) cont.

"Melhor é a repreensão franca do que o amor encoberto"
Provérbios 27:5

"Todo o que ama a disciplina ama o conhecimento, mas aquele que odeia a repreensão é tolo"
Provérbios 12:1

"Quem acolhe a disciplina mostra o caminho da vida, mas quem ignora a repreensão desencaminha outros"
Provérbios 10:17

5) Ter um bom nome

"Vale mais ter um bom nome do que muitas riquezas; e o ser estimado é melhor do que a riqueza e o ouro"
Provérbios 22:1

→ Boa reputação

6) Guardar dinheiro
"Há tesouro precioso e azeite na casa do sábio, mas o homem insensato os devora" Provérbios 21:20

7) Ter união:

"Como é bom e agradável quando os irmãos convivem em união!"

Salmos 133:1

(%

20210325

28/03 → Pessach 15° NISSAN.

1/6/1969 – 15° SIVAN, 5729.

TERRA

www	A	domain
✓ webmail2	CNAME	webmail2.terraempresas.com.br
✓ webmail	CNAME	webmail.terraempresas.com.br
ftp	A	198.171.234.15
ecp	CNAME	ecp.terraempresas.com.br
✓ @	NS	ns3.terraempresas.com.br
✓ @	NS	ns2.terraempresas.com.br
@	MX	mx.terraempresas.com.br
@	A	198.171.234.15

The one I love REM

@
A

Cmd
ping

216.239.32.21

Contador: mando da
Lantria

Rua Vasco da Gama, 25

perto Igreja do Brás

Misterious Ways U2

20210326

Quark botton

O modelo padrão não pode explicar a natureza da gravidade.

quark bottom devem decair em números iguais de partículas de elétrons e muons (uma partícula elementar semelhante ao elétrons)

Na realidade, produz mais elétrons do que múons.

Partícula ainda não descoberta → leptoquark

Padrão ouro → nível [illegible] chamado
cinco sigma
(1/3,5 milhões)

Acelerador é três sigma — 1/1000.

—

"A descoberta de uma nova força na natureza é o santo graal da física de partículas".

Att:

X17 → 5ª força da natureza?

Decaimento do elétron ([illegible])

↳ Hélio

As partículas do átomo foram divididas em ângulos de 115 graus, muito mais do que o esperado.

No O átomo decai, o excesso de momento que de energia produzida cria brevemente uma nova partícula desconhecida, que se decompõe rapidamente em pósitron e elétron.

33X mais massa que um elétron

= 17 megaeletronvolts.

Este decaimento também aconteceu com o berílio-8

Berílio

No berílio o ângulo foi 140 graus

Obs: À medida que a energia da luz produzida pelo berílio aumenta quando libera um elétron e um pósitron, o ângulo entre os dois deve diminuir.
Mas não foi o que aconteceu no experimento de 2016.

Pode ser um elo entre o mundo visível e a matéria escura.

Analysis Godement 4 vol.
Topological Methods,
Variational Methods.
Brezis

Pretty Fly the Offspring

⇒ Aniversário Stelvio 28/03

Pag. 67 Fábio Prieto de Souza (Desembargador)
www.fab.mil.br/
Download/arquivos/
medalhas/ALMANAQUE
AGRACIADOS OMA 2020.
pdf

3lib.net

Stupid Girl Garbage

20210401

STANACH

Piazza Santo Stefano (Bologna)

Basilica di San Domenico (Bologna)

—

⟹ Upload Books-2021-05

Ehrhard Behrends - π & Co.

—

Pi und Co (2016)

Pag. 157 (159)   1.3 Falta

=

—

veja

Ver

π und Co (2008) → Pag 149 (148)

Pag. 155 (158)

$\boxed{e\,d\pi \le 2 \le \pi\,d\pi}$

(%)
+,- * *. ! * # B! # G 7 #
! ! G 7 ! A H # 8 7#
+,- * *. ! * I # B! # G 7 #
! ! G % 7#
(@ @) (') '%

lichen Ergebnisse eintreten. Einen Fall, bei dem de Mérés Fehleinschätzung in spektakulärer Weise übertroffen wurde, erwähnt der amerikanische Glücksspielexperte John Scarne in seinem Buch *Complete guide to gambling*[14]: So soll ein Spieler – sich selbst im Vorteil wähnend – 1952 innerhalb von zwölf Stunden insgesamt 49000 Dollar bei Wetten darauf verloren haben, in jeweils 21 Würfen mindestens eine Doppel-Sechs zu erzielen. Tatsächlich ist die Gewinnwahrscheinlichkeit in Höhe von $1 - (35/36)^{21} = 0{,}4466$ deutlich geringer, als man vermutlich schätzen würde.

1.3 Lottotipps – „gleicher als gleich“?

Eine statistische Auswertung der insgesamt 1433 deutschen Lotto-Ausspielungen, die vom Oktober 1955 bis Anfang 1983 erfolgten, ergibt, dass – ohne Berücksichtigung der Zusatzzahlen – bei 76,4% der Ausspielungen mindestens eine der Zahlen von 1 bis 10 gezogen wurde. Getippte Zahlenreihen, die keine der Zahlen 1 bis 10 enthielten, hatten also allein aufgrund dieser Tatsache in 76,4% der Fälle keine Chance, einen Haupttreffer mit „Sechs Richtigen“ zu erzielen. Sollte man deshalb immer mindestens eine der Zahlen 1 bis 10 in seinem Lotto-Tipp berücksichtigen?

Das Zahlenlotto, das in Deutschland und in einigen anderen Ländern in der Form „6 aus 49“ gespielt wird, gehört heute zu den populärsten Glücksspielen. Und das nicht nur beim Publikum, sondern auch beim Staat, dessen Gewinn in der ungefähren Höhe des halben Einsatzes schon vor der Ziehung sicher ist. Entstanden ist das Lotto übrigens im 16. Jahrhundert in der Stadt Genua, wo damals jährlich fünf Senatoren per Los bestimmt wurden. Gleichzeitig konnte auf die zur Auswahl stehenden 110 Namen gewettet werden. Mit der Zeit verselbstständigte sich das Spiel und wurde dabei abstrahiert. Statt auf Namen setzte man nun auf Zahlen. Dem Siegeszug des Lottos konnten sich selbst die Regierungen des ehemaligen Ostblocks nicht entziehen[15]. Auch dort wurde das ursprünglich als kapitalistisch gebrandmarkte Spiel veranstaltet.

Aufgrund seiner Beliebtheit wurde das Lotto auch zum Gegenstand vieler Publikationen. In einem Lotto-Buch[16] wird die in der Fragestellung zitierte Auswertung wie folgt kommentiert:

> So gesehen muß man feststellen, daß Lotto eigentlich unlogisch ist. Wenn man darüber nachdenkt, ist es ganz einfach. Es haben nicht alle Zahlen beziehungsweise alle möglichen „Anfangszahlen“[17], die von 1 bis 44, die gleichen Chancen.
> Weil das so ist, haben nicht alle Tippreihen im Lottospiel die gleichen Chancen.
> ... wer Lotto spielt und dabei Reihen zusammenstellt, die mit einer Anfangszahl 11 und höher beginnen, verschenkt mehr als drei Viertel der Chancen, einen Sechser zu treffen. Selbst dann, wenn das Glück ihm eigentlich hold wäre. Er kann den Sechser deshalb nur

---

[14] John Scarne, *Complete guide to gambling*, New York 1974, S. 16.

[15] Die Entwicklung des Lottos in der DDR wird beschrieben in Wolfgang Paul, *Erspieltes Glück – 500 Jahre Geschichte der Lotterien und des Lotto*, Berlin 1978, S. 190-192.

[16] Rolf B. Alexander, *Das Taschenbuch vom Lotto*, München 1983; Zitate: S. 26, S. 68 f.

[17] Mit „Anfangszahl“ ist die kleinste Zahl einer getippten Sechser-Reihe gemeint.

© Springer-Verlag Berlin Heidelberg 2016
E. Behrends, G.M. Ziegler, P. Gritzmann (Hrsg.), *Pi und Co.*, DOI 10.1007/978-3-662-48872-0_22

> in einem knappen Viertel aller Ausspielungen treffen, weil seine Spielreihe nach der Formel „6 aus 49“ schlicht unvollständig ist. Der Spieler, der hohe Anfangszahlen für seine Reihe wagt, gleicht dem Lotteriespieler, der mit einem Viertellos die auf ein ganzes Los entfallende Million gewinnen will. Er kann sie einfach nicht bekommen.

Fast ist man geneigt, den Argumenten Glauben zu schenken und seine Tippreihen immer mit einer der Zahlen 1 bis 10 anfangen zu lassen. Andererseits ist aber jede Zahl und damit auch jede Lottoreihe theoretisch „gleichmöglich“, wie Laplace formulierte. Und warum sollten gerade die Zahlen von 1 bis 10 und nicht andere Zehner-Gruppen wie

- 34 bis 43 oder
- 4, 9, 14, 19, 24, 29, 34, 39, 44 und 49 oder
- 11, 16, 17, 22, 23, 25, 29, 32, 36 und 48

eine besondere Rolle spielen? Alles gut und schön, aber vielleicht doch nur graue Theorie? Schließlich kann das Ergebnis der statistischen Untersuchung doch nicht einfach ignoriert werden! Aber ist es wirklich so außergewöhnlich, wie es scheint? Und kann das statistische Ergebnis wirklich als Argument für die gegebene Empfehlung dienen?

Vergessen wir für einen Moment, dass die statistische Auswertung bereits vorliegt. Welches Ergebnis würden wir dann ungefähr erwarten? Das heißt, wie groß ist die Wahrscheinlichkeit und damit ungefähr die relative Häufigkeit dafür, dass eine Lottoausspielung mindestens eine der Lottozahlen 1 bis 10 enthält? Für eine Antwort könnte man einen Computer so programmieren, dass er alle Möglichkeiten und darunter die für das Ereignis „günstigen“ durchzählt. Dass es auch einfacher geht, verdanken wir einigen Formeln der **Kombinatorik**, einer mathematischen Teildisziplin, die sich mit den Möglichkeiten befasst, Dinge miteinander zu kombinieren oder anzuordnen. Der denkbar einfachste Fall betrifft die völlig freie Kombination von Merkmalen, etwa den Ergebnissen von zwei Würfeln: Jedes Ergebnis des einen Würfels kann mit jedem Ergebnis des anderen zusammentreffen, so dass es 6·6 = 36 Kombinationen gibt, wobei Würfelkombinationen wie 2-6 und 6-2 unterschieden sind.

Etwas komplizierter wird es, wenn Karten gemischt werden. Wie viele Möglichkeiten gibt es dafür, eine vorgegebene Anzahl von unterschiedlichen Karten anzuordnen? Handelt es sich nur um drei Karten, hier einfach 1, 2 und 3 genannt, dann sind die folgenden Reihenfolgen möglich:

1 2 3   1 3 2   2 3 1   2 1 3   3 1 2   3 2 1

Drei Karten lassen sich also auf 6 Arten sortieren, es gibt also 6 so genannte **Permutationen**. Bei 4 Karten gibt es bereits 24 und bei 5 Karten schon 120 Permutationen. Um diese Zahlen zu finden, müssen keineswegs alle Permutationen aufgelistet werden. So gibt es bei 5 Karten 5 Möglichkeiten für die erste Karte. Ist die erste Karte gewählt, kann die zweite Karte aus dem Rest von 4 Karten gewählt werden. Für die dritte Karte gibt es dann 3 und für die vierte Karte noch 2 Möglichkeiten. Am Schluss muss schließlich die einzig noch verbliebene Karte genommen werden. Die Zahl der Permutationen von 5 Karten oder anderen unterscheidbaren Dingen ist also gleich $5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$.

Die Anzahl von Permutationen hat eine so große Bedeutung, dass sie als eigenständige mathematische Operation interpretiert wird. Diese so genannte **Fakultät** wird mit einem Ausrufungszeichen „!“ abgekürzt: n!, gesprochen „n Fakultät“, steht für die Anzahl von Permutationen, die mit n unterschiedlichen Dingen gebildet werden können. Wie im Fall $n = 5$ berechnet man n Fakultät allgemein mit der Formel

$$n! = n \cdot (n-1) \cdot (n-2) \ldots 4 \cdot 3 \cdot 2 \cdot 1,$$

was für die Zahlen n = 1, 2, 3, 4, 5, 6 die Werte

$$1! = 1,\ 2! = 2,\ 3! = 6,\ 4! = 24,\ 5! = 120,\ 6! = 720$$

ergibt. Als praktisch hat sich die Festlegung des Wertes 1 für die Fakultät der Null erwiesen, das heißt

$$0! = 1.$$

Die 32 Karten eines Skatspiels können auf 32! = 32·31·... 4·3·2·1 verschiedene Arten gemischt werden, das ist eine immerhin 35-stellige Zahl, die selbst die vermutlich seit dem Urknall des Universums vergangenen Sekunden, eine 18-stellige Zahl, weit übertrifft:

$$32! = 263\ 130\ 836\ 933\ 693\ 530\ 167\ 218\ 012\ 160\ 000\ 000$$

So astronomisch diese Zahl ist, im Vergleich zu den 52! Permutationen eines Rommé- oder Poker-Spiels ist sie verschwindend klein: 52! ist nämlich eine 67-stellige Zahl – etwa so viel, wie es nach einer Schätzung Atome im gesamten Universum geben soll[18].

Beim Lotto werden, sieht man von der Zusatzzahl erst einmal ab, 6 von 49 Zahlen gezogen. Vergleichbare Auswahlen sind auch bei anderen Spielen üblich: Beim Pokern erhält ein Spieler 5 der 52 Karten, ein Skatspieler erhält 10 von 32 Karten. Allen Situationen gemeinsam ist, dass aus einer Gesamtzahl von unterscheidbaren Dingen eine festgelegte Zahl von Dingen zufällig ausgewählt wird. Auf die Reihenfolge, in der diese Dinge ausgewählt werden, kommt es dabei nicht an. Man spricht in solchen Fällen von **Variationen**.

Die Anzahl von möglichen Variationen kann ähnlich wie die von Permutationen bestimmt werden: Für die erste Kugel bei einer Lottoausspielung gibt es 49 Möglichkeiten. Wird die zweite Zahl gezogen, gibt es für diese noch 48 Möglichkeiten. Folglich können sich die ersten beiden Zahlen auf 49·48 verschiedene Weisen miteinander kombinieren. Bei den anschließend gezogenen Zahlen verringert sich die Anzahl der Möglichkeiten jeweils um 1. Daher existieren insgesamt 49·48·47·46·45·44 Sequenzen von 6 Lottozahlen, wobei sich einige Sequenzen nur in der Reihenfolge, nicht aber in den ausgewählten Zahlen unterscheiden. Dieser Sachverhalt lässt sich noch präzisieren: Jede Auswahl von 6 Lottozahlen kann als Permutation in insgesamt 6! = 720 Ziehungssequenzen auftreten, das heißt, die Zahl der Variationen beträgt

$$\frac{49\cdot 48\cdot 47\cdot 46\cdot 45\cdot 44}{6!} = 13983816.$$

Damit gibt es knapp 14 Millionen mögliche Tippreihen beim Lotto. Folglich beträgt die Wahrscheinlichkeit, „Sechs Richtige“ zu tippen, etwa 1 durch 14 Millionen. Dass trotz der geringen Wahrscheinlichkeit fast jede Woche Glückliche zu verzeichnen sind, liegt einzig an der riesigen Zahl von abgegebenen Lottotipps, welche die Anzahl der Mitspieler aufgrund mehrfacher Tipps sogar noch übersteigt.

---

[18] Wie stark Fakultäten wachsen, kann man gut anhand der so genannten Stirling'schen Formel sehen, mit der Fakultäten näherungsweise berechnet werden können. Die Stirling'sche Formel lautet

$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n},$$

wobei der relative Fehler für größere Werte n sehr klein ausfällt. Beschreiben lässt sich die Qualität der Approximation dadurch, dass der Quotient aus n! und der Stirling-Näherung zwischen $e^{1/(12n+1)}$ und $e^{1/12n}$ liegt. Beispielsweise für n=32 erhält man die Näherung $2{,}6245\cdot 10^{35}$; das ist nur 0,26% zu wenig.

Die allgemeine Formel für die Anzahl von Variationen ergibt sich ganz analog zum Lotto-Beispiel: Werden aus n unterschiedlichen Dingen k Dinge ausgewählt, dann gibt es insgesamt

$$\frac{n(n-1)(n-2)\ldots(n-k+1)}{k\,!}$$

verschiedene Möglichkeiten einer Auswahl. Der Bruch, der sich immer zu einer ganzen Zahl kürzen lässt, wird **Binomialkoeffizient** genannt. In seiner Schreibweise

$$\binom{n}{k}$$

spricht man auch von „n über k“. Die Zahl der möglichen Lottotipps beträgt

$$\binom{49}{6},$$

was genau den schon berechneten Wert ergibt.

Das Pascal’sche Dreieck

Die Gesamtheit der Binomialkoeffizienten lässt sich übersichtlich in einem Schema anordnen, das Pascal’sches Dreieck genannt wird:

$$\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1 \\
\ldots & & & & & \ldots & & & & & \ldots
\end{array}$$

Dabei steht der Binomialkoeffizient $\binom{n}{k}$ in der (n+1)-Zeile an (k+1)-ter Stelle, beispielsweise findet man $\binom{4}{2} = 6$ als dritten Wert in der fünften Zeile. Der Clou des Pascal’schen Dreiecks besteht darin, dass alle Werte ohne jede Multiplikation berechnet werden können, da jede Zahl gleich der Summe der beiden über ihr stehenden Zahlen ist. Warum das funktioniert, lässt sich schnell erklären: Um von n Karten k auszuwählen, nimmt man entweder die erste Karte und wählt die anderen k-1 aus den verbleibenden n-1 Karten aus oder man nimmt die erste Karte nicht und nimmt alle k Karten aus dem Rest. Das entspricht der Gleichung

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

Mit Binomialkoeffizienten ist es nun ein Leichtes, Wahrscheinlichkeiten beim Lotto zu berechnen. So gibt es unter den insgesamt knapp 14 Millionen möglichen Tipreihen nur $\binom{39}{6} = 3262623$, die sich bloß aus den 39 Zahlen von 11 bis 49 zusammensetzen. Das heißt, die Wahrscheinlichkeit, dass alle sechs ausgespielten Zahlen größer oder gleich 11 sind, beträgt 0,2333. Nach dem Gesetz der großen Zahlen ist es also auf Dauer zu erwarten, dass der Anteil der Ziehungen, die mindestens eine Zahl von 1 bis 10 enthalten, etwa gleich

76,67% beträgt. Das Ergebnis der statistischen Untersuchung von 76,4% ist damit alles andere als ungewöhnlich.

Wenn das statistische Ergebnis schon nicht ungewöhnlich ist, wie sieht es dann mit der ausgesprochenen Empfehlung aus, mindestens eine der Zahlen 1 bis 10 bei einem Tipp zu berücksichtigen? Man kann sie getrost vergessen, da sie auf einem Fehlschluss beruht! Die Aussage, dass man bei einem Tipp ohne die Zahlen 1 bis 10 mit fast 77-prozentiger Wahrscheinlichkeit unter anderem deshalb keinen Sechser erzielen wird, weil mindestens eine der Zahlen 1 bis 10 gezogen wird, ist schlicht uninteressant. Damit wird nämlich nur ausgesagt, dass bei Tipps, welche die Empfehlung nicht berücksichtigen, die Wahrscheinlichkeit auf „Sechs Richtige” kleiner als 0,2333 ist. Aber das ist ohnehin klar, denn die Wahrscheinlichkeit auf einen Sechser ist noch viel kleiner, nämlich gleich 0,0000000715.

Nicht überzeugt? Stellen wir uns vor, wir hätten die Zahlen 22, 25, 29, 31, 32, 38 getippt. Da wir die Ziehung im Fernsehen nicht selbst verfolgen können, bitten wir einen Bekannten, die Lottozahlen für uns aufzuschreiben. Immer noch nicht überzeugt, ob es richtig war, keine Zahl bis 10 zu tippen, fragen wir unseren Bekannten zunächst: „Ist eine Zahl von 1 bis 10 dabei?“. In knapp 77% der Fälle ist der Traum von den Millionen mit einem „Ja“ als Antwort bereits beendet. So weit so gut. In den anderen knapp 23% – und hier irrte der zitierte Buchautor – bleibt die Gewinnhoffnung aber nicht gleich, sondern sie steigt. Schließlich haben wir ja nur noch 6 aus 39 Zahlen richtig zu tippen.

Nachdem wir uns davon überzeugen konnten, dass der aus dem Buch zitierte Ratschlag unbegründet ist, bleibt noch die Frage danach, was einen Spieler erwartet, der der Empfehlung trotzdem folgt. Zunächst ist natürlich zu bemerken, dass die entsprechenden Zahlenkombinationen zwar nicht „besser” sind, wie der Autor vermutete, aber ebenso wenig sind sie „schlechter”, das heißt unwahrscheinlicher, als die anderen. Insofern ist die Empfehlung im Hinblick auf die Gewinnwahrscheinlichkeit keineswegs schädlich. Berücksichtigt man aber, dass die Gewinnhöhen beim Lotto stets davon abhängen, wie viele Mitspieler eine Gewinnklasse treffen, dann ändert sich das Bild[19]. Jede Zahlenkombination, die häufig getippte Zahlen und Zahlenkombinationen beinhaltet, ist danach weniger empfehlenswert, da sie im Gewinnfall zu vergleichsweise niedrigen Gewinnhöhen führt. Beispielsweise ergeben, da viele Lotto-Spieler ihren Tipp aus Datumsangaben ableiten, Ziehungen mit der Zahl 19 vergleichsweise niedrige Quoten. Entsprechendes gilt für Ziehungen, in denen die Zahlen von 1 bis 12 beziehungsweise 1 bis 31 besonders oft vorkommen. Andere Vorlieben erklären sich, wie bei der Glückszahl 7, eher aus der mit der Zahl verbundenen Symbolik. Und auch die geometrische Verteilung der Zahlen auf dem Spielschein hat sicher einen Einfluss.

---

[19] Die Gewinnquoten differieren zum Teil erheblich. So hat es bei der Gewinnklasse „Sechs Richtige” bisher zweimal besonders niedrige Quoten gegeben. Bei der Ziehung am 18.6.1977 waren es 205 vermeintliche „Glückspilze”, die alle die sechs gezogenen Zahlen, nämlich 9, 17, 18, 20, 29 und 40 richtig getippt hatten. Die Gewinnquote betrug aber nicht die ersehnte Million, sondern nur vergleichsweise klägliche 30737,80 DM. Was war passiert? Viele Tipper, vor allem im nordwestdeutschen Raum, hatten es sich angewöhnt, diejenigen Zahlen zu tippen, die in Holland eine Woche zuvor gezogen worden waren. Dies war, wie sich dann zeigte, ein schwerer Fehler, natürlich nicht deshalb, weil eine direkte Wiederholung unwahrscheinlicher war als eine andere Zahlenreihe, sondern einfach deshalb, weil zu viele Mitspieler die gleiche Idee hatten. Bei einer anderen Ziehung, nämlich am 23.1.1988, ergaben sich sogar 222 Volltreffer. Dafür verantwortlich war wohl die Regelmäßigkeit der gezogenen Zahlenreihe: 24, 25, 26, 30, 31 und 32.

Die Gewinnklassen im Lotto

Auch die Chance, beim Lotto eine bestimmte Gewinnklasse zu erzielen, kann mit Hilfe von Binomialkoeffizienten leicht berechnet werden. Beispielsweise erreicht man genau dann 4 Richtige, wenn

- 4 der 6 getippten Zahlen und
- 2 der 43 nicht getippten Zahlen

gezogen werden. Dazu gibt es, kombiniert man die Möglichkeiten für die richtigen wie die falsch getippten Zahlen miteinander, insgesamt genau

$$\binom{6}{4}\cdot\binom{43}{2} = 15\cdot 903 = 13545$$

Möglichkeiten. Das ergibt eine Wahrscheinlichkeit von 0,00097, ungefähr entsprechend 1/1032. Folglich bringt noch nicht einmal jeder tausendste Tipp 4 Richtige!

Gewinnklasse	Anzahl der Kombinationen	Wahrscheinlichkeit
6 Richtige	1	1/14 Mill.
5 Richtige mit Zusatzzahl	6	1/2,3 Mill.
5 Richtige	252	1/55491
4 Richtige	13545	1/1032
3 Richtige mit Zusatzzahl	17220	1/812
3 Richtige	229600	1/61
Verlust (Rest)	13723192	0,981
	13983816	

Aufgrund der zusätzlich ausgespielten „Superzahl“ unterteilt sich die höchste Gewinnklasse im Verhältnis von 9:1, so dass die Wahrscheinlichkeit für die höchste Gewinnklasse nur noch 1 zu 140 Millionen beträgt. Trotz des gestiegenen Umsatzes, nicht zuletzt auch aufgrund der deutschen Wiedervereinigung, wird so die höchste Gewinnklasse oft über mehrere Wochen nicht erreicht. Die dann nicht ausgespielten Gewinne bleiben in Form eines Jackpots für die nächste Ziehung stehen. 1994 erreichte dieser Jackpot immerhin einen Spitzenwert von 42 Millionen DM.

Hinweise darauf, welche Zahlen besonders oft getippt werden, lassen sich indirekt dadurch erhalten, dass man die wöchentlichen Gewinnquoten darauf untersucht, bei welchen gezogenen Zahlen sie höher und bei welchen sie niedriger ausfallen. Allerdings sind so aufgrund der vielschichtigen Einflüsse, etwa durch mehr oder weniger beliebte Teilkombinationen und einem nicht bei jeder Ziehung völlig gleichem Verhalten, nur eingeschränkte Aussagen zu erzielen[20]. Weit aussagekräftiger ist eine Auswertung, in der bei einer Ziehung des Jahres 1993 alle in Baden-Württemberg abgegebene Tipps einbezogen wurden[21]. Von diesen Tipps haben 80,7% eine Anfangszahl zwischen 1 und 10. Da davon ausgegangen werden kann, dass es sich bei diesem Wert nicht nur um eine Momentaufnahme handelt, sind die Kombinationen mit einer Anfangszahl von höchstens 10 weniger empfehlenswert. Sicher würde man dem zitierten Buch zu viel Ehre antun, wenn man das Verhalten auf seine Empfehlung zurückführen würde. Ursache ist sicher eher die schon genannte Vorliebe für Datumsangaben.

---

[20] Heinz Klaus Strick, *Zur Beliebtheit von Lottozahlen*, Praxis der Mathematik, 33 (1991), Heft 1, S. 15-22; Klaus Lange, *Zahlenlotto*, Ravensburg 1980, S. 61-110.

[21] Karl Bosch, *Lotto und andere Zufälle*, Braunschweig 1994, S. 201 ff.; Karl Bosch, *Glücksspiele: Chancen und Risiken*, München 2000, S. 57-70. Die Auswertung erstreckte sich über knapp 7 Millionen Tippreihen, so dass bei einer zufälligen Verteilung der Tipps jede der knapp 14 Millionen

Pokern

Beim Pokern erhält ein Mitspieler fünf der insgesamt 52 Karten, dafür gibt es

$$\binom{52}{5} = \frac{52\cdot 51\cdot 50\cdot 49\cdot 48}{1\cdot 2\cdot 3\cdot 4\cdot 5} = 2598960$$

Möglichkeiten. Will man unter diesen knapp 2,6 Millionen Kartenkombinationen beispielsweise die Anzahl der „Doppel-Zwillinge“ bestimmen, geht man am besten wie folgt vor: Ein Doppel-Zwilling besteht aus 5 Karten mit insgesamt 3 unterschiedlichen Kartenwerten, von denen 2 zweimal vorkommen. Ein Beispiel ist

- Herz-*Vier*, Kreuz-*Vier*, Herz-*Bube*, Pik-*Bube*, Pik-Dame.

Eindeutig bestimmt wird ein Doppel-Zwilling durch die folgenden Angaben:

- Die beiden Kartenwerte der Zwillinge (im Beispiel Vier, Bube),
- der einmal vorkommende Wert (Dame),
- die beiden Farben des niedrigwertigen Zwillings (Herz, Kreuz),
- die beiden Farben des höherwertigen Zwillings (Herz, Pik),
- die Spielfarbe des einmal vorkommenden Wertes (Pik).

Die Zahl der möglichen Kombinationen erhält man nun, wenn man die Anzahlen, die sich für diese Merkmale ergeben, miteinander multipliziert. Zu berücksichtigen dabei ist, dass nicht alle Merkmale frei miteinander kombinierbar sind. Im Einzelnen gibt es

- zunächst $\binom{13}{2} = 78$ Möglichkeiten für die Kartenwerte der beiden Zwillinge,
- dazu jeweils 11 Möglichkeiten für den einmal vorkommenden Wert,
- für die beiden Farben des ersten Zwillings $\binom{4}{2} = 6$ Möglichkeiten,
- für die Farben des zweiten Zwillings ebenfalls 6 Möglichkeiten und
- 4 Möglichkeiten für die Farbe des einmal vorkommenden Wertes.

Zahlenkombinationen etwa 0,5-mal zu erwarten ist. Allerdings gibt es bei den Tippreihen stark ausgeprägte Favoriten, von denen 24 mehr als tausend mal getippt wurden. Die häufigste Tippreihe, bestehend aus den Zahlen 7, 13, 19, 25, 31 und 37, wurde sogar 4004-mal getippt! Weniger die Zahlen selbst – bis auf die 25 allesamt Primzahlen – dürften für diese Vorliebe verantwortlich sein. Vielmehr bilden diese Zahlen auf dem Spielschein eine fast vollständige Diagonale, die oben rechts beginnt. Hochgerechnet auf ganz Deutschland dürfte diese Tippreihe bei jeder Ziehung über 30000-mal getippt werden. Kaum einer der betreffenden Spieler wird wohl ahnen, wie gering seine Gewinnaussichten sind.

Ähnliche Ergebnisse in anderen Ländern wurden von H. Riedwyl gefunden. Siehe: Hans Riedwyl, *Zahlenlotto – Wie man mehr gewinnt*, Bern 1990; Norbert Henze, Hans Riedwyl, *How to win more*, Natick 1998; Hans Riedwyl, *Gewinnen im Zahlenlotto*, Spektrum der Wissenschaft, 2002/3, S. 114-119.

Insgesamt existieren daher 78·11·6·6·4 = 123552 Doppel-Zwillinge. Werden aus einem gut gemischten Kartenblatt fünf Karten zufällig ausgeteilt, dann ist die Wahrscheinlichkeit, einen Doppel-Zwilling zu erhalten, gleich

$$123552/2598960 = 0{,}04754,$$

das heißt, etwa jedes 21. zufällig gezogene Blatt ergibt einen Doppel-Zwilling.

Die folgende Tabelle umfasst in der Reihenfolge der Wertigkeit die Anzahlen aller Poker-Kombinationen. Mit berücksichtigt ist auch die Würfelvariante, die mit fünf Würfeln gespielt wird, welche mit den Symbolen (Pik-)Neun, (Karo-)Zehn, Bube, Dame, König und (Kreuz-)Ass gekennzeichnet sind:

Poker-Kombination	**Anzahl bei ...**	
	5 Karten	**5 Würfeln**
Fünfling		6
Royal Flush (Zehn bis Ass einer Farbe)	4	
Straight Flush	36	
Vierling	624	150
Full House (Drilling und Zwilling)	3744	300
Flush (eine Farbe)	5108	
Straight (Straße: Werte in Reihenfolge)	10200	240
Drilling	54912	1200
Doppel-Zwilling	123552	1800
Zwilling	1098240	3600
Rest	1302540	480
	2598960	7776

Weiterführende Literatur zum Thema Lotto:

Norbert Henze, *2000mal Lotto am Samstag – gibt es Kuriositäten?*, Jahrbuch Überblicke der Mathematik, 1995, S. 7-25.

Glück im Spiel, Bild am Sonntag Buch, Hamburg, ca. 1987, S. 6-29.

Ralf Lisch, *Spielend gewinnen? – Chancen im Vergleich*, Berlin 1983, S. 38-54.

Günter G. Bauer (Hrsg.), *Lotto und Lotterie*, Homo Ludens – der spielende Mensch, Internationale Beiträge des Institutes für Spielforschung und Spielpädagogik an der Hochschule „Mozarteum" Salzburg, 7 (1997), München 1997.

( )"

Steiner Chain

Wikipedia

Ask Ethan: Where is the line between Mathematics and Physics?
https://br.pinterest.com/pin/340373684332567737/

=

https://www.symmetrymagazine.org/article/the-deconstructed-standard-model-equation

=

Curva e integral en Wα
https://br.pinterest.com/pin/709809591255709470/

= →

$$\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi} \rightarrow \pi$$

e y pi

relación normal

( )%

SPID
Sistema Pubblico di
Identità Digitale

reaquisição da cidadania

Art 13 Lei N.º 91
5 Fevereiro de 1992
(comma 1, letra ⓒ)

Declaração ao Consulado

Remessa ao Exterior

Banco Destinatário: Poste Italiane S.p.A
Nome do Beneficiário: Ministero
dell'Interno
D.L.C.I -
Cittadinanza
Finalidade: Legge n. 94/2009,
contributo Euro 250,00
IBAN: IT54D0 760103200000000809020

Bic/SWIFT: BPPIITRRXXX

(%

20210414 (20210414)

Hold that Rock

216

— Conjectura de Poincaré

Objeto homeomorfismo

"geometria flexível"
→ duas formas são a mesma se for possível transformar uma em outra sem quebrá-la.

"A monetização do êxito é o máximo insulto à matemática" Perelman G.

—

"Se a teoria está correta, não necessita de outro tipo de reconhecimento"
Perelman

"Não me interessa o dinheiro ou fama.
Não quero estar em exibição como um animal em um zoológico"
Perelman

═

( )

20210415

Russia 110.000 tropas (56 battalion)

Betty Faria e Dennis Carvalho (diretor)

↳ queimou Maria Gomes

porque ele saia com ela.

Medalhas da FAB para
desembargador TRF-3
(Presidente)
Fabio Prieto de Souza
(2014-2016)

1) www.fab.mil.br/Download/arquivos/medalhas/ALMANAQUEAGRACIADOSOMA 2020.pdf

Pag. 67 → Ordem do Mérito Aeronáutico

2) www.fab.mil.br/cabine/anexos/Almanaque_Agraciados_MSD_2019.pdf

Pag. 123 → Medalha Mérito Santos Dumont.

Rua Itamarataca, 77

Tio João + Nono Domenico
moraram na mesma
casa.

—

Contabilidade Dela Torre
R Vasco da Gama

—

→ vender p/ o marido da Lavínia.

—

Down – History of War - Issue 93, 2021

—

Google Workspace Business Enterprise
Price

—

Damaso Perez Prado – Rei Mambo

—

RonWyatt.com

—

(@

mathematics.test R$ 70
.technology R$ 70

.page R$ 120

---

upload Basics-002

Eva Wilma Buckup

C:\Data-NovoVivo-009-2021\
Books-009-2021\
20210419\Geometry the Line
and the Circle\
Plane Algebraic Curves\
Pencil of cubics...pdf

Pag. 32(63)

Fig 3.1

The eleven conic-walls
A, ... K

(%

20210422

fabricante cimento franco-suiço
Lafarge / Holcim
1400 funcionários - out.
maior fabricante de cimento no mundo.
90.000 funcion. mundo.

Estrelas fim da vida são sustentadas pelo Princ. de degenerescência.
→ Estrelas de Nêutrons.

—

"A vida é tudo o que você faz com ela!"

"Estude o que te emociona de forma quase íntima."

"Uma conquista só tem valor se ela te custar esforço!!" Janot

—

ROMA 21 Abril 753 AC.

Hoje: 2774 anos
(ontem)

= 970 A.C – 586 A.C

1ª destruição Templo de Jerusalém
586 A.C. pelos exércitos da Babilônia
Rei Nabucodonosor II

2ª destruição: 6 Ago 70
(575 AC – 70 DC) 8 set 70 (A12.com)

( )

2º Templo → reconstruído pelo Decreto
Ciro II, da Pérsia
(Livro de Esdras Cap 1, 1-4)

—

nó górdio → Górdio, rei lendário
da Frígia, antiga região
da Ásia Menor (Turquia)

—

Gordian Knot is a legend of Phrygian Gordium associated with Alexander the Great.
→ metaphor for an intractable problem solved easily by finding an approach to the problem that renders the perceived constraints of the problem moot
("cutting the Gordian Knot").

—

escárnio → zombaria

sarcasmo → zombaria nos opostos

ironia.

—

(@

20210428 1 US$ 5,40

R$ 20 000 000 $\div$ 5.40 =

$3703703,\overline{703}$

↳ ano de nascimento do meu pai.

$30.000.000 \div 5,40 =$

$5555555,\overline{5}$

Princípio de simetria da mecânica quântica.

"Nós somos o instrumento que o Universo tem para conhecer a si mesmo" Timote.

$69 \div 11 = 6,\overline{27}$

Dízima $0,\overline{63}$ $= 100x = 63,\overline{63}$

$x = 0,\overline{63}$

$99x = 63$

$x = \frac{63}{99}$

$0,\overline{69} = 100x = 69,69$

$x = 0,69$

$99x = 69$

$x = \frac{69}{99}$ ///

off: ~~$1969 \to 19\frac{69}{99}$~~ melhor $1969 = 19\frac{207}{3}$

—

$69 = \frac{207}{3} = \frac{138}{2}$

—

$0,\overline{5} = 100x = 5$[illegible]

$x = 0,5$ [illegible]

$99x = 49,5$

$\Downarrow$

$50 - 0,5 = 49,5$

$\frac{49,5}{99} = 0,\overline{5}$ ////

$45/99 = 0,\overline{45}$

$0,\overline{5} = 100x = 55,5$

$x = 0,5$

$99x = 55$

$x = \frac{55}{99}$

—

wolframphysics.org/
technical-introduction/

20210429m

Declaração de Balfour
02 Nov 1917

o governo britânico prometia ao povo judeu um "lar" na região da Palestina

---

Tratado de Sèvres
Agosto 1920

os otomanos cederam oficialmente as áreas acordadas, aceitando a criação dos mandatos britânicos e francês.

Palestina baseada na Bíblia

Lloyd George utilizou o Atlas da Geografia Histórica da Terra Santa, publicado em 1915 pelo reverendo escocês George Adam Smith, para definir os limites da região.

20210502

Data-NowVivo-010-2021\
20210502\
Trends in Differential Geometry,
Complex Analysis and
Mathematical Physics\
Trends in Differential Geometry
....pdf

Pág. 109 (126)

Some Geometric Properties
and Objects Related to
Bézier Curves
(Hristov)

Barycentric coordinates and
barycentric trajectories

→ real projective plane
$\mathbb{RP}_2$.

Pág. 111 (128)

1.2 Barycentric characterization
of the duality in the projective plane.

Projective isotomic conjugation

Import.

(""
+,- * *. ! * 5 # 6 ! 2 3/
7 J + ! 3 ! # ! ! ! 3 #) & ! ! 1 G 7
7 J / & 7#
( )

SOME GEOMETRIC PROPERTIES AND OBJECTS RELATED TO BÉZIER CURVES

M. J. HRISTOV

"St. Cyril and St. Methodius" University of Veliko Tarnovo,
Faculty of Education, Department of Algebra and Geometry
E-mail: m.hristov@uni-vt.bg

Classical Bézier curves (the parabolic case) with the natural barycentric description are considered. The explicit expressions for lenght, curvature, shape etc. are given. It is introduced the notion trajectory of Bézier curve and by means of the projective isotomic conjugation the corresponding envelope conic surfaces in the space are obtained.

Keywords: Barycentric coordinates and orbits; Projective isotomic conjugation; Bézier curves; Curvature; Shape; Envelope.

1. Preliminaries

1.1. *Barycentric coordinates and barycentric trajectories*

Let α be fixed euclidean plane, $\triangle ABC$ be fixed triangle in λ and $\mathbb{E}_2^*(\alpha)$ be the extended euclidean plane — the basic model of the real projective plane $\mathbb{RP}_2$. The set of points (resp. lines) in $\mathbb{E}_2^*(\alpha)$ will be denoted by Σ (resp. Λ). For $G \in \Sigma$, Λ_G will stand for the set of lines, incident (z) with G (known as star of lines with center G), for $g \in \Lambda$, Σ_g will stand for the set of points z g and U_g denotes the infinity (unproper) point of g. The infinity line in α is the set of all infinity points and will be denoted by ω.

For any point $P \in \Sigma$ and $p \in \mathbb{R}$ there exist unique $x, y, z \in \mathbb{R}$, so that

$$\begin{cases} p \cdot \overrightarrow{OP} = x \cdot \overrightarrow{OA} + y \cdot \overrightarrow{OB} + z \cdot \overrightarrow{OC} \\ x + y + z = p. \end{cases}$$

The ordered triple (x, y, z) is said to be triple of barycentric coordinates (b-coordinates) [4–6] of P with respect to basic Möbius triangle ABC and it will be written $P(x, y, z)$. The case $p = 0$ corresponds to $P \in \omega$ and $p \neq 0$ describes finite euclidean points. From projective point of view, b-coordinates of P are the affine coordinates with respect to space coordinate system

$$\mathcal{K} = \{O(\not\in\alpha), a_1(\|\overrightarrow{OA}), a_2(\|\overrightarrow{OB}), a_3(\|\overrightarrow{OC})\},$$

of a vector $p\|\overrightarrow{OP}$ and $\alpha : x + y + z = p$. That is why we do not use the terminology "homogeneous relative (absolute) b-coordinates" [4–6].

From algebraic point of view b-coordinates x, y, z of finite $P \in \Sigma$ are the unique solution of the system of linear equations (SLE)

$$\left\{\begin{array}{llll} & \lambda y & + z & = 0 \\ \mu x & + (1-\lambda)y & & = 0\ , \\ (1-\mu)x & & & = p \end{array}\right. \tag{1}$$

where $\lambda = (BCA') \overset{\mathrm{def}}{=} \dfrac{\overline{BA'}}{\overline{CA'}}$ and $\mu = (AA'P)$ are the affine ratios, always $\neq 1$ for each finite points P and $A' = PA \cap BC$ (it is assumed) and $\lambda = 1 \iff A' = U_{BC}$. We rewrite (1) in the matrix form $A.X_1 = B$, where

$$A = \begin{pmatrix} 1 & \lambda & 0 \\ 0 & 1-\lambda & \mu \\ 0 & 0 & 1-\mu \end{pmatrix}, \qquad B = \begin{pmatrix} 0 \\ 0 \\ p \end{pmatrix}, \qquad X_1 = \begin{pmatrix} z_1 \\ y_1 \\ x_1 \end{pmatrix}.$$

Thus, in transposed form,

$$X_1^T = B^T (A^{-1})^T = \left(\frac{p\lambda\mu}{(1-\lambda)(1-\mu)}; -\frac{p\mu}{(1-\lambda)(1-\mu)}; \frac{p}{1-\mu} \right).$$

The triangle matrix A is of stochastic type by columns (left(L)-stochastic-like) and generates the affine transfomation φ_A in $\mathbb{E}_2^*(\alpha)$. The reason to use "-like" is that A is not non-negative. The set $\mathcal{A}$ of all such matrixes is multiplicative, non-commutative group. We express the commutator $[A_1, A_2]$ of such matrixes in the form

$$[A_1, A_2] = \begin{vmatrix} \lambda_1 & \lambda_2 \\ \mu_1 & \mu_2 \end{vmatrix} \cdot D \tag{2}$$

where

$$D = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}, \qquad A_i = \begin{pmatrix} 1 & \lambda_i & 0 \\ 0 & 1-\lambda_i & \mu_i \\ 0 & 0 & 1-\mu_i \end{pmatrix}.$$

Moreover $\mathcal{A}$ is Lie group. We call $\mathcal{A}$ *L-stochastic-like Lie group with basic Möbius triangle* ABC.

By the consideration of the powers of $A \in \mathcal{A}$, we get the following

Definition 1.1. The orbit of the point $P(x, y, z)$ under the action of the matrix $A \in \mathcal{A}$, i.e. the set of points $\{P_n(x_n, y_n, z_n) = \varphi_A^{1-n}(P),\ n \in \mathbb{N}\}$, such that (x_n, y_n, z_n) is the unique solution of SLE

$$A^n X_n = B, \qquad X_n = \begin{pmatrix} z_n \\ y_n \\ x_n \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 0 \\ p \end{pmatrix}$$

is said to be *barycentric trajectory (b-trajectory)* of the point $P(\equiv P_1)$ with respect to basic $\triangle ABC$.

We get the explicit form of the b-trajectory in the form

$$\begin{aligned} z_n &= p + \frac{p}{\lambda - \mu}\left[\frac{\mu}{(1-\lambda)^n} - \frac{\lambda}{(1-\mu)^n}\right], \\ y_n &= \frac{p\mu}{\lambda - \mu}\left[\frac{1}{(1-\mu)^n} - \frac{1}{(1-\lambda)^n}\right], \\ x_n &= \frac{p}{(1-\mu)^n}. \end{aligned} \tag{3}$$

Any point $P_n(x_n, y_n, z_n) \in \alpha : x + y + z = p$ and determines

$$\lambda_n = (BCA_n) = -\frac{z_n}{y_n}, \qquad \mu_n = (AA_nP_n) = -\frac{y_n + z_n}{x_n}$$

where $A_n = AP_n \cap BC$. Thus, the position of P_n is determined by

$$\overrightarrow{CP_n} = \frac{1}{1-\mu_n}\overrightarrow{CA} - \frac{\mu_n}{1-\mu_n}\frac{1}{1-\lambda_n}\overrightarrow{CB}.$$

1.2. *Barycentric characterization of the duality in the projective plane. Projective isotomic conjugation*

Further we assume $P(x, y, z) : x + y + z = 1$, i.e. $\alpha : x + y + z = 1$.

In the following theorem we resume some well known facts from the analytic geometry of $\mathbb{E}_2^*(\alpha)$.

Theorem 1.1. *Let in the plane α of a basic $\triangle ABC$ be given in b-coordinates the points $A'(a_1, a_2, a_3)$, $B'(b_1, b_2, b_3)$ and $C'(c_1, c_2, c_3)$. Let*

$$\mathcal{M}(A'B'C') = \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{pmatrix} \qquad \textit{and} \qquad \Delta' = \det \mathcal{M}(A'B'C')$$

be the matrix and it's determinant, formed by the ordered triple (A', B', C'). Then:

(i) *A', B', C' are incident with a line* $\iff \Delta' = 0$;

(ii) *the ratio of the areas* $\dfrac{S_{A'B'C'}}{S_{ABC}} = |\Delta'|$;

(iii) *the barycentric equation of the line* $a' : \begin{cases} \mathsf{z}\ B'(b_1, b_2, b_3) \\ \mathsf{z}\ C'(c_1, c_2, c_3) \end{cases}$ *is*

$$a' : \begin{vmatrix} x & y & z \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix} = 0 \iff a' : \mathcal{A}_{23}.x + \mathcal{A}_{31}.y + \mathcal{A}_{12}.z = 0.$$

The minors $\mathcal{A}_{ij}$ are called b-coordinates of the line a' and it is written $a'[\mathcal{A}_{23},\, \mathcal{A}_{31},\, \mathcal{A}_{12}]$;

(iv) *the b-coordinates of the intersect point* $P = m \cap n$ *of the lines* $m[\mathcal{M}_1,\, \mathcal{M}_2,\, \mathcal{M}_3]$ *and* $n[\mathcal{N}_1,\, \mathcal{N}_2,\, \mathcal{N}_3]$ *are* $P\left(\dfrac{p_{23}}{\Delta}, \dfrac{p_{31}}{\Delta}, \dfrac{p_{12}}{\Delta}\right)$, *where* p_{ij} *are the minors, formed with $\{i, j\}$-columns of the matrix*

$$\mathcal{P}(m, n) = \begin{pmatrix} \mathcal{M}_1 & \mathcal{M}_2 & \mathcal{M}_3 \\ \mathcal{N}_1 & \mathcal{N}_2 & \mathcal{N}_3 \end{pmatrix} \qquad \text{and} \qquad \Delta = \begin{vmatrix} 1 & 1 & 1 \\ \mathcal{M}_1 & \mathcal{M}_2 & \mathcal{M}_3 \\ \mathcal{N}_1 & \mathcal{N}_2 & \mathcal{N}_3 \end{vmatrix};$$

(v) *the b-equation of a conic $\mathcal{C}$ is* $\mathcal{C} : (x, y, z).A.(x, y, z)^T = 0$, *where A is 3×3 real symmetric matrix. In particular, $\mathcal{C}$ passes through the vertexes of the basic triangle iff* $a_{ii} = 0$, $i = 1, 2, 3$.

There are several kinds of conjugate maps (isogonal, isotomic) with respect to fixed triangle, defined as colineations over $\mathbb{E}_2^*$, for example by means of point multiplication [6]. The theorems of Ceva and Menelaus give rise

Definition 1.2. *Projective* ($\mathbb{RP}_2$-) *isotomic conjugation* with respect to basic $\triangle ABC$ is the map

$$\begin{aligned} \pi\ : \Sigma \backslash \{\Sigma_{AB} \cup \Sigma_{BC} \cup \Sigma_{CA}\} &\longrightarrow \Lambda \backslash \{AB \cup BC \cup CA\} \\ P = AA' \cap BB' \cap CC' &\longmapsto \pi(P) = p\ \mathsf{z}\ A^*, B^*, C^*, \end{aligned}$$

where the points I', I^* are harmonically-conjugated to the pair (J, K), i.e. $I', I^*\, \mathsf{z}\ JK$ and $(JKI') = -(JKI^*)$ for $I, J, K \in \{A,\, B,\, C\}$. The Ceva point P is called *pole*, the Menelaus line $p = \pi(P)$ is called *polar line* and the map π is known also as *polarity* w.r. to $\triangle ABC$.

The image of the medicenter G is $\pi(G) = \omega$.

The following theorem is easy to check by using Theorem 1.1.

Theorem 1.2.

(i) *The* $\mathbb{RP}_2-$*isotomic conjugation is involutive corelation, i.e.* $\pi \neq \mathrm{id}$ *(the identity),* $\pi^2 = \mathrm{id}$ *(or* $\pi^{-1} \equiv \pi$*) and analytically in b-coordinates*

$$\Sigma \ni P(p_1,\, p_2,\, p_3) \xmapsto{\pi} \pi(P) = p\,[p_2p_3,\, p_3p_1,\, p_1p_2\,] \in \Lambda.$$

(ii) *The sets of stars of lines in the plane of basic* $\triangle ABC$ *and the set of conics, incident with the vertexes* A, B, C *are equivalent in the sence of the* $\mathbb{RP}_2$*-isotomic conjugation.*

2. Bézier curves and related geometric properties and objects

Bézier curves, as basic tools for Computer Aided Geometric Design [1], are intensively studied by many autors. We recall the well known

Definition 2.1. Let $\triangle A_0A_1A_2$ be fixed triangle with vertex-vector-positions $a_i = \overrightarrow{OA_i}$, $i = 0, 1, 2$. The polynomial curve of second degree

$$\mathcal{C} : r(t) = (1-t)^2 a_0 + 2t(1-t)a_1 + t^2 a_2, \quad 0 \leq t \leq 1$$

is said to be *Bézier curve* of $2^{-\mathrm{d}}$ power with basic $\triangle A_0A_1A_2$. It is also in use the denotation $\mathcal{C}_{\{A_0A_1A_2\}}$.

Generally, $\mathcal{C}_{\{A_0A_1\ldots A_n\}} : r(t) = \sum\limits_{k=0}^{n} \binom{n}{k} t^k (1-t)^{n-k}.a_k,\ t \in [0,1]$ is Bézier curve of $n^{-\mathrm{th}}$ power with basic symplex $A_0A_1\ldots A_n$, formed as the linear combination of $a_k = \overrightarrow{OA_k}$ with coefficients the basic Bernstein's polynomials $B_{k,n} = \binom{n}{k} t^k (1-t)^{n-k}$, $k = 0, 1, \ldots, n$. Point A_0 is called *initial*, A_n — *end-point*, all the rest A_k — *control points*. Any Bézier curve is invariant under the action of any affine transformation.

We use the following matrix forms for $\mathcal{C}_{\{A_0A_1A_2\}}$

$$\mathcal{C}_{\{A_0A_1A_2\}} : r(t) = (1,\, t,\, t^2) \cdot \begin{pmatrix} \overrightarrow{OA_0} \\ 2\overrightarrow{A_0A_1} \\ \overrightarrow{A_1A_0} + \overrightarrow{A_1A_2} \end{pmatrix} = (1,\, t,\, t^2) \cdot \begin{pmatrix} \overrightarrow{OA_0} \\ 2\overrightarrow{A_0A_1} \\ 2\overrightarrow{A_1M_1} \end{pmatrix},$$

where M_1 is the midpoint of A_0A_2, from which it's easy to calculate the derivatives $\dot{r}(t) = \frac{dr(t)}{dt}$ ets. and to proove the well known properties, used in the De'Casteljau algorithm:

Lemma 2.1. *Arbitrary Bézier curve* $\mathcal{C}_{\{A_0A_1A_2\}}$ *passes through* $A_0(t = 0)$ *and* $A_2(t = 1)$*, tangent to* $\overrightarrow{A_0A_1} \| \dot{r}(0)$*,* $\overrightarrow{A_1A_2} \| \dot{r}(1)$*, with curvatures at the* *initial point* A_0 *and the end-point* A_2

$$\varkappa(0) = \frac{S_{\triangle A_0A_1A_2}}{|\overrightarrow{A_0A_1}|^3}, \qquad \varkappa(1) = \frac{S_{\triangle A_0A_1A_2}}{|\overrightarrow{A_1A_2}|^3},$$

respectively, where $S_{\triangle A_0A_1A_2}$ *is the area of* $\triangle A_0A_1A_2$. *Moreover* $\mathcal{C}_{\{A_0A_1A_2\}}$ *intersects the median* A_1M_1 *at the midpoint* $M(t=\frac{1}{2})$ *and is tangent at* M *to the midsegment:* $\dot{r}(\frac{1}{2}) \| A_0A_2$.

From barycentric point of view any Bézier curve $\mathcal{C}_{\{A_0A_1A_2\}}$ is the set of points P with b-coordinates with respect to basic $\triangle A_0A_1A_2$ — the basic Bernstein polinomials of second degree $P\Big((1-t)^2,\, 2t(1-t),\, t^2\Big)$, $t \in [0,1]$.

Lemma 2.2.

(i) *Let* $P\Big((1-t)^2,\, 2t(1-t),\, t^2\Big)$, $t \in [0,1]$ *be arbitrary point of the Bézier curve* $\mathcal{C}_{\{A_0A_1A_2\}}$. *The b-coordinates of the points:* $B_0 = A_0P \cap A_1A_2$, $B_1 = A_1P \cap A_0A_2$, $B_2 = A_2P \cap A_1A_0$ *are*

$$B_0\left(0;\, \frac{2(1-t)}{2-t};\, \frac{t}{2-t}\right), \qquad B_1\left(\frac{(1-t)^2}{2t^2-2t+1};\, 0;\, \frac{t^2}{2t^2-2t+1}\right),$$
$$B_2\left(\frac{1-t}{1+t};\, \frac{2t}{1+t};\, 0\right).$$

(ii) *Any Bézier curve* $\mathcal{C}_{\{A_0A_1A_2\}}$ *generates the functional matrix* $A(t)$ *from the L-stochastic-like Lie group* $\mathcal{A}$ *with*

$$\lambda(t) = \frac{t}{2(t-1)} \qquad \text{and} \qquad \mu(t) = \frac{t(t-2)}{(t-1)^2}, \qquad t \in [0,\, 1).$$

Proof. The statement (i) follows by applying Theorem 1.1. and (ii) — by calculating the affine ratios $\lambda(t) = (A_2A_1B_0)$ and $\mu(t) = (A_0B_0P)$. □

For $\triangle A_0A_1A_2$ we denote: $l_i = |\overrightarrow{A_jA_k}|$, $m_i = |\overrightarrow{A_iM_i}|$, M_i — the midpoints of A_jA_k for $i \neq j \neq k \in \{0,1,2\}$, $\varphi = \measuredangle A_0A_1A_2$ and $\psi = \measuredangle A_0A_1M_1$.

Theorem 2.1. *Let* $\mathcal{C}_{\{A_0A_1A_2\}}$ *be Bézier curve.*

(i) *For any* $P \in \mathcal{C}_{\{A_0A_1A_2\}}$ *the ratio of the areas*

$$s(t) = \frac{S_{\triangle B_0B_1B_2}}{S_{\triangle A_0A_1A_2}} = \frac{4t^2(1-t)^2}{(2-t)(1+t)(2t^2-2t+1)} \in \left[0;\, \frac{2}{9}\right],$$

$\max\limits_{[0,1]} s = s(\frac{1}{2}) = \frac{2}{9} \iff P \equiv M(\frac{1}{4};\, \frac{1}{2};\, \frac{1}{4})$ — *the midpoint of the median* A_1M_1 *for* $\triangle A_0A_1A_2$;

(ii) *The length of the curve is*

$$l_{\mathcal{C}} = \frac{2m_1 - l_2\cos\psi}{2m_1}\sqrt{(2m_1 - l_2\cos\psi)^2 + l_2^2\sin^2\psi} + \frac{l_2^2\cos\psi}{2m_1} + \frac{l_2^2\sin^2\psi}{2m_1}\ln\frac{\sqrt{(2m_1 - l_2\cos\psi)^2 + l_2^2\sin^2\psi} + 2m_1 - l_2\cos\psi}{l_2(1-\cos\psi)}.$$

Proof. The statement (i) follows from Theorem 1.1.-(ii) and standard obtaining of the $\max s(t)$ over $[0,1]$. For (ii) we apply the lenght-formula

$$l_{\mathcal{C}} = \int_0^1 |\dot{r}(t)|dt \qquad \text{for} \qquad |\dot{r}(t)| = 2\sqrt{(1,2t).\Gamma(\overrightarrow{A_0A_1}\overrightarrow{A_1M_1}).\begin{pmatrix}1\\2t\end{pmatrix}}$$

where $\Gamma(\overrightarrow{A_0A_1}\overrightarrow{A_1M_1})$ is the Gramm-matrix for the pair $\{\overrightarrow{A_0A_1}, \overrightarrow{A_1M_1}\}$. □

Theorem 2.2. *The curvature $\varkappa(t)$ of any Bézier curve $\mathcal{C}_{\{A_0A_1A_2\}}$ is*

$$\varkappa(t) = S_{\triangle A_0A_1A_2}.[g(t)]^{-3/2},$$

where the function $g(t)$, defined over $[0,1]$, is of the form

$$g(t) = |(t-1)\overrightarrow{A_1A_0} + t\overrightarrow{A_1A_2}|^2 = (t-1\,;t)\begin{pmatrix} l_2^2 & l_0l_2\cos\varphi \\ l_0l_2\cos\varphi & l_0^2\end{pmatrix}\begin{pmatrix}t-1\\t\end{pmatrix}.$$

Proof. It follows by direct calculations, analogous to Theorem 2.1.-(ii), over the well known formula $\varkappa(t) = \dfrac{|\dot{r}\times\ddot{r}|}{|\dot{r}|^3}$. □

Theorem 2.2. implies immediately the well known

Corollary 2.1. *The Bézier curve $\mathcal{C}_{\{A_0A_1A_2\}}$ is a parabola with vertex at point K, such that $\overrightarrow{OK} = r(t_0)$, where*

$$t_0 = \frac{l_2^2 + l_0l_2\cos\varphi}{l_2^2 + 2l_0l_2\cos\varphi + l_0^2}$$

is the unique (global) maximum for $\varkappa(t)$: $\varkappa_{|_K} = \varkappa(t_0) = \max\limits_{[0,1]}\varkappa(t)$.

Moreover, the result of Theorem 2.2. gives additional geometric information of the shape of Bézier curve $\mathcal{C}_{\{A_0A_1A_2\}}$, considered and discussed in [2] and generalized in [3] for cubic Bézier curve.

By cosidering the action of $\mathbb{RP}_2$-isotomic conjugation over each point of $\mathcal{C}_{\{A_0A_1A_2\}}\setminus\{A_0, A_2\}$, jointly with Lemma 2.2. (i), we get

Theorem 2.3. *The* $\mathbb{RP}_2$*-isotomic-conjugated image of* $\mathcal{C}_{\{A_0A_1A_2\}}$ *is*

$$\pi(\mathcal{C}_{\{A_0A_1A_2\}}) = \Big\{p\,[2t^2;\ t(1-t);\ 2(1-t)^2]\ :\ t \in (0,1)\Big\},$$

i.e. is the 1-parameter set of polar lines

$$p = \pi(P) : 2t^2 \cdot x + t(1-t) \cdot y + 2(1-t)^2 \cdot z = 0, \quad t \in (0,1),\ u, v \in \mathbb{R}.$$

The lines, corresponding to $t = \frac{1}{3}, \frac{1}{2}, \frac{2}{3}$*, are the polar lines of the intersecting points of* $\mathcal{C}_{\{A_0A_1A_2\}}$ *with the medians through* A_2, A_1, A_0 *and are parallel to* A_1A_0*,* A_0A_2*,* A_2A_1*, respectively.*

From projective point of view w.r. to the space coordinate system $\mathcal{K} = \{O, a_1, a_2, a_3\}$ the obtained 1-parameter family of lines p, give rise 1-parameter family of the planes $\sigma : \begin{cases} \mathsf{z}\ O \\ \mathsf{z}\ p \end{cases}$. The envelope of the last one is conic surface $\mathcal{S}$ with vertex O and generator $\mathcal{E}_{\pi(\mathcal{C})} = \mathcal{S} \cap \alpha : x + y + z = 1$ — the envelope of $\pi(\mathcal{C}_{\{A_0A_1A_2\}})$. More precisely, after standard calculations, we get the following

Theorem 2.4. *The envelope* $\mathcal{S}$ *of the set of planes* $\sigma : \begin{cases} \mathsf{z}\ O \\ \mathsf{z}\ \pi(\mathcal{C}_{\{A_0A_1A_2\}}) \end{cases}$ *with respect to* $\mathcal{K} = \{O, a_1, a_2, a_3\}$ *is the real hyperbolic conic surface*

$$\mathcal{S} : 16xz - y^2 = 0$$

with generator in the basic plane $\alpha :\ x + y + z = 1$*, the envelope*

$$\mathcal{E}_{\pi(\mathcal{C})} : \left(x = \frac{(1-t)^2}{6t^2 - 6t + 1}; y = \frac{4t(t-1)}{6t^2 - 6t + 1}; z = \frac{t^2}{6t^2 - 6t + 1} \right)$$

(in b-coordinates) and $t \in (0,\ 1) \backslash \{\frac{1}{2} \pm \frac{1}{2\sqrt{3}}\}$.

Proof. From Theorem 2.3. the b-equation $F(x, y, z, t) = 0$, where $F(x, y, z, t) = 2t^2.x + t(1-t).y + 2(1-t)^2.z$ for $t \in (0,1)$ describes the 1-parameter set of polar lines. Then the envelope is obtained by the solutions of

$$\mathcal{E}_{\pi(\mathcal{C})} : \begin{cases} F(x, y, z, t) = 0 \\ \dfrac{\partial F}{\partial t}(x, y, z, t) = 0 \\ x + y + z = 1 \end{cases} \iff \begin{pmatrix} 2t^2 & t(1-t) & 2(1-t)^2 \\ 4t & 1-2t & 4(t-1) \\ 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

By solving the SLE with respect to x, y, z we proof the theorem. □

Analogous to Lemma 2.2, we get

Lemma 2.3. *The envelope* $\mathcal{E}_{\pi(\mathcal{C})}$ *of* $\mathbb{RP}_2$*-isotomic image of any Bézier curve* $\mathcal{C}_{\{A_0A_1A_2\}}$ *generates the functional L-stochastic-like matrix* $A_\varepsilon(t)$ *with*

$$\lambda(t) = \frac{t}{4(1-t)}, \qquad \mu(t) = \frac{t(4-5t)}{(t-1)^2}, \qquad t \in (0,\,1) : t \neq \frac{4}{5},\ \frac{1}{2} \pm \frac{1}{2\sqrt{3}}.$$

By applying (2) to Lemma 2.2 (ii) and Lemma 2.3, we calculate the commutators and obtaine

Lemma 2.4. *Let* $A(t)$ *and* $A_\varepsilon(t)$ *be the functional matrices in* $\mathcal{A}$ *of* $\mathcal{C}_{\{A_0A_1A_2\}}$ *and* $\mathcal{E}_{\pi(\mathcal{C})}$ *respectively. For* $t_1 \neq t_2$ *in* $(0,1)$*, the following relations for the commutators are valid*

$$[A(t_1),\, A(t_2)] = -2[A_\varepsilon(t_1),\, A_\varepsilon(t_2)] = \frac{t_1t_2(t_2-t_1)}{2(1-t_1)^2(1-t_2)^2}.D.$$

By the notion *b-trajectory* of a point we consider *b-trajectories* of the Bézier curve $\mathcal{C}_{\{A_0A_1A_2\}}$ as the family of lines $\mathcal{C}^{(n)}_{\{A_0A_1A_2\}}$ formed by the n^{th}–b-image of any point of $\mathcal{C}_{\{A_0A_1A_2\}}$, $\forall\, n \in \mathbb{N}$.

Theorem 2.5. *The b-trajectories of the Bézier curve* $\mathcal{C}_{\{A_0A_1A_2\}}$ *are the elements of the sequence of curves*

$$\mathcal{C}^{(n)}_{\{A_0A_1A_2\}} : r_n(t) = x_n(t).a_0 + y_n(t).a_1 + z_n(t).a_2, \quad t \in [0,\,1],$$

with

$$\begin{aligned}&\big(x_n(t); y_n(t); z_n(t)\big)\\ &= \left((t-1)^{2n};\ 2f(t)\begin{vmatrix}(t-1)^n & 2^n\\ 1 & (t-2)^n\end{vmatrix}; 1-f(t)\begin{vmatrix}(t-1)^{n+1} & 2^{n+1}\\ 1 & (t-2)^{n-1}\end{vmatrix}\right)\end{aligned} \tag{4}$$

and where $f(t) = \dfrac{(1-t)^n}{(t-3)(2-t)^{n-1}}$.

(i) *For arbitrary* $n \in \mathbb{N}$, $r_n(0) = a_0$, *i.e.* $A_0 \in \mathcal{C}^{(n)}_{\{A_0A_1A_2\}}$, $r_n(1) = a_2$ *and* $A_2 \in \mathcal{C}^{(n)}_{\{A_0A_1A_2\}}$.

(ii) *The tangent vectors to each* $\mathcal{C}^{(n)}_{A_0A_1A_2}$ *at* A_0 *and* A_2 *are*

$$\dot{r}_n(0) = 2n \cdot (a_1 - a_0) = 2n \cdot \overrightarrow{A_0A_1}, \qquad \forall n \in \mathbb{N},$$

$$\dot{r}_n(1) = \dot{f}(1).2^{n+1}(a_2 - a_1) = \begin{cases} 0, & \textit{for} \quad n \geq 2\\ 2(a_2 - a_1) = 2.\overrightarrow{A_1A_2}, & \textit{for} \quad n = 1.\end{cases}$$

(iii) *The envelope* $\mathcal{E}_n$ *of* $\pi(\mathcal{C}^{(n)}_{A_0A_1A_2})$, *in b-coordinates, is*

$$\mathcal{E}_n : \begin{cases} x = \Delta_1.\Delta^{-1} \\ y = \Delta_2.\Delta^{-1}, \\ z = \Delta_3.\Delta^{-1} \end{cases}$$

for all (t, n) *in order to have well defined*

$$\Delta_1 = \frac{-1}{y_n^2}\left(\frac{y_n}{z_n}\right)^{\bullet}, \quad \Delta_2 = \frac{-1}{z_n^2}\left(\frac{z_n}{x_n}\right)^{\bullet}, \quad \Delta_3 = \frac{-1}{x_n^2}\left(\frac{x_n}{y_n}\right)^{\bullet},$$

and $\Delta = \sum_{i=1}^{3} \Delta_i$.

Proof. By applying Lemma 2.2.(ii) jointly with (3) we get (4) and calculate (i). We express the derivatives $\dot{x}_n$, $\dot{y}_n$, $\dot{z}_n$ of (4) are as follows:

$$\dot{x}_n = 2n(t-1)^{2n-1},$$

$$\dot{y}_n = \frac{2f(t)}{(t-1)(t-2)(3-t)} \begin{vmatrix} (t-1)^n(t-2)^n & 2^n \\ t-1+n(t-3) & t-1+2n(t-2)(3-t) \end{vmatrix},$$

$$\dot{z}_n = \frac{2f(t)}{(t-1)(t-2)(3-t)} \begin{vmatrix} (t-1)^{n+1}(t-2)^{n-1} & -2^n \\ t-1+n(t-3) & (3-t)[1-n(t-2)]-1 \end{vmatrix}.$$

Thus we calculate (ii).

(iii) The $\mathbb{RP}_2-$ isotomic conjugated image of each $\mathcal{C}^{(n)}_{\{A_0A_1A_2\}}$ is the 1-parameter family of lines

$$\pi(\mathcal{C}^{(n)}_{\{A_0A_1A_2\}}) = p_n : \underbrace{\frac{1}{x_n(t)} \cdot x + \frac{1}{y_n(t)} \cdot y + \frac{1}{z_n(t)} \cdot z}_{=F_n(x,y,z,t)} = 0.$$

The envelope $\mathcal{E}_n$ of $\pi(\mathcal{C}^{(n)}_{\{A_0A_1A_2\}})$ is obtained by the solutions of the system

$$\mathcal{E}_n : \begin{cases} x+y+z=1 \\ F_n(x,y,z,t)=0 \\ \dfrac{\partial F_n}{\partial t}(x,y,z,t)=0 \end{cases} \iff \mathcal{E}_n : \begin{cases} x+y+z=1 \\ x_n^{-1}\cdot x + y_n^{-1}\cdot y + z_n^{-1}\cdot z = 0 \\ \dot{x}_n x_n^{-2}\cdot x + \dot{y}_n y_n^{-2}\cdot y + \dot{z}_n z_n^{-2}\cdot z = 0. \end{cases}$$

By applying the Cramer's rule for all t and n, each determinant to be well defined, we get the evelope $\mathcal{E}_n$ in b-coordinates. □

By (iii) of Theorem 2.5 in the space arises the sequence of conic surfaces $\mathcal{S}_n$ with vertex O and generators $\mathcal{E}_n$ — the envelopes of $\pi(\mathcal{C}^{(n)}_{\{A_0A_1A_2\}})$.

Acknowledgements

A support by Scientific researches fund of “St. Cyril and St. Methodius” University of V. Tarnovo under contract RD-491-08/2008 is gratefully acknowledged.

References

1. G. Farin, *Curves and Surfaces for Computer Aided Geometric Design*, 4th ed., Academic Press, London, 1997.
2. G. Georgiev, *Curve and Surface Design: Avignon 2006*, pp. 143, Nashboro Press, Brentwood, TN, 2007.
3. G. Georgiev, *Pure and Appl. Diff. Geom. PADGE 2007*, pp. 98, Shaker Verlag, Aachen, 2007.
4. K.R.S. Sastry, *Forum Geometricorum* **4**, 73 (2004).
5. V. Volenec, *Mathematical Communications* **8**, 55 (2003).
6. P. Yiu, *Int. J. Math. Educ. Sci. Technol.* **31**, 569 (2000).

20210502

Data-NowVivo-010-2021\
20210502\
Trends in Differential Geometry, Complex Analysis and Mathematical Physics\
Trends in Differential Geometry ....pdf

Pág. 109 (126)

Some Geometric Properties and Objects Related to Bézier Curves
(Hristov)

Barycentric coordinates and barycentric trajectories

→ real projective plane
$\mathbb{R}\mathbb{P}_2$.

Pág. 111 (128)

1.2 Barycentric characterization of the duality in the projective plane.

Projective isotomic conjugation

Import:

Pág. 112 (129)

There are several kinds of conjugate maps (isogonal, isotomic) with respect to fixed triangle, defined as colineations over E_2^{*} for example by means of point multiplication [6].

The theorems of Ceva and Menelaus give rise.

Definition 1.2.

Projective ($\mathbb{R}P_2$) – isotomic conjugation with respect to basic ΔABC is the map

P_2

Pág. 120 (137)

—

Canonical commutation relation

↳ canonical conjugate quantities

Conjugate variables are pairs of variables mathematically defined in such a way that they become Fourier transform duals, or more generally are related throught Pontryagin duality. The duality relations lead naturally to an uncertainty relation (Heisenberg uncertain).

Source: Pag. 120 (137)

Heisenberg relations in the general case.

=

Lo210503m

Heisenberg uncertainty principle

$$\Delta x \Delta p \geq \frac{h}{4\pi}$$

=

Planck's ~~[illegible]~~ constant.

Conjugate variables — are pairs of variables mathematically defined in such a way that they become Fourier transform dual or more generally are related through Pontryagin duality

(Heisenberg uncertainty principle)

Look: Noether's theorem

Which states that if the laws of physics are invariant with the respect to a change in one of the conjugate variable,

↓ Pappus theorem → conjugate points
(chacar)

then the other conjugate variable will not change with time
(it will be conserved)

Source:

canonical commutation relation
↔ canonical conjugate

Lei de Leavit → relação
período - luminosidade.

Astro mais brilhante a magnitude,
mais longo o período.

Em 2008 → membros da
comunidade astronômica,
votaram pela mudança do nome
do que antes era chamado de
relação período- luminosidade
para Lei de Leavit.

—

Vontade de Deus
+
liberdade do homem.

—

O homem morre pelo mau
uso da liberdade do homem.

—

A maioria das pessoas morrem
pelo mau uso da liberdade do
homem e não por vontade de Deus,
mas quando chega a hora é a
vontade de Deus?

—

d-dimensional Euclidean Space is
$s[d]r^d$ where
$s[d] = \pi^{d/2}/(d/2)!$ (the surface area is $d\, s[d] r^{d-1}$).
(The function $s[d]$ has a maximum around $d = 5.26$, then decreases rapidly with d.)

—

$$2\pi a^2(1-\mathrm{Cos}[r/a]) =$$
$$\pi r^2(1-r^2/(12a^2)+r^4/(360a^4)-\ldots)$$

—

emeraldcloudlab.com

$\underline{\underline{20210506}}_M$

575	240
1180	175
1755	415
	160
	575

1 US\$ = 5,20
US\$ = 337,50

personalpages.to.infn.it/~musich/

—

sphera → elipsoide (forças "gravidade")

(')

+,- * *. ! * 6 7 K# 4!

! ; L #ML !77 ! < 7 ! G) 7#

(% ""

téristique en fonction de la vitesse de la lumière sous la forme $w = \alpha_G c$, on peut faire apparaître, comme l'ont proposé Agnese et Festa en 1997, une constante de couplage gravitationnelle α_G qui a un statut en partie semblable, pour chaque sous-système auquel elle s'applique, à celui de la constante de structure fine de l'électromagnétisme.

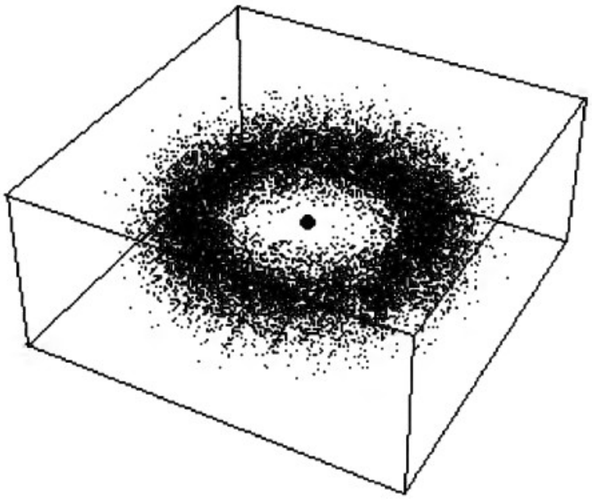

Figure 1.1. Exemple de sous-structure formée dans un disque protoplanétaire selon l'approche relativiste d'échelle.
(Transformation de l'équation fondamentale de la dynamique en une équation de forme schrödingerienne sous des conditions de fractalité.)

La résolution de l'équation de Schrödinger fondée sur un tel paramètre fournit des fonctions d'onde dont le carré du module donne la densité de probabilité attendue pour la distribution des corps dont on a ainsi décrit le mouvement. Cette distribution n'est pas plate, mais montre au contraire des pics de probabilité correspondant à des valeurs particulières des quantités conservatives comme l'énergie, le moment angulaire ou le vecteur de Runge-Lenz. Ainsi le disque initial est amené à s'organiser en sous-structures (fig. 1.1). Les planètes qui se formeront ensuite à partir de ces structures auront (une fois que le système se sera stabilisé, agrandissant ainsi l'horizon de prédictibilité) des éléments orbitaux plus probables qui pourront être calculés en fonction de ces quantités conservatives (qui sont donc communes aux deux régimes, classique et macroquantique).

On prédit ainsi que les demi-grands axes doivent montrer des pics de probabilité pour des valeurs : $a_n = (GM/w^2)\, n^2$

– où n est un entier, et que les excentricités sont elles-mêmes appelées à montrer des valeurs préférentielles données par : $e_{kn} = k/n$

2021 05 10za

Schrodinger

Source:

Data-NovoViro-010-2021\
Books-010-2021\
20210509\
Aide-Memoire de Mathématiques
Supérieures\
→ Des-fleurs-pour-Schr-dinger...pdf

Pag. 28(33)

Figure 1.1 →

Transformation de l'équation fondamentale de la dynamique en une équation de forme schrödingerienne sous de conditions de fractalité).

=

→ \Exercices-corrig-s-sur-la-Convexit.pdf\

Pag. 27(29) 1.2.2
Fonctions eulériennes

calculer l'intégrale

=

(@

20210513_M

I hope you're happy
Blue October

—

the Bandit King of Leon

—

the Killers My Own Soul's Warning
(Michael Hill video)

20200514_M

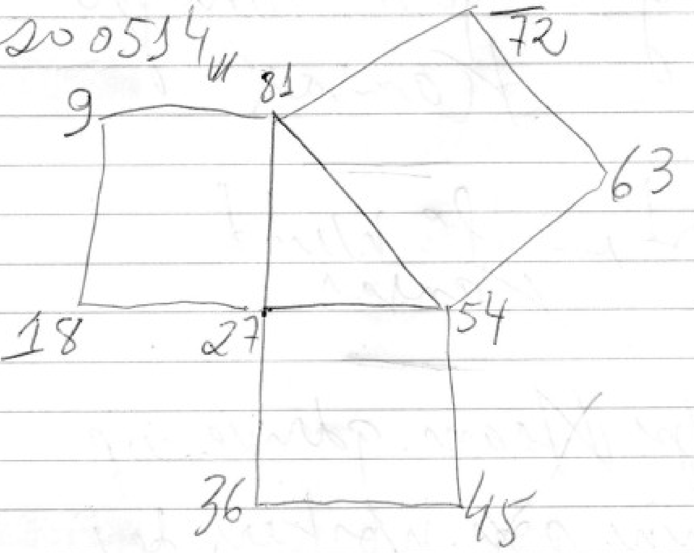

Source:
Books-10-2021\
20210514\
Faith Javare\
Javare Faith-
La Clave....pdf
(Capa)

—

20210515

veracrypt.fr (free)

R52 └3,78

13,756613 x24=330,158730

//

—

(@'

20210515

Eletromagnetismo (responsável pela luz, eletricidade, ímãs)

A força nuclear forte (responsável por manter tudo unido)

A força nuclear fraca (responsável pela radiatividade)

Para cada força existe uma ~~fo~~ partícula intermediadora.

Fótons são as partículas intermediadoras das forças eletromagnéticas

glúons → força forte

bósons pesados → força fraca.

Blocos de construção da matéria são formados pelas duas categorias:
bósons
férmions

Toda a matéria → 3 férmions:
2 Quarks
1 elétron.

2 tipos de Quarks no modelo padrão →

mas apenas 2 deles – os mais leves –
formam prótons e
nêutrons.

Off: Quarks beleza (decaimento)
geram
múons ou
elétrons
(produzindo ambos
em quantidades
iguais).

Nos experimentos
começam a provar algo
diferente.
Não se sabe se é por causa de:

a/ uma nova partícula ou
b/ nova força fundamental.

Quarks beleza < múons
> elétron
↑
decaíam com menos ou mais
frequência.

" (
+,- * *. ! * " ! ' 7! 7
' #?;
+,- * *. ! * " ! ' 7! 7 ' 7#
(' " 9

damental groups were used to distinguish them. A part of his table is depicted in Fig. 1.8.

1.4 Broken Surfaces Diagrams

Let a knotted surface $K = f(F)$ be given. Consider a 3-dimensional hyperplane $H \subset \mathbb{R}^4$ which we identify with $\mathbb{R}^3$ once and for all. The hyperplane H is thought of as the orthogonal complement to a vector $\mathbf{v} \in \mathbb{R}^4$, and the knotted surface K may be translated to be disjoint from H. The knotted surface K may also be perturbed slightly so that its projection p into $H = \mathbb{R}^3$ is a generic surface. Note that the projection p in this section is different from the projection $p_{\mathbf{v}}$ used in the previous sections.

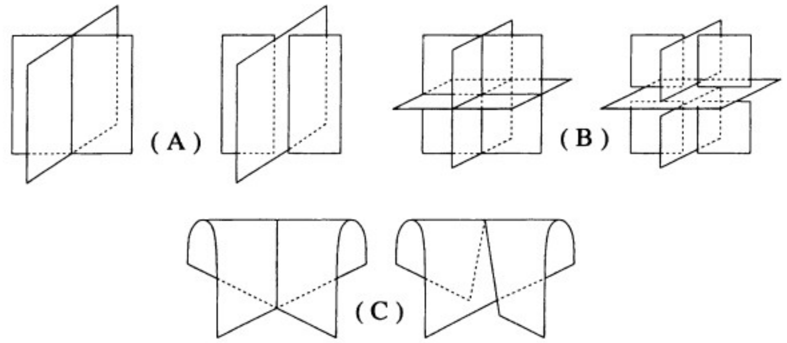

Fig. 1.9. Broken surface diagrams near singular points

This means that the image $p(f(F))$ has generic double points, triple points and branch points. That is, for any $y = p(f(x))$ there is a neighborhood N_y in $\mathbb{R}^3$ and a diffeomorphism $\psi : N_y \to \mathbb{R}^3$ such that the image of F under ψ looks like the intersection of 1, 2, or 3 coordinate planes or the cone on a figure 8 (Whitney umbrella). In these cases the point y is said to be a *non-singular point*, *a double point*, a *triple point*, or a *branch point*, respectively. Neighborhoods of double, triple, and branch points are depicted in Fig. 1.9, at the left of (A), (B), (C) respectively.

In N_y, when y is a double point, we can distinguish the relative height (with respect to the direction $\mathbf{v}$) of the two sheets in N_y as upper and lower. At a triple point we have an upper (top), middle, and lower (bottom) sheet. And at a branch point the upper and lower sheet converge at the branch point.

Now we describe the preimage of singular points

$$\{ \ x \in K \mid p(x) = p(y) \text{ for some } y \in K, x \neq y \ \}.$$

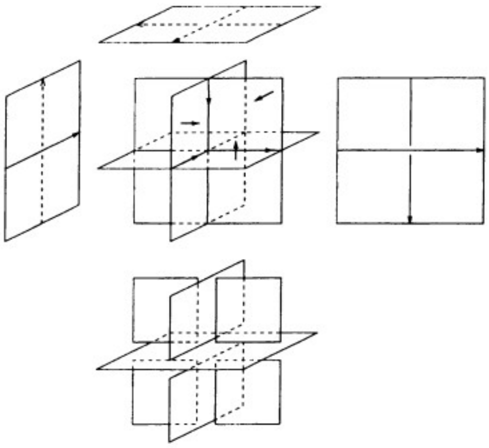

Fig. 1.10. Decker curves around a triple point

We let D, T, and B denote the sets of double, triple, and branch points in $\mathbb{R}^3$, respectively. We regard T and B as subsets of D. The preimages on K of the double points are denoted by D_a and D_b with labels representing above and below sheets. Thus D_b is lower with respect to the height direction $\mathbf{v}$ than D_a. The sets D_a and D_b are called upper and lower decker sets, respectively. We use the same terminology for the preimages in F, where $K = f(F)$, and call $f^{-1}(D_a)$ and $f^{-1}(D_b)$ the upper and lower decker sets as well. The transverse components of D_a and D_b are arcs that intersect at the preimages of triple points or branch points. The preimage on K of a branch point is a point that is in the common boundary of D_a and D_b. We use the same letter B to denote the branch points in $\mathbb{R}^3$ and on the knotted surface K. The preimage of a triple point, x, consists of three points x_t, x_m and x_b that lie in the top, middle, and bottom sheet, respectively. The preimage of the set T is the union of the three sets T_t, T_m, and T_b, where these are the collections of such top, middle, and bottom preimages of triple points, respectively. A point x_t is the interesction of two sub-arcs of D_a. The corresponding point x_m in the middle sheet is the intersection of an arc in D_a and an arc in D_b. The corresponding point x_b in the bottom sheet is the intersection of two arcs in D_b. Moreover, these six arcs are paired (top to middle, top to bottom, and middle to bottom) and project to the three arcs in the double point set that intersect at x. The situation at a triple point is depicted in Fig. 1.10. In the middle of the figure, a neighborhood of a triple point is depicted. The preimages for each sheet are depicted around it. The decker sets D_a and D_b are represented by dotted and solid lines, respectively. At the bottom of the figure, a broken surface diagram at a triple point is shown, which we explain below.

Now we generalize the convention for classical knot diagrams, where the under-arc is broken to indicate crossing information, to knotted surfaces. Let

N be a thin open neighborhood of $D_b \setminus B$, such that $B \subset \overline{N}$; where $\overline{N}$ denotes the closure. In other words, N is a thin strip along D_b, such that if an end point of an arc of D_b is a branch point, then the branch point is on the boundary of $\overline{N}$. Let $K' = K \setminus N$. Then the image $p(K)$ is a compact surface in $\mathbb{R}^3$ where the lower sheets are consistently removed along the double point curves. Thus this gives a diagram of a surface that is analogous to a classical knot diagram.

A *broken surface diagram* consists of a generic surface in $\mathbb{R}^3$ such that one of the sheets along the double point curve are consistently broken as shown in Fig. 1.9 (A) right. At triple and branch points, the surface is locally broken as indicated in Fig. 1.9 (B) and (C) right, respectively, and consistently extended to double point curves (see [CarSai98a] for more details). We have shown, in the preceding paragraph, that we obtain a broken surface diagram from a projection of a knotted surface, by breaking $p(K)$ along D_b. Similar conventions have been known for some time, and different colors or labels were used to indicate over and under crossing information (see [Fra87, HaKaShi99, Rose93, Yaji64], for example).

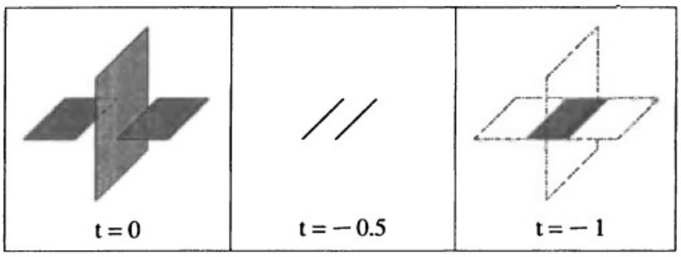

Fig. 1.11. Filling in a broken surface near a double point curve

On the other hand, given a broken surface diagram, an embedding of a surface is easy to construct as we now indicate. First we embed the broken surface, which is a surface with boundary, into $\mathbb{R}^3 = \mathbb{R}^3_0 = \{(x, y, z, 0)\}$. Next we will fill in the neighborhoods of the lower decker sets in hyperplanes $\mathbb{R}^3_t$ for $t < 0$. Figures 1.11, 1.12, and 1.13 indicate how the neighborhoods of the lower decker sets are embedded in $\mathbb{R}^3_{[0,-2]}$.

Near a double arc in the levels $\mathbb{R}^3_t$, for $t \in (-1, 0)$, a pair of arcs parallel to the boundary of the broken surface are embedded. At the level $\mathbb{R}^3_{-1}$, a strip is embedded. The union of the arcs and the strip form a 2-disk embedded in $\mathbb{R}^4$. Such a disk is embedded along any arc of double points, and these disks are chosen to match-up along closed double point arcs, or at triple points and branch points. The situation is depicted in Fig. 1.11.

Near a triple point, arcs parallel to the boundary of the broken surface are embedded in $\mathbb{R}^3_t$, for $t \in (-1, 0)$. The arcs in the bottom sheet look like the doubled "+" and the arcs in the middle sheet look like those for the

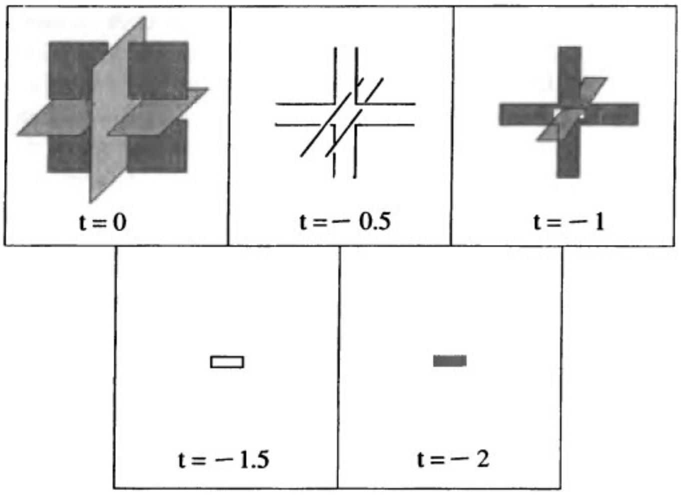

Fig. 1.12. Filling in a broken surface near a triple point

double curves described above. In Fig. 1.12, the middle picture at $t = -0.5$ depicts this situation. At the $t = -1$ level, these parallel arcs are filled in by rectangular (and "+" shaped) disks, except inside the small rectangle in the middle of "+" shaped arcs, through which the other rectangle goes. The rectangle continues in the levels $t \in (-2, -1)$, and is filled in by a smaller rectangular disk at $t = -2$ as depicted in the figure.

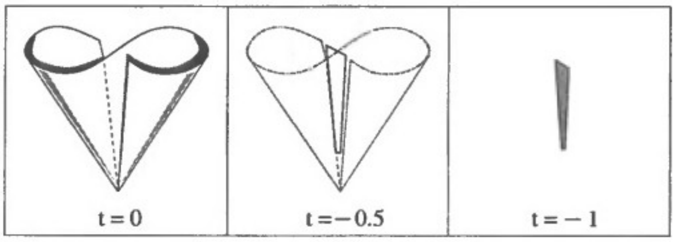

Fig. 1.13. Filling in a broken surface near a branch point

Near a branch point, an unfilled trapezoid is found in the levels $t \in (-1, 0)$, and a filled trapezoid is found in the $t = -1$ level, as depicted in Figure 1.13.

We summarize this convention as follows:

Proposition 1.21. *Any knotted surface gives rise to a broken surface diagram. Any broken surface diagram gives a knotted surface.*

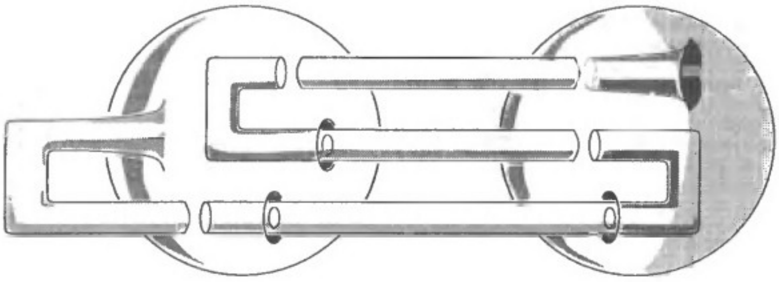

Fig. 1.14. The Fox-Milnor example as a broken surface

Example 1.22. Figure 1.14 depicts a broken surface diagram for the Fox-Milnor example in Fig. 1.1. In this figure, there are no triple points nor are there branch points.

Remark 1.23. On the other hand, it is not true that any generic surface in $\mathbb{R}^3$ is a projection of a knotted surface in $\mathbb{R}^4$. In other words, there are generic surfaces in $\mathbb{R}^3$ that do not lift to embeddings in $\mathbb{R}^4$. Such an example was described first in [Gill82]. Further criteria were found in [Ak96, CarSai98b, Sat00c], and higher (than 4) dimensional examples were given in [Oga01]. This is a phenomenon distinctive in higher dimensions, as any generically immersed circle in the plane obviously lifts to an embedding in $\mathbb{R}^3$.

Furthermore, in [Oga01], another remarkable fact was presented: there are projections of n-knots, for any $n > 2$, that do not lift to the unknot. The case $n = 2$ is open.

In case the knotted surface is oriented, the double point set D is oriented as follows. First, an orientation normal can be chosen in 3-space at every point of $p(K)$, except at branch points, in such a way that the ordered pair (tangent, normal) agrees with the fixed (say, right-handed) orientation of 3-space. A double point curve can be given an induced orientation via the triple (normal to upper sheet, normal to lower sheet, tangent to double curve) forming the right-handed orientation for 3-space.

Triple points are also oriented, in the sense that a sign, positive or negative, is assigned to each triple point as follows. A triple point is *positive* if and only if the ordered triple (normal to top sheet, normal to middle sheet, normal to bottom sheet) agrees with the orientation of 3-space.

1.5 Charts and Elementary String Interactions

In this section, we describe a planar graph, called a *chart*, that represents a given knotted surface. A chart will be a labeled planar graph. Consider a knotted surface K and perturb its image in $\mathbb{R}^4$ so that the restriction of

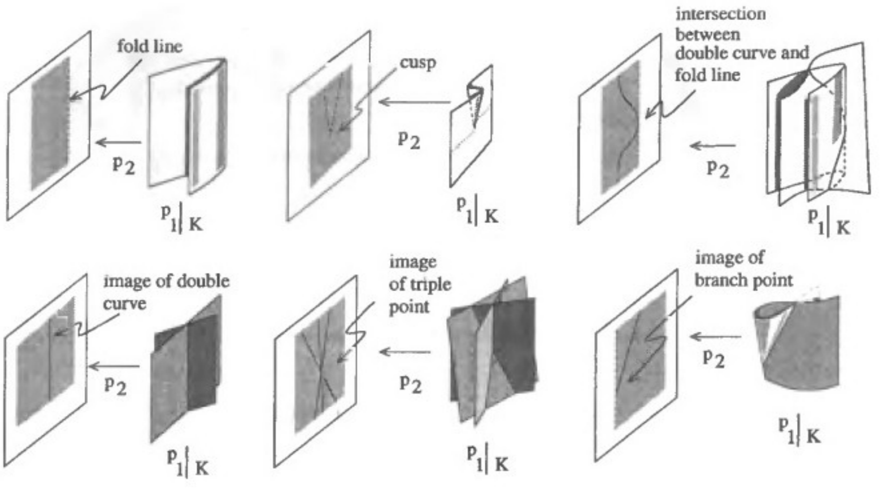

Fig. 1.15. The generic projections of knotted surfaces into planes

a projection $p : \mathbb{R}^4 \to \mathbb{R}^2$ onto a 2-dimensional affine subspace is a generic map between surfaces. The projection p can be factored (after a small perturbation if necessary) as the composition of a projection p_1 from $\mathbb{R}^4$ onto a 3-dimensional subspace $\mathbb{R}^3$ and another projection p_2 from $\mathbb{R}^3$ onto a 2-dimensional subspace $\mathbb{R}^2$, such that $p_1|_K$ and $p_2|_{(p_1(K))}$ are both generic.

Thus $p_1|_K$ has generic double points, triple points, and branch points. Meanwhile, the generic singularities of $p|_K$ consist of fold lines and cusps. The projection of a double point arc under p_2 is an arc in the plane. The projection of a triple point under p_2 is a valence 6 vertex, that is the intersection of three double point arcs. The double point curve could intersect with the fold lines in a single point, which looks like a point of tangency in $p|_K$ between a double point curve and a fold line. In this situation branch points are the end-points of double arcs that lie on fold lines. The situation of these projections is depicted in Fig. 1.15. In the figure, the projections of double point curves are depicted by solid lines, and those of fold lines are depicted by dotted lines.

We distinguish, in notation, the set D of double points of the projection, $p_1|_K$, from the image $p_2(D) = D^*$ which is a collection of edges in the plane. Similarly the triple point set, T, projects to $p_2(T) = T^*$ as the set of valence 6 vertices in the projection. The branch point set, B, projects to $p_2(B) = B^*$ which is the collection of boundary points of D^* that lie on the fold set $\mathcal{F}^*$. The set $\mathcal{F} = p_2^{-1}(\mathcal{F}^*)$ is called the *fold locus*. This is the union of the points at which p_2 has rank ≤ 1. The set, C, of cusps is the set of points at which the rank of p_2 is 0; its image in the plane is denoted by C^*. Also denote by DF the set of intersection $D \cap \mathcal{F}$, and $DF^* = p_2(DF)$.

In case the surface is oriented, the projection D^* of the double point set D is given an induced orientation inherited from the orientation of D. The fold lines are co-oriented as follows. The coorientation of a fold line is the image in $\mathbb{R}^2$ under p_2 of an oriented normal to the image $p_1(F) \subset \mathbb{R}^3$; see Fig. 1.17.

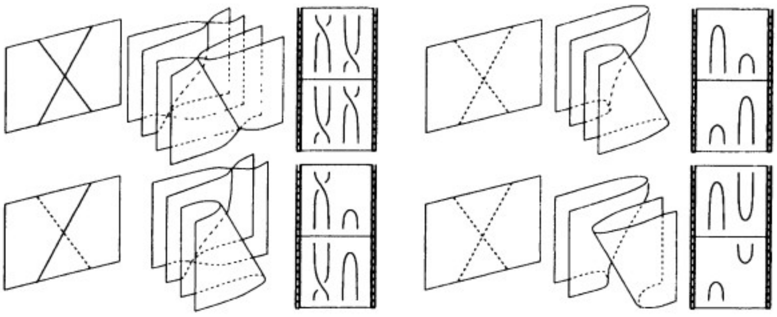

Fig. 1.16. Double intersection points of D^* and $\mathcal{F}^*$.

In summary, $D^* \cup \mathcal{F}^*$ forms a planar graph Γ with orientations for D^* and co-orientations for $\mathcal{F}^*$. The set of vertices V of such a graph Γ consists of T^*, B^*, C^*, DF^*, the double intersection points between D^* and $\mathcal{F}^*$, and double self-intersections between arcs in D^* or those in $\mathcal{F}^*$. See Fig. 1.16 for the latter double intersection points. The set E of edges are the closures of connected components of $(D^* \cup \mathcal{F}^*) \setminus V$. The types of vertices and incident edges are summarized in Table 1.1. Again, compare with Figs. 1.15 and 1.16.

Table 1.1. Vertices of a chart

Types of vertices	Valency	Types of incident edges
Cusp C^*	2	2 fold lines
Branch point B^*	3	2 fold lines, 1 double arc
Triple point T^*	6	6 double arcs
Tangential intersection of fold line and double curve DF^*	4	2 fold lines, 2 double arcs
Transverse intersection of fold line and double curve $D^* \cap \mathcal{F}^*$	4	2 fold lines, 2 double arcs
Self-intersection of double curves $D^* \cap D^*$	4	4 double arcs
Self-intersection of fold lines $\mathcal{F}^* \cap \mathcal{F}^*$	4	4 fold lines

Next, we define a label assigned to each edge of such graphs. A label (i, j) is a pair of non-negative integers. A line segment in $\mathbb{R}^3$ that is parallel to the direction p_2 intersects the surface $p_1(K)$ in a number of points. If we choose such a segment through a double point of $p_1(K)$, then we count the number of intersections of the line with the surface. The number, i, of such intersections between the double point and the plane into which the surface is projected is the first coordinate of the label on the projection of the double curve. The number, j, of points further from the plane of projection is the second coordinate of the label along the projection of the double curve. A point on the fold set is similarly labeled. This convention is slightly different from [CarSai98a] (the first and the second factors are switched). The difference in the conventions is not a problem since once a projection direction is chosen the plane can be a screen in front of or behind the surface.

Definition 1.24. A *chart* of a given knotted surface is the graph $G = (V, E)$ formed by $D^* \cup \mathcal{F}^*$, with the labels on edges, as described above.

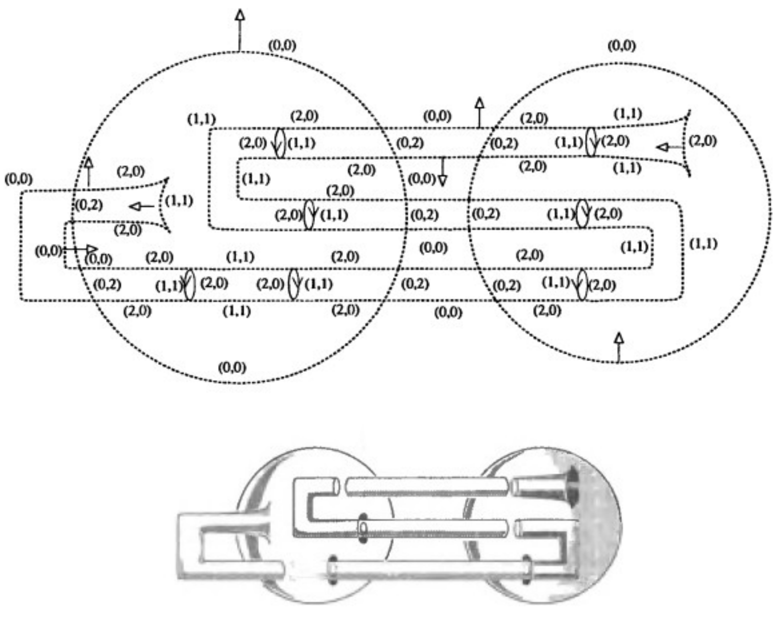

Fig. 1.17. The chart of the Fox-Milnor example

Example 1.25. A chart of the example given in Fig. 1.14 is given in Fig. 1.17.

The labels on the edges around a vertex are the same or distinct depending on the types of the vertex. The labels coincide at a double crossing of D^*, and at a branch point (on a fold line). Some labels are distinct at the rest of the vertices. At a cusp, the labels of the edges involved (fold lines) are $(i, j+1)$ and $(i+1, j)$. The six edges incident to a triple point have indices in cyclic order (i, j), $(i, j \pm 1)$, (i, j), $(i, j \pm 1)$, (i, j), and $(i, j \pm 1)$ (with all $(+)$s or all $(-)$s). At a crossing between $\mathcal{F}^*$, the labels are (i, j), (i', j'), $(i + 2, j)$, $(i', j' + 2)$, cyclically. At a crossing between D^* and $\mathcal{F}^*$, the label (i, j) on two fold lines is the same, and the labels on two edges in D^* are (i, j), $(i + 2, j)$ or (i, j), $(i, j + 2)$. At a vertex in DF^*, if the labels on the fold line are $(i, j + 1)$ and $(i + 1, j)$, then those for double arcs are $(i + 1, j)$ and $(i, j + 1)$. For the latter crossings, see Fig. 1.16.

For the rest of the section, we explain how to represent knotted surfaces as sequences of symbols (called elementary string interactions), using charts. In fact, this was a motivation for considering charts [CarSai98a].

A projection $p_3 : \mathbb{R}^3 \to \mathbb{R}$ can be chosen such that it restricts to a projection $p_3|_{\mathbb{R}^2} : \mathbb{R}^2 \to \mathbb{R}$ where $\mathbb{R}^2$ is the plane into which a chart is embedded (equivalently the plane into which the knotted surface K was projected) so that $p_3|_\Gamma$ is a generic Morse function with distinct critical values. More precisely, $p_3|_\Gamma$ satisfies the following conditions.

(i) The function $p_3|_\Gamma$ takes distinct values for vertices $V(\Gamma)$ of Γ.
(ii) Regard D^* and $\mathcal{F}^*$ as immersed 1-submanifolds. Then $p_3|_\Gamma$ has isolated non-degenerate critical points (local maxima/minima) on the interior of $(D^* \cup \mathcal{F}^*) \setminus V(\Gamma)$.
(iii) The above critical points have distinct critical values, that are also distinct from the values of $p_3(V(\Gamma))$.

Let t_i, $i = 1, \ldots, n$, be critical values of $p_3|_\Gamma$ described above, where we include $V(\Gamma)$ as the set of critical points. Let $[a_i, b_i]$ be small mutually disjoint intervals containing t_i, for $i = 1, \ldots, n$. Let $\mathbb{R}^2_t \subset \mathbb{R}^3$ denote the plane at height $t \in \mathbb{R}$, $\mathbb{R}^2_t = p_3^{-1}(t)$. The intersection $\mathbb{R}^2_t \cap \mathbb{R}^2$ (where the second $\mathbb{R}^2$ is where a chart lies) is a "horizontal" line, and the line passes through the critical points of $p_3|_\Gamma$ at $t = t_i$, $i = 1, \ldots, n$.

Let γ_t denote the immersed 1-manifold in $\mathbb{R}^2_t$ that is the intersection $p_1(K) \cap \mathbb{R}^2_t$. Then γ_{a_i} and γ_{b_i} differ by one of the small changes corresponding to the critical points of $p_3|_\Gamma$, that are called *elementary string interactions* (*ESIs*). These changes can be seen in Fig. 1.15 by looking at the top and bottom cross sections of $p_1|_K$ (right of each arrow representing p_2), and in Fig. 1.16 the changes are depicted in the form of movies. In each case, a height function on $\mathbb{R}^2_t$ is fixed for each t, in the direction parallel to the plane $\mathbb{R}^2$ in which the chart lies and perpendicular to the height function p_3. With respect to this height function, the curves γ_t have maxima and minima at fold lines. The ESIs are described in summary in Table 1.2.

The height function on the plane of the chart $p_3|_{\mathbb{R}^2}$ can be used to decompose a movie description into smaller pieces. A height function is fixed on each $\mathbb{R}^2_t$ as above, and the circles γ_t are given crossing information from a broken surface diagram of $p_1(K)$. In other words, except for $t = t_i$ $(i = 1, \cdots, n)$, $\mathbb{R}^2_t \cap p_1(K')$ is a classical link diagram γ'_t, where K' is K with a neighborhood of the lower decker set removed. The underlying immersed circles of γ'_t are γ_t.

Table 1.2. Elementary string intersections

Types of critical points	ESIs
Cusp C^*	Cancelation of local max/min
Branch point B^*	Reidemeister type I move
Triple point T^*	Reidemeister type III move
Tangential intersection of fold line and double curve DF^*	An arc going over (under) a local max (min)
Transverse intersection of fold line and double curve $D^* \cap \mathcal{F}^*$	Interchange of heights of a max or min and a crossing
Self-intersection of double curves $D^* \cap D^*$	Interchange of heights of distant crossings
Self-intersection of fold lines $\mathcal{F}^* \cap \mathcal{F}^*$	Interchange of heights of distant max and/or min
Minimum of $\mathcal{F}^*$	Birth of a circle, or a saddle
Maximum of $\mathcal{F}^*$	Death of a circle, or a saddle
Max/min of D^*	Reidemeister type II move

Right before/after a critical point $t = t_i$, the links γ'_{a_i} and γ'_{b_i} differ by an ESI.

Then each classical link diagram is represented symbolically, by using the symbols $\cup_{m,n}$, $\cap_{m,n}$, $X_{m,n}$, and $\overline{X}_{m,n}$ for a local maximum, a minimum, a positive crossing, and a negative crossing, respectively, where subscripts are the corresponding labels of the chart. Each classical link diagram is represented by a word in these symbols, and a knotted surface is represented by a sequence of words, where differences between consecutive words differ only by one of the ESIs.

The advantage to using such a movie is that no information is lost from still to still, such as rotations that we failed to describe in Remark 1.3 on Fig. 1.2. In some cases, it is necessary to explicitly mention which ESI connects two words. Specifically, there are pairs of words that can either be connected by saddles or births/deaths; in this case we must denote the ESI between them. In [CarSai98a] and [CRS97] this type of a combinatorial description of a knotted surface was given, and a complete set of moves among them that realizes equivalence was presented. Using such a combinatorial description, a 2-categorical description of knotted surfaces was obtained in [BL98].

Example 1.26. For a choice of a height function, a movie for the Fox-Milnor example is illustrated in Fig. 1.18, where changes between stills correspond to ESIs. The top still in the second film strip from the left in Fig. 1.18, for example, gives rise to a word

$$\cap_{0,0} \cap_{1,1} \overline{X}_{2,0} X_{2,0} \cup_{1,1} \cup_{0,0} \cap_{0,0} \cup_{0,0},$$

and the still below it corresponds to the word

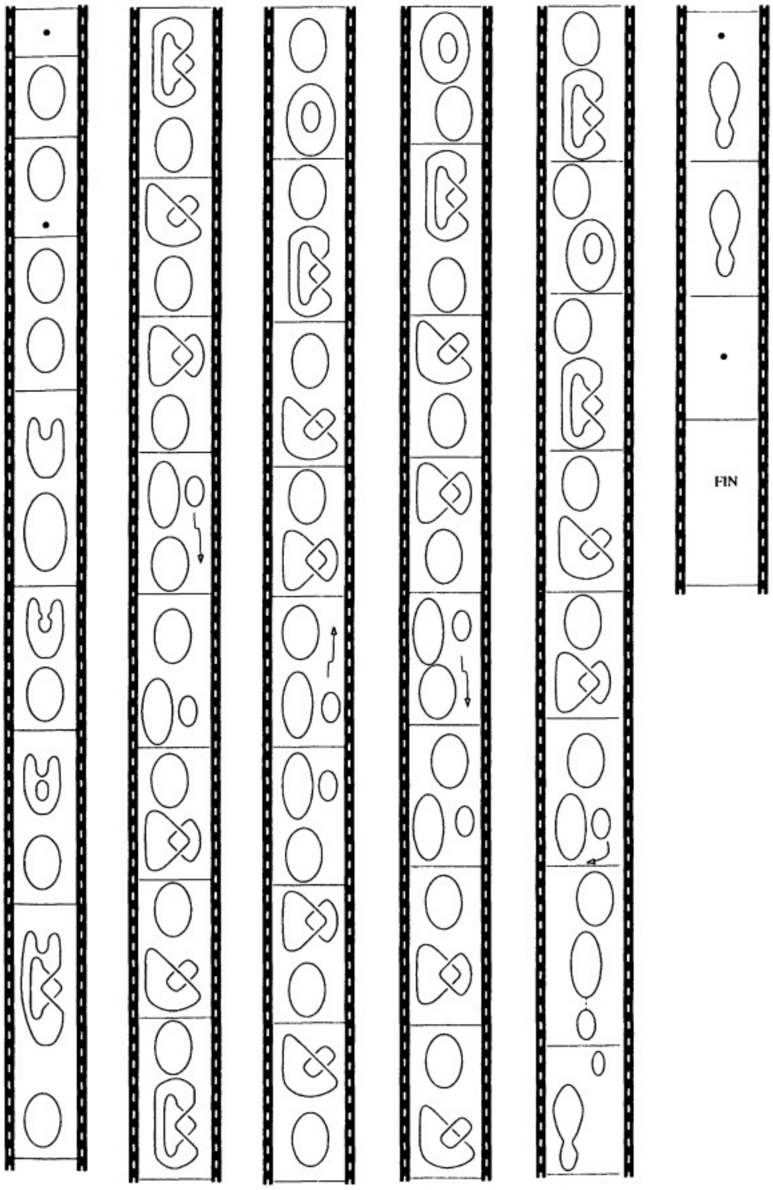

Fig. 1.18. A slow movie of the Fox-Milnor example

$$\cap_{0,0} \cap_{2.0} X_{1,1} X_{2,0} \cup_{1,1} \cup_{0,0} \cap_{0,0} \cup_{0,0}.$$

The subscripts correspond to the chart labels, and for example, $X_{1,1}$ means that there is one arc to the left and right of the crossing. The change from one of the above two words to the other corresponds to the tangential intersection of fold line and double curve, depicted in Fig. 1.15 top right.

1.6 Braids and Braid Charts

Let B and D be 2-disks. Fix a set of m interior points, $X_m = \{z_1, z_2, \ldots, z_m\}$, of B.

Definition 1.27. (cf. [CarSai94, Kama92a, Rud83, Viro90]) A *braided surface* (of degree m) is a properly embedded surface S in $B \times D$ such that

- the natural map $\pi : S \subset B \times D \to D$ is a branched covering map of D of degree m,
- ∂S is a closed braid in the solid torus $B \times \partial D$.

A *surface braid* or a *2-dimensional braid* is a braided surface such that

- ∂S is a trivial closed braid $X_m \times \partial D$ in the solid torus $B \times \partial D$.

The map $\pi : S \to D$ is called the *associated branched covering map* of S. A surface braid is *simple* if the associated branched covering is simple, i.e., $\#(\pi^{-1}(y)) = m$ or $m - 1$ for each $y \in D$.

Let S be a surface braid of degree m. Fix a bi-parametrization $I_3 \times I_4 \cong D$, where I_3 and I_4 are the unit intervals. For each $t \in I_4 = [0, 1]$, let b_t be a subset of the cylinder $B \times I_3$ such that the restriction of S to the cylinder $B \times I_3 \times \{t\} \subset B \times I_3 \times I_4 \cong B \times D$ is $b_t \times \{t\}$. The one-parameter family $\{b_t\}_{t \in [0,1]}$ is called a *motion picture* of the surface braid S.

For all t but a finite number of exceptional values, $b_t \subset B \times I_3$ is an m-braid such that $\partial b_t = X_m \times \partial I_3$. Moreover, each of b_0 and b_1 is the trivial braid $X_m \times I_3 \subset B \times I_3$. For an exceptional value t, b_t is a *singular braid* which is similar to a braid except there are intersection points among strings. Each intersection is called a *singular point* of b_t or of S, which corresponds to a singular point of the associated branched covering $S \to D$. If S is a simple surface braid, then each singular point is an intersection of two strings of b_t for some $t \in I_4$, and there are no other singular points of b_t on the same level of I_3. Conversely, if there is at least one singular point in the 2-disk $B \times (s, t)$ for each $(s, t) \in I_3 \times I_4 \cong D$ and it is the intersection of two strings of b_t, then S is simple. Deforming the bi-parametrization slightly, we may assume that there is a single singular point in b_t for each exceptional value $t \in I_4$. We sometimes assume this condition but not always. When we assume this condition and when S is simple, we call a motion picture $\{b_t\}$ a *braid movie*. See Figure 1.19. The motion picture presents a simple surface braid of degree

Inocero.com.br 6 14 Gb
46
Bans\(04_02)\ Geometry - Gauss\ 660

neologismo Stagflation → 1965
Iain Macleod
(dos tóries)

Inflação e desemprego → Curva de Phillips.

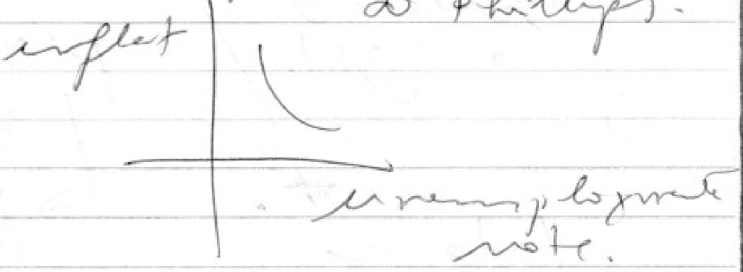

Portman

hawkish "Barber Boom"
dovish England

---

5 Forças fundamentais: eletromagnetismo
gravidade
força nuclear forte
força nuclear fraca
força de Higgs.

1 Higgs pesa ≅ átomo de iodo
↳ não tem spin.

As outras partículas ~~Não~~ têm spin mas tem um giro intrínseco.

obs: Higgs não tem giro ou spin.

Higgs → composição (vários partículas fund)
ou supersimetria

(solução para entender o Higgs).

píon → partícula sem spin.
→ prótons nêutrons

toda a partícula tem um par.

—

Universo tem dois tipos de partículas
{ matéria → { quarks → P N
{ força { léptons
→ { elétron
{ neutrino

—

[Quark decai em proporções diferentes
Para cada 100 decaimentos com elétrons houve 85 decaimentos com múons.

→ Quebra da simetria cosmica

eletrodinâmica quântica

Campo magnético intrínseco é medido em unidades típicas para partículas elementares.

$(g)_{experimental} = 2.00115965218073$

$(g)_{teórica} = 2.00115965218178$

Efeito Schwinger

Um dos motivos é a analogia do eletromagnetismo com a gravidade.

Na vizinhança de um buraco negro, pares de partículas são geradas para neutralizar o forte campo gravitacional do buraco negro

→ "efeito Hawking"

(%

2025 05 26

Hoje, às 07:53 h, sonhei com a minha mãe e ela me disse para eu não desistir. Ela me deu muitos beijos!!!

Adorei!!! Amei!!!

Obs: Quando acordei deste sonho maravilhoso, o meu pai chegou em casa e eu contei o meu sonho para ele.

mathematics.cloud (70)

ArrayPlot[CellularAutomaton
[254, {{1}, 0}, 40], Mesh→True]
↓
go

Joe Schmoe

~~@cOokiemon5ter~~

578
65
643
95
738
180
918
80
998

178
285
150
313
65
378
200
578

Books:

gallica.bnf.fr/ark:/12148/
bpt6k64545935

→ Exercices de géométrie
bpt6k6336853h

Cours de mathématiques élémentaires
Exercices de Géométrie

/bpt6k935782m
Géométrie Analytique

/bpt6k9376884

/bpt6k4226025h
Éléments de Géométrie
F. J. J.

$\sqrt{2}\ 360^{\circ}$ = pi radian.

Transient

$(360\sqrt{2})/((332e!)/39+23/3-$
$3907/(52e)+(8365e)/156)=\pm\pi$.

Toom's rule

'

! *+,-* * *. ! * '
D# D ! L ! N DL L ! D<J DO PD87 G "" %@" 7#

8 *+,-* * *. ! * '
< L D# D L L D 7 ! D# DO P9D0D87 G' @ 7#

+,- * *. ! * '
<J D# D L L DD87 G '@'@)"@ 7#

+,- * *. ! * '
L D#M! Q8 D#M! ! 3 D D# DO P !7 D2 D87 G ' @"" 7#

+,- * *. ! * '
L D# D L L D! ! 3 N DRDO P !7 D2 D87 G)"@ % 7#

(%

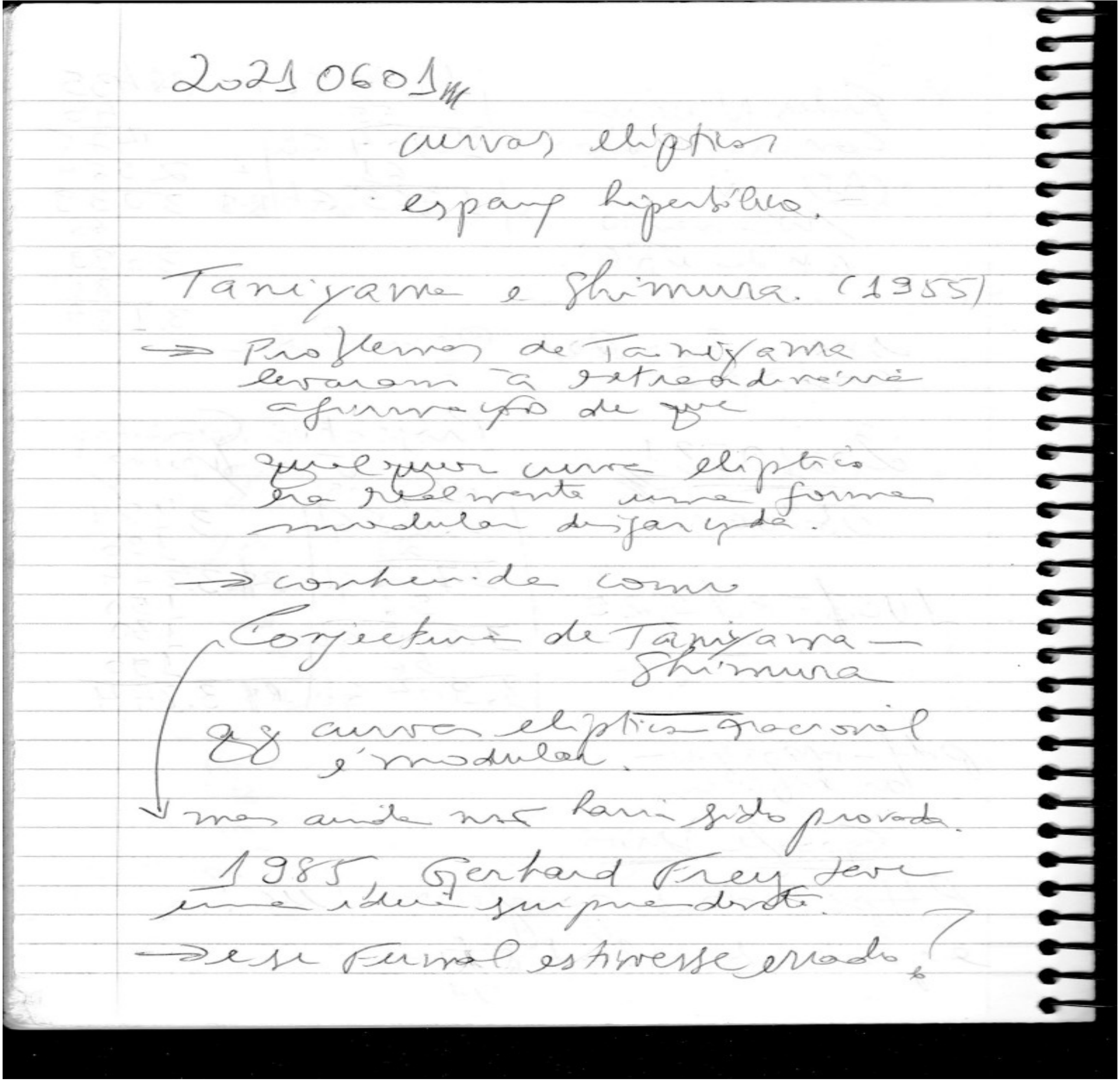

2021 0601

curvas elípticas

espaço hiperbólico.

Taniyama e Shimura. (1955)

→ Professores de Taniyama levaram à extraordinária afirmação de que

qualquer curva elíptica era realmente uma forma modular disfarçada.

→ conhecida como

Conjectura de Taniyama—Shimura

as curvas elípticas racional é modular.

mas ainda não havia sido provada.

1985, Gerhard Frey teve uma ideia surpreendente.

→ e se Fermat estivesse errado?

20210601 III

curvas elípticas

espaço hiperbólico.

Taniyama e Shimura. (1955)

⇒ Problemas de Taniyama levaram à extraordinária afirmação de que

qualquer curva elíptica era realmente uma forma modular disfarçada.

→ conhecida como

Conjectura de Taniyama — Shimura

as curvas elípticas racional é modular.

mas ainda não havia sido provada.

1985, Gerhard Frey teve uma ideia surpreendente.

→ e se Fermat estivesse errado?

$x^n + y^n = z^n$

→ se Fermat é falso Shimura-Taniyama também o é!

→ ele não provou, mas deu um argumento de plausibilidade.

Frey → conjectura épsilon.

Barry Mazur deu uma sugestão para envolver "provar completamente a conjectura de épsilon de Serre".
(20m23seg)

→ Barry Mazur sugeriu para Prof Ken Ribet

—

O truque de Andrew era transformar as curvas elípticas em algo chamado Representação de Galois,

Iwasawa theory → Fórmula do número de classes.
(26m23seg)

John Coates
→ supervisor
Andrew Wiles.

—

Utaka Taniyama
1958 died

Class Number Formula

Coates falou para Wiles → (91)
de um paper de Matthias Flach
↳ conseguiu uma
fórmula Class Number

Flach usando ideias de
Kolyvagin. (27m05seg)

Jan 1993 ⇒ Andrew conversou com
Nick Katz.

33m28seg - Wednesday June 23 1993
Elliptic Curves and
Galois Representations

Correção da Prova;
Erro sobre o Método de Flach
e Kolyvagin
(37m21seg)
Convidou Richard Taylor

Revelation → 41 min 14 seg.

É o mesmo problema que tinha tido na minha Teoria de Iwasawa, três anos antes.

=

Math envolvida:

Klein, Fricke, Hurwitz,
Hecke, Dirichlet, Dedekind,
Prova de Langland e Tunnell
Deligne, Rapoport, Katz
A Idéia de Mazur de usar a teoria de deformação das representações de Galois
Igusa, Eichler, Shimura,
Taniyama,
A redução de Frey
Bloch, Kato, Selmer, Frey,
Fermat.

( )

Kepler odiava das aulas
e uma das suas paixões era
a Geometria, que ele dizia existir
antes da Criação.

Kepler foi seu assistente de
Tycho Brahe (Praga)

3ª Lei Kepler:

"Os quadrados dos períodos
de translação dos planetas são
proporcionais aos cubos dos
semi-eixos maiores de suas
órbitas".

Kepler quase descobriu a Lei da
Gravidade, mas manteve a
força com o nome de
"magnetismo".

Kepler - 1571 - 1630
Newton - 1643 - 1727

Kepler na ótica determinou que a
intensidade de uma onda de luz diminui
com o quadrado da distância,
Também na neurofisiologia e foi
o primeiro a descobrir que as imagens
no olho humano são projetadas na
retina de cabeça para baixo.

( )

20210607

109,742,02 GB
113.704,51 GB (111,04 TB).

Já que o Direito não se endireita.
Eu ando direito.

—

wide awak → corpo largo

wide awake → totalmente acordado

—

linkedin.com/in/amanda-aldous-cd-adec-mds-mpa-pcsc-73736632

—

TIM
05/08/2021

—

⇒ ATOZ

( ))

20210610 Blog-Pastas com files

Matte1matics2@gmail.com

1-Website-Blog/Leibniz/ (19w)

2020/10/20201012/Leibniz-Werke-04-Volumes/

(16Qn)

20201031/Die Werke von Leibniz-Hannover

2020/11/Correspondance-de-Leibniz (1m0)

- 11/Leibniz-09-Volumes-Hannover (1v7w)
- 11/Leibniz-Google-16-Volumes (183B)

11/Leibniz-und-Landgraf (1nV0)

Obs:

~~2020/10/~~[illegible]

[illegible]/ 11/20201104-Leibniz-gwlb.de (15g_z14)

2021/01/Leibniz-Die-philosophischen (1vaS)
01/Leibniz-Oeuvres-7-Volumes (1N9d)
01/Leibniz-Opera-Omnia-Latin (15-BL)
01/Leibniz-Sämtliche _1923 (1ggB)

↘2021/02/ Gauss - Werke 12 - Volumes
(1 Dy 33)

—

Seqüência Blog (materiais)

De baixo para cima:

1) Leibniz's Works → October 12, 2020
(19WQ)
2) Die Werke von Leibniz - Hannover
→ October 31, 2020
(16Qn)
3) Leibniz und Landgraf von Hessen -
⇒ October 31, 2020
(1nVO)
4) Leibniz - Works - Google
→ Nov 4, 2020
(183B)
5) Correspondance de Leibniz avec l'électrice
↳ Nov 4, 2020
(1m0)
6) Leibniz's Works - Oeuvres de Leibniz
7 volumes French
→ Jan 17, 2021
(1N92)
7) Leibniz - Opera omnia - Latin - 1768
⇒ Jan 17, 2021
(15-BL)
8) Leibniz - Die philosophischen.. 1875
⇒ Jan 17, 2021
(1va5j)
9) Leibniz - Sämtliche Schriften
⇒ Jan 17, 2021
(1g9B)

10) Carl Friedrich Gauss Werke
→ Feb 25 2021
((1Dy33)

OK até aqui:

Atualização: (updated)

11) Leibniz-09-Volumes-Hannover
→ Jun 10 2021
(1V7W)

—

12) 20201104-Leibniz-gwlb.de
→ Jun 10 2021
(15g-z14)

20210611||| 70

130

The Earth Robber 200

Super Interplanetary Menace 230

Rendent The Cosmic Gladiators 430

—

germanroutes.com.br

—

MARIO SANTOS CPF 093913527-23

—

Protocolo 55.840.

10.738

430

20210612||| — R$ 11.168

100

= R$ 11.268

No Japão, o político corrupto se mata,
na China, o corrupto é executado,
na Itália, o corrupto é preso e
no Brasil o corrupto é solto e concorre
às eleições.
Ex: LULA.

—

Chapter 8

Volume-08---20211207

xn-matemhica-41a.org

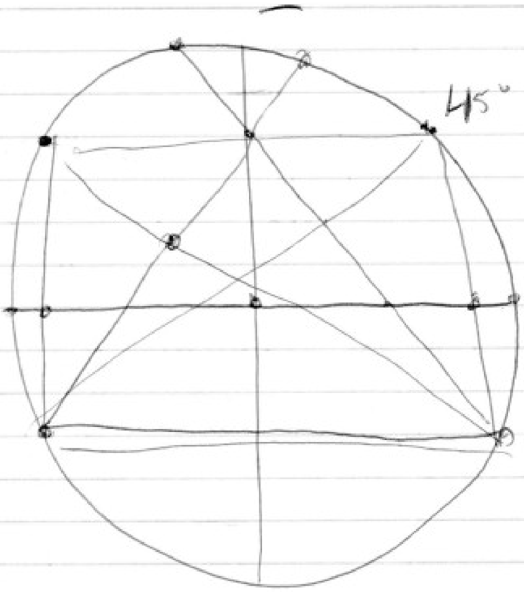

Pappus

Pappus La Collection Mathématique
Paul Ver Eecke

Fundamentals of Geometry Coxeter
(Fundamentos)

' $
! *+,-* * *. ! * / & " 0
1 0)) 2"

3 *+,-* * *. ! * / 4 5 " ! 0
! ! ! + ! + !)) 2"

+,- * *. ! * 6 0 , 7 " 8
! ! ! + ! + !)9 2"

" *+,-* * *. ! * :!7 8 6 0 ! " 8 2 0
/ 6 2 ; 7 0 2"

+,- * *. ! * / < = "! :
6 ! > ! 1 0) 2"

+,- * *. ! * / 8 , ! ? 7 !
1 ! +22 " / ! ! !@ 30 6 A B ! 2 ! 4
C B))$ 2"

+,- * *. ! * = " > 8 DEF E7
3 ! " ! 3 ! 7 ! 2 ? 7 2 ! ,) 2"

+,- * *. ! * = ! G +2E 0 " !
2 ? 7 2 ! 2 !2 ! " A 0 ! = " H !
) 2"

.3

+,- * *. ! * / < = "! :
6 ! > ! 1 0) 2"
' %($

Si aplicamos la misma idea al número complejo

$$z = \mu(\cos 1 + i \operatorname{sen} 1),$$

de módulo $\mu \neq 1$, vemos que sus potencias

$$z^\theta = \mu^\theta (\cos \theta + i \operatorname{sen} \theta)$$

quedan representadas por un punto que describe la espiral equiangular $r = \mu^\theta$ (§ 8.6).

EJERCICIO

Por medio del teorema de de Moivre, encuéntrense
(*a*) las dos raíces cuadradas de $3 + 4i$;
(*b*) las tres raíces cúbicas de 1, denotadas por 1; ω, ω^2;
(*c*) las seis raíces sextas de 1 (con la notación de ω);
(*d*) las doce raíces duodécimas de 1.

$\sqrt[n]{z} = \rho^{1/n}$ cis $\frac{\theta + 2k\pi}{n}$

$\sqrt{3+4i} = 5^{1/2}$ cis

9.5 LA FORMULA $e^{\pi i} + 1 = 0$

> *Hay una fórmula famosa –tal vez la más compacta y famosa de todas las fórmulas– que desarrolló Euler a partir de un descubrimiento de de Moivre:* $e^{i\pi} + 1 = 0$. . . . *Es igualmente atractiva para el místico, el científico, el filósofo y el matemático. Cada uno de ellos le encuentra un significado propio. Aunque ya llevaba más de un siglo de ser conocida, la fórmula de de Moivre llegó a Benjamin Pierce como una revelación. Al descubrirla un día se volvió a sus estudiantes. . . . "Señores," les dijo, "esto es cierto con seguridad, es absolutamente paradójico; no podemos entenderlo y no sabemos lo que significa, pero lo hemos demostrado y, por lo tanto, debe ser verdad."*
>
> E. Kasner y J. Newman
>
> [1, págs. 103-104]

Puesto que hemos aceptado potencias *reales*, como lo es $\mu^\theta = e^{\theta \log \mu}$, nos preguntamos con naturalidad el significado que se puede dar a las potencias *complejas*. Una respuesta parcial proviene de la teoría de series infinitas como

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \ldots,$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \ldots,$$

$$\operatorname{sen} x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots.$$

La primera serie, que es familiar cuando x es real, nos permite *definir* e^x en otros casos; por ejemplo

$$\begin{aligned} e^i &= 1 + i + \frac{i^2}{2!} + \frac{i^3}{3!} + \frac{i^4}{4!} + \frac{i^5}{5!} + \dots \\ &= 1 + i - \frac{1}{2!} - \frac{i}{3!} + \frac{1}{4!} + \frac{i}{5!} - \dots \\ &= \left(1 - \frac{1}{2!} + \frac{1}{4!} - \dots\right) + i\left(1 - \frac{1}{3!} + \frac{1}{5!} - \dots\right) \\ &= \cos 1 + i \operatorname{sen} 1. \end{aligned}$$

Estas series nos permiten expresar 9.41 en la forma concisa de Euler

9.51 $$\cos\theta + i \operatorname{sen}\theta = e^{\theta i},$$

que es un refinamiento de la identidad

$$\theta i = \log_e(\cos\theta + i \operatorname{sen}\theta)$$

de R. Cotes* (1682–1716), de quien dijo Newton después de su muerte prematura, "Si Cotes hubiera vivido, ¡tal vez hubiéramos llegado a saber algo! "

Al llevar $\theta = \pi$ a 9.51, obtenemos la "fórmula famosa"

$$e^{\pi i} = -1,$$

que relaciona de manera tan sorprendente los tres números importantes

$$e = 2.71828\dots, \qquad \pi = 3.14159\dots,$$

e i.

EJERCICIOS

1. Evalúese $e^{i\pi i}$. ¿Es i^i real?
2. Dedúzcase de 9.51 $\cos\theta = \dfrac{e^{\theta i} + e^{-\theta i}}{2}$, $\operatorname{sen}\theta = \dfrac{e^{\theta i} - e^{-\theta i}}{2i}$,

que son las conocidas fórmulas de $\cos(\theta + \alpha)$, $\operatorname{sen}(\theta + \alpha)$, y las derivadas de $\cos\theta$, $\operatorname{sen}\theta$.

9.6 RAICES DE ECUACIONES

> *Gauss . . . fue el primer matemático que usó los números complejos de manera realmente científica y segura.*
>
> G. H. Hardy
>
> [Hardy y Wright 1, pág. 188]

En el campo de los números complejos, podemos resolver cualquier ecuación cuadrática de coeficientes reales; por ejemplo, la ecuación 9.31 tiene las dos raíces i y $-i$. Es aún más notable que podamos resolver cualquier ecuación de coeficientes complejos; y no solamente una ecuación cuadrática, sino también una ecuación cualquiera de tercer o cuarto grado. Cuando hablamos de "resolver" una ecuación queremos decir que podemos encontrar expresiones explícitas de sus raíces en términos de sus coeficientes. E. Galois (que fue asesinado en 1832 cuando solamente tenía 20 años) demostró que la ecuación algebraica general 9.21 cuando $n > 4$ no se puede resolver en este sentido [Infeld 1]. A pesar de ello, el teorema fundamental del álgebra (que Gauss demostró en 1799) afirma la *existencia* de raíces para todos los valores de n, incluso cuando no dispongamos de expresiones explícitas. (Una bonita y pulcra demostración se encuentra en Birkhoff y MacLane [1, págs. 101-103].) De hecho, se pueden encontrar soluciones *numéricas* correctas hasta un número deseado de cifras decimales.

---

* *Harmonia mensurarum*, Cambridge, 1722, pág. 28.

EJERCICIO

Una escalera de mano que mide 24 metros de largo descansa en una fachada con el apoyo adicional de un cajón cúbico cuya arista mide 7 metros, que se ha colocado al pie de la fachada con una de sus aristas horizontales en contacto con la escalera. ¿A qué altura sube la escalera sobre la fachada? (*Indicación*: tómese $7x$ como la altura desde el extremo superior de la escalera a la tapa del cajón. Obténgase una ecuación cuya raíz relevante sea $7x = 9 + 4\sqrt{2}$.)

9.7 TRANSFORMACIONES CONFORMES

Vimos, en § 9.3, que la transformación

$$z' = z + b$$

(que suma a la variable compleja z la constante compleja b) es una traslación, por lo que

$$z' = az$$

(que multiplica z por la constante compleja a) ha de ser una rotación dilatativa alrededor del punto O, donde incluimos como casos especiales una dilatación (cuando a es real) y una rotación (cuando $|a| = 1$). En consecuencia, la rotación dilatativa alrededor del punto general c es

$$z' - c = a(z - c)$$

ó

$$z' = az + (1 - a)c.$$

Así, pues, la semejanza directa general, como se describió en § 5.5, es la transformación *lineal* general

$$z' = az + b$$

[Ford 1, pág. 3]; y tenemos aquí una traslación o una rotación dilatativa según sea $a = 1$ ó $a \neq 1$. (En el último caso, $c = b/(1 - a)$.)

Como el producto de una semejanza opuesta y una reflexión es directo, una semejanza opuesta cualquiera se puede expresar como producto de una

2. Estos 12 puntos dividen las 12 aristas del octaedro

$$(\pm\tau^2, 0, 0), \quad (0, \pm\tau^2, 0), \quad (0, 0, \pm\tau^2)$$

en la razón de $\tau : 1$.

3. Obténganse las coordenadas de los 20 vértices del dodecaedro regular [Coxeter 1, pág. 53].

11.3 LA ESPIRAL AUREA

> *Arquímedes, Leonardo, Newton: todos ellos hombres con mucho sentido práctico, pero con algo más. Una vocación a la pregunta, tal vez . . . Muy misterioso todo esto, muy misterioso. Y muy emocionante. Esto es, la emoción: se obtiene al ser Dios mientras se construye, y se obtiene al observar a Dios, es decir, al observar las cosas tal como son.*
>
> J. L. Synge [2, pág. 163]

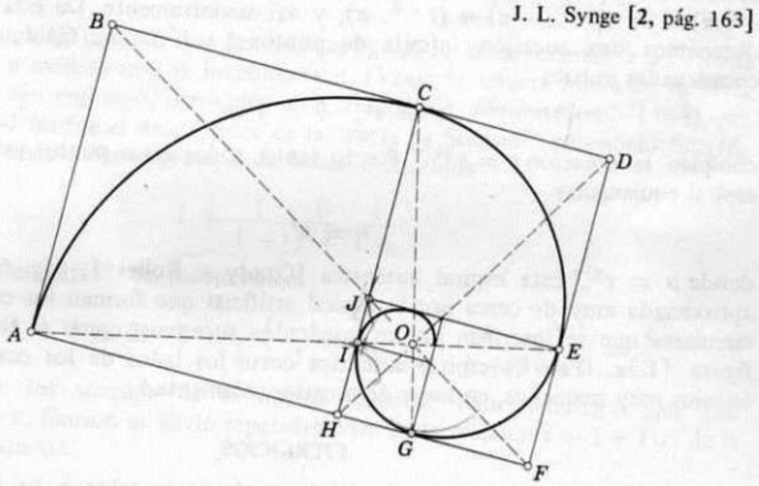

Figura 11.3a

La relación $\tau = 1 + 1/\tau$ nos señala que el rectángulo áureo $ABDF$ (figura 11.3*a*) se puede descomponer en dos piezas: el cuadrado $ABCH$ y un rectángulo áureo menor. (Se pueden identificar los puntos B, C, D con los Q, U, S de la figura 11.1*b*.) Del rectángulo pequeño $CDFH$ podemos cortar nuevamente otro cuadrado, de modo que dejemos un rectángulo aún menor, y continuar el proceso indefinidamente.

Puesto que el rectángulo $HJEF$ es homotético a $ABDF$, el vértice J del primero queda en la diagonal BF del segundo. De hecho, las rectas AE, BF, CG, DH contienen entre todas los vértices de todos los rectángulos. Para derivar $CDFH$ a partir de $ABDF$ podemos emplear una rotación dilatativa de centro O, el punto de intersección $BF \cdot DH$. Esta semejanza, que transforma cada uno de los puntos $A, C, E, G, I, \ldots$ en el siguiente, y cada uno de los puntos $B, D, F, H, J, \ldots$ en el siguiente, es el producto de un cuarto de giro negativo alrededor de O y la dilatación $O(\tau^{-1})$. Por lo tanto, DH es perpendicular a BF. La semejanza inversa, que transforma cada rectángulo en el siguiente mayor que él, es producto de un cuarto de giro positivo alrededor de O y la dilatación $O(\tau)$.

Puesto que

$$\frac{OB}{OD} = \tau = \frac{BC}{CD},$$

OC biseca el ángulo (recto) BOD. Así tenemos que las rectas CG, AE pasan por O y bisecan los ángulos entre BF, DH.

En términos de coordenadas polares con polo en O, la rotación dilatativa que transforma OE en OC transforma cualquier punto (τ, θ) en $(\tau r, \theta + \frac{1}{2}\pi)$. Si tomamos como recta inicial OE y la distancia OE como la unidad de medida, de manera que E sea $(1, 0)$, deducimos las coordenadas $(\tau, \frac{1}{2}\pi)$ de C,(τ^2, π) de A, $(\tau^3, \frac{3}{2}\pi)$ del vértice opuesto a A en el nuevo cuadrado que se colocaría sobre AF, y así sucesivamente. De la misma manera, G es $(\tau^{-1}, -\frac{1}{2}\pi)$, I es $(\tau^{-2}, -\pi)$ o (τ^{-2}, π), y así sucesivamente. De esta manera, obtenemos una sucesión infinita de puntos $\ldots, I, G, E, C, A, \ldots$ cuyas coordenadas polares

$$r = \tau^n, \quad \theta = \tfrac{1}{2}\pi n$$

cumplen la ecuación $r = \tau^{2\theta/\pi}$. Por lo tanto, todos estos puntos están en la espiral equiangular

$$r = \mu^\theta,$$

donde $\mu = \tau^{2/\pi}$ Esta espiral auténtica [Cundy y Rollet **1**, pág. 64] se ve aproximada muy de cerca por la espiral artificial que forman los cuadrantes circulares que se inscriben en los cuadrados sucesivos, como se tiene en la figura 11.3*a*. (Pero la espiral auténtica corta los lados de los cuadrados a ángulos muy pequeños, en lugar de tocarlos solamente.)

EJERCICIOS

1. Verifíquese la conclusión de § 5.4 cuando se la aplica a los rectángulos semejantes $ABDF$ y $CDFH$ de la figura 11.3*a*.
2. Obténganse coordenadas polares para los puntos J, H, F, D, B. Se encuentran en otra espiral equiangular que es congruente (¿mediante qué ángulo de rotación?) y homotética (¿mediante qué razón de amplificación?) con respecto a la espiral $IGECA$.

11.4 LOS NUMEROS DE FIBONACCI

> *Sint minim 1 et 1 quos imaginaberis inaequales. Adde, fient 2. cui adde maiorem 1 fient 3. cui adde 2 fient 5. cui adde 3 fient 8. cui adde 5 fient 13. cui adde 8 fient 21.*
>
> J. Kepler [**1**, pág. 270]

En 1202, Leonardo de Pisa, apodado Fibonacci ("hijo de buen carácter"), se encontró con su célebre sucesión de enteros f_n en relación con la cría de conejos.* Supuso que los conejos vivían para siempre y que cada mes una pareja procrea una nueva pareja que se vuelve productiva a la edad de dos meses. En el primer mes el experimento comienza con una pareja de recién nacidos. En el segundo mes, todavía hay una sola pareja. En el tercer mes, hay 2; en el cuarto, 3; en el quinto, 5; y así sucesivamente. Sea f_n el número de parejas de conejos en el enésimo mes. Los primeros valores (y sus razones sucesivas) se pueden tabular de la manera siguiente:

n:	0	1	2	3	4	5	6	7	8	9	10	. . .
f_n:	0	1	1	2	3	5	8	13	21	34	55	. . .
f_{n+1}/f_n:	∞	1	2	1.5	$1.\dot{6}$	1.6	1.625	1.6154	1.6190	1.6176	1.6182	. . .

Cuatro siglos después, Kepler afirmaba explícitamente lo que Fibonacci seguramente advirtió: que la suma de dos f cualquiera es igual a la siguiente, de manera que la sucesión queda determinada por la fórmula de recursión

11.41 $$f_0 = 0, \quad f_1 = 1, \quad f_k + f_{k+1} = f_{k+2}.$$

Kepler también observó que las razones $f_n : f_{n+1}$ se aproximan a $1 : \tau$ cada vez más, a medida que se incrementa n. (Véase la tercera hilera en la tabla anterior.) Sin embargo, otros cien años después, R. Simson (1687–1768), que en realidad no fue el descubridor de la "recta de Simson") reconoció f_{n+1}/f_n como el enésimo convergente de la fracción continua

$$1 + \frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{1+}\cdots.$$

(Los "convergentes" son los números

$$1, \quad 1 + 1 = 2, \quad 1 + \frac{1}{1+1} = \frac{3}{2}, \quad 1 + \frac{1}{1+}\,\frac{1}{1+1} = \frac{5}{3},$$

y los que los suceden de la misma manera.) Para demostrar que lím $(f_{n+1}/f_n = \tau$, Simson se sirvió repetidamente de la relación $\tau = 1 + 1/\tau$, de la manera siguiente:

$$\tau = 1 + \frac{1}{\tau} = 1 + \frac{1}{1+}\,\frac{1}{\tau} = 1 + \frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{\tau} = \cdots$$

Con respecto a f_n, E. Lucas (1842–1891) tomó en consideración los números relacionados g_n, que se definen por medio de

$$g_0 = 2, \quad g_1 = 1, \quad g_k + g_{k+1} = g_{k+2}.$$

Los primeros valores que obtuvo eran los siguientes:

n:	0	1	2	3	4	5	6	7	8	9	10	. . .
g_n:	2	1	3	4	7	11	18	29	47	76	123	. . .

Se demuestra facilmente por inducción que (cuando $n > 0$)

11.42 $$g_n = f_{n-1} + f_{n+1}.$$

---

* R. C. Archibald, Golden Section, *American Mathematical Monthly*, 25 (1918), págs. 232–238.

Dejemos esto como ejercicio para el lector y prosigamos para establecer las identidades de Lucas

11.43
$$f_{2n} = f_n\, g_n,$$

11.44
$$f_{2n+1} = f_n^2 + f_{n+1}^2,$$

que son obvias cuando $n = 0$ o 1. Para demostrar ambas por inducción, sumaremos las relaciones tentativas

$$f_{2k-1} = f_{k-1}^2 + f_k^2 \quad \text{y} \quad f_{2k} = f_k g_k,$$

con lo que obtenemos
$$\begin{aligned} f_{2k+1} &= f_{k-1}^2 + f_k(f_{k-1} + f_{k+1}) + f_k^2 \\ &= (f_{k-1} + f_k)f_{k+1} + f_k^2 \\ &= f_k^2 + f_{k+1}^2, \end{aligned}$$

y entonces sumamos

$$f_{2k} = f_k g_k \quad \text{y} \quad f_{2k+1} = f_k^2 + f_{k+1}^2\,,$$

con lo que obtenemos
$$\begin{aligned} f_{2k+2} &= f_k^2 + f_k(f_{k-1} + f_{k+1}) + f_{k+1}^2 \\ &= f_k f_{k+1} + f_{k+1} f_{k+2} \\ &= f_{k+1}\, g_{k+1}. \end{aligned}$$

De la misma manera, para establecer la identidad

11.45
$$\tau^n = f_n\tau + f_{n-1},$$

que es obvia cuando $n = 1$ ó 2, sumamos
$$\tau^k = f_k\tau + f_{k-1} \quad \text{y} \quad \tau^{k+1} = f_{k+1}\tau + f_k,$$
de modo que obtenemos
$$\tau^{k+2} = f_{k+2}\tau + f_{k+1}.$$

La identidad 11.45 sigue siendo válida cuando n es negativo, siempre y cuando definamos $f_{-k} = f_{-k+2} - f_{-k+1}$ (cuando $k > 0$), de manera que

$$f_{-k} = (-1)^{k+1} f_k.$$

Así,
$$\tau^{-k} = f_{-k}\tau + f_{-k-1}$$

11.46
$$= (-1)^{k+1}(f_k\tau - f_{k+1}).$$

y

11.47
$$(-\tau)^{-k} = f_{k+1} - f_k\tau$$

11.471
$$= f_{k-1} - f_k\tau^{-1}.$$

Al sumar 11.45 (donde k substituye a n) y 11.47, obtenemos

11.48
$$g_k = \tau^k + (-\tau)^{-k}\,.$$

De la misma manera, al substraer 11.471 de 11.45 (donde k substituye a n), obtenemos

$$f_k = \frac{\tau^k - (-\tau)^{-k}}{\tau + \tau^{-1}} = \frac{\tau^k - (-\tau)^{-k}}{\sqrt{5}}, \tag{11.49}$$

que es una fórmula explícita, descubierta por J. P. M. Binet en 1843.

A partir de 11.48 y 11.49 deducimos inmediatamente

$$\tau^k = \frac{g_k + f_k\sqrt{5}}{2}, \quad (-\tau)^{-k} = \frac{g_k - f_k\sqrt{5}}{2};$$

por ejemplo

$$\tau^3 = \sqrt{5} + 2, \qquad \tau^{-3} = \sqrt{5} - 2,$$
$$\tau^6 = 9 + 4\sqrt{5}, \qquad \tau^{-6} = 9 - 4\sqrt{5}.$$

EJERCICIOS

1. Exprésese la suma de la serie finita

$$1 + 1 + 2 + 3 + \ldots + f_n$$

en términos de f_{n+2}.

2. Demuéstrese la identidad de Simson

$$f_{n-1}\, f_{n+1} - f_n^2 = (-1)^n.$$

Esta es la base de un problema desconcertante y entretenido [Ball 1, pág. 85].*

3. Verifíquese la observación de Lagrange acerca de que los dígitos finales de los números de Fibonacci empiezan a recurrir después de un ciclo de sesenta:

$$1, 1, 2, 3, 5, 8, 3, 1, 4, \ldots, 7, 2, 9, 1, 0.$$

4. A partir de la identidad

$$(1 - t - t^2)(1 + t + 2t^2 + 3t^3 + \ldots + f_{n+1}\, t^n + \ldots) = 1,$$

que implica

$$\sum_0^\infty f_{n+1}\, t^n = \frac{1}{1 - t - t^2} = \sum_0^\infty (t + t^2)^k = \sum_0^\infty t^k(1 + t)^k = \sum_{k=0}^\infty \sum_{j=0}^k \binom{k}{j} t^{k+j},$$

dedúzcase la fórmula de Lucas para los números de Fibonacci en términos de coeficientes binomiales:

$$f_{n+1} = \binom{n}{0} + \binom{n-1}{1} + \binom{n-2}{2} + \cdots.$$

5. Dése el valor de 0.01 a t en la identidad $\Sigma f_{n+1}\, t^n = (1,-t-t^2)^{-1}$, y dedúzcase una manera fácil de escribir los 19 primeros dígitos en la forma decimal $\frac{10000}{9899}$.

6. Tabúlese g_{n+1}/g_n y encuéntrese su límite cuando n tiende a infinito.

---

* Véase también Coxeter, The golden section, phyllotaxis, and Wythoff's game, *Scripta Mathematica*, 19 (1953), pág. 139. Desde un punto de vista más general,

$$f_{n+h}\, f_{n+k} - f_n\, f_{n+h+k} = (-1)^n f_h\, f_k;$$

véase A. E. Danese, *American Mathematical Monthly*, 67 (1960), pág. 81.

data

2021 06 20 M

vsip.info doku.pub

Company GML

gmllimited.com

Av. Louise, 54, 1000

Bruxelles, Belgium

RHtps://mail.google.com/a/1temiomeyclon

Faltam:
→ 39-42
178

Fundamentos de Geometria Coxeter
Pag. 154 (173)

$e^{\pi i}+1=0$

Pag. 175 (195)
La Espiral Áurea.

ISBN 978-9681806415

the Real Projective Plane
H.S.M. Coxeter
ISBN 978-1-4612-7647-0

Geometria de Pasch.

www.amazon.com/Fundamentos-Geometria-H-S-Coxeter/dp/9681806417

US$ 169,49 M

R$ 660
120
R$ 780
580
R$ 1.360
380
R$ 1.740
160
R$ 1.900

1900
80
20
2000
140
2140

US$
170

3
170
5
850
2140
R$ 2990

data

S T Q Q S S D

Fundamentos de Matemática
Geometrias Analítica e Vetorial
Euclidianas e Não-Euclidiana
2 set 2008
Ana Chiummo
ISBN 978-852161635-1 316 pag

20210621III

idoc.pub / arpdf.com

Books-013-2021/Important/
François Apéry - Models of the
Real Projective Plane...pdf
Pág. 8 (VII)
→ image ✡
Pág. 18 (30)
Fig. 8
Fig. 9 Stereographic projection
from the sphere to
the plane.

Mathematicorum

' (
+,- * *. ! * = ! G +2E 0 " !
2 ? 7 2! 2 !2 ! " A 0 ! = " H !
) 2"
' &&

Preface

by *Egbert Brieskorn*

If you feel attracted by the beautiful figure on the cover of this book, your feeling is something you share with all geometers. "It is the delight in figures in a higher sense which distinguishes the geometer". This is a well known statement of the algebraic geometer Alfred Clebsch. Throughout the history of our science, great men like Plato and Kepler and Poincaré were inspired by the beauty and harmony of such figures. Whether we see them as elements or representations of some universal harmony or simply look at them as a structure accessible to our mind by its very harmony, there is no doubt that the contemplation and the creation of beautiful geometric figures has been and will be an integral part of creative mathematical thought.

Perception of beauty is not the work of the intellect alone – it needs sensitivity. In fact, it does need sensual experience. To seperate the ideas from the appearances, the structure from the surface, analytical thinking from holistic perception and science from art does not correspond to the reality of our mind and is simply wrong if it is meant as an absolute distinction. It would deprive our creative thought of one of its deepest sources of inspiration.

I remember a visit of the blind mathematician B. Morin in Göttingen. He gave us a lecture on the eversion of the sphere, the process which he had discovered for turning the sphere inside out by a continuous family of immersions. At that time, he had no analytic description of this very complicated process, but he had a very precise qualitative picture of its geometry in his mind, and he had pictures of models made according to his instruction. I was deeply impressed by the beauty of these figures, and I was moved by the fact that he had been able to see all this beautiful and complex geometry which we could visualize only with great difficulty even after it was shown to us.

In Göttingen we had a very nice collection of mathematical models dating from the days of Felix Klein and David Hilbert. You can find photographs of some of them in a very nice book of Gerd Fischer entitled "Mathematical models", which was published by Vieweg. Among these models there is one made in 1903 showing the surface of Boy. This is a surface obtained by an immersion of the real projective plane in 3-dimensional space. At the time when it was made it was known that the projective plane cannot be embedded as a smooth surface in 3-space. This is so because any smooth closed surface in 3-space divides the space into an interior part and an exterior. This implies that it is orientable: If you stand in the exterior with your feet on the surface, you know what it means to turn left. But the real projective plane is not orientable. This was a very remarkable discovery of Felix Klein in 1874, and it implies that the projective plane $\mathbb{P}^2$ cannot be embedded in the 3-dimensional space $\mathbb{R}^3$. So if we want to visualize $\mathbb{P}^2$ by means of a surface in $\mathbb{R}^3$, we have to be contented with less than a smooth surface. For instance we should admit that different parts of the surface penetrate each other. If the parts themselves are smooth and penetrate each other transversally, we call the surface transversally immersed. It is by no means obvious that there is a presentation of $\mathbb{P}^2$ as an immersed surface in $\mathbb{R}^3$. If there was none, we would be forced to allow the surface to have some singular points, i. e. points where it is not smooth. Such representations of $\mathbb{P}^2$ by surfaces with selfpenetrations and singular points were known to exist. One of them is a beautiful surface discovered by the German geometer Jacob Steiner. He called it his "Roman surface", because he discovered some of its properties during a stay at Tome in 1844. You can see a picture of this surface on **Plate 57** of this book and you can buy a pretty 3-dimensional model of it from Vieweg. It has tetrahedral symmetry, with six lines of selfpenetration issuing from the center of the tetrahedron in the direction of the midpoints. These singular points are of the simplest possible type. They are so called "Whitney-umbrellas" (**Plate 23**). These singularities are the simplest singularities (apart from selfpenetration) which can occur in this situation, and they are stable: a Whitney umbrella will not go away if the surface is deformed slightly.

Maybe it was the stability of these singularities that made David Hilbert suspicious so that he thought it would be impossible to present the projective plane $\mathbb{P}^2$ by an immersed surface without singular points. But the famous mathematician David Hilbert was mistaken. In 1901 his student W. Boy produced two surfaces which are immersions of $\mathbb{P}^2$ in $\mathbb{R}^3$. One of them is particularly pretty. It has a 3-fold axis of symmetry and this is the Boy surface which is the main theme of this book.

Of course, B. Morin knew Boy's surface when he visited us in Göttingen. He must have had a very clear picture of it in his mind. Nevertheless he asked us to take our model out of the show-case so that he could touch it and feel whether its curvature was beautiful. It is this kind of sensitivity for the genesis and transformation of forms which I think is necessary for the creation of ideas such as those of B. Morin or R. Thom leading to the beautiful constructions of F. Apéry that you find in this book.

Boy's surface, like Morin's halfway model of the eversion of the sphere at the time of his visit in Göttingen, was only described in terms of a qualitative geometry. Such descriptions are quite satisfactory from the point of view of differential topology. In fact since the days of H. Poincaré we know that in many situations such a qualitative description is all we can hope for. However, there is an older tradition in mathematics which asks for a more definite way of describing solutions of a problem. In algebraic or analytic geometry, a surface imbedded in 3-space is to be described by an equation or a parametrization by nice analytic functions such as polynomials or by some beautiful geometric construction. In a sense, a geometric figure can only be considered as a unique entity worthy of this name "figure" in the ancient tradition if it is well-defined in this way. This is what Kepler meant when he said: "That which is construction in geometry is consonance in music". Thus our desire for harmony in a geometric figure will only be fulfilled if we see it described in such a way. This beautiful book of F. Apéry tells the exciting history of the discovery of such a description for the surface of Boy.

It begins with very qualitative topological and combinatorical descriptions of the projective plane and the surface of Boy, and it shows us nice descriptions in terms of equations and parametrizations obtained by B. Morin 1978 (**Plate 33**), J.-P. Petit and J. Souriau (1981) (**Plates 31, 32**), J. F. Hughes 1985 (**Plate 34**) and J. Bryant (**Plate 35**). He then presents a very nice solution of his own originating from an idea of B. Morin. The idea is to construct the apparent contour of the Boy surface seen from infinity in the direction of its 3-fold axis by interpreting parts of it as a deformation of the "folded handkerchief singularity" mapping (x, y) to (x^2, y^2). This leads to a parametrization of a Boy surface in 3-space by 3 polynomials of degree 4 (proposition 4 on page 60, **Plates 36–38**). However, all these solutions have slight deficiencies. For instance the equation describing the surface may have other real zeros different from the points of the Boy surface.

But finally the author presents a truly perfect solution. It describes the Boy surface by a beautiful geometric construction, which is subtle and elementary, in the best tradition of classical geometry. In order to make it easy for you to understand his analysis, let me explain the construction in purely geometric terms. It generates the surface by a 1-parameter family of ellipses obtained as follows.

Let L be a line and p a point on L. This line will become the 3-fold axis of symmetry, and the pole p will be a point where L intersects the surface. Let P be the plane orthogonal to L passing through p, and let Q be a plane parallel to P at distance 1. In Q draw a circle K of radius $\sqrt{2}/3$ with axis L. Then draw a threecuspidal hypocycloid C such that K is its inscribed circle. C may be obtained as trace of a point of the circumference of a circle of radius $\sqrt{2}/3$ rolling inside a circle of radius $\sqrt{2}$ concentric with K. The curve C will be the apparent contour of the Boy surface that we are going to construct, if we project the surface from the pole p to the plane Q.

Next we draw an elongated hypocycloid H in the plane Q as follows. Let K' be a circle of radius 1 in Q concentric with K. Let M be a moving circle of radius 1/3 which rolls without gliding inside K'. Let t_0 be a point in the moving plane of M at distance $\sqrt{2}/3$ from the centre of M. As M rolls, t_0 will trace out an elongated hypocycloid in the plane Q. This curve will have three ordinary double points, and they lie on K. We choose the initial position of t_0 so that these double points are situated on the tangents of the 3 cusps of the 3-cuspidal hypocycloid C, as indicated by the following figure.

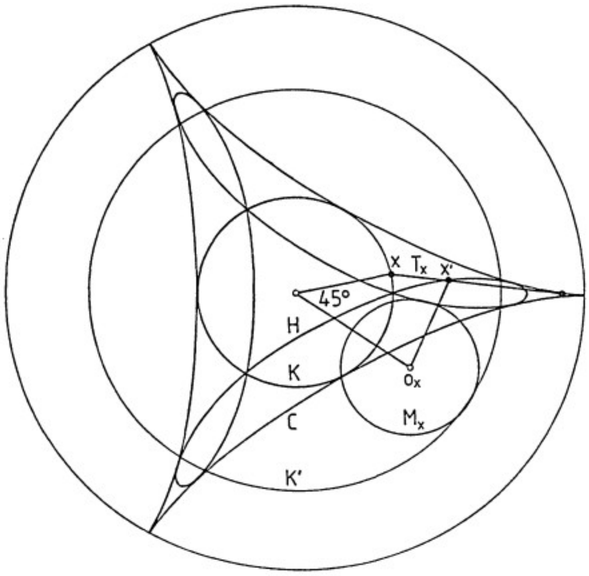

Now we define a map $K \to H$. Let x be a point moving on K in the positive sense. Let o_x be the centre of the moving circle M_x, where o_x is moving on a circle of radius 2/3 with the same angular velocity as x, but trailing behind 45°. The point t_0 for M_x determines a unique point x' on the elongated hypocycloid H, and our map is defined by associating this point x' to x.

It follows from the construction that x and x' lie on a tangent T_x to the hypocycloid C. Let P_x be the plane passing through this tangent T_x and the pole p. We define an ellipse E_x in the plane P_x by the following three conditions:

(a) E_x passes through x and x'.
(b) E_x is tangent to P in p.
(c) The distance d of x and x' and the maximal distance δ of a point on E_x from P are related as follows:

$$\delta = (1 - 9d^2/16)^{-1}.$$

These conditions define E_x uniquely.

And now our construction comes to its end: The surface S to be constructed is simply the union of all ellipses E_x. As x moves on K, the ellipses sweep out the surface. This is the Boy surface constructed by F. Apéry. Apéry proves that it is exactly the set of real zeroes of a polynomial equation of degree 6 which he can write down explicitly. Moreover he gives a nice description of the transversal immersion $\mathbb{P}^2 \to S \subset \mathbb{R}^3$. The ellipses are the images of a pencil of projective lines passing through the point in $\mathbb{P}^2$ which corresponds to the pole p. If we interpret $\mathbb{P}^2$ as quotient of the 2-sphere S^2 identifying antipodal points, the corresponding map $S^2 \to \mathbb{R}^3$ has a nice description in terms of trigonometric functions, and the inverse images of the ellipses are the great circles passing through a pair of antipodal points. **Plates 39–41** show this pretty surface.

Apéry's construction is beautiful, elementary, but subtle and extremely elegant. Moreover there is a similar construction for the Roman surface of Steiner, and one can pass from one surface to the other by a continuous family of surfaces, such that at a certain moment the six Whitney-umbrella-type singularities on the deformations of the Roman surface annihilate each other in pairs by a nice process called hyperbolic confluence. This process is illustrated on **Plates 28–30**, and the family of surfaces is shown on **Plates 51–56.**

Finally Apéry generalizes his approach and gets similar descriptions of other surfaces. Among them is the beautiful halfway model of B. Morin which you see on the cover of this book. All this is sheer beauty, and I hope that you will be delighted.

The compactness of $S^2/\sim$ implies j is a homeomorphism, so $S^2/\sim$ is a representation of the projective plane $\mathbb{P}^2$.

Using representation 4 we see that by j, a line corresponds to the quotient of a great circle of S^2 by the antipodal relation.

Representation 6 (Curt Reinhardt [RE]).

We start with the graph given by the edges and vertices of a regular octahedron, and we construct a 2-complex structure by adding four triangular faces of the octahedron, chosen so that any two have at most one common vertex, and three square faces corresponding to the diagonal of the octahedron (Fig. 7).

It is easy to prove we obtain a closed surface H, and we have

$$\chi(H) = 6 - 12 + 7 = 1$$

The 2-complex H is called the *Rheinhardt heptahedron.* It is homeomorphic to $\mathbb{P}^2$.

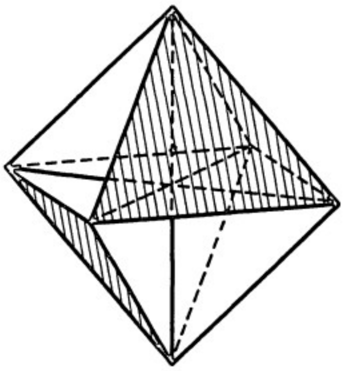

Fig. 7 Rheinhardt heptahedron.

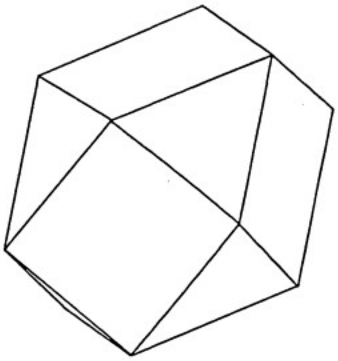

This heptahedron can be described as the quotient of the cuboctahedron by the antipodal map

Representation 7.

As in the preceding representation, we build a 2-complex whose vertices are those of a regular octahedron 0. We specify two antipodal vertices as the north and the south pole, so the four remaining vertices define the equatorial square. To the twelve edges of the octahedron, we add the two diagonals of the equatorial square. To complete the structure we add the four faces of 0 containing the south pole, two faces of 0 with just the north pole as common vertex, two new faces defined by the north pole and the diagonals of the equatorial square, and a quadrilateral face bounded by the two equatorial edges still free and the two previous diagonals (Fig. 8).

As previously, this 2-complex, called the *combinatorial cross-cap* and denoted by C, is a closed surface and its Euler characteristic yields

$$\chi(C) = 6 - 14 + 9 = 1 .$$

Thus C is homeomorphic to $\mathbb{P}^2$.

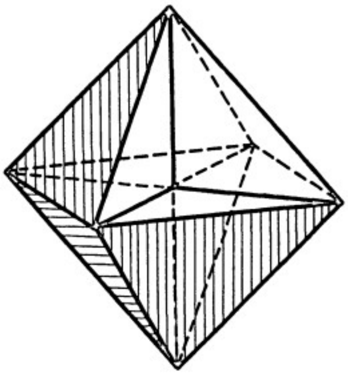

Fig. 8 Combinatorial cross-cap.

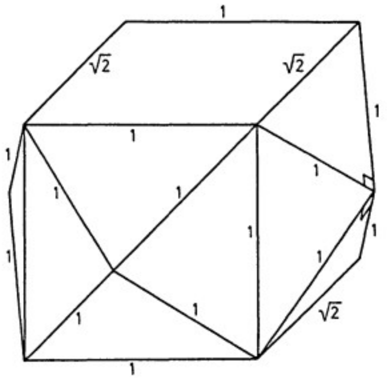

The combinatorial cross-cap can be obtained by making the quotient of this 2-complex by the antipodal map. We have replaced four lateral faces of a rectangular parallelepipet by four pyramids.

Representation 8.

We consider the equivalence relation on the *complex projective line* $\mathbb{C} \cup \{\infty\}$ given by

$$z \sim -1/\bar{z} \quad \text{and} \quad 0 \sim \infty$$

$\mathbb{C} \cup \{\infty\}$ is the Alexandroff compactification of $\mathbb{C}$; so prop. 1 sect. 1.1 implies the compactness of the quotient $\mathbb{C} \cup \{\infty\}/\sim$. This leads back to the spherical model of representation 5 by the stereographic projection from S^2 to $\mathbb{C} \cup \{\infty\}$ (Fig. 9):

$$p(x, y, z) = (x + iy)(1 - z)^{-1} \quad \text{and} \quad p(0, 0, 1) = \infty$$

This homeomorphism induces a homeomorphism between $(\mathbb{C} \cup \{\infty\})/\sim$ and the quotient of S^2 by the antipodal relation.

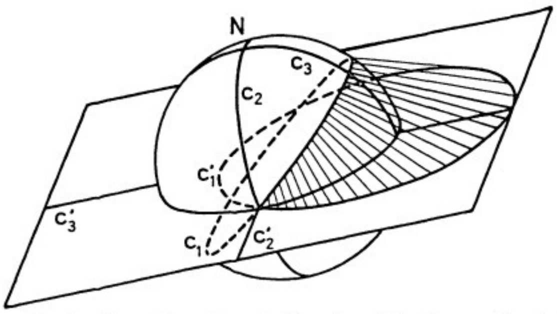

Fig. 9 Stereographic projection from the sphere to the plane. The image of a circle c is denoted by c′.

The image of a great circle on S^2 by p is either an affine line containing the origin 0 and completed by ∞, or a circle with center z and radius $1 + z\bar{z}$, according to the fact that the great circle goes or not through the north pole $(0, 0, 1)$ (Fig. 9). It gives the image of a line in this representation.

Representation 9.

Let K be the 2-complex, in Fig. 10. We consider the equivalence relation $\sim$ on K which identifies the edges according to the numbering and the orientation indicated on the figure. The particularity of this 2-complex is the possibility of realizing the former identifications by folding paper according to instructions given in Fig. 10.

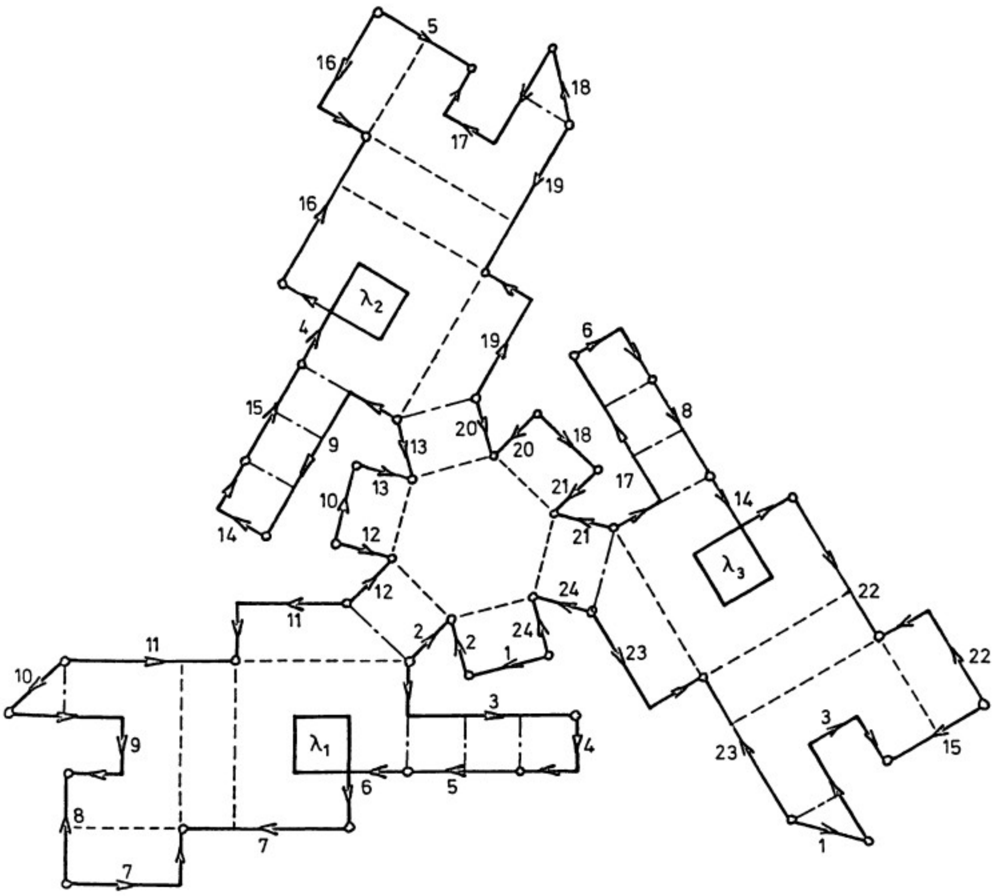

Fig. 10 The numbering of the edges of this 2-complex defines an equivalence relation which can be realized by folding the paper along the dashed lines. The two types of dashed lines indicate the direction of the folding: – – – – for ∨ folding —◦— for ∧ folding. The self-intersection curve in the paper model is the image of the loop defined by the sum $(6-) . \lambda_1 . (6+) . (8) . (14) . (15) . (4-) . \lambda_2 . (4+) . (5) . (6) . (8) . (14-) . \lambda_3 . (14+) . (15) . (4) . (5)$.
where $(6-) . (6+) = (6) \qquad (4-) . (4+) = (4) \qquad (14-) . (14+) = (14)$

It is easy to see that the quotient space $K/\sim$ is homeomorphic to the quotient of the disk $\bar{B}^2$ by the antipodal relation on the boundary (Fig. 11), so $K/\sim$ is homeomorphic to the projective plane $\mathbb{P}^2$ (see representation 3). By prop. 7 sect. 1.1 there are some points of self-intersection in the paper model in $\mathbb{R}^3$. In Fig. 10 these points come from the loop given by the sum

$$\begin{aligned}\lambda = {} & (6-)\,.\,\lambda_1\,.\,(6+)\,.\,(8)\,.\,(14)\,.\,(15)\,.\,(4-).\,\lambda_2\,.\,(4+)\,.\,(5)\,.\,(6)\,.\,(8)\,.\\ & .\,(14-)\,.\,\lambda_3\,.\,(14+)\,.\,(15)\,.\,(4)\,.\,(5)\end{aligned}$$

where

$$\begin{aligned}(6-)\,.\,\ (6+) &= \ (6)\\ (4-)\,.\,\ (4+) &= \ (4)\\ (14-)\,.\,(14+) &= (14)\end{aligned}$$

and (i) denotes the path whose number is i.

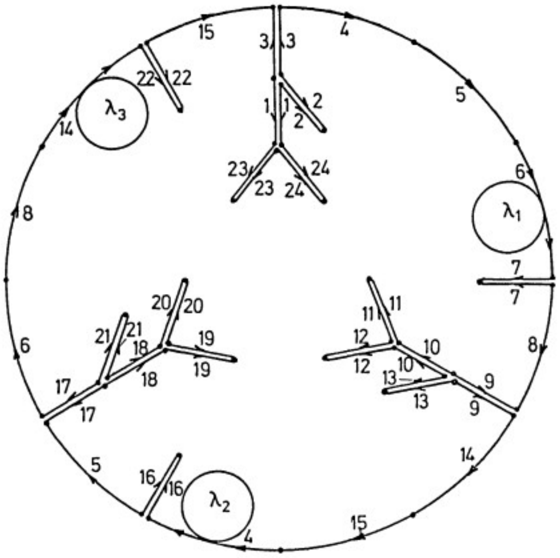
15
3
3
4
22
22
5
1
1
2
2
λ_3
14
23
24
23
24
6
8
λ_1
7
7
20
20
11
11
21
21
18
19
12
10
8
6
19
12
10
18
13
9
17
13
9
17
14
5
16
16
λ_2
15
4

Fig. 11 The model of Fig. 10 can be regarded as the quotient of a disk under the identifications shown.

The image of $K/\sim$ in $\mathbb{R}^3$ obtained by folding paper is a polyhedron called the *combinatorial Boy surface* (Fig. 12). The image of λ, that is the self-intersection curve, defines a loop in $\mathbb{R}^3$ which is called a *three-bladed propeller* (Fig. 13). It has only one multiple point, which is a triple point.

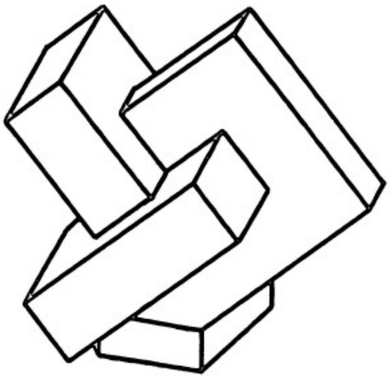

Fig. 12 Combinatorial model of the Boy surface constructed in Fig. 10.

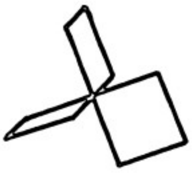

Fig. 13 Self-intersection curve of the model of Fig. 12: this is a three-bladed propeller.

Representation 10.

Let K be the 2-complex of Fig. 14. The complex K is isomorphic to the 2-complex K' of Fig. 15 (by an isomorphism of 2-complex we mean a homeomorphism which restricts to a homeomorphism on the underlying graphs and a bijection on the sets of vertices). The isomorphism is fixed by the numbering and the orientation of the edges on each figure.

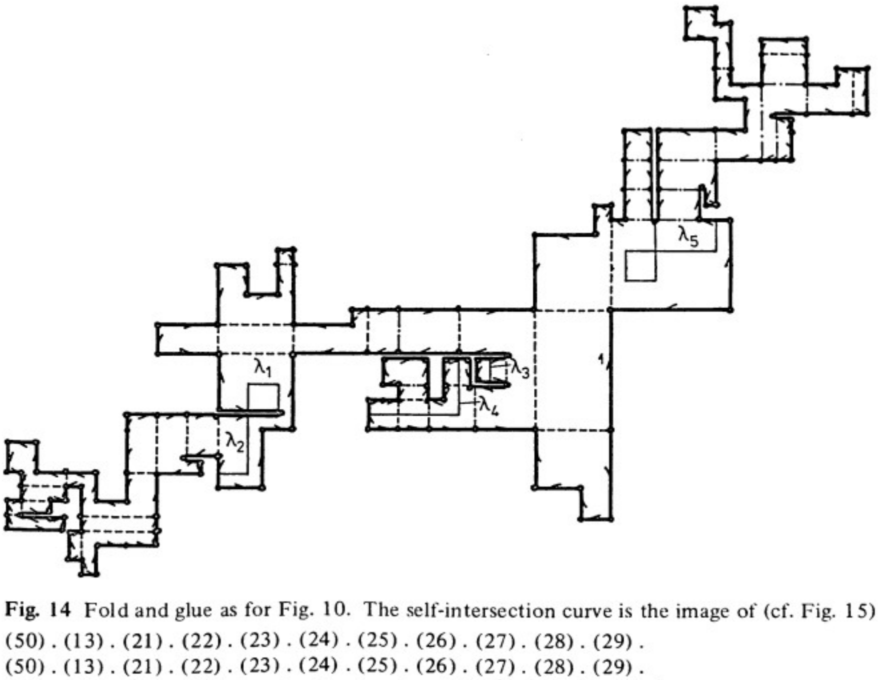

Fig. 14 Fold and glue as for Fig. 10. The self-intersection curve is the image of (cf. Fig. 15)
(50) . (13) . (21) . (22) . (23) . (24) . (25) . (26) . (27) . (28) . (29) .
(50) . (13) . (21) . (22) . (23) . (24) . (25) . (26) . (27) . (28) . (29) .
(33) . (45) . (47) . (49) . $\lambda_1 \cdot \lambda_2 \cdot \lambda_3 \cdot \lambda_4 \cdot \lambda_5$. $(32)^{-1}$. $(31)^{-1}$. $(30)^{-1}$

Spline ⇒ mecânica

Is a rectangular key fitting into grooves in the hub and shaft of a wheel, especially one formed integrally with the shaft which allows movement of the wheel. On the shaft.

Spline curves ⇒ Source:

Books-013-2021/20210623/
Aufgaben und...2/
Klaus Höllig - Approximation
and Modelling with B-Splines ...
(2014).pdf

Pág. 105(114)

Chapter 6 - Spline Curves.

20210624///

ID 2041318///

Chiarugi

segecosta@unibo.it

⇒ Gentile Sara Venturelli

recupero posti///

+,- * *. ! * I ! <J F KJ <J
+22 / ! ! " " A 2 0 & " ! ! " +22 " ! !
(2" ' % (!2 9 2 7

Chapter 6

Spline Curves

B-spline techniques have become the method of choice for modeling curves. Generalizing the Bézier form, spline representations possess analogous, very intuitive geometric properties. This was realized by Riesenfeld and Gordon, who saw the potential of B-splines for computer aided design. In contrast to polynomials, B-splines allow local shape control. Moreover, they provide approximations of arbitrary accuracy and smoothness.

In Section 6.1, we define polynomial and rational spline curves and, in Section 6.2, discuss basic properties. Then, we describe the fundamental knot insertion and refinement algorithms in Section 6.3. They play a key role for many applications such as conversion to Bézier form and numerical processing of spline curves. Section 6.4 is devoted to evaluation and differentiation. Finally, in Section 6.5, we explain how interpolation schemes for spline functions can be applied to curves.

6.1 ▪ Control Polygon

Spline curves are parametrized by linear combinations of B-splines. The coefficients form a control polygon with geometric properties familiar from the Bézier representation.

Spline Curve

A spline curve of degree $\le n$ in $\mathbb{R}^d$ has a parametrization

$$t \mapsto (p_1(t), \ldots, p_d(t)) = \sum_{k=0}^{m-1} c_k b_k(t), \quad t \in D_\tau^n,$$

with components p_ν in a spline space S_τ^n with finite knot sequence $\tau:\ \tau_0, \ldots, \tau_{m+n}$ and parameter interval $D_\tau^n = [\tau_n, \tau_m]$.

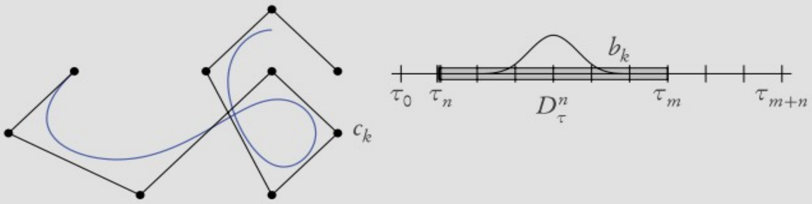

The coefficients $c_k = (c_{k,1}, \ldots, c_{k,d})$ can be combined into an $m \times d$ array C. They are called control points and form the control polygon c for p.

The shape of a spline curve is not only determined by the control points but also influenced by the placement of the knots. This is illustrated in the following example.

□ **Example:**

The figure shows two cubic spline curves together with their control polygons and knot sequences.

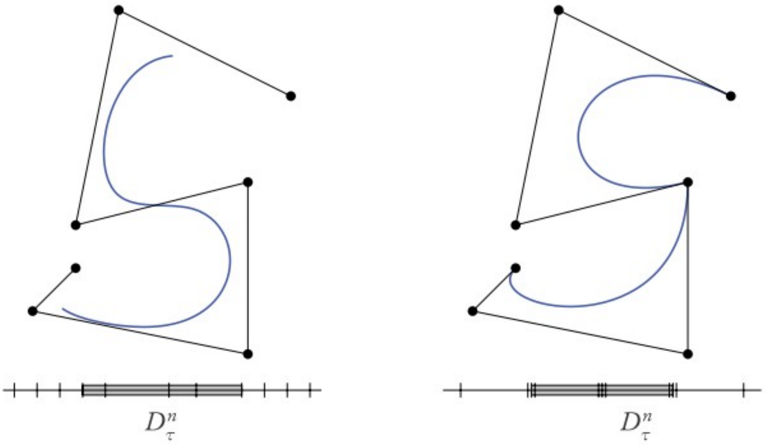

The left curve has simple knots; hence, the parametrization is twice continuously differentiable. The example on the right shows how nonsmooth curves can be modeled with the aid of multiple knots. At a knot with multiplicity n, the derivative of the parametrization need not be continuous. Hence, the tangent direction can change abruptly. Moreover, due to the n-fold knots at the endpoints of the parameter interval, the curve interpolates the first and the last control points. □

If the spline parametrization is periodic, the corresponding curve is closed. Accordingly, the control polygon is also closed, reflecting the periodicity of the B-spline coefficients. A precise definition, also introducing the proper notation, is given below.

Closed Spline Curve

A closed spline curve of degree $\le n$ in $\mathbb{R}^d$ has a parametrization

$$t \mapsto (p_1(t), \ldots, p_d(t)) = \sum_{k \in \mathbb{Z}} c_k \, b_k(t), \quad t \in \mathbb{R},$$

with components which are continuous T-periodic splines. This means that $p_\nu \in S^n_{\tau,T}$, $\tau = (\tau_0, \ldots, \tau_{M-1})$, and the B-splines b_k correspond to the periodically extended knot sequence $(\ldots, \tau - T, \tau, \tau + T, \ldots)$ and to M-periodic control points.

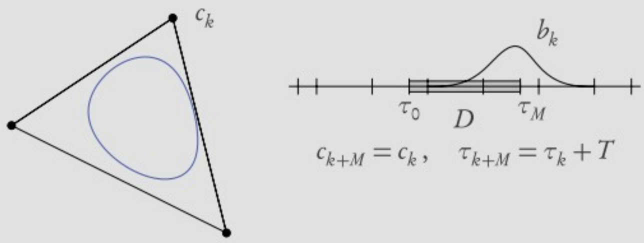

$$c_{k+M} = c_k\,, \quad \tau_{k+M} = \tau_k + T$$

According to the periodicity conditions, p is determined by M consecutive control points

$$C = \begin{pmatrix} c_0 \\ \vdots \\ c_{M-1} \end{pmatrix},$$

which form a closed control polygon for p.

The following example illustrates the representation of closed spline curves.

□ **Example:**

The figure shows two closed cubic spline curves together with their knots $(\tau_0, \dots, \tau_{M-1})$ in a periodicity interval $D = [\tau_0, \tau_0 + T)$.

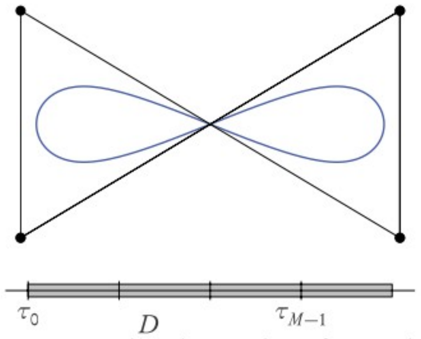

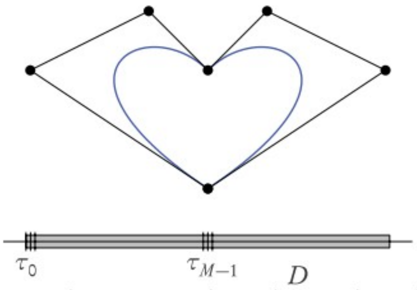

We note that the number of control points exactly corresponds to the number of knots in the half-open periodicity interval. For the example on the right, each knot has multiplicity 3. In particular, $\tau_0 = \tau_1 = \tau_2$ and $\tau_M = \tau_{M+1} = \tau_{M+2}$ are the two endpoints of D. □

As is illustrated in the figure below, we can describe a closed spline curve $p(t)$, $t \in D = [\tau_0, \tau_0 + T]$, with a nonperiodic parametrization. We adjoin to the knots $\tau_0, \dots, \tau_{M-1}$, $\tau_M = \tau_0 + T$, successively n knots on each side of D according to the periodicity condition:

$$\tau \to \tilde{\tau} : (\tau_{M-n} - T, \dots, \tau_{M-1} - T, \tau_0, \dots, \tau_{M-1}, \tau_0 + T, \dots, \tau_n + T\,.$$

The corresponding control points are

$$c_{M-n}, \dots, c_{M-1}, c_0, \dots, c_{M-1}\,,$$

yielding a parametrization with components in the spline space $S^n_{\tilde{\tau}}$. We note that for $n > M$ the auxiliary control points c_k have to be defined successively, for $k = -1, -2, \dots$.

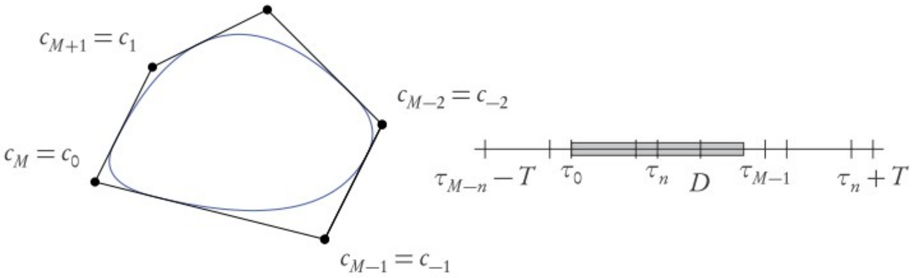

The alternative representation of a closed spline curve is convenient for numerical methods since algorithms are usually formulated for the nonperiodic case. With the above identification, closed curves can be handled with minor modifications only.

As for Bézier curves, we can use rational parametrizations. While the gain in accuracy and design flexibility is marginal, a number of standard shapes can be represented exactly. Therefore, this most general curve format has become a standard in many modeling systems.

Rational Parametrizations

A nonuniform rational B-spline parametrization (NURBS) $r = p/q$ is the quotient of a spline parametrization $t \mapsto p(t)$ with weighted control points

$$(c_k w_k) \in \mathbb{R}^d, \quad w_k > 0,$$

and a spline function $t \mapsto q(t)$ with coefficients w_k.

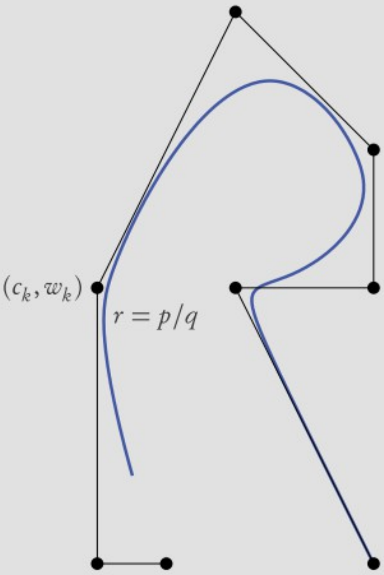

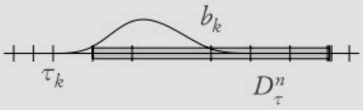

The weights w_k provide additional design flexibility. Increasing a weight pulls the curve towards the corresponding control point.

We can identify r with a spline curve in homogeneous coordinates parametrized by

$$t \mapsto (p(t)\,|\,q(t)) = \sum_k (c_k w_k \,|\, w_k)\, b_k(t) \in \mathbb{R}^{d+1}.$$

This interpretation is convenient for implementing algorithms such as knot insertion, evaluation, and differentiation.

As an example, we discuss the relationship to the rational Bézier form which is a special case of the NURBS representation.

□ **Example:**

Conic sections can be represented by rational quadratic Bézier curves. For example, the quarter circle with the standard trigonometric parametrization

$$(\cos\varphi, \sin\varphi), \quad \varphi \in [0, \pi/2],$$

has the Bézier control points and weights

$$(c_k \mid w_k): \quad (1,0 \mid 1), (1,1 \mid 1/\sqrt{2}), (0,1 \mid 1).$$

Accordingly, a full circle can be represented by a closed rational quadratic spline curve r in Bézier form with double knots

$$\tau_0 = 0, 0, 1, 1, 2, 2, 3, 3 = \tau_7, \quad \tau_{k+8} = \tau_k + 4$$

and control points and weights shown in the figure.

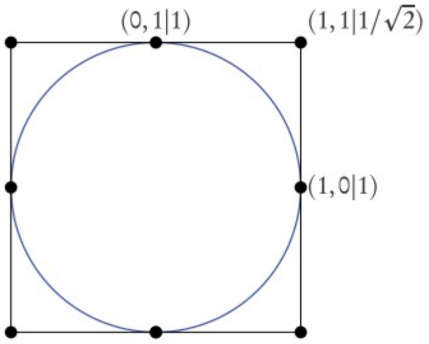

It is perhaps a bit surprising that the Bézier form does not correspond to a closed rational quadratic spline curve with simple uniform knots parametrized by

$$t \mapsto r(t) = \frac{\sum_k (c_k w_k)\, b^2(t-k)}{\sum_k w_k\, b^2(t-k)},$$

where

$$c_0 = (1,1),\ c_1 = (-1,1),\ c_2 = (-1,-1),\ c_3 = (1,-1), \quad c_{k+4} = c_k\,.$$

If this were the case, $b^2(1) = b^2(2) = 1/2$ implies

$$1 = |r(k)|^2 = \left| \frac{c_{k-2} w_{k-2}/2 + c_{k-1} w_{k-1}/2}{w_{k-2}/2 + w_{k-1}/2} \right|^2 .$$

In view of the specific form of the control points this yields

$$(w_{k-2} + w_{k-1})^2 + (w_{k-2} - w_{k-1})^2 = (w_{k-2} + w_{k-1})^2\,.$$

We conclude that all weights must be equal. But then the parametrization is polynomial which is not possible. □

6.2 ▪ Basic Properties

The geometric properties of the Bézier control polygon persist for spline curves. In particular, with appropriate choice of the knots, the endpoint interpolation property can be maintained.

Properties of the Control Polygon

The control polygon c of a spline curve parametrized by

$$p = \sum_{k=0}^{m-1} c_k\, b_k, \quad p_\nu \in S_\tau^n, \quad \tau:\ \tau_0, \ldots, \tau_{m+n},$$

qualitatively models the shape of p.

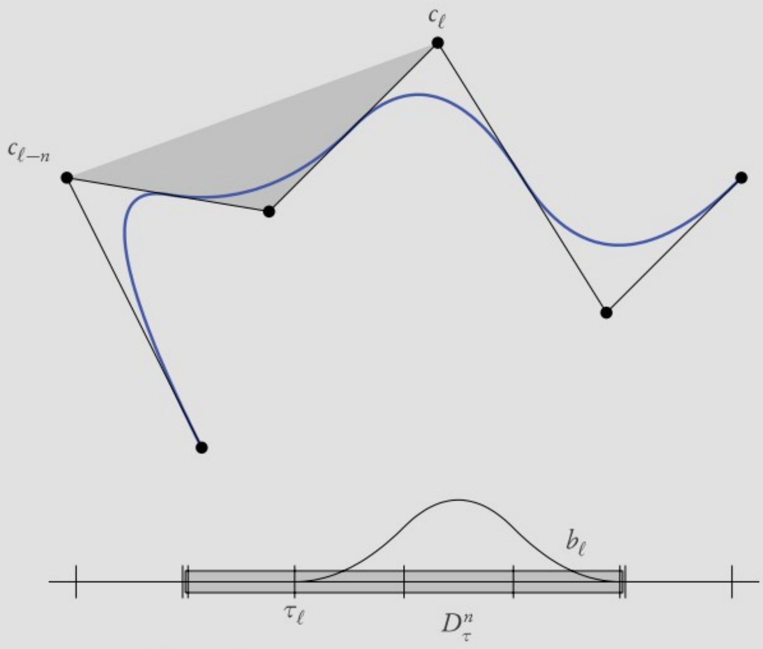

As is illustrated by the figure,

- for $\tau_\ell \le t \le \tau_{\ell+1}$ the point $p(t)$ lies in the convex hull of $c_{\ell-n}, \ldots, c_\ell$.

Moreover, if both endpoints of the parameter interval $D_\tau^n = [\tau_n, \tau_m]$ are knots with multiplicity n, then

- $p(\tau_n) = c_0$, $p(\tau_m) = c_{m-1}$,
- $p'(\tau_n^+) = \alpha_{1,\tau}^n (c_1 - c_0)$, $p'(\tau_m^-) = \alpha_{m-1,\tau}^n (c_{m-1} - c_{m-2})$

with $\alpha_{k,\tau}^n = n/(\tau_{k+n} - \tau_k)$. The last two properties referred to as endpoint interpolation imply that the control polygon is tangent to the spline curve, which is very useful for design purposes.

While the parametrization of a spline curve is continuous, the derivative can have jumps. This is taken into account in the formula for p', where the superscripts $+/-$ denote limits from the right/left.

The properties are easily verified.

Since only the B-splines b_k, $k = \ell - n, \ldots, \ell$, are relevant for $t \in [\tau_\ell, \tau_{\ell+1}]$, and they are positive and sum to one, the convex hull property is obvious.

To check the endpoint interpolation, we consider, e.g., the left endpoint. Because of the knot multiplicity n, all B-splines except b_0 vanish at τ_n. Moreover, $b_0(\tau_n) = 1$, so that $p(\tau_n)$ coincides with the control point c_0. By the differentiation formula,

$$p' = \sum_{k=1}^{n} \alpha_{k,\tau}^{n} (c_k - c_{k-1}) \, b_{k,\tau}^{n-1}, \quad \alpha_{k,\tau}^{n} = n/(\tau_{k+n} - \tau_k),$$

on the interval (τ_n, τ_{n+1}). Since

$$b_{k,\tau}^{n-1}(\tau_n^+) = \delta_{1,k},$$

the formula for $p'(\tau_n^+)$ follows.

□ **Example:**

The figure illustrates the convex hull property and endpoint interpolation for spline curves of degree 2 and 3.

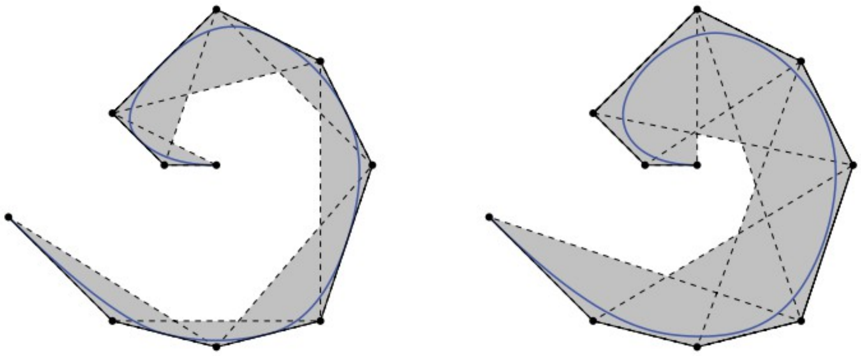

$n = 2$ $n = 3$

Obviously, for small degree, the convex hull provides a more accurate approximation. Moreover, if the control polygon contains nearly flat parts, the bounds are very tight. □

The convex hull property implies that the control polygon c is close to the corresponding curve p. In particular, if c is linear, so is p. Therefore, one expects that the distance to the spline curve is related to the second differences of the control points as a measure for the deviation from linearity. This is indeed the case as implied by the following precise bound due to Reif.

Distance to the Control Polygon

For a spline curve parametrized by

$$p = \sum_{k=0}^{m-1} c_k b_k, \quad p_\nu \in S^n_\tau ,$$

with $n > 1$, let c be a piecewise linear parametrization of the control polygon, which interpolates c_k at the knot averages $\tau_k^n = (\tau_{k+1} + \cdots + \tau_{k+n})/n$.

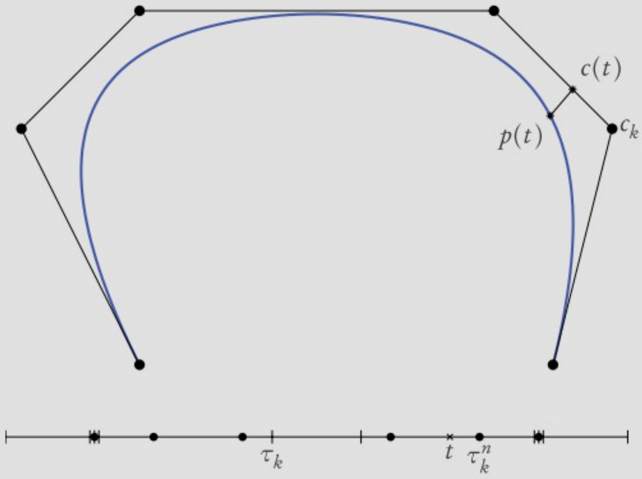

The distance of p to the control polygon can be bounded in terms of weighted second differences of the control points. For t in a nondegenerate knot interval $[\tau_\ell, \tau_{\ell+1}] \subseteq D_\tau^n$,

$$\|p(t) - c(t)\|_\infty \le \frac{1}{2n} \max_{\ell-n \le k \le \ell} \sigma_k^2 \max_{\ell-n+2 \le k \le \ell} \|\nabla_\tau^2 c_k\|_\infty,$$

where

$$\sigma_k^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\tau_{k+i} - \tau_k^n)^2$$

and $\nabla_\tau^2 c_k$ are the control points of the second derivative p''. Their explicit form is

$$\nabla_\tau^2 c_k = \frac{n-1}{\tau_{k+n-1} - \tau_k} \left(\frac{c_k - c_{k-1}}{\tau_k^n - \tau_{k-1}^n} - \frac{c_{k-1} - c_{k-2}}{\tau_{k-1}^n - \tau_{k-2}^n} \right),$$

where none of the denominators vanishes since the differences are at least as large as $\tau_{\ell+1} - \tau_\ell$.

The local estimate implies a global bound simply by taking the maximum of the right side over all k relevant for the entire parameter interval $D_\tau^n = [\tau_n, \tau_m]$ of the spline curve. In this case, $\nabla_\tau^2 c_k$ is set to 0 if $\tau_k = \tau_{k+n-1}$.

Clearly, it suffices to prove the estimate for each component separately. Therefore, we may consider the univariate case. By scaling the coefficients c_k, which appear linearly on both sides of the estimate, we can further assume that $\max_k |\nabla_\tau^2 c_k| = 1$.

Turning to the main arguments of the proof, we first show that the worst case corresponds to a parabola p^*,

$$p^*(t) = t^2/2 = \sum_k c_k^* b_k, \quad \nabla_\tau^2 c_k^* = 1\,.$$

In terms of the parametrizations $c(t)$ and $c^*(t)$ of the control polygons, this means that

$$-(p^* - c^*) \le p - c \le p^* - c^*$$

for any other spline $p = \sum_k c_k b_k$ with $\max_k |\nabla_\tau^2 c_k| = 1$.

Both inequalities follow from a simple geometric observation. Considering, e.g., an equivalent form of the right estimate

$$p - p^* = \sum_k (c_k - c_k^*)\, b_k \le c - c^*,$$

we observe that $\nabla_\tau^2 (c_k - c_k^*) \le 0$ in view of our assumption about the size of the weighted second differences. By definition, ∇_τ^2 forms differences of consecutive slopes of the polygon, which connects the points $(\tau_k^n,\, c_k - c_k^*)$ weighted by the positive factors $(n-1)/(\tau_{k+n-1} - \tau_k)$. Hence, the polygon is concave. Since, by Marsden's identity,

$$(t,\, p(t) - p^*(t)) = \sum_k (\tau_k^n,\, c_k - c_k^*)\, b_k(t),$$

the convex hull property implies that the graph of the spline $p - p^*$ lies below the polygon $c - c^*$, as claimed.

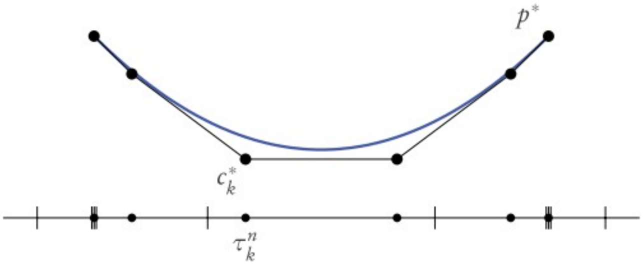

It remains to bound $p^* - c^*$, which then yields the estimate for $|p - c|$. Since the second derivative of this function, which is the difference between the parabola and the broken line shown in the figure, equals one on each interval (τ_k^n, τ_{k+1}^n), it follows that its maximum must occur at one of the knot averages. With the aid of Marsden's identity we can compute the explicit form of the control points for $p^*(t) = t^2/2$:

$$c_k^* = (\tau_k^n)^2/2 - \sigma_k^2/(2n)$$

(cf. the example in Section 4.3). Substituting this expression into

$$p^*(\tau_k^n) - c^*(\tau_k^n) = (\tau_k^n)^2/2 - c_k^*,$$

the desired estimate follows; i.e., we can bound $|p - c|$ by the maximum of $\sigma_k^2/(2n)$ for the relevant indices k.

□ **Example:**

As an example, we consider the quadratic spline curve $p = \sum_{k=0}^4 c_k b_k$ shown in the left figure with

$$C = \begin{pmatrix} 1 & 0 & 0 & 1 & 2 \\ 0 & 0 & 1 & 2 & 2 \end{pmatrix}^t, \quad \tau = (-1, 0, 0, 2, 6, 12, 14, 15).$$

The right diagram shows the error $e(t) = \|p(t) - c(t)\|_\infty$ for t in the parameter interval $D_\tau^2 = [0, 12]$ as well as the piecewise constant upper bound provided by the estimate.

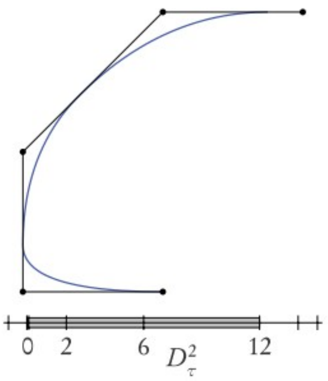

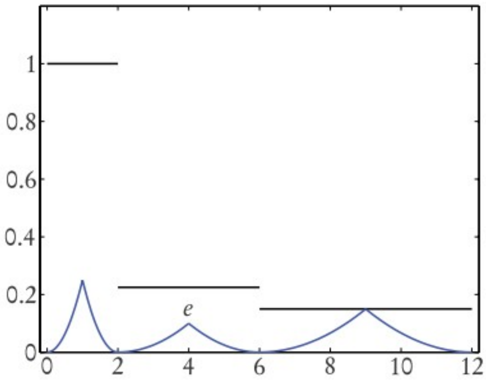

Specializing the general formulas for degree $n = 2$, we obtain

$$\tau_k^2 = (\tau_{k+1} + \tau_{k+2})/2, \quad \sigma_k^2 = (\tau_{k+2} - \tau_{k+1})^2/2.$$

The concrete values are

$$(\tau_0^2, \sigma_0^2) = (0,0), (1,2), (4,8), (9,18), (13,2) = (\tau_4^2, \sigma_4^2).$$

Finally, the weighted control point differences equal

$$\nabla_\tau c_k = \frac{c_k - c_{k-1}}{\tau_k^2 - \tau_{k-1}^2}: \quad (-1,0), (0,1/3), (1/5,1/5), (1/4,0),$$

and

$$\nabla_\tau^2 c_k = \frac{\nabla_\tau c_k - \nabla_\tau c_{k-1}}{\tau_{k+1} - \tau_k}: \quad (1/2,1/6), (1/20,-1/30), (1/120,-1/30)$$

with the index k starting from 1 and 2, respectively. This leads to the following estimates. For $t \in [0,2]$,

$$\|p(t) - c(t)\|_\infty \le \frac{1}{4} \cdot \max\{0,2,8\} \cdot \frac{1}{2} = 1.$$

Similarly, for the other two knot intervals $[2,6]$ and $[6,12]$, the bounds are

$$\frac{1}{4}\cdot 18\cdot\frac{1}{20}=\frac{9}{40},\quad \frac{1}{4}\cdot 18\cdot\frac{1}{30}=\frac{3}{20},$$

respectively.

The right figure shows that the estimate for the last knot interval is sharp. □

As a second example, we consider the case of uniform knots which permits slight simplifications.

□ **Example:**

For uniform knots τ_k with grid width h and odd degree $n=2m+1$,

$$\tau_k^n=\tau_k+(m+1)h,\quad \sigma_k^2=\frac{1}{n-1}\sum_{i=-m}^{m}(ih)^2\,.$$

Hence, the product of the first two factors in the error estimate for the control polygon equals

$$\frac{1}{2n}\,\frac{2h^2}{(n-1)}\sum_{i=1}^{m}i^2\,.$$

Substituting $m^3/3+m^2/2+m/6$ for the sum and cancelling common factors, we obtain the simplified expression

$$\frac{n+1}{24}\,h^2\,.$$

Therefore, since

$$\nabla_\tau^2 c_k=h^{-2}\Delta^2 c_{k-2},\quad \Delta^2 c_{k-2}=c_k-2c_{k-1}+c_{k-2}\,,$$

the error estimate for uniform knots has the form

$$||p(t)-c(t)||_\infty\le\frac{n+1}{24}\max_{\ell-n\le k\le\ell-2}||\Delta^2 c_k||_\infty$$

for $\tau_\ell\le t<\tau_{\ell+1}$.

An analogous computation shows that the same formula also holds for even degree $n=2m$. The fact that in this case the knot averages do not coincide with knots but are midpoints of knot intervals does not make any difference. □

6.3 ▪ Refinement

Refinement of the partition of the parameter interval is the standard procedure to increase the local accuracy or design flexibility of spline approximations. For spline curves, the corresponding algorithm for inserting new knots has a simple geometric interpretation in terms of the control polygon, as observed, independently, by Boehm and Cohen, Lyche, and Riesenfeld. It plays a fundamental role also for other applications.

Knot Insertion

Let $p = \sum_0^{m-1} c_k\, b_k \in S^n_\tau$ parametrize a spline curve. If we add a new knot s in the parameter interval D^n_τ and $s \in [\tau_\ell, \tau_{\ell+1})$, then the control points $\tilde{c}_k$ of p with respect to the refined knot vector

$$\tilde{\tau}: \quad \ldots,\, \tilde{\tau}_\ell = \tau_\ell,\, \tilde{\tau}_{\ell+1} = s,\, \tilde{\tau}_{\ell+2} = \tau_{\ell+1}, \ldots$$

are computed as follows.

On the segments $[c_{k-1}, c_k]$ with $\tau_k < s < \tau_{k+n}$ new control points are generated:

$$\tilde{c}_k = \gamma^n_{k,\tau} c_k + (1 - \gamma^n_{k,\tau}) c_{k-1}, \quad \gamma^n_{k,\tau} = \frac{s - \tau_k}{\tau_{k+n} - \tau_k}\,.$$

The other edges of the control polygon remain unchanged:

$$\tilde{c}_k = c_k \text{ for } \tau_{k+n} \le s, \quad \tilde{c}_k = c_{k-1} \text{ for } s \le \tau_k\,.$$

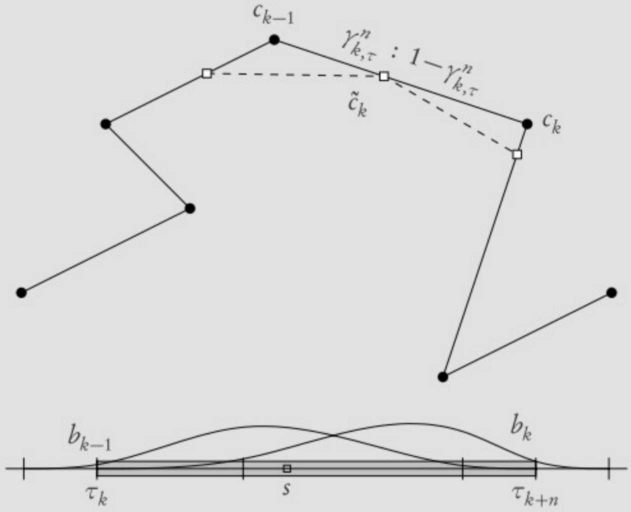

As illustrated in the figure, the new control point $\tilde{c}_k$ divides the segment $[c_{k-1}, c_k]$ in the same ratio as the parameter s divides the interval $[\tau_k, \tau_{k+n}]$, which is the intersection of the supports of the two associated B-splines. We note that, if s coincides with a knot ($s = \tau_\ell$), fewer control points need to be computed. More precisely, if s has multiplicity j in $\tilde{\tau}$, only $n + 1 - j$ convex combinations need to be formed.

Several new knots can be inserted by repeating the procedure. In particular, by raising the multiplicity of a knot to n, we obtain a point on the curve:

$$\tau_{\ell-n} < \tau_{\ell-n+1} = \cdots = \tau_\ell < \tau_{\ell+1} \implies p(\tau_\ell) = c_{\ell-n}\,.$$

Hence, the evaluation scheme for splines can be viewed as n-fold knot insertion.

To verify the knot insertion formula, we have to show that

$$\sum_k c_k\, b_k(t) = \sum_k \tilde{c}_k\, \tilde{b}_k(t), \quad \tilde{c}_k = \tilde{\gamma}_k c_k + (1 - \tilde{\gamma}_k) c_{k-1}\,,$$

where $\tilde{b}_k$ denote the B-splines of the refined knot sequence $\tilde{\tau}$ and

$$\tilde{\gamma}_k = \begin{cases} 1 & \text{if } \tau_{k+n} \le s, \\ \gamma_k & \text{if } \tau_k < s < \tau_{k+n}, \\ 0 & \text{if } s \le \tau_k. \end{cases}$$

The modified weights $\tilde{\gamma}_k$ incorporate the cases where the control points do not change.

As usual, for a proof, it is convenient to extend τ and $\tilde{\tau}$ to bi-infinite knot sequences and to restrict t to the parameter interval $D_\tau^n = [\tau_n, \tau_m]$ of the curve. Moreover, since the modified weights depend continuously on the knots, it is sufficient to consider simple knots.

With these notations and assumptions, the identity can be derived from general considerations. Since adding a knot enlarges the class of splines, any B-spline b_k can be written as a linear combination of the B-splines corresponding to $\tilde{\tau}$:

$$b_k = \sum_i d_{k,i} \tilde{b}_i .$$

Moreover, since b_k vanishes outside of the interval $[\tilde{\tau}_k, \tilde{\tau}_{k+n+2})$, it follows from the local linear independence of the B-splines that the coefficients can be nonzero only for $i = k$ and $i = k+1$. These observations imply that

$$\sum_k c_k\, b_k = \sum_k c_k\,(p_k \tilde{b}_k + q_k \tilde{b}_{k+1}) = \sum_k (c_k p_k + c_{k-1} q_{k-1})\, \tilde{b}_k .$$

Choosing $c_k = 1$, we obtain in particular

$$q_{k-1} = 1 - p_k$$

since the B-splines form a partition of unity.

It remains to determine the explicit form of the weight p_k. This can be done by analyzing the linear combination representing a single B-spline, which takes on the form

$$b_k = p_k \tilde{b}_k + q_k \tilde{b}_{k+1} .$$

We consider several cases.

(i) If $\tau_{k+n} \le s$, then b_{k-1} and $\tilde{b}_{k-1}$ have the same knots. Consequently, $b_{k-1} = \tilde{b}_{k-1}$, so that

$$1 = p_{k-1}, \quad 0 = q_{k-1} = 1 - p_k .$$

(ii) If $\tau_k < s < \tau_{k+n}$, then b_k has the knots $\tau_k, \ldots, \tau_{k+n+1}$, while $\tilde{b}_k$ has the knots $\tau_k, \ldots, s, \ldots, \tau_{k+n}$. It follows from the recursion for B-splines that

$$b_k(t) = \frac{t-\tau_k}{\tau_{k+1}-\tau_k} \cdots \frac{t-\tau_k}{\tau_{k+n}-\tau_k},$$
$$\tilde{b}_k(t) = \frac{t-\tau_k}{\tau_{k+1}-\tau_k} \cdots \frac{t-\tau_k}{s-\tau_k} \cdots \frac{t-\tau_k}{\tau_{k+n-1}-\tau_k}$$

for $t \in [\tau_k, \min(\tau_{k+1}, s))$. Since the B-spline $\tilde{b}_{k+1}$ vanishes on this interval,

$$p_k = \frac{s-\tau_k}{\tau_{k+n}-\tau_k} = \gamma_{k,\tau}^n .$$

(iii) If $s \le \tau_k$, then b_k and $\tilde{b}_{k+1}$ have the same knots. Hence, $b_k = \tilde{b}_{k+1}$ and

$$1 = q_k = 1 - p_{k+1}, \quad 0 = p_k\,.$$

In all cases,

$$p_k = \tilde{\gamma}_k\,,$$

proving the knot insertion formula.

□ **Example:**

We illustrate the knot insertion algorithm for the cubic spline curve $\sum_{k=0}^{6} c_k\, b_k$ with control points

$$C = \begin{pmatrix} 0 & 4 & 4 & 1 & 1 & 5 & 4 \\ 2 & 1 & 4 & 4 & 9 & 8 & 6 \end{pmatrix}^t$$

and knot sequence

$$\tau_0 = 1,\, 2,\, 2,\, 2,\, 3,\, 5,\, 5,\, 8,\, 9,\, 10,\, 11 = \tau_{10}\,,$$

shown in the figure.

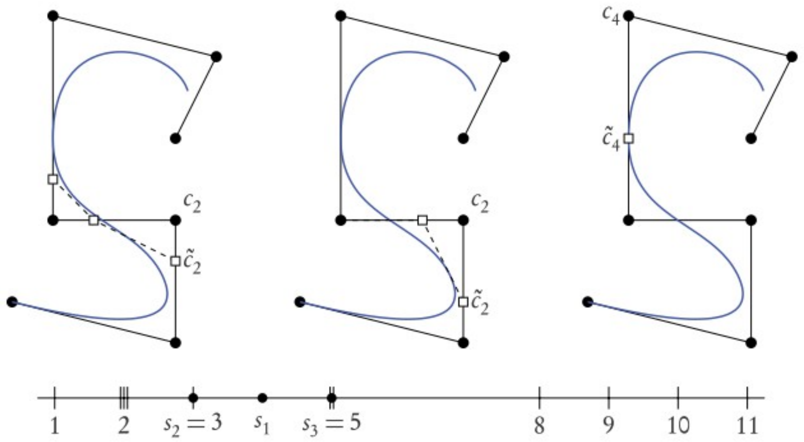

We consider three cases.

(i) $s = 4$ (multiplicity 1): Three segments of the control polygon are subdivided. The ratios

$$2:1,\; 2:1,\; 1:4$$

are determined by the relative position of s in the intervals

$$[\tau_k, \tau_{k+3}]:\; [2,5],\, [2,5],\, [3,8],$$

which are the intersections of the supports of the relevant B-splines. The resulting control points, corresponding to the refined knot sequence

$$\tilde{\tau}_0 = 1,\, 2,\, 2,\, 2,\, 3,\, 4,\, 5,\, 5,\, 8,\, 9,\, 10,\, 11 = \tilde{\tau}_{11}\,,$$

are

$$\tilde{C} = \begin{pmatrix} 0 & 4 & 4 & 2 & 1 & 1 & 5 & 4 \\ 2 & 1 & 3 & 4 & 5 & 9 & 8 & 6 \end{pmatrix}^t .$$

For example,

$$(4,3) = \tilde{c}_2 = \frac{2}{3}\,c_2 + \frac{1}{3}\,c_1$$

(cf. the left figure).

(ii) $s = 3$ (double knot): In this case, only the second and third segments are subdivided according to the ratios

$$1:2,\; 1:2\,,$$

as is shown in the middle figure.

(iii) $s = 5$ (triple knot): As is shown in the right figure, if the multiplicity after knot insertion is maximal, only one new control point is generated, which lies on the curve. In this example,

$$p(s) = \tilde{c}_4 = \frac{2}{5}\,(1,9) + \frac{3}{5}\,(1,4) = (1,6)\,. \qquad \square$$

Instead of inserting one knot at a time, several knots can be inserted simultaneously. For example, all knot intervals can be split in half to gain more approximation power on the entire parameter interval. The corresponding algorithm has a particularly elegant form for spline curves with uniform knots as was described by Lane and Riesenfeld.

Uniform Subdivision

Let

$$p = \sum_k c_k\, b^n_{k,\tau} \in S^n_\tau$$

be a parametrization of a uniform spline curve with grid width h. If new knots are simultaneously inserted at the midpoints of the knot intervals, the control points $\tilde{c}_k$ corresponding to the refined knot sequence $\tilde{\tau}$ with grid width $h/2$ and $\tilde{\tau}_{2k} = \tau_k$ can be computed as follows.

(i) The control points c_k, $k \sim \tau$, are doubled:

$$\tilde{c}_{2k} = \tilde{c}_{2k+1} = c_k\,.$$

(ii) Simultaneous averages of adjacent control points are formed,

$$\tilde{c}_k \leftarrow (\tilde{c}_k + \tilde{c}_{k-1})/2\,,$$

and this process is repeated n-times.

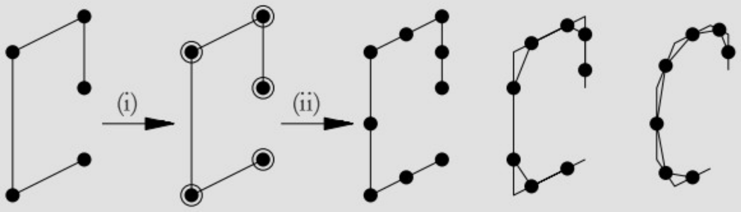

The explicit form of the new control points is

$$\tilde{c}_k = \sum_i s_{k-2i}\, c_i, \quad s_j = 2^{-n}\binom{n+1}{j},$$

where $s_j = 0$ for $j < 0$ or $j > n+1$ in accordance with the convention for binomial coefficients.

We note that for a finite knot sequence $\tau_0, \ldots, \tau_{m+n}$ the indices of the control points $\tilde{c}_k$ are not in the standard range $\{0, 1, \ldots\}$. By definition of $\tilde{\tau}$, the B-splines $b^n_{k,\tilde{\tau}}$, $k = n, \ldots, 2m-1$, are relevant for the parameter interval D^n_τ. We could have changed the definition of $\tilde{\tau}$. This would, however, result in less elegant subdivision formulas.

Turning to the derivation of the subdivision formula, it is sufficient to consider spline functions since the curve components are processed with identical operations. Moreover, in view of linearity, we need to check the identity only for a single B-spline b_ℓ. In this case, the coefficients are $c_k = \delta_{\ell-k}$ and the formula reduces to

$$\tilde{c}_k = s_{k-2\ell} = 2^{-n}\binom{n+1}{k-2\ell},$$

where $\tilde{c}_k$ are the coefficients for representing b_ℓ as a linear combination of the B-splines $\tilde{b}_k$ of the refined knot sequence $\tilde{\tau}$.

Using induction on the degree n, it is convenient to start with $n = 0$, although this case is excluded for splines. Since the B-splines of degree 0 are the characteristic functions of the half-open knot intervals and $\tilde{\tau}_{2\ell} = \tau_\ell$,

$$b^0_{\ell,\tau} = \tilde{b}^0_{2\ell,\tau} + \tilde{b}^0_{2\ell+1,\tau}.$$

This is in agreement with

$$\tilde{c}_{2\ell} = \tilde{c}_{2\ell+1} = 1, \quad \tilde{c}_k = 0,\, k \notin \{2\ell, 2\ell+1\},$$

as asserted by the subdivision formula.

For the induction step $n-1 \to n$, we have to show that

$$b^n_{\ell,\tau} = 2^{-n}\sum_k \binom{n+1}{k-2\ell} b^n_{k,\tilde{\tau}},$$

recalling that we are considering the special case $c_k = \delta_{\ell-k}$; i.e., the spline function $\sum_k c_k b^n_{k,\tau}$ is the single B-spline $b^n_{\ell,\tau}$. We differentiate the formula for subdividing $b^n_{\ell,\tau}$, using the simple recursion for the derivative of uniform B-splines. Since both sides vanish at $\tau_\ell = \tilde{\tau}_{2\ell}$, this yields the equivalent identity

$$B_\ell - B_{\ell+1} = 2^{1-n}\underbrace{\left[\sum_k \binom{n+1}{k-2\ell}(\tilde{B}_k - \tilde{B}_{k+1})\right]}_{S_1},$$

where we have set $B_k = b^{n-1}_{k,\tau}$ and $\tilde{B}_k = b^{n-1}_{k,\tilde{\tau}}$ to simplify notation. The factor 2^{-n} has changed since the differentiation formula involves the different factors h^{-1} and $(h/2)^{-1}$ for the B-splines on the left and right side, respectively. We now use the induction hypothesis and express the B-splines on the left side in terms of the B-splines $\tilde{B}_k$ corresponding to $\tilde{\tau}$. Noting that

$$B_{\ell+1}(x) = B_\ell(x-h), \quad \tilde{B}_k(x-h) = \tilde{B}_{k+2}(x),$$

this yields

$$2^{-(n-1)} \underbrace{\left[\sum_k \binom{n}{k-2\ell} (\tilde{B}_k - \tilde{B}_{k+2}) \right]}_{S_2} .$$

In both expressions $S_\nu = [\dots]$, we sum over $k \in \mathbb{Z}$, using the convention that $\binom{m}{j} = 0$ for $j < 0$ or $j > m$. The equality of these two sums follows by comparing the coefficients of $\tilde{B}_j$, which are

$$\binom{n+1}{j-2\ell} - \binom{n+1}{j-1-2\ell}$$

for S_1 and

$$\binom{n}{j-2\ell} - \binom{n}{j-2-2\ell}$$

for S_2. The expressions agree in view of the recursion for binomial coefficients:

$$\binom{n+1}{j-2\ell} = \binom{n}{j-2\ell} + \binom{n}{j-2\ell-1}, \quad \binom{n+1}{j-1-2\ell} = \binom{n}{j-1-2\ell} + \binom{n}{j-2-2\ell}.$$

It remains to derive the algorithmic definition of the coefficients $\tilde{c}_k$. Again, we can consider spline functions and make use of linearity. Hence, assuming that $c_k = \delta_{k-\ell}$, step (i) of the algorithm yields

$$\dots, \tilde{c}_{2\ell} = 1, 1 = \tilde{c}_{2\ell+1}, \dots,$$

where we list only nonzero coefficients. Applying step (ii) repeatedly, we obtain

$$\begin{array}{llllll} \tilde{c}_{2\ell} = & 1/2, & 2/2, & 1/2 & & \\ & 1/4, & 3/4, & 3/4, & 1/4 & \\ & 1/8, & 4/8, & 6/8, & 4/8, & 1/8 \\ & \dots . & & & & \end{array}$$

We recognize Pascal's triangle, i.e., the algorithm leads to the coefficients $2^{-n}\binom{n+1}{k-2\ell}$, as asserted by the subdivision formula.

☐ **Example:**

We apply the subdivision algorithm to the cubic closed spline curve with control points at the vertices $(\pm 1, \pm 1)$ of a square with side length 2.

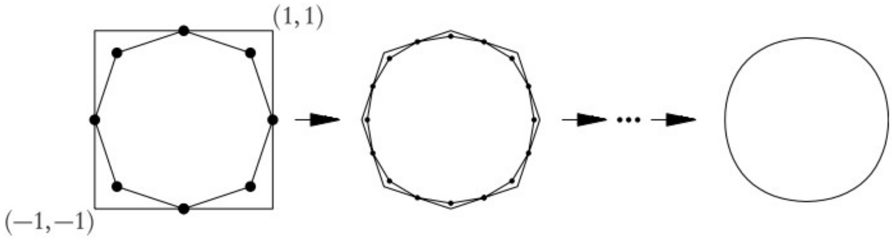

In this case, the subdivision rule is

$$\begin{array}{rclcl} \tilde{c}_{2\ell} & = & s_4 c_{\ell-2} + s_2 c_{\ell-1} + s_0 c_\ell & = & \frac{1}{8} c_{\ell-2} + \frac{3}{4} c_{\ell-1} + \frac{1}{8} c_\ell , \\ \tilde{c}_{2\ell+1} & = & s_3 c_{\ell-1} + s_1 c_\ell & = & \frac{1}{2} c_{\ell-1} + \frac{1}{2} c_\ell . \end{array}$$

Two steps of the corresponding refinements as well as the spline curve obtained in the limit are shown in the figure.

To analyze the iteration a little bit further, we derive an explicit formula for a control point sequence, emerging from one of the corners. To this end, we write the modification of three adjacent control points in matrix form:

$$\begin{pmatrix} \tilde{c}_{2\ell-1} \\ \tilde{c}_{2\ell} \\ \tilde{c}_{2\ell+1} \end{pmatrix} = \underbrace{\begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/8 & 3/4 & 1/8 \\ 0 & 1/2 & 1/2 \end{pmatrix}}_{A} \begin{pmatrix} c_{\ell-2} \\ c_{\ell-1} \\ c_\ell \end{pmatrix} .$$

The subdivision matrix has the eigenvalues 1, 1/2, and 1/4, and, by transformation to diagonal form, it can be shown that

$$A_\infty = \lim_{\ell\to\infty} A^\ell = \frac{1}{6} \begin{pmatrix} 1 & 4 & 1 \\ 1 & 4 & 1 \\ 1 & 4 & 1 \end{pmatrix} .$$

This proves, e.g., that the sequence of control points emerging from $(1,1)$ tends to the limit

$$\frac{1}{6} \begin{pmatrix} 1 & 4 & 1 \end{pmatrix} \begin{pmatrix} -1 & 1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix} = \frac{2}{3} \begin{pmatrix} 1 & 1 \end{pmatrix} ,$$

which is a point on the curve. □

As a more theoretical application of knot insertion, we discuss a measure for the number of oscillations of a spline curve.

Variation Diminution

The variation of a spline curve parametrized by $p = \sum_{k=0}^{m-1} c_k \, b_k$ with respect to a hyperplane H is not larger than the variation of its control polygon c:

$$V(p,H) \le V(c,H),$$

where V denotes the maximal number of pairs of consecutive points on opposite sides of H.

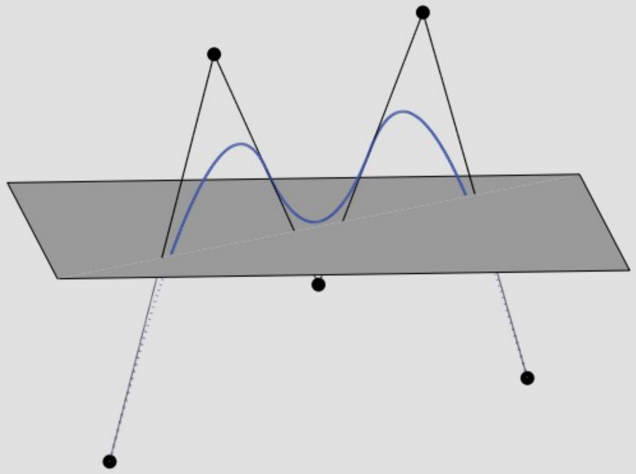

In particular, if the entire control polygon lies on one side of H, so does the spline curve.

This important property of spline curves can be proved by elegant geometric arguments, as was observed by Lane and Riesenfeld.

Since the spline curve consists of finitely many polynomial segments, its variation is finite. Hence, in the nontrivial case, when $V(p,H) > 0$, there exist pairs of points

$$p(t_{k-1}),\ p(t_k), \quad k = 1, \ldots, V(p,H),$$

lying on opposite sides of H. These points coincide with control points of a refined control polygon $\tilde{c}$ having the parameters t_k as knots of multiplicity n. Hence, it remains to show that knot insertion does not increase the variation since this implies

$$V(p,H) \le V(\tilde{c},H) \le V(c,H).$$

Without going into details, knot insertion repeats the following two steps (not necessarily in any particular order):

(i) placing an additional control point on an edge of the control polygon (splitting edges);

(ii) deleting a control point (cutting corners).

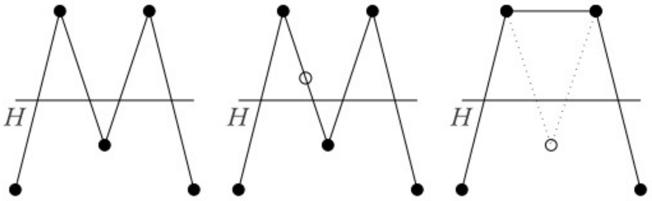

(i) splitting edges (ii) cutting corners

For example, when inserting an interior knot for a cubic spline curve, we place three points on adjacent edges of the control polygon and then delete two of the intermediate vertices. Obviously, neither of the two steps illustrated in the figure increases the variation.

□ **Example:**

Variation diminution can be used to locate the intersections of a spline curve p with a hyperplane H. To this end, we represent p as a union of Bézier segments, recalling that Bernstein polynomials are particular B-splines (cf. Section 4.1). After this preprocessing step, p is represented by a list of polynomial segments p^i, each described by its Bézier control polygon c^i. We now employ an elementary bisection strategy based on the variations

$$V^i = V(c^i, H),$$

which are upper bounds for variations of the Bézier segments.

If V^i is zero, we check whether accidentally one or both of the endpoints of p^i are points of intersection. Since $V(p^i, H) = 0$, intersections in the interior of the parameter interval for p^i are not possible, and we delete p^i from the list.

If $V^i \geq 1$ and the diameter of a bounding box for c^i is $>$ tol, we subdivide p^i at the midpoint of its parameter interval and append both parts in place of p^i to the list.

Upon termination of the subdivision procedure, there are two possibilities for the remaining Bézier segments p^i.

(i) If $V^i = 1$, then the endpoints of c^i, which coincide with the endpoints of p^i, must lie on opposite sides of H. Hence, an intersection of p^i with H exists. Since $V(p^i, H) \leq V^i$, this intersection is unique. Its location is determined by the bounding box for c^i with accuracy tol.

(ii) If $V^i > 1$, the existence of an intersection can be neither guaranteed nor excluded. A more subtle test is needed. This is not surprising since the case of more than one alternation includes the possibility of points of tangency, a numerically critical situation. Fortunately, for small tolerances, Bézier segments with $V^i > 1$ and small bounding boxes are relatively rare.

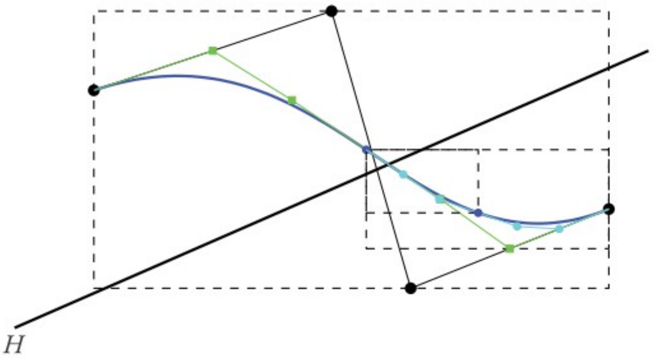

The figure illustrates the subdivision strategy for a planar cubic Bézier segment. Two subdivision steps are performed. We see that the bounding boxes provide an increasingly better approximation to the intersection. When subdivision is terminated, a Newton iteration can be started to determine the exact location. □

6.4 ▪ Algorithms

The algorithms for spline functions easily generalize to curves. We begin by considering the two principal operations.

Evaluation and Differentiation

A point

$$p(s) = \sum_{k=0}^{m-1} c_k\, b_k(s)$$

on a spline curve with knot sequence $\tau : \tau_0, \dots, \tau_{m+n}$ can be computed by repeatedly inserting s as a new knot until its multiplicity becomes n:

$$\tilde{\tau}_\ell < \tilde{\tau}_{\ell+1} = \cdots = \tilde{\tau}_{\ell+n} = s < \tilde{\tau}_{\ell+n+1} \implies p(s) = \tilde{c}_\ell\,,$$

where $\tilde{\tau}_\ell$ and $\tilde{c}_k$ denote the modified knots and control points, respectively.

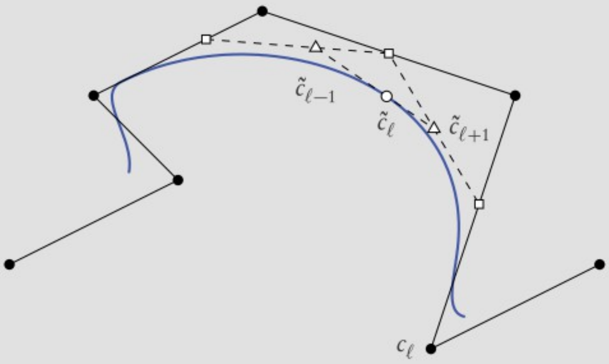

The refined control polygon $\tilde{c}$ is tangent to the curve:

$$p'(s^-) = \frac{n(\tilde{c}_\ell - \tilde{c}_{\ell-1})}{s - \tilde{\tau}_\ell}, \quad p'(s^+) = \frac{n(\tilde{c}_{\ell+1} - \tilde{c}_\ell)}{\tilde{\tau}_{\ell+n+1} - s},$$

where the one-sided derivatives coincide if s is not a knot with multiplicity n of the original knot sequence τ (i.e., if at least one knot is inserted). In this case,

$$p'(s) = \frac{n}{\tilde{\tau}_{\ell+n+1} - \tilde{\tau}_\ell} \left(\tilde{c}_{\ell+1} - \tilde{c}_{\ell-1}\right)$$

is an alternative formula for the tangent vector.

If $s \in (\tilde{\tau}_\ell, \tilde{\tau}_{\ell+n+1})$ is a knot with multiplicity n of $\tilde{\tau}$, the B-splines b_k with $k < \ell$ vanish for parameters $t \geq s$. Hence, $p(s)$ can be regarded as a left endpoint of the spline curve with control points $\tilde{c}_\ell$, $\tilde{c}_{\ell+1}$, ... The formulas for $p(s)$ and $p'(s^+)$ are thus a direct consequence of endpoint interpolation. Similarly, the formula for $p'(s^-)$ follows by regarding $p(s)$ as a right endpoint of the spline curve with control points $\dots, \tilde{c}_{\ell-1}, \tilde{c}_\ell$.

If the parametrization is continuously differentiable, $p'(s^-) = p'(s^+)$. Solving this equation for the control point $\tilde{c}_\ell$ and substituting the resulting expression in either of the formulas for the one-sided derivatives yields the formula for $p'(s)$ in terms of $\tilde{c}_{\ell-1}$ and $\tilde{c}_{\ell+1}$.

□ **Example:**

We consider the cubic spline curve p with control points

$$C = \begin{pmatrix} 10 & 4 & 0 & 20 & 20 \\ 10 & 4 & 20 & 20 & 0 \end{pmatrix}^{\mathrm{t}},$$

knots

$$\tau_0 = 0,\ 1,\ 1,\ 1,\ 2,\ 5,\ 6,\ 7,\ 8 = \tau_8,$$

and parameter interval $D^3_\tau = [1,5]$.

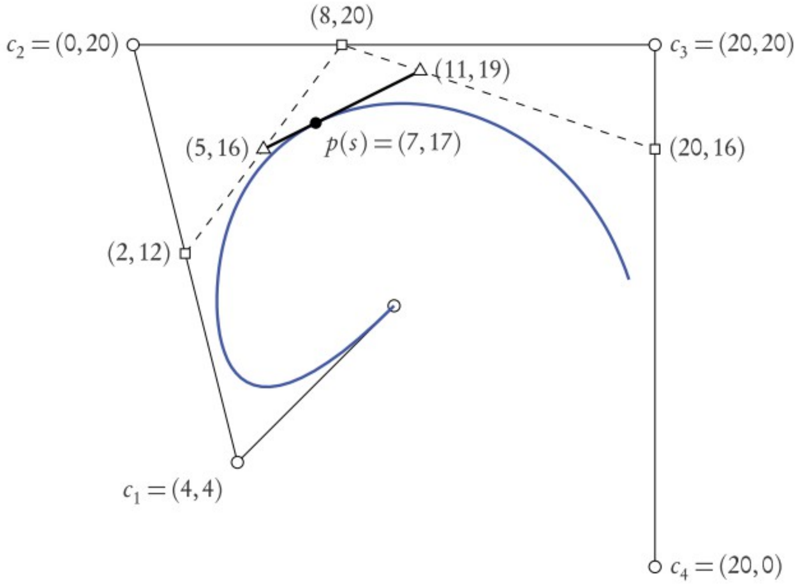

As is illustrated in the figure, we determine the point $p(s)$ and the tangent vector $p'(s)$ for $s = 3 \in [2,5)$. To this end, we insert s three times as new knot. Starting with the relevant control points $c_1, \ldots, c_4$ for the knot interval $[2,5)$, the new control points are generated with a triangular scheme. After having inserted s twice, we can compute

$$p'(3) = \frac{3}{5-2}((11,19) - (5,16)) = (6,3).$$

The point on the curve is then obtained by forming one last convex combination:

$$p(3) = \tilde{c}_4 = \frac{3-2}{5-2}(11,19) + \frac{5-3}{5-2}(5,16) = (7,17),$$

where the weights correspond to the relative position of s in the knot interval $[2,5)$. □

It is often possible to process the polynomial segments of a spline representation simultaneously. This can be done most efficiently by converting to Bézier form, using repeated knot insertion.

Bézier Form

The Bézier form of a spline curve parametrized by $p = \sum_{k=0}^{m-1} c_k\, b_k \in S_\tau^n$ is obtained by raising the multiplicity of each knot τ_k in the parameter interval $D_\tau^n = [\tau_n, \tau_m]$ to n. Then, for t in a nondegenerate parameter interval $[\tilde{\tau}_\ell, \tilde{\tau}_{\ell+1}] \subseteq D_\tau^n$ of the refined knot sequence $\tilde{\tau}$,

$$p(t) = \sum_{k=0}^{n} \tilde{c}_{\ell-n+k}\, b_k^n(s), \quad s = \frac{t - \tilde{\tau}_\ell}{\tilde{\tau}_{\ell+1} - \tilde{\tau}_\ell} \in [0,1],$$

where b_k^n are the Bernstein polynomials and $\tilde{c}_k$ the control points with respect to $\tilde{\tau}$. Hence, up to linear reparametrization (which is immaterial for the shape of the curve), the spline segments have Bézier form.

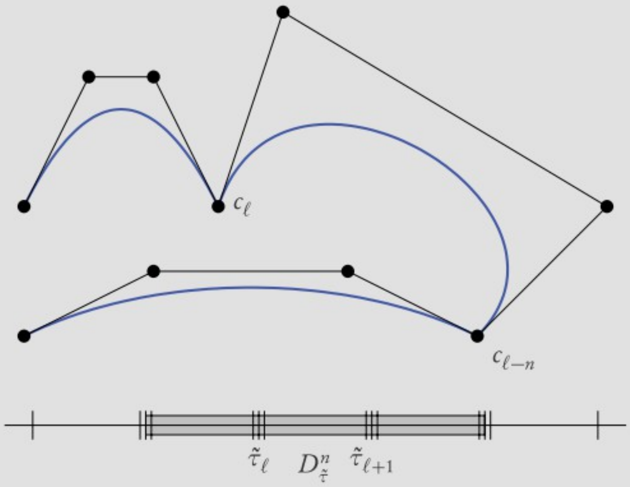

As shown in the figure, every nth control point lies on the curve separating the Bézier segments. Thus, by converting to Bézier form, we can apply polynomial algorithms simultaneously on the different knot intervals.

We already remarked in Section 4.1 that Bernstein polynomials are particular B-splines. If $\tilde{\tau}$ has n-fold knots at both endpoints of a knot interval $D_\ell = [\tilde{\tau}_\ell, \tilde{\tau}_{\ell+1}]$, i.e., if

$$\tilde{\tau}_{\ell-n} < \tilde{\tau}_{\ell-n+1} = \cdots = \tilde{\tau}_\ell < \tilde{\tau}_{\ell+1} = \cdots = \tilde{\tau}_{\ell+n} < \tilde{\tau}_{\ell+n+1},$$

then the relevant B-splines $\tilde{b}_k$ for D_ℓ ($k = \ell - n, \ldots, n$) coincide with the transformed Bernstein polynomials on D_ℓ. As a consequence of the B-spline recursion,

$$\tilde{b}_{\ell-n}(t) = \left(\frac{\tilde{\tau}_{\ell+1} - t}{\tilde{\tau}_{\ell+1} - \tilde{\tau}_\ell} \right)^n = b_0^n(s), \quad \tilde{b}_\ell(t) = \left(\frac{t - \tilde{\tau}_\ell}{\tilde{\tau}_{\ell+1} - \tilde{\tau}_\ell} \right)^n = b_n^n(s)$$

for $t \in D_\ell \Leftrightarrow s \in [0,1]$. We note that these two B-splines have support also on the neighboring intervals, thus ensuring continuity of the curve at the junction points. For $0 < i < n$, the B-spline $\tilde{b}_{\ell-n+i}$ has an $(n+1-i)$-fold knot at τ_ℓ and an $(i+1)$-fold knot at $\tau_{\ell+1}$ and, thus, $\tilde{b}_{\ell-n+i}(t) = b_i^n(s)$; those B-splines are supported entirely on D_ℓ.

□ **Example:**

As an example, we consider the closed quadratic spline curve with control points and extended knot sequence shown in the figure.

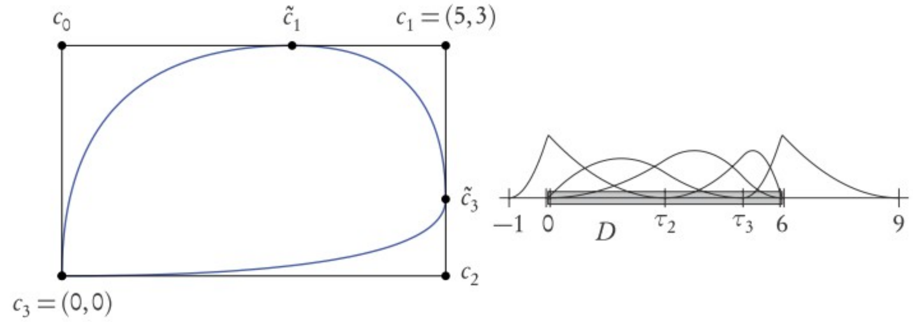

To convert to Bézier form, we double the knots $\tau_2 = 3$ and $\tau_3 = 5$, taking into account the periodicity of the extended partition and the control points:

$$\tau_{k+4} = \tau_k + 6, \quad c_{k+4} = c_k\,.$$

Applying the knot insertion algorithm first for $s = \tau_2$, only the index $k = 1$ satisfies

$$\tau_k < s = 3 < \tau_{k+2}\,.$$

Hence, inserting $s = 3$ yields only one new control point, namely,

$$\tilde{c}_1 = \frac{3}{5}(5,3) + \frac{2}{5}(0,3) = (3,3)\,.$$

Similarly, inserting $s = 5$, we compute the control point

$$\tilde{c}_3 = \frac{2}{3}(5,0) + \frac{1}{3}(5,3) = (5,1)$$

and obtain the Bézier control polygon

$$\tilde{c}_0 = c_0,\ \tilde{c}_1,\ \tilde{c}_2 = c_1,\ \tilde{c}_3,\ \tilde{c}_4 = c_2,\ \tilde{c}_5 = c_3\,.$$

We notice that the control points on the curve, which separate the spline segments, lie on the initial control polygon which touches the curve at these points. This is a particular feature of quadratic spline curves. The conversion is very simple in this case. The Bézier control points separating the curve segments at a simple knot τ_ℓ are computed by dividing the segment $[c_{\ell-2}, c_{\ell-1}]$ in the ratio

$$(\tau_\ell - \tau_{\ell-1}) : (\tau_{\ell+1} - \tau_\ell).$$

At a double knot $\tau_\ell = \tau_{\ell+1}$, no computations are necessary; $c_{\ell-1}$ is already a separating Bézier control point. □

As a further illustration, we consider the conversion for uniform knots. In this special case, the knot insertion procedure has an intriguing description via a combinatoric labeling scheme.

□ **Example:**

For spline curves parametrized by $p = \sum_k c_k b_k \in S^n_\tau$ with uniform knots τ_k the conversion to Bézier form has a beautiful geometric interpretation due to Ramshaw and is referred to as blossoming. It is illustrated in the figure below for a parametrization of degree $n = 4$.

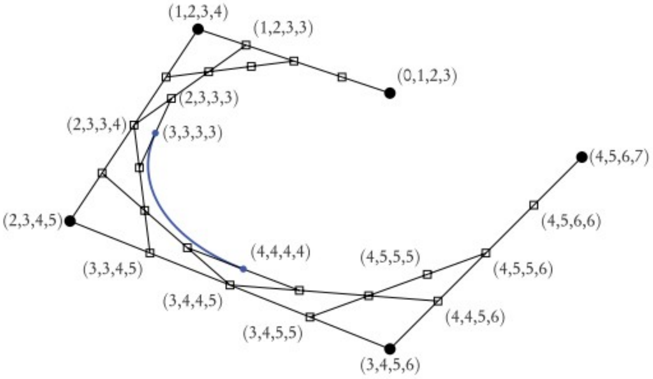

We label the control points by the indices of the inner knots of the corresponding B-splines. For example, c_3 has the label $(4,5,6,7)$ since b_3 has support $[\xi_3, \xi_8]$. Then, we connect pairs of points with labels

$$(a,b,c,d), \quad (b,c,d,e)$$

which have $n-1 = 3$ indices in common. On the resulting line segment, we place $e-a-1$ equally spaced new points and label them with

$$(a+1,b,c,d), \dots, (e-1,b,c,d),$$

reordering the indices if necessary. The process is repeated, at each stage utilizing the newly generated points, until all possible connections have been formed. In the final configuration, the Bézier segments consist of all points with labels, which have at most two different indices. In particular, the Bézier endpoints have labels with just one index of multiplicity $n = 4$. □

6.5 ▪ Interpolation

Interpolation methods for functions can be applied to curves in an obvious way. Each component is processed separately. Below, we list the most commonly used schemes.

Interpolation

Points p_k and tangent vectors d_k (if provided) can be interpolated with a spline curve at parameter values t_k, using any of the interpolation methods for spline functions. The univariate schemes are applied separately in each component to determine the components of the parametrization $p = \sum_k c_k b_k$. Standard choices are cubic Hermite interpolation and cubic spline interpolation with not-a-knot, natural, or clamped boundary conditions.

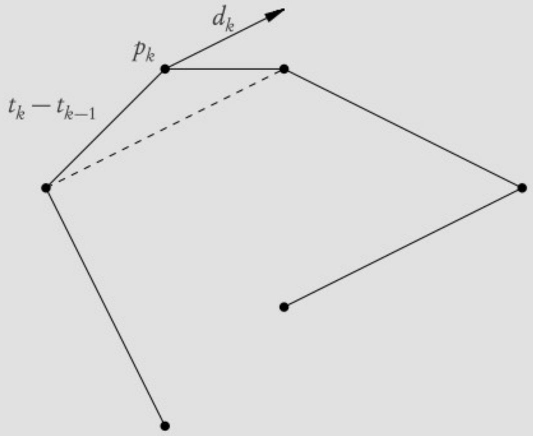

If only points are given, knots τ_j, parameter values t_k, and tangent vectors d_k (if required) have to be determined based on the available information. Basic choices are

- $t_k - t_{k-1} = |p_k - p_{k-1}|$;
- $t_k = \tau_{k+\ell}$ with the shift ℓ depending on the labeling of the knots;
- $d_k = (p_{k+1} - p_{k-1})/(t_{k+1} - t_{k-1})$.

More accurate derivative approximations employ local polynomial interpolation. The resulting formulas are used in particular at the endpoints of the parameter interval, where one-sided approximations are needed.

As a first example we consider cubic Hermite interpolation. While the interpolant is not curvature continuous, it can be constructed with a few simple operations without having to solve a linear system.

☐ **Example:**

If tangent vectors d_j are given at all points p_j, we can interpolate each pair of consecutive data with a cubic Bézier segment. By construction, these segments form a spline curve in Bézier form. In particular, two adjacent segments share a common control point lying on the curve.

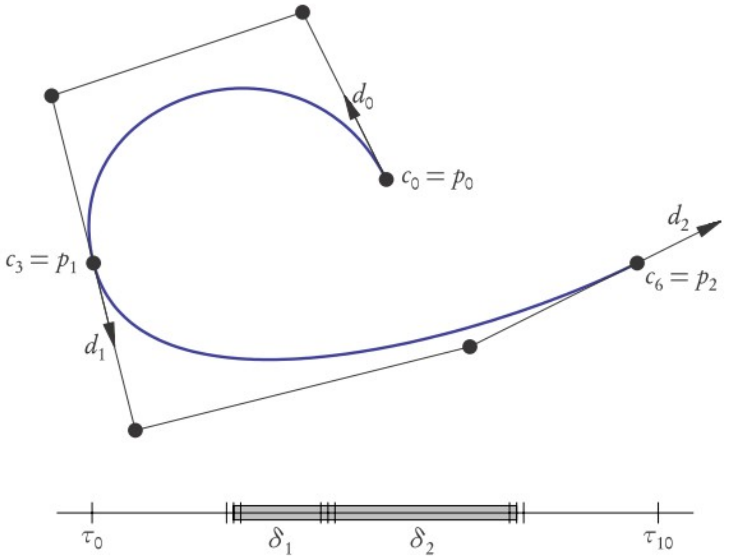

Considering, e.g., the interpolation by the segment interpolating p_0 and p_1, we choose the length of the parameter interval as $\tau_4 - \tau_3 = |p_1 - p_0| = \delta_1$. This means that the parametrization has the form

$$p(t) = \sum_{k=0}^{3} c_k\, b_k^3(s), \quad s = \frac{t - \tau_3}{\tau_4 - \tau_3},$$

where b_k^3 are the cubic Bernstein polynomials. Accordingly, the interpolation conditions imply

$$p_0 = c_0, \quad p_1 = c_3,$$

and

$$d_0 = \frac{3}{\delta_1}(c_1 - c_0), \quad d_1 = \frac{3}{\delta_1}(c_3 - c_2).$$

Solving the last equations for the two middle control points finally yields

$$c_1 = p_0 + \frac{\delta_1}{3} d_0, \quad c_2 = p_1 - \frac{\delta_1}{3} d_1. \qquad \square$$

Very popular interpolation schemes, also for curves, are natural spline interpolation and its variants. The implementation of these methods is discussed in the following example.

□ **Example:**

To interpolate points $p_0, \ldots, p_M$ by a natural cubic spline curve, we first make a standard selection of the parameter values. We choose t_k so that

$$t_{k+1} - t_k = |p_{k+1} - p_k|, \quad k = 0, \ldots, M-1.$$

In $D = [t_0, t_M]$, the knots $\tau_k = t_{k-3}$, $k = 3, \ldots, m$, $m = M + 3$, are used. Outside of D, we choose three equally spaced knots on each side, keeping the distance of the first and last knot interval. We can now apply the univariate scheme with the boundary conditions $p''(t_0) = p''(t_M) = 0$.

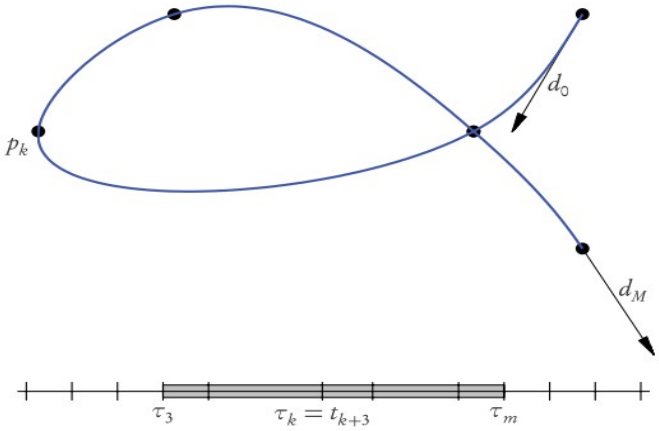

A more accurate approximation is obtained by using either clamped or not-a-knot boundary conditions. In the first case, derivatives are prescribed at the endpoints:

$$p'(t_0) = d_0, \quad p'(t_M) = d_M .$$

The not-a-knot conditions mean that the knots $\tau_4 = t_1$ and $\tau_{m-1} = t_{M-1}$ are removed. As a result, we interpolate using the B-splines $\tilde{b}_0, \ldots, \tilde{b}_M$ corresponding to the reduced knot sequence

$$\tilde{\tau}: \ \tau_0 < \cdots < \tau_3 = t_0 < \tau_5 < \tau_6 < \cdots < \tau_{m-3} < \tau_{m-2} < t_M = \tau_m < \cdots < \tau_{m+3}$$

Hence, the dimension $\tilde{m} = M + 1$ of the spline space $S^3_{\tilde{\tau}}$ matches the number of interpolation conditions and, as implied by the Schoenberg–Whitney condition, the interpolation problem is well posed. □

.
! *+,-* * *. ! * I ! <J F KJ <J + !3
" >J D <J ! ! B 2 2" L
3 *+,-* * *. ! * <J F I ! F <J F KJ
+ !3 " >J D <J ! ! B 2 2" L
+,- * *. ! * I ! <J F KJ <J + !3
" >J D <J ! ! B $ 2 A < " 3 M 2 2 B
) 2" L $

$ ')
! *+,-* * *. ! * $ 6 !+/ ! ! 2"
.3 & 2 ! NNN

3 *+,-* * *. ! * $ + 3 !> ! 2"

+,- * *. ! * $! 2"

" *+,-* * *. ! * $: ! "!" 2"

+,- * *. ! * $ 6 !+ ! ! 2"

+,- * *. ! * $ 6 !> ! 2"
' ! ! ! 2 !

+,- * *. ! * $ 4 3 2 / 2"

+,- * *. ! * $. 2 !"! 2"

+,- * *. ! * $ 3 ! A! 6 ! 2"

? *+,-* * *. ! * $ 3 ! 6 ! 2"
.3 > 7 ! NNN > " (2!

B *+,-* * *. ! * $ 3 ! 8 ! 2"

njwildberger.com/2015/12/02/the-alexander-horned-sphere-is-it-nonsense/

20210703 (20210703)

gauntlet = manopla.

A Cornucopia of Quadrilaterals - MAA Press - Vol 45

The midpoints of the sides of any quadrilateral are the vertices of a parallelogram, or that in a convex quadrilateral (not a parallelogram) the line through the midpoints of the diagonals (the Newton line) is equidistant from opposite vertices, or that, if your quadrilateral has an inscribed circle, its center lies (conversely) on the Newton line.

There are results dating back to Euclid: the side-lengths of a pentagon, a hexagon, and a decagon inscribed in a circle can be assembled into a right triangle (the proof uses a quadrilateral and circumscribing circle); and results dating to Erdős: from any point in a triangle the sum of the distances to the vertices is at least twice as large as the sum of the distances to the sides.

! *+,-* * *. ! * (+ 2 ! - !" ! !
+ ! ! ! ! 0 2"

3 *+,-* * *. ! * (! " + ! + 2 !
- !" ! ! + ! ! ! ! 0 2"

+,- * *. ! * (! " + ! ! !" !
! & ! ; " ! " ! ! ! ! ! + ! + ! 9 2"

" *+,-* * *. ! * (! " + ! ! !" !
! & ! ; " ! " ! ! 8 ! ! ! + ! + !
9 2"

+,- * *. ! * (& ! ! +
5/2 ! 8 0 I 0 & ! ! ! ! + ! + !M! ! ! + !
+ ! 2"
' $ (

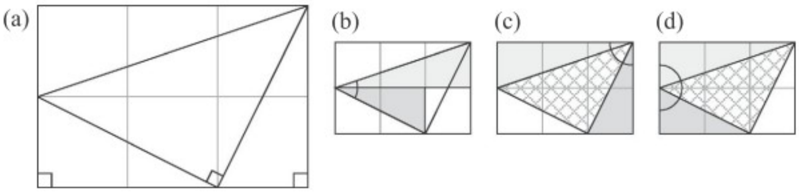

Figure 3.2.

The solution follows immediately from the Garfield trapezoid in Figure 3.1. The upper edge of the trapezoid has length $c\sqrt{2}$ (from the Pythagorean theorem just proven), and is at least as long as the base $a+b$. The two are equal if and only if the upper and lower edges are parallel, which occurs if and only if $a=b$.

If we divide both sides of (3.1) by 2 and recall that $c=\sqrt{a^2+b^2}$ we have

$$\frac{a+b}{2} \le \sqrt{\frac{a^2+b^2}{2}} \tag{3.2}$$

with equality if and only if $a=b$. The term on the right side in (3.2) is the root mean square, which we encountered in Challenge 2.5, and (3.2) is the arithmetic mean-root mean square inequality.

3.3 Trigonometric formulas and identities

Since the Garfield trapezoid is constructed from three triangles that share sides, it is remarkably well suited for illustrating a variety of trigonometric formulas and identities. In this section we often adjoin a fourth triangle to the trapezoid, and use the resulting rectangle to explore relationships among the sides and angles of the triangles in the configuration.

We begin by setting $a=1$ and $b=2$ in Figure 3.1 and embedding the icon in a 2-by-3 rectangle, as illustrated in Figure 3.2a. This enables us to evaluate certain angles marked $\measuredangle$ in the shaded triangles as arctangents and sums of arctangents.

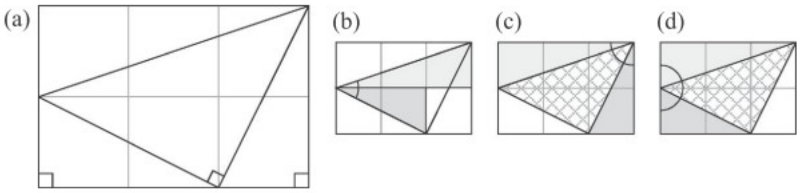

Figure 3.2.

From the marked angles in Figures 3.2b, 3.2c, and 3.2d we obtain

$$\begin{aligned} &\arctan(1/2)+\arctan(1/3)=\pi/4, \\ &\arctan(1)+\arctan(1/2)+\arctan(1/3)=\pi/2, \text{ and} \\ &\arctan(1)+\arctan(2)+\arctan(3)=\pi\,\text{[Wu, 2003]}. \end{aligned} \tag{3.3}$$

The argument can be easily modified to derive *Euler's arctangent identity*: For positive numbers p and q,

$$\arctan(1/p) = \arctan(1/(p+q)) + \arctan(q/(p^2+pq+1)). \quad (3.4)$$

See Figure 3.3 [Wu, 2004], and observe that $\alpha = \beta + \gamma$ where $\alpha = \arctan(1/p)$, $\beta = \arctan(1/(p+q))$, and $\gamma = \arctan(q/(p^2+pq+1))$.

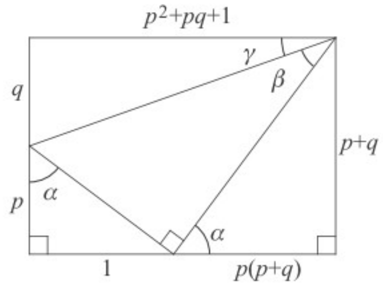
p^2+pq+1
γ
β
q
$p+q$
p
α
α
1
$p(p+q)$

Figure 3.3.

When $p = q = 1$, we obtain (3.3) as a special case of (3.4). See Challenges 3.3 and 3.4 for further arctangent identities.

The usual way to evaluate the trigonometric functions of 15° and 75° is with addition and subtraction formulas, e.g., $\sin(15°) = \sin(45° - 30°)$ or $\sin(60° - 45°)$, $\tan(75°) = \tan(30° + 45°)$, etc. However, they are readily evaluated geometrically using a modified Garfield trapezoid and a procedure adapted from [Hoehn, 2004]. In Figure 3.4a we construct the trapezoid from two isosceles right triangles, one with sides 1, 1, and $\sqrt{2}$ and the other with sides $\sqrt{3}$, $\sqrt{3}$, and $\sqrt{6}$. Since the sides containing the right angle in the gray triangle in Figure 3.4a are $\sqrt{2}$ and $\sqrt{6}$, the gray triangle has acute angles measuring 30° and 60°.

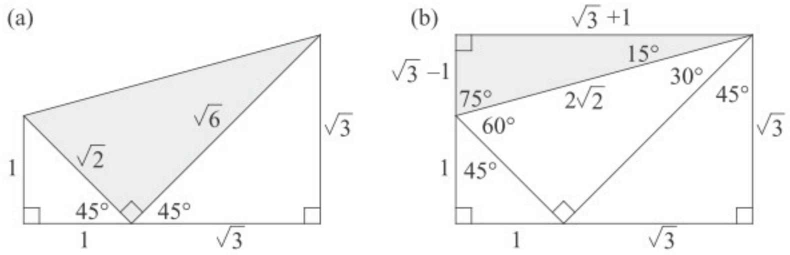
(a)
$\sqrt{6}$
$\sqrt{3}$
1
$\sqrt{2}$
45°
45°
1
$\sqrt{3}$
(b)
$\sqrt{3}+1$
$\sqrt{3}-1$
15°
30°
75°
$2\sqrt{2}$
45°
60°
$\sqrt{3}$
1
45°
1
$\sqrt{3}$

Figure 3.4.

Adjoin a right triangle to the top of the trapezoid to form the rectangle in Figure 3.4b. The acute angles of the adjoined triangle (shaded gray) are 15° and 75°; the sides are $\sqrt{3}-1$, $\sqrt{3}+1$, and $2\sqrt{2}$; and hence

$$\sin 15^\circ = \frac{\sqrt{3}-1}{2\sqrt{2}}, \tan 75^\circ = \frac{\sqrt{3}+1}{\sqrt{3}-1}, \text{etc.}$$

The procedure can be modified to illustrate the addition and subtraction formulas for the sine, cosine, and tangent, at least for small positive angles. Let α and β be acute angles whose sum is less than $\pi/2$, and arrange two triangles with acute angle α and one with acute angle β as shown in Figure 3.5.

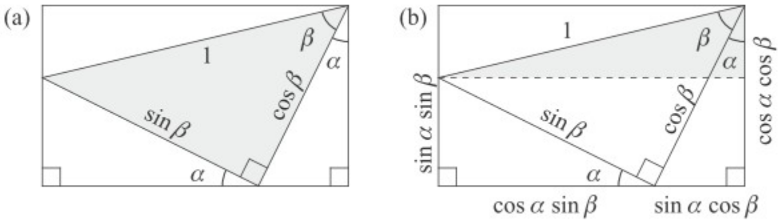

Figure 3.5.

If we let the sides of the gray triangle in Figure 3.5a be $\sin\beta$, $\cos\beta$, and 1, then it is easy to compute the lengths of the sides of the triangles with acute angle α, as shown in Figure 3.5b. Evaluating the lengths of the sides of the gray triangle in Figure 3.5b yields

$$\sin(\alpha+\beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta,$$
$$\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta.$$

The other formulas can be similarly illustrated, see Challenges 3.5 and 3.6.

We conclude this section with two little known identities for sums and products of tangents, again illustrated with modified Garfield trapezoids. In Figure 3.6 we show that for positive angles α, β, and γ such that $\alpha+\beta+\gamma=\pi$ (for instance, the angles in a triangle), we have

$$\tan\alpha + \tan\beta + \tan\gamma = \tan\alpha\tan\beta\tan\gamma.$$

The identity results from equating the two expressions for the vertical height of the rectangle.

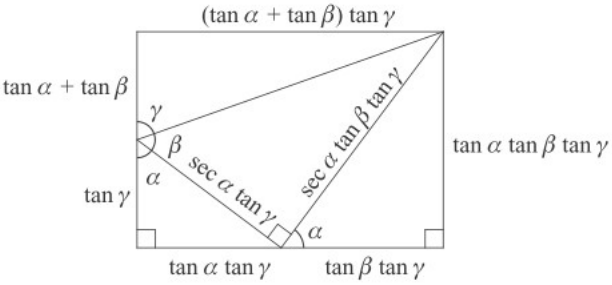
$(\tan\alpha + \tan\beta)\tan\gamma$
$\tan\alpha + \tan\beta$
γ
β
α
$\tan\gamma$
$\sec\alpha\tan\gamma$
$\sec\alpha\tan\beta\tan\gamma$
$\tan\alpha\tan\beta\tan\gamma$
α
$\tan\alpha\tan\gamma$
$\tan\beta\tan\gamma$

Figure 3.6.

In Figure 3.7 we see a complementary result: for positive angles α, β, and γ such that $\alpha + \beta + \gamma = \pi/2$, we have

$$\tan\alpha\tan\beta + \tan\beta\tan\gamma + \tan\gamma\tan\alpha = 1.$$

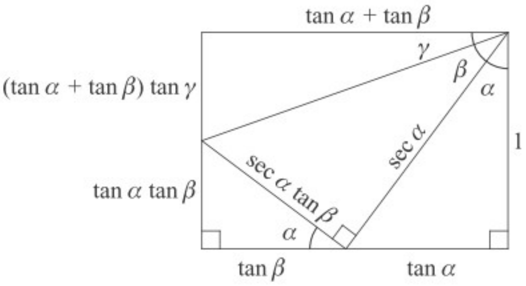
$\tan\alpha + \tan\beta$
γ
β
α
$(\tan\alpha + \tan\beta)\tan\gamma$
1
$\sec\alpha$
$\sec\alpha\tan\beta$
$\tan\alpha\tan\beta$
α
$\tan\beta$
$\tan\alpha$

Figure 3.7.

3.4 Challenges

3.1. Use Garfield's trapezoid to show that for any θ, $|\sin\theta + \cos\theta| \leq \sqrt{2}$.

3.2. Use Garfield's trapezoid to illustrate the AM-GM inequality.

3.3. Let $0 < a < b$. Use a diagram similar to Figure 3.2a to show

$$\arctan(a/b) + \arctan((b-a)/(b+a)) = \pi/4.$$

3.4. Let $\{F_n\}_{n=1}^{\infty}$ denote the sequence of Fibonacci numbers, defined by $F_1 = F_2 = 1$ and $F_n = F_{n-1} + F_{n-2}$ for $n \geq 3$. Prove

$$\arctan\frac{1}{F_{2n}} = \arctan\frac{1}{F_{2n+1}} + \arctan\frac{1}{F_{2n+2}}$$

for $n \geq 1$. [Hint: Use Euler's arctangent identity and *Cassini's identity* for the Fibonacci numbers: $F_{k-1}F_{k+1} - F_k^2 = (-1)^k$ for $k \geq 2$. See Challenge 18.2 for a proof of Cassini's identity.]

3.5. Use a diagram similar to Figure 3.4 to illustrate the subtraction formulas for the sine and cosine:

$$\sin(\alpha - \beta) = \sin\alpha\cos\beta - \cos\alpha\sin\beta,$$
$$\cos(\alpha - \beta) = \cos\alpha\cos\beta + \sin\alpha\sin\beta.$$

3.6. Use diagrams similar to Figure 3.4 to illustrate the addition and subtraction formulas for the tangent:

$$\tan(\alpha + \beta) = \frac{\tan\alpha + \tan\beta}{1 - \tan\alpha\tan\beta} \quad \text{and} \quad \tan(\alpha - \beta) = \frac{\tan\alpha - \tan\beta}{1 + \tan\alpha\tan\beta}.$$

3.7. Let a, b, c be the sides of a right triangle, with c the hypotenuse. Show that three right triangles (each similar to the given triangle) with sides a^2, ab, ac; ab, b^2, bc; and ac, bc, c^2; can be arranged to form a Garfield trapezoid that is actually a rectangle, yielding another proof of the Pythagorean theorem.

3.8. An integration technique often presented in calculus is the *Weierstrass substitution*, useful for integrals of rational functions of the sine and cosine. The change of variables $z = \tan(\theta/2)$ yields rational functions of z for $\sin\theta$ and $\cos\theta$. Use the Garfield trapezoid in Figure 3.8 to evaluate $\sin\theta$ and $\cos\theta$ as functions of z.

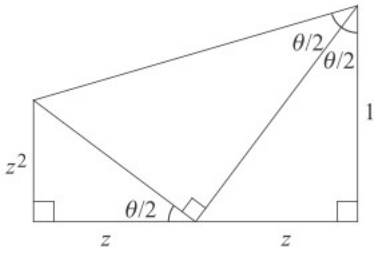

Figure 3.8.

3.9. Given $a, b > 0$ consider for any $k > 0$ the trapezoid in Figure 3.9, where P_k is the midpoint of the slanted edge. Describe the position of P_k as a function of k.

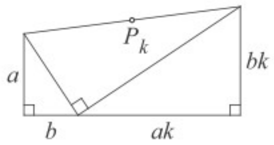

Figure 3.9.

3.10. It is an immediate consequence of the AM-GM inequality (see Section 2.2 or Challenge 3.2) that for x positive, $x + (1/x) \geq 2$. Show that this inequality can be improved to

$$x + \frac{1}{x} \geq \frac{1}{2}\left(\sqrt{x} + \frac{1}{\sqrt{x}}\right)^2 \geq 2. \tag{3.5}$$

CHAPTER 4

The Semicircle

The eye is the first circle; the horizon which it forms is the second; and throughout nature this primary figure is repeated without end. It is the highest emblem in the cipher of the world.

Ralph Waldo Emerson, *Essays*

The semicircle has long played an important role in architecture and art. The Romans employed semicircular arches in their constructions such as the Pont du Gard aqueduct near Nîmes, France (see Figure 4.1a). Similar arches can be found in Romanesque architecture throughout Europe. The Moors also used semicircular arches in their most impressive buildings, such as in the interior of the Mezquita mosque in Córdoba, Spain (see Figure 4.1b). Semicircles also appear in paintings of modern artists such as Wassily Kandinsky (*Semicircle*, 1927) and Robert Mangold (*Semi-Circle I–IV*, 1995).

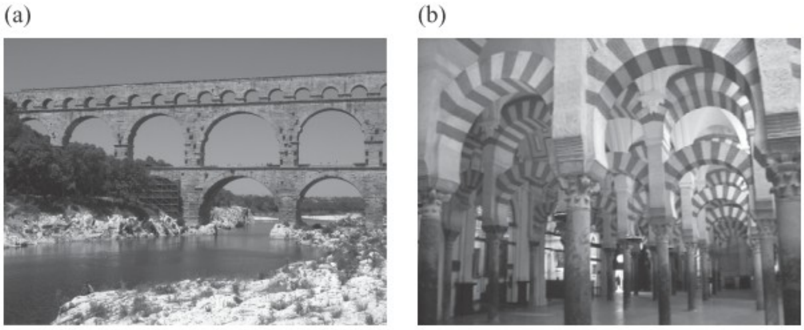
(a) (b)

Figure 4.1.

The icon for this chapter is a semicircle, simple and easily constructed with straightedge and compass but nonetheless remarkably useful in geometry and trigonometry. It figures prominently in the work of Greek geometers such as Thales, Euclid, Archimedes, Hippocrates, and Pappus. The semicircle also appears in the solution of the first optimization problem in literature, Dido's problem. We also encounter it in visual demonstrations of a variety of trigonometric identities, and in an exploration of the five Platonic solids inscribed in a sphere.

4.1 Thales' triangle theorem

Perhaps the first occurrence of the semicircular icon in geometry is in a theorem named after Thales of Miletus (circa 624–546 BCE), one of the earliest of the great Greek mathematicians. The theorem also appears as Proposition 31 in Book III of Euclid's *Elements*. There are two theorems commonly called "Thales' theorem"—one about a right angle inscribed in a semicircle, which we call *Thales' triangle theorem*, and another about similar triangles, which we call *Thales' proportionality theorem* and present in Chapter 5.

Thales' triangle theorem. *A triangle inscribed in a semicircle is a right triangle.* See Figure 4.2a.

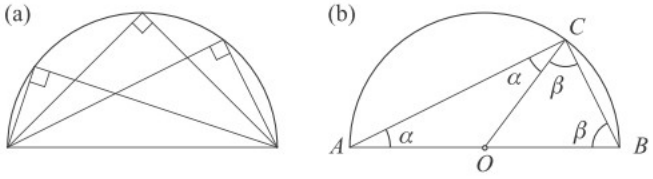

Figure 4.2.

Label the vertices of the triangle as shown in Figure 4.2b, and let O denote the center of the semicircle. Then $|AO| = |BO| = |CO|$ since each is the length of a radius, and thus triangles AOC and BOC are isosceles. Hence $\angle OAC = \angle OCA = \alpha$ and $\angle OBC = \angle OCB = \beta$. The angles of triangle ABC sum to 180°, so $2\alpha + 2\beta = 180°$, or $\alpha + \beta = 90°$. Thus C is a right angle as claimed.

Also, $2\alpha + 2\beta = 180°$ implies that $\angle BOC = 2\alpha$.

Thales' triangle theorem is used repeatedly throughout this chapter. The converse of the theorem also holds, see Challenge 4.1.

4.2 The right triangle altitude theorem and the geometric mean

In Section 2.2 we encountered the *geometric mean* $\sqrt{ab}$ of two positive numbers a and b (and we will meet it again in Sections 13.4 and 18.4). Why is it called geometric? The three numbers a, $\sqrt{ab}$, b are in geometric progression, just as the three numbers a, $(a+b)/2$, b (where $(a+b)/2$ is the arithmetic mean of a and b) are in arithmetic progression, but it is more likely that the answer lies in the purely geometric construction found in Proposition 13 in Book VI of the *Elements* of Euclid, and illustrated in Figure 4.3a.

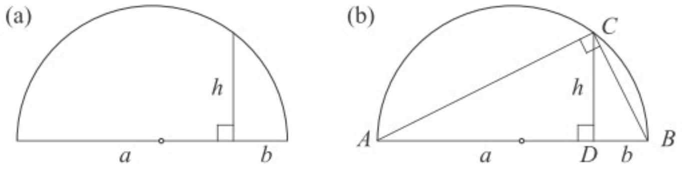

Figure 4.3.

The construction in Figure 4.3a is based on the corollary to Proposition 8 in that Book, which is also known as the

Right triangle altitude theorem: *The length of the altitude to the hypotenuse of a right triangle is the geometric mean of the lengths of the resulting two segments of the hypotenuse.*

Thus in Figure 4.3a we have $h = \sqrt{ab}$.

From Thales' triangle theorem, ABC is a right triangle and the half-chord CD partitions ABC into two triangles ACD and BCD, each similar to ABC. Ratios of corresponding sides in ACD and BCD yields $a/h = h/b$ so that $h = \sqrt{ab}$ is the geometric mean of a and b.

Since $h^2 = ab$, h is the side of a square equal in area to a rectangle with sides a and b, yielding a geometric average of a and b. The arithmetic mean $(a+b)/2$ of a and b appears in Figure 4.3b as the radius of the semicircle. Since the radius to C (not drawn) is at least as long as h, we have another proof of the arithmetic mean-geometric mean inequality $(a+b)/2 \geq \sqrt{ab}$.

When we construct a square equal in area to a given figure, we say we have *squared* the figure. Figure 4.3a shows how to square rectangles. To square a triangle, first construct a rectangle equal in area to the triangle. To square a convex n-gon, first partition it into triangles and square them. By virtue of the Pythagorean theorem, we can construct a square equal in area to two squares, and repeating this procedure yields a square equal in area to any number of squares, which completes the squaring of the n-gon.

4.3 Queen Dido's semicircle

The legend of Dido comes to us from the epic poem *Aeneid*, written by the Roman poet Virgil (70–19 BCE). Dido was a princess from the Phoenician city of Tyre (in present-day Lebanon). Dido fled the city after her brother murdered her husband and arrived in Africa in about 900 BCE, near the bay of Tunis. Dido decided to purchase land from the local leader, King Jarbas of Numidia, so she and her people could settle there. She paid Jarbas a sum of money for as much land as she could enclose with the hide of an ox. Virgil (as translated by Sarah Ruden) describes the scene as follows:

> They came here, where you now see giant walls
> And the rising citadel of newborn Carthage.
> They purchased land, 'as much as one bull's hide
> Could reach around,' and called the place 'the Bull's Hide.'

To obtain as much land as possible, Dido cut the ox skin into thin strips and tied them together, as illustrated in the 17th century woodcut in Figure 4.4. This plot of land would later become the site of the city of Carthage.

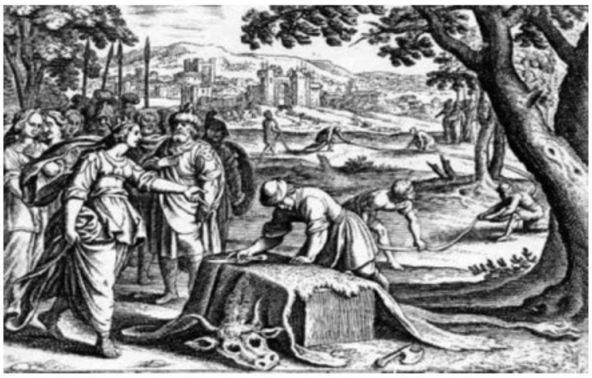

Figure 4.4.

This brings us to *Dido's problem*: How should she lay the strips on the ground to enclose as much land as possible? If we assume that the ground is flat and the Mediterranean shore is a straight line, then the optimal solution is to lay the strips in the shape of a semicircle, which legend tells us is what Dido did. Our proof is from [Niven, 1981], who attributes it to Jakob Steiner (1796–1863).

To prove that the semicircle is the optimal solution we first need the elementary result: *Of all triangles with two sides of fixed length and a third side of arbitrary length, the triangle with the largest area is the right triangle with the arbitrary side as hypotenuse*. If the fixed sides have lengths a and b and if θ denotes the angle between them, then the area T of the triangle equals $(ab\sin\theta)/2$. To maximize T we need only maximize $\sin\theta$, which occurs if and only if $\theta = 90°$.

Dido's problem is closely related to the *isoperimetric problem* ("isoperimetric" means "having the same perimeter"): Among all closed curves of a given length, which encloses the greatest area? The difference is that in Dido's case only the length of the strips of oxhide was fixed, so she was free to use as much of the straight shoreline as she wished. We call the solution

Dido's theorem. *Suppose a curve C of fixed length and a straight line L enclose a region. If C is not a semicircle, then it can be replaced by another curve C' of the same length so that C' and L enclose a region with greater area. Hence if there is a curve enclosing (with L) a region with maximal area, it must be a semicircle.*

Let A and B be the points of L which are the endpoints of the curve C, as illustrated in Figure 4.5a. From Thales' triangle theorem we know that if C is not a semicircle, then we can find a point P on C where $\angle APB \neq 90°$. Then the region enclosed by C and L consists of three parts: the region R enclosed by the line segment AP and C, the region S enclosed by the line segment BP and C, and the triangle APB denoted by T.

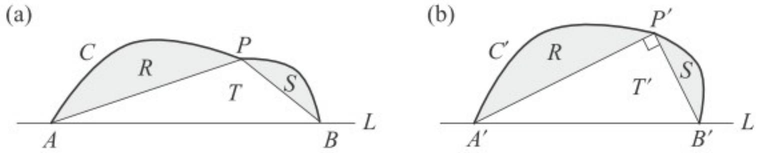

Figure 4.5.

With AP and BP fixed at the point P, slide A and B along L to A' and B' with $|A'P'| = |AP|$ and $|B'P'| = |BP|$ so that $\angle A'P'B' = 90°$, as shown in Figure 4.5b. By construction, the shaded areas R and S remain the same, but the area of T' is greater that the area of T, so that the curve C has been transformed into a curve C' with the same length as C but enclosing a greater area with L.

As noted in the statement of Dido's theorem, we must assume that a curve enclosing the maximum area exists. The existence of such a curve was proven formally by Karl Theodor Wilhelm Weierstrass (1815–1897).

4.4 The semicircles of Archimedes

The *Book of Lemmas* (or *Liber Asumptorum*) is a collection of fifteen propositions and their proofs attributed to Archimedes (287–212 BCE). It survives today from an Arabic translation by Thabit ibn Qurra (836–901). Several of the propositions in the *Book of Lemmas* concern semicircles and we present four of them, Propositions 3, 4, 8, and 14. Proposition 3 presents a bisection of the vertical line segment in our icon, proposition 8 uses the semicircle to trisect an angle, and propositions 4 and 14 concern area of figures known as an *arbelos* (shoemaker's knife) and a *salinon* (a salt-cellar). See [Heath, 1897] for all the propositions and their proofs.

Archimedes' Proposition 3. *Let AB be the diameter of a semicircle, and let the tangents to it at B and at any other point D on it meet at T. If now DE is drawn perpendicular to AB, and if AT, DE meet at F,* then $|DF| = |FE|$. See Figure 4.6a.

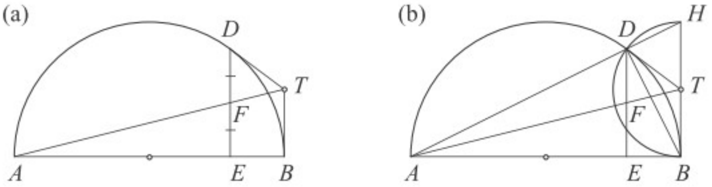

Figure 4.6.

Draw AD, and extend it and BT to meet at H, as shown in Figure 4.6b. Since $\angle ADB$ is a right angle, so is $\angle BDH$. In addition we have $|BT| = |TD|$. Hence T is the center of a semicircle with diameter BH, and $|BT| = |TH|$. Since DE and BH are parallel, it follows that $|DF| = |FE|$.

Archimedes' Proposition 4. *Let P, Q, and R be three points on a line, with Q lying between P and R. Semicircles are drawn on the same side of the line with diameters PQ, QR, and PR. An arbelos is the figure bounded by the three semicircles. Draw the perpendicular to PR at Q, meeting the* *largest semicircle at S. Then the area A of the arbelos equals the area C of the circle with diameter QS.* See Figure 4.7.

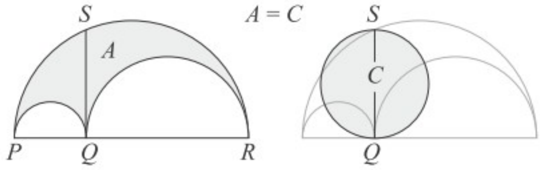

Figure 4.7.

There is a direct algebraic proof of Proposition 4 using the geometric mean, which we present as Challenge 4.2. Here we present a geometric proof [Nelsen, 2002b] based on the Pythagorean theorem. In Proposition 31 of Book VI of the *Elements*, Euclid writes: *In right-angled triangles the figure on the side opposite the right angle equals the sum of the similar and similarly described figures on the sides containing the right angle*. For similar figures on the sides, we now use semicircles. See Figure 4.8.

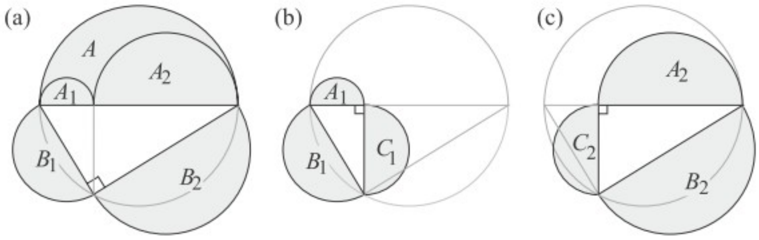

Figure 4.8.

In Figure 4.8a we have $A + A_1 + A_2 = B_1 + B_2$, in Figure 4.8b, $B_1 = A_1 + C_1$, and in Figure 4.8c, $B_2 = A_2 + C_2$. Together the equations yield $A + A_1 + A_2 = A_1 + C_1 + A_2 + C_2$, which simplifies to $A = C_1 + C_2 = C$.

In Proposition 8 Archimedes uses a circle to trisect an angle. We present a slight variation with a semicircle [Aaboe, 1964].

Archimedes' Proposition 8. *In a semicircle with center O and diameter AB, let $\angle AOE$ be the angle to be trisected. Let C be the point on the extension of AB so that the line CE intersects the circle at D with $|CD|$ equal to the radius of the circle. Then $\angle ACE = (1/3)\angle AOE$.* See Figure 4.9.

Let $\alpha = \angle AOE$. Since $|CD| = |OD| = |OE|$, triangles DOE and ODC are isosceles, and hence have equal base angles β and γ, respectively,

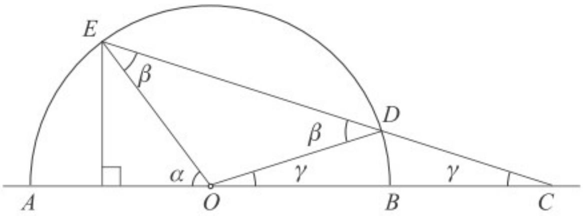

Figure 4.9.

as indicated in Figure 4.9. Hence by the exterior angle theorem (Proposition 32 in Book I of Euclid's *Elements*) $\beta = 2\gamma$ and $\alpha = \beta + \gamma = 3\gamma$.

Figure 4.9 can also be used to derive the triple angle formulas for the sine and cosine. See Section 7.5.

Archimedes' Proposition 14. *Let P, Q, R, S be four points on a line (in that order) such that $PQ = RS$. Semicircles are drawn above the line with diameters PQ, RS, and PS, and another semicircle with diameter QR is drawn below the line.* A salinon *is the figure bounded by the four semicircles. Let the axis of symmetry of the salinon intersect its boundary at M and N. Then the area A of the salinon equals the area C of the circle with diameter MN.* See Figure 4.10.

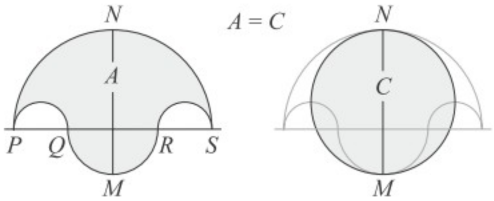

Figure 4.10.

In our proof [Nelsen, 2002a], we use the fact that the area of a semicircle is $\pi/2$ times the area of the inscribed isosceles right triangle (see Figure 4.11).

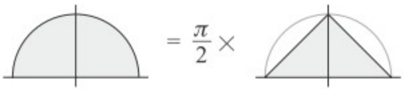

Figure 4.11.

Consequently, the area of the salinon is $\pi/2$ times the area of a square (see Figure 4.12), which equals the area of a circle and proves the proposition.

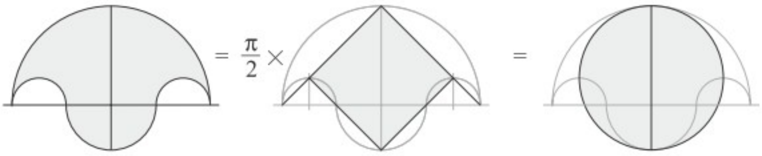

Figure 4.12.

4.5 Pappus and the harmonic mean

The *harmonic mean* of two positive numbers a and b is often defined as the reciprocal of the arithmetic mean of the reciprocals of a and b, i.e., $([(1/a)+(1/b)]/2)^{-1}$, which simplifies to $2ab/(a+b)$. The name harmonic derives from its connection with the harmonic sequence $1, 1/2, 1/3, 1/4, \ldots$, which is important in music. In the harmonic sequence, each term is the harmonic mean of its predecessor and its successor, e.g., $1/3$ is the harmonic mean of $1/2$ and $1/4$.

We have seen how the arithmetic and geometric means of a and b appear as line segments in a semicircle with diameter $a+b$ in Figure 4.3a: the radius is the arithmetic mean $(a+b)/2$ and the half-chord perpendicular to the diameter is the geometric mean $\sqrt{ab}$. Pappus of Alexandria (circa 290–350), in Book III of his *Collection*, added one more line segment to illustrate the harmonic mean as well, as shown in Figure 4.13.

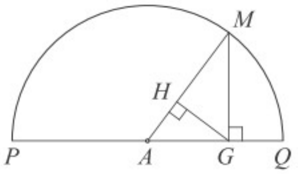

Figure 4.13.

When $|PG| = a$ and $|GQ| = b$, the arithmetic and geometric means are $|AM| = (a+b)/2$ and $|GM| = \sqrt{ab}$ respectively. Since GH is perpendicular to AM, triangles AGM and GHM are similar, so that $|HM|/|GM| = |GM|/|AM|$. Thus the harmonic mean is $|HM| = 2ab/(a+b)$, with $|HM| \leq |GM| \leq |AM|$. Also, the geometric mean of a and b equals the geometric mean of $(a+b)/2$ and $2ab/(a+b)$.

4.6 More trigonometric identities

The trigonometric functions can be expressed both in terms of the coordinates of points on the unit circle and as ratios of sides of right triangles. The two approaches combine with our semicircle icon to produce illustrations of a variety of trigonometric identities.

In Figure 4.2b, $\angle BOC = 2\alpha$ as a consequence of the exterior angle theorem, which immediately yields the two half-angle formulas for the tangent function.

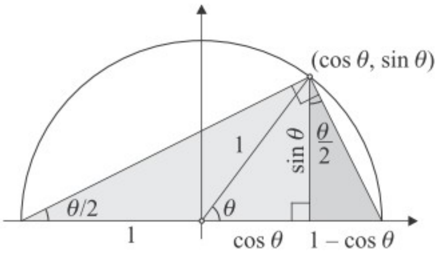
$(\cos\theta, \sin\theta)$
1
$\sin\theta$
$\frac{\theta}{2}$
$\theta/2$
θ
1
$\cos\theta$
$1-\cos\theta$

Figure 4.14.

In the light gray triangle in Figure 4.14 we have $\tan(\theta/2) = \sin\theta/(1 + \cos\theta)$, and in the dark gray one we have $\tan(\theta/2) = (1 - \cos\theta)/\sin\theta$ [Walker, 1942]. Figure 4.14 can also be used to establish certain inverse trigonometric identities: see Challenge 4.3.

If we replace θ by 2θ in Figure 4.14, we can obtain the double-angle formulas for the sine and cosine. See Figure 4.15.

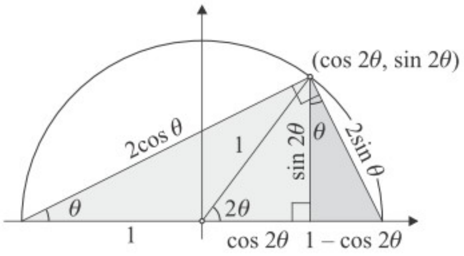
$(\cos 2\theta, \sin 2\theta)$
$2\cos\theta$
1
$\sin 2\theta$
θ
$2\sin\theta$
θ
2θ
1
$\cos 2\theta$
$1-\cos 2\theta$

Figure 4.15.

In the light gray triangle, we have $\sin\theta = \sin 2\theta/2\cos\theta$, hence $\sin 2\theta = 2\sin\theta\cos\theta$, $\cos\theta = (1+\cos 2\theta)/2\cos\theta$ or $\cos 2\theta = 2\cos^2\theta - 1$. In the dark gray triangle, we have $\sin\theta = (1-\cos 2\theta)/2\sin\theta$, so that $\cos 2\theta =$

$1 - 2\sin^2\theta$ [Woods, 1936]. See Challenge 4.4 for another derivation of the formulas.

In Challenge 3.8 we encountered the Weierstrass substitution, useful for integrals of rational functions of the sine and cosine. The change of variables $z = \tan(\theta/2)$ yields rational functions of z for $\sin\theta$ and $\cos\theta$, as illustrated in Figure 4.16.

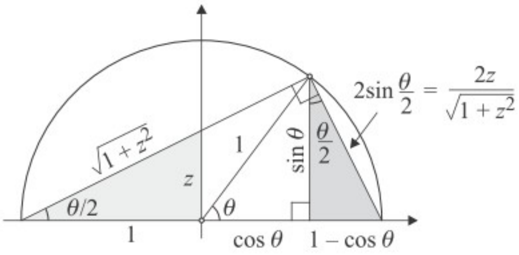

Figure 4.16.

Since $z = \tan(\theta/2)$, the sides of the light gray triangle are 1, z, and $\sqrt{1+z^2}$, and consequently the hypotenuse of the dark gray triangle is $2\sin(\theta/2) = 2z/\sqrt{1+z^2}$. Since the shaded triangles are similar, we have

$$\frac{\sin\theta}{2z/\sqrt{1+z^2}} = \frac{1}{\sqrt{1+z^2}} \text{ and } \frac{1-\cos\theta}{2z/\sqrt{1+z^2}} = \frac{z}{\sqrt{1+z^2}},$$

and hence $\sin\theta = 2z/(1+z^2)$ and $\cos\theta = (1-z^2)/(1+z^2)$ [Deiermann, 1998].

4.7 Areas and perimeters of regular polygons

There is a nice relationship between the perimeter of a regular polygon with n sides (an *n-gon*) and the area of a regular polygon with $2n$ sides when both are inscribed in a circle of radius r. If we let s_n denote the side length, $P_n = ns_n$ the perimeter, and A_n the area of a regular n-gon, then the area $A_{2n}/2n$ of the shaded triangle in Figure 4.17 is equal to one-half its base r times its altitude $s_n/2$, i.e., $A_{2n}/2n = rs_n/4$.

Hence the area of the $2n$-gon is $A_{2n} = rns_n/2 = rP_n/2$, i.e., half the product of the circumradius and perimeter of the n-gon. For example, the perimeter of a regular hexagon inscribed in a circle of radius r is $P_6 = 6r$, and hence the area of a regular dodecagon inscribed in the same circle is $A_{12} = rP_6/2 = 3r^2$.

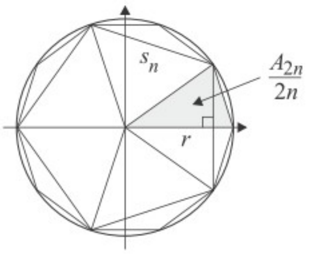

Figure 4.17.

As a consequence we can conclude that the area of a circle is equal to one-half the product of its radius and its circumference, a fact known to the Indian mathematician Bhāskara (circa 1114–1185).

4.8 Euclid's construction of the five Platonic solids

The last of the thirteen Books of the *Elements* of Euclid concerns the five Platonic solids—the tetrahedron, cube, octahedron, dodecahedron, and icosahedron. After numerous Propositions about the individual solids Proposition 18, the final Proposition of the final Book, provides a construction for the edge length of each of the five solids when inscribed in the same sphere. The construction is remarkably simple, based on a semicircle and several segments perpendicular to its diameter. Figure 4.18 is a simplified version of a diagram that appears in many translations of the *Elements*.

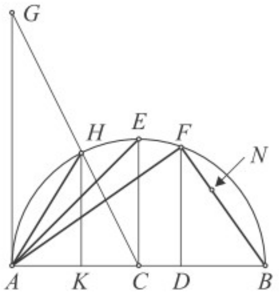

Figure 4.18.

Draw a semicircle with AB as diameter, and locate points C and D on AB so that $|AC| = |CB|$ and $|AD| = 2|DB|$. Draw AG perpendicular to

AB with $|AG| = |AB|$, let GC intersect the semicircle at H, and draw HK perpendicular to AB. Draw CE and DF perpendicular to AB, draw AH, AE, AF, and BF, and locate N on BF so that $|BF| = \phi|BN|$ where $\phi = (1 + \sqrt{5})/2$ is the golden ratio, the positive root of the quadratic equation $\phi^2 = \phi + 1$. (Euclid describes how to do this in Proposition II.11.) Then

AF is the edge of the tetrahedron,
BF is the edge of the cube,
AE is the edge of the octahedron,
BN is the edge of the dodecahedron, and
AH is the edge of the icosahedron.

We justify the last claim and leave the others to Challenge 4.8. The twelve vertices of an icosahedron with edge s are determined by three s-by-$s\phi$ *golden rectangles*. See Figure 4.19.

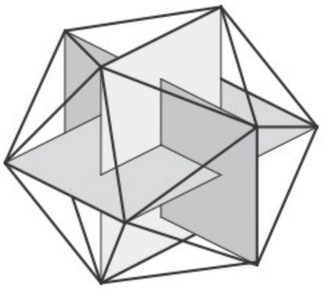

Figure 4.19.

If $|AB| = d$ in Figure 4.18, then triangle geometry yields $|AH| = d/\sqrt{2+\phi}$. The diagonal of the golden rectangle is also the diameter d of the sphere, and hence $d = \sqrt{s^2 + s^2\phi^2} = s\sqrt{2+\phi}$. It now follows that $s = d/\sqrt{2+\phi} = |AH|$, as required.

4.9 Challenges

4.1. Prove the converse of Thales' triangle theorem: the hypotenuse of a right triangle is a diameter of its circumcircle.

4.2. Prove Proposition 4 from the *Book of Lemmas* using the fact that, in Figure 4.7, $|QS|$ is the geometric mean of $|PQ|$ and $|QR|$.

4.3. Many identities for inverse trigonometric functions are equivalent to identities for trigonometric functions. Establish the following identities by relabeling sides and angles in Figure 4.14:

(a) $\dfrac{\arcsin x}{2} = \arctan\dfrac{x}{1+\sqrt{1-x^2}} = \arctan\dfrac{1-\sqrt{1-x^2}}{x}$;

(b) $\dfrac{\arccos x}{2} = \arctan\dfrac{1-x}{\sqrt{1-x^2}} = \arctan\dfrac{\sqrt{1-x^2}}{1+x} = \arctan\sqrt{\dfrac{1-x}{1+x}}$;

(c) $\dfrac{\arctan x}{2} = \arctan\dfrac{x}{1+\sqrt{1+x^2}} = \arctan\dfrac{\sqrt{1+x^2}-1}{x}$.

4.4. Use Figure 4.20 to derive the double angle formulas for the sine and cosine. (Hints: Compute the area of the shaded triangle in two ways, and express the length of the chord in two ways.)

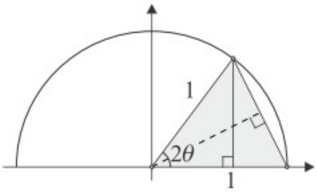

Figure 4.20.

4.5. Using the semicircle icon show that $\arctan x \le \arcsin x$ for x in [0,1].

4.6. Show that the area of the shaded region in Figure 4.21 depends only on a.

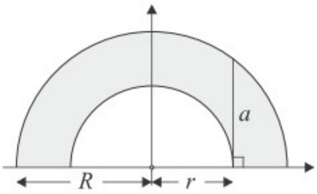

Figure 4.21.

4.7. If we draw semicircles on the legs of a right triangle inscribed in a semicircle, we create a *lune* on each leg, the region inside the small semicircle and outside the large one, as illustrated in light gray in Figure 4.22. Show that the combined area of the two lunes equals the area of the right triangle, a fact known to Hippocrates of Chios (circa 470–410 BCE).

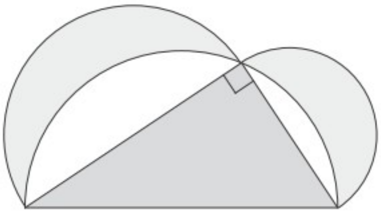

Figure 4.22.

4.8. In Figure 4.18, show that AF, BF, AE, and BN are, respectively, the edges of the tetrahedron, cube, octahedron, and dodecahedron inscribed in a sphere of diameter AB.

4.9. Another curve bounded by three semicircular arcs, like the arbelos, is the heart-shaped curve called the *cardioid of Boscovich* (Roger Boscovich, 1711–1787) illustrated in Figure 4.23. Show that any line through the cusp bisects its perimeter.

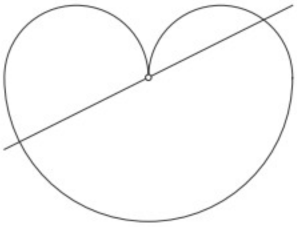

Figure 4.23.

Book: 20210703/ Oeuvres Tchebychef
-2 Volumes/
Oeuvres_de_P_L_Tchebychef_
417.pdf.

Pág. 129 (167) - Sur les engrenages
Pág. 165 (203) - Sur les quadratures

Obs: Important.

Source: DATA-NovoVivo-014-2021/
Books-014-2021/

Other: Posta:
/A Cornucopia of Quadrilaterals

/Wildberger/ Divine Proportions...
pdf

/Oeuvres Tchebychef - 2 Volumes/

Obs: 2 volumes - Oeuvres
Tchebychef
4 volumes - Oeuvres de
Charles Hermite.

Imp: No capítulo Sur les Quadratures
fala da [illegible] de M. Hermite (valeurs
de l'intégrale).

%
! *+,-* * *. ! * % . 7 M" M M>M8 30 MO P
30 7M ! ?M32 9B))9(9%B 2" L
3 *+,-* * *. ! * % . 7 M" M M>M8 30 MO P
30 7M ! ?M32 9B) 9(2" L
') 9 !

Sur les engrenages*).

§ 1. Dans les recherches théoriques sur le tracé des engrenages on suppose ordinairement donnée la forme du profil de la dent de l'une des roues et on en déduit celle du profil de la dent de l'autre. On peut trouver ainsi une infinité de différentes formes des engrenages, mais il n'y en a que très peu, qui sont employées dans la pratique. Pour que la pratique puisse, sans se borner à ces formes particulières, se servir de telle forme qui est la plus avantageuse dans le cas donné, il faut avoir une méthode générale pouvant donner un moyen facile pour tracer les profils des dents d'un couple des roues qui satisfont le mieux aux exigences de la pratique. Si l'on veut tracer les profils des dents mathématiquement exacts, on rencontre des obstacles insurmontables, mais la pratique n'exige pas cette rigueur; pour lui suffisent les procédés approximatifs, qui consistent en ce, qu'on donne aux faces et aux flancs des dents la forme des arcs de cercle convenablement choisis. Quand on n'a en vue que les profils des dents en arcs de cercle (ces profils peuvent en général remplacer tous les autres avec une approximation suffisante pour la pratique), la question de la détermination de la forme des dents le mieux satisfaisant aux exigences de la pratique devient considérablement simplifiée, parce qu'il ne s'agit alors que de déterminer les centres et les rayons des quelques cercles. En étudiant le mouvement des engrenages profilés en arcs de cercle, on voit qu'ils ne peuvent jamais donner lieu à l'invariabilité du rapport des vitesses angulaires, qui présente une condition nécessaire pour que le mouvement de l'engrenage soit regulier; par suite en déterminant le profil de ces engrenages il faut se proposer pour but de diminuer autant qu'il est possible les irrégularités du mouvement de l'engrenage, qui proviennent de la variabilité plus ou moins considérable du rapport des vitesses angulaires.

---

*) Ce mémoire a paru en français sous le titre: «Mémoire sur les engrenages par M. Tchebycheff. Traduction de MM. Dwelshauvers-Dery et Couharevitch.» dans la «Revue universelle des mines» 1875, pp. 523—546; la traduction de M. Mescherskiy se distingue de la précédente en ce qu'elle reproduit de plus près le texte original.

Tel est notre objet; nous allons montrer, comment on trouve, en satisfaisant aux exigences de la pratique, la forme des faces et des flancs des dents de l'engrenage profilé en arcs de cercle, pour laquelle les irrégularités du mouvement de l'engrenage sont les moindres possibles.

§ 2. Soient: C, C_1 (fig. 1) les centres des roues dentées (que nous appellerons en abrégé: la roue C et la roue C_1), et de leurs circonférences primitives; $R = CD$, $R_1 = C_1D$ les rayons de ces dernières; $\rho = AD$, $\rho_1 = BD$ les rayons des arcs de cercle DE et DF, dont le premier est le profil de la face de la dent de la roue C et le second celui du flanc de la dent de la roue C_1; A et B les positions respectives des centres de ces arcs au moment, où leur point de contact est sur la ligne des centres CC_1 au point D; A_1 et B_1 les positions de ces centres à un autre moment, lorsque les arcs sont en contact au point quelconque G. Les angles $ACA_1 = \lambda$ et $BC_1B_1 = \mu$ sont ceux, dont les roues C et C_1 tournent en même temps que le point de contact des dents se déplace de D en G. L'angle μ est une fonction de λ que nous désignerons par $F(\lambda)$; la dérivée de cette fonction $F'(\lambda)$ représentera alors le rapport des vitesses angulaires des roues C et C_1 correspondant aux différentes valeurs de $\lambda = ACA_1$.

Fig. 1.

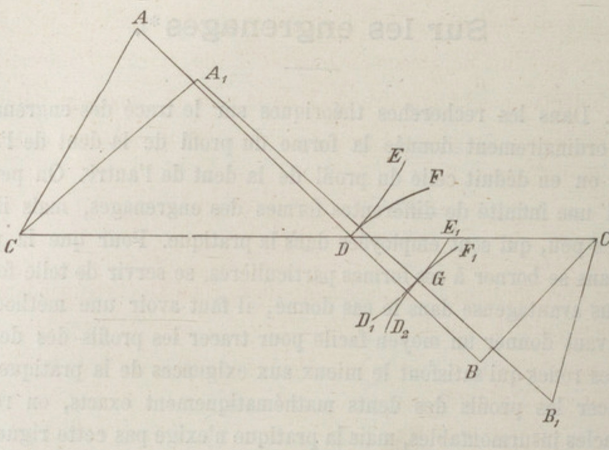

Pour $\lambda = 0$ on a $\mu = 0$, donc la fonction $F(\lambda)$ devient nulle quand $\lambda = 0$; d'autre part il est facile de voir que pour cette valeur $\lambda = 0$ la dérivée $F'(\lambda)$ doit être égale à $\frac{R}{R_1}$. En effet, quand $\lambda = 0$, le point de contact des dents est sur la ligne des centres au point D; donc le rapport des vitesses angulaires des roues C et C_1 pour cette valeur de λ, comme il suit d'une propriété générale des engrenages, est égal à $\frac{CD}{C_1D}=\frac{R}{R_1}$; mais nous avons vu que ce rapport est égal à la dérivée $F'(\lambda)$, par conséquent $F'(\lambda)=\frac{R}{R_1}$ pour $\lambda=0$.

Avant d'examiner les irrégularités du mouvement des roues C et C_1, — inévitables, quand les dents sont profilées en arcs de cercle, — remarquons que dans le cas considéré, où les rayons des circonférences primitives sont R et R_1, le rapport des vitesses angulaires devrait être constant et égal à $\frac{R}{R_1}$, de manière qu'à un déplacement $\lambda=ACA_1$ de la roue C devrait correspondre un déplacement angulaire $\frac{R}{R_1}\lambda$ de la roue C_1, mais le déplacement réel de cette roue, comme nous avons posé plus haut, est égal à $\mu=F(\lambda)$; par suite la différence

$$F(\lambda)-\frac{R}{R_1}\lambda$$

représente l'erreur dans le déplacement de la roue C_1 correspondant au déplacement angulaire λ de la roue C_1.

Or, nous venons de voir que $F(\lambda)$ s'annule pour $\lambda=0$, donc pour cette valeur de λ la différence

$$F(\lambda)-\frac{R}{R_1}\lambda$$

devient nulle également. Pour les autres valeurs de λ cette différence n'est pas nulle en général, mais elle s'écarte plus ou moins de zéro pendant que λ croît de zéro jusqu'à sa limite l qu'elle atteint, lorsque les extrémités E et F des arcs DE et DF, profils des dents, sont en contact.

Donc pour diminuer autant que possible les irrégularités du mouvement de l'engrenage considéré il faut choisir les arcs de cercle DE et DF de manière que la différence

$$F(\lambda)-\frac{R}{R_1}\lambda$$

s'écarte le moins possible de zéro, tandis que λ varie entre les limites $\lambda=0$ et $\lambda=l$.

La détermination exacte des arcs de cercle DE et DF, qui rempliraient cette condition est excessivement difficile à cause de la complication de la forme de la fonction $F(\lambda)$; mais, si l'on se borne à une approximation des dents est sur la ligne des centres au point D; donc le rapport des vitesses angulaires des roues C et C_1 pour cette valeur de λ, comme il suit d'une propriété générale des engrenages, est égal à $\frac{CD}{C_1D} = \frac{R}{R_1}$; mais nous avons vu que ce rapport est égal à la dérivée $F'(\lambda)$, par conséquent $F'(\lambda) = \frac{R}{R_1}$ pour $\lambda = 0$.

Avant d'examiner les irrégularités du mouvement des roues C et C_1,— inévitables, quand les dents sont profilées en arcs de cercle, — remarquons que dans le cas considéré, où les rayons des circonférences primitives sont R et R_1, le rapport des vitesses angulaires devrait être constant et égal à $\frac{R}{R_1}$, de manière qu'à un déplacement $\lambda = ACA_1$ de la roue C devrait correspondre un déplacement angulaire $\frac{R}{R_1}\lambda$ de la roue C_1, mais le déplacement réel de cette roue, comme nous avons posé plus haut, est égal à $\mu = F(\lambda)$; par suite la différence

$$F(\lambda) - \frac{R}{R_1}\lambda$$

représente l'erreur dans le déplacement de la roue C_1 correspondant au déplacement angulaire λ de la roue C_1.

Or, nous venons de voir que $F(\lambda)$ s'annule pour $\lambda = 0$, donc pour cette valeur de λ la différence

$$F(\lambda) - \frac{R}{R_1}\lambda$$

devient nulle également. Pour les autres valeurs de λ cette différence n'est pas nulle en général, mais elle s'écarte plus ou moins de zéro pendant que λ croît de zéro jusqu'à sa limite l qu'elle atteint, lorsque les extrémités E et F des arcs DE et DF, profils des dents, sont en contact.

Donc pour diminuer autant que possible les irrégularités du mouvement de l'engrenage considéré il faut choisir les arcs de cercle DE et DF de manière que la différence

$$F(\lambda) - \frac{R}{R_1}\lambda$$

s'écarte le moins possible de zéro, tandis que λ varie entre les limites $\lambda = 0$ et $\lambda = l$.

La détermination exacte des arcs de cercle DE et DF, qui rempliraient cette condition est excessivement difficile à cause de la complication de la forme de la fonction $F(\lambda)$; mais, si l'on se borne à une approximation suffisante pour la pratique, on pourra bien se débarrasser des difficultés, provenant de la complication de la fonction $F(\lambda)$, en substituant à l'expression

$$F(\lambda)-\frac{R}{R_1}\lambda$$

son développement suivant les puissances ascendantes de λ; puisque entre les limites 0 et l l'angle λ est toujours fort petit, on n'a pas besoin de prendre un nombre considérable de termes de ce développement; pour la résolution des questions, qui se présentent dans la pratique, il suffit, comme on verra, s'arrêter à la troisième ou à la quatrième puissance de λ, de manière que l'expression de la différence

$$F(\lambda)-\frac{R}{R_1}\lambda$$

prendra la forme de l'un des deux polynomes suivants:

$$K_3\lambda^3+K_2\lambda^2+K_1\lambda+K_0,$$

ou

$$K_4\lambda^4+K_3\lambda^3+K_2\lambda^2+K_1\lambda+K_0.$$

§ 3. D'après ce qui précède, la détermination des arcs DE et DF, qui forment le profil de l'engrenage le plus régulier, se réduit à la recherche des coefficients des polynomes

$$K_3\lambda^3+K_2\lambda^2+K_1\lambda+K_0,$$
$$K_4\lambda^4+K_3\lambda^3+K_2\lambda^2+K_1\lambda+K_0,$$

pour qu'ils se rapprochent le plus possible de zéro entre des limites données de λ. Ces polynomes ressemblent à ceux qui se présentent, quand on cherche la meilleure forme du parallélogramme de Watt: ceux-là ne diffèrent de ceux-ci que par quelques particularités, résultant de ce que nous avons remarqué plus haut par rapport aux valeurs de la fonction $F(\lambda)$ et de sa dérivée $F'(\lambda)$ pour $\lambda=0$. La fonction $F(\lambda)$ s'annule pour $\lambda=0$, donc dans le développement de la différence

$$F(\lambda)-\frac{R}{R_1}\lambda$$

n'existe pas le terme, qui ne contient pas λ; la dérivée $F'(\lambda)$ pour $\lambda=0$ devient égale à $\frac{R}{R_1}$, par conséquent le coefficient de λ dans ce développement est égal à zéro. Il s'en suit que nos polynomes prennent l'une des deux formes:

$$K_3\lambda^3 + K_2\lambda^2,$$

$$K_4\lambda^4 + K_3\lambda^3 + K_2\lambda^2.$$

Dans ces derniers polynomes comme dans ceux qui se rapportent à la théorie du parallélogramme de Watt, le premier coefficient étant supposé connu, les autres doivent être choisis de manière que les valeurs des polynomes s'écartent le moins possible de zéro entre les limites $\lambda = 0$ et $\lambda = l$. D'après ce qui a été démontré dans notre mémoire intitulé: *Théorie des mécanismes connus sous le nom de parallélogrammes* *), la détermination de tels polynomes dans le cas général n'est possible qu'à l'aide des fonctions elliptiques, mais dans le cas des polynomes du troisième et du quatrième degré elle est très facile. En s'arrêtant aux millièmes dans les valeurs des coefficients, on trouve que les polynomes prennent les formes suivantes:

$$K_2(\lambda^3 - 0{,}894\,l\lambda^2),$$

$$K_4(\lambda^4 - 1{,}559\,l\lambda^3 + 0{,}578\,l^2\lambda^2)\ ^{**)}$$

Ces polynomes entre les limites $\lambda = 0$ et $\lambda = l$ s'écartent de zéro moins que tous les autres de la même espèce, cependant il n'est pas difficile de voir que leur écart de zéro devient le plus grand, lorsque λ atteint sa valeur limite l, par conséquent dans les engrenages, où les variations du rapport des vitesses angulaires s'expriment par les polynomes ci-dessus, la poussée est la plus defectueuse au moment, où une dent échappe et où la suivante va se remettre en prise à la ligne des centres. Cette circonstance est très desavantageuse dans la pratique, qui exige, qu'à la fin comme au commencement de la prise le rapport des vitesses soit le même; c'est pourquoi il faut remplacer les polynomes, trouvés plus haut, par les autres, qui,

---

*) T. I, pag. 111—143.

**) Les expressions exactes de ces polynomes sont:

$$K_3(\lambda^3 - al\lambda^2),$$

$$K_4\left(\lambda^4 - \frac{4pl}{q}\lambda^3 + \frac{(13p^2-4)\,l^2}{2q^2}\lambda^2\right),$$

où a est déterminée par l'équation: $a^3 + \frac{27}{4}(a-1) = 0$,

$$p = \sqrt{\frac{20 + 24\sqrt{3}}{83}},$$

$$q = \frac{p}{2} + \sqrt{1-p^2} + \sqrt{p^2 + 2p\sqrt{1-p^2}}.$$

s'écartant le moins de zéro entre les limites $\lambda = 0$ et $\lambda = l$, s'annuleraient non seulement pour $\lambda = 0$, mais aussi pour $\lambda = l$. D'après le § 9 du mémoire précité on trouvera les polynomes cherchés, si l'on remplace l par $\frac{l}{0,894}$ dans l'expression écrite plus haut du polynome du troisième degré et par $\frac{l}{0,952}$ dans celle du polynome du quatrième degré. On obtient ainsi

$$K_3(\lambda^3 - l\lambda^2),$$
$$K_4(\lambda^4 - 1,638 l\lambda^3 + 0,638 l^2\lambda^2),$$

polynomes, qui serviront à déterminer les profils des dents remplissant la condition que le rapport des vitesses angulaires aie la même valeur normale à la fin ainsi qu'au commencement de la prise et qui pendant la durée de la prise s'écarte les moins possible de cette valeur.

4. Nous commencerons par le cas, où dans le choix des profils des dents, les exigences de la pratique étant satisfaites, il ne reste qu'une quantité disponible, qu'on peut déterminer de manière, que les irrégularités du mouvement de l'engrenage soient les moindres possibles. Dans ce cas, en nous arrêtant à la troisième puissance dans le développement de la différence

$$F(\lambda) - \frac{R}{R_1}\lambda$$

nous devons poser

$$F(\lambda) - \frac{R}{R_1}\lambda = K_3(\lambda^3 - l\lambda^2). \tag{1}$$

Or, en vertu de la nature de la fonction $F(\lambda)$, le premier membre de l'équation (1) peut être réduit toujours à la forme suivante exacte jusqu'à λ^3

$$K_3\lambda^3 + K_2\lambda^2,$$

comme nous l'avons vu au § 3; donc l'équation (1) nous donne une relation entre les coefficients K_2 et K_3 et, par conséquent, aussi entre les quantités, qui déterminent le profil cherché.

Pour obtenir cette relation nous pourrions trouver l'expression de la fonction $F(\lambda)$ et puis comparer les coefficients de λ^2 et de λ^3 dans les deux membres de l'équation (1), mais on peut procéder plus simplement, en remarquant, que l'équation (1) différenciée donne

$$F'(\lambda) - \frac{R}{R_1} = K_3(3\lambda^2 - 2l\lambda)$$

ce qui suppose avec une approximation jusqu'à λ^2, que la différence

$$F'(\lambda) - \frac{R}{R_1}$$

devient nulle pour $\lambda = 0$ et $\lambda = \frac{2}{3} l$.

La première valeur $\lambda = 0$, annule toujours la différence

$$F'(\lambda) - \frac{R}{R_1},$$

comme nous l'avons vu au § 2; l'annulation de cette différence pour $\lambda = \frac{2}{3} l$ nous servira à trouver la relation entre les quantités, lesquelles contient la fonction $F(\lambda)$.

Dans ce but rappelons que, d'après le § 2, la dérivée $F'(\lambda)$ représente le rapport réel des vitesses angulaires des roues considérées et la fraction $\frac{R}{R_1}$ — la valeur normale de ce rapport; si donc la différence

$$F'(\lambda) - \frac{R}{R_1}$$

s'annule pour $\lambda = \frac{2}{3} l$, c'est qu'à ce moment le rapport des vitesses angulaires a sa valeur normale et par conséquent la normale commune aux deux profils en prise passe par le point de contact D des circonférences primitives sur la ligne des centres CC_1. Or, il est facile de trouver l'inclinaison de la normale commune sur la ligne des centres dans la position, où $\lambda = \frac{2}{3} l$, ce qui servira à établir la relation entre les quantités qui déterminent le profil cherché.

§ 5. Conservant les mêmes notations que dans la figure 1, nous aurons (fig. 2): $CD = R$, $C_1D = R_1$ les rayons des circonférences primitives, $AD = \rho$, $BD = \rho_1$, les rayons des arcs de cercle DE et DF; A et B les positions des centres des arcs DE et DF au moment, où leur point de contact est sur la ligne des centres CC_1 au point D; A_1 et B_1 les positions respectives de ces centres après une rotation de l'angle $\lambda = ACA_1 = \frac{2}{3} l$ de la roue C, quand la normale commune A_1B_1 aux profils D_1E_1 et D_2F_1, comme il est dit au paragraphe précédent, passera par la point de contact primitif D; soient: $r = AC = A_1C$ et les angles

$$L = ACD, \quad N = ADC, \quad v = ADA_1.$$

Le triangle A_1DC donne la relation

$$\operatorname{tang}(ADC - ADA_1) = \frac{A_1C \, . \, \sin A_1CD}{CD - A_1C \, . \, \cos A_1CD};$$

Fig. 2.

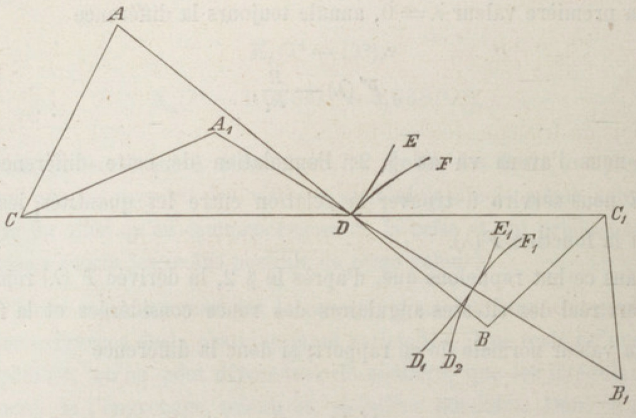

en employant les notations adoptées on a donc

$$\operatorname{tang}(N - v) = \frac{r \sin\left(L - \frac{2}{3}l\right)}{R - r\cos\left(L - \frac{2}{3}l\right)}$$

ou

$$\frac{\operatorname{tang} N - \operatorname{tang} v}{1 + \operatorname{tang} N . \operatorname{tang} v} = \frac{r \sin\left(L - \frac{2}{3}l\right)}{R - r\cos\left(L - \frac{2}{3}l\right)}$$

En résolvant cette équation par rapport à tang v, on trouve

$$\operatorname{tang} v = \frac{\left[R - r\cos\left(L - \frac{2}{3}l\right)\right] \operatorname{tang} N - r \sin\left(L - \frac{2}{3}l\right)}{R - r\cos\left(L - \frac{2}{3}l\right) + r \sin\left(L - \frac{2}{3}l\right) \operatorname{tang} N},$$

ou

$$(2) \qquad \operatorname{tang} v = \frac{R \sin N - r \sin\left(L + N - \frac{2}{3}l\right)}{R \cos N - r \cos\left(L + N - \frac{2}{3}l\right)}.$$

Développant ici

$$\sin\left(N + L - \frac{2}{3}l\right), \ \cos\left(N + L - \frac{2}{3}l\right)$$

suivant les puissances ascendantes de l, on exprime par les séries suivantes: le numérateur

$$R \sin N - r \sin(L+N) + \frac{2}{3} r \cos(L+N) . l + \frac{2}{9} r \sin(L+N) . l^2 + \ldots ,$$

et le dénominateur

$$R \cos N - r \cos(L+N) - \frac{2}{3} r \sin(L+N) . l + \frac{2}{9} r \cos(L+N) . l^2 + \ldots .$$

Mais du triangle ACD on tire

$$CD . \sin ADC = AC . \sin CAD.$$

$$CD . \cos ADC + AC . \cos CAD = AD,$$

d'où, en remplaçant l'angle CAD par

$$\pi - (ADC + ACD),$$

avec les notations adoptées on reçoit

$$R \sin N = r \sin(L+N),$$

$$R \cos N - r \cos(L+N) = \rho,$$

ce qui réduit les expressions trouvées du numérateur et du dénominateur dans la formule (2) aux formes suivantes plus simples:

$$\frac{2}{3}(R \cos N - \rho)\, l + \frac{2}{9} R \sin N . l^2 + \ldots .$$

$$\rho - \frac{2}{3} R \sin N . l + \frac{2}{9}(R \cos N - \rho) l^2 + \ldots .$$

En divisant la première de ces séries par la seconde on trouve

$$\frac{2}{3}\left(\frac{R}{\rho} \cos N - 1\right) l + \frac{4}{9} \frac{R}{\rho}\left(\frac{R}{\rho} \cos N - \frac{1}{2}\right) \sin N . l^2 + \ldots ,$$

et par suite, d'après l'équation (2), la valeur de tang v exacte jusqu'à la deuxième puissance inclusivement sera

$$\operatorname{tang} v = \frac{2}{3}\left(\frac{R}{\rho} \cos N - 1\right) l + \frac{4}{9} \frac{R}{\rho}\left(\frac{R}{\rho} \cos N - \frac{1}{2}\right) \sin N . l^2 .$$

Mais l'approximation admise permet de remplacer la tangente par son arc, ce qui donne

$$(3)\qquad v=\frac{2}{3}\left(\frac{R}{\rho}\cos N-1\right)l+\frac{4}{9}\,\frac{R}{\rho}\left(\frac{R}{\rho}\cos N-\frac{1}{2}\right)\sin N.l^2.$$

Telle doit être la variation de l'inclinaison de la normale commune aux profils en prise après une rotation de l'angle $ACA_1=\frac{2}{3}l$ de la roue C, pour que cette normale passe par le point de contact des circonférences primitives lorsque $ACA_1=\frac{2}{3}l$, ce qui doit avoir lieu pour l'engrenage considéré.

Cela posé, il est facile de trouver une relation entre les quantités

$$\rho,\ \rho_1,\ N,$$

qui déterminent les arcs DE et DF; à cet effet il faut connaître la valeur de v pour des valeurs quelconques de

$$\rho,\ \rho_1,\ N.$$

C'est ce que nous allons chercher.

§ 6. Soient (fig. 3) A et B les positions des centres des arcs consi-

Fig. 3.

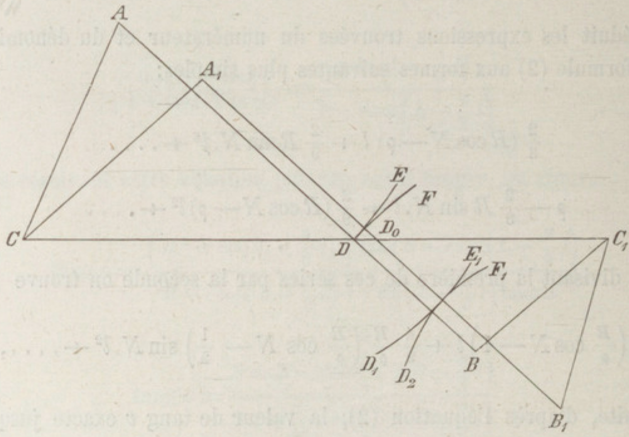

dérés, quand leur point de contact est sur la ligne des centres;

$$ACD=L,\ ADC=N,\ CD=R,\ C_1D=R_1,$$

A_1 et B_1 les positions des centres des arcs après que la roue C a tourné de l'angle quelconque $ACA_1 = \lambda$,

$$A_1CD = L - \lambda, \quad A_1D_0C = N - v;$$

soient aussi $BC_1D = M$ et μ l'accroissement de cet angle correspondant à l'angle de la rotation $ACA_1 = \lambda$ de la roue C; alors

$$B_1C_1D = M + \mu.$$

Pour obtenir l'équation qui détermine la valeur de v pour une valeur quelconque de λ, projetons le polygone $CA_1B_1C_1$ sur la direction de la ligne des centres et sur une perpendiculaire à cette direction; nous aurons

$$A_1C.\cos A_1CD + A_1B_1 \cos A_1D_0C + B_1C_1 \cos B_1C_1D = CC_1$$
$$A_1C \sin A_1CD - A_1B_1 \sin A_1D_0C + B_1C_1 \sin B_1C_1D = 0,$$

d'où, en remarquant que

$$A_1C = r, \quad B_1C_1 = r_1, \quad A_1B_1 = \rho + \rho_1, \quad CC_1 = R + R_1$$
$$A_1CD = L - \lambda, \quad A_1D_0C = N - v, \quad DC_1B_1 = M + \mu,$$

on trouve

$$r\cos(L-\lambda) + (\rho+\rho_1)\cos(N-v) + r_1\cos(M+\mu) = R + R_1,$$
$$r\sin(L-\lambda) - (\rho+\rho_1)\sin(N-v) + r_1\sin(M+\mu) = 0.$$

Éliminant $(M + \mu)$ entre ces équations, on a:

$$r_1^2 = [R + R_1 - r\cos(L-\lambda) - (\rho+\rho_1)\cos(N-v)]^2$$
$$+ [-r\sin(L-\lambda) + (\rho+\rho_1)\sin(N-v)]^2,$$

ou en supprimant les parenthèses

$$r_1^2 = (R+R_1)^2 + (\rho+\rho_1)^2 + r^2 + 2r(\rho+\rho_1)\cos(L+N-\lambda-v)$$
$$-2(R+R_1)\left[r\cos(L-\lambda) + (\rho+\rho_1)\cos(N-v)\right].$$

Pour les valeurs particulières $\lambda = 0$, $v = 0$, cette égalité devient:

$$r_1^2 = (R + R_1)^2 + (\rho + \rho_1)^2 + r^2 + 2r(\rho + \rho_1)\cos(L + N)$$
$$- 2(R + R_1)\left[r \cos L + (\rho + \rho_1)\cos N\right].$$

Soustrayant cette équation de la précédente, on obtient

$$\frac{\cos(L + N - \lambda - v) - \cos(L + N)}{R + R_1} - \frac{\cos(N - v) - \cos N}{r} - \frac{\cos(L - \lambda) - \cos L}{\rho + \rho_1} = 0.$$

Pour en déduire la valeur de l'angle v développée suivant les puissances ascendantes de l'angle λ posons

$$v = K_1\lambda + K_2\lambda^2 + \ldots.$$

et déterminons les coefficients K_1, K_2, ...

En portant cette expression de v dans l'équation précédente et développant suivant les puissances de λ, nous aurons

$$\left.\begin{array}{l} \left\{\dfrac{\sin(L + N)}{R + R_1}(K_1 + 1) - \dfrac{\sin N}{r}K_1 - \dfrac{\sin L}{\rho + \rho_1}\right\}\lambda \\ + \left\{\begin{array}{c} \dfrac{2\sin(L + N)K_2 - \cos(L + N)(K_1 + 1)^2}{R + R_1} \\ - \dfrac{2\sin N.K_2 - \cos N.K_1^2}{r} + \dfrac{\cos L}{\rho + \rho_1} \end{array}\right\}\dfrac{\lambda^2}{2} + \ldots \end{array}\right\} = 0,$$

d'où pour la détermination des coefficients K_1, K_2 on tire les équations suivantes:

$$\frac{\sin(L + N)}{R + R_1}(K_1 + 1) - \frac{\sin N}{r}K_1 - \frac{\sin L}{\rho + \rho_1} = 0,$$

$$\frac{2\sin(L + N)K_2 - \cos(L + N)(K_1 + 1)^2}{R + R_1} - \frac{2\sin N.K_2 - \cos N.K_1^2}{r} + \frac{\cos L}{\rho + \rho_1} = 0,$$

. . . .

qui donnent

$$K_1 = \frac{\dfrac{\sin L}{\rho + \rho_1} - \dfrac{\sin(L + N)}{R + R_1}}{\dfrac{\sin(L + N)}{R + R_1} - \dfrac{\sin N}{r}};$$

$$K_2 = \frac{\dfrac{\cos(L + N)}{R + R_1}(K_1 + 1)^2 - \dfrac{\cos N}{r}K_1^2 - \dfrac{\cos L}{\rho + \rho_1}}{2\left(\dfrac{\sin(L + N)}{R + R_1} - \dfrac{\sin N}{r}\right)}.$$

Afin de chasser de ces expressions les angles

$$L = ACD \text{ et } L + N = ADC + ACD$$

remarquons, que le triangle ACD donne

$$\sin ACD = \frac{AD}{AC} \sin ADC,$$

$$\sin (ADC + ACD) = \sin CAD = \frac{CD}{AC} \sin ADC;$$

$$\cos ACD = \frac{CD - AD . \cos ADC}{AC},$$

$$\cos (ADC + ACD) = - \cos CAD = - \frac{AD - CD . \cos ADC}{AC},$$

ou, d'après la notation adoptée,

$$\sin L = \frac{\rho}{r} \sin N; \ \sin (L + N) = \frac{R}{r} \sin N;$$

$$\cos L = \frac{R - \rho \cos N}{r}; \ \cos (L + N) = - \frac{\rho - R \cos N}{r}.$$

En portant ces valeurs de

$$\sin L, \ \sin (L + N), \ \cos L, \ \cos (L + N)$$

dans les expressions précédentes des coefficients K_1, K_2, on trouve, toutes les réductions faites,

$$K_1 = \frac{R \rho_1 - R_1 \rho}{R_1 (\rho + \rho_1)};$$

$$K_2 = \frac{(\rho - R \cos N)(K_1 + 1)^2 + (R + R_1) \left[\cos N . K_1{}^2 + \frac{R - \rho \cos N}{\rho + \rho_1} \right]}{2 R_1 \sin N}.$$

Telles sont les valeurs des coefficients K_1 et K_2 de la série

$$v = K_1 \lambda + K_2 \lambda^2 + \ldots, \tag{4}$$

qui donne la valeur de l'angle v correspondant à une valeur quelconque de l'angle λ pour toutes les valeurs de

$$\rho, \ \rho_1, \ N.$$

§ 7. En posant dans la formule trouvée

$$\lambda = \frac{2}{3} l,$$

nous aurons la valeur correspondante de l'angle v

$$v = \frac{2}{3} K_1 l + \frac{4}{9} K_2 l^2 + \ldots .$$

Or dans le cas de l'engrenage considéré d'après la formule (3) (§ 5) cette valeur particulière de v est

$$\frac{2}{3}\left(\frac{R}{\rho}\cos N - 1\right) l + \frac{4}{9}\,\frac{R}{\rho}\left(\frac{R}{\rho}\cos N - \frac{1}{2}\right)\sin N \,.\, l^2.$$

Comparant ces deux valeurs du même angle on obtient une équation, qui, étant divisée par $\frac{2}{3} l$, prend la forme

$$(5)\qquad K_1 - \frac{R}{\rho}\cos N + 1 + \frac{2}{3}\left[K_2 - \frac{R}{\rho}\left(\frac{R}{\rho}\cos N - \frac{1}{2}\right)\sin N\right] l = 0,$$

où K_1 et K_2 ont les valeurs indiquées au § 6.

La valeur de K_1, comme nous avons vu, est très simple:

$$(6)\qquad K_1 = \frac{R\rho_1 - R_1\rho}{R_1(\rho + \rho_1)}.$$

Quant à l'expression de la valeur de K_2 elle peut être considérablement simplifiée en vertu de l'approximation, à laquelle on se borne.

Remarquons que dans le calcul de cette valeur on peut négliger même les premières puissances de l, puisque cette quantité n'entre comme facteur que dans le dernier des termes retenus. En outre, d'après l'équation (5), si l'on néglige le second terme, on a

$$(7)\qquad K_1 - \frac{R}{\rho}\cos N + 1 = 0,$$

équation, à l'aide de laquelle on simplifie facilement l'expression de la valeur K_2.

En effet le numérateur de cette expression, K_1 étant remplacé par la différence

$$\frac{R}{\rho}\cos N - 1,$$

devient

$$(\rho + R_1 \cos N)\frac{R^2}{\rho^2}\cos^2 N - 2\,(R + R_1)\frac{R}{\rho}\cos^2 N + \frac{R + R_1}{\rho + \rho_1}(R + \rho_1 \cos N).$$

Mais la formule (7), d'après (6), donne

$$\frac{R\rho_1 - R_1\rho}{R_1(\rho + \rho_1)} - \frac{R}{\rho}\cos N + 1 = 0,$$

formule, qui se réduit à

$$\frac{R + R_1}{\rho + \rho_1} = \frac{RR_1}{\rho\rho_1}\cos N. \tag{8}$$

En remplaçant le facteur

$$\frac{R + R_1}{\rho + \rho_1}$$

dans le dernier terme de l'expression précédente par

$$\frac{RR_1}{\rho\rho_1}\cos N,$$

on la réduit à la suivante:

$$\left[\frac{R^2R_1}{\rho^2}\cos^2 N - \frac{(R + R_1)R}{\rho}\cos N + \frac{R^2R_1}{\rho\rho_1}\right]\cos N,$$

qui après la substitution

$$\frac{RR_1(\rho + \rho_1)}{\rho\rho_1}\cos N$$

au lieu de

$$R + R_1$$

devient

$$\frac{R^2R_1}{\rho\rho_1}(1 - \cos^2 N)\cos N$$

ou

$$\frac{R^2R_1}{\rho\rho_1}\sin^2 N . \cos N.$$

Remplaçons ici

$$\frac{\cos N}{\rho\rho_1}$$

d'après la formule (8) par

$$\frac{R + R_1}{\rho + \rho_1}\,\frac{1}{RR_1}.$$

et nous aurons

$$\frac{R(R + R_1)}{\rho + \rho_1}\sin^2 N,$$

expression réduite du numérateur de K_2, donné à la fin du § 6.

Par suite on obtient pour la détermination du coefficient K_2 la formule suivante très simple:

$$K_2 = \frac{R}{2R_1} \frac{R+R_1}{\rho+\rho_1} \sin N. \tag{9}$$

Portant les valeurs trouvées de K_1 et K_2 dans l'équation (5) et remplaçant dans son dernier terme, d'après (8),

$$\frac{R.\cos N}{\rho}$$

par

$$\frac{R+R_1}{R_1} \frac{\rho_1}{\rho+\rho_1},$$

on obtient, la réduction faite,

$$\cos N - \frac{R+R_1}{RR_1} \frac{\rho\rho_1}{\rho+\rho_1} + \frac{1}{3R_1} \frac{(R_1+2R)\,\rho_1 - (R+2R_1)\,\rho}{\rho+\rho_1} \sin N.l = 0. \tag{10}$$

§ 8. Cette équation exprime une propriété géometrique très-simple des arcs de cercle qui forment les profils de la face et du flanc de la dent de l'engrenage considéré. Mais avant de la faire ressortir cherchons la formule qui sert à déterminer les longueurs de ces arcs.

Conservant les notations précédentes des arcs DE, DF et de leurs centres A et B au moment, où le point de contact des dents est sur la ligne des centres, soient (fig. 4) D_1E_1, D_2E_1, A_1, B_1, les positions extrêmes des

Fig. 4.

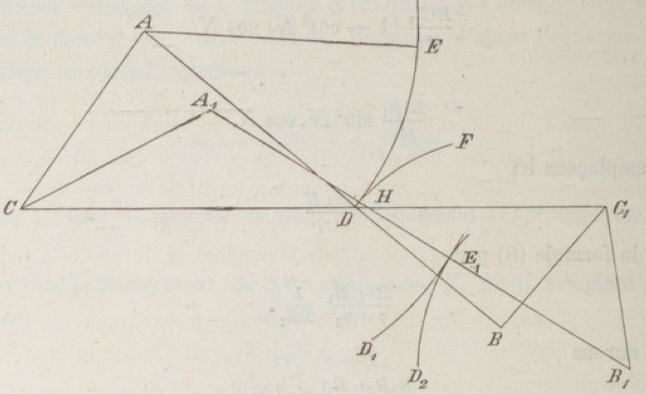

arcs et de leurs centres au moment, où les dents vont se quitter, ce qui correspond, comme on l'a vu, à la rotation de l'angle $ACA_1 = l$ de la roue C.

Traçons la droite AE; l'angle CAE est égal à CA_1E_1, parceque A et A_1 de même que E et E_1 représentent les mêmes points de la roue C dans ses deux positions; donc l'angle DAE qui est égal à $CAE - CAD$ sera égal aussi à $CA_1H - CAD$; or

$$\begin{aligned} CA_1H &= \pi - A_1CH - A_1HC, \\ CAD &= \pi - ACD - ADC, \end{aligned}$$

donc

$$DAE = ACD - A_1CH + ADC - A_1HC.$$

D'après nos notations:

$$ACD - A_1CH = ACA_1 = l,$$

$$ADC - A_1HC = N - (N - v) = v,$$

où la valeur de v s'obtient de la formule (4) pour $\lambda = l$; il s'ensuit, que si nous appellons ω l'angle DAE, qui détermine la longueur de l'arc DE, profil de la face de la dent de l'engrenage considéré, on a pour la valeur de cet angle

$$\omega = l + v = (K_1 + 1)\, l + K_2 l^2,$$

ce qui, après la substitution des valeurs de K_1 et K_2 trouvées au § 7, se réduit à la forme suivante:

$$\omega = \frac{R + R_1}{R_1} \frac{\rho_1}{\rho + \rho_1}\, l + \frac{R}{2R_1} \frac{R + R_1}{\rho + \rho_1} \sin N . l^2. \tag{11}$$

A l'aide de cette expression de l'angle $\omega = DAE$ il est facile à faire voir, que l'équation (10) représente une simple propriété géométrique de l'arc DE, à savoir:

«Au moment où le point de contact des deux dents est sur la ligne des centres, le point extrême E de l'arc DE (fig. 5) se trouve sur la circonférence, qui passe par le point D de contact primitif et par les extrémités P et P_1 des deux arcs DP et DP_1 des circonférences primitives, qui correspondent à l'angle de rotation de la roue C pendant la durée de la prise des arcs DE et DF».

Pour le démontrer traçons la circonférence, passant par les points D, P et P_1, et la corde DE; soit E_1 le point de leur intersection; cherchons la longueur DE_1.

D'après nos notations

$$ADC = N; \quad DAE = \omega, \quad DCP = l, \quad CD = R, \quad C_1D = R_1,$$

et l'angle DC_1P_1 est celui dont a tourné la roue C_1 pendant toute la durée de la prise des arcs DE et DF, de sorte que

$$DC_1P_1 = \frac{R}{R_1}\,DCP = \frac{R}{R_1}\,l.$$

Fig. 5.

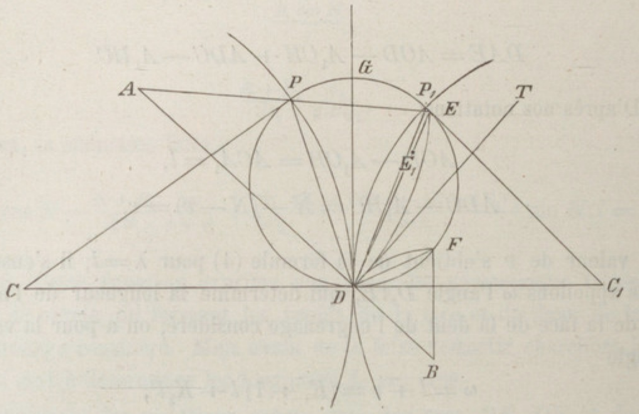

Menons par le point D la droite DG perpendiculaire à CC_1 et DT perpendiculaire à AD; ces droites étant tangentes aux arcs DP, DP_1 et DE respectivement font avec les cordes de ces arcs les angles

$$PDG = \tfrac{1}{2}\,DCP,\ P_1DG = \tfrac{1}{2}\,DC_1P_1,\ EDT = \tfrac{1}{2}\,DAE,$$

ou d'après les notations adoptées

$$PDG = \frac{l}{2};\ P_1DG = \frac{Rl}{2R_1};\ EDT = \frac{\omega}{2}.$$

Quant à l'angle GDT, ses côtés étant perpendiculaires à ceux de l'angle ADC, on a $GDT = ADC$ et par suite $GDT = N$.

On en déduit:

$$PDP_1 = PDG + P_1DG = \frac{l}{2} + \frac{R}{2R_1}\,l,$$

$$P_1DE_1 = GDT - GDP_1 - EDT = N - \frac{R}{2R_1}\,l - \frac{\omega}{2},$$

$$PDE_1 = GDT + PDG - EDT = N + \frac{l}{2} - \frac{\omega}{2}.$$

D'autre part ayant

$$DCP = l,\ DC_1P_1 = \frac{R}{R_1}\, l$$
$$DC = R,\ DC_1 = R_1,$$

on trouve les longueurs des cordes DP et DP_1:

$$DP = 2\,.\,CD\,.\sin\frac{PCD}{2} = 2R\sin\frac{l}{2},$$
$$DP_1 = 2C_1D\,.\sin\frac{DC_1P_1}{2} = 2R_1\sin\frac{Rl}{2R_1}.$$

Connaissant les longueurs de ces cordes ainsi que les angles qu'elles forment avec la corde DE_1, on trouve facilement la longueur de cette dernière à l'aide de l'équation:

$$DP\,.\sin(P_1DE_1) - DP_1\sin(PDE_1) + DE_1\sin(PDP_1) = 0,$$

qui a lieu pour trois cordes quelconques passant par un même point de la circonférence.

En y portant les valeurs trouvées des longueurs de DP et DP_1 et des angles P_1DE_1, PDE_1, PDP_1, on obtient pour la détermination de la longueur de DE_1 l'équation suivante:

$$(12)\quad 2R\sin\frac{l}{2}\sin\left(N - \frac{Rl}{2R_1} - \frac{\omega}{2}\right) - 2R_1\sin\frac{Rl}{2R_1}\sin\left(N + \frac{l}{2} - \frac{\omega}{2}\right) + DE_1\sin\left(\frac{l}{2} + \frac{Rl}{2R_1}\right) = 0.$$

Pour en déduire la longueur de DE_1 avec une approximation du deuxième ordre, développons les deux premiers termes jusqu'au troisième ordre et le facteur de DE_1 jusqu'au deuxième; nous aurons:

$$\frac{R+R_1}{2R_1}DE_1\,.\,l - \frac{(R+R_1)R}{2R_1}\cos N\,.\,l^2 + \left[\frac{R_1{}^2 - R^2}{12R_1{}^2}Rl - \frac{(R+R_1)R}{4R_1}\,\omega\right]\sin N\,.\,l^2 = 0,$$

d'où

$$DE_1 = R\cos N\,.\,l + R\left\{\frac{\omega}{2} + \frac{R-R_1}{6R_1}\,l\right\}\sin N\,.\,l,$$

Remplaçant ω par sa valeur (11) on trouve

$$DE_1 = R\cos N\,.\,l + R\,\frac{(4R+2R_1)\,\rho_1 + (R-R_1)\,\rho}{6R_1(\rho+\rho_1)}\sin N\,.\,l^2.$$

expression exacte jusqu'à la deuxième puissance de l inclusivement de la longueur de DE_1, qui détermine la position du point E_1, où la circonférence passant par les trois points D, P, P_1 est coupée par la corde DE.

En déterminant d'après les formules trouvées plus haut la longueur de la corde DE, on voit qu'avec une approximation du deuxième ordre on peut remplacer cette corde par l'arc DE et par suite exprimer sa longueur par le produit

$$\rho \, . \, \omega.$$

A l'aide de la formule (11) on trouve

$$DE = \frac{R+R_1}{R_1} \frac{\rho\rho_1}{\rho+\rho_1} l + \frac{R\rho}{2R_1} \frac{R+R_1}{\rho+\rho_1} \sin N . l^2.$$

Nous aurons donc pour la différence des longueurs des cordes DE et DE_1 l'expression suivante exacte jusqu'à la deuxième puissance de l:

$$DE - DE_1 =$$

$$\left(\frac{R+R_1}{R_1} \frac{\rho\rho_1}{\rho+\rho_1} - R \cos N\right) l - \frac{R}{3R_1} \frac{(2R+R_1)\rho_1 - (2R_1+R)\rho}{\rho+\rho_1} \sin N . l^2.$$

Comparant cette expression avec le premier membre de l'équation (10) on voit, qu'elle ne diffère de celui-ci que par un facteur $-lR$; par suite en vertu de l'équation (10) on a

$$DE - DE_1 = 0,$$

donc le point E_1 coïncide avec le point E, qui est par conséquent sur la circonférence considérée, ce qu'il fallait démontrer.

§ 9. Répétant relativement à l'arc DF, profil du flanc de la dent de la roue C_1, tout ce qui vient d'être dit relativement à l'arc DE, profil de la face de la dent de la roue C, on arrive aux formules, qui se déduisent des formules précédentes en remplaçant des lettres relatives à la roue C par celles relatives à la roue C_1 et réciproquement.

Quant à la quantité l, qui représente l'angle de la rotation de la roue C pendant la prise des arcs DE et DF, il faut la remplacer par

$$-\frac{R}{R_1} l,$$

parce que la roue C_1 en même temps tourne de l'angle $\frac{R}{R_1} l$ dans le sens opposé.

Au lieu de l'angle $\omega = DAE$, qui mesure l'inclinaison réciproque des rayons AD et AE passant par les extrémités de l'arc DE, nous aurons

$$\omega_1 = DBF$$

avec le signe — à cause de la position inverse des rayons BD et BF, passant par les extrémités de l'arc DF.

L'équation (11) deviendra par ces changements

$$\omega_1 = \frac{R+R_1}{R} \frac{\rho}{\rho+\rho_1} l - \frac{R}{2R_1} \frac{R+R_1}{\rho+\rho_1} \sin N . l^2; \tag{13}$$

l'équation (10) reste la même, quand on passe de l'arc DE à l'arc DF; ces deux équations nous conduisent au même résultat relativement à l'arc DF, que nous avons obtenu dans le paragraphe précédent relativement à l'arc DE, c'est à dire, l'extrémité F est située sur la circonférence passant par les points D, P et P_1.

On arrive ainsi à la conclusion suivante, concernant l'engrenage conconsidéré:

«Au moment, où le point de contact des dents est sur la ligne des centres, les cinq points suivants sont sur une même circonférence: les extrémités des profils de la face de l'une et de la partie vive du flanc de l'autre des deux dents qui sont en contact sur la ligne des centres, les points des circonférences primitives qui viennent à la ligne des centres à la fin de la prise de ces dents, et le point de contact des circonférences primitives».

Cette propriété n'est exacte que jusqu'à la troisième puissance de l, puisque en déterminant les positions des différents points nous avons négligé dans nos formules les puissances de l supérieures à la seconde. Mais cette approximation est suffisante dans la pratique en vertu de la petitesse de la quantité l, qui représente l'angle de la rotation de la roue pendant la prise d'un couple des dents de l'un ou de l'autre côté de la ligne des centres.

En se basant sur les formules (11) et (13), on peut trouver une autre propriété de l'engrenage considéré. Ces formules étant ajoutées donnent

$$\omega + \omega_1 = \frac{R+R_1}{R_1} l,$$

mais d'après nos notations

$$\omega = DAE, \quad \omega_1 = DBF, \quad l = DCP,$$

et comme nous avons remarqué plus haut

$$\frac{Rl}{R_1} = DC_1P_1;$$

donc l'équation trouvée se réduit à la suivante:

$$DAE + DBF = DCP + DC_1P_1.$$

Considérant les angles que forment les cordes des arcs DP, DP_1, DE et DF avec leurs tangentes DG et DT, on voit que

$$PDG = \frac{1}{2}\, DCP,\ \ P_1DG = \frac{1}{2}\, DC_1P_1,$$

$$EDT = \frac{1}{2}\, DAE,\ \ FDT = \frac{1}{2}\, DBF,$$

par suite l'égalité précédente donne

$$EDT + FDT = PDG + P_1DG,$$

d'où

$$EDF = PDP_1,$$

c'est-à-dire que les cordes des arcs DE et DF font entre elles le même angle que les cordes des arcs de retraite DP et DP_1; les points P, P_1, F, E étant situés sur la même circonférence passant par le point D, il en résulte, que les droites PP_1 et EF sont égales entre elles.

On peut énoncer cette propriété de l'engrenage considéré comme il suit:

«La distance entre les extrémités des profils de la face de l'une et de la partie vive du flanc de l'autre des deux dents, dont le point de contact est sur la ligne des centres, est égale à la distance des points des circonférences primitives, qui viennent à la ligne des centres à la fin de la prise de ces dents».

Puisque dans le tracé des engrenages, on donne les circonférences primitives et les arcs de retraite et d'approche, on pourra toujours, d'après ce que nous avons montré plus haut, tracer la circonférence, où doivent être situées les extrémités des profils de la face et de la partie vive du flanc des deux dents et on connaîtra la distance de ces extrémités au moment, où les dents sont en contact sur la ligne des centres.

Soient (fig. 6) C et C_1 les centres des circonférence primitives, D le point de leur contact. Ayant en vue de déterminer le tracé des profils des dents des roues C et C_1 portons les longueurs DP et DP_1 des arcs d'approche sur les circonférences primitives; puis traçons la circonférence passant par les points D, P, P_1, sur laquelle doivent se trouver les extrémités E et F des profils de la face et de la partie vive du flanc des dents au moment, où le point de leur contact et sur la ligne des centres; la distance de ces extrémités doit être égale à PP_1. Quant au choix de la position des points E et F sur la circonférence, il dépend des exigences de la pratique: plus ces points seront éloignés du point D, plus longue sera la dent de la roue C et plus profond sera le creux de la roue C_1; en même temps sera plus mince le bout de la dent de la roue C et plus étroit près du fond le creux de la roue C_1.

Après avoir choisi les positions des points E et F sur la circonférence et la direction de la normale commune aux profils des dents en contact au point D, nous aurons les profils cherchés, en traçant par le points D et E, D et F les arcs de cercle, ayant au point D la normale choisie.

Fig. 6.

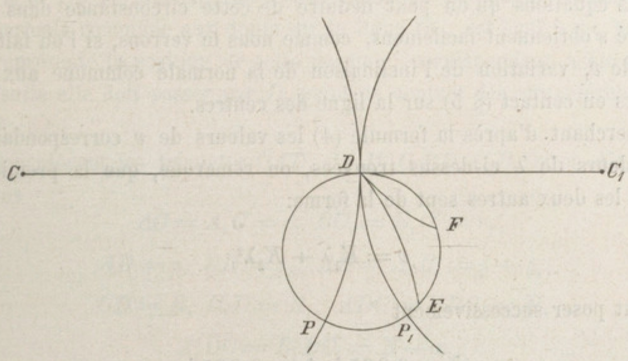

§ 10. Passons maintenant au cas, où l'on dispose de deux quantités dans le choix des arcs de cercle formant les profils des dents, en vue de diminuer autant qu'il est possible les irrégularités, qui sont inévitables une fois que l'engrenage est profilé en arcs de cercle.

Développant la différence:

$$F(\lambda) - \frac{R}{R_1}\lambda$$

suivant les puissances ascendantes de λ jusqu'à la quatrième puissance, on trouve, que d'après le § 3 cette différence se réduit au polynome:

$$K_4(\lambda^4 - 1,638\,l\lambda^3 + 0,638\,l^2\lambda^2).$$

D'où il suit que pour déterminer les valeurs de λ, qui annulent la dérivée

$$F'(\lambda) - \frac{R}{R_1}$$

on à l'équation:

$$4\lambda^3 - 1,638.3.l\lambda^2 + 0,638.2.l^2\lambda = 0.$$

dont les racines sont:

$$\lambda = 0,\ \lambda = 0,365\,l,\ \lambda = 0,865\,l.$$

On conclut de là, d'après le § 4, que la normale commune aux profils des dents en contact passe par le point de contact des circonférences primitives pour les valeurs suivantes de l'angle λ:

$$\lambda = 0,\quad \lambda = 0{,}365\,l,\quad \lambda = 0{,}865\,l.$$

Les équations qu'on peut déduire de cette circonstance dans le cas considéré s'obtiennent facilement, comme nous le verrons, si l'on fait usage de l'angle v, variation de l'inclinaison de la normale commune aux profils des dents en contact (§ 5) sur la ligne des centres.

Cherchant d'après la formule (4) les valeurs de v correspondant aux trois valeurs de λ ci-dessus trouvées, on remarque, que la première est nulle et les deux autres sont de la forme:

$$v = K_1\lambda + K_2\lambda^2,$$

où il faut poser successivement

$$\lambda = 0{,}365\,l,\quad \lambda = 0{,}865\,l.$$

En désignant ces deux valeurs de v par v_1 et v_2 nous aurons

$$v_1 = 0{,}365\,K_1 l + 0{,}365^2 K_2 l^2,$$
$$v_2 = 0{,}865\,K_1 l + 0{,}865^2 K_2 l^2.$$

Soient, conformément aux notations précédentes, C et C_1 (fig. 7) les

Fig. 7.

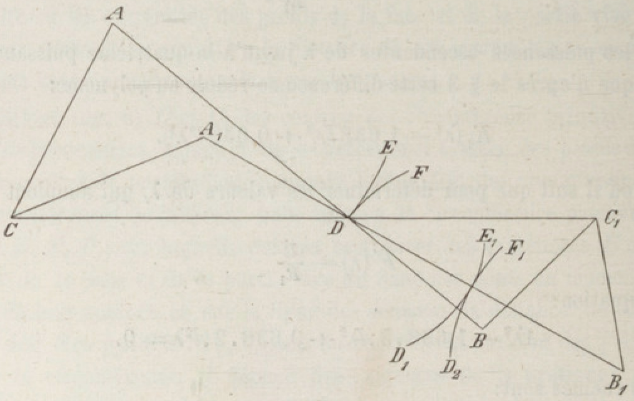

centres des circonférences primitives, $CD = R$, $C_1D = R_1$ leurs rayons; A

et B les centres des arcs de cercle DE et DF, lorsque leur point de contact est sur la ligne des centres; $AD=\rho$, $BD=\rho_1$ les rayons de ces arcs; $AC=r$, $BC_1=r_1$ les distances des points A, B aux centres C, C_1; $ADC=N$ l'angle, qui mesure l'inclinaison de la normale commune aux arcs DE et DF sur la ligne des centres. Si A_1 et B_1 sont les positions des centres des arcs DE et DF au moment, où le décroissement de l'angle de l'inclinaison de la normale commune aux dents sur la ligne CC_1 est égal à v_1, la droite A_1B_1 représente la position de cette normale, correspondant à $\lambda=0{,}365\,l$; et par suite elle doit passer par D, point de contact des circonférences primitives.

Considérant les triangles CAD, CA_1D, C_1BD, C_1B_1D, où d'après nos notations

$$AC=A_1C=r,\ BC_1=B_1C_1=r_1,$$
$$AD=\rho,\ BD=\rho_1,\ AB=A_1B_1=\rho+\rho_1,$$
$$CD=R,\ C_1D=R_1,\ ADC=BDC_1=N,$$
$$A_1DC=B_1DC_1=N-v_1,$$

et posant

$$AD-A_1D=B_1D-BD=h,$$

nous aurons les égalités suivantes:

$$r^2=\rho^2+R^2-2\rho R\cos N,$$
$$r^2=(\rho-h)^2+R^2-2(\rho-h)R\cos(N-v_1),$$
$$r_1{}^2=\rho_1{}^2+R_1{}^2-2\rho_1R_1\cos N,$$
$$r_1{}^2=(\rho_1+h)^2+R_1{}^2-2(\rho_1+h)R_1\cos(N-v_1),$$

qui, après l'élimination de r et r_1, nous donnent:

$$\rho^2+R^2-2\rho R\cos N=(\rho-h)^2+R^2-2(\rho-h)R\cos(N-v_1),$$
$$\rho_1{}^2+R_1{}^2-2\rho_1R_1\cos N=(\rho_1+h)^2+R_1{}^2-2(\rho_1+h)R_1\cos(N-v_1).$$

Pour éliminer de ces équations la quantité h on peut les résoudre par rapport à

$$\rho-h \text{ et } \rho_1+h$$

et ajouter les résultats.

On arrive ainsi à l'équation:

$$\rho+\rho_1-(R+R_1)\cos(N-v_1)=$$
$$\sqrt{\rho^2-2R\rho\cos N+R^2\cos^2(N-v_1)}+\sqrt{\rho_1{}^2-2R_1\rho_1\cos N+R_1{}^2\cos^2(N-v_1)},$$

qui peut être réduite à la forme:

$$\begin{aligned}\rho+\rho_1-(R+R_1)\cos(N-v_1)= \\ \sqrt{(\rho-R\cos N)^2+R^2\left[\cos^2(N-v_1)-\cos^2 N\right]} \\ +\sqrt{(\rho_1-R_1\cos N)^2+R_1^2\left[\cos^2(N-v_1)-\cos^2 N\right]}.\end{aligned}$$

Posons pour abréger

$$\frac{\cos^2(N-v_1)-\cos^2 N}{\cos^2 N}=t_1, \tag{14}$$

ce qui donne

$$\cos(N-v_1)=\cos N\sqrt{1+t_1},$$

et nous pourrons écrire l'équation précédente comme il suit:

$$\begin{gathered}\rho+\rho_1-(R+R_1)\cos N\sqrt{1+t_1}= \\ \sqrt{(\rho-R\cos N)^2+R^2\cos^2 N.t_1}+\sqrt{(\rho_1-R_1\cos N)^2+R_1^2\cos^2 N.t_1}.\end{gathered}$$

Développant les radicaux suivant les puissances de t_1, on ramène tous les termes dans le premier membre; divisant par t_1, on trouve:

$$\left.\begin{aligned}&\frac{R^2\cos N}{\rho-R\cos N}+\frac{R_1^2\cos N}{\rho_1-R_1\cos N}+R+R_1 \\ -\frac{1}{4}&\left[\frac{R^4\cos^3 N}{(\rho-R\cos N)^3}+\frac{R_1^4\cos^3 N}{(\rho_1-R_1\cos N)^3}+R+R_1\right]t_1 \\ +\frac{1}{8}&\left[\frac{R^6\cos^5 N}{(\rho-R\cos N)^5}+\frac{R_1^6\cos^5 N}{(\rho_1-R_1\cos N)^5}+R+R_1\right]t_1^2 \\ +&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\end{aligned}\right\}=0.$$

Faisons pour abréger

$$\frac{R\cos N}{\rho-R\cos N}=X, \quad \frac{R_1\cos N}{\rho_1-R_1\cos N}=Y, \tag{15}$$

et l'équation ci-dessus prendra la forme suivante:

$$\left.\begin{aligned}RX+R_1Y+R+R_1-\tfrac{1}{4}(RX^3+R_1Y^3+R+R_1)\,t_1 \\ +\tfrac{1}{8}(RX^5+R_1Y^5+R+R_1)\,t_1^2+\ldots\end{aligned}\right\}=0. \tag{16}$$

En répétant les mêmes opérations pour l'angle v_2, qui correspond à l'angle de la rotation $\lambda = 0{,}865\, l$, pour lequel la normale commune doit passer aussi par le point D et posant:

$$\frac{\cos^2 (N - v_2) - \cos^2 N}{\cos^2 N} = t_2, \tag{17}$$

nous obtenons l'équation

$$\left. \begin{array}{l} RX + R_1 Y + R + R_1 - \frac{1}{4} (RX^3 + R_1 Y^3 + R + R_1)\, t_2 \\ \quad + \frac{1}{8} (RX^5 + R_1 Y^5 + R + R_1)\, t_2^2 + \ldots \end{array} \right\} = 0, \tag{18}$$

analogue à l'équation (16).

Des équations (16) et (18) on déduit les valeurs de X et Y; en suite à l'aide des formules (15) on trouve les valeurs de ρ et ρ_1, rayons des arcs de cercle formant les profils des dents.

§ 11. Pour faciliter la résolution des équations (16) et (18) nous négligerons d'abord les termes contenant les puissances de t_1 et t_2 supérieures à la première. Nous aurons pour les valeurs approchées de X et Y les équations:

$$RX + R_1 Y + R + R_1 - \frac{1}{4} (RX^3 + R_1 Y^3 + R + R_1)\, t_1 = 0,$$

$$RX + R_1 Y + R + R_1 - \frac{1}{4} (RX^3 + R_1 Y^3 + R + R_1)\, t_2 = 0$$

qui se réduisent aux équations suivantes:

$$RX + R_1 Y + R + R_1 = 0,$$

$$RX^3 + R_1 Y^3 + R + R_1 = 0.$$

On déduit de la première

$$Y = -\frac{R}{R_1} X - \frac{R}{R_1} - 1. \tag{19}$$

Substituant cette valeur de Y dans la seconde on arrive à l'équation:

$$RX^3 + R_1 \left(-\frac{R}{R_1} X - \frac{R}{R_1} - 1 \right)^3 + R + R_1 = 0,$$

qui, les réductions faites, aura la forme:

$$(R_1 - R)\, X^3 - 3RX^2 - 3\, (R + R_1)\, X - R - 2R_1 = 0.$$

Cette équation a deux racines égales à -1 et la troisième, qui est égale à

$$\frac{R+2R_1}{R_1-R}.$$

Or, remarquant que d'après la première des formules (15) X ne peut devenir égal à -1 que si le rayon ρ devient nul, on conclut que les racines égales à -1 ne sont pas admissibles dans notre problème; donc il faut prendre pour la valeur cherchée de X la racine

$$\frac{R+2R_1}{R_1-R}.$$

Substituant cette valeur de X dans l'équation (19) on trouve

$$Y=\frac{R_1+2R}{R-R_1}.$$

Telles sont les premières valeurs approchées de X et de Y. Pour les déterminer avec plus d'exactitude, posons:

$$(20)\qquad \left\{\begin{aligned} X&=\frac{R+2R_1}{R_1-R}+\alpha,\\ Y&=\frac{R_1+2R}{R-R_1}+\beta,\end{aligned}\right.$$

où α et β sont les quantités négligées dans la première approximation.

Introduisant ces nouvelles valeurs de X et de Y dans les équations (16) et (18) et supprimant les termes, qui contiennent les puissances des quantités α, β, t_1, t_2 supérieures à la première, on obtient:

$$\left.\begin{aligned} &R\alpha+R_1\beta-\frac{3}{4}\left[R\left(\frac{R+2R_1}{R_1-R}\right)^2\alpha+R_1\left(\frac{R_1+2R}{R-R_1}\right)^2\beta\right]t_1\\ &\quad+\frac{1}{8}\left[R\left(\frac{R+2R_1}{R_1-R}\right)^5+R_1\left(\frac{R_1+2R}{R-R_1}\right)^5+R+R_1\right]t_1^2\end{aligned}\right\}=0,$$

$$\left.\begin{aligned} &R\alpha+R_1\beta-\frac{3}{4}\left[R\left(\frac{R+2R_1}{R_1-R}\right)^2\alpha+R_1\left(\frac{R_1+2R}{R-R_1}\right)^2\beta\right]t_2\\ &\quad+\frac{1}{8}\left[R\left(\frac{R+2R_1}{R_1-R}\right)^5+R_1\left(\frac{R_1+2R}{R_1-R}\right)^5+R+R_1\right]t_2^2\end{aligned}\right\}=0.$$

Si l'on se borne à la première puissance de t_1 et t_2 dans la détermination des valeurs de α et β, les deux équations précédentes donnent immédiatement l'équation

$$R\alpha+R_1\beta=0.$$

Pour obtenir une seconde équation simple entre α et β soustrayons l'une de l'autre les mêmes équations et, supprimant le facteur commun $\frac{3}{4}(t_2 - t_1)$, nous aurons:

$$R\left(\frac{R+2R_1}{R_1-R}\right)^2\alpha + R_1\left(\frac{R_1+2R}{R-R_1}\right)^2\beta =$$

$$\frac{1}{6}\left[R\left(\frac{R+2R_1}{R_1-R}\right)^5 + R_1\left(\frac{R_1+2R}{R-R_1}\right)^5 + R + R_1\right](t_1+t_2).$$

Remplaçant dans cette équation

$$\beta = -\frac{R}{R_1}\alpha,$$

on a

$$R\left[\left(\frac{R+2R_1}{R_1-R}\right)^2 - \left(\frac{R_1+2R}{R_1-R}\right)^2\right]\alpha =$$

$$\frac{1}{6}\left[R\left(\frac{R+2R_1}{R_1-R}\right)^5 + R_1\left(\frac{R_1+2R}{R-R_1}\right)^5 + R + R_1\right](t_1+t_2),$$

d'où l'on tire

$$\alpha = \frac{R_1(R+2R_1)(R_1+2R)}{2(R_1-R)^3}(t_1+t_2),$$

et par suite

$$\beta = \frac{R(R+2R_1)(R_1+2R)}{2(R_1-R)^3}(t_1+t_2).$$

En portant ces valeurs de α et β, exactes jusqu'à la première puissance de t_1 et t_2 dans les formules (20), on obtient avec le même degré d'approximation:

$$X = \frac{R+2R_1}{R_1-R}\left[1 + \frac{R_1(R_1+2R)}{2(R_1-R)^2}(t_1+t_2)\right],$$

$$Y = \frac{R_1+2R}{R-R_1}\left[1 + \frac{R(R+2R_1)}{2(R_1-R)^2}(t_1+t_2)\right].$$

§ 12. Substituant les valeurs trouvées de X et Y dans les formules (15), on aura pour la détermination des valeurs ρ et ρ_1 les équations suivantes:

$$\frac{R\cos N}{\rho - R\cos N} = \frac{R+2R_1}{R_1-R}\left[1 + \frac{R_1(R_1+2R)}{2(R_1-R)^2}(t_1+t_2)\right],$$

$$\frac{R_1\cos N}{\rho_1 - R_1\cos N} = \frac{R_1+2R}{R-R_1}\left[1 + \frac{R(R+2R_1)}{2(R-R_1)^2}(t_1+t_2)\right].$$

Si l'on se borne aux termes du premier ordre relativement à t_1 et t_2, on en tire:

$$\rho = \frac{3RR_1\cos N}{R+2R_1}\left[1 - \frac{(R_1+2R)(t_1+t_2)}{6(R_1-R)}\right],$$

$$\rho_1 = \frac{3RR_1\cos N}{R_1+2R}\left[1 - \frac{(R+2R_1)(t_1+t_2)}{6(R-R_1)}\right].$$

Quant à la valeur de la somme $t_1 + t_2$, qui entre dans ces formules, on remarque que les expressions (14) et (17), étant développées suivant les puissances de v, donnent avec approximation jusqu'à la deuxième puissance

$$t_1 = \frac{2 \cos N . \sin N}{\cos^2 N} v_1 = 2 \operatorname{tang} N . v_1 ; \quad t_2 = \frac{2 \cos N . \sin N}{\cos^2 N} v_2 = 2 \operatorname{tang} N . v_2 .$$

Mais d'après le § 10 on a avec approximation jusqu'à la deuxième puissance de l

$$v_1 = 0{,}365\, K_1 l, \quad v_2 = 0{,}865\, K_1 l,$$

donc avec la même approximation

$$t_1 + t_2 = 2\,(0{,}365 + 0{,}865)\, K_1 \operatorname{tang} N . l = 2{,}460\, K_1 \operatorname{tang} N . l,$$

où d'après (6)

$$K_1 = \frac{R\rho_1 - R_1 \rho}{R_1 (\rho + \rho_1)}$$

Si dans cette expression de K_1 on remplace ρ et ρ_1 par leurs valeurs trouvées ci-dessus, on a avec approximation jusqu'à la première puissance de t_1 et t_2 exclusivement

$$K_1 = \frac{R - R_1}{3R_1} .$$

et par suite

$$t_1 + t_2 = \frac{2{,}460}{3} \frac{R - R_1}{R_1} \operatorname{tang} N . l$$

$$= 0{,}820 . \frac{R - R_1}{R_1} \operatorname{tang} N . l .$$

Substituant cette valeur de $t_1 + t_2$ dans les expressions des valeurs de ρ et ρ_1 trouvées plus haut, nous obtenons

$$\rho = \frac{3RR_1 \cos N}{R + 2R_1} \left[1 + 0{,}137 \frac{R_1 + 2R}{R_1} \operatorname{tang} N . l \right],$$

$$\rho_1 = \frac{3RR_1 \cos N}{R_1 + 2R} \left[1 - 0{,}137 \frac{R + 2R_1}{R_1} \operatorname{tang} N . l \right].$$

Tels sont donc les rayons des arcs de cercle, dont l'un forme le profil de la face de la dent de la roue C et l'autre celui du flanc de la dent de roue C_1.

§ 13. Pour déterminer les longueurs de ces arcs, on a d'après le § 8 avec approximation jusqu'à l^3

$$\omega = (K_1 + 1)\, l + K_2 l^2,$$

où ω est la longueur de l'arc de cercle, profil de la face de la dent de la roue C_1, supposant que le rayon de cet arc est pris pour l'unité, et les coefficients K_1 et K_2 d'après le § 6 ont les valeurs suivantes:

$$K_1 = \frac{R\rho_1 - R_1\rho}{R_1(\rho + \rho_1)};$$

$$K_2 = \frac{(\rho - R\cos N)(K_1 + 1)^2 + (R + R_1)\left[\cos N . K_1^2 + \frac{R - \rho\cos N}{\rho + \rho_1}\right]}{2R_1 \sin N}.$$

Remplaçant ρ et ρ_1 par leurs valeurs trouvées dans le § 12, nous supprimons les termes contenant, comme facteurs, l^2 dans l'expression de K_1 et l dans celle de K_2; nous aurons ainsi:

$$K_1 + 1 = \frac{R + 2R_1}{3R_1}\left(1 - 0{,}137\,\frac{R_1 + 2R}{R_1}\,\operatorname{tang} N . l\right),$$

$$K_2 = \frac{(R + 2R_1)(R_1 + 2R)}{18R_1^2}\operatorname{tang} N.$$

Après la substitution de ces valeurs dans l'expression ci-dessus de ω on obtient avec approximation jusqu'à l^3

$$\omega = \frac{R + 2R_1}{3R_1}\left[1 + \left(\frac{1}{6} - 0{,}137\right)\frac{R_1 + 2R}{R_1}\operatorname{tang} N . l\right] l.$$

Remplaçant dans cette formule ω par $-\omega_1$, l par $-\frac{R}{R_1}\,l$, R par R_1, ρ par ρ_1 et réciproquement on parvient d'après le § 9 à la formule pour déterminer la longueur de l'arc de cercle, qui forme le profil du flanc de la dent de la roue C_1:

$$\omega_1 = \frac{R_1 + 2R}{3R_1}\left[1 - \left(\frac{1}{6} - 0{,}137\right)\frac{R + 2R_1}{R_1}\operatorname{tang} N . l\right] l,$$

où ω_1 désigne la longueur de l'arc, son rayon étant pris pour l'unité.

On en déduit que la somme

$$\omega + \omega_1$$

est égale à

$$\frac{R + R_1}{R_1}\, l$$

et le produit

$$\rho . \omega$$

avec approximation jusqu'à l^3 se réduit à

$$R\cos N . l + \frac{R(R_1 + 2R)}{6R_1}\sin N . l^2.$$

Cette expression est la même que celle de la corde DE_1 (§ 8), qui résulte de la formule

$$DE_1 = R\cos N.l + R\left(\frac{\omega}{2} + \frac{R-R_1}{6R_1}\,l\right)\sin N.l,$$

après la substitution de la valeur de ω. On conclut de là, que dans le cas présent subsistent les propriétés que nous avons demontrées (§ 8, § 9) pour le cas, où dans le choix des arcs de cercle formant le profil de l'engrenage il n'y a qu'une seule quantité dont on peut disposer pour diminuer les irrégularités du mouvement.

Nous avons vu dans le § 9 que dans ce cas on peut modifier suivant les exigences de la pratique la direction de la normale commune aux dents qui sont en contact sur la ligne des centres, sans changer les positions choisies sur la circonférence DPP_1 (fig. 6) pour les extrémités des profils de la face de la dent de la roue C et de la partie vive du flanc de la dent de la roue C_1. Au contraire, dans le cas que nous examinons à présent la position de ces extrémités définit complétement la direction de la normale commune mentionnée, puisque d'après le § 8 (fig. 5) on a:

$$N = GDT = GDE + EDT = GDE + \frac{\omega}{2}.$$

En portant l'expression connue de ω, où ρ et ρ_1 doivent être remplacés par leurs valeurs, on trouve une équation, qui détermine l'angle N, si l'angle GDE est donné. Si l'on a égard à ce fait que c'est dans ce cas notamment que l'on obtient la plus grande régularité possible avec des engrenages profilés en arcs de cercle, on conclura de ce qui précède que toutes les fois, lorsqu'on choisit la direction de la normale commune AB conformément aux exigences de la pratique, il faut chercher donner à cette normale la direction qui s'écarte le moins possible de celle que l'équation précédente fait connaître.

Les valeurs trouvées de ω et ω_1 avec approximation jusqu'à l donnent

$$\frac{\omega_1}{\omega} = \frac{R_1 + 2R}{R + 2R_1},$$

or d'après le § 8 (fig. 5)

$$\omega = DAE = 2EDT,$$
$$\omega_1 = DBF = 2FDT,$$

donc on a avec approximation jusqu'à la première puissance de l

$$\frac{FDT}{EDT} = \frac{R_1 + 2R}{R + 2R_1}.$$

D'où il ressort que, si la normale commune aux dents en contact AB a la direction indiquée, leur tangente commune DT divisé l'arc EF en deux parties, qui sont entre elles dans le rapport

$$\frac{R_1+2R}{R+2R_1},$$

ce rapport étant exact jusqu'à la première puissance de l.

On en déduit une construction géométrique très-simple pour déterminer approximativement la direction de la tangente DT, correspondant à celle de la normale AB, pour laquelle les irregularites du mouvement de l'engrenage considéré deviennent les moindres possibles.

Remarquons pour terminer, que les angles que les cordes PD et P_1D (fig. 5) font avec le diamètre passant par le point D ont pour valeurs approximatives jusqu'à l^2, respectivement:

$$\frac{R+2R_1}{6R_1}\,l,\ \frac{R_1+2R}{6R_1}\,l.$$

En comparant ces valeurs avec celles des angles

$$PDG=\frac{\omega}{2},\ GDP_1=\frac{\omega_1}{2},$$

qui résultent des formules trouvées, on remarque qu'elles sont respectivement égales avec approximation jusqu'à la deuxième puissance de l. On voit de là, qu'avec ce degré d'approximation le diamètre du cercle DPP_1, passant par le point D, fait avec les cordes DP et DP_1 des angles respectivement égaux à ceux, que la tangente DT aux arcs DE et DF doit faire avec le cordes DE et DF, pour que l'irregularité de la transmission du mouvement par l'engrenage soit réduite au minimum possible.

Sur les quadratures.

1. Dans l'Ouvrage très-important que M. Hermite vient de publier sur l'Analyse mathématique, l'illustre géomètre donne une nouvelle formule pour évaluer approximativement la valeur de l'intégrale

$$\int_{-1}^{+1} \frac{\varphi(x)}{\sqrt{1-x^2}}\, dx.$$

Dans cette formule toutes les valeurs de la fonction $\varphi(x)$ entrent avec un même coefficient; c'est ce qui apporte une différence essentielle entre la formule de M. Hermite et celle de Gauss, et ce qui en rend très-commode l'application numérique. L'utilité des formules approximatives de ce genre m'engage à présenter quelques réflexions sur la recherche de ces formules.

Nous supposerons que, la fonction $F(x)$ étant donnée, on cherche à exprimer le plus près possible les intégrales de la forme

$$\int_{-1}^{+1} F(x)\, \varphi(x)\, dx,$$

quelle que soit la fonction $\varphi(x)$, par la formule

$$k\left[\varphi(x_1) + \varphi(x_2) + \ldots + \varphi(x_n)\right],$$

où $k, x_1, x_2, \ldots x_n$ sont des valeurs qui ne dépendent pas de la fonction $\varphi(x)$. Comme cette formule ne contient que $n+1$ quantités $k, x_1, x_2, \ldots x_n$ dont on puisse disposer, il est impossible de l'identifier avec la valeur de l'intégrale

$$\int_{-1}^{+1} F(x)\, \varphi(x)\, dx$$

au delà des termes qui contiennent les n premières dérivées de la fonction $\varphi(x)$, et, par conséquent, on aura

$$(1)\quad \int_{-1}^{+1} F(x)\,\varphi(x)\,dx - k\left[\varphi(x_1)+\varphi(x_2)+\ldots+\varphi(x_n)\right] = k_1\,\varphi^{(n+1)}(0)+k_2\,\varphi^{(n+2)}(0)+\ldots,$$

en designant par $k_1, k_2, \ldots$ les coefficients de $\varphi^{(n+1)}(0)$, $\varphi^{(n+2)}(0), \ldots$ dans l'expression de la différence

$$\int_{-1}^{+1} F(x)\,\varphi(x)\,dx - k\left[\varphi(x_1)+\varphi(x_2)+\ldots+\varphi(x_n)\right],$$

que l'on trouve en développant, d'après la formule de Maclaurin, la fonction $\varphi(x)$ sous le signe de l'intégrale et les valeurs $\varphi(x_1)\ \varphi(x_2), \ldots, \varphi(x_n)$ qui sont hors ce signe.

2. Pour trouver, d'après la formule (1), que l'on suppose possible, la valeur du coefficient k et les valeurs $x_1, x_2, \ldots, x_n$ de la variable x, nous remarquons que cette formule, dans le cas particulieur où

$$\varphi(x) = \frac{1}{z-x},$$

z étant une quantité quelconque, se réduit à l'égalité

$$\int_{-1}^{+1} \frac{F(x)}{z-x}\,dx - k\left(\frac{1}{z-x_1}+\frac{1}{z-x_2}+\ldots+\frac{1}{z-x_n}\right) =$$
$$1.2.3\ldots(n+1)k_1 z^{-n-2} + 1.2.3\ldots(n+2)k_2 z^{-n-3} + \ldots$$

où $k, k_1, k_2, \ldots, x_1, x_2, \ldots, x_n$ sont des valeurs qui ne dépendent pas de z. D'autre part, en dénotant par $f(z)$ le produit

$$(z-x_1)(z-x_2)\ldots(z-x_n),$$

on a

$$\frac{f'(z)}{f(z)} = \frac{1}{z-x_1}+\frac{1}{z-x_2}+\ldots+\frac{1}{z-x_n},$$

et, par suite, la formule précédente se réduit à celle-ci:

$$(2)\quad \int_{-1}^{+1} \frac{F(x)}{z-x}\,dx = k\frac{f'(z)}{f(z)} + \frac{1.2.3\ldots(n+1)k_1}{z^{n+2}} + \frac{1.2.3\ldots(n+2)k_2}{z^{n+3}} + \ldots$$

En multipliant cette formule par z et en remarquant que, pour $z = \infty$, les valeurs

$$z \int_{-1}^{+1} \frac{F(x)}{z-x}\,dx = \int_{-1}^{+1} \frac{F(x)}{1-\frac{x}{z}}\,dx,$$

$$\frac{z f'(z)}{f(z)} = \frac{1}{1-\frac{x_1}{z}} + \frac{1}{1-\frac{x_2}{z}} + \ldots + \frac{1}{1-\frac{x_n}{z}},$$

$$\frac{1.2.3\ldots(n+1)k_1}{z^{n+1}}, \qquad \frac{1.2.3\ldots(n+2)k_2}{z^{n+2}}, \ldots$$

sont respectivement égales à

$$\int_{-1}^{+1} F(x)\,dx, \quad n, \quad 0, \quad 0, \ldots,$$

on parvient à cette égalité

$$\int_{-1}^{+1} F(x)\,dx = nk,$$

ce qui nous donne, pour la détermination du coefficient k, la formule suivante:

$$k = \frac{1}{n} \int_{-1}^{+1} F(x)\,dx.$$

3. Pour déterminer la fonction

$$f(z) = (z-x_1)(z-x_2)\ldots(z-x_n),$$

nous remarquons que la formule (2), étant intégrée par rapport à z, nous donne

$$\int_{-1}^{+1} F(x) \log(z-x)\,dx = k \log\frac{f(z)}{C} - \frac{1.2\,3\ldots nk_1}{z^{n+1}} - \frac{1.2.3\ldots(n+1)k_2}{z^{n+2}} - \ldots,$$

où C est une constante, et de là on trouve

$$f(z)\, e^{-\frac{1.2.3\ldots nk_1}{kz^{n+1}} - \frac{1.2.3\ldots(n+1)k_2}{kz^{n+2}} - \ldots} = Ce^{\frac{1}{k}\int_{-1}^{+1} F(x)\log(z-x)\,dx}$$

Comme la fonction cherchée

$$f(z) = (z-x_1)(z-x_2)\ldots(z-x_n)$$

est de degré n et que l'expression

$$e^{-\frac{1.2.3\ldots nk_1}{kz^{n+1}} - \frac{1.2.3\ldots(n+1)k_2}{kz^{n+2}} - \ldots}$$

ne diffère de 1 que par les puissances de z inférieures à z^{-n}, il est clair que la partie entière du premier membre de la formule trouvée est égale à la fonction $f(z)$, et que, par conséquent, on aura

$$f(z) = ECe^{\frac{1}{k}\int_{-1}^{+1} F(x)\log(z-x)\,dx}$$

ou

$$f(z) = CEe^{\frac{1}{k}\int_{-1}^{+1} F(x)\log(z-x)\,dx}, \tag{3}$$

en désignant par E la partie entière de la fonction mise sous ce signe. Dans cette formule, la valeur de la constante k, comme nous l'avons vu, est donnée par l'équation

$$k = \frac{1}{n}\int_{-1}^{+1} F(x)\,dx. \tag{4}$$

Quant à la constante C, on trouvera aisément sa valeur en remarquant que le coefficient de z^n, dans la fonction cherchée, est égal à 1; mais nous n'insisterons pas sur la recherche de la valeur de cette constante, vu qu'elle peut être toujours supprimée sans modifier l'équation

$$f(z) = CEe^{\frac{1}{k}\int_{-1}^{+1} F(x)\log(z-x)\,dx} = 0,$$

dont les racines présentent les valeurs de $x_1, x_2, \ldots, x_n$ dans la formule en question

$$\int_{-1}^{+1} F(x)\,\varphi(x)\,dx = k\left[\varphi(x_1) + \varphi(x_2) + \ldots + \varphi(x_n)\right].$$

4. Passant aux applications, nous ferons d'abord

$$F(x) = \frac{1}{\sqrt{1-x^2}},$$

ce qui est le cas de M. Hermite, et où l'intégrale

$$\int_{-1}^{+1} F(x)\,\varphi(x)\,dx$$

se réduit à

$$\int_{-1}^{+1} \frac{\varphi(x)}{\sqrt{1-x^2}}\,dx.$$

ne différe de 1 que par les puissances de z inférieures à z^{-n}, il est clair que la partie entière du premier membre de la formule trouvée est égale à la fonction $f(z)$, et que, par conséquent, on aura

$$f(z) = ECe^{\frac{1}{k}\int_{-1}^{+1} F(x)\log(z-x)\,dx}$$

ou

$$f(z) = CEe^{\frac{1}{k}\int_{-1}^{+1} F(x)\log(z-x)\,dx}, \tag{3}$$

en désignant par E la partie entière de la fonction mise sous ce signe. Dans cette formule, la valeur de la constante k, comme nous l'avons vu, est donnée par l'équation

$$k = \frac{1}{n}\int_{-1}^{+1} F(x)\,dx. \tag{4}$$

Quant à la constante C, on trouvera aisément sa valeur en remarquant que le coefficient de z^n, dans la fonction cherchée, est égal à 1; mais nous n'insisterons pas sur la recherche de la valeur de cette constante, vu qu'elle peut être toujours supprimée sans modifier l'équation

$$f(z) = CEe^{\frac{1}{k}\int_{-1}^{+1} F(x)\log(z-x)\,dx} = 0,$$

dont les racines présentent les valeurs de $x_1, x_2, \ldots, x_n$ dans la formule en question

$$\int_{-1}^{+1} F(x)\,\varphi(x)\,dx = k\left[\varphi(x_1) + \varphi(x_2) + \ldots + \varphi(x_n)\right].$$

4. Passant aux applications, nous ferons d'abord

$$F(x) = \frac{1}{\sqrt{1-x^2}},$$

ce qui est le cas de M. Hermite, et où l'intégrale

$$\int_{-1}^{+1} F(x)\,\varphi(x)\,dx$$

se réduit à

$$\int_{-1}^{+1} \frac{\varphi(x)}{\sqrt{1-x^2}}\,dx.$$

quand on fait $k = \frac{2}{n}$, et que l'on prend pour $x_1, x_2, \ldots, x_n$ les racines de l'équation

$$Ee^{\frac{n}{2}\left(\log\frac{(z+1)^{z+1}}{(z-1)^{z-1}} - 2\right)} = 0 \quad \text{ou} \quad E\frac{(z+1)^{\frac{n(z+1)}{2}}}{(z-1)^{\frac{n(z-1)}{2}}} = 0,$$

équation qu'on peut mettre, par le développement en séries, sous la forme suivante:

$$(5) \qquad Ez^n e^{-\frac{n}{2.3\,z^2} - \frac{n}{4.5\,z^4} - \frac{n}{6.7\,z^6} - \ldots} = 0.$$

6. En donnant à n les valeurs les plus simples, telles que

$$n = 2,\ 3,\ 4,\ 5,\ 6,\ 7,$$

on trouve que, pour ces valeurs de n, l'équation (5) devient respectivement

$$z^2 - \frac{1}{3} = 0,$$

$$z^3 - \frac{1}{2}z = 0,$$

$$z^4 - \frac{2}{3}z^2 + \frac{1}{45} = 0,$$

$$z^5 - \frac{5}{6}z^3 + \frac{7}{72}z = 0,$$

$$z^6 - z^4 + \frac{1}{5}z^2 - \frac{1}{105} = 0,$$

$$z^7 - \frac{7}{6}z^5 + \frac{119}{360}z^3 - \frac{149}{6480}z = 0,$$

et, en resolvant ces équations, on obtient les systèmes suivants des valeurs de $x_1, x_2, \ldots, x_n$:

$$n = 2.$$

$$x_1 = -0{,}577350,$$

$$x_2 = +0{,}577350;$$

$$n = 3.$$

$$x_1 = -0{,}707107,$$

$$x_2 = 0,$$

$$x_3 = +0{,}707107;$$

$n = 4.$

$$x_1 = -0{,}794654,$$
$$x_2 = -0{,}187592,$$
$$x_3 = +0{,}187592,$$
$$x_4 = +0{,}794654;$$

$n = 5.$

$$x_1 = -0{,}832497,$$
$$x_2 = -0{,}374541,$$
$$x_3 = \quad 0,$$
$$x_4 = +0{,}374541,$$
$$x_5 = +0{,}832497;$$

$n = 6.$

$$x_1 = -0{,}866247,$$
$$x_2 = -0{,}422519,$$
$$x_3 = -0{,}266635,$$
$$x_4 = +0{,}266635,$$
$$x_5 = +0{,}422519,$$
$$x_6 = +0{,}866247;$$

$n = 7.$

$$x_1 = -0{,}883862,$$
$$x_2 = -0{,}529657,$$
$$x_3 = -0{,}323912,$$
$$x_4 = \quad 0,$$
$$x_5 = +0{,}323912,$$
$$x_6 = +0{,}529657,$$
$$x_7 = +0{,}883862.$$

Avec ces valeurs de $x_1, x_2, \ldots, x_n$, la formule

$$\frac{2}{n}\left[\varphi(x_1) + \varphi(x_2) + \ldots + \varphi(x_n)\right]$$

donne l'expression approximative de l'intégrale

$$\int_{-1}^{+1} \varphi(x)\, dx,$$

qui, dans certains cas, est plus commode pour les applications que ne l'est celle de Gauss; car, dans cette dernière formule, les valeurs $\varphi(x_1), \varphi(x_2), \ldots, \varphi(x_n)$ entrent munies de divers coefficients. Comme notre expression de l'intégrale $\int_{-1}^{+1} \varphi(x)\, dx$ n'est exacte que jusqu'aux termes en $\varphi^{(n+1)}(0)$, $\varphi^{(n+2)}(0), \ldots$, on devra y prendre, en général, plus de termes que dans la formule de Gauss. Néanmoins, dans le cas où les valeurs de $\varphi(x_1)$, $\varphi(x_2), \ldots, \varphi(x_n)$, d'aprés lesquelles on détermine l'intégrale $\int_{-1}^{+1} \varphi(x)\, dx$, sont affectées d'erreurs inconnues, notablement plus grandes que celle qui résulte des termes rejetés, la formule approchée que nous venons de trouver doit être préférée à celle de Gauss même par rapport au degré de précision, vu que, dans cette formule approchée, la somme des carrés de coefficients, à cause de leur égalité, a la plus petite valeur possible.

7. Revenant au cas résolu par M. Hermite, nous remarquons que l'intégrale

$$\int_{-1}^{+1} \frac{\varphi(x)}{\sqrt{1-x^2}}\, dx$$

pour $x = \cos\theta$ se réduit à

$$\int_{0}^{\pi} \varphi(\cos\theta)\, d\theta;$$

donc la formule donnée par lui peut avoir des applications trés-utiles dans la recherche des valeurs approchées du premier terme du développement de $\varphi(\cos\theta)$ en série

$$A_0 + A_1 \cos\theta + A_2 \cos 2\theta + \ldots$$

Pour trouver une expression pareille du coefficient A_1, on devrait faire, dans les formules du n^0 3,

$$F(x) = \frac{x}{\sqrt{1-x^2}}.$$

Or, pour cette valeur de $F(x)$, l'intégrale $\int_{-1}^{+1} F(x)\, dx$, qui entre dans les formules du n^0 3, se réduit à zéro, ce qui fait voir clairement que, pour le cas en question, ces formules ne sont pas applicables. Nous allons montrer le parti qu'on peut cependant tirer, dans ce cas, de la méthode exposée plus haut.

En remplaçant, dans l'intégrale

$$\int_{-1}^{+1} F(x)\,\varphi(x)\,dx,$$

la fonction $\varphi(x)$ par son développement en série

$$\varphi(0) + \frac{\varphi'(0)}{1}x + \frac{\varphi''(0)}{1.2}x^2 + \frac{\varphi'''(0)}{1.2.3}x^3 + \ldots,$$

le terme du résultat qui contient $\varphi(0)$ s'annule toutes les fois que l'intégrale $\int_{-1}^{+1} F(x)\,dx$ est égale à zéro; mais il n'en est plus ainsi, évidemment, de son expression sous la forme

$$k\left[\varphi(x_1) + \varphi(x_2) + \ldots + \varphi(x_n)\right]$$

à moins qu'on ne prenne, dans cette formule, la moitié des termes avec le signe —. Nous allons donc chercher à exprimer la valeur approchée de l'intégrale

$$\int_{-1}^{+1} F(x) = 0,$$

dans la supposition de

$$\int_{-1}^{+1} F(x)\,\varphi(x)\,dx$$

par la formule

$$k\left[\varphi(x_1) + \varphi(x_2) + \ldots + \varphi(x_m) - \varphi(x_{m+1}) - \varphi(x_{m+2}) - \ldots - \varphi(x_{2m})\right],$$

où il y a m termes avec le signe + et m termes avec le signe —.

8. Comme, dans la formule

$$k\left[\varphi(x_1) + \varphi(x_2) + \ldots + \varphi(x_m) - \varphi(x_{m+1}) - \varphi(x_{m+2}) - \ldots - \varphi(x_{2m})\right],$$

il y a $2m+1$ valeurs, savoir: $k, x_1, x_2, \ldots, x_m, x_{m+1}, x_{m+2}, \ldots, x_{2m}$ dont on peut disposer, et que par sa composition, le terme en $\varphi(0)$ s'annule, on peut identifier cette formule avec l'intégrale $\int_{-1}^{+1} F(x)\,\varphi(x)\,dx$ jusqu'aux termes qui contiennent les $2m+1$ premières dérivées de $\varphi(x)$, ce qui nous donne l'équation

$$\int_{-1}^{+1} F(x)\,\varphi(x)dx = k\left[\varphi(x_1) + \varphi(x_2) \ldots + \varphi(x_m) - \varphi(x_{m+1}) - \varphi(x_{m+2}) \ldots - \varphi(x_{2m})\right]$$
$$+ k_1\varphi^{2m+2}(0) + k_2\varphi^{2m+3}(0) + \ldots$$

le cas en question, ces formules ne sont pas applicables. Nous allons montrer le parti qu'on peut cependant tirer, dans ce cas, de la méthode exposée plus haut.

En remplaçant, dans l'intégrale

$$\int_{-1}^{+1} F(x)\,\varphi(x)\,dx,$$

la fonction $\varphi(x)$ par son développement en série

$$\varphi(0) + \frac{\varphi'(0)}{1}x + \frac{\varphi''(0)}{1.2}x^2 + \frac{\varphi'''(0)}{1.2.3}x^3 + \ldots,$$

le terme du résultat qui contient $\varphi(0)$ s'annule toutes les fois que l'intégrale $\int_{-1}^{+1} F(x)\,dx$ est égale à zéro; mais il n'en est plus ainsi, évidemment, de son expression sous la forme

$$k\left[\varphi(x_1) + \varphi(x_2) + \ldots + \varphi(x_n)\right]$$

à moins qu'on ne prenne, dans cette formule, la moitié des termes avec le signe —. Nous allons donc chercher à exprimer la valeur approchée de l'intégrale

$$\int_{-1}^{+1} F(x) = 0,$$

dans la supposition de

$$\int_{-1}^{+1} F(x)\,\varphi(x)\,dx$$

par la formule

$$k\left[\varphi(x_1) + \varphi(x_2) + \ldots + \varphi(x_m) - \varphi(x_{m+1}) - \varphi(x_{m+2}) - \ldots - \varphi(x_{2m})\right],$$

où il y a m termes avec le signe + et m termes avec le signe —.

8. Comme, dans la formule

$$k\left[\varphi(x_1) + \varphi(x_2) + \ldots + \varphi(x_m) - \varphi(x_{m+1}) - \varphi(x_{m+2}) - \ldots - \varphi(x_{2m})\right],$$

il y a $2m+1$ valeurs, savoir: $k, x_1, x_2, \ldots, x_m, x_{m+1}, x_{m+2}, \ldots, x_{2m}$ dont on peut disposer, et que par sa composition, le terme en $\varphi(0)$ s'annule, on peut identifier cette formule avec l'intégrale $\int_{-1}^{+1} F(x)\,\varphi(x)\,dx$ jusqu'aux termes qui contiennent les $2m+1$ premières dérivées de $\varphi(x)$, ce qui nous donne l'équation

$$\int_{-1}^{+1} F(x)\,\varphi(x)dx = k\left[\varphi(x_1) + \varphi(x_2) \ldots + \varphi(x_m) - \varphi(x_{m+1}) - \varphi(x_{m+2}) \ldots - \varphi(x_{2m})\right]$$
$$+ k_1\varphi^{2m+2}(0) + k_2\varphi^{2m+3}(0) + \ldots$$

on trouve

$$\frac{f_0(z)}{f_1(z)}\, e^{-L_1 z^{-2m-2} - L_2 z^{-2m-3} - \dots} = e^{\frac{1}{k}\int_{-1}^{+1} F(x) \log(z-x)\, dx}.$$

Les fonctions $f_0(z)$, $f_1(z)$ étant de même degré, la fraction $\frac{f_0(z)}{f_1(z)}$ est du degré zéro; de plus, l'expression

$$e^{-L_1 z^{-2m-2} - L_2 z^{-2m-3} - \dots}$$

ne diffère de 1 que par les puissances de z inférieures à z^{-2m-1}; par conséquent, l'équation trouvée nous montre que la fraction $\frac{f_0(z)}{f_1(z)}$ ne diffère de l'expression

$$e^{\frac{1}{k}\int_{+1}^{+1} F(x) \log(z-x)\, dx}$$

que par les termes qui renferment les puissances de z moins élevées que z^{-2m-1} et, par suite, moins élevées que le degré de la fraction $\frac{1}{z\,[f_1(z)]^2}$; car la fonction $f_1(z)$, comme nous l'avons vu, n'est que du degré m. Mais, on le sait, la fraction $\frac{f_0(z)}{f_1(z)}$ ne peut donner une valeur approchée d'une fonction quelconque exacte jusqu'à l'ordre de $\frac{1}{z\,[f_1(z)]^2}$, à moins qu'elle ne soit l'une des fractions convergentes qu'on trouve par le développement en fraction continue, et que le quotient complet, correspondant à cette fraction convergente, ne soit dépourvu du terme en $\frac{1}{z}$. En partant de là, il est aisé de trouver et la constante k et les fonctions $f_0(z)$, $f_1(z)$, qui déterminent les valeurs de $x_1, x_2, \dots, x_m, x_{m+1}, x_{m+2}, \dots, x_{2m}$. A cet effet, on développera l'expression

$$e^{\frac{1}{k}\int_{-1}^{+1} F(x) \log(z-x)\, dx}$$

en fraction continue, en s'arrêtant au quotient qui correspond à une fraction convergente dont les termes sont du degré m. Égalant à zéro le coefficient de $\frac{1}{z}$ dans l'expression complète de ce quotient, on aura l'équation qui déterminera la valeur de la constante k, et, en mettant la valeur de k ainsi déterminée dans deux termes de la fraction convergente, on aura les fonctions cherchées $f_0(z)$, $f_1(z)$.

9. Pour montrer, sur un exemple, l'usage de ce que nous venons d'exposer, supposons qu'il s'agisse de trouver l'expression approximative de l'intégrale

$$\int_{-1}^{+1} x\varphi(x)\, dx.$$

Pour cela on posera, dans les formules du numéro précédent,

$$F(x) = x.$$

Pour cétte valeur de $F(x)$, on obtient

$$\int_{-1}^{+1} F(x) \log (z - x)\, dx = \int_{-1}^{+1} x \log (z - x)\, dx = \frac{z^2 - 1}{2} \log \frac{z + 1}{z - 1} - z,$$

$$e^{\frac{1}{k} \int_{-1}^{+1} F(x) \log (z - x)\, dx} = e^{\frac{z^2 - 1}{2k} \log \frac{z + 1}{z - 1} - \frac{z}{k}}.$$

En développant la dernière expression en fraction continue, on trouve que le quotient complet, correspondant à une fraction convergente dont les termes sont du premier degré, est égal à

$$3kz + 1 - \left(\frac{3}{5} k - \frac{1}{9k}\right) \frac{1}{z} + \dots,$$

et que cette fraction est égale à

$$\frac{3kz - 1}{3kz + 1},$$

d'où nous concluons que, dans l'expression approximative de l'intégrale

$$\int_{-1}^{+1} x \varphi(x)\, dx$$

par la formule

$$k \left[\varphi(x_1) - \varphi(x_2) \right]$$

on doit prendre, pour k, une racine de l'équation

$$\frac{3}{5} k - \frac{1}{9k} = 0,$$

et, pour x_1, x_2, respectivement, les racines des équations

$$3kz - 1 = 0, \quad 3kz + 1 = 0.$$

On trouve ainsi deux valeurs de k:

$$k = +\sqrt{\frac{5}{27}}, \quad k = -\sqrt{\frac{5}{27}}$$

et deux systèmes des valeurs de x_1, x_2:

$$x_1 = +\sqrt{\frac{3}{5}}, \quad x_2 = -\sqrt{\frac{3}{5}},$$

$$x_1 = -\sqrt{\frac{3}{5}}, \quad x_2 = +\sqrt{\frac{3}{5}};$$

mais de ces doubles valeurs de k, x_1, x_2, il ne résulte évidemment qu'une seule valeur de l'expression cherchée, savoir:

$$\sqrt{\frac{5}{27}}\left[\varphi\left(\sqrt{\frac{3}{5}}\right)-\varphi\left(-\sqrt{\frac{3}{5}}\right)\right].$$

Pour trouver une expression approximative à quatre termes de l'intégrale

$$\int_{-1}^{+1} x\varphi(x)\,dx$$

on prendra le quotient de la même fraction continue qui correspond à la fraction convergente dont les termes sont du second degré. Comme la valeur complète de ce quotient s'exprime par la série

$$105\,k^2 z - \frac{5832\,k^4 - 945\,k^2 + 35}{243\,k^2 - 45}\,\frac{1}{z} + \dots$$

et qu'il correspond à la fraction convergente

$$\frac{270\,k^2 z^2 - 90\,kz + 10 - 54\,k^2}{270\,k^2 z^2 + 90\,kz + 10 - 54\,k^2},$$

on trouvera les valeurs de la constante k et de x_1, x_2, x_3, x_4 par les équations

$$5832\,k^4 - 945\,k^2 + 35 = 0,$$

$$270\,k^2 z^2 - 90\,kz + 10 - 54\,k^2 = 0,$$

$$270\,k^2 z^2 + 90\,kz + 10 - 54\,k^2 = 0.$$

En les résolvant, on parvient à ces deux expressions approximatives de l'intégrale en question

$$0{,}23937\left[\varphi(0{,}89224)+\varphi(0{,}50030)-\varphi(-0{,}50030)-\varphi(-0{,}89224)\right],$$

et

$$0{,}32363\left[\varphi(0{,}84905)+\varphi(0{,}18093)-\varphi(-0{,}18093)-\varphi(-0{,}84905)\right].$$

10. En passant à la recherche des expressions approximatives de l'intégrale

$$\int_0^{\pi} \cos\theta\varphi(\cos\theta)\,d\theta \quad \text{ou} \quad \int_{-1}^{+1} \frac{x}{\sqrt{1-x^2}}\,\varphi(x)\,dx,$$

nous poserons dans nos formules

$$F(x) = \frac{x}{\sqrt{1-x^2}}$$

Pour cette valeur de $F(x)$ on obtient

$$\int_{-1}^{+1} F(x)\log(z-x)\,dx = \int_{-1}^{+1} \frac{x}{\sqrt{1-x^2}}\log(z-x)\,dx = -\pi\left(z-\sqrt{z^2-1}\right),$$

$$e^{\frac{1}{k}\int_{-1}^{+1} F(x)\log(z-x)\,dx} = e^{-\frac{\pi}{k}\left(z-\sqrt{z^2-1}\right)}.$$

Développant la dernière expression en fraction continue, on trouve les fractions convergentes

$$\frac{4kz-\pi}{4kz+\pi},\quad \frac{48k^2z^2-12k\pi z+\pi^2-12k^2}{48k^2z^2+12k\pi z+\pi^2-12k^2},\ldots,$$

qui correspondent aux quotients complets

$$4kz+\pi-\frac{12k^2-\pi^2}{12k}\,\frac{1}{z}+\ldots,\quad 12kz-\frac{720k^4-60\pi^2k^2+\pi^4}{20\,(12k^2-\pi^2)\,k}\,\frac{1}{z}+\ldots,$$

d'où, d'après le n° 8: 1°, pour la détermination de k, x_1, x_2 dans l'expression approximative de l'intégrale

$$\int_0^\pi \cos\theta\,\varphi(\cos\theta)\,d\theta$$

par la formule

$$k\left[\varphi(x_1)-\varphi(x_2)\right]$$

résultent ces équations

$$12k^2-\pi^2=0,\quad 4kz-\pi=0,\quad 4kz+\pi=0;$$

et 2°, pour la détermination d'une expression semblable à quatre termes, les équations

$$720k^4-60\pi^2k^2+\pi^4=0,$$

$$48k^2z^2-12k\pi z+\pi^2-12k^2=0,$$

$$48k^2z^2+12k\pi z+\pi^2-12k^2=0.$$

Les expressions approximatives de l'intégrale

$$\int_0^\pi \cos\theta\,\varphi(\cos\theta)\,d\theta,$$

que l'on obtient d'après ces équations, se réduisent à celles qui suivent:

$$\int_0^\pi \cos\theta\,\varphi(\cos\theta)\,d\theta = \frac{\pi}{\sqrt{12}}\left\{\varphi(\cos\theta_0) - \varphi\left[\cos(\pi + \theta_0)\right]\right\},$$

$$\int_0^\pi \cos\theta\,\varphi(\cos\theta)\,d\theta =$$

$$0{,}151765\,\pi\left\{\varphi(\cos\theta_1) + \varphi(\cos\theta_2) - \varphi\left[\cos(\pi + \theta_1)\right] - \varphi\left[\cos(\pi + \theta_2)\right]\right\},$$

où

$$\theta_0 = 30^\circ,\quad \theta_1 = 12^\circ,\quad \theta_2 = 48^\circ\ ^{*)}.$$

On trouverait des expressions plus approchées de l'intégrale

$$\int_0^\pi \cos\theta\,\varphi(\cos\theta)\,d\theta,$$

en prenant plus de termes dans la formule

$$k\left[\varphi(x_1) + \varphi(x_2) + \ldots + \varphi(x_m) - \varphi(x_{m+1}) - \varphi(x_{m+2}) - \ldots - \varphi(x_{2m})\right].$$

Nous n'insisterons pas sur la recherche de ces formules; nous remarquerons seulement que, en remplaçant dans toutes ces formules les termes de la forme $\varphi(\cos\theta_\lambda)$ par

$$\frac{1}{l}\left[\varphi\left(\cos\frac{\theta_\lambda}{l}\right) + \varphi\left(\cos\frac{2\pi + \theta_\lambda}{l}\right) + \ldots + \varphi\left(\cos\frac{2(l-1)\pi + \theta_\lambda}{l}\right)\right],$$

où l est un nombre entier, on obtient les expressions approximatives de l'intégrale

$$\int_0^\pi \cos l\theta\,\varphi(\cos\theta)\,d\theta.$$

Ainsi, en partant des formules précédentes, qui donnent les valeurs approchées, à deux et quatre termes, de l'intégrale

$$\int_0^\pi \cos\theta\,\varphi(\cos\theta)\,d\theta,$$

on passerait aux expressions approximatives à $2l$ et $4l$ termes de l'intégrale

$$\int_0^\pi \cos l\theta\,\varphi(\cos\theta)\,d\theta,$$

---

*) Hormis cette expression, il existe une autre aussi aux quatres termes, où

$$k = 0{,}245562\,\pi,\qquad \theta_1 = 24^0,\qquad \theta_2 = 84^0.$$

A. L.

expressions qui peuvent être présentées ainsi:

$$\frac{\pi}{\sqrt{12}\,l}\sum_{\mu=0}^{\mu=2l}(-1)^{\mu}\,\varphi\left(\cos\frac{\mu\pi+\theta_0}{l}\right),$$

$$\frac{0{,}151765\,\pi}{l}\sum_{\mu=0}^{\mu=2l}(-1)^{\mu}\left[\varphi\left(\cos\frac{\mu\pi+\theta_1}{l}\right)+\varphi\left(\cos\frac{\mu\pi+\theta_2}{l}\right)\right],$$

où les signes de sommations s'étendent à $\mu=0,\ 1,\ 2,\ldots,\ 2l-1$.

9 '
+,- * *. ! * 9
5/ M" M E E MM32 9B9(%(%)$% 2"
' ((8 E R

Exercice.

27. Théorème. — Cercle des neuf points. *Dans un triangle, les milieux des côtés, les pieds des hauteurs et les milieux des droites qui joignent les sommets au point de concours des hauteurs, sont situés sur une même circonférence.*

EULER*, *Mémoires de Saint-Pétersbourg*, 1765.

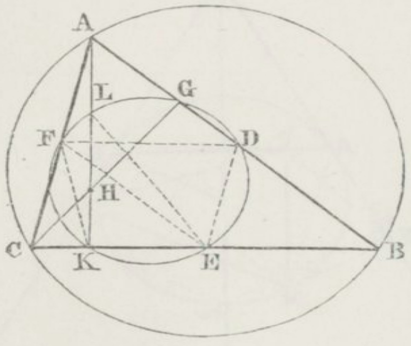

Fig. 11.

Soient D, E, F les points milieux des côtés; AK, CG deux des hauteurs, et L le point milieu de AH.

Circonscrivons une circonférence au triangle DEF.

Pour démontrer le théorème, il suffit de prouver que cette circonférence passe par le pied K, d'une hauteur quelconque, et par le point L, milieu de AH.

1° La droite FK, qui joint le sommet K de l'angle droit au point milieu F de l'hypoténuse, égale la moitié de cette hypoténuse (G., n° 222); donc FK = FC = donc DE.

Ainsi le trapèze EDFK est isocèle; par suite, la circonférence qui passe par E, D, F, passe aussi par le quatrième sommet K.

2° La droite FL, qui joint les points milieux F, L des côtés du triangle ACH, est parallèle à la base CH; d'ailleurs FE est aussi parallèle à AB; donc l'angle EFL égale l'angle AGC, égale donc 90 degrés.

Le quadrilatère EKFL est inscriptible à cause des angles droits EKL, EFL; donc la circonférence qui passe par les trois sommets E, K, F, passe aussi par le quatrième sommet L. *C. Q. F. D.*

28. Scolies. I. *Le centre du cercle des neuf points est au milieu de la droite qui joint le point de concours des hauteurs au centre du cercle circonscrit à ce triangle.*

En effet, le centre se trouve sur les perpendiculaires élevées au milieu de FG et de KE; or ces perpendiculaires passent par le point M, milieu de OH (fig. 12).

---

* EULER, né à Bâle en 1707, mort en 1783, célèbre analyste; perfectionna le *calcul intégral* et fit connaître les cinq surfaces du second degré. (G., n° 857.)

II. — *Le rayon du cercle des neuf points est la moitié du rayon du cercle circonscrit.*

Car $EM = \frac{1}{2} EL = \frac{1}{2} AO$

Cela résulte aussi des triangles semblables EOM, AHO.

III. *La tangente* EJ, *au point milieu d'un côté, et ce même côté sont anti-parallèles par rapport à l'angle opposé.*

Les tangentes EJ, AT sont parallèles, car elles sont perpendiculaires aux rayons parallèles EM, AO ; de plus, l'angle CAT = CBA.

Donc AT et CB ou EJ et CB sont anti-parallèles par rapport à l'angle A.

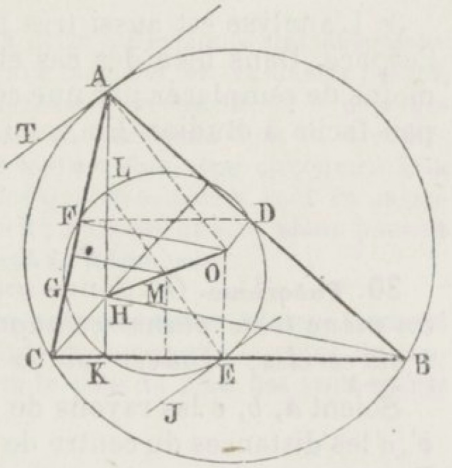

Fig. 12.

Note. Deux droites AT et CB (fig. 12) sont *anti-parallèles* par rapport à un angle CAB, lorsque l'angle CAT, que la première forme avec l'un des côtés de l'angle donné, égale l'angle que forme la seconde droite avec l'autre côté. Ainsi l'angle CAT égale ABC.

La dénomination de lignes anti-parallèles est due à A. ARNAULD *, vers 1667. (*Nouvelles Annales mathématiques* **, 1850, page 409.)

29. Remarque. 1° Le théorème du *cercle des neuf points* peut être démontré de plusieurs manières différentes, mais il n'en est pas de plus élémentaire et de plus rapide que la précédente (n° 27).

2° C'est par l'emploi judicieux de l'analyse que l'on découvre les relations les plus simples qui rattachent entre elles les diverses parties d'une même question et que l'on trouve, par suite, le meilleur mode de démonstration.

---

* ANTOINE ARNAULD, né à Paris en 1612, mort à Bruxelles en 1694, est beaucoup plus connu par son séjour à Port-Royal et par ses œuvres de polémique que par ses *Nouveaux Éléments de géométrie*, publiés en 1667, mais sans nom d'auteur.

** Le journal connu sous le nom de *Nouvelles Annales mathématiques* a été fondé, en 1842, par MM. TERQUEM et GERONO.

M. Terquem, mort en 1862, a été successivement remplacé par MM. E. PROUHET, J. BOURGET, CH. BRISSE, tandis que l'honorable et savant M. GERONO continue à rédiger le journal qu'il a fondé il y a quarante ans.

Nous aurons à citer fréquemment les *Nouvelles Annales*, car cet ouvrage nous a fourni de nombreuses et intéressantes questions et d'utiles renseignements bibliographiques. Les renvois seront indiqués par (N. A., année ..., page ...).

TERQUEM, né à Metz en 1782, mort à Paris en 1862; fut admis à l'École polytechnique en 1801 ; il occupa la chaire de mathématiques transcendantes, au lycée de Mayence, de 1804 à 1814. A partir de cette époque, il fut bibliothécaire au dépôt d'artillerie à Paris, publia divers ouvrages, et collabora assidûment au journal de M. GERONO.

3° L'analyse est aussi très utile lorsqu'il s'agit de la géométrie dans l'espace. Dans bien des cas elle permet de se passer de figure, ou du moins de remplacer par une construction simple une figure compliquée peu facile à étudier. En voici quelques exemples.

Exercice.

30. Théorème. *On donne une sphère et un point fixe* P; *par ce point on mène trois plans rectangulaires deux à deux et qui déterminent trois cercles; prouver que la somme de ces trois cercles est constante.*

Soient a, b, c les rayons de ces cercles, r le rayon de la sphère et a', b', c' les distances du centre de la sphère aux cercles; il faut prouver que l'on a :

$$\pi a^2 + \pi b^2 + \pi c^2 = \text{constante}$$

ou, ce qui revient au même, $a^2 + b^2 + c^2 =$ une valeur constante

mais $a^2 = r^2 - a'^2;\quad b^2 = r^2 - b'^2;\quad c^2 = r^2 - c'^2$

on a donc $a^2 + b^2 + c^2 = 3r^2 - (a'^2 + b'^2 + c'^2)$

Il suffit de prouver que la quantité à soustraire est constante.

Or les trois distances a', b', c', perpendiculaires deux à deux, menées du centre O sur les trois plans rectangulaires, dont P est le point commun, sont les trois arêtes latérales d'un parallélépipède rectangle ayant PO pour diagonale; par suite, la somme des carrés de ces arêtes égale PO^2 (G., n° 439), et le théorème est démontré.

31. Remarque. La détermination de la valeur constante n'offre aucune difficulté.

Ainsi $a^2 + b^2 + c^2 = 3r^2 - PO^2$

donc $\pi a^2 + \pi b^2 + \pi c^2 = 3\pi r^2 - \pi PO^2$

La somme des trois cercles déterminés par le trièdre tri-rectangle dont P est le sommet, égale trois grands cercles moins le cercle qui aurait PO pour rayon.

Exercice.

32. Théorème. *Lorsque les arêtes opposées d'un octaèdre inscrit dans une sphère sont dans un même plan, les trois diagonales de l'octaèdre se coupent au même point. En menant un plan tangent à la sphère par chaque sommet de l'octaèdre, on forme un hexaèdre circonscrit dont les faces, prises quatre à quatre, concourent en un même point.* (G., n° 429.)

1° Les arêtes opposées étant dans un même plan, les deux diagonales qui joignent les extrémités des arêtes opposées se coupent, car elles sont deux à deux dans un même plan. Les trois diagonales de l'octaèdre ne peuvent être

')
! *+,-* * *. ! * = :!7 " ! & ! 8
> ! " 8 :!7 " A A! A B))9 2"

3 *+,-* * *. ! * > ! S & !
1 ! - ! 0 , 7 ! +3 < 1 " > 7 : 2!B 2 ! $ 2"

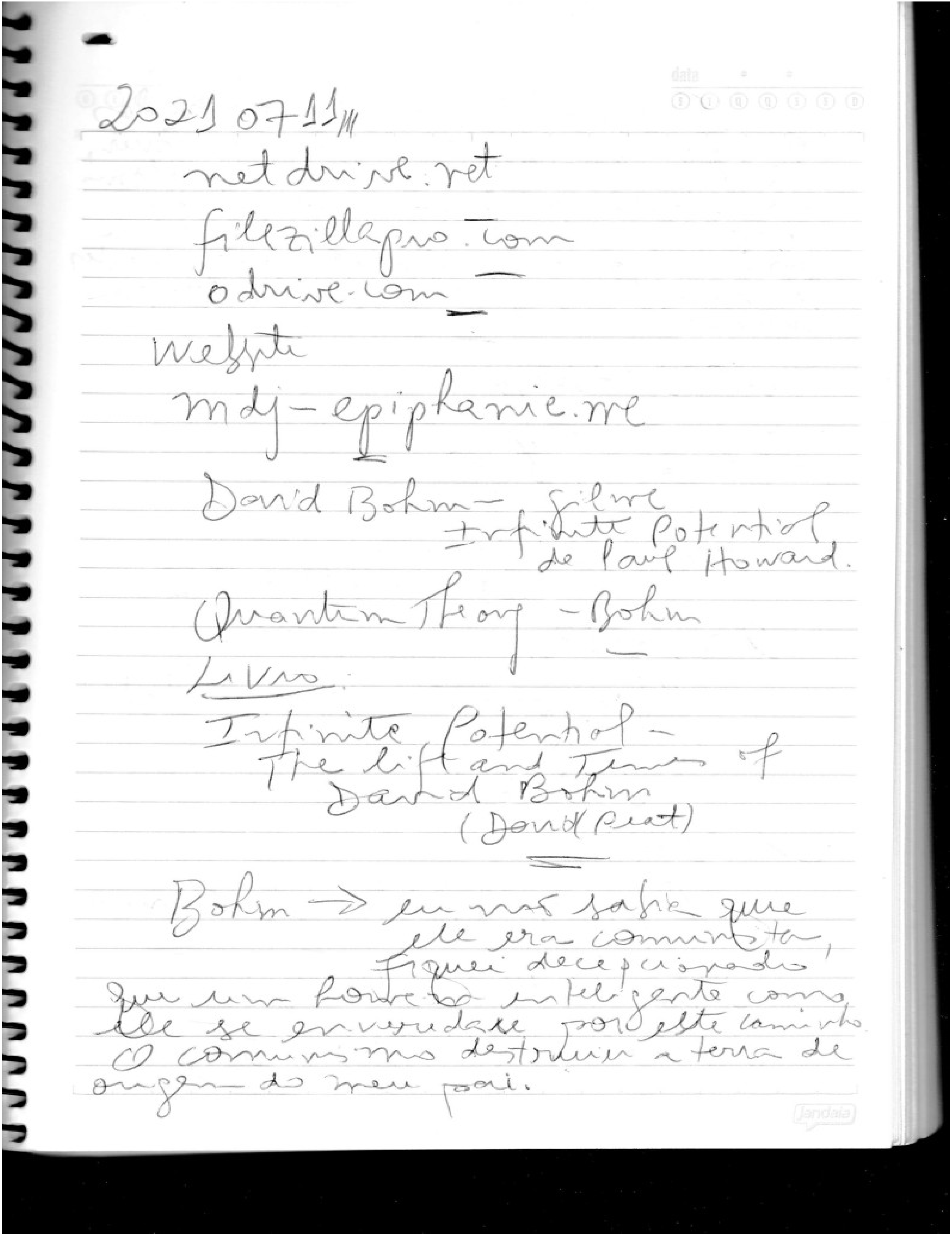

2021 07 11

netdrive.net

filezillapro.com

odrive.com

Website

mdj-epiphanie.me

David Bohm — filme
Infinite Potential
de Paul Howard.

Quantum Theory - Bohm

Livro

Infinite Potential -
The life and Times of
David Bohm
(David Peat)

Bohm → eu não sabia que
ele era comunista,
fiquei decepcionado
que um homem tão inteligente como
ele se enveredasse por este caminho
O comunismo destruiu a terra de
origem do meu pai.

O capitalismo é a melhor das formas imperfeitas para se viver, a liberdade gerada pelo capitalismo compensa as imperfeições do sistema. Não estamos no céu e portanto, as imperfeições são presentes e transitórias, se acaba uma surgem outras.

20210712///.

OAB 80502//
Tel 2503-2288

Ilha de Itacuruçá — Mangaratiba RJ.

20210713///

"tangping" ⇒ "ficar deitado"

"9-9-6"

20210714///

Pappus-Guldinus Theorem

$3456 \div 11 = 3{,}14\overline{18}$//

$3456789 \div 11 = 3{,}142535\overline{45}$//

$3456789/7 = 4{,}93270$

$3456/7 = 4{,}93\overline{714285}$

$345/6 = 5{,}75$

Diego Rubiera-Garcia

drgarcia@fc.ul.pt

Bob.Palais@uvu.edu

palais@math.utah.edu

musich.art

cut-the-knot.org

20210715

Gambarogno Swiss

↳ Lago Maggiore

Calculus in 3D Nitecki

STENIOMUSICH.COM

& ' (.2 > A!

Books:

Luigi Bianchi

Opere vol. 3 - Sistemi tripli ortogonali

2 - Applicabilità e problemi di deformazione

5 - Trasformazioni delle superficie e delle curve

7 - Problemi di rotolamento

8 - Classi speciali di superficie

6 - Congruenze di rette e di sfere e loro deformazioni

4.1 => Deformazioni delle quadriche

---

bn1lib.org → library.lol

z-lib.org

smo01-internet@yahoo.com

Tes@33

user: smo01

@singlelogin.me/

:! ! 4 7 7 $ #A B $ # #> A !
! # L " ! Q 7 NNN
! *+,-* * *. ! * D " ! 2"

3 *+,-* * *. ! * D " ! 2"

+,- * *. ! * D " $ B 2"

" *+,-* * *. ! * > A ! > D ! !
" D " 7! !3 2 ! " D 5 2) 9 2"

' %

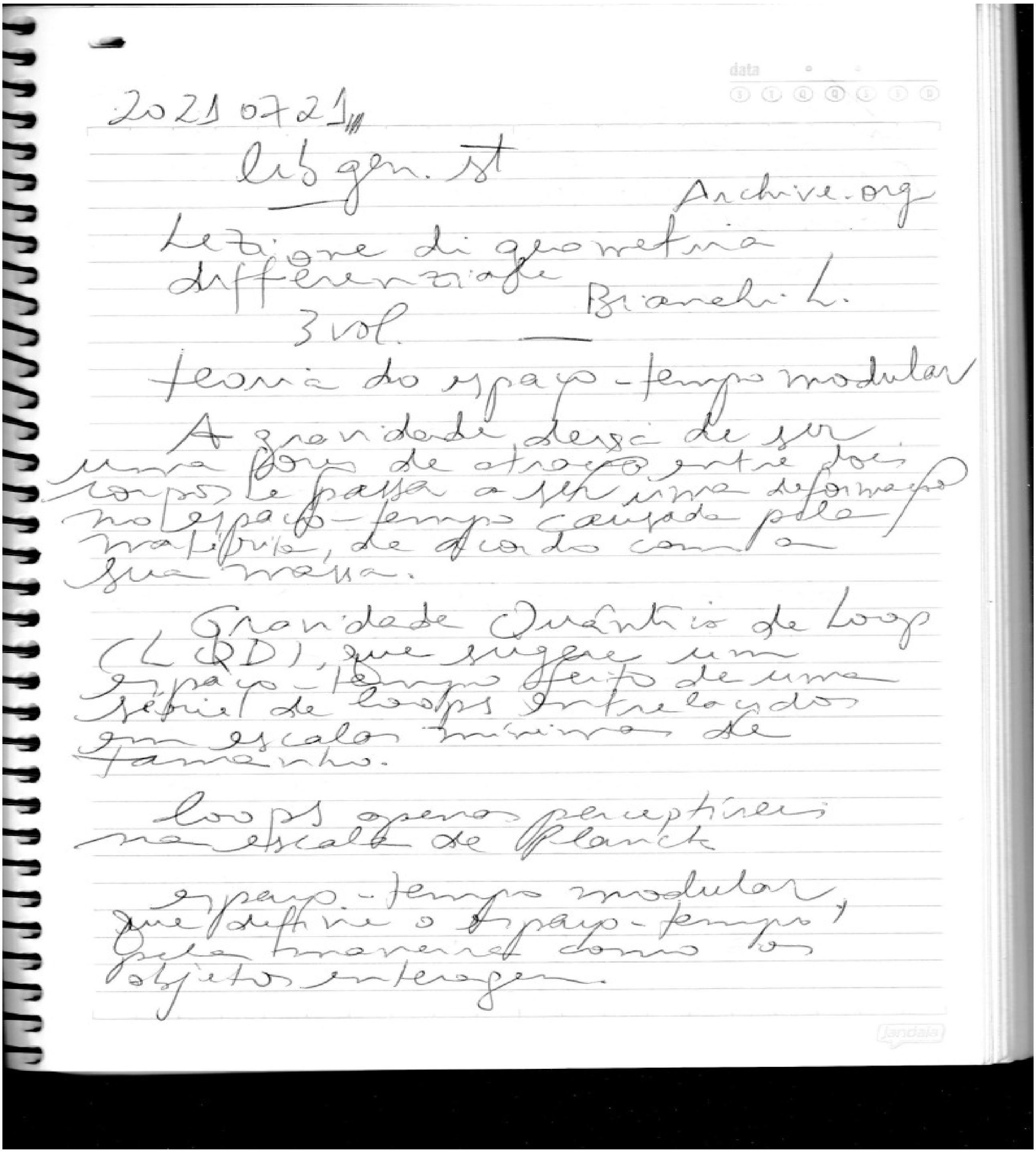

2021 07 21

lib gen. st

Archive.org

Lezione di geometria differenziale

Bianchi L.

3 vol.

teoria do espaço-tempo modular

A gravidade deixa de ser uma força de atração entre dois corpos e passa a ser uma deformação no espaço-tempo causada pela matéria, de acordo com a sua massa.

Gravidade Quântica de Loop (LQD), que sugere um espaço-tempo feito de uma série de loops entrelaçados em escalas mínimas de tamanho.

loops apenas perceptíveis na escala de Planck

espaço-tempo modular, que define o espaço-tempo pela maneira como os objetos interagem.

Assim, o espaço-tempo se torna uma consequência do próprio mundo quântico, não palco que precisa ser conciliado com ele.

Source (Freidel, Leigh e Minic)
space.com

2021 07 22 ///

wolframcloud.com

Recognize Handwritten Digits

digit = Classify[{1 → 1, 2 → 2, ...}]

digit[8, "TopProbabilities"]

↳ o resultado mostra a frequência na qual o fenômeno se apresenta.

---

wolfram.com/open-materials/

→ Sandbox

Cryptografia assimétricos → RSA.

stenio.munich@studio.unibo.it

20210723

"misshoki" or "hinomaru" → bandeira nac. japão.

"kyokujitsuki" → "rising sun"

20210725

Jennifer Connelly

Construindo uma carreira

Livros:

pdfdrive.com

Books:

Books-015-2021/_Important/

Geometry Illuminated An Illustrated Introduction to Euclidean ... pdf

Visual Complex Analysis.pdf

Pág. 216 (242)

The Cartesian Form

Pág. 222 (248)

Visual Differentiation of log(z)

Pág. ~~25~~ 224 (250)

Inverse Functions

→ twist (torção) of f

MHHMY71.pdf

Pág. 30 (56)

38 - The Circle

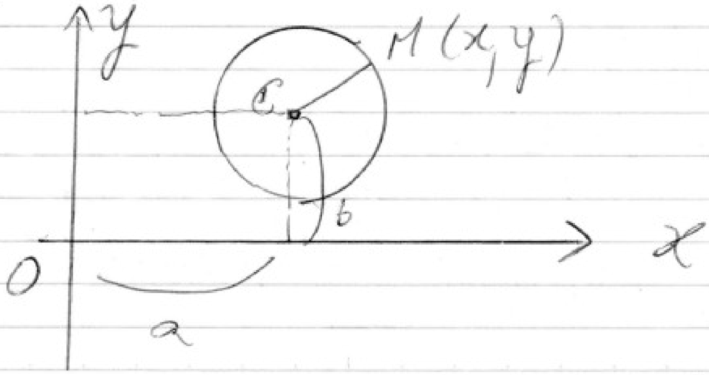

)
! *+,-* * *. ! *) 6 0 & ! " + & ! "
& " 5 " ! ! " <02 3 ! 6 0 2"
3 *+,-* * *. ! *) 6 0 & ! "M + & ! "
& " 5 " ! ! " <02 3 ! 6 0 2"

'

! *+,-* * *. ! * 8 ! 4 " ! ! 2 /
+ ! 0 ! " T ./ "))) 2"

3 *+,-* * *. ! * 8 ! 4 " ! ! 2 /
+ ! 0 .;)) "?7

' 9 (8 ! ! =

5

Further Geometry of Differentiation

I Cauchy-Riemann Revealed

1 Introduction

In the previous chapter we began to investigate the remarkable nature of analytic functions in $\mathbb{C}$ by studying mappings in the less structured realm of $\mathbb{R}^2$. In particular the Jacobian provided us with a painless way of deriving the Cauchy-Riemann characterization of analytic functions, and also of computing their amplitwists. However, this approach was rather indirect. In this chapter we will instead study differentiation directly in the complex plane, primarily through the use of infinitesimal geometry. Our first application of this approach will be the rederivation of the Cauchy-Riemann (henceforth " CR) equations, and the discovery of new forms that they can take on.

2 The Cartesian Form

Consider a very fine mesh of squares aligned with the real and imaginary axes. See the top left of [1]. Under an analytic mapping each infinitesimal square will be amplitwisted to produce an image that is also square. We will show that the CR equations are nothing more than a symbolic restatement of this geometric fact.

Zoom in on an individual square and its image, as depicted in the bottom half

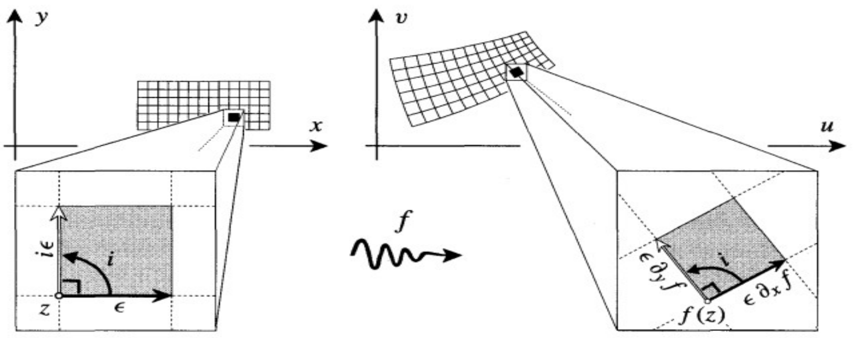

Figure [1]

of [1]. Suppose, as drawn, that the initial square has side ϵ. If we start at z and then move a distance ϵ in the x-direction, the image will move along a complex number given by

$$\text{(change in } x)\cdot\text{(rate of change with x of the image f)} = \epsilon\, \partial_x f.$$

Similarly, if the point moves along the vertical edge by going ϵ in the y-direction, then its image will move along $\epsilon\, \partial_y \mathbf{f}$. Now since these two image vectors span a square they must be related by a simple rotation of $\pi/2$, that is by multiplication with i. After cancelling ϵ, we thus obtain

$$i\,\partial_x f = \partial_y f,$$

et voilà! That this is indeed a compact form of the CR equations may be seen by inserting $\mathbf{f} = u + \mathrm{i}v$:

$$i\,\partial_x(u + \mathrm{i}v) = \partial_y(u + \mathrm{i}v),$$

and then equating real and imaginary parts to yield

$$\partial_x u = \partial_y v \quad \text{and} \quad \partial_x v = -\partial_y u, \tag{1}$$

just as before. To obtain the amplitwist itself, we recall that each infinitesimal arrow is taken to its image by multiplication with $\mathbf{f}'$. Now, since we know what the images are for the two sides of the square, we can deduce

$$\begin{aligned} \epsilon \longmapsto \epsilon\, f' &= \epsilon\, \partial_x f \\ \Rightarrow f' &= \partial_x f \end{aligned}$$

and

$$\begin{aligned} i\epsilon \longmapsto i\epsilon\, f' &= \epsilon\, \partial_y f \\ \Rightarrow f' &= -i\, \partial_y f. \end{aligned}$$

3 The Polar Form

Equation (1) is the most common way of writing CR, but it isn't the only way. It took this form because we chose to describe both complex planes in terms of their real and imaginary parts, that is with Cartesian coordinates. Thus we could briefly describe (1) as being the Cart.-Cart. form. In Ex. 10, p. 212 we retained Cartesian coordinates for the first plane but employed polar coordinates in the image plane; this led to another form (Cart.-Polar) of CR. As the next example of our geometric method we will derive the Polar-Cart. form of the equations.

In order to do this, we begin with an infinitesimal square adapted to polar coordinates. See [2]. If we start at z and increase r by $\mathrm{d}r$, then we obtain $e^{i\theta}\,\mathrm{d}r$

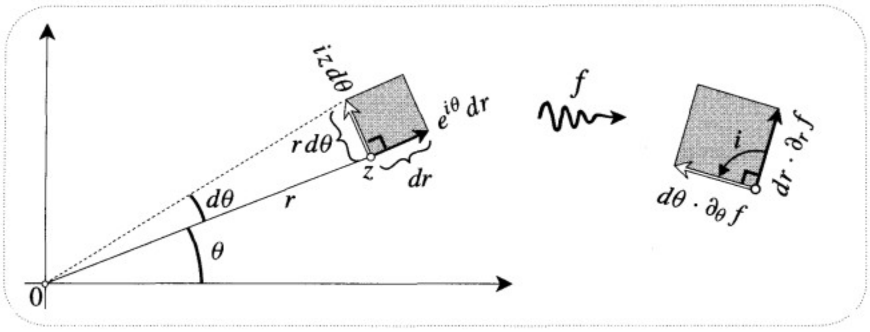

Figure [2]

as the radial edge. If, on the other hand, we increase 0 by $d\theta$, then the point will move in the perpendicular direction given by ie^{iB}. As $d\theta$ tends to zero, this edge is ultimately equal to an infinitesimal arc of circle of length $r\,d\theta$; the complex number describing it will therefore be i $e^{iB}\,r\,d\theta = i\,z\,d\theta$. It's also clear from our picture that

$$\text{initially square} \qquad \mathrm{d\,r} = r\,d\theta. \tag{2}$$

Now look at the image. Just as before, if we increase r by $\mathrm{d\,r}$ then the image will move along $\mathrm{d\,r} \cdot \partial_r \mathrm{f}$; likewise, changing θ by $d\theta$ will move the image along $d\theta \cdot \partial_\theta \mathbf{f}$. If the mapping is analytic then these again span a square, and so the latter must be i times the former:

$$d\theta \cdot \partial_\theta f = i dr \cdot \partial_r f.$$

Substituting (2) into this, and cancelling $d\theta$, we obtain

$$\partial_\theta f = ir\,\partial_r f \tag{3}$$

as the new compact form of CR. By inserting $\mathbf{f} = u + \mathrm{i}v$, the reader may verify that **(3)** is equivalent to the following pair of Polar-Cart. equations:

$$\partial_\theta v = +r\,\partial_r u \tag{4}$$

$$\partial_\theta u = -r\,\partial_r v. \tag{5}$$

By examining the amplitwist that carries each arrow to its image we can also obtain two expressions for the derivative:

$$\begin{aligned} e^{i\theta}\,dr \longmapsto e^{i\theta}\,dr \cdot f' &= dr \cdot \partial_r f \\ \Longrightarrow f' &= e^{-i\theta}\,\partial_r f \end{aligned} \tag{6}$$

and

$$i\,z\,d\theta \longmapsto i\,z\,d\theta \cdot f' \;=\; d\theta \cdot \partial_\theta f$$

$$\Longrightarrow f' \quad = \quad -(i/z)\,\partial_\theta f. \qquad (7)$$

As a simple example let's take $z^3 = r^3\,e^{3i\theta}$. From *(6)* we obtain

$$(z^3)' = e^{-i\theta}\,3r^2\,e^{3i\theta} = 3r^2\,e^{2i\theta} = 3z^2,$$

while from (7) we obtain

$$(z^3)' = -(i/z)\,r^3\,3i\,e^{3i\theta} = -(i/z)\,3i\,z^3 = 3z^2.$$

In obtaining the same answer from both these expressions we have also verified that z^3 actually was analytic in the first place.

Of the four possible ways of writing CR, only one now remains to be found, namely the Polar-Polar form. We leave it to the reader to verify that if we write $f = R\,e^{i\Psi}$ (cf. Ex. *10,* p. 212) then CR takes the form

$$\partial_\theta R = -r\,R\,\partial_r \Psi \quad \text{and} \quad R\,\partial_\theta \Psi = r\,\partial_r R.$$

II An Intimation of Rigidity

A recurring theme in complex analysis is the "rigidity" of analytic functions. By this we mean that their highly structured nature (everywhere locally an amplitwist) enables us to pin down their precise behaviour from very limited information. For example, even if we are only told the effect of an analytic function on a small region, then its definition can be extended beyond these confines in a unique way—like a crystal grown from a seed. In fact, given even the meagre knowledge of how an analytic mapping affects a closed curve (just the points *on* the curve mind you), we can predict precisely what happens to each point *inside!* See [3]. Later we will justify these wild claims, and in Chapter 9 we will even find an explicit formula (due to Cauchy) for w in terms of A, B, C, etc. For the moment, though, we will

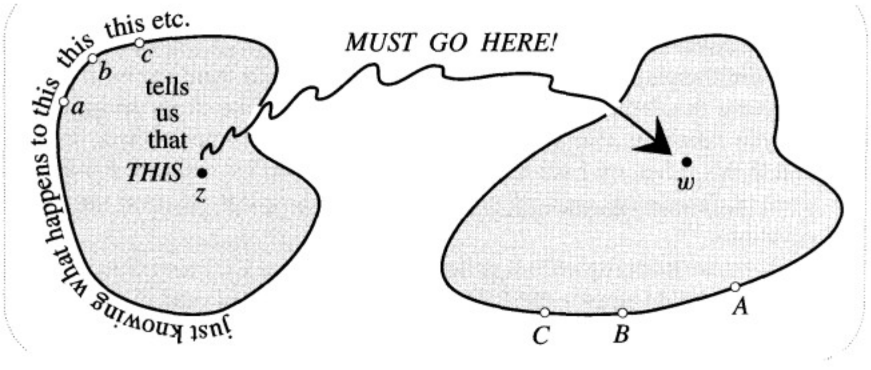

Figure **[3]**

'

+,- * *. ! * << C 2"

Visual Complex Analysis.pdf

Pag. 216 (242)

the Cartesian Form

Pag. 222 (248)

Visual Differentiation of log (z)

Pag. ~~25~~ 224 (250)

Inverse Functions

→ twist (torque) of f

MHHMY71.pdf

Pag. 30 (56)

38 – The Circle

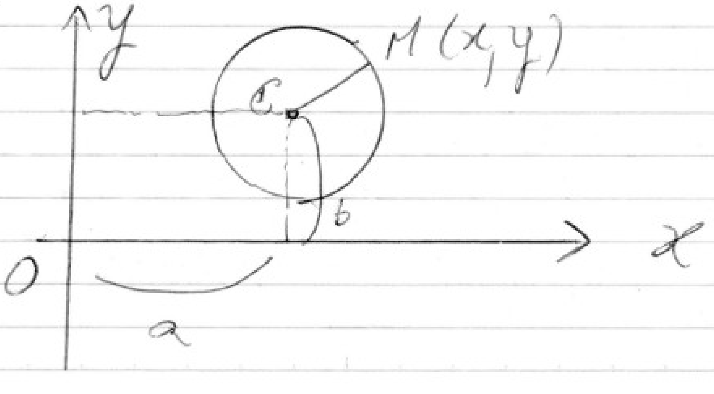

A circle of radius R with centre at the point C(a, b) is described by the equation

$$(x-a)^2+(y-b)^2=R^2$$

rewritten

$$x^2+y^2-2ax-2by+a^2+b^2-R^2=0$$

(20210725)

Mathematics Modern Calculus and Analytic Geometry. djvu (2012)

Pag. 527 (545)

Fig. 9.2

circle $x^2+y^2=a^2$, with parametric representation

$$x=a\cos t,\quad y=a\sin t,\quad (0\leq t\leq 2\pi)$$

ou

(cf Pag. 305)

There is nothing unique about the parametric representation of a curve

Ex:

$x = a \sin t$, $y = a \cos t$ $(0 \leq t < 2\pi)$ (2π)

or even

$x = a \cos \ln t$, $y = a \sin \ln t$

$(1 \leq t \leq e^{2\pi})$

represent the same circle

$$x^2 + y^2 = a^2$$

20210726th

reproduzir vídeos ao contrário.

Snow Patrol Open your Eyes

inspirado no filme "Era um encontro" (C'était un rendez-vous) 1976 por Claude Lelouch

---

')
+,- * *. ! * 7 ! F , ! " + "
! ! " + ! 0 6 0 : 7 3 ! (2"
' % !2)
+,- * *. ! * , ! " + 7 ! F ! !
" ! ! " + ! 0 6 0 : 7 3 ! "?7
' % %(% =)

CHAPTER 9 CURVES AND VECTORS IN R^2

67. CURVES IN GENERAL. PARAMETRIC EQUATIONS

In Sec. 30 we introduced the notion of a curve as a geometric figure which can be drawn without lifting pen from paper. Then we imposed the extra condition that the figure be the graph of a function (of x). Geometrically, this means that no line parallel to the y-axis intersects the curve more than once, or equivalently, that the pen moves steadily in one direction, from left to right or from right to left. To remove this unnatural restriction, which was only imposed temporarily to keep things simple, we make the point P traced out by the pen a function of a new real variable t, which can be thought of as the *time* at which P occupies one position or another. Thus Figure 9.1 shows a curve C and indicates the positions of a variable point P of C at various times.

More generally, we can think of t as any *parameter*, i.e., any real variable (not necessarily the time) uniquely determining the position of P. This is in keeping with the definition of a parameter on p. 446, as an “arbitrary constant each value of which characterizes a particular member of a set.” For example, in Figure 9.1 we can specify the position of P by giving the values of $t + 1$ or even of e^t, since knowledge of either $t + 1$ or e^t uniquely determines t and hence the position of P. Let x and y be the coordinates of P in a given rectangular coordinate system. Then specifying P as a function of t is the same as specifying x and y as functions of t, i.e., writing down a pair of equations

$$x = f(t), \qquad y = g(t), \tag{1}$$

called *parametric equations* of the curve since they involve a parameter t. Here it is tacitly assumed that the functions f and g have the same domain T. It is also assumed that f and g are not both constant, since otherwise C would reduce to a point. By the *graph* of (1), we mean the geometric figure obtained by plotting the point $P = (x, y) = (f(t), g(t))$ for every $t \in T$. If the graph of $x = f(t)$, $y = g(t)$ is a curve, we often save words by saying merely “the curve $x = f(t)$, $y = g(t)$.”

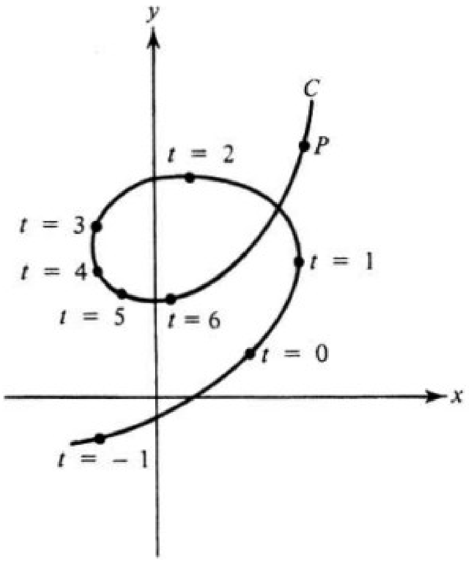

FIGURE 9.1

As on p. 217, we now require that T be some interval (closed, open or half-open, finite or infinite) and that f and g both be continuous in T. In other words, by a *curve* we mean the graph of a pair of parametric equations (1) where *the domain of f and g is an interval in which f and g are both continuous*. More concisely, a curve is the "continuous image of an interval." Roughly speaking, this means that the graph of (1) has no "gaps" and can be drawn without lifting pen from paper.[†] However, the graph of (1) may have no gaps even if the italicized condition is violated. (Why can't a similar complication occur in the case of the graph of a function $y = f(x)$?) For example, the symbol ∅ is most easily drawn by lifting pen from paper, corresponding to a pair of equations (1) with discontinuous f and g, or with a domain T other than an interval. On the other hand, it can also be drawn *without* lifting pen from paper, since we are now allowed to "trace back" over portions of a curve. In other words, ∅ is a curve, since it is the graph of some other pair of equations

$$x = f^*(t), \qquad y = g^*(t) \qquad (t \in T^*),$$

where T^* is an interval with f^* and g^* both continuous in T^*. Hence in talking about a curve $x = f(t)$, $y = g(t)$, $t \in T$, there is no loss of generality in assuming from the outset that T is an interval with f and g both continuous in T, and *we will do so from now on, writing I (for interval) instead of T.*

REMARK. By an *arc* of a curve with parametric equations

$$x = f(t), \qquad y = g(t) \qquad (t \in I),$$

we mean any curve with parametric equations

$$x = f(t), \qquad y = g(t) \qquad (t \in I')$$

where I' is a subinterval of I.

Example 1. The curve

$$y = \varphi(x) \qquad (a \le x \le b)$$

can be “represented in parametric form” by identifying x with the parameter t:

$$x = t, \qquad y = \varphi(t) \qquad (a \le t \le b).$$

Similarly, the graph of

$$x = \psi(y) \qquad (a \le t \le b),$$

(where ψ is continuous) is a curve, as we see by identifying y with the parameter t:

$$x = \psi(t), \qquad y = t \qquad (a \le t \le b).$$

Example 2. The curve with polar equation

$$r = f(\theta) \qquad (\alpha \le \theta \le \beta)$$

can be represented in parametric form with the angle θ as parameter by noting that

$$x = r\cos\theta, \qquad y = r\sin\theta,$$

and hence

$$x = f(\theta)\cos\theta, \qquad y = f(\theta)\sin\theta \qquad (\alpha \le \theta \le \beta)$$

(recall p. 505).

Example 3. The curve

$$x = a\cos t, \qquad y = b\sin t \qquad (0 \le t \le 2\pi) \tag{2}$$

is an ellipse if $a > 0$, $b > 0$, $a \neq b$. In fact, eliminating the parameter t from the two equations, we have

$$\frac{x}{a} = \cos t, \qquad \frac{y}{b} = \sin t,$$

and hence

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = \cos^2 t + \sin^2 t = 1, \tag{3}$$

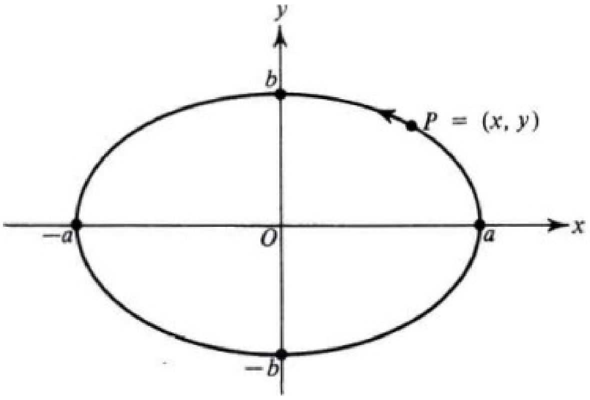

FIGURE 9.2

which we immediately recognize as the equation of the ellipse shown in Figure 9.2. Hence any point satisfying (2) belongs to the ellipse (3).

To see that the graph of (2) *coincides* with the ellipse, we examine the behavior of the variable point $P = (x, y) = (a \cos t, b \sin t)$ as t increases from 0 to 2π. For $t = 0$, P coincides with the vertex $(a, 0)$ of the ellipse. As t increases from 0 to $\pi/2$, x decreases to 0 and y increases to b, so that P describes the part of the ellipse in the first quadrant. As t increases further, P describes the rest of the ellipse, returning to the point $(a, 0)$ for $t = 2\pi$. If t is increased still further, P begins to trace out the ellipse once again. In fact, since cos t and sin t are both periodic, with fundamental period 2π, it is clear that P makes one complete circuit around the ellipse in the counterclockwise direction as P increases from t_0 to $t_0 + 2\pi$i, where r_0 is arbitrary. However, if $\delta < 2\pi$, then P describes only part of the ellipse as t increases from t_0 to $t_0 + \delta$.

It is easy to see that (2) represents the same ellipse (3) even if a or b is allowed to be negative, except that the ellipse is traced out in the clockwise direction if $ab < 0$. If $a = b$, (3) becomes the circle $x^2 + y^2 = a^2$, with parametric representation

$$x = a \cos t, \qquad y = a \sin t \qquad (0 \le t \le 2\pi)$$

(cf. p. 305). There is nothing unique about the parametric representation of a curve. For example, the equations

$$x = a \sin t, \qquad y = a \cos t \qquad (0 \le t \le 2\pi)$$

or even

$$x = a \cos \ln t, \qquad y = a \sin \ln t \qquad (1 \le t \le e^{2\pi})$$

represent the same circle $x^2 + y^2 = a^2$.

If $a = 0$, $b \neq 0$, the curve (2) reduces to the vertical line segment joining the points $(0, -b)$ and $(0, b)$, while if $a \neq 0$, $b = 0$, (2) reduces to the horizontal line segment joining the points $(-a, 0)$ and $(0, a)$, both segments being traversed twice. The case $a = b = 0$ is excluded, since then the curve reduces to the single point $(0, 0)$.

Example 4. Consider the curve

$$x = a \cosh t, \qquad y = b \sinh t \qquad (-\infty < t < +\infty), \tag{4}$$

where $a > 0$, $b > 0$. Eliminating t from the two equations, we get

$$\frac{x}{a} = \cosh t, \qquad \frac{y}{b} = \sinh t,$$

hence

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = \cosh^2 t - \sinh^2 t = 1, \tag{5}$$

which we immediately recognize as the equation of the hyperbola shown in Figure 9.3. Hence any point satisfying (4) belongs to the hyperbola (5). However, there are points of the hyperbola which do not satisfy (4), and in fact (4) represents only the right-hand branch of the hyperbola (the solid curve in Figure 9.3). To see this, we merely note that as t increases from $-\infty$ to 0, x decreases from $+\infty$ to a and y decreases from $-\infty$ to 0 (recall the properties of the hyperbolic functions from Sec. 40), while as t increases from 0 to $+\infty$, x increases from a to $+\infty$ and y increases from 0 to $+\infty$. In other words, the variable point $P = (x, y) = (a \cosh t, b \sinh t)$ moves upward along the right-hand branch of (5). If b is negative, then (4) gives the right-hand branch of (5) traversed in the *downward* direction. On the other hand, if a is replaced by $-a$ ($a > 0$), (4) gives the left-hand branch of (5) (the dashed curve in the figure), traversed in the upward direction if $b > 0$ and in the downward direction if $b < 0$. Moreover, the graph of (4) is the y-axis if $a = 0$, $b = 0$, the interval $(a, +\infty)$ if $a > 0$, $b = 0$, and the interval $(-\infty, -a]$ if a is replaced by $-a$, $b = 0$ (check all these claims).

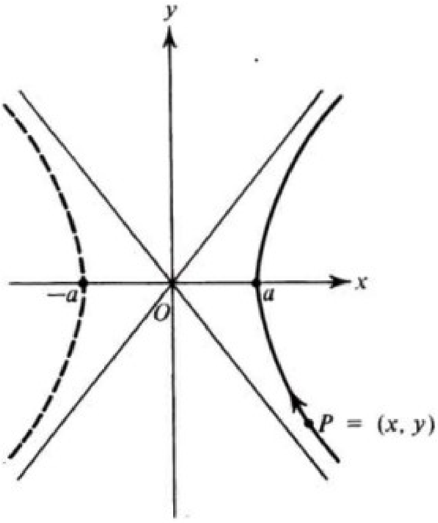

FIGURE 9.3

Example 5. Suppose a circle of radius a rolls without sliding along a horizontal straight line. Find the curve described by a fixed point P of its circumference.

Solution. Let the straight line be the x-axis, and let t be the angle through which the circle has rotated starting from a position in which P coincides with the origin O. Then the curve described by P, called a *cycloid*, has parametric equations

$$x = a(t - \sin t), \qquad y = a(1 - \cos t) \qquad (-\infty < t < +\infty). \tag{6}$$

In fact, examining Figure 9.4, we find that P has abscissa

$$x = |OQ| = |OO'| - a \sin t.$$

But $|OO'|$ equals the length of the arc $\widehat{O'P}$, since the circle rolls without sliding. Thus $|OO'| = at$, and hence

$$x = at - a \sin t = a(t - \sin t).$$

Similarly, we have

$$y = |O'A'| - a \cos t = a(1 - \cos t).$$

Note that this curve is the graph of a function $y = \varphi(x)$, where φ is periodic with period $2\sigma a$. However, to find φ explicitly, we would have to eliminate t from the equations (6) and then solve the resulting equation, of the form $F(x, y) = 0$, for y as a function of x. This is a formidable calculation, and a pointless one at that, since we can find out anything we want to know about the cycloid directly from the parametric equations (6).

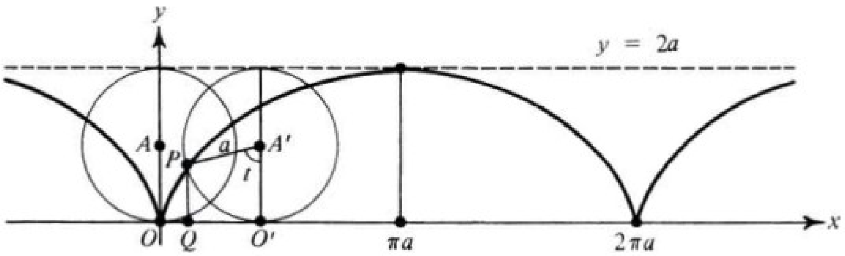

FIGURE 9.4

In general, of course, the curve with parametric equations

$$x = f(t), \qquad y = g(t) \qquad (a \le t \le b)$$

will not be the graph of any function

$$y = \varphi(x) \qquad (A \le x \le B).$$

For example, the unit circle

$$x = \cos t, \qquad y = \sin t \qquad (0 \le t \le 2\pi)$$

cannot be the graph of a function, since every line $x = c$, $-1 < c < 1$ cuts the circle *twice*. In fact, as noted in Example 2, p. 86, the circle is the union of the graph of

$$y = \sqrt{1 - x^2}$$

(the upper unit semicircle) and the graph of

$$y = -\sqrt{1 - x^2}$$

(the lower unit semicircle). On the other hand, every point P of the circle except the points $(-1, 0)$ and $(1, 0)$ belongs to either the upper semicircle or to the lower semicircle, i.e., to an arc of the circle which is the graph of a function.

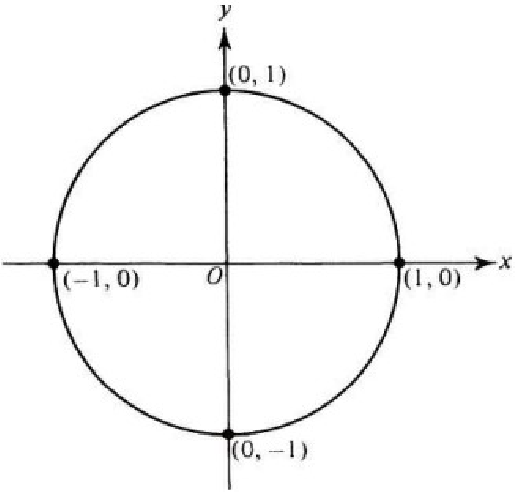

FIGURE 9.5

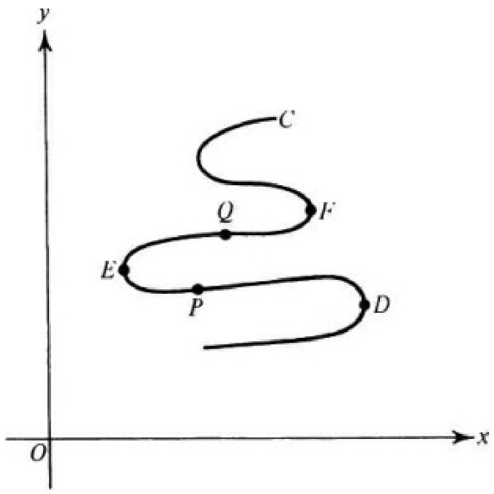

FIGURE 9.6

The same cannot be said of the points (−1, 0) and (1, 0), since any arc of the circle containing one of these points must overlap both the upper and the lower semicircles (see Figure 9.5). Similarly, the point P of the curve C shown in Figure 9.6 belongs to the arc $\widehat{DE}$ which is the graph of a function, the point Q belongs to the arc $\widehat{EF}$ which is the graph of a function, and so on, although the "extreme points" D, E, F themselves cannot be ascribed to arcs of the curve which are graphs of functions. In other words, an arc of a curve may be the graph of a function even though the curve *as a whole* is not. This is the substance of

THEOREM 9.1. *Given a curve*

$$x = f(t), \qquad y = g(t) \qquad (a \le t \le b),$$

where the functions f and g are differentiable in an open interval $(\alpha, \beta) \subset [a, b]$, *suppose the derivative*

$$\frac{dx}{dt} = f'(t)$$

does not change sign or vanish in (α, β). *Then the arc*

$$x = f(t), \qquad y = g(t) \qquad (\alpha < t < \beta) \tag{7}$$

(*part of the original curve*) *is the graph of a differ entiable function*

$$y = \varphi(x) \qquad (A < x < B),$$

and

$$\frac{dy}{dx} = \frac{g'(t)}{f'(t)}. \tag{8}$$

Moreover, if the functions f and g can be differentiated n times, so can the function φ.

Proof. Suppose $f'(t) > 0$ in (α, β). Then it follows from Theorem 6.24, p. 317, that f is increasing in $[\alpha, \beta]$. Therefore, by Theorem 6.9, p. 273, the inverse function $t = f^{-1}(x)$ exists and is continuous in $[f(\alpha), f(\beta)]$. Hence, as asserted, the arc (7) is the graph of the function

$$y = g(t) = g(f^{-1}(x)) \equiv \varphi(x) \qquad (A < x < B),$$

where $A = f(\alpha)$, $B = f(\beta)$.† Moreover, by the rule for differentiating an inverse function (Theorem 6.10, p. 280), dt/dx exists and equals

$$\frac{1}{\dfrac{dx}{dt}} = \frac{1}{f'(t)},$$

since $f'(t)$ does not vanish in (α, β). Hence

$$\frac{dy}{dx} = \frac{dy}{dt}\,\frac{dt}{dx} = \frac{\dfrac{dy}{dt}}{\dfrac{dx}{dt}} = \frac{g'(t)}{f'(t)},$$

by the rule for differentiation of a composite function (Theorem 5.6, p. 232). (The same argument has already been used on p. 506 to find the slope of a curve with polar equation $r = f(\theta)$.) If f and g are differentiable n times in (α, β), then

$$\begin{aligned} \frac{d^2y}{dx^2} &= \frac{d\left(\dfrac{dy}{dx}\right)}{dt}\,\frac{dt}{dx} = \frac{1}{f'(t)}\,\frac{d}{dt}\left[\frac{g'(t)}{f'(t)}\right] \\ &= \frac{g''(t)f'(t) - g'(t)f''(t)}{[f'(t)]^3}, \end{aligned} \tag{9}$$

and similarly for higher-order derivatives. ▌

REMARK 1. The conclusion that the arc (7) can be represented as the graph of $y = \varphi(x)$ remains valid even if $f'(t)$ vanishes at points of (α, β), provided $f'(f)$ is nonnegative (or nonpositive) and does not vanish in any subinterval of (α, β). This follows at once from Theorem 6.24. However, we can only use formula (8) at points where $f'(t) \neq 0$.

REMARK 2. The variables x and y are on a completely equal footing here. Thus suppose the derivative

$$\frac{dy}{dt} = g'(t)$$

does not change sign or vanish in (α, β). Then the arc (7) is the graph of a function of the form

$$x = \psi(y) \qquad (A < y < B)$$

(where now x is the dependent variable and y the independent variable), and

$$\frac{dx}{dy} = \frac{f'(t)}{g'(t)}.$$

For example, every point of the unit circle except the points (0, −1) and (0, 1) belongs to an arc of the circle which is the graph of a function of y, i.e., either to the right-hand semicircle

$$x = \sqrt{1 - y^2}$$

or to the left-hand semicircle

$$x = -\sqrt{1 - y^2}.$$

The same cannot be said about the points (0, −1) and (0, 1), since every arc of the circle containing one of these points must overlap both the right-hand and the left-hand semicircles (see Figure 9.5). However, *every* point of the unit circle belongs to an arc of the circle which is the graph of a function if we allow both functions of x and functions of y. Similarly, the points D, E, F of the curve C shown in Figure 9.6 now belong to arcs of the curve which are graphs of functions of the form $x = \psi(y)$.

Example 6. For the cycloid (6), we have

$$\frac{dx}{dt} = a(1 - \cos t) > 0 \tag{10}$$

unless

$$t = 2n\pi \qquad (n = 0, \pm 1, \pm 2, \ldots),$$

in which case

$$\frac{dx}{dt} = 0.$$

It follows from Theorem 9.1 and Remark 1 that the cycloid *as a whole* is the graph of a function $y = \varphi(x)$, as already noted. Moreover, (8) gives

$$\frac{dy}{dx} = \frac{\dfrac{dy}{dt}}{\dfrac{dx}{dt}} = \frac{\sin t}{1 - \cos t} = \cot\frac{t}{2} \tag{11}$$

if $t \neq 2n\pi$, i.e., if $x \neq 2n\pi a$. If $t = 2n\pi$, we cannot use (11), but then a direct calculation shows that $dy/dx = \infty$, corresponding to the vertical tangents to the cycloid at the points $(2n\pi a, 0)$ (see Figure 9.4).†

Since (10) holds in any interval $(2n\pi, (2n + 1)\pi)$ and the functions (6) can be differentiated any number of times, it follows from Theorem 9.1 that y can be differentiated any number of times with respect to x. In particular, using (9), we get

$$\frac{d^2y}{dx^2} = \frac{(1 - \cos t)\cos t - \sin^2 t}{a(1 - \cos t)^3} = -\frac{1}{a(1 - \cos t)^2} < 0.$$

Hence the curve is concave downward in every interval $(2n\pi, (2n + 1)\pi)$, as confirmed by Figure 9.4.

Problem Set 67

1. For what values of the parameter t does each of the following points lie on the indicated curve:
  - a) $(-9, 0)$ on $x = 3(2 \cos t - \cos 2t)$, $y = 3(2 \sin t - \sin 2t)$;
  - b) $(3, 2)$ on $x = t^2 + 2t$, $y = t^3 + t$;
  - c) $(2, 2)$ on $x = 2 \tan t$, $y = 2 \sin^2 t + \sin 2t$;
  - d) $(0, 0)$ on $x = t^2 - 1$, $y = t^3 - t$?

2. Write parametric equations for the square with vertices (0, 0), (1, 0), (1, 1) and (0, 1).

3. Find dy/dx if
  - a) $x = t^2$, $y = 2t$;
  - b) $x = \cos t$, $y = t + \sin t$;
  - c) $x = a \cos^2 t$, $y = b \sin^2 t$;
  - d) $x = \cos^3 t$, $y = \sin^3 t$;
  - e) $x = e^t \sin t$, $y = e^t \cos t$.

4. Prove that if

$$x = \frac{1 + \ln t}{t^2}, \qquad y = \frac{3 + 2 \ln t}{t} \qquad (t > 0),$$

then

$$y\frac{dy}{dx} = 1 + 2x\left(\frac{dy}{dx}\right)^2.$$

5. Find the slope of the tangent to the curve
  - a) $x = 3 \cos t$, $y = 4 \sin t$ at $(\frac{3}{2}\sqrt{2}, 2\sqrt{2})$;
  - b) $x = t - t^4$, $y = t^2 - t^3$ at $(0, 0)$;
  - c) $x = t^3 + 1$, $y = t^2 + t + 1$ at $(1, 1)$;
  - d) $x = 2 \cos t$, $y = \sin t$ at $(1, -\frac{1}{2}\sqrt{3})$.

6. Find d^2y/dx^2 if
  - a) $x = 2t^2$, $y = 3t^3$;
  - b) $x = a \cos t$, $y = b \sin t$;
  - c) $x = \cos^3 t$, $y = \sin^3 t$;
  - d) $x = t \cos t$, $y = t \sin t$.

7. Prove that if $x = \sin t$, $y = \sin at$, then

$$(1 - x^2)\frac{d^2y}{dx^2} - x\frac{dy}{dx} + a^2y = 0.$$

' $

www.mackenzie.br/gerti/
estudantes-e-funcionarios/softwares/
wolfram-mathematica.

Um buraco negro é menor que o raio de Schwarzschild.

MIT — supercondutor chamado grafeno de três camadas torcidos por ângulo mágico.

→ tripleto de spin

↓

a energia de ambos os elétrons muda na mesma direção, mantendo a supercondutividade independente da intensidade do campo magnético ao qual estão submetidos.

3 camadas de grafeno e a camada do meio foi girada em 1,56 graus em relação às camadas externas.

ângulo Gauss?

↳ curvatura

3 camadas torcidas poderá ser aplicado no desenvolvimento de computadores quânticos mais robustos.

Supercondutividade topológica
↳ ingrediente chave para fazer isso seriam supercondutores de tripletos de spin.

O grafeno poderá ser usado para modular os sinais de telecomunicações em velocidades ultrarrápidas podendo ser utilizados para bloquear frequências de comunicações indesejadas.

Grafeno ferroelétrico condutor
→ servirá para armazenar dados (memórias superrápidas para computadores)

Grafeno →
uma das formas cristalinas do carbono,
organizado em uma única camada de átomos de carbono dispostos em hexágonos.

NB – Nitreto de Boro →

(dois)
camada do meio entre 2 grafenos.
↳ ângulo ligeiramente diferente entre si.

→ Padrão moiré
(moiré)

→ padrão formado quando duas grades apresentam movimento relativo entre si.

Padrão moiré
↳ compensa a curvatura espacial de Gauss?

→ importante!!!

⇓

Primeiro ferroelétrico condutor,

"é a primeira demonstração que relata ferroeletricidade eletrônica pura, que exibe polarização de carga, sem movimento iônico na rede subjacente". Philip Kim
(Univ. Harvard).

Https:// corsi.unibo.it/magistrale/
matematics

Coordinatore: Marilena Barnabei

' $(

https://www.ifsc.usp.br/
~strontium/Publication/Scripts/
QuantumMechanicsScript.pdf

---

studenti.unibo.it

20210727

Angstrom = 0,1 nm

10^{-10} m   10.000.000.000
10^{-8} cm

1 Å = 100 picômetros.

1 Å = usado para medir o espectro eletromagnético
(electromagnetic spectrum)

Albert A. Michelson defined so that the red line of cadmium was equal to 6438.47 Å.

1907 → 15°C dry air (hydrogen scale)
760 mm Hg
9.8067 m/s² (gravity)

$= 1\,\text{Å} = 6438.4696$

angstrom remained a secondary unit of length for use in spectroscopy.

$7.2^{\wedge 2} = 50$ $\quad 7.2^{\wedge 7.2} = 5{,}184$

$7.1^{\wedge 2} = 50{,}41$ $\quad 7.2^{\wedge 2} = 51{,}84$

$\sqrt{50} = 7{,}071\ldots$

$7{,}07^{\wedge 2} \overset{!!}{=} 49{,}9849$

$\downarrow$ $\qquad \sqrt{0{,}0151} = 0{,}12288\ldots$

$7{,}071$

$0{,}001 \rightarrow \sqrt{0{,}001} = 0{,}0316227\ldots$

$\sqrt{100} = 10$

$\sqrt{50} = \quad \rangle\ 7{,}5$

$\sqrt{25} = 5$ $\qquad 7{,}5^{\wedge 2} = 56{,}25$

$\rightarrow$ error!!!

$100 - 25 = 75$ $\quad \sqrt{75} = 8{,}66025\ldots$

vejamos:

$\sqrt{100} = 10$ } 1 33974...

$\sqrt{75} = 8,66025...$ } 1'589186...

$\sqrt{50} = 7,07106...$ } 2,07...

$\sqrt{25} = 5'$

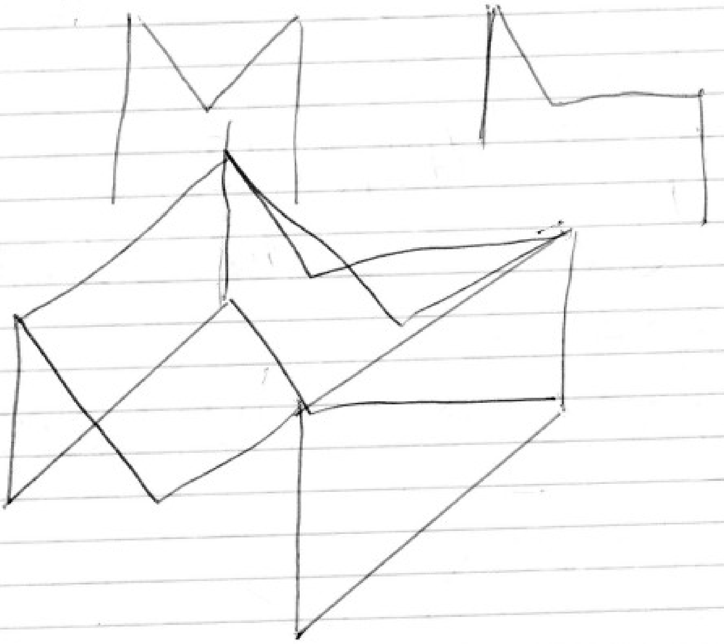

→ cubo

' $

data

20210727

WILLIAM OUCHI
SOLLEDADE M.

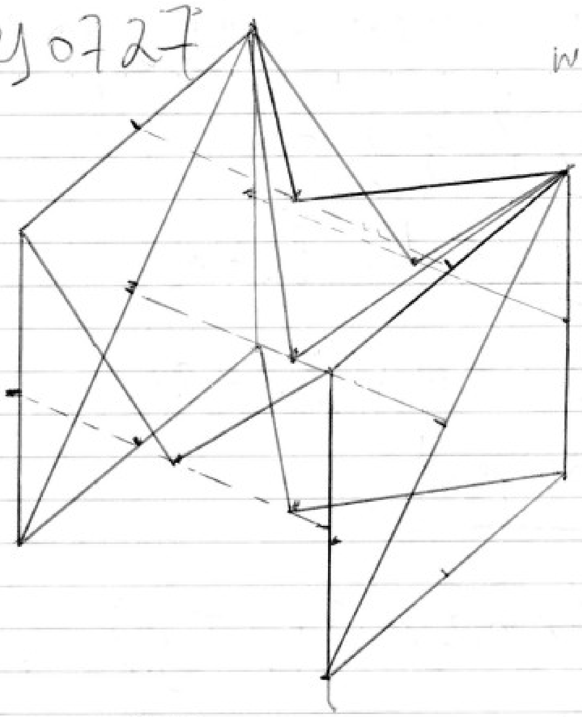

Line $-x+y=0$

Important!

r_2 : Line (E_1, F_1)

$\rightarrow -4.2x + 4.27y = -5,87$

(F_1, E_1)

$\rightarrow 4.2x - 4.27y = 5.87$

Jandaia

Alinhamento de 3 pontos

$G_1 (3, 1.5)$

$H_1 (5.12, 3.62)$

$I_1 (7.24, 5.74)$

Important:

$g_2 = \text{Line}(G_1, I_1)$

$\rightarrow -4.24x + 4.24y = -6.36$

$$\left.\begin{array}{l} G_1H_1 = 3 \\ H_1I_1 = 3 \end{array}\right\} \text{distance}$$

Fazer 3 perpendiculares nos pontos G_1 H_1 I_1

clicando novos pontos em cima destes teremos:

20210728 III

G_1 H_1 I_1 $\Rightarrow$ J_1 $\boxed{K_1}$ L_1

↳ centro

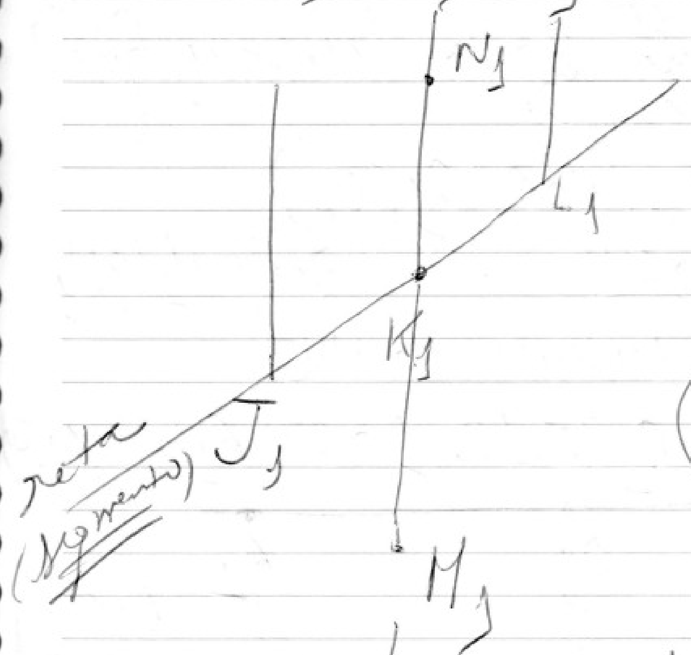

N_1
L_1
K_1
J_1
reta (segmento)
M_1

M_1 K_1 N_1

↳ 3 pontos

Portanto:

J_1 K_1 L_1 e M_1 K_1 N_1

→ Pappus Theorem

⇒ Temos 2 retas

Important: Point O_1 (ápice)

O_1 = Intersect(g_1, n_1)

→ (4.24, 10.24).

Fazer uma linha perpendicular sobre a reta criada pelos pontos $M_1K_1N_1$ que passa pelo ponto O_1

PerpendicularLine(O_1, s_2)

$Y = 10.24$ ↳ reta $M_1K_1N_1$

MUS: Parte superior do MUS

$P_1(3, 10.24)$ $Q_1(5.12, 10.24)$ $R_1(7.24, 10.24)$

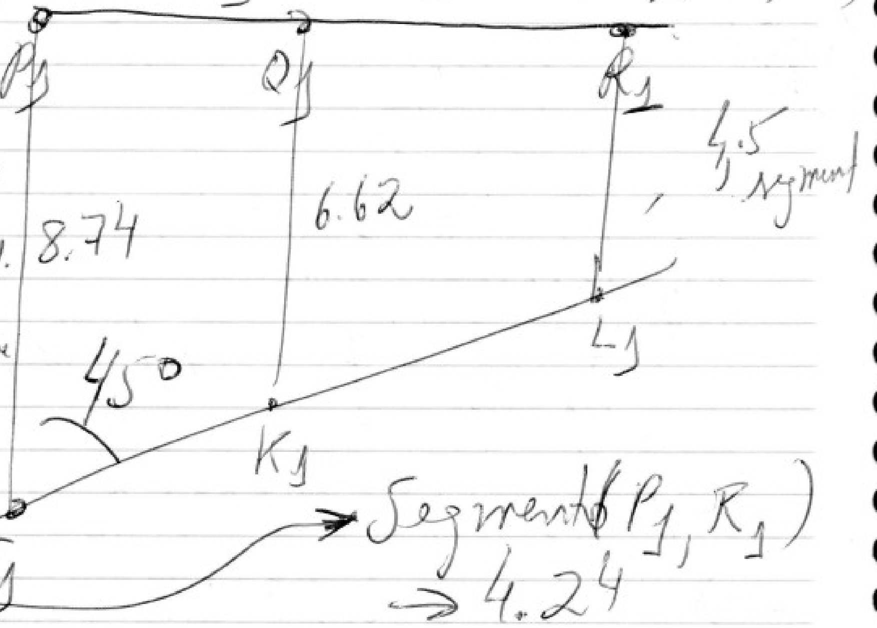

Comprimento reta J_1L_1 = 6 unid.

Diferença entre $J_1P_1 - L_1R_1$ = 8.74 − 4,5 = $\boxed{4,24}$ =

Segment(P_1, R_1)

→ 4.24

Source:

ARQ_IMPORTANT_03/Geogebra-M01-20210727/

m01-20210727-13.ggb

20210728 III

Fazer este mesmo desenho com as arestas com 8 unidades.

Simular também em 3D (forma cubo)

Copiar os comandos dados para se criar este modelo do MUS.

ex: A= Intersect(xAxis, yAxis)
→ (0,0)

B = Point(yAxis)
→ (0,6) etc

COMANDOS GEOGEBRA

A= Intersect(xAxis, yAxis)
→ (0,0)

B = Point(yAxis) → (0,6)

C = (6, -3)

D = (6, 3)

c = Circle(C, Segment(C,D)) →
$(x-6)^2+(y+3)^2=36$

d: Circle(A, Segment(A,B)) →

$x^2+y^2=36$

$E=(1,1)$

$F=(7,-2)$

G=(Point(Circle(A,6)) $\Rightarrow (6,0)$

f: Line(A,E) $\Rightarrow -x+y=0$

H=Intersect(d,f,2) $\Rightarrow$ $(4.24, 4.24)$

g= Segment(A,H) $\Rightarrow 6$

h: Line(G,F) $\Rightarrow -x+y=-9$

I=Intersect(c,h,2) $\Rightarrow$ (1a24, 1.24)

i= Segment(G,I) $\Rightarrow 6$

j= Segment(A,B) $\Rightarrow 6$

k= Segment(G,D) $\Rightarrow 6$

J= Point(j) $\Rightarrow$ (9 3)

l= Segment(G,J) $\Rightarrow 6.71$

m= PerpendicularBisector (G,J) $\Rightarrow 6x-3y=13.5$

K=Intersect(m, l) $\Rightarrow$ $(3, 1.5)$

n= Segment(K,B) $\Rightarrow 5.41$

p= Segment(K,D) $\Rightarrow 3.35$

q: PerpendicularBisector (G,I) $\Rightarrow -4.24x-4.24y=-30.73$

L= Intersect(q, i) $\Rightarrow$ $(8.12, -0.88)$

r= Perpendicular Bisector(A,H) $\Rightarrow -4.24x-4.24y=-18$

M= Intersect(r, g) $\Rightarrow$ $(2.12, 2.12)$

s= Segment(M,L) $\Rightarrow 6.71$

t: Perpendicular Bisector (M,L) $\Rightarrow -6x+3y=28{,}86$

N= Intersect(t, s) $\Rightarrow$ $(5.12, 0.62)$

a= Segment(I,N) $\Rightarrow 5.16$

b= Segment(N,H) $\Rightarrow 3.73$

$O = (1, 7)$

$P = (7, 4)$

e : Line(B, O) ⇒
$-x + y = 6$

f_1 : Line(D, P) ⇒
$-x + y = -3$

c_1 : Circle(B, Radius(d))
$\Rightarrow x^2 + (y-6)^2 = 36$

Q = Intersect(c_1, e, 2) ⇒
(4.24, 10.24)

g_1 = Segment(B, Q) → 6

d_1 : Circle(D, Radius(c))
$\rightarrow (x-6)^2 + (y-3)^2 = 36$

R = Intersect(d_1, f_1, 2) →
(10.24, 7.24)

h_1 = Segment(D, R) →
6

i_1 : PerpendicularBisector
(B, Q) →
$-4.24x - 4.24y =$
-43.46

j_1 : PerpendicularBisector
(D, R) ⇒ $-4.24x - 4.24y =$
-56.18

S = Intersect(e, i_1) ⇒
(2.12, 8.12)

T = Intersect(f_1, j_1) →
(8.12, 5.12)

k_1 : Line(S, T) ⇒ $3x + 6y = 55.09$

l_1 : PerpendicularBisector(S, T) ⇒
$-6x + 3y = -10.86$

U = Intersect(k_1, l_1) ⇒ (5.12, 6.62)

m_1 = Segment(U, R) ⇒ 5.16

n_1 = Segment(U, Q) ⇒ 3.73

p_1 = Segment(H, Q) → 6

q_1 = Segment(I, R) → 6

r_1 : Perpendicular Bisector(I, R) →
$y = 14.24$

s_1 : Perpendicular Bisector(H, Q) →
$y = 7.24$

V = Intersect(r_1, q_1) ⇒
(10.24, 4.24)

W = Intersect(s_1, p_1) ⇒ (4.24, 7.24)

t_1 = Segment(W, V) ⇒ 6.71

a_1 : Perpendicular Bisector
$(V, W) \Rightarrow 6x - 3y = 26.23$

z = Intersect$(t_1, k_1) \rightarrow$
undefined.

A_1 = Intersect$(a_1, t_1) \rightarrow$
$(7.24, 5.24)$

b_1 = Segment$(Q, A_1) \rightarrow$
5.41

e_1 = Segment$(R, A_1) \Rightarrow 3.35$

f_2 = Segment$(A, Q) \Rightarrow 11.09$

g_2 = Segment$(G, R) \Rightarrow 11.09$

h_2 : Perpendicular Bisector$(A, Q) \Rightarrow$
$-14.24x - 10.24y = -61.46$

B_1 = Intersect$(g, h_2) \Rightarrow$
$(12.12, 5.12)$

i_2 : Perpendicular Bisector(G, R)
$\Rightarrow -4.24x - 10.24y = -56.18$

C_1 = Intersect$(i_1, i_2) \rightarrow$
$(8.12, 2.12)$

j_2 : Line$(B_1, C_1) \rightarrow$
$3x + 6y = 37.09$

k_2 : Perpendicular Bisector(B_1, C_1)
$\rightarrow -6x + 3y = -19.86$

D_1 = Intersect$(j_2, k_2) \Rightarrow$
$(5.12, 3.62)$

l_2 = Segment$(D_1, R) \rightarrow$
6.27 (6.27)

m_2 = Segment$(D_1, Q) \rightarrow$
6.68

E_1 = Point$(n) \Rightarrow (2.97, 1.54)$

F_1 = Intersect$(b_1, e_1) \Rightarrow$
$(7.24, 5.24)$

n_2 : Line$(E_1, F_1) \rightarrow$
$-4.2x + 14.27y =$
-5.87

p_2 : Line$(F_1, E_1) \rightarrow$
$4.2x - 4.27y =$
5.87

G_1 = Intersect$(n, p) \Rightarrow$
$(3, 1.5)$

H_1 = Intersect(l_2, m_2)
$\rightarrow (5.12, 3.62)$

I_1 = Intersect(b_1, e_1)
$\rightarrow (7.24, 5.74)$

q_2 = Line$(G_1, I_1) \rightarrow$
$-4.24x + 4.24y =$
-6.36

Jandala

J_1 = Intersect(n, p) → (3, 1.5)

r_2 : Line(J_1, k) → $x = 3$

K_1 = Intersect (l_2, m_2) → (5.12, 3.62)

s_2 : Line(K_1, k) → $x = 5.12$

L_1 = Intersect(b_1, e_1) → (7.24, 5.74)

t_2 : Line(L_1, k) → $x = 7.24$

M_1 = Intersect(s_2, a) → (5.12, 0.62)

N_1 = Intersect (s_2, m_1) → (5.12, 6.62)

O_1 = Intersect (g_1, n_1) → (4.24, 10.24)

a_2 : PerpendicularLine (O_1, l_2) → $y = 10.24$

P_1 = Intersect(r_2, a_2) → (3, 10.24)

Q_1 = Intersect(s_2, a_2) → (5.12, 10.24)

R_1 = Intersect(t_2, a_2) → (7.24, 10.24)

S_1 = Intersect(r_2, n) → (3, 1.5)

b_2 = Segment(S_1, P_1) → 8.74

T_1 = Intersect(s_2, l_2) → (5.12, 3.62)

c_2 = Segment(T_1, Q_1) → 6.62

U_1 = Intersect(t_2, b_1) → (7.24, 5.24)

d_2 = Segment(U_1, R_1) → 4.5

e_2 = Segment(P_1, R_1) → 4.24

V_1 = Intersect(b_1, e_1) → (7.24, 5.74)

f_3 = Segment(S_1, V_1) → 6

α = Angle(V_1, S_1, P_1) → 45°

Source: 8 unidades em 3D →

(2) m01-3D-20210728-05.ggb

Checagem dos passos: (20210728)

A = (0, 0)

B = (0, 6)

C = (6, -3)

D = (6, 3)

E = (1, 1)

F = (7, -2)

G = (1, 7)

H = (7, 4)

f: Ray(A, E) → $-x+y=0$

g: Ray(C, F) → $-x+y=-9$

h: Ray(B, G) → $-x+y=6$

i: Ray(D, H) → $-x+y=-3$

Circles nos pontos
A, B, C e D
usando o segmento
$\overline{AB}$.

Ex:

c : Circle(A, Segment(A, B))
→ $x^2+y^2=36$

d : Circle(B, Segment(A, B))
→ $x^2+(y-6)^2=36$

e : Circle(C, Segment(A, B))
→ $(x-6)^2+(y+3)^2=36$

k : Circle(D, Segment(A, B))
→ $(x-6)^2+(y-3)^2=36$

Intersect das
retas com os
círculos:

I = Intersect(c, f, 1)
→ (4.24, 4.24)

J = Intersect(d, h, 1)
→ (4,24, 10.24)

K = Intersect(e, g, 1)
→ (10.24, 1.24)

L = Intersect(k, i, 1)
→ 10,24, 7.24)

Não deixe visível os
Raios e os Círculos,
deixando os pontos de
intersecção visíveis para
podermos fazer as
retas do cubo.
Apague também os
pontos (esconder):
E, F, G e H.

Erro:

Distância $\overline{AC}=6,71$

tinha que ser 6,00

vou refazer
colocando

2 arestas sobre o
eixo x e y.

(menos chance de
erro)

Check 02 (20210728).

A = Intersect (xAxis, yAxis) - ⇒ (0, 0)

B = Point(xAxis) → (6, 0)

C = Point(yAxis) → (0, 6)

D = (6, 6)

—— 45°

E = (1, 1)

F = (7, 1)

G = (1, 7)

H = (7, 7)

Criar os Rays:

f: Ray(A, E) → $-x + y = 0$

Usarei a mesma reta no ponto D → continuidade da reta

Obs: a mesma reta AE continua em DH.

g: Ray(B, F) → $-x + y = -6$

h: Ray(C, G) → $-x + y = 6$

Criar os círculos e as interseções sobre os Rays $\overline{AE}$, $\overline{BF}$, $\overline{CG}$ e $\overline{DH}$.

Usando a função COMPASS:

c: Circle(A, Segment(A, C))
→ $x^2+y^2=36$

d: Circle(B, Segment(A, C))
→ $(x-6)^2+y^2=36$

e: Circle(C, Segment(A, C))
→ $x^2+(y-6)^2=36$

k: Circle(D, Segment(A, C))
→ $(x-6)^2+(y-6)^2=36$

obs: Alterar as cores destes círculos.

Usar o comando Intersect:

I = Intersect(c, f, 1) → (4.24, 4.24)
J = Intersect(d, g, 1) → (10.24, 4.24)
K = Intersect(e, h, 1) → (4.24, 10.24)
L = Intersect(k, f, 2) → (10.24, 10.24)

Obs: Esconda os círculos e os Rays, deixando apenas as interseções.

Depois, criaremos as retas do cubo.

Fazer os segmentos do Cubo:

i = Segmento(A, C) → 6 (altera a cor já)

j = Segment(B, D) → 6

Crie o ponto M=(3, 3)

Crie os novos segmentos:

l = Segment(C, M) → 4.24

m = Segment(M, D) → 4.24

n = Segment(A, I) → 6 (preto)

p = Segment(B, J) → 6 (preto)

q = Segment(C, K) → 6 (preto)

r = Segment(D, L) → 6 (preto)

s = Segment(I, K) → 6 (vermelho)

t = Segment(J, L) → 6 (Vermelho)

Novos comandos:

a: PerpendicularBisector(I, K) → y = 7.24

obs: Vale para o segmento s(I, K) e t(J, L)

Intersect:

N = Intersect(a, s) → (4.24, 7.24)

O = Intersect(a, t) → (10.24, 7.24)

Obs: Esconda a linha **a**.

Important: Crie:

b: PerpendicularBisector(N, O) → x = 7.24

Importante observar que este ponto coincide com a reta r(D, L) → somente na perspectiva.

Obs: reta a cobre a reta r. (interseção visual, mas não real).

Criemos a interseção:

P = Intersect(a, b) → (7.24, 7.24)

Obs: Agora sim, esconda o segmento **a e b**.

Crie os segmentos:

f_1 = Segment(K, P) → 4.24 (vermelho)

g_1 = Segment(P, L) → 4.24 (vermelho)

Divisão de segmentos:

Segmentos $\overline{CK}$ e $\overline{DL}$ com o comando Bisector

h_1: PerpendicularBisector(C, K):
$\Rightarrow -4.24x - 4.24y = -43.46$

i_1: Perpendicular Bisector(D, L):
$\Rightarrow -4.24x - 4.24y = -68.91$

Intersect:

R = Intersect(h_1, g) → (2.12, 8.12)

S = Intersect(i_1, r) → (8.12, 8.12)

Obs: Esconda h_1 e I_1.

Crie Segment:

j_1 = Segment(R, S) → 6

k_1: Perpendicular Bisector(R, S) ⇒
$x = 5.12$

T = Intersect(k_1, j_1) → (5.12, 8.12)

Esconda j_1 e k_1.

Outros comandos (base cubo)

l_1 : Perpendicular Bisector (A, I) →
$-4.24x - 4.24y = -18$

m_1 : Perpendicular Bisector (B, J) →
$-4.24x - 4.24y = -43.46$

Intersect:

$U = \text{Intersect}(l_1, n) \rightarrow (2.12, 2.12)$
$V = \text{Intersect}(m_1, p) \rightarrow (8.12, 2.12)$
$n_1 = \text{Segment}(U, V) \rightarrow 6$
p_1 : Perpendicular Bisector (U, V) →
$x = 5.12$
$W = \text{Intersect}(p_1, n_1) \rightarrow (5.12, 2.12)$

Obs: Apague o resto e deixe somente este ponto.

$q_1 = \text{Segment}(I, W) \rightarrow 2.3$
$r_1 = \text{Segment}(W, J) \rightarrow 5.54$
$s_1 = \text{Segment}(K, T) \rightarrow 2.3$
$t_1 = \text{Segment}(T, L) \rightarrow 5.54$

Other:

Segmento Inclinado:

a_1 = Segment(A, K) → 11.09

b_1 = Segment(B, L) → 11.09

c_1: Perpendicular Bisector(A, K) →

$-424x - 10.24y = -61.46$

d_1: Perpendicular Bisector(B, L) →

$-4.24x - 10.24y = -86.91$

Z = Intersect(c_1, a_1) → (2.12, 5.12)

A_1 = Intersect(d_1, b_1) → (8.12, 5.12)

e_1 = Segment(Z, A_1) → 6

f_2: Perpendicular Bisector(Z, A_1) →

$x = 5.12$

B_1 = Intersect(e_1, f_2) → (5.12, 5.12)

Important:

Os Pontos M, B_1 e P estão em linha reta.

Distância entre MP → 6 unidades.

Parte Final:

Jandaia

Crie a linha, ligando os pontos K L

i_2 : Line (K, L) → $y = 10.24$

Criemos 3 segmentos com o comando Parallel Line baseado no segmento AC (6 unidades)

j_2 : Line(M, i) → $x = 3$

k_2 : Line(B_1, j_2) → $x = 5.12$

I_2 : Line(P, k_2) → $x = 7.24$ (7.24)

Intersect:

C_1 = Intersect(i_2, j_2) → (3, 10.24)

D_1 = Intersect(i_2, k_2) → (5.12, 10.24)

E_1 = Intersect(i_2, I_2) → (7.24, 10.24)

FINAL:

m_2 = Segment(M, C_1) → 7.24 ⟩ 2.12

n_2 = Segment(B_1, D_1) → 5.12 ⟩ 2.12

p_2 = Segment(P, E_1) → 3

soma

q_2 = Segment(C_1, E_1) → 4.24 =

Source:
ARQ-IMPORTANT-03/
Geogebra-M01-20210727/
m01-20210728-check-01-012.ggb
(arquivo final).

Obs: O próximo arquivo será criado para excluir os pontos e os nomes dos elementos, ficando apenas a imagem final.

m01-20210728-check-01-014.ggb

Imagens Criadas:

m01-20210728-check-01.png
m01-20210728-check-01.pdf

Duração: 12:00 - 17:45h
(20210728)

' %

20210729

Estreito de Málaca (Singapura)
Estreito de Sunda (Jacarta)
Estreito de Lombok.

—

"O homem que morre rico, morre desonrado" Carnegie.

—

scientificmagazines.top

—

20210730

New simulators:

File: m01-20210730-check_01-001.ggb

c_2: Circle(B_1, M) →

$$(x-5.12)^2+(y-5.12)^2=9$$

Important:

3 Points:

1) Point P (7.24, 7.24)
2) Point B_1 (5.12, 5.12) e
3) Point M (3, 3)

Estes três pontos estão alinhados dentro do círculo c_2 e dois destes pontos estão sobre a circunferência.

Outros 3 pontos:

1) Point T(5.12, 812),
2) Point B_1(5.12, 5.12) e
3) Point W(5.12, 2.12)

⇒ 3 pontos alinhados e 2 pontos sobre a circunferência.

Todos os 6 pontos se encontram num mesmo ponto B_1, ou seja, as retas $\overline{PM}$ e $\overline{TW}$ fazem intersecção no ponto B_1.

De acordo com Pappus Theorem observamos uma reta criada a partir de 3 pontos.
O ângulo entre estas retas é de 45°.

Obs: Simulação em 3D:
m01-3D-20210730-01.ggb

Usando a imagem 3D, utilizei o comando: Circle through 3 Points

Points: M_1, K_1, N_1

c_2: Circle(M_1, K_1, N_1) ⇒

$\Rightarrow X=(4,4,4)+(0,-2{,}83\cos(t)-2.83\sin(t),$
$2.83\cos(t)-2{,}83\sin(t))$

Localização dos pontos:

Jandaia

$M_1 = \text{Intersect}(j_1, n_2) \rightarrow (4,8,4)$

$K_1 = \text{Intersect}(b, m_2) \rightarrow (4,4,8)$

$N_1 = \text{Intersect}(j, I_2) \rightarrow (4,0,4)$

Marcando as 2 retas que se intersectavam no ponto (B_1) J_1
↳2D ↳3D.

a) $M_1(j_1, n_2)$, $J_1(4,4,4)$ (m_2, g_2) e $N_1(j, I_2)$

↓ ↓ ↓

$(4,8,4)$ $(4,4,4)$ $(4,0,4)$

b) $K_1(b, m_2)$, $J_1(m_2, g_2)$ e $L_1(s_1, r_1)$

↓

$(4,4,8)$ $(4,4,4)$ $(4,4,0)$

Estudar:

https://en.m.wikipedia.org/wiki/
Squaring_the_circle

obs: File 01 → Squaring the circle –
Ramanujan – 1914.gif

File:01-Squaring_the_circle-Ramanujan-1914.gif

Circulo horizontal:

d_2: Circle(J_1, 4, m_2) ⇒
$x = (4, 4, 4) + (4\cos(t), 4\sin(t), 0)$

Outros dados:

g_2 = Segment(N_1, K_1) ⇒ 5.66
O_1 = Intersect(edgeBC, k) → (8, 0, 8)
r_2 = Segment(O_1, J_1) ⇒ 6,93
Source:
m01-3D-20210730-03.ggb

20210801

Sir Chandrasekhara Venkata Raman, físico indiano que em 1930 se tornou o primeiro asiático a ganhar o Prêmio Nobel de Física.

Trabalho de Raman é centrado na dispersão da luz, que mostra que quando a luz atravessa um material transparente, parte da luz dispersada muda o comprimento de onda e amplitude, conhecido como

Efeito Raman princípio fundamental no campo da óptica

Ele descobriu que a cor do oceano é azul não por uma simples reflexão do céu, mas porque a água em si faz com que a luz azul se espalhe.

20210802

http://www.edocente.com.br/pnld/2018/
2019
2020

---

http://ramanujan.math.trinity.edu/wtrench/texts/TRENCH_FREE_DIFFEQ_I.pdf
II/

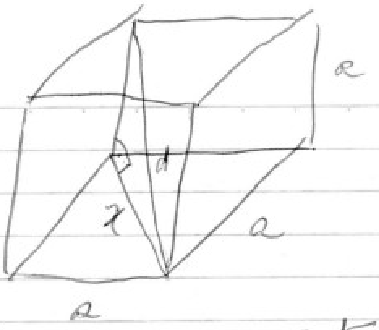

Cubo é um caso particular de paralelepípedo reto retangular

$$d = \sqrt{a^2+b^2+c^2} = \sqrt{3a^2} = a\sqrt{3}$$
$$d = a\sqrt{3}$$

Source : Books-016-2021/20210802/ Math/Math-Dante-3-Volumes/ Matematica Cont Aplic_2_MP_ 0008P18023_PNLD2018.pdf
Pág. 177 (179)

others :

Books-016-2021/20210802/Math/ ebook_final_isbn.pdf
Pág. 119 (132) - capítulo 6
Atividades Envolvendo Circunferência com o Uso do Geogebra

—

Books-016-2021/20210802/ Differential-Equations/ math135.pdf

npde10.pdf

Jandaia

$! 9
! *+,-* * *. ! * $
! ! ! +2 M M M $M 4>: 2"
3 *+,-* * *. ! * $
! ! ! +2 M M M $M 4>: 2"
+,- * *. ! * $
! ! ! +2 M$M M $M 4>: 2"
.3
+,- * *. ! * $
! ! ! +2 M M M $M 4>: 2"
')

Cálculo da diagonal de um paralelepípedo retângulo e de um cubo

No paralelepípedo de dimensões a, b e c, temos:

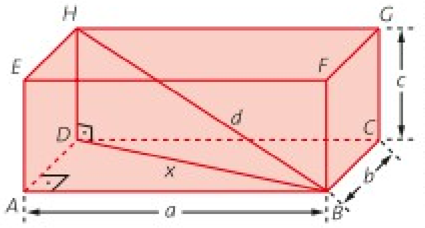

Ilustrações técnicas desta página: Banco de imagens/Arquivo da editora

d = medida da diagonal do paralelepípedo
x = medida da diagonal da base

Na figura podemos localizar dois triângulos retângulos:

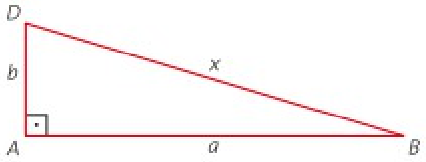

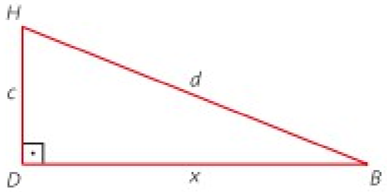

- Como o triângulo ABD é retângulo em A, temos, pela relação de Pitágoras:
  $x^2 = a^2 + b^2$ Ⓘ
- Como o triângulo DBH é retângulo em D, temos, pela relação de Pitágoras:
  $d^2 = x^2 + c^2$ Ⓘ Ⓘ
- Substituindo Ⓘ em Ⓘ Ⓘ, vem:
  $d^2 = x^2 + c^2 = a^2 + b^2 + c^2 \Rightarrow d = \sqrt{a^2 + b^2 + c^2}$

No cubo, como ele é um caso particular de paralelepípedo reto retangular, temos:

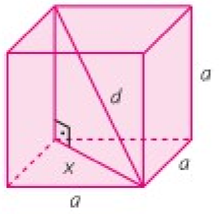

$$d = \sqrt{a^2 + a^2 + a^2} = \sqrt{3a^2} = a\sqrt{3}$$

$$d = a\sqrt{3}$$

20210803

Archimedes (240 A.C)

$22/7 < \pi < 223/71$

William Oughtred foi o primeiro a empregar a letra grega π como símbolo del coeficiente entre las longitudes de uma circunferência e (comprimento) seu diâmetro.

off: $\pi \cong 335/113$

Tsu Ch'ung-chih (480 A.C)

Source: 3con14.com/32-útiles/curiosidades/71-decimales-de-pi.html

---

Milagre!!!

Kombi Quebrou 20191104 (2 dias depois do (Seg-Feira) Dia de Finados)

Google (mark) - -23.61661, -46.62734

Temperatura: 37°C

Na esquina tinha um cara vendendo água gelada

SP-160 → Av. Prof. Abraão de Morais

200 metros do Shopping Plaza Sul.

(embaixo da árvore que tem recuo para estacionar - 50 m General Santos

As peças para consertar a Kombi foram compradas na R. Gen. Chagas Santos → a primeira rua à esquerda.

Na loja no cruzamento da R. General Chagas Santos e a Rua Curimina parema → Auto Peças Carvalho (Kombi à loja → 300 metros)

' 9%

20210805 ///

Spacetime - Marcus Kriele
The Ricci Flow in Riemannian
Geometry - Ben Andrews

Método catenária → conexão do
sistema pantográfico aos cabos
superiores (energia elétrica).

Schwartz surface.

Joseph Raphson x Newton
(50 anos)

→ cycle contraction for 42+6
results in 22/7 as 48 and 42 finds
a simple pi in between, for 7
others: https://community.wolfram.com/
groups/-/m/t/2311778

→ Newton - Raphson method. (1690)
method for approximating
the roots of an equation.

Raphson → Analysis Aequationum
Universalis (1690)

Newton → Method of Fluxions (1671),

but this work would not be published until 1736, nearly 50 years after Raphson's Analysis.)

Raphson was a staunch supporter of Newton's claim, and not that of Gottfried Leibniz, to be the sole inventor of calculus.

In addition, Raphson translated Newton's Arithmetica Universalis into English.

—

Raphson – De Spatio Reali published in 1697.

—

Standard candles and the value that is calculated using the lambda-cold-dark-matter (Λ CDM) model of cosmology.

—

300.000.000 planetas ⇔ em nossa galáxia

↓

1 planeta habitável:

Kepler-442b (KOI-4742.01) descoberto em 2015.

(condições semelhante às da Terra).

20210806

Data-NovoVivo-016-2021/Books-016-2021/20210806\Doing Mathematics Convection, Subject, Calculation, Analogy/

Martin H. Krieger - Constitution of Matter .... djvu

Pág. 37 (57) - the thermodynamic limit.

---

to + verbo infinitivo

for + gerúndio

Ex: She called to talk to you.

Conversations are good for qualifying what's wrong.

I will write a letter to her.

We gave the new to them.

He always carries the books for her.

She will check that for me.

to → person.

(
+,- * *. ! * (! < I
! M ! ! ! 0 " 57 0"!0 0 ! ! ; 7 0 .
! 2"
+,- * *. ! * (! < I
! M ! ! ! 0 " 57 0"!0 0 ! ! ; 7 0 .
! "?7
' $ % 8 8 "0 ! >

features of a situation often appears to have other deep facts built into it, facts that are made obvious only in the technical working out of the formalism.

THE THERMODYNAMIC LIMIT

The existence of a temperature for bulk matter, a measure of the average energy per degree of freedom, is a statement as well that bulk matter might possess what we might call pervasive properties, those size-independent intensive variables. The next task is to demonstrate how and when bulk matter also possesses properties linearly dependent on N, extensive variables. Given that the hydrogen atom, the multielectron atom, and matter composed of lots of atoms are each stable to collapse, matter composed of lots of atoms being bulky (having a size that goes at least as $N^{1/3}$), we need to ask whether some properties of bulk matter can scale with N (rather than being just bounded by N)—what is called the thermodynamic limit—and whether such matter is mechanically stable (so that when it is compressed, its density becomes greater).[35]

We shall first consider an archetype of such proofs, one that does not, on its face, apply to the Coulomb interaction between charged particles. The problem is to divide up bulk matter in such a way that it is seen to be composed of more-or-less independent bulkettes, and so some of its properties will scale with N. Put differently, the task is to build up matter from those bulkettes, combining them to form larger units, ad infinitum. We then consider how that archetype is adjusted so that the Coulomb force, with its infinite range, will still provide for extensive properties ($\approx N$) rather than properties dependent on N^2 (just the number of two-body interparticle interactions, were forces of very long range). In such an account, we find that we are saying something about how we might think of the constitution of matter, in terms of boxes (or balls) geometrically increasing in size and consisting of similar components, and as we get to larger and larger sizes the effect of N on the thermodynamic free energy becomes more and more linear. As we shall see in the next chapter, other accounts of crucial properties of matter suggest rather different constitutions, each of the constitutions to be taken as an aspect of bulk matter, a way we encounter and think about it.

Recall that, through a series of detailed model-dependent calculations, H. A. Kramers (about 1934–1936) came to realize that thermodynamics is a limit science, that thermodynamic properties (intensive variables, such as temperature and pressure, and extensive ones, such as entropy or free energy), and distinctive thermodynamic phenomena such as phase transitions, appear only in the infinite volume limit.[36] While, following Planck and J. W. Gibbs (1902), we might compute the partition function, Q (what Planck called "the sum over states," in German *Zustandsumme*, hence the conventional use of Z for the partition function), a measure of the number of ways of partitioning the energy of the system among its components, what we actually measure and is well defined is the free energy density, f: $\lim_{N \to \infty, N/V \text{constant}} - 1/\beta N \ln Q$ (namely, $\exp - \beta N f = (\exp - \beta f)^{n_B N_B} = Q$, where n_B is the number of atoms in an independent bulkette, and N_B is the number of bulkettes), an infinite volume limit object. (We have seen a similar situation in the Darwin-Fowler derivation in which the temperature $[1/\beta]$ may be in effect "pulled out" as a good intensive variable, when the number of systems considered goes to infinity, the $N_S \to \infty$ limit. Here we might consider each of our bulkettes as a system and so identify N_S with N_B.)

But, having shown that matter is stable to collapse, just which sort of limit are we to take when we are concerned with extensive properties? What is the technical/mathematical construction of "the constitution of matter" (as Lieb and Lebowitz call it) that would enable us to demonstrate the existence of such limits? Put differently, just what is the meaning of a "linear law," as they refer to the thermodynamic limit? (Here I shall be following Lieb and Lebowitz's argument, an epitome and ingenious development of previous work.)[37]

As I have indicated, for these purposes matter's constitution might be archetypally modeled by a hierarchy of bulkettes, perhaps cubes, say each twice the dimensions of its predecessor—at each scale filling all of space; or space is built up out of larger and larger cubes—and so we might go to infinite volume.[38] (See figure 3.2.) Now, no bulkette has more than a limited number of atoms within it. The atoms are in effect impenetrable hard spheres. For recall that we already know, from a proof, that bulk matter is stable, having a lower bound of its energy proportional to N, and bulky, its linear dimensions going as $N^{1/3}$. The atoms' nonzero, finite size (radius a, $\infty > a > 0$), as we shall see, provides for the uniform convergence to the infinite volume limit of the series of approximations for the energy density or for the free energy density;[39] now each component bulkette (or subbulkette) is interacting with its neighbors at best through surface interactions, whose contribution to the energy density or free energy density proves increasingly negligible, so that the series converges, and so that there are extensive properties ($\approx N$) and in general no $N^{1+\epsilon}$, $\epsilon > 0$, properties, which might eventually lead to explosion.

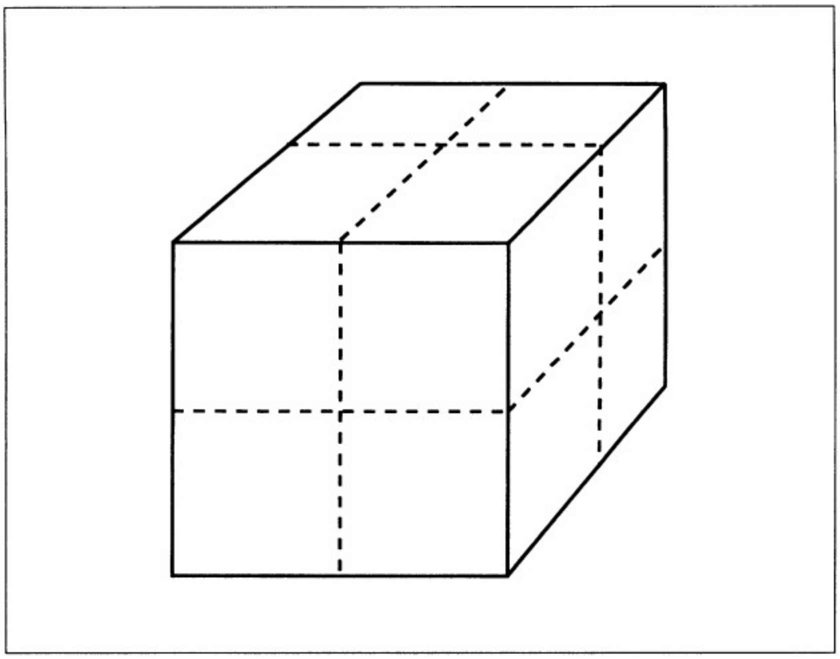

FIGURE 3.2 Hierarchical cubic decomposition of space.

Van Hove, and Yang and Lee, in their respective demonstrations of the existence of a well-defined pressure in ordinary matter (that is, the equivalence of the equation of state of matter in the grand canonical ensemble to that in the canonical ensemble—that one might measure a property of matter, such as pressure, in situ, rather than by isolating the matter ahead of time), and of the impossibility of phase transitions except in the infinite volume limit (if at all), used the apparatus of a decomposition into bulkettes, composed of presumably stable atoms (equal to hard spheres), and interacting through finite-range forces.[40] Again, for the Coulomb interaction ($\approx 1/r$), the finite-range assumption is not at all obvious.

I should note that there are at least two kinds of decomposition in these proofs.[41] The first allows the size of the component cubic bulkettes to grow more slowly than does the volume, so that as one goes to the infinite volume limit the number of components goes to infinity, but more slowly. The second is a geometric series of nested cubes, each twice the size of its predecessor, and therefore eight smaller cubes fit into the next larger cube. As we shall see, when we use balls or spheres as our bulkettes, we have a packing problem and we fill in the interstices (and the interstices of interstices) with increasingly smaller balls (down to a minimum-sized one).

Again, the thermodynamic limit is a large-distance or large-*r* problem, unlike the stability to collapse I discussed earlier, which concerns the small-*r* possibilities. Put simply, if the range of an attractive interaction between particles is very large or "infinite," then we might expect that if we put pieces of matter together there might well be an explosion, the energy per particle rising so rapidly that, rather than a diffusive loss of excess energy through steady state heat conduction, there is shock wave and mechanical explosion. (If the interaction is repulsive, then it would be impossible to produce bulk matter in the first place. And if there were something like critical-mass effects as in radioactive matter, then adding on more matter allows even more matter for mutual interaction, nonlinearly increasing the energy of interaction.) With long-range interactions, bulkettes could not be independent of each other, and putting two in close proximity would produce a bulkette that is qualitatively different than its components. Practically, we could not hope that fluctuations in density would have little effect on thermodynamic properties in equilibrium (rather than, say, leading to explosion, which we surely do worry about for subcritical nuclear assemblies).

We shall want small but still sufficiently large bulkettes to be more or less separated and independent of each other, so that the energy per particle of bulk matter so composed is roughly constant. As we might infer from the Darwin-Fowler derivation, the statistical independence of the bulkettes should allow for sharply defined values of intensive variables and for the volume properties of the bulkettes just adding up; the fact that there is a unique temperature (or saddle point in the Darwin-Fowler derivation) suggests that the convergence to the infinite volume limit would be robust, namely, a uniform convergence, comparatively insensitive to details of the particular actual system or mode of convergence.

Can we so model bulk matter? Each of the proofs I shall discuss depends on finding a canonical way of decomposing matter into bulkettes, that way working out because atoms are of finite size, and so a limited number fit into a finite-sized bulkette—and hence there cannot be, along the way to the infinite volume limit, some bunching up of all the atoms within just one bulkette, or a few. And forces are of finite range, so that bulkette-bulkette interactions are surface- rather than volume-dependent. Such matter will have those extensive properties, its component atoms will not collapse, and the bulk will not explode.

As we approach the thermodynamic limit, it may be useful to recall that proofs of the stability of matter itself, that there is a linear lower bound of the energy, are quantum mechanical, with no need to resort to statistical mechanics as such. (The Thomas-Fermi model employs an average charge density, $n(r)$, but does not use statistical mechanics. There is no partition function. Thomas-Fermi theory is just a quantum mechanical theory.) As we might expect from its name, the thermodynamic limit is in fact statistical, and the partition function is the object of interest; that there is a linear limit for the energy is a statistical mechanical statement. The fluctuations here are not quantum mechanical, but refer to thermal fluctuations in particle number, for example.

The proofs that the energy and that the free energy possess thermodynamic limits are quite similar. In what follows, the free energy plays the central role. In each case, I show a linear N-dependent lower bound (I have already done so for the energy), and then an effectively decreasing sequence of estimates—and so prove the existence of a limit. Note, however, that interactions among the bulkettes contribute differently to the energy (they just add in) than they do to the partition function, $\Sigma \exp - \beta E$ (where they affect the combinatorics as well). In sum, to model and analyze the constitution of matter is to prove that a partition function (and so a free energy density) computed from that model possesses the desirable properties in the thermodynamic limit.

Proving a Thermodynamic Limit: The Free Energy[42]

Technically, one way of saying that matter is composed of bulkettes that are more or less independent is to say that the grand partition function, *GPF*—the partition function of systems when not only the energy but also the number of particles may vary—of some bulk matter (volume V) might be decomposed into the grand partition function of bulkettes of volume V/γ: $GPF(z,V) = \Sigma z^N Q_N(V) = [GPF(z, V/\gamma)]^\gamma$,[43] where the surface contributions to the grand partition function are comparatively small. And, in the infinite volume limit (V to infinity, V/γ constant and suitably large, so that each bulkette has lots of particles—noting that this kind of definition is just what is crucial and an invention), the above relationship holds with no surprises along the way.[44]

The proofs of the theorems of van Hove and of Yang and Lee, concerning the behavior of the partition function and of the grand partition function as the volume goes to infinity, along the way implement these features by several technical moves:[45] If the range of the interparticle forces is fixed and finite (and so volume independent), r_0, then each bulkette can (more or less) only interact with another bulkette along its surface. One sets up r_0-deep "corridors" between the bulkettes, let us say. (See figure 3.3. That one can do this means that there cannot be an explosion, since the interaction among bulkettes goes as $N^{2/3}$ rather than as $N^{1+\epsilon}$, $\epsilon > 0$.) As long as surface to volume ratios are of the usual sort, the contributions of these corridor interactions to the partition function are sufficiently small, in the sense that their contribution to the free energy density or to the grand canonical pressure ($P = (1/\beta V) \ln GPF$), each a logarithm of a partition function then divided by the volume, is negligible. The construction to prove this is by now conventional, but is a crucial invention—namely, how we might break up space into weakly interacting bulkettes: a geometric decomposition into smaller cubes, either an infinite number of such cubes or an infinite nesting of cubes.[46] Given this construction, and a finite range, then the contributions to the partition function (and so to the free energy density or to the grand canonical pressure) from the states in which corridor interactions are allowed stays sufficiently small and is bounded by a function that depends on r_0 alone.

Moreover, one wants to be sure nothing "interesting" happens on the way to infinite volume, that the limits converge nicely enough, hopefully uniformly. What might happen interestingly? There could be some weak $N^{1+\epsilon}$ effect, which when N is large enough starts becoming important. But I have precluded that by the finite-range assumption. Or perhaps all the particles might literally conglomerate into one bulkette, leaving all other bulkettes absent of particles. Now, we might just note that actual matter is bulky, and moreover it does not implode, and so this bunching up is impossible. What turns out to be crucial in the actual proofs, however, is that the particles are presumed to be hard spheres, of nonzero radius, a (namely, they are stable energetically); therefore, only a limited number of them could be packed into any bulkette.

The polynomial in z that is the grand partition function for a bulkette—$GPF = \Sigma z^n Q_n$—has a maximum degree, the maximum number of particles that could fit into a bulkette. The grand partition function for a bulkette or for a finite number of bulkettes is therefore bounded on a compact set for finite z (called the activity or fugacity). (Note that the activity is a measure of the [Gibbs free] energy needed, μ, to add one more particle to the system: $z = \exp \beta\mu$.) This boundedness turns out to be just what is needed to be sure of uniform convergence (namely, bounded derivatives) and so continuity.[47]

In sum, if r_0 were not finite, then interactions among bulkettes would not be bounded. If a were zero, then a single bulkette could have a very large fraction of the particles, and then fluctuation and macroscopic graininess would dominate the bulkiness of matter.

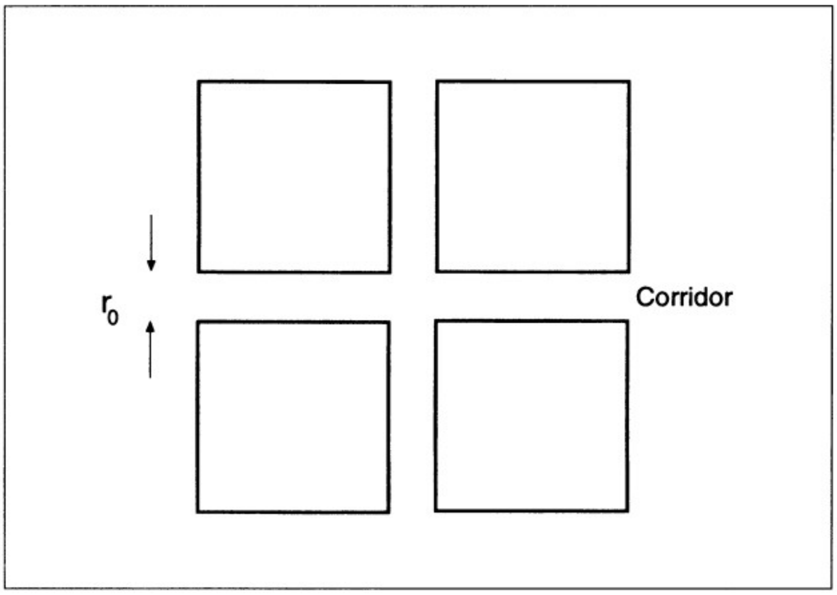

FIGURE 3.3 Corridors between cubes.

As a model for a proof of the thermodynamic limit for Coulomb forces, I have pointed to technical features of some of the proofs of the thermodynamic limit for finite-range forces: the capacity to bound the corridor contributions in a volume-independent way, and the boundedness of the partition function—which are crucial formally and mathematically. And then I suggest that those features reflect deep physical facts about matter, their absence implying a state of matter very different than the one we see everyday.

Still, we do not yet have the wherewithal to provide for a thermodynamic limit. Given the infinite range of the Coulomb force, how are we to apply the technology we have developed for finite-range forces?

Screening

For Coulomb forces, the thermodynamic limit is quite peculiar.[48] For example, were electrons positively charged, there would be no question about the stability of bulk matter: the excess charge would go to the surface, the minimum electrostatic energy would go as $N^2/N^{1/3}$ ($\approx e^2/r$), namely $N^{5/3}$, and there would be no thermodynamic limit: the energy per particle is not at all constant. Moreover, in general, only if the potential falls off as $r^{-(3+\epsilon)}$ would there be such a limit—except here ($1/r$, Coulomb) we are saved by the opposite signs of the charges of the electrons and the protons: "different parts of the system far from each other must be approximately independent, despite the long range nature of the Coulomb force. The fundamental physical, or rather electrostatic, fact that underlies this is *screening* [in effect, there is an r_0 that is finite]; the distribution of the particles must be sufficiently neutral and isotropic locally so that according to Newton's theorem the electric potential far away will be zero. The problem is to express this idea in precise mathematical form."[49]

While "everyone knows" that the Coulomb force is shielded in bulk neutral matter (no net force at a distance, since in net neutral charged matter the distant effects of the positive and the negative charges cancel out), it is not at all apparent how what everyone knows might be used to show that such matter possesses a thermodynamic limit. The crucial insight is that the external effect of a sphere or a ball, isotropically filled with charge, is equivalent to having a point charge at the center ("Newton's [or Gauss's] theorem"). And if that net charge were (almost) zero, there would be (almost) no external force—in effect a finite r_0.[50] The task here is to pack space with such balls, and (as in van Hove and in Yang and Lee) the invention that is crucial is a mode of such packing—so that the spherical bulkettes or balls, at various scales, are of fixed uniform density, something that is automatic when we use cubes but that is not so obvious for spheres, since the leftover unpacked spaces are nontrivial. The development of a workable geometric construction—one that insures that the density of matter is constant, the infinite volume limit in practice—is as crucial here as was the earlier partitioning into cubes.[51] Since any but the smallest ball is filled with smaller balls, the partitioning looks something like figure 3.4. Again, such a partitioning is not only a mathematical device; it also shows how we might think of the constitution of matter, the neutral balls reminding us poignantly about the significance of screening in this realm.

Such a packing of space with neutral balls is not quite a claim that matter is actually composed of neutral balls, that matter is so constituted. It is a very useful assumption, one that is in fact not quite the case. But it is an assumption that can be relaxed enough so that the proof is quite believable as a model of actual matter. (One proves the limit for neutral balls and then shows that any surface charges due to nonneutrality that vanish in the infinite volume limit do not matter.) Namely, for each of the component balls, as long as the excess charge on each "neutral" ball is sufficiently small (its surface-charge density goes to zero) at each scale, and the net total charge on all balls is zero (macroscopic neutrality), all is OK. Hence, the dimensions of the characteristic charge fluctuations in actual matter have to be small compared to the size of the smallest ball considered. It can also be shown that these nonneutral configurations do not contribute much to

.3 6! U 8

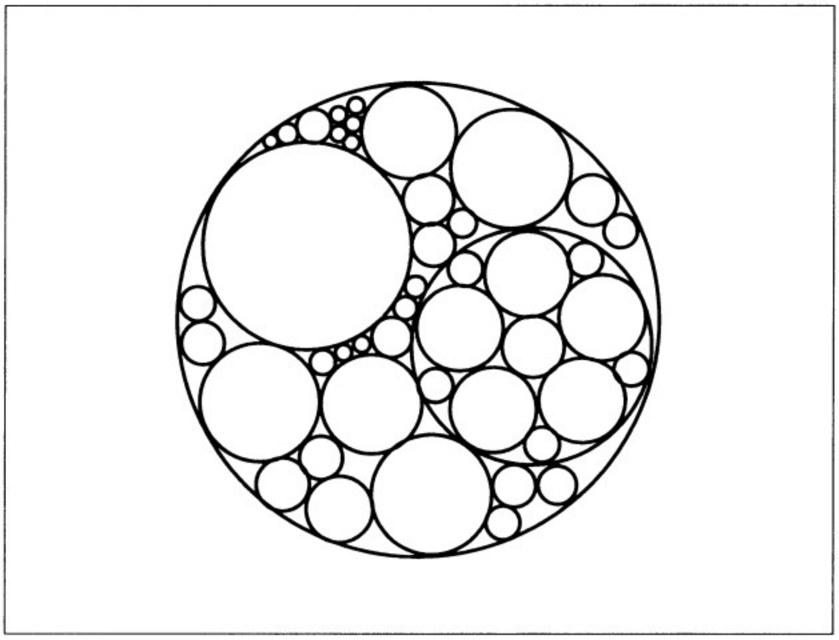

FIGURE 3.4 Filling space with balls.

the grand canonical pressure, the nonneutral parts being much like the corridor contributions in other constructions (as long as the surface-charge density goes to zero as the volume goes to infinity). It is a *façon de parler* to say that matter is constituted of neutral balls, justified by these physical facts.

The neutral-ball construction leads to a suitably convergent series of approximations to the partition function, and so it provides a constitution of matter. More precisely, first, it leads to a recursive sequence of estimates for the free energy density:[52] the free energy density of a ball at size l, f_l, is less than or equal to the free energy density computed by packing the neutral l-sized ball with smaller neutral balls (each of whose free energy density has also been bounded from above by its component balls' free energies), and then summing their contribution to the free energy: namely, $f_l \leq \Sigma_{i=0,\ldots,l-1} f_i \upsilon_i$ (where υ_i is the fraction of volume of the l-sized ball that is covered by the i-sized balls). Now, let us allow for almost neutral balls.

Second, in order to show that the residual Coulomb interaction energy of these almost neutral balls is not out of bounds and possesses a limit, one needs to pack space with l-sized neutral balls (each packed with $(l-1)$–sized and smaller almost neutral balls). What makes the proof go through is that, in computing a bound on the partition function for finite N and V, a sum or average for all configurations, one need consider only the average interaction among all the bulkettes, that is, "the *average* interdomain Coulomb energy in a canonical ensemble in which the Coulomb interaction is present in each subdomain but the. . . domains are independent of each other."[53] And it turns out that, for all overall neutral volumes, that average interdomain energy is zero, by the Newton-Gauss law. Even if fluctuations do give various balls small amounts of nonneutrality, that does not matter, since at each level one is demanding overall neutrality.[54] And when one goes to the next larger size ball, one can relax the neutrality assumption for the now smaller ball.

Crucially, the bulkettes must be statistically independent of each other, again what we would expect of bulk matter that is stable and in equilibrium.[55] Technically, that independence might be understood as an averaging that leads to $1/\sqrt{N}$ effects, damping charge fluctuations. More formally, $Q \geq (\exp - \beta U)\Pi_j Q_j$: the partition function for a finite system is greater than or equal to just $\exp - \beta$(Average Interaction Energy, U), representing the interactions among the bulkettes, times the product of the partition functions of each of the bulkettes that compose it.[56] (Note that, since the average energy is zero, the exponential equals one.) This inequality then defines a sequence of in-effect decreasing estimates for the free energy density as the volume gets larger ($f \leq \Sigma f_j$), and, given the stability of bulk matter (which leads to a bound on the partition function and so to a lower bound on the free energy, for which see the section below, "Limits Are Physical"), we are led to the proof of the existence of an infinite volume limit, the thermodynamic limit we are seeking.[57]

Now, say the actual matter itself were not electrically neutral. Would it still exhibit a thermodynamic limit? As I have suggested already, if the excess charge is small, it should have little effect. Moreover, the long-range character of the Coulomb force demands that, if there is to be a conventional thermodynamic limit, then the system must be effectively neutral—otherwise shape and size effects become important.

Technically, taking into account the amount of excess charge (over neutrality) in taking the thermodynamic limit: if the surface charge density goes to zero, then the limit is the same as for the neutral system; if the surface charge density goes to infinity, then there is no limit (as one might expect); and if it goes to a finite value, to the neutral free energy is added an electrostatic surface energy.[58]

In sum, the thermodynamic limit separates out N-dependence through the use of independent bulkettes. The proofs all depend on the hard-core or nonzero size of the atoms and the finite (that is, noninfinite) range of the interparticle forces. And screening does the finite ranging work for the Coulomb force, once we pack space with effectively neutral balls.

THERMODYNAMIC STABILITY AND MATHEMATICAL PHYSICS AS EMPIRICAL

These mathematical proofs are in effect philosophical analyses of thermodynamics, showing in various contexts the meaning of the infinite volume limit through the demands of technical constructions of the constitution of matter. The technical and mathematical features of the proofs depend upon three contingent physical facts, contingent in the sense that they need not be the case, but surely are the case in our everyday world: (1) The energy is bounded from below, namely, matter is stable, whether that is presumed by some hard-sphere repulsion or derived from the Pauli exclusion principle. The free energy, too, is bounded from below. (2) Bulkette-bulkette interaction terms are small; namely, they are surface terms, their contribution to the energy or the free energy densities being negligible. This smallness depends on the effective finiteness of the range of the interaction. And so space might be usefully partitioned into independent bulkettes—the remaining task being to find a partitioning that in some sense converges, that nicely reaches the infinite volume limit. (3) Hence, nothing interesting will happen, such as all the particles going to one bulkette (granularity), when one allows for larger- (or smaller-) scale bulkettes.[59] The latter fact is, essentially, the uniformity of convergence. Technically, at each stage in the approximation, on the way to the infinite volume limit, one can show that the free energy per unit volume, f_l, l being the scale, converges nicely enough for there to be uniform convergence: $|f_{2l}-f_l| < c/l$.[60]

The convergence to a limit (independent of the mode of sequencing or the shape of the volumes) can be shown to have rather deep and generic physical consequences: matter becomes denser when compressed, and its temperature is raised when heat is added to it (it is "thermodynamically stable"); and in general (except perhaps at phase transition), measurements of properties as a function of density or temperature for large samples will be a good approximation to the infinite volume limit.

For, even if there is a lower bound on the energy, so that matter is energetically stable, and even if there is a thermodynamic limit (asymptotic $1/N$ limit), we have yet to show there is such thermodynamic stability: that the system will not exhibit a negative compressibility, Δdensity$/\Delta$pressure < 0, so that pressing upon it would decrease its

Google
steniomusic.com

ID de cliente
C03VWgYKO

09 Jun 2021 → Drive 70,98 GB
Drive Compartilhados 128,32 TB

Books1-016-2021/20210806/
The Geometry of the Universe/
The Geometry of the Universe - Colin Rourke.pdf

Pág. 73 (88)
5.4 Conservation of angular momentum
Figure 5.3 - Proof of conservation of angular momentum

Books1-016-2021/20210806/
Invisible-Universe/
The Invisible Universe Dark Matter ... pdf

Pág. 138 (151) 6.7 The Higgs Mechanism

%
! *+,-* * *. !* % 8 6 0 . 8 ; 7
, B 2"
' $ %(7! ! !

made is the *metrical embodiment of the hypothesised inertial drag field*. It is not hard to see that the inertial frame at a point rotates about a line parallel to the z-axis with angular velocity the value of ω at that point. This is clear if ω is constant and in general, provided ω is continuous, it follows from the locality of inertial frames. So to fit with inertial drag as formulated in (5.1) it is necessary to set $\omega = A/(r+K)$ (at least in the (x,y)-plane). However it is easy to work with a general function ω and specialise when needed. The orbits of particles moving on geodesics in the equatorial plane will now be investigatied and, provided ω decreases like A/r as $r \to \infty$, the orbits will be found to fit observed rotation curves.

The reader might find Figure 5.2 helpful for visualising geodesics in the inertial drag metric and understanding the inertial drag effects. It shows geodesics on a typical cylinder $r = z =$const.

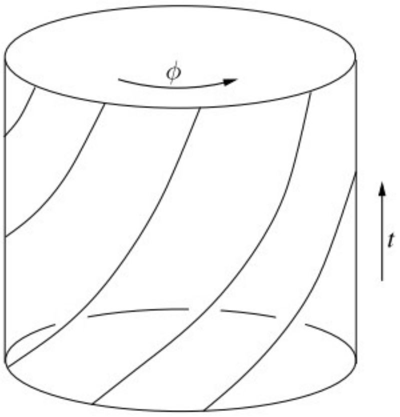

Figure 5.2: Geodesics on the cylinder $r = z =$const in the inertial drag metric

5.4 Conservation of angular momentum

Here now is the analytic derivation of the rotation curve (the relation between v and r) for the orbit of a particle in the equatorial plane moving with total velocity $\mathbf{v}$, which has tangential component v (perpendicular to the line through the origin). Recall from the discussion in Section 5.2 above that there are two opposing effects at work: the slingshot effect, which tends to increase v with r and conservation of angular momentum which tends to decrease it. These two effects are calculated together. The proof works in any inertial drag metric, where the particle moves along a geodesic. (The special case of flat Minkowski space with inertial drag effects was motivated in Section 5.2.)

The derivation starts with a proof of conservation of angular momentum, which is a property of any system with a central force (or space-time geometry which simulates a central force). It is not restricted to Newtonian physics. The proof is adapted from Newton's proof of the equal area law for planetary orbits (which is exactly the same as conservation of angular momentum, see Figure 1.4). For the time being ignore ω (or set it equal to zero).

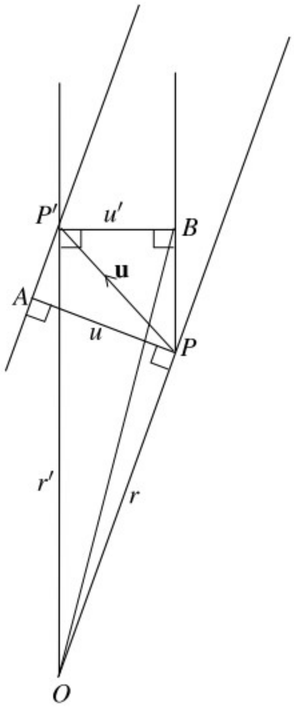

Figure 5.3: Proof of conservation of angular momentum

The idea is to replace the central force by a series of central impulses at equally spaced (small) intervals of time. Consider Figure 5.3. At a particular time the particle (of small unit mass) is at P and has just received a central impulse resulting in velocity $\mathbf{u}$. Its tangential velocity at P is $u = |AP|$. One small interval of time later the particle is at P' and receives another central impulse (along the line OP') which does not change its tangential velocity $u' = |P'B|$. But the triangle OPP' can be regarded as having base $r = |OP|$ and height u or base $r' = |OP'|$ and height u' hence

$$ur = u'r' \tag{5.4}$$

in other words the angular momentum at P is the same as that at P'.

To obtain the result for an arbitrary continuous central force, take the limit of a sequence of central impulses. Note the proof does not use any property of the central force other than that it acts towards the centre. Nor does it assume that r represents a genuine distance in the metric under consideration. All that is needed is that Euclidean geometry correctly describes the relationship between r and distances perpendicular to radii near $P$and P' which is precisely how r was chosen.

5.5 The fundamental relation

Now reinstate ω. Note that "force" in the model is a property of local space-time geometry. In the case that ω is constant, the inertial frame (rotating with ω) is the same as the unrotated case and in this frame the force is central. Therefore by locality it is central in the general case in the inertial frame. Therefore the proof just given makes sense in the inertial frame at P', in other words rotating with angular velocity $\omega' = \omega(P')$, though, as will be seen, in the limit the same result is obtained if it is assumed that the frame is rotating with angular velocity $\omega(P)$. To find the required relationship between v and r write v for the full tangential velocity at P and v' at P'. Since the frame is rotating at ω', $v = u + \omega' r$ and $v' = u' + \omega' r'$. Write $v' = v + \delta v$, $u' = u + \delta u$, $r' = r + \delta r$ and $\omega' = \omega + \delta\omega$.

Since $ur = u'r'$ (Equation 5.4), substituting for u', v' and simplifying gives

$$u\,\delta r + r\,\delta u = 0\,. \tag{5.5}$$

But

$$\delta u = u' - u = v' - \omega' r' - (v - \omega' r) = v' - v - \omega'(r' - r) = \delta v - \omega'\,\delta r$$

and substituting for $u, \delta u$ in (5.5) gives

$$(v - \omega' r)\,\delta r + r(\delta v - \omega'\,\delta r) = 0$$

which gives

$$r\,\delta v = 2r\omega'\,\delta r - v.$$

It is now possible to replace ω' by ω to first order (as forecast) and going to the limit yields the *fundamental relation* between v and r:

$$\boxed{\frac{dv}{dr} = 2\omega - \frac{v}{r}} \tag{5.6}$$

The fundamental relation can be understand intuitively as follows. The slingshot effect intuitively produces an acceleration $dv/dr = \omega$. On the other hand $v_{\text{inert}} = v - \omega r$ is the "inertial" tangential velocity (corrected for rotation of the local inertial frame) and therefore conservation of angular momentum produces a deceleration in v of v_{inert}/r or an acceleration $dv/dr = \omega - v/r$. Adding the two effects gives the relation.

5.6 Solving to find rotation curves

Given ω as a function of r, (5.6) can be solved to give v as a function of r. Rewrite it as

$$r\,\frac{dv}{dr} + v = 2\omega r\,.$$

The LHS is $d/dr\,(rv)$ and the general solution is

$$v = \frac{1}{r}\Big(\int 2\omega r\,dr + \text{const}\Big). \tag{5.7}$$

It is now clear that any prescribed differentiable rotation curve can be obtained by making a suitable choice of continuous ω.

Of interest here are solutions which, like observed rotation curves, are asymptotically constant and inspecting (5.7) this happens precisely when $\int 2\omega r\,dr$ is asymptotically equal to Qr for some Q and this happens precisely when 2ω is asymptotically equal to Q/r. This proves the following result.

Theorem *The equatorial geodesics in the inertial drag metric* (5.3) *have tangential velocity asymptotically equal to constant* Q *if and only if* ω *is asymptotically equal to* A/r *where* $Q = 2A$.

5.7 The basic model

Now specialise to the case $\omega = A/(r+K)$ which gives the value of inertial drag formulated in (5.1).

From (5.7)

$$v = \frac{1}{r}\left(\int \frac{2Ar}{r+K}\,dr + C\right) = \frac{2A}{r}\left(\int 1 - \frac{K}{r+K}\,dr\right) + \frac{C}{r}$$

$$= 2A - \frac{2AK}{r}\log\left(\frac{r}{K}+1\right) + \frac{C}{r} \tag{5.8}$$

where C is a constant depending on initial conditions. For a particle ejected from the centre with $v = r\omega_0$ for r small, $C = 0$, and for general initial conditions there is a contribution C/r to v which does not affect the behaviour for large r. For the solution with $C = 0$ there are two asymptotes. For r small, $v \approx r\omega_0$ and the curve is roughly a straight line through the origin. And for r large the curve approaches the horizontal line $v = 2A$. A rough graph is given in Figure 5.4 (left) where $K = A = 1$. The similarity with a typical rotation curve, Figure 5.4 (right), is obvious. Note that no attempt has been made here to use meaningful units on the left. See Figure 5.6 below for curves from the model using sensible units.

There are other shapes for rotation curves; see [96] for a survey. All agree on the characteristic horizontal straight line. Figure 5.5 is reproduced from [96] and gives a good selection of rotation curves superimposed. In Figure 5.6 is a selection of rotation curves again superimposed, sketched using Mathematica[1] and the model given here. The different curves correspond to choices of A, K and C. The similarity is again obvious. The units used differ. In the model given here natural units are used so that a velocity of .001 is 300km/s and a distance of 45,000 is 15Kpc approx.

---

[1] The notebook `Rots.nb` used to draw this figure can be collected from [4] and the values of the parameters used read off.

3 *+,-* * *. ! * % 8 &7 3 ; 7 :! B ! F
:! B 5 0F ! " . ! " 5 " ; 7 2"
' $ % 9 8 < !

6.7 The Higgs Mechanism

The particles of the standard model can be divided into two groups. On the one hand, there are the particles that measure the interactions, namely photons, W^+, W^-, Z^0, and gluons, and on the other, the particles of matter, namely quarks and leptons. The particles that have a mass are W^+, W^-, Z^0, quarks, and leptons.

Unfortunately, a problem that plagued the theory was the fact that the particles of the model had to have zero mass for reasons of symmetry. Some mechanism was necessary to break the symmetry, but in the symmetry breaking, the well-known massless Nambu–Goldstone bosons appeared. A sort of trick was needed to make them disappear or hide. This "trick" is the Higgs mechanism, in which Goldstone bosons are "eaten" by gauge bosons, which become massive. How does this happen?

As we know, particles have a spin, which is often described in classical mechanics and in early quantum theory as a rotation.[7] If we consider an electron, its spin is ½ and in measurements we can find only two *spin states*, or *polarization states*, or *degrees of freedom*, which represent the ways in which a field can vibrate: ½ and −½. Each polarization state can be considered as a different particle or a different degree of freedom. *A fundamental point to remember is that massive particles have three degrees of freedom.*

For a particle with spin 1, we can have two possibilities: if the particle is massive (e.g., in the case of W and Z bosons), we could find the values −1, 0, 1. The particle has three degrees of freedom. This means that the fields of massive particles can have an up-down vibration (and vice versa), a left-right vibration (and vice versa), and forward-backward vibration (and vice versa). If the particle has no mass, as in the case of photons, we would find –1 and +1, that is, only two degrees of freedom.[8] So the fields of massless particles can vibrate only in two ways: up-down (and vice versa) and from right to left (and vice versa). These oscillations are shown in Appendix C (Fig. C1).

---

[7] As already reported, the notion of spin as rotation around an axis is incorrect. This quantity was not foreseen by non-relativistic quantum mechanics; it had been introduced as a result of several experiments. It is automatically obtained in the relativistic version of quantum mechanics, by the Dirac equation, and does not correspond to a rotation around an axis.

To summarize, a scalar field, with spin 0, has only one degree of freedom. Vector fields with spin 1 have two degrees of freedom if they are massless and have excitations perpendicular to the direction of propagation (1, –1). If they have mass, they have three degrees of freedom, with two perpendicular excitations (1, –1) and one longitudinal (0) (see Appendix C).

So how do the particles W^+, W^- and Z^0 acquire their mass?

Before the symmetry breaking, the electroweak interaction was mediated by four fields: the fields B, W_1, W_2, and W_3 (Fig. 6.9) and the corresponding bosons were all massless.

We can consider them as the parents of the fields of the photon γ and those of W^+, W^-, and Z^0. Consequently, they have only two degrees of freedom, excitations perpendicular to the direction of propagation, and zip at the speed of light. To make them massive, one needs a third degree of freedom, the longitudinal one (forward-backward oscillation). We need three scalar particles that, combining with W_1, W_2, W_3, provide each with the missing longitudinal degree of freedom. These scalar particles, the *Goldstone bosons*, are produced at the end of the electroweak era, when the universe had 10^{-12} s and temperature was 10^{15} K in the electroweak symmetry breaking phase mentioned in Chapter 2. As described in Appendix C, the Higgs potential underwent a transformation. The field "slipped" from the false vacuum to the true vacuum, like a ball that descends from the top of a hill to the valley (see Fig. C3).

Fig. 6.9. The electroweak interaction mediators.

---

[8] It should be noted that if 0 is found, this does not mean that the particle has no spin.

In this event, four scalar particles (see Fig. 6.10) were generated which we refer to as H_1, H_2, H_3, and H, or also as H^0, H^+, H^-, and h three of which (H^0, H^+, H^-) are Goldstone bosons, eager to help W_1, W_2, W_3 to become massive. H_1, H_2, H_3, alias H^0, H^+, H^-, provide the degree of longitudinal freedom needed by W_1, W_2, W_3 to become massive.

After W_1, W_2, W_3, absorb H_1, H_2, H_3, that is, the missing degree of freedom, immediately become massive. In Figs. 6.9–6.13, we show some cartoons describing how the bosons W_1, W_2, W_3, by eating H^0, H^+, H^-, transform into W^+, W^-, and Z^0. W_1 and W_2 combine and, eating H^+, give rise to W^+ (Fig. 6.11).

Similarly, W_1 and W_2 combine and eating H^-, give rise to W^- (Fig. 6.12).

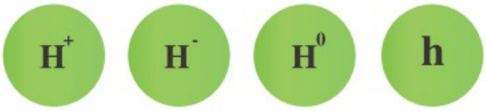

Fig. 6.10. The three massless Goldstone bosons H^+, H^-, H^0, and the (massive) Higgs boson, h.

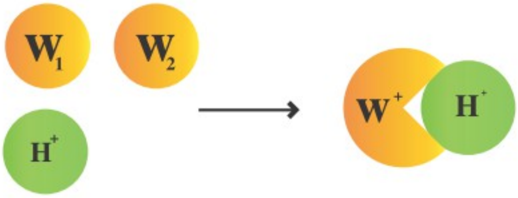

Fig. 6.11. Generation of W^+ from the combination of W_1 and W_2 and eating of H^+. Credit: Modification of work by Flip Tanedo, Quantum Diaries.

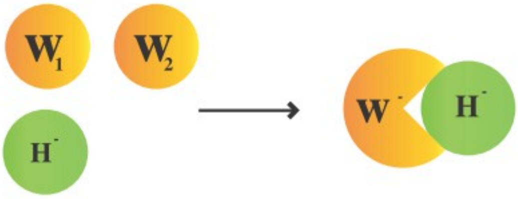

Fig. 6.12. Generation of W^- from the combination of W_1 and W_2 and eating of H^-. Credit: Modification of work by Flip Tanedo, Quantum Diaries.

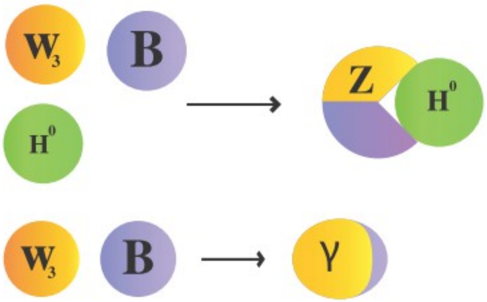

Fig. 6.13. Top: Generation of Z^0 from the combination of W_3 and B and eating of H^0. Bottom: Generation of the photon, γ, from the combination of W_3 and B. Credit: Modification of work by Flip Tanedo, Quantum diaries.

The boson Z is obtained when W_3 and B combine and eat H^0. W_3 and B also combine (Fig. 6.13). No Goldstone boson is left, and then the photon (combination of W_3 and B) remains massless, and h, which is *the Higgs boson* discovered in 2012, remains free.

In Appendix C, a more precise description of the Higgs mechanism is given.

In the case of fermions (quarks and leptons), mass acquisition occurs in a different way, with the so-called extension of the Higgs mechanism to *Yukawa's interaction*, an interaction that couples the fermion field with that of the Higgs. Intuitively, in their motion these particles "feel" the presence of the Higgs field, which fills the whole space, and begin to interact with it. When fermions move in space, they interact with the Higgs field. Using a not very correct, but effective, analogy, the field of Higgs slows them down like honey does with an object that moves in it. Entangled in the Higgs field, the particles find it more difficult to move, as if they had mass and their speed becomes less than that of light. Particles such as electrons assume mass when the Higgs boson "intertwines" the two polarizations of the electron. The operation requires energy, supplied by the Higgs field, which is perceived as the resting mass of the particle. The mass grows as the intensity of the interaction grows.

If we could disable the Higgs field, we would go back to the initial state with four massless bosons. The quarks would lose their mass, the two electron components with different polarization would decouple and lose their mass, and the atoms and matter would be annihilated.

A legitimate question is, why the other fields cannot do what the Higgs field does? What distinguishes the Higgs field from other fields, for example from the electromagnetic one? This difference was described in more detail in section 6.2 when we talked about fields. Unlike the other fields, the Higgs field is always different from zero in its state of minimum activity, $\phi = 0$, as seen in Fig. 6.4 (right panel), and Fig. 6.14. It even occurs that the Higgs field has more energy when it is null, that is, when $\phi = 0$, than when it is not null. Contrary to the usual fields, if we want the Higgs field not to invade space with its bosons, we must supply energy. This is the difference that distinguishes it from other fields. Particles, sensitive to it, always have a certain probability of interacting with it and become massive.

The intuitive description of the Higgs mechanism is due to David Miller, a physicist from University College London. He won the prize offered by Minister William Waldegrave (a bottle of champagne) to the physicist who had managed to explain in an intuitive and exhaustive way how the mechanism works and what it was used to find out.

Miller imagined a salon full of people evenly distributed in space, chatting among themselves as a VIP arrives. As the VIP

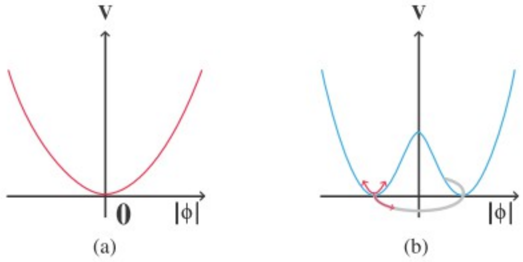

Fig. 6.14. The potential V of a field like the electromagnetic one (left) and that of the Higgs field (right). When the field "falls down" from the hill to the minimum, it can have an up and down oscillation generating the boson H and, on the circumference of the minimum, generating the bosons H_1, H_2, H_3.

moves, he attracts people and a crowd forms. The crowding produces resistance to the movement of the VIP, and it looks as if the VIP is acquiring mass. A particle moving in the Higgs field faces a similar situation.

The concept of the Higgs boson is explained by imagining that someone is circulating a voice in the lounge. Those who listened to it first will report it to the neighbors. The room will be crossed by a wave of "crowds of people" that, like the VIP, "gain mass." The Higgs boson would resemble this crowding.

It is often said that the Higgs boson gives mass to objects. This is not exactly correct. The Higgs mechanism provides mass to only the elementary particles. An important point to note is that the mass of any object, such as a table, is due in minimal part to the Higgs field. Everyday objects are made up of protons and neutrons.

Each proton is made up of two up-quarks, with mass 2.3 MeV, and one down-quark, with mass 4.8 Mev. So if the mass of a proton were all due to quarks, it would have to weigh 2.3 + 2.3 + 4.8 = 9.4 MeV. In reality, the proton has a much larger mass, 938 MeV. This means that the Higgs field that provides mass to the quarks contributes only about 1% to the mass of the proton. The mass of the proton is basically due to the fluctuations of the vacuum (Fig. 6.15). The quarks that

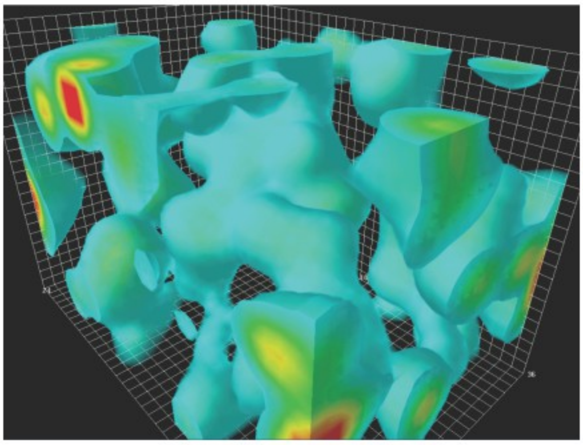

Fig. 6.15. Fluctuations of the vacuum in a proton. Credit: Wikipedia, Derek B. Leinweber.

make up the proton are confined in the proton, within which the quantum vacuum is subject to fluctuations that create and destroy quarks and gluons continuously and in a very short time. The interaction of the quarks that make up the proton with the fluctuations of the vacuum generates energy (see the Casimir, Lamb, and Schwinger effect in Chapter 6).

If the Higgs field did not exist, the quarks and leptons would have no mass and would move in space at the speed of light. The consequence is that protons, neutrons, nuclei, and atoms would not form the universe and, finally, us.

So far, we have talked about the theory of how elementary particles gain mass. Is there evidence that this is really the case? The answer is yes. A large part (3/4) of the Higgs mechanism was discovered in 1983 at CERN by Carlo Rubbia and Simon van der Meer, with the discovery of the W^+, W^-, and Z^0 bosons. The cherry on the cake, or the Higgs boson, was discovered on July 4, 2012, thus completing the construction of the standard model.

6.8 The Dynamics of Colors and the Atomic Nucleus

Meanwhile, the study of cosmic rays had allowed to identify several particles. In 1936 Carl Anderson, the positron discoverer, together with Seth Neddermeyer, discovered a negatively charged particle, like the electron, but about 200 times heavier. It was called *μ meson*.

A meson is a particle made up, as we will see, of two other elementary particles: a quark and an antiquark. These particles "feel" the strong nuclear force.

The reception of the new particle was not the warmest, since it was not clear why there should have been a heavier cousin than the electron. Physicists were so stunned that Isaac Rabi learned of the discovery and exclaimed, "And who ordered it?"

Moreover, physicists had not understood that in reality the particle was not a meson but precisely a cousin of the electron, about 200 times heavier. It belongs, together with neutrinos, to the family of leptons that "feel" the weak nuclear force. In the 1970s, history repeated itself and another cousin of the electron and muon was discovered, the *tau*, or *tau meson*, having a mass approximately 3,500 times that of the electron. Only in 1947 were the mesons predicted by Yukawa really discovered: the *pion*, or π *meson*, and the *kaon*, or *k meson*. To study the deep structure of matter, in the 1950s, particle accelerators began to be built. Accelerators allow to produce collisions with energies much greater than those of the particles involved in the collision, thus producing particles of greater mass. According to the equation $E = mc^2$, for a particle to be generated it is necessary that the energy be greater than its mass at rest. As the accelerator energies increased, the number of particles produced became so large that there was talk of a *zoo of particles*.

Murray Gell-Mann and Yuval Ne'eman showed that the particles could be arranged in a two-dimensional scheme that indicated a wider but also more approximate symmetry, called "the way of the octet," since they remembered the geometric structure of the octets.

The new particles belonged to the hadron family. Physicists began to think that these particles were not really elementary, and in 1964 Gell-Mann and Zweig proposed the *quark model*, confirmed in 1968 with experiments at the Stanford Linear Accelerator (SLAC), distinguishing the properties of the quarks with the terms *up* (u), *down* (d), and *strange* (s). The model described the proton as consisting of the three *uud* quarks and the neutron constituted by the *udd* combination. Particles such as pions are composed of a quark and an antiquark. However, the model had a flaw. The particle (baryon) Δ^{++} consisted of three up-quarks, and this violated the Pauli exclusion principle. The solution proposed by Han, Nambu, and Greenberg was to assume that there is a new quantum number for quarks called color. It was later recognized that it was a sort of charge (similar to the notion of electric charge, but at the same time with various differences with it), called *color*. So each of the quarks had three different colors, and suddenly the number of quarks tripled. Han and Nambu suggested that the strong interaction was mediated by eight gauge bosons, called *gluons*. These are the ideas behind the strong interaction theory, *quantum chromodynamics*, which takes its name from color, in Greek *chróma*. Gross, Wilczek, and Politzer showed that the color force, unlike electromagnetism, weakened with distance, an effect called *asymptotic freedom*. In addition, the color strength increases when quarks are trying to move away, resulting in the inability to see free quarks. This phenomenon is called *confinement*. Today, in addition to the up, down, and strange quarks, we know that there are three other quarks: *charm* (c), *bottom* (b) (or beauty), and *top* (t) (or truth). The quarks are combined into three pairs — (u, d), (c, s), (t, b) — of increasing mass, and each with the same pair of charges, 2/3 e^-, –1/3 e^-. The quarks therefore have three generations of particles and similarly the leptons: (e^-, ν_e), (μ, ν_μ), (τ, ν_τ). Quarks were discovered between the '70s and the '90s. The last and heaviest of the quarks, the top, was observed at Fermilab in 1995.

6.9 Summary: Particles and Interactions of the Standard Model

6.9.1 *Particles*

The fundamental building blocks that make up the world are the fields. However, their oscillations manifest themselves as particles. The standard model contains two large families of particles: fermions and bosons.

Fermions take their name from Enrico Fermi and obey the *Pauli exclusion principle*. They follow the *Fermi–Dirac statistic* and therefore have a half-integer spin (1/2 $\hbar$, 3/2 $\hbar$, 5/2 $\hbar$ …) and are the constituents of matter. There are two types of fermions: the quarks, constituting the protons and neutrons, and the leptons, constituting electrons, muons, tauons, neutrinos, and their respective antiparticles.

The *bosons* take their name from Satyendranath Bose, they follow the *Bose-Einstein statistic* and from the *spin-statistic theorem,* they have integer spin (0 $\hbar$, 1 $\hbar$, 2 $\hbar$…).

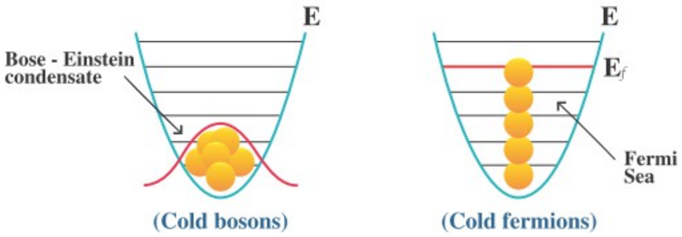

Fig. 6.16. Bosons and fermions.
Source: Modification of work by Fred Bellaiche.

Fermions, following the Pauli exclusion principle, occupy much more space than bosons, as shown in Fig. 6.16. On the left side is represented a gas of bosons and on the right one of fermions at temperatures close to absolute zero (–273.16 degrees centigrade). The boson gas collapses to form a *Bose–Einstein condensate*. Fermions cannot place themselves in a state that takes up little space. They arrange to occupy energetic states of increasing energy up to a state of maximum energy called *Fermi energy*. This is why fermions are the constituents of matter, while bosons are the vectors of interactions

Bosons are distinguished in *gauge bosons* or vectors that mediate the forces and in *mesons*, stable particles consisting of a quark and its antiparticle, antiquark. Particles composed of a large number of particles, such as protons and neutrons, and atomic nuclei can behave like bosons or fermions depending on the total spin. Another classification of the particles according to the forces to which they are subjected or their mass is as follows:

1. *Hadrons* (from the Greek "strong") are particles subject to strong nuclear force and can be divided into

    - *Barions* (from the Greek "heavy"): They are fermions. They consist of nuclear particles, the baryons, such as protons, neutrons, the particle *Λ*, *Sigma*, Δ, Ξ, Ω (and their respective antiparticles), and heavier and non-nuclear particles, *hyperons*. Baryons are not elementary particles, but made up of *quarks*. There are six types of quarks: *up* (u), *down* (d), *strange* (s),

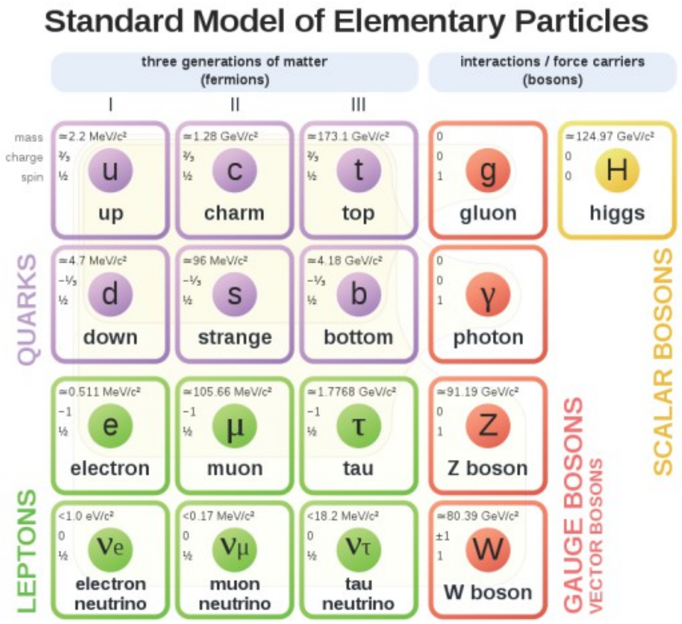
Standard Model of Elementary Particles
three generations of matter (fermions)
interactions / force carriers (bosons)
I
II
III
mass
charge
spin
≃2.2 MeV/c²
⅔
½
u
up
≃1.28 GeV/c²
⅔
½
c
charm
≃173.1 GeV/c²
⅔
½
t
top
0
0
1
g
gluon
≃124.97 GeV/c²
0
0
H
higgs
≃4.7 MeV/c²
−⅓
½
d
down
≃96 MeV/c²
−⅓
½
s
strange
≃4.18 GeV/c²
−⅓
½
b
bottom
0
0
1
γ
photon
≃0.511 MeV/c²
−1
½
e
electron
≃105.66 MeV/c²
−1
½
μ
muon
≃1.7768 GeV/c²
−1
½
τ
tau
≃91.19 GeV/c²
0
1
Z
Z boson
<1.0 eV/c²
0
½
νe
electron neutrino
<0.17 MeV/c²
0
½
νμ
muon neutrino
<18.2 MeV/c²
0
½
ντ
tau neutrino
≃80.39 GeV/c²
±1
1
W
W boson
QUARKS
LEPTONS
GAUGE BOSONS
VECTOR BOSONS
SCALAR BOSONS

Fig. 6.17. Particles and interactions in the standard model. Credit: Wikipedia.

charm (c), *bottom* (b) (or *beauty*), and *top* (t) (or *truth*). The quarks are combined in three pairs — (u, d), (c, s), (t, b) — of increasing mass, and each with the same pair of charges 2/3 e^-, −1/3 e^- (where e^- is the electron charge). Each quark has *three colors*, so we have 18 quarks. Protons consist of two up- (u) and one down-quark (d): (uud). Neutrons from one up- and two down-quarks (udd). Protons have a mass of about 938 MeV and neutrons are slightly more massive, 939 MeV, and have a size of the order of 10^{-15} m. Outside the nucleus, neutrons are unstable and have an average life of approximately 15 minutes. Protons are almost immortal particles, given that their average life is greater than 10^{34} years, much greater than the age of the universe. The average life of the proton is important because some Grand Unification Theories (GUTs), theories that unify strong, weak, and electromagnetic nuclear forces, require that the proton decay. The baryons are assigned the so-called *baryonic number*, which is $B = +1$ for baryons and $B = -1$ for anti-baryons. For the other particles, mesons and fermions, $B = 0$. The baryon number is an almost always preserved quantity, that is, the baryon number before a reaction is the same as the one after.

- *Mesons* (from the Greek "middle"): They are bosons subject to a strong nuclear force with intermediate masses between the baryons and the leptons. They consist of a quark and an antiquark.
- *Leptons* (from the Greek "light"). They have no strong interaction, are elementary particles, and are fermions with spin 1/2. Examples of this family are charged particles such as electrons with a mass of about 10^{-30} kg (0.511 MeV) and dimensions less than 10^{-22} m, *muons* with a mass 207 times that of the electron, *tauons* with a mass about 3500 times larger than that of electrons, and their respective antiparticles. The family also contains neutral particles, the *neutrinos*. Neutrinos have a very small mass, between 10,000 and 1 million times smaller than that of the electron. They come in three flavors: *electron*, *muon*, and *tau neutrino* and can oscillate, that is, transform into each other (*neutrino oscillations*). Like quarks, leptons have three generations of particles $(e, \nu e)$, $(\mu, \nu\mu)$, (τ, ν_{τ}). Like baryons, leptons are assigned a leptonic number: $L = +1$ for particles, $L = -1$ for antiparticles, and $L = 0$ for hadrons.

Ultimately, we have three families of leptons, six particles, and three of baryons, another six particles. To each of the particles corresponds an antiparticle of equal mass and opposite charge.

A very interesting point is that matter is made of fermions and only of particles of the first family, namely quarks u and d and electrons. A legitimate question at this point is, why are there two other families of particles if they do not constitute matter, and why is there a mass hierarchy? This and many other questions, as we will see at the end of this section, show the limitations of the *standard model*.

Another discrepancy of the model is that neutrinos have no mass, while the phenomenon of neutrino oscillation, that is, the transformation of one of the three types of neutrinos into another type, implies that they have mass, however small. One of the possible solutions, better described in Appendix D, is that there is another neutrino, the *sterile neutrino*, which interacts with other particles only with gravity and can interact with other neutrinos also with the mechanism of neutrino oscillations. The *seesaw mechanism* (Fig. 6.18) ensures that the mass of the neutrino is inversely proportional to that of the sterile neutrino. If this latter has a mass of the order of 10^{16} GeV, the neutrino has a mass of about 10^{-2} eV, which is the expected mass for neutrinos.

This mechanism is based on the existence of new physics at the scale of the Grand Unification Theory, 10^{15}–10^{16} GeV.

6.9.2 *Interactions*

In nature, there are four interactions or forces: *gravitational*, *electromagnetic*, *strong nuclear*, and *weak nuclear force*.

The standard model only describes the last three.

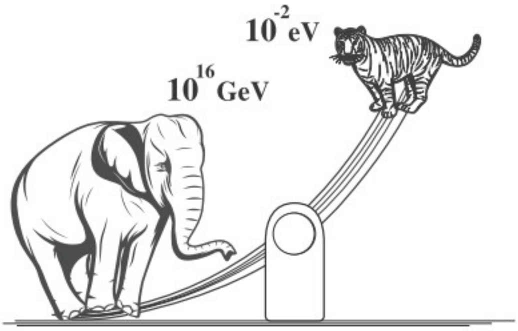

Fig. 6.18. The seesaw mechanism. The larger the mass of the right-handed, sterile, neutrino (see Appendix D), the smaller the mass of the usual neutrino.

- *Electromagnetic force*: It can be attractive and repulsive and decays with the square of the distance. It is responsible for electromagnetic phenomena (emission and absorption of radiation) and chemical reactions. It holds electrons around nuclei. The electromagnetic force is mediated by the *photon*.
- *Strong nuclear force*: It is the most intense. It is responsible for the existence of the atomic nucleus, has a range of action in the order of the size of the proton, and grows as the distance decreases. It is mediated by eight bosons, the *gluons*.
- *Weak nuclear force*: It is responsible for radioactive decay and contributes to making the stars shine. It has three mediators, the bosons W^+, W^-, Z^0, which, unlike the mediators of the other forces, have mass, respectively 90 and 100 times heavier than that of the proton. Their mass comes directly from the Higgs mechanism. The weak and electromagnetic forces are manifestations of a single force, the *electroweak force*.
- *Gravitational force*: Although not described by the standard model, it is of fundamental importance. It has a never-observed theoretical mediator, the *graviton*. It is the weakest of forces, 10^{25} times less intense than the weak one, 10^{36} than the electromagnetic one, and 10^{38} than the strong one, but it dominates in large spaces and in the universe. It is always attractive and responsible for the formation of structures and the rotation of planets around stars, galaxies, and so on.

The *Higgs boson*, discovered in 2012, supplies mass to elementary particles such as quarks and W^+, W^-, Z^0 bosons. Its measured mass is 125 GeV, much smaller than the theoretical expected value. This problem, which we discuss in the next chapter, is usually indicated with the *hierarchy problem*, and its solution implies the existence of physics beyond the standard model.

The 12 elementary particles (not counting quark color, which would bring them to 24) do not feel the various forces in the same way. All "feel" the gravitational force, while only the charged ones perceive the electromagnetic force. Strong force is felt only by quarks.

Protons, neutrons, and electrons form *atoms*. The simplest one, the hydrogen atom, is made of a proton and an electron. A gold atom has 79 protons and 118 neutrons, and for electrical neutrality, 79 electrons. An atom is basically an empty structure in quantum mechanics.[9] To get an idea, if the core had the diameter of a tennis ball (7 cm in diameter), the closest electron would be at a distance of 3 km.

Although the standard model is a very successful theory with incredible precision, it is not a complete theory, because it is not able to fully explain the nature of the world. Several open questions remain. For example, why is the world made only by a generation of quarks and leptons but in nature there are three of them? Are quarks and leptons really elementary particles? The mass of the particles is not given by the model but is obtained experimentally. Is there any way to get it from theory? That is, why does the standard model not foresee the mass of the particles? And again, the universe contains only matter, and if so, why? How can gravitation be integrated into the model? Is it possible to build a theory in which all forces are unified? How can we solve the hierarchy problem (we will discuss this issue in Chapter 7).

Finally, the standard model does not provide the existence of dark matter and energy. Other theories besides the standard model, which we will illustrate later, have been proposed that introduce models for dark matter and dark energy. However, they have never been confirmed experimentally.

---

[9] In quantum field theory, this is no longer the case. The space between electrons and protons contains virtual particles.

= # # %$
V/ & 9 ' 9) !" (" V/ ! Q 7
! "

! " !" # # $ %$ % & ' (!
'$ " !") " ! $# # *)
! +,-.+ + +/ ! + ') "%) $ "01
! +,-.+ + +/ ! + ') "%) $ 2"
(!2 &&

CHAPTER VII

RIEMANN'S SYMBOLS AND PROPERTIES RELATING TO CURVATURE; RICCI'S AND EINSTEIN'S SYMBOLS; GEODESIC DEVIATION

1. Cyclic displacement and the relations between parallelism and curvature.

Schouten,[1] by his vector methods, and independently of him Pérès,[2] by ordinary calculus methods, have demonstrated the great importance, for determining the geometrical properties of a V_n, of the displacement of a direction round a closed circuit; in particular the importance of infinitesimal circuits in investigating local properties at a generic point P.

Consider a generic direction (a unit vector) $\mathbf{u}$ drawn from P, and give it a parallel displacement round a closed curve T of infinitesimal length so that it comes back again to P; after the displacement we shall have a direction $\mathbf{u}_1$, also drawn from P, but not in general coinciding with $\mathbf{u}$. The change in the contravariant components u^r due to the displacement round the circuit will in general depend on the *area* of the circuit, on its *configuration* (i.e. on the orientation in V_n of the element of surface on which the circuit is drawn), and on the *metrical properties* of the V_n at P. The influence of the last-named properties is exerted through the first and second derivatives of the a_{ik}'s; these derivatives occur in certain characteristic groups which are called *Riemann's symbols*, and which are composed of Christoffel's symbols and their first derivatives. In the particular case of a surface, these expressions reduce to one, which is that known in geometry as the (Gaussian) *curvature* of the surface; for any V_n the consideration of Riemann's symbols provides a convenient way of extending the notion of curvature.

In this chapter we shall first consider displacement round a particular form of infinitesimal circuit, namely, an elementary parallelogram. We shall then discuss some of the properties of Riemann's symbols, which occur in the investigation of the displacement, and shall use these properties to obtain the formula for changing the order of two successive covariant differentiations, by determining the difference between the derivatives. Lastly, we shall return to the question of displacement round any circuit whatever, and shall deduce from it the notion of curvature, first for a surface, then for any V_n whatever.

---

[1] "Die direkte Analysis zur neueren Relativitätstheorie", in *Verh. der Kon. Ak. van Wet. te Amsterdam*, Deel 12, No. 6, 1919. Cf. also the same writer's *Der Ricci-Kalkül* (Berlin, Springer, 1924), II, §§ 12–16.

[2] "Le parallélisme de M. Levi-Civita et la courbure riemannienne", in *Rend. della R. Acc. dei Lincei*, (5), Vol. XXVII (first half-year, 1919). pp. 425–428.

2. **Cyclic displacement round an elementary parallelogram.**

Let two elementary vectors, δP, $\delta' P$, be drawn from a generic point P of a V_n. We shall interpret the first as an elementary displacement PP_1, and give the vector $\delta' P$ a parallel displacement along it; let Q be the position of the extremity of $\delta' P$ after this displacement. If we apply the same process to δP, and give it a parallel displacement along the path PP_2, we reach the same point Q (as we have already shown in Chapter V, § 19), even if we retain terms of the type $\delta\delta' P$, $\delta'\delta P$, while neglecting terms of the second order in δP and $\delta' P$. We can therefore, in any V_n, consider an *elementary parallelogram* PP_1QP_2.

We shall adopt the obvious convention of representing by δq the change in any quantity q (scalar or vector) in passing from P to P_1 and by $\delta' q$ the analogous change in passing from P to P_2. For a vector, we shall calculate these changes by the formulæ of parallelism.

Now let $D_1 q$ be the change in q on passing from P to Q along the path PP_1Q, and $D_2 q$ the analogous change on passing along the other pair of sides PP_2Q which with the first pair make up the circuit.

It will be seen at once that (neglecting second order terms as explained above) the total change Δq on going round the entire circuit in the sense PP_1QP_2P is $D_1 q - D_2 q$. We shall first examine $D_1 q$.

The change denoted by δ corresponds to the displacement along PP_1; hence, if the value of our quantity was q at P, its value at P_1 will be

$$q + \delta q = q_1.$$

The displacement along P_1Q changes q_1 into $q_1 + \delta' q_1$, so that at Q we shall have the quantity

$$\begin{aligned} &q + \delta q + \delta'(q + \delta q) \\ &= q + \delta q + \delta' q + \delta'\delta q \end{aligned}$$

so that

$$D_1 q = \delta q + \delta' q + \delta'\delta q.$$

As $D_2 q$, by its definition, differs from $D_1 q$ only by interchanging P_1 and P_2, and therefore δ and δ', we get

$$D_2 q = \delta' q + \delta q + \delta\delta' q.$$

It follows that the change caused by the displacement round the circuit is

$$\Delta q = (\delta'\delta - \delta\delta')q. \quad . \quad . \quad . \quad . \quad (1)$$

We must now find an explicit form for this expression, supposing that the quantity q is a vector $\mathbf{u}$, and calculating the increments δ and δ' by the formulæ of parallelism. By these formulæ, the changes δu^r of the contravariant components will be given by the Pfaffian (Chapter V, equation (52))

$$\delta u^r = -\sum_1^n{}_{ih} \{ih, r\}\, u^i\, \delta x_h, \quad . \quad . \quad . \quad (2)$$

while the changes $\delta' u^r$ will be given by the same Pfaffian relative to the increments $\delta' x_h$. From (1) we see that we have to calculate the bilinear covariant (cf. Chapter II, § 4) of this Pfaffian.

Differentiating (2) with the symbol δ' we get

$$\delta'\delta u^r = -\sum_1^n{}_{ih} \delta'\{ih, r\}\, u^i\, \delta x_h - \sum_1^n{}_{ih} \{ih, r\}\, \delta' u^i\, \delta x_h$$
$$- \sum_1^n{}_{ih} \{ih, r\}\, u^i\, \delta'\delta x_h.$$

To expand the first sum, we note that the expressions $\{ih, r\}$ are functions of the x's, and therefore

$$\delta'\{ih, r\} = \sum_1^n{}_k \frac{\partial}{\partial x_k} \{ih, r\}\, \delta' x_k.$$

The second, on substituting for $\delta' u^i$ the expression analogous to (2), becomes

$$\sum_1^n{}_{ihkl} \{ih, r\} \{kl, i\}\, u^l\, \delta' x_k\, \delta x_h,$$

or, interchanging i and l in order to get the factor u^i here too,

$$\sum_1^n{}_{ihkl} \{lh, r\} \{ki, l\}\, u^i\, \delta' x_k\, \delta x_h.$$

We have, therefore, taking out the factors $u^i\,\delta x_h\,\delta' x_k$ in the first two sums,

$$\delta'\delta u^r = -\sum_1^n{}_{ihk}\left[\frac{\partial}{\partial x_k}\{ih,\,r\} - \sum_1^n{}_l\{lh,\,r\}\{ki,\,l\}\right]u^i\,\delta x_h\,\delta' x_k$$
$$-\sum_1^n{}_{ih}\{ih,\,r\}\,u^i\,\delta'\delta x_h.$$

The expression for $\delta\delta' u^r$ can be obtained from this by interchanging δ and δ'; in the first sum we shall also interchange h and k, giving

$$\delta\delta' u^r = -\sum_1^n{}_{ihk}\left[\frac{\partial}{\partial x_h}\{ik,\,r\} - \sum_1^n{}_l\{lk,\,r\}\{hi,\,l\}\right]u^i\,\delta' x_k\,\delta x_h$$
$$-\sum_1^n{}_{ih}\{ih,\,r\}\,u^i\,\delta\delta' x_h.$$

In taking the difference $\delta'\delta u^r - \delta\delta' u^r$ the third sum cancels out, because $\delta\delta' x_h = \delta'\delta x_h$ (cf. Chapter II, § 4), and there remain the terms involving the indices i, h, k, in which $u^i\,\delta x_h\,\delta' x_k$ can be taken out as a common factor. If we introduce *Riemann's symbols of the second kind*,

$$\left.\begin{aligned}\{ir,\,hk\} = \frac{\partial}{\partial x_k}\{ih,\,r\} - \frac{\partial}{\partial x_h}\{ik,\,r\}\\ -\sum_1^n{}_l[\{lh,r\}\{ik,l\} - \{lk,r\}\{ih,l\}]\end{aligned}\right\}(i,\,r,\,h,\,k = 1,\,2,\,\ldots\,n),\quad (3)$$

we shall therefore have

$$\Delta u^r = (\delta'\delta - \delta\delta')u^r = -\sum_1^n{}_{ihk}\{ir,\,hk\}\,u^i\,\delta x_h\,\delta' x_k. \quad . \quad (4)$$

This formula shows that the required increment $\Delta\mathbf{u}$ depends on the vector $\mathbf{u}$, on the two vectors δP, $\delta' P$ which define the parallelogram, and lastly, on the metric of the manifold, through the quantities $\{ir,\,hk\}$. From (4) it follows as a particular case that for Euclidean spaces Riemann's symbols as just defined are all zero, whatever may be the co-ordinates x chosen for reference. In fact, for such a manifold, we have $\Delta u^r = 0$ $(r = 1,\,2,\,\ldots\,n)$, since any vector resumes its original value after parallel displacement round any closed circuit whatever. Hence the right-hand side of (4) vanishes for every r, and for any value of the vector **u** and of the displacements δP, $\delta' P$, i.e. for any values of the arguments u^i, δx_h, $\delta' x_k$. The coefficients $\{ir, hk\}$ must therefore vanish separately.

It will be useful to point out at once the following two properties of the operator Δ:

(*a*) When applied to a product, it behaves like a symbol of ordinary differentiation, i.e.

$$\Delta(\psi\phi) = \psi\Delta\phi + \phi\Delta\psi,$$

which can be verified directly, by calculating first $\delta(\psi\phi)$, then $\delta'\delta(\psi\phi)$, &c.;

(*b*) when applied to a function of position, the result is zero, as is obvious from the meaning of the symbol.

If instead of the increments Δu^r we wish to find those of the covariant components, we can use the relation

$$u_j = \sum_1^n{}_r\, a_{jr}\, u^r,$$

and therefore, from the properties of the operator Δ,

$$\begin{aligned}\Delta u_j &= \sum_1^n{}_r\, a_{jr}\, \Delta u^r \\ &= -\sum_1^n{}_{rihk}\, a_{jr}\, \{ir, hk\}\, u^i\, \delta x_h\, \delta' x_k.\end{aligned}$$

If we introduce *Riemann's symbols of the first kind*,

$$(ij, hk) = \sum_1^n{}_r\, a_{jr} \{ir, hk\}, \quad . \quad . \quad . \quad . \quad (5)$$

we can sum with respect to r, and can then write

$$\Delta u_j = -\sum_1^n{}_{ihk}\, (ij, hk)\, u^i\, \delta x_h\, \delta' x_k, \quad . \quad . \quad (4')$$

which is analogous to (4).

Solving (5) we get Riemann's symbols of the second kind in terms of those of the first kind by the formula (the inverse of (5))

$$\{ir, hk\} = \sum_1^n{}_j\, a^{jr}\, (ij, hk.) \quad . \quad . \quad . \quad (5')$$

3. **Fundamental properties of Riemann's symbols of the second kind.**

As can be seen from their definition (3), Riemann's symbols of the second kind are functions of position, depending on the coefficients a_{ik}, their first derivatives (contained in Christoffel's symbols) and second derivatives (contained in the derivatives of Christoffel's symbols). They have the following fundamental properties:

(*a*) They are antisymmetrical in the last two indices, i.e.

$$\{ir, hk\} = -\{ir, kh\}, \quad . \quad . \quad . \quad (6)$$

whence in particular

$$\{ir, hh\} = 0.$$

This property follows immediately from (3).

(*b*) They constitute a mixed tensor,[1] contravariant with respect to the second index and covariant with respect to the other three, so that the symbol $\{ir, hk\}$ could also be denoted (as is sometimes done) by a^r_{ihk}. To prove this, consider the invariant

$$F = \sum_1^n{}_r \, p_r u^r,$$

where the p's are given (but completely arbitrary) functions of position, so that $\Delta p_r = 0$. If we give F a displacement round an infinitesimal circuit, we find (remembering the behaviour of the operator Δ)

$$\Delta F = \sum_1^n{}_r \, (u^r \Delta p_r + p_r \Delta u^r)$$
$$= \sum_1^n{}_r \, p_r \Delta u^r.$$

As F is invariant, this quantity must also be so; replacing Δu^r in it by its expression (4), we get the quadrilinear form

$$\Delta F = -\sum_1^n{}_{ihkr} \{ir, hk\} \, p_r u^i \delta x_h \delta' x_k, \quad . \quad . \quad (7)$$

which expresses the required property of the Riemann's symbols, since the simple systems p_r, u^i, δx_h, $\delta' x_k$ are all arbitrary.

---

[1] Very generally, especially in works on the Theory of Relativity, called the *Riemann-Christoffel tensor.*

We can use the tensor character of Riemann's symbols to obtain a second proof of the fact that Riemann's symbols are all zero for a Euclidean V_n (the first proof is an immediate consequence of (4), as was shown in the preceding section). In fact, the definition (3) shows at once that they vanish in Cartesian co-ordinates, and in consequence they vanish in any other system of co-ordinates.

(*c*) They have an important cyclic property with respect to the three indices of covariance, namely,

$$\{ir, hk\} + \{hr, ki\} + \{kr, ih\} = 0. \quad . \quad . \quad (8)$$

To prove this, we again take F and formula (7), but we suppose that in them the p's are derivatives of an invariant function f of position (whose numerical values are otherwise arbitrary), and we also take as vector $\mathbf{u}$ an infinitesimal displacement with components $u^r = dx_r$, which is also arbitrary. With this choice, F becomes

$$F = \sum_1^n{}_r \frac{\partial f}{\partial x_r} dx_r = df,$$

and (7) becomes

$$(\delta'\delta - \delta\delta')\, df = -\sum_1^n{}_{ihkr} \{ir, hk\}\, p_r\, dx_i\, \delta x_h\, \delta' x_k. \quad . \quad (9)$$

Interchanging cyclically the three infinitesimal vectors denoted by the operators d, δ, δ', we get the other two formulæ:

$$(d\delta' - \delta' d)\, \delta f = -\sum_1^n{}_{ihkr} \{ir, hk\}\, p_r\, \delta x_i\, \delta' x_h\, dx_k, \qquad (9')$$

$$(\delta d - d\delta)\, \delta' f = -\sum_1^n{}_{ihkr} \{ir, hk\}\, p_r\, \delta' x_i\, dx_h\, \delta x_k. \quad (9'')$$

Now on the right-hand side of these last two formulæ we can arrange to have the product $dx_i\, \delta x_h\, \delta' x_k$ in the general term, merely by a suitable interchange of the indices of summation. We can then add (9), (9′), and (9″); the left-hand side gives 0, since the terms cancel out in pairs (e.g. $\delta'\delta df - \delta' d\delta f = \delta'(\delta df - d\delta f) = 0$, since f is a function of position); and we get

$$0 = \sum_1^n{}_{ihkr} [\{ir, hk\} + \{hr, ki\} + \{kr, ih\}]\, p_r\, dx_i\, \delta x_h\, \delta' x_k.$$

As p_r, dx_i, δx_h, $\delta' x_k$ are arbitrary, (8) follows at once.

4. Fundamental properties and number of Riemann's symbols of the first kind.

Riemann's symbols of the first kind, as defined by (5) (i.e. the quantities obtained by compounding the quadruple system of the symbols of the second kind with the system of coefficients a_{ik}) have the following properties:

(*a*) They are *covariant with respect to all four indices*, so that they may be denoted, as is often done, by $a_{ij,\,hk}$; this follows from the definition, in consequence of the law of contraction.

The remark that a Euclidean V_n has all its Riemann's symbols zero is true here too, whatever may be the system of reference.

(*b*) They are *antisymmetrical* with respect to each pair of indices, so that we have identically

$$(ij,\, hk) = -(ij,\, kh), \quad . \quad . \quad . \quad . \quad . \quad (10)$$

$$(ij,\, hk) = -(ji,\, hk). \quad . \quad . \quad . \quad . \quad . \quad (11)$$

The identity (10) follows at once from (5), and from the analogous property of the symbols of the second kind. To prove (11) we shall follow a method analogous to that used in § 3 (*b*), taking as invariant the scalar product of two arbitrary vectors **u**, **v**,

$$F = \overset{n}{\underset{1}{\Sigma}}{}_j\, v_j\, u^j.$$

Applying the operator Δ and remembering that in a parallel displacement the scalar product does not change (so that $\Delta F = 0$) we get

$$0 = \overset{n}{\underset{1}{\Sigma}}{}_j\, v_j\, \Delta u^j + \overset{n}{\underset{1}{\Sigma}}{}_j\, u^j\, \Delta v_j. \quad . \quad . \quad . \quad (12)$$

The expression for Δu^j is given by (4), by writing j in it instead of r; that for Δv_j is given by (4′). Substituting, we get

$$0 = \overset{n}{\underset{1}{\Sigma}}{}_{jihk}\, v_j \{ij,\, hk\}\, u^i\, \delta x_h\, \delta' x_k + \overset{n}{\underset{1}{\Sigma}}{}_{jihk}\, u^j\, (ij,\, hk)\, v^i\, \delta x_h\, \delta' x_k. \qquad (12')$$

In the first sum we express v_j in terms of the contravariant components v_r, and then, remembering (5), we sum with respect to j; we get successively

$$\overset{n}{\underset{1}{\Sigma}}{}_{jihkr}\, a_{jr}\, v^r \{ij,\, hk\}\, u^i\, \delta x_h\, \delta' x_k = \overset{n}{\underset{1}{\Sigma}}{}_{ihkr}\, v^r\, (ir,\, hk)\, u^i\, \delta x_h\, \delta' x_k;$$

lastly, changing the indices i and r into j and i respectively (to get the term in the same form as the second part of (12′)), we get

$$\sum_{1}^{n}{}_{jihk}\, v^i\,(ji,\, hk)\; u^j\,\delta x_h\,\delta' x_k.$$

We can now return to (12′), and taking out the common factor $v^i\, u^j\, \delta x_h\, \delta' x_k$ we get

$$0 = \sum_{1}^{n}{}_{jihk}\, [(ji,\, hk) + (ij,\, hk)]\; v^i\, u^j\,\delta x_h\,\delta' x_k; \qquad (13)$$

from this, since u^j, v^i, δx_h, $\delta' x_k$ are arbitrary, we get formula (11).

(*c*) Riemann's symbols of the first kind have also a *cyclic property* analogous to that of the symbols of the second kind and immediately deducible from the latter, namely,

$$(ij,\, hk) + (hj,\, ki) + (kj,\, ih) = 0, \quad . \quad . \quad (14)$$

where the second index remains fixed and the other three are permuted cyclically. This formula follows directly from (8), on multiplying by a_{jr} and summing with respect to r.

As each of the terms in this sum is antisymmetrical, we can at once obtain from (14) a similar identity

$$(ij,\, hk) + (ih,\, kj) + (ik,\, jh) = 0, \quad . \quad . \quad (14')$$

in which the first index remains fixed and the other three are permuted cyclically.

(*d*) Lastly, for the symbols of the first kind, there is a property of *permutability*, which is a consequence of the foregoing properties, and according to which we can interchange the two pairs of indices without altering the value of the symbol; namely,

$$(ij,\, hk) = (hk,\, ij). \quad . \quad . \quad . \quad . \quad (15)$$

To prove this, take (14′) and the three other identities obtained from it by cyclic permutation of the four indices in the order i, j, h, k, or

$$\begin{aligned}
\underline{(ij,\, hk)} + (ih,\, kj) + (ik,\, jh) &= 0,\\
(jh,\, ki) + (jk,\, ih) + \underline{(ji,\, hk)} &= 0,\\
\underline{(hk,\, ij)} + (hi,\, jk) + (hj,\, ki) &= 0,\\
(ki,\, jh) + (kj,\, hi) + (kh,\, ij) &= 0.
\end{aligned}$$

Adding the first and fourth and subtracting the second and third of these identities, and using the property of antisymmetry, we see that the terms cancel out in pairs, except the four underlined, which give

$$2(ij,\,hk) - 2(hk,\,ij) = 0,$$

whence the required property follows.

We shall now calculate the *number of independent Riemann's symbols of the first kind.* A quadruple system (cf. Chapter IV, § 2) has in general n^4 elements, if there are n independent variables. The number of distinct Riemann's symbols of each kind is smaller, however, as these symbols are connected by the identities we have just proved. We shall determine this number for the symbols of the first kind, dividing them into three classes, and counting separately those in each class, as follows:

(1) Symbols with only two different indices: these are of the type $(ij,\ ij)$, since the other possible arrangements are either reducible to this, or give zero values. Each pair of unequal indices i, j therefore gives a single symbol of this class, which thus contains

$$\frac{n(n-1)}{2};$$

(2) Symbols with three different indices: these are of the type $(ij,\,ih)$, since here too the other possible arrangements are reducible to this or give zero values. Every triplet of unequal indices will give three symbols of this type (since the repeated index may be any one of the three); since there are $\frac{n(n-1)(n-2)}{1\cdot 2\cdot 3}$ triplets, the number of distinct symbols of the class we are considering amounts to

$$\frac{n(n-1)(n-2)}{2};$$

(3) Symbols with four different indices: a set of four different indices i, j, h, k will give the three symbols

$$(ij,\,hk),\quad (ih,\,kj),\quad (ik,\,jh),$$

while every other possible arrangement gives a symbol reducible to one of these. But these three are not independent, on account of the cyclic relation (14′). It follows that each of the

$\dfrac{n(n-1)(n-2)(n-3)}{1 \cdot 2 \cdot 3 \cdot 4}$ sets of four indices gives two distinct symbols, so that the total number of these is

$$\frac{n(n-1)(n-2)(n-3)}{12}.$$

Adding these three partial totals, and simplifying, we get the total number N of independent Riemann's symbols of the first kind:

$$N = \frac{n^2(n^2-1)}{12}.$$

Thus for an ordinary surface ($n = 2$) we have $N = 1$; for a three-dimensional space, $N = 6$; for a four-dimensional space, $N = 20$.

5. **Bianchi's identities.**[1]

Bianchi's identities are cyclic relations between the covariant derivatives of Riemann's symbols of both the first and second kinds. They are obtained as follows:

Take formula (3), which defines the symbol of the second kind $\{ir, hk\}$, and differentiate it with respect to x_l. We note, however, that on differentiation the last part, which consists of terms of the second order in Christoffel's symbols, gives terms made up of the product of one such symbol by the derivative of another: the essential point for us is that, with reference to a specified point P, by choosing co-ordinates which are geodesic at that point, we can make all these terms vanish. We cannot, however, treat the first part in the same way, as the geodesic co-ordinates make Christoffel's symbols vanish but not their derivatives. We shall therefore have the formula, valid at the point P for co-ordinates geodesic at P,

$$\frac{\partial}{\partial x_l}\{ir, hk\} = \frac{\partial^2}{\partial x_k \, \partial x_l}\{ih, r\} - \frac{\partial^2}{\partial x_h \, \partial x_l}\{ik, r\}.$$

---

[1] These identities were stated without proof by PADOVA, on the strength of a verbal communication of RICCI (cf. "Sulle deformazioni infinitesime", in *Rend. della R. Acc. dei Lincei*, (4), Vol. V (first half-year, 1889), p. 176). They were then forgotten even by Ricci himself. BIANCHI rediscovered them and published a proof obtained by direct calculation in 1902 (*Ibid.*, (5), Vol. XI (first half-year, 1902), pp. 3–7).

Write down also the two other formulæ obtained from this by cyclic permutation of the indices h, k, l, leaving i and r fixed;

$$\frac{\partial}{\partial x_h}\{ir, kl\} = \frac{\partial^2}{\partial x_l \partial x_h}\{ik, r\} - \frac{\partial^2}{\partial x_k \partial x_h}\{il, r\},$$

$$\frac{\partial}{\partial x_k}\{ir, lh\} = \frac{\partial^2}{\partial x_h \partial x_k}\{il, r\} - \frac{\partial^2}{\partial x_l \partial x_k}\{ih, r\}.$$

Adding the terms on the left and right of these three equations we get

$$\frac{\partial}{\partial x_l}\{ir, hk\} + \frac{\partial}{\partial x_h}\{ir, kl\} + \frac{\partial}{\partial x_k}\{ir, lh\} = 0, \quad (16)$$

which holds at the point P, in the particular system of co-ordinates chosen. Now consider the following mixed tensor of rank five

$$A^{r}_{ihkl} = \{ir, hk\}_l + \{ir, kl\}_h + \{ir, lh\}_k,$$

in which the suffixes outside the brackets denote covariant differentiation. This system, referred to the point P and to co-ordinates geodesic at P, is identical with the left-hand side of (16), since in these conditions the covariant and ordinary derivatives are identical; all its elements are therefore equal to zero, and it will therefore be identically zero whatever may be the system of reference (cf. Chapter IV, § 12). We have thus proved the identity

$$\{ir, hk\}_l + \{ir, kl\}_h + \{ir, lh\}_k = 0, \quad . \quad (17)$$

which is a first form of the result established by Bianchi.

For the symbols of the first kind the analogous relation is easily proved from the definition of these symbols given by (5). In fact, taking the covariant derivative of this formula, and using Ricci's lemma, we find that

$$(ij, hk)_l = \sum_{1}^{n}{}_r \, a_{jr}\{ir, hk\}_l.$$

From this, permuting cyclically the indices h, k, l and summing, we get, by (17),

$$(ij, hk)_l + (ij, kl)_h + (ij, lh)_k = 0, \quad . \quad . \quad (17')$$

which is the second form of Bianchi's identity.

6. Commutation rule for the second covariant derivatives.

An important application of Riemann's symbols occurs in the formula which gives the relation between the two systems

$$A^{(j)}_{(i)hk} \quad \text{and} \quad A^{(j)}_{(i)kh},$$

obtained by double covariant differentiation from a generic tensor $A^{(j)}_{(i)}$, where (i) stands for the aggregate of indices of covariance $i_1 \ldots i_m$ and (j) for the aggregate of indices of contravariance $j_1 \ldots j_\mu$. To simplify the formulæ we shall consider a mixed double system A^j_i, with the remark that the procedure is similar if there are more indices.

We start as usual from the bilinear form

$$F = \overset{n}{\underset{1}{\Sigma}}{}_{ij}\, A^j_i\, \xi^i\, u_j,$$

the invariance of which determines the law of transformation of the A's; the ξ^i's are the contravariant components of an arbitrary vector $\boldsymbol{\xi}$, and the u_j's are the covariant components of another arbitrary vector $\mathbf{u}$. The procedure will consist in calculating, in two different ways, the quantity ΔF corresponding to a cyclic displacement round an elementary parallelogram (cf. § 2), and in equating the two expressions so obtained.

A first way of calculating ΔF is as follows. We associate the increments δ, δ' with two sides of the parallelogram, as in § 2, and use (1). Note first that from the definition of the covariant derivative (cf. Chapter VI, § 1) δF is given by

$$\delta F = \overset{n}{\underset{1}{\Sigma}}{}_{ijh}\, A^j_{i|h}\, \xi^i\, u_j\, \delta x_h.$$

Similarly, applying the symbol δ' to this expression, we shall get

$$\delta'\delta F = \overset{n}{\underset{1}{\Sigma}}{}_{ijhk}\, A^j_{i|hk}\, \xi^i\, u_j\, \delta x_h\, \delta' x_k.$$

From this, interchanging δ and δ', we get

$$\delta\delta' F = \overset{n}{\underset{1}{\Sigma}}{}_{ijhk}\, A^j_{i|hk}\, \xi^i\, u_j\, \delta' x_h\, \delta x_k;$$

subtracting these two equations, after interchanging the indices h and k in the second, we get

$$\Delta F = \overset{n}{\underset{1}{\Sigma}}{}_{ijhk} \left(A^j_{i\,|\,hk} - A^j_{i\,|\,kh}\right) \xi^i\, u_j\, \delta x_h\, \delta' x_k. \quad . \quad (18)$$

The other method of calculating this quantity consists in applying the operator Δ directly to the expression for F. Remembering the fundamental properties of Δ (§ 2) we shall get

$$\Delta F = \overset{n}{\underset{1}{\Sigma}}{}_{ij} A^j_i \left(u_j\, \Delta \xi^i + \xi^i\, \Delta u_j\right),$$

or, substituting for $\Delta \xi^i$ and Δu_j the expressions given by (4) and (4′),

$$\Delta F = -\overset{n}{\underset{1}{\Sigma}}{}_{ijlhk} A^j_i \{li,\, hk\}\, u_j\, \xi^l\, \delta x_h\, \delta' x_k$$
$$- \overset{n}{\underset{1}{\Sigma}}{}_{ijphk} A^j_i\, (pj,\, hk)\, u^p\, \xi^i\, \delta x_h\, \delta' x_k.$$

In order to get the second sum in the same form as the first, we shall express it in terms of symbols of the second kind and of the covariant components of $\mathbf{u}$; to do this we first use the property of antisymmetry with respect to the first pair of indices, and express the u^p's in terms of the u_l's, so that the sum becomes

$$+ \overset{n}{\underset{1}{\Sigma}}{}_{ijphkl} A^j_i\, (jp,\, hk)\, a^{pl}\, \xi^i\, u_l\, \delta x_h\, \delta' x_k.$$

Summing with respect to p, and using (5′), we get

$$\overset{n}{\underset{1}{\Sigma}}{}_{ijhkl} A^j_i \{jl,\, hk\}\, \xi^i\, u_l\, \delta x_h\, \delta' x_k.$$

Now, in order to reconstruct the second expression for ΔF in a suitable form, we shall first interchange some indices, so as to be able to take out the factor $\xi^i\, u_j\, \delta x_h\, \delta' x_k$ which occurs in (18). We must interchange l and i in the first sum, and l and j in the (modified) second; we shall then get, collecting both sums under a single summation sign,

$$\Delta F = -\overset{n}{\underset{1}{\Sigma}}{}_{ijhkl} \left[\, A^j_l \{il,\, hk\} - A^l_i \{lj,\, hk\} \right] \xi^i\, u_j\, \delta x_h\, \delta' x_k.$$

Comparing this expression with (18), and remembering that $\boldsymbol{\xi}$, $\mathbf{u}$, the δx's, and the $\delta' x$'s are arbitrary, we get the *commutation formula*

$$A^{j}_{i|hk} - A^{j}_{i|kh} = -\sum_{1}^{n}{}_{l} \left[A^{j}_{l} \{il, hk\} - A^{l}_{i} \{lj, hk\} \right]. \quad (19)$$

If the system from which we start has m indices of covariance and μ of contravariance, we must consider m vectors $\boldsymbol{\xi}$, determined by their contravariant components, and μ vectors $\mathbf{u}$, determined by their covariant components; by an analogous process we shall find

$$\left.\begin{aligned} A^{(j)}_{(i)|hk} - A^{(j)}_{(i)|kh} = -\sum_{1}^{n}{}_{l} \Bigg[& \sum_{1}^{m}{}_{r} A^{(j)}_{i_1 \ldots i_{r-1} l i_{r+1} \ldots i_m} \{i_r l, hk\} \\ & - \sum_{1}^{\mu}{}_{\rho} A^{j_1 \ldots j_{\rho-1} l j_{\rho+1} \ldots j_\mu}_{(i)} \{l j_\rho, hk\} \Bigg]. \end{aligned}\right\} \quad (20)$$

7. Cyclic displacement round any infinitesimal circuit.

We now return to the order of ideas interrupted at § 2. Given a direction $\mathbf{u}$ at a point P, we shall give $\mathbf{u}$ a parallel displacement (in V_n) round a closed curve T, infinitesimal, but of any shape whatever, and of course passing through P; we propose to calculate the change Du^r in a generic parameter of $\mathbf{u}$ caused by the cyclic displacement. The formula we shall find will be merely a generalization of (4), and must reduce to (4) if as a particular case we take for T an infinitesimal parallelogram.

For an elementary displacement dx_h we know that the change in u^r is

$$du^r = -\sum_{1}^{n}{}_{ih} \{ih, r\} u^i \, dx_h,$$

so that it has the form of a Pfaffian

$$\psi_r = \sum_{1}^{n}{}_{h} X_{rh} \, dx_h, \quad . \quad . \quad . \quad . \quad (21)$$

in which the X's are functions of position (since they contain Christoffel's symbols) and of the u^i's which are defined *along* *the curve T* by the equations of parallelism $du^r = \psi_r$. We have to calculate the integral

$$Du^r = \int_T \psi_r = \int_T \sum_1^n{}_h X_{rh}\, dx_h, \quad . \quad . \quad (22)$$

where we use the operator D to indicate the increment resulting from displacement right round the circuit. We shall now consider any surface σ containing the curve T, and we shall call Γ the region of this surface which is within T, and is such that T constitutes its complete boundary. We propose to transform the integral round the circuit T, which occurs in (22), into an integral over the surface Γ. To do this, we shall first introduce a system of co-ordinates q_1 and q_2 on the surface in question, defining them by the parametric equations

$$x_i = x_i(q_1, q_2) \qquad (i = 1, 2, \ldots n).$$

The dx's will consequently be linear functions of dq_1 and dq_2; substituting their expressions in the Pfaffian (21), this will take the form

$$du^r = \psi_r = Q_1\, dq_1 + Q_2\, dq_2, \quad . \quad . \quad (21')$$

where the quantities Q_1 and Q_2, like the X's, are defined along the curve T. The integral to be calculated will thus be

$$Du^r = \int_T (Q_1\, dq_1 + Q_2\, dq_2). \quad . \quad . \quad (22')$$

We shall suppose the curvilinear co-ordinates q_1, q_2 so chosen that the sense of integration round T is the same as that determined on Γ (at a generic point) by the rotation (through the angle less than 180°) from the positive direction of the line q_1 (i.e. in the sense of q_1 increasing) to the positive direction of the line q_2.

The transformation of the line integral (22′) into a surface integral taken over Γ could be effected at once if the Q's were defined as functions of position in the interior of Γ as well as on its boundary. But instead of this they contain the u^r's, which are given at P, and at points on T have values resulting from the parallel displacement along T itself, but are not defined for a point M within Γ, their values at M depending on the path followed in the parallel displacement of **u** from P to M. We shall, however, show that if Γ is infinitesimal the influence of the path of displacement on the values of the u^r's at M is negligible, and therefore we may consider the u^r's, and in consequence the X's or the Q's, as functions of position over the whole area Γ, which will enable us to make the required transformation of (22').[1]

We shall make some preliminary remarks on orders of magnitude, using for this purpose the general existence theorem for integrals of systems of ordinary differential equations. Such a system is constituted by the equations

$$du^r = \psi_r \qquad (r = 1, 2, \ldots n),$$

which define the functions u^r, along a generic line T, starting from a given set of initial values at P (cf. Chapter II, § 5).

Now the existence theorem assures us that in general (i.e. when certain not very restrictive conditions of continuity and differentiability are satisfied) the initial values define the integrals uniquely, and these integrals and their derivatives are continuous functions for values of the independent variable within an interval which is not shorter than some assignable quantity.

In our case, granting, what will naturally be the case, that the coefficients a_{ik} of ds^2 and the reciprocals a^{ik} are finite and continuous, as well as their first and second derivatives, in a certain region round P, and supposing also that the length of the vector **u** at P is limited, i.e. is not greater than some specified, but arbitrary, constant U, it can easily be deduced from the above-mentioned existence theorem that—considering the arc of the curve of displacement as the independent variable—the u^r's are defined (as continuous, differentiable, &c., functions), with P as starting-point, along any curve whatever, for a segment of the curve of length not greater than a certain quantity Λ

---

[1] We shall in fact here limit the discussion to an indication of the general lines of the argument, without pausing over the details needed to justify the various steps of the proof with complete rigour. There is an exhaustive proof in the article by H. TIETZE: "Ueber Parallelverschiebung in Riemann'schen Raümen", in *Math. Zeitschrift*, Vol. 16, 1923, pp. 308–317.

which depends exclusively on the metric of the manifold and on U. Hence it follows in particular that the differences between the u^r's and their initial values are of the same order[1] of magnitude as the length L of the arc along which the integration of the system $du^r = \psi_r$ is effected.

Further, the area Γ which we are considering on the surface σ is infinitesimal, in the sense that we propose to make it diminish indefinitely. It is therefore perfectly legitimate to suppose that it is already so small that every point in it can be reached from P by a line of length not greater than Λ and that the length of the whole contour T is also less than Λ.

We thus have the result that if **u** undergoes a parallel displacement from P to a point M within the area Γ, or on its boundary T, the u^r's at M differ from their values at P by quantities of the same order as L, if $L\,(\leqslant \Lambda)$ is the maximum length of the lines considered. As a first approximation, in which quantities of the same order as L are neglected, we can therefore take the components u^r as constant and $= u^{r(0)}$ over the whole area Γ, including its boundary.

We can find a closer approximation if we calculate the u^r's at M by integrating the Pfaffian (21) along a curve PQ, substituting, however, for the coefficients X_{rh} their values at P. This process involves an error of order L in the values of these coefficients, and therefore an error of order L^2 in the values found for the u^r's at M; the choice of the curve PM is indifferent, as in this case the Pfaffian becomes an exact differential. As the coefficients are constants, the integration can be at once effected, and will give linear functions of the x's for the u^r's. We shall thus have obtained the u^r's as functions of position at all points (including the boundary) of the area Γ, neglecting quantities of order L^2.

We can get a third approximation by substituting these approximate linear expressions of the u^r's in the coefficients X_{rh}. The X's will thus become functions of position, as was required, defined throughout Γ, including its boundary T, and coinciding on T (if we neglect L^2) with their accurate values as already defined. These values of the X's can be used to calculate the integral (22), which will give the value of Du^r, the error being now of order L^3. It was necessary to carry the approximation thus far, since the other two approximations make ψ an exact differential, and would give the value zero for Du^r, the obvious meaning of this being that Du^r is a quantity of order L^2.

---

[1] This means that the differences in question are not greater than the product of L by a certain *finite* coefficient, which does not depend on L or on the curve of integration, but only on P, U, and the metric of the manifold.

We shall therefore return to (22), giving the X's the meaning just explained, and transform it into the form (22′) by using the parametric equations of the surface σ; Q_1 and Q_2 will now represent functions of position defined at all points of Γ. In consequence[1] we can transform the line integral into a surface integral, getting

$$Du^r = \int_\Gamma \left(\frac{\partial Q_2}{\partial q_1} - \frac{\partial Q_1}{\partial q_2}\right) dq_1\, dq_2,$$

or also

$$Du^r = \int_\Gamma \left[\frac{\partial}{\partial q_1}(Q_2\, dq_2)\, dq_1 - \frac{\partial}{\partial q_2}(Q_1\, dq_1)\, dq_2\right]. \qquad (23)$$

We must now find the value of the integrand on the right of (23), which can be done by means of the following considerations, without writing out the expressions for Q_1 and Q_2 at full length.

Let the operator δ denote the increment of a quantity corresponding to a displacement along the line $q_2 =$ constant, when

---

[1] By the ordinary formulæ,

$$\int_\Gamma \frac{\partial f}{\partial q_1} dq_1 dq_2 = \int_T f dq_2, \quad \int_\Gamma \frac{\partial f}{\partial q_2} dq_1 dq_2 = -\int_T f dq_1, \quad . \quad (G)$$

where f is a function of q_1 and q_2 which is continuous, together with its first derivatives. Usually in these formulæ q_1 and q_2 are interpreted as Cartesian co-ordinates in a plane, and the sense in which the curve T is described is defined by the condition that the pair of directions s, n (s the tangent to the boundary in the sense in which it is described, n the normal to T drawn inwards) is congruent in the plane with the pair q_1, q_2. The formulæ obviously hold, however, independently of this interpretation in plane geometry, and can therefore be applied even if q_1, q_2 are any curvilinear co-ordinates whatever, the sense of describing the curve being determined by an analogous criterion to that just explained, provided that we introduce, at a generic point of the boundary curve, the directions tangential to q_1 (i.e. $q_2 =$ constant) and q_2, in the sense in which the respective parameters increase. Now we have already supposed (p. 187) that the auxiliary curvilinear co-ordinates q_1, q_2 behave as regards sense in precisely this way. Hence the equations (G) hold both in magnitude and in sign.

q_1 increases by dq_1; similarly, δ' corresponds to an increase dq_2 in q_2 alone. We shall therefore have

$$\left.\begin{aligned}\delta x_i &= \frac{\partial x_i}{\partial q_1}\,dq_1,\\ \delta' x_i &= \frac{\partial x_i}{\partial q_2}\,dq_2,\end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \qquad (24)$$

and analogous expressions for any function of position. Now note that in (21′) the first term represents precisely the increment of u^r due to a displacement of the first type ($dq_2 = 0$), so that

$$\delta u^r = Q_1\,dq_1,$$

and, the second term having an analogous meaning,

$$\delta' u^r = Q_2\,dq_2.$$

We can therefore write (23) in the form

$$\begin{aligned}Du^r &= \int_\Gamma \left[\frac{\partial}{\partial q_1}(\delta' u^r)\,dq_1 - \frac{\partial}{\partial q_2}(\delta u^r)\,dq_2\right]\\ &= \int_\Gamma [\delta\delta' u^r - \delta'\delta u^r];\end{aligned}$$

and from this, remembering equation (4), we get

$$Du^r = \int_\Gamma \sum_1^n{}_{ihk} \{ir, hk\}\, u^i\,\delta x_h\,\delta' x_k.$$

Now let us put ξ, η for the parameters (in V_n) of the lines $q_2 =$ constant, $q_1 =$ constant, i.e.

$$\xi^i = \frac{\delta x_i}{\delta s}, \quad \eta^i = \frac{\delta' x_i}{\delta' s},$$

where δs, $\delta' s$ are the lengths of the displacements, along the lines $q_2 =$ constant, $q_1 =$ constant, whose components are δx_i, $\delta' x_i$. Then we can also write

$$Du^r = \int_\Gamma f_r\,\delta s\,\delta' s, \quad . \quad . \quad . \quad . \qquad (25)$$

where for shortness we have put

$$f_r = \sum_1^n {}_{ihk} \{ir, hk\} u^i \xi^h \eta^k.$$

We can now remark that if ϑ is the angle between the coordinate lines q_1, q_2, the element of area (cf. Chapter V, § 7) is

$$d\Gamma = \delta s\, \delta' s \sin\vartheta,$$

and therefore (25) becomes

$$Du^r = \int_\Gamma \frac{f_r}{\sin\vartheta}\, d\Gamma.$$

By the mean value theorem, if $\left[\frac{f_r}{\sin\vartheta}\right]_{M_0}$ denotes the value of the function of position $\frac{f_r}{\sin\vartheta}$ at a suitable point M_0 (not known *a priori*) within Γ, and $D\Gamma$ the area of the region Γ, we can also write

$$Du^r = \left[\frac{f_r}{\sin\vartheta}\right]_{M_0} D\Gamma.$$

Now the value at M_0 of the function of position $\frac{f_r}{\sin\vartheta}$ differs from its value at P by quantities of order L (since the distance M_0P is of this order); since the area $D\Gamma$ is of order L^2, the error caused by substituting the value at P for that at M is of order L^3, which we have agreed to neglect. We shall therefore have

$$Du^r = \frac{f_r}{\sin\vartheta} D\Gamma,$$

or

$$Du^r = \frac{D\Gamma}{\sin\vartheta} \sum_1^n {}_{ihk} \{ir, hk\} u^i \xi^h \eta^k. \quad . \quad . \quad (26)$$

In this formula the quantities ξ^h and η^k represent the parameters of the lines q_1, q_2 at P, and ϑ is the angle between these lines; the values of the u^i's and of the Riemann's symbols refer to the point P. It will be seen that the influence of the circuit of displacement appears in this formula in three geometrical elements, which serve substantially to determine the circuit itself, namely: two directions $\boldsymbol{\xi}$, $\boldsymbol{\eta}$ (*a priori* any whatever) which determine the section on which the circuit is supposed drawn, together with the angle between them; and the area $D\Gamma$ of the circuit itself (measured according to the metric of V_n).

8. Pérès's formula.

From (26) we immediately get the fundamental formula which serves as a link between parallelism and curvature.

Take any (fourth) direction $\mathbf{v}$ drawn from P. Let α be the angle between $\mathbf{v}$ and $\mathbf{u}$, and $\alpha + D\alpha$ the angle between $\mathbf{v}$ and $\mathbf{u} + D\mathbf{u}$; we propose to calculate $D\alpha$. To do this we take the scalar product

$$\mathbf{u} \times \mathbf{v} = \cos\alpha$$

(assuming that $\mathbf{u}$, like $\mathbf{v}$, is a unit vector), and differentiate with the symbol D, remembering that $D\mathbf{v} = 0$ since $\mathbf{v}$ is a fixed vector. We get

$$\mathbf{v} \times D\mathbf{u} = -\sin\alpha\, D\alpha,$$

or, substituting for the scalar product on the left its expression $\sum_1^n{}_r\, v_r\, Du^r$, and using (26),

$$\sin\alpha\, D\alpha = -\frac{D\Gamma}{\sin\vartheta} \sum_1^n{}_{irhk} \{ir,\, hk\}\, u^i\, v_r\, \xi^h\, \eta^k.$$

If in this formula we express the v_r's in terms of the v^j's, and sum with respect to r (remembering (5)), we get

$$\sin\alpha\, D\alpha = -\frac{D\Gamma}{\sin\vartheta} \sum_1^n{}_{ijhk} (ij,\, hk)\, u^i\, v^j\, \xi^h\, \eta^k.$$

Now if the directions $\mathbf{u}$, $\mathbf{v}$ coincide or are opposite, this formula reduces to an identity of no interest, since on the left we have $\sin\alpha = 0$, and the right-hand side vanishes from the antisymmetry of the Riemann's symbols. Excluding this case, we can divide the whole equation by $\sin\alpha$, and we get *Pérès's formula*

$$D\alpha = \frac{-D\Gamma}{\sin\alpha \sin\vartheta} \sum_1^n{}_{ijhk} (ij,\, hk)\, u^i\, v^j\, \xi^h\, \eta^k. \quad . \quad (27)$$

9. Application to surfaces. Gaussian curvature of a V_2.

In considering the particular case of a V_2, i.e. an ordinary surface, the directions $\mathbf{u}$, $\mathbf{v}$ must of course be contained in the section (the only one there is) defined by $\boldsymbol{\xi}$, $\boldsymbol{\eta}$. Since the only purpose of these last two vectors is to specify the section on which the circuit is drawn, we can make them coincide with **u** and **v** without loss of generality; (27) will then become

$$D\alpha = \frac{-D\Gamma}{\sin^2\alpha} \sum_{1}^{n}{}_{ijhk} (ij,\, hk)\, u^i v^j u^h v^k. \quad . \quad . \quad (27')$$

Since for $n = 2$ Riemann's symbols which do not vanish are represented by the single arrangement of indices (12, 12), this formula can be further reduced to

$$D\alpha = \frac{-D\Gamma}{\sin^2\alpha} (12,\, 12)\, (u^1 v^2 - u^2 v^1)^2;$$

or finally, remembering the expression for $\sin\alpha$ in terms of the parameters of the two directions **u**, **v** (cf. Chapter V, § 4, formula (9′)), we get

$$D\alpha = -\frac{D\Gamma}{a} (12,\, 12).$$

It is usual to put

$$\frac{(12,\, 12)}{a} = K, \quad . \quad . \quad . \quad . \quad . \quad (28)$$

so that the foregoing formula becomes

$$\frac{D\alpha}{D\Gamma} = -K. \quad . \quad . \quad . \quad . \quad . \quad (29)$$

From this it will be seen that the function of position K defined by (28) is an invariant; it depends on the coefficients a_{ik} and their first and second derivatives, and is identical with the quantity which in the theory of surfaces is known as the *total*, or *Gaussian*, *curvature* (the product of the curvatures of the principal sections).[1] The equation (29) can be put in a more instructive form if we introduce the *angle of parallelism* ϵ, i.e. the angle between **u** and $\mathbf{u} + D\mathbf{u}$ (or between **u** and its parallel obtained by displacement round the circuit), measured in the sense in which the circuit is described. We can also say that ϵ is the angle through which **u** has been rotated as a result of the cyclic displacement. It is obvious that ϵ has the same absolute value as $D\alpha$, but we shall see that as regards sign the precise relation is

---

[1] See also below, Chapter IX, § 5.

$$\epsilon = -D\alpha.$$

To show this we need only remember the convention adopted above (§ 7) that the circuit is to be described in the positive sense with respect to the co-ordinate lines q_1, q_2, or from the versor $\boldsymbol{\xi}$ to the versor $\boldsymbol{\eta}$. As $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ now coincide with $\mathbf{u}$ and $\mathbf{v}$ the sense in which the circuit is described is from $\mathbf{u}$ to $\mathbf{v}$ (through the convex angle). Accordingly (see fig. 3) if $D\alpha > 0$, $\mathbf{u} + D\mathbf{u}$ is outside the convex angle $\widehat{\mathbf{uv}}$, and is therefore reached from $\mathbf{u}$ by moving in the negative sense ($\epsilon < 0$), and vice versa. We can therefore write (29) in the form

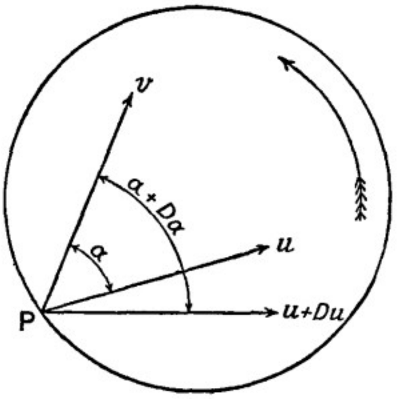

Fig. 3

$$\frac{\epsilon}{D\Gamma} = K, \quad . \quad . \quad . \quad . \quad . \quad (29')$$

which gives the following important interpretation of the curvature K: K *is the ratio of the angle of parallelism* (taken with its proper sign in relation to the sense of describing the circuit) *to the area of the circuit.*

In the case of a Euclidean V_2 Riemann's symbol is zero (cf. § 4) and therefore $K = 0$. This can also be deduced from the geometrically obvious fact (already used in § 2) that the parallel displacement is integrable (i.e. that the result does not depend on the curve of displacement).

10. **Riemannian curvature of a V_n.**

If instead of a surface we consider any V_n whatever, the notion of curvature becomes less simple. If P is a fixed point of the V_n, then with every section through P determined by two arbitrary directions $\boldsymbol{\xi}$, $\boldsymbol{\eta}$ drawn from P we can associate an invariant K, which is called the *Riemannian curvature* of the V_n at P with respect to the section considered. Following Riemann, we construct the geodesic surface determined by the point P and the two directions $\boldsymbol{\xi}$, $\boldsymbol{\eta}$, and then take the Gaussian curvature K of this geodesic surface as the curvature of the V_n at the point and in the section in question. In general the Riemannian curvature differs in the different sections.

The foregoing considerations enable us to give another important definition of the Riemannian curvature, and to find the analytical expression for it.

Given the elements P, $\boldsymbol{\xi}$, $\boldsymbol{\eta}$, construct the geodesic surface g defined by them and consider an infinitesimal circuit on g, passing through P, of area $D\Gamma$. Take one of the given directions, say $\boldsymbol{\xi}$, and give it a parallel displacement with respect to the surface g round the circuit, in the sense $\boldsymbol{\xi} \rightarrow \boldsymbol{\eta}$. Now calculate by Pérès's formula the change $D\alpha$ in the angle between $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$, i.e. the difference between its values before and after the displacement. The curvature K will then be given by (29). Now from Severi's theorem that an infinitesimal parallel displacement with respect to the surface g (using the metric of g) can be replaced by the analogous infinitesimal parallel displacement in V_n, it follows at once that this method of calculating K does not really involve the use of the geodesic surface g. Hence the Riemannian curvature K can be defined as *the ratio* (*with sign changed*) *of* $D\alpha$ *to* $D\Gamma$, *where* $D\alpha$ *is the change in the angle between the given directions* $\boldsymbol{\xi}$, $\boldsymbol{\eta}$ *caused by the parallel displacement in* V_n *of one of these directions round an infinitesimal circuit of area* $D\Gamma$, *belonging to the section* $\boldsymbol{\xi}$, $\boldsymbol{\eta}$, and described in the sense $\boldsymbol{\xi} \rightarrow \boldsymbol{\eta}$. We therefore have

$$K = -\frac{D\alpha}{D\Gamma}, \quad . \quad . \quad . \quad . \quad . \quad (30)$$

as in the V_2.

The explicit expression for K corresponding to this can be obtained from (27′), and we get

$$K = \frac{1}{\sin^2\alpha} \sum_{1}^{n}{}_{ijhk} (ij, hk)\, u^i v^j u^h v^k. \quad . \quad . \quad (31)$$

The symmetry of the right-hand side in $\mathbf{u}$ and $\mathbf{v}$ provides formal confirmation of the fact that it is immaterial whether we displace **u** or **v**, as we get in either case the same value for $D\alpha$ and the same value for K.

A third definition of the Riemannian curvature can be obtained from the proof of the following lemma.

Take any three points P, P', P'' on a surface V_2, and join them in pairs by arcs of geodesics, forming a so-called *geodesic triangle*. If ϕ, ϕ', ϕ'' are the angles of this triangle, the quantity

$$\epsilon = \phi + \phi' + \phi'' - \pi \quad . \quad . \quad . \quad . \quad (32)$$

is called the *geodesic excess*.

Now take any point on the triangle, and give a parallel displacement round the triangle to the direction of the side at that point, or of one of the sides if the point is a vertex; e.g. starting from P, the direction **u** at P of the side PP' (taken in the sense $P \rightarrow P'$).

We want to show that the angle between the initial and final positions of **u** (measured from the initial position in the sense which at each vertex is dependent on the sense $PP'P''$ in which the circuit is described), i.e. *the angle of parallelism* (relative in this case to a circuit of a special kind, but without the restriction of being infinitesimal), *is the same as the geodesic excess* ϵ.

For the proof we shall follow **u** in a cyclic displacement round the circuit $PP'P''$, noting in the first place that from P to P' **u** remains tangential to the side PP', on account of the autoparallelism of geodesics (cf. Chapter V, § 11). On arriving at P', **u** is thus inclined at an angle $\pi - \phi'$ (outside the triangle at P') to the side $P'P''$ (in the sense indicated by these letters); more precisely, **u** is behind the tangent to the side $P'P''$ by an angle of $\pi - \phi'$, the sense of rotation at P', as we have already said, being determined by the sense of description $PP'P''$. In the displacement from P' to P'' this angle remains unchanged; at P'' there will be a further loss of $\pi - \phi''$ (with respect to the new side $P''P$); and finally at P yet another loss of $\pi - \phi$ (with respect to PP'). Taking all these together, we see that in its final position the parallel to **u** has been rotated away from its initial position through an angle of $3\pi - (\phi + \phi' + \phi'')$ in the negative direction, or $\epsilon - 2\pi$ in the positive direction. Now in the pencil of directions at a point an angle is determined geometrically by a quantity of the form $\theta \pm 2n\pi$, where n is any integer; hence we have proved that *one value of the angle of parallelism is the geodesic excess* ϵ. Now in a Euclidean space the geodesic excess is zero, and in any manifold whatever, for an infinitesimal triangle, the excess is infinitesimal; we therefore see that for reasons of continuity the value adopted for the angle of parallelism is in fact the most suitable, being that which tends to zero with the triangle.

The lemma just proved is rigorously true in a V_2, whatever may be the magnitude of the geodesic triangle considered. If in particular we apply it to an infinitesimal triangle, we can substitute ϵ for $-D\alpha$ in (29), so getting

$$K = \frac{\epsilon}{D\Gamma},$$

a formula which defines the Riemannian curvature as *the ratio of the geodesic excess to the area of an infinitesimal geodesic triangle lying in the section considered, and having one vertex at the given point* P. It will be seen that this is an obvious extension to n-dimensional manifolds with any metric of an elementary theorem in spherical geometry (the area of a spherical triangle = the spherical excess × the square of the radius); the latter theorem, however, unlike the former, holds also for a finite triangle.

We must confine ourselves to a mere reference to the important researches of Professors Schouten[1] and Bompiani[2] on the simultaneous cyclic displacement of several directions, and even of the whole sheaf of directions drawn from a single point. Their work throws light on the theory of Riemannian curvature under various new aspects.

11. Case of a V_3. The tensors a_{ik} of Ricci and G_{ik} of Einstein.

For a manifold of three dimensions the symbols of the first species (ij, hk) which do not vanish reduce substantially, in virtue of (10), (11), and (14), to the scheme

$$(i+1 \;\; i+2, k+1 \;\; k+2) \qquad (i, k = 1, 2, 3),$$

---

[1] See references given in note (1), p. 172.

[2] "Studi sugli spazi curvi", in *Atti del R. Ist. Veneto*, Vol. LXXX, 1921, pp. 355–386, 839–859, 1113–1145.

with the convention (Chapter VI, § 9) of regarding as equivalent two indices which differ by a multiple of 3.

Using this notation, we now introduce after Ricci the double symmetric system

$$\alpha^{ik} = \frac{(i+1\ \ i+2,\ \ k+1\ \ k+2)}{a} \qquad (i, k = 1, 2, 3), \quad (33)$$

which constitutes, as we shall now show, a *contravariant tensor*. In fact, if we make use of the contravariant system ϵ (Chapter VI, § 8) we see that (33) is equivalent to

$$\alpha^{ik} = \tfrac{1}{4} \overset{3}{\underset{1}{\Sigma}}{}_{pqrs}\, \epsilon^{ipq}\, \epsilon^{krs}\, (pq, rs) \qquad (i, k = 1, 2, 3): \ . \quad (33')$$

which proves the assertion. The verification of this last formula is immediate, when we remember that of the various determinations *a priori* possible for the pair p, q, there are only two,

$$p = i + 1, \ \ q = i + 2$$

and $$p = i + 2, \ \ q = i + 1,$$

corresponding to which ϵ^{ipq} has a value different from zero, viz. $\frac{1}{\sqrt{a}}$ in the former case, $-\frac{1}{\sqrt{a}}$ in the latter. Similarly, for the other pair r, s, the only determinations to which there corresponds a non-vanishing ϵ^{krs} are

$$r = k + 1, \ \ s = k + 2$$

and $$r = k + 2, \ \ s = k + 1.$$

The sum $\overset{3}{\underset{1}{\Sigma}}{}_{pqrs}$ reduces therefore to four terms all expressible as

$$\frac{(i+1\ \ i+2,\ \ k+1\ \ k+2)}{a},$$

thus giving the result stated in (33).

The name *Ricci's symbols* is sometimes given to the α^{ik}'s just defined, or, more specifically, to *the elements*

$$\alpha_{ik} = \overset{3}{\underset{1}{\Sigma}}{}_{jh}\, a_{ij}\, a_{hk}\, \alpha^{jh} \qquad (i, k = 1, 2, 3) \quad . \quad (34)$$

of the reciprocal covariant tensor.

It is worth noting that equations (33), if we take account of the properties of Riemann's symbols in respect of symmetry and antisymmetry, can be solved so as to give the values of these symbols in the form

$$(ij, hk) = \sum_{1}^{3}{}_{\nu\rho}\, a^{\nu\rho}\, \epsilon_{\nu ij}\, \epsilon_{\rho hk}. \quad . \quad . \quad . \quad (33'')$$

In fact, expanding the second member and attending to the definition of the covariant system ϵ, we find the value zero corresponding to every set of four i, j, h, k for which the first member vanishes, and the value aa^{ik} corresponding to a set of four of the type $i+1, i+2, k+1, k+2$; (33″) therefore follows from (33).

By contracting the Ricci tensor α by means of the fundamental tensor (the coefficients of ds^2, or their reciprocals) we obtain the *linear invariant* of the tensor α

$$\mathcal{M} = \sum_{1}^{3}{}_{ik}\, a_{ik}\, \alpha^{ik} = \sum_{1}^{3}{}_{ik}\, a^{ik}\, \alpha_{ik}. \quad . \quad . \quad . \quad (35)$$

We may point out another formal relation of which use will be made in Chapter XII. For any V_n whatever we can derive from the Riemannian tensor, by contraction with respect to two indices, the covariant double tensor

$$G_{ik} = \sum_{1}^{n}{}_{jh}\, a^{jh}\, (ij, hk), \quad . \quad . \quad . \quad . \quad (36)$$

which, in virtue of equations (5), can also be written in the form

$$G_{ik} = \sum_{1}^{n}{}_{h} \{ih, hk\}. \quad . \quad . \quad . \quad . \quad (36')$$

This tensor was noticed by Ricci, who applied it to the study of the local distribution of curvatures in a V_n; it was afterwards taken up by Einstein, who gave it a fundamental place in the theory of relativity (in which $n = 4$): it is commonly known as the *Einstein tensor*.

For $n = 3$ the α_{ik}'s are related in a simple way to the G_{ik}'s. To bring out the connexion simply and neatly, it is convenient to make use of two properties of the ternary systems ϵ: one expressed by the identity (which can be verified immediately)

$$\sum_{1}^{3}{}_{j}\, \epsilon^{pqj}\, \epsilon_{rsj} = \delta^p_r \delta^q_s - \delta^p_s \delta^q_r \qquad (p, q, r, s = 1, 2, 3), \quad . \quad (37)$$

the δ's having the usual meaning (0 for different indices, 1 for equal indices); and the other translating, as it were, the definition of a^{jh} as a reciprocal element

$$a^{jh} = \frac{1}{2} \sum_{1}^{3}{}_{pqrs} \epsilon^{pqj} \epsilon^{rsh} a_{pr} a_{qs}.$$

Substitute in equations (36), in place of the a^{jh}'s, the second member here; and, in place of the (ij, hk)'s, the expression for them in (33″), with $\epsilon_{\rho hk}$ changed into $-\epsilon_{\rho kh}$. Taking account of (37) we find

$$G_{ik} = \frac{1}{2} \sum_{1}^{3}{}_{pqrs\nu\rho} a_{pr} a_{qs} \alpha^{\nu\rho} (\delta^p_\nu \delta^q_i - \delta^p_i \delta^q_\nu)(\delta^r_k \delta^s_\rho - \delta^r_\rho \delta^s_k).$$

Of the four terms obtained by expanding the product of the bracketed expressions, the two which are positive are merely interchanged by interchange of p with q and of r with s, which does not alter the product $a_{pr} a_{qs}$; and similarly with the negative terms. We can therefore confine our attention to one term only of each kind, and suppress the factor $\frac{1}{2}$. If we take, e.g.

$$\delta^p_\nu \delta^q_i (\delta^r_k \delta^s_\rho - \delta^r_\rho \delta^s_k),$$

and bear in mind the meaning of the symbols δ, we find that the sum reduces to one with respect to ν and ρ only, giving

$$G_{ik} = \sum_{1}^{3}{}_{\nu\rho} (a_{\nu k} a_{i\rho} \alpha^{\nu\rho} - a_{\nu\rho} a_{ik} \alpha^{\nu\rho}),$$

or finally, having regard to (34), (35),

$$G_{ik} = \alpha_{ik} - \mathcal{M} a_{ik} \qquad (i, k = 1, 2, 3). \quad . \quad (38)$$

12. Curvature of a manifold of three dimensions around a point. Principal directions and invariants.

Consider in a V_3 a generic section or *facet*, f, defined by two directions (versors) **u**, **v**, whose parameters are u^i, v^i $(i = 1, 2, 3)$, issuing from the same point P. Let **w** be their vector product (Chapter VI, § 8), the moments of which are

$$w_\nu = \sum_{1}^{3}{}_{ij} \epsilon_{\nu ij} u^i v^j \qquad (\nu = 1, 2, 3). \quad . \quad . \quad (39)$$

Corresponding to the section f in our manifold V_3 we have the

Riemannian curvature K given by (31), with $n = 3$. In the sum $\overset{3}{\underset{1}{\Sigma}}_{ijhk}$ of (31) it is convenient to introduce, in place of Riemann's symbols (ij, hk) the expressions for them given in (33″). Taking account of (39), we find immediately

$$K = \frac{1}{\sin^2 \alpha} \overset{3}{\underset{1}{\Sigma}}_{\nu\rho}\, a^{\nu\rho}\, w_\nu\, w_\rho. \quad . \quad . \quad . \quad (40)$$

Deferring for a moment the illustration of this formula, we recall the fact that in general the length w of the vector $\mathbf{u} \wedge \mathbf{v}$ (the vector product of $\mathbf{u}$ and $\mathbf{v}$) is given by the product of the lengths of $\mathbf{u}$, $\mathbf{v}$ by the sine of the angle between them. For the Euclidean V_3 this implies, as already mentioned (Chapter VI, § 8), the identity of the moments (39) referred to Cartesian co-ordinates with the ordinary components (orthogonal projections) of the vector product $\mathbf{u} \wedge \mathbf{v}$. For a general V_3, it is sufficient to remember that we can choose co-ordinates which are *locally Cartesian* at any assigned point P, so that the a_{ik}'s have the values δ_i^k (and so that also—though this is not important for the present purpose—their first derivatives vanish). Now both in the measures of the lengths and the angular distances of vectors proceeding from the point P, and in the definition (39) of the w_ν's, there enter only the components of the vectors and the values of the a_{ik}'s at P. Locally then everything is the same as for Euclidean space.[1]

Turn now to our case, in which $\mathbf{u}$ and $\mathbf{v}$ are unit vectors. It follows that $w = \sin\alpha$, whence $\frac{w_i}{w} = \frac{w_i}{\sin\alpha}$ constitute the

---

[1] We can also of course calculate the length of $\mathbf{w}$. Thus we note that, besides (39), we have the equivalent formulæ defining the contravariant components:

$$w^\nu = \overset{3}{\underset{1}{\Sigma}}_{hk}\, \epsilon^{\nu hk}\, u_h\, v_k.$$

Hence, in view of (37) (the index of summation in which we may suppose transferred to the first place),

$$\begin{aligned} w^2 &= \overset{3}{\underset{1}{\Sigma}}_\nu\, w^\nu w_\nu = \overset{3}{\underset{1}{\Sigma}}_{ijhk}\, u^i v^j u_h v_k \left(\delta_i^h \delta_j^k - \delta_j^h \delta_i^k\right) \\ &= \overset{3}{\underset{1}{\Sigma}}_i\, u^i u_i \overset{3}{\underset{1}{\Sigma}}_j\, v^j v_j - \overset{3}{\underset{1}{\Sigma}}_i\, u^i v_i \overset{3}{\underset{1}{\Sigma}}_j\, v^j u_j \\ &= u^2 v^2 - (uv \cos\alpha)^2 = u^2 v^2 \sin^2\alpha. \end{aligned}$$

moments λ_i of the direction $\boldsymbol{\lambda}$ normal to the section *f*. The *sense* assigned to the normal by (39) is characterized by the system ϵ or, geometrically, by the sense of the trihedron formed at P by the positive directions of the co-ordinate lines. In fact, from these equations (39), if we suppose, for example, that the lines 1, 2 are taken in the directions **u**, **v**, there results

$$w_1 = w_2 = 0, \qquad w_3 = \frac{\sqrt{a}}{\sqrt{a_{11}\, a_{22}}} > 0,$$

so that (Chapter V, § 22) **w** makes an acute angle with the positive direction of the line 3, whose parameters are

$$s_3^1 = s_3^2 = 0, \qquad s_3^3 = \frac{1}{\sqrt{a_{33}}} > 0.$$

Thus $\boldsymbol{\lambda}$ is perpendicular to f at P, and directed so that the trihedron **u, v,** $\boldsymbol{\lambda}$ has the same sense as the trihedron formed by the positive directions of the co-ordinate lines at the same point P.

Equation (40) now takes the form, given by Ricci,

$$K = \sum_1^3{}_{\nu\rho}\, a^{\nu\rho}\, \lambda_\nu\, \lambda_\rho = \sum_1^3{}_{\nu\rho}\, a_{\nu\rho}\, \lambda^\nu\, \lambda^\rho, \quad . \quad . \quad (40')$$

which defines the curvature of a variable section f *through* P *as a homogeneous quadric in the parameters* λ^ν *(or in the moments* λ_ν*) of the normal to* f, the sense of the normal being indifferent, since the expressions in (40′) do not change when we change the signs of the λ's.

The dependence of K on the direction $\boldsymbol{\lambda}$ of the section with centre at P is of purely local nature. We can therefore, in accordance with an observation made above, make use of the elementary criteria of analytical geometry just as if it were a question of ordinary space—we have only to take co-ordinates which are *Cartesian and orthogonal at* P. The λ^ν's then become direction cosines, and we have for K (except for a different signification of the coefficients $a_{\nu\rho}$) the same expression as the one which characterizes the distribution of moments of inertia (of an assigned material system) with respect to the ∞^2 axes coinciding with the lines of the versors $\boldsymbol{\lambda}$ proceeding from P. As we know, the law of variation of K becomes expressible geometrically if we introduce the ellipsoid of inertia E, of centre P: to any direction of $\boldsymbol{\lambda}$ there belongs a value of K, viz. $\frac{1}{PQ^2}$, where Q is the intersection of the line of $\boldsymbol{\lambda}$ with the ellipsoid E; the three axes of E correspond to stationary values of K (with respect to neighbouring directions), the axes of maximum and minimum length in particular corresponding respectively to minimum and maximum values of K.

That being so, the same law of variation will hold for K the curvature; but in the general V_3 interpretation is necessary. This implies merely that the ellipsoid E (like, for example, Dupin's indicatrix for the case of an ordinary surface) takes us in thought outside the V_3, as an auxiliary representative element to be associated with the Euclidean three-dimensional space which is tangential to the V_3 at P when, as is always allowable (Chapter V, § 21), we imagine the V_3 immersed in a Euclidean S_N $(N \geqslant 6)$.

The outstanding result is that *there exist at every point* P *at least three mutually orthogonal directions* $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$, $\boldsymbol{\lambda}_3$ *to which* (or rather to the normal sections perpendicular to which) *belong curvatures which are stationary with respect to those of adjacent sections.* These directions are called *principal directions of curvature*, and the corresponding values ω_1, ω_2, ω_3 of K *principal curvatures.* In general, that is when the three ω's are distinct, the principal directions are uniquely determinate (ellipsoid with three unequal axes); when two principal curvatures are equal but differ from the third (ellipsoid of revolution), e.g. $\omega_1 = \omega_2 \neq \omega_3$, only the principal direction $\boldsymbol{\lambda}_3$ is uniquely determinate, while every pair of directions $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$ orthogonal to $\boldsymbol{\lambda}_3$ and to each other can be considered principal. If the three principal curvatures coincide (sphere) the curvature K is the same for all sections, and every set of three mutually orthogonal directions is a principal set.

All this of course can be established by purely algebraic methods: we have only to avail ourselves of the theory of quadratic forms and their transformations. Let

$$\phi = \sum_{1}^{n}{}_{ij}\, a_{ij} \lambda^i \lambda^j, \quad \psi = \sum_{1}^{n}{}_{ij}\, a_{ij} \lambda^i \lambda^j \quad . \quad . \quad (41)$$

be two quadrics, *one of which at least is definite,* say ϕ; the independent variables being λ^i ($i = 1, 2, \ldots n$). Then the following facts are well-known[1]:

(1) If we consider the ratio

$$\omega = \frac{\psi}{\phi}$$

as a function of the λ's, and look for the values of these variables which make $\delta\omega = 0$, we are led to the system of n linear homogeneous equations

$$\sum_1^n{}_j (\alpha_{ij} - \omega a_{ij}) \lambda^j = 0 \qquad (i = 1, 2, \ldots n); \quad . \quad (42)$$

these are satisfied by values not all null of the λ's if and only if the determinant of the coefficients vanishes, so that the ω's are roots of the equation of degree n:

$$\| \alpha_{ij} - \omega a_{ij} \| = 0, \quad . \quad . \quad . \quad . \quad (43)$$

called the *characteristic equation.*

(2) If λ^i, λ'^i are two solutions of equations (42) corresponding to two distinct roots ω, ω' of the characteristic equation, there exists between them the relation (of orthogonality)

$$\sum_1^n{}_{ij} a_{ij} \lambda^i \lambda'^j = 0.$$

(3) The characteristic equation (43) has its n roots $\omega_1, \omega_2, \ldots \omega_n$ *all real* (distinct or coincident).

(4) To each simple root ω_h of (43) corresponds, in the manifold V_n the ds^2 of which has the a_{ij}'s as its coefficients, one and only one direction $\boldsymbol{\lambda}_h$ whose parameters λ_h^i satisfy (42) for $\omega = \omega_h$. With each root of multiplicity μ (> 1) we can in an infinite number of ways associate μ mutually orthogonal directions in V_n, whose parameters are independent solutions of equations (42) when we give ω the value of the said multiple root.

From all this there results that it is possible in every case to set up at least one set of n mutually orthogonal directions $\boldsymbol{\lambda}_h$ ($h = 1, 2, \ldots n$), (uniquely determined in the general case, in which the roots of (43) are all simple), the parameters λ_h^j of which satisfy the system (42), with ω equal to the corresponding ω_h.

---

[1] For proofs, see for example: BIANCHI, *Lezioni di geometria analitica,* Appendix; (Pisa: Spoerri, 1915): or BROMWICH, *Quadratic Forms and their Classification* . . . (Cambridge tracts on Mathematics . . . No. 3).

By means of these $\boldsymbol{\lambda}_h$'s we can obtain the so-called canonical expressions. We begin by introducing the moments

$$\lambda_{h|i} = \sum_1^n{}_j\, a_{ij}\, \lambda_h^j \qquad (h, i = 1, 2, \ldots n), \quad . \quad (44)$$

and observe that, by associating with the n identities

$$\sum_1^n{}_j\, \lambda_h^j\, \lambda_{h|j} = 1 \qquad (h = 1, 2, \ldots n)$$

the $\frac{1}{2}n(n - 1)$ conditions of orthogonality of two different directions $\boldsymbol{\lambda}_h$, $\boldsymbol{\lambda}_k$ of the n-ple set

$$\sum_1^n{}_j\, \lambda_h^j\, \lambda_{k|j} = 0 \qquad (h \neq k),$$

we obtain altogether the n^2 relations

$$\sum_1^n{}_j\, \lambda_h^j\, \lambda_{k|j} = \delta_h^k \qquad (h, k = 1, 2, \ldots n),$$

which express the noteworthy fact that *the* n^2 *parameters* λ_h^j *of an* n*-ple orthogonal set are the elements reciprocal (in an algebraic sense) to the* n^2 *moments in the determinant* $\| \lambda_{h|j} \|$ *which they form; and vice versa that the moments are the reciprocals of the parameters in the corresponding determinant* $\| \lambda_h^j \|$ (cf. Chapter IV, § 6).

Further, besides the relations just written which refer to the columns, their analogues with respect to the rows also hold good; these may be written

$$\sum_1^n{}_h\, \lambda_h^j\, \lambda_{h|i} = \delta_i^j \qquad (i, j = 1, 2, \ldots n). \quad . \quad (45)$$

Taking account of this, if we multiply (44) by $\lambda_{h|k}$ and sum with respect to h, there results immediately

$$a_{ik} = \sum_1^n{}_h\, \lambda_{h|i}\, \lambda_{h|k} \qquad (i, k = 1, 2, \ldots n), \quad . \quad (46)$$

which are *expressions for the fundamental tensor* a$_{ik}$ *in terms of the moments of any orthogonal* n*-ple set.*

Consider in particular the n-ple set (or one of the n-ple sets) $\boldsymbol{\lambda}_h$, the parameters of which, λ_h^j, satisfy (42) where, for each index h, $\omega = \omega_h$. These equations, in virtue of (44), may be written

$$\sum_1^n{}_j\, a_{ij}\,\lambda_h^j = \omega_h\,\lambda_{h|i} \qquad (h, i = 1, 2, \ldots n).$$

Multiplying by $\lambda_{h|k}$, and summing with respect to the index h, while attending to (45), we obtain the canonical expression for the a_{ik}'s to be associated with (46), viz.

$$a_{ik} = \sum_1^n{}_h\, \omega_h\,\lambda_{h|i}\,\lambda_{h|k} \qquad (i, k = 1, 2, \ldots n). \quad . \quad (47)$$

After this the simultaneous reduction of the two quadrics to orthogonal form is easily effected by substituting for the original variables λ^i the n linear combinations

$$z_h = \sum_1^n{}_i\, \lambda_{h|i}\,\lambda^i \qquad (h = 1, 2, \ldots n). \quad . \quad (48)$$

In fact, when we substitute in (41) the values (46), (47) of the coefficients, taking account of (48), there results

$$\phi = \sum_1^n{}_h\, z_h{}^2, \quad \psi = \sum_1^n{}_h\, \omega_h\, z_h{}^2. \quad . \quad . \quad . \quad (49)$$

The condition that the λ^i's should be parameters, that is to say that the expression (41) for ϕ should have the value 1, becomes in the new variables z, $\sum_1^n{}_h\, z_h{}^2 = 1$. The mode of variation of ψ, when the λ's are parameters of direction, is identical with that of the ratio $\frac{\psi}{\phi}$, when the variables are independent; or, if we wish, of the quadric $\psi = \sum_1^n{}_h\, \omega_h\, z_h{}^2$, when the z's are connected the relation $\sum_1^n{}_h\, z_h{}^2 = 1$.

Moreover, the stationary values of ψ in these cases are precisely the roots of the characteristic equation (43).

The form

$$\psi = \sum_1^n{}_{ij}\, a_{ij}\,\lambda^i\,\lambda^j$$

is an obvious generalization, for any value of n, of the expression (40′) for K, which defines the local distribution of curvature in a V_3.

The above considerations manifestly apply also to the behaviour of the curvature K when the direction of the section is supposed to vary, a topic which has already been discussed in a more elementary way.

13. **Geodesics infinitely near a given geodesic.**

We shall conclude this chapter with a discussion of the extension to n dimensions of a classical formula due to Jacobi, which defines in a very simple way the aggregate of those geodesics g of a surface which are infinitely near a given geodesic B, called the geodesic *base*. Jacobi gives the linear equation

$$\frac{d^2y}{d\sigma^2} + Ky = 0, \tag{J}$$

where y denotes the distance (normal) of any point M of g from the base, σ the arc of the base measured from an arbitrary origin O up to the projection P of M upon B, and $K(\sigma)$ the Gaussian curvature of the surface at P.

(J) is simply, in Poincaré's phrase, the *equation of variations* of geodesics starting from B. There can be deduced from it, as we know,[1] some very remarkable consequences with respect to the behaviour of geodesics in the immediate neighbourhood of the base, the nature of the surface intervening only through its total curvature K. This is obviously an *intrinsic* question, depending entirely on the metric of the surface (as defined by its ds^2), and not at all on the different configurations which the surface can present in space.

It naturally suggests itself that we should try to extend the study of this subject of geodesic deviation to a Riemannian manifold V_n of any number of dimensions. We have long had, of course, the equations of Lagrange defining the geodesics of a V_n, in a form the convenience of which, whether from the point of view of theory or of notation, is all that could be desired.

---

[1] See, for example, DARBOUX, *Théorie des surfaces*, Vol. III (Paris, Gauthier-Villars, 1894; new impression 1923), Chap. V; or BLASCHKE, *Vorlesungen über Differentialgeometrie*, Vol. I (2nd edition, Berlin, Springer, 1924), §§ 83–88.

These equations we may therefore use for the purpose of forming the equations of variations. Then with the help of Bianchi's idea (Chapter V, § 26) of the *derivative* of a vector attached to the points of a line of V_n, we reduce these equations to a condensed form, geometrically suggestive and of course invariant. The actual construction of the equations (linear of the second order, n in all, with the same number of unknowns) requires no further data than the base B and the metric of the manifold (especially Riemann's symbols) along that curve.

We find that this system of equations admits a linear first integral, which in its turn leads to a linear relation in finite terms among the unknowns. We are thus left with a system of $n - 1$ equations or, in the special case of an ordinary surface, with a single equation, as in Jacobi's classical result.

To bring the final system of equations to as simple a form as possible we have to make a suitable choice of variables. Now we have already seen (Chapter VI, § 11) that it is possible, beginning with any co-ordinates x, to define new co-ordinates y in terms of which ds^2 becomes locally Cartesian, in the sense that the derivatives of the coefficients a_{ik} all vanish at an assigned point O. We have also seen (Chapter VI, § 11, footnote) that it is possible, any curve B being given, to choose co-ordinates y for which ds^2 has this locally Cartesian character *at every point of* B. When B is a geodesic, the system of co-ordinates y which (§ 17) will finally be used has the following properties, which we merely state here without proof.[1] Let M be any point in the immediate neighbourhood of B, P the orthogonal projection of M upon B. Then y_n is the arc of the base B measured from an arbitrary origin O up to P; the y_α's ($\alpha = 1, 2, \ldots n - 1$) may be regarded as components of the elementary vector PM in $n - 1$ directions mutually perpendicular and all perpendicular to B, chosen arbitrarily at O and carried by parallelism along B.

14. Geodesic deviation in an n-dimensional manifold.

Consider, along with the geodesic B, any other geodesic g (more precisely, an arc of g) belonging to the immediate neighbourhood of B.[1] Corresponding to every point M of g, take the point P of B having the same y_n as M and the rest of its y's $= 0$. It is important for what follows to fix precisely the relation between an elementary arc ds of g and the corresponding arc $d\sigma = dy_n$ of B. Throughout the neighbourhood of B we have, for the coefficients b_{ik} of ds^2 in terms of the co-ordinates y, the Euclidean values

---

[1] For a complete development of this point, compare the paper "Sur l'écart géodésique" in *Mathematische Annalen* (Vol. 97).

$$b_{ik} = 0 \quad (i \neq k), \quad b_{ii} = 1,$$

neglecting quantities of the second order.

It follows that for any curve whatever

$$ds = dy_n \sqrt{1 + \sum_{1}^{n-1}{}_{a} \left(\frac{dy_a}{dy_n}\right)^2},$$

the quantity under the radical differing from its exact value only by terms of the second order. For g, both the y_a and the $\frac{dy_a}{dy_n}$ may be regarded as infinitely small. It follows that

$$\frac{ds}{dy_n} = \frac{ds}{d\sigma} = 1$$

to a second approximation—a generalization of the elementary fact that a segment infinitely near (with respect to direction) a given right line differs from its projection on the line by an infinitesimal of the second order.

15. **Invariant form of the equations defining geodesic deviation.**

We have now to form the equations defining in general co-ordinates the behaviour of any geodesic g infinitely near B.

Put

$$x_i = \phi_i(\sigma) + \xi^i, \quad . \quad . \quad . \quad . \quad (50)$$

where the ξ^i's and their derivatives with respect to σ are infinitely small.

The ξ^i's represent the increments of the co-ordinates x_i of a point P of B which passes to a corresponding point M of g: they can be regarded as the contravariant components of the elementary vector $PM = \boldsymbol{\xi}$. Adopting a more general point of view, we shall now regard this vector as not necessarily perpendicular to B, its orientation depending on the law of correspondence between the points P and M of our two geodesics. We can therefore no longer assume that the elementary arc ds of g is (as in § 14) equal to the corresponding arc $d\sigma$ of B; but, the displacement in question being always infinitely small, we can foresee all the same that, if we put

---

[1] In the *strict* sense, that is to say, with the understanding that not only are changes of position of corresponding points on B and g to be small, but also changes of direction of corresponding tangents.

$$\frac{ds}{d\sigma} = 1 + \lambda, \quad . \quad . \quad . \quad . \quad . \quad (51)$$

the *elongation* (or coefficient of dilatation) λ remains infinitely small with the ξ^i's and their derivatives. This will be proved formally in a moment. Meantime, differentiate the formulæ (50) with respect to σ. We have

$$\frac{ds}{d\sigma}\dot{x}_i = \dot{\phi}_i + \frac{d\xi^i}{d\sigma}, \quad . \quad . \quad . \quad . \quad . \quad (50')$$

where $\dot{x}_i$ is written for $\frac{dx_i}{ds}$, and $\dot{\phi}_i$ for $\frac{d\phi_i}{d\sigma}$.

A second differentiation gives

$$\left(\frac{ds}{d\sigma}\right)^2 \ddot{x}_i + \frac{d^2s}{d\sigma^2}\dot{x}_i = \ddot{\phi}_i + \frac{d^2\xi^i}{d\sigma^2}.$$

From (51),
$$\frac{d^2s}{d\sigma^2} = \frac{d\lambda}{d\sigma},$$

and (50′) can be written

$$\dot{x}_i - \dot{\phi}_i = -\lambda\dot{x}_i + \frac{d\xi^i}{d\sigma};$$

all this holds without any assumption concerning the smallness of λ.

Considering now λ *and* $\frac{d\lambda}{d\sigma}$ *as infinitesimal*, and neglecting all terms of the second order, we may replace $\frac{d^2s}{d\sigma^2}\dot{x}_i = \frac{d\lambda}{d\sigma}\dot{x}_i$ by $\frac{d\lambda}{d\sigma}\dot{\phi}_i$, and consequently take the equation for $\ddot{x}_i$ to be

$$\left(\frac{ds}{d\sigma}\right)^2 \ddot{x}_i = \ddot{\phi}_i + \frac{d^2\xi^i}{d\sigma^2} - \frac{d\lambda}{d\sigma}\dot{\phi}_i. \quad . \quad . \quad (50'')$$

Now (Chapter V, § 24) for any geodesic g we have

$$\ddot{x}_i = -\sum_1^n{}_{jh}\{jh, i\}\,\dot{x}_j\,\dot{x}_h \qquad (i = 1, 2, \ldots n).$$

Multiply these equations by $\left(\frac{ds}{d\sigma}\right)^2$, and make the following substitutions: (50) in the $\{jh, i\}$, (50′) in the $\frac{ds}{d\sigma}\,\dot{x}_j$, (50″) in the $\left(\frac{ds}{d\sigma}\right)^2\ddot{x}_i$.

Remembering that

$$\ddot{\phi}_i = -\sum_1^n{}_{jh}\{jh, i\}\,\dot{\phi}_j\,\dot{\phi}_h,$$

since B is a geodesic, and of course retaining terms of the first order only, we find

$$\frac{d^2\xi^i}{d\sigma^2} - \frac{d\lambda}{d\sigma}\dot{\phi}_i = -\sum_1^n{}_{jhk}\frac{\partial\{jh, i\}}{\partial x_k}\,\xi^k\,\dot{\phi}_j\,\dot{\phi}_h - 2\sum_1^n{}_{jh}\{jh, i\}\,\dot{\phi}_j\,\frac{d\xi^h}{d\sigma}.$$

Next, for the sake of showing explicitly the contravariance of the $\dot{\phi}_i$ (parameters of the direction tangential to B), put

$$\dot{\phi}_i(\sigma) = b^i \qquad (i = 1, 2, \ldots n). \quad . \quad . \quad (52)$$

The equations just obtained (if we also change the indices i and j into r and i) become

$$\frac{d^2\xi^r}{d\sigma^2} - \frac{d\lambda}{d\sigma}b^r + 2\sum_1^n{}_{jh}\{jh, r\}b^j\frac{d\xi^h}{d\sigma}$$
$$= -\sum_1^n{}_{ihk}\frac{\partial\{ih, r\}}{\partial x_k}\,b^i\,b^h\,\xi^k. \quad (53)$$

We proceed to transform these equations for the sake of showing their invariant structure with reference to any change of coordinates. For this purpose we introduce Bianchi's conception of the *derivative* vector of a vector $\boldsymbol{\xi}$ given as a function of the points of a line (Chapter V, § 26).

If B is the line, the contravariant components of the vector $D\boldsymbol{\xi}$, the derivative of $\boldsymbol{\xi}$, are given by the equations

$$(D\boldsymbol{\xi})^r = \frac{\partial\xi^r}{\partial\sigma} + \sum_1^n{}_{ih}\{ih, r\}\,b^i\,\xi^h \qquad (r = 1, 2, \ldots n). \quad (54)$$

For co-ordinates which are Cartesian either rigorously, or else locally along the line B, we have simply

$$(D\boldsymbol{\xi})^r = \frac{d\xi^r}{d\sigma}.$$

With the help of the vector $\boldsymbol{\xi}$, it is easy to give an explicit expression for the elongation λ, without making any hypothesis as to its order of magnitude, the equations (50) and (50′), which we shall use, being exact.

On g we have identically

$$\overset{n}{\underset{1}{\Sigma}}{}_{ir}\, a_{ir}\, \dot{x}_i\, \dot{x}_r = 1.$$

If we multiply on both sides by $\left(\frac{ds}{d\sigma}\right)^2$, and replace (in the a_{ir}'s) the x_i's by their values (50) and the $\frac{ds}{d\sigma}\dot{x}_i$'s by (50′), we find at once (neglecting terms of the second order with respect to ξ)

$$\left(\frac{ds}{d\sigma}\right)^2 = 1 + 2\overset{n}{\underset{1}{\Sigma}}{}_{ir}\, a_{ir}\, \dot{\phi}_i \frac{d\xi^r}{d\sigma} + \overset{n}{\underset{1}{\Sigma}}{}_{irh} \frac{\partial a_{ir}}{\partial x_h} \dot{\phi}_i\, \dot{\phi}_r\, \xi^h.$$

We have, moreover, the well-known identities (Chapter V, § 16)

$$\frac{\partial a_{ik}}{\partial x_l} = \overset{n}{\underset{1}{\Sigma}}{}_h \left[a_{hi}\{kl, h\} + a_{hk}\{il, h\}\right].$$

If in the last sum of the expression for $\left(\frac{ds}{d\sigma}\right)^2$ we replace the indices r, h by k, l and take account of the identity just written down, replacing also, as in (52),

$$\overset{n}{\underset{1}{\Sigma}}{}_i\, a_{ir}\, \dot{\phi}_i, \quad \overset{n}{\underset{1}{\Sigma}}{}_i\, a_{hi}\, \dot{\phi}_i, \quad \overset{n}{\underset{1}{\Sigma}}{}_k\, a_{hk}\, \dot{\phi}_k,$$

$$\text{by } b_r,\ b_h,\ b_h,$$

we find, on account of (54),

$$\left(\frac{ds}{d\sigma}\right)^2 = 1 + 2\overset{n}{\underset{1}{\Sigma}}{}_r\, (D\boldsymbol{\xi})^r\, b_r.$$

The vector $D\boldsymbol{\xi}$ is infinitely small at the same time as the ξ^i's and their derivatives. We have, therefore, by extracting the root, neglecting terms of the second order, and attending to the definition (51) of the elongation λ,

$$\lambda = \overset{n}{\underset{1}{\Sigma_r}}\, (D\boldsymbol{\xi})^r\, b_r, \qquad . \quad . \quad . \quad . \quad . \quad (51')$$

which shows its infinitesimal character.

Naturally $D\boldsymbol{\xi}$ admits in its turn a derivative vector $D^2\boldsymbol{\xi}$. Its contravariant components are defined, from (54), by

$$(D^2\boldsymbol{\xi})^r = \frac{d(D\boldsymbol{\xi})^r}{d\sigma} + \overset{n}{\underset{1}{\Sigma_{kl}}}\, \{kl, r\}\, b^k\, (D\boldsymbol{\xi})^l.$$

Introducing on the right-hand side the expressions (54) for $(D\boldsymbol{\xi})^r$ and $(D\boldsymbol{\xi})^l$ (after making some literal changes in the indices), we obtain

$$(D^2\boldsymbol{\xi})^r = \frac{d^2\xi^r}{d\sigma^2} + \frac{d}{d\sigma}\left(\overset{n}{\underset{1}{\Sigma_{ih}}}\, \{ih, r\}\, b^i\, \xi^h\right)$$
$$+ \overset{n}{\underset{1}{\Sigma_{kl}}}\, \{kl, r\}\, b^k \frac{d\xi^l}{d\sigma} + \overset{n}{\underset{1}{\Sigma_{ihkl}}}\, \{kl, r\}\, \{ih, l\}\, b^k\, b^i\, \xi^h, \quad (55)$$

the second members, like the first, constituting a contravariant system. Bringing out explicitly the differentiation with respect to σ, and making some changes of indices, we can write

$$(D^2\boldsymbol{\xi})^r = \frac{d^2\xi^r}{d\sigma^2} + 2\,\overset{n}{\underset{1}{\Sigma_{jh}}}\, \{jh, r\}\, b^j \frac{d\xi^h}{d\sigma} + \Xi^{(r)}, \qquad (55')$$

where we put for brevity

$$\Xi^{(r)} = \overset{n}{\underset{1}{\Sigma_{ihk}}}\, \frac{\partial\{ik, r\}}{\partial x_h}\, b^i\, b^h\, \xi^k + \overset{n}{\underset{1}{\Sigma_{lk}}}\, \{lk, r\}\, \xi^k \frac{db^l}{d\sigma}$$
$$+ \overset{n}{\underset{1}{\Sigma_{ihkl}}}\, \{lh, r\}\, \{ik, l\}\, b^i\, b^h\, \xi^k.$$

We may note in passing that, in the auxiliary quantities $\Xi^{(r)}$, the index r is purely ordinal: we have placed it above, but in brackets, so as to avoid the suggestion that the $\Xi^{(r)}$'s form a contravariant system, which they do not.

For our purpose it is sufficient to replace in $\Xi^{(r)}$ the derivatives $\frac{db^l}{d\sigma} = \ddot{\phi}_l$ by their values

$$-\overset{n}{\underset{1}{\Sigma}}_{ih}\,(ih,\,l)\,b^i\,b^h,$$

as given by the equations for geodesics (Chapter V, § 24), so that we may write

$$\Xi^{(r)} = \overset{n}{\underset{1}{\Sigma}}_{ihk}\,\frac{\partial\,\{ik,\,r\}}{\partial\,x_h}\,b^i\,b^h\,\xi^k$$
$$+\overset{n}{\underset{1}{\Sigma}}_{ilhk}\,[\{lh,\,r\}\,\{ik,\,l\} - \{lk,\,r\}\,\{ih,\,l\}]\,b^i\,b^h\,\xi^k. \quad (56)$$

If then we add $\Xi^{(r)}$ to the two members of equations (53), attending to (55′) and to the definition of Riemann's symbols (§ 2), the equations take the *invariant form*

$$(D^2\boldsymbol{\xi})^r - \frac{d\lambda}{d\sigma}\,b^r = -\overset{n}{\underset{1}{\Sigma}}_{ihk}\,\{ir,hk\}\,b^i b^h\,\xi^k \qquad (r = 1, 2, \ldots n). \quad (57)$$

16. Geodesic deviation. Specification of the differential system. First integral. Linear relation in finite terms.

The system (57), taken along with the value of λ given in (51′), contains $n + 1$ equations, with the same number of unknowns; but it is easy to foresee, from the method of obtaining it, that it cannot by itself determine completely all the unknowns; there ought to remain an element of indeterminateness arising from the arbitrary nature of the law of correspondence between the points P of B and M of g. More definitely, we can prove that the definition (51′) of λ, or rather the relation derived from it by differentiation

$$\frac{d}{d\sigma}\left\{\lambda - \overset{n}{\underset{1}{\Sigma}}_r\,(D\boldsymbol{\xi})^r\,b_r\right\} = 0 \quad . \quad . \quad . \quad (51'')$$

is a consequence of equations (57) themselves. To establish this, note first that, for any vector $\mathbf{v}$ whose contravariant components are v^r, we have

$$\frac{d}{d\sigma}\overset{n}{\underset{1}{\Sigma}}_r\,v^r\,b_r = \overset{n}{\underset{1}{\Sigma}}_r\,\frac{dv^r}{d\sigma}\,b_r + \overset{n}{\underset{1}{\Sigma}}_r\,v^r\,\frac{db_r}{d\sigma}.$$

But the derivatives $\frac{db_r}{d\sigma}$ of the moments of a geodesic B satisfy the equations (Chapter V, § 24 and § 26 (52′))

$$\frac{db_r}{d\sigma} = \overset{n}{\underset{1}{\Sigma}}{}_{ih} \{ir, h\} b^i b_h,$$

which express, we may say, the autoparallelism of the geodesic B in terms of its moments b_r. The preceding identity, after interchange of the indices h and r in the last sum, therefore takes the form

$$\frac{d}{d\sigma} \overset{n}{\underset{1}{\Sigma}}{}_r v^r b_r = \overset{n}{\underset{1}{\Sigma}}{}_r (D\mathbf{v})^r b_r. \qquad . \quad . \quad . \qquad (58)$$

In virtue of this equation, the first member of (51″) now becomes (if we replace the vector $\mathbf{v}$ by $D\boldsymbol{\xi}$)

$$\frac{d\lambda}{d\sigma} - \overset{n}{\underset{1}{\Sigma}}{}_r (D^2\boldsymbol{\xi})^r b_r.$$

That this expression vanishes we can easily prove from equation (57), making use of the properties of Riemann's symbols, as follows. Multiply both sides of (57) by b_r and sum with respect to r, noting that $\overset{n}{\underset{1}{\Sigma}}{}_r b_r b^r = 1$. The right-hand member can be written

$$- \overset{n}{\underset{1}{\Sigma}}{}_{ihkrp} \{ir, hk\} a_{pr} b^p b^i b^h \xi^k.$$

We now sum with respect to r, thus changing the Riemann symbol to the symbol of the first kind, and then make use of the antisymmetry of (ip, hk) in i and p (§ 4), from which it follows that the sum is zero.

Equation (51′) is therefore simply a particular integral (or invariant relation) of the system (57); its role reduces to that of fixing one of the constants of integration. As the system (57) contains $n + 1$ unknowns, to make it determinate it is necessary to associate with it some other condition—a circumstance easy to understand from the geometrical point of view, since we have still to fix the law of punctual correspondence between g and the base.

From a formal point of view the easiest way to complete the system (57) is to cut out the unknown λ by putting

$$\frac{d\lambda}{d\sigma} = 0.$$

The equations (57) are thus reduced to the normal form (Chapter III, § 2)

$$(D^2\boldsymbol{\xi})^r = -\sum_{1}^{n}{}_{ihk}\{ir, hk\}\, b^i b^h \xi^k \qquad (r = 1, 2, \ldots n); \quad \text{(I)}$$

and we see that, on account of the identity (51″), *the system* (I) *admits the first integral*

$$\sum_{1}^{n}{}_r (D\boldsymbol{\xi})^r b_r = \lambda = \text{constant}, \quad . \quad . \quad . \quad \text{(II)}$$

expressing the fact that there is a constant linear dilatation when we pass from any arc of B to the corresponding arc of g.

Since, on account of the identity (58), the first member of the integral (II) is simply the derivative of $\sum_{1}^{n}{}_r \xi^r b_r$, it follows that every solution of the differential system (I) gives also

$$\sum_{1}^{n}{}_r \xi^r b_r = \lambda\sigma + C, \quad . \quad . \quad . \quad . \quad \text{(III)}$$

where C is a second constant.

If in particular we take $\lambda = 0$, we see that we can associate with the differential system the relation

$$\sum_{1}^{n}{}_r \xi^r b_r = C.$$

This gives the translation into analysis of the obvious geometrical fact that we can assign the correspondence between the points M of g and P of B in such a way that the (infinitely small) vector PM will have its orthogonal projection upon the tangent $\mathbf{t}$ to the base at the point P equal to a constant C. Such a law of correspondence implies, in virtue of (II), that there is no alteration of length ($\lambda = 0$) as between the arcs of B and the corresponding arcs of g. To particularize still further, if $C = 0$, we arrive at the orthogonal law of correspondence (PM perpendicular to B) considered in § 13.

It is scarcely necessary to add that, in order to substitute other geometrical laws of correspondence, we have only to associate with the system (57) the analytical translation of the law chosen, instead of the law $\frac{d\lambda}{d\sigma} = 0$. For example, if we wish PM to be inclined to B at an angle ψ (constant or a given function of σ) the additional equation will be

$$\sum_{1}^{n}{}_{r}\, \xi^{r} b_{r} = \xi \cos\psi,$$

where
$$\xi = \left| \sqrt{\sum_{1}^{n}{}_{ik}\, a_{ik}\, \xi^{i} \xi^{k}} \right|$$

represents the length of the vector $\boldsymbol{\xi}$.

In a case like this, some slight supplementary discussion of the complete system will be needed—its reduction to the normal form, determination of the number of constants of integration, &c.

17. Reduced form of the differential system (I) in terms of the co-ordinates y.

We now return to the co-ordinates y, and fix definitely on the orthogonal law of correspondence between the base B and any geodesic in its neighbourhood. As we have just seen, such a correspondence is expressed analytically by the differential system (I), with the specifications $\lambda = C = 0$ of the constants of integration connected with (II) and (III).

As remarked in § 14, the co-ordinate y_n of M is identical with that of P. Since the other co-ordinates y_α $(\alpha = 1, 2, \ldots n-1)$ of P are 0, the variations η^i of the co-ordinates y are respectively

$$\eta^{\alpha} = y_{\alpha} \quad (\alpha = 1, 2, \ldots n-1), \quad \eta^{n} = 0,$$

thus justifying the name of Cartesian components of the normal displacement or *deviation* $PM = \boldsymbol{\eta}$ which we give to the y_α's.

Moreover, the parameters $b^i = \frac{dy_i}{d\sigma}$ of the base B vanish for $i = 1, 2, \ldots n-1$; and $b^n = 1$. Christoffel's symbols also vanish along B, as well as their first derivatives with respect to

σ (or, what comes to the same thing, to y_n), and consequently

$$(D\boldsymbol{\eta})^r = \frac{d\eta^r}{d\sigma}.$$

Equations (I) thus become

$$\frac{d^2y_\alpha}{d\sigma^2} = -\sum_1^{n-1}{}_\beta \{n\alpha, n\beta\} y_\beta \quad (\alpha = 1, 2, \ldots n-1), \quad \text{(I')}$$
$$0 = -\sum_1^{n-1}{}_\beta \{nn, n\beta\} y_\beta,$$

where, in both sums, we have suppressed the term corresponding to the value n of the index, since every Riemann's symbol which has its last indices equal vanishes (§ 3).

The first group (I′) (comprising $n-1$ linear equations of the second order) *defines the* n — 1 *Cartesian components of the* (*normal*) *deviation* PM. The last equation reduces to an identity, as may be seen as follows. Riemann's symbols of the second kind are in all cases connected to those of the first kind by the relations

$$(ij, hk) = \sum_1^n{}_r\, a_{jr} \{ir, hk\};$$

we have, moreover, for the symbols of the first kind (§ 4),

$$(ij, hk) = -(ji, hk).$$

In our case the coefficients a_{jr} of ds^2 reduce, *on B*, to 0 (for $r \neq j$) and to 1 (for $r = j$). We have therefore, on the base, equality between symbols of the two kinds whose indices are the same; and, in particular,

$$\{nn, n\beta\} = (nn, n\beta) = 0. \qquad \text{Q.E.D.}$$

Of course, the integral relations (II) and (III) are of no further account, being now identities on account of the vanishing of the b^i's ($i = 1, 2, \ldots n-1$) and of the nth component $\xi^n = \eta^n$ of the displacement PM.

18. Case of $n = 2$ — formula of Jacobi.

For $n = 2$, that is to say for an ordinary surface, if B is the geodesic base, $y_2 =$ the arc σ, and y_1 ($= y$) the normal distance from M to B, the system (I′) reduces to the single equation

$$\frac{d^2y}{d\sigma^2} = -\{21, 21\} y.$$

Now (§ 9) for any co-ordinates whatever, the Gaussian curvature K of a V_2 is expressed by the ratio

$$\frac{(12,\,12)}{a} = \frac{(21,\,21)}{a},$$

a denoting the discriminant of the ds^2 of V_2.

For our co-ordinates, which are Cartesian along B, $a = 1$, and Riemann's symbols of the second kind are (§ 17) the same as their homologues of the first kind. The equation defining y is therefore none other than the equation of Jacobi (§ 13)

$$\frac{d^2y}{d\sigma^2} + Ky = 0.$$

CHAPTER VIII

RELATIONS BETWEEN TWO DIFFERENT METRICS REFERRED TO THE SAME PARAMETERS; MANIFOLDS OF CONSTANT CURVATURE

1. **Differences between Christoffel's symbols relative to two different metrics assigned to the same analytical manifold.**

We introduced in Chapter IV notions of *tensor*, *covariance*, &c., relative to an analytical manifold V_n, i.e. to the aggregate of n variables $x_1,\ x_2,\ \ldots\ x_n$; we then, in the third part of the following chapter, considered the metrical manifolds obtained by associating with an analytical manifold V_n a specified (but arbitrary) positive and definite differential quadratic form.

There is clearly no reason against assigning in turn to the same analytical manifold two distinct metrical determinations, defined by the two quadratic forms[1]

$$ds^2 = \sum_1^n{}_{ik}\, a_{ik}\, dx_i\, dx_k, \quad . \quad . \quad . \quad . \quad (1)$$

$$ds'^2 = \sum_1^n{}_{ik}\, a'_{ik}\, dx_i\, dx_k \quad . \quad . \quad . \quad . \quad (1')$$

---

[1] As a geometrical interpretation, we can think of two distinct V_n's whose points are in one-to-one correspondence, so that a set of n values assigned to $x_1, x_2, \ldots x_n$ can be represented either by a point P of one, or by the corresponding

3 +,-.+ + +/ ! + ')))%*)' "01
3 +,-.+ + +/ ! + ')))%*)' 2"
3% +,-.+ + +/ ! + ' 6 7 8 3 ! ! !
! " "! ! ! ! 9 ! " $'% "01
3) +,-.+ + +/ ! + ' 6 7 8 3 ! ! !
! " "! ! ! ! 9 ! " $'% 2"
($!2

Chapter 8

GÖDEL'S THEOREMS ON THE INHERENT LIMITATIONS OF FORMAL SYSTEMS

1 Gödel's new metamathematical method

We come now, in this chapter, to the celebrated metamathematical theorems of Gödel, which first revealed a fundamental disparity between formal axiomatic theories, of the type considered by Hilbert, and the intuitive mathematics that they are constructed to represent. The phase of formalism that lasted until 1930, and that was dominated by Hilbert's constructive work on metamathematics or theory of proof (*Beweistheorie*), can now be seen as a comparatively naive phase—intensely creative, and productive of the basic conceptions and techniques on which most later work has been based, but at the same time relatively uncritical of its own foundations and ignorant of its inherent limitations. By 1930, however, enough progress had been made on the constructive side for the process of formalization itself to be a possible object of critical study; and in a terse but incisive paper,[1] distinguished alike by astonishing originality, profundity of conception, and mastery of intricate detail, Gödel carried metamathematics over at a single stride into its second and more reflective phase.

Hilbert himself, in originally conceiving the idea of the formalist programme, had already taken a decisive step in the study of the formal aspect of mathematics. Even before this, the desirability of formalizing mathematical theories had been recognized by Peano, who had evolved a suitable technique for the purpose, but for Peano formalization was no more than an essential aid to rigour within mathematics. Hilbert, on the contrary, formalized mathematical theories in order to turn them into well-defined objects of discussion, thus making possible the new kind of investigation to which he gave the name 'metamathematics'. In his *Beweistheorie*, which we have outlined in the previous chapter, the mathematics is treated formally while the metamathematical discussion, though kept strictly within the bounds of finitary argument, proceeds in a more intuitive way. The further step which Gödel took was to devise a treatment of metamathematics that is itself essentially formal in character; and this new advance was made possible by a simple but ingenious device that he introduced for the 'arithmetization' of metamathematical reasoning.

---

[1] K. Gödel: 'Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I', *Monatsh Math. Phys.*, **38** (1931), 173–198.

A formalized mathematical theory, as understood by Hilbert, is an edifice of symbolic expressions, made up of standard symbols that are used again and again, and such that the constitution of its component expressions (the well-formed formulae) and the manner in which these expressions may be assembled into derivations are regulated by precisely formulated syntactic rules. What Gödel realized was that, from a theoretical point of view, the particular nature of the signs that are chosen is totally irrelevant, so that we may as well adopt a standard system of 'coding', in which each symbol that is used is represented by a fixed natural number. Any formula, which is simply a finite string of symbols, will then be represented by a corresponding string of numbers; and by using the key to the code we shall be able to get back from the string of numbers to the formula. The coding can be arranged, moreover, in such a way that the rules which govern both the formation of well-formed formulae and the arrangement of such formulae in derivations can be expressed in terms of simple arithmetical operations on the representative numbers. When this is done, the metamathematical assertions that we are mostly interested in no longer require formulation in terms that refer specifically to the particular formal theory to which they relate, but they are all of a single uniform type, being without exception arithmetical propositions. Thus metamathematics, no less than the mathematics which forms its object, can be formalized, and it then becomes much more sharply defined than it was previously, when no limitation was imposed on it beyond that of being finitary.

When the new metamathematical technique is applied with formalized arithmetic itself as the primary theory, the way is made open for a most significant development. Since the metamathematics is now of the same nature as the mathematics, both of them consisting of arithmetical propositions, every metamathematical assertion can be expressed within the primary theory (provided that this is sufficiently comprehensive); and in this way the metamathematics of the formal system becomes formalizable within that system itself. A situation is thus created which allows of the sort of exploitation of self-reference that formed the basis of Cantor's proof that no class is equipotent with the class of all its subclasses, and also of Richard's paradoxical argument (see p. 127) and of the even simpler paradox of the man who declares 'The statement that I am now making is untrue'. Seizing upon this circumstance, Gödel was able to show that, if a system of formalized arithmetic is wide enough, then (i) the system is necessarily incomplete, in the sense that there exists a formula $\mathfrak{A}$ of the system such that neither $\mathfrak{A}$ nor its negation is derivable, and (ii) if the system is consistent, then no proof of its consistency is possible which can be formalized within it.

2 Gödel's heuristic argument

In the introduction to his paper of 1931, Gödel explained the general idea of his method by means of the following heuristic argument, formulated in relation to the logico-mathematical system of *Principia Mathematica*.

Consider all the formulae $R(x)$ which involve a single free variable x of the same type as the natural numbers. These formulae can be arranged, lexicographically for example, as a sequence $\{R_n(x)\}$.[1] Now the statement 'The formula $R_n(n)$, obtained when "n" is substituted for "x" in the nth formula of the sequence, is not derivable within the system' can itself be formalized within the system, and it is accordingly represented there by some formula $S(n)$. But the formula $S(x)$ then involves a single free variable x of the same type as the natural numbers, and therefore $S(x)$ coincides with $R_q(x)$ for some q. Thus $R_q(n)$ is a formal equivalent of 'The formula $R_n(n)$ is not derivable'; and consequently the formula $R_q(q)$ means 'The formula $R_q(q)$ is not derivable'. We therefore have, in $R_q(q)$, a formula that asserts its own underivability.

It now readily follows that this formula $R_q(q)$ is undecidable in the formal system. For suppose, first of all, that $R_q(q)$ were derivable. Then the proposition '$R_q(q)$ is underivable', expressed by this derivable formula, would be true. The assumption that $R_q(q)$ is derivable is therefore untenable. But now let us suppose that the negation $\sim R_q(q)$ of $R_q(q)$ were derivable. In that case the corresponding proposition 'not-[$R_q(q)$ is underivable]' would be true, and $R_q(q)$ would thus be derivable. If we assume that the formal system is consistent, we once again have an impossibility. Thus neither $R_q(q)$ nor $\sim R_q(q)$ can be derivable, and $R_q(q)$ is accordingly an undecidable formula of the system.

As Gödel says, this heuristic argument is plainly lacking in rigour, not only because we take for granted in it that the metamathematics of *Principia Mathematica* can be formalized within the *Principia* itself, but also because essential use is made of the non-formal semantic concept of truth of a proposition. In order to eliminate these deficiencies, Gödel proceeded in the body of his paper to develop the same line of argument fully and in a strictly formal manner, basing it not on *Principia Mathematica* itself but on a slightly different formal system, namely an arithmetical system P that he constructed by combining together the extended calculus of predicates (with the simple theory of types) and Peano's axioms for the natural numbers.

---

[1] The convention on the use of symbols that we have adopted in earlier chapters would require $\{R_n(x)\}$ to be written as $\{\mathfrak{R}_n(x)\}$; but in this chapter we prefer to adhere closely to the usage of Gödel in order to facilitate comparison of our summary with his original paper.

3 The formal system P and its arithmetized metamathematics

In summarizing Gödel's main argument, we must begin with the specification of the formal system P on which it is based. The primitive symbols of this system include both constants and variables, as follows:
I. Constants: '$\sim$' (not), '$\vee$' (or), 'Π' (for all), '0' (zero), 'f' (the successor of), and the two brackets '(' and ')'.
II. Variables: variables of the first type 'x_1', 'y_1', 'z_1', . . . (for individuals, i.e. natural numbers); variables of the second type 'x_2', 'y_2', 'z_2', . . . (for classes of individuals); variables of the third type 'x_3', 'y_3', 'z_3', . . . (for classes of classes of individuals); and so forth.[1]

In place of the symbolic expressions ϕx and $(x)\phi x$ that are used in *Principia Mathematica* for the propositional function 'x has the property ϕ' and the proposition 'Every individual has the property ϕ', we have in the system P the expressions $x_2(x_1)$ and $x_1\Pi(x_2(x_1))$.

The primitive symbols are to be supplemented by further constant symbols used as abbreviations. In particular, the symbols '.', '$\supset$', '$\equiv$', '(Ex)', '$=$' are used as '.', '$\supset$', '$\equiv$', '$(\exists x)$', '$=$' are used in *Principia Mathematica.*

By an *expression (Zeichen) of the first type* is meant an expression of one of the forms $a, fa, ffa, fffa, \ldots$, where a is either 0 or else a variable of the first type. An *expression of the nth type* ($n > 1$) can only be a variable of that type.

An *elementary formula* is an expression $a(b)$, where b is an expression of some type n and a is an expression of the next higher type $n + 1$.

The class of *formulae* is the smallest class that contains all the elementary formulae and is closed with respect to negation, disjunction, and universal quantification. By this we mean that, if a and b are formulae and x is any variable, then $\sim(a)$, $(a) \vee (b)$, and $x\Pi(a)$ are also formulae.

A formula with no free variables is said to be a *propositional formula.* A formula with exactly n free individual variables and no other free variable is said to be a *relation-formula with n arguments.* A relation-formula with one argument is also called a *class-formula.*

If a is any formula, v is a variable, and b is an expression of the same type as v, we are to understand by

$$\text{Subst } a \begin{pmatrix} v \\ b \end{pmatrix}$$

---

[1] No primitive symbols are needed for predicates. A predicate with one argument can be interpreted as a class of individuals, a predicate with two arguments as a class of ordered pairs of individuals, and so on; and an ordered pair $a \downarrow b$ can in turn be identified with a class of classes, namely $\{\{a\},\{a, b\}\}$ (cf. p. 171).

the formula which results when b is substituted for all the free occurrences (if there are any) of v in a.

A formula a is said to be obtained from a formula b by *elevation of type* if it is obtained when the type of every variable that occurs in b is raised by the same amount.

The axioms of the system P are arranged in five groups as follows:

I. 1. $\sim(fx_1 = 0)$,
2. $fx_1 = fy_1 \supset x_1 = y_1$,
3. $x_2(0) . x_1\Pi(x_2(x_1) \supset x_2(fx_1)) \supset x_1\Pi(x_2(x_1))$.

II. Every formula that can be obtained from the following schematic forms by substituting arbitrary formulae for p, q, r:

1. $p \vee p \supset p$,
2. $p \supset p \vee q$,
3. $p \vee q \supset q \vee p$,
4. $(p \supset q) \supset (r \vee p \supset r \vee q)$.

III. Every formula that can be obtained from the following schematic forms by making substitutions of the kind specified below for a, b, c, v, and, in the case of 1, carrying out the substitution indicated:

1. $v\Pi(a) \supset \text{Subst } a \begin{pmatrix} v \\ c \end{pmatrix}$,
2. $v\Pi(b \vee a) \supset b \vee v\Pi(a)$.

For a we may substitute an arbitrary formula, for v an arbitrary variable, for b a formula in which v does not occur free, and for c an expression of the same type as v, provided that c contains no variable that is bound in a where v is free.

IV. Every formula that can be obtained from the schematic form

1. $(Eu)(v\Pi(u(v) \equiv a))$

by substituting for v and u variables of types n and $n + 1$ respectively, and for a a formula in which u does not occur free. (This axiom, which is related to the 'principle of comprehension' for classes, as formulated below on p. 285, takes the place in the system P of the axiom of reducibility in *Principia Mathematica*.)

V. The following formula, and every formula that is obtainable from it by elevation of type:

1. $x_1\Pi(x_2(x_1) \equiv y_2(x_1)) \supset x_2 = y_2$.

This axiom asserts that a class is completely determined by its members; cf. the axiom of determination in the theory of sets, as stated on p. 288 below.

No rule of substitution is required, since substitutions are already taken care of by the adoption of axiom-schemata in preference to specific axioms (cf. p. 44). The other rules of procedure are incorporated in the system P in the following manner: A formula c is said to be an *immediate consequence* of a and b (or of a) if a is the formula $(\sim(b)) \vee (c)$ (or if c is the formula $v\Pi(a)$, where v is any variable). The class of *derivable formulae* is defined as the smallest class of formulae which contains the axioms and is closed with respect to the relation of immediate consequence.

Exercise. Rewrite the specification of Gödel's system P in the notation of the present book, using German letters as syntactic variables. Compare Gödel's axioms with those of (a) *Principia Mathematica*, (b) Peano's treatment of arithmetic, (c) Hilbert's formalization of logic and arithmetic, as given above in Chapters 3 and 7.

In order to arithmetize the metamathematics of the system P, Gödel had first of all to rewrite the entire formal system, using numbers everywhere in place of the special symbols $\sim$, $\vee$, Π, 0, f, $(\,,)$, x_1, $y_1, \ldots$; and this he did in principle by defining what has since come to be known as a *Gödel numbering* of the formal expressions. He first attached numbers to the constant primitive symbols in accordance with the following table:

Symbol	0	f	$\sim$	$\vee$	Π	(	)
Number	1	3	5	7	9	11	13

Then, making use of the natural sequence of prime numbers $p_1, p_2, p_3, \ldots$ (i.e. the sequence $2, 3, 5, \ldots$), he assigned to the variables $x_n, y_n, z_n, \ldots$ of the nth type the numbers $p_7^n, p_8^n, p_9^n, \ldots$, i.e. the numbers $17^n, 19^n, 23^n, \ldots$.

The coding of the symbols that is defined in this way is such that every primitive symbol of the system has a well-defined Gödel number, and distinct symbols always have distinct numbers. To any finite sequence $a_1, a_2, \ldots, a_k$ of primitive symbols there then corresponds a finite sequence $n_1, n_2, \ldots, n_k$ of natural numbers; and with this latter sequence we may associate in turn the single natural number $p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$. By this process a one-one mapping is set up between the totality of finite sequences a of primitive symbols on the one hand, and a certain subset of the set of natural numbers on the other. We may denote this mapping by $a \longleftrightarrow \Phi(a)$. Then, since every formula a of the system P may be treated as a mere string of symbols, it has a well-defined Gödel number $\Phi(a)$.

EXAMPLE. Consider axiom I 2, which runs

$$fx_1 = fy_1 \supset x_1 = y_1.$$

Written in terms of primitive symbols only, without use of abbreviations, this becomes

$$\sim(x_2\Pi(\sim x_2(fx_1) \vee x_2(fy_1))) \vee x_2\Pi(\sim x_2(x_1) \vee x_2(y_1)),$$

which is a string of 34 symbols. From this string we obtain a sequence of 34 numbers 5, 11, 17^2, 9, 11, 5, $17^2, \ldots$, 13; and the Gödel number of the formula is therefore

$$p_1^5 p_2^{11} p_3^{17^2} \cdots p_{34}^{13}.$$

Not every integer corresponds to a well-formed formula, of which it is the Gödel number, or even to a string of primitive symbols; but if a number is the Gödel number of a formula, this formula can actually be found by resolution of the number into its prime factors.

The Gödel coding can readily be extended from single formulae to finite sequences of formulae—among which, of course, are included derivations. Let $a_1, \ldots, a_l$ be given formulae, and let the Gödel number $\Phi(a_i)$ of a_i be m_i, $i = 1, \ldots, l$. Then we assign to the sequence $\{a_1, \ldots, a_l\}$ the single number $p_1^{m_1} \ldots p_l^{m_l}$. When this has been done, every symbol, every formula, and every finite sequence of formulae in the system P is represented without ambiguity by a Gödel number.

Suppose now that we have some metamathematical predicate or relation $R(a_1, \ldots, a_n)$ that is defined for sets of n symbols, formulae, or sequences of formulae $a_1, \ldots, a_n$. In the special case $n = 1$ the 'relation' is to be understood, as usual, as a class of symbols, etc. Then the relation $R(a_1, \ldots, a_n)$ determines an 'induced' relation $R'(x_1, \ldots, x_n)$, defined for sets of natural numbers. This induced relation holds for a particular set of numbers $x_1, \ldots, x_n$ if and only if there exist $a_1, \ldots, a_n$ in the formal system such that, for $i = 1, \ldots, n$, a_i has Gödel number x_i—in the appropriate sense—and $R(a_1, \ldots, a_n)$ is true. Gödel represents the arithmetical relation R' that is derived in this way from any particular metamathematical relation R by putting the name of R into italics. Thus, for example, *formula* is the class (i.e. predicate) whose members are the Gödel numbers of formulae; and 'm is a *derivation* of n' is the relation between two numbers m and n which holds whenever m is the Gödel number of a sequence of formulae that is a derivation, the end-formula of which has Gödel number n.

When Gödel had set up his numbering in the manner described, he had to go on to show that the arithmetized metamathematics of P was all contained in P; that is to say, he had to show that the various italic relations between numbers that are obtained from such metamathematical relations as 'formula', 'immediate consequence', and 'derivable' admit of formalization within P. He was able to do this with the aid of recursive arithmetic. Only the elements of recursive arithmetic were needed for this purpose, and Gödel gave the following explanation of what he understood by recursive functions and recursive relations. The explanation is expressed intuitively, of course, not formalized within the system P.

An arithmetical function $\varphi(x_1,x_2,\ldots,x_n)$ is said to be *recursively defined* in terms of a pair of given functions $\psi(x_1,x_2,\ldots,x_{n-1})$ and $\mu(x_1,x_2,\ldots,x_{n+1})$ if, for all $x_2,\ldots,x_n$, k,

$$\varphi(0,x_2,\ldots,x_n) = \psi(x_2,\ldots,x_n),$$
$$\varphi(k+1,x_2,\ldots,x_n) = \mu(k,\varphi(k,x_2,\ldots,x_n),x_2,\ldots,x_n).$$

A *function* φ is said to be *recursive*[1] if there exists a finite sequence of functions $\varphi_1, \varphi_2, \ldots, \varphi_n$ such that $\varphi_n = \varphi$, and every function φ_k of the sequence is either recursively defined in terms of two of its predecessors, or is obtained by substitution in one of its predecessors, or else is either a constant or the successor function $x + 1$.

A *relation* $R(x_1,\ldots,x_n)$ between natural numbers is said to be *recursive* if there exists a recursive function $\varphi(x_1,\ldots,x_n)$ such that, for any numbers $x_1,\ldots,x_n$, the relation $R(x_1,\ldots,x_n)$ holds if and only if $\varphi(x_1,\ldots,x_n) = 0$.

Recursiveness having thus been defined, the demonstration that the metamathematics of the system P can be formalized in P proceeds in two stages. First it is shown that all the relevant metamathematical relations are recursively expressible in terms of the Gödel numbering; and then recourse is had to a general theorem which asserts that every recursive relation is representable in P (cf. the definition of formal representability on p. 222 above).

Gödel states and proves from first principles for his formal system P the following theorem, which may be compared with the corresponding theorem for Hilbert's system (Z), which we give in Chapter 10 (p. 269):

THEOREM 1. *If* $R(x_1,\ldots,x_n)$ *is any relation in recursive arithmetic, there exists in* P *a relation-formula* $\varrho(x_1^{(1)},\ldots,x_1^{(n)})$, *with n free variables, such that, for any set of n numbers* $m_1,\ldots,m_n$, (*i*) *if* $R(m_1,\ldots,m_n)$ *holds then* $\varrho(m_1',\ldots,m_n')$ *is derivable in* P, *and* (*ii*) *if* $R(m_1,\ldots,m_n)$ *does not hold then* $\sim\varrho(m_1',\ldots,m_n')$ *is derivable in* P, m_i' *denoting the symbol* $ff\ldots f0$ *that corresponds in* P *to the number* m_i.

This theorem appears in Gödel's paper as *Satz* V (p. 186).

That all the relevant metamathematical relations are recursive was shown by Gödel in a series of 46 lemmas. Typical of the relations that he considers are those which we can express (using the device of italic names to which we referred on p. 235) by the words 'x is a *variable of the nth type*', 'x is an *axiom*', and 'x is a *derivation* of the *formula* y'.

---

[1] I.e. primitive recursive in current terminology (cf. Chapter 10, §6).

4* The central argument of Gödel's paper

In attempting to summarize Gödel's argument in such a way that we can give some indication of the essential steps without going into too many formal details, we shall adopt a device of denoting the sequence of symbols (i.e the expression) that has a given Gödel number x by $\tilde{x}$, and the sequence of expressions that has a given Gödel number x by $\tilde{\tilde{x}}$. Then the statement 'x is a *derivation* of the *formula* y', for instance, can be expressed in the alternative form '$\tilde{\tilde{x}}$ is a sequence of formulae that is a derivation of the formula $\tilde{y}$'. We shall also use the notation 'x'', as in the theorem stated at the end of §3, for the expression $ff\ldots f0$ which represents the number x in the formal system P.

Gödel found it necessary to postulate that the system P is *ω-consistent*, or consistent in a stronger sense than that which is ordinarily envisaged, and the concept of ω-consistency must be explained before we can state his main theorem. An arithmetical formalism is consistent in the ordinary sense—or *simply consistent*—if no two formulae $\mathfrak{A}$ and $\overline{\mathfrak{A}}$ are both derivable in it; but consistency in this sense does not exclude the more subtle kind of contraction that occurs when, for some formula $\mathfrak{A}(n)$ that contains a free variable n, $\mathfrak{A}(\mathfrak{n})$ is derivable for each particular number $\mathfrak{n}$, and $\overline{(x)\mathfrak{A}(x)}$ is also derivable. This possibility was first pointed out by Alfred Tarski. It is easily seen that a system that is ω-consistent is also simply consistent, for if it were not simply consistent then every formula would be derivable in it (cf. p. 211); but Tarski has shown by an example that a system can be simply consistent without being ω-consistent.[1]

Gödel formulated his main argument in terms of any formal system that can be obtained by supplementation of the axioms of the basic system P by an arbitrary recursive class $\varkappa$ of additional formulae. He first defined the *consequential set* (*Folgerungsmenge*) of $\varkappa$ as the smallest set of formulae that contains the axioms of P and the formulae belonging to $\varkappa$, and is at the same time closed with respect to the relation of immediate consequence. If $\varkappa$ is empty, its consequential set plainly reduces to the set of derivable formulae of P.

The two senses of consistency can now be redefined, with reference to such a class $\varkappa$, as follows: A class of formulae $\varkappa$ is said to be *simply consistent* if its consequential set does not contain any formula together with its negation; and $\varkappa$ is said to be *ω-consistent* if there does not exist any class-formula $a(v_1)$ such that (i) $a(n')$ belongs to the consequential set of $\varkappa$ for every natural number n, and (ii) $\sim(v_1\Pi(a(v_1)))$ also belongs to this consequential set.

---

[1] A. Tarski: 'Einige Betrachtungen über die Begriffe der ω-Widerspruchsfreiheit und der ω-Vollständigkeit', *Monatsh. Math. Phys.*, **40** (1933), 97–112. In this paper Tarski claims priority in the introduction of the concept of ω-consistency though not in the use of this name for it.

We can now state Gödel's key theorem (his *Satz* VI) in the following terms:

THEOREM 2. *If $\varkappa$ is any ω-consistent recursive class of formulae of the system* P, *there exists a formula $\varrho(x_1)$, which determines a recursive class of numbers, and which is such that neither $v_1\Pi(\varrho(v_1))$ nor $\sim(v_1\Pi(\varrho(v_1)))$ belongs to the consequential set of $\varkappa$.*

Proof. In outlining Gödel's proof, we shall use the term '$\varkappa$-derivation' to mean a derivation in which any formulae belonging to $\varkappa$ may be used as initial formulae.

We consider first of all the metamathematical propositional function '$\tilde{\tilde{x}}$ is not a $\varkappa$-derivation of the formula which results when y' is substituted for y_1 in the formula $\tilde{y}$'. The 'y_1' that is here referred to is the particular symbol of the system P that has Gödel number 19, and the letter 'y' in it has no connexion with the 'y' that is one of the variables involved in the propositional function. This propositional function is a relation, $Q(x,y)$ say, built up out of constituent notions such as 'im mediate consequence' and 'substitution', whose italic counterparts ar proved in the lemmas of Gödel's paper to be recursive; and it is therefore represented in P by a relation-formula $\tilde{q}(x_1,y_1)$, in such a way tha (i) if $\tilde{\tilde{x}}$ is not a $\varkappa$-derivation of $\tilde{y}(y')$, then $\tilde{q}(x',y')$ is derivable, and (ii) i $\tilde{\tilde{x}}$ is a $\varkappa$-derivation of $\tilde{y}(y')$, then $\sim(\tilde{q}(x',y'))$ is derivable.

We now put

$$\tilde{p}(y_1) \equiv x_1\Pi(\tilde{q}(x_1,y_1))$$

and

$$\tilde{r}(x_1) \equiv \tilde{q}(x_1,p').$$

Then

$$\tilde{p}(p') \equiv x_1\Pi(\tilde{q}(x_1,p'))$$
$$\equiv x_1\Pi(\tilde{r}(x_1)),$$

and

$$\tilde{q}(x',p') \equiv \tilde{r}(x').$$

Substituting p for y in (i) and (ii) above, we have: (i)$'$ if $\tilde{\tilde{x}}$ is not a $\varkappa$-derivation of $\tilde{p}(p')$, then $\tilde{q}(x',p')$ is derivable; and (ii)$'$ if $\tilde{\tilde{x}}$ is a $\varkappa$-derivation of $\tilde{p}(p')$, then $\sim(\tilde{q}(x',p'))$ is derivable. In other words, (i)$''$ if $\tilde{\tilde{x}}$ is not a $\varkappa$-derivation of $x_1\Pi(\tilde{r}(x_1))$, then $\tilde{r}(x')$ is derivable; and (ii)$''$ if $\tilde{\tilde{x}}$ is a $\varkappa$-derivation of $x_1\Pi(\tilde{r}(x_1))$, then $\sim(\tilde{r}(x'))$ is derivable.

At this point the argument divides into two parts:

1. We suppose that $x_1\Pi(\tilde{r}(x_1))$ is $\varkappa$-derivable. Then it follows that $\tilde{r}(n')$ is $\varkappa$-derivable for any number n whatever. But, in view of (ii)$''$,

there is an n such that $\sim(\tilde{r}(n'))$ is $\varkappa$-derivable. We thus have a $\varkappa$-derivable contradiction; and this proves, even if $\varkappa$ is assumed merely to be *simply* consistent, that the formula $x_1\Pi(\tilde{r}(x_1))$ is not $\varkappa$-derivable.
2. We have shown in 1. that $x_1\Pi(\tilde{r}(x_1))$ is not $\varkappa$-derivable, and thus, for any n whatever, $\tilde{\tilde{n}}$ is not a $\varkappa$-derivation of $x_1\Pi(\tilde{r}(x_1))$. Hence, by (i)'', $\tilde{r}(n')$ is derivable for every choice of n. If, therefore, $\varkappa$ is assumed to be ω-consistent, the formula $\sim(x_1\Pi(\tilde{r}(x_1)))$ is not $\varkappa$-derivable.

1. and 2. together show that neither the formula $x_1\Pi(\tilde{r}(x_1))$ nor its negation is $\varkappa$-derivable, and this proves the theorem.

Remarks (i) In the above version of the proof we have used the expression $\tilde{q}(x_1,y_1)$ to stand for the formula with Gödel number q, and to indicate at the same time that this formula contains x_1 and y_1 as free variables. We have then used $\tilde{q}(x_1,p')$ to stand for the formula

$$\text{Subst } \tilde{q}(x_1,y_1)\begin{pmatrix} y_1 \\ p' \end{pmatrix}.$$

(ii) It has been shown by Barkley Rosser (*J. Symb. Logic*, **1** (1936), 87–91) that Gödel's original proof can be modified in such a way that the existence of an undecidable formula is shown to follow merely from simple consistency of $\varkappa$, the heavier requirement of ω-consistency being in fact superfluous.

(iii) After proving his main theorem on the existence of undecidable formulae, Gödel made the following comment on its range of application:

'In the proof of Theorem VI, no properties of the system P were used except the following:
1. The class of axioms and the rules of inference (i.e. the relation of "immediate consequence") are recursively definable (as soon as the primitive symbols have been replaced in some way by natural numbers).
2. Every recursive relation is definable within the system P (in the sense of Theorem V).

It follows that, in every formal system that satisfies the hypotheses 1 and 2 and is ω-consistent, there are undecidable propositions of the form $(x)F(x)$, where F is a recursively defined property of natural numbers, and the same is true of every extension of such a system by means of a recursively definable ω-consistent class of axioms. The systems that satisfy the hypotheses 1 and 2 include, as may easily be verified, the Zermelo-Fraenkel and von Neumann systems of axioms for the theory of sets, and also the system of axioms for the theory of numbers that consists of the Peano axioms, recursive definitions, and the logical rules. Hypothesis 1 is satisfied by any system whatever that has the usual rules of inference and whose axioms arise (as in the case of P) by substitution in a finite number of schemata.'

5 The impossibility of an 'internal' proof of consistency

In §4 of his paper, Gödel turned his attention to the second of his famous theorems, the one that establishes the impossibility of proving certain formal systems consistent from within. The general idea of the argument (of which only the outline is sketched in the original paper) may be explained in the following terms.

The argument depends on the fact that in the system P we are able to formalize intuitive arithmetical reasoning. That is to say, given an intuitive arithmetical proof, we can construct in P a corresponding formal derivation. Suppose then, once again, that $\varkappa$ is an arbitrary recursive class of formulae of the system P. We already know, by 1. on p. 238, that from the hypothesis that $\varkappa$ is simply consistent it follows that the formula $x_1\Pi(\tilde{r}(x_1))$ is not $\varkappa$-derivable. More precisely, we are able to give an intuitive proof (which is effectively arithmetical in view of our arithmetization of the metamathematics of the system P) that this conclusion follows. Hence we can establish, by intuitive reasoning, the implication

($\varkappa$ is consistent) $\rightarrow$ (the formula $x_1\Pi(\tilde{r}(x_1))$ is not $\varkappa$-derivable),
i.e. ($\varkappa$ is consistent) $\rightarrow$ (x)($\tilde{\tilde{x}}$ is not a $\varkappa$-derivation of $x_1\Pi(\tilde{r}(x_1))$),
i.e. ($\varkappa$ is consistent) $\rightarrow$ (x)($\tilde{\tilde{x}}$ is not a $\varkappa$-derivation of $\tilde{p}(p')$),
i.e. ($\varkappa$ is consistent) $\rightarrow$ $(x)Q(x,p)$.

The intuitive proof of this implication can now be formalized by an equivalent derivation in P. The metamathematical proposition '$\varkappa$ is consistent' will be formalizable in P by a certain propositional formula $\tilde{w}$; and since $Q(x,p)$ is formalizable, as we already know, by $\tilde{r}(x')$, it follows that $(x)Q(x,p)$ is formalizable by $x_1\Pi(\tilde{r}(x_1))$. The derivation in P will accordingly be a derivation of the formal implication

$$\tilde{w} \supset x_1\Pi(\tilde{r}(x_1)).$$

If, now, $\tilde{w}$ were to be derivable in P, we would at once have a derivation of $x_1\Pi(\tilde{r}(x_1))$—and this we know to be impossible if $\varkappa$ is consistent. Thus $\tilde{w}$ is not derivable in P, and so we have the following theorem:

THEOREM 3. *If $\varkappa$ is any simply consistent recursive class of formulae in* P, *there does not exist any proof, formalizable in* P, *of the consistency of the system which results when the formulae in $\varkappa$ are adjoined to* P *as additional initial formulae.*

If, in particular, we take for $\varkappa$ the empty class, we see that Theorem 3 applies to the system P itself.

It was the publication of this theorem that effectively brought to a close the initial phase of the formalist movement, as we have indicated in §1 of this chapter. Although, as Gödel was careful to point out, the theorem only rules out *formalizable* proofs of consistency, and does not say anything directly about *finitary* proofs, it does not seem likely that a finitary procedure could ever be devised that was not also formalizable.

Hilbert and Bernays have discussed Gödel's theory and its implications at length in the second volume of their *Grundlagen der Mathe-*

matik (1939). They modify the details of Gödel's paper, using an alternative way of arithmetizing metamathematics which brings the theory into more immediate relation with Hilbert's treatment of formalized arithmetic, but the basis of the theory remains substantially unchanged. The proofs are, of course, given in full; and for each of Gödel's theorems precise conditions are formulated which are sufficient to ensure that a formal system is one to which the theorem applies. It is eventually shown (*Grundlagen*, II, §5, 2, e) that the theorem on the impossibility of an internally formalizable proof of consistency holds, in particular, for the system (Z) of formalized arithmetic, as specified on p. 223 above.

We see, then, that the consistency of formalized arithmetic can only be proved by methods which are more general than any that are provided for in the arithmetical system itself. When Gödel had established this fact, it was natural that attempts should be made to prove the consistency of arithmetic by means which did not go too far beyond the previously recognized limits of strict finitariness, and in 1936 Gentzen succeeded in constructing a proof of consistency which depended on the application of transfinite induction. For this further development of metamathematics see Note 1, immediately following, and also Section 2 of the Appendix, p. 385.

SUPPLEMENTARY NOTES ON CHAPTER 8

1. *Gentzen's proof of the consistency of arithmetic.* It follows from Gödel's theorem on proofs of consistency (Theorem 3, p. 240) that, for any reasonably adequate formalization of arithmetic to be proved consistent, more powerful types of metamathematical argument must be allowed than those originally envisaged by Hilbert. The proof of consistency will then no longer provide the clear justification of arithmetic as a 'significant' formal theory that was at one time hoped for, and its value in this respect will vary inversely with the power of the methods of demonstration which it employs. But, whatever weight may now be given to proofs of consistency as an ultimate guarantee of the validity of mathematical theories, any precise indication of means that are sufficient to establish the consistency of formalized arithmetic is of major interest in its own right. In the proof of consistency which Gentzen constructed, finitary reasoning of the earlier kind was supplemented by a limited use of transfinite induction.

Gentzen's argument depended on the introduction of a process of reduction (cf. p. 211) which can be applied to any derivation in the arithmetical system that he considered. This process was defined in a recursive manner, and to complete the specification Gentzen had to show that the reduction of any given derivation can be accomplished in a finite number of steps. To show that the process of reduction always terminates, he introduced ordinal numbers as measures of the complexity of formulae and then applied transfinite induction to these ordinal numbers. For this purpose he needed transfinite induction for a certain segment of Cantor's second number class, namely the series of all ordinals below the first ε-number, i.e. the smallest ordinal number α such that $\omega^\alpha = \alpha$. A different proof of the consistency of arithmetic, also dependent on transfinite induction up to the first ε-number, has been given by Ackermann [2]. This proof makes essential use of the selection operator ε_x, which Hilbert originally introduced with the aim of extending to the arithmetical system (Z) the proofs of consistency that he had already constructed for weaker systems.

2. *Alternative accounts of Gödel's investigation.* A detailed presentation in English of the substance of Gödel's fundamental paper is contained in the mimeographed notes [3] of lectures which Gödel gave at Princeton in 1934. These notes were compiled by Kleene and Rosser. Rosser has published also, in the *Journal of Symbolic Logic* in 1939, an informal paper [2] which gives the gist of Gödel's work.

Gödel's theory is presented afresh by Mostowski in his book [1], again in considerable detail. A more elementary summary is that of N. R. Hanson [1]; and there is also a popular account by Nagel and Newman [1].

data
S T Q Q S S D

20210807$_{m}$ 20210806$_{m}$

Books-016-2021/20210804/
The Absolute Differential Calculus (Calculus of Tensors)/
0486634019.djvu

Pág. 172(187) - Chapter VII

Riemann's Symbols and Properties relative to Curvature:
Ricci's and Einstein's Symbols;
Geodesic Deviation

1. Cyclic displacement and the relations between parallelism and curvature

20210804/Mathematical Logic.../
0442043546.djvu

Pág. 224(119) - Supplementary Notes on Chapter 7

Foundations of projective geometry

Pág. 229(121) Chapter 8
Gödel's theorems on the inherent Limitations of formal Systems

1. Gödel's new metamathematical method.

Jandaia

data

Pág. 191(102) Chapter 7
Formalized Mathematics and
Metamathematics

Hilbert's new approach to the
foundations of mathematics

—

Pág. 152(83) Supplementary Notes on
Chapter 5

1. Peano's space-filling curve.

—

20210807///

heal our wounds →
curar nossas feridas.

—

uploaded.net 99€/2year

—

Academia.edu
profprofsm@yahoo.com
Pass@8790 —

Joseph Raphson - Math
and Newton
×
Leibniz —

Chapter 7

FORMALIZED MATHEMATICS AND METAMATHEMATICS

1 Hilbert's new approach to the foundations of mathematics

For all the inspiration that *Principia Mathematica* has communicated to the logicians and philosophers of mathematics of the twentieth century, and for all its rich fecundity as a source of concepts and symbolic devices, this great work remains, in the literature of the foundations of mathematics, a lone classic without progeny.[1] That such should be the case may well seem puzzling; and indeed there does not appear to be any simple reason why the impetus of the logistic movement should have failed with the publication of the third volume of the *Principia*. It is an undoubted fact, however, that the leadership of research into the foundations of mathematics passed at that time from Russell to Hilbert, and that it is from Hilbert that most of the subsequent developments directly derive. Before turning to the details of Hilbert's work we must try to see, as best we can, how this sudden change came about.

The Russell of the *Principia* was both a mathematician and a logician, though inclining, as we may infer from the outlook revealed in *The Principles of Mathematics*, more to the side of logic and philosophy than to that of mathematics, whereas Hilbert was one of the most professional of mathematicians. This difference may go a long way towards accounting for the greater readiness of mathematicians interested in the foundations of mathematics to align themselves with Hilbert than to follow Russell into the unfamiliar territory of philosophy.

There were also certain specific deficiencies in *Principia Mathematica* which must have made it appear that the solution there offered to the problems raised by the foundations of mathematics was not a complete or final one. The controversial nature of the theory of types, and in particular the intrusion of the admittedly arbitrary axiom of reducibility, constituted an obvious imperfection; and there was an even more fundamental weakness in the logistic view itself that had been shown up by the discovery of the vicious-circle paradoxes. The known paradoxes could indeed all be eliminated from the logistic system by the theory of types, but this still did not ensure that no other paradoxes would arise in the future, and the most that could be claimed was that the logic of the *Principia* was as rigorous as anyone then knew how to make it. But if Russell had not happened to stumble upon his contradiction, the *Principia* would presumably have been written (as indeed was Frege's *Grundgesetze der Arithmetik*) without any separation of types, and it is unlikely that logicians would have suspected that there might be something wrong with it. Elimination of known fallacies is therefore not by itself a sufficient safeguard, and the ultimate logical justification of mathematics must include some more positive guarantee of its consistency. The main concern of Hilbert, in the decades following the publication of *Principia Mathematica*, was with the search for such a guarantee.

---

[1] On the logical side, the *Principia* has a direct descendant in Quine's *Mathematical Logic* (1940, revised 1951).

The approach of Hilbert to the problems of the foundations of mathematics was so closely bound up with his personal experience of the activity of mathematical creation that we can scarcely hope to understand it unless we see his work as a whole. By good fortune we are in a position to trace the mathematical origin of his preoccupation with questions of foundations, for we possess an early though fully mature work of his in which we can see this interest in process of development. This work is the famous *Grundlagen der Geometrie* [Foundations of Geometry] of 1899, which we shall have to consider in some detail. But before turning to Hilbert's writings we must first summarize the main facts of his career.

David Hilbert was born in 1862 at Königsberg in East Prussia—the native city of Immanuel Kant, the great philosopher—and he became in turn a student, a lecturer, and a professor at the university of Königsberg. In 1895 he went as professor to Göttingen, where he remained until his death in 1943; and he thus spent more than half his life at the university whose fame as a centre of mathematical research had begun with the presence there of the illustrious figure of Gauss and had been maintained by a succession of mathematicians of the highest eminence, from Dirichlet and Riemann to Klein. Indeed, Hilbert himself must be numbered among the greatest creative mathematicians since Gauss, for besides making discoveries of fundamental importance in many different branches of mathematics, he influenced decisively the mathematical outlook of his whole generation.

In some fields of study Hilbert introduced concepts and methods that enabled research to go forward in entirely new directions, as with his theorem, described by Hermann Weyl[1] as 'one of the simplest and most important theorems in the whole of algebra', which asserts that every ideal in a polynomial ring has a finite basis, and also with the concept that is now known by the name of 'Hilbert space'. In the case of the classical theory of invariants he was able, by solving a fundamental problem that had been the ultimate objective of a vast amount of earlier work, to bring research virtually to an end. Among the most important and influential of his publications (excluding those devoted specifically to the foundations of mathematics) are the book on geometry that we are about to discuss; the famous *Zahlbericht* of 1897, a report on the state of knowledge in the theory of algebraic numbers, drawn up for the *Deutsche Mathematiker-Vereinigung*; a book, *Grundzüge einer allgemeinen Theorie der linearen Integralgleichungen* (Leipzig and Berlin, 1912), in which are reprinted six fundamental papers on integral equations, originally published in the *Göttinger Nachrichten* between 1904 and 1910; and the two volumes of *Methoden der mathematischen Physik* (Berlin, 1924 & 1937), written in collaboration with Richard Courant, in which essentially algebraic methods are made the basis of a unified treatment of partial differential equations, integral equations, and the calculus of variations.

---

[1] See H Weyl: *The Classical Groups* (Princeton, 1939), p 251.

2 Hilbert's 'Grundlagen der Geometrie'

Hilbert's epoch-making book of 1899 was offered as a new presentation of euclidean geometry, written with the intention of taking fully into account the enhanced understanding of the axiomatic method that mathematicians had achieved in the centuries that had elapsed since the time of Euclid, and more particularly in the closing decades of the nineteenth century. Not only did Hilbert profit by this recently acquired insight, but he himself raised the axiomatic method to a new level of efficiency; and of his book Hermann Weyl subsequently wrote as follows:

'The soil was well prepared, especially by the Italian school of geometers. Yet it was as if over a landscape, wherein but a few men with a superb sense of orientation had found their way in murky twilight, the sun had risen all at once. Clear and clean-cut we find stated the axiomatic concept according to which all geometry is a hypothetical deductive system; it depends on the implicit definitions of the concepts of spatial objects and relations which the axioms contain, and not on a description of their intuitive content. A complete and natural system of geometric axioms is set up. They are required to satisfy the logical demands of consistency, independence, and completeness, and by means of quite a few peculiar geometries, constructed *ad hoc*, the proof of independence is furnished in detail.'[1]

---

[1] H. Weyl: 'David Hilbert, 1862–1943', *Obituary Notices of Fellows of the Royal Society*, **4** (1942–1944), 547–553. In this notice, Weyl gave an account of Hilbert's mathematical achievements and showed clearly why it was that his influence on the development of mathematics was so great. A more detailed survey of Hilbert's work is contained in the longer obituary notice [5] that Weyl wrote for the American Mathematical Society

In the opening chapter of the *Grundlagen der Geometrie* Hilbert laid his foundation for a new axiomatic treatment of euclidean geometry that was intended to be absolutely rigorous, although expressed without the aid of any special symbolism and in language no more technical than that of Euclid. The later chapters of the book were then devoted to investigations, with the system of geometry itself as their object, of the kind described nowadays as *metamathematical*. Chief among these investigations was one that established the consistency of geometry; and the others were mainly concerned with general questions of logical dependence of propositions on other propositions—whether certain important theorems are deducible from such and such axioms, and whether the axioms themselves are independent.

The formal system was constructed in accordance with the principles of Peano, although this feature was somewhat masked by Hilbert's use of ordinary language in preference to logical symbolism. Hilbert had his reasons for this choice of language, the main consideration presumably being that he wished his work to be seen as a continuation of the age-old striving of geometers to evolve a purely rational theory of the spatial structure of the world. That this was his motive is confirmed by the following words from the introduction to the book:

> 'Geometry, like arithmetic, requires for its systematic development only a small number of simple basic principles. These principles are known as axioms of geometry. Setting up axioms for geometry and investigating the way in which they are connected is a problem that has been discussed ever since the time of Euclid in numerous admirable contributions to the literature of mathematics. The problem in question amounts to logical analysis of our intuition of space.'

The concluding sentence of this paragraph was no merely conventional remark, but a deliberate affirmation of ultimate intention. Hilbert made this quite clear by placing at the head of the introduction, as a motto for the whole book, these famous words from Kant's *Critique of Pure Reason*: 'Thus all human knowledge begins with intuitions, goes from there to concepts, and ends with ideas.' It had been Kant's firm conviction that theoretical knowledge and immediate intuition are inseparable, and that even the most abstruse scientific reasoning derives its significance and its objectivity from its continuity with the basic intuitions of the active mind. In a rather similar way Hilbert believed that, although geometry is entirely self-sufficient as a deductive system, it is nevertheless dependent on the immediacy of intuition for the interest that makes it a recurrent challenge to the human mind instead of merely one among a whole infinity of arbitrary intellectual exercises.

Thus the language of the *Grundlagen* was the traditional language of elementary geometry; and where this language offered convenient synonyms, such as 'lies on' or 'belongs to' for 'is a point of', Hilbert used them freely to denote one and the same abstract concept. From the point of view of the strict logician, however—and this was one of the points of view that Hilbert so skilfully kept distinct and yet at the same time never entirely separated—the concepts denoted by the geometrical terms were primitive concepts in the sense of Peano. Among such concepts were 'point', 'line' (*Gerade*), and 'plane', conceived as three separate kinds of entity, and a number of basic relations such as 'belongs to' (i.e. 'is incident with'), 'is between', 'is congruent to', and 'is parallel to'. The axioms or primitive propositions which comprised the implicit definitions of the primitive concepts were twenty in number, and they were classified in five sets as follows: I, 1–8, axioms of connexion; II, 1–4, axioms of order; III, 1–5, axioms of congruence; IV, the axiom of parallels; V, 1–2, axioms of continuity. These axioms are a sufficient foundation for a strictly deductive treatment of Euclid's system of geometry, in which stated axioms take over the ill-defined function of such semi-intuitive procedures as establishment of congruence by the method of superposition. We shall now give them in full.

I. *The axioms of connexion:*

1. If A and B are two points, there is always a line a that is incident with both of them.
2. If A and B are two points, there is not more than one line that is incident with both of them.
3. On a line there are always at least two points. There are at least three points which do not lie on a line.
4. If A, B, C are any three points which do not lie on one and the same line, there is always a plane α which is incident with them all. For every plane, there is a point that is incident with it.
5. If A, B, C are any three points which do not lie on one and the same line, there is not more than one plane that is incident with them all.
6. If two points A, B of a line a lie in a plane α, then every point of a lies in the plane α.
7. If two planes α, β have a point A in common, then they have at least one other point B in common.
8. There are at least four points which do not all lie in a plane.

II. *The axioms of order:*

1. If a point B lies between a point A and a point C, then A, B, C are three distinct points of a line, and B then also lies between C and A.
2. If A and C are two points, there is always at least one point B on the line AC which is such that C lies between A and B.
3. For any three points of a line, there is not more than one that lies between the other two.
4. Let A, B, C be three points which do not lie on a line, and let a be a

line in the plane ABC which does not pass through any of the points A, B, C. Then if the line a goes through a point of the segment AB [i.e. a point between A and B] it must go also through a point of the segment AC or through a point of the segment BC.

III. *The axioms of congruence:*
1. If A and B are two points on a line a, and if, further, A' is a point on the same or another line a', then a point B' can be found on the line a', on a given side of A', which is such that the segment AB is congruent or equal to the segment $A'B'$; in symbols:

$$AB \equiv A'B'.$$

2. If a segment $A'B'$ and a segment $A''B''$ are congruent to the same segment AB, then the segment $A'B'$ is congruent to the segment $A''B''$; in short, if two segments are congruent to a third, then they are congruent to each other.
3. Let AB and BC be two segments, without any points in common, on a line a, and also let $A'B'$ and $B'C'$ be two segments on the same or another line a', again without common points. If $AB \equiv A'B'$ and $BC \equiv B'C'$, then $AC \equiv A'C'$ also.
4. Let $\sphericalangle\,(h,k)$ be an angle in a plane α, and let a line a' in a plane α' and also a particular side of a' in α' be given. If h' denotes one ray of the line a' which issues from a point O', then in the plane α' there is one and only one ray k' such that the angle $\sphericalangle\,(h,k)$ is congruent or equal to the angle $\sphericalangle\,(h',k')$, and at the same time all the internal points of the angle $\sphericalangle\,(h',k')$ lie on the given side of a'; in symbols:

$$\sphericalangle\,(h,k) \equiv \sphericalangle\,(h'k').$$

Every angle is congruent to itself, i.e. always

$$\sphericalangle\,(h,k) \equiv \sphericalangle\,(h,k).$$

5. If, for two triangles ABC and $A'B'C'$, the congruences

$$AB \equiv A'B', \quad AC \equiv A'C', \quad \sphericalangle\,BAC \equiv \sphericalangle\,B'A'C'$$

all hold, then the congruence

$$\sphericalangle\,ABC \equiv \sphericalangle\,A'B'C'$$

is always satisfied also.

IV. *The axiom of parallels (Euclidean axiom):*
Let a be an arbitrary line and let A be a point that is not on a. Then, in the plane that is determined by a and A, there is at most one line that passes through A and does not meet a.

V. *The axioms of continuity:*

1. (*Axiom of measurement* or *Archimedean axiom.*) Let AB and CD be any segments whatever. Then there is a natural number n such that, if the segment CD is stepped off n times in succession along the ray from A which passes through B, a point beyond B will be reached.
2. (*Axiom of linear completeness.*) The system of points on a line, with its relations of order and congruence, does not admit of any extension that preserves all relations between existing elements, and also all the fundamental properties of linear order and congruence which follow from the axioms I–III, and that preserves V, 1.

The twenty axioms have been stated here in the original sequence, but the later ones will be seen to involve concepts that have not been formally introduced. This is because Hilbert himself interspersed between the separate sets of axioms various deductions of theorems and discussions of derivative notions. For details of these, and of the way in which Euclid's basic theorems were deduced from the new set of axioms, the original memoir must be consulted. It will be found that, despite its appearance of comparative informality, Hilbert's system of euclidean geometry is developed with great formal rigour. In any case, important as was Hilbert's achievement of rewriting Euclid in the spirit of Peano, the *Grundlagen der Geometrie* was to prove even more significant as a metamathematical essay than as a treatise on geometry; and it is this aspect of the book that we need chiefly to discuss at present.

The first metamathematical topic that Hilbert considered was the consistency of his axioms, that is to say the freedom from contradiction of the deductive theory that is built upon them. The method that he adopted for proving consistency was one of constructing an arithmetical model of the abstract system; and the use of models has since come to be very important in metamathematical investigations. By a *model* of a formal axiomatic theory we understand a realization of that theory, constructed in some mathematical domain which is already available. The formal theory will be based, as always, on one or more postulated totalities of entities, certain constant predicates and functions, and suitable axioms which define these predicates and functions implicitly; and to construct a model of it we replace the postulated totalities of entities by certain specific mathematical totalities, and the constant predicates and functions by specific predicates and functions defined for these totalities. If we are then able to show that, with this particular identification of the primitive elements of the formal theory, all the axioms become true propositions, we have succeeded in producing a model of the original theory. A model of a formal theory is thus a well-defined mathematical system with the particular structure that is characterized by the theory.

What Hilbert did in the case of axiomatic euclidean geometry was to construct a model of the geometrical system within the domain of arithmetic. He could then argue that, if a contradiction were deducible from the geometrical axioms, a contradiction could also be reached by purely arithmetical reasoning. It follows that, when once the consistency of arithmetic is granted, the consistency of the system of euclidean geometry is assured. Thus, although the method of models does not establish the consistency of geometry in an absolute sense, it does nevertheless furnish a proof of the consistency of geometry relative to that of arithmetic—and even this is a substantial achievement.

Hilbert obtained two main results concerning freedom from contradiction. He began by taking a certain number field Ω, consisting of all the algebraic numbers that can be obtained by 'beginning with the number 1 and applying a finite number of times the four rational operations: addition, subtraction, multiplication, division, and the fifth operation $|\sqrt{(1 + \omega^2)}|$, where ω denotes any number already obtained by means of the five operations'. By defining a point as an ordered pair (x,y) of numbers in Ω, a line as a set of ratios $u:v:w$ with not both of u and v zero, the relation of incidence by the equation $ux + vy + w = 0$, the order of points on a line in terms of the order of magnitude of their coordinates, and so forth—prompted always by the familiar usages of coordinate geometry—he was able to construct a model that satisfied all those of his axioms that apply to plane geometry, with the single exception of V, 2 (the axiom of linear completeness); and the same method could readily be extended to the axioms for three-dimensional space. And if the restricted field Ω was replaced simply by the field of all real numbers, so that the model became identical with ordinary cartesian geometry, all the axioms were then satisfied, V, 2 included. The reason why Hilbert thought it worth while to give a separate proof of the less comprehensive consistency theorem was that the 'arithmetic' to which the consistency of geometry was made relative was logically much simpler in that case, and therefore even less open to philosophical doubt than the full arithmetic of the real continuum.

Having demonstrated the consistency of his axioms, Hilbert went on to show also that these were, in all essential respects, independent. This he did, in Weyl's phrase, 'by means of quite a few peculiar geometries, constructed *ad hoc*'; for a possible way of showing that some given axiom in a set is independent of the other axioms is by constructing a model for which all the axioms are satisfied, with the single exception of this particular one. Hilbert first of all inferred the independence of the axiom of parallels by appealing to the known model of non-euclidean geometry that can be set up by taking as the 'space' of the model the interior of a sphere in cartesian space, and defining 'congruence' in it by means of the collineations of the cartesian space that transform the sphere into itself.[1]

Besides the axiom of parallels, there are two more of the twenty axioms whose independence is of especial interest, namely the last of the axioms of congruence, III, 5, and the axiom of Archimedes, V, 1; and Hilbert dealt with these in the following way. He first of all modified three-dimensional cartesian geometry by defining 'distance' by means of the formula

$$\sqrt{[(x_1 - x_2 + y_1 - y_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2]},$$

instead of by the usual formula derived from the theorem of Pythagoras. This gave him a 'geometry' which satisfies all of the axioms except III, 5. And finally, in order to demonstrate the independence of axiom V, 1, he constructed a field $\Omega(t)$ of algebraic functions of t by proceeding exactly as in the construction of the earlier field Ω of algebraic numbers, but beginning now with the real variable t instead of with the number 1. Since every element of $\Omega(t)$ has a constant sign for all sufficiently large t, an ordering relation for this field may be defined by taking '$a(t) < b(t)$' to mean '$a(t) - b(t) < 0$, in the ordinary sense, for all sufficiently large t'. But, with this definition of order, no finite multiple $n.1$ of 1 can exceed t, and the field $\Omega(t)$ is therefore non-archimedean. The 'cartesian' geometry based on this field then satisfies all the twenty axioms except V, 1.

Two further metamathematical results, established in the *Grundlagen*, that have become part of common geometrical knowledge are concerned with the theorems of Pappus (or Pascal) and Desargues. The first of these theorems asserts that, if A, B, C and A', B', C' are triads of points on two intersecting lines, then the three points $BC'.B'C$, $CA'.C'A$, $AB'.A'B$ are collinear; and the second asserts that, if ABC and $A'B'C'$ are coplanar triangles such that the lines AA', BB', CC' are concurrent, then the three points $BC.B'C'$, $CA.C'A'$, $AB.A'B'$ are collinear. These theorems, which involve no relations but those of incidence, belong more properly to projective geometry than to euclidean geometry; and the well-known facts about them are that Desargues's theorem can be proved from axioms of incidence for space but not from axioms of incidence for the plane only, while Pappus's theorem cannot be proved at all from axioms of incidence alone, since it demands a commutative ground-field.[2]

---

[1] This model is due to Klein. Its existence establishes the consistency of non-euclidean geometry in much the same way as the existence of the cartesian model establishes the consistency of euclidean geometry. An alternative model for the same type of non-euclidean geometry was devised by Poincaré. On the subject of these models see Note 2, p. 227.

[2] H. F. Baker discusses the theorems of Pappus and Desargues in relation to the axioms of geometry in his *Principles of Geometry*, I, Foundations (1922).

Hilbert confined his discussion of the two theorems to the special cases in which the lines of collinearity that are asserted to exist are at infinity. For Desargues's theorem he proved the following result:

'There exists a plane geometry in which all the axioms I, 1–3, II, III, 1–4, IV*, V hold, i.e. all the linear and plane axioms except for the axiom of congruence III, 5, and for which Desargues's theorem is not true. Desargues's theorem is therefore not a consequence of the specified axioms; it requires for its proof either the space axioms or axiom III, 5 on the congruence of triangles.' The axiom IV* referred to is a strengthened form of the axiom of parallels IV, with the words 'at most one line' replaced by 'one and only one line'.

The treatment of Pappus's theorem depended on a purely algebraic result, to the effect that an ordered skew-field (i.e. an algebraic system for which all the axioms for an ordered field are postulated, with the single exception of the commutative law of multiplication $ab = ba$) is necessarily commutative if it is archimedean. The main conclusion that Hilbert was able to draw from this result was that Pappus's theorem can be proved without the axioms of congruence provided that a form of the Archimedean axiom V, 1 is taken, but not otherwise.

3 Axiomatic theories and their significance

With the publication in 1899 of his *Grundlagen der Geometrie*, Hilbert improved substantially upon the best that had previously been done by way of logical justification of euclidean geometry; and having achieved this initial success, he devoted the major part of his working life from that time on to the much greater task of applying his new metamathematical method to pure mathematics as a whole (though continuing at the same time to pursue more orthodox mathematical research as well). His conception of metamathematics became more precise as the work progressed, and his standard of rigour became correspondingly more stringent—so much so, indeed, that in the end even so modest an objective as the attainment of perfect rigour in the arithmetic of the natural numbers proved to be for ever out of reach. But in spite of the ultimate breakdown of the undertaking, the work done and the results obtained have greatly enhanced our knowledge and understanding of the formal articulation of both mathematics and logic; and through this work, no less than by his more strictly mathematical innovations, Hilbert has done much to shape the mathematical tradition inherited by the mid-twentieth century.

Hilbert evolved a distinctive personal philosophy of mathematics, which had very different presuppositions from the philosophy of Russell that we considered in Chapter 6; and his system is often referred to by the name of *formalism* in order to distinguish it from Russell's logistic system. The results of the researches that have centred round formalism are set forth with admirable lucidity in the two volumes of *Grundlagen der Mathematik*, which Hilbert published in 1934 and 1939 in collaboration with Paul Bernays, whose own further contribution to the foundations of mathematics we shall consider in Chapter 11 below. The account of formalism that now follows is directly based on this standard work, and the details that we are often obliged to omit are all to be found in its pages.

Whereas Russell has consistently maintained that the whole of pure mathematics is reducible to logic, Hilbert firmly believed mathematics to be an autonomous activity of the mathematician; and so formalism is more mathematical than philosophical in spirit, and it reflects very clearly the tendency towards total abstraction that was becoming dominant in mathematics at the time when the formalist outlook was being developed. The general tendency in mathematics in the latter part of the nineteenth century had been in the direction of more and more extensive use of the axiomatic method, and Hilbert characteristically thought of mathematics in axiomatic terms. The entire formalist programme was in fact based on a further development of the axiomatic principle; and in order to appreciate its aims we accordingly need to begin by clarifying the notion of an axiomatic theory.

At the beginning of the *Grundlagen der Mathematik* an important distinction is drawn between two accepted senses of the word 'axiomatic', there distinguished as 'concrete' (*inhaltlich*) and 'formal'. In applying the *concrete* axiomatic method we take some body of empirical knowledge with which we are already tolerably familiar, and we try to make this knowledge systematic by idealizing the concepts which it involves and picking out from among the known facts a small number of basic principles from which all else can be derived by deduction. This is the method that was applied by Euclid to geometry and by Newton to mechanics; and it has been used more recently, though not yet with complete success, in the theory of probability. The *formal* axiomatic method on the other hand, which is the method of pure mathematics *par excellence*, is not tied to any existing factual knowledge, since it is a method of constructing an abstract theory by beginning with arbitrary primitive ideas and primitive propositions, and then developing formal consequences without reference to any sort of meaning.

The concrete axiomatic method is a most powerful intellectual instrument, which can be used whenever a new science is taking shape—that is to say when familiarity with and understanding of some range of phenomena have reached a point at which general laws that govern it begin to be discernible. Use of this method raises mere factual knowledge to the level of scientific theory; and in natural science the method is fully adequate to what is required of it. But in mathematics, where empirical fact provides only the initial prompting for a mathematical investigation (and perhaps the material for subsequent application of the resulting theory), and where both logistic and formalist philosophers are agreed in recognizing only logic as final arbiter in questions of mathematical correctness, it is the *formal* axiomatic method that is needed. Nevertheless, most mathematicians would admit that this formal method is not sufficient by itself. We cannot be content with a mathematics that is imprisoned within a world of total abstraction,[1] a mathematics in which any axiomatic theory that may be arbitrarily set up is as good as any other. Mathematics may indeed recognize no criteria of *correctness* but those of logic, but the *significance* of mathematical theories can hardly be altogether separated from the possibility of their application in wider realms of knowledge. Mathematics is, in fact, indissolubly linked with scientific as well as with everyday thinking, and this connexion forms part of its significance as knowledge. It is for this reason that the literature of foundations of mathematics contains important discussions relating to the *Anwendungsproblem* or problem of application; and this problem obviously calls for use of the concrete axiomatic method. Pure mathematics provides logical analyses of various abstract structural patterns, and a particular mathematical theory can find an application in a particular range of experience if the pattern with which it deals is in fact present in that experience. To reveal the possibility of the application, therefore, and to assess its precision, we need to isolate the pattern of the experience itself, by treating this axiomatically in the concrete manner. That is a topic to which we shall return in Part III (cf. Chapter 13).

Quite apart from the question of whether a given abstract theory applies to a particular body of experience, we may also want to ask whether such a theory is significant *in itself*, that is to say as an abstract theory. A formal axiomatic theory—which is nothing but an edifice of propositions that exists as an organic whole for no other reason than that the propositions are all deducible from certain propositions which have been adopted, by arbitrary fiat, as the axioms of the theory—is not 'significant' in the same sense as a concrete theory, rooted in already familiar knowledge; and for it some special sort of vindication is plainly needed. The most obvious suggestion as to how a formal theory should be exhibited as non-trivial is that we ought to show that the structure which it analyses actually exists in some concrete domain —in other words, that we ought to point to some realization of the theory. But although to do this would be quite sufficient to satisfy mathematicians and scientists that a particular branch of abstract mathematics is significant, no philosophically acceptable demonstration of significance is in fact possible by such means; for almost every body of knowledge bristles with philosophical difficulties, and to try to justify abstract euclidean geometry, say, by pointing to empirically known facts concerning drawing and measuring would be to make the stronger partner lean upon the weaker. It is only in a few cases, moreover, that the theories which make up abstract mathematics have grown directly out of more intuitive knowledge at all, and there are now many branches of pure mathematics that have no immediate intuitive counterpart. Looking at the question of significance from the point of view of philosophy of mathematics, therefore, we must ask with reference to any given abstract mathematical theory not 'Has this theory a realization in some more concrete domain of knowledge?' but rather 'Does this theory in principle admit of realization?'. In other words, we shall ask if it is theoretically possible to replace the primitive concepts of the theory by well-defined specific concepts in such a way that the primitive propositions all become true assertions. Any such interpretation of the undetermined concepts will be said to yield a *realization* of the abstract theory (i.e. a model, in the sense of p. 197 above).

---

[1] It is of some interest to recall that Hilbert himself was an author of a book on intuitive geometry—D. Hilbert and S. Cohn-Vossen: *Anschauliche Geometrie* (Berlin, 1932); English translation *Geometry and the Imagination* (New York, 1956).

The idea of realizability that we have here is obviously related to the idea of satisfiability introduced in Chapter 3, but it applies to entire formal theories instead of to isolated formulae. We said earlier (p. 71) that a formula of the calculus of predicates is *satisfiable* for a given domain of individuals if predicates defined for this domain can be substituted for the predicate variables in it, and particular individuals and propositions for any individual and propositional variables that may be present, in such a way that the formula yields a true proposition. We now say that a formal theory is *realizable* if a domain of individuals can be chosen, and constant individuals and predicates taken, in such a way that, with this particular specification of the primitive elements of the theory, all the axioms are true propositions.

The most direct way of showing that an abstract theory is realizable, and ultimately the only one, is by actually producing a realization, though not necessarily in any concrete domain; but here we already find ourselves confronted with a fundamental difficulty. Only in a very few cases, in fact, can we claim actually to *produce* a realization, namely in those cases in which we can do this with a finite domain of individuals. For, unless the domain of individuals is finite, its very existence raises philosophical problems of the same order as those relating to the abstract system which our construction of the realization is designed to justify. If we are seeking a realization of the abstract theory of groups, say, we need only take some finite group, specified by a multiplication table which we can write out in full, and there, beyond all possibility of doubt, is a demonstration that the theory of groups is realizable. Or again, we have a completely satisfactory realization of abstract projective geometry of two dimensions in the well-known 7-point geometry (see Note 1, p. 224), and this model of the abstract theory can be defined combinatorially by a finite table of incidence relations. To produce in such a way a finite model is to give a perfect proof of realizability, but unfortunately there are quite simple systems of axioms for which no finite model can exist (cf. the example given on p. 72). This means that we are forced to deal with infinite domains of individuals; and so we must consider how these are to be handled.

4 The domain of numerals, treated by finitary means

The simplest infinite totality is that of the natural numbers, a progression which is usually defined implicitly by means of the Peano axioms. But if we adopt such a definition we make the domain of numbers itself dependent on an abstract axiomatic theory, which in turn gives rise to its own question of realizability; and so our whole method is vitiated by circularity. If we are to use the set of natural numbers, or some equivalent totality, as a domain of individuals for the construction of realizations, we must find a way of actually exhibiting this set directly, as an object of immediate perception or intuition, instead of characterizing it implicitly by means of axioms. Admittedly we cannot have the whole infinite collection before us, but we must at least be able to survey as much of it as is needed at any particular moment.

We can make a start by taking as the individuals of our progression the symbols |, ||, |||, . . . , which we obtain one after another by starting with a single vertical stroke and then appending a further stroke at each stage of the construction. For the sake of clarity in the following discussion we shall call these stroke-symbols *numerals*, and we shall use small German letters (in metastatements, or statements about the primary theory) to stand for unspecified numerals. Given any two numerals, we can compare them with regard to length. By crossing off one stroke from each of them again and again we can decide in a finite number of steps whether or not they are of the same length; and if they are not, we find out in this way which is the shorter. We are thus able to define an order among the distinct numerals, by putting $\mathfrak{a} < \mathfrak{b}$ whenever the symbol $\mathfrak{a}$ is shorter than the symbol $\mathfrak{b}$. If $\mathfrak{a}$ and $\mathfrak{b}$ are any two numerals, we define their sum $\mathfrak{a} + \mathfrak{b}$ as the numeral obtained when $\mathfrak{b}$ is written immediately to the right of $\mathfrak{a}$, and we define their product $\mathfrak{a} . \mathfrak{b}$ as the numeral that is obtained when $\mathfrak{a}$ is substituted for each of the strokes in the symbol denoted by $\mathfrak{b}$.

In order that the numerals may provide a non-hypothetical domain of individuals for realizations of axiomatic theories, we must be able to prove, by reasoning that never goes beyond the immediately given, all the arithmetical properties that we shall need in the construction of such realizations. Such non-hypothetical reasoning is referred to by Hilbert and Bernays as 'finitary inference' (*das finite Schließen*). To quote their words:

> 'We shall always use the word "finitary" to indicate that the discussion, assertion, or definition in question is kept within the bounds of thorough-going producibility of objects and thorough-going practicability of processes, and may accordingly be carried out within the domain of concrete inspection.'[1]

What is essential to finitary reasoning is thus that (a) the entities talked about are produced and not merely postulated, and (b) no process of definition or calculation is admitted unless it can be guaranteed to terminate in a finite number of steps, and a definite bound can be set to this number in advance.

All the definitions that we have so far laid down in relation to numerals are finitary; and we can also give finitary interpretations to the two fundamental arithmetical procedures of proof by mathematical induction and definition by recursion. A proof by induction of a theorem $\mathfrak{A}(n)$ gives us a means of verifying $\mathfrak{A}(\mathfrak{n})$, for any assigned numeral $\mathfrak{n}$, in a number of steps to which a bound can be assigned in advance; and a recursive definition of a function $\mathfrak{f}(n)$ similarly enables us actually to evaluate $\mathfrak{f}(\mathfrak{n})$, for any particular $\mathfrak{n}$, also in a previously bounded number of steps. With these basic procedures at our disposal, we can now reformulate in finitary terms the usual arguments by which it is customary to establish the associative, commutative, and distributive laws for addition and multiplication, we can introduce the concept of prime number, and we can prove the 'fundamental theorem of arithmetic', the theorem which asserts that every integer is representable, in one and essentially only one way, as a product of prime factors.

Adoption of the finitary point of view first reveals itself as an irksome limitation on arithmetical thinking when we have occasion to work with propositions that involve quantifiers. An assertion $(x)\mathfrak{A}(x)$ can only be understood as meaning that, for any choice of a numeral $\mathfrak{n}$, the proposition $\mathfrak{A}(n)$ can be shown finitarily to be true; and an assertion $(Ex)\mathfrak{A}(x)$ has to be interpreted in such a way that it can only be made when we actually know some particular numeral $\mathfrak{n}$ for which $\mathfrak{A}(\mathfrak{n})$ is true, or at the very least when we have a process of calculation that can be guaranteed to lead to the determination of such a numeral in a previously bounded number of steps.

---

[1] 'Wir [wollen] allemal mit dem Worte „finit" zum Ausdruck bringen, daß die betreffende Überlegung, Behauptung oder Definition sich an die Grenzen der grundsatzlichen Ausführbarkeit von Prozessen hält und sich somit im Rahmen konkreter Betrachtung vollzieht.' (*Grundlagen der Mathematik*, I, p. 32.)

Finitary interpretation of non-elementary assertions presents even greater difficulty when the propositions that we wish to assert involve negation. The negation of an elementary proposition, that is to say one without quantifiers, is entirely unproblematic, but the situation is altogether changed when quantifiers are admitted. Let us consider first a proposition of the form $\overline{(Ex)\mathfrak{A}(x)}$. The most natural finitary interpretation of the assertion of this proposition is that, for any $\mathfrak{n}$ whatever, $\mathfrak{A}(\mathfrak{n})$ is false. With this interpretation, however, the alternatives $(Ex)\mathfrak{A}(x)$ and $\overline{(Ex)\mathfrak{A}(x)}$ are not exhaustive, for it may well happen that we have no means of determining an $\mathfrak{n}$ for which $\mathfrak{A}(\mathfrak{n})$ is true, and yet are unable to prove the falsity of $\mathfrak{A}(\mathfrak{n})$ for arbitrary choice of $\mathfrak{n}$. The two decisions, in fact, do not involve the same situation. Thus we find that, in the finitary logic of non-elementary propositions, the law of excluded middle 'A or not-A' (*tertium non datur*) fails to hold. Much the same considerations apply also to propositions of the form $\overline{(x)\mathfrak{A}(x)}$. There is more than one possible finitary interpretation of such an assertion—for instance availability of a counter-example, or derivability of a contradiction from the assumption that $\mathfrak{A}(\mathfrak{n})$ holds for arbitrary choice of $\mathfrak{n}$—but in any case the law of excluded middle is invalid.

It will be clear from what we have just seen that the virtual infallibility of finitary reasoning is only achieved at the cost of a heavy sacrifice of deductive power; and it turns out that this type of argument is too weak to enable us to develop a theory of numerals comparable, either in range or in logical simplicity, with ordinary arithmetic.

We see, then, that the arithmetic of the numerals I, II, III, . . . can never be a complete substitute for arithmetic as ordinarily understood by mathematicians, and it cannot even provide a realization of any formal theory that adequately represents ordinary arithmetic. In spite of this, however, the numerals themselves are of great value in metamathematics, for they do at least provide a potentially infinite system of entities, many properties of which can be treated by purely finitary means. The sequence of numerals I, II, III, . . ., in fact, is now available as a finitary counterpart of the sequence of natural numbers 1, 2, 3, ... If, moreover, as is often the case, we wish the progression of numbers to begin with 0 instead of 1, we can easily modify the numerals so that they behave in the corresponding way. All that we need to do is to replace the sequence of expressions I, II, III, . . . by the sequence 0, 0I, 0II, . . ., and to adapt the definitions of addition and multiplication suitably. The sum $\mathfrak{a} + \mathfrak{b}$ of two numerals in the extended sense is obtained by removing the initial '0' from $\mathfrak{b}$ and appending what remains to the right of $\mathfrak{a}$; and the product $\mathfrak{a} . \mathfrak{b}$ is obtained by removing the '0' from $\mathfrak{a}$ and substituting what is left (which may be nothing at all) for each stroke in the constitution of $\mathfrak{b}$.

5 The metamathematics of formalized theories

We now return to the fundamental question of mathematical significance. Pure mathematics, as it is understood nowadays, belongs to what Kant called the realm of ideas—a domain of pure thought, where the mind is no longer bound in any way by the conditions of sensory intuition or the exigencies of contingent fact—and mathematical theories can therefore only be conceived as abstract axiomatic theories. Such a theory must certainly be accepted as significant if a realization of it can be exhibited. When a finite realization is available, significance is guaranteed in the strongest possible way; but we shall also be satisfied if an infinite realization, based on the system of numerals discussed in §4, can be constructed by finitary means. In view of the fact, however, that only so limited a part of the arithmetic of the numerals is finitary, we can scarcely hope to produce even this kind of realization in any but a few favourable cases. The direct method of establishing significance thus soon fails, and we are obliged to have recourse to some weaker form of guarantee. The weakest criterion that can be held to meet the need at all is the one on which Hilbert based his philosophy of formalism, namely demonstrable consistency. Hilbert was prepared to accept any axiomatic theory as significant—that is to say as characterizing a possible structural pattern—if a proof could be given that no contradiction is deducible from its axioms.

Here, then, we have the essential problem of foundations of mathematics as Hilbert understood it in his later work. *A proof must be given, for each branch of mathematics, that the permitted procedures of demonstration can never yield two theorems $\mathfrak{A}$ and not-$\mathfrak{A}$.* This proof of consistency, furthermore, must be logically unassailable; and the metamathematical argument is accordingly required to be strictly finitary.

Now here there is a difficulty; for an ordinary axiomatic mathematical theory is a relatively ill-defined object of thought, whereas it is of the essence of finitary reasoning that it can only be applied to definite and constructively given entities. It was this consideration that produced a clear break in Hilbert's work on foundations of mathematics, which separates the earlier metamathematical discussion of euclidean geometry in the *Grundlagen der Geometrie* from the mature development of the philosophy of formalism, as presented in Hilbert's later papers and in the *Grundlagen der Mathematik*. In his earlier book, Hilbert first of all expounded an axiomatic treatment of geometry which was to be taken as presenting euclidean geometry in its most modern form—the abstract science that had grown out of pre-Euclidean empirical study of space—and the metamathematics that followed was intended simply to demonstrate by ordinary mathematical argument that this theory had certain important characteristics, especially that of consistency. In the *Grundlagen der Mathematik*, on the contrary, the essential task was to *justify* mathematics in its final abstract form, by exclusively finitary means. Here Hilbert was obliged to apply his metamathematical argument not to the mathematical theory itself, which would be built up merely in conformity with the ordinary canons of mathematical rigour, but to a much stricter formalized theory or *formalism*, of the type that we have described in Chapter 4, §3, chosen in such a way that it would faithfully represent the mathematical theory as far as its totality of theorems was concerned.

Formal theories, as set up by Hilbert, are required to be strictly syntactic in their construction. This means that any such theory must be based on a postulated domain of individuals, a finite set of initial formulae, and certain explicitly formulated rules of derivation (say those belonging to the syntactic treatment of the restricted calculus of predicates, as given on p. 66). Its derivable formulae will then be all those formulae that can be obtained from the initial formulae by a finite number of applications of the rules of derivation. The formal theory is to be set up in such a way that every provable theorem of the original mathematical theory has its counterpart in a derivable formula of the formal theory; and if this can be done, then we shall only need to demonstrate the consistency of the formal theory in order to justify (in the sense explained) the mathematical theory that it represents.

Throughout mathematics, frequent and essential use is made of non-finitary arguments, and the formal systems which are constructed to represent mathematical theories will consequently have to provide rules of derivation which correspond to non-finitary inferences. Hilbert considered, when he devised his metamathematical method, that this did not matter, because even if the intuitive interpretation of a formal system is not finitary, the system itself, treated as a totality of derivations (i.e. finite sequences of formulae, put together in accordance with stated rules) is constructively defined, and is therefore a possible object of finitary discussion. The formalist programme thus offered a reasonable hope that a finitary metamathematics would be able to justify a non-finitary mathematics. We must now examine in detail how the programme was put into effect.

6 Consistency of the restricted calculus of predicates

As a preliminary to Hilbert's main undertaking, it was necessary first of all to apply the new method to symbolic logic, by making the logical calculus that was to be used into a strictly formal system, and then establishing its freedom from contradiction by finitary means. The system of logic that was chosen was, in fact, the restricted calculus of predicates, treated syntactically in the manner that we have already specified in Chapter 3. In order to complete our account of this initial phase in the working out of the formalist philosophy of mathematics, therefore, we have only to show how the consistency of the restricted calculus of predicates can be demonstrated. This is a simple matter.

THEOREM 1. *The restricted calculus of predicates is free from contradiction, in the sense that two formulae* $\mathfrak{A}$ *and* $\overline{\mathfrak{A}}$ *cannot both be derived in the system.*

Proof. (i) We consider first of all the propositional calculus, as presented syntactically in Chapter 2, §6, which forms the most elementary part of the restricted calculus of predicates. The only undefined symbols that occur in it are propositional variables (capital Latin letters) and the logical symbols '$\overline{\ }$' and '$\vee$'. Taking two metalogical symbols 'T' and 'F', we can draw up a truth-table for any formula of the calculus, and in this way we can determine finitarily whether or not the formula is an identical one, i.e. one that has 'T' for every entry in its truth-table. It can now readily be shown (again by finitary argument) that (1) the axioms (a)–(d) are identical formulae, and (2) the property of being identical is left invariant by all applications of the rules of procedure (α) and (β). It follows from this that every derivable formula is identical, and hence that two formulae $\mathfrak{A}$ and $\overline{\mathfrak{A}}$ are never both derivable.

(ii) Turning now from the propositional calculus to the full calculus to which the theorem refers, we find that the same proof continues to apply if we simply strike out from every formula of the calculus of predicates any arguments and quantifiers that may be present in it, retaining only the bare capital letters and treating these as if they were propositional variables. The axiom (e) for example, which reads $(x)F(x) \rightarrow F(y)$, is then changed into the identical formula $F \rightarrow F$, i.e. $\overline{F} \vee F$; and the rules of procedure become trivial in a similar way.

Remark. The proof just given is plainly finitary. The idea behind it may be expressed in general terms in the following way. In order to show that two formulae $\mathfrak{A}$ and $\overline{\mathfrak{A}}$ are not both derivable, we introduce constructively a property of symbolic expressions which (i) belongs to all the axioms, (ii) is invariant with respect to the rules of procedure, and is therefore inherited by every derived formula, and (iii) cannot belong to both $\mathfrak{A}$ and $\overline{\mathfrak{A}}$. It is an extension of this same idea that makes possible the proofs of consistency, shortly to be discussed, that Hilbert devised for various arithmetical formalisms.

7 Consistency of arithmetic

After carrying out his programme in the simple case of logic, Hilbert turned his attention to arithmetic, first of all taking a very small part of arithmetic and justifying it, and then by degrees strengthening the axioms so that more and more of ordinary arithmetic was represented in the formal system. The successive phases of this inquiry are discussed in the first volume of *Grundlagen der Mathematik*.

The systems that are considered are all based on Peano's formalization of arithmetic in terms of the three primitive concepts of number, zero, and successor. A domain of individuals is postulated, corresponding to Peano's class N_0 of numbers; a particular individual of this domain is designated by the (constant) symbol '0'; and a mathematical function a', which associates with every term $\mathfrak{a}$ a uniquely defined term $\mathfrak{a}'$, serves to represent the successor function in the formalism.

In the first and weakest of the successive arithmetical formalisms—which we shall call the system (A_0), although it is unnamed in the *Grundlagen*—there are two constant predicates, each with two arguments, namely the logical predicate of identity, $a = b$, and the arithmetical predicate of inequality, $a < b$. Seven axioms are taken in addition to the axioms (a)–(f) of the restricted calculus of predicates. These are the two axioms (J_1) and (J_2) for identity, three axioms $(<_1)$, $(<_2)$, $(<_3)$ for order of magnitude, and two of the Peano axioms, (P_1) and (P_2), for the successor function. We thus have the following set of axioms:

$$\left.\begin{array}{ll}
(\mathrm{J}_1) & a = a, \\
(\mathrm{J}_2) & a = b \rightarrow (A(a) \rightarrow A(b)), \\
(<_1) & \overline{a < a}, \\
(<_2) & (a < b \;\&\; b < c) \rightarrow a < c, \\
(<_3) & a < a', \\
(\mathrm{P}_1) & a' \neq 0, \\
(\mathrm{P}_2) & a' = b' \rightarrow a = b.
\end{array}\right\} \quad (\mathrm{A}_0)$$

As we can easily see (cf. p. 72), the three axioms for order of magnitude are not satisfiable with a finite number of individuals, and the domain of individuals now postulated must accordingly be infinite.

The rules of derivation for the arithmetical formalism are those of the restricted calculus of predicates, trivially modified where necessary to take account of the fact that we now have certain constant symbols '0', '′', '=', '<' in addition to the symbols of the logical calculus itself. As usual, we allow all identical formulae of the propositional calculus to be used as initial formulae in derivations (cf. p. 68).

In order to prove that a formal system is consistent, we have to show that no two contradictory formulae $\mathfrak{A}$ and $\overline{\mathfrak{A}}$ can both be derived. This is something that is often best done indirectly, by showing that some particular formula is underivable in the system. In view of the usefulness of the principle that allows us to transform the problem in this manner, we shall begin our discussion of consistency by expressing this principle formally in a metamathematical theorem:

THEOREM 2. *A formal system* F *is consistent if and only if there is some formula of the system that is not derivable.*

Proof. (i) If F is consistent then, for any arbitrary choice of a formula $\mathfrak{A}$, it is impossible that both $\mathfrak{A}$ and $\overline{\mathfrak{A}}$ should be derivable, and it follows from this that $\mathfrak{A} \;\&\; \overline{\mathfrak{A}}$ is an underivable formula of the system.

(ii) If F is inconsistent, there are some two formulae $\mathfrak{A}$ and $\overline{\mathfrak{A}}$ that are both derivable. But then, if $\mathfrak{B}$ is any formula whatever, $\mathfrak{B}$ can be derived with the aid of the identical formula $X \rightarrow (\overline{X} \rightarrow Y)$ of the propositional calculus.

To establish the consistency of the arithmetical system (A_0), then, we require to show that there is some formula that cannot be derived. The formula chosen for this role by Hilbert was $0 \neq 0$, and we shall now give an outline of Hilbert's metamathematical proof that no derivation in the system (A_0) can have $0 \neq 0$ for its end-formula.

The proof hinges on the discovery of a property of formulae of the system that can be shown to belong to every derivable formula but not to $0 \neq 0$. When proving the consistency of the propositional calculus, we were able to obtain a property of the required kind by turning to the semantic interpretation of propositional formulae as truth-functions of their arguments; and we could do this because the use of truth-tables is an essentially finitary procedure. In the present situation we again find ready to hand an intuitive interpretation of the symbolic expressions—namely their interpretation as statements in elementary arithmetic—and this suggests that we might try to use the elementary interpretation in order to define a suitable formal property. But this interpretation, as we have already seen in §4, is not wholly finitary. Indeed, if it were finitary we would hardly need to apply the indirect procedures of formalist metamathematics to this part of mathematics. Some means has therefore to be found of extracting an appropriate finitary property of formal expressions from their intuitive arithmetical meaning. This has in fact been done by Hilbert, with the aid of an ingenious process of *reduction*, which plays a most important part in the whole metamathematical treatment of arithmetic.

What reduction achieves is the transformation of formulae that contain bound variables, without essential change of their intuitive truth-content, into formulae that are free from such variables. This is just what is needed, as it is precisely the presence of bound variables that makes arithmetic non-finitary. We shall see, moreover, that a process of reduction is only needed for formulae that do not contain propositional or predicate variables.

To illustrate the nature of the process of reduction, let us take the simple formula

$$(Ex)(\mathfrak{a} < x'' \ \& \ x'' < \mathfrak{b}),$$

in which $\mathfrak{a}$ and $\mathfrak{b}$ are supposed to be given terms. To say that there is an x such that $\mathfrak{a} < x'' < \mathfrak{b}$ is to say that there is a $y \geqslant 0''$ such that $\mathfrak{a} < y < \mathfrak{b}$; and this is the case if and only if $\mathfrak{b}$ exceeds $0''$ and there is at least one number between $\mathfrak{a}$ and $\mathfrak{b}$. We can accordingly replace the original existential statement by the simpler statement

$$0'' < \mathfrak{b} \ \& \ \mathfrak{a}' < \mathfrak{b}.$$

This is, in fact, the formula that would be obtained if the process of reduction, now to be defined, were applied to the original formula. It will be found that the fact that the natural numbers form a progression makes it always possible for us to eliminate existential quantifiers from formulae of the system (A_0) without altering their import as arithmetical statements.

Suppose, then, that $\mathfrak{F}$ is some formula of (A_0) which does not contain any propositional or predicate variable. If $\mathfrak{F}$ has no bound variables it requires no reduction, but if bound variables are present in it we proceed in the following manner: We first of all replace every universal quantifier by an existential quantifier, by the device of substituting $\overline{(E\mathfrak{x})\overline{B(\mathfrak{x})}}$ for $(\mathfrak{x})B(\mathfrak{x})$. Then we select an 'innermost' component of $\mathfrak{F}$ of the form $(E\mathfrak{x})\mathfrak{A}(\mathfrak{x})$, that is to say a component of this form which is such that $\mathfrak{A}(\mathfrak{x})$ does not involve any further existential quantifier. $\mathfrak{A}(\mathfrak{x})$ is therefore a combination, by means of the logical operators of the propositional calculus, of equalities $\mathfrak{a} = \mathfrak{b}$ and inequalities $\mathfrak{a} < \mathfrak{b}$, where $\mathfrak{a}$ and $\mathfrak{b}$ denote numbers (i.e. symbols of the form 0, $0'$, $0''$, etc.) or individual variables, free or bound, possibly with appended dashes. In th metamathematical discussion which follows, we shall find it convenien to write a term $\mathfrak{a}''^{\cdots\prime}$, in which $\mathfrak{t}$ dashes are appended to $\mathfrak{a}$, as $\mathfrak{a}^{(\mathfrak{t})}$.

Having selected a suitable component $(E\mathfrak{x})\mathfrak{A}(\mathfrak{x})$ of $\mathfrak{F}$, we apply certain transformations to $\mathfrak{A}(\mathfrak{x})$, as follows:

(i) We first put $\mathfrak{A}(\mathfrak{x})$ into disjunctive normal form, so that it becomes a disjunction $\mathfrak{C}_1 \vee \ldots \vee \mathfrak{C}_r$, in which each component $\mathfrak{C}_i$ is a conjunction of equalities and inequalities and negations of equalities and inequalities.

(ii) We next eliminate all the negations, by replacing $\mathfrak{a} \neq \mathfrak{b}$ by $\mathfrak{a} < \mathfrak{b} \vee \mathfrak{b} < \mathfrak{a}$ and $\overline{\mathfrak{a} < \mathfrak{b}}$ by $\mathfrak{a} = \mathfrak{b} \vee \mathfrak{b} < \mathfrak{a}$; and then we put the expression once again into disjunctive normal form $\mathfrak{F}_1 \vee \ldots \vee \mathfrak{F}_n$.

(iii) In every equality or inequality of the form $\mathfrak{x}^{(\mathfrak{t})} = \mathfrak{x}^{(\mathfrak{l})}$ or $\mathfrak{x}^{(\mathfrak{t})} < \mathfrak{x}^{(\mathfrak{l})}$, with $\mathfrak{x}$ occurring on both sides, we substitute 0 for $\mathfrak{x}$, thus obtaining $0^{(\mathfrak{t})} = 0^{(\mathfrak{l})}$ or $0^{(\mathfrak{t})} < 0^{(\mathfrak{l})}$.

(iv) We now determine the maximum number of dashes, $\mathfrak{t}$ say, appended to $\mathfrak{x}$ anywhere in the formula $\mathfrak{F}_1 \vee \ldots \vee \mathfrak{F}_n$, as modified in (iii), and we then make up the number of dashes on every $\mathfrak{x}$ in the formula to $\mathfrak{t}$ by adding dashes to both sides of an equality or inequality wherever necessary. For example, the formula

$$(0 < \mathfrak{x}' \mathbin{\&} \mathfrak{x} < a) \lor (\mathfrak{x}''' = a'')$$

would have to be changed into

$$(0'' < \mathfrak{x}''' \mathbin{\&} \mathfrak{x}''' < a''') \lor (\mathfrak{x}''' = a'').$$

The formula $\mathfrak{A}(\mathfrak{x})$ will now have been transformed into a formula which we may write as

$$\mathfrak{C}_1(\mathfrak{x}^{(\mathfrak{t})}) \lor \ldots \lor \mathfrak{C}_m(\mathfrak{x}^{(\mathfrak{t})}) \lor \mathfrak{C}_{m+1} \lor \ldots \lor \mathfrak{C}_n,$$

each disjunctive component $\mathfrak{C}_i(\mathfrak{x}^{(\mathfrak{t})})$, $i = 1, \ldots, m$, being a conjunction of expressions of the possible forms $\mathfrak{x}^{(\mathfrak{t})} = \mathfrak{a}$, $\mathfrak{x}^{(\mathfrak{t})} < \mathfrak{a}$, $\mathfrak{a} < \mathfrak{x}^{(\mathfrak{t})}$, $\mathfrak{a} = \mathfrak{b}$, $\mathfrak{a} < \mathfrak{b}$; and we accordingly replace $(E\mathfrak{x})\mathfrak{A}(\mathfrak{x})$ by the disjunction

$$(E\mathfrak{x})\mathfrak{C}_1(\mathfrak{x}^{(\mathfrak{t})}) \lor \ldots \lor (E\mathfrak{x})\mathfrak{C}_m(\mathfrak{x}^{(\mathfrak{t})}) \lor \mathfrak{C}_{m+1} \lor \ldots \lor \mathfrak{C}_n.$$

Then, for each of the m formulae $(E\mathfrak{x})\mathfrak{C}_i(\mathfrak{x}^{(\mathfrak{t})})$, there are two possibilities to be considered.

(1) At least one equality $\mathfrak{x}^{(\mathfrak{t})} = \mathfrak{a}$ occurs in the conjunction $\mathfrak{C}_i(\mathfrak{x}^{(\mathfrak{t})})$, so that $(E\mathfrak{x})\mathfrak{C}_i(\mathfrak{x}^{(\mathfrak{t})})$ has the form

$$(E\mathfrak{x})(\mathfrak{x}^{(\mathfrak{t})} = \mathfrak{a} \mathbin{\&} \mathfrak{C}_i^*(\mathfrak{x}^{(\mathfrak{t})})).$$

It is understood, of course, that the residual part $\mathfrak{C}_i^*(\mathfrak{x}^{(\mathfrak{t})})$ of the quantified expression need not be present at all, and if it is present it need not actually involve $\mathfrak{x}$. When $(E\mathfrak{x})\mathfrak{C}_i(\mathfrak{x}^{(\mathfrak{t})})$ is of the form just given, we replace it by the formula

$$(0^{(\mathfrak{t})} = \mathfrak{a} \lor 0^{(\mathfrak{t})} < \mathfrak{a}) \mathbin{\&} \mathfrak{C}_i^*(\mathfrak{a}),$$

in which there no longer appears either the variable $\mathfrak{x}$ or the quantifier $(E\mathfrak{x})$.

(2) The conjunction $\mathfrak{C}_i(\mathfrak{x}^{(\mathfrak{t})})$ involves $\mathfrak{x}^{(\mathfrak{t})}$ in inequalities only. In this case we may write $(E\mathfrak{x})\mathfrak{C}_i(\mathfrak{x}^{(\mathfrak{t})})$ in the form

$$\mathfrak{H} \mathbin{\&} (E\mathfrak{x})(\mathfrak{a}_1 < \mathfrak{x}^{(\mathfrak{t})} \mathbin{\&} \ldots \mathbin{\&} \mathfrak{a}_k < \mathfrak{x}^{(\mathfrak{t})} \mathbin{\&} \mathfrak{x}^{(\mathfrak{t})} < \mathfrak{b}_1 \mathbin{\&} \ldots \mathbin{\&} \mathfrak{x}^{(\mathfrak{t})} < \mathfrak{b}_p),$$

where $\mathfrak{H}$ is the conjunction of all the components of $\mathfrak{C}_i(\mathfrak{x}^{(\mathfrak{t})})$ that do not involve $\mathfrak{x}^{(\mathfrak{t})}$. We then replace $(E\mathfrak{x})\mathfrak{C}_i(\mathfrak{x}^{(\mathfrak{t})})$ by the conjunction

$$\begin{aligned}
\mathfrak{H} \mathbin{\&} 0^{(\mathfrak{t})} &< \mathfrak{b}_1 \mathbin{\&} \ldots \mathbin{\&} 0^{(\mathfrak{t})} < \mathfrak{b}_p \mathbin{\&} \\
\mathbin{\&} \mathfrak{a}_1' &< \mathfrak{b}_1 \mathbin{\&} \ldots \mathbin{\&} \mathfrak{a}_1' < \mathfrak{b}_p \mathbin{\&} \\
\mathbin{\&} \mathfrak{a}_2' &< \mathfrak{b}_1 \mathbin{\&} \ldots \mathbin{\&} \mathfrak{a}_2' < \mathfrak{b}_p \mathbin{\&} \\
&\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\
\mathbin{\&} \mathfrak{a}_k' &< \mathfrak{b}_1 \mathbin{\&} \ldots \mathbin{\&} \mathfrak{a}_k' < \mathfrak{b}_p.
\end{aligned}$$

If there is no component of the form $\mathfrak{a} < \mathfrak{x}^{(\mathfrak{t})}$, then only the first line is required; and if there is no component of the form $\mathfrak{x}^{(t)} < \mathfrak{b}$, then $(E\mathfrak{x})\mathfrak{E}_i(\mathfrak{x}^{(t)})$ may be replaced simply by $\mathfrak{H}$ if this expression is non-null and by $0 = 0$ otherwise.

When the process just described has been completed, one quantifier will have been eliminated from the original formula $\mathfrak{F}$ without change of the *intuitive* truth or falsity of the arithmetical statement to which $\mathfrak{F}$ corresponds if it contains no free variables, and without alteration in $\mathfrak{F}$ as a 'truth-function' otherwise; and a finite number of applications of the process therefore suffices to remove all the quantifiers from $\mathfrak{F}$. The formula which finally results is then free from bound variables; it contains exactly the same free variables as the original formula $\mathfrak{F}$; and its 'truth content' is the same as that of $\mathfrak{F}$.

Exercise. Apply the process of reduction to the formula

$$(x)(x < a \rightarrow b < x) \rightarrow a = 0,$$

in which a and b are free variables.

We can now use the procedure of reduction in order to define the metamathematical concept of verifiability for formulae of the arithmetical system (A_0) which do not contain propositional or predicate variables.

DEFINITION. (a) A numerical formula (i.e. one without variables) is *verifiable* if it is true when interpreted in terms of the finitary arithmetic of numerals.

(b) A formula with free individual variables only is *verifiable* if a finitary proof can be given that it becomes a true statement whenever definite numerals are substituted for the variables in it.

(c) A formula with bound variables is *verifiable* if it can be transformed by the process of reduction into a formula that is verifiable in the sense of (a) or (b).

It should be observed that part (c) of the definition requires justification. The process of reduction is not unique (because of the ambiguity of the disjunctive normal form) and we need to show that all possible reductions of a formula lead to the same decision as to its verifiability. The proof of this is long and rather tedious but full details are given by Hilbert and Bernays.

The property of verifiability, as just defined, does not take us outside the sphere of finitary argument, and we are therefore at liberty to use it for metamathematical purposes. By its aid we are able, in fact, to adapt to the formal system (A_0) the general method for proving consistency that we have already described (p. 211). We require to show that the formula $0 \neq 0$ is not derivable; and this will be accomplished if we can demonstrate by finitary means that, if a formula is derivable, then it is also verifiable.

THEOREM 3. *The formalized system of arithmetic* (A_0) *is free from contradiction.*

Proof. Let ($\mathfrak{F}_i$) be any derivation in the system (A_0). We prove that the end-formula of this derivation cannot be $0 \neq 0$; and to do so we divide the proof into two parts.

(i) In the first part of the proof we assume that no bound variables occur in the given derivation ($\mathfrak{F}_i$). Thus the available initial formulae are the identical formulae of the propositional calculus and the seven special axioms (J_1), (J_2), ($<_1$), ($<_2$), ($<_3$), (P_1), (P_2); and the only schema that can be used is the rule of inference (β), as given on p. 67. The derivation ($\mathfrak{F}_i$) is therefore a finite sequence of formulae, each of which is either (i) an identical formula of the propositional calculus or one of the seven special axioms, (ii) a repetition of an earlier formula of the sequence ($\mathfrak{F}_i$) or else a formula obtained by making a substitution in an earlier formula, or (iii) a formula obtained by use of the schema (β).

Our first step is one of unravelling the given derivation ($\mathfrak{F}_i$) into its separate strands (*Auflösung in Beweisfäden*), and this is achieved by starting with the end-formula and working back systematically to the initial formulae of the derivation. We set out the result of the process schematically in the following manner. First of all we put down the end-formula in the middle of the page. Then above it we place the formula or pair of formulae from which it is immediately derived; above each such formula we place in the same way the formula or formulae from which it is immediately derived; and so on, until we have worked back in every instance to an initial formula. If a formula $\mathfrak{W}$ in the resolved derivation is derived from two formulae $\mathfrak{U}$ and $\mathfrak{B}$ by use of the schema (β) we put

$$\begin{tikzcd}
\mathfrak{U} \arrow[dr, dash] & & \mathfrak{B} \arrow[dl, dash] \\
& \mathfrak{W},
\end{tikzcd}$$

while if a formula $\mathfrak{W}$ is derived from a formula $\mathfrak{B}$ by repetition or substitution we put

$$\begin{tikzcd}
\mathfrak{B} \arrow[d, dash] \\
\mathfrak{W}.
\end{tikzcd}$$

In order to avoid confusion in the diagram it is desirable to begin by numbering off the successive formulae of the sequence ($\mathfrak{F}_i$), so that these can be represented by their numbers in the resolved derivation ($\mathfrak{F}_i'$) when this is displayed schematically.

As an illustration of the process of resolution into strands, we give here an example which Hilbert and Bernays discuss on p. 226 of their first volume. The formula $0'''' \neq 0''$ may be derived as follows:

$$\begin{array}{rc}
(1) & (A \rightarrow B) \rightarrow (\bar{B} \rightarrow \bar{A}) \\
(2) & (a' = b' \rightarrow a = b) \rightarrow (a \neq b \rightarrow a' \neq b') \\
(3) & a' = b' \rightarrow a = b \\
\cline{2-2}
(4) & a \neq b \rightarrow a' \neq b' \\
(5) & a' \neq 0 \rightarrow a'' \neq 0' \\
(6) & a' \neq 0 \\
\cline{2-2}
(7) & a'' \neq 0' \\
(8) & a'' \neq 0' \rightarrow a''' \neq 0'' \\
\cline{2-2}
(9) & a''' \neq 0'' \\
(10) & 0'''' \neq 0''.
\end{array}$$

Key:

(1): [identical formula]
(2): [(1) $a' = b'/A$, $a = b/B$]
(3): [(P_2)]
(4): [(3), (2): (β)]
(5): [(4) a'/a, $0/b$]
(6): [(P_1)]
(7): [(6), (5): (β)]
(8): [(4) a''/a, $0'/b$]
(9): [(7), (8): (β)]
(10): [(9) $0'/a$].

The derivation ($\mathfrak{F}_i$) that we have here, when written schematically as a resolved derivation ($\mathfrak{F}_i'$), takes the form:

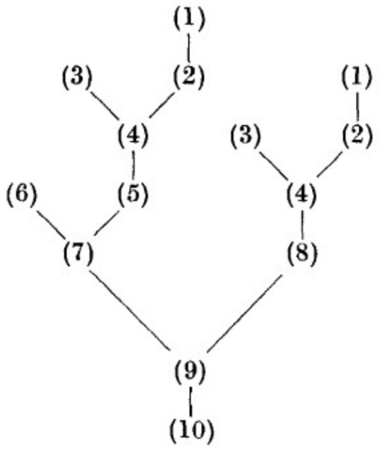

The formula (10) is the end-formula, and (1), (3), (6) are the initial formulae that are used.

Having unravelled the strands of ($\mathfrak{F}_i$), we now work backwards through $(\mathfrak{F}_i')$, taking each strand in turn and going always from the end-formula of the derivation right back to the initial formula in which the strand originates. In the example given, we would have to follow five different strands: (i) (10) → (9) → (7) → (6); (ii) (10) → (9) → (7) → → (5) → (4) → (3); (iii) (10) → (9) → (7) → (5) → (4) → (2) → (1); (iv) (10) → (9) → (8) → (4) → (3); (v) (10) → (9) → (8) → (4) → (2) → → (1). As we work along the strands, we carry all substitutions back to the initial formulae. In other words, whenever we come to a link

$$\begin{tikzcd}
\mathfrak{B} \arrow[d, dash] \\
\mathfrak{W}
\end{tikzcd}$$

which is such that $\mathfrak{W}$ is obtained from $\mathfrak{B}$ by one or more substitutions, we make the inverse substitutions in $\mathfrak{B}$ and in all the formulae that stand higher than $\mathfrak{B}$ in the strands which pass through $\mathfrak{B}$. Let the derivation $(\mathfrak{F}_i')$, as thus modified, be denoted by $(\mathfrak{F}_i'')$. Then, whenever we have a pair of formulae

$$\begin{tikzcd}
\mathfrak{B} \arrow[d, dash] \\
\mathfrak{W}
\end{tikzcd}$$

in $(\mathfrak{F}_i'')$, $\mathfrak{W}$ is simply a repetition of $\mathfrak{B}$; and whenever we have a triad of formulae

$$\begin{tikzcd}
\mathfrak{U} \arrow[dr, dash] & & \mathfrak{B} \arrow[dl, dash] \\
& \mathfrak{W} &
\end{tikzcd}$$

in $(\mathfrak{F}_i'')$, the passage from $\mathfrak{U}$ and $\mathfrak{B}$ to $\mathfrak{W}$ is a valid application of the rule (β). Thus $(\mathfrak{F}_i'')$ only differs from a proper derivation in that its initial formulae are not restricted to the identical formulae of the propositional calculus and the seven special axioms, but may be any formulae derived from these by substitution.

It is possible that some free variables may still be left in $(\mathfrak{F}_i'')$. This would be the case, for example, if $(\mathfrak{F}_i')$ contained some such sub-derivation as

$$\begin{tikzcd}
A \vee \overline{A} \arrow[d, dash] & & \\
A(a) \vee \overline{A(a)} \arrow[dr, dash] & & (A(a) \vee \overline{A(a)}) \rightarrow B \arrow[dl, dash] \\
& B \arrow[d, dash] & \\
& {}. &
\end{tikzcd}$$

We accordingly eliminate all remaining free variables from $(\mathfrak{F}_i)$ by making the following substitutions: $0 = 0$ for every propositional variable; $\mathfrak{a} = \mathfrak{a}$ for every predicate variable with a single argument $\mathfrak{a}$; $(\mathfrak{a} = \mathfrak{a} \,\&\ldots\&\, \mathfrak{t} = \mathfrak{t})$ for every predicate variable with the arguments

$\mathfrak{a}, \ldots, \mathfrak{t}$; and 0 for every individual variable. Then the array of formulae that we are left with is still a derivation in the same extended sense as $(\mathfrak{F}_i'')$; and every formula in the array is a numerical formula, that is to say a combination by means of the operators of the propositional calculus of equalities $\mathfrak{s} = \mathfrak{t}$ and inequalities $\mathfrak{s} < \mathfrak{t}$ between definite 'numbers' 0, $0'$, $0''$, etc. If, moreover, the original derivation $(\mathfrak{F}_i)$ had $0 \neq 0$ as its end-formula, the derivation yielded by the process that we have just described would also end with $0 \neq 0$.

Now it can be verified, by finitary argument, that (1) the initial formulae of the modified derivation are all true statements when they are given an interpretation based on the finitary arithmetic of the numerals 0, 0|, 0||, ...; (2) both the operation of repeating a formula and the operation of applying the rule (β) leave truth in this sense invariant; and (3) the formula $0 \neq 0$ is false in the corresponding sense. We thus have the required finitary proof of the underivability of the formula $0 \neq 0$, when 'derivable' is understood as meaning 'derivable without use of bound variables'. Our further task, in part (ii) of the proof, is to remove this limitation.

(ii) We now suppose that the derivation $(\mathfrak{F}_i)$ with which we begin uses the full resources of the restricted calculus of predicates. This means that the axioms (e) and (f), p. 66, which involve quantification, can now occur as initial formulae, and the two schemata ($\gamma 1$) and ($\gamma 2$) can now be used in addition to the schema (β).

We begin, as in part (i) of the proof, by unravelling the strands of the derivation, and by carrying back all substitutions to the initial formulae. Having done this, we eliminate all propositional and predicate variables as in case (i). But this time, because of the free variables that are essential to the schemata ($\gamma 1$) and ($\gamma 2$), we cannot eliminate all free *individual* variables without invalidating the derivation. Neither can we eliminate the bound (individual) variables that may occur. Consequently we are unable to transform the given sequence of formulae $(\mathfrak{F}_i)$ into an array of numerical formulae that retains the structure of a derivation, and so to prove as in (i) that the end-formula cannot be the false formula $0 \neq 0$. We can, however, use verifiability in place of numerical truth; and then a similar method of proof applies. We show that the initial formulae are all verifiable and that the property of verifiability is inherited by every formula in the modified derivation, and it then follows that $0 \neq 0$ cannot be the end-formula. The details of the argument, which we omit from this summary, are given in full by Hilbert and Bernays.

Here, then, we have in outline the proof of the first consistency theorem for formalized arithmetic, and we must now assess the value of what has been established. According to the formalist programme, in order to justify ordinary mathematics we are to formalize it and then to show that the formal system is free from contradiction. The proof just given has shown that the initial formalization of part of arithmetic by the system (A_0) is indeed consistent; and we must now ask how adequate this formalization is as a representation of intuitive arithmetic. In other words, we must raise the question of the completeness of the system. A formal theory may be said to be (semantically) complete if every intuitively valid statement that can be made in the symbolism of the theory is derivable; and in the present situation we shall make the notion of 'intuitively valid' precise by interpreting it to mean verifiable in the sense already laid down. The question, therefore, is whether every verifiable formula of the system (A_0) is derivable; and the answer is that this is not the case. Hilbert and Bernays prove (*Grundlagen*, p. 252) that the verifiable formula

$$a < b \rightarrow (a' = b \vee a' < b)$$

is not derivable from the axioms. To do this, they use the standard method for proving independence (cf. p. 198 above), devising a suitably distorted property of 'verifiability', which holds for the axioms and is invariant with respect to the rules of procedure, but does not hold for the formula in question. They then go on to prove (pp. 253–260) that all formulae that are verifiable in the original proper sense become derivable if the following five formulae are added to the axioms:

$$\begin{aligned}
&a \neq b \rightarrow (a < b \vee b < a),\\
&a < b \rightarrow a' < b',\\
&0 = a \vee 0 < a,\\
&0 < a \rightarrow (Ex)(x' = a),\\
&(a < c \;\&\; c < b) \rightarrow a' < b.
\end{aligned}$$

When these new axioms are added, some of the original ones become derivable, and hence superfluous. The extended list of axioms can eventually be condensed into the following set of seven only:

$$\left.\begin{aligned}
&a = b \rightarrow (A(a) \rightarrow A(b)),\\
&(a < b \;\&\; b < c) \rightarrow a < c,\\
&a \neq b \rightarrow (a < b \vee b < a),\\
&a < a',\\
&a < b \rightarrow \overline{b < a'},\\
&\overline{a < 0},\\
&a \neq 0 \rightarrow (Ex)(x' = a).
\end{aligned}\right\} \quad (\mathrm{A})$$

It is proved in the *Grundlagen* that these seven axioms are independent, and further that *a formula without propositional or predicate* *variables is derivable from them if and only if it is verifiable.* This last result shows that the system (A) is both consistent and complete.

All of Peano's five axioms except the axiom of induction are derivable in the system (A), but the axiom of induction is not derivable. This last statement is certainly plausible, for the axiom of induction

$$[A(0) \,\&\, (x)(A(x) \rightarrow A(x'))] \rightarrow A(a)$$

involves a predicate variable, and the only one of the axioms (A) in which any variable other than an individual variable occurs is the second axiom for identity (J_2). A formal proof of the independence of the axiom of induction is given by Hilbert and Bernays (p. 273).

We may note, incidentally, that instead of representing the principle of mathematical induction by the axiom just given, we could equally well formalize it by a schema:

$$\frac{\begin{array}{c}\mathfrak{A}(0)\\ \mathfrak{A}(a) \rightarrow \mathfrak{A}(a')\end{array}}{\mathfrak{A}(a).}$$

The two alternative ways of incorporating the principle in the formal system can be shown to have identical deductive power (*Grundlagen*, p. 266).

If we add the axiom of induction to the set of axioms (A), some of these axioms can be weakened or omitted altogether; and we thus arrive at the following revised set of axioms:

$$\left.\begin{array}{l} a = b \rightarrow (A(a) \rightarrow A(b)),\\ \overline{a < a},\\ (a < b \,\&\, b < c) \rightarrow a < c,\\ a < a',\\ a < b \rightarrow (a' = b \vee a' < b),\\ 0 = 0,\\ [A(0) \,\&\, (x)(A(x) \rightarrow A(x'))] \rightarrow A(a). \end{array}\right\} \quad \text{(B)}$$

The two systems (A) and (B) can be shown to be equivalent *for the derivation of formulae without propositional or predicate variables*—that is to say, any such formula that is derivable in one of the systems is also derivable in the other—but the systems are not equivalent for the derivation of formulae in which propositional and predicate variables may occur. The last assertion follows from the non-derivability of the axiom of induction from the set of axioms (A).

In the system (B) we are able to derive the very important arithmetical *principle of the least number* (*Grundlagen*, p. 284), expressed by the formula

$$A(a) \rightarrow (Ex)[A(x) \,\&\, (y)(A(y) \rightarrow (x = y \vee x < y))].$$

This principle is best utilized in conjunction with the ι-symbol, which provides an explicit expression for the least number in question. Let us denote by $\mathfrak{M}(c)$ the formula

$$A(c)\ \&\ (y)(A(y) \rightarrow (c = y \vee c < y)).$$

Then the formula just quoted may be written

$$A(a) \rightarrow (Ex)\mathfrak{M}(x).$$

By means of this formula we can derive (*Grundlagen*, p. 395) the two uniqueness formulae for the formula

$$((z)\overline{A(z)} \rightarrow a = 0)\ \&\ ((Ez)A(z) \rightarrow \mathfrak{M}(a)),$$

in which a is a free variable; and we are therefore at liberty to introduce the corresponding ι-term by means of the explicit definition

$$\mu_x A(x) \equiv \iota_x[((z)\overline{A(z)} \rightarrow x = 0)\ \&\ ((Ez)A(z) \rightarrow \mathfrak{M}(x))].$$

The term $\mu_x A(x)$ is the formal counterpart of the intuitive concept 'the least number with the property A if there is any such number, and 0 otherwise'.

The theorem on the eliminability of the ι-symbol (see above, p. 95) now permits us to assert, without further discussion, that *any formula, not itself involving the μ-symbol, that can be derived from the set of axioms* (B) *with the aid of the μ-symbol is also derivable from these axioms without use of the μ-symbol.*

Looking now at the way in which the axioms (B) have been obtained, we see that all of Peano's axioms are incorporated in the formal system, either implicitly through the symbolism or directly as derivable formulae; and this may seem to imply that the system is no less comprehensive than Peano's arithmetic. Such a conclusion is unwarranted, however, since Peano's treatment of arithmetic is actually based on more than the bare axioms. In the *Formulaire* extensive use is made of definitions, and we have already noted (p. 147) that as well as explicit definitions, which serve merely to introduce abbreviations, there are also *recursive definitions*, which are definitions in quite another sense and which make substantial additions to the system.

A recursive definition of a function $\mathfrak{f}(n)$ of one variable takes the form of a pair of equations

$$\mathfrak{f}(0) = \mathfrak{a},$$
$$\mathfrak{f}(n') = \mathfrak{b}(n, \mathfrak{f}(n)),$$

where $\mathfrak{a}$ is a term that does not involve any variable, $\mathfrak{b}(x,y)$ is a term that involves the variables x and y only, and neither $\mathfrak{a}$ nor $\mathfrak{b}(x,y)$ involves the function $\mathfrak{f}(-)$. For a function $\mathfrak{f}(a, \ldots, k, n)$, with parameters $a, \ldots, k$, the definition takes the more general form

$$\begin{aligned} \mathfrak{f}(a, \ldots, k, 0) &= \mathfrak{a}(a, \ldots, k), \\ \mathfrak{f}(a, \ldots, k, n') &= \mathfrak{b}(a, \ldots, k, n, \mathfrak{f}(a, \ldots, k, n)). \end{aligned}$$

The adoption, within a given formal system, of a particular recursive definition always adds to the formal resources which the system provides, that is to say to the available means of symbolic expression, since we thereby acquire a new function-symbol; but it does not necessarily render the system any more comprehensive. We need, therefore, to distinguish between essential and non-essential extension of a formal system, the extension occasioned by a particular recursive definition being regarded as non-essential if the new function is one already representable in some alternative way in the original system. It is clearly important to give precise metamathematical meaning to the notion of formal representability of a function, and Hilbert and Bernays adopt the following definition (p. 352):

DEFINITION. Let a new function of one or more variables be adjoined to a given formal system by the introduction of a function-symbol $\mathfrak{f}(a, \ldots, k)$, together with associated formulae which permit calculation of the value of the function for any assigned numerical values of its arguments. The function is said to be *representable* in the original system if there is a formula $\mathfrak{A}(a, \ldots, k, l)$ of the system such that, for every substitution of numerical values $\mathfrak{a}, \ldots, \mathfrak{k}, \mathfrak{l}$ for the variables $a, \ldots, k, l$, the formula $\mathfrak{A}(\mathfrak{a}, \ldots, \mathfrak{k}, \mathfrak{l})$ is derivable if $\mathfrak{l}$ coincides with the value of $\mathfrak{f}(\mathfrak{a}, \ldots, \mathfrak{k})$ and the formula $\overline{\mathfrak{A}(\mathfrak{a}, \ldots, \mathfrak{k}, \mathfrak{l})}$ is derivable otherwise.

It can be shown (*Grundlagen*, p. 356) that the sum function $a + b$, which is definable by the equations of recursion

$$\begin{aligned} a + 0 &= a, \\ a + n' &= (a + n)', \end{aligned}$$

is not representable in the system (B). We accordingly take '+' as a new primitive symbol and incorporate this recursive definition in the system of axioms; and the usual sort of reorganization then leads to the following strengthened system:

$$\left.\begin{aligned} & a = a, \\ & a = b \rightarrow (A(a) \rightarrow A(b)), \\ & a' \neq 0, \\ & a' = b' \rightarrow a = b, \\ & a + 0 = a, \\ & a + b' = (a + b)', \\ & [A(0) \mathbin{\&} (x)(A(x) \rightarrow A(x'))] \rightarrow A(a). \end{aligned}\right\} \quad \text{(D)}$$

The system (D) does not contain '$<$' among its primitive symbols, but the predicate $a < b$ can be introduced by explicit definition, thus:

$$a < b \leftrightarrow (Ex)(x \neq 0 \; \& \; a + x = b).$$

The new system, as we have said, is stronger than the old; but once again we are able to assert that the system that we have before us is both consistent and complete. The process of reduction that we used before can be modified in such a way that it applies to formulae in (D), and we can then show that the totality of verifiable formulae of this system coincides with the totality of derivable formulae.

The assertion that the systems (B) and (D) are both complete, although the second is demonstrably wider than the first, has an air of paradox. There is not really any conflict, however, since completeness, as we have been interpreting this term, is always relative to the symbolic resources of the system concerned. And, even though the system (D) is complete in this internal sense, it is still far from complete as a formalization of intuitive arithmetic. We can show, in fact, that the product function $a.b$ is not representable in (D) (*Grundlagen*, p. 369). We accordingly add the symbol '$.$' and the relevant equations of recursion to the primitive elements of the system (D), passing thus to the following extended system of axioms:

$$\left.\begin{array}{l}
a = a, \\
a = b \rightarrow (A(a) \rightarrow A(b)), \\
a' \neq 0, \\
a' = b' \rightarrow a = b, \\
a + 0 = a, \\
a + b' = (a + b)', \\
a.0 = 0, \\
a.b' = a.b + a, \\
[A(0) \; \& \; (x)(A(x) \rightarrow A(x'))] \rightarrow A(a).
\end{array}\right\} \quad \text{(Z)}$$

Although the passage from (D) to (Z) appears to be of exactly the same kind as that from (B) to (D), involving nothing but the addition to the existing axioms of the recursive definition of one new function, this latest extension of the formalism in reality produces an entirely new situation. With addition and multiplication both at our disposal, we are now able to formalize the whole of the elementary theory of numbers, and we can in consequence represent in the system (Z) various propositions concerning the natural numbers that are as yet undecided, and that may even be undecidable by elementary means.

A famous undecided conjecture is Goldbach's hypothesis, according to which every even integer greater than 2 is expressible as a sum of two prime integers. In order to translate this assertion into the symbolism of the system (Z), we first of all formalize the propositional function 'n is prime' by defining a predicate $\mathfrak{Pr}(n)$ explicitly as follows:

$$\mathfrak{Pr}(n) \leftrightarrow n \neq 0 \;\&\; n \neq 0' \;\&\; (u)(v)((u \neq 0' \;\&\; v \neq 0') \rightarrow u.v \neq n).$$

Then we can express Goldbach's hypothesis in the form

$$(x)[0' < x \rightarrow (Ey)(Ez)(\mathfrak{Pr}(y) \;\&\; \mathfrak{Pr}(z) \;\&\; 0''.x = y + z)].$$

A second well-known conjecture concerning prime numbers is that there are infinitely many prime-pairs, or pairs of prime numbers which differ by only 2. It comes to the same thing to assert that there are arbitrarily large prime-pairs, and so we arrive at the formula

$$(x)(Ey)(x < y \;\&\; \mathfrak{Pr}(y) \;\&\; \mathfrak{Pr}(y'')).$$

The very fact that the formal system (Z) is wide enough to allow such translations into its symbolism to be made, which indicates that we are at last within sight of completeness of a mathematically interesting kind, signalizes also the breakdown of the method for proving consistency that has been applicable up to this point. That method depended on the existence of a process of reduction which could be used to define verifiability; and the effect of reduction was to convert any formula, without essential change of its intuitive truth-content, into a formula that could be handled by finitary means. Now since there are, among the formulae of the system (Z), some which correspond to problems in arithmetic that have for centuries defied all efforts to solve them, it seems very unlikely that any process of reduction for this system can be found.

The failure for the system (Z) of one particular method for proving consistency does not of itself imply that no finitary proof of consistency can be given for this system; but such is in fact now known to be the case on other grounds. One of the most famous theorems in metamathematics, proved in 1931 by Kurt Gödel, placed this virtually beyond doubt, and so brought to an abrupt end the phase of the development of formalism that we have been discussing in this chapter. The work of Gödel on the limitations of formalism is the topic of the chapter that now follows.

SUPPLEMENTARY NOTES ON CHAPTER 7

1. *Foundations of projective geometry.* The study of foundations of mathematics can be interpreted in a restricted sense, as an examination of the logical structure of deductive systems, or more widely, as a partly historical and partly critical consideration of the development of mathematical concepts and methods. From the latter point of view the history of projective

geometry is especially instructive, because of the progressive change of outlook that has taken place as this subject has become more thoroughly understood.

The story of projective geometry begins with Gérard Desargues (1593–1662), who in 1639 published his essay *Brouillon proiect d'une atteinte aux éuénemens des rencontres d'un cone auec un plan* [Sketch of an inquiry into the results of the intersection of a cone with a plane]. In this work Desargues introduced a number of ideas characteristic of projective geometry, as for instance point at infinity, pole and polar, and involution of pairs of points on a line. His argument, however, was too abstruse for all but one or two mathematicians of the time, and the essay was soon entirely forgotten.

After a lapse of nearly two centuries, projective geometry was studied again, independently of Desargues, by Jean-Victor Poncelet (1788–1867). A military engineer by profession, Poncelet served with Napoleon's army; and when he eventually published his geometrical discoveries, in his *Traité des propriétés projectives des figures* (1822), he remarked of the book that it was 'the outcome of the investigations which I have carried on, since the spring of 1813, in the prisons of Russia'. In this book Poncelet developed the conception of duality, making use of the transformation by reciprocal polars determined by a fixed conic; and he also introduced (in an intuitive manner) such completely new ideas as that of the circular points at infinity.

A more systematic, though no less original contribution to projective geometry was made by August Ferdinand Möbius in *Der barycentrische Calcul* (1827). Möbius devised a way of representing the points of the plane (or of space) by homogeneous coordinates referred to an arbitrarily selected triangle (or tetrahedron) of reference—a way, moreover, that is essentially projective and that permits projective relationships to be expressed simply in algebraic terms. He defined cross ratio (as a quotient of position ratios) and he established the main properties of this important invariant. He also studied the 'net' of points—now known as a *Möbius net*—which is obtained from the vertices of a quadrangle when the two operations of joining two points of the plane by a line and taking the point of intersection of two given lines are repeated indefinitely often; and he was able to show that (i) in any coordinate system based on three of the vertices of the quadrangle as fundamental points and the remaining vertex as unit point, the points of the net are precisely those points which have rational coordinates, and (ii) every neighbourhood of any given point of the plane contains points of the net.

Up to this stage in its history, projective geometry had been conceived as just one particular division of the study of figures in ordinary space, namely that part which deals exclusively with relationships that are projectively invariant; but now a more radical way of treating the subject emerged in the work of Georg Karl Christian von Staudt (1798–1868), who was professor of mathematics at Erlangen. Realizing that projective geometry is essentially non-metrical, von Staudt determined to treat it from the outset as a 'geometry of position'; and he carried this project through in his *Geometrie der Lage* (1847) and *Beiträge zur Geometrie der Lage* (1856–1860). He began with an intuitively conceived space, in which the concepts of straight line and plane had meaning, but no concept of distance was introduced; and he then attempted to build up the whole of projective geometry in terms of relations of incidence. He had no means of introducing the metrical notion of cross ratio, but he was nevertheless able to define the harmonic relation by the quadrangle construction and then to make use of the properties of the Möbius net. He defined two ranges of points (and similarly two two-dimensional or three-dimensional systems) to be projectively related when there is a correspondence between them which is such that harmonic tetrads correspond always to harmonic tetrads. Appealing to considerations of continuity, he then attempted to justify the fundamental theorem that a projectivity between ranges is uniquely determined by three corresponding pairs; but in this one particular he went slightly wrong, since the argument that he used is invalid unless the line can be asserted to be a continuum in the sense of Dedekind. The need for a special postulate to ensure this was pointed out in 1873 by Felix Klein, then in his turn professor at Erlangen.

Von Staudt's methodological principle that since projective geometry is logically independent of notions of spatial magnitude it ought to be developed non-metrically gained general acceptance; and towards the end of the nineteenth century many attempts were made to improve still further on the rigour of von Staudt's treatment by making the presentation strictly axiomatic. An early contribution to this movement was made by Moritz Pasch, in his *Vorlesungen über neuere Geometrie* (1882). Although Pasch believed that geometry ought to be grounded in intuition—and in consequence of this belief he chose to begin his treatment with finite linear segments rather than with straight lines of infinite extent—he considere the same time that the logical basis of the deductive system must be absolutely precise. He accordingly based the development on a numbe *Kernsätze*, in which were expressed the fundamental facts concerning incidence. Other mathematicians (in particular Peano and Hilbert) soon went further still, recognizing that the logical structure of geometry is in fact independent of intuition. They saw, in other words, that only *abstract* primitive notions and primitive propositions are needed, and that geometrical theorems are valid for all systems of entities that satisfy the appropriate axioms.

The formal properties of the relation of incidence constitute the most elementary part of projective geometry; and when incidence geometry is presented axiomatically it is seen to be a common substructure of many different systems of geometry, for instance of euclidean geometry and of the non-euclidean geometries as well as of projective geometry. In particular, metrical geometries can be obtained when a suitable theory of congruence is set up. For this aspect of the foundations of geometry see Hilbert [2] and Whitehead [2, 3].

A very careful investigation of the minimal assumptions that will suffice for the construction of a system of projective geometry of n-dimensional space was made by Gino Fano in a paper 'Sui postulati fondamentali della geometria proiettiva', published in 1891. Fano showed that an explicit existential postulate is necessary in order to ensure that a line always contains more than two points, and that a further postulate is then necessary in order to ensure that points are not always their own harmonic conjugates. He did this by setting up models for which the postulates in question fail to hold, showing, for instance, that a two-dimensional projective geometry is possible in which there are only seven points, lying in threes on seven lines.

A complete axiomatic presentation of projective geometry, worked out with the aid of Peano's logical symbolism, but retaining few of the symbols in the published version, is contained in a paper of Mario Pieri, 'I principii della geometria di posizione composti in sistema logico deduttivo' (1899). Pieri's system is based on two primitive notions (point and the join of two points) and nineteen primitive propositions or postulates. The earlier postulates are concerned with properties of incidence, while the later ones deal with the relations which give rise to an order of the points of a line; and among these later postulates there is one which ensures that the line is a continuous series in the sense of Dedekind. Pieri considers that his set of postulates is complete, since he is able to derive all the principles used by von Staudt, and von Staudt has already derived the further theorems of projective geometry in detail, and has shown moreover that his points can be represented by real coordinates.

The attempts to make projective geometry rigorous that we have mentioned so far have all set out from a totality of 'geometrical' points, either supposed given intuitively or else postulated as primitive; but an analytical approach is also a possibility, and this has on the whole been preferred in recent times. The relation between the two possible approaches may be expressed in the following way. In a synthetic treatment such as that of Pieri, coordinates may be introduced by von Staudt's method. Rational coordinates are first of all determined for all the points of a Möbius net, and then real coordinates are determined by continuity for all other points. When this is done, points become in effect pairs of ratios $x:y:z$ of real numbers. But the same result would have been obtained more directly if a point had initially been *defined* as such a pair of ratios, a straight line as the totality of points which satisfy a given homogeneous equation, and so forth; and this is precisely the analytical approach.

When all geometrical entities are defined algebraically, it appears that there is no theoretical necessity for the coordinates to be *real* numbers, since the algebraic development requires only that they should be drawn from a field. Every field F thus yields corresponding spaces of 2, 3, or n dimensions; and if F happens to be a Galois field $GF(p^s)$, with only a finite number of elements, the corresponding projective spaces have each only a finite number of points. Since there are infinitely many Galois fields, one for each prime-power p^s, there are infinitely many finite projective geometries of given dimensionality. In the simplest case of all, when the ground-field is $GF(2)$, the projective plane has seven points, and projective space of three dimensions has fifteen. The finite geometries were first discussed in relation to Galois fields by Veblen and Bussey [1] in 1906.

2. *Models of non-euclidean geometry.* A brief account of the models of Klein and Poincaré for two-dimensional hyperbolic geometry is given by

Courant and Robbins, [1], pp. 219–224. Poincaré's model is mentioned also by Pedoe, [1], p. 58.

Klein developed the ideas involved in his model in his fundamental paper 'Über die sogenannte Nicht-Euklidische Geometrie' (1871), and they are treated very fully in his book *Vorlesungen über Nicht-Euklidische Geometrie* (1928). Hyperbolic geometry results when, in the real projective plane, a non-virtual conic is taken as absolute conic and only the points of its interior (i.e. those points from which no real tangent can be drawn) are treated as actual points. Angular measure can be introduced by the natural generalization of Laguerre's formula for euclidean geometry, with the proper absolute conic taking the place of the absolute point-pair (I,J); and distance can be defined in terms of cross ratio by the dual procedure. Collineations which leave the absolute conic invariant can then be interpreted as non-euclidean rotations. If a virtual conic is taken instead of a non-virtual one as the absolute conic, an elliptic metric is obtained instead of a hyperbolic metric.

3. *The literature of Hilbert's metamathematics.* Whereas Hilbert's *Grundlagen der Geometrie* has been translated into English (see the Bibliography), the *Grundlagen der Mathematik* of Hilbert and Bernays, which supplies the definitive account of Hilbert's later metamathematical work, at present exists in German only. Much of the substance of this book is contained, however, together with material from other sources, in Kleene's *Introduction to Metamathematics*. Kleene's book is also provided with a full and very useful bibliography.

Hilbert's many-sided contribution to mathematics, and his enormous influence on the creative work of other mathematicians, form the subject of a long commemorative article 'David Hilbert and his mathematical work', contributed by Hermann Weyl to the *Bulletin of the American Mathematical Society*, **50** (1944), 612–654. Weyl gives a clear and concise summary of Hilbert's achievement in each of the many fields in which he worked. There is a shorter notice by Weyl in the *Obituary Notices of Fellows of the Royal Society*, **4** (1942–1944), 547–553, and this also conveys clearly the unique position of Hilbert in the mathematical world of his time.

tremely useful and even indispensable from the didactic point of view in the early teaching of the calculus if too much time is not to be lost. But no one will deny that this way of introducing the calculus can make no claim to being systematic. My feeling of dissatisfaction at that time was so overwhelming that I firmly resolved to go on reflecting until I had found a purely arithmetical and absolutely rigorous basis for the principles of the calculus. It is so often said that the calculus deals with continuous quantities, and yet no explanation of this continuity is ever given, and even the strictest presentations of the calculus do not base their proofs on continuity, but either appeal more or less consciously to geometrical or geometrically inspired ideas or else make the proofs rest on theorems which are never themselves proved arithmetically. One such theorem is the one that I have mentioned, and closer investigation convinced me that this or any equivalent theorem can be taken as a sufficient foundation for infinitesimal analysis. It only remained to discover its origin in the elements of arithmetic, and so at the same time to obtain a real definition of the essence of continuity. This I succeeded in doing on 24 November 1858. . . .'

Dedekind found the solution of the problem that had been troubling him in his new definition of real numbers as sections of the ordered set of rational numbers, and in his theorem that to every section of the ordered set of real numbers there corresponds a real number. This is only one example of the fundamental work that he did in giving strictly conceptual definitions of mathematical ideas that had hitherto been vague or intuitive. Another example is his replacement of Kummer's fictitious ideal factors, in the theory of algebraic number fields, by ideals, in the now familiar sense, defined as certain sets of actual numbers (see Note 2, p. 153).

We shall have occasion to say more about Dedekind in the next chapter (p. 158), when we discuss the logical analysis of mathematical concepts, but here we must turn to another great pioneer in the field of foundations of mathematics, Giuseppe Peano.

5 Peano's 'Formulaire de Mathématique'

Even though Dedekind had a more lively interest in questions of logical analysis than men such as Cauchy and Weierstrass, who were only concerned with logic in so far as it was involved in their purely mathematical work, he was still essentially a mathematician, and not a logician or a philosopher. But a turning point in the whole critical movement was reached with Peano, who, after having begun his career as a critical mathematician of a somewhat similar type to Dedekind, devoted himself specifically to the study of foundations of mathematics for its own sake. For this reason alone, Peano must be looked upon as a pivotal figure in the history of this particular branch of study. His was, moreover, an unusual and engaging personality; and since some knowledge of the man is necessary for the proper understanding of his thought, we give here a few biographical details.

Giuseppe Peano was born in 1858 (the very year of Dedekind's great discovery) and died in 1932, having spent almost the whole of his life at Turin, where for many years he was professor of infinitesimal analysis. He had wide mathematical interests, as well as major interests outside mathematics, and the number of his published papers exceeds two hundred. Although he was a meticulous scholar, who always liked, for example, to trace statements and discoveries to their original sources, he was no pedant; and he is also said to have been a remarkably enthusiastic and inspiring teacher. He never tired of trying to persuade other teachers that they fail in their task unless they are able to make every fact which they have to communicate come to life in the student's mind, so that he sees it as something directly relevant to his own thinking.

Peano's career falls naturally into three phases, since his chief interest was successively in the infinitesimal calculus, in foundations of mathematics, and in linguistic studies; but all three preoccupations were in some measure with him throughout his life.

Like Dedekind, Peano was made aware of the need for a higher standard of rigour in mathematics by his experience of teaching the infinitesimal calculus. His first book[1]—which he published under the name of Angelo Genocchi, his revered former teacher at Turin—was a treatise on the calculus, of the customary form but very much improved in the details of the exposition. In all his mathematical work Peano strove towards a twofold ideal of simplicity and rigour, and his book on the calculus is exemplary in both these respects. Beginning with a strict definition of real number, essentially that of Dedekind, he develops the calculus systematically, formulating every theorem with the greatest possible accuracy and precision, and strictly avoiding in the proofs any illegitimate appeal to intuitive properties of curves. When the customary enunciations of theorems are too loose, or conditions that need to be satisfied are not as a rule clearly stated, Peano often constructs counter-examples to show that assertions made in standard textbooks are incomplete or erroneous.

Both Peano and Dedekind traced much of the still prevalent lack of rigour in analysis to unwarranted reliance upon intuition. The languages of analysis and geometry are to some extent translatable into each other, by reason of the connexion between functions and their graphs, but their equivalence is by no means as complete as has often been supposed. At a comparatively naive level of mathematical thinking (corresponding historically to the state of mathematics before the beginning of the critical movement), where the concepts of function and curve are interpreted in a common-sense way, it is reasonable to regard these concepts as interchangeable. As a result of the critical movement, however, the concepts of number and function are now no longer intuitive, and they have been found to require elaborate logical definition. If, therefore, we reason intuitively about curves—arguing, for example, that if a continuous curve has points on both sides of a straight line then it must 'obviously' have at least one point of intersection with the line—we are not justified in translating our conclusions into statements about functions in the sense in which they are now understood in analysis. Since the concepts of analysis are abstract concepts, having strictly logical definitions, we can only prove the theorems of analysis by arguing exclusively from these definitions. The intuitive properties of graphs may suggest to us new analytical theorems, and they may also make existing theorems more readily intelligible, but they can never be grounds of proof.

---

[1] A. Genocchi: *Calcolo differenziale e principii di calcolo integrale, pubblicato con aggiunte dal D.r Giuseppe Peano* (Turin, 1884).

Peano gave a striking illustration of the way in which intuitively evident conclusions may cease to be valid if the concepts involved in them are idealized theoretically when he devised his famous space-filling curve. He published this discovery in a short paper entitled 'Sur une courbe, qui remplit toute une aire plaine' (*Math. Ann.*, **36** (1890), 157–160), introduced as follows:

> 'In this note we determine two one-valued and continuous functions x and y of a single (real) variable t, which, as t varies in the interval (0,1), take all pairs of values such that $0 \leqslant x \leqslant 1$, $0 \leqslant y \leqslant 1$. If, in accordance with custom, we use the term *continuous curve* for the locus of those points whose coordinates are given continuous functions of one variable, we obtain in this way a curvilinear arc which passes through all the points of a square. Hence, if an arc of a continuous curve is given, about which nothing further is assumed known, it is not always possible to enclose the arc in an arbitrarily small area.'

It is thus clearly demonstrated that idealization of the notion of function makes impossible the retention of the simple intuitive notion of the dimensionality of a set of points. Details of the functions which Peano used in the definition of his curve are to be found in Note 1, p. 152, at the end of the present chapter.

As a result of his work on the properties of functions, Peano arrived at the conclusion that, in the refounding of mathematics, half-measures are useless. It is of no avail to tinker with common sense and intuition; for we are now committed to defining such notions as number and function in a more formal way, and this means that intuitive mathematics and abstract mathematics have entirely different presuppositions. As mathematicians, therefore, we must renounce intuition altogether, and develop mathematics strictly by its own proper means. With the clear recognition of this principle, Peano entered upon the second phase of his career, devoted to the transformation of mathematics into a self-contained formal system.

Peano followed the ancient Greek precedent in presenting his system of mathematics in demonstrative form; but his abandonment of intuition as a sufficient foundation made it necessary for him to modif the Greek method in several essential respects. In the first place, th entities to be considered—numbers, functions, points, etc.—are neithe entities given in intuition nor entities whose properties are dependen on their being idealizations of something given in intuition (althoug they may in fact have originated in that way); and it follows from this that they must all be either postulated or defined. Again, since mathematics is to be logically independent of intuition the axioms of the demonstrative system cannot be justified by appeal to self-evidence. They too must be simply postulated, and they neither require nor admit of any further justification within the formal theory. When once the entities in the system and their initial properties have been postulated, the theorems must all be derived by purely logical processes; and here it is particularly necessary to guard against involuntary appeal to intuition. To this end Peano introduced two safeguards, each of which inaugurated an enduring tradition in the study of mathematics and its foundations. In the first place, he renounced ordinary language and determined to rewrite mathematics entirely in terms of a new symbolism. He had a number of reasons for doing this—the fact that ordinary language is often imprecise, and that its structure does not always match the structure of mathematical demonstration; the desirability of using entirely neutral symbols, so that there is no danger of appealing tacitly to the intuitive associations of common words (as the older analysts were led to do when they inferred properties of functions from those of curves); and the universality of mathematical symbolism, as compared with the diversity of national languages. And secondly, he formalized the logic of mathematical argument as well as the language in which mathematics is expressed, undertaking for this purpose the task of replacing intuitive inference by application of a limited number of stated logical rules. In other words, he constructed and used a symbolic logic, by the aid of which the inferring of a conclusion from premisses was replaced by the derivation of one symbolic expression from other such expressions by quasi-algebraic calculation.

The new form that the theories of pure mathematics were to assume, then, was of the following nature. Each theory would involve a small number of primitive ideas, to be represented by suitably chosen symbols; and these ideas would be left undefined, simply being enumerated at the beginning of the exposition. There would also be a small number of primitive propositions, containing the primitive ideas, which would likewise be enumerated. Development of the theory would be wholly by derivation of symbolically formulated propositions, either from the primitive propositions themselves or from propositions already derived, in accordance with the stated principles of symbolic logic. New entities might be introduced in the course of the development, either by explicit definition as some particular combination of given entities or, in the case of arithmetic, by the special process of 'recursive definition' (see p. 147).

The work of formalization of mathematics involved the rewriting of existing theories in symbolic form, and it also involved a great deal of amplification and reorganization of the deductive sequence in order to make the system conform to the new logical canons, which were much more exacting than any that had been recognized previously. This vast work was carried out by Peano, in collaboration with a number of colleagues and assistants, and the results were published in the *Rivista di Matematica* and the *Formulaire de Mathématique*. The first of these two publications was a journal, founded by Peano in 1891 for the dissemination of the results obtained in his circle, and published until 1906, while the second was a cumulative exposition of the whole system as worked out to date.

The *Formulaire* appeared in five successive editions or volumes. The series began in 1894, with an introductory pamphlet entitled 'Notations de logique mathématique. Introduction au Formulaire de Mathématique', and in 1895 there followed the first full volume of the *Formulaire* itself, in nine sections written by Peano and his collaborators: I. Mathematical logic; II. Algebraic operations; III. Arithmetic; IV. Theory of magnitudes (*grandeurs*) [Burali-Forti]; V. Classes of numbers; VI. Theory of sets (*ensembles*) [Vivanti]; VII. Limits [Bettazzi]; VIII. Series [Giudice]; IX. Contributions to the theory of algebraic numbers [Fano]. The second volume of the *Formulaire* was published in three parts in 1897, 1898, and 1899; and the third, fourth, and fifth volumes appeared respectively in 1901, 1902–1903, and 1908.

The relation between the different volumes is confused, since these are neither complementary volumes of a single work, like the three volumes of *Principia Mathematica*, nor successive editions of one and the same work. Central topics, such as mathematical logic and arithmetic, are treated again and again, now briefly and now at much greater length, but peripheral theories may appear only once, when first formalized. Thus the earlier volumes are not altogether superseded by the later ones, although the different volumes overlap to a considerable extent.

The version of the *Formulaire* that we shall discuss here is that of 1901, which is arranged in the following five parts: I. Mathematical logic; II. Arithmetic; III. Analytic functions; IV. Complex numbers; V. Vectors.

The section on mathematical logic is now of no more than historical interest; for Peano did not develop symbolic logic beyond a comparatively rudimentary stage, and the propositional calculus and the calculus of predicates, as formalized by him, were first worked out in full detail, as we shall describe in Chapter 6, by Russell and Whitehead in *Principia Mathematica* (1910). Nevertheless, the credit for introducing the essential notions and symbols belongs to Peano, as Russell himself has freely acknowledged. Peano used letters to denote propositions and propositional functions, devised notations[1]—of which some are still in use, while others have now been supplanted—for the logical constants 'not', 'implies', 'implies for all x', etc., and pointed out the fundamental significance of the relation of class-membership, reserving for it the initial letter of the Greek word ἐστί (*is*).

Apart from the section devoted to logic, the most important of the five divisions of the *Formulaire* of 1901 is undoubtedly the second one, on arithmetic. To this formulation of the basic properties of numbers (which owes much to Dedekind's essay *Was sind und was sollen die Zahlen?*, to be discussed in §2 of the next chapter) almost all subsequent work on the foundations of arithmetic can ultimately be traced back. Peano bases his treatment on three primitive ideas:

(i) N_0, the class of natural numbers;
(ii) 0, the particular number zero;
(iii) $a+$, the successor of the number a.

There are six primitive propositions, which we may write in a slightly modernized form as follows:

(i) $\mathrm{N}_0 \in \mathrm{Cls}$;
(ii) $0 \in \mathrm{N}_0$;
(iii) $a \in \mathrm{N}_0 \rightarrow a+ \in \mathrm{N}_0$;
(iv) $[s \in \mathrm{Cls}\ \&\ 0 \in s\ \&\ (x)(x \in s \rightarrow x+ \in s)] \rightarrow \mathrm{N}_0 \subset s$;
(v) $(a,b \in \mathrm{N}_0\ \&\ a+ = b+) \rightarrow a = b$;
(vi) $a \in \mathrm{N}_0 \rightarrow a+ \neq 0$.

With the exception of (i), which belongs more to Peano's way of presenting arithmetic than to arithmetic itself, these primitive propositions are simply the well-known 'Peano axioms' for the natural numbers.[1] They may be expressed in familiar language as follows: (i) N_0 is a class; (ii) 0 is a number; (iii) The successor of any number is a number; (iv) If a class s is such that (a) it contains 0, and (b) if it contains any number a then it also contains the successor $a+$ of that number, then s includes the whole of N_0; (v) No two numbers have the same successor; (vi) 0 is not the successor of any number.

---

[1] Peano, and Russell likewise, wrote 'p implies q' as '$p \supset q$'. Peano also wrote '$p\,.\!\supset_x.\,q$' to mean 'for all x, p implies q', where p and q are propositional functions involving the variable x; and Russell continued to use this notation, in the more explicit form '$\phi x \supset_x \psi x$', as an alternative to '$(x)\,.\phi x \supset \psi x$'. Hilbert subsequently replaced '$\supset$' by '$\rightarrow$', at the same time discontinuing the use of subscripts to indicate universal quantification. For 'x is a member of a' Peano wrote '$x \varepsilon a$'. and for 'the class of entities x which satisfy the propositional function p' he wrote '$x \ni p$'. Russell changed this last usage to '$\hat{x}(\phi x)$'.

The fourth of the primitive propositions is the axiom of induction, whose function in the system is to formalize the demonstrative procedure known as *mathematical induction*. This is the mode of argument that we commonly use in demonstrating a theorem $\mathfrak{A}(n)$, the statement of which involves an undetermined integer n as a parameter. We prove (a) that $\mathfrak{A}(0)$ holds, and (b) that if $\mathfrak{A}(n)$ holds then so also does $\mathfrak{A}(n+)$; and we then conclude that $\mathfrak{A}(n)$ is valid for every value of n. Proof by mathematical induction is fundamental to Peano's whole system, and indeed it is the only specifically arithmetical form of argument that is provided by the primitive propositions. In the working out of the system, Peano also makes essential use of the constructive counterpart of this demonstrative procedure, namely *recursive definition*. An arithmetical function $\mathfrak{f}(n)$ is said to be defined recursively if (a) an explicit definition of $\mathfrak{f}(0)$ is given, and (b) $\mathfrak{f}(n+)$ is defined in terms of $\mathfrak{f}(n)$ by means of functions that are already available. Interpreted intuitively, a recursive definition of $\mathfrak{f}(n)$ plainly provides a means of determining in succession the values $\mathfrak{f}(0)$, $\mathfrak{f}(1)$, $\mathfrak{f}(2), \ldots$ of the function. Peano uses the method in order to define the arithmetical sum and product, $a + b$ and $a \times b$, doing this in each case by recursion with respect to b:

$$a + 0 = a, \qquad a + (b+) = (a + b)+,$$

and

$$a \times 0 = 0, \qquad a \times (b+) = (a \times b) + a.$$

The familiar associative, commutative, and distributive laws can now be established in the formal system, and much of the elementary theory of numbers is then readily derivable by formal means.

Having proceeded thus far with the systematic construction of his system of arithmetic, Peano goes on to indicate much more briefly how rational numbers and real numbers can be obtained formally from the natural numbers; and he then turns, in the next two sections of the *Formulaire*, to a consideration of various topics in elementary analysis.

---

[1] In the literature, the 'natural numbers' are taken sometimes as 0, 1, 2, ... and sometimes as 1, 2, 3, ... Peano began the series with 0, and Hilbert adopted the same convention. Dedekind, on the contrary, took 1 as the first natural number, and this is now the normal practice in mathematics. The choice between the two possible conventions is of no theoretical importance, and we shall ordinarily conform to the usage of the author whose work we are discussing.

The concluding section of the work, which has the heading 'Vectors', is devoted to geometry.

Geometry was indeed yet another subject to which Peano gave a considerable amount of attention, and on which he even wrote a separate book[1]; but although his work on geometry is by no means without interest, it is of much less significance in the history of foundations of mathematics than his contribution to arithmetic. It arose out of the dissatisfaction which he, in common with certain other nineteenth-century mathematicians, felt with the way in which modern geometry had developed. The use of cartesian coordinates had made geometry less elementary than it was in classical times, and also, by rendering it dependent on processes of calculation, had reduced to a minimum its intuitive content. But the cartesian method relies on the introduction of a set of arbitrarily chosen axes, having no connexion with the figure under investigation; and during the nineteenth century several attempts were made to perfect a symbolic calculus which would apply more directly to geometrical entities, without intervention of an extraneous frame of reference. Möbius with his barycentric calculus (1827), Hamilton with his quaternions (1843), and Grassmann with his *Ausdehnungslehre* or calculus of extension (1844 and 1862) had all made contributions to this movement; and prompted by their example, Peano gave a new formal treatment of certain parts of geometry by means of a calculus of vectors, similar to the system now widely used in theoretical physics.

The adoption of vectorial ideas in physics and applied mathematics is not, however, due directly to Peano's influence, since the use of vectors was developed independently of him by Josiah Willard Gibbs (1839–1903) and Oliver Heaviside (1850–1925); and already in 1881 Willard Gibbs had some notes on vectors printed for circulation among his students. And even in the field of geometry, Peano's vectorial treatment has perhaps become something of a historical curiosity,[2] for geometry has in the meantime developed along quite other lines, becoming less and less dependent on geometrical intuition of any kind. Differential geometry is now handled, in fact, by the calculus of tensors, and algebraic geometry by means of a coordinate representation over a postulated ground-field, which makes possible the application to it of the no less powerful methods of modern algebra and topology.

In his later years Peano turned away from mathematics and its foundations, and devoted almost all of his time and energy to the task of winning support for a new international auxiliary language, Interlingua, that he had invented. His interest in linguistic studies had already been awakened much earlier in his career, and this part of his activity, seemingly remote from mathematics, made a far from negligible contribution to the success of his mathematical and logical researches.

---

[1] G. Peano: *Calcolo Geometrico, secondo l'Ausdehnungslehre di H. Grassmann, preceduto dalle operazioni della logica deduttiva* (Turin, 1888).

[2] This is not to say that it must always remain so. New ideas in mathematics are not always significant in the way that their originators intend, and history shows that an apparently barren innovation may become abundantly fruitful when a later age provides it with fresh soil.

Peano had expert knowledge of classical languages, as well as a lively interest in comparative philology. At one time he made a number of attempts to reduce the grammatical structure of various languages to essentially mathematical form, in order to see if he could find any common structure belonging to all languages, and so, perhaps, inherent in language as such. He became convinced in the end that no such structure exists, and that both grammar and idiom are wholly dependent on the fortuitous circumstances of history. As time went on, he became more and more concerned with communication in its various aspects. The problems presented by communication of abstract or mathematical knowledge were manageable, and Peano solved them so successfully that his symbolic methods led directly to the creation of modern mathematical logic. Communication of more concrete experience—thoughts, impressions, emotions, etc.—is very much more difficult, since the substance of what is to be communicated can exhibit infinite variety, and so linguistic representation of elements in experience can only be approximate. In many of its functions, language is evocative rather than representational; but even in so far as it is representational its effective use is much impeded by the existence of a multiplicity of different national forms. In his mathematical writing, addressed to mathematicians at large, Peano changed over early on from his own language of Italian to French, which is more widely known; but even French is not fully international, and he looked back with regret to the time, not so very distant, when Latin had been the universal language of scholars.[1] In his day, unfortunately, comparatively few people could write Latin, and the difficulty of the language was a bar to its revival; but the idea came to him that most of the difficulty could be avoided by the simple expedient of dispensing with the complexity of Latin grammar (which is in any case historically fortuitous, and in no way essential to the language as a medium of communication). Peano thus arrived at his auxiliary international language of *latino sine flexione*, afterwards named *Interlingua*. The vocabulary is that which is to be found in any Latin dictionary, and there is virtually no grammar. Nouns are undeclined and verbs unconjugated, every noun taking the invariable form of the Latin ablative singular and every verb that of the singular imperative. Case is indicated, as in many modern languages, by means of prepositions, and tense by means of adverbs. In his later mathematical publications, Peano consistently used Interlingua. The *Rivista di Matematica* was then called *Revista de Mathematica*, and when the final version of the *Formulaire* appeared in 1908 it was as *Formulario Mathematico*.

---

[1] In 1889 there had in fact appeared an essay of twenty-nine pages, written in Latin, with the title *Arithmetices Principia nova methodo exposita, a Ioseph Peano.*

It emerges clearly, even from the present superficial account of his work, that Peano was a man of quite remarkable originality—original not merely in his ability to carry out mathematical research of the customary kind, but even more in his capacity for detaching himself from prevailing traditions and looking at them critically and dispassionately. He was able to confront in this way some of the most widely accepted presuppositions of his age, and to ask himself whether they really corresponded to the needs of the actual situation. He considered afresh, as we have already described, the whole question of communication between one mind and another, and he was led by his inquiries to advocate radical reforms in the two very different spheres of mathematics and linguistics. His mathematical innovations were to exercise a profound influence on the thought of mathematicians and philosophers of mathematics; but his long championship of Interlingua as an auxiliary language met with no lasting response. In mathematics, he recognized that there is an unbridgeable gulf between the intuitive outlook of common sense and the logical rigour that, by his day, had come to be looked upon as essential in the construction of mathematical theories. He came to the conclusion that mathematics must be constructed, independently of intuition, in a way that guarantees absolutely the validity of its theorems; and in order to satisfy this requirement he set to work to rewrite mathematics in symbolic form as an axiomatic system, based exclusively on postulated primitive notions and primitive propositions. This was an epoch-making step to take; and we shall see how heavily later writers on the foundations of mathematics have been indebted to Peano for his initiative.

6 The symbolic language of Leibniz

Although Peano was a man of outstanding ability he also had, to a most unusual degree, the quality of humility. This was so strong an element in his personality that he could never bring himself to accept the credit for his own achievements.[1] He would certainly never have admitted to being an original thinker, and it was his nature to see his own work always as nothing but a continuation or fulfilment of someone else's—as for instance of Genocchi's when he wrote his book on the infinitesimal calculus. In the case of symbolic logic, it was Leibniz who received the credit. Peano had especial reverence for Leibniz, which is partly accounted for by the fact that he was probably the first man ever fully to appreciate some of the things that Leibniz did or conceived the idea of doing. In the preface to the *Formulaire* of 1901 we find these words:

---

[1] It is wholly characteristic of Peano that he expressed the wish that when he died, honoured though he was as a man of acknowledged eminence in the world of learning, he should be buried in the part of the cemetery that was reserved for the paupers of the city of Turin. And in the brief report of Peano's funeral that is printed in *Schola et Vita*, **7** (1932), 103, we read that indeed, after a ceremony in the great hall of the university, 'funere move ad cæmeterio, ubi nostro caro et venerato prof. Peano es sepulto in campo commune, 3° sud, in fossa n. 674, inter pauperes'.

'Mathematical logic studies the properties of logical operations and relations, which it denotes by symbols. A few of the principles of this science are to be found in general logic (see Aristotle). Its true founder is Leibniz, who enunciated the principal properties of the ideas now represented by $\cap$, $\cup$, $-$ [not], $=$, $\supset$, Λ.'

Leibniz, as is generally acknowledged, may be compared to Aristotle in the range of his abilities. He had prodigiously wide interests and made highly original contributions to many different branches of learning; and among his numerous projects was the construction of a symbolic logical language. But the men of his time did not take him seriously, and he himself never found the leisure needed to carry through the immense task of working out a suitable symbolism in detail, starting from the very beginning. Although, however, his project remained largely a dream, he worked at it intermittently and obtained a number of fragmentary results. These were preserved among his manuscripts in the royal library at Hanover, and some of them were at length published in selections from his works; but once again their significance was not appreciated. Eventually, when Peano had been led by his own researches into the foundations of mathematics to reflect independently on the same matters, he realized for the first time the meaning of the results noted down by his great predecessor. And so at last, just before the close of the nineteenth century, Peano's colleague Giovanni Vacca went to Hanover to make an examination of the manuscripts there. The extracts which Vacca published in the *Formulaire* of 1899 inspired Louis Couturat to go more fully into the matter; and in 1901 Couturat published *La Logique de Leibniz*, which was followed in 1903 by a large volume containing the logical fragments themselves.

Now that Peano had actually produced a logical calculus, and had thus demonstrated that the project was not illusory, the true significance and importance of Leibniz's work could at last be appreciated. That Leibniz himself had seen clearly, all the time, what he was about, is proved by some words that he wrote in an often-quoted letter of 1714:

'If I had been less distracted, or if I had been younger or assisted by young men in a position to help me, I would have hoped to give a kind of *spécieuse générale*,[1] in which all truths of reason would be reduced to a calculus. This would be at the same time a sort of universal speech or writing, but infinitely different from all those proposed hitherto; for the characters, and the words themselves, would be governed by Reason; and errors, with the exception of those of fact, could only be mistakes in calculation. It would be very difficult to create or invent this language or characteristic, but very easy to learn it without the aid of a dictionary.'

The symbolic logic of Peano, though limited to a narrow range of discourse (namely the exposition of demonstrative theories) is just such an instrument as Leibniz had in mind.

SUPPLEMENTARY NOTES ON CHAPTER 5

1. *Peano's space-filling curve.* Peano's procedure for defining a curve which passes through every point of the square $0 \leqslant x \leqslant 1$, $0 \leqslant y \leqslant 1$ (cf. p. 143) may be summarized as follows. Let t be any real number in the range $0 \leqslant t \leqslant 1$, expressed in the scale of 3 in the form

$$a_1/3 + a_2/3^2 + \ldots + a_n/3^n + \ldots.$$

Denoting by $\mathbf{k}$ an operator which converts 0, 1, or 2 into its complement with respect to 2, so that $\mathbf{k}0 = 2$. $\mathbf{k}1 = 1$, $\mathbf{k}2 = 0$, we define two sequences $\{b_n\}$ and $\{c_n\}$ in the following way:

$$\begin{aligned} b_1 &= a_1, \\ b_2 &= \mathbf{k}^{a_2} a_3, \\ &\cdot \quad \cdot \quad \cdot \quad \cdot \\ b_n &= \mathbf{k}^{a_2 + a_4 + \ldots + a_{2n-2}} a_{2n-1}; \\ &\cdot \quad \cdot \quad \cdot \quad \cdot \end{aligned}$$

and

$$\begin{aligned} c_1 &= \mathbf{k}^{a_1} a_2, \\ c_2 &= \mathbf{k}^{a_1 + a_3} a_4, \\ &\cdot \quad \cdot \quad \cdot \quad \cdot \\ c_n &= \mathbf{k}^{a_1 + a_3 + \ldots + a_{2n-1}} a_{2n}, \\ &\cdot \quad \cdot \quad \cdot \quad \cdot \end{aligned}$$

If, now, we put

$$X = b_1/3 + b_2/3^2 + \ldots, \qquad Y = c_1/3 + c_2/3^2 + \ldots$$

X and Y are continuous functions of t; and the parametric representation $x = X$, $y = Y$ then defines a curve with the required property. A number of details still have to be taken care of, but even when this is done Peano's entire paper occupies only four pages.

---

[1] Literally, 'generalized algebra'. The name *arithmétique spécieuse* was formerly used for algebra; cf. the article *Spécieuse* in the *Encyclopédie* of Diderot and d'Alembert (Paris, 1751–1765).

2. *Ideal numbers and ideals.* Kummer introduced ideal numbers into arithmetic with the aim of simplifying the discussion of algebraic number fields

By an *algebraic number* is meant a (complex) number which satisfies an algebraic equation

$$x^n + a_1x^{n-1} + \ldots + a_n = 0,$$

in which the coefficients $a_1, \ldots, a_n$ are rational numbers, i.e. elements of the rational field Q. If, more specifically, these coefficients happen to be rational integers, the algebraic number is said to be an *algebraic integer*. Now the set of all algebraic integers is a ring, and the set of all algebraic numbers is the field of quotients of this ring; but the arithmetical analogy with the rational integers and the rational numbers does not go very far, since there are no algebraic integers that are irreducible, i.e. prime.

The situation becomes much more like that in ordinary arithmetic if, instead of the field of *all* algebraic numbers, some restricted field $Q(\theta)$ is taken, that is to say the smallest field, containing Q, which has some specified algebraic number θ as one of its elements. The algebraic integers which belong to $Q(\theta)$ then form a ring, and $Q(\theta)$ is the field of quotients of this ring.

For some choices of θ, the arithmetic of $Q(\theta)$ is very similar to that of Q. In the case $\theta = \sqrt{3}$, for instance, we have a direct analogue of the theorem that every rational integer is expressible in one and essentially only one way as a product of prime factors; but in other cases there is no such analogy. Thus the number 21 has two distinct resolutions into factors which are irreducible in the field $Q(\sqrt{-5})$, namely 3.7 and $(1 + 2\sqrt{-5})(1 - 2\sqrt{-5})$. This is possible because, although the algebraic integers 3, 7, $1 + 2\sqrt{-5}$, $1 - 2\sqrt{-5}$ are all irreducible in the field $Q(\sqrt{-5})$, they have as common factors certain algebraic integers which do not belong to this field. We have, in fact,

$$1 + 2\sqrt{-5} = \sqrt{\lambda}\sqrt{-\varkappa'}, \quad 1 - 2\sqrt{-5} = \sqrt{\lambda'}\sqrt{-\varkappa},$$
$$3 = \sqrt{\lambda}\sqrt{\lambda'}, \quad 7 = \sqrt{\varkappa}\sqrt{\varkappa'},$$

where

$$\varkappa = 2 + 3\sqrt{-5}, \quad \varkappa' = 2 - 3\sqrt{-5}; \quad \lambda = 2 + \sqrt{-5}, \quad \lambda' = 2 - \sqrt{-5}.$$

The arithmetic of $Q(\sqrt{-5})$ is thus anomalous in that common factors, which have an effect on the multiplicative properties of the integers of the field, are sometimes missing from the field; and Kummer's idea was to postulate 'ideal' numbers of the field in order to fill any gaps that may exist. The field that he originally considered was not in fact $Q(\sqrt{-5})$ but a cyclotomic field, generated from Q by adjunction to it of a complex pth root of unity. In his own words:

> 'Es ist mir gelungen, die Theorie derjenigen complexen Zahlen, welche aus höheren Wurzeln der Einheit gebildet sind und welche bekanntlich in der Kreistheilung, in der Lehre von den Potenzresten und den Formen höherer Grade eine wichtige Rolle spielen, zu vervollständigen und zu vereinfachen; und zwar durch Einführung einer eigenthümlichen Art imaginärer Divisoren, welche ich *ideale complexe Zahlen* nenne; worüber eine kurze Mittheilung zu machen ich mir erlaube.

Wenn α eine imaginäre Wurzel der Gleichung $\alpha^\lambda = 1$, λ eine Primzahl ist und a, a_1, a_2, etc. ganze Zahlen sind, so ist $f(\alpha) = a + a_1\alpha + a_2\alpha^2 + \ldots a_{\lambda-1}\alpha^{\lambda-1}$ eine complexe ganze Zahl. Eine solche complexe Zahl kann entweder in Factoren derselben Art zerlegt werden; oder auch nicht. Im ersten Fall ist sie eine zusammengesetzte Zahl: im andern Fall ist sie bisher eine complexe Primzahl genannt worden. Ich habe nun aber bemerkt, daß, wenn auch $f(\alpha)$ auf keine Weise in complexe Factoren zerlegt werden kann, sie deshalb noch nicht die wahre Natur einer complexen Primzahl hat, weil sie schon gewöhnlich der ersten und wichtigsten Eigenschaft der Primzahlen ermangelt: nämlich, daß das Product zweier Primzahlen durch keine von ihnen verschiedene Primzahl theilbar ist. Es haben vielmehr solche Zahlen $f(\alpha)$, wenn gleich sie nicht in complexe Factoren zerlegbar sind, dennoch die Natur der zusammengesetzten Zahlen; die Factoren aber sind alsdann nicht wirkliche, sondern *ideale complexe Zahlen*.' ('Zur Theorie der complexen Zahlen', *J. reine angew. Math.*, **35** (1847), 319–326.)

The further step taken by Dedekind was the recognition that, in order to handle the arithmetic of a given algebraic field $Q(\theta)$, we do not need to be able actually to produce irreducible factors, since we can do all that is required by arguing in terms of those totalities of integers of the field that behave as if they were divisible by some common irreducible factor. These totalities are what are now known as *ideals*. The procedure here adopted by Dedekind may be compared with his own identification of real numbers with totalities of rational numbers which behave as if separated off by boundary numbers, and also with Frege's identification of the cardinal number of a class with the totality of all classes which behave as if possessing just as many members as that class.

For the use of ideals in the study of algebraic number fields see E. Hecke: *Vorlesungen über die Theorie der algebraischen Zahlen*, Chapter 5.

3. *Sources of historical information.* (a) *General.* The standard history of Greek mathematics is Heath [3].

A very comprehensive survey of the history of mathematics up to 1799, in four volumes, is M. Cantor [1]. For the subsequent development of modern mathematics, in the course of the nineteenth century, see Klein [2].

A collection of essays on the greatest mathematicians from the fifth century B.C. onwards, which is perceptive and informative beneath a somewhat racy style, is E. T. Bell: *Men of Mathematics*. The same author has also written a less popular work [2], in which the emphasis is more on mathematical discoveries than on biographical anecdotes.

Biographical details of modern mathematicians, as well as information concerning their work, are often most readily obtained from the obituary notices published by learned societies. See, in particular, the series *Obituary Notices of Fellows of the Royal Society* (1932–1955) and its continuation since 1956 as *Biographical Memoirs of Fellows of the Royal Society*, and also the obituary notices in the *Journal of the London Mathematical Society*.

(b) *For Peano.* There are a number of articles which give details of Peano's life and work. His contributions to mathematics are outlined in Levi [1] and his contributions to logic in Cassina [3]. Cassina's more elaborate notice [2] covers the whole of Peano's mathematical and logical work, and includes a full list of his publications. Yet another account of Peano's life and work, Cassina [1], is essentially the same as an article which Cassina wrote in 1928, with Peano's assistance, to mark Peano's seventieth birthday. For Peano's own writings, see the three volumes of *Opere scelte*, published 1957–1959.

+,-.+ + +/ ! + '
7 ! < <! "< 3 < ! 2"

" +,-.+ + +/ ! + '
,9,4& & < ,7= ,7& /< ,4 >4/<&<= 2 !<- 2"

+,-.+ + +/ ! + '
7 < <! "<? < 3 <, 3 ! < <! "<@A2 2"

+,-.+ + +/ ! + ' $ *%$ 2"

+,-.+ + +/ ! + ') " 2"
(*

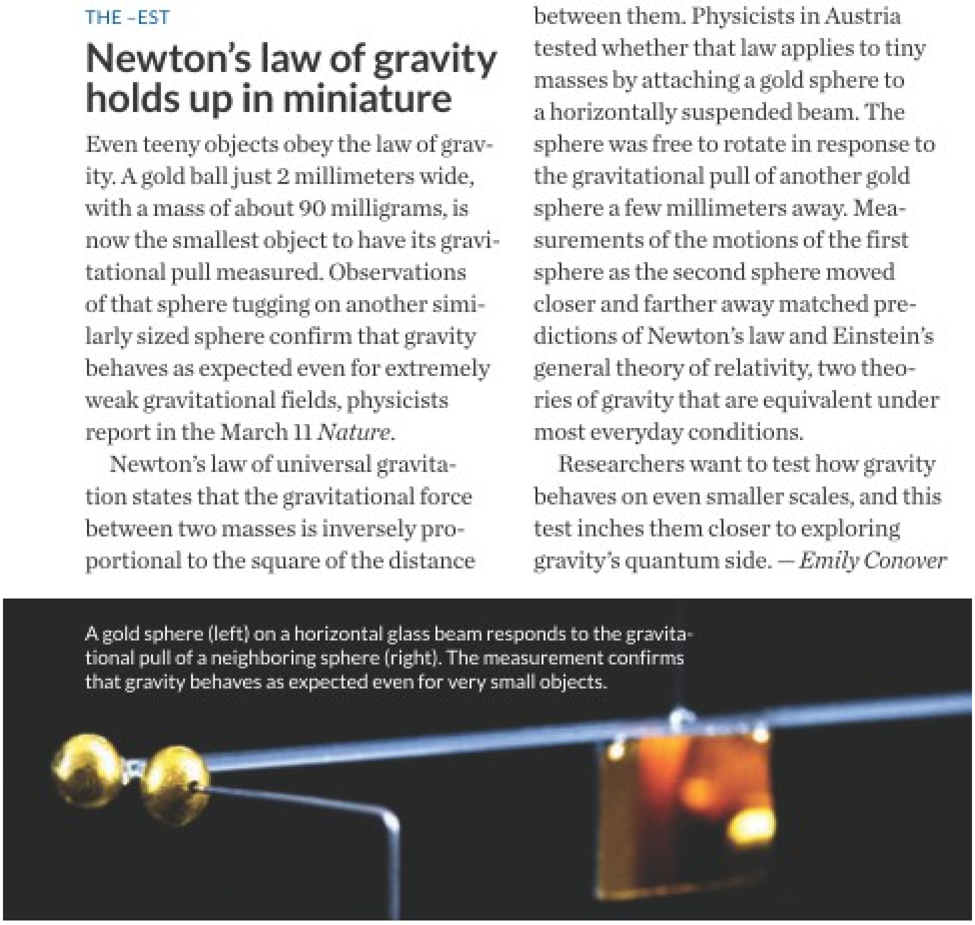

THE –EST

Newton's law of gravity holds up in miniature

Even teeny objects obey the law of gravity. A gold ball just 2 millimeters wide, with a mass of about 90 milligrams, is now the smallest object to have its gravitational pull measured. Observations of that sphere tugging on another similarly sized sphere confirm that gravity behaves as expected even for extremely weak gravitational fields, physicists report in the March 11 *Nature*.

Newton's law of universal gravitation states that the gravitational force between two masses is inversely proportional to the square of the distance between them. Physicists in Austria tested whether that law applies to tiny masses by attaching a gold sphere to a horizontally suspended beam. The sphere was free to rotate in response to the gravitational pull of another gold sphere a few millimeters away. Measurements of the motions of the first sphere as the second sphere moved closer and farther away matched predictions of Newton's law and Einstein's general theory of relativity, two theories of gravity that are equivalent under most everyday conditions.

Researchers want to test how gravity behaves on even smaller scales, and this test inches them closer to exploring gravity's quantum side. — *Emily Conover*

A gold sphere (left) on a horizontal glass beam responds to the gravitational pull of a neighboring sphere (right). The measurement confirms that gravity behaves as expected even for very small objects.

+,-.+ + +/ ! + ' $%%))' 2 3

+,-.+ + +/ ! + '
& " < <4! <, 3 !<B <,22 ! < <,&< ! <- <& ! <, ! A
2"

0 +,-.+ + +/ ! + ' 6 ! ! ?! 7 > 1
2 !C 9 3 < @ " ! - 1 ! 9! D E 2 ?!3
?!3 $ 2 3

C +,-.+ + +/ ! + ' $)'%%)' 2 3
C +,-.+ + +/ ! + ' $%%))' 2 3

data

Books-016-2021/20210808/

→ Math/Analisis_Matematico_Calculo_I_Espinosa.pdf

→ The center and Focus Problem.../
The_Center_and_Focus_Problem....pdf

→ Heat and Mass Transfer Ed 2/
9811539871.pdf

→ Science News - 31 July 2021/
sciencenews20210410f-dl.pdf

Pág. 5 (7) - Newton's law of gravity
hold up in miniature.

→ A Geometrical Framework for
Understanding the Universe/
1793304467.epub

→ Introduction to Lattice Algebra.../
Introduction to Lattice Algebra...pdf

→ The Universe Speaks in Numbers/
Graham Farmelo - The Universe
Speaks in Numbers.epub

→ Towards the Holy Grail of Physics/
1794633146.epub
usar junto com A Geometrical Framework
for Understanding the Universe.

→ Modern Classical physics optics, fluids.../
Modern Classical physics optics...pdf

Jandaia

(%
! +,-.+ + +/ ! + , ! A ! 6 A E ! " %E
E?E 1 2"

3 +,-.+ + +/ ! + , 6! !A " ! !
,"1 & ! & 4 ! ! ! 7 ! 3A ! ! > 1
> 1 2"

3 +,-.+ + +/ ! + , 6! !A " ! !
,"1 & ! && 3A ! ! -! ! 0! ! ! ! A > 1 2"
/3
(!2 %

Chapter 13

Geometric Iterations

13.1 Introduction

In this chapter the focus of study, which till now has been confined to arithmetic and algebra, shifts to geometry. We start by giving a few illustrations that show how iterations arise in geometry.

Reflections. Let f denote reflection in a line ℓ (so that ℓ acts like a mirror). Then $f \circ f = I$, because two successive reflections in the same line take us right back to the starting point; here I denotes the identity operation, which in geometry means "*do nothing, let the point stay where it is*". So reflection in a line is a 2-cycle function.

Half-turns. Let H denote a half-turn (a rotation through 180°) about a fixed point O; then $H \circ H = I$ (for two successive half-turns about the same point are equivalent to a 360° turn and therefore to the identity map I). So a half-turn too is a 2-cycle function.

On the other hand, two successive half-turns about two distinct points, P and Q, are equivalent to a displacement through the vector $2\overrightarrow{PQ}$.

More generally, if θ is any angle ($\theta \neq 0^\circ, \theta \neq 360^\circ$) then the effect of two successive rotations f,g, with f centered at P and through an angle θ, and g centered at Q and through an angle $-\theta$, is a displacement. This fact comes of use when we have to move heavy furniture!

Quarter-turns. Let Q denote a quarter-turn about a fixed, point O; then $Q \circ Q = H$ and $Q(4) = H \circ H = I$. Here we have an example of a 4-cycle function.

An identity involving half-turns. For any quadrilateral ABCD, let P,Q,R,S be the midpoints of the sides AB,BC,CD,DA respectivey. Then we have the curious relation

$$HP \circ HQ \circ HR \circ HS = I.$$

i.e., $HP \circ HQ \circ HR \circ HS(X) = X$ for all points X.

Frieze patterns. Here is an illustration from art. Let a design $\mathcal{B}$ (a "motif") be selected, and let the sequence $\langle \mathcal{B};f \rangle$ be sketched for some appropriate function f. This simple action can result in rich geometrical patterns, depending on the choice of $\mathcal{B}$ and f. Many of the famous artifacts of Islamic art (e.g., those found in the Alhambra fortress in Granada, Spain, or in parts of the Islamic world) can be studied and appreciated from this perspective.

We give an example of such a sequence. Choose the motif to be the clubs symbol, ♣, and let f be, '*move the motif one unit to the right*'. Then the sequence ⟨♣;f⟩ is an infinite strip:

···♣♣♣♣♣♣♣♣♣···

Designs of this type are known as frieze patterns, and they are often seen on temple walls, saree borders, pullovers, and so on.

Wall-paper patterns and crystals. The idea of a repeating one-dimensional pattern can easily be extended to two dimensions, resulting in wall-paper patterns. The three-dimension version of this is obviously a crystal; here a structure gets repeated in some definite manner in three dimensions. Other objects not typically referred to as crystals may possess such repetitive structures; for example, the DNA double helix, and in general other spiral forms that occur in nature.

We now study a few iterations involving geometric themes; *en route* we shall discover some pretty mathematics!

13.2 Perpendiculars

Let △ABC be acute-angled, and let P0 be a point on BC. We locate another point P1 on BC (i.e., on the same side) by dropping perpendiculars as follows: P0Q, from P0 to *AB*; QR, from *Q* to CA; and then RP1, from R back to BC. The position of P1 clearly depends on that of P0, so we write P1 = f(P0), where f is a function that 'acts' on the points on line BC, permuting them amongst themselves (see Figure 13.1). We regard the movement from P0 to P1 as the first round of an iteration.

We now repeat this sequence of actions: we drop perpendiculars in turn to AB, to CA, to BC,…, and obtain the points

$$P2 = f(P1), P3 = f(P2), P4 = f(P3), \ldots.$$

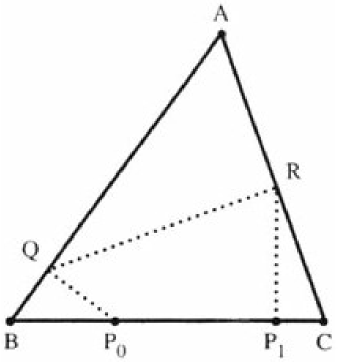

Figure 13.1. *After 1 round*

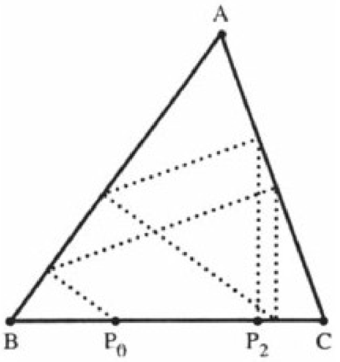

Figure 13.2. *After 2 rounds*

with the points P2,P3,P4,… all lying on BC. Figure 13.2 shows the path taken during the first two rounds of the iteration.

After a few.iterations, we notice something curious—*the path starts* to *close in on itself.* More precisely, the path converges to a triangular path inscribed in $\triangle ABC$. Note that this means that the sequence of points ⟨P0,P1,P2,P3,…⟩ tends to some limiting position on BC. Figure 13.3 displays the limiting path as an inscribed triangle (shown in a thick line). Observe that the inscribed triangle is similar to the original triangle (its sides are perpendicular to the sides of the original triangle, so the angles of the two triangles have the same measures; this explains the similarity).

It is natural to ask what happens when we change the position of P. Experimentation reveals that we reach the very same inscribed triangle each time! Figure 13.4 shows the path followed for another choice of P.

You may wonder what happens if the foot of the perpendicular lies not on the side on which it is dropped, but on its extension. The answer is that it does not matter. In Figure 13.5 the choice of figure is such that for the first few iterations, the feet of the perpendiculars lie on the extensions of the sides. After a few iterations, however, the situation "corrects" itself: the feet of the perpendiculars once again lie upon the sides, and we reach the very same inscribed triangle.

* * * * * * * * * * * * * * * * * * * *

Till now our iteration has followed a clockwise direction: we dropped a perpendicular first to AB, then to CA, then to BC,.... But we could also opt for an anticlockwise direction, dropping a perpendicular first to CA, then to AB, then to BC,.... What happens if we do it this way?

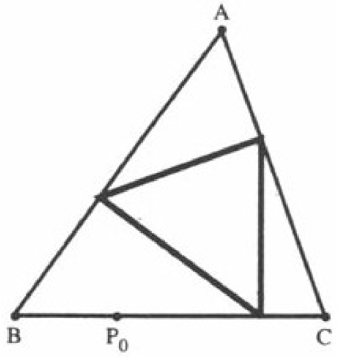

Figure 13.3. *Limiting path–1*

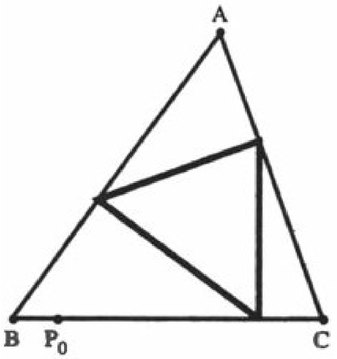

Figure 13.4. *Limiting path–2*

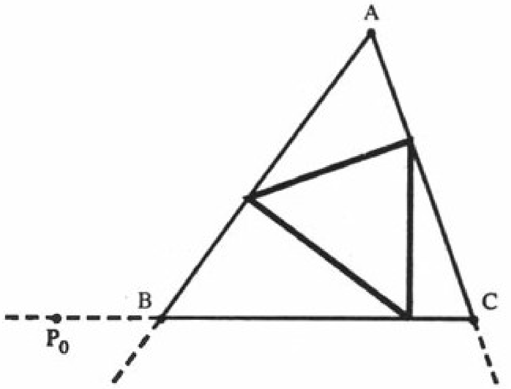

Figure 13.5. *Starting from a "wrong point"*

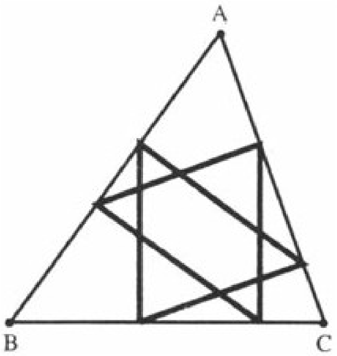

Figure 13.6.

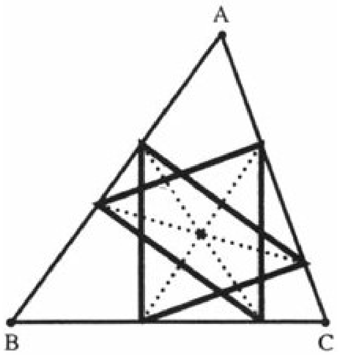

Figure 13.7.

Not surprisingly, we reach an inscribed triangle yet again; but, most pleasingly, the triangle turns out to be congruent to (and yet distinct from) the one got earlier! (See Figure 13.6.)

A surprising connection

An examination of the figure reveals that corresponding sides of these two triangles are parallel to one another (this owes to the perpendicularity relations noted earlier). Pairs of such objects are said to be *homothetic* to one another. If we draw the joins of corresponding pairs of points belonging to homothetic figures, the joins meet in a point called the *homothetic center* of the two figures. For example, an enlargement mapping yields pairs of homothetic figures, and the center of homothety is simply the center of the enlargement. A half-turn also yields pairs of homothetic figures, and the center of homothety in this case is the center or "pivot" of the half-turn ; it bisects the joins of corresponding pairs of points. (Comment: A half-turn is the same as an enlargement with scale factor equal to –1. Note that a half-turn is the only homothety in which the input figure is congruent to the output figure.)

In the current example the two figures are congruent to one another, so the homothety is a half-turn . The homothetic center turn s out to be a "known" point— it is the *symmedian point* of the triangle. The point is constructively defined as follows: we locate G, the centroid of the triangle, draw the medians GA,GB and. GC, then reflect them, respectively, in the lines internally bisecting the angles at A,B and C. The reflected segments are known as *symmedians*, and they meet in the symmedian point of the triangle.

Another way of finding the same point is as follows: draw the circumcircle of triangle ABC, then draw tangents to the circle at A,B and C. Let the triangle enclosed by the tangents be XY Z, with the vertices labelled so that X is opposite A (i.e., the tangents at B and C meet at X), Y is opposite B and Z is opposite C; then the lines AX, *B*Y and CZ meet in the symmedian point.

Exercises

13.2.1 Prove that the inscribed triangle referred to above is similar to the original triangle.

13.2.2 What happens if the original triangle is right-angled? How many iterations are needed in this case to reach the limiting path?

13.2.3 What happens if the original triangle is obtuse-angled?

13.2.4 What happens if we work with a convex quadrilateral instead of a triangle?

13.2.5 Prove that the sequence

$$\langle P,f(P),f(f(P)),\ldots,\rangle$$

converges to a limiting point, which moreover is the same for all points P on BC. (Hint: Use trigonometry to find the distances from B to the sequence of iterates of P.)

Explanations

We now justify the claim made that the two inscribed triangles obtained “in the limit” above are congruent and homothetic to one another, the center of homothety being the symmedian point. Our proof will be based on computation.

We note firstly that the sides of the inscribed triangles are respectively perpendicular, in some order, to the sides of the parent triangle, implying that each triangle is similar to the parent triangle. Therefore, the two triangles are similar to one another; and homothetic too, as their sides are parallel to one another.

In Figure 13.8, the two inscribed triangles are labelled A′B′C′ and A″B″C″, respectively; for convenience, we refer to them as T′ and T″, respectively, and to $\triangle$ABC as T. The labelling of T′ and T″ has been done in the obvious manner, to draw attention to their similarity (note how their angles correspond to one another). Thus we have angle equalities such as

$$\angle B'A'C' = \angle BAC = \angle B''A''C'', \angle A'B'C' = \angle ABC = \angle A''B''C''$$

and so on.

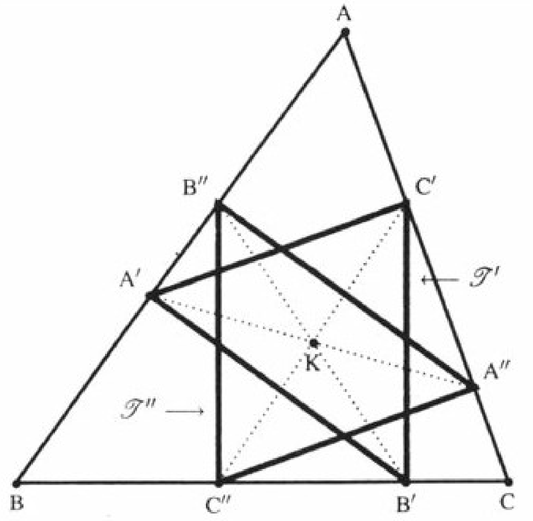

Figure 13.8. *Study of the two homothetic triangles*

We first show that triangles T′ and T″ are congruent to one another. Write a′,b′,c′ for the sides of T′, and a″,b″,c" for the sides of T″. Using elementary trigonometry we obtain the following relations:

$$a = BC = BB' + B'C = c' \sin B + a' \tan C'$$

$$b = AC = AC' + C'C = b' \tan A + a' \sin C,$$

$$c = AB = AA' + A'B = b' \sin A + c' \tan B.$$

The relations may be displayed in matrix form thus:

cotC 0 cscBcscC cotA0 0 cscAcotB · a′ b′ c′ = a b c .

Using well known formulas such as

$$\cos A = b2 + c2 - a2\ 2bc\ ,\sin A = a\ 2R,$$

where R is the radius of the circumcircle of the triangle, and doing a bit of simplification, the equations take the form shown below (here $\Delta = abc/4R$ denotes the area of the triangle):

$$1\ 4\Delta \cdot a2 + b2 - c2\ 0\ 2ca\ 2ab\ b2 + c2 - a2\ 0\ 0\ 2bc\ c2 + a2 - b2\ \cdot a'\ b'\ c' = ab\ c$$

The equations may be solved for a',b',c', and we get a most pleasing result:

$$a'\ b'\ c' = 4\Delta\ a2 + b2 + c2 \cdot ab\ c$$

Using the (certainly not too well known!) fact that

$$4\Delta\ a2 + b2 + c2 = \tan\omega,$$

where ω is the Brocard angle of the triangle, we may write this as

$$a'\ b'\ c' = \tan\omega \cdot ab\ c$$

The symmetry of the result tells us immediately that the very same equations will be obtained for a'',b'',c''. It follows that $a' = a'',b' = b''$ and $c' = c''$, and the congruence of the two inscribed triangles follows.

As segments $B''C''$ and $C'B'$ are congruent and parallel to one another, their center of homothety, K, lies at the center of rectangle $B''C''B'C'$. This means in particular that the perpendicular distance of K from side BC is

$$1\ 2a' = 1\ 2\tan\omega\ \cdot\ a = ka, \quad \text{where } k = 1\ 2\tan\omega.$$

We find similarly that the distances of K from sides CA and AB are kb and kc respectively. The ratios of the distances of K from the sides BC,CA,AB are therefore as $a : b : c$. But this is precisely the property which characterizes the symmedian point! The claim we made earlier therefore stands justified.

13.3 Angle bisectors

The iteration we study next deals with angle bisectors. We start by defining the relevant terms.

Incenter of a Triangle The internal angle bisectors of a triangle meet in a point—the *incenter* of the triangle, or the center of the circle that fits snug within the triangle. The circle is the *incircle* of the triangle. (To see why the bisectors meet in a point, use the fact that each point on an angle bisector is equidistant from the two arms of the angle; see Figure 13.9.)

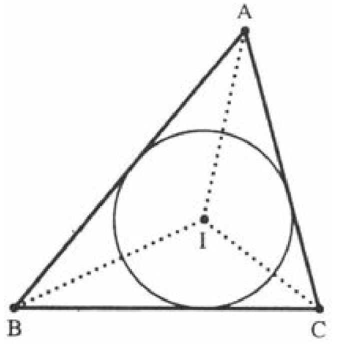

Figure 13.9.

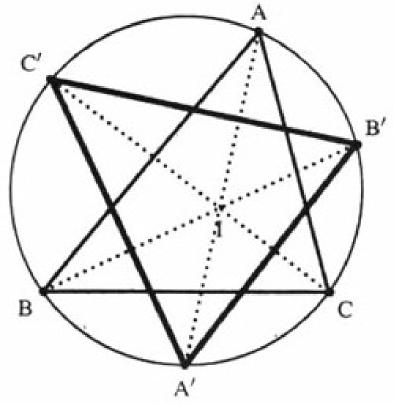

Figure 13.10.

The iteration

Let A,B,C be three distinct points on a circle Γ, and let $\triangle ABC$ be called T0; this triangle will serve as the seed of the iteration. Let the angle bisectors of T0 be drawn. and let them be extended beyond the sides of the triangle to meet the circle Γ. Let T1 be the triangle whose vertices are the three points of intersection so formed. (See Figure 13.10; the angle bisectors are shown in a dotted line; T1 is $\triangle A'B'C'$.)

Take the actions done so far to be one round of an iteration f; that is, take T1 = f(T0) to be the output produced with T0 as input. Let T2,T3,… be computed as follows:

$$T2 = f(T1), T3 = f(T2), T4 = f(T3), \ldots.$$

(Thus, to get T3 from T2, the angle bisectors of T2 are drawn and extended till they meet Γ. The three points of intersection are the vertices of T3; and so on.) What happens now? Figures 13.11 and 13.12 show the results of the iteration after 5 and 10 rounds respectively.

Observations

An interesting fact emerges when we examine these sketches: as we continue the iterations, *the triangles start to look more and more equilateral in shape*, and this happens regardless of the shape of the original triangle. Figures 13.13 shows the result of 10 rounds of the iteration using a differently shaped triangle as seed; the original triangle is shown in a light line and the final triangle in a thick line.

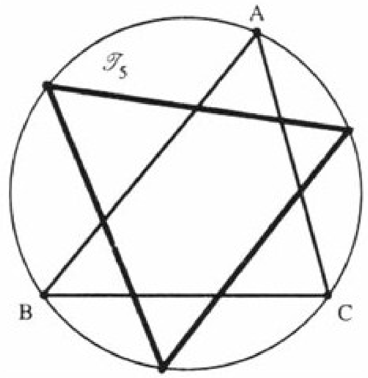

Figure 13.11. *After* 5 *rounds*

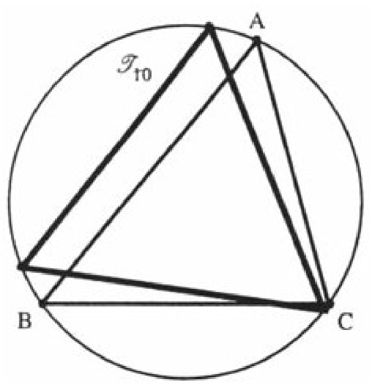

Figure 13.12. *After* 10 *rounds*

Note that the triangle itself does not stay fixed: T100 is distinct from T101, but they are both close to being equilateral. But T100 and T102 are "almost" the same triangle (they very nearly coincide), and likewise for T101 and T103.

Analysis

To see why convergence takes place, we compute the angles of T1 and check whether they are closer to one another than the angles of T0. This is easy to do.

In Figure 13.14, T0 is △ABC and T1 is △DEF. We compute the angles of T1 in terms of the angles of T0, using the well known theorem concerning "angles in the same segment." (If P,Q are points on an arc XY of a circle, then $\angle XPY = \angle XQY$.) We argue as follows. We have, $\angle FDE = \angle FDA + \angle ADE$. Next,

$$\angle FDA = \angle FCA = C\ 2\ , \angle ADE = \angle ABE = B\ 2\ ,$$

therefore $\angle FDE = 1\ 2(B + C) = 90\ \text{-}1\ 2A$. Similarly we get, $\angle FED = 90\ \text{-}1\ 2B$ and $\angle DFE = 90\ \text{-}1\ 2C$. Therefore the angles of △DEF are

$$90^\circ\text{-}A\ 2\ ,90^\circ\text{-}B\ 2\ ,90^\circ\text{-}C\ 2\ .$$

(Check that the angle sum is 180°.) So we have an iteration on triples that acts as follows:

$$(x,y,z) \mapsto 90\ \text{-}x\ 2,90\ \text{-}y\ 2,90\ \text{-}y\ 2.$$

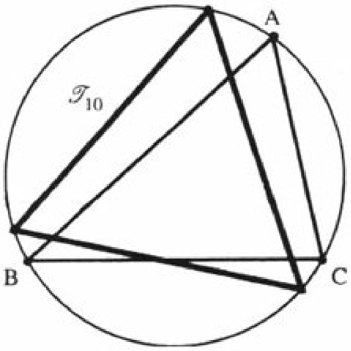

Figure 13.13. *A different* T0

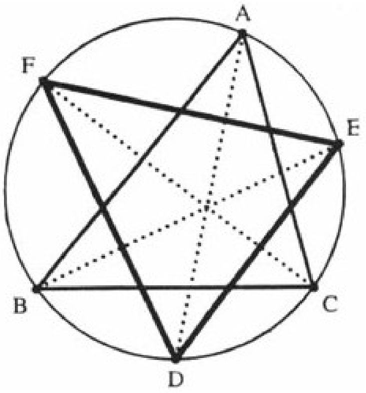

Figure 13.14. *Angle computations*

Consider the function f given by f(x) = 90 -1 2x. Let any number x0 be chosen, and let its f-orbit ⟨x0;f⟩ be computed. Does the sequence converge to a limit? We do a few trials to find out.

- x0 = 20 60.31,59.84,... The sequence produced is: 20, 80, 50, 65, 57.5, 61.25, 59.38,
- x0 = 40 60.16,59.92,... The sequence produced is: 40, 70, 55, 62.5, 58.75, 60.63, 59.69,
- x0 = 5 59.14,60.43,59.78,... The sequence produced is: 5, 87.5, 46.25, 66.88, 56.56, 61.72,

The figures clearly suggest that regardless of the seed chosen, the sequence converges to a limiting value of 60. Indeed this is so, and the exercise set below asks you to find a proof. As the iteration is the same for each angle of the triangle, it follows that each angle converges to 60°, and so the triangles do indeed converge to an equilateral shape.

Degree of 'non-equilateralness' of a triangle

Here is one way (the exercise set asks you to find another way!) of showing that the f-orbit of any starting number converges to a limiting value of 60. We have already seen that if T0 has angles $x^{\circ}, y^{\circ}, z^{\circ}$, then T1 has angles

$$90 - x\ 2^{\circ},\ 90 - y\ 2^{\circ},\ 90 - z\ 2^{\circ}$$

Consider the *differences* between the angles of T0; they are

$$|x - y|^{\circ}, |y - z|^{\circ} \text{ and } |z - x|^{\circ}$$

respectively. In the case of T1, the differences come out to be

$$|x - y|^{\circ}\ 2\ ,\ |y - z|^{\circ}\ 2 \text{ and } |z - x|^{\circ}\ 2\ ,$$

respectively. *So the differences have all been halved.*

Given any triangle, the extent to which it is *not* equilateral can be measured; it seems reasonable, surely, to describe a triangle with angles of $55^{\circ}, 60^{\circ}, 65^{\circ}$ as "closer to being equilateral" than one with angles of $10^{\circ}, 60^{\circ}, 110^{\circ}$. Given a triangle T with angles α, β, γ, a possible index for measuring how close it is to being equilateral is the quantity δ (T) given by

$$\delta(T) = |\alpha - \beta| + |\beta - \gamma| + |\gamma - \alpha|.$$

Observe that $\delta \geq 0$ for all triangles, and moreover that $\delta(T) = 0$ if and only if $\alpha = \beta = \gamma$; that is, if and only if T is equilateral.

In terms of this measure, the result obtained above can be stated succinctly as:

$$\delta(T1) = 1\ 2\delta(T0).$$

Similarly, $\delta(T2,) = \delta(T1)/2$, and so on. Since the δ-value is halved each time, it follows that $\delta(Tn)$ tends to 0 as n increases indefinitely. so the T's do indeed tend towards the equilateral shape.

Exercises

13.3.1 Let I0,I1,I2,… be, respectively, the incenters of the triangles T0,T1,T2, …. Does the sequence

$$\langle I0, I1, I2, \ldots \rangle$$

possess a limit?—i.e., do the incenters converge to a point? If so, what is this point?

13.3.2 Find another proof for the claim that for the iteration function given by f(x) = 90 -1 2x, the f-orbit corresponding to any seed converges to 60.

13.4 Medians

From angle bisectors we move on medians. Here are two definitions.

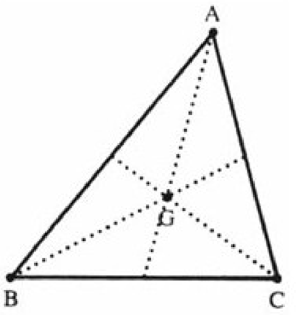

Figure 13.15.

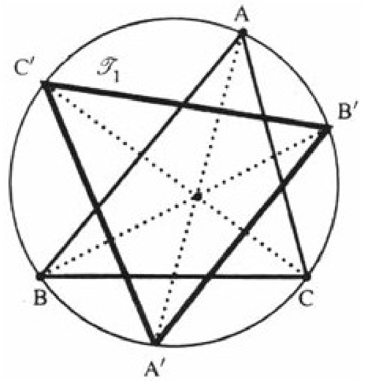

Figure 13.16.

Median of a Triangle; Centroid The line segment connecting a vertex of a triangle to the mid-point of the opposite side is called a *median* of the triangle. The three medians of any triangle meet in a point, called the *centroid* of the triangle; see Figure 13.15.

The iteration

We start with triangle T0, with vertices A,B and C, inscribed in a circle Γ. Let its medians be drawn, and let them be extended, as was done with the angle bisectors, to meet Γ. Let T1 be the triangle whose vertices are the three points of intersection; see Figure 13.16.

Let the actions performed so far be considered as one round of an iteration f; take T1 = f(T0) to be the output produced from the input T0.

Since T0 and T1 share the same circumcircle, the iteration can be continued. So we get

$$T2 = f(T1), T3 = f(T2), T4 = f(T3), \ldots,$$

and so on. (To produce T3 from T2, the medians of T2 are drawn and extended till they meet Γ; the three points of intersection are the vertices of T3.)

What happens now? Figures 13.17 and 13.18 show the results of the iteration after 5 and 15 rounds respectively. (In each case T0 has been drawn in a light line and the final Ti in a heavy line.)

Remarkably, we find exactly the same thing happening as in the angle-bisector iteration: *the triangles start to look more and more equilateral in shape*. And as earlier, this happens regardless of the shape of the seed triangle. Figure 13.19 shows the result after 15 rounds of the iteration, with a differently shaped triangle as seed.

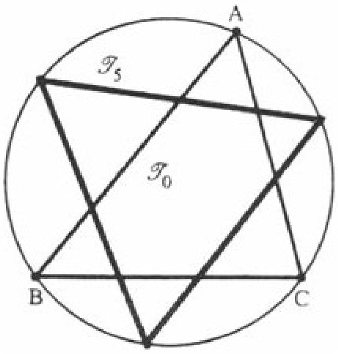
A
$\mathcal{T}_5$
$\mathcal{T}_0$
B
C

Figure 13.17. *After* 5 *rounds*

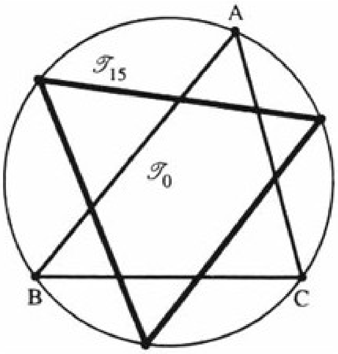
A
$\mathcal{T}_{15}$
$\mathcal{T}_0$
B
C

Figure 13.18. *After* 15 *rounds*

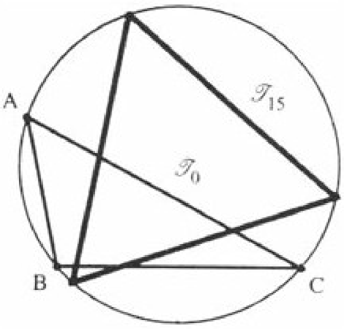
A
$\mathcal{T}_{15}$
$\mathcal{T}_0$
B
C

Figure 13.19. *After* 15 *rounds, with another* T0

⋆ ⋆ ⋆ ⋆ ⋆ ⋆ ⋆ ⋆ ⋆ ⋆ ⋆ ⋆ ⋆ ⋆ ⋆ ⋆ ⋆ ⋆ ⋆ ⋆

In order to do an analysis of this problem, we must compute the angles of

T1 and check whether they are in some sense closer to one another than the angles of the parent triangle; or we could work with the sides rather than the angles. This was easy to do in the case of the angle-bisector iteration. Unfortunately the analysis in this case is rather more involved. The ambitious reader may want to try finding a proof independently, before proceeding to the analysis given at the end of this chapter (in the Appendix).

Exercises

13.4.1 Let Gi denote the centroid of Ti,i = 0,1,2,…, and consider the sequence

$$\langle G0,G1,G2,G3,\ldots\rangle.$$

Does the sequence possess a limit? That is, do the centroids converge to some limiting position?

13.4.2 Instead of drawing medians, we could draw lines according to the following scheme.

Let k denote a fixed number, $0 < k < 1$. Labelling T0 as ABC, we locate points P on BC,Q on CA and R on AB according to the rule

$$BP\ PC = CQ\ QA = AR\ RB = k,$$

draw the lines AP,BQ,CR, then extend them to meet the circle Γ; the three points of intersection are then the vertices of T1. Likewise we generate T2 from T1,T3 from T2, and so on.

Let Gi be the centroid of Ti. Does the sequence ⟨G0,G1,G2,…⟩ converge to a point? Does convergence depend on the choice of k? Use a computer to do experiments and make a guess based on your findings. (Observe that $k = 0.5$ corresponds to the original iteration.)

Numerical analysis

As mentioned earlier, the analysis of the medians iteration is rather involved. We shall give the full details later. For now, we shall only show how the iteration can be described in algebraic terms; this then allows a numerical investigation.

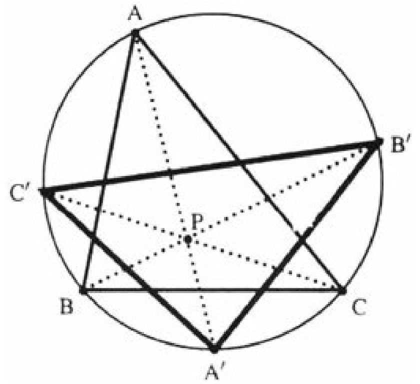

Figure 13.20. *A subsidiary result*: $B'C' : C'A' : A'B' = AP \cdot BC : BP \cdot CA : CP \cdot AB$

We first establish a subsidiary result. In Figure 13.20, we see a triangle ABC, inscribed in its circumcircle Γ, with an interior point P. The lines AP,BP and CP, when extended, meet the circumcircle again in the points A′,B′ and C′, respectively. We shall show that

$$B'C' : C'A' : A'B' = AP \cdot BC : BP \cdot CA : CP \cdot AB.$$

The proof is as follows. From the chord theorem, we get:

$$AP \cdot PA' = BP \cdot PB' = CP \cdot PC'.$$

Let k be the common value of these three products. From the similarity of triangles B′PC′ and CPB, we get:

$$B'C' \; BC = PB' \; PC \,, \therefore B'C = BC \cdot PB' \; PC = k \cdot BC \; BP \cdot CP = k \cdot AP \cdot BC \; AP \cdot BP \cdot CP.$$

It follows that $B'C' : C'A' : A'B' = AP \cdot BC : BP \cdot CA : CP \cdot AB$, as claimed.

★★★★★★★★★★★★★★★★★★★★

If P happens to be the centroid of triangle ABC, then the lengths AP,BP,CP are two-thirds of the lengths of the medians through A,B,C, respectively. Writing ma,mb,mc for the lengths of these three medians, and a′,b,c′ for the lengths of the sides of triangle A′B′C′, we see that

$$a' : b' : c' = am\ a : bmb : cmc,$$

certainly a pretty and compact result!

From **Apollonius's theorem** we know that

ma2 = b2 2 + c2 2 -a2 4 ,mb2 = c2 2 + a2 2 -b2 4 ,mc2 = a2 2 + b2 2 -c2 4 .

Using these relations, we are now able to make a numerical study of the convergence profile of the medians-iteration. Writing the relations a′ : b′ : c′ = ama : bmb : cmc as

a′2 : b′2 : c′2 = a2m a2 : b2m b2 : c2m c2,

we see that a′2 : b′2 : c′2 is the same as

a2(2b2 + 2c2 - a2) : b2(2c2 + 2a2 - b2) : c2(2a2 + 2b2 - c2).

It will be evident now that it is more convenient to deal with the *squares* of the sides of the triangle than with the sides themselves. Further, as it is only the *ratios* that the sides bear to one another that are of relevance to us, we define quantities x,y,z,x ′,y′,z′ as follows:

x = a2 a2 + b2 + c2,y = b2 a2 + b2 + c2,z = b2 a2 + b2 + c2,

and

x′ = a′2 a′2 + b′2 + c′2,y′ = b′2 a′2 + b′2 + c′2,z′ = b′2 a′2 + b′2 + c′2;

then x + y + z = 1 = x′ + y′ + z′,x : y : z = a2 : b2 : c2, and x′ : y′ : z′ = a′2 : b′2 : c′2.

Now we describe what iteration "does" to the triple (x,y,z). From the above we see that the iteration maps (x,y,z) to the triple (x′,y′,z′), where

x′ = x(2y + 2z - x) 4(xy + yz + zx) - (x2 + y2 + z2),

y′ = y(2z + 2x - y) 4(xy + yz + zx) - (x2 + y2 + z2),

z′ = z(2x + 2y - z) 4(xy + yz + zx) - (x2 + y2 + z2).

The common denominator here is the sum

x(2y+2z-x)+y(2z+2x-y)+z(2x+2y-z) = 4(xy+yz+zx)-(x2+y2+z2).

The important consequence of this formulation is that we have at hand a fully numerical version of the iteration. We now test it out.

Example 1. Consider a triangle with sides a = 3,b = 4,c = 5. As a2 : b2 : c2 = 9 : 16 : 25 and 9 + 16 + 25 = 50, the starting triple is (x0,y0,z0) = (0.18,0.32,0.50). The subsequent triples, computed via the formula displayed above, are:

$$(0.1800, 0.3200, 0.5000), \quad (0.3108, 0.3936, 0.2956),$$
$$(0.3374, 0.3279, 0.3347), \quad (0.3333, 0.3333, 0.3333),$$
$$(0.3333, 0.3333, 0.3334), \quad (0.3333, 0.3333, 0.3333).$$

The speed with the triple (0.3333, 0.3333, 0.3333) (which corresponds to the equilateral triangle) is reached is most impressive! (There is a small amount of round-off error in the displays: the last two triples sum to 0.9999 rather than to 1.)

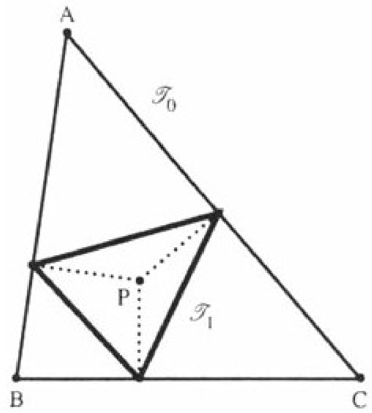

Figure 13.21. *The pedal triangle*

Example 2. Now consider a triangle with sides a = 2,b = 5,c = 6. We get the following result:

$$(0.0615, 0.3846, 0.5538), \quad (0.1789, 0.5210, 0.3001),$$
$$(0.3194, 0.2778, 0.4028), \quad (0.3410, 0.3322, 0.3268),$$
$$(0.3333, 0.3334, 0.3333), \quad (0.3333, 0.3333, 0.3333).$$

Once again the swift rate of convergence is in full view.

★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★

The full proof that the sequence of iterates converges to an equilateral shape is given at the end of the chapter.

13.5 Pedal triangles

We start with an acute angled triangle T0, with vertices A,B,C, and a point P lying within the region enclosed by the triangle. Let perpendiculars be dropped from P to each side of T0, and let the feet of these perpendiculars be the vertices of a new triangle T1. This triangle is the *pedal triangle* of P relative to T0. Note that T1 is "inscribed" in T0, for its vertices lie on the sides of T0, one to each side (see Figure 13.21).

As P lies within T1, we can repeat these actions and construct the pedal triangle of P relative to T1, by dropping perpendiculars from P to the sides of T1 and joining the feet of the perpendiculars to one another. In this manner we obtain another triangle T2, and P lies within this triangle too. We now construct the pedal triangle of P relative to T2, obtain another triangle, T3, and we continue thus, obtaining a sequence ⟨T0,T1,T2,T3,…⟩, in which each Ti is inscribed within the preceding one. The triangles shrink in size very rapidly, and soon the eye is unable to distinguish the triangles from the point P itself. Figure 13.22 shows the T's till T3.

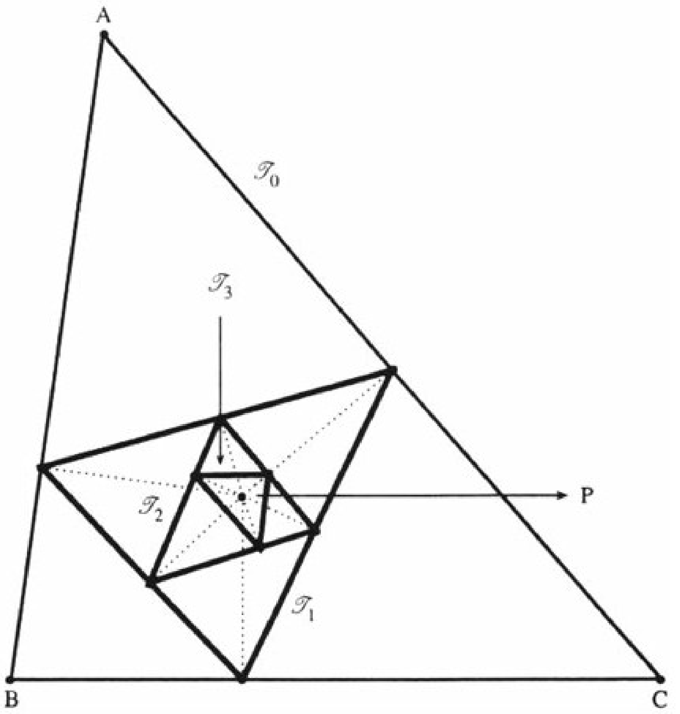

Figure 13.22. *Inscribed pedal triangles*

There is a nice feature about the sequence that could easily escape the eye: T3 is similar to T0! Examine the figure carefully; you should be able to see the similarity on your own. (The two triangles are in fact homothetic to one another, and P is the center of homothety.)

Exercises

13.5.1 Prove that T3 is similar to T0. (*Hint*. Compute the angles of T1. Do not draw the pedal triangles after T1, else the figure becomes too confusing!)

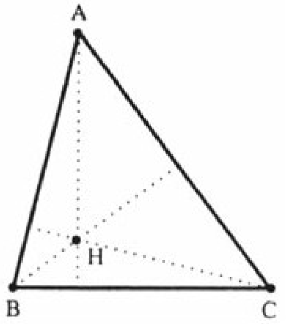

Figure 13.23.

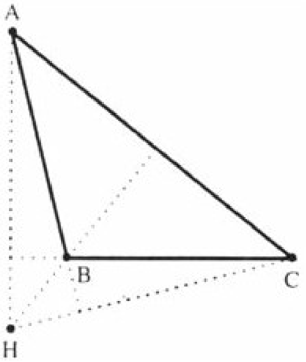

Figure 13.24.

13.6 Altitudes

Having studied iterations associated with the angle bisectors and the medians of a triangle, it is natural to consider the case of altitudes. The iteration turn s out to be quite exotic, producing complete chaos! The earlier iterations were both fairly "well behaved." in terms of yielding predictable results, but that is not the case here. Before we start we need two definitions.

Altitude, Orthocenter The line drawn from a vertex of a triangle perpendicular to the opposite side is called an *altitude* of the triangle. The three altitudes of a triangle always meet in a point, called the *orthocenter* of the triangle. The orthocenter can lie either inside or outside the triangle (see Figures 13.23 and 13.24).

The iteration

We start the iteration with triangle ABC inscribed in a circle Γ; call the triangle T0, and let its orthocenter be H0. The altitudes of T0 are drawn and extended till they meet Γ. Let T1 be the triangle whose vertices are the three new points of intersection, and let H1 be its orthocenter. Take T1 to be the output arising from the input T0.

We now iterate these actions: we draw the altitudes of T1, extend them till they meet Γ, and join the three points of intersection to one another; then the resulting triangle is T2, and its orthocenter is H2. We continue in this manner, obtaining a sequence of triangles,

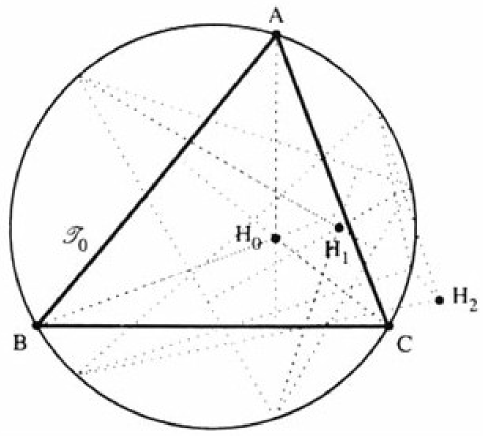

Figure 13.25. *Altitude iteration, after* 2 *rounds*

$$\langle T0,T1,T2,T3,\ldots\rangle,$$

and a corresponding sequence of orthocenters,

$$\langle H0,H1,H2,H3,\ldots\rangle.$$

Having studied in some detail the iterations with angle bisectors and medians, we may now be tempted to guess that the sequence of orthocenters will converge to a point. But we are in for a surprise!

Figure 13.25 shows the result after 2 rounds of the iteration (the orthocenters are shown, but some details have been hidden so as to avoid a visual mess).

Figure 13.26 shows the result after 25 rounds of the iteration (the labelling of the Hi's has been suppressed, for obvious reasons; likewise for the construction lines used). Note the dust-like scatter.

Figure 13.27 shows the result after 85 rounds. There appears now to be a random scatter! (There is a curious star-like effect, with streaks in six directions symmetrically placed about the circle; but this is hard to explain.) If we continue the iterations still further, we find a circular region entirely covered with dots. A careful analysis reveals that the radius of this circle is three times the radius of the circumcircle of T0.

Remark. For some choices of the starting triangle T0, a dust-like effect is *not* obtained; instead, we get a cycle. The reader may want to investigate under what conditions this will happen.

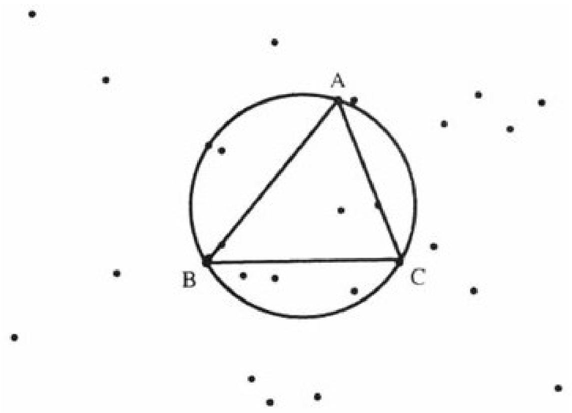

Figure 13.26. *After* 25 *rounds*

Exercises

13.6.1 For which triangles does the orthocenter lie (i) within the triangle (ii) on one side of the triangle (iii) one side of the triangle? Can the orthocenter ever lie in the interior of a side of the triangle?

13.6.2 Show that the altitudes of a triangle meet in a point.

13.6.3 In the iteration just considered, do the Ti's converge to any limiting shape?

13.6.4 Consider all possible triangles inscribed in a circle Γ, and let $\mathcal{H}$ denote the set of their orthocenters. Show that $\mathcal{H}$ occupies a circular region whose radius is three times that of Γ. You will need to use the following well known result due to Euler:

> *For any triangle, the circumcenter O centroid G and orthocenter H lie on a straight line the "Euler line" of the triangle, with H between O and G and moreover,* OH = 3OG.

13.6.5 It had been noted above that "for some choices of the starting triangle T0, a dust-like effect is *not* obtained; instead, we get a cycle." Investigate under what conditions this will happen.

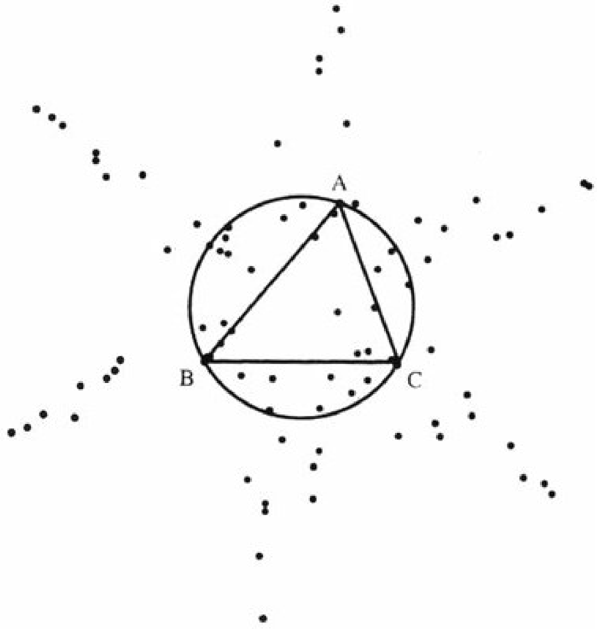

Figure 13.27. *After* 85 *rounds*

13.7 Pretty pictures

Next we show how simple iterative actions can give rise to very attractive pictures.

13.7.1 Nested Squares

Let k be a number lying between 0 and 1. Let S0 be a square, and let its vertices be labelled in cyclic order as A,B,C,D. Let points P,Q,R,S, respectively, be located on the sides *AB*, BC,CD,DA of S0, according to the following rule:

$$AP\ AB = BQ\ BC = CR\ CD = DS\ DA = k.$$

(If k = 0.5, then P,Q,R,S are the mid-points of the sides of S0.) The segments PQ,QR,RS,SP are drawn, giving another square, S1, which we take to be the output of a function with input S0. Observe that S1 is smaller than S0 and is wholly contained within it; it has the same center as its parent square, but has been turned through a small angle. To draw S2 we temporarily rename S1 as *ABCD* and locate points P,Q,R,S on its sides as per our rule; then S2 is the square *PQRS*; and so on. We thus obtain a sequence of squares

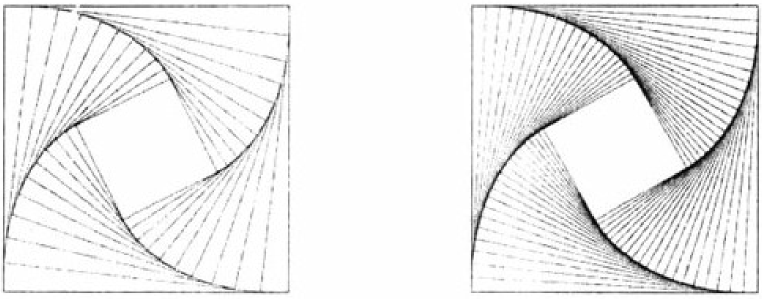

$k = 0.1$, 10 *iterations*

$k = 0.05$, 20 *iterations*

Figure 13.28. *Nested squares*

$$\langle S0,S1,S2,S3,\ldots\rangle,$$

in which each square lies inside all the preceding ones. It is easy to see that the squares converge to the center of symmetry of the square. The picture that results is a pretty one; two such pictures have been portrayed in Figure 13.28.

Exercises

13.7.1 Prove that Si is a square for each i.

13.7.2 What picture results when k = 0.5?

13.7.3 Which values of k generate the nicest pictures? (You will need to do some experimentation to find out!)

13.7.4 If you work with non-square quadrilaterals instead of squares, what do you find?

13.7.5 If you carefully study the pictures generated above, you will see many spiralling curves. This is actually an optical effect: even though we have not drawn any curve, it so happens that the lines that we have drawn are all tangent to a particular curve, and it is this curve that we “see. So in this sense the curve is not merely an illusion but has been created by the lines. (The lines are said to *envelope* the curve.)

List some daily life situations in which straight lines create the effect of a curve. (Hint. Think of the “tea-cup curve.”)

13.7.6 Four men (A,B,C,D) are stationed at the comers of a square, one at each vertex, and at a given instant they start walking towards one other: A towards B,B towards C,C towards D and D towards A, all at the same speed. As they are all in movement, each one must continuously alter his direction of movement. What paths do the men trace out? Where do they meet?

13.7.7 A dog spots a rabbit and starts to chase it; the rabbit heads straight for its burrow. What sort of path does the dog follow? (Experiment with different possibilities. If the dog and rabbit run in the same direction, then the path is simply a straight line. What happens if the rabbit initially starts in a direction which is perpendicular to the dog’s path?)

13.7.2 Nested Triangles

We do exactly the same thing with triangles. Let T0 be a equilateral triangle and let k be any number lying between 0 and 1. We label the vertices of T0 as A,B,C, then locate points P,Q,R on the sides of T0 using the following rule:

$$BP\ BC = CQ\ CA = AR\ AB = k;$$

then we declare T1 to be △PQR. We find that T1 is equilateral, but of smaller size that T0; it has the same center of symmetry as its parent triangle, and is wholly contained within it. We similarly generate T2 from T1,T3 from T2, and so on. The sequence

$$\langle \mathrm{T0},\mathrm{T1},\mathrm{T2},\mathrm{T3},\ldots \rangle$$

then consists of nested triangles, one within the other, converging to the center of symmetry of the triangle. Figure 13.29 shows two sample pictures.

This example and the previous one ("nested squares") form the basis of,many pretty *Nail-and-thread* art pieces that can be found in the craft center of a school.

Exercises

13.7.8 Suppose that T0 is not equilateral. What can be said about T1,T2,T3, … ? Are the triangles further down the sequence 'more equilateral' than the earlier ones?

13.8 Fractals

Fractals constitute a relatively modem discovery; they date to the latter half of the 20th century and may be said to have originated in the work of the French mathematician and geometer Benoit Mandelbrot. In the bargain a new class of mathematical objects has emerged, one that holds considerable relevance to the study of the natural world. The concept is a relatively simple one, but it is remarkable how far it takes us. We illustrate the idea with a simple example —that of the *snowflake curve*.

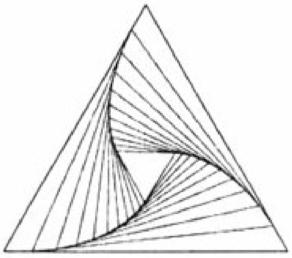

$k = 0.1$, 10 *iterations*

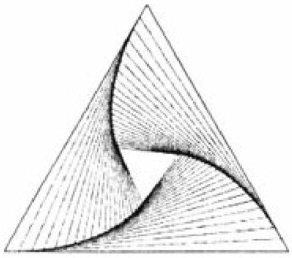

$k = 0.05$, 20 *iterations*

Figure 13.29. *Nested triangles*

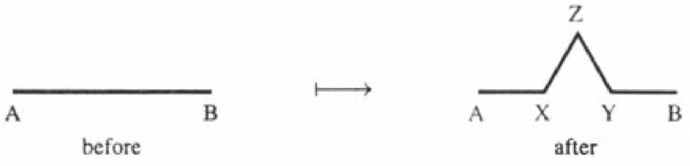

Figure 13.30. *Action of* f *on a line segment*

★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★

We first describe how the function f acts. Let a line segment ℓ = AB be given, and let f act on ℓ thus: we locate the points X,Y which trisect AB (with X nearer to A than to B, so the points occur on ℓ in the order A,X,Y,B), and then locate Z such that ΔXYZ is equilateral. The ambiguity regarding the side of ℓ on which Z is located is resolved thus: in ΔXYZ, the vertex order (X,Y,Z) is required to be *anticlockwise*. Following this, segment XY is deleted and replaced by segments XZ and Y Z (see Figure 13.30).

Now let a geometric figure G0 be given, composed only of line segments; for instance, G0 could be a polygon. We let f act on each edge of G0, and obtain a new geometric figure, say G1. The same treatment is given to G1, and so on, and we thus obtain a sequence of geometric figures,

$$\langle G0, G1, G2, G3, \ldots \rangle,$$

each of which is composed only of line segments, which get smaller and smaller as we go down the sequence. The following sequence of figures shows what happens when G0 is an equilateral triangle. Note the beauty of the figures! It is possible in the minds' eye to imagine what happens 'in the limit', but it is impossible to draw it. The limiting object is known as the *snowflake curve*— for obvious reasons, as this is what the object resembles it its broad outline and symmetry. (We have stopped at G5 because, as you will readily see, drawing even the next shape is going to be rather tedious.)

Is it not remarkable how such a simple rule can give rise to so pretty a shape?

13.9 Map-in-a-map

A USAMO problem Our last example of iterations in geometry will be based on the following elegant problem which appeared in the USA mathematical olympiad (USAMO) of 1978.

> *We are given two square maps of the same country, drawn to different scales. The smaller map is placed on and within the boundary of the larger one. Show that there is precisely one point on the small map which represents the same place of the country as the point of the large map that lies directly below it. Give a way of locating this point.*

The use of the word "square" may suggest that the result holds only for maps which are square in shape, but this is clearly no restriction, for any map can be embedded in a larger map that is square in shape.

Notation

Let ABCD denote the larger map, and A′B′C′D′ the smaller one. We use the prime symbol (′) to show the correspondence between the maps: if X is any point on the large map, then the point on the small map corresponding to X will be denoted by X′. We also write $X' = f(X)$, where f is the function describing how the two maps correspond to one another (in Figure 13.32, $A' = f(A)$,$B' = f(B)$, and so on).

Next, we denote by r the scale factor of the small map relative to the large map ($0 < r < 1$). The scaling applies uniformly to all parts of the map, so if two distinct points P and Q of the large map correspond to (distinct) points P′ and Q′ of the small map, then $P'Q' = r \cdot PQ < PQ$. So if "dist" represents the distance function, with dist (P,Q) = the distance between the points P and Q, then the function f has the property that

$$\text{dist}\,(f(P),f(Q)) = r \cdot \text{dist}\,(P,Q) < \text{dist}\,(P,Q),$$

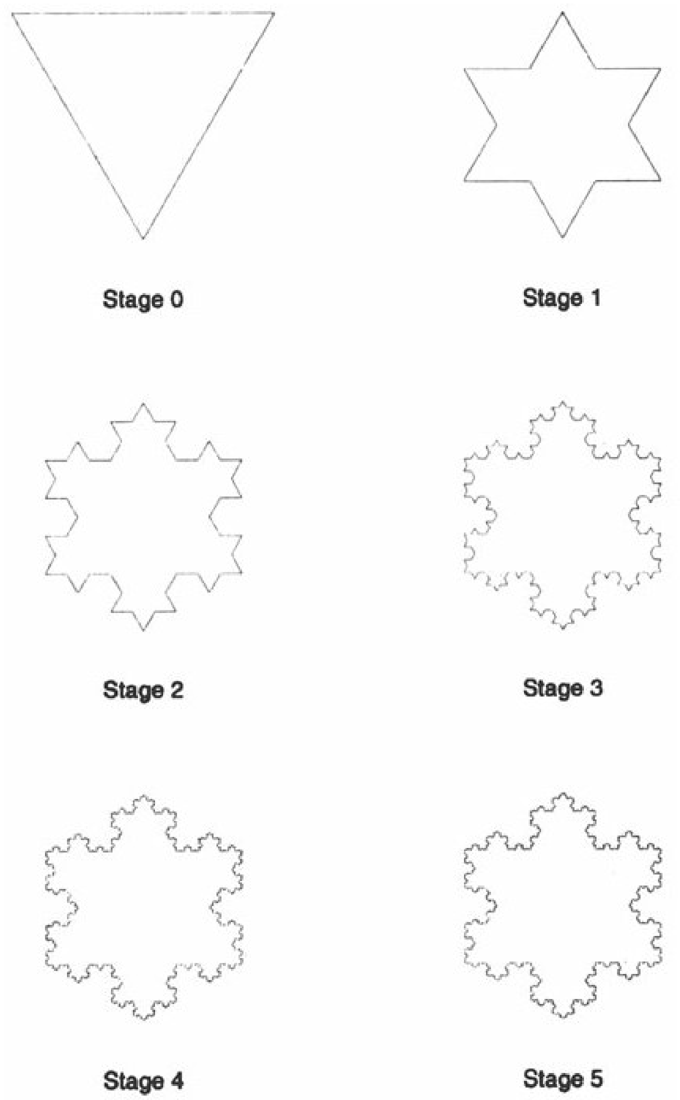
Stage 0
Stage 1
Stage 2
Stage 3
Stage 4
Stage 5

Figure 13.31. *Stages* 0 *to* 5 *of the snowflake curve*

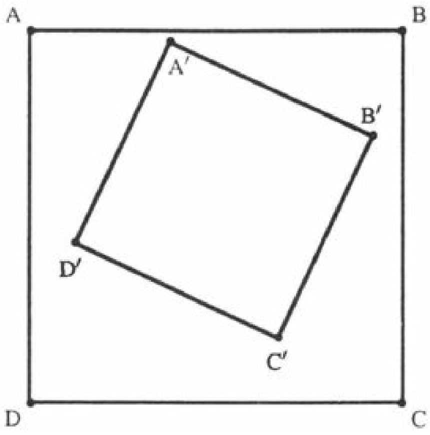

Figure 13.32. *Map-in-a-map*

for all points P≠Q of the large map. Note in particular that for all points P≠Q of the large map, the distance between f(P) and f(Q) is strictly less than the distance between P and Q.

The problem asks us to show that there is one and only one point O of the large map such that f(O) = O (i.e., f has one and only one fixed point), and to show how the point can be located geometrically.

Solution to the problem

Observe that the 'only one' pan is easy: there certainly cannot be *more than one* such point. For if P and Q were both fixed points (P≠Q), then we would have P = f(P) and Q = f(Q), i.e., P′ = P and Q′ = Q, implying that P′Q′ = PQ; but this contradicts the relation P′Q′ < PQ obtained above. So there can be at most one fixed point.

When f is applied to the entire square M = ABCD, we get the smaller square M′. What happens if we apply f to M′? We obviously get a square M″ which lies within M′ and bears exactly the same relation to M′ as M′ does to M. The scale of M″ is r times that of M′, and therefore r2 times that of M. If we are given only M and M′, we can easily construct M″ using Euclidean tools. For, point A″ must lie inside M′ and be located such that ΔA″A′B′~ΔA′AB. We can easily do this using the tools of Euclidean geometry.

This can clearly be continued indefinitely: we can apply f to M'' and get M''', and so on. The squares thus obtained shrink at a geometric rate, and each lies within the preceding square. So we get a "nested" sequence of squares (see Figure 13.33).

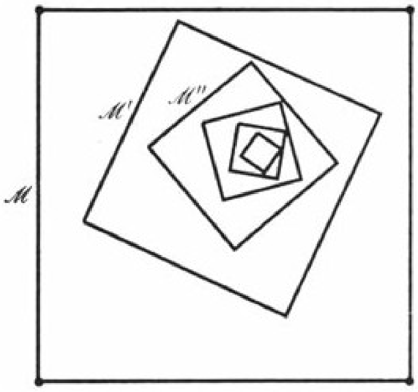

Figure 13.33. *Nested maps*

If we focus attention (see Figure 13.34) on the f-iterates of the vertex A alone, we see that the sequence of iterates

$$A, A', A'', A''', A'''', \ldots$$

quite visibly converges to a limiting point *O. This is the fixed point of* f.

Locating the fixed point geometrically

It is possible to find the fixed point O without too much work. Indeed, we need only two iterates, A' and A''. (If we are not given A'', then we first construct it as indicated above.) We give below a geometric procedure for locating O.

Observe firstly that for any three points P,Q and R, the image of $\triangle PQR$ under f, namely $\triangle P'Q'R'$, is similar to $\triangle PQR$. Since O is a fixed point ($O' = O$), we infer that $\triangle OAA'$ is similar to $\triangle OA'A''$, and $\angle OA'A'' = \angle OAA'$. Therefore $A'A''$ is tangent (at A') to the circumcircle of $\triangle OAA'$. This circle can readily be drawn, even though we do not know where 0 lies; for its center must lie on the line through A' perpendicular to $A'A''$, and it must also lie on the perpendicular bisector of segment AA'.

Next, we have ∠OA′A = ∠OA″A′, therefore AA′ is tangent (at A′) to the circumcircle of △OA″A′. This circle too can be drawn readily, even though we have not yet found O, by applying the same logic: its center must lie on the line through A′ perpendicular to AA′ and also on the perpendicular bisector of segment A′A″.

The two circles meet at A′ and O, therefore O is located; see Figure 13.35.

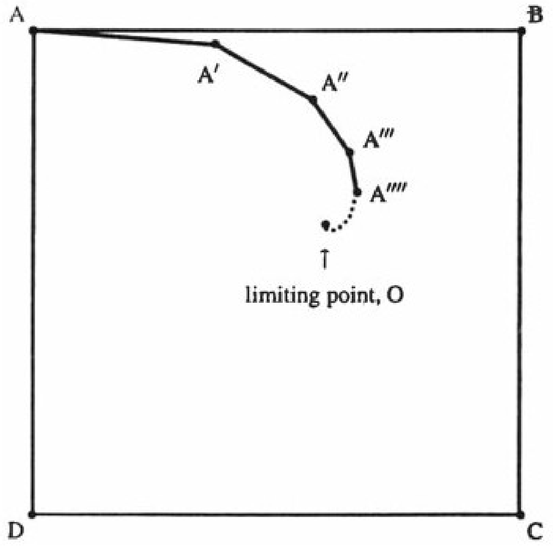

Figure 13.34. *Successive iterates of A*

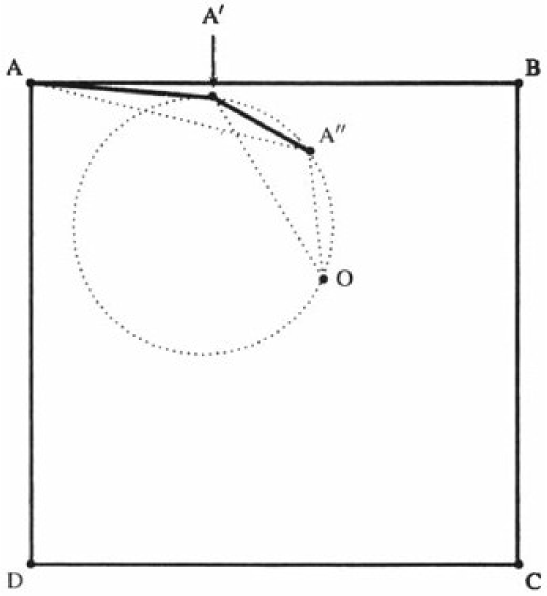

Figure 13.35. *Locating the fixed point* A

Closing remarks

The subject of this section is related to what is known as *fixed point theory*, a very active area of mathematics today—it has whole journals and websites devoted to it! Fixed point theory is part of a subject called *topology*.

The map-in-a-map result is a special case of an extremely important result in this field called Brouwer's *fixed point theorem*. The theorem has all kinds of nice applications in subjects that would seem very far removed from mathematics; for example, economics. The theorem was discovered by a Dutch mathematician, L E J Brouwer (1881–1966).

The "map" result is an application of Brouwer's theorem in *two* dimensions. Even the one dimensional case of the theorem is elegant:

> *Let* f *be a continuous map from a finite interval I into itself; then* f *has a fixed point.*

The two dimensional version of the theorem states the following:

> *Let* f *be a continuous function from a convex, bounded set* S *into itself;*

then fhos *a fixed point.*

The word “continuous” in these two statements here has a technical meaning which we will not go into. In the context of our “map” result, what it means is that while we are allowed to shrink the map as much as we wish, and place it inside the larger map at any orientation, even upside down, we are not permitted to *tear* it. Tearing the map would qualify as a “discontinuous” operation.

The theorem may be similar extended to cover the case of three dimensions.

Here is a pretty application of the theorem; it is called the *coffee cup theorem*. It says that if you take a cup of coffee and stir the coffee with a spoon, then at the end there is some coffee particle that is in exactly the same place that it was before the stirring. If you “push” this particle out of its position, then somewhere else in the cup a particle will end up in its original position!

Exercises

13.9.1 Consider the case when the sides of the small map are parallel respectively to the sides of the large map. Two subcases of interest are: (a) $A'B' \parallel AB$, $B'C' \parallel BC$; (b) $A'B' \perp AB$, $B'C' \perp BC$.

Find simpler ways of locating the fixed point O in these two subcases.

13.9.2 What happens if the small map juts outside the large map? Does a fixed point necessarily exist?

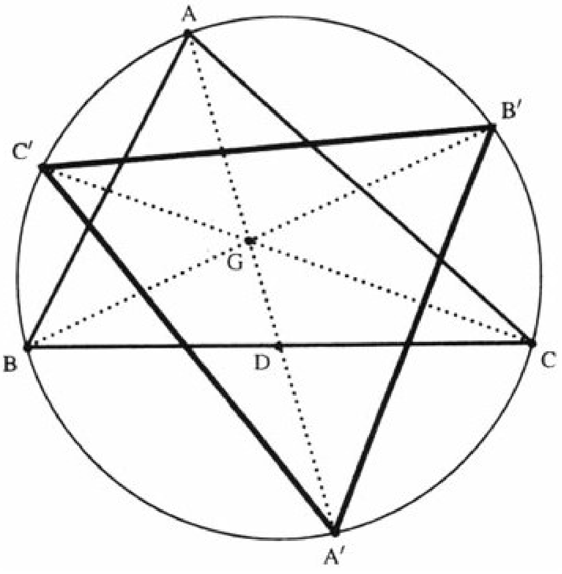

Figure 13.36.

13.9.3 I take a walk as follows: from my house I walk 1 km East, then 1/2 km North, then 1/4 km West, then 1/8 km South, then 1/16 km East, then 1/32 km North, and so on. Where do I find myself eventually? (Give a geometric argument for locating the limiting position.)

13.10 Appendix—analysis of the medians-iteration

We now take up the analysis of the medians iteration studied in Section 13.4. We had said then that the analysis was rather challenging. We now give the full details. Our approach will be.to translate the geometric iteration into an algebraic one, and then recast it into a geometric context using the idea of BARYCENTRIC COORDINATES. (For those who are unfamiliar with this term, a full definition is given below.)

★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★

In Figure 13.36, AD,BE,CF are the medians of triangle ABC, and their lengths are ma,mb,mc, respectively. The medians, when extended, meet the circumcircle Γ at points A′,B′,C′. Let the sides of triangle ABC be a,b,c, and let the lengths of the medians be ma,mb,mc, respectively. Then $a' : b' : c' =$

$ama : bmb : cmc$ (this was proved earlier). It follows that

$$a'2 : b'2 : c'2 = a2(2b2 + 2c2 - a2) : b2(2c2 + 2a2 - b2) : c2(2a2 + 2b2 - c2).$$

Let $x : y : z = a2 : b2 : c2$ and $x' : y' : z' = a'2 : b'2 : c'2$. Then, we have:

$$x' : y' : z' = x(2y + 2z - x) : y(2z + 2x - y) : z(2x + 2y - z).$$

Noting that the sum of the numbers in the triple on the right side is:

$$x(2y+2z-x)+y(2z+2x-y)+z(2x+2y-z) = 4(xy+yz+z\kappa)-(x2+y2+z2),$$

we define a mapping on triples of real numbers whose sum is 1, as follows. If (x,y,z) is such a triple, i.e., with $x + y + z = 1$, then its image is $(x',y',z') = f(x,y,z)$, where

$x' = x(2y + 2z - x)\ 4(xy + yz + zx) - (x2 + y2 + z2)$, $y' = x(2z + 2x - y)\ 4(xy + yz + zx) - (x2 + y2 + z2)$, $z' = x(2x + 2y - z)\ 4(xy + yz + zx) - (x2 + y2 + z2)$, (13.1)

We need to show that the iterates tend to the triple (1/3, 1/3, 1/3).

Transformation of problem, using barycentric coordinates

Let us regard (x,y,z) as the *barycentric coordinates* of a point P with reference to some fixed equilateral triangle T. Before going on, we define what this means.

Oriented area Given a polygon P with its boundary (usually denoted by ∂P) traversed in an anticlockwise or clockwise direction, its *oriented area* (denoted by [P]) is its ordinary area affixed with a *positive* sign if the direction is anticlockwise, and a *negative* sign if the direction is clockwise.

From the definition, it follows that [ABC] = -[ACB] for any triangle ABC, for the "sense" of the closed path A → B → C → A is opposite to that of the path A → C → B→ A.

Barycentric coordinates Given a fixed triangle T with vertices A,B,C (we shall call it as the *triangle of reference*), for any point P in the plane of the triangle, let

$$x : y : z = [PBC] : [PCA] : [PAB] .$$

We shall assume that the labelling of triangle ABC is done in an anticlockwise order (so [ABC] is positive). The location of P is uniquely determined once the ratios x : y : z are known (this follows from elementary geometry). We call the triple (x,y,z), or the ratios x : y : z, the *barycentric coordinates* of P. Examples: The barycentric coordinates of vertex A are (1,0,0), those of the mid-point of side BC are (1/2, 1/2, 0) , and those of the centroid of the triangle are (1/3, 1/3, 1/3). The only triple that does not correspond to any point in the plane is (0,0,0).

Note that the triple (kx,ky,kz) corresponds to the same point as (x,y,z). (So the centroid may also be described by the triple (1, 1, 1), and the mid-point of side BC by the triple (1, 1, 0) .) If x,y,z all have the same sign, then the point P lies inside the triangle, and if $x < 0, y > 0, z > 0$, then P lies within the angle determined by the rays $\overrightarrow{AB}$and $\overrightarrow{AC}$but on the opposite side of BC as compared with A.

★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★

Coming back to our iteration, we see that the image of a point P, determined by a triple (x,y,z), is a point P′, determined by the triple (x′,y′,z′), where x′,y′,z′ are computed using equation (13.1). We need to check whether the iterates of P under this mapping tend to the centroid G of the triangle of reference, T.

We assign the following coordinates to the vertices of T = △UV W:

$$U = (0,3), V = (-1,0), W = (1,0).$$

Then. the centroid of T is G = (0,1⁄3), and the coordinates of the point P corresponding to the triple (x,y,z) are

$$x \cdot U + y \cdot V + z \cdot W x + y + z = z - y x + y + z, x3 x + y + z$$

We first establish two preliminary results (“lemmas”).

Lemma 1. *The equation of the incircle of* T *is* x2 + y2 + z2 = 2(xy + yz + zx). *Points within the incircle have* x2 + y2 + z2 < 2(xy + yz + zx); *points outside the incircle have* x2 + y2 + z2 > 2(xy + yz + zx).

For, the square of the distance between P and G is

$$PG2 = z - y x + y + z2 + x3 x + y + z - 1 32$$

= 4(x2 + y2 + z2 - xy - yz - zx) 3(x + y + z)2 .

Since the radius of the incircle is 1⁄3, we compute the quantity PG2 - 1⁄3:

PG2 -1 3 = 4(x2 + y2 + z2 - xy - yz - zx) 3(x + y + z)2 -1 3

= x2 + y2 + z2 - 2(xy + yz + zx) (x + y + z)2 .

So x2 + P + z2 - 2(xy + yz + zx) is negative, zero or positive, according to whether P is within, on, or outside the incircle of the triangle. ❑

Lemma 2. *The equation of the circumcircle of* T *is* xy + yz + zx = 0. *Points within the circumcircle have* xy + yz + zx > 0, *and points outside the circumcircle have* xy + yz + zx < 0.

Since the radius of the circumcircle of the triangle is 2⁄3, we compute the quantity PG2 - 4⁄3, using the expression for PG2 from above. After simplification, we get:

PG2 -4 3 = 4(x2 + y2 + z2 - xy - yz - zx) 3(x + y + z)2 -4 3 = -4(xy + yz + zx) (x + y + z)2 .

Therefore, xy + yz + zx is positive, zero or negative, according to whether P is within, on, or outside the circumcircle. ❑

Theorem 13.1 *If P lies within the incircle, the image point P ′ lies closer to the center of the incircle than P, i. e ., GP′ < GP. If P lies on the incircle, then so does P′; i. e the distance from G stays fixed. If P lies outside the incircle, then P ′ too lies outside the incircle. If P lies on the circumcircle, then so does P′.*

Proof. The coordinates of the image point P ′, with x ′,y′,z′ given by equations (13.1), are

(y - z) · (2x - y - z) 4(xy + yz + zx) - (x2 + y2 + z2) , x · (x - 2y + 2z)3 4(xy + yz + zx) - (x2 + y2 + z2).

The square of the distance between P′ and G(i.e.,P′G2) is therefore given by

(y - z) · (2x - y - z) 4(xy + yz + zx) - (x2 + y2 + z2)2 + 1 3 · 2x2 - y2 - z2 - 2xy - 2xz + 4yz 4(xy + yz + zx) - (x2 + y2 + z2)2.

This simplifies (after some effort!) to

P′ G2 = 4 3 · (x2 + y2 + z2 - xy - yz - zx)2 [4(xy + yz + zx) - (x2 + y2 + z2)]2.

Therefore,

P′G2 -1 3 = (x + y + z)2 · (x2 + y2 + +z2 - 2xy - 2yz - 2zx) [4(xy + yz + zx) - (x2 + y2 + z2)]2 . (13.2)

Observe that if x;y : z = 1 : 1 : 1, then P′G = 0, i.e., P′ = G. Further , if P≠G, then

P′G2 PG2 - 1 = 9 · (x2 + y2 + z2 - 2xy - 2yz - 2zx) · (xy + yz + zx) [4(xy + yz + zx) - (x2 + y2 + z2)]2 . (13.3)

From equations (13.2) and (13.3), we see that the following possibilities can arise.

- If P lies within the incircle, then we have

  xy + yz + zx > 0,x2 + y2 + z2 < 2(xy + yz + zx),

  so P′G < PG. Therefore, points within the incircle are pulled *towards* the center of the triangle.

- If P lies on the incircle, then

  x2 + y2 + z2 = 2(xy + yz + zx),

  giving P′G2 - 1⁄3 = 0; therefore, P′ too lies on the incircle.

- If P lies outside the incircle, i.e.,

  x2 + y2 + z2 > 2(xy + yz + zx),

  then P′G2 - 1⁄3 > 0, which means that P′ too lies outside the incircle.

- If P lies outside the in circle but within the circumcircle, then

  xy + yz + zx > 0,x2 + y2 + z2 > 2(xy + yz + zx),

  so P′G > PG. Therefore, points in this region are pushed *away* from the center of the triangle.

- If P lies on the circumcircle, then P′G = PG, therefore, P′ too lies on the circumcircle.

- If P lies outside the circumcircle, then P′G < PG. So such points are once again pulled *towards* the center of the triangle.

Convergence of the iteration. Let the quantity t be defined by

$$t = x2 + y2 + z2\ xy + yz + zx. \qquad (13.4)$$

Note that we cannot have t < 1, as x,y,z are real. We now write the relation

$$P'G2\ PG2 = (x + y + z)2 \cdot (x2 + y2 + z2 - xy - yz - zx)\ [4(xy + yz + zx) - (x2 + y2 + z2)]2 \qquad (13.5)$$

in terms of t. Since

$$(x + y + z)2 = (xy + yz + zx) \cdot (t + 2),$$

$$x2 + y2 + z2 - xy - yz - zx = (xy + yz + zx) \cdot (t - 1),$$

$$4(xy + yz + zx) - (x2 + y2 + z2) = (xy + yz + zx) \cdot (4 - t),$$

we get:

$$P'G2\ PG2 = (t + 2) \cdot (t - 1)\ (4 - t)2\ . \qquad (13.6)$$

Let $g(t) = (t + 2) \cdot (t - 1)/(4 - t)2$. Then $g(2) = 1$; this confirms what we already know: that points on the incircle stay on the incircle.

A table of values of g is of interest:

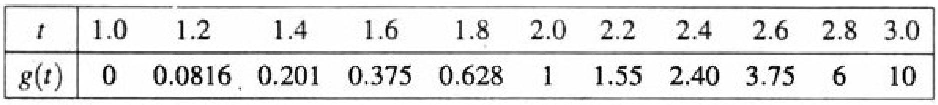

t	1.0	1.2	1.4	1.6	1.8	2.0	2.2	2.4	2.6	2.8	3.0
$g(t)$	0	0.0816	0.201	0.375	0.628	1	1.55	2.40	3.75	6	10

We see that as t drops from 2 to 1, g(t) decreases from 1 to 0; moreover, the slope of g(t) at t = 0 is 0. This explains why the iterates converge to the equilateral shape so rapidly.

Does the starting point always lie within the incircle? We must check that at the start of the iteration, the point P does lie inside the incircle. But this is easy. Recall that at the start we had

$$x : y : z = a2 : b2 : c2.$$

Therefore, to show that $x2 + y2 + z2 < 2(xy + yz + zx)$ we must show that

$$a4 + b4 + c4 < 2(a2b2 + b2c2 + c2a2).$$

In other words, we must show that

$$a4 + b4 + c4 - 2(a2b2 + b2c2 + c2a2) < 0.$$

Rather conveniently for us, the expression on the side factorizes exactly, as

$$-(a + b + c) \cdot (a + b - c) \cdot (b + c - a) \cdot (c + a - b).$$

In this factorization, the quantities $a + b - c, b + c - a$ and $c + a - b$ are all positive (strictly), by virtue of the triangle inequality ("any two sides of a triangle are together greater than the third side"); also, trivially, $a + b + c$ is positive. It follows that the quantity

$$a4 + b4 + c4 - 2(a2b2 + b2c2 + c2a2)$$

is negative, and therefore, the inequality $a4 + b4 + c4 < 2(a2b2 + b2c2 + c2a2)$ is true.

This means that the triple (x,y,z) with which we start the iteration satisfies the relation

$$x2 + y2 + z2 < 2(xy + yz + zx).$$

So, at the start of the iteration, the point P lies strictly within the incircle, and the iteration pulls if towards the center of the triangle, at an ever increasing rate. , ❑

★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★ ★

We may anticipate from the above that the triangles for which convergence is slow are those for which the ratio $(a2 + b2 + c2)/(ab + bc + ca)$ is close to 2; the closer the ratio is to 2, the slower the convergence. This will be the case for an isosceles triangle whose base is small in relation to the equal sides; for example, a triangle with sides $a = 100 = b$ and $c = 1$, with ratio $20001/10200 \approx 1.961$.

Let us try the iteration using this triangle. The initial values of x,y,z are given by:

$$X = 1002\ 1002 + 1002 + 12 \approx 0.499975, y \approx 0.499975, z \approx 0.00005.$$

Starting with this triple, the first few iterates are the following:

(0.49998, 0.49998, 0.00005), (0.49990, 0.49990, 0.00020),

(0.49960, 0.49960, 0.00080), (0.49841, 0.49841, 0.00317),
(0.49380, 0.49380, 0.01240), (0.47731, 0.47731, 0.04538),
(0.43254, 0.43254, 0.13492), (0.36922, 0.36922, 0.26157),
(0.33729, 0.337 29, 0.32542), (0. 33338, 0.33338,0.33324),
(0.33333, 0.33333, 0.33333), (0.33333, 0.33333 ,0.33333).

We see that even with such a skewed starting triple, convergence to the equilateral shape is very rapid.

Remark. As noted above, if the point P lies in the annular region between the incircle and the circumcircle of T, then $P'G > PG$; it is "pushed away" from the center of the triangle. And if P lies outside the circumcircle, then $P'G < PG$; it is "pulled back" towards the center of the triangle.

This has the following remarkable consequence: if the starting point lies out-side the incircle of T, then its iterates cannot converge to any position whatever. Instead, they are forced to "wander around" aimlessly outside the incircle; neither can they enter within the circle, nor can they recede very far from it—a strange situation, indeed !

+,-.+ + +/ ! + , ? 7 A
E?E 1 2"

" +,-.+ + +/ ! + 3 C 2"

+,-.+ + +/ ! + F @ @ 33! "< G! 3! ! G C
@ 33! " ! H 4 ! , 3 !H ! " E ! ? , > " ,22 ! ! I
=" $ 2"

+,-.+ + +/ ! + F @ 33! "< G! 3! ! G C
@ 33! " ! H 4 ! , 3 !H ! " E ! ? < , > " ,22 ! ! I
=" * 2"

( )
! +,-.+ + +/ ! + % A*) A A%* 2"
/3 ! @ 33! "
( )%)) ! J

4.4 MEASURE ZERO

There is measure in all things.—Horace

The measure theory approach to integration, *Lebesgue integration*, is discussed in Section 4.11. It is superior to Riemann integration from several points of view. It makes it possible to integrate otherwise nonintegrable functions, and it is better behaved than Riemann integration with respect to limits.

However, Lebesgue integrals are essentially impossible to compute, unless you treat the integral as a Riemann integral or a limit of such integrals.

We mentioned in Section 4.3 that the criterion for integrability given by Theorem 4.3.9 is not sharp. It is not necessary that a function (bounded and with bounded support) be continuous except on a set of volume 0 to be integrable: it is necessary and sufficient that it be continuous except on a set of *measure* 0.

Measure theory is a big topic, discussed in Appendix A21. Fortunately, the notion of *measure* 0 is much more accessible. It is a subtle notion with some bizarre consequences; it gives us a way, for example, of saying that the rational numbers "don't count". Thus it allows us to use Riemann integration to integrate some quite interesting functions, including one we explore in Example 4.4.7 as a reasonable model for space averages in statistical mechanics.

In Definition 4.4.1 of measure 0, a *box* B in $\mathbb{R}^n$ of sidelength $\delta > 0$ is a cube of the form

$$\{\, \mathbf{x} \in \mathbb{R}^n \mid a_i < x_i < a_i + \delta,\ i = 1, \dots, n \,\}. \qquad 4.4.1$$

There is no requirement that the a_i or δ be dyadic.

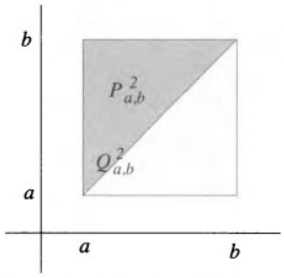

Figure for Exercise 4.3.4

Exercise 4.3.4, part a: Permutations are discussed in Section 4.8. Exercise 4.5.17 gives further applications of Exercise 4.3.4.

4.3.4 For any real numbers $a < b$, let

$$Q^n_{a,b} = \{\, \mathbf{x} \in \mathbb{R}^n \mid a \le x_i \le b \text{ for all } 1 \le i \le n \,\},$$

and let $P^n_{a,b} \subset Q^n_{a,b}$ be the subset where $a \le x_1 \le x_2 \le \cdots \le x_n \le b$. (The case where $n = 2$ is shown in the figure in the margin.)

a. Let $f : \mathbb{R}^n \to \mathbb{R}$ be an integrable function that is symmetric in the sense that

$$f\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = f\begin{pmatrix} x_{\sigma(1)} \\ \vdots \\ x_{\sigma(n)} \end{pmatrix} \quad \text{for any permutation } \sigma \text{ of the symbols } 1, 2, \dots, n.$$

Show that

$$\int_{Q^n_{a,b}} f(\mathbf{x})\, |d^n\mathbf{x}| = n! \int_{P^n_{a,b}} f(\mathbf{x})\, |d^n\mathbf{x}|.$$

b. Let $f : [a, b] \to \mathbb{R}$ be an integrable function. Show that

$$\int_{P^n_{a,b}} f(x_1) f(x_2) \cdots f(x_n) |d^n\mathbf{x}| = \frac{1}{n!}\left(\int_a^b f(x)|dx|\right)^n.$$

Exercise 4.3.5: You may either apply theorems or prove the result directly. If you use theorems, you must show that they apply.

4.3.5 Let P be the region $x^2 < y < 1$. Prove that $\int_P \sin(y^2)\, |dx\, dy|$ exists.

4.4 MEASURE ZERO

There is measure in all things.—Horace

The measure theory approach to integration, *Lebesgue integration*, is discussed in Section 4.11. It is superior to Riemann integration from several points of view. It makes it possible to integrate otherwise nonintegrable functions, and it is better behaved than Riemann integration with respect to limits.

However, Lebesgue integrals are essentially impossible to compute, unless you treat the integral as a Riemann integral or a limit of such integrals.

We mentioned in Section 4.3 that the criterion for integrability given by Theorem 4.3.9 is not sharp. It is not necessary that a function (bounded and with bounded support) be continuous except on a set of volume 0 to be integrable: it is necessary and sufficient that it be continuous except on a set of *measure* 0.

Measure theory is a big topic, discussed in Appendix A21. Fortunately, the notion of *measure* 0 is much more accessible. It is a subtle notion with some bizarre consequences; it gives us a way, for example, of saying that the rational numbers "don't count". Thus it allows us to use Riemann integration to integrate some quite interesting functions, including one we explore in Example 4.4.7 as a reasonable model for space averages in statistical mechanics.

In Definition 4.4.1 of measure 0, a *box* B in $\mathbb{R}^n$ of sidelength $\delta > 0$ is a cube of the form

$$\{\, \mathbf{x} \in \mathbb{R}^n \mid a_i < x_i < a_i + \delta,\ i = 1, \dots, n \,\}. \qquad 4.4.1$$

There is no requirement that the a_i or δ be dyadic.

Definition 4.4.1 (Measure 0). A set $X \subset \mathbb{R}^n$ has *measure* 0 if for every $\epsilon > 0$, there exists an infinite sequence of open boxes B_i such that

$$X \subset \cup B_i \quad \text{and} \quad \sum \mathrm{vol}_n(B_i) \leq \epsilon. \tag{4.4.2}$$

You may have noticed that while our dyadic cubes are semi-open, the boxes in Definition 4.4.1 are open. But the theory could be built just as well with closed cubes and boxes (see Exercise 4.4.1).

The theory also applies to boxes with other shapes: Definition 4.4.1 works with the B_i as arbitrary sets with well-defined volume. Exercise 4.4.2 asks you to show that you can use balls; Exercise 4.9 asks you to show that you can use arbitrary pavable sets.

Definition 4.4.2 (Almost everywhere, almost all). The term *almost everywhere*, abbreviated *a.e.*, means *except on a set of measure 0*. A property true of all **x** except for **x** on a set of measure 0 is said to be true for *almost all* **x**.

The crucial difference between measure and volume is the word "infinite" in Definition 4.4.1. A set with volume 0 can be contained in a *finite* sequence of cubes whose total volume is arbitrarily small. A set with volume 0 necessarily has measure 0, but it is possible for a set to have measure 0 but not to have a defined volume, as shown in Example 4.4.3. Note that when we speak of measure 0 of a subset $X \subset \mathbb{R}^n$, we mean n-dimensional measure 0.

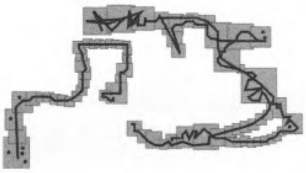

FIGURE 4.4.1.
The set X, shown as a black line, is covered by boxes that overlap.

Remark. We speak of boxes rather than cubes to avoid confusion with the cubes of our dyadic pavings. In dyadic pavings, we considered "families" of cubes all the same size: the cubes at a particular resolution N, and fitting the dyadic grid. The boxes B_i of Definition 4.4.1 get small as i increases, since their total volume is less than ϵ, but it is not necessarily the case that any particular box is smaller than the one immediately preceding it. The boxes can overlap, as illustrated in Figure 4.4.1, and they are not required to square with any particular grid. △

Exercise A21.1 in the appendix explores some of the bizarre properties of the set U_ϵ of Example 4.4.3. This set is a good one to keep in mind while trying to picture the boxes B_i, because it helps us to see that while the sequence $i \mapsto B_i$ is made of $B_1, B_2, \dots$, in order, these boxes may skip around: if B_1 is centered at 1/2, then B_2 is centered at 1/3, B_3 at 2/3, B_4 at 1/4, and so on. Some boxes may be contained in others: for example, depending on the choice of ϵ, the interval centered at 17/32 may be contained in the interval centered at 1/2.

Example 4.4.3 (A set with measure 0, undefined volume). The set of rational numbers in the interval $[0,1]$ has measure 0. You can list them in order, for instance, as 1, 1/2, 1/3, 2/3, 1/4, 2/4, 3/4, 1/5, ... (The list is infinite and includes some numbers more than once.) Center an open interval of length $\epsilon/2$ at 1, an open interval of length $\epsilon/4$ at 1/2, an open interval of length $\epsilon/8$ at 1/3, and so on. Call U_ϵ the union of these intervals. The sum of the lengths of these intervals is $\epsilon\,(1/2 + 1/4 + 1/8 + \cdots) = \epsilon$, and the length of the union is $< \epsilon$, since some of the intervals overlap.

You can place all the rationals in $[0,1]$ in intervals that are infinite in number but whose total length is arbitrarily small! The set U_ϵ thus has measure 0. But it does not have a defined volume because every interval that is not a single point contains both rational and irrational numbers. △

A similar argument proves a stronger result:

Theorem 4.4.4 (A countable union of sets of measure 0 has measure 0). *Let $i \mapsto X_i$ be a sequence of sets of measure 0. Then*

$$X_1 \cup X_2 \cup \dots \quad \textit{is a set of measure 0.}$$

Example 4.4.3 (and, more generally, any countable set) corresponds to the case of Theorem 4.4.4 where each X_i is a single point.

Theorem 4.4.4 is of course not true for arbitrary unions: any set is the union of its points, which all have measure 0. Thus measure theory depends on distinguishing between countable and uncountable infinities, and could only come after Cantor's work (see Section 0.6). Indeed Riemann integration, which doesn't depend on Cantor's work, came before, but Lebesgue integration, which does, comes after.

Proof of Theorem 4.4.4: We can turn the sequences

$$B_{1,i}, \dots, B_{j,i}, \dots$$

into a single sequence by listing the boxes in some order, just as we did for the rational numbers in Example 4.4.3.

For instance, one could list first the boxes where $i+j=2$, then those where $i+j=3$, and so on:

$$B_{1,1}, B_{1,2}, B_{2,1}, B_{1,3}, B_{2,2}, \dots.$$

Proof of Theorem 4.4.4. Since there exist infinite sequences $i \mapsto B_{j,i}$ of boxes (one sequence for each j) such that

$$\begin{aligned} X_1 &\subset B_{1,1} \cup B_{1,2} \cup \dots, \quad \text{and} \quad \sum \operatorname{vol} B_{1,i} \le \frac{\epsilon}{2} \\ X_2 &\subset B_{2,1} \cup B_{2,2} \cup \dots, \quad \text{and} \quad \sum \operatorname{vol} B_{2,i} \le \frac{\epsilon}{4} \\ &\vdots \\ X_j &\subset B_{j,1} \cup B_{j,2} \cup \dots, \quad \text{and} \quad \sum \operatorname{vol} B_{j,i} \le \frac{\epsilon}{2^j}, \end{aligned} \qquad 4.4.3$$

there exists an infinite sequence of open boxes $B_{j,i}$ such that

$$(X_1 \cup X_2 \cup \dots) \subset \bigcup_{j,i} B_{j,i} \quad \text{and} \quad \sum_{i,j} \operatorname{vol} B_{i,j} \le \epsilon. \qquad \square$$

Corollary 4.4.5. *Let $B_R(\mathbf{0})$ be the ball of radius R centered at $\mathbf{0}$. If for all $R \ge 0$ the subset $X \subset \mathbb{R}^n$ satisfies $\operatorname{vol}_n(X \cap B_R(\mathbf{0})) = 0$, then X has measure 0.*

Proof. Since $X = \cup_{m=1}^{\infty} (X \cap B_m(\mathbf{0}))$, it is a countable union of sets of volume 0, hence measure 0. $\square$

Proposition 4.4.6. *Let X be a subspace of $\mathbb{R}^n$ of dimension $k < n$. Then X has measure 0, and any translate of X has measure 0.*

We could also state Proposition 4.4.6 in terms of affine subspaces: an affine subspace $X \in \mathbb{R}^n$ of dimenion $k < n$ has measure 0.

Proof. The first statement follows from Proposition 4.3.5 and Corollary 4.4.5; the second from Proposition 4.1.22. $\square$

The set described in Example 4.4.3 was already encountered in Example 4.3.3, where we found that we could not integrate the function that is 1 at the rational numbers in $[0,1]$ and 0 elsewhere. This function is discontinuous everywhere; in every interval, no matter how small, it jumps from 0 to 1 and from 1 to 0. In Example 4.4.7 we see a function that looks similar but is very different. This function *is* continuous except over a set of measure 0 and thus *is* integrable. It arises in real life (statistical mechanics, at least).

Example 4.4.7: "In lowest terms" traditionally includes the requirement that the denominator be positive, so f is well defined for negative numbers: $-\frac{1}{2}$ can be written only as $\frac{-1}{2}$, not as $\frac{1}{-2}$.

Example 4.4.7 (An integrable function with discontinuities on a set not of volume 0). Consider the function

$$f(x) = \begin{cases} \frac{1}{q} & \text{if } x = \frac{p}{q} \text{ is rational, written in lowest terms with } q > 0 \text{ and } |x| \le 1 \\ 0 & \text{if } x \text{ is irrational or } |x| > 1. \end{cases}$$

This function is integrable although it is discontinuous at values of x for which $f(x) \ne 0$. For instance, $f(3/4) = 1/4$, but arbitrarily close to

3/4 we have irrational numbers, giving $f(x) = 0$. But such values form a set of measure 0. The function is continuous at the irrationals: arbitrarily close to any irrational number x you will find rational numbers p/q, but you can choose a neighborhood of x that includes only rational numbers with arbitrarily large denominators q, so that f will be arbitrarily small in such a neighborhood. $\triangle$

FIGURE 4.4.2.
Ludwig Boltzmann (1844–1906) is credited with the invention of statistical mechanics. A mathematical physicist, he also gave lectures in Vienna on philosophy so popular that no lecture hall large enough could be found.

"The most ordinary things are to philosophy a source of insoluble puzzles," he wrote. "With infinite ingenuity it constructs a concept of space or time and then finds it absolutely impossible that there be objects in this space or that processes occur during this time ... the source of this kind of logic lies in excessive confidence in the so-called laws of thought." In Hegel he found an "unclear thoughtless flow of words"; in Kant, "there were many things that I could grasp so little that ... I almost suspected that he was pulling the reader's leg or was even an imposter."

For the KAM theorem, see "A proof of Kolmogorov's theorem on the conservation of invariant tori," J. Hubbard and Y. Ilyashenko, *Discrete and Continuous Dynamical Systems*, vol. 10, N.1 & 2, Jan. and March 2004, pp. 367–385.

The function of Example 4.4.7 is important because it is a model for functions that show up in an essential way in statistical mechanics (unlike the function of Example 4.3.3, which, as far as we know, is only a pathological example, devised to test the limits of mathematical statements).

In statistical mechanics, one tries to describe a system, typically a gas enclosed in a box, made up of perhaps 10^{25} molecules. Quantities of interest might be temperature, pressure, concentrations of various chemical compounds, etc. A state of the system is a specification of the position and velocity of each molecule (and rotational velocity, vibrational energy, etc., if the molecules have inner structure); to encode this information, one might use a point in some gadzillion-dimensional space.

Mechanics tells us that at the beginning of our experiment, the system is in some state that evolves according to the laws of physics, "exploring" as time proceeds some part of the total state space (and exploring it quite fast relative to our time scale: particles in a gas at room temperature typically travel at several hundred meters per second and undergo millions of collisions per second). The guess underlying thermodynamics is that the quantity one measures, which is really a time average of the quantity as measured along the trajectory of the system, should be nearly equal in the long run to the average over all possible states, called the *space average*. (The "long run" is quite a short run by our clocks.)

This equality of time averages and space averages is called *Boltzmann's ergodic hypothesis*. It is the key hypothesis that connects statistical mechanics to thermodynamics, and physicists believe that it holds in great generality, although there aren't many mechanical systems where it is mathematically proved to be true, and the KAM theorem shows that in some cases where you might expect it to be true, it is not.[4]

What does this have to do with the function f of Example 4.4.7? Even if you believe that a generic time evolution will explore state space fairly evenly, there will always be some trajectories that don't. Consider the (considerably simplified) model of a single particle, moving without friction on a square billiard table, with ordinary bouncing when it hits an edge (the angle of incidence equal to the angle of reflection). Most trajectories (those that start with irrational slope) will evenly fill up the table. But those with rational slopes emphatically will not: they form closed trajectories, which go over and over the same closed path.

---

[4] The KAM theorem, due to Andrei Kolmogorov, Vladimir Arnold, and Jürgen Moser, applies to all frictionless systems of classical mechanics. It shows that under some conditions, these systems are inherently stable. The difference between stability and chaos can hinge on a delicate question in number theory: to what extent irrational numbers can be approximated by rational numbers.

Still, as shown in Figure 4.4.3, these closed paths visit more and more of the table as the denominator of the slope becomes large.

Suppose further that the quantity to be observed is some function f on the table with average 0, which is positive near the center and very negative near the corners, as indicated by the lack of shading near the corners in Figure 4.4.3. Suppose we start our particle at the center of the table but don't specify its direction.

Trajectories through the center of the table, and with slope 0, will have positive time averages, as will trajectories with slope ∞; the trajectories will miss the corners. Similarly, we believe that the average, over time, of each trajectory with rational slope will also be positive. But trajectories with irrational slope will have 0 time averages: given enough time, these trajectories will visit each part of the table equally. And trajectories with rational slopes with large denominators will have time averages close to 0.

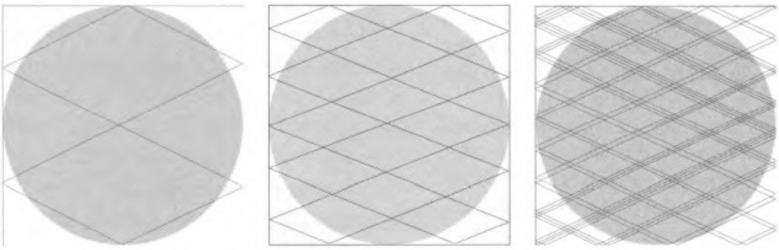

FIGURE 4.4.3. The trajectory with slope 2/5, at center, visits more of the square than the trajectory with slope 1/2, at left. The slope of the trajectory at right closely approximates an irrational number; if allowed to continue, this trajectory would visit nearly every part of the square. (The square would soon become black.)

Because the rational numbers have measure 0, their contribution to the average does not matter; in this case, at least, Boltzmann's ergodic hypothesis seems correct. △

This billiard table example is a caricature of the kind of thing we might run into in statistical mechanics, for instance, having one gas in half a box and another in another half, and removing the partition between them.

Statistical mechanics is an attempt to apply probability theory to large systems of particles, to estimate average quantities, like temperature, pressure, etc., from the laws of mechanics. Thermodynamics, on the other hand, is a completely macroscopic theory. It tries to relate the same macroscopic quantities (temperature, pressure, etc.) on a phenomenological level. Clearly, one hopes to explain thermodynamics by statistical mechanics.

Questions of rational versus irrational numbers are central to understanding the intricate interplay of chaotic and stable behavior exhibited, for example, by the *lakes of Wada*. For more on this topic, see J.H. Hubbard, "What it Means to Understand a Differential Equation," *The College Mathematics Journal*, Vol. 25 (Nov. 5, 1994), 372–384.

Integrability of "almost" continuous functions

We are now ready to state and prove Theorem 4.4.8. This theorem makes it clear that whether or not a function is integrable does not depend on where it is placed on some arbitrary grid.

Theorem 4.4.8 (What functions are integrable). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be bounded with bounded support. Then f is integrable if and only if it is continuous except on a set of measure 0.*

Theorem 4.4.8 is stronger than Theorem 4.3.9, since any set of volume 0 also has measure 0, but not vice versa. It is also stronger than Theorem 4.3.6. It is not stronger than Theorem 4.3.1. Theorems 4.4.8 and 4.3.1 both give an "if and only if" condition for integrability; they are exactly equivalent. But it is often easier to verify that a function is integrable using Theorem 4.4.8.

Proof. We will start with the harder direction: if $f : \mathbb{R}^n \to \mathbb{R}$, bounded with bounded support, is continuous almost everywhere, then it is integrable. We will use the criterion for integrability given by Theorem 4.3.1; thus we want to prove that for all $\epsilon > 0$ there exists N such that the cubes $C \in \mathcal{D}_N$ over which $\text{osc}_C(f) > \epsilon$ have a combined volume less than ϵ. (Recall from Definition 4.1.4 that $\text{osc}_{B_i}(f)$ is the oscillation of f over B_i.)

Denote by Δ the set of points where f is not continuous, and choose $\epsilon > 0$ (which will remain fixed for the duration of the proof). By Definition 4.4.1 of measure 0, there exists an infinite sequence $i \mapsto B_i$ of open boxes such that

$$\Delta \subset \cup B_i \quad \text{and} \quad \sum \text{vol}_n B_i < \epsilon. \qquad 4.4.4$$

We will choose our boxes so that no box is contained in any other.

The proof is fairly involved. First, we want to get rid of infinity.

We prune the list of boxes by throwing away any box that is contained in an earlier one. We could prove our result without pruning the list, but it would make the argument more cumbersome.

Lemma 4.4.9. *Let $i \mapsto B_i$ be a sequence satisfying $\Delta \subset \cup B_i$ and $\sum \text{vol}_n B_i < \epsilon$. Then $\text{osc}_{B_i}(f) > \epsilon$ on only finitely many boxes B_i.*

Denote such boxes by B_{i_j}; denote the union of the B_{i_j} by L. (In Figure 4.4.4, the B_i are the lightly shaded boxes; the B_{i_j} are slightly darker.)

Proof of Lemma 4.4.9. Assume the lemma is false. Then there exist an infinite subsequence of boxes $j \mapsto B_{i_j}$ and two infinite sequences of points, $j \mapsto \mathbf{x}_j, j \mapsto \mathbf{y}_j$ in B_{i_j}, such that $|f(\mathbf{x}_j) - f(\mathbf{y}_j)| > \epsilon$.

The sequence $j \mapsto \mathbf{x}_j$ is bounded, since the support of f is bounded and $\mathbf{x}_j$ is in the support of f. So (by Theorem 1.6.3) it has a convergent subsequence $\mathbf{x}_{j_k}$ converging to some point $\mathbf{p}$. Since $|\mathbf{x}_j - \mathbf{y}_j| \to 0$ as $j \to \infty$, the subsequence $\mathbf{y}_{j_k}$ also converges to $\mathbf{p}$.

The function f is certainly not continuous at $\mathbf{p}$, so $\mathbf{p} \in \Delta$ and by formula 4.4.4, $\mathbf{p}$ has to be in a particular box; we will call it $B_\mathbf{p}$. (Since the boxes can overlap, it could be in more than one, but we just need one.) Since $\mathbf{x}_{j_k}$ and $\mathbf{y}_{j_k}$ converge to $\mathbf{p}$, and since the B_{i_j} get small as j gets big (their total volume being less than ϵ), then all $B_{i_{j_k}}$ after a certain point are contained in $B_\mathbf{p}$. But this contradicts our assumption that we had pruned our list of B_i so that no one box was contained in any other. Therefore, Lemma 4.4.9 is correct: there are only finitely many B_i on which $\text{osc}_{B_i} f > \epsilon$. $\square$ Lemma 4.4.9

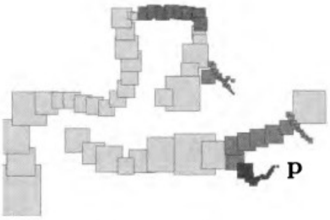

FIGURE 4.4.4.

The collection of boxes B_i covering Δ is lightly shaded. The boxes B_{i_j} with osc $> \epsilon$ are shaded slightly darker. A convergent subsequence of those is shaded darker yet. The key to Lemma 4.4.9 is that the point to which they converge must belong to some box B_i.

Lemma 4.4.10. *There exists N such that if $C \in \mathcal{D}_N(\mathbb{R}^n)$ and $\text{osc}_C f > \epsilon$, then $C \subset L$.*

If we prove this, we will be finished, because by Theorem 4.3.1, a bounded function with bounded support is integrable if there exists N at which the total volume of cubes with osc $> \epsilon$ is less than ϵ. We know that L is a finite set of B_i, and that the B_i have total volume $< \epsilon$.

Proof of Lemma 4.4.10. We will again argue by contradiction. Suppose the lemma is false. Then for every N, there exists a C_N not a subset of L such that $\text{osc}_{C_N} f > \epsilon$. In other words,

$$\exists \text{ points } \mathbf{x}_N, \mathbf{y}_N, \mathbf{z}_N \text{ in } C_N, \text{ with } \mathbf{z}_N \notin L \text{ and } |f(\mathbf{x}_N) - f(\mathbf{y}_N)| > \epsilon. \qquad 4.4.5$$

Since $\mathbf{x}_N, \mathbf{y}_N$, and $\mathbf{z}_N$ are infinite sequences (for $N = 1, 2, \dots$), there exist convergent subsequences $\mathbf{x}_{N_i}$, $\mathbf{y}_{N_i}$, and $\mathbf{z}_{N_i}$, all converging to the same point, which we will call $\mathbf{q}$.

What do we know about $\mathbf{q}$?

- $\mathbf{q} \in \Delta$: i.e., it is a discontinuity of f – no matter how close $\mathbf{x}_{N_i}$ and $\mathbf{y}_{N_i}$ get to $\mathbf{q}$, we have $|f(\mathbf{x}_{N_i}) - f(\mathbf{y}_{N_i})| > \epsilon$. Therefore (since all the discontinuities of f are contained in the B_i), it is in some box B_i, which we'll call $B_{\mathbf{q}}$.
- $\mathbf{q} \notin L$. (The set L is open, since it is a union of open sets; see Exercise 1.5.3. Therefore its complement is closed; since no point of the sequence $\mathbf{z}_{N_i}$ is in L, its limit, $\mathbf{q}$, is not in L either.)

Since $\mathbf{q} \in B_{\mathbf{q}}$, and $\mathbf{q} \notin L$, we know that $B_{\mathbf{q}}$ is not one of the boxes with osc $> \epsilon$. But $\mathbf{x}_{N_i}$ and $\mathbf{y}_{N_i}$ are in $B_{\mathbf{q}}$ for N_i large enough, so that $\text{osc}_{B_{\mathbf{q}}} f < \epsilon$ contradicts $|f(\mathbf{x}_{N_i}) - f(\mathbf{y}_{N_i})| > \epsilon$. □ Lemma 4.4.10

You may ask, how do we know they converge to the same point? Because $\mathbf{x}_{N_i}$, $\mathbf{y}_{N_i}$, and $\mathbf{z}_{N_i}$ are all in the same cube, which is shrinking to a point as $N \to \infty$.

Recall that Δ is the set of points where f is not continuous, and L is the finite union of the boxes B_i on which $\text{osc}_{B_i}(f) > \epsilon$.

The boxes B_i making up L are in $\mathbb{R}^n$, so the complement of L is $\mathbb{R}^n - L$. Recall (Definition 1.5.4) that a closed set $C \subset \mathbb{R}^n$ is a set whose complement $\mathbb{R}^n - C$ is open.

The proof of Lemma 4.4.10 means that we have proved Theorem 4.4.8 in one direction: if a bounded function with bounded support is continuous except on a set of measure 0, then it is integrable.

Now we need to prove the other direction: if a function $f : \mathbb{R}^n \to \mathbb{R}$, bounded and with bounded support, is integrable, then it is continuous almost everywhere. This is easier, but the fact that we chose our dyadic cubes half-open and our boxes open introduces a little complication.

Definition 4.4.1 specifies that the boxes B_i are open. Equation 4.1.12 defining dyadic cubes shows that they are half-open: x_i is greater than or equal to one amount, but strictly less than another:

$$\frac{k_i}{2^N} \le x_i < \frac{k_i + 1}{2^N}.$$

Choose $\delta > 0$. Since f is integrable, we know (Theorem 4.3.1) that for any $\epsilon > 0$, there exists N such that the finite union of cubes

$$\{\, C \in \mathcal{D}_N(\mathbb{R}^n) \mid \text{osc}_C(f) > \epsilon \,\} \qquad 4.4.6$$

has total volume less than ϵ. So we can choose N_1 such that if $\mathcal{C}_{N_1}$ denotes the finite collection of cubes $C \in \mathcal{D}_{N_1}(\mathbb{R}^n)$ with $\text{osc}_C f > \delta/4$, these cubes have total volume less than $\delta/4$.

Now let $\mathcal{C}_{N_2}$ be the finite collection of cubes with $\text{osc}_C f > \delta/8$; these cubes have total volume less than $\delta/8$. Continue with $\delta/16, \dots$. Finally, consider the infinite sequence of open boxes $B_1, B_2, \dots$ obtained by listing first the interiors of the cubes in $\mathcal{C}_{N_1}$, then those of the cubes in $\mathcal{C}_{N_2}$, etc.

This almost solves our problem: the sums of the volumes of the boxes in our sequence is at most $\delta/4 + \delta/8 + \cdots = \delta/2$. The problem is that discontinuities on the boundary of dyadic cubes may go undetected by oscillation on dyadic cubes: the value of the function over one cube could be 0, and the value over an adjacent cube could be 1; in each case the oscillation over the cube would be 0, but the function would be discontinuous at points on the border between the two cubes.

This is easily dealt with: the union of all the boundaries of all dyadic cubes (over all N) has measure 0. To see this, denote by $\delta\mathcal{D}_N(\mathbb{R}^n)$ the union of the boundaries of the dyadic cubes of $\mathcal{D}_N$. Then

1. For each N, the boundary $\delta\mathcal{D}_N(\mathbb{R}^n)$ is a countable union of translates of subspaces of dimension $n - 1$, hence has measure 0 by Theorem 4.4.4 and Corollary 4.4.5.

2. The set $\cup_{N=1}^{\infty} \delta \mathcal{D}_N(\mathbb{R}^n)$ has measure 0, since it is a countable union of sets of measure 0. □ Theorem 4.4.8.

Corollary 4.4.11. *Let f and g be integrable functions on $\mathbb{R}^n$ such that $f \geq g$ and $\int f(\mathbf{x})|d^n\mathbf{x}| = \int g(\mathbf{x})|d^n\mathbf{x}|$. Then*

$$\{\,\mathbf{x} \mid f(\mathbf{x}) \neq g(\mathbf{x})\,\} \text{ has measure 0.}$$

Proof. The function $f - g$ is integrable, hence continuous almost everywhere. Thus if $f > g$ on a set not of measure 0, there exists $\mathbf{x}_0$ such that $f(\mathbf{x}_0) > g(\mathbf{x}_0)$ and $f - g$ is continuous at $\mathbf{x}_0$. Then there exist $\epsilon > 0$ and $r > 0$ such that $f(\mathbf{x}) - g(\mathbf{x}) > \epsilon$ when $|\mathbf{x} - \mathbf{x}_0| \leq r$. We then have

$$\int_{\mathbb{R}^n} (f-g)(\mathbf{x})|d^n\mathbf{x}| \geq \int_{B_r(\mathbf{x}_0)} (f-g)(\mathbf{x})|d^n\mathbf{x}| \geq \epsilon \ \mathrm{vol}_n(B_r(\mathbf{x}_0)) > 0, \quad 4.4.7$$

contradicting our assumption that $\int f(\mathbf{x})|d^n\mathbf{x}| = \int g(\mathbf{x})|d^n\mathbf{x}|$. □

In Proposition 4.1.14 we saw that the sum of two integrable functions is integrable, and that the product of a real number and an integrable function is integrable. It is also true that if two functions $f, g : \mathbb{R}^n \to \mathbb{R}$ are both integrable, their product is integrable (although the integral of the product is in general *not* the product of the integrals).

Corollary 4.4.12: The integral of the product is usually *not* the product of the integrals. However (Proposition 4.1.16), if f and g are functions of *different* variables, then the integral of the product over the product of the domains *is* the product of the integrals.

Corollary 4.4.12 (Product of integrable functions is integrable). *If two functions $f, g : \mathbb{R}^n \to \mathbb{R}$ are both integrable, then fg is also integrable.*

Proof. If f is continuous except on a set X_f of measure 0, and g is continuous except on a set X_g of measure 0, then (Proposition 1.5.29) fg is continuous except on a subset of $X_f \cup X_g$. By Theorem 4.4.4, this subset has measure 0. Thus by Theorem 4.4.8, fg is integrable. □

Corollary 4.4.13 follows from Theorem 4.4.8.

Corollary 4.4.13 (Integration is translation invariant). *If a function $f : \mathbb{R}^n \to \mathbb{R}$ is integrable, then for any $\vec{\mathbf{v}} \in \mathbb{R}^n$, the function $\mathbf{x} \mapsto f(\mathbf{x} - \vec{\mathbf{v}})$ is integrable.*

EXERCISES FOR SECTION 4.4

4.4.1 Use Definition 4.1.17 of n-dimensional volume to show that the same sets have measure 0 whether you define measure 0 using open boxes or closed boxes.

4.4.2 Show that $X \subset \mathbb{R}^n$ has measure 0 if and only if for any $\epsilon > 0$ there exists an infinite sequence of balls

$$B_i = \{\,\mathbf{x} \in \mathbb{R}^n \mid |\mathbf{x} - \mathbf{a}_i| < r_i\,\} \text{ with } \sum_{i=1}^{\infty} r_i^n < \epsilon \text{ such that } X \subset \cup_{i=1}^{\infty} B_i.$$

4.4.3 Prove that any compact subset of $\mathbb{R}^n$ of measure 0 has volume 0.

Hint for Exercise 4.4.3: Use the Heine Borel theorem in Appendix A3.

****4.4.4** Consider the subset $U \subset [0,1]$ that is the union of the open intervals $\left(\frac{p}{q} - \frac{C}{q^3}, \frac{p}{q} + \frac{C}{q^3}\right)$ for all rational numbers $p/q \in [0,1]$. Show that for $C > 0$ sufficiently small, U is not pavable. What would happen if the 3 were replaced by 2? (This is really hard.)

4.4.5 Show that Corollary 4.4.11 is false if you replace "measure" by "volume".

4.4.6 Show that if the boundary of $A \subset \mathbb{R}^n$ has a well-defined n-dimensional volume, that volume is 0: $\text{vol}_n\, \partial A = 0$. *Hint*: Use Theorem 4.8.8.

4.4.7 Show that the set of numbers which in base 10 can be written using only the digits 1–6 has measure 0.

4.5 FUBINI'S THEOREM AND ITERATED INTEGRALS

FIGURE 4.5.1.

The Italian mathematician Guido Fubini (1879–1943) was forced to retire when the Manifesto of Fascist Racism made anti-Semitism official policy of Italy in 1938. He joined the Institute for Advanced Study in Princeton the following year.

We now know – in principle, at least – how to determine whether a function is integrable. Assuming it is, how do we integrate it? Fubini's theorem allows us to compute multiple integrals by hand, or at least reduce them to the computation of one-dimensional integrals. It asserts that if $f : \mathbb{R}^n \to \mathbb{R}$ is integrable, then

$$\int_{\mathbb{R}^n} f(\mathbf{x})\, |d^n\mathbf{x}| = \int_{-\infty}^{\infty} \left(\dots \left(\int_{-\infty}^{\infty} f \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} dx_1 \right) \dots \right) dx_n. \qquad 4.5.1$$

That is, first we hold the variables $x_2, \dots, x_n$ constant and integrate with respect to x_1; then we integrate the resulting (no doubt complicated) function with respect to x_2, and so on. Note that the expression on the left side of equation 4.5.1 doesn't specify the order in which the variables are taken, so the iterated integral on the right can be written in any order: we could integrate first with respect to x_n, or any other variable. This is important for both theoretical and computational uses of Fubini's theorem. Note also that in one dimension, equation 4.5.1 becomes equation 4.1.30.

Remark. Equation 4.5.1 is not quite correct, because some of the functions in parentheses on the right side of equation 4.5.1 may not be integrable; this problem is discussed in Appendix A17. We state Fubini's theorem correctly at the end of this section. For now, assume that we are in the (common) situation where equation 4.5.1 works. △

In practice, the main difficulty in setting up a multiple integral as an iterated one-dimensional integral is dealing with the "boundary" of the region over which we wish to integrate the function. We tried to sweep difficulties like the fractal coastline of Britain under the rug by choosing to integrate over all of $\mathbb{R}^n$, but of course those difficulties are still there. This is where we have to come to terms with them: we have to figure out the upper and lower limits of the integrals.

If the domain of integration looks like the coastline of Britain, it is not obvious how to go about this. For domains of integration bounded by smooth curves and surfaces, formulas exist in many cases that are of interest (particularly during calculus exams), but this is still the part that gives students the most trouble. So before computing any multiple integrals, let's see how to set them up. While a multiple integral is *computed* from inside out – first with respect to the variable in the inner parentheses – we recommend *setting up* the problem from outside in, as shown in Examples 4.5.1 and 4.5.2.

By "integrate over the triangle" we mean that we imagine that the function f is defined by some formula inside the triangle, and by $f = 0$ outside the triangle.

Example 4.5.1 (Setting up multiple integrals: an easy example). Suppose we want to integrate a function $f : \mathbb{R}^2 \to \mathbb{R}$ over the triangle

$$T = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2 \,\middle|\, 0 \le 2x \le y \le 2 \right\} \qquad 4.5.2$$

shown in Figure 4.5.2. This triangle is the intersection of the three regions (here, halfplanes) defined by the inequalities $0 \le x,\ 2x \le y,$ and $y \le 2$.

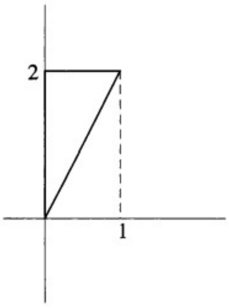

FIGURE 4.5.2.
The triangle defined by equation 4.5.2.

Say we want to integrate first with respect to y. We set up the integral as follows, temporarily omitting the limits of integration:

$$\int_{\mathbb{R}^2} f \begin{pmatrix} x \\ y \end{pmatrix} |dx\, dy| = \int \left(\int f\, dy \right) dx. \qquad 4.5.3$$

(We just write f for the function, as we don't want to complicate issues by specifying a particular function.) Starting with the outer integral – thinking first about x – we hold a pencil parallel to the y-axis and roll it over the triangle from left to right. We see that the triangle (the domain of integration) starts at $x = 0$ and ends at $x = 1$, so we write in those limits:

$$\int_0^1 \left(\int f\, dy \right) dx. \qquad 4.5.4$$

Now think about what happened as you rolled the pencil from $x = 0$ to $x = 1$: for each value of x, what values of y were in the triangle? In this simple case, for every value of x, the pencil intersects the triangle in an interval going from the hypotenuse to $y = 2$. So the upper value is $y = 2$, and the lower value is $y = 2x$, the line on which the hypotenuse lies. Thus we have

$$\int_0^1 \left(\int_{2x}^2 f\, dy \right) dx. \qquad 4.5.5$$

If we want to start by integrating f with respect to x, we write

$$\int_{\mathbb{R}^2} f \begin{pmatrix} x \\ y \end{pmatrix} |dx\, dy| = \int \left(\int f\, dx \right) dy; \qquad 4.5.6$$

starting with the outer integral, we hold our pencil parallel to the x-axis and roll it from the bottom of the triangle to the top, from $y = 0$ to $y = 2$. As we roll the pencil, we ask what are the lower and upper values of x for each value of y. The lower value is always $x = 0$, and the upper value is set by the hypotenuse, but we express it as $x = y/2$. This gives

$$\int_0^2 \left(\int_0^{\frac{y}{2}} f\,dx \right) dy. \qquad 4.5.7$$

Now suppose we are integrating over only part of the triangle, as shown in Figure 4.5.3. What limits do we put in the expression $\int(\int f\,dy)\,dx$? Try it yourself before checking the answer in the footnote.[5] △

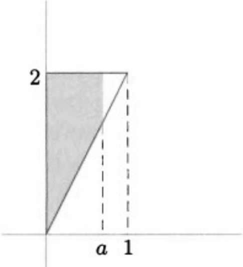

FIGURE 4.5.3.

The shaded area represents a truncated part of the triangle of Figure 4.5.2.

Example 4.5.2 (Setting up multiple integrals: a somewhat harder example). Now let's integrate an unspecified function $f:\mathbb{R}^2 \to \mathbb{R}$ over the area bordered on the top by the parabolas $y = x^2$ and $y = (x-2)^2$ and on the bottom by the straight lines $y = -x$ and $y = x - 2$, as shown in Figure 4.5.4.

Let's start again by sweeping our pencil from left to right, which corresponds to the outer integral being with respect to x. The limits for the outer integral are clearly $x = 0$ and $x = 2$, giving

$$\int_0^2 \left(\int f\,dy \right) dx. \qquad 4.5.8$$

As we sweep our pencil from left to right, we see that the lower limit for y is set by the straight line $y = -x$, and the upper limit by the parabola $y = x^2$, so we are tempted to write $\int_0^2 \left(\int_{-x}^{x^2} f\,dy \right) dx$, but at $x = 1$, we have a problem. The lower limit is now set by the straight line $y = x - 2$, and the upper limit by the parabola $y = (x-2)^2$. How can we express this? Try it yourself before looking at the answer in the footnote.[6] △

Exercise 4.5.3 asks you to set up the multiple integral for Example 4.5.2 when the outer integral is with respect to y. The answer will be a sum of integrals.

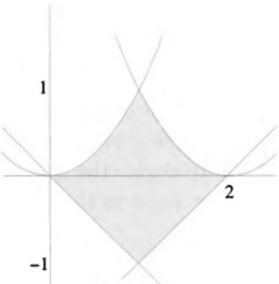

FIGURE 4.5.4.

The region of integration for Example 4.5.2.

Example 4.5.3 (Setting up a multiple integral in $\mathbb{R}^3$). Already in $\mathbb{R}^3$ this kind of visualization becomes much harder. Suppose we want to integrate a function over the pyramid P shown in Figure 4.5.5 and given by the formula

$$P = \left\{ \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{R}^3 \mid 0 \le x;\ 0 \le y;\ 0 \le z;\ x + y + z \le 1 \right\}. \qquad 4.5.9$$

---

[5] When the domain of integration is the truncated triangle in Figure 4.5.3, the integral is written

$$\int_0^a \left(\int_{2x}^2 f\,dy \right) dx.$$

The other direction is harder; we will return to it in Exercise 4.5.2.

[6] We need to break up this integral into a sum of integrals:

$$\int_0^1 \left(\int_{-x}^{x^2} f\,dy \right) dx + \int_1^2 \left(\int_{x-2}^{(x-2)^2} f\,dy \right) dx.$$

Exercise 4.5.1 asks you to justify ignoring that we have counted the line $x = 1$ twice.

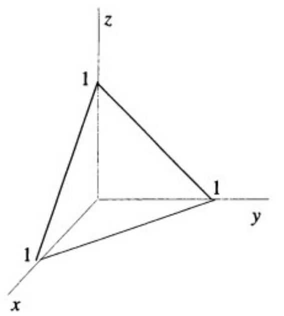

FIGURE 4.5.5.
Pyramid for Example 4.5.3.

We want to figure out the limits of integration for the multiple integral

$$\int_P f\begin{pmatrix}x\\y\\z\end{pmatrix}|dx\,dy\,dz| = \int\left(\int\left(\int f\,dx\right)dy\right)dz. \qquad 4.5.10$$

There are six ways to apply Fubini's theorem, which in this case because of symmetries will result in the same expressions with the variables permuted. Let us vary z first. For instance, we can lift a piece of paper and see how it intersects the pyramid at various heights. Clearly the paper will only intersect the pyramid when its height is between 0 and 1. This gives $\int_0^1 (\quad)dz$, where the space needs be filled in by the double integral of f over the part of the pyramid P at height z, pictured at left in Figure 4.5.6, and again at right, this time drawn flat.

Note that when setting up a multiple integral, the upper and lower limits for the outer integral are just numbers. Then (going from outer to inner) the limits for the next integral may depend on the variable of the outer integral. See, for instance the upper limit $y/2$ in integral 4.5.7, and the upper limit $1 - z$ in integral 4.5.11. The limits for the next integral may depend on the variables of the two outer integrals; see the upper limit $1 - y - z$ in formula 4.5.12.

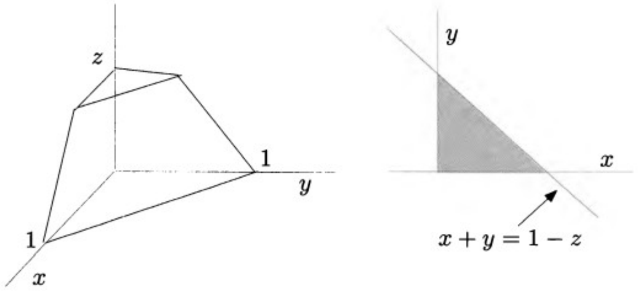

FIGURE 4.5.6. LEFT: The pyramid of Figure 4.5.5, truncated at height z. RIGHT: The plane at height z, put flat.

This time we are integrating over a triangle (which depends on z); see the right side of Figure 4.5.6. So we are in the situation of Example 4.5.1. Let's vary y first. Holding a pencil parallel to the x-axis and rolling it over the triangle from bottom to top, we see that the relevant y-values are between 0 and $1 - z$ (the upper value is the largest value of y on the part of the line $y = 1 - z - x$ that is in the triangle). So we write

$$\int_0^1\left(\int_0^{1-z}(\qquad\qquad)dy\right)dz, \qquad 4.5.11$$

where now the blank represents the integral over part of the horizontal line segment at height z and "depth" y. These x-values are between 0 and $1 - z - y$, so finally the integral is

$$\int_0^1\left(\int_0^{1-z}\left(\int_0^{1-y-z} f\begin{pmatrix}x\\y\\z\end{pmatrix}dx\right)dy\right)dz. \quad \triangle \qquad 4.5.12$$

Formula 4.5.12: When we integrate with regard to x, why is the upper value $1-y-z$, when the upper value with regard to y is $1-z$, not $1-z-x$? Since the inner integral, first to be computed, is with regard to x, we must consider all values of y and z. But when we integrate with regard to y, we have already integrated with regard to x. Clearly the value of x for which $y = 1 - z - x$ is largest is $x = 0$, giving upper value $y = 1 - z$.

Example 4.5.4 (A less visual example). In the preceding examples, we could deduce the upper and lower limits from pictures of the region over which we want to integrate. Here is a case where the picture doesn't give the answer. Let us set up a multiple integral over the ellipse E defined by

$$E = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \mid x^2 + xy + y^2 \le 1 \right\}. \qquad 4.5.13$$

The boundary of E is defined by $y^2 + xy + x^2 - 1 = 0$, so on the boundary,

$$y = \frac{-x \pm \sqrt{x^2 - 4(x^2-1)}}{2} = \frac{-x \pm \sqrt{4 - 3x^2}}{2}. \qquad 4.5.14$$

Such values exist only if $4 - 3x^2 \ge 0$, i.e., $|x| \le 2\sqrt{3}/3$. So to set up the integral of a function f over E we write

Here, x and y play symmetric roles, so to integrate first with respect to x, we would just replace every y by x and vice versa.

$$\int_{-2\sqrt{3}/3}^{2\sqrt{3}/3} \left(\int_{\frac{-x-\sqrt{4-3x^2}}{2}}^{\frac{-x+\sqrt{4-3x^2}}{2}} f\begin{pmatrix} x \\ y \end{pmatrix} dy \right) dx. \quad \triangle \qquad 4.5.15$$

Now let's actually compute a few multiple integrals.

Example 4.5.5 (Computing a multiple integral). Let $f:\mathbb{R}^2 \to \mathbb{R}$ be the function $f\begin{pmatrix} x \\ y \end{pmatrix} = xy\,\mathbf{1}_S\begin{pmatrix} x \\ y \end{pmatrix}$, where S is the unit square, as shown in Figure 4.5.7. Then

$$\begin{aligned} \int_{\mathbb{R}^2} f\begin{pmatrix} x \\ y \end{pmatrix} |dx\,dy| &= \int_0^1 \left(\int_0^1 xy\,dx \right) dy \\ &= \int_0^1 \left[\frac{x^2y}{2} \right]_{x=0}^{x=1} dy = \int_0^1 \frac{y}{2}\,dy = \frac{1}{4}. \quad \triangle \end{aligned} \qquad 4.5.16$$

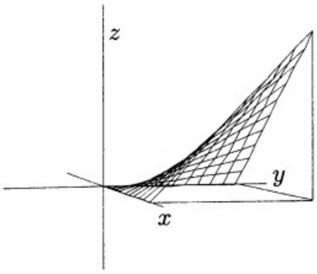

FIGURE 4.5.7.

The integral in equation 4.5.16 is 1/4: the volume under the surface defined by $f\begin{pmatrix} x \\ y \end{pmatrix} = xy$ and above the unit square is 1/4.

In Example 4.5.5 it is clear that we could have taken the integral in the opposite order and found the same result, since our function is $f\begin{pmatrix} x \\ y \end{pmatrix} = xy$. Fubini's theorem says that this is always true as long as the functions involved are integrable. This fact is useful; sometimes a multiple integral can be computed in elementary terms when written in one direction, but not in the other, as you will see in Example 4.5.6. It may also be easier to determine the limits of integration if the problem is set up in one direction rather than another, as we saw in the case of the truncated triangle shown in Figure 4.5.3.

Example 4.5.6 (Choose the easy direction). Let us integrate the function e^{-y^2} over the triangle shown in Figure 4.5.8:

$$T = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2 \mid 0 \le x \le y \le 1 \right\}. \qquad 4.5.17$$

Fubini's theorem gives us two ways to write this integral as an iterated one-dimensional integral:

$$\int_0^1 \left(\int_x^1 e^{-y^2} dy \right) dx \qquad \text{and} \qquad \int_0^1 \left(\int_0^y e^{-y^2} dx \right) dy. \qquad 4.5.18$$

The first cannot be computed in elementary terms, since e^{-y^2} does not have an elementary antiderivative. But the second can:

$$\int_0^1 \left(\int_0^y e^{-y^2} dx \right) dy = \int_0^1 ye^{-y^2}\, dy = -\frac{1}{2}\left[e^{-y^2}\right]_0^1 = \frac{1}{2}\left(1 - \frac{1}{e}\right). \qquad 4.5.19$$

(The y in ye^{-y^2} makes it possible to integrate in elementary terms.) △

Older textbooks contain many examples of such computational miracles. We believe it is sounder to take a serious interest in the numerical theory and go lightly over computational tricks, which do not work in any great generality in any case.

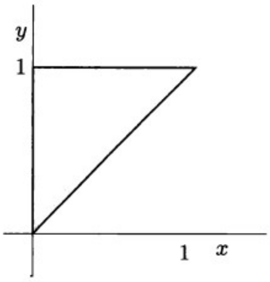

FIGURE 4.5.8.
The triangle

$$\left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2 \mid 0 \le x \le y \le 1 \right\}$$

of Example 4.5.6.

Example 4.5.7: The ball $B_1^n(\mathbf{0})$ is the ball of radius 1 in $\mathbb{R}^n$, centered at the origin; the ball

$$B^{n-1}_{\sqrt{1-x_n^2}}(\mathbf{0})$$

is the ball of radius $\sqrt{1-x_n^2}$ in $\mathbb{R}^{n-1}$, still centered at the origin.

In the first line of equation 4.5.20 we imagine slicing the n-dimensional ball horizontally and computing the $(n-1)$-dimensional volume of each slice.

Example 4.5.7 (Volume of a ball in $\mathbb{R}^n$). Let $B_R^n(\mathbf{0})$ be the ball of radius R in $\mathbb{R}^n$, centered at $\mathbf{0}$, and let $b_n(R)$ be its volume. By Proposition 4.1.24, $b_n(R) = R^n b_n(1)$. We will denote by $\beta_n = b_n(1)$ the volume of the unit ball. By Fubini's theorem,

$$\begin{aligned}
\underbrace{\beta_n}_{\substack{\text{vol. of}\\ \text{unit ball in } \mathbb{R}^n}} &= \int_{B_1^n(\mathbf{0})} |d^n\mathbf{x}| = \int_{-1}^{1} \overbrace{\left(\int_{B^{n-1}_{\sqrt{1-x_n^2}}(\mathbf{0})} |d^{n-1}\mathbf{x}| \right)}^{\substack{(n-1)\text{-dimensional vol.}\\ \text{of one slice of } B_1^n(\mathbf{0})}} dx_n \\
&= \int_{-1}^{1} \underbrace{b_{n-1}\left(\sqrt{1-x_n^2}\right)}_{\substack{\text{vol. ball of radius}\\ r=\sqrt{1-x_n^2} \text{ in } \mathbb{R}^{n-1}}} dx_n = \int_{-1}^{1} \underbrace{(1-x_n^2)^{\frac{n-1}{2}}}_{r^{n-1}} \underbrace{\beta_{n-1}}_{\substack{\text{vol. ball of}\\ \text{radius 1}\\ \text{in } \mathbb{R}^{n-1}}} dx_n \\
&= \underbrace{\beta_{n-1}}_{\substack{\text{vol. of unit}\\ \text{ball in } \mathbb{R}^{n-1}}} \int_{-1}^{1} (1-x_n^2)^{\frac{n-1}{2}}\, dx_n. \qquad 4.5.20
\end{aligned}$$

This reduces the computation of β_n to computing the integral

$$c_n = \int_{-1}^{1} (1-t^2)^{\frac{n-1}{2}}\, dt. \qquad 4.5.21$$

This is a standard tricky problem from one-variable calculus: Exercise 4.5.4 asks you to show that $c_0 = \pi$ and $c_1 = 2$ and that for $n \ge 2$,

$$c_n = \frac{n-1}{n} c_{n-2}. \qquad 4.5.22$$

This allows us to make Table 4.5.9. It is easy to continue the table. (What is β_6? Check below.[7]) △

If you enjoy inductive proofs, you might try Exercise 4.5.5, which asks you to find a formula for β_i for all i. Exercise 4.5.6 asks you to show that as n increases, the volume of the n-dimensional unit ball becomes a smaller and smaller proportion of the smallest n-dimensional cube that contains it.

---

[7] $c_6 = \frac{5}{6}c_4 = \frac{15\pi}{48}$, so $\beta_6 = c_6\beta_5 = \frac{\pi^3}{6}$.

n	$c_n = \frac{n-1}{n} c_{n-2}$	Volume of ball $\beta_n = c_n \beta_{n-1}$
0	π	
1	2	2
2	$\frac{\pi}{2}$	π
3	$\frac{4}{3}$	$\frac{4\pi}{3}$
4	$\frac{3\pi}{8}$	$\frac{\pi^2}{2}$
5	$\frac{16}{15}$	$\frac{8\pi^2}{15}$

TABLE 4.5.9. Computing the volume of the unit ball in $\mathbb{R}^1$ through $\mathbb{R}^5$.

Computing probabilities using integrals

We saw in Section 4.2 that integrals are useful in computing probabilities.

Example 4.5.8 (Using Fubini to compute probabilities). Recall (equation 4.2.12) that the expectation (expected value) of a random variable f is given by $E(f) = \int_S f(s)\mu(s)\,|ds|$, where the function μ weights outcomes according to how likely they are, and S is the sample space. If we choose two points x, y in the interval $[0, 1]$, without privileging any part of that interval, what is the expected value of the function $f\begin{pmatrix} x \\ y \end{pmatrix} = (x-y)^2$?

"Without privileging any part of that interval" means that $\mu = 1$. Thus

$$\begin{aligned} E\big((x-y)^2\big) &= \int_0^1 \int_0^1 (x-y)^2 dx\,dy = \int_0^1 \left(\int_0^1 x^2 - 2xy + y^2\,dx \right) dy \\ &= \int_0^1 \left[\frac{x^3}{3} - x^2 y + y^2 x \right]_0^1 dy = \int_0^1 \frac{1}{3} - y + y^2\,dy \\ &= \left[\frac{y}{3} - \frac{y^2}{2} + \frac{y^3}{3} \right]_0^1 = \frac{1}{6}. \end{aligned} \qquad 4.5.23$$

Given x, y in the same interval, what is the expected value of the function $f\begin{pmatrix} x \\ y \end{pmatrix} = |x-y|$? Now we want to include positive values of $x - y$ and positive values of $y - x$. To do this, let's assume that $x > y$ (so the integral for y goes from 0 to x) and double the integral to account for the cases where $y > x$:

$$\begin{aligned} E\big(|x-y|\big) &= 2\int_0^1 \left(\int_0^x x - y\,dy \right) dx = 2\int_0^1 \left[xy - \frac{y^2}{2} \right]_0^x dx \\ &= 2\int_0^1 x^2 - \frac{x^2}{2}\,dx = 2\left[\frac{x^3}{3} - \frac{x^3}{6} \right]_0^1 = \frac{1}{3}. \quad \triangle \end{aligned} \qquad 4.5.24$$

Exercise 4.5.10 asks you to show that if two points $\mathbf{x}, \mathbf{y}$ are chosen in the unit square, without privileging any part of the square, then $E\Big(|\mathbf{x} - \mathbf{y}|^2\Big) = 1/3$.

Note that the expected value of $|x - y|$ is more complicated to compute than the expected value of $(x-y)^2$. Asked to give the expected distance between x and y, a probabilist would most likely compute $E((x-y)^2)$ and take the square root, getting $1/\sqrt{6}$, which is not the actual expected distance of $1/3$. Computing $E((x-y)^2)$ and taking the square root is analogous to computing the standard deviation; computing $E(|x-y|)$ is analogous to computing the mean absolute deviation.

3 +,-.+ + +/ ! + %% $*) 2"
(%))

Homeomorphisms Between the Circular Disc and the Square

6

Chamberlain Fong

Contents

Introduction 124
Canonical Mapping Space 125
Mapping Diagram with Equations 125
Some Mathematical Details 127
  Fernandez-Guasti Squircle 127
  Tapered2 Squircular Mapping 127
  Lamé Squircle 129
  Elliptical Grid Mapping 131
  Conformal Square Mapping via Schwarz-Christoffel 134
  A Complex Class of Squircles 137
Application: Squaring the Poincaré Disk 139
  Hyperbolic Tilings 142
Application: Elliptification of Rectangular Imagery 144
  Size Versus Shape Distortions 147
Conclusion 147
Cross-References 147
References 147

Abstract

The circle and the square are among the most common shapes used by mankind. Consequently, it is worthwhile to study the mathematical correspondence between the two. This chapter discusses three different ways of mapping a circular region to a square region and vice versa. Each of these mappings has nice closed-form invertible equations and different interesting properties. In addition, this chapter will present artistic applications of these mappings such as

---

C. Fong (✉)
exile.org, San Francisco, CA, USA
e-mail: chamberlain@alum.berkeley.edu

© Springer Nature Switzerland AG 2021

B. Sriraman (ed.), *Handbook of the Mathematics of the Arts and Sciences*,
https://doi.org/10.1007/978-3-319-57072-3_27

converting the Poincaré disk to a square as well as molding rectangular artworks into oval-shaped ones.

Keywords

Conformal square · Escheresque artworks · Invertible mappings · Non-Euclidean geometry · Poincaré disk · squircles

Introduction

In topology, a *homeomorphism* is equivalence relation between two geometric objects that can be continuously deformed into one another. Two objects are homeomorphic if there is a continuous invertible mapping between them. An opt-repeated humorous example of this is the homeomorphism between the donut and the coffee cup. Joking aside, a coffee cup is not at all easy to define mathematically. Instead, this chapter will focus on two objects that are simple and well-defined mathematically – namely, the circular disc and the square. We shall study ways to map the circular disc to a square region and vice versa (Fig. 1).

Needless to say, there are infinitely many ways to map a circular disc to a square. Of particular interest to us are mappings with nice closed-form invertible equations. This chapter will discuss three such mappings. In addition, this chapter will also introduce and discuss different curious properties of these mappings. These properties include concepts such as uniform grids, radial constraints, and conformality. But before we proceed, this is an important lingering issue that needs to be addressed. This mathematical problem sounds all too familiar. Are we dealing with a classic problem in mathematics? Is this problem just "squaring the circle" under a modern guise?

Indeed, there is a famous classic problem in mathematics called "squaring the circle." This well-known geometric construction problem involves using a straight-

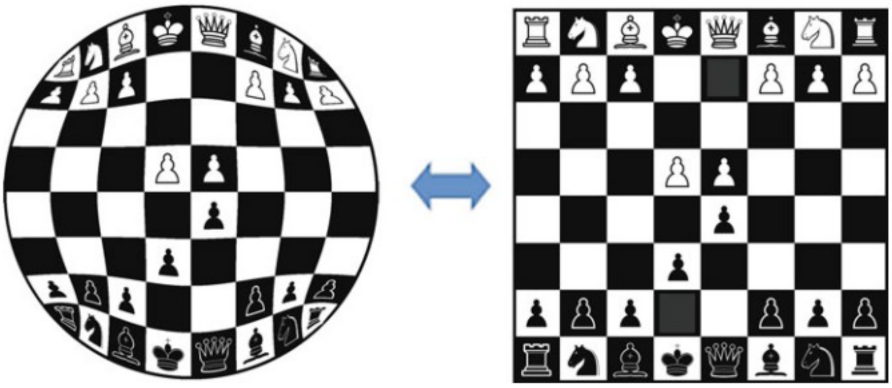

Fig. 1 A chessboard mapped to a circular disc

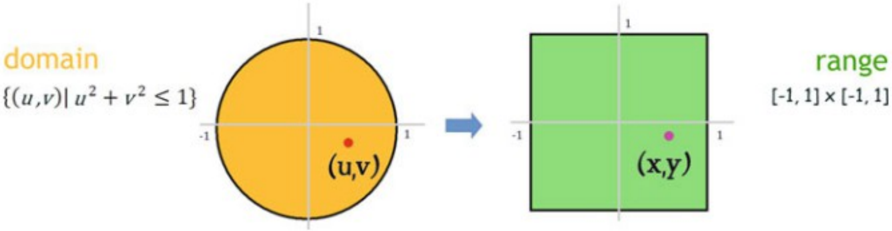

Fig. 2 Canonical mapping space for the circular disc and the square

edge and a compass to produce a circle and square with equal area. At first glance, our mapping problem seems similar to "squaring the circle." However, the two problems are only superficially alike and ultimately quite different. Our mapping problem involves finding two-dimensional mapping equations that a computer can calculate, whereas the classic problem has to do with geometric construction using a draftsman's tools.

Canonical Mapping Space

In order to mathematically describe the mappings, we need to first cover notation and introduce some variables in the canonical mapping space. For the mappings, the domain is the unit disc centered at the origin, and the range is the square with corners at $(\pm 1, \pm 1)$. This is shown in Fig. 2. We shall denote (u, v) as a point contained in the unit disc and (x, y) as the corresponding point contained in the square after the mapping.

Mathematically speaking, we want to find functions f that maps every point (u, v) in the circular disc to a point (x, y) in the square region and vice versa. In others word, we want to derive equations for f such that $(u, v) = f(x, y)$ and $(x, y) = f^{-1}(u, v)$.

Mapping Diagram with Equations

Figure 3 shows the three mappings that are at the core of this chapter. These are the *Tapered2 Squircular*, *Elliptical Grid*, and the *Conformal Square mapping*. In order to illustrate the visual properties of these mappings, there are diagrams for a disc with a radial grid converted to a square. These appear on the left side of Fig. 3. Similarly, on the right side, there are diagrams for a rectilinear square grid converted to a circular disc. Each of these mappings has corresponding forward and inverse equations accompanying the diagrams. From here, one can observe that these equations come in varying degrees of mathematical sophistication.

For the Tapered2 Squircular mapping, it is convenient to use vector notation. In contrast, for the Conformal Square mapping, it is most appropriate to use complex

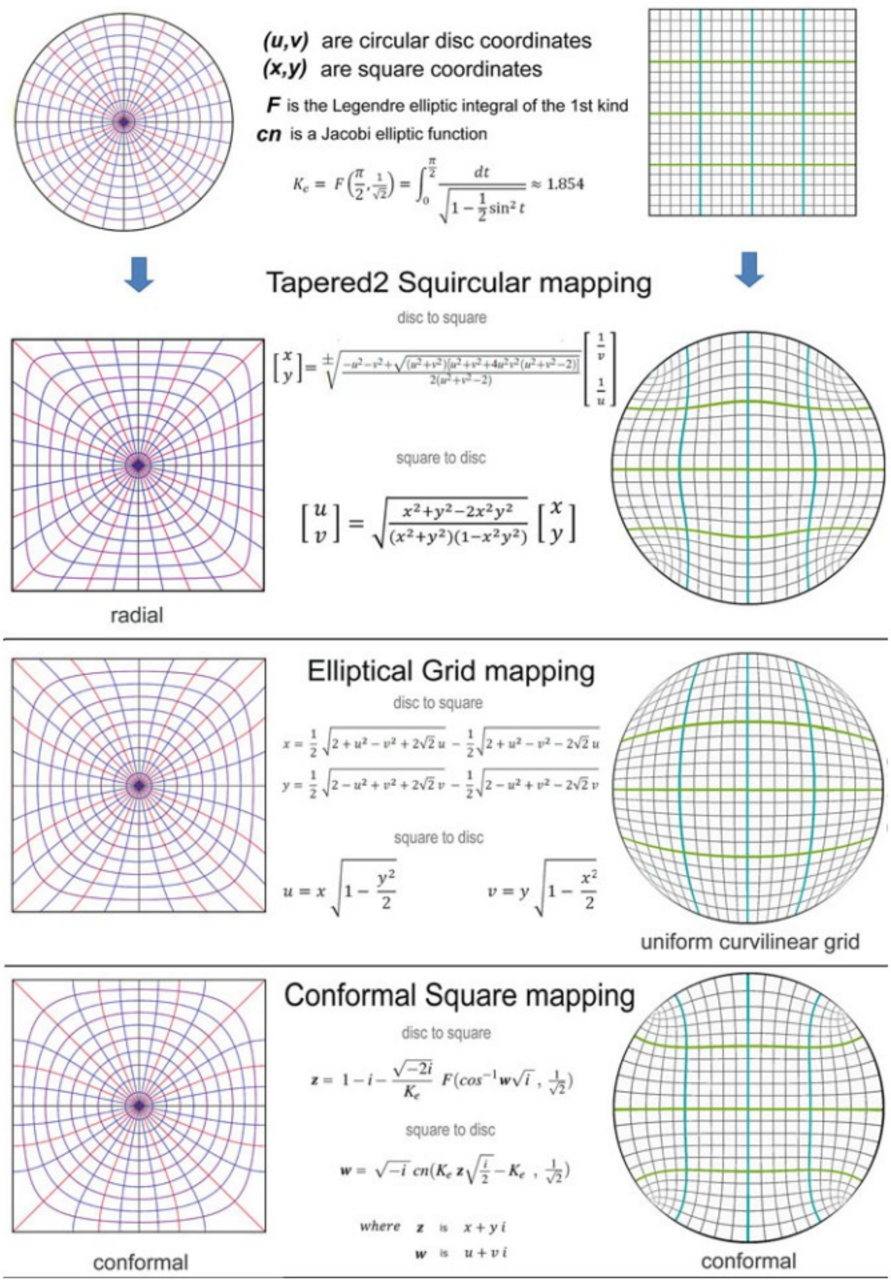
(u,v) are circular disc coordinates
(x,y) are square coordinates
F is the Legendre elliptic integral of the 1st kind
cn is a Jacobi elliptic function
$K_e = F\left(\frac{\pi}{2}, \frac{1}{\sqrt{2}}\right) = \int_0^{\frac{\pi}{2}} \frac{dt}{\sqrt{1-\frac{1}{2}\sin^2 t}} \approx 1.854$
Tapered2 Squircular mapping
disc to square
$\begin{bmatrix} x \\ y \end{bmatrix} = \pm\sqrt{\frac{-u^2-v^2+\sqrt{(u^2+v^2)[u^2+v^2+4u^2v^2(u^2+v^2-2)]}}{2(u^2+v^2-2)}} \begin{bmatrix} \frac{1}{v} \\ \frac{1}{u} \end{bmatrix}$
square to disc
$\begin{bmatrix} u \\ v \end{bmatrix} = \sqrt{\frac{x^2+y^2-2x^2y^2}{(x^2+y^2)(1-x^2y^2)}} \begin{bmatrix} x \\ y \end{bmatrix}$
radial
Elliptical Grid mapping
disc to square
$x = \frac{1}{2}\sqrt{2+u^2-v^2+2\sqrt{2}\,u} - \frac{1}{2}\sqrt{2+u^2-v^2-2\sqrt{2}\,u}$
$y = \frac{1}{2}\sqrt{2-u^2+v^2+2\sqrt{2}\,v} - \frac{1}{2}\sqrt{2-u^2+v^2-2\sqrt{2}\,v}$
square to disc
$u = x\sqrt{1-\frac{y^2}{2}}$
$v = y\sqrt{1-\frac{x^2}{2}}$
uniform curvilinear grid
Conformal Square mapping
disc to square
$z = 1-i-\frac{\sqrt{-2i}}{K_e} F(\cos^{-1} w\sqrt{i}, \frac{1}{\sqrt{2}})$
square to disc
$w = \sqrt{-i}\, cn(K_e z\sqrt{\frac{i}{2}} - K_e, \frac{1}{\sqrt{2}})$
where z is $x + y i$
w is $u + v i$
conformal
conformal

Fig. 3 Some mappings to convert a disc to a square and vice versa

numbers to come up with relatively compact equations. On the other hand, the Elliptical Grid mapping can be expressed simply by using plain algebraic equations. Each of these mappings will be discussed in more detail in the next section.

It is important to review the signum function at this point. This function is abbreviated as $sgn(x)$ and is defined as

$$sgn(x) = \begin{cases} -1 & if \;\; x < 0 \\ 0 & if \;\; x = 0 \\ 1 & if \;\; x > 0 \end{cases}$$

Some Mathematical Details

Fernandez-Guasti Squircle

In 1992, Manuel Fernandez-Guasti discovered a plane algebraic curve that is an intermediate shape between the circle and the square (Fernandez-Guasti, 1992). His curve is represented by a quartic equation.

$$x^2 + y^2 - \frac{s^2}{r^2} x^2 y^2 = r^2 \tag{1}$$

His equation includes a parameter s that can be used to blend between the circle and the square smoothly. Figure 4 illustrates the Fernandez-Guasti squircle at varying values of s.

The squareness parameter s can have any value between 0 and 1. When $s = 0$, the equation produces a circle with radius r. When $s = 1$, the equation produces a square with side length $2r$. In between, the equation produces a smooth curve that is a geometric hybrid of the circle and the square.

Tapered2 Squircular Mapping

The Fernandez-Guasti squircle can be used as groundwork for creating mappings between the circular disc and the square. The key idea is to match circular contours inside the circular disc to squircular contours inside the square. This is shown in Fig. 5.

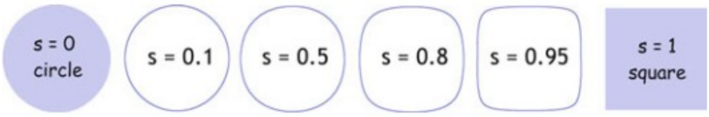

Fig. 4 The Fernandez-Guasti squircle at varying squareness

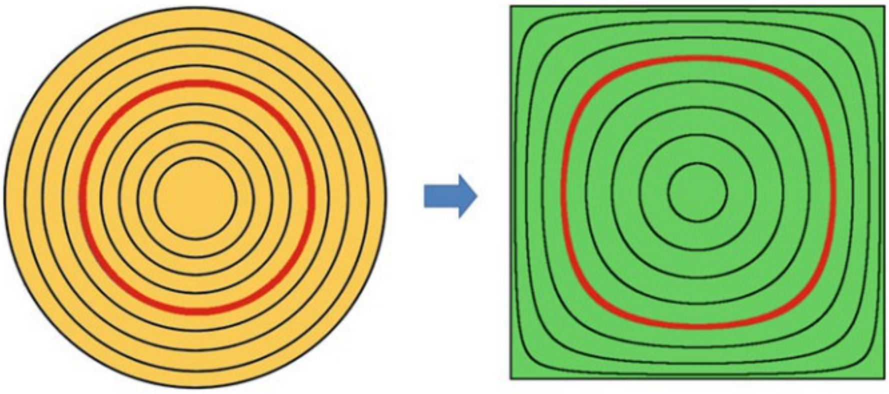

Fig. 5 Mapping based on the squircular continuum

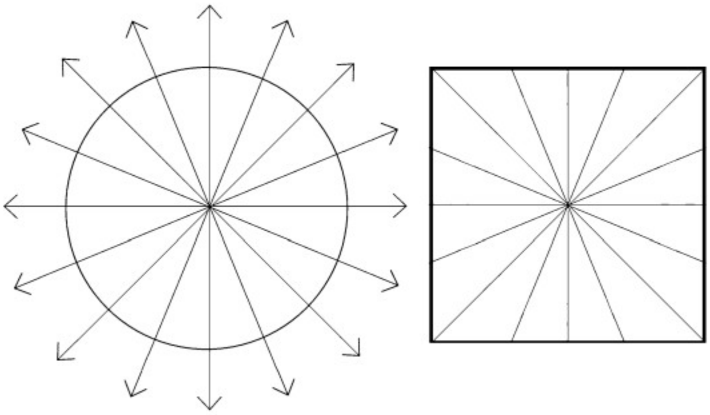

Fig. 6 Radial constraint on the mapping

A circular disc can be considered as a continuum of concentric circles. Likewise, a square region can be considered as a continuum of concentric squircles with increasing squareness. These concentric squircles become increasingly square-like as they approach the bounding rim. We shall denote this phenomena as the *squircular continuum* of the square.

In addition to matching contours inside the circular disc and the square, the Tapered2 Squircular mapping has a restriction called the *radial constraint*. In a nutshell, this means that points are only allowed to move radially during the mapping process. This is illustrated in Fig. 6.

By coupling the radial constraint with a nonlinear squircular continuum, it is possible to derive a set of equations for mapping the circular disc to a square. This is a geometric overview of how the Tapered2 Squircular mapping was derived. For more details, see Fong (2019).

The mapping equations provided in Fig. 3 are not exactly precise. In the interest of brevity, some important details were intentionally left out. For example, in the square-to-disc mapping equation, there are degenerate points that cause a division by zero. These degenerate points are located at the center and the four corners of the square. At these locations, the mapping equation results in the indeterminate form $\frac{0}{0}$. However, the limiting value of the mapping for these points is well-defined.

In view of these and other missing details, here are the amended mapping equations for the Tapered2 Squircular mapping.

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{cases} sgn(uv)\sqrt{\frac{-u^2-v^2+\sqrt{(u^2+v^2)[u^2+v^2+4u^2v^2(u^2+v^2-2)]}}{2(u^2+v^2-2)}} \begin{bmatrix} \frac{1}{u} \\ \frac{1}{v} \end{bmatrix} & if \quad \begin{matrix} u \neq 0 \\ v \neq 0 \end{matrix} \\ \begin{bmatrix} u \\ v \end{bmatrix} & otherwise \end{cases}$$

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{cases} \sqrt{\frac{x^2+y^2-2x^2y^2}{(x^2+y^2)(1-x^2y^2)}} \begin{bmatrix} x \\ y \end{bmatrix} & if \quad \begin{matrix} (x, y) \neq (0, 0) \\ (x, y) \neq (\pm 1, \pm 1) \end{matrix} \\ \begin{bmatrix} sgn(x)\ \frac{1}{2}\sqrt{2} \\ sgn(y)\ \frac{1}{2}\sqrt{2} \end{bmatrix} & if \quad (x, y) = (\pm 1, \pm 1) \\ \begin{bmatrix} 0 \\ 0 \end{bmatrix} & if \quad (x, y) = (0, 0) \end{cases}$$

Lamé Squircle

The Fernandez-Guasti squircle is by no means the only shape that is a parameterized hybrid of the circle and square. In fact, there is a much more famous curve known as the superellipse that also has this property. The Lamé squircle is a special case of the superellipse with no eccentricity. This plane algebraic curve was originally studied by Gabriel Lamé in 1818. Qualitatively, the curve appears very similar to the Fernandez-Guasti squircle, but there are several important differences which will be discussed later. The Lamé squircle has the equation

$$|x|^p + |y|^p = r^p \tag{2}$$

There are two parameters in this equation: p and r. The power parameter p allows the shape to interpolate between the circle and the square. When $p = 2$, the equation produces a circle with radius r. As $p \to \infty$, the equation produces a square with a side length of $2r$. In between, the equation produces a smooth planar curve that resembles both the circle and the square. Figure 7 shows the Lamé curve at increasing powers.

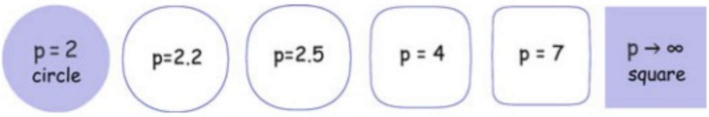

Fig. 7 Lamé squircle at varying polynomial powers

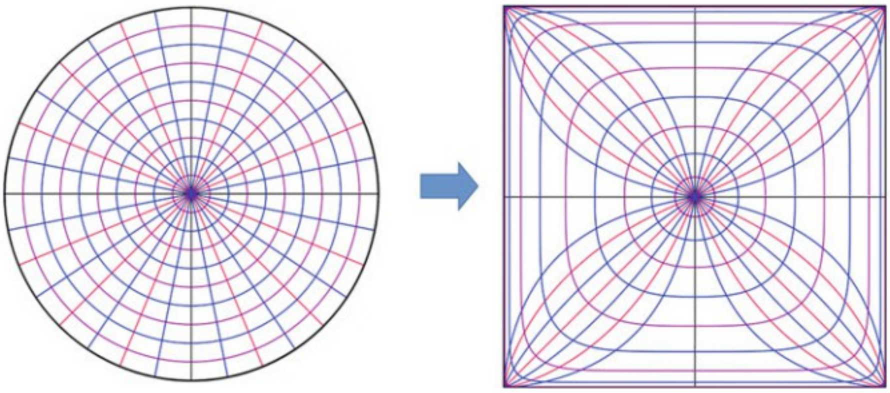

Fig. 8 Disc-to-square mapping based on the Lamé squircle

Although the Lamé squircle exhibits qualitative similarities with the Fernandez-Guasti squircle, there is a major difference between the two. The Lamé squircle can only approximate the square. It requires an infinite exponent in order to fully realize the square. Moreover, since the polynomial equation for the Lamé curve has unbounded exponents, it is unwieldy and difficult to manipulate algebraically. Nevertheless, the Lamé squircle can be used as groundwork for devising another disc-to-square mapping.

The Lamé squircle has a parametric form in which x and y coordinates of points on the curve can be specified in terms of parameter t. The parametric equations are

$$\begin{aligned} x(t) &= sgn(\cos t)\ |\cos t|^{\frac{2}{p}} \\ y(t) &= sgn(\sin t)\ |\sin t|^{\frac{2}{p}} \end{aligned} \tag{3}$$

Using these parametric equations of the Lamé squircle, it is possible to devise another disc-to-square mapping. The key idea is to make the substitutions $\mathbf{u} = \cos \mathbf{t}$ and $\mathbf{v} = \sin \mathbf{t}$ to represent points in the circular disc. This mapping is shown in Fig. 8 with corresponding equations as

$$\begin{aligned} x &= sgn(u)\ |u|^{1-u^2-v^2} \\ y &= sgn(v)\ |v|^{1-u^2-v^2} \end{aligned} \tag{4}$$

This mapping based on the Lamé squircle has several undesirable properties that make it less useful than the other mappings shown in Fig. 3. For one thing, explicit inverse equations are not provided. It is quite likely that there are no closed-form equations for expressing the inverse mapping. This means that the inverse mapping can only be computed using iterative root-solving techniques in numerical analysis.

Another egregious property of this mapping is that it only works on the open circular disc and the open square region. This means that mapping does not include the circular boundary of the disc or the rim of the square region. Mathematically, the domain of the mapping is the open circular disc $\{(u, v) | u^2 + v^2 < 1\}$, and the range of the mapping is the $(-1, 1) \times (-1, 1)$ open square region.

Elliptical Grid Mapping

The square can easily be subdivided into a grid of smaller squares. In stark contrast, there is no easy way to subdivide the circle into a grid of smaller nonoverlapping circles. Nevertheless, the circular disc can be subdivided into a uniform grid based on elliptical arcs. This is shown on the right of Fig. 9. Using this observation, it is possible to devise a square-to-disc mapping. The key idea is to match cells inside the square grid with cells inside the circular grid.

In 2005, Philip Nowell (2005) came up with very simple equations for mapping a square to a circular disc. The derivation basically boils down to two mathematical constraints needed to match square cells with elliptical grid cells. The first constraint comes from matching vertical line segments inside the square to vertically oriented elliptical arcs inside the circular disc. This is illustrated in Fig. 10 where a vertical line and its corresponding elliptical arc are shown in red. This mapping relationship can be summarized by the equation:

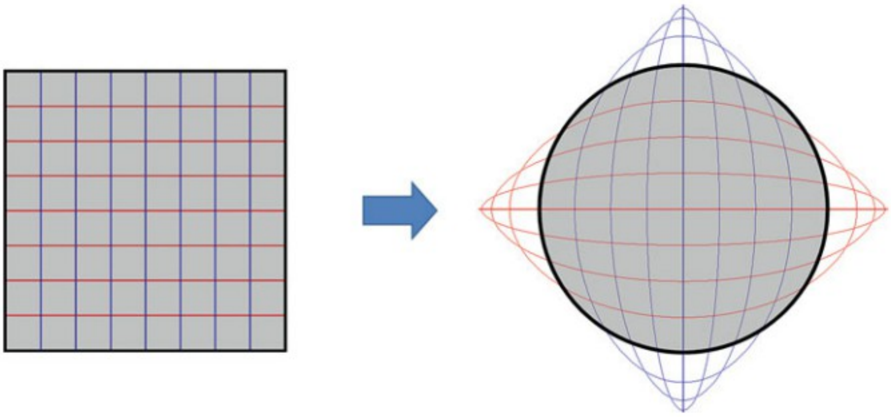

Fig. 9 Overview of the Elliptical Grid mapping

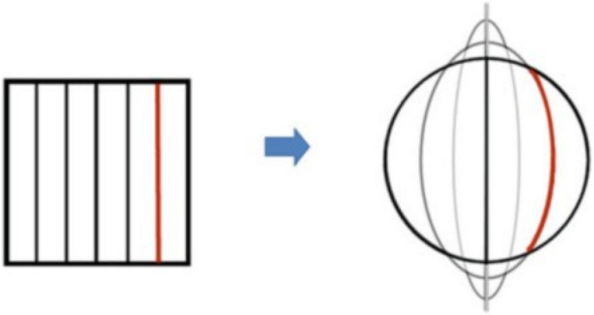

Fig. 10 Vertical constraint of the Elliptical Grid mapping

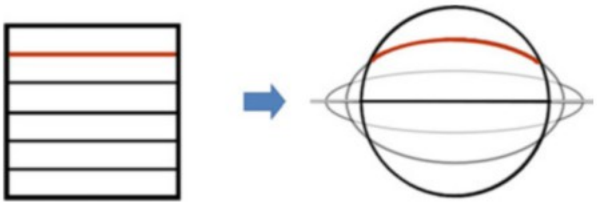

Fig. 11 Horizontal constraint of the Elliptical Grid mapping

$$1 = \frac{u^2}{x^2} + \frac{v^2}{2 - x^2} \tag{5}$$

Essentially, for each vertical line segment of constant x inside the square, there is a corresponding equation of an ellipse centered in the circle. The bounds, eccentricity, and semi-major and semi-minor axial lengths of this ellipse vary depending on the value of x. Specifically, the left and right vertex tips of the ellipse follow the value of x. Meanwhile, the top and bottom vertex tips vary as $\pm\sqrt{2 - x^2}$.

The second constraint of the mapping comes from matching horizontal line segments inside the square to sideway-oriented elliptical arcs inside the circular disc. This is illustrated in Fig. 11 where a horizontal line and its corresponding elliptical arc are shown in red. This mapping relationship can also be summarized by the equation:

$$1 = \frac{u^2}{2 - y^2} + \frac{v^2}{y^2} \tag{6}$$

Essentially, for each horizontal line segment of constant y in the square, there is a corresponding equation of an ellipse centered in the circle. The bounds, eccentricity, and semi-major and semi-minor axial lengths of this ellipse vary depending on the value of y. Specifically, the top and bottom vertex tips of the ellipse follow the value of y. Meanwhile, the left and right vertex tips vary as $\pm\sqrt{2 - y^2}$.

Basically, the mapping assigns the grid of perpendicular vertical and horizontal lines inside the square to a grid of elliptical arcs inside the circular disc. As shown in Fig. 9, this is essentially matching square grid cells with elliptical grid cells . A curvilinear grid of elliptical arcs can be formed by superimposing the vertically

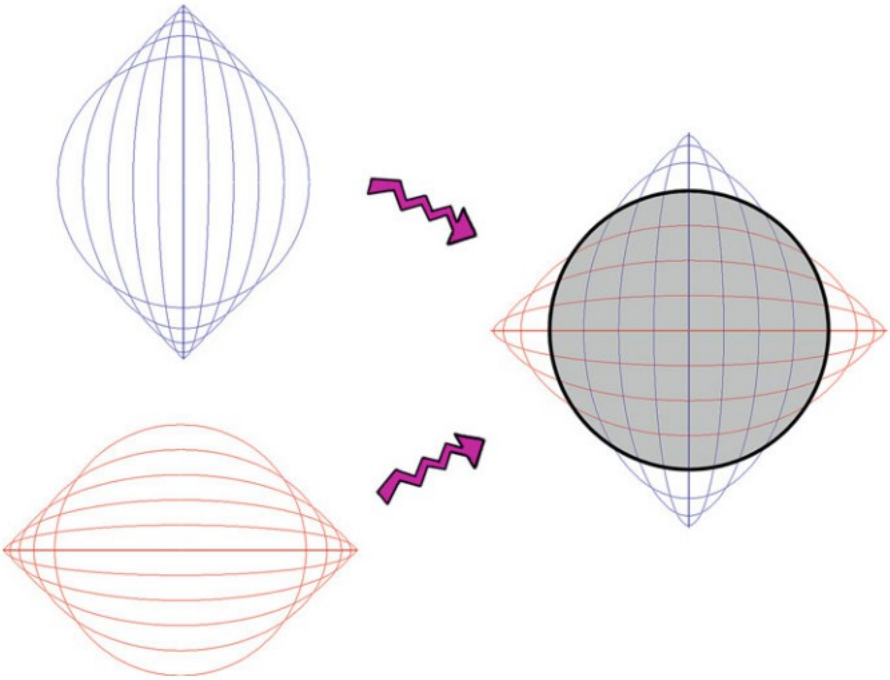

Fig. 12 Superimposing the vertical and horizontal arcs to create a curvilinear elliptical grid

oriented and horizontally oriented elliptical arcs inside the circular disc. Figure 12 shows the resulting curvilinear grid resulting from the superimposition.

Note that the ellipses get more and more eccentric as x or y approach zero. Also, the ellipses get more circular as x or y approach ± 1.

Mathematically, we can mix the vertical constraint equation with the horizontal constraint equation to get algebraic expressions for u and v in terms of x and y. This can be done by starting with the vertical constraint equation and doing some algebraic manipulations to isolate u^2.

$$1 = \frac{u^2}{x^2} + \frac{v^2}{2 - x^2} \quad \Longrightarrow \quad 1 - \frac{v^2}{2 - x^2} = \frac{u^2}{x^2} \quad \Longrightarrow \quad u^2 = x^2 \left(1 - \frac{v^2}{2 - x^2} \right)$$

We can then plug this u^2 value into the horizontal constraint equation

$$1 = \frac{u^2}{2 - y^2} + \frac{v^2}{y^2} \quad \Longrightarrow \quad 1 = \frac{x^2 \left(1 - \frac{v^2}{2 - x^2} \right)}{2 - y^2} + \frac{v^2}{y^2}$$

Multiply both sides of the equation by $(2 - x^2)(2 - y^2)y^2$ to remove fractions and get

$$(2-x^2)(2-y^2)y^2 = x^2y^2(2-x^2-v^2) + v^2(2-x^2)(2-y^2)$$

After which, one can isolate v^2 into one side of the equation

$$\begin{aligned}
(2-x^2)(2-y^2)y^2 - x^2y^2(2-x^2) = -x^2y^2v^2 + (2-x^2)(2-y^2)v^2 &\implies \\
v^2(2-x^2)(2-y^2-x^2y^2) = (2-x^2)y^2(2-y^2-x^2) &\implies \\
v^2 = y^2\frac{(2-x^2)(2-y^2-x^2)}{(2-x^2)(2-y^2)-x^2y^2} &\implies \\
v^2 = y^2\frac{(2-x^2)(2-y^2-x^2)}{4-2x^2-2y^2} = y^2\frac{2-x^2}{2} &\implies \\
v = y\sqrt{\frac{2-x^2}{2}} \implies \mathbf{v} = \mathbf{y}\sqrt{\mathbf{1}-\frac{\mathbf{x}^2}{\mathbf{2}}}
\end{aligned}$$

Similarly, in a symmetric fashion, one can solve for u as

$$\mathbf{u} = \mathbf{x}\sqrt{\mathbf{1}-\frac{\mathbf{y}^2}{\mathbf{2}}}$$

It is also possible to derive the inverse equations for this mapping, but the algebraic manipulation needed is much more elaborate. Instead, we refer the reader to Fong (2014) for a lengthy derivation of denested inverse equations.

Conformal Square Mapping via Schwarz-Christoffel

A conformal map is a mapping that preserves angles between geometric features after the mapping operation is performed. In other words, a conformal mapping does not distort angles. This section will discuss a disc-to-square mapping that is conformal. However, in order to do this, we need to take a small detour through the complex plane beforehand and discuss some theory first.

One of the most celebrated results in nineteenth-century complex analysis is the Riemann mapping theorem (Frederick and Schwarz, 1990). It states that there exists a conformal map between the open unit disc and every simply connected subset of the complex plane. Moreover, it states that this conformal map is unique if we fix a point and the orientation of the mapping. In theory, the Riemann mapping theorem is nice, but it is only an existence theorem. It does not specify how to find the conformal mapping. The next important breakthrough came with the works of Hermann Schwarz and Elwin Christoffel. In the 1860s, Schwarz and Christoffel independently developed a formula for a conformal mapping between the unit disc and simple polygonal regions in the complex plane. The formula is complicated and involves an integral in the complex plane.

In order to show how complicated this formula is, it is listed below in its full glory. We will not bother explaining what each of the variables mean in the complex integral. It is suffice to say that the general Schwarz-Christoffel formula is quite complicated.

$$f(z) = \int^{z} \prod_{k=1}^{n} \left(1 - \frac{\zeta}{z_k}\right)^{\alpha_k - 1} d\zeta \tag{7}$$

Furthermore, for most polygons, the integral can only be approximated numerically. Fortunately, for the special case of the square, the Schwarz-Christoffel formula can be reduced to an explicit analytical expression involving complex-valued elliptic integrals and elliptic functions (Langer and Singer, 2011).

Legendre Elliptic Integrals

It is appropriate to make a segue here and briefly discuss elliptic integrals and elliptic functions. Of particular interest to us are Legendre elliptic integrals and Jacobi elliptic functions. Specifically, we are interested in the incomplete Legendre elliptic integral of the first kind called F and its closely related inverse function, the Jacobi elliptic function cn.

Mathematically, the incomplete Legendre elliptic integral of the first kind is a two-parameter function defined as

$$F(\phi, k) = \int_{0}^{\phi} \frac{1}{\sqrt{1 - k^2 sin^2(t)}}\, dt \tag{8}$$

Note that this integral cannot be simplified using any of the standard techniques covered in freshman calculus classes. This integral was originally studied in the context of measuring the arc length of an ellipse. That is the reason why it is called an elliptic integral (Rice and Brown, 2012).

The two arguments provided with this function are also intimately tied to the ellipse. The first parameter ϕ is some sort of angular parameter. It originates from the angle subtended by an arc of the ellipse. The second parameter k is closely related to the eccentricity of the ellipse.

Meanwhile, the Jacobi elliptic function cn can be considered as some sort of inverse to F. Mathematically, it is related to F by the following equation:

$$cn(\eta, k) = \cos(F^{-1}(\eta, k)) \tag{9}$$

A Fundamental Conformal Map

Without getting into the nitty-gritty details of the Schwarz-Christoffel mapping, Fig. 13 shows a fundamental conformal map between the circular disc and the square in the complex plane. This mapping can be derived by simplifying the Schwarz-Christoffel integral for the square and using the doubly periodic nature of the Jacobi elliptic function cn on the complex plane. In essence, one could map every point

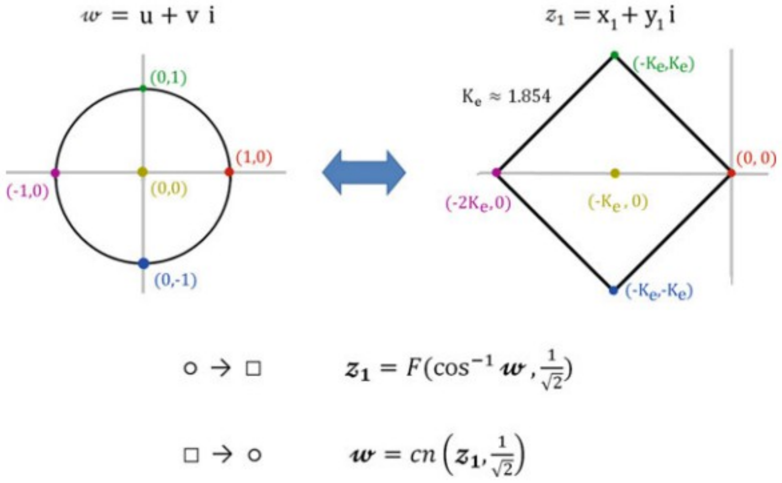

Fig. 13 A conformal map between the disc and square in the complex plane

inside the unit disc to a square region conformally by just an evaluation of the complex-valued Jacobi elliptic function *cn*. Furthermore, the inverse of the mapping can be calculated using the incomplete Legendre elliptic integral of the first kind F.

Canonical Alignment

The main drawback of the diagram on the complex plane is that the x and y coordinates are not in the canonical mapping space. Figure 13 shows a square with corner coordinates in terms of a constant K_e instead of the ±1 that is desired. Moreover, the square is tilted by 45° and off-center from the origin. In order to get this mapping into the canonical mapping space, one needs to perform a series of affine transformations on the square. These include centering the square to the origin and scaling it down to have a side length value of 2. In order to do this, one has to introduce a rotational factor for the 45° tilt as well as K_e offsets and scale factors. This is exactly what happens in the explicit equations for the Conformal Square mapping. Basically, this is the canonized mapping equation in the complex plane

$$\mathbf{z} = 1 - i - \frac{\sqrt{-2i}}{K_e} \;\; F(cos^{-1}\mathbf{w}\sqrt{i},\;\; \tfrac{1}{\sqrt{2}}) \tag{10}$$

and its inverse is

$$\mathbf{w} = \sqrt{-i} \;\; cn(K_e\mathbf{z}\sqrt{\tfrac{i}{2}} - K_e,\;\; \tfrac{1}{\sqrt{2}}) \tag{11}$$

where $\boldsymbol{z}$ is the complex number $x + y\,\boldsymbol{i}$
and $\boldsymbol{w}$ is the complex number $u + v\,\boldsymbol{i}$.

It is probably appropriate to explain the $\sqrt{\pm i}$ factors that appear throughout the equations. These multiplicative constants are just a compact way of representing the ±45° rotational adjustments needed to align the equations to the canonical mapping space in Fig. 2.

In the complex plane, rotation can be done simply by multiplication with the complex number $e^{i\theta}$. For 45° rotation, the multiplicative factor is

$$e^{i\frac{\pi}{4}} \;=\; \cos\frac{\pi}{4} + i\;\sin\frac{\pi}{4} \;=\; \frac{1}{\sqrt{2}}(1+i) \;=\; \sqrt{i} \tag{12}$$

For −45° rotation, the multiplicative factor is

$$e^{-i\frac{\pi}{4}} \;=\; \cos\frac{\pi}{4} - i\;\sin\frac{\pi}{4} \;=\; \frac{1}{\sqrt{2}}(1-i) \;=\; \sqrt{-i} \tag{13}$$

Software Implementation

In order to implement this mapping on a computer, one needs to be able to calculate special functions such as cn and F. Source code for numerical computation of these functions is readily available in open-source libraries such as Boost and the GNU Scientific Library. Also, there is a reference implementation appearing in the popular Numerical Recipes book (Press et al., 1992).

One possible pitfall in software implementation is that the mapping requires complex-valued versions of the special functions. These complex variants are typically not included in open-source libraries. Nonetheless, these complex variants are well-defined mathematically. The formulas for computing complex-valued cn and F are given by L.M. Milne-Thomson in the classic AMS-55 reference book (Abramowitz and Stegun, 1972).

It is worth mentioning here that there are well-established fast and robust algorithms for computing the F and cn special functions (Carlson, 1977). As a matter of fact, Gauss has shown that elliptic integrals can be calculated quickly using the arithmetic-geometric mean (Hancock, 1958). It is suffice to say that these special functions can be calculated about as fast as standard trigonometric functions.

A Complex Class of Squircles

The Fernandez-Guasti squircle was used as groundwork in the development of the Tapered2 Squircular mapping. This section will discuss using a similar idea but acting in reverse to derive another type of squircle. Specifically, one can start with the Conformal Square mapping and come up with a different type of squircle called the *complex squircle*.

Let us revisit the squircular continuum in Fig. 5 which was originally discussed in the derivation of the Tapered2 Squircular mapping. By working backward this time and starting from the Conformal Square mapping, it is possible to derive equations for a curve that is an intermediate shape between the circle and the square.

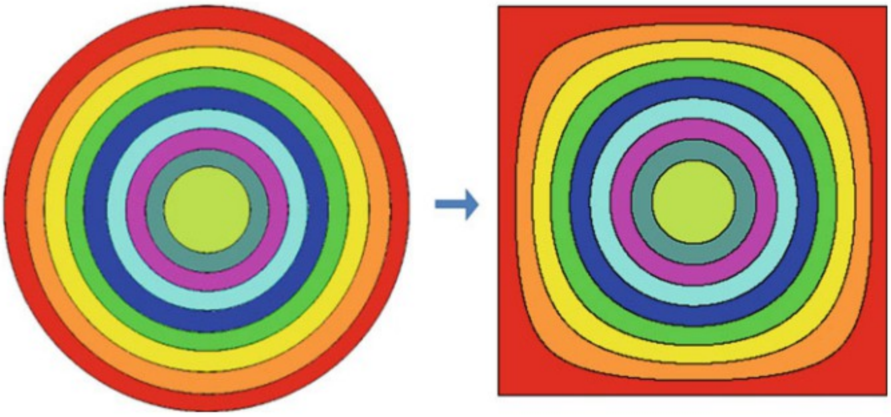

Fig. 14 The squircular continuum revisited

Figure 14 shows a circular disc subdivided into concentric rings of different colors. This circular disc is then mapped to a square via the Conformal Square mapping. The resulting square is still subdivided by concentric rings, but the enclosing shapes are not quite circular. Consequently, it is logical to ask what sort of shape encompasses the boundaries of the concentric rings inside the square.

One can then intuitively surmise from Fig. 14 that the concentric shapes inside the square are some sort of squircular curve. Indeed, these concentric shapes exhibit all of the defining characteristics of a hybrid curve between the circle and the square. Therefore, it makes sense to classify them as yet another type of squircle – the *complex squircle*.

The complex squircle can be written in parametric form by applying Conformal Square mapping to the parametric equation of the circle. In the complex plane, the parametric form of the unit circle is e^{it}, with $0 \leq t \leq 2\pi$. By applying the Conformal Square mapping on the circle, one can come up with a complex-valued function ψ that serves as an auxiliary representation of the complex squircle.

$$\psi(t,q) \;=\; 1 - i - \frac{\sqrt{-2i}}{K_e} F\left(\cos^{-1}(q e^{it}\sqrt{i}),\, \tfrac{1}{\sqrt{2}}\right) \tag{14}$$

ψ is a two-parameter complex function that is based on the disc-to-square equation of the Conformal Square mapping. The first argument t is used for curve parametrization. The second argument q is just a squareness parameter analogous to the s parameter of the Fernandez-Guasti squircle.

Using this ψ function, the complex squircle can be defined as the curve arising from these parametric equations:

$$\begin{aligned} x(t) \;&=\; \Re[\psi(t,q)]\, \tfrac{r}{\psi(0,q)} \\ y(t) \;&=\; \Im[\psi(t,q)]\, \tfrac{r}{\psi(0,q)} \end{aligned} \tag{15}$$

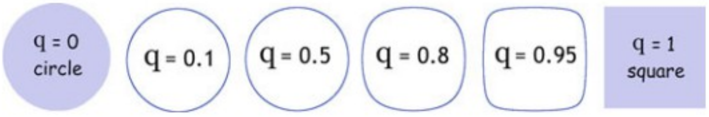

Fig. 15 The complex squircle at varying squareness values

Table 1 Three different types of squircles discussed in this chapter

Name	Squareness	Equation	Key property
Fernandez-Guasti squircle	$s \in [0, 1]$	$x^2 + y^2 - \frac{s^2}{r^2}x^2y^2 = r^2$	Quartic polynomial equation
Lamé squircle	$p \in [2, \infty)$	$\lvert x\rvert^p + \lvert y\rvert^p = r^p$	Unbounded polynomial power
Complex squircle	$q \in [0, 1]$	$\psi(t,q) = 1 - i - \frac{\sqrt{-2i}}{K_e} F(cos^{-1}(qe^{it}\sqrt{i}), \frac{1}{\sqrt{2}})$ $x(t) = \Re[\psi(t, q)]\ \frac{r}{\psi(0,q)}$ $y(t) = \Im[\psi(t, q)]\ \frac{r}{\psi(0,q)}$	Complex parametric equations with parameter $t \in [0, 2\pi]$

As with the other types of squircles previously discussed, there are two parameters for the complex squircle: q and r. The squareness parameter q allows this shape to interpolate between the circle and the square. When $q = 0$, the equation produces a circle with radius r. When $q = 1$, the equation produces a square with a side length of $2r$. In between, the equation produces a smooth planar curve that resembles both the circle and the square. This is shown in Fig. 15.

In summary, mathematical shapes known as squircles play an important role in the development of disc-to-square mappings. This chapter discussed three different types of squircles along with associated mappings. These squircles are summarized in Table 1.

Application: Squaring the Poincaré Disk

This section will discuss how to convert circular Escheresque artwork into squares. Figure 16 shows an example of a circular tiling with interlocking angels and devils. The pattern on the right was inspired from M.C. Escher's famous *Circle Limit IV* (1960). The pattern on the left is its conversion to a square.

In order to delve more into Escheresque artworks (Dunham, 2009), some mathematical background in non-Euclidean geometry is necessary. In particular, we need to discuss a type of non-Euclidean geometry called hyperbolic geometry. To make things simple, we shall restrict ourselves to a two-dimensional construct of hyperbolic geometry called the hyperbolic plane. One can intuitively think of the hyperbolic plane as a surface with negative curvature everywhere (Taimina, 2009). In contrast, the Euclidean plane is a flat surface with zero curvature everywhere.

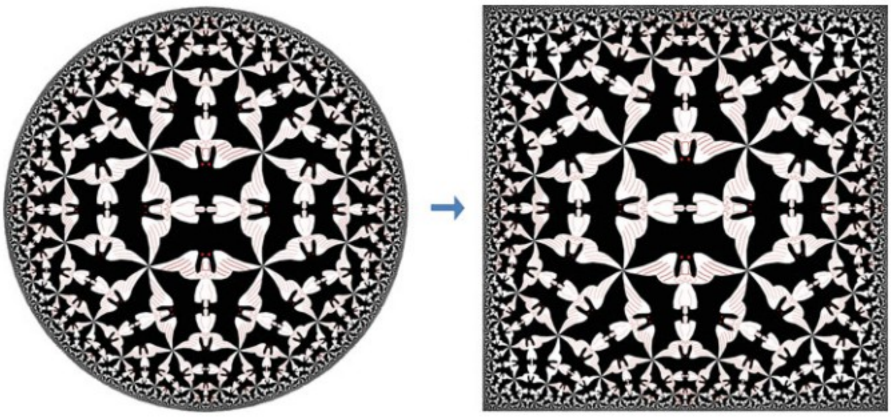

Fig. 16 A circular angels & devils pattern converted to a square

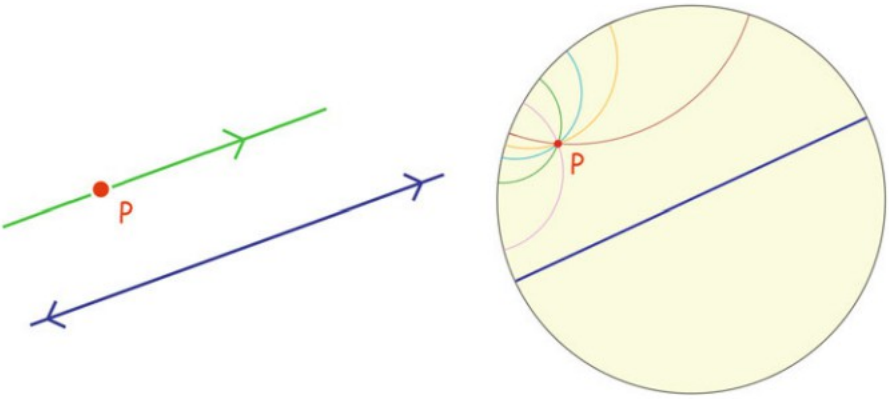

Fig. 17 Playfair's axiom on the Euclidean plane (left) and the hyperbolic plane (right)

There is a long and storied history of non-Euclidean geometry that this chapter will only gloss over. Non-Euclidean geometry arises from the negation of Euclid's fifth postulate – also known as the parallel postulate. A modern formulation of this postulate states that

> In a plane, given a line and a point not on it, at most one line parallel to the given line can be drawn through the point.

This formulation is known as Playfair's axiom. It is illustrated in the left diagram of Fig. 17. Meanwhile, hyperbolic geometry arises when Playfair's axiom is overturned with the following statement:

> In a hyperbolic plane, given a line and a point not on it, there are several lines parallel to the given line than can be drawn through the point.

In order to qualify this statement in hyperbolic geometry, one needs to have well-defined and consistent notions of points, lines, and parallelism in the hyperbolic plane. First, let us introduce the Poincaré disk.

Mathematicians have come up with many different models of the hyperbolic plane in order to study hyperbolic geometry. The most popular model is probably the Poincaré disk. The Poincaré disk can be considered as some sort of projection of the hyperbolic plane onto a unit circular disc in the Euclidean plane. It has several interesting properties that make it desirable for representing the hyperbolic plane.

One important property of the Poincaré disk is that it presents the entire hyperbolic plane within the confines of a finite circular disc in IR^2. Every point in the hyperbolic plane is represented by a unique point inside the Poincaré disk. This makes it easy to visualize the entire hyperbolic plane within the model.

Another important property of the Poincaré disk is conformality. In a conformal model, the hyperbolic measure of angle is the same as the Euclidean measure of angle. In other words, this hyperbolic model does not distort angles. One can measure hyperbolic angles between geometric entities inside the Poincaré disk by simply making Euclidean angular measurements. Unfortunately, the same does not hold true for hyperbolic distance. Hyperbolic distances are greatly distorted in the Poincaré disk.

In the Poincaré disk model, the origin at $(0, 0)$ lies at the center of the circular disc. The bounding circle at the rim of disc is infinitely far away from the origin. The hyperbolic distance of a point away from the origin gets larger and larger as the point gets closer to the bounding circle. As a consequence of this, geometric entities appear smaller as they approach the bounding circle.

One can think of the Poincaré disk in analogy to the Cartesian coordinate system. Just like the Cartesian coordinate system faithfully models the Euclidean plane, the Poincaré disk is a meaningful representation of the hyperbolic plane. The Cartesian coordinate system allows one to convert geometric statements about the Euclidean plane into algebraic equations. In a similar manner, the Poincaré disk allows one to do the same for the hyperbolic plane.

Table 2 lists various hyperbolic entities along with their Euclidean counterparts from within the Poincaré disk model. Using this table, one can make sense of Playfair's axiom in the hyperbolic plane.

Table 2 Hyperbolic entities in the Poincaré disk

Geometric entity	Corresponding feature inside the Poincaré disk
Hyperbolic point	This corresponds to a Euclidean point in the interior of the bounding circle of the Poincaré disk model
Hyperbolic line	This corresponds to a circular arc orthogonal to the bounding circle. Circular arcs can be part of circles of arbitrary radius
Hyperbolic angle	The angle between two hyperbolic lines is the same as the Euclidean angle between their corresponding circular arcs inside the Poincaré disk model
Hyperbolic parallelism	Hyperbolic lines are parallel if and only if they do not intersect in the interior of the Poincaré disk model

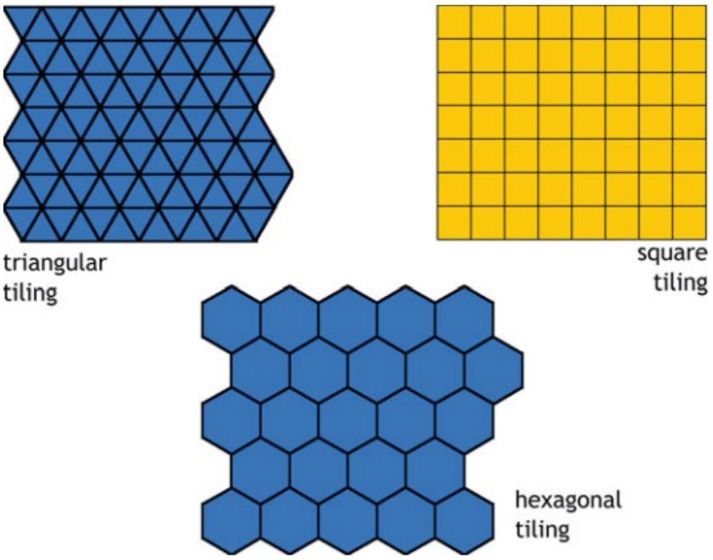

Fig. 18 The three regular Euclidean tilings

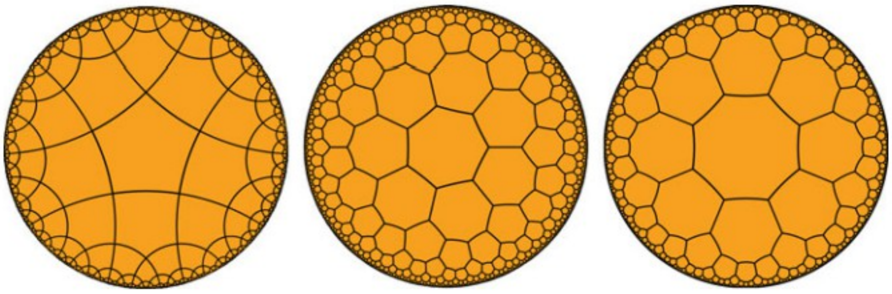

Fig. 19 Possible pentagonal, heptagonal, and octagonal tilings of the Poincaré disk

Hyperbolic Tilings

It is well-known that only three types of regular polygons can tile the Euclidean plane – namely, the triangle, the square, and the hexagon. These tilings are shown in Fig. 18. It is lesser known that all regular polygons can tile the hyperbolic plane! To illustrate this, Fig. 19 shows the Poincaré disk tiled with hyperbolic pentagons, heptagons, and octagons.

The stop sign is probably the most common octagonal shape that people encounter in their daily lives. It was ratified as an international standard by the United Nations in 1968. Being a regular octagon, the stop sign can tile the entire hyperbolic plane. This is shown in Fig. 20 as a Poincaré disk pattern. Furthermore,

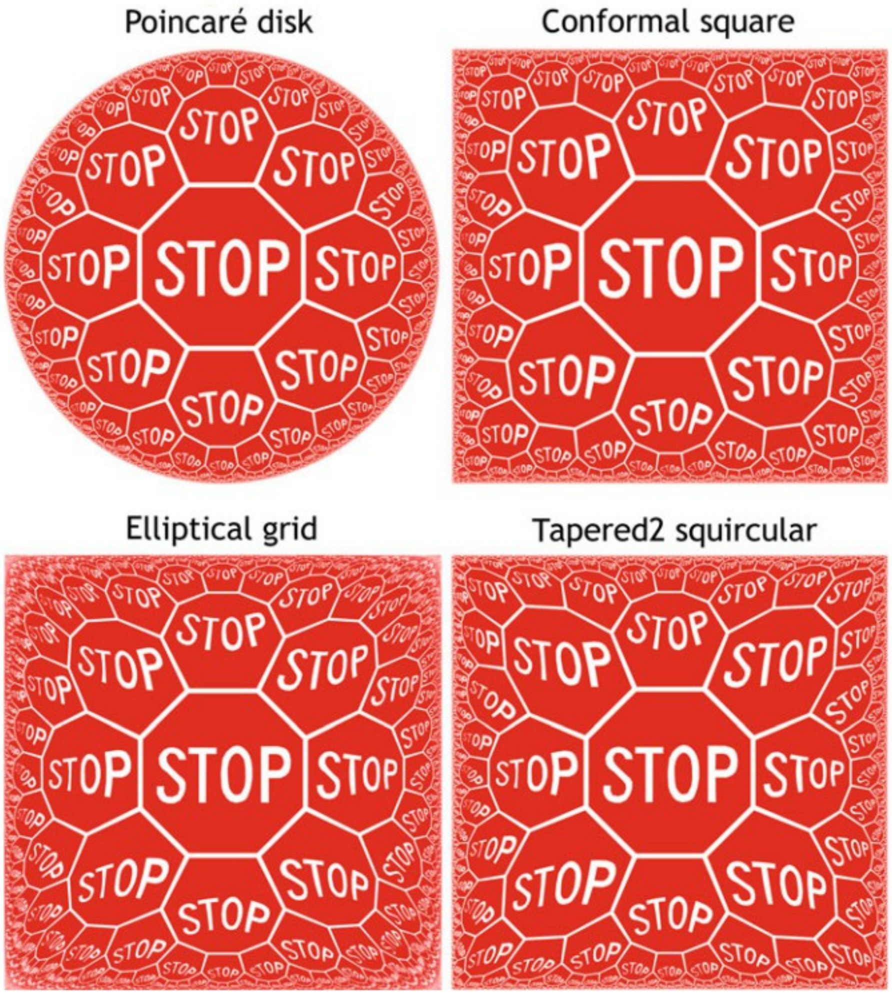

Poincaré disk

Conformal square

Elliptical grid

Tapered2 squircular

Fig. 20 Hyperbolic stop signs

it is possible to convert this pattern into squares by using any of the three mappings discussed in this chapter.

The most relevant and appropriate way to convert the Poincaré disk to a square is via Conformal Square mapping. There is a very simple reason for this – conformality. Recall that the Poincaré disc model is a conformal model of the hyperbolic plane. It follows that the conformal square is a natural extension of the Poincaré disc to a square region. In essence, undistorted Euclidean angles in the Poincaré disk remain intact in the conformal square.

The two other mappings can also be used to convert the Poincaré disk to a square, but the results are less than stellar. For the Elliptical Grid mapping, the corners

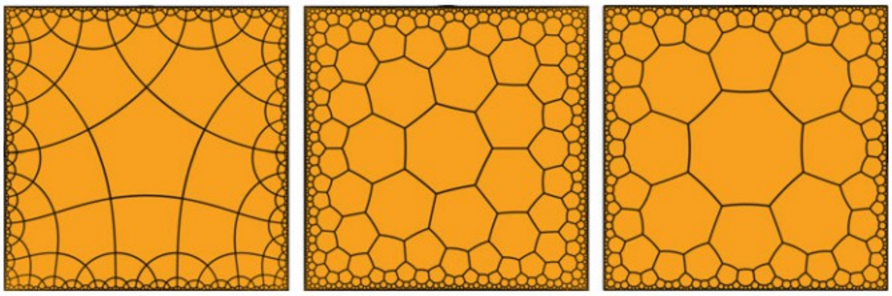

Fig. 21 Hyperbolic tilings mapped to the conformal square

of the square appear as a muddled mess. Meanwhile, the Tapered2 Squircular mapping produces fairly decent results, but there is significant shape distortion of the octagons.

To summarize, the conformal square is the natural extension of the Poincaré disk to a square. In order to further illustrate this, Fig. 21 shows the conformal square versions of the hyperbolic tilings from Fig. 19.

Application: Elliptification of Rectangular Imagery

Most of the world's photographs are rectangular. However, one might want to convert them into circular or elliptical images for artistic reasons. This section will extend the previously discussed square-to-disc mappings to handle ellipses and rectangles. After which, one can apply these mappings to rectangular artworks to produce elliptical ones. Henceforth, this process of molding a rectangle into an ellipse will be referred to as *elliptification*.

Elliptification is very simple if one already has a square-to-disc mapping. Elliptification can be done by simply removing the eccentricity and then reintroducing it back after the square-to-disc mapping is performed. This procedure is shown in the top diagram in Fig. 22 along with an illustrative example showing the elliptification of the United States flag at the bottom.

The artistic process of elliptification is not new. For centuries, artists would paint on oval-shaped canvases, and photographers would cut their pictures into oval regions. Creating oval-shaped artworks is an important form of artistic expression and stylization. This has traditionally been done by just cropping or cutting out the corner regions of the artwork.

This chapter promotes the idea of using explicit mathematical mappings to create oval-shaped artworks. This is possible using any of the square-to-disc mappings covered in this chapter and related papers (Fong, 2014, 2019; Shirley and Chiu, 1997). To demonstrate this effect, Figs. 23 and 24 show Edouard Manet's last major work *"A Bar at the Folies-Bergere"* (1882) converted into an oval region using

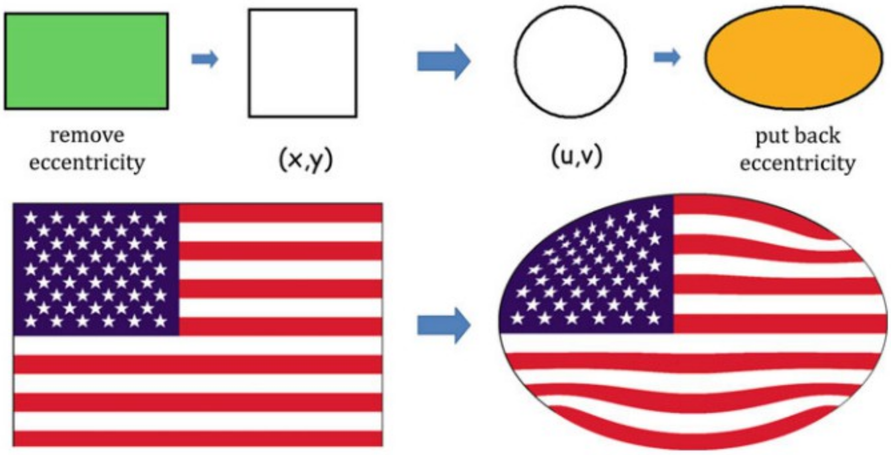
remove
eccentricity
(x,y)
(u,v)
put back
eccentricity

Fig. 22 Elliptification: converting rectangular imagery to oval regions

Fig. 23 Manet's *"A Bar at the Folies-Bergere"* (1882) and its cropped version

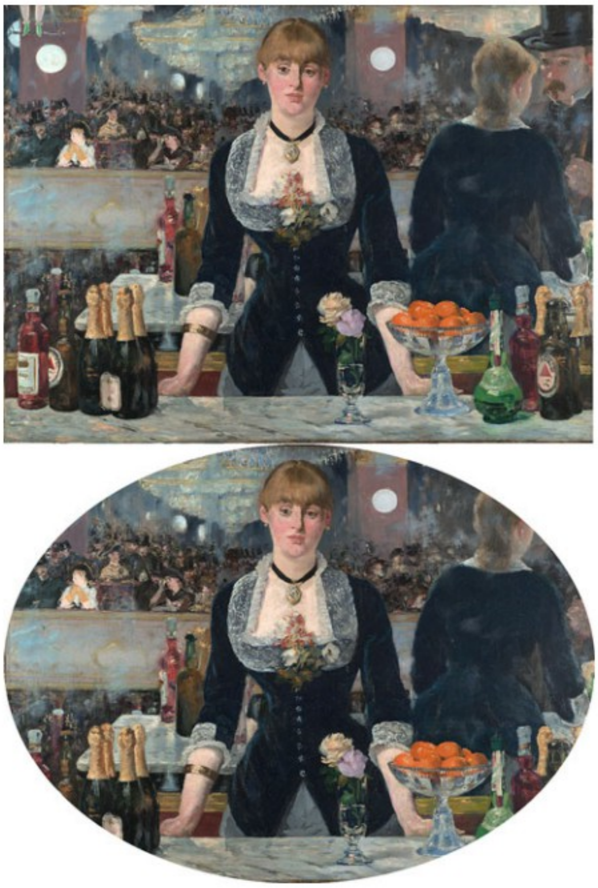

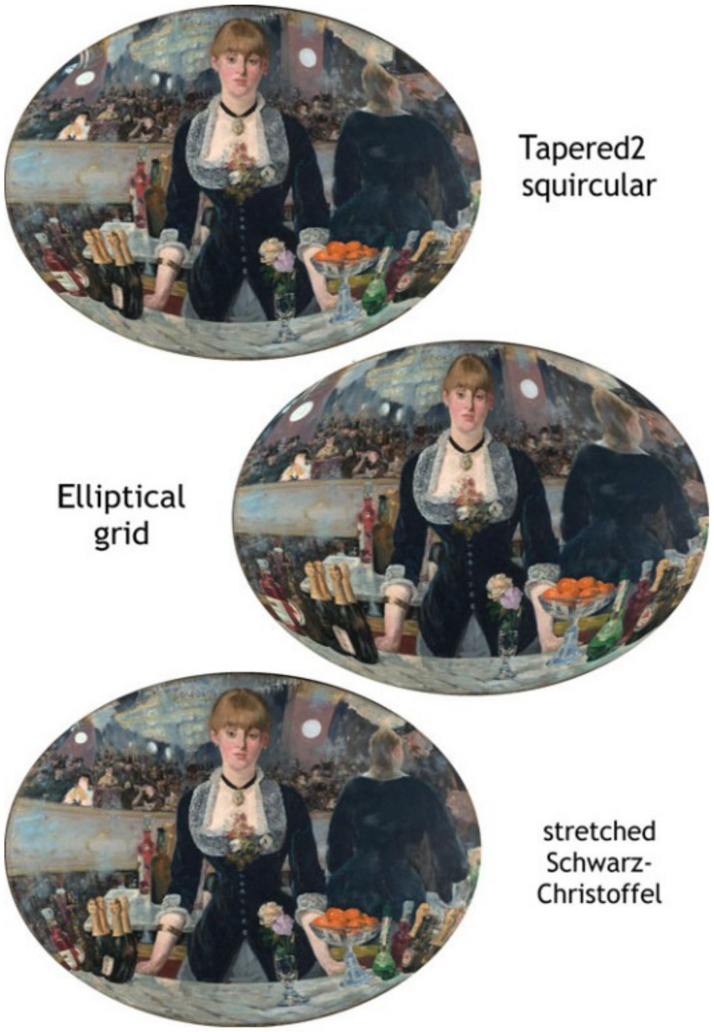
Tapered2
squircular
Elliptical
grid
stretched
Schwarz-
Christoffel

Fig. 24 Elliptification of Manet's famous impressionist painting

the different mappings discussed in this chapter. The traditional process of simply cropping the picture to produce an oval region is quite unacceptable because it removes many noteworthy features near the corners of the painting. For example, observe that in Manet's cropped painting, the bar patron with the top hat is gone. Also, the dangling legs of the trapeze artist at the top left corner of the painting are nowhere to be found. In contrast, the mathematically mapped paintings keep all these features intact, albeit distorted.

Size Versus Shape Distortions

In differential geometry, two-dimensional distortion is generally categorized into two major types: size and shape distortions (Floater and Hormann, 2005). The elliptification of Manet's painting highlights the difference between these two types of distortion. The stretched Schwarz-Christoffel based on the Conformal Square mapping has significant size distortions near the corner. This is quite evident by observing the dangling legs of the trapeze artist, which has shrunk to the point of being barely visible. In contrast, the Elliptical Grid mapping has significant shape distortions near the corner. Specifically, the gentleman with a top hat appears considerably deformed. On the other hand, the Tapered2 Squircular mapping offers a good compromise between size and shape distortions. It provides the best result among the three mappings for Manet's painting. In fact, the circularized chessboard shown in Fig. 1 also uses the Tapered2 Squircular mapping. One can observe that the size and shape fidelity of the corner rooks is reasonably intact.

Conclusion

This chapter discussed three explicit methods for mapping the circular disc to a square and vice versa. In addition, some artistic applications of these mappings were provided. For hyperbolic art, the Conformal Square mapping gives the best results. However, for other artistic applications such as elliptification, this is not necessarily the case because the mapping has sizeable distortions near the four corners.

Cross-References

- ▶ Geometric and Aesthetic Concepts based on Pentagonal Structures
- ▶ Mathematics and Art: Unifying Perspectives
- ▶ Tessellated, Tiled, and Woven Surfaces in Architecture
- ▶ The Beauty of Blaschke Products

References

Abramowitz M, Stegun I (1972) Handbook of mathematical functions. Dover Publications Inc., New York

Carlson B (1977) Special functions of applied mathematics. Academic, New York

Dunham D (2009) Hamiltonian paths and hyperbolic patterns. Contemp Math 479:51–65

Fernandez-Guasti M (1992) Analytic geometry of some rectilinear figures. Int J Math Educ Sci Technol 23:895–901

Floater M, Hormann K (2005) Surface parameterization: a tutorial and survey. In: Dodgson N, Floater M, Sabin M (eds) Advances in multiresolution for geometric modelling. Springer, New York, pp 157–186

Fong C (2014) Analytical methods for squaring the disc. In: Seoul ICM 2014

Fong C (2019) Elliptification of rectangular imagery. In: Joint mathematics meeting SIGMAA-ARTS

Frederick C, Schwarz E (1990) Conformal image warping. IEEE Comput Graph Appl 10(2):54–61

Hancock H (1958) Elliptic integrals. Dover Publications Inc., New York

Langer J, Singer D (2011) The lemniscatic chessboard. Forum Geometricorum 11:183–199

Nowell P (2005) Mapping a square to a circle (blog). http://mathproofs.blogspot.com/2005/07/mapping-square-to-circle.html

Press W, Flannery B, Teukolsky S, Vetterling W (1992) Numerical recipes in C. The art of scientific computing, 2nd edn. Cambridge University Press, Cambridge

Rice A, Brown E (2012) Why ellipses are not elliptic curves. Math Mag 85(3):163–176

Shirley P, Chiu K (1997) A low distortion map between disk and square. J Graph Tools 2:45–52

Taimina D (2009) Crocheting adventures with hyperbolic planes. AK Peters, Wellesley

2021 08 15 ///

2-Geral

Book1-016-2021/20210815/

/Handbook of the Mathematics of the Arts and Sciences/

3319570714.pdf

Pág. 123(144)

Homeomorphisms Between the Circular Disc and the Square.

Pág. 126(147)-

Fig. 3 Some mapping to convert a disc to a square and vice versa

---

Pág. 57(79)

Fig. 15 Isomorphisms between the hemisphere model, the Klein model, and the Poincaré models.

Pág. 58(80)/

Fig. 16 Lines in Poincaré disk model and the half plane model

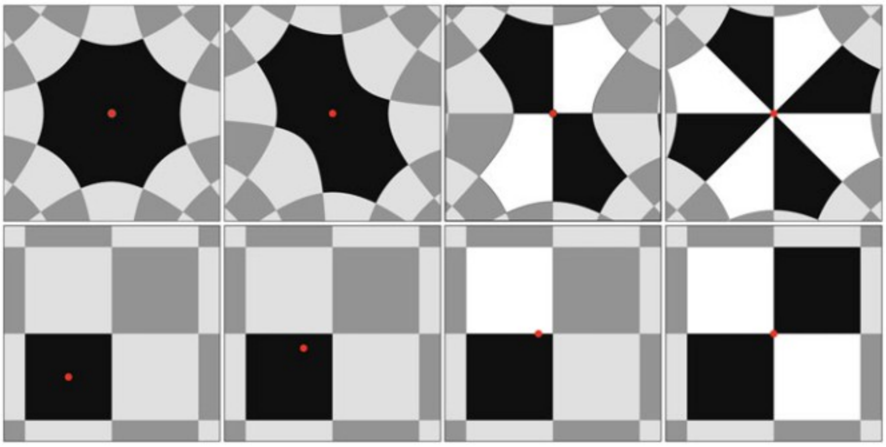

Fig. 13 Some prototypes of exceptional tiles (shown in black)

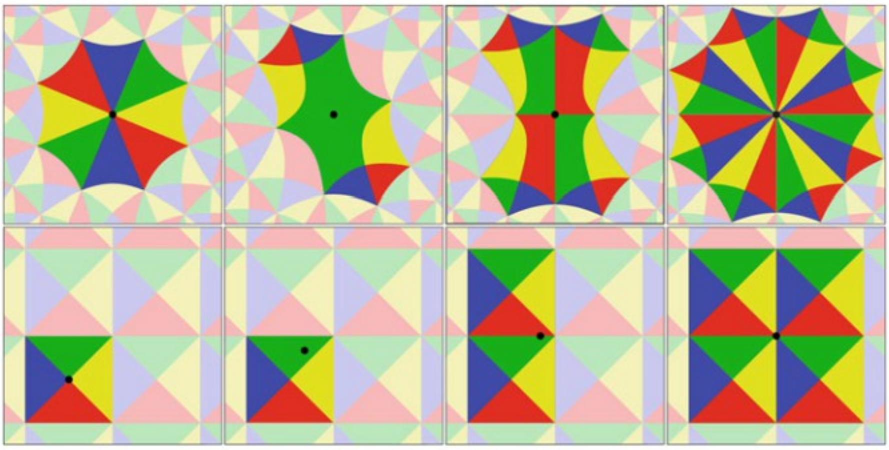

Fig. 14 Symmetries near critical points arising from subdivision of tiles

respectively, in Fig. 14. As in Fig. 13, the images in the upper row are constructed by pulling back the images in the lower row via $w = f(z)$.

Hyperbolic Geometry

One strategy used to prove the parallel postulate was to replace it by its negation and then to search for a contradiction. Carl Friedrich Gauss first realized that this search would be in vain and that, in fact, there is a geometry that satisfies all Euclidean axioms but with the fifth postulate negated. Although he had developed a fairly advanced *non-Euclidean geometry* in the year 1816, Gauss did not publish it and confined himself to occasional remarks in private letters (see Stäckel 1933 for details). The first publications about what is now called *hyperbolic geometry* are by Nikolai Ivanovich Lobachevsky (1829) and János Bolyai (1832). Yet, even later, whether or not this new and counterintuitive geometry might contain a contradiction remained an open question. These doubts could be dispelled by providing a Euclidean model for hyperbolic geometry; that is, as discussed in the introduction, the terms line and point, among others, are associated with certain objects from Euclidean geometry that are then shown to satisfy the axioms. It follows that non-Euclidean geometry is consistent if Euclidean geometry is.

The first model described by Eugenio Beltrami in 1868 was the *pseudosphere model* that represents only a portion of the hyperbolic plane and will therefore not be discussed further. That same year Beltrami first described the Poincaré and Klein models, even in n dimensions (see Milnor 1982). In two dimensions these models can be introduced most conveniently using projections from a sphere, as is visualized in Fig. 15.

(1) The points of the *hemisphere model* are the points lying strictly in the upper half of a unit sphere in $\mathbb{R}^3$; the lines are half-circles orthogonal to the equator. In Fig. 15, one such line is depicted in yellow. It can be verified easily that each pair of distinct points determines exactly one line that meets them both. Moreover, to a given line L and a point P not on L, there exist an infinite number of lines through P that do not intersect L, so that this geometry does not satisfy Playfair's axiom.

(2) The *Poincaré disk model* is obtained from the hemisphere model by stereographic projection from the south pole to the equator plane. Here points are identified with points in the open unit disk $\mathbb{D}$, and lines are either open arcs orthogonal to the unit circle or open diameters. In Fig. 15, points and lines in the Poincaré model are depicted in red.

Figure 16 on the left shows a few hyperbolic lines in the Poincaré model. Again, through two distinct points in $\mathbb{D}$, there is exactly one (hyperbolic) line. The endpoints on the unit circle, which are not part of the line, are referred to as *ideal points*. Hence, two different lines terminating at the same ideal point are

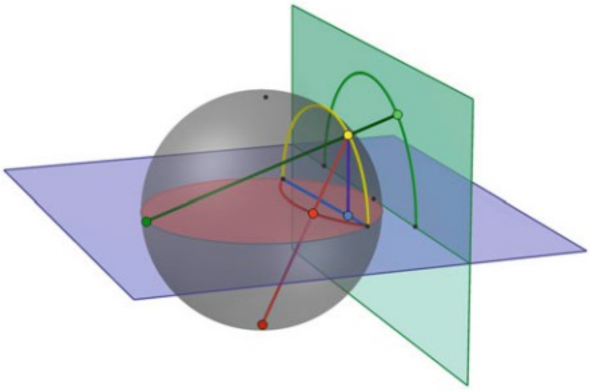

Fig. 15 Isomorphisms between the hemisphere model, the Klein model, and the Poincaré models

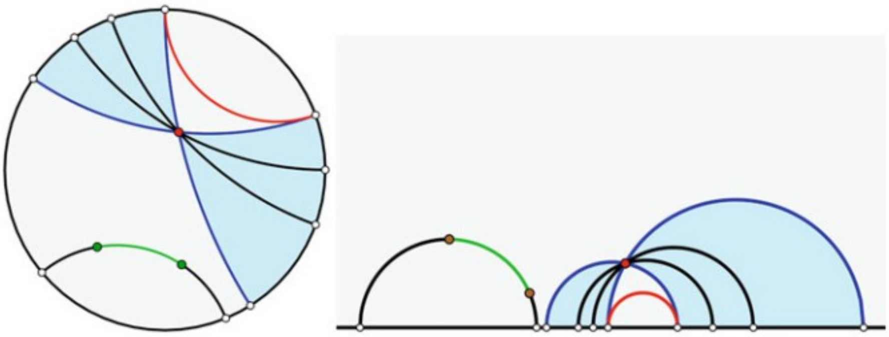

Fig. 16 Lines in the Poincaré disk model and the half-plane model

parallel because they have no common point in $\mathbb{D}$. This distinguishes them from *ultraparallel lines*, which do not even tend to common ideal points. Figure 16 also illustrates the fact that Playfair's axiom is violated.

(3) Projecting the hemisphere orthogonally onto the equator plane yields the *Klein model*, *Cayley-Klein model*, or *Beltrami-Klein model*. Its points are the points in the open unit disk, and its lines are open line segments connecting ideal points on the boundary of the unit disk. A (hyperbolic) line is depicted in blue in Fig. 15. Composing the two mappings just introduced yields an isomorphism between the Poincaré and the Klein model. Its explicit form in complex coordinates is

$$F(z) = \frac{2z}{1 + |z|^2}. \tag{1}$$

For more details about the Klein model, the reader is referred to Baldus (1944) and Greenberg (2007).

(4) If, as in Fig. 15, the projection is performed from a point on the equator of the Riemann sphere onto a plane tangent to the sphere at an opposite point, the model that lives in a half plane, usually identified with the upper half plane $\mathbb{H} := \{x + iy : y > 0\}$ of the complex plane, is obtained. This model is the *Poincaré half-plane model* in which lines are either open semicircles or open half-lines orthogonal to the real line; see Fig. 16, right. Elementary considerations also show a mapping that directly relates both Poincaré models. Such a mapping turns out to be the *Cayley transform*, a special Möbius transformation that maps $\mathbb{D}$ onto $\mathbb{H}$:

$$F(z) = \mathrm{i}\frac{1 - z}{1 + z}. \tag{2}$$

See, for example, the exposition in Anderson (2005), which is based on the Poincaré half-plane model.

Hilbert's axioms for plane geometry introduce only five undefined or *primitive terms* by the properties they are supposed to possess. These are "point," "line," "incidence," "betweenness," and "congruence." So, in the description of the models, all these terms must be addressed, not only points and lines. Incidence and betweenness have their usual Euclidean meaning. It is therefore possible to define a *line segment* $\overline{AB}$ as the collection of all points on the line through A and B that are located between A and B or coincide with one of the points. Such a line segment is colored in green in the two Poincaré models shown in Fig. 16.

It is also possible to define (hyperbolic) convexity: A set is *hyperbolically convex* if it contains all line segments between any two of its points. As in Euclidean geometry, the intersection of all convex sets containing a set M is again convex and will be called the *hyperbolic convex hull* of M.

The notion of congruence and length of line segments in the Poincaré model is more subtle, since it must be compatible with Hilbert's axioms. To this end, we define the *hyperbolic distance* between z and w in $\mathbb{D}$ by

$$d(z, w) = \log \frac{1 + \rho(z, w)}{1 - \rho(z, w)}, \quad \text{where } \rho(z, w) = \left| \frac{z - w}{1 - \overline{w}z} \right|. \tag{3}$$

The function ρ is called the *pseudo-hyperbolic distance*.

This formula has interesting consequences: If z is fixed and w tends to the boundary of the unit disk, their distance will go to infinity. In other words, the Poincaré disk is unbounded and ideal points have infinite distance from points inside the disk. If both z and w tend to $\partial\mathbb{D}$ while their Euclidean distance remains fixed, their hyperbolic distance will go to infinity. The other way around also works: An object of fixed size in the hyperbolic world that moves toward the boundary of the unit disk, (which is not a boundary of the hyperbolic disk) becomes tiny for an observer in the Euclidean plane.

The conformality of the Poincaré disk model and the Möbius map (2) imply that the Poincaré half-plane model is also conformal. On the other hand, the transition map (1) is not conformal. Therefore, the hyperbolic size of an angle between two lines in the Klein model is not equal to its Euclidean size.

The final relevant notion is that of *isometries*, that is, self-mappings of the hyperbolic plane that leave the length of line segments and, therefore, angles invariant. Isometries form a group under composition that acts on the hyperbolic plane. In the Poincaré disk model, there are two types of isometries: conformal self-maps of the unit disk, which are the special Möbius transformations that can be written in the form

$$b(z) = c \frac{z - z_0}{1 - \overline{z}_0 z} \tag{4}$$

with $|c| = 1$ and $z_0 \in \mathbb{D}$, and anti-conformal self-maps of $\mathbb{D}$, which can be written as the composition $b(\overline{z})$. Recall that $z \mapsto \overline{z}$ is the reflection in the real axis and it reverses orientation. A function of the form (4) is called a *Blaschke factor*.

When considered in the complex plane, b has a pole at $1/\overline{z_0}$, as can be seen on the left of Fig. 17. The image in the middle of the figure shows a Blaschke factor.

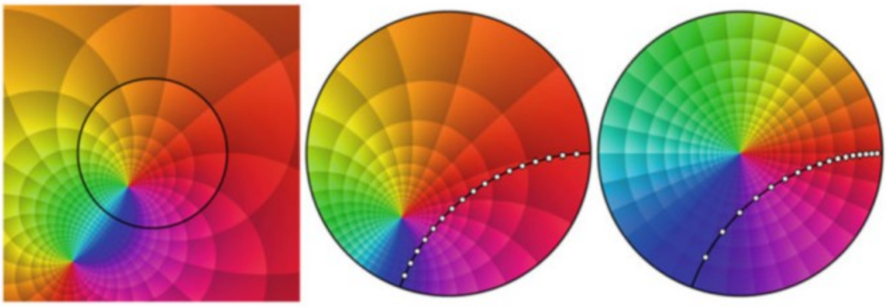

Fig. 17 A conformal self-map of the disk (domain on the left and in the middle, image at right)

This self-map of $\mathbb{D}$ is a *hyperbolic translation*; that is, there is a line that is invariant under b (which is the black line with the white dots in the middle and right images of the figure).

Blaschke Products

A *complex polynomial* p of *degree* n has the form $p(z) = a_n z^n + \cdots + a_1 z + a_0$ with complex coefficients $a_0, \ldots, a_n$, where $a_n \neq 0$ is assumed. It follows from the fundamental theorem of algebra that, for any $w \in \mathbb{C}$, the equation $p(z) = w$ has exactly n solutions, if they are counted according to multiplicities. Since p is also holomorphic, we can say that a polynomial is a self-map of $\mathbb{C}$ that maps $\mathbb{C}$ onto an n-fold covering of itself. In view of the fact that $\lim_{|z|\to\infty} |p(z)| = \infty$, polynomials have a continuous extension onto the Riemann sphere by defining $p(\infty) := \infty$. Recalling that $z \mapsto 1/z$ interchanges 0 and ∞, it is even possible to speak of the *multiplicity of* ∞, which is then defined to be the multiplicity of the zero of the analytic function $z \mapsto 1/p(1/z)$, and this multiplicity is easily seen to be equal to n. Hence, polynomials are also self-maps of the Riemann sphere of valency n. (It should be noted that they are not the only ones.)

What are the self-maps of the unit disk of valency n? The answer is that these are exactly the functions

$$B(z) = c \prod_{k=1}^{n} \frac{z - z_k}{1 - \overline{z}_k z} \tag{5}$$

with $|c| = 1$ and $z_1, \ldots, z_n \in \mathbb{D}$. These functions are called *Blaschke products* or *hyperbolic polynomials* of degree n. Blaschke products share many properties with polynomials. Again by the fundamental theorem of algebra, a polynomial of degree n can be factored as

$$p(z) = c\,(z - z_1) \cdots (z - z_n)$$

where $z_1, \dots, z_n \in \mathbb{C}$ are its zeros, each repeated according to its multiplicity. Recalling that conformal self-maps of $\mathbb{C}$ are of the form $z \mapsto az + b$, it follows that polynomials of degree n are exactly those functions that can be written as a product of n conformal self-maps of $\mathbb{C}$. Comparing the conformal self-maps of $\mathbb{D}$ appearing in (4) with the general form of a Blaschke product of degree n as in (5), it follows that Blaschke products are exactly those functions that can be written as a product of n conformal self-maps of $\mathbb{D}$.

The derivative of a polynomial p of degree n is a polynomial of degree $n - 1$ and therefore has $n - 1$ zeros, called *critical points* of p. The derivative of a Blaschke product B of degree n is not a Blaschke product, but it has $n - 1$ critical points in $\mathbb{D}$ (counting, of course, possible multiplicities). If the points $\zeta_1, \dots, \zeta_{n-1} \in \mathbb{C}$ are prescribed, and a polynomial p with these points as critical points is to be constructed, it is apparent that $p'(z) = a(z - \zeta_1) \cdots (z - \zeta_{n-1})$ with $a \neq 0$ arbitrary, and thus

$$p(z) = \int p'(z)\, dz = a \int (z - \zeta_1) \cdots (z - \zeta_{n-1})\, dz + b;$$

that is, p exists and is uniquely determined up to post-composition with a conformal self-map $z \mapsto az + b$ of the plane $\mathbb{C}$. The analogous property for hyperbolic polynomials is a nontrivial result that was first proved by Heins: Given $n - 1$ points in $\mathbb{D}$, there is a Blaschke product B of degree n with these points as critical points, Heins (1962, 1986). The function B is unique up to post-composition with conformal self-maps of $\mathbb{D}$. Semmler and Wegert (2019) gave a simple proof using equilibriums of charge configurations. For further references on this topic, see Kraus and Roth (2008, 2013) and Garcia et al. (2018, Chapter 6).

Blaschke products live most naturally in the Poincaré disk model because of their conformal character. This is demonstrated by the following theorem due to Walsh, which is the hyperbolic counterpart of the well-known Gauss-Lucas theorem, Walsh (1950, 1952). The Gauss-Lucas theorem states that the critical points of any complex polynomial lie in the convex hull of its zeros. Walsh's theorem says that the critical points (in $\mathbb{D}$) of a hyperbolic polynomial (i.e., a Blaschke product) lie in the hyperbolic convex hull of its zeros (for an alternative proof, see Wegert 2011).

In the picture on the left of Fig. 18, the 10 zeros of a Blaschke product are given (black points), along with their hyperbolic convex hull and the critical points in $\mathbb{D}$ of the Blaschke product (gray points).

In fact, from Formula (5) it is possible to define the Blaschke product B not only in the unit disk but also on the Riemann sphere. In this case, B is a rational function that maps the Riemann sphere onto itself. More precisely, each of the upper hemisphere, the lower hemisphere, and the equator is mapped onto an n-fold covering of itself, respectively. In particular, a Blaschke product of degree n maps $\mathbb{D}$ onto itself n times, and it wraps the unit circle around itself n times with strictly monotone argument.

The Blaschke product B also shows some nice symmetry: The value $B(1/\overline{z})$ at a point $1/\overline{z}$, which is the reflection of z in the unit circle, is the reflection of $B(z)$ in

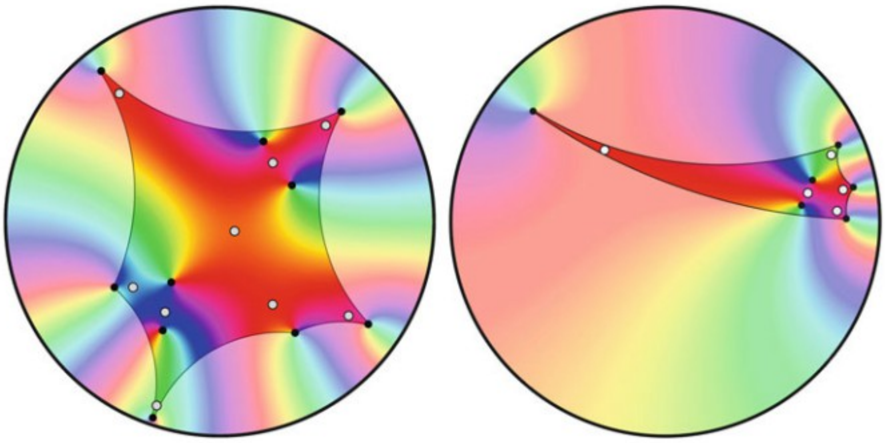

Fig. 18 Walsh's theorem on the location of critical points of a Blaschke product

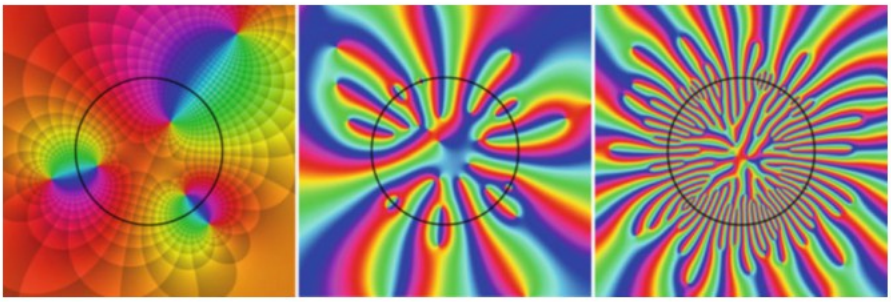

Fig. 19 Blaschke products of degrees 3, 20, and 70 in the square $|\operatorname{Re} z| \leq 2$, $|\operatorname{Im} z| \leq 2$

the unit circle,

$$B(1/\overline{z}) = 1/\overline{B(z)}, \qquad z \in \mathbb{C}. \tag{6}$$

This has the consequence that the zeros z_k of B inside the unit circle correspond to poles of B at the reflected points $1/\overline{z_k}$ outside the unit disk. This is illustrated in Fig. 19. (Not all poles can be seen because z is restricted to the square $|\operatorname{Re} z| \leq 2$, $|\operatorname{Im} z| \leq 2$.)

The perfect symmetry induced by relation (6) can be seen more readily by viewing the Blaschke product on the Riemann sphere, as shown in Fig. 20. Since reflection reverses orientation, the orientation of the color spectrum in the upper and lower hemisphere is reversed, and the zeros in the upper hemisphere appear as poles in the lower hemisphere.

This brief discussion demonstrates the surprising analogy between Blaschke products and polynomials, justifying the name hyperbolic polynomials. More substantial results in this direction can be found in Ng and Tsang (2013).

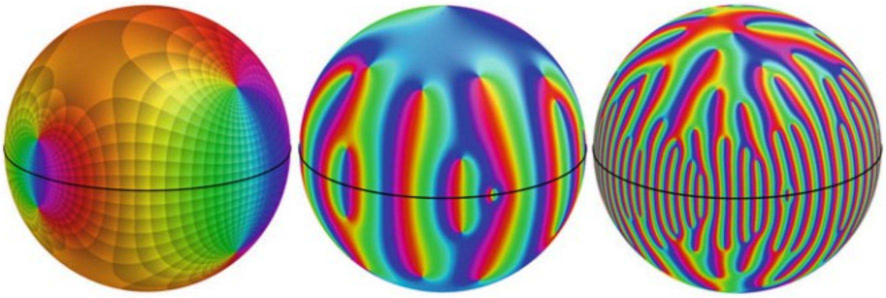

Fig. 20 Blaschke products of degrees 3, 20, and 70 on the Riemann sphere

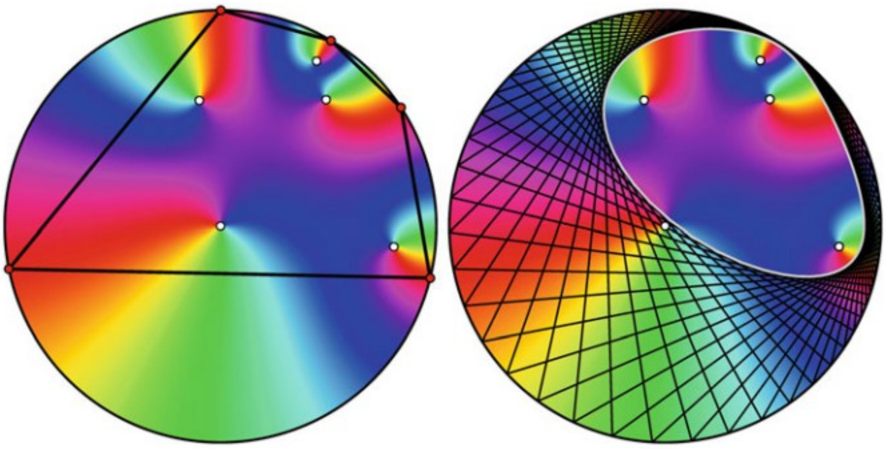

Fig. 21 A Blaschke product of degree 5 with some inscribed polygons

Blaschke Products and Ellipses in the Euclidean Plane

Recall from the previous section that a Blaschke product of degree n wraps the unit circle, $\partial\mathbb{D}$, around itself n times. Therefore, given any color, there will be exactly n points on the unit circle in the phase plot that have this color; in other words, these points are mapped to the same $\lambda \in \partial\mathbb{D}$. Now draw the convex n-gon with these points as vertices. In Fig. 21, this process is applied once in the figure on the left and then repeated on the right, moving λ around the unit circle. As this happens, the envelope of the n-gons obtained produces a closed curve in the unit disk. The phase portrait of the Blaschke product appears to have no obvious symmetries, but what about the curve? Can the curve be described in a nice way? Does it possess symmetries? And, if so, do these symmetries say anything about the Blaschke product?

To form a *Blaschke curve*, points of the same color on the unit circle are joined to their nearest neighbors using line segments. The envelope of these line segments yields the curve. The curve appearing in Fig. 21 is obtained from a Blaschke product of degree 5, and, perhaps as expected, it does not have any obvious symmetry. Here it is assumed that $B(0) = 0$. If this is not the case, it is possible to choose a Blaschke factor M mapping $B(0)$ to 0 and then to consider the Blaschke product defined by $B_1 := M \circ B$. Now B_1 and B are of the same degree, $B_1(0) = 0$, and the curve obtained from B_1 is the same as the one obtained from B. One thing is surprising though: The four nonzero zeros of the Blaschke product (the white dots not in the center) lie inside the curve. It is not difficult to show that if $B(0) = B(a) = 0$ and $a \neq 0$, then for each $\lambda \in \partial\mathbb{D}$ the point a lies in the convex hull of the points z for which $B(z) = \lambda$, Gorkin et al. (1994, Proposition 4.8).

Consider the Blaschke product $B(z) = \mathrm{i}\, z\,(z - 1/2)\,/\,(1 - z/2)$. In the picture of Fig. 22, there are two red points on the unit circle and these are the two points z_1 and z_2 for which $B(z_1) = B(z_2) = 1$. The result of connecting these two points is the picture on the left.

Continue connecting the two points on the circle of the same color and drawing the line segments; something unexpected happens as a result of this process. Looking at the picture on the right in Fig. 22, all line segments seem to intersect in one point, namely, the point $1/2$, the nonzero zero of B.

This is always true: Let B be a Blaschke product of degree 2 with one zero at 0 and one zero at a point a in $\mathbb{D}$, and consider the line segment joining the two points $z_{1,\lambda}$ and $z_{2,\lambda}$ for which $B(z_{1,\lambda}) = B(z_{2,\lambda}) = \lambda$. Then for all λ on the unit circle, these line segments will pass through the point a. In this simplest case, the enveloping curve will degenerate to a point. While this might be considered a symmetric object, it is not a very interesting one. The obvious question now is: What happens if B has zeros 0, a, and b and the points B identifies on the unit circle are connected?

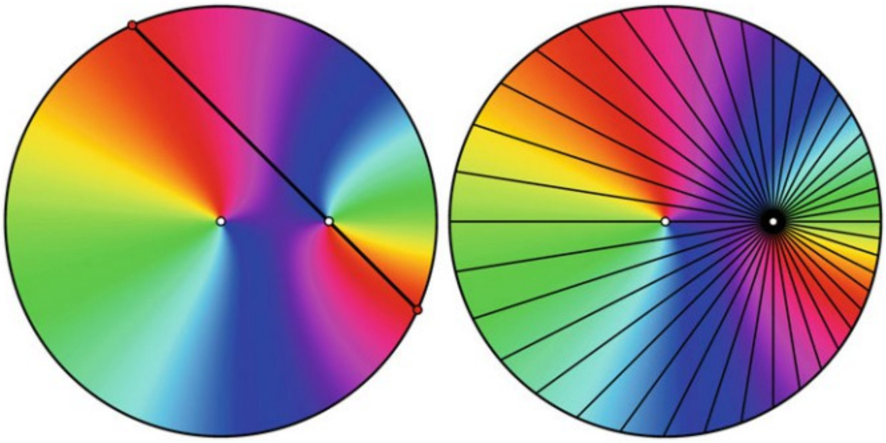

Fig. 22 Line segments with points on the unit circle identified by a Blaschke product of degree 2

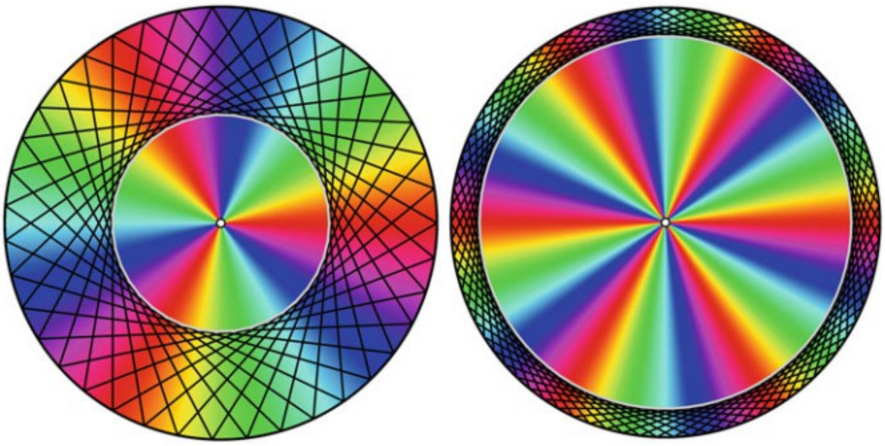

Fig. 23 The circles associated to $B_1(z) = z^3$ and $B_2(z) = z^6$

Start with the simplest degree-3 Blaschke product, namely, $B(z) = z^3$. As in Fig. 23, the points B identifies are equally spaced on the unit circle, and it follows from this that the curve obtained from this process is a circle, a perfectly symmetric curve. In fact, any Blaschke product of the form cz^n for $|c| = 1$ and a positive integer $n \geq 3$ will be associated with a circle centered at the origin. The bigger the n is, the larger the radius of the circle. This is illustrated by the picture on the right of Fig. 23.

Now consider the case in which the degree-3 Blaschke product has a nonzero zero. For example, suppose the two zeros are $a = 0.8 - 0.1\mathrm{i}$ and $b = 0.5 + 0.5\mathrm{i}$.

In Fig. 24, the three points B maps to the point 1 are connected to obtain the single triangle appearing in the picture on the left. For the picture on the right, 20 such triangles are formed. As before, a and b lie inside all the triangles. But more seems to be true here.

In fact, it seems that every triangle circumscribes an ellipse and that the foci of the ellipse are located at these two zeros of B. A classical theorem explains why such a result might be expected: Siebeck's theorem says that given three noncollinear points z_1, z_2, and z_3, the zeros of the function

$$F(z) = \frac{m_1}{z - z_1} + \frac{m_2}{z - z_2} + \frac{m_3}{z - z_3}, \text{ where } m_1, m_2, m_3 > 0$$

are the foci of the ellipse inscribed in the triangle formed by z_1, z_2, z_3. Applying Siebeck's theorem to the function

$$F_\lambda(z) = \frac{B(z)/z}{B(z) - \lambda}, \text{ for } \lambda \in \partial\mathbb{D},$$

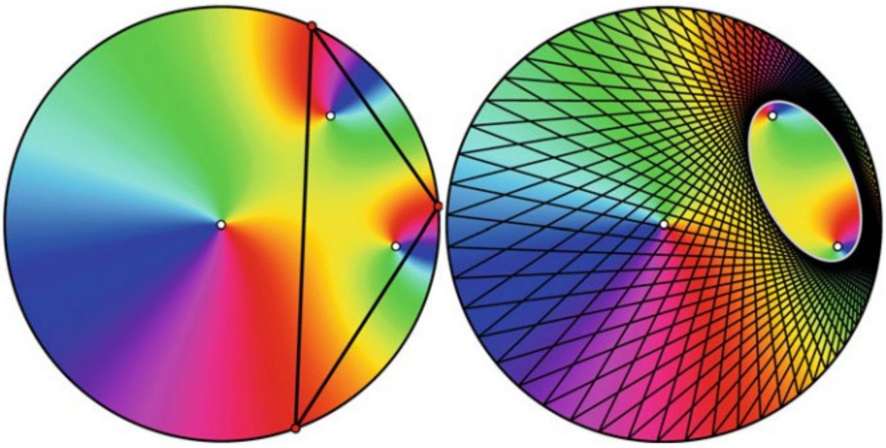

Fig. 24 A Blaschke product of degree 3

one sees that for each $\lambda \in \partial\mathbb{D}$, the triangle with vertices at the three points B maps to λ circumscribes an ellipse with foci at the zeros of $B(z)/z$. But these ellipses may change as λ changes. So the question is: Why do these triangles always circumscribe the same ellipse? For the answer, see Daepp et al. (2002, 2018). For Blaschke products of degree 3, then, the result is a well-known curve with two axes of symmetry, and the two foci of the curve are the zeros of $B(z)/z$.

The next case to consider is, of course, degree 4. Here, when the convex quadrilaterals with vertices at the points the Blaschke product identifies are formed, the result is a curve – but not a familiar one, in general. However, there is a very special case in which some symmetry appears: By a theorem of Fujimura, the associated curve of a degree-4 Blaschke product is an ellipse if and only if the Blaschke product is the composition of two degree-2 Blaschke products, Fujimura (2013) (see also Gorkin and Wagner 2017). For example, the picture on the left of Fig. 25 is associated with a Blaschke product that is a composition, $B = D \circ C$, where

$$C(z) = z\frac{z - (0.2 - 0.5\mathrm{i})}{1 - (0.2 + 0.5\mathrm{i})z} \quad \text{and} \quad D(z) = z\frac{z - (0.5 + 0.3\mathrm{i})}{1 - (0.5 - 0.3\mathrm{i})z}.$$

On the right, an asymmetrical picture associated with a degree-4 Blaschke product outlining a clearly non-elliptical form appears, and Fujimura's result says that the Blaschke product cannot be written as a composition of two nontrivial Blaschke products.

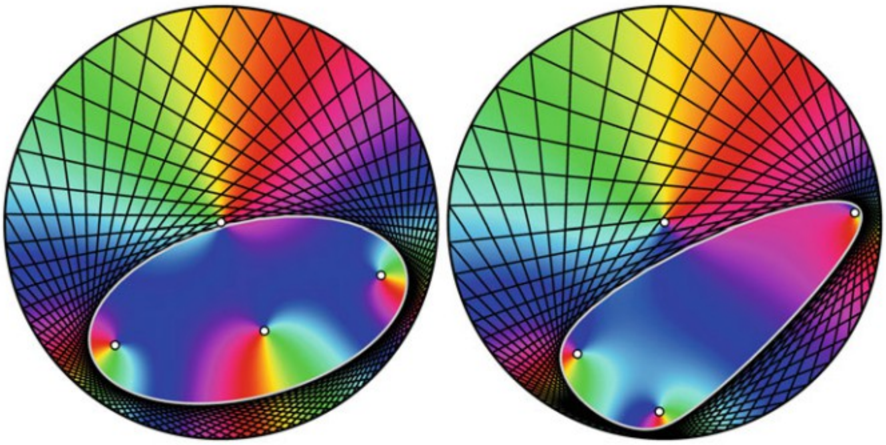

Fig. 25 Curves associated to degree-4 Blaschke products

Blaschke Products and Ellipses in the Poincaré Disk Model

So far, the focus has been on the envelope of the line segments joining the points that the Blaschke product B identifies. But Blaschke products are hyperbolic polynomials, so it is natural to view this in terms of hyperbolic geometry. In particular, what happens when the Euclidean lines of the previous section are replaced with hyperbolic lines in the Poincaré disk model?

We again start with the simplest case, a Blaschke product B of degree 2. The ideal points on the unit circle that are identified by B are now joined by the hyperbolic line of the Poincaré disk model. These lines all intersect in one point, but this time they intersect at the critical point of B; see Fig. 26. In the Euclidean model, when $B(0) \neq 0$, a Blaschke factor could be used to move points so that zero mapped to zero. In the hyperbolic case, there is no need to require that $B(0) = 0$, because the critical points are invariant under post-composition by a Blaschke factor.

Formula (3) for distance in the Poincaré disk model was introduced in the section entitled *Hyperbolic Geometry.* In view of this formula, it is reasonable to ask what a *hyperbolic circle* with center $a \in \mathbb{D}$ is. The answer is that it is the set of points z for which $d(z, a) = r$, for a constant r with $r > 0$ – the same as in the Euclidean setting, but using a different distance function. Since $d(z, a) = r$ if and only if $\rho(z, a) = s$ for some s with $0 < s < 1$, a hyperbolic circle with center a and ρ-radius s is

$$D_\rho(a, s) := \{z : \rho(z, a) = s\}.$$

If $a = 0$, then (as the reader should check) the hyperbolic circle centered at zero and of radius r is also a Euclidean circle with center zero and Euclidean radius s, where $s = (e^r - 1)/(e^r + 1)$. Something like this is true in general: Consider $D_\rho(a, s)$ and the Blaschke factor defined by

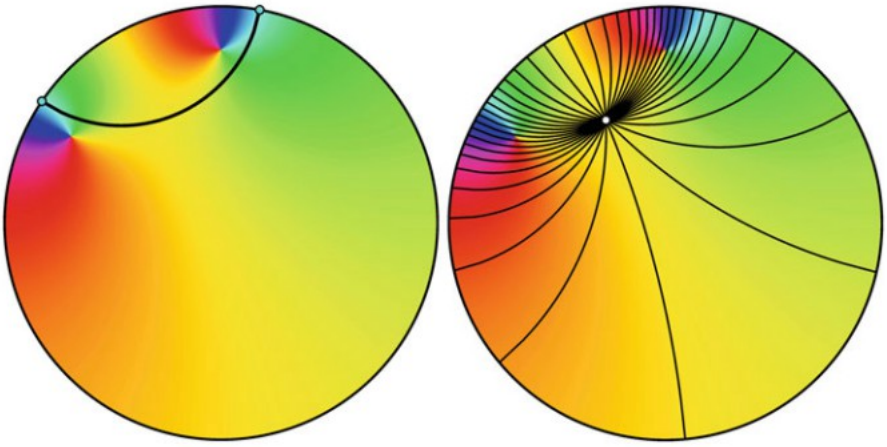

Fig. 26 Hyperbolic lines with ideal points identified by a degree-2 Blaschke product

$$M_a(z) = \frac{z-a}{1-\overline{a}z}$$

that maps a to 0. Since M_a preserves hyperbolic circles and distances, M_a maps the circle $C_\rho(a, s)$ to the circle $C_\rho(0, s)$. But M_a also preserves Euclidean circles and so does M_a^{-1}. Therefore $C_\rho(a, s) = M_a^{-1}(C_\rho(0, s))$ is also a Euclidean circle. Thus, a hyperbolic circle is also a Euclidean circle (and conversely), though the center and radius are not usually the same. Formulas for the center and the radius can be found in Garnett (1981). The Poincaré disk model bounded by the black unit circle with five blue circles of hyperbolic radius 1 appears in Fig. 27. The blue hyperbolic centers are located at five distinct points, namely

$$\left(1-\frac{1}{2^n}\right)\mathrm{e}^{(2n-1)\pi\,\mathrm{i}/4} \qquad \text{for } n = 0, \dots, 4.$$

The black dots are the Euclidean centers of the circles, and it should be noted that the Euclidean and hyperbolic centers of the circle centered at the origin coincide. The curve in the figure on the right is also a hyperbolic circle produced by the Blaschke product

$$B(z) = \left(\frac{z-(1+\mathrm{i})/5}{1-z(1-\mathrm{i})/5}\right)^3.$$

In the previous section, the focus was on ellipses (Euclidean ellipses). It is time to consider ellipses again, but this time in the hyperbolic setting. A *hyperbolic ellipse* is the set of points such that the sum of the distances to two fixed points a_1 and a_2

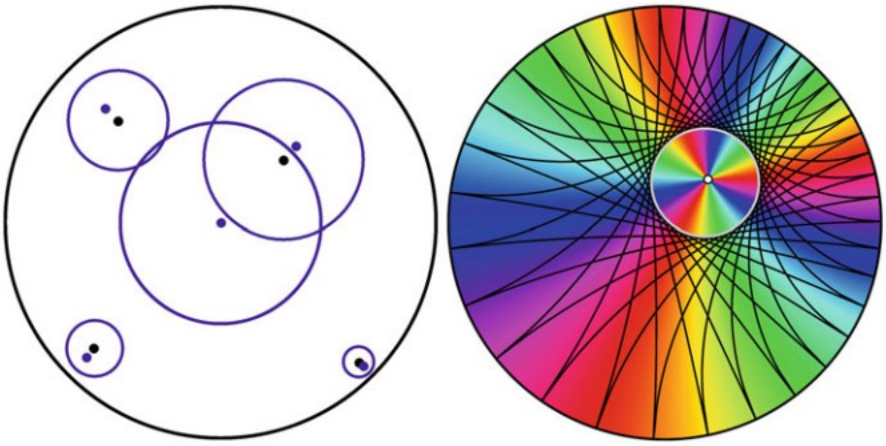

Fig. 27 Circles in the Poincaré disk model

is a constant. While this may sound familiar, the point is that the distance is now the hyperbolic distance.

Recall that lines in this model (the Poincaré disk model) are represented by circular arcs (geodesics) inside the unit circle that are orthogonal to the circle at their points of intersection. In the Euclidean model, starting with a Blaschke product of degree 3 with a zero at zero and connecting the points it identified, the resulting triangle circumscribed an ellipse. Consider now the hyperbolic version of this: Connect the three points a Blaschke product of degree 3 (not necessarily with a zero at 0) identifies with geodesics, forming three parallel lines. Is the result interesting? Indeed it is. On the left of Fig. 28, the ideal points that are identified by a Blaschke product of degree 3 are connected by three parallel lines. On the right, these lines envelope a curve when the procedure is repeated with other points on the unit circle.

In fact, Singer showed that if γ denotes the curve in $\mathbb{D}$ that is the envelope of the non-Euclidean geodesics joining points a Blaschke product B of degree n identifies, then γ is part of an algebraic curve for which the real foci in $\mathbb{D}$ are the critical points of B in $\mathbb{D}$ (together with their inverses with respect to $\mathbb{D}$), Singer (2006). If the Blaschke product is of degree 3, then the curve γ is a non-Euclidean ellipse, and the geometric foci are the two algebraic foci in $\mathbb{D}$, which are the two critical points of B in $\mathbb{D}$.

The curve on the right in Fig. 28 is thus a hyperbolic ellipse with foci marked in the figure. Our eye, trained by Euclidean geometry, is no longer able to recognize the symmetry of this curve because the distance function seems “distorted”; thus, to understand this, it is necessary to expand the traditional vision (i.e., Euclidean vision) of an ellipse.

We now also have a better understanding of the circle on the right of Fig. 27. The curve associated with this Blaschke product is a hyperbolic ellipse, but the two critical points coincide, and thus the ellipse becomes a hyperbolic circle. This, however, is also a Euclidean circle, as discussed above. It may be difficult to see in the picture, but the zero of the Blaschke product, which is also the critical point, is not in the Euclidean center of this circle. It is the circle's center with respect to the Poincaré disk model.

Figure 29 shows curves that correspond to Blaschke products of degrees 4 and 7, respectively. These are no longer ellipses, but the critical points of the Blaschke products that lie in the Poincaré disk are real foci of these curves. There is another piece of the curve and other critical points outside the disk; the relation between the points is that those outside are the symmetric points, with respect to the unit circle, of those inside.

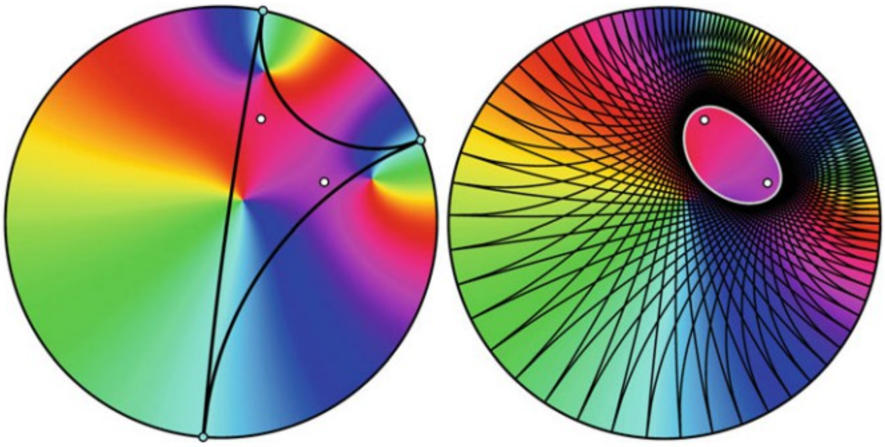

Fig. 28 A Blaschke product of degree 3 in the Poincaré disk model

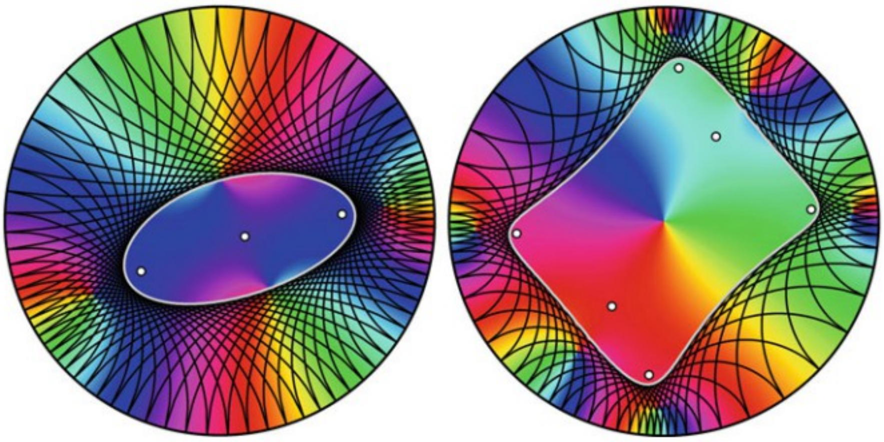

Fig. 29 Curves associated to Blaschke products of degree 4 and 7 in the Poincaré model

Compositions of Blaschke Products

This section demonstrates how visualization can provide new insight into the algebraic structure of mathematical objects. Such structures arise, for instance, from the factorization of natural numbers, which is the source of the rich field of number theory. In contrast to this, factorization of Blaschke products in the usual sense is trivial. However, there is another algebraic operation that gives rise to a number of challenging questions – this is the *composition* $g \circ f$ of two functions f and g, defined by $(g \circ f)(z) := g(f(z))$. It can easily be verified that the composition $B = g \circ f$ of Blaschke products f and g of degree m and n, respectively, is a Blaschke product of degree mn, so composition is an operation on the class of (finite) Blaschke products.

In analogy to multiplication of integers, it is easy to build a composition – but to decide whether or not a given function is composite, and to decompose it into its "factors," is a nontrivial task. There has been a great deal of interest in this problem: For polynomials a breakthrough appears in Ritt (1922). For a brief account of the history of the composition problem for Blaschke products, the reader, is referred to Cowen (2012). One approach to the problem will be described in detail below. For a different, but related, approach, see Chalendar et al. (2018).

Figure 30 shows enhanced phase plots of two Blaschke products. Knowing that one is a composition and the other one is not, it is not too difficult to guess which is which: A composition should have more structure than an arbitrary function – and looking at the phase plots with some patience, one finds that the candidate on the left is the better choice.

In fact, the relevant structure is encoded in the exceptional tiles that have more than 4 corners (see the section *Seeing Complex Functions*). The reason for this is

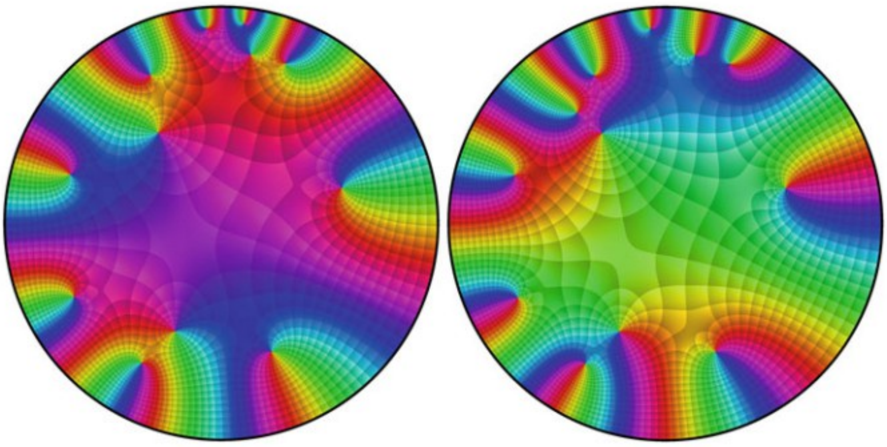

Fig. 30 Which of these phase plots represents a composition of Blaschke products?

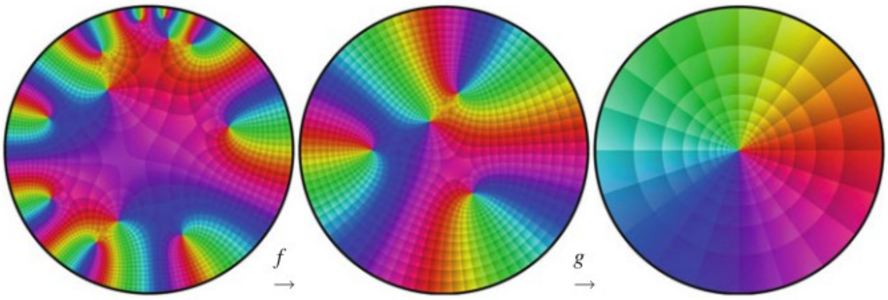

Fig. 31 Construction of the phase plot of a composition $g \circ f$, to be read from right to left

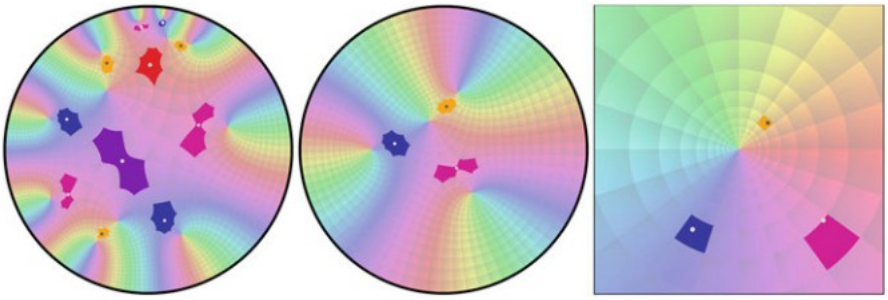

Fig. 32 Generation of critical tiles in the phase plots of $g \circ f$ and g from tiles in the z plane

explained in Fig. 31, which illustrates the composition of two Blaschke products. Starting with a standard polar pattern on the right-hand side, the phase plot of g in the middle is its pullback via the mapping g. The $n-1$ critical points of g are located in the exceptional tiles. The phase plot of $g \circ f$ on the left is the pullback of the image in the middle via f. Its exceptional tiles are of different origin: Tiles of the first type contain the critical points of f – they are pre-images of regular tiles of the plot in the middle – while tiles of the second type are generated as pullback of exceptional tiles in the phase plot of g. Since the image of f is an m-fold covering of the disk, each of the latter tiles is replicated m times.

As a consequence, the exceptional tiles in the phase plot of $g \circ f$ form $n-1$ groups, each consisting of m conformally equivalent tiles, and the remaining $m-1$ tiles contain the critical points of f (see Fig. 32). The arithmetic background of this observation is the chain rule

$$(g \circ f)' = (g' \circ f) \cdot f'.$$

The explanations above are somewhat simplified and have to be modified if f or g have multiple critical points or if their critical points and critical values are in special positions. Working out the details yields the following decomposition result.

A finite Blaschke product B is decomposable as $B = g \circ f$ with Blaschke products f and g of degree $m \geq 2$ and $n \geq 2$, respectively, if and only if the critical points of B can be partitioned into multisets[1] $A_0, A_1, \dots, A_{n-1}$ such that:

(i) The set A_0 contains $m - 1$ elements, and each set $A_1, \dots, A_{n-1}$ contains m elements.
(ii) Two critical points of B have the same multiplicity whenever they belong to the same set A_k for some $k = 1, \dots, n - 1$,
(iii) Let f_0 be (one and then any) Blaschke product of degree m with A_0 as set of critical points. Then f_0 is constant on each A_k for $k = 1, \dots, n - 1$.

If these conditions are satisfied, then B can be decomposed as $B = g_0 \circ f_0$, and the general form of such decompositions is

$$B = (g_0 \circ h^{-1}) \circ (h \circ f_0)$$

with a conformal disk automorphism h, Daepp et al. (2015).

In many cases the determination of candidates for the sets $A_1, \dots, A_{n-1}$ is rather easy: If $B = g_0 \circ f_0$ with f_0 satisfying condition (iii), then B must also be constant on each set, say $B(z) = c_k$ for all $z \in A_k$. Typically, these $n - 1$ critical values c_k of B are distinct, and they are also different from the critical values of B attained at the critical points of f_0. Thus, the splitting of critical points into the sets A_k corresponds to the values of B at these points. However, one should be aware that the existence of such a partitioning is a necessary condition that does not by itself guarantee that B is decomposable. Moreover, there are (exceptional) cases in which B also attains the critical value c_k at critical points not belonging to A_k, for instance, at points in $A_j \neq A_k$ or at critical points of f_0. This happens, for example, if g is itself a composition.

Verification of the crucial third condition relies on Heins' theorem on the existence and uniqueness of a Blaschke product f_0 of degree m with prescribed set A_0 of $m - 1$ (not necessarily distinct) critical points in $\mathbb{D}$ (see the section *Blaschke products*). The algorithmic aspects of this result are quite challenging, but in the case at hand, this can be circumvented in the following way: Since the set of Blaschke products (of any fixed degree) is invariant with respect to pre- and post-composition with conformal automorphisms h of the disk, one can assume that $B(0) = 0$ and then replace a (possible) composition $B = g \circ f$ by $B = (g \circ h^{-1}) \circ (h \circ f) = g_0 \circ f_0$, such that also $f_0(0) = 0$. This implies $g_0(0) = 0$. Hence, if $f_0(z_0) = 0$ at a point $z_0 \in \mathbb{D}$, then $B(z_0) = 0$. Consequently f_0 must be a sub-product of B, and it is only necessary to test the third criterion (iii) against all such f_0.

---

[1] In contrast to sets, multisets may contain elements repeatedly.

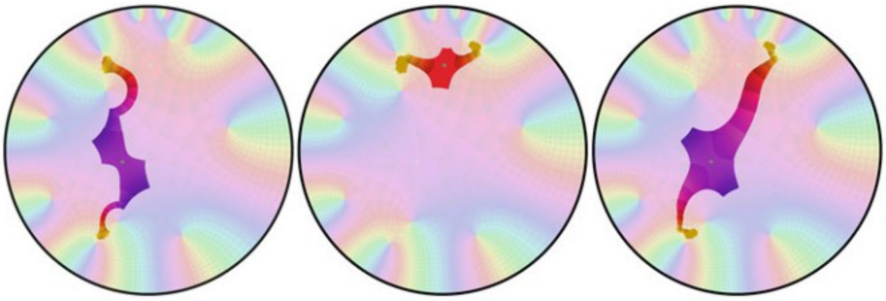

Fig. 33 Some hidden symmetries in the phase plot of a composition

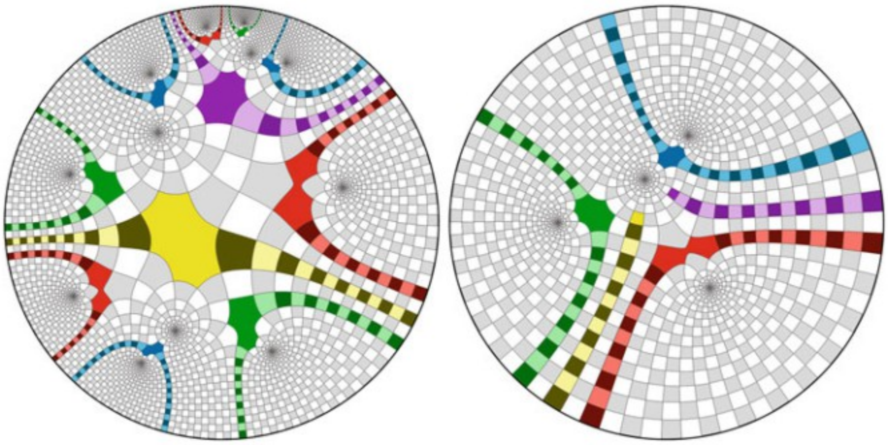

Fig. 34 Discrete conformal analysis of Blaschke products revealing symmetries

The governing principle behind these observations is a kind of hidden symmetry that is inherent in the compositions of Blaschke products and can be read off from their enhanced phase plots. Such symmetries are visualized in Fig. 33, which illustrates the fact that the critical points of $B = g \circ f$ belonging to the same set A_k (see the decomposition result) are, in a sense, symmetric with respect to the critical points of f. The reason for the appearance of these symmetries becomes plausible in Fig. 34, which shows "polar chessboard plots" of $g \circ f$ (left) and g (right). The yellow and the violet exceptional tiles contain the $m - 1 = 2$ critical points of f. Each of these belongs to a colored chain that connects opposite edges of that tile with the boundary of the disk. These two chains split the disk into $m = 3$ subdomains S_1, S_2, S_3, and f maps each subdomain onto (a copy of) $\mathbb{D}$ (shown on the right). These mappings are injective, except for the (colored) tiles of the bounding chains, in which two symmetric tiles have the same image. The exceptional (yellow and violet) tiles are mapped onto (double coverings) of the corresponding terminal tiles of the (yellow and violet) chains in the image on the right.

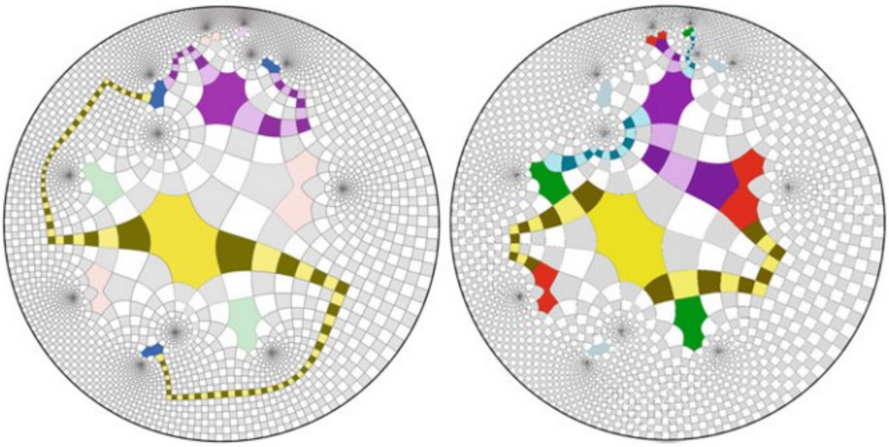

Fig. 35 Symmetric paths connecting tiles associated with the sets A_1 (left) and A_2 (right)

Each of the subdomains S_j contains $n - 1 = 2$ exceptional tiles (red, green, blue) that are the pre-images of the (correspondingly colored) critical tiles in the chessboard plot of g with respect to the mapping f. Finally, each group $A_1, \dots, A_{n-1}$ of critical points in the decomposition result is represented by exactly one member in each subdomain S_j. The color of the corresponding exceptional tile indicates the group to which its critical point belongs.

Polar chessboard plots are also convenient for finding the relevant symmetries, as illustrated in Fig. 35. This figure shows several chains of tiles that are symmetric with respect to the critical points of f (contained in the yellow and violet exceptional tile) and connect exceptional tiles (red, green, and blue) that are associated with the same set A_1, A_2, and A_3, respectively. If such symmetries are absent at some level of discretization, the Blaschke product is indecomposable. Theoretically, it would also work the other way around, provided that one could verify these symmetries exactly. This requires two things: (1) finding appropriate symmetric paths connecting the exceptional tiles and (2) verifying that the connected critical tiles are conformally equivalent. The first step is basically a counting exercise, but the second is impracticable. On the other hand, if a Blaschke product is not a composition, this can always be detected by choosing a sufficiently fine discretized chessboard plot.

Conclusion

The goal of this chapter has been to demonstrate how visualization can help to understand concepts and results of a mathematical theory without recourse to formulas. In fact, many mathematicians consider their research to be one of discovery, and seeing the objects of their investigations can be a source of

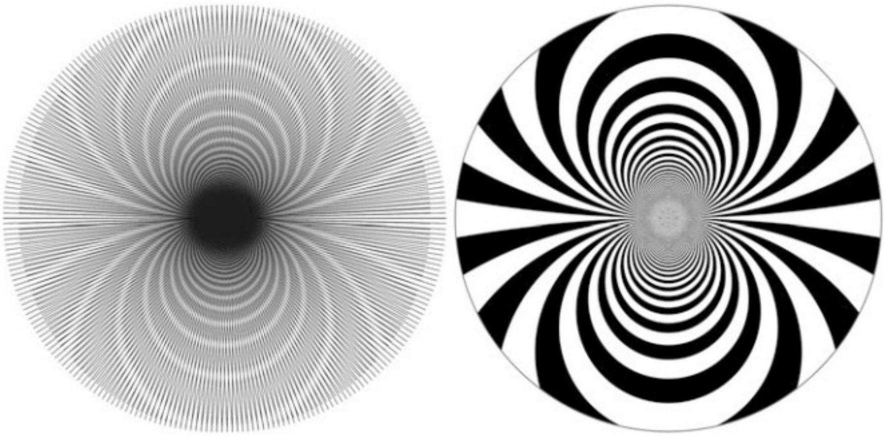

Fig. 36 Carsten Nicolai's stellar grating MI-RG-090 and a plot of $f(z) = \exp(1/z)$

inspiration. Beauty and symmetry are often guiding principles in these explorations. While beauty remains a matter of individual taste, experience, and convention, the concept of symmetry has a precise mathematical meaning, formally expressed in the language of group theory and reaching far beyond the common interpretation of this notion.

For centuries in the visual arts, beauty was a dominating criterion for the reception of artwork. In particular, during the Renaissance, strong composition rules were often based on symmetry and mathematical principles, and ingenious artists like Leonardo da Vinci or Albrecht Dürer made important contributions to mathematics.

During the twentieth century, the fundamental paradigm that art must be beautiful was abandoned, and art was assigned a new role as part of the human project of exploring nature, reality, and society. In the preface to the catalogue of Carsten Nicolai's exhibition "unidisplay," Eva Huttenlauch writes: "Art has become a category of thinking rather than seeing … ", Huttenlauch (2013, p. 13). As the aims of artists expanded, the border between mathematics and arts began to blur.

Artists like Victor Vasareli, Sol LeWitt, and Carsten Nicolai and scientists like Heinrich Heesch, Anatoly Fomenko, and Isaac Amidror are building bridges between arts and mathematics. Approaching a topic from opposite sides, emphasizing different aspects, and using tools specific to each discipline, mathematicians and artists sometimes seem to explore one and the same general phenomenon, indicating the presence of universal principles. For example, it is not an accident that the "stellar grating MI-RG-090" from the work "moiré index" (Nicolai, 2010b) resembles a black-and-white phase plot of the function $f(z) = e^{1/z}$ (Fig. 36). In the same vein, it would be interesting to do a comparative study of Niclolai's systematic listing of grids and tilings, Nicolai (2010a), (especially the semi-regular and irregular ones), with Frank Farris' wallpaper patterns, Farris (2015), and Isaac Amidror's theoretical foundation of moiré patterns, Amidror (2007, 2009).

Visualization is important for both the arts and mathematics. For the former, this has a long tradition. For the latter, tradition has focused on the subfield of geometry. However, advances of computer graphics opened up new possibilities for other branches of mathematics as well. In this chapter, it is the field of complex analysis that benefits from visual representations of its objects. In many mathematical subjects, the visual tradition has only just begun.

Acknowledgments Since August 2018, Pamela Gorkin has been serving as a Program Director in the Division of Mathematical Sciences at the National Science Foundation (NSF), USA, and as a component of this position, she received support from NSF for research, which included work on this paper. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Amidror I (2007) The theory of the Moiré Phenomenon: Volume II: Aperiodic Layers. Springer, Dordrecht

Amidror I (2009) The theory of the Moiré Phenomenon: Volume I: Periodic Layers. Springer, London

Anderson JW (2005) Hyperbolic geometry. Springer undergraduate mathematics series, 2nd edn. Springer, London

Arnold DN, Rogness J (2008) Möbius transformations revealed. Notices Am Math Soc 55(10):1226–1231

Baldus R (1944) Nichteuklidische Geometrie: hyperbolische Geometrie der Ebene, 2nd edn. Sammlung Göschen, De Gruyter

Chalendar I, Gorkin P, Partington JR, Ross WT (2018) Clark measures and a theorem of Ritt. Math Scand 122(2):277–298. https://doi.org/10.7146/math.scand.a-104444

Cowen CC (2012) Finite Blaschke products as compositions of other finite Blaschke products. arXiv:1207.4010

Daepp U, Gorkin P, Mortini R (2002) Ellipses and finite Blaschke products. Am Math Mon 109(9):785–795. https://doi.org/10.2307/3072367

Daepp U, Gorkin P, Shaffer A, Sokolowsky B, Voss K (2015) Decomposing finite Blaschke products. J Math Anal Appl 426(2):1201–1216. https://doi.org/10.1016/j.jmaa.2015.01.039

Daepp U, Gorkin P, Shaffer A, Voss K (2018) Finding Ellipses: What Blaschke Products, Poncelet's Theorem, and the Numerical Range Know about Each Other. The Carus mathematical monographs, No. 34. The Mathematical Association of America/American Mathematical Association, Providence

Euclid (1956) The thirteen books of Euclid's Elements translated from the text of Heiberg. Vol. I: Introduction and Books I, II. Vol. II: Books III–IX. Vol. III: Books X–XIII and Appendix. Dover Publications, Inc., New York, translated with introduction and commentary by Thomas L. Heath, 2nd edn

Farris FA (1998) Reviews: Visual Complex Analysis. Am Math Mont 105(6):570–576. https://doi.org/10.2307/2589427

Farris FA (2015) Creating symmetry: The Artful Mathematics of Wallpaper Patterns. Princeton University Press, Princeton. https://doi.org/10.1515/9781400865673

Fujimura M (2013) Inscribed ellipses and Blaschke products. Comput Methods Funct Theory 13(4):557–573. https://doi.org/10.1007/s40315-013-0037-8

Garcia SR, Mashreghi J, Ross WT (2018) Finite Blaschke products and their connections. Springer, Cham. https://doi.org/10.1007/978-3-319-78247-8

Garnett JB (1981) Bounded analytic functions, Pure and Applied Mathematics, vol 96. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York/London

Gorkin P, Wagner N (2017) Ellipses and compositions of finite Blaschke products. J Math Anal Appl 445(2):1354–1366. https://doi.org/10.1016/j.jmaa.2016.01.067

Gorkin P, Laroco L, Mortini R, Rupp R (1994) Composition of inner functions. Results Math 25(3-4):252–269. https://doi.org/10.1007/BF03323410

Greenberg MJ (2007) Euclidean and non-euclidean geometries: development and history, 4th edn. W. H. Freeman, New York

Heins M (1962) On a class of conformal metrics. Nagoya Math J 21:1–60

Heins M (1986) Some characterizations of finite Blaschke products of positive degree. J Analyse Math 46:162–166. https://doi.org/10.1007/BF02796581

Huttenlauch E (2013) Art as an explanatory model of universal complexity: on Carsten Nicolai's unidisplay, in: Carsten Nicolai, unidisplay, pp 9–13 (80), Gestalten, Berlin

Jahnke E, Emde F (1909) Funktionentafeln mit Formeln und Kurven. Mathematisch-physikalische Schriften für Ingenieure und Studierende, B.G. Teubner, Leipzig und Berlin

Kraus D, Roth O (2008) Critical points of inner functions, nonlinear partial differential equations, and an extension of Liouville's theorem. J Lond Math Soc (2) 77(1):183–202. https://doi.org/10.1112/jlms/jdm095

Kraus D, Roth O (2013) Critical points, the Gauss curvature equation and Blaschke products. In: Blaschke products and their applications. Fields institute communications, vol 65, Springer, New York, pp 133–157. https://doi.org/10.1007/978-1-4614-5341-3_7

Milnor J (1982) Hyperbolic geometry: the first 150 years. Bull Am Math Soc 6(1):9–24

Ng TW, Tsang CY (2013) Polynomials versus finite Blaschke products. In: Blaschke products and their applications. Fields institute communications, vol 65. Springer, New York, pp 249–273. https://doi.org/10.1007/978-1-4614-5341-3_14

Nicolai C (2010a) Grid Index. Die Gestalten Verlag, Berlin

Nicolai C (2010b) Moiré Index. Die Gestalten Verlag, Berlin

Ritt JF (1922) Prime and composite polynomials. Trans Am Math Soc 23(1):51–66. https://doi.org/10.2307/1988911

Semmler G, Wegert E (2019) Finite Blaschke products with prescribed critical points, Stieltjes polynomials, and moment problems. Anal Math Phys 9(1):221–249. https://doi.org/10.1007/s13324-017-0193-5

Singer DA (2006) The location of critical points of finite Blaschke products. Conform Geom Dyn 10:117–124. https://doi.org/10.1090/S1088-4173-06-00145-7

Stäckel P (1933) Gauß als Geometer. In: Carl Friedrich Gauß Werke Band X.2 Abhandlungen über Gauss' wissenschaftliche Tätigkeit auf den Gebieten der reinen Mathematik und Mechanik. Springer Berlin

Walsh JL (1950) The location of critical points of analytic and harmonic functions. Colloquium Publications, vol 34. American Mathematical Society

Walsh JL (1952) Note on the location of zeros of extremal polynomials in the non-euclidean plane. Acad Serbe Sci, Publ Inst Math 4:157–160

Wegert E (2011) Phase diagrams of meromorphic functions. Comput Methods Funct Theory 10(2):639–661

Wegert E (2012) Visual complex functions: An Introduction with Phase Portraits. Birkhäuser/Springer Basel AG, Basel. https://doi.org/10.1007/978-3-0348-0180-5

Wegert E, Semmler G (2011) Phase plots of complex functions: a journey in illustration. Notices Am Math Soc 58(6):768–780

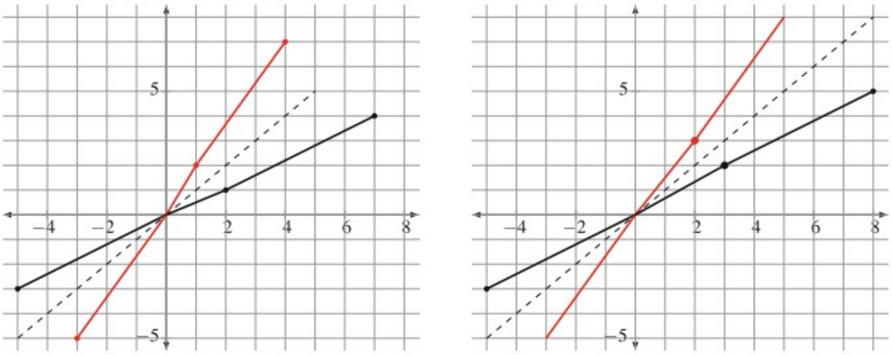

Black line is type A frequency $3/5$.
Red line is Type B frequency $5/3$

Black line is type B frequency $3/5$.
Red line is Type A frequency $5/3$

Fig. 8 Pairs of conjugate skew Sturmian sequences with different types and reciprocal frequencies

introduce an equivalence relation defined by $A_{\frac{q}{p}} \sim B_{\frac{p}{q}}$ and $B_{\frac{q}{p}} \sim A_{\frac{p}{q}}$. Given the geometric nature of prime numbers, it is natural to wonder about the geometrical behavior of these equivalence classes of a given skew Sturmian sequence. To this purpose, introduce the subset

$$Z_{\frac{q}{p}} = A_{\frac{q}{p}} \cup B_{\frac{p}{q}} \cup B_{\frac{q}{p}} \cup A_{\frac{p}{q}}$$

containing the (skew Sturmian) dynamical information associated to (p, q).

Since $Z_{\frac{q}{p}}$ has cardinality eight, it seemed more interesting to plot the points of $Z_{\frac{q}{p}}$ when the points (p, q) belong to a certain line.

Ford Circles and Farey Sequences

Ford circles (Ford, 1938) are "geometric picturizations" of fractions. For a given rational number $\frac{q}{p}$ with p and q coprime numbers, the Ford circle associated to $\frac{q}{p}$ is given by the equation

$$\left(x - \frac{q}{p}\right)^2 + \left(y - \frac{1}{2p^2}\right)^2 = \left(\frac{1}{2p^2}\right)^2.$$

A remarkable property of Ford circles is the following: If $\frac{Q}{P}$ is an adjacent fraction to $\frac{q}{p}$, that is to say, if $|qP - pQ| = 1$, then the Ford circles associated to $\frac{q}{p}$ and $\frac{Q}{P}$ are tangent. For example, the fractions $\frac{1}{2}, \frac{2}{3}$, and $\frac{4}{7}$ are adjacent to $\frac{3}{5}$. The Ford circles corresponding to these fractions are illustrated in Fig. 9.

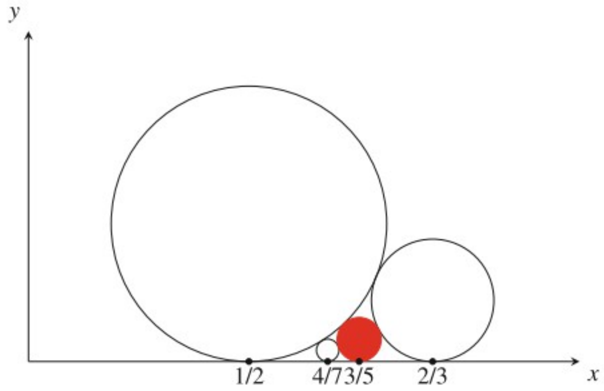

Fig. 9 Ford circles for $\frac{3}{5}$, in red, and the Ford circles of its adjacent fractions $\frac{1}{2}$, $\frac{4}{7}$, and $\frac{2}{3}$. Notice that the Ford circles are an infinite family of circles tangent to each other

Ford proved (Ford, 1938, Theorem 4) that any such irreducible fraction $\frac{q}{p}$ with $|p| > 1$ has exactly two adjacent fractions with denominators smaller than p. It follows that $A(p, q) = (a, b)$ and $B(p, q) = (c, d)$ are the only other two pairs with $0 < a, c < p$ such that both $\frac{b}{a}$ and $\frac{d}{c}$ are the *closest adjacent fractions* to $\frac{q}{p}$. Recall that a pth-Farey sequence consists of all proper reduced fractions with denominator up to p. Hence, $A(p, q)$ and $B(p, q)$ can be seen as the left and right neighbors, respectively, of $\frac{q}{p}$ in the p-th Farey sequence (Conway and Guy, 1996, Page 152). Recall that the mediant fraction of $\frac{b}{a}$ and $\frac{d}{c}$ is defined as $\frac{b+d}{a+c}$. Suppose that $A(p, q) = (a, b)$ and $B(p, q) = (c, d)$. Since $\frac{q}{p}$ is the mediant fraction of its neighbors, it follows that $\frac{q}{p} = \frac{b+d}{a+c}$ which implies $1 = aq - bp = a(b + d) - b(a + c) = ad - bc$. Therefore the pairs (p, q), $A(p, q)$, and $B(p, q)$ have mutually adjacent associated fractions. Thus, the Ford circles for (p, q), $A(p, q)$, and $B(p, q)$ are mutually tangent, as shown in Fig. 9.

Bézout Graphs

Let p be a positive integer. In (Itzá-Ortiz et al, 2020), the Bézout coefficients of the coprime pair (p, q) as $B_1(p, q) = A(p, q)$ and $B_{-1}(p, q) = B(p, q)$ were defined. Then the Bézout set $\mathscr{B}_p$ for p is

$$\mathscr{B}_p = \bigcup_{\substack{0<q<p \\ \gcd(p,q)=1}} Z_{\frac{q}{p}}.$$

In other words, the Bézout set $\mathscr{B}_p$ of p is the union of the sets $Z_{\frac{q}{p}}$ as the pairs (p, q) of coprime numbers move along the segment from $(p, 0)$ to (p, p). Notice

Pág. 136(157)

Fig. 13 A conformal map between the disc and square in the complex plane.

Pág. 140(161)

Fig. 17 Playfair's axiom on the Euclidean plane (left) and the hyperbolic plane (right)

Pág. 158(179)

Ford Circles and Farey Sequences

→ "geometric picturizations" of fractions.

Pág. 159(180) -

Fig. 9 Ford circles for $\frac{3}{5}$, in red, and the Ford circles of its adjacents fractions

$\frac{1}{2}$, $\frac{4}{7}$, and $\frac{2}{3}$. Notice that the Ford circles are an infinite family of circles tangent to each other.

Ford proved (1938, theorem 4) that any such (qualquer tal) irreducible

irreducible fraction q with
$|p| > 1$ has exactly two adjacent
fractions with denominator
smaller than p. ...

Important:

Pág. 173 (193)

Each Surface admits more
than one geometric shape

Pág. 174 (194)

Fig. 2 Moduli space for the
surface of genus 2

Fig. 3 From a regular hexagon
on the plane to the
torus.

Pág. 183 (203) ~~[illegible]~~

Anamorphosis Forward Fast

Fig. 3 Two very different 3D objects with the
same visual core.

Fig. 4 Two very different 3D objects with the same
visual core one surrounding the other.
Anamorphosis is radial and isotropic, with no

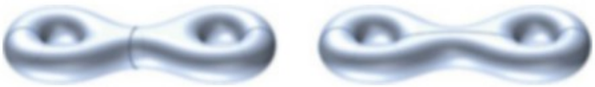

Fig. 1 Two different pants decompositions of a surface of genus 2

glue, you end up with the closed surface of genus 2. Figure 1 is a picture of two kinds of pants decompositions of the surface of genus 2.

If you use four pairs of pants, then you get the closed surface of genus 3. In fact, the surface of genus g can be obtained from 2g-2 pairs of pants by gluing 6g-6 boundary components of those pants pairwise. It does not matter how you glue them – you always get the same surface up to topological equivalence.

As an exercise, I invite the readers to try to decompose the closed surface of genus g into 2g-2 pairs of pants by drawing 3g-3 loops appropriately. Try with g = 2, 3, 4, … and see if you can generalize your idea for arbitrary genus. This is a fun mental exercise, and it is the base of the idea of building closed surfaces out of hexagons.

Hyperbolic Plane and Negative Curvature

In this section, we will briefly talk about the hyperbolic plane $\mathbb{H}^2$. This is like the usual plane in some way but very different in other ways, since it is not flat but curved weirdly. At every point, a small neighborhood looks like the saddle. It is curved in the way that in some direction, it looks like convex, and in other direction, it looks like concave.

In the world of hyperbolic plane, there are more "rooms" out there. For a straight line L and a point p which is not on the line L, there are infinitely many straight line passing through p but never meet the line L we started with! If you recall what you learned about the geometry of the usual plane, there should be a unique such line, but in the hyperbolic plane, there are infinitely many.

To grasp the idea how the hyperbolic plane actually looks like, I encourage the readers to check out the famous paintings of the tessellation of the hyperbolic plane by Escher (1959).

The reason why I mention the hyperbolic plane in this section is that the "natural geometric shape" of an arbitrary closed surface of genus at least two is like the hyperbolic plane. So far, we have always considered surfaces up to topological equivalence which ignore the precise shape such as the distance between two points and the angle between two lines. But each surface can be made into a geometric object by specifying the shape, and it is a remarkable fact that any closed surface of genus at least two has a geometric shape such that every point has a small neighborhood where the world looks exactly like the hyperbolic plane. When the surface is in such a shape, we call it a hyperbolic surface.

Once you have a decomposition of a hyperbolic surface of genus g into $2g - 2$ pairs of pants via $3g-3$ loops on the surface, we can straighten up the loops so that at every point on the loop, it locally looks like a piece of a straight line in the hyperbolic plane. On each pair of pants, we have three straight line boundary components. For each pair of boundary components, one can connect them via a straight line segment which meets the boundary components at the right angle. After connecting all three pairs, now one sees that the pair of pants is decomposed into two hexagons. Each of this hexagon geometrically lives in the hyperbolic plane, in the sense the it is exactly the same hexagon with some hexagon in the hyperbolic plane where each side is a straight line segment.

This is already an interesting fact, because now we have a pair of pants made out of two hexagons where each side is a straight line segment and each internal angle is the right angle. We call such a hexagon a right-angled hexagon. A right-angled hexagon cannot live in our usual world, since this would mean the total sum of the internal angles is 3π, while the total sum of the internal angles of a hexagon in the Euclidean plane is 4π. This is another feature of negative curvature. In fact, in the hyperbolic plane, one can have a hexagon whose internal angles sum up to any number between 0 and 4π.

Above we expressed a hyperbolic surface of genus g as a collection of $4g - 4$ hexagons in the hyperbolic plane glued along their boundary. An elementary hyperbolic geometry can show that the geometric shape of a right-angled hexagon in the hyperbolic plane is completely determined by the length of the three alternating sides. Hence, the two hexagons forming one pair of pants are in the exactly same shape. In particular, this means, up to geometric equivalence, we only need $2g - 2$ hexagons to build a closed surface of genus g.

Each Surface Admits More Than One Geometric Shape

We have so far shown that all but two closed surfaces are built out of hexagons – more precisely, any closed surface of genus at least 2 can be obtained from gluing hyperbolic hexagons along their boundary. Since the surface is built upon pieces which live in the hyperbolic plane, it has the natural geometric shape as a hyperbolic surface. But the geometric shape of the surface is not unique, and by varying the shapes of hexagons, we can build surface which are topologically the same but geometrically different.

Perhaps this can be best explained in the pants decomposition point of view. For each loop used in the pants decomposition of a surface, one can change the length of the loop (but not all loops can be really short or really long due to our geometric constraints), and also when the lengths of the loops are fixed, there is still a room for further deformation of the shape. In fact, at each loop, one can think two pairs of pants are glued along the loop. But this gluing can be also altered by cutting it along the loop again and twist by an arbitrary angle and glue back. Mathematicians showed that each twisting will yield a surface of different geometric shape (although this is totally nontrivial). We have $3g - 3$ curves for a pants decomposition on the surface of genus g, and each curve has two degrees of freedom for the deformation – the length and the twisting. This means we have $6g - 6$ parameters to determine the geometric shape of the surface. In fact, if we collect all possible (hyperbolic) geometric shapes that a closed surface of genus g can admit, we get a huge $6g - 6$-dimensional space called the moduli space (Fig. 2).

Fig. 2 Moduli space for the surface of genus 2

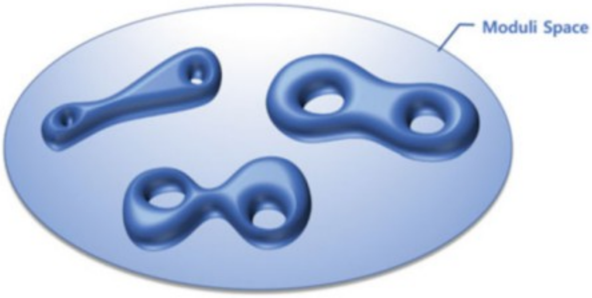

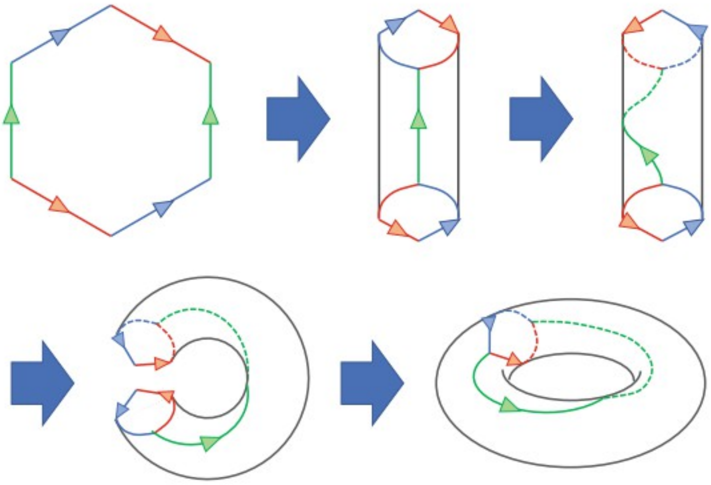

Fig. 3 From a regular hexagon on the plane to the torus

To conclude our journey, we go back to the case of the torus and see that the torus can be also obtained from a hexagon. In this case, it is obtained from a flat hexagon (i.e., a hexagon in the usual plane) by gluing sides pairwise.

Consider a regular hexagon in the plane, and glue the opposite sides pairwise. Then we get a torus – like a magic! Seeing is believing, so I describe the procedure of this side-gluing with a picture. See Fig. 3. To learn more about the torus and its symmetry, the readers are encouraged to see also the first chapter of Thurston (1997).

References

Bredon G (2013) Topology and geometry. Springer, Berlin

Escher MC, (1959) Circle Limit III, website (https://mcescher.com/gallery/mathematical/)

Farb B, Margalit D (2011) A primer on mapping class groups. Princeton University Press, Princeton/Oxford

Thurston W (1997) Three-dimensional geometry and topology. Princeton University Press, Princeton

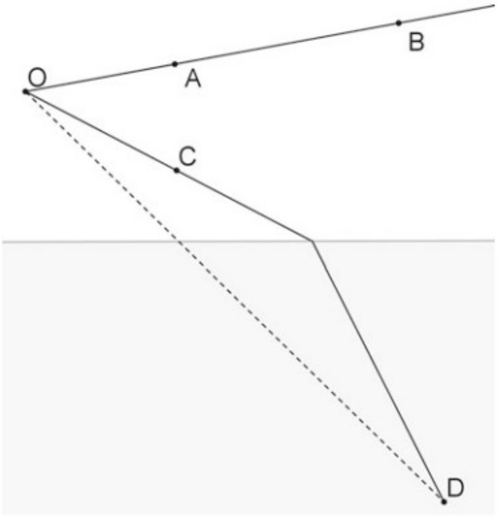

Fig. 2 Points A and B are both geometrically and visually aligned relative to the observer at O. Points C and D are visually but not geometrically aligned relative to O due to refraction on the transition between optical media (e.g., water and air)

it will extend is an open question. So, more than explicit statements of absolute truth, these principles are implicit definitions of the gamut of experimental conditions under which they are verified. And they are interesting concepts if that gamut is interesting. That is the case with radial occlusion. Although far from universally valid, it is the default supposition of our interaction with light and the world. We will no longer be concerned in this section on the intricacies of when the principle works, but will reason on its consequences. All that lies ahead follows geometrically from this single assumption of radial occlusion that we will later abstract into our definition of anamorphosis (Definition 4).

Anamorphosis Formed Fast

We make a quick and informal summary here, before we proceed to the formal definitions and results.

If we accept the principle of radial occlusion, then it follows that many different $3D$ objects will look the same from O. The points of an object define a cone of rays from the observer O, the visual cone. If two objects define the same cone of rays, they should look the same. Such objects we will call *anamorphs* of each other relative to O (in fact we will require a technical condition for this *anamorphic equivalence* or *anamorphosis* that is slightly looser than equality of visual cones, but that is the subject of the next section; here we speak loosely).

The important things to note is that, unlike in perspective, the equivalence here is between 3D objects, and not just between objects and drawings. There is no need here to mention projection surfaces.

For instance, take the wireframe cube of Fig. 3 and an observer at O. If we slide each vertex along its ray from O, we get the object to the right of the cube. These

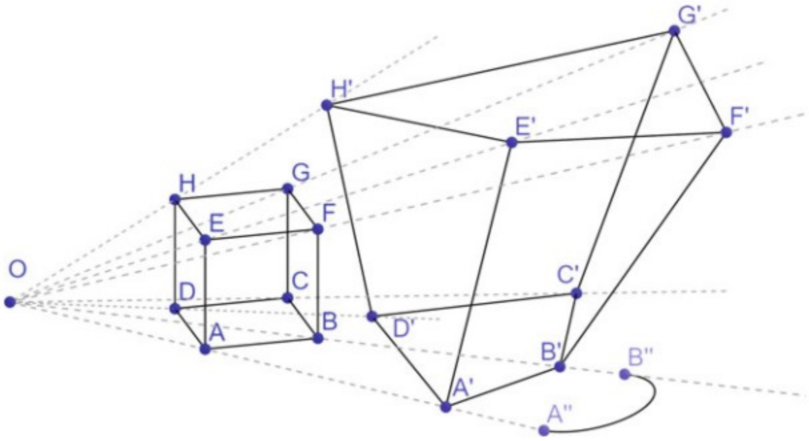

Fig. 3 Two very different 3D objects with the same visual cone. They will look the same when seen from O if the principle of radial occlusion is valid

Fig. 4 Two very different 3D objects with the same visual cone, one surrounding the other. Anamorphosis is radial and isotropic, with no preferential axis

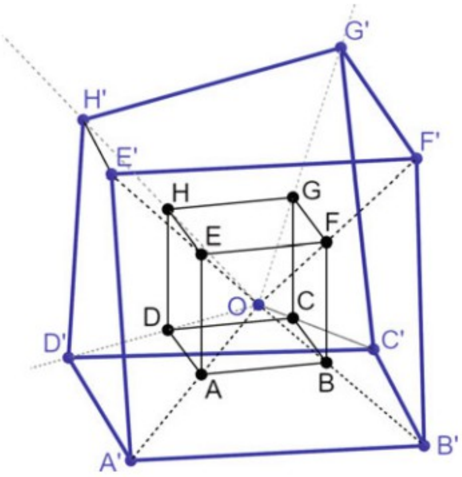

two shapes are very different, but they define exactly the same cone of rays from O, so, if we accept the principle of radial occlusion, we must accept that the two objects will look the same from O. Of course, the segments joining the vertices don't have to be straight. You could bend each segment freely within the cone of rays it generates, and it would preserve the equivalence. For instance, arc $A''B''$ in Fig. 3 is equivalent to rays AB and $A'B'$ as it lies within the cone of AB. This cone is just the angle $\angle AOB$.

This equivalence is radial and isotropic in nature. Unlike classical perspective, where there is a preferential axis defined by the perpendicular to the plane of the picture, here there is no such axis, plane, or even picture, hence no limitation of field of view: this is just an equivalence between objects. So the example of Fig. 4 works just as well as the previous one. This allows to give equal footing to immersive anamorphoses, that is, anamorphs that completely surround the observer (in turn, this proves foundational for spherical perspectives (Araújo 2018c)).

Remark 1. The limitation of a preferred axis is often kept in the literature from attachment either to the habits of perspective or to the convenient availability of the mathematical framework of projective geometry (see, for instance, (Sánchez-Reyes and Chacón 2016)). But in projective geometry all points on a line are equivalent, and this is a limited framework for anamorphosis. In anamorphosis the classes of equivalence are rays, not lines, so the sphere and not the projective plane will represent the visual data. The projective plane ignores exactly half of this visual data.

The principle of radial occlusion implies the existence of 2D anamorphs to any 3D object. Suppose S is a surface, not containing O. Then the intersection of the visual cone of a 3D object X with S is a 2D object that subtends the same cone as X itself. Hence it is a "drawing" that looks exactly like S when seen from O.

You could loosely say that the principle of radial occlusion is what makes *drawing* possible.

The most obvious case is the intersection of an object's cone with a plane. We see two examples in Fig. 5, the same object being projected onto a "vertical" and a "horizontal" plane.

There is no fundamental difference between horizontal and vertical, of course, since these are just named relative to the gravity field, which is quite irrelevant for optics. And yet the former anamorph is usually called a classical perspective, while the latter is called an "oblique anamorphosis." What is the difference? Any projection plane defines a preferred axis, the perpendicular to the plane through O. The difference is the angular distance of the object's cone with regard to that axis. When the object's cone is closely aligned with the main axis, the drawing is called a perspective; when the angular separation is "large," it is called an oblique anamorphosis. We see that the difference is one of degree, and both are

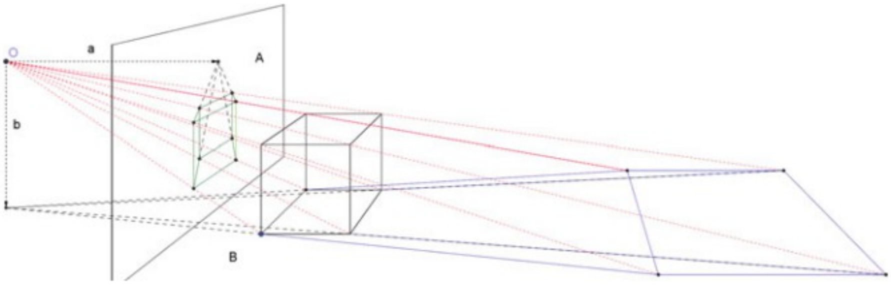

Fig. 5 Projection onto a plane. There is no absolute difference between the linear perspective and the oblique anamorphosis. Both are anamorphosis onto a plane; the latter merely projects the object to a region further away from the foot of the perpendicular

special cases of anamorphosis, the projection being just another anamorph that happens to be contained in a plane, whatever the plane's orientation may be. The projection is operationally the same in both cases; what makes it a "perspective" or "oblique anamorphosis" is the relative position of the body being projected, not the functional form of the projection itself. It is also quite indefensible to say, as is usually naively put, that the anamorphosis is "deformed," while the perspective isn't. What deformation is meant? By the principle of radial occlusion, none of them is *optically* deformed, being indistinguishable from O, while metrically both of them are deformed; the elongation of the edges in the oblique anamorphosis has a counterpart in their compression in the perspective. Of course some type of metrical deformation is inevitable if you lose one dimension.

Let us move to more complex setups. Nothing stops us from using several planes at once, as in Fig. 6. The cone projects onto this union of planes as several images that should all join seamlessly when viewed from O. We will see that this changes the global properties (such as the number of vanishing points) in a fundamental manner while preserving the local properties of classical perspective. This multi-plane projection is historically used in the so-called perspective boxes (Čučaković and Paunović 2016; Gay and Cazzaro 2019; Spencer 2018; Verweij 2010).

But of course we are not limited to planes. If we intersect the visual cone of an object with a curved surface, we still obtain a 2D anamorph. Now the line projections will themselves be curved in space, yet they will seem straight when seen from O. Ahead we will see how to investigate the shapes of these lines analytically and geometrically and how to draw them in practice. In Fig. 7 we see an example of

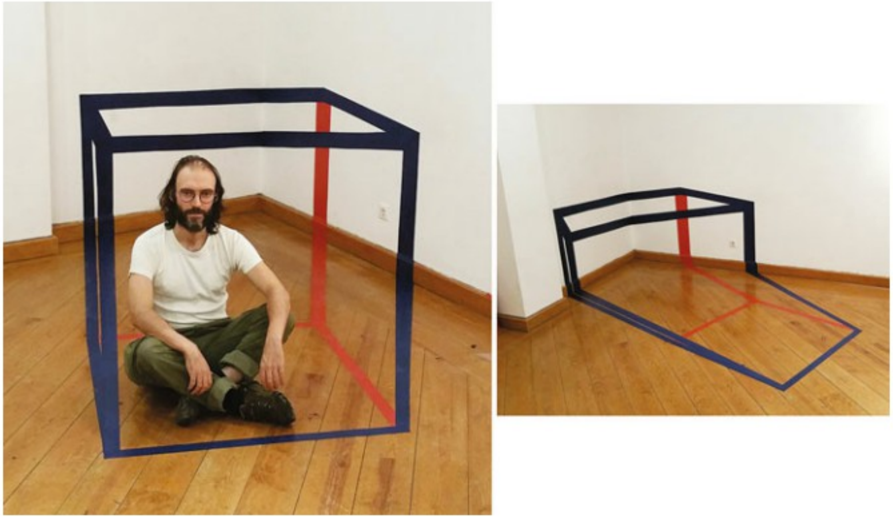

Fig. 6 Anamorphosis of a cube onto three planes. Locally these are just plane perspectives but globally, some lines in such multi-plane configurations may present up to two vanishing points. Temporary installation in Óbidos, Portugal, joint work by the author and Maria Bianchi Aguiar

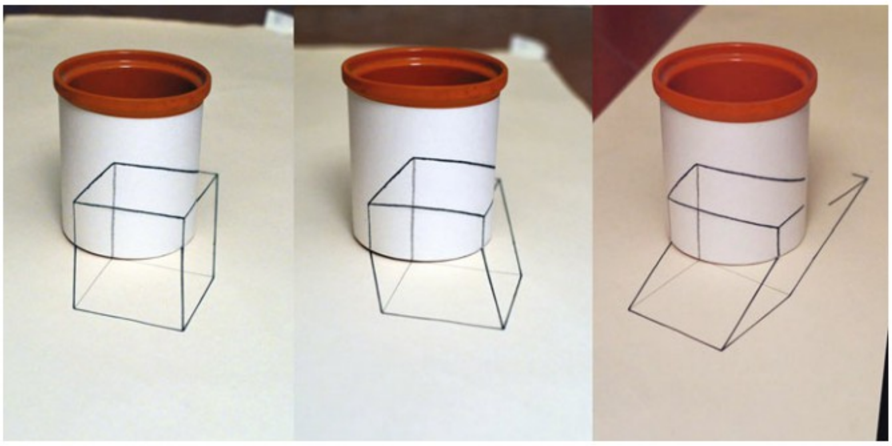

Fig. 7 Anamorphosis of a cube onto a union of a plane and a cylinder. (Drawing by the author)

an anamorph onto a composite surface made up of a cylinder and a plane. We notice that the lines of the anamorph don't even have to be connected.

Further, if we wish to obtain an immersive anamorphosis, that is, an anamorphosis that surrounds the viewer, we may place *O* inside a surface, for example, in the axis of a cylinder. Then the observer may look in all directions around the vertical axis and have the illusion of being surrounded by an immersive landscape. This is the principle behind the famous panoramas of the nineteenth century, illusions so striking at the time that the word em *panorama*, originally created to signify these illusory pictures, became in time associated to the physical landscape itself. In fact the word seems to appear in print for the first time in 1791 in an advertisement promoting Robert Barker's large displays of immersive, cylindrical anamorphoses at his "panorama building," (Fig. 8) erected on Leicester square, in London (Huhtamo 2013). From the platform at the center of this building, large paintings were displayed, to immerse the spectators in foreign landscapes or representations of famous battles, such as that of Waterloo, that was pictured in the Leicester Square panorama in 1816, having been fought less than a year before. Such panoramas became widespread worldwide in the nineteenth century, and even today a panorama of the battle of Waterloo is on display at the museum erected on the field of battle, and a few depictions of civil war battles survive in the United States. It is important to note here that Barker's panoramas, striking as they were, were not strict cylindrical anamorphoses. According to Kemp (1990, p. 214) they were made by joining a series of adjacent linear drawings, with some care to soften the transitions. These are perfectly adequate as long as there are no exceedingly long architectural lines involved, which is not a problem for wide vistas where each building will occupy a small angle of view and the general landscape is one of natural forms that the eye cannot judge for exactness.

Of course, the most isotropic anamorphic view would be a spherical one, and this too has been exploited in various ways. The visual illusions of the planetarium

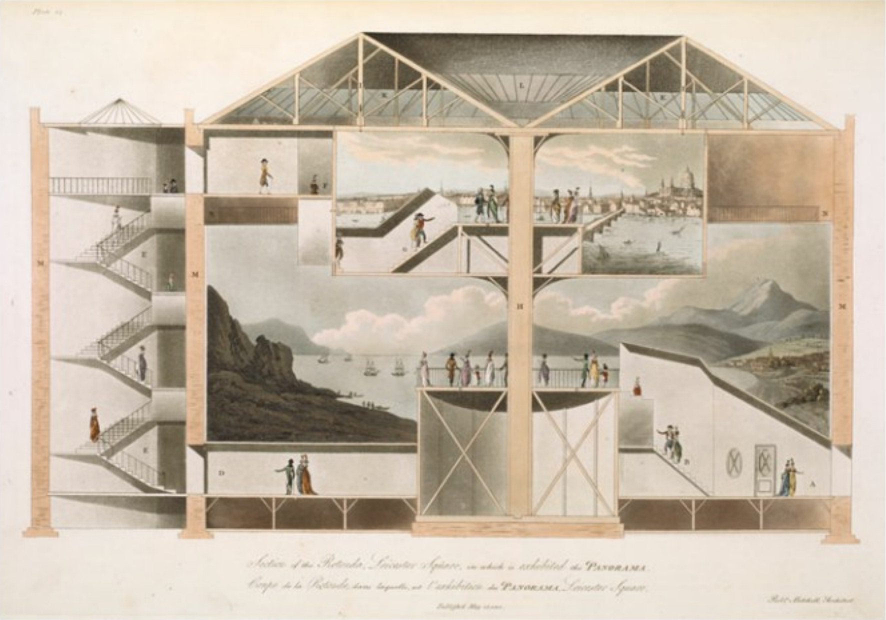
Section of the Rotunda, Leicester Square, in which is exhibited the PANORAMA
Coupe de la Rotunde, dans laquelle, est l'exhibition du PANORAMA Leicester Square.

Fig. 8 Cross section of the Rotunda in Leicester Square, showing the viewing platforms for the main and secondary panoramas. A dark passage between the two cleared the viewing palate. The building was designed by Robert Mitchell to exhibit the panoramas of Robert Barker. Etching/Colored Aquatint by Robert Mitchell, 1801 (Mitchell 1801a, b)

or of a modern, digital full dome allow for hemispherical anamorphoses, and no more than architectural convenience and economy stops us from applying the same principle to build a full sphere, to provide a truly immersive anamorphosis, completely surrounding the observer. In fact, fully spherical immersive illusions were tried architecturally. Charles Delangard proposed the concept of the *Georama* to the Geographical Society of Paris in 1822, and the first exemplar was erected in that city in 1826 (Figs. 9) – a sphere of approximately 12 meters in diameter, whose interior was painted with a representation of the Earths geography, seen from within (Belisle 2015; Oettermann and Schneider 1997). Delangard's Georama, like the panorama before it, became widespread and was implemented elsewhere , the most famous example being James Wyld's Great Globe in Leicester Square. The concept was also used to represent the celestial globe. There was one such globe in the Paris exhibition of 1900 that visitors could enter to see the firmament rotate around their station point. Some Georamas paired these geographic and astronomical views, as in the case of Wyld's Great Globe, whose interior displayed the Earth's surface while having the celestial firmament painted on the exterior dome. This was to be expected, since the map making tradition has a long-standing pairing between the charts of the stars and those of the Earth (see, for instance, the ▸ Chap. 19, "Spherical Perspective" in the present volume). Again, like the

Fig. 9 Illustration of Delangard's Georama

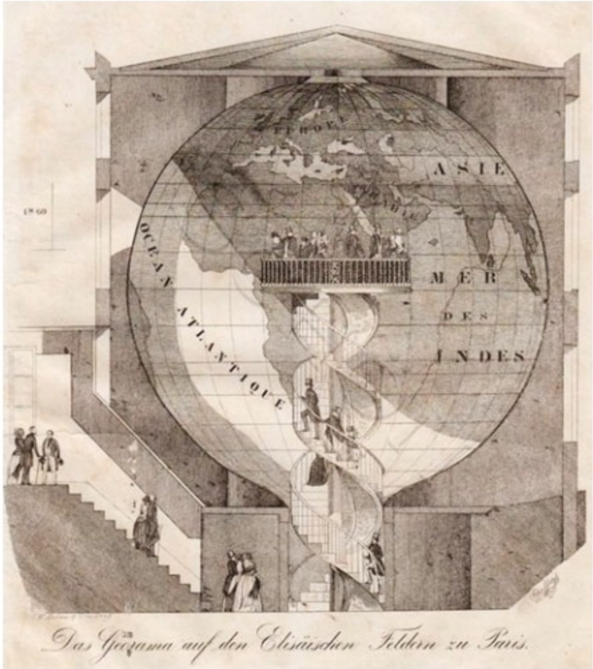
EUROPE
ASIE
OCEAN ATLANTIQUE
MER DES INDES
Das Georama auf den Elisäischen Feldern zu Paris.

panoramas before them, the Georamas were not strict anamorphoses, having often deformations to facilitate particular readings, or putting the viewers in observational platforms not necessarily located at the center of the sphere (Belisle 2015). Still, they demonstrate the architectural possibilities of a completely immersive anamorphosis. Of course in the present day, a type of immersive anamorphoses is daily achievable through digital means. Fist-person games or simulations work by creating a flat anamorphosis dynamically onto the screens of the 3D glasses, centered on a moving observational point O, the screen serving as a moving window into the virtual world that will be in anamorphosis if the distance to the screen is adequate. Of course, since the distances are small, the principle of radial occlusion will not really be verified, so the illusion is not really immersive. Better results are obtained with the help of 3D helmets, where lenses create a sense of size and distance and illusion pairs are created for each eye at a moving pair of observational points O_L and O_R. This is but another step in the evolution of moving panoramas.

Some Considerations on Anamorphosis

The Point of Observation

It is important to stress that anamorphic equivalence is always related to an observer O. It makes no sense to ask if something is an anamorphosis without context. Everything is an anamorphosis of some other things relative to some point. It only makes sense to ask if object X is an anamorphosis of object Y with relation to O. This is what is implied when in the usual dictionary definition of anamorphosis, we say that the object "looks right" from a certain point. But this vagueness causes all sorts of misconceptions. The issue is not that it looks "correct" (meaning recognizable as "something") but that it looks like something that has be prescribed in advance. The question is how much can be prescribed or, to put it more rigorously, how much does a prescribed anamorph determine the set of all possible anamorphs of it with relation to given viewpoints.

Multiple Points of Observation

We can prescribe the different appearances for the same object from two different points within some rather forgiving bounds. When one prescribes that X be anamorphic to Y_1 from O_1, this determines a cone of rays from O_1 to X, but does not determine X itself. Each point of X on that cone may still be freely moved along its ray. That degree of freedom allows one to specify another object (under constraints) Y_2 such that X will be anamorphic to Y_2 from another point O_2.

As beautifully demonstrated by the works of Kokichi Sugihara, the degrees of freedom of the conic projection allow quite some leeway. Sugihara wonderfully explored such double anamorphic constructions with ambiguous 3D objects that from one point may look cylindrical and from another box-like or star-shaped (Sugihara 2015a, b, 2016), and not only the geometry of such ambiguous cylinders but even their topology may be changed, when, for instance, a pair of cylinders may appear to be intersecting or not according to the choice of observation point (Sugihara 2018).

"Impossible" Objects

It may seem paradoxical, but a real object may be anamorphic to an *impossible* one, that is, it may look like an object that makes sense locally, but not globally, like Penrose's triangle (Draper 1978; Penrose and Penrose 1958), M. C. Escher's Belvedere cube (Escher 1958, 1972), or Huffman's various specimens of wireframe constructions (Huffman 1968, 1971). This is achieved by visually mimicking incompatible self-connections and other impossible features through hidden discontinuities or deformations of a real object (see Sánchez-Reyes and Chacón (2020) for a modern software-based treatment of this problem). Of course the apparent impossibility comes not so much from what we see as from what we assume, that is, not from the geometric constraints determined by the appearance of the object but from the psychological assumptions of how our brain makes sense of the limited visual information supplied by the view from a single point (Kulpa 1983; Sugihara 1982, 2000). Human visual processing relies on shortcuts and rules-of-thumb meant to work on the generic rather than the general case. For a simple example of an apparently impossible object, note that the rightmost object of Fig. 3 might be classified as one of Huffman's impossible polyhedra, but only if we assume that the implied faces are planar. The reader will verify that there is in fact no polyhedron whose faces are delimited by the edges shown. However, if we triangulate across the vertices, as in Fig. 10, then get a perfectly consistent polyhedron, only the number

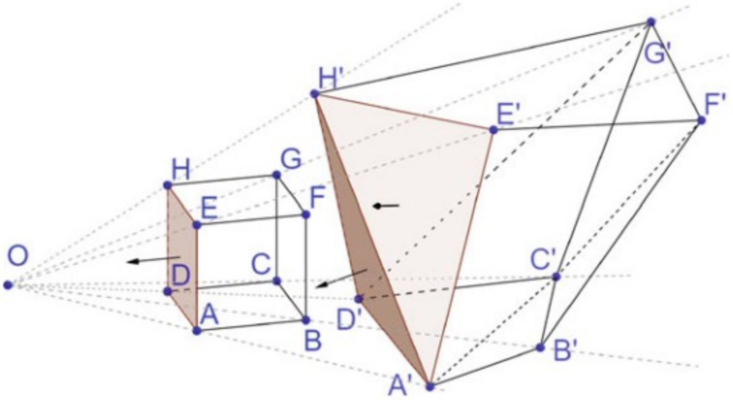

Fig. 10 The rightmost object of Fig. 3 can be made into a consistent polyhedron by a choice of triangulation. Corresponding triangulations would result in the anamorphically equivalent edges in the cube and the right side object, but the faces themselves would be shaded differently under a light source. In a Lambert model of illumination, the value (gray level) of the two faces $A'H'D'$ and $A'E'H'$ will vary with the inner product of the faces normals to the incidence direction of the light. This could be compensated by painting the faces with different shades of gray so as to make them seem uniform when seen from O, achieving "color anamorphosis"

of planar faces is greater than the six originally assumed. Due to the ambiguity of radial occlusion, it is no wonder that even the simplest of wireframe objects, like Necker's cube (Necker 1832), should make our brains oscillate between two readings undecisively; the wonder is that we may make any reading at all with confidence from among the infinite anamorphs that are mathematically consistent.

On Color

It is important to note that anamorphosis in our present sense refers only to mere equivalence of contours or outlines. This limited principle is necessary but not sufficient for proper anamorphoses in the usual sense, as most optical illusions going by that name – be it Pozzo's dome at St. Ignazio's or Holbein's painting of the ambassadors – require also mimesis of color to be effective. This aspect is usually unmentioned in the literature, as we assume that color will take care of itself once we use conical anamorphoses to define the contours of the objects we wish to represent in painting. This is true with careful diffuse lighting, but in a general setup color will certainly *not* take care of itself. Color anamorphoses (so to speak) are a whole new layer of complexity for mimesis that we will not treat it here with any detail, but would like to mention briefly. Color anamorphosis would require consideration of models of illumination, material properties, and, of course, consideration of color itself as a perceptual theory with its own principles. Color has its own mimetic principle, its own "anamorphosis" of sorts. The space of visual color stimuli may be identified with the set of all power distributions in the visible spectrum, which, as a vector space of functions, is of very high dimension (we could say infinite dimensional if we identify it with the set of all functions on the interval [0, 1], but there are restrictions so the space is probably better seen as very high – but finite – dimensional). This space is reduced by our perception to only a convex set in a space of three dimensions, identifiable (for instance) as something like value, saturation, and hue. This reduction in dimension is analogous to the way anamorphosis reduces the three dimensions of the real space to a two-dimensional perceptual space (though the reduction is much more drastic in color).

But all of this is beyond our scope. To consider the simplest of cases, take again the two anamorphic wireframe objects of Fig. 10. The situation is simple enough if we just consider the wireframe objects with ideal segments joining the points. If we consider a real object, then already we have some complications, as those lines will need some thickness to be visibly represented. We can naively conceive of them optically as painted matte black cylinders (with no specular reflections), thin, and featureless. If we now consider the polyhedron defined by a triangulation then both objects would just look like a black mass defined by their identical outlines. Let us now go a bit further and suppose the objects are made of triangles painted in some neutral gray. Suppose also they reflect light equally in every direction and with an intensity proportional to the inner product of the surface normal to the incident light ray direction. This is a simple Lambert model of illumination, adequate for rendering matte surfaces (Foley et al 1996).

Within such a model, the gray level of each face will be determined by its angle to the light source. Hence, mimicry of color will require us to paint each face in such a way as to compensate the difference in their normal vectors. For instance, in Fig. 10, face $ADHE$ of the cube is anamorphic to the union of the two triangular faces $A'D'H'$ and $A'H'E'$. Notice the difference in the normal vectors on the faces. If the faces were to be painted equally, and assuming light comes from front and high, the face that points downward would look darker than the cube's face, and the face that points upward would look brighter. So to compensate for this and achieve apparent equality of value ("color anamorphosis"), we would have to paint the upward facing triangle darker and the downward one lighter than the square face.

In Fig. 11 we see a simple hand-drawn example of color anamorphosis: a plane anamorph of a cube, side by side with a real cube of the same apparent dimensions. The physical cube is painted uniformly with white paint. The plane anamorph's normal vector coincides with the normal vector of the top face, so these corresponding faces can be painted with the same value. But the other "faces" of the plane anamorph had to be painted darker in order to match the appearance of the faces of the 3D cube that are turned at a larger angle from the incident light.

The rules of this color anamorphosis game depend on many factors and will vary with the type of light, its form, its location, the type of reflection we get from the materials, and so on. In most cases it will be impossible to match apparent color by simply changing the local color, as the ranges will be too large. Lambert reflection and a uniform light is just about the easiest situation for matching. Diffuse lighting helps, as cast shadows will create additional complications (or artistic opportunities).

If we were to attempt a formal treatment of color anamorphosis, we would start with something akin to the following definition: Two objects X_1 and X_2 are color-

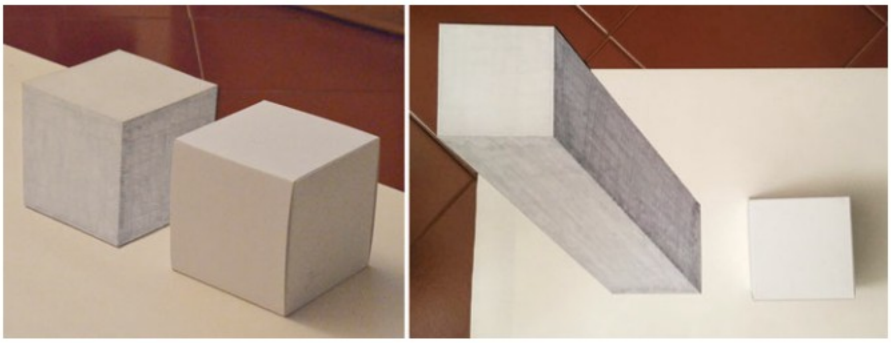

Fig. 11 A 3D cube and a plane anamorph

anamorphic relative to point O and light sources L_1, L_2 if X_1 seen from O under light source L_1 is visually indistinguishable from X_2 seen from point O under light source $L2$. Hence we have a dependence on the light sources quite analogous to the dependence on O in conical anamorphosis.

Such a concept allows us to speak, for instance, of a ball under an incandescent bulb being color-anamorphic to a disc under a halogen lamp, the disc being painted so as to simulate the color gradations of the ball. We will not, however, attempt here any formal development of this, but the concept is intuitively what any painter does when he paints a daylight scene to be seen under gallery lights and, more analytically, what a restoration expert has in mind when making sure that he uses not just the right color but the right pigment to restore a patch of color, so as to ensure color matching under change of light source spectrum.

We can sum up the observations of this section by saying (unpardonably boldy and omitting many qualifiers) anamorphosis is what makes drawing possible; color anamorphosis is what makes painting possible.

Binocular Anamorphoses

The abstract eye of anamorphosis is cyclopic, yet we can still easily work with binocular vision. Anamorphosis acts before the light hits the eye – it is a manipulation not so much of the eye, but of the light that reaches it. For this reason, arguments too concerned with what happens at the retina or the brain are missing the point. Whatever the complex workings of the visual system, feeding it equivalent inputs will result in equal perceptions. Hence, we don't need to understand how the brain combines binocular images; we just need to fool each eye separately, by figuring out what light packets it would be receiving from its position if the imaginary object was present. Then the brain will take care of the mixing. The problem is reduced to the technical one of showing each eye its own anamorph, built in the standard way, tailored to its specific position. The biggest problem is a suppressive one: to make sure that each eye *only* gets its corresponding picture. This can be achieved simply (but with limitations) by using a physical wall (septum) to

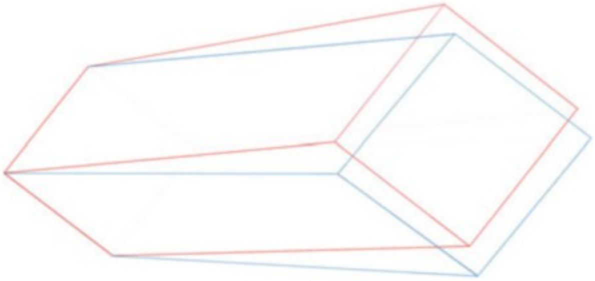

Fig. 12 Anamorphic anaglyph of a cube, to be viewed with red-blue 3D glasses

constrain each eye to its own compartment. This simple method, used by Dutour (1760) in the eighteenth century, is still an element used in contemporary solutions. Wheatstone (1838, 1852) achieved the separation through his mirror stereoscope in 1832 (Wade and Ono 2012). From that idea comes a long line of evolutions and simplifications, through the Holmes stereoscope in the 1860s and then the view-master in the late 1930s. These are not very different from modern VR viewers like Google Cardboard, except that today the illusions can move. All of these use in some measure either diffraction or reflection, so strictly speaking they are not working according to conical anamorphosis (although the distinction is superficial). A more strictly conical device is the anaglyph. Perspective anaglyphs were first published by Vuibert (1912) in a book with drawings by Henri Richard (Cabezos Bernal 2015). We are all familiar with the red-blue or red-cyan 3D anaglyphic glasses. The filter over each eye eliminates the opposite image, therefore separating the two anamorphoses, resulting in pictures that pop out of their frames. Doing the same procedure with an “oblique anamorphosis” instead of a “perspective” creates a far more startling result (Araújo 2017a; Cabezos Bernal 2015). For instance, Fig. 12, seen from the correct point, at a sharp angle to the page, will look like a cube popping out of the page, and young students get surprisingly amazed – in spite or perhaps because of their familiarity of computer graphics – to see what resembles a wireframe cube floating midair and wobbling as they move their head slightly. Moving the head in a vertical over the observation point gives the illusion of a bar chart animation, with a parallelepiped growing before one’s eyes. We will not remark further upon binocular vision, as it can be handled by this adaptation of monocular considerations. The reader can refer to Cabezos Bernal (2015) for more on this matter.

Anamorphosis Formally Reformed

Mathematical Preliminaries

We now proceed to the mathematical development of what we loosely described above.

We need some geometric background. We will model our three-dimensional world abstractly as an Euclidean three-dimensional space, denoted $\mathscr{E}$. You can think of $\mathscr{E}$ as $\mathbb{R}^3$. A point O of $\mathscr{E}$ will represent the location of the observer's eye. Let $\mathscr{E}_O = \mathscr{E}_O \setminus \{O\}$ denote the three-dimensional space except for point O. We will want to model objects. We define an *object* to mean a closed set of $\mathscr{E}_O$ and a *scene* to mean a finite union of objects. The term "closed" is a topological term. We recall some terminology briefly for the mathematician and then explain it at length. We beg the artist for bravery at this point. Hold the line for these will all be made more concrete later.

Definition 1. Let S be a subset of a topological space. We say that P is a *limit point* of S if every neighborhood of P contains a point Q of S other than P. The *closure* of S is the union of S with its limit points and is denoted by $cl(S)$. S is *closed* if $S = cl(S)$. The *residue* of S is the closure of S minus the set itself, denoted $Res(S) = cl(S) \setminus S$.

Intuitively: a limit point of a set is a point that can be approached indefinitely close without leaving the set. Closed sets are those that contain all their limit points. A point, a segment, a line, a plane, a circle, and a sphere are all closed sets. For a counterexample, take the set $S = \overline{AB} \setminus \{A, B\}$, a segment minus its endpoints. Both A and B can still be approached indefinitely close without leaving S, so they are limit points of S. But they are not in S; hence S is not closed.

We will find ahead sets that naturally will be missing limit points. For various reasons we do not like that, so we will add them in. Mathematicians have a fancy name for adding in those missing points – they call it *taking the closure*. The *closure* of S is the smallest closed set that contains it. So, going back to our example of the segment without endpoints S, the closure of S, denoted by $cl(S)$, is the full segment $\overline{AB}$, which contains all its limit points and is therefore a closed set. Finally, we call residue of S to the set of limit points that we have to add to a set S in order to close it. So, $Res(S) = cl(S) \setminus S$. In our example, the residue was the set of the missing endpoints of the segment, $\{A, B\}$.

We cannot really speak about closed sets, compacts, and so on, without having a topology. A topology on a set is a specification of which subsets are open (open sets are the complement of closed sets). In what follows we will be interested in studying the set of visual rays from the eye O. We will see how to put a topology on the set of these rays. First let us define what a ray is.

We say that a set S is *convex* if for any two points A and B of S, the segment $\overline{AB}$ is in S.

A point O on a line l divides $l \setminus \{O\}$ into two disjoint convex sets l_1, l_2. We say that l_1 and l_2 are each a ray over the line l, with origin (or vertex) in O. We interpret rays as directed half-lines going from O to infinity (in the case of l_1 and l_2, going in opposite directions from O). Note that rays are missing the point of origin O. Vectorially, if $\overrightarrow{v}$ is a vector, a ray is the set $O + a\overrightarrow{v}, a \in \mathbb{R}^+$. For $P \neq O$, $\overrightarrow{OP}$ denotes the ray with origin in O that passes through P. It is the same notation as for a vector, but we will allow context to disambiguate. Note that the usual definition of ray includes the origin point; ours doesn't because we don't want different rays to intersect each other. So a ray with origin in O is contained in $\mathscr{E}_O$.

Let $\mathscr{R}_O$ denote the set of all rays from O. Let S^2_O denote the unit sphere with center O. There is a canonical isomorphism between the set of rays and the sphere: we identify each ray with the point where it intersects the sphere. This bijection endows the set of rays with the topology of the sphere, giving it a notion of closed sets, limit points, and so on. We will therefore freely refer to rays or points on the sphere interchangeably.

Anamorphosis as a Mathematical Object

The principle of radial occlusion defined above expresses a notion of equivalence between points with regard to a reference point O, the eye of the observer. When the principle is valid, two points that are geometrically aligned along a ray from the eye are equivalent as far as visual perception goes. It follows that two objects should be equivalent if all their points are aligned or, equivalently, if they generate the same set of visual rays from O. We express this set as follows:

Definition 2. We say that the visual cone of an object Σ relative to O is the set of rays subtended by Σ with origin in O, which we denote $C_O(\Sigma)$. Hence $C_O(\Sigma) = \{\overrightarrow{OP}, P \in \Sigma\}$.

Now, from radial occlusion, it is natural to define that two sets "look the same" if they have the same visual cone. Later we will show that a somewhat looser equivalence is more interesting, which we will call *anamorphosis*. For now, for the purpose of motivation, we will concentrate on the stricter notion of conical equivalence.

Definition 3. We say that two objects Σ_1 and Σ_2 are *conically equivalent* relative to O if they have the same visual cone relative to O. We write this as $\Sigma_1 \overset{O}{=} \Sigma_2$.

It is clear that conical equivalence is indeed an equivalence relation between 3D objects, meaning a relation that is reflexive, symmetric, and transitive. We are interested in studying the classes of equivalence by this relation and how to construct the members of these classes.

Proposition 1. *Let P, Q be points of $\mathscr{E}_O$. $P \overset{O}{=} Q$ if and only if P and Q are on the same ray from O.*

Proposition 1 shows that the principle of radial equivalence for pairs of points is a particular case of Definition 3. This trivially follows from the fact that the cone of a point P is the ray $\overrightarrow{P}$.

Proposition 2. *if P, Q, P', Q' are points of $\mathscr{E}_O$ such that $P' \circeq P$ and $Q' \circeq Q$ then $\overline{PQ} \circeq \overline{P'Q'}$.*

Proof. By Proposition 1, triangles POQ and $P'OQ'$ define the same angle with vertex at O. Each ray inside the angle intersects sides $\overline{PQ}$ and $\overline{P'Q'}$ at a single point, which establishes a bijection between both the sides and the rays of the cone.

We can state this as the cone of rays of a segment AB is the angle $\angle AOB$.

Remark 2. The proof does not work for lines as the triangles do not capture the full cone of the line. In fact the proposition is false for lines.

Corollary 1. *If γ is a continuous curve with parametrization $g : [0, 1] \mapsto \angle AOB$ with $g(0) \circeq A$ and $g(0) \circeq B$ then $\gamma \circeq \overline{AB}$.*

Remark 3. This is the case of curve $A''B''$ in Fig. 3.

Proposition 3. *if ΔPQR and $\Delta P'Q'R'$ are triangles with vertices in $\mathscr{E}_O$, such that $P' \circeq P$, $Q' \circeq Q$, and $R' \circeq R$ then $\Delta PQR \circeq \Delta P'Q'R'$.*

Proof. For the sides of the triangles, the result follows from Proposition 2. Let I be an interior point of one of the triangles, say triangle ΔPQR. Then the ray from A through I finds a point J on $\overline{QR}$. Since $\overline{QR} \circeq \overline{Q'R'}$, there is a point J' on $\overline{Q'R'}$ such that $J' \circeq J$.

As in the previous proposition, this does not generalize to the full plane defined by a triangle.

Note that segments are convex combinations of two points and triangles are convex combinations of three. The result is valid for general convex combinations.

As a corollary of the propositions above, if Σ is a graph composed of vertices and segments (say, the cube of Fig. 3), we can obtain slide every vertex freely along its ray from O, and the graph obtained will have the same cone as the original graph. The same is true of the polyhedra obtained from the graph by triangulation (there are several choices of flat faces compatible with a graph). Moving the vertices will create a new polyhedron, equivalent to the first.

Corollary 2. *If S is a continuous surface with boundary on a continuous closed curve anamorphic to triangle $\triangle ABC$ and S is contained in the solid angle defined by $\triangle ABC$, then $S \circeq \triangle ABC$.*

Remark 4. When dealing in computer graphics, the propositions above can be used to obtain anamorphs very easily from a reference object. For instance, it is very easy, using the Geogebra software (Hohenwarter 2002; Hohenwarter et al 2013), to slide points on their rays from O to obtain anamorphic polyhedra as in Fig. 3 and even to deform these sides into simple curves. Sánchez-Reyes and Chacón (2016) do something analogous to obtain anamorphic deformations of a reference object using Bézier triangles and NURBS.

As we have seen, Proposition 2 does not generalize to full lines. A line l on a plane H divides the plane into a disjoint union of convex sets $\pi = \pi_1 \cup l \cup \pi_2$, where π_1 and π_2 the half-planes on either side of l. The cone of l is characterized by the following proposition.

Proposition 4. *Let l be a line in $\mathscr{E}_O$. O and l define a plane $\pi_O(l)$. Let l_O be the line parallel to l passing through O. l_O divides $\pi_O(l)$ in two disjoint half-planes. Then $C_O(l)$ is the set of rays through the half-plane of $\pi_O(l)$ that contains l.*

We remark that although l is closed in $\mathscr{E}$, its cone is not closed in $\mathscr{R}_O$. This will be important ahead. There are two rays missing to make the cone closed, these being the two diametrically opposite rays whose union is l_O. These mark the limits of the cone of l but do not belong to it. Intuitively, the line of sight approaches parallelism to l more and more as it follows it to infinity, but never reaches it. The reader probably recognizes line l_O as the line that in classical perspective defines the vanishing point of l by intersection with the plane of projection. Here we have no plane of projection, but this line is still important:

Proposition 5. *Let l and l' be lines in $\mathscr{E}_O$. Then $l \overset{O}{=} l'$ if and only if l and l' define the same plane with O and $l_O = l'_O$.*

In linear perspective line l_O is used to construct the projection of l. The construction is made by joining the intersection of l and l_O with the projection plane. This is a fundamental theorem of linear perspective (Andersen 1992, page 12), but it cannot even be stated for anamorphoses since we have no projection plane. Further we will see ahead that a general projection surface will not intersect l anyway. But we have the following analogous construction that uses l_O and l to obtain an arc of circle that is related to l in a canonical way.

Proposition 6. *There is a single circle $\mathscr{C}$ of center O that is tangent to l. Let $\mathscr{C}_O(l)$ be the intersection of $\mathscr{C}$ with the half-plane of $\pi_O(l)$ on the same side of l_O as l. Then $\mathscr{C}_O(l)$ and l define the same cone of rays from O.*

Proof. The perpendicular to l from O intersects l at the point I which is minimizes the distance from l. Hence the circle through I of center O will be tangent to l (Fig. 13). This circle intersect l_O at two points V and $V^\star$ diametrically opposite from O. Consider the half-circle obtained from $VIV^\star$ by excluding its endpoints V and $V^\star$. There is a one-to-one correspondence between points on this arc and points on the cone $C_O(l)$, obtained by intersection of the arc with $\overrightarrow{OP}$, for each P in l.

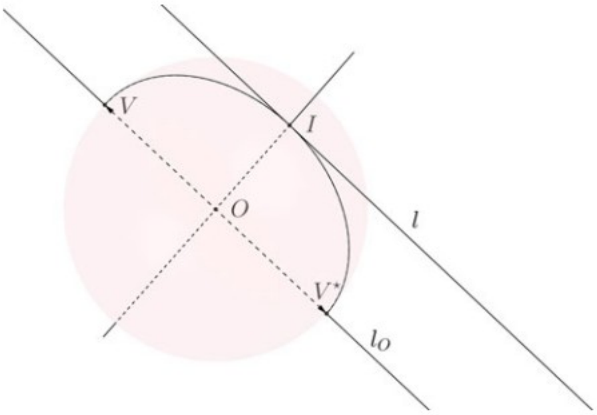

Fig. 13 The canonical meridian representative of a line's visual cone

This is nice result at first sight. There is a canonical arc of circle that has the same cone as the line. But there is a problem. The reader may have noticed that we didn't say that these were equivalent objects. That is false both for the arc with the endpoints and for the arc without them. The arc without the endpoints has the same cone of rays as the line, but it is *not* a closed set; hence it is not what we defined as an object. The arc with the endpoints is closed, but it is not conically equivalent to the line, as the rays corresponding to the endpoints are not a part of the line's cone.

We could solve this in several ways, for instance, by changing the definition of object, but the most satisfying seems to be to change our notion of equivalence. Hence we define anamorphic equivalence by slightly relaxing conical equivalence:

Definition 4. We say that two objects Σ_1 and Σ_2 are *equivalent by conical anamorphosis* relative to O; their visual cones are equal up to topological closure, that is, if $cl(C_O(\Sigma_1)) = cl(C_O(\Sigma_2))$. We write this as $\Sigma_1 \overset{\circ}{\sim} \Sigma_2$.

Some nomenclature, "equivalent by conical anamorphosis relative to O," is a mouthful, so for short we'll just say that two objects are *anamorphic*, or that they are *anamorphs* (of each other), or that one is an anamorphosis of the other. This definition fulfills well both the common usage of these expressions and our technical needs. These concessions to brevity and custom should not make us forget that any anamorphic equivalence is always relative to a specific observation point, so each of these phrases should be mentally ended with "relative to O."

Proposition 7. *If Σ_1, Σ_2 are objects such that $\Sigma_1 \overset{\circ}{=} \Sigma_2$, then $\Sigma_1 \overset{\circ}{\sim} \Sigma_2$.*

Obviously, conic equivalence implies anamorphic equivalence. Hence the results we obtained for conical equivalence above are still valid for anamorphosis.

Now let us reconsider the canonical semicircle we defined above. If we add to it its endpoints, by taking its closure, it becomes an anamorph to the line l. These two points are very important. They are the intersection of l_O with the sphere; hence they are the same for every line parallel to l. Hence they will be meeting points for all lines parallel to l. The two rays of l_O generalize the notion of vanishing points from classical perspective.

We can generalize the construction we did for the line, so as to get a canonical anamorph of any object, as well as a vanishing set:

Definition 5. We say that the *anamorphosis* of an object Σ relative to O is $\Lambda_O(\Sigma) = cl(C_O(\Sigma))$. We identify it with its projection onto the unit sphere, $cl(C_O(\Sigma)) \cap S^2_O$, when context is clear. The latter we also call the *canonical anamorph* of Σ.

It follows trivially from Definition 4 that an object is anamorphic to its anamorphosis thus defined. Its spherical anamorph is called canonical because it is the most natural of its 2D anamorphs, arising from the identification of the space of rays with the sphere.

Definition 6. We say that $\mathscr{V}_O(\Sigma) = Res(C_O(\Sigma)) = cl(C_O(\Sigma)) \setminus C_O(\Sigma)$ is the *vanishing set* of object Σ relative to O. We call *vanishing* points to the points of the vanishing set. We identify the rays of the vanishing set with the corresponding points on the unit sphere S^2_O.

The vanishing set is the residue of the visual cone of the object, that is, the points you must add to make it closed. This definition has the advantage that it works for any object and, when restricted to lines and planes, reduces to what you would expect from the classical definition, as we will see.

Remark 5. When teaching the concept of vanishing points to artists, it is hard to deal with topological concepts explicitly, especially on a short timeframe, such as a workshop. Then it is more practical to generalize the tradition of perspective started by Taylor and simply define vanishing points of lines as the points associated with the two opposite rays that make up l_O and then notice that these are at the end of the canonical semicircle; then it follows from the definition that parallel lines have the same vanishing points and that these are meeting points for those lines.

Proposition 8. *The canonical anamorph of a line $l \subset \mathscr{E}_O$ is a meridian of the unit sphere. The vanishing set of l is a pair of antipodal points located at the ends of the meridian, where l_O, the parallel to l through O, intersects the sphere.*

This proposition turns perspective into spherical geometry, so we can apply the tools of that discipline. See Catalano (1986) and Araújo (2018c). In particular we will be talking a lot about antipodal points, so let us recall that two points are called *antipodal* if they are diametrically opposite on a sphere. Given a point P on the sphere we denote its antipodal point by $P^\star$. The notion extends trivially to the corresponding rays through the center of the sphere.

This proposition has the obvious corollary that not only do parallel lines have the same vanishing points, but these are actual meeting points of their anamorphosis. This is a very concrete alternative to the abstract notion of parallel lines meeting at infinity – their spherical anamorphs actually meet on the sphere.

Corollary 3. *If l and l' are parallel lines, their canonical anamorphs intersect at their vanishing points.*

This is a beautifully unifying view of vanishing points, unlike the classical perspective view where a line has sometimes one and sometimes no vanishing points, or even the (hemi)spherical perspective of Barre and Flocon (Barre and Flocon 1968) where lines have either one or two vanishing points. This search for constancy in the number of vanishing points can be compared to the insistence of mathematicians in redefining numbers to ensure that an n-degree equation always has exactly n roots. The projective plane view of perspective answers the same desire for symmetry, but there we have always a single vanishing point. Our present notion of anamorphosis (and later of perspective) splits that point in two as we want no preferential axis and wish to capture the whole visual environment.

Now we consider the matter of construction of these anamorphs. Since the canonical anamorph of a line is an arc of circle, we might expect it to require three points for its construction. But since it is a meridian, and the two vanishing points are antipodal, two points are enough.

Recall that two points on a sphere define a plane through its center. The intersection of that plane with the sphere is a circle with the same radius as the sphere and to this is called a *great circle* of the sphere, or a *geodesic* . A meridian is a connected half of a great circle, so it can be specified by choosing two antipodal points and then picking one of the two possible halves of the geodesic. This picking of a meridian can be done by choosing a point in its interior. But since the two endpoints are mutually antipodal (hence knowing one is enough), all we really need to know to determine a meridian is two points of it, as long as one of them is specified to be an endpoint. For the same reason, only two points are needed to determine the image of a line, as long as one of them is specified to be a vanishing point.

Proposition 9. *The anamorphosis of a line l of $\mathscr{E}_O$ is determined by one vanishing point and by the projection on the sphere of one point of l.*

Proof. The vanishing point V and the measured point P' projected from l onto the sphere determine a plane through O, hence a geodesic. On this geodesic, the line image will be the meridian $VP'V^{\star}$.

Note that if the line l crosses O then its image is just a set of two points, these being located at the place of the vanishing points, which in this case are actual points of the projection. It is the line itself that vanishes. Taylor remarked on much the same situation in classical perspective, and that is one of the justifications for the terms "vanishing point," as the line "vanishes" into it in this degenerate case (Andersen, 2007, page 175). There is a technical difference in our definitions here, since by Definition 6, these two points are not in our vanishing set at all, while in classical perspective the vanishing point exists and coincides with the single projected point.

Example 1. In Figs. 14 and 15 we see two examples of pairs of spatial lines l and j relating to their vanishing points and to their spherical anamorphs.

In Fig. 14 we see two parallel lines. They have therefore the same translation to the origin, so $l_O \equiv j_O$; hence they have the same pair of antipodal vanishing points, V and $V^{\star}$. Their canonical anamorphs on the sphere are meridians ending at these vanishing points. Hence the vanishing points are actual meeting points for the canonical anamorphs.

As for construction of these spherical anamorphs, two projected points P and Q from l and j, respectively, define each meridian, as in Proposition 9. These points may be measured anywhere, but one way would be at the intersection of the meridian with the geodesic plane orthogonal to l_O.

In Fig. 15 we see two perpendicular lines. Now lines l_O and j_O are themselves orthogonal, crossing at O, and we get two pairs of vanishing points. The meridians are obtained from these by measuring a single additional point, the common point where the two lines cross. Hence we see that only vanishing point V_1 and point P need be measured to obtain the whole construction.

In Fig. 16 we see the spherical anamorph of a cube. The cube defines a set of six distinct vanishing points, which are meeting points for the lines that prolong the edges of the cube.

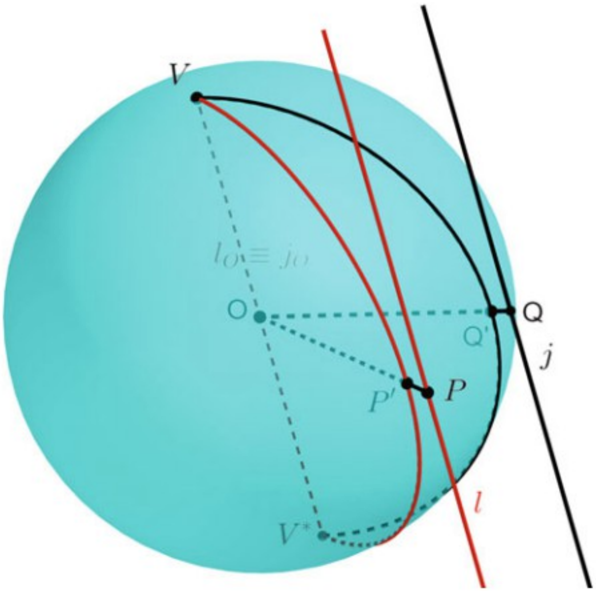

Fig. 14 Anamorphosis onto the sphere of two parallel lines. The lines project as meridians that meet at two common vanishing points V and $V^{\star}$, antipodal to each other

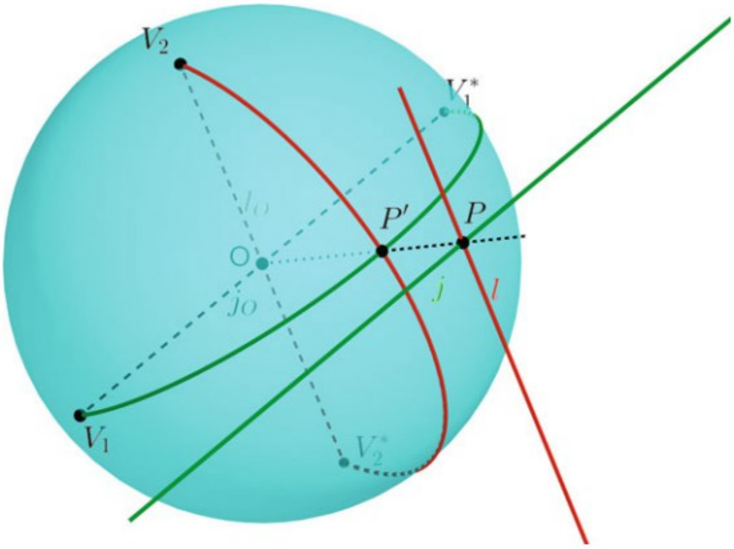

Fig. 15 Two lines l and j intersecting at a right angle. Their canonical anamorphs on the sphere go to pairs of vanishing points distributed at a regular separation of 90 degrees along the geodesic defined by the orthogonal lines l_O and j_O

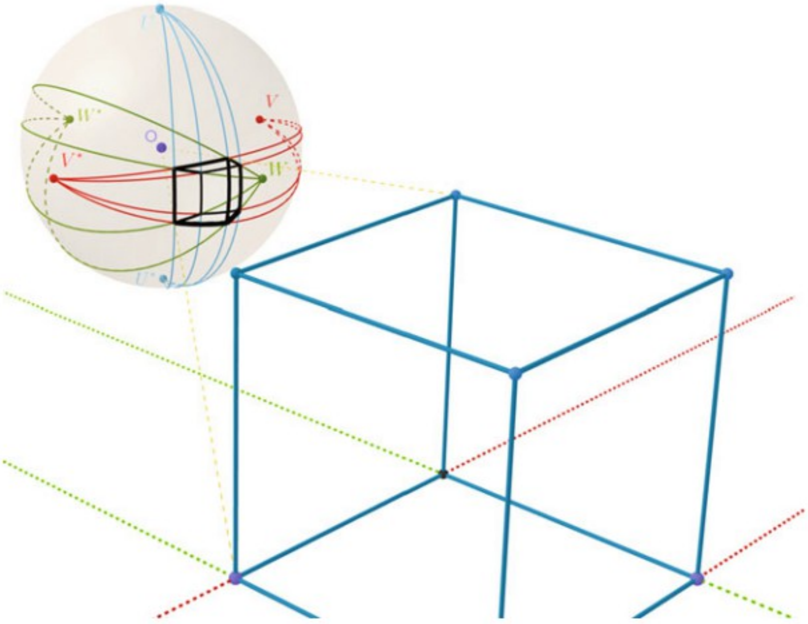

Fig. 16 Anamorph of a cube on a sphere. The cube defines a set of six distinct vanishing points, two for each set of parallel edges

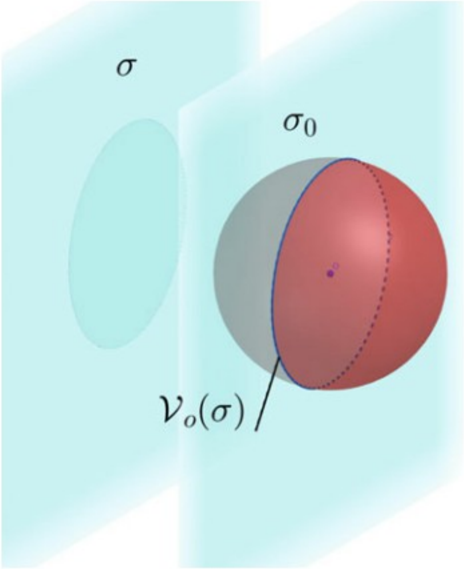

Fig. 17 A plane σ and its vanishing set $\mathcal{V}_O(\sigma)$ on the unit sphere

Proposition 10. *Let σ be a plane not containing O. Let σ_O be the plane parallel to σ through O. Then σ_O defines a great circle of the sphere, which is the vanishing set of σ (Fig. 17). This great circle divides the sphere in two hemispheres. The anamorphosis of σ is the hemisphere on the side of σ_O that contains σ.*

Proof. The construction is analogous to that of the canonical circle. Take a perpendicular from O to σ, and this defines a sphere tangent to σ, half of which will be the image of the visual cone of σ. To get the anamorphosis, scale it down to unit radius.

The vanishing points of lines and planes are quite analogous to the linear perspective case, in that they are obtained by translation to the observation point followed by intersection with the projection surface. But our Definition 6 applies also to general objects that cannot be treated in this fashion, such as the following example.

Example 2. Let l be the curve $t \mapsto (t, t^2, 0)$ and let $O = (0, 1, 0)$. Then the cone of l will the great circle defined by $z = 0$, minus point $(0, 0, 1)$. But the closure of this set on the sphere is the whole of the great circle. So the vanishing set is $(0, 0, 1)$ and the anamorphosis is the great circle through $z = 0$.

Example 3. Let S be the surface $z = x^2 + y^2$ and let $O = (0, 0, 1)$. Then the cone of S will the whole sphere, minus the point $(0, 0, 1)$. But the closure of this set on the sphere is the whole of the sphere, so the vanishing set is $(0, 0, 1)$ and the anamorphosis is the sphere itself.

This is a different perspective from the more traditional notion of "following the line to infinity." Although the vanishing set is the same, we can look for the vanishing points in the anamorphic projection itself, by looking for "missing points" or, if you will, at "vanished" points. This is also a better definition when we consider sets whose limit rays cannot be obtained by simple curve parametrizations like in our second example above.

We would now like to define an anamorphosis onto an arbitrary surface and its vanishing points.

In what follows we will define surface to mean a manifold of dimension two, that is, something that locally maps to a region of a plane. These will usually be manifolds with boundary, such as a sheet of paper or a half-sphere. Most of the time they will be smooth, but we only require them to be topological manifolds. Most of the time they will be connected, but not always.

We use the canonical anamorphosis onto the sphere as the blueprint from which all others are derived as projected images, usually partial ones. The idea is simple: the anamorphosis of an object onto a surface is just the closure of the conical projection of the object onto the surface. The vanishing points on the surface will be the subset of the canonical vanishing points that happen to fall on the surface. Technically, it takes some care not to count points multiple times if the surface has a complex shape that folds over itself or has several connected components.

To avoid these problems, and because that work is centered on spherical perspectives, Araújo (2018c) defines the anamorphosis only for compact starred surfaces, which are just radial deformations of the sphere. These are quite enough for most purposes, so we start with them and then discuss possible generalizations briefly.

Definition 7. We say that a compact surface S is starred relative to O, or that it is an O-starred surface, if every ray from O touches S at most once. A compact surface is said to be locally O-starred at $P \in S$ if there is a neighborhood B of P such that every ray through B intersects $S \cap B$ at most once.

You can think of an O-starred surface as being locally defined by $f(u) = h(u)u, u \in S^2_O$ where $h : U \subset S^2_O \to \mathbb{R}^+$ on a region U of the sphere S^2_O. Intuitively, $h(u)$ represent a ratio that pushes each point of the sphere closer or further away from the center O.

We can define anamorphosis onto an O-starred surface S as follows.

Definition 8. The anamorphosis relative to O of an object Σ onto an O-starred surface S is $\Lambda_{O,S}(\Sigma) = cl(C_O(\Sigma) \cap S)$. The vanishing set of Σ on S is $\mathscr{V}_{O,S}(\Sigma) = \Lambda_{O,S}(\Sigma) \setminus (C_O(\Sigma) \cap S)$.

Since the conical projection is a continuous map onto an O-starred surface, we have the following result.

Proposition 11. *Let S be a locally starred surface at V. Then V is a meeting point for the anamorphic images of lines with vanishing point $\overrightarrow{OV} \in \mathscr{R}_O$.*

When S is a starred surface, we can express these objects very simply as intersections.

Proposition 12. *Let S be an O-starred surface. Then $\Lambda_{O,S}(\Sigma) = \Lambda_O(\Sigma) \cap S$, and $\mathscr{V}_{O,S}(\Sigma) = \mathscr{V}_{O,S}(\Sigma) \cap S$.*

That is, the anamorphosis (resp. vanishing set) of Σ on S are just the intersection of the rays of the anamorphosis (resp. vanishing set) with S. When considering the actual construction of an anamorphosis on a surface, we can construct one anamorph – the simplest one we can find – and then project it onto the required surface. In particular, the canonical anamorphosis onto the unit sphere is a good candidate as a prototype for others, since it is so symmetrical. So it is natural to solve the anamorphosis there and then project it conically onto the required surface, using the fact that $\Lambda_{S,O}(\Sigma) \overset{\varrho}{\sim} C_{0,S}(\Lambda_{S^2,O}(\Sigma))$.

For this reason, we are compelled to study the projections of lines and planes on the sphere with special interest, as well as its projections onto various kinds of surfaces. Lines, as we have discussed, project as meridians onto the sphere, each meridian ending at two antipodal vanishing points. We will now consider how these meridians in turn project onto various surfaces.

Example 4. Consider the anamorphosis of a line l onto a compact region S of a plane $\pi \in \mathscr{E}_O$: this can be identified with (a compact subset of) linear perspective or with so-called oblique anamorphosis. Let π_O be the plane through O such that $\pi_O \parallel \pi$. l projects on the sphere as a meridian with vanishing set $\{V, V^\star\}$. Suppose the vanishing set is on π_O. Then if l is on the side of π, it projects onto a line; if on the other side, it projects as the empty set. In either case the rays corresponding to the vanishing points do not intersect π, so the line has no vanishing points on π. If the vanishing points are not on π_O, then the line intersect the plane π at a point I and one and only one of the vanishing points projects onto π as a point V_π. Then the line projects into the ray with origin at V_π that passes through I, and its vanishing point on π is V_π. Since we assume S to be a compact (hence bounded) subset of π, both the vanishing point and point I (as well as an unbounded section of the line projection) may be actually outside the anamorphosis proper, but they can still be used for the construction.

Remark 6. We can contrast classical perspective (or plane anamorphosis) with spherical anamorphosis by saying that in the former a line projects either as a line or as a ray, while in the latter a line projects as a "segment" (i.e., a meridian ending at vanishing points).

Example 5. Suppose S is a compact region of three intersecting planes disposed as the floor and walls in the corner of a room (Fig. 6). This is locally the same as a plane anamorphosis, but globally very different, as some lines may have two vanishing points. In rendering lines, we may consider the following strategy: in each plane plot two points of the line, thus finding two lines joining at the intersection of the projection planes. More efficiently, find only three points per line, one point on each plane and one on the intersection of the planes. But since two non-antipodal points define a geodesic, in fact two points anywhere on the planes must be enough to define the line across all projection planes. These constructions are maximally symmetric in the case of the anamorphosis onto a cube, which has been studied in Araújo et al (2019b) as a special case of spherical perspective. Then the geodesic associated with a line l projects as a set of six or of four segments on the cube, obtainable from any two points through simple descriptive geometry constructions.

Example 6. Let S be a compact region of a cylinder, with O outside the cylinder (Figs. 18 and 7). This is a good example of how we can make do with only starred surfaces. Although the cylinder is not starred from O we can consider as projection

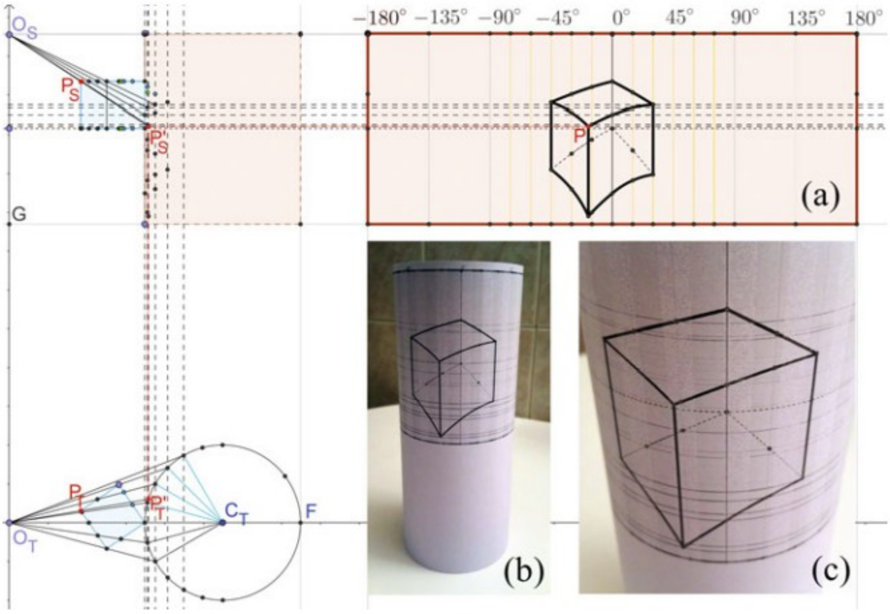

Fig. 18 Descriptive geometry construction of the anamorphosis of a cube onto a cyilinder. Each point is found on the top view (Bottom left) then lifted to the side view (Upper Left) and found on the developed (cut and unrolled) cylinder by transforming the angular coordinates from the top view and transporting the heights from the side view (**a**). The rectangle can be cut and rolled up to obtain the cylindrical anamorphosis (**b**) which will be correctly observed from the point O determined by the two projections O_S and O_T. From this point the edges of the cube appear straight (**c**)

surface only the proximal region of the cylinder, which is visible from O and exclude the distal, occluded part, thus obtaining a starred subset. A generic non-vertical line will define a plane that cuts the cylinder in an ellipse. A line will project as an arc of an ellipse, with at most one vanishing point. A large class of lines will have no vanishing points at all on the cylinder, as l_O will not intersect it. If the cylinder were to be cut at a vertical line and opened isometrically (unrolled), the ellipses would transform into segments of sinusoidal curves. See Apostol and Mnatsakanian (2007).

Example 7. Let S again be a cylinder as in the previous example, but now O is on its axis (Fig. 19). Now generic geodesics will project onto ellipses that are half above and half below the horizontal plane through O. The line will project as an arc of its ellipse, ended by two diametrically opposed vanishing points. Lines on the horizontal plane through O are special cases, projecting as horizontal half-circles. Lines with vanishing points along the axis of the cylinder will project as verticals.

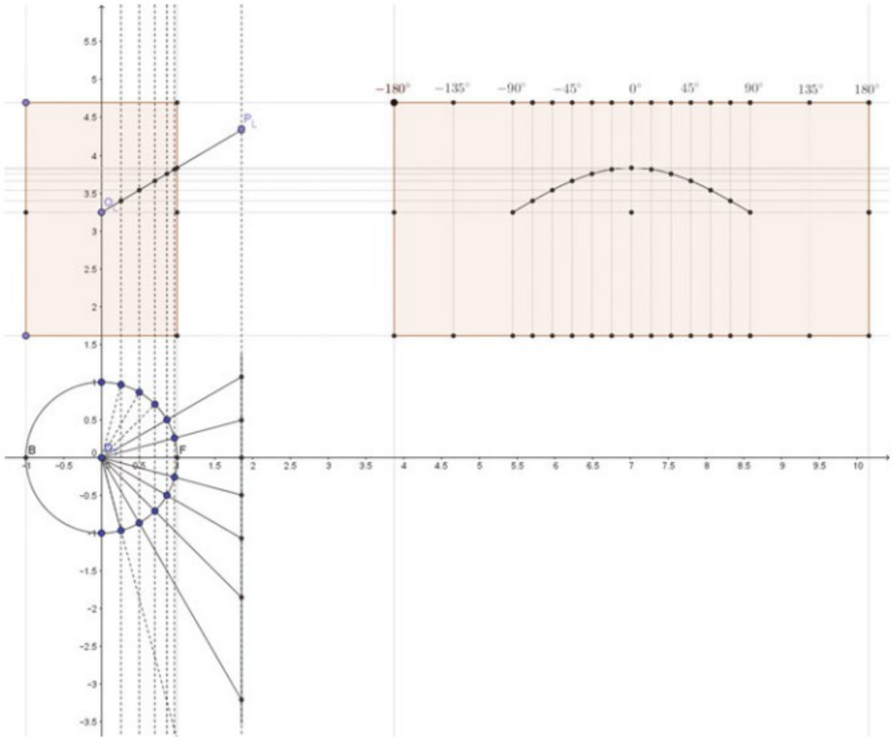

Fig. 19 Descriptive geometry construction of the anamorphosis of a line onto a cylinder with O on the axis. The image of the line on the unrolled cylinder is one half of a sinusoid whose axis of symmetry is the horizon line, at the height of O

All geodesics, projected as sinusoids in the opened cylinder, will share the same horizontal axis of symmetry, as the planes of their geodesics all intersect the axis of the cylinder at O. This was not true in the previous case, where the plane of a generic line could intersect the cylinder's axis at any point.

More General Surfaces

The following section is technically nitpicking that should be ignored by the artist (and probably by everyone on a first reading).

Although starred surfaces are quite enough for most purposes and provide the most elegant theoretical results, they can hardly model the large variety of real surfaces on which artists have constructed anamorphoses. Often these can be rather complex surfaces that fold over themselves in such a way as cut a ray from O multiple times; hence they are non-starred. Often they will have several connected components, sometimes so many that they might be better modeled as a point cloud.

It is in fact quite up for debate what the most desirable definition would be for a general anamorphic projection surface. We will not aim at closing that discussion here but just at exploring it a little.

First of all, we note that we could certainly extend our definitions to include all sorts of surfaces, even non-compact surfaces such as the whole Euclidean plane (in that case the formalism under consideration just becomes that of linear perspective). We would lose some of the elegance but things still work. It is interesting to consider what happens when we remain with compacts but relax conditions just a little.

One way to deal with non-starred surfaces is by cropping or changing the surfaces into their minimal starred equivalents by considering only the subset that is visible from O. We did this informally with the cylinder in Example 6. We can formalize this procedure by generalizing conical projection to take occlusion into account, in the following way: given a point P and a surface S, the *proximal conical projection* sends P to the intersection with S which is closest to O (the reasoning is that the distal ones are occluded).

Definition 9. The proximal conical projection from O to a surface $S \subset \mathscr{E}_O$ is the map from $\mathscr{R}_O$ to S defined by $\varphi_{O,S}(\overrightarrow{r}) = \{P \in \overrightarrow{r} \cap S : \forall Q \in \overrightarrow{r} \cap S, |OQ| \geq |OP|\}$

Remark 7. When the surface is O-starred, the proximal conical map reduces to the ordinary conical projection $\overrightarrow{r} \mapsto \overrightarrow{r} \cap S$. We identify the map from $\mathscr{R}_O$ with the corresponding map from $\mathscr{E}_O$ defined by $P \mapsto \overrightarrow{OP} \cap S$ that maps each spatial point to its conical projection on the surface.

Identifying point of S with their rays, we can use the proximal conical projection map to define the minimal starred equivalent of S itself, i.e., the smallest locally starred surface that has the same anamorphosis as S. In fact $\varphi_{O,S}(S)$ has no double points, preserving only the *proximal* (closest to O) intersection of each ray going

Fig. 20 The object composed by the rectangles A, B, C, D is not starred. Taking the proximal projection $\varphi_{O,S}(S)$ eliminates sections B and C, leaving two connected components A and D, minus segment h_2, which is recovered by taking the closure

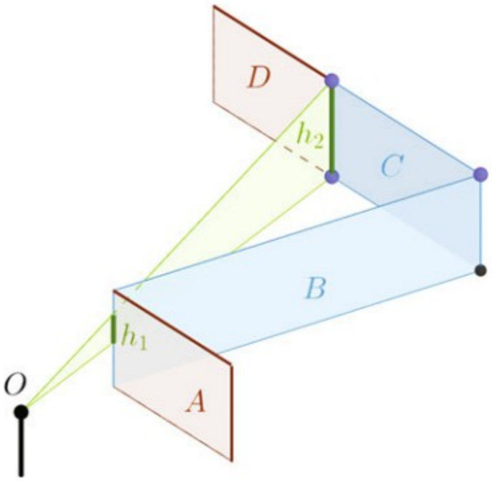

through a point of S. But since we insist on compact sets, we define the minimal O-starred equivalent of S to be $S^\star_O = cl(\varphi_{O,S}(S))$. Because of the possible existence of multiple connected components, this surface will generally have a non-empty set of repeated intersections, although one of measure zero. It will be locally starred *almost everywhere*, i.e., there will be a set δ of measure zero such that $S^\star_O$ is locally starred in $S^\star_O \setminus \delta$. We say that a surface with this property is *quasi-starred*.

Example 8. The object of Fig. 20, defined by the union of rectangles A, B, C, and D, is not O-starred. Taking the proximal projection $\varphi_{O,S}(S)$ eliminates sections B and C, which are occluded by A, leaving two connected components A and B, minus the segment h_2 of B, which is occluded by segment h_1 of A. This segment is recovered by taking the closure, so that $S^\star_O = A \cap B$ is the disjoint union of two compact rectangles. It is only quasi-starred, since $h_1 \overset{O}{\sim} h_2$ is a set of double points relative to O.

These double points have to be handled carefully. Continuing Example 8, consider Fig. 21. The vanishing points V and V', which are aligned with O, both belong to the surface and correspond to the same vanishing point on $\mathscr{R}_O$, but are meeting points for different sets of lines. We must take some care in defining vanishing points for general surfaces, to avoid duplication or the appearance of false vanishing points, as the equivalent of Proposition 12 is no longer valid.

Let us consider our requirements: the vanishing points of a set Σ on S are determined by projection of $\mathscr{V}_O(\Sigma)$, so clearly $\mathscr{V}_{O,S}(\Sigma)$ is a subset of the intersection of $\mathscr{V}_O(\Sigma) \subset \mathscr{R}_O$ with S. But this cannot be an equality otherwise we'd have double vanishing points. Also, we cannot define them as $\varphi_{O,S}(\mathscr{V}_O(\Sigma))$ as sometimes the proximal intersection with S of the vanishing rays are not the meeting

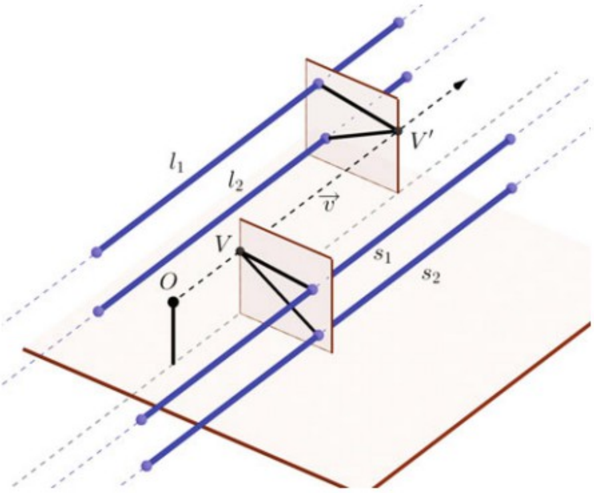

Fig. 21 Anamorphosis onto a quasi-starred set with two connected components. Points V and V' are in the ray of the same canonical vanishing point, but Definition 10 ensures that V (resp. V') is only a vanishing point for the family parallel lines s_1, s_2 (resp. l_1, l_2)

point of the projected lines, which is located in a distal connected component. The following solves all these difficulties:

Definition 10. The anamorphosis of an object Σ onto a projection surface S relative to O is $\Lambda_{O,S}(\Sigma) = cl(\varphi_{O,S}(C_O(\Sigma) \cap S^\star_O))$, where $S^\star_O = \varphi_{O,S}(S)$ is the minimal compact starred subset of S. The vanishing set of Σ on S is $\mathscr{V}_{O,S}(\Sigma) = \Lambda_{O,S} \setminus (C_O(\Sigma) \cap S^\star_O)$.

Let us see how this works with the example of Fig. 21. $S = S^\star_O$ is the union of two disjoint compact rectangles. Points V and V' are on the same ray from O, on the edge of the proximal and distal rectangle, respectively. Let Σ be the set of lines s_1, s_2. By Definition 10, the vanishing set is V', the meeting point of the lines projected on the distal rectangle, and V doesn't figure in either the vanishing set or the anamorphic image. That is because neither V nor V' are present in $\varphi_{O,S}(C_O(\Sigma) \cap S^\star_O)$, the latter only appearing when we take the closure. In the same way, V' will be ignored when calculating the vanishing set of lines s_1, s_2 by Definition 10, and only V will remain. This is why we define the vanishing set through topological closure and not mere projection from the canonical vanishing set. In this way we both avoid duplication of points and ensure that the vanishing set preserves its meaning as a location for the meeting of parallel line images.

Proposition 13. *If S is a surface, and its minimal quasi-starred equivalent $S^\star_O$ is locally starred at a point V, then V is a meeting point for the anamorphic images of lines that have the ray $\overrightarrow{OV}$ as a vanishing point.*

Note that if V is in $\delta \subset S^\star_O$ then it will be a meeting point for only a subset of those lines (see Fig. 21).

Simplifications: Talking to Artists

We would like to keep a balance between the concerns of artists and those of mathematicians. The technical details of Definition 10 are somewhat overwhelming for the former, so it behooves us to find some simplification. The trick is how to simplify without outright lying. The matter can be presented thus to an artist.

We are interested in drawing an anamorphosis of an object Σ onto a surface S. The surface can be anything. The artist is instructed that for the purposes of drawing, only the part that is visible from O must be considered. Points of S in the conical shadow of other points of S which are closer to O are to be discarded. The exception are points at the edge (boundary) of the shadow. These are allowed to remain. From this cropping we get a working surface S', the minimal surface, in which we will actually draw.

Each line l determines exactly two vanishing "points," V and $V^\star$, which are the rays from O that are parallel to l_O (you can also call them vanishing rays for clarity). These are always exactly two, and diametrically opposite to each other. These vanishing rays can be represented as actual points in a canonical way by intersecting them with the surface of an the imaginary "visual sphere" around O. When drawing an anamorphosis onto the minimal surface S', each of the two vanishing rays may intersect it several times. The multiple intersections of each ray with S' should be seen as multiple manifestations of the same vanishing point. So, for instance, V will manifest as points $V_1, \ldots, V_k$, which we can order by distance to O. Then all lines parallel to l which meet at V on the sphere will meet at one of the V_i on S'. Which one can be found a posteriori from drawing the lines near to their meeting points and seeing in which subsurface they are found (it will be the one that contains the projection of the germ of l around V – but here we'd be going all abstract again, so we won't).

This statement avoids the most difficult terminology and requires minimal abstraction. It can be easily explained graphically to even very young students, and yet it carries the essentials, as far as drawing is concerned, of the discussion above. In a later section we will see how to implement these concepts graphically through descriptive geometry constructions.

On Compactness

Why have we insisted on compact projection surfaces? If we think about it, this is quite strange. This was introduced by Araújo (2018c) for the needs of immersive perspectives as it is a very natural setting for the most usual such perspectives, namely, the cylindrical and spherical cases. Rather shockingly, however, it excludes classical perspective in its usual formulation, as the Euclidean plane, infinite and unbounded, is not a compact set. So it reformulates linear perspective onto a bounded region of the plane, though this region may be as large as you want it. You can imagine this region as a rectangle, for instance, representing a sheet of paper. Since we want it to be closed, it will technically be a manifold with boundary. Classical perspective will then become a degenerate limiting case as the region grows to infinity. This is analogous to the way in which orthogonal projection is usually considered a limiting case of conical projection as O becomes more and more distant from the projected object.

As we have seen in Example 5, classical perspective still remains, in our view, a fundamental object, as our constructions of the compact plane anamorphosis are better done considering the full plane, but only as a construction device, with the final result being cropped to a region that is bounded, though as large as we want it (reminding us of the Aristotelian philosophical distinction between *potential* and *actual* infinity). Even so, this is a frightful demotion of the infinite plane of classical perspective, so we must consider what we get in return.

Compactness is a desirable property. Ask a mathematician what linear perspective is as a mathematical object, and he will insist it is the projective plane. Like the sphere, the projection plane is a compact. Like the sphere, it gives you a definition of vanishing points that attributes a constant number of vanishing points to a line (this number being one, in the projective case, two in our case). This uniformity is pleasing to the mathematician's eye, like that of the number of roots in a polynomial equation of given degree. In the projective plane case, this is done by identifying the actual perspective drawing as a chart of the projective plane. Hence the drawing lives naturally in the infinite plane. On the other hand this implies that the vanishing point of lines parallel to the chart's plane actually lies nowhere in the chart. It lives only on the "point at infinity" of the chart, that is, on the one-dimensional projective line that lies outside of the chart's domain. We find that this is a rather immaterial satisfaction. The projective plane is not a surface that the draughtman can "see"; the gained vanishing point is more than a little ethereal. By contrast, the sphere is not only an abstraction, it lives in the actual scene to be drawn; it is as concrete a surface as the plane itself – in fact more so, because it is bounded. It is the natural object that requires the least abstraction, just a bounded surface of ordinary Euclidean space, and yet affords a more pleasing result than either the infinite plane or the comparably abstract projective plane, affording a distinction between directions that encompasses a fully immersive view.

What is formalized by the anamorphosis onto a compact surface is the notion of a *drawing* itself: the result is a real, drawing that can be materially executed and yet captures two vanishing points for every line. Mathematicians, of course, have no problem with the abstract notion of the infinite plane, but sometimes we want to deliberately keep to the finite. Just as we can abstract a notion of an infinite proof and yet for the most part decide that finite, actually executable proofs are our main object of interest, so here too, although plane perspective is certainly not a problematic concept – we could certainly accept the full Euclidean plane as a projection surface and still do our work – we decide that compact surfaces are the object of most interest to abstract the notion of drawing. This in fact corresponds neatly to the way in which plane perspective, or plane anamorphosis, is actually used by draughtsmen. Every drawing is made onto a finite canvas, although the canvas may be extended at will with no absolute limitation. And the draughstman knows the difference well between the bounded plane of the drawing and the unbounded plane that contains it and may be useful as a construction device though not part of the drawing itself. Often the draughtsman will have to use methods that find a vanishing point outside the drawing sheet in order to construct the lines inside it. And it is nice to have an abstraction of perspective that relates directly to the practical distinction.

Let us be more general: in a way we can say that compact – bounded and closed – objects are all that we can draw. Can we really draw a line? No. It is unbounded. We cannot draw infinite things. We draw a segment and end it with little dots to signify continuation. This is representing infinity, not drawing it. But when we project the line conically onto the sphere around O, it becomes a meridian. Can we now draw it? Not yet. As we have seen, it is not a closed meridian – it is missing its endpoints, as the vanishing points do not belong to the visual cone. We cannot draw open ended arcs. We can again only represent them, usually by drawing a small circles to signify that the ends are missing, but we cannot really perform the feat of drawing an arc without its endpoints. How to solve this? We could just add on those endpoints. In mathematics we do that and call it that fancy name: taking the closure. By taking the closure of the conical projection of the line, the cone becomes a closed half-plane, the meridian gets its endpoints. We now have a closed meridian on the closed and bounded – compact – sphere. A compact arc of circle. And compact sets are the right mathematical abstraction for that which can actually be drawn.

Descriptive Geometry Construction of Anamorphoses

In this section we will be concerned with constructing anamorphoses onto surfaces from the point of view of the draughtsman, using simple descriptive geometry. We focus on the mechanical device of a modified Dürer machine, both in physical form and abstracted through descriptive geometry constructions.

This is a matter intrinsic interest for our purposes, but we start with some lateral considerations of a didactic nature.

Handmade vs Digital Anamorphoses

It is undeniable that at present, digital means greatly simplify practical construction of anamorphosis, both in 2D, with *projection mapping* (Inglis 2018; Monroe and Redmann 1994) and in sculpture through 3D modeling and printing (we have already mentioned (Sugihara 2015a), but see also Ballegooijen and Kuiper's work in Dunham (2019)). These are the quickest and more efficient options for production purposes, and there is nothing wrong with that; the working artist knows that there is no such thing as cheating and anyway a vast gamut of anamorphosis would be unreachable or at least untried without such means. However, premature or excessive reliance on technological shortcuts not only lacks charm and intellectual appeal, but can lead to limiting results, as the tools that happen to be on offer define a scope of action and thought, the terms of the language of creativity, so to speak (Papert and Turkle 1991). This is always so, even if the tool is the brush, but digital tools, due to their complexity and opaqueness, tend to so in a more insidious manner. The expressive scope thus delimited may be inadequate for some artists , who then have little recourse to developers if they lack the understanding of the basics that is required for communication. The tendency will be for artists to adapt to tools rather than the other way around and end up chasing ever-changing flawed interfaces (Coates et al 2010; Norman 1990), so it is hard to say how much effort is really saved if time economies obtained by neglecting fundamentals are spent chasing transient knowledge of tools that never really *just work* as advertised (Kim and Chin 2019). The tendency is for these interfaces to help artists work with implicit principles they do not understand deeply, but there are limits to what can be achieved before the confused artist hits the walls of his ignorance. For instance, Tran Luciani and Lundberg (2016) reports both on the achievements of digital interfaces in helping artists create spherical panoramas and on the difficulties that always eventually arise from not knowing the essentials of the spherical perspective underlying the interface. Araújo (2019a) has argued that there is a need for more interfaces that force (and help) the artist to learn the fundamentals rather than helping them to avoid such learning; and that production software should ideally come in after the user has already mastered the discipline in its rudimentary form, so as to avoid the black box frame of mind. Further, the knowledge of the mathematical and geometrical fundamentals underlying their tools is better acquired, for most artists, not through formalism, but through manual practice in an embodied form. For this purpose, Araújo (2017b) proposed anamorphosis, in the form here discussed, as the central concept behind technologies such as virtual and augmented reality, 360-degree photography, and projection mapping; further, it proposed the modified Dürer machine, both physical and "virtualized" through descriptive geometry constructions, as an adequate physical embodyment for artistic exploration of the principle. This resulted in a teaching process that has been tested in students of a Ph.D. program in digital media arts and yet also, reworked and simplified, as a new way of introducing young students to the subjects of linear perspective and descriptive geometry (Araújo 2017c). In the latter case, making anamorphosis the fundamental concept and linear perspective the derived concept, in a reversal of the usual practice, and expressing this concept with constructions that are handmade, yet display their anamorphic properties through digital means (such as photography and VR), resulted in an enticing process to students that have grown with a view centered on digital media. This has shown promise in preliminary tests with Portuguese 9th grade students (approximately 15 years old). In what follows we present the descriptive geometry constructions of anamorphosis in a way that reflects the learning paths here mentioned.

Dürer Machines Running Back and Forth

Descriptive geometry constructions are adequate for constructing both planar and curvilinear anamorphoses. We will describe a path of growing generality leading swiftly from plane anamorphosis to curvilinear perspectives. This is both a descriptive path for our subject and a learning path a student might follow with exercises of growing difficulty and generality. Using conical anamorphosis as the central concept integrates the process and simplifies it to the point where the whole apparatus of both anamorphosis and spherical perspectives can be learned by an artist in a few weeks, even if starting from no formal knowledge of perspective (Araújo 2017b).

A practical way to teach anamorphic constructions to artists is to start by implementing a simple Dürer machine with a thread fixed on a tripod or other fixed point that will serve as the observer's point *O* (Fig. 22). We make the Dürer machine work forward to the ground plane rather than, as usual, backward to the perspective plane. For each reference point *P* on the object, we extend the thread from *O* through *P* and mark the image point *R* where the thread hits the ground plane. In this way Dürer's perspective machine becomes a machine for making anamorphoses, which can be used to obtain an oblique anamorphosis such as those of Figs. 11 and 23. The same setup can also be used to obtain curvilinear anamorphoses such as that of Fig. 7, onto a cylinder, only requiring more points for drawing each line segment that will now project as a curve.

This mechanical implementation of the Dürer machine is important to establish in the mind of the artist a concrete, physical, and operational notion of anamorphic equivalence. This makes it easier to then abstract it diagramatically, doing away with the thread. A simple diagram with a side view and top view – the most basic of descriptive geometry concepts – allows for construction of the oblique

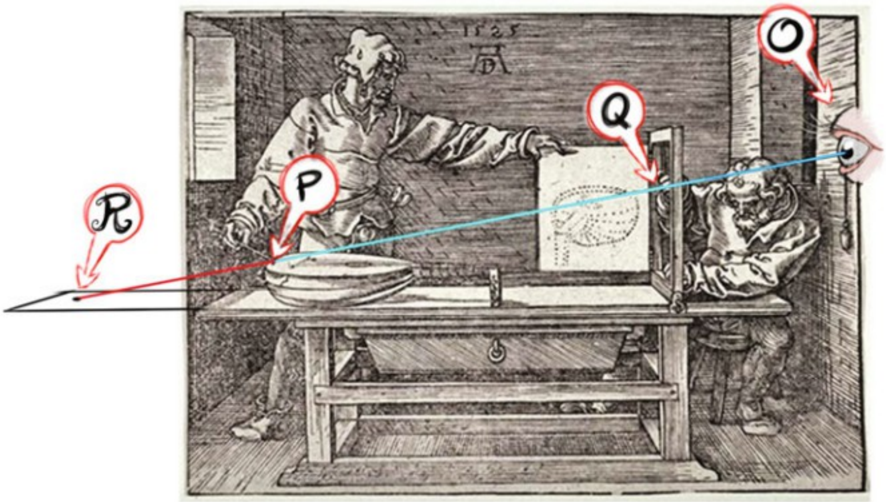

Fig. 22 A Dürer perspective machine can be run back or forth, to make either a "perspective" onto the vertical plane or an "oblique anamorphosis" onto the horizontal one. Points P, R, and Q are all anamorphically equivalent from *O*, so should be indistinguishable to a viewer at *O* (assuming radial occlusion). (Reproduction of a print by Albrech Dürer, defaced by the author)

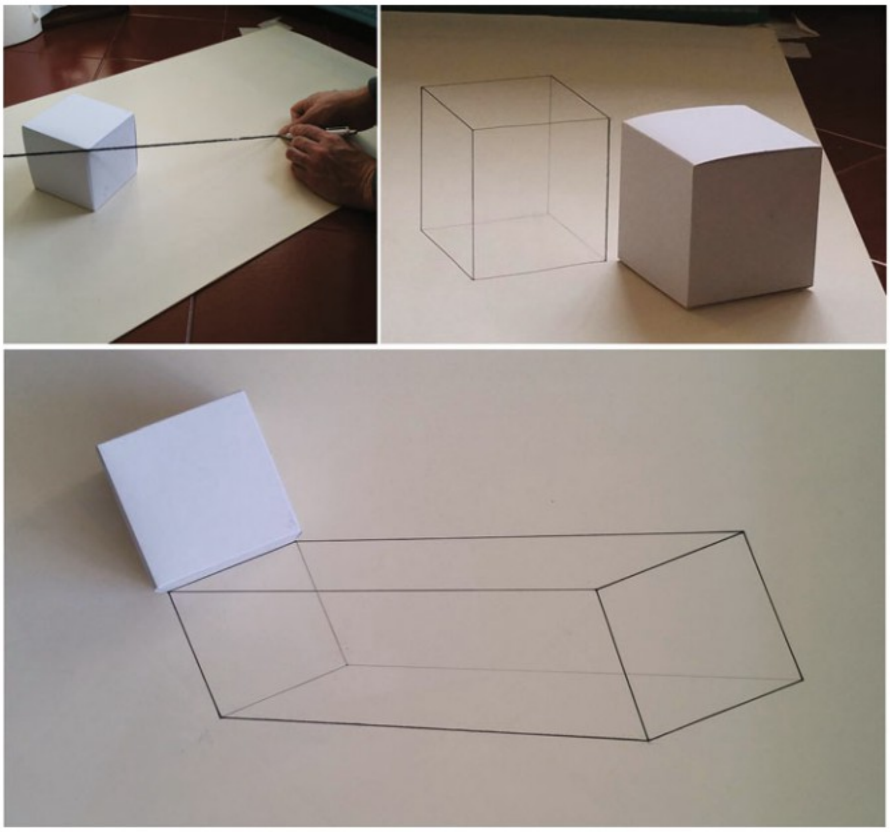

Fig. 23 Making a cube anamorph with a thread fixed to a point

anamorphosis with diagrammatic "thread" (see Fig. 24). The process integrates the anamorph and its construction on the same piece of paper.

Take the example of Fig. 24, a student's exercise. We start by establishing orthographic top and side views (a "plan and elevation") of the object to be projected, using the ground plane as a horizontal folding line. We establish the position of O by plotting its projection O_T on the horizontal plane and fixing its height at O_S in the side view. For the example presented, it is easier to start the object's drawing with the top view of the vertices of the base and then lifting them to the side view, through the folding line, for instance, starting from the top view projection of point P_T in the figure, lifting it through a vertical to get its side view P_S, by setting its height above the folding line. Then to plot the anamorph of the object onto the ground plane, start with the side view, as rays in the side view have *true* intersections with the ground plane. An intersection is said to be *true* if the intersection of the projection is the projection of the intersection. For instance, taking the vertex P, $\overrightarrow{OP}$ projects as $\overrightarrow{OP_S}$, and since the ground plane is

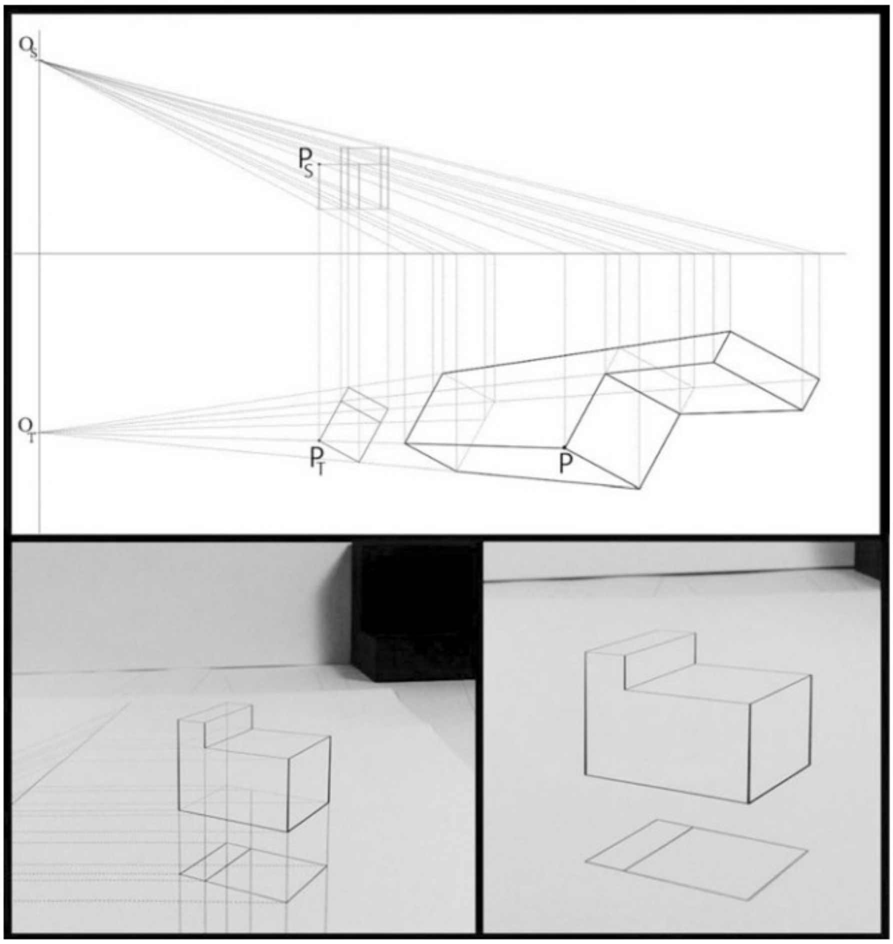

Fig. 24 Top: Abstracting a Dürer machine with plan and elevation views of the threads going through a cube's vertices. The anamorphosis is realized on the top view and should be observed with the eye above O_T at the height of O_S. Bottom: The anamorphic view from O. (Student work by Manuel Flores)

perpendicular to the folding line on the side view, ray $\overrightarrow{OP_S}$ has a *true* intersection with the ground line, that is, the side view of the intersection of $\overrightarrow{OP}$ with the ground plane is the intersection of the projected ray $\overrightarrow{OP_S}$ with the ground line. By contrast, the intersection of that ray with a vertical edge of the cube, for instance, would not necessarily be true. Obtaining the intersection with the ground line, drop it on a vertical to intersect ray $\overrightarrow{OP_T}$. Since $\overrightarrow{OP_T}$ is the projection of the vertical plane through $\overrightarrow{OP}$, the intersection is true and equal to the top view of the anamorphic image of P onto the ground plane.

In this way we obtain the image of all the vertices on the top view of the ground plane, which we can connect with line segments to get the full anamorph of the cube. Then the anamorphic illusion will be obtained if we observe it with our eye over point O_T at the height determined by O_S.

Of course this effect only happens in our perception if the principle of radial occlusion is verified. This will not be verified by the naked eye if the anamorphosis is too small, and the typical classroom anamorphosis will generally be far too small for the principle to be valid. However, the camera – especially the small sensor camera of mobile phones, with their large depth of field and close focusing distance – is just the right tool to simulate the eye of the perfect monocular observer. Most anamorphosis we have displayed here, being drawn in small A4 to A3 sheets of paper, are, in this situation, small constructions that cannot really fool the naked eye, but which display their magic perfectly on camera. Students, especially younger students, seem to find this extra step not a mere concession to convenience, but actually an added charm, a mathematical trick that is inherently *instagrammable*.

We can extend this descriptive geometry construction to many other situations; in fact we could use it as a pretext to teach all the usual operations of descriptive geometry, projecting objects against planes slanted at arbitrary angles, or against cones, cylinders, spheres, or various other classes of surfaces. One can also consider the problem of shadows, both volumetric shadows as we have considered above and also projected shadows (See Fig. 25) that can be constructed in much the same way as we construct oblique anamorphosis, as the mechanism is exactly the same, requiring only that we treat a point light source as an additional conic apex in the diagram of Fig. 24. Anaglyphic anamorphoses (Fig. 12) are also easily constructed in this way.

So many are the achievable constructions and the descriptive geometry techniques put to use that Araújo (2017c) proposes anamorphoses as a path into descriptive geometry for young students, given its motivational possibilities, and the

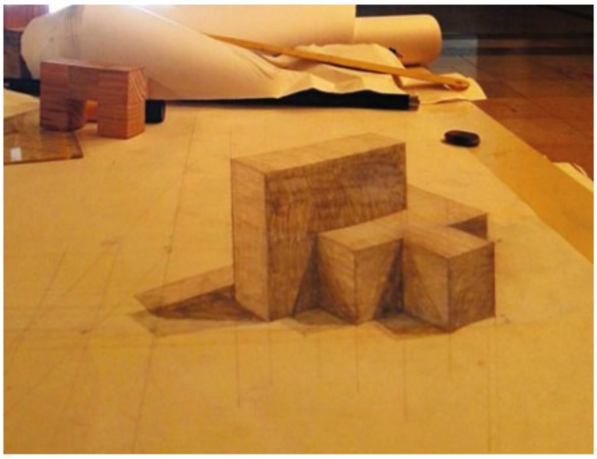

Fig. 25 Anamorphic structure with projected shadows, using orthographic projection. (Student work by Maria Bianchi Aguiar)

fact that, unlike many other constructions in this discipline, anamorphoses provides exercises without a need for an *oracle*, meaning that a student can, for the most part, independently *see* if the exercise was well executed, without a teacher to tell him so. This solves a good deal of the difficulties of descriptive geometry, the main ones being the student's inability to visualize the spatial construction that is being achieved, the other the lack of compelling incentives to achieving the result.

Let us consider, among the many constructions accessible through this scheme, a couple that will be of use in our transition to the study of perspective. We revisit, with the aid of descriptive geometry, the case of the cylinder.

First, revisit Example 6 (Fig. 18), where a cube is projected onto a cylinder with O located outside of the cylinder. In this case, and in contrast to the case of the oblique anamorphosis onto the ground plane, it is the top view that provides for the easier starting point, as rays from O_T have true intersections with the cylinder wall, which projects as a circle. So, as we see in Fig. 18a, a vertex P of the cube, with top view P_T may be projected from O_T in the top view until it hits a point P'_T in the cylinder's circular boundary; raise this point on a vertical to meet the side view of the same ray, $\overrightarrow{O_S P_S}$, finding P'_S. Transport the height of P'_S by sending a horizontal across to the flattened cylinder view, and find there the image of P where this horizontal intersect the vertical that marks the position of the radial plane through O defining angle $\alpha = \angle(P_T C_T F)$ in the top view, where point F is located at 180 degrees to the point where the cylinder is cut.

Consider next the case of Example 7 (Figs. 19 and 26) in which O is in the cylinders axis. The projection is made in the same way, illustrated in the figure with the plot of a vertex P of a cube. Consider, for instance, the horizontal edge of the cube going through point P and ending above point F, in the figure. The segment is contained in a line that projects as the curve $VPV^\star$.

As mentioned in Example 7 this curve is a sinusoid. It is easy to see why using descriptive geometry, in the particular case of horizontals (and the general case is easily obtained through a little trigonometry). One can always rotate the horizontal to assume it projects as a point in side view, as in Fig. 19. Then any point Q along that line, whose orthogonal projection Q_T subtends an angle $\alpha = Q_T O_T F$ in the top view, will project as $\cos(\alpha)$ on the horizontal line $O_T F$ in top view, and this measure will be transferred up to the line $O_L P_L$ on the side view, hence being equal to $m \cos(\alpha)$ with m determined by the slope of $O_L P_L$ and hence constant for all Q on the line. Once the cylinder is flattened, the vertical measurement is preserved, and the horizontal measurements are linear with the angle α, so the curve obtained as Q slides along the line is proportional to $\cos(\alpha)$, hence a sinusoidal curve.

Returning to Fig. 19 and to the line defined by the horizontal edge through P, its vanishing points V and $V^\star$ are obtained by translation to O and intersection with the cylinder. There are several ways to plot the sinusoid itself. In a variation of what was done in Fig. 19, one can simply plot points corresponding to regular angular intervals on the top view circle and project these onto the flat view, then interpolate. More interesting may be to note that the sinusoid is defined by any two of its (non-antipodal) points, for instance, V and P. A good choice is V and the apex of the

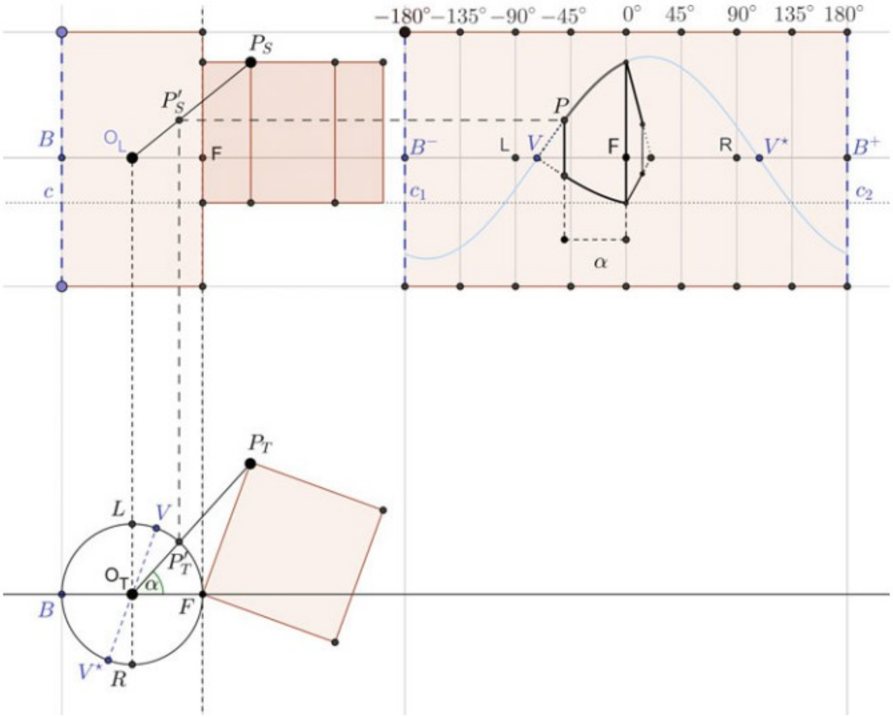

Fig. 26 Cylindrical anamorphosis of a cube onto a cylinder. The image on the flattened cylinder (upper right) is a cylindrical perspective. It transforms non-vertical lines into sinusoidal curves. Notice the blue sinusoidal curve defining the cube's top edge through point *P*

curve, where it reaches it maximum. This apex projects in top view at the point where the top view of the line meets its orthogonal through O_T. Given V and the apex image, it is easy to plot the sinusoid in the flat view.

Perspectives

We have talked from the start of considering perspectives, both linear and curvilinear, as a derived concept from anamorphosis. We are now in a position to develop that idea.

Let us begin once more intuitively and loosely: in our view, a perspective is just a flat representation of a surface anamorphosis, like a map is a flat representation of the Earth's globe. That is all. We want these maps to have certain properties, so we will need some technical details, but this is the main idea to keep in mind.

The notion of *curvilinear perspective* has a complicated past with a history that is often confusingly reported upon. The difficulties come not so much from the misconceptions of the past, but from those remaining in the present, under whose burden the historian must work. It is hard to argue whether Leonardo or Jean Fouquet are precursors of curvilinear perspective (Andersen 2007) without a proper definition of what those perspectives are. Often both the original sources and the historian fail to distinguish the perspective from the corresponding anamorphosis. This distinction, that so clarifies the subject, is our focus here. The matter of a perspective being "curvilinear" is comparatively minor. We will define as curvilinear any perspective that projects spatial straight lines into plane curves. This of course includes the cylindrical perspective and the usual spherical ones. Note that "curves" cannot mean only smooth curves; otherwise this has the curious result that at least one perspective (the cubical case) manages to be spherical without being curvilinear (in this perspective spatial lines project as sets of line segments). But we will not concern ourselves much with this. Curves will arise from Definition 11 as side effects of the need to project compact anamorphoses onto compact plane sets while preserving their main topological characteristics, such as vanishing points. The perspectives we will define are not so much characterized by being *curvilinear* as they are by being *central* (a side effect of being derived from conical anamorphosis) and *compact*.

We will motivate our definition of perspective by considering the simplest of curvilinear perspectives: cylindrical perspective. It has a long history. Whether you consider it defined by Herdman's 1853 treatise (Herdman 1853) or implicitly defined by Baldassare Lanci's 1557 (Kemp 1990, p. 175) perspective machine, curvilinear perspective was certainly textbook matter in the 1900 Ware (1900), and in the very early twentieth century was used for military balloon airmen spotting for artillery (Dept. of Military Aeronautics 1918). Its aspect is familiar to the modern reader, as it corresponds to that of the common panoramic photography.

We have already silently met cylindrical perspective, in Figs. 19 and 26. We have seen that in order to construct the anamorphosis of an object (in this case a cube) onto the cylinder, with O at the axis, it is convenient to first draw the scene in plan and elevation, flatten the cylinder into a rectangle by cutting and unrolling it on the plane, and project the object onto this rectangle (upper right corner of Fig. 26). Then, once the projection is drawn, the rectangle can be rolled up again and glued at the vertical edges to form the compact cylinder in 3D space, and the resulting picture on its surface will thus become the cylindrical anamorphosis. That auxiliary plane drawing on the rectangle, from which the anamorphosis is rolled up, is what we call the *cylindrical perspective*. Note that this is purely an affair of convenience. By definition, the anamorphosis is obtained on the spatial cylinder, through conical projection. But in practice, we find it convenient to draw on planes. This is analogous to cartography, where the exact shape and metric of features on a globe is sacrificed to the convenience of a plane chart. So, to us, perspective is merely this: a plane chart – a flat representation – of the spatial anamorphosis. The flat representation is not only easier to draw; it allows for they eye to capture the whole picture at a glance, at the price of introducing optical deformations. This is often the reason to use it, as in the case of Jean Fouquet's *Arrival of the Emperor Charles IV at the Basilica in Saint Denis* , where a deformation quite similar to cylindrical perspective allows us to see at a glance further down the street than would be possible with linear perspective.

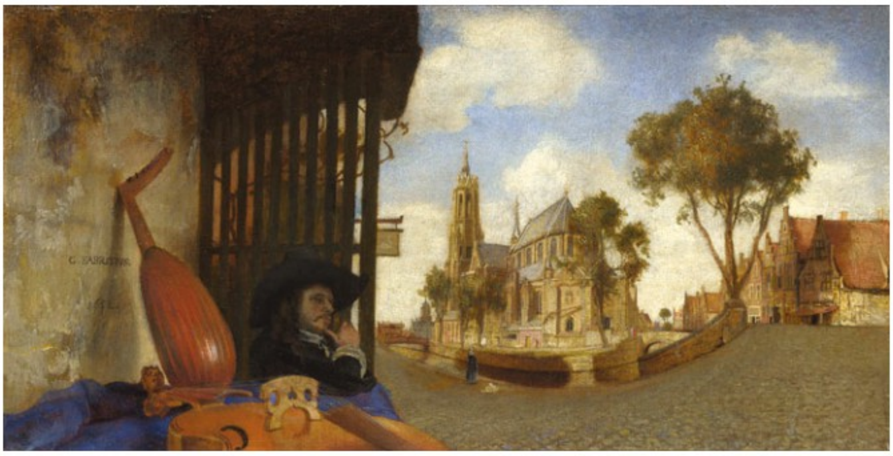

Fig. 27 A View of Delft, with a Musical Instrument Seller's Stall. Carel Fabritius, 1652. Currently at the National Gallery in London. Apparently a cylindrical perspective, although nothing is known regarding the method of its construction

Analogously to the case of cartography, we must pay for this convenience with deformations that result in perception inaccuracies. This is true even in the case of the cylinder that unrolls isometrically onto the plane. See, for instance, Fabritius' *View of Delft* (1652) (Fig. 27).The flat cylindrical perspective drawing, although it contains all the information of the cylindrical anamorphosis, and is isometric to it, is no longer an optical illusion; there is no point O from which the points can recreate the effect of the original object – the cones no longer align; hence it is no longer an anamorphosis. So the straight lines in the picture all get deformed. The line image of Fig. 19 looks nothing like a line (as we have seen, it is a sinusoidal curve), and the cube on Fig. 26 looks nothing like a cube. The non-vertical lines are bent; that in itself is not the problem, as they are also bent in the anamorphosis, and still they will look straight when seen from O. The problem is that in the perspective they are not just metrically but fundamentally *optically* deformed; there is no point O' from which they can all look straight. The perspective drawing is a *representation*, a storage of visual information – but it is not mimetic.

Of course it can still be *read* up to a point, even without training. But I challenge anyone to distinguish with perfect accuracy – a task easy in classical perspective – which lines in a cylindrical perspective are really bent and which are only bent through projection. The task can hardly be done by the untrained eye unless with an extreme reliance on context. The cylindrical perspective still evokes the scene, but no longer mimics it.

Let us be more precise. We view the perspective as a map from 3D space to a compact region of the plane, defined by a two-step process: the first step being the anamorphosis, from the 3D space to the surface S, the second step being the flattening of S. The anamorphosis is completely defined by the choice of S and O;

the flattening is a much more arbitrary map, just like in cartography. Because we will want to preserve the visual *information* , if not the visual aspect, we have some requirements. First, we would like it to be a bijection, so that no information is lost. Then, we'd like it to be a homeomorphism (continuous with continuous inverse) so as to preserve the topology and hence the vanishing sets and their meeting properties. Finally, we would like it to be as smooth as possible. We have a second set of more vague, yet crucial requirements: we would like it to preserve as much of the mimesis as possible; though it will not be an anamorphosis, it should be in some way recognizable as a picture; it should summon the visual presence – or at least the intellectual recognition – of what it represents. Finally, it should be solvable by elementary means. By this we mean that you should be able to draw it by hand, by ruler and compass, or some other simple means with adequate precision and in reasonable time, i.e., we would like it to be a perspective for humans and not just for computers.

We find that even the first three technical requirements are asking too much. We need to interpret them with some care. Look at the case of the cylinder in Fig. 26 and it will become apparent why. The cylinder is a developable surface; it can be unrolled onto the plane, preserving distances on the surface. Intuitively we would like to say that the flattening is the unrolling map that sends the cylinder to the rectangle. But this is not a bijection on the compact rectangle. Let c be the segment through B where the cylinder is cut, and let c_1 and c_2 be the corresponding vertical edges of the rectangle. Each point of c is sent to a point on c_1 and another equivalent one on c_2, so this "projection" is not even a map at all.

We solve this by noting first that the flattening π is well defined *almost everywhere*, meaning, on an open dense subset of the cylinder S. Namely, it is well defined on S minus the union of its boundary with the vertical segment c. This projects to an open rectangle that excludes c_1, c_2 and is a homeomorphism onto that open rectangle. Further, the inverse π^{-1} can be extended uniquely to a continuous map $\tilde{\pi}$ from the full compact rectangle to the complete cylinder. We just map both c_1 and c_2 to c, since in this direction the map has no problem of uniqueness. Apart from technicalities, we have the map we wanted. We define the cylindrical perspective to be the composition of π with the conical projection, and the perspective image/vanishing set to be the image of the anamorphosis/vanishing set, where we identify as the same point any points with equal images through $\tilde{\pi}$.

We will use this example as a template for the general case and say that π is a flattening if it can be extended to a continuous map in the way we just exemplified. We formalize this by the following definition, adapted from Araújo (2015):

Definition 11. Let S be an O-starred surface. Let U be an open dense subset of S. We say that $\pi : U \to \mathbb{R}^2$ is a *flattening* of S if π is an homeomorphism onto $\pi(U)$, and there is a continuous map $\tilde{\pi} : cl(\pi(U)) \to S$ such that $\tilde{\pi}|_{\pi(U)} = \pi^{-1}$. We say that $p = \pi \circ \varphi_{O,S}$ is the *perspective* associated to the flattening π, where $\varphi_{O,S}$ is the conical projection. Let $\tilde{p} = \varphi_{O,S}^{-1} \circ \tilde{\pi}$. Given an object Σ, we say that $\tilde{p}^{-1}(\Sigma)$ is the strict perspective image of Σ, that $\tilde{p}^{-1}(\mathscr{V}_O(\Sigma))$ is the vanishing set of Σ, and that the perspective image of Σ is the union of its strict perspective image with its vanishing set.

The following set is often useful in discussing perspectives. It is the minimal closed set we must take out of S for π to be well defined.

Definition 12. Given a flattening π of S, We call blowup to the subset of points of $cl(\pi(S))$ where $\tilde{\pi}$ is not injective.

The geometer may notice that we are here abusing the term "blowup," common in other areas of geometry, where it means, roughly, the replacement of a point by the projective line. Here it is in fact just the set where gluing (identification) of edges takes place. We use the term blowup by analogy to the important case of fisheye perspective where a point will identify to a circle or more precisely to the set of rays or directions defined by that circle. This is analogous to the usual sense of blowup of a point, which roughly is the replacement of the point with a projective line.

Example 9. Take again the case of the cylinder in Fig. 26. Let $O = (0, 0, 0)$; let the radius of the cylinder be $r = 1$ and its top boundary lie at $z = \pm h$. Suppose we cut the cylinder at the vertical segment c that passes through $B = (-1, 0, 0)$. Then, in cylindrical coordinates, we define $\tilde{\pi}(\theta, z) = (\cos(\theta), \sin(\theta), z)$, so the blowup set is $\mathscr{B} = c$, and $\tilde{\pi}$ implicitly defines π in the $S \setminus c$. Having defined π, the cylindrical perspective is defined by the entailment $p = \pi \circ \varphi_{O,S}$.

As seen above, spatial verticals are anamorphic to verticals on the cylinder and non-verticals to ellipses. Their perspective images by π are the curves we obtained above through the cylinder's unfolding: arcs of sinusoidal curves. For instance, in Fig. 26, the line that goes through P and has vanishing points V and $V^{\star}$ defines a plane through O that intersects the cylinder on an ellipse and that ellipse maps in perspective to the full sinusoid in blue. The line itself is the arc $VPV^{\star}$, one half of the sinusoid. These can be seen as the images through the cylindrical flattening of the great circle (resp. meridian) of the canonical anamorph of the plane (resp. line) in question.

Notice that although π, the flattening, is the more intuitive concept, $\tilde{\pi}$, the closure of its inverse, is the more natural map, defined everywhere and usually easier to work with analytically, as in the Example 9. This is very common in geometry.

Definition 11 is just a technical way of making sure that the perspective image of an object and its vanishing points are just the sets in the perspective image that correspond to the ones previously defined in the anamorphosis. As we stated before, everything important is already defined at the level of anamorphosis, perspective being relegated to a final step of convenient representation. This of course does not in deny the importance, both technical and aesthetic, of perspective, nor the intricacies of its construction, but it clarifies it meaning. In particular it does away with long-standing conceptual mistakes expressed in questions such as "do we see in classical/spherical perspective" or "does linear perspective cause deformations." It is not so much that we answer these questions; instead we show they are badly posed. There is no meaning to the question because we simply do not "see in perspective," whatever that perspective may be. Perspective is a map of the information in an anamorphosis and the anamorphosis is, when radial occlusion is verified, a mimetic object (a *trompe l'oeil*).

It is true that we can speak, more or less vaguely, of some perspectives preserving the aspect of the scene better than others, that is, of being "closer" to the mimesis than others. How close depends on the measure we use. (Araújo 2016) speaks of the *reading mode* of a perspective. For instance, a cylindrical perspective has a natural reading mode in the sense that when the cylinder of radius r is unrolled, we can think of the eye point O being transformed into a horizontal line floating parallel to the central axis of the picture, at a constant distance r. The natural reading mode is for the eye to travel horizontally, scanning the picture as it moves, one vertical thin strip at a time. At each point of the motion the vertical line in front of the viewer – and only that one – will be in anamorphosis. By contrast, in a fisheye perspective, only the central point in anamorphosis and a small neighborhood around it will show minimal differences if overlapped with the observation of the actual object. Notice that we take care to say "with the observation of the actual object" not with "its image," as otherwise we would be in the difficult situation of trying to define what the "image" means. Mimesis is checked not by comparing "images" but by looking at pairs of objects (that may be volumes or drawings on surfaces) and being unable to tell the difference between them, at least in some well-specified experimental sense (e.g., do the vertices of a wireframe anamorph of a cube seem to occupy the same spot as the vertices of the real cube it seeks to mimic).

In this sense we see that indeed classical perspective occupies a special place in the bestiary of perspectives, as it is the only one that is still an anamorphosis. This happens because the anamorphosis is already on a plane, so the flattening map is just the identity map. The reading mode of a perspective is the same as that of the anamorphosis: it consists of putting your eye in the point O and rotating it freely toward any spot of the plane of the picture. The so-called perspective deformations that happen for large angles of vision happen simply because the pictures in question are made in such a scale that they make the viewer unable to occupy point O without breaking the requirements of the principle of radial occlusion. Usually this happens because the perspective is draw at such a scale that the eye must be too close to the picture plane. And of course, the illusion will not work of the viewer leaves point O to look closer at some detail, or if he looks at the picture from some arbitrary point in a crowded room. But anamorphosis only works from point O. To blame it for not working otherwise is to blame a fork for not being a spoon.

Spherical Perspectives

We have seen that among anamorphoses, the spherical one holds a special place, due to its natural identification with the space of rays $\mathscr{R}_O$, or with the concept of *visual sphere*. For the same reason, the spherical case also holds special importance among perspectives.

Of course, while spherical anamorphosis is unique, spherical perspectives are innumerable. Every chart of the sphere defines a flattening and hence a spherical perspective according to Definition 11. In fact every central perspective (even classical perspective, if restricted to a compact subset of the plane) can be seen formally as a *spherical* perspective; hence the term is only as interesting as the qualifications we add to it: for instance, we can distinguish *total* spherical perspectives, i.e., one those in which all points of the sphere have a perspective image. Each spherical perspective, although it holds the same visual information and would generate the same anamorphosis, holds its own visual characteristics and artistic possibilities as a drawing type. These visual characteristics hold all sort of representational and expressive possibilities that have been explored in traditional media (Araújo 2019b; Barre and Flocon 1968; Casas 1983; Michel 2013; Moose 1986) as well as in purely digital visualizations (Correia et al 2013) and are more and more being investigated as *hybrid immersive media* (Araújo et al 2019a; Olivero and Sucurado 2019), a mode of artistic expression that allows handmade drawings to be visualized through immersive, digital means.

It is interesting to ask how we can solve a spherical perspective, where by *solving* we mean giving a systematic method to find and draw all vanishing sets of lines and planes through simple geometrical constructions (such as ruler and compass constructions).

Notice that since by Proposition 10 the perspective images of lines are subsets of the vanishing sets of planes, and since plane images are delimited by their vanishing sets, obtaining all vanishing sets also obtains all line and plane images; hence we can say that solving vanishing sets is the true subject of perspective.

We have seen that the vanishing sets of spherical perspective are all determined by the geodesics of the sphere. This suggests, if not an algorithm, at least a strategy for solving a spherical perspective: solving a given spherical perspective should be always attempted by focusing not on lines themselves, but on full geodesics, on classifying the geodesics according to the properties of their projection by the flattening and on finding an efficient method to plot the geodesics of each class.

In this, we will find there are features common to all spherical perspectives, and features peculiar to each. What common is the first step of the perspective projection, since it is just anamorphosis onto the sphere; what is particular is that each spherical perspective will have its own metric characteristics, which may vary widely. It will also have its own topological properties, but these are by design more limited in variation than the metric ones. Definition 11 ensures that π is continuous outside of the blowup set, so the topological variations will depend entirely on this set, which means that the blowup is a fundamental feature of perspective.

This view of perspective, first stated and applied to the case of the azimuthal equidistant perspective (Araújo 2018c), was later used to solve the equirectangular perspective case (Araújo 2018b) and then the case of cubical perspective (Araújo et al 2019b). The splitting of the anamorphosis from the flattening, and the focus on full geodesics instead of individual lines, proved an elegant solution for these perspectives. Previously, the only spherical perspective that had been solved systematically was the azimuthal equidistant in the hemispherical case, meaning that in fact only half of it had been solved. This is the perspective of Barre and Flocon (Barre and Flocon 1968), solved in the 1960s, which renders into images similar to those of a "fisheye" camera. The full azimuthal equidistant case had only been solved in either a qualitative way (Casas 1983) or through informal grid methods (Michel 2013) which lack the generality of Barre and Flocon's solution. So there stood this curious situation in which the first spherical perspective was not (fully) spherical, and the subsequent ones were not (formal) perspectives.

The approach here described and implemented in (Araújo 2018c) was able not only to extend Barre and Flocon's perspective to the full sphere but also simplified their solution in the frontal hemisphere, reducing it to two distinct classes, the focus on geodesics instead of lines providing a clearer, more unified view of the problem, for even when plotting only the frontal part of a line, it is helpful to be able to call on the antipode of every point (and the pair of any vanishing point) for auxiliary constructions.

As for the equirectangular case, it had previously been treated only computationally or through fixed grid constructions, the equirectangular grid being calculated at fixed intervals and used as a guide for drawing. The more general approach in Araújo (2018b) leads naturally to an exploitation of the translational symmetries of the perspective , which can be used with a system of dynamical moving grids that allow for the plotting of all lines in good approximation (Araújo 2018a, 2019a).

The cubical perspective case is an interesting one, since seeing it as a spherical perspective (Araújo et al 2019b) greatly simplifies previous approaches (Olivero et al 2019; Rossi et al 2018) that saw it as a conjunction of six related classical perspectives. It might not at all seem obvious that seeing a set of classical perspectives as a spherical one would be a simplification.

In all these three cases, the approach passes through understanding the flattening and its blowup set, classifying the geodesics of the sphere according to how they are projected in the plane, and finding a method to render each class through simple constructions. For this, the understanding of antipodes and how to plot them is essential , as the duality of vanishing point pairs usually results in symmetries that can be exploited. As these are curvilinear perspectives, approximations will be required to draw line projections, and in these the number of operations should be kept to a minimum. This will require a careful exploitation of the symmetries arising from the flattening. A general philosophical principle is that not only the perspective drawings themselves but also the auxiliary drawings should happen in a compact subset of the plane; in short, everything should be bounded.

We will not explore in any detail here the actual constructions of spherical perspectives, leaving that for another chapter in the present volume (see ▶ Chap. 19, "Spherical Perspective"). We will merely relate the fundamental properties of two of these cases, to see how they fit on our general scheme described above.

In the case of the azimuthal equidistant perspective (Fig. 28 (top)), the flattening works by choosing a reference point B (for "Back") and loosening the meridians there, straightening them without stretching to obtain a disc centered on the point

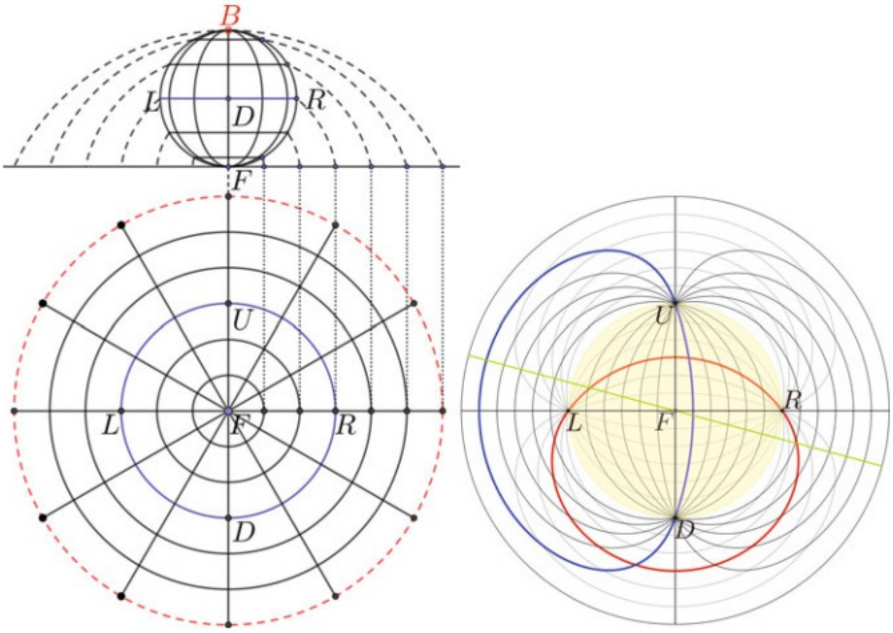

Fig. 28 Azimuthal equidistant flattening of the sphere. Left: Blowing up point B maps the sphere to a disc centered on the image of F, the antipode of B. Geodesics through F map to diameters of the disc. Right: The green line is a geodesic through F. Geodesics not through F (blue and red lines) map to closed convex curves, well approximated by arcs of circles in the yellow inner disc that represents the space in front of the observer

F (for “’Forward”) at the antipode of B. If we reference F and B as the “poles” of the sphere, then FB meridians flatten as rays of the disc and “parallels” flatten as circles, with the image of the equator separating an inner circle (the frontal view, rendered yellow on Fig. 28 (right)) from an outer ring, where the back view is rendered (see the examples in Fig. 29). The flattening is one-to-one at all points except at B. Taking the closure we get a compact disc, with the outer boundary circle (the blowup set) mapping entirely onto B. The flattening preserves the metric along each individual FB meridian, though of course not globally. This preservation of the metric is what allows measurements to be made, so drawing this perspective requires careful attention to these special meridians in the construction process. The natural measurements for this perspective are angle pairs, each point on the sphere being measured by one angle choosing the FB meridian where it lies and the other the length traversed from F along that chosen meridian. As for geodesics, these can be classified in two classes: geodesics through F render as diameters and all others as closed curves (see Fig. 28 (right)). For the latter, half the geodesics will be rendered in the anterior disc (yellow in Fig. 28 (right)) as an arc of circle c in good approximation, as shown by Barre and Flocon (1968); the other half can be rendered, as the locus of a point P that moves so as to keep at a distance of half

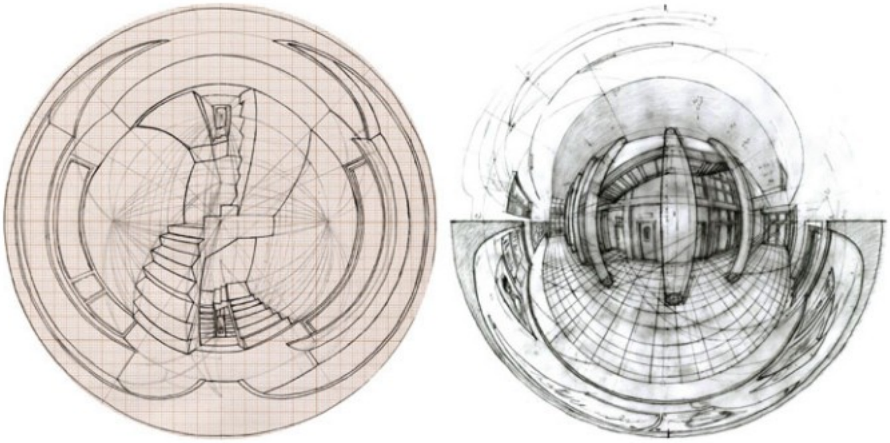

Fig. 29 Two handmade drawings of urban scenes, obtained through the process described in Araújo (2018c) (left) and Araújo (2019b) (right). All lines are contained in geodesic images of the azimuthal equidistant flattening. (Drawings by the author)

a diameter from the point of c that is across from it along the line that joins it to F. This locus can be drawn using a mechanical process involving a ruler sliding on a fixed point, as explained (Araújo 2018c) and in the ▶ Chap. 19, "Spherical Perspective" in the present volume. This can be simplified by a moving grid system using the rotational symmetries of the perspective around F (Araújo 2019b).

In the case of equirectangular perspective, we blowup the sphere at two points U and D (representing for instance the directions "up" and "down") and straighten the meridians, keeping them fixed at the circle halfway between U and D (the "equator"). In this way we obtain a cylinder, which we can now cut and unroll onto a rectangle. Unlike the case of cylindrical perspective, this rectangle will contain a record of the whole view around O. There are two classes of geodesics: geodesics through U that render onto straight vertical lines and all the others. The first (resp. second) type includes the image of vertical lines (resp. horizontals) but is not restricted to these. If we consider the set of all geodesics through two antipodal points on the equator (say, points L and R, corresponding to the observer's Left and Right), we can generate all other geodesics from translations of elements of this set. This allows for a construction of complex scenes such as that of Fig. 30 without explicit use of ruler and compass and without the limitations of a fixed grid (see (Araújo 2018a, 2019a)). As pointed shown in Araújo (2018b). These geodesics are very similar to sinusoidal curves when their apex (point of highest angular elevation) is low; hence this perspective looks like cylindrical perspective for low elevations (see Fig. 30). As the elevation rises, the geodesic images converge to a square wave, but even for high elevation they can still be rendered in good approximation with descriptive geometry constructions. In Fig. 30 we can see a drawing done on location of an equirectangular spherical perspective using the sliding grid method of (Araújo 2018a). On top of it we see a spherical anamorph of the same scene; the drawing was

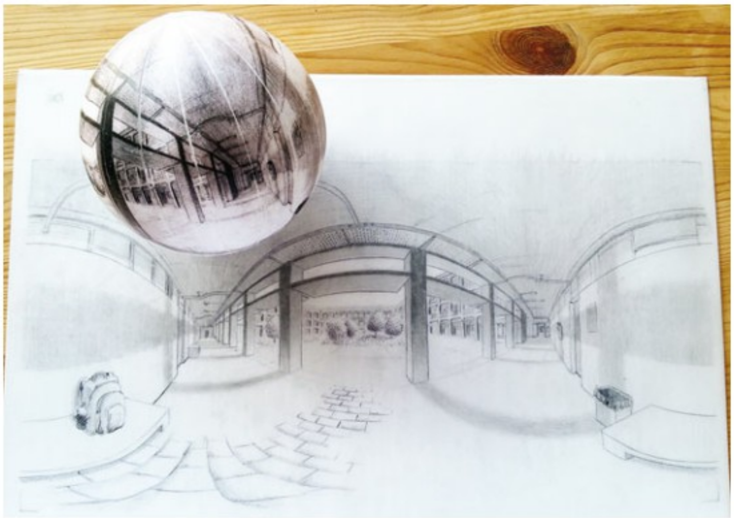

Fig. 30 A spherical anamorphosis and its equirectangular perspective. The spherical perspective drawing was done on location by the author. It was scanned, transformed on a computer, and then printed and glued on a ball to obtain the corresponding spherical anamorphosis. The author would like to thank Ph.D. student Lucas F. Olivero for making the 3D model on the top left of the picture

transformed by a computer into a sinusoidal projection, and then printed and glued onto a ball. This again shows that often the point of a perspective is to be a blueprint for the anamorphosis, instead of the other way around. We defined the perspective as the flattening of the anamorphosis because the anamorphosis is the uniquely defined object; but just as often the anamorphosis will be in practice treated as the folding of the flat perspective.

The Problem with Perspective

The view of anamorphosis we studied here was motivated mainly by the wish to solve conceptual problems with perspective. We will start by discussing what those problems are.

One might be surprised to hear are any conceptual problems with perspective at all. Linear perspective is universally used today to represent a 3D scene on a plane. Most working artists and architects take it for granted as an objective picture of an environment, identifying it with photographic representation. Mathematicians take it for a well-understood concept, made rigorous as an application of projective geometry (see ▶ Chap. 4, “Looking Through the Glass”).

Yet the nature and meaning of perspective has been a point of contention since its inception. We can see this as a contention between the views of Euclid and those of Leonardo.

Euclid and Psychophysics

We could argue that our view of anamorphosis in the present work is a recovery of concepts first formulated in Euclid's Optics with the addition of vanishing points. Or, put in another way, that the fourth axiom of Euclid's optics is the earliest known precedent for the concept of anamorphosis.

Axiom 1. *(4th axiom of Euclid's Optics) and that those things seen within a larger angle appear larger, and those seen within a smaller angle appear smaller, and those seen within equal angles appear to be of the same size (Burton 1945)*

Euclid here is postulating that for visual perception, the visual angle subtended by an object and its perceived size are the same thing. This is clearly true for a simple object in isolation of other stimuli. That is, if I take a black ball on a white background and I make it subtend progressively larger angles an observer will certainly report a larger apparent size. But it is too daring a proposition when related to complex visual environments. For instance, Ebbinghaus illusion (Ebbinghaus 1902) shows that the perceived size of a colored circle can depend strongly on the disposition of other colored circle's in its vicinity. What exactly is happening is still a matter for contention, with several theories extant (Roberts et al 2005). This is not uncommon for visual phenomena whenever global variables – rather than very well-controlled local experiments – are taken into account. Analogous problems occur in color theory, where the local phenomena are well understood and mathematically elegant, while the global theory is confusing and qualitative. So we must not take the term "apparent size" at face value. Astronomers use the term in the Euclidean sense to mean simply the size of the solid angle subtended, while in studies of perception, apparent size means the subjective impression of size as reported by human subjects.

In the latter sense, Euclid asserts too much, ignoring contextual matters. Still, a careful localized interpretation of the fourth axiom does hold water as empirical fact. Like we said before, it defines an implicit empirical scope wherein it is true, and therefore it is as interesting a concept as that scope happens to be relevant.

Still, in our own work here, we took a different, although related, path. We are not concerned with size but with form; not with the size of the solid angle subtended by the cone of an object but with the cone itself. Our principle of radial occlusion neither implies nor is implied by Euclid's 4th axiom of optics. But the emphasis on the visual angle, without reference to any projection surface, the isotropy of view, and the notion of visual cone as a function of the object rather than on some central axis is already present in Euclid, and a reading of his theorems shows that he has in mind something similar to us. The difference is his avoidance of infinity. Therefore we might say that the work here developed clarifies Euclid and adds to it the notion of points at infinity – of compactness, if you will. It is impressive to read Euclid in retrospect and notice the generality of his view at such an early time and how the late arrival of linear perspective seems by comparison a step back in generality, being a theory of representation of too particular a kind.

Leonardo's Axiom and Paradox

Leonardo considered there were many conceptual problems with perspective as formulated by Alberti. He distinguished between "natural" and "accidental" perspective. This is related (but not identical) with the more usual distinction between *natural* perspective (meaning optics, as a study of both light and visual perception) and *artificial* perspective (meaning linear perspective). Leonardo objected to the requirement in linear perspective of eye placement at single point of observation and thought that a spherical canvas would make for a more natural perspective; and he seems to have thought that in the absence of this a very distant canvas might do. In this he seemed to be concerned with the inclination of the surface normal with regard to the angle of view (in Kemp's (Kemp 1990) view, the foreshortening of the canvas just as much as the foreshortening of the picture). "Foreshortening of the canvas" is of course no more than a reference to the effects of oblique anamorphosis.

Leonardo's greatest difficulty, or at least the one that survived to our time, was based on the following axiom, which he considered a requirement of a natural perspective:

Axiom 2. *Leonardo's axiom: Among objects of equal size that which is most remote from the eye will look the smallest.*

This in itself is not contentious. But he then confusingly identified this notion with the completely different and in fact incompatible one that they should therefore be represented by metrically smaller projections. This in contrast with Euclid's fourth axiom of optics.

He gave the example still called *Leonardo's paradox* (Fig. 31): the conical projection of a sphere that moves along the plane projection actually grows as it moves away from the observer (see Araújo 2016; **?**). This "paradox" really isn't one. True, the perspective picture grows, but so does the distance of this projection to the observer, in such a way that the apparent size of the projection diminishes. By Euclid's fourth axiom, both the distal sphere and its projection will look smaller than the proximal one. Leonardo's real objection may be the disparity between the properties of the metric representation and the angular/visual properties. He would like a perspective that carried the proper visual properties of mimesis without the requirement of standing at a single point O. And he would like that the metric properties, along with the angular ones, should satisfy his axiom. Hence his preference for a spherical canvas. In this, there is a conflating of the two concepts that we separated in the present work: perspective and anamorphosis. This conflation would endure and persist as a philosophical difficulty until the twentieth century. It is in fact still a difficulty today in the way these affairs are discussed. This leads to asking the wrong questions such as whether a "perspective" has "deformations" or whether it is "natural" or whether it corresponds to the way we see. These questions vanish if we only separate the two concepts of mimesis and representation, anamorphosis, and its flattening into perspective. So that the point may be made that the paradox is no such thing – it does not point out to any failure of linear perspective in achieving what it sets out to do, but is merely a complaint

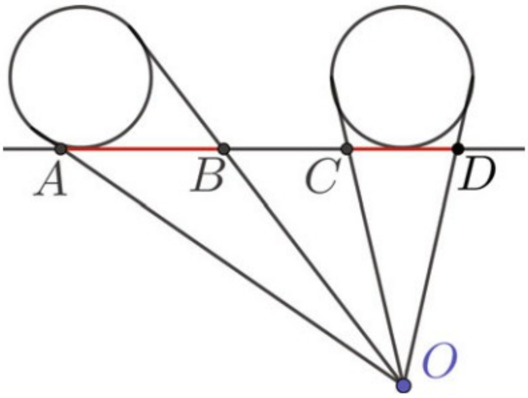

Fig. 31 Leonardo's paradox: $|AB| > |CD|$ even though the corresponding sphere is further away from O

that linear perspective does not do as much as Leonardo wishes from a perspective. But it is hard to see this as a flaw of perspective as much as a failure of Leonardo to settle on a consistent set of requirements. One cannot ask at the same time to preserve both angular and linear measurements. That linear perspective (as plane anamorphosis) does not preserve the latter is not a bug but a feature mathematically required to preserve the former.

This enduring failure to distinguish the mimetic role of anamorphosis from the representational role of perspective may be due to the fact that linear perspective and plane anamorphosis are the same physical object (the "flattening" being trivial); this identification between the anamorphosis and the corresponding perspective only happens in the case of linear perspective, and this coincidence probably made it all the harder to understand the distinction between the two concepts. It is rather unfortunate that the most "obvious" of perspectives is in fact a quite peculiar one. The matter was still contentious in dealing with cylindrical perspective, much later, and even when the practical difficulties of the drawing itself were well handled, mechanically and theoretically (Kemp 1990, p. 175).

Effects on the Development of Spherical Perspective

In the twentieth century, the discussion started by Leonardo took a particularly confusing turn and produced reams of papers in a polemic between the *realist* and *conventionalist* schools of thought. A simplified view of the matter is that the realist position, represented by Ernst Gombrich, attributes an objective quality to the ability of perspective pictures to mimic reality. This ability comes from the fact that the picture realized by the rules of perspective furnishes a properly located observer with the same bundle of light rays that he would receive from the original object. The conventionalist position, as championed by Goodman, argues that instead we see classical perspective pictures as realistic simply because we have been acculturated to them. Like in most such debates, things are not so clear as in this rapid caricature. First of all Gombrich's "realist" position is hardly clear-cut. In *Art and Illusion* (Gombrich 1960) he considered that there is no reality without interpretation and that seeing is inseparable from conceptualization: there is no *innocent eye*. Gilman (1992) argues that there is no *naturalistic* position on either side of the debate and sees it rather as a contention between conventionalist factions, Gombrich being merely a conventionalist with particular scruples where perspective is concerned, which Goodman saw as puzzling contradiction within the context of Gombrich's general position. Gombrich's objection was not so much against the contention that perception required interpretation but that this interpretation was just a cultural construct. According to Mitrović (Mitrović 2013b) it was this strong thesis of cultural relativism that the totality of perceived reality is culturally constructed, as well as associated collectivist notions that repulsed Gombrich. The positions are complex, and Vergsten (Verstegen 2011) argues that both Gombrich and Goodman are "guilty of letting extraneous issues color their discussion of the (non)conventionality of perspective." Be all this as it may, the opposing positions have been argued confusingly and in so many skirmishes that scholar's careers can be erected on the mere accountancy of blows. As in most such debates, the boxers have reached belatedly and by exhaustion the obvious conclusion that, while fresh, they refused to reach by good sense: that neither extreme position is tenable (Frigg and Hunter 2010, page xx) and that the interesting question is how exactly purely optical mimesis and convention interact.

Fortunately, however, we are not here concerned with fleshing out such intricate business. Goodman argued famously that representation cannot be mere resemblance, since resemblance is symmetric (A resembles B implies B resembles A) while representation is not (a picture may represent a horse but a horse doesn't represent its picture).

We argue that in our present formulation we have cut this gordian knot by separating the matter of *resemblance* (anamorphic equivalence, which is indeed symmetric) from that of *representation* (perspective, which is not; A being a perspective of B doesn't make B a perspective of A).

So we are not interested in dealing with this debate in the terms set by Goodman. We have answered it: linear perspective *is not* conventional – it is the only perspective that is also an anamorphosis, and anamorphosis is an optical principle that stands on top of an objective empirical fact: the principle of linear occlusion.

We still care about this debate, however, because it influenced contemporary understanding of the purely mimetic possibilities of perspective and anamorphosis. This is especially unfortunate as Goodman was at its weakest when intruding into the field of mimesis proper and has since been extensively refuted in detail – see, for instance, Mitrović (2013a). Consider the infamous phrase from *Languages of Art* (Goodman 1968). Goodman concedes for the moment the mimetic properties of perspective, but decried what he calls the "remarkable" requirements for such an illusion to occur:

> The picture must be viewed through a peephole, face on, from a distance, with one eye closed and the other motionless. The object also must be observed through a peephole, from a given (but not usually the same) angle and distance, and with a single unmoving eye. Otherwise, the light rays will not match (Goodman 1968, page 12).

The conditions of observation, as described, seem indeed remarkable, even reminiscent of an infamous scene from Kubrik's *A Clockwork Orange* when Alex is bound to a chair with eyes clamped wide open, his gaze forcedly fixed forward. But more remarkable is that the argument easily crumbles not only under the scrutiny of the empiricist but merely under that of the tourist. For all it takes is a casual stroll into the church of Saint Ignazio de Loyola in Rome and a look at its illusionary ceiling, where Andrea Pozzo indeed used "perspective" (anamorphosis in our view) as an alternative to architecture proper (Fig. 1). The Jesuits wanted a grander building than it could afford, and Pozzo constructed it in virtuality, going through all the steps of an architectural project, with plan and elevation constructions, perspective renderings, and finally anamorphic constructions . Here, the tourist looks to painted columns that extend the real ones so expertly that the seams are hard to identify. So expertly are these illusions constructed that it is hard to discern the true form of the curved ceiling. There is no peephole here, no need to close an eye and to keep the other unmoving. Goodman is stuck on the experience of small-scale anamorphoses and confuses the practical limitations of size with theoretical limitations. With a high ceiling at his disposal Pozzo constructs an *immersive* anamorphosis. At that large scale, the observer's two eyes work monocularly in good approximation. Also, the head can and is expected to rotate freely to contemplate the surrounding illusion. It is rather deceiving to say that the anamorphosis has to be seen "from a certain angle and distance." It has to be seen *from a certain point*, but on that point the eye is allowed to rotate freely. In such a large scale illusion as in Pozzo's ceiling at St. Ignazio's (or as in the large-scale nineteenth-century panoramas), the observer can even break a little with the theoretical impositions and get away with it. A large disk marks the spot from where the illusion of Pozzo should be observed, but a couple of steps away from it (or a few heads too tall or short) won't break the illusion substantially. It is quite remarkable that the arguments of Goodman could survive scrutiny in the twentieth century when Pozzo not only understood in practice the full possibilities of anamorphosis but wrote on them so eloquently and clearly in his treatise. If the eye is not innocent, far less is the mind so. It was simply the right cultural moment for such an argument to be entertained, in spite of the evidence of the eye. In the present work we would hardly engage with these arguments, which in our view have aged badly, except that they were also used in much the same way by none other than Barre and Flocon as a justification for their development of spherical perspective (Barre and Flocon 1968) or at least in the paper they wrote with Bouligand aiming at its mathematical and conceptual justification (Barre et al 1964). In the latter work the author's seek to justify their choice of spherical projection (the azimuthal equidistant projection, limited to one hemisphere) in the context of the search for a natural perspective, which in their view would satisfy the rather vague requirement of "ordering visible elements on a surface to form an image that causes on the spectator a sensation of volume and space" ("Ordonner sur une surface des éléments visibles, formant une image qui procure au spectateur des sensations de volumes e d'espace"). This desideratum (called "demande Δ") being somewhat vague, the authors seek to approach it by the constraint of an “axiome fondamental” A, a more geometrical axiom that should be, in their view, a necessary condition for such a perspective. This “axiome A” turns out to be just Leonardo’s axiom (Axiom 2). Watching what the authors proceed to do one finds that the “surface” mentioned in the requirement Δ is in fact assumed to be a plane and that “seems smaller” means metrically smaller. This is the exact same conflation of concepts that was made by Leonardo. The author’s give much the same objections to linear perspective as Leonardo, as Goodman, and in particular, two objections usually attributed to Panofsky (Panofsky 1927, 1991), but actually much older, which pretend to justify that whatever the “natural” perspective is, it can’t be classical perspective.

The first objection is that the projection of lines in the retina is curved, not straight; hence we must in fact not see lines as straight. This is a rather blunt misconception that implicitly assumes some sort of theater of the mind where some homunculus in the visual brain sits to see the retinal image (and then we might ask, in infinite regression, how does *its* visual system work?). This is in fact an old objection already aimed at Kepler’s model of vision, which Descartes answered by pointing out that the shape of the stimulus in the retina and the sensation caused by the stimuli are two different things. The retinal image is also inverted and yet we don’t see the world upside down (Kemp 1990, p. 234).

A second objection is that the example of two parallel lines going to infinity goes against Leonardo’s axiom, as we know from experience that the lines seem to converge yet in linear perspective the distance between their projections remains constant. This confuses the metric properties of the projection with the perception of these when seen from O. Araújo (2016) refutes this by pointing out that if we put the plane of projection on top of the lines then the lines will coincide with their perspective. Hence to claim that the perspective of the lines does not seem to converge would be to claim that the lines themselves don’t seem to converge. To put it more strikingly, if the perspective drawing of a long wall is not visually realistic then the real wall itself is not visually realistic. Of course the answer is that both the lines and their perspectives do appear to converge because the subtended angle from O decreases with distance in exactly the right way precisely because the drawn lines remain at a constant distance from each other. But much the same question had been raised by Schickhardt in 1623 and refuted by Kepler in a similar fashion (Kemp 1990, p.477-8). We can add that Leonardo’s paradox of the spheres has exactly the same answer: it doesn’t matter that the distal sphere has the largest projection, since the projection is seen from further away, in exactly the right measure to make it mimic the real sphere.

Being charitable and removing the layers of confusion and contradiction, one finds in the arguments of Barre, Flocon, and Bouligand one concrete and well-defined objective: to find a projection onto the plane that should cover a 180-degree angle of vision around an axis, project onto a bounded region of the plane, and be anamorphic in the whole representation. This would imply that the metric properties should be such as to save appearances (the angular properties). This is a concrete interpretation of desideratum Δ and Leonardo’s axiom. The only problem is that it can’t be done, which is also, in the end, the conclusion of Bouligand. So the azimuthal equidistant projection is chosen as simply the better compromise among the available candidates. It is all about minimizing metric deformations as far as possible. Hence the vagueness of desideratum Δ – the more strict requirement of actual mimesis could not be satisfied.

But of course, no matter what word juggling we use, nobody in the end mistakes a spherical perspective for the real thing. There is certainly a visual evocation of the object, but not mimesis in the strict sense. And this is not, as the authors seem to assume, a problem that comes merely from the sphere being non-developable. The cylinder is developable (isometric to the plane) and yet unrolling it destroys the anamorphic effect. In the case of Barre and Flocon's spherical perspective, there is simply no point O from which the drawing has a mimetic effect in the sense of the cone of the representation being the same as that of the represented object. The authors have it exactly the wrong way around: spherical perspective, unlike classical perspective (and unlike spherical *anamorphosis*), is entirely conventional. It requires an agreed upon, rather sophisticated intellectual translation in order to be read. It's achievements are of a different order: it is a flat record of one half of a spherical anamorphosis, which has good metrical properties and – crucially – is easy to draw by hand, since its line projections are arcs of circle. The authors do injustice to their own work by chasing a quixotic goal while underselling their true achievements – the rational description of a construction method for the first rigorous spherical perspective.

These philosophical misconceptions are not merely academic. Araújo (2016) argues that looking exclusively at the minimization of distortion is probably the main reason why Barre and Flocon did not extend their perspective beyond the hemispheric view. Extending their perspective to a full 360-degree view would require acceptance of rather egregious metric deformations, which would defeat their standard objections against linear perspective.

In the view we present in this work, these deformations are not a problem, as the spherical perspective does not intend to contend for the place of "natural perspective" and is by definition flawed in its mimetic properties. Mimesis is left to the anamorphic step, and we bind all the complicated phenomenological, optical, and physiological concepts purely within our axiom of radial occlusion. Rules of interpretation and empirical motivations are not allowed inside once the mathematical development begins.

In this view there is no contradiction between anamorphic mimesis and metric deformations on the perspective, just as there is no paradox between surveying and map making. The view that perspective is an entailment of an anamorphosis and a flattening also clarifies the fundamental role of anamorphosis and the arbitrary nature of perspective, giving a dual role to spherical anamorphosis and linear perspective: among anamorphoses, it gives the spherical one the canonical role, the points of the sphere being identified with rays, hence with the classes of anamorphic equivalence of spatial points; and to linear perspective it gives a special place, justifying the realist's view in its most naturalistic bent: linear perspective is indeed special, because it is the only perspective for which the flattening is trivial, that is, the only perspective that is still an anamorphosis. It therefore keeps a dual role of representation of visual information and of mimetic object, objectively creating a visual illusion, independently of any convention.

Conclusion

We have presented a theoretical framework that separates and delimits the notions of perspective and anamorphosis in such a way as to render mute long points of contention regarding whether or not curvilinear perspectives are "natural" or "correspond to the way we see." The question is found to be not well posed, and the difficulties vanish upon reformulation. Our view of anamorphosis may be seen as an update of Euclid's optics with a reworking of the assumptions involved and the addition of vanishing points, hence taming the notion of infinity, which Euclid is careful to avoid. We make a single empirical assumption – the principle of linear occlusion – abstracted into a simple geometrical notion, anamorphic equivalence, and proceed from there to obtain both practical geometrical constructions and an elegant, symmetric notion of vanishing set. We derive from this notion of anamorphosis a definition of compact central perspectives that includes all the usual curvilinear perspectives. This family, strictly speaking, excludes classical perspective, since the infinite plane is not a compact, yet it includes the restrictions of linear perspective to any compact, with classical perspective as a degenerate limiting case. Yet this notion – by viewing perspectives as an entailment of two separate steps, one being mimetic (anamorphosis) and the other representational (flattening) – clarifies the special role of linear perspective, giving it a distinguished position in the bestiary of perspectives, as the perspective that is simultaneously an anamorphosis. In this way we provide a language that naturally dissolves a philosophical misconception dating back at least to the Renaissance, simply by making the terms of discourse correspond to the operational practice of perspective.

Cross-References

- ► Anamorphosis: Between Perspective and Catoptrics
- ► Looking Through the Glass
- ► Spherical Perspective

Acknowledgments A. B. Araújo was funded by FCT Portuguese national funds through project UIDB/Multi/04019/2020

References

Andersen K (1992) Brook Taylor's Work on Linear Perspective: A Study of Taylor's Role in the History of Perspective Geometry. Including Facsimiles of Taylor's Two Books on Perspective. Springer, New York, vol 10, pp 1–67. https://doi.org/10.1007/978-1-4612-0935-5

Andersen K (2007) The geometry of an art: the history of the mathematical theory of perspective from Alberti to Monge. Springer Science & Business Media, New York

Apostol TM, Mnatsakanian MA (2007) Unwrapping curves from cylinders and cones. Am Math Mon 114(5):388–416

Araújo A (2015) A construction of the total spherical perspective in ruler, compass and Nail. https://arxiv.org/abs/1511.02969

Araújo A (2017a) Anamorphosis: optical games with perspective's playful parent. In: Silva JN (ed) Proceedings of the Recreational Mathematics Colloquium V (2017) – G4G Europe. Associação Ludus, Lisbon, pp 71–86

Araújo A (2017b) Cardboarding mixed reality with Durer machines. In: Proceedings of the 5th Conference on Computation, Communication, Aesthetics and X, pp 102–113

Araújo A (2017c) A geometria (descritiva) da anamorfose e das perspectivas curvilíneas. In: Workshop "Matemática e Arte", Sociedade Portuguesa de Matemática, pp 101–108

Araújo A (2017d) Guidelines for Drawing Immersive Panoramas in Equirectangular Perspective. In: Proceedings of the 8th International Conference on Digital Arts – ARTECH2017, ACM Press, Macau, China, pp 93–99. https://doi.org/10.1145/3106548.3106606

Araújo A (2018a) Let's Sketch in 360º: Spherical Perspectives for Virtual Reality Panoramas. In: Bridges 2018 Conference Proceedings, Tessellations Publishing, pp 637–644

Araújo AB (2016) Topologia, anamorfose, e o bestiário das perspectivas curvilíneas. Convocarte–Revista de Ciências da Arte 2:51–69

Araújo AB (2018b) Drawing equirectangular VR panoramas with ruler, compass, and protractor. J Sci Technol Arts 10(1):2–15. https://doi.org/10.7559/citarj.v10i1.471

Araújo AB (2018c) Ruler, compass, and nail: constructing a total spherical perspective. J Math Arts 12(2-3):144–169. https://doi.org/10.1080/17513472.2018.1469378

Araújo AB (2019a) Eq A sketch 360, a serious toy for drawing Equirectangular spherical perspectives. In: Proceedings of the 9th International Conference on Digital and Interactive Arts, ACM, Braga Portugal, pp 1–8. https://doi.org/10.1145/3359852.3359893

Araújo AB (2019b) A fisheye gyrograph: taking spherical perspective for a spin. In: Goldstine S, McKenna D, Fenyvesi K (eds) Proceedings of Bridges 2019: Mathematics, Art, Music, Architecture, Education, Culture, Tessellations Publishing, Phoenix, Arizona, pp 659–664

Araújo AB, Olivero LF, Antinozzi S (2019a) HIMmaterial: exploring new hybrid media for immersive drawing and collage. In: Proceedings of the 9th International Conference on Digital and Interactive Arts, ACM, Braga Portugal, pp 1–4. https://doi.org/10.1145/3359852.3359950

Araújo AB, Rossi A, Olivero LF (2019b) Boxing the visual sphere: towards a systematic solution of the cubical perspective. UID per il disegno (2019):33–40. https://doi.org/10.36165/1004

Baltrušaitis J (1983) Les Perspectives Dépravées. Flammarion

Barnard ST (1983) Interpreting perspective images. Artif Intell 21(4):435–462

Barre A, Flocon A (1968) La Perspective Curviligne. Flammarion, Paris

Barre A, Flocon A, Bouligand G (1964) 'Etude comparée de différentes méthodes de perspective, une perspective curviligne. Bulletin de la Classe des Sciences de La Académie Royale de Belgique 5(L)

Belisle B (2015) Nature at a glance: Immersive maps from panoramic to digital. Early Popular Visual Culture 13(4):313–335

Burton HE (1945) Euclid's optics. J Opt Soc 35(5):357–372

Cabezos Bernal PM (2015) Imágenes estereoscópicas aplicadas a la representación arquitectónica. PhD Thesis, Universitat Politècnica de València

Casas F (1983) Flat-sphere perspective. Leonardo 16(1):1–9

Catalano G (1986) Prospettiva Sferica. Università degli Studi di Palermo

Coates P, Arayici Y, Koskela LJ, Kagioglou M, Usher C, O' Reilly K (2010) The limitations of BIM in the architectural process. In: First International Conference on Sustainable Urbanization (ICSU 2010), Hong Kong, China

Collins DL (1992) Anamorphosis and the eccentric observer: inverted perspective and construction of the gaze. Leonardo 25(1):73–82

Correia JV, Romão L, Ganhão SR, da Costa MC, Guerreiro AS, Henriques DP, Garcia S, Albuquerque C, Carmo MB, Cláudio AP, Chambel T, Burgess R, Marques C (2013) A New Extended Perspective System for Architectural Drawings. In: Zhang J, Sun C (eds) Global design and local materialization, vol 369. Springer, Berlin/Heidelberg, pp 63–75. https://doi.org/10.1007/978-3-642-38974-0_6

Correia V, Romão L (2007) Extended perspective system. In: Proceedings of the 25th eCAADe International Conference, pp 185–192

Čučaković A, Paunović M (2016) Perspective in stage design: an application of principles of anamorphosis in spatial visualisation. Nexus Netw J 18(3):743–758. https://doi.org/10.1007/s00004-016-0297-5

Dept of Military Aeronautics USMADA (1918) Panoramic drawing, one-point and cylindrical perspective. G.P.O.

Draper SW (1978) The Penrose triangle and a family of related figures. Perception 7(3):283–296. https://doi.org/10.1068/p070283

Dunham D (2019) The Bridges 2018 mathematical art exhibitions. J Math Arts 1–15. https://doi.org/10.1080/17513472.2019.1654330

Dutour É (1760) Discussion d'une question d'optique. l'Académie des Sciences. Memoires de Mathematique et de physique presentes par Divers Savants 3:514–530

Ebbinghaus H (1902) Grundzüge Der Psychologie. Verlag von Viet & Co., Leipzig

Escher MC (1958) Belvedere

Escher MC (1972) The Graphic work of M. C. Escher – Introduced And Explained By The Artist, New, Revised and Expanded Edition. Ballantine Books, New York

Fasolo M, Mancini MF (2019) The 'Architectural' Projects for the Church of St. Ignatius by Andrea Pozzo. diségno (4):79–90. https://doi.org/10.26375/disegno.4.2019.09

Foley JD, Van FD, Van Dam A, Feiner SK, Hughes JF, Angel E, Hughes J (1996) Computer graphics: principles and practice, vol 12110. Addison-Wesley Professional

Frigg R, Hunter M (2010) Beyond Mimesis and Convention: Representation in Art and Science, vol 262. Springer

Gay F, Cazzaro I (2019) Venetian perspective boxes: When the images become environments. In: Luigini A (ed) Proceedings of the 1st International and Interdisciplinary Conference on Digital Environments for Education, Arts and Heritage. EARTH 2018. Advances in Intelligent Systems and Computing, vol 919. Springer, Cham pp 95–105. http://doi-org-443.webvpn.fjmu.edu.cn/10.1007/978-3-030-12240-9_11

Gilman D (1992) A new perspective on pictorial representation. Aust J Philos 70(2). https://doi.org/10.1080/00048409212345061

Gombrich EH (1960) Art and illusion; a study in the psychology of pictorial representation. Pantheon Books, New York

Goodman N (1968) Languages of art: an approach to a theory of symbols, 2nd edn. Hackett Publishing Company, Indianapolis

Grau O (1999) Into the belly of the image: historical aspects of virtual reality. Leonardo 32(5):365–371. https://doi.org/10.1162/002409499553587

Herdman WG (1853) A treatise on the curvilinear perspective of nature; and its applicability to art. John Weale & co., London

Hohenwarter M (2002) GeoGebra: Ein Softwaresystem für dynamische Geometrie und Algebra der Ebene. PhD thesis, Paris Lodron University, Salzburg, Austria

Hohenwarter M, Borcherds M, Ancsin G, Bencze B, Blossier M, Delobelle A, Denizet C, Éliás J, Fekete Á, Gál L, Konečný Z, Kovács Z, Lizelfelner S, Parisse B, Sturr G (2013) GeoGebra 4.4

Huffman DA (1968) Decision criteria for a class of 'impossible objects'. In: Proceedings of the First Hawaii International Conference on System Sciences, Honolulu

Huffman DA (1971) Impossible objects as nonsense sentences, Machine Intelligence 6. Machine Intelligence 6:295–323

Huhtamo E (2013) Illusions in motion – media archaeology of the moving panorama and related spectacles, 1st edn. Leonardo Book Series, The MIT Press

Inglis T (2018) Constructing 3D perspective anamorphosis via surface projection. In: Bridges 2018 Conference Proceedings, Tessellations Publishing, pp 91–98

Kemp M (1990) The science of art. Yale University Press, New Haven and London

Kim Y, Chin S (2019) An analysis of the problems of BIM-based drawings and implementation during the construction document phase. In: 36th International Symposium on Automation and Robotics in Construction, Banff. https://doi.org/10.22260/ISARC2019/0025

Kulpa Z (1983) Are impossible figures possible? Signal Process 5(3):201–220. https://doi.org/10.1016/0165-1684(83)90069-5

Michel G (2013) 'L'oeil, au Centre de la Sphere Visuelle. Boletim da Aproged (30)

Mitchell R (1801a) Plans, and views in perspective, with descriptions, of buildings erected in England and Scotland: and also an essay, to elucidate the Grecian, Roman and gothic architecture, accompanied with designs. Wilson & Company

Mitchell R (1801b) Section of the Rotunda, Leicester Square | British Library – Picturing Places. https://www.bl.uk/collection-items/section-of-the-rotunda-leicester-square

Mitrović B (2013a) Nelson Goodman's arguments against perspective: a geometrical analysis. Nexus Netw J 15(1):51–62. https://doi.org/10.1007/s00004-012-0133-5

Mitrović B (2013b) Visuality after Gombrich: the innocence of the eye and modern research in the philosophy and psychology of perception. Zeitschrift für Kunstgeschichte 76(H. 1):71–89

Monroe MM, Redmann WG (1994) Apparatus and method for projection upon a three-dimensional object

Moose M (1986) Guidelines for constructing a fisheye perspective. Leonardo 19(1):61–64

Morehead JC Jr (1955) Perspective and projective geometries: a comparison. Rice Institute Pamphlet-Rice University Studies 42(1):1–25

Necker LA (1832) Observations on some remarkable optical phænomena seen in Switzerland; and on an optical phænomenon which occurs on viewing a figure of a crystal or geometrical solid. Lond Edinb Dublin Philos Mag J Sci 1(5):329–337. https://doi.org/10.1080/14786443208647909

Norman DA (1990) Why interfaces don't work. The art of human-computer interface design 218

Oettermann S, Schneider DL (1997) The panorama: history of a mass medium, vol 2. Zone Books, New York

Olivero LF, Sucurado B (2019) Inmersividad analógica: Descubriendo el dibujo esférico entre subjetividad y objetividad. Estoa Revista de la Facultad de Arquitectura y Urbanismo de la Universidad de Cuenca 8(16):80–109

Olivero LF, Rossi A, Barba S (2019) A codification of the cubic projection to generate immersive models. diségno (4):53–63. https://doi.org/10.26375/disegno.4.2019.07

Panofsky E (1927) Die Perspektive asl'symbolische Form'. Vortrage der Bibliothek Warburg 1924–1925, vol 320

Panofsky E (1991) Perspective as symbolic form. Zone Books, New York

Papert S, Turkle S (1991) Epistemological Pluralism. In: Idit, Papert, Harel S (eds) Constructionism, Ablex Publishing Co., pp 161–191

Penrose LS, Penrose R (1958) Impossible objects: a special type of visual illusion. Br J Psychol 49(1):31–33. https://doi.org/10.1111/j.2044-8295.1958.tb00634.x

Roberts B, Harris MG, Yates TA (2005) The Roles of Inducer Size and Distance in the Ebbinghaus Illusion (Titchener Circles). Perception 34(7):847–856. https://doi.org/10.1068/p5273

Rossi A, Olivero LF, Barba S (2018) "CubeME", a variation for an immaterial rebuilding. In: Rappresentazione/Materiale/Immateriale Drawing as (in) Tangible Representation, Cangemi Editore, pp 31–36

Rossi M (2016) Architectural perspective between image and building. Nexus Netw J 18(3):577–583. https://doi.org/10.1007/s00004-016-0311-y

Sánchez-Reyes J, Chacón JM (2016) Anamorphic free-form deformation. Comput Aided Geom Des 46:30–42

Sánchez-Reyes J, Chacón JM (2020) How to make impossible objects possible: Anamorphic deformation of textured NURBS. Computer Aided Geom Des 78:101826. https://doi.org/10.1016/j.cagd.2020.101826

Spencer J (2018) Illusion as ingenuity: Dutch perspective boxes in the Royal Danish Kunstkammer's 'Perspective Chamber'. J Hist Collections 30(2):187–201

Sugihara K (1982) Classification of impossible objects. Perception 11(1):65–74. https://doi.org/10.1068/p110065

Sugihara K (2000) "Impossible objects" Are not necessarily impossible – mathematical study on optical illusion –. In: Goos G, Hartmanis J, van Leeuwen J, Akiyama J, Kano M, Urabe M (eds) Discrete and Computational Geometry, vol 1763, Springer, Berlin/Heidelberg, pp 305–316. https://doi.org/10.1007/978-3-540-46515-7_27

Sugihara K (2015a) Ambiguous cylinders: a new class of impossible objects. Comput Aided Drafting Des Manuf 25(4):19–25

Sugihara K (2015b) Height reversal generated by rotation around a vertical axis. J Math Psychol 68:7–12

Sugihara K (2016) Ambiguous Cylinder illusion. https://www.youtube.com/watch?v=oWfFco7K9v8

Sugihara K (2018) Topology-disturbing objects: a new class of 3D optical illusion. J Math Arts 12(1):2–18. https://doi.org/10.1080/17513472.2017.1368133

Termes D (1998) New perspective systems. self-published

Tomilin MG (2001) Anamorphoses-optical oddities of the Renaissance or sources of the science of image processing? J Opt Technol 68(9):723. https://doi.org/10.1364/JOT.68.000723

Tran Luciani D, Lundberg J (2016) Enabling designers to Sketch Immersive Fulldome presentations. In: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems – CHI EA '16, ACM Press, San Jose, California, USA, pp 1490–1496. https://doi.org/10.1145/2851581.2892343

Verstegen I (2011) Come dire oggettivamente che la prospettiva è relativa. Rivista di estetica (48):217–235

Verweij A (2010) Perspective in a box. In: Architecture, mathematics and perspective. Springer, pp 47–62

Vuibert H (1912) Les Anaglyphes Géométriques. Librairie Vuibert, Paris

Wade NJ, Ono H (2012) Early studies of binocular and stereoscopic vision 1. Jpn Psychol Res 54(1):54–70

Ware W (1900) Modern perspective: a treatise upon the principles and practice of plane and cylindrical perspective. The Macmillan company, New York; Macmillan & co., ltd., London

Wheatstone C (1838) Contributions to the physiology of vision. Part the first. on some remarkable, and hitherto unobserved, phenomena of binocular vision. Philos Trans R Soc Lond 128:371–394

Wheatstone C (1852) Contributions to the physiology of vision. Part the Second. On some remarkable, and hitherto unobserved, phenomena of binocular vision (continued). Philos Trans R Soc Lond 142:1–17

irreducible fraction q with $|p| > 1$ has exactly two adjacent fractions with denominator smaller than p. ...

Important:

Pág. 173 (193)

Each Surface admits more than one geometric shape

Pág. 174 (194)

Fig. 2 Moduli space for the surface of genus 2

Fig. 3 From a regular hexagon on the plane to the torus.

Pág. 183 (203) ~~Ana~~

Anamorphosis Forward Fast

Fig. 3 Two very different 3D objects with the same visual core.

Fig. 4 Two very different 3D objects with the same visual core, one surrounding the other.
Anamorphosis is radial and isotropic, with no

preferential axis.

Fig. 5 (184)(204)

Projection onto a plane.

There is no absolute difference between the linear perspective and the oblique anamorphosis.

Both are anamorphosis onto a plane: the latter merely projects the object to a region further away from the foot of the perpendicular

Pág. 190 (210)

Fig. 10 choice of triangulation.

Pág. 201 (221)

Fig. 14 Anamorphosis onto the sphere of two parallel lines.

The lines project as meridians that meet at two common vanishing points V and V*, antipodal to each other.

Pág. 202 (222)

Fig. 15, ~~the~~ two lines l and j intersecting at a right triangle.

Fig. 16 Anamorph of a cube on a sphere.

data

the cube defines a set of
distinct vanishing points
two for each set of parallel
edges (arestas)

Pág. 205 (226)

Fig. 18

Descriptive geometry construction
of the anamorphosis of a cube onto
a cylinder.

⇒ Projection
⇒ point the edges of the cube
appear straight (c)

! " # !# $% % & ' "$ ($) !
' # !# & # ! $% % & ' "
) '

Metamath.org

1) Pi is a real number:
http://us.metamath.org/mpeuni/pire.html

2) 2 < pi < 4.
http://us.metamath.org/mpeuni/pigt2lt4.html

3) 3 =< pi
http://us.metamath.org/mpeuni/pige3.html

$sen_{\alpha} = \frac{b}{a}\left(\frac{co}{h}\right)$

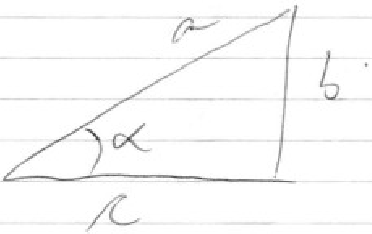

$\cos_{\alpha} = \frac{c}{b}$

$\tan_{\alpha} = \frac{b}{c}$

Seno de um ângulo

O seno de um ângulo é a razão entre o cateto oposto ao referido ângulo e a hipotenusa de um B triângulo retângulo.

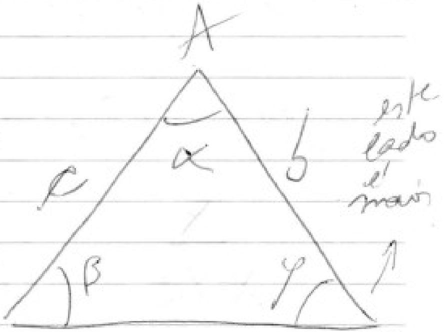

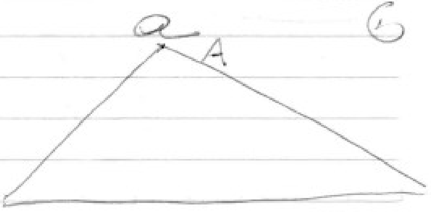

Ângulos notáveis

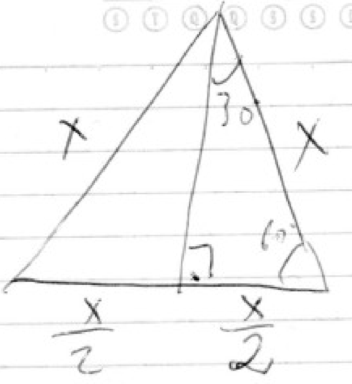
x
30°
x
60°
$\frac{x}{2}$
$\frac{x}{2}$

$\text{seno } 30° = \frac{CO}{h}$

$\text{sen } 60° = \frac{CA}{h}$

$\text{sen}(30°) = (x/2)/x = x/2 \cdot 1/x = 1/2$

$$\frac{\frac{x}{2}}{x} = \cancel{\frac{x}{2} \cdot \frac{1}{x}} = \frac{1}{2}$$

$\text{sen}(60°) = \frac{x}{h}$

$\text{sen } 45° =$

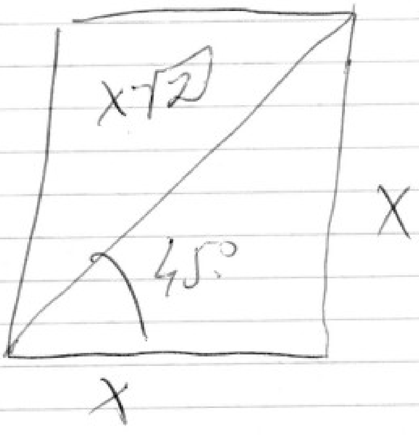
$x\sqrt{2}$
45°
x
x

$\text{sen } 45° =$

$$\frac{\cancel{x}}{\cancel{x}\sqrt{2}} = \frac{1}{\sqrt{2}}$$

Definição do seno de um ângulo.

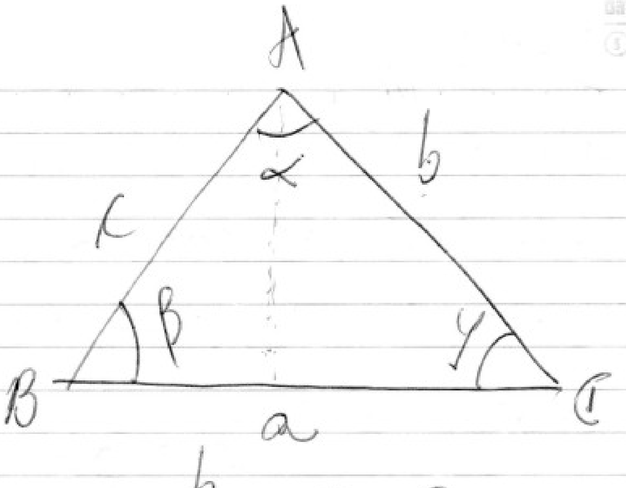

$$\frac{a}{sen\,\alpha} = \frac{b}{sen.\,\beta} = \frac{c}{sen\,\gamma}$$

↳ $\frac{co}{h}$

O seno de um ângulo é a razão entre o cateto oposto ao referido ângulo e a hipotenusa de um triângulo retângulo.

∴ divisão da medida do cateto oposto pela medida da hipotenusa

$$\boxed{sen = \frac{CO}{h}}$$

) $

Celestial Mechanics – Hagihara
(5 volumes)

Source:
\Arq_Barcos\04_02\

\Geometry_Gauss

Norman Megill
(Megill)
nm@alum.mit.edu

---

fusão nuclear com confinamento inercial.

→ obtenção do "plasma ardente" em que as próprias reações de fusão dão calor para mais fusões.

other:
fusão por confinamento magnético.

20210823///
https://www.cs.ru.nl/~freek/100/
Formalizing 100 Theorems

Megill's completeness theorem

) $

689895236653934908/

→ Geometria do Ovo

=

hence - portanto

lies - encontra-se em
on

—

transitivite (transitivity) theorem
for parallel lines

Pág. 158(177)

Source: Books-017-2021/
20210825/
Fermat/
Methods of Geometry.pdf

—

grandade híbrida generalizada -
Palatini.

—

ring light (LED)

=

Books: Source

Books-017-2021/20210825/

⇒

$
! *+,-* * *. ! * $ # / 0 1#

2 *+,-* * *. ! * $ 3 ! 4 5 34!
6 7! 8 0 1 ! ! # / ! , ! 7 0 8 1#
) 9 $

and so

$$\begin{pmatrix} t' \\ \boldsymbol{x}' \end{pmatrix} = \begin{pmatrix} 1 & \boldsymbol{o}^\mathsf{T} \\ \boldsymbol{o} & \boldsymbol{D} \end{pmatrix} \begin{pmatrix} t \\ \boldsymbol{x} \end{pmatrix} + \begin{pmatrix} t_o \\ \boldsymbol{x}_o \end{pmatrix}. \tag{1.13}$$

By moving the origin of the new coordinate system with the velocity $\boldsymbol{v}$ as mentioned above, one finally gets the general Galilei transformation:

$$\vec{\boldsymbol{x}}' = \begin{pmatrix} 1 & \boldsymbol{o}^\mathsf{T} \\ -\boldsymbol{v} & \boldsymbol{D} \end{pmatrix} \vec{\boldsymbol{x}} + \vec{\boldsymbol{x}}_o, \tag{1.14}$$

with

$$\vec{\boldsymbol{x}} \overset{\text{def}}{=} \begin{pmatrix} t \\ \boldsymbol{x} \end{pmatrix} \quad \text{and} \quad \vec{\boldsymbol{x}}_o \overset{\text{def}}{=} \begin{pmatrix} t_o \\ \boldsymbol{x}_o \end{pmatrix}.$$

This is an *affine mapping*, or an *affine transformation*.

One gets a *linear transformation* by introducing the extended vector:

$$\begin{pmatrix} t \\ \boldsymbol{x} \\ 1 \end{pmatrix} \in \mathbb{R}^5, \tag{1.15}$$

namely

$$\begin{pmatrix} t' \\ \boldsymbol{x}' \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & \boldsymbol{o}^\mathsf{T} & t_0 \\ -\boldsymbol{v} & \boldsymbol{D} & \boldsymbol{x}_0 \\ 0 & \boldsymbol{o}^\mathsf{T} & 1 \end{pmatrix} \begin{pmatrix} t \\ \boldsymbol{x} \\ 1 \end{pmatrix}. \tag{1.16}$$

Newton's fundamental laws are invariant also with respect to such a transformation. A general Galilei transformation is determined by 10 parameters: t_o, $\boldsymbol{x}_o \in \mathbb{R}^3$, $\boldsymbol{v} \in \mathbb{R}^3$ and $\boldsymbol{D} \in \mathbb{R}^{3\times 3}$. The rotation matrix $\boldsymbol{D}$, in fact, has only three main parameters since any general rotation consists of successive rotations performed around the x_1-, x_2- and x_3-axis, so the whole rotation is characterized by the three angles φ_1, φ_2 and φ_3, where, for example, the rotation around the x_1-axis is achieved by the matrix

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\varphi_1 & \sin\varphi_1 \\ 0 & -\sin\varphi_1 & \cos\varphi_1 \end{pmatrix}.$$

1.1.3 Maxwell's Equations and Galilei Transformation

The situation is completely different with Maxwell's equations of electromagnetic dynamics. They are not invariant with respect to a Galilei transformation! Indeed, in an inertial system $\mathcal{X}$, a static charge q generates only a static electric field; in an inertial frame, moving with the speed $\boldsymbol{v}$, $q\boldsymbol{v}$ is an electric current which generates there a magnetic field!

In the nineteenth century, it was believed that all physical phenomena are mechanical and electromagnetic forces could be traced to the stress states of a world-aether, the Maxwell's tensions. It was assumed that even a vacuum must be filled with aether. This aether then is the carrier of the electromagnetic phenomena.

Suppose an *inertial frame* is a reference system in which Galilei's principle of inertia is valid. Then Einstein in his *general relativity principle* claims:

The laws of nature take on the same form in all inertial systems.

For the fundamental law of mechanics, this is derived above. The principle of relativity applies neither to electrodynamics nor to optics. How should the basic equations of electrodynamics be modified so that the relativity principle is valid? This is the content of Einstein's *Theory of Special Relativity*. He expanded it in the *Theory of General Relativity*. This theory treats how the natural laws must be modified so that they are also valid in *accelerating* or not uniformly against each other moving reference systems.

1.2 Lorentz Transformation

1.2.1 Introduction

At the end of the nineteenth century, experiments were conceived which should determine the velocity of the Earth with respect to the resting cosmic aether. This speed relative to the aether can only be measured by an electromagnetic effect, e.g. the light wave propagation. But in the Michelson–Morley experiment in 1881 and 1887 no drift velocity was found. Einstein concluded:[1]

The speed of light c is always constant.

Independently of the movement of the light source and the observer, light has the same speed value in every inertial frame.

Assume a light pulse is produced at time $t = t' = 0$ in the two axis-parallel reference systems $\mathcal{X}$ and $\mathcal{X}'$ with a common origin. If the light spreads with the speed of light c in the reference system $\mathcal{X}'$, then, for example, it is true that $x_1' = c \cdot t$. From (1.7) it follows for the x_1-direction that if $\boldsymbol{v}$ has the x_1-directional component, then

$$x_1 = x_1' + vt = (c + v)t,$$

yielding, contrary to the Michelson–Morley experiment, a propagation velocity for light of $c + v > c$. Therefore, a transformation must be valid, different from the Galilei transformation. We try a linear transformation:

$$t = ft' + \boldsymbol{e}^{\mathsf{T}} \boldsymbol{x}', \tag{1.17}$$

---

[1] c from Latin celeritas: speed.

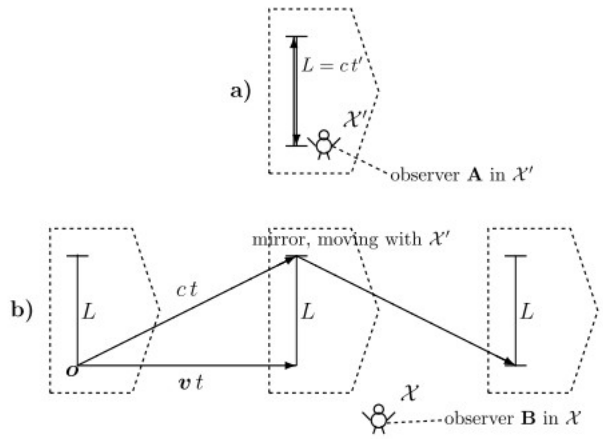
a)
$L = ct'$
$\mathcal{X}'$
observer **A** in $\mathcal{X}'$
b)
mirror, moving with $\mathcal{X}'$
ct
L
L
L
o
$v\,t$
$\mathcal{X}$
observer **B** in $\mathcal{X}$

Fig. 1.3 Different ways of light: (**a**) seen by an observer in $\mathcal{X}'$, (**b**) seen by an observer in $\mathcal{X}$

$$\boldsymbol{x} = \boldsymbol{b}t' + \boldsymbol{A}\boldsymbol{x}', \tag{1.18}$$

i.e.

$$\vec{\boldsymbol{x}} = \hat{\boldsymbol{L}}'\vec{\boldsymbol{x}}' \in \mathbb{R}^4,$$

with

$$\hat{\boldsymbol{L}}' \overset{\text{def}}{=} \begin{pmatrix} f & \boldsymbol{e}^{\mathsf{T}} \\ \boldsymbol{b} & \boldsymbol{A} \end{pmatrix} \in \mathbb{R}^{4\times 4}.$$

1.2.2 Determining the Components of the Transformation Matrix

That t' is different from t (according to Galilei this was not the case, he took $t' = t$) is shown by the following reasoning. We consider two observers; observer **A** moves relative to the observer **B** in a spaceship with velocity $\boldsymbol{v}$ (Fig. 1.3). The spaceship with observer **A** has the inertial system $\mathcal{X}'$, and the observer **B** on Earth has the inertial system $\mathcal{X}$. A beam of light moves from the origin $\boldsymbol{x} = \boldsymbol{x}' = \boldsymbol{o}$ of the reference systems $\mathcal{X}$ and $\mathcal{X}'$ at time $t = t' = 0$ perpendicular to the velocity $\boldsymbol{v}$ and reaches after t' seconds for the observer **A** in the moving reference system $\mathcal{X}'$ a mirror moving also with the reference system $\mathcal{X}'$. For the observer **B** in the stationary reference system $\mathcal{X}$, the light beam reaches the mirror, which has been moved in the $\boldsymbol{v}$ direction a distance of $v \cdot t$ after t seconds. Since in all initial systems the speed of light is constant and equal to c, by Pythagorean theorem, one has

$$(ct)^2 = (vt)^2 + (L)^2 = (vt)^2 + \left(ct'\right)^2,$$

or, after solving for t,

$$\underline{\underline{t = \gamma t'}} \tag{1.19}$$

with

$$\boxed{\gamma \overset{\text{def}}{=} \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}.} \tag{1.20}$$

A comparison of (1.19) with (1.17) provides for $\boldsymbol{x}' = \boldsymbol{o}$

$$\underline{\underline{f = \gamma}}. \tag{1.21}$$

For $\boldsymbol{x}' = \boldsymbol{o}$ one has $\boldsymbol{x} = \boldsymbol{v}t$, and from (1.18) it follows that $\boldsymbol{x} = \boldsymbol{b}t'$. Thus, $\boldsymbol{b}t' = \boldsymbol{v}t$, i.e. $\boldsymbol{b} = \boldsymbol{v}\frac{t}{t'}$. From (1.19) it follows, on the other hand, that $\frac{t}{t'} = \gamma$, so

$$\underline{\underline{\boldsymbol{b} = \gamma \boldsymbol{v}}}. \tag{1.22}$$

Up to now, the following transformation equations were determined:

$$t = \gamma t' + \boldsymbol{e}^{\mathsf{T}} \boldsymbol{x}', \tag{1.23}$$

$$\boldsymbol{x} = \gamma \boldsymbol{v} t' + \boldsymbol{A}\boldsymbol{x}'. \tag{1.24}$$

Equations (1.23) and (1.24) deliver a transformation of $\mathcal{X}'$ to $\mathcal{X}$. If one wants to invert this transformation, one must replace $\boldsymbol{v}$ by $-\boldsymbol{v}$, $\boldsymbol{x}$ by $\boldsymbol{x}'$, and vice versa, and t by t' and vice versa (as $\boldsymbol{A}$ and $\boldsymbol{e}$ may depend on $\boldsymbol{v}$, in the following we first write $\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{e}}$):

$$t' = \gamma t + \tilde{\boldsymbol{e}}^{\mathsf{T}} \boldsymbol{x},$$

$$\boldsymbol{x}' = -\gamma \boldsymbol{v} t + \tilde{\boldsymbol{A}}\boldsymbol{x},$$

which can be combined to

$$\vec{\boldsymbol{x}}' = \begin{pmatrix} \gamma & \tilde{\boldsymbol{e}}^{\mathsf{T}} \\ -\gamma \boldsymbol{v} & \tilde{\boldsymbol{A}} \end{pmatrix} \vec{\boldsymbol{x}} \overset{\text{def}}{=} \hat{\boldsymbol{L}}\vec{\boldsymbol{x}}. \tag{1.25}$$

Both transformations performed one after another must result in the identity matrix:

$$\hat{\boldsymbol{L}}'\hat{\boldsymbol{L}} \overset{!}{=} \boldsymbol{I}. \tag{1.26}$$

For the top left (1, 1)-element of the matrix product $\hat{\boldsymbol{L}}'\hat{\boldsymbol{L}}$ one gets:

$$\left(\gamma, \boldsymbol{e}^{\mathsf{T}}\right) \begin{pmatrix} \gamma \\ -\gamma \boldsymbol{v} \end{pmatrix} = \gamma^2 - \gamma \boldsymbol{e}^{\mathsf{T}} \boldsymbol{v} \overset{!}{=} 1.$$

Hence,

$$\gamma \boldsymbol{e}^{\mathsf{T}} \boldsymbol{v} = \gamma^2 - 1. \tag{1.27}$$

Taking for $\boldsymbol{e}$

$$\boldsymbol{e}^{\mathsf{T}} = \alpha \boldsymbol{v}^{\mathsf{T}} \tag{1.28}$$

and using (1.27) yields

$$\gamma \alpha v^2 = \gamma^2 - 1,$$

and moreover,

$$\gamma \alpha = \frac{1}{v^2}\left(\frac{c^2}{c^2 - v^2} - 1\right) = \frac{1}{c^2 - v^2} = \frac{\gamma^2}{c^2},$$

i.e.

$$\alpha = \frac{\gamma}{c^2}. \tag{1.29}$$

Equation (1.29) used in (1.28) finally yields

$$\underline{\underline{\boldsymbol{e}^\mathsf{T} = \frac{\gamma}{c^2}\boldsymbol{v}^\mathsf{T}.}} \tag{1.30}$$

Thus, till now we have calculated:

$$\hat{\boldsymbol{L}}' = \begin{pmatrix} \gamma & \frac{\gamma}{c^2}\boldsymbol{v}^\mathsf{T} \\ \gamma \boldsymbol{v} & \boldsymbol{A} \end{pmatrix}.$$

Obviously,

$$\tilde{\boldsymbol{e}}^\mathsf{T} = -\frac{\gamma}{c^2}\boldsymbol{v}^\mathsf{T}.$$

Suppose now that in (1.25) we have $\tilde{\boldsymbol{A}} = \boldsymbol{A}$. Then for the matrix element in the lower right corner of the matrix product $\hat{\boldsymbol{L}}'\hat{\boldsymbol{L}}$ in (1.26) we obtain

$$(\gamma \boldsymbol{v}, \boldsymbol{A}) \begin{pmatrix} -\frac{\gamma}{c^2}\boldsymbol{v}^\mathsf{T} \\ \boldsymbol{A} \end{pmatrix} = -\frac{\gamma^2}{c^2}\boldsymbol{v}\boldsymbol{v}^\mathsf{T} + \boldsymbol{A}^2 \overset{!}{=} \boldsymbol{I},$$

i.e.

$$\boldsymbol{A}^2 = \boldsymbol{I} + \frac{\gamma^2}{c^2}\boldsymbol{v}\boldsymbol{v}^\mathsf{T}. \tag{1.31}$$

From (1.27), by inserting (1.30), follows

$$\frac{\gamma^2}{c^2}v^2 = \gamma^2 - 1.$$

Plugging this into (1.31) provides

$$\boldsymbol{A}^2 = \boldsymbol{I} + \left(\gamma^2 - 1\right)\frac{\boldsymbol{v}\boldsymbol{v}^\mathsf{T}}{v^2}. \tag{1.32}$$

Since $(\gamma - 1)^2 = \gamma^2 - 2(\gamma - 1) - 1$, for $\gamma^2 - 1$ one can write

$$\gamma^2 - 1 = (\gamma - 1)^2 + 2(\gamma - 1). \tag{1.33}$$

Equation (1.33) inserted into (1.32) yields

$$\boldsymbol{A}^2 = \boldsymbol{I} + 2(\gamma - 1)\frac{\boldsymbol{v}\boldsymbol{v}^\intercal}{v^2} + (\gamma - 1)^2\frac{\boldsymbol{v}\boldsymbol{v}^\intercal}{v^2} = \left(\boldsymbol{I} + (\gamma - 1)\frac{\boldsymbol{v}\boldsymbol{v}^\intercal}{v^2}\right)^2,$$

where

$$\frac{\boldsymbol{v}\boldsymbol{v}^\intercal\boldsymbol{v}\boldsymbol{v}^\intercal}{v^4} = \frac{\boldsymbol{v}(\boldsymbol{v}^\intercal\boldsymbol{v})\boldsymbol{v}^\intercal}{v^4} = \frac{\boldsymbol{v}\boldsymbol{v}^\intercal}{v^2}$$

was used. Therefore,

$$\underline{\underline{\boldsymbol{A} = \boldsymbol{I} + (\gamma - 1)\frac{\boldsymbol{v}\boldsymbol{v}^\intercal}{v^2}}}. \tag{1.34}$$

It is, in fact, true that $\boldsymbol{A}(-\boldsymbol{v}) = \boldsymbol{A}(\boldsymbol{v})$, i.e. the above assumption that $\tilde{\boldsymbol{A}} = \boldsymbol{A}$ is correct.

Thus, the matrix $\hat{\boldsymbol{L}}$ of the Lorentz transformation is determined completely as

$$\boxed{\hat{\boldsymbol{L}} = \left(\begin{array}{c|c} \gamma & -\frac{\gamma}{c^2}\boldsymbol{v}^\intercal \\ \hline -\gamma\boldsymbol{v} & \boldsymbol{I} + (\gamma - 1)\frac{\boldsymbol{v}\boldsymbol{v}^\intercal}{v^2} \end{array}\right).} \tag{1.35}$$

For $c \to \infty$ we get $\gamma = 1$ and the Lorentz transformation turns into the Galilei transformation with $\boldsymbol{T}_{\text{Galilei}}$.

In the often in the textbooks treated special case when the velocity $\boldsymbol{v}$ is towards the x_1-axis, i.e.

$$\boldsymbol{v} = \begin{pmatrix} v \\ 0 \\ 0 \end{pmatrix},$$

one obtains

$$\hat{\boldsymbol{L}} = \left(\begin{array}{c|c} \gamma & -\frac{\gamma}{c^2}(v, 0, 0) \\ \hline -\gamma\begin{pmatrix} v \\ 0 \\ 0 \end{pmatrix} & \boldsymbol{I} + \frac{(\gamma-1)}{v^2}\begin{pmatrix} v^2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \end{array}\right) = \begin{pmatrix} \gamma & -\frac{\gamma v}{c^2} & 0 & 0 \\ -\gamma v & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

so

$$\boxed{\begin{aligned} t' &= \gamma t - \tfrac{\gamma}{c^2} v x_1, \\ x_1' &= -\gamma v t + \gamma x_1, \\ x_2' &= x_2, \\ x_3' &= x_3. \end{aligned}} \tag{1.36}$$

Introducing as the time component in the vector $\vec{x}$ the time multiplied with the speed of light c, i.e. $x_0 = ct$, we obtain from (1.35) the transformation

$$\vec{x}' \overset{\text{def}}{=} \begin{pmatrix} ct' \\ \boldsymbol{x}' \end{pmatrix} = \left(\begin{array}{c|c} \gamma & -\frac{\gamma}{c}\boldsymbol{v}^\mathsf{T} \\ \hline -\frac{\gamma}{c}\boldsymbol{v} & \boldsymbol{I} + (\gamma - 1)\frac{\boldsymbol{v}\boldsymbol{v}^\mathsf{T}}{v^2} \end{array}\right) \begin{pmatrix} ct \\ \boldsymbol{x} \end{pmatrix}, \tag{1.37}$$

i.e. the new transformation matrix

$$\boxed{\boldsymbol{L}(\boldsymbol{v}) \overset{\text{def}}{=} \left(\begin{array}{c|c} \gamma & -\frac{\gamma}{c}\boldsymbol{v}^\mathsf{T} \\ \hline -\frac{\gamma}{c}\boldsymbol{v} & \boldsymbol{I} + (\gamma - 1)\frac{\boldsymbol{v}\boldsymbol{v}^\mathsf{T}}{v^2} \end{array}\right)} \tag{1.38}$$

is now a *symmetric matrix*, and so

$$\boxed{\begin{aligned} ct' &= \gamma ct - \frac{\gamma}{c}\boldsymbol{v}^\mathsf{T}\boldsymbol{x}, \\ \boldsymbol{x}' &= \boldsymbol{x} + (\gamma - 1)\frac{\boldsymbol{v}^\mathsf{T}\boldsymbol{x}}{v^2}\boldsymbol{v} - \gamma\boldsymbol{v}t. \end{aligned}} \tag{1.39}$$

1.2.3 Simultaneity at Different Places

It will be seen that events at different places which are simultaneous for an observer in $\mathcal{X}$ are not in general simultaneous for a moving observer in $\mathcal{X}'$. This is caused by the finite speed of light. The reference system $\mathcal{X}'$ may move towards the stationary reference system $\mathcal{X}$ with velocity $\boldsymbol{v}$. If the two events E_1 and E_2 in $\mathcal{X}$ have the coordinates $\vec{x}_1$ and $\vec{x}_2$, then they are *simultaneous* if $t_1 = t_2$. Do these two events then happen also at the same time for an observer in the reference system $\mathcal{X}'$? Due to (1.37),

$$ct_1' = \gamma ct_1 - \frac{\gamma}{c}\boldsymbol{v}^\mathsf{T}\boldsymbol{x}_1$$

and

$$ct_2' = \gamma ct_2 - \frac{\gamma}{c}\boldsymbol{v}^\mathsf{T}\boldsymbol{x}_2.$$

Dividing both equations by c and subtracting one from another yields

$$t_1' - t_2' = \gamma(t_1 - t_2) - \frac{\gamma}{c^2}\boldsymbol{v}^\mathsf{T}(\boldsymbol{x}_1 - \boldsymbol{x}_2),$$

so

$$\underline{\underline{t_2' - t_1' = \frac{\gamma}{c^2}\boldsymbol{v}^\mathsf{T}(\boldsymbol{x}_1 - \boldsymbol{x}_2).}}$$

The events $\vec{x}_1'$ and $\vec{x}_2'$ are also simultaneous only if the velocity $\boldsymbol{v}$ is perpendicular to the local difference $\boldsymbol{x}_1 - \boldsymbol{x}_2$. Conclusion:

Fig. 1.4 Simultaneity

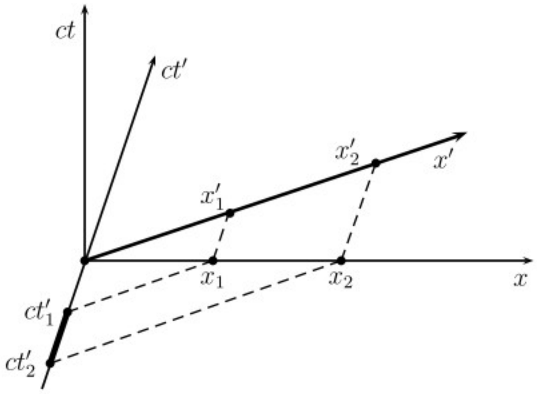

> Events at different locations, which in the reference system $\mathcal{X}$ are simultaneous, need *not* be simultaneous when seen from the reference system $\mathcal{X}'$.

Example For the special case $\boldsymbol{v} = [v, 0, 0]^\mathsf{T}$ one only needs to consider the ct- and the x-coordinate (Minkowski diagram, Fig. 1.4). The Lorentz transformation does not change the x_2- and x_3-component. It is therefore possible to restrict the consideration to the two-dimensional transformations:

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \begin{pmatrix} \gamma & -\frac{\gamma}{c}v \\ -\frac{\gamma}{c}v & \gamma \end{pmatrix} \begin{pmatrix} ct \\ x \end{pmatrix}.$$

The event $\binom{0}{x_1}$ is transformed to

$$\begin{pmatrix} ct'_1 \\ x'_1 \end{pmatrix} = \begin{pmatrix} \gamma & -\frac{\gamma}{c}v \\ -\frac{\gamma}{c}v & \gamma \end{pmatrix} \begin{pmatrix} 0 \\ x_1 \end{pmatrix} = \begin{pmatrix} -\frac{\gamma}{c}vx_1 \\ \gamma x_1 \end{pmatrix}$$

and accordingly the event $\binom{0}{x_2}$, simultaneous in $\mathcal{X}$, is transformed to $\binom{-\frac{\gamma}{c}vx_2}{\gamma x_2}$ in $\mathcal{X}'$. In Fig. 1.4, the difference on the ct'-axis is then

$$ct'_2 - ct'_1 = \frac{\gamma}{c}v(x_1 - x_2) \neq 0,$$

if $v \neq 0$ and $x_1 \neq x_2$, i.e. in the moving reference system $\mathcal{X}'$ the two events E_1 and E_2 are no longer simultaneous.

1.2.4 Length Contraction of Moving Bodies

Einstein (1905) was the first who completely demonstrated that length contraction is an effect due to the change in the notions of space, time and simultaneity brought about by special relativity. Length contraction can simply be derived from the Lorentz transformation. Let the reference system $\mathcal{X}$ be stationary and the reference system $\mathcal{X}'$ moving towards it with the speed $\boldsymbol{v}$. A yardstick has in the stationary system the two endpoints $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Its resting length is

$$\boldsymbol{l}_0 = \boldsymbol{x}_2 - \boldsymbol{x}_1, \tag{1.40}$$

i.e.

$$l_0^2 = (\boldsymbol{x}_2 - \boldsymbol{x}_1)^\mathsf{T}(\boldsymbol{x}_2 - \boldsymbol{x}_1). \tag{1.41}$$

At time t', the endpoints of the yardstick in the moving reference system $\mathcal{X}'$ have the coordinates $\boldsymbol{x}'_1$ and $\boldsymbol{x}'_2$. With (1.38) we obtain

$$\boldsymbol{l}_0 = \boldsymbol{x}_2 - \boldsymbol{x}_1 = \boldsymbol{A}\big(\boldsymbol{x}'_2 - \boldsymbol{x}'_1\big) \overset{\text{def}}{=} \boldsymbol{A}\boldsymbol{l}'. \tag{1.42}$$

It follows that

$$l_0^2 = \boldsymbol{l}_0^\mathsf{T}\boldsymbol{l}_0 = \boldsymbol{l}'^\mathsf{T}\boldsymbol{A}^2\boldsymbol{l}'. \tag{1.43}$$

With (1.32) one gets

$$l_0^2 = \boldsymbol{l}'^\mathsf{T}\left(\boldsymbol{I} + \big(\gamma^2 - 1\big)\frac{\boldsymbol{v}\boldsymbol{v}^\mathsf{T}}{v^2}\right)\boldsymbol{l}' = \boldsymbol{l}'^\mathsf{T}\boldsymbol{l}' + \big(\gamma^2 - 1\big)\frac{(\boldsymbol{v}^\mathsf{T}\boldsymbol{l}')^2}{v^2}. \tag{1.44}$$

In the product $\boldsymbol{v}^\mathsf{T}\boldsymbol{l}'$, only the component $\boldsymbol{l}'_\parallel$ of $\boldsymbol{l}'$ parallel to the velocity $\boldsymbol{v}$ comes into effect, i.e. it is true that $\boldsymbol{v}^\mathsf{T}\boldsymbol{l}' = \boldsymbol{v}^\mathsf{T}\boldsymbol{l}'_\parallel$. Thus,

$$l_0^2 = l'^2 + \big(\gamma^2 - 1\big)\frac{(\boldsymbol{v}^\mathsf{T}\boldsymbol{l}'_\parallel)^2}{v^2} = l'^2 + \big(\gamma^2 - 1\big){l'_\parallel}^2,$$

so

$$\underline{\underline{l'^2 = l_0^2 - \big(\gamma^2 - 1\big)l'^2_\parallel.}} \tag{1.45}$$

Since always $\gamma^2 - 1 \geq 0$, it follows from (1.45) that

$$\underline{\underline{l' \leq l_0.}} \tag{1.46}$$

The result of the yardstick length measurement therefore depends on the reference system in which the length measurement was made.

If the yardstick is parallel to the velocity $\boldsymbol{v}$, then $\boldsymbol{l}' = \boldsymbol{l}'_\parallel$ and (1.45) becomes $\gamma l' = l_0$, or

$$\boxed{l' = \frac{l_0}{\gamma} = l_0\sqrt{1 - \frac{v^2}{c^2}},} \tag{1.47}$$

i.e. as $v \to c$ one gets $l' \to 0$. For example, if the velocity is $v = 0.8c$ (i.e. 80 % of the speed of light), then the length is $l' = 0.6l_o$.

Einstein in Matrix Form
Exact Derivation....pdf

Pág. 6 (19) 1.13 Maxwell's Eq.
and Galilei Transformation
5 (19) 1.1.2
General Galilei Transformation

+

Yehuda B. Band ...- Quantum
mechanics ... (2013).pdf

Pág. 208 (229) 5.2.2
Quantum Entanglement
and Bell States

$(\nearrow \sqrt{2})$

Problem 5.9

Write the Bell states (5.25a) as four-component vectors using the notation of Problem 5.8.

$$|\phi^+\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\0\\0\\1\end{pmatrix}, \ldots$$

+,- * *. ! * $: #! ; ;! #< : ! +7 !
- ! ! !11 ! ! 0 ! # ! + !#
" 1#
) $ '

5.2.2 QUANTUM ENTANGLEMENT AND BELL STATES

Quantum entanglement for two-particle systems in a pure state was introduced in Sec. 1.3.3. It will be useful to redefine it, so that the definition also applies to *multiqubit states* of the form (5.23), as a property of states in a product Hilbert space. Moreover, it will be useful to redefine it for mixed states as well and for arbitrary n-level systems. First, consider a product of two spaces $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$. A pure state $|\Psi\rangle \in \mathcal{H}$ is said to be entangled if it cannot be written as a single tensor product of states $|\psi_1\rangle \in \mathcal{H}_1$ and $|\psi_2\rangle \in \mathcal{H}_2$, i.e., $|\Psi\rangle \neq |\psi_1\rangle \otimes |\psi_2\rangle$. Rather, it must be written as a *sum* of such products, for example, $|\Psi\rangle = a|\psi_1\rangle \otimes |\psi_2\rangle + b|\phi_1\rangle \otimes |\phi_2\rangle$.

Quantum entanglement occurring in multipartite qubit states is a powerful computation and information resource. For example, entanglement allows for dense coding of quantum information (see Sec. 5.2.5) and teleportation of quantum information (see Sec. 5.2.7). One of the central questions in quantum information is how to quantify entanglement for arbitrary n qubit states. For two-qubit pure states, there is a set of four *maximally entangled* states called *Bell states* denoted as $|BS\rangle$ and defined as follows:

$$|\Phi^+\rangle = \frac{1}{\sqrt{2}}[|00\rangle + |11\rangle], \quad |\Phi^-\rangle = \frac{1}{\sqrt{2}}[|00\rangle - |11\rangle], \tag{5.25a}$$

$$|\Psi^+\rangle = \frac{1}{\sqrt{2}}[|01\rangle + |10\rangle], \quad |\Psi^-\rangle = \frac{1}{\sqrt{2}}[|01\rangle - |10\rangle]. \tag{5.25b}$$

Clearly, Bell states cannot be written as a tensor product of single qubit states. In Sec. 5.2.3, we will present methods for explicitly constructing Bell states.

The Einstein Podolsky Rosen (EPR) paradox [20], Bell inequalities [52–55], and many quantum information and quantum computation algorithms use the Bell states extensively. Bell states and EPR concepts play an important role in testing fundamental concepts in quantum mechanics. EPR and Bell inequalities will be discussed in Secs 5.7 and 5.8.

An important property of Bell states $|BS\rangle$ (and entangled pure states in general) is: When a pure state density matrix $\rho_{\mathrm{BS}} = |\mathrm{BS}\rangle\langle\mathrm{BS}|$ is constructed from a Bell state, and is partially traced over (say) the second qubit, $\mathrm{Tr}_2\,[|\mathrm{BS}\rangle\langle\mathrm{BS}|] = \frac{1}{2}\,[|0\rangle\langle 0| + |1\rangle\langle 1|]$, the result is an incoherent state (mixed density matrix). In the case of Bell states, the first qubit, left after tracing out the second qubit, has 50% probability to be in either state $|0\rangle$ or $|1\rangle$.

The entangled pair of qubits in a Bell state could be spin 1/2 particles, where the state $|\downarrow\rangle$ can represent the qubit $|0\rangle$ and $|\uparrow\rangle$ can represent the qubit $|1\rangle$, or alternatively, the qubits could be photons, where a horizontal photon $|\leftrightarrow\rangle$ can represent the qubit $|0\rangle$ and a vertical photon $|\updownarrow\rangle$ can represent the qubit $|1\rangle$ as shown in Fig. 5.17 in Sec. 5.6.5. A method of entangling photons using degenerate down-conversion (degenerate difference frequency generation) will be described in Sec. 5.6.5. All one-qubit unitary transformations can be implemented using beam splitters and phase shifters, which will be described in Sec. 5.6.5. Moreover, beam splitters and phase shifters can be used to turn photons that are entangled in one of the Bell states into one of the other Bell states. So, photons are viable qubits that can be used for quantum information processing. The only problem is that photons do not interact (at least not in vacuum), so it might appear that two-qubit controlled gates (see below) cannot be engineered for photons, and these are essential for universal quantum computation (a universal set of gates are a set of gates such that any function can be computed using these gates). It was shown that *non-deterministic* two-qubit controlled gates for photons can be engineered [56], as described in Sec. 5.6.5.

An example of an entangled n-qubit state can be constructed using n photons in n different spatial modes (think of photon modes as being specified by the wavevector $\mathbf{k}$, such that photons in different modes correspond to photons with different wavevectors). Since photons are bosons, their state vector must be symmetrized (see Chapter 8). Such symmetrized states can be constructed by applying the symmetrization operator, $\mathcal{S}$,[4] to the state $|\mathbf{k}_1\rangle_1|\mathbf{k}_2\rangle_2 \ldots |\mathbf{k}_n\rangle_n$:

$$|\Psi_{\mathcal{S}}\rangle \equiv \sqrt{n!}\,\mathcal{S}\,|\mathbf{k}_1\rangle_1|\mathbf{k}_2\rangle_2 \ldots |\mathbf{k}_n\rangle_n = \frac{1}{\sqrt{n!}}\big[|\mathbf{k}_1\rangle_1|\mathbf{k}_2\rangle_2 \ldots |\mathbf{k}_n\rangle_n + |\mathbf{k}_2\rangle_1|\mathbf{k}_1\rangle_2 \ldots |\mathbf{k}_n\rangle_n + \cdots + |\mathbf{k}_n\rangle_1|\mathbf{k}_{n-1}\rangle_2 \ldots |\mathbf{k}_1\rangle_n\big]. \tag{5.26}$$

---

[4] The symmetrization operator, $\mathcal{S}$, [see Sec. 8.1, Eq. (8.7)], is the sum of all possible permutations of the n-qubits divided by $n!$. The additional factor of $\sqrt{n!}$ in Eq. (5.26) is for normalization, as detailed in Sec. 8.6.

There are $n!$ terms in Eq. (5.26) corresponding to the $n!$ permutations of the photons. One of the challenges of all-optical quantum computing is how to engineer such states, but, as mentioned earlier, another challenge is how to construct two-qubit gates for photons, since they do not interact in vacuum.

Problem 5.9

(a) Write the Bell states (5.25a) as four-component vectors using the notation of Problem 5.8.
(b) Write the density matrix for $|\Phi^-\rangle$ and $|\Psi^-\rangle$ as 4×4 matrices.

Answer:

$$|\Phi^+\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\0\\0\\1\end{pmatrix}, \quad |\Phi^-\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\0\\0\\-1\end{pmatrix}, \quad |\Psi^+\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\1\\1\\0\end{pmatrix}, \quad |\Psi^-\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\1\\-1\\0\end{pmatrix}. \tag{5.27}$$

$$\rho_{\Phi^-} = \frac{1}{2}\begin{pmatrix}1 & 0 & 0 & -1\\0 & 0 & 0 & 0\\0 & 0 & 0 & 0\\-1 & 0 & 0 & 1\end{pmatrix}, \quad \rho_{\Psi^-} = \frac{1}{2}\begin{pmatrix}0 & 0 & 0 & 0\\0 & 1 & -1 & 0\\0 & -1 & 1 & 0\\0 & 0 & 0 & 0\end{pmatrix}. \tag{5.28}$$

Entanglement Entropy

Let us try to quantify the degree of entanglement in a bipartite state, a topic that was briefly discussed in Sec. 2.5.2. We introduce a quantity called *entanglement entropy* that serves as a measure for the degree of entanglement for a bipartite pure state. Let $|\Psi\rangle \in \mathcal{H}_1 \otimes \mathcal{H}_2$ be an entangled state (e.g., $|\Psi\rangle = |BS\rangle$) and let $\rho_{12} = |\Psi\rangle\langle\Psi|$ be the pure state density matrix constructed from $|\Psi\rangle$. The reduced mixed density matrices for the two subsystems is obtained by partial tracing: $\rho_1 = \mathrm{Tr}_2\rho_{12}$, $\rho_2 = \mathrm{Tr}_1\rho_{12}$. The entanglement entropy of the bipartite pure state ρ_{12} is defined as,

$$\mathcal{E}(\rho_{12}) \equiv \mathcal{S}(\rho_1) = \mathcal{S}(\rho_2), \tag{5.29}$$

where $\mathcal{S}(\rho) = -\mathrm{Tr}\,\rho\log_2\rho$ is the von Neumann entropy of ρ. The proof that $\mathcal{S}(\rho_1) = \mathcal{S}(\rho_2)$ will be given below after discussing the Schmidt decomposition.

Problem 5.10

(a) Show that the entanglement entropy of $|BS\rangle\langle BS|$ equals 1.
(b) Consider the entangled state $|\Psi\rangle = 0.8|00\rangle + 0.6|11\rangle$. Show that the entanglement entropy $\mathcal{E}(\rho = |\Psi\rangle\langle\Psi|) = 0.699722$.
(c) Show that the result $\mathcal{E}(|BS\rangle\langle BS|) > \mathcal{E}(\rho = |\Psi\rangle\langle\Psi|)$ is not accidental, namely, $|BS\rangle\langle BS|$ has the maximal entanglement entropy $\mathcal{E}(|BS\rangle\langle BS|)$ among all pure states constructed from entangled states.

Greenberger–Horne–Zeilinger States

Entangled states can be composed of more than two qubits, e.g., the three-qubit state known as the *Greenberger–Horne–Zeilinger* (GHZ) state [52],

$$|\mathrm{GHZ}\rangle = \frac{1}{\sqrt{2}}(|000\rangle + |111\rangle), \tag{5.30}$$

or the n-qubit GHZ state,

$$|\text{GHZ}\rangle = \frac{|0\rangle^{\otimes n} + |1\rangle^{\otimes n}}{\sqrt{2}}. \tag{5.31}$$

These states are not a simple tensor product of the states of each qubit but rather a superposition that describes quantum correlations (entanglement) between the qubits. The entanglement between the n-qubits means that the state of the system cannot be specified in terms of the state of each of the n qubits individually (the state is not a product state).[5] Note that this sort of state is what we called a Schrödinger cat state at the end of Sec. 1.3.1. There we noted that these kinds of states tend to decohere quickly by virtue of interactions with an outside environment.

Although there is no standard measure for multi-qubit state entanglement (particularly for mixed states), the GHZ states are believed to be maximally entangled. Note that the trace over the third qubit (denoted by Tr_3) in a three-qubit GHZ state yields a maximally mixed two-qubit state:

$$\text{Tr}_3\left[(|000\rangle + |111\rangle)(\langle 000| + \langle 111|)\right] = \frac{1}{2}\left[|00\rangle\langle 00| + |11\rangle\langle 11|\right].$$

Entanglement of qubits will be discussed from the perspective of correlation between the particles in Sec. 6.1.8, where we consider two-qubit entanglement, three-qubit entanglement, mutli-qubit entanglement, and also qutrit entanglement (Sec. 6.2) as well as "continuous-variable entanglement" (Sec. 6.5).

Schmidt Decomposition

An important and useful result concerning pure states of a bipartite system emerges from the *Schmidt decomposition theorem* of linear algebra (see below). We have in mind two Hilbert spaces $\mathcal{H}_A$ and $\mathcal{H}_B$ of respective dimensions n_A and n_B and their tensor product space $\mathcal{H}_{AB} = \mathcal{H}_A \otimes \mathcal{H}_B$. The theorem states that for pure states $|\psi\rangle \in \mathcal{H}_{AB}$, there exist two orthonormal sets of states $\{|\xi_{i,A}\rangle \in \mathcal{H}_A\}$ and $\{|\varphi_{i,B}\rangle \in \mathcal{H}_B\}$ with $i = 1, 2, \ldots, \mathcal{N}_S$, where $\mathcal{N}_S = \min(n_A, n_B)$, such that the state $|\psi\rangle$ can be written as a sum,

$$\boxed{|\psi\rangle = \sum_{i=1}^{\mathcal{N}_S} \lambda_i \, |\xi_{i,A}\rangle|\varphi_{i,B}\rangle, \quad 0 \le \lambda_i \le 1, \ \sum_{i=1}^{\mathcal{N}_S} \lambda_i^2 = 1.} \tag{5.32}$$

The numbers λ_i are called *Schmidt coefficients*, and their number $\mathcal{N}_S$ is called the *Schmidt number*. The basis functions $|\xi_{i,A}\rangle$ and $|\varphi_{i,B}\rangle$ are called *Schmidt basis functions*. Note that we can write Eq. (5.32) as $|\psi\rangle = \sum_i \sqrt{p_i}\, |\xi_{i,A}\rangle|\varphi_{i,B}\rangle$, where p_i are the occupation probabilities for the states $|\xi_{i,A}\rangle$ and $|\varphi_{i,A}\rangle$, i.e., $\lambda_i = \sqrt{p_i}$.

The proof of Eq. (5.32) uses the singular value decomposition theorem [57]. Consider the expansion of the pure state $|\psi\rangle \in \mathcal{H}_A \otimes \mathcal{H}_B$ in terms of bases $\{|j\rangle \in \mathcal{H}_A$ and $\{|k\rangle \in \mathcal{H}_B$, $|\psi\rangle = \sum_{j,k} a_{jk}|j\rangle|k\rangle$. If $\dim \mathcal{H}_A \neq \dim \mathcal{H}_B$, then the matrix $a = \{a_{jk}\}$ is not a square matrix and therefore cannot be diagonalized. The theorem states that the matrix a can be written as a product $a = u\, d\, v^\dagger$, where u and v are unitary matrices and d is a diagonal matrix with nonnegative elements d_{ii}; these nonzero diagonal elements of d are called the *singular values* of the matrix a. Thus,

$$|\psi\rangle = \sum_{j,k} a_{jk}|j\rangle|k\rangle = \sum_{i=1}^{\mathcal{N}_S} \sum_{j,k} u_{ji} d_{ii} v_{ki}^\dagger |j\rangle|k\rangle. \tag{5.33}$$

Letting $|\xi_{i,A}\rangle = \sum_j u_{ji}|j\rangle$, $|\varphi_{i,B}\rangle = \sum_k v_{ki}^\dagger|k\rangle$, and $\lambda_i = d_{ii}$, yields Eq. (5.32). The singular value decomposition of a matrix a is pictorially shown in Fig. 5.2.

---

[5] The state in Eq. (5.31) is sometimes called a *noon* state, since it can be written as $|\text{noon}\rangle = (|n\rangle_0|0\rangle_1 + |0\rangle_0|n\rangle_1)/\sqrt{2}$, which is a superposition of n particles in state 0 and zero particles in state 1, and vise versa. A more general form of Eq. (5.31) that contains a phase angle is $|\text{GHZ}\rangle = (|0\rangle^{\otimes n} + e^{i\theta}|1\rangle^{\otimes n})/\sqrt{2}$.

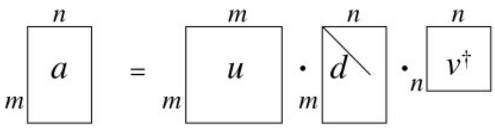

FIG 5.2 Singular value decomposition of an $m \times n$ matrix a into $a = u\,d\,v^\dagger$. The bullets imply matrix multiplication. u is an $m \times m$ matrix, d is an $m \times n$ matrix, and v is an $n \times n$ matrix.

Problem 5.11

(a) Show that $\{|\xi_{i,A}\rangle\}$ and $\{|\varphi_{i,B}\rangle\}$ are orthonormal sets of vectors using the unitarity of u and v.
(b) Under what condition on the amplitudes a_{jk}, is the condition $\sum_i \lambda_i^2 = 1$ satisfied?

Problem 5.12

If ABC is a composite system composed of three subsystems, A, B, and C, are there pure states $|\psi\rangle$ of ABC that cannot be written in the form $|\psi\rangle = \sum_i \lambda_i\, |\xi_{i,A}\rangle|\varphi_{i,B}\rangle|\chi_{i,C}\rangle$, where the λ_i are nonnegative, $\sum_i \lambda_i^2 = 1$ and the basis states are orthonormal? That is, can we always write $|\psi\rangle_{ABC} = \sum_i \sqrt{p_i}\, |\xi_{i,A}\rangle|\varphi_{i,B}\rangle|\chi_{i,C}\rangle$ with orthogonal basis functions?

Answer: Schmidt decomposition of tripartite pure states is not possible in general. If it was possible, the partial trace of the density matrix ρ_{ABC} over A and B would have the same "spectrum" of probabilities as would the trace of ρ_{ABC} over C (this is called a strong "equal spectrum" condition of the partial traces). Now, consider the state $\frac{1}{2}(|00\rangle_{AB} + |11\rangle_{AB})|0\rangle_C$, and show that this "equal spectrum" condition is not satisfied.

Schmidt Decomposition and Entanglement Entropy: The Schmidt decomposition theorem can be used to determine the entanglement entropy of an entangled pure state [defined in Eq. (5.29)],

$$\rho_{AB} = |\psi\rangle\langle\psi| = \sum_{i=1}^{\mathcal{N}_S} \lambda_i\, |\xi_{i,A}\rangle|\varphi_{i,B}\rangle \sum_{j=1}^{\mathcal{N}_S} \lambda_j\, \langle\xi_{j,A}|\langle\varphi_{j,B}|. \tag{5.34}$$

Taking the partial trace of ρ_{AB} over B yields,

$$\rho_A = \mathrm{Tr}_B[\rho_{AB}] = \sum_{i=1}^{\mathcal{N}_S} \lambda_i^2\, |\xi_{i,A}\rangle\langle\xi_{i,A}|, \tag{5.35}$$

which is a mixed state of subsystem A. Similarly, taking partial trace of ρ_{AB} over A yields,

$$\rho_B = \mathrm{Tr}_A[\rho_{AB}] = \sum_{i=1}^{\mathcal{N}_S} \lambda_i^2\, |\varphi_{i,B}\rangle\langle\varphi_{i,B}|, \tag{5.36}$$

which is a mixed state of subsystem B. We conclude that there is spectral equality of ρ_A and ρ_B, i.e., since $\{|\xi_{i,A}\rangle \in \mathcal{H}_A\}$ and $\{|\varphi_{i,B}\rangle \in \mathcal{H}_B\}$ are orthonormal sets, then $\{\lambda_i^2\}$, $(i = 1, 2, \ldots \mathcal{N}_S)$ are the nonzero eigenvalues of both ρ_A acting in $\mathcal{H}_A$ and ρ_B acting in $\mathcal{H}_B$. Thus,

If $|\psi\rangle \in \mathcal{H}_A \otimes \mathcal{H}_B$ and $\rho_{AB} = |\psi\rangle\langle\psi|$ is *the pure density operator constructed from* $|\psi\rangle$, then $\rho_A = \mathrm{Tr}_B[\rho_{AB}]$ and $\rho_B = \mathrm{Tr}_A[\rho_{AB}]$ have the same spectrum, $\{\lambda_i^2\}$, $(i = 1, 2, \ldots \mathcal{N}_S)$.

Since λ_i^2 is the probability for observing the states $|\xi_{iA}\rangle\langle\xi_{iA}|$ and $|\varphi_{iB}\rangle\langle\varphi_{iB}|$,

$$\mathcal{S}_A = \mathcal{S}_B = -\sum_{i=1}^{\mathcal{N}_S} \lambda_i^2 \log_2 \lambda_i^2, \tag{5.37}$$

i.e., the von Neumann entropies of the subsystems of a pure state are equal. These range between 0 (for a pure state, $\mathcal{N}_S = 1$) and $\log_2 \mathcal{N}_S$. Unless only one of the λ_i is nonzero, i.e., the Schmidt number is unity, the purity of the subsystems is not unity, and their von Neumann entropy are non-zero. Clearly, this "*entanglement entropy*" is not the entropy we consider in statistical mechanics and thermodynamics; here, we start with a system in a pure state, divide the system into two subsystems, and find that the entropy of the whole system is not the sum of the entropies of the subsystems; this von Neumann entropy is not extensive; hence, it is not always identical to the thermodynamic entropy.

Problem 5.13

(a) Take the partial trace over subsystem B of the pure state density matrix $\rho = |\psi\rangle\langle\psi|$ with $|\psi\rangle = \sum_{j,k} a_{jk}|j\rangle|k\rangle$ and show that $\rho_A = \mathrm{Tr}_B\ \rho = aa^\dagger$.
(b) Find ρ_A for the Bell state $|\Phi^+\rangle$.
(c) Calculate the entropy $\mathcal{S}_A$.

Answers: (a) $\rho_A = \mathrm{Tr}_B\ \rho = \sum_k a_{jk}a^*_{kj}$. (b) For the Bell state $|\Phi^+\rangle$, $a = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{\sqrt{2}} \end{pmatrix}$, so $\rho_A = aa^\dagger = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$.

(c) $\mathcal{S}_A = 1$.

The Schmidt coefficients, λ_i, can be used to distinguish separable pure states from entangled pure states. A separable pure state is characterized by a vector of Schmidt coefficients with only one nonvanishing entry, whereas the Schmidt vector of an entangled state has at least two nonvanishing components. A pure state has maximal entanglement entropy if its Schmidt coefficient vector is

$$\{\lambda_i\} = \{1/\sqrt{\mathcal{N}_S}, \ldots, 1/\sqrt{\mathcal{N}_S}\}, \tag{5.38}$$

where $\mathcal{N}_S$ is the Schmidt number. The Schmidt coefficients relate the degree of entanglement of pure bipartite states to the von Neumann entropy of the corresponding reduced density matrices of the two subsystems composing the system—a pure reduced density matrix with vanishing von Neumann entropy corresponds to a separable state, whereas a maximally entangled state leads to the maximum von Neumann entropy density matrix. It is easy to verify that Eq. (5.38) maximizes the entropy of the subsystems, $\mathcal{S}(\{\lambda_i\}) = -\sum_{i=1}^{\mathcal{N}_S} \lambda_i^2 \log_2 \lambda_i^2$, by maximizing the von Neumann entropy $\mathcal{S}(\{\lambda_i\})$ with respect to the λ_i under the density matrix trace constraint condition $\mathcal{G}(\{\lambda_i^2\}) \equiv \sum_{i=1}^{\mathcal{N}_S} \lambda_i^2 = 1$ by forming the function $\mathcal{F}(\{\lambda_i\}) = \mathcal{S} + \gamma\mathcal{G}$, where γ is a Lagrange multiplier, setting $\frac{\partial \mathcal{F}}{\partial \lambda_i} = 0$ for all i, and determining γ, such that the constraint condition is satisfied [25].

Mixed-State Entanglement

Not only pure states but also mixed states of a quantum system consisting of two or more subsystems can be entangled. Consider two quantum systems A and B with corresponding density matrices ρ_A and ρ_B, and let ρ_{AB} be the density matrix of the composite system. Is $\rho_{AB} = \rho_A \otimes \rho_B$? The answer is affirmative only if there is no correlation between the two systems. (Correlation between systems will be discussed in Sec. 6.1.8.) The nature of the correlations between systems is related to the notion of *separability*. Two systems A and B are said to be separable if the density matrix ρ_{AB} can be written as a convex sum of product states,

$$\rho_{AB} = \sum_{i=1}^{M} p_i\, \rho_A^i \otimes \rho_B^i, \quad p_i > 0, \quad \sum_{i=1}^{M} p_i = 1, \tag{5.39}$$

where ρ_A^i and ρ_B^i $(i = 1, 2, \ldots, M)$ are proper density matrices (hermiticity, unit trace, and semipositivity). A mixed state was defined by Werner [58] to be entangled if it is *inseparable*, i.e., if it *cannot* be written in the form (5.39).

1.3. Integrals involving vectors

Circulation of a vector field

The circulation of a vector field $\mathbf{E}$ in a displacement $d\mathbf{r} = dx\,\mathbf{e}_x + dy\,\mathbf{e}_y + dz\,\mathbf{e}_z$ is $\mathbf{E}.d\mathbf{r}$ (Figure 1.2a). The work of a force is a typical example of circulation. The circulation of $\mathbf{E}$ along a curve $\mathcal{C}$ going from $\mathbf{r}$ to $\mathbf{r}_o$ is the *line integral*

$$C = \int_{\mathcal{C}} d\mathbf{r}.\mathbf{E} = \int_{\mathbf{r}}^{\mathbf{r}_o} d\mathbf{r}.\mathbf{E} = \int_{\mathbf{r}}^{\mathbf{r}_o} dr\, E_{//} = \int_{\mathbf{r}}^{\mathbf{r}_o} [dx\, E_1 + dy\, E_2 + dz\, E_3]\,, \qquad [1.24]$$

where $d\mathbf{r}$ is the infinitesimal displacement along the path $\mathcal{C}$ and $E_{//}$ is the tangential component of $\mathbf{E}$. The circulation is a scalar quantity defined as the limit of the sum of the scalar products $d\mathbf{r}_n.\mathbf{E}_n$ of the infinitesimal elements $d\mathbf{r}_n$ of $\mathcal{C}$ and the fields $\mathbf{E}_n$ at these elements. To calculate the integral in the general case, a parametric representation of the curve $x = x(u)$, $y = y(u)$ and $z = z(u)$ may be used, where u is any parameter with u and u_o corresponding to the extreme positions $\mathbf{r}$ and $\mathbf{r}_o$. The components E_α become functions of u and the circulation becomes an integral over u

$$C = \int_{\mathcal{C}} d\mathbf{r}.\mathbf{E} = \int_{u}^{u_o} du\ [\ (dx/du)\, E_1(u) + (dy/du)\, E_2(u) + (dz/du)\, E_3(u)\ ]. \qquad [1.25]$$

If the field has a uniform tangential component $E_{//} = E$ along the path $\mathcal{C}$, its circulation is $C = EL_{\mathcal{C}}$, where $L_{\mathcal{C}}$ is the length of the path. On the other hand, if $\mathbf{E} = E\,\mathbf{e}_z$ is uniform in the direction Oz, its circulation is $\int_{\mathcal{C}} d\mathbf{r}.\mathbf{E} = \int_{z}^{z_o} dz\ E = (z_o - z)E$.

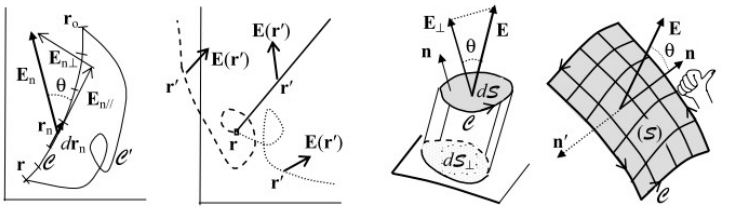

(a) (b) (c) (d)

Figure 1.2. *a) Circulation of* $\mathbf{E}$ *along a path* $\mathcal{C}$ *going from* $\mathbf{r}$ *to* $\mathbf{r}_o$. If $\mathbf{E}$ *is conservative, this circulation is equal to* $V(\mathbf{r}) - V(\mathbf{r}_o)$ for any $\mathcal{C}$. *b) Setting* $V(\infty) = 0$, $V(\mathbf{r})$ *is the circulation of* $\mathbf{E}$ *along an arbitrary path going from* $\mathbf{r}$ *to infinity. c) The flux of* $\mathbf{E}$ *through an infinitesimal surface* $d\mathcal{S}$. *d) The flux through an open surface* $\mathcal{S}$ *bounded by an oriented contour* $\mathcal{C}$

Flux of a vector field

Consider the integral over a surface $\mathcal{S}$

$$\Phi = \iint_{\mathcal{S}} d\mathcal{S}\,\mathbf{n}.\mathbf{E}(\mathbf{r}) = \iint_{\mathcal{S}} d\mathcal{S}\,E(\mathbf{r})\cos\theta = \iint_{\mathcal{S}} d\mathcal{S}\,E_\perp(\mathbf{r}), \qquad [1.26]$$

where $\mathbf{E}(\mathbf{r})$ is a vector field, $\mathbf{n}$ is the unit vector normal to the surface $\mathcal{S}$ at the running point $\mathbf{r}$, θ is the angle of $\mathbf{E}$ with $\mathbf{n}$ and $E_\perp$ is the component of $\mathbf{E}$ in the direction of $\mathbf{n}$. This integral is the *flux* of $\mathbf{E}$ through $\mathcal{S}$. The flux $d\Phi = d\mathcal{S}\,\mathbf{n}.\mathbf{E}(\mathbf{r})$ through the infinitesimal area $d\mathcal{S}$ is a scalar quantity and so is the flux. Note that we may write $d\Phi = d\mathcal{S}\,E_\perp(\mathbf{r})$, where $E_\perp(\mathbf{r})$ is the normal component of $\mathbf{E}$, or $d\Phi = d\mathcal{S}_\perp\,E(\mathbf{r})$, where $d\mathcal{S}_\perp$ is the projection of $d\mathcal{S}$ on the normal plane to $\mathbf{E}$ (Figure 1.2c). $d\Phi$ is positive or negative, depending on whether θ is acute or obtuse, and it vanishes if $\mathbf{E}$ is tangent to $\mathcal{S}$. Note also that $\mathbf{n}$ has two possible orientations; by changing the direction of $\mathbf{n}$, we change the sign of Φ. In the case of an open surface, which is bounded by an oriented closed curve $\mathcal{C}$, we choose $\mathbf{n}$ according to the right-hand rule (Figure 1.2d). In the case of a closed surface $\mathcal{S}$, we choose $\mathbf{n}$ oriented outward; Φ is then the outgoing flux.

The flux is additive both for the vector field and for the area. In the particular case of a field having a uniform component in the direction of $\mathbf{n}$, its flux is $\Phi = B_\perp \mathcal{S}$. Another physically interesting case is that of a radial field $\mathbf{E} = Kq\mathbf{r}/r^3$ of a charge q. Its flux through a closed surface $\mathcal{S}$ is $\Phi = Kq \iint_{\mathcal{S}} d\mathcal{S}\,\mathbf{n}.\mathbf{r}/r^3 = Kq\Omega$, where Ω is the solid angle of the cone, whose apex is at q and which is subtended by $\mathcal{S}$: it is equal to 4π if q is inside $\mathcal{S}$ and equal to 0 if q is outside $\mathcal{S}$.

1.4. Gradient and curl, conservative field and scalar potential

The work of a force $\mathbf{F}$ acting on a particle of mass m in a displacement $d\mathbf{r}$ is $dW = \mathbf{F}.d\mathbf{r}$. This work is transformed into kinetic energy if no other force acts on the particle. Conversely, to displace the particle without acquiring kinetic energy dU_K, an external agent must exert a force $\mathbf{F}' = -\mathbf{F}$ and supply a work $dW' = -\mathbf{F}.d\mathbf{r}$. If the force is *conservative*, this work is transformed into potential energy dU_P of the particle in the field of force $\mathbf{F}$. This analysis can be repeated for any vector field $\mathbf{E}$. Its circulation along a path $\mathcal{C}$ going from $\mathbf{r}$ to $\mathbf{r}_o$ depends in general on $\mathbf{r}$ and $\mathbf{r}_o$ and also on the path $\mathcal{C}$. Its circulation on a closed path is not necessarily equal to zero. The differential form $dx\,E_1 + dy\,E_2 + dz\,E_3$ is a total differential if the components E_α are the partial derivatives of a scalar function $-V$ where V is called the *scalar potential* corresponding to the field of force $\mathbf{F}$. Then, we have $E_1 = -\partial_1 V$, $E_2 = -\partial_2 V$, and $E_3 = -\partial_3 V$, which we write in the vector form

) $&

20210825/ Electromagnetism
Maxwell.../
Electromagnetism, Maxwell
Equation... pdf

Pag. 7(27)

1.3. Integrals involving vectors

Circulation of a vector field

Fig 1.2.

20210825/ Fermat/
Methods of Geometry... pdf

Pag. 159(178)

Theorem 3 (Angle Bisectors theorem)

The angle bisectors of ΔABC meet at an interior I equidistant from the edge lines

Fig. 5.1.2/ gu

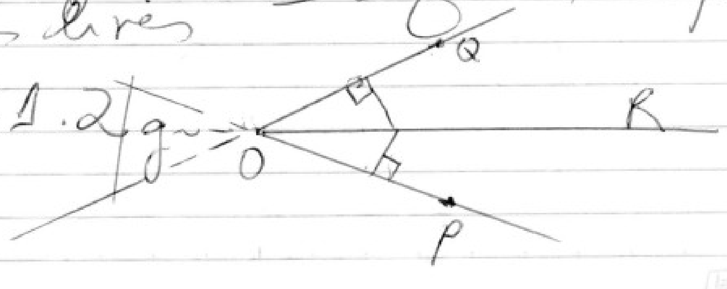

+,- * *. ! * $ # / 0 656 7 1#
) '$

trigonometry in section 5.5 in many contexts. The material at the end of the chapter is aimed straight at a wonderful result, Feuerbach's theorem about five circles related to a triangle. This tour de force shows the depth you can achieve with these methods.

En route you'll encounter more strands of advanced Euclidean geometry that you could follow to many fascinating parts of the subject. To limit the excursions, these strands are generally abandoned quickly. The exercises in section 5.11 will pursue some of them. Return to others as you please in later years, and follow them into the vast literature of this fascinating part of geometry!

5.1 Four concurrence theorems

Concepts

- Standard triangle terminology and notation
- Edge bisectors of ΔABC
- Circumcircle, circumcenter O, and circumradius R
- Angle bisectors
- Incircle, incenter I, and inradius r
- Medians and centroid G
- Altitudes and orthocenter H

The theorems in this section often constitute the most intricate part of an elementary geometry text. They're related to almost all advanced work in triangle and circle geometry.

For triangle ΔABC as in figure 5.1.1, let A', B', and C' denote the midpoints of the edges a, b, and c opposite vertices A, B, and C. That notation is used throughout this chapter. The perpendicular bisectors of a and b meet at a point O; if they coincided or were parallel, the transitivity theorem for parallel lines and the alternate interior angles theorem would imply $a \parallel b$. By the perpendicular bisector theorem, O is equidistant from

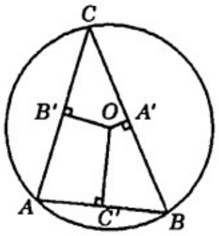

Figure 5.1.1 ΔABC, its circumcenter O, and its circumcircle

B and C and from C and A, hence from A and B, so it lies on the perpendicular bisector of c. This proves

***Theorem 1* (*Edge bisectors theorem*).** The perpendicular bisectors of the edges of ΔABC meet at a point O equidistant from its vertices.

O is called the *circumcenter* of ΔABC. The circle with center O through A, B, and C is the *circumcircle* of ΔABC. Its radius R is the *circumradius.*

Theorem 2. Any three noncollinear points lie on a unique circle.

Proof. Theorem 1 shows that the circle exists. To prove its uniqueness, note that the center of any circle through the vertices of a triangle is equidistant from them, hence must lie on the edge bisectors, so must coincide with the circumcenter. ♦

The *bisector* of angle $\angle POQ$ is the line g through its vertex O such that, if R is a point on g inside $\angle POQ$, then $\angle POR \cong \angle ROQ$. An argument suggested by figure 5.1.2 shows that the bisector consists of O and all points equidistant from lines $\overleftrightarrow{OP}$ and $\overleftrightarrow{OQ}$ that are interior to $\angle POQ$ or its vertical counterpart.

Now consider the bisectors of the angles of ΔABC, as in figure 5.1.3. Since the ends of edge a lie on different rays of $\angle A$, that angle's bisector meets a at some point X. For the same reason, the bisector of $\angle C$ meets segment $\overline{AX}$. The bisectors' intersection I is equidistant from edge lines $\overleftrightarrow{AB}$ and $\overleftrightarrow{CA}$ and from $\overleftrightarrow{CA}$ and $\overleftrightarrow{BC}$, hence from $\overleftrightarrow{BC}$ and $\overleftrightarrow{AB}$, so it lies on the bisector of $\angle B$. This proves

***Theorem 3* (*Angle bisectors theorem*).** The angle bisectors of ΔABC meet at an interior point I equidistant from the edge lines.

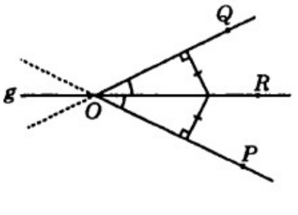

Figure 5.1.2 Bisector of $\angle POQ$

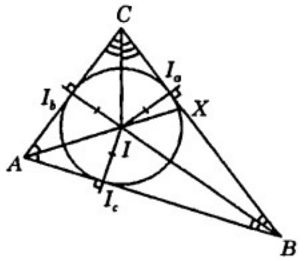

Figure 5.1.3 ΔABC, its incenter I, and its incircle

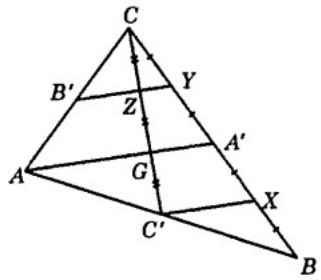

Figure 5.1.4 Proof of lemma 4: ΔABC, two medians, and centroid G

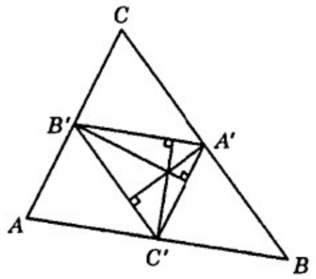

Figure 5.1.5 ΔABC and its medial triangle $\Delta A'B'C'$

I is called the *incenter* of ΔABC. The common distance r from I to the edge lines is the triangle's *inradius*. The circle with center I and radius r is the *incircle*. It's tangent to the edge lines at points I_a, I_b, and I_c within edges a, b, and c. You can show easily that the incircle is the only circle tangent to all three lines at points interior to the edges. Three more circles are tangent to these lines, however, at points outside the edges. They're studied in detail in section 5.9.

Segments $\overline{AA'}$, $\overline{BB'}$, and $\overline{CC'}$ in ΔABC are called its *medians*.

Lemma 4. $\overline{AA'}$ and $\overline{CC'}$ meet at a point G two thirds of the way from C to C'.

Proof. As shown in figure 5.1.4, C and C' lie on different legs of $\angle A$ and $\overrightarrow{AA'}$ lies in the interior of $\angle A$, so $\overline{AA'}$ and $\overline{CC'}$ meet. Let X and Y be the midpoints of $\overline{BA'}$ and $\overline{CA'}$. By the same argument, $\overline{YB'}$ and $\overline{CC'}$ meet at a point Z. Then $\overleftrightarrow{YB'} \parallel \overleftrightarrow{AA'}$ since Y and B' are edge midpoints of $\Delta AA'C$. Similarly, $\overleftrightarrow{AA'} \parallel \overleftrightarrow{XC'}$, hence $CZ = ZG = GC'$. ♦

Theorem 5 (Medians theorem). The medians of ΔABC meet at a point G two thirds of the way from any vertex to the midpoint of the opposite side.

Proof. Apply lemma 4 to ΔBAC: $\overline{BB'}$ and $\overline{CC'}$ meet at the same point G as $\overline{AA'}$ and $\overline{CC'}$. ♦

G is called the *centroid* of ΔABC. $\Delta A'B'C'$ is its *medial* triangle. The following properties of these triangles, shown in figure 5.1.5, are easy to verify.

Theorem 6. The edges of the medial triangle are parallel to those of ΔABC and half their length. Also,

$$\Delta ABC \sim \Delta A'B'C' \cong \Delta AB'C' \cong \Delta A'BC' \cong \Delta A'B'C.$$

The medians of ΔABC contain those of the medial triangle, hence these triangles have the same centroid. The edge bisectors of ΔABC contain the altitudes of the medial triangle.

The feet of the altitudes of ΔABC through vertices A, B, and C are denoted by D, E, and F. From a given ΔABC it's simple to construct a new triangle whose medial triangle is ΔABC. Since the altitudes of ΔABC lie in the edge bisectors of the new triangle, the altitude lines are concurrent. This proves

***Corollary 7* (*Altitudes theorem*).** The altitude lines of triangle ΔABC meet at a point H.

H is called the *orthocenter*[1] of ΔABC. It's shown in figure 5.1.6.

***Corollary** 8.* The circumcenter of a triangle is the orthocenter of its medial triangle.

A set of lines that pass through a single point is called *concurrent*.[2] Euclid described the concurrence of the edge bisectors and angle bisectors of a triangle.[3] He didn't mention concurrence of altitudes or medians. But the altitudes' concurrence is so closely related to that of the angle bisectors that it must have been familiar then. The medians' concurrence is included as a very elementary result in Archimedes' treatise "On plane equilibria."[4] The terminology *circumcircle*, *inradius*, etc., was introduced in the late

Figure 5.1.6 ΔABC and its orthocenter H

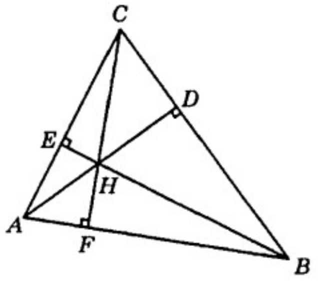

---

[1] The prefix *ortho-* occurs in many areas of mathematics. It stems from a Greek word for *upright*. Here that refers to the altitudes.

[2] From the Latin verb *curro* and prefix *con-*, which mean *run* and *together*.

[3] Euclid [1908] 1956, book IV, propositions 4 and 5.

[4] Archimedes [1912] n.d., book I, proposition 14.

nineteenth century.[5] At that time the following notation for points and distances related to ΔABC was standardized:

$A\ B\ C$ .... vertices,
$a\ b\ c$ ..... opposite edges,
$A'\ B'\ C'$ ... midpoints of opposite edges,
$D\ E\ F$ .... feet of corresponding altitudes,
$I_a\ I_b\ I_c$ .... opposite incircle tangency points,
$G\ H$ ...... centroid, orthocenter,
$O\ I$ .... circumcenter, incenter,
$R\ r$ .... circumradius, inradius.

Further standard notation is introduced later in this chapter.

5.2 Menelaus' theorem

Concepts

Directed distances
Menelaus' product
Menelaus' theorem

Menelaus' theorem is included in this chapter for two reasons. In the next section, it leads to Desargues' theorem, which was used in section 1.1 to settle a question in perspective drawing. In section 5.4, Menelaus yields Ceva's theorem, a generalization of three of the four concurrence theorems that started this chapter.

Menelaus' and several related theorems involve both the order of points on various lines and some ratios of distances between them. To consider these simultaneously, it's convenient to use coordinate geometry on one line g at a time. Give g a scale c. In place of the distance $PQ = |c(P) - c(Q)|$ between points P and Q on g consider their *directed distance*

$$P \text{ to } Q = c(Q) - c(P),$$

so that

$$P \text{ to } Q = -(Q \text{ to } P) \qquad PQ = |P \text{ to } Q|.$$

Although the scale c is suppressed in this notation, you need it to compute the directed distance. Most authors use the same symbol PQ for directed and undirected distances; you must rely on the context to distinguish the concepts. However, directed distances see only limited use in this text, and

---

[5] See W. H. H. H. 1883.

.2 ! ?

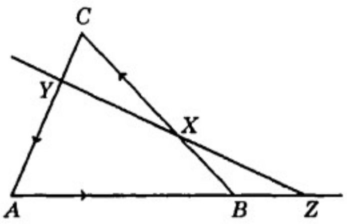

Figure 5.2.1
Menelaus' theorem

it seems worthwhile to make them visibly different. In a diagram, you can indicate the use of an unnamed scale for a line g by drawing an arrow from some point on g to a point with a larger coordinate, like this: —>— .

Directed distances usually appear by twos, in ratios. That explains why the scale is suppressed from the notation: For any points P, Q, R, and S on g with $R \neq S$, the ratio $(P \text{ to } Q)/(R \text{ to } S)$ doesn't depend on the scale. That is, if d is another scale for g, then

$$\frac{P \text{ to } Q}{R \text{ to } S} = \frac{c(Q) - c(P)}{c(S) - c(R)} = \frac{d(Q) - d(P)}{d(S) - d(R)} .$$

It's easy to prove that; see exercise 4.1.7. Signs of ratios like this are often used to convey information about the order of points on g. For example, if P, Q, and R lie on g and $Q \neq R$, then

$$\frac{P \text{ to } Q}{Q \text{ to } R} \geq 0 \leftrightarrow P\text{-}Q\text{-}R.$$

Menelaus' theorem involves a product of ratios of directed distances between points on the edge lines of a triangle ΔABC. Assign scales to lines $\overleftrightarrow{BC}$, $\overleftrightarrow{CA}$, and $\overleftrightarrow{AB}$, as indicated by the arrows in figure 5.2.1. Then select points X to Z on these lines but distinct from the vertices. The product

$$\frac{A \text{ to } Z}{Z \text{ to } B} \, \frac{B \text{ to } X}{X \text{ to } C} \, \frac{C \text{ to } Y}{Y \text{ to } A}$$

of ratios of two directed distances on each line is called *Menelaus' product*. It doesn't depend on the choice of scales. Its sign gives information about how many of the points X, Y, and Z may lie between the vertices. The product is positive if all three lie within the edges or just one does, but negative if just two lie within the edges.

***Theorem 1* (*Menelaus' theorem*).** X, Y, and Z are collinear if and only if Menelaus' product equals -1.

Proof. Suppose X, Y, and Z lie on a line g, as in figure 5.2.2. Find a line $h \neq \overleftrightarrow{AB}, \overleftrightarrow{CA}$ through A not parallel to $\overleftrightarrow{BC}$ nor g. Construct lines j and k parallel to h through B and C. Then h, j, and k intersect

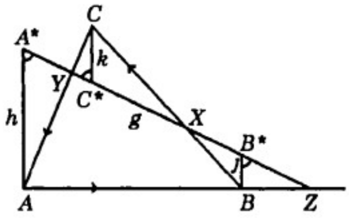

Figure 5.2.2
Proof of Menelaus' theorem

g at points A^*, B^*, and C^* that form three pairs of similar triangles with associated ratios as follows:

$$\Delta AZA^* \sim \Delta BZB^* \qquad \Delta BXB^* \sim \Delta CXC^* \qquad \Delta CYC^* \sim \Delta AYA^*$$

$$\frac{A \text{ to } Z}{Z \text{ to } B} = -\frac{AA^*}{BB^*} \qquad \frac{B \text{ to } X}{X \text{ to } C} = \frac{BB^*}{CC^*} \qquad \frac{C \text{ to } Y}{Y \text{ to } A} = \frac{CC^*}{AA^*}.$$

These equations imply

$$\frac{A \text{ to } Z}{Z \text{ to } B}\,\frac{B \text{ to } X}{X \text{ to } C}\,\frac{C \text{ to } Y}{Y \text{ to } A} = -1.$$

This computation depends on the positions of X, Y, and Z. Since they're collinear, they can't all lie within the edges, nor can just one of them. Thus, just one or all of them must lie outside ΔABC. Figure 5.2.2 shows the case when just one—namely, Z —lies outside. If all three lay outside, then all three ratios of directed distances would be negative, and their product would still equal -1.

Conversely, suppose Menelaus' product equals -1. Then all three or just one of X, Y, and Z must lie outside ΔABC. You can assume that X and Y both lie within the edges, as in figure 5.2.2, or both lie outside. Let $g = \overleftrightarrow{XY}$. If g / c, then

$$\frac{B \text{ to } X}{X \text{ to } C} = \frac{Y \text{ to } A}{C \text{ to } Y}$$

Mathematics flourished at Alexandria for many centuries after Euclid, waning gradually with the rise of Christianity and its antipathy toward Greek culture. MENELAUS worked there around A.D. 100. Principally an astronomer, he studied the precession of the equinoxes and other phenomena. To further that work he founded the subject of spherical trigonometry with a book, *Sphaerica*. Menelaus authored several other texts on geometry and astronomy as well. This section's theorem carrying his name is related to his work in spherical trigonometry, but here it's used as a key to some important theorems in plane geometry.

hence

$$\frac{A \text{ to } Z}{Z \text{ to } B} = -1,$$

and it would follow that A to $Z = B$ to Z, hence $A = B$ —contradiction! Therefore, g must intersect c at a point $Z' \neq A, B$. By the previous paragraph,

$$\frac{A \text{ to } Z'}{Z' \text{ to } B}\,\frac{B \text{ to } X}{X \text{ to } C}\,\frac{C \text{ to } Y}{Y \text{ to } A} = -1,$$

hence

$$\frac{A \text{ to } Z'}{Z' \text{ to } B} = \frac{A \text{ to } Z}{Z \text{ to } B}.$$

This equation clearly implies $Z = Z'$, so Z also falls on g. ♦

5.3 Desargues' theorem

Concepts

Pencils of lines
Euclidean forms of Desargues' theorem
Does Desargues' theorem require metric notions?

Section 1.1 used a form of Desargues' theorem to solve a fundamental problem in descriptive geometry. Desargues' theorem underlies various fundamental techniques in that field; in fact, it was originally discovered in that context.[6] There are several forms of the theorem in Euclidean geometry. They can all be interpreted as specific instances of a single result in projective geometry, the theory that was developed to serve as a basis for the techniques of perspective drawing. That unified form of the theorem plays a central role in the modern axiomatic development of projective geometry.

In this section, however, Desargues' theorem is considered only in the context of Euclidean geometry, hence some forms must be treated separately. Some of the distinctions have to do with situations where three lines are parallel or have a common intersection. To simplify the terminology, the families of all lines parallel to a given line and of all lines through a given point are called parallel and concurrent *pencils*.[7] A set of lines is called *copencilar* if they all belong to the same pencil. Several styles of proof are

---

[6] The original publication is available in translation: Desargues 1648.

[7] This usage stems from the Latin word *penicillus* for *brush*.

.2 6 ! ?

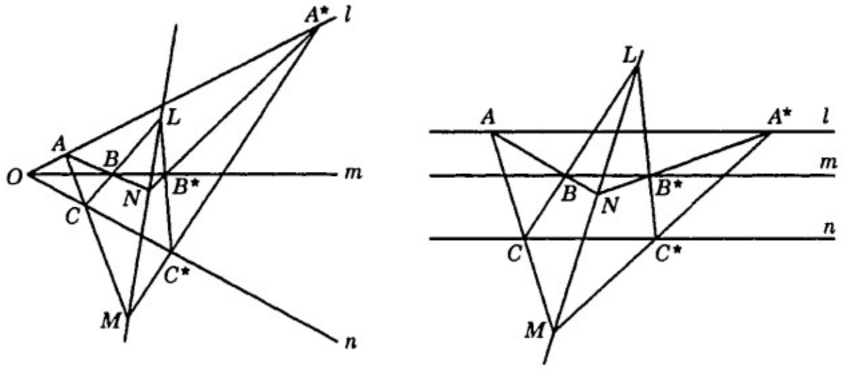

Concurrent case Parallel case

Figure 5.3.1 First form of Desargues' theorem

useful for various forms of Desargues' theorem. The first form considered here can be proved neatly from Menelaus' theorem.

Theorem 1 ***(First form of Desargues' theorem and its converse).*** Consider coplanar triangles ΔABC and $\Delta A^*B^*C^*$ as in figure 5.3.1, with three distinct lines l, m, and n and three distinct points L, M, and N such that

A and A^* lie on l but not on m nor n,
B and B^* lie on m but not on n nor l,
C and C^* lie on n but not on l nor m,

$$L = \overleftrightarrow{BC} \cap \overleftrightarrow{B^*C^*}, \quad M = \overleftrightarrow{CA} \cap \overleftrightarrow{C^*A^*}, \quad N = \overleftrightarrow{AB} \cap \overleftrightarrow{A^*B^*}.$$

Then l, m, and n are copencilar if and only if L, M, and N are collinear.

Proof. The hypotheses imply $A \neq A^*$, $B \neq B^*$, and $C \neq C^*$. In order to use directed distances, assign scales to all lines under consideration. Suppose l, m, and n all pass through a point O. Apply Menelaus' theorem first to the collinear points L, B^*, and C^* on the edge lines of ΔOBC to get

$$\frac{O \text{ to } B^*}{B^* \text{ to } B} \, \frac{B \text{ to } L}{L \text{ to } C} \, \frac{C \text{ to } C^*}{C^* \text{ to } O} = -1,$$

then to the collinear points M, C^*, and A^* on the edge lines of ΔOAC to get

$$\frac{C \text{ to } M}{M \text{ to } A} \, \frac{A \text{ to } A^*}{A^* \text{ to } O} \, \frac{O \text{ to } C^*}{C^* \text{ to } C} = -1,$$

Gérard DESARGUES was born in 1593 in Lyon, the son of a minor government official. Nothing is known of his education. By 1626 he'd moved to Paris and developed a reputation as a mathematician and engineer, particularly interested in problems arising in graphic arts and architecture. There he met Descartes. They became lifelong friends, and influenced each other's mathematical work. Richelieu appointed Desargues an army engineer officer in 1628. He assisted in the design of the fortifications and dikes at La Rochelle. During the late 1620s and 1630s, Desargues researched and gave mathematics lectures in Paris. His principal interest was the analysis and application of perspective techniques, and he pioneered the projective theory of conic sections. His work was published in the late 1630s and 1640s, but in scattered fashion, and sometimes only as appendices to his students' works. His style was novel and overly concise. For these reasons, and because of the great attention given to the 1637 publication of Descartes' elements of analytic geometry, Desargues' work received little note during his lifetime and in fact for the next two hundred years. His student Blaise Pascal, however, achieved fame with a treatise on conic sections based on Desargues' work. Desargues continued to practice engineering and architecture until he retired to his country estate near Lyon in 1650. He died there in 1662.

and to the collinear points N, A^*, and B^* on the edge lines of ΔOAB to get

$$\frac{O \text{ to } A^*}{A^* \text{ to } A}\,\frac{A \text{ to } N}{N \text{ to } B}\,\frac{B \text{ to } B^*}{B^* \text{ to } O} = -1.$$

Now multiply these three equations and simplify to get

$$\frac{A \text{ to } N}{N \text{ to } B}\,\frac{B \text{ to } L}{L \text{ to } C}\,\frac{C \text{ to } M}{M \text{ to } A} = -1.$$

By Menelaus' theorem, points N, L, and M on the edge lines of ΔABC are collinear. The proof that concurrence of l, m, and n implies collinearity of L, M, and N is complete.

Next, suppose that L, M, and N are collinear but l, m, and n aren't all parallel. Then two of these lines, say n and l, intersect at a point O. Collinearity of B^*, O, and B follows by the previous paragraph's argument with

A, B, C, A^*, B^*, C^*, L, M, N, and O replaced by
C, L, C^*, A, N, A^*, B^*, O, B, and M.

That is, O lies on m as well. Thus, collinearity of L, M, and N implies copencilarity of l, m, and n.

Finally, suppose that l, m, and n are parallel. Considering similar triangles, you get

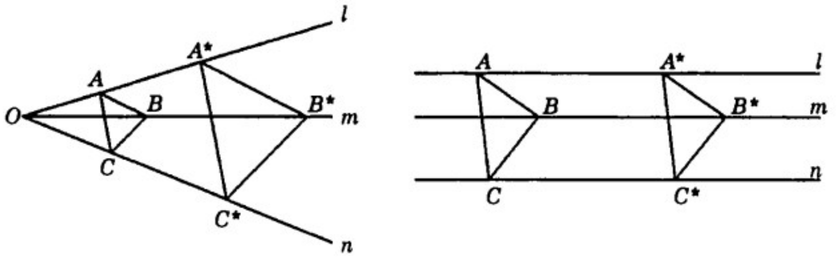

Concurrent case Parallel case

Figure 5.3.2 Second form of Desargues' theorem

$$\frac{A \text{ to } N}{N \text{ to } B}\,\frac{B \text{ to } L}{L \text{ to } C}\,\frac{C \text{ to } M}{M \text{ to } A} = \left[-\frac{A \text{ to } A^*}{B \text{ to } B^*}\right]\left[-\frac{B \text{ to } B^*}{C \text{ to } C^*}\right]\left[-\frac{C \text{ to } C^*}{A \text{ to } A^*}\right] = -1.$$

By Menelaus' theorem, points N, L, and M on the edge lines of ΔABC are collinear. ♦

In the second form of Desargues' theorem, ΔABC and $\Delta A^*B^*C^*$ have parallel edges. You can supply the proof by considering similar triangles.

***Theorem 2* (*Second form of Desargues' theorem and its converse*).** Consider coplanar triangles ΔABC and $\Delta A^*B^*C^*$ as in figure 5.3.2, with three distinct lines l, m, and n such that

A and A^* lie on l but not on m nor n,
B and B^* lie on m but not on n nor l,
C and C^* lie on n but not on l nor m,
$\overleftrightarrow{AB} \parallel \overleftrightarrow{A^*B^*}$, $\overleftrightarrow{BC} \parallel \overleftrightarrow{B^*C^*}$.

Then l, m, and n are copencilar if and only if $\overleftrightarrow{CA} \parallel \overleftrightarrow{C^*A^*}$.

The statements of Desargues' theorem involve only incidence notions (collinearity, concurrence, parallelism, etc.), but the proofs given and suggested above involve metric notions (directed distances, similar triangles, etc.). When the figures are three-dimensional, however, as in the following form of the theorem, you can avoid metric techniques.

***Theorem 3* (*Third form of Desargues' theorem and its converse*).** Consider noncoplanar triangles ΔABC and $\Delta A^*B^*C^*$ with three distinct lines l, m, and n and three distinct points L, M, and N such that

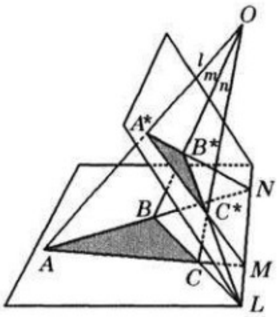

Figure 5.3.3 Third form of Desargues' theorem: noncoplanar case

A and A^* lie on l but not on m nor n,
B and B^* lie on m but not on n nor l,
C and C^* lie on n but not on l nor m,

$L = \overleftrightarrow{BC} \cap \overleftrightarrow{B^*C^*}$, $M = \overleftrightarrow{CA} \cap \overleftrightarrow{C^*A^*}$, $N = \overleftrightarrow{AB} \cap \overleftrightarrow{A^*B^*}$.

Then l, m, and n are copencilar if and only if L, M, and N are collinear.

Proof. If l, m, and n are copencilar, then L, M, and N are collinear because they all lie in the planes of both triangles. See figure 5.3.3 for the concurrent case; you may draw a figure for the parallel case. You may also supply the proof that collinearity of L, M, and N implies copencilarity of l, m, and n. ♦

In 1907, Alfred North Whitehead showed how to use a three-dimensional construction to prove the coplanar case of Desargues' theorem—theorem 1—without using metric concepts.[8] Suppose ΔABC and $\Delta A^*B^*C^*$ lie in the same plane ε, and l, m, and n intersect in a common point O, as in figure 5.3.4. (To enhance the figure's legibility, lines $\overleftrightarrow{CA}$ and $\overleftrightarrow{C^*A^*}$ were omitted.) There's no loss in assuming that A lies between O and A^* as in the figure—otherwise, just relabel the triangles. Select a point P not on ε and a point R between but different from O and P. The plane $\delta = LMR$ intersects $\overleftrightarrow{PA}$, $\overleftrightarrow{PB}$, and $\overleftrightarrow{PC}$ at points A'', B'', and C'' between but different from P and A, B, and C. Lines $\overleftrightarrow{A^*A''}$ and $\overleftrightarrow{OP}$ intersect at a point Q between but different from O and P. Several triples of points are collinear because they lie in two distinct planes:

B'', C'', L in δ and BCP
C'', A'', M in δ and CAP
C^*, C'', Q in $A^*A''M$ and C^*OP
B^*, B'', Q in $C^*C''L$ and B^*OP.

---

[8] Whitehead [1907] 1971, 16–17.

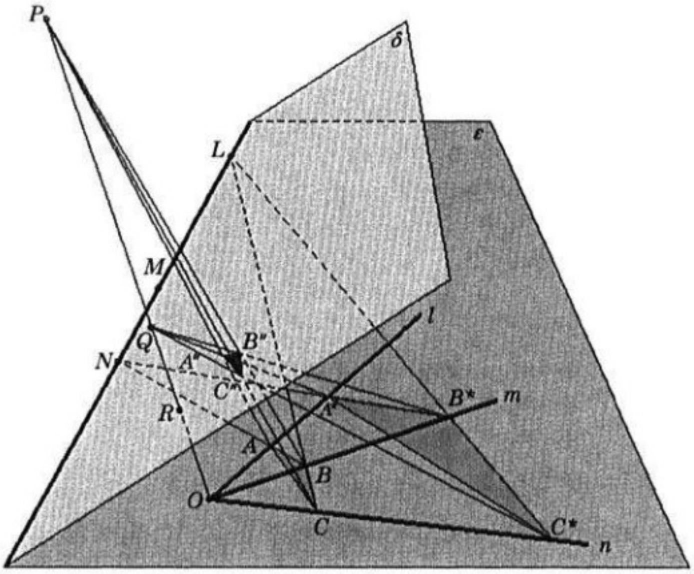

Figure 5.3.4
Third form of Desargues' theorem: coplanar case

By the proof of noncoplanar case of theorem 3, lines $\overleftrightarrow{AB}$ and $\overleftrightarrow{A''B''}$ must be parallel or intersect in a point on $\overleftrightarrow{LM}$, and lines $\overleftrightarrow{A^*B^*}$ and $\overleftrightarrow{A''B''}$ must be parallel or intersect in a point on $\overleftrightarrow{LM}$. One of $\overleftrightarrow{AB} \parallel \overleftrightarrow{A''B''}$ and $\overleftrightarrow{A^*B^*} \parallel \overleftrightarrow{A''B''}$ must fail because $\overleftrightarrow{AB}$ and $\overleftrightarrow{A^*B^*}$ aren't parallel. Assume $\overleftrightarrow{AB}$ and $\overleftrightarrow{A''B''}$ intersect at a point N^* on $\overleftrightarrow{LM}$; you can handle the other possibility similarly. Points A^*, B^*, and N^* are collinear because they all lie in the distinct planes δ and $A''B''Q$. This implies $N = N^*$, hence N lies in $\overleftrightarrow{LM}$. The proof that concurrence of l, m, and n implies collinearity of L, M, and N is complete. You may supply a similar nonmetric proof that L, M, and N are collinear if l, m, and n are parallel. The proof of theorem 1 included a nonmetric argument that in this planar situation, collinearity of L, M, and N implies copencilarity of l, m, and n.

Several years before Whitehead's work, David Hilbert (1899, section 23) had already shown that *every* proof of the plane Desargues theorem requires either metric concepts or a three-dimensional construction.

Alfred North WHITEHEAD was born in 1861 in Ramsgate. His father was a schoolteacher and Anglican clergyman. Alfred's brother became a bishop. In grammar school, Alfred excelled in sports and mathematics. In 1880, he entered Trinity College, Cambridge; all his courses there were in mathematics. Upon graduation he became a fellow of the college, and soon was appointed senior lecturer in mathematics. In 1890 he married Evelyn Wade; they had three children.

Whitehead's early mathematical work had to do with group theory and algebra. He published a book *Universal algebra* that developed in a roundabout way material that would later be included in the theories of linear and multilinear algebra. Around 1900 he began working more in foundations of mathematics, and published two small works on axiom systems for projective and descriptive geometry.

During this period at Cambridge he guided the work of Bertrand Russell as a student, including Russell's dissertation on foundations of geometry. (Whitehead's father had been the Russells' vicar, and in that capacity had been called on to convince the young boy that the earth was round!) Early in the new century they began collaboration on *Principia mathematica*, a monumental work that attempted to demonstrate the *logistic thesis*, that all mathematics can be derived from logical principles alone. A major achievement in mathematical logic, the book provided a framework for much significant research, although Gödel showed later that logicism was inadequate as a philosophy of mathematics. (No system like *Principia* can entail all true arithmetic statements but no false ones.)

By 1910, Whitehead's interests had turned almost entirely to philosophy of science, and he moved to the University of London. During his London period he produced three books on foundations of physics, especially relativity theory. Influenced by the tragedies of World War I, including the death of his son, an aviator, he turned even more to philosophical studies. He crossed the Atlantic in 1924 to become professor of philosophy at Harvard. There he wrote a series of works on philosophy of science, education, and religion, finally retiring in 1937 at age 76. Whitehead died in 1947.

5.4 Ceva's theorem

Concepts

Ceva's theorem
Deriving the medians, altitudes, and angle bisectors theorems
Proving Ceva's theorem by mechanics

A generalization of three of the four concurrence theorems in section 5.1 was found about nineteen centuries after them. It's phrased in terms of Menelaus' product:

.2 7!?

***Theorem 1* (*Ceva's theorem*).** In order to use directed distances, assign scales to the edge lines of ΔABC. If points X, Y, and Z lie on these lines and are distinct from the vertices, then

$$\frac{A \text{ to } Z}{Z \text{ to } B}\,\frac{B \text{ to } X}{X \text{ to } C}\,\frac{C \text{ to } Y}{Y \text{ to } A} = 1$$

if and only if lines $l = \overleftrightarrow{AX}$, $m = \overleftrightarrow{BY}$, and $n = \overleftrightarrow{CZ}$ are copencilar.

Proof. Each of these statements will be proved in turn:

i. if X, Y, and Z lie within the edges, then l, m, and n aren't all parallel; if they concur, then Menelaus' product equals 1;

ii. if X, Y, and Z lie within the edges and Menelaus' product equals 1, then l, m, and n concur;

iii. if X and Y lie outside the edges and Z within, and l, m, and n concur, then Menelaus' product equals 1;

iv. if X and Y lie outside the edges and Z within, and l, m, and n are parallel, then Menelaus' product equals 1;

v. if X and Y lie outside the edges and Z within, and Menelaus' product equals 1, then l, m, and n are copencilar;

vi. if just one or all three of X, Y, and Z lie within the edges, then Menelaus' product is negative and l, m, and n aren't copencilar.

Parts (*iii*) to (*v*) cover all cases where exactly one of X, Y, and Z lies within an edge of ΔABC; if X or Y instead of Z lies inside, just relabel the triangle.

Part (*i*). Suppose X, Y, and Z lie within the edges, as in figure 5.4.1. You can easily see that l, m, and n can't be all parallel. Suppose they meet at a point W. Then W lies inside the triangle. Apply Menelaus' theorem to the collinear points Z, C, and W on the edge lines of ΔABX:

$$\frac{A \text{ to } Z}{Z \text{ to } B}\,\frac{B \text{ to } C}{C \text{ to } X}\,\frac{X \text{ to } W}{W \text{ to } A} = -1.$$

Apply it again to the collinear points W, B, and Y on the edge lines of ΔAXC:

$$\frac{A \text{ to } W}{W \text{ to } X}\,\frac{X \text{ to } B}{B \text{ to } C}\,\frac{C \text{ to } Y}{Y \text{ to } A} = -1.$$

Multiply these equations to get

$$\frac{A \text{ to } Z}{Z \text{ to } B}\,\frac{B \text{ to } X}{X \text{ to } C}\,\frac{C \text{ to } Y}{Y \text{ to } A} =$$

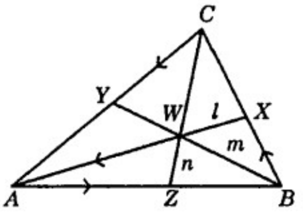

Figure 5.4.1 Ceva's theorem, parts (*i*) and (*ii*)

$$\left[\frac{A\text{ to }Z}{Z\text{ to }B}\,\frac{B\text{ to }C}{C\text{ to }X}\,\frac{X\text{ to }W}{W\text{ to }A}\right]\left[\frac{A\text{ to }W}{W\text{ to }X}\,\frac{X\text{ to }B}{B\text{ to }C}\,\frac{C\text{ to }Y}{Y\text{ to }A}\right]=(-1)(-1)=1.$$

Part (*ii*). Suppose X, Y, and Z lie within the edges, as in figure 5.4.1. Then lines l and m meet at a point W inside the triangle, and line $\overleftrightarrow{CW}$ meets edge $\overline{AB}$ at a point Z'. By part (*i*),

$$\frac{A\text{ to }Z'}{Z'\text{ to }B}\,\frac{B\text{ to }X}{X\text{ to }C}\,\frac{C\text{ to }Y}{Y\text{ to }A}=1.$$

If Menelaus' product equals 1, i.e.

$$\frac{A\text{ to }Z}{Z\text{ to }B}\,\frac{B\text{ to }X}{X\text{ to }C}\,\frac{C\text{ to }Y}{Y\text{ to }A}=1,$$

then

$$\frac{A\text{ to }Z}{Z\text{ to }B}=\frac{A\text{ to }Z'}{Z'\text{ to }B}.$$

This implies $Z=Z'$, so W lies on n as well as l and m.

Part (*iii*). Suppose X and Y lie outside the edges and Z within, and l, m, and n meet at a point W, as in figure 5.4.2. Then W lies outside the triangle. The proof for part (*i*) is valid in this case as well.

Part (*iv*). Suppose X and Y lie outside the edges and Z within, and l, m, and n are parallel, as in figure 5.4.3. Reasoning with similar triangles, you get

$$\frac{A\text{ to }Z}{Z\text{ to }B}\,\frac{B\text{ to }X}{X\text{ to }C}\,\frac{C\text{ to }Y}{Y\text{ to }A}=\frac{A\text{ to }Z}{Z\text{ to }B}\,\frac{B\text{ to }A}{A\text{ to }Z}\,\frac{Z\text{ to }B}{B\text{ to }A}=1.$$

Part (*v*). Suppose X and Y lie outside the edges and Z within, and l, m, and n aren't all parallel. Then at least two of these lines meet; suppose l meets m at a point W. Then W lies outside the triangle, and the lines $\overleftrightarrow{CW}$ and $\overleftrightarrow{AB}$ meet at a point Z'. Proceed as in part (*ii*) to show that $Z=Z'$, hence W lies on n. The arguments for l or m meeting n are similar.

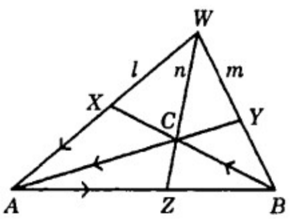

Figure 5.4.2 Ceva's theorem, part (*iii*)

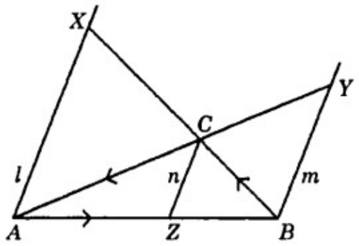

Figure 5.4.3 Ceva's theorem, part (*iv*)

Part (*vi*). If just one or all three of X, Y, and Z lie inside the edges, then Menelaus' product is negative, and you can verify that l, m, and n can't be parallel or concurrent. ♦

Ceva's theorem directly implies concurrence of the medians of ΔABC. Let X, Y, and Z be the edge midpoints A', B', and C', so that

$$\frac{A \text{ to } Z}{Z \text{ to } B}\,\frac{B \text{ to } X}{X \text{ to } C}\,\frac{C \text{ to } Y}{Y \text{ to } A} = \frac{A \text{ to } C'}{C' \text{ to } B}\,\frac{B \text{ to } A'}{A' \text{ to } C}\,\frac{C \text{ to } B'}{B' \text{ to } A} = 1\cdot 1\cdot 1 = 1.$$

By Ceva's theorem the medians are copencilar. They can't be parallel because the midpoints lie within the edges, so they must concur.

To derive the altitudes' concurrence from Ceva's theorem, first label the vertices of ΔABC so that $\angle A$ and $\angle B$ are acute. Then construct the circle with diameter $\overline{AB}$, as in figure 5.4.4. The feet D and E of the altitudes through A and B lie on this circle. Inscribed angles $\angle CAD$ and $\angle CBE$ are congruent because they subtend the same arc $\widehat{DJE}$, hence $\Delta ADC \sim \Delta BEC$, so

$$\frac{CE}{DC} = \frac{BC}{CA}\,.$$

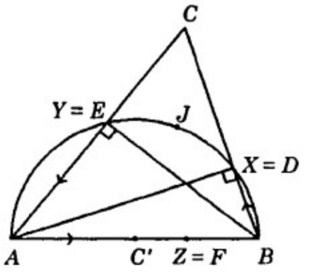

Figure 5.4.4 Deriving the altitudes' concurrence from Ceva's theorem

Giovanni CEVA was born around 1647 in Milan, to a wealthy family. He studied in Pisa, and later in life received a regular stipend from the archduke of Tuscany as a sort of court mathematician. In his 1678 book *De lineis rectis* is found the triangle theorem known by his name, and many elaborations, treated in the framework of mechanics. In several other books, he continued detailed studies in geometrical mechanics, many of which would be supplanted soon by calculus applications. He died in 1734. His Jesuit brother Tomasso was also a mathematician, and a poet.

Similarly,

$$\frac{BD}{FB} = \frac{AB}{BC} \qquad \frac{AF}{EA} = \frac{CA}{AB},$$

where F is the foot of the altitude through C. Now, choose scales for the edge lines as in figure 5.4.4, and let $X, Y, Z = D, E, F$. If $\angle C$ is acute, as in the figure,

$$\frac{A \text{ to } Z}{Z \text{ to } B} \, \frac{B \text{ to } X}{X \text{ to } C} \, \frac{C \text{ to } Y}{Y \text{ to } A} = \frac{AF}{FB} \, \frac{BD}{DC} \, \frac{CE}{EA}$$

$$= \frac{AF}{EA} \, \frac{BD}{FB} \, \frac{CE}{DC} = \frac{CA}{AB} \, \frac{AB}{BC} \, \frac{BC}{CA} = 1.$$

If $\angle C$ is obtuse, then you must place minus signs before two of the fractions in each of these products, but you get the same result. By Ceva's theorem, the altitudes are copencilar. Since they're perpendicular to the edges of a triangle, they can't be parallel, so they must concur.

To derive the concurrence of the angle bisectors of ΔABC from Ceva's theorem, you need a preliminary result. As in figure 5.4.5, the bisector g of $\angle A$ meets the opposite edge at a point X. Let $B^\star$ and $C^\star$ be the feet of the perpendiculars to g through B and C. If $B^\star \neq X \neq C^\star$, then $\Delta XBB^\star \sim \Delta XCC^\star$ and $\Delta ABB^\star \sim \Delta ACC^\star$. From these similarities it follows that

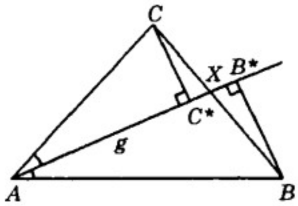

Figure 5.4.5 Deriving the angle bisectors' concurrence from Ceva's theorem

$$\frac{B \text{ to } X}{X \text{ to } C} = \frac{BB^*}{CC^*} = \frac{AB}{CA}\,.$$

On the other hand, if $B^* = X$ or $C^* = X$, then $\overleftrightarrow{BC} \perp g$, ΔABC is isosceles, and you get the same result:

$$\frac{B \text{ to } X}{X \text{ to } C} = 1 = \frac{AB}{CA}\,.$$

By the same argument, the bisectors of $\angle B$ and $\angle C$ meet the opposite edges at points Y and Z such that

$$\frac{C \text{ to } Y}{Y \text{ to } A} = \frac{BC}{AB} \qquad \frac{A \text{ to } Z}{Z \text{ to } B} = \frac{CA}{BC}\,.$$

The previous three equations entail

$$\frac{A \text{ to } Z}{Z \text{ to } B}\,\frac{B \text{ to } X}{X \text{ to } C}\,\frac{C \text{ to } Y}{Y \text{ to } A} = \frac{CA}{BC}\,\frac{AB}{CA}\,\frac{BC}{AB} = 1\,.$$

By Ceva's theorem, the angle bisectors are copencilar. They can't be parallel because X, Y, and Z lie within the edges, so they must concur.

The original derivation of Ceva's theorem was quite different from that given in this section. Ceva didn't base it directly on Euclid; instead, he used some principles governing the center of gravity of a system of weights. Given points X, Y, and Z within the edges of ΔABC opposite angles A, B, and C, Ceva assigned weights to the vertices as follows:

Vertex	*Weight*
A	$CY \cdot BX$
B	$YA \cdot XC$
C	$YA \cdot BX$.

(See figure 5.4.1.) Assuming that the supporting triangle has no mass, he then computed the center of gravity of the system of three weights. In such a computation, you can replace a subsystem of two weights u and v by a single weight $u + v$ at its center of gravity. This is the point P about which the moments of the weights u and v are opposite; they're on opposite sides of P, and the products of u and v by their distances from P are equal. Since $BX \cdot (YA \cdot XC) = XC \cdot (YA \cdot BX)$, the moments about X of the weights at B and C are opposite, so X is their center of gravity. You can replace those two weights by a single one at X. It follows that the center of gravity of the original system coincides with that of the new one consisting of weights at A and X, hence it lies on $\overline{AX}$. Similarly, since $CY \cdot (YA \cdot BX) = YA \cdot (CY \cdot BX)$, you can replace the weights at C and A by a single one at Y, so the center of gravity of the original system lies on $\overline{BY}$. Thus, it's the intersection of $\overline{AX}$ and $\overline{BY}$. Now, Ceva reasoned, segments $\overline{AX}$, $\overline{BY}$, and $\overline{CZ}$ are concurrent if and only if the center of gravity of the original system also falls on $\overline{CZ}$. That happens just in case you can replace the weights at A and B by a single weight at Z, i.e., just when their moments about Z are opposite: $AZ \cdot (CY \cdot BX) = ZB \cdot (YA \cdot XC)$. This condition simply says Menelaus' product equals 1:

$$\frac{A \text{ to } Z}{Z \text{ to } B} \frac{B \text{ to } X}{X \text{ to } C} \frac{C \text{ to } Y}{Y \text{ to } A} = 1.$$

This argument, of course, is not supported by the axiomatic foundation of chapter 3. But that foundation can be extended to describe the behavior of systems of weights, and justify the argument.

If $X, Y, Z = A', B', C'$ in the preceding argument, so that figure 5.4.1 represents the concurrence of the medians of ΔABC, then the weights assigned A, B, and C are all equal. Thus, in that case, Ceva's argument is identical to one used often in mechanics courses to demonstrate the medians' concurrence.

5.5 Trigonometry

Concepts

- Law of sines
- Law of cosines
- Determining other parts of a triangle from ASA, SAS, or SSS data
- Using SSA data
- Cosine and sine sum and difference formulas
- Cosine and sine double and half angle formulas
- SAS and ASA area formulas; Hero's SSS area formula
- Area formulas involving the circumradius
- Inequalities for approximating sine values

This section completes the study of plane trigonometry begun in chapter 3. In addition to the material you'd find in standard trigonometry texts, it presents several striking area formulas that go beyond what you'll find there. The section concludes with a method for approximating sine values, which also plays a major role in applying calculus to trigonometry problems.

Classical trigonometry shows how to determine some parts of ΔABC —measures of $\angle A$, $\angle B$, and $\angle C$ and lengths a, b, and c of the opposite edges—when you're given others. Theorem 3.13.15 and exercise 4.5.4 summarized this for the case where $\angle C$ is right. In that situation, given any two of a, b, c, $m\angle A$, and $m\angle B$, but not just the two angle measures, you can use Pythagoras' theorem, some simple equations involving the trigonometric functions, and computed values of these functions to find the remaining parts. For oblique triangles, you need more sophisticated

20210825/ Electromagnetism
Maxwell.../
Electromagnetism Maxwell
Equation ... pdf
Pag. 7 (27)
1.3. Integrals involving
vectors
Circulation of a vector
field
Fig 1.2.

20210825/ Fermat/
Methods of Geometry ... pdf
Pag. 159 (178)
Theorem 3 (Angle Bisectors
theorem)

The angle bisectors of ΔABC meet at an interior I equidistant from the edge lines

Fig. 5.1.2

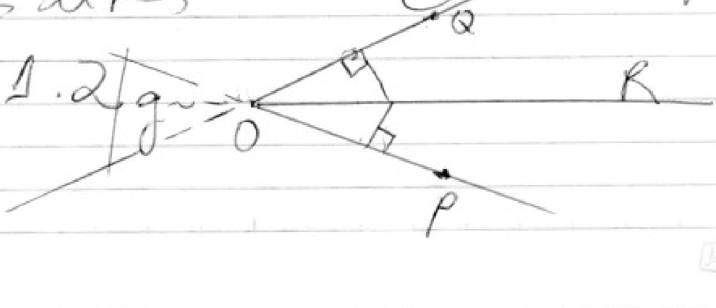

( ) $'
+,- * *. ! * ! @ ! 7 0 1#
) &$ 9 !1 7!

7

Curvature

We have seen, for example in rotating reference frames that the geometry in a space with non-vanishing acceleration of gravity, may be non-Euclidean. It is easy to visualize curves and surfaces in three-dimensional space but it is difficult to grasp visually what curvature means in three-dimensional space, or worse still, in four-dimensional space-time. However the curvature of such spaces may be discussed using the lower dimensional analogues of curves and surfaces. It is therefore important to have a good knowledge of the differential geometry of surfaces. Also the formalism used in describing surfaces may be taken over with minor modifications, when we are going to describe the geometric properties of curved space-time.

7.1 Curves

Let us consider curves in Euclidean three-dimensional space, $\mathbb{E}^3$ and let $\mathbf{r}(s)$ be the position vector of points parametrized by the *arc length* s. The unit tangent vector along the curve is

$$\mathbf{t} = \frac{d\mathbf{r}}{ds} = \frac{dx^i}{ds}\mathbf{e}_i. \tag{7.1}$$

The faster the unit tangent vector changes the direction along the curve the more curved the curve is. The *curvature vector* of the curve is defined by

$$\mathbf{k} = \frac{d\mathbf{t}}{ds}. \tag{7.2}$$

Since $\mathbf{t} \cdot \mathbf{t} = 1$ we have by differentiation $\mathbf{t} \cdot \mathbf{k} = 0$, thus the curvature vector is always orthogonal to the tangent vector. The length of the curvature vector is called the curvature of the curve, and is denoted by κ:

$$\kappa \equiv |\mathbf{k}|. \tag{7.3}$$

A curve with vanishing curvature is a straight line.

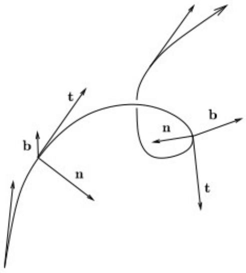

Figure 7.1: A curve in three-dimensional space

Example

Example 7.1 (The curvature of a circle)
Consider a circle of radius R. The tangent vector is

$$\mathbf{t} = \frac{d\mathbf{r}}{ds} = \frac{dr}{ds}\mathbf{e}_r + \frac{d\theta}{ds}\mathbf{e}_\theta = \frac{1}{R}\mathbf{e}_\theta \tag{7.4}$$

since $r = R$ and $s = R\theta$ along the circle. The curvature vector can now be found with

$$\mathbf{k} = \frac{d\mathbf{t}}{ds} = \frac{1}{R^2}\frac{d\mathbf{e}_\theta}{d\theta} = -\frac{1}{R}\mathbf{e}_r. \tag{7.5}$$

Thus the curvature of the circle is $\kappa = \frac{1}{R}$.

The unit vector $\mathbf{n}$ defined by

$$\boxed{\frac{d\mathbf{t}}{ds} = \kappa\mathbf{n}} \tag{7.6}$$

is called the *principal normal vector* of the curve. The vectors $\mathbf{t}$ and $\mathbf{n}$ span a plane which is called the *osculating plane* of the curve. This plane turns as we move along the curve. The unit normal vector $\mathbf{b}$ of this plane, defined by

$$\mathbf{b} = \mathbf{t} \times \mathbf{n} \tag{7.7}$$

is called the *binormal vector* of the curve. The rate of turning of the osculating plane is given by $\frac{d\mathbf{b}}{ds}$. Since $\mathbf{t} \cdot \mathbf{b} = 0$, it follows by differentiation that

$$\frac{d\mathbf{t}}{ds} \cdot \mathbf{b} + \mathbf{t} \cdot \frac{d\mathbf{b}}{ds} = 0. \tag{7.8}$$

Combining this with equations (7.6) and (7.7) it follows that $\frac{d\mathbf{b}}{ds}$ is orthogonal to $\mathbf{t}$. Since $\mathbf{b}$ has constant length, $\frac{d\mathbf{b}}{ds}$ has no component along $\mathbf{b}$ either. Thus $\frac{d\mathbf{b}}{ds}$ points in the $\mathbf{n}$ direction. The *torsion* τ of a curve is defined by the equation

$$\boxed{\frac{d\mathbf{b}}{ds} = -\tau\mathbf{n}.} \tag{7.9}$$

The vectors $\{\mathbf{t}, \mathbf{n}, \mathbf{b}\}$ represent three orthonormal basis vector fields along the curve. They are related by eq. (7.7) together with

$$\mathbf{t} = \mathbf{n} \times \mathbf{b}, \qquad \mathbf{n} = \mathbf{b} \times \mathbf{t}. \tag{7.10}$$

The variation of $\mathbf{n}$ along the curve is now given by

$$\frac{d\mathbf{n}}{ds} = \frac{d\mathbf{b}}{ds} \times \mathbf{t} + \mathbf{b} \times \frac{d\mathbf{n}}{ds} = -\tau\mathbf{n} \times \mathbf{t} + \kappa\mathbf{b} \times \mathbf{n} \tag{7.11}$$

so

$$\boxed{\frac{d\mathbf{n}}{ds} = \tau\mathbf{b} - \kappa\mathbf{t}.} \tag{7.12}$$

Equations (7.6), (7.9) and (7.12) are called the *Serret-Frenet equations*.

7.2 Surfaces

Consider a two-dimensional surface embedded in three-dimensional Euclidean space. Let u and v be coordinates (or parameters) on the surface. Then, at every point on the surface the basis vectors

$$\mathbf{e}_u = \frac{\partial}{\partial u}, \qquad \mathbf{e}_v = \frac{\partial}{\partial v} \tag{7.13}$$

define a tangent plane of the surface.

The line element on the surface is

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu \tag{7.14}$$

where $x^1 = u$ and $x^2 = v$. This is often called *the first fundamental form* of the surface.

The directional derivative of $\mathbf{e}_\mu$ along $\mathbf{e}_\nu$ has generally one component in the tangent plane of the surface and one component orthogonal to the surface,

$$\boxed{\mathbf{e}_{\mu,\nu} = \Gamma^\alpha_{\ \mu\nu}\mathbf{e}_\alpha + K_{\mu\nu}\mathbf{n}} \tag{7.15}$$

where $\Gamma^\alpha_{\ \mu\nu}$ is the connection coefficients of the u, v system and $\mathbf{n}$ is the unit normal vector field on the surface. The coefficients $K_{\mu\nu}$ are defined by this equation. Since $\mathbf{e}_{\mu,\nu} = \frac{\partial^2}{\partial x^\mu \partial x^\nu} = \frac{\partial^2}{\partial x^\nu \partial x^\mu} = \mathbf{e}_{\nu,\mu}$ these coefficients are symmetric. Note also that the eq. (7.15) provides us with an interpretation of the covariant derivative and the connection coefficients $\Gamma^\alpha_{\ \mu\nu}$. The covariant derivative of a surface embedded in an Euclidean space is the ordinary derivative in Euclidean space, projected onto the tangent space of the surface.

The equation (7.15) is usually called *Gauss' equations*. Using

$$K_{\mu\nu} = \mathbf{e}_{\mu,\nu} \cdot \mathbf{n} \tag{7.16}$$

we can calculate the coefficients $K_{\mu\nu}$. Let now $\mathbf{u}$ be the tangent vector of a curve in the surface, parametrized by λ. Using equation (7.15) we have

$$\frac{d\mathbf{u}}{d\lambda} = u^\mu_{\ ;\nu} u^\nu \mathbf{e}_\mu + K_{\mu\nu} u^\mu u^\nu \mathbf{n}. \tag{7.17}$$

The λ-independent coefficient in front of $\mathbf{n}$, $K_{\mu\nu} dx^\mu dx^\nu$ is called *the second fundamental form* of the surface. Whereas the first fundamental form determines the intrinsic geometry of the surface, the second fundamental form reflects the extrinsic geometry, i.e. how the surface curves in the ambient three-dimensional Euclidean space in which it is embedded.

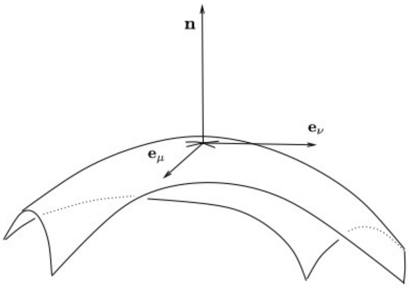

Figure 7.2: A two-dimensional surface embedded in three-dimensional Euclidean space

If λ is the arc-length, we can write equation (7.17) as

$$\frac{d\mathbf{u}}{d\lambda} = \kappa_g \mathbf{e} + \kappa_n \mathbf{n}, \quad \kappa_g \equiv |u^\mu{}_{;\nu} u^\nu \mathbf{e}_\mu|, \quad \kappa_n \equiv K_{\mu\nu} u^\mu u^\nu \tag{7.18}$$

where $\mathbf{e}$ is a unit vector. The quantities κ_g and κ_n are called *the geodesic curvature* and *the normal curvature* of the surface, respectively.

Since $\frac{d\mathbf{u}}{d\lambda}$ is orthogonal to $\mathbf{u}$ and $\mathbf{n}$, the unit vector $\mathbf{e}$ in the surface is given by

$$\mathbf{e} = \pm \mathbf{n} \times \mathbf{u}. \tag{7.19}$$

We see that the normal curvature can also be written

$$\kappa_n = \frac{d\mathbf{u}}{d\lambda} \cdot \mathbf{n}. \tag{7.20}$$

By differentiating $\mathbf{u} \cdot \mathbf{n}$ we get

$$\mathbf{u} \cdot \frac{d\mathbf{n}}{d\lambda} = -\kappa_n = -K_{\mu\nu} u^\mu u^\nu. \tag{7.21}$$

This is *Weingarden's equation*.

The geodesic and normal curvatures, taken separately, characterize the extrinsic geometry of the surface. We shall now find a quantity, called the Gaussian curvature, which is a measure of the intrinsic geometry of the surface.

At an arbitrary point on the surface we consider geodesic curves through the point with tangent vectors $\mathbf{u} = u^\mu \mathbf{e}_\mu$. In order to compare the normal curvature of the geodesics having different directions, we choose tangent vectors of unit length,

$$\mathbf{u} \cdot \mathbf{u} = g_{\mu\nu} u^\mu u^\nu = 1. \tag{7.22}$$

The directions with maximal and minimal values of the normal curvature are found by extremizing κ_n as given by eq. (7.18) with the constraint eq. (7.22). Thus we have to solve the variational problem δF for arbitrary u^μ, where

$$F = K_{\mu\nu} u^\mu u^\nu - k(g_{\mu\nu} u^\mu u^\nu - 1). \tag{7.23}$$

Here k is a Lagrange multiplier. Variation with respect to u^μ gives

$$\delta F = 2(K_{\mu\nu} - k g_{\mu\nu}) u^\nu \delta u^\mu. \tag{7.24}$$

Thus we must have

$$(K_{\mu\nu} - kg_{\mu\nu})u^{\nu} = 0. \tag{7.25}$$

This set of equations has non-trivial solutions whenever

$$\det(K_{\mu\nu} - kg_{\mu\nu}) = 0, \tag{7.26}$$

which yields the following quadratic equation for k

$$k^2 \det(g_{\mu\nu}) - k(g_{11}K_{22} - 2g_{12}K_{12} + g_{22}K_{11}) + \det(K_{\mu\nu}) = 0 \tag{7.27}$$

with solutions k_1 and k_2. These are the extremal values of k.

To see the meaning of k we multiply eq. (7.25) by u^{μ} and use eq. (7.18). This gives

$$k = \kappa_n. \tag{7.28}$$

The extremal values of k are called the *principal curvatures* of the surface. The *Gaussian curvature* of the surface is defined as

$$K = k_1 k_2. \tag{7.29}$$

From eq. (7.27) it follows that

$$\boxed{K = \frac{\det(K_{\mu\nu})}{\det(g_{\mu\nu})}.} \tag{7.30}$$

Let the directions corresponding to the principal curvatures be characterized by the vectors **u** and **v**. From equation (7.25) and the symmetry of $K_{\mu\nu}$ follows

$$(k_1 - k_2)(\mathbf{u} \cdot \mathbf{v}) = 0. \tag{7.31}$$

For $k_1 \neq k_2$ this implies that **u** and **v** are orthogonal. The principal curvatures are found in orthogonal directions.

A positive Gaussian curvature means that the principal curvatures are of the same sign. The surface is locally similar to a distorted sphere, and the geometry is locally elliptic. If the Gaussian curvature vanishes, one of the principal curvatures have to vanish and the geometry is locally planar and the geometry is locally Euclidean. Finally, if the Gaussian curvature is negative the surface is locally saddle-shaped, i.e. like a hyperbolic surface. In this case the geometry is said to be locally hyperbolic.

In the following sections we will show that the Gaussian curvature represents a measure of the intrinsic geometry of the surface. This will be achieved by expressing it in terms of the components of the Riemann tensor, which we will introduce in the next section. Note however that curves have no intrinsic curvature.

7.3 The Riemann Curvature Tensor

In this section we will not restrict ourselves to two-dimensional surfaces, but describe spaces of arbitrary number of dimensions. We will introduce the important concept of curvature in an invariant way, described by the Riemann curvature tensor.

Consider two nearby points Q and P connected by a vector $\delta\lambda\mathbf{v}$ of infinitesimal length. Let $\mathbf{A}$ be a vector field. Then the difference between the vector field at Q, denoted $\mathbf{A}_Q$, and the vector $\mathbf{A}_P$ parallel-transported from P to Q (which is denoted $\mathbf{A}_{PQ}$) is to first order in $\delta\lambda$ given by the directional derivative at Q of $\mathbf{A}$ in the $\mathbf{v}$-direction

$$\delta\lambda\nabla_{\mathbf{v}}\mathbf{A} \approx \mathbf{A}_Q - \mathbf{A}_{PQ} \tag{7.32}$$

Thus

$$\mathbf{A}_{PQ} \approx \left(1 - \delta\lambda\nabla_{\mathbf{v}}\right)\mathbf{A}_Q \tag{7.33}$$

To second order $\mathbf{A}_{PQ}$ is given by the first terms of the Taylor expansion

$$\mathbf{A}_{PQ} \approx \left(1 - \delta\lambda\nabla_{\mathbf{v}} + \frac{1}{2}\delta\lambda^2\nabla_{\mathbf{v}}\nabla_{\mathbf{v}}\right)\mathbf{A}_Q \tag{7.34}$$

We are now going to parallel-transport the vector $\mathbf{A}_P$ around the polygon shown in Fig.7.3. Parallel-transporting $\mathbf{A}_{PQ}$ from Q to R gives

$$\mathbf{A}_{PQR} \approx \left(1 - \delta\lambda\nabla_{\mathbf{u}} + \frac{1}{2}\delta\lambda^2\nabla_{\mathbf{u}}\nabla_{\mathbf{u}}\right)\left(1 - \delta\lambda\nabla_{\mathbf{v}} + \frac{1}{2}\delta\lambda^2\nabla_{\mathbf{v}}\nabla_{\mathbf{v}}\right)\mathbf{A}_R. \tag{7.35}$$

Proceeding around the polygon gives

$$\begin{aligned}\mathbf{A}_{PQRSTP} \approx\ & \left(1 + \delta\lambda\nabla_{\mathbf{u}} + \frac{1}{2}\delta\lambda^2\nabla_{\mathbf{u}}\nabla_{\mathbf{u}}\right)\left(1 + \delta\lambda\nabla_{\mathbf{v}} + \frac{1}{2}\delta\lambda^2\nabla_{\mathbf{v}}\nabla_{\mathbf{v}}\right) \\ \times\ & \left(1 - \delta\lambda^2\nabla_{[\mathbf{u},\mathbf{v}]}\right) \qquad (7.36) \\ \times\ & \left(1 - \delta\lambda\nabla_{\mathbf{u}} + \frac{1}{2}\delta\lambda^2\nabla_{\mathbf{u}}\nabla_{\mathbf{u}}\right)\left(1 - \delta\lambda\nabla_{\mathbf{v}} + \frac{1}{2}\delta\lambda^2\nabla_{\mathbf{v}}\nabla_{\mathbf{v}}\right)\mathbf{A}_R.\end{aligned}$$

Thus, to second order, the change of the vector after parallel-transport around the polygon is

$$\delta\mathbf{A} = \mathbf{A}_{PQRSTP} - \mathbf{A}_P = \left([\nabla_{\mathbf{u}}, \nabla_{\mathbf{v}}] - \nabla_{[\mathbf{u},\mathbf{v}]}\right)\delta\lambda^2\mathbf{A}_P \tag{7.37}$$

In flat space there would be no such change of $\mathbf{A}$. This change is due to the curvature of the space. It can be shown that $\delta\mathbf{A}$ as given in eq. (7.37) is a vector which is linear in $\mathbf{A}$, $\mathbf{u}$ and $\mathbf{v}$ (see e.g. [vW81]). Thus $\delta\mathbf{A}$ may be expressed by a tensor of rank $\{^1_3\}$. This tensor $\mathbf{R}$ is called the *Riemann curvature tensor* and is defined by

$$\boxed{\mathbf{R}(\mathbf{u}, \mathbf{v})\mathbf{A} \equiv \left([\nabla_{\mathbf{u}}, \nabla_{\mathbf{v}}] - \nabla_{[\mathbf{u},\mathbf{v}]}\right)\mathbf{A}.} \tag{7.38}$$

The components of this tensor are given by

$$\mathbf{e}_\mu R^\mu{}_{\nu\alpha\beta} = \left([\nabla_\alpha, \nabla_\beta] - \nabla_{[\mathbf{e}_\alpha,\mathbf{e}_\beta]}\right)\mathbf{e}_\nu \tag{7.39}$$

It follows that the Riemann curvature tensor is antisymmetric in $\mathbf{u}$ and $\mathbf{v}$ i.e. in α and β. We can therefore define a matrix of two-forms

$$\boxed{\mathbf{R}^\mu_\nu = \frac{1}{2}R^\mu{}_{\nu\alpha\beta}\boldsymbol{\omega}^\alpha \wedge \boldsymbol{\omega}^\beta} \tag{7.40}$$

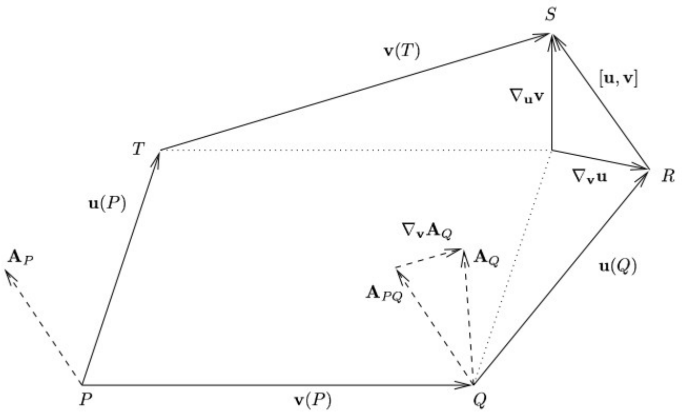
S
$\mathbf{v}(T)$
$\nabla_{\mathbf{u}}\mathbf{v}$
$[\mathbf{u},\mathbf{v}]$
T
$\nabla_{\mathbf{v}}\mathbf{u}$
R
$\mathbf{u}(P)$
$\nabla_{\mathbf{v}}\mathbf{A}_Q$
$\mathbf{A}_P$
$\mathbf{A}_Q$
$\mathbf{u}(Q)$
$\mathbf{A}_{PQ}$
P
$\mathbf{v}(P)$
Q

Figure 7.3: The vector $\mathbf{A}_P$ parallel transported around the polygon $PQRSTP$

which are called the *curvature forms*.

Equation (7.37) may now be written

$$\delta\mathbf{A} = \mathbf{e}_\mu R^\mu{}_{\nu\alpha\beta} A^\nu u^\alpha v^\beta. \tag{7.41}$$

The infinitesimal area of the polygon is to lowest order in $\mathbf{u}$ and $\mathbf{v}$

$$\Delta S^{\alpha\beta} = u^\alpha v^\beta - u^\beta v^\alpha. \tag{7.42}$$

Using the antisymmetry of $R^\mu{}_{\nu\alpha\beta}$ we may then write

$$\delta\mathbf{A} = \frac{1}{2} A^\nu R^\mu{}_{\nu\alpha\beta} \Delta S^{\alpha\beta} \mathbf{e}_\mu. \tag{7.43}$$

This shows that the change of a vector by parallel transport around a closed curve is proportional to the curvature of the space and to the area enclosed by the curve.

We shall now show how $R^\mu{}_{\nu\alpha\beta}$ may be expressed by the connection- and structure coefficients of a given basis. We find

$$\begin{aligned} \mathbf{e}_\mu R^\mu{}_{\nu\alpha\beta} &= \left([\nabla_\alpha, \nabla_\beta] - \nabla_{[\mathbf{e}_\alpha, \mathbf{e}_\beta]}\right)\mathbf{e}_\nu \\ &= \left(\nabla_\alpha\nabla_\beta - \nabla_\beta\nabla_\alpha - c^\rho{}_{\alpha\beta}\nabla_\rho\right)\mathbf{e}_\nu \\ &= \left(\Gamma^\mu{}_{\nu\beta,\alpha} + \Gamma^\rho{}_{\nu\beta}\Gamma^\mu{}_{\rho\alpha} - \Gamma^\mu{}_{\nu\alpha,\beta} - \Gamma^\rho{}_{\nu\alpha}\Gamma^\mu{}_{\rho\beta} - c^\rho{}_{\alpha\beta}\Gamma^\mu{}_{\nu\rho}\right)\mathbf{e}_\mu \end{aligned} \tag{7.44}$$

which implies that

$$\boxed{R^\mu{}_{\nu\alpha\beta} = \Gamma^\mu{}_{\nu\beta,\alpha} - \Gamma^\mu{}_{\nu\alpha,\beta} + \Gamma^\rho{}_{\nu\beta}\Gamma^\mu{}_{\rho\alpha} - \Gamma^\rho{}_{\nu\alpha}\Gamma^\mu{}_{\rho\beta} - c^\rho{}_{\alpha\beta}\Gamma^\mu{}_{\nu\rho}.} \tag{7.45}$$

The connection coefficients and structure coefficients are calculated from a basis vector field in the space considered, without reference to any higher dimensional space. Thus eq. (7.45) shows that the Riemann tensor describes the intrinsic geometry of space.

The curvature forms may now be written in component form as

$$\mathbf{R}^{\mu}{}_{\nu} = \left(\Gamma^{\mu}{}_{\nu\beta,\alpha} - \frac{1}{2} c^{\rho}{}_{\alpha\beta} \Gamma^{\mu}{}_{\nu\rho} + \Gamma^{\rho}{}_{\nu\beta} \Gamma^{\mu}{}_{\rho\alpha} \right) \boldsymbol{\omega}^{\alpha} \wedge \boldsymbol{\omega}^{\beta}. \tag{7.46}$$

Using equations (6.136) and (6.165) we find

$$\boxed{\mathbf{R}^{\mu}{}_{\nu} = \mathbf{d}\boldsymbol{\Omega}^{\mu}{}_{\nu} + \boldsymbol{\Omega}^{\mu}{}_{\lambda} \wedge \boldsymbol{\Omega}^{\lambda}{}_{\nu}.} \tag{7.47}$$

This is *Cartan's 2nd structural equation.*

We shall now deduce some identities fulfilled by the Riemann tensor. Using eqs. (6.118) and (6.144) we get

$$\begin{aligned} A^{\mu}{}_{;\beta\alpha} \quad = \quad & A^{\lambda} \Gamma^{\mu}{}_{\lambda\beta,\alpha} - A^{\lambda} \Gamma^{\tau}{}_{\lambda\alpha} \Gamma^{\mu}{}_{\tau\beta} - A^{\mu}{}_{;\lambda} \Gamma^{\lambda}{}_{\beta\alpha} + A^{\mu}{}_{,\beta\alpha} \\ & + \left(A^{\lambda}{}_{;\alpha} \Gamma^{\mu}{}_{\lambda\beta} + A^{\lambda}{}_{;\beta} \Gamma^{\mu}{}_{\lambda\alpha} \right). \end{aligned} \tag{7.48}$$

The two latter terms are symmetric in α and β and do not contribute to the antisymmetric combination $A^{\mu}{}_{;\beta\alpha} - A^{\mu}{}_{;\alpha\beta}$. Hence, by using eqs. (7.45) and (6.134), we get

$$A^{\mu}{}_{;\beta\alpha} - A^{\mu}{}_{;\alpha\beta} = A^{\lambda} R^{\mu}{}_{\lambda\alpha\beta} - A^{\mu}{}_{;\lambda} T^{\lambda}_{\alpha\beta}. \tag{7.49}$$

Since this is a tensor equation it is valid in an arbitrary basis. Eq. (7.49) is called the *Ricci identity*. In Riemannian geometry it reduces to

$$\boxed{A^{\mu}{}_{;\beta\alpha} - A^{\mu}{}_{;\alpha\beta} = R^{\mu}{}_{\lambda\alpha\beta} A^{\lambda}.} \tag{7.50}$$

Combining this with eq. (6.163) leads to

$$\mathbf{d}^2 \mathbf{A} = \frac{1}{2} R^{\mu}{}_{\nu\alpha\beta} A^{\nu} \mathbf{e}_{\mu} \otimes \boldsymbol{\omega}^{\alpha} \wedge \boldsymbol{\omega}^{\beta} \tag{7.51}$$

By exterior differentiation of Cartan's first structural equation, eq. (6.180) combined with Cartan's second structural equation and Poincaré's Lemma, we find

$$\mathbf{R}^{\mu}{}_{\nu} \wedge \boldsymbol{\omega}^{\nu} = \mathbf{d}\mathbf{T}^{\mu} + \boldsymbol{\Omega}^{\mu}{}_{\nu} \wedge \mathbf{T}^{\nu}. \tag{7.52}$$

By means of eq. (6.216), this equation takes the form

$$\mathbf{R}^{\mu}{}_{\nu} \wedge \boldsymbol{\omega}^{\nu} = \mathbf{D}\mathbf{T}^{\mu}. \tag{7.53}$$

This is *Bianchi's first identity*. In Riemannian geometry it reduces to

$$\boxed{\mathbf{R}^{\mu}{}_{\nu} \wedge \boldsymbol{\omega}^{\nu} = 0.} \tag{7.54}$$

On component form this equation is

$$\boxed{R^{\mu}{}_{[\nu\alpha\beta]} = 0.} \tag{7.55}$$

By exterior differentiation of Cartan's second structural equation, and using Poincaré's Lemma, we get

$$\mathbf{d}\mathbf{R}^{\mu}{}_{\nu} = \mathbf{d}\boldsymbol{\Omega}^{\mu}{}_{\lambda} \wedge \boldsymbol{\Omega}^{\lambda}{}_{\nu} - \boldsymbol{\Omega}^{\mu}{}_{\lambda} \wedge \mathbf{d}\boldsymbol{\Omega}^{\lambda}{}_{\nu} = \mathbf{R}^{\mu}{}_{\lambda} \wedge \boldsymbol{\Omega}^{\lambda}{}_{\nu} - \boldsymbol{\Omega}^{\mu}{}_{\lambda} \wedge \mathbf{R}^{\lambda}{}_{\nu}, \tag{7.56}$$

which is more usually written

$$\mathbf{dR}^{\mu}_{\ \nu} + \mathbf{\Omega}^{\mu}_{\ \lambda} \wedge \mathbf{R}^{\lambda}_{\ \nu} - \mathbf{R}^{\mu}_{\ \lambda} \wedge \mathbf{\Omega}^{\lambda}_{\ \nu} = 0. \tag{7.57}$$

Applying eq. (6.214) to a matrix of two-forms this equation may be written

$$\boxed{\mathbf{DR}^{\mu}_{\ \nu} = 0.} \tag{7.58}$$

This is *Bianchi's second identity*. On component form this identity becomes

$$\boxed{R^{\mu}_{\ \nu[\alpha\beta;\gamma]} = 0.} \tag{7.59}$$

An additional symmetry of the Riemann tensor is most easily found by decomposing it in an orthonormal basis field. Applying eq. (6.169) and Cartan's second structural equation (7.47), we get

$$\boxed{R_{\mu\nu\alpha\beta} = -R_{\nu\mu\alpha\beta}} \tag{7.60}$$

The fourth and last symmetry of the Riemann tensor is found by applying geodesic normal coordinates. From eq. (7.45), (6.110) and (6.232) we then get

$$R_{\mu\nu\alpha\beta} = \frac{1}{2}(g_{\mu\beta,\nu\alpha} - g_{\mu\alpha,\nu\beta} + g_{\nu\alpha,\mu\beta} - g_{\nu\beta,\mu\alpha}). \tag{7.61}$$

It follows that

$$\boxed{R_{\mu\nu\alpha\beta} = R_{\alpha\beta\mu\nu}.} \tag{7.62}$$

The four symmetries of the Riemann tensor reduce its number of independent components in an n-dimensional space from n^4 to $\frac{1}{12}n^2(n^2-1)$, in four-dimensional space-time from 256 to 20.

We shall now construct a curvature tensor of rank $\{^0_2\}$ by contraction of the Riemann tensor. Note first that

$$R^{\alpha}_{\ \alpha\mu\nu} = 0 \tag{7.63}$$

because of the antisymmetry in the first two indices. Furthermore

$$R^{\alpha}_{\ \mu\nu\alpha} = -R^{\alpha}_{\ \mu\alpha\nu}. \tag{7.64}$$

Thus there exists only one independent non-vanishing contraction of the Riemann tensor. This is called the *Ricci tensor* and is usually written

$$\boxed{R_{\mu\nu} = R^{\alpha}_{\ \mu\alpha\nu}.} \tag{7.65}$$

From eq. (7.62) follows that it is symmetrical. It has $\frac{1}{2}n(n+1)$ independent components in an n-dimensional space; 10 components in four-dimensional space-time.

Contraction of the Ricci tensor gives the *Ricci scalar*

$$\boxed{R = R^{\mu}_{\ \mu}.} \tag{7.66}$$

Let us calculate the divergence of the Ricci tensor. Contraction of Bianchi's second identity eq.(7.58) (μ with α), yields

$$\begin{aligned} R^{\mu}_{\ \nu\mu\beta;\gamma} + R^{\mu}_{\ \nu\gamma\mu;\beta} + R^{\mu}_{\ \nu\beta\gamma;\mu} &= \\ R_{\nu\beta;\gamma} - R_{\nu\gamma;\beta} + R^{\mu}_{\ \nu\beta\gamma;\mu} &= 0. \end{aligned} \tag{7.67}$$

Raising the index ν and contracting with γ leads to

$$R^{\nu}{}_{\beta;\nu} - R_{;\beta} + R^{\mu}{}_{\beta;\mu} = 0. \tag{7.68}$$

Hence the divergence of the Ricci tensor is

$$R^{\nu}{}_{\mu;\nu} = \frac{1}{2}\left(\delta^{\nu}{}_{\mu}R\right)_{;\nu}. \tag{7.69}$$

Thus the tensor

$$E^{\nu}{}_{\mu} = R^{\nu}{}_{\mu} - \frac{1}{2}\delta^{\nu}{}_{\mu}R \tag{7.70}$$

is divergence free. This is *Einstein's curvature tensor*. Its covariant components are

$$\boxed{E_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu}.} \tag{7.71}$$

It follows immediately that this tensor is symmetric.

In an n-dimensional space the vanishing of the divergence of the Einstein tensor represents n equations. Thus the Einstein tensor has $\frac{1}{2}n(n-1)$ independent components in general, and 6 independent components in four-dimensional space-time.

The Weyl Curvature Tensor

Let us now focus on the case where the dimension of the manifold is 4, like our four-dimensional spacetime. The symmetries of the Riemann tensor imply that the Riemann tensor has 20 independent components in four dimensions. The Ricci tensor, on the other hand, has only 10 independent components. The components of the Riemann tensor which is not captured in the Ricci part form what is called the *Weyl curvature tensor*.

In four dimensions the Weyl curvature tensor is defined by

$$C_{\alpha\beta\gamma\delta} = R_{\alpha\beta\gamma\delta} - g_{\alpha[\gamma}R_{\delta]\beta} + g_{\beta[\gamma}R_{\delta]\alpha} + \frac{1}{3}Rg_{\alpha[\gamma}g_{\delta]\beta}. \tag{7.72}$$

It possesses the same symmetries as the Riemann tensor,

$$C_{\alpha\beta\gamma\delta} = C_{\gamma\delta\alpha\beta}, \quad C_{\alpha\beta\gamma\delta} = -C_{\beta\alpha\gamma\delta}, \quad C_{\alpha[\beta\gamma\delta]} = 0,$$

and in addition, *contraction over any pair of indices yields zero*:

$$C^{\alpha}{}_{\beta\alpha\delta} = 0. \tag{7.73}$$

Hence, when contracting the Riemann tensor over two indices only the Ricci part of it will survive. This gives, as we will see later, a physical interpretation of the Weyl tensor. Einstein's equations, which will be introduced in the next chapter, will only involve the Ricci tensor and hence the Weyl tensor represents the *free gravitational field*. Thus even if the Ricci tensor is zero, there can be a free gravitational field encoded in the Weyl tensor. This property gives rise to many interesting phenomena; many of which will be discussed in this book. Two important examples are gravitational waves, and the Schwarzschild vacuum solution.

The definition of the Weyl tensor in any dimension, see section 18.4. This section also discusses some further properties of the Weyl tensor.

7.4 Extrinsic and Intrinsic Curvature

From eq. (7.45) it is seen that the components of the Riemann tensor may be calculated from the components of the metric tensor. They are defined in terms of basis-vectors in a space which has a curvature specified by the Riemann tensor. Thus the specification of the metric does not presuppose any embedding of the curved space in a higher-dimensional flat space. The Riemann curvature tensor represents intrinsic geometric properties of space, which may be measured by inhabitants of that space. (Such inhabitants are always assumed to be creatures with the same number of dimensions as that of the space they inhabit.) Therefore one says that the Riemann tensor is a measure of the *intrinsic* curvature of space.

In the case of a two-dimensional surface, say a balloon, the intrinsic geometry is measured by two-dimensional creatures on the surface, "flatlanders". In general, if the Riemann tensor in a space vanishes, the space is flat. For a two-dimensional surface this means that the surface may be rolled onto an Euclidean plane without any local changes of the geometry. If the surface is an elastic membrane, no stresses or strains are introduced by this "rolling out" of it. The intrinsic geometry of a cylindrical surface for example, is Euclidean, and has no intrinsic curvature.

However, as seen from an external Euclidean three-dimensional space, the cylindrical surface looks curved. The surface has external curvature. We shall now introduce a tensor measuring the external curvature of a space, which is embedded in a space of one more dimension.

In the following we shall consider a curved n-dimensional space $\mathcal{M}^n$ embedded in an $(n+1)$-dimensional space $\mathcal{M}^{n+1}$ which also may be curved. Such a space embedded in an space with one dimensional higher, is called a *hypersurface*. Furthermore, Greek indices are associated with $\mathcal{M}^{n+1}$ and Latin indices with $\mathcal{M}^n$.

A measure of the extrinsic curvature of a space is obtained by considering how the direction of the unit normal vector $\mathbf{N}$ to the hypersurface changes with position on the hypersurface. *The extrinsic curvature tensor* $\mathbf{K}$ is a tensor on $\mathcal{M}^n$ of rank $\left\{ {0 \atop 2} \right\}$ defined up to a sign ambiguity by

$$\boxed{K_{ab} = -\mathbf{e}_b \cdot \nabla_a \mathbf{n}} \tag{7.74}$$

where the covariant derivative is taken in the ambient space $\mathcal{M}^{n+1}$. Since $\mathbf{N}$ is orthogonal to the basis vectors $\mathbf{e}_b$ on $\mathcal{M}^n$, so that $\nabla_a(\mathbf{e}_b \cdot \mathbf{n}) = 0$ we get

$$K_{ab} = \mathbf{n} \cdot \nabla_a \mathbf{e}_b = \mathbf{n} \cdot \mathbf{e}_\alpha \Gamma^{\alpha}_{\ ba} = n^\alpha g_{\alpha\beta} \Gamma^{\beta}_{\ ba} = n_\alpha \Gamma^{\alpha}_{\ ba}, \tag{7.75}$$

where $\mathbf{n} = n^\alpha \mathbf{e}_\alpha$. We introduce an orthonormal basis $\{\mathbf{e}_{\hat{a}}\}$ on $\mathcal{M}^n$ and a normal unit vector $\mathbf{n} = \mathbf{e}_{\hat{n}}$. By defining $\mathbf{n} \cdot \mathbf{n} = \epsilon = \pm 1$ we get

$$K_{\hat{a}\hat{b}} = \Gamma^{\hat{n}}_{\ \hat{a}\hat{b}}\,, \qquad g_{\hat{n}\hat{n}} = \epsilon. \tag{7.76}$$

Equation (7.75) shows that the extrinsic curvature is symmetric.

Let us now deduce the relation between the Riemann tensors of $\mathcal{M}^{n+1}$ and $\mathcal{M}^n$ and the extrinsic curvature of $\mathcal{M}^n$.

We will calculate the Riemann tensors using orthonormal frames. The Riemann tensor in $\mathcal{M}^{n+1}$ projected onto $\mathcal{M}^n$ is given by

$$\left(\mathbf{d}^2 \mathbf{e}_{\hat{b}}\right)_\perp = \left(\frac{1}{2} {}^{(n+1)}R^{\hat{\lambda}}_{\ \hat{b}\hat{\alpha}\hat{\beta}} \mathbf{e}_{\hat{\lambda}} \otimes \boldsymbol{\omega}^{\hat{\alpha}} \wedge \boldsymbol{\omega}^{\hat{\beta}}\right)_\perp = \frac{1}{2} {}^{(n+1)}R^{\hat{a}}_{\ \hat{b}\hat{c}\hat{d}} \mathbf{e}_{\hat{a}} \otimes \boldsymbol{\omega}^{\hat{c}} \wedge \boldsymbol{\omega}^{\hat{d}}. \tag{7.77}$$

We can also write this in another way, using instead the Riemann tensor in $\mathcal{M}^n$

$$\begin{aligned}\left(\mathbf{d}^2\mathbf{e}_{\hat{b}}\right)_\perp &= \left(\mathbf{d}\left[\mathbf{e}_{\hat{\alpha}}\otimes\mathbf{\Omega}^{\hat{\alpha}}{}_{\hat{b}}\right]\right)_\perp = \left(\mathbf{d}\mathbf{e}_{\hat{\alpha}}\otimes\mathbf{\Omega}^{\hat{\alpha}}{}_{\hat{b}}\right)_\perp + \left(\mathbf{e}_{\hat{\alpha}}\otimes\mathbf{d}\mathbf{\Omega}^{\hat{\alpha}}{}_{\hat{b}}\right)_\perp \\ &= \mathbf{e}_{\hat{a}}\otimes\left(\mathbf{d}\mathbf{\Omega}^{\hat{a}}{}_{\hat{b}} + \mathbf{\Omega}^{\hat{a}}{}_{\hat{\alpha}}\wedge\mathbf{\Omega}^{\hat{\alpha}}{}_{\hat{b}}\right)_\perp . \end{aligned} \tag{7.78}$$

We can now use Cartan's second structural equation in $\mathcal{M}^n$ by decomposing the wedge product

$$\begin{aligned}\left(\mathbf{d}^2\mathbf{e}_{\hat{b}}\right)_\perp &= \mathbf{e}_{\hat{a}}\otimes\left(\mathbf{d}\mathbf{\Omega}^{\hat{a}}{}_{\hat{b}} + \mathbf{\Omega}^{\hat{a}}{}_{\hat{\lambda}}\wedge\mathbf{\Omega}^{\hat{\lambda}}{}_{\hat{b}} + \mathbf{\Omega}^{\hat{a}}{}_{\hat{n}}\wedge\mathbf{\Omega}^{\hat{n}}{}_{\hat{b}}\right)_\perp \\ &= \mathbf{e}_{\hat{a}}\otimes\left({}^{(n)}\mathbf{R}^{\hat{a}}{}_{b} + \left[\mathbf{\Omega}^{\hat{a}}{}_{\hat{n}}\wedge\mathbf{\Omega}^{\hat{n}}{}_{\hat{b}}\right]_\perp\right) \\ &= \frac{1}{2}\left({}^{(n)}R^{\hat{a}}{}_{\hat{b}\hat{c}\hat{d}} + \Gamma^{\hat{a}}{}_{\hat{n}\hat{c}}\Gamma^{\hat{n}}{}_{\hat{b}\hat{d}}\right)\mathbf{e}_{\hat{a}}\otimes\boldsymbol{\omega}^{\hat{c}}\wedge\boldsymbol{\omega}^{\hat{d}} \\ &= \frac{1}{2}\left({}^{(n)}R^{\hat{a}}{}_{\hat{b}\hat{c}\hat{d}} \pm K^{\hat{a}}{}_{\hat{c}}K_{\hat{b}\hat{d}}\right)\mathbf{e}_{\hat{a}}\otimes\boldsymbol{\omega}^{\hat{c}}\wedge\boldsymbol{\omega}^{\hat{d}}. \end{aligned} \tag{7.79}$$

From equations (7.77) and (7.79) it follows that

$${}^{(n)}R^{\hat{a}}{}_{\hat{b}\hat{c}\hat{d}} = {}^{(n+1)}R^{\hat{a}}{}_{\hat{b}\hat{c}\hat{d}} \pm 2K^{\hat{a}}{}_{[\hat{c}}K_{\hat{d}]\hat{b}}. \tag{7.80}$$

This is an expression of *Gauss' Theorema Egregium*, which says that an inhabitant of, say, a three-dimensional space, may perform measurements, within the three-dimensional space, which reveals to him the curvature of that space. It is not necessary to embed the space in a higher-dimensional one.

Let us write equation (7.80) in a covariant manner. The Riemann tensor on the right side in eq. (7.80) is the projected Riemann tensor of the ambient space $\mathcal{M}^{n+1}$. If we split the metric tensor $g_{\alpha\beta}$ into

$$g_{\alpha\beta} = h_{\alpha\beta} + \epsilon n_\alpha n_\beta, \tag{7.81}$$

where n_α are the components of the unit normal vector $\mathbf{n}$, then the tensor $h_{\alpha\beta}$ will act as a projection tensor onto the space $\mathcal{M}^n$. In addition, the tensor $h_{\alpha\beta}$ will be the metric tensor of the space $\mathcal{M}^n$ for vectors on $\mathcal{M}^n$. Since the exterior curvature and the Riemann tensor in $\mathcal{M}^n$ already are projected we have

$$h^{\lambda}{}_{\alpha}K_{\lambda\beta} = K_{\alpha\beta} \tag{7.82}$$

i.e. they are eigentensors to the projection map $h^{\lambda}{}_{\alpha}$. The equation (7.80) can now be written covariantly as

$$\boxed{{}^{(n)}R^{\alpha}{}_{\beta\mu\nu} = {}^{(n+1)}R^{\lambda}{}_{\gamma\sigma\rho}h^{\alpha}{}_{\lambda}h^{\gamma}{}_{\beta}h^{\sigma}{}_{\mu}h^{\rho}{}_{\nu} + \epsilon\left(K^{\alpha}{}_{\mu}K_{\beta\nu} - K^{\alpha}{}_{\nu}K_{\beta\mu}\right).} \tag{7.83}$$

Similarly, we can show that

$$\boxed{{}^{(n)}\nabla_\alpha K_{\beta\mu} - {}^{(n)}\nabla_\mu K_{\beta\alpha} = {}^{(n+1)}R^{\lambda}{}_{\sigma\rho\delta}n_\lambda h^{\sigma}{}_{\beta}h^{\rho}{}_{\alpha}h^{\delta}{}_{\mu}} \tag{7.84}$$

where ${}^{(n)}\nabla_\alpha = h^{\beta}{}_{\alpha}\nabla_\beta$ is the n-dimensional connection. The derivation of this is left as a problem (see problem 7.4). This equation is called the *Codazzi equation*.

In the special case where $\mathcal{M}^{n+1}$ is flat, eq. (7.80) reduce to

$${}^{(n)}R_{\hat{a}\hat{b}\hat{c}\hat{d}} = \pm\left(K_{\hat{a}\hat{c}}K_{\hat{d}\hat{b}} - K_{\hat{a}\hat{d}}K_{\hat{c}\hat{b}}\right). \tag{7.85}$$

For a two dimensional surface $\mathcal{M}^2$ these equations reduce to the single equation

$$R_{\hat{1}\hat{2}\hat{1}\hat{2}} = K_{\hat{1}\hat{1}} K_{\hat{1}\hat{2}} - (K_{\hat{1}\hat{2}})^2 = \det(K_{\hat{a}\hat{b}}). \tag{7.86}$$

Comparing with eq. (7.30) of a surface (with $\det(g_{\hat{a}\hat{b}}) = 1$) we obtain

$$K = R_{\hat{1}\hat{2}\hat{1}\hat{2}}. \tag{7.87}$$

Thus, the Gaussian curvature represents the *intrinsic* geometry of the surface.

Example 7.2 (The curvature of a straight circular cone) **Example**

From the figure 7.4 it is seen that the line element on the conical surface is

$$ds^2 = dl^2 + r^2 d\theta^2 = dl^2 + \left(\frac{R}{H}\right)^2 l^2 d\theta^2. \tag{7.88}$$

Here are l and θ coordinates on the surface, and r a coordinate normal to the axis of the cone.

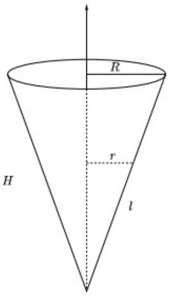

Figure 7.4: A cone is intrinsically flat

We introduce an orthonormal basis on the surface

$$\boldsymbol{\omega}^{\hat{l}} = \mathbf{dl}, \quad \boldsymbol{\omega}^{\hat{\theta}} = \left(\frac{R}{H}\right) \mathbf{d}\boldsymbol{\theta}. \tag{7.89}$$

Exterior differentiation yields

$$\mathbf{d}\boldsymbol{\omega}^{\hat{l}} = 0, \quad \mathbf{d}\boldsymbol{\omega}^{\hat{\theta}} = \frac{1}{l}\boldsymbol{\omega}^{\hat{l}} \wedge \boldsymbol{\omega}^{\hat{\theta}}. \tag{7.90}$$

Cartan's first structural equation now give the non-vanishing connection forms

$$\boldsymbol{\Omega}^{\hat{\theta}}{}_{\hat{l}} = -\boldsymbol{\Omega}^{\hat{l}}{}_{\hat{\theta}} = \frac{1}{l}\boldsymbol{\omega}^{\hat{\theta}}. \tag{7.91}$$

Thus

$$\mathbf{d}\boldsymbol{\Omega}^{\hat{\theta}}{}_{\hat{l}} = \left(\mathbf{d}\frac{1}{l}\right)\boldsymbol{\omega}^{\hat{\theta}} + \frac{1}{l}\mathbf{d}\boldsymbol{\omega}^{\hat{\theta}} = 0. \tag{7.92}$$

and Cartan's second structural equation gives

$$\mathbf{R}^{\hat{\theta}}{}_{\hat{l}} = \mathbf{R}^{\hat{l}}{}_{\hat{\theta}} = 0 \tag{7.93}$$

which shows that the intrinsic geometry of the conical surface is Euclidean.

The extrinsic curvature is given by eq. (7.75). In the present case the single non-vanishing component of this tensor is

$$K_{\hat{\theta}\hat{\theta}} = \Gamma^{\hat{n}}{}_{\hat{\theta}\hat{\theta}}. \tag{7.94}$$

In example 6.9 we found that $\Gamma^{\hat{r}}{}_{\hat{\theta}\hat{\theta}} = -\frac{1}{r}$, so

$$\mathbf{d}\mathbf{e}_{\hat{\theta}} = -\frac{1}{r}\mathbf{e}_{\hat{r}} \otimes \boldsymbol{\omega}^{\hat{\theta}} \tag{7.95}$$

From Fig.7.4 we see that

$$\mathbf{e}_{\hat{r}} = \frac{H}{L}\mathbf{e}_{\hat{n}} + \frac{R}{L}\mathbf{e}_{\hat{l}}, \quad L = (H^2 + L^2)^{\frac{1}{2}} \tag{7.96}$$

This yields

$$\mathbf{d}\mathbf{e}_{\hat{\theta}} = -\frac{H}{Lr}\mathbf{e}_{\hat{n}} \otimes \boldsymbol{\omega}^{\hat{\theta}} - \frac{R}{Lr}\mathbf{e}_{\hat{l}} \otimes \boldsymbol{\omega}^{\hat{\theta}} \tag{7.97}$$

which shows that

$$\Gamma^{\hat{n}}{}_{\hat{\theta}\hat{\theta}} = -\frac{H}{Lr} \tag{7.98}$$

Hence, the conical surface has a non-vanishing extrinsic curvature given by

$$K_{\hat{\theta}\hat{\theta}} = -\left(1 + \frac{R^2}{H^2}\right)^{-\frac{1}{2}} \cdot \frac{1}{r} \tag{7.99}$$

The limit $H \longrightarrow \infty, r \longrightarrow R$ represents a straight cylinder with $K_{\hat{\theta}\hat{\theta}} = -\frac{1}{R}$. The limit $H \longrightarrow 0$ represents a plane with $K_{\hat{\theta}\hat{\theta}} = 0$.

7.5 The equation of geodesic deviation

Consider two nearby geodesic curves, both parametrised by a parameter λ. Let s be a vector connecting points on the two curves with the same value of λ. The connecting vector s is said to measure the *geodesic deviation* of the curves.

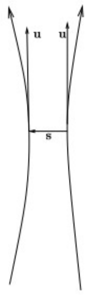

Figure 7.5: The two solid lines are neighbouring geodesics. They are connected by an infinitesimal vector s that obeys the equation of geodesic deviation

In order to deduce an equation describing how the geodesic deviation varies along the curves, we consider the covariant directional derivative of s along the curve $\nabla_{\mathbf{u}}\mathbf{s}$ where $\mathbf{u}$ is the tangent vector to the curve.

Let $\mathbf{u}$ and s be coordinate basis vectors of a coordinate system. Then $[\mathbf{s},\mathbf{u}] = 0$, so that

$$\nabla_{\mathbf{u}}\mathbf{s} = \nabla_{\mathbf{s}}\mathbf{u} \tag{7.100}$$

giving

$$\nabla_{\mathbf{u}}\nabla_{\mathbf{u}}\mathbf{s} = \nabla_{\mathbf{u}}\nabla_{\mathbf{s}}\mathbf{u}. \tag{7.101}$$

Furthermore

$$\begin{aligned}\mathbf{R}(\mathbf{u},\mathbf{s})\mathbf{u} &= \left([\nabla_{\mathbf{u}},\nabla_{\mathbf{s}}] - \nabla_{[\mathbf{u},\mathbf{s}]}\right)\mathbf{u} \\ &= [\nabla_{\mathbf{u}},\nabla_{\mathbf{s}}]\,\mathbf{u}.\end{aligned} \tag{7.102}$$

Thus

$$\nabla_{\mathbf{u}}\nabla_{\mathbf{u}}\mathbf{s} = \nabla_{\mathbf{s}}\nabla_{\mathbf{u}}\mathbf{u} + \mathbf{R}(\mathbf{u},\mathbf{s})\mathbf{u}. \tag{7.103}$$

Since the curves are geodesics $\nabla_{\mathbf{u}}\mathbf{u} = 0$, and the equation reduces to

$$\boxed{\nabla_{\mathbf{u}}\nabla_{\mathbf{u}}\mathbf{s} + \mathbf{R}(\mathbf{s},\mathbf{u})\mathbf{u} = 0} \tag{7.104}$$

where we have used the antisymmetry of the Riemann tensor. Equation (7.104) is called the *equation of geodesic deviation*. The component form of the equation is

$$\left(\frac{d^2\mathbf{s}}{d\lambda^2}\right)^{\mu} + R^{\mu}{}_{\alpha\nu\beta}u^{\alpha}s^{\nu}u^{\beta} = 0. \tag{7.105}$$

This equation shows that the Riemann tensor can be determined entirely from measurements of geodesic deviation.

In comoving geodesic normal coordinates with $\mathbf{u} = (1,0,0,0)$ the equation reduces to

$$\left(\frac{d^2\mathbf{s}}{d\lambda^2}\right)^{i} + R^{i}{}_{0j0}s^{j} = 0. \tag{7.106}$$

7.6 Spaces of constant curvature

We have in section 6.11 seen how important Killing vectors may be for the solvability of the equations that govern particle motion. Highly symmetric spaces are therefore important both in mathematics and in physics. We will study maximally symmetric spaces with a Riemannian metric, S^n, $\mathbb{P}^n$, $\mathbb{E}^n$ and $\mathbb{H}^n$. The first 2 of these differ only at a global scale, that is they are locally the same but their topology is different. Maybe the mathematically most interesting one is the hyperbolic space $\mathbb{H}^n$. We will start out with the most familiar one, the Euclidean space $\mathbb{E}^n$.

The Euclidean space, $\mathbb{E}^n$

We are all familiar with this space. The metric can be written

$$ds^2 = g_{ab}\mathbf{dx}^a \otimes \mathbf{dx}^b \tag{7.107}$$

where $g_{ab} = \delta_{ab}$. We can easily generalize this to the Minkowskian case by letting the components of the metric tensor be of either sign, $g_{ab} = \pm\delta_{ab}$. As claimed, this is a maximally symmetric space, thus having $\frac{1}{2}n(n+1)$ linearly independent Killing vectors. We should stress that when we say linearly independent in this context, we mean linearly independent *solutions* of Killing's equations. The Killing vectors are not linearly independent in the space itself. Killing's equations do in this case reduce to

$$\xi_{i,j} + \xi_{j,i} = 0 \tag{7.108}$$

because the connection coefficients all vanish identically. We can easily find n linearly independent solutions to this equation,

$$\boldsymbol{\xi}_{(a)} = \frac{\partial}{\partial x^b}. \tag{7.109}$$

Another set of solutions are[1]

$$\boldsymbol{\xi}_{(ab)} = x^b\frac{\partial}{\partial x^a} - x^a\frac{\partial}{\partial x^b} \tag{7.110}$$

The solutions $\boldsymbol{\xi}_{(ab)}$ are antisymmetric in the indices so they represent $\frac{1}{2}n(n-1)$ linearly independent solutions. Thus all in all we have found $n + \frac{1}{2}n(n-1) = \frac{1}{2}n(n+1)$ linearly independent solutions of Killing's equation. Hence, as claimed, the Euclidean spaces admit $\frac{1}{2}n(n+1)$ linearly independent Killing vectors and is therefore maximally symmetric.

The one-parameter group of diffeomorphisms associated to these Killing vectors can be found by solving the equations

$$\frac{d}{dt}\phi_t = \boldsymbol{\xi}. \tag{7.111}$$

Representing ϕ_t by a vector $\mathbf{V}$, the equations for $\boldsymbol{\xi}_{(a)}$ can be written

$$\frac{d}{dt}\mathbf{V}_{(a)} = \mathbf{e}_a. \tag{7.112}$$

It has solutions

$$\mathbf{V}_{(a)} = \mathbf{V}_0 + t\mathbf{e}_a \tag{7.113}$$

where $\mathbf{V}_0$ is a constant vector corresponding to the initial condition. These mappings are mere translations a distance t in the a-direction. The Euclidean spaces have translation invariance.

Defining $\mathsf{J}_{(ab)}$ as the antisymmetric matrix with indices

$$\left(\mathsf{J}_{(ab)}\right)_{ij} = \delta_{ai}\delta_{bj} - \delta_{bi}\delta_{aj} \tag{7.114}$$

---

[1] Since we use Cartesian coordinates we have $x_i \equiv g_{ij}x^j = x^i$. Hence, the position of the index does not matter.

we can write for the Killing vectors $\boldsymbol{\xi}_{(ab)}$

$$\frac{d}{dt}\mathbf{V}_{(ab)} = \mathsf{J}_{(ab)}\mathbf{V}_{(ab)}. \tag{7.115}$$

This equation have solutions

$$\mathbf{V}_{(ab)} = e^{t\mathsf{J}_{(ab)}} \cdot \mathbf{V}_0 \tag{7.116}$$

where

$$e^{t\mathsf{J}_{(ab)}} = 1 + t\mathsf{J}_{(ab)} + \frac{1}{2}t^2\mathsf{J}^2_{(ab)} + ... = \mathsf{R}_{(ab)}(t) \tag{7.117}$$

is the rotation matrix through an angle t with respect to the (ab)-plane. Showing this is left as an exercise, see the problems in the end of this chapter. Thus, the Euclidean spaces are also rotationally symmetric. The three-dimensional Euclidean space has 6 linear independent Killing vectors, representing translation and rotational invariance. There is one significant difference between these operations. While the translations move "the whole space", the rotations leave an axis fixed. For a point p, the subgroup of the isometry group that leaves p fixed, is called *the isotropy subgroup*. In the case of $\mathbb{E}^3$, the isotropy subgroup of any point is the group of rotations with respect to this point. We will come back to these concepts in a later chapter, where we will define these groups more rigorously.

The elliptic spaces, S^n and $\mathbb{P}^n$

Let us first start out by defining the spheres S^n. Consider the $(n+1)$-dimensional Euclidean space and look at the hypersurface

$$(x^1)^2 + (x^2)^2 + ... + (x^{n+1})^2 = 1. \tag{7.118}$$

This hypersurface with the induced metric is the sphere S^n. We claim now

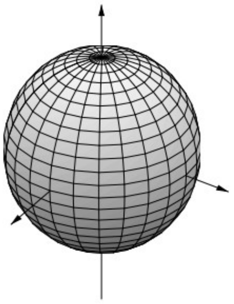

Figure 7.6: The two-dimensional sphere embedded in a Euclidean space

that the Killing vectors of the sphere is the Killing vectors from the Euclidean space that generate maps that map the sphere onto itself. These are the rotations with respect to the origin, with corresponding Killing vectors

$$\boldsymbol{\xi}_{(\alpha\beta)} = x^{\beta}\frac{\partial}{\partial x^{\alpha}} - x^{\alpha}\frac{\partial}{\partial x^{\beta}} \tag{7.119}$$

where Latin indices run from 1 to $(n+1)$. Here we have $\frac{1}{2}n(n+1)$ of them, and when constrained to the sphere, they form the Killing vectors of the sphere. Thus this space is maximally symmetric. Let us use Gauss' Theorema Egregium to calculate its curvature properties. Given a point on the sphere, then for one of the coordinates $x^a \neq 0$. Assume that $x^{n+1} \neq 0$. Then we can define the following basis vectors

$$\mathbf{e}_a = \boldsymbol{\xi}_{(a(n+1))} = x^{n+1}\frac{\partial}{\partial x^a} - x^a\frac{\partial}{\partial x^{n+1}} \tag{7.120}$$

where Greek indices run from 1 to n. The metric in this basis is

$$\begin{aligned} g_{ab} &= \mathbf{e}_a \cdot \mathbf{e}_b = \left(x^{n+1}\frac{\partial}{\partial x^a} - x^a\frac{\partial}{\partial x^{n+1}}\right)\cdot\left(x^{n+1}\frac{\partial}{\partial x^b} - x^b\frac{\partial}{\partial x^{n+1}}\right) \\ &= x^a x^b + \left(x^{n+1}\right)^2\delta_{ab}. \end{aligned} \tag{7.121}$$

The radial vector

$$\mathbf{e}_r = x^\alpha\frac{\partial}{\partial x^\alpha} \tag{7.122}$$

has unit length on the sphere. Using this we can calculate the components of the extrinsic curvature

$$\begin{aligned} K_{ab} &= \mathbf{e}_r \cdot (\mathbf{d}\mathbf{e}_a(\mathbf{e}_b)) = \mathbf{e}_r \cdot \left[\frac{\partial}{\partial x^a}\left(\mathbf{d}\mathbf{x}^{n+1}(\mathbf{e}_b)\right) - \frac{\partial}{\partial x^{n+1}}\left(\mathbf{d}\mathbf{x}^a(\mathbf{e}_b)\right)\right] \\ &= \mathbf{e}_r \cdot \left[x^b\frac{\partial}{\partial x^a} + \delta_{ab}(x^{n+1})\frac{\partial}{\partial x^{n+1}}\right] = x^a x^b + \left(x^{n+1}\right)^2\delta_{ab}. \end{aligned} \tag{7.123}$$

Thus we have

$$K_{ab} = g_{ab}. \tag{7.124}$$

The Gauss' Theorema Egregium now gives us the Riemann tensor

$$R_{abcd} = g_{ac}g_{bd} - g_{ad}g_{bc}. \tag{7.125}$$

Contracting once,

$$R_{bd} = (n-1)g_{bd} \tag{7.126}$$

we see that the Ricci tensor is proportional to the metric[2]. Thus in an orthonormal frame, the Ricci tensor will be everywhere positive and constant. Contracting once more to obtain the Ricci scalar

$$R = n(n-1). \tag{7.127}$$

The projective space $\mathbb{P}^n$ can now be obtained by identifying opposite points on the sphere S^n. The projective space is therefore basically half of the sphere. There is a pathology however, for n even, the projective space is *non-orientable*. That means there is no globally defined unit normal vector field on the space. However, for n odd (including therefore dimension 3), the space is orientable, and no problem of such kind exists.

---

[2] More generally, spaces for which the Ricci tensor can be written $R_{ab} = \lambda g_{ab}$ where λ is a constant, are called *Einstein spaces*. All the constant curvature spaces are Einstein spaces.

Let us derive a useful form of the metric of the sphere. We introduce a radial coordinate r by

$$r^2 = \sum_a (x^a)^2 \tag{7.128}$$

so that

$$x^{n+1} = \pm\sqrt{1-r^2}. \tag{7.129}$$

Hence,

$$\left(dx^{n+1}\right)^2 = \frac{r^2dr^2}{1-r^2}. \tag{7.130}$$

Assuming now that $d\Omega_{n-1}$ is the metric on the $(n-1)$-dimensional sphere, S^{n-1}, then

$$\sum_a (dx^a)^2 = dr^2 + r^2 d\Omega_{n-1}^2 \tag{7.131}$$

which is just the expression for the Euclidean metric in spherical coordinates. The metric on the sphere, S^n, is now

$$ds^2 = \left(dx^{n+1}\right)^2 + \sum_a (dx^a)^2 = \frac{dr^2}{1-r^2} + r^2 d\Omega_{n-1}^2. \tag{7.132}$$

Note that this metric only covers half of the sphere. For the case $\mathbb{P}^n$ it covers the whole space, except for the points on the equator which forms a set of measure zero.

The Hyperbolic spaces, $\mathbb{H}^n$

As for the sphere, the hyperbolic space can be viewed upon as a hypersurface in a flat $(n+1)$-dimensional space. But now we have to use the flat $(n+1)$-dimensional Minkowski space with metric

$$ds^2 = \eta_{\alpha\beta}\mathbf{dx}^\alpha \otimes \mathbf{dx}^\beta = -\mathbf{dx}^{n+1} \otimes \mathbf{dx}^{n+1} + \sum_a \mathbf{dx}^a \otimes \mathbf{dx}^a. \tag{7.133}$$

The hyperbolic space is defined as the hyperboloid

$$-(x^{n+1})^2 + (x^1)^2 + ... + (x^n)^2 = -1, \quad x^{n+1} > 0. \tag{7.134}$$

This surface is space-like and has a Riemannian metric. Its symmetries can be found by using the same argument as for the sphere. The symmetries for the Minkowski space that leave the hyperboloid invariant, are the Lorentz transformations in $(n+1)$ dimensions. The Killing vectors for the Lorentz transformations come in two classes, boosts

$$\boldsymbol{\xi}_{(a)} = x^{n+1}\frac{\partial}{\partial x^a} + x^a\frac{\partial}{\partial x^{n+1}} \tag{7.135}$$

and rotations

$$\boldsymbol{\xi}_{(ab)} = x^b\frac{\partial}{\partial x^a} - x^a\frac{\partial}{\partial x^b}. \tag{7.136}$$

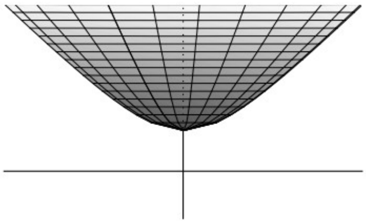

Figure 7.7: The Hyperbolic plane embedded in three-dimensional Minkowski space

As for the sphere we can now choose a set of basis vectors by

$$\mathbf{e}_a = \boldsymbol{\xi}_{(a)} \tag{7.137}$$

but these will now be globally defined since $x^{n+1} \geq 1$. The metric in this basis is

$$\begin{aligned} g_{ab} &= \mathbf{e}_a \cdot \mathbf{e}_b = \left(x^{n+1} \frac{\partial}{\partial x^a} + x^a \frac{\partial}{\partial x^{n+1}} \right) \cdot \left(x^{n+1} \frac{\partial}{\partial x^b} + x^b \frac{\partial}{\partial x^{n+1}} \right) \\ &= -x^a x^b + \left(x^{n+1} \right)^2 \delta_{ab} \end{aligned} \tag{7.138}$$

since $\frac{\partial}{\partial x^{n+1}} \cdot \frac{\partial}{\partial x^{n+1}} = -1$. The radial vector

$$\mathbf{e}_r = x^\alpha \frac{\partial}{\partial x^\alpha} \tag{7.139}$$

has on the hyperboloid a length

$$\mathbf{e}_r \cdot \mathbf{e}_r = -(x^{n+1})^2 + \sum_a (x^a)^2 = -1. \tag{7.140}$$

Following an almost identical procedure to the case of the sphere, we can calculate the extrinsic curvature

$$\begin{aligned} K_{ab} &= \mathbf{e}_r \cdot (\mathbf{d}\mathbf{e}_a(\mathbf{e}_b)) \\ &= g_{ab}. \end{aligned} \tag{7.141}$$

The Gauss's Theorema Egregium now turns into (we have to choose the negative sign due to the negative length of the normal unit vector)

$$R_{abcd} = -(g_{ac} g_{bd} - g_{ad} g_{bc}). \tag{7.142}$$

Contracting once,

$$R_{bd} = -(n-1) g_{bd}. \tag{7.143}$$

This space has negative curvature, and basically just the same curvature properties as for the sphere, just with an opposite sign of the curvature. A metric on the hyperbolic space can written by the following. Introduce spherical coordinates,

$$r^2 = \sum_a (x^a)^2 \tag{7.144}$$

and following the procedure as for the sphere, the metric can be written

$$ds^2 = \frac{dr^2}{1+r^2} + r^2 d\Omega_{n-1}^2 \tag{7.145}$$

where $d\Omega_{n-1}^2$ is the metric on the $(n-1)$-dimensional sphere. In this case the metric covers the whole of $\mathbb{H}^n$.

As we see, this space has infinite volume in contrast to the sphere. Note that there exist (and a lot of them) compact hyperbolic spaces as well, but these breaks the isometries for the $\mathbb{H}^n$ at a global scale. Locally, however, they have a hyperbolic metric and are locally isometric to $\mathbb{H}^n$.

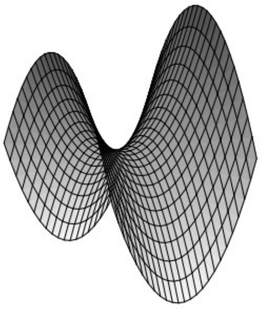

Figure 7.8: The hyperbolic plane can be seen upon as a saddle-surface

A more intuitive view of the hyperbolic plane is by considering it as the surface of a saddle, Fig. 7.8. A saddle surface in Euclidean three-space has a negative curvature and has therefore much of the same properties of the constantly curved hyperbolic plane. We should emphasize that the whole hyperbolic plane given by eq. (7.134) for $n = 2$ cannot be embedded in the three-dimensional Euclidean space. So the surface in Fig. 7.8 has not a constant curvature.

We can find a constant negatively curved space in Euclidean space. Consider the surface generated by the rotation around the z-axis of the *tractrix*

$$z = \ln\left|\frac{1 \pm \sqrt{1-r^2}}{r}\right| \mp \sqrt{1-r^2} \tag{7.146}$$

where $r^2 = x^2 + y^2$. This space will have constant negative curvature and are sometimes called a *pseudo-sphere*. The tractrix and the pseudo-sphere are depicted in Fig.7.9.

The name tractrix is due to the fact that this is the curve which is traced out by an object on the end of a rope of unit length held by a running child along the z-axis. If the object and the child start in the xz-plane at $(1, 0)$ and $(0, 0)$, respectively, then the curve is precisely the curve eq. (7.146) with $y = 0$. The points along the circle $r = 1$ correspond to a singular line, and hence, this cannot be the whole hyperbolic space.

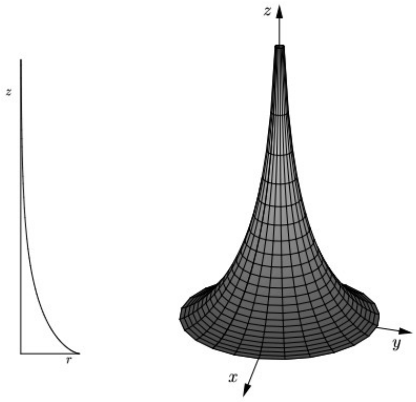

(a) The tractrix (b) The pseudo-sphere

Figure 7.9: The tractrix and the pseudo-sphere. The pseudo-sphere is obtained by rotating the tractrix around the z-axis.

Problems

7.1. *Rotation matrices*
Show that

$$e^{t\mathsf{J}_{(ab)}} = 1 + t\mathsf{J}_{(ab)} + \frac{1}{2}t^2\mathsf{J}_{(ab)}^2 + \dots = \mathsf{R}_{(ab)}(t), \tag{7.147}$$

i.e. verify eq. (7.117).

7.2. *Inverse metric on S^n*
In this problem we will use the basis vectors $\mathbf{e}_a$ defined in eq. (7.120):

$$\mathbf{e}_a = \boldsymbol{\xi}_{(a(n+1))} = x^{n+1}\frac{\partial}{\partial x^a} - x^a\frac{\partial}{\partial x^{n+1}}. \tag{7.148}$$

(a) Assume that the basis forms on the sphere S^n has the form

$$\boldsymbol{\omega}^a = C^a_{\ \mu}\mathbf{dx}^\mu \tag{7.149}$$

Find the coefficients $C^a_{\ \mu}$.

(b) Find g^{ab}, and verify that $g^{ab}g_{bc} = \delta^a_{\ c}$.

7.3. *The curvature of a curve*
Consider a curve $y = y(x)$ in a two-dimensional plane. Utilize that the differential of the arc length is $ds = \left(1+y'^2\right)^{1/2}dx$ and show that the tangent vector, the curvature vector and the curvature of the curve are given by

$$\begin{aligned} \mathbf{t} &= \left(1+y'^2\right)^{-1/2}\left(\mathbf{e}_x + y'\mathbf{e}_y\right), \\ \mathbf{k} &= \left(1+y'^2\right)^{-2}y''\left(y'\mathbf{e}_x + \mathbf{e}_y\right), \\ \kappa &= \left(1+y'^2\right)^{-3/2}y'', \end{aligned} \tag{7.150}$$

respectively.

7.4. *The Gauss-Codazzi equations*
The Gauss-Codazzi equations are of great importance and tells us much about the geometry of hypersurfaces. We will need these equations for a later chapter, and in this problem we will show an important relation between the Ricci curvatures in the ambient space and on the hypersurface.

(a) We will first show that the projection operator $h^{\alpha}_{\ \beta}$, defined by eq. (7.81) is what it is claimed to be. We define $h_{\alpha\beta}$ with eq. (7.81):

$$g_{\alpha\beta} = h_{\alpha\beta} + \epsilon n_{\alpha} n_{\beta} \tag{7.151}$$

where $g_{\alpha\beta}$ is the metric in the ambient space, and n_{α} is the components of the unit normal vector. The operator $h^{\alpha}_{\ \beta}$ is now defined by $h^{\alpha}_{\ \beta} = g^{\alpha\lambda} h_{\lambda\beta}$. Show the following:

$$\begin{aligned} h_{\alpha\beta} &= h_{\beta\alpha} \\ h^{\alpha}_{\ \lambda} h^{\lambda}_{\ \beta} &= h^{\alpha}_{\ \beta}. \end{aligned}$$

These properties defines a projection map. We now have to show that it projects onto the space that is orthogonal to n^{α}. Hence, if v^{β} is a vector, we have to show that

$$\begin{aligned} h^{\alpha}_{\ \beta} v^{\beta} &= 0, && \text{iff } v^{\beta} \text{ is parallel to } n^{\beta}. \\ h^{\alpha}_{\ \beta} v^{\beta} &= v^{\alpha}, && \text{iff } v^{\beta} \text{ is orthogonal to } n^{\beta}. \end{aligned}$$

Prove this. Explain now we can define the projection of an arbitrary tensor by projection the tensor, index by index.

(b) Verify the Codazzi equation, eq. (7.84). (Hint: Calculate $\left(\mathbf{d}^2\mathbf{e}_{\hat{b}}\right) \cdot \mathbf{n}$.)

(c) Show that

$$^{(n+1)}R = {}^{(n)}R - \epsilon\left(K^2 - K^{\alpha\beta}K_{\alpha\beta}\right) + 2\epsilon^{(n+1)}R_{\alpha\beta}n^{\alpha}n^{\beta}, \tag{7.152}$$

which can be written

$$-2\epsilon^{(n+1)}E_{\alpha\beta}n^{\alpha}n^{\beta} = {}^{(n)}R - \epsilon\left(K^2 - K^{\alpha\beta}K_{\alpha\beta}\right). \tag{7.153}$$

7.5. *The Poincaré half-space*
Consider half of $\mathbb{R}^3$, $z > 0$ with metric

$$ds^2 = \frac{1}{z^2}\left(dx^2 + dy^2 + dz^2\right) \tag{7.154}$$

(a) Calculate the connection forms and the curvature forms using the structural equations of Cartan.

(b) Calculate the Riemann tensor, the Ricci tensor and the Ricci scalar.

(c) Show that

$$R_{abcd} = -(g_{ac}g_{bd} - g_{ad}g_{bc}) \tag{7.155}$$

Compare this with the three-dimensional hyperbolic space. Are there any way we can differentiate between these two cases? Are they different manifestations of the same space?

7.6. *The pseudo-sphere*
Show that the tractrix, eq. (7.146), obeys the differential equation

$$\left(\frac{dr}{dz}\right)^2 = \frac{r^2}{1-r^2}. \tag{7.156}$$

Substitute this into the line-element for flat space, do the substitution $r = \sin\theta$, and show that the metric on the pseudo-sphere can be written

$$ds^2 = d\theta^2 + \sinh^2\theta d\phi^2. \tag{7.157}$$

Show that this metric has constant negative curvature.

7.7. *A non-Cartesian coordinate system in two dimensions*
Consider the following metric on a two-dimensional surface:

$$ds^2 = v^2 du^2 + u^2 dv^2. \tag{7.158}$$

We will show, in two different ways, that this is only the flat Euclidean plane in disguise.

(a) Use the orthonormal frame approach and find the connection one-forms $\mathbf{\Omega}^{\hat{a}}_{\ \hat{b}}$. Find also the curvature two-forms $\mathbf{R}^{\hat{a}}_{\ \hat{b}}$ and show that they are identically zero.

(b) Show that the metric can be put onto the form

$$ds^2 = dx^2 + dy^2$$

by finding a transformation matrix $\mathsf{M} = \left(M^i_{\ a}\right)$ connecting the basis vectors $\mathbf{e}_u$ and $\mathbf{e}_v$, and $\mathbf{e}_x$ and $\mathbf{e}_y$. This can be done using the following relations

$$\begin{aligned} g_{ab} &= g_{ij} M^i_{\ a} M^j_{\ b} \\ \frac{\partial M^i_{\ a}}{\partial x^b} &= \frac{\partial M^i_{\ b}}{\partial x^a}. \end{aligned} \tag{7.159}$$

Where do these relations come from?

7.8. *The curvature tensor of a sphere*
Introduce an orthonormal basis on the sphere, S^2, and use Cartan's structural equations to find the physical components of the Riemann curvature tensor.

7.9. *The curvature scalar of a surface of simultaneity*
The spatial line-element of a rotating disc is

$$d\ell^2 = dr^2 + \frac{r^2}{1 - \frac{\omega^2 r^2}{c^2}} d\phi^2. \tag{7.160}$$

Introduce an orthonormal basis on this surface and use Cartan's structural equations to find the Ricci scalar.

7.10. *The tidal force pendulum and the curvature of space*
We will again consider the tidal force pendulum from Example 1.3. Here we shall use the equation for geodesic deviation, eq. (7.104), to find the period of the pendulum.

(a) Why can the equation for geodesic equation be used to find the period of the pendulum in spite of the fact that the particles do not move along geodesics? Explain also why the equation can be used even though the centre of the pendulum does not follow a geodesic.

(b) Assume that the centre of the pendulum is fixed at a distance R from the centre of mass of the Earth. Introduce an orthonormal basis $\{\mathbf{e}_{\hat{a}}\}$ with the origin at the centre of the pendulum (see Fig.7.10).

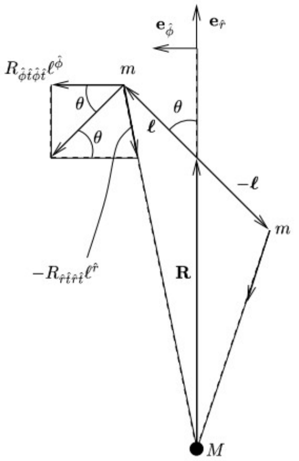
$\mathbf{e}_{\hat{\phi}}$
$\mathbf{e}_{\hat{r}}$
$R_{\hat{\phi}\hat{t}\hat{\phi}\hat{t}}\ell^{\hat{\phi}}$
m
θ
θ
θ
$\boldsymbol{\ell}$
θ
$-\boldsymbol{\ell}$
m
$-R_{\hat{r}\hat{t}\hat{r}\hat{t}}\ell^{\hat{r}}$
$\mathbf{R}$
M

Figure 7.10: The tidal force pendulum

Show that, to first order in v/c and ϕ/c^2, where v is the three-velocity of the masses and ϕ the gravitational potential at the position of the pendulum, that the equation of geodesic equation takes the form

$$\frac{d^2\ell_{\hat{i}}}{dt^2} + R_{\hat{i}\hat{0}\hat{j}\hat{0}}\ell^{\hat{j}} = 0. \tag{7.161}$$

(c) Find the period of the pendulum expressed in terms of the components of Riemann's curvature tensor.

7.11. *The Weyl tensor vanishes for spaces of constant curvature*
Use the definition eq. (7.72) to show that the constant curvature spaces S^4, $\mathbb{E}^4$, and $\mathbb{H}^4$ all have zero Weyl tensor.

20210825/ Quantum/

general-relativity.pdf

Pág. 149(162) - Chapter 7

Curvature

Pág. 152(165) Figure 7.2

A two-dimensional surface embedded in three-dimensional Euclidean space.

Pág. 153(166) 7.3

The Riemann Curvature Tensor

Pág. 158(171)

The Weyl Curvature Tensor

Pág. 159(172) 7.4

Extrinsic and Intrinsic Curvature

Pág. 162(175) - 7.5

The Equation of geodesic deviation

( )
+,- * *. ! * "9' " 9' 1#
) &" '

Somit gilt

$$ax^2 + bx + c = a\left[\left(x + \frac{b}{2a}\right)^2 - \left(\frac{b}{2a}\right)^2\right] + c = a\left(x + \frac{b}{2a}\right)^2 - a\left(\frac{b}{2a}\right)^2 + c$$

und folglich

$$d = \frac{b}{2a} \quad \text{sowie} \quad e = -a \cdot \left(\frac{b}{2a}\right)^2 + c = c - ad^2 \,.$$

Damit haben wir die Idee des folgenden Algorithmus erklärt.

Algorithmus 1.2 (Quadratische Ergänzung II)
Überführung des Terms

$$ax^2 + bx + c$$

in die Form

$$a(x+d)^2 + e$$

1. Berechne

$$d = \frac{b}{2a} \,.$$

2. Berechne

$$e = c - ad^2 \,.$$

Beispiel 1.14
Wir überführen den Term

$$ax^2 + bx + c = 9x^2 + 36x + 12$$

in die Form $a(x+d)^2 + e$. Gemäß Algorithmus 1.2 berechnen wir mit $a = 9$, $b = 36$ und $c = 12$

$$d = \frac{b}{2a} = \frac{36}{2 \cdot 9} = 2 \quad \text{und} \quad e = c - ad^2 = 12 - 9 \cdot 2^2 = -24 \,.$$

Damit ist

$$9x^2 + 36x + 12 = 9(x+2)^2 - 24 \,.$$

■

1.6 Die irrationalen Zahlen π und e

In diesem Abschnitt beschäftigen wir uns mit zwei irrationalen Zahlen, die in der Mathematik eine besonders große Rolle spielen. Dies sind die Kreiszahl

$$\pi = 3.141592653589793238\ldots \tag{1.79}$$

und die Eulersche Zahl

$$\mathrm{e} = 2.718281828459045235\ldots \qquad (1.80)$$

Die Zahl π tritt beispielsweise bei der Definition des Bogenmaßes von Winkeln (siehe Abschnitt 3.2), bei der Berechnung des Flächeninhalts ebener Flächen und Flächen im Raum, bei Volumenberechnungen von Körpern sowie der Berechnung der Länge von Kurven auf. Aus der Schulmathematik sind sicher noch die Formeln zur Berechnung des Flächeninhalts F und des Umfangs U eines Kreises mit dem Durchmesser d bekannt, nämlich

$$F = \frac{\pi}{4}\,d^2 \quad \text{und} \quad U = \pi d\,.$$

Die Zahl e benötigt man im Zusammenhang mit der sogenannten e-Funktion, mit welcher man z. B. Wachstumsprozesse mathematisch beschreiben kann (siehe Beispiel 3.29 im Abschnitt 3.2). Sie tritt auch auf, wenn man den Durchhang eines nur durch sein Eigengewicht belasteten Seils (siehe Beispiel 3.30) berechnet oder spiralförmige Kurven mathematisch beschreibt.

Ausgehend von der Formel für den Umfang eines Kreises mit dem Radius d kann man einen Näherungswert für π folgendermaßen ermitteln. Man markiert auf einem Kreis mit dem Durchmesser $d = 1$, der eine Gerade berührt, den Punkt in welchem die Berührung erfolgt. Dann lässt man den Kreis auf dieser Geraden solange abrollen bis der Markierungspunkt auf dem Kreis wieder Berührungspunkt des Kreises mit der Geraden ist (siehe Abbildung 1.1). Der Abstand zwischen den beiden Berührungspunkten auf der Geraden ist gleich π, denn bei einem Kreis mit dem Durchmesser $d = 1$ ist der Umfang U gleich π.

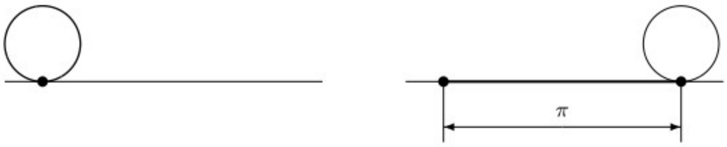

Abb. 1.1 Bestimmung von π

Im Folgenden geben wir einige Möglichkeiten an, wie man Näherungswerte für π und e berechnen kann. Unser Ziel ist es dabei, Näherungswerte auf 15 Stellen (14 Nachkommastellen) genau anzugeben. Dies ist die Genauigkeit in der Zahlendarstellung, welche man hat, wenn man in den üblichen Programmiersprachen wie z. B. der Programmiersprache C doppelt genaue Zahlen (double precision) nutzt (siehe auch Bemerkung 5.4 im Abschnitt 5.2).

Archimedes[3] ermittelte Näherungswerte für π, indem er Vielecke (6-Eck, 12-Eck, …, 96-Eck) in einen Kreis einbeschrieb bzw. den Kreis mit den Vielecken umschrieb. Dabei erhielt er mittels der Umfänge der Vielecke die Einschachtelung

$$\boxed{3+\frac{10}{71}=\frac{223}{71}\approx 3.140845<\pi<3+\frac{10}{70}=\frac{22}{7}\approx 3.142857\,.}$$

Ein Vergleich mit (1.79) zeigt, dass man so die Zahl π mit zwei Nachkommastellen Genauigkeit erhält.

Im Jahr 1593 entwickelte Viète[4] die folgende Formel mit unendlich vielen Faktoren zur Darstellung von $\frac{2}{\pi}$.

$$\frac{2}{\pi}=\frac{\sqrt{2}}{2}\cdot\frac{\sqrt{2+\sqrt{2}}}{2}\cdot\frac{\sqrt{2+\sqrt{2+\sqrt{2}}}}{2}\cdot\frac{\sqrt{2+\sqrt{2+\sqrt{2+\sqrt{2}}}}}{2}\cdot\ldots\,. \tag{1.81}$$

Berechnet man nur das Produkt der ersten n Faktoren (n eine natürliche Zahl, welche man sich vorgibt), dann erhält man einen Näherungswert für $\frac{2}{\pi}$ und daraus einen Näherungswert für π. In der folgenden Tabelle 1.1 geben wir an, wieviele Faktoren man im obigen Produkt benötigt, um einen Näherungswert für π mit k exakten Stellen zu bestimmen:

Tab. 1.1 Unter Nutzung der Formel von Viète berechnete Näherungswerte für π

Näherungswert für π	Anzahl n der Faktoren in (1.81)	Anzahl k der exakten Stellen
3.061467458920718	2	1
3.121445152258051	3	2
3.140331156954752	5	3
3.141277250932772	6	4
3.141513801144300	7	5
3.141591421511199	10	6
3.141592345570117	11	7
3.141592634338562	13	8
3.141592652386590	15	9
3.141592653288992	16	10
3.141592653514593	17	11
3.141592653585093	19	12
3.141592653589500	21	13
3.141592653589720	22	14
3.141592653589792	25	15

---

[3] Archimedes von Syrakus, geb. um 287 v. Chr., gest. 212 v. Chr. vermutlich in Syrakus (Sizilien), griechischer Mathematiker, Physiker und Ingenieur

[4] François Viète, latinisiert Franciscus Vieta, geb. 1540 in Fontenay-le-Comte, gest. 1603 in Paris, französischer Advokat und Mathematiker

Wallis[5] entwickelte 1655 die folgende Darstellungsformel für $\frac{\pi}{2}$:

$$\boxed{\frac{\pi}{2} = \prod_{k=1}^{\infty} \frac{2k}{2k-1} \cdot \frac{2k}{2k+1} = \frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdot \frac{6}{7} \cdot \dots \,. \qquad (1.82)}$$

Nutzt man nur die ersten n Faktoren im obigen Produkt erhält man Näherungswerte für $\frac{\pi}{2}$ und damit für π (siehe Tabelle 1.2).

Tab. 1.2 Unter Nutzung der Formel von Wallis berechnete Näherungswerte für π

Näherungswert für π	Anzahl n der Faktoren in (1.82)	Anzahl k der exakten Stellen
3.555555555555555	3	1
3.198184286610794	27	2
3.149959034324206	187	3
3.141000179743544	2650	4
3.141500009935133	16954	5
3.141599999947270	213819	6
3.141592000000156	2403322	7
3.141592600000000	29254976	8
3.141592659999999	199932851	9
3.141592653000000	383761692	10
3.141592653500000	386800998	11
3.141592653580000	387350266	12
3.141592653589000	387413282	13
3.141592653589700	387418192	14
3.141592653589790	387418826	15

Im Jahr 1674 veröffentlichte Leibniz[6] die folgende Darstellungsformel für $\frac{\pi}{4}$:

$$\boxed{\frac{\pi}{4} = \sum_{k=1}^{\infty} (-1)^{k-1} \cdot \frac{1}{2k-1} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots + (-1)^{n-1} \cdot \frac{1}{2n-1} \dots \,. \quad (1.83)}$$

Berechnen wir nur die Summe der ersten n Summanden, ergibt sich ein Näherungswert für $\frac{\pi}{4}$ und somit für π. Diese Näherungswerte sind in der Tabelle 1.3 angegeben.

---

[5] John Wallis, geb. 3.12.1616 in Ashford, Kent, gest. 8.11.1703 in Oxford, englischer Mathematiker

[6] Gottfried Wilhelm Leibniz, geb. 1.7.1646 in Leipzig, gest. 14.11.1716 in Hannover, Philosoph, Mathematiker, Physiker, Historiker, 1672/73 Konstruktion einer Rechenmaschine für die vier Grundrechenarten, Entwickler der Differential- und Integralrechnung

Tab. 1.3 Unter Nutzung der Formel von Leibniz berechnete Näherungswerte für π

Näherungswert für π	Anzahl n der Summanden in (1.83)	Anzahl k der exakten Stellen
3.466666666666666	3	1
3.194187909231942	19	2
3.149995866593469	119	3
3.141000236580159	1688	4
3.141500009528465	10794	5
3.141599999994785	136121	6
3.141592000000232	1530012	7
3.141592600000000	18660304	8
3.141592659999999	155973041	9
3.141592653000000	1700659132	10

Wie aus der Tabelle ersichtlich ist, benötigt man ungefähr 1.7 Milliarden Summanden, damit man wenigstens zehn exakte Ziffern erhält. Mit dieser Formel ist es also extrem aufwändig, sehr genaue Näherungswerte für π zu bestimmen.

Euler[7] publizierte im Jahr 1748 eine Reihe von Darstellungsformeln für π, z. B. die Formel

$$\frac{\pi^2}{6} = \sum_{k=1}^{\infty} \frac{1}{k^2} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots + \frac{1}{n^2} + \cdots . \tag{1.84}$$

Berechnet man nur die Summe aus den ersten n Summanden, dann erhält man einen Näherungswert für $\frac{\pi^2}{6}$ und daraus einen Näherungswert für π. Derartige Näherungswerte sind in der Tabelle 1.4 angegeben.

Tab. 1.4 Unter Nutzung der Formel von Euler (1.84) berechnete Näherungswerte für π

Näherungswert für π	Anzahl n der Summanden in (1.84)	Anzahl k der exakten Stellen
3.011773947846214	7	1
3.100697301395177	23	2
3.140002027036458	600	3
3.141000025786488	1611	4
3.141500008065816	10307	5
3.141590000000166	359863	6
3.141592000000416	1461054	7
3.141592600000001	17820378	8

---

[7] Leonhard Euler, geb. 15.4.1707 in Basel, gest. 18.9.1783 in Sankt Petersburg, Mathematiker und Physiker

Auch bei dieser Formel benötigt man sehr viele Summanden, um einen recht genauen Näherungswert für π zu erhalten. Da die Summanden für große Werte von n sehr klein sind, haben sie ab einem gewissen n keinen Einfluss mehr bei der Addition in der Computerarithmetik (siehe Abschnitt 5.2, insbesondere auch Beispiel 5.11), so dass sich der Näherungswert für π auch nicht mehr verbessern lässt.

Eine weitere von Euler veröffentlichte Darstellungsformel für π ist die Formel

$$\boxed{\frac{\pi^2}{8} = \sum_{k=1}^{\infty} \frac{1}{(2k-1)^2} = 1 + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \cdots + \frac{1}{(2n-1)^2} + \cdots . \qquad (1.85)}$$

Näherungswerte für π, welche man bei Nutzung dieser Formel erhält, sind in der Tabelle 1.5 ersichtlich.

Tab. 1.5 Unter Nutzung der Formel von Euler (1.85) berechnete Näherungswerte für π

Näherungswert für π	Anzahl n der Summanden in (1.85)	Anzahl k der exakten Stellen
3.034615113797611	3	1
3.101600903283900	8	2
3.140000704127709	200	3
3.141000943973225	538	4
3.141500012560330	3436	5
3.141590000011233	119955	6
3.141592000000234	487018	7
3.141592600000000	5939849	8
3.141592659240904	4801919418	9

Wie wir aus dieser Tabelle erkennen können, ist es ebenfalls schwierig mittels der Formel (1.85) sehr genaue Näherungswerte für π zu bestimmen. Die folgende Formel von Euler ist wesentlich besser zur Bestimmung von Näherungswerten für π geeignet.

$$\boxed{\begin{aligned} \frac{\pi - 3}{4} &= \sum_{k=1}^{\infty} (-1)^{k-1} \cdot \frac{1}{2k \cdot (2k+1) \cdot (2k+2)} \\ &= \frac{1}{2 \cdot 3 \cdot 4} - \frac{1}{4 \cdot 5 \cdot 6} + \frac{1}{6 \cdot 7 \cdot 8} - \cdots (-1)^{n-1} \cdot \frac{1}{2n \cdot (2n+1) \cdot (2n+2)} \cdots \\ & \qquad (1.86) \end{aligned}}$$

Die Tabelle 1.6 enthält Näherungswerte für π, welche man unter Nutzung der obigen Formel erhält.

Tab. 1.6 Unter Nutzung der Formel von Euler (1.86) berechnete Näherungswerte für π

Näherungswert für π	Anzahl n der Summanden in (1.86)	Anzahl k der exakten Stellen
3.166666666666666	1	2
3.145238095238095	3	3
3.141254823607764	8	4
3.141518985595275	14	5
3.141599007405716	33	6
3.141592011095012	72	7
3.141592601798007	168	8
3.141592659950400	339	9
3.141592653004256	752	10
3.141592653500038	1406	11
3.141592653580005	2944	12
3.141592653589000	6806	13
3.141592653589700	13940	14
3.141592653589799	31003	15

Von Bailey[8], Borwein[9] und Plouffe[10] wurde 1995 die folgende Formel (BBP-Formel) zur Berechnung von π entwickelt (siehe auch Bailey et al. [1997]).

$$\boxed{\begin{aligned} \pi &= \sum_{k=0}^{\infty} \frac{1}{16^k}\left(\frac{4}{8k+1} - \frac{2}{8k+4} - \frac{1}{8k+5} - \frac{1}{8k+6}\right) \\ &= \left(4 - \frac{2}{4} - \frac{1}{5} - \frac{1}{6}\right) + \frac{1}{16}\left(\frac{4}{9} - \frac{2}{12} - \frac{1}{13} - \frac{1}{14}\right) \qquad (1.87) \\ &\quad + \cdots + \frac{1}{16^n}\left(\frac{4}{8n+1} - \frac{2}{8n+4} - \frac{1}{8n+5} - \frac{1}{8n+6}\right) + \cdots . \end{aligned}}$$

Wie die Tabelle 1.7 zeigt, erhält man schon mit sehr wenigen Summanden dieser Summe sehr genaue Näherungswerte für π.

Nachdem wir verschiedene Möglichkeiten zur Berechnung von Näherungswerten für π vorgestellt haben, diskutieren wir im Folgenden die näherungsweise Berechnung der Eulerschen Zahl e. Die Eulersche Zahl wird als Grenzwert der Zahlenfolge

$$(a_n)_{n=1}^{\infty} = \left(\left(1 + \frac{1}{n}\right)^n\right)_{n=1}^{\infty} \tag{1.88}$$

---

[8] David Harold Bailey, geb. 14.8.1948, US-amerikanischer Mathematiker und Informatiker

[9] Peter Benjamin Borwein, 5.10.1953 in St Andrews, kanadischer Mathematiker

[10] Simon Plouffe, geb. 11.6.1956 in Saint-Jovite, Québec, kanadischer Mathematiker

Tab. 1.7 Unter Nutzung der BBP-Formel berechnete Näherungswerte für π

Näherungswert für π	Anzahl $n+1$ der Summanden in (1.86)	Anzahl k der exakten Stellen
3.133333333333333	1	2
3.141422466422466	2	4
3.141587390346581	3	5
3.141592457567435	4	7
3.141592645460336	5	8
3.141592653228087	6	10
3.141592653572880	7	11
3.141592653588972	8	12
3.141592653589752	9	14
3.141592653589791	10	15

definiert (siehe auch Papula [2014], Jung [2021b], d. h.

$$\mathrm{e} = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n$$

Man kann zeigen, dass für wachsende Werte von n die sogenannten Folgeglieder $a_n = \left(1+\frac{1}{n}\right)^n$ immer größer werden, aber die Zahl e nicht überschreiten (siehe z. B. Meyberg und Vachenauer [2001a], Jung [2021b]). Deshalb kann man erwarten, dass man für hinreichend große Werte von n gute Näherungswerte für e erhält. Dieser Idee sind aber auf jedem Computer Grenzen gesetzt. Ab einem gewissen n^* führt eine weitere Vergrößerung von n zu keinem genaueren Näherungswert für die Eulersche Zahl e (siehe Beispiel 5.11 im Abschnitt 5.2). Wie groß man n wählen muss, damit man einen Näherungswert für e mit k exakten Stellen erhält, ist in der Tabelle 1.8 ersichtlich.

Eine andere Möglichkeit zur näherungsweisen Berechnung von e ist die Nutzung der Summe

$$\begin{aligned} \mathrm{e} &= \sum_{k=0}^{\infty}\frac{1}{k!} = \frac{1}{0!}+\frac{1}{1!}+\frac{1}{2!}+\frac{1}{3!}+\cdots+\frac{1}{n!}+\cdots \\ &= 1+1+\frac{1}{2}+\frac{1}{6}+\cdots+\frac{1}{1\cdot 2\cdot\cdots\cdot(n-1)\cdot n}+\cdots \end{aligned} \tag{1.89}$$

(siehe auch Papula [2014], Meyberg und Vachenauer [2001a], Jung [2021b] bezüglich der Taylorentwicklung der e-Funktion). In der Tabelle 1.9 geben wir an, wie groß n

Tab. 1.8 Unter Nutzung der Zahlenfolge (1.88) berechnete Näherungswerte für e

Näherungswert für e	n in (1.88)	Anzahl k der exakten Stellen
2.000000000000000	1	1
2.700139678846832	74	2
2.710040437932754	164	3
2.718000019542346	4822	4
2.718200001379015	16609	5
2.718280000070016	743198	6
2.718281000071943	1639286	7
2.718281800011852	31064651	8
2.718281820087948	48423998	9
2.718281828014652	51978772	10
2.718281828403200	52012934	11
2.718281828458244	52105193	12
2.718281828459916	67217174	13
2.718281828459070	67374142	14
2.718281828459041	82162639	15

sein muss, damit die Summe der ersten $n+1$ Summanden in der Formel (1.89) für e einen Näherungswert liefert, welcher k exakte Stellen hat.

Tab. 1.9 Unter Nutzung der Summe (1.89) berechnete Näherungswerte für e

Näherungswert für e	Anzahl $n+1$ der Summanden in (1.89)	Anzahl k der exakten Stellen
2.000000000000000	2	1
2.708333333333333	5	2
2.716666666666666	6	3
2.718055555555555	7	4
2.718253968253968	8	5
2.718281525573192	10	7
2.718281801146384	11	8
2.718281826198492	12	9
2.718281828286168	13	10
2.718281828446759	14	11
2.718281828458230	15	12
2.718281828459042	17	15
2.718281828459045	18	16

) "
! *+,-* * *. ! * ! @ + , 5
1#

2 *+,-* * *. ! * ! B 1 7 !
!11 ! ! 2 # C 7 0 9 1#
2 *+,-* * *. ! * ! B 1 7 !
!11 ! ! 2 # C 7 0 9 #D7

+,- * *. ! * ! 1
! ! 1#
+,- * *. ! * ! 1
! ! #D7

.2
+,- * *. ! * ! @ + , 5 1#
) ' 5

$$B = \{x : 2x^3 - x^2 - 3x = 0\},$$

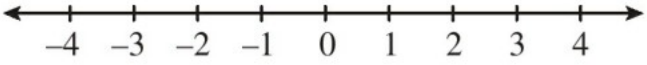

Chapter 1

Preliminaries

1.1 Numbers and Sets

The numbers we will use in this book are the *real numbers*. These are all the numbers that can be written in decimal notation. We often think of them as corresponding to points on a number line (see Figure 1.1). The simplest real numbers are the *integers*: the numbers 0, 1, −1, 2, −2, 3, −3, and so on. A real number is said to be *rational* if it can be written as an integer divided by an integer; for example, 2/3 and −13/5 are rational numbers. Notice that every integer is also a rational number, since, for example, we can write 3 as 3/1. Real numbers that are not rational are called *irrational*. For example, it can be shown that $\sqrt{2} = 1.41421\ldots$ and $\pi = 3.14159\ldots$ are irrational. Both rational and irrational numbers are spread throughout the number line; in fact, between any two real numbers there are infinitely many rational numbers and also infinitely many irrational numbers.

We will also often work with collections of numbers. In mathematics, a collection of objects is called a *set*, and the objects in the collection are called *elements* of the set. The simplest way to specify a set is to list the elements of the set between braces. For example, $\{-1, 0, 3/2\}$ is the set whose elements are the three numbers −1, 0, and 3/2. If we let the letter A stand for this set, then we write $3/2 \in A$ to say that 3/2 is an element of A, while $4 \notin A$ means that 4 is not an element of A.

Another way to specify a set is to give a rule for determining which objects belong to the set and which do not. For example, if we write

$$B = \{x : 2x^3 - x^2 - 3x = 0\}, \tag{1.1}$$

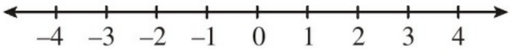

Figure 1.1: The number line.

then this means that B is the set whose elements are all values of x that satisfy the equation $2x^3 - x^2 - 3x = 0$. Equation (1.1) is read "B is equal to the set of all x such that $2x^3 - x^2 - 3x = 0$." The equation $2x^3 - x^2 - 3x = 0$ that appears in the definition of B is a statement that is true for some values of x and false for others. You should think of it as an elementhood test for the set B; those values of x that make the equation true pass the test and are elements of B, while those that make the equation false are not. To determine which numbers belong to B we simply have to solve the equation, which we can do by factoring the left-hand side. We have

$$2x^3 - x^2 - 3x = x(x + 1)(2x - 3),$$

so the equation can be rewritten $x(x + 1)(2x - 3) = 0$, and the solutions are 0, -1, and 3/2. These are the elements of the set B. Notice that these are exactly the same as the elements of the set A defined earlier. Thus $B = A$; they are both the same collection of numbers, described in two different ways.

Although we will usually use the letter x when writing an elementhood test to define a set, as we did in the definition of B, in fact any letter can be used. For example, the set $C = \{y : 2y^3 - y^2 - 3y = 0\}$ is the set of all values of y that satisfy the equation $2y^3 - y^2 - 3y = 0$. Of course, this is the same equation that we used in the definition of B, but with x replaced by y. The values of y that satisfy the equation are therefore once again the numbers 0, -1, and 3/2. Therefore $C = B = A$; we have the same set of numbers described in yet another way.

Here is another example of a set defined by an elementhood test: $I = \{x : 2 < x < 5\}$. This time the elementhood test is $2 < x < 5$, which is a shorthand way of saying that $2 < x$ and $x < 5$. In this case $3 \in I$, since the statement $2 < 3 < 5$ is true, but $5 \notin I$, since the statement $2 < 5 < 5$ is false. The elements of I are all the numbers strictly between 2 and 5. There are infinitely many numbers in this range, so we cannot list all the elements of I, as we did for A. But we can mark them on a number line, as in Figure 1.2.

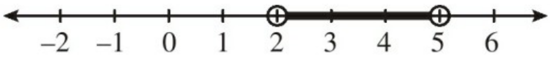

Figure 1.2: $I = \{x : 2 < x < 5\}$. The open circles indicate that 2 and 5 are not elements of I, while the thick line shows that all numbers between 2 and 5 are elements.

The set I is an example of a kind of set called an *open interval*. For any numbers a and b with $a < b$, the set of all numbers strictly between a and b is an open interval, and it is denoted (a, b). In other words,

$$(a, b) = \{x : a < x < b\}.$$

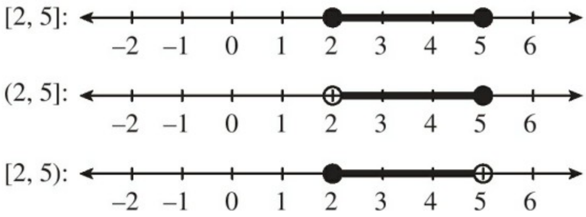

Figure 1.3: The closed interval [2, 5], and the half-open intervals (2, 5] and [2, 5). Black dots indicate points that are included in a set.

The numbers a and b are called the *endpoints* of the interval. Thus $I = (2, 5)$; it is the open interval with endpoints 2 and 5.

The endpoints of an open interval are not elements of the interval. But sometimes we will want to include the endpoints, so we define

$$[a, b] = \{x : a \leq x \leq b\}.$$

This set is called a *closed interval*. For example, $[2, 5] = \{x : 2 \leq x \leq 5\}$. This set is exactly the same as the open interval (2, 5) considered earlier, except that it includes the endpoints 2 and 5 as elements. If we include only one endpoint, we get a *half-open interval*. As you might guess, we write half-open intervals like this:

$$(a, b] = \{x : a < x \leq b\},$$
$$[a, b) = \{x : a \leq x < b\}.$$

In general, we use a square bracket to indicate that an endpoint is included in an interval, and a parenthesis to indicate that it is not. Figure 1.3 shows examples of closed and half-open intervals.

The *interior* of an interval is the set containing all numbers in the interval except the endpoints. Thus, the interior of the closed interval [2, 5] is the open interval (2, 5). The interiors of (2, 5] and [2, 5) are also (2, 5), and we will even say that the interior of (2, 5) is (2, 5).

Finally, we sometimes want to consider intervals that extend infinitely far in some direction, so we introduce the notation:

$$(a,\infty) = \{x : x > a\},$$
$$[a,\infty) = \{x : x \geq a\},$$
$$(-\infty, b) = \{x : x < b\},$$
$$(-\infty, b] = \{x : x \leq b\}.$$

Some examples of infinite intervals are shown in Figure 1.4. The interior of $[a,\infty)$ is (a,∞), and the interior of $(-\infty, b]$ is $(-\infty, b)$. The set of all real numbers is often denoted R, but we could also think of it as the interval $(-\infty,\infty)$. We consider any interval that does not include its endpoints to be an open interval. Thus, the intervals (a,∞), $(-\infty, b)$, and $(-\infty,\infty)$ are open intervals, and the interior of any interval is an open interval.

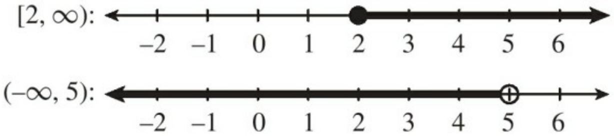

Figure 1.4: Infinite intervals.

Since this is our first use of the infinity symbol∞, it might be worthwhile to pause at this point to explain what this symbol means. The most important thing to understand about the infinity symbol is that *there is no such number as infinity*. You might wonder, then, how it can be correct to use this symbol in mathematical notation like (a,∞). The answer is that, according to the definition we have given, this notation is simply a shorthand for $\{x : x > a\}$, and this last expression makes no mention of infinity. Every time we make a statement using the symbol ∞, it will be a similar shorthand for a statement that makes no mention of infinity. We will never use the symbol ∞ as if it stood for a number. Thus, for example, wewould never set x equal to ∞in some formula, and we would never talk about the “closed interval” $[2,\infty]$.

There are two ways of combining sets that we will sometimes make use of. If A and B are sets, then the *intersection* of A and B, denoted $A \cap B$, is the set whose elements are those objects that belong to both A and B. Thus

$$A \cap B = \{x : x \in A \text{ and } x \in B\}.$$

For example,

$$\begin{aligned}[2,\infty) \cap (-\infty, 5) &= \{x : x \in [2,\infty) \text{ and } x \in (-\infty, 5)\}\\ &= \{x : 2 \leq x \text{ and } x < 5\} = [2, 5).\end{aligned}$$

Looking at Figure 1.4, you can see that the elements of $[2,\infty) \cap (-\infty, 5)$ are those numbers that are in the overlap of the sets $[2,\infty)$ and $(-\infty, 5)$. In general, you can think of $A \cap B$ as the overlap of A and B.

The *union* of A and B, denoted $A \cup B$, is the set whose elements are all those objects that are elements of either A or B (or both). That is,

$$A \cup B = \{x : x \in A \text{ or } x \in B\}.$$

You could think of $A \cup B$ as the set you get if you take all the elements of A, and all the elements of B, and throw them together into one set. For example, if we take all the numbers in the intervals $(2, 4]$ and $[4, 5]$ and put them together into one set, we get the interval $(2, 5]$. That is,

$$\begin{aligned}(2, 4] \cup [4, 5] &= \{x : x \in (2, 4] \text{ or } x \in [4, 5]\}\\ &= \{x : 2 < x \leq 4 \text{ or } 4 \leq x \leq 5\} = \{x : 2 < x \leq 5\} = (2, 5].\end{aligned}$$

If A and B are sets, then A is called a *subset* of B if every element of A is also an element of B. We write $A \subseteq B$ to indicate that A is a subset of B. For example, $(2, 4) \subseteq [2, 4]$, and $[2, 4] \subseteq (1, 5)$.

One reason that intervals are important in calculus is that they often come up as solution sets of inequalities. In particular, we will often be concerned with inequalities involving absolute values. Recall that the *absolute value* of a number x is defined as follows:

$$|x| = \begin{cases} x, & \text{if } x \geq 0,\\ -x, & \text{if } x < 0.\end{cases}$$

This notation means that if $x \geq 0$ then $|x| = x$, and if $x < 0$ then $|x| = -x$. For example, $|3| = 3$ and $|-4| = -(-4) = 4$.

The fact that $|x|$ is defined by cases, with one formula when $x \geq 0$ and a different formula when $x < 0$, suggests a method that can be used when solving any problem involving absolute values: reasoning by cases. As a simple example of this kind of reasoning, notice that if $x \geq 0$ then we have $|x| = x \geq 0$, and if $x < 0$ then $|x| = -x > 0$. In both cases the statement $|x| \geq 0$ is true, so we conclude that for every number x, $|x| \geq 0$. You can also use reasoning by cases to show that for every number x, $\sqrt{x^2} = |x|$ (see Exercise 15).

Here's an example of how reasoning by cases can be used to solve an inequality involving absolute values:

Example 1.1.1. Solve:

$$|x| < 3.$$

Solution. Motivated by the definition of $|x|$, we will consider $x \geq 0$ and $x < 0$ separately.

Case 1. $x \geq 0$. In this case, according to the definition of absolute value we have $|x| = x$, so the inequality $|x| < 3$ means $x < 3$.

Case 2. $x < 0$. Now the definition of absolute value says that $|x| = -x$, and substituting this into our inequality $|x| < 3$ gives us $-x < 3$. Multiplying by -1 (and remembering that when multiplying an inequality by a negative number, we must reverse the direction of the inequality) we get $x > -3$.

So what's the solution to our inequality? Is it $x < 3$, as we found in case 1, or $x > -3$, as in case 2? To answer this question, we must think about what it means to solve an inequality. To solve the inequality $|x| < 3$ means to determine which values of x make the inequality true. Our reasoning in case 1 shows that, for $x \geq 0$, the inequality means $x < 3$. Thus the inequality is true if $0 \leq x < 3$ and false if $x \geq 3$. We can't tell from this reasoning whether the inequality is true or false when $x < 0$; that's the purpose of case 2. Case 2 tells us that, for $x < 0$, the inequality will be true precisely when $x > -3$. Thus, the inequality is true if $-3 < x < 0$ and false if $x \leq -3$. Putting all this information together, we conclude that the inequality is true if $-3 < x < 3$ and false if either $x \geq 3$ or $x \leq -3$, as shown in Figure 1.5. This means that the *solution set* of the inequality is an open interval:

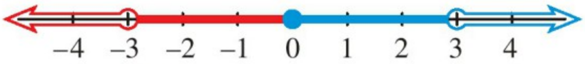

Figure 1.5: The solution to Example 1.1.1.The parts of the number line marked with solid lines are in the solution set, and outlined areas are not. The blue lines were determined in case 1 of the solution, and the red lines in case 2.

$$\{x : |x| < 3\} = \{x : -3 < x < 3\} = (-3, 3).$$

Notice that in case 1 we determined that all numbers in the interval $[0, 3)$ are in the solution set, and in case 2 we determined that the numbers in $(-3, 0)$ are also in the solution set. The solution set is therefore the union of these two intervals: $[0, 3) \cup (-3, 0) = (-3, 3)$.

Here's another way of describing the answer to Example 1.1.1. Our reasoning shows that the statements $|x| < 3$ and $-3 < x < 3$ are true for exactly the same values of x; the two statements are equivalent. In other words, for any number x, if $|x| < 3$, then $-3 < x < 3$, and if $-3 < x < 3$, then $|x| < 3$. Mathematicians usually describe this situation by saying that $|x| < 3$ is true *if and only if* $-3 < x < 3$. The phrase "if and only if" comes up often in mathematics, and we will see it many times later in this book.

Of course, there is nothing special about the number 3 in this example. Similar reasoning, using the variable y in place of the number 3, can be used to establish the following theorem. Parts 3 and 4 of the theorem follow directly from parts 1 and 2, by negating the statements involved.

Theorem 1.1.2. *For any numbers x and y, the following statements are true:*

1. $|x| < y$ *if and only if* $-y < x < y$.
2. $|x| \leq y$ *if and only if* $-y \leq x \leq y$.
3. $|x| \geq y$ *if and only if either* $x \leq -y$ *or* $x \geq y$.
4. $|x| > y$ *if and only if either* $x < -y$ *or* $x > y$.

The most important use of absolute values in calculus is to compute distances between numbers on the number line. To find the distance between two numbers, we subtract the smaller number from the larger. For example, the distance from -4 to 3 on the number line is $3-(-4) = 7$. In general, if $a \leq b$ then the distance from a to b is $b-a$. But if $a > b$ then the distance is $a - b$. Is there a single formula that gives the distance from a to b no matter which of the numbers is larger?

It turns out that the formula $|b-a|$ does the trick. We can see this by once again using reasoning by cases.

Case 1. $a \leq b$. Then $b-a \geq 0$, so $|b-a| = b-a$, which, in this case, is the distance from a to b on the number line.

Case 2. $a > b$. Now $b-a < 0$, so $|b-a| = -(b-a) = a - b$, which is, once again, the distance from a to b in this case.

Thus, no matter which of the numbers a and b is larger, we have

$$|b-a| = \text{the distance from } a \text{ to } b \text{ on the number line.}$$

This fact provides a nice way to see why our solution to Example 1.1.1 makes sense. We have

$$|x| = |x - 0| = \text{distance from 0 to } x \text{ on the number line.}$$

With this interpretation for $|x|$, the inequality $|x| < 3$ can be thought of as saying

$$(\text{distance from 0 to } x \text{ on the number line}) < 3.$$

It is clear geometrically that the values of x for which this statement is true are the numbers in the open interval $(-3, 3)$, exactly as we found in our solution to Example 1.1.1.

Example 1.1.3. Solve:

$$|3t + 2| \leq 4.$$

Solution. By part 2 of Theorem 1.1.2, the inequality to be solved is equivalent to

$$-4 \leq 3t + 2 \leq 4.$$

Subtracting 2 all the way through the inequality gives us

$$-6 \leq 3t \leq 2,$$

and then dividing by 3 we get

$$-2 \leq t \leq 2/3.$$

Thus the solution set for this inequality is a closed interval:

$$\{t : |3t + 2| \leq 4\} = \{t : -2 \leq t \leq 2/3\} = [-2, 2/3].$$

Sometimes it is useful to be able to simplify absolute values of complicated expressions. In such situations, the following theorem can be helpful.

Theorem 1.1.4. *For any numbers x and y, the following statements are true:*

1. $|xy| = |x||y|$.
2. *If* $y \neq 0$ *then* $\left|\frac{x}{y}\right| = \frac{|x|}{|y|}$.

Proof. We will just prove part 1; the proof of part 2 is similar. First note that if either x or y is 0 then both sides of the equation $|xy| = |x||y|$ are 0, so the equation is true.
If neither of them is 0, then each is either positive or negative. This leaves us with four cases to consider:

Case 1. $x > 0, y > 0$. Then by the definition of absolute value, $|x| = x$ and $|y| = y$. Also, since the product of two positive numbers is positive, $xy > 0$. Therefore $|xy| = xy = |x||y|$.

Case 2. $x > 0, y < 0$. Then $|x| = x$ and $|y| = -y$. Since a positive number times a negative number is negative, $xy < 0$. Thus $|xy| = -xy = x(-y) = |x||y|$.

Case 3. $x < 0, y > 0$. Then $|x| = -x$, $|y| = y$, and $xy < 0$, so $|xy| = -xy = |x||y|$.

Case 4. $x < 0, y < 0$. Then $|x| = -x$ and $|y| = -y$. Since a negative times a negative is positive, we have $xy > 0$, so $|xy| = xy = (-x)(-y) = |x||y|$.

To illustrate the use of Theorem 1.1.4, we briefly revisit Example 1.1.3. Notice that $|3t + 2| = |3(t + 2/3)| = |3||t + 2/3| = 3|t - (-2/3)|$, where we have used part 1 of Theorem 1.1.4 in the second step. Thus, the inequality in Example 1.1.3 can be rewritten

$$3|t - (-2/3)| \leq 4.$$

Dividing through by 3, this is equivalent to

$$|t - (-2/3)| \leq 4/3,$$

which can be interpreted as meaning

$$\text{(distance from } t \text{ to } -2/3 \text{ on the number line)} \leq 4/3.$$

Thus the solution set of the inequality is

$$[-2/3-4/3, -2/3+4/3] = [-2, 2/3],$$

exactly as we found in Example 1.1.3.

Theorem 1.1.4 tells us how to simplify an absolute value of a product or a quotient. What about the absolute value of a sum? Is it always true that $|x+y| = |x| + |y|$? A little experimentation shows that, unfortunately, this equation is not always true. For example, $|5+(-3)| = 2$, but $|5| + |-3| = 8$. However, we do have the following important fact, which is known as the *triangle inequality*.

Theorem 1.1.5 (Triangle Inequality). *For all numbers x and y,* $|x+y| \leq |x| + |y|$.

Proof. According to part 2 of Theorem 1.1.2, the triangle inequality is equivalent to the statement

$$-(|x|+|y|) \leq x+y \leq |x|+|y|, \tag{1.2}$$

so it will suffice to prove this inequality. We will leave it to you to verify (using reasoning by cases) that

$$-|x| \leq x \leq |x|$$

and

$$-|y| \leq y \leq |y|$$

(see Exercise 18). Adding these two inequalities gives inequality (1.2).

The theorems we have proven about absolute values could be regarded as shortcuts that allow us to solve problems more easily. For example, using Theorem 1.1.2 we were able to solve Example 1.1.3 without having to resort to reasoning by cases. This is a pattern that we will see repeated many times in this book. When a new concept is introduced, we will initially rely on the definition to tell us how to solve problems involving that concept. But often solutions based on the definition will be long and complicated, so we will develop theorems that provide shortcuts that allow us to solve problems more easily.

You may wonder, why bother learning the definitions? Why not just learn the shortcuts, if they provide easier ways of solving problems? One answer is that shortcuts, while helpful, are also often somewhat limited. They usually allow us to solve only a restricted range of problems. Sometimes you come across a problem for which the shortcuts are not helpful, and then you have no choice but to return to the definitions of the relevant concepts to solve the problem.

As an illustration of this, we close this section by solving an inequality involving absolute values for which our various shortcuts don't seem to be very helpful.

Example 1.1.6. Solve:

$$2|x|-3 \geq |x-1|.$$

Solution. Although it is possible to use Theorem 1.1.2 to solve this inequality (see Exercise 16), the solution is not easy. Here we use the more straightforward approach of working from the definition of absolute value, which suggests reasoning by cases. Since we need to work with both $|x|$ and $|x-1|$, we are led to consider three cases: $x < 0$, $0 \leq x < 1$, and $x \geq 1$.

Case 1. $x < 0$. In this case we also have $x-1 < 0$, so according to the definition of absolute value, $|x| = -x$ and $|x-1| = -(x-1) = 1-x$. Substituting into our inequality, we find that we must solve

$$-2x-3 \geq 1-x.$$

Adding $2x$ to both sides and subtracting 1, we get $-4 \geq x$. Thus, for negative values of x, the inequality is true if and only if $x \leq -4$. In other words, it is true if $x \leq -4$ and false if $-4 < x < 0$.

Case 2. $0 \leq x < 1$. We have $x \geq 0$, so $|x| = x$, and $x < 1$, so $x-1 < 0$ and therefore $|x-1| = 1-x$. Thus in this case the inequality means

$$2x-3 \geq 1-x,$$

which can be simplified to $x \geq 4/3$. But this inequality is false for all values of x in the range $0 \leq x < 1$. Thus, none of the numbers in the interval $[0, 1)$ will be included in the solution set of our inequality.

Case 3. $x \geq 1$. Then $x \geq 0$ and $x-1 \geq 0$, and therefore $|x| = x$ and $|x-1| = x-1$. Filling in these formulas in our inequality we get

$$2x - 3 \geq x - 1,$$

and simplifying leads to $x \geq 2$. Thus, the inequality is true for $x \geq 2$ and false for $1 \leq x < 2$.

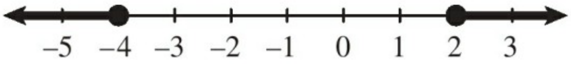

Figure 1.6: The solution set for Example 1.1.6 is $(-\infty,-4] \cup [2,\infty)$.

Combining the information from all three cases, we find that the solution set is the set containing all numbers in the interval $(-\infty,-4]$ and also all numbers in the interval $[2,\infty)$. In other words,

$$\{x : 2|x| - 3 \geq |x - 1|\} = \{x : \text{either } x \leq -4 \text{ or } x \geq 2\} = (-\infty,-4] \cup [2,\infty).$$

Exercises 1.1

1–13: Solve the inequality. If possible, write the solution set as an interval or a union of intervals.

1. $|x - 5| < 7$.
2. $|4x + 2| \leq 6$.
3. $|5 - 2t| \geq 4$.
4. $|x - 4| < x$.
5. $2|x - 4| < x$.
6. $|x + 4| < x$.
7. $|x + 4| < 2x$.
8. $|6 - 2u| - 3 < u$.
9. $|3y| \leq |y| + 10$.
10. $|x - 3| \geq |x| - 1$.
11. $|2x - 3| \geq |x| - 1$.
12. $|2x - 2| > |x| - 1$.
13. $|3x - 6| \leq |3 - 6x|$.
14. Write the following sets as intervals or unions of intervals. Use

interval notation.

(a) $\{x : x^2 < 64\}$.

(b) $\{x : x^3 < 64\}$.

(c) $\{x : 4 \le x^2 < 9\}$.

(d) $(-\infty, 5] \cap (3, \infty)$.

(e) $(-3, 5] \cup (3, 7]$.

15. Use reasoning by cases to prove that for every number x, $\sqrt{x^2} = |x|$.
16. Solve Example 1.1.6 by using Theorem 1.1.2.
17. Prove part 1 of Theorem 1.1.2. (Hint: Treat the cases $y > 0$ and $y \le 0$ separately.)
18. Prove that for every number x, $-|x| \le x \le |x|$. (This fact is used in the proof of Theorem 1.1.5.)
19. Prove that for all numbers x and y, $|x - y| \ge |x| - |y|$. (Hint: Start by applying the triangle inequality to $|(x - y) + y|$.)
20. Use the triangle inequality to show that for every number x, if $|x| \le 2$ then $|x^3 - 7x + 3| \le 25$.
21. Show that for every number x, if $|x| \le 2$ then $\left|\frac{x^3 - 7x + 3}{9 - x^2}\right| \le 5$. (Hint: Use

    Exercise 19 to show that $|9 - x^2| \ge 5$.)

1.2 Graphs in the Plane

In Section 1.1 we considered statements involving a single variable, usually x, and we illustrated these statements by marking on a number line the values of the variable for which the statement is true. Often in calculus we will work with statements involving two variables, usually x and y. To illustrate such statements we use the coordinate plane. Each point in the plane represents an ordered pair of numbers (x, y), as illustrated in Figure 1.7.[1]

For example, consider the statement

$$1 < x \le 3 \text{ and } -2 \le y < 1. \tag{1.3}$$

+,- * *. ! * + (# 1 4 + ! 0
! # / 0 + ! ! ! ! 0 1#
) '

Multiplying both sides by -1 and using 2) yields

$$0 = (-1)0 = (-1)(1 + (-1)) = (-1)1 + (-1)^2 = -1 + (-1)^2.$$

Thus $(-1)^2$ is an additive inverse to -1; of course 1 also is. By the uniqueness of additive inverses, we see that $(-1)^2 = 1$. The proof of 4) is similar. Start with $0 = 1 + (-1)$ and multiply by x to get $0 = x + (-1)x$. Thus $(-1)x$ is an additive inverse of x and the result follows by uniqueness of additive inverses. Finally, to prove 5), we assume that $xy = 0$. If $x = 0$, the conclusion holds. If $x \neq 0$, we can multiply by $\frac{1}{x}$ to obtain

$$y = (\frac{1}{x}x)y = \frac{1}{x}(xy) = \frac{1}{x}0 = 0.$$

Thus, if $x \neq 0$, then $y = 0$, and the conclusion also holds. □

We note a point of language, where mathematics usage may differ with common usage. For us, the phrase "either $x = 0$ or $y = 0$" allows the possibility that **both** $x = 0$ and $y = 0$.

Example 2.1. A field with two elements. Let $\mathbf{F}_2$ consist of the two elements 0 and 1. We put $1 + 1 = 0$, but otherwise we add and multiply as usual. Then $\mathbf{F}_2$ is a field.

This example illustrates several interesting things. For example, the object 2 (namely $1 + 1$), can be 0 in a field. This possibility will prevent the quadratic formula from holding in a field for which $2 = 0$. In Theorem 2.1 we will derive the quadratic formula when it is possible to do so.

First we make a simple observation. We have shown that $(-1)^2 = 1$. Hence, when $-1 \neq 1$, it follows that 1 has two square roots, namely ± 1. Can an element of a field have more than two square roots? The answer is no.

Lemma 2.1. *In a field, an element t can have at most two square roots. If x is a square root of t, then so is $-x$, and there are no other possibilities.*

Proof. If $x^2 = t$, then $(-x)^2 = t$ by 3) and 4) of Proposition 2.1. To prove that there are no other possibilities, we assume that both x and y are square roots of t. We then have

$$0 = t - t = x^2 - y^2 = (x - y)(x + y). \tag{2}$$

By 5) of Proposition 2.1, we obtain either $x - y = 0$ or $x + y = 0$. Thus $y = \pm x$ and the result follows. □

The difference of two squares law stating that $x^2 - y^2 = (x - y)(x + y)$ is a gem of elementary mathematics. For example, suppose you are asked to multiply 88 times 92 in your head. You imagine $88 * 92 = (90 - 2)(90 + 2) = 8100 - 4 = 8096$ and impress some audiences. One can also view this algebraic identity for positive integers simply by removing a small square of dots from a large square of dots and rearranging the dots to form a rectangle. The author once used this kind of method when doing volunteer teaching of multiplication to third graders. See Figure 1.1 for a geometric interpretation of the identity in terms of area.

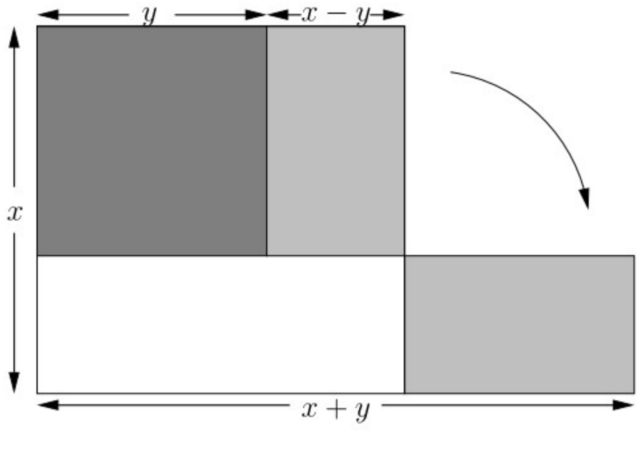

Figure 1.1. Difference of two squares.

We pause to make several remarks about square roots. The first remark concerns a notational convention; the discussion will help motivate the notion of ordered field defined below. The real number system will be defined formally below, and we will prove that positive real numbers have square roots. Suppose $t > 0$. We write $\sqrt{t}$ to denote the *positive* x for which $x^2 = t$. Thus both x and $-x$ are square roots of t, but the notation $\sqrt{t}$ means the positive square root. For the complex numbers, things will be more subtle. We will prove that each nonzero complex number z has two square roots, say $\pm w$, but there is no sensible way to prefer one to the other. We emphasize that the existence of square roots depends on more than the field axioms. Not all positive rational numbers have rational square roots, and hence it must be proved that each positive real number has a square root. The proof requires a limiting process. The quadratic formula, proved next, requires that the expression $b^2 - 4ac$ be a square. In an arbitrary field, the expression $\sqrt{t}$ usually means any x for which $x^2 = t$, but the ambiguity of signs can cause confusion. See Exercise 1.4.

Theorem 2.1. *Let* $\mathbf{F}$ *be a field. Assume that* $2 \neq 0$ *in* $\mathbf{F}$. *For* $a \neq 0$ *and arbitrary* b, c *we consider the quadratic equation*

$$ax^2 + bx + c = 0. \tag{3}$$

Then x *solves* (3) *if and only if*

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \tag{4}$$

If $b^2 - 4ac$ *is not a square in* $\mathbf{F}$, *then* (3) *has no solution.*

Proof. The idea of the proof is to *complete the square*. Since both a and 2 are nonzero elements of $\mathbf{F}$, they have multiplicative inverses. We therefore have

$$ax^2 + bx + c = a(x^2 + \frac{b}{a}x) + c = a(x^2 + \frac{b}{a}x + \frac{b^2}{4a^2}) + c - \frac{b^2}{4a}$$

$$= a(x + \frac{b}{2a})^2 + \frac{4ac - b^2}{4a}. \tag{5}$$

We set (5) equal to 0 and we can easily solve for x. After dividing by a, we obtain

$$(x + \frac{b}{2a})^2 = \frac{b^2 - 4ac}{4a^2}. \tag{6}$$

The square roots of $4a^2$ are of course $\pm 2a$. Assuming that $b^2 - 4ac$ has a square root in $\mathbf{F}$, we solve (6) for x by first taking the square root of both sides. We obtain

$$x + \frac{b}{2a} = \pm\frac{\sqrt{b^2 - 4ac}}{2a}. \tag{7}$$

After a subtraction and simplification we obtain (4) from (7). □

The reader surely has seen the quadratic formula before. Given a quadratic polynomial with real coefficients, the formula tells us that the polynomial will have no real roots when $b^2 - 4ac < 0$. For many readers the first exposure to complex numbers arises when we introduce square roots of negative numbers in order to use the quadratic formula.

► **Exercise 1.3.** Show that additive and multiplicative inverses in a field are unique.

► **Exercise 1.4.** A subtlety. Given a field, is the formula

$$\sqrt{u}\sqrt{v} = \sqrt{uv}$$

always valid? In the proof of the quadratic formula, did we use this formula implicitly? If not, what did we use?

Example 2.2. One can completely analyze quadratic equations with coefficients in $\mathbf{F}_2$. The only such equations are $x^2 = 0$, $x^2 + x = 0$, $x^2 + 1 = 0$, and $x^2 + x + 1 = 0$. The first equation has only the solution 0. The second has the two solutions 0 and 1. The third has only the solution 1. The fourth has no solutions. We have given a complete analysis, even though Theorem 2.1 cannot be used in this setting.

Before introducing the notion of ordered field, we give a few other examples of fields. Several of these examples use modular (clock) arithmetic. The phrases *add modulo p* and *multiply modulo p* have the following meaning. Fix a positive integer p, called the *modulus*. Given integers m and n, we add (or multiply) them as usual and then take the remainder upon division by p. The remainder is called the sum (or product) modulo p. This natural notion is familiar to everyone; five hours after nine o'clock is two o'clock; we added modulo twelve. The subsequent examples can be skipped without loss of continuity.

Example 2.3. Fields with finitely many elements. Let p be a prime number, and let $\mathbf{F}_p$ consist of the numbers $0, 1, ..., p-1$. We define addition and multiplication modulo p. Then $\mathbf{F}_p$ is a field.

In Example 2.3, p needs to be a prime number. Property 5) of Proposition 2.1 fails when p is not a prime. We mention without proof that the number of elements in a finite field must be a power of a prime number. Furthermore, for each prime p and positive integer n, there exists a finite field with p^n elements.

► **Exercise 1.5.** True or false? Every quadratic equation in $\mathbf{F}_3$ has a solution.

Fields such as $\mathbf{F}_p$ are important in various parts of mathematics and computer science. For us, they will serve only as examples of fields. The most important examples of fields for us will be the real numbers and the complex numbers. To define these fields rigorously will take a bit more effort. We end this section by giving an example of a field built from the real numbers. We will not use this example in the logical development.

Example 2.4. Let K denote the collection of rational functions in one variable x with real coefficients. An element of K can be written $\frac{p(x)}{q(x)}$, where p and q are polynomials, and we assume that q is not the zero polynomial. (We allow $q(x)$ to equal 0 for some x, but not for all x.) We add and multiply such rational functions in the usual way. It is tedious but not difficult to verify the field axioms. Hence K is a field. Furthermore, K contains $\mathbf{R}$ in a natural way; we identify the real number c with the constant rational function $\frac{c}{1}$. As with the rational numbers, many different fractions represent the same element of K. To deal rigorously with such situations, one needs the notion of equivalence relation, discussed in Section 5.

3. Inequalities and ordered fields

Comparing the sizes of a pair of integers or of a pair of rational numbers is both natural and useful. It does not make sense however to compare the sizes of elements in an arbitrary ring or field. We therefore introduce a crucial property shared by the integers $\mathbf{Z}$ and the rational numbers $\mathbf{Q}$. For x, y in either of these sets, it makes sense to say that $x > y$. Furthermore, given the pair x, y, one and only one of three things must be true: $x > y$, $x < y$, or $x = y$. We need to formalize this idea in order to define the real numbers.

Definition 3.1. A field $\mathbf{F}$ is called *ordered* if there is a subset $P \subset \mathbf{F}$, called the *set of positive elements* of $\mathbf{F}$, satisfying the following properties:

1) For all x, y in P, we have $x + y \in P$ and $xy \in P$ (closure).

2) For each x in $\mathbf{F}$, one and only one of the following three statements is true: $x = 0$, $x \in P$, $-x \in P$ (trichotomy).

► **Exercise 1.6.** Let $\mathbf{F}$ be an ordered field. Show that $1 \in P$.

► **Exercise 1.7.** Show that the trichotomy property can be rewritten as follows. For each x, y in $\mathbf{F}$, one and only one of the following three statements is true: $x = y$, $x - y \in P$, $y - x \in P$.

The rational number system is an ordered field; a fraction $\frac{p}{q}$ is positive if and only if p and q have the same sign. Note that q is never 0 and that a rational number is 0 whenever its numerator is 0. It is of course elementary to check in

This equivalence relation partitions the set $\mathbf{R}[t]$ into equivalence classes; the situation is strikingly similar to modular arithmetic. Given a polynomial $p(t)$, we use the division algorithm to write $p(t) = q(t)(1+t^2) + r(t)$, where the remainder r has degree at most one. Thus $r(t) = a + bt$ for some a, b, and this r is the unique first-degree polynomial equivalent to p. In the case of modular arithmetic we used the remainder upon division by the modulus; here we use the remainder upon division by $t^2 + 1$.

► **Exercise 1.31.** Verify the transitivity property of equivalence modulo I.

Standard notation in algebra writes $\mathbf{R}[t]/(1+t^2)$ for the set of equivalence classes. We can add and multiply in $\mathbf{R}[t]/(1+t^2)$. As usual, the sum (or product) of equivalence classes P and Q is defined to be the equivalence class of the sum $p+q$ (or the product pq) of members; the result is independent of the choice. An equivalence class then can be identified with a polynomial $a+bt$, and the sum and product of equivalence classes satisfies (15) and (16). In this setting we define $\mathbf{C}$ as the collection of equivalence classes with this natural sum and product:

$$\mathbf{C} = \mathbf{R}[t]/(1+t^2). \tag{25}$$

Definition (25) allows us to set $t^2 = -1$ whenever we encounter a term of degree at least two. The irreducibility of t^2+1 matters. If we form $\mathbf{R}[t]/(p(t))$ for a reducible polynomial p, then the resulting object will not be a field. The reason is precisely parallel to the situation with modular arithmetic. If we consider $\mathbf{Z}/(n)$, then we get a field (written $\mathbf{F}_n$) if and only if n is prime.

► **Exercise 1.32.** Show that $\mathbf{R}[t]/(t^3+1)$ is not a field.

► **Exercise 1.33.** A polynomial $\sum_{k=0}^{d} c_k t^k$ in $\mathbf{R}[t]$ is equivalent to precisely one polynomial of the form $A + Bt$ in the quotient space. What is $A + Bt$ in terms of the coefficients c_k?

► **Exercise 1.34.** Prove the division algorithm in $\mathbf{R}[t]$. In other words, given polynomials p and g, with g not the zero polynomial, show that one can write $p = qg + r$ where either $r = 0$ or the degree of r is less than the degree of g. Show that q and r are uniquely determined by p and g.

► **Exercise 1.35.** For any polynomial p and any x_0, show that there is a polynomial q such that $p(x) = (x - x_0)q(x) + p(x_0)$.

6. A glimpse at metric spaces

Both the real number system and the complex number system provide intuition for the general notion of a metric space. This section can be omitted without impacting the logical development, but it should appeal to some readers.

Definition 6.1. Let X be a set. A *distance function* on X is a function $\delta : X \times X \to \mathbf{R}$ such that the following hold:

1) $\delta(x, y) \geq 0$ for all $x, y \in X$ (distances are nonnegative).

2) $\delta(x, y) = 0$ if and only if $x = y$ (distinct points have positive distance between them; a point has 0 distance to itself).

3) $\delta(x, y) = \delta(y, x)$ for all $x, y \in X$ (the distance from x to y is the same as the distance from y to x).

4) $\delta(x, w) \leq \delta(x, y) + \delta(y, w)$ for all $x, y, w \in X$ (the triangle inequality).

Definition 6.2. Let X be a set, and let δ be a distance function on X. The pair (X, δ) is called a *metric space*.

Mathematicians sometimes write that "the distance function δ makes the set X into a metric space". That statement is not completely precise. It ignores a somewhat pedantic point: the set X is not the metric space; a metric space consists of the set and the distance function. Many different distance functions are possible for most sets X.

Sometimes we use the word *metric* on its own to mean the distance function on a metric space. The most intuitive example is the real number system; putting $\delta(x, y) = |x - y|$ makes $\mathbf{R}$ into a metric space. There are many other possibilities for metrics on $\mathbf{R}$; for example, putting $\delta(x, y)$ equal to 1 whenever $x \neq y$ gives another possible distance function. In Chapter 2 we formally define the absolute value function for complex numbers. Then $\mathbf{C}$ becomes a metric space when we put $\delta(z, w) = |z - w|$. Again, many other distance functions exist. We mention these ideas now for primarily one reason. The basic concepts involving sequences and limits can be developed in the metric space setting. The concept of completeness is then based upon Cauchy sequences; a metric space (X, δ) is *complete* if and only if every Cauchy sequence in X has a limit in X. Given a metric space that is not complete in this sense, it is possible to enlarge it by including all limits of Cauchy sequences. This approach provides one method for defining the real numbers in terms of the rational numbers. Recall that our development assumes the existence of the real numbers and defines the rational numbers as a particular subset of the real numbers.

In a metric space there are notions of *open* and *closed* balls. Given $p \in X$, we write $B_r(p)$ for the set of points whose distance to p is less than r; we call $B_r(p)$ the open ball of radius r about p. The closed ball also includes points whose distance to p equals r. Depending on the metric δ, these sets might not resemble our usual geometric picture of balls. A subset S of a metric space is *open* if, for each $p \in S$, there is an $\epsilon > 0$ such that $B_\epsilon(p) \subset S$. In most situations what counts is the collection of open subsets of a metric space. It is possible and also appealing to define all the basic concepts (limit, continuous function, bounded, compact, connected, etc.) in terms of the collection of open sets. The resulting subject is called *point set topology*. In order to keep this book at an elementary level, we will not do so; instead we rely on the metric in $\mathbf{C}$ defined by $\delta(z, w) = |z - w|$. The distance between points in this metric equals the usual Euclidean distance between them.

Occasionally we will require that an open set be connected, and hence we give the definition here. In Definition 6.3 and a few other times in this book, the symbol $\varnothing$ denotes the empty set. The empty set is open; there are no points in $\varnothing$, and hence the definition of open is satisfied, albeit vacuously.

Definition 6.3. Let (X, δ) be a metric space. A subset S of X is called connected if the following holds: whenever $S = A \cup B$, where A and B are open and $A \cap B = \varnothing$, then either $A = \varnothing$ or $B = \varnothing$.

In Chapter 6 we will use the following fairly simple result. In it and in the subsequent comment we assume **C** is equipped with the usual metric.

Theorem 6.1. *An open subset Ω of* **C** *is connected if and only if each pair of points $z, w \in \Omega$ can be joined by a polygonal path whose sides are parallel to the axes.*

Proof. (Sketch.) The result holds (and is uninteresting) when Ω is the empty set. We therefore first assume Ω is connected and nonempty. Choose $z \in \Omega$. Consider the set S of points that can be reached from z by such a polygonal path. Then S is not empty, as $z \in S$. Also, S is open; if we can reach w, then we can also reach points in a ball about w. On the other hand, the set T of points we cannot reach from z is also open. By the definition of connectedness, Ω is not the union of disjoint open subsets unless one of them is empty. Thus, as S is nonempty, T must be empty. Hence $S = \Omega$ and the conclusion holds.

The proof of the converse statement is similar; we prove its contrapositive. Assume Ω is not connected and nonempty. Then $\Omega = A \cup B$, where A and B are open, $A \cap B = \varnothing$, but neither A nor B is empty. One then checks that no polygonal path in Ω connects points in A to points in B. Hence the contrapositive of the converse statement holds. □

We pause to state the intuitive characterization of connected subsets of the real line. A subset of **R** (in the usual metric) is connected if and only if it is an interval. By convention, the word interval includes the entire real line, the empty set, and semi-infinite intervals. We leave the proof as an exercise, with the following hints. Given a bounded connected set S, let α be its greatest lower bound and let β be its least upper bound. We claim that S must be the interval between α and β, perhaps including one or both of these end points. To check this assertion, suppose $\alpha < x_0 < \beta$. If x_0 is not in S, then $\{x \in S : x < x_0\}$ and $\{x \in S : x > x_0\}$ are open nonempty subsets of S violating the definition of connectedness. The same idea works when S is unbounded below or above. The converse (an interval is connected) proceeds by writing the interval as $A \cup B$, where A and B are nonempty *closed* sets. (Their complements are open.) Choosing points in each and successively bisecting the interval between them creates two monotone sequences with a common limit, which must then be in both sets. Hence $A \cap B$ is nonempty.

► **Exercise 1.36.** Complete the proof that a subset of **R** is connected if and only if it is an interval.

Finally we mention one more concept, distinct from *connectedness*, but with a similar name. Roughly speaking, an open and connected subset S of **C** is called *simply connected* if it has no holes. Intuitively speaking, S has a hole if there is a closed curve in S which surrounds at least one point not in S. For example, the complement of a point is open and connected, but it is not simply connected. The set of z for which $1 < |z| < 2$ is open and connected but not simply connected. In

Chapter 8 we will define the *winding number* of a closed curve about a point. At that time we give a precise definition of simple connectivity.

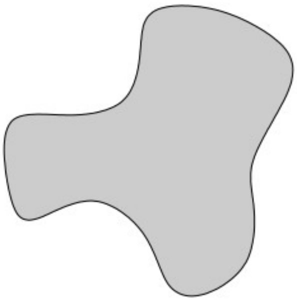

Figure 1.6. A connected and simply connected set.

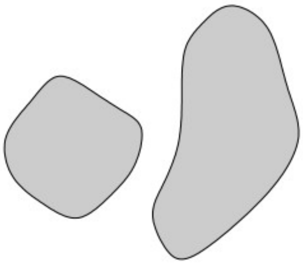

Figure 1.7. A disconnected set.

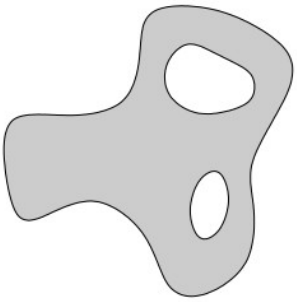

Figure 1.8. A connected but not simply connected set.

Books-018-2021/20210901/
→ Calculus-Math/
Calculus A Rigorous First Course.pdf
→ An Introduction to Complex Analysis and Geometry/
0821852744.pdf

Pag. 5(17) Fig 1.1 Difference of two squares

Pag. 21(33) 6. A glimpse at metric spaces.

20210902

7300 Townline Rd Lakeview,
MI 48850

pdfdrive.com
libgen.st
sanet.ws/books/
ebook3000.com/Science/

abakcus.com

Park yoshemite (Sequoias)

Troca de registro da água.

https://gallica.bnf.fr/ark:/
12148/bpt6k9730303m

→ Journal de Mathématiques
Pures et Appliquées.

109 années disponibles -
168 numéros

Source:
https://gallica.bnf.fr/ark:/12148/
cb343487840/date

O amor de Deus nos
prepara para a Eternidade

https://gallica.bnf.fr/ark:/
12148/bpt6k96431405 →

Application de lánalyse à la géométrie - G. Monge

→ 12148/bpt6k6227759f →
Recherches générales sur les surfaces courbes - Gauss
→ 12148/bpt6k3942g →
La théorie des parallèles
Lobatchevski

Nikolaj Ivanovič Lobačevskij
(1792-1856)

Vectors, St. Lines, Planes and Mechanics

Narayana Dash
ASIN: B00MQCNJLM
August 14, 2014

Terra : 4,54 bilhões anos ± 50 m

) $

20210906 (20210905)

Batalha de Waterloo 18 Jun 1815

Napoleão morreu em 1831 e os seus (ossos) restos mortais foram encaminhados para o Royal United Services Institute, onde foi preservado.

Marechal Grouchy ter é visto como o culpado pela derrota de Bonaparte por não ter evitado que os prussianos se unissem aos ingleses e por não ter acorrido à Waterloo com seus 30 mil homens, quando ouviu a canhonaria da batalha.

gbfproofreading.co.uk

mathsjam.com/cities/canterbury/

www.solipsys.co.uk/cgi-bin/MJ_Wiki.py

Sam Hartburn
sam.hartburn@gbfproofreading.co.uk
Phone 07594 602060.
(65)

20210905m (20210905)

Data-NovoLivro-018-2021/
Ramanujam-Works/
KSR chap3.pdf

3. Ramanujan's Notebooks

Pag. 7

"Modular equations and
approximations to π"

π like:

$$\frac{63}{25}\,\frac{17+15\sqrt{5}}{7+15\sqrt{5}}=3.1415926538\ldots$$

etc.

Source:

www.imsc.res.in/~rao/
ramanujan/images/
KSR Chap3.pdf (KSRchap3.pdf)

Obs: Prof. Rao (Srinivasa)
Institute of Mathematics Sciences
Chennai-600 020, India

Source: www.imsc.res.in/~rao/

3. Ramanujan's Notebooks

The history of the notebooks, in brief, is the following: Ramanujan had noted down the results of his researches, without proofs, (as in *A Synopsis of Elementary Results, a book on pure Mathematics,* by G.S. Carr), in three notebooks, between the years 1903 - 1914, before he left for England. These were the notebooks which he showed to his benefactors to convince them about his abilities as a mathematician. The results in these notebooks were organized by him. The first notebook has 16 chapters in 134 pages. The second is a revised, enlarged version of the first, containing 21 chapters in 252 pages. The third notebook contains 33 pages of unorganized material (included at the end of volume 2 of the facsimile edition [VII]). Ramanujan took these notebooks with him to Cambridge. But, in one of his letters to a friend, he wrote that he had no time to look into them and most probably he did not put them to use during his five year stay abroad.

The Preface to the *Collected Papers of Srinivasa Ramanujan* ([III], p.ix) contains the following appraisal of the work in the notebooks:

> *There is still a large mass of unpublished material. None of the contents of Ramanujan's notebooks has been printed, unless incorporated in later papers, except that one chapter, on generalized hypergeometric series, was analyzed by Hardy* [37] *in the Proceedings of the Cambridge Philosophical Society. This chapter is sufficient to show that, while the notebooks are naturally unequal in quality, they contain much which should certainly be published. It would be a very formidable task to work through them systematically, select particular passages, and edit these with adequate comment, and it is impossible to print the notebooks as they stand without further monetary assistance. The singular quality of Ramanujan's work and the romance which surrounds his career, encourage us to hope that this volume may enjoy sufficient success to make possible the publication of another.*

Hardy tried to persuade the University of Madras to undertake such a task and in turn, Prof. G.N. Watson was requested by the Madras University, in 1931, to edit the notebooks in a suitable form for publication. This was a formidable task since the notebooks contained about 4000 theorems.

Prof. Watson gave a lecture [39] on Ramanujan's notebooks at the London Mathematical Society in 1931, and published several papers between 1928–36, motivated by Ramanujan's work. Watson's Presidential address to the London Mathematical Society, in 1935, was on *mock-theta functions*, discovered by Ramanujan during the period when he was terminally ill (March 1919 – April 1920) after his return to India. Watson undertook the task of editing the notebooks with Prof. B.M. Wilson and

> *they focussed their attention on the second notebook. Chapters 2–13 were to be edited by Wilson, and Watson was to examine chapters 14–21. Unfortunately, Wilson passed away prematurely in 1935 at the age of 38.* (Berndt [40]).

This possibly put an end to the joint effort of Watson and Wilson to edit the notebooks.

In 1957, with monetary assistance from Sir Dadabai Naoroji Trust, at the instance of Prof. Homi J. Bhabha and Prof. K. Chandrasekaran of the Tata Institute of Fundamental Research, the Tata Institute published a facsimile edition of the notebooks of Ramanujan in two volumes. The reproduction is thoroughly faithful to the original. About a thousand copies were made then and in the centenary year of Ramanujan's birth, 1987, Springer - Narosa reprinted the same.

The formidable task of truly editing the notebooks –*viz.* to either prove each of the results or provide references to the literature where the proofs may be found – which was started by G. N. Watson and B. M. Wilson but never completed, was taken up in right earnest by Prof. Bruce C. Berndt, who has been publishing a series of papers, since 1981, by himself or in collaboration with others, including Indian Mathematicians Padmini T. Joshi, C. Adiga and S. Bhargava. He has also utilized the earlier work of G.N. Watson and B.M. Wilson on the notebooks with the Trinity College. This dedicated work of Berndt, published by Springer-Verlag [VIII], is now available in five parts, the first of which appeared in 1985 and the fifth, in 1997.

At the beginning of his work, in Part I, Berndt [34] points out:

> *The notebooks were originally intended primarily for Ramanujan's own personal use and not for publication ... Some of Ramanujan's incorrect 'theorems' in number theory found in his letters to*

> *Hardy have been well publicized. Thus, perhaps, some think that Ramanujan was prone to making errors. However, such thinking is erroneous. The notebooks contain scattered errors. Especially if one takes into account the roughly hewn nature of the material and his frequently formal arguments, Ramanujan's accuracy is amazing.*

Resurgence of interest

After Ramanujan returned to India, though he was seriously ill and bedridden for most of the time, he was working continuously. His life-long habit had been to work on a slate and transfer the final results to the notebook or sheets of paper. His brother-in-law has stated [59] that *a vast quantity of papers containing his notes was handed over to the University after his death.* Three days after Ramanujan died, his brother S. Lakshminarasimhan, in a letter to Prof. Hardy, dated 29th April 1920, informing him about his demise, wrote :

> *All his Mss that were in his trunk were handed over on the day of his death to Mr. Ramachandra Row's son in law, since the former is at Nilgiris. Not only those but also the journals, magazines - all he possessed except the books which he had, were taken away by them.*

In a letter to Prof. Hardy, dated 3 Dec. 1920, Mr. R. Ramachandra Row, wrote:

> *Ramanujan's M.S.S whatever they are are with me and will be handed over intact to the University of Madras who I understand is already in correspondence with you regarding the methods of publication.*

These papers reached Prof. Hardy, who should have passed them on to Prof. G.N. Watson and hence to their final destination – viz. the estate of Prof. Watson, in the underground cellar of Trinity College – after he wrote several papers inspired Ramanujan's results, until their discovery in 1976 by George E. Andrews.

Prof. Watson died in Feb. 1965. A decade later, in the spring of 1976, when Prof. George Andrews [53] of the Pennsylvania State University was going through the estate of Watson, he discovered a box of papers.

> *The most interesting item in this box was a manuscript of more*

> *than one hundred pages in Ramanujan's distinctive handwriting which contains over six hundred mathematical formulae listed one after the other without proof. It is my contention that this manuscript, or notebook, was written during the last year of Ramanujan's life after his return to India from England. My evidence for this assertion is all indirect; in the words of Stephen Leacock, 'It is what we call circumstantial evidence — the same that people are hanged for'* .

These are now referred to as the contents of the *Lost* notebook of Ramanujan. Andrews assessed ([101], p.xii):

> *While it is impossible to categorize the various formulas completely, a rough approximation of its contents is the following:*
>
> *q-series and related topics including mock θ-functions: 60%*
> *Modular equations and relation, singular moduli: 30%*
> *Integrals, Dirichlet series, congruences, asymptotics, misc.: 10%*
>
> *I give only this rough break down because of the chaotic nature of this manuscript. On many pages there are fragments of numerical computations and infinite series running off in all directions. It may be possible eventually to make sense out of some of these pages; if so, the above percentage breakdown may change. ...*
> *In any event it is clear that q-series investigations make up the bulk of the work in in the Lost Notebook. Indeed I count about 380 formulae that I consider to belong more to q-series than to modular relations or other topics.*

Sixty-seven years after the death of Ramanujan, due to the discovery of the 'lost' notebook by Prof. George Andrews, in 1976, and the editing of the three Notebooks of Ramanujan by Prof. Bruce Berndt, there has been a resurgence of interest in the work of Ramanujan. These notebooks of Ramanujan have formed the basis for numerous papers by many mathematicians, who gave proofs of the theorems and conjectures of Ramanujan obtained by him through his intuition and sheer brilliance. This is a singular and unparalleled phenomenon in the annals of mathematics. In the centenary year of his birth, 1987, several conferences were held in many countries to focus attention on his work and the recent developments. Mystery still surrounds *some of his* *work, mostly in the theory of elliptic functions*, wherein

> *it is not possible, after all the work of Watson and Mordell, to draw the line between what he may have picked up somehow and what he must have found for himself.* ([IV], p.10).

Hardy [60] regretted that he could easily have but did not find out all the details from Ramanujan himself since he

> *was quite able and willing to give a straight answer to a straight question, and not in the least disposed to make a mystery of his achievements.*

However, Hardy [60] argues that Ramanujan

> *was not particularly interested in his own history or psychology; he was a mathematician anxious to get on with the job. And after all I too was a mathematician, and a mathematician meeting Ramanujan had more interesting things to think about than historical research. It seemed ridiculous to worry him about how he had found this or that known theorem, when he was showing me half a dozen new ones almost every day.*

A Page from the second Notebook

To provide a feeling for Ramanujan's notebooks, we reproduce here a page [61] from his second Notebook, Chapter XVIII. There is no particular reason for choosing this page except that it contains an entry which has become a 'folklore' which illustrates the attachment Ramanujan had to numbers prompting Littlewood to state that to Ramanujan every number is a *personal friend.*

There are 2 parts shown in this page. The first part is concerning the geometrical construction of a square whose area is equal to that of a given circle. This is a rare piece of Ramanujan's work in geometry, since he did very little work in geometry and perhaps did not have much interest in it. Several attempts on this problem have been made from early times by (ancient Greek and Hindu) mathematicians and it has been shown that it is impossible

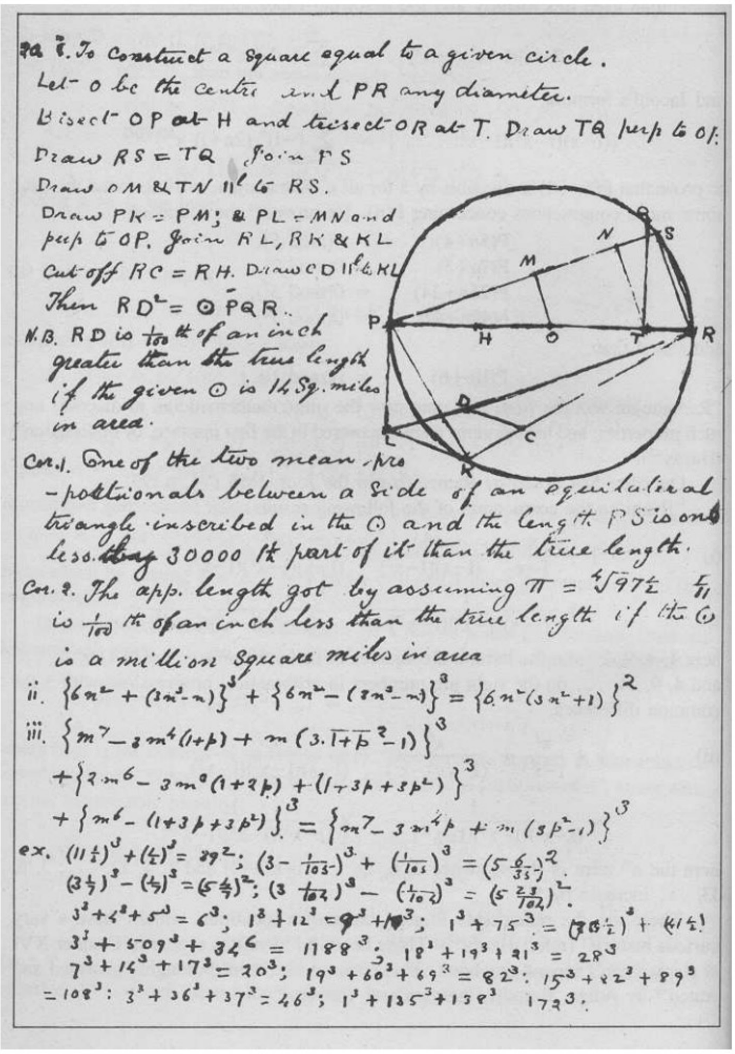

7a i. To construct a square equal to a given circle.
Let O be the centre and PR any diameter.
Bisect OP at H and trisect OR at T. Draw TQ perp to OP.
Draw RS = TQ Join PS
Draw OM & TN $\|^{l}$ to RS.
Draw PK = PM; & PL = MN and perp to OP. Join RL, RK & KL
Cut off RC = RH. Draw CD $\|^{l}$ to KL
Then $RD^2 = \odot PQR$.
N.B. RD is $\frac{1}{100}$th of an inch greater than the true length if the given $\odot$ is 14 Sq. miles in area.
Cor. 1. One of the two mean proportionals between a side of an equilateral triangle inscribed in the $\odot$ and the length PS is one ~~than~~ less. 30000th part of it than the true length.
Cor. 2. The app. length got by assuming $\pi = \sqrt[4]{97\tfrac{1}{2} - \tfrac{1}{11}}$ is $\frac{1}{100}$th of an inch less than the true length if the $\odot$ is a million square miles in area

ii. $\{6n^2 + (3n^3 - n)\}^3 + \{6n^2 - (3n^3 - n)\}^3 = \{6n^2(3n^2+1)\}^2$

iii. $\{m^7 - 3m^4(1+p) + m(3\cdot\overline{1+p}^2 - 1)\}^3$
$+ \{2m^6 - 3m^3(1+2p) + (1+3p+3p^2)\}^3$
$+ \{m^6 - (1+3p+3p^2)\}^3 = \{m^7 - 3m^4 p + m(3p^2-1)\}^3$

ex. $(11\tfrac{1}{2})^3 + (\tfrac{1}{2})^3 = 39^2$; $(3 - \tfrac{1}{105})^3 + (\tfrac{1}{105})^3 = (5\tfrac{6}{35})^2$.
$(3\tfrac{1}{7})^3 - (\tfrac{4}{7})^3 = (5\tfrac{4}{7})^2$; $(3\tfrac{1}{102})^3 - (\tfrac{1}{102})^3 = (5\tfrac{23}{102})^2$.
$3^3 + 4^3 + 5^3 = 6^3$; $1^3 + 12^3 = 9^3 + 10^3$; $1^3 + 75^3 = (38\tfrac{1}{2})^3 + (41\tfrac{1}{2})$
$3^3 + 509^3 + 34^6 = 1188^3$; $18^3 + 19^3 + 21^3 = 28^3$
$7^3 + 14^3 + 17^3 = 20^3$; $19^3 + 60^3 + 69^3 = 82^3$: $15^3 + 82^3 + 89^3$
$= 108^3$: $3^3 + 36^3 + 37^3 = 46^3$; $1^3 + 135^3 + 138^3$ 172^3.

Fig. caption: The last three lines of p. 224 and the p. 225 from the second notebook of Ramanujan, chapter XVIII. ([61]).

to *square the circle*, in the Euclidian sense[1]. Hardy has commented that he found in Ramanujan's collection of books in Cambridge some books by 'quacks' on this theme.

Ramanujan published a paper [62] entitled, "*Squaring the circle*" in the Jour. of the I.M.S. in 1913, in which he gives a geometrical construction for finding the length of the side of a square whose area equals that of the circle. He also reproduced this and another geometrical construction for π in his later paper [63] on "*Modular equations and approximations to* π". In this paper, he deduced several formulae[2] for π like:

$$\frac{63}{25}\frac{17+15\sqrt{5}}{7+15\sqrt{5}} = 3.14159265380...$$

$$\pi = \frac{24}{\sqrt{142}} log \left\{ \sqrt{\left(\frac{10+11\sqrt{2}}{4}\right)} + \sqrt{\left(\frac{10+7\sqrt{2}}{4}\right)} \right\}$$

$$\pi = 12\sqrt{190}\ log\left\{(2\sqrt{2}+\sqrt{10})(3+\sqrt{10})\right\}$$

formulae correct to 9, 16 and 18 decimal places; and

$$\frac{1}{\pi} = 2\sqrt{2}\left\{\frac{1103}{99^2} + \frac{27493}{99^6}\cdot\frac{1}{2}\cdot\frac{1.3}{4^2} + \frac{53883}{99^10}\cdot\frac{1.3}{2.4}\cdot\frac{1.3.5.7}{4{\cdot}8^2} + \cdots\right\}$$

$$= 2\sqrt{2}\sum_{n=0}^{\infty}\frac{(1/4)_n(1/2)_n(3/4)_n}{(1)_n(1)_n(1)_n}(1103+26390n)(\frac{1}{99})^{4n+2},$$

which he asserted would be "*rapidly convergent*", for the first term itself gives the sum to 8 decimal places. In 1986, Borwein and Borwein [64], two computer scientists used a version of Ramanujan's formula to calculate π to

---

[1] It is the ancient Greeks who set-up the circle-squaring problem with the following two conditions [83]: the solution should use only a straightedge and compass, so that the proofs can be reduced to Euclid's theorems and the solution must use only a finite number of steps. If, in fact, $\pi r^2 = a^2$, where r is the radius of the circle and a is the length of the side of a square, then it follows that $\pi = (a/r)^2$, or that π is rational. In fact, π has been shown to be not only an *irrational number* but also, proved by Ferdinand von Lindermann, in 1882, to be a *transcendental* number. Furthermore, the phrase 'squaring the circle' in common usage suggests a project doomed to failure!

[2] Note that Ramanujan has used the equality sign for his approximations to π even though he states explicitly that the formulae are correct only to a certain number of decimal places.

17 million places and found that the formula converges on the exact value with far greater efficiency than any previous method[3]. This success proved that Ramanujan's insight was correct.

The next entries on this page in Ch. XVIII of his second note book are solutions of Diophantine equations:

$$\begin{aligned} X^3+Y^3 &= U^2 \\ \text{and} \qquad X^3+Y^3+Z^3 &= U^3 \quad \text{(Euler's equation).} \end{aligned}$$

which are followed by numerical examples for these equations – 6 solutions for the first and 12 solutions for the Euler's equation and one which looks like but is not a solution of the Euler's equation. Characteristically, Ramanujan does not give the values of the parameters for the examples (ref. [VIII], part III, p.199 for the examples with the values of the parameters appended).

Ramanujan's solution of Euler's equation in rational numbers given here is:

$$\begin{aligned} X &= m^7 - 3m^4(1+p) + m(2+6p+3p^2), \\ Y &= 2m^6 - 3m^3(1+2p) + 1 + 3p + 3p^2, \\ Z &= m^6 - 1 - 3p - 3p^2, \\ U &= m^7 - 3m^4 p + m(3p^2-1), \end{aligned}$$

where m and p denote arbitrary numbers. He has also given another solution of this equation elsewhere (see Question 441 which follows) but neither of these two parameter solutions is the general solution[4] which contains three parameters, prompting Hardy [65] to comment that

> *Diophantine equations should have suited him, but he did comparatively little with them, and what he did was not his best.*

The second of the 12 examples given for the Euler equation on this page:

$$1^3 + 12^3 = 9^3 + 10^3 \quad (= \ 1729)$$

---

[3] David and Gregory Chudnovsky are two mathematicians who hold several world records for calculating the highest number of digits of π: first up to 450 million digits, then 1 billion and then 2 billion digits. The current record is over 51 billion digits (cf. For a brief history of π, [83]).

[4] The last of the examples $1^3 + 6^3 + 8^3 = 9^3$ does not belong to this class of solutions, since no values of m and p yield this solution to the Euler's equation. So, its presence reveals that perhaps, Ramanujan was aware that his solution was not the most general for Euler's equation.

corresponds to ($m = 2$, $p = 3$). This has been made famous by the following anecdote associated with it, due to Hardy [66]:

> *He could remember the idiosyncrasies of numbers in an almost uncanny way. It was Littlewood who said that every positive integer was one of Ramanujan's personal friends. I remember going to see him once when he was lying ill in Putney. I had ridden in taxi-cab number 1729, and remarked that the number seemed to me rather a dull one, and that I hoped that it was not an unfortunate omen. 'No', he replied, 'it is a very interesting number, it is the smallest number expressible as a sum of two cubes in two different ways'. I asked him, naturally, whether he could tell me the solution of the corresponding problem for fourth powers; and he replied, after a moment's thought, that he knew no obvious example, and supposed that the first such number must be very large.*

The simplest known solution of:

$$x^4 + y^4 = z^4 + t^4$$

is Euler's: $59^4 + 158^4 = 133^4 + 134^4 = 635318657.$
Euler gave a solution involving two parameters, but no 'general' solution is known.

The anecdote reveals Ramanujan's remarkable feeling for numbers and his sharp memory which made him recall one entry out of several thousands he had made in his notebook just like that and the fact that he had not recorded in his notebook the observation he made about 1729 which came from him only when Prof. Hardy made an innocuous statement regarding the taxi number.

Though the number 1729 itself finds no explicit mention in the notebooks, the Euler's equation and/or its solution: $1^3 + 12^3 = 9^3 + 10^3$ finds mention[5] in three Questions to the Journal of the I.M.S. ([III], p.331) and in two places in his notebooks ([VII], Vol. 2, p.266 and 387). These are indicated below:

---

[5] I am thankful to Mr. P.K. Srinivasan for drawing my attention to these and his publications [81].

- **Question 441** (V, 39)[6]:
Shew that

$$(6a^2-4ab+4b^2)^3=(3a^2+5ab-5b^2)^3+(4a^2-4ab+6b^2)^3+(5a^2-5ab-3b^2)^3,$$

and find other quadratic expressions satisfying similar relations.
[*Solution* by S. Narayanan, VI, 226.]

- **Question 661** (VII, 119):
Solve in integers $x^3+y^3+z^3=u^6$,
and deduce the following:

$$\begin{array}{rclrcl} 6^3-5^3-3^3 &=& 2^6, & 8^3+6^3+1^3 &=& 3^6,\\ 12^3-10^3+1^3 &=& 3^6, & 46^3-37^3-3^3 &=& 6^6,\\ 174^3+133^3-45^3 &=& 14^6, & 1188^3-509^3-3^3 &=& 34^6. \end{array}$$

[*Solutions by* N.B. Mitra, XIII, 15 – 17. Additional solution and remarks by N.B. Mitra, XIV, 73 – 77.]

- **Question 681** (VII, 160):
Solve in integers $x^3+y^3+z^3=1$,
and deduce the following:

$$\begin{array}{rclrcl} 6^3+8^3 &=& 9^3-1, & 9^3+10^3 &=& 12^3+1,\\ 135^3+138^3 &=& 172^3-1, & 791^3+812^3 &=& 1010^3-1,\\ 11161^3+11468^3 &=& 14258^3+1, & 65601^3+67402^3 &=& 83802^3+1. \end{array}$$

[*Partial solution by* N.B. Mitra, XIII, 17. See also N.B. Mitra, XIV, 73–77 (76–77).]

- If p,q,r are quantities so taken that $p+3a^2=q+3ab=r+3b^2=(a+b)^2$ and m and n are any two quantities, then

$$n(mp+nq)^3+m(mq+nr)^3=m(np+mq)^3+n(nq+mr)^3.$$

A particular case of the above theorem is:

$$(3a^2+5ab-5b^2)^3+(4a^2-4ab+6b^2)^3+(5a^2-5ab-3b^2)^3=(6a^2-4ab+4b^2)^3.$$

([VIII], Vol.2, p.266)

---

[6] J. of the IMS, vol. V, p.39.

- If $\alpha^2 + \alpha\beta + \beta^2 = 3\lambda\gamma^2$, then

$$(\alpha + \lambda^2\gamma)^3 + (\lambda\beta + \gamma)^3 = (\lambda\alpha + \gamma)^3 + (\beta + \lambda^2\gamma)^3.$$

([VIII], Vol.2, p.387)

Note that in the above 4-parameter solution for the Euler's equation, it has been observed (cf. [95] (c), p.43) that for $\alpha = 3, \beta = 0, \lambda = 3$ and $\gamma = 1$, we get $12^3 + 1^3 = 10^3 + 9^3 = 1729$. The fact that Ramanujan could come out spontaneously with the statement that 1729 in the smallest number which could be expressed as the sum of two cubes in two different ways, something which he had not jotted down in his notebooks, proves that every integer is a 'personal friend' of Ramanujan as remarked by Prof. Littlewood.

Hardy's statement that two thirds of Srinivasan Ramanujan's work in India, (contained in the notebooks), consisted of rediscoveries, is refuted by Prof. Bruce C. Berndt, who states after his extensive studies that Hardy's estimate is too high. Berndt's main goal was: "*To prove each of Ramanujan's theorems*" and for known results refer to literature where proofs may be found. Tables of the contents of "Ramanujan's Notebooks", Parts I to V, edited by Prof. Berndt are given here in the following pages. The titles of the chapters gives the reader an idea about the nature and range of problems and the number of results Ramanujan had recorded in each chapter.

In the Introduction to Part V, Berndt states:

> *This volume, however, should not be regarded as the closing chapter on Ramanujan's notebooks. Instead, it is just the first milestone on our journey to understanding Ramanujan's ideas. ... It is our fervent wish that these volumes will serve as springboards for further investigations by mathematicians intrigued by Ramanujan's remarkable ideas. ... Ramanujan remarked that several of his series arose from alternative theories of elliptic functions ... The first of the three alternative theories is the most interesting and the most important, and we feel that a large body of work remains to be discovered here.*

Ramanujan's Note Books, Part I, Bruce C. Berndt ([VIII], 1985):

Ch.	Subject	# of Results
1.	Magic Squares	43
2.	Sums Related to the Harmonic series or the Inverse Tangent Function	68
3.	Combinatorial Analysis and Series Inversions	86
4.	Iterates of the Exponential Function and an Ingenius Formal Technique	50
5.	Eulerian Polynomials and Numbers, Bernoulli Numbers and the Riemann Zeta-Function	94
6.	Ramanujan's Theory of Divergent series	61
7.	Sums of Powers, Bernoulli Numbers and the Γ Function	110
8.	Analogues of the Gamma Function	108
9.	Infinite Series Identities, Transformations, and Evaluations	139
	Ramanujan's Quarterly Reports	

Ramanujan's Note Books, Part II, Bruce C. Berndt ([VIII], 1989):

Ch.	Subject	# of Results
10.	Hypergeometric Series I	116
11.	Hypergeometric Series II	103
12.	Continued Fractions	113
13.	Integrals and Asymptotic Expansions	92
14.	Infinite Series	87
15.	Asymptotic Expansions Modular Forms	94

Ramanujan's work on hypergeometric series, contained in chapters X and XI of his second notebook was edited by Hardy, as stated earlier. In chapter 4 of this book, we take a closer look at some of the beautiful work of Ramanujan on hypergeometric series.

Ramanujan's Notebooks, Part III, Bruce C. Berndt ([VIII], 1991):

Ch.	Subject	# of Results
16.	q-Series and Theta Functions	134
17.	Fundamental Properties of Elliptic Functions	162
18.	The Jacobian Elliptic Functions	135
19.	Modular Equations of Degrees 3, 5 and 7 and Associated Theta Function Identities	185
20.	Modular Equations of Higher & Composite Degrees	173
21.	Eisenstein series	45

The 100 Pages of Unorganized material after Chapter 21 of Srinivasa Ramanujan's second Notebook and the 33 Pages of his third Notebook were organized into chapters 22 to 31 by Berndt.

Ramanujan's Notebooks, Part IV, Bruce C. Berndt ([VIII], 1994):

Ch.	Subject	# of Results
22	Elementary results	47
23.	Number Theory	108
24.	Theory of Prime numbers	24
25.	Theta Function and Modular Equations	86
26.	Inversion Formulas for Lemniscate & other functions.	10
27.	Q series	9
28.	Integrals	63
29.	Special Functions	39
30.	Partial Fraction Expansions	15
31.	Elementary and miscellaneous analysis	36
	16 Chapters of first Notebook	54

In the concluding part V, Berndt examines the unorganized pages in all the three notebooks of Ramanujan on continued fractions, alternative theories of elliptic functions, class invariants and singular moduli, explicit values of theta-functions, modular equations, infinite series, approximations and asymptotic expansions, etc. Berndt points out that *very few claims in this volume pertain to Ramanujan's published papers and problems.*

Ramanujan‘s Notebooks, Part V, Bruce C. Berndt ([VIII], 1997)[7]:

Ch.	Subject	# of Results
32.	Continued Fractions	73
33.	Ramanujan's Theories of Elliptic Functions to Alternative Bases	62
34.	Class Invariants and Singular Moduli	196
35.	Values of Theta-Functions	24
36.	Modular Equations and Theta-Function Identities in Notebook 1	87
37.	Infinite Series	53
38.	Approximations and Asymptotic Expansions	46
39.	Miscellaneous Results in the First Notebook	24

Many theorems communicated by Ramanujan to Hardy in his Jan.16, Feb. 27, 1913 letters are found in Chapters.22 – 31. In all, $759+605+834+491+565 = 3254$ theorems have been studied in the parts I–V by Prof. Berndt. Hardy had estimated that the notebooks of Ramanujan contained approximately 3000 – 4000 statements of theorems and besides editing one chapter of Ramanujan's second Notebook, in 1923, he tried to get them edited. His intentions have at long last been fulfilled by the tireless efforts of the indefatigable Bruce C. Berndt. For the motivated student of Mathematics, this five part publication will provide a plethora of results for further studies. As stated elsewhere [XV], it is no exaggeration to say that as long as people do mathematics, the work of Ramanujan and the stupendous effort of Prof. Bruce C. Berndt in editing the Ramanujan notebooks will be appreciated.

---

[7] The author thanks Prof. Berndt for readily communicating the Preface and the Contents pages of Part V, by e-mail, on request.

R.P. Agarwal's work on the notebooks

Prof. Ratan P. Agarwal is the founder of a school of ordinary and basic hypergeometric series in India. In recent times, he has consolidated a part of his work in a publication entitled: *Resonance of Ramanujan's Mathematics* [82]. This two volume work concerns some of the results of Srinivasa Ramanujan (1887 - 1920), the renowned Indian Mathematician, recorded in his notebooks, which continue to be a perennial source of inspiration for generations of Mathematicians.

Volume I has five chapters. Chapter 1 concerns certain summation and asymptotic formulae for generalized (ordinary and basic) hypergeometric series. Chapter 2 is on Rogers-Ramanujan identities. Chapters 3 - 5 deal with certain definite Integrals of interest to Ramanujan.

Prof. Agarwal opines, like Hardy, that *a major part of Ramanujan's systematic work on ordinary hypergeometric series has been in the nature of a rediscovery of the work already done by previous authors* – the ${}_7F_6(1)$ summation formula of Dougall (1907) and particular cases of this sum, such as, Pfaff (1797) - Saalschütz (1890) theorem for the ${}_3F_2(1)$ and the Gauss and Kummer summation theorems for the ${}_2F_1(1)$ and the ${}_2F_1(-1)$, respectively, and Dixon's summation theorem for a well-poised ${}_3F_2(1)$. These and Ramanujan's entries in his notebooks concerning the deep results on the asymptotic relations for a finite number of terms of a variety of ordinary hypergeometric series, which center around the behaviour of the logarithmic solutions of the hypergeometric differential equation are presented in Chapter 1.

Chapter 2 deals with the derivations of the celebrated Rogers–Ramanujan identities[8], their combinatorial interpretations, their generalizations and their applications in Statistical Mechanics for the solution of the two-dimensional hard-hexagonal model (by R.J. Baxter). Several proofs of these identities and the interesting mathematical developments have been described very well in this cogently written chapter.

---

[8] Refer chapter 2 of this work for a discussion about these famous identities.

The next three chapters of Volume I deal with some of the integrals studied by Ramanujan. Prof. Agarwal presents definite integrals associated with Fourier transforms in the second Notebook and self-reciprocal functions given in his 'lost' Notebook in Chapter 3. He has shown how some of Ramanujan's results on integrals in his Notebook can be established with the help of other entries of his

> *that every entry given in the notebooks by him has a purpose, although sometimes a particular entry may look out of context at the first instance. ... We have advocated elsewhere also that it may be interesting many a time to try to prove 'unproven' entry of Ramanujan with the help of some other entry given by him earlier or later on.*

Chapter 4 is focused on Ramanujan's Master Theorem for the evaluation of definite integrals, viz.:

$$\int_0^\infty x^{n-1}F(x)dx = \Gamma(n)\phi(-n), \qquad \text{where } F(x) = \sum_{k=0}^{\infty} \frac{\phi(k)(-x)^k}{k!},$$

in some neighbourhood of $x = 0$ and for n not necessarily a positive integer. The three Quarterly Reports submitted by Ramanujan to the Madras University, as a requirement stipulated by the University for awarding him the *First Research Scholarship in Mathematics*, contain various applications of this theorem to the *evaluation of a huge variety of integrals* and expansion formulae. The appropriate conditions of convergence for its validity have been provided by Hardy in his lectures contained in *Ramanujan* [IV]. The q-extension of the Master Theorem by Jackson (1951) and its extension to two variables by R.P. Agarwal (1974) are presented. The fundamental gamma and beta functions of mathematical analysis have been ingeniously extended by Ramanujan and these are related to the 'incomplete' gamma and beta functions (in which the upper limit of the integrals is finite and not ∞). The q-extensions of these works by several mathematicians including

Thomae (1879), F.H. Jackson (1904), R. Askey (1978), G.E. Andrews and R. Askey (1981), A. Verma and V.K. Jain (1992) are also included.

Chapter 5 is devoted to the study of integrals of the type:

$$\int_{-\infty}^{\infty} \frac{e^{at^2+bt}}{e^{ct}+d} dt$$

for particular values of a, b, c, d, evaluated by Ramanujan. Mordell (1933) classified these according to the values of the parameters and evaluated them in terms of Jacobi's theta and other related functions, using the method of complex contour integration. Professor Agarwal presents the general method of evaluation of different standard forms of this integral and relates them to those studied by Ramanujan, who used transform calculus to evaluate his integrals. Integrals associated with Riemann's zeta functions and elliptic modular relations are not discussed, as pointed out by the author himself. The chapter concludes with the mention of the example of a definite integral with fractional derivatives in the integrand stated in the third Quarterly Report. Ramanujan's formal deduction is based on his Master theorem and it agrees with the classical Liouville's definition of a fractional derivative of a function.

In Volume II, Prof. Agarwal presents a critical appraisal of Ramanujan's extensive and intriguing work on elliptic functions. Results pertaining to theta functions, partial theta functions, 'mock' theta functions (discovered by Ramanujan during the last year of his life on his return to India), as well as Lambert series and their relationship with elliptic functions, 'mock' theta functions and allied functions. While theta functions and partial theta functions are covered in chapter 1 of this volume and Lambert series and related functions are dealt with in chapter 5, the major part of this volume is concerned with the 'mock' theta function results of Ramanujan, presented in three extensive chapters (2, 3 and 4), contained in the 'Lost' Notebook, which was made available to the world in a facsimile edition, by Narosa, for the first time in the centenary year of Ramanujan's birth 1987 (though it was discovered in the estate of G.N. Watson by Prof. George E. Andrews of the Pennsylvania State University, in Spring 1976).

After giving the definitions of the four Jacobi theta functions and the 'false' theta functions (viz. theta functions with the 'wrong' signs for the terms) defined by L.J. Rogers, Prof. Agarwal introduces the Ramanujan notation for the theta function in terms of his symmetric function:

$$f(a,b) = \sum_{k=-\infty}^{\infty} (ab)^{k(k-1)/2}(a^k + b^k), \quad |ab| < 1.$$

In terms of this symmetric function, Ramanujan defines:

$$f(qe^{2iz}, qe^{-2iz}) = \theta_3(z,q), \text{ where } q = e^{i\pi\tau} \text{ and } |q| < 1.$$

Several general properties of the function $f(a,b)$ are then presented. Jacobi's triple product identity and other such identities due to D. Hickerson (1988) and G.E. Andrews and D. Hickerson (1991) are derived explicitly, due to their importance to simplify calculations in later chapters on 'mock' theta functions. A number of theta function expansions in the 'lost' notebook pertain to 'partial' theta functions. Some of these are stated and G.E. Andrews (1981) derived a general basic hypergeometric identity and showed the results of Ramanujan as special cases of it.

Ramanujan defined four third order mock theta functions and to this set three more were added by Watson. Watson used the transformations of basic hypergeometric series for obtaining new definitions for all the seven functions. Agarwal himself has contributed significantly to the theory of 'mock' theta functions by showing that most of the general identities of Andrews (referred above) belong to a very general class of basic hypergeometric transformations. His results are presented and those of Andrews deduced from them. Furthermore, Agarwal defines the third order mock theta functions in terms of a ${}_2\Phi_1$ basic hypergeometric series and deduces many of the mock theta function properties from the well-known properties of the ${}_2\Phi_1$. Ramanujan gave ten mock theta functions of order five, in two groups of five each, and three functions of order seven. He asserted that the members of each of the two groups of mock theta functions of order five are related amongst those belonging to the same group only, while the three mock theta functions of order seven are not related to each other. The works of Andrews, Agarwal and M. Gupta on the mock theta functions of order five are presented in detail. A. Gupta and Agarwal showed that mock theta functions of order five and seven are defined through basic hypergeometric series of type ${}_3\Phi_2$ and ${}_4\Phi_3$. They have also made attempts to develop the theory of the above types of basic hypergeometric series to obtain new transformations and definitions for certain mock theta functions of orders five and seven. It is pointed out that the absence of a general transformation theory for the ${}_3\Phi_2(a, b, c; e, f; z)$ series [except when the argument z is q (the base parameter itself) or ef/abc], is the main reason why there exists no general transformation theory for mock theta functions of order five.

Observations of Agarwal on the relationship between the 'mock' theta functions and the basic hypergeometric series lead him to define:

> *the order of a mock theta function as* $(2r+1)$ *if it is expressible in terms of a* ${}_{r+1}\Phi_r$ *series, on a single base* q^k, $k \leq r+1$. *There may be in the definition of the mock theta function an additive term with* ${}_{r+1}\Phi_r$ *consisting of* θ*-products, which do not affect the order.*

While this is an intersting observation by itself, it is simplistic, to say the least. For, mystery still surrounds the incomplete work of Ramanujan in this area contributing to the belief that Ramanujan was working on a general theory of mock theta functions, whose preliminary results are the ones in the 'lost' notebook of his and that he was snatched away by fate before he could reveal his grand scheme!

This two volume work of Prof. Agarwal is a deep study of some of the work of Ramanujan on ordinary and basic hypergeometric series, on definite integrals and theta, partial theta and mock theta functions. It is very original in parts and has a minimum overlap with the comprehensive five part work on the notebooks of Ramanujan by Bruce C. Berndt[9]. This work presents, in as coherent a way as is possible, the work of Agarwal and his school of students and coworkers. It is a valuable supplement for the research scholar and for mathematicians and presents the author's understanding of the 'how' and 'why' of what Ramanujan did.

---

[9] The author's reference to these Parts I to V as the Berndt Notebooks is unfortunate, especially in view of the significance of the Ramanujan notebooks and the 'Lost' notebook, which are the source materials for both Berndt and Agarwal.

Dyson on Ramanujan's notebooks

Freeman J. Dyson, the renowned Physicist ([XVI], p.25), concludes his article entitled: *A Walk through Ramanujan's Garden*, with the following advise:

> *In conclusion, I would like to urge all of you who are working in the many fields of mathematics which have been enriched by Ramanujan's ideas to go back to the source the collected papers and the notebooks. . . . The notebooks . . . are now appearing in a splendidly annotated version edited by Bruce Berndt [VIII]. The "lost" notebook is now accessible to us through the devoted labors of George Andrews ([53], [56]). When I started my walk through Ramanujan's garden 47 years ago, only the collected papers were available. A year after I chose Hardy and Wright's "Theory of Numbers" (Oxford, Clarendon Press, 1938) as a school prize, I won another prize. For the second prize I chose Ramanujan's collected papers. The collected papers have traveled with me from England to America and are still as fresh to-day as they were in 1940. Whenever I am angry or depressed, I pull down the collected papers from the shelf and take a quiet stroll in Ramanujan's garden. I recommend this therapy to all of you who suffer from headaches or jangled nerves. And Ramanujan's papers are not only a good therapy for headaches. They also are full of beautiful ideas which may help you to do more interesting mathematics.*

Certainly several of us will not find this therapy useful. Since Dyson, Erdös and Selberg were among the very best in their chosen fields, they were greatly benefited by an early introduction to the Collected Papers. The rest can derive the benefit of the extensive study of Ramanujan's work by Hardy [IV] and Berndt [VIII] to begin appreciating the prolific, original work of the mathematical genius in his published work and in his notebooks.

2 *+,-* * *. ! * " ,! ! D!
1#

Prof. Stenio Musich
www.steniomusich.com

20210906

Published
by
Journal of The Indian Mathematical Society

https://www.imsc.res.in/~rao/ramanujan

https://www.imsc.res.in/~rao/ramanujan/collectedpapers/collectedright1.htm
https://www.imsc.res.in/~rao/ramanujan/collectedpapers/modularEqn/modularEqn1.htm

MODULAR EQUATIONS AND APPROXIMATIONS TO π.

By S. Ramanujan.

1. If we suppose that

$$(1)\quad (1+e^{-\pi\sqrt{n}})(1+e^{-3\pi\sqrt{n}})(1+e^{-5\pi\sqrt{n}})\ldots = 2^{\frac14}e^{-\pi\sqrt{n}/24}G_n$$

and

$$(2)\quad (1-e^{-\pi\sqrt{n}})(1-e^{-3\pi\sqrt{n}})(1-e^{-5\pi\sqrt{n}})\ldots = 2^{\frac14}e^{-\pi\sqrt{n}/24}g_n,$$

then G_n and g_n can always be expressed as roots of algebraical equations when n is any rational number. For we know that

$$(3)\quad (1+q)(1+q^3)(1+q^5)\ldots = 2^{\frac16}q^{\frac1{24}}(kk')^{-\frac1{12}}$$

and

$$(4)\quad (1-q)(1-q^3)(1-q^5)\ldots = 2^{\frac16}q^{\frac1{24}}k^{-\frac1{12}}k'^{\frac16}.$$

Now the relation between the moduli k and l, which makes

$$n\frac{K'}{K} = \frac{L'}{L},$$

where $n = r/s$, r and s being positive integers, is expressed by the modular equation of the rs^{th} degree. If we suppose that $k = l'$, $k' = l$, so that $K = L'$, $K' = L$, then

$$q = e^{-\pi L'/L} = e^{-\pi\sqrt{n}},$$

and the corresponding value of k may be found by the solution of an algebraical equation.

From (1), (2), (3), and (4) it may easily be deduced that

$$(5)\quad g_{4n} = 2^{\frac14}g_nG_n,$$

$$(6)\quad G_n = G_{1/n},\quad 1/g_n = g_{4/n},$$

$$(7)\quad (g_nG_n)^8(G_n^8 - g_n^8) = \tfrac14.$$

I shall consider only integral values of n. It follows from (7) that we need only consider one of G_n or g_n for any given value of n; and from (5) that we may suppose n not divisible by 4. It is most convenient to consider g_n when n is even, and G_n when n is odd.

2. Suppose then that n is odd. The values of G_n and g_{2n} are got from the same modular equation. For example, let us take the modular equation of the 5th degree, viz.,

$$(8)\quad \left(\frac{u}{v}\right)^3 + \left(\frac{v}{u}\right)^3 = 2\left(u^2v^2 - \frac{1}{u^2v^2}\right),$$

where $2^{\frac14}q^{\frac1{24}}u = (1+q)(1+q^3)(1+q^5)\ldots$

and $2^{\frac14}q^{\frac5{24}}v = (1+q^5)(1+q^{15})(1+q^{25})\ldots.$

By changing q to $-q$ the above equation may also be written as

$$(9)\quad \left(\frac{v}{u}\right)^3 - \left(\frac{u}{v}\right)^3 = 2\left(u^2v^2 + \frac{1}{u^2v^2}\right),$$

where $2^{\frac14}q^{\frac1{24}}u = (1-q)(1-q^3)(1-q^5)\ldots$

and $2^{\frac14}q^{\frac5{24}}v = (1-q^5)(1-q^{15})(1-q^{25})\ldots.$

If we put $q = e^{-\pi/\sqrt5}$ in (8), so that $u = G_{1/5}$ and $v = G_5$, and hence $u = v$, we see that

$$v^4 - v^{-4} = 1.$$

Hence $\quad v^4 = \dfrac{1+\sqrt5}{2},\quad G_5 = \left(\dfrac{1+\sqrt5}{2}\right)^{\frac14}.$

Similarly, by putting $q = e^{-\pi\sqrt{(2/5)}}$, so that $u = g_{2/5}$ and $v = g_{10}$, and hence $u = 1/v$, we see that

$$v^6 - v^{-6} = 4.$$

Hence $\quad v^2 = \dfrac{1+\sqrt5}{2},\quad g_{10} = \sqrt{\left(\dfrac{1+\sqrt5}{2}\right)}.$

Similarly it can be shown that

$$G_9 = \left(\frac{1+\sqrt3}{\sqrt2}\right)^{\frac13},\quad g_{18} = (\sqrt2+\sqrt3)^{\frac13},$$

$$G_{17} = \sqrt{\left(\frac{5+\sqrt{17}}{8}\right)} + \sqrt{\left(\frac{\sqrt{17}-3}{8}\right)},$$

$$g_{34} = \sqrt{\left(\frac{7+\sqrt{17}}{8}\right)} + \sqrt{\left(\frac{\sqrt{17}-1}{8}\right)},$$

and so on.

3. In order to obtain approximations for π we take logarithms of (1) and (2). Thus

$$(10)\quad \pi = \frac{24}{\sqrt{n}}\log(2^{\frac14}G_n),$$

$$\pi = \frac{24}{\sqrt{n}}\log(2^{\frac14}g_n),$$

approximately, the error being nearly $\dfrac{24}{\sqrt{n}}e^{-\pi\sqrt{n}}$ in both cases. These equations may also be written as

$$(11)\quad e^{\pi\sqrt{n}/24} = 2^{\frac14}G_n,\quad e^{\pi\sqrt{n}/24} = 2^{\frac14}g_n.$$

In those cases in which G_n^{12} and g_n^{12} are simple quadratic surds we may use the forms

$$(G_n^{12}+G_n^{-12})^{\frac1{12}},\quad (g_n^{12}+g_n^{-12})^{\frac1{12}},$$

instead of G_n and g_n; for we have

$$g_n^{12} = \tfrac18 e^{\frac12\pi\sqrt{n}} - \tfrac32 e^{-\frac12\pi\sqrt{n}},$$

approximately, and so

$$g_n^{12} + g_n^{-12} = \tfrac18 e^{\frac12\pi\sqrt{n}} + \tfrac{13}{2}e^{-\frac12\pi\sqrt{n}},$$

approximately, so that

$$(12)\quad \pi = \frac{2}{\sqrt{n}}\log\{8(g_n^{12}+g_n^{-12})\},$$

the error being about $\dfrac{104}{\sqrt{n}}e^{-\pi\sqrt{n}}$, which is of the same order as the error in the formulæ (10). The formula (12) often leads to simpler results. Thus the formula (10) gives

$$e^{\pi\sqrt{18}/24} = 2^{\frac14}g_{18}$$

or

$$e^{\frac14\pi\sqrt{18}} = 10\sqrt2 + 8\sqrt3.$$

But if we use the formula (12), or

$$e^{\pi\sqrt{n}/24} = 2^{\frac14}(g_n^{12}+g_n^{-12})^{\frac1{12}},$$

we get a simpler form, viz.,

$$e^{\frac14\pi\sqrt{18}} = 2\sqrt7.$$

4. The values of g_{2n} and G_n are obtained from the same equation. The approximation by means of g_{2n} is preferable to that by G_n for the following reasons.

(*a*) It is more accurate. Thus the error when we use G_{65} contains a factor $e^{-\pi\sqrt{65}}$, whereas that when we use g_{130} contains a factor $e^{-\pi\sqrt{130}}$.

(*b*) For many values of n, g_{2n} is simpler in form than G_n; thus

$$g_{130} = \sqrt{\left\{(2+\sqrt5)\left(\frac{3+\sqrt{13}}{2}\right)\right\}},$$

while

$$G_{65} = \left\{\left(\frac{1+\sqrt5}{2}\right)\left(\frac{3+\sqrt{13}}{2}\right)\right\}^{1/4} \times \sqrt{\left\{\sqrt{\left(\frac{9+\sqrt{65}}{8}\right)} + \sqrt{\left(\frac{1+\sqrt{65}}{8}\right)}\right\}}.$$

(*c*) For many values of n, g_{2n} involves quadratic surds only, even when G_n is a root of an equation of higher order. Thus G_{23}, G_{29}, G_{31} are roots of cubic equations, G_{47}, G_{79} are those of quintic equations, and G_{71} is that of a septic equation, while g_{46}, g_{58}, g_{62}, g_{94}, g_{142} and g_{158} are all expressible by quadratic surds.

5. Since G_n and g_n can be expressed as roots of algebraical equations with rational coefficients, the same is true of G_n^{24} or g_n^{24}. So let us suppose that

$$1 = ag_n^{-24} - bg_n^{-48} + \ldots,$$

or

$$g_n^{24} = a - bg_n^{-24} + \ldots.$$

But we know that

$$64e^{-\pi\sqrt{n}}g_n^{24} = 1 - 24e^{-\pi\sqrt{n}} + 276e^{-2\pi\sqrt{n}} - \ldots,$$

$$64g_n^{24} = e^{\pi\sqrt{n}} - 24 + 276e^{-\pi\sqrt{n}} - \ldots,$$

$$64a - 64bg_n^{-24} + \ldots = e^{\pi\sqrt{n}} - 24 + 276e^{-\pi\sqrt{n}} - \ldots,$$

$$64a - 4096be^{-\pi\sqrt{n}} + \ldots = e^{\pi\sqrt{n}} - 24 + 276e^{-\pi\sqrt{n}} - \ldots,$$

that is

$$(13)\quad e^{\pi\sqrt{n}} = (64a+24) - (4096b+276)e^{-\pi\sqrt{n}} + \ldots.$$

Similarly, if

$$1 = aG_n^{-24} - bG_n^{-48} + \ldots,$$

then

$$(14)\quad e^{\pi\sqrt{n}} = (64a-24) - (4096b+276)e^{-\pi\sqrt{n}} + \ldots.$$

From (13) and (14) we can find whether $e^{\pi\sqrt{n}}$ is very nearly an integer for given values of n, and ascertain also the number of 9's or 0's in the decimal part. But if G_n and g_n be simple quadratic surds we may work independently as follows. We have for example

$$g_{22} = \sqrt{(1+\sqrt2)}.$$

Hence $\quad 64g_{22}^{24} = e^{\pi\sqrt{22}} - 24 + 276e^{-\pi\sqrt{22}} - \ldots,$

$\quad\quad\quad 64g_{22}^{-24} = 4096e^{-\pi\sqrt{22}} + \ldots,$

so that

$$64(g_{22}^{24}+g_{22}^{-24}) = e^{\pi\sqrt{22}} - 24 + 4372e^{-\pi\sqrt{22}} + \ldots = 64\{(1+\sqrt2)^{12} + (1-\sqrt2)^{12}\}.$$

Hence $\quad e^{\pi\sqrt{22}} = 2508951.9982\ldots.$

Again $\quad G_{37} = (6+\sqrt{37})^{\frac14},$

$$64G_{37}^{24} = e^{\pi\sqrt{37}} + 24 + 276e^{-\pi\sqrt{37}} + \ldots,$$

$$64G_{37}^{-24} = 4096e^{-\pi\sqrt{37}} - \ldots,$$

so that

$$64\left(G_{37}^{24}+G_{37}^{-24}\right)=e^{\pi\sqrt{37}}+24+4372e^{-\pi\sqrt{37}}-\ldots$$
$$=64\{(6+\sqrt{37})^{6}+(6-\sqrt{37})^{6}\}.$$

Hence $\qquad e^{\pi\sqrt{37}}=199148647.999978\ldots.$

Similarly, from $\qquad g_{58}=\sqrt{\left(\frac{5+\sqrt{29}}{2}\right)},$

we obtain

$$64\left(g_{58}^{24}+g_{58}^{-24}\right)=e^{\pi\sqrt{58}}-24+4372e^{-\pi\sqrt{58}}+\ldots$$
$$=64\left\{\left(\frac{5+\sqrt{29}}{2}\right)^{12}+\left(\frac{5-\sqrt{29}}{2}\right)^{12}\right\}.$$

Hence $\qquad e^{\pi\sqrt{58}}=24591257751.99999982\ldots.$

6. I have calculated the values of G_n and g_n for a large number of values of n. Many of these results are equivalent to results given by Weber; for example,

$$G_{13}^{4}=\frac{3+\sqrt{13}}{2},\qquad G_{25}=\frac{1+\sqrt{5}}{2},$$
$$g_{30}^{4}=(2+\sqrt{5})(3+\sqrt{10}).\quad G_{37}^{4}=6+\sqrt{37},$$
$$G_{49}=\frac{7^{1/4}+\sqrt{(4+\sqrt{7})}}{2},\qquad g_{58}^{2}=\frac{5+\sqrt{29}}{2},$$
$$g_{70}^{2}=\frac{(3+\sqrt{5})(1+\sqrt{2})}{2},$$
$$G_{73}=\sqrt{\left(\frac{9+\sqrt{73}}{8}\right)}+\sqrt{\left(\frac{1+\sqrt{73}}{8}\right)},$$
$$G_{85}=\left(\frac{1+\sqrt{5}}{2}\right)\left(\frac{9+\sqrt{85}}{2}\right)^{1/4},$$
$$G_{97}=\sqrt{\left(\frac{13+\sqrt{97}}{8}\right)}+\sqrt{\left(\frac{5+\sqrt{97}}{8}\right)},$$
$$g_{190}^{2}=(2+\sqrt{5})(3+\sqrt{10}),$$
$$G_{385}^{2}=\tfrac{1}{8}(3+\sqrt{11})(\sqrt{5}+\sqrt{7})(\sqrt{7}+\sqrt{11})(3+\sqrt{5}),$$

and so on. I have also many results not given by Weber. I give a complete table of new results. In Weber's notation, $G_n=2^{-\frac{1}{4}}f\{\sqrt{(-n)}\}$ and $g_n=2^{-\frac{1}{4}}f_1\{\sqrt{(-n)}\}$.

TABLE I.

$$g_{62}+\frac{1}{g_{62}}=\tfrac{1}{2}\{\sqrt{(1+\sqrt{2})}+\sqrt{(9+5\sqrt{2})}\},$$

$$G_{65}^{2}=\sqrt{\left\{\left(\frac{1+\sqrt{5}}{2}\right)\left(\frac{3+\sqrt{13}}{2}\right)\right\}}\times\left\{\sqrt{\left(\frac{1+\sqrt{65}}{8}\right)}+\sqrt{\left(\frac{9+\sqrt{65}}{8}\right)}\right\},$$

$$g_{66}^{2}=\sqrt{(\sqrt{2}+\sqrt{3})}(7\sqrt{2}+3\sqrt{11})^{1/6}\times\left\{\sqrt{\left(\frac{7+\sqrt{33}}{8}\right)}+\sqrt{\left(\frac{\sqrt{33}-1}{8}\right)}\right\},$$

$$G_{69}^{2}=(3\sqrt{3}+\sqrt{23})^{1/4}\left(\frac{5+\sqrt{23}}{4}\right)^{1/6}\times\left\{\sqrt{\left(\frac{6+3\sqrt{3}}{4}\right)}+\sqrt{\left(\frac{2+3\sqrt{3}}{4}\right)}\right\},$$

$$G_{77}^{2}=\{\tfrac{1}{2}(\sqrt{7}+\sqrt{11})(8+3\sqrt{7})\}^{1/4}\times\left\{\sqrt{\left(\frac{6+\sqrt{11}}{4}\right)}+\sqrt{\left(\frac{2+\sqrt{11}}{4}\right)}\right\},$$

$$G_{81}^{3}=\frac{(2\sqrt{3}+2)^{1/3}+1}{(2\sqrt{3}-2)^{1/3}-1},$$

$$g_{90}=\{(2+\sqrt{5})(\sqrt{5}+\sqrt{6})\}^{1/6}\times\left\{\sqrt{\left(\frac{3+\sqrt{6}}{4}\right)}+\sqrt{\left(\frac{\sqrt{6}-1}{4}\right)}\right\},$$

$$g_{94}+\frac{1}{g_{94}}=\tfrac{1}{2}\{\sqrt{(7+\sqrt{2})}+\sqrt{(7+5\sqrt{2})}\},$$

$$g_{98}+\frac{1}{g_{98}}=\tfrac{1}{2}\{\sqrt{2}+\sqrt{(14+4\sqrt{14})}\},$$

$$g_{114}^{2}=\sqrt{\{\sqrt{2}+\sqrt{3})}(3\sqrt{2}+\sqrt{19})^{1/6}\times\left\{\sqrt{\left(\frac{23+3\sqrt{57}}{8}\right)}+\sqrt{\left(\frac{15+3\sqrt{57}}{8}\right)}\right\},$$

$$G_{117}=\tfrac{1}{2}\left(\frac{3+\sqrt{13}}{2}\right)^{1/4}(2\sqrt{3}+\sqrt{13})^{1/6}\{3^{1/4}+\sqrt{(4+\sqrt{3})}\},$$

$$\begin{cases}G_{121}+\dfrac{1}{G_{121}}=\left(\dfrac{11}{2}\right)^{1/6}\left\{\left(3+\dfrac{1}{3\sqrt{3}}\right)^{1/3}+\left(3-\dfrac{1}{3\sqrt{3}}\right)^{1/3}\right\}\\ \dfrac{1}{G_{121}}=\dfrac{1}{3\sqrt{2}}\Big[(11-3\sqrt{11})^{1/3}\{(3\sqrt{11}+3\sqrt{3}-4)^{1/3}\\ \qquad\qquad+(3\sqrt{11}-3\sqrt{3}-4)^{1/3}\}-2\Big],\end{cases}$$

$$g_{126}=\sqrt{\left(\frac{\sqrt{3}+\sqrt{7}}{2}\right)}(\sqrt{6}+\sqrt{7})^{1/6}\times\left\{\sqrt{\left(\frac{3+\sqrt{2}}{4}\right)}+\sqrt{\left(\frac{\sqrt{2}-1}{4}\right)}\right\}^{2},$$

$$g_{138}^{2}=\sqrt{\left(\frac{3\sqrt{3}+\sqrt{23}}{2}\right)}(78\sqrt{2}+23\sqrt{23})^{1/6}\times\left\{\sqrt{\left(\frac{5+2\sqrt{6}}{4}\right)}+\sqrt{\left(\frac{1+2\sqrt{6}}{4}\right)}\right\},$$

$$G_{141}^{2}=(4\sqrt{3}+\sqrt{47})^{1/4}\left(\frac{7+\sqrt{47}}{\sqrt{2}}\right)^{1/6}\times\left\{\sqrt{\left(\frac{18+9\sqrt{3}}{4}\right)}+\sqrt{\left(\frac{14+9\sqrt{3}}{4}\right)}\right\},$$

$$G_{145}^{2}=\sqrt{\left\{\frac{(2+\sqrt{5})(5+\sqrt{29})}{2}\right\}}\times\left\{\sqrt{\left(\frac{17+\sqrt{145}}{8}\right)}+\sqrt{\left(\frac{9+\sqrt{145}}{8}\right)}\right\},$$

$$\frac{1}{G_{147}}=2^{-1/12}\left[\frac{1}{2}+\frac{1}{\sqrt{3}}\left\{\sqrt{\left(\frac{7}{4}\right)}-(28)^{1/6}\right\}\right],$$

$$G_{153}=\left\{\sqrt{\left(\frac{5+\sqrt{17}}{8}\right)}+\sqrt{\left(\frac{\sqrt{17}-3}{8}\right)}\right\}^{2}\times\left\{\sqrt{\left(\frac{37+9\sqrt{17}}{4}\right)}+\sqrt{\left(\frac{33+9\sqrt{17}}{4}\right)}\right\}^{1/3},$$

$$g_{154}^{2}=\sqrt{\left\{(2\sqrt{2}+\sqrt{7})\left(\frac{\sqrt{7}+\sqrt{11}}{2}\right)\right\}}\times\left\{\sqrt{\left(\frac{13+2\sqrt{22}}{4}\right)}+\sqrt{\left(\frac{9+2\sqrt{22}}{4}\right)}\right\}.$$

$$g_{158}+\frac{1}{g_{158}}=\tfrac{1}{2}\{\sqrt{(9+\sqrt{2})}+\sqrt{(17+13\sqrt{2})}\},$$

$$\begin{cases}G_{169}+\dfrac{1}{G_{169}}=\left(\dfrac{13}{4}\right)^{1/6}\left\{\left(1+\dfrac{1}{3\sqrt{3}}\right)^{1/3}+\left(1-\dfrac{1}{3\sqrt{3}}\right)^{1/3}\right\}^{2},\\ \dfrac{1}{G_{169}}=\dfrac{1}{3}\left[(\sqrt{13}-2)+\left(\dfrac{13-3\sqrt{13}}{2}\right)^{1/3}\right.\\ \qquad\left.\times\left\{\left(3\sqrt{3}-\dfrac{11-\sqrt{13}}{2}\right)^{1/3}-\left(3\sqrt{3}+\dfrac{11-\sqrt{13}}{2}\right)^{1/3}\right\}\right],\end{cases}$$

$$g_{174}=\sqrt{(1+\sqrt{2})}(4\sqrt{2}+\sqrt{33})^{1/6}\times\left\{\sqrt{\left(\frac{9+\sqrt{33}}{8}\right)}+\sqrt{\left(\frac{1+\sqrt{33}}{8}\right)}\right\},$$

$$G_{205}=\left(\frac{1+\sqrt{5}}{2}\right)\left(\frac{3\sqrt{5}+\sqrt{41}}{2}\right)^{1/4}\times\left\{\sqrt{\left(\frac{7+\sqrt{41}}{8}\right)}+\sqrt{\left(\frac{\sqrt{41}-1}{8}\right)}\right\},$$

$$G_{213}^{2}=(5\sqrt{3}+\sqrt{71})^{1/4}\left(\frac{59+7\sqrt{71}}{4}\right)^{1/6}\times\left\{\sqrt{\left(\frac{21+12\sqrt{3}}{2}\right)}+\sqrt{\left(\frac{19+12\sqrt{3}}{2}\right)}\right\},$$

$$G_{217}^{2}=\left\{\sqrt{\left(\frac{9+4\sqrt{7}}{2}\right)}+\sqrt{\left(\frac{11+4\sqrt{7}}{2}\right)}\right\}\times\left\{\sqrt{\left(\frac{12+5\sqrt{7}}{4}\right)}+\sqrt{\left(\frac{16+5\sqrt{7}}{4}\right)}\right\},$$

$$G_{225}=\left(\frac{1+\sqrt{5}}{4}\right)(2+\sqrt{3})^{1/3}\{\sqrt{(4+\sqrt{15})}+15^{1/4}\},$$

$$g_{238}=\left\{\sqrt{\left(\frac{1+2\sqrt{2}}{4}\right)}+\sqrt{\left(\frac{5+2\sqrt{2}}{4}\right)}\right\}\times\left\{\sqrt{\left(\frac{1+3\sqrt{2}}{4}\right)}+\sqrt{\left(\frac{5+3\sqrt{2}}{4}\right)}\right\},$$

$$G_{265}^{2}=\sqrt{\left\{(2+\sqrt{5})\left(\frac{7+\sqrt{53}}{2}\right)\right\}}\times\left\{\sqrt{\left(\frac{89+5\sqrt{265}}{8}\right)}+\sqrt{\left(\frac{81+5\sqrt{265}}{8}\right)}\right\},$$

$$G_{289}=\left[\sqrt{\left\{\frac{17+\sqrt{17}+17^{1/4}(5+\sqrt{17})}{16}\right\}}+\sqrt{\left\{\frac{1+\sqrt{17}+17^{1/4}(5+\sqrt{17})}{16}\right\}}\right]^{2},$$

$$G_{301}^{2}=\left\{(8+3\sqrt{7})\left(\frac{23\sqrt{43}+57\sqrt{7}}{2}\right)\right\}^{1/4}\times\left\{\sqrt{\left(\frac{46+7\sqrt{43}}{4}\right)}+\sqrt{\left(\frac{42+7\sqrt{43}}{4}\right)}\right\},$$

$$g_{310}=\left(\frac{1+\sqrt{5}}{2}\right)\sqrt{(1+\sqrt{2})}\times\left\{\sqrt{\left(\frac{7+2\sqrt{10}}{4}\right)}+\sqrt{\left(\frac{3+2\sqrt{10}}{4}\right)}\right\},$$

$$\begin{cases}G_{325}=\left(\dfrac{3+\sqrt{13}}{2}\right)^{1/4}t,\ \text{where}\\ t^{3}+t^{2}\left(\dfrac{1-\sqrt{13}}{2}\right)^{2}+t\left(\dfrac{1+\sqrt{13}}{2}\right)^{2}+1\\ \qquad=\sqrt{5}\left\{t^{3}-t^{2}\left(\dfrac{1+\sqrt{13}}{2}\right)+t\left(\dfrac{1-\sqrt{13}}{2}\right)-1\right\},\end{cases}$$

$$G_{333}=\tfrac{1}{2}(6+\sqrt{37})^{1/4}(7\sqrt{3}+2\sqrt{37})^{1/6}\times\{\sqrt{(7+2\sqrt{3})}+\sqrt{(3+2\sqrt{3})}\},$$

$$\begin{cases}G_{363}=2^{\frac{1}{12}}t,\ \text{where}\\ 2t^{3}-t^{2}\{(4+\sqrt{33})+\sqrt{(11+2\sqrt{33})}\}\\ \qquad-t\{1+\sqrt{(11+2\sqrt{33})}\}-1=0,\end{cases}$$

$$G_{441}^{2}=\left(\frac{\sqrt{3}+\sqrt{7}}{2}\right)(2+\sqrt{3})^{1/3}\times\left\{\frac{2+\sqrt{7}+\sqrt{(7+4\sqrt{7})}}{2}\right\}\left\{\frac{\sqrt{(3+\sqrt{7})}+(6\sqrt{7})^{1/4}}{\sqrt{(3+\sqrt{7})}-(6\sqrt{7})^{1/4}}\right\},$$

$$G_{445}=\sqrt{(2+\sqrt{5})}\left(\frac{21+\sqrt{445}}{2}\right)^{1/4}\times\sqrt{\left\{\left(\frac{13+\sqrt{89}}{8}\right)+\sqrt{\left(\frac{5+\sqrt{89}}{8}\right)}\right\}},$$

$$G_{465}^{2}=\sqrt{\left\{(2+\sqrt{3})\left(\frac{1+\sqrt{5}}{2}\right)\left(\frac{3\sqrt{3}+\sqrt{31}}{2}\right)\right\}}(5\sqrt{5}+2\sqrt{31})^{1/6}\times\left\{\sqrt{\left(\frac{2+\sqrt{31}}{4}\right)}+\sqrt{\left(\frac{6+\sqrt{31}}{4}\right)}\right\}\times\left\{\sqrt{\left(\frac{11+2\sqrt{31}}{2}\right)}+\sqrt{\left(\frac{13+2\sqrt{31}}{2}\right)}\right\},$$

$$G_{505}^{2}=(2+\sqrt{5})\sqrt{\left\{\left(\frac{1+\sqrt{5}}{2}\right)(10+\sqrt{101})\right\}}\times\left\{\left(\frac{5\sqrt{5}+\sqrt{101}}{4}\right)+\sqrt{\left(\frac{105+\sqrt{505}}{8}\right)}\right\},$$

$$g_{522}=\sqrt{\left(\frac{5+\sqrt{29}}{2}\right)}(5\sqrt{29}+11\sqrt{6})^{1/6}\times\left\{\sqrt{\left(\frac{9+3\sqrt{6}}{4}\right)}+\sqrt{\left(\frac{5+3\sqrt{6}}{4}\right)}\right\},$$

$$G_{553}^{2}=\left\{\sqrt{\left(\frac{96+11\sqrt{79}}{4}\right)}+\sqrt{\left(\frac{100+11\sqrt{79}}{4}\right)}\right\}\times\left\{\sqrt{\left(\frac{141+16\sqrt{79}}{2}\right)}+\sqrt{\left(\frac{143+16\sqrt{79}}{2}\right)}\right\},$$

$$g_{630}=(\sqrt{14}+\sqrt{15})^{1/6}\sqrt{\left\{(1+\sqrt{2})\left(\frac{3+\sqrt{5}}{2}\right)\left(\frac{\sqrt{3}+\sqrt{7}}{2}\right)\right\}}\times\left\{\sqrt{\left(\frac{\sqrt{15}+\sqrt{7}+2}{4}\right)}+\sqrt{\left(\frac{\sqrt{15}+\sqrt{7}-2}{4}\right)}\right\}\times\left\{\sqrt{\left(\frac{\sqrt{15}+\sqrt{7}+4}{8}\right)}+\sqrt{\left(\frac{\sqrt{15}+\sqrt{7}-4}{8}\right)}\right\},$$

$$G_{765}^{2}=\left(\frac{3+\sqrt{5}}{2}\right)(16+\sqrt{255})^{1/6}\sqrt{\left\{(4+\sqrt{15})\left(\frac{9+\sqrt{85}}{2}\right)\right\}}\times\left\{\sqrt{\left(\frac{6+\sqrt{51}}{4}\right)}+\sqrt{\left(\frac{10+\sqrt{51}}{4}\right)}\right\}\times\left\{\sqrt{\left(\frac{18+3\sqrt{51}}{4}\right)}+\sqrt{\left(\frac{22+3\sqrt{51}}{4}\right)}\right\},$$

$$G_{777}^{2}=\sqrt{\left\{(2+\sqrt{3})(6+\sqrt{37})\left(\frac{\sqrt{3}+\sqrt{7}}{2}\right)\right\}}(246\sqrt{7}+107\sqrt{37})^{1/6}\times\left\{\sqrt{\left(\frac{6+3\sqrt{7}}{4}\right)}+\sqrt{\left(\frac{10+3\sqrt{7}}{4}\right)}\right\}\times\left\{\sqrt{\left(\frac{15+6\sqrt{7}}{2}\right)}+\sqrt{\left(\frac{17+6\sqrt{7}}{2}\right)}\right\},$$

$$G_{1225} = \left(\frac{1+\sqrt{5}}{2}\right)(6+\sqrt{35})^{1/4}\left\{\frac{7^{1/4}+\sqrt{(4+\sqrt{7})}}{2}\right\}^{3/2}$$
$$\times\left[\sqrt{\left\{\frac{43+15\sqrt{7}+(8+3\sqrt{7})\sqrt{(10\sqrt{7})}}{8}\right\}}+\sqrt{\left\{\frac{35+15\sqrt{7}+(8+3\sqrt{7})\sqrt{(10\sqrt{7})}}{8}\right\}}\right],$$

$$G^2_{1353} = \sqrt{\left\{(3+\sqrt{11})(5+3\sqrt{3})\left(\frac{11+\sqrt{123}}{2}\right)\right\}}$$
$$\times\left(\frac{6817+321\sqrt{451}}{4}\right)^{1/6}$$
$$\times\left\{\sqrt{\left(\frac{17+3\sqrt{33}}{8}\right)}+\sqrt{\left(\frac{25+3\sqrt{33}}{8}\right)}\right\}$$
$$\times\left\{\sqrt{\left(\frac{561+99\sqrt{33}}{8}\right)}+\sqrt{\left(\frac{569+99\sqrt{33}}{8}\right)}\right\},$$

$$G^2_{1645} = (2+\sqrt{5})\sqrt{\left\{(3+\sqrt{7})\left(\frac{7+\sqrt{47}}{2}\right)\right\}}\left(\frac{73\sqrt{5}+9\sqrt{329}}{2}\right)^{1/4}$$
$$\times\left\{\sqrt{\left(\frac{119+7\sqrt{329}}{8}\right)}+\sqrt{\left(\frac{127+7\sqrt{329}}{8}\right)}\right\}$$
$$\times\left\{\sqrt{\left(\frac{743+41\sqrt{329}}{8}\right)}+\sqrt{\left(\frac{751+41\sqrt{329}}{8}\right)}\right\}.$$

7. Hence we deduce the following approximate formulæ.

TABLE II.

$$e^{\frac{1}{8}\pi\sqrt{18}} = 2\sqrt{7},\quad e^{\pi\sqrt{22}/12} = 2+\sqrt{2},\quad e^{\frac{1}{4}\pi\sqrt{30}} = 20\sqrt{3}+16\sqrt{6},$$
$$e^{\frac{1}{4}\pi\sqrt{34}} = 12(4+\sqrt{17}),\quad e^{\frac{1}{2}\pi\sqrt{46}} = 144(147+104\sqrt{2}),$$
$$e^{\frac{1}{4}\pi\sqrt{42}} = 84+32\sqrt{6},\quad e^{\pi\sqrt{58}/12} = \frac{5+\sqrt{29}}{\sqrt{2}},$$
$$e^{\frac{1}{4}\pi\sqrt{70}} = 60\sqrt{35}+96\sqrt{14},\quad e^{\frac{1}{4}\pi\sqrt{78}} = 300\sqrt{3}+208\sqrt{6},$$
$$e^{\pi\sqrt{35}/24} = \frac{1+\sqrt{(3+2\sqrt{5})}}{\sqrt{2}},\quad e^{\frac{1}{4}\pi\sqrt{102}} = 800\sqrt{3}+196\sqrt{51},$$
$$e^{\frac{1}{4}\pi\sqrt{130}} = 12(323+40\sqrt{65}),\quad e^{\pi\sqrt{190}/12} = (2\sqrt{2}+\sqrt{10})(3+\sqrt{10}),$$
$$\pi = \frac{12}{\sqrt{130}}\log\left\{\frac{(2+\sqrt{5})(3+\sqrt{13})}{\sqrt{2}}\right\},$$

where $A(k)$ denotes an algebraic function of the special class described above. I shall use the letter A generally to denote a function of this type.

Now, if we put $k = l'$ and $k' = l$ in (16), we have

$$(17)\quad n\left\{1-24\left(\frac{1}{e^{2\pi\sqrt{n}}-1}+\frac{2}{e^{4\pi\sqrt{n}}-1}+\dots\right)\right\}$$
$$-\left\{1-24\left(\frac{1}{e^{2\pi/\sqrt{n}}-1}+\frac{2}{e^{4\pi/\sqrt{n}}-1}+\dots\right)\right\} = \left(\frac{K}{\pi}\right)^2 A(k).$$

The algebraic function $A(k)$ of course assumes a purely numerical form when we substitute the value of k deduced from the modular equation. But by substituting $k = l'$ and $k' = l$ in (15) we have

$$n^{\frac{1}{4}}e^{-\pi\sqrt{n}/12}(1-e^{-2\pi\sqrt{n}})(1-e^{-4\pi\sqrt{n}})(1-e^{-6\pi\sqrt{n}})\dots$$
$$= e^{-\pi/(12\sqrt{n})}(1-e^{-2\pi/\sqrt{n}})(1-e^{-4\pi/\sqrt{n}})(1-e^{-6\pi/\sqrt{n}})\dots.$$

Differentiating the above equation logarithmically we have

$$(18)\quad n\left\{1-24\left(\frac{1}{e^{2\pi\sqrt{n}}-1}+\frac{2}{e^{4\pi\sqrt{n}}-1}+\dots\right)\right\}$$
$$+\left\{1-24\left(\frac{1}{e^{2\pi/\sqrt{n}}-1}+\frac{2}{e^{4\pi/\sqrt{n}}-1}+\dots\right)\right\} = \frac{6\sqrt{n}}{\pi}.$$

Now, adding (17) and (18), we have

$$(19)\quad 1-\frac{3}{\pi\sqrt{n}}-24\left(\frac{1}{e^{2\pi\sqrt{n}}-1}+\frac{2}{e^{4\pi\sqrt{n}}-1}+\dots\right) = \left(\frac{K}{\pi}\right)^2 A(k).$$

But it is known that

$$1-24\left(\frac{q}{1+q}+\frac{3q^3}{1+q^3}+\frac{5q^5}{1+q^5}+\dots\right) = \left(\frac{2K}{\pi}\right)^2\left(1-2k^2\right),$$

so that

$$(20)\quad 1-24\left(\frac{1}{e^{\pi\sqrt{n}}+1}+\frac{3}{e^{3\pi\sqrt{n}}+1}+\dots\right) = \left(\frac{K}{\pi}\right)^2 A(k).$$

Hence, dividing (19) by (20), we have

$$(21)\quad \frac{1-\frac{3}{\pi\sqrt{n}}-24\left(\frac{1}{e^{2\pi\sqrt{n}}-1}+\frac{2}{e^{4\pi\sqrt{n}}-1}+\dots\right)}{1-24\left(\frac{1}{e^{\pi\sqrt{n}}+1}+\frac{3}{e^{3\pi\sqrt{n}}+1}+\dots\right)} = R,$$

Similarly we can find the sum of (19) when $n = 21, 33, 57, 93$, etc., by combining the values of $f(3)$ and $f(7)$, $f(3)$ and $f(11)$, and so on, obtained from Table III.

11. The errors in (22) and (24) being about

$$8\pi e^{-\pi\sqrt{n}}(\pi\sqrt{n}-3),\quad 24\pi(10\pi\sqrt{n}-31)e^{-2\pi\sqrt{n}},$$

we cannot expect a high degree of approximation for small values of n. Thus, if we put $n = 7, 9, 16$, and 25 in (24), we get

$$\frac{19}{16}\sqrt{7} = 3.14180\dots,$$
$$\frac{7}{3}\left(1+\frac{\sqrt{3}}{5}\right) = 3.14162\dots,$$
$$\frac{99}{80}\left(\frac{7}{7-3\sqrt{2}}\right) = 3.14159274\dots,$$
$$\frac{63}{25}\left(\frac{17+15\sqrt{5}}{7+15\sqrt{5}}\right) = 3.14159265380\dots,$$

while $\pi = 3.14159265358\dots.$

But if we put $n = 25$ in (22), we get only

$$\frac{9}{5}+\sqrt{\frac{9}{5}} = 3.14164\dots.$$

12. Another curious approximation to π is

$$\left\{9^2+\frac{19^2}{22}\right\}^{\frac{1}{4}} = 3.14159265262\dots.$$

This value was obtained empirically, and it has no connection with the preceding theory.

The actual value of π, which I have used for purposes of calculation, is

$$\frac{355}{113}\left(1-\frac{.0003}{3533}\right) = 3.1415926535897943\dots,$$

which is greater than π by about 10^{-15}. This is obtained by simply taking the reciprocal of $1-(113\pi/355)$.

In this connection it may be interesting to note the following simple geometrical constructions for π. The first merely gives the ordinary value $355/113$. The second gives the value $(9^2+19^2/22)^{1/4}$ mentioned above.

$$f(2) = \frac{4KL}{\pi^2}(k'+l),$$
$$f(3) = \frac{4KL}{\pi^2}(1+kl+k'l'),$$
$$f(4) = \frac{4KL}{\pi^2}(\sqrt{k'}+\sqrt{l})^2,$$
$$f(5) = \frac{4KL}{\pi^2}(3+kl+k'l')\sqrt{\left(\frac{1+kl+k'l'}{2}\right)},$$
$$f(7) = \frac{12KL}{\pi^2}(1+kl+k'l'),$$
$$f(11) = \frac{8KL}{\pi^2}\{2(1+kl+k'l')+\sqrt{(kl)}+\sqrt{(k'l')}-\sqrt{(kk'll')}\},$$
$$f(15) = \frac{4KL}{\pi^2}[\{1+(kl)^{\frac{1}{2}}+(k'l')^{\frac{1}{2}}\}^2-\{1+kl+k'l'\}],$$
$$f(17) = \frac{4KL}{\pi^2}\sqrt{\{44(1+k^2l^2+k'^2l'^2)+168(kl+k'l'-kk'll')}$$
$$\overline{-102(1-kl-k'l')(4kk'll')^{\frac{1}{3}}-192(4kk'll')^{\frac{2}{3}}\}},$$
$$f(19) = \frac{24KL}{\pi^2}\{(1+kl+k'l')+\sqrt{(kl)}+\sqrt{(k'l')}-\sqrt{(kk'll')}\},$$
$$f(23) = \frac{4KL}{\pi^2}[11(1+kl+k'l')-16(4kk'll')^{\frac{1}{6}}$$
$$\times\{1+\sqrt{(kl)}+\sqrt{(k'l')}\}-20(4kk'll')^{\frac{1}{3}}],$$
$$f(31) = \frac{12KL}{\pi^2}[3(1+kl+k'l')+4\{\sqrt{(kl)}+\sqrt{(k'l')}+\sqrt{(kk'll')}\}$$
$$-4(kk'll')^{\frac{1}{4}}\{1+(kl)^{\frac{1}{4}}+(k'l')^{\frac{1}{4}}\}],$$
$$f(35) = \frac{4KL}{\pi^2}[2\{\sqrt{(kl)}+\sqrt{(k'l')}-\sqrt{(kk'll')}\}$$
$$+(4kk'll')^{-\frac{1}{6}}\{1-\sqrt{(kl)}-\sqrt{(k'l')}\}^2].$$

Thus the sum of the series (19) can be found in finite terms, when $n = 2, 3, 4, 5$, etc., from the equations in Table III. We can use the same table to find the sum of (19) when $n = 9, 25, 49$, etc.; but then we have also to use the equation

$$\frac{3}{\pi} = 1-24\left(\frac{1}{e^{2\pi}-1}+\frac{2}{e^{4\pi}-1}+\frac{3}{e^{6\pi}-1}+\dots\right),$$

which is got by putting $k = k' = 1/\sqrt{2}$ and $n = 1$ in (18).

where R can always be expressed in radicals if n is any rational number. Hence we have

$$(22)\qquad \pi = \frac{3}{(1-R)\sqrt{n}}$$

nearly, the error being about $8\pi e^{-\pi\sqrt{n}}(\pi\sqrt{n}-3)$.

9. We may get a still closer approximation from the following results.

It is known that

$$1+240\sum_{r=1}^{r=\infty}\frac{r^3q^{2r}}{1-q^{2r}} = \left(\frac{2K}{\pi}\right)^4(1-k^2k'^2),$$

and also that

$$1-504\sum_{r=1}^{r=\infty}\frac{r^5q^{2r}}{1-q^{2r}} = \left(\frac{2K}{\pi}\right)^6(1-2k^2)(1+\tfrac{1}{2}k^2k'^2).$$

Hence, from (19), we see that

$$(23)\quad \left\{1-\frac{3}{\pi\sqrt{n}}-24\sum_{r=1}^{r=\infty}\frac{r}{e^{2\pi r\sqrt{n}}-1}\right\}\left\{1+240\sum_{r=1}^{r=\infty}\frac{r^3}{e^{2\pi r\sqrt{n}}-1}\right\}$$
$$= R'\left\{1-504\sum_{r=1}^{r=\infty}\frac{r^5}{e^{2\pi r\sqrt{n}}-1}\right\},$$

where R' can always be expressed in radicals for any rational value of n. Hence

$$(24)\qquad \pi = \frac{3}{(1-R')\sqrt{n}},$$

nearly, the error being about $24\pi(10\pi\sqrt{n}-31)e^{-2\pi\sqrt{n}}$.

It will be seen that the error in (24) is much less than that in (22), if n is at all large.

10. In order to find R and R' the series in (16) must be calculated in finite terms. I shall give the final results for a few values of n.

TABLE III.

$$q = e^{-\pi K'/K},\quad q^n = e^{-\pi L'/L},$$
$$f(q) = n\left(1-24\sum_1^\infty\frac{q^{2mn}}{1-q^{2mn}}\right)-\left(1-24\sum_1^\infty\frac{q^{2m}}{1-q^{2m}}\right),$$

$$\pi = \frac{24}{\sqrt{142}}\log\left\{\sqrt{\left(\frac{10+11\sqrt{2}}{4}\right)}+\sqrt{\left(\frac{10+7\sqrt{2}}{4}\right)}\right\},$$
$$\pi = \frac{12}{\sqrt{190}}\log\{(2\sqrt{2}+\sqrt{10})(3+\sqrt{10})\}.$$
$$\pi = \frac{12}{\sqrt{310}}\log[\tfrac{1}{4}(3+\sqrt{5})(2+\sqrt{2})\{(5+2\sqrt{10})+\sqrt{(61+20\sqrt{10})}\}].$$
$$\pi = \frac{4}{\sqrt{522}}\log\left[\left(\frac{5+\sqrt{29}}{\sqrt{2}}\right)^3(5\sqrt{29}+11\sqrt{6})\right.$$
$$\left.\times\left\{\sqrt{\left(\frac{9+3\sqrt{6}}{4}\right)}+\sqrt{\left(\frac{5+3\sqrt{6}}{4}\right)}\right\}^6\right].$$

The last five formulæ are correct to 15, 16, 18, 22, and 31 places of decimals respectively.

8. Thus we have seen how to approximate to π by means of logarithms of surds. I shall now show how to obtain approximations in terms of surds only. If

$$n\frac{K'}{K} = \frac{L'}{L},$$

we have

$$\frac{n\,dk}{kk'^2K^2} = \frac{dl}{ll'^2L^2}.$$

But, by means of the modular equation connecting k and l, we can express dk/dl as an algebraic function of k, a function moreover in which all coefficients which occur are algebraic numbers. Again,

$$q = e^{-\pi K'/K},\quad q^n = e^{-\pi L'/L},$$

and

$$(15)\quad \frac{q^{\frac{1}{12}}(1-q^2)(1-q^4)(1-q^6)\dots}{q^{\frac{1}{12}n}(1-q^{2n})(1-q^{4n})(1-q^{6n})\dots} = \left(\frac{kk'}{ll'}\right)^{\frac{1}{6}}\sqrt{\left(\frac{K}{L}\right)}.$$

Differentiating this equation logarithmically, and using the formula

$$\frac{dq}{dk} = \frac{\pi^2q}{2kk'^2K^2},$$

we see that

$$(16)\quad n\left\{1-24\left(\frac{q^{2n}}{1-q^{2n}}+\frac{2q^{4n}}{1-q^{4n}}+\dots\right)\right\}$$
$$-\left\{1-24\left(\frac{q^2}{1-q^2}+\frac{2q^4}{1-q^4}+\dots\right)\right\} = \frac{KL}{\pi^2}A(k),$$

(1) Let AB (fig. 1) be a diameter of a circle whose centre is O.

Bisect AO at M and trisect OB at T.

Draw TP perpendicular to AB and meeting the circumference at P.

Draw a chord BQ equal to PT and join AQ.

Draw OS and TR parallel to BQ and meeting AQ at S and R respectively.

Draw a chord AD equal to AS and a tangent $AC=RS$.

Join BC, BD, and CD and draw EX, parallel to CD, meeting BC at X.

Then the square on BX is very nearly equal to the area of the circle, the error being less than a tenth of an inch when the diameter is 40 miles long.

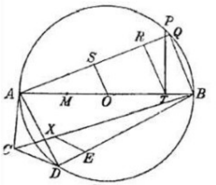

Fig. 1.

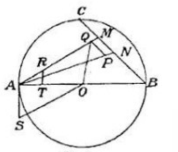

Fig. 2.

(2) Let AB (fig 2) be a diameter of a circle whose centre is O.

Bisect the arc ACB at C and trisect AO at T.

Join BC and cut off from it CM and MN equal to AT.

Join AM and AN and cut off from the latter AP equal to AM.

Through P draw PQ parallel to MN and meeting AM at Q.

Join OQ and through T draw TR, parallel to OQ, and meeting AQ at R.

Draw AS perpendicular to AO and equal to AR, and join OS.

Then the mean proportional between OS and OB will be very nearly equal to a sixth of the circumference, the error being less than a twelfth of a inch when the diameter is 8000 miles long.

13. I shall conclude this paper by giving a few series for $1/\pi$.

It is known that, when $k \leq 1/\sqrt{2}$,

$$(25)\qquad \left(\frac{2K}{\pi}\right)^2 = 1+\left(\frac{1}{2}\right)^2(2kk')^2+\left(\frac{1.3}{2.4}\right)^2(2kk')^4+\dots.$$

Hence we have

$$(26)\qquad q^{\frac{1}{3}}(1-q^2)^4(1-q^4)^4(1-q^6)^4\dots$$
$$=\left(\frac{1}{4}kk'\right)^{\frac{2}{3}}\left\{1+\left(\frac{1}{2}\right)^2(2kk')^2+\left(\frac{1.3}{2.4}\right)^2(2kk')^4+\dots\right\}.$$

Differentiating both sides in (26) logarithmically with respect to k, we can easily show that

$$(27)\qquad 1-24\left(\frac{q^2}{1-q^2}+\frac{2q^4}{1-q^4}+\frac{3q^6}{1-q^6}+\dots\right)$$
$$=(1-2k^2)\left\{1+4\left(\frac{1}{2}\right)^2(2kk')^2+7\left(\frac{1.3}{2.4}\right)^2(2kk')^4+\dots\right\}.$$

But it follows from (19) that, when $q=e^{-\pi\sqrt{n}}$, n being a rational number, the left-hand side of (27) can be expressed in the form

$$A\left(\frac{2K}{\pi}\right)^2+\frac{B}{\pi},$$

where A and B are algebraic numbers expressible by surds. Combining (25) and (27) in such a way as to eliminate the term $(2K/\pi)^2$, we are left with a series for $1/\pi$. Thus, for example,

$$(28)\qquad \frac{4}{\pi}=1+\frac{7}{4}\left(\frac{1}{2}\right)^3+\frac{13}{4^2}\left(\frac{1.3}{2.4}\right)^3+\frac{19}{4^3}\left(\frac{1.3.5}{2.4.6}\right)^3+\dots,$$
$$(q=e^{-\pi\sqrt{3}},\ 2kk'=\tfrac{1}{2}),$$

$$(29)\qquad \frac{16}{\pi}=5+\frac{47}{64}\left(\frac{1}{2}\right)^3+\frac{89}{64^2}\left(\frac{1.3}{2.4}\right)^3+\frac{131}{64^3}\left(\frac{1.3.5}{2.4.6}\right)^3+\dots,$$
$$(q=e^{-\pi\sqrt{7}},\ 2kk'=\tfrac{1}{8}),$$

$$(30)\qquad \frac{32}{\pi}=(5\sqrt{5}-1)+\frac{47\sqrt{5}+29}{64}\left(\frac{1}{2}\right)^3\left(\frac{\sqrt{5}-1}{2}\right)^8+\dots,$$
$$\frac{89\sqrt{5}+59}{64^2}\left(\frac{1.3}{2.4}\right)^3\left(\frac{\sqrt{5}-1}{2}\right)^{16}+\dots,$$
$$\left[q=e^{-\pi\sqrt{15}},\ 2kk'=\frac{1}{8}\left(\frac{\sqrt{5}-1}{2}\right)^4\right];$$

here $(5\sqrt{5}-1)$, $(47\sqrt{5}+29)$, $(89\sqrt{5}+59)$, etc. are in arithmetical progression.

14. The ordinary modular equations express the relations which hold between k and l when $nK'/K=L'/L$, or $q^n=Q$, where

$$q=e^{-\pi K'/K},\quad Q=e^{-\pi L'/L},$$
$$K=1+\left(\frac{1}{2}\right)^2k^2+\left(\frac{1.3}{2.4}\right)^2k^4+\dots.$$

There are corresponding theories in which q is replaced by one or other of the functions

$$q_1=e^{-\pi K_1'\sqrt{2}/K_1},\quad q_2=e^{-2\pi K_2'/(K_2\sqrt{3})}\quad q_3=e^{-2\pi K_3'/K_3},$$

where

$$K_1=1+\frac{1.3}{4^2}k^2+\frac{1.3.5.7}{4^2.8^2}k^4+\frac{1.3.5.7.9.11}{4^2.8^2.12^2}k^6+\dots,$$
$$K_2=1+\frac{1.2}{3^2}k^2+\frac{1.2.4.5}{3^2.6^2}k^4+\frac{1.2.4.5.7.8}{3^2.6^2.9^2}k^6+\dots.$$
$$K_3=1+\frac{1.5}{6^2}k^2+\frac{1.5.7.11}{6^2.12^2}k^4+\frac{1.5.7.11.13.17}{6^2.12^2.18^2}k^6+\dots.$$

From these theories we can deduce further series for $1/\pi$, such as

$$(31)\qquad \frac{27}{4\pi}=2+17.\frac{1}{2}.\frac{1}{3}.\frac{2}{3}.\left(\frac{2}{27}\right)$$
$$+32.\frac{1.3}{2.4}.\frac{1.4}{3.6}.\frac{2.5}{3.6}.\left(\frac{2}{27}\right)^2+\dots,$$

$$(32)\qquad \frac{15\sqrt{3}}{2\pi}=4+37.\frac{1}{2}.\frac{1}{3}.\frac{2}{3}.\left(\frac{4}{125}\right)$$
$$+70.\frac{1.3}{2.4}.\frac{1.4}{3.6}.\frac{2.5}{3.6}.\left(\frac{4}{125}\right)^2+\dots,$$

$$(33)\qquad \frac{5\sqrt{5}}{2\pi\sqrt{3}}=1+12.\frac{1}{2}.\frac{1}{6}.\frac{5}{6}.\left(\frac{4}{125}\right)$$
$$+23.\frac{1.3}{2.4}.\frac{1.7}{6.12}.\frac{5.11}{6.12}.\left(\frac{4}{125}\right)^2+\dots,$$

$$(34)\qquad \frac{85\sqrt{85}}{18\pi\sqrt{3}}=8+141.\frac{1}{2}.\frac{1}{6}.\frac{5}{6}.\left(\frac{4}{85}\right)^3$$
$$+274.\frac{1.3}{2.4}.\frac{1.7}{6.12}.\frac{5.11}{6.12}.\left(\frac{4}{85}\right)^6+\dots,$$

$$(35)\qquad \frac{4}{\pi}=\frac{3}{2}-\frac{23}{2^3}.\frac{1}{2}.\frac{1.3}{4^2}+\frac{43}{2^6}.\frac{1.3}{2.4}.\frac{1.3.5.7}{4^2.8^2}-\dots,$$

$$(36)\qquad \frac{4}{\pi\sqrt{3}}=\frac{3}{4}-\frac{31}{3.4^3}.\frac{1}{2}.\frac{1.3}{4^2}$$
$$+\frac{59}{3^2.4^5}.\frac{1.3}{2.4}.\frac{1.3.5.7}{4^2.8^2}-\dots,$$

$$(37)\qquad \frac{4}{\pi}=\frac{23}{18}-\frac{283}{18^3}.\frac{1}{2}.\frac{1.3}{4^2}$$
$$+\frac{543}{18^5}.\frac{1.3}{2.4}.\frac{1.3.5.7}{4^2.8^2}-\dots,$$

$$(38)\qquad \frac{4}{\pi\sqrt{5}}=\frac{41}{72}-\frac{685}{5.72^3}.\frac{1}{2}.\frac{1.3}{4^2}$$
$$+\frac{1329}{5^2.72^5}.\frac{1.3}{2.4}.\frac{1.3.5.7}{4^2.8^2}-\dots,$$

$$(39)\qquad \frac{4}{\pi}=\frac{1123}{882}-\frac{22583}{882^3}.\frac{1}{2}.\frac{1.3}{4^2}$$
$$+\frac{44043}{882^5}.\frac{1.3}{2.4}.\frac{1.3.5.7}{4^2.8^2}-\dots,$$

$$(40)\qquad \frac{2\sqrt{3}}{\pi}=1+\frac{9}{9}.\frac{1}{2}.\frac{1.3}{4^2}+\frac{17}{9^2}.\frac{1.3}{2.4}.\frac{1.3.5.7}{4^2.8^2}+\dots,$$

$$(41)\qquad \frac{1}{2\pi\sqrt{2}}=\frac{1}{9}+\frac{11}{9^3}.\frac{1}{2}.\frac{1.3}{4^2}+\frac{21}{9^5}.\frac{1.3}{2.4}.\frac{1.3.5.7}{4^2.8^2}+\dots,$$

$$(42)\qquad \frac{1}{3\pi\sqrt{3}}=\frac{3}{49}+\frac{43}{49^3}.\frac{1}{2}.\frac{1.3}{4^2}$$
$$+\frac{83}{49^5}.\frac{1.3}{2.4}.\frac{1.3.5.7}{4^2.8^2}+\dots,$$

$$(43)\qquad \frac{2}{\pi\sqrt{11}}=\frac{19}{99}+\frac{299}{99^3}.\frac{1}{2}.\frac{1.3}{4^2}$$
$$+\frac{579}{99^5}.\frac{1.3}{2.4}.\frac{1.3.5.7}{4^2.8^2}+\dots,$$

$$(44)\qquad \frac{1}{2\pi\sqrt{2}}=\frac{1103}{99^2}+\frac{27493}{99^6}.\frac{1}{2}.\frac{1.3}{4^2}$$
$$+\frac{53883}{99^{10}}.\frac{1.3}{2.4}.\frac{1.3.5.7}{4^2.8^2}+\dots.$$

In all these series the first factors in each term form an arithmetical progression; *e.g.*, 2, 17, 32, 47, ..., in (31), and 4, 37, 70, 103, ..., in (32). The first two series belong to the theory of q_2, the next two to that of q_3, and the rest to that of q_1.

The last series (44) is extremely convergent. Thus, taking only the first term, we see that

$$\frac{1103}{99^2}=.11253953678\dots,$$
$$\frac{1}{2\pi\sqrt{2}}=.11253953951\dots.$$

15. In concluding this paper I have to remark that the series

$$1-24\left(\frac{q^2}{1-q^2}+\frac{2q^4}{1-q^4}+\frac{3q^6}{1-q^6}+\dots\right),$$

which has been discussed in §§ 8–13, is very closely connected with the perimeter of an ellipse whose eccentricity is k. For, if a and b be the semi-major and the semi-minor axes, it is known that

$$(45)\qquad p=2\pi a\left\{1-\frac{1}{2^2}k^2-\frac{1^2.3}{2^2.4^2}k^4-\frac{1^2.3^2.5}{2^2.4^2.6^2}k^6-\dots\right\},$$

where p is the perimeter and k the eccentricity. It can easily be seen from (45) that

$$(46)\qquad p=4ak'^2\left\{K+k\frac{dK}{dk}\right\}.$$

But, taking the equation

$$q^{\frac{1}{12}}(1-q^2)(1-q^4)(1-q^6)\dots=(2kk')^{\frac{1}{6}}\sqrt{(K/\pi)},$$

and differentiating both sides logarithmically with respect to k, and combining the result with (46) in such a way as to eliminate dK/dk, we can show that

$$(47)\qquad p=\frac{4a}{3K}\left[K^2(1+k'^2)+(\tfrac{1}{2}\pi)^2\right.$$
$$\left.\times\left\{1-24\left(\frac{q^2}{1-q^2}+\frac{2q^4}{1-q^4}+\dots\right)\right\}\right].$$

But we have shown already that the right-hand side of (47) can be expressed in terms of K if $q=e^{-\pi\sqrt{n}}$, where n is any rational number. It can also be shown that K can be expressed in terms of Γ-functions if q be of the forms $e^{-\pi n}$, $e^{-\pi n\sqrt{2}}$ and $e^{-\pi n\sqrt{3}}$, where n is rational. Thus, for example, we have

$$(48)\qquad \begin{cases} k=\sin\dfrac{\pi}{4},\quad q=e^{-\pi}, \\ \qquad p=a\sqrt{\left(\dfrac{\pi}{2}\right)}\left\{\dfrac{\Gamma(\frac{1}{4})}{\Gamma(\frac{3}{4})}+\dfrac{\Gamma(\frac{3}{4})}{\Gamma(\frac{5}{4})}\right\}, \\ k=\tan\dfrac{\pi}{8},\quad q=e^{-\pi\sqrt{2}}, \\ \qquad p=a\sqrt{\left(\dfrac{\pi}{4}\right)}\left\{\dfrac{\Gamma(\frac{1}{8})}{\Gamma(\frac{5}{8})}+\dfrac{\Gamma(\frac{5}{8})}{\Gamma(\frac{9}{8})}\right\}, \\ k=\sin\dfrac{\pi}{12},\quad q=e^{-\pi\sqrt{3}}, \\ \qquad p=a\sqrt{\left(\dfrac{\pi}{\sqrt{3}}\right)}\left\{\left(1+\dfrac{1}{\sqrt{3}}\right)\dfrac{\Gamma(\frac{1}{3})}{\Gamma(\frac{5}{6})}+2\dfrac{\Gamma(\frac{5}{6})}{\Gamma(\frac{1}{3})}\right\}, \\ \dfrac{b}{a}=\tan^2\dfrac{\pi}{8},\quad q=e^{-2\pi}, \\ \qquad p=(a+b)\sqrt{\left(\dfrac{\pi}{2}\right)}\left\{\dfrac{1}{2}\dfrac{\Gamma(\frac{1}{4})}{\Gamma(\frac{3}{4})}+\dfrac{\Gamma(\frac{3}{4})}{\Gamma(\frac{5}{4})}\right\}, \end{cases}$$

and so on.

16. The following approximations for p were obtained empirically :—

$$(49)\qquad p=\pi[3(a+b)-\sqrt{\{(a+3b)(3a+b)\}}+\epsilon],$$

where ϵ is about $ak^{12}/1048576$;

$$(50)\qquad p=\pi\left\{(a+b)+\frac{3(a-b)^2}{10(a+b)+\sqrt{(a^2+14ab+b^2)}}+\epsilon\right\},$$

where ϵ is about $3ak^{20}/68719476736$.

Squaring the Circle

https://www.imsc.res.in/~rao/ramanujan/collectedpapers/squaring/squaring.htm

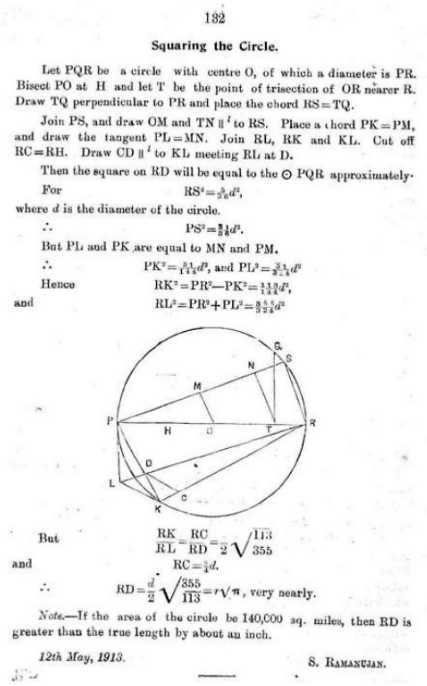

Squaring the Circle.

Let PQR be a circle with centre O, of which a diameter is PR. Bisect PO at H and let T be the point of trisection of OR nearer R. Draw TQ perpendicular to PR and place the chord RS = TQ.

Join PS, and draw OM and TN $\|^l$ to RS. Place a chord PK = PM, and draw the tangent PL = MN. Join RL, RK and KL. Cut off RC = RH. Draw CD $\|^l$ to KL meeting RL at D.

Then the square on RD will be equal to the ⊙ PQR approximately.

For $$RS^2 = \tfrac{5}{36}d^2,$$

where d is the diameter of the circle.

$$\therefore \quad PS^2 = \tfrac{31}{36}d^2.$$

But PL and PK are equal to MN and PM.

$$\therefore \quad PK^2 = \tfrac{31}{144}d^2, \text{ and } PL^2 = \tfrac{31}{324}d^2$$

Hence $$RK^2 = PR^2 - PK^2 = \tfrac{113}{144}d^2,$$

and $$RL^2 = PR^2 + PL^2 = \tfrac{355}{324}d^2$$

But $$\frac{RK}{RL} = \frac{RC}{RD} = \frac{3}{2}\sqrt{\frac{113}{355}}$$

and $$RC = \tfrac{3}{4}d.$$

$$\therefore \quad RD = \frac{d}{2}\sqrt{\frac{355}{113}} = r\sqrt{\pi}, \text{ very nearly.}$$

Note.—If the area of the circle be 140,000 sq. miles, then RD is greater than the true length by about an inch.

12th May, 1913.

S. RAMANUJAN.

Published
by
Journal of The Cambridge University Press

imsc.res.in/~rao/ramanujan

https://www.imsc.res.in/~rao/ramanujan/CamUnivCpapers/collectedright1.htm

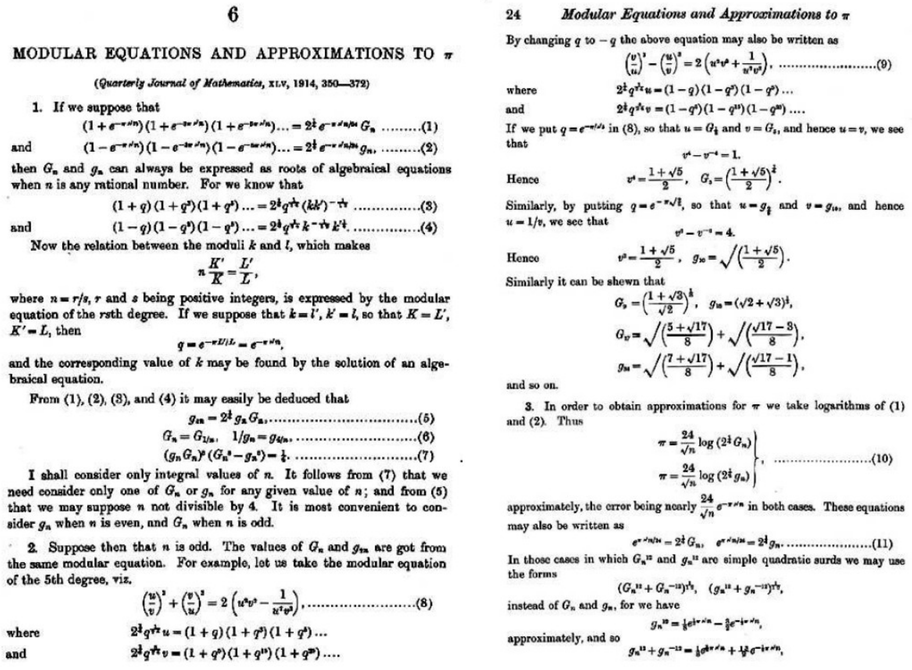

MODULAR EQUATIONS AND APPROXIMATIONS TO π

(*Quarterly Journal of Mathematics*, XLV, 1914, 350—372)

1. If we suppose that

$$(1+e^{-\pi\sqrt{n}})(1+e^{-3\pi\sqrt{n}})(1+e^{-5\pi\sqrt{n}})\ldots=2^{\frac14}e^{-\pi\sqrt{n}/24}G_n \quad\ldots\ldots\ldots(1)$$

and

$$(1-e^{-\pi\sqrt{n}})(1-e^{-3\pi\sqrt{n}})(1-e^{-5\pi\sqrt{n}})\ldots=2^{\frac14}e^{-\pi\sqrt{n}/24}g_n, \quad\ldots\ldots\ldots(2)$$

then G_n and g_n can always be expressed as roots of algebraical equations when n is any rational number. For we know that

$$(1+q)(1+q^3)(1+q^5)\ldots=2^{\frac16}q^{\frac1{24}}(kk')^{-\frac1{12}} \quad\ldots\ldots\ldots\ldots\ldots(3)$$

and

$$(1-q)(1-q^3)(1-q^5)\ldots=2^{\frac16}q^{\frac1{24}}k^{-\frac1{12}}k'^{\frac16}. \quad\ldots\ldots\ldots\ldots\ldots(4)$$

Now the relation between the moduli k and l, which makes

$$n\frac{K'}{K}=\frac{L'}{L},$$

where $n=r/s$, r and s being positive integers, is expressed by the modular equation of the rsth degree. If we suppose that $k=l'$, $k'=l$, so that $K=L'$, $K'=L$, then

$$q=e^{-\pi L'/L}=e^{-\pi\sqrt{n}},$$

and the corresponding value of k may be found by the solution of an algebraical equation.

From (1), (2), (3), and (4) it may easily be deduced that

$$g_{4n}=2^{\frac14}g_nG_n, \quad\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

$$G_n=G_{1/n},\quad 1/g_n=g_{4/n}, \quad\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(6)$$

$$(g_nG_n)^8(G_n^8-g_n^8)=\tfrac14. \quad\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(7)$$

I shall consider only integral values of n. It follows from (7) that we need consider only one of G_n or g_n for any given value of n; and from (5) that we may suppose n not divisible by 4. It is most convenient to consider g_n when n is even, and G_n when n is odd.

2. Suppose then that n is odd. The values of G_n and g_{4n} are got from the same modular equation. For example, let us take the modular equation of the 5th degree, viz.

$$\left(\frac{u}{v}\right)^3+\left(\frac{v}{u}\right)^3=2\left(u^2v^2-\frac{1}{u^2v^2}\right), \quad\ldots\ldots\ldots\ldots\ldots\ldots\ldots(8)$$

where

$$2^{\frac14}q^{\frac1{24}}u=(1+q)(1+q^3)(1+q^5)\ldots$$

and

$$2^{\frac14}q^{\frac5{24}}v=(1+q^5)(1+q^{15})(1+q^{25})\ldots.$$

By changing q to $-q$ the above equation may also be written as

$$\left(\frac{v}{u}\right)^3-\left(\frac{u}{v}\right)^3=2\left(u^2v^2+\frac{1}{u^2v^2}\right), \quad\ldots\ldots\ldots\ldots\ldots\ldots\ldots(9)$$

where

$$2^{\frac14}q^{\frac1{24}}u=(1-q)(1-q^3)(1-q^5)\ldots$$

and

$$2^{\frac14}q^{\frac5{24}}v=(1-q^5)(1-q^{15})(1-q^{25})\ldots.$$

If we put $q=e^{-\pi/\sqrt5}$ in (8), so that $u=G_{\frac15}$ and $v=G_5$, and hence $u=v$, we see that

$$v^4-v^{-4}=1.$$

Hence

$$v^4=\frac{1+\sqrt5}{2},\quad G_5=\left(\frac{1+\sqrt5}{2}\right)^{\frac14}.$$

Similarly, by putting $q=e^{-\pi\sqrt{\frac25}}$, so that $u=g_{\frac25}$ and $v=g_{10}$, and hence $u=1/v$, we see that

$$v^6-v^{-6}=4.$$

Hence

$$v^2=\frac{1+\sqrt5}{2},\quad g_{10}=\sqrt{\left(\frac{1+\sqrt5}{2}\right)}.$$

Similarly it can be shewn that

$$G_9=\left(\frac{1+\sqrt3}{\sqrt2}\right)^{\frac13},\quad g_{18}=(\sqrt2+\sqrt3)^{\frac13},$$

$$G_{17}=\sqrt{\left(\frac{5+\sqrt{17}}{8}\right)}+\sqrt{\left(\frac{\sqrt{17}-3}{8}\right)},$$

$$g_{34}=\sqrt{\left(\frac{7+\sqrt{17}}{8}\right)}+\sqrt{\left(\frac{\sqrt{17}-1}{8}\right)},$$

and so on.

3. In order to obtain approximations for π we take logarithms of (1) and (2). Thus

$$\left.\begin{aligned}\pi&=\frac{24}{\sqrt n}\log(2^{\frac14}G_n)\\ \pi&=\frac{24}{\sqrt n}\log(2^{\frac14}g_n)\end{aligned}\right\}, \quad\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(10)$$

approximately, the error being nearly $\frac{24}{\sqrt n}e^{-\pi\sqrt n}$ in both cases. These equations may also be written as

$$e^{\pi\sqrt{n}/24}=2^{\frac14}G_n,\quad e^{\pi\sqrt{n}/24}=2^{\frac14}g_n. \quad\ldots\ldots\ldots\ldots\ldots\ldots\ldots(11)$$

In those cases in which G_n^{12} and g_n^{12} are simple quadratic surds we may use the forms

$$(G_n^{12}+G_n^{-12})^{\frac1{12}},\quad (g_n^{12}+g_n^{-12})^{\frac1{12}},$$

instead of G_n and g_n, for we have

$$g_n^{12}=\tfrac18e^{\frac12\pi\sqrt n}-\tfrac32e^{-\frac12\pi\sqrt n},$$

approximately, and so

$$g_n^{12}+g_n^{-12}=\tfrac18e^{\frac12\pi\sqrt n}+\tfrac{13}{2}e^{-\frac12\pi\sqrt n},$$

approximately, so that

$$\pi = \frac{2}{\sqrt{n}} \log \{8(g_n^{12}+g_n^{-12})\}, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(12)$$

the error being about $\frac{104}{\sqrt{n}}e^{-\pi\sqrt{n}}$, which is of the same order as the error in the formulæ (10). The formula (12) often leads to simpler results. Thus the second of formulae (10) gives

$$e^{\pi\sqrt{18}/24} = 2^{\frac14} g_{18}$$

or

$$e^{\frac14\pi\sqrt{18}} = 10\sqrt2 + 8\sqrt3.$$

But if we use the formula (12), or

$$e^{\pi\sqrt{n}/24} = 2^{\frac14}(g_n^{12}+g_n^{-12})^{\frac{1}{12}},$$

we get a simpler form, viz.

$$e^{\frac18\pi\sqrt{18}} = 2\sqrt7.$$

4. The values of g_{2n} and G_n are obtained from the same equation. The approximation by means of g_{2n} is preferable to that by G_n for the following reasons.

(*a*) It is more accurate. Thus the error when we use G_{65} contains a factor $e^{-\pi\sqrt{65}}$, whereas that when we use g_{130} contains a factor $e^{-\pi\sqrt{130}}$.

(*b*) For many values of n, g_{2n} is simpler in form than G_n; thus

$$g_{130} = \sqrt{\left\{(2+\sqrt5)\left(\frac{3+\sqrt{13}}{2}\right)\right\}},$$

while

$$G_{65} = \left\{\left(\frac{1+\sqrt5}{2}\right)\left(\frac{3+\sqrt{13}}{2}\right)\right\}^{\frac14} \sqrt{\left\{\sqrt{\left(\frac{9+\sqrt{65}}{8}\right)} + \sqrt{\left(\frac{1+\sqrt{65}}{8}\right)}\right\}}.$$

(*c*) For many values of n, g_{2n} involves quadratic surds only, even when G_n is a root of an equation of higher order. Thus G_{23}, G_{29}, G_{31} are roots of cubic equations, G_{47}, G_{79} are those of quintic equations, and G_{71} is that of a septic equation, while g_{46}, g_{58}, g_{62}, g_{94}, g_{142}, and g_{158} are all expressible by quadratic surds.

5. Since G_n and g_n can be expressed as roots of algebraical equations with rational coefficients, the same is true of G_n^{24} or g_n^{24}. So let us suppose that

$$1 = ag_n^{-24} - bg_n^{-48} + \ldots,$$

or

$$g_n^{24} = a - bg_n^{-24} + \ldots.$$

But we know that

$$64e^{-\pi\sqrt{n}}g_n^{24} = 1 - 24e^{-\pi\sqrt{n}} + 276e^{-2\pi\sqrt{n}} - \ldots,$$

$$64g_n^{24} = e^{\pi\sqrt{n}} - 24 + 276e^{-\pi\sqrt{n}} - \ldots,$$

$$64a - 64bg_n^{-24} + \ldots = e^{\pi\sqrt{n}} - 24 + 276e^{-\pi\sqrt{n}} - \ldots,$$

$$64a - 4096be^{-\pi\sqrt{n}} + \ldots = e^{\pi\sqrt{n}} - 24 + 276e^{-\pi\sqrt{n}} - \ldots,$$

that is

$$e^{\pi\sqrt{n}} = (64a+24) - (4096b+276)e^{-\pi\sqrt{n}} + \ldots. \quad \ldots\ldots\ldots(13)$$

Similarly, if

$$1 = aG_n^{-24} - bG_n^{-48} + \ldots,$$

then

$$e^{\pi\sqrt{n}} = (64a-24) - (4096b+276)e^{-\pi\sqrt{n}} + \ldots. \quad \ldots\ldots\ldots(14)$$

From (13) and (14) we can find whether $e^{\pi\sqrt{n}}$ is very nearly an integer for given values of n, and ascertain also the number of 9's or 0's in the decimal part. But if G_n and g_n be simple quadratic surds we may work independently as follows. We have, for example,

$$g_{22} = \sqrt{(1+\sqrt2)}.$$

Hence

$$64g_{22}^{24} = e^{\pi\sqrt{22}} - 24 + 276e^{-\pi\sqrt{22}} - \ldots,$$

$$64g_{22}^{-24} = 4096e^{-\pi\sqrt{22}} + \ldots,$$

so that

$$64(g_{22}^{24}+g_{22}^{-24}) = e^{\pi\sqrt{22}} - 24 + 4372e^{-\pi\sqrt{22}} + \ldots = 64[(1+\sqrt2)^{12} + (1-\sqrt2)^{12}].$$

Hence

$$e^{\pi\sqrt{22}} = 2508951{\cdot}9982\ldots.$$

Again

$$G_{37} = (6+\sqrt{37})^{\frac14},$$

$$64G_{37}^{24} = e^{\pi\sqrt{37}} + 24 + 276e^{-\pi\sqrt{37}} + \ldots,$$

$$64G_{37}^{-24} = 4096e^{-\pi\sqrt{37}} - \ldots,$$

so that

$$64(G_{37}^{24}+G_{37}^{-24}) = e^{\pi\sqrt{37}} + 24 + 4372e^{-\pi\sqrt{37}} - \ldots = 64[(6+\sqrt{37})^6 + (6-\sqrt{37})^6].$$

Hence

$$e^{\pi\sqrt{37}} = 199148647{\cdot}999978\ldots.$$

Similarly, from

$$g_{58} = \sqrt{\left(\frac{5+\sqrt{29}}{2}\right)},$$

we obtain

$$64(g_{58}^{24}+g_{58}^{-24}) = e^{\pi\sqrt{58}} - 24 + 4372e^{-\pi\sqrt{58}} + \ldots$$
$$= 64\left\{\left(\frac{5+\sqrt{29}}{2}\right)^{12} + \left(\frac{5-\sqrt{29}}{2}\right)^{12}\right\}.$$

Hence

$$e^{\pi\sqrt{58}} = 24591257751{\cdot}99999982\ldots.$$

6. I have calculated the values of G_n and g_n for a large number of values of n. Many of these results are equivalent to results given by Weber; for example,

$$G_{13}^4 = \frac{3+\sqrt{13}}{2}, \qquad G_{25} = \frac{1+\sqrt5}{2},$$

$$g_{30}^6 = (2+\sqrt5)(3+\sqrt{10}), \qquad G_{37}^4 = 6+\sqrt{37},$$

$$G_{49} = \frac{7^{\frac14}+\sqrt{(4+\sqrt7)}}{2}, \qquad g_{58}^2 = \frac{5+\sqrt{29}}{2},$$

$$g_{70}^2 = \frac{(3+\sqrt5)(1+\sqrt2)}{2},$$

$$G_{73} = \sqrt{\left(\frac{9+\sqrt{73}}{8}\right)} + \sqrt{\left(\frac{1+\sqrt{73}}{8}\right)},$$

$$G_{85} = \left(\frac{1+\sqrt5}{2}\right)\left(\frac{9+\sqrt{85}}{2}\right)^{\frac14},$$

$$G_{97} = \sqrt{\left(\frac{13+\sqrt{97}}{8}\right)} + \sqrt{\left(\frac{5+\sqrt{97}}{8}\right)},$$

$$g_{190}^2 = (2+\sqrt5)(3+\sqrt{10}),$$

$$G_{385}^2 = \tfrac18(3+\sqrt{11})(\sqrt5+\sqrt7)(\sqrt7+\sqrt{11})(3+\sqrt5),$$

and so on. I have also many results not given by Weber. I give a complete table of new results. In Weber's notation, $G_n = 2^{-\frac14}f\{\sqrt{(-n)}\}$ and $g_n = 2^{-\frac14}f_1\{\sqrt{(-n)}\}$.

TABLE I.

$$g_{62} + \frac{1}{g_{62}} = \tfrac12\{\sqrt{(1+\sqrt2)} + \sqrt{(9+5\sqrt2)}\},$$

$$G_{65}^2 = \sqrt{\left\{\left(\frac{1+\sqrt5}{2}\right)\left(\frac{3+\sqrt{13}}{2}\right)\right\}}\left\{\sqrt{\left(\frac{1+\sqrt{65}}{8}\right)} + \sqrt{\left(\frac{9+\sqrt{65}}{8}\right)}\right\},$$

$$g_{66}^2 = \sqrt{(\sqrt2+\sqrt3)}(7\sqrt2+3\sqrt{11})^{\frac16}\left\{\sqrt{\left(\frac{7+\sqrt{33}}{8}\right)} + \sqrt{\left(\frac{\sqrt{33}-1}{8}\right)}\right\},$$

$$G_{69}^2 = (3\sqrt3+\sqrt{23})^{\frac14}\left(\frac{5+\sqrt{23}}{4}\right)^{\frac16}\left\{\sqrt{\left(\frac{6+3\sqrt3}{4}\right)} + \sqrt{\left(\frac{2+3\sqrt3}{4}\right)}\right\},$$

$$G_{77}^2 = \{\tfrac12(\sqrt7+\sqrt{11})(8+3\sqrt7)\}^{\frac14}\left\{\sqrt{\left(\frac{6+\sqrt{11}}{4}\right)} + \sqrt{\left(\frac{2+\sqrt{11}}{4}\right)}\right\},$$

$$G_{81}^3 = \frac{(2\sqrt3+2)^{\frac13}+1}{(2\sqrt3-2)^{\frac13}-1},$$

$$g_{90} = \{(2+\sqrt5)(\sqrt5+\sqrt6)\}^{\frac16}\left\{\sqrt{\left(\frac{3+\sqrt6}{4}\right)} + \sqrt{\left(\frac{\sqrt6-1}{4}\right)}\right\},$$

$$g_{94} + \frac{1}{g_{94}} = \tfrac12\{\sqrt{(7+\sqrt2)} + \sqrt{(7+5\sqrt2)}\},$$

$$g_{98} + \frac{1}{g_{98}} = \tfrac12\{\sqrt2 + \sqrt{(14+4\sqrt{14})}\},$$

$$g_{114}^2 = \sqrt{(\sqrt2+\sqrt3)}(3\sqrt2+\sqrt{19})^{\frac16}\left\{\sqrt{\left(\frac{23+3\sqrt{57}}{8}\right)} + \sqrt{\left(\frac{15+3\sqrt{57}}{8}\right)}\right\},$$

$$G_{117} = \tfrac12\left(\frac{3+\sqrt{13}}{2}\right)^{\frac14}(2\sqrt3+\sqrt{13})^{\frac16}\{3^{\frac14} + \sqrt{(4+\sqrt3)}\},$$

$$\left.\begin{aligned} G_{121} + \frac{1}{G_{121}} &= \left(\frac{11}{2}\right)^{\frac16}\left\{\left(3+\frac{1}{3\sqrt3}\right)^{\frac13} + \left(3-\frac{1}{3\sqrt3}\right)^{\frac13}\right\} \\ \frac{1}{G_{121}} &= \frac{1}{3\sqrt2}[(11-3\sqrt{11})^{\frac13}\{(3\sqrt{11}+3\sqrt3-4)^{\frac13} \\ &\qquad + (3\sqrt{11}-3\sqrt3-4)^{\frac13}\} - 2] \end{aligned}\right\},$$

$$g_{126} = \sqrt{\left(\frac{\sqrt3+\sqrt7}{2}\right)}(\sqrt6+\sqrt7)^{\frac16}\left\{\sqrt{\left(\frac{3+\sqrt2}{4}\right)} + \sqrt{\left(\frac{\sqrt2-1}{4}\right)}\right\}^2,$$

$$g_{138}^2 = \sqrt{\left(\frac{3\sqrt3+\sqrt{23}}{2}\right)}(78\sqrt2+23\sqrt{23})^{\frac16}$$
$$\times\left\{\sqrt{\left(\frac{5+2\sqrt6}{4}\right)} + \sqrt{\left(\frac{1+2\sqrt6}{4}\right)}\right\},$$

$$G_{141}^2 = (4\sqrt3+\sqrt{47})^{\frac14}\left(\frac{7+\sqrt{47}}{\sqrt2}\right)^{\frac16}\left\{\sqrt{\left(\frac{18+9\sqrt3}{4}\right)} + \sqrt{\left(\frac{14+9\sqrt3}{4}\right)}\right\},$$

$$G_{145}^2 = \sqrt{\left\{\frac{(2+\sqrt5)(5+\sqrt{29})}{2}\right\}}\left\{\sqrt{\left(\frac{17+\sqrt{145}}{8}\right)} + \sqrt{\left(\frac{9+\sqrt{145}}{8}\right)}\right\},$$

$$\frac{1}{G_{147}} = 2^{-\frac{1}{12}}\left[\frac12 + \frac{1}{\sqrt3}\left\{\sqrt{\left(\frac74\right)} - (28)^{\frac16}\right\}\right],$$

$$G_{153} = \left\{\sqrt{\left(\frac{5+\sqrt{17}}{8}\right)} + \sqrt{\left(\frac{\sqrt{17}-3}{8}\right)}\right\}^2$$
$$\times\left\{\sqrt{\left(\frac{37+9\sqrt{17}}{4}\right)} + \sqrt{\left(\frac{33+9\sqrt{17}}{4}\right)}\right\}^{\frac13},$$

$$g_{154}^2 = \sqrt{\left\{(2\sqrt2+\sqrt7)\left(\frac{\sqrt7+\sqrt{11}}{2}\right)\right\}}$$
$$\times\left\{\sqrt{\left(\frac{13+2\sqrt{22}}{4}\right)} + \sqrt{\left(\frac{9+2\sqrt{22}}{4}\right)}\right\},$$

$$g_{158} + \frac{1}{g_{158}} = \tfrac12\{\sqrt{(9+\sqrt2)} + \sqrt{(17+13\sqrt2)}\},$$

$$\left.\begin{aligned} G_{169} + \frac{1}{G_{169}} &= \left(\frac{13}{4}\right)^{\frac16}\left\{\left(1+\frac{1}{3\sqrt3}\right)^{\frac13} + \left(1-\frac{1}{3\sqrt3}\right)^{\frac13}\right\}^2 \\ \frac{1}{G_{169}} &= \frac13\left[(\sqrt{13}-2) + \left(\frac{13-3\sqrt{13}}{2}\right)^{\frac13}\right. \\ &\qquad \left.\times\left\{\left(3\sqrt3-\frac{11-\sqrt{13}}{2}\right)^{\frac13} - \left(3\sqrt3+\frac{11-\sqrt{13}}{2}\right)^{\frac13}\right\}\right] \end{aligned}\right\},$$

$$g_{198} = \sqrt{(1+\sqrt2)}(4\sqrt2+\sqrt{33})^{\frac16}\left\{\sqrt{\left(\frac{9+\sqrt{33}}{8}\right)} + \sqrt{\left(\frac{1+\sqrt{33}}{8}\right)}\right\},$$

$$G_{205} = \left(\frac{1+\sqrt5}{2}\right)\left(\frac{3\sqrt5+\sqrt{41}}{2}\right)^{\frac14}\left\{\sqrt{\left(\frac{7+\sqrt{41}}{8}\right)} + \sqrt{\left(\frac{\sqrt{41}-1}{8}\right)}\right\},$$

$$G_{213}^2 = (5\sqrt3+\sqrt{71})^{\frac14}\left(\frac{59+7\sqrt{71}}{4}\right)^{\frac16}$$
$$\times\left\{\sqrt{\left(\frac{21+12\sqrt3}{2}\right)} + \sqrt{\left(\frac{19+12\sqrt3}{2}\right)}\right\},$$

$$G_{217}^2 = \left\{\sqrt{\left(\frac{9+4\sqrt7}{2}\right)} + \sqrt{\left(\frac{11+4\sqrt7}{2}\right)}\right\}$$
$$\times\left\{\sqrt{\left(\frac{12+5\sqrt7}{4}\right)} + \sqrt{\left(\frac{16+5\sqrt7}{4}\right)}\right\},$$

$$G_{225} = \left(\frac{1+\sqrt5}{4}\right)(2+\sqrt3)^{\frac13}\{\sqrt{(4+\sqrt{15})} + 15^{\frac14}\},$$

$$g_{238}=\left\{\sqrt{\left(\frac{1+2\sqrt2}{4}\right)}+\sqrt{\left(\frac{5+2\sqrt2}{4}\right)}\right\}$$
$$\times\left\{\sqrt{\left(\frac{1+3\sqrt2}{4}\right)}+\sqrt{\left(\frac{5+3\sqrt2}{4}\right)}\right\},$$

$$G_{265}^2=\sqrt{\left\{(2+\sqrt5)\left(\frac{7+\sqrt{53}}{2}\right)\right\}}\left\{\sqrt{\left(\frac{89+5\sqrt{265}}{8}\right)}+\sqrt{\left(\frac{81+5\sqrt{265}}{8}\right)}\right\},$$

$$G_{289}=\left[\sqrt{\left\{\frac{17+\sqrt{17}+17^{\frac14}(5+\sqrt{17})}{16}\right\}}+\sqrt{\left\{\frac{1+\sqrt{17}+17^{\frac14}(5+\sqrt{17})}{16}\right\}}\right]^2,$$

$$G_{301}^2=\left\{(8+3\sqrt7)\left(\frac{23\sqrt{43}+57\sqrt7}{2}\right)\right\}^{\frac14}$$
$$\times\left\{\sqrt{\left(\frac{46+7\sqrt{43}}{4}\right)}+\sqrt{\left(\frac{42+7\sqrt{43}}{4}\right)}\right\},$$

$$g_{310}=\left(\frac{1+\sqrt5}{2}\right)\sqrt{(1+\sqrt2)}\left\{\sqrt{\left(\frac{7+2\sqrt{10}}{4}\right)}+\sqrt{\left(\frac{3+2\sqrt{10}}{4}\right)}\right\},$$

$$\left.\begin{aligned}G_{325}&=\left(\frac{3+\sqrt{13}}{2}\right)^{\frac14}t,\text{ where}\\ &t^3+t^2\left(\frac{1-\sqrt{13}}{2}\right)^2+t\left(\frac{1+\sqrt{13}}{2}\right)^2+1\\ &\qquad=\sqrt5\left\{t^3-t^2\left(\frac{1+\sqrt{13}}{2}\right)+t\left(\frac{1-\sqrt{13}}{2}\right)-1\right\}\end{aligned}\right\},$$

$$G_{333}=\tfrac12(6+\sqrt{37})^{\frac14}(7\sqrt3+2\sqrt{37})^{\frac16}\{\sqrt{(7+2\sqrt3)}+\sqrt{(3+2\sqrt3)}\},$$

$$\left.\begin{aligned}G_{363}&=2^{\frac1{12}}t,\text{ where}\\ &2t^3-t^2\{(4+\sqrt{33})+\sqrt{(11+2\sqrt{33})}\}\\ &\qquad-t\{1+\sqrt{(11+2\sqrt{33})}\}-1=0\end{aligned}\right\},$$

$$G_{441}^2=\left(\frac{\sqrt3+\sqrt7}{2}\right)(2+\sqrt3)^{\frac13}\left\{\frac{2+\sqrt7+\sqrt{(7+4\sqrt7)}}{2}\right\}\left\{\frac{\sqrt{(3+\sqrt7)}+(6\sqrt7)^{\frac14}}{\sqrt{(3+\sqrt7)}-(6\sqrt7)^{\frac14}}\right\},$$

$$G_{445}=\sqrt{(2+\sqrt5)}\left(\frac{21+\sqrt{445}}{2}\right)^{\frac14}\sqrt{\Big\{\Big(\frac{13+\sqrt{89}}{8}\Big)}+\sqrt{\left(\frac{5+\sqrt{89}}{8}\right)}\Big\},$$

$$G_{465}^2=\sqrt{\left\{(2+\sqrt3)\left(\frac{1+\sqrt5}{2}\right)\left(\frac{3\sqrt3+\sqrt{31}}{2}\right)\right\}}(5\sqrt5+2\sqrt{31})^{\frac16}$$
$$\times\left\{\sqrt{\left(\frac{2+\sqrt{31}}{4}\right)}+\sqrt{\left(\frac{6+\sqrt{31}}{4}\right)}\right\}$$
$$\times\left\{\sqrt{\left(\frac{11+2\sqrt{31}}{2}\right)}+\sqrt{\left(\frac{13+2\sqrt{31}}{2}\right)}\right\},$$

$$G_{505}^2=(2+\sqrt5)\sqrt{\left\{\left(\frac{1+\sqrt5}{2}\right)(10+\sqrt{101})\right\}}$$
$$\times\left\{\left(\frac{5\sqrt5+\sqrt{101}}{4}\right)+\sqrt{\left(\frac{105+\sqrt{505}}{8}\right)}\right\},$$

$$g_{522}=\sqrt{\left(\frac{5+\sqrt{29}}{2}\right)}(5\sqrt{29}+11\sqrt6)^{\frac16}\left\{\sqrt{\left(\frac{9+3\sqrt6}{4}\right)}+\sqrt{\left(\frac{5+3\sqrt6}{4}\right)}\right\},$$

$$G_{553}^2=\left\{\sqrt{\left(\frac{96+11\sqrt{79}}{4}\right)}+\sqrt{\left(\frac{100+11\sqrt{79}}{4}\right)}\right\}$$
$$\times\left\{\sqrt{\left(\frac{141+16\sqrt{79}}{2}\right)}+\sqrt{\left(\frac{143+16\sqrt{79}}{2}\right)}\right\},$$

$$g_{630}=(\sqrt{14}+\sqrt{15})^{\frac16}\sqrt{\left\{(1+\sqrt2)\left(\frac{3+\sqrt5}{2}\right)\left(\frac{\sqrt3+\sqrt7}{2}\right)\right\}}$$
$$\times\left\{\sqrt{\left(\frac{\sqrt{15}+\sqrt7+2}{4}\right)}+\sqrt{\left(\frac{\sqrt{15}+\sqrt7-2}{4}\right)}\right\}$$
$$\times\left\{\sqrt{\left(\frac{\sqrt{15}+\sqrt7+4}{8}\right)}+\sqrt{\left(\frac{\sqrt{15}+\sqrt7-4}{8}\right)}\right\},$$

$$G_{765}^2=\left(\frac{3+\sqrt5}{2}\right)(16+\sqrt{255})^{\frac16}\sqrt{\left\{(4+\sqrt{15})\left(\frac{9+\sqrt{85}}{2}\right)\right\}}$$
$$\times\left\{\sqrt{\left(\frac{6+\sqrt{51}}{4}\right)}+\sqrt{\left(\frac{10+\sqrt{51}}{4}\right)}\right\}$$
$$\times\left\{\sqrt{\left(\frac{18+3\sqrt{51}}{4}\right)}+\sqrt{\left(\frac{22+3\sqrt{51}}{4}\right)}\right\},$$

$$G_{777}^2=\sqrt{\left\{(2+\sqrt3)(6+\sqrt{37})\left(\frac{\sqrt3+\sqrt7}{2}\right)\right\}}(246\sqrt7+107\sqrt{37})^{\frac16}$$
$$\times\left\{\sqrt{\left(\frac{6+3\sqrt7}{4}\right)}+\sqrt{\left(\frac{10+3\sqrt7}{4}\right)}\right\}$$
$$\times\left\{\sqrt{\left(\frac{15+6\sqrt7}{2}\right)}+\sqrt{\left(\frac{17+6\sqrt7}{2}\right)}\right\},$$

$$G_{1225}=\left(\frac{1+\sqrt5}{2}\right)(6+\sqrt{35})^{\frac14}\left\{\frac{7^{\frac14}+\sqrt{(4+\sqrt7)}}{2}\right\}^{\frac32}$$
$$\times\left[\sqrt{\left\{\frac{43+15\sqrt7+(8+3\sqrt7)\sqrt{(10\sqrt7)}}{8}\right\}}\right.$$
$$\left.+\sqrt{\left\{\frac{35+15\sqrt7+(8+3\sqrt7)\sqrt{(10\sqrt7)}}{8}\right\}}\right],$$

$$G_{1353}^2=\sqrt{\left\{(3+\sqrt{11})(5+3\sqrt3)\left(\frac{11+\sqrt{123}}{2}\right)\right\}}$$
$$\times\left(\frac{6817+321\sqrt{451}}{4}\right)^{\frac16}$$
$$\times\left\{\sqrt{\left(\frac{17+3\sqrt{33}}{8}\right)}+\sqrt{\left(\frac{25+3\sqrt{33}}{8}\right)}\right\}$$
$$\times\left\{\sqrt{\left(\frac{561+99\sqrt{33}}{8}\right)}+\sqrt{\left(\frac{569+99\sqrt{33}}{8}\right)}\right\},$$

$$G_{1645}^2=(2+\sqrt5)\sqrt{\left\{(3+\sqrt7)\left(\frac{7+\sqrt{47}}{2}\right)\right\}}\left(\frac{73\sqrt5+9\sqrt{329}}{2}\right)^{\frac14}$$
$$\times\left\{\sqrt{\left(\frac{119+7\sqrt{329}}{8}\right)}+\sqrt{\left(\frac{127+7\sqrt{329}}{8}\right)}\right\}$$
$$\times\left\{\sqrt{\left(\frac{743+41\sqrt{329}}{8}\right)}+\sqrt{\left(\frac{751+41\sqrt{329}}{8}\right)}\right\}.$$

7. Hence we deduce the following approximate formulæ.

TABLE II.

$$e^{\frac18\pi\sqrt{18}}=2\sqrt7,\quad e^{\pi\sqrt{22}/12}=2+\sqrt2,\quad e^{\frac14\pi\sqrt{30}}=20\sqrt3+16\sqrt6,$$
$$e^{\frac14\pi\sqrt{34}}=12(4+\sqrt{17}),\quad e^{\frac12\pi\sqrt{46}}=144(147+104\sqrt2),$$
$$e^{\frac14\pi\sqrt{42}}=84+32\sqrt6,\quad e^{\pi\sqrt{58}/12}=\frac{5+\sqrt{29}}{\sqrt2},$$
$$e^{\frac14\pi\sqrt{70}}=60\sqrt{35}+96\sqrt{14},\quad e^{\frac14\pi\sqrt{78}}=300\sqrt3+208\sqrt6,$$
$$e^{\pi\sqrt{55}/24}=\frac{1+\sqrt{(3+2\sqrt5)}}{\sqrt2},\quad e^{\frac14\pi\sqrt{102}}=800\sqrt3+196\sqrt{51},$$
$$e^{\frac14\pi\sqrt{130}}=12(323+40\sqrt{65}),\quad e^{\pi\sqrt{190}/12}=(2\sqrt2+\sqrt{10})(3+\sqrt{10}),$$
$$\pi=\frac{12}{\sqrt{130}}\log\left\{\frac{(2+\sqrt5)(3+\sqrt{13})}{\sqrt2}\right\},$$
$$\pi=\frac{24}{\sqrt{142}}\log\left\{\sqrt{\left(\frac{10+11\sqrt2}{4}\right)}+\sqrt{\left(\frac{10+7\sqrt2}{4}\right)}\right\},$$
$$\pi=\frac{12}{\sqrt{190}}\log\{(2\sqrt2+\sqrt{10})(3+\sqrt{10})\},$$
$$\pi=\frac{12}{\sqrt{310}}\log[\tfrac14(3+\sqrt5)(2+\sqrt2)\{(5+2\sqrt{10})+\sqrt{(61+20\sqrt{10})}\}],$$
$$\pi=\frac{4}{\sqrt{522}}\log\left[\left(\frac{5+\sqrt{29}}{\sqrt2}\right)^3(5\sqrt{29}+11\sqrt6)\right.$$
$$\left.\times\left\{\sqrt{\left(\frac{9+3\sqrt6}{4}\right)}+\sqrt{\left(\frac{5+3\sqrt6}{4}\right)}\right\}^6\right].$$

The last five formulæ are correct to 15, 16, 18, 22, and 31 places of decimals respectively.

8. Thus we have seen how to approximate to π by means of logarithms of surds. I shall now shew how to obtain approximations in terms of surds only. If

$$n\frac{K'}{K}=\frac{L'}{L},$$

we have
$$\frac{ndk}{kk'^2K^2}=\frac{dl}{ll'^2L^2}.$$

But, by means of the modular equation connecting k and l, we can express dk/dl as an algebraic function of k, a function moreover in which all coefficients which occur are algebraic numbers. Again,

$$q=e^{-\pi K'/K},\quad q^n=e^{-\pi L'/L},$$
$$\frac{q^{\frac1{12}}(1-q^2)(1-q^4)(1-q^6)\ldots}{q^{\frac{n}{12}}(1-q^{2n})(1-q^{4n})(1-q^{6n})}=\left(\frac{kk'}{ll'}\right)^{\frac16}\sqrt{\left(\frac{K}{L}\right)}.\quad\ldots\ldots(15)$$

Differentiating this equation logarithmically, and using the formula

$$\frac{dq}{dk}=\frac{\pi^2q}{2kk'^2K^2},$$

we see that

$$n\left\{1-24\left(\frac{q^{2n}}{1-q^{2n}}+\frac{2q^{4n}}{1-q^{4n}}+\ldots\right)\right\}$$
$$-\left\{1-24\left(\frac{q^2}{1-q^2}+\frac{2q^4}{1-q^4}+\ldots\right)\right\}=\frac{KL}{\pi^2}A(k),\quad\ldots(16)$$

where $A(k)$ denotes an algebraic function of the special class described above. I shall use the letter A generally to denote a function of this type.

Now, if we put $k=l'$ and $k'=l$ in (16), we have

$$n\left\{1-24\left(\frac{1}{e^{2\pi\sqrt n}-1}+\frac{2}{e^{4\pi\sqrt n}-1}+\ldots\right)\right\}$$
$$-\left\{1-24\left(\frac{1}{e^{2\pi/\sqrt n}-1}+\frac{2}{e^{4\pi/\sqrt n}-1}+\ldots\right)\right\}=\left(\frac{K}{\pi}\right)^2A(k).\quad\ldots(17)$$

The algebraic function $A(k)$ of course assumes a purely numerical form when we substitute the value of k deduced from the modular equation. But by substituting $k=l'$ and $k'=l$ in (15) we have

$$n^{\frac14}e^{-\pi\sqrt n/12}(1-e^{-2\pi\sqrt n})(1-e^{-4\pi\sqrt n})(1-e^{-6\pi\sqrt n})\ldots$$
$$=e^{-\pi/(12\sqrt n)}(1-e^{-2\pi/\sqrt n})(1-e^{-4\pi/\sqrt n})(1-e^{-6\pi/\sqrt n})\ldots.$$

Differentiating the above equation logarithmically we have

$$n\left\{1-24\left(\frac{1}{e^{2\pi\sqrt n}-1}+\frac{2}{e^{4\pi\sqrt n}-1}+\ldots\right)\right\}$$
$$+\left\{1-24\left(\frac{1}{e^{2\pi/\sqrt n}-1}+\frac{2}{e^{4\pi/\sqrt n}-1}+\ldots\right)\right\}=\frac{6\sqrt n}{\pi}.\quad\ldots\ldots\ldots(18)$$

Now, adding (17) and (18), we have

$$1-\frac{3}{\pi\sqrt n}-24\left(\frac{1}{e^{2\pi\sqrt n}-1}+\frac{2}{e^{4\pi\sqrt n}-1}+\ldots\right)=\left(\frac{K}{\pi}\right)^2A(k).\quad\ldots\ldots(19)$$

But it is known that

$$1-24\left(\frac{q}{1+q}+\frac{3q^3}{1+q^3}+\frac{5q^5}{1+q^5}+\ldots\right)=\left(\frac{2K}{\pi}\right)^2(1-2k^2),$$

so that
$$1-24\left(\frac{1}{e^{\pi\sqrt n}+1}+\frac{3}{e^{3\pi\sqrt n}+1}+\ldots\right)=\left(\frac{K}{\pi}\right)^2A(k).\quad\ldots\ldots\ldots(20)$$

Hence, dividing (19) by (20), we have

$$\frac{1-\dfrac{3}{\pi\sqrt n}-24\left(\dfrac{1}{e^{2\pi\sqrt n}-1}+\dfrac{2}{e^{4\pi\sqrt n}-1}+\ldots\right)}{1-24\left(\dfrac{1}{e^{\pi\sqrt n}+1}+\dfrac{3}{e^{3\pi\sqrt n}+1}+\ldots\right)}=R,\quad\ldots\ldots\ldots(21)$$

where R can always be expressed in radicals if n is any rational number. Hence we have

$$\pi = \frac{3}{(1-R)\sqrt{n}}, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(22)$$

nearly, the error being about $8\pi e^{-\pi\sqrt{n}}(\pi\sqrt{n}-3)$.

9. We may get a still closer approximation from the following results. It is known that

$$1 + 240\sum_{r=1}^{r=\infty}\frac{r^3q^{2r}}{1-q^{2r}} = \left(\frac{2K}{\pi}\right)^4(1-k^2k'^2),$$

and also that

$$1 - 504\sum_{r=1}^{r=\infty}\frac{r^5q^{2r}}{1-q^{2r}} = \left(\frac{2K}{\pi}\right)^6(1-2k^2)(1+\tfrac{1}{2}k^2k'^2).$$

Hence, from (19), we see that

$$\left\{1-\frac{3}{\pi\sqrt{n}}-24\sum_{r=1}^{r=\infty}\frac{r}{e^{2\pi r\sqrt{n}}-1}\right\}\left\{1+240\sum_{r=1}^{r=\infty}\frac{r^3}{e^{2\pi r\sqrt{n}}-1}\right\}$$
$$= R'\left\{1-504\sum_{r=1}^{r=\infty}\frac{r^5}{e^{2\pi r\sqrt{n}}-1}\right\}, \ \ldots\ldots(23)$$

where R' can always be expressed in radicals for any rational value of n. Hence

$$\pi = \frac{3}{(1-R')\sqrt{n}}, \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(24)$$

nearly, the error being about $24\pi(10\pi\sqrt{n}-31)e^{-2\pi\sqrt{n}}$

It will be seen that the error in (24) is much less than that in (22), if n is at all large.

10. In order to find R and R' the series in (16) must be calculated in finite terms. I shall give the final results for a few values of n.

TABLE III.

$$q = e^{-\pi K'/K}, \quad q^n = e^{-\pi L'/L},$$

$$f(q) = n\left(1-24\sum_1^\infty\frac{q^{2mn}}{1-q^{2mn}}\right) - \left(1-24\sum_1^\infty\frac{q^{2m}}{1-q^{2m}}\right).$$

$$f(2) = \frac{4KL}{\pi^2}(k'+l),$$

$$f(3) = \frac{4KL}{\pi^2}(1+kl+k'l'),$$

$$f(4) = \frac{4KL}{\pi^2}(\sqrt{k'}+\sqrt{l})^2,$$

$$f(5) = \frac{4KL}{\pi^2}(3+kl+k'l')\sqrt{\left(\frac{1+kl+k'l'}{2}\right)},$$

$$f(11) = \frac{8KL}{\pi^2}\{2(1+kl+k'l')+\sqrt{(kl)}+\sqrt{(k'l')}-\sqrt{(kk'll')}\},$$

$$f(15) = \frac{4KL}{\pi^2}[\{1+(kl)^{\frac{1}{2}}+(k'l')^{\frac{1}{2}}\}^2-\{1+kl+k'l'\}],$$

$$f(17) = \frac{4KL}{\pi^2}\sqrt{\{44(1+k^2l^2+k'^2l'^2)+168(kl+k'l'-kk'll')}$$
$$-102(1-kl-k'l')(4kk'll')^{\frac{1}{2}}-192(4kk'll')^{\frac{3}{4}}\},$$

$$f(19) = \frac{24KL}{\pi^2}\{(1+kl+k'l')+\sqrt{(kl)}+\sqrt{(k'l')}-\sqrt{(kk'll')}\},$$

$$f(23) = \frac{4KL}{\pi^2}[11(1+kl+k'l')-16(4kk'll')^{\frac{1}{6}}\{1+\sqrt{(kl)}+\sqrt{(k'l')}\}$$
$$-20(4kk'll')^{\frac{1}{3}}],$$

$$f(31) = \frac{12KL}{\pi^2}[3(1+kl+k'l')+4\{\sqrt{(kl)}+\sqrt{(k'l')}+\sqrt{(kk'll')}\}$$
$$-4(kk'll')^{\frac{1}{4}}\{1+(kl)^{\frac{1}{4}}+(k'l')^{\frac{1}{4}}\}],$$

$$f(35) = \frac{4KL}{\pi^2}[2\{\sqrt{(kl)}+\sqrt{(k'l')}-\sqrt{(kk'll')}\}$$
$$+(4kk'll')^{-\frac{1}{6}}\{1-\sqrt{(kl)}-\sqrt{(k'l')}\}^2].$$

Thus the sum of the series (19) can be found in finite terms, when $n = 2, 3, 4, 5, \ldots$, from the equations in Table III. We can use the same table to find the sum of (19) when $n = 9, 25, 49, \ldots$; but then we have also to use the equation

$$\frac{3}{\pi} = 1-24\left(\frac{1}{e^{2\pi}-1}+\frac{2}{e^{4\pi}-1}+\frac{3}{e^{6\pi}-1}+\cdots\right),$$

which is got by putting $k = k' = 1/\sqrt{2}$ and $n = 1$ in (18).

Similarly we can find the sum of (19) when $n = 21, 33, 57, 93, \ldots$, by combining the values of $f(3)$ and $f(7)$, $f(3)$ and $f(11)$, and so on, obtained from Table III.

11. The errors in (22) and (24) being about

$$8\pi e^{-\pi\sqrt{n}}(\pi\sqrt{n}-3), \quad 24\pi(10\pi\sqrt{n}-31)e^{-2\pi\sqrt{n}},$$

we cannot expect a high degree of approximation for small values of n. Thus, if we put $n = 7, 9, 16$, and 25 in (24), we get

$$\frac{19}{16}\sqrt{7} = 3{\cdot}14180\ldots,$$

$$\frac{7}{3}\left(1+\frac{\sqrt{3}}{5}\right) = 3{\cdot}14162\ldots,$$

$$\frac{99}{80}\left(\frac{7}{7-3\sqrt{2}}\right) = 3{\cdot}14159274\ldots,$$

$$\frac{63}{25}\left(\frac{17+15\sqrt{5}}{7+15\sqrt{5}}\right) = 3{\cdot}14159265380\ldots,$$

while $\pi = 3{\cdot}14159265358\ldots.$

But if we put $n = 25$ in (22), we get only

$$\frac{9}{5}+\sqrt{\left(\frac{9}{5}\right)} = 3{\cdot}14164\ldots.$$

12. Another curious approximation to π is

$$\left(9^2+\frac{19^2}{22}\right)^{\frac{1}{4}} = 3{\cdot}14159265262\ldots.$$

This value was obtained empirically, and it has no connection with the preceding theory.

The actual value of π, which I have used for purposes of calculation, is

$$\frac{355}{113}\left(1-\frac{{\cdot}0003}{3533}\right) = 3{\cdot}1415926535897943\ldots,$$

which is greater than π by about 10^{-16}. This is obtained by simply taking the reciprocal of $1-(113\pi/355)$.

In this connection it may be interesting to note the following simple geometrical constructions for π. The first merely gives the ordinary value 355/113. The second gives the value $(9^2+19^2/22)^{\frac{1}{4}}$ mentioned above.

(1) Let AB (Fig. 1) be a diameter of a circle whose centre is O.

Bisect AO at M and trisect OB at T.

Draw TP perpendicular to AB and meeting the circumference at P.

Draw a chord BQ equal to PT and join AQ.

Draw OS and TR parallel to BQ and meeting AQ at S and R respectively.

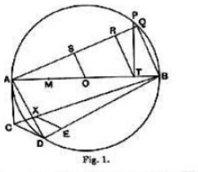

Fig. 1.

Draw a chord AD equal to AS and a tangent $AC = RS$.

Join BC, BD, and CD; cut off $BE = BM$, and draw EX, parallel to CD, meeting BC at X.

Then the square on BX is very nearly equal to the area of the circle, the error being less than a tenth of an inch when the diameter is 40 miles long.

(2) Let AB (Fig. 2) be a diameter of a circle whose centre is O.

Bisect the arc ACB at C and trisect AO at T.

Join BC and cut off from it CM and MN equal to AT.

Join AM and AN and cut off from the latter AP equal to AM.

Through P draw PQ parallel to MN and meeting AM at Q.

Join OQ and through T draw TR, parallel to OQ, and meeting AQ at R.

Draw AS perpendicular to AO and equal to AR, and join OS.

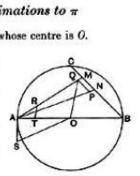

Fig. 2.

Then the mean proportional between OS and OB will be very nearly equal to a sixth of the circumference, the error being less than a twelfth of an inch when the diameter is 8000 miles long.

13. I shall conclude this paper by giving a few series for $1/\pi$.

It is known that, when $k \leqslant 1/\sqrt{2}$,

$$\left(\frac{2K}{\pi}\right)^2 = 1+\left(\frac{1}{2}\right)^3(2kk')^2+\left(\frac{1.3}{2.4}\right)^3(2kk')^4+\ldots. \ \ldots\ldots\ldots(25)$$

Hence we have

$$q^{\frac{1}{3}}(1-q^2)^4(1-q^4)^4(1-q^6)^4\ldots$$
$$= \left(\frac{1}{4}kk'\right)^{\frac{2}{3}}\left\{1+\left(\frac{1}{2}\right)^3(2kk')^2+\left(\frac{1.3}{2.4}\right)^3(2kk')^4+\ldots\right\}. \ \ldots\ldots(26)$$

Differentiating both sides in (26) logarithmically with respect to k, we can easily shew that

$$1-24\left(\frac{q^2}{1-q^2}+\frac{2q^4}{1-q^4}+\frac{3q^6}{1-q^6}+\ldots\right)$$
$$= (1-2k^2)\left\{1+4\left(\frac{1}{2}\right)^3(2kk')^2+7\left(\frac{1.3}{2.4}\right)^3(2kk')^4+\ldots\right\}. \ \ldots(27)$$

But it follows from (19) that, when $q = e^{-\pi\sqrt{n}}$, n being a rational number, the left-hand side of (27) can be expressed in the form

$$A\left(\frac{2K}{\pi}\right)^2+\frac{B}{\pi},$$

where A and B are algebraic numbers expressible by surds. Combining (25) and (27) in such a way as to eliminate the term $(2K/\pi)^2$, we are left with a series for $1/\pi$. Thus, for example,

$$\frac{4}{\pi} = 1+\frac{7}{4}\left(\frac{1}{2}\right)^3+\frac{13}{4^2}\left(\frac{1.3}{2.4}\right)^3+\frac{19}{4^3}\left(\frac{1.3.5}{2.4.6}\right)^3+\ldots,$$
$$(q = e^{-\pi\sqrt{3}},\ 2kk' = \tfrac{1}{2}). \ \ldots\ldots\ldots\ldots\ldots(28)$$

$$\frac{16}{\pi} = 5+\frac{47}{64}\left(\frac{1}{2}\right)^3+\frac{89}{64^2}\left(\frac{1.3}{2.4}\right)^3+\frac{131}{64^3}\left(\frac{1.3.5}{2.4.6}\right)^3+\ldots,$$
$$(q = e^{-\pi\sqrt{7}},\ 2kk' = \tfrac{1}{8}). \ \ldots\ldots\ldots\ldots\ldots(29)$$

$$\frac{32}{\pi}=(5\sqrt5-1)+\frac{47\sqrt5+29}{64}\left(\frac12\right)^3\left(\frac{\sqrt5-1}{2}\right)^8$$
$$+\frac{89\sqrt5+59}{64^2}\left(\frac{1.3}{2.4}\right)^3\left(\frac{\sqrt5-1}{2}\right)^{16}+\dots,$$
$$\left[q=e^{-\pi\sqrt{30}},\ 2kk'=\frac18\left(\frac{\sqrt5-1}{2}\right)\right];\ \dots\dots\dots(30)$$

here $5\sqrt5-1$, $47\sqrt5+29$, $89\sqrt5+59$, ... are in arithmetical progression.

14. The ordinary modular equations express the relations which hold between k and l when $nK'/K=L'/L$, or $q^n=Q$, where

$$q=e^{-\pi K'/K},\quad Q=e^{-\pi L'/L},$$
$$K=1+\left(\frac12\right)^2k^2+\left(\frac{1.3}{2.4}\right)^2k^4+\dots.$$

There are corresponding theories in which q is replaced by one or other of the functions

$$q_1=e^{-\pi K_1'\sqrt2/K_1},\quad q_2=e^{-2\pi K_2'/(K_2\sqrt3)},\quad q_3=e^{-2\pi K_3'/K_3},$$

where

$$K_1=1+\frac{1.3}{4^2}k^2+\frac{1.3.5.7}{4^2.8^2}k^4+\frac{1.3.5.7.9.11}{4^2.8^2.12^2}k^6+\dots,$$
$$K_2=1+\frac{1.2}{3^2}k^2+\frac{1.2.4.5}{3^2.6^2}k^4+\frac{1.2.4.5.7.8}{3^2.6^2.9^2}k^6+\dots,$$
$$K_3=1+\frac{1.5}{6^2}k^2+\frac{1.5.7.11}{6^2.12^2}k^4+\frac{1.5.7.11.13.17}{6^2.12^2.18^2}k^6+\dots.$$

From these theories we can deduce further series for $1/\pi$, such as

$$\frac{27}{4\pi}=2+17\,\frac12\,\frac13\,\frac23\left(\frac{2}{27}\right)$$
$$+32\,\frac{1.3}{2.4}\,\frac{1.4}{3.6}\,\frac{2.5}{3.6}\left(\frac{2}{27}\right)^2+\dots,\ \dots\dots(31)$$

$$\frac{15\sqrt3}{2\pi}=4+37\,\frac12\,\frac13\,\frac23\left(\frac{4}{125}\right)$$
$$+70\,\frac{1.3}{2.4}\,\frac{1.4}{3.6}\,\frac{2.5}{3.6}\left(\frac{4}{125}\right)^2+\dots,\ \dots\dots(32)$$

$$\frac{5\sqrt5}{2\pi\sqrt3}=1+12\,\frac12\,\frac16\,\frac56\left(\frac{4}{125}\right)$$
$$+23\,\frac{1.3}{2.4}\,\frac{1.7}{6.12}\,\frac{5.11}{6.12}\left(\frac{4}{125}\right)^2+\dots,\ \dots(33)$$

$$\frac{85\sqrt{85}}{18\pi\sqrt3}=8+141\,\frac12\,\frac16\,\frac56\left(\frac{4}{85}\right)^3$$
$$+274\,\frac{1.3}{2.4}\,\frac{1.7}{6.12}\,\frac{5.11}{6.12}\left(\frac{4}{85}\right)^6+\dots,\ \dots(34)$$

$$\frac4\pi=\frac32-\frac{23}{2^3}\,\frac12\,\frac{1.3}{4^2}+\frac{43}{2^5}\,\frac{1.3}{2.4}\,\frac{1.3.5.7}{4^2.8^2}-\dots,\ \dots\dots\dots\dots\dots(35)$$

$$\frac{4}{\pi\sqrt3}=\frac34-\frac{31}{3.4^3}\,\frac12\,\frac{1.3}{4^2}+\frac{59}{3^2.4^5}\,\frac{1.3}{2.4}\,\frac{1.3.5.7}{4^2.8^2}-\dots\ \dots\dots\dots(36)$$

$$\frac4\pi=\frac{23}{18}-\frac{283}{18^3}\,\frac12\,\frac{1.3}{4^2}+\frac{543}{18^5}\,\frac{1.3}{2.4}\,\frac{1.3.5.7}{4^2.8^2}-\dots,\ \dots\dots\dots(37)$$

$$\frac{4}{\pi\sqrt5}=\frac{41}{72}-\frac{685}{5.72^3}\,\frac12\,\frac{1.3}{4^2}+\frac{1329}{5^2.72^5}\,\frac{1.3}{2.4}\,\frac{1.3.5.7}{4^2.8^2}-\dots,\ \dots\dots(38)$$

$$\frac4\pi=\frac{1123}{882}-\frac{22583}{882^3}\,\frac12\,\frac{1.3}{4^2}+\frac{44043}{882^5}\,\frac{1.3}{2.4}\,\frac{1.3.5.7}{4^2.8^2}-\dots,\ \dots(39)$$

$$\frac{2\sqrt3}{\pi}=1+\frac99\,\frac12\,\frac{1.3}{4^2}+\frac{17}{9^2}\,\frac{1.3}{2.4}\,\frac{1.3.5.7}{4^2.8^2}+\dots,\ \dots\dots\dots\dots\dots(40)$$

$$\frac{1}{2\pi\sqrt2}=\frac19+\frac{11}{9^3}\,\frac12\,\frac{1.3}{4^2}+\frac{21}{9^5}\,\frac{1.3}{2.4}\,\frac{1.3.5.7}{4^2.8^2}+\dots,\ \dots\dots\dots\dots(41)$$

$$\frac{1}{3\pi\sqrt3}=\frac{3}{49}+\frac{43}{49^3}\,\frac12\,\frac{1.3}{4^2}+\frac{83}{49^5}\,\frac{1.3}{2.4}\,\frac{1.3.5.7}{4^2.8^2}+\dots\ \dots\dots\dots(42)$$

$$\frac{2}{\pi\sqrt{11}}=\frac{19}{99}+\frac{299}{99^3}\,\frac12\,\frac{1.3}{4^2}+\frac{579}{99^5}\,\frac{1.3}{2.4}\,\frac{1.3.5.7}{4^2.8^2}+\dots,\ \dots\dots\dots(43)$$

$$\frac{1}{2\pi\sqrt2}=\frac{1103}{99^2}+\frac{27493}{99^6}\,\frac12\,\frac{1.3}{4^2}+\frac{53883}{99^{10}}\,\frac{1.3}{2.4}\,\frac{1.3.5.7}{4^2.8^2}+\dots\ \dots(44)$$

In all these series the first factors in each term form an arithmetical progression; e.g. 2, 17, 32, 47, ..., in (31), and 4, 37, 70, 103, ..., in (32). The first two series belong to the theory of q_2, the next two to that of q_3, and the rest to that of q_1.

The last series (44) is extremely rapidly convergent. Thus, taking only the first term, we see that

$$\frac{1103}{99^2}=\cdot11253953678\dots,$$
$$\frac{1}{2\pi\sqrt2}=\cdot11253953951\dots.$$

15. In concluding this paper I have to remark that the series

$$1-24\left(\frac{q^2}{1-q^2}+\frac{2q^4}{1-q^4}+\frac{3q^6}{1-q^6}+\dots\right),$$

which has been discussed in §§ 8—13, is very closely connected with the perimeter of an ellipse whose eccentricity is k. For, if a and b be the semi-major and the semi-minor axes, it is known that

$$p=2\pi a\left\{1-\frac{1}{2^2}k^2-\frac{1^2.3}{2^2.4^2}k^4-\frac{1^2.3^2.5}{2^2.4^2.6^2}k^6-\dots\right\},\ \dots\dots\dots(45)$$

where p is the perimeter and k the eccentricity. It can easily be seen from (45) that

$$p=4ak'^2\left\{K+k\frac{dK}{dk}\right\}.\ \dots\dots\dots\dots\dots\dots\dots\dots(46)$$

But, taking the equation

$$q^{\frac1{12}}(1-q^2)(1-q^4)(1-q^6)\dots=(2kk')^{\frac16}\sqrt{(K/\pi)},$$

and differentiating both sides logarithmically with respect to k, and combining the result with (46) in such a way as to eliminate dK/dk, we can shew that

$$p=\frac{4a}{3K}\left[K^2(1+k'^2)+(\tfrac12\pi)^2\left\{1-24\left(\frac{q^2}{1-q^2}+\frac{2q^4}{1-q^4}+\dots\right)\right\}\right].\ (47)$$

But we have shewn already that the right-hand side of (47) can be expressed in terms of K if $q=e^{-\pi\sqrt n}$, where n is any rational number. It can also be shewn that K can be expressed in terms of Γ-functions if q be of the forms $e^{-\pi n}$, $e^{-\pi n\sqrt2}$ and $e^{-\pi n\sqrt3}$, where n is rational. Thus, for example, we have

$$\left.\begin{aligned}
&k=\sin\frac\pi4,\quad q=e^{-\pi},\\
&\qquad p=a\sqrt{\left(\frac\pi2\right)}\left\{\frac{\Gamma(\frac14)}{\Gamma(\frac34)}+\frac{\Gamma(\frac34)}{\Gamma(\frac54)}\right\},\\
&k=\tan\frac\pi8,\quad q=e^{-\pi\sqrt2},\\
&\qquad p=a\sqrt{\left(\frac\pi4\right)}\left\{\frac{\Gamma(\frac18)}{\Gamma(\frac58)}+\frac{\Gamma(\frac58)}{\Gamma(\frac98)}\right\},\\
&k=\sin\frac{\pi}{12},\quad q=e^{-\pi\sqrt3},\\
&\qquad p=a\sqrt{\left(\frac{\pi}{\sqrt3}\right)}\left\{\left(1+\frac{1}{\sqrt3}\right)\frac{\Gamma(\frac13)}{\Gamma(\frac56)}+2\frac{\Gamma(\frac56)}{\Gamma(\frac13)}\right\},\\
&\frac ba=\tan^2\frac\pi8,\quad q=e^{-2\pi},\\
&\qquad p=(a+b)\sqrt{\left(\frac\pi2\right)}\left\{\frac12\frac{\Gamma(\frac14)}{\Gamma(\frac34)}+\frac{\Gamma(\frac34)}{\Gamma(\frac14)}\right\},
\end{aligned}\right\}\ \dots\dots(48)$$

and so on.

16. The following approximations for p were obtained empirically:

$$p=\pi\left[3(a+b)-\sqrt{\{(a+3b)(3a+b)\}}+\epsilon\right],\ \dots\dots\dots\dots\dots(49)$$

where ϵ is about $ak^{12}/1048576$;

$$p=\pi\left\{(a+b)+\frac{3(a-b)^2}{10(a+b)+\sqrt{(a^2+14ab+b^2)}}+\epsilon\right\},\ \dots\dots\dots(50)$$

where ϵ is about $3ak^{20}/68719476736$.

Squaring the Circle

https://www.imsc.res.in/~rao/ramanujan/CamUnivCpapers/Cpaper5/page1.htm

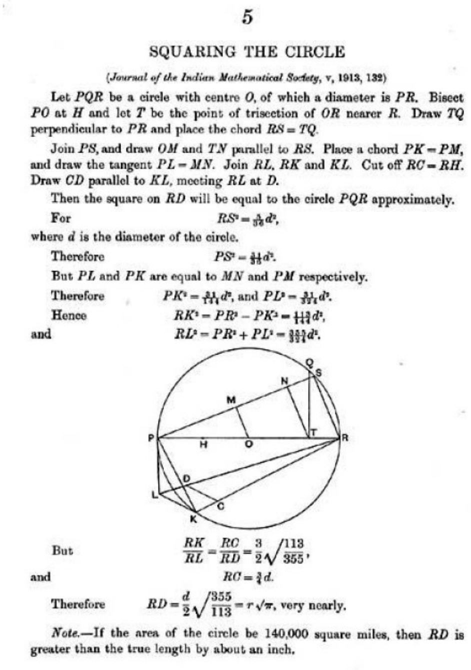

5

SQUARING THE CIRCLE

(*Journal of the Indian Mathematical Society*, v, 1913, 132)

Let PQR be a circle with centre O, of which a diameter is PR. Bisect PO at H and let T be the point of trisection of OR nearer R. Draw TQ perpendicular to PR and place the chord $RS = TQ$.

Join PS, and draw OM and TN parallel to RS. Place a chord $PK = PM$, and draw the tangent $PL = MN$. Join RL, RK and KL. Cut off $RC = RH$. Draw CD parallel to KL, meeting RL at D.

Then the square on RD will be equal to the circle PQR approximately.

For $$RS^2 = \tfrac{5}{36} d^2,$$
where d is the diameter of the circle.

Therefore $$PS^2 = \tfrac{31}{36} d^2.$$

But PL and PK are equal to MN and PM respectively.

Therefore $$PK^2 = \tfrac{31}{144} d^2, \text{ and } PL^2 = \tfrac{31}{324} d^2.$$

Hence $$RK^2 = PR^2 - PK^2 = \tfrac{113}{144} d^2,$$

and $$RL^2 = PR^2 + PL^2 = \tfrac{355}{324} d^2.$$

But $$\frac{RK}{RL} = \frac{RC}{RD} = \frac{3}{2}\sqrt{\frac{113}{355}},$$

and $$RC = \tfrac{3}{4} d.$$

Therefore $$RD = \frac{d}{2}\sqrt{\frac{355}{113}} = r\sqrt{\pi}, \text{ very nearly.}$$

Note.—If the area of the circle be 140,000 square miles, then RD is greater than the true length by about an inch.

The End

20210906

)
1 %% %F ! % ! ! D! %

1 %% %F ! % ! ! D! % ! C 7 1!1 % #
.2 = !

1 %% %F ! % ! ! D! % ! C 7 1!1 % 1!1 9%1!
.2 +11 4 !
! G
1 %% %F ! % ! ! D! % ! C 7 1!1 % 1!1 9%1!

20210905m (20210905)

Data-NovoLivro-018-2021/
Ramanujam-Works/
KSR chap3.pdf

3. Ramanujan's Notebooks

Pag. 7

"Modular equations and approximations to π"

π like:

$$\frac{63}{25}\,\frac{17+15\sqrt{5}}{7+15\sqrt{5}} = 3.1415926538\ldots$$

etc.

Source:

www.imsc.res.in/~rao/
ramanujan/images/
KSR Chap3.pdf (KSRchap3.pdf)

Obs: Prof. Rao (Srinivasa)
Institute of Mathematics Sciences
Chennai - 600 020, India

Source: www.imsc.res.in/~rao/

ramanujan/contentindex.html

Import:

imsc.res.in/~rao/ramanujan/
CamUnivCpapers/collectedright1.htm

5. Squaring the circle
(Journal of the Indian Mathematical
Society, 5 (1913), 132

Source:

imsc.res.in/~rao/ramanujan/
CamUnivCpapers/Cpaper5/
page1.htm

6. Modular equations and
approximations to π
(Quarterly Journal of Mathematics,
45 (1914), 350-372.)

Source:

www.imsc.res.in/~rao/ramanujan/
CamUnivCpapers/Cpaper6/
page1.htm → pag17.htm

data / /
S T Q Q S S D

other:
www.imsc.res.in/~rao/
ramanujan/collectedpapers/
collectedright1.htm

obs: Journal of the Indian Mathematical Society.

/collectedpaper/squaring/squaring.htm

/collectedpaper/modularEqn/
modularEqn1.htm
Eqn23.htm

Término: 06/09/2021 00:00h

Cópia dos arquivos:
Data-NovoVivo-018-2021/
Books-018-2021/Important/
Ramanujan/
obs: 44 files

&) &
! *+,-* * *. ! * &
H!@ G @# @1! ! I @@@J KH 2! 7 L D@M L !D@21 9L"$& 1#

2 *+,-* * *. ! * & / 0 N! # 3
! ! ! 1#

+,- * *. ! * & 0 1 . 2 1 #
.4 # .4 # C 7 0 1#

+,- * *. ! * & 5 ! O / 2
- ! 5 # 8 0 > 5 0 > L # ! 2 # C 7 0 1#

+,- * *. ! * & /! 0 ; 8 / # ,!
8 6 7 ; ! 0 ! ! ,! 2 1#

+,- * *. ! * & "' ' " 1#

→2) Physics of Spin-Orbit-Coupled Oxides/
0199602026.pdf

→3) Problems in Quantum Field Theory
with Fully-Worked Solutions/
Problems_in_Quantum_Field_Theory.pdf

→4) The Golden Ratio The Divine Beauty
of Mathematics PDF/
1536134546.pdf

→5) Leibniz an the invention of
Mathematical Transcendence/
3515120823.pdf

Other:

Books-018-2021/Important/Ramanujan/
Ramanujan-Prof. Steno Musich www.
stenomusich.com.pdf

- Books-018-2021/20210906/
Kurt Godel - Works - 5 Volumes/

- Data-NovoVivo-018-2021/Books-018-2021/
Important/

.2 ;..E A*;..E *; L 1 1! *E / #
> L ' *

20210909
Books cont.

- Books-018-2021/Advanced Engineering Mathematics...../
0471488852.pdf
Pág. 35(54) 1.6 Orthogonal Trajectories.

An important type of problem in physics (or geometry) is to find a family of curves that intersect a given family of curves at right angles..

- Books-018-2021/20210909/
Nanohertz Gravitational Wave Astronomy/
Nanohertz Gravitational Wave Astronomy.pdf
Pág. 12(25) 2.2.2.
Waves from Geometry.
Pág 14(27) Figure 2.2
The periodic deformation of spacetime caused a GW is illustrated for the two distinct polarizations permitted

within GR.

BLOGGER:

1) stenomunich.com →
https//www.blogger.com/blog/
stats/day/
41 06.987.220.885697.888
Q Q T B M

2) matematics.org →
770366.718.644.342.674
Q T B M

3) mathematics.pictures →
1.878.416.941.151.684.229
Q Q T B M

4) mathematics.page →
9.161.514.591.326.450.059
Q Q T B M

---

christopher.carl.munich@gmail.com
donaldson.com.br

) "' '& 9 . ! 8 !D

1.6 Orthogonal Trajectories. *Optional*

An important type of problem in physics or geometry is to find a family of curves that intersect a given family of curves at right angles. The new curves are called **orthogonal trajectories** of the given curves (and conversely). Examples are curves of equal temperature (*isotherms*) and curves of heat flow, curves of equal altitude (*contour lines*) on a map and curves of steepest descent on that map, curves of equal potential (*equipotential curves*, curves of equal voltage—the concentric circles in Fig. 22), and curves of electric force (the straight radial segments in Fig. 22).

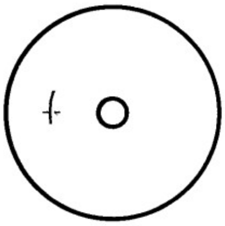

Fig. 22. Equipotential lines and curves of electric force (dashed) between two concentric (black) circles (cylinders in space)

Here the **angle of intersection** between two curves is defined to be the angle between the tangents of the curves at the intersection point. *Orthogonal* is another word for *perpendicular*.

In many cases orthogonal trajectories can be found by using ODEs, as follows. Let

$$G(x, y, c) = 0 \tag{1}$$

be a given family of curves in the xy-plane, where each curve is specified by some value of c. This is called a **one-parameter family of curves**, and c is called the *parameter* of the family. For instance, a one-parameter family of quadratic parabolas is given by (Fig. 23)

$$y = cx^2 \qquad \text{or, written as in (1),} \qquad G(x, y, c) = y - cx^2 = 0.$$

Step 1. Find an ODE for which the given family is a general solution. Of course, this ODE must no longer contain the parameter c. In our example we solve algebraically for c and then differentiate and simplify; thus,

$$\frac{y}{x^2} = c, \qquad \frac{y'x^2 - 2xy}{x^4} = 0,$$

hence

$$y' = \frac{2y}{x}.$$

The last of these equations is the ODE of the given family of curves. It is of the form

$$y' = f(x, y). \tag{2}$$

Step 2. Write down the ODE of the orthogonal trajectories, that is, the ODE whose general solution gives the orthogonal trajectories of the given curves. This ODE is

$$\tilde{y}' = -\frac{1}{f(x, \tilde{y})} \tag{3}$$

with the same f as in (2). Why? Well, a given curve passing through a point (x_0, y_0) has slope $f(x_0, y_0)$ at that point, by (2). The trajectory through (x_0, y_0) has slope $-1/f(x_0, y_0)$ by (3). The product of these slopes is -1, as we see. From calculus it is known that this is the condition for orthogonality (perpendicularity) of two straight lines (the tangents at (x_0, y_0)), hence of the curve and its orthogonal trajectory at (x_0, y_0).

Step 3. Solve (3).

For our parabolas $y = cx^2$ we have $y' = 2y/x$. Hence their orthogonal trajectories are obtained from $\tilde{y}' = -x/2\tilde{y}$ or $2\tilde{y}\tilde{y}' + x = 0$. By integration, $\tilde{y}^2 + \frac{1}{2}x^2 = c^*$. These are the ellipses in Fig. 23 with semi-axes $\sqrt{2c^*}$ and $\sqrt{c^*}$. Here, $c^* > 0$ because $c^* = 0$ gives just the origin, and $c^* < 0$ gives no real solution at all.

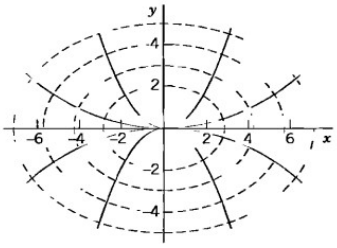

Fig. 23. Parabolas and orthogonal trajectories (ellipses) in the text

---

[illegible]

1–12 ORTHOGONAL TRAJECTORIES

Sketch or graph some of the given curves. Guess what their orthogonal trajectories may look like. Find these trajectories.

(Show the details of your work.)

1. $y = 4x + c$ **2.** $y = c/x$

3. $y = cx$ **4.** $y^2 = 2x^2 + c$

5. $x^2y = c$ **6.** $y = ce^{-3x}$

7. $y = ce^{x^2/2}$ **8.** $x^2 - y^2 = c$

9. $4x^2 + y^2 = c$ **10.** $x = c\sqrt{y}$

11. $x = ce^{y/4}$ **12.** $x^2 + (y - c)^2 = c^2$

13–15 OTHER FORMS OF THE ODEs (2) AND (3)

13. **(y as independent variable)** Show that (3) may be written $dx/d\tilde{y} = -f(x, \tilde{y})$. Use this form to find the orthogonal trajectories of $y = 2x + ce^{-x}$.

14. **(Family $g(x, y) = c$)** Show that if a family is given as $g(x, y) = c$, then the orthogonal trajectories can be obtained from the following ODE, and use the latter to solve Prob. 6 written in the form $g(x, y) = c$.

$$\frac{d\tilde{y}}{dx} = \frac{\partial g/\partial \tilde{y}}{\partial g/\partial x}$$

15. **(Cauchy–Riemann equations)** Show that for a family $u(x, y) = c = const$ the orthogonal trajectories $v(x, y) = c^* = const$ can be obtained from the following *Cauchy–Riemann equations* (which are basic in complex analysis in Chap. 13) and use them to find the orthogonal trajectories of $e^x \sin y = const$. (Here, subscripts denote partial derivatives.)

$$u_x = v_y, \qquad u_y = -v_x$$

16–20 **APPLICATIONS**

16. **(Fluid flow)** Suppose that the *streamlines* of the flow (paths of the particles of the fluid) in Fig. 24 are $\Psi(x, y) = xy = const$. Find their orthogonal trajectories (called **equipotential lines,** for reasons given in Sec. 18.4).

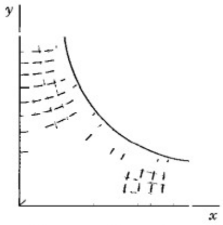

Fig. 24. Flow in a channel in Problem 16

17. **(Electric field)** Let the electric equipotential lines (curves of constant potential) between two concentric cylinders (Fig. 22) be given by $u(x, y) = x^2 + y^2 = c$. Use the method in the text to find their orthogonal trajectories (the curves of electric force).

18. **(Electric field)** The lines of electric force of two opposite charges of the same strength at $(-1, 0)$ and $(1, 0)$ are the circles through $(-1, 0)$ and $(1, 0)$. Show that these circles are given by $x^2 + (y - c)^2 = 1 + c^2$. Show that the **equipotential lines** (orthogonal trajectories of those circles) are the circles given by $(x + c^*)^2 + \tilde{y}^2 = c^{*2} - 1$ (dashed in Fig. 25).

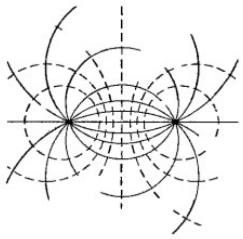

Fig. 25. Electric field in Problem 18

19. **(Temperature field)** Let the **isotherms** (curves of constant temperature) in a body in the upper half-plane $y > 0$ be given by $4x^2 + 9y^2 = c$. Find the orthogonal trajectories (the curves along which heat will flow in regions filled with heat-conducting material and free of heat sources or heat sinks).

20. **TEAM PROJECT. Conic Sections.** **(A)** State the main steps of the present method of obtaining orthogonal trajectories.

(B) Find conditions under which the orthogonal trajectories of families of ellipses $x^2/a^2 + y^2/b^2 = c$ are again conic sections. Illustrate your result graphically by sketches or by using your CAS. What happens if $a \to 0$? If $b \to 0$?

(C) Investigate families of hyperbolas $x^2/a^2 - y^2/b^2 = c$ in a similar fashion.

(D) Can you find more complicated curves for which you get ODEs that you can solve? Give it a try.

1.7 Existence and Uniqueness of Solutions

The initial value problem

$$|y'| + |y| = 0, \qquad y(0) = 1$$

has no solution because $y = 0$ (that is, $y(x) = 0$ for all x) is the only solution of the ODE. The initial value problem

$$y' = 2x, \qquad y(0) = 1$$

2 *+,-* * *. ! * ' 1 , 8!0 M! A
/ !7 ! ! >!7 + 0 , 1#
) '

2.2.2 Waves from Geometry

Let's dig into some illuminating mathematics to grasp what these gravitational waves (GWs) really are. General Relativity is enshrined within the *Einstein Field Equations*,

$$G_{\mu\nu} := R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} = 8\pi T_{\mu\nu}, \tag{2.1}$$

where $G_{\mu\nu}$ is the Einstein tensor, $R_{\mu\nu} := R^{\lambda}_{\mu\lambda\nu}$ is the Ricci tensor derived from the Riemann curvature tensor, $R = g^{\mu\nu}R_{\mu\nu}$ is the Ricci scalar, $g_{\mu\nu}$ is the metric, $T_{\mu\nu}$ is the stress-energy tensor, and we assume natural units such that $G = c = 1$.

We consider the linearized treatment of GWs that Einstein originally studied (11; 12). In the following, Greek indices refer to 4-D spacetime coordinates, while Roman indices refer to 3-D spatial coordinates. The derivation presented here closely follows the treatment in Maggiore (29). We start with spatially-flat 4-D Minkowski spacetime, with $\eta_{\mu\nu} = \mathrm{diag}(-1, 1, 1, 1)$, which upon adding a perturbation results in

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}, \tag{2.2}$$

where $h_{\mu\nu}$ is a tensor metric perturbation quantity that we treat only in the weak field case, $|h_{\mu\nu}| \ll 1$. This metric perturbation is assumed to be so small that the usual index raising and lowering operations can be performed with the Minkowski metric. Expanding the Einstein tensor to linear order in $h_{\mu\nu}$ in this perturbed spacetime gives,

$$G_{\mu\nu} = \frac{1}{2}\left(\partial_\mu\partial^\alpha h_{\alpha\nu} + \partial_\nu\partial^\alpha h_{\alpha\mu} - \partial_\mu\partial_\nu h - \Box h_{\mu\nu} + \eta_{\mu\nu}\Box h - \eta_{\mu\nu}\partial^\alpha\partial^\beta h_{\alpha\beta}\right), \tag{2.3}$$

where $\partial_\mu \equiv \partial/\partial x^\mu$; $\partial^\mu \equiv \eta^{\mu\nu}\partial_\nu$; $h = \eta^{\mu\nu}h_{\mu\nu}$ is the trace of $h_{\mu\nu}$; and $\Box = \eta^{\mu\nu}\partial_\mu\partial_\nu$ is the flat space d'Alembertian operator. This is a bit cumbersome, so we change variables in an effort to tidy this up. We define the trace-reversed perturbation, $\bar{h}_{\mu\nu} = h_{\mu\nu} - n_{\mu\nu}h/2$, which also implies that $\bar{h} = \eta^{\mu\nu}\bar{h}_{\mu\nu} = h - 2h = -h$, such that $h_{\mu\nu} = \bar{h}_{\mu\nu} - n_{\mu\nu}\bar{h}/2$. The Einstein tensor then becomes

$$G_{\mu\nu} = \frac{1}{2}\left(\partial_\mu\partial^\alpha \bar{h}_{\alpha\nu} + \partial_\nu\partial^\alpha \bar{h}_{\alpha\mu} - \Box\bar{h}_{\mu\nu} - \eta_{\mu\nu}\partial^\alpha\partial^\beta \bar{h}_{\alpha\beta}\right). \tag{2.4}$$

Similar to electromagnetism, we can ditch spurious degrees of freedom from our equations by choosing an appropriate gauge. If we consider the coordinate transformation $x^\alpha \mapsto x^\alpha + \xi^\alpha$, the transformation of the metric perturbation to first order is $h_{\mu\nu} \mapsto h_{\mu\nu} - (\partial_\mu\xi_\nu + \partial_\nu\xi_\mu)$. Asserting $\partial_\mu\xi_\nu$ to be of the same order as $|h_{\mu\nu}|$, the transformed metric perturbation retains the condition $|h_{\alpha\beta}| \ll 1$. This symmetry allows us to choose the *Lorenz gauge* (sometimes referred to as the *De Donder gauge*, or *Hilbert gauge*, or *harmonic gauge*) where $\partial^\nu\bar{h}_{\mu\nu} = 0$, such that the Einstein tensor reduces to the much more compact

$$G_{\mu\nu} = -\frac{1}{2}\Box\bar{h}_{\mu\nu}. \tag{2.5}$$

Note that in choosing the Lorenz gauge we have imposed 4 conditions that reduce the number of degrees of freedom in the symmetric 4×4 $h_{\mu\nu}$ matrix from 10 down to 6. Finally, the field equations reduce to

$$\Box \bar{h}_{\mu\nu} = -16\pi T_{\mu\nu}, \tag{2.6}$$

which can be generally solved using the radiative Green's function:

$$\bar{h}_{\mu\nu}(\vec{x}, t) = 4 \int d^4x' \frac{T_{\mu\nu}(\vec{x}', t)}{|\vec{x} - \vec{x}'|}. \tag{2.7}$$

Examining Eq. 2.6 far from the source gives $\Box \bar{h}_{\mu\nu} = 0$, whose solution is clearly wave-like with propagation speed c. We might think that all components of the metric perturbation are radiative. This is a gauge artefact (13), where, in general, we can split the metric perturbation into (i) gauge degrees of freedom, (ii) physical, radiative degrees of freedom, and (iii) physical, non-radiative degrees of freedom. The Lorenz gauge is preserved under a coordinate transformation $x^\mu \mapsto x^\mu + \xi^\mu$, provided that $\Box \xi_\mu = 0$, which also implies that $\Box \xi_{\mu\nu} \equiv \Box(\partial_\mu \xi_\nu + \partial_\nu \xi_\mu - \eta_{\mu\nu} \partial_\rho \xi_\rho) = 0$. This means that the 6 independent components of $\bar{h}_{\mu\nu}$ can have $\xi_{\mu\nu}$ subtracted, which depend on 4 independent functions ξ_μ, thereby distilling the GW information down to 2 independent components of the metric perturbation. The conditions imposed on $\bar{h}_{\mu\nu}$ are such that $\bar{h} = 0$ (therefore $h_{\mu\nu} = \bar{h}_{\mu\nu}$), and $h^{0i} = 0$, which lead to the following properties of the metric perturbation that define the *transverse-traceless* (TT) gauge:

$$h^{0\mu} = 0, \quad h^i_i = 0, \quad \partial^j h_{ij} = 0. \tag{2.8}$$

Choosing a coordinate system such that we have a plane GW propagating in the z-direction in a vacuum, we can write the solution to $\Box \bar{h}_{\mu\nu} = 0$ as

$$h^{\mathrm{TT}}_{\mu\nu}(t, z) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & h_+ & h_\times & 0 \\ 0 & h_\times & -h_+ & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \cos\left[\omega(t - z)\right], \tag{2.9}$$

where h_+ and $h_\times$ are the amplitudes of the two distinct polarizations of GWs permitted within general relativity, denoted as "plus" (+) and "cross" (×) modes for how they tidally deform a circular ring of test masses in the plane perpendicular to the direction of propagation.

The tidal deformation of spacetime caused by a GW as it propagates is at the heart of all ground-based (LIGO-Virgo-KAGRA), Galactic-scale (pulsar-timing arrays), and planned space-borne (LISA) detection efforts. Consider the *coordinate* separation of two spatially-separated test masses as a GW sweeps by. The test masses fall along geodesics of the perturbed spacetime, such that in the weak-field regime, and to first order in amplitude of the wave, the coordinate separation of the test masses remains unchanged. However, the

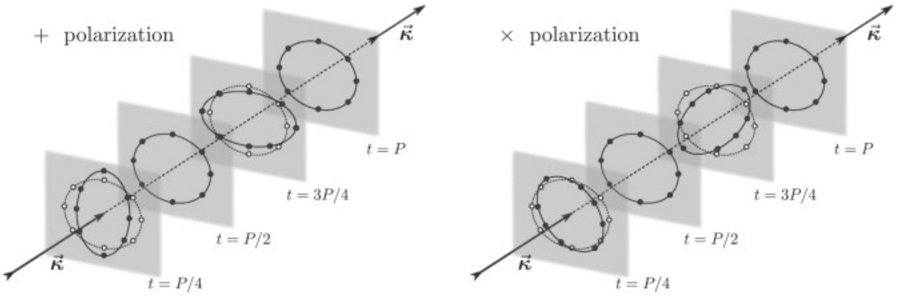

Figure 2.2: The periodic deformation of spacetime caused by a GW is illustrated for the two distinct polarizations permitted within GR. The influence of the GW is entirely in the plane perpendicular to the direction of propagation. Figure reproduced with permission from Ref. (30).

proper separation between the test masses *is* affected, and depends on the wave properties. The fractional change in the proper distance between two test masses separated by $\Delta x = L$ on the x-axis of a coordinate system due to the passage of a GW is given by $\delta L/L \simeq h_{xx}/2$, leading to a definition of the GW amplitude as the *strain*. For a periodic signal such that $h_{xx}(t, z = 0) = h_+ \cos \omega t$, we see that this proper distance separation oscillates according to $\delta \ddot{L} = -2\pi^2 f^2 L h_+ \cos \omega t$. We can see from Eq. 2.9 that the $+$-polarisation will lengthen distances along the x-axis while simultaneously contracting distances along the y-axis. The influence of the $h_\times$ mode is similar, but rotated by $\pi/4$ degrees counter-clockwise in the xy-plane. This tidal deformation in the plane perpendicular to the direction of propagation is illustrated for a full wave-cycle in Fig. 2.2.

2.2.2.1 The Quadrupole Formula

In the weak-field ($v \ll 1$) limit, far from a source, the leading order contribution to the solution of Eq. 2.6 is a function of the accelerating quadrupole moment of the source's mass distribution. Stress-energy conservation implies that the monopole makes no contribution due to conservation of the system's total energy, while the dipole moment makes no contribution due to conservation of momentum of the system's center-of-mass. Therefore the *quadrupole radiation formula* for the spatial components of the metric perturbation is (12),

$$h_{ij} = \frac{2}{r}\Lambda_{ij,kl}\frac{d^2 Q_{kl}}{dt^2}, \tag{2.10}$$

where r is the distance to the source, and Q_{kl} is the *reduced* quadrupole moment of $\rho(t, \vec{x})$ (the source's mass density), which is defined as,

$$Q_{ij} = \int d^3x\, \rho(t, \vec{x}) \left(x_i x_j - \frac{1}{3} r^2 \delta_{ij} \right), \tag{2.11}$$

and $\Lambda_{ik,jl}$ is the *Lambda tensor* that projects the metric perturbation in the Lorenz gauge into the TT gauge (see Ref. (29)).

The most relevant source system for us is a binary composed of compact objects (COs), whether white dwarfs, neutron stars, or black holes. Let us consider that each CO (of mass M) orbits one another at a distance $R(t)$ from their common center of mass (assumed to be far enough apart that we can ignore any tidal disruption effects) with slowly varying angular velocity $\omega(t)$. The COs are assumed to be moving at non-relativistic speeds to simplify the calculation of the quadrupole moment tensor. Kepler's third law gives us $\omega^2 = M/4R^3$, such that the total energy of the system is $E = -M^2/4R$. The orbital geometry is defined such that it lies in the $x-y$ plane with the origin coinciding with the system's center of mass. Each CO's coordinates at $t = 0$ are thus

$$x_1 = -x_2 = R\cos\omega t, \quad y_1 = -y_2 = R\sin\omega t, \quad z_1 = z_2 = 0. \tag{2.12}$$

Evaluating the second time-derivative of the reduced quadrupole mass-moment, and plugging into Eq. 2.10, the radiative components of the GWs from this system are

$$h_{\mu\nu} = -\frac{8MR^2\omega^2}{r}\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \cos 2\omega t & \sin 2\omega t & 0 \\ 0 & \sin 2\omega t & -\cos 2\omega t & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{2.13}$$

where the GW frequency is *twice* the binary orbital frequency due to the quadrupolar nature of the emission. Note that the distance to the source is directly encoded in the amplitude of this leading-order radiative term.

Just like electromagnetic waves, or even waves on a string, the energy density in GWs is $\propto \omega^2 h^2$, where h is the amplitude of the wave. The amplitude is already proportional to a second time derivative of the reduced quadrupole moment, where the the factor of ω implies an additional time derivative. Therefore, based on simple scaling arguments, the energy density in GWs should be proportional to a quadratic combination of the third time-derivative of the quadrupole mass moment. In a background spacetime that is approximately flat far from the source, the Issacson stress-energy tensor of a GW is

$$T_{\mu\nu} = \frac{1}{32\pi}\langle \partial_\mu \hat{h}_{\alpha\beta} \partial_\nu \hat{h}^{\alpha\beta} \rangle, \tag{2.14}$$

where $\langle\cdot\rangle$ denotes an average over several wave cycles. Evaluating the energy-flux elements of Eq. 2.14 in the quadrupole approximation, and integrating over the sphere, the GW luminosity of the source system

$$L_{\mathrm{GW}} \equiv \frac{dE_{\mathrm{GW}}}{dt} = \frac{1}{5}\left\langle \frac{d^3 Q_{ij}}{dt^3}\frac{d^3 Q^{ij}}{dt^3} \right\rangle, \tag{2.15}$$

which, for a CO binary gives

$$L_{\mathrm{GW}} = \frac{128}{5} M^2 R^4 \omega^6 = \frac{128}{5} 4^{1/3} \left(\frac{\pi M}{P}\right)^{10/3} \tag{2.16}$$

By equating the loss in the binary's orbital energy with GW emission, we can get a qualitative understanding of the frequency and strain evolution as the binary inspirals toward an eventual merger. The rate of change of the binary's orbital period is $dP/dt \propto P^{-5/3}$, such that the GW frequency evolution is $df_{\mathrm{GW}}/dt \propto f_{\mathrm{GW}}^{11/3}$. The strain amplitude is $h \propto R^2 f_{\mathrm{GW}}^2 \propto f_{\mathrm{GW}}^{2/3}$. Hence the orbital evolution, as driven by GW emission, causes the strain amplitude to increase in tune with the frequency, with both said to be "chirping" as the binary progresses toward merger.

2.3 STOCHASTIC GRAVITATIONAL WAVE BACKGROUNDS

Imagine yourself at a crowded party. All the guests are mingling and chatting to create a background hum of the usual social small-talk. You perk up your ears to attempt to hear whether your friend is stuck in an awkward conversation across the room, but alas, all you can hear is that damned background hum. Nevertheless, if we brought the lens of statistical inference to this party banter, there would be interesting information – the distribution of laughter could tell us whether this is a party worth sticking around at, and by localizing regions of high laughter one may be able to zero-in on the life and soul of this get-together. Occasionally, someone who may have had a bit too much to drink may forget to regulate their volume, creating a distinct sentence that can be heard above the fray.

Now imagine translating this scenario to GW signals. A cosmological population of systems emitting GWs of a similar frequency and comparable amplitude may not be able to be individually resolved by a detector. In such a scenario, the signals sum incoherently to produce a stochastic background of GWs (SGWB). "Stochastic" here formally means that we treat this background as a random process that is only studiable in terms of its statistical properties. As a function of time, the SGWB may look like random fluctuations without any discernible information, but if we dig deeper and look at its spectral information then it begins to reveal its secrets. It may have more power at lower frequencies than at higher frequencies, indicating that it varies on longer timescales. It may have a sharp uptick at a few frequencies, potentially revealing a few of those loud "voices" at the party. Let us now see how we describe a SGWB, and how it manifests as signals in our detectors. Much of the material in the following subsections has been adapted from the excellent treatment in Romano & Cornish (31).

20210910 Donaldson.

5511997076689

⇒ 4894-6340

---

osdoptics.com —

"relógios de rede óptica"

15 bilhões de anos para perder 1 seg.

"É um instrumento que permite sondar o tecido básico do espaço-tempo no Universo".

1967 – 1 seg =
9.192.631.770 oscilações de um átomo de Césio.

Novo Relógio Atômico: Estrôncio

Obs: Se os átomos caíssem devido à gravidade ou se movessem, haveria uma perda de precisão e a relatividade causaria distorções nas cronometragens.

---

plouffe.fr/Simon Plouffe.htm

1992: $e^{\pi} - \pi = 19.999099979\ldots$

2006: New formulas for Pi and 1/Pi:

$$\frac{1}{\pi} = 8\sum_{n=1}^{\infty}\frac{n}{e^{\pi n}-1} - 40\sum_{n=1}^{\infty}\frac{n}{e^{2\pi n}-1} + 32\sum_{n=1}^{\infty}\frac{n}{e^{4\pi n}-1}$$

$$\pi = 72\sum_{n=1}^{\infty}\frac{n^{-1}}{e^{\pi n}-1} - 96\sum_{n=1}^{\infty}\frac{n^{-1}}{e^{2\pi n}-1} + 24\sum_{n=1}^{\infty}\frac{n^{-1}}{e^{4\pi n}-1}$$

(wrong!!!)

Source:
https://vixra.org/abs/1409.0078
Identities Inspired by the
Ramanujan Notebook,
Second Series

https://vixra.org/abs/1410.0054

2011 - A formula for mass ratio of the neutron/proton.

$$\frac{8}{27}\left(-\sqrt{3}+\frac{5}{\frac{1}{8}(-1+\sqrt{5})+\frac{1}{4}\sqrt{\frac{3}{2}(5+\sqrt{5})}}\right)$$

Obs. title:

The Art of Inspired Guessing

→ one is related to the mass ratio of the neutron and proton.

1.00137841917

simon.plouffe@univ-nantes.fr

simon.plouffe@sympatico.ca

Não me respondeu em nenhum destes emails.

As equações estão equivocados.

https://cs.uwaterloo.ca/journals/JIS/compars.html.

/construction 503.gif

Ele não teve nem a consideração de responder, pois ele sabe que o que eu estou falando é verdade, mas ele não consegue admitir e nem dar razão à pura realidade inquestionável.

Para muitas pessoas admitir que está equivocado é um duro golpe para o seu ego.

Admitir a verdade, liberta a alma

20210911 (20210810)

https://archive.org/details/

- einleitungindieh00kleiuoft

- vorlgeovonalfre01clebrich

others:

Leçons sur la géométrie 3 vol.

www.archive.org/details/

leonssurlago01clebuoft
03

9) 9

! 1 %%! 7 %# ! % ! 2
.2
+,- * *. ! * 9 ! 2 1#

! 1 %%! 7 %# ! % ! 2
.2
+,- * *. ! * 9 ! 2 1#

!" 1 %%! 7 %# ! % ! " 2
.2
+,- * *. ! * 9 ! " 2 1#

2 *+,-* * *. ! * 9 !1 = ! 1#
.2
G # /G G ! Q 1!

2 *+,-* * *. ! * 9 !1 = ! 1#

+,- * *. ! * 9 ! 0 1#
.2 H O ! /G G #

+,- * *. ! * 9 ! # ! 1#
.2
8 ! G # /G G 1G !

+,- * *. ! * 9 ! # ! 1#

/construction 503.gif

Ele não teve nem a consideração de responder pois ele sabe que o que eu estou falando é verdade, mas ele não consegue admitir e nem dar razão à pura realidade inquestionável.

Para muitas pessoas admitir que está equivocado é um duro golpe para o seu ego.

Admitir a verdade, liberta a alma

20210910 (20210810)

https://archive.org/details/

- einleitungindieh00kleiuoft

- vorlgeovonalfre01clebuich

others:

Leçons sur la géométrie 3 vol.

www.archive.org/details/

leconssurlageo01clebuoft
03

)

/vorlesungenber11 deb
12 deb
21 deb

/rieflachvolesung00 klierich

triângulo escalino → 3 lados ≠s.

triângulo acutângulo → < 90°
triângulo retângulo → = 90°
triângulo obtusângulo → > 90°

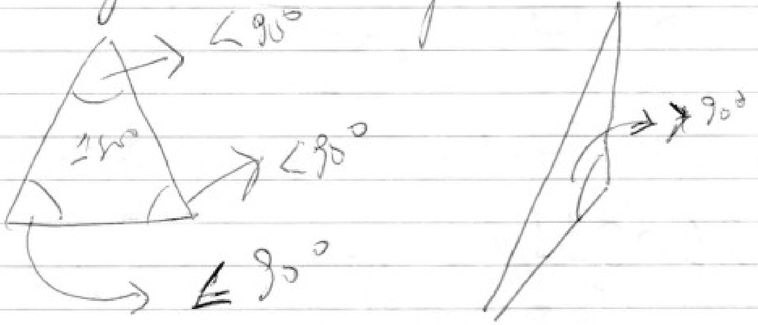
< 90°
150
< 90°
< 90°
> 90°

triângulo equilátero
triângulo isósceles
triângulo escaleno

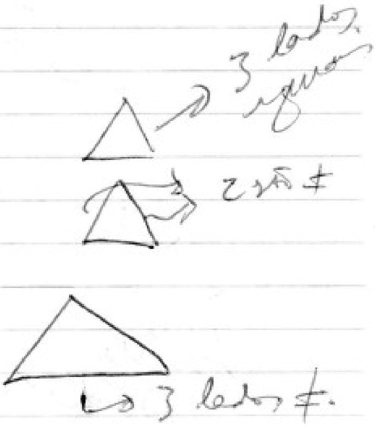
3 lados iguais
2 são ≠
3 lados ≠.

/planealgebraiccu00hilt_0

/3edtreatiseonhighesalmnoft

Treatise on the higher plane curves
Salmon

/treatisehigherpl00salmrich
(rnrich)

/ Coursemathematic0202gourrich

—

trapézio isósceles, reto, escaleno (scalene triangles)
↳ 2 lados = (paralelos)

—

Soma dos ângulos de um quadrilátero = 360°

20210911///
Samantha

—

Sb 11,20 Catecismo 299.

Deus cria um mundo ordenado e bom

299. Uma vez que Deus cria com sabedoria, a

! 1 %%! 7 %# ! % ! 7 L

+,- * *. ! * ! 7 L @D1 A 1

! 1 %%! 7 %# ! % ! L

+,- * *. ! * ! L @D1 A 1

2 1 %%! 7 %# ! % ! 1 2 $ L

1 %%! 7 %# ! % ! ! L

+,- * *. ! * ! ! L 1#

1 %%! 7 %# ! % ! ! L

+,- * *. ! * ! ! L 1#

" 1 %%! 7 %# ! % ! ! "L

+,- * *. ! * ! ! "L 1#

1 %%! 7 %# ! %1 ! ! 2 ! @

1 %%! 7 %# ! %" # ! ! %

+,- * *. ! * " # ! ! 1#

1 %%! 7 %# ! % ! 1 !

1 %%! 7 %# ! % ! !

+,- * *. ! * ! ! 1#

1 %%! 7 %# ! % ! 7 L

1 %%! 7 %# ! % ! 1 2 $ L

+,- * *. ! * ! 1 2 $ L 1#

D 1 %%! 7 %# ! % 7 ! !

+,- * *. ! * 7 ! ! 1#
D 1 %%! 7 %# ! % ! 0 L

+,- * *. ! * ! 0 L 1#

L 1 %%! 7 %# ! %1 ! ! 2 ! @

+,- * *. ! * 1 ! ! 2 ! @ 1#

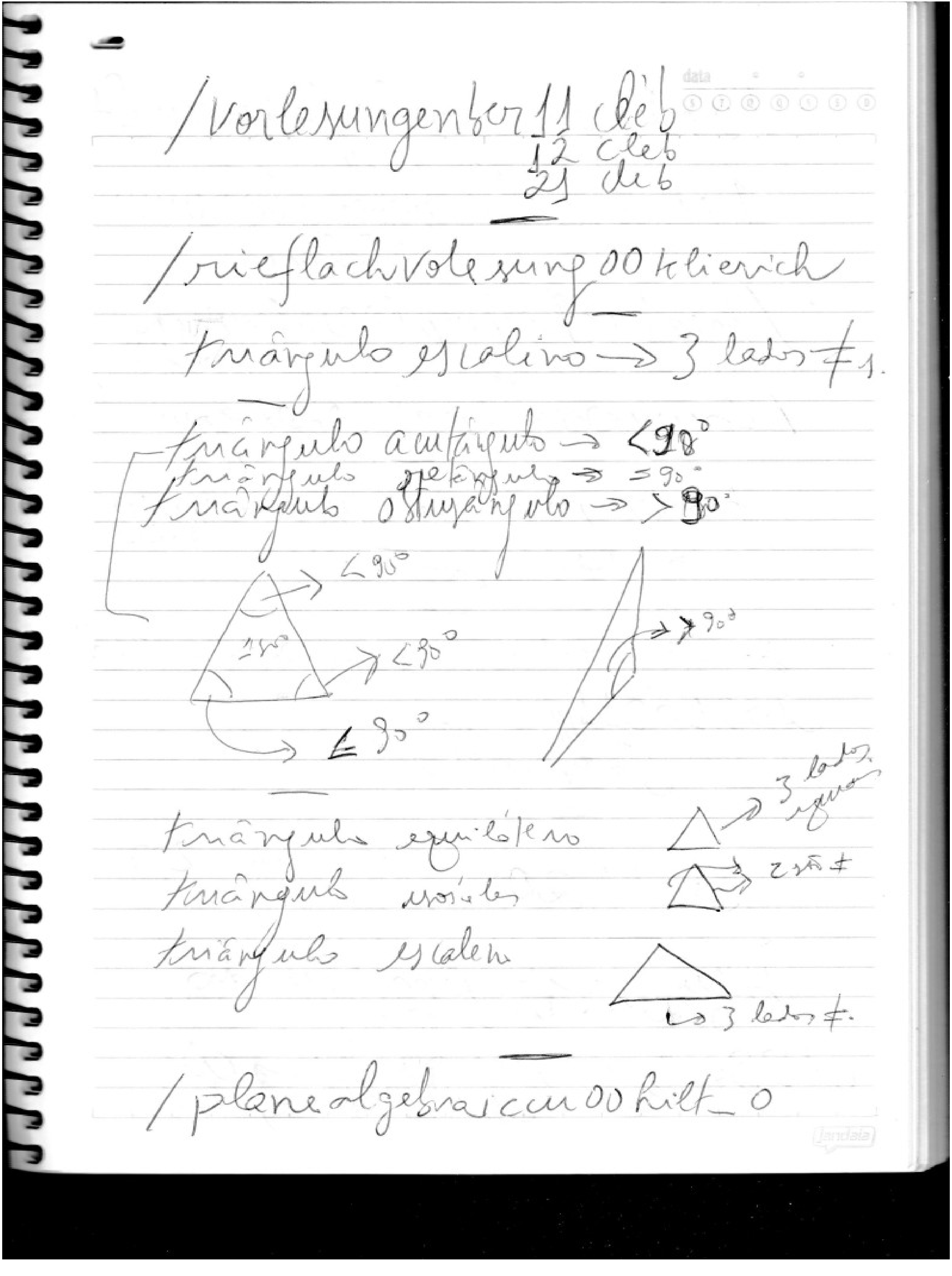

/vorlesungenbor11 deb
12 deb
21 deb

/rieflachvolesung00klierich

triângulo escalino → 3 lados ≠.

triângulo acutângulo → <90°
triângulo retângulo → =90°
triângulo obtusângulo → >90°

<90°
180°
<90°
≤90°
>90°

triângulo equilátero → 3 lados iguais
triângulo isósceles → 2 são ≠
triângulo escaleno → 3 lados ≠.

/planealgebraicum00hilt_0

/3edtreatiseonhighesalmnoft

Treatise on the higher plane curves
Salmon

/treatisehigherpl00salmrich
(Tmrich)

/Coursemathematic0202gourrich

—

trapézio isósceles, reto, escaleno (scalene triangles)
↳ 2 lados = (paralelos)

—

Soma dos ângulos de um quadrilátero = 360°

—

20210911///
Samantha

—

Sb 11,20   Catecismo 299.

Deus cria um mundo ordenado e bom

299. Uma vez que Deus cria com sabedoria, a

criação possui ordem.

"Dispusestes tudo com medida, número e peso" (Sb 11,20)

Sb 11,20 => "E, mesmo sem isto, eles poderiam perecer por um sopro, perseguidos pela justiça e aniquilados pelo vento de vosso poder; mas, dispusestes tudo com medida, quantidade (número) e peso."

IMPORTANT:

Data-NovoVivo-018-2021/Books-018-2021/20210911/Quadradutee theory)/ (Quadrature theory)/

Tangent_N:49 Les maths de l'impossible la quadrature du cercle!.pdf (ISBN 9782848841564) (ISSN 2263-4908) Novembre 2013.

Pág. 12(14)

La quadrature du cercle

) $

criação possui ordem.

"Dispusestes tudo com medida, número e peso" (Sb 11,20)

Sb 11,20 => "E, mesmo sem isto, eles poderiam perecer por um sopro, perseguidos pela justiça e aniquilados pelo vento de vosso poder; mas, dispusestes tudo com medida, quantidade e peso."
(número)

IMPORTANT:

Data-NovoVivo-018-2021/Books-018-2021/20210911/Quadradutee theory)/
(Quadrature theory)/

Tangent. N°49 Les maths de l'impossible la quadrature du cercle! .pdf (ISBN 9782848841564)
(ISSN 2263-4908)
Novembre 2013.

Pág. 12(14)

La quadrature du cercle

Pag. 16(18) →

La trisection de l'angle

Pag. 19(20) →

La quadratrice de Dinostrate

Dinostrate et la quadrature du cercle ($\sqrt{\pi}$).

/Quadrature Theory/

Advanced Differential Quadrature Methods... pdf

Pag. 44(68) –

2.1 Differential quadrature in the complex plane.

/Quadrature theory/

3.375 in 0.02 in Gaussian Quadrature.pdf

/20210911/classical Mechanics/
1439893691.pdf
↳ Compliant Mechanisms

Pág. 141 (158)
4. Compliances of Curvilinear-Axis Flexible Hinges

/20210911/Tangent/
Plane Geometry Practice.pdf

Pág. 214

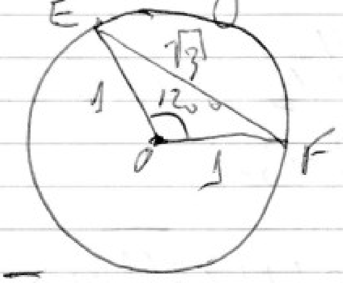

/Tangent/Tangent_-_HS74_-2020-04.pdf
→ Courbes et trajectoires

/Tangent/tangent198_-2021-02_03.pdf
→ Roger Penrose.

! *+,-* * *. ! * 8! < B G MR &$ H !
? 1 2 H! = !# ! # 1#
) & H! = !# ! #

par Hervé Lehning

La quadrature du cercle

« Quadrature du cercle » a pris le sens de « problème impossible ». Pourtant, elle est réalisable ! De manière approchée, elle correspond au calcul du nombre π. De façon exacte, à l'invention de divers instruments ou techniques. Alors qu'est-ce qui est impossible dans la quadrature ?

Pour calculer l'aire d'un domaine, les mathématiciens de l'Antiquité cherchaient à construire un carré de même aire que le domaine donné. Cette opération est simple si on part d'un rectangle, par exemple. Dans ce cas, la construction se fait à la règle et au compas (voir *le Cercle*, Bibliothèque Tangente 36, 2009).

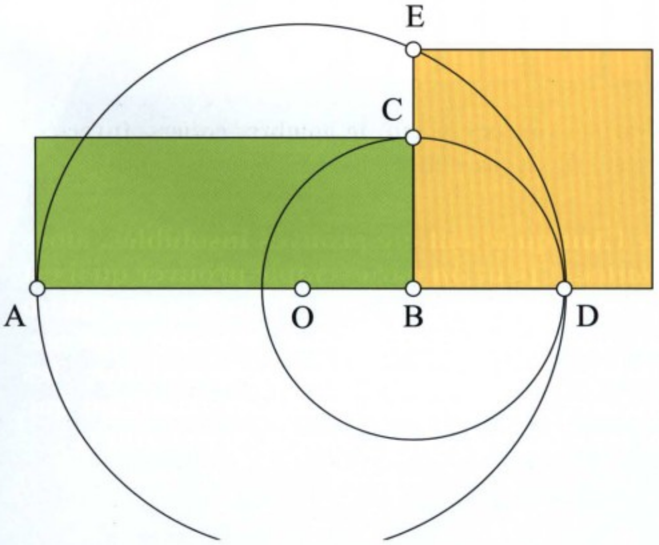

De façon générale, pour les Anciens Grecs, résoudre un problème revenait souvent à construire sa solution à la règle et au compas. De façon précise, la procédure est la suivante. On se donne une figure composée d'un certain nombre de points (au moins deux). Ensuite, on demande de construire, à la règle (non graduée) et au compas, la solution d'un problème associée à cette figure. Par exemple, on peut construire la diagonale d'un carré en n'en connaissant qu'un côté, [AB], en utilisant trois fois le compas et deux fois la règle. Cette construc-

Quadrature du rectangle à la règle et au compas : le carré jaune a même aire que le rectangle vert. Pour construire le carré à partir du rectangle, on trace d'abord le cercle de centre B passant par C. Il coupe le prolongement de [AB] en D, puis le milieu O de [AD] et le cercle de centre O passant par D. Il coupe le prolongement de [BC] en E. On construit alors un carré de côté [BE].

tion permet de réaliser la duplication du carré, c'est-à-dire de construire un carré d'aire double d'un carré donné.

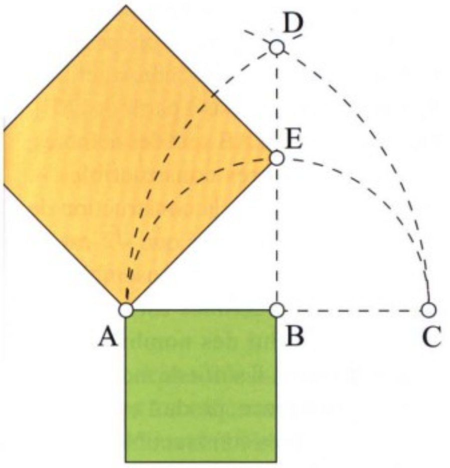

Duplication du carré : le carré jaune a une aire double de celle du carré vert. Pour construire le jaune à partir du vert, on trace le cercle de centre B passant par A, il recoupe (AB) en C. On trace alors les cercles de centres A et C passant par C et A. Ils se coupent en un point D. On trace la droite (BD). Elle coupe le premier cercle en un point E. [AE] est un côté du carré d'aire double, qui se trouve ainsi complètement déterminé.

Sur le papier, une telle construction donne une valeur approchée médiocre. La précision totale est impossible, sauf en idéal. Pour les mathématiciens grecs comme pour leurs successeurs, derrière ces objets (règle et compas) se cachent des concepts. Ceux-ci sont définissables de manière abstraite à partir de la notion de distance, plus court chemin d'une part et points équidistants de l'autre. Dans le monde des idées dont parle Platon (au IV^e siècle avant notre ère), la construction à la règle et au compas est exacte. Cette justification est souvent avancée pour expliquer l'utilisation exclusive de ces instruments. Son défaut

De la construction à la règle et au compas

Partons de deux points distincts O et I. Réaliser une construction géométrique à la règle et au compas à partir de ces deux points revient à définir une suite de points $(A_n)_{n \geq 0}$ avec $A_0 = O$, $A_1 = I$, et, pour tout $n > 1$, A_n déterminé par intersection de droites ou de cercles construits à partir des points précédents. Munissons le plan d'un repère orthonormé dont O est l'origine et I le point de coordonnées $(1, 0)$. Les droites et les cercles du plan possèdent des équations cartésiennes du type $ax + by = c$ et $(x - a)^2 + (y - b)^2 = c^2$, où les paramètres a, b et c sont déterminés par opérations algébriques sur les coordonnées des points définissant ces droites et ces cercles. Les coordonnées d'un point A_n de la suite s'obtiennent donc par résolution de systèmes de la forme :

$$\begin{cases} ax + by = c \\ a'x + b'y = c' \end{cases}, \quad \begin{cases} ax + by = c \\ (x - a')^2 + (y - b')^2 = c'^2 \end{cases},$$

$$\text{ou} \begin{cases} (x - a)^2 + (y - b)^2 = c \\ (x - a')^2 + (y - b')^2 = c'^2 \end{cases}.$$

En résolvant ces systèmes, on observe que les coordonnées du point A_n sont solutions d'équations algébriques de degré 1 ou 2, équations dont les coefficients sont obtenus à partir des coordonnées des points A_0, $A_1 \ldots A_{n-1}$. Cela permet de justifier le résultat de Wantzel : les coordonnées des points constructibles sont obtenues par succession d'additions, de soustractions, de multiplications, de divisions et de passages à la racines carrés à partir de nombres entiers. Inversement, il est possible de construire un point dont les coordonnées soient de la forme ci-dessus.

Réaliser la quadrature du cercle revient à construire à la règle et au compas une suite finie de points $(A_n)_{n \geq 0}$ dont deux éléments A_k et A_m vérifient. Si cette quadrature était possible, les coordonnées des points étant des nombres constructibles, π serait lui aussi un nombre constructible. En prouvant en 1882 que π est transcendant, Lindemann établit par la même occasion que π n'est pas constructible : la quadrature du cercle est impossible.

D. D.

L'impossibilité de la quadrature tient dans les moyens autorisés…

est de s'appliquer également à d'autres, comme l'équerre, que les mathématiciens grecs refusaient pour leurs constructions. De fait, la raison profonde de l'utilisation exclusive de la règle et du compas est d'ordre mystique, pas mathématique. Le cercle et la droite sont considérés comme les seules formes parfaites. Chez les pythagoriciens, le cercle est associé au « un », donc à la divinité, la droite à la rectitude morale, aux hommes droits. Les mathématiciens grecs s'engluèrent très vite dans la pensée que toute grandeur doit être construite à la règle et au compas, idée abandonnée de nos jours même si d'aucuns veulent l'adapter à l'ère informatique. L'idée est aussi pertinente que de remplacer les automobiles par des diligences tirées par des robots.

Cette question d'un autre temps permet simplement de motiver celle de l'existence de nombres non constructibles à la règle et au compas, π en particulier, qui est le rapport de l'aire d'un cercle au carré construit sur l'un de ses rayons. Autrement dit, la quadrature du cercle à la règle et au compas en un nombre fini d'étapes est impossible.

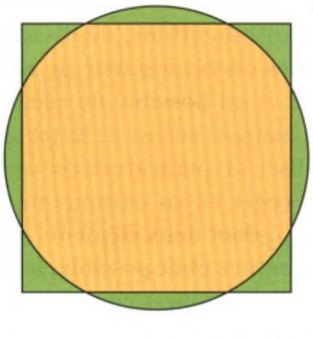

La quadrature du cercle consiste à construire, à la règle et au compas et en un nombre fini d'étapes, un carré de même surface.

Si on change les règles, la quadrature du cercle est possible, soit de façon approchée, soit en utilisant d'autres moyens que la règle et le compas.

Corps des nombres constructibles

La construction à la règle et au compas se transpose parmi les nombres. On se donne un segment [AB] et on considère tous les segments [CD] constructibles à la règle et au compas à partir de [AB]. Les rapports CD / AB sont des nombres, appelés « nombres constructibles ». Ainsi, on déduit de la construction de la diagonale d'un carré que $\sqrt{2}$ est un nombre constructible. L'ensemble des nombres constructibles est un corps, inclus dans celui des nombres réels. Pour le prouver, il suffit de montrer que somme, différence, produit et quotient de deux nombres constructibles le sont de même. Les deux premiers cas se résolvent facilement avec un coup de règle et deux de compas. Pour les deux autres, il est nécessaire d'utiliser le théorème de Thalès. De plus, ce théorème permet de montrer que les nombres rationnels sont tous constructibles.

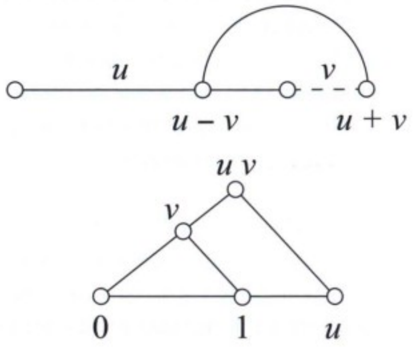

Constructions à la règle et au compas : si *u* et *v* sont constructibles, alors *u* + *v* et *u* – *v* le sont grâce à une utilisation du compas. Le produit *u v* l'est d'après le théorème de Thalès. L'inverse de *u* se construit en faisant coïncider *u v* avec 1 dans la figure du bas.

Par ailleurs, en utilisant les équations de droites et de cercles, on démontre que les nombres constructibles sont

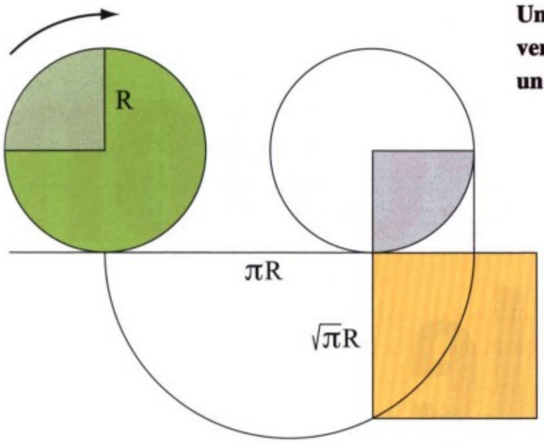

Une quadrature du cercle. On fait tourner le cercle vert d'un demi-tour, ce qui introduit le nombre π, puis un carré de même aire que le cercle initial.

Carl von Lindemann a montré l'impossibilité de la quadrature du cercle en prouvant la transcendance du nombre π.

tous algébriques, c'est-à-dire solutions d'équations à coefficients entiers. Pierre-Laurent Wantzel (1814–1848) précisa ce corps des nombres constructibles et prouva ainsi l'impossibilité de certaines constructions, comme la duplication du cube, qui consiste à construire un cube de volume double d'un cube donné. Dans ce cas, il suffit de montrer que la racine cubique de 2 ne peut être construite, du fait des équations qu'elle vérifie.

Les premiers nombres non algébriques, c'est-à-dire *transcendants*, ont été découverts en 1844 par Joseph Liouville (1809–1882), mais il a fallu attendre Charles Hermite (1822–1901) en 1873 pour que la transcendance de la constante d'Euler *e* soit démontrée. En 1882, Ferdinand von Lindemann (1852–1939) montra que π est transcendant, donc non constructible. Sa preuve, très technique, dépasse largement les ambitions de cet article. Elle prouve l'impossibilité de la quadrature du cercle, ce qui n'a pas tari l'imagination des quadrateurs, qui continuent aujourd'hui encore à envoyer à *Tangente* leurs « solutions » à un problème que l'on sait insoluble depuis longtemps.

H. L.

Un peu d'histoire

Pour les mathématiciens de la Grèce antique, il n'existait pas d'autres nombres que ceux que nous nommons de nos jours les entiers naturels (et encore… 1 n'était pas considéré comme nombre, mais comme principe de tout nombre). Les premiers mathématiciens, comme Pythagore, pensaient que *« tout est nombre »*, c'est-à-dire que deux grandeurs (longueurs, aires…) avaient toujours une commune mesure, en d'autres termes étaient commensurables.

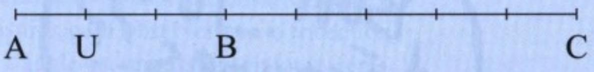

Commensurabilité : AB et AC sont commensurables s'il existe un point U tel que AB et AC soient multiples de AU.

Malheureusement pour sa doctrine, Pythagore prouva lui-même qu'il existe des grandeurs incommensurables, le côté et la diagonale d'un carré par exemple. Les longueurs commensurables correspondent à ce que nous nommons *nombres rationnels*, les autres aux *nombres irrationnels*. La diagonale du carré est cependant constructible à la règle et au compas, c'est pourquoi les mathématiciens grecs remplacèrent leur idéal de commensurabilité en celui de construction à la règle et au compas. Ils nous ont ainsi légué le problème de la quadrature du cercle, dont on n'a suspecté l'impossibilité que vers le XVI^{e} siècle.

La trisection de l'angle

À l'aide d'une règle et d'un compas, il est assez facile de séparer un angle en deux parts égales en construisant une bissectrice. Est-il de même possible de le séparer en trois ? C'est le problème de la trisection de l'angle, qui date de l'antiquité grecque.

Bon ...tu découpes ??
Heu...
CATOUNE

De nombreux mathématiciens se sont investis dans le problème de la trisection de l'angle, proposant d'abord des solutions partielles, jusqu'à parvenir à établir que la trisection de l'angle est en fait impossible en toute généralité.

La trisectrice de Maclaurin

Parmi les solutions proposées pour amorcer une trisection à la règle et au compas, plusieurs sont basées sur l'introduction de courbes planes dont les points seraient constructibles à la règle et au compas. C'est le cas de la trisectrice de Maclaurin, introduite au XVIII^e siècle. Pour la définir, considérons deux points distincts O et Ω du plan et introduisons la médiatrice Δ du segment [OΩ]. Pour un point P variant sur Δ, on considère le point M déterminé par la deuxième l'intersection de la droite (PΩ) et du cercle de centre O passant par P.

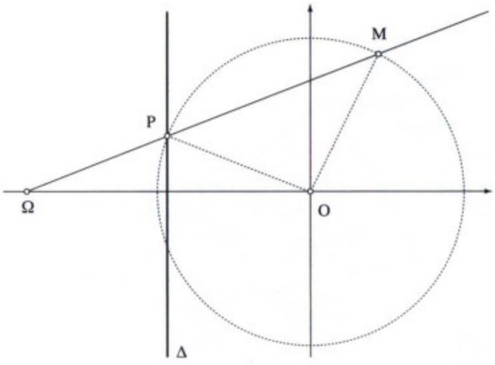

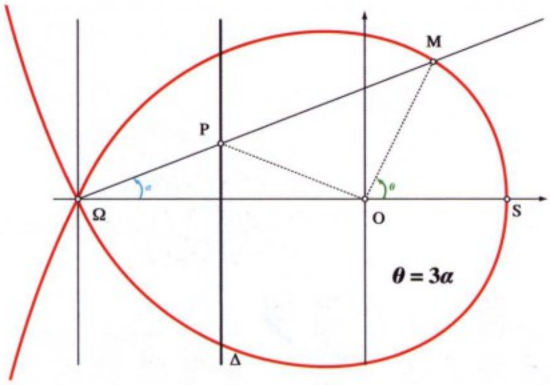

Ce point M parcourt la courbe trisectrice qui nous intéresse. Si l'angle à couper en trois est $\widehat{SOM}$, sa trisection apparaît comme étant $\widehat{S\Omega M}$.

En fait, cette courbe est inspirée d'une idée déjà introduite par Archimède. Considérons un cercle de rayon R dont le centre est le sommet de l'angle à découper. Définissons l'axe des abscisses par un côté de cet angle ; le cercle introduit coupe alors l'autre côté en un point M. À l'aide de notre règle graduée, traçons la droite issue de M coupant l'axe des abscisses en un point Ω et le cercle en un point P de sorte que ΩP = R. Par cette dernière condition, nous exploitons une graduation de la règle ; ce n'est donc pas une construction à la règle et au compas au sens strict. En observant que les triangles OPΩ et POM sont isocèles, on établit aisément que l'angle $\widehat{SOM}$ est égal au triple de $\widehat{S\Omega M}$.

La trisection de l'angle à la règle et au compas est cependant impossible. Certes, il est certains angles qu'il est facile de couper en trois ; c'est par exemple le cas de l'angle plat, de mesure π. En effet, puisque $\cos(\pi/3) = 1/2$, il n'est pas difficile de construire un angle de mesure $\pi/3$ (et c'est même très pratique pour découper un gâteau en six parts sensiblement égales…). Ensuite, pour observer que la trisection de l'angle est impossible en toute généralité, observons qu'un angle de mesure $\pi/3$ ne peut, lui, être séparé en trois. Pour cela, considérons le nombre

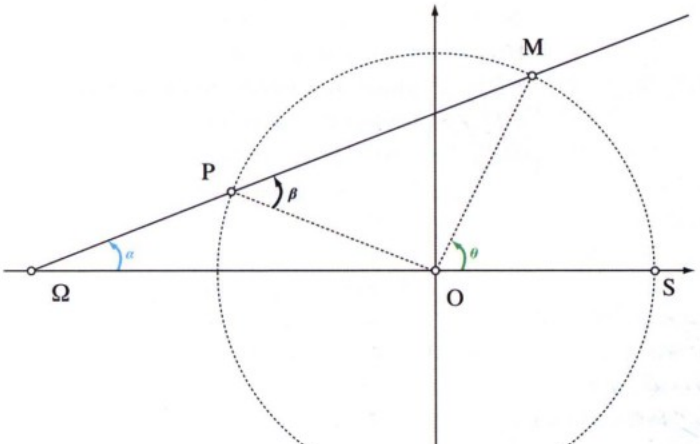

En considérant le triangle OPΩ, $2\alpha = \beta$.
En considérant le triangle POM, $2\beta = \alpha + \theta$.

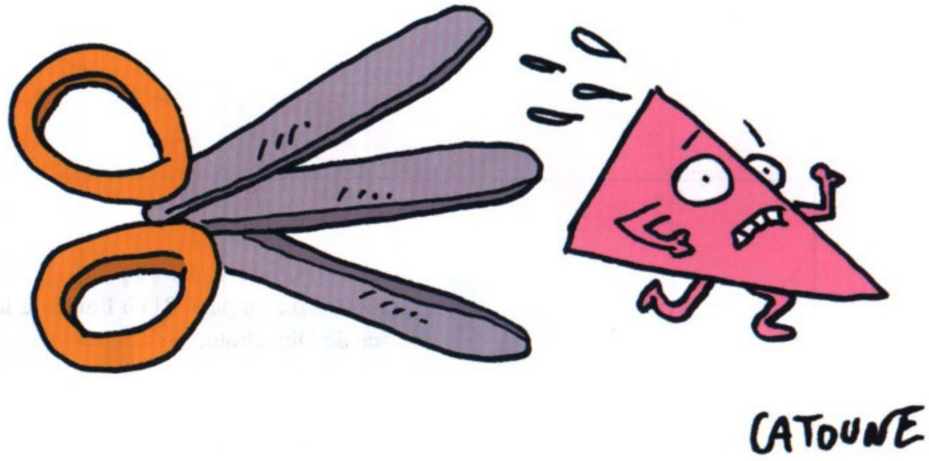

Le triangle de Morley

Les bissectrices intérieures d'un triangle concourent en un point : c'est le centre du cercle inscrit dans ce triangle. Les trisectrices intérieures d'un triangle se coupent elles aussi de façon remarquable : leurs points d'intersection déterminent un triangle... équilatéral ! C'est le triangle de Morley, découvert par Frank Morley à la toute fin du XIX^e siècle.

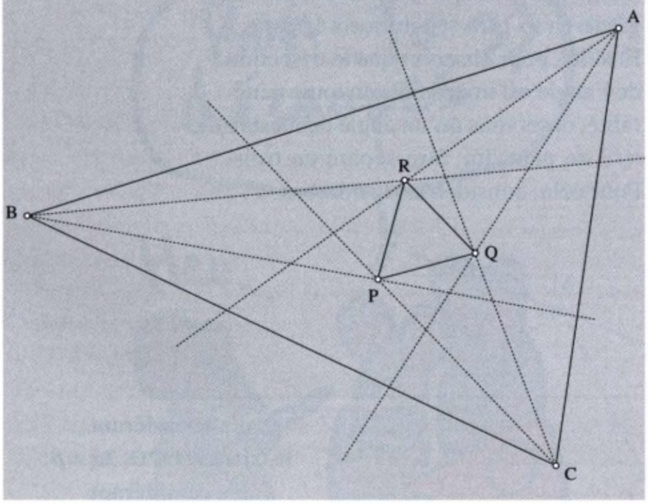

$x = \cos(\pi/9)$. En exploitant l'identité $\cos(3a) = 4\cos^3 a - 3\cos a$, on établit que x est solution de l'équation $8x^3 - 6x - 1 = 0$. C'est une équation algébrique construite à partir d'une expression polynomiale à coefficients entiers (plus précisément, c'est une équation algébrique de degré 3).

Au XIX^e siècle, Wantzel a caractérisé les points constructibles à la règle et au compas en un nombre fini d'étapes à partir d'une unité donnée. Ces points sont ceux dont les coordonnées sont racines d'équations algébriques de degré une puissance de 2, mais qui ne sont racines d'aucune de degré strictement inférieur. Par exemple, le point de coordonnées $(1 + \sqrt{5})/2$ et $2^{1/4}$ est constructible, car ses coordonnées sont respectivement racines des équations $x^2 - x - 1 = 0$ et $x^4 - 2 = 0$.

Si l'on pouvait construire un angle de mesure $\pi/9$ à la règle et au compas, il en serait de même du point de coordonnées $(\cos(\pi/9), \sin(\pi/9))$. Or, son abscisse x est solution d'une équation de degré 3, et on peut justifier qu'elle n'est solution d'aucune équation de degré strictement inférieur : la construction à la règle et au compas d'un angle de mesure $\pi/9$ est donc tout bonnement impossible. Il en est alors de même de la trisection de l'angle en général.

D. D.

par Jean-Alain Roddier

La quadratrice de Dinostrate

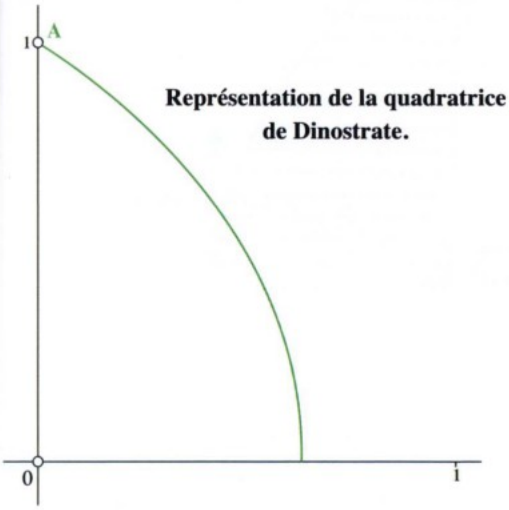

Représentation de la quadratrice de Dinostrate.

À l'origine, la quadratrice de Dinostrate est une courbe servant à la trisection approchée des angles. Elle aurait été définie par le grec Hippias d'Élis au V^e siècle avant notre ère, avant d'être reprise un peu plus tard par le mathématicien grec Dinostrate pour effectuer la quadrature approchée du cercle.

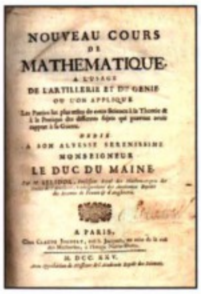

NOUVEAU COURS
DE
MATHEMATIQUE
A L'USAGE
DE L'ARTILLERIE ET DU GENIE
OU L'ON APPLIQUE
DEDIÉ
A SON ALTESSE SERENISSIME
MONSEIGNEUR
LE DUC DU MAINE
A PARIS,
M. DCC. XXV.

En 1725, dans son Nouveau Cours de mathématique à l'usage de l'artillerie et du génie, le mathématicien Bernard Forest de Bélidor (1698–1761) expose une méthode de construction point par point de la quadratrice de Dinostrate, puis il explique comment se servir de cette courbe construite « mécaniquement » pour réaliser la trisection de l'angle (voir l'article sur ce sujet). Une représentation paramétrique du support de la quadratrice de Dinostrate est

$$\begin{cases} x = \dfrac{t}{\tan\left(\dfrac{\pi}{2}t\right)}, & \text{avec } 0 < t \leq 1. \\ y = t \end{cases}$$

Dinostrate et la quadrature du cercle

La quadrature du cercle consiste à construire un carré d'aire π à partir d'un cercle de rayon 1, ce qui revient à construire un segment de longueur $\sqrt{\pi}$. On peut résoudre ce problème de manière exacte (voir en page 21) à l'aide de la quadratrice de Dinostrate.

Notons M(t) le point de l'arc paramétré ci-contre correspondant à la valeur t du paramètre. Pour la valeur $t = 0$, le point M(t) n'est pas défini mais admet une position limite lorsque t tend vers 0. Un calcul de limite élémentaire montre que l'arc peut être prolongé par continuité en lui adjoignant le point S de coordonnées $(2/\pi, 0)$.

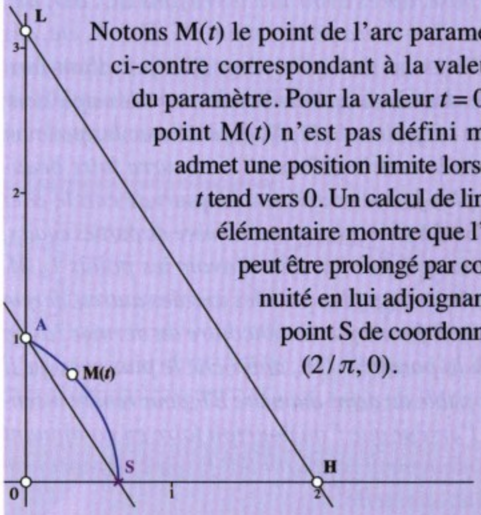

Considérons le point H (2, 0), puis le point L d'intersection de l'axe des ordonnées avec la parallèle à la droite (AS) passant par H (voir la figure ci-dessus). L a pour coordonnées (0, π). On peut alors construire le segment [BF], qui a pour longueur $\sqrt{\pi}$. Pour cela, on considère le point M(–1, 0), puis le cercle de diamètre [ML]. Dans le triangle MFL rectangle en F, nous avons $BF^2 = BM\ BL = \pi$.

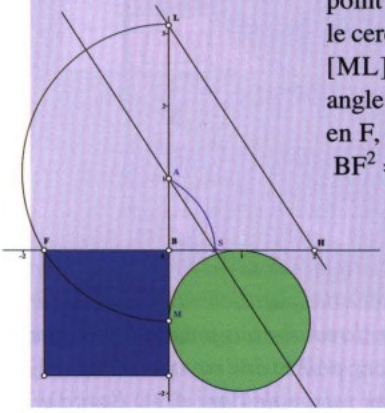

par Jean-Alain Roddier

Le texte de Bélidor

Bernard Forest de Bélidor a décrit la construction point par point de la quadratrice de Dinostrate dans le texte suivant.

« Pour décrire cette courbe, il faut diviser le rayon AB en un grand nombre de parties égales ; de manière que le quart de cercle AT puisse être divisé en un même nombre de parties égales ; ainsi nous supposerons que l'on a divisé le quart de cercle en seize parties égales ; aussi bien que le rayon AB. Cela posé, après avoir tiré les rayons BC, BD, BE, BF, etc. l'on tirera par les points G, H, I, K, etc. des parallèles au demi-diamètre BT, qui allant rencontrer les rayons qui divisent le quart de cercle donneront les points L, M, N, O, etc. avec lesquels on tracera la courbe AS, que l'on pourra faire beaucoup plus juste, en divisant le quart de cercle & et le rayon BA en un plus grand nombre de parties égales, que l'on n'a fait ici, afin d'avoir les points L, M, N, O, beaucoup plus près les uns des autres, & que le point R formé par la rencontre du dernier rayon BP, & la parallèle QR, approche le plus près qu'il est possible du demi-diamètre BT, pour rendre insensible l'erreur que l'on pourroit faire en continuant mécaniquement la courbe AR jusqu'à la rencontre du demi-diamètre. »

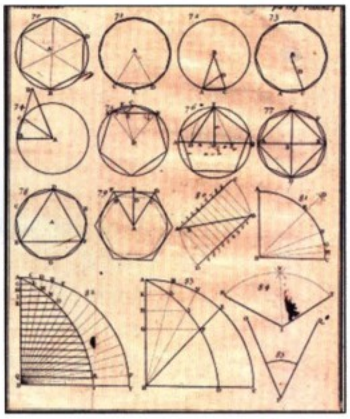

« Il faut remarquer que, par la génération de cette courbe, si l'on mène des parallèles HM & KO, qui ailleurs rencontrer la courbe aux points M & O, que si l'on tire par ces points des rayons BD & BF, qu'il y aura même raison de l'arc AD à l'arc DF, que de la ligne AH à la ligne HK. »

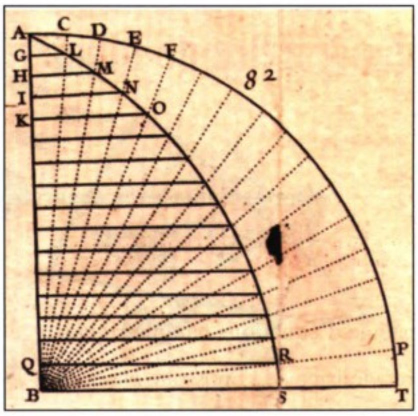

Autrement dit : $\frac{\text{longueur }(\overset{\frown}{AD})}{\text{longueur }(\overset{\frown}{DF})} = \frac{AH}{HK}$.

De là à la trisection de l'angle, il n'y a qu'un pas :

« Supposant que l'on ait tracé sur un morceau de corne ou de carton bien uni, la courbe AD de la façon qu'on vient de l'enseigner, on propose de diviser l'Angle OPQ en trois parties égales. Pour résoudre ce Problème, supposant que la courbe accompagnée de son quart de cercle AC, je fais l'angle ABE égal à l'angle donné, & au point F, où le rayon BE coupe la courbe AD, j'abaisse la perpendiculaire FG sur le diamètre AB, qui me donne la partie AG, que je divise en autant de parties égales que l'angle donné soit divisé. »

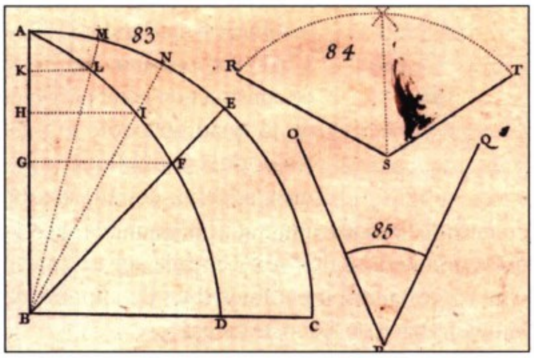

Pour ce qui relève de la trisection de l'angle obtus $\widehat{RST}$, Bélidor propose de diviser cet angle en deux angles égaux puis d'opérer la trisection de l'angle aigu ainsi obtenu.

Impossibles, ces figures ? (4)

Équation paramétrique de la quadratrice de Dinostrate

Les coordonnées du point P (voir les notations en page ci-contre) sont $(\cos(\pi t/2), \sin(\pi t/2))$. La droite (BP) a donc pour équation réduite $y = x\tan(\pi t/2)$. Le point R a pour ordonnée t (la même que celle du point Q) et il appartient à la droite (BP) ;on en déduit les coordonnées de R, qui sont : $\left(\dfrac{t}{\tan\left(\dfrac{\pi}{2}t\right)}, t\right)$.

Un système d'équations paramétriques de la quadratrice de Dinostrate est donc $\begin{cases} x = \dfrac{t}{\tan\left(\dfrac{\pi}{2}t\right)}, \\ y = t \end{cases}$ avec $t \in \mathbb{Q} \cap \,]0,1]$.

On peut ainsi obtenir une équation du support de la quadratrice de Dinostrate : $x = y/\tan(\pi y/2)$.
Pour construire cette courbe, il suffit de tracer la courbe d'équation $y = x/\tan(\pi x/2)$ puis de prendre sa symétrique par rapport à droite d'équation $y = x$. On se restreint ensuite à [0, 1].

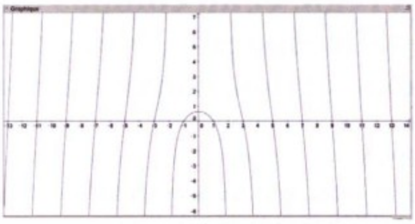

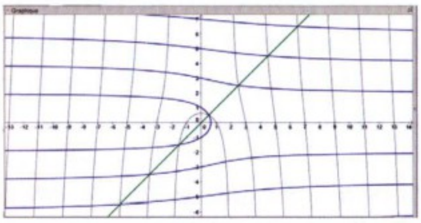

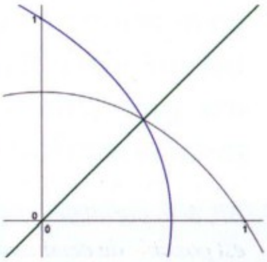

Réaliser le tripoutre

Pour réaliser un tripoutre, on peut commencer par reproduire, découper et assembler le modèle en suivant les simples indications de la figure ci-contre. Si tout se passe bien, vous obtiendrez une sorte de maquette avec quatre barres (en papier) à angle droit. Maintenant, il s'agit de se positionner avec un appareil photo face à la maquette et de choisir un point de vue, en jouant sur la distance qui sépare l'objectif de la maquette, afin de fusionner visuellement la barre supérieure proche du photographe avec la barre du fond (voir les différents essais ci-contre). Il est parfois nécessaire de rogner l'angle de la barre supérieure (marqué avec un point rouge) pour le faire coïncider parfaitement avec la barre à l'arrière-plan. Avec un peu de patience, et beaucoup d'essais, vous obtiendrez d'excellents résultats !

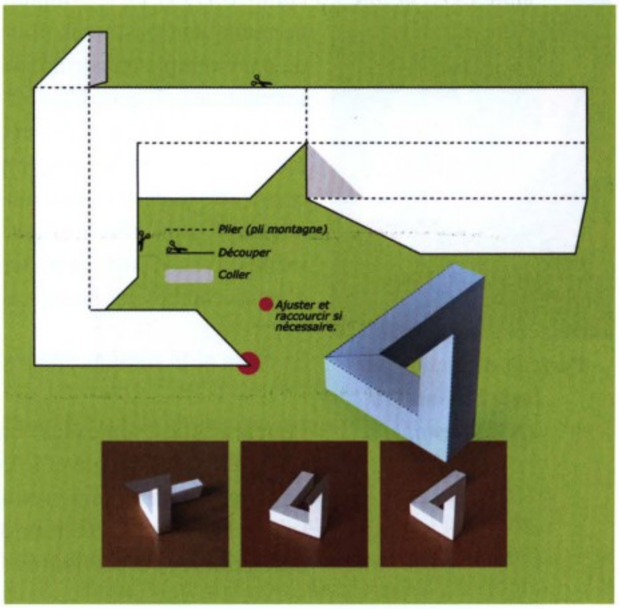

L'illustration ci-dessus montre comment notre cerveau perçoit un seul et unique triangle ABC alors que les droites qui le représentent dans l'espace réel ont une orientation fort différente. Il s'ensuit qu'il est possible de fabriquer de nombreuses variantes de la maquette pour atteindre le même résultat visuel.

2 *+,-* * *. ! * +#7! # 6 ! - !# !
!1 ! P B! @ , +11 # ! ! P M ! 1#
) && 9 6 ! = !# ! 1 4 1 !

(2) Use of analytic function or Airy's stress function in the form of complex variable; and

(3) Application of boundary conditions to determination of the unknowns.

The last step involves solving a system of linear equations, which can be neglected if conformal mapping, a powerful technique able to transform a complicated physical problem into a readily analyzable one, is employed. On the contrary, conformal mapping can be readily performed if Lagrange interpolation in the complex plane is used together. We will thus in section 2.4 develop a methodology which combines conformal mapping technique and complex DQ method for easily solving some two-dimensional potential problems.

2.1 Differential quadrature in the complex plane

Lagrange interpolation on the real line is one of the simplest interpolation methods and has been widely investigated over the past two centuries (Buchanan and Turner, 1997). It is possible to generalize Lagrange interpolation to the complex plane. Some changes are, however, necessary to account for more complicated cases special in the complex plane. In what follows, extension of Lagrange interpolation to the complex plane and differentiation of a function of a complex variable (or more exactly analytical function) by use of complex Lagrange polynomial are presented. Because books on complex variable theory are plenty, the reader is assumed to be familiar with basic knowledge of complex variables and analytic functions. Several books are listed in the reference section and the reader is referred to Chapter 8.

It is assumed that in the domain S in the two-dimensional complex plane encircled by its boundary Γ, an analytic function $\varphi(z)$ is defined. On the boundary Γ, N nodes are specified as shown in Fig. 2.1. These nodes are assumed to be a sequence in the form of

$$z_1, z_2, \ldots, z_N \tag{2.1}$$

on which the function values of the analytic function $\varphi(z)$ are known

$$\varphi_1, \varphi_2, \cdots, \varphi_N \tag{2.2}$$

Our purpose is to construct a polynomial of $z = x + iy$, where $i = \sqrt{-1}$ is the imaginary unit, in the form of

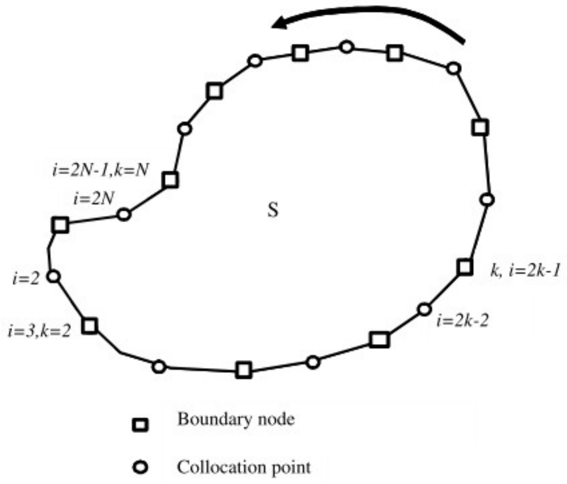

FIGURE 2.1: Boundary nodes and collocation points.

$$\varphi(z) = \sum_{k=1}^{N} \lambda_k(z)\varphi_k \tag{2.3}$$

satisfying

$$\lambda_k(z) = \begin{cases} 1 & z = z_k \\ 0 & z = z_j \neq z_k \end{cases} \quad k = 1, \cdots, N, \quad j = 1, \cdots, N. \tag{2.4}$$

This is the Kronecker delta function. It is its property that guarantees the interpolating polynomial $\varphi(z)$ equal to the function value φ_k at each node z_k. Direct extension of Lagrange polynomial on the real line to the complex plane leads to the following equations (Buchanan and Turner, 1997):

$$\lambda_k(z) = \frac{L_k(z)}{L_k(z_k)}, \quad L_k(z) = \prod_{\substack{j=1 \\ j \neq k}}^{N} (z - z_j), \tag{2.5}$$

where the symbol $\prod$ represents multiplication of the terms following the symbol.

It is straightforward to find the first- and second-order derivatives by differentiating Eq. (2.5) with respect to z

$$\varphi'(z) = \frac{d\varphi(z)}{dz} = \sum_{k=1}^{N} a_k(z)\varphi_k \text{ and } \varphi''(\mathrm{z}) = \frac{\mathrm{d}^2\varphi}{\mathrm{dz}^2} = \sum_{\mathrm{k}=1}^{\mathrm{N}} \mathrm{b_k(z)}\varphi_{\mathrm{k}} \tag{2.6}$$

where $a_k(z)$ are the coefficients for the first-order derivative. After some manipulations, we obtain the coefficients for $z \neq z_k$ as

$$a_k(z) = \frac{d\lambda_k(z)}{dz} = \frac{1}{z - z_k} \sum_{\substack{\ell=1 \\ \ell \neq k}}^{N} \frac{L_\ell(z_\ell)}{L_k(z_k)}, \quad k = 1, \cdots, N. \tag{2.7a}$$

And for $z = z_k$

$$a_k(z) = \sum_{\substack{\ell=1 \\ \ell \neq k}}^{N} a_\ell(z) \frac{L_\ell(z_\ell)}{L_k(z_k)}, \quad k = 1, \cdots, N. \tag{2.7b}$$

The coefficients for the second-order derivative can be obtained in the similar manner as

$$b_k(z) = \frac{da_k(z)}{dz} = \frac{d^2\lambda_k(z)}{dz^2} = \frac{z - z_k}{1} \sum_{\substack{\ell=1 \\ \ell \neq k}}^{N} a_\ell(z) \frac{L_\ell(z_\ell)}{L_k(z_k)} - \frac{z - z_k}{a_k(z)} \tag{2.8a}$$

$k = 1, \cdots, N, \text{if } \mathrm{z} \neq \mathrm{z_k}$, and

$$b_k(z) = \frac{1}{2} \sum_{\substack{\ell=1 \\ \ell \neq k}}^{N} b_\ell(z) \frac{L_\ell(z_\ell)}{L_k(z_k)}, \quad k = 1, \cdots, N, \quad \text{if} \quad \mathrm{z} = \mathrm{z_k}. \tag{2.8b}$$

Equations (2.7) and (2.8) are the most important formulas in this chapter. It is clearly shown that both first- and second-order derivatives are reduced to the weighted sum of the function values at the nodes. This greatly simplifies the numerical differentiation process without deteriorating numerical accuracy.

It is obvious that if N nodes are used, the interpolating polynomial degree is $(N-1)$. Using this idea, it can be shown that the approximation errors for the first-order derivative (R^1) and the second-order one (R^2) are

$$R^1(z_k) = \frac{\varphi^{(N)}(\xi) W^{(1)}(z_k)}{N!}, \quad k = 1, \ldots, N \tag{2.9a}$$

$$R^2(z_k) = \frac{\varphi^{(N)}(\xi) W^{(2)}(z_k)}{N!}, \quad k = 1, \ldots, N \tag{2.9b}$$

where $W(z) = \prod_{j=1}^{N} (z - z_j)$. These residual estimates show that it is possible to achieve high accuracy if the number of grid points N is large. Moreover accuracy is proportional to N or even to its powers.

The aforementioned formulas are identical to those on the real line if complex variable z is replaced by real variable x. But a complex polynomial is really different from a real polynomial in the sense that the former also allows for the appearance of the terms of negative powers. This is because a Laurent's series, in which terms of negative powers appear, in the complex plane replaces the Taylor's series on the real line. Terms of negative powers in a Laurent's series bring in convenience to treat infinite domain as these terms vanish or are constants at infinity. In "domain" type methods such as Finite Element Method (FEM), an infinite domain is approximated by a very large domain. This is not cost-effective. On the other hand, infinite domain problems can be easily handled by boundary methods such as Boundary Element Method (BEM). In the following, we will construct a Lagrange polynomial which contains terms of negative powers, and thus allows for effective treatment of infinite domain.

It is proposed to construct the Lagrange interpolation for infinite domain in the following way:

$$\lambda_k(z) = \frac{\Lambda_k(z)}{\Lambda_k(z_k)}, \quad \Lambda_k(z) = \prod_{\substack{j=1 \\ j\neq k}}^{N} \left(\frac{1}{z} - \frac{1}{z_j}\right). \tag{2.10}$$

It is clearly indicated that $\lambda_k(z)$ possesses the Kronecker delta function property defined by Eq. (2.4). After some algebraic operations, Eq. (2.10) turns out to be a rational interpolation. Hence, Lagrange interpolation in the complex plane is composed of two parts: polynomial and rational interpolations. Note that polynomial and rational approximations are two individual techniques on the real line, but each of them is a part of Lagrange interpolation in the complex plane.

This above procedure is justified by the fact that there are two types of analytical functions: holomorphic and meromorphic (Chapter 8). The former is differentiable in an open subset, and thus can be approximated by Lagrange polynomial in which all the terms have positive powers. The latter is differentiable except at finitely many points and thus can be approximated by a complex ration number.

To find the first- and second-order derivatives of $\Lambda_k(z)$, multiply both sides of Eq. (2.10) by $(1/z - 1/z_k)$ and then differentiate both sides with respect to z. The results after simple manipulations are

$$\varphi'(z) = \frac{d\varphi(z)}{dz} = \sum_{k=1}^{N} a_k(z)\varphi_k \tag{2.11}$$

for the first derivative, and

$$\varphi''(z) = \frac{d^2\varphi}{dz^2} = \sum_{k=1}^{N} b_k(z)\varphi_k \tag{2.12}$$

for the second derivative. The coefficients for the first-order derivative are

$$a_k(z) = \frac{d\lambda_k(z)}{dz} = -\frac{1}{z - z^2/z_k} \sum_{\substack{\ell=1 \\ \ell\neq k}}^{N} \frac{\Lambda_\ell(z_\ell)}{\Lambda_k(z_k)}, \quad k = 1, \cdots, N, \text{if} \quad \mathrm{z} \neq \mathrm{z_k} \tag{2.13}$$

and

$$a_k(z) = \sum_{\substack{\ell=1 \\ \ell\neq k}}^{N} a_l(z) \frac{\Lambda_\ell(z_\ell)}{\Lambda_k(z_k)}, \quad k = 1, \cdots, N, \text{if} \quad \mathrm{z} = \mathrm{z_k}. \tag{2.14}$$

The coefficients for the second-order derivative are

$$b_k(z) = \frac{da_k(z)}{dz} = \frac{d^2\lambda(z)}{dz^2} = -\frac{1}{z - z^2/z_k}\Big[\sum_{\substack{\ell=1 \\ \ell\neq k}}^{n} a_\ell(z) \frac{\Lambda_\ell(z_\ell)}{\Lambda_k(z_k)} - (1 - \frac{2z}{z_k}) a_k(z)\Big] \tag{2.15}$$

if $\quad z \neq z_k, k = 1, \cdots, N$ and

$$b_k(z) = \frac{z_k}{2 + z_k} \sum_{\substack{\ell=1 \\ \ell\neq k}}^{N} b_\ell(z) \frac{\Lambda_\ell(z_\ell)}{\Lambda_k(z_k)}, \text{if} \quad \mathrm{z} = \mathrm{z_k}, \mathrm{k} = 1, \cdots, \mathrm{N}. \tag{2.16}$$

Equations (2.6) to (2.8) are suitable for finite domain and Eqs. (2.11) to (2.16) are suitable for infinite domain.

Example 2.1 *Lagrange interpolation in the complex plane*

Lagrange interpolation in the complex plane is an extension of Lagrange interpolation on the real line. It plays a crucial role in complex DQ method, and the accuracy related with interpolation and numerical differentiation is of major concern. This example is designed to check its accuracy.

Suppose an analytic function $\varphi(z)$ is given in a domain encircled by its boundary Γ. Because the analytic function is given, its values on the boundary Γ are also known. Based on the function values on N nodes on the boundary Γ and Lagrange interpolation in Eqs. (2.5) to (2.8), the function value and its first- and second-order derivatives at any point within the domain can be obtained. Comparison of numerical and analytical results will reveal the accuracy of Lagrange interpolation in the complex plane.

Take $\varphi(z) = z^3$ in a unit circle bounded by Γ: $|z| = 1$ for example. N nodes on Γ are specified. The relative errors are defined

$$e_1 = \frac{|\varphi(z) - \varphi_0(z)|}{|\varphi_0(z)|}, \quad e_2 = \frac{|\varphi'(z) - \varphi_0'(z)|}{|\varphi_0'(z)|}, \quad e_3 = \frac{|\varphi''(z) - \varphi_0''(z)|}{|\varphi_0''(z)|} \quad (2.17)$$

where the subscript "$_0$" denotes the known function values, and thus $\varphi_0 = z^3$, $\varphi_0' = 3z^2$ and $\varphi_0'' = 6z$. The letters in Eq. (2.17) without subscripts denote the numerical results.

The relative errors defined by Eq. (2.17) for the point $z = \sqrt{2}(1+i)/4$ are shown in Fig. 2.2. When $N = 5$ the relative errors are up to 3%. As N increases to 10, the relative errors are dramatically reduced to the order of $10^{-5} \sim 10^{-4}$ %. As N increases to larger numbers, the relative errors increase rather than decrease. The reason is that as N increases, the order of Lagrange polynomial $L(z)$ in Eq. (2.5) is correspondingly raised and more higher-order terms appear in Lagrange polynomial. From numerical analysis, it is well known that higher-order monomials (of the type z^n) might introduce local oscillations to the numerical results, as a consequence, deteriorating numerical accuracy. However, the numerical accuracies as shown in Fig. 2.2 remain below 10^{-2}% for the second-order derivatives, 10^{-3}% for the first-order one and 10^{-4}% for the function values over the whole range. Such accuracy is encouraging.

If more similar analyses are performed at other points, which is omitted here for brevity, we are able to obtain the same conclusions (Zong, 2003a). Thus, as on the real line, Lagrange interpolation in the complex plane may also yield highly accurate results by use of a small number of nodes.

2.2 Complex DQ method for potential problems

In this and the next sections, the applicability and validity of the complex DQ method are demonstrated through the two-dimensional potential problems governed by Laplace's equation and plane linear elastic problems governed by bi-harmonic equation.

Consider a domain S encircled by its boundary Γ. An analytic function defined on S is of the form

$$\varphi(z) = u(x, y) + iv(x, y) \quad (2.18)$$

where $i = \sqrt{-1}$ is the imaginary unit, u,v are real and imaginary parts, respectively, and $z = x + iy$.

+,- * *. ! * " " ' /! !
- !# ! 1#

+,- * *. ! * M ! H 2 1 !
! 6 5 4 B , 1#
) & ' & 1 ! 7 ! +4 5 4 2 B

4 Compliances of Curvilinear-Axis Flexible Hinges

Flexible hinges with curved/circular axis are employed in applications where coupling between axial and bending deformation, as well as between bending and torsion, are necessary. Analytical modeling and design of circular-axis hinges for in-plane mechanism applications with either small or large deformations have been studied by Hsiao and Lin [1], Lee and Ziegert [2], Shisheng et al. [3], Callegari et al. [4], Howell [5] or Jagirda [6], for instance. Similar analysis/design methods have also been developed to characterize the out-of-plane (diaphragm) behavior of circular-axis elastic members in a wide array of applications, including opto-mechanics, precision positioning or robotics – see Trylinski [7], Sydenham [8], Vukobratovich et al. [9], Kyusogin and Sagawa [10], Hermann [11], Mohamad et al. [12], Duong and Kazerounian [13], Shielpiekandula and Youcef-Toumi [14], Yong and Reza-Mohemani [15], Awtar [16], Awtar and Slocum [17], Trease et al. [18], Ishaquddin et al. [19] and Kress et al. [20]. Finite element methods as detailed in Ishaquddin et al. [19], Kress et al. [20], also see Chapter 8 – and compliance matrix-based approaches – as detailed in Lobontiu and Cullin [21], Lobontiu et al. [22], Lobontiu [23–26], Lobontiu et al. [27], Lobontiu [28] – have been utilized to model curvilinear-axis flexible hinges and mechanisms incorporating them.

This chapter formulates the in-plane and out-of-plane compliance matrices of planar flexible hinges that are fully formed of curvilinear-axis segments or combine curvilinear-axis segments with straight-axis segments. By means of translation, addition, rotation and mirroring, the local-frame compliance matrices of the individual segments, like those studied in Chapter 2, are mixed either in series or in parallel to produce the compliance matrices of the final hinge configurations. Flexible-hinge designs, which are symmetric or antisymmetric with respect to an axis, are studied in this chapter by means of compliance matrices of simpler hinge portions. Folded hinges are also analyzed; they comprise curvilinear-axis segments in arrangements that are inscribed in geometric envelopes of straight/curvilinear profile or may be of radial design.

4.1 COMPLIANCE MATRIX TRANSFORMATIONS

In order to formulate the in-plane and out-of-plane compliance matrices of flexible hinges that incorporate curvilinear-axis segments, the local-frame segment compliance matrices need to be transferred to the global frame; this is mainly achieved by matrix translation and rotation, but it can also include mirroring.

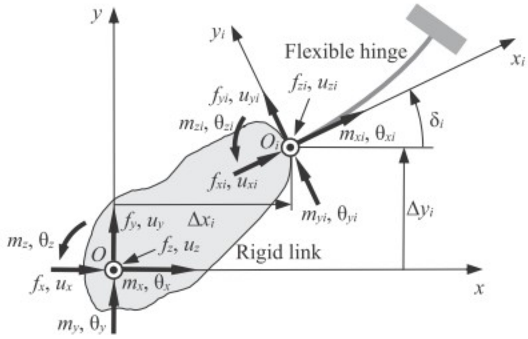

FIGURE 4.1 Rigid link connected to planar curvilinear-axis flexible hinge in global and local reference frames under the action of load displaced from its free end.

Assume that a load formed of f_x, f_y, m_z, m_x, m_y and f_z is applied at a point O that is rigidly connected to the free end O_i of a generic curvilinear-axis flexible hinge, which is fixed at the opposite end, as illustrated in Figure 4.1. The aim is to evaluate the displacements u_x, u_y, θ_z, θ_x, θ_y and u_z in the global reference frame $Oxyz$ in terms of the load at O and to express the flexible-hinge compliance matrices with respect to O based on the hinge native compliance matrices that are formulated in the $O_ix_iy_iz_i$ local frame. The load/displacement vectors are formally separated into in-plane and out-of-plane components.

4.1.1 IN-PLANE COMPLIANCES

The in-plane loads f_x, f_y and m_z statically transfer to O_i from their original location O as:

$$\begin{cases} f_{xi} = f_x \cdot \cos\delta_i + f_y \cdot \sin\delta_i \\ f_{yi} = -f_x \cdot \sin\delta_i + f_y \cdot \cos\delta_i \\ m_{zi} = f_x \cdot \Delta y_i - f_y \cdot \Delta x_i + m_z \end{cases} . \tag{4.1}$$

Equation (4.1) can be written in matrix form as:

$$\left[f^{(i)}_{O_i,ip} \right] = \left[R^{(i)} \right] \left[T^{(i)}_{OO_i,ip} \right] \left[f^{(i)}_{O_i,ip} \right]$$

$$\text{with}: \begin{cases} \left[f^{(i)}_{O_i,ip} \right] = \begin{bmatrix} f_{xi} & f_{yi} & m_{zi} \end{bmatrix}^T; \quad \left[f^{(i)}_{O,ip} \right] = \begin{bmatrix} f_x & f_y & m_z \end{bmatrix}^T \\ \left[R^{(i)} \right] = \begin{bmatrix} \cos\delta_i & \sin\delta_i & 0 \\ -\sin\delta_i & \cos\delta_i & 0 \\ 0 & 0 & 1 \end{bmatrix}; \quad \left[T^{(i)}_{OO_i,ip} \right] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \Delta y_i & -\Delta x_i & 1 \end{bmatrix} \end{cases} . \tag{4.2}$$

where $[R^{(i)}]$ is the rotation matrix and $\left[T^{(i)}_{OO_i,ip}\right]$ is the in-plane translation matrix.

The global-frame displacements u_x, u_y and θ_z are expressed in terms of the local-frame displacements u_{xi}, u_{yi} and θ_{zi} as follows:

$$\begin{cases} u_x = u_{xi} \cdot \cos\delta_i - u_{yi} \cdot \sin\delta_i + \Delta y_i \cdot \theta_{z_i} \\ u_y = u_{xi} \cdot \sin\delta_i + u_{yi} \cdot \cos\delta_i - \Delta x_i \cdot \theta_{z_i} \, . \\ \theta_z = \theta_{x_i} \end{cases} \tag{4.3}$$

Equations (4.3) are rewritten in the matrix form as:

$$\left[u^{(i)}_{O,ip}\right] = \left[T^{(i)}_{OO_i,ip}\right]^T \left[R^{(i)}\right]^T \left[u^{(i)}_{O_i,ip}\right] \text{ with } \begin{cases} \left[u^{(i)}_{O_i,ip}\right] = \begin{bmatrix} u_{xi} & u_{yi} & \theta_{zi} \end{bmatrix}^T \\ \left[u^{(i)}_{O,ip}\right] = \begin{bmatrix} u_x & u_y & \theta_z \end{bmatrix}^T \end{cases} . \tag{4.4}$$

The local-frame loads and displacements are related by means of the hinge in-plane compliance matrix by the known relationship:

$$\left[u^{(i)}_{O_i,ip}\right] = \left[C^{(i)}_{O_i,ip}\right]\left[f^{(i)}_{O_i,ip}\right]. \tag{4.5}$$

Combining Eqs. (4.2), (4.4) and (4.5) yields:

$$\left[u^{(i)}_{O,ip}\right] = \left[C^{(i)}_{O,ip}\right]\left[f^{(i)}_{O,ip}\right] \quad \text{with} \quad \left[C^{(i)}_{O,ip}\right] = \left[T^{(i)}_{OO_i,ip}\right]^T \left[R^{(i)}\right]^T \left[C^{(i)}_{O_i,ip}\right]\left[R^{(i)}\right]\left[T^{(i)}_{OO_i,ip}\right]. \tag{4.6}$$

where $\left[C^{(i)}_{O,ip}\right]$ is the global-frame, in-plane compliance matrix of the generic curvilinear-axis flexible hinge.

4.1.2 OUT-OF-PLANE COMPLIANCES

A similar procedure is applied to determine the relationship between global-frame out-of-plane loads and displacements. The out-of-plane loads m_x, m_y and f_z are statically reduced (moved) to O_i from their original location O as:

$$\begin{cases} m_{xi} = m_x \cdot \cos\delta_i + m_y \cdot \sin\delta_i + f_z \cdot \left(\Delta x_i \cdot \sin\delta_i - \Delta y_i \cdot \cos\delta_i\right) \\ m_{yi} = -m_x \cdot \sin\delta_i + m_y \cdot \cos\delta_i + f_z \cdot \left(\Delta x_i \cdot \cos\delta_i + \Delta y_i \cdot \sin\delta_i\right). \\ f_{zi} = f_x \end{cases} \tag{4.7}$$

which are collected in the matrix form:

$$\left[f^{(i)}_{O_i,op}\right]=\left[R^{(i)}\right]\left[T^{(i)}_{OO_i,op}\right]\left[f^{(i)}_{O,op}\right]$$

$$\text{with}: \begin{cases} \left[f^{(i)}_{O_i,op}\right]=\begin{bmatrix} m_{xi} & m_{yi} & f_{zi} \end{bmatrix}^T; \ \left[f^{(i)}_{O,op}\right]=\begin{bmatrix} m_x & m_y & f_z \end{bmatrix}^T \\ \left[T^{(i)}_{OO_i,op}\right]=\begin{bmatrix} 1 & 0 & -\Delta y_i \\ 0 & 1 & \Delta x_i \\ 0 & 0 & 1 \end{bmatrix} \end{cases}. \tag{4.8}$$

The rotation matrix $[R^{(i)}]$ is calculated in Eq. (4.2) and $\left[T^{(i)}_{OO_i,op}\right]$ is the out-of-plane translation matrix. The displacements θ_x, θ_y and u_z are expressed in terms of the displacements θ_{xi}, θ_{yi} and u_{zi} as:

$$\begin{cases} \theta_x=\theta_{xi}\cdot\cos\delta_i-\theta_{yi}\cdot\sin\delta_i \\ \theta_y=\theta_{xi}\cdot\sin\delta_i+\theta_{yi}\cdot\cos\delta_i \\ u_z=\theta_{xi}\cdot\left(\Delta x_i\cdot\sin\delta_i-\Delta y_i\cdot\cos\delta_i\right)+\theta_{yi}\cdot\left(\Delta x_i\cdot\cos\delta_i+\Delta y_i\cdot\sin\delta_i\right)+u_{zi} \end{cases}. \tag{4.9}$$

Equations (4.9) can also be written in matrix form as:

$$\left[u^{(i)}_{O,op}\right]=\left[T^{(i)}_{OO_i,op}\right]^T\left[R^{(i)}\right]^T\left[u^{(i)}_{O_i,op}\right] \quad \text{with} \quad \begin{cases} \left[u^{(i)}_{O_i,op}\right]=\begin{bmatrix} \theta_{xi} & \theta_{yi} & u_{zi} \end{bmatrix}^T \\ \left[u^{(i)}_{O,ip}\right]=\begin{bmatrix} u_x & u_y & \theta_z \end{bmatrix}^T \end{cases}. \tag{4.10}$$

By definition, the out-of-plane compliance matrix connects the corresponding loads and displacements in the local frame as:

$$\left[u^{(i)}_{O_i,op}\right]=\left[C^{(i)}_{O_i,op}\right]\left[f^{(i)}_{O_i,op}\right]. \tag{4.11}$$

Equations (4.8), (4.10) and (4.11) are combined into:

$$\left[u^{(i)}_{O,op}\right]=\left[C^{(i)}_{O,op}\right]\left[f^{(i)}_{O,op}\right]\text{with}\left[C^{(i)}_{O,op}\right]=\left[T^{(i)}_{OO_i,op}\right]^T\left[R^{(i)}\right]^T\left[C^{(i)}_{O_i,op}\right]\left[R^{(i)}\right]\left[T^{(i)}_{OO_i,op}\right]. \tag{4.12}$$

where $\left[C^{(i)}_{O,op}\right]$ is the global-frame, out-of-plane compliance matrix of the generic curvilinear-axis flexure hinge.

It is important to note that the transformation equations derived in this section for a curvilinear-axis hinge are also valid for a straight-axis hinge. The only difference is that the in-plane and out-of-plane, local-frame compliance matrices of a straight-axis hinge need to be used instead of those of a curvilinear-axis hinge.

1 Introduction

This introductory chapter gives a brief description of flexible hinges and flexure-based compliant mechanisms for macro- and micro-scale applications by highlighting the main traits of these mechanical members/devices. An outline of the subject matter and the associated approach in this book is also sketched here in order to identify and possibly locate the work in the context of similar dedicated information that has already been published.

Derived from the term *compliance*, utilized by Paros and Weisbord [1] to identify the reciprocal of stiffness, the *compliant* qualifier is utilized in this book to identify any mechanical device that comprises at least one elastically deformable joint, called *flexible* (or *flexure*) *hinge*. Such a mechanism usually includes other rigid links, which ensure the necessary rigidity and serve as input/output ports. The flexible-hinge *mechanisms* employ their mobility to perform conversions between electric, thermal, magnetic, etc. and mechanical energies. Many of these mechanisms, which are typically of monolithic architecture, are designed to undergo relatively low levels of rotation/deformation in order to generate accurate, quasi-continuous output motion. This work covers the flexure-based compliant mechanisms possessing all these traits.

Figure 1.1 shows the schematic representation of two microaccelerometers – sensing microdevices used to evaluate external acceleration. Both devices have a planar structure and comprise rigid links (fixed or movable) that are connected by

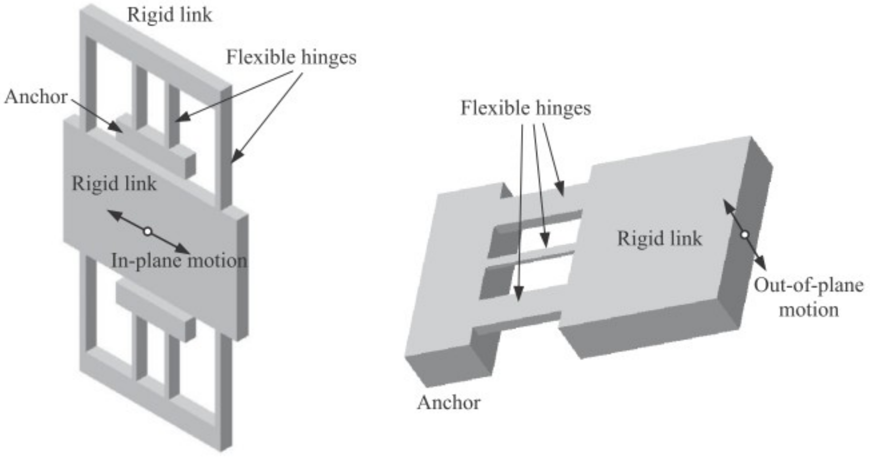

(a)　(b)

FIGURE 1.1 Flexible-hinge microaccelerometers with: (a) in-plane motion; (b) out-of-the-plane motion.

flexible hinges of (constant) rectangular cross-section. The thinner in-plane cross-sectional dimension of the eight hinges of the microaccelerometer, as shown in Figure 1.1a, also *known as* folded-beam accelerometer (Bhushan [2]), enables the central rigid link (the proof mass) to undergo in-plane, back-and-forth translatory motions when its anchor links are attached to an external vibrating source along the motion direction. Alternatively, in order to sense acceleration along a direction perpendicular to the device plane, the design of Figure 1.1b, named resonant-beam accelerometer (see Tudor et al. [3]), utilizes flexible hinges whose out-of-plane dimension is smaller than the in-plane cross-sectional thickness; consequently, these hinges can bend along an in-plane axis and enable the mobile rigid link (also known as seismic mass) to displace out of the plane. Both devices use additional sensing (capacitive, for instance) to convert the mechanical acceleration into electrical voltage.

Flexible hinges are largely utilized in the auto/aviation industry in applications such as acceleration/speed/position sensors, adaptable seats, airbags, fault-tolerant connectors, single-surface-independent aircraft control devices, actuators for configurable-geometry foils, steering columns, antifriction bearings, suspension systems, satellite small-angle tilting mechanisms, laser-beam communication systems between spacecraft, and flexible couplings. The biomedical industry is also a beneficiary of mechanisms that are based on flexible hinges, and applications in this category include devices for vascular catheters, urethral compression devices, intravascular endoprostheses, cardiac massage apparatuses, orthotic devices and biopsy devices. The computers and fiber optics industries have also applications that incorporate flexible hinges, and the examples in this category include disk drive suspensions, laser systems, optical mirrors, optical disks, microscopes, cameras, print heads, optical scanning equipment, vibrating beam accelerometers, keyboard assemblies, kinematic lens mountings and rotary actuators for disk drives. Various other fields utilize flexible-hinge mechanisms, and a few application examples are coin packaging systems, systems for remotely playing percussion musical instruments, collapsible fishing net mechanisms, table tennis ball-retrieving systems, snow blade attachments, foot propulsion devices for float tube users, bicycle seats, steerable wheels for skate/roller ski, respiratory masks, grinding/polishing machines, fluid jet cutting machines and flywheels. The compliant micro-/nano-electromechanical systems (MEMS/NEMS) are almost entirely based on microdevices that generate their motions by means of flexible hinges. Examples in this industry comprise optical switches, miniature load cells, flexible mounts for imaging masks, load-sensitive resonators, gyroscopes, gravity gradiometers, disk memory head positioners, wire bonding heads, microfluidic devices, accelerometers, scan modules for bar code readers and cantilevers for microscopy.

A flexible hinge is a relatively thin member that provides the relative rotation between two adjacent rigid links mainly through flexing (bending), as sketched in Figure 1.2, where a conventional rotational joint is shown next to a flexible hinge.

In terms of this rotary function, a flexible hinge can be considered the structural correspondent of a bearing with limited rotation capability, as illustrated in

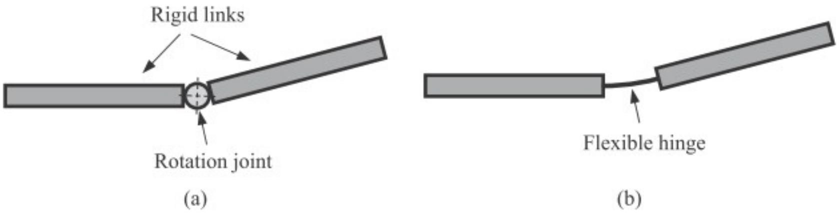

FIGURE 1.2 Joints enabling relative rotation in mechanisms: (a) conventional rotation joint; (b) skeleton representation of a generic flexible hinge.

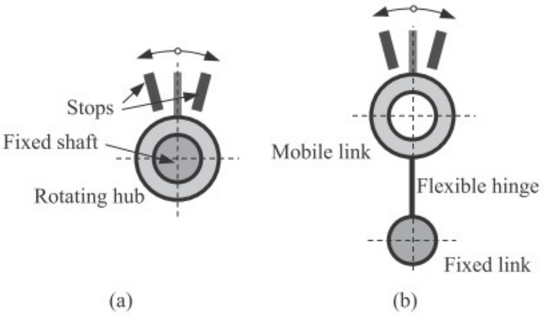

FIGURE 1.3 Functional similarity between rotary bearings and flexible hinges: (a) collocated (concentric) rotation produced by a rotary bearing; (b) non-collocated rotation produced by a flexible hinge.

Figure 1.3. In a classical rotary bearing, the relative rotation takes place between a shaft and its housing, these mating parts are concentrically located, and the rotation can be limited to a specific angular sector, as indicated in Figure 1.3a. A flexible hinge can provide a similar limited rotary output, and the only difference is attributable to the fact that the "centers" of the two adjacent members undergoing the relative rotation are no longer collocated, as shown in Figure 1.3b.

Physically, a flexible hinge can mainly be achieved in two different ways:

- Use an independently fabricated member (like a strip or shim in two-dimensional applications or a cylinder-like part in three-dimensional applications) to connect two rigid members that are designed to undergo a relative rotation;
- Generate the flexible hinge by either subtractive or additive fabrication procedures. In this case, the flexure is integral (or monolithic) to the rigid links it connects.

Among the benefits provided by the flexible hinges, the most notable are:

- No friction losses;
- No need of lubrication;

- No hysteresis;
- Compactness;
- Capacity to be utilized in small-scale applications;
- Ease of fabrication;
- Virtually no maintenance needed;
- No need of repair since the mechanism will operate until something will fail (usually the hinges) because of fatigue or overloading;
- Producing small-increment, quasi-continuous and very precise motion.

Flexible hinges do have limitations, however, and a few examples of such drawbacks are:

- The hinges, particularly the notch ones, provide relatively low levels of rotations due to stress limitations;
- The rotation is not pure because the deformation of a flexible hinge is complex as it is produced by axial loading, shear and torsion in addition to bending;
- The flexible hinge is usually sensitive to temperature variations, and therefore, its dimensions change as a result of thermal expansion/contraction, which leads to modifications in the original hinge elastic properties.

For two-dimensional applications where the flexible hinge is fabricated by removing material from a blank piece, the manufacturing processes that are being utilized for this purpose include end milling, electrodischarge machining (EDM), laser cutting, water/liquid jetting, metal stamping, three-dimensional printing and photolithographic/microfabrication techniques for MEMS/NEMS.

In two-dimensional, planar flexible mechanisms, the hinge is designed to be compliant about a single axis only (the input, compliant or sensitive axis), along which the relative rotation between the adjacent rigid links is taking place, and stiff (as much as possible) about all other axes and motions – see the designs of Figure 1.4a and c. These configurations are also known as *single-axis* flexible hinges. For other hinge architectures that have rotational symmetry (they are revolute and their cross-section is circular), like the one pictured in Figure 1.4b, bending is indiscriminately possible about any axis that is perpendicular to the axial direction, and therefore, the compliant axis can instantaneously be set by the loading/boundary conditions of the three-dimensional compliant mechanism. Consistent with the terminology that has already been utilized such flexible hinges are also called *multi-axis*. Single-axis flexible hinges are specifically intended for two-dimensional compliant mechanisms that have in-plane and out-of-plane motion, whereas multi-axis hinges are implemented in three-dimensional applications in order to take advantage of their capacity to produce relative rotation about two or more compliant axes.

While the designs of Figure 1.4a and b have straight longitudinal axes, flexible hinges with curvilinear longitudinal axis, like the one shown in Figure 1.4c, have also been designed and utilized in various applications.

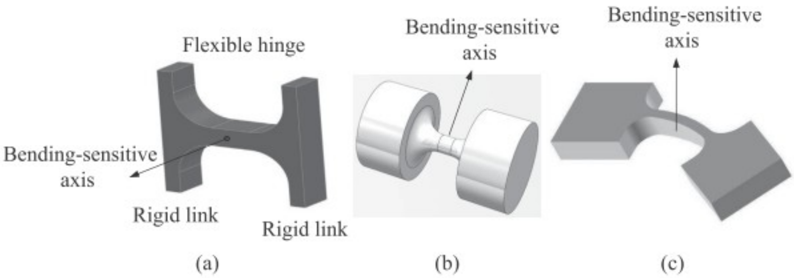
Flexible hinge
Bending-sensitive axis
Rigid link
Rigid link
Bending-sensitive axis
Bending-sensitive axis
(a)
(b)
(c)

FIGURE 1.4 (a) Straight-axis flexible hinge of rectangular cross-section; (b) straight-axis flexible hinge of circular cross-section; (c) curvilinear-axis flexible hinge of rectangular cross-section.

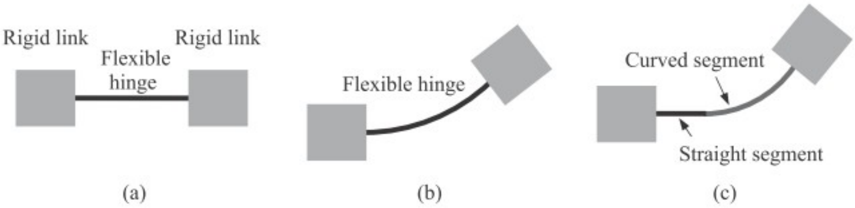
Rigid link
Flexible hinge
Rigid link
Flexible hinge
Curved segment
Straight segment
(a)
(b)
(c)

FIGURE 1.5 Skeleton representation of: (a) straight-axis flexible hinge; (b) curvilinear-axis flexible hinge; (c) flexible hinge with straight- and curvilinear-axis segments.

Oftentimes, it is convenient to use a simplified, skeleton representation of flexible hinges by just drawing the longitudinal axis of the hinge, as illustrated in Figure 1.5.

A considerable number of compliant mechanisms utilize a reduced pool of flexible-hinge configurations. Apart from the constant cross-section (either rectangular or circular) designs that present the inconvenience of high stress concentration at the areas joining the rigid links, two other geometries – the circular design (Paros and Weisbord [1]) and the right circularly corner-filleted configuration (Lobontiu et al. [4]) (these basic configurations are sketched in Figure 1.6) – occupy an ample space of choices for flexible-hinge applications. The simplicity of their geometry and the relative ease of fabrication are two main reasons that prompted the widespread use of these two types of hinges. In

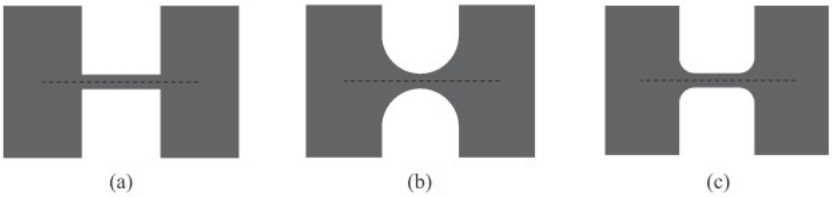

FIGURE 1.6 Commonly utilized flexible-hinge configurations with straight longitudinal axis: (a) constant cross-section; (b) right circular; (c) right circularly corner-filleted.

addition, both configurations come with the advantage of being able to reduce the stress concentration levels through the filleted regions at their corners (ends).

A thorough analytical presentation of single-axis and two-axis circular flexure hinges is provided in Paros and Weisbord [1]. The authors of this paper interchangeably utilized the terms *flexure hinge* and *flexure* to denote a mechanical member, which is "compliant in bending about one axis but rigid about the cross-axes", and mentioned that flexure hinges are incorporated in applications where angular motion is required about an axis – the compliant axis. The key merit of this seminal work consisted in the clear mathematical definition of a flexure hinge as a spring element, which displays two distinct behaviors: it is compliant about one axis, in order to produce the desired rotation, and stiff (as much as possible) in all other motions about the other axes in order to prevent/minimize the respective parasitic motions. This paper provided analytical equations, both exact and approximate, for the spring rates (reciprocal to compliances) of single-axis and two-axis circular symmetric flexures in terms of the motions generated through the consideration of bending and axial effects. The alternative terminology of *flexural pivot* has also been utilized over the years in referring to the same mechanical member. The *notch hinge* has also been employed by several authors; see Smith [5], for instance, especially for single-axis flexible hinges. While the term *flexure* aptly depicts the primary function and operation mode of an elastic joint to flex (bend), there are also hinge configurations whose main intended/functional deformation is torsional. In order to cover all possible deformation modes, the terminology *flexible hinge* is used throughout this book. The term *compliance* (also known as *flexibility*) is utilized here in the original definition given in Paros and Weisbord [1] such that the term refers to the quantity which is the reciprocal of one flexure's spring rate (or stiffness). As a terminology extension from *compliance* – see also previous mention at the beginning of this chapter – the qualifier *compliant* is used in this book to denote the behavior of a mechanical device/mechanism that moves by the elastic deformation of its flexible hinges in order to realize a design objective (actuation or sensing).

Expanding the flexible-hinge design domain with new geometries and configurations is prompted by practical need. The compliant behavior of flexible hinges and flexure-based mechanisms largely depend on the specific geometry of the hinge for a given material. Slight alterations in geometry and dimensions can sensibly modify the elastic response. This aspect is particularly important in mechanisms where precision is a key performance parameter or where a finely tuned, quasi-continuous output is expected in terms of displacement, force or frequency (resonant) response. This book provides compliance equations for a large variety of flexible-hinge configurations that have been developed and reported in the specialty literature over the past years or which are new. It also qualifies and compares the flexible hinges in terms of level and precision of the output motion, sensitivity to parasitic (undesired) loading and stress levels.

The book essentially develops not only a process of discretization through which the compliance (stiffness) is primarily derived, but also studies the inertia and damping characteristics that are derived for a large variety of flexible hinges

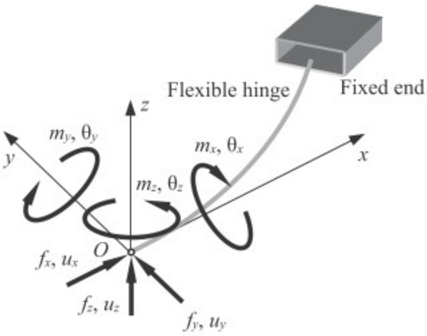
Flexible hinge
Fixed end
z
m_y, θ_y
y
m_x, θ_x
x
m_z, θ_z
O
f_x, u_x
f_z, u_z
f_y, u_y

FIGURE 1.7 Skeleton fixed-free flexible hinge with six DOF at the free end.

both analytically and by means of the finite element technique. With the analytical approach, a distributed-parameter flexible hinge is equivalently transformed into a discrete one with properties defined individually and independently in terms of its degrees of freedom (DOF). For a three-dimensional flexible hinge, one of its ends can move with respect to the other one (presumed fixed) by three translations and three rotations, respectively, as sketched in Figure 1.7. Consequently, a flexible hinge can be modeled as a six-DOF member in the three-dimensional space when only the position of one end relative to the opposite one is of interest.

Planar flexible hinges of rectangular cross-sections can be designed to operate in their original plane, as the microaccelerometer of Figure 1.1a does, or out of their plane, as the case is with the mechanism of Figure 1.1b.

For in-plane operation, a planar flexible hinge, like the curvilinear-axis one shown in Figure 1.8a, has its in-plane dimension thinner than its (constant) out-of-plane dimension w. As a result, the free end O can translate along the x and y axes and rotate around the z axis. The quasi-static behavior of such a hinge, as shown in Chapters 2–4, is characterized by an *in-plane compliance matrix* connecting the load vector formed of the forces f_x and f_y and the moment m_z to the corresponding

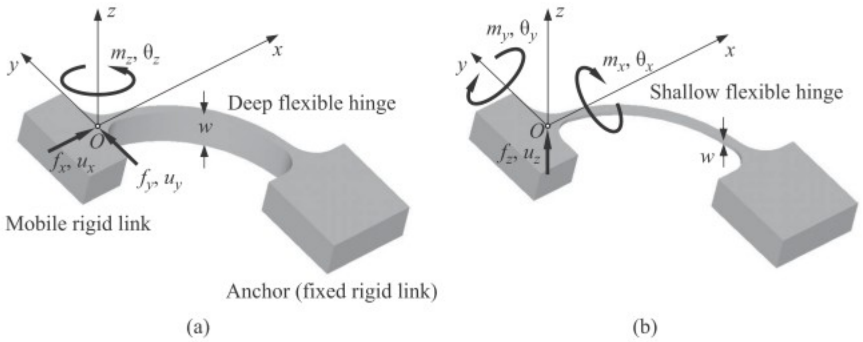
z
m_z, θ_z
x
y
Deep flexible hinge
w
O
f_x, u_x
f_y, u_y
Mobile rigid link
Anchor (fixed rigid link)
(a)
m_y, θ_y
z
m_x, θ_x
x
y
Shallow flexible hinge
O
f_z, u_z
w
(b)

FIGURE 1.8 Curvilinear-axis flexible hinges for: (a) in-plane operation; (b) out-of-plane operation.

displacement vector defined by the translations u_x and u_y and the rotation θ_z. The hinge deformation includes bending around the z axis, axial deformation along the x axis and shear along the y axis. Out-of-plane deformations produced by the other three loads, namely, m_x, m_y and f_z, are undesired and therefore parasitic.

Conversely, the flexible hinge shown in Figure 1.8b, with an out-of-plane dimension w smaller than the in-plane thickness, can easily deform out of its original structural plane; as such, an out-of-plane compliance matrix can be derived to connect the load vector defined by the moments m_x and m_y and the force f_z to the resulting displacement vector formed of the rotations θ_x and θ_y and displacement u_z. This particular hinge mainly deforms in bending around the y axis, torsion around the x axis and shear along the z axis. All other deformations resulting from the in-plane loads f_x, f_y and m_z are undesired (parasitic).

It should be noted again that flexible hinges are approached in this book as members that are subjected to small deformations. This assumption stems from the reality that in numerous engineering applications the flexible hinges actually behave according to this model, which enables utilization of the linear (or first-order) bending theory and derivation of the spring rates for various hinge configurations.

In addition to the compliance-based approach utilized to model the static and dynamic behaviors of flexible-hinge mechanisms, which is a variant of the *force* (or *flexibility matrix*) *method*, this book also studies the flexible hinges and their mechanisms by means of the finite element technique using the classical *displacement* (*stiffness matrix*) *method*. A flexible hinge of either straight or curvilinear/circular axis is modeled as a two- or three-node line element with six DOF per node. A generic finite element formulation is developed by defining the elemental stiffness, mass and damping matrices. Explicit forms of these elemental matrices are derived for a few flexible-hinge configurations. These equations are subsequently utilized to characterize the static or dynamic response of various flexible-hinge mechanism examples.

Several other related topics, including stress concentration, combined stresses and theories of failure, flexible-hinge precision of rotation, layered flexible hinges, temperature effects and piezoelectric actuation/sensing, are also studied in this book.

REFERENCES

1. Paros, J.M. and Weisbord, L., How to design flexure hinges, *Machine Design*, November, 151, 1965.
2. Bhushan, B. (editor), *Springer Handbook of Nanotechnology*, 3rd ed., Springer, Berlin, 2010.
3. Tudor, M.J. et al., Silicon resonator sensors: interrogation techniques and characteristics, *IEEE Proceedings on Control Theory and Applications*, 135 (5), 364, 1988.
4. Lobontiu, N. et al., Corner-filleted flexure hinges, *ASME Journal of Mechanical Design*, 123, 346, 2001.
5. Smith, S.T., *Flexures – Elements of Elastic Mechanisms*, Gordon and Breach Science Publishers, Amsterdam, 2000.

+,- * *. !* ! / 0 ! 1#
) 9< " &

6

Chords and Circular Segments

This chapter discusses chords, including theorems that relate to intersecting chords, and also explores the area of a circular segment which is formed by a chord.

Chapter 6 Concepts

A **chord** (pronounced with a silent "h" so that it sounds like "cord") is a line segment that connects two points that lie on the circumference of a circle. A **circular segment** is a region enclosed by an arc and a chord. For example, in the diagram on the left below, $\overline{AB}$ is a chord and the shaded region between $\overline{AB}$ and arc length s is a circular segment.

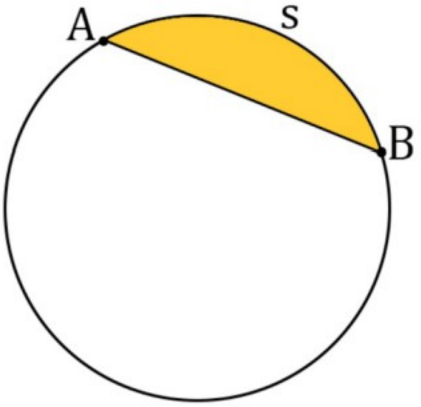

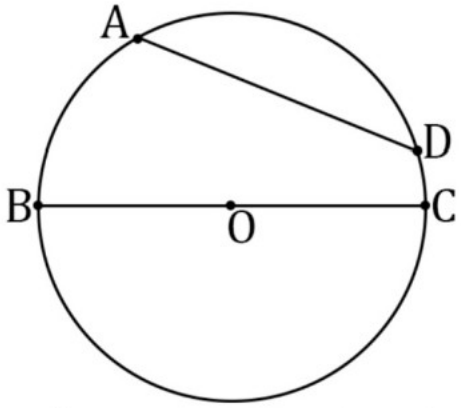

A diameter is a special chord. A **diameter** is a chord that passes through the center of the circle. A chord that does not pass through the center of the circle is shorter than the diameter. For example, in the diagram above on the right, point O lies at the center of the circle, $\overline{BC}$ is a diameter, $\overline{AD}$ is a chord that does not pass through O, and $AD < BC$.

If two **chords** in the same circle are **congruent**, the chords subtend **congruent arcs**. For example, if $\overline{AD} \cong \overline{BC}$ in the diagram below, then arc lengths s and t are equal.

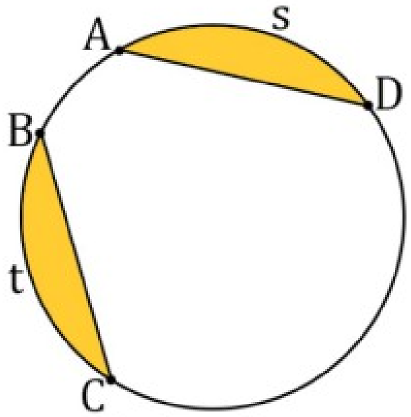

Tip: When working with chords or circular segments, it is often helpful to **draw line segments connecting the endpoints of the chord to the center of the circle**. Since the endpoints of the chord lie on the circumference of the circle, each line segment that connects an endpoint of a chord to the center of a circle is a radius. A triangle bounded by a chord and two radii is an isosceles triangle. For example, in the diagram on the next page, points A and B are the endpoints of chord $\overline{AB}$, point O lies at the center of the circle, $\overline{AO}$ and $\overline{BO}$ are radii, ∠AOB is a central angle, and ΔABO is isosceles.

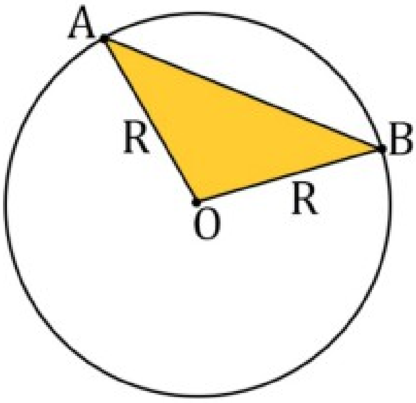

Note the distinction between a **circular segment** and a **sector** (Chapter 3). A circular segment refers to the region between a chord and an arc (left figure on the next page), whereas a sector is a pie slice cut from the center of the circle (right figure on the next page). The **area** of the **circular segment** shown below on the left can be found by subtracting the area of ΔABO from the area of the sector shown on the right. Recall from Chapter 3, Problem 7 that the **area** of a **sector** is $\frac{\theta R^2}{2}$, provided that θ (which is ∠AOB in the diagram below) is expressed in radians. Recall that π rad = 180°. Also recall that the area of a triangle is $\frac{bh}{2}$ (Volume 1, Chapter 6).

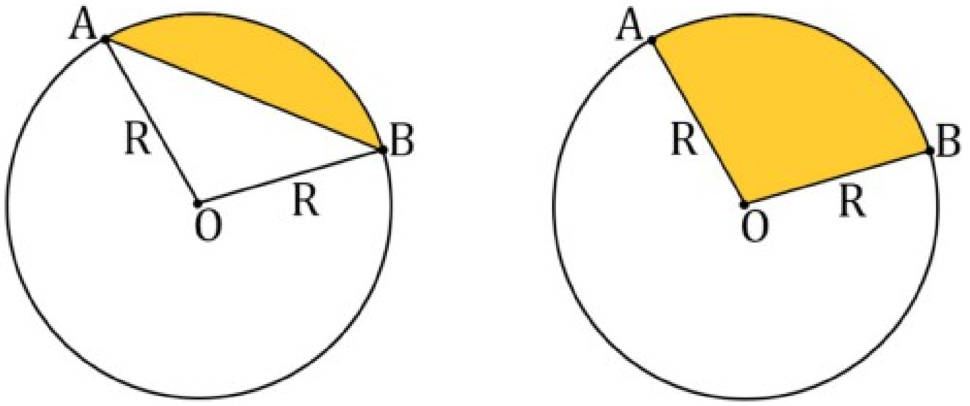

When **two chords** of the same circle **intersect**, the angle between the chords is equal to the average value of the two central angles that subtend the same arcs. For example, chords $\overline{AC}$ and $\overline{BD}$ intersect at point E in the diagram below, $\alpha \cong \gamma$ (vertical angles), point O lies at the center of the circle, and central angles θ and φ subtend the same arcs (s and t) as α and γ. Angles α, γ, θ, and φ are related by $\alpha = \gamma = \frac{\theta+\varphi}{2}$.

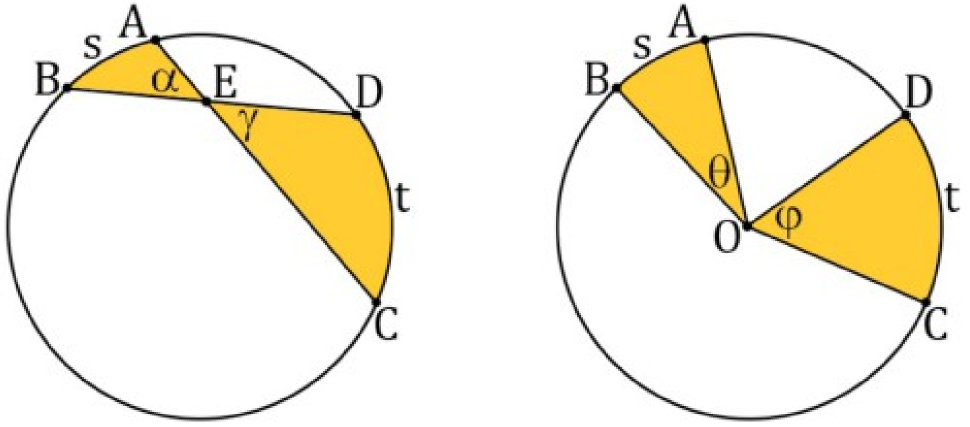

According to the **intersecting chords theorem**, the distances in the left diagram above are related by the following formula, which involves multiplying distances together: $(AE)(CE) = (BE)(DE) = R^2 - EO^2$. As shown below, EO is the distance from the point where the chords intersect (E) to the center of the circle (O). The radius of the circle is R.

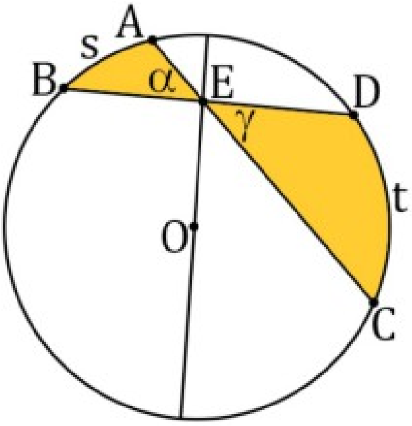

Notation: In this book, a lowercase letter of the English alphabet (like s or q) labeling an arc refers to a distance equal to the **arc length**. This book uses one of the following three common method to indicate an angle: an angle symbol followed by three uppercase English letters (like ∠ABC), an angle symbol followed by a number (like ∠1), or a lowercase Greek letter (like θ). A lowercase English letter (like s or q) in this book can only be a distance.

When reading a book that labels an **angle** on an arc, this is usually the central angle for the arc (even if the central angle is not drawn).

Chapter 6 Examples

Example 1. In the diagram below, $AB = 5\sqrt{2}$ and the diameter of the circle is 10. Find the central angle that subtends the (minor) arc from A to B.

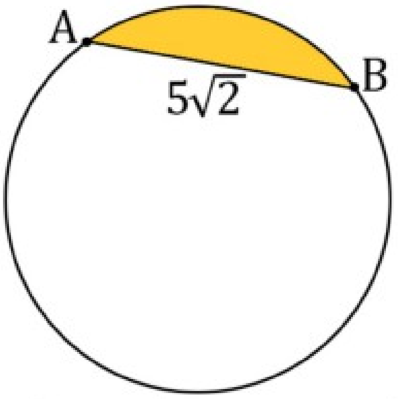

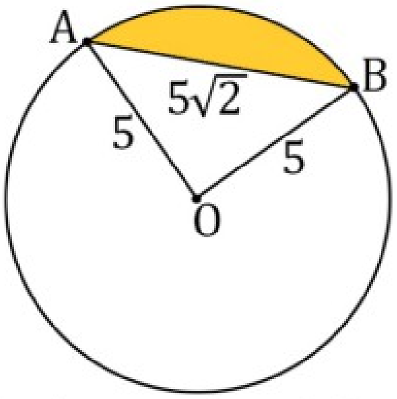

First, draw the center of the circle (point O) and draw line segments $\overline{AO}$ and $\overline{BO}$. Since $D = 10$, the radius is $R = \frac{D}{2} = \frac{10}{2} = 5$. Note that $AO = BO = R = 5$. The sides of ΔABO come in the ratio $AO{:}BO{:}AB = 5{:}5{:}5\sqrt{2} = 1{:}1{:}\sqrt{2}$. (The ratio was reduced by dividing each side by 5.) Since the sides of a 45°-45°-90° triangle come in the ratio $1{:}1{:}\sqrt{2}$ (see Volume 1, Chapter 5), this shows that ΔABO is a 45°-45°-90° triangle. The hypotenuse (AB) is opposite to the right angle. Therefore, $\angle AOB = 90°$.

Example 2. Find the area of the shaded circular segment in the diagram below. As in Example 1, $\text{AB} = 5\sqrt{2}$ and the diameter of the circle is 10.

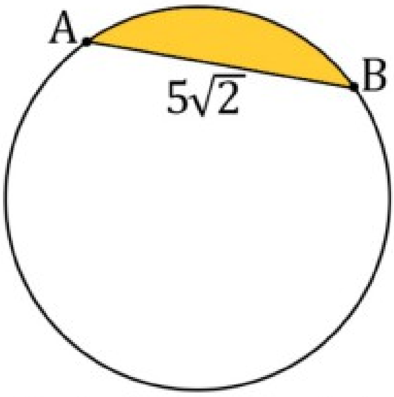

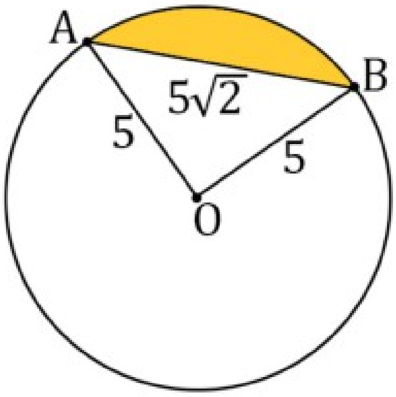

First find the area of the sector (Chapter 3). Recall from Example 1 that $\angle\text{AOB} = 90°$.

Since $360° \div 90° = 4$, the sector has $1/4$ of the area of a full circle: $\frac{\pi R^2}{4} = \frac{\pi 5^2}{4} = \frac{25\pi}{4}$. Since ΔABO is a right triangle, the two legs (AO and BO) serve as the base and height. The area of ΔABO equals $\frac{1}{2}(\text{AO})(\text{BO}) = \frac{1}{2}(5)(5) = \frac{25}{2}$. Subtract the area of ΔABO from the area of the sector to find the area of the circular segment: $A = \frac{25\pi}{4} - \frac{25}{2} = \frac{25}{4}(\pi - 2)$ is the exact area, which was found by factoring out $\frac{25}{4}$. Check the answer by distributing: $\frac{25}{4}(\pi - 2) = \frac{25}{4}\pi - \frac{25}{4}(2) = \frac{25}{4}\pi - \frac{25}{2}$. The area of the circular segment is approximately equal to

$\frac{25}{4}(3.14 - 2) \approx 7.13$ when rounded to three significant figures.

Example 3. The same circle is shown twice below. On the right, point O lies at the center of the circle. Determine α and γ.

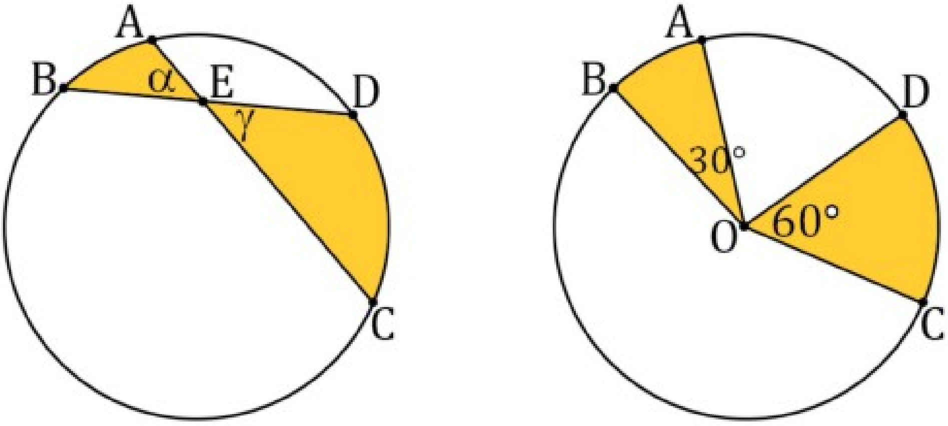

First, $\alpha \cong \gamma$ because α and γ are vertical angles. Second, $\overline{AC}$ and $\overline{BD}$ are intersecting chords, which means that α and γ are each equal to the average of the central angles that subtend the same arcs as α and γ.

$$\alpha = \gamma = \frac{\angle AOB + \angle COD}{2} = \frac{30° + 60°}{2} = \frac{90°}{2} = 45°$$

Example 4. In the diagram below, $AE = 2$, $BE = 4$, and $CE = 10$. Find DE.

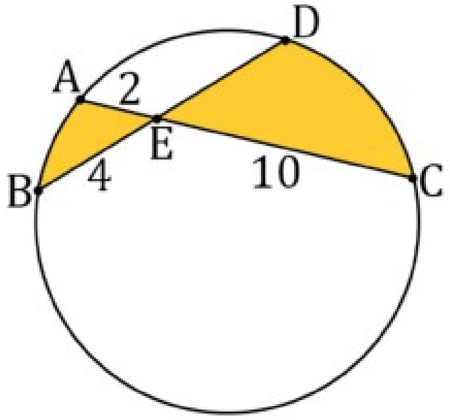

Apply the intersecting chords theorem:

$$(AE)(CE) = (BE)(DE)$$

$$(2)(10) = (4)(DE)$$

$$\frac{20}{4} = 5 = DE$$

Example 5. In the diagram below, $AD = 9$, $CD = 32$, and the radius of the circle equals 22. Find DO.

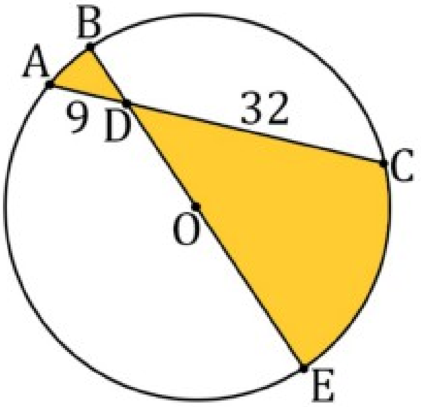

Apply the intersecting chords theorem:

$$(AD)(CD) = R^2 - DO^2$$

$$(9)(32) = 22^2 - DO^2$$

$$288 = 484 - \mathrm{DO}^2$$

$$\mathrm{DO}^2 + 288 = 484$$

$$\mathrm{DO}^2 = 484 - 288 = 196$$

$$\mathrm{DO} = \sqrt{196} = 14$$

Check: $(\mathrm{BD})(\mathrm{DE}) = (22 - 14)(22 + 14) = (8)(36) = 288$.

Example 6. In the diagram below, point O lies at the center of the circle and $\overline{\mathrm{AD}} \cong \overline{\mathrm{BC}}$. Prove that s ≅ t.

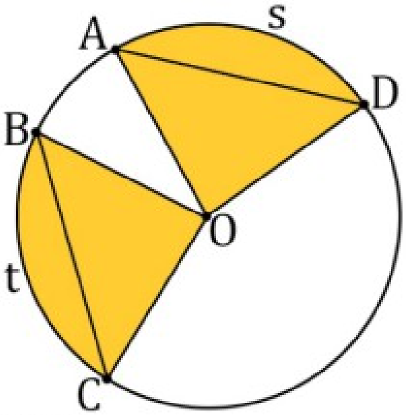

$\overline{\mathrm{AO}} \cong \overline{\mathrm{DO}} \cong \overline{\mathrm{BO}} \cong \overline{\mathrm{CO}}$ because each is a radius and $\overline{\mathrm{AD}} \cong \overline{\mathrm{BC}}$ is given. $\Delta\mathrm{ADO} \cong \Delta\mathrm{BCO}$ according to SSS (recall Volume 1, Chapter 3). $\angle\mathrm{AOD} \cong \angle\mathrm{BOC}$ via the CPCTC. Since arc length equals radius times the central angle (Chapter 3), it follows that s ≅ t (because $\angle\mathrm{AOD}$ and $\angle\mathrm{BOC}$ are the central angles that subtend arc lengths s and t). This proves that if two chords are congruent, the corresponding minor arcs are also congruent.

Chapter 6 Problems

Note: The diagrams are not drawn to scale.

1. In the diagram below, point O lies at the center of the circle, $AO = 12$, and $\angle AOB = 60°$. Determine the length of chord AB and the area of the shaded circular segment.

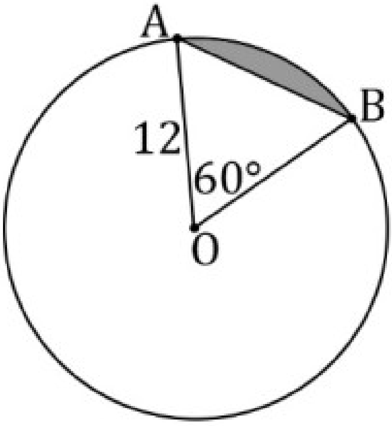

2. The diagram below shows a quarter circle. Determine the length of chord CD and the area of the shaded circular segment.

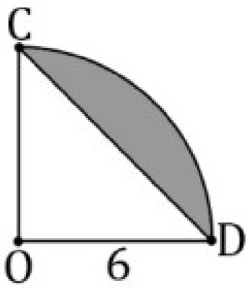

3. In the diagram below, $EF = \sqrt{3}$ and the radius of the circle is 1. Determine the central angle that subtends the minor arc from E to F and the area of the shaded circular segment.

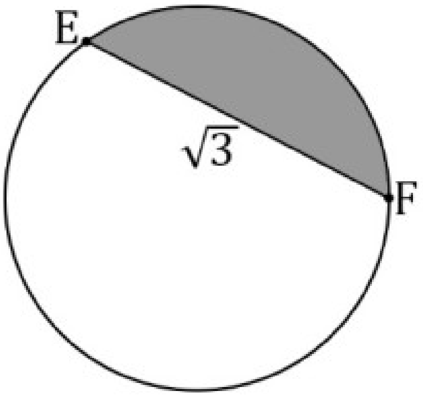

4. In the diagram below, DF is a diameter of the circle. Determine the length of chord DE and the combined area of the two shaded circular segments.

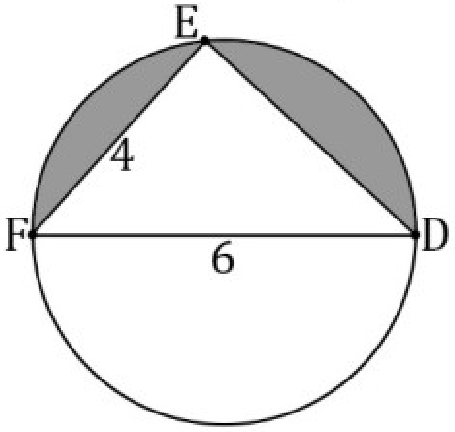

5. In the diagram below, point O lies at the center of the circle and the major arc length from K to L is 10π. Determine the length of chord KL and the area of the shaded region.

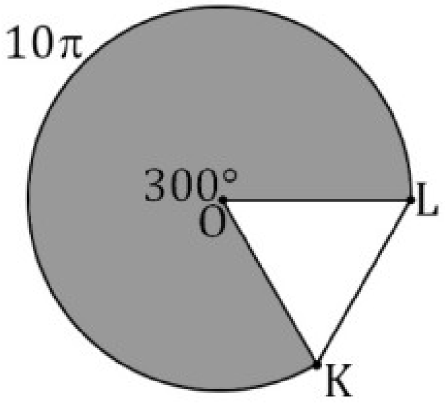

6. In the diagram below, point O lies at the center of the circle, $AO = 2$, and $\angle AOB = 30°$. Determine the length of chord AB and the area of the shaded circular segment. **Tip**: Draw a vertical line segment down from A to make two right triangles. Draw these triangles in a larger diagram off to the side.

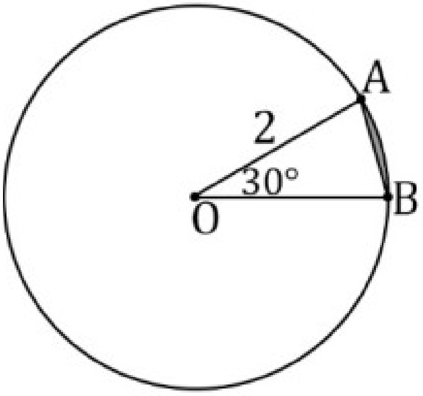

7. The diagram below shows square EFGH and two semicircles that have $\overline{EF}$ and $\overline{GH}$ as diameters. Find the combined area of the shaded regions.

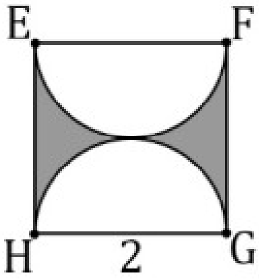

8. The diagram below shows square ABCD and two quarter circles that have points A and C as centers. Find the area of the shaded region.

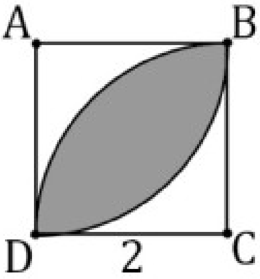

9. In the diagram below, ⊙W and ⊙X intersect at points Y and Z, ⊙W and ⊙X have the same radius (which is 6), and the radius of ⊙W and ⊙X is equal to YZ. Find the area of the shaded region.

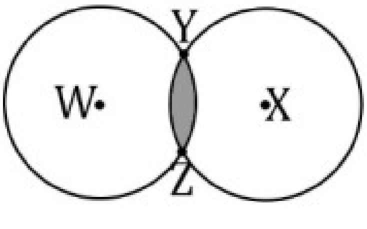

10. In the diagram below, ⊙P, ⊙Q, and ⊙R each have a radius of 3, ⊙P and ⊙Q touch only at point T, ⊙Q and ⊙R touch only at point U, and ⊙P and ⊙R touch only at point S. Find the area of the shaded region.

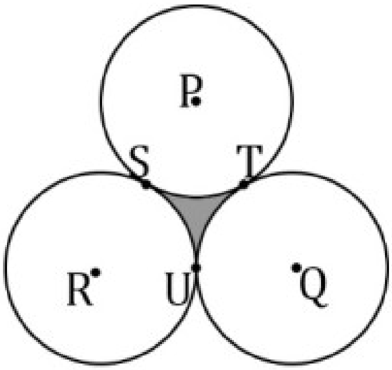

11. The diagram below shows ΔABC and three arcs of circles that have points D, E, and F as centers. Points D, E, and F are midpoints of their respective sides and AB = AC = BC = 2. Find the area of the shaded region.

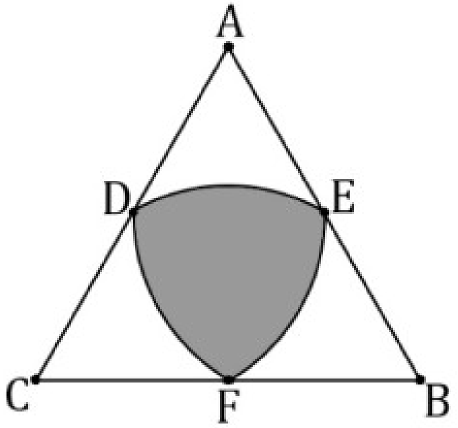

12. The crescent-shaped region formed between two circular arcs such as the shaded region below is called a **lune**. In the diagram below, LK and MN are the diameters of the two semicircles. Find the area of the shaded lune.

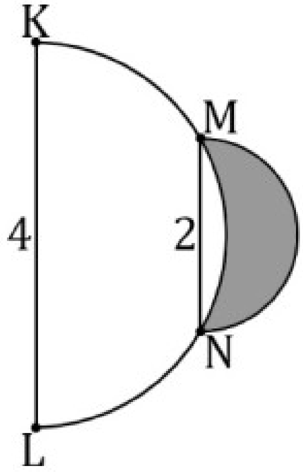

13. In the diagram below, point O lies at the center of the circle, $\overline{AD}$ is a diameter, and $AB = CD = 7$. Determine α, β, γ, δ, ε, φ, η, and κ.

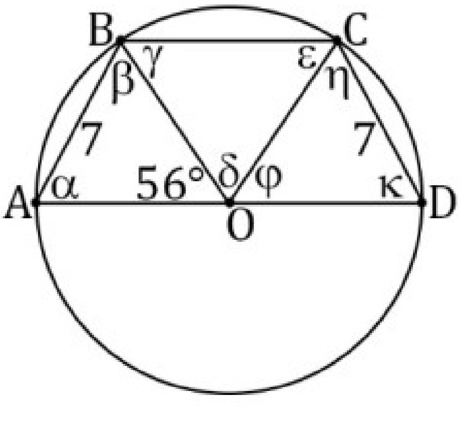

14. The same circle is shown twice below. On the right, point O lies at the center of the circle. Determine α, β, γ, and δ.

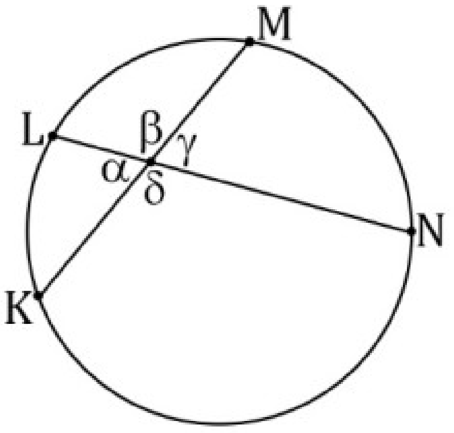

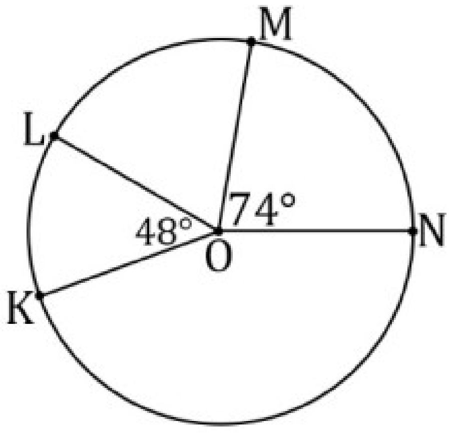

15. The same circle is shown twice below. On the right, point O lies at the center of the circle. Determine α.

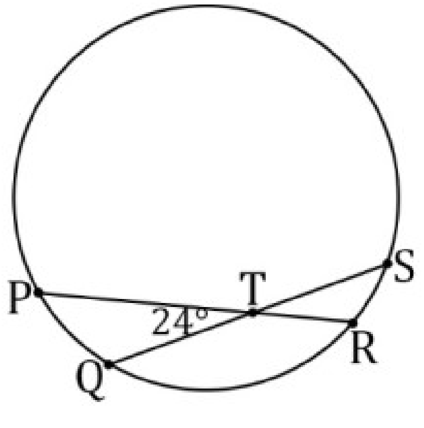

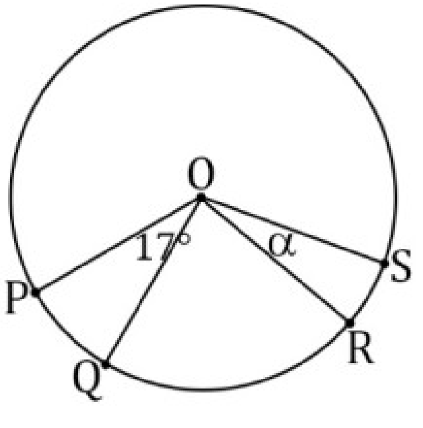

16. In the diagram below, $\angle AFB = 74°$, $\angle COD = 108°$, and point O lies at the center of the circle. Determine $\angle AOF$.

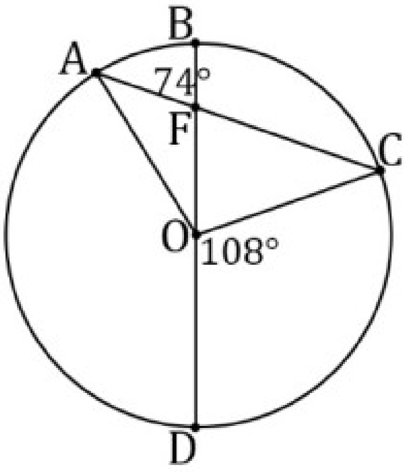

17. In the diagram below, $KN = 8$, $LN = 9$, $MN = 27$, and the diameter of the circle is 36. Find JN and the distance from N to the center of the circle.

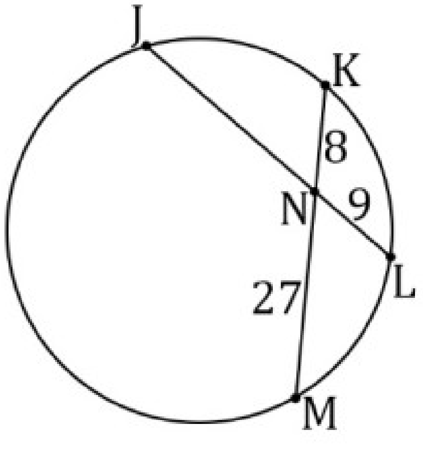

18. In the diagram below, $VX = 6$, $OV = 7$, $VZ = 12$, and point O lies at the center of the circle. Find VY, OY, and WY.

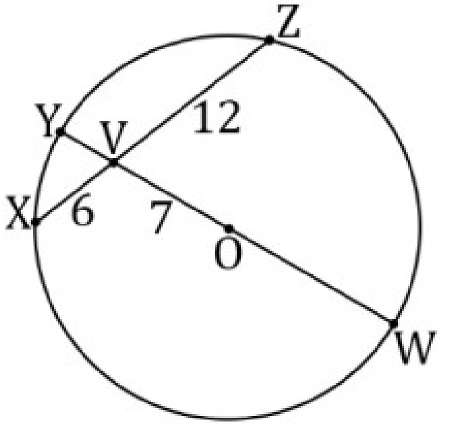

19. In the diagram below, point O lies at the center of the circle, $AE = 1$, $DO = \sqrt{2}$, and $\overline{AC} \perp \overline{BD}$. Find $\angle AOE$, $\angle COD$, AO, EO, and BE without using the formulas below. Use these answers and the information given to verify that $(BE)(DE) = (AE)(CE) = R^2 - EO^2$ and that $\angle BEA = \frac{\angle AOE + \angle COD}{2}$.

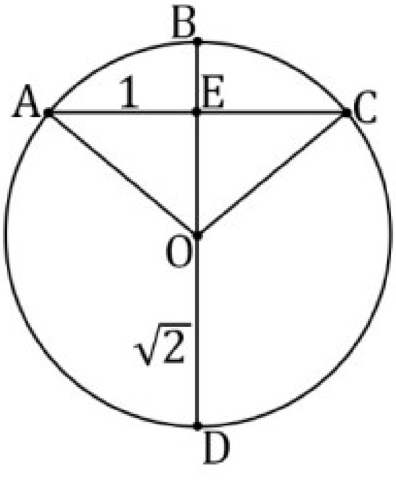

20. In the diagram below, $\overline{VW} \cong \overline{VZ}$. Prove that $\overline{WZ} \parallel \overline{XY}$. Note: Point V does NOT lie at the center of the circle and NONE of the chords shown is a diameter.

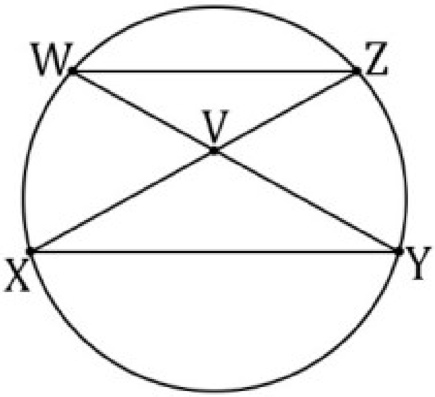

21. Prove that $\Delta ABE \sim \Delta CDE$ and $\Delta ADE \sim \Delta BCE$ in the diagram below. Notes: Point E may NOT lie at the center of the circle. Do NOT assume that any of the chords are parallel. Do NOT use the intersecting chords theorem. Hint: Review Chapter 4.

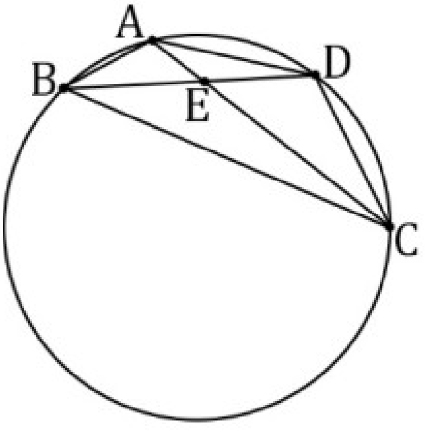

22. Prove that $(\mathrm{AE})(\mathrm{CE}) = (\mathrm{BE})(\mathrm{DE})$ for the diagram below. Notes: Point E may NOT lie at the center of the circle. Do NOT assume that any of the chords are parallel.

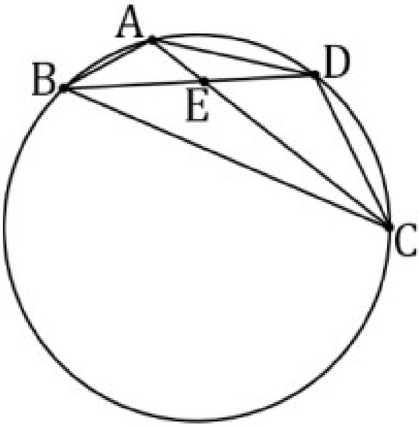

23. In the diagram below, point O lies at the center of the circle. Prove that $(\mathrm{AE})(\mathrm{CE}) = (\mathrm{BE})(\mathrm{DE}) = \mathrm{R}^2 - \mathrm{EO}^2$.

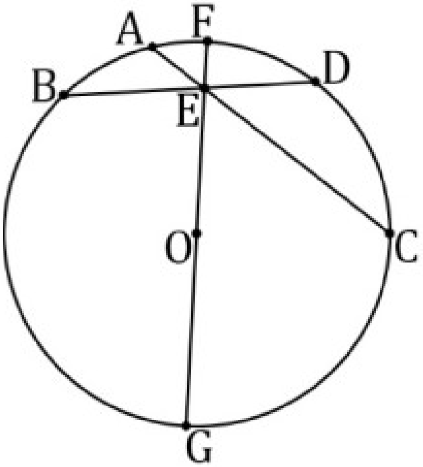

24. In the diagram below, point O lies at the center of the circle. Prove that $\angle AEB = \angle CED = \frac{\angle AOB + \angle COD}{2}$.

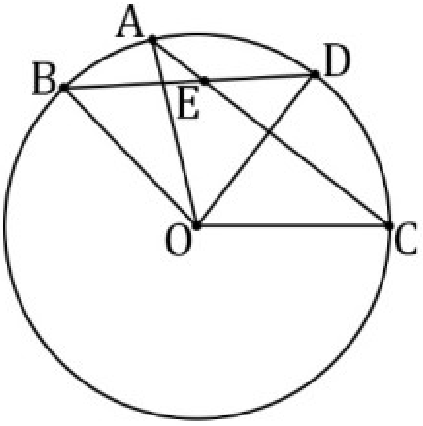

Chapter 6 Answers

1. $AB = 12$ and $A = 24\pi - 36\sqrt{3} = 12(2\pi - 3\sqrt{3}) \approx 13$. Notes: Since $360° \div 6 = 60°$, the area of the sector is $\frac{\pi R^2}{6} = \frac{\pi 12^2}{6} = \frac{144\pi}{6} = 24\pi$. Since $AO = BO = R$ and $\angle AOB = 60°$, it follows that ΔABO is an equilateral triangle. The edge length of ΔABO is the radius of the circle: $L = R = 12$. Recall from Volume 1, Chapter 6 that the area of an equilateral triangle with edge length L is $\frac{L^2\sqrt{3}}{4} = \frac{12^2\sqrt{3}}{4} = \frac{144\sqrt{3}}{4} = 36\sqrt{3}$. The area of the sector minus the area of the triangle equals the area of the circular segment:

$$A = 24\pi - 36\sqrt{3} = 12(2\pi - 3\sqrt{3})$$

In the last step, 12 was factored out. (Check this by distributing the 12.)

2. $CD = 6\sqrt{2}$ and $A = 9\pi - 18 = 9(\pi - 2) \approx 10.3$. Notes: The area of the quarter circle is $\frac{\pi R^2}{4} = \frac{\pi 6^2}{4} = \frac{36\pi}{4} = 9\pi$. Since $CO = DO = R$ and $\angle COD = 90°$, it follows that ΔCDO is a 45°-45°-90° triangle. The area of ΔCDO is $\frac{1}{2}bh = \frac{1}{2}R^2 = \frac{1}{2}(6)^2 = \frac{36}{2} = 18$. The area of the sector

minus the area of the triangle equals the area of the circular segment:

$$A = 9\pi - 18 = 9(\pi - 2)$$

In the last step, 9 was factored out. (Check this by distributing the 9.)

3. $\angle EOF = 120°$ (see the diagram on the next page) and $A = \frac{\pi}{3} - \frac{\sqrt{3}}{4} \approx 0.614$. Notes: Since EO:FO:EF = $1:1:\sqrt{3}$, ΔEOF (where O lies at the center) is a 30°-30°-120° triangle (recall Volume 1, Chapter 5, Problem 12). In the right diagram on the next page, ΔDEO and ΔDEF are 30°-60°-90° triangles. Since the sides of a 30°-60°-90° triangle come in the ratio $1:\sqrt{3}:2$, $DE = \frac{\sqrt{3}}{2}$ (since DE is opposite to 30°, DE must be one-half of EF) and $DO = \frac{1}{2}$ (since DO is opposite to 30°, it must be one-half EO). The base of ΔEFO is $FO = 1$ and the height of ΔEFO is DE $= \frac{\sqrt{3}}{2}$. The area of ΔEFO is $\frac{1}{2}(1)\left(\frac{\sqrt{3}}{2}\right) = \frac{\sqrt{3}}{4}$. Since $360° \div 3 = 120°$, the area of the sector is $\frac{\pi R^2}{3} = \frac{\pi 1^2}{3} = \frac{\pi}{3}$. The area of the sector minus the area of the triangle equals the area of the circular segment:

$$A = \frac{\pi}{3} - \frac{\sqrt{3}}{4}$$

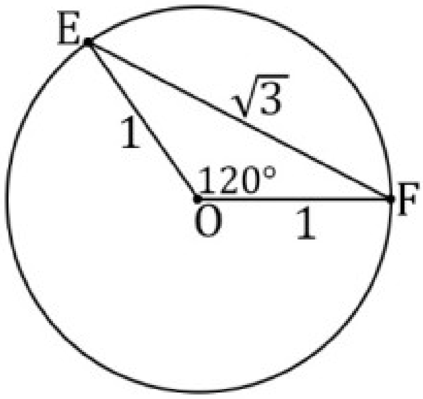

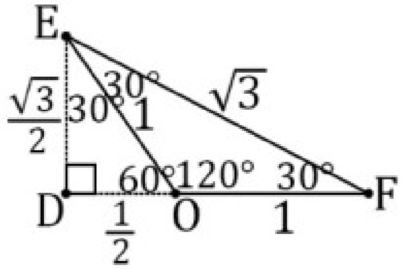

4. $\mathrm{DE} = 2\sqrt{5}$ and $\mathrm{A} = \frac{9\pi}{2} - 4\sqrt{5} \approx 5.19$. Notes: $\Delta\mathrm{DEF}$ is a right triangle according to Thales's theorem (Chapter 5). Use the Pythagorean theorem to check that $\mathrm{DE} = 2\sqrt{5}$:

$$\mathrm{DE}^2 + \mathrm{EF}^2 = \left(2\sqrt{5}\right)^2 + 4^2 = 2^2\left(\sqrt{5}\right)^2 + 4^2$$
$$= 4(5) + 16 = 20 + 16 = 36 = 6^2 = \mathrm{DF}^2$$

The area of $\Delta\mathrm{DEF}$ is $\frac{1}{2}\mathrm{bh} = \frac{1}{2}(\mathrm{DE})(\mathrm{EF}) = \frac{1}{2}\left(2\sqrt{5}\right)(4) = 4\sqrt{5}$. The radius of the circle is $\mathrm{R} = \frac{\mathrm{DF}}{2} = \frac{6}{2} = 3$. The area of the semicircle is $\frac{\pi \mathrm{R}^2}{2} = \frac{\pi 3^2}{2} = \frac{9\pi}{2}$. The area of the sector minus the area of the triangle equals the **combined area** of the **two** circular segments:

$$\mathrm{A} = \frac{9\pi}{2} - 4\sqrt{5}$$

5. $KL = 6$ and $A = 30\pi \approx 94.2$. Notes: $300° = 300° \times \frac{\pi}{180°} = \frac{30\pi}{18} = \frac{5\pi}{3}$ rad (since 30 and 18 are each evenly divisible by 6) is the central angle for the major arc. The given value 10π is the corresponding major arc length: $s = 10\pi$. (The problem clearly states that 10π is an **arc length**, NOT an angle.) Use the arc length formula (Chapter 3).

$$s = R\theta$$

$$10\pi = R\frac{5\pi}{3}$$

$$30\pi = 5\pi R$$

$$6 = R$$

The minor central angle $\angle KOL = 360° - 300° = 60°$. Since $KO = LO = R = 6$ and $\angle KOL = 60°$, ΔKLO is an equilateral triangle. Therefore, $KL = R = 6$. Since $\frac{300°}{360°} = \frac{5}{6}$, the shaded region is $5/6$ of a full circle. The area of the shaded region (which is a **sector**, NOT a circular segment) is $\frac{5\pi R^2}{6} = \frac{5\pi 6^2}{6} = \frac{5\pi 36}{6} = 5\pi 6 = 30\pi$.

6. $AB = \sqrt{8 - 4\sqrt{3}} = 2\sqrt{2 - \sqrt{3}} = \sqrt{6} - \sqrt{2} \approx 1.04$ (all three answers are equivalent, but the form $\sqrt{6} - \sqrt{2}$ is preferred because it does not have nested square roots)

and $A = \frac{\pi}{3} - 1 = \frac{\pi-3}{3} \approx 0.0472$. Notes: ΔABO is isosceles: $AO = BO = R = 2$. (ΔABO is NOT a right triangle.) Draw vertical line segment $\overline{AC}$, as shown on the next page, to create right triangles ΔACO and ΔABC. Since ΔACO is a 30°-60°-90° triangle, $AC{:}CO{:}AO = 1{:}\sqrt{3}{:}2$. Since $CO = \sqrt{3}$ and $BO = 2$, it follows that $BC = 2 - \sqrt{3}$. Use the Pythagorean theorem to find AB.

$$BC^2 + AC^2 = AB^2$$

$$\left(2 - \sqrt{3}\right)^2 + 1^2 = AB^2$$

Apply the "foil" method: $(t - u)^2 = t^2 - 2tu + u^2$.

$$4 - 4\sqrt{3} + 3 + 1 = AB^2$$

$$8 - 4\sqrt{3} = 4\left(2 - \sqrt{3}\right) = AB^2$$

$$\sqrt{8 - 4\sqrt{3}} = 2\sqrt{2 - \sqrt{3}} = AB$$

Following is a clever way to avoid the nested square roots given above.

$$8 - 4\sqrt{3} = 8 - 2(2)\sqrt{3} = 8 - 2\sqrt{4}\sqrt{3} = 6 - 2\sqrt{12} + 2$$

$$= \left(\sqrt{6} - \sqrt{2}\right)^2$$

$$AB = \sqrt{8 - 4\sqrt{3}} = \sqrt{6} - \sqrt{2}$$

Since $\frac{30°}{360°} = \frac{1}{12}$, the area of the sector is $1/12$ the area of a full circle:

$$\frac{\pi R^2}{12} = \frac{\pi 2^2}{12} = \frac{4\pi}{12} = \frac{\pi}{3}$$

Since ΔABO has a base of BO = 2 and a height of AC = 1, the area of ΔABO is

$$\frac{1}{2}bh = \frac{1}{2}(BO)(AC) = \frac{1}{2}(2)(1) = 1$$

and the area of the circular segment is

$$A = \frac{\pi}{3} - 1 = \frac{\pi - 3}{3}$$

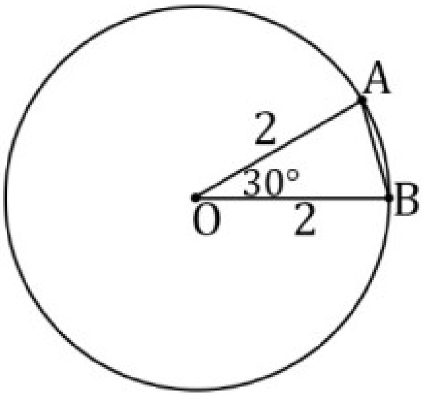

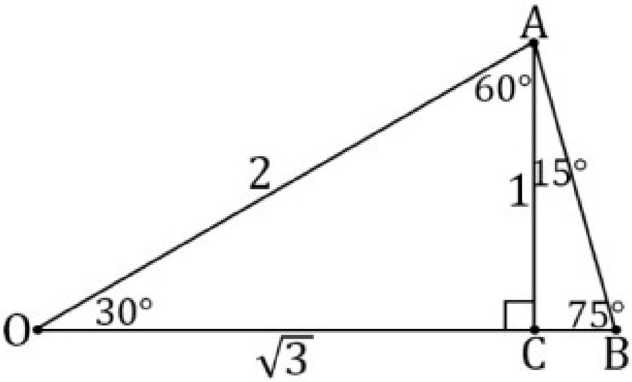

7. $A = 4 - \pi \approx 0.858$. Notes: The area of the square is $2^2 = 4$. The combined area of the two semicircles equals the area of one circle of radius $R = \frac{2}{2} = 1$. The combined area of the two semicircles is $\pi R^2 = \pi 1^2 = \pi$. To find the combined area of the shaded regions, subtract the area of the two semicircles from the area of the square.

8. $A = 2\pi - 4 = 2(\pi - 2) \approx 2.28$. Method 1: Cut the square in half along a diagonal as shown below. Find the

area of each shaded circular segment shown below (recall the solution to Problem 2). The area of the quarter circle is $\frac{\pi R^2}{4} = \frac{\pi 2^2}{4} = \frac{4\pi}{4} = \pi$. The area of the right triangle is $\frac{1}{2} bh = \frac{1}{2} R^2 = \frac{1}{2}(2)^2 = \frac{4}{2} = 2$. The area of the sector minus the area of the triangle equals the area of the circular segment: $\pi - 2$. The two circular segments shown below have equal area. The total area is $A = 2\pi - 4 = 2(\pi - 2)$.

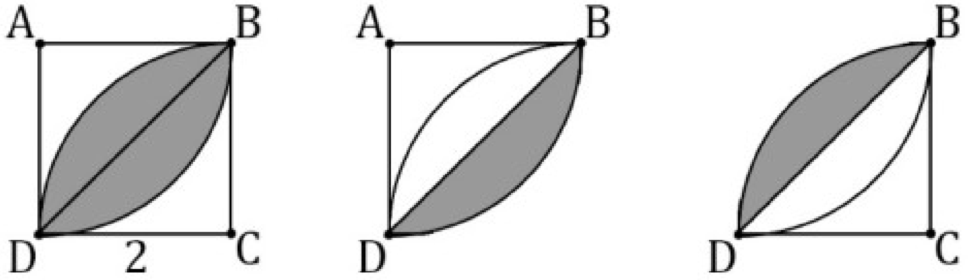

Method 2: The area of the two quarter circles shown below added together equals the area of the square plus the area of the desired shaded region.

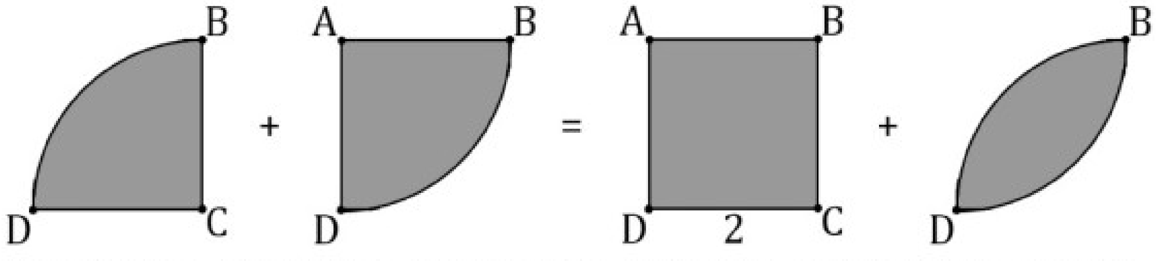

Therefore, the area of the two quarter circles minus the area of the square equals the area of the desired shaded region.

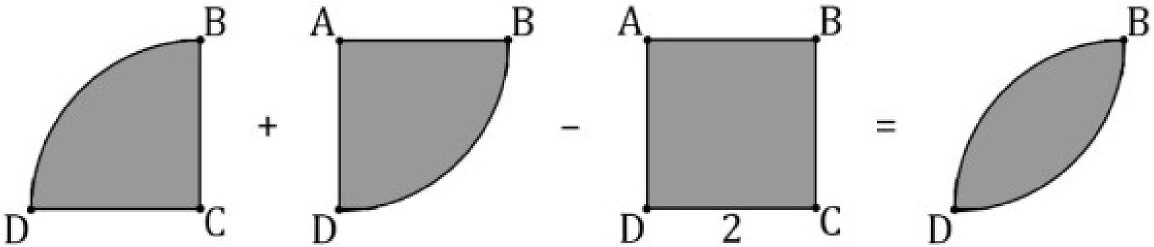

$$\frac{\pi R^2}{4} + \frac{\pi R^2}{4} - R^2 = A$$

$$A = \frac{\pi 2^2}{4} + \frac{\pi 2^2}{4} - 2^2 = \frac{4\pi}{4} + \frac{4\pi}{4} - 4 = 2\pi - 4$$

$$= 2(\pi - 2)$$

9. $A = 12\pi - 18\sqrt{3} = 6\left(2\pi - 3\sqrt{3}\right) \approx 6.52$. Notes: $WY = WZ = YZ = XY = XZ = R = 6$. $\Delta WYZ \cong \Delta XYZ$ and both are equilateral triangles. Find the area of the circular segment shown below (like the solution to Problem 1) and then double this. The area of the sector is $\frac{\pi R^2}{6} = \frac{\pi 6^2}{6} = \frac{36\pi}{6} = 6\pi$. Recall from Volume 1, Chapter 6 that the area of an equilateral triangle with edge length L is $\frac{L^2\sqrt{3}}{4} = \frac{6^2\sqrt{3}}{4} = \frac{36\sqrt{3}}{4} = 9\sqrt{3}$. The area of the sector minus the area of the triangle equals the area of the circular segment: $6\pi - 9\sqrt{3} = 3\left(2\pi - 3\sqrt{3}\right)$. Double this to find the total area: $A = 12\pi - 18\sqrt{3} = 6\left(2\pi - 3\sqrt{3}\right)$.

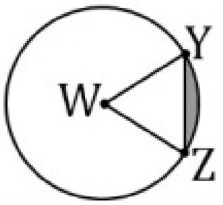

10. $A = 9\sqrt{3} - \frac{9\pi}{2} = 9\left(\sqrt{3} - \frac{\pi}{2}\right) \approx 1.45$. Notes: ΔPQR is an equilateral triangle with edge length $PQ = PR = QR = 2R = 2(3) = 6$. Recall from Volume 1, Chapter 6 that the area of an equilateral triangle with edge length L is $\frac{L^2\sqrt{3}}{4} = \frac{6^2\sqrt{3}}{4} = \frac{36\sqrt{3}}{4} = 9\sqrt{3}$. Each of the three sectors (for vertices P, Q, and R) has a 60° angle (like Problems 1 and 9). Each sector has area $\frac{\pi R^2}{6} = \frac{\pi 3^2}{6} = \frac{9\pi}{6} = \frac{3\pi}{2}$. (Note that the radius of each sector is $R = 3$, whereas the edge length of equilateral triangle ΔPQR is $2R = 6$.) The combined area of the three sectors is $3\left(\frac{3\pi}{2}\right) = \frac{9\pi}{2}$. The area of ΔPQR minus the area of the three sectors equals the area of the shaded region.

$$A = 9\sqrt{3} - \frac{9\pi}{2} = 9\left(\sqrt{3} - \frac{\pi}{2}\right)$$

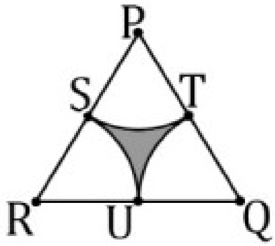

11. $A = \frac{\pi}{2} - \frac{\sqrt{3}}{2} = \frac{\pi - \sqrt{3}}{2} \approx 0.705$. Notes: Draw ΔDEF as shown on the next page. ΔDEF is an equilateral triangle with edge length $DE = DF = EF = \frac{AB}{2} = \frac{AC}{2} = \frac{BC}{2} = \frac{2}{2} = 1$

(recall Volume 1, Chapter 7, "Tips for Working with Medians"). The area of ΔDEF is $\frac{L^2\sqrt{3}}{4} = \frac{1^2\sqrt{3}}{4} = \frac{\sqrt{3}}{4}$ (recall Volume 1, Chapter 6). The area of the **sector** from vertex E that includes both ΔDEF and the shaded circular segment in the diagram on the right below is $\frac{\pi R^2}{6} = \frac{\pi 1^2}{6} = \frac{\pi}{6}$ (like the solution to Problem 1). The area of the sector minus the area of ΔDEF equals the area of the shaded circular **segment** shown in the diagram on the right below: $\frac{\pi}{6} - \frac{\sqrt{3}}{4}$. To find the shaded region given in the problem, add the area of ΔDEF to three times the area of the circular segment shown below on the right.

$$\frac{\sqrt{3}}{4} + 3\left(\frac{\pi}{6} - \frac{\sqrt{3}}{4}\right) = \frac{\sqrt{3}}{4} + \frac{3\pi}{6} - \frac{3\sqrt{3}}{4} = \frac{3\pi}{6} - \frac{2\sqrt{3}}{4}$$

$$= \frac{\pi}{2} - \frac{\sqrt{3}}{2} = \frac{\pi - \sqrt{3}}{2}$$

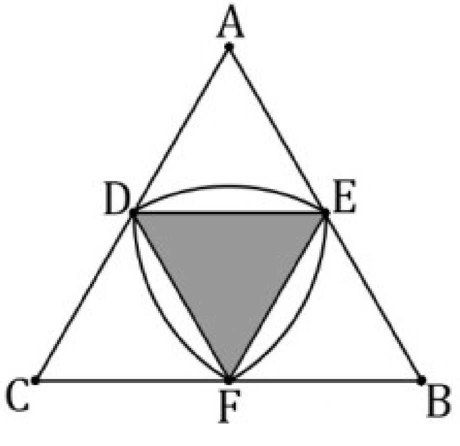

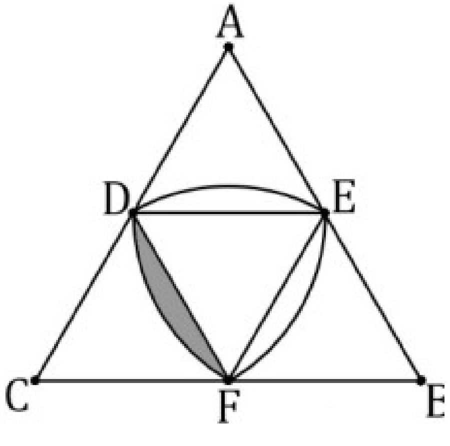

12. $A = \sqrt{3} - \frac{\pi}{6} \approx 1.21$. Notes: The area of the small semicircle shaded in the diagram on the left on the next page minus the area of the circular segment shaded in the middle diagram equals the area of the lune shaded in the diagram on the right. The radius of the small semicircle is $R_{\text{small}} = \frac{2}{2} = 1$. The area of the small semicircle is $\frac{\pi R_{\text{small}}^2}{2} = \frac{\pi 1^2}{2} = \frac{\pi}{2}$. Each sector in the middle diagram on the following page has a 60° central angle, a radius of $R_{\text{large}} = \frac{4}{2} = 2$, and an area of $\frac{\pi R_{\text{large}}^2}{6} = \frac{\pi 2^2}{6} = \frac{4\pi}{6} = \frac{2\pi}{3}$. ΔMNO is an equilateral triangle. The area of ΔMNO is $\frac{L^2\sqrt{3}}{4} = \frac{2^2\sqrt{3}}{4} = \frac{4\sqrt{3}}{4} = \sqrt{3}$ (Volume 1, Chapter 6). Subtract the area of ΔMNO from the area of the sector to find the area of the circular segment shaded in the middle diagram on the following page: $\frac{2\pi}{3} - \sqrt{3}$. To find the area of the lune, subtract the area of the circular segment from the area of the small semicircle shaded in the left diagram.

$$\frac{\pi}{2} - \left(\frac{2\pi}{3} - \sqrt{3}\right) = \frac{\pi}{2} - \frac{2\pi}{3} + \sqrt{3} = \sqrt{3} + \frac{3\pi}{6} - \frac{4\pi}{6}$$
$$= \sqrt{3} - \frac{\pi}{6}$$

The minus sign was distributed. For example,
$a - (b - c) = a - b - (-c) = a - b + c$.

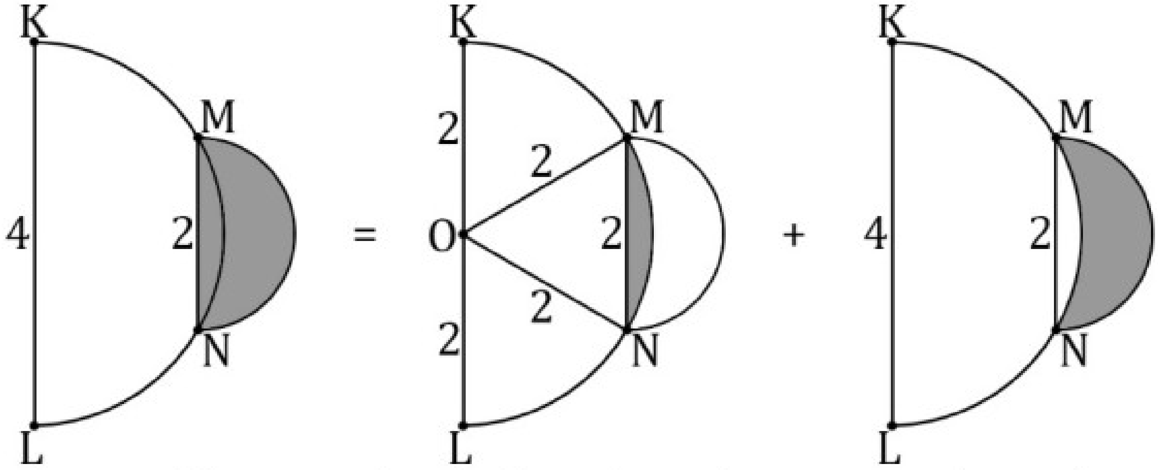

This problem involved a lune based on an equilateral triangle (ΔMNO). Don't confuse this lune with another common and important lune, which instead features a 45°-45°-90° triangle. The lune based on the 45°-45°-90° triangle is important because the π's cancel out for that lune. Compare the two lunes shown on the next page. The lune on the left was featured in this problem, and involves an equilateral triangle. The lune on the right involves a 45°-45°-90° triangle. For the lune on the right, the π's cancel out. Try it!

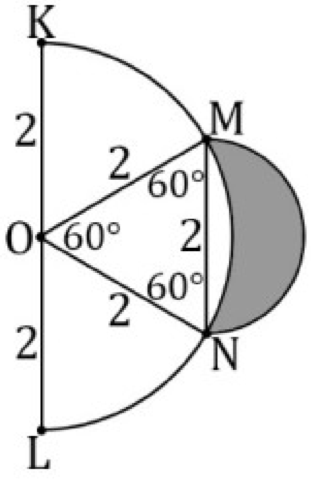

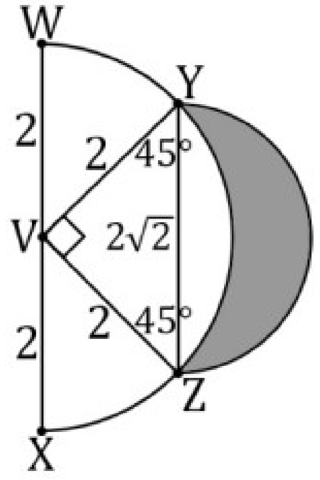

13. $\alpha = 62°$, $\beta = 62°$, $\gamma = 56°$, $\delta = 68°$, $\varepsilon = 56°$, $\varphi = 56°$, $\eta = 62°$, and $\kappa = 62°$. Notes: Central angles that subtend congruent arcs are congruent. Since congruent chords subtend congruent arcs, central angles that subtend congruent chords are congruent. Therefore, since $AB = CD = 7$, $\varphi = 56°$. $AO = BO = CO = DO = R$. Although the value of R is not given, it is clear that ΔABO, ΔBCO, and ΔCDO are isosceles triangles, such that $\alpha = \beta$, $\gamma = \varepsilon$, and $\eta = \kappa$. Checks: $56° + \delta + \varphi = 56° + 68° + 56° = 180°$ (straight angle), $\alpha + \beta + 56° = 62° + 62° + 56° = 180°$, $\delta + \gamma + \varepsilon = 68° + 56° + 56° = 180°$, $\varphi + \eta + \kappa = 56° + 62° + 62° = 180°$. Parallel notes: It turns out that $\overline{BC} \parallel \overline{AD}$ (which is why $\gamma = 56°$ and $\varepsilon = \varphi$ are two pairs of congruent alternate interior angles), but $\overline{BO}$ is NOT parallel to $\overline{CD}$ (η does NOT equal δ) and $\overline{AB}$ is NOT parallel to $\overline{CO}$ (δ does NOT equal β).

14. $\alpha = \gamma = 61°$ and $\beta = \delta = 119°$. Notes: $\alpha = \gamma = \frac{48°+74°}{2} = \frac{122°}{2} = 61°$ (intersecting chords) and $\beta = \delta$ (vertical angles). Check: $\alpha + \beta + \gamma + \delta = 61° + 119° + 61° + 119° = 360°$.

15. $\alpha = 31°$. Check: $\angle PTQ = \angle STR = \frac{\alpha+17°}{2} = \frac{31°+17°}{2} = \frac{48°}{2} = 24°$. Note: To solve for α, first write $24° = \frac{\alpha+17°}{2}$ and then multiply both sides of the equation by 2.

16. $\angle AOF = 40°$. Check: $\angle AFB = \angle CFD = \frac{\angle AOB+\angle COD}{2} = \frac{40°+108°}{2} = \frac{148°}{2} = 74°$. Notes: $\angle AOB \cong \angle AOF$. $\overline{AC}$ and $\overline{BD}$ are the intersecting chords. In this problem, $\overline{BD}$ happens to be a diameter.

17. $JN = 24$ and $NO = 6\sqrt{3}$. Checks: $(JN)(LN) = (24)(9) = 216$, $(KN)(MN) = (8)(27) = 216$, and $R^2 - NO^2 = 18^2 - \left(6\sqrt{3}\right)^2 = 324 - (6)^2\left(\sqrt{3}\right)^2 = 324 - 36(3) = 324 - 108 = 216$.

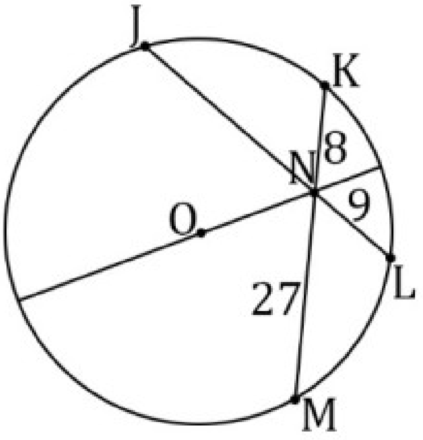

18. VY = 4, OY = 11, and WY = 22. Checks: (VX)(VZ) = (6)(12) = 72 and $R^2 - OV^2 = 11^2 - 7^2 = 121 - 49 = 72$. Notes: Apply the intersecting chords theorem to $\overline{WY}$ and $\overline{XZ}$. In this problem, $\overline{WY}$ happens to be a diameter. VY + OV = OY = R.

19. ∠AOE = 45°, ∠COD = 135°, AO = $\sqrt{2}$, EO = 1, and BE = $\sqrt{2} - 1$. Notes: AO = BO = CO = DO = R = $\sqrt{2}$. ΔAEO and ΔCEO are 45°-45°-90° triangles. AE:EO:AO = CE:EO:CO = 1:1:$\sqrt{2}$. ∠AOE = ∠COE = 45°. ∠COE + ∠COD = 180°. BE + EO = BO. DO + EO = DE = 1 + $\sqrt{2}$. Checks: (BE)(DE) = $\left(\sqrt{2} - 1\right)\left(1 + \sqrt{2}\right) = \sqrt{2} + 2 - 1 - \sqrt{2} = 2 - 1 = 1$, (AE)(CE) = (1)(1) = 1, $R^2 - EO^2 = \left(\sqrt{2}\right)^2 - 1^2 = 2 - 1 = 1$, and $\frac{\angle AOE + \angle COD}{2} = \frac{45° + 135°}{2} = \frac{180°}{2} = 90° = \angle BEA$.

20. Main ideas: $(\mathrm{VW})(\mathrm{VY}) = (\mathrm{VZ})(\mathrm{VX})$ since $\overline{\mathrm{WY}}$ and $\overline{\mathrm{XZ}}$ are intersecting chords. Since $\overline{\mathrm{VW}} \cong \overline{\mathrm{VZ}}$, this becomes $(\mathrm{VW})(\mathrm{VY}) = (\mathrm{VW})(\mathrm{VX})$, which reduces to $\mathrm{VY} = \mathrm{VX}$. $\Delta\mathrm{WVZ} \sim \Delta\mathrm{VXY}$ according to SAS similarity ($\mathrm{VW}/\mathrm{VZ} = 1$, $\mathrm{VY}/\mathrm{VX} = 1$, and the angle between the sides is congruent; recall Volume 1, Chapter 4). $\angle\mathrm{WZV} \cong \angle\mathrm{VXY}$ are congruent alternate interior angles (because $\Delta\mathrm{WVZ}$ is similar to $\Delta\mathrm{VXY}$ and each triangle is isosceles). Note: Recall from Volume 1 that $\sim$ means similarity, whereas $\cong$ means congruence.

21. Main ideas: $\angle\mathrm{AEB} \cong \angle\mathrm{CED}$ (vertical angles) and $\angle\mathrm{ABD} \cong \angle\mathrm{ACD}$ (these inscribed angles subtend the same arc length from A to D; recall Chapter 4). $\Delta\mathrm{ABE} \sim \Delta\mathrm{CDE}$ according to AA (recall Volume 1, Chapter 4). It can similarly be shown that $\Delta\mathrm{ADE} \sim \Delta\mathrm{BCE}$.

22. Main ideas: $\Delta\mathrm{ABE} \sim \Delta\mathrm{CDE}$ (see the solution to Problem 21). AE:BE = DE:CE because the corresponding sides of similar triangles come in the same proportion (Volume 1, Chapter 4).

$$\frac{\mathrm{AE}}{\mathrm{BE}} = \frac{\mathrm{DE}}{\mathrm{CE}}$$

Cross multiply. (Multiply both sides by BE and by CE.)

$$(\mathrm{AE})(\mathrm{CE}) = (\mathrm{BE})(\mathrm{DE})$$

23. Main ideas: $(AE)(CE) = (BE)(DE)$ was proven in Problem 22. Apply the intersecting chords theorem (proven in Problems 21-22) to chords $\overline{AC}$ and $\overline{FG}$: $(AE)(CE) = (EF)(EG)$. Since $EF + EO = FO = GO = R$, it follows that $EF = R - EO$ and $EG = R + EO$. Therefore, $(AE)(CE) = (EF)(EG)$ may be rewritten as:

$$(AE)(CE) = (R - EO)(R + EO)$$

Apply the "foil" method: $(a + b)(c + d) = ac + ad + bc + bd$.

$$(AE)(CE) = R^2 + R(EO) - (EO)R - EO^2 = R^2 - EO^2$$

24. Main ideas: Draw line segment $\overline{AD}$, as shown on the next page. $\angle AED + \angle DAE + \angle ADE = 180°$ (interior angles of ΔADE). $\angle AED + \angle CED = 180°$ (supplementary angles). $\angle CED = \angle DAE + \angle ADE$ (set the first two equations equal and subtract $\angle AED$ from both sides.) $\angle DAE = \angle DAC = \frac{\angle COD}{2}$ and $\angle ADE = \angle ADB = \frac{\angle AOB}{2}$ (an inscribed angle is one-half of a central angle that subtends the same arc; recall Chapter 4). Plug $\angle DAE = \frac{\angle COD}{2}$ and $\angle ADE = \frac{\angle AOB}{2}$ into $\angle CED = \angle DAE + \angle ADE$ to get $\angle CED = \frac{\angle COD}{2} + \frac{\angle AOB}{2} = \frac{\angle COD + \angle AOB}{2}$.

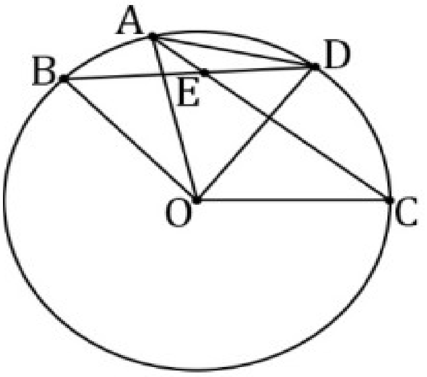
A
D
B
E
O
C

+,- * *. ! * 8! @ @B &@ @ & 1#
Q 2 !D
+,- * *. ! * 8! $ @ @ @ " 1#
Q ,) " '

« une belle endormie » qui mobilisait peu.

Dans les années 1960, Roger Penrose a révolutionné la relativité générale en introduisant des techniques de topologie différentielle et de géométrie algébrique (voir l'article suivant). À cette époque, les trous noirs étaient des objets décrits par certaines solutions des équations de la relativité générale et avait fait l'objet de différents modèles décrits par l'astrophysicien allemand Karl Siegmund Schwarzschild (1873–1916), par Roy Kerr et par le physicien américain Julius Robert Oppenheimer (1904–1967). Il n'y avait pas alors de preuve expérimentale ou observationnelle de leur existence.

Le grand apport de Penrose a été d'utiliser la relativité générale pour théoriser la formation de trous noirs par effondrement gravitationnel des étoiles. Il proposait donc, pour la première fois, une histoire de la formation des trous noirs, ce qui pouvait apporter une preuve théorique supplémentaire de leur existence. Il publiait ainsi en 1965 un article dans lequel il démontrait, en utilisant la relativité générale, que les trous noirs pouvaient se former dans des conditions réalistes et qu'il existait, au-delà de l'horizon des évènements, une singularité qui ne pouvait être décrite par les théories physiques actuelles. Avec Stephen Hawking, d'abord son étudiant, puis son collaborateur, les recherches sur les trous noirs connurent des développements continus et importants, marqués aussi par des controverses heuristiques sur l'« entropie des trous noirs » (voir *la Guerre des savants*, Leonard Susskind, Robert Laffont, 2010).

Le diagramme de Penrose

Cette vue cavalière représente une étoile qui s'effondre en deux coordonnées d'espace. Un cercle est une sphère. Le temps s'écoule vers le haut du diagramme : le cercle en bas représente la surface de l'étoile, et son rayon va diminuer (vers le haut) en fonction du temps.

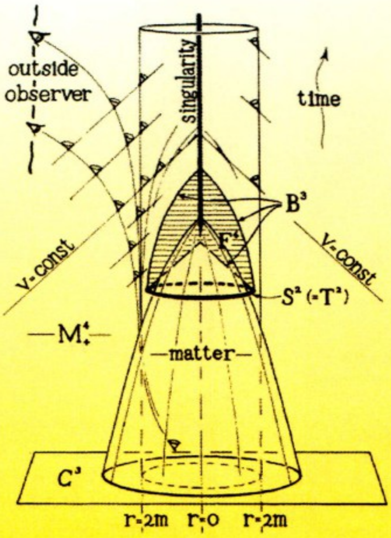

)© Roger Penrose, Physical Review Letters, 1965

À un moment donné, la surface atteint le cercle en gras, qui représente le rayon de Schwarzschild.

La diminution de la surface continue jusqu'à la singularité formée par la ligne verticale figurée en gras. Ce diagramme montre l'effondrement d'une étoile vers un trou noir.

La théorie de l'effondrement gravitationnel

Roger Penrose démontrait en 1965 un théorème de géométrie qui éclairait l'effondrement gravitationnel non sphérique des étoiles (*Gravitational collapse and space-time singularities*, *Physical Review Letters* 14). Interprété dans le langage de la physique, le résultat est le suivant : tout comme dans un effondrement sphérique, dans un effondrement « réaliste » non sphérique, un stade critique est atteint au-delà duquel aucune communication avec l'extérieur ne peut être établie. Une fois que ce stade est passé, un ou deux évènements peuvent survenir : soit la géométrie de l'espace-temps produit une singularité (cas d'un effondrement sphérique), soit, de manière plus étonnante, l'effondrement met en contact notre univers avec un autre univers dont nous ignorions jusque-là l'existence.

Il y a une différence importante entre la singularité engendrée par un effondrement sphérique et celle qui pourrait être créée lors d'un effondrement non sphérique. Dans le premier cas, lorsqu'elle rencontre la singularité, la matière qui constitue l'étoile est « écrasée » en une densité infinie, dans un volume nul. Dans le cas général, le théorème de Penrose n'affirme rien quant au comportement de la matière d'une étoile en effondrement : cette matière pourrait émerger dans une autre région de l'univers, ou dans un autre univers. Ces résultats sur l'effondrement au-delà du rayon critique sont assez étranges pour inciter certains scientifiques à abandonner la relativité générale d'Einstein au profit d'une théorie qui prédit des phénomènes moins décapants. Penrose a cependant indiqué que la plupart des autres théories géométriques de la gravitation prévoient les mêmes types de phénomènes : son théorème repose principalement sur le mouvement des photons à travers l'espace-temps courbe, et ce mouvement est approximativement le même dans toutes les théories géométriques de la gravitation. Les phénomènes étonnants prédits à l'intérieur du rayon critique ne peuvent pas être vus par un observateur extérieur. C'est pourquoi on peut ignorer ces prédictions. Toutefois, l'évolution de notre univers ressemble

La formation d'une singularité

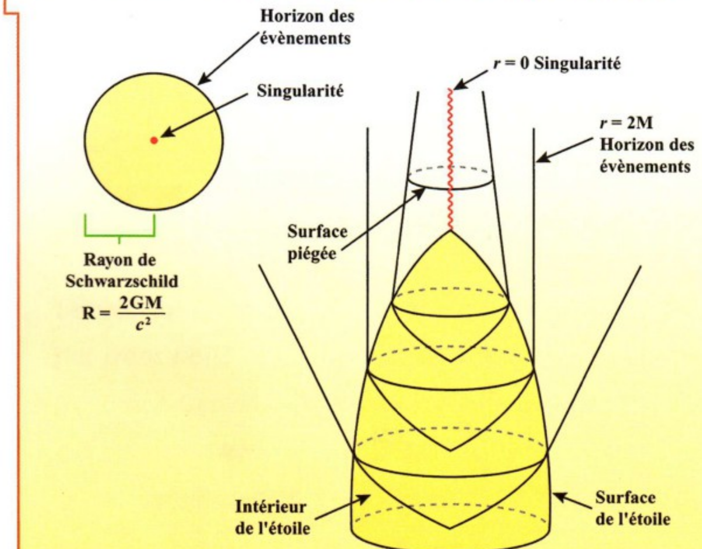

Tout en bas, le cercle représente la coupe d'une sphère matérielle de rayon r qui commence son effondrement gravitationnel. L'effondrement continue en s'élevant dans le schéma (le temps t augmente). Le rayon r diminue jusqu'à une limite (le *rayon de Schwarzschild*) où la sphère est si dense (et sa gravité si forte) que même la lumière ne peut échapper de sa surface. Un trou noir se forme. La position radiale à partir de laquelle la lumière ne peut plus s'échapper délimite une surface de l'espace-temps, l'*horizon des évènements*. Lorsque $r = 0$, on aboutit à une singularité, point où la matière est infiniment dense et où la courbure de l'espace-temps devient infinie. L'équation du champ d'Einstein cesse d'être valide, la théorie de la relativité générale ne s'applique plus. À l'intérieur de l'horizon des évènements, tout mouvement de matière ou de lumière est dirigé vers la singularité.

É.T.

globalement, et de bien des manières, à une étoile qui s'effondre. Si la densité de l'univers est assez grande, il ne s'étendra pas indéfiniment. Il va tôt ou tard arrêter son expansion et commencer à se contracter. La phase de contraction imiterait assez fidèlement l'effondrement d'une étoile. Par conséquent, on ne peut s'inquiéter du sort ultime de l'univers sans s'intéresser à l'effondrement gravitationnel des étoiles.

RÉFÉRENCES

- *L'esprit, l'ordinateur et les lois de la physique.* Roger Penrose, InterÉditions, 1993.
- *Le trésor des ondes gravitationnelles.* Pierre Spagnou, CNRS Éditions, 2020.

✢ La théorie des trous noirs

C'est entre 1967 et 1975 que la théorie des trous noirs progresse vraiment. Un article de Penrose publié en 1972 montre le chemin parcouru vers une compréhension beaucoup plus profonde du phénomène. Pour commencer, l'auteur prend soin de placer l'effondrement gravitationnel et la formation des trous noirs dans le contexte de l'évolution stellaire, soucieux qu'il est de faire sortir les trous noirs du champ de la spéculation pure pour les ancrer dans une réalité astronomique. Revenant ensuite sur le terrain théorique, Penrose introduit le concept de *cône de lumière* (voir encadré) pour décrire la structure de l'espace-temps. Il s'attache ensuite à décrire des solutions de type trou noir plus élaborées que celle trouvée par Schwarzschild : des trous noirs qui ne sont pas exactement sphériques, ou qui possèdent une charge électrique. Il énonce alors le *théorème fondamental de la physique des trous noirs* : le trou noir isolé le plus général est pratiquement dépourvu de propriétés.

Le trou noir isolé le plus général est formellement entièrement décrit par trois paramètres : sa masse M, sa charge électrique Q et son moment angulaire J (son « sens de rotation » : comme une étoile est en rotation, elle possède un moment cinétique, qui est conservé lorsque l'étoile change de volume et dont hérite donc le trou noir *in fine*). En conséquence, il n'existe qu'un nombre réduit de types différents de trous noirs : ceux de Schwarzschild (sphériques, obtenus à partir d'étoiles sans mouvement cinétique, donc sans mouvement de rotation, et sans charge électrique : $J = Q = 0$), ceux de Kerr (non sphériques, avec mouvement de rotation, mais sans charge électrique : $J \neq 0$, $Q = 0$), ceux de Reissner–Nordström (sans rotation et chargés électriquement : $J = 0$, $Q \neq 0$) et ceux de Kerr–Newman (à la fois en rotation et chargés : $J \neq 0$, $Q \neq 0$).
Tout comme la physique quantique, la relativité générale résiste bien aux multiples tests expérimentaux qui sont effectués depuis plusieurs dizaines d'années et qui sont récompensés aujourd'hui par un prix Nobel.

M.L.

©123RF / Igor Zhuravlov

par Michel Criton et Daniel Lignon

Le triangle de Penrose

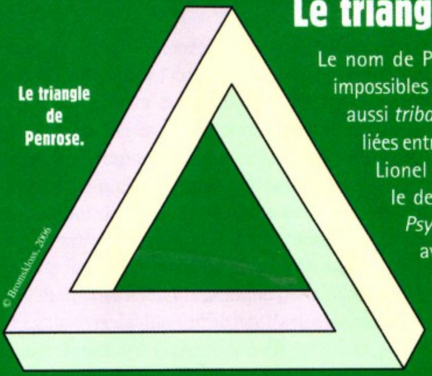
Le triangle de Penrose.

© Bromskloss, 2006

Le nom de Penrose est attaché à plusieurs « objets impossibles ». C'est le cas du *triangle de Penrose*, appelé aussi *tribarre* ou *tripoutre*, constitué de trois poutres liées entre elles et représentées en 2D. Avec son père, Lionel Penrose (1898–1972), ils en publièrent le dessin dans un article du *British Journal of Psychology* en février 1958. Roger Penrose en avait eu l'idée après avoir découvert le travail du graveur néerlandais Maurits Cornelis Escher (1898–1972) alors qu'il se trouvait à Amsterdam en 1954 pour le Congrès international des mathématiciens.

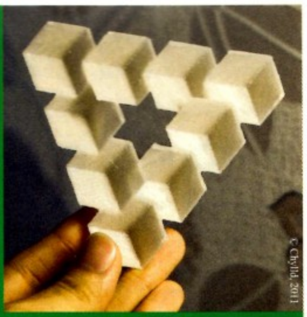
© Chylld, 2011

Représentation de l'illusion de Reutersvärd (devinez l'astuce mise en œuvre pour la réalisation !).

En fait, c'est l'artiste suédois Oscar Reutersvärd (1915–2002) qui, le premier, a dessiné une série de neuf cubes dans la même configuration. C'était en 1934 pendant un cours de latin où, selon ses dires, il s'ennuyait.

Cette découverte changea le cours de sa vie. Au cours de sa carrière de graphiste, il dessina de très nombreuses « figures impossibles » : il est d'ailleurs parfois surnommé le « père de l'impossible ». Pour lui rendre hommage, la Suède a émis en 1982 une série de timbres. Manifestement, ni les Penrose, ni Escher n'avaient eu connaissance des figures créées, des années auparavant, par Reutersvärd.

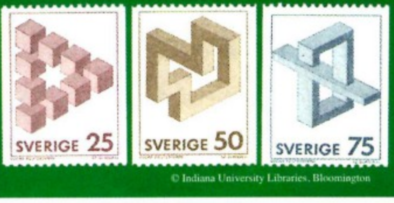

© Indiana University Libraries, Bloomington

Les timbres suédois en hommage à Reutersvärd.

en bref

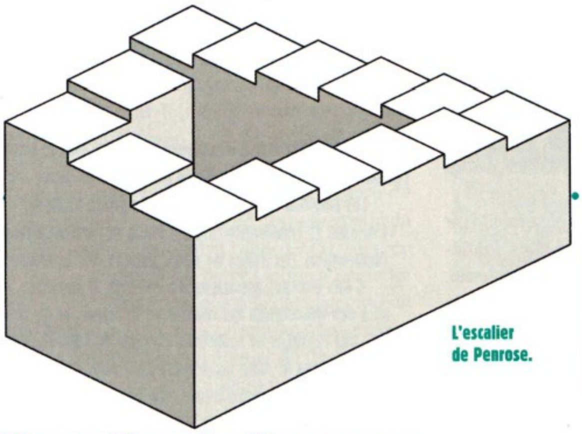
© Sakurambo, 2005

L'escalier de Penrose.

L'escalier de Penrose

Une autre « figure impossible » est présente dans l'article écrit par les Penrose en 1958 : l'escalier. Mais, là encore, cette figure avait déjà été décrite par Reutersvärd, en 1937 !

Escher a connaissance de cet objet suite à des échanges avec Roger Penrose. Dès 1960, il en tirera une lithographie restée célèbre, *Montée et Descente*. Andrew Lipson et Daniel Shiu ont réussi à recréer cette œuvre (et bien d'autres, voir leur site, www.andrewlipson.com/lego.htm) avec des briques danoises...

Références :

Maths et arts plastiques. Bibliothèque Tangente 23, 2018.
Inventions mathématiques. Jean-Paul Delahaye, Belin, 2014.
Les secrets des dimensions. Bibliothèque Tangente 66, 2019.

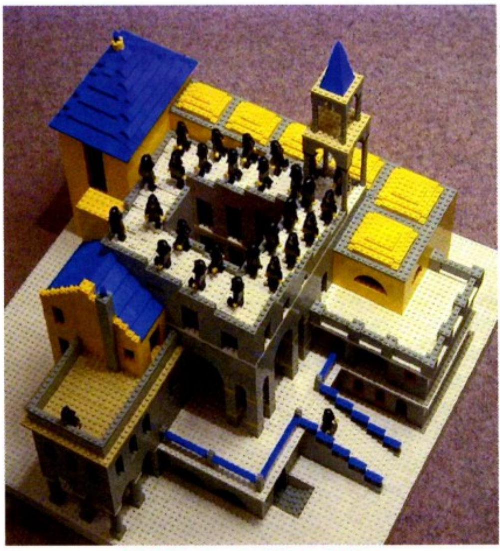
© A. Lipson, 2002 (andrewlipson.com/lego.htm)

Réalisation avec des briques danoises de l'œuvre *Montée et Descente* d'Escher (trouvez l'artifice que les artistes ont utilisé !).

par Édouard Thomas

Les singularités
en relativité générale

L'année 1965 marque l'arrivée du premier théorème moderne, dû à Roger Penrose, sur les singularités en relativité générale. Il s'agit d'un résultat géométrique, qui aura d'importantes répercussions concernant l'origine de l'univers et l'effondrement d'étoiles massives.

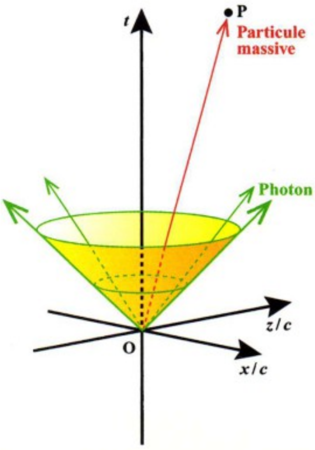
t
P
Particule massive
Photon
z / c
O
x / c

Demi-cône de lumière dirigé vers le futur. L'ensemble des possibles suite à l'évènement O se situe dans ce demi-cône. Les deux coordonnées d'espace *x* et *z* sont divisées par *c* de sorte que les lignes d'univers des photons soient inclinées à 45°. Une *ligne d'univers* (droite si la particule se déplace à vitesse uniforme, courbe sinon) décrit l'histoire d'une particule tout au long de son existence.

© E. Thomas, 2015

Les équations de la relativité générale (voir encadré) comptent parmi les plus fondamentales de toute la physique moderne. Introduites en 1915 et 1916 par Albert Einstein (1879–1955), on les appelle plus simplement aujourd'hui *l'équation de champ d'Einstein*, au singulier. Cette dernière offre d'innombrables défis mathématiques redoutables de toute beauté (conjecture du cerceau, stabilité non linéaire de la famille de Kerr, conjecture BKL…), dont les fameuses conjectures relatives au phénomène de censure cosmique (voir plus loin) : l'équation est loin d'avoir révélé tous ses secrets !

Du point de vue mathématique, un « espace-temps » est une *variété lorentzienne* muni d'une métrique. Le « chemin le plus court » (pour la métrique considérée) entre deux points de la variété est appelé une *géodésique*. Enfin, la notion de « courbure » de l'espace-temps étudié est encodée et quantifiée à l'aide d'un *tenseur de courbure* (ici, $R_{\mu,\nu}$). Plus l'espace-temps est « courbe », plus les géodésiques vont s'éloigner ou se rapprocher rapidement. Enfin, si $g_{\mu,\nu}(V) > 0$, le vecteur tangent V est *de genre espace*, si $g_{\mu,\nu}(V) = 0$, il est *de genre lumière*, et si $g_{\mu,\nu}(V) < 0$, V est *de genre temps*.

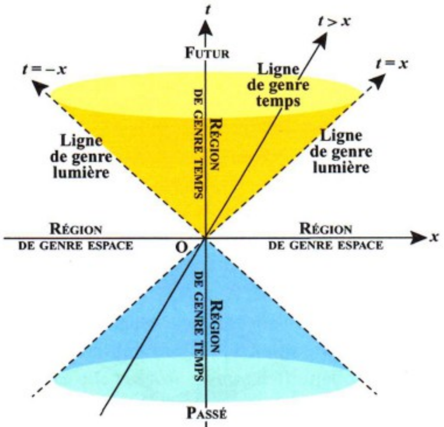
t
t > x
Futur
t = −x
t = x
Ligne de genre temps
Ligne de genre lumière
Région de genre temps
Ligne de genre lumière
Région de genre espace
O
Région de genre espace
x
Région de genre temps
Passé

L'axe *x* représente une dimension d'espace, *t* est le temps. Les particules matérielles suivent des lignes de genre temps. Les photons suivent des lignes de genre lumière. L'origine O représente la localisation et l'instant d'une observation. Les lignes de genre espace correspondent à des vitesses supérieures à celle de la lumière, ou à des trajectoires qui remontent le temps, et ne sont donc pas réalistes en relativité. L'avenir de l'observateur se situe dans le *cône de lumière* supérieur (en jaune), sur une ligne de genre temps, son passé se situe dans le cône de lumière inférieur (en bleu).

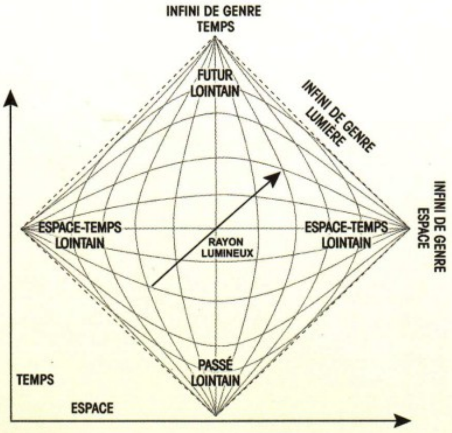
INFINI DE GENRE TEMPS
FUTUR LOINTAIN
INFINI DE GENRE LUMIÈRE
INFINI DE GENRE ESPACE
ESPACE-TEMPS LOINTAIN
RAYON LUMINEUX
ESPACE-TEMPS LOINTAIN
PASSÉ LOINTAIN
TEMPS
ESPACE

L'arrivée des singularités en physique

Le premier « vrai » théorème sur les singularités porte précisément sur l'équation du champ d'Einstein. Il a été obtenu en 1965 dans un article de moins de trois pages, assez dense, par Sir Roger Penrose. Son importance est telle qu'il a valu à son auteur de recevoir, mardi 6 octobre dernier, le prix Nobel de physique pour *« la découverte que la formation d'un trou noir est une prévision robuste de la théorie de la relativité générale »*. Le célèbre mathématicien britannique partage ce prix à part égale avec le duo formé par l'astrophysicien allemand Reinhard Genzel (né en 1952) et par l'astronome américaine Andrea Mia Ghez (née en 1965), pour leur *« découverte d'un objet compact supermassif au centre de notre galaxie »*, à savoir Sagittarius A*. Découverte qui doit beaucoup aux résultats de Penrose…

Ce sont donc des contributions théoriques (l'introduction de concepts clés en physique mathématique et la démonstration d'un théorème) qui valent au Britannique la fameuse distinction. La démonstration du théorème sur les singularités va conduire dans les années qui suivent à une théorie physique « mûre », dotée de résultats puissants.

L'existence d'« irrégularités catastrophiques » en relativité générale, appelées *singularités*, n'avait pas été anticipée par Albert Einstein. Pourtant, dès 1916 (voir plus loin), les premiers résultats théoriques relatifs à une masse centrale (le Soleil) de Karl Schwarzschild, sans aller jusqu'à imaginer des « trous noirs », ouvrent la voie à de nouvelles explorations. En 1922, le mathématicien russe Alexandre Friedmann (1888–1925), suivi en 1927 par l'astrophysicien belge Georges Lemaître (1894–1966), obtiennent des résultats accréditant l'hypothèse d'un univers en expansion, posant dès lors la question de son « point de départ ». En 1955, Amal Kumar Raychaudhuri (1923–2005) prouve une sorte de théorème de singularité pour certaines poussières stellaires, qui sera généralisé l'année suivante par Arthur Komar (1931–2011). Les singularités ont le vent en poupe.

Un théorème de géométrie

En 1965, Roger Penrose est un expert du calcul tensoriel et de l'équation du champ d'Einstein. Travaillant sur les singularités, il prouve que les écarts ou déviations par rapport à la symétrie sphérique ne peuvent pas empêcher, à eux seuls, la formation de singularités au sein de la relativité générale. Le résultat ne fait pas intervenir les hypothèses d'homogénéité ou de symétrie de ses prédécesseurs : il surprend par sa généralité. Il affirme que si un espace-temps est « suffisamment régulier » (au sens que $g_{\mu,\nu}$ est deux fois continûment différentiable) et qu'il vérifie trois conditions, alors il contient une géodésique causale mais *incomplète* (que l'on ne peut pas prolonger à $\mathbb{R}$ tout entier). C'est un résultat de pure géométrie.

Les hypothèses sont de nature géométrique pour l'une (sur la courbure de l'espace-temps : la densité totale de masse et d'énergie ne doit jamais être négative), causale pour la deuxième (« le futur doit toujours être distinct du passé », afin d'éviter les boucles temporelles), et analytique pour la troisième (condition initiale ou condition de bord : l'espace-temps doit avoir un volume infini avant que l'effondrement ne commence). Les deux premières sont « réalistes » sur le plan de la physique. La troisième pose encore problème : on ignore en effet si notre espace-temps a un volume infini. La plupart des théoriciens de la relativité conjecturent cependant que le théorème de Penrose est valide, que l'espace-temps soit infini ou non. Mais le résultat reste à généraliser et à prouver dans le cas d'un volume fini. •

L'équation du champ d'Einstein

Dans son apparence la plus classique, l'équation d'Einstein prend la forme suivante :

$$\mathrm{R}_{\mu,v} - \frac{1}{2}\mathrm{R}g_{\mu,v} + \Lambda g_{\mu,v} = \frac{8\pi}{c^4}\mathrm{G}\mathrm{T}_{\mu,v}.$$

Elle fait intervenir $\mathrm{R}_{\mu,\nu}$ le *tenseur de Ricci* (tenseur d'ordre 2 qui encode la déformation de l'espace-temps) ; R la *courbure scalaire* (un nombre qui mesure la courbure de l'espace-temps, c'est la trace de $\mathrm{R}_{\mu,\nu}$) ; $g_{\mu,\nu}$ le *tenseur métrique*, à savoir un tenseur d'ordre 2 qui permet d'associer un produit scalaire à deux vecteurs quelconques en chaque point de l'espace-temps (il est de signature (+, –, –, –), caractéristique de la géométrie lorentzienne) ; $\mathrm{T}_{\mu,\nu}$ le *tenseur énergie–impulsion*, qui permet de caractériser la répartition de masse et d'énergie dans l'espace-temps ; Λ une constante physique fondamentale (la *constante cosmologique*), G la constante gravitationnelle ($\mathrm{G} = 6{,}67430 \times 10^{-11}\ \mathrm{m}^3\ \mathrm{kg}^{-1}\ \mathrm{s}^{-2} \pm 0{,}00015 \times 10^{-11}\ \mathrm{m}^3\ \mathrm{kg}^{-1}\ \mathrm{s}^{-2}$) et c la vitesse de la lumière dans le vide ($c = 299\,792\,458\ \mathrm{m\ s}^{-1}$).

L'étude mathématique de l'équation du champ d'Einstein est très ardue. Déjà, comment en saisir la signification ? Les *tenseurs* sont des applications multilinéaires qui, à des vecteurs et covecteurs d'un espace vectoriel, associent un scalaire. Ce qu'il convient de comprendre, c'est que l'équation du champ d'Einstein permet de connaître le tenseur métrique ($g_{\mu,\nu}$, indispensable pour mesurer des longueurs et des angles dans l'espace-temps) étant donnée une distribution de matière et d'énergie (exprimée sous la forme du tenseur énergie–impulsion $\mathrm{T}_{\mu,\nu}$). La géométrie de l'espace-temps (décrite par $g_{\mu,\nu}$) est déterminée localement par son contenu en énergie–impulsion (donné par $\mathrm{T}_{\mu,\nu}$) ; l'espace-temps dit à la matière comment se déplacer (*via* $\mathrm{R}_{\mu,\nu}$) ; la matière dit à l'espace-temps comment se courber (*via* R).

Références :

Les équations de la physique moderne. Bibliothèque Tangente 71, 2020.
Vecteurs et espaces vectoriels. Bibliothèque Tangente 65, 2018.

L'espace-temps de Gödel

En 1949 et 1952, le célèbre logicien Kurt Gödel (1906–1978), qui s'intéresse alors à la relativité générale, démontre que la rotation de l'espace-temps à elle seule peut suffire à empêcher la formation de singularités. *L'espace-temps de Gödel* est l'un des rares exemples d'une solution *régulière* (c'est-à-dire sans singularité) aux équations d'Einstein. Il est par ailleurs en rotation et *géodésiquement complet* (au sens que le paramètre définissant un point courant de ses géodésiques est défini sur $\mathbb{R}$ tout entier, ce qui équivaut à dire, par le théorème de Hopf–Rinow, que les parties fermées et bornées sont compactes). De fait, il contient, mathématiquement parlant, des *courbes de genre temps fermées*, ce qui signifie, du point de vue de la physique, que le voyage dans le temps y serait permis. Il s'agit donc d'une solution « artificielle », avec boucles temporelles. Or, si l'on pose le principe de causalité comme prérequis à tout modèle physique, il ne saurait y avoir de boucles temporelles. L'espace-temps de Gödel conserve toutefois d'intéressantes vertus pédagogiques.

© G. Guerrerio, Logische Paradoxien und mathematische Wahrheit

Kurt Gödel.

Penrose introduit l'idée d'incomplétude pour les géodésiques afin de décrire d'éventuels espaces-temps singuliers. Il introduit également la notion de surface fermée S *piégée* (au sens que deux familles de rayons lumineux émergeant de manière orthogonale de S et dirigées vers le futur convergent initialement : elles se trouvent « piégées » et ne *peuvent s'échapper de S), une contribution qui* s'avérera fondamentale pour toute la physique des champs gravitationnels. De fait, ses idées ont infusé en géométrie différentielle, en analyse, en géométrie algébrique et en topologie différentielle. Une singularité n'appartient pas « physiquement » à l'espace-temps considéré, il s'agit d'une pure abstraction mathématique (en effet, par définition, un espace-temps physique est constitué de points réguliers). Il s'agit d'une demi-géodésique, de genre temps ou de genre lumière, dirigée vers le futur, et qui est *incomplète* (son intervalle de définition est borné) : l'histoire d'une particule de cette semi-géodésique commence à un temps fini (le Big Bang) et s'achève en un temps fini (en pénétrant dans un trou noir).

Le théorème sur les singularités de 1965 accrédite l'hypothèse d'un état final singulier d'une étoile massive. Étendant ce résultat, les travaux de Hawking en 1966 et de Hawking–Penrose en 1970 indiquent qu'une singularité dans le passé était inévitable en relativité générale : l'hypothèse d'un Big Bang se trouve confortée. En 1972 enfin, Stephen Hawking développe une théorie des trous noirs : il devient en effet nécessaire de compléter la théorie de la relativité générale par des considérations issues de la physique quantique lorsqu'une gravité extrême est en jeu.

C'est cosmique !

Les théorèmes sur les singularités en relativité générale, qui stipulent tous que, pour une raison non spécifiée, soit l'observateur disparaît, soit le rayon lumineux s'arrête, ne marquent pas la fin de l'histoire. En effet, ils ne caractérisent pas la *singularité. De quelle nature est cette géodésique* causale incomplète ? Est-elle localisée ? Est-elle unique ? Survient-elle dans le passé ou dans le futur ? Surtout, d'un point de vue « pratique » : peut-on observer une singularité ?

Dans la solution exacte à l'équation du champ d'Einstein obtenue en 1916 par Schwarzschild, une singularité a été mise au jour ; elle se niche au sein d'une étoile ; elle n'est donc pas visible. Dans cette solution, on fait entre autres l'hypothèse que l'espace-temps possède une symétrie sphérique

La notion de surface piégée

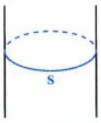

La notion de surface fermée piégée (ci-contre, à gauche, représentée comme un cercle), introduite par Penrose, formalise l'idée intuitive d'une surface fermée bordant une région de l'espace dont la lumière ne peut s'échapper. C'est une notion délicate dans la mesure où l'idée intuitive fait référence à un espace absolu qui ne saurait exister dans un contexte relativiste général. Dans le schéma qui suit, la dimension d'espace z n'est pas représentée.

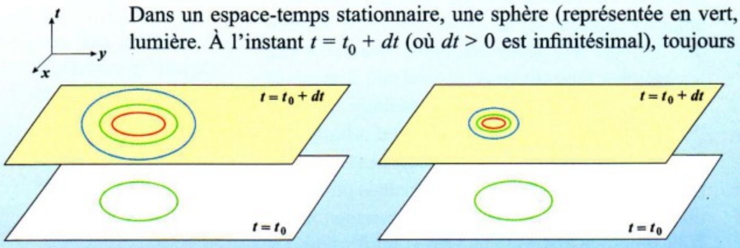

Espace-temps stationnaire **Espace-temps en contraction**

Dans un espace-temps stationnaire, une sphère (représentée en vert, en coupe, à l'instant $t = t_0$) émet des rais de lumière. À l'instant $t = t_0 + dt$ (où $dt > 0$ est infinitésimal), toujours en coupe (dans le même plan), on observe la sphère initiale (en vert), une deuxième sphère (en bleu) correspondant au lieu défini par les rais de lumière émis vers l'extérieur de la sphère initiale, et une troisième sphère (en rouge) correspondant au lieu défini par les rais de lumière émis vers l'intérieur de la sphère initiale. L'aire A_0 de la sphère originale reste constante dans le temps. L'aire de la sphère bleue, supérieure à A_0, augmente avec le temps. L'aire de la sphère rouge, inférieure à A_0, diminue avec le temps.

Dans un univers en contraction, les sphères bleue et rouge peuvent avoir une aire constamment inférieure à A_0, y compris celle de la sphère bleue, bien qu'elle croisse en permanence ! Cette sphère est alors un exemple d'une *surface fermée piégée dans le futur*.

parfaite et on fixe $\Lambda = 0$ dans l'équation du champ d'Einstein. Dès lors, l'apparition d'une singularité aurait pu résulter d'hypothèses très contraignantes et peu réalistes. Il n'en est rien : le théorème sur les singularités de 1965 a prouvé que des situations bien plus générales forcent l'émergence de singularités au sein d'étoiles ou de trous noirs. Penrose a alors posé les questions naturelles suivantes :

- Les singularités produites à la suite d'un effondrement gravitationnel sont-elles toujours cachées au sein d'une étoile ou d'un trou noir, derrière un horizon des évènements ? (*censure cosmique faible*)
- Peut-on exclure que, à l'exception du Big Bang, les singularités produites à la suite d'un effondrement gravitationnel sont de genre temps ? (*censure cosmique forte*)

Ces questions posent en fait le problème de l'adéquation entre une formulation mathématique précise et rigoureuse (l'étude de l'équation du champ d'Einstein) et le phénomène astrophysique d'un effondrement gravitationnel. Faut-il prendre en compte la nature quantique du champ gravitationnel ? Dans la négative, on est conduit à réaliser des approximations, qui cessent d'être valides à proximité d'une singularité de l'espace-temps (à cause de la courbure très grande de ce dernier).

Si les réponses aux deux questions posées par Penrose sont négatives (on parle de *singularité nue*, d'où le terme humoristique de « censure »), alors les effets de la gravité quantique seraient visibles en situation d'effondrement gravitationnel, et la gravité quantique deviendrait nécessaire pour comprendre la nature profonde des singularités.

C'est ainsi que sont nées, sous la plume de Penrose, la *conjecture faible de la censure cosmique* dès 1969 et la *conjecture forte* en 1976. La censure cosmique est un phénomène physique hypothétique donnant naissance à une singularité nue. Selon ces conjectures, il n'existerait pas d'autre singularité nue que celle associée au Big Bang.

Mathématiquement parlant, les deux « conjectures » sont indépendantes et ne sont pas énoncées de manière extrêmement précise ; il s'agit plus de programmes de recherche. Il n'existe pas de singularité de genre temps dans un espace-temps globalement hyperbolique. Dans un tel espace-temps, la conjecture de la censure cosmique forte est automatiquement vérifiée. L'idée de la version « forte » est alors que l'ensemble E des données initiales qui conduisent à un espace-temps globalement hyperbolique remplit tout l'espace-temps (E ne peut pas être étendu sous des conditions génériques).

Cette conjecture est l'une des plus fameuses de toute la théorie des équations aux dérivées partielles. Elle est également très difficile à attaquer, sauf à introduire des hypothèses simplificatrices de type symétrie de l'espace-temps.

Pour l'étudier, Penrose a introduit l'*inégalité de Penrose*. Il s'agit d'une sorte d'inégalité isopérimétrique, selon laquelle (dans les notations de l'équation du champ d'Einstein) l'aire de l'horizon apparent S d'une hypersurface de Cauchy Σ de masse totale M, de moment angulaire J et de charge électrique Q est inférieure à $16\pi\left(\frac{GM}{c^2}\right)^2$, soit $4\pi(r_{\text{Sch}})^2$ avec $r_{\text{Sch}} = \frac{2GM}{c^2}$ le *rayon de Schwarschild*.

L'*horizon apparent* S est la surface externe de Σ piégée. L'« aire de S » désigne en fait l'aire minimale de toute surface de Σ contenant S.

Pour un trou noir, on a $Q \leq \frac{G}{c}M$ ou $J \leq \frac{G}{c}M^2$.

Pour une singularité nue,
on a $Q > \frac{G}{c}M$ ou $J > \frac{G}{c}M^2$.

Comme l'inégalité de Penrose ne dépend que de Σ, qui est un objet mathématique, elle peut être étudiée géométriquement pour elle-même et généralisée en toute dimension.

Une violation de l'inégalité de Penrose entraînerait la visibilité d'effets quantiques dans le phénomène macroscopique d'effondrement gravitationnel (et donc une absence de prédictibilité de la relativité générale telle qu'elle est développée actuellement). Ce serait un contre-exemple à la conjecture faible de la censure cosmique.

À ce jour, l'inégalité de Penrose a été prouvée dans certains cas particuliers, mais elle n'a été ni confirmée, ni infirmée dans le cas général.

É.T.

Relativité générale : les contributions de Penrose

En dehors de ses théorèmes sur les singularités (celui de 1965, et ceux qui suivront, en particulier avec Stephen Hawking), Sir Roger Penrose a durablement marqué la relativité générale de son empreinte. Les *diagrammes de Penrose* ou *de Penrose–Carter* sont des diagrammes bidimensionnels facilitant l'étude des propriétés causales d'un espace-temps ; ils permettent de représenter plusieurs métriques spatio-temporelles en supprimant systématiquement deux dimensions d'espace (voir un exemple dans le précédent article). Ils sont *conformes* : les lignes horizontales sont du genre espace, les lignes verticales sont du genre temps, la lumière (géodésique de genre nul) se déplace sur les *lignes de lumière*, orientées à 45° et –45°. Le *formalisme de Penrose* ou *de Newman–Penrose* est un ensemble de notations introduites pour traiter la relativité générale en termes de notations avec des spineurs, qui remplace certaines variables réelles par des variables complexes. Enfin, Penrose travaille avec acharnement, depuis plus d'un demi-siècle, à une grande théorie géométrique, la *théorie des twisteurs*, qui permettrait de formaliser sur des bases solides une gravité quantique. L'espace-temps y serait remplacé par l'*espace des twisteurs*, un espace géométrique très sophistiqué dans lequel le déplacement des photons serait représenté à l'aide de quatre nombres complexes.

© Weston Library, Oxford

Stephen Hawking (1942 – 2018).

par C. Laurens, M. Criton et É. Thomas

Des Nobel de physique et des maths

Le plus célèbre des lauréats du prix Nobel de physique est sans doute Albert Einstein, récompensé en 1921 pour des travaux autour de l'effet photoélectrique – et non pour sa fameuse théorie de la relativité. À ses côtés, dans la liste des lauréats, on retrouve plusieurs scientifiques ayant œuvré du côté des mathématiques. L'un des domaines d'étude largement représentés parmi eux est la physique quantique, pour laquelle ont été mis à l'honneur Max Karl Ernst Ludwig Planck en 1918 *« pour la découverte des quanta d'énergie »*, Werner Heisenberg en 1932 *« pour la création de la mécanique quantique »*, puis Wolfgang Pauli en 1945, Max Born en 1954 ou encore Julian Seymour Schwinger et Richard Phillips Feynman en 1965.

Par ailleurs, bon nombre de ces aventuriers nobélisés ont été récompensés pour des découvertes en physique atomique ou subatomique, comme Niels Bohr en 1922, Louis de Broglie en 1929, Erwin Schrödinger et Paul Adrien Maurice Dirac en 1933, ou encore Eugene Paul Wigner en 1963.

On recense également des prix Nobel décernés à Hendrik Antoon Lorentz en 1902 pour des travaux en électromagnétisme, à John William Strutt Rayleigh en 1904 pour des études en physique des gaz, à Wilhelm Wien en 1911 pour de la thermodynamique, à Lev Davidovitch Landau en 1962 pour des découvertes en physique de la matière condensée ou à Subrahmanyan Chandrasekhar en 1983 pour des travaux en astrophysique. Une telle variété souligne, au-delà de la pluridisciplinarité des lauréats, l'intrication profonde des mathématiques et de la physique.

en bref

Penrose et l'affaire Kleenex

Dans les années 1970, Roger Penrose avait déposé des brevets pour un certain nombre de ses pavages apériodiques (!) et il avait créé une société, Pentaflex, chargée de gérer les droits de ces brevets.

En 1997, il découvre que des rouleaux de papier hygiénique de la firme Kleenex, achetés par son épouse, portent des motifs inspirés de l'un de ses pavages, sans son accord. Il porte l'affaire en justice et attaque la firme américaine Kimberly-Clark Corporation, détentrice de la marque déposée Kleenex.

L'issue du procès est incertaine. Selon certaines sources, Penrose gagna le procès et obligea la firme à retirer ses rouleaux en vente et à changer de décoration ; selon d'autres sources, il trouva un arrangement financier avec le fabricant. De fait, un motif géométrique peut-il être soumis à des droits d'auteur ?

En 1995, un informaticien américain, Roger Schlafly, déposa un brevet pour deux nombres premiers (US Patent 5 373 560), l'un de cent cinquante chiffres et l'autre de trois cents chiffres, utilisables dans des procédés de codage. Penrose lui-même considéra qu'un tel dépôt de brevet était abusif.

Un engagement dans la diffusion des connaissances

Sir Roger Penrose est un vulgarisateur hors pair, à l'écrit comme à l'oral. Son ouvrage *l'Esprit, l'Ordinateur et les Lois de la physique* (Interéditions, 1993) est ainsi un modèle du genre et permet de comprendre finement la pensée originale et les idées scientifiques de l'auteur.

On sait moins que Penrose est très impliqué dans l'Archive Trust for Research in Mathematical Sciences and Philosophy (**https://v1.archmathsci.org**), dont il est l'un des présidents d'honneur depuis sa création. Cette fondation caritative britannique fondée en 2007 repose sur un projet que porte son directeur, le physicien Michael Wright, depuis 1973. L'objet est de créer des archives sonores et vidéo de milliers de conférences, colloques, séminaires, cours, discussions et entretiens (formels ou informels), assurés par des universitaires de tout premier plan, et qui ont été inlassablement filmés, par Wright et son équipe parcourant le monde, depuis près de cinquante ans.

À ce jour, l'Archive Trust compte plus de quarante mille tels documents, accompagnés de plus de deux cent cinquante mille pages de transcriptions, notes de cours et correspondances. Les mathématiques pures et la physique théorique forment l'essentiel du fonds documentaire, qui couvre également la philosophie, la cosmologie, l'histoire des sciences… Au moins vingt pour cent des archives sont en français.

Le travail de numérisation, de catalogage et de référencement par des métadonnées pertinentes est considérable et nécessite des moyens financiers substantiels. À terme, l'intégralité de l'archive est destinée à être mise à disposition gracieusement auprès de la communauté des chercheurs.

© adrian finn (greatbear-digitising)

Des milliers de documents d'exception sont stockés en attendant d'être numérisés.

+,- * *. ! * ! 3 !
1 ! H!2 ! 0 341 ! # # < = +11 ! 3 7 1#
) 9 ' +11 # 4 ((6 8 H 7 1!

Appendix II.D The Liouville space

The Liouville (line) space $\mathcal{L}$ (see also Refs. [11,67]) is defined as the product $\mathcal{H} \otimes \mathcal{H}^+$ where $\mathcal{H}$ is the Hilbert space and $\mathcal{H}^+$ is its associated dual space.

Vectors in $\mathcal{L}$. If the ket vectors $|\alpha\rangle$ (and associated bra's $\langle\alpha|$) form a basis of $\mathcal{H}$, the $|\alpha\rangle\langle\beta|'$s, denoted $|\alpha\beta\rangle\rangle$, (and $\langle\langle\alpha\beta| \equiv |\beta\rangle\langle\alpha|$) which are operators in $\mathcal{H}$, form a basis in $\mathcal{L}$. The product of two vectors of $\mathcal{L}$ is defined by:

$$\langle\langle A'|A\rangle\rangle = \mathrm{Tr}[A'A], \tag{II.D1}$$

where Tr[...] designates the trace. One thus has:

$$\langle\langle\alpha'\beta'|\alpha\beta\rangle\rangle = \mathrm{Tr}[|\beta'\rangle\langle\alpha'|\alpha\rangle\langle\beta|] = \sum_{\gamma}\langle\gamma|\beta'\rangle\langle\alpha'|\alpha\rangle\langle\beta|\gamma\rangle. \tag{II.D2}$$

Hence, if the $|\alpha\rangle'$s form an orthonormalized basis of $\mathcal{H}$ (i.e., $\langle\alpha'|\alpha\rangle = \delta_{\alpha,\alpha'}$), the $|\alpha\beta\rangle\rangle'$s are also orthonormalized in $\mathcal{L}$, since Eq. (II.D2) straightforwardly leads to:

$$\langle\langle\alpha'\beta'|\alpha\beta\rangle\rangle = \delta_{\alpha,\alpha'}\delta_{\beta,\beta'}. \tag{II.D3}$$

Furthermore, one also has the following "closure" relation:

$$\sum_{\alpha,\beta}|\alpha\beta\rangle\rangle\langle\langle\alpha\beta| = 1. \tag{II.D4}$$

If A designates an operator in $\mathcal{H}$, and hence a vector in $\mathcal{L}$, one has:

$$\begin{aligned}|A\rangle\rangle &= \sum_{\alpha,\beta}\langle\langle\alpha\beta|A\rangle\rangle \times |\alpha\beta\rangle\rangle = \sum_{\alpha,\beta}\mathrm{Tr}[|\beta\rangle\langle\alpha|A] \times |\alpha\beta\rangle\rangle = \sum_{\alpha,\beta,\gamma}\langle\gamma|\beta\rangle\langle\alpha|A|\gamma\rangle \times |\alpha\beta\rangle\rangle \\ &= \sum_{\alpha,\beta}\langle\alpha|A|\beta\rangle \times |\alpha\beta\rangle\rangle,\end{aligned} \tag{II.D5}$$

yielding:

$$\langle\langle\alpha\beta|A\rangle\rangle = \langle\alpha|A|\beta\rangle. \tag{II.D6}$$

Operators in $\mathcal{L}$. If $\hat{O}$ is a (super) operator in $\mathcal{L}$, constructed from the operator O of $\mathcal{H}$, the action of $\hat{O}$ on any vector $|A\rangle\rangle$ of $\mathcal{L}$ is given by (A being an operator in $\mathcal{H}$):

$$\hat{O}|A\rangle\rangle = [O, A], \tag{II.D7}$$

where [,] is the commutator. One can then define the matrix elements in the Liouville space. Indeed, Eq. (II.D5) leads to:

$$\hat{O}|\alpha\beta\rangle\rangle = [O, |\alpha\rangle\langle\beta|] = O|\alpha\rangle\langle\beta| - |\alpha\rangle\langle\beta|O, \tag{II.D8}$$

leading, from Eq. (II.D6), to:

$$\langle\langle\alpha'\beta'|\hat{O}|\alpha\beta\rangle\rangle = \langle\alpha'|O|\alpha\rangle\langle\beta|\beta'\rangle - \langle\alpha|\alpha'\rangle\langle\beta|O|\beta'\rangle. \tag{II.D9}$$

Hence, if O is a Hermitian operator in $\mathcal{H}$ and the $|\alpha\rangle$'s are its orthonormalized eigenvectors with associated eigenvalues o_α, one has:

$$\langle\langle\alpha'\beta'|\hat{O}|\alpha\beta\rangle\rangle = (o_\alpha - o_\beta) \times \delta_{\alpha,\alpha'}\delta_{\beta,\beta'}. \tag{II.D10}$$

This equation is often used in the present book in the particular case of the operator $\hat{O} = L_a$ of $\mathcal{L}$ constructed from the operator $O = H_a/\hbar$, where H_a is the Hamiltonian describing the internal states of the radiator. If $|i\rangle$ and $|f\rangle$ designate eigenvectors of H_a associated with the (energy) eigenvalues E_i and E_f, then the matrix element:

$$\langle\langle fi|L_a|fi\rangle\rangle = (E_f - E_i)/\hbar \equiv \omega_{fi}, \tag{II.D11}$$

directly gives the frequency ω_{fi} [or, in wavenumbers, $\sigma_{fi} = \omega_{fi}/(2\pi c)$] of the line associated with the f←i optical transition.

Evolution operators in $\mathcal{L}$. In the Hilbert space, the Heisenberg representation of the time evolution of an operator O is given by:

$$O(t) = U^{+}(t)OU(t), \tag{II.D12}$$

where U(t) is the evolution operator (U^{+} being its adjoint), given by:

$$U(t) = e^{-iHt/\hbar}, \tag{II.D13}$$

H being the total Hamiltonian for the considered system. Let us introduce the Liouville operator L constructed from $H/\hbar$. Note that, due to Eq. (II.D7), the use of L in the equation giving the evolution of the density operator $\rho(t)$, that is:

$$i\hbar\frac{\partial}{\partial t}\rho(t) = [H, \rho], \tag{II.D14}$$

leads to the following compact form:

$$i\frac{\partial}{\partial t}\rho(t) = L\rho, \tag{II.D15}$$

which naturally calls for the introduction of the Liouville operator $\hat{U}(t)$, defined by:

$$\hat{U}(t) = e^{-iLt}. \tag{II.D16}$$

From Eq. (II.D12) and the definition of L, it is then straightforward to show that:

$$O(t) = \hat{U}(t)^{+}O. \tag{II.D17}$$

One thus has, successively using Eqs. (II.D17), (II.D4), and (II.D6):

$$\begin{aligned}\langle\langle\beta'\alpha'|O(t)\rangle\rangle &= \langle\langle\beta'\alpha'|\hat{U}^{+}(t)O\rangle\rangle = \sum_{\alpha,\beta}\langle\langle\beta'\alpha'|\hat{U}^{+}(t)|\beta\alpha\rangle\rangle\langle\langle\beta\alpha|O\rangle\rangle \\ &= \sum_{\alpha,\beta}\langle\langle\beta'\alpha'|\hat{U}^{+}(t)|\beta\alpha\rangle\rangle\langle\beta|O|\alpha\rangle.\end{aligned} \tag{II.D18}$$

From Eqs. (II.D6) and (II.D12), and the closure relation in the Hilbert space, one can write:

$$\begin{aligned}\langle\langle\beta'\alpha'|O(t)\rangle\rangle &= \langle\beta'|O(t)|\alpha'\rangle = \langle\beta'|U^{+}(t)OU(t)|\alpha'\rangle \\ &= \sum_{\alpha,\beta}\langle\beta'|U^{+}(t)|\beta\rangle\langle\beta|O|\alpha\rangle\langle\alpha|U(t)|\alpha'\rangle.\end{aligned} \tag{II.D19}$$

Identifying Eqs. (II.D18) and (II.D19) leads to:

$$\langle\langle\beta'\alpha'|\hat{U}^{+}(t)|\beta\alpha\rangle\rangle = \langle\beta'|U^{+}(t)|\beta\rangle\langle\alpha|U(t)|\alpha'\rangle = \langle\beta|U(t)|\beta'\rangle^{*}\langle\alpha|U(t)|\alpha'\rangle. \tag{II.D20}$$

This shows that $\langle\langle fi|\hat{U}^{+}(t)|f'i'\rangle\rangle = \langle f'|U(t)|f\rangle^{*}\langle i'|U(t)|i\rangle$, and, finally:

$$\langle\langle f'i'|\hat{U}(t)|fi\rangle\rangle = \langle f'|U(t)|f\rangle\langle i'|U(t)|i\rangle^{*}, \tag{II.D21}$$

in agreement with Eq. (II.45).

Appendix II.E The resolvent approach

II.E1 Spectral-shape expression

The starting point of the Fano and Ben-Reuven formalism[44–46,137] is Eqs. (II.7) and (II.8), which define the spectral density $F(\omega)$ as the Fourier transform of the autocorrelation function $\Phi(t)$ of the radiator dipole moment $\vec{d}$ (when considering, of course, electric dipole absorption spectra):

$$F(\omega) = \frac{1}{\pi}\mathrm{Re}\left\{\int_{0}^{\infty} dt e^{i\omega t}\mathrm{Tr}_{S_M}\left[\rho_M(0)e^{+iH_M t/\hbar}\vec{d}^{+}e^{-iH_M t/\hbar}\vec{d}\right]dt\right\}. \tag{II.E1}$$

Here, H_M and $\rho_M(0)$ are the Hamiltonian and the initial density operator of the system S_M composed of one radiator in a bath of N_p perturbers. Note that, implicit in Eq. (II.E1), is the assumption that the translational motion of the radiating molecule can be

20210912

Deanna Padilla.

https://abandonedsoutheast.com/2021/08/09/lynnewood-hall/

Flatland Dover
(Abbott)

Andrew Beal $3^6 + 18^3 = 3^8$

3 is a common factor.

Kind of arithmetic procedural error

$\frac{16}{64} = \frac{1\cancel{6}}{\cancel{6}4} = \frac{1}{4}$ $\frac{26}{65} = \frac{2\cancel{6}}{\cancel{6}5} = \frac{2}{5}$

$\frac{19}{95} = \frac{1\cancel{9}}{\cancel{9}5} = \frac{1}{5}$ $\frac{49}{98} = \frac{4\cancel{9}}{\cancel{9}8} = \frac{4}{8}$

1690 - Giovanni Cassini was the first person to suggest that Jupiter could not be a solid planet. He noted that the planet was rotating at different rates depending on the latitude (differential rotation).

Here's the formula for the largest number of pieces that can be produced from slicing a doughnut with n cuts.

13 pieces with 3 cuts.

$$\frac{n^3 + 3n^2 + 8n}{6}$$

$$\sin 2x = 2 \sin x \cos x$$

$$\cos 2x = \cos^2 x - \sin^2 x$$

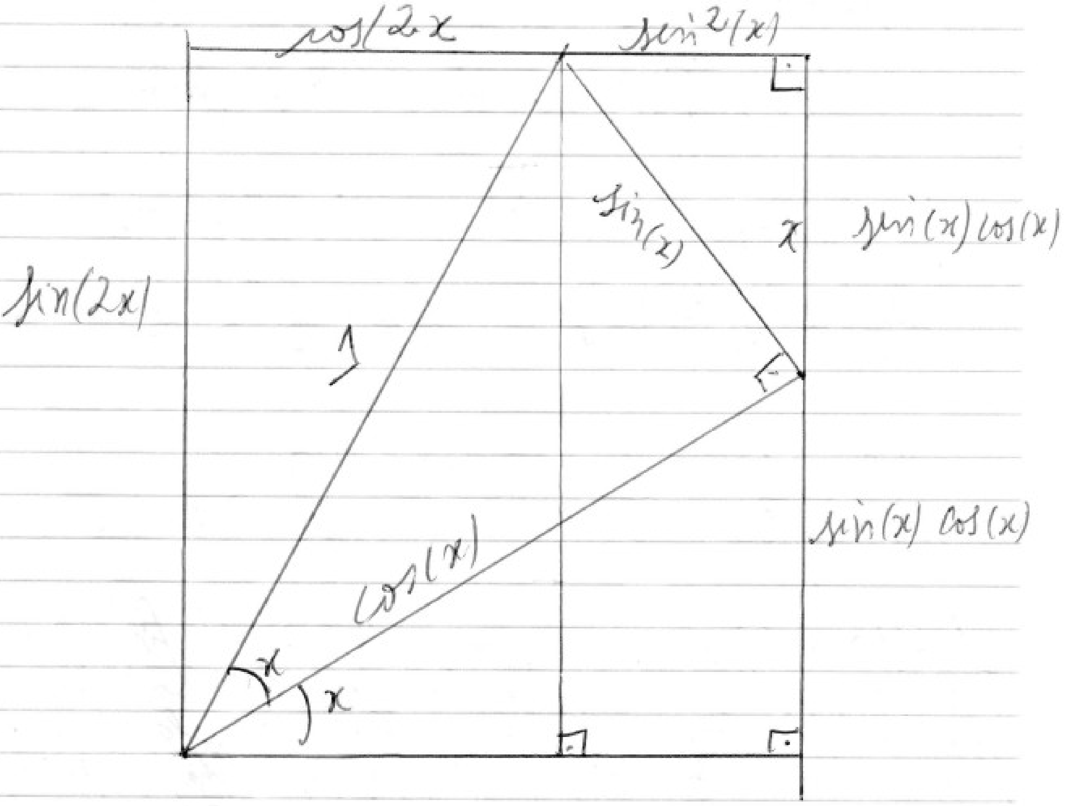

cos(2x
sin²(x)
sin(x)
x
sin(x) cos(x)
sin(2x)
1
cos(x)
x
x
sin(x) cos(x)

$$\frac{\pi}{2} = \frac{2 \cdot 2 \cdot 4 \cdot 4 \cdot 6 \cdot 6 \cdot 8 \cdot 8 \ldots}{1 \cdot 3 \cdot 3 \cdot 5 \cdot 5 \cdot 7 \cdot 7 \cdot 9 \ldots}$$

$$a^2 + b^3 = c^7$$

this is the only solution in positive integers:

$$15312283^2 + 9262^3 = 113^7$$

"A mathematician is a blind man in a dark room looking for a black cat which isn't there"

Charles Darwin

→ Não concordo pois é a matemática que fornece a luz que desvenda toda a escuridão das outras áreas.

Heni Munich.

Andrew Wiles

Andrew Wiles on what it feels like to do mathematics

data

"Perhaps I could best describe my experience of doing mathematics in terms of entering a dark mansion.

One goes into the first room, and it's dark, completely dark.

One stumbles around bumping into the furniture, and gradually, you learn where each piece of furniture is, and finally, after six months or so, you find the light switch.

You turn it on, and suddenly, it's all illuminated! You can see exactly where you were".

Liouville (property of the divisors)

$$(1+2+2+4)^2 = 81 = 1^3+2^3+2^3+4^3$$

Source:

fermat

www.twitter.com/fermatlibrary

20210913///

Mensch (wen)

Books:

Jandaia

) * %
! +,-.+ + +/ ! +) % 0 1

1 +,-.+ + +/ ! +) 2 3 4 ! ! !
4 ! ! ! 1 # 5 6 3 ' 0#
*)" ' " 0 7 8 6

4 Complex projective spaces

> In the house of mathematics there are many mansions and of these the most elegant is projective geometry.
>
> *Morris Kline*

An attentive reader of Chapter 3 will have noticed that there must exist a Hopf fibration of any odd-dimensional sphere; it is just that we studiously avoided to mention what the resulting base space is. In this chapter it will be revealed that

$$\mathbb{C}\mathbf{P}^n = \mathbf{S}^{2n+1}/\mathbf{S}^1. \tag{4.1}$$

The space on the left-hand side is called *complex projective space* or – when it is used as a space of pure states in quantum mechanics – *projective Hilbert space*. It is not a sphere unless $n = 1$. The study of projective spaces was actually begun by artists during the Renaissance, and this is how we begin our story.

4.1 From art to mathematics

Painters, engaged in projecting the ground (conveniently approximated as a flat plane) onto a flat canvas, discovered that parallel lines on the ground intersect somewhere on a *vanishing line* or *line at infinity* that exists on the canvas but not on the ground, where parallel lines do not meet. Conversely, a line exists on the ground (right below the artist) that does not appear on the canvas at all. When the subject was taken over by mathematicians this led to the theory of the *projective plane* or *real projective 2-space* $\mathbb{R}\mathbf{P}^2$. The idea is to consider a space whose points consists of the one-dimensional subspaces of a three-dimensional vector space $\mathbb{R}^3$. Call them *rays*. The origin is placed at the eye of the artist, and the rays are his lines of sight. The ground and the canvas serve as two *affine coordinate planes* for $\mathbb{R}\mathbf{P}^2$. A *projective line* is a certain one parameter family of points in the projective plane, and is defined as the space of rays belonging to some two-dimensional subspace $\mathbb{R}^2$. In itself, this is the real projective 1-space $\mathbb{R}\mathbf{P}^1$. To see how all this works out in formulae, we observe that any triple of real numbers defines a one-dimensional subspace of $\mathbb{R}^3$. But the latter does not determine the triple of numbers uniquely. Hence

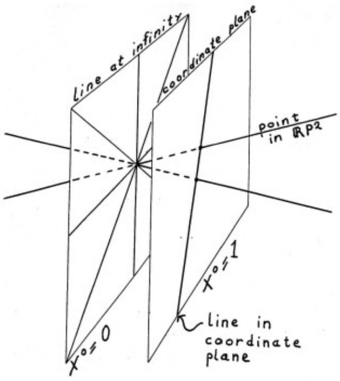

Figure 4.1. Affine coordinates for (almost all of) the projective plane.

a point of $\mathbb{R}\mathbf{P}^2$ is given by an equivalence class

$$(X^0, X^1, X^2) \sim k(X^0, X^1, X^2)\ ; \qquad k \in \mathbb{R} \quad k \neq 0\ . \tag{4.2}$$

The numbers X^α are known as *homogeneous coordinates* on $\mathbb{R}\mathbf{P}^2$. If we want to use true coordinates, one possibility is to choose a plane in $\mathbb{R}^3$ such as $X^0 = 1$, representing an infinite canvas, say. We think of this plane as an affine plane (a vector space except that the choice of origin and scalar product is left open). We can then label a ray with the coordinates of the point where it intersects the affine plane. This is an affine coordinate system on the projective plane; it does not cover the entire projective plane since the rays lying in the plane $X^0 = 0$ are missing. But this plane is a vector space $\mathbb{R}^2$, and the space of rays lying in it is the projective line $\mathbb{R}\mathbf{P}^1$. Hence the affine coordinates display the projective plane as an infinite plane to which we must add an extra projective line 'at infinity'.

Why did we define our rays as one-dimensional subspaces, and not as directed lines from the origin? The answer does not matter to the practising artist, but for the mathematician our definition results in simple *incidence properties* for points and lines. The projective plane will now obey the simple axioms that

(1) Any two different points lie on a unique line.
(2) Any two different lines intersect in a unique point.

There are no exceptional cases (such as parallel lines) to worry about. That axiom (2) holds is simply the observation that two planes through the origin in $\mathbb{R}^3$ always intersect, in a unique line.

There is nothing special about the line at infinity; the name is just an artefact of a special choice of affine coordinates. Every projective line is the space of rays in some two-dimensional subspace of $\mathbb{R}^3$, and it has the topology of a circle – to its image in the affine plane we must add one point 'at infinity'.

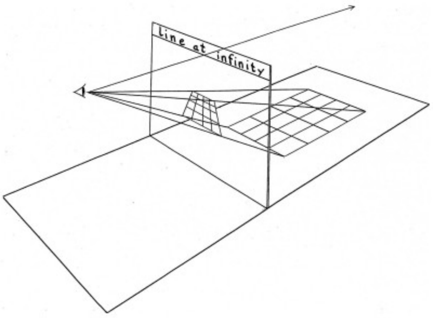

Figure 4.2. Renaissance painter, engaged in changing affine coordinates.

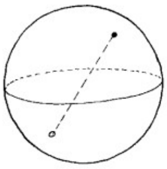

Figure 4.3. Real projective space is the sphere with antipodal points identified.

It is interesting to observe that the space of all lines in $\mathbb{R}\mathbf{P}^2$ is another $\mathbb{R}\mathbf{P}^2$ since one can set up a one-to-one correspondence between rays and planes through the origin in $\mathbb{R}^3$. In fact there is a complete *duality* between the space of points and the space of lines in the projective plane. The natural way to select a specific one-to-one correspondence between $\mathbb{R}\mathbf{P}^2$ and the dual $\mathbb{R}\mathbf{P}^2$ is to introduce a metric in the underlying vector space, and to declare that a ray in the vector space is dual to that plane through the origin to which it is orthogonal. But once we think of the vector space as a metric space as well as a linear one we will be led to the round metric as the natural metric to use on $\mathbb{R}\mathbf{P}^2$. Indeed a moment's reflection shows that $\mathbb{R}\mathbf{P}^n$ can be thought of as the sphere $\mathbf{S}^n$ with antipodal points identified,

$$\mathbb{R}\mathbf{P}^n = \mathbf{S}^n/\mathbb{Z}_2 \ . \tag{4.3}$$

This is one of the reasons why the real projective plane is so famous: it is a space of constant curvature where every pair of geodesics intersect once (as opposed to twice, as happens on a sphere). In fact all of Euclid's axioms for points and straight lines are valid, except the fifth, which shows that the parallel axiom is independent of the others.

In general a specific one-to-one map between a projective space and its dual is known as a *polarity*, and choosing a polarity singles out a particular metric on the projective space as being the most natural one.

There are other geometrical figures apart from points and lines. A famous example is that of the *conic section*. Consider a (circular, say) cone of rays

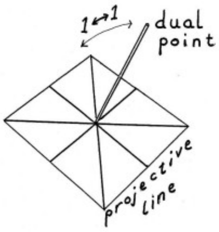

Figure 4.4. Polarity: a metric for the projective plane.

with its vertex at the origin of $\mathbb{R}^3$. In one affine coordinate system the points of intersection of the cone with the affine plane will appear as a circle, in another as a hyperbola. Hence circles, ellipses and hyperbolae are projectively equivalent. In homogeneous coordinates such a conic section appears as

$$(X^1)^2 + (X^2)^2 = (X^0)^2 \ . \tag{4.4}$$

This equation defines a submanifold of $\mathbb{R}\mathbf{P}^2$ because it is homogeneous in the X^α, and therefore unaffected by scaling. If we use affine coordinates for which $X^0 = 1$ it appears as a circle, but if we use affine coordinates for which $X^1 = 1$ it appears as a hyperbola. We can use this fact to give us some further insight into the somewhat difficult topology of $\mathbb{R}\mathbf{P}^2$: if we draw our space as a disc with antipodal points on the boundary identified – this would be the picture of the topology implied by Figure 4.3 – then we see that the hyperbola becomes a topological circle because we are adding two points at infinity. Its interior is an ordinary disc while its exterior is a *Möbius strip* because of the identifications at the boundary – if you glue a disc to the boundary of a Möbius strip you obtain a projective plane. By the way $\mathbb{R}\mathbf{P}^2$, unlike $\mathbb{R}\mathbf{P}^1$, is a *non-orientable* space; this means that a 'right-handed' basis in the tangent space can be turned into a 'left-handed' one by moving it around so that in fact the distinction between right- and left-handedness cannot be upheld. See Figure 4.5.

If you think a little bit further about the $\mathbb{R}\mathbf{P}^2$ topology you will also notice a – possibly reassuring – fact, namely that there is a topological difference between projective lines and conic sections. Although they are both circles intrinsically, the conic sections have the property that they can be continually deformed to a point within the projective plane, while the lines cannot. Indeed if you cut the projective plane open along a projective line it remains connected, while it splits into two pieces if you cut along a conic section.

We could go on to discuss n-dimensional real projective spaces $\mathbb{R}\mathbf{P}^n$, but we confine ourselves to two topological remarks: that $\mathbb{R}\mathbf{P}^n = \mathbf{S}^n/\mathbb{Z}_2$ (obvious), and that $\mathbb{R}\mathbf{P}^{2n+1}$ is orientable while $\mathbb{R}\mathbf{P}^{2n}$ is not (easy to prove).

We can change our *field* from the real numbers $\mathbb{R}$ to rational, complex or quaternionic numbers. In fact one can also use finite fields, such as the integers modulo p where p is some prime number. If we repeat the construction of

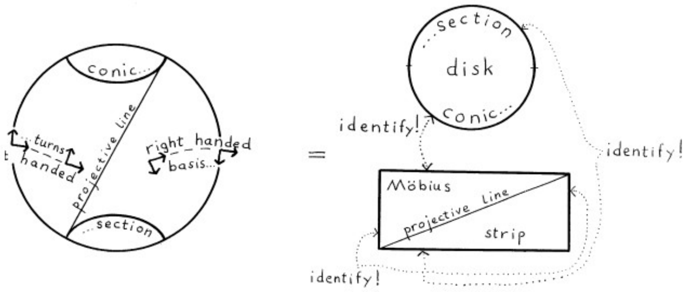

Figure 4.5. The topology of the real projective plane. For a different view, see Figure 4.12.

the projective plane starting from a three-dimensional vector space over the field of integers modulo p then we will obtain a projective plane with exactly $p+1$ points on each line. This is an interesting construction, but like real and quaternionic projective spaces it is of only secondary importance in quantum mechanics, where complex projective space occupies the centre of the stage.

4.2 Complex projective geometry

Complex projective space is in some respects easier to study than is its real cousin, mainly because the complex numbers form a *closed field* (that is to say, every polynomial equation has a root). The most effective presentation of $\mathbb{C}\mathbf{P}^n$ starts directly in the vector space $\mathbb{C}^N = \mathbb{C}^{n+1}$, known as *Hilbert space* in quantum mechanics; our convention is always that $N = n+1$. By definition $\mathbb{C}\mathbf{P}^n$ is the space of rays in $\mathbb{C}^{n+1}$, or equivalently the space of equivalence classes of $n+1$ complex numbers, not all zero, under

$$(Z^0, Z^1, \dots, Z^n) \sim \lambda(Z^0, Z^1, \dots, Z^n) \ ; \qquad \lambda \in \mathbb{C} \qquad \lambda \neq 0 \ . \tag{4.5}$$

We will use Greek indices to label the homogeneous coordinates. We can cover $\mathbb{C}\mathbf{P}^n$ with affine coordinate patches like

$$Z^\alpha = (1, z^1, \dots, z^n) \ , \qquad Z^\alpha = (z^{n'}, 1, z^{1'}, \dots, z^{(n-1)'}) \tag{4.6}$$

and so on. This leads to two important observations. The first is that one coordinate patch covers all of $\mathbb{C}\mathbf{P}^n$ except that subset which has (say)

$$Z^\alpha = (0, Z^1, \dots, Z^n) \ . \tag{4.7}$$

But the set of all such rays is a $\mathbb{C}\mathbf{P}^{n-1}$. Hence we conclude that topologically $\mathbb{C}\mathbf{P}^n$ is like $\mathbb{C}^n$ with a $\mathbb{C}\mathbf{P}^{n-1}$ attached 'at infinity'. Iterating this observation we obtain the *cell decomposition*

$$\mathbb{C}\mathbf{P}^n = \mathbb{C}^n \cup \mathbb{C}^{n-1} \cup \dots \cup \mathbb{C}^0 \ . \tag{4.8}$$

Evidently $\mathbb{C}\mathbf{P}^0$ is topologically a point and $\mathbb{C}\mathbf{P}^1$ is a 2-sphere, while for $n > 1$

we get something new that we will have to get used to as we proceed. The second observation is that in a region where the coordinate systems overlap, they are related by (say)

$$z^{a'} = \frac{Z^{a+1}}{Z^1} = \frac{Z^0}{Z^1}\frac{Z^{a+1}}{Z^0} = \frac{z^{a+1}}{z^1} \, , \tag{4.9}$$

which is clearly an analytic function on the overlap region. Hence $\mathbb{C}\mathbf{P}^n$ is a complex manifold.

The *linear subspaces* of $\mathbb{C}\mathbf{P}^n$ are of major importance. They are defined as the images of the subspaces of $\mathbb{C}^{n+1}$ under the natural map from the vector space to the projective space. Equivalently they are given by a suitable number of linear equations in the homogeneous coordinates. Thus the *hyperplanes* are $(n-1)$-dimensional submanifolds of $\mathbb{C}\mathbf{P}^n$ defined by the equation

$$P_\alpha Z^\alpha = 0 \tag{4.10}$$

for some fixed set of $n+1$ complex numbers P_α. This definition is unaffected by a change of scale for the homogeneous coordinates, and also by a change of scale in P_α. Hence the $n+1$ complex numbers P_α themselves are homogeneous coordinates for a $\mathbb{C}\mathbf{P}^n$; in other words the hyperplanes in a projective n-space can be regarded as the points of another projective n-space which is dual to the original.

If we impose a set of $m \leq n$ independent linear equations on the Z^α we obtain a linear subspace of complex dimension $n-m$ which is itself a $\mathbb{C}\mathbf{P}^{n-m}$. Geometrically this is the intersection of m hyperplanes. The space whose points consist of all such $(n-m)$-dimensional subspaces of $\mathbb{C}\mathbf{P}^n$ is known as a *Grassmannian* – it is only in the case of hyperplanes (and the trivial case of points) that the Grassmannian is itself a $\mathbb{C}\mathbf{P}^n$. A linear subspace of complex dimension one is known as a *complex projective line*, a linear subspace of dimension two is a complex projective plane and so on, while $\mathbb{C}\mathbf{P}^0$ is just a point. The complex projective line is a $\mathbb{C}\mathbf{P}^1$ – topologically this is a sphere and this may boggle some minds, but it is a line in the sense that one can draw a unique line between any pair of points (this is essentially the statement that two vectors in $\mathbb{C}^{n+1}$ determine a unique two-dimensional subspace). It also behaves like a line in the sense that two projective lines in a projective space intersect in a unique point if they intersect at all (this is the statement that a pair of two-dimensional subspaces are either disjoint or they share one common ray or they coincide). In general the intersection of two linear subspaces A and B is known as their *meet* $A \cap B$. We can also define their *join* $A \cup B$ by taking the linear span of the two subspaces of the underlying vector space to which A and B correspond, and then go back down to the projective space. These two operations – meet and join – turn the set of linear subspaces into a partially ordered structure known as a *lattice*, in which every pair of elements has a greatest lower bound (the meet) and a least upper bound (the join). This is the starting point of the subject known as quantum logic (Jauch, 1968; Varadarajan, 1985). It is also the second time we encounter a lattice – in Section 1.1 we came across the lattice of faces of a convex body.

In Chapter 8 we will find a convex body whose lattice of faces agrees with the lattice of subspaces of a vector space. By then it will be a more interesting lattice, because we will have an inner product on $\mathbb{C}^N$, so that a given subspace can be associated with its orthogonal complement.

But we have not yet introduced any inner product in $\mathbb{C}^N$, or any metric in $\mathbb{C}\mathbf{P}^n$. In fact we will continue to do without it for some time; projective geometry is precisely that part of geometry that can be done without a metric. There is an interesting group theoretical view of this, originated by Felix Klein. All statements about linear subspaces – such as when two linear subspaces intersect – are invariant under general linear transformations of $\mathbb{C}^N$. They form the group $GL(N, \mathbb{C})$, but only a subgroup acts *effectively* on $\mathbb{C}\mathbf{P}^n$. (Recall that $N = n+1$. A transformation is said to act effectively on some space if it moves at least one point.) Changing the matrix with an overall complex factor does not change the transformation effected on $\mathbb{C}\mathbf{P}^n$, so we can set its determinant equal to one and multiply it with an extra complex Nth root of unity if we wish. The *projective group* is therefore $SL(n+1, \mathbb{C})/\mathbb{Z}_N$.

According to Klein's conception projective geometry is completely characterized by the projective group; its subgroups include the group of affine transformations that preserves the $\mathbb{C}\mathbf{P}^{n-1}$ at infinity and this subgroup characterizes affine geometry. A helpful fact about the projective group is that any set of $n+2$ points can be brought to the standard position $(1, 0, \ldots, 0)$, $\ldots$, $(0, \ldots, 0, 1)$, $(1, 1, \ldots, 1)$ by means of a projective transformation. For $\mathbb{C}\mathbf{P}^1$ this is the familiar statement that any triple of points on the complex plane can be transformed to 0, 1, and ∞ by a Möbius transformation.

4.3 Complex curves, quadrics and the Segre embedding

The equation that defines a subspace of $\mathbb{C}\mathbf{P}^n$ does not have to be linear; any homogeneous equation in the homogeneous coordinates Z^α gives rise to a well-defined submanifold. Hence in addition to the linear subspaces we have *quadrics, cubics, quartics* and so on, depending on the degree of the defining polynomial. The locus of a number of homogeneous equations

$$w_1(Z) = w_2(Z) = \cdots = w_m(Z) = 0 \tag{4.11}$$

is also a subspace known as an *algebraic* (or projective) *variety*, and we move from projective to *algebraic* geometry.[1] *Chow's theorem* states that every non-singular algebraic variety is a complex submanifold of $\mathbb{C}\mathbf{P}^n$, and conversely every compact complex submanifold is the locus of a set of homogeneous equations. On the other hand it is not true that every complex manifold can be embedded as a complex submanifold in $\mathbb{C}\mathbf{P}^n$.

There are two kinds of submanifolds that are of immediate interest in quantum mechanics. One of them is the *complex curve*; by definition this is a map of $\mathbb{C}\mathbf{P}^1$ into $\mathbb{C}\mathbf{P}^n$. In real terms it is a 2-surface in a $2n$-dimensional space. This does not sound much like a curve, but once it is accepted that $\mathbb{C}\mathbf{P}^1$ deserves the name of line it will be admitted that the name curve is reasonable here. Let us first choose $n = 2$, so that we are looking for a complex curve in the projective plane. Using (u, v) as homogeneous coordinates on $\mathbb{C}\mathbf{P}^1$ we clearly get a map into $\mathbb{C}\mathbf{P}^2$ if we set

---

[1] A standard reference, stressing the setting of complex manifolds, is Griffiths and Harris (1978). For an alternative view see the book by Harris alone (Harris, 1992).

$$(u, v) \quad \rightarrow \quad (u^2, uv, v^2) \ . \tag{4.12}$$

(In Section 6.4 we will adjust conventions a little.) This is a well-defined map because the expression is homogeneous in the u and v. Evidently the resulting complex curve in $\mathbb{C}\mathbf{P}^2$ obeys the equation

$$Z^0 Z^2 \ = \ Z^1 Z^1 \ . \tag{4.13}$$

Hence it is a quadric hypersurface as well, and indeed any quadric can be brought to this form by projective transformations. In the projective plane a quadric is also known as a conic section.

In higher dimensions we encounter complex curves that are increasingly difficult to grasp since they are not quadrics. The next simplest case is the *twisted cubic* curve in $\mathbb{C}\mathbf{P}^3$, defined by

$$(u, v) \quad \rightarrow \quad (u^3, u^2 v, uv^2, v^3) \ . \tag{4.14}$$

We leave it aside for the moment though.

A class of submanifolds that is of particular interest in quantum mechanics arises in the following way. Suppose that the complex vector space is given as a *tensor product*

$$\mathbb{C}^{n+1} \otimes \mathbb{C}^{m+1} \ = \ \mathbb{C}^{(n+1)(m+1)} \ . \tag{4.15}$$

Then there should be an embedded submanifold

$$\mathbb{C}\mathbf{P}^n \times \mathbb{C}\mathbf{P}^m \ \in \ \mathbb{C}\mathbf{P}^{(n+1)(m+1)-1} \ . \tag{4.16}$$

Indeed this is true. In terms of homogeneous coordinates the submanifold can be parametrized as

$$Z^\alpha = Z^{\mu\mu'} = P^\mu Q^{\mu'}, \tag{4.17}$$

in a fairly obvious notation – the P^μ and $Q^{\mu'}$ are homogeneous coordinates on $\mathbb{C}\mathbf{P}^n$ and $\mathbb{C}\mathbf{P}^m$, respectively. The construction is known as the *Segre embedding.* The submanifold is a Cartesian product with (complex) dimension $n+m$, and it follows from Chow's theorem that it can be defined as the locus of nm homogeneous equations in the large space. Indeed it is easy to see from the definition that the submanifold will be the intersection of the quadrics

$$Z^{\mu\mu'} Z^{\nu\nu'} - Z^{\mu\nu'} Z^{\nu\mu'} = 0 \ . \tag{4.18}$$

In quantum mechanics – and in Section 15.2 – the Segre submanifold reappears as the set of separable states of a composite system.

Let us consider the simplest case

$$\mathbb{C}\mathbf{P}^1 \times \mathbb{C}\mathbf{P}^1 \ \in \ \mathbb{C}\mathbf{P}^3 \ , \tag{4.19}$$

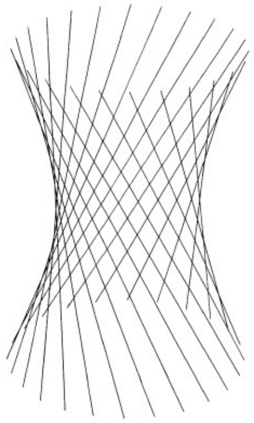

Figure 4.6. The real Segre embedding: a hyperboloid ruled by straight lines.

when only one equation is needed. Write

$$(Z^0, Z^1, Z^2, Z^3) = (Z^{00'}, Z^{01'}, Z^{10'}, Z^{11'}) \ . \tag{4.20}$$

Then the submanifold $\mathbb{C}\mathbf{P}^1 \times \mathbb{C}\mathbf{P}^1$ is obtained as the quadric surface

$$Z^0Z^3 - Z^1Z^2 = 0 \ . \tag{4.21}$$

In general the non-degenerate quadrics in $\mathbb{C}\mathbf{P}^3$ are in one-to-one correspondence to all possible embeddings of $\mathbb{C}\mathbf{P}^1 \times \mathbb{C}\mathbf{P}^1$. It is interesting to consider the projection map from the product manifold to one of its factors. This means that we hold $Q^{\mu'}$ (say) fixed and vary P^μ. Then the fibre of the map – the set of points on the quadric that are projected to the point on $\mathbb{C}\mathbf{P}^m$ that is characterized by that particular $Q^{\mu'}$ – will be given by the equations

$$\frac{Z^{00'}}{Z^{01'}} = \frac{P^0Q^{0'}}{P^0Q^{1'}} = \frac{Q^{0'}}{Q^{1'}} = \frac{P^1Q^{0'}}{P^1Q^{1'}} = \frac{Z^{10'}}{Z^{11'}} \ . \tag{4.22}$$

Since $Q^{\mu'}$ is fixed this implies that there is a complex number λ such that

$$Z^0 = \lambda Z^1 \quad \text{and} \quad Z^2 = \lambda Z^3 \ . \tag{4.23}$$

This pair of linear equations defines a projective line in $\mathbb{C}\mathbf{P}^3$. Projecting down to the other factor leads to a similar conclusion. In this way we see that the quadric is ruled by lines, or in other words that through any point on the quadric there goes a pair of straight lines lying entirely in the quadric.

For visualization, let us consider the real Segre embedding

$$\mathbb{R}\mathbf{P}^1 \times \mathbb{R}\mathbf{P}^1 \in \mathbb{R}\mathbf{P}^3 \ . \tag{4.24}$$

This time we choose to diagonalize the quadric; define

$$Z^0 = X + U \ , \quad Z^1 = X - U \ , \quad Z^2 = V + Y \ , \quad Z^3 = V - Y \ . \tag{4.25}$$

We then obtain

$$Z^0Z^3 - Z^1Z^2 = X^2 + Y^2 - U^2 - V^2 = 0 \ . \qquad (4.26)$$

Now let us choose affine coordinates by dividing through with V. The quadric becomes

$$x^2 + y^2 - u^2 \ = \ 1 \ . \qquad (4.27)$$

This is a hyperboloid of one sheet sitting in a three-dimensional real space. The fact that such a surface can be *ruled* by straight lines is a surprising fact of elementary geometry (first noted by Sir Christopher Wren).

4.4 Stars, spinors and complex curves

The *stellar representation* is a delightful way of visualizing $\mathbb{C}\mathbf{P}^n$ in real terms. It works for any n and simplifies some problems in a remarkable way.[2] Here we will develop it a little from a projective point of view, while Chapter 7 discusses the same construction with the added benefit of a metric. The idea is that vectors in $\mathbb{C}^{n+1}$ are in one-to-one correspondence with the set of nth degree polynomials in one complex variable z, such as

$$w(z) = Z^0z^n + Z^1z^{n-1} + \cdots + Z^n \ . \qquad (4.28)$$

(The important point is that we have a polynomial of the nth degree; in Chapter 7 we will have occasion to polish the conventions a little.) We can rescale the vector Z^α so that $Z^0 = 1$; therefore points in $\mathbb{C}\mathbf{P}^n$ will be in one-to-one correspondence with unordered sets of n complex numbers, namely with the complex roots of the equation

$$Z^0z^n + Z^1z^{n-1} + \cdots + Z^{n+1} = 0 = Z^0(z - z_1)(z - z_2)\ldots(z - z_n) \ . \qquad (4.29)$$

Multiple roots are allowed. If $Z^0 = 0$ then infinity counts as a root (of multiplicity m if $Z^1 = \cdots = Z^{m-1} = 0$). Finally, by means of a stereographic projection the roots can be represented as unordered sets of n points on an ordinary 2-sphere – the points are called 'stars' and thus we have arrived at the stellar representation, in which points in $\mathbb{C}\mathbf{P}^n$ are represented by n unordered stars on a 'celestial' sphere. As a mathematical aside it follows that $\mathbb{C}\mathbf{P}^1 = \mathbf{S}^2$, $\mathbb{C}\mathbf{P}^2 = \mathbf{S}^2 \times \mathbf{S}^2/S_2$ and in general that $\mathbb{C}\mathbf{P}^n = \mathbf{S}^2 \times \mathbf{S}^2 \times \cdots \times \mathbf{S}^2/S_n$, where S_n is the *symmetric group* of permutations of n objects.

There is a piece of notation that is conveniently introduced at this juncture. Let us denote the homogeneous coordinates on $\mathbb{C}\mathbf{P}^1$ by

$$\zeta^A = (u, v) \ . \qquad (4.30)$$

So we use a capital Latin letter for the index and we will refer to ζ^A as a

---

[2] The first reference appears to be to Majorana (1932). The result has been rediscovered many times within (Bacry, 1974) and without (Penrose, 1960) the context of quantum mechanics.

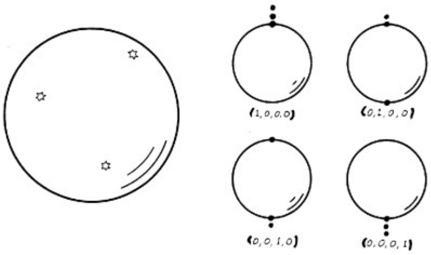

Figure 4.7. Any state from $\mathbb{C}\mathbf{P}^3$ may be represented by three stars on the sphere. Some special cases, where the stars coincide, are shown on the right.

spinor. The overall scale of the spinor is irrelevant to us, so we can introduce an affine coordinate z by

$$z = \frac{v}{u} \ : \qquad \zeta^A \sim (1, z) \ . \tag{4.31}$$

A spinor for which $u = 0$ then corresponds to the south pole on the Riemann sphere. We will use the totally anti-symmetric tensor ϵ_{AB} (the symplectic structure on $\mathbf{S}^2$, in fact) to raise and lower indices according to

$$\zeta_A \equiv \zeta^B \epsilon_{BA} \qquad \Leftrightarrow \qquad \zeta^A = \epsilon^{AB} \zeta_B \ . \tag{4.32}$$

Due to the fact that ϵ_{AB} is anti-symmetric there is a definite risk that sign errors will occur when one uses its inverse ϵ^{AB}. Simply stick to the convention just made and all will be well. Note that $\zeta_A \zeta^A = 0$. So far then a spinor simply denotes a vector in $\mathbb{C}^2$. We do not think of this as a Hilbert space yet because there is no inner product. On the contrary the formalism is manifestly invariant under the full projective group $SL(2, \mathbb{C})/\mathbb{Z}^2$. A special linear transformation on $\mathbb{C}^2$ gives rise to a *Möbius transformation* on the sphere;

$$\zeta^A \rightarrow \begin{bmatrix} u' \\ v' \end{bmatrix} = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \qquad \Rightarrow \qquad z \rightarrow z' = \frac{\alpha z + \beta}{\gamma z + \delta} \ , \tag{4.33}$$

where $\alpha\delta - \beta\gamma = 1$ is the condition (on four complex numbers) that guarantees that the matrix belongs to $SL(2, \mathbb{C})$ and that the transformation preserves ϵ_{AB}. The group of Möbius transformations is exactly the group of projective transformations of $\mathbb{C}\mathbf{P}^1$.

We can go on to consider totally symmetric *multispinors*

$$\Psi^{AB} = \Psi^{(AB)} \ , \qquad \Psi^{ABC} = \Psi^{(ABC)} \tag{4.34}$$

and so on. (The brackets around the indices mean that we are taking the totally symmetric part.) It is easy to see that when the index ranges from zero to one then the number of independent components of a rank n totally symmetric multispinor is $n + 1$, just right so that the multispinor can serve as homogeneous coordinates for $\mathbb{C}\mathbf{P}^n$. To see that this works, consider the equation

$$\Psi_{AB\dots M}\zeta^A\zeta^B\dots\zeta^M = 0 \ . \tag{4.35}$$

If we now choose the scale so that $\zeta^A = (1, z)$ then the above equation turns into an nth degree polynomial in z, having n complex roots. We can use this fact together with Eq. (4.29) to translate between the $\Psi^{AB\dots M}$ and the Z^α, should this be needed. We can also use the fundamental theorem of algebra to rewrite the polynomial as a product of n factors. In analogy with Eq. (4.29) we find that

$$\Psi_{AB\dots M}\zeta^A\zeta^B\dots\zeta^M = 0 = (\alpha_0 + \alpha_1 z)(\beta_0 + \beta_1 z)\dots(\mu_0 + \mu_1 z) \ . \tag{4.36}$$

The conclusion is that a rank n multispinor can be written – uniquely except for an overall factor of no interest to us – as a symmetrized product of n spinors:

$$\Psi_{AB\dots M}\zeta^A\zeta^B\dots\zeta^M = \alpha_A\zeta^A\beta_B\zeta^B\dots\mu_M\zeta^M \quad \Rightarrow \quad \Psi_{AB\dots M} = \alpha_{(A}\beta_B\dots\mu_{M)} \ . \tag{4.37}$$

The factors are known as *principal spinors*, and they are of course the n unordered points (note the symmetrization) on the sphere in slight disguise.

The stellar representation, or equivalently the spinor notation, deals with linear subspaces in an elegant way. Consider a complex projective line (a $\mathbb{C}\mathbf{P}^1$) in $\mathbb{C}\mathbf{P}^2$ for definiteness. A general point in $\mathbb{C}\mathbf{P}^2$ is described by a pair of unordered points on the 2-sphere, or equivalently as a spinor

$$\Psi^{AB} = \alpha^{(A}\beta^{B)} \equiv \frac{1}{2}\big(\alpha^A\beta^B + \alpha^B\beta^A\big) \ . \tag{4.38}$$

Evidently we get a complex projective line by holding one of the points fixed and letting the other vary, that is by holding one of the principal spinors (say β^A) fixed and letting the other vary.

The spinor notation also deals elegantly with complex curves in general. Thus we get a conic section in $\mathbb{C}\mathbf{P}^2$ as the set of points for which the principal spinors coincide. That is to say that

$$\Psi^{AB} = \Psi^A\Psi^B \tag{4.39}$$

for some spinor Ψ^A. Through any point on the quadric (for which $\Psi^A = \alpha^A$ say) there goes a complex projective line

$$\Psi^{AB} = \zeta^{(A}\alpha^{B)} \tag{4.40}$$

(where ζ^A varies). This line is *tangent* to the quadric since it touches the quadric in a unique point $\alpha^A\alpha^B$. It is moreover rather easy to see that a pair of tangent lines always intersect in a unique point. See Figure 4.8.

4.5 The Fubini–Study metric

Now we want a notion of distance, and indeed a Riemannian metric, in $\mathbb{C}\mathbf{P}^n$. How do we do it? Pick an arbitrary pair of points. The distance between them

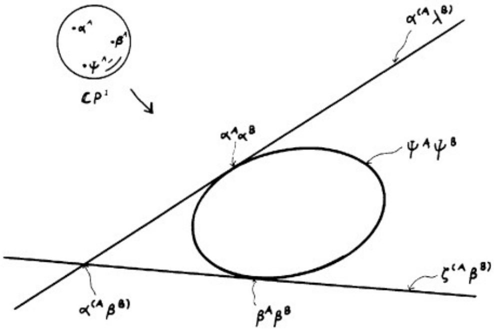

Figure 4.8. A conic section in $\mathbb{C}\mathbf{P}^2$ and a pair of tangent lines.

will be the length of the geodesic curve connecting them. On the other hand we know that there is a unique projective line connecting them; topologically this is $\mathbb{C}\mathbf{P}^1 = \mathbf{S}^2$. Suppose that we insist that this 2-sphere is a round 2-sphere, with the ordinary round metric. Let us also insist that the metric on $\mathbb{C}\mathbf{P}^n$ is such that the projective lines are *totally geodesic*. Technically a submanifold is said to be totally geodesic if a geodesic with respect to the induced metric on the submanifold is also a geodesic with respect to the metric on the embedding space, or equivalently if a geodesic that starts out parallel to the submanifold stays in the submanifold. But a geodesic on a complex projective line is simply a great circle on a 2-sphere, so we have found a way to define geodesics between two arbitrary points in $\mathbb{C}\mathbf{P}^n$. The resulting notion of distance is called the *Fubini–Study distance.*[3]

It only remains to make this definition of the metric on $\mathbb{C}\mathbf{P}^n$ explicit. Since the geodesic lives on some complex projective line, we can write down its equation using nothing but the results of Chapter 3. Let us recall Eq. (3.85) for a geodesic on an odd-dimensional sphere:

$$Z^\alpha(\sigma) = Z^\alpha(0)\cos\sigma + \dot{Z}^\alpha(0)\sin\sigma \ , \tag{4.41}$$

where

$$Z(0)\cdot\bar{Z}(0) = \dot{Z}(0)\cdot\dot{\bar{Z}}(0) = 1 \ , \qquad Z(0)\cdot\dot{\bar{Z}}(0) + \dot{Z}(0)\cdot\bar{Z}(0) = 0 \ . \tag{4.42}$$

If $\dot{Z}^\alpha(0) = \mathrm{i}Z^\alpha(0)$ this is a Hopf circle, and projects to a point on $\mathbb{C}\mathbf{P}^1$ $\mathbb{C}\mathbf{P}^1 = \mathbf{S}^2$. In general we can write

$$Z^\alpha(0) \equiv m^\alpha \ , \qquad \dot{Z}^\alpha(0) = m^\alpha\cos a + n^\alpha\sin a \ , \qquad m\cdot\bar{n} = 0 \ . \tag{4.43}$$

Since the constant vectors m^α and n^α have been chosen orthogonal, they lie antipodally on the particular complex projective line that they span. We can now appeal to the fibre bundle perspective (Section 4.8) to see that this curve is horizontal, and hence projects to a geodesic on $\mathbb{C}\mathbf{P}^1$, if and only if

---

[3] It was first studied by two leading geometers from a century ago; Fubini (1903) and Study (1905).

$\dot{Z} \cdot \bar{Z} = 0$, that is to say if and only if $\cos a = 0$. Alternatively we can use a direct calculation to show that for general a they project to latitude circles, and to great circles if and only if $\cos a = 0$. Either way, we conclude that our definition implies that a geodesic on $\mathbb{C}\mathbf{P}^n$ is given by

$$Z^\alpha(\sigma) = m^\alpha \cos\sigma + n^\alpha \sin\sigma \ , \qquad m \cdot \bar{m} = n \cdot \bar{n} = 1 \ , \ m \cdot \bar{n} = 0 \ . \qquad (4.44)$$

More precisely, this is the horizontal lift of a geodesic on $\mathbb{C}\mathbf{P}^n$ to the odd-dimensional sphere $\mathbf{S}^{2n+1}$, in the fibre bundle $\mathbb{C}\mathbf{P}^n = \mathbf{S}^{2n+1}/\mathbf{S}^1$.

We will now use this set-up to define the Fubini–Study distance D_{FS} between any two points on $\mathbb{C}\mathbf{P}^n$. Since everything takes place within a complex projective line, which lifts to a 3-sphere, we can use the expression for the distance arrived at in our discussion of the Hopf fibration of $\mathbf{S}^3$, namely Eq. (3.95). The Fubini–Study distance D_{FS} must be given by

$$\cos^2 D_{\text{FS}} \ = \ \kappa \ , \qquad (4.45)$$

where the *projective cross-ratio* between two points P^α and Q^α is given by

$$\kappa \ = \ \frac{P \cdot \bar{Q}\ Q \cdot \bar{P}}{P \cdot \bar{P}\ Q \cdot \bar{Q}} \ . \qquad (4.46)$$

We give the same formula in standard quantum mechanical notation in Eq. (5.16). There is one new feature: in Section 3.5 we assumed that $P \cdot \bar{P} = Q \cdot \bar{Q} = 1$, but this assumption has now been dropped. As in Section 4.1 there is a polarity – that is to say a map from the space of points to the space of hyperplanes – hidden here: the polarity maps the point P^α to that special hyperplane consisting of all points Q^α with

$$Q \cdot \bar{P} \ \equiv \ Q^\alpha \bar{P}_\alpha \ = \ 0 \ . \qquad (4.47)$$

Interestingly complex conjugation appears here for the first time. Another interesting fact is that our definitions imply that $\mathbb{C}\mathbf{P}^n$ is a C_π manifold: all geodesics are closed, and have the same circumference. This is a quite exceptional property, related to the fact that the isotropy group – the subgroup of isometries leaving a given point invariant – is so large that it acts transitively on the space of directions there. Spheres have this property too. The circumference of a geodesic on $\mathbb{C}\mathbf{P}^n$ equals π because $\mathbb{C}\mathbf{P}^1$ is a 2-sphere of radius $1/2$; the geodesic on $\mathbf{S}^{2n+1}$ has circumference 2π and doubly covers the geodesic on $\mathbb{C}\mathbf{P}^n$. We also note that the maximal distance between two points is equal to $\pi/2$ (when the two vectors are orthogonal), and that all geodesics meet again at a point of maximal distance from their point of origin.

To obtain the line element in a local form we assume that

$$Q^\alpha \ = \ P^\alpha + \mathrm{d}P^\alpha \qquad (4.48)$$

and expand to second order in the vector $\mathrm{d}P^\alpha$ (and to second order in D_{FS}). The result is the *Fubini–Study metric*

$$\mathrm{d}s^2 = \frac{\mathrm{d}P \cdot \mathrm{d}\bar{P}\ P \cdot \bar{P} - \mathrm{d}P \cdot \bar{P}\ P \cdot \mathrm{d}\bar{P}}{P \cdot \bar{P}\ P \cdot \bar{P}} \ . \qquad (4.49)$$

Before we are done we will become very familiar with this expression.

It is already clear that many formulae will be simplified if we normalize our vectors so that $P \cdot \bar{P} = 1$. In quantum mechanics, where the vectors are the usual Hilbert space state vectors, this is usually done. In this chapter we will use normalized vectors now and then in calculations, but on festive occasions – say when stating a definition – we do not.

An important definition follows immediately, namely that of the 2-form

$$\Omega \;=\; \mathrm{i}\, \frac{P \cdot \bar{P}\; \mathrm{d}P \cdot \wedge \mathrm{d}\bar{P} - \mathrm{d}P \cdot \bar{P} \wedge P \cdot \mathrm{d}\bar{P}}{P \cdot \bar{P}\, P \cdot \bar{P}} \;. \tag{4.50}$$

This is clearly a relative of the metric and the suggestion is that it is a (closed) symplectic form and that $\mathbb{C}\mathbf{P}^n$ is a Kähler manifold for all n. And this is true. To prove the Kähler property it is helpful to use the affine coordinates from Section 4.2. When expressed in terms of them the Fubini–Study metric becomes

$$\mathrm{d}s^2 = \frac{1}{1+|z|^2}\left(\mathrm{d}z^a \mathrm{d}\bar{z}_a - \frac{\bar{z}_a \mathrm{d}z^a \mathrm{d}\bar{z}_b z^b}{1+|z|^2}\right) \equiv 2 g_{a\bar{b}}\; \mathrm{d}z^a \mathrm{d}\bar{z}^{\bar{b}} \;, \qquad |z|^2 \equiv z^a \bar{z}_a \;. \tag{4.51}$$

(We are using $\delta_{a\bar{b}}$ to change $\bar{z}^{\bar{a}}$ into $\bar{z}_a$.) Now

$$2 g_{a\bar{b}} = \frac{1}{1+|z|^2}\left(\delta_{a\bar{b}} - \frac{\bar{z}_a z_{\bar{b}}}{1+|z|^2}\right) = \partial_a \partial_{\bar{b}} \ln\left(1+|z|^2\right) \;. \tag{4.52}$$

Or more elegantly

$$g_{a\bar{b}} \;=\; \frac{1}{2} \partial_a \partial_{\bar{b}} \ln Z \cdot \bar{Z} \;, \tag{4.53}$$

where it is understood that the homogeneous coordinates should be expressed in terms of the affine ones. As we know from Section 3.3 this proves that $\mathbb{C}\mathbf{P}^n$ is a Kähler manifold. Every complex submanifold is Kähler too. For $n = 1$ we recognize the metric on $\mathbf{S}^2$ written in stereographic coordinates.

We will use a different choice of coordinate system to explore the space in detail, but the affine coordinates are useful for several purposes, for instance if one wants to perform an explicit check that the curve (4.44) is indeed a geodesic, or when one wants to compute the Riemann curvature tensor. Thus:

$$g^{a\bar{b}} = 2(1+|z|^2)(\delta^{a\bar{b}} + z^a \bar{z}^{\bar{b}}) \tag{4.54}$$

$$\Gamma_{bc}{}^{a} \;=\; -\frac{1}{1+|z|^2}\left(\delta^a_b \bar{z}_c + \delta^a_c \bar{z}_b\right) \tag{4.55}$$

$$R_{a\bar{b}c\bar{d}} \;=\; -2(g_{a\bar{b}} g_{c\bar{d}} + g_{a\bar{d}} g_{c\bar{b}}) \tag{4.56}$$

$$R_{a\bar{b}} \;=\; 2(n+1) g_{a\bar{b}} \;. \tag{4.57}$$

All components not related to these through index symmetries or complex conjugation are zero. From the expression for the Ricci tensor $R_{a\bar{b}}$ we see that its traceless part vanishes; the Fubini–Study metric solves the Euclidean Einstein equations with a cosmological constant. The form of the full Riemann tensor $R_{a\bar{b}c\bar{d}}$ shows that the space has constant holomorphic sectional curvature (but it does not show that the curvature is constant, and in fact it is not – our result is weaker basically because $g_{a\bar{b}}$ is just one 'block' of the metric tensor). There is an important theorem that says that a simply connected complex manifold of constant holomorphic sectional curvature is necessarily isometric with $\mathbb{C}\mathbf{P}^n$ or else it has the topology of an open unit ball in $\mathbb{C}^n$, depending on the sign of the curvature. (If it vanishes the space is $\mathbb{C}^n$.) The situation is clearly analogous to that of simply connected two-dimensional manifolds of constant sectional curvature, which are either spheres, flat spaces, or hyperbolic spaces with the topology of a ball.

We observe that both the Riemannian and the symplectic geometry relies on the Hermitian form $Z \cdot \bar{Z} = Z^\alpha \bar{Z}_\alpha$ in $\mathbb{C}^N$. Therefore distances and areas are invariant only under those projective transformations that preserve this form. These are precisely the unitary transformations, with anti-unitary transformations added if we allow the transformations to flip the sign of the symplectic form. In quantum mechanics this theorem goes under the name of Wigner's theorem (Wigner, 1959):

Theorem 4.1 (Wigner's) *All isometries of* $\mathbb{C}\mathbf{P}^n$ *arise from unitary or anti-unitary transformations of* $\mathbb{C}^N$.

Since only a part of the unitary group acts effectively the connected component of the isometry group is in fact $SU(N)/\mathbb{Z}_N$. For $N = 2$ we are dealing with a sphere and $SU(2)/\mathbb{Z}_2 = SO(3)$, that is the group of proper rotations. The full isometry group $O(3)$ includes reflections and arises when anti-unitary transformations are allowed.

We obtain an infinitesimal isometry by choosing an Hermitian matrix – a generator of $U(N)$ – and writing

$$\mathrm{i}\,\dot{Z}^\alpha = H^\alpha{}_\beta Z^\beta\ . \tag{4.58}$$

This equation – to reappear as the *Schrödinger equation* in Section 5.1 – determines a Killing flow on $\mathbb{C}\mathbf{P}^n$. A part of it represents 'gauge', that is changes in the overall phase of the homogeneous coordinates. Therefore we write the projective equation

$$\mathrm{i}Z^{[\alpha}\dot{Z}^{\beta]} = Z^{[\alpha}H^{\beta]}{}_\gamma Z^\gamma\ , \tag{4.59}$$

where the brackets denote anti-symmetrization. This equation contains all the information concerning the Killing flow on $\mathbb{C}\mathbf{P}^n$ itself, and is called the *projective Schrödinger equation* (Hughston, 1995, p. 59). The fixed points of the flow occur when the right-hand side vanishes. Because of the anti-symmetrized indices this happens when Z^α is an eigenvector of the Hamiltonian,

$$H^\alpha{}_\beta Z^\beta = EZ^\alpha\ . \tag{4.60}$$

A picture will be given later.

There are other ways of introducing the Fubini–Study metric. We usually understand $\mathbb{C}\mathbf{P}^1$, that is the 2-sphere, as a subset embedded in $\mathbb{R}^3$. Through the Hopf fibration we can also understand it as the base space of a fibre bundle whose total space is either $\mathbf{S}^3$ or $\mathbb{C}^2$ with the origin excluded. Both of these pictures can be generalized to $\mathbb{C}\mathbf{P}^n$ for arbitrary n. The fibre bundle picture will be discussed in more detail later. The embedding picture relies on an embedding into $\mathbb{R}^{(n+1)^2-1}$. The dimension is rather high but we will have to live with this. To see how it works, use homogeneous coordinates to describe $\mathbb{C}\mathbf{P}^n$ and form the matrix

$$\rho^{\alpha}{}_{\beta} \ \equiv \ \frac{Z^{\alpha}\bar{Z}_{\beta}}{Z \cdot \bar{Z}} \ . \tag{4.61}$$

This is a useful way to represent a point in $\mathbb{C}\mathbf{P}^n$ since the reduncancy of the homogeneous coordinates has disappeared. The resulting matrix is Hermitian and has trace unity. Hence we have an embedding of $\mathbb{C}\mathbf{P}^n$ into the space of Hermitian $N \times N$ matrices with trace unity, which has the dimension advertised, and can be made into a flat space in a natural way. We simply set

$$D^2(A,B) \ = \ \frac{1}{2}\mathrm{Tr}(A-B)^2 \ . \tag{4.62}$$

This expression defines a distance D between the matrices A and B (equal to the *Hilbert–Schmidt distance* introduced in Chapter 8). It is analogous to the chordal distance between two points on the sphere when embedded in a flat space in the standard way.

To see that the embedding gives rise to the Fubini–Study metric we take two matrices representing different points of $\mathbb{C}\mathbf{P}^n$, such as

$$\rho^{\alpha}_{P\ \beta} = \frac{P^{\alpha}\bar{P}_{\beta}}{P \cdot \bar{P}} \qquad \rho^{\alpha}_{Q\ \beta} = \frac{Q^{\alpha}\bar{Q}_{\beta}}{Q \cdot \bar{Q}} \ , \tag{4.63}$$

and compute

$$D^2(\rho_P, \rho_Q) = 1 - \kappa = 1 - \cos^2 D_{\mathrm{FS}} \ , \tag{4.64}$$

where κ is the projective cross-ratio (4.46) and D_{FS} is the Fubini–Study distance between P and Q along a curve lying within the embedded $\mathbb{C}\mathbf{P}^n$. If the points are very close we can expand the square of the cosine to second order and obtain

$$D \ = \ D_{\mathrm{FS}} + \text{higher order terms} \ . \tag{4.65}$$

This proves our point. The Riemannian metric

$$\mathrm{d}s^2 = \frac{1}{2}\,\mathrm{Tr}\ \mathrm{d}\rho\,\mathrm{d}\rho, \tag{4.66}$$

is precisely the Fubini–Study metric, provided that it is evaluated at a point where the matrix ρ is an outer product of two vectors – and $\mathrm{d}\rho$ is the 'infinitesimal' difference between two such matrices, or more precisely a tangent vector to $\mathbb{C}\mathbf{P}^n$. This embedding will reappear in Chapter 8 as the embedding of pure states into the space of density matrices.

4.6 $\mathbb{C}\mathbf{P}^n$ illustrated

At this point we still do not know what $\mathbb{C}\mathbf{P}^n$ 'looks like', even in the modest sense that we feel that we know what an n-sphere 'looks like' for arbitrary n. There is a choice of coordinate system that turns out to be surprisingly helpful in this regard (Barros e Sá, 2001a; Bengtsson, Brännlund and Życzkowski, 2002). Define

$$(Z^0, Z^1, \ldots, Z^n) = (n_0, n_1 \mathrm{e}^{\mathrm{i}\nu_1}, \ldots, n_n \mathrm{e}^{\mathrm{i}\nu_n}) \ , \tag{4.67}$$

where $0 \leq \nu_i < 2\pi$ and the real numbers n_0, n_i are non-negative, $n_0 \geq 0$, $n_i \geq 0$, and obey the constraint

$$n_0^2 + n_1^2 + \cdots + n_n^2 = 1 \ . \tag{4.68}$$

We call such coordinates *octant coordinates*, because $n_0, n_1, \ldots$ clearly form the positive hyperoctant of an n-sphere. The phases ν_i form an n-torus, so we already see a picture of the topology of $\mathbb{C}\mathbf{P}^n$ emerging: we have a set of tori parametrized by the points of a hyperoctant, with the proviso that the picture breaks down at the edges of the hyperoctant, where the phases are undefined.[4]

To go all the way to a local coordinate system we can set

$$\begin{cases} n_0 &= \cos\vartheta_1 \sin\vartheta_2 \sin\vartheta_3 \ \ldots \ \sin\vartheta_n \\ n_1 &= \sin\vartheta_1 \sin\vartheta_2 \sin\vartheta_3 \ \ldots \ \sin\vartheta_n \\ n_2 &= \qquad \cos\vartheta_2 \sin\vartheta_3 \ \ldots \ \sin\vartheta_n \\ \ \vdots & \ \vdots \qquad\qquad \vdots \\ n_n &= \qquad\qquad \cos\vartheta_n \end{cases} \qquad 0 < \vartheta_i < \frac{\pi}{2} \ . \tag{4.69}$$

This is just like Eq. (1.13), except that the range of the coordinates has changed. We can also use n_i as orthographic coordinates. Alternatively we can set

$$y_i \ = \ n_i^2 \tag{4.70}$$

and use the n coordinates y_i. This looks contrived at this stage but will suggest itself later. If we also define $y_0 = n_0^2$ we clearly have

$$y_0 + y_1 + \cdots + y_n = 1 \ . \tag{4.71}$$

A probability interpretation is not far behind ... .

To see how it all works, let us consider $\mathbb{C}\mathbf{P}^2$. The point is that octant coordinates are quite well adapted to the Fubini–Study metric, which becomes

$$\begin{aligned} \mathrm{d}s^2 = {} & \mathrm{d}n_0^2 + \mathrm{d}n_1^2 + \mathrm{d}n_2^2 \\ & + n_1^2(1 - n_1^2)\,\mathrm{d}\nu_1^2 + n_2^2(1 - n_2^2)\,\mathrm{d}\nu_2^2 - 2n_1^2 n_2^2 \ \mathrm{d}\nu_1\,\mathrm{d}\nu_2 \ . \end{aligned} \tag{4.72}$$

The first piece here, given Eq. (4.68), is recognizable as the ordinary round metric on the sphere. The second part is the metric on a flat torus, whose shape depends on where we are on the octant. Hence we are justified in thinking of

---

[4] There is an entire branch of mathematics called 'toric geometry' whose subject matter, roughly speaking, consists of spaces that can be described in this way. See Ewald (1996) for more on this.

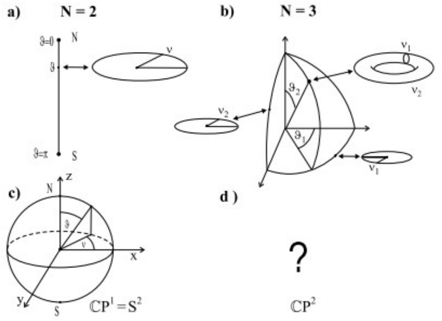

Figure 4.9. $\mathbb{C}\mathbf{P}^2$ may be visualized as the positive octant of a 2-sphere. Each point inside the octant represents a torus T^2 spanned by the phases $(\nu_1,\ \nu_2)$ (b). Each point at the edges of the octant denotes a circle, so each of three edges representes a 2-sphere. For comparison we plot $\mathbb{C}\mathbf{P}^1$ in the same manner, in panel (a). A realistic view of $\mathbb{C}\mathbf{P}^1$ is shown in panel (c); an analogous view of $\mathbb{C}\mathbf{P}^2$ would be more difficult (d).

$\mathbb{C}\mathbf{P}^2$ as a set of flat 2-tori parametrized by a round octant of a 2-sphere. There is an evident generalization to all n, and in fact we can draw accurate pictures of $\mathbb{C}\mathbf{P}^3$ if we want to.

For $n = 1$ we obtain a one parameter family of circles that degenerate to points at the end of the interval; a moment's thought will convince the reader that this is simply a way to describe a 2-sphere. Just to make sure, use the angular coordinates and find that Eq. (4.72) becomes

$$\mathrm{d}s^2 = \mathrm{d}\vartheta_1^2 + \frac{1}{4}\sin^2(2\vartheta_1)\,\mathrm{d}\nu_1^2 = \frac{1}{4}\left(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\right)\,, \tag{4.73}$$

where we used $\theta = 2\vartheta_1$ and $\phi = \nu_1$ in the second step.

To make the case $n = 2$ quite clear we make a map of the octant, using a stereographic or a gnomonic projection (Section 3.1). The latter is quite useful here and it does not matter that only half the sphere can be covered since we need to cover only one octant anyway. It is convenient to centre the projection at the centre of the octant and adjust the coordinate plane so that the coordinate distance between a pair of corners of the resulting triangle equals one. The result is shown in Figure 4.10.

This takes care of the octant. We obtain a picture of $\mathbb{C}\mathbf{P}^2$ when we remember that each interior point really represents a flat torus, conveniently regarded as a parallelogram with opposite sides identified. The shape of the parallelogram – discussed in Section 3.3 – is relevant. According to Eq. (4.72) the lengths of the sides are

$$L_1 = \int_0^{2\pi} \mathrm{d}s = 2\pi n_1\sqrt{1-n_1^2} \quad \text{and} \quad L_2 = 2\pi n_2\sqrt{1-n_2^2}\,. \tag{4.74}$$

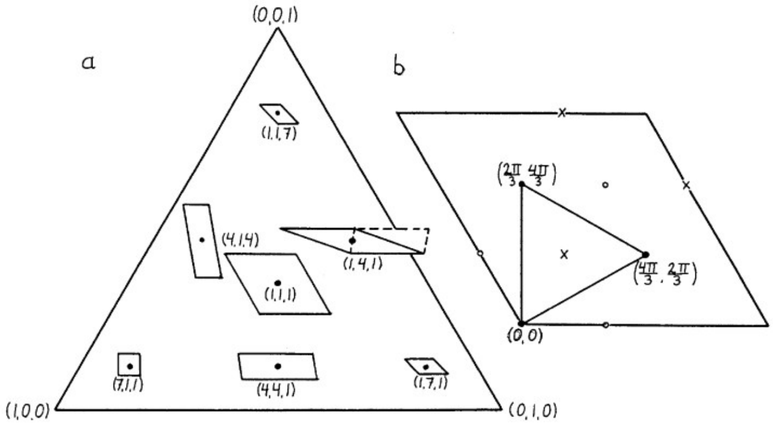

Figure 4.10. In (a) we indicate how the torus lying over each interior point changes with position in the octant. The position in the octant is given by an unnormalized vector. At the edges the tori degenerate to circles so edges are complex projective lines. The corners of the octant represent orthogonal points. It may be convenient to perform some cutting and gluing of the parallelogram before thinking about the shape of the torus it defines, as indicated with dashed lines for the torus lying over the point $(1, 4, 1)$. The size of the octant relative to that of the tori is exaggerated in the picture. To bring this home we show, in (b), the largest torus – the one sitting over $(1, 1, 1)$ – decorated with three times three points (marked with crosses and filled or unfilled dots). Each such triple corresponds to an orthogonal basis. The coordinates (ν_1, ν_2) are given for one of the triples.

The angle between them is given by

$$\cos\theta_{12} = -\frac{n_1 n_2}{\sqrt{1-n_1^2}\,\sqrt{1-n_2^2}} \ . \tag{4.75}$$

The point is that the shape depends on where we are on the octant. So does the total area of the torus,

$$A = L_1 L_2 \sin\theta_{12} = 4\pi^2 n_0 n_1 n_2 \ . \tag{4.76}$$

The 'biggest' torus occurs at the centre of the octant. At the boundaries the area of the tori is zero. This is because there the tori degenerate to circles. In effect an edge of the octant is a one parameter family of circles, in other words it is a $\mathbb{C}\mathbf{P}^1$.

It is crucial to realize that there is nothing special going on at the edges and corners of the octant, whatever the impression left by the map may be. Like the sphere, $\mathbb{C}\mathbf{P}^n$ is a homogeneous space and looks the same from every point. To see this, note that any choice of an orthogonal basis in a three-dimensional Hilbert space gives rise to three points separated by the distance $\pi/2$ from each other in $\mathbb{C}\mathbf{P}^2$. By an appropriate choice of coordinates we can make any

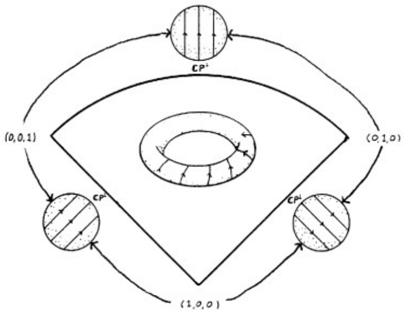

Figure 4.11. Killing vector flows on $\mathbb{C}\mathbf{P}^2$. Stereographic coordinates are used for the octant.

such triplet of points sit at the corners of an octant in a picture identical to the one shown in Figure 4.11.

To get used to the picture we begin by looking at some one real dimensional curves in it. We choose to look at the flow lines of an isometry – a Killing field. Since the isometries are given by unitary transformations of $\mathbb{C}^N$, we obtain an infinitesimal isometry by choosing an Hermitian matrix. Any such matrix can be diagonalized, so we can choose a basis in which the given Hermitian 3×3 matrix takes the form

$$H^{\alpha}{}_{\beta} = \begin{bmatrix} E_0 & 0 & 0 \\ 0 & E_1 & 0 \\ 0 & 0 & E_2 \end{bmatrix} . \tag{4.77}$$

We can therefore arrange our octant picture so that the fixed points of the flow – determined by the eigenvectors – occur at the corners. If we exponentiate Eq. (4.58) we find that the isometry becomes

$$\begin{bmatrix} n^0 \\ n^1 \mathrm{e}^{\mathrm{i}\nu_1} \\ n^2 \mathrm{e}^{\mathrm{i}\nu_2} \end{bmatrix} \quad \rightarrow \quad \begin{bmatrix} \mathrm{e}^{-\mathrm{i}E_0 t} & 0 & 0 \\ 0 & \mathrm{e}^{-\mathrm{i}E_1 t} & 0 \\ 0 & 0 & \mathrm{e}^{-\mathrm{i}E_2 t} \end{bmatrix} \begin{bmatrix} n^0 \\ n^1 \mathrm{e}^{\mathrm{i}\nu_1} \\ n^2 \mathrm{e}^{\mathrm{i}\nu_2} \end{bmatrix} \tag{4.78}$$

where t is a parameter along the flow lines. Taking out an overall phase, we find that this implies that

$$n^I \rightarrow n^I , \quad \nu_1 \rightarrow \nu_1 + (E_0 - E_1)t , \quad \nu_2 \rightarrow \nu_2 + (E_0 - E_2)t . \tag{4.79}$$

Hence the position on the octant is preserved by the Killing vector flow; the movement occurs on the tori only. At the edges of the octant (which are spheres, determined by the location of the fixed points) the picture is the expected one. At a generic point the orbits wind around the tori and, unless the frequencies are rational multiples of each other, they will eventually cover the tori densely. Closed orbits are exceptional.

We go on to consider some submanifolds. Every pair of points in the complex projective plane defines a unique complex projective line, that is a $\mathbb{C}\mathbf{P}^1$,

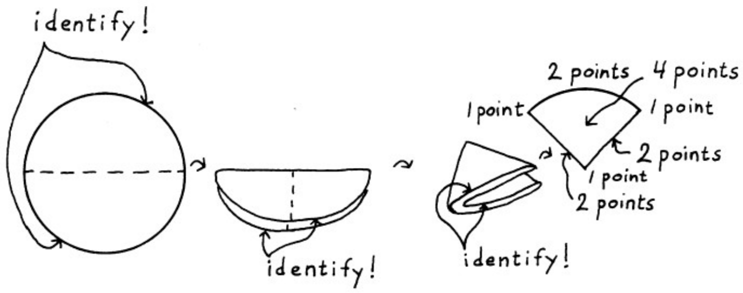

Figure 4.12. Using stereographic coordinates we show how the octant picture of the real submanifold $\mathbb{R}\mathbf{P}^2$ is related to the standard description as a hemisphere with antipodal points on the equator identified.

containing the pair of points. Conversely a pair of complex projective lines always intersect at a unique point. Through every point there passes a 2-sphere's worth of complex projective lines, conveniently parametrized by the way they intersect the line at infinity, that is the set of points at maximal distance from the given point. This is easily illustrated provided we arrange the picture so that the given point sits in a corner, with the line at infinity represented by the opposite edge. An interesting fact about $\mathbb{C}\mathbf{P}^2$ follows from this, namely that it contains *incontractible* 2-spheres – a complex projective line can be deformed so that its radius grows, but it cannot be deformed so that its radius shrinks because it has to intersect the line at infinity in a point. (The topological reasons for this can be seen in the picture.)

Another submanifold is the real projective plane $\mathbb{R}\mathbf{P}^2$. It is defined in a way analogous to the definition of $\mathbb{C}\mathbf{P}^2$ except that real rather than complex numbers are used. The points of $\mathbb{R}\mathbf{P}^2$ are therefore in one-to-one correspondence with the set of lines through the origin in a three-dimensional real vector space and also with the points of $\mathbf{S}^2/\mathbb{Z}_2$, that is to say the sphere with antipodal points identified. In its turn this is a hemisphere with antipodal points on the equator identified. $\mathbb{R}\mathbf{P}^2$ is clearly a subset of $\mathbb{C}\mathbf{P}^2$. It is illuminating to see how the octant picture is obtained, starting from the stereographic projection of a hemisphere (a unit disc) and folding it twice.

Next choose a point. Adjust the picture so that it sits at a corner of the octant. Surround the chosen point with a 3-sphere consisting of points at constant distance from that point. In the picture this will appear as a curve in the octant, with an entire torus sitting over each interior point on the curve. This makes sense: from the Hopf fibration we know that $\mathbf{S}^3$ can be thought of as a one parameter family of tori with circles at the ends. The 3-sphere is round if the tori have a suitable rectangular shape. But as we let our 3-sphere grow, its tori get more and more 'squashed' by the curvature of $\mathbb{C}\mathbf{P}^2$, and the roundness gradually disappears. When the radius reaches its maximum value of $\pi/2$ the 3-sphere has collapsed to a 2-sphere, namely to the projective line

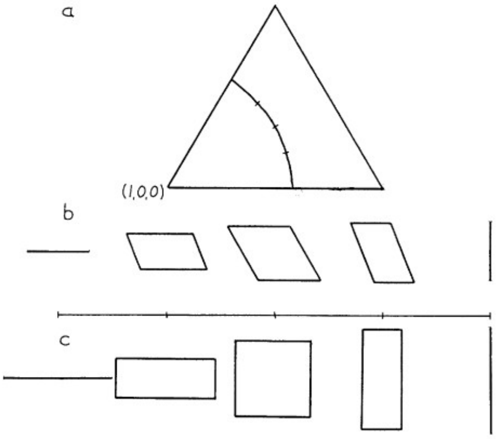

Figure 4.13. The set of points at constant distance from a corner form a squashed 3-sphere. In (a) we show how such a submanifold appears in the octant. All points in the torus lying over a point on the curve are included. In (b) we show how the size and shape of the torus change as we move along the curve in the octant; at the ends of the interval the tori collapse to circles. For comparison, in (c) we show the corresponding picture for a round 3-sphere.

at infinity. In equations, set

$$n_1 = \sin r \cos \frac{\theta}{2} , \quad n_2 = \sin r \sin \frac{\theta}{2} , \quad \tau = \nu_1 + \nu_2 , \quad \phi = \nu_1 - \nu_2 , \quad (4.80)$$

where $0 \leq r \leq \pi/2$, $0 \leq \theta \leq \pi$. When we express the Fubini–Study metric in these coordinates we find that they are geodesic polar coordinates with the coordinate r measuring the distance from the origin – curves with affine parameter equal to r are manifestly geodesics. Indeed

$$\mathrm{d}s^2 \;=\; \mathrm{d}r^2 + \sin^2 r \, (\Theta_1^2 + \Theta_2^2 + \cos^2 r \, \Theta_3^2) \; , \quad (4.81)$$

where the Θ_I are the invariant forms introduced in Section 3.7; our squashed spheres at constant distance are Berger spheres, as defined in the same section.

Finally, a warning: the octant picture distorts distances in many ways. For instance, the distance between two points in a given torus is shorter than it looks, because the shortest path between them is not the same as a straight line within the torus itself. Note also that we have chosen to exaggerate the size of the octant relative to that of the tori in the pictures. To realize how much room there is in the largest torus we note that one can find four sets of orthonormal bases in the Hilbert space $\mathbb{C}^3$ such that the absolute value of the scalar product between members of different bases is $1/\sqrt{3}$ – a kind of sphere packing problem in $\mathbb{C}\mathbf{P}^2$. If one basis is represented by the corners of the octant, the remaining 3×3 basis vectors are situated in the torus over the centre of the octant,[5] as illustrated in Figure 4.10. The message is that the biggest torus is really big.

4.7 Symplectic geometry and the Fubini–Study measure

So far the symplectic form has not been illustrated, but the octant coordinates are very useful for this purpose too. If we use $n_1, n_2, \ldots, n_n$ as orthographic coordinates on the octant we find that the symplectic form (4.50) is

$$\Omega \;=\; 2(n_1 \mathrm{d}n_1 \wedge \mathrm{d}\nu_1 + \cdots + n_n \mathrm{d}n_n \wedge \mathrm{d}\nu_n) \; . \tag{4.82}$$

Even better, we can use $y_i = n_i^2$ as coordinates. Then Ω takes the canonical form

$$\Omega = \mathrm{d}y_1 \wedge \mathrm{d}\nu_1 + \cdots + \mathrm{d}y_n \wedge \mathrm{d}\nu_n \; . \tag{4.83}$$

In effect (y_i, ν_i) are action-angle variables.

Given a symplectic form we can construct a phase space volume by wedging it with itself enough times. This is actually simpler than to compute the determinant of the metric tensor, and the two notions of volume – symplectic and metric – agree because we are on a Kähler manifold. Thus, in form language, the *Fubini–Study volume element* on $\mathbb{C}\mathbf{P}^n$ is

$$\mathrm{d}\tilde{\Omega}_n \;=\; \frac{1}{n!}\,(\frac{1}{2}\Omega) \wedge (\frac{1}{2}\Omega) \wedge \cdots \wedge (\frac{1}{2}\Omega) \; , \tag{4.84}$$

where we take n wedge products. Equivalently we compute the square root of the determinant of the metric. We have decorated the volume element with a tilde because in Section 7.6 we will divide it with the total volume of $\mathbb{C}\mathbf{P}^n$ to obtain the *Fubini–Study measure* $\mathrm{d}\Omega_n$, which is a probability distribution whose integral over $\mathbb{C}\mathbf{P}^n$ equals unity. Any measure related to the Fubini–Study volume element by a constant is distinguished by the fact that it gives the same volume to every ball of given radius, as measured by the Fubini–Study metric, and by the fact that it is unitarily invariant, in the sense that $\mathrm{vol}(A) = \mathrm{vol}\big(U(A)\big)$ for any subset A of $\mathbb{C}\mathbf{P}^n$.

It is a volume element worth studying in several coordinate systems. Using octant coordinates we obtain

$$\mathrm{d}\tilde{\Omega}_n \;=\; n_0 n_1 \ldots n_n \, \mathrm{d}V_{\mathbf{S}^n} \, \mathrm{d}\nu_1 \ldots \mathrm{d}\nu_n \; , \qquad 0 < \nu_i < 2\pi \tag{4.85}$$

where $\mathrm{d}V_{\mathbf{S}^n}$ is the measure on the round n-sphere. If we use orthographic coordinates on the round octant, remember that $\sqrt{g} = 1/n_0$ in these coordinates, and use the coordinates $y_i = n_i^2$, we obtain

$$\mathrm{d}\tilde{\Omega}_n = n_1 \ldots n_n \, \mathrm{d}n_1 \ldots \mathrm{d}n_n \, \mathrm{d}\nu_1 \ldots \mathrm{d}\nu_n = \frac{1}{2^n} \, \mathrm{d}y_1 \ldots \mathrm{d}y_n \, \mathrm{d}\nu_1 \ldots \mathrm{d}\nu_n \; . \tag{4.86}$$

---

[5] It is not known to what extent this property generalizes to arbitrary Hilbert space dimension. Ivanović (1981) and Wootters and Fields (1989) have shown that $N + 1$ orthonormal bases with the modulus of the scalar product between members of different bases always equal to $1/\sqrt{N}$ can be found if the dimension $N = p^k$, where p is a prime number.

All factors in the measure have cancelled out! As far as calculations of volumes are concerned $\mathbb{C}\mathbf{P}^n$ behaves like a Cartesian product of a flat simplex and a flat torus of a fixed size.

The angular coordinates ϑ can also be used on the octant. We just combine Eqs. (4.85) with Eq. (1.14) for the measure on the round octant. Computing the total volume of $\mathbb{C}\mathbf{P}^n$ with respect to its Fubini–Study volume element then leads to an easy integral:

$$\mathrm{vol}(\mathbb{C}\mathbf{P}^n) = \int_{\mathbb{C}\mathbf{P}^n} \mathrm{d}\tilde{\Omega}_n = \prod_{i=1}^{n} \int_0^{\frac{\pi}{2}} \mathrm{d}\vartheta_i \int_0^{2\pi} \mathrm{d}\nu_i \, \cos\vartheta_i \, \sin^{2i-1}\vartheta_i = \frac{\pi^n}{n!} \ . \quad (4.87)$$

We fixed the linear scale of the space by the requirement that a closed geodesic has circumference π. Then the volume goes to zero when n goes to infinity. Asymptotically the volume of $\mathbb{C}\mathbf{P}^n$ even goes to zero somewhat faster than that of $\mathbf{S}^{2n}$, but the comparison is not fair unless we rescale the sphere until the great circles have circumference π. Having done so we find that the volume of $\mathbb{C}\mathbf{P}^n$ is always larger than that of the sphere (except when $n = 1$, when they coincide).

It is curious to observe that

$$\mathrm{vol}(\mathbb{C}\mathbf{P}^n) = \mathrm{vol}(\Delta_n) \times \mathrm{vol}(\mathbf{T}^n) = \frac{\mathrm{vol}(\mathbf{S}^{2n+1})}{\mathrm{vol}(\mathbf{S}^1)} \ . \quad (4.88)$$

The first equality expresses the volume as the product of the volumes of a flat simplex and a flat torus and is quite surprising, while the second is the volume of $\mathbf{S}^{2n+1}$, given in Eq. (1.17), divided by the volume 2π of a Hopf circle as one would perhaps expect from the Hopf fibration. The lesson is that the volume does not feel any of the more subtle topological properties involved in the fibre bundle construction of $\mathbb{C}\mathbf{P}^n$.

4.8 Fibre bundle aspects

Although we did not stress it so far, it is clear that $\mathbb{C}\mathbf{P}^n$ is the base manifold of a bundle whose total space is $\mathbf{S}^{2n+1}$ or even $\mathbb{C}^N = \mathbb{C}^{n+1}$ with its origin deleted. The latter bundle is known as the *tautological bundle* for fairly obvious reasons. A lightning review of our discussion of fibre bundles in Chapter 3 consists, in effect, of the observation that our expression for the Fubini–Study metric is invariant under

$$\mathrm{d}P^\alpha \ \rightarrow \ \mathrm{d}P^\alpha + zP^\alpha \ , \quad (4.89)$$

where $z = x + \mathrm{i}y$ is some complex number. But the vector xP^α is orthogonal to the sphere, while $\mathrm{i}yP^\alpha$ points along its Hopf fibres. Because it is unaffected by changes in these directions the Fubini–Study metric really is a metric on the space of Hopf fibres. In some ways it is convenient to normalize our vectors to $Z \cdot \bar{Z} = 1$ here because then we are dealing with a principal bundle. The tautological bundle is not a principal bundle; technically it is called a *Hermitian line bundle*.

A fibre bundle can always be equipped with a connection that allows us to lift curves in the base manifold to the bundle space in a unique manner. Here the preferred choice is

$$\omega \;=\; \frac{\mathrm{i}}{2}\,\frac{Z\cdot \mathrm{d}\bar{Z} - \mathrm{d}Z\cdot\bar{Z}}{Z\cdot\bar{Z}} \;=\; -\mathrm{i}\,\mathrm{d}Z\cdot\bar{Z}\ . \tag{4.90}$$

In the last step we used normalized vectors. The equation $\omega = 0$ now expresses the requirement that $\mathrm{d}Z^\alpha$ be orthogonal to Z^α, so that the lifted curve will be perpendicular to the fibres of the bundle. A minor calculation confirms that

$$\mathrm{d}\omega \;=\; \Omega\ , \tag{4.91}$$

where Ω is the symplectic 2-form defined in Eq. (4.50).

Let us lift some curves on $\mathbb{C}\mathbf{P}^n$ to Hilbert space. We have already done so for a geodesic on the base manifold – the result is given in Eq. (4.44). Since $\dot{Z}\cdot\bar{Z} = 0$ along this curve, it follows that $\omega = 0$ along it, so indeed it is a horizontal lift. Another example that we have encountered already: in Eq. (4.58) we wrote the Schrödinger equation, or equivalently the flow line of some isometry, as

$$\mathrm{i}\,\mathrm{d}Z^\alpha = H^\alpha{}_\beta Z^\beta \mathrm{d}t\ . \tag{4.92}$$

Evidently

$$\omega = \frac{1}{2}(Z^\alpha \bar{H}_\alpha{}^\beta \bar{Z}_\beta - \bar{Z}_\alpha H^\alpha{}_\beta Z^\beta)\ \mathrm{d}t = 0 \tag{4.93}$$

along the resulting curve in the bundle, so that this is the horizontal lift of a curve defined in the base manifold by the projective Schrödinger equation (4.59).

Now consider an arbitrary curve $C(\sigma)$ in $\mathbb{C}\mathbf{P}^n$, and an arbitrary lift of it to $\mathbb{C}^N$. Let the curve run from σ_1 to σ_2. Define its *geometric phase* as

$$\phi_{\mathrm{g}} = \arg\big(Z(\sigma_2)\cdot\bar{Z}(\sigma_1)\big) - \int_C \omega\ . \tag{4.94}$$

From now on we assume that the length of the curve is less than $\pi/2$ so that the argument that defines the *total phase* – the first term on the right-hand side – of the curve is well defined. Suppose that we change the lift of the curve by means of the transformation

$$Z^\alpha \to \mathrm{e}^{\mathrm{i}\lambda}Z^\alpha \qquad \Rightarrow \qquad \omega \to \omega + \mathrm{d}\lambda\ , \tag{4.95}$$

where λ is some function of σ. Although both the connection and the total phase change under this transformation, the geometric phase does not. Indeed it is easy to see that

$$\phi_{\mathrm{g}} \to \phi_{\mathrm{g}} + \lambda(\sigma_2) - \lambda(\sigma_1) - \int_{\sigma_1}^{\sigma_2}\mathrm{d}\lambda = \phi_{\mathrm{g}}\ . \tag{4.96}$$

The conclusion is that the geometric phase does not depend on the particular lift that we use. Hence it is genuinely a function of the original curve in

$\mathbb{C}\mathbf{P}^n$ itself. The total phase on the other hand is not – although it equals the geometric phase for a horizontal lift of the curve.

An interesting observation, with consequences, is that if we compute the total phase along the horizontal lift (4.44) of a geodesic in $\mathbb{C}\mathbf{P}^n$ we find that it vanishes. This means that the geometrical phase is zero for any lift of a geodesic. A geodesic is therefore a *null phase curve*. The converse is not true – we will find further examples of null phase curves in Chapter 6. To see what the consequences are, consider a *geodesic triangle* in $\mathbb{C}\mathbf{P}^n$, that is three points P, Q and R connected with geodesic arcs not longer than $\pi/2$. In flat space the lengths of the sides determine the triangle uniquely up to isometries, but in curved spaces this need not be so. As it happens a geodesic triangle in $\mathbb{C}\mathbf{P}^n$ is determined up to isometries by its side lengths and its *symplectic area*, that is the integral of the symplectic form Ω over any area bounded by the three geodesic arcs (Brehm, 1990). To such a triangle we can associate the *Bargmann invariant*[6]

$$\Delta_3(P,Q,R) = \frac{P\cdot\bar{Q}Q\cdot\bar{R}R\cdot\bar{P}}{P\cdot\bar{P}Q\cdot\bar{Q}R\cdot\bar{R}} = \cos D_{PQ}\cos D_{QR}\cos D_{RP}\mathrm{e}^{-\mathrm{i}\Phi} \ . \tag{4.97}$$

Here D_{PQ} is the Fubini–Study distance between the points whose homogeneous coordinates are P and Q, and so on. But what is the phase Φ? Since each side of the triangle is a geodesic arc the geometric phase of each side is zero, and by Eq. (4.94) the total phase of the side is equal to the integral along the arc of ω. Adding the contribution of all the sides (and normalizing the vectors) we get

$$\arg P\cdot\bar{Q}Q\cdot\bar{R}R\cdot\bar{P} = -\arg Q\cdot\bar{P} - \arg R\cdot\bar{Q} - \arg P\cdot\bar{R} = -\oint_{\partial\Delta}\omega \ . \tag{4.98}$$

It follows from Stokes' theorem that the phase Φ is the symplectic area of the triangle,

$$\Phi = \int_{\Delta}\Omega \ . \tag{4.99}$$

We can go on to define Bargmann invariants for four, five and more points in the same cyclic manner and – by triangulation – we find that their phases are given by the symplectic area of the geodesic polygon spanned by the points.

The geometric phase also goes under the name of the *Berry phase*; the case that Berry studied was that of an eigenstate of a Hamiltonian H that is carried around a loop by adiabatic cycling of H.[7]

---

[6] It is invariant under phase changes of the homogenous coordinates and attracted Bargmann's attention (Bargmann, 1964) because, unlike the Fubini–Study distance, it is a complex valued function defined on $\mathbb{C}\mathbf{P}^n$. Blaschke and Terheggen (1939) considered this early on. That its argument is the area of a triangle was proved by both physicists (Mukunda and Simon, 1993) and mathematicians (Hangan and Masala, 1994).

[7] The original paper by Berry (1984) is reprinted in a book edited by Shapere and Wilczek (1989). See Anandan and Aharonov (1990) and Mukunda and Simon (1993) for the point of view that we take here, and Chruściński and JamioŁkowski (2004) for much more.

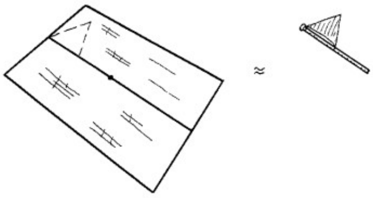

Figure 4.14. A flag may consist of a one-dimensional subspace of a two-dimensional subspace of a three-dimensional space – it is clear why it is called that!

4.9 Grassmannians and flag manifolds

Projective space is only the first of its kind. Starting from an N-dimensional vector space V, real or complex, we can consider nested sequences of subspaces V_i of dimension d_i, such that $V_1 \subset V_2 \subset \cdots \subset V_r$. This is called a *flag* of subspaces (as explained in Figure 4.14). The space of all flags of a given kind is known as a *flag manifold* and denoted by $\mathbf{F}^{(N)}_{d_1 \dots d_r}$; projective space is the easy case where a flag consists of a single one-dimensional subspace only. The next simplest case is that where the flags consist of a single M-dimensional subspace; such a flag manifold is known as the *Grassmannian* $\mathbf{F}^{(N)}_{M} = \mathbf{Gr}^{\mathbb{C}}_{M,N}$. The notation also tells us which field ($\mathbb{R}$ or $\mathbb{C}$) we are using; if there is no label then the complex numbers are implied.

To see how this works let us consider the case of $\mathbf{Gr}_{2,4}$, which is the space of 2-planes in a four-dimensional vector space. We fix a 2-plane in $\mathbb{C}^4$ by fixing two linearly independent vectors spanning that plane, and collect them into a rank two $N \times 2$ matrix

$$\begin{bmatrix} Z_{0,0} & Z_{0,1} \\ Z_{1,0} & Z_{1,1} \\ Z_{2,0} & Z_{2,1} \\ Z_{3,0} & Z_{3,1} \end{bmatrix} \sim \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ z_{2,0} & z_{2,1} \\ z_{3,0} & z_{3,1} \end{bmatrix} . \tag{4.100}$$

The entries in the matrix on the left are homogeneous coordinates on the Grassmannian. Then we exercised our right to perform linear combinations and rescalings of the columns to get the matrix in a standard form. The remaining four complex numbers serve as affine coordinates on the four complex dimensional Grassmannian. Note that if the upper two rows were linearly dependent this form could not have been reached, but since the matrix as a whole has rank two we can introduce a similar coordinate system by singling out another pair of rows. (Indeed it is not hard to see that the Grassmannian is a complex manifold that can be completely covered by six coordinate patches of this kind.) From a geometrical point of view it is sometimes advantageous to think of the Grassmannian of 2-planes as the space of projective lines in a projective space of dimension $N - 1$; it is an interesting exercise to convince oneself directly that the space of lines in 3-space is four-dimensional. On the other hand the use of an $N \times M$ matrix as homogeneous coordinates for $\mathbf{Gr}_{M,N}$ has advantages too and leads to an immediate proof that it is a complex manifold of $M(N-M)$ complex dimensions.[8]

When we get to more involved examples of flag manifolds any kind of explicit description of the space will get involved, too. For this reason it is a good idea to fall back on a description as coset spaces. Let us begin with $\mathbf{F}_1^{(4)} = \mathbb{C}\mathbf{P}^3$. The group $GL(4, \mathbb{C})$ acts transitively on the underlying vector space. Pick any point, say $p = (1, 0, 0, 0)$ and find its isotropy group – that subgroup of the transitive group that leaves the given point invariant. Here the isotropy group consists of all matrices h of the form

$$h = \begin{bmatrix} \bullet & \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet & \bullet \end{bmatrix} , \tag{4.101}$$

where $\bullet$ is any complex number. This defines a subgroup of $GL(4, \mathbb{C})$, and we can now define $\mathbb{C}\mathbf{P}^3$ as a coset space following the recipe in Section 3.8. We can use $SL(4, \mathbb{C})$ as our starting point instead; the resulting coset will be the same once the form of h is restricted so that it has unit determinant. Call the resulting isotropy group $P_1^{(4)}$. Indeed we can restrict ourselves to $U(N)$ since the unitary subgroup of $GL(4, \mathbb{C})$ also acts transitively; when h is restricted to belong to $U(N)$ it becomes a block diagonal matrix representing $U(1) \times U(3)$. The argument clearly generalizes to any dimension, so that we have proved that

$$\mathbf{F}_1^{(N)} \equiv \mathbb{C}\mathbf{P}^{N-1} = \frac{SL(N, \mathbb{C})}{P_1^{(N)}} = \frac{U(N)}{U(1) \times U(N-1)} \ . \tag{4.102}$$

For real projective space we just have to replace $SL(\mathbb{C})$ with $SL(\mathbb{R})$ and the unitary groups by orthogonal groups.

The same argument repeated for the Grassmannian $\mathbf{Gr}_{2,4}$ reveals that the isotropy group can be written as the set $P_2^{(4)}$ of special linear matrices of the form

$$h = \begin{bmatrix} \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ 0 & 0 & \bullet & \bullet \\ 0 & 0 & \bullet & \bullet \end{bmatrix} , \tag{4.103}$$

or as block diagonal unitary matrices belonging to $U(2) \times U(2)$ if only unitary transformations are considered. For the flag manifold $\mathbf{F}_{1,2,3}^{(N)}$ the isotropy group is a *Borel subgroup* B of $SL(4, \mathbb{C})$; by definition the *standard Borel subgroup* is the group of upper triangular matrices and a general Borel subgroup is given by conjugation of the standard one. If the group is $U(N)$ the Borel subgroup is the subgroup $U(1) \times U(1) \times U(1) \times U(1)$, given by diagonal matrices in the standard case. The isotropy groups considered above are examples of *parabolic*

---

[8] The affine coordinates are particularly useful if one wants to discuss the natural (Fubini–Study like) metric properties of the Grassmannians. See Wong (1967) for this.

subgroups P of the group G, represented by 'block upper triangular' matrices such that $B \subset P \subset G$. Thus we can write, in complete generality,

$$\mathbf{F}^{(N)}_{d_1 d_2 \dots d_r} = \frac{SL(N, \mathbb{C})}{P^{(N)}_{d_1 d_2 \dots d_r}} = \frac{U(N)}{[U(k_1) \times U(k_2) \dots U(k_{r+1})]} \ , \tag{4.104}$$

where $k_1 = d_1$ and $k_{i+1} = d_{i+1} - d_i$ (and $d_{r+1} = N$). Given that the real dimension of $U(N)$ is N^2 the real dimension of an arbitrary flag manifold is

$$\dim(\mathbf{F}^{(N)}_{d_1 d_2 \dots d_r}) = N^2 - \sum_{i=1}^{r+1} k_i^2 \ . \tag{4.105}$$

Equivalently it is simply twice the number of zeroes in the complex matrices representing the parabolic subgroup $P^{(N)}_{d_1 d_2 \dots d_r}$. For real flag manifolds, replace the unitary groups with orthogonal groups.

The two descriptions of flag manifolds given in Eq. (4.104) are useful in different ways. To see why we quote two facts: the quotient of two complex groups is a complex manifold, and the quotient of two compact groups is a compact manifold. Then the conclusion is that all flag manifolds are compact complex manifolds. Indeed they are also Kähler manifolds. This has to do with another way of arriving at them, based on *adjoint orbits* in a Lie algebra. This aspect of flag manifolds will be discussed at length in Chapter 8 but let us reveal the point already here. The Lie algebra of the unitary group is the set of Hermitian matrices, and such matrices can always be diagonalized by unitary transformations. The set of Hermitian matrices with a given spectrum $(\lambda_1, \lambda_2, \dots, \lambda_N)$ can therefore be written as

$$H_{\lambda_1 \lambda_2 \dots \lambda_N} = U H_{\text{diag}} U^{-1} \ . \tag{4.106}$$

This is an adjoint orbit and a little bit of thinking reveals that it is also a flag manifold $\mathbf{F}^{(N)}$; which particular one depends on whether there are degeneracies in the spectrum. For the special case when $\lambda_1 = 1$ and all the others are zero we came across this fact when we embedded $\mathbb{C}\mathbf{P}^n$ in the flat vector space of Hermitian matrices; see Eq. (4.61). The theme will be further pursued in section 8.5.[9]

A variation on the theme deserves mention too. A *Stiefel manifold* is by definition the space of sets of M orthonormal vectors in an N-dimensional vector space. It is not hard to see that these are the complex (real) homogeneous spaces

$$\mathcal{S}t^{(\mathbb{C})}_{M,N} = \frac{U(N)}{U(N-M)} \quad \text{and} \quad \mathcal{S}t^{(\mathbb{R})}_{M,N} = \frac{O(N)}{O(N-M)}. \tag{4.107}$$

This is not quite the same as a Grassmannian since the definition of the latter is insensitive to the choice of basis in the M-dimensional subspaces that are the points of the Grassmannian. As special cases we get $\mathcal{S}t^{(\mathbb{C})}_{1,N} = \mathbf{S}^{2N-1}$ and $\mathcal{S}t^{(\mathbb{R})}_{1,N} = \mathbf{S}^{N-1}$; see Section 3.8.

---

[9] For further information on flag manifolds, in a form accessible to physicists, consult Picken (1990). The adjoint orbits should really be coadjoint, but the distinction is irrelevant here.

Problems

◇ **Problem 4.1** Consider $n+2$ ordered points on the plane, not all of which coincide. Consider two such sets equivalent if they can be transformed into each other by translations, rotations and scalings. Show that the topology of the resulting set is that of $\mathbb{C}\mathbf{P}^n$. What does $\mathbb{C}\mathbf{P}^n$ have to do with archaeology?

◇ **Problem 4.2** If you manage to glue together a Möbius strip and a hemisphere you get $\mathbb{R}\mathbf{P}^2$. What will you obtain if you glue two Möbius strips together?

◇ **Problem 4.3** Carry through the argument needed to prove that a complex projective line cannot be shrunk to a point within $\mathbb{C}\mathbf{P}^n$, using formulae, and using the octant picture.

1 +,-.+ + +/ ! +) 2 3 . ! ! ,
(# . ! 9 ! ! 1 # 5 6 3 0#
* % " 0 7 8 6

* %

data

Data - Novo Livro - 018 - 2021/
Books - 018 - 2021/
Importantíssimo/

- 1887187308.epub
  → Art of the Golden Ratio
  ISBN 978-1-887187-20-6 = Hallo.

- Geometry of quantum states ....(2006).pdf (429 pág) (a)
- Geometry of Quantum States....(2017).pdf (636 pág) (b)

(a) 94 (106) 4 Complex Projective Spaces.

● (b) 103 (120) 4 Complex Projective Spaces.

"In the house of mathematics there are many mansions and these the most elegant is projective geometry." Morris Kline.

→ ver pág seguinte. ✡

Andrew Wiles usou a ideia de Kline para explicar o seu modo de entender a matemática.

Morris Kline - Mathematics in Western Culture
(George Allen & Unwin)
1954

* %

- Quantum Mechanics 1-3...pdf (2425 pages) Claude Cohen-Tannoudji.
- Quantum Mechanics, volume 3... pdf (788 pages)

/Importantíssimo/Morris-Kline/

- Morris Kline - Mathematical thought from Ancient to Modern Times

1972 - 3 volumes - .djvu.

- Morris Kline - Mathematics in Western Culture

✡ 1964 - 513 pages.

- Morris Kline - Mathematics for the nonmathematician (1985). pdf
- Morris Kline - Calculus... 1998. pdf

Dream of the Golden Ratio and i

$$\left(\frac{\sqrt{2+\varphi}}{2}+\frac{i}{2\varphi}\right)^5 = i$$

/diegorattaggi

% * %
! +,-.+ + +/ ! + % . ! ! % :!
0 ; < ; ! # ,00 ! , ! ; 0 ; ! # ,00 7 ! = ;
: ; ; ! ; ! # 9 ! 0#

1 +,-.+ + +/ ! + % . ! ! ; %
= ; : ; ; ! ; ! # 9 ! % > 3 ?) 0#

+,-.+ + +/ ! + % @ ! ! !
< , # < /7 # 5 6 3) #86

+,-.+ + +/ ! + % @ ! ! !
< , # < /7 # 5 6 3 ; 5 ,) #86

% +,-.+ + +/ ! + % @ ! ! !
< , # < % /7 # 5 6 3 ; 5 ,) #86

" +,-.+ + +/ ! + % @ ! ! !
< , # < 6 A A% /7 # 5 6 3) 0#

+,-.+ + +/ ! + % @ ! !
! ! ! B 6 1 !) & 0#

+,-.+ + +/ ! + % @ ! , (6
! # 3 ! ,00 ! B 6 1 !)) 0#

*   %

- Quantum Mechanics 1-3... pdf (2425 pages) Claude Cohen-Tannoudji.
- Quantum Mechanics, volume 3... pdf (788 pages)

/Importantissimo/ Morris-Kline/

- Morris Kline - Mathematical Thought from Ancient to Modern Times -
  1972 - 3 volumes - .djvu.
- Morris Kline - Mathematics in Western Culture
  ✡ 1964 - 513 pages.
- Morris Kline - Mathematics for the nonmathematician (1985). pdf
- Morris Kline - Calculus... 1998. pdf

Dream of the Golden Ratio and i

$$\left(\frac{\sqrt{2+\varphi}}{2}+\frac{i}{2\varphi}\right)^{5}=i$$

/diegorattaggi

Corredor de Suwalki,
a fronteira de 65 km entre
Lituânia e Polônia que separa
a Belarus do território russo
ocidental de Kaliningrado no Báltico.

—

The Disc Embedding Theorem
Stefan Behrens
www.math.uni-bielefeld.de/
~sbehrens/

Source:
www.quantamagazine.org/
new-math-book-rescues-landmark-
topology-proof-20210909/

—

www.quantamagazine.org/
triangulation-conjecture-
disproved-20150113/

—

www.quantamagazine.org

—

% * "
! +,-.+ + +/ ! + % < B 9 1 ## <
/7 # 5 6 3 0#

1 +,-.+ + +/ ! + % 2 2 ! B ! !
> CD C! , # E! # < C 2 ! 9# 0 0 C
0#
* %% "& ,11 % "

3.3 Sinus, Kosinus, Tangens

Die Winkelfunktionen Sinus, Kosinus und Tangens spielen in der Mathematik eine tragende Rolle. Sie sind die Basis aller Berechnungen, die auf Winkelmessung beruhen.

Betrachten wir den Einheitskreis: Ein Punkt P darauf habe in einem kartesischen Koordinatensystem die Koordinaten (x, y). Der zugehörige *Polarwinkel* sei φ (Abb. 3.22). Dann wird definiert:

$$x = \cos\varphi, \quad y = \sin\varphi, \quad \tan\varphi = \frac{y}{x} = \frac{\sin\varphi}{\cos\varphi} \tag{3.6}$$

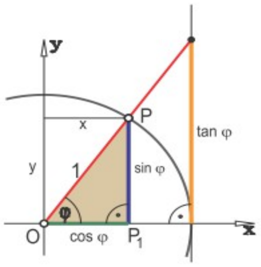

Abb. 3.22 Sinus, Kosinus, Tangens

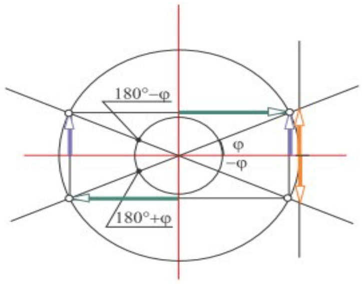

Abb. 3.23 $\sin(\pm\varphi)$, $\sin(180° \pm \varphi)$

Es folgt nach Pythagoras die wichtige Beziehung

$$\sin^2\varphi + \cos^2\varphi = 1 \tag{3.7}$$

Die Proportion $\tan\varphi : 1 = \sin\varphi : \cos\varphi$ zeigt, dass sich in Abb. 3.22 auch der Tangens auf der zur y-Achse parallelen (vertikalen) rechten Kreistangente $x = 1$ ablesen lässt.

Seltener wird der *Kotangens* verwendet (weil am Taschenrechner nicht verfügbar). Es handelt sich dabei um den Kehrwert des Tangens:

$$\cot\varphi = \frac{1}{\tan\varphi} = \frac{\cos\varphi}{\sin\varphi}$$

Aus Abb. 3.23 lassen sich folgende Beziehungen erkennen:

$$\begin{aligned} \sin(180° - \varphi) &= \sin\varphi & \qquad \sin(-\varphi) &= -\sin\varphi \\ \cos(180° - \varphi) &= -\cos\varphi & \qquad \cos(-\varphi) &= \cos\varphi \\ \tan(180° - \varphi) &= -\tan\varphi & \qquad \tan(-\varphi) &= -\tan\varphi \end{aligned} \tag{3.8}$$

Wir betrachten nun das rechtwinklige Dreieck OP_1P in Abb. 3.22. Es wird gebildet von:

- der dem Winkel φ anliegenden Kathete („Ankathete“) A mit der Länge x,

- von der φ gegenüberliegenden Kathete („Gegenkathete“) G mit der Länge y und

- der Hypothenuse H mit der Länge 1.

Es gilt:

$$\cos\varphi = \frac{A}{H}, \quad \sin\varphi = \frac{G}{H}, \quad \tan\varphi = \frac{G}{A} \tag{3.9}$$

Das gilt in jedem rechtwinkligen Dreieck mit Schenkelwinkel φ (ähnliche Dreiecke!).

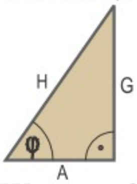

Damit gilt im *rechtwinkligen Dreieck*

$$A = H\cos\varphi, \quad G = H\sin\varphi \tag{3.10}$$

Weiters gilt $\alpha = 90^\circ - \beta$. Durch Vertauschen der Bezeichnungsweisen erhält man die wichtige Beziehung

$$\cos\alpha = \sin(90^\circ - \alpha) \quad \text{bzw.} \quad \sin\alpha = \cos(90^\circ - \alpha) \tag{3.11}$$

Insbesondere gilt wegen $\cos 45^\circ = \sin 45^\circ$ und $\cos^2 45^\circ + \sin^2 45^\circ = 1$

$$\cos 45^\circ = \sin 45^\circ = \frac{1}{\sqrt{2}}. \tag{3.12}$$

Aus dem halben gleichseitigen Dreieck erkennt man sofort

$$\cos 60^\circ = \sin 30^\circ = \frac{1}{2}$$

und damit wegen $\cos^2 60^\circ + \sin^2 60^\circ = 1$

$$\sin 60^\circ = \cos 30^\circ = \frac{\sqrt{3}}{2}$$

Insgesamt haben wir jetzt die folgenden Hauptwerte, die man sich mit der Zeit einprägen sollte:

x	0°	30°	45°	60°	90°
$\cos x$	1	$\sqrt{3}/2$	$\sqrt{2}/2$	$1/2$	0
$\sin x$	0	$1/2$	$\sqrt{2}/2$	$\sqrt{3}/2$	1
$\tan x$	0	$1/\sqrt{3}$	1	$\sqrt{3}$	∞

(3.13)

∘ Als „Eselsbrücke“ kann man sich die Tabelle für den Sinus leicht merken (die Tabelle für den Kosinus ist dann „rückläufig“):

x	0°	30°	45°	60°	90°
$\sin x$	$\sqrt{0}/2$	$\sqrt{1}/2$	$\sqrt{2}/2$	$\sqrt{3}/2$	$\sqrt{4}/2$

Die folgenden *Additionstheoreme* brauchen wir uns nicht auswendig zu merken, wir benötigen sie aber bei den Drehungen in Abschnitt 6.1:

$$\begin{aligned}\sin(\alpha+\beta) &= \sin\alpha\cos\beta+\cos\alpha\sin\beta, \\ \cos(\alpha+\beta) &= \cos\alpha\cos\beta-\sin\alpha\sin\beta\end{aligned} \qquad (3.14)$$

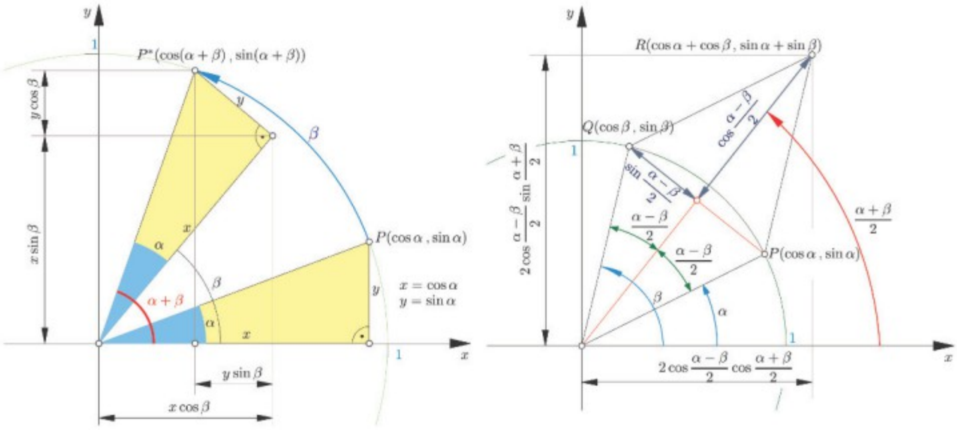

y
1
$P^*(\cos(\alpha+\beta)\,,\sin(\alpha+\beta))$
$y\cos\beta$
y
β
$x\sin\beta$
x
α
$P(\cos\alpha\,,\sin\alpha)$
β
$\alpha+\beta$
α
y
$x=\cos\alpha$
$y=\sin\alpha$
x
1
x
$y\sin\beta$
$x\cos\beta$
y
$R(\cos\alpha+\cos\beta\,,\sin\alpha+\sin\beta)$
$Q(\cos\beta\,,\sin\beta)$
$\cos\frac{\alpha-\beta}{2}$
1
$2\cos\frac{\alpha-\beta}{2}\sin\frac{\alpha+\beta}{2}$
$\sin\frac{\alpha-\beta}{2}$
$\frac{\alpha-\beta}{2}$
$\frac{\alpha-\beta}{2}$
$\frac{\alpha+\beta}{2}$
$P(\cos\alpha\,,\sin\alpha)$
β
α
1
x
$2\cos\frac{\alpha-\beta}{2}\cos\frac{\alpha+\beta}{2}$

Abb. 3.24 Beweise für die trigonometrischen Additionstheoreme

∘ *Beweis*:
Seien $P_\alpha(\cos\alpha/\sin\alpha)$ und $P_\beta(\cos\beta/\sin\beta)$ jene Punkte am Einheitskreis, die zu den Winkeln α und β gehören. Wir verdrehen nun P_β um α und erhalten so $P_{\alpha+\beta}(\cos(\alpha+\beta)/\sin(\alpha+\beta))$. Seine Koordinaten sind aber gemäß Abb. 3.24 links auch

$$P_{\alpha+\beta}(\cos\alpha\cos\beta-\sin\alpha\sin\beta/\sin\alpha\cos\beta+\cos\alpha\sin\beta).$$

■

Anwendung: Man zeige mit den Additionstheoremen

$$\cos 2x = 1-2\sin^2 x \text{ und damit } \sin\frac{x}{2} = \pm\sqrt{\frac{1}{2}\,(1-\cos x)} \qquad (3.15)$$

∘ *Beweis*:

$$\cos 2x = \cos(x+x) = \cos^2 x - \sin^2 x = (1-\sin^2 x) - \sin^2 x = 1-2\sin^2 x$$

$$\Rightarrow \sin x = \sqrt{\frac{1}{2}\,(1-\cos 2x)} \Rightarrow \sin\frac{\alpha}{2} = \sqrt{\frac{1}{2}\,(1-\cos\alpha)}$$

(Auf die Bezeichnungsweise kommt es nicht an.) ■

■

Anwendung: Wie wurden früher Tabellen für die Winkelfunktionen erstellt?

Die Frage ist eher vom historischen Standpunkt aus interessant. Ohne diese

Tabellen konnte man bis vor wenigen Jahrzehnten keine genauen Berechnungen machen! Schon die alten Griechen haben Tabellen für die Winkelfunktionen angelegt, die Araber haben diese übernommen und verfeinert. Auch die Wikinger sollen die geografische Breite auf See mit Tangens-Tabellen ermittelt haben. Wie der Taschenrechner bzw. Computer heute die Werte berechnet, besprechen wir bei der Differentialrechnung im Abschnitt über Reihenentwicklung (Anwendung S. 389).

Lösung:
Aus den Hauptwerten (3.13) lassen sich nach den Additionstheoremen (3.14) die Zwischenwerte

$$\sin 15^\circ = \cos 75^\circ = \sin(60^\circ - 45^\circ) \quad \text{und} \quad \sin 75^\circ = \cos 15^\circ = \sin(90^\circ - 15^\circ)$$

berechnen. Durch Übergang zum halben Winkel (3.15) bekommt man neue Werte, die ihrerseits mittels der Additionstheoreme zu neuen Werten führen, usw. Es war zwar viel Arbeit, aber man konnte beliebig viele Zwischenwerte auf diese Art ermitteln. Wenn man es „noch genauer wissen wollte", *interpolierte* man zwei Nachbarwerte linear (in sehr kleinen Intervallen kann man die Funktionen getrost durch eine Gerade ersetzen).
Die Tangenswerte ergeben sich als Quotient aus Sinus und Kosinus. ■

Gelegentlich braucht man die sog. *zweiten Additionstheoreme*

$$\begin{aligned}\sin\alpha + \sin\beta &= 2\sin\frac{\alpha+\beta}{2}\cos\frac{\alpha-\beta}{2}\\ \cos\alpha + \cos\beta &= 2\cos\frac{\alpha+\beta}{2}\cos\frac{\alpha-\beta}{2}\end{aligned} \tag{3.16}$$

∘ *Beweis*:
Wir betrachten Abb. 3.24 rechts: Die Punkte O, P, R und Q sollen eine Raute mit Seitenlänge 1 bilden. R hat zunächst die Koordinaten $R(\cos\alpha + \cos\beta / \sin\alpha + \sin\beta)$. Die halbe Diagonale OM hat die Länge $\cos\frac{\alpha-\beta}{2}$. Damit lassen sich die Koordinaten von R auch angeben durch

$$R(2\cos\frac{\alpha-\beta}{2}\cos\frac{\alpha+\beta}{2} / 2\cos\frac{\alpha-\beta}{2}\sin\frac{\alpha+\beta}{2})$$

■

Anwendung: Drehstrom oder Dreiphasenstrom (Abb. 3.25)
Bei der Erzeugung von Drehstrom rotiert ein magnetischer Läufer innerhalb dreier Induktionsspulen und induziert dabei drei phasenverschobene Wechselströme. Diese können einzeln oder gemeinsam verwendet werden. Man berechne die Summe aller Spannungen $y = \sin x + \sin(x + \frac{2\pi}{3}) + \sin(x + \frac{4\pi}{3})$ sowie die Spannungsdifferenz $y = \sin x - \sin(x + \frac{2\pi}{3})$ zwischen zwei aufeinander folgenden Phasen (vgl. dazu Anwendung S. 179).

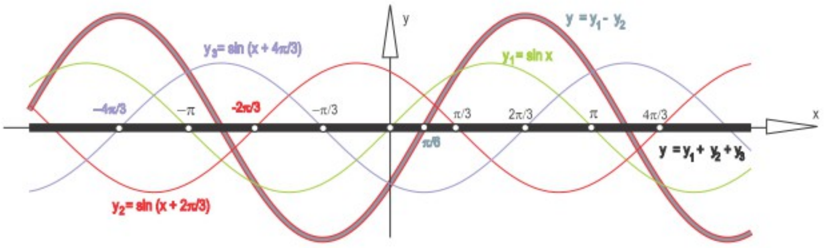

Abb. 3.25 Wellen können einander verstärken oder auslöschen

Lösung:
Wir zeigen: Die Summe zweier phasenverschobener Sinuslinien (Verschiebung x_0, gleiche Amplitude und Frequenz) ist wieder eine Sinuslinie gleicher Frequenz, aber unterschiedlicher Amplitude:
Nach den zweiten Additionstheoremen gilt nämlich

$$\sin x + \sin(x + x_0) = 2\,\sin\tfrac{x+(x+x_0)}{2}\,\cos\tfrac{x-(x+x_0)}{2} = 2\cos\tfrac{-x_0}{2}\,\sin(x+\tfrac{x_0}{2}).$$

Für $x_0 = \frac{2\pi}{3}$ ($\Rightarrow 2\cos\frac{-\pi}{3} = 1$) ergibt sich die Beziehung

$$\sin x + \sin(x + \tfrac{2\pi}{3}) = \sin(x + \tfrac{\pi}{3}).$$

Wegen $\sin x = -\sin(x+\pi) \Rightarrow \sin(x+\frac{\pi}{3}) = -\sin(x+\frac{4\pi}{3})$ ist diese Summe entgegengesetzt gleich der dritten Funktion, und die Gesamtsumme verschwindet:

$$\sin x + \sin(x + \frac{2\pi}{3}) + \sin(x + \frac{4\pi}{3}) = 0 \tag{3.17}$$

Die Spannungsdifferenz zwischen zwei Nachbarphasen ergibt sich durch

$$\sin x - \sin(x + x_0) = \sin x + \sin(-x - x_0) = \\ = 2\,\sin\tfrac{x-(x+x_0)}{2}\,\cos\tfrac{x+(x+x_0)}{2} = 2\sin\tfrac{-x_0}{2}\,\cos(x+\tfrac{x_0}{2}).$$

Für $x_0 = \frac{2\pi}{3}$ haben wir eine Verstärkung der Amplitude mit dem Faktor $2\sin -\frac{\pi}{3} = -\sqrt{3}$. Die Differenzspannung wird beschrieben durch

$$y = -\sqrt{3}\,\cos(x + \frac{\pi}{3}) = -\sqrt{3}\,\sin(\frac{\pi}{2} - (x + \frac{\pi}{3})) = \sqrt{3}\,\sin(x - \frac{\pi}{6})$$

Bei Einzelspannungen von 230 Volt ergibt sich damit eine maximale Spannungsdifferenz von etwa 400 Volt. ■

Anwendung: Schiefe Ebene (Abb. 3.26)
Um ein Fass von 100 kg Masse nicht auf eine Rampe heben zu müssen, rollt man es auf einem Brett, das $\alpha = 30°$ geneigt ist, hinauf. Welche Kraft $F = |\vec{F}|$ ist nötig (Reibung vernachlässigt), um die treibende Komponente $T = |\vec{T}|$ des Gewichts G zu kompensieren?

Lösung:

$$T = F = G\,\sin\alpha$$

$$G = m\,g = 100\text{kg} \cdot 9{,}81\text{m/s}^2 = 981\,\text{N}$$

$$\Rightarrow F = 981\text{N} \cdot \frac{1}{2} = 490{,}5\,\text{N}$$

■

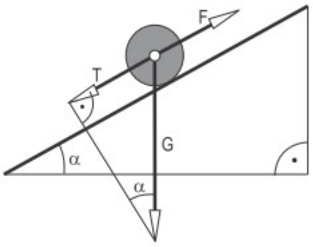

Abb. 3.26 Schiefe Ebene in Theorie...

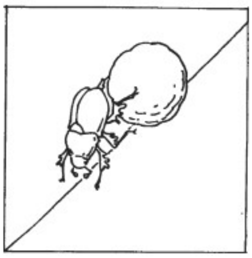

Abb. 3.27 ...und Praxis

Anwendung: Pyramidenrampen

Die alten Ägypter haben sicherlich schiefe Ebenen („Rampen") benutzt, um die schweren Steinblöcke zu heben (auf Schlitten zu ziehen). In Anwendung S. 71 haben wir gesehen, dass weitaus der meiste Teil der Quader in den unteren Schichten verbaut wurde. Für diese niedrigen Schichten bot sich durchaus eine – mit der Höhe des Pyramidenstumpfs wachsende – Rampe an. Rampenmodelle für größere Höhen scheitern allerdings an der praktischen Durchführbarkeit solcher Modelle. Offensichtlich haben die damaligen Baumeister mit wesentlich ausgefeilteren Methoden gearbeitet („Kippschlitten" u. Ä.).

Der Steinbruch der Cheops-Pyramide lag etwa 400 m südlich der Pyramide. Wie hoch konnte man auf dieser Wegstrecke die Kalksteinblöcke bei 8% Rampenneigung anheben? Wie lange hätte eine solche Rampe sein müssen, um auch den allerletzten Baustein der Pyramide, das „Pyramidion", auf die Spitze der Cheops-Pyramide zu transportieren?

Lösung:

Die Steigung der Rampe (Neigungswinkel α) ist $\tan\alpha = 8 : 100$. Bei 400 m Basisstrecke bedeutet dies einen Hub von $400 \cdot \frac{8}{100}\,\mathrm{m} = 32\,\mathrm{m}$. Immerhin ist das gut ein Fünftel der Bauhöhe, und etwa die Hälfte aller Blöcke ist bis in diese Höhe verbaut. Um auch das Pyramidion über die Rampe zu transportieren, müsste diese allerdings über eine fünf mal so lange Basisstrecke verfügen (2 000 m). Allein die Dämme der Rampe hätten dann ein Vielfaches des Volumens der gesamten Pyramide erfordert! ■

Anwendung: Keilkräfte (Abb. 3.28)

Ein symmetrischer Keil mit dem Spitzwinkel α wird mit der Kraft F in die Unterlage gedrückt. Wie groß sind die Normalkräfte F_1 bzw. F_2 gegen die Wände?

Lösung:

$$F_1 = F_2 = \frac{\frac{F}{2}}{\sin\varphi} = \frac{F}{2\sin\frac{\alpha}{2}}$$

F_1 wird somit umso größer, je kleiner $\sin\varphi$ wird, also je kleiner α ist. ■

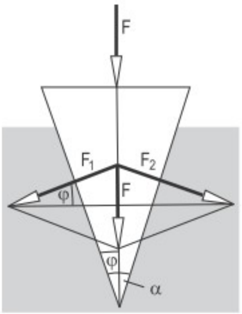

Abb. 3.28 Keilkräfte in Theorie…

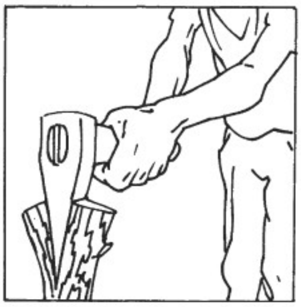

Abb. 3.29 … und Praxis

Anwendung: Keilriemen (Abb. 3.30)
Gegeben seien die Radien R und r sowie der Zentralabstand d. Wie lang ist der Riemen?

Lösung:
So einfach das Beispiel scheint, benötigen wir doch einen kleinen Trick, den man „sehen" muss: Durch Parallelverschieben des gemeinsamen Tangentenstücks der beiden Kreise ergibt sich ein rechtwinkliges Dreieck, in dem dann der Pythagoras anwendbar ist: $t = \sqrt{d^2 - (R-r)^2}$. Weiters ist dort $\sin\varphi = (R-r)/d \;\Rightarrow \varphi$ (im Bogenmaß!). Damit ergibt sich die Riemenlänge mit

$$L = 2\left[R\left(\frac{\pi}{2} + \varphi\right) + t + r\left(\frac{\pi}{2} - \varphi\right)\right]$$

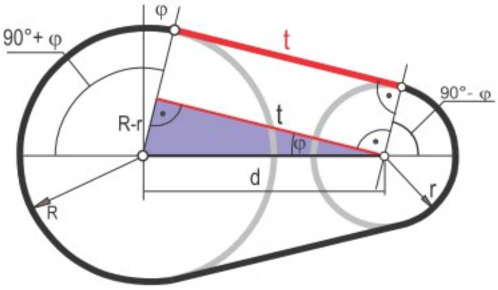

Abb. 3.30 Keilriemen

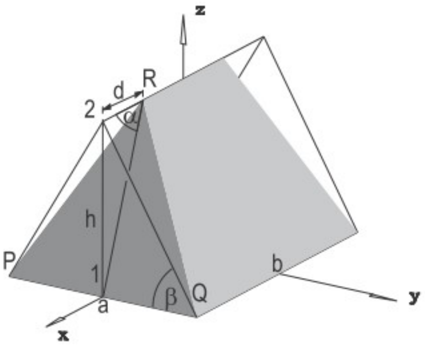

Abb. 3.31 Walmdach

Anwendung: Walmdach (Abb. 3.31)
Man berechne das Volumen des skizzierten Walmdachraumes bei gegebenen Traufenlängen a, b und gegebenen Dachneigungswinkeln α, β.

Lösung:
Im Dreieck $Q12$ gilt

$$h = \frac{a}{2}\tan\beta,$$

Damit ist die Fläche A des Dreiecks $PQ2$ bestimmt:

$$A = \frac{ah}{2}.$$

Das Volumen ergibt sich aus der Differenz des Volumens des dreiseitigen Prismas mit Querschnittsfläche A und Länge b und des doppelten Volumens des Tetraeders $PQ2R$, dessen Dreieck $PQ2$ ebenfalls die Fläche A hat, und dessen dazu normale Kante $R2$ die Länge d hat. d ergibt sich aus dem rechtwinkligen Dreieck $12R$ mit

$$d = h/\tan\alpha$$

$$\Rightarrow V = A\,b - 2\,\frac{A\,d}{3} = A(b - \frac{2d}{3})$$ ■

Anwendung: Kurvenlage eines Motorrads (Abb. 3.32)
Wenn ein Motorrad (Gewicht $G = m\,g$) mit der Geschwindigkeit v eine Kurve (Radius r) durchfährt, muss es sich um den Winkel α neigen, um nicht durch die Fliehkraft ($F = mv^2/r$) umzukippen.

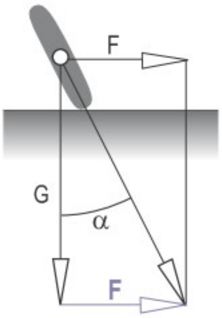

Abb. 3.32 Kurvenlage in Theorie...

Abb. 3.33 ...und Praxis

Lösung:
Es ist

$$\tan\alpha = \frac{F}{G} = \frac{v^2}{gr}$$

(unabhängig von der Masse, Abb. 3.34). Außerdem muss die Reibungskraft größer als F sein.

Zahlenbeispiel:

1) Aus einem Foto (Abb. 3.33 links) sind der Neigungswinkel α und der Kurvenradius r bekannt. Wie schnell war das Moped auf dem Gartenweg unterwegs?
$\alpha = 15°,\ r = 10\,\text{m} \Rightarrow v = \sqrt{g\,r\,\tan\alpha} \approx 5\,\text{m/s} \approx 18\,\text{km/h}$
2) Wie groß muss der Winkel sein, wenn v und r gegeben sind?
$r = 70\,\text{m},\ v = 72\,\text{km/h} = 20\,\text{m/s} \Rightarrow \tan\alpha \approx 0{,}58 \Rightarrow \alpha \approx 30°$

3) Der abgebildete Extrem-Carver (Abb. 3.35) fährt eine Kurve mit einem Radius von $r = 5$ m mit einer Geschwindigkeit von $v = 12$ m/s. Daraus ergibt sich $\tan\alpha = 2{,}88$, was bedeutet, dass das Brett unter einem Winkel von 71° geneigt sein muss, damit der Carver sein Gleichgewicht halten kann. Dabei wirkt mehr als das Dreifache der Erdanziehungskraft auf den Sportler. Das Berühren des Schnees wie auf dem Foto ist eine physikalische Notwendigkeit bei einer Standard-Hangneigung von 20°. Der Athlet ist gezwungen, sich buchstäblich über den Schnee zu ziehen, wenn er es um die Kurve schaffen will. ■

Abb. 3.34 Der Neigungswinkel hängt von der Geschwindigkeit, nicht aber von der Masse ab!

Abb. 3.35 Extremes Snowboarding

Anwendung: Verebnung eines Drehkegels (Abb. 3.36)
Ein Drehkegel (Radius r, Höhe h) geht bei Abwicklung („Verebnung“) in einen Kreissektor mit dem Radius s und dem Zentriwinkel ω über. Man berechne ω in Abhängigkeit vom halben Öffnungswinkel α des Kegels.

Lösung:
Gemäß Abb. 3.36 müssen die Bogenlängen am Basiskreis des Drehkegels und jene am Kreissektor übereinstimmen, also gelten:

$$2\pi\, r = s\,\omega \Rightarrow \omega = 2\pi\,\frac{r}{s}$$

Mit $\dfrac{r}{s} = \sin\alpha$ ergibt sich die Formel

$$\omega = 2\pi\,\sin\alpha \quad \text{bzw.} \quad \omega^\circ = 360^\circ\,\sin\alpha \tag{3.18}$$

Die zweite Variante der Formel entsteht dabei aus der ersten durch Multiplikation mit $\frac{180}{\pi}$. ■

Anwendung: Schubkurbelgetriebe (Otto-Motor) (Abb. 3.37 links)
Der Kolben treibt über den Kolbenbolzen B und die Schubstange AB den Kurbelzapfen A an und bewirkt die Drehung der Kurbelwelle L. Man be-

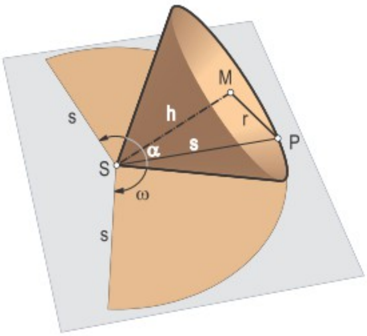

Abb. 3.36 Verebnung des Drehkegels

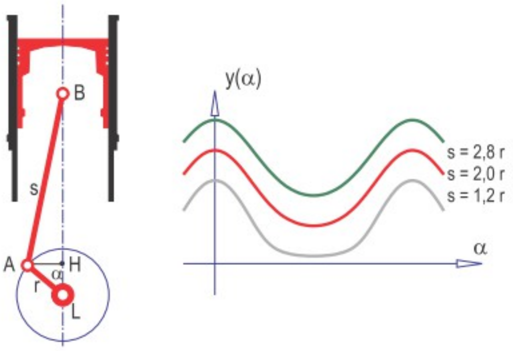

Abb. 3.37 *Otto*-Motor, Zeit-Weg-Diagramme

rechne den Abstand $\overline{LB}$ in Abhängigkeit vom Drehwinkel α – diese Distanz ist der Schlüssel zur gesamten kinematischen Untersuchung des Getriebes (Geschwindigkeits- und Beschleunigungsverhältnisse).

Lösung:

Die Strecke $y = \overline{LB}$ ist natürlich vom Drehwinkel α abhängig, und dieser ist wiederum proportional zur Zeit t. Man erhält also mit der Gleichung $y = y\big(\alpha(t)\big) = y(t)$ ein Weg-Zeit-Diagramm von B.

$$\overline{LB} = \overline{LH} + \overline{HB}$$

$$\overline{LH} = r\,\cos\alpha,\ \overline{AH} = r\,\sin\alpha,\ \overline{HB} = \sqrt{s^2 - \overline{AH}^2}$$

$$y = \overline{LB} = r\,\cos\alpha + \sqrt{s^2 - r^2\sin^2\alpha}$$

Abb. 3.37 rechts zeigt solche Diagramme $y = y(\alpha)$ für verschiedene Proportionen $r : s$. Für $s >> r$ (sprich: s viel größer als r) gilt

$$\sqrt{s^2 - r^2\sin^2\alpha} = s\sqrt{1 - \left(\frac{r}{s}\right)^2\sin^2\alpha} \approx s$$

und das Weg-Zeit-Diagramm nähert sich der gewöhnlichen Sinuslinie (die Graphen von $y = \sin x$ und $y = \cos x$ unterscheiden sich nur durch Parallelverschiebung längs der x-Achse um $\frac{\pi}{2}$. In beiden Fällen spricht man daher von einer Sinuslinie)

$$y = r\,\cos\alpha + s.$$

Das Demo-programm `otto_engine.exe` auf der Webseite zum Buch (`www.uni-ak.ac.at/math`) wurde unter Verwendung von $y(\alpha)$ erstellt. Für den Maschinenbauer ist zusätzlich noch ein Geschwindigkeitsdiagramm wichtig. Dieses erhält man, wie wir im Kapitel *Differentialrechnung* sehen werden, durch „Ableiten" der Funktion (Anwendung S. 281). ■

Relativ häufig brauchen wir folgende Formel:

$$\sin 2\alpha = 2 \sin\alpha \cos\alpha \tag{3.19}$$

∘ *Beweis*:
Wir betrachten das rechtwinklige Dreieck ABC in Abb. 3.38 (mit $\overline{AC} = 1$). Es hat die Fläche

$$A_1 = \frac{1}{2} \sin\alpha \cos\alpha$$

Wir spiegeln es an AB und erhalten ein doppelt so großes (gleichschenkeliges) Dreieck ACC^* (Fläche $A_2 = \sin\alpha\cos\alpha$). Die Höhe CH in diesem Dreieck ist $\sin 2\alpha$. Damit lässt sich die Fläche A_2 auch wie folgt berechnen:

$$A_2 = \frac{1 \cdot \sin 2\alpha}{2} = \sin\alpha\cos\alpha = 2\,A_1$$

■

Anwendung: Fläche des regelmäßigen n-Ecks (Abb. 3.39)
Man berechne die Fläche eines regelmäßigen n-Ecks mit Umkreisradius r.

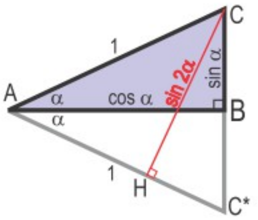

Abb. 3.38 $\sin 2\alpha = 2\sin\alpha\cos\alpha$

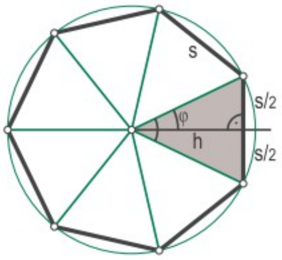

Abb. 3.39 Fläche des reg. n-Ecks

Lösung:
Es gilt

$$\frac{s}{2} = r\,\sin\varphi,\ \ h = r\,\cos\varphi \quad \text{mit} \quad \varphi = \frac{360^\circ}{2n}.$$

Daraus berechnet sich die Fläche

$$A = n \cdot \frac{s}{2} \cdot h = n\,r^2\,\sin\varphi\cos\varphi = n\,r^2 \frac{\sin 2\varphi}{2} \ \Rightarrow$$

$$A = \frac{n\,r^2}{2} \sin\frac{360^\circ}{n} \tag{3.20}$$

∘ Lassen wir n „gross“ werden. Dann „konvergiert“ die Fläche des Vielecks gegen die Kreisfläche $\pi\,r^2$. Schon *Archimedes* hat vor mehr als 2 000 Jahren auf ähnliche Weise – nämlich mit dem *Umfang* des regelmäßigen n-Ecks – die Zahl π recht gut angenähert.
Verwenden wir die „Grenzwertschreibweise“, dann gilt zunächst

$$A_\circ = r^2 \lim_{n\to\infty} \frac{n}{2} \sin\frac{360^\circ}{n} = r^2\,\pi.$$

Durch Übergang zum Bogenmaß ($360° = 2\pi$) erhalten wir nach Kürzen durch r^2 und Division durch π

$$\lim_{n\to\infty} \frac{n}{2\pi} \sin \frac{2\pi}{n} = 1.$$

Schreiben wir statt $\frac{2\pi}{n}$ noch x (sodass x gegen 0 konvergiert, wenn n gegen unendlich strebt), dann erhalten wir eine wichtige Beziehung, die wir in der Differentialrechnung brauchen:

$$\lim_{x\to 0} \frac{1}{x} \cdot \sin x = \lim_{x\to 0} \frac{\sin x}{x} = 1 \tag{3.21}$$

In Worten ausgedrückt:

Für sehr kleine Winkel stimmen Bogenmaß und Sinus überein.

■

Anwendung: Schiefer Wurf nach oben und maximale Wurfweite

Man errechnet die Höhe h nach t Sekunden mittels Modifizierung von Formel (1.17) in Anwendung S. 32:

$$h = G(t) - B(t) = (v_0 \sin\alpha\,)t - \frac{g}{2}t^2 \tag{3.22}$$

v_0 und g sind wieder Anfangsgeschwindigkeit und Erdbeschleunigung. α bezeichnet den Anfangswinkel zur Horizontalen. Der Luftwiderstand wird vernachlässigt.

Man berechne die Wurfweite bei gegebenem v_0 und α. Für welches α erzielt man die größte Wurfweite?

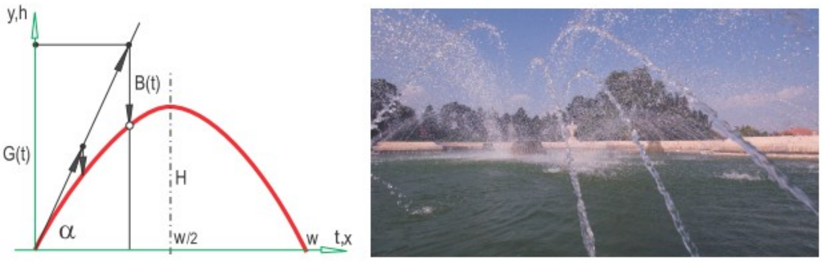

Abb. 3.40 Schiefer Wurf nach oben

Lösung:

Durch Lösen der quadratischen Gleichung (3.22) erhält man

$$t_{1,2} = \frac{1}{g}\left(v_0 \sin\alpha \pm \sqrt{v_0^2 \sin^2\alpha - 2gh}\right)$$

Der höchste Punkt wird für $D = v_0^2 \sin^2\alpha - 2gh = 0$, also

$$\text{zum Zeitpunkt } t_0 = \frac{v_0 \sin\alpha}{g} \text{ in der Höhe } H = \frac{v_0^2 \sin^2\alpha}{2g}$$

erreicht. Die Gesamtzeit bis zum Aufprall in der Ebene $h = 0$ ist dann $2t_0$. Für die horizontale Komponente (die x-Koordinate des Wurfobjekts) gilt $x(t) = t\,v_0 \cos\alpha$. Aus der Gesamtzeit ergibt sich die Wurfweite:

$$w = 2t_0 v_0 \cos\alpha = \frac{2\sin\alpha\cos\alpha\, v_0^2}{g} = \frac{1}{g} v_0^2 \sin 2\alpha$$

Abb. 3.41 *Galileos* Experiment

Abb. 3.42 Hoher Weitsprung oder weiter Hochsprung?

Die theoretisch maximale Wurfweite ergibt sich für maximales $\sin 2\alpha$. Der Sinus nimmt bei 90° seinen Maximalwert 1 an. Aus $\sin 2\alpha = 1$ folgt $2\alpha = 90°$ und daher ist für $\alpha = 45°$ die Wurfweite maximal. In Anwendung S. 386 wird die maximale Wurfweite am schrägen Hang besprochen.

∘ In der Praxis muss noch – spätestens seit *Galileos* Experiment mit den Kugeln aus Eisen und Holz vom Turm zu Pisa (Abb. 3.41) – der Luftwiderstand einkalkuliert werden. Die Wurfparabel wird dann eine Parabel 3. Ordnung. Beim Weitwerfen eines Schlagballs, beim Kugelstoßen, beim Speerwurf, aber auch beim Weitsprung (Abb. 3.42), ist für einen Anfangswinkel von $\alpha = 45°$ meist aus physiologischen Gründen nicht die größte Anfangsgeschwindigkeit v_0 erreichbar. Diese ist aber noch wichtiger als der Anfangswinkel: Die Wurfweite (Sprungweite) wächst ja *quadratisch* mit v_0. Deshalb sind die besten Sprinter oft auch die besten Weitspringer (Carl Lewis). ■

Gleichungen vom Typ $P \sin x + Q \cos x + R = 0$

Dieser Gleichungstyp kommt in der Praxis recht häufig vor (Anwendung S. 144, Anwendung S. 340) und führt auf eine quadratische Gleichung:
Wir formen zunächst um und erhalten

$$P \sin x = -Q \cos x - R \Rightarrow P^2 \sin^2 x = (Q \cos x + R)^2$$

$$\Rightarrow P^2 (1 - \cos^2 x) = Q^2 \cos^2 x + R^2 + 2QR \cos x$$

Setzen wir zur Abkürzung $u = \cos x$, dann haben wir die quadratische Gleichung

$$(P^2 + Q^2)\, u^2 + 2QR\, u + R^2 - P^2 = 0$$

mit den beiden (unter Umständen auch zusammenfallenden bzw. nicht reellen) Lösungen für $u = \cos x$

$$\begin{aligned} u_{1,2} &= \frac{-2QR \pm 2\sqrt{Q^2R^2 + (P^2+Q^2)(P^2-R^2)}}{2(P^2+Q^2)} = \\ &= \frac{-QR \pm \sqrt{P^4 + P^2Q^2 - P^2R^2}}{P^2+Q^2} = \frac{-QR \pm P\sqrt{P^2+Q^2-R^2}}{P^2+Q^2} \end{aligned}$$

Anwendung: Schnitt zweier konfokaler Kegelschnitte

In Kapitel 6 brauchen wir des Öfteren die Lösung der Gleichung (z.B. in Anwendung S. 340, Anwendung S. 340)

$$\frac{r_1}{1+\varepsilon_1\,\cos x} = \frac{r_2}{1+\varepsilon_2\,\cos(x+\delta)}$$

Sie tritt beim Schnitt zweier Kegelschnitte mit einem gemeinsamen Brennpunkt auf. Der konstante Wert δ ist der „Verdrehungswinkel" der beiden Kegelschnitte. Man führe die Gleichung mittels der Additionstheoreme auf die Form $P\,\sin x + Q\,\cos x + R = 0$, deren Lösung ja bekannt ist.

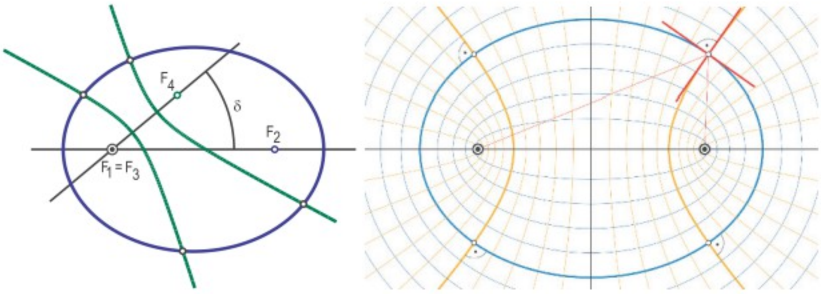

Abb. 3.43 Schnitt von Kegelschnitten mit einem bzw. zwei gemeinsamen Brennpunkten (rechts)

Lösung:
Durch Ausmultiplizieren erhalten wir zunächst

$$r_2(1+\varepsilon_1\,\cos x) = r_1(1+\varepsilon_2\,\cos(x+\delta)).$$

Setzen wir $\cos(x+\delta) = \cos x\,\cos\delta - \sin x \sin\delta = c\,\cos x - s\,\sin x$, dann haben wir

$$r_2 + r_2\,\varepsilon_1\,\cos x = r_1 + r_1\,\varepsilon_2\,(c\,\cos x - s\,\sin x)$$

und damit

$$\underbrace{(r_2\,\varepsilon_1 - r_1\,\varepsilon_2\,c)}_{P}\,\cos x + \underbrace{r_1\,\varepsilon_2\,s}_{Q}\,\sin x + \underbrace{r_2 - r_1}_{R} = 0$$

∘ Sind beide Brennpunkte identisch (Abb. 3.43 rechts), schneiden einander die Kurven (ohne Beweis) stets rechtwinklig. Man spricht in diesem Zusammenhang von orthogonalen Kegelschnittsbüscheln (vgl. Anwendung S. 544). ■

3.4 Das schiefwinklige Dreieck

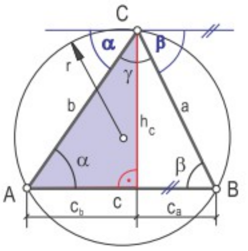

Abb. 3.44 Winkelsumme im Dreieck

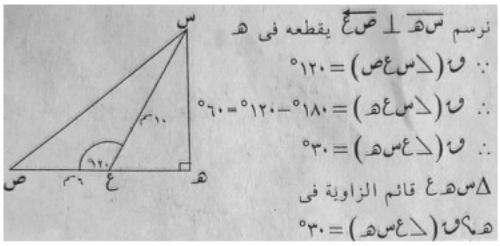

نرسم سه ⊥ صع يقطعه فى ه
∵ ق(∠سعص) = ١٢٠°
∴ ق(∠سعه) = ١٨٠° − ١٢٠° = ٦٠°
∴ ق(∠عسه) = ٣٠°
Δسهع قائم الزاوية فى
ه ∴ ق(∠عسه) = ٣٠°

Abb. 3.45 Aus einem arabischen Lehrbuch …

Zur eindeutigen Festlegung eines schiefwinkligen Dreiecks benötigen wir genau *drei Angabeelemente.*
Ausnahme: Die Angabe der drei Winkel α, β und γ genügt nicht, weil ja zwischen den drei Winkeln folgender Zusammenhang besteht: *Die Winkelsumme im Dreieck beträgt* 180°.

$$\boxed{\alpha + \beta + \gamma = 180^\circ} \tag{3.23}$$

∘ *Beweis*:
Wir zeichnen in C die Parallele zu AB (Abb. 3.44). Dort treten α und β als Parallelenwinkel noch einmal auf. ■
Aus dieser Beziehung folgt übrigens $\sin\gamma = \sin\big(180^\circ - (\alpha+\beta)\big)$ und somit

$$\sin\gamma = \sin(\alpha + \beta) \tag{3.24}$$

∘ Die sog. elliptische Geometrie beschäftigt sich mit Räumen, in denen man keine Parallele zu einer Geraden ziehen kann. Dort beträgt die Winkelsumme im Dreieck nicht 180°, sondern ist größer und nicht für jedes Dreieck gleich. Dies ist in der sphärischen Geometrie von Bedeutung (sphärische Dreiecke, siehe dazu Abb. 4.71), spielt aber auch in der modernen Physik eine wichtige Rolle (Schlagwort „gekrümmter Raum“).
Seien nun die Seiten und Winkel des Dreiecks bekannt. Wir berechnen die Hilfsgrößen h_c, c_a und c_b:

$$h_c = a\sin\beta,\ \ c_a = a\cos\beta,\ \ c_b = c - c_a = c - a\cos\beta. \tag{3.25}$$

Nun gilt nach Pythagoras

$$c_b^2 + h_c^2 = b^2 \Rightarrow (c - a\cos\beta)^2 + (a\sin\beta)^2 = b^2$$
$$\Rightarrow\ c^2 + a^2\underbrace{(\cos^2\beta + \sin^2\beta)}_{1} - 2ac\cos\beta = b^2,$$

was zum

Kosinussatz führt, der eine Verallgemeinerung des Pythagoreischen Lehrsatzes darstellt (die beiden unteren Formeln entstehen durch „zyklisches Vertauschen“ der Bezeichnungen):

$$\begin{aligned} b^2 &= c^2 + a^2 - 2ca\cos\beta \\ c^2 &= a^2 + b^2 - 2ab\cos\gamma \\ a^2 &= b^2 + c^2 - 2bc\cos\alpha \end{aligned} \tag{3.26}$$

Um sich von Bezeichnungsweisen zu lösen, merkt man sich den Satz am besten in Worten:

Das Quadrat über einer Seite ist gleich der Summe der Quadrate der beiden anderen Seiten, vermindert um das doppelte Produkt der anderen Seitenlängen mal dem Kosinus des eingeschlossenen Winkels.

Bei Angabe der Dreieckseiten lassen sich damit die Winkel berechnen, z.B.

$$\cos\alpha = \frac{b^2 + c^2 - a^2}{2bc}$$

Anwendung: Kräfte im Gleichgewicht (Abb. 3.46)
Drei komplanare Kräfte (also Kräfte, die in einer Ebene wirken)

$$F_1 = 60\text{N}, F_2 = 45\text{N}, F_3 = 30\text{N}$$

greifen in einem Punkt an und halten einander das Gleichgewicht. Berechnen wir die Winkel, die sie paarweise miteinander einschließen!
Lösung:

$$a = 60,\ b = 45,\ c = 30 \Rightarrow \text{Kosinussatz} \Rightarrow$$

$$\alpha = 104{,}5^\circ,\ \beta = 46{,}6^\circ \Rightarrow$$

$$\varphi = \alpha + \beta = 151{,}1^\circ,\ \psi = 180^\circ - \alpha = 75{,}5^\circ,\ \sigma = 180^\circ - \beta = 133{,}4^\circ. \qquad ■$$

Sind zwei Seiten und der von ihnen eingeschlossene Winkel gegeben, wendet man ebenfalls den Kosinussatz an.

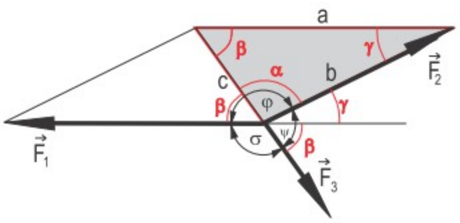

Abb. 3.46 Kräfte im Gleichgewicht

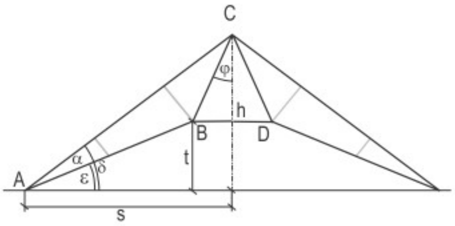

Abb. 3.47 Belgischer Dachbinder

Anwendung: Belgischer Dachbinder (Abb. 3.47)
Gegeben sind Dachneigungswinkel δ, Untergurtneigungswinkel ε, Gesamtbreite $2s$, Untergurtlänge $\overline{AB}$. Gesucht sind die anderen Balkenlängen.

Lösung:
Ermittelt man die Sparrenlänge $\overline{AC} = \dfrac{s}{\cos\delta}$, so sind vom Dreieck ABC zwei Seiten AB, AC und der eingeschlossene Winkel $\alpha = \delta - \varepsilon$ bekannt. Die Seite $\overline{BC}$ liefert der Kosinussatz:

$$\overline{BC} = \sqrt{\overline{AB}^2 + \overline{AC}^2 - 2\overline{AB}\,\overline{AC}\cos\alpha}$$

$$\overline{BD} = 2(s - \overline{AB}\,\cos\varepsilon)$$

■

Anwendung: Schnitt zweier Kugeln (Abb. 3.48)
Der Schnitt zweier Kugeln Σ_1, Σ_2 (Mittelpunkte M_1, M_2, Radien r_1, r_2, „Zentralabstand“ $\overline{M_1M_2} = d$) ist ein Kreis k mit Mitte M und Radius r in einer Normalebene zur Achse M_1M_2. Man berechne r und $e = \overline{MM_1}$. Für welches d ist $r = 0$ bzw. $r^2 < 0$ ($\Rightarrow$ r nicht reell)?

Lösung:
In Abb. 3.48 erkennen wir ein schiefwinkliges Dreieck M_1M_2S mit den bekannten Seitenlängen r_1, r_2 und d. Für dieses gilt:

$$r_2^2 = r_1^2 + d^2 - 2r_1d\,\cos\alpha \Rightarrow \cos\alpha = \frac{r_1^2 + d^2 - r_2^2}{2r_1d}$$

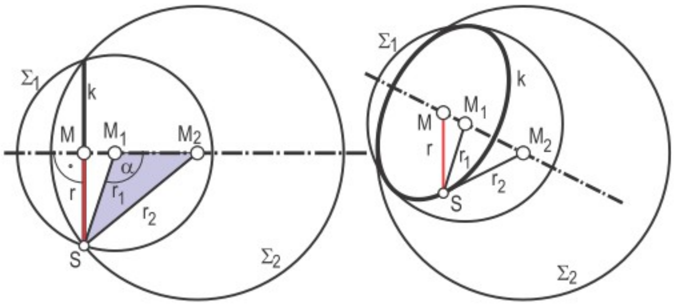

Abb. 3.48 Schnittkreis zweier Kugeln (spezielle und allgemeine Ansicht)

Es muss gelten:

$$\cos\alpha \leq 1 \Rightarrow \frac{r_1^2 + d^2 - r_2^2}{2r_1d} \leq 1$$

$$\Rightarrow r_1^2 + d^2 - r_2^2 \leq 2r_1d \Rightarrow r_1^2 - 2r_1d + d^2 \leq r_2^2 \Rightarrow (r_1 - d)^2 \leq r_2^2$$

Für $(r_1 - d)^2 > r_2^2$ wird der Schnittkreis nicht reell sein. Für

$$r_1 - d = \pm r_2 \Rightarrow r_1 - r_2 = d \text{ oder } r_1 + r_2 = d$$

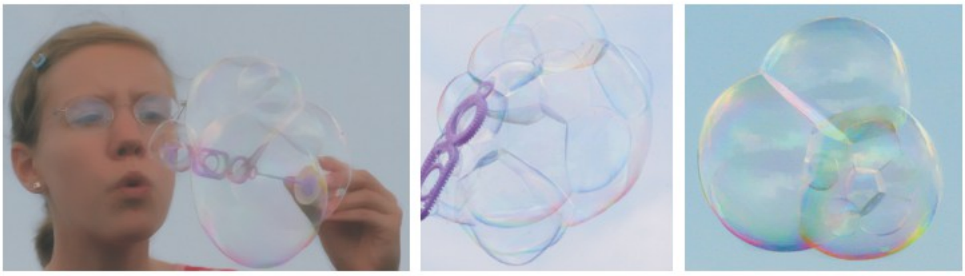

Abb. 3.49 Entstehung von Seifenblasen aus mehreren Kugeln

berühren sich die Kugeln in einem Punkt („Nullkreis"). Im Fall eines reellen Schnittkreises ergeben sich nun der Radius r und der Abstand e des Mittelpunkts M von M_1:

$$r = r_1 \sin\alpha \quad \text{und} \quad e = r_1 \cos\alpha$$

∘ Die Ebene, in welcher der Schnittkreis k liegt, heißt *Potenzebene*. Der Schnitt zweier Kugeln ist immer ein Kreis – reell oder imaginär. Mit Verweis auf Anwendung S. 544 sei erwähnt: Alle Kugeln haben zusätzlich den sog. *absoluten Kreis* gemeinsam, der imaginär ist und noch dazu „im Unendlichen" liegt. ■

Fläche des Dreiecks:

Zunächst gilt für die Fläche A des Dreiecks (mit den Bezeichnungen von Abb. 3.44)

$$A = \frac{c\,h_c}{2} = \frac{c\,b\,\sin\alpha}{2}$$

Weil die Beschriftung vertauscht werden kann, gilt ebenso

$$\boxed{A = \frac{a\,c\,\sin\beta}{2} = \frac{b\,a\,\sin\gamma}{2} = \frac{c\,b\,\sin\alpha}{2}} \tag{3.27}$$

Aus (3.27) folgt unmittelbar der

Sinussatz:

$$\boxed{\frac{a}{\sin\alpha} = \frac{b}{\sin\beta} = \frac{c}{\sin\gamma}} \tag{3.28}$$

In Worten:

Der Quotient aus Seitenlänge und Sinus des gegenüberliegenden Winkels ist konstant. Oder: Das Verhältnis zweier Seiten ist gleich dem Verhältnis der Sinus-Werte der gegenüberliegenden Winkel.

Damit lassen sich folgende Aufgaben lösen:

+,-.+ + +/ ! + % F! ! C ! ! / 1
! G , /1 6 ! ! # ? ! ,#6 0 0#
* & & =

Tycho and Mathematics

Geometry was an essential tool to convert Tycho's measurements into celestial positions. Living in the computer age as I do, I have to admit that I never gave this a second thought until I realized that he worked before logarithms were invented. Multiplication and division were therefore much more severe and labour-intensive problems than they were once logarithms became available.

The key art was spherical trigonometry, which is a shorthand name for the analysis of triangles on the surfaces of spheres. Of course, the sphere the early scientists thought of was a faraway sphere on whose inside surface were the "fixed" stars.

The parameters governing such a triangle are built up in Figs. 1.5, 1.6, 1.7, and 1.8.

Thus, our triangle on the surface of a sphere can be uniquely defined entirely in terms of angles, provided that its sides all lie on great circles.

Now let me define some planar triangles to help me analyse the spherical triangle (Fig. 1.9). I do this by drawing lines tangential to the two great circles that meet at A.

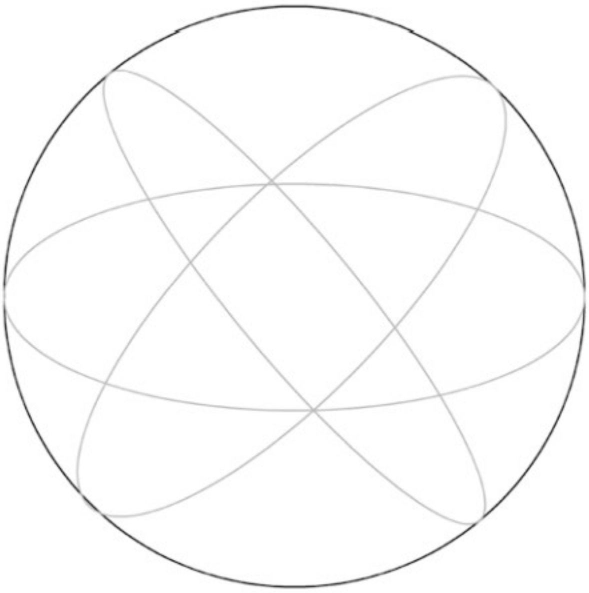

Fig. 1.5 Imagine a sphere onto whose surface three great circles are drawn. A great circle is a circle with the same radius as the sphere, whose centre is the centre of the sphere. (Image: Author)

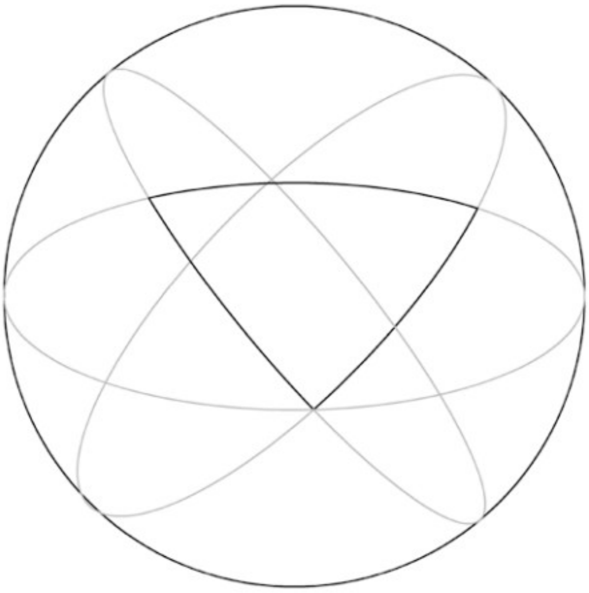

Fig. 1.6 These three great circles form the sides of a triangle on the sphere. They actually host another triangle on the other side of the sphere, but we will not consider that for now. (Image: Author)

Then by the cosine rule, these two equations are true.

$$\begin{aligned} \mathrm{DE}^2 &= \mathrm{AD}^2 + \mathrm{AE}^2 - 2(\mathrm{AD})(\mathrm{AE})\cos A; \\ \mathrm{DE}^2 &= \mathrm{OD}^2 + \mathrm{OE}^2 - 2(\mathrm{OD})(\mathrm{OE})\cos a. \end{aligned} \tag{1.2}$$

Because we chose tangents, the angles ∠OAD and ∠OAE are right angles. Thus,

$$\begin{aligned} \mathrm{OD}^2 &= \mathrm{OA}^2 + \mathrm{AD}^2; \\ \mathrm{OE}^2 &= \mathrm{OA}^2 + \mathrm{AE}^2. \end{aligned} \tag{1.3}$$

Substitute the equations of Eq. (1.3) into Eq. (1.2):

$$\begin{aligned} \mathrm{DE}^2 &= \mathrm{AD}^2 + \mathrm{AE}^2 - 2(\mathrm{AD})(\mathrm{AE})\cos A; \\ \mathrm{DE}^2 &= \mathrm{OA}^2 + \mathrm{AD}^2 + \mathrm{OA}^2 + \mathrm{AE}^2 - 2(\mathrm{OD})(\mathrm{OE})\cos \boldsymbol{a}. \end{aligned} \tag{1.4}$$

Now subtract the upper from the lower of Eq. (1.4).

$$0 = 2\left(\mathrm{OA}^2\right) + 2(\mathrm{AD})(\mathrm{AE})\cos A - 2(\mathrm{OD})(\mathrm{OE})\cos a. \tag{1.5}$$

Rearranging Eq. (1.5) gives

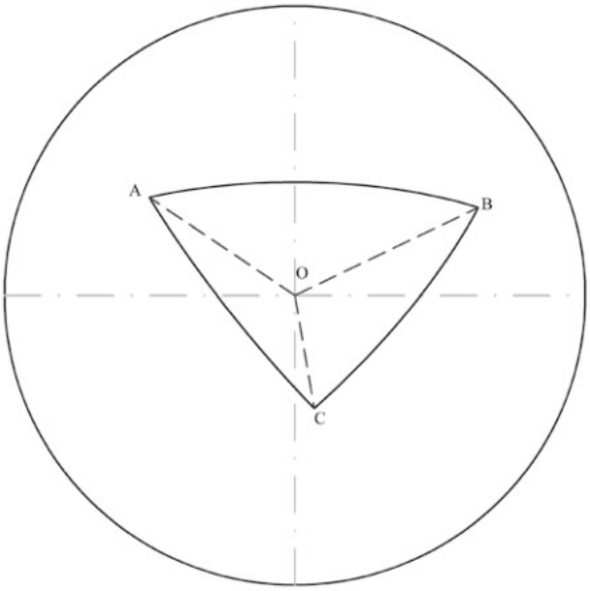

Fig. 1.7 Let this triangle have vertices A, B and C, all in upper case. Let the centre of the sphere be at O. (Image: Author)

$$\begin{aligned}\cos a &= \frac{\mathrm{OA}}{\mathrm{OE}}\frac{\mathrm{OA}}{\mathrm{OD}}+\frac{\mathrm{AE}}{\mathrm{OE}}\frac{\mathrm{AD}}{\mathrm{OD}}\cos A \\ \text{so } \cos a &= \cos b\cos c+\sin b\sin c\cos A\end{aligned} \tag{1.6}$$

using the definitions of sine as opposite/hypotenuse and cosine as adjacent/hypotenuse. Rearranging Eq. (1.6) gives

$$\cos A=\frac{\cos a-\cos b\cos c}{\sin b\sin c}\ . \tag{1.7}$$

In the pre-logarithmic age, the very useful Eq. (1.7) would have been messy to manipulate. It involves two multiplications and a division.

It is therefore worth some effort to see if we can turn the multiplications into additions or subtractions.

We will only deal with the cases where angles are right angles or less because Tycho's measurements did not use angles >90°. By following the logic in Fig. 1.10, it can be seen that

$$\sin\left(g+h\right)=\sin g\,cos\,h\ +\ \cos g\,sin\,h \tag{1.8}$$

and

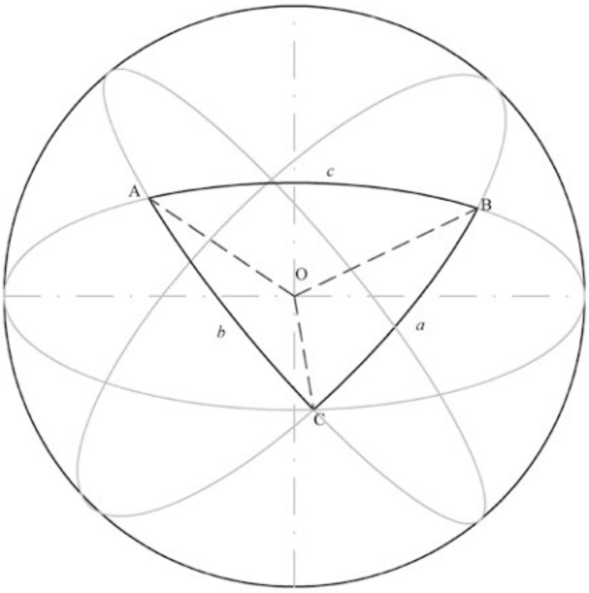

Fig. 1.8 Let the side opposite A has length *a*, let that opposite B has length *b* and let that opposite C has length *c*. A better way to analyse the triangle is to regard *a*, *b* and *c* as angles. *a* is the angle BOC; *b* is the angle COA and *c* is the angle BOA. (Image: Author)

$$\cos(g+h)=\cos g\cos h\ -\ \sin g\sin h \tag{1.9}$$

Proofs of these formulae for all angles, whether greater than 90° or not, are given in Appendix 3.

Eq. (1.8) can be used like this:

$$\sin g\sin\ h=\frac{1}{2}\left[\cos(f-g)-\cos(f+g)\right], \tag{1.10}$$

and Eq. (1.9) can be used like this:

$$\cos g\cos\ h=\frac{1}{2}\left[\cos(f-g)+\cos(f+g)\right], \tag{1.11}$$

These relations can be used in Eq. (1.7) to eliminate the multiplications but not the division:

$$\cos A=\frac{2\cos a-\left[\cos(b-c)+\cos(b+c)\right]}{\cos(b-c)-\cos(b+c)}. \tag{1.12}$$

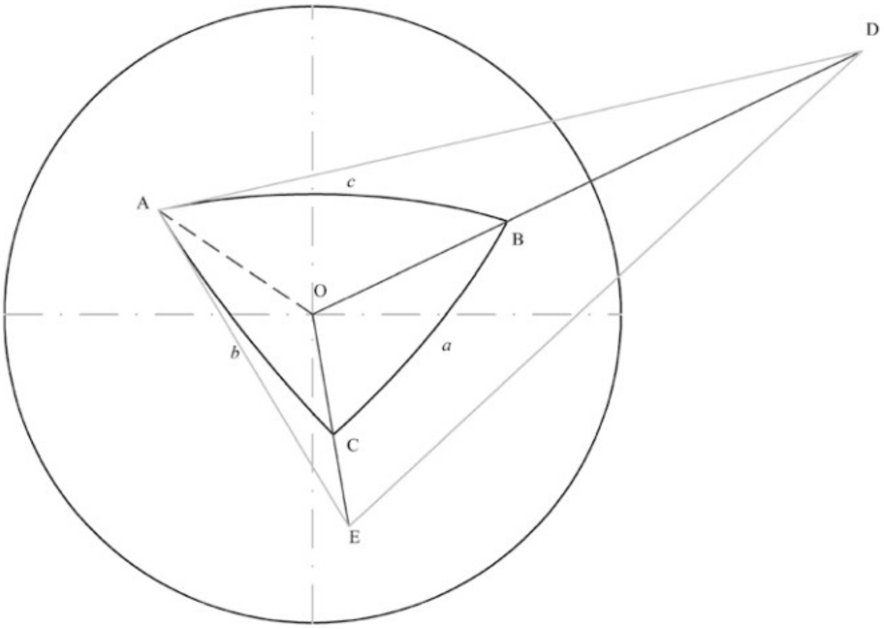

Fig. 1.9 Two lines are drawn tangent to the two great circles that cross at A. The lines OB and OC are projected until they meet these tangent lines at D and E, respectively. (Image: Author)

This elimination of multiplication was well worth the bother when you could not use logarithms. Indeed, the process was given a name: *prosthaphaeresis*. This compound word was built from the Greek words *prosthesis* and *aphaeresis* (subtract).

In those days, Eqs. (1.2)–(1.12) were not written out as formulae. They were written in words [24].

Tycho claimed priority for this technique. For this he has been in hot water with historians ever since. It's the perfect storm in a teacup. There does not appear to be enough evidence to settle the matter decisively [24], but it seems likely that Tycho was introduced to the formula by an assistant, Johannes Wittich [25]. Thoren [25] reports that Wittich brought the method to Tycho's observatory, having developed it beforehand. Tycho, who was nothing if not competitive, seems to have viewed prosthaphaeresis as something of a trade secret. He was not best pleased when he found out that, after leaving his employment, Wittich published the technique [24]. Such a driven man as Tycho venting his anger may well have been an impressive sight.

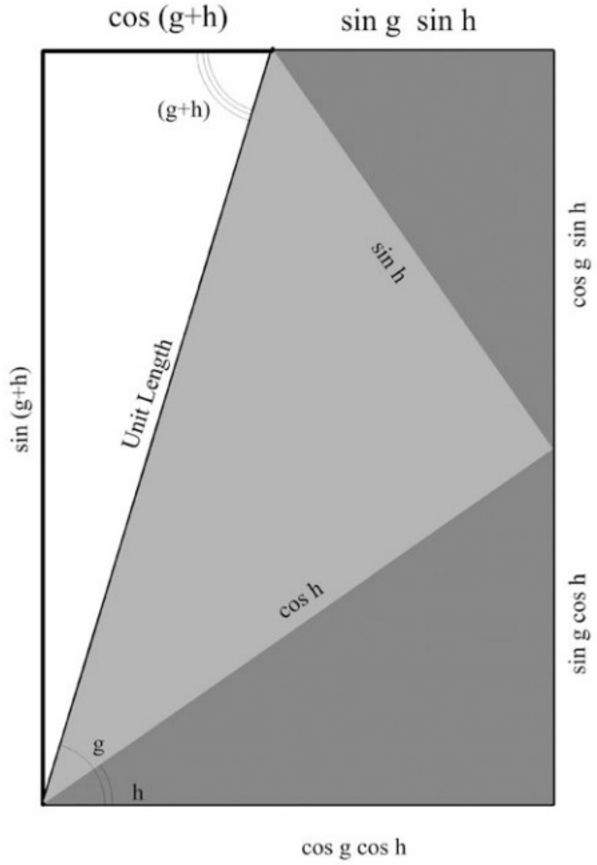

Fig. 1.10 Diagram showing how to manipulate sines and cosines of sums of angles

Tycho's Instruments

Tycho was no lone genius. In fact, you could even make a case for saying that his operation at his observatory in Uraniborg, Hveen, the Danish island now part of Sweden and known as Ven, was the first example of "big science". By the standards of today's big science, where the biggest projects employ thousands, Tycho's operation was small, yet his grant from the King of Denmark was 1% of the total royal expenditure. That was not to be sneezed at.

His first big challenge was to observe a supernova in 1572. He was careful to measure the position of the supernova relative to those of the other stars in Cassiopeia, within which constellation it had appeared. He looked

+,-.+ + +/ ! + % @ ! !
> /7 # 5 6 3 ; 5 ,)"' 0#
* %' '&

this disparity of emphasis that we are impelled to seek the reasons for it. There are several.

We mentioned earlier that in the classical period industry, commerce, and finance were conducted by slaves. Hence the educated people, who might have produced new ideas and new methods for handling numbers, did not concern themselves with such problems. Why worry about the use of numbers in measurement if one doesn't measure, or in trading if one dislikes trade? Nor do philosophers need the numerical dimensions of even one rectangle to speculate about the properties of all rectangles.

Like most philosophers the Greeks were star-gazers. They studied the heavens to penetrate the mysteries of the universe. But the use of astronomy in navigation and calendar reckoning hardly concerned the Greeks of the classical period. For their purposes, shapes and forms were more relevant than measurements and calculations, and so geometry was favored. Of these forms, the circle and sphere, suggested of course by superficial observation of the sun, moon, and planets, received the major share of attention. Hence their astronomical interests, too, led the classical Greeks to favor geometry.

The twentieth century seeks reality by breaking matter down—witness our atomic theories. The Greeks preferred to build matter up. For Aristotle and other Greek philosophers the form of an object is the reality to be found in it. Matter as such is primitive and shapeless; it is significant only when it has shape. It is no wonder, then, that geometry, the study of forms, was the special concern of the Greeks.

Finally, it was the solution of a vital mathematical problem that drove the Greek mathematicians into the camp of the geometers. We have already spoken of the fact that the Babylonian civilization, as well as earlier ones, used integers and fractions. The Babylonians were familiar also with a third type of number which arose through the application of a theorem on right triangles.

First, let us examine the theorem. If a right triangle has arms of lengths 3 and 4, the hypotenuse, or side opposite the right angle (*AB* in fig. 2) has length 5. Now the square of 5, namely 25, is the sum of the squares of 3 and 4, i.e., $5^2 = 3^2 + 4^2$. This relationship among the sides of a right triangle, that is, that the square of the length of the hypotenuse is equal to the sum of the squares of the lengths of the other two sides, is commonly known as the Pythagorean theorem.

To the Babylonians and Egyptians the fact, if not a proof, of this relationship was known.

Suppose now that the arms of a right triangle both have length 1 (fig. 3). What would be the length of the hypotenuse? Let us call the hypotenuse x. Then according to the Pythagorean theorem its length must be such that

$$x^2 = 1^2 + 1^2 = 2.$$

Hence x, the length of the hypotenuse, must be a number whose square is 2. We indicate the number whose square is 2 by $\sqrt{2}$ and call it the square root of 2. But what number equals $\sqrt{2}$? That is, what number multiplied by itself gives 2?

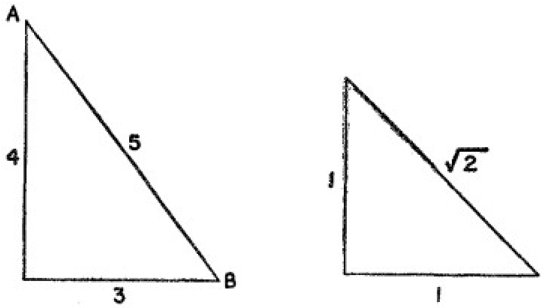

Figures 2 and 3. Two right triangles

The answer, as the Pythagorean school of mathematicians discovered to its great dismay, is that there is no whole number or fraction whose square is 2. $\sqrt{2}$ is a new kind of number, and they called it *irrational* because it could not be expressed exactly as a ratio of whole numbers, as 4/3 or 3/2. By contrast, whole numbers and fractions are called *rational* numbers. These terms are in use today.

The irrational number is a much neglected topic in the history of thought and a troublesome member of our number system. We have just seen that such numbers must be used in order to represent lengths and they are, moreover, explicitly and implicitly involved in almost all of mathematics. Yet how can we add, subtract, multiply, or divide such numbers? For example, how can we add 2 and $\sqrt{2}$? How do we divide $\sqrt{7}$ by $\sqrt{2}$?

The Babylonians had a makeshift, though practical, solution of these difficulties. They approximated the value of $\sqrt{2}$. For example, since the square of 14/10 or 1.4 is 1.96, and since 1.96 is nearly equal to 2, 1.4 must be nearly equal to $\sqrt{2}$. An even better approximation to $\sqrt{2}$ is 1.41 because the square of 1.41 is 1.988.

The Babylonian approximation to $\sqrt{2}$ does not permit exact reasoning with irrational numbers, for no matter how many decimal places we are willing to use we cannot write a rational number whose square is *exactly* 2. Yet, if mathematics is to merit its claim to being an exact study, it must evolve a method of working with $\sqrt{2}$ itself and not an approximation of it. To the Greek mind, this difficulty was as genuine and as prepossessing as the problem of food to a castaway on a coral reef.

Not content to use the less scrupulous method of the Babylonians, the Greeks undertook to face the logical difficulty squarely. In order to think about irrational numbers with exactness they conceived the idea of working with all numbers geometrically. They started out this way. A length was chosen to represent the number 1. Other numbers were then represented in terms of this length. To represent $\sqrt{2}$, for example, they used a length equal to the hypotenuse of a right triangle whose sides were one unit in length. The sum of 1 and $\sqrt{2}$ was a length formed by adjoining a unit segment to the length representing $\sqrt{2}$. In this geometrical form the sum of a whole number and an irrational one is no more difficult to conceive than the sum of one and one.

Similarly the product of two numbers, 3 and 5 for example, was expressed geometrically as the area of the rectangle with dimensions 3 and 5. In the case of 3 and 5 the use of area as a way of thinking about the product may be no great advantage. But one can also think of the product of 3 and $\sqrt{2}$ as an area. To think about this second rectangle is no more difficult than to think of the first one; yet it provides an exact way of working with the product of an integer and an irrational number or, for that matter, two irrational numbers.

The Greeks not only operated with numbers in the geometric manner but went so far as to solve equations involving unknowns by series of geometrical constructions. The answers to these constructions were line segments whose lengths were the unknown values. The thoroughness of their conversion to geometry may be judged from the fact that the product of four numbers was unthinkable in classical Greece because there was no geometric figure to represent it in the manner that area and volume represented the product of two and three numbers respectively. Incidentally, we still speak of a number such as 25 as the *square* of 5 and of 27 as the *cube* of 3 in conformity with Greek thought.

The preference of the Greeks for geometry was so marked that during his travel in Laputa, Gulliver was again forced to comment:

> The knowledge I had in mathematics gave me great assistance in acquiring their phraseology, which depended much upon that science and music; and in the latter I was not unskilled. Their ideas are perpetually conversant in lines and figures. If they would, for example, praise the beauty of a woman, or any other animal, they describe it by rhombs, circles, parallelograms, ellipses, and other geometrical terms, or by words of art drawn from music, needless here to repeat. I observed in the King's kitchen all sorts of mathematical and musical instruments, after the figures of which they cut up the joints that were served to his Majesty's table.

Because the Greeks converted arithmetical ideas into geometrical ones and because they devoted themselves to the study of geometry, that subject dominated mathematics until the nineteenth century, when the difficulties in treating irrational numbers on an exact, purely arithmetical basis were finally resolved. In view of the clumsiness and complexity of arithmetical operations geometrically performed, this conversion was, from a practical standpoint, a highly unfortunate one. The Greeks not only failed to develop the number system and algebra which industry, commerce, finance, and science must have, but they also hindered the progress of later generations by influencing them to adopt the more awkward geometrical approach. Europeans became so habituated to Greek forms and fashions that Western civilization had to wait for the Arabs to bring a number system from far-off India.

Unfortunate as this Greek perversion of the number system and of algebra may appear to us with our understanding of progress, it still should not invoke on Greek heads the condemnation that has sometimes been heaped there. The one backward step the Greeks took was in itself thoroughly reasonable; moreover, the damage done is heavily outweighed by the incomparable good of their other accomplishments.

When most people describe the Greek contributions to modern civilization, they talk in terms of art, philosophy, and literature. No doubt the Greeks deserve the highest praise for what they bequeathed to us in these fields. Greek philosophy is as alive and significant today as it was then. Greek architecture and sculpture, especially the latter, are more beautiful to the average educated person of the twentieth century than the creations of his own age. Greek plays still appear on Broadway. Nevertheless, the contribution of the Greeks that did most to determine the character of present-day civilization was their mathematics. By altering the nature of the subject in the manner we have related, they were able to proffer their supreme gift. This we proceed to examine.

Books-018-2021/ Importantíssimo/
Morris-Kline/ Morris Kline -
Mathematics in Western Culture...
(1964).pdf.
Pág. 36(65)

Two right triangles

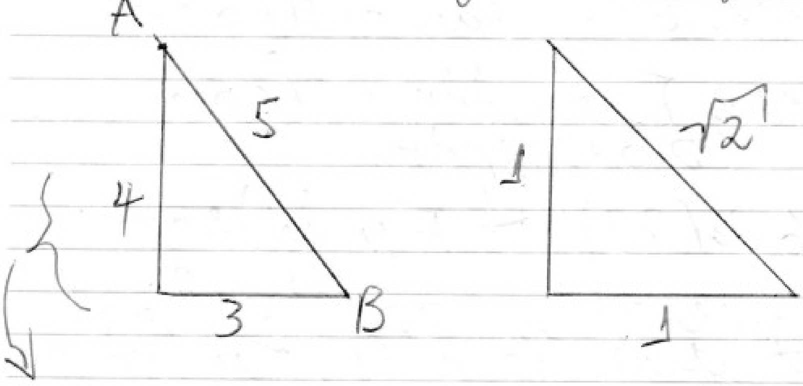

$3 \rightarrow$
$4 \rightarrow 3 + \textcircled{1} = 4$
$\hookrightarrow 1/3$ de 3

$5 \rightarrow 3 + \textcircled{2} = 5$
$\hookrightarrow 2/3$ de 3

$4 \rightarrow 4 + \textcircled{1} = 5$
$\hookrightarrow 1/4$ de 4

1
$1 + 1/3 = 1,33$
$1 + 2/3 = 1,66$

mantendo a mesma proporção.

data

$\sqrt{2} = 1,4142...$

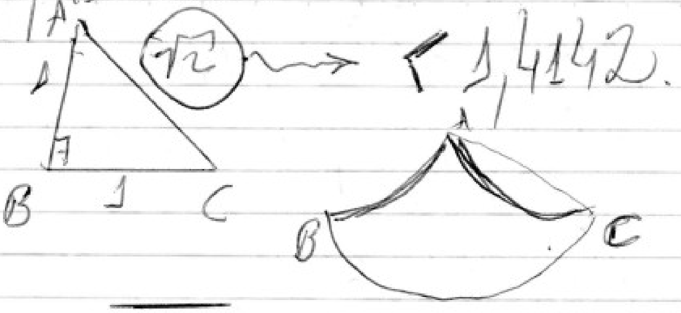

Books-018-2021/20210913/
The Disc Embedding Theorem/
The Disc Embedding Theorem...
(2021).pdf

Pág. 62(81) → Euclidean.

so we only need to show that it is locally Euclidean.

Let $E := J(S^{d-1} \times [0, \frac{1}{2}])$.

Pág. 92(111) → degree of the Gauss map $S^1 \times S^1 \to S^2$

Pág. 133(152) → Figure 10.1

A modification of the Cantor function. On the right, we see that certain vertical segments (red) are created.

jandaia

—

Pág. 219(238) - Fig 15.3
Tubing multiple points of intersection.

—

20210914 ////

Alexandre Augusto Bellei
CPF 541 656 009 49

—

persmatematik.blogpot.com
desmos.com

—

They want to see you succeed,
but never better than them.

—

Three-Dimensional Geometry
and Topology
William P. Thurston

—

Gravitation Misner.

—

Books:

Data-NovoLivo-018-2021/
Books-018-2021/

+,-.+ + +/ ! + % < B 9 1 ## <
/7 # 5 6 3 0#
* ' 9 # !

and fixing some smaller neighbourhood $W \subsetneq V$ of a. Recall that $D \subseteq S$. Define a map $\psi\colon D \to S$, as follows.

$$\psi(x) = \begin{cases} x & \text{if } x \in f^{-1}(W) \\ f^{-1} \circ h \circ f(x) & \text{if } x \in D \smallsetminus A. \end{cases}$$

The function above is well defined and continuous, since f is injective away from A and B and $f^{-1} \circ h \circ f(x) = x$ on $f^{-1}(W) \smallsetminus A$.

We also check that $B \subseteq \operatorname{Int} D$ is the only inverse set of ψ. This follows, since h maps $f(D)$ into V and f is injective away from A and B. To finish, we need to show that $\psi(\operatorname{Int} D)$ is open in S. We check by hand that $\psi(\operatorname{Int} D) = f^{-1} \circ h \circ f(\operatorname{Int} D)$. We know from before that $f(\operatorname{Int} D)$ is open in T. Then $f^{-1} \circ h \circ f(\operatorname{Int} D)$ is open, since h is a homeomorphism and f is continuous. Now apply Lemma 3.28 to the map $\psi\colon D \to S$ to conclude that B is cellular. A similar proof shows that A is cellular. □

3.4 Brown's Proof of the Schoenflies Theorem

After our lengthy interlude in the previous section, we return to the Schoenflies theorem, which we restate below in an equivalent form.

Theorem 3.29 (Schoenflies theorem [Bro60]) *Let $i\colon S^{d-1} \hookrightarrow S^d$ be a continuous embedding admitting a bicollar. Then the closure of each component of $S^d \smallsetminus i(S^{d-1})$ is homeomorphic to D^d.*

Proof By the bicollar hypothesis, there exists $J\colon S^{d-1} \times [-1,1] \to S^d$ such that $J|_{S^{d-1}\times\{0\}}$ equals i. From elementary homology computations, as in the proof of Theorem 3.5, it follows that the complement of the image of J in S^d has exactly two connected components. Denote their closures by A and B, where A meets $J(S^{d-1} \times \{1\})$.

Observe that the quotient space $S^d/\{A, B\}$ is homeomorphic to S^d due to the existence of the bicollar. In other words, $S^d/\{A, B\}$ can be identified with the (unreduced) suspension of S^{d-1}, which is homeomorphic to S^d, as indicated in Figure 3.8. Thus we have the composition

$$f\colon S^d \xrightarrow{\pi} S^d/\{A, B\} \xrightarrow{\cong} S^d,$$

where $\cong$ denotes homeomorphism and π denotes the quotient map. This map $f\colon S^d \to S^d$ has exactly two inverse sets, namely A and B, and is surjective. By Proposition 3.27, each of A and B is cellular.

Let D denote the closed northern hemisphere of S^d, thought of as a subset of the codomain S^d. By definition, D is a copy of the d-dimensional ball. Let $U := A \cup (J(S^{d-1} \times (0,1]))$; that is, U is the component of $S^d \smallsetminus i(S^{d-1})$ containing A. Then we have the restriction $f|_{\overline{U}}\colon \overline{U} \to D$, whose unique inverse set is A, which is cellular in S^d and thus in $\overline{U}$ by Proposition 3.16.

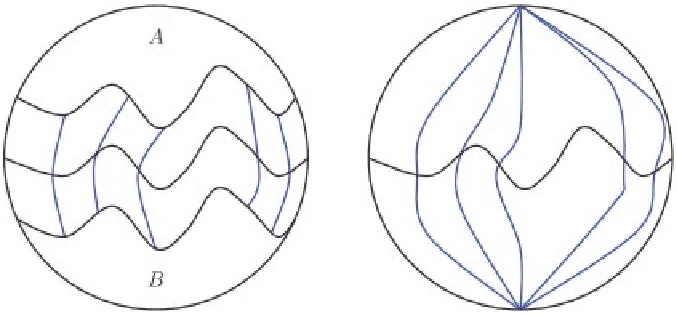

Figure 3.8 Brown's proof of the Schoenflies theorem. Going from left to right, the regions labelled A and B are collapsed to a point each, stretching out a neighbourhood of the equator in the process.

Our goal is to apply Proposition 3.22 to $\overline{U}$. To do so, we first need to show that $\overline{U}$ is a manifold (with boundary). The only possible failure could be near the boundary. As a subspace of S^d, $\overline{U}$ is already Hausdorff and second countable, so we only need to show that it is locally Euclidean. Let $E := J(S^{d-1} \times [0, \frac{1}{2}])$. Note that $\overline{U} = U \cup E$. Moreover, f restricts to a continuous bijection from E to some collar of ∂D. This collar of D is closed by the closed map lemma (Lemma 3.23). Therefore $f|_E$ is a homeomorphism, so $\overline{U}$ is a manifold, as needed.

Then we have the following diagram:

$$\begin{tikzcd}
\overline{U} \arrow[r, "f"] \arrow[d, "\pi"'] & D \\
\overline{U}/A. \arrow[ur, "\overline{f}"'] &
\end{tikzcd}$$

The map f is constant on the fibres of the quotient map $\pi\colon \overline{U} \to \overline{U}/A$ and thus descends to the map $\overline{f}$, which is a homeomorphism by the closed map lemma (Lemma 3.23). Next, since A is cellular, the map π is approximable by homeomorphisms by Proposition 3.22. Let $\widetilde{\pi}\colon \overline{U} \to \overline{U}/A$ be any such approximating homeomorphism. Then $\overline{f} \circ \widetilde{\pi}\colon \overline{U} \to D$ is a homeomorphism. It follows that $\overline{U}$ is homeomorphic to the d-dimensional ball $D \cong D^d$, as claimed. $\square$

$\widetilde{A}_2$, the core $\widetilde{q}$ of $\widetilde{Q}$ gives a path from a point in $\widetilde{A}_1$ to a point in $\widetilde{A}_2$, which must intersect at least one of D and D'. This contradicts that $\widetilde{Q}$ is disjoint from D and D'. Therefore, $\widetilde{Q}$ has substantial intersection with at least one of $\widetilde{B}_1$ and $\widetilde{B}_2$.

We see that if A does not have substantial intersection with P, then both of $\widetilde{A}_1$, $\widetilde{A}_2$, and at least one of $\widetilde{B}_1$ and $\widetilde{B}_2$, have substantial intersection with $\widetilde{Q}$. Pushing these intersections downstairs, we conclude that A and B both have substantial intersection with Q, completing the proof of Proposition 6.4. ☐

Now we are ready to show that the decomposition $\mathcal{B}_2$ does not shrink.

Proof of Theorem 6.1 Recall that we write T for the initial solid torus in the defining sequence for $\mathcal{B}_2$, and we call the two nested solid tori in the defining pattern P and Q, as shown in Figure 6.1. Let A and B be embedded meridional discs in T, union thin annuli of width $\delta > 0$ in $S^3 \smallsetminus \operatorname{Int} T$. Suppose that A and B are distance $\kappa > 0$ apart. Choose

$$\varepsilon < \min\{\kappa, \delta/2\}.$$

Write $\pi\colon S^3 \to S^3/\mathcal{B}_2$ for the quotient map. We want to choose a metric on $S^3/\mathcal{B}_2$, which we may do, since $\mathcal{B}_2$ is upper semi-continuous (Corollary 4.13). Hausdorff [Hau30] showed that for a metrizable space E with a subspace F, and prescribed metric on F, one may find a metric on E that extends the given metric on F. Apply this to $S^3/\mathcal{B}_2$, to obtain a metric that agrees with the standard metric on $\overline{S^3 \smallsetminus T}$. In other words, for all $x, y \in \overline{S^3 \smallsetminus T}$, we have $d_{S^3}(x,y) = d_{S^3/\mathcal{B}_2}(\pi(x), \pi(y))$. We have used that $\pi|_{\overline{S^3 \smallsetminus T}}$ is a homeomorphism.

Now let $h\colon S^3 \to S^3$ be a homeomorphism such that $d_{S^3/\mathcal{B}_2}(\pi(x), \pi(h(x))) < \varepsilon$ for every $x \in S^3$; that is, h satisfies (i) of Definition 4.16. Then, by the construction of the metric $d_{S^3/\mathcal{B}_2}$, for $x \in T$ and $y \in S^3 \smallsetminus T$, the distance $d_{S^3/\mathcal{B}_2}(x,y) \geq d_{S^3/\mathcal{B}_2}(\partial T, y)$. It follows that for every $x \in S^3$ whose distance from T is at least ε, both $x, h(x) \in \overline{S^3 \smallsetminus T}$ and so $d_{S^3}(h(x), x) < \varepsilon$.

For $\nu > 0$, let $N_\nu(T)$ denote a ν-neighbourhood of T, namely all the points of distance at most ν from T.

Claim *$h^{-1}(\partial A) \subseteq N_{3\delta/2}(T) \smallsetminus T$ and $h^{-1}(\partial B) \subseteq N_{3\delta/2}(T) \smallsetminus T$. Moreover, both curves have linking number one with the core of T.*

The first sentence is immediate from the properties of h and the choice of ε and δ. The homeomorphism h moves points of ∂A less than ε. The linking number between ∂A (respectively, $h^{-1}(\partial A)$) and the core of T can be computed as the degree of the Gauss map $S^1 \times S^1 \to S^2$ given by sending (x,y), with $x \in \partial A$ (respectively, $h^{-1}(\partial A)$) and y in the core of T, considered as vectors in $\mathbb{R}^3$, to $(x-y)/\|x-y\|$. But a perturbation of the Gauss map, small relative to the diameter of S^2, does not change its degree. This proves the claim.

We next prove that some decomposition element intersects both of the discs $A_0 := h^{-1}(A)$ and $B_0 := h^{-1}(B)$. By the claim, and linking number considerations, both A_0 and B_0 have substantial intersections with T. In more detail, ∂A_0 and the core of T have linking number one, so form a nontrivial link. The argument of Lemma 6.5, with the linking number used instead of the triple linking number, shows that A_0 and B_0 has a substantial intersection with T, and similarly for B. By Lemma 6.3, applied after making a small isotopy of ∂T so that A_0 and ∂T intersect transversely (we can make such an isotopy without affecting the decomposition elements), A_0 contains a meridian of T and therefore contains a smaller disc A_0' with boundary a meridian of T. Similarly, B_0 contains a smaller disc B_0' with boundary a meridian of T. By Proposition 6.4, at least one of P and Q has substantial intersection with both A_0' and B_0', and therefore with A_0 and B_0. Without loss of generality, let it be P.

By Lemma 6.3, A_0 and B_0 both contain discs A_1 and B_1 in S^3 bounded by meridians for P. Let $P_1, Q_1 \subseteq P$ be the solid tori of the next stage of the defining sequence for $\mathcal{B}_2$. Apply Proposition 6.4 again to see that at least one of P_1 and Q_1 has substantial intersection with both of A_1 and B_1, and so has substantial intersection with both A_0 and B_0. Again, we assume without loss of generality that P_1 has these substantial intersections.

Iterate this procedure to obtain a sequence $P \supsetneq P_1 \supsetneq P_2 \supsetneq \cdots$ in the defining sequence with substantial (and in particular nonempty) intersection with A_0 and B_0. Thus the decomposition element $\Delta = P \cap P_1 \cap P_2 \cap \cdots$ intersects both $A_0 = h^{-1}(A)$ and $B_0 = h^{-1}(B)$. This implies that $h(\Delta)$ intersects both A and B. Therefore, $\operatorname{diam}_{S^3} h(\Delta)$ is at least as large as the minimal distance from a point in A to a point in B; that is, $\operatorname{diam} h(\Delta) > \kappa > \varepsilon$.

Now $h\colon S^3 \to S^3$ was chosen to be some homeomorphism satisfying (i) of Definition 4.16. We have shown that there exists an ε such that for any such h, (ii) of Definition 4.16, that $\operatorname{diam}_{S^3} h(\Delta) < \varepsilon$ for every decomposition element $\Delta \in \mathcal{B}_2$, is not satisfied. Therefore, $\mathcal{B}_2$ does not shrink. This completes the proof of Theorem 6.1. $\square$

(c) The map f restricts to a homeomorphism $f|_{f^{-1}(E)}\colon f^{-1}(E) \to E$.

Then f can be approximated by homeomorphisms that agree with f on $f^{-1}(E)$.

The fact that f restricts to a homeomorphism from $S^3 = \partial D^4$ to itself implies that f has degree one and is thus surjective.

Observe that the inverse sets of f form an upper semi-continuous decomposition $\mathcal{D}$ of D^4 by Lemma 9.3, since every inverse set is closed in the compact space D^4. Let $\pi\colon D^4 \to D^4/\mathcal{D}$ be the quotient map. We obtain a well defined continuous function $\overline{f}\colon D^4/\mathcal{D} \to D^4$, since f is constant on the fibres of π. The function $\overline{f}$ is a continuous bijection from a compact space to a Hausdorff space and is thus a homeomorphism by the closed map lemma (Lemma 3.23). The ball to ball theorem establishes that f is approximable by homeomorphisms $f_n\colon D^4 \to D^4$. By composing with $\overline{f}^{-1}$, we see that the quotient map π is approximable by the homeomorphisms $\overline{f}^{-1} \circ f_n\colon D^4 \to D^4/\mathcal{D}$. This used that $\overline{f}^{-1}$ is uniformly continuous by the Heine–Cantor theorem (Theorem 3.25). The following diagram shows the maps. The diagram commutes when the top arrow denotes the function f.

$$\begin{tikzcd}
D^4 \arrow[r, "f", "f_n"'] \arrow[d, "\pi"'] & D^4 \\
D^4/\mathcal{D} \arrow[ur, "\overline{f}"'] &
\end{tikzcd}$$

We also remark that the restriction to dimension four is completely irrelevant and the proof we give will work in all dimensions. However, a result of Siebenmann [Sie73] shows that functions between manifolds of dimension five or higher, with far fewer restrictions than in Theorem 10.1, are approximable by homeomorphisms. This was known prior to the ball to ball theorem, which would have been an easy special case in high dimensions. Thus the 4-dimensional case is what requires a special argument.

10.1 The Main Idea of the Proof

Before going into the details, we outline the main idea of the proof. First, in an attempt to reduce confusion, we re-label the source and the target copies of D^4 as X and Y and consider the function $f\colon X \to Y$.

The proof will directly use the definition of a function being approximable by homeomorphisms (Definition 3.19). That is, given $\varepsilon > 0$, we will find a homeomorphism $h\colon X \to Y$ such that $d(h, f) < \varepsilon$. Note that f fails to be a homeomorphism exactly at the inverse sets, and at first sight the ones with largest diameter are the most offensive. With the goal of constructing a homeomorphism h as above, we wish to control the inverse sets with diameter greater than some fixed ε. By the hypothesis that the collection of inverse sets is null, we have only finitely many such inverse sets to worry about.

Consider the Cantor function (Figure 10.1) which has a large horizontal spot in the middle, demonstrating its non-injectivity. We present a strategy to approximate the Cantor function by homeomorphisms. Finding such an approximation for the Cantor function is

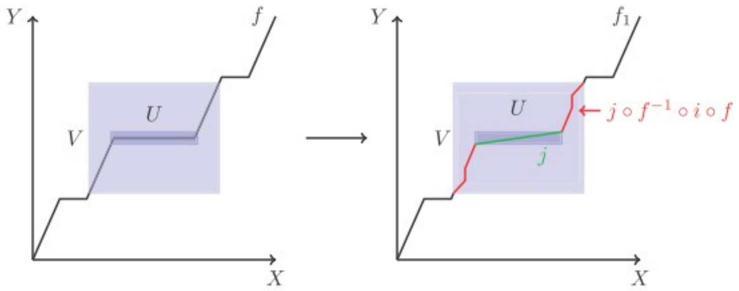

Figure 10.1 A modification of the Cantor function. On the right, we see that certain vertical segments (red) are created.

in general not hard, but the complicated strategy we are about to give has the advantage of working in the 4-dimensional context. We work with the graph of the Cantor function as in Figure 10.1 and find small neighbourhoods $U \subseteq V$ of the region in the graph corresponding to the largest inverse set, namely the horizontal spot in the middle. We then modify the function. It remains the same outside of V. Within U, we tilt it slightly in order to eliminate the inverse set. However, the price we pay is the creation of certain vertical cliffs in the region $V \smallsetminus U$. In other words, the result of the modification no longer has the large inverse set but is also no longer a function. It is instead a *relation*, meaning that it is a multi-valued function. We will define relations carefully soon. On the other hand, the vertical regions are located in small neighbourhoods and we may hope to be able to control them.

The strategy of the proof is then as follows. We perform the modification above to all the large inverse sets we wish to control. This replaces large horizontal steps with smaller vertical steps. Then we utilize the symmetry of the situation and consider the graph for the inverse function (really, relation). This has a number of vertical steps, upon which we can apply the same modification as above, before using the symmetry again. This process creates some new horizontal steps, but these have small diameter. By performing this iteration over and over again, we eventually reduce all the horizontal regions of our graph to small diameter, as desired.

In preparation for the proof, we give the precise formulae for the first step. Assume that some inverse set of f maps to $0 \in Y$, and suppose this is the inverse set we presently wish to control. Take two concentric, closed, round balls $U \subsetneq V \subseteq Y$ with distinct radii around $0 \in Y$, where U has very small radius. Define a stretching homeomorphism $i\colon V \to Y$ that is the identity on U and sends $V \smallsetminus U$ onto $Y \smallsetminus U$. In particular, $i|_{\partial V}\colon \partial V \to \partial Y$ is a homeomorphism. We could define such a function i as a radial expansion for now, but in later steps we will need more flexibility.

Note that the composition $i^{-1} \circ f\colon X \to V$ restricted to ∂X is a homeomorphism onto ∂V. Therefore, $(i^{-1} \circ f)|_{\partial X}$ can be extended to a homeomorphism $j\colon X \to V$ by the Alexander trick.

Finally, we define a relation $f_1 \colon X \to Y$ as follows.

$$f_1 = \begin{cases} f & \text{on } X \smallsetminus f^{-1}(V) \\ j \circ (f^{-1} \circ i \circ f) & \text{on } f^{-1}(V \smallsetminus U) \\ j & \text{on } f^{-1}(U), \end{cases} \tag{10.1}$$

using that $f(f^{-1}(V \smallsetminus U)) \subseteq V$. We check that the definitions match on the overlaps $\partial f^{-1}(V)$ and $\partial f^{-1}(U)$. Recall that, since f is continuous, $\partial f^{-1}(V) \subseteq f^{-1}(\partial V)$ and $\partial f^{-1}(U) \subseteq f^{-1}(\partial U)$. First, $i = \mathrm{Id}$ on U, so

$$j \circ f^{-1} \circ i \circ f = j \circ f^{-1} \circ f = j$$

on $f^{-1}(\partial U)$. Next, $j = i^{-1} \circ f$ on ∂X. Therefore,

$$j \circ f^{-1} \circ i \circ f = i^{-1} \circ f \circ f^{-1} \circ i \circ f = f$$

on $f^{-1}(\partial V)$. Observe that f_1 no longer has a singular point at 0, since $f_1^{-1}(0) = j^{-1}(0)$ and j is a homeomorphism. However, f_1 need not be a function, as we now explain. Consider a point $y \in V \smallsetminus U \subseteq Y$ such that $i(y)$ lies in the singular image of f. Recall that f is surjective. Then, given $x \in f^{-1}(V \smallsetminus U)$ with $y = f(x)$, the set $f^{-1}(i(f(x)))$ has cardinality strictly greater than one. This cardinality is preserved under the homeomorphism j. In other words, the relation f_1 maps x to the set $\{j \circ f^{-1} \circ i \circ f(x)\}$, which has more than one element, and as a result, f_1 is not a function. Helpfully, this failure only occurs for points mapping to V, and we could choose V to be arbitrarily small, as long as it is bigger than U. Thus the places where f_1 fails to be a function can be controlled.

This describes the principal idea behind the proof of the ball to ball theorem. In the next two sections, we address some necessary preliminaries.

10.2 Relations

Given sets X and Y, a subset of $X \times Y$ is said to be a *relation* from X to Y. For the rest of this section, we restrict ourselves to relations where, for every $x \in X$, there is some $y \in Y$ with $(x, y) \in R$. These are denoted by $R \colon X \to Y$, expressing the fact that the *domain* of R is all of X. We think of a relation $R \colon X \to Y$ as a multi-valued function which assigns to every $x \in X$ the subset $R\{x\} := \{y \in Y \mid (x, y) \in R\}$. Of course, any function from X to Y is a relation from X to Y. More precisely, we define the *graph* of a function $f \colon X \to Y$ as the set $\{(x, y) \mid x \in X, y \in Y, y = f(x)\} \subseteq X \times Y$. A function corresponds uniquely to its graph and thus corresponds to a relation. We will frequently refer to the graph of a function as simply the function for convenience. A relation $R \colon X \to Y$ is (the graph of) a function if and only if $R\{x\}$ is a singleton set for all $x \in X$, in which case we write $R(x)$ for the image of $x \in X$; that is, $R\{x\} = \{R(x)\}$ for all $x \in X$.

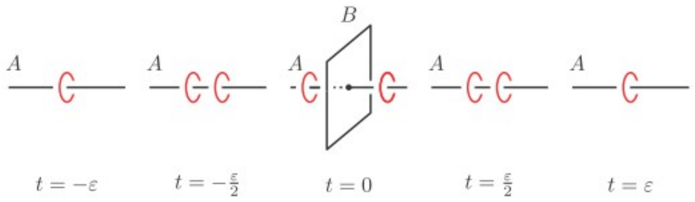

Figure 15.1 A Clifford torus at a double point of the surfaces A and B. As usual, the fourth coordinate is interpreted as time, t. The double point is seen at the $t = 0$ time slice. The red circles depicted trace out the Clifford torus as we move backwards and forwards in time. Note that each S^1 factor of the torus corresponds to a meridian of either A or B, and as such bounds a meridional disc that intersects the respective surface transversely at a single point. Compare with Figure 1.5.

Given a group G and $g \in G$, let $\langle\langle g \rangle\rangle$ denote the normal subgroup of G generated by g. We have that

$$\begin{aligned}\pi_1(M \smallsetminus (A' \cup B)) &\cong \pi_1(M \smallsetminus (A' \cup B \cup D))/\langle\langle \mu_D \rangle\rangle \\ &\cong \pi_1(M \smallsetminus (A' \cup B \cup D))/\langle\langle \gamma_T \rangle\rangle \\ &\cong \pi_1(M \smallsetminus (A \cup B))/\langle\langle [\mu_A, \mu_B] \rangle\rangle,\end{aligned}$$

as asserted.

15.2 Elementary Geometric Techniques

The following operations will be described on (immersed) surfaces. Where necessary, these operations can be extended to framed surfaces, and therefore to stages of capped or uncapped gropes or towers. However, we will not comment on this every time, unless there is extra particular care that needs to be taken.

15.2.1 Tubing

Tubing was described in [FQ90, Sections 1.8 and 1.9]. Suppose we have an immersed connected surface, A. Let Σ and Σ' be two other immersed surfaces intersecting A transversely at points p and p', respectively, with $p \neq p'$. Let γ be a smooth, embedded arc in A joining p and p'. Consider a tubular neighbourhood of γ intersecting A and B in small discs about p and p'. Cut out these discs from $A \cup B$ and glue on the rest of the boundary of the neighbourhood of γ to $\Sigma \cup \Sigma'$. In other words, we are gluing in a meridional annulus for γ. This process is called *tubing* Σ *into* Σ' (*along* $\gamma \subseteq A$) or vice versa. (In [FQ90], the resulting surface is called the *sum* of Σ and Σ'.) We shall at times also refer to this process as tubing p into the surface Σ'. If Σ and Σ' are distinct framed immersed surfaces, then the result of tubing is itself immersed and inherits a framing coming from the framing of the annulus, of Σ, and of Σ'. If Σ and Σ' are the same immersed framed surface, we need additionally that the intersection points p and p' have opposite signs, otherwise we would obtain a

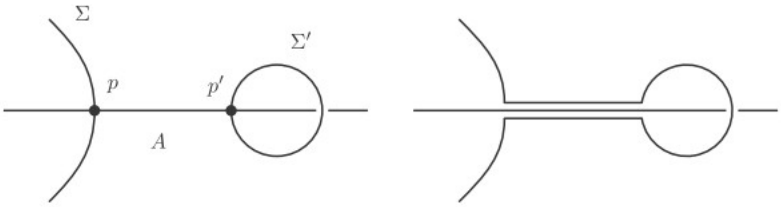

Figure 15.2 Tubing into a transverse sphere.

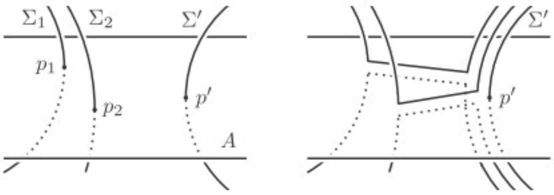

Figure 15.3 Tubing multiple points of intersection.

nonorientable surface. Note that the notion of 'opposite signs' is still meaningful even if the ambient space M is nonorientable, by using the arc γ to transport a local orientation at p to one at p'.

We will often tube surfaces into geometrically transverse spheres; that is, Σ' will be a framed sphere such that p' is its unique point of intersection with A. This situation is depicted in Figure 15.2. Note that this operation then decreases the number of points of intersection between A and Σ by one.

In most cases, we will tube into a parallel push-off of Σ' instead of Σ' itself. This will allow us to tube multiple times, as follows. Suppose that Σ' is a framed surface geometrically transverse to A; that is, A and Σ' intersect exactly once, at the point p'. Now suppose we have surfaces Σ_i intersecting A at the points p_i, for $i = 1, \ldots, k$. Then take k parallel push-offs of Σ', using distinct nonvanishing sections of the normal bundle (of the form $\Sigma' \times \{p\}$ in the given framing). These intersect A at k distinct points. Find pairwise disjoint, embedded arcs on A joining these points to the $\{p_i\}$ and use these to tube each Σ_i into a distinct parallel push-off of Σ'. This eliminates the intersection points $p_1, \ldots, p_k$, and also leaves the original Σ' unchanged for further tubing or other constructions. This process is described in Figure 15.3.

Note that if some surface B intersects Σ', then now Σ_i intersects B for each i. In particular, if Σ' has double points, then the tubing creates new points in $\Sigma_i \pitchfork \Sigma_j$, for all i, j, as well as in $\Sigma_i \pitchfork \Sigma'$.

15.2.2 Boundary Twisting

Boundary twisting was introduced in [FQ90, Section 1.3]. Suppose we have two immersed surfaces A and B in a 4-manifold M such that part of the boundary of B is embedded in A, as shown in Figure 15.4(a), and this part of ∂B lies in the interior of M. For us, this situation usually arises when B is a Whitney disc pairing intersection points of A with itself

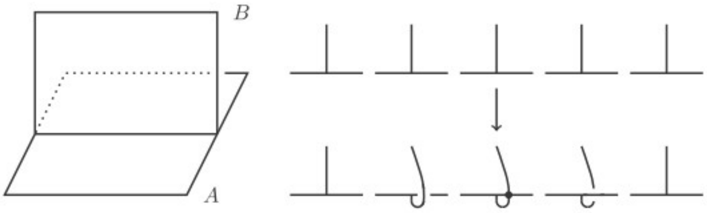

Figure 15.4 Boundary twisting. Left: Before boundary twisting. Right: Cross sections for the picture on the left, before and after boundary twisting. Each cross section shows a 3-dimensional time slice.

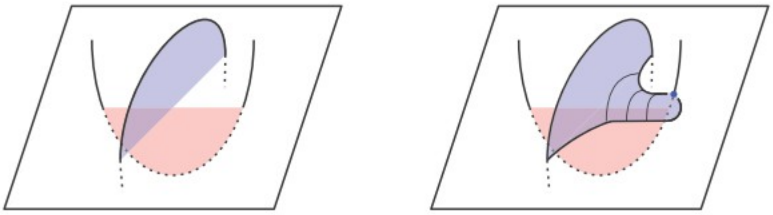

(a) (b)

Figure 15.5 Boundary push off to ensure Whitney circles are disjoint. We perform a regular homotopy of one Whitney circle until it becomes disjoint from the other, introducing a new intersection point for the corresponding Whitney disc.

or some other surface. Another commonly occurring situation is that of a cap of a grope or tower attached to a lower stage. The operation of *boundary twisting* B about A consists of changing a collar of B near a point in its boundary on A, as depicted in Figure 15.4(b). Note that this creates a new point of intersection between A and B and changes the framing of B by a full twist.

15.2.3 Making Whitney Circles Disjoint

We saw in Chapter 11 that, given two immersed surfaces A and B, if $\lambda(A, B) = 0$, the intersection points between A and B can be paired up with Whitney discs (Proposition 11.10). *A priori*, the corresponding Whitney circles may intersect one another. Performing the Whitney trick with intersecting Whitney circles leads to new intersections, which we would like to avoid. We ensure that Whitney circles are disjoint by pushing one Whitney circle along the other, as shown in Figure 15.5. This is a regular homotopy, and as we see in the figure, leads to new intersections between a Whitney disc and either A or B.

To achieve this in general for surfaces A and B, enumerate the Whitney arcs, and then work on the arcs in order. For the ith arc, push other arcs with index greater than i off the ith arc, starting with one of the arcs closest to the endpoint, until the ith arc is disjoint from all

$\sqrt{2} = 1,4142...$

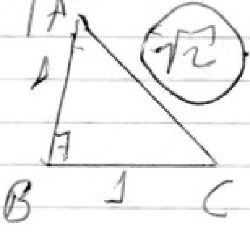

$\boxed{\sqrt{2}} \rightsquigarrow$ Γ 1,4142.

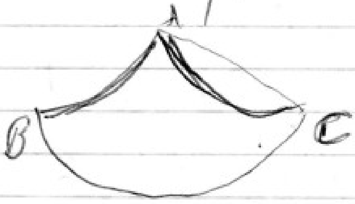

Books-018-2021/20210913/
The Disc Embedding Theorem/
The Disc Embedding Theorem...
(2021).pdf

Pág. 62(81) → Euclidean.

so we only need to show that it is locally Euclidean.

Let $E := J(S^{d-1} \times [0, \frac{1}{2}])$.

Pág. 92(111) → degree of the Gauss map $S^1 \times S^1 \to S^2$

Pág. 133(152) → Figure 10.1

A modification of the Cantor function. On the right, we see that certain vertical segments (red) are created.

—

Pág. 219(238) - Fig 15.3
Tubing multiple points of intersection.

—

20210914 III

Alexandre Augusto Bellei
CPF 541 656 00949

—

persmatematik.blogspot.com
desmos.com

—

They want to see you succeed, but never better than them.

—

Three-Dimensional Geometry and Topology
William P. Thurston

—

Gravitation Misner.

—

Books:

Data-NovoLivro-018-2021/
Books-018-2021/

% * "%
+H(%+://@ D+://@ +: C
0 0! +(0 ! I +

+H(%+://@ D+://@ +: C
0 0! +(0 ! I +2! +

+H(%+://@ D+://@ +: C
0 0! +(0 ! I +2! +2 3+

+H(%+://@ D+://@ +: C
0 0! +(0 ! I +2! +2 !6 ! +

+H(%+://@ D+://@ +: C
0 0! +(0 ! I +2! + +

+H(%+://@ D+://@ +: C
0 0! +(0 ! I +2 3 < 0 3+

! +,-.+ + +/ ! + % 2 3 2 3
J ; , ! ; ! # < B ! 0! 0#

! +,-.+ + +/ ! + % 2 3 < 0 3 !
= # 2 3 < B ! 0! ; 0#
* '& = "

!% +,-.+ + +/ ! + % 2 3 % <
! 0! ; % E # ! 2 3 ! # 6! 0#

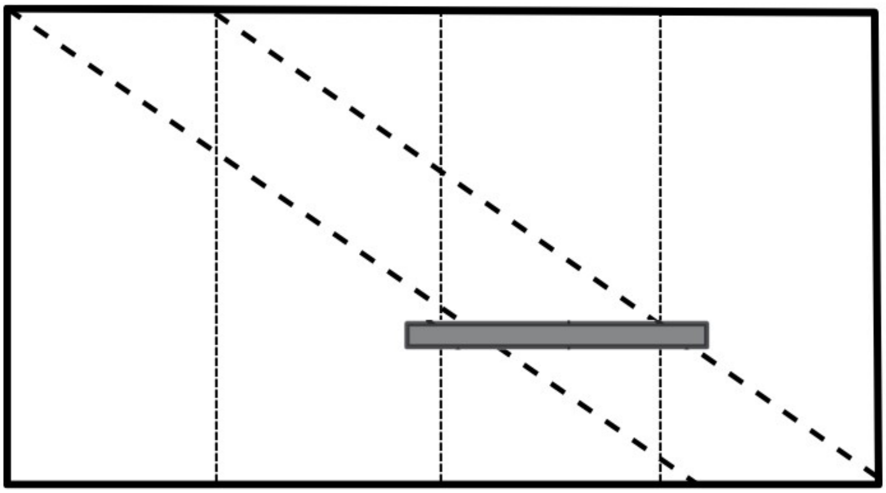

FIGURE 3. Blocking rays through a rectangle

We set $X_1 = \bigcup_{j=1}^{k_1} X_1(j)$ and proceed by induction. Assuming that X_i has been defined, find a collection of rectangles in each rectangle of X_i that block the bottom from the top, each of the new rectangles having diameter less than $1/(i+1)$. The union of these smaller rectangles is X_{i+1}.

The intersection $\bigcap_{i=0}^{\infty} X_i$ is an opaque Cantor set in the sense that every segment from the bottom of X_0 to the top must intersect the Cantor set.

6.3.2. A Fat Cantor Set. The total length of the segments removed from $[0, 1]$ in forming the middle-thirds Cantor set is

$$\frac{1}{3} + \left(\frac{1}{3}\right) \cdot \left(\frac{2}{3}\right) + \left(\frac{1}{3}\right) \cdot \left(\frac{2}{3} \cdot \frac{2}{3}\right) + \cdots = 1.$$

We conclude that the middle-thirds Cantor set has length $1 - 1 = 0$.

But the only critical condition required in the construction of the general Cantor set is that the pieces remaining at each stage have diameters approaching 0 in the limit. In order to satisfy that requirement, it suffices, for example, to make sure that each remaining piece have diameter less than one-half the diameter of the preceding pieces. Thus we may remove a total length as little as we please. As a result, there are Cantor sets in $[0, 1]$ with length as close to 1 as we choose.

6.3.3. A Wild Cantor Set. A Cantor set X in 3-space $\mathbb{R}^3$ is *tame* if there is a homeomorphism of $\mathbb{R}^3$ to itself that takes X to the middle-thirds Cantor set Y viewed as a subset of the x-axis. Otherwise X is called *wild*.

Although each Cantor set is 0-dimensional, there are Cantor sets in 3-space that are so complicated and tangled that one could form a necklace out of them that could be placed around the neck and would not fall off. The standard example X is, in fact, called *Antoine's Necklace*. The Cantor set X must be wild because it does not share the following property with the middle-thirds Cantor set Y.

EXERCISE 6.20. View the standard middle-thirds Cantor set Y as a subset of the x-axis in 3-space $\mathbb{R}^3$. Let $f : \mathbb{S}^1 \to \mathbb{R}^3 \setminus Y$. Then the map f can be extended to a map $F : \mathbb{B}^2 \to \mathbb{R}^3 \setminus Y$.

Here is a description of Antoine's necklace; see Figure 4:

We take as X_0 a solid torus (a solid doughnut). Inside of it, forming X_1, is a finite collection of smaller disjoint solid tori linked together in a chain that circles the interior of X_0. The set X_2 is formed by a collection of even smaller chains, one such chain circling the interior of each of the links making up X_1, and so forth. The necklace X is the set of points that lie in each of the smaller and smaller sets X_0, X_1, X_2, ...

Here is a property that distinguishes X from Y topologically:

Consider the circle C that goes through the hole in the doughnut X_0. Then C does not bound even a singular disk in $\mathbb{R}^3 \setminus X$. (Proving this fact is nontrivial.)

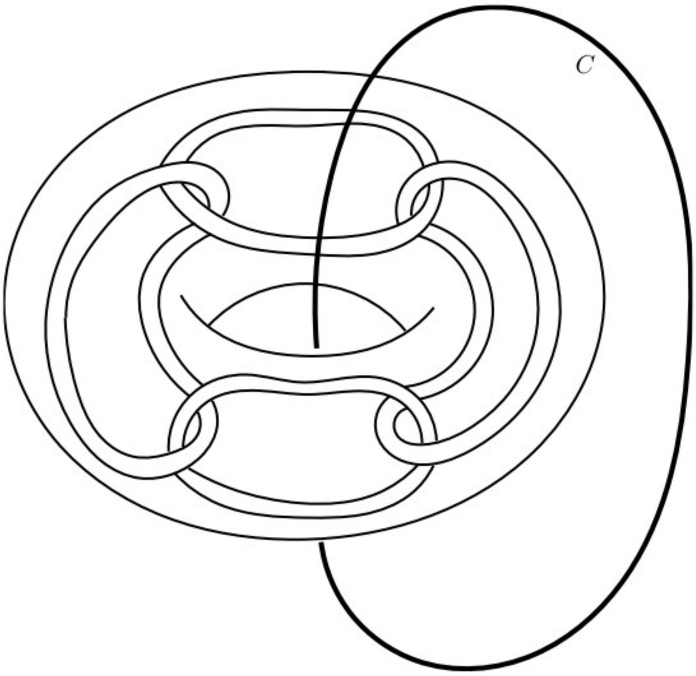

FIGURE 4. Antoine's Necklace

6.3.4. Hausdorff Dimension and a Fractal Cantor Set. Felix Hausdorff suggested another measure of dimension based on measure rather than topology. The middle-thirds Cantor set which we introduced in Definition 6.16 has, according to Hausdorff, dimension equal to $\log 2/\log 3 \approx .63092975...$ as opposed to covering dimension 0. Mandelbrot has defined a *fractal* as a set whose Hausdorff dimension is not equal to its Lebesgue covering dimension [86, 87, 88]. Since the covering dimension of the middle-thirds Cantor set is equal to 0, that set is, according to Mandelbrot's definition, a fractal.

Hausdorff observed that the volume of an n-dimensional cube is proportional to the nth power of its diameter and suggested that, for an arbitrary set X, the nth power of the diameter be taken as an approximation to the *n-dimensional volume* of X, regardless of the fine-scale shape of the object X. To refine the approximation, take an open (or closed) cover U of X by small sets and define the approximate n-dimensional volume $H^n(X, U)$ to be the sum of the nth powers $|u|^n$ of the diameters $|u|$ of the elements $u \in U$. Hausdorff noted that, in this manner it is possible to assign every subset of a separable metric space an n-dimensional volume for every n. Here are the details:

In order to make the n-dimensional volume of X well-defined, the covers of X by small open sets should be as efficient as possible. To that end, fix a positive number δ, consider all possible countable open covers U of X by sets of diameter $< \delta$, and define the δ-approximate n-dimensional volume $H^n_\delta(X)$ of X to be the infimum of the approximations $H^n(X, U)$. Finally, define the n-dimensional volume $H^n(X)$ of X by the formula

$$H^n(X) = \lim_{\delta \to 0} H^n_\delta(X).$$

There is nothing in this definition that requires that n be a nonnegative integer. Thus Hausdorff noted that one could assign an r-dimensional volume $H^r(X)$ to X for every nonnegative number r by adding rth powers of diameters.

Hausdorff proved as a consequence of the following theorem that, given the set X, there is a unique nonnegative number d such that $H^r(X) = \infty$ for $r < d$ and $H^r(X) = 0$ for $r > d$. He called this number d the *dimension* of X. We call it the *Hausdorff dimension* of X.

THEOREM 6.21. *Suppose that r and s are nonnegative real numbers such that $r < s$ and that $H^r(X) < \infty$. Then $H^s(X) = 0$.*

PROOF. Pick a positive number P such that $H^r(X) < P$. Then, for each $\eta > 0$, there is a countable open cover U_η of X by sets of diameter $< \eta$ such that $H^r(X, U_\eta) < P$. Thus, for each $\delta > 0$ and for each $\eta < \delta$,

$$H^s_\delta(X) \le H^s(X, U_\eta) = \sum\{|u|^s : u \in U_\eta\}$$

$$\le \eta^{s-r} \sum\{|u|^r : u \in U_\eta\} = \eta^{s-r} H^r(X, U_\eta) < \eta^{s-r} P.$$

Hence, for each $\delta > 0$, $H^s_\delta(X) \le \eta^{s-r} P$. It follows that $H^s_\delta(X) = 0$ and hence that $H^s(X) = 0$. □

Hausdorff gave interesting examples of sets whose (Hausdorff) dimension is not an integer. Here is one of his examples.

THEOREM 6.22. *The Hausdorff dimension of the middle-thirds Cantor set C is $r = \log 2/\log 3$. Furthermore, $H^r(C) = 1$.*

PROOF. We first show why it is natural to conjecture that $r = \log 2/\log 3$ is the Hausdorff dimension of C:

It is hard to imagine covers of C that are more efficient than the defining covers C_n, where $C = \bigcap_{n=1}^\infty C_n$ and C_n is the disjoint union of 2^n intervals of length $(1/3^n)$. The 2^n intervals of C_n define a closed cover U_n of C that, for the arbitrary dimension s, gives the following approximation to $H^s(C)$:

$$J^s(C, U_n) = \sum\{|u|^s \,:\, u \in U_n\} = \left(\frac{2}{3^s}\right)^n = e^{n(\log 2 - s\log 3)},$$

where we have used the identity $a^b = e^{b\log a}$.

If $\log 2 - s\log 3 = 0$, which implies $s = \log 2/\log 3$, we have $H^s(C, U_n) = e^{n\cdot 0} = 1$, which is neither ∞ nor 0. For $s < \log 2/\log 3$, $H^s(C, U_n) \to \infty$. For $x > \log 2/\log 3$, $H^s(C, U_n) \to 0$. We conclude that $s = (\log 2/\log 3 = r)$ is the most probable dimension of C In view of the previous theorem, it remains only to show that $H^r(C) = 1$.

The estimates that we have already made show that $H^r(C) \leq 1$. It remains to prove that $H^r(C) \geq 1$.

To that end, we suppose to the contrary that $H^r(C) < 1$. We fix a positive number δ, and consider an open cover U of C by sets of diameter $< \delta$ such that $H^r(C, U) < 1$. Since C is compact, we may assume that the cover U is finite. Without increasing the sum, we may assume that each element of U is an open interval. Expanding each element of U a tiny bit if necessary, we may assume that the end points of the interval lie in the complement of C. We replace these open intervals by their corresponding closed intervals, and proceed to modify this cover by closed intervals as follows.

If $I = [a, b]$ is one of the intervals, we may shorten I until its endpoints a and b actually lie in C. Let E denote the largest gap in C between the endpoints of I. This gap divides I into three subintervals P', E, and P'', with $a \in P'$ and $b \in P''$.

We need three facts:

(1) $3^r = 2$. (Take the logarithm of both sides.)

(2) $(3/2)(|P'| + |P''|) \leq |P'| + |E| + |P''|$. (This follows from the fact that $|P'|, |P''| \leq |E|$. In order to see that $|P'|, |P''| \leq |E|$, consider the last $n \geq 0$ such that the endpoints a and b of the interval I lie in the same interval J of C_n. Then E is the middle third of J, P' is a subinterval of the first third of J, and P'' is a subinterval of the last third of J.)

(3) $((|P'| + |P''|)/2)^r \geq (|P'|^r + |P''|^r)/2$. (This follows since $r < 1$ so that the function x^r is concave. See Figure 5)

From (1), (2), and (3) we deduce

$$|I|^r = (|P'| + |E| + |P''|)^r \geq (\frac{3}{2}(|P'| + |P''|))^r$$
$$= 2((|P'| + |P''|)/2)^r \geq 2(|P'|^r + |P''|^r)/2 = |P'|^r + |P''|^r.$$

Thus we may replace I by P' and P''.

A finite number of such operations can be used to reduce the covering to one by intervals of the same C_n. But at that stage, the approximate area is ≥ 1 by the calculation of the first half of the proof, contradicting our (false) assumption that the approximate area was < 1, a contradiction, as we wished to show. Hence $H^r(C, U) \geq 1$.

Thus $H^r(C) = 1$; and the proof is complete. □

6.3.5. The Cantor Function. There is a natural map from the Cantor set onto the unit interval. See an approximation to the graph of this function in Figure 7. This mapping is at the basis of the Cantor function: a continuous function that has slope 0 almost everywhere, yet increases in such a way that it maps the unit interval $[0, 1]$ *onto* itself.

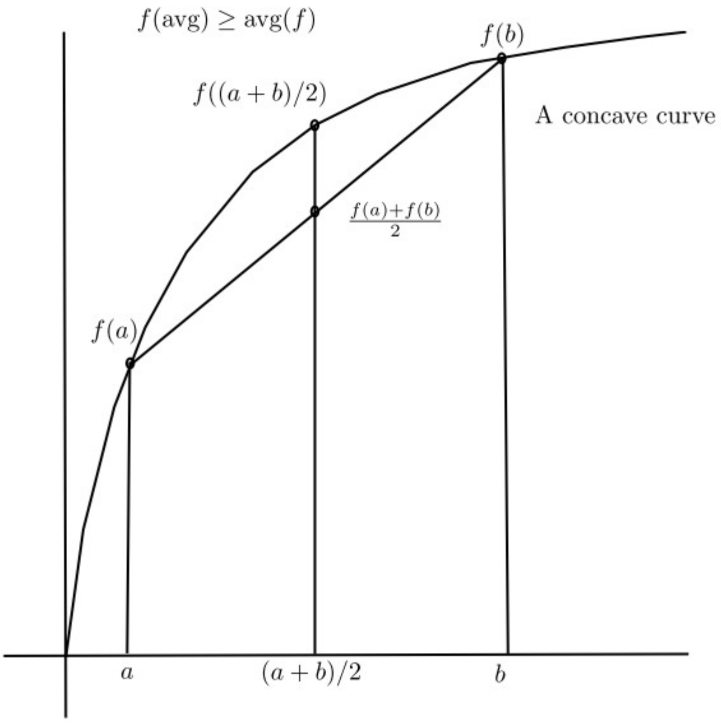

FIGURE 5. A concave curve

The idea of the construction is to create a monotone decomposition of the interval. Recall the definition of monotone decomposition in Section 5.2, Definition 5.8, and both Corollary 5.7 and Theorem 5.9 from that section. We describe the construction in generality:

We partition $[0,1]$ into a collection X of points and closed subintervals. See Figure 6. In the case of the Cantor set, the closed subintervals are the closures of the open intervals removed in defining the Cantor set. The elements of X are then equivalence classes of an equivalence relation on $[0,1]$. We assume that there are at least two, hence uncountably many, equivalence classes, most of which contain only a single point of $[0,1]$. At most countably many of the equivalence classes are closed intervals, since each contains an open set and there can be no more than countably many disjoint open sets.

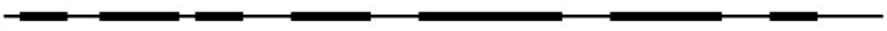

FIGURE 6. Monotone decomposition of an interval

We recall the *identification space* or *decomposition space* whose points are the elements of X. That is, each element of X, even if it consists of more than one point of $[0,1]$, is abstractly considered to be one point of X. We assigned a topology to X in the usual manner as follows. We considered the function $p : [0,1] \to X$

which assigned to each point $x \in [0,1]$ the equivalence class $p(x) = [x] \in X$ which contains x. We declared a subset $U \subset X$ to be an open subset of X iff the union $U^* = p^{-1}(U)$ of the elements of U formed an open subset of $[0,1]$. Note that, by this definition, the map $p : [0,1] \to X$ is continuous. We proved the following earlier.

THEOREM 6.23. *The decomposition space X is an arc.*

DEFINITION 6.24. Let I denote the set of closed intervals $[x_\alpha, y_\alpha]$ in $[0,1]$ that are closures of the open intervals deleted from $[0,1]$ in defining the middle-third Cantor set C. Let X denote the partition of $[0,1]$ into the elements of I and the singleton sets containing the points not included in any of the intervals of I. We assign X the identification-space topology. Let $p : [0,1] \to X$ denote the continuous function that assigns each point $x \in [0,1]$ the element $p(x) = [x] \in X$ which contains x. Note that X is an arc by the previous theorem. Hence we may identify X with $[0,1]$ and p with a map from $[0,1]$ to itself which fixes 0 and 1.

THEOREM 6.25. *Let I, $X = [0,1]$, and $p : [0,1] \to X = [0,1]$ be as in the previous definition. Then we may assume that $p(I)$ is the set of rational numbers in the open interval $(0,1)$.*

PROOF. The set $p(I)$ is a countable dense subset of $(0,1) \subset [0,1] = X$. Every countable dense subset of $(0,1)$ is embedded in $[0,1]$ in exactly the same way. □

COROLLARY 6.26. *The set of points of the middle-thirds Cantor set C that are not endpoints of any of the open intervals removed in forming C is homeomorphic to the set of irrational points between 0 and 1.*

THEOREM 6.27. *There is a (weakly) monotone continuous function $f : [0,1] \to [0,1]$, with $f(0) = 0$ and $f(1) = 1$ that has derivative 0 almost everywhere.*

PROOF. Map the first middle third to 1/2. Map the next larger intervals to 1/4 and 3/4. Map the four next intervals to 1/8, 3/8, 5/8, and 7/8, etc. Extend continuously to the remaining points of $[0,1]$. Since the sum of the lengths of the intervals defining the Cantor set is

$$\frac{1}{3} \cdot \left(1 + (2/3) + (2/3)^2 + \cdots + (2/3)^n + \cdots\right) = \frac{1}{3} \cdot \frac{1}{1-(2/3)} = 1,$$

the graph is horizontal on a set of measure 1 yet rises from 0 to 1.

□

6.4. Cantor Sets in the Plane Are Tame

We saw that there are Cantor sets in 3-dimensional space which are so entangled that they can act like a necklace and not fall off when placed around the neck. Are there Cantor sets in the plane with similar entanglement? We show in this section that there are no such Cantor sets in the plane. Not only are Cantor sets in the plane homeomorphic by definition to the standard middle-thirds Cantor set, but the homeomorphisms can be realized as the restrictions of homeomorphisms from the plane to itself. That is, every embedding of the Cantor set in the plane is equivalent to the standard embedding. We call all standard embeddings of an object *tame embeddings*.

!% +,-.+ + +/ ! + % 2 3 % <
! 0! ; % E # ! 2 3 ! # 6! 0#
* && ' !0 %

CHAPTER 3

Gravity As Curvature

We now turn to the description of curvature.

3.1. Einstein Identified Gravity with the Curvature of Space-Time

Albert Einstein identified gravity with the curvature of space-time: A massive object distorts, or curves, space-time; and other objects respond gravitationally by following (locally) shortest paths in that curved space, paths that deserve to be called straight. Misner, Thorne, and Wheeler, in their book "Gravitation" [97], compare the path followed by an object acting under the influence of gravity to a straight cut in the surface of an apple where this cut, near the stem of the apple, dips toward the stem as if that stem were a source of gravitational force.

Two problems must be solved in application of Einstein's ideas: (1) We usually recognize the curvature of an object by observing how it curves in some larger space (= extrinsic curvature). Since we live *within* space-time, how can we recognize and describe its curvature from within space-time itself (= intrinsic curvature)? (2) How do we recognize and describe the paths in a curved object that are locally shortest, hence deserve to be called *straight* (= geodesic)?

Gauss showed us how these problems can be solved for 2-dimensional surfaces. Riemann generalized Gaussian curvature to higher dimensions by considering the Gaussian curvature of each 2-dimensional slice in higher dimensional space.

In this introduction to curvature, we will restrict ourselves, as Gauss did, to 2-dimensional surfaces that reside in 3-dimensional Euclidean space. But we will not follow Gauss's path. For both questions, we will first approximate the surface by a polyhedral surface where it is easy to answer both questions: How flat is a polyhedral surface? How straight is an edge path in a polyhedral surface? The corresponding concepts in smooth surfaces will then follow by a limiting process.

CHAPTER 4

Curvature by Polyhedral Approximation

4.1. Approximating Smooth Surfaces by Polyhedra

Gauss described a smooth surface in $\mathbb{R}^3$ locally as the image of a map $S : \mathbb{R}^2 \to \mathbb{R}^3 : (x, y) \mapsto (x, y, z)$, where z is a many-times differentiable function of x and y. It is an easy matter to triangulate such a surface.

Key Idea 4.1. *Every smooth surface S in $\mathbb{R}^3$ can be closely approximated locally by polyhedral surfaces S''.*

We restrict our attention to piecewise smooth disks in the surface S.

Definition 4.2. A disk D in a smooth surface S is called *piecewise smooth* if there are finitely many points $p_0, p_1, \ldots, p_n = p_0$ in the boundary ∂D of D that divide ∂D into arcs $p_0p_1, p_1p_2, \ldots, p_{n-1}p_n$ each of which is smooth and such that there is a well-defined angle of turning in ∂D at each of the points $p_1, p_2, \ldots, p_n$.

Theorem 4.3. *If $D \subset \mathbb{R}^2$ is a disk that is mapped homeomorphically by $S : \mathbb{R}^2 \to \mathbb{R}^3$ to a piecewise smooth disk $S(D) \subset S$, then $S(D)$ can be closely approximated by polyhedral disks $D' \subset \mathbb{R}^3$.*

Proof. Choose points $q_0, q_1, \ldots, q_m = q_0$ in cyclic order on ∂D that divide ∂D into tiny arcs. Since D is piecewise smooth, the points q_i will be the vertices of a polygonal simple closed curve J in $\mathbb{R}^2$ that closely approximates ∂D. The curve J bounds a disk $D' \subset \mathbb{R}^2$ that approximates D. The disk D' can be triangulated with very small triangles T_j. If A, B, and C are the vertices of T_j, then S maps them to vertices $A' = S(A)$, $B' = S(B)$, and $C' = S(C)$. The vertices A', B', and C' are the vertices of a triangle $T'_j \subset \mathbb{R}^3$. The triangles T'_j are the triangles of a polyhedral disk D' that closely approximates $S(D)$. ☐

4.2. The Curvature of a Polyhedral Disk

In this chapter we define the curvature of a polyhedral disk in $\mathbb{R}^3$. We shall see that curvature is the average obstruction to flattening that disk into the plane locally without distorting lengths or areas.

4.3. How Flat Is a Disk?

We consider a polyhedral disk D, not assumed to be flat, triangulated by finitely many Euclidean triangles. Combinatorially, we can picture the disk as lying in the flat plane as in Figure 1, but to do this we may have to change angles and side lengths of the individual triangles.

If we fix our intention on one of the interior vertices v of our disk, then the triangles containing that vertex form a polyhedral cone with that vertex as tip of the cone. This polyhedral cone is called the *star* of the vertex v in D and is denoted

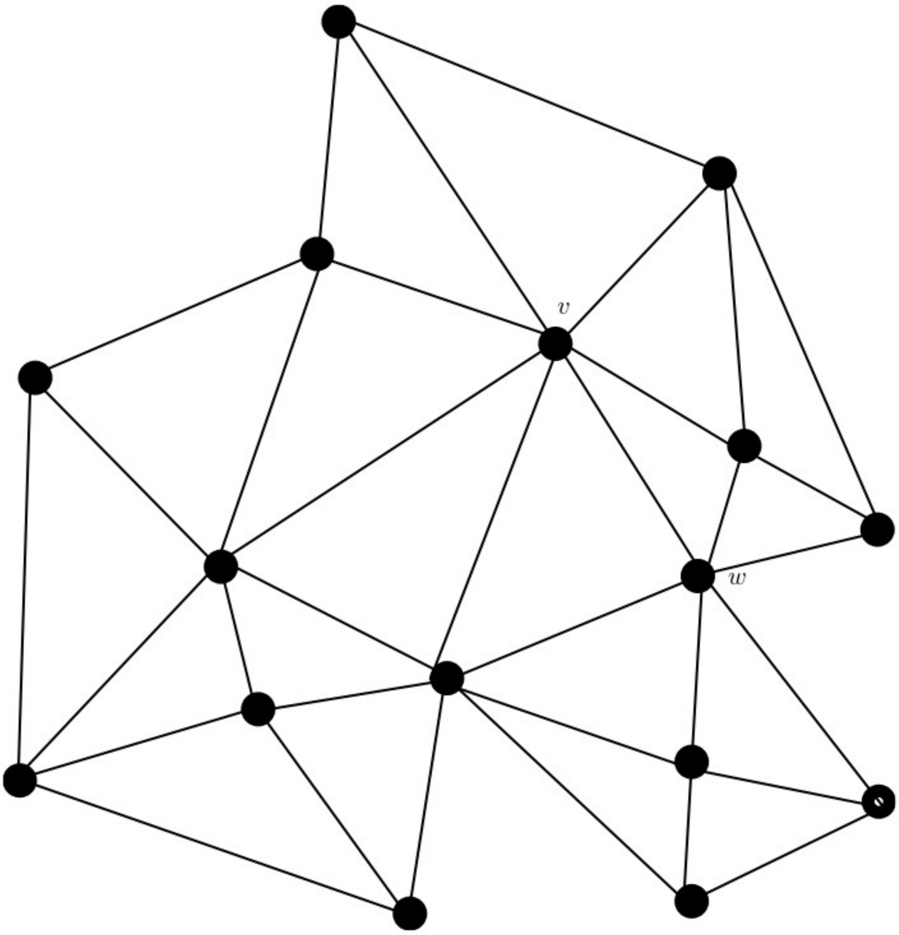

FIGURE 1. A triangulated disk in the flat plane

by star(v, D). This star can be drawn in the plane as a disk if we allow ourselves to distort lengths and angles (left side of Figure 2). If we cut star(v, D) along one of the edges emanating from v, then we can embed the triangles of the result in the plane, one by one, without distorting lengths or angles (right side of the Figure 2).

The cut star, as laid out in the plane, will close up naturally to form a planar disk if an only if the sum $i(v)$ of the angles at v is exactly 2π ($= 360$ degrees). We may take the difference $2\pi - i(v)$ as a measure of the degree to which the star of v in D fails to be flat. The angle $i(v)$ can potentially take on any value in the range $(0, \infty)$ so that $2\pi - i(v)$ can lie anywhere in the range $(-\infty, 2\pi)$. If the angle $i(v)$ is larger than 2π, then the undistorted triangles will spiral around the vertex v any number of times.

DEFINITION 4.4. Let $I(D)$ denote the set of interior vertices of D, and let $|I(D)|$ denote the number of those vertices. Let $i(v)$ denote the angle sum at each

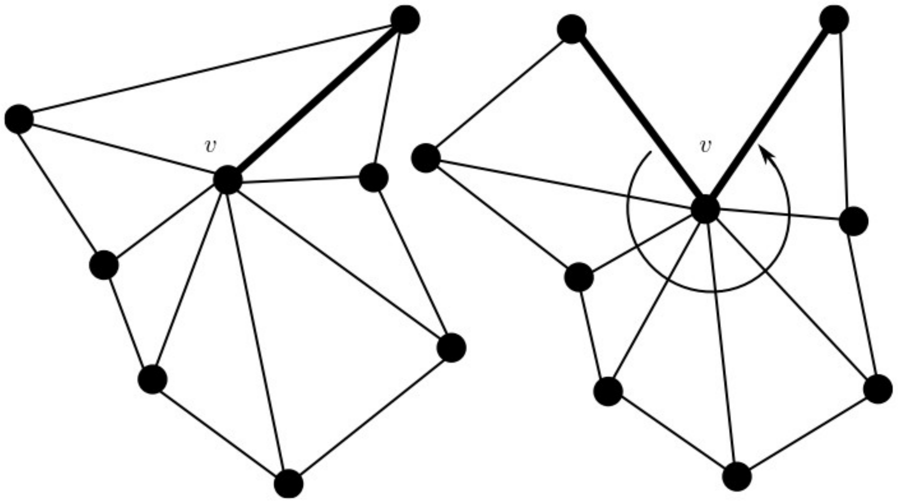

FIGURE 2. The star of an interior vertex

element $v \in I(D)$, and let $i(D) = \sum\{i(v) \mid v \in I(D)\}$. Then we may take the sum

$$\operatorname{def}(D) = \sum\{2\pi - i(v) \mid v \in I(D)\} = 2\pi \cdot |I(D)| - i(D)$$

as a global measure of the degree to which D fails to be flat. We call $\operatorname{def}(D)$ the *angle defect* of D.

4.4. How Straight Is a Disk Boundary?

We again consider a polyhedral disk D, not assumed to be flat, triangulated by finitely many Euclidean triangles. Combinatorially, we picture the disk as lying in the flat plane as before, but repeated in Figure 3. However, to do this we may have to change angles and side lengths of the individual triangles.

If we fix our attention on one of the boundary vertices w of our disk, then the triangles containing that vertex form a polyhedral cone with that vertex as tip of the cone. This polyhedral cone is called the *star* of the vertex w in D and is denoted by $\operatorname{star}(w, D)$. This star can be drawn in the plane as a disk with $\partial D \cap \operatorname{star}(w, D)$ straight if we allow ourselves to distort lengths and angles (left side of Figure 4). If we are careful to embed each triangle in the plane, one by one, without distorting lengths or angles (right side of Figure 4), the result will be a planar disk with $\partial D \cap \operatorname{star}(v, D)$ straight if and only if the sum $b(w)$ of the angles at w is exactly π ($= 180$ degrees).

We may take the difference $\pi - b(w)$ as a measure of the degree to which the boundary of D fails to be straight at w. The angle $b(w)$ can potentially take on any value in the range $(0, \infty)$ so that $\pi - b(w)$ can lie anywhere in the range $(-\infty, \pi)$. If the angle $b(w)$ is larger than 2π, then the undistorted triangles will spiral around the vertex w any number of times.

DEFINITION 4.5. Let $B(D)$ denote the set of boundary vertices of D, and let $|B(D)|$ denote the number of those vertices. Let $b(w)$ denote the angle sum at each element $w \in B(D)$, and let $b(D) = \sum\{b(w) \mid w \in B(D)\}$. Then we may take

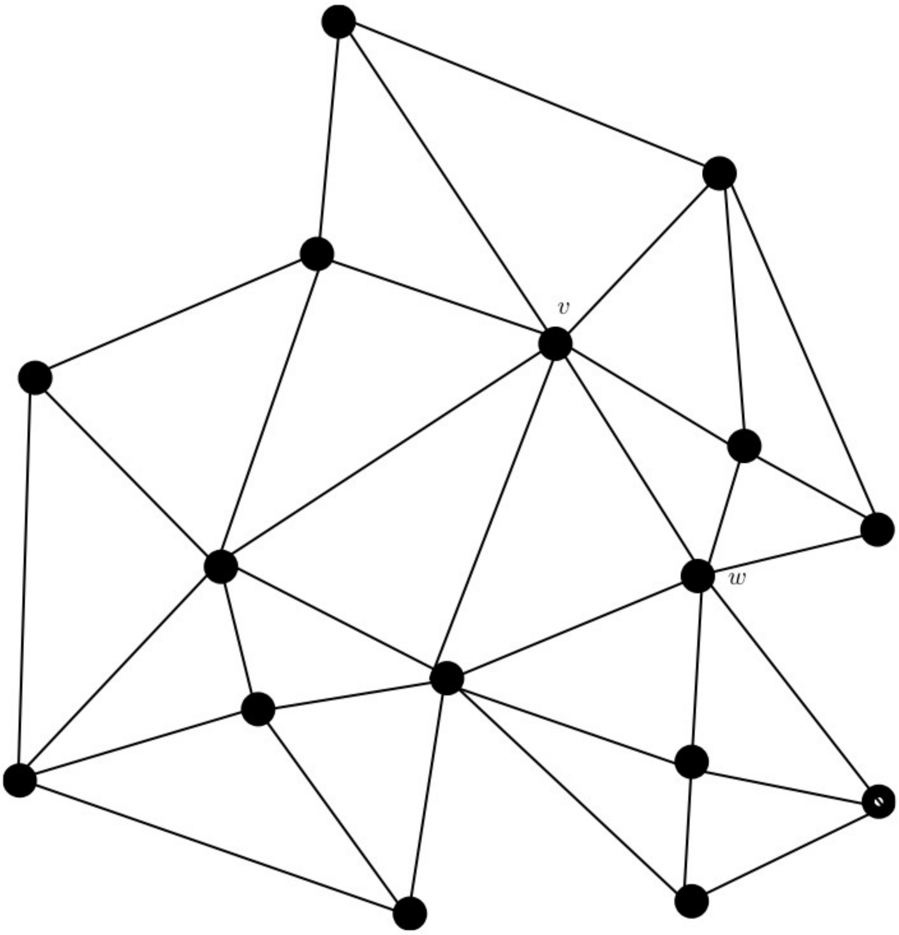

Figure 3. The triangulated disk again

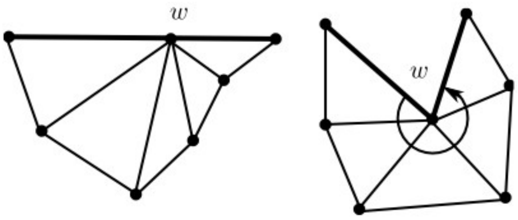

Figure 4. The star of a boundary vertex

the sum

$$\partial\mathrm{def}(D) = \sum\{\pi - b(w) \mid w \in B(D)\} = \pi \cdot |B(D)| - b(D)$$

as a global measure of the degree to which ∂D fails to be straight. (Of course, we do not expect a disk boundary to be straight.) We call $\partial\mathrm{def}(D)$ the *boundary defect* of D.

4.5. Duality Theorem: Angle Defect + Boundary Defect = 2π

As before, we assume that D is a polyhedral disk, not necessarily flat, triangulated by finitely many Euclidean triangles.

THEOREM 4.6 (Duality Theorem). *In a polyhedral disk D,*

$$\mathrm{def}(D) + \partial\mathrm{def}(D) = 2\pi.$$

REMARK. This easy but important theorem shows that angle defect and boundary defect are equally effective for measuring the degree to which a disk D fails to be flat. This duality between angle defect and boundary defect is very important. Angle defect gives a very intuitive picture of how difficult it is to flatten a surface into the plane, hence captures the notion of curvature. Boundary defect, on the other hand, captures the notion of how straight a path is and leads to beautiful derivative formulas by which we can study curvature.

A first consequence of this duality, as explained in a subsequent section, will be the famous theorem known as the Umlauf Theorem, Theorem 4.7, which, for simple closed polygonal curves in a planar surface, states that the boundary defect is always 2π since the angle defect of a planar surface is always 0. That is, as we travel along a planar simple closed curve, turning sometimes to the left and sometimes to the right, in the end we always turn precisely through a net angle of 2π.

A second consequence of the duality is a formula for the area of a disk in the 2-sphere $\mathbb{S}^2$ of radius 1. The area is precisely $2\pi - \partial\mathrm{def}$ (Theorem 4.9). This result will be an important ingredient in our proof of Gauss's *Theorema Egregium* (= *remarkable theorem*), Theorems 6.4 and 6.5, which shows that intrinsic and extrinsic views of curvature are equivalent. *End remark.*

PROOF OF THE DUALITY THEOREM, THEOREM 4.6. We review the notation of the preceding sections:

$I(D)$ is the set of interior vertices of D and $|I(D)|$ the number of such vertices.
$B(D)$ is the set of boundary vertices of D and $|B(D)|$ the number of such vertices.
$i(D)$ is the sum of all interior angles.
$b(D)$ is the sum of all boundary angles.
$\mathrm{def}(D) = |I(D)| \cdot 2\pi - i(D)$.
$\partial\mathrm{def}(D) = |B(D)| \cdot \pi - b(D)$.

We proceed by induction on the number T of triangles in D. See Figure 5.

If $T = 1$ so that D consists of a single triangle, D has no interior vertices, which implies that $\mathrm{def}(D) = 0$. On the other hand, $\partial\mathrm{def}(D) = 3 \cdot \pi - \pi = 2\pi$. Thus the theorem is true for $T = 1$.

We assume that $T > 1$ and that the theorem is true for all disks D with fewer than T triangles. If we take one triangle ABC with edge AC in the boundary of D, then at least one of the other edges, say AB is an interior edge of D. We consider two cases according to whether the vertex B is a boundary vertex or an interior vertex.

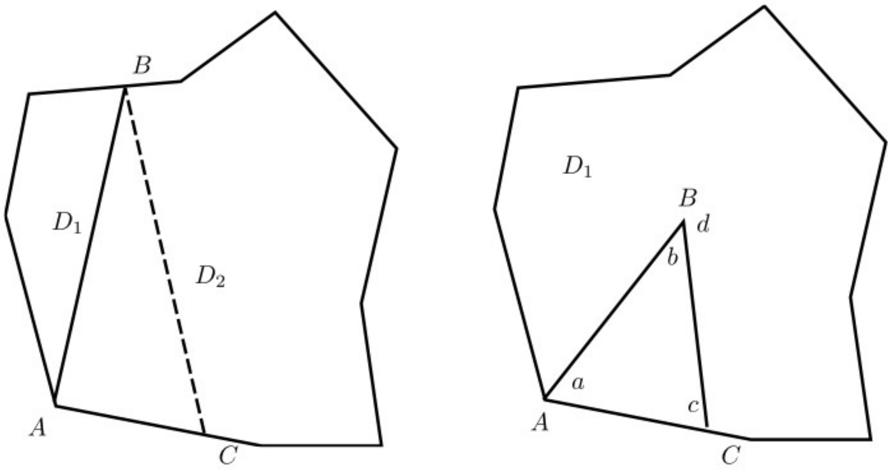

FIGURE 5. The two cases in the proof of the Duality Theorem

Case 1. Suppose that B is a boundary vertex of D. Then AB divides D into two subdisks D_1 and D_2 that share the edge AB. We calculate:

$$\begin{align*}
\operatorname{def}(D) + \partial\operatorname{def}(D) &= \Big(|I(D)| \cdot 2\pi - i(D)\Big) + \Big(|B(D)| \cdot \pi - b(D)\Big) \notag\\
&= \Big(|I(D_1)| + |I(D_2)|\Big) \cdot 2\pi - \Big(i(D_1) + i(D_2)\Big) \notag\\
&\quad + \Big(|B(D_1)| + |B(D_2)| - 2\Big) \cdot \pi - \Big(b(D_1) + b(D_2)\Big) \notag\\
&= \Big(\operatorname{def}(D_1) + \partial\operatorname{def}(D_1)\Big) + \Big(\operatorname{def}(D_2) + \partial\operatorname{def}(D_2)\Big) - 2\pi \notag\\
&= 2\pi + 2\pi - 2\pi \tag{4.1}\\
&= 2\pi. \notag
\end{align*}$$

Case 2. Suppose that B is an interior vertex of D. Then $AB \cup BC$ divides the triangle ABC with angles a, b, c at vertices A, B, C from a subdisk D_1 with angle d at B. We calculate:

$$\begin{align*}
\operatorname{def}(D) + \partial\operatorname{def}(D) &= \Big(|I(D)| \cdot 2\pi - i(D)\Big) + \Big(|B(D)| \cdot \pi - b(D)\Big) \notag\\
&= \Big(|I(D_1)| + 1\Big) \cdot 2\pi - \Big(i(D_1) + b + d\Big) \notag\\
&\quad + \Big(|B(D_1)| - 1\Big)\pi - \Big(b(D_1) + a + c - d\Big) \notag\\
&= \Big(\operatorname{def}(D_1) + \partial\operatorname{def}(D_1)\Big) + \big(2\pi - b - d - \pi - a - c + d\big) \notag\\
&= 2\pi + \Big(\pi - (a + b + c)\Big) \tag{4.2}\\
&= 2\pi. \notag
\end{align*}$$

□

4.6. The Curvature of a Polyhedral Disk

We have seen that both $\operatorname{def}(D)$ and $\partial\operatorname{def}(D)$ are reasonable measures of how difficult it is to flatten a polyhedral disk in the plane without distorting lengths and areas. Both of these measures are scale invariant. Hence we normalize by dividing by area. We define the *curvature* $K(D)$ of D to be the average

$$K(D) = \frac{\operatorname{def}(D)}{\operatorname{Area}(D)} = \frac{2\pi - \partial\operatorname{def}(D)}{\operatorname{Area}(D)}.$$

4.7. Applications of the Duality Between Angle Defect and Boundary Defect

In this section we use the duality between angle defect and boundary defect to trivially deduce a famous about the flat Euclidean plane which is known as the *Umlauf Theorem.* We then use the same duality to develop a formula for the area of a disk in the 2-sphere $\mathbb{S}^2$ of radius 1.

4.7.1. The Umlauf Theorem.

THEOREM 4.7 (Umlauf Theorem). *If D_0 is a polyhedral planar disk, then $\partial\operatorname{def}(D_0) = 2\pi$. That is, if one traverses the boundary of a planar disk, then the net angle, positive and negative, through which one turns is 2π. (In German, "laufen" means "to run" and "um" is a preposition meaning "around".)*

PROOF. Since D_0 is planar, it is obvious that the angle defect of D_0 is 0. Hence $2\pi = \operatorname{def}(D_0) + \partial\operatorname{def}(D_0) = \partial\operatorname{def}(D_0)$. □

4.7.2. The Area of a Disk in the 2-Sphere $\mathbb{S}^2$.

DEFINITION 4.8. A disk $D_0 \subset \mathbb{S}^2$ is said to be *spherically polyhedral* if its boundary is a finite union of segments lying in great circles of $\mathbb{S}^2$. The reason for calling its boundary polygonal is this: Given two points $x, y \in \mathbb{S}^2$, with $x \neq -y$, the shortest path between x and y on $\mathbb{S}^2$ is unique and lies on a great circle. That is, great circles deserve to be called *straight* on $\mathbb{S}^2$. We shall discuss this at some length later. A *spherical triangle* is a spherically polyhedral disk with three edges. A *spherical triangulation* of a spherically polyhedral disk is a union of spherical triangles, where each pair of triangles that intersect meet either in a vertex or a common edge.

REMARK. It is important to note that polyhedral and spherically polyhedral approximations to disks in $\mathbb{S}^2$ are closely related. Supposing that the vertices of a polyhedral approximation lie in $\mathbb{S}^2$, projection from the center of $\mathbb{S}^2$ will take the polyhedral approximation to a spherically polyhedral approximation. *End remark.*

THEOREM 4.9. *If $D_0 \subset \mathbb{S}^2$ is a disk that is spherically polyhedral, then* $\operatorname{Area}(D_0) = 2\pi - \partial\operatorname{def}(D_0)$.

PROOF. Since boundary defect measures the angles through which a path turns relative to a straight path, and since great circle segments on a sphere are straight, the boundary defect of a spherically polygonal path is simply the sum of the turning angles at the vertices between successive great-circle segments. Recall that the turning angle is the difference $\pi - b(w)$ between straight (π) and the angle sum ($b(w)$) in D_0 at boundary vertex w. We employ our previous notations:

$$\partial\mathrm{def}(D_0) = \sum_w \{\pi - b(w)\} = \pi \cdot |B(D_0)| - b(D_0),$$

where the sum is taken over the vertices w of the spherical polygonal boundary path, $b(w)$ is the angle sum of D_0 at w, $|B(D_0)|$ is the number of vertices w, and $b(D_0)$ is the sum of all the $b(w)$.

We assume that D_0 has a spherical triangulation. We proceed by induction on the number N of spherical triangles in this triangulation. The key case is that of a single triangle, where $N = 1$:

LEMMA 4.10 (The Triangle Lemma). *Suppose that* $\Delta = ABC \subset \mathbb{S}^2$ *is a triangle whose sides are great-circle segments on the* 2*-sphere* $\mathbb{S}^2$ *of radius* 1 *with angles* α*,* β*, and* γ *at the vertices* A*,* B*, and* C*, respectively. Then the area of* Δ *is given by the formula*

$$\mathrm{Area}(\Delta) = \alpha + \beta + \gamma - \pi.$$

The Triangle Lemma gives the desired result for $N = 1$ since, in that case, the turning angles at A, B, and C are, respectively, $\pi - \alpha$, $\pi - \beta$, and $\pi - \gamma$ and

$$2\pi - \partial\mathrm{def} = 2\pi - 3\pi + \alpha + \beta + \gamma = \alpha + \beta + \gamma - \pi.$$

We assume the Triangle Lemma for the moment and complete the induction on N as follows. If $N > 1$, then we assume the result inductively for fewer than N triangles. Exactly as in the proof of the Duality Theorem, we take one triangle ABC with edge AC in the boundary of D_0, where at least one of the other edges, say AB is an interior edge of D_0. We consider two cases acording to whether the vertex B is a boundary vertex or an interior vertex. See Figure 6.

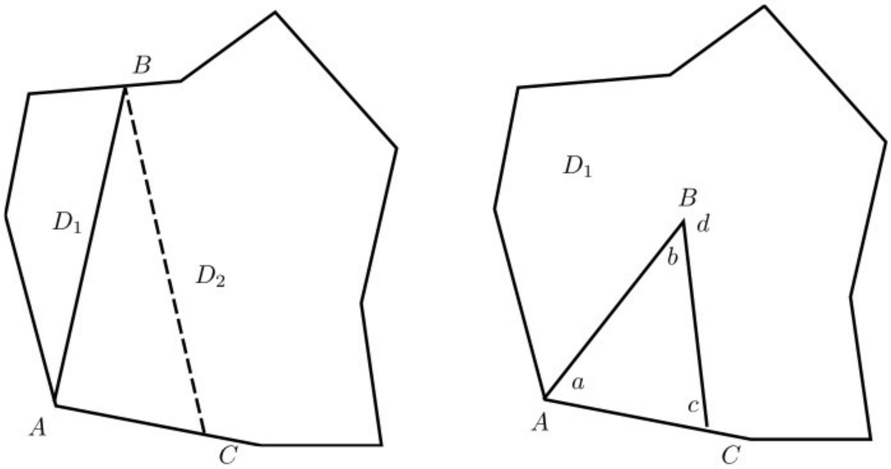

FIGURE 6. The two cases in the proof of the Triangle Lemma

Case 1. Suppose that B is a boundary vertex of D. Then AB divides D into two subdisks D_1 and D_2 that share the edge AB. We note that $|B(D_0)| =$

$|B(D_1)| + |B(D_2)| - 2$ and that $b(D_0) = b(D_1) + b(D_2)$. Hence, we calculate:

$$\begin{aligned} 2\pi - \big(\pi \cdot |B(D_0)| - b(D_0)\big) &= 2\pi - \pi B(D_1) - \pi B(D_2) + 2\pi + b(D_1) + b(D_2) \\ &= 2\pi - \big(\pi B(D_1) - b(D_1)\big) \\ &\quad + 2\pi - \big(\pi B(D_2) - b(D_2)\big) \\ &= \mathrm{Area}(D_1) + \mathrm{Area}(D_2) \\ &= \mathrm{Area}(D_0). \end{aligned}$$

Case 2. Suppose that B is an interior vertex of D_0. Then $AB \cup BC$ divides the triangle ABC with angles a, b, c at vertices A, B, C from a subdisk D_1 with angle d at B. We note that $|B(D_0)| = |B(D_1)| - 1$ and $b(D_0) = b(D_1) + a + c - d = b(D_1) + a + c + b - 2\pi$. (Note that $b + d = 2\pi$ since D_0 is smooth at B.) Hence, we calculate:

$$\begin{aligned} 2\pi - \big(\pi \cdot |B(D_0)| - b(D_0)\big) &= 2\pi - \pi|B(D_1)| + \pi + b(D_1) + a + b + c - 2\pi \\ &= 2\pi - (\pi \cdot |B(D_1)| - b(D_1)) + (a + b + c - \pi) \\ &= \mathrm{Area}(D_1) + \mathrm{Area}(ABC) \\ &= \mathrm{Area}(D_0). \end{aligned}$$

□

We turn now to the proof of the Triangle Lemma.

LEMMA 4.11. *If great circles I and J on $\mathbb{S}^2$ meet at angle α, then the union $I \cup J$ bounds two spherical digons on $\mathbb{S}^2$ of total area $(2\alpha/2\pi) \cdot 4\pi = 4\alpha$. (A spherical digon is a two-sided "polygon" bounded by portions of two great circles. See Figure 7.)*

PROOF. We lose no generality in assuming that I and J are longitudes that meet at the north pole N and the south pole S. In Figure 7, we view the northern hemisphere and the southern hemisphere separately, side by side, with the two digons formed from the shaded regions. It is obvious geometrically that the two shaded digons compose the fraction $(2\alpha/2\pi)$ of the total area of the sphere $\mathbb{S}^2$. Since the sphere $\mathbb{S}^2$ has area 4π, the desired result follows. □

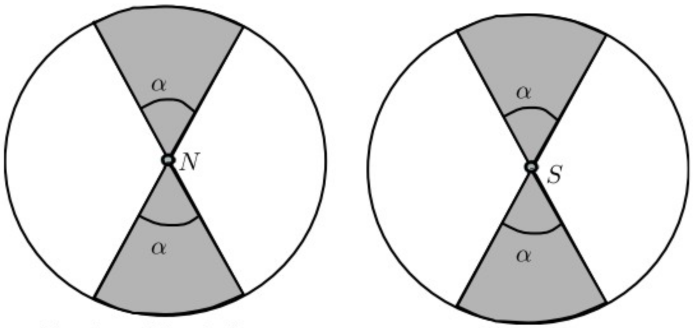

Northern Hemisphere Southern Hemisphere

FIGURE 7. The double lune between two great circles

Proof of the Triangle Lemma. The triangle Δ is bounded by portions of three great circles $I \supset AB$, $J \supset AC$, and $K \supset BC$. See Figure 8.

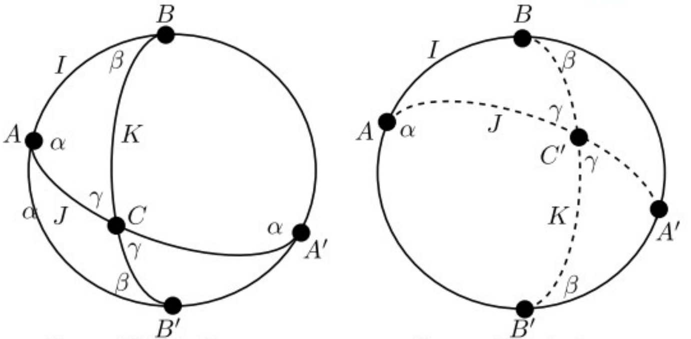

Upper Hemisphere Lower Hemisphere

Figure 8. Analyzing the spherical triangle

In the figure, we take I as bounding an upper hemisphere and a lower hemisphere, which we view side by side. We have labelled each of the resulting angles in the diagram that are given by α, β, and γ. Great circles I and J meet at A and its antipodal point $A' = -A$; I and K meet at B and its antipode $B' = -B$; J and K meet at C and its antipode $C' = -C$.

Let $D(\alpha)$ denote the union of the two digons with angles α bounded by $I \cup J$. Let $D(\beta)$ denote the union of those with angles β bounded by $I \cup K$. Let $D(\gamma)$ denote the union of those with angles γ bounded by $J \cup K$. Note that

$$\text{Area}(D(\alpha)) = 4\alpha; \quad \text{Area}(D(\beta)) = 4\beta; \quad \text{Area}(D(\gamma)) = 4\gamma$$

by the lemma. Then $D(\alpha)$, $D(\beta)$, and $D(\gamma)$ together cover each of the congruent triangles $\Delta = ABC$ and $\Delta' = A'B'C'$ three times and cover each of the complementary triangles exactly one time. Taking one of the copies of Δ and of Δ' and adding them to the complement, we see that $D(\alpha)$, $D(\beta)$, and $D(\gamma)$ cover the entire sphere $\mathbb{S}^2$ once, and each of Δ and Δ' twice. Thus

$$\text{Area}(D(\alpha)) + \text{Area}(D(\beta)) + \text{Area}(D(\gamma)) = \text{Area}(\mathbb{S}^2) + 2\text{Area}(\Delta) + 2\text{Area}(\Delta').$$

Consequently, since Δ and Δ' have the same area,

$$4\alpha + 4\beta + 4\gamma = 4\pi + 4\text{Area}(\Delta)$$

or

$$\text{Area}(\Delta) = \alpha + \beta + \gamma - \pi,$$

as claimed. □

4.8. The Curvature of a Smooth Disk

We define the angle defect of a smooth disk to be the limit of the angle defects of its polyhedral approximations. We define the boundary defect of a smooth disk to be the limit of the boundary defects of its polyhedral approximations.

We define the curvature $K(D_0)$ of a smooth disk D_0 as a limit of the curvatures of polyhedral approximations D:

$$K(D_0) = \lim_{D \to D_0} K(D),$$

where the limit is taken over polyhedral approximations D to D_0 and

$$K(D) = \frac{\operatorname{def}(D)}{\operatorname{Area}(D)} = \frac{2\pi - \partial\operatorname{def}(D)}{\operatorname{Area}(D)}.$$

We define curvature $K(S, p)$ of a smooth surface S at a point $p \in S$ by taking a second limit:

$$K(S, p) = \lim_{D_0 \to p} K(D_0).$$

REMARK. We have already developed enough information to see that these limits exist for the plane and the 2-sphere. That $K(D)$ exists for a smooth disk in a smooth surface will follow in a subsequent section when we develop an integral formula for boundary defect. Finally, that $K(S, p)$ exists will be a consequence of our proof of Gauss's famous *Theorema Egregium*, Theorems 6.4 and 6.5. *End remark.*

Using angle defect, with angle defect in the plane being 0, we have the immediate corollaries,

COROLLARY 4.12. *For every planar disk D_0 and every point p in the plane $\mathbb{R}^2$,*

$$K(D_0) = 0 \quad \textit{and} \quad K(\mathbb{R}^2, p) = 0.$$

COROLLARY 4.13. *The Umlauf Theorem is true not just for planar disks that are polyhedral but also for all planar disks.*

Using boundary defect and the area formula for spherical polyhedral disks D which gives $\operatorname{Area}(D) = 2\pi - \partial\operatorname{def}(D)$, we find

COROLLARY 4.14. *For every spherical disk D_0 and every point $p \in \mathbb{S}^2$,*

$$K(D_0) = 1 \quad \textit{and} \quad K(\mathbb{S}^2, p) = 1.$$

*   "&

and Curvature

Pág. 68 — 55(68)

Chapter 3 Gravity as Curvature

obs: Gravitation - Misner

Source: Importantíssimo/
Gauss/
Gravitation Misner/
– Gravitation....(1973).pdf
– Misner - Gravitation - Princeton University Press (2017).pdf

d) Lesson in Geometry - Vol 1....pdf
Jacques Hadamard

Source: Books-018-2021/
Importantíssimo/
Geometry-Topology/
Jacques-Hadamard-Leçons-de-Géométrie-élémentaire/
Jacques Hadamard - Leçons de géométrie élémentaire
I. pdf (322 pág).
II. pdf (602 pág).

1 +,-.+ + +/ ! + % 2 !6 ! > ? = !
) % 0#

1 +,-.+ + +/ ! + % 2 !6 !
5 6 3 0#

+,-.+ + +/ ! + % J 2 3; !
2 3 0# K F! 4 ?!#! ! #

+,-.+ + +/ ! + % F! 4 ?!#! ! # J L
M M M M ! (M M 0 ! , ! #) 0#
K ?!#! ! #

+,-.+ + +/ ! + % F! 4 ?!#! ! # J L
M M M M ! ((M M #! N 0! , ! #) 0#
K ?!#! ! #

+,-.+ + +/ ! + % B ! 2 3 ! #
! ! ! 3 ! (! # ; J 2 0 ! # ?! ! 3 0#

+,-.+ + +/ ! + % B ! 2 3 ! #
! ! ! 3 ! ((= 1 : # ; < 0 3 ! # 2! = # 0#
* " % 6! ! # ? 3

adapted to the decompositions (2.2.9) and (2.2.15) and calculate the image explicitly using (2.2.5).[17]

2.2.6 Prove that the mappings $\mathrm{pr}^{1,0} \circ \iota$ and $\mathrm{pr}^{0,1} \circ \iota$, defined by (2.2.13), are $\mathbb{C}$-linear and $\mathbb{C}$-anti-linear, respectively. Show that (2.2.14) holds.

2.2.7 Prove the second assertion in Corollary 2.2.15. *Hint.* Use point 2 of Proposition 2.2.14.

2.2.8 Prove that h defined by (2.2.27) is linear in the first and anti-linear in the second entry.

2.2.9 Prove formula (2.2.31).

2.2.10 Give an alternative proof of Proposition 2.2.24 by using (2.2.37) and (2.1.33).

2.2.11 Prove formula (2.2.47).

2.2.12 Prove formula (2.2.51). *Hint.* Prove that $\mathsf{g}((\nabla_X \mathsf{J})Y, Z) = \mathsf{g}(\nabla_X(\mathsf{J}Y), Z) + \mathsf{g}(\nabla_X Y, \mathsf{J}Z)$ and rewrite the terms on the right hand side according to (2.2.40). Use formula I/4.1.6. Alternatively, the proof can be found in [381], see Proposition 4.2 in Chap. IX.

2.2.13 Prove that for $H = \mathrm{Sp}(n, \mathbb{R})$, the cokernel of the mapping (2.2.2) is isomorphic to $\bigwedge^3(\mathbb{R}^{2n})^*$. Show that the corresponding intrinsic torsion coincides with the exterior derivative of the almost symplectic form, cf. Example 2.2.21.

2.3 Curvature and Holonomy

In this section, we continue the discussion of connections compatible with H-structures. Here, we consider exclusively torsion-free connections and ask which holonomy groups may occur for such a connection. This question has first been studied systematically by Berger, see [68, 69].

At this point, the reader may wish to recall the basic notions from the general holonomy theory as presented in Sect. 1.7. For a linear connection Γ in $L(M)$, let $P_{u_0}(\Gamma)$ be the holonomy bundle of Γ with base point $u_0 \in L(M)$. By Proposition 1.7.12, Γ is reducible to $P_{u_0}(\Gamma)$ and thus, for any $u \in P_{u_0}(\Gamma)$, the curvature Ω of Γ takes values in the Lie algebra $\mathfrak{h}_{u_0}(\Gamma)$ of the holonomy group $\mathscr{H}_{u_0}(\Gamma) \subset \mathrm{GL}(n, \mathbb{R})$. By the Ambrose-Singer Theorem 1.7.15, we have

$$\mathfrak{h}_{u_0}(\Gamma) = \mathrm{span}\left\{\Omega_u(X, Y) : u \in P_{u_0}(\Gamma)\,,\ \ X, Y \in \Gamma_u\right\}\,. \tag{2.3.1}$$

It is the condition of torsion-freeness which makes the above question nontrivial. If we drop this assumption, then any closed Lie subgroup $H \subset \mathrm{GL}(n, \mathbb{R})$ may occur as the holonomy group of a linear connection on some n-dimensional manifold M, see [283]. However, in general, such a connection will have a nontrivial torsion. By the Bianchi identity (2.1.17), vanishing of the torsion implies

---

[17] Cf. also Example 2.2.19.

$$\Omega \wedge \theta = 0\,, \tag{2.3.2}$$

and, by the Ambrose-Singer Theorem, this yields a nontrivial restriction on the holonomy. Now, let $P \subset L(M)$ be an H-structure on an n-dimensional manifold M, let ω be an H-compatible connection and let Ω be its curvature. For simplicity, let us denote $\mathbb{R}^n \equiv V$. By Remark 2.1.16, we may represent Ω equivalently by the curvature mapping

$$\mathscr{R} : P \to \bigwedge^2 V^* \otimes \mathfrak{h} \tag{2.3.3}$$

fulfilling the equivariance condition (2.1.25) with respect to the natural representation $\sigma : H \to \operatorname{Aut}\left(\bigwedge^2 V^* \otimes \mathfrak{h}\right)$ given by

$$\sigma_a((\xi \wedge \tau) \otimes A) := \left((a^{-1})^{\mathrm{T}}\xi \wedge (a^{-1})^{\mathrm{T}}\tau\right) \otimes \operatorname{Ad}(a)A\,. \tag{2.3.4}$$

Since the exterior products of the components θ^i of θ span the spaces of horizontal forms, (2.3.2) implies that $\mathscr{R}$ takes values in the kernel $\mathfrak{K}(\mathfrak{h})$ of the mapping

$$\delta : \bigwedge^2 V^* \otimes \mathfrak{h} \to \bigwedge^3 V^* \otimes V\,, \quad \delta = (a \otimes \mathrm{id}) \circ (\mathrm{id} \otimes \iota_{\mathfrak{h}})\,, \tag{2.3.5}$$

where a is the anti-symmetrization mapping, cf. (2.2.2). Clearly,

$$\mathfrak{K}(\mathfrak{h}) = \left\{F \in \bigwedge^2 V^* \otimes \mathfrak{h} : F(\mathbf{x}, \mathbf{y})\mathbf{z} + F(\mathbf{y}, \mathbf{z})\mathbf{x} + F(\mathbf{z}, \mathbf{x})\mathbf{y} = 0\,,\ \ \mathbf{x}, \mathbf{y}, \mathbf{z} \in V\right\}\,.$$

The space $\mathfrak{K}(\mathfrak{h})$ is called the space of curvature mappings.

Lemma 2.3.1 *The subspace*

$$\underline{\mathfrak{h}} := \operatorname{span}\left\{F(\mathbf{x}, \mathbf{y}) \in \mathfrak{h} : F \in \mathfrak{K}(\mathfrak{h})\,,\ \ \mathbf{x}, \mathbf{y} \in V\right\} \tag{2.3.6}$$

is an ideal of $\mathfrak{h}$.

Proof Let $F(\mathbf{x}, \mathbf{y}) \in \underline{\mathfrak{h}}$ and let $A \in \mathfrak{h} \subset \operatorname{End}(V)$. Then, we may write

$$[F(\mathbf{x}, \mathbf{y}), A] = \tilde{F}(\mathbf{x}, \mathbf{y}) - F(A\mathbf{x}, \mathbf{y}) - F(\mathbf{x}, A\mathbf{y})\,,$$

where

$$\tilde{F}(\mathbf{x}, \mathbf{y}) = [F(\mathbf{x}, \mathbf{y}), A] + F(A\mathbf{x}, \mathbf{y}) + F(\mathbf{x}, A\mathbf{y})\,.$$

One checks by direct inspection that $\tilde{F} \in \mathfrak{K}(\mathfrak{h})$. ∎

Note that $\tilde{F}$ corresponds exactly to the action of A on F obtained by differentiating the equivariance condition (2.1.25).[18] Thus, by the Ambrose-Singer Theorem, for the Lie algebra $\mathfrak{h}_{u_0}(\Gamma)$ of the holonomy group of a torsion-free connection Γ, we have

$$\underline{\mathfrak{h}_{u_0}(\Gamma)} = \mathfrak{h}_{u_0}(\Gamma)\,.$$

We conclude that a Lie subalgebra $\mathfrak{h} \subset \mathfrak{gl}(n, \mathbb{R})$ can occur as the Lie algebra of the holonomy group of a torsion-free connection only if it coincides with the ideal $\underline{\mathfrak{h}}$. This is commonly referred to as the first criterion of Berger. It yields a necessary condition for a Lie subalgebra to be the holonomy Lie algebra of a torsion-free connection.

Next, let us analyze the Bianchi identity (2.1.16) in terms of $\mathscr{R}$. The covariant derivative $D\mathscr{R} = \mathrm{d}\mathscr{R} + \sigma'(\omega)\mathscr{R}$ is a horizontal 1-form on P with values in $\mathfrak{K}(\mathfrak{h})$.

Definition 2.3.2 A torsion-free connection fulfilling $D\mathscr{R} = 0$ is called locally symmetric.

Decomposing $D\mathscr{R}$ with respect to the horizontal frame $\{\theta^i\}$, we obtain a function $D\mathscr{R} : P \to V^* \otimes \mathfrak{K}(\mathfrak{h})$. Using the fact that the commutators of horizontal standard vector fields corresponding to a torsion-free connection are vertical (Exercise 2.3.1), we calculate

$$\begin{aligned} D\Omega(B(\mathbf{x}), B(\mathbf{y}), B(\mathbf{z})) &= \mathrm{d}\Omega(B(\mathbf{x}), B(\mathbf{y}), B(\mathbf{z})) \\ &= B(\mathbf{x})\big(\Omega(B(\mathbf{y}), B(\mathbf{z})\big) - \Omega([B(\mathbf{x}), B(\mathbf{y})], B(\mathbf{z})) + \text{cycl.} \\ &= \mathrm{d}\mathscr{R}(B(\mathbf{x}))(\mathbf{y} \wedge \mathbf{z}) + \text{cycl.} \\ &= D\mathscr{R}(\mathbf{x})(\mathbf{y} \wedge \mathbf{z}) + \text{cycl.}\ . \end{aligned}$$

Thus, by the Bianchi identity $D\Omega = 0$, we conclude that the function $D\mathscr{R}$ takes values in the kernel of the mapping

$$\delta' : V^* \otimes \mathfrak{K}(\mathfrak{h}) \to \textstyle\bigwedge^3 V^* \otimes \mathfrak{h}\,, \tag{2.3.7}$$

defined as the composition

$$V^* \otimes \mathfrak{K}(\mathfrak{h}) \to V^* \otimes \textstyle\bigwedge^2 V^* \otimes \mathfrak{h} \to \textstyle\bigwedge^3 V^* \otimes \mathfrak{h}$$

of the inclusion and the anti-symmetrization mappings. Clearly, the kernel of δ' is

$$\mathfrak{K}^1(\mathfrak{h}) := \left\{\Phi \in V^* \otimes \mathfrak{K}(\mathfrak{h}) : \Phi(\mathbf{x})(\mathbf{y}, \mathbf{z}) + \Phi(\mathbf{y})(\mathbf{z}, \mathbf{x}) + \Phi(\mathbf{z})(\mathbf{x}, \mathbf{y}) = 0\,,\ \ \mathbf{x}, \mathbf{y}, \mathbf{z} \in V\right\}\,.$$

Thus, if $\mathfrak{h}$ is the holonomy Lie algebra of a torsion-free linear connection that is not locally symmetric, then necessarily $\mathfrak{K}^1(\mathfrak{h}) \neq 0$. This is usually referred to as the second Berger criterion.

---

[18] Clearly, this is the action of the Killing vector field generated by A.

Definition 2.3.3 A Lie subalgebra $\mathfrak{h} \subset \mathrm{End}(V)$ is called a Berger algebra if $\underline{\mathfrak{h}} = \mathfrak{h}$. A Berger algebra is called symmetric if $\mathfrak{K}^1(\mathfrak{h}) = 0$ and non-symmetric otherwise. Correspondingly, a Lie subgroup $H \subset \mathrm{Aut}(V)$ is referred to as a (symmetric or non-symmetric) Berger group if its Lie algebra is a (symmetric or non-symmetric) Berger algebra.

By the above discussion, we have the following.

Proposition 2.3.4 (Berger) *Let $\mathfrak{h} \subset \mathrm{End}(V)$ be a Lie subalgebra. Then,*

1. *If $\mathfrak{h}$ is the Lie algebra of the holonomy group of a torsion-free connection on some manifold, then $\mathfrak{h}$ is a Berger algebra.*
2. *If $\mathfrak{K}^1(\mathfrak{h}) = 0$, then any torsion-free connection on a manifold whose holonomy Lie algebra is contained in $\mathfrak{h}$ must be locally symmetric.*

Based upon these criteria, Berger started to tackle the above classification problem. It is natural to distinguish between the following two classes:

(a) Lie subalgebras $\mathfrak{h}$ lying in some $\mathfrak{o}(\eta)$, where η is some non-degenerate bilinear form on V. In this case, the associated H-structure defines a pseudo-Riemannian manifold. Therefore, this is called the metric case.
(b) Lie subalgebras which are not contained in any orthogonal Lie algebra. This is called the non-metric case.

Within this general analysis, Berger obtained a list of candidates for Lie subalgebras of type (a) and also an (incomplete) list for type (b).[19] These lists where refined and completed by the work of Alekseevski [14], Bryant [108, 109], Chi [132], Merkulov and Schwachhöfer [569]. The final full classification of irreducible holonomies of torsion-free affine connections was obtained by Merkulov and Schwachhöfer [441]. For an exhaustive discussion, we refer to the reviews of Bryant [110] and Schwachhöfer [570] and to the textbooks of Besse [76], Joyce [353] and Salamon [555]. In [110], the reader can find the complete classification list (divided into four parts) together with a lot of information on methods for proving that a given group in the list really occurs as a holonomy. It turns out that every such group is realized at least locally.[20]

In the remainder of this section, we exclusively consider the metric case. That is, we consider (pseudo-)Riemannian manifolds (M, g), endowed with their unique torsion-free metric connection (the Levi-Civita connection). Under this assumption, the frame bundle reduces to the orthonormal frame bundle $O(M)$ and the whole theory may be described in terms of objects living on $O(M)$. Consequently, in the case under consideration, the holonomy group is a subgroup of the structure group $\mathrm{O}(k, l)$. If the Levi-Civita connection is locally symmetric, we call (M, g) locally symmetric.

---

[19] The list provided by Theorem 2.3.19 below is included in type (a).

[20] The appropriate method working for three of the above mentioned four tables is to describe torsion-free connections with a given holonomy as solutions to an exterior differential system and to apply Cartan's existence theorem.

Definition 2.3.5 Let (M, g) be a pseudo-Riemannian manifold. The curvature mapping

$$\mathscr{R} : O(M) \to \bigwedge^2 V^* \otimes \mathfrak{o}(k, l)$$

of the Levi-Civita connection of g is called the Riemann curvature mapping. Correspondingly, the curvature tensor R is called the Riemann curvature of (M, g).

Comparing with the general case, $\mathscr{R}$ has some additional properties coming from the fact that we may use the metric η to identify V with V^*. In particular, $\mathfrak{o}(k, l) \cong \bigwedge^2 V^*$, and thus

$$\mathscr{R}(u) \in \bigwedge^2 V^* \otimes \bigwedge^2 V^* , \tag{2.3.8}$$

for every $u \in O(M)$.

Proposition 2.3.6 *The Riemann curvature mapping* $\mathscr{R}$ *of a pseudo-Riemannian manifold has the following algebraic properties. For any* $\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w} \in V$,

$$\mathscr{R}(\mathbf{x}, \mathbf{y}) = -\mathscr{R}(\mathbf{y}, \mathbf{x}) , \tag{2.3.9}$$

$$\eta(\mathscr{R}(\mathbf{x}, \mathbf{y})\mathbf{z}, \mathbf{w}) = -\eta(\mathscr{R}(\mathbf{x}, \mathbf{y})\mathbf{w}, \mathbf{z}) , \tag{2.3.10}$$

$$\eta(\mathscr{R}(\mathbf{x}, \mathbf{y})\mathbf{z}, \mathbf{w}) = \eta(\mathscr{R}(\mathbf{z}, \mathbf{w})\mathbf{x}, \mathbf{y}) , \tag{2.3.11}$$

$$\mathscr{R}(\mathbf{x}, \mathbf{y})\mathbf{z} + \mathscr{R}(\mathbf{y}, \mathbf{z})\mathbf{x} + \mathscr{R}(\mathbf{z}, \mathbf{x})\mathbf{y} = 0 . \tag{2.3.12}$$

Proof Formulae (2.3.9) and (2.3.10) follow immediately from (2.3.8) and formula (2.3.12) is a direct consequence of the fact that $\mathscr{R}$ takes values in the kernel $\mathfrak{K}(\mathfrak{h})$ of the mapping (2.3.5). It remains to prove (2.3.11). For that purpose, we write down the following four versions of (2.3.12).

$$\begin{aligned}
0 &= \eta(\mathscr{R}(\mathbf{x}, \mathbf{y})\mathbf{z}, \mathbf{w}) + \eta(\mathscr{R}(\mathbf{y}, \mathbf{z})\mathbf{x}, \mathbf{w}) + \eta(\mathscr{R}(\mathbf{z}, \mathbf{x})\mathbf{y}, \mathbf{w}) , \\
0 &= \eta(\mathscr{R}(\mathbf{y}, \mathbf{z})\mathbf{w}, \mathbf{x}) + \eta(\mathscr{R}(\mathbf{z}, \mathbf{w})\mathbf{y}, \mathbf{x}) + \eta(\mathscr{R}(\mathbf{w}, \mathbf{y})\mathbf{z}, \mathbf{x}) , \\
0 &= -\eta(\mathscr{R}(\mathbf{z}, \mathbf{w})\mathbf{x}, \mathbf{y}) - \eta(\mathscr{R}(\mathbf{w}, \mathbf{x})\mathbf{z}, \mathbf{y}) - \eta(\mathscr{R}(\mathbf{x}, \mathbf{z})\mathbf{w}, \mathbf{y}) , \\
0 &= -\eta(\mathscr{R}(\mathbf{w}, \mathbf{x})\mathbf{y}, \mathbf{z}) - \eta(\mathscr{R}(\mathbf{x}, \mathbf{y})\mathbf{w}, \mathbf{z}) - \eta(\mathscr{R}(\mathbf{y}, \mathbf{w})\mathbf{x}, \mathbf{z}) .
\end{aligned}$$

Summation of these equations and using (2.3.9) and (2.3.10) yields the assertion. ■

Remark 2.3.7

1. By Proposition 2.3.6,

   $$\mathscr{R} : O(M) \to S^2\left(\bigwedge^2 V^*\right) , \tag{2.3.13}$$

   where $S^2\left(\bigwedge^2 V^*\right) = \bigwedge^2 V^* \overset{s}{\otimes} \bigwedge^2 V^*$ is the symmetrized tensor product. By (2.1.25), $\mathscr{R}$ has the following equivariance property, see Exercise 2.3.2,

   $$\mathscr{R}(\Psi_a(u))(\mathbf{x}, \mathbf{y}, \mathbf{u}, \mathbf{v}) = \mathscr{R}(u)(a\mathbf{x}, a\mathbf{y}, a\mathbf{u}, a\mathbf{v}) , \tag{2.3.14}$$

for $a \in \mathrm{O}(k,l)$ and $\mathbf{x}, \mathbf{y}, \mathbf{u}, \mathbf{v} \in V$.

2. By (2.1.27), the Riemann curvature R fulfils identities corresponding to (2.3.9)–(2.3.12) with $\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w} \in V$ replaced by $X, Y, Z, W \in \mathrm{T}_m M$ and η replaced by g. Thus, in particular, $\mathsf{R} \in \Gamma^\infty\big(S^2(\bigwedge^2 \mathrm{T}^*M)\big)$. For a local frame $\{e_i\}$, using (2.1.52) we write
$$\mathsf{R}_{ijkl} \equiv \mathsf{g}(\mathsf{R}(e_i, e_j)e_k, e_l) = \mathsf{R}^m{}_{ijk}\, \mathsf{g}_{ml} \,.$$

In this notation, the algebraic properties (2.3.9)–(2.3.12) read
$$\mathsf{R}_{ijkl} = -\mathsf{R}_{jikl}\,, \quad \mathsf{R}_{ijkl} = -\mathsf{R}_{ijlk}\,, \quad \mathsf{R}_{ijkl} = \mathsf{R}_{klij}\,, \tag{2.3.15}$$
$$\mathsf{R}_{ijkl} + \mathsf{R}_{jkil} + \mathsf{R}_{kijl} = 0 \tag{2.3.16}$$

◆

Using the above properties, the space of Riemann curvature mappings $\mathfrak{K}(\mathfrak{o}(k,l))$ may be characterized as follows. By standard representation theory of the group $\mathrm{O}(k,l)$, for $n \geq 4$, one obtains the following decompositions into $\mathrm{O}(k,l)$-irreducible modules [76, 555]:
$$\textstyle\bigwedge^3 V^* \otimes V^* = \bigwedge^2 V^* \oplus \bigwedge^4 V^* \oplus U\,, \tag{2.3.17}$$
$$S^2\left(\textstyle\bigwedge^2 V^*\right) = \mathbb{R} \oplus \Sigma_0^2 \oplus \textstyle\bigwedge^4 V^* \oplus W\,, \tag{2.3.18}$$

where Σ_0^2 stands for the space of traceless endomorphisms of $\mathbb{R}^n$ (viewed as symmetric 2-tensors) and where U and W are orthogonal complements. By dimension counting, U and W are not isomorphic.

Proposition 2.3.8 *The space of Riemann curvature mappings is given by*
$$\mathfrak{K}(\mathfrak{o}(k,l)) = \ker\varphi \,\cap\, S^2\left(\textstyle\bigwedge^2 V^*\right)\,, \tag{2.3.19}$$

where
$$\varphi : \textstyle\bigwedge^2 V^* \otimes \bigwedge^2 V^* \to \bigwedge^4 V^*\,, \quad \varphi(\xi \otimes \tau) := \xi \wedge \tau\,. \tag{2.3.20}$$

Proof Under the identifications $\mathfrak{o}(k,l) \cong \bigwedge^2 V^*$ and $V \cong V^*$, $\mathfrak{K}(\mathfrak{o}(k,l))$ coincides with the kernel of the mapping $\chi : \bigwedge^2 V^* \otimes \bigwedge^2 V^* \to \bigwedge^3 V^* \otimes V^*$ given by
$$\chi(\alpha \otimes (\zeta \wedge \tau)) := (\alpha \wedge \zeta) \otimes \tau - (\alpha \wedge \tau) \otimes \zeta\,.$$

Now, consider the decompositions (2.3.17) and (2.3.18). Viewing χ as an $\mathrm{O}(k,l)$-intertwining mapping and using Schur's Lemma, together with the fact that χ is surjective, we conclude that χ must be zero on the irreducible subspaces $\mathbb{R}$, Σ_0^2 and W. By dimension counting, these subspaces span the kernel of χ. Moreover, restricted to $S^2(\bigwedge^2 V^*)$, χ maps onto $\bigwedge^4 V^*$ and coincides with φ. ■

Combining (2.3.19) and (2.3.18), for $n \geq 4$, we obtain[21]

$$\mathfrak{K}(\mathfrak{o}(k,l)) = \mathbb{R} \oplus \Sigma_0^2 \oplus W\,. \tag{2.3.21}$$

This yields a decomposition of the Riemann curvature into its irreducible components with respect to the action of $\mathrm{O}(k,l)$. The component Σ_0^2 corresponds to the contraction to $V^* \otimes V^*$ defined by taking the trace of the mapping $\mathbf{z} \mapsto \mathscr{R}(\mathbf{z},\mathbf{x})\mathbf{y}$ and restricting it to $S^2(V^*)$.

Definition 2.3.9 (*Ricci tensor*) Let (M,g) be a pseudo-Riemannian manifold and let $\mathscr{R}$ be its Riemann curvature mapping. The mapping

$$\widetilde{\mathsf{Ric}} : O(M) \to S^2(V^*)\,, \quad \widetilde{\mathsf{Ric}}(u)(\mathbf{x},\mathbf{y}) := \operatorname{tr}\{\mathbf{z} \mapsto \mathscr{R}(u)(\mathbf{z},\mathbf{x})\mathbf{y}\} \tag{2.3.22}$$

is called the Ricci curvature mapping. Correspondingly,

$$\mathsf{Ric} : \mathrm{T}_m M \times \mathrm{T}_m M \to \mathbb{R}\,, \quad \mathsf{Ric}(X,Y) := \operatorname{tr}\{Z \mapsto \mathsf{R}(Z,X)Y\} \tag{2.3.23}$$

is called the Ricci tensor of (M,g).

Note that Ric is of the same geometric type as the metric. Thus, viewing it as a mapping $\mathrm{T}_m M \to \mathrm{T}_m^* M$ and using $\mathsf{g}^{-1} : \mathrm{T}_m^* M \to \mathrm{T}_m M$, we can define a scalar on M.

Definition 2.3.10 (*Scalar curvature*) Let (M,g) be a pseudo-Riemannian manifold and let Ric be its Ricci tensor. The function

$$\mathsf{Sc} : M \to \mathbb{R}\,, \quad \mathsf{Sc}(m) := \operatorname{tr}(\mathsf{g}^{-1} \circ \mathsf{Ric})(m) \tag{2.3.24}$$

is called the scalar curvature of (M,g). The corresponding equivariant function $\widetilde{\mathsf{Sc}} : O(M) \to \mathbb{R}$ is called the scalar curvature mapping.

The scalar curvature corresponds to the first component in the decomposition (2.3.21). The component corresponding to the third summand is called the Weyl tensor. In Sect. 2.8, the above decomposition will be discussed in detail for the case $n = 4$.

Remark 2.3.11 Denoting $\mathsf{R}_{ij} = \mathsf{Ric}(e_i, e_j)$, we obtain the following local expressions for the Ricci tensor and the scalar curvature,

$$\mathsf{R}_{ij} = \mathsf{g}^{kl}\,\mathsf{R}_{kijl}\,, \quad \mathsf{Sc} = \mathsf{g}^{ij}\,\mathsf{R}_{ij}\,. \tag{2.3.25}$$

In particular, for a holonomic frame, we obtain

$$\mathsf{R}_{ij} = \partial_i\,\Gamma^l{}_{jl} - \partial_j\,\Gamma^l{}_{il} + \Gamma^l{}_{jm}\Gamma^m{}_{il} - \Gamma^l{}_{im}\Gamma^m{}_{jl}\,. \tag{2.3.26}$$

---

[21] For $k + l = 3$, one obtains $\mathfrak{K}(\mathfrak{o}(k,l)) = \mathbb{R} \oplus \Sigma_0^2$. For $k + l = 4$, this result belongs to Singer and Thorpe [592].

For an orthonormal local frame, we have $\mathsf{R}_{ij} = \eta^{kl}\,\mathsf{R}_{kijl}$. This yields the following useful formula

$$\mathsf{Ric}(X, Y) = \eta^{kl}\mathsf{g}(\mathsf{R}(e_k, X)Y, e_l)\,, \quad X, Y \in \mathfrak{X}(M)\,. \tag{2.3.27}$$

◆

There is an important special class of Riemannian manifolds characterized by the fact that their curvature has a trivial Σ_0^2-component in the decomposition (2.3.21).

Definition 2.3.12 (*Einstein manifold*) A (pseudo-)Riemannian manifold (M, g) is called Einstein if its Ricci tensor is a constant multiple of the metric at each point of M.

Note that for an n-dimensional Einstein space (M, g) we have

$$\mathsf{Ric} = \frac{\mathsf{Sc}}{n}\,\mathsf{g}\,, \tag{2.3.28}$$

where Sc is constant. In Sect. 2.5, we will see a large class of Einstein manifolds.

In the next step, we show which impact the above additional structures have on the analysis of the Berger criteria in the metric case. For a chosen orthonormal frame $u_0 \in P_{u_0}(\Gamma)$, let us consider the holonomy bundle $P_{u_0}(\Gamma) \subset O(M)$. Let us denote

$$H = \mathscr{H}_{u_0}(\Gamma)\,, \quad \mathfrak{h} = \mathfrak{h}_{u_0}(\Gamma)\,.$$

On $P_{u_0}(\Gamma)$, the curvature takes values in $\mathfrak{h} \subset \mathfrak{o}(k, l) \cong \bigwedge^2(V^*)$. This fact, together with (2.3.19), implies the following.

Proposition 2.3.13 *For any point $u \in P_{u_0}(\Gamma)$, the Riemann curvature $\mathscr{R}(u)$ belongs to the space*

$$\mathfrak{K}(\mathfrak{h}) = \ker\varphi \,\cap\, S^2(\mathfrak{h})\,. \tag{2.3.29}$$

■

It turns out that for many subgroups $H \subset \mathrm{O}(k, l)$, the restriction of φ to $S^2(\mathfrak{h})$ is injective. This implies $\mathfrak{K}(\mathfrak{h}) = 0$ and, thus, $\underline{\mathfrak{h}} = 0$. Then, the first Berger criterion implies that, in this case, H cannot occur as a holonomy group.

In the same way, the covariant derivative $D\mathscr{R}$ may be dealt with. By the above discussion, we have the following.

Proposition 2.3.14 *For any point $u \in P_{u_0}(\Gamma)$, the covariant derivative of $\mathscr{R}(u)$ takes values in*

$$\mathfrak{K}^1(\mathfrak{h}) = \ker\delta' \,\cap\, \big(V^* \otimes \mathfrak{K}(\mathfrak{h})\big)\,, \tag{2.3.30}$$

where $\delta' : V^* \otimes \mathfrak{K}^1(\mathfrak{o}(k, l)) \to \bigwedge^3 V^* \times \mathfrak{o}(k, l)$, cf. formula (2.3.7). ■

As already mentioned above, the condition $\mathfrak{K}^1(\mathfrak{h}) = 0$ distinguishes a special class of possible candidates. By Proposition 2.3.4, in this case the Riemannian manifold is necessarily locally symmetric. We exclude this class of spaces for a while, postponing its presentation to Sect. 2.5.

Finally, we show that we may limit our attention to the case where the representation of the holonomy group H on $V \equiv \mathbb{R}^n$ is irreducible. We consider the Riemannian metric case and comment on the pseudo-Riemannian case at the end. Under this assumption, the holonomy group is a subgroup of $\mathrm{O}(n)$. Let us assume, on the contrary, that the representation of H is reducible, that is, there exists a proper subspace $W \subset V$ invariant under H. Since we assume that η be definite, there exists an invariant orthogonal complement $W^\perp \subset V$. Proceeding further in this manner, we obtain an invariant orthogonal decomposition

$$V = W_0 \oplus W_1 \oplus \ldots \oplus W_k \,, \tag{2.3.31}$$

with W_0 carrying the trivial representation[22] (acting as the identity) and W_k carrying nontrivial irreducible representations of H for all $k \geq 1$. The following theorem belongs to de Rham [150]. It simplifies the holonomy classification problem essentially.

Theorem 2.3.15 (de Rham Splitting Theorem) *Let* (M, g) *be a Riemannian manifold. If the holonomy group* H *acts reducibly on* $\mathbb{R}^n$*, then the restricted holonomy group*[23] H^0 *of* (M, g) *is isomorphic to a product,*

$$H^0 = \{e\} \times H_1 \times \ldots \times H_k \,,$$

and M *is locally isomorphic to a product of Riemannian manifolds,*

$$M_0 \times M_1 \times \ldots \times M_k \,,$$

with M_0 *being flat.*

Proof By the above discussion, $\mathscr{R} : O(M) \to \bigwedge^2 V^* \otimes \mathfrak{o}(n)$ and $\mathscr{R}(u)(\mathbf{x}, \mathbf{y})$ takes values in $\mathfrak{h} \equiv \mathfrak{h}_{u_0}(\Gamma)$, for any $u \in P_{u_0}(\Gamma)$ and any $\mathbf{x}, \mathbf{y} \in V$. Since the decomposition (2.3.31) is invariant, we have

$$\mathscr{R}(u)(\mathbf{x}, \mathbf{y})_{\restriction W_0} = 0 \,, \quad \mathscr{R}(u)(\mathbf{x}, \mathbf{y})_{\restriction W_i} \subset W_i \,, \tag{2.3.32}$$

for $1 \leq i \leq k$. We decompose $\mathbf{x} = \sum \mathbf{x}_i$ and $\mathbf{y} = \sum \mathbf{y}_i$ with respect to (2.3.31) and insert this decomposition into $\mathscr{R}(u)(\mathbf{x}, \mathbf{y})$. This yields

$$\mathscr{R}(u)(\mathbf{x}, \mathbf{y}) = \sum_i \mathscr{R}(u)(\mathbf{x}_i, \mathbf{y}_i) + \sum_{i \neq j} \mathscr{R}(u)(\mathbf{x}_i, \mathbf{y}_j) \,.$$

---

[22] Clearly, W_0 may be zero.

[23] By Theorem 1.7.9, this is the identity connected component of H.

By (2.3.12) and (2.3.32), we have $\mathscr{R}(u)(W_i, W_j)W_k = 0$ for i, j and k pairwise distinct. Next, consider the case $i = j \neq k$. Then, again by (2.3.12),

$$\mathscr{R}(u)(\mathbf{x}_i, \mathbf{y}_i)\mathbf{z}_k = 0 \,, \quad \mathscr{R}(u)(\mathbf{y}_i, \mathbf{z}_k)\mathbf{x}_i = -\mathscr{R}(u)(\mathbf{z}_k, \mathbf{x}_i)\mathbf{y}_i \,.$$

The first of these equations implies $\mathscr{R}(u)(W_i, W_i)W_k = 0$ for $i \neq k$. Using (2.3.11), from the second equation we obtain

$$\eta\big(\mathscr{R}(u)(\mathbf{z}_k, \mathbf{x}_i)\mathbf{y}_i, \mathbf{x}_i\big) = \eta\big(\mathscr{R}(u)(\mathbf{z}_k, \mathbf{y}_i)\mathbf{x}_i, \mathbf{x}_i\big) = \eta\big(\mathscr{R}(u)(\mathbf{x}_i, \mathbf{x}_i)\mathbf{z}_k, \mathbf{y}_i\big) \,,$$

and the anti-symmetry of $\mathscr{R}$ implies $\mathscr{R}(u)(W_k, W_i)W_i = 0$ for $i \neq k$. We conclude

$$\mathscr{R}(u)(\mathbf{x}, \mathbf{y}) = \sum_i \mathscr{R}(u)(\mathbf{x}_i, \mathbf{y}_i) \,.$$

Now, according to the equivariance of $\mathscr{R}$, as u ranges over $\pi^{-1}(m) \cap P_{u_0}(\Gamma)$ and $\mathbf{x}$, $\mathbf{y}$ over V, for every i, the mappings $\mathscr{R}(u)(\mathbf{x}_i, \mathbf{y}_i)$ span an ideal $\mathfrak{h}_i(m) \subset \mathrm{End}(W_i)$ of $\mathfrak{h}$. Finally, varying m yields ideals $\mathfrak{h}_i$ and, by (2.3.1), the decomposition

$$\mathfrak{h} = \mathfrak{h}_1 \oplus \ldots \oplus \mathfrak{h}_k \,.$$

This proves the first assertion. To prove the second assertion, first note that the splitting (2.3.31) induces a splitting of the horizontal distribution Γ on $P_{u_0}(\Gamma)$,

$$\Gamma = \Gamma_1 \oplus \ldots \oplus \Gamma_k \,, \quad \Gamma_i := \Gamma \cap \theta^{-1}(W_i) \,.$$

By H-equivariance, this splitting induces a family of distributions $D_i = \pi'(\Gamma_i)$ on M such that

$$\mathrm{T}M = D_1 \oplus \ldots \oplus D_k \,.$$

Moreover, corresponding to (2.3.31), let us decompose

$$\theta = \theta_1 + \ldots + \theta_k \,, \quad \omega = \omega_1 + \ldots + \omega_k \,, \quad \Omega = \Omega_1 + \ldots + \Omega_k \,,$$

with $\theta_i \in \Omega^1(P_{u_0}(\Gamma)) \otimes W_i$ and ω_i , $\Omega_i \in \Omega^*(P_{u_0}(\Gamma)) \otimes \mathfrak{h}_i$. We define the distributions

$$\hat{\Gamma}_i := \Gamma_i \oplus V_i$$

on $P_{u_0}(\Gamma)$, with V_i being the vertical distribution spanned by the Killing vector fields generated from elements of $\mathfrak{h}_i$. Clearly, Γ_i is spanned by the horizontal standard vector fields generated by any basis of W_i. Thus, $\hat{\Gamma}_i$ annihilates both θ_j, ω_j, and Ω_j for any $j \neq i$ and, by point 2 of Remark 2.1.14 and (1.4.5), for every i the distribution $\hat{\Gamma}_i$ is involutive. Consequently, by the Frobenius Theorem, it is integrable and, for every i, we have

$$\mathrm{d}\theta_i + \omega_i \wedge \theta_i = 0\,, \quad \Omega_i = \mathrm{d}\omega_i + \frac{1}{2}[\omega_i, \omega_i]\,. \qquad (2.3.33)$$

Let $P_i \subset P_{u_0}(\Gamma)$ be an integral manifold of $\hat{\Gamma}_i$. Integrability of $\hat{\Gamma}_i$ clearly induces integrability of D_i and the integral manifolds U_i of D_i fulfil $U_i = \pi(P_i) \subset M$. Moreover, for every i, the restriction $\pi_i : P_i \to U_i$ of π defines a principal H_i-bundle and, by (2.3.33), ω_i is a torsion-free connection on P_i with restricted holonomy group H_i.

To summarize, for every $m \in M$, there exists a neighbourhood $U \cong U_1 \times \ldots \times U_k$ of m in M, with the U_i being integral manifolds of D_i, and the Levi-Civita connection restricted to U being a product of the Levi-Civita connections on the components U_i. ■

Definition 2.3.16 A Riemannian manifold (M, g) which is locally isomorphic to a product of Riemannian manifolds is called locally reducible. It is called irreducible if it is not locally reducible.

Clearly, by Theorem 2.3.15, if (M, g) is irreducible, then the restricted holonomy group necessarily acts irreducibly. Under additional assumptions, de Rham [150] was able to prove the following global version of Theorem 2.3.15.

Theorem 2.3.17 (Global de Rham Splitting Theorem) *Let (M, g) be a geodesically complete simply connected Riemannian manifold and assume that the holonomy group*[24] *of the Levi-Civita connection acts reducibly. Then, (M, g) is the direct product of geodesically complete simply connected irreducible Riemannian manifolds (M_i, g_i),*

$$(M, \mathsf{g}) = (M_0, \mathsf{g}_0) \times (M_1, \mathsf{g}_1) \times \ldots \times (M_k, \mathsf{g}_k)\,.$$

Here, (M_0, g_0) is a Euclidean vector space whose dimension is possibly zero. ■

Remark 2.3.18 Both versions of the de Rham Splitting Theorem have been extended to the case of an indefinite metric by Wu [682, 683]. ♦

Summarizing our discussion, for finding the possible holonomy groups of a Riemannian manifold (M, g), it is reasonable to make the following assumptions:

(a) M is simply connected. This ensures that the holonomy group is connected and that it coincides with the restricted holonomy group.
(b) (M, g) is irreducible. This implies that the holonomy group acts irreducibly.
(c) (M, g) is not locally symmetric. This requires $\mathfrak{K}^1(\mathfrak{h}) \neq 0$.

Under these assumptions, for the Riemannian case, Berger obtained the following.

Theorem 2.3.19 (Berger) *Let (M, g) be an n-dimensional simply connected irreducible Riemannian manifold which is not locally symmetric. Then, its holonomy group H belongs to one of the following classes:*

---

[24] By Remark 1.7.11, if M is simply connected, then the holonomy group and the restricted holonomy group coincide.

1. $H = \mathrm{SO}(n),\ n \geq 2$, *(generic Riemannian manifold)*
2. $H = \mathrm{U}(m),\ n = 2m \geq 4$, *(generic Kähler manifold)*
3. $H = \mathrm{SU}(m),\ n = 2m \geq 4$, *(special Kähler manifold)*
4. $H = \mathrm{Sp}(m) \cdot \mathrm{Sp}(1),\ n = 4m \geq 8$, *(quaternionic Kähler manifold)*
5. $H = \mathrm{Sp}(m),\ n = 4m \geq 8$, *(Hyper-Kähler manifold)*
6. $H = G_2,\ n = 7$, *(special holonomy)*
7. $H = \mathrm{Spin}(7),\ n = 8$, *(special holonomy).* ■

For the proof, which is beyond the scope of this book, we refer to [68, 69, 555].

Remark 2.3.20

1. An elegant proof of Theorem 2.3.19 is obtained from the following result of Simons [591]: if M is irreducible, then either the holonomy group H acts transitively on S^{n-1} or its identity component acts trivially on the space of curvature tensors $\mathfrak{K}(\mathfrak{h})$. Then, Theorem 2.3.19 is obtained by using the classification of simple Lie algebras and their representations.
2. According to Examples 2.2.22 and 2.2.27, it was clear from the beginning that the groups $\mathrm{SO}(n)$ and $\mathrm{U}(n)$ must occur in the above list. For a detailed discussion of examples for all the groups occuring in Theorem 2.3.19, we refer to [555]. ♦

Exercises

2.3.1 Show that the commutators of horizontal standard vector fields corresponding to a torsion-free connection are vertical.

2.3.2 Confirm the equivariance property (2.3.14). *Hint:* Under the identification $\mathfrak{o}(n) \cong (\mathbb{R}^n)^* \wedge (\mathbb{R}^n)^*$, the adjoint representation is mapped onto the second exterior power of the dual of the basic representation.

2.3.3 Show that, in terms of the Riemann curvature R, the Bianchi identity (2.1.16) reads

$$(\nabla_X \mathsf{R})(Y, Z) + (\nabla_Y \mathsf{R})(Z, X) + (\nabla_Z \mathsf{R})(X, Y) = 0\,. \tag{2.3.34}$$

2.4 Sectional Curvature

In this section, we discuss a generalization of the classical Gaussian curvature of surfaces in $\mathbb{R}^3$. It reduces the study of the Riemann curvature to the study of real valued functions. Let (M, g) be a pseudo-Riemannian manifold. Let $\Sigma_m \subset \mathrm{T}_m M$ be a 2-dimensional subspace such that $\mathsf{g}_{\restriction \Sigma_m}$ is non-degenerate. Let $\{X, Y\}$ be an arbitrary basis of Σ_m. We put

$$\mathsf{K}(\Sigma_m) := \frac{\langle \mathsf{R}(X, Y)Y, X\rangle}{\| X \|^2 \| Y \|^2 - \langle X, Y\rangle^2}\,, \tag{2.4.1}$$

where $\| \cdot \|^2$ and $\langle \cdot, \cdot \rangle$ are the quadratic form and the bilinear form, respectively, induced from g. It can be easily shown that $\mathsf{K}(\Sigma_p)$ is well defined, that is,

(a) the right hand side of (2.4.1) does not depend on the choice of the basis. This is a simple consequence of the symmetry properties of R given by point 2 of Remark 2.3.7 and is, thus, left to the reader (Exercise 2.4.1).
(b) Σ_m is non-degenerate iff $\| X \|^2 \| Y \|^2 - \langle X, Y \rangle^2 \neq 0$, (Exercise 2.4.2).

Note that K may be viewed as a mapping from the Graßmann manifold $G_2(\mathrm{T}_m M)$ to $\mathbb{R}$. Let $G_2^0(\mathrm{T}_m M) \subset G_2(\mathrm{T}_m M)$ be the subset of non-degenerate subspaces.

Definition 2.4.1 The mapping $\mathsf{K} : G_2^0(\mathrm{T}_m M) \to \mathbb{R}$ given by (2.4.1) is called the sectional curvature of the pseudo-Riemannian manifold at $m \in M$.

Clearly, in the Riemannian case, every 2-dimensional subspace of $\mathrm{T}_m M$ is non-degenerate.

Proposition 2.4.2 *The curvature tensor* R *is completely determined by the sectional curvature. If the mapping* K *is constant, that is,* $\mathsf{K}(\Sigma_m) = k(m)$ *for every* $\Sigma_m \in G_2^0(\mathrm{T}_m M)$, *then*

$$\mathsf{R}_m(X, Y)Z = k(m)\big(\langle Y, Z \rangle X - \langle X, Z \rangle Y\big) . \tag{2.4.2}$$

Conversely, if (2.4.2) *is fulfilled, then all non-degenerate planes have sectional curvature* $k(m)$.

Proof The proof of the first assertion is the consequence of the following simple polarization argument. Denote $\alpha(X, Y) := \langle \mathsf{R}(X, Y)X, Y \rangle$, for any $X, Y \in \mathrm{T}_m M$. Then, by direct inspection,

$$\begin{aligned} -6\langle \mathsf{R}(X, Y)Z, W \rangle = {} & \alpha(X + W, Y + Z) - \alpha(X + W, Y) - \alpha(X + W, Z) \\ & - \alpha(X, Y + Z) - \alpha(W, Y + Z) + \alpha(X, Z) + \alpha(W, Y) \\ & - \alpha(Y + W, X + Z) + \alpha(Y + W, X) + \alpha(Y + W, Z) \\ & + \alpha(Y, X + Z) + \alpha(W, X + Z) - \alpha(Y, Z) - \alpha(W, X) , \end{aligned}$$

showing that R is determined by α and, thus, by K. We prove the second statement. For that purpose, denote

$$\mathsf{R}_0(X, Y)Z := \langle Y, Z \rangle X - \langle X, Z \rangle Y .$$

Note that R_0 shares the symmetry properties (2.3.9), (2.3.10) and (2.3.12) of R.[25] Assume that $\mathsf{K}(\Sigma_m) = k(m)$ for all non-degenerate planes. If X, Y span a non-degenerate plane, then by (2.4.1),

$$\langle \mathsf{R}(X, Y)Y, X \rangle = k(m)\big(\langle Y, Y \rangle X - \langle X, Y \rangle Y\big) = \langle k(m)\mathsf{R}_0(X, Y)Y, X \rangle .$$

---

[25] It also shares the symmetry property (2.3.11), but this is not needed here.

Thus, the tensor $\hat{\mathsf{R}} := \mathsf{R} - k(m)\mathsf{R}_0$ has the above symmetry properties and fulfils

$$\langle \hat{\mathsf{R}}(X, Y)Y, X\rangle = 0\,. \tag{2.4.3}$$

If X and Y span a degenerate plane, we can choose sequences $X_n \to X$ and $Y_n \to Y$ of tangent vectors such that X_n and Y_n span non-degenerate planes for each n.[26] Then, $\langle \hat{\mathsf{R}}(X_n, Y_n)Y_n, X_n\rangle = 0$ for all n and, thus, (2.4.3) holds for degenerate planes as well. Finally, note that this equation is also true for pairs X, Y which are linearly dependent. We conclude that (2.4.3) holds for all $X, Y \in \mathrm{T}_m M$. Now, the assertion is a consequence of the following simple algebraic fact (Exercise 2.4.3): If

$$\tilde{\mathsf{R}} : \mathrm{T}_m M \times \mathrm{T}_m M \times \mathrm{T}_m M \times \mathrm{T}_m M \to \mathbb{R}$$

is a quadrilinear mapping sharing the symmetry properties (2.3.9), (2.3.10) and (2.3.12) of R, then $\langle \tilde{\mathsf{R}}(X, Y)Y, X\rangle = 0$ implies $\tilde{\mathsf{R}} = 0$.

The converse statement is trivial. ■

Proposition 2.4.2 leads us to an important class of pseudo-Riemannian manifolds.

Definition 2.4.3 If $\mathsf{K}(\Sigma_m) = k(m)$ for every $\Sigma_m \in G_2^0(\mathrm{T}_m M)$, then we say that (M, g) is a space of constant curvature at m. Let k be a real number. We say that (M, g) is a space of constant curvature k if $\mathsf{K}(\Sigma_m) = k$ at every point $m \in M$.

Remark 2.4.4

1. By the proof of Proposition 2.4.2, for a space of constant curvature, we have

$$\mathsf{R}(X, Y)Z = k\big(\langle Y, Z\rangle X - \langle X, Z\rangle Y\big)\,, \quad k \in \mathbb{R}\,. \tag{2.4.4}$$

2. By a theorem of Schur, see Theorem 2.2. in Chap. V of [381], if (M, g) is a space of constant curvature at every point of M and $\dim M \geq 3$, then M is a space of constant curvature, that is, the mapping $m \to k(m)$ is constant.
3. It is not hard to construct models of spaces of constant curvature. The simplest Riemannian example is the n-sphere of radius r embedded in the standard way in $\mathbb{R}^{n+1}$. This is a space of constant curvature equal to $\frac{1}{r^2}$. The simplest pseudo-Riemannian model is the pseudo-Euclidean space $(\mathbb{R}^n_s, \mathsf{g}^n_s)$ with the signature $(n - s, s)$. It is easy to show that this is a space of constant curvature equal to 0. In Sect. 2.5, we will see a large class of spaces of constant curvature. For an exhaustive presentation of this subject we refer to [676].
4. In the indefinite case, there is a lot of subtleties and there is quite a number of classical papers on that subject, see [63, 145, 257, 395, 490] and further references therein. ♦

---

[26] By property (b) above, in any fixed basis of $\mathrm{T}_m M$, $\| X \|^2 \| Y \|^2 - \langle X, Y\rangle^2$ is a polynomial in the components of X and Y whose zero set does not contain any open subset.

Exercises

2.4.1 Prove that (2.4.1) does not depend on the choice of the basis.

2.4.2 Show that the restriction of a pseudo-Riemannian metric to a 2-dimensional subspace $\Sigma_m \subset \mathrm{T}_m M$ is non-degenerate iff $\parallel X \parallel^2 \parallel Y \parallel^2 - \langle X, Y\rangle^2 \neq 0$.

2.4.3 Prove the following. If $\tilde{\mathsf{R}} : \mathrm{T}_m M \times \mathrm{T}_m M \times \mathrm{T}_m M \times \mathrm{T}_m M \to \mathbb{R}$ is a quadrilinear mapping sharing the symmetry properties (2.3.9), (2.3.10) and (2.3.12) of R, then $\langle \tilde{\mathsf{R}}(X, Y)Y, X\rangle = 0$ implies $\tilde{\mathsf{R}} = 0$.

2.5 Symmetric Spaces

In this section, we take up the discussion from Sect. 2.3. We analyze the special case $\mathfrak{K}^1(\mathfrak{h}) = 0$, that is, we analyze the condition

$$D\mathscr{R} = 0\,, \tag{2.5.1}$$

defining locally symmetric manifolds, cf. Definition 2.3.2. Thus, we give up assumption (c) prior to Theorem 2.3.19, but we keep on assuming the following.

(a) M is simply connected, which ensures that the holonomy group H is connected and that it coincides with the restricted holonomy group.
(b) (M, g) is irreducible, which implies that H acts irreducibly.

Moreover, as above, we limit our attention to the Riemannian metric case, that is, $H \subset \mathrm{O}(n)$ is a compact Lie subgroup acting irreducibly on $V \equiv \mathbb{R}^n$. Then, by the Holonomy Principle, cf. Proposition 1.7.20, the space of parallel sections of

$$E = O(M) \times_{\mathrm{O}(n)} S^2\left(\bigwedge\nolimits^2 V^*\right)$$

is in one-to-one correspondence with the space of holonomy-invariant vectors in $S^2\left(\bigwedge^2 V^*\right)$ as follows. Any $\mathscr{R}$ satisfying (2.5.1) is constant on $P_{u_0}(\Gamma)$ and, restricted to $P_{u_0}(\Gamma)$, it takes values in $\mathfrak{K}(\mathfrak{h})$ given by (2.3.29). Thus, the Holonomy Principle assigns to $\mathscr{R}$ the H-invariant element

$$F := \mathscr{R}(u) \in \mathfrak{K}(\mathfrak{h}), \quad u \in P_{u_0}(\Gamma)\,. \tag{2.5.2}$$

Lemma 2.5.1 *Let $H \subset \mathrm{O}(n)$ be a closed subgroup and let $F \in \mathfrak{K}(\mathfrak{h})$ be an H-invariant element. Then, $\mathfrak{g} = \mathfrak{h} \oplus V$ carries the structure of a Lie algebra given by*

$$[A, \mathbf{x}] = -[\mathbf{x}, A] = A\mathbf{x}\,, \quad [\mathbf{x}, \mathbf{y}] = -F(\mathbf{x}, \mathbf{y})\,, \quad A \in \mathfrak{h}\,,\ \mathbf{x}, \mathbf{y} \in V\,.$$

Proof Bilinearity and anti-symmetry are obvious. We prove that the Jacobi identity holds. For that purpose, we have to consider three cases:
(a) Let $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$. Since $F(\mathbf{x}, \mathbf{y}) \in \mathfrak{h} \subset \mathrm{End}(V)$, the definition of $\mathfrak{K}(\mathfrak{h})$ implies

$$[[\mathbf{x}, \mathbf{y}], \mathbf{z}] + [[\mathbf{y}, \mathbf{z}], \mathbf{x}] + [[\mathbf{z}, \mathbf{x}], \mathbf{y}] = 0 .$$

(b) Let $\mathbf{x}, \mathbf{y} \in V$. By the H-invariance of F, cf. (2.1.25), we have

$$F(\mathbf{x}, \mathbf{y}) = \mathrm{Ad}(a^{-1}) \circ F(a\mathbf{x}, a\mathbf{y}) , \quad a \in H \subset \mathrm{O}(n) .$$

Differentiating this equation, we obtain

$$[F(\mathbf{x}, \mathbf{y}), A] + F(A\mathbf{x}, \mathbf{y}) + F(\mathbf{x}, A\mathbf{y}) = 0$$

for any $A \in \mathfrak{h}$. This implies

$$[[\mathbf{x}, \mathbf{y}], A] + [[\mathbf{y}, A], \mathbf{x}] + [[A, \mathbf{x}], \mathbf{y}] = 0 .$$

(c) Let $\mathbf{x} \in V$ and $A, B \in \mathfrak{h}$. Then, by definition of the Lie bracket of $\mathfrak{h} \subset \mathrm{End}(V)$,

$$[A, B](\mathbf{x}) = A(B\mathbf{x}) - B(A\mathbf{x}) .$$

This proves the third case. ■

To make contact with the standard notation, we denote $V = \mathfrak{m}$. Then,

$$\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m} \tag{2.5.3}$$

and the commutation relations of $\mathfrak{g}$ fulfil:

$$[\mathfrak{h}, \mathfrak{h}] \subset \mathfrak{h} , \quad [\mathfrak{h}, \mathfrak{m}] \subset \mathfrak{m} , \quad [\mathfrak{m}, \mathfrak{m}] \subset \mathfrak{h} . \tag{2.5.4}$$

Moreover, by the Ambrose-Singer Theorem,

$$[\mathfrak{m}, \mathfrak{m}] = \mathfrak{h} . \tag{2.5.5}$$

Associated with the decomposition (2.5.3), there is a linear mapping

$$\lambda : \mathfrak{g} \to \mathfrak{g} , \quad \lambda(A, \mathbf{x}) := (A, -\mathbf{x}) , \quad A \in \mathfrak{h} , \mathbf{x} \in \mathfrak{m} . \tag{2.5.6}$$

By (2.5.4), λ is an involutive Lie algebra homomorphism (Exercise 2.5.1). Conversely, we have the following.

Lemma 2.5.2 *Any involutive Lie algebra homomorphism* λ *of a Lie algebra* $\mathfrak{g}$ *induces a decomposition* $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$ *fulfilling* (2.5.4).

Proof Since $\lambda^2 = \mathrm{id}$, λ is diagonalizable and we may decompose $\mathfrak{g}$ into the eigenspaces $\mathfrak{h}$ and $\mathfrak{m}$ of λ corresponding to the eigenvalues $+1$ and -1, respectively. Now, the first relation in (2.5.4) is obvious. To check the remaining two, we calculate

$$\lambda([A, \mathbf{x}]) = [\lambda(A), \lambda(\mathbf{x})] = -[A, \mathbf{x}] , \quad A \in \mathfrak{h} , \ \mathbf{x} \in \mathfrak{m} ,$$

that is, $[A, \mathbf{x}] \in \mathfrak{m}$. Similarly, $\lambda([\mathbf{x}, \mathbf{y}]) = [\mathbf{x}, \mathbf{y}] \in \mathfrak{h}$ for any $\mathbf{x}, \mathbf{y} \in \mathfrak{m}$. ■

Definition 2.5.3 Let $\mathfrak{g}$ be a Lie algebra and let λ be an involutive automorphism of $\mathfrak{g}$. Then, the pair $(\mathfrak{g}, \lambda)$ is called a symmetric Lie algebra. In addition,

1. if the set of fixed points $\mathfrak{h}$ of λ is a compactly embedded Lie subalgebra[27] of $\mathfrak{g}$, then $(\mathfrak{g}, \lambda)$ is called an orthogonal symmetric Lie algebra,
2. if $\mathfrak{h} \cap \mathfrak{z} = \{0\}$, where $\mathfrak{z}$ is the center of $\mathfrak{g}$, then $(\mathfrak{g}, \lambda)$ is called effective.
3. if $(\mathfrak{g}, \lambda)$ is effective and $\mathrm{ad}([\mathfrak{m}, \mathfrak{m}])$ acts irreducibly on $\mathfrak{m}$, then $(\mathfrak{g}, \lambda)$ is called irreducible.

Proposition 2.5.4 *The Lie algebra* $\mathfrak{g}$ *constructed in Lemma 2.5.1, endowed with the involutive automorphism* λ *given by* (2.5.6)*, is an irreducible orthogonal symmetric Lie algebra.*

Proof By construction, $(\mathfrak{g}, \lambda)$ is symmetric. Since, by assumption, $H \subset \mathrm{O}(n)$ is a compact Lie subgroup acting faithfully on $\mathbb{R}^n$, $\mathrm{ad}(\mathfrak{h})$ is compact and, thus, $(\mathfrak{g}, \lambda)$ is orthogonal. Suppose $A \in \mathfrak{h} \cap \mathfrak{z}$. Then,

$$A\mathbf{x} = [A, \mathbf{x}] = 0$$

for every $\mathbf{x} \in \mathfrak{m}$ and, thus, $A = 0$. Thus, $(\mathfrak{g}, \lambda)$ is effective. Finally, by assumption, H acts irreducibly on $\mathfrak{m}$. Thus, $\mathrm{ad}(\mathfrak{h})$ acts irreducibly on $\mathfrak{m}$, too. This, together with (2.5.5) implies that $(\mathfrak{g}, \lambda)$ is irreducible. ■

In the sequel, the pair $(\mathfrak{g}, \lambda)$ constructed above will be called the canonical symmetric Lie algebra associated with the locally symmetric Riemannian manifold we started with. The decomposition (2.5.3) will be called the canonical decomposition of $(\mathfrak{g}, \lambda)$.

The following proposition characterizes irreducible symmetric Lie algebras.

Proposition 2.5.5 *Let* $(\mathfrak{g}, \lambda)$ *be an irreducible symmetric Lie algebra and let* $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$ *be the decomposition induced by* λ*. Then, one of the following cases occurs:*

1. $\mathfrak{g}$ *is a simple Lie algebra.*
2. $\mathfrak{g} = \tilde{\mathfrak{g}} \oplus \tilde{\mathfrak{g}}$ *with* $\tilde{\mathfrak{g}}$ *simple, fulfilling* $\mathfrak{h} = \big\{(A, A) : A \in \tilde{\mathfrak{g}}\big\}$ *and* $\lambda(A, B) = (B, A)$ *for any* $A, B \in \tilde{\mathfrak{g}}$.
3. $[\mathfrak{m}, \mathfrak{m}] = 0$.

For the proof we refer the reader to [381].[28]

---

[27] That is, the group of transformations of $\mathfrak{g}$ generated by $\mathrm{ad}(\mathfrak{h})$ is compact.

[28] Cf. Proposition 7.5 in Vol. 2, Chap. XI of [381].

Remark 2.5.6

1. Assume that either point 1 or point 2 of Proposition 2.5.5 holds. Then, since $[\mathfrak{m}, \mathfrak{m}] \oplus \mathfrak{m}$ is an ideal in $\mathfrak{g}$, we have $\mathfrak{h} = [\mathfrak{m}, \mathfrak{m}]$. Thus, an effective symmetric Lie algebra is irreducible iff $\mathfrak{h} = [\mathfrak{m}, \mathfrak{m}]$, that is, iff $\mathfrak{g}$ is of the form described either by point 1 or by point 2. In particular, if $(\mathfrak{g}, \lambda)$ is irreducible, then $\mathfrak{g}$ is semisimple.
2. Conversely, if $(\mathfrak{g}, \lambda)$ is an orthogonal symmetric Lie algebra and $\mathfrak{g}$ is simple, then $\mathrm{ad}(\mathfrak{h})$ acts irreducibly on $\mathfrak{m}$, see Proposition 7.4 in Vol. 2, Chap. XI of [381]. ♦

Proposition 2.5.5 and property (2.5.5) imply that the canonical symmetric Lie algebra $(\mathfrak{g}, \lambda)$ is semisimple. Consequently, by Proposition I/5.4.10, the Killing form

$$\mathsf{k} : \mathfrak{g} \times \mathfrak{g} \to \mathbb{R}, \quad \mathsf{k}(X, Y) = \mathrm{tr}(\mathrm{ad}(X)\mathrm{ad}(Y)),$$

of $\mathfrak{g}$ is non-degenerate. Moreover, the relations (2.5.4) imply that the decomposition (2.5.3) is orthogonal with respect to k (Exercise 2.5.2). Equivalently, k is λ-invariant. This implies that the restrictions $\mathsf{k}^{\mathfrak{h}}$ and $\mathsf{k}^{\mathfrak{m}}$ of k to $\mathfrak{h}$ and $\mathfrak{m}$, respectively, are both non-degenerate and λ-invariant, too. Moreover, they have the following properties:

(a) By Corollary I/5.5.8, $\mathsf{k}^{\mathfrak{h}}$ is negative semidefinite and, since $(\mathfrak{g}, \lambda)$ is effective, it is negative definite.

(b) Since $\mathrm{ad}(\mathfrak{h})$ acts irreducibly on $\mathfrak{m}$ and since both $\mathsf{k}^{\mathfrak{m}}$ and the scalar product η on $\mathfrak{m}$ induced from the metric g are $\mathrm{ad}(\mathfrak{h})$-invariant, by Schur's Lemma, they must be proportional to each other,

$$\eta(\mathbf{x}, \mathbf{z}) = -c\, \mathsf{k}^{\mathfrak{m}}(\mathbf{x}, \mathbf{z}), \quad \mathbf{x}, \mathbf{z} \in \mathfrak{m}, \ c \in \mathbb{R}, c \neq 0. \tag{2.5.7}$$

Thus, since η is positive definite, $\mathsf{k}^{\mathfrak{m}}$ is either positive or negative definite.

Definition 2.5.7 An effective orthogonal symmetric Lie algebra $(\mathfrak{g}, \lambda)$ with $\mathfrak{g}$ semisimple is said to be of compact or of non-compact type, if the restriction of the Killing form of $\mathfrak{g}$ to $\mathfrak{m}$ is, respectively, negative definite or positive definite.

Remark 2.5.8 Combining Proposition 2.5.5 with Propositions 7.4 and 7.5 in in Vol. 2, Chap. XI of [381], one can show that any irreducible orthogonal symmetric Lie algebra is either of compact or of non-compact type. ♦

Next, we show that, given an irreducible orthogonal symmetric Lie algebra $(\mathfrak{g}, \lambda)$, one can construct a special type of homogeneous Riemannian manifold.

Let $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$ be the decomposition induced from λ. Let $\tilde{G}$ be the connected simply connected Lie group with Lie algebra $\mathfrak{g}$ and let $\tilde{H}$ be the connected Lie subgroup corresponding to $\mathfrak{h}$. Then, the space of left cosets $M := \tilde{G}/\tilde{H}$ is a simply connected manifold endowed with the natural left $\tilde{G}$-action given by left translations. Let

$$\tilde{Z} = \left\{ g \in \tilde{G} : g(m) = m \ \text{ for all } \ m \in M \right\}$$

be the kernel of this action. Since, by assumption, $(\mathfrak{g}, \lambda)$ is effective, $\tilde{Z}$ must be discrete. Thus, M is an almost effective $\tilde{G}$-manifold. We pass to an effective action by setting $G := \tilde{G}/\tilde{Z}$ and $H := \tilde{H}/\tilde{Z}$. Then, $M = G/H$, G and H are connected, and we have the natural left effective action

$$\delta : G \times G/H \to G/H\,, \quad (a, [g]) \mapsto \delta_a([g]) := [ag]\,.$$

By point 4 of Example 1.1.4, the natural projection $\pi : G \to M$ endows G with the structure of a principal H-bundle P and the tangent mapping π' identifies $\mathfrak{m}$ and $\mathrm{T}_{[\mathbb{1}]}M$ as vector spaces. Under this identification, the isotropy representation

$$H \to \mathrm{Aut}(\mathrm{T}_{[\mathbb{1}]}M)\,, \quad h \mapsto (\delta_h)'_{\mathbb{1}}\,,$$

is given by $\mathrm{Ad}(H)$ acting on $\mathfrak{m}$, cf. point 1 of Remark I/6.2.10. Correspondingly,

$$G \times_{\mathrm{Ad}(H)} \mathfrak{m} \to \mathrm{T}M\,, \quad [(a, \mathbf{x})] \mapsto [\mathrm{L}'_a(\mathbf{x})], \tag{2.5.8}$$

is an isomorphism. Since $(\mathfrak{g}, \lambda)$ is orthogonal and irreducible, there exists an $\mathrm{Ad}(H)$-invariant scalar product η on $\mathfrak{m}$ which is unique up to a positive factor. Clearly, η induces an H-invariant scalar product on $\mathrm{T}_{[\mathbb{1}]}M$ which, using the left G-action δ, can be extended to a G-invariant Riemannian metric g on M. To summarize, we have constructed a simply connected transitive and effective G-manifold (M, g) with G acting by isometries.

Consider the bundle of orthonormal frames $O(M)$ of (M, g). Note that any η-orthonormal basis $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ of $\mathfrak{m}$ induces via π' a g-orthonormal frame $(e_1, \ldots, e_n)$ at $[\mathbb{1}] \in M$ and, thus, an injective bundle morphism

$$\vartheta : P \to O(M)\,, \quad \vartheta(a) := (\delta'_a(e_1), \ldots, \delta'_a(e_n))\,, \tag{2.5.9}$$

projecting onto the identical diffeomorphism of M. The corresponding Lie group homomorphism $\tau : H \to O(n) \subset \mathrm{GL}(n, \mathbb{R}) \cong \mathrm{Aut}(\mathrm{T}_m M)$ is given by the adjoint action of H on $\mathfrak{m} \cong \mathrm{T}_{[\mathbb{1}]}M$. To summarize, P is a subbundle of $O(M)$.

Now, decompose the Maurer–Cartan form $\theta^G \in \Omega^1(G, \mathfrak{g})$ with respect to (2.5.3):

$$\theta^G = \theta_{\mathfrak{h}} + \theta_{\mathfrak{m}}\,.$$

By Example 1.3.19, $\theta_{\mathfrak{h}}$ coincides with the canonical G-invariant connection[29] ω^c on P. Recall that the corresponding horizontal distribution is generated by $\mathfrak{m}$, that is, by left invariant vector fields $a \mapsto (\mathbf{x}_*)_a = L'_a(\mathbf{x})$ with $\mathbf{x} \in \mathfrak{m}$.

Lemma 2.5.9 *Under the morphism* (ϑ, τ), $\theta_{\mathfrak{m}}$ *corresponds to the soldering form* θ *on* $O(M)$, *that is,* $\vartheta^*\theta = \theta_{\mathfrak{m}}$.

---

[29] Note that this is a special case of the canonical invariant connection defined in point 2 of Remark 1.9.14. It is obtained by setting $G = H$ and $\lambda = \mathrm{id}$ there.

Proof By $\mathfrak{m}$-valuedness of $\theta_\mathfrak{m}$ and horizontality of θ, both $\vartheta^*\theta$ and $\theta_\mathfrak{m}$ vanish on the left invariant vector fields generated by elements of $\mathfrak{h}$. Thus, let $\mathbf{x}_*$ be generated by $\mathbf{x} \in \mathfrak{m}$. Then, clearly $\theta_\mathfrak{m}(\mathbf{x}_*) = \mathbf{x}$. On the other hand,

$$(\vartheta^*\theta)_g(\mathbf{x}_*) = \vartheta(g)^{-1}(\rho' \circ \vartheta'(\mathbf{x}_*)) = \vartheta(g)^{-1}(\pi'(\mathbf{x}_*)) = \vartheta(g)^{-1}(\delta'_g \circ \pi'(\mathbf{x})) = \mathbf{x}\,,$$

where $\rho : O(M) \to M$ is the canonical projection. ■

Proposition 2.5.10 *The Riemannian manifold (M, g) has the following properties:*

1. *Under the morphism (ϑ, τ), the Levi-Civita connection ω^0 of (M, g) corresponds to the canonical connection ω^c, that is, $\vartheta^*\omega^0 = \omega^c$.*
2. *The Riemann curvature of (M, g) is constant and given by the linear mapping*

$$F : \bigwedge{}^2\mathfrak{m} \to \mathfrak{h}\,, \quad F(\mathbf{x}, \mathbf{y}) = -[\mathbf{x}, \mathbf{y}]\,. \tag{2.5.10}$$

3. *The holonomy group based at $\vartheta(\mathbb{1})$ of ω^0 is H and the holonomy bundle coincides with P.*
4. *The Riemann curvature of (M, g) is parallel, that is, (M, g) is locally symmetric.*
5. *For any $\mathbf{x} \in \mathfrak{m}$, $t \mapsto \pi(L_g \exp(t\mathbf{x}))$ is a geodesic through $[g] \in M$. Conversely, every geodesic through $[g]$ is of this form. In particular, M is geodesically complete.*

Proof 1. We decompose the commutator $[\theta^G, \theta^G] \in \Omega^2(G, \mathfrak{g})$ with respect to (2.5.3). By (2.5.4),

$$[\theta^G, \theta^G]_\mathfrak{h} = [\theta_\mathfrak{h}, \theta_\mathfrak{h}] + [\theta_\mathfrak{m}, \theta_\mathfrak{m}]\,, \quad [\theta^G, \theta^G]_\mathfrak{m} = 2[\theta_\mathfrak{h}, \theta_\mathfrak{m}]\,. \tag{2.5.11}$$

Since the Levi-Civita connection is uniquely characterized by its covariant derivative D_{ω^0} on $\mathrm{T}M$, it is enough to show that the covariant derivative D_{ω^c} induced by ω^c via the isomorphism (2.5.8) coincides with D_{ω^0}. This is done by showing that the extension of ω^c to $O(M)$ is metric and torsionless. By Proposition 1.2.6, we may view any vector field X on M as an H-equivariant mapping $\tilde{X} : G \to \mathfrak{m}$ and, thus,

$$D_{\omega^c}\tilde{X} = \mathrm{d}\tilde{X} + \mathrm{ad}(\omega^c) \circ \tilde{X} = \mathrm{d}\tilde{X} + [\theta_\mathfrak{h}, \tilde{X}]\,,$$

cf. Eq. (1.4.2). Let η be the (unique up to a positive factor) $\mathrm{Ad}(H)$-invariant scalar product on $\mathfrak{m}$. By $\mathrm{Ad}(H)$-invariance, we obtain

$$\eta(D_{\omega^c}\tilde{X}, \tilde{Y}) + \eta(\tilde{X}, D_{\omega^c}\tilde{Y}) = \mathrm{d}(\eta(\tilde{X}, \tilde{Y}))\,.$$

This shows that the extension of ω^c to $O(M)$ is metric. It remains to show that this extension is torsionless: restricting the Maurer–Cartan equation to $\mathfrak{m}$ and using (2.5.11) we get

$$D_{\omega^c}\theta_\mathfrak{m} = \mathrm{d}\theta_\mathfrak{m} + [\theta_\mathfrak{h}, \theta_\mathfrak{m}] = 0\,.$$

But, by Lemma 2.5.9, $\vartheta^*\theta = \theta_\mathfrak{m}$ and, thus, $\vartheta^*\Theta = 0$. By uniqueness of the Levi-Civita connection, the assertion follows.

2. By the Structure Equation, the curvature form of ω^c is given by[30]

$$\Omega^c = -\frac{1}{2}[\theta_\mathfrak{m}, \theta_\mathfrak{m}]\,.$$

By point 1, $\vartheta^*\Omega^0 = \Omega^c$. These two facts immediately imply (2.5.10).

3. By point 2 and by the Ambrose-Singer Theorem, the Lie algebra of the holonomy group of ω^0 is $[\mathfrak{m}, \mathfrak{m}]$. By point 1 of Remark 2.5.6, $[\mathfrak{m}, \mathfrak{m}] = \mathfrak{h}$ and, thus, the Lie algebra of the holonomy group of ω^0 coincides with $\mathfrak{h}$. Since, by construction, M is simply connected, the holonomy group of ω^0 is connected and coincides with the restricted holonomy group. On the other hand, since H is connected, too, we obtain the assertion. It follows that P coincides with the holonomy bundle of ω^0.

4. Since the curvature is constant on P and, thus, H-invariant, the Holonomy Principle 1.7.20 implies the assertion.

5. By Proposition 2.1.22, the geodesics of (M, g) are given by the projections of integral curves of horizontal standard vector fields on $L(M)$. Since they are horizontal, these curves may be chosen to lie in P. The restriction of $B(\mathbf{y})$, $\mathbf{y} \in \mathbb{R}^n$, to P is given by the left-invariant vector field generated by $\mathbf{x} = y^i\mathbf{e}_i \in \mathfrak{m}$, where $\{\mathbf{e}_i\}$ is a basis in $\mathfrak{m}$. Thus, here, the geodesics are given as projections of (global) one-parameter subgroups $t \mapsto \exp(t\mathbf{x})$ and their left translates by arbitrary group elements $g \in G$. ■

By point 3 of Proposition 2.5.10, the irreducibility of $(\mathfrak{g}, \lambda)$ implies that (M, g) is irreducible. Together with points 4 and 5, this yields the following.

Corollary 2.5.11 *(M, g) is a complete irreducible locally symmetric Riemannian manifold.* ■

Next, we show that the involutive automorphism λ induces a special symmetry for any point $m \in M$. Since any automorphism of a Lie algebra is the differential of a unique automorphism of the corresponding simply connected Lie group,[31] λ induces a unique automorphism σ of $\tilde{G}$. By (2.5.6), it fulfils $\sigma(\tilde{H}) = \tilde{H}$. Thus, σ descends to an involutive diffeomorphism $s : M \to M$. By construction,

$$s'_{[\mathbb{1}]} : \mathrm{T}_{[\mathbb{1}]}M \to \mathrm{T}_{[\mathbb{1}]}M\,, \quad s'_{[\mathbb{1}]}(X) = -X\,. \tag{2.5.12}$$

Thus, under the identification $\mathrm{T}_{[\mathbb{1}]}M = \mathfrak{m}$, we have $s'_{[\mathbb{1}]} = \lambda_{\restriction\mathfrak{m}}$.

Lemma 2.5.12 *The origin $[\mathbb{1}]$ of M is an isolated fixed point of s. Moreover, s is an isometry of the Riemannian metric g.*

---

[30] Since ω^c is a G-invariant connection, this is a special case of point 4 of Remark 1.9.14.

[31] For a proof, see e.g. Theorem 3.27 in [652].

Proof The proof of the first assertion is left to the reader (Exercise 2.5.4). To prove the second statement, we have to show that the mapping

$$s'_m : \mathrm{T}_m M \to \mathrm{T}_m M$$

is isometric. For the point $m = [\mathbb{1}]$, this follows immediately from (2.5.12), because at the origin g coincides with η and the latter is λ-invariant. To prove the invariance for an arbitrary point $m = [g]$, note that for any $g, h \in G$,

$$s(\delta_g[h]) = s([gh]) = [\sigma(g)\sigma(h)] = \delta_{\sigma(g)}[\sigma(h)] = \delta_{\sigma(g)} s([h]) ,$$

that is, $s \circ \delta_g = \delta_{\sigma(g)} \circ s$. Differentiation of this identity yields

$$s'_{[g]} \circ (\delta_g)'_{[\mathbb{1}]} = (\delta_{\sigma(g)})'_{[\mathbb{1}]} \circ s'_{[\mathbb{1}]} .$$

By construction, g is G-invariant and, thus, $(\delta_g)'_{[\mathbb{1}]}$ and $(\delta_{\sigma(g)})'_{[\mathbb{1}]}$ leave g invariant. This yields the assertion. ■

Remark 2.5.13 For every $g \in \tilde{Z}$, we have $\big(\sigma(g)\big)(m) = s \circ g \circ s(m) = s^2(m) = m$. Hence, $\sigma(\tilde{Z}) = \tilde{Z}$ and σ descends to an automorphism of G, denoted by the same symbol. One has $\sigma(H) = H$. ♦

Next, for any $m = [g] \in M$, we define[32]

$$s_m : M \to M , \quad s_m := \delta_g \circ s \circ \delta_{g^{-1}} . \tag{2.5.13}$$

Differentiating (2.5.13), we obtain $s'_m = \delta'_g \circ s'_{[\mathbb{1}]} \circ \delta'_{g^{-1}}$ for any $m = [g] \in M$. Thus, by Lemma 2.5.12, by formula (2.5.12) and by the G-invariance of g, for any $m \in M$, s_m is an involutive isometry of g fulfilling (Exercise 2.5.5)

$$s_m(m) = m , \quad (s_m)'_m = -\,\mathrm{id} . \tag{2.5.14}$$

The following remark yields a geometric interpretation of the symmetry s_m.

Remark 2.5.14 Let $t \to \gamma(t)$ be a geodesic of (M, g) with $\gamma(0) = m$. Since an isometry transforms geodesics to geodesics, $t \mapsto \tau(t) := s_m(\gamma(t))$ is a geodesic, too. By (2.5.14), its tangent vector at $t = 0$ satisfies

$$\dot{\tau}(0) = (s_m)'_m \, \dot{\gamma}(0) = -\dot{\gamma}(0) . \tag{2.5.15}$$

Now, the uniqueness property of geodesics, see Corollary 2.1.23, implies $\tau(t) = \gamma(-t)$. Thus, for any $m \in M$,

---

[32] Clearly, this definition does not depend on the choice of the representative.

$$s_m(\gamma(t)) = \gamma(-t)\,, \tag{2.5.16}$$

that is, s_m reverses the geodesics through m. ◆

Definition 2.5.15 (*Riemannian globally symmetric space*) A Riemannian manifold (M, g) is called globally symmetric if for each $m \in M$ there exists an involutive isometry $s_m : M \to M$ such that m is an isolated fixed point of s_m. The mapping s_m is called the symmetry of (M, g) at m.

Taking into account that, in the above construction of (M, g), the scalar product on $\mathfrak{m}$ is unique up to a positive constant and that a change of this constant implies a conformal transformation of g, we obtain the following.

Proposition 2.5.16 *To any irreducible*[33] *orthogonal symmetric Lie algebra* $(\mathfrak{g}, \lambda)$ *there corresponds a unique homothetic equivalence class* $(M, [\mathsf{g}])$ *of simply connected irreducible Riemannian globally symmetric spaces.* ■

It should be clear that the locally symmetric Riemannian manifold we started with and the Riemannian globally symmetric space constructed here are deeply related. Indeed, let (M, g) be a locally symmetric space. Let $(\mathfrak{g}, \lambda)$ be its canonical symmetric Lie algebra with canonical decomposition $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$. Let $\eta \in S^2(\mathfrak{m}^*)$ be the scalar product on $\mathfrak{m}$ defined by g and let $F \in \mathfrak{K}(\mathfrak{h}) \subset \bigwedge^2 \mathfrak{m}^* \otimes \mathfrak{h}$ be the Riemann curvature of (M, g). Let G/H be the Riemannian globally symmetric space constructed from $(\mathfrak{g}, \lambda)$. Then, for any chosen point $m \in M$, via

$$\mathrm{T}_m M \cong \mathfrak{m} \cong \mathrm{T}_{[1]} G/H$$

we obtain an isometric isomorphism between $\mathrm{T}_m M$ and $\mathrm{T}_{[1]} G/H$ and, by point 2 of Proposition 2.5.10, M and G/H have the same Riemann curvature given by the mapping F. By standard arguments,[34] this implies the following.

Corollary 2.5.17 *Every point of a locally symmetric space* (M, g) *admits a neighbourhood isometric to a neighbourhood of the origin of the Riemannian globally symmetric space constructed from the canonical symmetric Lie algebra of* (M, g).

Note, however, that not every locally symmetric space is a Riemannian globally symmetric space. It is even not necessarily homogeneous. As an example,[35] let M be a compact Riemann surface with genus ≥ 2, equipped with a Riemannian metric of constant curvature equal to -1. Then, the isometry group of M is finite and, thus, M is not homogeneous and, consequently, also not globally symmetric.

As an immediate consequence of the existence of the symmetries s_m, we obtain

Proposition 2.5.18 *Any Riemannian globally symmetric space* (M, g) *is complete.*

---

[33] Remember that irreducibility includes effectiveness, cf. Definition 2.5.3.

[34] See Theorem 7.4 in Chap. VI of [381].

[35] This example is taken from [73].

Proof Consider any geodesic $t \mapsto \gamma(t)$ defined on the interval $[0, t_0[$. Apply the symmetry $s_{\gamma(t_0-\varepsilon)}$ to γ with some ε fulfilling $0 < \varepsilon < \frac{t_0}{2}$. By (2.5.16), this operation extends the domain of γ to $[0, 2t_0 - 2\varepsilon[$. Continuing this procedure, we obtain completeness of (M, g). ■

Next, given a Riemannian globally symmetric space (M, g), for every geodesic $t \mapsto \gamma(t)$ we consider the family of isometries

$$T^{\gamma}_t := s_{\gamma(\frac{t}{2})} \circ s_{\gamma(0)}\,, \tag{2.5.17}$$

called the transvections along γ. The following properties are immediate consequences of (2.5.15) and (2.5.16) and are, therefore, left to the reader (Exercise 2.5.3).

Proposition 2.5.19 *Let (M, g) be a Riemannian globally symmetric space and let $t \mapsto \gamma(t)$ be a geodesic. Then,*

1. *T^{γ}_t acts on γ by translations, that is, $T^{\gamma}_t(\gamma(s)) = \gamma(t+s)$.*
2. *$(T^{\gamma}_t)'_{\gamma(s)}$ acts by parallel translation from $\gamma(s)$ to $\gamma(t+s)$ along γ, that is, for any parallel vector field X along γ,*

$$(T^{\gamma}_t)'_{\gamma(s)}(X(\gamma(s)) = X(\gamma(t+s))\,.$$

3. *$\left\{T^{\gamma}_t\right\}_{t\in\mathbb{R}}$ is a 1-parameter group of isometries, that is, $T^{\gamma}_{t+s} = T^{\gamma}_t \circ T^{\gamma}_s$.* ■

Recall from Example 2.2.16 that the isometry group $I(M)$ of a Riemannian manifold M is a Lie group. Let us denote its identity component by $I_0(M)$. By point 3 of Proposition 2.5.19, for any geodesic γ, the transvections T^{γ}_t form a subgroup (called the transvection group) of $I_0(M)$. On the other hand, by a classical theorem of Hopf and Rinow,[36] any two points of a complete Riemannian manifold may be joined by a geodesic. Using these two facts, we obtain the following.

Corollary 2.5.20 *Let (M, g) be a Riemannian globally symmetric space. Then,*

1. *Geodesics in M are images of 1-parameter groups of isometries.*
2. *The identity component $I_0(M)$ acts transitively on M.* ■

Proposition 2.5.21 *Let (M, g) be an irreducible Riemannian globally symmetric space and let G be a Lie group acting transitively and isometrically on M. If G acts effectively, then G coincides with $I_0(M)$.*

Proof Clearly, $I_0(M)$ is the largest connected group of isometries of (M, g). Denote $G' = I_0(M)$ and let $\mathfrak{g}'$ be its Lie algebra. Conjugation by s defines an automorphism σ' of G' which clearly restricts to the automorphism σ of G, cf. Remark 2.5.13. The canonical decompositions $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$ and $\mathfrak{g}' = \mathfrak{h}' \oplus \mathfrak{m}'$ necessarily fulfil $\mathfrak{m}' = \mathfrak{m}$. Here $\mathfrak{h}'$ is the Lie algebra of the stabilizer of the chosen point on M under G'. Thus, by Remark 2.5.6,

---

[36] See, e.g. [352].

$$\mathfrak{h} = [\mathfrak{m}, \mathfrak{m}] = [\mathfrak{m}', \mathfrak{m}'] = \mathfrak{h}' .$$

This implies $\mathfrak{g}' = \mathfrak{g}$ and, thus, $G' = G$. ■

Thus, in the construction leading to Proposition 2.5.16, the Lie group G actually coincides with $I_0(M)$. Now, we are able to prove the converse of Proposition 2.5.16.

Proposition 2.5.22 *To any simply connected irreducible Riemannian globally symmetric space there corresponds a unique irreducible orthogonal symmetric Lie algebra.*

Proof Let (M, g) be a simply connected irreducible Riemannian globally symmetric space. By Corollary 2.5.20, $G = I_0(M)$ acts transitively and effectively on M. Let H be the isotropy group of this Lie group action at a chosen point $o \in M$. By the homotopy sequence of the fibration $H \to G \to G/H$, the simply-connectedness of G/H and the connectedness of G imply that H is connected. Moreover, by Theorem 3.4 in Chap. VI of [381], the isotropy subgroup $I(M)_m$ at any point $m \in M$ is compact. Hence, $H = G \cap I(M)_o$ is compact, too. Thus, $M = G/H$ and, by standard arguments, $\pi : G \to M$ is a submersion. In particular, $\pi' : \mathrm{T}_1 G \to \mathrm{T}_o M$ is an H-equivariant surjective linear mapping whose kernel coincides with $\mathrm{T}_1 H$.

Let s be the symmetry at o. Since s is an involutive diffeomorphism, the mapping $g \mapsto \sigma(g) := s \circ g \circ s^{-1}$ defines an involutive automorphism of G. Let $\mathfrak{g}$ and $\mathfrak{h}$ be the Lie algebras of G and H, respectively. Clearly, $\lambda := \sigma'$ is an involutive automorphism of $\mathfrak{g}$. Let $\mathfrak{m}$ be the eigenspace of λ corresponding to the eigenvalue -1. By (2.5.14), $\pi'(\mathfrak{m}) = \mathrm{T}_o M$. We prove that $\mathfrak{h}$ is the eigenspace of λ corresponding to the eigenvalue $+1$: let

$$G^\sigma := \{g \in G : \sigma(g) = g\}$$

be the fixed point set of σ. By (2.5.14), s_o' commutes with the isotropy representation of H at o and, thus, H is contained in G^σ. Conversely, if $g \in G^\sigma$, then it commutes with s and, thus, for any 1-parameter subgroup $t \mapsto g_t$ of G^σ,

$$s \circ g_t(o) = g_t \circ s(o) = g_t(o) ,$$

that is, the orbit $g_t(o)$ is left invariant pointwise by s. Now, by Lemma 2.5.12, o is an isolated fixed point. Thus, $g_t(o)$ must coincide with o. But, $g_t(o) = o$ implies that the 1-parameter subgroup $t \mapsto g_t$ is contained in H. Since a connected Lie group is generated by its 1-parameter subgroups, we have $(G^\sigma)^0 \subset H$. Thus,

$$(G^\sigma)^0 \subset H \subset G^\sigma .$$

This relation implies that $\mathfrak{h}$ coincides with the $(+1)$-eigenspace of λ, indeed. To summarize, the decomposition $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$ is canonical with respect to λ, that is, $(\mathfrak{g}, \lambda)$ is a symmetric Lie algebra. Since H is compact, $\mathrm{ad}(\mathfrak{h})$ is a compactly embedded Lie subalgebra of $\mathfrak{g}$, that is, $(\mathfrak{g}, \lambda)$ is orthogonal. It remains to prove that $(\mathfrak{g}, \lambda)$ is irreducible. Since g is G-invariant, we are in the situation described by Proposition

2.5.10. By this proposition, H coincides with the holonomy group of the Levi-Civita connection of g. Thus, the irreducibility of (M, g), together with the effectiveness of the action of G on M, implies the irreducibility of $(\mathfrak{g}, \lambda)$. ■

Remark 2.5.23 In the course of the above proof, we have found the following structure: a triple (G, H, σ) fulfilling

1. G is a connected Lie group and H is a closed subgroup,
2. σ is an involutive automorphism of G such that $(G^\sigma)^0 \subset H \subset G^\sigma$,
3. $\mathrm{Ad}(H)$ is compact,

is called a Riemannian symmetric pair. This notion clearly constitutes a link between symmetric spaces and symmetric Lie algebras. ♦

Combining Proposition 2.5.16 with Proposition 2.5.22, we obtain the following.

Theorem 2.5.24 *The homothetic equivalence classes of simply connected irreducible Riemannian globally symmetric spaces are in one-to-one correspondence with the irreducible orthogonal symmetric Lie algebras.* ■

This theorem reduces the classification of symmetric spaces of the above type to the classification of irreducible symmetric Lie algebras of compact or of non-compact type. According to a beautiful duality,[37] the problem further reduces to the classification of irreducible symmetric Lie algebras of the non-compact type. The latter can be shown to be in one-to-one correspondence with the real simple Lie algebras of non-compact type. If the complexification of such a Lie algebra is simple as a complex Lie algebra, then M is said to be of type III, otherwise M is said to be of type IV. The corresponding compact irreducible symmetric spaces are obtained by duality and are referred to as of type I and II, respectively. The complete list of simply connected irreducible symmetric spaces with symmetry group being a classical Lie group is given in Tables 2.1 and 2.2.[38] Here, $\mathrm{SO}_0(p, q)$ denotes the identity component of $\mathrm{SO}(p, q)$ and $\mathrm{SO}^*(2n)$ is the subgroup of $\mathrm{SO}(2n, \mathbb{C})$ satisfying

$$g^{\mathrm{T}} \mathsf{J}_0 \overline{g} = \mathsf{J}_0 , \quad g^{\mathrm{T}} g = \mathbb{1}_{2n} .$$

For the corresponding list with exceptional Lie groups we refer to the textbook of Helgason [293]. As already mentioned, there the reader may find an exhaustive presentation of the whole subject.

Remark 2.5.25 Note that in our considerations, we have excluded the class of symmetric Lie algebras fulfilling $[\mathfrak{m}, \mathfrak{m}] = 0$, cf. case 3 in Proposition 2.5.5. Symmetric Lie algebras with this property are said to be of Euclidean type. By point 2 of Proposition 2.5.10, they are necessarily flat. One can show that if G/H is simply connected,

---

[37] See Sect. 8 of Chap. XI in [381] or Sect. 2 of Chap. V in [293].

[38] By definition, the rank is the dimension of some maximal Abelian subspace of $\mathfrak{m}$. Any two maximal Abelian subspaces of $\mathfrak{m}$ are $\mathrm{Ad}(H)$-conjugate.

Table 2.1 Classical symmetric spaces of types I and III

Type I	Type III	Dimension	Rank
$\mathrm{SU}(n)/\mathrm{SO}(n)$	$\mathrm{SL}(n,\mathbb{R})/\mathrm{SO}(n)$	$(n-1)(n+2)/2$	$n-1$
$\mathrm{SU}(2n)/\mathrm{Sp}(n)$	$\mathrm{SL}(n,\mathbb{H})/\mathrm{Sp}(n)$	$(n-1)(2n+1)$	$n-1$
$\mathrm{SU}(p+q)/S(\mathrm{U}(p)\times\mathrm{U}(q))$	$\mathrm{SU}(p,q)/S(\mathrm{U}(p)\times\mathrm{U}(q))$	$2pq$	$\min(p,q)$
$\mathrm{SO}(p+q)/(\mathrm{SO}(p)\times\mathrm{SO}(q))$	$\mathrm{SO}_0(p,q)/(\mathrm{SO}(p)\times\mathrm{SO}(q))$	pq	$\min(p,q)$
$\mathrm{SO}(2n)/\mathrm{U}(n)$	$\mathrm{SO}^*(2n)/\mathrm{U}(n)$	$n(n-1)$	$[n/2]$
$\mathrm{Sp}(n)/\mathrm{U}(n)$	$\mathrm{Sp}(n,\mathbb{R})/\mathrm{U}(n)$	$n(n+1)$	n
$\mathrm{Sp}(p+q)/(\mathrm{Sp}(p)\times\mathrm{Sp}(q))$	$\mathrm{Sp}(p,q)/(\mathrm{Sp}(p)\times\mathrm{Sp}(q))$	$4pq$	$\min(p,q)$

Table 2.2 Classical symmetric spaces of types II and IV. For type II, see Proposition X.1.2 and Sect. IV.6 in [293]

Type II	Type IV	Dimension	Rank
$\mathrm{SU}(n+1)$	$\mathrm{SL}(n+1,\mathbb{C})/\mathrm{SU}(n+1)$	$n(n+2)$	n
$\mathrm{Spin}(2n+1)$	$\mathrm{SO}(2n+1,\mathbb{C})/\mathrm{SO}(2n+1)$	$n(2n+1)$	n
$\mathrm{Sp}(n)$	$\mathrm{Sp}(n,\mathbb{C})/\mathrm{Sp}(n)$	$n(2n+1)$	n
$\mathrm{Spin}(2n)$	$\mathrm{SO}(2n,\mathbb{C})/\mathrm{SO}(2n)$	$n(2n-1)$	n

then a symmetric space of this type is isometric to some Euclidean space $\mathbb{R}^n$. Clearly, $\mathbb{R}^n$ itself provides the simplest example, with the symmetry at the origin given by $s : \mathbf{x} \to -\mathbf{x}$. ◆

Next, we show that Riemannian symmetric spaces provide Riemannian manifolds of certain types met before. Recall that if $(\mathfrak{g},\lambda)$ is irreducible, then $\mathfrak{g}$ is necessarily semisimple and thus, the Killing form k is non-degenerate. As already noted, this implies

$$\eta(\mathbf{x},\mathbf{z}) = -c\,\mathsf{k}^{\mathfrak{m}}(\mathbf{x},\mathbf{z})\,, \quad \mathbf{x},\mathbf{z}\in\mathfrak{m}\,, \tag{2.5.18}$$

for some $c\in\mathbb{R}$, $c\neq 0$, cf. (2.5.7). Recall from point 2 of Proposition 2.5.10 that the curvature mapping $\mathscr{R}$ is given by the mapping F, cf. formula (2.5.10). Substituting $\mathbf{x} = F(\mathbf{u},\mathbf{v})\mathbf{w}$ into (2.5.18) and using the $\mathrm{ad}(\mathfrak{h})$-invariance of k, we obtain

$$\eta(F(\mathbf{u},\mathbf{v})\mathbf{w},\mathbf{z}) = c\,\mathsf{k}^{\mathfrak{m}}([[\mathbf{u},\mathbf{v}],\mathbf{w}],\mathbf{z}) = c\,\mathsf{k}^{\mathfrak{h}}([\mathbf{u},\mathbf{v}],[\mathbf{w},\mathbf{z}])\,. \tag{2.5.19}$$

Setting $\mathbf{x}=\mathbf{u}=\mathbf{z}$ and $\mathbf{y}=\mathbf{v}=\mathbf{w}$ in (2.5.19), we immediately obtain the following formula for the sectional curvature:

$$\eta(F(\mathbf{x},\mathbf{y})\mathbf{y},\mathbf{x}) = -c\,\mathsf{k}^{\mathfrak{h}}([\mathbf{x},\mathbf{y}],[\mathbf{x},\mathbf{y}])\,. \tag{2.5.20}$$

This yields useful formulae for the Ricci tensor and for the scalar curvature. For any orthonormal basis $\{\mathbf{e}_i\}$ of $\mathfrak{m}$,

$$\mathsf{Ric}(\mathbf{e}_i, \mathbf{e}_j) = -\sum_k \eta([[\mathbf{e}_k, \mathbf{e}_i], \mathbf{e}_j], \mathbf{e}_k)\,, \quad \mathsf{Sc} = -\sum_{k,l} \eta([[\mathbf{e}_k, \mathbf{e}_l], \mathbf{e}_l], \mathbf{e}_k)\,. \tag{2.5.21}$$

Proposition 2.5.26 *Let (M, g) be an irreducible Riemannian globally symmetric space and let $(\mathfrak{g}, \lambda)$ be the corresponding irreducible orthogonal symmetric Lie algebra.*

1. *If $(\mathfrak{g}, \lambda)$ is of compact type, then (M, g) is a compact Einstein manifold with non-negative sectional curvature and positive definite Ricci tensor.*
2. *If $(\mathfrak{g}, \lambda)$ is of non-compact type, then (M, g) is a simply connected Einstein manifold with non-positive sectional curvature and negative definite Ricci tensor. Moreover, M is diffeomorphic to a Euclidean space.*

Proof Let $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$ be the canonical decomposition. By Theorem 2.5.24, $G = I_0(M)$ acts transitively and effectively on M and g is G-invariant. Since $k^{\mathfrak{h}}$ is negative definite, the statements about the sectional curvature K follow immediately from (2.5.20). Since the Ricci tensor Ric is a symmetric $\mathrm{ad}(\mathfrak{h})$-invariant bilinear form on $\mathfrak{m}$ and since $\mathrm{ad}(\mathfrak{h})$ acts irreducibly on $\mathfrak{m}$, Ric must be proportional to the metric, that is, (M, g) is an Einstein space.

1. Let $(\mathfrak{g}, \lambda)$ be of compact type. Then, K is non-negative and, thus, Ric is semi-positive definite. Since M is Einstein, Ric is either positive definite or zero. But if Ric is zero, then (2.3.27) implies that K must also be zero, which contradicts the non-degeneracy of k and, thus, the irreducibility of $(\mathfrak{g}, \lambda)$. Finally, since $\mathsf{k}^{\mathfrak{m}}$ is negative definite, k is negative definite and, since $\mathfrak{g}$ is semisimple, G is compact. Thus, M is compact.

2. Let $(\mathfrak{g}, \lambda)$ be of non-compact type. Then, by similar arguments, M is Einstein with negative definite Ricci tensor. The remaining statement follows from Theorem 8.3 in Chap. VIII of [381]. ■

In the remainder of this section, we present the symmetric space structure of a few of the types in Table 2.1 explicitly. By Theorem 2.5.24, it is enough to exhibit the corresponding symmetric Lie algebra structure. For a much more detailed discussion of examples we refer to Chap. XI of [381] and to [692]. We leave it to the reader to check the statements below (Exercise 2.5.6).

Example 2.5.27

1. Consider type I in lines 3, 4, and 7 of Table 2.1. Lines 3 and 7 correspond to the Graßmann manifolds

$$G_{\mathbb{K}}(k, n) \cong \mathrm{U}_{\mathbb{K}}(n)/(\mathrm{U}_{\mathbb{K}}(n-k) \times \mathrm{U}_{\mathbb{K}}(k))\,, \quad \mathbb{K} = \mathbb{C}, \mathbb{H}\,,$$

and line 4 corresponds to the Graßmann manifold of oriented subspaces of $\mathbb{R}^{p+q}$.[39] The corresponding symmetric Lie algebra is given by

$$\mathfrak{u}_{\mathbb{K}}(p+q) = (\mathfrak{u}_{\mathbb{K}}(p) \oplus \mathfrak{u}_{\mathbb{K}}(q)) \oplus \mathfrak{m}\,,$$

where

$$\begin{aligned} \mathfrak{u}_{\mathbb{K}}(p) \oplus \mathfrak{u}_{\mathbb{K}}(q) &= \left\{ \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} \in \mathfrak{u}_{\mathbb{K}}(p+q) : A \in \mathfrak{u}_{\mathbb{K}}(p)\,,\ B \in \mathfrak{u}_{\mathbb{K}}(q) \right\}\,, \\ \mathfrak{m} &= \left\{ \begin{bmatrix} 0 & -X^{\dagger} \\ X & 0 \end{bmatrix} \in \mathfrak{u}_{\mathbb{K}}(p+q) : X \in L(\mathbb{K}^p, \mathbb{K}^q) \right\}\,. \end{aligned}$$

The action of $\mathrm{Ad}(H)$ on $\mathfrak{m}$ is given by

$$X \mapsto hXk^{-1}\,, \quad h \in \mathrm{U}_{\mathbb{K}}(q)\,,\ k \in \mathrm{U}_{\mathbb{K}}(p)\,,$$

and the involutive automorphism λ acts via

$$\begin{bmatrix} A & -X^{\dagger} \\ X & B \end{bmatrix} \mapsto \begin{bmatrix} A & X^{\dagger} \\ -X & B \end{bmatrix}\,.$$

The corresponding involutive automorphism σ is given by conjugation with

$$\mathbb{1}_{p,q} = \begin{bmatrix} -\mathbb{1}_p & 0 \\ 0 & \mathbb{1}_q \end{bmatrix}\,. \tag{2.5.22}$$

2. Consider the special case $p = n$ and $q = 1$ for type I in line 4 of Table 2.1:

$$\mathrm{S}^n = S_{\mathbb{R}}(1, n+1) = \mathrm{SO}(n+1)/\mathrm{SO}(n)\,.$$

The underlying symmetric Lie algebra is given by

$$\mathfrak{o}(n+1) = \mathfrak{o}(n) \oplus \mathfrak{m}\,, \tag{2.5.23}$$

where

$$\begin{aligned} \mathfrak{o}(n) &= \left\{ \begin{bmatrix} 0 & 0 \\ 0 & A \end{bmatrix} \in \mathfrak{o}(n+1) : A \in \mathfrak{o}(n) \right\}\,, \\ \mathfrak{m} &= \left\{ \begin{bmatrix} 0 & -\mathbf{x}^{\mathrm{T}} \\ \mathbf{x} & 0 \end{bmatrix} \in \mathfrak{o}(n+1) : \mathbf{x} \in \mathbb{R}^n \right\}\,. \end{aligned}$$

Then, $\mathrm{Ad}(\mathrm{SO}(n))$ gets identified with the basic representation of $\mathrm{SO}(n)$ on $\mathbb{R}^n$ and, under the identification $\mathfrak{m} \cong \mathbb{R}^n$, the Euclidean scalar product on $\mathbb{R}^n$ yields a scalar product on $\mathfrak{m}$ which coincides with the restriction of the Killing form on $\mathfrak{o}(n+1)$ to $\mathfrak{m}$ up to the factor $-2(n-1)$. The involutive automorphisms are read off from the previous point.

---

[39] Cf. Example I/7.5.6.

3. Consider type I in line 5 of Table 2.1. One easily shows that $\mathrm{SO}(2n)/\mathrm{U}(n)$ is the space of orthogonal complex structures on the $2n$-dimensional Euclidean space.[40] Here we decompose[41]

$$\mathfrak{o}(2n) = \mathfrak{u}(n) \oplus \mathfrak{m}\,,$$

with

$$\begin{aligned}\mathfrak{u}(n) &= \left\{ \begin{bmatrix} X & Y \\ -Y & X \end{bmatrix} \in \mathfrak{o}(2n) : X, Y \in \mathfrak{gl}(n,\mathbb{R})\,,\ X = -X^{\mathrm{T}}\,,\ Y = Y^{\mathrm{T}} \right\}\,,\\ \mathfrak{m} &= \left\{ \begin{bmatrix} X & Y \\ Y & -X \end{bmatrix} \in \mathfrak{o}(2n) : X, Y \in \mathfrak{gl}(n,\mathbb{R})\,,\ X = -X^{\mathrm{T}}\,,\ Y = -Y^{\mathrm{T}} \right\}\,.\end{aligned}$$

The involutive automorphism $\lambda : \mathfrak{o}(2n) \to \mathfrak{o}(2n)$ corresponding to this decomposition is given by conjugation with the matrix

$$\mathsf{J}_0 = \begin{bmatrix} 0 & -\mathbb{1} \\ \mathbb{1} & 0 \end{bmatrix}\,.$$

4. Consider type I in line 1 of Table 2.1. Recall from Sect. 7.6 of Part I that $\mathrm{U}(n)/\mathrm{O}(n)$ is the space of Lagrangian subspaces of $\mathbb{R}^{2n}$ endowed with its canonical symplectic structure. Correspondingly, $\mathrm{SU}(n)/\mathrm{SO}(n)$ is called the space of special Lagrangian subspaces. Here, we decompose

$$\mathfrak{su}(n) = \mathfrak{o}(n) \oplus \mathfrak{m}\,,$$

with

$$\begin{aligned}\mathfrak{o}(n) &= \left\{ \begin{bmatrix} X & 0 \\ 0 & X \end{bmatrix} \in \mathfrak{su}(n) : X \in \mathfrak{gl}(n,\mathbb{R})\,,\ X = -X^{\mathrm{T}}\,,\ \operatorname{tr} X = 0 \right\}\,,\\ \mathfrak{m} &= \left\{ \begin{bmatrix} 0 & Y \\ -Y & 0 \end{bmatrix} \in \mathfrak{su}(n) : Y \in \mathfrak{gl}(n,\mathbb{R})\,,\ Y = Y^{\mathrm{T}} \right\}\,.\end{aligned}$$

Here, we have used the embedding $\mathfrak{u}(n) \subset \mathfrak{o}(2n)$ from the previous point. Under this embedding, the involutive automorphism $\lambda : \mathfrak{su}(n) \to \mathfrak{su}(n)$ is given by

$$\begin{bmatrix} X & Y \\ -Y & X \end{bmatrix} \mapsto \begin{bmatrix} X & -Y \\ Y & X \end{bmatrix}\,.$$

---

[40] Cf. Example I/7.5.5.

[41] Cf. Example 2.2.19.

5. Consider type III in line 4 of Table 2.1 with $p = 1$, that is, $M = \mathrm{SO}_0(1,n)/\mathrm{SO}(n)$. On the level of Lie algebras, we have to consider the pseudo-Euclidean space $(\mathbb{R}^{1,n}, \eta)$ with $\eta = \mathbb{1}_{1,n}$ given by (2.5.22). Then,

$$\mathfrak{o}(n,1) = \left\{X \in \mathfrak{gl}(n+1,\mathbb{R}) : X^{\mathrm{T}}\mathbb{1}_{1,n} + \mathbb{1}_{1,n}X = 0\right\} .$$

Embedding $\mathfrak{o}(n) \subset \mathfrak{o}(1,n)$ via $Y \mapsto \begin{bmatrix} 1 & 0 \\ 0 & Y \end{bmatrix}$, we obtain the canonical decomposition

$$\mathfrak{o}(1,n) = \mathfrak{o}(n) \oplus \mathfrak{m} , \quad \mathfrak{m} = \left\{ \begin{bmatrix} 0 & \mathbf{u}^{\mathrm{T}} \\ \mathbf{u} & 0 \end{bmatrix} \in \mathfrak{o}(1,n) : \ \mathbf{u} \in \mathbb{R}^n \right\} .$$

It is obvious that M may be identified with the hypersurface $H_+(1,n) \subset \mathbb{R}^{1,n}$ defined by

$$\eta(\mathbf{u},\mathbf{u}) = -1 , \quad u^0 \geq 1 .$$

Therefore, M is referred to as the hyperbolic space form of $(\mathbb{R}^{1,n}, \eta)$. ◆

Remark 2.5.28 Consider the example of the n-sphere above. By Example 1.1.18, under the identification $\mathfrak{m} \cong \mathbb{R}^n$, the bundle of orthonormal frames $O(\mathrm{S}^n)$ coincides with the principal $\mathrm{SO}(n)$-bundle $\mathrm{SO}(n+1) \to \mathrm{SO}(n+1)/\mathrm{SO}(n)$ and, by Proposition 2.5.10, the Levi-Civita connection on S^n with respect to the natural metric coincides with the $\mathrm{SO}(n+1)$-invariant canonical connection on this bundle. The curvature (2.5.10) reads $F(\mathbf{x},\mathbf{y}) = \mathbf{x} \wedge \mathbf{y}$. Comparing with (2.4.2), this shows that S^n has a constant sectional curvature equal to 1. ◆

For applications of the theory of symmetric spaces in this book, see Sects. 6.8 and 7.9.

Exercises

2.5.1 Prove that λ defined by (2.5.1) is an involutive Lie algebra homomorphism.

2.5.2 Prove that the decomposition (2.5.3) is orthogonal with respect to the Killing form.

2.5.3 Prove Proposition 2.5.3.

2.5.4 Prove Lemma 2.5.12.

2.5.5 Prove the following. For an involutive isometry s with isolated fixed point m, one has $s'_m = -\,\mathrm{id}$. *Hint.* Use the eigenspace decomposition of s'_m.

2.5.6 Check the statements in Example 2.5.27.

+,-.+ + +/ ! + % < B ! 2 3 ! #
< 0 3; 0#
*))) =)

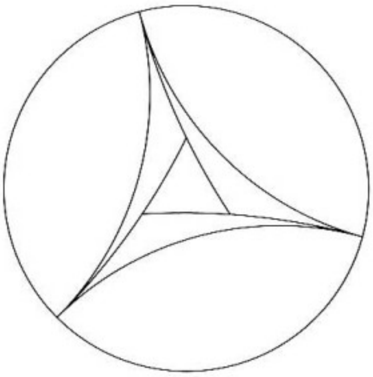

Figure 2.28. Area of general hyperbolic triangles. If you subtract a finite hyperbolic triangle from a suitable ideal triangle, you get three $\frac{2}{3}$-ideal triangles. Adding up angles and areas gives Proposition 2.4.13.

Exercise 2.4.17. What is a good definition for the area of a non-simple polygon? What is the formula for the area of a non-simple hyperbolic polygon? Compare Problem 2.3.12.

2.5. Hyperbolic Isometries

We now turn to the qualitative study of the isometries of hyperbolic space, which, as we saw in Problem 2.2.14, form a large group. We could use linear algebra for this purpose, but we'll instead use direct geometric constructions, the better to develop our hyperbolic intuition. The algebraic approach is taken in Problem 2.5.24.

We start in three dimensions. Let $g : \mathbf{H}^3 \to \mathbf{H}^3$ be an orientation-preserving isometry other than the identity. An *axis* of g is any line L that is invariant under g and on which g acts as a (possibly trivial) translation.

Proposition 2.5.1 (axis is unique). *A non-trivial orientation-preserving isometry of* $\mathbf{H}^3$ *can have at most one axis.*

Proof of 2.5.1. Suppose that L and M are distinct axes for an orientation-preserving isometry g. If g fixes both L and M pointwise, take a point x on M but not on L. Then g fixes the plane containing L and x, because it fixes three non-collinear points on it. Since g preserves orientation, it is the identity.

If, on the other hand, L is translated by g, we have $d(x, M) = d(g(x), M)$ for $x \in L$, so the function $d(x, M)$ is periodic and therefore bounded. But two distinct lines cannot remain a bounded distance from each other in both directions, since that would imply they have the same two endpoints on the sphere at infinity. $\boxed{2.5.1}$

Exercise 2.5.2. Find a non-trivial orientation-preserving isometry of $\mathbf{H}^3$ that leaves invariant more than one line.

Any orientation-preserving isometry of $\mathbf{E}^3$ is either a translation, a rotation about some axis, or a *screw motion*, that is, a rotation followed by a translation along the axis of rotation. (Exercise 2.5.6 asks you to prove this.) The situation in $\mathbf{H}^3$ is somewhat richer, and has its own special terminology.

If a non-trivial orientation-preserving isometry g of $\mathbf{H}^3$ has an axis that is fixed pointwise, it is called an *elliptic* isometry, or a *rotation* about its axis. In this case the orbit of a point p off the axis—the set of points $g^k(p)$, for $k \in \mathbf{Z}$—lies on a circle around the axis.

If g has an axis that is translated by a non-trivial amount, it is called *hyperbolic*. There are two possibilities here: the orbit of a point off the axis may lie on a plane, always on the same side of the axis; it is in fact contained in an equidistant curve, like the one shown in Figure 1.10. In this case we say that g is a *translation*. Alternatively, the orbit can be the vertices of a polygonal helix centered around the axis; in this case g is a screw motion, as can be seen by applying a compensatory translation. (An alternate usage of the word "hyperbolic" specializes it to what we're calling translations, in which case "loxodromic" designates a screw motion. This distinction is not very useful, and we'll not adopt it.)

By the proposition above, no transformation can be at the same time elliptic and hyperbolic. But there are isometries that are neither elliptic nor hyperbolic: they are called *parabolic*. For instance, any isometry of $\mathbf{H}^3$ that appears as a Euclidean translation parallel to the bounding plane in the upper half-space model is parabolic.

The proof of the next proposition describes a geometric construction to locate the axis of a non-trivial orientation-preserving isometry of $\mathbf{H}^3$, if it has one. We will need an elementary fact about pairs of lines; recall (cf. Exercise 2.3.10) that two lines in $\mathbf{H}^3$ are called *parallel* if they have a common endpoint on S^2_∞, or, equivalently, if the distance between them approaches zero at one end or the other.

Lemma 2.5.3 (common perpendicular for lines in $\mathbf{H}^3$). *Two distinct lines in $\mathbf{H}^3$ are either parallel, or they have a unique common perpendicular.*

Proof of 2.5.3. Let the lines be X and Y, and consider the distance function $d(x, y)$ between points $x \in X$ and $y \in Y$. If the lines are not parallel, this function goes to ∞ as either x or y or both go to ∞; therefore it has a minimum, attained at points x_0 and y_0.

If the minimum is zero, that is, if the lines cross, any common perpendicular must go through the intersection point, otherwise we'd have a triangle with two right angles, which is impossible by Proposition 2.4.13. As there is a unique line orthogonal to the plane spanned by X and Y and passing through their intersection point, the lemma is proved in this case. (Notice that this part is false in dimension greater than three.)

If the minimum distance is not zero, the line between x_0 and y_0 is a perpendicular by Exercise 2.4.6. If there were another common perpendicular, we'd obtain a quadrilateral in space with all right angles (although conceivably its sides could cross). Subdividing the quadrilateral by a diagonal, we'd get two plane triangles, whose angles add to at least 2π; this is again impossible. $\boxed{2.5.3}$

If L is a line in $\mathbf{H}^3$, we denote by r_L the reflection in L, which is the rotation of π about L.

Proposition 2.5.4 (finding the axis). *Any non-trivial orientation-preserving isometry g of $\mathbf{H}^3$ can be written in the form $g = r_L \circ r_M$, where the lines L and M are parallel, secant or neither depending on whether g is parabolic, elliptic or hyperbolic. The axis of g is the common perpendicular of L and M, if it exists.*

Proof of 2.5.4. Take any point p such that $g(p) \neq p$. If $g^2(p) = p$, the midpoint q of the line segment $\overline{p\,g(p)}$ is fixed by g, and the plane through q perpendicular to the line $p\,g(p)$ is invariant. Since g reverses the orientation of this plane, it must act on it as a reflection, fixing a line K. Therefore, $g = r_K$ is elliptic of order two. In this case, one can take L and M to be two orthogonal lines, both orthogonal to K at a point $x \in K$.

If p, $g(p)$ and $g^2(p) \neq p$ are collinear, p is fortuitously on the axis of g, and g is hyperbolic. We can replace p by some other point not on this axis, and reduce to the next case.

In the remaining case, we let M be the bisector of the angle $p\,g(p)g^2(p)$, so that r_M fixes $g(p)$ and interchanges p with $g^2(p)$. To define L, we look at the dihedron with edge $g(p)g^2(p)$ whose sides contain p and $g^3(p)$, respectively, and take L as the line that bisects this dihedron and is also a perpendicular bisector of the segment $\overline{p\,g(p)}$, as shown in Figure 2.29. (The dihedral angle along $p\,g(p)$ may be 0 or π, but this doesn't cause problems.) By symmetry, r_L interchanges $g(p)$ with $g^2(p)$, and also p with $g^3(p)$. Therefore, $r_L \circ r_M$ sends p to $g(p)$, $g(p)$ to $g^2(p)$ and $g^2(p)$ to $g^3(p)$. Since $r_L \circ r_M$ and g agree at three non-collinear points, they agree on the whole plane containing these three points. Therefore they agree everywhere, since they both preserve orientation.

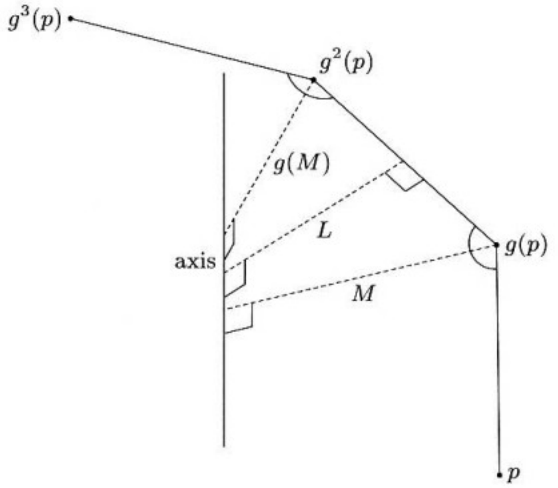

Figure 2.29. The axis of a three-dimensional isometry. The axis of an isometry g of $\mathbf{H}^3$ may be constructed, in the generic case, by connecting the orbit of a point p in a polygonal path.

The common perpendicular of L and M, if it exists, is the axis of g, because it is invariant under $r_L \circ r_M$, which acts on it as a translation. The sorting into cases now follows from Lemma 2.5.3 and from the definitions of parabolic, elliptic and hyperbolic transformations. $\boxed{2.5.4}$

Exercise 2.5.5. Using the decomposition of Proposition 2.5.4, show that any parabolic orientation-preserving isometry of $\mathbf{H}^3$ is conjugate to a Euclidean translation of the upper half-space model.

Exercise 2.5.6 (isometries of $\mathbf{E}^3$). Prove that any orientation-preserving isometry of $\mathbf{E}^3$ is either a translation, a rotation or a screw motion. Give a geometric construction for the axis of the transformation, in the latter two cases.

The proof of Proposition 2.5.4 also applies to any two-dimensional isometry, whether it preserves or reverses orientation, since such an isometry can be extended to a three-dimensional orientation-preserving isometry. But three-dimensional orientation-reversing isometries, or arbitrary isometries in higher dimension, can have fixed-point sets other than lines: for example, a reflection in a plane of any dimension, or a compound rotation about a plane of any even codimension. Any transformation with a fixed point in hyperbolic space is still called *elliptic*. An isometry that translates an axis is *hyperbolic*, and as before the axis is unique. Any other non-trivial transformation is *parabolic*.

Problem 2.5.7 (orientation-reversing isometries of $\mathbf{H}^3$). Classify orientation-reversing isometries of $\mathbf{H}^3$. One approach is to modify the construction above, and express an orientation-reversing isometry as the composition of reflection in a plane with reflection in a line. Another approach is to exploit the fact that the square of an orientation-reversing isometry is orientation-preserving. Which isometries of $\mathbf{H}^3$ have orientation-reversing square roots? What are the square roots of the identity? What are the square roots of other isometries that have them?

To study isometries in arbitrary dimension, we continue with the geometric point of view and develop a method, based on the convexity of the distance function, that will be important later, because it can be adapted to other negatively curved metrics on $\mathbf{R}^n$. For another, algebraic, method, see Problem 2.5.24.

A *convex function* on a Riemannian manifold is a real-valued function f such that, for every geodesic γ, parametrized at a constant speed, the induced function $f \circ \gamma$ is convex. In other words, for every $t \in (0, 1)$,

$$f \circ \gamma(t) \leq t f \circ \gamma(0) + (1 - t) f \circ \gamma(1).$$

If the inequality is strict for all non-constant γ, we say that f is *strictly convex*.

The *product* of two Riemannian manifolds is the product manifold, with the Riemannian metric that is the sum of the metrics on the factors. It is easy to show that a curve in the product manifold is a geodesic (parametrized at constant speed) if and only if each projection in a factor is also a geodesic (parametrized at constant speed). In particular, a geodesic in $\mathbf{H}^n \times \mathbf{H}^n$ is a curve each of whose projections is a hyperbolic line or a point.

Theorem 2.5.8 (distance function is convex). *The distance function $d(x, y)$, considered as a map $d : \mathbf{H}^n \times \mathbf{H}^n \to \mathbf{R}$, is convex.*

+,-.+ + +/ ! + % 2 !6 3 ! # 9 N
! ! 6 3 0#
+,-.+ + +/ ! + % % % &"%&) 0#
* = : 0 0 ! 4 1 94

Chapter 1
Introduction

The field of quantum information encompasses the study of the representation, storing, processing and accessing of information by quantum mechanical systems. The field grew from the investigations of the physical limits to computation initiated by Bennett and Landauer. One of the first questions studied was whether quantum mechanics imposes any limits on what a computer can do, and it was shown by Feynman that it does not. Earlier work by Paul Benioff had explored the possibilities of quantum Turing machines. Shortly after Feynman's work, David Deutsch realized that not only is quantum mechanics not a problem for computation, it can also be an advantage. The major breakthrough in the field was Peter Shor's factoring algorithm, which showed that a quantum computer can find the prime factors of integers in a time that scales as a polynomial of the size (number of digits) of the integer.

1.1 The Qubit

The basic unit of classical information is the bit, which can be 0 or 1. The corresponding object in quantum information is the qubit, which is a two-level quantum system. The two levels are often denoted by $|0\rangle$ and $|1\rangle$, which correspond to logical 0 and 1, respectively. Natural physical systems that can be used to represent a qubit are electronic or nuclear spins, and the polarization of a photon. The key difference between a bit and a qubit, is that the latter can exist in a superposition state

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle, \tag{1.1}$$

while the former is definitely either 0 or 1. This leads to significant differences in what can be done with information represented by bits and that represented by qubits.

© Springer Nature Switzerland AG 2021

J. A. Bergou et al., *Quantum Information Processing*, Graduate Texts in Physics,
https://doi.org/10.1007/978-3-030-75436-5_1

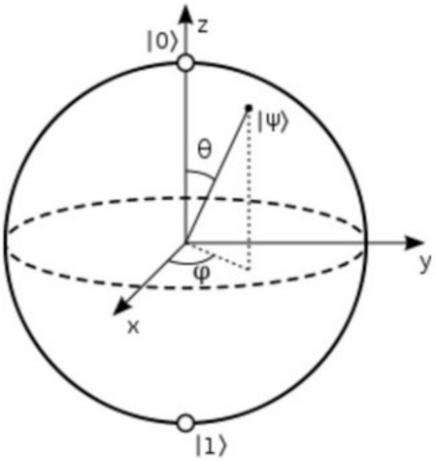

Fig. 1.1 Bloch sphere representation of the qubit in Eq. (1.2)

A convenient representation of the state of a qubit is given by the Bloch sphere. The state is parameterized by two angles, $0 \le \theta \le \pi$ and $0 \le \phi < 2\pi$,

$$|\psi\rangle = \cos(\theta/2)|0\rangle + e^{i\phi}\sin(\theta/2)|1\rangle. \tag{1.2}$$

It is represented by a point on the unit sphere whose polar angle is θ and whose azimuthal angle is ϕ. That is, the vector from the origin to the point representing the state makes an angle of θ with the z axis, and its component in the x-y plane makes an angle of ϕ with the x axis. The state $|0\rangle$ is the North Pole of the sphere, and the state $|1\rangle$ is the South Pole. The Bloch sphere of a qubit is shown in Fig. 1.1. The Bloch sphere is often useful in illustrating how the state of a qubit is changed by a quantum map.

The state of N qubits is spanned by the tensor product basis

$$\begin{aligned}|0\rangle \otimes \ldots |0\rangle \otimes |0\rangle &= |0\ldots 00\rangle\\ |0\rangle \otimes \ldots |0\rangle \otimes |1\rangle &= |0\ldots 01\rangle\\ &\vdots\\ |1\rangle \otimes \ldots |1\rangle \otimes |1\rangle &= |1\ldots 11\rangle \end{aligned} \tag{1.3}$$

Note that we have expressed these states as $|x\rangle$, where x is an N-digit binary number. The most general N-qubit state can be expressed as

$$|\Psi\rangle = \sum_{x=0}^{2^N-1} c_x|x\rangle. \tag{1.4}$$

+,-.+ + +/ ! + %) % 0#
* " !0 ! D!

1.1 PHOTON POLARIZATION

The electric/magnetic field of plane linearly polarized waves is described as follows [13]:

$$\boldsymbol{A}(\boldsymbol{r},t)=\boldsymbol{p}A_0\exp[j(\omega t-\boldsymbol{k}\cdot\boldsymbol{r})],\quad \boldsymbol{A}\in\{\boldsymbol{E},\boldsymbol{H}\} \tag{1.1}$$

where $\boldsymbol{E}(\boldsymbol{H})$ denotes the electric (magnetic) field, $\boldsymbol{p}$ denotes the polarization orientation, $\boldsymbol{r}=x\boldsymbol{e}_x+y\boldsymbol{e}_y+z\boldsymbol{e}_z$ is the position vector, and $\boldsymbol{k}=k_x\boldsymbol{e}_x+k_y\boldsymbol{e}_y+k_z\boldsymbol{e}_z$ denotes the wave propagation vector whose magnitude is $k=2\pi/\lambda$ (λ is the operating wavelength). For the x-polarization waves ($\boldsymbol{p}=\boldsymbol{e}_x$, $\boldsymbol{k}=k\boldsymbol{e}_z$), Eq. (1.1) becomes:

$$\boldsymbol{E}_x(z,t)=\boldsymbol{e}_xE_{0x}\cos(\omega t-kz), \tag{1.2}$$

while for y-polarization ($\boldsymbol{p}=\boldsymbol{e}_y$, $\boldsymbol{k}=k\boldsymbol{e}_z$) it becomes:

$$\boldsymbol{E}_y(z,t)=\boldsymbol{e}_yE_{0y}\cos(\omega t-kz+\delta), \tag{1.3}$$

where δ is the relative phase difference between the two orthogonal waves. The resultant wave can be obtained by combining Eqs. (1.2) and (1.3) as follows:

$$\boldsymbol{E}(z,t)=\boldsymbol{E}_x(z,t)+\boldsymbol{E}_y(z,t)=\boldsymbol{e}_xE_{0x}\cos(\omega t-kz)+\boldsymbol{e}_yE_{0y}\cos(\omega t-kz+\delta). \tag{1.4}$$

The *linearly polarized* wave is obtained by setting the phase difference to an integer multiple of 2π, namely $\delta=m\cdot 2\pi$:

$$\begin{aligned}&\boldsymbol{E}(z,t)=(\boldsymbol{e}_xE\cos\theta+\boldsymbol{e}_yE\sin\theta)\cos(\omega t-kz);\quad E=\sqrt{E_{0x}^2+E_{0y}^2},\\&\theta=\tan^{-1}\frac{E_{0y}}{E_{0x}}.\end{aligned} \tag{1.5}$$

By ignoring the time-dependent term, we can represent the linear polarization as shown in Fig. 1.1A. On the other hand, if $\delta\neq m\cdot 2\pi$, the *elliptical polarization* is obtained. From Eqs. (1.2) and (1.3), by eliminating the time-dependent term we obtain the following equation of ellipse:

$$\left(\frac{E_x}{E_{0x}}\right)^2+\left(\frac{E_y}{E_{0y}}\right)^2-2\frac{E_x}{E_{0x}}\frac{E_y}{E_{0y}}\cos\delta=\sin^2\delta,\tan 2\phi=\frac{2E_{0x}E_{0y}\cos\delta}{E_{0x}^2-E_{0y}^2}, \tag{1.6}$$

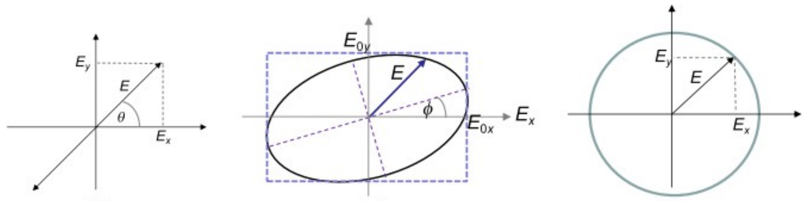

(A) (B) (C)

FIGURE 1.1

Various forms of polarizations: (A) linear polarization, (B) elliptic polarization, and (C) circular polarization.

which is shown in Fig. 1.1B. By setting $\delta = \pm\pi/2, \pm 3\pi/2, \ldots$ the equation of ellipse becomes:

$$\left(\frac{E_x}{E_{0x}}\right)^2 + \left(\frac{E_y}{E_{0y}}\right)^2 = 1. \tag{1.7}$$

By setting further $E_{0x} = E_{0y} = E_0$; $\delta = \pm\pi/2, \pm 3\pi/2, \ldots$ the equation of ellipse becomes the circle:

$$E_x^2 + E_y^2 = 1, \tag{1.8}$$

and corresponding polarization is known as *circular polarization* (Fig. 1.1C). *The right circularly polarized* wave is obtained for $\delta = \pi/2 + 2m\pi$:

$$\boldsymbol{E} = E_0[\boldsymbol{e}_x \cos(\omega t - kz) - \boldsymbol{e}_y \sin(\omega t - kz)]. \tag{1.9}$$

Otherwise, for $\delta = -\pi/2 + 2m\pi$, the polarization is known as left circularly polarized.

Very often, the *Jones vector representation* of the polarization wave is used:

$$\boldsymbol{E}(t) = \begin{bmatrix} E_x(t) \\ E_y(t) \end{bmatrix} = E \begin{bmatrix} \sqrt{1-\kappa} \\ \sqrt{\kappa} e^{j\delta} \end{bmatrix} e^{j(\omega t - kz)}, \tag{1.10}$$

where κ is the power-splitting ratio between states of polarizations (SOPs), with the complex phasor term being typically omitted in practice.

Another interesting representation is *Stokes vector representation*:

$$\boldsymbol{S}(t) = \begin{bmatrix} S_0(t) \\ S_1(t) \\ S_2(t) \\ S_3(t) \end{bmatrix}, \tag{1.11}$$

where the parameter S_0 is related to the optical intensity by:

$$S_0(t) = |E_x(t)|^2 + |E_y(t)|^2. \tag{1.12a}$$

The parameter $S_1 > 0$ is related to the preference for horizontal polarization and is defined by:

$$S_1(t) = |E_x(t)|^2 - |E_y(t)|^2. \tag{1.12b}$$

The parameter $S_2 > 0$ is related to the preference for $\pi/4$ SOP:

$$S_2(t) = E_x(t)E_y^*(t) + E_x^*(t)E_y(t). \tag{1.12c}$$

Finally, the parameter $S_3 > 0$ is related to the preference for right circular polarization and is defined by:

$$S_3(t) = j\left[E_x(t)E_y^*(t) - E_x^*(t)E_y(t)\right]. \tag{1.12d}$$

The parameter S_0 is related to other Stokes parameters by:

$$S_0^2(t) = S_1^2(t) + S_2^2(t) + S_3^2(t). \tag{1.13}$$

The *degree of polarization* is defined by:

$$p = \frac{\left[S_1^2 + S_2^2 + S_3^2\right]^{1/2}}{S_0}, \quad 0 \le p \le 1 \tag{1.14}$$

For $p = 1$ the polarization does not change in time. The Stokes vector can be represented in terms of Jones vector parameters as:

$$s_1 = 1 - 2\kappa, \;\; s_2 = 2\sqrt{\kappa(1-\kappa)}\cos\delta, \;\; s_3 = 2\sqrt{\kappa(1-\kappa)}\sin\delta. \tag{1.15}$$

After *normalization* with respect to S_0, the normalized Stokes parameters are given by:

$$s_i = \frac{S_i}{S_0}; \;\; i = 0, 1, 2, 3. \tag{1.16}$$

If the normalized Stokes parameters are used, the polarization state can be represented as a point on a *Poincaré sphere*, as shown in Fig. 1.2. The points located at the opposite sides of the line crossing the center represent the orthogonal polarizations.

The polarization ellipse is very often represented in terms of ellipticity and azimuth, which are illustrated in Fig. 1.3. The ellipticity is defined by the ratio of half-axes lengths. The corresponding angle is called the ellipticity angle and denoted by ε. The small ellipticity means that the polarization ellipse is highly elongated, while for zero ellipticity the polarization is linear. For $\varepsilon = \pm\pi/4$, the polarization is circular. For $\varepsilon > 0$, the polarization is right elliptical. On the other hand, the azimuth angle η defines the orientation of the main axis of ellipse with respect to E_x.

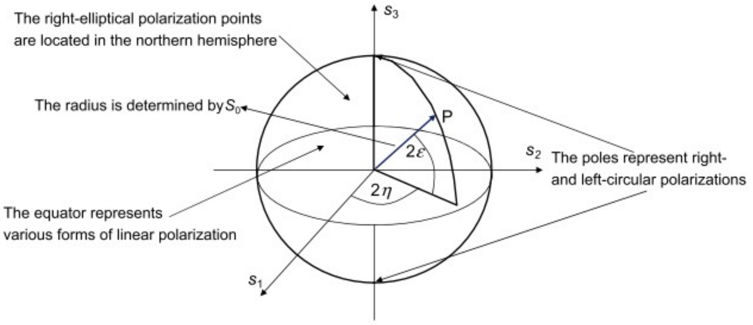
The right-elliptical polarization points are located in the northern hemisphere
The radius is determined by S_0
The equator represents various forms of linear polarization
s_3
s_2
s_1
P
2ε
2η
The poles represent right- and left-circular polarizations

FIGURE 1.2

Representation of polarization state as a point on a Poincaré sphere.

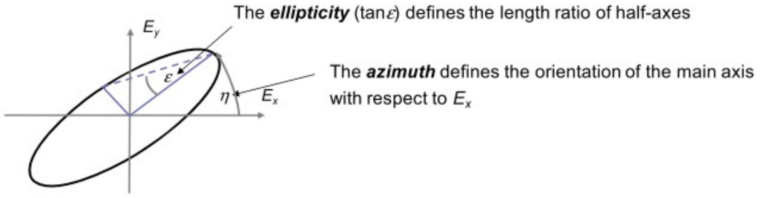
The ellipticity (tanε) defines the length ratio of half-axes
E_y
ε
η
E_x
The azimuth defines the orientation of the main axis with respect to E_x

FIGURE 1.3

The ellipticity and azimuth of the polarization ellipse.

The parameters of the polarization ellipse can be related to the Jones vector parameters by:

$$\sin 2\,\varepsilon = 2\sqrt{\kappa(1-\kappa)}\sin\delta, \quad \tan 2\,\eta = \frac{2\sqrt{\kappa(1-\kappa)}\sin\delta}{1-2\kappa}. \tag{1.17}$$

Finally, the parameters of the polarization ellipse can be related to the Stokes vector parameters by:

$$s_1 = \cos 2\eta\,\cos 2\varepsilon,\;\; s_2 = \sin 2\eta\,\cos 2\varepsilon,\;\; s_3 = \sin 2\varepsilon, \tag{1.18}$$

and corresponding geometrical interpretation is provided in Fig. 1.2.

Let us now observe the *polarizer—analyzer ensemble*, shown in Fig. 1.4. When an electromagnetic wave passes through the polarizer, it can be represented as a vector in the x—y plane, transversal to the propagation direction, as given by Eq. (1.5), where the angle θ depends on the filter orientation. By introducing the unit vector $\widehat{p} = (\cos\theta, \sin\theta)$, Eq. (1.5) can be rewritten as:

$$\boldsymbol{E} = E_0\widehat{p}\cos(\omega t - kz). \tag{1.19}$$

If $\theta = 0$ rad, the light is polarized along the x-axis, while for $\theta = \pi/2$ rad it is polarized along the y-axis. The *natural light* is *unpolarized* as it represents

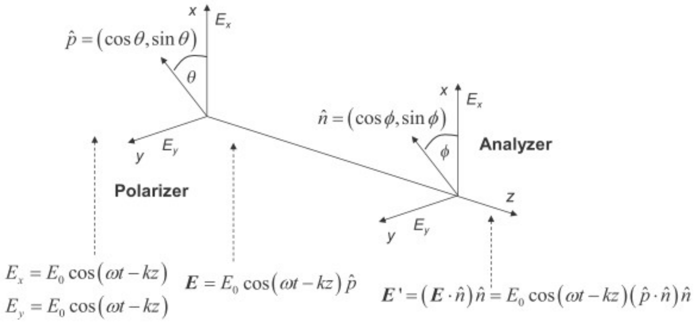
x
E_x
$\hat{p} = (\cos\theta, \sin\theta)$
θ
y
E_y
Polarizer
x
E_x
$\hat{n} = (\cos\phi, \sin\phi)$
ϕ
Analyzer
z
y
E_y
$E_x = E_0\cos(\omega t - kz)$
$E_y = E_0\cos(\omega t - kz)$
$\boldsymbol{E} = E_0\cos(\omega t - kz)\,\hat{p}$
$\boldsymbol{E}' = (\boldsymbol{E}\cdot\hat{n})\hat{n} = E_0\cos(\omega t - kz)(\hat{p}\cdot\hat{n})\hat{n}$

FIGURE 1.4

The polarizer—analyzer ensemble for the study of photon polarization.

incoherent superposition of 50% of light polarized along the x-axis and 50% of light polarized along the y-axis. After the analyzer, whose axis makes an angle ϕ with respect to the x-axis, which can be represented by the unit vector $\widehat{n} = (\cos\phi, \sin\phi)$, the output electric field is given by:

$$\begin{aligned} \boldsymbol{E}' = (\boldsymbol{E}\cdot\widehat{n})\widehat{n} &= E_0 \cos(\omega t - kz)(\widehat{p}\cdot\widehat{n})\widehat{n} \\ &= E_0 \cos(\omega t - kz)[(\cos\theta, \sin\theta)\cdot(\cos\phi, \sin\phi)]\widehat{n} \\ &= E_0 \cos(\omega t - kz)[\cos\theta\cos\phi + \sin\theta\sin\phi]\widehat{n} \\ &= E_0 \cos(\omega t - kz)\cos(\theta - \phi)\widehat{n}. \end{aligned} \tag{1.20}$$

The intensity of the analyzer output field is given by:

$$I' = |\boldsymbol{E}'|^2 = I\cos^2(\theta - \phi), \tag{1.21}$$

which is commonly referred to as *Malus' law*.

The polarization decomposition by a birefringent plate is now studied (Fig. 1.5). Experiment shows that photodetectors PD_x and PD_y are never triggered simultaneously, which indicates that an entire photon reaches either PD_x or PD_y (a photon never splits). Therefore the corresponding probabilities that a photon is detected by photodetector PD_x and PD_x can be determined by:

$$p_x = \Pr(\mathrm{PD}_x) = \cos^2\theta, \quad p_y = \Pr(\mathrm{PD}_y) = \sin^2\theta. \tag{1.22}$$

If the total number of photons is N, the number of detected photons in x-polarization will be $N_x \cong N\cos^2\theta$, and the number of detected photons in y-polarization will be $N_y \cong N\sin^2\theta$. In the limit, as $N \to \infty$ we would expect Malus' law to be obtained.

Let us now study polarization decomposition and recombination by means of birefringent plates, as illustrated in Fig. 1.6. Classical physics prediction of total probability of a photon passing the polarizer–analyzer ensemble is given by:

$$p_{\mathrm{tot}} = \cos^2\theta\cos^2\phi + \sin^2\theta\sin^2\phi \neq \cos^2(\theta - \phi), \tag{1.23}$$

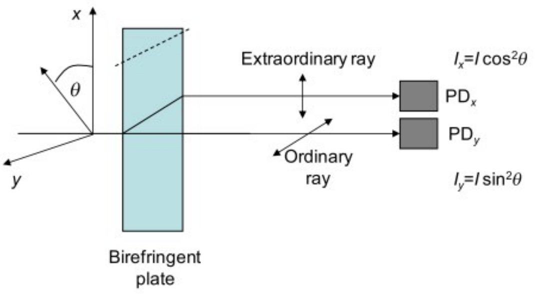

FIGURE 1.5

Polarization decomposition by a birefringent plate. *PD*, Photodetector.

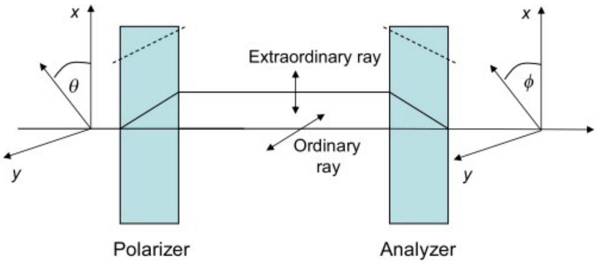

FIGURE 1.6

Polarization decomposition and recombination by a birefringent plate.

which is inconsistent with Malus' law, given by Eq. (1.21). To reconstruct the results from wave optics, it is necessary to introduce into quantum mechanics a concept of probability amplitude that α is detected as β, which is denoted as $a(\alpha \rightarrow \beta)$, and is a complex number. The probability is obtained as the squared magnitude of probability amplitude:

$$p(\alpha \rightarrow \beta) = |a(\alpha \rightarrow \beta)|^2. \tag{1.24}$$

The relevant probability amplitudes related to Fig. 1.6 are:

$$\begin{aligned} a(\theta \rightarrow x) &= \cos\theta \quad a(x \rightarrow \phi) = \cos\phi \\ a(\theta \rightarrow y) &= \sin\theta \quad a(x \rightarrow \phi) = \sin\phi \end{aligned} \tag{1.25}$$

The basic principle of quantum mechanics is to sum up the probability amplitudes for indistinguishable paths:

$$a_{\text{tot}} = \cos\theta\cos\phi + \sin\theta\sin\phi = \cos(\theta - \phi). \tag{1.26}$$

The corresponding total probability is:

$$p_{\text{tot}} = |a_{\text{tot}}|^2 = \cos^2(\theta - \phi), \tag{1.27}$$

and this result is consistent with Malus' law!

Based on the previous discussion, the *state vector* of photon polarization is given by:

$$|\psi\rangle = \begin{pmatrix} \psi_x \\ \psi_y \end{pmatrix}, \tag{1.28}$$

where ψ_x is related to x-polarization and ψ_y is related to y-polarization, with the normalization condition as follows:

$$|\psi_x|^2 + |\psi_y|^2 = 1. \tag{1.29}$$

In this representation, the x- and y-polarization photons can be represented by:

$$|x\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, |y\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{1.30}$$

and the right and left circular polarization photons are represented by:

$$|R\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ j \end{pmatrix}, |L\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -j \end{pmatrix}. \tag{1.31}$$

In Eqs. (1.28)–(1.31) we used Dirac notation to denote the column vectors (kets). In Dirac notation, with each column vector ("ket") $|\psi\rangle$, we associate a row vector ("bra") $\langle\psi|$ as follows:

$$\langle\psi| = \begin{pmatrix} \psi_x^* & \psi_y^* \end{pmatrix}. \tag{1.32}$$

The *scalar (dot) product* of ket $|\phi\rangle$ bra $\langle\psi|$ is defined by "brackets" as follows:

$$\langle\phi|\psi\rangle = \phi_x^*\psi_x + \phi_y^*\psi_y = \langle\psi|\phi\rangle^*. \tag{1.33}$$

The normalization condition can be expressed in terms of scalar product by:

$$\langle\psi|\psi\rangle = 1. \tag{1.34}$$

Based on Eqs. (1.30) and (1.31) it is evident that:

$$\langle x|x\rangle = \langle y|y\rangle = 1, \quad \langle R|R\rangle = \langle L|L\rangle = 1.$$

Because the vectors $|x\rangle$ and $|y\rangle$ are *orthogonal*, their dot product is zero:

$$\langle x|y\rangle = 0, \tag{1.35}$$

and they form the *basis*. Any state vector $|\psi\rangle$ can be written as a *linear superposition* of basis kets as follows:

$$|\psi\rangle = \begin{pmatrix} \psi_x \\ \psi_y \end{pmatrix} = \psi_x|x\rangle + \psi_y|y\rangle. \tag{1.36}$$

We can now use Eqs. (1.34) and (1.36) to derive an important relation in quantum mechanics, known as *completeness relation*. The projections of state vector $|\psi\rangle$ along basis vectors $|x\rangle$ and $|y\rangle$ are given by:

$$\langle x|\psi\rangle = \psi_x\underbrace{\langle x|x\rangle}_{1} + \psi_y\underbrace{\langle x|y\rangle}_{0} = \psi_x, \quad \langle y|\psi\rangle = \psi_x\langle y|x\rangle + \psi_y\langle y|y\rangle = \psi_y. \tag{1.37}$$

By substituting Eq. (1.37) into Eq. (1.36) we obtain:

$$|\psi\rangle = |x\rangle\langle x|\psi\rangle + |y\rangle\langle y|\psi\rangle = \underbrace{(|x\rangle\langle x| + |y\rangle\langle y|)}_{I}|\psi\rangle, \tag{1.38}$$

and from the right side of Eq. (1.38) we derive the completeness relation:

$$|x\rangle\langle x| + |y\rangle\langle y| = I. \tag{1.39}$$

The probability that the photon in state $|\psi\rangle$ will pass the x-polaroid is given by:

$$p_x = \frac{|\psi_x|^2}{|\psi_x|^2 + |\psi_y|^2} = |\psi_x|^2 = |\langle x|\psi\rangle|^2, \tag{1.40}$$

the probability amplitude of the photon in state $|\psi\rangle$ to pass the x-polaroid is:

$$a_x = \langle x|\psi\rangle. \tag{1.41}$$

Let $|\phi\rangle$ and $|\psi\rangle$ be two physical states. The probability amplitude of finding ϕ in ψ, denoted as $a(\phi \rightarrow \psi)$, is given by:

$$a(\phi \rightarrow \psi) = \langle \phi|\psi\rangle, \tag{1.42}$$

and the probability for ϕ to pass the ψ test is given by:

$$p(\phi \rightarrow \psi) = |\langle \phi|\psi\rangle|^2. \tag{1.43}$$

1.2 THE CONCEPT OF QUBIT

Based on the previous section, it can be concluded that the quantum bit, also known as qubit, lies in a 2D Hilbert space H, isomorphic to C^2-space, where C is the complex number space, and can be represented as:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}; \quad \alpha, \beta \in C; \quad |\alpha|^2 + |\beta|^2 = 1, \tag{1.44}$$

where $|0\rangle$ and $|1\rangle$ states are computational basis (CB) states, and $|\psi\rangle$ is a superposition state. If we perform the measurement of a qubit, we will get $|0\rangle$ with probability $|\alpha|^2$ and $|1\rangle$ with probability $|\beta|^2$. Measurement changes the state of a qubit from a superposition of $|0\rangle$ and $|1\rangle$ to the specific state consistent with the measurement result. If we parametrize the probability amplitudes α and β as follows:

$$\alpha = \cos\left(\frac{\theta}{2}\right), \quad \beta = e^{j\phi} \sin\left(\frac{\theta}{2}\right), \tag{1.45}$$

where θ is a polar angle and ϕ is an azimuthal angle, we can geometrically represent the qubit by the Bloch sphere (or the Poincaré sphere for the photon) as illustrated in Fig. 1.7. (Note that the Bloch sphere from Fig. 1.7 is a little bit different from that of the Poincaré sphere from Fig. 1.2.) Bloch vector coordinates are given by ($\cos\phi \sin\theta$, $\sin\phi \sin\theta$, $\cos\theta$). This Bloch vector representation is related to CB by:

$$|\psi(\theta, \phi)\rangle = \cos(\theta/2)|0\rangle + e^{j\phi} \sin(\theta/2)|1\rangle \doteq \begin{pmatrix} \cos(\theta/2) \\ e^{j\phi} \sin(\theta/2) \end{pmatrix}, \tag{1.46}$$

and Curvature
Pág. 68 — 55(68)
Chapter 3 Gravity as Curvature
obs: Gravitation-Misner
Source: Importantíssimo/
Gauss/
Gravitation Misner/
— Gravitation....(1973).pdf
— Misner-Gravitation-Princeton University Press(2017).pdf

d) Lesson in Geometry - Vol 1....pdf
Jacques Hadamard
Source: Books-018-2021/
Importantíssimo/
Geometry-Topology/
Jacques-Hadamard-Lecons-de-géométrie-élementaire/

Jacques Hadamard-
Lecons de géométrie élementaire
I. pdf (322 pág).
II. pdf (602 pág).

2) Importantíssimo/
Geometry-Topology/
Differential Geometry and
Mathematical Physics-
Part I ...pdf (473 pág.)
Part II ...pdf (837 pág.)
Pág. 127(142)
2.3 Curvature and
Holonomy

- Importantíssimo/
Three-Dimensional/
Three-Dimensional Geometry and
Topology, Vol. 1 ...pdf
Author: William P. Thurston
• Pág. 89(99) Figure 2.29
The axis of a
three-dimensional isometry.

- Books-018-2021/2021 09 14/
Gauss/Geometry/
Gravity an introduction to
Einstein's general relativity.
pdf

—

Other source:

Importantíssimo/
Gravity an Introduction.../
Gravity....(2013).pdf

—

Books-018-2021/20210914/
Quantum Information Processing
Theory and Implementation/

— 3030754359.pdf
Pág. 2(17) Fig 1.1 Bloch sphere
representation of the qubit
in Eq (1.2)

— 0128219823.pdf
Pág. 4(11) Chapter 1
1.1 Photon Polarization.
Figure 1.1.
(A) linear polarization,
(B) elliptic polarization, and
(C) circular polarization.

Important:

Pág. 06 (13)

Figure 1.2

Representation of polarization state as a point on a Poincaré sphere.

Obs: Use o book:

Books-018-2021/
Importantíssimo/
The Disc Embedding Theorem/
The Disc Embedding Theorem (2021).pdf

...

Pág. 07 (14)

Figure 1.3

The ellipticity and azimuth of the polarization ellipse

Obs: Jones vector parameters

Figure 1.4

The polarizer-analyzer ensemble for the study of photon polarization

(%% * ")
! +,-.+ + +/ ! + %%) %)" ' 0#
/1 = ! ! < 3

1 ://@ D+://@ +: C
0 0! +(0 ! I +< # ! ! ! > C < (! 6 2 3
! +
/1
1 +,-.+ + +/ ! + %% FO - 2 1 , ! #
B # 2 3 0#

1 +,-.+ + +/ ! + %% FO - 2 1
2 C! CO 0 ! : ? # 1) 0#

1% +,-.+ + +/ ! + %% FO - 2 1 0 6
8 6 2 3 , 2 # # < < - ! ! # 0 7 2 3 0 !
: ? # 1 0#

1" +,-.+ + +/ ! + %% FO - 2 1 !
0 6 8 6 2 3 , 2 # # < < - ! ! # 0 7 2 3
0 ! : ? # 1 0#

1& +,-.+ + +/ ! + %% FO - 2 1 ; 5 ?
@ C! 0 ! < # ! ! ! P > C < (! 6 2 3 !
0 ! : ? # 1 0#

1' +,-.+ + +/ ! + %% < # ! ! ! P > C
< (! 6 2 3 ! ; 0 ! : ? # 1 0#

+,-.+ + +/ ! + %% B ! 2 3 < 3;
; ! # ,00 ! % 0#
* & " ' (! 0 ! B # < #

the 6 radical planes. Its barycentric coordinates are computed from Eq. (4.14) and (4.15). The point z_0 is interior to the convex domain of the 4-order intersection. This latter is partitioned into four parts, each one being the intersection of one sphere and one trihedron originating in z_0.

These trihedra, which intersect respectively the spheres 1, 2, 3, 4, are:

$$\{z_0 \to x^{(123)}, z_0 \to x^{(124)}, z_0 \to z^{(134)}\},$$
$$\{z_0 \to x^{(123)}, z_0 \to x^{(124)}, z_0 \to z^{(234)}\},$$
$$\{z_0 \to x^{(123)}, z_0 \to x^{(134)}, z_0 \to z^{(234)}\},$$
$$\{z_0 \to x^{(124)}, z_0 \to x^{(134)}, z_0 \to z^{(234)}\}.$$

It follows that the calculation of V_{1234} and S_{1234} can be done through the calculation of the portion of the area of a sphere defined by its intersection with a trihedron having its origin in the interior of the sphere. This latter calculation is done analytically in Sect. 4.2.6 using the Gauss–Bonnet theorem. An analytical treatment of the four-sphere intersection without invoking the Gauss–Bonnet theorem was reported [27].

4.2.6 Intersection of a Sphere with a Dihedron or with Trihedron

We consider a sphere of center c and radius R intersecting a salient trihedron of origin z inside the sphere. The trihedron intersects the boundary of the sphere at x_1, x_2, x_3, these three points being ordered in the direct sense when referred to z as origin. We look for the calculation of the volume V_{st} of the intersection and for the surface S_{st} of the spherical triangle on the boundary of the sphere. This spherical triangle is bounded by the oriented arcs of circle (x_1,x_2), (x_2,x_3), (x_3,x_1). These arcs have respective radii h_{12}, h_{23}, h_{31} and respective angles b_{12}, b_{23}, b_{31}, each of these angles being in $[0;\pi]$. The trihedron intersects each of the planes respectively tangent to the sphere at x_1, x_2, x_3, thus defining the respective angles a_1, a_2, a_3, each of these angles being in $[0;\pi]$. We define c_{12}, c_{23}, c_{31} as the projections of c on the planes containing the respective triplets (x_1,z,x_2), (x_2,z,x_3), (x_3,z,x_1). We also define the angles associated to the three lens: $\theta_{ij} = \widehat{x_i, c, c_{ij}} = \widehat{x_j, c, c_{ij}}$, $(i,j) = (1,2)$ or $(i,j) = (2,3)$ or $(i,j) = (3,1)$, such that $\cos\theta_{ij}$ is positive when the vector $c_{ij} - c$ of origin c has the same sense than the direct normal to the plane of the oriented arc (x_i,x_j), and $\cos\theta_{ij}$ is negative when $c_{ij} - c$ has a sense opposite to this normal. We have also $\sin\theta_{ij} = h_{ij}/R$. The sphere has a constant Gaussian curvature and each arc of circle has a constant geodesic curvature. Thus, the Gauss–Bonnet theorem [26, 52] leads to a simple expression of S_{st} in Eq. (4.22), in which the first term in parenthesis is the spherical excess:

$$\frac{S_{st}}{R^2} = (a_1 + a_2 + a_3 - \pi) - (b_{12}\cos\theta_{12} + b_{23}\cos\theta_{23} + b_{31}\cos\theta_{31}). \qquad (4.22)$$

V_{zx} being the signed volume of the oriented tetrahedron $(x_1 - z, x_2 - z, x_3 - z)$ and V_{cx} being the signed volume of the oriented tetrahedron $(x_1 - c, x_2 - c, x_3 - c)$, the volume V_{st} is given below [37]:

$$V_{st} = \frac{R^3}{6}\left[\cos\theta_{12}\sin^2\theta_{12}(\sin b_{12} - b_{12}) \right. \\ + \cos\theta_{23}\sin^2\theta_{23}(\sin b_{23} - b_{23}) \\ \left. + \cos\theta_{31}\sin^2\theta_{31}(\sin b_{31} - b_{31})\right] + V_{zx} - V_{cx} + \frac{S_{st}R}{3}. \tag{4.23}$$

When z lies at c, we retrieve the expression given by Girard's theorem [48]: $\frac{S_{st}}{R^2}$ is the spherical excess and $V_{st} = \dfrac{S_{st}R}{3}$.

Remark: rather than looking for the validity of Eq. (4.23) from integrating S_{st} as stated in [37], we give a hint of proof as follows. We define the point x_0 at the intersection of the half line of origin c passing through z with the boundary of the sphere, and we add the three signed contributions to V_{st} due to the trihedra $\{z \to x_1, z \to x_2, z \to x_0\}$ and $\{z \to x_2, z \to x_3, z \to x_0\}$ and $\{z \to x_3, z \to x_1, z \to x_0\}$, a contribution being negative when the bounding arcs are not oriented in the conventional sense. For each trihedra, say, $\{z \to x_1, z \to x_2, z \to x_0\}$, the volume of the intersection with the sphere is $V_{sc012} - V_{cz12} - V_{c12}$, in which V_{sc012} is the volume of the domain bounded by $c - x_0$, $c - x_1$, $c - x_2$ and by the spherical triangle (x_0, x_1, x_2) intercepted by the trihedron $\{z \to x_1, z \to x_2, z \to x_0\}$ and of area computable by Eq (4.22). Despite that this domain is not bounded by a trihedron, V_{sc012} is computable by trivial integration in respect to the radius of the sphere. V_{cz12} is the volume of the tetrahedron (c, z, x_1, x_2). V_{c12} is the volume of the portion of the cone of origin c with height $R\cos\theta_{12}$ and of basis the circular segment of area $(b_{12} - \sin b_{12})(R\sin\theta_{12})^2/2$, i.e., $V_{c12} = \cos\theta_{12}\sin^2\theta_{12}(b_{12} - \sin b_{12})R^3/6$. The rest of the calculation is easy.

The case of the six trihedra listed at the end of Sect. 4.2.4 is of interest. It is such that one of the arcs, say, (x_2, x_3), lies on a great circle of the sphere: z, x_2, x_3, c, c_{12}, c_{31}, are coplanar, and c_{23} coincides with c, then $a_2 = \pi/2$, $a_3 = \pi/2$, and $\theta_{23} = \pi/2$.

Grouping each trihedra with its symmetric image through the plane of the great circle let us to the expressions of the volume V_{sd} of the intersection of a sphere and a dihedron and for the surface S_{sd} of the associated portion of spherical area. Keeping the notations used in Eqs. (4.22) and (4.23):

$$\frac{S_{sd}}{R^2} = 2a_1 - 2(b_{12}\cos\theta_{12} + b_{31}\cos\theta_{31}), \tag{4.24}$$

$$V_{sd} = \frac{R^3}{3}\left[\cos\theta_{12}\sin^2\theta_{12}(\sin b_{12} - b_{12}) \right. \\ \left. + \cos\theta_{31}\sin^2\theta_{31}(\sin b_{31} - b_{31})\,\right] + 2(V_{zx} - V_{cx}) + \left.\frac{S_{sd}R}{3}\right]. \tag{4.25}$$

Care: in Eqs. (4.24) and (4.25), b_{12} and b_{31} are the angles of the half arcs (not the full arcs) bounding the intersection of the sphere with the dihedron. It is also recalled that the arcs must be correctly oriented and that $\cos\theta_{12}$ and $\cos\theta_{31}$ can be negative.

When z is outside the sphere, the intersection with a dihedron reduces to simpler cases such as a sphere minus two lens, and the intersection with a trihedron can be reduced to intersections with appropriate dihedra or trihedra for which Eqs. (4.22)–(4.25) can be used. More generally we can extend volumes and surfaces computations to the intersections of a sphere with a tetrahedron, a convex polyhedron, and with any union of convex polyhedra. These computations are based on Eqs. (4.22)–(4.25) but are not needed in the framework of this chapter.

4.2.7 Intersections of Order 5 and Higher

By iterating Eqs. (4.12) and (4.13) until order n, the volume and the surface of the union of the n spheres can be expressed from the inclusion–exclusion principle as follows:

$$V = \sum_{1\le i\le n} V_i - \sum_{1\le i_1<i_2\le n} V_{i_1 i_2} + \sum_{1\le i_1<i_2<i_3\le n} V_{i_1 i_2 i_3} + \cdots + (-1)^{n-1} V_{i_1 i_2 \cdots i_n}, \quad (4.26)$$

$$S = \sum_{1\le i\le n} S_i - \sum_{1\le i_1<i_2\le n} S_{i_1 i_2} + \sum_{1\le i_1<i_2<i_3\le n} S_{i_1 i_2 i_3} + \cdots + (-1)^{n-1} S_{i_1 i_2 \cdots i_n}. \quad (4.27)$$

Setting $d = 3$ and $n \geq 5$, the existence of intersections of order n is deduced from Helly's theorem [14]. It means that the intersection of five spheres is not empty if and only if each of the five subsets of four spheres gives rise to a non-empty intersection. If $n \geq 6$, the intersection of six spheres is not empty if and only if each of the six subsets of five spheres gives rise to a nonempty intersection.

For practical computations, we need Theorem 4.5:

Theorem 4.5. *When $n \geq 5$ spheres in E^3 have a common non-empty intersection, there are m (with $1 \leq m \leq 3$) of these spheres such that their union J contains the intersection I of the $n - m$ remaining spheres.*

Proof. Assume initially that $n = 5$ spheres have a common non-empty intersection. From Helly's theorem, each of the five subsets of four spheres gives rise to a nonempty intersection. If one of these subsets of four spheres is such that the union of $m = 1, 2$, or three spheres contains the intersection of the $4-m$ remaining spheres, the union of these m spheres contains the intersection of the $5-m$ remaining spheres and the theorem stands. As stated in Sect. 4.2.5, a set of four spheres giving rise to a nonempty intersection which is not in one of these latter configurations is in the general configuration.

Now we need to prove the theorem when the five subsets of four spheres are in the case of a general configuration. For this latter, the four contact point pairs between three spheres are such that one of these contact points lies inside the fourth sphere

20210914$_{M}$ (20210914$_{M}$)

- Books-018-2021/Importantíssimo/Fractional-Trigonometry/
  1119139406.pdf
- The Cinderella.2 Manual..../
  The Cinderella.2.Manual....pdf
- Jürgen Richter-Gebert-Perspectives on Projective Geometry.... (2011).pdf

- Importantíssimo/Distance-Geometry/
  Distance Geometry-Theory, Methods, and Applications-(2013).pdf

Pág. 71(85)

4.2.6 Intersection of a Sphere with a Dihedron or with Trihedron.

The sphere has a constant Gaussian curvature and each arc of circle has a constante

(%" * &
! +,-.+ + +/ ! + %" 6! ! B! ! , ! 3 ! 7
! # ! 0 0 0#
* " " 9 # ! 0! ! 7 ! #

It can be further generalized to

$$\left\| \sum_{i=1}^{n} x_i \right\| \leq \sum_{i=1}^{n} \|x\| \ .$$

Now, let $\| \ \|$ be the norm derived from the above inner product. By direct calculations, one can check that the following *parallelogram identity*,

$$\|x+y\|^2 + \|x-y\|^2 = 2(\|x\|^2 + \|y\|^2) \ , \tag{2.7}$$

is valid (in any Hilbert space). However, if it is not valid for a particular norm, then we can conclude that this norm *cannot* be obtained from an inner product. The identity (2.7) implies the *parallelogram inequality*:

$$\|x+y\|^2 \leq 2(\|x\|^2 + \|y\|^2) \ , \tag{2.8}$$

which sometimes might be more useful than the triangle inequality (2.6).

A clear and brief reference to such topics can be found in (Rudin, 1976, Ch 1,2,4) or more advance sources, e.g. (Dieudonné, 1969a, Ch I–VI) and (Kantorovich and Akilov, 1982, Part I).

2.4 Euclidean spaces of matrices and their norms

Let $x, y \in \mathbb{R}^n$ be written in coordinate form as $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, respectively. Then $\mathbb{R}^n$ becomes an Euclidean space with the following (Euclidean) inner product of any two $x, y \in \mathbb{R}^n$:

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i \ . \tag{2.9}$$

The vectors x and y are called *orthogonal* when $\langle x, y \rangle = 0$. With $x = y$, (2.9) becomes $\langle x, x \rangle = \sum_{i=1}^{n} x_i^2$, which defines the Euclidean norm of $x \in \mathbb{R}^n$:

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \ . \tag{2.10}$$

The matrices in $\mathcal{M}(n)$ and $\mathcal{M}(n,p)$ can be considered vectors (points) in $\mathbb{R}^{n^2}$ and $\mathbb{R}^{np}$, respectively, with the corresponding definitions of scalar product and norm. Thus, $\mathcal{M}(n)$ and $\mathcal{M}(n,p)$ possess all of the nice properties of the Euclidean space. For example, they are complete normed spaces (as any other finite-dimensional normed space). Also, every bounded and closed set $\mathcal{U} \subset \mathbb{R}^n$ is *compact*, i.e. every infinite sequence of points $x_1, \dots, x_n, \dots \in \mathcal{U}$ contains a subsequence converging in $\mathcal{U}$. The compact sets are very useful. Indeed, consider compact $\mathcal{U} \subset \mathbb{R}^n$ and continuous $f : \mathbb{R}^n \to \mathbb{R}^p$, then

- $f(\mathcal{U}) \subset \mathbb{R}^p$ is compact;
- for $p = 1$, $f(\mathcal{U})$ is a closed interval $[a,b] \in \mathbb{R}$, and thus it reaches its minimal and maximal values, a and b, respectively;
- f is uniformly continuous on $\mathcal{U}$.

There are a plenty of other norms that can be defined in $\mathbb{R}^n$. The most popular ones are

- ℓ_1-*norm*, also known as the *sum-norm*, *Manhattan* or *taxicab norm*:

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \ ; \tag{2.11}$$

- $\ell_\infty-$*norm*, also known as the *max-norm*:

$$\|x\|_\infty = \max_{1 \le i \le n} |x_i| \ ; \tag{2.12}$$

- ℓ_p-*norm*, for $1 < p < \infty$:

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \ . \tag{2.13}$$

The following *Hölder inequality* for ℓ_p-norms is very important:

$$|\langle x, y \rangle| \le \|x\|_p \|y\|_q \ , \text{ for } \frac{1}{p} + \frac{1}{q} = 1 \ , \tag{2.14}$$

and gives the Cauchy–Schwartz inequality (2.4) for $p = q = 2$.

The good news is that all norms in $\mathbb{R}^n$ (and in any finite-dimensional space) are equivalent in the following sense. If $\| \ \|_\dagger$ and $\| \ \|_\ddagger$ are two norms, then there exist constants α and β, such that

$$\alpha\| \ \|_\dagger \le \| \ \|_\ddagger \le \beta\| \ \|_\dagger \ .$$

For example, if $x \in \mathbb{R}^n$, then

$$\begin{gathered}\|x\|_2 \le \|x\|_1 \le \sqrt{n}\|x\|_2 \ , \\ \|x\|_\infty \le \|x\|_2 \le \sqrt{n}\|x\|_\infty \ , \\ \|x\|_\infty \le \|x\|_1 \le n\|x\|_\infty \ .\end{gathered}$$

This implies that convergence and continuity in $\mathbb{R}^n$ do not depend on the adopted norm. For example, proving the continuity of the linear mapping $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^p)$ is particularly simple with respect to the max-norm (2.12) in $\mathbb{R}^n$. In fact, the continuity of A at $0_n \in \mathbb{R}^n$ is sufficient for its continuity in every point. Indeed, for any $x \in \mathbb{R}^n$, i.e. $x = \sum_1^n x_i e_i$, we have

$$\begin{aligned}\|Ax\| &= \left\| A\left(\sum_{i=1}^n x_i e_i\right)\right\| = \left\|\sum_{i=1}^n x_i A e_i\right\| \\ &\le \left(\sum_{i=1}^n \|Ae_i\|\right) \max_{1\le i\le n} |x_i| \le \lambda\|x\|_\infty \ ,\end{aligned}$$

which implies continuity at 0_n. Now, for any $y \in \mathbb{R}^p$, substitute above $x := x - y$, which leads to

$$\|Ax - Ay\| \le \lambda\|x - y\|_\infty \ ,$$

implying continuity at y. The inequality is valid for all $x, y \in \mathbb{R}^n$, and thus it implies *uniform continuity*. As the constant λ is the same for all $x, y \in \mathbb{R}^n$, then A is also *Lipschitz continuous* with *Lipschitz constant* λ. The inequality $\|Ax\| \le \lambda\|x\|_\infty$ (for all $x \in \mathbb{R}^n$) also shows that $\|A\| \le \lambda$, i.e. A is *bounded*. The opposite is also true: if A is linear and bounded, then it is continuous at 0_n.

The case $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R})$, when A is linear functional, is of particular interest. Let $x \in \mathbb{R}^n$ and $x = \sum_1^n x_i e_i$. Then, making use of the inner product (2.9), we find

$$Ax = A\left(\sum_{i=1}^n x_i e_i\right) = \sum_{i=1}^n x_i A(e_i) = \sum_{i=1}^n x_i a_i = \langle a, x\rangle \ ,$$

which gives the general form of any linear functional $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R})$. The space of all linear functionals $\mathcal{L}(\mathbb{R}^n, \mathbb{R})$ is called the *dual or conjugate space* of $\mathbb{R}^n$ and is denoted by $(\mathbb{R}^n)^*$. In general, $\mathcal{L}(\mathcal{E}, \mathbb{R})$ is denoted by $\mathcal{E}^*$. If $\mathbb{R}^n$ is Euclidean with (2.9), then $(\mathbb{R}^n)^*$ is also $n-$dimensional Euclidean space. The vector $a = (Ae_1, \ldots, Ae_n)$ is specific for every A, and can be considered as a vector from $(\mathbb{R}^n)^*$. In tensor algebra, vectors from both $\mathbb{R}^n$ and $(\mathbb{R}^n)^*$ are considered simultaneously. In order to distinguish them, the vectors from $(\mathbb{R}^n)^*$ are called *covariant*, while the "standard" vectors of $\mathbb{R}^n$ are called *contravariant*. Consider the following Cartesian product $\mathcal{F}$ of p copies of the real vector space $\mathcal{E}$ and q copies of its dual $\mathcal{E}^*$:

$$\mathcal{F} = \underbrace{\mathcal{E} \times \ldots \times \mathcal{E}}_{p} \times \underbrace{\mathcal{E}^* \times \ldots \times \mathcal{E}^*}_{q} \ .$$

Then, the linear functional (some authors prefer bi-linear functional):

$$T : \mathcal{F} \to \mathbb{R} \ (\text{or } T : (\mathcal{E})^p \times (\mathcal{E}^*)^q \to \mathbb{R}) \tag{2.15}$$

is called $(p, q)-$tensor. All such tensors (linear functionals) form a vector space denoted by $\mathcal{T}_p^q(\mathcal{E})$. The tensor theory is important part of (multi)linear algebra, e.g. (Bourbaki, 1974, Ch3) or (Kostrikin and Manin, 1997, Ch4). Details related to modern applications of tensors in data analysis and their computational aspects can be found in (Hackbusch, 2012; Landsberg, 2012). Some elements of tensor decomposition applied to MDA problems are discussed briefly in Section 10.8.

The above definitions of inner product (2.9) and norm (2.10) can be elaborated to treat $\mathcal{M}(n)$ and $\mathcal{M}(n, p)$ directly as metric or normed spaces without going through vectorization. For any $A, B \in \mathcal{M}(m, n)$, one can define

$$\langle A, B \rangle_F = \text{trace}(A^\top B) \ , \tag{2.16}$$

which is known as the *Frobenius inner product*. Then, the corresponding norm in $\mathcal{M}(n)$ and $\mathcal{M}(m, n)$ is the *Frobenius norm* defined as

$$\|A\|_F = \sqrt{\text{trace}(A^\top A)} \ , \tag{2.17}$$

and one can check that it fulfils the three conditions for a norm.

Then, the corresponding metric in $\mathcal{M}(n)$ and $\mathcal{M}(m, n)$ is defined by

$$d(A, B) = \|A - B\|_F = \sqrt{\text{trace}(A^\top B)} \ . \tag{2.18}$$

Let λ_i be the eigenvalues of $A \in \mathcal{S}(n)$. It follows from (2.17) that $\|A\|_F = \sqrt{\sum_i \lambda_i^2}$. This implies that $|\lambda| \leq \|A\|_F$ for any eigenvalue λ of A, as well as $|\mathrm{trace}(A)| \leq \sqrt{n}\,\|A\|_F$. Moreover, if $\|A\|_F < 1$, then $I_n + A > 0$ (is PD).

Such type of norms (with matrix argument) are called *matrix norms* (Horn and Johnson, 2013, 5.6). Other popular choices, apart from (2.17), are

- 1-norm of $A \in \mathcal{M}(m, n)$:

$$\|A\|_1 = \sum_{i,j} |a_{ij}| \ ; \tag{2.19}$$

- ∞-norm of $A \in \mathcal{M}(m, n)$:

$$\|A\|_\infty = \max_{i,j} |a_{ij}| \ ; \tag{2.20}$$

- p-norm of $A \in \mathcal{M}(m, n)$, for $1 < p < \infty$:

$$\|A\|_p = \max_{\|x\|_p = 1} \|Ax\|_p \ , \tag{2.21}$$

  where $\| \ \|_p$ on the right-hand side of (2.21) is the corresponding ℓ_p-norm of $n \times 1$ vectors. Note that $\|A\|_2 \leq \sqrt{\|A\|_1 \|A\|_\infty}$. The special case of $p = 2$ is very important and is also well known as the *spectral norm* of $A \in \mathcal{M}(m, n)$, because $\|A\|_2 = \sqrt{\lambda_{\max}(A^\top A)}$ or

$$\|A\|_2 = \sigma_{\max}(A), \text{ the largest singular value of } A \ . \tag{2.22}$$

The equivalence of the different matrix norms is quantified by inequalities, many of which are listed in (Horn and Johnson, 2013, 5.6.P23). Several other useful inequalities between matrix norms are provided in (Golub and Van Loan, 2013, Ch 2.3). For example, if $A \in \mathcal{M}(m, n)$, then

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{\mathrm{rank}(A)}\|A\|_2 \leq \sqrt{\min\{m, n\}}\|A\|_2 \ .$$

Further details about norms of matrices and matrix inequalities can be found in (Bhatia, 1997, Ch IV), (Horn and Johnson, 2013, Ch 5), (Magnus and Neudecker, 1988, Ch 11), as well as in (Marshall et al., 2011).

1 +,-.+ + +/ ! + %") % %)"' ' 0#
/1 2 !0 < ! !
* B

2. *Construct $T(A) \xleftarrow{\bar{n}} F \xrightarrow{g} C$ as the pullback of $T(A) \xrightarrow{T(f)} T(B) \xleftarrow{\bar{m}} C$ (cf. square (2) below); by the universal property of pullbacks, this in addition entails the existence of a morphism $A \xrightarrow{n} F$.*

Then (n, g) is the FPC of (f, m), and n is in $\mathcal{M}$.

$$\begin{tikzcd}
 & & & A \arrow[r, "f"] \arrow[d, dashed, tail, "\exists! n"] \arrow[dd, tail, bend right, "\eta_A"'] & B \arrow[d, tail, "m"] & \\
A \arrow[r, "f"] & B \arrow[d, tail, "m"] & {} \arrow[r, squiggly] & F \arrow[r, dashed, "g"] \arrow[d, "\bar{n}"] \arrow[dr, phantom, "\text{PB} \; (2)"] & C \arrow[d, "\bar{m}"] \arrow[dr, phantom, "\text{PB} \; (1)"] & B \arrow[l, tail, "m"'] \arrow[d, equal] \\
 & C & & T(A) \arrow[r, "T(f)"] & T(B) & B \arrow[l, tail, "\eta_B"']
\end{tikzcd} \tag{3}$$

This guarantee for the existence of FPCs will prove quintessential for constructing $\mathcal{M}$-multi-pushout complements, which are defined as follows:

Definition 6. *For a category* $\mathbf{C}$ *with an $\mathcal{M}$-partial map classifier, the $\mathcal{M}$-multi-pushout complement ($\mathcal{M}$-multi-POC) $\mathcal{P}(f, b)$ of a composable sequence of morphisms $A \xrightarrow{f} B$ and $B \xrightarrow{b} D$ with $b \in \mathcal{M}$ is defined as*

$$\mathcal{P}(f, b) := \{(A \xrightarrow{a} P, P \xrightarrow{d} D) \in \mathsf{mor}(\mathbf{C})^2 \mid a \in \mathcal{M} \wedge (d, b) = \mathsf{PO}(a, f)\} \,. \tag{4}$$

Proposition 2. *In a quasi-topos* $\mathbf{C}$ *and for $\mathcal{M} = \mathsf{rm}(\mathbf{C})$ the class of regular monomorphisms, let $\mathcal{P}(f, b)$ be an $\mathcal{M}$-multi-POC.*

- **Universal property of** $\mathcal{P}(f, b)$: *for every diagram such as in (5)(i) where (1) + (2) is a pushout along an $\mathcal{M}$-morphism n, and where $m = m' \circ b$ for some $m', b \in \mathcal{M}$, there exists an element (a, d) of $\mathcal{P}(f, b)$ and an $\mathcal{M}$-morphism $p \in \mathcal{M}$ such that the diagram commutes and (2) is a pushout. Moreover, for any $p' \in \mathcal{M}$ and for any other element $(a', d')$of $\mathcal{P}(f, b)$ with the same property, there exists an isomorphism $\delta \in \mathsf{iso}(\mathbf{C})$ such that $\delta \circ a = a'$ and $d' \circ \delta = d$.*
- **Algorithm to compute** $\mathcal{P}(f, b)$*:*
  1. *Construct (n, g) in diagram (5)(ii) by taking the FPC of (f, b).*
  2. *For every pair of morphisms (a, p) such that $a \in \mathcal{M}$ and $a \circ p = n$, take the pushout (1), which by universal property of pushouts induces an arrow $D \xrightarrow{e} C$; if $e \in \mathsf{iso}(\mathbf{C})$, (a, d) is a contribution to the $\mathcal{M}$-multi-POC of (f, b).*

$$\begin{tikzcd}
A \arrow[r, "f"] \arrow[d, dashed, tail, "\exists! a"] \arrow[dd, tail, bend right, "n"'] \arrow[dr, phantom, "(1)"] & B \arrow[d, tail, "b"] \arrow[dd, tail, bend left, "m"] & & A \arrow[r, "f"] \arrow[d, tail, "a"] \arrow[dd, tail, bend right, "n"'] \arrow[dr, phantom, "\text{PO} \; (1)"] & B \arrow[d, tail, "b"] \arrow[dd, tail, bend left, "m"] \\
P \arrow[r, dashed, "\exists! d"] \arrow[d, dashed, "\exists! p"] \arrow[dr, phantom, "(2)"] & D \arrow[d, dashed, tail, "m'"] & & P \arrow[r, "d"] \arrow[d, "p"] \arrow[dr, phantom, "(2)"] & D \arrow[d, "e"] \\
Q \arrow[r, "q"] & E & & F \arrow[r, "g"] & C
\end{tikzcd} \tag{5}$$

(i) (ii)

Proof. The universal property of $\mathcal{P}(f,b)$ follows from pushout-pullback decomposition: pushouts along $\mathcal{M}$-morphisms are pullbacks, so (1) + (2) is a pullback; taking the pullback (p,d) of (q,m') yields by the universal property of pullbacks a morphism a (which is unique up to isomorphism), and thus by pullback-pullback decomposition that (1) and (2) are pullbacks. By stability of $\mathcal{M}$-morphisms under pullbacks, both a and p are in $\mathcal{M}$, and finally by pushout-pullback decomposition, both (1) and (2) are pushouts. This proves that (a,d) is in $\mathcal{P}(f,b)$.

To prove that the algorithm provided indeed computes $\mathcal{P}(f,b)$, note first that by the universal property of FPCs, whenever in a diagram as in (5)(ii) we have that $D \cong C$ and $b \in \mathcal{M}$, since pushouts along $\mathcal{M}$-morphisms are pullbacks, square (1) is a pullback, which entails by the universal property of FPCs that there exists a morphism p such that $p \circ a = n$. By stability of $\mathcal{M}$-morphisms under pullbacks, we find that a must be in $\mathcal{M}$, so indeed every possible contribution to $\mathcal{P}(f,b)$ must give rise to a diagram as in (5)(ii), which proves the claim.

An example of an $\mathcal{M}$-multi-POC construction both in **SGraph** and in **Graph** is given in the diagram below. Note that in **Graph**, the $\mathcal{M}$-multi-POC does not contain the FPC contribution (since in **Graph** the pushout of the relevant span would yield to a graph with a multi-edge).

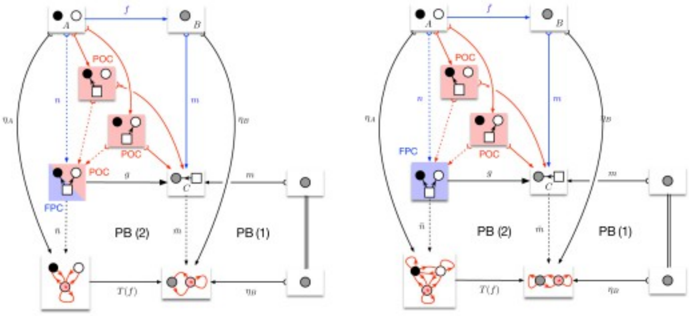

(6)

Definition 7 ($\mathcal{M}$-FPC-augmentations). *In a quasi-topos*[3] **C** *with* $\mathcal{M} = \mathsf{rm}(\mathbf{C})$*, consider a pushout square along an* $\mathcal{M}$*-morphism such as square* (1) *in the diagram below (where* $\alpha, \bar{\alpha} \in \mathcal{M}$*):*

$$\begin{tikzcd}
A \arrow[r, "a"] \arrow[d, tail, "\alpha"] \arrow[dd, tail, bend right=60, "n\circ\alpha"'] \arrow[dr, phantom, "\substack{\text{PO}\\(1)}"] & B \arrow[d, tail, "\bar{\alpha}"] \arrow[dd, tail, bend left=60, "e\circ\bar{\alpha}"] \\
C \arrow[r, "\bar{a}"] \arrow[d, tail, "n"] \arrow[dr, phantom, "(2)"] & D \arrow[d, "e"] \\
F \arrow[r, "f"] & E
\end{tikzcd} \tag{7}$$

---

[3] As demonstrated in [26, Fact 3.4], every finitary $\mathcal{M}$-adhesive category **C** possesses an (extremal $\mathcal{E}$, $\mathcal{M}$)-factorization, so if **C** is known to possess FPCs as required by the construction, this might allow to generalize the $\mathcal{M}$-FPC-PO-augmentation construction to this setting.

We define an $\mathcal{M}$-FPC augmentation (FPA) *of the pushout square* (1) *as a diagram formed from an epimorphism* $e \in \mathsf{epi}(\mathbf{C})$ *and that satisfies the following properties:*

- *The morphism* $e \circ \bar{\alpha}$ *is an* $\mathcal{M}$*-morphism.*
- $(\bar{\alpha}, id_B)$ *is a pullback of* $(e, e \circ \bar{\alpha})$*.*
- *Square* (1) + (2) *is an FPC, and the induced morphism* n *that exists*[4] *by the universal property of FPCs, here w.r.t. the FPC* $(n \circ \alpha, f)$ *of* $(a, e \circ \bar{\alpha})$*, is an* $\mathcal{M}$*-morphism.*

For a pushout as in (1)*, we denote by* $\mathsf{FPA}(\alpha, a)$ *its class of FPAs:*

$$\mathsf{FPA}(\alpha, a) := \{(n, f, e) \mid e \in \mathsf{epi}(\mathbf{C}) \wedge e \circ \bar{\alpha}, n \in \mathcal{M} \wedge (f, n \circ \alpha) = FPC(a, e \circ \bar{\alpha})\} \quad (8)$$

As induced by the properties of pushouts and of FPCs, FPAs are defined up to universal isomorphisms (in D*,* E *and* F*), and for a given pushout square there will in general exist multiple non-isomorphic such augmentations.*

The final technical ingredient for our rewriting theoretic constructions is a notion of *multi-sum* adapted to the setting of quasi-topoi, a variation on the general theory of multi-(co-)limits due to Diers [20].

Definition 8. *In a quasi-topos* $\mathbf{C}$*, the multi-sum* $\sum_{\mathcal{M}}(A, B)$ *of two objects* $A, B \in \mathsf{obj}(\mathbf{C})$ *is defined as a family of cospans of regular monomorphisms* $A \xrightarrow{f} Y \xleftarrow{g} B$ *with the following universal property: for every cospan* $A \xrightarrow{a} Z \xleftarrow{b} B$ *with* $a, b \in \mathsf{rm}(\mathbf{C})$*, there exists an element* $A \xrightarrow{f} Y \xleftarrow{g} B$ *in* $\sum_{\mathcal{M}}(A, B)$ *and a regular monomorphism* $Y \xrightarrow{y} Z$ *such that* $a = y \circ f$ *and* $b = y \circ g$*, and moreover* (f, g) *as well as* y *are unique up to universal isomorphisms.*

$$\begin{tikzcd}
A \arrow[dr, "in_A"] \arrow[ddr, "y_A"'] \arrow[dddr, bend right, "a"'] & & B \arrow[dl, "in_B"'] \arrow[ddl, "y_B"] \arrow[dddl, bend left, "b"] \\
& A+B \arrow[d, "e"] \arrow[dd, bend left, "{[a,b]}"] & \\
& Y \arrow[d, "m"] & \\
& Z &
\end{tikzcd}
\quad \rightsquigarrow \quad
\begin{tikzcd}
& & X & & \\
& & \varnothing \arrow[u, "\iota_X"] \arrow[dll, "\iota_A"'] \arrow[drr, "\iota_B"] & & \\
A \arrow[from=uurr, bend left, "x_A"'] \arrow[drr, "in_A"] \arrow[ddr, "y_A"'] \arrow[ddrrr, "p_A"] \arrow[dddrrrr, "q_A"] \arrow[ddddrr, bend right, "a"'] & & & & B \arrow[from=uull, bend right, "x_B"] \arrow[dll, "in_B"'] \arrow[ddlll, "y_B"] \arrow[ddl, "p_B"'] \arrow[ddd, bend left, "q_B"] \arrow[ddddll, bend left, "b"] \\
& & A+B \arrow[dl, "e"'] \arrow[dr, "e_P"] & & \\
& Y \arrow[ddr, "m"'] & & P \arrow[dr, "q"] \arrow[ddl, dashed, "\exists ! z"'] & \\
& & & & Q \arrow[dll, "m_Q"] \\
& & Z & &
\end{tikzcd}
\quad (9)$$

(*i*) (*ii*)

---

[4] Note that square (1) pasted with the pullback square formed by the morphisms $\overline{\alpha}, id_B, e, e \circ \bar{\alpha}$ yields a pullback square that is indeed of the right form to warrant the existence of a morphism n into the FPC square (1) + (2).

Lemma 1. *If* $\mathbf{C}$ *is a quasi-topos, the multi-sum* $\sum_{\mathcal{M}}(A, B)$ *arises from the epi-*$\mathcal{M}$*-factorization of* $\mathbf{C}$ *(for* $\mathcal{M} = \mathsf{rm}(\mathbf{C})$*; compare [29]).*

- ***Existence:*** *Let* $A \xrightarrow{in_A} A + B \xleftarrow{in_B} B$ *be the disjoint union of* A *and* B*. Then for any cospan* $A \xrightarrow{a} Z \xleftarrow{b} B$ *with* $a, b \in \mathcal{M}$*, the epi-*$\mathcal{M}$*-factorization of the induced arrow* $A + B \xrightarrow{[a,b]} Z$ *into an epimorphism* $A + B \xrightarrow{e} Y$ *and an* $\mathcal{M}$*-morphism* $Y \xrightarrow{m} Z$ *yields a cospan* $(y_A = e \circ in_A, y_B = e \circ in_B)$*, which by the decomposition property of* $\mathcal{M}$*-morphisms is a cospan of* $\mathcal{M}$*-morphisms (cf. (9)(i)).*
- ***Construction:*** *For objects* $A, B \in \mathsf{obj}(\mathbf{C})$*, every element* $A \xrightarrow{q_A} Q \xleftarrow{q_B} B$ *in* $\sum_{\mathcal{M}}(A, B)$ *is obtained from a pushout of some span* $A \xleftarrow{x_A} X \xrightarrow{x_B} B$ *with* $x_A, x_B \in \mathcal{M}$ *and a morphism* $P \xrightarrow{q} Q$ *in* $\mathsf{mono}(\mathbf{C}) \cap \mathsf{epi}(\mathbf{C})$ *(cf. (9)(ii)).*

Proof. See [7, Appendix B].

Since in an adhesive category all monos are regular [34], in this case the multi-sum construction simplifies to the statement that every monic cospan can be uniquely factorized as a cospan obtained as the pushout of a monic span composed with a monomorphism. It is however worthwhile emphasizing that for generic quasi-topoi $\mathbf{C}$ one may have $\mathcal{M} \neq \mathsf{mono}(\mathbf{C})$, as is the case in particular for the quasi-topos **SGraph** of simple graphs. We illustrate this phenomenon in the diagram on the right via presenting the multi-sum construction for $A = B = \bullet$. Note in particular the monic-epis that extend the two-vertex graph S_0 into the graphs S_1, S_2 and S_3, all of which have the same vertices as S_0 (recalling that a morphism in **SGraph** is monic/epic if it is so on vertices), yet additional edges, so that in particular none of the morphisms $S_0 \to S_j$ for $j = 1, 2, 3$ is edge-reflecting.

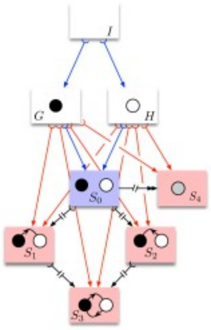

3 Non-linear Sesqui- and Double-Pushout Rewriting

In much of the traditional work on graph- and categorical rewriting theories [22], while it was appreciated early in its development that in particular SqPO-rewriting permits the cloning of subgraphs [17], and that both SqPO- and DPO-semantics permit the fusion of subgraphs (i.e. via input-linear, but output-non-linear rules), the non-uniqueness of pushout complements along non-monic morphisms for the DPO- and the lack of a concurrency theorem in the SqPO-case in general has prohibited a detailed development of non-linear rewriting theories to date. Interestingly, the SqPO-type concurrency theorem for linear rules as developed in [2] exhibits the same obstacle for the generalization to non-linear rewriting as the DPO-type concurrency theorem, i.e., the non-uniqueness

* &

geodesic curvature.

20210914 (20210915)

mdj-epiphanie.me

m01-20210825

matematics-00-20210814

trickle down → gotejar.

Books!

1) Books-018-2021/20210915/
- Multivariate Data Analysis ../
3030769739.pdf
Pàg.27(47) 2.4
Euclidean spaces of matrices and their norms
- Graph Transformation/
10.1007_978-3-030-78946-6.pdf
Pàg. 10(22) → Definition 7

-

M-FPC-augmentations

→ M-morphism such as
square(1)
in the diagram below.

- Handbook of Financial Econometrics, Mathematics.../
Handbook.....(2020).pdf

2) Books-018-2021/20210915/

- Mathematics of the Arts.../
Bharath Sriraman - Handbook of the Mathematics ... (2021).pdf
pág.711(725) →

- Wasan Geometry
→ important.

3) Books-018-2021/20210915/
Handbook of Mathematics/

- Handbook of Elliptic and Hyperelliptic Curve...pdf
843 pages - 2006,
ISBN 978-1-58488-518-4.

* &
+,-.+ + +/ ! + %" : ! ! ! ! ?! #1 C
! ! , ! # 0 0#
+,-.+ + +/ ! + %" : ! ! ! ! #
?! #1 C ! ! , ! # 0 0#
* " >! ! 3

Collection. Aida, the founder of Saijō school, left behind several hundreds volumes, which are on line at *Yamagata University Academic Repository*. Toyoyoshi proposed many problems involving congruent circles, whose manuscripts can be seen in *Digital Library Department of Mathematics, Kyoto University*. More than a hundred titles are online at *Kotenseki Sogo Database in Waseda University*. Also several hundreds of books and manuscripts are online at *Shimoura Collection in Museum of Science Tokyo University of Science*. The first and the last collections can also be found at the site of *National Institute of Japanese literature*.

Wasan Geometry

Let us see the catalog pages at the opening of a popular book of formulas (Yamamoto, 1841) (see Figs. 1, 2, 3, 4, 5, and 6). The figures are arranged to be read top-to-bottom and right-to-left. There are a hundred and two figures, and a formula is stated on later pages for each of the figures. For example, for the leftmost figure at the top in Fig. 5 consisting of an ellipse and a square, the formula states that the side of the square equals $ab/\sqrt{a^2+b^2}$, where a and b are the major axis and the minor axis of the ellipse. As this example shows, which is also stated as a problem and its answer, most geometric problems attempt to discover certain dimensions of geometric figures such as the diameter of a circle, a major (or minor) axis of an ellipse, the side of a rectangle, etc. The catalog also shows that figures involving touching circles constitute the major parts of Wasan geometry. Touching spheres were also considered (see Fig. 5, right). As in Figs. 5 and 6, ellipses, which were considered as sections of cylinders, were studied often, but hyperbolas and parabolas were not outside of several exceptions such as Iwai (no date), a sangaku problem (Gunma Wasan Kenkyūkai, 1987, p. 89), and also see Mikami (1913-2). Also rhombuses were considered, but parallelograms barring a couple of exceptions, such as Gunma Wasan Kenkyūkai (1987, p. 72) and Kimura (1855), were not. Wasan plane geometry is not a triangle geometry, i.e., Wasan mathematicians did not study about one triangle but studied some relationships which arose when several elementary figures such as triangles, rectangles, and circles got together. For a lot of notable problems in Wasan geometry, see the nice collection (Fukagawa and Pedoe, 1989).

Problems Involving Congruent Circles

We state results of Wasan geometry as problems and answers to distinguish them from the others. In this section, we consider problems of Wasan geometry involving several congruent circles. Those problems were considered especially by mathematicians in Shiseisanka school and Toyoyoshi in Miki school. Since the problems are easily generalized today, we consider the problems together with their generalization. We will see that the 3-4-5 triangle is often associated with several interesting special cases.

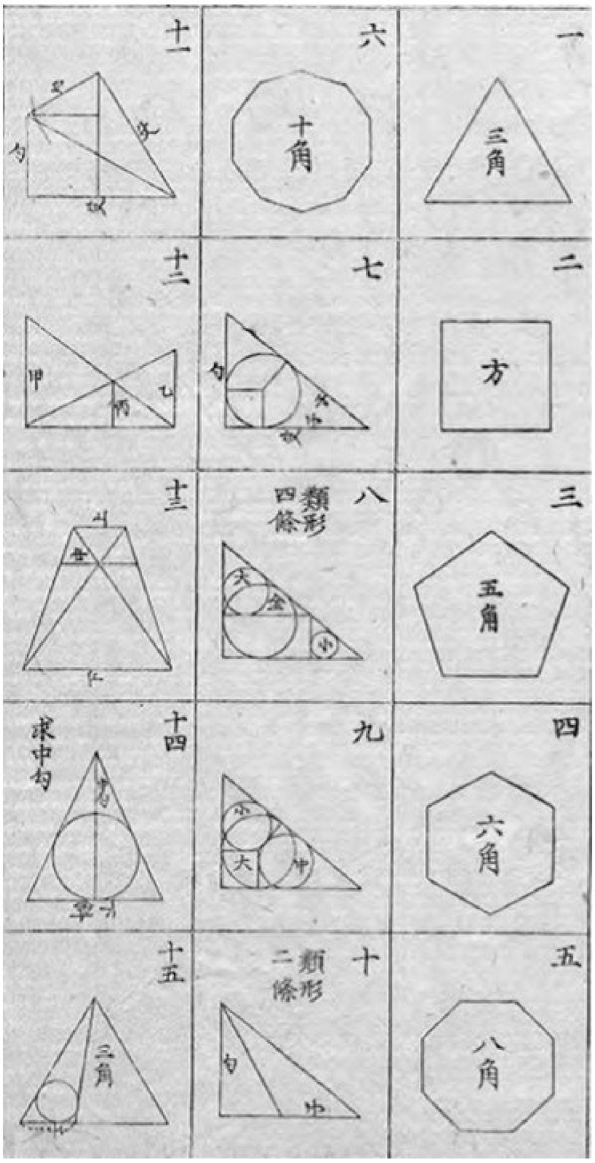
一
三角
二
方
三
五角
四
六角
五
八角
六
十角
七
勾
中
八
類形四條
大
全
小
九
小
大
中
十
類形二條
勾
中
十一
勾
十二
甲
乙
丙
十三
全
十四
求中勾
十五
三角

Fig. 1 Catalog page 1

Congruent circles in Wasan geometry are circles in line in many cases as in Fig. 7, which we define as follows. Let $\alpha_1, \alpha_2, \cdots, \alpha_n$ be congruent circles touching a line t from the same side such that α_1 and α_2 touch, and $\alpha_i (\neq \alpha_{i-2})$ $(i = 3, 4, \cdots, n)$ touches α_{i-1}. In this case, we call the circles $\alpha_1, \alpha_2, \cdots, \alpha_n$ congruent circles on a line or congruent circles on t. The next proposition is very useful for problems in Wasan geometry involving externally touching circles.

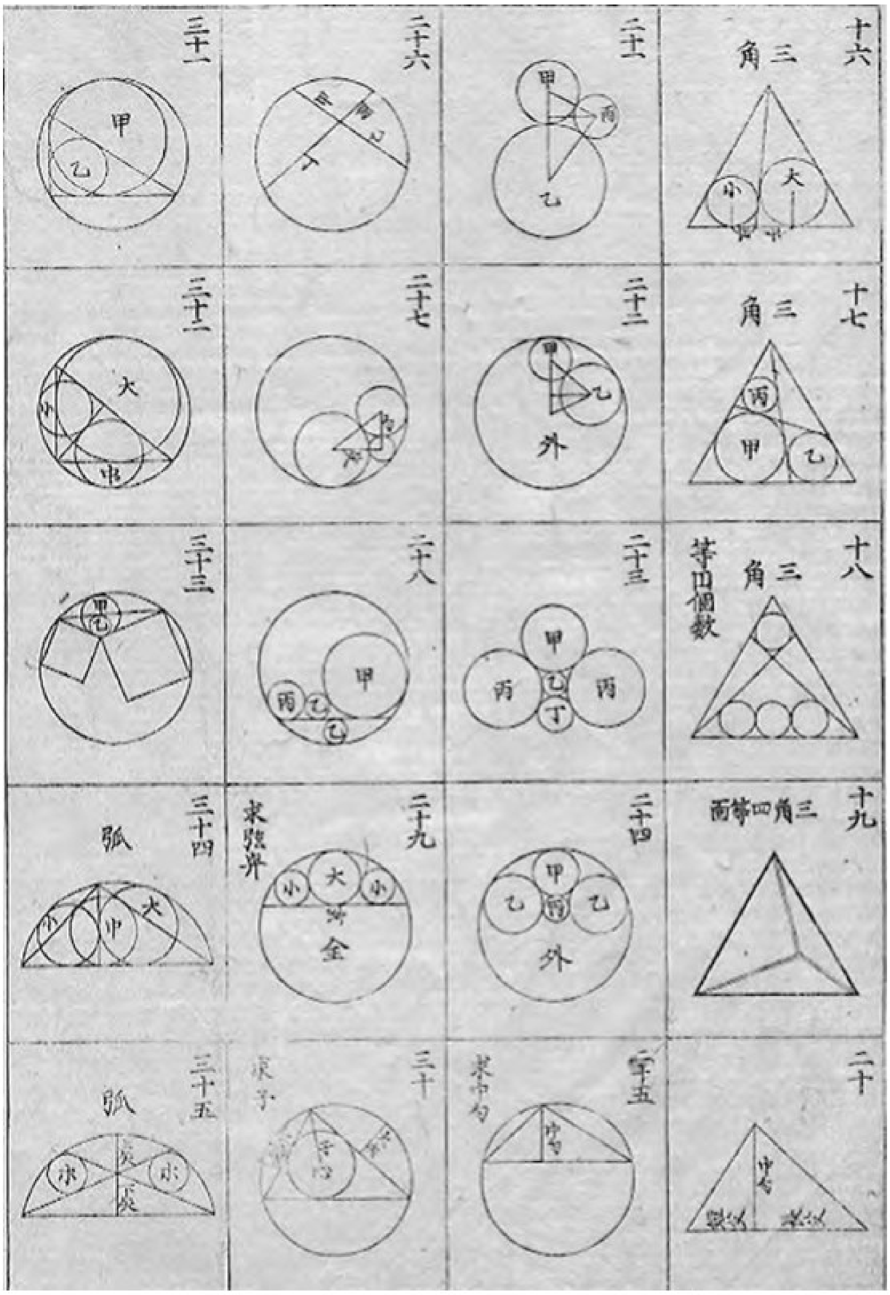
十六
三角
小
大
十七
三角
丙
甲
乙
十八
三角
等円個数
十九
三角四等面
二十
中勾
二十一
甲
乙
丙
二十二
甲
乙
外
二十三
甲
丙
乙
丙
丁
二十四
甲
乙
丙
乙
外
二十五
求中勾
中勾
二十六
二十七
二十八
甲
丙
乙
乙
二十九
求弦幷
小
大
小
全
三十
求子
三十一
甲
乙
三十二
大
小
中
三十三
甲
乙
三十四
弧
小
中
大
三十五
弧
小
小

Fig. 2 Catalog page 2

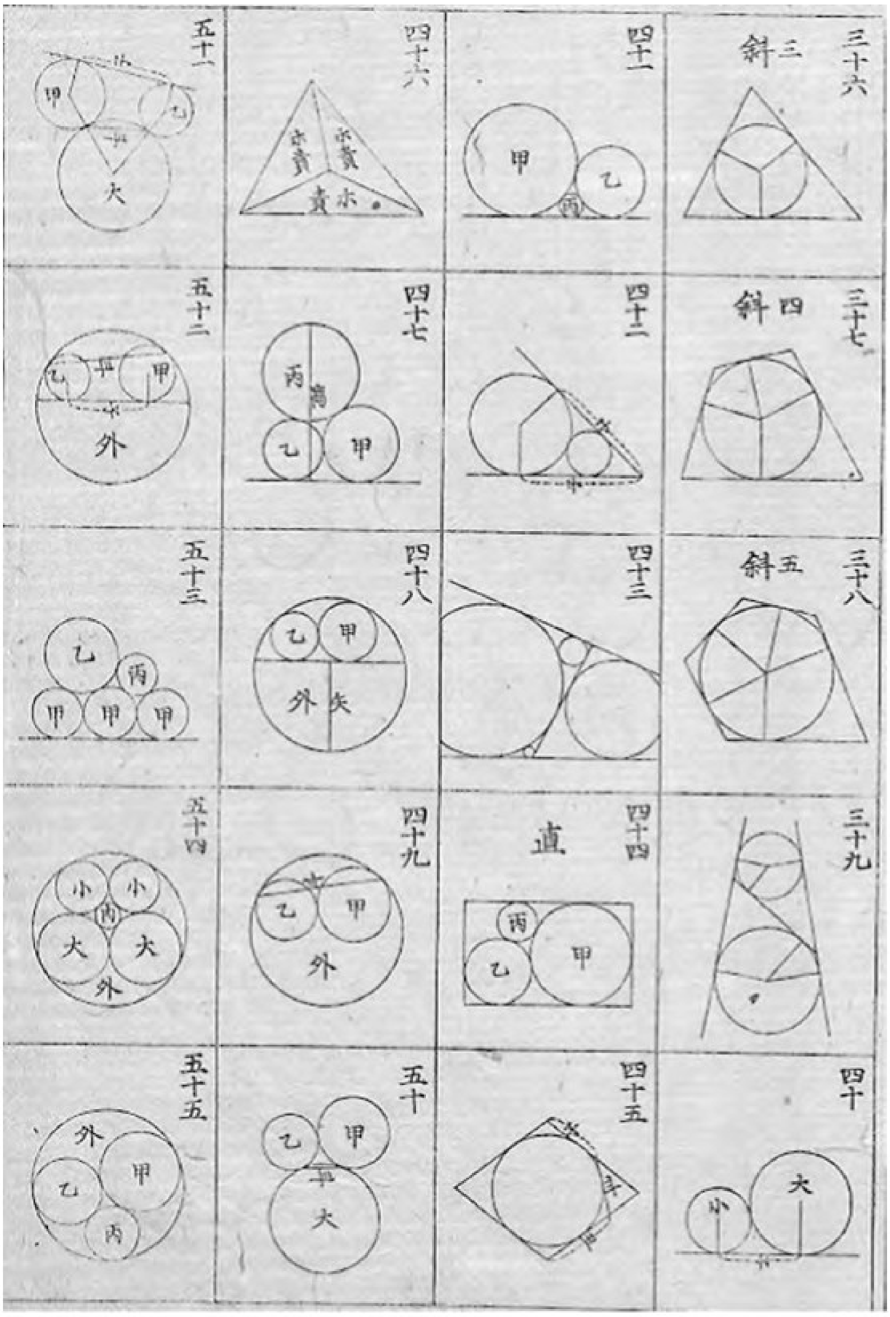
三十六
斜三
三十七
斜四
三十八
斜五
三十九
四十
大
小
四十一
甲
乙
丙
四十二
四十三
四十四
直
丙
甲
乙
四十五
四十六
四十七
丙
乙
甲
四十八
乙
甲
外
矢
四十九
乙
甲
外
五十
乙
甲
大
五十一
甲
乙
大
五十二
乙
甲
外
五十三
乙
丙
甲
甲
甲
五十四
小
小
丙
大
大
外
五十五
外
甲
乙
丙

Fig. 3 Catalog page 3

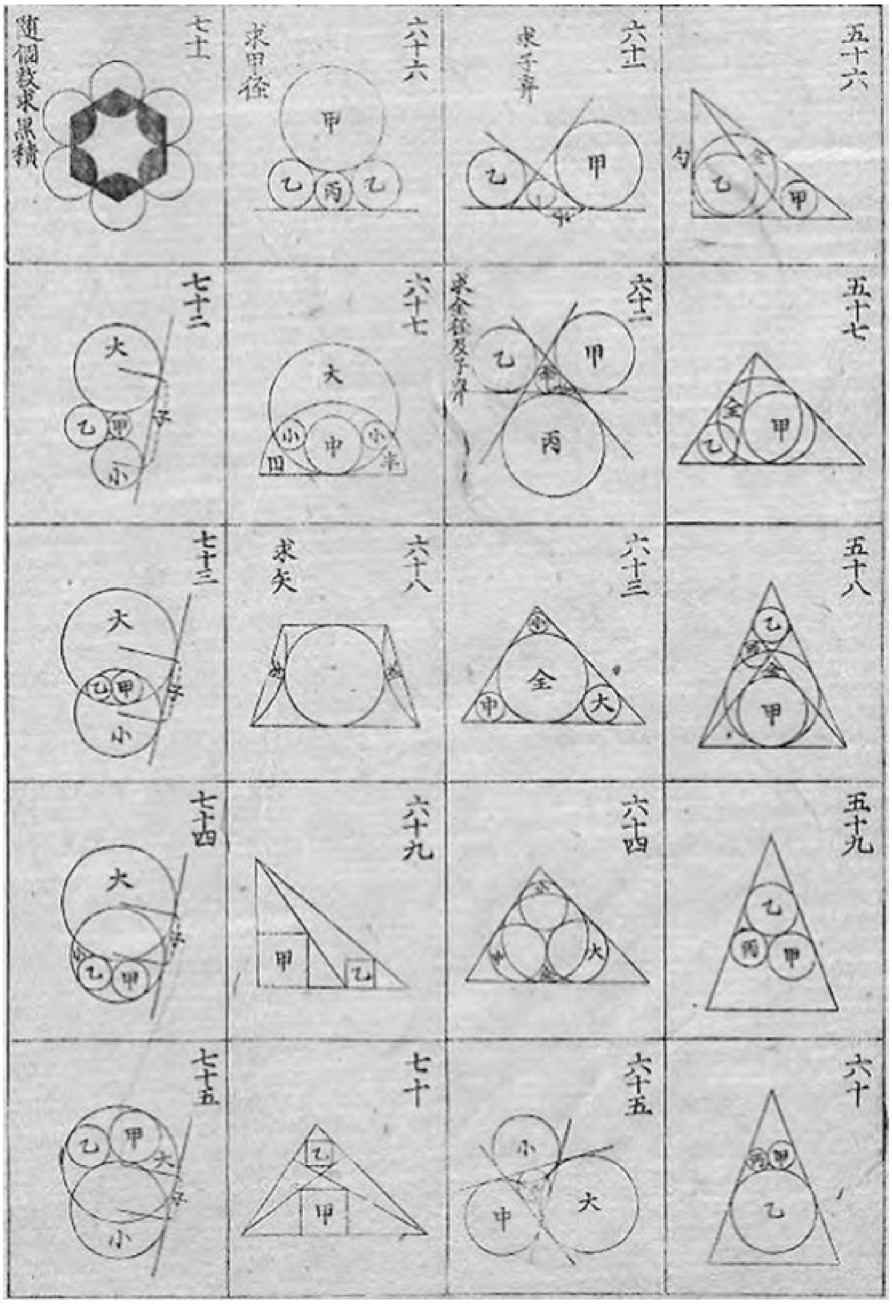
五十六
勾
全
乙
甲
六十一
求子弁
甲
乙
六十六
求甲径
甲
乙
丙
乙
七十一
随個数求黒積
五十七
全
甲
乙
六十二
求全径及子弁
乙
甲
丙
六十七
大
小
中
小
七十二
大
乙
甲
小
五十八
乙
全
甲
六十三
全
甲
大
六十八
求矢
七十三
大
乙
甲
小
五十九
乙
丙
甲
六十四
大
六十九
甲
乙
七十四
大
乙
甲
六十
丙
甲
乙
六十五
小
大
中
七十
乙
甲
七十五
乙
甲
大
小

Fig. 4 Catalog page 4

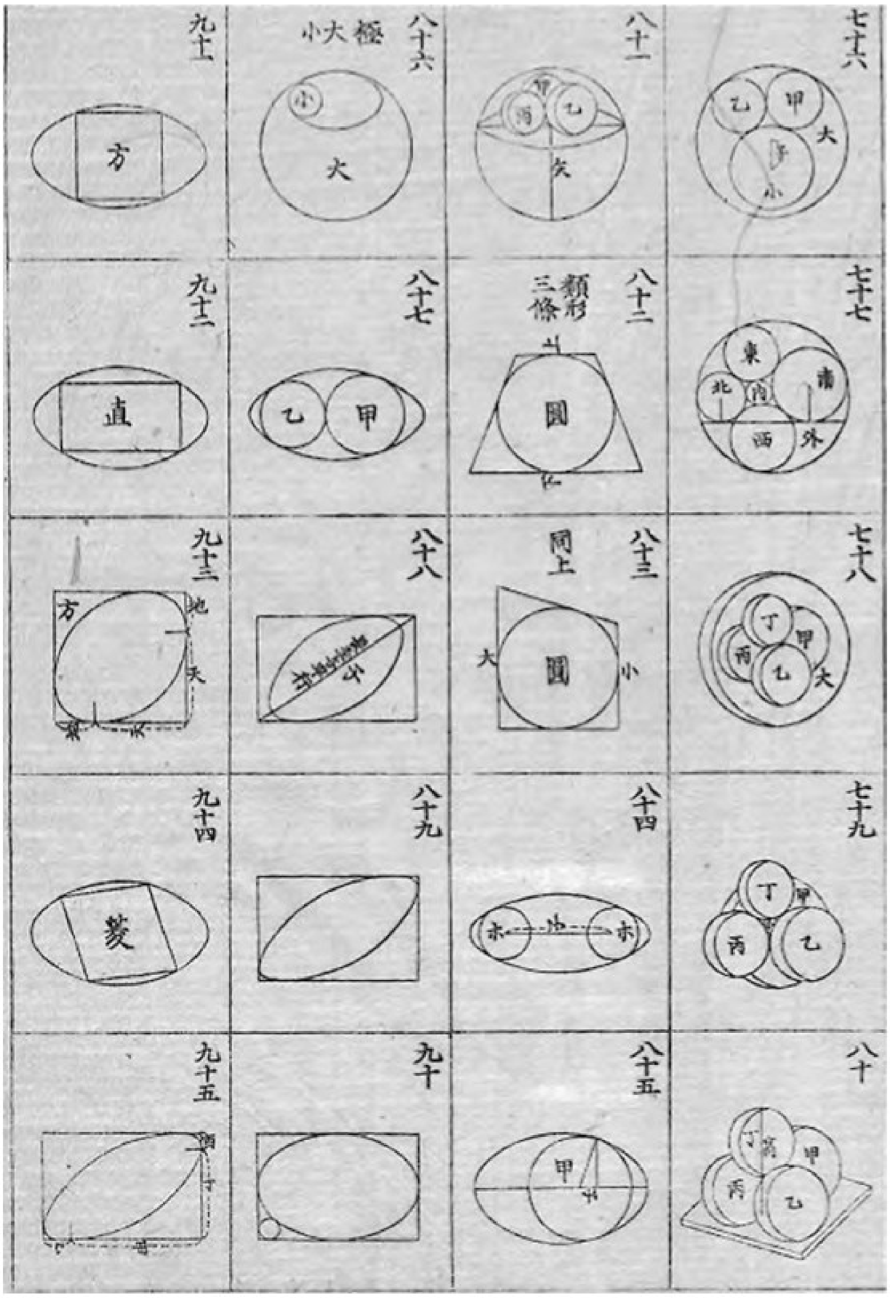
七十六
乙
甲
大
小
七十七
東
北
丙
南
西
外
七十八
丁
甲
丙
乙
大
七十九
丁
甲
丙
乙
八十
丁
甲
丙
乙
八十一
丙
乙
矢
八十二
類形三條
圓
八十三
同上
大
圓
小
八十四
八十五
甲
八十六
極大小
小
大
八十七
乙
甲
八十八
八十九
九十
九十一
方
九十二
直
九十三
方
地
天
九十四
菱
九十五

Fig. 5 Catalog page 5

Fig. 6 Catalog page 6

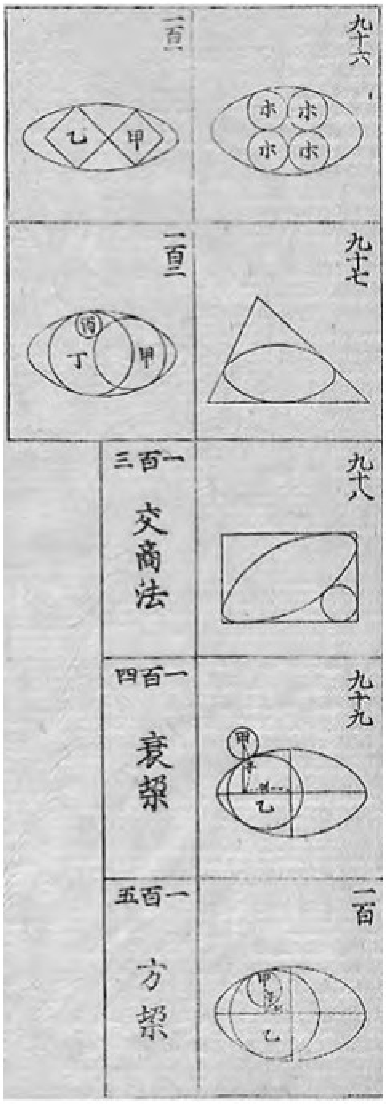
九十六
一百一
乙
甲
九十七
一百二
丙
丁
甲
九十八
一百三
交商法
九十九
一百四
衰垜
甲
乙
一百
一百五
方垜
甲
乙

Fig. 7 Congruent circles on t

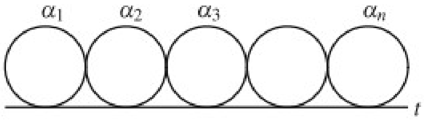
α_1
α_2
α_3
α_n
t

Proposition 1. *If two externally touching circles of radii r_1 and r_2 touch a line at two points P and Q, then $|PQ| = 2\sqrt{r_1 r_2}$.*

Congruent Circles on a Line and a Circle

If $\beta_1, \beta_2, \cdots, \beta_n$ $(n \geq 2)$ are congruent circles of radius b on a line t, and a circle α of radius a touches β_1, β_n, and t, we denote the configuration consisting of α, $\beta_1, \beta_2, \cdots, \beta_n$ and t by $A(n)$ (see Fig. 8). If $n = 2$, the circle α coincides with the incircle of the curvilinear triangle made by β_1, β_2, and t. If $n = 3$, the circles α and β_2 coincide, and we get the trivial relation $a = b$. The relation between a and b was considered in the case $n = 2, 4, 5$ for $A(n)$ in Wasan geometry. The next theorem can easily be proved by Proposition 1 (Okumura, 2017a):

Theorem 1. *A circle of radius a and n $(n \geq 2)$ congruent circles of radius b on a line can form $A(n)$ if and only if*

$$\frac{a}{b} = \left(\frac{n-1}{2}\right)^2. \tag{1}$$

Figure 9 shows $A(5)$ appeared in the problem proposed by Shinohara in 1809 (Shimura, no date). Since $a/b = 1/4$ for $A(2)$, and $a/b = 4$ for $A(5)$, we can construct a recursive configuration arising from Shinohara's figure (see Fig. 10), where the horizontal parallel segments are removed. The configuration $A(n)$ has several interesting properties, one of which is as follows (Okumura, 2017a):

Theorem 2. *Assume that $\beta_1^n, \beta_2^n, \cdots, \beta_n^n$ $(n \geq 2)$ are congruent circles on a line t forming $A(n)$ with a circle α and t. If γ is the incircle of the curvilinear triangle made by α, β_1^n and t, then there are congruent circles $\gamma = \beta_1^{n+2}, \beta_2^{n+2}, \cdots, \beta_{n+2}^{n+2}$ on t such that they form $A(n+2)$ with α and t.*

Assume that a circle α, its tangent t, and congruent circles β_1^3, $\beta_2^3 (= \alpha)$, β_3^3 on t form $A(3)$. Starting with this figure, and using Theorem 2 repeatedly, we get Fig. 11. With this construction, we can define the configuration $A(1)$ as the figure consisting of α, t, and β_1^1, where β_1^1 is the tangent of α parallel to t (see Fig. 12).

Fig. 8 $A(5)$

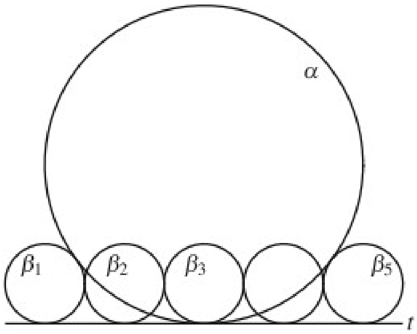

Fig. 9 $A(5)$ in Shinohara's problem (Shimura, no date)

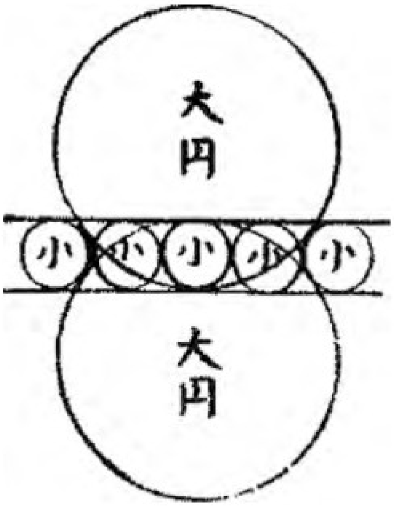

Fig. 10 Configuration arising from $A(5)$

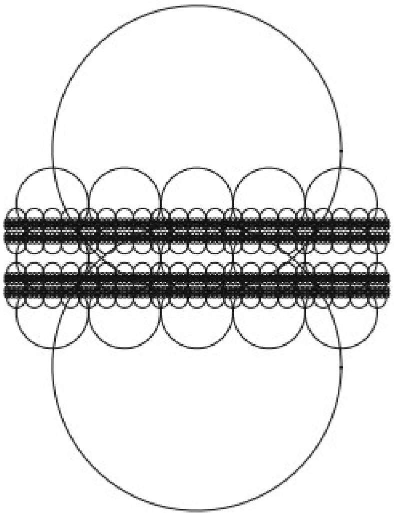

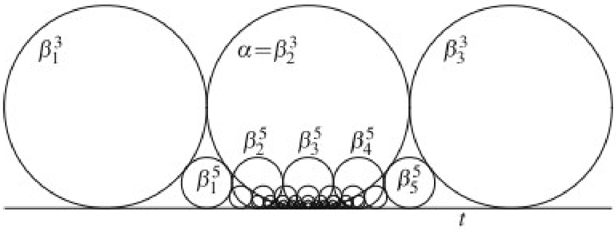
β_1^3
$\alpha=\beta_2^3$
β_3^3
β_2^5
β_3^5
β_4^5
β_1^5
β_5^5
t

Fig. 11 $A(3) \cup A(5) \cup A(7) \cdots$

Fig. 12 $A(1)$

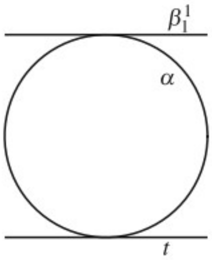

Fig. 13 $B(5)$

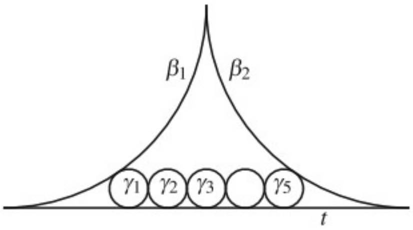

Fig. 14 Toyoyoshi (no date)

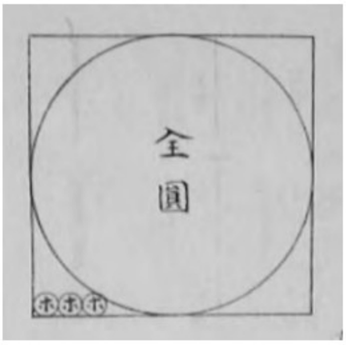

Congruent Circles on a Line with Two Congruent Circles on a Line

Let β_1 and β_2 be congruent circles of radius b on a line t, and let $\gamma_1, \gamma_2, \cdots, \gamma_n$ be congruent circles of radius c on t such that they lie inside of the curvilinear triangle made by β_1, β_2, and t and γ_1 touches β_1 and γ_n touches β_2. The configuration consisting of β_1, β_2, γ_1, γ_2, $\cdots$, γ_n, and t is denoted by $B(n)$ (see Fig. 13). The configuration $B(1)$ coincides with $A(2)$. The relation between b and c was considered in the case in which $n = 1, 5$ and n is even in Wasan geometry. If we divide $B(n)$ by the perpendicular from the point of tangency of β_1 and β_2 to t, we get two congruent figures. Toyoyoshi considered one of the resulting figure and got a relation essentially the same as (2) in the next theorem in the case n being even (Toyoyoshi, no date) (see Fig. 14).

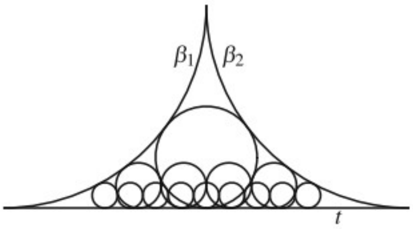

Fig. 15 $B(1) \cup B(4) \cup B(9)$

Theorem 3. *Two circles of radius b and n congruent circles of radius c on a line can form* $B(n)$ *if and only if*

$$\frac{b}{c} = (\sqrt{n} + 1)^2. \tag{2}$$

A similar theorem to Theorem 2 for $B(n)$ also holds (Okumura, 2017a) (see Fig. 15).

Theorem 4. *Assume that* $\gamma_1, \gamma_2, \cdots, \gamma_{n^2}$ *are congruent circles on a line t forming* $B(n^2)$ *with two congruent circles* β_1 *and* β_2 *and t for a positive integer n, where* β_1 *touches* γ_1*. If* δ *is the incircle of the curvilinear triangle made by* β_1, γ_1, *and t, then there are congruent circles* $\delta = \delta_1, \delta_2, \cdots, \delta_{(n+1)^2}$ *on t such that they form* $B((n+1)^2)$ *with* β_1, β_2, *and t.*

Congruent Circles on a Line and Congruent Squares

Congruent squares are derived from the configurations $A(n)$ and $B(n)$. We consider the following problem (Aida, 1788) (see Fig. 16).

Problem 1. Let β_1 and β_2 be congruent circles of radius b on a line t. If $ABCD$ is a square inside of the curvilinear triangle made by β_1, β_2, and t such that the side DA lies on t and the points C and B lie on β_1 and β_2, respectively. Show that $2b = 5|AB|$.

The problem is the case $n = 1$ in the next theorem (Okumura, 2018a) (see Fig. 17).

Theorem 5. *Assume that congruent circles* β_1 *and* β_2 *of radius b on a line t and congruent circles* $\gamma_1, \gamma_2, \cdots, \gamma_n$ *on t form* $B(n)$ *with t, where* γ_1 *and* γ_n *touch* β_1 *and* β_2 *at points C and B, respectively. If A is the foot of perpendicular from B to t, the following relations hold.*

Fig. 16 $n = 1$

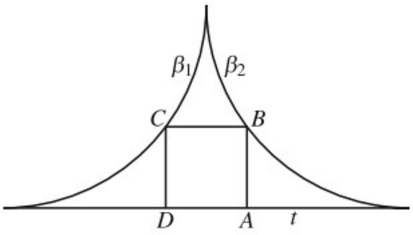
β_1
β_2
C
B
D
A
t

Fig. 17 $n = 3$

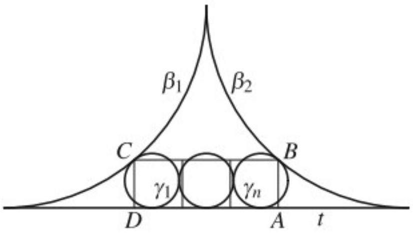
β_1
β_2
C
B
γ_1
γ_n
D
A
t

Fig. 18 $A(4)$

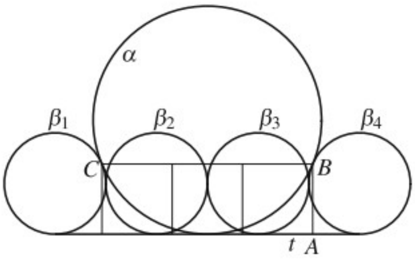
α
β_1
β_2
β_3
β_4
C
B
t
A

(i) $n|AB| = |BC|$.
(ii) $2b = (1 + (\sqrt{n} + 1)^2)|AB|$.

If D is the foot of perpendicular from C to t in the theorem, then the rectangle $ABCD$ can be divided into n congruent squares of side length $|AB|$. The theorem also shows that the incircle of the curvilinear triangle formed by β_1, β_2, and t touches β_1 and β_2 at C and B, respectively, in Problem 1. For the configuration $A(n)$, we also have a similar theorem (see Fig. 18).

Theorem 6. *Assume that a circle α of radius a and congruent circles $\beta_1, \beta_2, \cdots, \beta_n$ on t form $A(n)$ with t, where α touches β_1 and β_n at points C and B, respectively. If A is the foot of perpendicular from B to t, the following relations hold.*

(i) $(n - 1)|AB| = |BC|$.
(ii) $2a = \left(1 + \left(\frac{n-1}{2}\right)^2\right)|AB|$.

Two Congruent Circles on a Line

We consider the next problem (Furuya, 1854) (see Fig. 19). There are several sangaku problems and books stating the same problem.

Problem 2. Let α and β be externally touching circles of radii a and b, respectively, with an external common tangent t. Let γ and γ' be two congruent circles on t of radius c such that they lie inside of the curvilinear triangle made by α, β, and t, γ touches α, and γ' touches β. Find c in terms of a and b.

The part (iii) of the next theorem gives an answer of the problem (Okumura, 2017b) (see Fig. 20).

Theorem 7. *Let α, β, γ, γ', and t be as in Problem 2. Assume that δ and δ' are two congruent circles on t of radius d such that α and β lie inside of the curvilinear triangle made by δ, δ', and t, δ touches α, and δ' touches β. Then the following relations hold.*

(i) $\sqrt{a} + \sqrt{b} = \sqrt{d} - \sqrt{c}$.

(ii) $ab = cd$.

(iii) $c = \dfrac{w - \sqrt{w^2 - 4ab}}{2}$ *and* $d = \dfrac{w + \sqrt{w^2 - 4ab}}{2}$, *where* $w = a + b + 4\sqrt{ab}$.

Fig. 19 A pair of two congruent circles on t

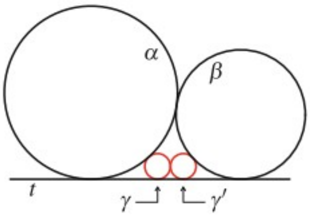

Fig. 20 Two pairs of two congruent circles on t

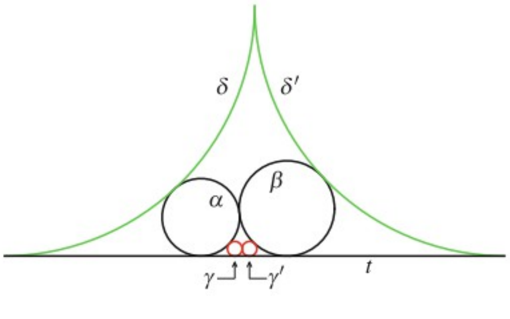

Congruent Circles on a Line with Two Intersecting Congruent Circles

We generalize the next problem (Takeda, 1825), which can also be found in several Wasan books (see Fig. 21).

Problem 3. For two intersecting circles δ_1 and δ_2 of radius s, there are four congruent smaller circles of radius r such that two of them touch each other and δ_1 and δ_2 internally, each of the other two circles touches δ_1 and δ_2 externally, and one of the external common tangents of δ_1 and δ_2. Show $s = 6r$.

Problem 3 is the case $n = 1$ in the next theorem (Okumura, 2018g) (see Fig. 22).

Theorem 8. *For a rectangle $ABCD$ satisfying $s = |BC| > |AB|$, let δ be the circle of center B passing through C. If $\gamma_1, \gamma_2, \cdots, \gamma_n$ are congruent circles of radius r on DA lying inside of the curvilinear triangle made by CD, DA, and δ such that γ_1 touches CD and γ_n touches δ, and $\gamma'_1, \gamma'_2, \cdots, \gamma'_n$ are congruent circles of radius r on DA touching DA from the same side as γ_1 such that γ'_1 touches δ internally from the side opposite to D and γ'_n touches AB from the same side as D, then the following statements hold.*

(i) $s = 2(2n + 1)r$.
(ii) *There is a circle of radius r touching γ_n, γ'_1 and DA.*

Two Sets of Congruent Circles on a Line and Two Circles

We consider the next sangaku problem stated in a collection of problems proposed by mathematicians in Shiseisanka school (no author's name, no datea) (see Fig. 23). The problem is also cited in Saitama Prefectural Library (1969).

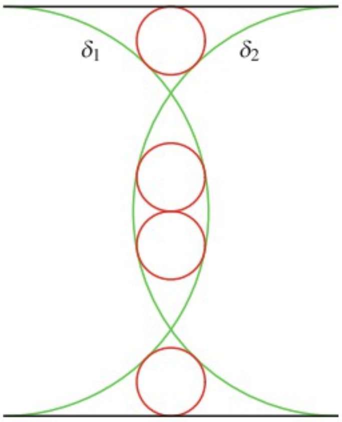

Fig. 21 $n = 1$

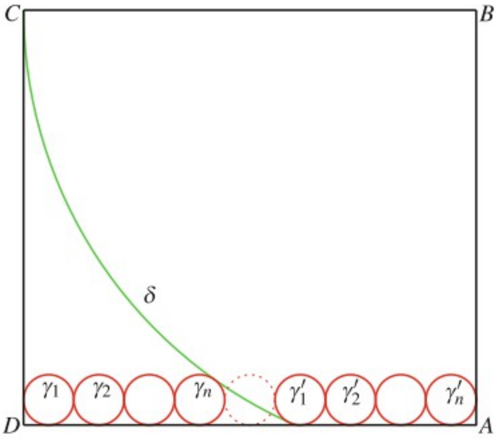

Fig. 22 $n = 4$

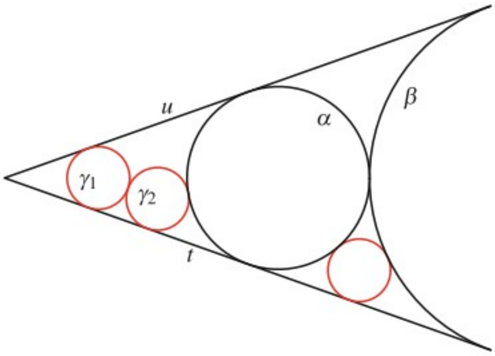

Fig. 23 Two congruent circles on t and one circle on t of the same radius

Problem 4. Let α and β be externally touching circles of radii a and b ($a < b$), respectively, with external common tangents t and u. Congruent circles γ_1 and γ_2 on t lie inside of the curvilinear triangle made by t, u, and α such that γ_1 touches u and γ_2 touches α, and the incircle of the curvilinear triangle made by α, β, and t is congruent to γ_1. Show that $b = 2a$.

The problem can be solved by Proposition 1. It is involving two larger circles and two sets of congruent circles on a line of the same radius, one is $\{\gamma_1, \gamma_2\}$, and the other is the set consisting of incircle of the curvilinear triangle made by α, β, and t, though the latter consists of only one circle. It suggests that there may be some interesting simple relationships between the ratio of the radii of the two larger circles and the numbers of congruent circles of the two sets of the circles of the same radius, if we consider similar figures. The next theorem shows one of such relationships (Okumura and Sodeyama, 1998) (see Fig. 24).

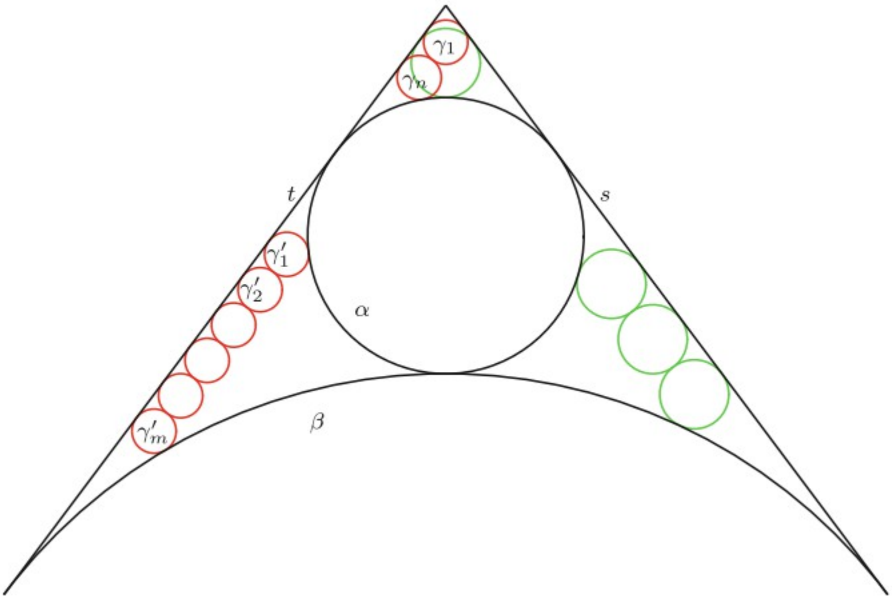

Fig. 24 $(n, m) = (2, 6)$ (left), $(n, m) = (1, 3)$ (right)

Theorem 9. *Let α and β be externally touching circles of radii a and b $(a < b)$, respectively, with external common tangents t and u. Congruent circles $\gamma_1, \gamma_2, \cdots, \gamma_n$ on t lie inside of the curvilinear triangle made by t, u, and α such that γ_1 touches u and γ_n touches α, also congruent circles $\gamma'_1, \gamma'_2, \cdots, \gamma'_m$ on t lie inside of the curvilinear triangle made by α, β and t such that γ'_1 touches α and γ'_m touches β, where γ_1 and γ'_1 are also congruent. Then $b = 4a$ if and only if $3n = m$.*

Notice that the point of intersection of t and u, the center of α, and the point of tangency of α and t form a 3-4-5 triangle in Fig. 24.

A Square and Three Congruent Circles in an Isosceles Triangle

We consider the next problem (Takigawa, 1827), which can be found in several Wasan books (see Fig. 25).

Problem 5. For an isosceles triangle EFG with base EF, let $ABCD$ be a square such that B and C lie on the sides FG and GE, respectively, and DA lies on the side EF, where the triangles ABF and BCG have the same inradius r. Show that $4r = |AB|$.

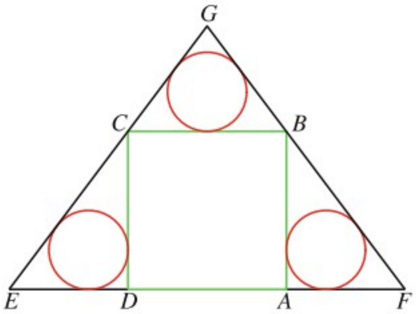

Fig. 25 $n = 1$

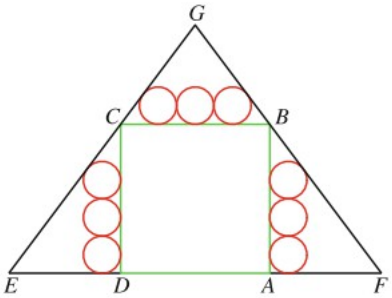

Fig. 26 $n = 3$

For a triangle EFG, let $\gamma_1, \gamma_2, \cdots, \gamma_n$ be congruent circles on EF of radius r lying inside of EFG such that γ_1 touches GE, and γ_n touches FG. In this case we say that EFG has n circles of radius r on EF. Problem 5 is the case $n = 1$ in the next theorem (Okumura, 2018b) (see Fig. 26).

Theorem 10. *For an isosceles triangle EFG with base EF, let $ABCD$ be a square such that B and C lie on the sides FG and GE, respectively, and DA lies on the side EF. If the triangle ABF has n circles of radius r on AB and the triangle BCG has n circles of radius r on BC, then $2(n+1)r = |AB|$ and $|EF| : |FG| = 6 : 5$.*

The theorem shows that the shape of the isosceles triangle EFG is fixed for any positive integer n, which is divided into two 3-4-5 triangles by the perpendicular bisector of the base.

Congruent Circles in a Rectangle

Let $ABCD$ be a rectangle of center O and circumcircle δ of radius s. If the triangle OAB has n circles of radius r on AB and a circle of radius r touches the side BC at the midpoint and the minor arc of δ cut by BC, then the configuration is denoted by $S(n)$. Furthermore if the triangle OBC has m circles of radius r on BC for $S(n)$, we denote the configuration by $S(n, m)$. The following problem was proposed on the

Fig. 27 $S(2), k = 1$

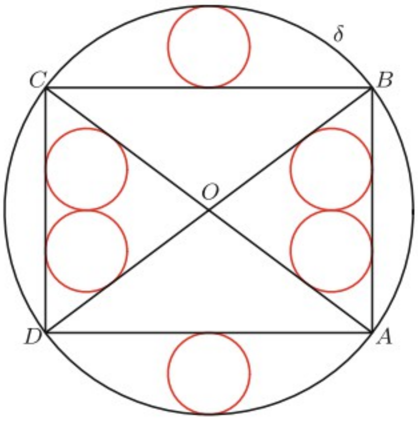

Fig. 28 $S(12), k = 3$

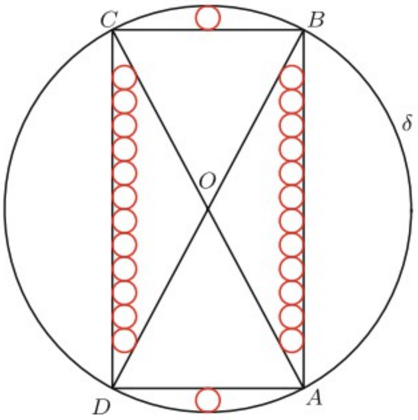

occasion of the 150th anniversary of the death of Kōwa (or Takakazu) Seki (Okayu, 1855) (see Fig. 27).

Problem 6. Show $s/r = 5$ for $S(2)$.

Problem 6 is the case $k = 1$ in the next theorem (Okumura, 2018d) (see Fig. 28).

Theorem 11. *The configuration $S(n)$ can be constructed for any positive integer n, and the shape is determined uniquely by n, where s/r is an integer if and only if there is a positive integer k such that $n = k(k+1)$. In this event $s/r = (k+1)^2 + 1$.*

The configuration $S(n, m)$ exists if and only if $(n, m) = (2, 2), (6, 5)$, where the triangle ABC is a 3-4-5 triangle in both the cases (Okumura, 2018e) (see Figs. 29 and 30).

Fig. 29 $S(2, 2)$

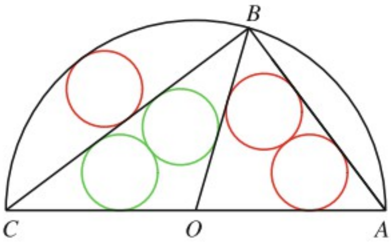

Fig. 30 $S(6, 5)$

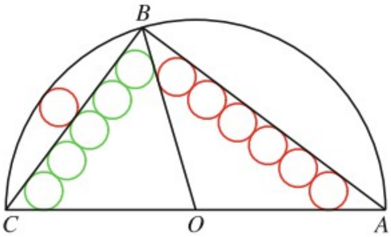

The Arbelos in Wasan Geometry

In this section we consider problems of Wasan geometry involving an arbelos. For a point O on the segment AB, let α, β, and γ be the semicircles of diameters AO, BO, and AB, respectively, constructed on the same side of AB (see Fig. 31). The area surrounded by the three semicircles is called an arbelos and denoted by (α, β, γ). We call the perpendicular to AB at O the axis. The axis divides the arbelos into two curvilinear triangles with congruent incircles. The two congruent circles are believed to be studied by Archimedes and called the twin circles of Archimedes. Circles congruent to the twin circles are said to be Archimedean. Today's research on the arbelos is mostly concerning with finding new Archimedean circles. However the third Archimedean circle was not found until the twentieth century. Meanwhile Leon Bankoff found an Archimedean circle, which is sometimes called Bankoff triplet circle (Bankoff, 1974). After the publication of the notable paper entitled "Those ubiquitous Archimedean circles" (Dodge et al., 1999), which presented dozens of Archimedean circles, many mathematicians were interested in the arbelos, and many papers on the arbelos have been published.

Since the arbelos was also considered in Wasan geometry, it is an very good example to see the different approaches to studying the same figure between the East and the West. Ordinarily the arbelos is described by three semicircles so that the axis is vertical as in Fig. 31. But in Wasan geometry, it is described by three mutually touching circles with collinear centers such that the line passing through the centers is vertical (see Fig. 32). However we describe figures of Wasan geometry in ordinary way to reduce the space here. As Fig. 32 shows, twin circles of Archimedes were also considered in Wasan geometry (Akabane, 1998; Fukushimaken Wasan Kenkyū Hozonkai, 1989; Hirayama and Matsuoka, 1966); however there are few problems

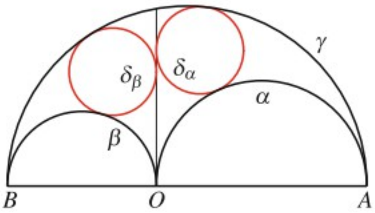

Fig. 31 (α, β, γ) and the twin circles

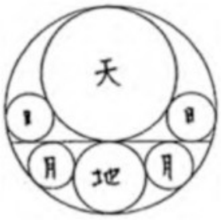

Fig. 32 Fujisawa (1874)

involving other Archimedean circles. We will see two sangaku problems involving such Archimedean circles at the end of this section. On the other hand, various kinds of non-Archimedean circles were considered. We generalize those problems or show that the figures of the problems have some interesting properties.

We assume $|AO| = 2a$, $|BO| = 2b$. The radius of Archimedean circles equals $r_\mathrm{A} = ab/(a+b)$. The twin circles of Archimedes touching α and β are denoted by δ_α and δ_β, respectively.

Two Sangaku Problems Involving a Circle of the Same Radius

We consider two sangaku problems involving a circle of the same radius. The part (i) of the next problem was proposed by Izumiya in 1866 (Saitama Prefectural Library, 1969) (see Fig. 33), and the part (ii) was proposed by Naitoh in 1983 (this sangaku seems to be made in modern days) (Fukushimaken Wasan Kenkyū Hozonkai, 1989) (see Fig. 34).

Problem 7. Assume that $a = b$ for (α, β, γ).
(i) The tangent of α from B meet γ again in a point D. Show that the inradius of the curvilinear triangle made by α, γ and the perpendicular from D to AB equals $a/9$.
(ii) Show that the radius of the circle touching the remaining external common tangent of α and δ_α and the minor arc of γ cut by the tangent at the midpoint equals $a/9$.

The next theorem shows that there are several congruent circles including the two circles in the problem even in the case $a \neq b$ (Okumura, 2019d) (see Fig. 35).

Theorem 12. *We assume that t is the external common tangent of α and β, u is the remaining external common tangent of α and δ_α, v is the tangent of α from B*

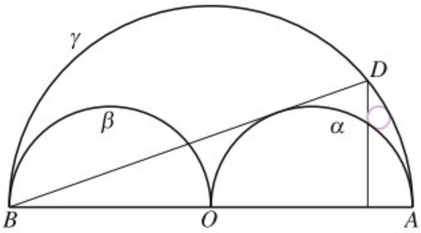
γ
D
β
α
B
O
A

Fig. 33 A small circle of radius $a/9$

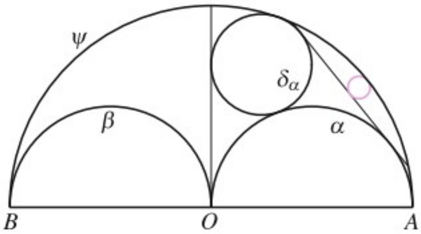
ψ
δ_α
β
α
B
O
A

Fig. 34 A small circle of radius $a/9$

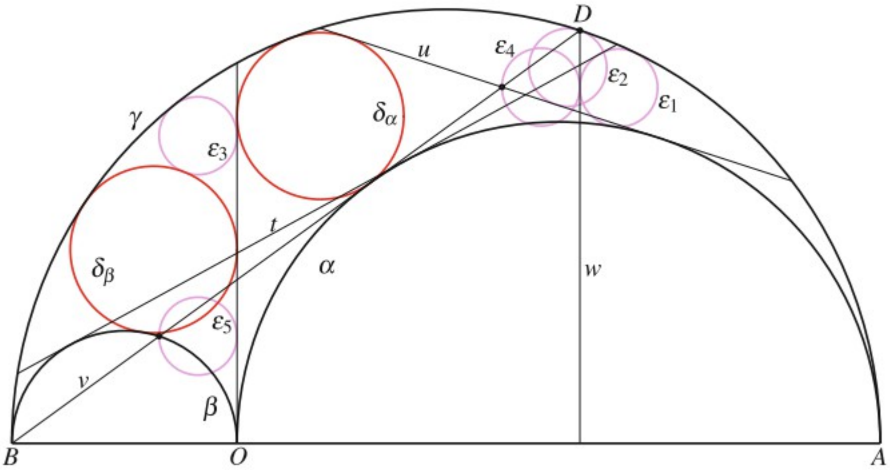
D
u
ε_4
ε_2
ε_1
γ
ε_3
δ_α
t
δ_β
α
w
ε_5
v
β
B
O
A

Fig. 35 A generalization

meeting γ again in a point D, and w is the perpendicular from D to AB. Then the following five circles are congruent and have radius

$$\frac{a^2b}{(a+2b)^2}.$$

ε_1*: the incircle of the curvilinear triangle made by* α, γ, *and* w
ε_2*: the circle touching* u *and the minor arc of* γ *cut by* u *at the point* D
ε_3*: the incircle of the curvilinear triangle made by* γ, δ_β, *and the axis*

ε_4*: the smallest circle touching* w *and passing through the point of intersection of* u *and* v

ε_5*: the smallest circle touching the axis and passing through the point of intersection of* β *and* v

The circle ε_1 also touches the line t, and D is the midpoint of the minor arc of γ cut by u in the theorem.

Two Congruent Circles Touching a Perpendicular to AB

We consider the next sangaku problem, which was proposed by Tamura in 1898 (Saitama Prefectural Library, 1969) (see Fig. 36).

Problem 8. Assume that h is a perpendicular to AB intersecting α and the incircle of the curvilinear triangle made by α, γ, and h is congruent to the circle touching α and β externally and h. Show that the common radius equals $|AB|/10$ in the case $a = b$.

For this problem, we are interested in a necessary and sufficient condition, in which the two congruent circles are obtained. The next theorem gives such a condition (Okumura, 2019e) (see Fig. 37).

Theorem 13. *Let* T *be the point of tangency of the incircle of* (α, β, γ) *and* α. *Let* h *be a perpendicular from* P *to* AB *for a point* P *on* α. *The incircle of the curvilinear triangle made by* α, γ, *and* h *is congruent to the circle touching* α *and* β *externally and* h *if and only if* P *coincides with* T. *In this event if* h *meets* AB *in a point* U,

Fig. 36 Tamura's problem

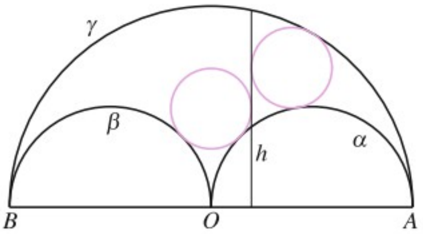

Fig. 37 Three congruent circles

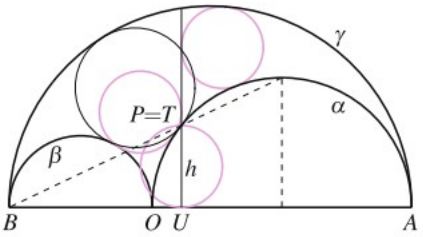

then the circle of diameter TU is also congruent to the two circles and the common radius equals

$$\frac{ab(a+b)}{b^2+(a+b)^2}.$$

Notice that the line BT meets α again in the farthest point on α from AB in the theorem.

Two Circles Touching a Perpendicular to AB at the Same Point

We consider two sangaku problems and assume that h is the perpendicular to AB at H for a point H on the segment AO, δ is the circle touching α and β externally and h, and ε is the incircle of the curvilinear triangle made by α, γ and h. The part (i) of the next problem was proposed by Satoh in 1883 (Fukushimaken Wasan Kenkyū Hozonkai, 1989) (see Fig. 38). The part (ii) was proposed by Chiba in 1880 (Yasutomi, 1987) (see Fig. 39).

Problem 9. (i) The circle ε is congruent to the circle of diameter OH and $a = b$. Show that the common radius is triple the inradius of the curvilinear triangle made by α, γ and the external common tangents of α and β.

(ii) The circle δ is congruent to the circle of diameter AH. Show that the radius of ε equals $ab/(a+2b)$.

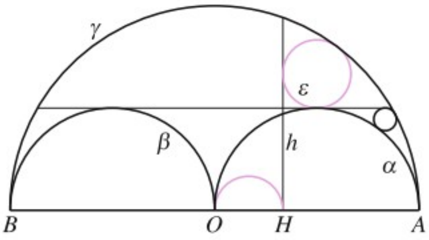

Fig. 38 Two congruent circles

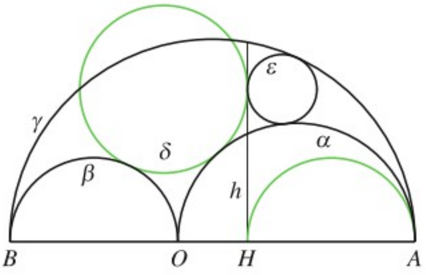

Fig. 39 Two congruent circles

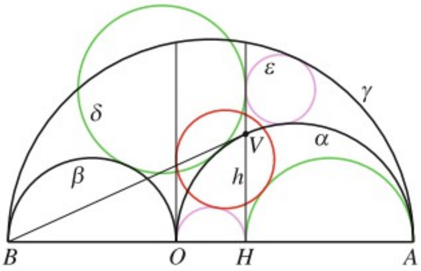

Fig. 40 Circles derived from the tangent of α from B touching α at V

The incircle of the curvilinear triangle made by α, γ, and the external common tangents of α and β in (i) is the circle ε_1 in Theorem 12, i.e., its radius equals $a^2b/(a+2b)^2$ in the case $a \neq b$. We also consider necessary and sufficient conditions, in which each of the pairs of congruent circles in the problem is obtained. The next theorem gives several such conditions (Okumura, 2019b,i) (see Fig. 40).

Theorem 14. *Let V be the point of tangency of α and its tangent from B. The following statements are equivalent.*

(i) *The line h passes through the point V.*
(ii) *The circle δ and the circle of diameter AH are congruent.*
(iii) *The circle ε and the circle of diameter HO are congruent.*
(iv) *The circles δ and ε touch.*
(v) *The two circles touching the two circles of diameters HO and AH externally and the axis are Archimedean.*

If h passes through the point V, then $|HO| = 2ab/(a+2b)$ and the radius of the circle δ equals $a(a+b)/(a+2b)$ in the theorem.

Two Congruent Circles Touching an Inclined Line to AB

We denote the point of intersection of γ and the axis by I. Let ρ be the reflection in the line AB. We assume that ε_1 is the incircle of the curvilinear triangle made by α, γ, and the line IP for a point P on the semicircle α^ρ and ε_2 is the circle touching the semicircles α^ρ, γ^ρ, and IP from the side opposite to A (see Fig. 41). We consider the next problem (Akita, 1833).

Problem 10. The circles ε_1 and ε_2 are congruent. Find the common radius in terms of a and $|IO|$.

Considering the power of the point O with respect to γ, we get $|IO| = 2\sqrt{ab}$. Therefore Problem 10 essentially demands to find the common radius of the two circles in terms of a and b. However b was not used in the problem because of the

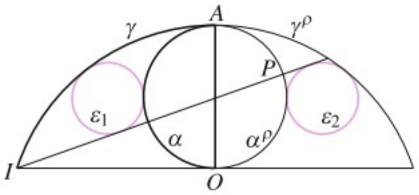
γ
A
γ^ρ
P
ε_1
ε_2
α
α^ρ
I
O

Fig. 41 Two congruent circles touching an inclined line to AB

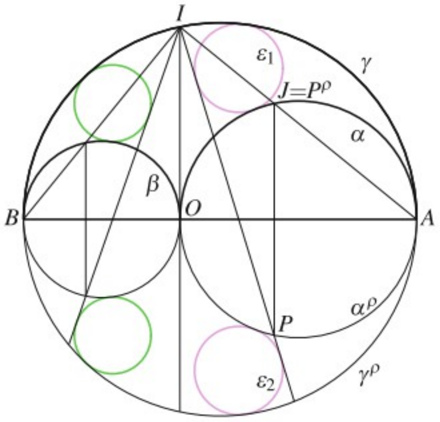
I
ε_1
γ
$J{=}P^\rho$
α
β
B
O
A
P
α^ρ
γ^ρ
ε_2

Fig. 42 Two congruent circles touching α, γ and IJ^ρ

absence of the semicircle β from the figure. Let J be the point of intersection of α and the segment AI. The next theorem gives a condition, in which the pair of congruent circles in the problem is obtained (Okumura, 2019a) (see Fig. 42).

Theorem 15. *The circles ε_1 and ε_2 are congruent if and only if the point P^ρ coincides with the point J. In this event the common radius equals*

$$\frac{4a^2b}{(2a+b)^2}.$$

The point J is one of the most notable points on α, because the external common tangent of α and β touches α at J, the distances from the center of β to I and J are the same, and the distance between J and the axis equals $2r_{\mathrm{A}}$, where recall that r_{A} is the radius of Archimedean circles.

Congruent Circles Touching a Circle Passing Through the Center of α

The next sangaku problem was proposed by Matsuda (no author's name, no dateb) (see Fig. 43).

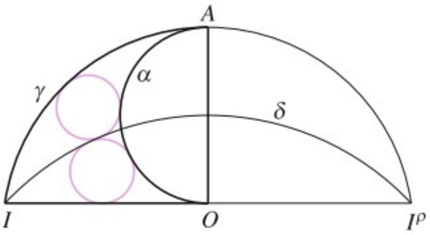

Fig. 43 Matsuda's problem

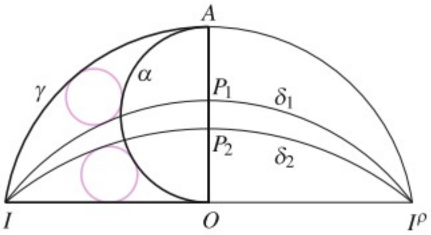

Fig. 44 A generalization

Problem 11. A circle δ passes through the points I and I^ρ and divides the curvilinear triangle made by α, γ, and IO into two curvilinear triangles with congruent incircles. Show that δ passes through the center of α.

The problem is generalized as follows (Okumura, 2019c) (see Fig. 44).

Theorem 16. *For a point P_i $(i = 1, 2)$ on the segment AO, let δ_i be the circle passing through the points I, I^ρ, and P_i. Then the incircle of the curvilinear triangle made by α, γ, and δ_1 is congruent to the incircle of the curvilinear triangle made by α, IO, and δ_2 if and only if the center of α coincides with the midpoint of P_1P_2.*

Reflection in the Axis

Let $'$ be the reflection in the axis. The next sangaku problem was proposed by Satoh in 1850 (Yasutomi, 1987) (see Fig. 45).

Problem 12. Show that the radius of the circle touching the semicircles β' externally, γ internally, and the axis from the side opposite to B equals $a/2$.

The problem is obtained if $n = 1$ in the next theorem (Okumura, 2019f) (see Fig. 46).

Theorem 17. *Let (n) be the composition of' and the homothety with center O and ratio n for a non-negative real number n, where we consider $X^{(0)} = O$ for any point X. If α_2 (resp. β_2) is the circle touching $\beta^{(n)}$ (resp. $\alpha^{(n)}$) externally, γ internally, and*

Fig. 45 Satoh's problem

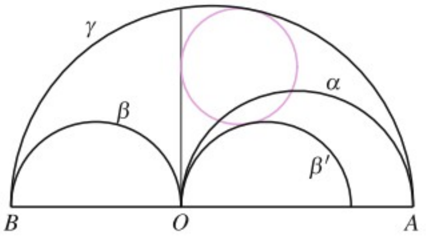

Fig. 46 $n = 0.8$

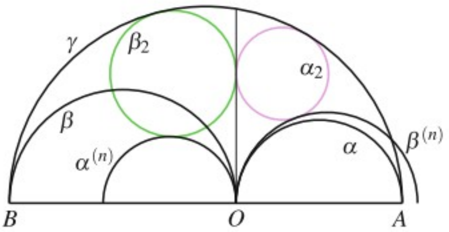

Fig. 47 C_1

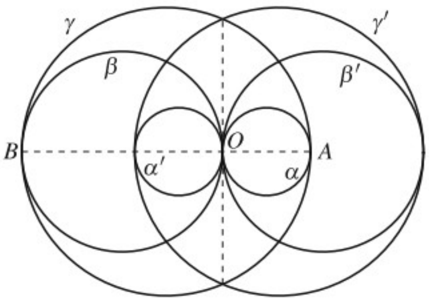

the axis from the side opposite to B (resp. A), then the circles α_2 *and* β_2 *touch and have radii* $a/(n+1)$ *and* $b/(n+1)$*, respectively.*

We construct a recursive circle configuration by Theorem 17. For this purpose we consider the arbelos together with its reflection in AB, i.e., we assume that α, β, and γ are the circles of diameters AO, BO, and AB, respectively, and denote the configuration consisting of the three circles by (α, β, γ). Let $C_1 = (\alpha, \beta, \gamma) \cup (\alpha, \beta, \gamma)'$ (see Fig. 47). We call C_1 a symmetric arbelos of diameter $|AB|$, where we assume $|AB| = 1$. Let α_2 (resp. β_2) be one of the circles touching β' (resp. α') externally, γ internally, and the axis from the side opposite to B (resp. A), where we assume that α_2 and β_2 lie on the same side of AB, i.e., they touch by Theorem 17. If γ_2 is the smallest circle touching the two circles internally, then the figure $(\alpha_2, \beta_2, \gamma_2) \cup (\alpha_2, \beta_2, \gamma_2)'$ is a symmetric arbelos of diameter $1/2$ similar to C_1. We call $(\alpha_2, \beta_2, \gamma_2) \cup (\alpha_2, \beta_2, \gamma_2)'$ with its reflection in AB the two small copies of C_1 (see Fig. 48). If the configuration C_k has been constructed, which consists of C_1, two symmetric arbeloi of diameter $1/2$ similar to C_1, four symmetric arbeloi

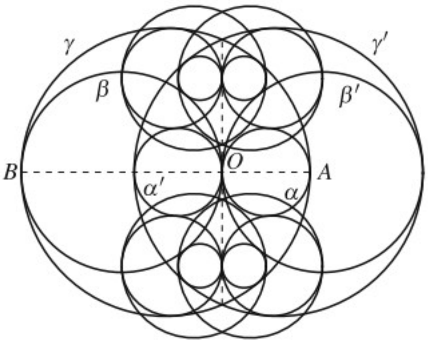

Fig. 48 C_1 with the two small copies $= C_2$

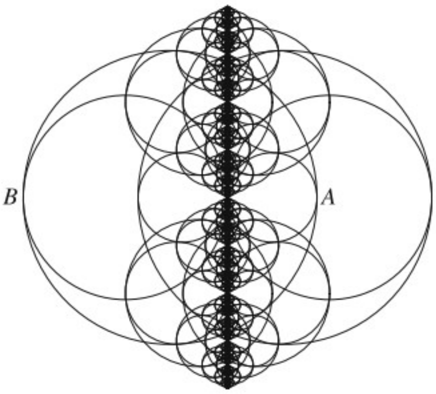

Fig. 49 C_0

of diameter $1/4$ similar to $C_1, \cdots$, and 2^{k-1} symmetric arbeloi of diameter $1/2^{k-1}$ similar to C_1, then we add the two small copies of each of the symmetric arbeloi of diameter $1/2^{k-1}$ and let C_{k+1} be the resulting configuration. The configuration C_n is now defined for any positive integer n. Then we define the configuration $C_0 = \bigcup_{i \geq 1} C_i$ (see Fig. 49).

Golden Arbelos

If $a/b = \phi^{\pm 1}$, then (α, β, γ) is called a golden arbelos, where $\phi = (1 + \sqrt{5})/2$. We consider a problem involving a golden arbelos. Let σ be the reflection in the perpendicular bisector of AB. Recall that the radius of Archimedean circles equals $r_{\mathrm{A}} = ab/(a+b)$. We consider the next problem (Ono, 1855) (see Fig. 50). The same sangaku problem can be found in Fukushimaken Wasan Kenkyū Hozonkai (1989).

Problem 13. Let ε be the circle touching the semicircles γ internally, α^{σ} externally, and the axis from the side opposite to A such that ε and α have the same radius. Find the radius of ε in terms of the radii of γ and δ_{α}.

Fig. 50 The case α^{σ} and ε being congruent

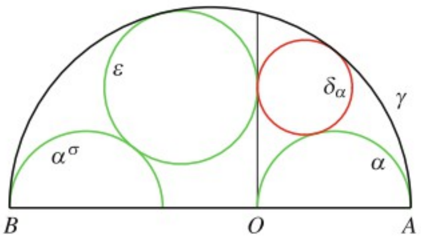

Fig. 51 A generalization

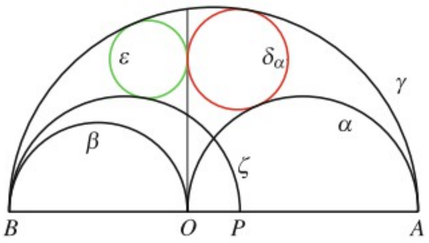

We show that the circles δ_{α} and ε touch in the problem. Let ζ be a semicircle of diameter BP for a point P on the segment AB constructed on the same side of AB as γ, and let ε be the circle touching γ internally, ζ externally, and the axis from the side opposite to A (see Fig. 51). The next theorem gives conditions in which the circles δ_{α} and ε touch (Okumura, 2019h).

Theorem 18. *The following statements are equivalent.*

(i) *The circles δ_{α} and ε touch.*
(ii) *The circle ε has radius $b - r_{\mathrm{A}}$.*
(iii) *The semicircle ζ coincides with α^{σ}.*

Assume that ζ and ε have radius a as in the problem. Then the circles δ_{α} and ε touch, and we get $a = b - r_{\mathrm{A}}$ by the theorem, which implies $2a = a + b - r_{\mathrm{A}}$. Hence we get the answer of the problem $a = (a + b - r_{\mathrm{A}})/2$. On the other hand, $a = b - r_{\mathrm{A}}$ is equivalent to $b = \phi a$. Therefore (α, β, γ) is a golden arbelos and $r_{\mathrm{A}}, a, b, a + b$ form a geometric progression with common ratio ϕ. Also $a = b - r_{\mathrm{A}}$ shows that the circle of diameter OO^{σ} is Archimedean. The circle also touches ε, since it is concentric to γ (see Fig. 52). The next theorem shows that the Archimedean circle can also be obtained in the case $b \neq \phi a$ and gives a characterization of the golden arbelos using the Archimedean circle (Okumura, 2019h) (see Fig. 53).

Theorem 19. *Let η be the circle touching the circle ε externally and the axis at O from the side opposite to A. Then η is Archimedean if and only if ζ and ε have the same radius. In this event, (α, β, γ) is a golden arbelos such that $b = \phi a$ if and only if ζ and η touch.*

Fig. 52 The Archimedean circle of diameter OO^{σ}

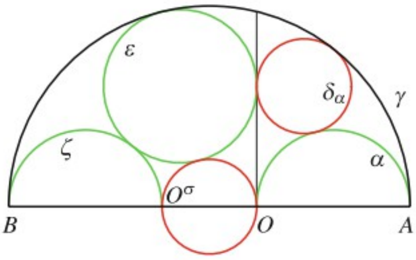
ε
δ_{α}
γ
ζ
α
O^{σ}
B
O
A

Fig. 53 The case η being Archimedean

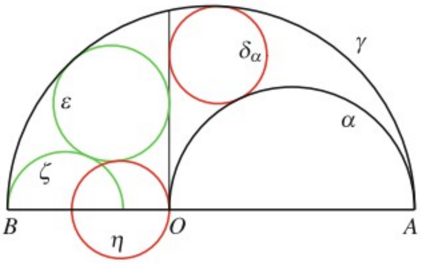
δ_{α}
γ
ε
α
ζ
B
O
A
η

Fig. 54 $G \cup G^{\tau} = G_1 \cup G_2$

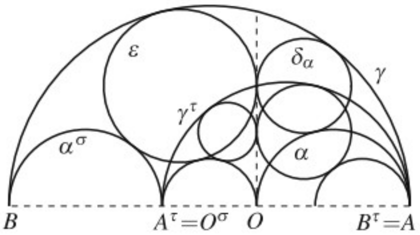
ε
δ_{α}
γ
γ^{τ}
α^{σ}
α
B
$A^{\tau}{=}O^{\sigma}$
O
$B^{\tau}{=}A$

We assume $b = \phi a$ and construct a self-similar circle configuration. Let τ be the composition of σ and the homothety of center A and ratio $1/\phi$, where σ is applied at first. Then τ coincides with the homothety of center O and ratio $-1/\phi$ for a point on the line AB. Therefore the axis is fixed by τ. Also γ^{τ} passes through the point of tangency of δ_{α} and ε by Proposition 1, because $(2\sqrt{ar_{\mathrm{A}}})^2 = 2a \cdot 2\phi^{-1}a = |B^{\tau}O||A^{\tau}O|$ (see Fig. 54).

Let G be the figure consisting of α, α^{σ}, γ, δ_{α}, and ε, which is obtained by removing the axis and AB from Fig. 50. Let $G_i = G^{\tau^{i-1}}$ for $i = 1, 2, 3, \cdots$, and $G_0 = \bigcup_{i \geq 1} G_i$. Following to the custom of Wasan geometry, we describe G_0 so that AB is a vertical line with its reflection in AB (see Fig. 55).

Arbelos with Overhang

Aida (1747–1817) considered a deformed arbelos formed by a circle and two semicircles in the circle as in Fig. 56 and left several notable results, some of which are stated as follows (Aida, no date) (see Fig. 57).

Fig. 55 G_0 with its reflection in AB

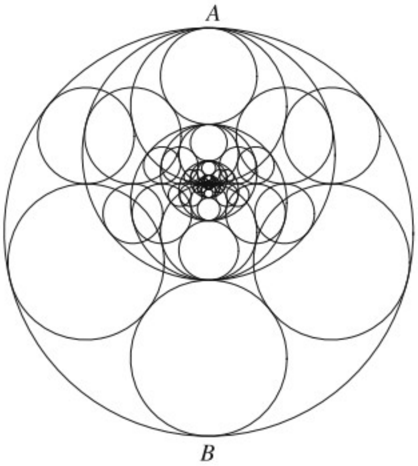

Fig. 56 Aida (no date)

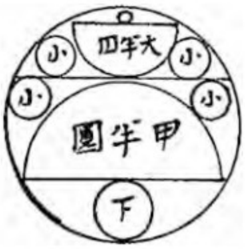

Fig. 57 Aida's deformed arbelos

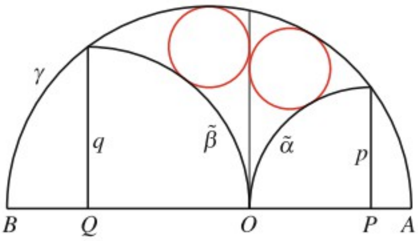

Problem 14. Assume that $\tilde{\alpha}$ (resp. $\tilde{\beta}$) is the quarter circle of radius p (resp. q) and center P (resp. Q) on the segment AO (resp. BO) such that one of the endpoints is O and the other lies on γ. Show that the following statements hold.

(i) The axis divides the curvilinear triangle made by $\tilde{\alpha}$, $\tilde{\beta}$, and γ into two curvilinear triangles with common inradius r satisfying $q = 2(p-r)r/(p-2r)$.
(ii) $c = |AP| + q = |BQ| + p$, where c is the radius of γ.
(iii) $p^2 + q^2 = c^2$.
(iv) If s is the inradius of the curvilinear triangle made by $\tilde{\alpha}$, $\tilde{\beta}$, and γ, then $2s = (p+q)^2/\sqrt{p^2+q^2} - (p+q)$.

Solving the equation in (i) for r, we get $r = (p+q-\sqrt{p^2+q^2})/2 = (p+q-c)/2$ by (iii). Therefore we have

$$2r + c = p + q.$$

By the last equation and (iv), we have $2s = (p+q)^2/c - (p+q) = 2(p+q)r/c$. Hence we get

$$\frac{s}{r} = \frac{p+q}{c}.$$

Essentially the same relation as (iii) was rediscovered (Jobbings, 2011). Aida's deformed arbelos is a special case of a generalized arbelos called the arbelos with overhang (Okumura, 2014). Let A_h (resp. B_h) be a point on the half line OA (resp. OB) with initial point O such that $|OA_h| = 2(a+h)$ (resp. $|OB_h| = 2(b+h)$) for a real number h satisfying $-\min(a,b) < h$. Let α_h (resp. β_h) be the semicircle of diameter A_hO (resp. B_hO) constructed on the same side of AB as γ. The configuration of three semicircles α_h, β_h, and γ is called an arbelos with overhang h and denoted by $(\alpha_h, \beta_h, \gamma)$ (see Figs. 58 and 59). The circle touching α_h (resp. β_h) externally, γ internally, and the axis from the side opposite to B (resp. A) has radius

$$\frac{ab}{a+b+h}.$$

Circles of the same radius are called Archimedean circles of $(\alpha_h, \beta_h, \gamma)$.

Let V (resp. W) be the point of intersection of α_h (resp. β_h) and γ in the case $h \geq 0$. Since

Fig. 58 $-\min(a,b) < h < 0$

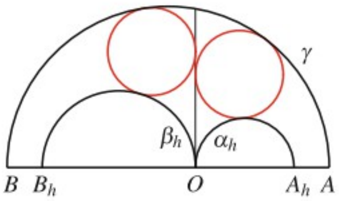

Fig. 59 $h \geq 0$

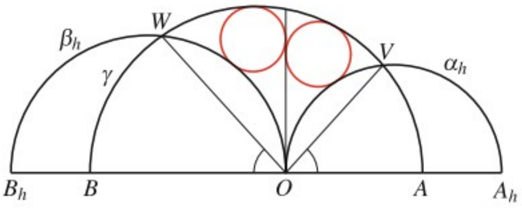

$$\cos \angle VOA = \cos \angle WOB = \sqrt{\frac{ab}{(a+h)(b+h)}},$$

Aida's deformed arbelos is obtained if $\cos \angle VOA = 1/\sqrt{2}$, which is equivalent to

$$h = \frac{f-a-b}{2}, \tag{3}$$

where $f = \sqrt{a^2+6ab+b^2}$. While we easily have $h = p-a = q-b$. Therefore

$$p = a+h = \frac{f+a-b}{2}, \quad q = b+h = \frac{f+b-a}{2}, \quad p+q = f.$$

The equation (3) is equivalent to $h = ab/(a+b+h)$. Therefore we get $r = h$, and the circles of diameters AA_h and BB_h are Archimedean circles of $(\alpha_h, \beta_h, \gamma)$.

Arbeloi Determined by a Chord

Let δ be a circle of radius s. A chord of δ determines two congruent arbeloi, if we consider two circles touching the chord at the midpoint and the circle δ (see Fig. 60). We denote the radius of Archimedean circles of the arbeloi by $\overline{p}$ if the length of the chord equals p. There is a simple relation between p and $\overline{p}$ (Okumura, 2019g):

$$p = 4\sqrt{s\overline{p}}. \tag{4}$$

The next sangaku problem was proposed by Kawada (or Kawata) in 1816 (no author's name, no datea) and cited in Saitama Prefectural Library (1969) (see Fig. 61).

Fig. 60 $p = |QR|$

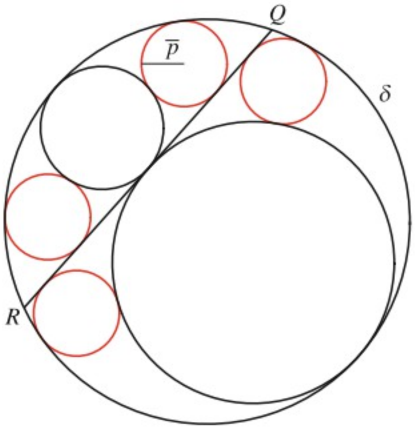

Fig. 61 $\overline{p} = \overline{q} + \overline{r}$

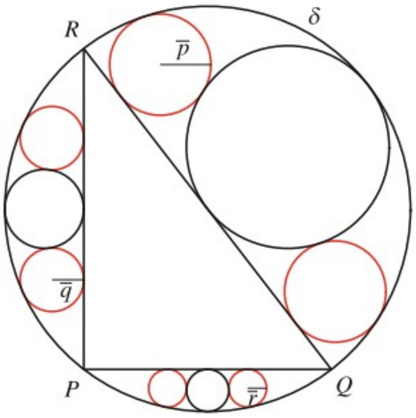

Problem 15. A triangle PQR with right angle at P is inscribed in the circle δ. Show that $\overline{p} = \overline{q} + \overline{r}$ holds, where $p = |QR|$, $q = |RP|$ and $r = |PQ|$.

The relation (4) gives a solution of the problem and shows that the converse is also true. A problem stating the relationship between s and $\overline{p}$, $\overline{q}$, $\overline{r}$ for a general triangle can be found in Aida (1788). Also a problem stating a relationship between s and the four radii of Archimedean circles of the arbeloi determined by four chords forming an inscribed quadrilateral in δ can be found in Aida (1788). Those relationships can also be derived by (4) from the relation between the lengths of the sides of a triangle or a cyclic quadrilateral and its circumradius.

A Sangaku Problem Involving an Archimedean Circle

We consider the next problem proposed by Ootoba, which can be found in the sangaku hung in 1853 at Takenobu Inari Shrine in Kyoto (Kinki Sūgakushigakukai, 1992) (see Fig. 62).

Problem 16. Let H be the point of intersection of the tangent of γ parallel to AB and the tangent of α at the point A. Show that the incircle of the triangle made by the tangent of γ parallel to AB, the axis, and the remaining tangent of α from H is Archimedean.

The circle corresponding to γ is not described in the original sangaku problem, but a square is described (see Fig. 63). As stated before, the twin circles of Archimedes were considered in Wasan geometry. But it is little to see the other Archimedean circles. Moreover there seems no other mathematicians who found the same circle as to Ootoba's Archimedean circle until today. Therefore this circle is very notable. As stated at the beginning of this section, it is believed that Bankoff's circle is the thirdly discovered Archimedean circle (Bankoff, 1974). But the date

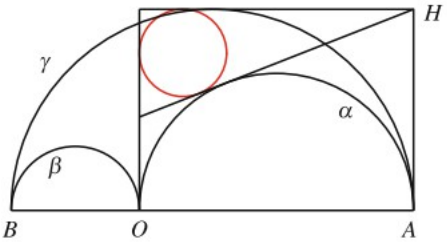

Fig. 62 Ootoba's Archimedean circle

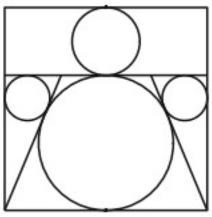

Fig. 63 Original figure

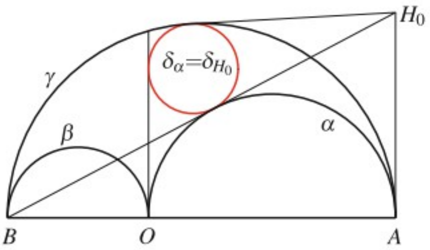

Fig. 64 Generalized Ootoba's Archimedean circle closest to the line AB

of the sangaku suggests that the history of studying Archimedean circles should be updated. The problem is generalized as follows (Okumura, 2019j) (see Figs. 64 and 65).

Theorem 20. *Let H_0 be the point of intersection of the tangent of α from B and the common tangent of α and γ at A. Let H_1 and H_2 be points different from A on the half line AH_0 with initial point A. If the incircle of the triangle made by the remaining tangents of α and γ from H_i and the axis is denoted by δ_{H_i} for $i = 0, 1, 2$, then the following statements hold.*

(i) *The circle δ_{H_i} is Archimedean.*
(ii) *The circle δ_{H_1} is closest to AB if and only if $H_1 = H_0$, where the circles δ_{H_0} and δ_α coincide.*
(iii) *The circles δ_{H_1} and δ_{H_2} coincide if and only if H_2 is the inverse of H_1 in the circle of center A passing through H_0*

Since there are infinitely many choices of the point H_1, we get infinitely many Archimedean circles touching the axis from the side opposite to B. Similar infinitely

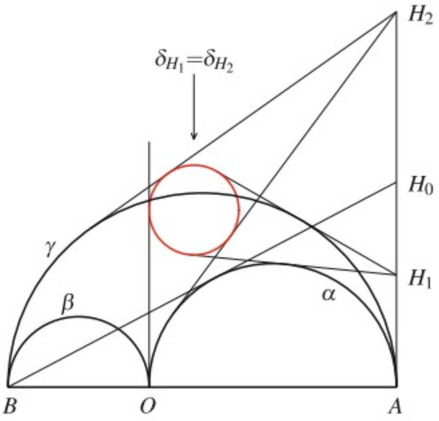

Fig. 65 A generalization

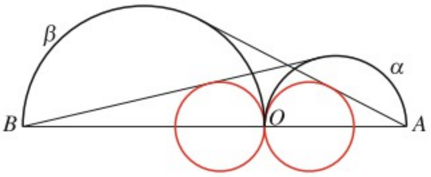

Fig. 66 Nagata's Archimedean circles

many Archimedean circles are also obtained from points on the common tangent of β and γ at B.

A Sangaku Problem Involving Two Archimedean Circles

We consider the next problem proposed by Nagata described in the sangaku hung in 1831 at Atsuta Shrine in Nagoya (Morishima, no date) (see Fig. 66).

Problem 17. Show that the circle touching α at O internally and the tangent of β from A is congruent to the circle touching β at O internally and the tangent of α from B.

The common radius of the two circles equals r_{A} in the problem. However the problem can be solved without expressing the radius of the two circles in terms of a and b. Therefore we cannot be sure that Nagata noticed that the radius of the two small circles equaled r_{A}. In any case, Nagata's Archimedean circles should also be noticed in the history of studying Archimedean circles. The problem is stated in a slightly generalized form in Fukagawa and Pedoe (1989), and a generalization of this can be found in Okumura (1995).

We give another generalization of the problem. For two points P and Q, we denote the circle of diameter PQ by (PQ). Let P and Q be points on the line AB lying in the order P, B, A, Q, or B, P, Q, A or B, Q, P, A such that O is the point

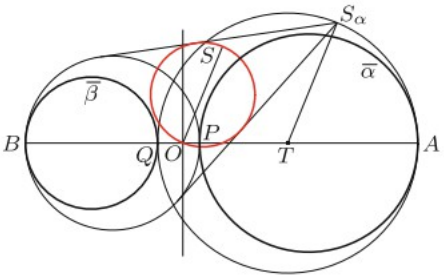

Fig. 67 A generalization

of intersection of AB and the radical axis of the circles (AP) and (BQ). In this case we call the configuration consisting of the three circles $\overline{\alpha} = (AP)$, $\overline{\beta} = (BQ)$, and $\gamma = (AB)$ a collinear arbelos and denote by $(\overline{\alpha}, \overline{\beta}, \gamma)$. Let s and t be the radii of the circles (AQ) and (BP), respectively. Circles of radius $st/(s+t)$ are called Archimedean circles of $(\overline{\alpha}, \overline{\beta}, \gamma)$. The collinear arbelos and its Archimedean circles coincide with the ordinary arbelos and its Archimedean circles if $P = Q = O$. The problem is generalized as follows (Okumura, 2013) (see Fig. 67).

Theorem 21. *For a collinear arbelos $(\overline{\alpha}, \overline{\beta}, \gamma)$ and a point $S \neq O$, let S_α be the external center of similitude of the circles (OS) and (BP), where (OS) does not touch the circle β internally in the case $P = Q = O$. Then (OS) is an Archimedean circle of $(\overline{\alpha}, \overline{\beta}, \gamma)$ if and only if the point S_α lies on the circle (AQ) and the vectors $\overrightarrow{OS}$ and $\overrightarrow{TS_\alpha}$ are parallel with the same direction, where T is the center of (AQ).*

Since there are infinitely many choices of the point S_α, we get infinitely many Archimedean circles of $(\overline{\alpha}, \overline{\beta}, \gamma)$. Similar infinitely many Archimedean circles can also be obtained from points on the circle (BP). Another generalization of Problem 17 to a generalized arbelos called a skewed arbelos can also be found in Okumura (2013).

Wasan Geometry and Division by Zero

In this section we consider applications of recently made definition of division by zero (Kuroda et al., 2014) to Wasan geometry. The definition is stated as follows:

Definition 1.

$$\frac{z}{0} = 0 \text{ for any real number } z. \tag{5}$$

In other words, division is stated by

$$z \div x = z \cdot x^{-1} \quad (x \neq 0), \quad z \div x = z \cdot x \quad (x = 0),$$

which can also be considered in any field. Since the definition is very primitive, we have to check whether it is meaningful or not to our mathematics. The definition should be discarded if it can conclude less meaningful results. Hence we have to consider what the definition implies carefully. While Saburou Saitoh, who advocated the definition more than five years ago, has got more than a thousand meaningful results derived by the definition at the present time of writing. Meanwhile a recent study using the proof assistant Isabelle/HOL also derived (5) (Nipkow et al., 2015). In this section we will show two simple applications of the definition to Wasan geometry. For an extensive introduction of division by zero, see Okumura and Saitoh (2018b), and its application to Wasan geometry, see Kanai and Okumura (2017), Matsuura et al. (2019), Okumura (2018f, 2019h), Okumura and Saitoh (2018a).

The Configuration $A(1)$

In Theorem 1, we did not consider the case $n = 1$ for the configuration $A(n)$ consisting of a circle α of radius a with a tangent t and the other tangent parallel to t, because the radius of a line could not be considered (see Fig. 12). Circles and lines are represented by the equation $e(x^2 + y^2) + 2fx + 2gy + h = 0$. If $e \neq 0$, it represents a circle of radius

$$\sqrt{\frac{f^2 + g^2 - eh}{e^2}}.$$

Hence we can consider that *the radius of a line equals* 0 by (5), since the equation expresses a line if $e = 0$. Hence if $n = 1$, the left side of (1) equals $a/0 = 0$ by (5). Therefore (1) still holds for $A(1)$.

A Three-Circle Problem

We consider the following problem (Aida, 1788; Okumura, 2018c) (see Fig. 68).

Problem 18. Assume that t and u are parallel tangents of a circle β of radius b, α is a circle of radius a touching β externally and t from the same side as β, and γ is the circle of radius c touching α and β externally and u from the same side as β. Show that the following relation holds:

$$c = \frac{b^2}{4a}. \tag{6}$$

We consider the case $a = 0$ for the problem, where notice that the assumption $b > a$ stated in Fukagawa and Pedoe (1989) is unnecessary. Let O be the point of

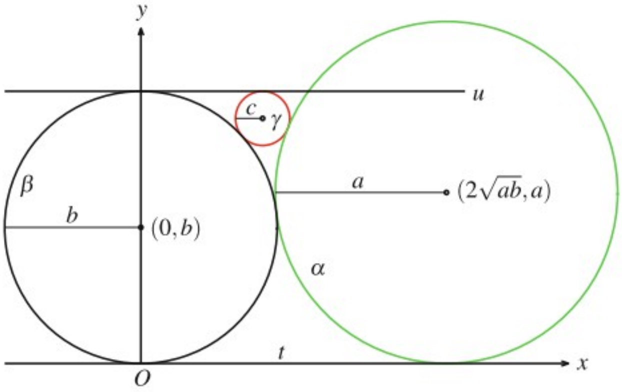

Fig. 68 A three-circle problem

Fig. 69 The circle α (dotted lines and the point)

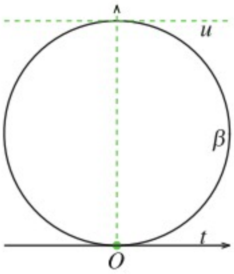

tangency of β and t. We use a rectangular coordinate system with origin O such that the center of β has coordinates $(0, b)$. We may assume that the circle α has an equation

$$\left(x - 2\sqrt{ab}\right)^2 + (y - a)^2 - a^2 = 0 \tag{7}$$

by Proposition 1. The equation can be arranged as

$$\frac{x^2 + y^2}{\sqrt{a}} - 4x\sqrt{b} - 2\sqrt{a}(y - 2b) = 0,$$
$$\frac{x^2 + y^2}{a} - 4x\sqrt{\frac{b}{a}} - 2(y - 2b) = 0.$$

Therefore if $a = 0$, we have $x^2 + y^2 = 0$, $x = 0$, and $y - 2b = 0$, respectively, by (5). The three equations express the origin, the y-axis, and the line $y = 2b$ (the line u), respectively (see Fig. 69). We can consider that the origin and the line u have radius 0 and touch β and t, i.e., they satisfy the assumption of the problem. But the y-axis is perpendicular to β and t, though it has radius 0. However we can still consider that *a line perpendicular to a circle and a line touches them*, because (5) implies $\tan(\pi/2) = 0$. Therefore we can consider that the y-axis has radius 0 and touches β and t, i.e., it also satisfies the assumption of the problem.

The circle γ has an equation $(x - 2\sqrt{bc})^2 + (y - 2b + c)^2 = c^2$. Substituting $c = b^2/(4a)$ in this equation and rearranging, we get the following three equations.

$$a(x^2 + (y - 2b)^2) - 2bx\sqrt{ab} + \frac{b^2 y}{2} = 0, \tag{8}$$

$$\sqrt{a}(x^2 + (y - 2b)^2) - 2bx\sqrt{b} + \frac{b^2 y}{2\sqrt{a}} = 0,$$

$$x^2 + (y - 2b)^2 - 2bx\sqrt{\frac{b}{a}} + \frac{b^2 y}{2a} = 0.$$

Hence if $a = 0$, we get $y = 0$, $x = 0$, and $x^2 + (y - 2b)^2 = 0$. The last three equations express the x-axis (the line t), the y-axis, and the point of coordinates $(0, 2b)$, respectively (see Fig. 70).

When α approaches to the origin, γ approaches to the line t. Hence we consider that if α coincides with the origin, then γ coincides with the line t (see Fig. 71). Symmetrically we also consider that if α coincides with the line u, γ coincides with the point of coordinates $(0, 2b)$ (see Fig. 72). Therefore we can consider if α coincides with the y-axis, γ also coincides with the y-axis as the remaining case (see Fig. 73). In all the cases, α and γ satisfy the assumption of the problem. Moreover (6) still holds by (5), since $a = 0$ implies $c = 0$.

Notice that the three cases described in Figs. 71, 72, and 73 are obtained if we consider the left sides of (7) and (8) as polynomials of $\sqrt{a}$ and assume that the constant terms, the coefficients of a, and the coefficients of $\sqrt{a}$ equal 0, respectively. Also notice that the case described in Fig. 73 could not be obtained without (5) before.

The first example for $A(n)$ shows that the definition enables us to include some cases which have been excluded before. The second example shows that the definition still enables us to see some unexpected but meaningful phenomena which cannot be seen without the definition. It seems that studies on such unexpected phenomena using (5) will open the door to the new mathematics in the near feature.

Fig. 70 The circle γ (dotted lines and the point)

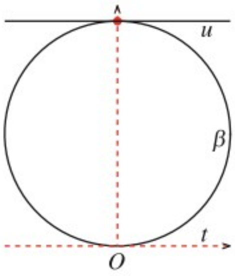

Fig. 71 The case $\alpha = O$

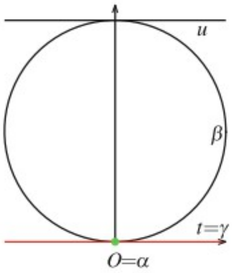

Fig. 72 The case $\alpha = u$

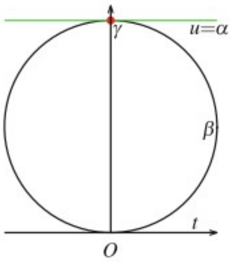

Fig. 73 The case α coinciding with the y-axis

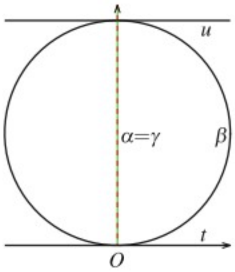

Practical Side

In this section we show several examples of application of Wasan and Wasan geometry. As stated before, many Wasan mathematician also studied astronomy, surveying, and the art of divination. They published many books on such areas. On the other hand, a popular book Jinkōki was published in early time of Edo era, which was a manual for the calculating tool called soroban. It also included other topics such as recreational mathematics and applications of mathematics in daily life (see Fig. 74) (Yoshida, 1632).

Figures 75 and 76 are taken from a book for surveying (Okumura, 1836). Figure 75 shows a frequently used diagram to get values of the six trigonometric functions. If the side length of the square equals 1, then the values of the six trigonometric functions of the angle between the horizontal side of the square and the oblique segment are obtained as the lengths of certain segments in the diagram. For example, the tangent equals the length of the oblique segment.

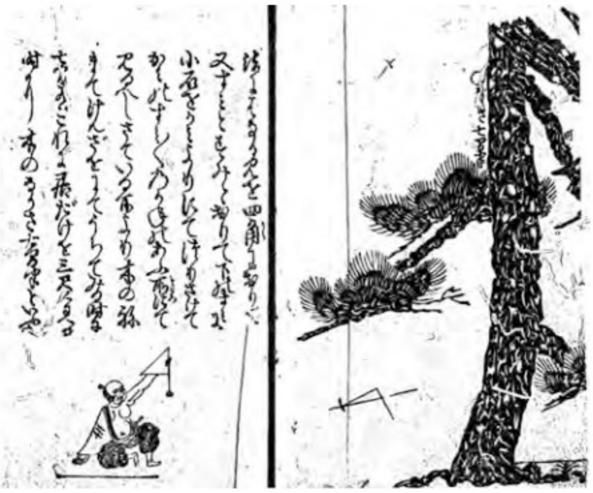

Fig. 74 (Yoshida, 1632)

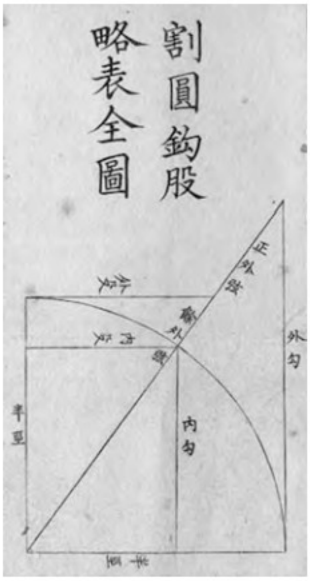

Fig. 75 (Okumura, 1836)

Trigonometric tables were imported in the eighteenth century and used in surveying and astronomy. Figure 77 shows an example of trigonometric table used for surveying written on a fan.

Wasan mathematician also published books for sailing the oceans. Ishiguro published a book of celestial navigation (Ishiguro, 1835). It explains how to use quadrant (see Fig. 78). It also explains how to determine the altitude of the celestial north pole using Triones (see Fig. 79). As we have just seen, the concept of angles was frequently used in surveying and astronomy, but it was not used in Wasan geometry.

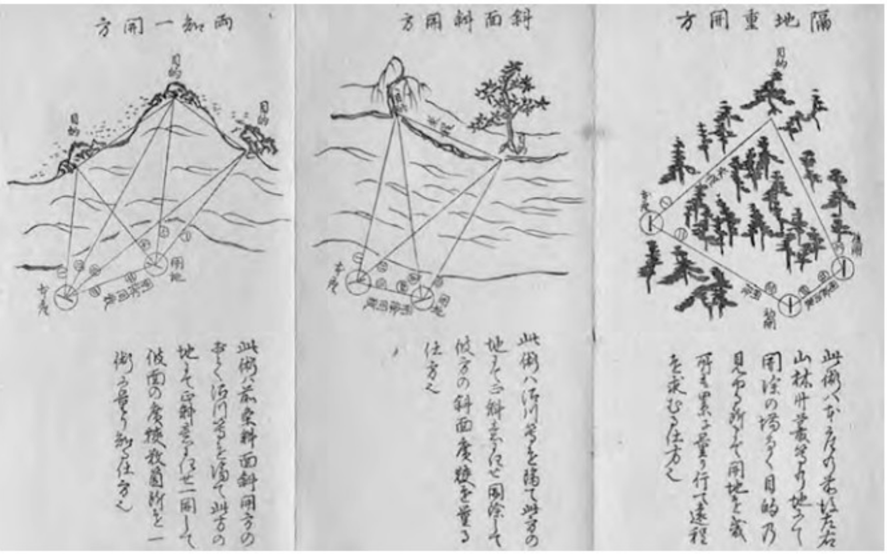
隔地重用方
斜面斜用方
両斜一用方

Fig. 76 (Okumura, 1836)

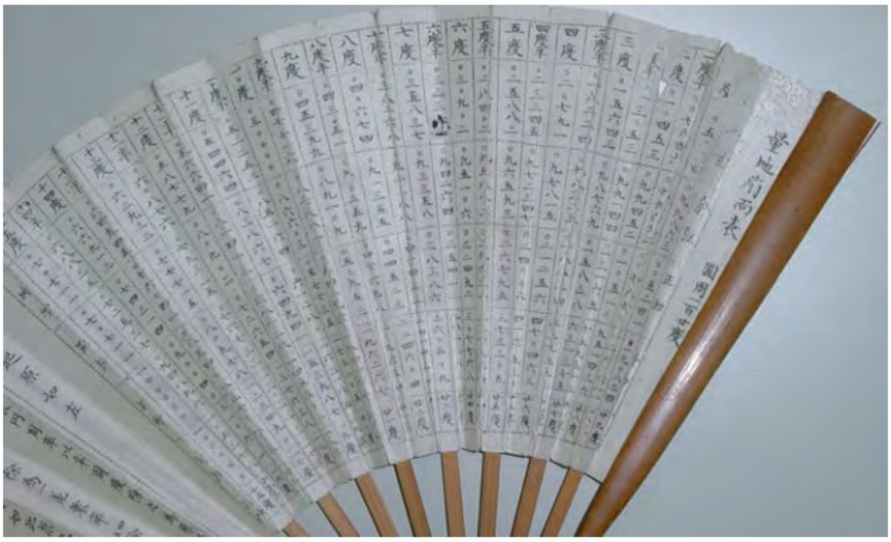

Fig. 77 Trigonometric table written on a fan (Nakasone, 1864)

On December 9 in 1872, Meiji Government declared that the old calendar system would be abolished and new Gregorian calendar system was used from the January first in 1873. Hanai published the book explaining the new calendar system in January in the same year (Hanai, 1873) (see Fig. 80). It seems that the publication was made with remarkably short preparation time.

Fig. 78 (Ishiguro, 1835)

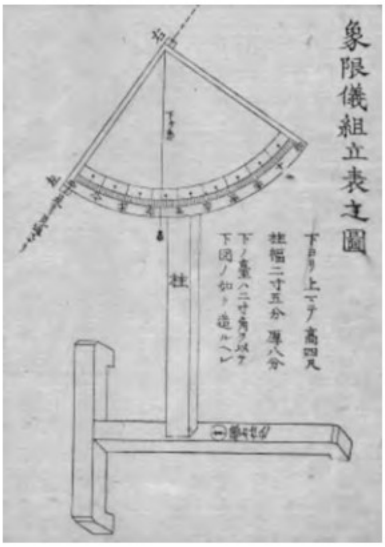

Fig. 79 (Ishiguro, 1835)

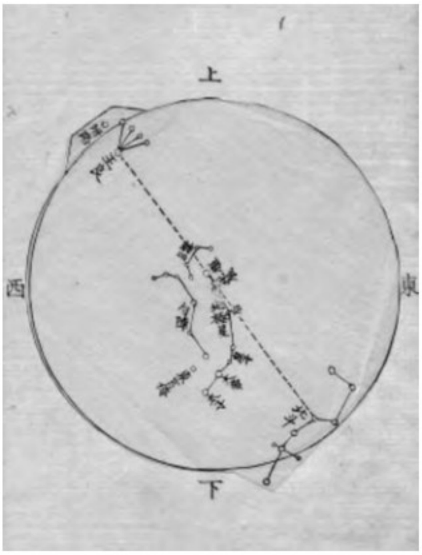

The change of the social system in Japan at the beginning of the Meiji era was so drastic and urgent as this example shows, but Wasan mathematicians supported it well. Surveying was done in all over the country due to the land-tax reformation in the same year 1873 with the aid of many Wasan mathematicians in each of the local areas.

Japanese traditional buildings were made of wood. Figure 81 is taken from a pattern book for carpenters (Suzuki, 1721), in which the lengths of several parts are indicated. Figure 82 is also taken from the same book, which shows architectural

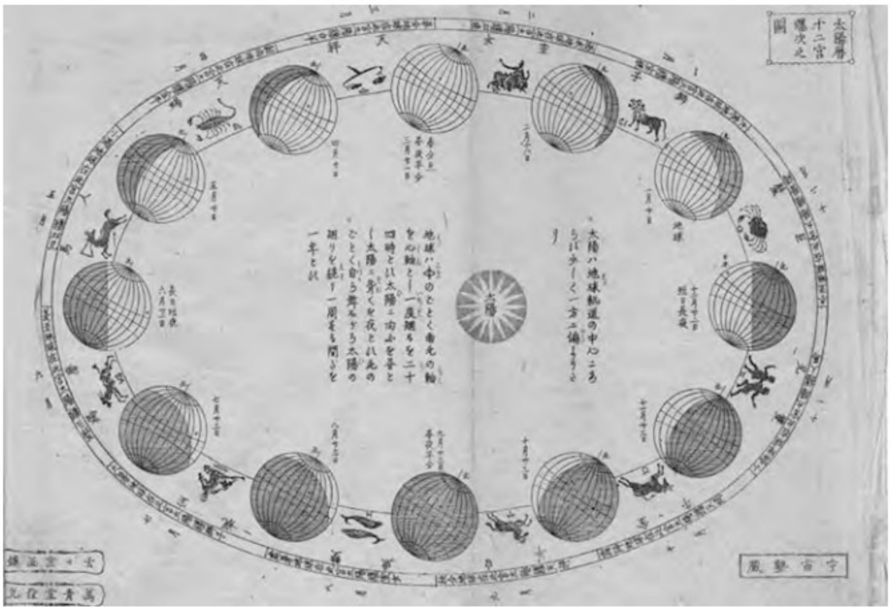

Fig. 80 (Hanai, 1873)

Fig. 81 (Suzuki, 1721)

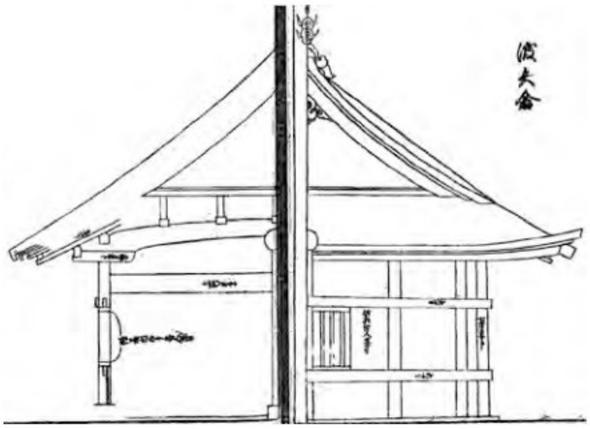

ornaments. L-shaped carpenter's square were used as the main tool to design each of the architectural elements. Its basic principle has been supposed to be the similarity of right triangles, but the contents were not disclosed as secret methods used in each of the families of carpenters. Meanwhile Heinouchi published two books explaining how to use the carpenter's square in terms of Wasan geometry (Heinouchi, 1848, 1833). He gave a mathematical foundation of the methods and made them accessible. Figure 83 is taken from his 1833 book.

Fig. 82 (Suzuki, 1721)

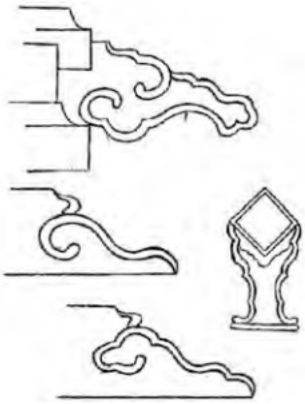

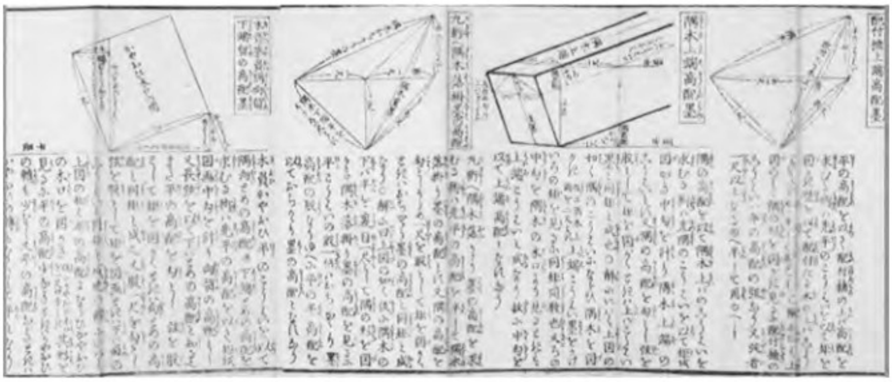

Fig. 83 (Heinouchi, 1833)

Study of Wasan Geometry: Past and Present

Wasan was in decline through Meiji era, and several attempts to gain perspective on Wasan were made. Tsuruichi Hayashi studied Wasan mathematically and published many papers including many geometric ones. Seki's famous work on determinant was discovered by him. Hayashi's works were collected and published as a book in two volumes consisting of about two thousands of total number of pages (Hayashi, 1937).

One of the most important works on Wasan geometry in the second half of the twentieth century was made by the geometer Shikō Iwata, who published a dictionary of geometry in six volumes with two supplement volumes (Iwata, 1993, 1988, 1982, 1980, 1978, 1976, 1974, 1971). It is a collection of results in plane and solid geometry including a huge number of results in Wasan geometry together with their modern proofs and references. The total number of pages of the 8 volumes is more than 4000 pages.

The works of notable Wasan mathematicians such as Seki were collected and published and studied by modern mathematicians and historians. However Wasan geometry were contributed by a lot of mathematicians and amateurs. The results were neither collected nor classified. As if this isn't bad enough, there were some difficulties to access Wasan books and manuscripts because of their preciousness. Thereby it had been difficult to access and check enough number of results of Wasan geometry till the end of the twentieth century. The opening of the several digital libraries at the beginning of the twenty-first century mentioned in the section of Wasan is one of the most noticeable changes for those who want to study Wasan and Wasan geometry, because it makes an immense amount of Wasan materials accessible to all the people. It seems that the physical environment to study Wasan geometry is made up for all the people in the world today.

References

Aida (会田安明) (ed) (no date) Sampō Chikusaku Jutsu (算法逐索術), vol 2. Tohoku University Digital Collection

Aida (会田安明) (ed) (1788) Sampō Tenshōhō (算法天生法). Tohoku University Digital Collection

Akabane (赤羽千鶴) et al (ed) (1998) The sangaku in Nagano prefecture (enlarged and revised edition) (増補長野県の算額). Kyōiku Shokan (教育書館)

Akita (秋田義蕃) (ed) (1833) Sampō Tenzan Tebikigusa Furoku (算法点竄手引草附録). Tohoku University Digital Collection

Bankoff L (1974) Are the twin circles of Archimedes really twins?. Math Mag 72: 214–218

Dodge C W, Schoch T, Woo P Y, Yiu P (1999) Those ubiquitous Archimedean circles. Math Mag 72: 202–213

Fujisawa (藤沢忠親) (ed) Sampō Shinsho (算法審書). Tohoku University Digital Collection

Fukagawa H, Pedoe D (1989) Japanese Temple Geometry Problems. Charles Babbage Research Centre Winnipeg Canada

Fukushimaken Wasan Kenkyū Hozonkai (福島県和算研究保存会) (ed) (1989) The sangaku in Fukushima (福島の算額). Sōju Shuppan (蒼樹出版) Fukushima

Furuya (古谷道生) (ed) (1854) Sampō Tsūsho (算法通書). Tohoku University Digital Collection

Gunma Wasan Kenkyūkai (群馬和算研究会) (ed) (1987) The sangaku in Gunma (群馬の算額). Gunma Wasan Kenkyūkai

Hanai (花井静) (ed) (1873) Taiyō Reki Zokkai (太陽暦俗解). In Wasansho Shūsei (和算書集成) (Wasan collection) Okumura H (ed) (2001) Iwanami Shoten

Hayashi T (林鶴一) (1937) Collected papers on the old Japanese Mathematics (和算研究集録) volumes 1 and 2. Tokyo Kaiseikan (東京開成館)

Heinouchi (平内廷臣) (ed) (1848) Shouka Kujutsu Shinsho (匠家矩術新書). Tohoku University Digital Collection (non-online)

Heinouchi (平内廷臣) (ed) (1833) Shouka Kujutsu Youkai (匠家矩術要解). Kotenseki Sogo Database in Waseda University

Hirayama (平山諦), Matsuoka (松岡元久) (1966) The sangaku in Yamagata (山形の算額).

Ishiguro (石黒信由) (1835) Tokai Hyōteki (渡海標的). In Wasansho Shūsei (和算書集成) (Wasan collection) Okumura H (ed) (2001) Iwanami Shoten

Iwai (岩井好算) (no date) Keihoukei Kou (圭宝形考). Tohoku University Digital Collection

Iwata S (岩田至康) (1993) The unabridged dictionary of geometry (幾何学大辞典) supplement, vol 2. Maki Shoten (槇書店)

Iwata S (岩田至康) (1988) The unabridged dictionary of geometry (幾何学大辞典) supplement, vol 1. Maki Shoten (槇書店)

Iwata S (岩田至康) (1982) The unabridged dictionary of geometry (幾何学大辞典), vol 6. Maki Shoten (槇書店)
Iwata S (岩田至康) (1980) The unabridged dictionary of geometry (幾何学大辞典), vol 5. Maki Shoten (槇書店)
Iwata S (岩田至康) (1978) The unabridged dictionary of geometry (幾何学大辞典), vol 4. Maki Shoten (槇書店)
Iwata S (岩田至康) (1976) The unabridged dictionary of geometry (幾何学大辞典), vol 3. Maki Shoten (槇書店)
Iwata S (岩田至康) (1974) The unabridged dictionary of geometry (幾何学大辞典), vol 2. Maki Shoten (槇書店)
Iwata S (岩田至康) (1971) The unabridged dictionary of geometry (幾何学大辞典), vol 1. Maki Shoten (槇書店)
Jobbings A K (2011) Two semicircles fill half a circle. Math Gaz 95:538–540
Kanai Y, Okumura H (2017) A three tangent congruent circle problem. Sangaku J Math 1:16–20
Kinki Sūgakushigakukai (近畿数学史学会) (ed) (1992) The sangaku in Kinki (近畿の算額). Osaka Kyōiku Tosho (大阪教育図書)
Kimura (木村尚寿) (1828) Onchi Sansō (温知算叢). Tohoku University Digital Collection
Kuroda M, Michiwaki H, Saitoh S, and Yamane M (2014) New meanings of the division by zero and interpretations on $100/0 = 0$ and on $0/0 = 0$. Int J Appl Math 27:191–198
Matsuura T, Okumura H, Saitoh S (2019) Division by zero calculus and Pompe's theorem. Sangaku J Math 3:36–40
Mikami Y (1913-2) The parabola and hyperbola in Japanese mathematics. Tohoku Mathematical Journal, First Series 3:29–37
Mikami Y (1913) The development of Mathematics in China and Japan. Chelsea, New York
Morishima (森島忠造) (ed) (no date) Kansai Shachū Sansou (簡斎社中算草). Tohoku University Digital Collection
Nakasone (中曽根) (1864) Ryōchi Senmen Hyō (量地扇面表). In Wasansho Shūsei (和算書集成) (Wasan collection) Okumura H (ed) (2001) Iwanami Shoten
Nipkow T, Paulson L C, Wenzel M (2015) A Proof Assistant for Higher-Order Logic. Springer-Verlag
Okayu (御粥安本) (ed) (1855) Hōnōcho Sekisensei Sandai Kujō (奉納箸隻先生算題九条). Tohoku University Digital Collection
Okumura (奥村喜三郎) (1836) Ryōchi Kodo Sampō (量地弧度算法). In Wasansho Shūsei (和算書集成) (Wasan collection) Okumura H (ed) (2001) Iwanami Shoten
Okumura H (1995) Two similar triangles. Math Gaz 79:569–571
Okumura H (1999) Geometries in the East and the West in the 19th century. Symmetry: Culture and Science : 9:189–197
Okumura H (2013) Archimedean circles of the collinear arbelos and the skewed arbelos. J Geom Graph 17:31–52
Okumura H (2014) The arbelos with overhang. KoG 18:19–27
Okumura H (2017a) Configurations of congruent circles on a line. Sangaku J Math 1:24–34
Okumura H (2017b) Theorems on two congruent circles on a line. Sangaku J Math 1:35–38
Okumura H (2018a) A note on a problem involving a square in a curvilinear triangle. Sangaku J Math 2:3–5
Okumura H (2018b) A note on an isosceles triangle containing a square and three congruent circles. Sangaku J Math 2:8–10
Okumura H (2018c) Wasan geometry with the division by 0. Int J Geom 8:17–20
Okumura H (2018d) Solution to 2017-3 Problem 5. Sangaku J Math 2:17–21
Okumura H (2018e) Solution to the problem proposed in "Solution to 2017-3 Problem 5". Sangaku J Math 2:24–26
Okumura H (2018f) Solution to 2017-1 Problem 4 with division by zero. Sangaku J Math 2: 27–30
Okumura H (2018g) Solution to Problem 2018-3-2. Sangaku J Math 2:54–56
Okumura H (2019a) A note on the arbelos in Wasan geometry, a problem in Sampō Tenzan Tebikigusa Furoku. Sangaku J Math 3:12–14

Okumura H (2019b) A note on the arbelos in Wasan geometry, Satoh's problem. Sangaku J Math 3:15–16

Okumura H (2019c) A note on the arbelos in Wasan geometry, Matsuda's problem. Sangaku J Math 3:19–23

Okumura H (2019d) The arbelos in Wasan geometry, problems of Izumiya and Naitō. J Class Geom 4:1–5

Okumura H (2019e) The arbelos in Wasan geometry, Tamura's problem. Glob J Adv Res Class Mod Geom 8:33–36

Okumura H (2019f) A note on the arbelos in the Wasan geometry: Satoh's problem and a circle pattern. Mathematics and Informatics 62:301–304

Okumura H (2019g) Arbeloi determined by a chord and solutions to problems 2017-3-8 and 2019-3-4. Sangaku J Math 3:41–50

Okumura H (2019h) A characterization of the golden arbelos involving an Archimedean circle. Sangaku J Math 3:67–71

Okumura H (2019i) The arbelos in Wasan geometry: Chiba's problem. Glob J Adv Res Class Mod Geom 8:98–104

Okumura H (2019j) The arbelos in Wasan geometry: Ootoba's problem and Archimedean circles. Sangaku J Math 3:91–97

Okumura H, Saitoh S (2018a) Applications of the division by zero calculus to Wasan geometry. Glob J Adv Res Class Mod Geom 7:44–49

Okumura H, Saitoh S (2018b) Wasan geometry and division by zero calculus. Sangaku J Math 2:57–73

Okumura H, Sodeyama C (1998) A surprising property of successively touching circles. Mathematics Plus: 6:17–18

Ono (小野以正) (ed) (1855) Keiteki Sampō Shinan Taisei (啓廸算法指南大成). Tohoku University Digital Collection.

Saitama Prefectural Library (埼玉県立図書館) (ed) (1969) The sangaku in Saitama (埼玉の算額). Saitama Prefectural Library (埼玉県立図書館)

Shimura (志村昌義) et al (ed) (no date) Kiōshū (淇澳集). Tohoku University Digital Collection

Smith D E, Mikami Y (1914) A history of Japanese mathematics. Open Court Chicago

Suzuki (鈴木重春) (ed) (1721) Daiku Tekagami (大匠手鑑). Tohoku University Digital Collection

Takeda (武田真元) (1825) Sampō Benran (算法便覧). Tohoku University Digital Collection

Takigawa (滝川有人) et al (ed) Mishō Sampō (未詳算法), vol 9. Tohoku University Digital Collection

Toyoyoshi (豊由周齋) (ed) (no date) Chikusaku (逐索). Digital Library Department of Mathematics Kyoto University

Yamamoto (山本賀前) (ed) (1841) Sampō Jojutsu (算法助術). In Wasansho Shūsei (和算書集成) (Wasan collection) Okumura H (ed) (2001) Iwanami Shoten

Yasutomi Y (安富有恒) (ed) (1987) The extant sangaku in Iwate (和算–岩手の現存算額のすべて). Seijisha (青磁社), Tokyo

Yoshida Y (吉田光由) (1632) Jinkōki (塵劫記). Tohoku University Digital Collection

no author's name (no datea) Kururisha Sandaishū (久留里社算題集). Tohoku University Digital Collection

no author's name (no dateb) Sendai Kokuuzoudou Sandai (仙台虚空蔵堂算題). Tohoku University Digital Collection

M-FPG-augmentations

→ M-morphism such as

in the square (1) diagram below.

- Handbook of Financial Econometrics, Mathematics.../

Handbook.....(2020).pdf

2) Books-018-2021/20210915/

- Mathematics of the Arts.../

Bharath Sriraman - Handbook of the Mathematics ... (2021).pdf

pag. 711(725) →

- Wasan Geometry

→ important.

3) Books-018-2021/20210915/

Handbook of Mathematics/

- Handbook of Elliptic and Hyperelliptic Curve...pdf

842 pages - 2006,

ISBN 978-1-58488-518-4.

+,-.+ + +/ ! + %" ?! #1 C 9 0 !
?30 0 6 30 !0 3 0#
+,-.+ + +/ ! + %" ?! #1 C 9 0 !
?30 0 6 30 !0 3 0#
* ' % (((,

Chapter 13

Arithmetic of Elliptic Curves

Christophe Doche and Tanja Lange

Contents in Brief

13.1 Summary of background on elliptic curves 268
First properties and group law • Scalar multiplication • Rational points • Torsion points • Isomorphisms • Isogenies • Endomorphisms • Cardinality

13.2 Arithmetic of elliptic curves defined over $\mathbb{F}_p$ 280
Choice of the coordinates • Mixed coordinates • Montgomery scalar multiplication • Parallel implementations • Compression of points

13.3 Arithmetic of elliptic curves defined over $\mathbb{F}_{2^d}$ 289
Choice of the coordinates • Faster doublings in affine coordinates • Mixed coordinates • Montgomery scalar multiplication • Point halving and applications • Parallel implementation • Compression of points

Elliptic curves constitute one of the main topics of this book. They have been proposed for applications in cryptography due to their fast group law and because so far no subexponential attack on their discrete logarithm problem (cf. Section 1.5) is known. We deal with security issues in later chapters and concentrate on the *group* arithmetic here. In an actual implementation this needs to be built on an efficient implementation of finite field arithmetic (cf. Chapter 11).

In the sequel we first review the background on elliptic curves to the extent needed here. For a more general presentation of elliptic curves, see Chapter 4. Then we address the question of efficient implementation in large odd and in even characteristics. We refer mainly to [HaMe$^+$ 2003] for these sections.

Note that there are several softwares packages or libraries able to work on elliptic curves, for example `PARI/GP` [Pari] and **apecs** [Apecs]. The former is a linkable library that also comes with an interactive shell, whereas the latter is a `Maple` package. Both come with full sources. The computer algebra systems `Magma` [Magma] and `SIMATH` [Simath] can deal with elliptic curves, too.

Elliptic curves have received a lot of attention throughout the past almost 20 years and many papers report experiments and timings for various field sizes and coordinates. We do not want to repeat the results but refer to [Ava 2004a, CoMi$^+$ 1998] and Section 14.7 for odd characteristic and [HaLó$^+$ 2000, LóDa 1998, LóDa 1999] for even characteristic. Another excellent and comprehensive reference comparing point multiplication costs and implementation results is [HaMe$^+$ 2003, Tables 3.12, 3.13 and 3.14 and Chap. 5].

13.1 Summary of background on elliptic curves

13.1.1 First properties and group law

We start with a practical definition of the concept of an elliptic curve.

Definition 13.1 An *elliptic curve* E *over a field* K denoted by E/K is given by the *Weierstraß equation*

$$E : y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6 \tag{13.1}$$

where the coefficients $a_1, a_2, a_3, a_4, a_6 \in K$ are such that for each point (x_1, y_1) with coordinates in $\overline{K}$ satisfying (13.1), the partial derivatives $2y_1 + a_1x_1 + a_3$ and $3x_1^2 + 2a_2x_1 + a_4 - a_1y_1$ do not vanish simultaneously.

The last condition says that an elliptic curve is *nonsingular* or *smooth*. A point on a curve is called *singular* if both partial derivatives vanish (cf. the Jacobi criterion 4.94). For shorter reference we group the coefficients in (13.1) to the equation

$$E : y^2 + h(x)y = f(x), \quad h(x), f(x) \in K[x], \quad \deg(h) \leqslant 1, \deg(f) = 3 \text{ with } f \text{ monic.}$$

The smoothness condition can also be expressed more intrinsically. Indeed, let

$$\begin{aligned} &b_2 = a_1^2 + 4a_2, \;\; b_4 = a_1a_3 + 2a_4,\\ &b_6 = a_3^2 + 4a_6, \;\; b_8 = a_1^2a_6 - a_1a_3a_4 + 4a_2a_6 + a_2a_3^2 - a_4^2. \end{aligned}$$

In odd characteristic, the transformation $y \mapsto y - (a_1x + a_3)/2$ leads to an isomorphic curve given by

$$y^2 = x^3 + \frac{b_2}{4}x^2 + \frac{b_4}{2}x + \frac{b_6}{4}\cdot \tag{13.2}$$

The cubic polynomial above has only simple roots over the algebraic closure $\overline{K}$ if and only if its discriminant is nonzero. The equation of the discriminant is therefore useful to determine if (13.2) is an elliptic curve or not. In addition, it is relevant for characteristic 2 fields as well.

Definition 13.2 Let E be a curve defined over K by (13.1) and let b_2, b_4, b_6 and b_8 as above. The *discriminant of the curve* E denoted by Δ satisfies

$$\Delta = -b_2^2b_8 - 8b_4^3 - 27b_6^2 + 9b_2b_4b_6.$$

The curve E is nonsingular, and thus is an elliptic curve, if and only if Δ is nonzero. In this case, we introduce the j-*invariant* of E, that is $j(E) = (b_2^2 - 24b_4)^3/\Delta$.

Example 13.3 In $\mathbb{F}_p$ with $p = 2003$, an elliptic curve is given by

$$E_1 : y^2 + 2xy + 8y = x^3 + 5x^2 + 1136x + 531. \tag{13.3}$$

Indeed, we have $b_2 = 24, b_4 = 285, b_6 = 185, \Delta = 1707 \neq 0$ and $j = 171$.

We now show how to turn the set of points of E into a group with group operation denoted by $\oplus$. For this we visualize it over the reals as in Figure 13.1 and assume $h(x) = 0$.

Figure 13.1 Group law on elliptic curve $y^2 = f(x)$ over $\mathbb{R}$.

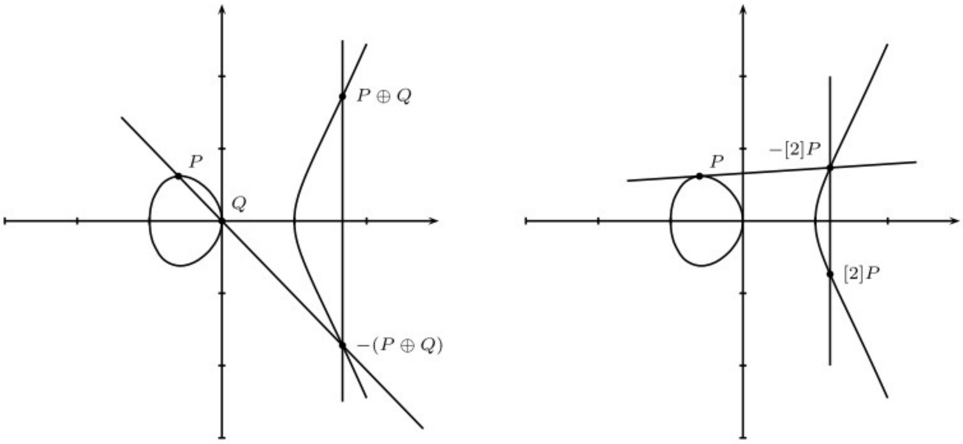

To add two points $P = (x_1, y_1)$ and $Q = (x_2, y_2)$ in general position one draws a line connecting them. There is a third point of intersection. Mirroring this point at the x-axis gives the sum $P \oplus Q$. The same construction can be applied to double a point where the connecting line is replaced by the tangent at P.

Furthermore, we need to define the sum of two points with the same x-coordinate since for them the group operation cannot be performed as stated. As $y^2 = f(x)$ there are at most 2 such points (x_1, y_1) and $(x_1, -y_1)$. Furthermore, we have to find the neutral element of the group.

The way out is to include a further point P_∞ called the *point at infinity*. It can be visualized as lying far out on the y-axis such that any line $x = c$, for some constant c, parallel to the y-axis passes through it. This point is the neutral element of the group. Hence, the line connecting (x_1, y_1) and $(x_1, -y_1)$ passes through P_∞. As it serves as the neutral element, the inflection process leaves it unchanged such that $(x_1, y_1) \oplus (x_1, -y_1) = P_\infty$, i.e., $(x_1, -y_1) = -P$.

This explanation might sound a little like hand-waving and only applicable to $\mathbb{R}$. We now derive the addition formulas for an arbitrary field K, which hold universally. For a proof we refer to Chapter 4.

Take $P \neq Q$ with $x_1 \neq x_2$ as above and let us compute the coordinates of $R = P \oplus Q = (x_3, y_3)$. The intersecting line has slope

$$\lambda = \frac{y_1 - y_2}{x_1 - x_2}$$

and passes through P. Its equation is thus given by

$$y = \lambda x + \frac{x_1 y_2 - x_2 y_1}{x_1 - x_2}.$$

We denote the constant term by μ and remark $\mu = y_1 - \lambda x_1$. The intersection points with the curve are obtained by equating the line and E

$$(\lambda x + \mu)^2 + (a_1 x + a_3)(\lambda x + \mu) = x^3 + a_2 x^2 + a_4 x + a_6.$$

This leads to the equation $r(x) = 0$ where

$$r(x) = x^3 + (a_2 - \lambda^2 - a_1\lambda)x^2 + (a_4 - 2\lambda\mu - a_3\lambda - a_1\mu)x + a_6 - \mu^2 - a_3\mu.$$

We already know two roots of $r(x)$, namely the x-coordinates of the other two points. Since

$$r(x) = (x - x_1)(x - x_2)(x - x_3)$$

one has $\lambda^2 + a_1\lambda - a_2 = x_1 + x_2 + x_3$. As x_1, x_2 are defined over K so is x_3 and $\tilde{y}_3 = \lambda x_3 + \mu$. The inflection at the x-axis has to be translated to the condition that the second point has the same x-coordinate and also satisfies the curve equation. We observe that if $P = (x_1, y_1)$ is on the curve then so is $(x_1, -y_1 - a_1x_1 - a_3)$, which corresponds to $-P$ since the point at infinity is the neutral element for this law. Accordingly, we find $y_3 = -\lambda x_3 - \mu - a_1 x_3 - a_3$.

Doubling $P = (x_1, y_1)$ works just the same with the slope obtained by implicit derivating. Thus we have $P \oplus Q = (x_3, y_3)$ and

$$\begin{aligned} -P &= (x_1, -y_1 - a_1x_1 - a_3), \\ P \oplus Q &= (\lambda^2 + a_1\lambda - a_2 - x_1 - x_2, \lambda(x_1 - x_3) - y_1 - a_1x_3 - a_3), \text{ where} \\ \lambda &= \begin{cases} \dfrac{y_1 - y_2}{x_1 - x_2} & \text{if } P \neq \pm Q, \\ \dfrac{3x_1^2 + 2a_2x_1 + a_4 - a_1y_1}{2y_1 + a_1x_1 + a_3} & \text{if } P = Q. \end{cases} \end{aligned}$$

It is immediate from the pictorial description that this law is commutative, has the point at infinity as neutral element, and that the inverse of (x_1, y_1) is given by $(x_1, -y_1 - a_1x_1 - a_3)$. The associativity can be shown to hold by simply applying the group law and comparing elements. We leave the lengthy computation to the reader. Note that Chapter 4 gives extensive background showing in an abstract way the group of points on E to form a group. For a more geometrical proof, relying on Bezout's theorem, see e.g., [CAS 1991].

Example 13.4 One can easily check that the points $P_1 = (1118, 269)$ and $Q_1 = (892, 529)$ lie on the curve $E_1/\mathbb{F}_p$ as defined by (13.3). Then

$$\begin{aligned} -P_1 &= (1118, 1493), \\ P_1 \oplus Q_1 &= (1681, 1706), \\ [2]P_1 &= (1465, 677) \end{aligned}$$

are also on E_1.

The point at infinity can be motivated by giving an alternative description of elliptic curves. Equation (13.1) expresses the curve in *affine coordinates*. The same elliptic curve E in *projective coordinates* is then given by the equation

$$E : Y^2Z + a_1XYZ + a_3YZ^2 = X^3 + a_2X^2Z + a_4XZ^2 + a_6Z^3.$$

Let us denote by $(X_1 : Y_1 : Z_1)$ an element of the *projective* 2*-space* $\mathbb{P}^2/K$, i.e., a class of $\overline{K}^3 \setminus \{(0,0,0)\}$ modulo the relation

$$(X_1 : Y_1 : Z_1) \sim (X_2 : Y_2 : Z_2) \iff \text{there is } \lambda \in \overline{K}^* \mid X_2 = \lambda X_1,\ Y_2 = \lambda Y_1 \text{ and } Z_2 = \lambda Z_1.$$

By abuse of notation, we identify a class with any of its representatives and call $(X_1 : Y_1 : Z_1)$ a *projective point*. We remark that only a single point of E satisfies $Z_1 = 0$, namely the *point at infinity*, which is in this case $P_\infty = (0 : 1 : 0)$. When $Z_1 \neq 0$, there is a simple correspondence between the projective point $(X_1 : Y_1 : Z_1)$ and the affine point (x_1, y_1) using the formula

$$(x_1, y_1) = (X_1/Z_1, Y_1/Z_1) \tag{13.4}$$

As the representation $(X_1 : Y_1 : Z_1)$ is not normalized, one can perform arithmetic in projective coordinates without any inversion. Note also that *generalized projective coordinates* involving suitable powers of Z_1 in (13.4) are commonly used, cf. Sections 13.2.1 and 13.3.1.

Example 13.5 The point $P_1' = (917 : 527 : 687)$ lies on the curve E_1 of equation (13.3) expressed in projective coordinates, i.e.,

$$E_1 : Y^2Z + 2XYZ + 8YZ^2 = X^3 + 5X^2Z + 1136XZ^2 + 531Z^3.$$

In fact, P_1' is in the same class as $(1118 : 269 : 1)$ and thus corresponds to the affine point $P_1 = (1118, 269)$.

13.1.2 Scalar multiplication

Take $n \in \mathbb{N} \setminus \{0\}$ and let us denote the *scalar multiplication by n on E* by $[n]$, or $[n]_E$ to avoid confusion. Namely,

$$\begin{aligned} [n] : E &\to E \\ P &\mapsto [n]P = \underbrace{P \oplus P \oplus \cdots \oplus P}_{n \text{ times}}. \end{aligned}$$

This definition extends trivially to all $n \in \mathbb{Z}$, setting $[0]P = P_\infty$ and $[n]P = [-n](-P)$ for $n < 0$. Chapter 9 deals with exponentiation, i.e., the computation of x to some power n. In the context of elliptic curves, this corresponds to $[n]P$. Thus multiplications, squarings, and divisions are replaced by additions, doublings, and subtractions on E.

As an example, we give the analogue of Algorithm 9.10 with additive notation.

Algorithm 13.6 Sliding window scalar multiplication on elliptic curves

INPUT: A point P on an elliptic curve E, a nonnegative integer $n = (n_{l-1} \dots n_0)_2$, a parameter $k \geqslant 1$ and the precomputed points $[3]P, [5]P, \dots, [(2^k - 1)]P$.
OUTPUT: The point $[n]P$.

```
$Q \leftarrow P_\infty$ and $i \leftarrow l - 1$
while $i \geqslant 0$ do
    if $n_i = 0$ then $Q \leftarrow [2]Q$ and $i \leftarrow i - 1$
    else
        $s \leftarrow \max(i - k + 1, 0)$
        while $n_s = 0$ do $s \leftarrow s + 1$
        for $h = 1$ to $i - s + 1$ do $Q \leftarrow [2]Q$
        $u \leftarrow (n_i \dots n_s)_2$                [$n_i = n_s = 1$ and $i - s + 1 \leqslant k$]
        $Q \leftarrow Q \oplus [u]P$              [$u$ is odd so that $[u]P$ is precomputed]
```

```
            $i \leftarrow s - 1$
    return $Q$
```

Remark 13.7 Since subtractions can be obtained in a straightforward way, signed-digit representations of n are well suited to compute $[n]P$, cf. Section 9.1.4.

Example 13.8 With the settings of Example 13.4, let us compute $[763]P_1$ with Algorithm 13.6 and a window of size 3. We precompute $[3]P_1 = (1081, 1674)$, $[5]P_1 = (851, 77)$, $[7]P_1 = (663, 1787)$ and since $763 = (\underset{5}{\underline{101}}\,\underset{7}{\underline{111}}\,\underset{5}{\underline{101}}\,\underset{1}{\underline{1}})_2$, the intermediate values of Q are

$$\begin{aligned}
[5]P_1 &= (851, 77), & [10]P_1 &= (4, 640), & [20]P_1 &= (836, 807),\\
[40]P_1 &= (1378, 1696), & [47]P_1 &= (1534, 747), & [94]P_1 &= (1998, 1094),\\
[188]P_1 &= (1602, 1812), & [376]P_1 &= (478, 1356), & [381]P_1 &= (1454, 981),\\
[762]P_1 &= (1970, 823), & [763]P_1 &= (1453, 1428).
\end{aligned}$$

Using the NAF expansion of $763 = (\underset{3}{\underline{10\bar{1}}}\,00000\,\underset{-5}{\underline{\bar{1}0\bar{1}}})_s$ instead, one obtains

$$\begin{aligned}
[3]P_1 &= (1081, 1674), & [6]P_1 &= (255, 1499), & [12]P_1 &= (459, 1270),\\
[24]P_1 &= (41, 1867), & [48]P_1 &= (1461, 904), & [96]P_1 &= (1966, 1808),\\
[192]P_1 &= (892, 529), & [384]P_1 &= (1928, 1803), & [768]P_1 &= (799, 1182),\\
[763]P_1 &= (1453, 1428).
\end{aligned}$$

The last step, namely $[763]P_1 = [768]P_1 \oplus [-5]P_1$, needs $[-5]P_1 = (851, 216)$ which is obtained directly from $[5]P_1$.

13.1.3 Rational points

When we consider a point P on an elliptic curve E/K, it is implicit that P has its coordinates in $\overline{K}$. To stress that P has its coordinates in K, we introduce a new concept.

Definition 13.9 Let E be an elliptic curve defined over K. The points lying on E with coordinates in K form the set of *K-rational points of E* denoted by $E(K)$. We have

$$E(K) = \{(x_1, y_1) \in K^2 \mid y_1^2 + a_1x_1y_1 + a_3y_1 = x_1^3 + a_2x_1^2 + a_4x_1 + a_6\} \cup \{P_\infty\}.$$

The structure of the group of $\mathbb{F}_q$-rational points is easy to describe. Indeed, by Corollary 5.77, $E(\mathbb{F}_q)$ is either cyclic or isomorphic to a product of two cyclic groups, namely $E(\mathbb{F}_q) \simeq \mathbb{Z}/d_1\mathbb{Z} \times \mathbb{Z}/d_2\mathbb{Z}$ where $d_1 \mid d_2$ and $d_1 \mid q - 1$.

For cryptographic applications one usually works in a subgroup of prime order ℓ. Hence, one is interested in curves and finite fields such that $|E(\mathbb{F}_q)| = c\ell$ for some small *cofactor* c. See [GAMC 2000] for conjectural probabilities that the number of points is a prime or has a small cofactor.

Finding a random $\mathbb{F}_q$-rational point P on an elliptic curve $E/\mathbb{F}_q$ is quite easy. See Sections 13.2 and 13.3 for examples. If the curve has a cofactor $c > 1$ then this random point needs not lie inside the group of order ℓ. However, the point $Q = [c]P$ either equals P_∞, in which case one has to try with a different random point P, or is a point in the prime order subgroup.

Example 13.10 Let us consider the curve E_1 as defined by (13.3). One can check that $|E_1(\mathbb{F}_p)| = 1956 = 12 \times 163$. So, there are 1955 affine points with coordinates in $\mathbb{F}_p$ and the point at infinity P_∞ lying on E_1. The point $P_1 = (1118, 269)$ is of order 1956 which implies that the group $E_1(\mathbb{F}_p)$ is cyclic generated by P_1. The point $Q_1 = (892, 529)$ is of prime order 163.

+,-.+ + +/ ! + %" ?! #1 C / < ? 3 /
! ! /7 # 5 6 3 ; 5 ,) 0#

+,-.+ + +/ ! + %" ?! #1 C = < ! !
! !0 ! ! # ?! P - 0#
/1 (0 ! GGG

+,-.+ + +/ ! + %" ?! #1 C 0 # - ! !
3 ! # 0 3 3 9 0#

+,-.+ + +/ ! + %" ?! #1 C ! ! ! 2 # 3
= ! , ! 3 ! # ! < # : C Q 0#
* ' =

Handbook of Mathematical Geodesy
Geosystems Mathematics, 1–163
© Springer International Publishing AG, part of Springer Nature 2018

Gauss as Scientific Mediator Between Mathematics and Geodesy from the Past to the Present

Willi Freeden, Thomas Sonar, and Bertold Witte

Abstract. The objective of the paper is to document the pioneer dimension of Gauss's ideas, concepts, and methods in a twofold direction based on selected case examples, namely to demonstrate his mediation function between mathematics and geodesy to explain the historic development over the past centuries from the initial Gaussian ignition to modern characteristics and tendencies.

Keywords. Gauss, geometric number theory, numerical integration, integral theorems and boundary value problems, least squares adjustment.

1. Introduction

1.1. C.F. Gauss (1777–1855): A brief look at his life

Carl Friedrich Gauss, born on 30 April 1777 in Brunswick (Braunschweig), came from a humble background whose parents were only able to enroll him in a basic writing and counting school. His mathematics teacher discovered his exceptional arithmetic skills and became an advocate for the talented student to facilitate the placement in a grammar school. At the age of fourteen the young Gauss was introduced to the Duke of Brunswick who pledged to finance the education of the extraordinarily talented boy. At the age of 15, Gauss had the genial idea to transfer the principles of logarithm tables to the prime number theory. It was the first time in the history of prime number research that prime number probability became object of research for increasing number ranges. However, he was not able to prove his prime number assumption that the number $\pi(x)$ smaller than x behaves asymptomatically as the quotient from x and $\log(x)$. Still this became the starting point for a variety of number theoretical examinations of renowned mathematicians until present. In the year 1795 Gauss started his studies at the University of Göttingen. After a brief period of time he decided to study mathematics. In

1796 Gauss discovered the solution to an old geometric problem, namely the answer to the question "which straight lines, regular polygons can be exclusively constructed by compass and straightedge". The evidence for the constructability of the regular hexadecagon led Gauss to continue groundbreaking research work, the results of which he compiled to a significant number theoretical work already during his study period which, alas, was only published in 1801 because it had to be translated into Latin as it was customary at that time. This opus published under the title "Disquisitiones arithmeticae" (cf. Figure 2.1) assured Gauss the recognition, in particular, amongst the leading French mathematicians of the era (Cauchy, Laplace, d'Alembert, Laguerre). He was attributed to the best mathematicians of his time due to his number theory. The "Disquisitiones arithmeticae" contain many significant research results such as, e.g., the celebrated Fundamental Theorem of Gauss, or the Law of Quadratic Reciprocity of Legendre. The crowning result of his contributions was the complete solution of binomial equations, and a most unexpected achievement in placing the imaginary unit on a firm basis. He actually was the first to use the imaginary symbol "i", giving it the interpretation of a geometric mean (see [27] for more details). After having graduated from Göttingen – upon the Duke's request who continued to grant him financial support – he switched to the University of Helmstedt where he studied under Prof. Pfaff and finished his thesis in 1799 in which he provided the exact evidence of the "Fundamental Theorem of Algebra" (Latin title: "Demonstratio nova theorematis, omnem functionem algebraicam rationalem integram unius variabilis in factores reales primi vel secundi gradus resolvi posse"). In the following time Gauss almost exclusively dealt with strictly mathematic or geometric questions. Applications were secondary during this epoch although he had already been working on the least squares method since 1794 with a multitude of practical studies. The deeper he penetrated into mathematics, the more fully he was persuaded that its true meaning lies in its application to practical life and natural science (cf. [27]). On January 1st, 1801 the astronomer Piazzi from Palermo discovered a small planet named Ceres. He could only track it for a time period of 41 days. Due to the low number of measuring values, the known methods at that time did not allow to calculate the precise orbit of Ceres in order to locate this planetoid again. The interest in the new planet caused Gauss to temporarily lay aside his purely mathematical researches. Now he created applicable methods for orbit determinations. He sought the orbit which would fit the observations as good as possible using an ellipse by applying his method of least squares. In doing so, his way resulted in the solution of two different problems: first, to find an approximate orbit; second, to correct this orbit in such a way that it "satisfies" the observations as well as possible (for more details the reader is referred to [27]). At the turn of the year 1801/1802 von Zach in Gotha and Olbers in Bremen rediscovered the planetoid, its location agreeing exactly with the ephemeris computed by Gauss. The discovery of Ceres introduced him to the world as an astronomer of the highest order (cf. [27]).

Gauss received a lot of praise and recognition for his scientific achievements; in January 1802 the Petersburg Academy of Sciences nominated him Correspond-

FIGURE 1.1. Medal (headside (left), backside (right)) of the 150th anniversary of the death of C.F. Gauss, 2005 (from [122]).

ing Member and in September 1802 he was invited to teach at the Academy which he declined as other offers later. In 1807 he accepted the call to Göttingen to become a professor for astronomy and the director of the observatory. In the years to follow Gauss mainly worked on astronomic topics but still on the theory of numbers. In the meantime he refocused his research works on application-related topics of mathematics which he found in astronomy as well as in geodesy. From 1818 until approximately 1830 geodesy became the center of his activities. Gauss made many geodetic observations during the arc measurements and the land surveying (1821–1844) in the Kingdom of Hanover in which he personally participated. As a matter of fact, Gauss alone managed the comprehensive calculations for this surveying project. Geodesy derived large direct and indirect profits from this practical activity which, for instance, was lamented by Bessel because of the entailing large time commitment and physical strain. During these works of arc measurements and geodetic surveying Gauss also proved to be a gifted practitioner: Particular mention should be made to the heliotrope (cf. Figure 6.5) invented by him which significantly facilitated and accelerated surveying works. The fundamental studies in classical differential geometry, in particular on the theory of surfaces, and his contributions to potential theory and the further development of the least squares method can be attributed to his practical activity as geodesist and astronomer. The fundamental surface studies, the so-called "Disquisitiones generales circa superficies curvas" (Gauss's Works, Vol. IV), due to their content, cannot be allocated to the actual geodetic works (see [141]) but more to the mathematical fundus. However, they are closely related to geodesy. The significance for geodesy constitutes in the detailed elaboration on the theory of geodetic lines for which preparatory studies had been done by Bernoulli, Euler, and Clairaut, but not with the strong comprehensiveness and thoroughness Gauss applied. The "Disquisitiones" laid the grounds for suitable arithmetic formulas and coordinate systems on the geodetically relevant rotation ellipsoid. Gauss published practical applications for geodetic questions in his "Studies on Topics of Higher Geodesy" in

1843 and 1846 (see Gauss's Works, Vol. IV). Their importance for geodesy can be concisely expressed by the following words: Gauss is the founder of higher geodesy.

Since the collaboration with the Göttingen physicist Weber in the year 1831, Gauss published contributions to classical mechanics, Earth's magnetism, geometric optics, and electrodynamics. In mechanics, for example, he developed the principle of least constraint which was named after him and which has the advantage in comparison to other approaches that it can be demonstratively construed. The theory of optical systems of Gauss for the first time shows the exact principles of the passage of one light beam through a coaxial lens system when it is slightly tilted to the axis. The Gauss ocular is still used for autocollimation nowadays, e.g., for specific tasks in engineering surveying. Gauss's magnetic studies are recognized as his most significant contributions to physics. Gauss and Weber jointly developed the first electromagnetic telegraph.

The three areas of geomagnetism in which Gauss made great contributions were those related to the absolute measurement of the field, the analysis in terms of spherical harmonics, and the organization and equipping of magnetic observatories. Because of the insufficiency of observations over the globe at the time he worked, many of the investigations which he proposed on the basis of the spherical harmonic analysis had in spite of their tremendous significance to await later workers. These include the quantitative separation of internal and external sources, the effect of the Earth's ellipsoidal shape, the possible non-vanishing of the constant term in the expression for the potential and the possible existence of a non-potential portion of the field (see, e.g., [70, 179] and the references therein).
Following W.K. Bühler [17], Gauss did not venture deeply into electrodynamics. In this field he is an outsider fascinated by interesting ideas and phenomena. His legacy contains several interesting recordings, however, no reasonably rounded theory. In fragments regarding the nature of the electromagnetic field, Gauss attempts to describe the theory of long-distance effects – a theory which was then further developed by Weber and Neumann and finally superseded by Maxwell's theory of electromagnetism.

Measured by his rich mathematic knowledge, Gauss published a relatively small number of papers. After his death on 23 February 1855, large amount of unpublished mathematic ideas was detected in his legacy, among others, recordings on non-Euclidean geometry. Gauss supposedly asked himself the question in view of the different geometries, which geometry correctly depicts the physical reality. The experience taught that if the dimensions are sufficiently small, the Euclidean geometry applied regarding the measuring precision, based on which Gauss assumed that the Euclidean geometry also applies to the "infinite small", but that deviations occur with larger dimensions. Gauss wished to know whether our space is curved, i.e., non-Euclidean. For this purpose he measured the inner angles in a large triangle (cf. Figure 6.1). More concretely, Gauss took measurements from three mountains in Germany, Hohenhagen, near Göttingen, Brocken in the Harz Mountains and Inselsberg in the Thüringer Wald to the south. The three lines joining these locations form a great triangle, the angle at Hohenhagen is close to a right angle, so the area of the triangle is close to half the product of the two short sides. Gauss assumed that light propagates along geodesics. During his high-precision measurement campaign he was not able to detect deviations from the Euclidean internal angle amount, implying that the Euclidean geometry can be applied for relatively large distances in the physical space. This result was negative for Gauss which may have been the reason for his decision not to publish his studies on non-Euclidean geometry. It is doubtful whether this interpretation is correct; Gauss presumably knew that potential deviations would only become obvious in triangles of astronomic dimensions (see [240] for more details). At a later time, Riemann, who was inspired by Gauss, created the Riemann geometry, the basis on which Einstein was able to build up his relativity theory which also finds its application, e.g., in satellite geodesy (see also [224]).

In conclusion, it has to be stated that it was extremely fortunate for geodesy that the world-renowned mathematician (*"Princeps Mathematicorum"*), Carl Friedrich Gauss (cf. Figure 1.2), had been so much fascinated by geodesy from early on in his mathematical career to which he dedicated a significant part of this work time over his life. In a letter to the astronomer Olbers in January 1802 he expressed it as follows: *"The most refined geometer and the perfect astronomer these are two separate titles which I highly esteem with all my heart, and which I worship with passionate warmth whenever they are united"* (see [27, 141] for more details).

1.2. Scientific bridge between mathematics and geodesy

Evidently, a contribution concerned with Gauss as scientific mediator between mathematics and geodesy (as intended by this publication) has to follow an obligatory line, namely to provide an appropriate platform within which mathematically/geodetically interrelated features can be better motivated and understood, thereby canonically leading to an interdisciplinary palette of solution procedures in diverse areas of application. In accordance with this intention, criteria must be found relative to which the scope and limitations of the various methods can be assessed. This is important both in theory and practice since there generally is no cure-all method for most of the problems in reality; it is imperative to be able to clarify why a certain method works in some context as well as when not to use that method.

The authors of this work are convinced that Gauss's suggestions and ideas as exemplary work is still helpful to understand the intuitive principles and interrelations that underlie the various methods and procedures to be needed for the solution of problems and desiderata. So, in summary, the present article may be regarded as an attempt to justify this basic assertion.

1.3. Mathematical circuit: abstraction and concretization

What is it exactly that enables mathematicians to build a bridge between geodesy and their discipline? What is exactly that enables the mathematicians to provide the transfer from concrete geodetic measurements and observables to abstract

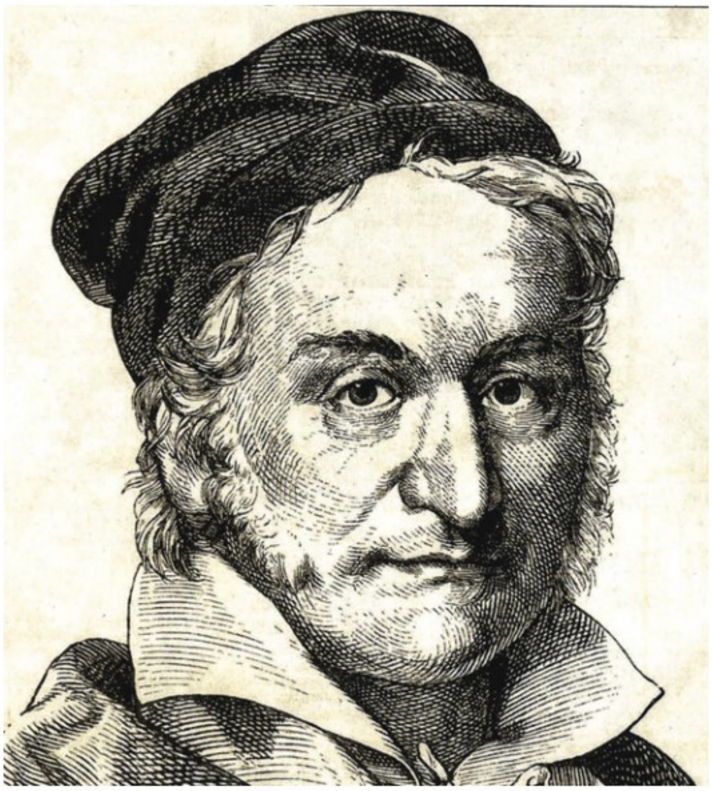

FIGURE 1.2. Johann Carl Friedrich Gauss, born, 30 April 1777 Brunswick, Duchy of Brunswick-Wolfenbüttel, Holy Roman Empire, died, 23 February 1855 (aged 77) Göttingen, Kingdom of Hanover.

mathematical formalisms and models? What is exactly that enables the mathematicians to bridge the time gap from historic nomenclature to modern notation? Some answers should be given already at this early stage: The mathematics world of numbers and structures contains efficient tokens by which the rule-like aspect of geodetic problems can be described appropriately. In fact, this description includes as essential step a simplification by abstraction. Essential properties of the problem are separated from unimportant ones, further specified, and afterwards included into a solution scheme. The "eye for similarities" often enables mathematicians to recognize a posteriori that an adequately reduced problem may also arise from very different situations in various application areas, so that the resulting solutions may be applicable to multiple cases after an adequate adaptation or concretization. Without this ingredient, the abstraction remains essentially useless. The interaction between abstraction and concretization characterizes the history of mathematics and its current development as a common language and independent standards in a unified setup. A problem reduced by abstraction has to be considered as a new "concrete" problem to be solved within a general framework, that determines the validity of a possible solution, not only in geodesy, but also in quite different sciences. The more examples and facets one knows, the more one recognizes the causality between the abstractness of mathematical concepts and their impact and cross-sectional significance.

As an immediate consequence, two important questions arise in a work making the attempt to pursue geodetic ideas and concepts from Gaussian time up to resulting mathematical settings in our days.

(i) What can be specified as mathematical structures and settings by a process of abstraction, i.e., a reduction to essential features?
(ii) What can be achieved from the reduced context obtained by abstraction as a new field of scientific interest, e.g., in geodesy or other sciences?

The purpose of this contribution is to discuss these questions for a selected collection of case studies. The interplay between abstraction and concretization will be explained for each of the problems, thereby offering new perspectives for future challenges. Dilemmas and methodologies will be indicated during the resolution process.

1.4. Specific strategies imposed on our work

The purpose of our work is not to describe the entire opus of the "science titan" Gauss in all its facets from the past until today. Such an approach would be doomed to fail in view of the enormous significance and the immense amount of ideas and works of Gauss. The objective of the authors is rather to document the pioneer dimension of Gauss's ideas, concepts, and methods in a twofold direction based on selected case examples, to demonstrate his mediation function between mathematics and geodesy firstly and secondly the historic development over the past centuries from the initial ignition by Gauss to the modern characteristics and tendencies in mathematics and/or geodesy. The authors consciously limited themselves to special topics in which they feel competent to focus on based on their scientific formation. It includes lattice point number theory with its basics for the development of modern sampling methods, Gauss integration with its impacts on modern fast algorithms for global modeling of geodetically and geophysically relevant quantities, the potential theory with the Gauss law as the initial point for geodetic boundary value problems to determine the physically defined Earth's figure and finally the least squares method in its canonic continuation into the issues of present ill-posed and inverse problems, e.g., in terms of pseudodifferential equations. Due to the reference of this contribution to mathematical geodesy, Gauss's contributions to geomagnetics were not reflected here (in this conjunction, reference is made to, e.g., publications of G.D. Garland [70], K. Reich, E. Roussanova [179] and their cross-references). In addition, a difficulty arose with translating Gauss terms into modern mindsets. The mathematic/geodetic language and forms of expression have developed and changed over the past two centuries. The authors, alas, were not able to convey and suitably illustrate the development process of the notation. For reasons of legibility, a standardized unified notation was selected (except quotes) adjusted to the requirements of modern times as the last link of the term-defining process.

2. From Gaussian circle problem to geosampling

There is no doubt that the theory of numbers was Gauss's favourite subject. In a much quoted dictum, he asserted that

> *"Mathematics is the Queen of the Sciences and the Theory of Numbers is the Queen of Mathematics".*

Moreover, in the introduction to "Eisenstein's Mathematische Abhandlungen" Gauss wrote:

> *"The Higher Arithmetic presents us with an inexhaustible storehouse of interesting truths – of truths, too, which are not isolated but stand in the closest relation to one another, and between which, with each successive advance of the science, we continually discover new and sometimes wholly unexpected points of contact. A great part of the theories of Arithmetic derive an additional charm from the peculiarity that we easily arrive by induction at important propositions which have the stamp of simplicity upon them but the demonstration of which lies so deep as not to be discovered until after many fruitless efforts; and even then it is obtained by some tedious and artificial process while the simpler methods of proof long remain hidden from us."*

All this is well illustrated by what is perhaps Gauss's most profound publication, namely his "Disquisitiones arithmeticae"(cf. Figure 2.1). It has been described, quite justifiably, as the "Magna Carta of Number Theory", and the depth and originality of thoughts to be manifested in this work are particularly remarkable considering that they were written when Gauss was only about eighteen years of age. In view of the great impact Gauss had on large areas of modern number theory, anything even approaching a comprehensive representation of their influence seems untenable. It is not surprising that there is a huge amount of literature concerned with Gauss's number theoretical results, and his influence on modern mathematics is enormous. However, the obvious problem for our purposes here is the question if the "Queen of Mathematics" is actually able to show impacts to modern mathematical geodesy.

2.1. Lattice points inside circles

We start our bridge from Gaussian concepts of number theory to modern geodetically relevant sampling with a recapitulation of some results on the number of lattice points inside circles

$$\mathbb{S}^1_N = \{x \in \mathbb{R}^2 : |x| = N\} \tag{2.1}$$

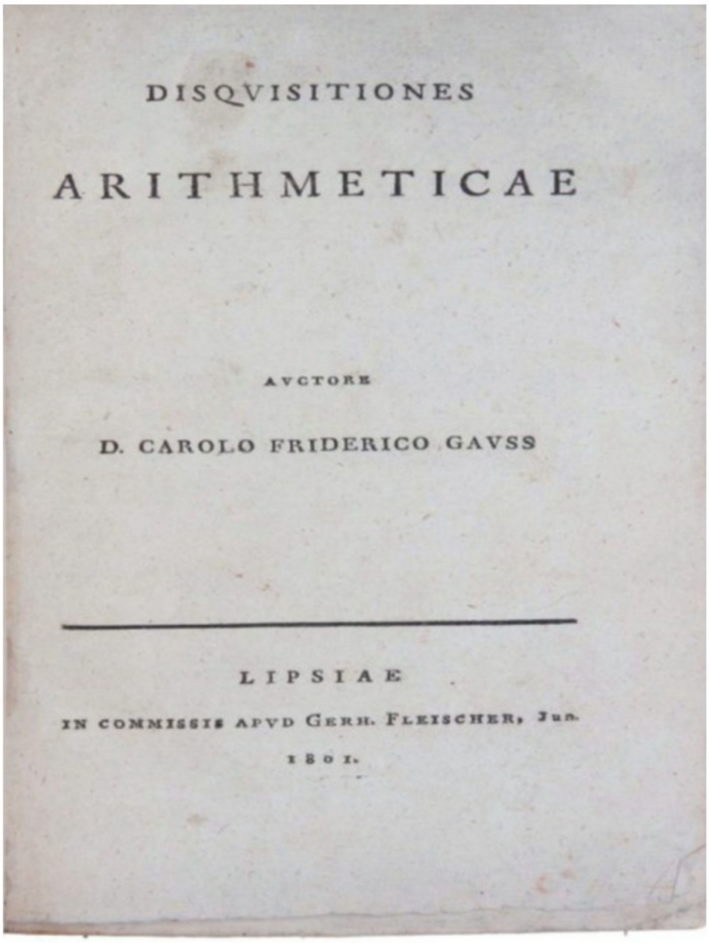

DISQVISITIONES

ARITHMETICAE

AVCTORE

D. CAROLO FRIDERICO GAVSS

LIPSIAE

IN COMMISSIS APVD GERH. FLEISCHER, Iun.

1801.

FIGURE 2.1. Title-page of **"Disquisitiones arithmeticae"**, original latin edition by Gerhard Fleischer, Lipsiae (Leipzig) 1801 (668 pages), first reprint as first issue of "Gesamtausgabe Carl Friedrich Gauss: Werke. Band 1, Dieterich, Göttingen 1863", reprint by Springer-Verlag, New York Heidelberg 1986, ISBN 0-387 96254-9 (English translation by Arthur A. Clarke, 1986, in revised form by William C. Waterhouse).

of radii $N > \frac{\sqrt{2}}{2}$ around the origin 0; more accurately, we deal with closed disks

$$\overline{\mathbb{B}_N^2} = \{x \in \mathbb{R}^2 : |x| \leq N\} \tag{2.2}$$

of radii $N > \frac{\sqrt{2}}{2}$ (for more background material and deeper number theoretical concepts the reader is referred, e.g., to the monographs [47, 66]).

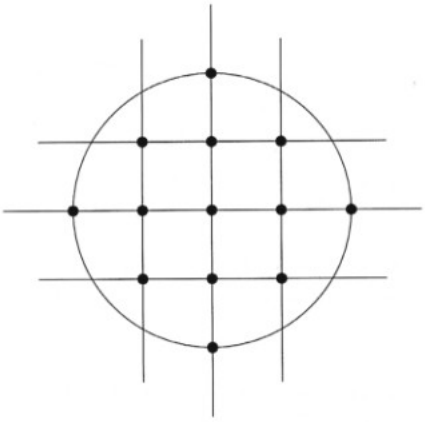

FIGURE 2.2. Lattice points inside a circle.

The problem of determining the total number of lattice points of $\mathbb{Z}^2$ inside and on a circle with radius N, i.e., the determination of the quantity

$$\sharp_{\mathbb{Z}^2}\left(\overline{\mathbb{B}^2_N}\right) = \sharp\left\{(n_1,n_2)^T \in \mathbb{Z}^2 : n_1^2+n_2^2 \leq N^2\right\} \tag{2.3}$$

reaches back to L. Euler [33]. In today's nomenclature it can be equivalently expressed as a sum in the form

$$\sharp_{\mathbb{Z}^2}\left(\overline{\mathbb{B}^2_N}\right) = \sum_{\substack{n_1^2+n_2^2\leq N^2;\\ (n_1,n_2)^T\in\mathbb{Z}^2}} 1. \tag{2.4}$$

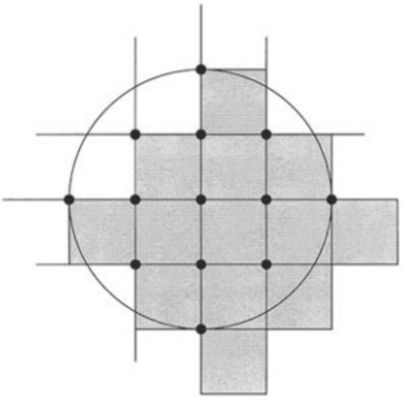

FIGURE 2.3. The polyhedral set $\mathbb{P}^2_N$.

Gauss [1801] found a simple, but efficient method for its estimation (cf. Figures 2.2 and 2.3): associate to every square the Northwest edge as lattice point. The union of all squares with lattice points inside $\overline{\mathbb{B}_N^2}$ defines a polyhedral set $\mathbb{P}_N^2$ with area

$$\|\mathbb{P}_N^2\| = \sharp_{\mathbb{Z}^2}(\overline{\mathbb{B}_N^2}) \tag{2.5}$$

(cf. Figure 2.3). Since the diagonal of each square is $\sqrt{2}$, the geometry of Figure 2.3 tells us that

$$\pi\left(N - \frac{\sqrt{2}}{2}\right)^2 \leq \sharp_{\mathbb{Z}^2}\left(\overline{\mathbb{B}_N^2}\right) \leq \pi\left(N + \frac{\sqrt{2}}{2}\right)^2. \tag{2.6}$$

Therefore, $\sharp_{\mathbb{Z}^2}\left(\overline{\mathbb{B}_N^2}\right) - \pi N^2$ after division by N is bounded for $N \to \infty$, which is usually written with Landau's O-symbol as

$$\sharp_{\mathbb{Z}^2}\left(\overline{\mathbb{B}_N^2}\right) = \pi N^2 + O(N). \tag{2.7}$$

In other words, the number of lattice points in $\overline{\mathbb{B}_N^2}$ is equal to the area of that circle plus a remainder of the order of the boundary. In particular,

$$\sharp_{\mathbb{Z}^2}\left(\overline{\mathbb{B}_N^2}\right) \sim \pi N^2 \tag{2.8}$$

so that a method of determining the irrational, transcendent number π becomes obvious (for alternative approaches to π within the history of analysis the reader is referred to [204]):

$$\lim_{N\to\infty} \frac{\sharp_{\mathbb{Z}^2}\left(\overline{\mathbb{B}_N^2}\right)}{N^2} = \pi. \tag{2.9}$$

C.F. Gauss [71] illustrated his result by taking $N^2 = 100\,000$. In this case he calculated

$$\sum_{\substack{|g|^2 \leq 100\,000;\\ g \in \mathbb{Z}^2}} 1 = 314\ 197. \tag{2.10}$$

This calculation determines the number π up to three decimals after the comma.

2.2. Circle problem and Hardy's conjecture

The formula (2.7) due to C.F. Gauss [71] allows the following representation in the nomenclature of Landau's O-symbols:

$$\sharp_{\mathbb{Z}^2}\left(\overline{\mathbb{B}_N^2}\right) = \pi N^2 + O(N). \tag{2.11}$$

The so-called *circle problem* is concerned with the question of determining the bound

$$\alpha_2 = \inf\left\{\gamma : \sharp_{\mathbb{Z}^2}\left(\overline{\mathbb{B}_{\sqrt{N}}^2}\right) = \pi N + O(N^\gamma)\right\}. \tag{2.12}$$

Until now, we knew from (2.11) that $\alpha_2 \leq \frac{1}{2}$. An improvement of the Gaussian result, however, turns out to be very laborious, in fact, requiring a great effort.

A first remarkable result is due to W. Sierpinski [201], who proved by use of a method of his teacher G. Voronoi [220] that

$$\sharp_{\mathbb{Z}^2}\left(\overline{\mathbb{B}^2_{\sqrt{N}}}\right) \quad = \pi N + O\left(N^{\frac{1}{3}}\right), \tag{2.13}$$

i.e., $\alpha_2 \leq \frac{1}{3}$. The proof of Sierpinski is elementary (see, e.g., [47, 66] for more details); it is a link between geometry and number theory.

By use of advanced methods on exponential sums (based on the work by, e.g., H. Weyl [228], H.R. Chen [18], and many others) the estimate $\frac{1}{3}$ could be strengthened to some extent. It culminated in the publication by G. Kolesnik [128], who had as his sharpest result with these techniques

$$\sharp_{\mathbb{Z}^2}\left(\overline{\mathbb{B}^2_{\sqrt{N}}}\right) \quad - \pi N = O\left(N^{\frac{139}{429}}\right). \tag{2.14}$$

M.N. Huxley [117] devised a substantially new approach (not discussed here); his strongest result was the estimate

$$\sharp_{\mathbb{Z}^2}\left(\overline{\mathbb{B}^2_{\sqrt{N}}}\right) \quad - \pi N = O\left(N^{\frac{131}{416}}\right). \tag{2.15}$$

(note that $\frac{139}{429} = 0.324009\ldots$, while $\frac{131}{416} = 0.315068\ldots$).

Hardy's conjecture claims

$$\sharp_{\mathbb{Z}^2}\left(\overline{\mathbb{B}^2_{\sqrt{N}}}\right) - \pi N = O\left(N^{\frac{1}{4}+\varepsilon}\right) \tag{2.16}$$

for every $\varepsilon > 0$. This conjecture seems to be still a challenge for future work. However, in the year 2007, S. Cappell and J. Shaneson deposited a paper entitled "Some Problems in Number Theory I: The Circle Problem" in the arXiv:math/0702613 claiming to prove the bound of $O(N^{\frac{1}{4}+\varepsilon})$ for $\varepsilon > 0$.

0.250000	Gauss (1801)
0.083333 . . .	G. Voronoi (1903), Sierpinski (1906)
0.080357 . . .	J.E. Littlewood, A. Walfisz (1924)
0.079268 . . .	J.G. van der Corput (1928)
0.074324 . . .	J.-R. Chen (1963)
0.074009 . . .	G. Kolesnik (1985)
0.064903 . . .	M.N. Huxley (2003)

TABLE 1. Incremental improvements for the value ε_2 in the estimate (2.17).

Table 1 lists incomplete incremental improvements for the quantity ε_2 of the upper limit for the circle problem

$$\sharp_{\mathbb{Z}^2}\left(\overline{\mathbb{B}^2_{\sqrt{N}}}\right) - \pi N = O\left(N^{\frac{1}{4}+\varepsilon_2}\right) \; . \tag{2.17}$$

For all recent improvements, the proofs became rather long and made use of some of the more heavy machinery in hard analysis.

Summarizing our results about lattice points inside circles (cf. [47]) we are confronted with the following situation:

$$\frac{1}{4} \;\leq\; \alpha_2 \;\leq\; \frac{1}{4} + \varepsilon_2 \tag{2.18}$$

and

$$\sharp_{\mathbb{Z}^2}\left(\overline{\mathbb{B}^2_{\sqrt{N}}}\right) - \pi N \neq O\left(N^{\frac{1}{4}}\right), \tag{2.19}$$

$$\sharp_{\mathbb{Z}^2}\left(\overline{\mathbb{B}^2_{\sqrt{N}}}\right) - \pi N = O\left(N^{\frac{1}{4}+\varepsilon_2}\right), \tag{2.20}$$

where $0 < \varepsilon_2 \leq \frac{1}{4}$ (for example, Huxley's bound $\varepsilon_2 = 0.064903\ldots$).

2.3. Variants of the circle problem

There are many perspectives to formulate variants of the Gaussian lattice point problem for the circle. It already was the merit of E. Landau [135] to point out particularly interesting areas, such as

- *General two-dimensional lattices*

$$\Lambda = \{g = ng_1 + mg_2 \;:\; n, m \in \mathbb{Z}\} \tag{2.21}$$

with $g_1, g_2 \in \mathbb{R}^2$ linearly independent (see Figure 2.4) can be used instead of the unit lattice $\mathbb{Z}^2$.

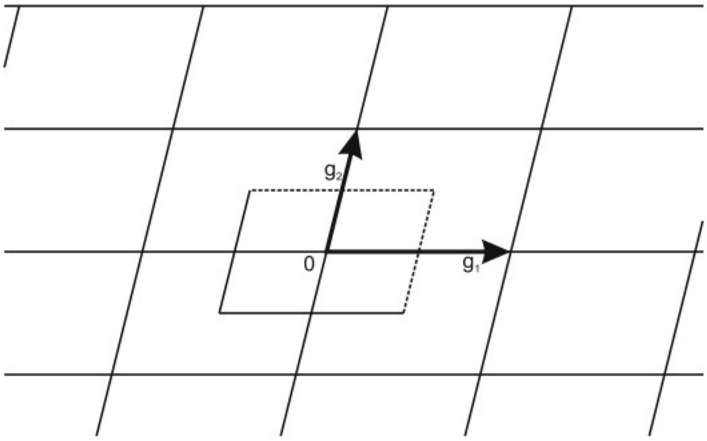

FIGURE 2.4. Two-dimensional lattice Λ generated by $g_1, g_2 \in \mathbb{R}^2$.

- The *remainder term* can be represented as alternating series, called *Hardy–Landau series* in terms of the Bessel function J_1 of order 1 (for the different facets of the proof see [103, 134, 160])

$$\sum_{\substack{|g|\leq N\\ g\in\Lambda}}{}' \, 1 = \frac{\pi N^2}{\|\mathcal{F}_\Lambda\|} + \lim_{R\to\infty} \frac{\pi N^2}{\|\mathcal{F}_\Lambda\|} \sum_{\substack{0<|h|\leq R\\ h\in\Lambda^{-1}}} \frac{J_1(2\pi|h|N)}{\pi|h|N}, \tag{2.22}$$

where Λ is an arbitrary lattice in $\mathbb{R}^2$ and

$$\mathcal{F}_\Lambda = \left\{ x = x_1 g_1 + x_2 g_2 \in \mathbb{R}^2 : -\frac{1}{2} \leq x_i < \frac{1}{2},\ i = 1,2 \right\} \tag{2.23}$$

is the *fundamental cell of* $\Lambda \subset \mathbb{R}^2$ with

$$\|\mathcal{F}_\Lambda\| = \sqrt{\det\left((g_i \cdot g_j)_{i,j=1,2}\right)} \tag{2.24}$$

as the area $\|\mathcal{F}_\Lambda\|$ *of* $\mathcal{F}_\Lambda$. Moreover, the following convention

$$\sum_{\substack{|g| \leq N \\ g \in \Lambda}}{}' 1 = \sum_{\substack{|g| < N \\ g \in \Lambda}} 1 + \frac{1}{2} \sum_{\substack{|g| = N \\ g \in \Lambda}} 1 \tag{2.25}$$

is used in lattice point theory (note that the last sum only occurs if there is a lattice point $g \in \Lambda$ with $|g| = N$).

- Lattice points can be affected by non-constant weights (see [35])

$$\begin{aligned} &\sum_{\substack{|a+g| \leq N \\ g \in \Lambda}}{}' e^{2\pi i y \cdot (a+g)} F(a+g) \\ &\quad = \lim_{R \to \infty} \frac{1}{\|\mathcal{F}_\Lambda\|} \sum_{\substack{|h-y| \leq R \\ h \in \Lambda^{-1}}} e^{2\pi i a \cdot h} \int_{\substack{|x| \leq N \\ x \in \mathbb{R}^2}} F(x)\, e^{-2\pi i x \cdot (h-y)}\, dx, \end{aligned} \tag{2.26}$$

where dx is the volume element, $a, y \in \mathbb{R}^2$, F is twice continuously differentiable in $\overline{\mathbb{B}^2_N}$, $N > 0$, and the following convention has been used analogously to (2.25)

$$\sum_{\substack{|a+g| \leq N \\ g \in \Lambda}}{}' \cdots = \sum_{\substack{|a+g| < N \\ g \in \Lambda}} \cdots + \frac{1}{2} \sum_{\substack{|a+g| = N \\ g \in \Lambda}} \cdots . \tag{2.27}$$

Note that, for $F = 1$, this formula leads back to

$$e^{2\pi i a \cdot y} \sum_{\substack{|g+a| \leq N \\ g \in \Lambda}}{}' e^{2\pi i g \cdot y} = \lim_{R \to \infty} \frac{\pi N^2}{\|\mathcal{F}_\Lambda\|} \sum_{\substack{|h-y| \leq R \\ h \in \Lambda^{-1}}} e^{2\pi i a \cdot h} \frac{J_1(2\pi |h-y| N)}{\pi |h-y| N}. \tag{2.28}$$

For $a = y = 0$ we obtain the *classical Hardy–Landau identity*, i.e., the identity

$$\sum_{\substack{|g| \leq N \\ g \in \Lambda}}{}' 1 = \lim_{R \to \infty} \frac{\pi N^2}{\|\mathcal{F}_\Lambda\|} \sum_{\substack{|h| \leq R \\ h \in \Lambda^{-1}}} \frac{J_1(2\pi |h| N)}{\pi |h| N}. \tag{2.29}$$

holds true. Observe that J_1 satisfies the asymptotic relation $J_1(r) = \frac{r}{2} + \cdots$, so that

$$\sum_{\substack{|g| \leq N \\ g \in \Lambda}}{}' 1 = \frac{\pi N^2}{\|\mathcal{F}_\Lambda\|} + \lim_{R \to \infty} \frac{\pi N^2}{\|\mathcal{F}_\Lambda\|} \sum_{\substack{0 < |h| \leq R \\ h \in \Lambda^{-1}}} \frac{J_1(2\pi |h| N)}{\pi |h| N}. \tag{2.30}$$

- Generalizations to lattices $\Lambda \subset \mathbb{R}^q$ and regular regions $\mathcal{G} \subset \mathbb{R}^q, q \geq 2$, and continuous functions on $\overline{\mathcal{G}} = \mathcal{G} \cup \partial\mathcal{G}$ can be formulated in Gaussian summability (see [47] for the proof and a more detailed study)

  $$\sum_{\substack{a+g\in\overline{\mathcal{G}}\\ g\in\Lambda}}{}' e^{2\pi i y\cdot(a+g)} F(a+g) = \lim_{\substack{\tau\to 0\\ \tau>0}} \frac{1}{\|\mathcal{F}_\Lambda\|} \sum_{h\in\Lambda^{-1}} e^{-\tau\pi^2 h^2} e^{2\pi i h\cdot a} \int_{\mathcal{G}} F(x) e^{-2\pi i x\cdot(h-y)}\, dx, \ \ a, y \in \mathbb{R}^q, \quad (2.31)$$

  where a *regular region* $\mathcal{G}$ *in* $\mathbb{R}^q$ is understood to be an open and connected set $\mathcal{G} \subset \mathbb{R}^q$, $q \geq 2$, for which
  (i) its boundary $\partial\mathcal{G}$ constitutes an orientable, piecewise smooth Lipschitzian manifold of dimension $q-1$,
  (ii) the origin is contained in $\mathcal{G}$,
  (iii) $\mathcal{G}$ divides $\mathbb{R}^q$ into the "inner space" $\mathcal{G}$ and the "outer space" $\mathbb{R}^q \backslash \overline{\mathcal{G}}$, $\overline{\mathcal{G}} = \mathcal{G} \cup \partial\mathcal{G}$.

  Clearly,

  $$\sum_{\substack{a+g\in\overline{\mathcal{G}}\\ g\in\Lambda}}{}' e^{2\pi i y\cdot(a+g)} F(a+g) = \frac{1}{\|\mathcal{F}_\Lambda\|} \int_{\mathcal{G}} F(x) e^{2\pi i x\cdot y}\, dx + \lim_{\substack{\tau\to 0\\ \tau>0}} \frac{1}{\|\mathcal{F}_\Lambda\|} \sum_{\substack{0<|h|\leq R\\ h\in\Lambda^{-1}}} e^{-\tau\pi^2 h^2} e^{2\pi i h\cdot a} \int_{\mathcal{G}} F(x) e^{-2\pi i x\cdot(h-y)}\, dx, \ \ a, y \in \mathbb{R}^q. \quad (2.32)$$

  The following abbreviation has been used consistently

  $$\sum_{\substack{a+g\in\overline{\mathcal{G}}\\ g\in\Lambda}}{}' \cdots = \sum_{\substack{a+g\in\mathcal{G}\\ g\in\Lambda}} \cdots + \sum_{\substack{a+g\in\partial\mathcal{G}\\ g\in\Lambda}} \alpha(a+g) \cdots \quad (2.33)$$

  with $\alpha(a+g)$ denoting the *solid angle* subtended by $\partial\mathcal{G}$ at $a+g$ (note that, as geoscientifically relevant regular regions, we may choose the interior of the (actual) Earth's body or parts of it, the interior of geoscientifically relevant surfaces such as the geoid, telluroid, etc., but also ball, ellipsoid, cube, polyhedral bodies, etc. are included in accordance with the above definition; in potential-theoretic jargon, the solid angle will be explained by Definition 4.1 in a more detailed way).

2.4. Multivariate Shannon sampling

Let us continue with the observation that, for every $y \in \mathbb{R}^q$, the (for dimensions $q \geq 3$ formally understood) series

$$a \mapsto \lim_{N\to\infty} \frac{1}{\|\mathcal{F}_\Lambda\|} \sum_{\substack{|h-y|\le N \\ h\in\Lambda^{-1}}} e^{2\pi i a\cdot h} \underbrace{\int_{\mathcal{G}} F(x) e^{-2\pi i x\cdot(h-y)}\,dx}_{F_{\mathcal{G}}^{\wedge}(h-y)},\ a\in\mathbb{R}^q, \tag{2.34}$$

as well as the finite sum

$$a \mapsto \sum_{\substack{a+g\in\overline{\mathcal{G}} \\ g\in\Lambda}}{}' e^{2\pi i y\cdot(a+g)} F(a+g),\ a\in\mathbb{R}^q, \tag{2.35}$$

show Λ-periodicity, i.e., as functions of the variable $a\in\mathbb{R}^q$ they are periodic with respect to the lattice $\Lambda\subset\mathbb{R}^q$. As a consequence (see [57] for the details), Shannon-type sampling procedures can be obtained by formal integration of the lattice point identity (2.31) over a regular region $\mathcal{H}$ that is not necessarily equal to $\mathcal{G}$

$$\begin{aligned}
&\int_{\mathcal{H}} \sum_{\substack{a+g\in\overline{\mathcal{G}} \\ g\in\Lambda}}{}' e^{-2\pi i y\cdot(a+g)} F(a+g)\,da \\
&\quad = \sum_{\substack{(\mathcal{F}_\Lambda+\{g'\})\cap\overline{\mathcal{H}}\neq\emptyset \\ g'\in\Lambda}} \underbrace{\int_{\mathcal{G}\cap\bigcup_{g\in\Lambda}(((\overline{\mathcal{H}}\cap(\mathcal{F}_\Lambda+\{g'\}))-\{g'\})+\{g\})} F(x)e^{-2\pi i y\cdot x}\,dx}_{=F^{\wedge}_{\mathcal{G}\cap\bigcup_{g\in\Lambda}(((\overline{\mathcal{H}}\cap(\mathcal{F}_\Lambda+\{g'\}))-\{g'\})+\{g\})}(y)} \\
&\quad = \lim_{\substack{\tau\to0 \\ \tau>0}} \frac{1}{\|\mathcal{F}_\Lambda\|} \sum_{h\in\Lambda^{-1}} e^{-\tau\pi^2h^2} \underbrace{\int_{\mathcal{G}} F(x)e^{-2\pi i h\cdot x}\,dx}_{=F_{\mathcal{G}}^{\wedge}(h)}\ \underbrace{\int_{\mathcal{H}} e^{2\pi i a\cdot(h-y)}\,da}_{=K_{\mathcal{H}}(h-y)}.
\end{aligned} \tag{2.36}$$

The identity (2.36) has many interesting properties. For example, by virtue of the Gaussian summability, the convergence of the cardinal-type series on the right-hand side of (2.36) may be exponentially accelerated. All manifestations of over- and undersampling can be explicitly analyzed by the finite sum of Fourier transforms on the left side of the identity (see [57]), dependent on the geometric configurations of the chosen regular regions $\mathcal{G},\mathcal{H}$ (note that the identity (2.36) also seems to be unknown for the uni-variate case in this generality).

The Gaussian summability of the cardinal series on the right-hand side of (2.36) is of great importance from numerical point of view; it enables a fast computation of the series. Nonetheless, W. Freeden, M.Z. Nashed [57]) show that the identity (2.36) additionally holds true in ordinary sense, i.e., we have

$$\begin{aligned}
&\int_{\mathcal{H}} \sum_{\substack{a+g\in\overline{\mathcal{G}} \\ g\in\Lambda}}{}' e^{-2\pi i y\cdot(a+g)} F(a+g)\,da \\
&\quad = \sum_{\substack{(\mathcal{F}_\Lambda+\{g'\})\cap\overline{\mathcal{H}}\neq\emptyset \\ g'\in\Lambda}} \underbrace{\int_{\mathcal{G}\cap\bigcup_{g\in\Lambda}(((\overline{\mathcal{H}}\cap(\mathcal{F}_\Lambda+\{g'\}))-\{g'\})+\{g\})} F(x)e^{-2\pi i y\cdot x}\,dx}_{=F^{\wedge}_{\mathcal{G}\cap\bigcup_{g\in\Lambda}(((\overline{\mathcal{H}}\cap(\mathcal{F}_\Lambda+\{g'\}))-\{g'\})+\{g\})}(y)}
\end{aligned}$$

$$= \frac{1}{\|\mathcal{F}_\Lambda\|} \sum_{h \in \Lambda^{-1}} \underbrace{\int_{\mathcal{G}} F(x) e^{-2\pi i h \cdot x} \, dx}_{= F^\wedge_{\mathcal{G}}(h)} \; \underbrace{\int_{\mathcal{H}} e^{2\pi i a \cdot (h-y)} \, da}_{= K_{\mathcal{H}}(h-y)}. \tag{2.37}$$

A simple, but significant case of (2.37) in Euclidean space $\mathbb{R}^q$ showing no phenomenon of aliasing is obtained under the special choice $\overline{\mathcal{G}} = \overline{\mathcal{H}} \subset \overline{\mathcal{F}_\Lambda}$ leading to the Shannon-type identity

$$F^\wedge_{\mathcal{G}}(y) = \frac{1}{\|\mathcal{F}_\Lambda\|} \sum_{h \in \Lambda^{-1}} F^\wedge_{\mathcal{G}}(h) \;\; K_{\mathcal{G}}(h-y). \tag{2.38}$$

In fact, the identity (2.38) is a multi-variate variant of the Shannon sampling theorem (cf. [198]), but now for (geoscientifically relevant) regions $\mathcal{G}$. The principal impact of Shannon sampling on information theory is that it allows the replacement of a bandlimited signal $F^\wedge_{\mathcal{G}}$ related to $\mathcal{G}$ by a discrete sequence of its samples without loss of any information. Also it specifies the lowest rate, i.e., the Nyquist rate (cf. [57]), that it enables to reproduce the original signal. In other words, Shannon sampling provides the bridge between continuous and discrete versions of a bandlimited function.

The Shannon sampling theorem has many applications in engineering and physics, for example, in signal processing, data transmission, cryptography, constructive approximation such as spectral analysis by Slepian functions (see, e.g., [202, 203]), partial differential equations such as Boltzmann equation (see, e.g., [171] and the references therein), and inverse problems such as the multi-variate discussion of a Fourier transform over a (geodetically relevant) regular region (see, e.g., [164] and the references therein for the uni-variate study of the antenna problem).

Finally, it should be remarked that sampling theory (in the sense of the representation of an analog signal in terms of its samples) has attracted considerable interest in the past three decades. Major advances in the theory and applications of sampling expansion in different function spaces have been made by mathematicians and engineers. There are several journals on signal processing and computational harmonic analysis that publish papers on sampling theory and its applications. The field also sports its own journal "Sampling Theory in Signal and Image Processing" and has its own "professional society". We quote from Wikipedia:

> *"SampTA (Sampling Theory and Applications) is a biennial interdisciplinary conference for mathematicians, engineers, and applied scientists. The main purpose of SampTA is to exchange recent advances in sampling theory and to explore new trends and directions in related areas of applications. The SampTA conference series began as a small workshop in* 1995 *in Riga, Latvia, but the meetings grew into full-fledged conferences attracting an even mix of mathematicians and engineers as the interest in sampling theory blossomed."*

Figure 2.3 The polyhedral set $\mathbb{P}_N$.

Pág. 11(26) – 2.2

Circle problem and Hardy's conjecture.

Pág. 12(27) –

2007, S. Cappell and J. Shaneson →

"Some Problems in Number Theory I: The Circle Problem" in the

arXiv.com

math/0702613

claiming to prove the bound of

$O(N^{\frac{1}{4}+\epsilon})$ for $\epsilon > 0$.

→ Table 1.

Incremental improvements for the value θ_2 in the estimate (2.17).

Obs: Ver livro todo!!!

Pág. 28(43) Gaussian Quadrature

where

$$G_{0,\ldots,m}((\mathrm{L}+\lambda)^p;x,t) = \sum_{\substack{k=0\\ \lambda\neq \mathrm{L}^\wedge(k)}}^{m} \frac{1}{(\lambda-\mathrm{L}^\wedge(k))^p} P_k^*(x)P_k^*(t). \tag{3.49}$$

$G^{\perp}_{0,\ldots,m}((\mathrm{L}+\lambda)^p;\cdot,\cdot)$ is called *mth truncated Legendre (Green) function with respect to the operator* $(\mathrm{L}+\lambda)^p$. These functions admit an easy reformulation of Theorem 3.2 after application of partial integration.

Theorem 3.3. *Suppose that $p\in\mathbb{N}$, $m\in\mathbb{N}_0$. Assume that $\lambda\in\mathbb{R}$ with*

$$\lambda\neq \mathrm{L}^\wedge(m+1), \mathrm{L}^\wedge(m+2),\ldots.$$

Then, for all $x\in(-1,1)$ and $F\in C^{(2p)}([-1,1])$,

$$\begin{aligned} F(x) &= \sum_{n=0}^{m}\int_{-1}^{+1} F(t)P_n^*(t)\;dt\;P_n^*(x)\\ &\quad+\int_{-1}^{1} G^{\perp}_{0,\ldots,m}((\mathrm{L}+\lambda)^p;x,t)\;(\mathrm{L}_t+\lambda)^p F^{\perp}_{0,\ldots,m}(t)\;dt, \end{aligned}\tag{3.50}$$

where we have used the abbreviation

$$F^{\perp}_{0,\ldots,m} = F - \underbrace{\sum_{n=0}^{m}\int_{-1}^{1} F(t)P_n^*(t)\;dt\;P_n^*}_{=F_{0,\ldots,m}}. \tag{3.51}$$

Theorem 3.3 enables a comparison between the functional value $F(x)$ at the point x and the mth truncated orthogonal expansion of F in terms of Legendre polynomials at x with explicit representation of the remainder term in integral form.

Later on, this formula will be used to formulate adaptive estimates of the remainder terms in Gaussian numerical integration.

Gaussian Quadrature. Let $x_{1,n},\ldots,x_{n,n}$ with $x_{1,n}<\cdots<x_{n,n}$ be the zeros of the Legendre polynomial P_n of degree n. Then it is well known that there is precisely one zero of the polynomial P_{n+1} in each interval $(-1,x_{1,n}),\ldots,(x_{n,n},1)$.

For later use we want to prove the following preparatory result.

Lemma 3.5. *The matrix*

$$A=\Big(P_k^*(x_{j,n})\Big)_{\substack{k=0,\ldots,n-1\\ j=1,\ldots,n}} \tag{3.52}$$

is non-singular.

Proof. Assume that the matrix is singular. Then there exist coefficients $C_0,\ldots,C_{n-1}$ with $c=(C_0,\ldots,C_{n-1})^T\neq(0,\ldots,0)^T$ such that $c^TA=0$, i.e., the following polynomial of degree $\leq n-1$

$$Q(x)=\sum_{l=0}^{n-1} C_l P_l^*(x) \tag{3.53}$$

has as zeros $x_{1,n}, \ldots, x_{n,n}$. Hence, $Q = 0$. Now, for $k = 0, \ldots, n-1$,

$$0 = \int_{-1}^{1} Q(x) P_k^*(x)\ dx = C_k. \tag{3.54}$$

But this is a contradiction. Thus, Lemma 3.5 must be true. □

We are now interested in the *"Gaussian integration formula"* involving the "*Gaussian n-point formula*" $\mathcal{G}_n$

$$\mathcal{I}(F) = \int_{-1}^{1} F(t)\ dt \simeq \mathcal{G}_n(F) = \sum_{i=1}^{n} w_{i,n} F(x_{i,n}), \tag{3.55}$$

where the knots $x_{i,n}$, $i = 1, \ldots, n$, are the zeros of the Legendre polynomial P_n and the weights $w_{i,n}, i = 1, \ldots, n$, are determined by the linear system

$$\begin{aligned}
\sum_{i=1}^{n} w_{i,n} P_0^*(x_{i,n}) &= \int_{-1}^{1} P_0(t) P_0^*(t)\ dt &&= \sqrt{2},\\
\sum_{i=1}^{n} w_{i,n} P_1^*(x_{i,n}) &= \int_{-1}^{1} P_1(t) P_0^*(t) dt &&= 0,\\
&\ \ \vdots && \ \ \vdots\\
\sum_{i=1}^{n} w_{i,n} P_{n-1}^*(x_{i,n}) &= \int_{-1}^{1} P_{n-1}(t) P_0^*(t)\ dt = 0.
\end{aligned} \tag{3.56}$$

From Lemma 3.5 it is clear that the system (3.56) is uniquely solvable in the unknown weights $w_{1,n}, \ldots, w_{n,n}$.

Central for our considerations is the following well-known theorem (see, e.g., [206, 226]).

Theorem 3.4. *Let $x_{1,n}, \ldots, x_{n,n}$ be the zeros of the Legendre polynomial P_n. Furthermore, let $w_{1,n}, \ldots, w_{n,n}$ be the (unique) solution of the linear system* (3.56).

(i) *Then, the weights are positive, i.e., $w_{i,n} > 0$ for $i = 1, \ldots, n$, and we have*

$$\int_{-1}^{+1} P(t)\ dt = \sum_{i=1}^{n} w_{i,n} P(x_{i,n}) \tag{3.57}$$

for all $P \in \mathrm{Pol}_{0,\ldots,2n-1}$, i.e., for all algebraic polynomials of degrees $\leq 2n-1$,

(ii) *Conversely, if* (3.57) *is valid for real numbers $w_{i,n}, x_{i,n}$, $i = 1, \ldots, n$, then the knots $x_{i,n}$, $i = 1, \ldots, n$, are the zeros of the Legendre polynomial P_n and the weights $w_{i,n}, i = 1, \ldots, n$, satisfy the linear equations* (3.56).

(iii) *Finally, there exist no real numbers $x_{i,n}, w_{i,n}, i = 1, \ldots, n$, such that* (3.57) *holds true for all $P \in \mathrm{Pol}_{2n}$.*

Proof. Consider a polynomial $P \in \mathrm{Pol}_{0,\ldots,2n-1}$. Then there exist $Q, R \in Pol_{n-1}$ with

$$Q = \sum_{r=0}^{n-1} A_r P_r^* \tag{3.58}$$

and

$$R = \sum_{r=0}^{n-1} B_r P_r^* \tag{3.59}$$

such that

$$P = P_n^* Q + R. \tag{3.60}$$

Because of the orthogonality of the Legendre polynomials it is clear that

$$\begin{aligned}\int_{-1}^{+1} P(t)\, dt &= \int_{-1}^{1} P(t) \underbrace{P_0(t)}_{=1}\, dt \\ &= \int_{-1}^{1} Q(t) P_n^*(t)\, dt + \int_{-1}^{1} R(t) P_0(t)\, dt \\ &= \int_{-1}^{1} R(t) P_0(t)\, dt. \end{aligned}\tag{3.61}$$

In connection with (3.59) this implies

$$\int_{-1}^{1} P(t)\, dt = \sqrt{2}\, B_0. \tag{3.62}$$

Moreover, because of $P_n^*(x_{i,n}) = 0$, $i = 1, \ldots, n$, we obtain

$$\begin{aligned}\sum_{i=1}^{n} w_{i,n} P(x_{i,n}) &= \sum_{i=1}^{n} w_{i,n} \underbrace{P_n^*(x_{i,n})}_{=0} Q(x_{i,n}) + \sum_{i=1}^{n} w_{i,n} R(x_{i,n}) \\ &= \sum_{i=1}^{n} w_{i,n} \sum_{k=0}^{n-1} B_r P_r^*(x_{i,n}) \\ &= \sqrt{2}\, B_0. \end{aligned}\tag{3.63}$$

This shows us that (3.57) holds true.

Inserting the polynomials $\tilde{T}_k \in \mathrm{Pol}_{0,\ldots,2n-1}$, $k = 1, \ldots, n$, given by

$$\tilde{T}_k(x) = \prod_{\substack{j=1 \\ j \neq k}}^{n} (x - x_{j,n})^2, \quad x \in [-1, 1], \tag{3.64}$$

into (3.57) we get for $k = 1, \ldots, n$

$$0 < \int_{-1}^{+1} \tilde{T}_k(t)\, dt = \sum_{i=1}^{n} w_{i,n} \tilde{T}_k(x_{i,n}) = \sum_{i=1}^{n} w_{i,n} \prod_{\substack{j=1 \\ j \neq k}}^{n} (x_{i,n} - x_{j,n})^2, \tag{3.65}$$

hence, it is clear that

$$w_{i,n} > 0, \quad i = 1, \ldots, n. \tag{3.66}$$

This proves (i).

Next we verify (iii). Suppose that $w_{i,n}, x_{i,n}$, $i = 1, \ldots, n$, are given in such a way that (3.57) is valid for all $P \in \mathrm{Pol}_{0,\ldots,2n}$. Choose $H \in \mathrm{Pol}_{0,\ldots,2n}$ of the form

$$H(x) = \prod_{j=1}^{n} (x - x_{j,n})^2. \tag{3.67}$$

Then

$$0 < \int_{-1}^{1} H(x)\ dx = \sum_{i=1}^{n} w_{i,n} H(x_{i,n}) = 0. \tag{3.68}$$

This is a contradiction.

In order to guarantee (ii) we apply the Gauss rule (3.57) especially to the Legendre polynomials P_k^*, $k = 0, \ldots, n-1$, such that

$$\sum_{i=1}^{n} w_{i,n} P_k^*(x_{i,n}) = \int_{-1}^{1} P_k^*(x)\ dx = \int_{-1}^{1} P_k^*(x) \underbrace{P_0(x)}_{=1}\ dx. \tag{3.69}$$

In other words, the weights have to satisfy (3.56), i.e.,

$$\sum_{i=1}^{n} w_{i,n} P_k^*(x_{i,n}) = \left\{ \begin{array}{rl} \sqrt{2}, & k = 0, \\ 0, & k = 1, \ldots, n-1. \end{array} \right. \tag{3.70}$$

We now consider $P = P_k^* P_n^*$ for $k = 0, \ldots, n-1$, and apply (3.57). This gives

$$0 = \int_{-1}^{1} P_k^*(t) P_n^*(t)\ dt = \sum_{i=1}^{n} w_{i,n}\ P_k^*(x_{i,n})\ P_n^*(x_{i,n}) \tag{3.71}$$

for $k = 0, \ldots, n-1$. Consequently, the vector $a = (A_1, \ldots, A_n)^T$ given by

$$(A_1, \ldots, A_n)^T = (w_{1,n} P_n^*(x_{1,n}), \ldots, w_{n,n} P_n^*(x_{n,n}))^T \tag{3.72}$$

satisfies the linear system

$$0 = \sum_{i=1}^{n} A_i P_k^*(x_{i,n}), \quad k = 0, \ldots, n-1, \tag{3.73}$$

where the coefficient matrix is non-singular. But this means that $A_i = 0$, $i = 1, \ldots, n$, hence, $P_n(x_{i,n}) = 0$, $i = 1, \ldots, n$, as required. □

Remainder Terms Involving Green's Function. Now, the integral formula (Theorem 3.3) comes into play. For all values $\lambda \in \mathbb{R} \backslash \{\mathrm{L}\hat{}(2n), \mathrm{L}\hat{}(2n+1), \ldots\}$ and for functions $F \in C^{(2p)}([-1,+1])$, $p \in \mathbb{N}$, we have

$$\begin{aligned} \sum_{i=1}^{n} w_{i,n} F(x_{i,n}) = & \int_{-1}^{1} F(t)\ dt \\ & + \sum_{i=1}^{n} w_{i,n} \int_{-1}^{1} G^{\perp}_{0,\ldots,2n}((\mathrm{L}+\lambda)^p; x_{i,n}, t)\ (\mathrm{L}_t + \lambda)^p F(t)\ dt. \end{aligned} \tag{3.74}$$

In other words, the remainder

$$\mathcal{E}_n(F) = \mathcal{I}(F) - \mathcal{G}_n(F), \tag{3.75}$$

when the integral

$$\mathcal{I}(F) = \int_{-1}^{1} F(t)\ dt, \quad F \in C^{(2p)}([-1,1]), \tag{3.76}$$

is replaced by the "*Gaussian n-point formula*"

$$\mathcal{G}_n(F) = \sum_{i=1}^{n} w_{i,n} F(x_{i,n}), \tag{3.77}$$

is expressible by means of the $2n$th truncated Legendre (Green) function in integral form

$$\mathcal{E}_n(F) = -\sum_{i=1}^{n} w_{i,n} \int_{-1}^{1} G^{\perp}_{0,\ldots,2n}((\mathrm{L}+\lambda)^p; x_{i,n}, t)\ (\mathrm{L}_t+\lambda)^p F(t)\ dt \tag{3.78}$$

for all $\lambda \in \mathbb{R}\backslash\{\mathrm{L}^\wedge(2n), \mathrm{L}^\wedge(2n+1), \ldots\}$.

Applying the Cauchy–Schwarz inequality to (3.78) we obtain

$$|\mathcal{E}_n(F)| \le \sqrt{A^{(p)}_{\lambda}(n)} \sqrt{\int_{-1}^{1} |(\mathrm{L}_t+\lambda)^p F(t)|^2\ dt}, \tag{3.79}$$

for $\lambda \in \mathbb{R}\backslash\{\mathrm{L}^\wedge(2n), \mathrm{L}^\wedge(2n+1), \ldots\}$, where we have used the abbreviation

$$A^{(p)}_{\lambda}(n) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,n} w_{j,n} G^{\perp}_{0,\ldots,2n}((\mathrm{L}+\lambda)^{2p}; x_{i,n}, x_{j,n}) \tag{3.80}$$

with

$$\begin{aligned} &G^{\perp}_{0,\ldots,2n}((\mathrm{L}+\lambda)^{2p}; x_{i,n}, x_{j,n}) \\ &\quad = \int_{-1}^{1} G^{\perp}_{0,\ldots,2n}((\mathrm{L}+\lambda)^p; x_{i,n}, t) G^{\perp}_{0,\ldots,2n}((\mathrm{L}+\lambda)^p; t, x_{j,n})\ dt. \end{aligned} \tag{3.81}$$

Summarizing our results we obtain the following remainder estimate.

Theorem 3.5. *Let $x_{i,n}, i = 1, \ldots, n$, be the zeros of the Legendre polynomial P_n, $n \ge 1$, and let $w_{i,n}, i = 1, \ldots, n$, be the unique solution of the linear system*

$$\sum_{i=1}^{n} w_{i,n} P_k(x_{i,n}) = \int_{-1}^{1} P_0(t) P_k(t)\ dt, \tag{3.82}$$

$k = 0, \ldots, n-1$.

Then, for values $\lambda \in \mathbb{R}\backslash\{\mathrm{L}^\wedge(2n), \mathrm{L}^\wedge(2n+1), \ldots\}$ and for functions $F \in C^{(2p)}[-1,1]$, we have

$$\left| \int_{-1}^{1} F(t)\ dt - \sum_{i=1}^{n} w_{i,n} F(x_{i,n}) \right| \le \sqrt{A^{(p)}_{\lambda}(n)} \sqrt{\int_{-1}^{1} \left|(\mathrm{L}_t+\lambda)^p F(t)\right|^2\ dt}, \tag{3.83}$$

where

$$A_\lambda^{(p)}(n) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,n}w_{j,n} \sum_{k=2n}^{\infty} \frac{1}{(\lambda - \mathrm{L}^\wedge(k))^{2p}} P_k^*(x_{i,n})P_k^*(x_{j,n}). \tag{3.84}$$

Remainder Term Estimation. From (3.78), we get the estimate

$$|\mathcal{E}_n(F)| \le \sqrt{A_\lambda^{(p)}(n)}\sqrt{\int_{-1}^{1} \left|(\mathrm{L}_t + \lambda)F_{0,\dots,2n-1}^{\perp}(t)\right|^2 \, dt}, \tag{3.85}$$

where

$$F_{0,\dots,2n-1}^{\perp}(t) = F(t) - \underbrace{\sum_{n=0}^{2n-1} \int_{-1}^{1} F(u)P_n^*(u)\, du\; P_n^*(t)}_{=F_{0,\dots,2n-1}}, \tag{3.86}$$

$t \in [-1,1]$.

Observing the estimate $|P_n(t)| \le 1$ for all $n \in \mathbb{N}_0$ and $t \in [-1,1]$ and the properties of the Gaussian weights we get

$$|A_\lambda^p(n)| \le \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,n}w_{j,n} \sum_{k=2n}^{\infty} \frac{2k+1}{(\lambda - \mathrm{L}^\wedge(k))^{2p}} = 2\sum_{k=2n}^{\infty} \frac{2k+1}{(\lambda - \mathrm{L}^\wedge(k))^{2p}} \tag{3.87}$$

for all $\lambda \in \mathbb{R}\backslash\{\mathrm{L}^\wedge(2n), \mathrm{L}^\wedge(2n+1), \dots\}$.

Theorem 3.6. *For* $F \in C^{(2n)}([-1,1])$,

$$\begin{aligned}\int_{-1}^{1} F(t)\, dt - \sum_{i=1}^{n} w_{i,n}F(x_{i,n}) \\ = \sum_{i=1}^{n} w_{i,n} \int_{-1}^{1} G_{2n}^{\perp}\big((L+\lambda)^n; x_{i,n}, t\big)(\mathrm{L}_t + \lambda)^n F_{0,\dots,2n-1}^{\perp}(t)\, dt,\end{aligned} \tag{3.88}$$

where $F_{0,\dots,2n-1}^{\perp}(t)$ *is defined by* (3.86).

Convergence of Gaussian Quadrature. From Theorem 3.5 we obtain

$$\left|\int_{-1}^{1} F(t)\, dt - \sum_{i=1}^{n} w_{i,n}F(x_{i,n})\right| \le \sqrt{A_0^{(p)}(n)}\sqrt{\int_{-1}^{1} (\mathrm{L}_t F(t))^2\, dt}, \tag{3.89}$$

where

$$A_0^{(p)}(n) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,n}w_{j,n} \sum_{k=2n}^{\infty} \frac{1}{(k(k+1))^{2p}} P_k^*(x_{i,n})P_k^*(x_{j,n}). \tag{3.90}$$

The sequence

$$\left\{2\sum_{k=2n}^{\infty} \frac{2k+1}{(k(k+1))^{2p}}\right\}_{n\in\mathbb{N}} \tag{3.91}$$

is (strictly) monotonously decreasing and positive. Furthermore, we have the rough estimate

$$0 \leq A_0^{(p)}(n) \leq 2 \sum_{k=2n}^{\infty} \frac{2k+1}{(k(k+1))^{2p}}. \tag{3.92}$$

Therefore, we come to the conclusion.

Theorem 3.7. *The sequence* $\{A_o^{(p)}(n)\}_{n\in\mathbb{N}}$ *has the limit*

$$\lim_{n\to\infty} A_0^p(n) = 0 \tag{3.93}$$

such that, for all $F \in C^{(2p)}([-1,1])$,

$$\lim_{n\to\infty} \mathcal{E}_n(F) = 0, \tag{3.94}$$

i.e., the Gaussian quadrature rule is convergent:

$$\lim_{n\to\infty} \sum_{i=1}^{n} w_i F(x_{i,n}) = \int_{-1}^{1} F(t)\ dt. \tag{3.95}$$

Comparing different numerical integration techniques with the Gaussian method we see that the last method assures the highest degree of polynomial precision. The critical point of the estimate (3.85), however, is to find an estimate for $(\mathrm{L}+\lambda)^{2p}F(t)$ with $(L+\lambda)^{2p}$ being appropriately adapted to the specific properties of the function such that the bound of $|\mathcal{E}_n(F)|$ becomes small. Moreover, it should be critically mentioned that the pointsets $\{x_{1,n},\ldots,x_{n,n}\}_{n\in\mathbb{N}}$ are not hierarchically distributed.

Gauss's work was simplified by Jacobi and further developed by F.G. Mehler [149], E.B. Christoffel [19], and others through much of the 19th century. Eventually, there emerged a coherent theory which received its first systematic expositions by E.B. Christoffel [19], R. Radau [177], and E. Heine [106] in his book on spherical functions. An outline of the developments in numerical quadrature that took place in a period of approximately 200 years from Newton via Gauss to early developments in the 20th century can be found in [188], and a German edition of the four principal memoirs (of Newton, Cotes, Gauss, and Jacobi) is due to A. Kowalewski [129]. A more detailed survey on developments up to the second half of the 20th century is given by W. Gautschi [85], culminating in the following comment:

> *"Gauss's discovery must be rated as one of the most significant events of the* 19*th century in the field of numerical integration and perhaps in all of numerical analysis. The result not only has great beauty and power, but also influenced many later developments in computing and approximation. It soon inspired contemporaries, such as Jacobi and Christoffel, to perfect Gauss's method and to develop it into new directions. Towards the end of the century, it inspired K. Heun* [109] *to generalize Gauss's idea to ordinary differential equations, which in turn led to significant developments in the numerical solution of differential equations, notably the discovery of the Runge–Kutta method* (*cf.* [133])*. Gauss's influence continues into the* 20*th century and is still felt today, ..."*

Whether or not the Gauss method had actually been widely used in practice up to the seventieth of the last century is a matter of some doubt, since the method requires the evaluation of functions at irrational arguments, hence, tedious interpolation. All this changed when powerful digital computers entered the scene, which generated a phase of renewed interest in Gaussian quadrature. The formulas began to be routinely applied, and increased usage, in turn, led to important new theoretical developments.

3.2. Periodic approximate integration

Next we are interested in trapezoidal sums based on Euler-type summation formulas for approximate integration of one-dimensional periodic functions.

To this end we first list some obvious results on *"τ-dilated lattices"*

$$\tau\mathbb{Z} = \{\tau g : g \in \mathbb{Z}, \tau > 0\}. \tag{3.96}$$

The *fundamental cell* $\mathcal{F}_{\tau\mathbb{Z}}$ of the lattice $\tau\mathbb{Z}$ is given by

$$\mathcal{F}_{\tau\mathbb{Z}} = \left\{ x \in \mathbb{R} : -\frac{1}{2}\tau \le x < \frac{1}{2}\tau \right\}. \tag{3.97}$$

A function $F : \mathbb{R} \to \mathbb{C}$ is called *$\tau\mathbb{Z}$-periodical* if $F(x+g) = F(x)$ holds for all $x \in \mathcal{F}_{\tau\mathbb{Z}}$ and $g \in \tau\mathbb{Z}$. The function $\Phi_h : \mathbb{R} \to \mathbb{C},\ h \in \frac{1}{\tau}\mathbb{Z}$, given by

$$x \mapsto \Phi_h(x) = \frac{1}{\sqrt{\tau}} e^{2\pi i h x} \tag{3.98}$$

is $\tau\mathbb{Z}$-periodical and satisfies the identity

$$\int_{\mathcal{F}_{\tau\mathbb{Z}}} \Phi_h(x) \Phi_{h'}(x)\ dx = \left\{ \begin{array}{ll} 1, & h = h' \\ 0, & h \ne h'. \end{array} \right. \tag{3.99}$$

The space of all $F \in C^{(m)}(\mathbb{R})$ that are $\tau\mathbb{Z}$-periodical is denoted by $C^{(m)}_{\tau\mathbb{Z}}(\mathbb{R})$, $0 \le m \le \infty$. Clearly, the space $L^2_{\tau\mathbb{Z}}(\mathbb{R})$ is the completion of $C^{(0)}_{\tau\mathbb{Z}}(\mathbb{R})$ with respect to the norm $\|\cdot\|_{L^2_{\tau\mathbb{Z}}(\mathbb{R})}$. The system $\{\Phi_h\}_{h \in \frac{1}{\tau}\mathbb{Z}}$ is orthonormal with respect to the $L^2_{\tau\mathbb{Z}}(\mathbb{R})$-inner product. By convention, we say that λ is an *eigenvalue of the lattice $\tau\mathbb{Z}$ with respect to the operator* Δ of the second-order derivative (i.e., the one-dimensional Laplace operator), if there is a non-trivial solution U of the differential equation $(\Delta + \lambda)U = 0$ satisfying the "boundary condition" of periodicity $U(x+g) = U(x)$ for all $x \in \mathcal{F}_{\tau\mathbb{Z}}$ and $g \in \tau\mathbb{Z}$. From classical Fourier analysis we know that the operator Δ has a half-bounded and discrete eigenspectrum $\{\Delta^\wedge(h)\}_{h \in \frac{1}{\tau}\mathbb{Z}} \subset \mathbb{R}$ such that $(\Delta_x + \Delta^\wedge(h))\,\Phi_h(x) = 0, x \in \mathcal{F}_{\tau\mathbb{Z}}$, with eigenvalues $\Delta^\wedge(h)$ given by $\Delta^\wedge(h) = 4\pi^2 h^2, h \in \frac{1}{\tau}\mathbb{Z}$, and eigenfunctions $\Phi_h(x) = \tau^{-\frac{1}{2}} e^{2\pi i h x}, h \in \frac{1}{\tau}\mathbb{Z}, x \in \mathcal{F}_{\tau\mathbb{Z}}$.

Consequently, the *eigenspectrum of the operator* Δ (with respect to $\tau\mathbb{Z}$) is given by

$$\mathrm{Spect}_\Delta(\tau\mathbb{Z}) = \left\{ \Delta^\wedge(h) : \Delta^\wedge(h) = 4\pi^2 h^2,\ h \in \frac{1}{\tau}\mathbb{Z} \right\}. \tag{3.100}$$

Trapezoidal Integration Rule. For purposes of *numerical integration of one-dimensional periodic functions* we introduce the *definition of the $\tau\mathbb{Z}$-lattice Green* both points Gauss's idea proved to be very fertile. His formulas are distinguished from other solutions due to their superior convergence, but also because Gauss had proven their accuracy and the size of the neglected terms. Gauss derived these formulas in two different ways: through a conformal mapping of the ellipsoid to the sphere and through power series expansions stemming from the differential equations of the geodesic on the ellipsoid of rotation. Gauss, however, did not present the formulas. This was done by F.R. Helmert [108] based on Gauss's formulas of the mean width (cf. [141]). A disadvantage of the series ansatz using arguments of means lies in the fact that the coefficients of the series have to be newly computed for each new pair of points, while the coefficients of the Legendre series have to be computed only once if the coordinates of multiple points have to be determined with respect to the same pole.

Approximately since 1965 the possibilities of numerical methods to solve the main tasks of geodesy, provided by standard methods of numerical mathematics, can be exploited due to the beginning development of electronic data processing computers. The variety of potential solutions can be divided roughly in four categories (according to B. Heck [105]):

(i) The first kind of approach rests on classical solutions of O. Schreiber [196] and H. Boltz [12]. Since poles are singular points of the geographical coordinate system Legendre's series cannot be used in the vicinity of poles. Moreover, these series show slow convergence in case of large distances so that this approach makes sense only in case of a short geodesic up to 150 km and up to 400 km if Gauss's formulas are used.

(ii) The second category rests on Bessel's approach (cf. [6]) of exploiting elliptic integrals which are today directly computed by numerical quadrature rules (see, e.g., [193]). This approach can also be used in the vicinity of poles and is accurate and advisable even in case of large distances of > 500 km.

(iii) The solution approaches in the third category are based on Gauss's conformal mapping of the ellipsoid to the sphere or the plane (cf. [84]) which are considered as auxiliary surfaces on which the main tasks of geodesy can be accomplished by means of closed formulas from spherical and plane trigonometry, respectively. Subsequently, the inverse mapping may be used to transfer back to the ellipsoid.

(vi) In the fourth category we find methods which were already used by A.M. Legendre [139] in a special case, namely that a given geodetic line is replaced by another line connecting start and end point of the geodesic. This other line may be a spatial chord, for example. Due to geodetic use of Earth satellites this three-dimensional ansatz has grown in importance.

6.2. Gaussian conformal mapping of the Earth's surface

For quite some time Gauss had dealt with the question of the method best suited to coordinate triangulation points because the knowledge of geographical coordinates is not sufficient. It seemed most convenient to him to exploit the coordinates with which every point could be represented in a plane (Gauss Werke, Vol. V, p. 367).

About 1815 he recognized the conformal mapping of the triangular points onto the plane as being the most appropriate solution to the problem. The term "konform" was used by him since 1843. The task formulated by Gauss, to project a given surface onto another one in a fairly general manner so that the image becomes similar to the preimage in its smallest parts was formulated as a contest question for the year 1821 by the Scientific Society in Copenhagen and reformulated the next year. Gauss submitted his solution (cf. [78]) on the 11, December 1822 and won the prize. His solution was distinguished from other known work in that it was valid in case of arbitrary surfaces and that he had fully laid open the conditions of conformity, the constancy of the augmenting relation in a certain point, and the equal angle condition (cf. [141]). Gauss applied his general solution to the following particular cases:

- *Conformal mapping between two planes.* As Gauss himself remarked (cf. Art. 8 in [78]), this is a useful method for the geodetic praxis, in order "to transform a map based on mediocre measurements which may be good in small details but is generally somehow distorted into a better one, if one knows the correct loci of a number of points."
- Conformal mapping of the sphere by transversal Mercator's projection or other mappings, e.g., the *conforming Lambert mapping.*
- *Conformal mappings of the ellipsoid of rotation onto the sphere.* In this mapping one can choose a suitable radius of the sphere to keep the differences between the ellipsoidal and spherical geometries relatively small. Hence, for practical applications spherical relations can be used which allow closed form solution. In case of local computations reasonable accuracy can be gained by letting the radius of the sphere depend on the geographic latitude. To this end two "replacement spheres" have proved their worth, namely the image sphere of Soldner, formerly used in Bavarian land surveying, and the Gaussian osculating sphere. The latter was used by Gauss (see [84]) with the radius M_0N_0 with regard to a point P_0, M_0 being the radius of curvature of the meridian and N_0 being the oblique radius of curvature.

  Since the ellipsoid shows the same measure of curvature in the central point P_0 in this mapping the differences in scale between ellipsoid and sphere stay small for regions not too large. As Gauss has shown the metric on a surface is determined by the Gaussian curvature. Hence, line segments and angles agree on two surfaces with the same Gaussian curvature. The surface of an ellipsoid, however, shows a variable Gaussian curvature so that the relationship of curvatures are only equivalent in the small (cf. Figures 6.2 and 6.3).

Under all mappings associated with Gauss's name the so-called *Gauss–Krüger projection* (and similar ones) have enjoyed wide distribution. In this projection the main meridian is mapped length-preserving to a straight line and is the abscissa of the plane system. Gauss had chosen this projection for the land surveying of Hanover (1828–1844) (cf. [141]). The size of the distortion depends only on the

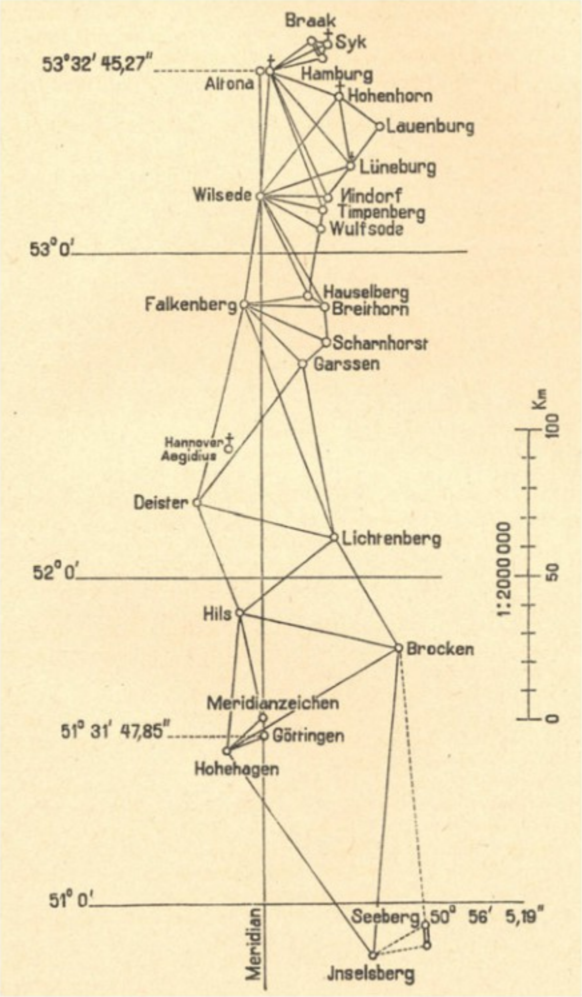
Braak
Syk
53°32′ 45,27″
Altona
Hamburg
Hohenhorn
Lauenburg
Lüneburg
Wilsede
Nindorf
Timpenberg
Wulfsode
53° 0′
Hauselberg
Falkenberg
Breithorn
Scharnhorst
Garssen
Hannover Aegidius
Deister
Lichtenberg
52° 0′
Hils
Brocken
Meridianzeichen
51° 31′ 47,85″
Göttingen
Hohehagen
51° 0′
Seeberg
50° 56′ 5,19″
Jnselsberg
Meridian
1:2000000
0
50
100 Km

FIGURE 6.1. Triangulation of Hanoverian arc measurement between Göttingen and Altona (now a suburb of Hamburg) carried out under Gauss's supervision 1821–1825, Collected Works, Vol. 9, p. 347 (Illustration taken from [118]).

distance between the points and the main meridian and is independent from the north-south dilatation. Today one follows the idea that meridian strips of 6° width with a difference in longitude of 3° from the main meridian to the boundary meridians suffice for tolerable distortions (this is today's UTM-coordinate system).

Gauss's work on surface theory was important for the later nineteenth century in the sense of Sophus Lie, while until his time in geometry only finite groups of transformations had been considered. Gauss paved the way for the general theory

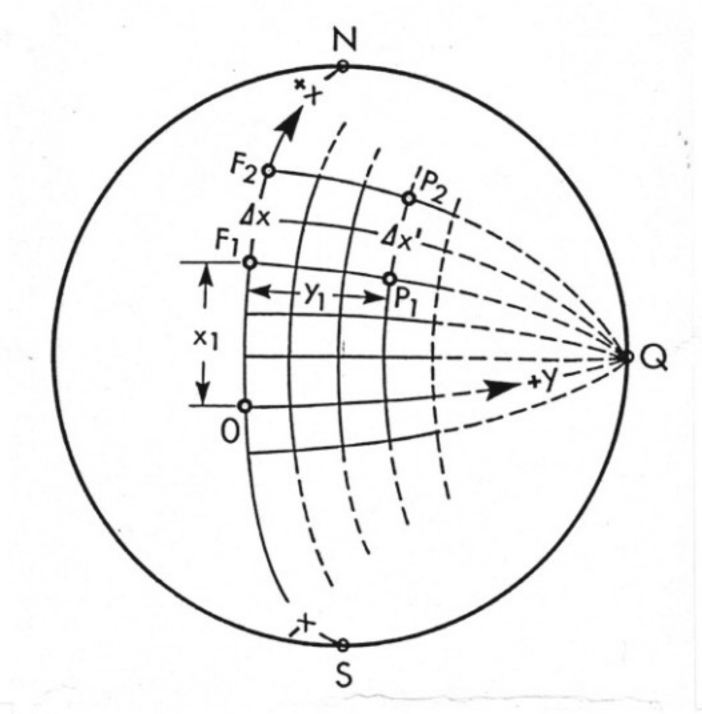
N
+x
F_2
P_2
Δx
Δx'
F_1
y_1
P_1
x_1
Q
+y
0
+
S

FIGURE 6.2. Ellipsoidal orthogonal coordinates (Illustration taken from [233]).

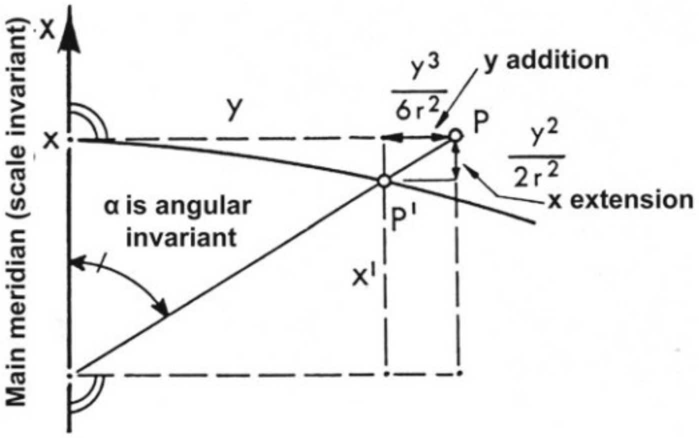
X
Main meridian (scale invariant)
x
y
$\frac{y^3}{6r^2}$
y addition
P
$\frac{y^2}{2r^2}$
x extension
P'
α is angular invariant
x'

FIGURE 6.3. Differential projection distortions (Illustration taken from [233]).

of the multiply extended manifolds, or n-dimensional space (see [27] for more details).

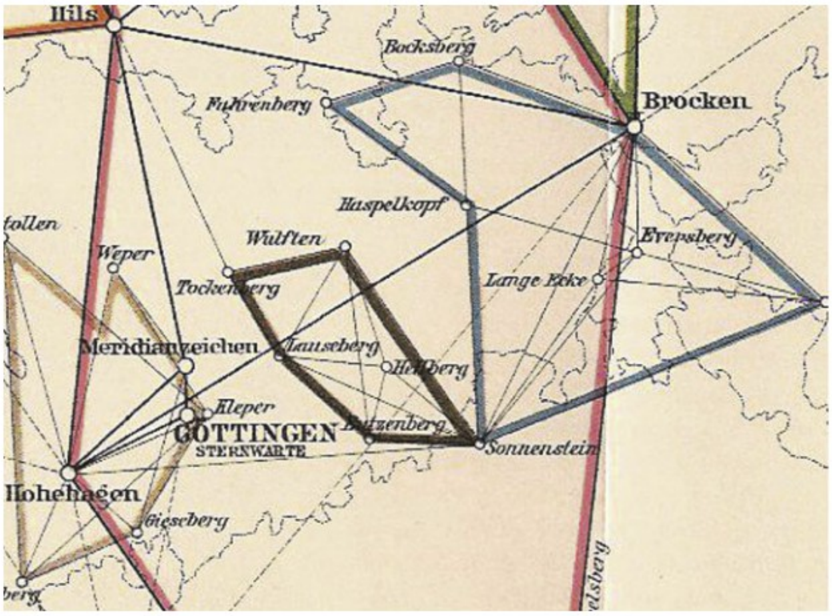
Hils
Bocksberg
Fuhrenberg
Brocken
Haspelkopf
tollen
Wulften
Weper
Tockenberg
Erepsberg
Lange Ecke
Meridianzeichen
Lauseberg
Hellberg
Kleper
GÖTTINGEN
STERNWARTE
Butzenberg
Sonnenstein
Hohenhagen
Gieseberg
berg
elsberg

FIGURE 6.4. A detail of the triangulation of Hanover carried out under Gauss's supervision 1821–1825 (Collected Works, Vol. 9, p. 347). Note that the three lines joining the locations Hohenhagen, Brocken, and Inselsberg (this point does not appear here, only a part of the connecting lines is displayed) form a great triangle, the angle at Hohenhagen is close to a right angle, so that the area of the triangle is close to half the product of the two short sides.

The Hanover surveying work also stimulated Gauss's interest in the study of curves and surfaces in three-dimensional differential geometry in Euclidean space (today usually called *Gaussian geometry*). Gauss also was led to the Gaussian curvature (an intrinsic measure of curvature, dependent only on how distances are measured on the surface, not on the way it is embedded in space).

6.3. Claims and perspectives

All in all, while engaged on a surveying task for the Royal House of Hanover in the years after 1818, Gauss was also concerned with the shape of the Earth. He started to formulate revolutionary ideas, like the concept of the geoid. He questioned one of the central tenets of the whole of mathematics, Euclidean geometry, which was clearly premised on a flat, and not a curved, universe. He later claimed to have considered a non-Euclidean geometry, which was internally consistent and free of contradiction. Unwilling to court controversy, however, Gauss decided not to pursue or publish any of his far-reaching concepts in non-Euclidean geometry. János Bolyai independently discovered non-Euclidean geometry in 1829; his work was published in 1832. After seeing it, Gauss wrote to Farkas Bolyai, the father of János Bolyai:

> *"To praise it would amount to praising myself. For, the entire content of the work ... coincides almost exactly with my own meditations which have occupied my mind for the past thirty or thirty-five years."*

The long history of the discussion of the parallel postulate started probably sooner than Euclid published his *Elements* about 300 BC as his fifth postulate [32], p. 155:

> *"That, if a straight line falling on two straight lines make the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which are the angles less than the two right angles."*

Many modern authors have commented on the meaning of this postulate; we mention only R. Bonola [13], J. Gray [90], B.A. Rosenfeld [184]. In fact, even in our time the fifth Euclidean postulate seems odd and we may understand why so many authors after Euclid tried to prove the fifth postulate (and thus making it a theorem) from the other axioms and postulates. Already Proclus (412–485), who wrote a Commentary on the First Book of Euclid's Elements, tried to prove the fifth postulate exploiting an argument given by Aristoteles to show the finiteness of the universe [176], p. 291, and failed, cf. [90], p. 39. Proclus informs us, that the first of the Ptolemies under whose reign Euclid lived, wrote himself a book on the fifth postulate and proved it, but Proclus points out the fallacy of this attempt. John Playfair (1748–1819) gave an alternative formulation of the fifth postulate in 1795 which is now known as *Playfair's Axiom* but which he himself attributed to William Ludlam (1717–1788). This formulation is the one which most people refer to nowadays when speaking of the "axiom of parallels" [239, p. 16]:

> *"Given a line a and a point A not lying on a, then there exists, in the plane determined by a and A, one and only one line which contains A but not any point of a."*

Or, even shorter,

> *"Through a given point can be drawn only one parallel to a given line."*

Many mathematicians tried to get hold of a proof of the fifth postulate; famous names over the centuries being Nasir al-Din al-Tusi (1201–1274) in Persia, John Wallis (1616–1703) in England, Giovanni Girolamo Saccheri (1667–1733) in Italy, the Swiss mathematician Johann Heinrich Lambert (1728–1777) in Germany, and Adrien-Marie Legendre (1752–1833) in France. At the beginning of the 19th century the invention of a Non-Euclidean Geometry seemed to have been in the air. If so many and different attempts to prove the fifth postulate a mere conclusion of other axions had failed the question whether there exists a geometry in which the fifth postulate was wrong became sensible. This may explain why it was not only Gauss who found the key to this new geometry but also Nikolai Lobachevsky (1792–1856) and János Bolyai (1802–1860) independently. However, Gauss was the first, although he did not publish his results. It is now known that it was he who coined the denotation "Non-Euclidean" for the new geometries which emerge out of a neglect of the fifth postulate. Already in 1804 Gauss received a letter of his friend Farkas Bolyai (1775–1856), father of János Bolyai, in which Bolyai presented a proof that the fifth postulate could be deduced from other axioms of geometry. Gauss praised the work of his friend, but found a flaw in the argument. He wrote (cf. [16, p. 100]):

> *"You want to know my sincere and frank opinion. And this is that your method does not satisfy me. I will try to make the critical point* (*which belongs to the same kind of obstacles which made my own efforts so futile*) *as clear as I can. I still hope that these cliffs will be navigated eventually, and this, before I die. For now, I am, however, extremely busy with other things…"*

Although Gauss remarked in 1846 that he knew about the existence of Non-Euclidean geometries for the last fifty years it is not before 1816 that we see written evidence. In that year Gauss reflected on different attempts to prove the fifth postulate in a book review. Although he was too careful to express his own opinion we can be sure that the reason why he reviewed these attempts can be seen in the fact that Gauss was already convinced in 1816 that Non-Euclidean geometries existed.

Perhaps the Non-Euclidean geometry which can be most easily understood is the spherical geometry. The navigators, map makers, and naval mathematicians of the 16th century were very well aware of this type of geometry in which every triangle (build from parts of three great circles on the sphere) shows a sum of inner angles of larger than 180° and is thus today seen as a simple model of elliptic geometry. The Non-Euclidean geometries by Gauss, Lobachevsky, and Bolyai are nowadays classified as being hyperbolic geometries, in which a triangle shows an angular sum of less than 180°.

Spherical geometry was Gauss's bread-and-butter occupation as early as 1803 when he started to survey the Duchy of Brunswieck by means of a triangulation and met his wife to be, Johanna Osthoff. After Heinrich Christian Schumacher (1780–1850), astronomer and geodesist, informed Gauss of the Danish arc measurement and stimulated a surveying of the Hanoverian lands to fit the Danish measurements in the north, Gauss became enthusiastic. Eventually, Schumacher succeeded in persuading the English king Georg IV, who was head of the house of Hanover, to authorize Gauss with the surveying work, and Gauss started field work in 1821 (see Figures 6.1 and 6.4) concerned with measurements from three mountains in Germany, Hohenhagen, near Göttingen, Brocken in the Harz Mountains and Inselberg in the Thüringer Wald to the south. The three lines joining these locations form a great triangle, the angle at Hohenhagen is close to a right angle, so the area of the triangle is close to half the product of the two short sides). Gauss made the important invention of the heliotrope (cf. Figure 6.5) to ensure measurements of hitherto unknown accuracy. In fact, heliotropes were used

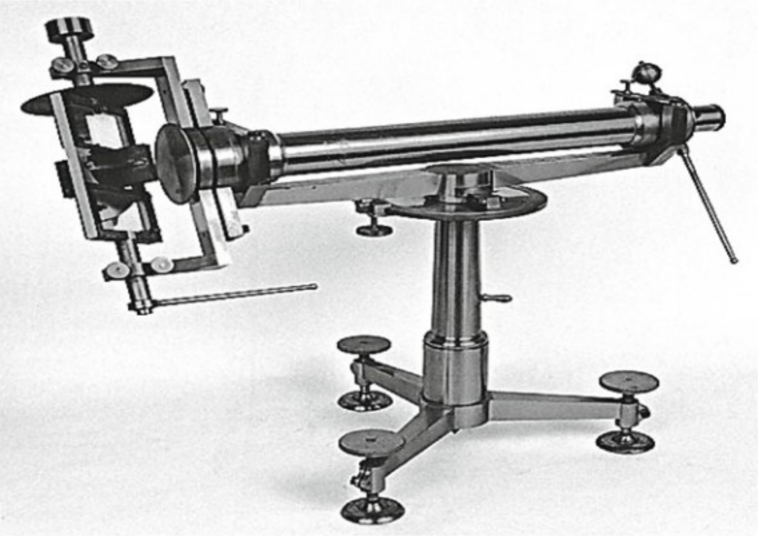

FIGURE 6.5. Heliotrope (Geophysical Institute of Göttingen University).

in surveys from Gauss's survey in Germany in 1821 through the late 1980s, when GPS measurements replaced the use of the heliotrope in long distance surveys. Surveying field work lasted until 1825 when Gauss withdraw from working in the countryside. However, he oversaw the further surveying work going on in 1828 and lasting until 1844 and did all the number crunching necessary. Friends, including Friedrich Wilhelm Bessel (1784–1846), criticized him for wasting his time in these computations instead of creating new theorems, but Gauss saw deeper. In a letter to Bessel dated March 14, 1824, he wrote (cf. [8]):

> *"... you accused me of loosing my time and wished me luck that the loss of time might be over soon. Great God, how wrong you judge me. ... Certainly, I also think like you in that matter. All measurements in the world do not outweigh a single theorem, with which the science of eternal truths will be truly advanced. But you should not judge over the absolute, but over the relative worth* [of measurements]. ... *And however small you estimate this worth, in my eyes it is higher than those concerns which are interrupted by them."*

Surveying problems also motivated Gauss to develop his thoughts on least squares and more general problems of what is now called mathematical statistics. The result was the definitive exposition of his mature concepts in the note "Theoria combinationis observationum erroribus minimis obnoxiae" (1823, with supplement in 1826). In "Bestimmung des Breitenunterschiedes zwischen den Sternwarten von Göttingen and Altona durch Beobachtungen am Ramsdenschen Zenithsector" of the year 1828 he summed up his ideas on the figure of the Earth, instrumental errors, and the calculus of observations. Furthermore, his publication "Disquisitiones generales circa superficies curvas" (1828), which grew out of his meditations in surveying and geodesy of three decades, represented the seed of more than a century of work on differential geometry.

Finally it should be mentioned that resulting research led to, among other things, Einstein's theory of general relativity, which describes the universe as non-Euclidean.

References

[1] Adams, D.R., Hedberg, L.I.: *Function Spaces and Potential Theory.* Grundlehren der Mathematischen Wissenschaften, 314, Springer, Berlin, (1996).

[2] Adrain, R.: *Research Concerning the Probabilities of the Errors which Happen in Making the Observations.* The Analyst of Mathematical Museum, Vol. I, W.P. Farrand and Co., Philadelphia, (1808).

[3] Augustin, M., Freeden, W.: A Survey on Classical Boundary Value Problems in Physical Geodesy. Encyclopedia of Geodesy, E.W. Grafarend, ed., Springer Int. Publ. Switzerland, DOI 10.1007/978-3-319-02370-0-1117-1, (2015).

[4] Baer, R.: *Linear Algebra and Projective Geometry.* Academic Press, New York, (1952).

[5] Ben-Israel, A., Greenville, T.N.E., *Generalized Inverses: Theory and Applications.* New York: Wiley-Interscience, (1974).

[6] Bessel, W.: Über die Berechnung der geographischen Längen und Breiten aus geodätischen Vermessungen. Astronomische Nachrichten, Bd. 4, Nr. 86, Altona, (1826).

[7] Bialas, V.: *Erdgestalt, Kosmologie und Weltanschauung.* Konrad Wittwer Stuttgart, (1982).

[8] Biermann, K.-R: *Carl Friedrich Gauss – Der "Fürst der Mathematiker" in Briefen und Gesprächen.* Verlag C.H. Beck, München, (1990).

[9] Bjerhammar, A.: Rectangular Reciprocal Matrices with Special Reference to Geodetic Calculations. Bulletin Géodesique, 188–220, (1951).

[10] Bjerhammar, A.: *Gravity Reduction to an Internal Sphere.* Division of Geodesy, Stockholm, Sweden, (1962).

[11] Börsch, A.: Simon, P.: Abhandlungen zur Methode der kleinsten Quadrate von Carl Friedrich Gauss (in deutscher Sprache herausgegeben). Berlin, (1887).

[12] Boltz, H.: *Formeln und Tafeln zur numerischen (nicht logarithmischen) Berechnung geographischer Koordinaten aus den Richtungen und Längen der Dreiecksseiten erster Ordnung.* Veröff. Geod. Inst. Potsdam, Neue Folge, Nr. 110, (1942).

[13] Bonola, R: *Non-Euclidean Geometry – With a Supplement Containing 'The Theory of Parallels' by Nicholas Lobachevski and 'The Science of Absolute Space' by John Bolyai.* Engl. transl. of the first edition 1912. Dover Publication, New York, (1955).

[14] Boullion, T.L., Odell, P.L.: *Generalized Inverse Matrices.* Wiley-Interscience, New York, (1971).

[15] Bruns, H.: Die Figur der Erde "Ein Beitrag zur europäischen Gradmessung". P. Stankiewicz, Berlin, (1878).

[16] Bühler, W.K.: *Gauss. A Biographical Study.* Springer-Verlag, Berlin, Heidelberg, New York, (1981).

[17] Bühler, W.K.: *Gauss – Eine Biographische Studie.* Springer, Berlin. Heidelberg, (1987).

[18] Chen, J.-R.: The Lattice Points in a Circle. Sci. Sinica, **12**:633–649, (1963).

[19] Christoffel, E.B.: Über die Gaussische Quadratur und eine Verallgemeinerung derselben. J. Reine Angew. Math., **55**:61–82. (Ges. Math. Abhandlungen I, 65–87), (1858).

[20] Cui J., Freeden W.: Equidistribution on the Sphere. SIAM J. Sci. Stat. Comp., **18**:595–609, (1997).

[21] Davis, C.H.: Theory of the Motion of Heavenly Bodies Moving about the Sun in Conic Sections, translation of "Theoria motus corporum coelestium in sectionibus conicis solem ambientium" by C.F. Gauss, reprinted Dover, New York, (1963).

[22] Davis, C.H., Robinson, D.W.: Generalized Inverses of Morphisms, Linear Algebra and Its Applications, **5**:319–328, (1972).

[23] Davis, P.J.: *Interpolation and Approximation.* Blaisdell, New York, (1963).

[24] Davis, P.J., Rabinowitz, P.: *Numerical Integration.* Blaisdell, Toronto, London, (1967).

[25] Dedekind, R.: Gauß in seiner Vorlesung über die Methode der kleinsten Quadrate. Festschrift zur Feier des hundertfünfzigjährigen Bestehens der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Weidmannsche Buchhandlung, (1901).

[26] Driscoll, J.R., Healy, R.M.: Computing Fourier Transforms and Convolutions on the 2-Sphere. Advances in Applied Mathematics, **15**:202–250, (1994).

[27] Dunnington, G.W.: Carl Friedrich Gauss: Titan of Science A Study of his Life and Work. Exposition Press, New York, (1955).

[28] Edgeworth, F.Y.: On Observations Relating to Several Quantities. Hermathena, **6**:279–285, (1887).

[29] Engl, H.: Discrepancy Principles for Tikhonov Regularization of Ill-posed Problems, Leading to Optimal Convergence Rates. J. Optim. Theory Appl., **52**:209–215, (1987).

[30] Engl, H.: *Integralgleichungen.* Springer, Berlin, (1997).

[31] Eisenhart, C., Carl Friedrich Gauss. International Encyclopedia of Social Sciences, New York, Vol. VI, 74–81, (1986)

[32] Euclid: *The Thirteen Books of The Elements.* Vol. 1. 2nd Edition, translated with introduction and commentary by Sir Thomas L. Heath. Dover Publication, New York, (1956).

[33] Euler, L.: Methodus universalis serierum convergentium summas quam proxime inveniendi, Commentarii Academiae Scientiarum Petropolitanae. **8**:3–9, Opera Omnia (XIV), 101–107, (1736).

[34] Fredholm I.: Sur une classe d'équations functionelles. Acta Mathematica **27**:365–390, (1903).

[35] Freeden, W.: Über eine Verallgemeinerung der Hardy–Landauschen Identität. Manuscr. Math., **24**:205–216, (1978).

[36] Freeden, W.: Über die Gaußsche Methode zur angenäherten Berechnung von Integralen. Math. Method. Appl. Sci., **2**:397–409, (1980).

[37] Freeden, W.: On the Approximation of External Gravitational Potential with Closed Systems of (Trial) Functions. Bull. Géod., **54**, 1–20, (1980a).

[38] Freeden, W.: On Integral Formulas of the (Unit) Sphere and Their Application to Numerical Computation of Integrals. *Computing*, **25**:131–146, 1980b.

[39] Freeden, W.: On Spherical Spline Interpolation and Approximation. Math. Meth. in the Appl. Sci., **3**:551–575, (1981a).

[40] Freeden, W.: On Approximation by Harmonic Splines. Manuscr. Geod., **6**:193–244, (1981b).

[41] Freeden, W.: Least Squares Approximation by Linear Combinations of (Multi)-Poles. Report **344**, Department of Geodetic Science and Surveying, The Ohio State University, Columbus, (1983).

[42] Freeden, W.: Spherical Spline Interpolation: Basic Theory and Computational Aspects. J. Comput. Appl. Math., **11**:367–375, (1984).

[43] Freeden, W.: A Spline Interpolation Method for Solving Boundary Value Problems of Potential Theory From Discretely Given Data. Math. Part. Diff. Equations, **3**:375–398, (1987).

[44] Freeden, W.: Interpolation by Multidimensional Periodic Splines. J. Approx. Theory **55**:104–117, (1988).

[45] Freeden, W.: *Multiscale Modelling of Spaceborne Geodata*. B.G. Teubner, Stuttgart, Leipzig, (1999).

[46] Freeden, W., Schreiner, M.: Spherical Functions of Mathematical Geosciences – A Scalar, Vectorial, and Tensorial Setup. Springer, Heidelberg, (2009).

[47] Freeden, W.: *Metaharmonic Lattice Point Theory*. CRC Press, Taylor & Francis Group, Boca Raton, (2011).

[48] Freeden, W.: Geomathematics: Its Role, Its Aim, and Its Potential. in: Handbook of Geomathematics, Vol. 1, 2nd ed., W. Freeden, M.Z. Nashed, T. Sonar, Eds., 3–79, Springer, New York, Berlin, Heidelberg, (2015).

[49] Freeden, W., Gerhards, C.: *Geomathematically Oriented Potential Theory*. Chapman and Hall, CRC Press, Boca Raton, London, New York, (2013).

[50] Freeden, W., Gutting, M.: *Special Functions of Mathematical (Geo)Physics*. Birkhäuser, Basel, (2013).

[51] Freeden, W., Kersten, H.: *The Geodetic Boundary-Value Problem Using the Known Surface of the Earth*. Veröff Geod Inst RWTH Aachen, **29**, (1980).

[52] Freeden, W., Kersten, H.: A Constructive Approximation Theorem for the Oblique Derivative Problem in Potential Theory. Math. Meth. Appl. Sci., **4**:104–114, (1981).

[53] Freeden, W., Kersten, H.: An Extended Version of Runge's Theorem. *Manuscr. Geod.*, **7**:267–278, (1982).

[54] Freeden, W., Maier, T.: On Multiscale Denoising of Spherical Functions: Basic Theory and Numerical Aspects. Electronic Transactions on Numerical Analysis (ETNA), **14**:40–62, (2002).

[55] Freeden, W., Mayer, C.: Multiscale Solution for the Molodensky Problem on Regular Telluroidal Surfaces, Manuscr. Math., Acta Geod. Geophys. Hung., **41**:55–86, (2006).

[56] Freeden, W., Michel, V.: *Multiscale Potential Theory* (*With Applications to Geoscience*). Birkhäuser, Boston, (2004).

[57] Freeden, W., Nashed, M.Z.: Multi-Variate Hardy-Type Lattice Point Summation and Shannon-Type Sampling. Int. J. Geomath., **6**:163–249, (2015).

[58] Freeden, W., Schneider, F.: Regularization Wavelets and Multiresolution. Inverse Problems, **14**:225–243, (1998).

[59] Freeden, W., Schneider, F.: Runge–Walsh Wavelet Approximation for the Helmholtz Equation. J. Math. Anal. Appl., **235**, 533–566, (1999).

[60] Freeden, W., Schreiner, M.: Orthogonal and Non-Orthogonal Multiresolution Analysis, Scale Discrete and Exact Fully Discrete Wavelet Transform on the Sphere. Const. Approx., **14**:493–515, (1998).

[61] Freeden, W., Schreiner, M.: Local Multiscale Modelling of Geoid Undulations from Deflections of the Vertical. J. Geod., **79**:641–651, (2006).

[62] Freeden, W. and Witte, B.: A Combined (Spline-)Interpolation and Smoothing Method for the Determination of the Gravitational Potential from Heterogeneous Data. Bull. Geod., **56**:53–62, (1982).

[63] Freeden, W., Wolf, K.: Klassische Erdschwerefeldbestimmung aus der Sicht moderner Geomathematik. *Math. Semesterb.*, **56**: 53–77, (2008).

[64] Freeden, W., Gervens, T., Schreiner, M.: *Constructive Approximation on the Sphere* (*With Applications to Geomathematics*). Oxford Science Publications, Clarendon, Oxford, (1998).

[65] Freeden, W., Fehlinger, T., Klug, M., Mathar, D., Wolf, K.: Classical Globally Reflected Gravity Field Determination in Modern Locally Oriented Multiscale Framework. *J. Geodesy*, **83**:1171–1191, (2009).

[66] Fricker, F.: *Einführung in die Gitterpunktlehre*, Birkhäuser, Basel, Boston, Stuttgart, (1981).

[67] Friedrich, K.: Allgemeine für die Rechenpraxis geeignete Lösung für die Aufgaben der kleinsten Absolutsumme und der günstigsten Gewichtsverteilung, Zeitschrift für Vermessungswesen, 337–358, (1937).

[68] Friedrichs, K.O.: *Functional Analysis and Applications.* pp. 144–147, 151–153, New York University, mimeographed lecture notes.

[69] Galle, A.: Über die geodätischen Arbeiten von Gauß, I. Abschnitt: Die Methode der kleinsten Quadrate. Berlin, (1924), (Gauß Werke Bd. XI 2. Abteilung).

[70] Garland, G.D.: The Contributions of Carl Friedrich Gauss. Historia Mathematica, **6**:5–29, (1979).

[71] Gauss, C.F.: Untersuchungen über Höhere Arithmetik (Disquisitiones Arithmeticae (Latin)), Gerhard Fleischer jun., Leipzig, auch: Gauss Werke, Band 1, (1801).

[72] Gauss, C.F.: Theoria motus corporum coelestium in sectionibus conicis solem ambientium, Hamburg, (1809), Werke 7. Translated into English by C.H. Davis, (1963).

[73] Gauss, C.F.: Disquisitio de elementis ellipticis Paladis ex opositionibus annorum 1803, 1804, 1805, 1807, 1808, 1809, Göttingen, (1810).

[74] Gauss, C.F.: Disquisitiones generales circa seriem infinitam $1+((\alpha\cdot\beta)/\,(1-\gamma))x+\cdots$ etc., “Commentationes Societatis Regiae Scientiarium Gottingensis Recentiores”, 2. (Werke Ill, 123–162], (1812).

[75] Gauss, C.F.: Theoria attractionis corporum sphaeroidicorum ellipticorum homogeneorum methodo nova tractata, Commentationes societatis regiae scientiarium Gottingensis recentiores, 2, 355–378, (1813).

[76] Gauss, C.F.: Methodus nova integralium valores per approximationem inveniendi. Commentationes Societatis Regiae Scientiarium Gottingensis Recentiores 3, (Werke III, 163–196), (1814).

[77] Gauss, C.F.: Theoria combinationis observationum erroribus minimis obnoxiae, 1. Teil, Göttingen, (1821).

[78] Gauss, C.F.: Allgemeine Auflösung der Aufgabe: Die Teile einer gegebenen Fläche auf einer andern gegebenen Fläche so abzubilden, dass die Abbildung dem Abgebildeten in den kleinsten Theilen ähnlich wird. In H.C. Schumacher (Hrsg.): Astronomische Abhandlungen, Heft 3, Altona 1825, S. 5–30, (1822).

[79] Gauss, C.F.: Theoria combinationis observationum erroribus minimis obnoxiae. 2. Teil, Göttingen, (1823).

[80] Gauss, C.F.: Supplementum theoriae combinationis observationum erroribus minimis obnoxiae. Göttingen, (1826).

[81] Gauss, C.F.: Allgemeine Theorie des Erdmagnetismus. Resultate aus den Beobachtungen des magnetischen Vereins, Göttingen, (1838).

[82] Gauss, C.F.: Werke, Vierter Band. Hrsg. Königl. Gesellschaft der Wissenschaft zu Göttingen, 109–117, B.G. Teubner, Leipzig, (1880).

[83] Gauss, C.F.: Werke, Neunter Band, Hrsg. Königl. Gesellschaft der Wissenschaften zu Göttingen, B.G. Teubner, Leipzig, (1903).

[84] Gauss, C.F.: Untersuchungen über Gegenstände der Höheren Geodäsie. Hrsg. J. Frischauf, Ostwald’s Klassiker der exakten Wissenschaften, Nr. 177, Leipzig, (1910).

[85] Gautschi, W.: A Survey of Gauss–Christoffel Quadrature Formulas. In: E.B. Christoffel: The Influence of His Work on Mathematics and the Physical Sciences. P.L. Butzer and F. Feher, Eds., Birkhäuser, Basel, (1981).

[86] Gerling, C.L.: Ausgleichungsrechnungen der praktischen Geometrie oder die Methode der kleinsten Quadrate. Hamburg und Gotha, (1843).

[87] Gnedenko, B.W.: Über die Arbeiten von C.F. Gauß zur Wahrscheinlichkeitsrechnung, 193–204, in: C.F. Gauß Leben und Werk, Hrsg. H. Reichardt, Haude & Spenersche Verlagsbuchhandlung, Berlin, (1960).

[88] Goldstine, H.H.: A History of Numerical Analysis from the 16th Through the 19th Century. Springer-Verlag, New York, (1977).

[89] Grafarend, C.F., Klapp, M., Martinec, M.: Spacetime Modelling of the Earths Gravity Field by Ellipsoidal Harmonics. Handbook of Geomathematics, W. Freeden, Z. Nashed, T. Sonar, Eds., Springer, Heidelberg, 2nd Edition, 381–496, (2015).

[90] Gray, J: *Ideas of Space. Euclidean, Non-Euclidean, and Relativistic.* Clarendon Press, Oxford, (1979).

[91] Green, G.: *An Essay on the Application of Mathematical Analysis to the Theories of Electricity and Magnetism.* T. Wheelhouse, Nottingham, England, (1838).

[92] Groetsch, C.W.: *Generalized Inverses of Linear Operators.* Marcel Dekker. Inc., New York, (1977).

[93] Groetsch, C.W.: *Inverse Problems in the Mathematical Science.* Vieweg, Braunschweig, (1993).

[94] Groten, E.: *Geodesy and the Earth's Gravity Field* I + II. Dümmler, Bonn, (1979).

[95] Gutting, M.: *Fast Multipole Methods for Oblique Derivative Problems.* Ph.D.-thesis, Geomathematics Group, University of Kaiserslautern, (2007).

[96] Gutting, M.: *Fast Multipole Accelerated Solution of the Oblique Derivative Boundary Value Problem.* GEM Int. J. Geomath., **3**:223–252, (2012).

[97] Hammer, E.: Beitrag zur Geschichte der Ausgleichungsrechnung. *Zeitschrift für Vermessungswesen*, 613–628, (1900).

[98] Hampel, F.R., Ronschetti, E.M, Rousseeuw, P.R., Stahel, W.A.: *Robust Statistics.* Wiley, New York, (1986).

[99] Hansen, P.A.: Von der Methode der kleinsten Quadrate im Allgemeinen und in ihrer Anwendung auf die Geodäsie. *Abhandlungen der Mathematisch-Physikalischen Klasse der Königlich-Sächsischen Gesellschaft der Wissenschaften, Band* 8, 571–806, S. Hirzel, Leipzig, (1867).

[100] Hansen, P.A: Kurz gefaßte rationelle Ableitung des Ausgleichsverfahrens eines Dreiecksnetzes nach der Methode der kleinsten Quadrate. Berichte der Mathematisch-Physikalischen Klasse der Königlich-Sächsischen Gesellschaft der Wissenschaften, Band 20, 129–150, S. Hirzel, Leipzig, (1868).

[101] Hansen, P.A: Fortgesetzte geodätische Untersuchungen bestehend aus zehn Supplementen zur Abhandlung von der Methode der kleinsten Quadrate. *Abhandlungen der Mathematisch-Physikalischen Klasse der Königlich-Sächsischen Gesellschaft der Wissenschaften, Band* 9, 1–185, S. Hirzel, Leipzig, (1871).

[102] Hansen, P.A.: Entwicklung eines neuen veränderten Verfahrens zur Ausgleichung eine Dreiecksnetzes mit besonderer Beachtung des Falles, in welchem gewisse Winkel im voraus bestimmte Werte erhalten sollen. *Abhandlungen der Mathematisch-Physikalischen Klasse der Königlich-Sächsischen Gesellschaft der Wissenschaften, Band* 9, 185–289, S. Hirzel, Leipzig, (1871).

[103] Hardy, G.H.: On the Expression of a Number as the Sum of Two Squares. Quart. J. Math. (Oxford), **46**:263–283, (1915).

[104] Harter, H.L.: Nonuniqueness of Least Absolute Values Regression. Commun. Statist. Theor. Meth., **A6** (9):829–838, (1977).

[105] Heck, B.: *Rechenverfahren und Auswertemodelle der Landesvermessung Klassische und moderne Methoden.* 3. Auflage, Wichmann Verlag, Heidelberg, (2003).

[106] Heine, E.: *Anwendungen der Kugelfunctionen und der verwandten Functionen.* 2nd Edition, Reimer, Berlin. (I. Theil: Mechanische Quadratur, 1–31), (1881).

[107] Heiskanen, W.A., Moritz, H.: *Physical Geodesy.* Freeman, San Francisco, (1967).

[108] Helmert, F.R.: Lotabweichungen. Heft 1, Veröff. d. Kgl. Preuss. Geod. Inst., Berlin, (1886).

[109] Heun, K.: Neue Methode zur approximativen Integration der Differentialgleichungen einer unabhängigen Veränderlichen. Z. Math. Phys., **45**:23–38, (1900).

[110] Heuser, H.: *Funktionalanalysis.* 4. Auflage, Teubner, (1975).

[111] Hilbert, D.: *Grundzüge einer allgemeinen Theorie der linearen Integralgleichungen.* Teubner, Leipzig, (1912).

[112] Hofmann-Wellenhof, B., Moritz, H.: *Physical Geodesy.* Springer, Wien, New York, 2nd Edition, (2006).

[113] Hörmander, L.: The Boundary Problems of Physical Geodesy. The Royal Institute of Technology, Division of Geodesy, Report 9, Stockholm, (1975).

[114] Hotine, M.: *Mathematical Geodesy.* ESSA Monography 2, U.S. Department of Commerce, Washington (Reprint 1992 by Springer), (1969).

[115] Huber, P.J.: *Robust Statistics.* Wiley, New York, (1981).

[116] Hurwitz, W.A.: On the Pseudo-Resolvent to the Kernel of an Integral Equation Transactions of the American Mathematical Society **13**: 405–418, (1912).

[117] Huxley, M.N.: Exponential Sums and Lattice Points III. Proc. London Math., **87**:591–609, (2003).

[118] Jordan, W., Eggert, O.: *Handbuch der Vermessungskunde.* Erster Band, Ausgleichungsrechnung nach der Methode der kleinsten Quadrate, 9. Aufl., J.B. Metzler, Stuttgart, (1948).

[119] Kanani, E.: *Robust Estimators for Geodetic Transformations and GIS.* Ph.D.-thesis, ETH No. 13521, Institut für Geodäsie und Photogrammetrie, Eidg. Technische Hochschule Zürich, Mitteilung Nr. 70, Zürich, (2000).

[120] Kantorowitsch, L.W., Akilow, G.P.: *Funktionalanalysis in normierten Räumen.* Akademie-Verlag, Berlin, (1964).

[121] Kellogg, O.D.: *Foundations of Potential Theory.* Frederick Ungar Publishing Company, New York, (1929).

[122] Kertscher, D.: Carl Friedrich Gauss – "Genie, Gigant, Titan". Z. f. Vermessungswesen, 130. Jg., **6**:372–378, (2005).

[123] Koch, K.R.: *Parameterschätzung und Hypothesentests.* Dümmler Verlag, Bonn (1997).

[124] Koch, K.R.: *Parameter Estimation and Hyothesis Testing in Linear Models.* 2nd ed., Springer, Berlin, (1999).

[125] Koch, K.R., Pope, A.J.: Uniqueness and Existence for the Geodetic Boundary Value Problem Using the Known Surface of the Earth. Bulletin Géodésique, **106**:467–476, (1972).

[126] Koch, K.R., Witte, B.: The Earth's Gravity Field Represented by a Simple Layer Potential From Doppler Tracking of Satellites. NOAA Technical Memorandum NOS 9, National Oceanic and Atmospheric Administration, National Ocean Service, Rockville, Md., (1971).

[127] Koch, K.R., Witte, B.: The Earth's Gravity Field Represented by a Simple Layer Potential From Doppler Tracking of Satellites. Journal of Geophysical Research, 76(35): 8471–8479, The American Geophysical Union, (1971).

[128] Kolesnik, G.: On the Method of Exponent Pairs. Acta Arith., **45**: 115–143, (1985).

[129] Kowalewski, A.: *Newton, Cotes, Gauss, Jacobi: Vier grundlegende Abhandlungen über Interpolation und genäherte Quadratur.* von Veit, Leipzig, (1917).

[130] Krarup, T.: *A Contribution to the Mathematical Foundation of Physical Geodesy.* Danish Geodetic Institute, Report No. 44, Copenhagen, (1969).

[131] Krarup, T.: *Letters on Molodensky's Problem* I–IV. Communication to the members of the IAG Special Study Group 4.31, (1973).

[132] Krüger, L.: Über die Ausgleichung von bedingten Beobachtungen in zwei Gruppen. Veröff. d. Kgl. Preu. Inst., N. F., Nr. 18, Potsdam, (1905).

[133] Kutta, W.: Beitrag zur näherungsweisen Integration totaler Differentialgleichungen. Z. Math. Phys., **46**: 435–453, (1901).

[134] Landau, E., Über die Gitterpunkte in einem Kreis (Erste Mitteilung). Gött. Nachr., 148–160, (1915).

[135] Landau, E.: Über die Gitterpunkte in einem Kreis IV. Gött. Nachr., 58–65, (1924).

[136] Laplace, P.S.: Traité de mécanique céleste. tome 2, Paris, (1799).

[137] Laplace, P.S.: Théorie analytique des probabilités. Livre II, Chap. IV. Paris, (1812).

[138] Laplace, P.S.: Théorie analytique des probabilités. Œuvres, tome VII, p. 353, (1812).

[139] Legendre A.M.: Nouvelles méthodes pour la détermination des orbites des comètes. Paris, (1806).

[140] Legendre, A.M.: Analyse des triangles tracés sur la surface d'un sphéroide. Tome VII de la I Série des mémoires de l'Académie des Sciences, Paris, 131, (1806).

[141] Lehmann, G.: Gauss theoretische Arbeiten. In C.F. Gauss und die Landesvermessung in Niedersachsen, herausgegeben von der Niedersächsischen Vermessungs- und Katasterverwaltung, (1955).

[142] Lehmann, R., Lösler, M.: Multiple Outlier Detection: Hypothesis Tests versus Model Selection by Information Criteria. Journal of Surveying Engineering, **142**: DOI: http://dx.doi.org/10.1061/(ASCE)SU.1943-5428.0000189, (2016).

[143] Listing J.B.: Über unsere jetzige Kenntnis der Gestalt und Größe der Erde. Dietrichsche Verlagsbuchhandlung, Göttingen, (1873).

[144] Magnus, W., Oberhettinger, F., Soni, R.P.: *Formulas and Theorems for the Special Functions of Mathematical Physics.* In: Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen, Band 52, Springer, Berlin, 3. Auflage, (1966).

[145] Maire, Ch., Boscović, R.J.: De Litteraria Expeditione per Pontificiam Ditionem ad Dimentiendos duos Meridiani Gradus, Rom, 409–503, (1755).

[146] Markina, I., Vodopyanov, S.K.: Fundamentals of the Nonlinear Potential Theory for Subelliptic Equations. I, II. Siberian Adv. Math., **7**:32–62, (1997).

[147] Marx, C.: Outlier Detection by Means of Monte Carlo Estimation Including Resistant Scale Estimation. J. of Applied Geodesy, **9**:123–141, (2015).

[148] McEwen, J.D., Puy, G., Thiran, J.-P., Vandergheinst, P.: Sampling Theorems and Compressive Sensing on the Sphere. Wavelets and Sparsity XIV, Papadakis, M., van de Ville, D., Goyal, V.K., Eds., Proc. of SPIE, Vol. 8138, F1–F9, (2011).

[149] Mehler, F.G.: Bemerkungen zur Theorie der mechanischen Quadraturen. J. Reine Angew. Math., **63**:152–157, (1864).

[150] Meissl, P.: *On the Linearization of the Geodetic Boundary Value Problem.* Department of Geodetic Science, No. 152, The Ohio State University, Columbus, OH, (1971).

[151] Molodensky, M.S., Eremeev, V.F., Yurkina, M.I.: *Methods for Study of the External Gravitational Field and Figure of the Earth.* Trudy TSNIIGAiK, Geodezizdat,

Moscow, 131, (1960), English translat.: Israel Program for Scientific Translation, Jerusalem, (1962).

[152] Moore, E.H.: On the Reciprocal of the General Algebraic Matrix. Bulletin of the American Mathematical Society **26**:394–395, (1920).

[153] Moore, E.H.: General Analysis. Memoirs of the American Mathematical Society, **1**:197–209, (1935).

[154] Moritz, H.: Statistische Methoden in der gravimetrischen Geodäsie. Zeitschrift für Vermessungswesen, 409–416, (1963).

[155] Moritz, H.: Der Begriff der mathematischen Erdgestalt seit Gauss. Allgemeine Vermessungs-Nachrichten, 133–138, (1977).

[156] Moritz, H.:*Advanced Physical Geodesy*. Wichmann Verlag, Karlsruhe, (1980).

[157] Moritz, H.: Geodesy and Mathematics. Zeszyty Naukowe Akademii Görniczo-Hutniczej IM. Stanislawa Staszica, No. 780, Geodezja, 63, Kraków, 38–43, (1981).

[158] Moritz, H.: *Advanced Physical Geodesy*, Wichmann, Karlsruhe, 2nd Edition, (1989).

[159] Moritz, H.: Classical Physical Geodesy. Handbook of Geomathematics, W. Freeden, Z. Nashed, T. Sonar, Eds., Springer, Heidelberg, 2nd Edition, 253–290, (2015).

[160] Müller, C.: Eine Verallgemeinerung der Eulerschen Summenformel und ihre Anwendung auf Fragen der analytischen Zahlentheorie. Abh. Math. Sem. Univ. Hamburg, **19**:41–61, (1954).

[161] Müller, C., *Analysis of Spherical Symmetries in Euclidean Spaces*. Springer, New York, Berlin, Heidelberg, (1998).

[162] Nashed, M.Z.: Generalized Inverses, Normal Solvability and Iteration for Singular Operator Equations. In: *Nonlinear Functional Analysis and Applications*, L.B. Rall, ed., Academic, New York, 311–359, (1971).

[163] Nashed, M.Z.: Aspects of Generalized Inverses in Analysis and Regularization. In: Generalized Inverses and Applications. Academic Press, New York, 193–244, (1976).

[164] Nashed, M.Z.: *Generalized Inverses and Applications*. Academic Press, New York, (1976).

[165] Nashed, M.Z.: Operator-Theoretic and Computational Approaches to Ill-Posed Problems with Applications to Antenna Theory. IEEE Transactions on Antennas and Propagation, **29**:220–231, (1981).

[166] Nashed, M.Z.: Inner, Outer, and Generalized Inverses in Banach and Hilbert Spaces. Numer. Funct. Anal. Optim., **9**:261–326, (1987).

[167] Nashed, M.Z., Walter, G.G.: General Sampling Theorems for Functions in Reproducing Kernel Hilbert Spaces. Math. Control Signals Systems, **4**:363–390, (1991).

[168] Neumann, F., *Vorlesungen über die Theorie des Potentials und der Kugelfunktionen*, Teubner, Leipzig, 135–154, (1887).

[169] Neumann, J.V.: On Regular Rings. Proceedings of the National Academy of Science of the United States of America, **22**:707–713, (1936).

[170] Nutz, H.: *A Unified Setup of Gravitational Field Observables*. Ph.D.-thesis, Geomathematics Group, University of Kaiserslautern, Shaker, Aachen, (2002).

[171] Palczewski, A., Schneider, J., Bobylev, A.: A Consistency Result for a Discrete Velocity Model of the Boltzmann Equation. SIAM J. Numer. Anal., **34**:1865–1883, (1997).

[172] Penrose, R.: A Generalized Inverse for Matrices. Proceedings of the Cambridge Philosophical Society 51, 406–413, (1955).

[173] Petrini, H.: Sur l'existence des dérivées secondes du potentiel. C.R. Acad. Sci. Paris, **130**:233–235, (1900).

[174] Pizzetti, P.: Sopra il Calcoba Tesrico delle Deviazioni del Geoide dall'Ellissoide. *Att. R Accad. Sci. Torino*, **46**:331–350, (1910).

[175] Plackett, R.L.: A Historical Note on the Method of Least Squares. Biometrika, **36**:458–460, (1949).

[176] Proclus: *A Commentary on the First Book of Euclid's Elements*. Translated with Introduction and Notes, by Glenn R. Morrow. Princeton University Press, (1970).

[177] Radau, R.: Étude sur les formules d'approximation qui servent à calculer la valeur numérique d'une intégrale définie. J. Math. Pures Appl., **3**(6):283–336, (1880).

[178] Rado, R.: Note on Generalized Inverses of Matrices. Proceedings of the Cambridge Philosophical Society **52**:600–601, (1956).

[179] Reich, K., Roussanova, E.: Visualising Geomagnetic Data by Means of Corresponding Observations. GEM International J. Geomath., **4**:1–25, (2013).

[180] Reid, W.T.: Generalized Inverses of Differential and Integral Operators. Theory and Applications of Generalized Inverses of Matrices (T.L. Boullion and P.L. Odell 1–25). Symposium Proceedings, Texas Tech University Mathematics Series, Vol. 4., Lubbock; Texas Tech Press, (1965).

[181] Reinsch, C.: Smoothing by Spline Functions. Numer. Math., **10**:177–183, (1967).

[182] Reuter, R.: *Über Integralformeln der Einheitssphäre und harmonische Splinefunktionen*. Ph.D.-thesis, Veröff. Geod. Inst. RWTH Aachen, Report No. 33, (1982).

[183] Robinson, D.W.: Gauss and Generalized Inverses. Historia Mathematics, **7**:118–125, (1980).

[184] Rosenfeld, B.A.: *A History of Non-Euclidean Geometry – Evolution of the Concept of a Geometric Space*. Springer Science+Business Media, (1988).

[185] Rummel, R.: Geodesy. Encyclopedia of Earth System Science, W.A. Nierenberg, ed., Academic Press, **2**:253–262, (1992).

[186] Rummel, R.: Spherical Spectral Properties of the Earth's Gravitational Potential and Its First and Second Derivatives. In: Lecture Notes in Earth Science, Springer, Berlin, Vol. **65**:359–404, (1997).

[187] Runge, C.: Zur Theorie der eindeutigen analytischen Funktionen, *Acta Math.* 6, 229–234, (1885).

[188] Runge, C., Willers, F.A.: Numerische und graphische Integration. Encyklopädie der Math. Wiss., **2**(3):47–176, Leipzig, (1915).

[189] Sanso, F.: A Note on Density Problems and the Runge Krarup's Theorem. Bolletino di Geodesia e Science Affini **41**, 422–477, (1982).

[190] Sard, A.: Integral Representations of Remainders. Duke Math. J. **15**:333–345, (1948).

[191] Schauder, J.: Potentialtheoretische Untersuchungen, Math. Z., **35**:536–538, (1931).

[192] Schlossmacher, E.J.: An Iterative Technique for Absolute Deviations Curve Fitting. J. Americ. Statistical Association, **68**:857–859, (1973).

[193] Schmidt; H.: Lösung der geodätischen Hauptaufgaben auf dem Rotationsellipsoid mittels numerischer Integration. Zeitschrift für Vermessungswesen **123**:121–128, (1999).

[194] Schreiber, O.: Anordnung der Winkelbeobachtungen im Göttinger Basisnetz. Zeitschrift für Vermessungswesen, 129–161, (1882).

[195] Schreiber, O.: *Die konforme Doppelprojektion der trigonometrischen Abteilung der Königlich Preussischen Landesaufnahme.* Formeln und Tafeln. Selbstverlag, Berlin, (1897).

[196] Schreiber, O.: Formeln und Tafeln zur Berechnung der geographischen Coordinaten aus Richtungen und Längen der Dreiecksseiten, Königl. Preuss. Landesaufn., Berlin, (1878).

[197] Schreiner, M.: *Wavelet Approximation by Spherical Up Functions.* Habilitation Thesis, Geomathematics Group, University of Kaiserslautern, Shaker, Aachen, (2004).

[198] Shannon, C.E.: Communication in the Presence of Noise. Proc. Institute of Radio Engineers, **37**:10–21, (1949).

[199] Shure, L., Parker, R.L., Backus, G.E.: Harmonic Splines for Geomagnetic Modelling. Phys. Earth Planet. Inter. **28**:215–229, (1982).

[200] Siegel, C.L.: Über die analytische Theorie der quadratischen Formen, III. Annals of Mathematics **38**:212–291, (1937).

[201] Sierpinski, W.: O pewnem zagadnieniu z rachunku funckcyj asmptotycznych (Über ein Problem des Kalküls der asymptotischen Funktion (polnisch)), Prace Math.-Fiz., **17**:77–118, (1906).

[202] Simons, F.J., Plattner, A.: Scalar and Vector Slepian Functions, Spherical Signal Estimation and Spectral Analysis. Handbook of Geomathematics (W. Freeden, M.Z. Nashed, T. Sonar, Eds.), 2nd Edition, Vol. 3, 2563–2608, (2015).

[203] Slepian, D.: Some Comments on Fourier Analysis, Uncertainty and Modeling. SIAM Rev., **25**(3):379–393, (1983).

[204] Sonar, T., 3000 *Jahre Analysis.* Springer, Berlin, Heidelberg, New York, (2011).

[205] Stewart G.W.: Theoria Combinationis Observationum Erroribus Minimis Obnoxiae. Pars Prior et Pars Posterior and Supplementum, Theory of the Combination of Observations Least subject to Errors, Part One, Part Two, Supplement by C.F. Gauß, Society for Industrial and Applied Mathematics, (1995).

[206] Stoer, J., *Numerische Mathematik* 1. 5. Auflage, Springer-Verlag, Berlin, Heidelberg, New York, (1989).

[207] Stokes, G.G.: On the Variation of Gravity on the Surface of the Earth. Trans. Camb. Phil. Soc., **8**:672–695, (1849).

[208] Svensson, S.L.: Pseudodifferential Operators. A New Approach to the Boundary Value Problems of Physical Geodesy. Manuscr. Geod., **8**:1–40, (1983).

[209] Szegö G.: *Orthogonal Polynomials.* American Mathematical Society Colloquium Publications, 4th edition, Volume 23, American Mathematical Society, Providence, Rhode Island, (1975).

[210] Taylor, A.E., Lay, D.: *Functional Analysis.* Wiley, New York, (1979).

[211] Tikhonov, A.N.: On the Stability of Inverse Problems. *Dokl. Akad. Nauk. SSSR*, **39**:195–198, (1943).

[212] Tikhonov, A.N.: On the Solution of Incorrectly Formulated Problems and the Regularization Method, *Dokl. Akad Nauk SSSR*, **151**:501–504, (1963).

[213] Tikhonov, A.N.: On Methods of Solving Incorrect Problems, *Amer. Math. Soc. Transl.*, **2**, 70: 222–224, (1968).

[214] Tikhonov, A.N., Arsenin V.Ya.: *Solutions of Ill-Posed Problems.* Wiley, New York, (1977).

[215] Tikhonov, A.N., Goncharsky, A.V., Stepanov, V.V., Yagola, A.G.: *Numerical Methods for the Solution of Ill-Posed Problems.* Kluwer, Dordrecht, (1995).

[216] Todhunter, I., A History of the Mathematical Theories of Attraction and of the Figure on the Earth from the Time of Newton to that of Laplace, Macmillan, London (reprint 1962 by Dover Publications, New York), (1878).

[217] Vening Meinesz, F.A.: A Formula Expressing the Deflection of the Plumb-Line in the Gravity Anomalies and some Formulae for the Gravity Field and the Gravity Potential Outside the Geoid. Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, **31**(3):315–331, (1928).

[218] Vekua, I.N.: Über die Vollständigkeit des Systems harmonischer Polynome im Raum. *Dokl. Akad. Nauk* 90, 495–498, (1953).

[219] Vogler, C.A.: Lehrbuch der praktischen Geometrie. Bd. II 1. Halbband, Braunschweig, (1894).

[220] Voronoi, G.: Sur un problème du calcul des fonctions asymptotiques. JRAM, **126**:241–282, (1903).

[221] Wahba, G.: *Spline Models for Observational Data.* SIAM Philadelphia, (1990).

[222] Walsh, J.L.: The Approximation of Harmonic Functions by Harmonic Polynomials and by Harmonic Rational Functions, *Bull. Amer. Math. Soc.* 35, 499–544, (1929).

[223] Wangerin, A.: *Theorie des Potentials und der Kugelfunktionen*, Walter de Gruyter & Co, Berlin, Leipzig, (1921).

[224] Weizel, R.: C.F. Gauß, Angewandte Mathematik 200 Jahre nach seiner Geburt. Zeitschrift für Vermessungswesen, 520–525, (1977).

[225] Wermer, J.: *Potential Theory.* Lecture Notes in Mathematics, 408, Springer, Berlin, (1974).

[226] Werner, J., Numerische Mathematik 1. Vieweg Studium, Braunschweig, (1991).

[227] Werner, D.: *Funktionalanalysis.* 4. Auflage; Springer, Berlin, (2002).

[228] Weyl, H.: Über die Gleichverteilung von Zahlen mod. Eins. Math. Ann., **77**:313–352, (1916).

[229] Witte, B.: Die Berechnung von Schwerestörungen im Außenraum mit Hilfe des Greenschen Satzes. Zeitschrift für Vermessungswesen (ZfV), 135–141, Stuttgart, (1969).

[230] Witte, B.: Vergleich verschiedener Methoden für die Berechnung von Schwerestörungen im Außenraum. Schweizerische Zeitschrift für Vermessung, Photogrammetrie und Kulturtechnik, 161–168, (1969).

[231] Witte, B.: Die Bestimmung von Horizontalableitungen der Schwere im Außenraum aus einer Weiterentwicklung der Stokesschen Funktion. Gerlands Beiträge zur Geophysik, 87–94, Leipzig, (1970).

[232] Witte, B.: Computational Procedures for the Determination of a Simple Layer Model of the Geopotential From Doppler Observations. NOAA Technical Report NOS-42, National Oceanic and Atmospheric Administration, National Ocean Service, Rockville, Md., (1971).

[233] Witte, B., Sparla, P.: *Vermessungskunde und Grundlagen der Statistik für das Bauwesen.* 8. Auflage, Wichmann – VDE-Verlag, Berlin, (2015).

[234] Wolf, H.: *Ausgleichungsrechnung nach der Methode der kleinsten Quadrate.* Dümmler Verlag, Bonn, (1968).

[235] Wolf, H.: *Ausgleichungsrechnung. Formeln zur praktischen Anwendung.* Dümmler Verlag, Bonn, (1975).

[236] Wolf, H.: C.F. Gauß und die Methode der kleinsten Quadrate. Allgemeine Vermessungs-Nachrichten, 125–133, (1977).

[237] Wolf, H.: Die deutsche Landesvermessung in ihrer historischen Entwicklung. Festschrift 100 Jahre Geodätische Lehre und Forschung in Hannover. Wissenschaftliche Arbeiten der Fachrichtung Vermessungswesen der Universität Hannover, Nr. 100, 47–53. Hannover, (1981).

[238] Wolf, H.: Neues Altes in der Ausgleichungsrechnung. Vermessung, Photogrammetrie, Kulturtechnik, 233–240, (1983).

[239] Wolfe, H.E.: *Introduction to Non-Euclidean Geometry.* Reprint of the first edition 1945. Dover Publications, Mineola, New York, (2012).

[240] Wußing, H.: *Carl Friedrich Gauss.* B.G. Teubner Verlag, Leipzig, (1989).

[241] Yoshida, K.: *Functional Analysis.* 5th Edition, Springer, Berlin, (1980).

Willi Freeden
Geomathematics Group
University of Kaiserslautern
MPI-Gebäude, Paul-Ehrlich-Str. 26
D-67663 Kaiserslautern, Germany
e-mail: freeden@rhrk.uni-kl.de

Thomas Sonar
Institute for Computational Mathematics
University of Brunswick
Pockelsstraße 14
D-38106 Braunschweig, Germany
e-mail: t.sonar@tu-bs.de

Bertold Witte
Institute for Geodesy and Geoinformation
University of Bonn
Nussallee 17
D-53115 Bonn, Germany
e-mail: bertold.witte@uni-bonn.de

*   &"

Pag. 147 (162) → Figure 6.4

A detail of the triangulation of Hannover, carried out under Gauss's supervision 1821-1825

(Collected Works, Vol. 9, p. 347)

⇒ Gaussian curvature

Gauss also was led to the Gaussian Curvature →

an intrinsic measure of curvature dependent only on how distances are measured on the surface, not on the way it is embedded in space.

Vajên:

Pag. 143 (158) 6.2.

Gaussian conformal mapping of the Earth's surface.

→ Strasbourg Master Class on Geometry ... (2012).pdf

+,-.+ + +/ ! + %" ! 1 ! !
2 3 9 0 ! ! ! ! 3 0#
*)) E 9 # ! 2 3
/1 , # GGG * &&

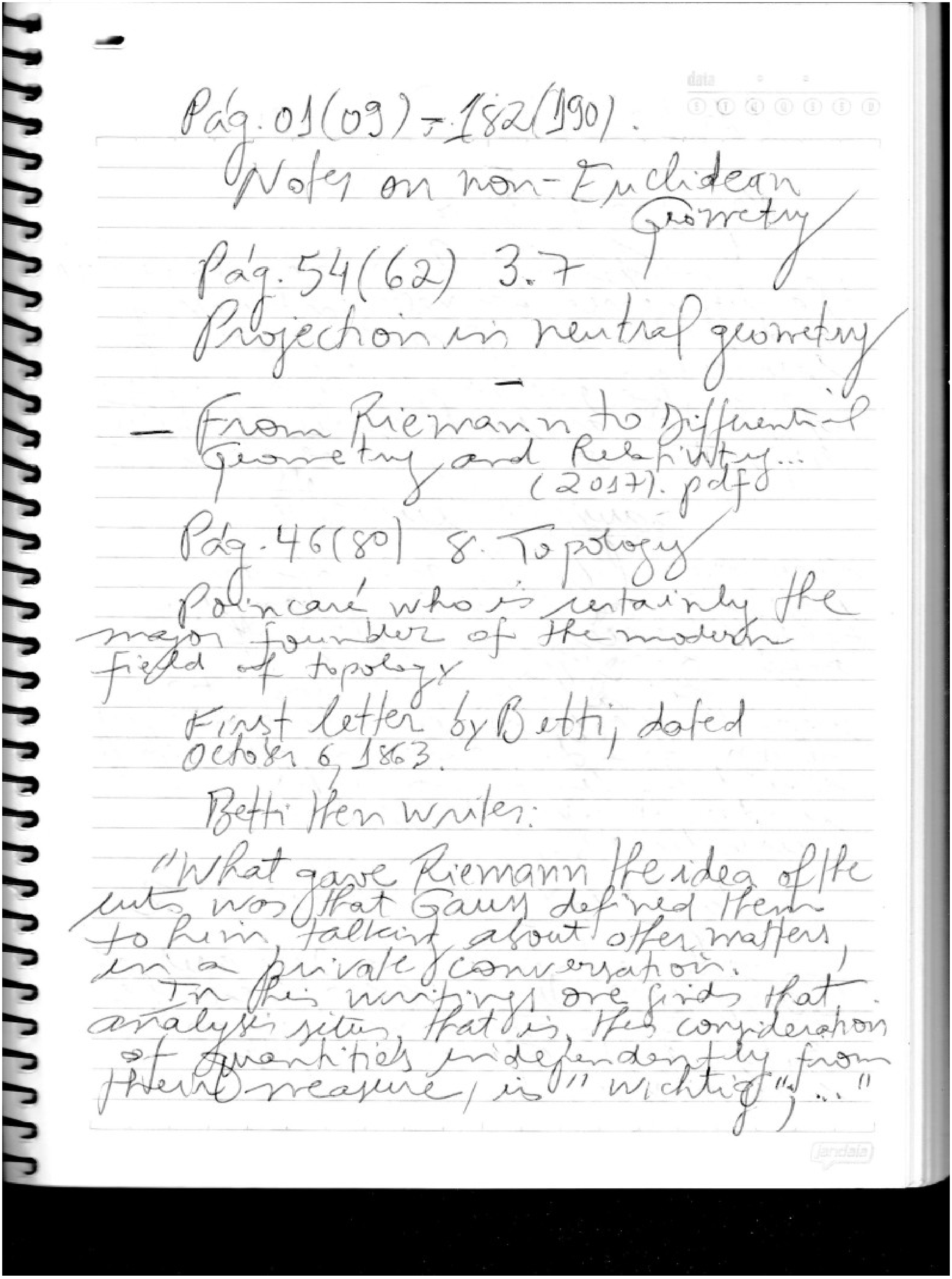

Pág. 01(09) – 182(190).

Notes on non-Euclidean geometry

Pág. 54(62) 3.7

Projection in neutral geometry

— From Riemann to Differential Geometry and Relativity... (2017). pdf

Pág. 46(80) 8. Topology

Poincaré who is certainly the major founder of the modern field of topology

First letter by Betti, dated October 6, 1863.

Betti then writes:

"What gave Riemann the idea of the cuts was that Gauss defined them to him, talking about other matters, in a private conversation.

In his writings one finds that analysis situs, that is, the consideration of quantities independently from their measure, is " wichtig"..."

8 +,-.+ + +/ ! + %" = - ! # !
3 ! # ! 6 3 0 0#
* "' < 0 3

of the mind." On Kant, Gauss had the advantage of being a mathematician. In a letter to his friend Bessel, dated April 9, 1830, Gauss writes (translation from [30] p. 13):

> We must confess in all humility that a number is *solely* a product of our mind. Space, on the other hand, possesses also a reality outside of our minds, the laws of which we cannot fully prescribe a priori.

In another letter, sent to Wolfgang Bolyai on March 6, 1832 and published in his *Collected Works*, Gauss writes, concerning the two hypotheses on the angle sum in a triangle, that it is precisely in the difficulty of this decision that "lies the clearest proof that Kant was wrong in asserting that space is just a form of our perception."

Gauss was also very critical of Kant's argument based on symmetries in the text we quoted above ("There being in perfectly similar and equal but incongruous solids, such as the right and the left hand, conceived of solely as to extent,…it is patent that only by pure intuition can the difference, namely, incongruity, be noticed"). We further discuss this in Chap. 6 of the present volume [190].

It is not surprising that Riemann declares, in his habilitation lecture, that, concerning his ideas on space, he is influenced by Gauss.

Riemann's ideas on space were discussed by Clifford, the first mathematician who translated into English Riemann's habilitation text, cf. [46].

8 Topology

Poincaré, who is certainly the major founder of the modern field of topology,[64] declares in his "Analysis of his own works" (*Analyse des travaux scientifiques de Henri Poincaré faite par lui-même*), [202] p. 100, that he has two predecessors in the field, namely, Riemann and Betti. The latter, in his correspondence with his friend and colleague Placido Tardy reports on several conversations he had with Riemann on topology. Two letters from Betti to Tardy on this subject are reproduced and translated in the book [206] by Pont, in the article [257] by Weil, and prior to them, by Loria in his obituary on Tardy [178].

The first of these two letters by Betti, dated October 6, 1863, starts with the following (Weil's translation): "I have newly talked with Riemann about the connectivity of spaces, and have formed an accurate idea of the matter," and he goes on explaining to his friend the notion of connectivity and that of the order of connectivity. Betti then writes:

> What gave Riemann the idea of the cuts was that Gauss defined them to him, talking about other matters, in a private conversation. In his writings one finds that analysis situs, that is, this consideration of quantities independently from their measure, is "wichtig"; in the last

---

[64] We may quote P. S. Alexandrov, who declared in a talk he gave at a celebration of the centenary of Poincaré's birth [9]: "To the question of what is Poincaré's relationship to topology, one can reply in a single sentence: he created it." On Poincaré and Riemann, Alexandrov, in the same talk, says the following: "The close connection of the theory of functions of a complex variable, which Riemann has observed in embryonic form, was first understood in all its depth by Poincaré."

> years of his life he has been much concerned with a problem in analysis situs, namely: given a winding thread and knowing, at every one of its self-intersections, which part is above and which below, to find whether it can be unwound without making knots; this problem he did not succeed in solving except in special cases ...

The second letter, dated October 16, 1863, starts with: "Riemann proves quite easily that every space can be reduced to an SC space by means of 1-cuts and SC 2-cuts." In the same letter, Betti elaborates on this subject, giving many examples in n dimensions. He concludes the letter by noting that the number of line sections is equal to the number of periodicity moduli of an $(n-1)$-integral, the number of simply connected surface sections to the number of periodicity moduli of an $(n-2)$-integral, and so on.

This should make clear the parentage, for what concerns topology, from Riemann to Poincaré, potentially including Betti. In this section, we go further back in the history of topological ideas, and we review some of the important works done before Riemann in this field.

René Thom considers that topology was born in ancient Greece. He expanded on this idea in several articles, cf. [246, 247]. This is a perfectly reasonable theory. In fact, the question depends on what sense we give to the word "topology." If the matter concerns the notions of limit and convergence, then the roots of this field are indeed in Greek antiquity, and more especially, in the writings of Zeno, which do not survive, but which were quoted by his critics and commentators, including Plato, Aristotle and Simplicius. Likewise, if the question concerns the notion of space, and the related notion of place, then the roots also are in Greek science. We already alluded to this fact in the previous section. The Greeks made a distinction between space and place and the notion of place (situs) is at the basis of topology. The three words place, situs and $\tau\acute{o}\pi o\varsigma$ are synonyms. To the best of our knowledge, a systematic investigation of the origin of topology in Greek antiquity has never been conducted. A whole book may be written on that subject. Failing to do this now, we shall start our exposition of the roots of topology with Leibniz, as it is usually done. Indeed, it is commonly accepted that the first explicit mention of topology as a mathematical field was made by him.

Even though no purely topological result can be attributed to Leibniz, he had the privilege to express for the first time, back in the seventeenth century, the need for a new branch of mathematics, which would be "a geometry that is more general than the rigid Euclidean geometry and the analytic geometry of Descartes." Leibniz describes his geometry as purely qualitative and concerned with the study of figures independently of their metrical properties. In a letter to Christiaan Huygens, sent on September 8, 1679 (cf. [176] pp. 578–569 and [144] vol. VIII n$^{\text{o}}$ 2192), he writes:

> After all the progress I have made in these matters, I am still not happy with Algebra, because it provides neither the shortest ways nor the most beautiful constructions of Geometry. This is why when it comes to that, I think that we need another analysis which is properly geometric or linear, which expresses to us directly *situm*, in the same way as algebra expresses *magnitudinem*. And I think that I have the tools for that, and that we might represent figures

> and even engines and motion in character, in the same way as algebra represents numbers in magnitude.[65]

In the same letter ([176] p. 570), Leibniz adds:

> I found the elements of a new characteristic, completely different from Algebra and which will have great advantages for the exact and natural mental representation, although without figures, of everything that depends on the imagination. Algebra is nothing but the characteristic of undetermined numbers or magnitudes. But it does not directly express the place, angles and motions, from which it follows that it is often difficult to reduce, in a computation, what is in a figure, and that it is even more difficult to find geometrical proofs and constructions which are enough practical even when the Algebraic calculus is all done.[66]

Together with his letter to Huygens, Leibniz included the manuscript of an essay he wrote on the new subject. He writes, in the same letter ([176] p. 571):

> But since I don't see that anybody else has ever had the same thought, which makes me fear that it might be lost if I do not get enough time to complete it, I will add here an essay which seems to me important, and which will suffice at least to rendre my aim more credible and easier to conceive, so that if something prevents its realization now, it will serve as a monument for posterity and give the possibility to somebody else to finish it.[67]

He then explains in more detail his vision of this new domain of mathematics, and where it stands with respect to algebra and geometry, giving several examples of a formalism to denote loci, showing how this formalism expresses statements such that the intersection of two spherical surfaces is a circle, and the intersection of two planes is a line.

Leibniz' letter ends with the words ([174] p. 25):

---

[65] Après tous les progrès que j'ai faits en ces matières, je ne suis pas encore content de l'Algèbre, en ce qu'elle ne donne ni les plus courtes voies, ni les plus belles constructions de Géométrie. C'est pourquoi lorsqu'il s'agit de cela, je crois qu'il nous faut encore une autre analyse proprement géométrique ou linéaire, qui nous exprime directement *situm*, comme l'algèbre exprime *magnitudinem*. Et je crois d'en avoir le moyen, et qu'on pourrait représenter des figures et même des machines et mouvements en caractères, comme l'algèbre représente les nombres en grandeurs. [We have modernized the French.].

[66] J'ai trouvé quelques éléments d'une nouvelle caractéristique, tout à fait différente de l'Algèbre, et qui aura de grands avantages pour représenter à l'esprit exactement et au naturel, quoique sans figures, tout ce qui dépend de l'imagination. L'Algèbre n'est autre chose que la caractéristique des nombres indéterminés ou des grandeurs. Mais elle n'exprime pas directement la situation, les angles et les mouvements, d'où vient qu'il est souvent difficile de réduire dans un calcul ce qui est dans la figure, et qu'il est encore plus difficile de trouver des démonstrations et des constructions géométriques assez commodes lors même que le calcul d'Algèbre est tout fait.

[67] Mais comme je ne remarque pas que quelqu'autre ait jamais eu la même pensée, ce qui me fait craindre qu'elle ne se perde, si je n'y ai pas le temps de l'achever, j'ajouterai ici un essai qui me paraît considérable, et qui suffira au moins à rendre mon dessein plus croyable et plus aisé à concevoir, afin que si quelque hasard en empêche la perfection à présent, ceci serve de monument à la postérité, et donne lieu à quelque autre d'en venir à bout.

> I have only one remark to add, namely, that I see that it is possible to extend the characteristic to things which are not subject to imagination. But this is too important and it would lead us too far for me to be able to explain myself on that in a few words.[68]

When Leibniz started his correspondence with Huygens, the latter was already a well established scientist whose achievements were behind him, and it was not easy to convince him of the usefulness of a new theory. Huygens thought that the theory was too abstract and he remained skeptical about it. He was above all a geometer working on concrete geometrical problems.

One may recall that when Leibniz sent him the above letter, Huygens was considered as a world authority in geometry and physics. He was settled in Paris since 15 years, and he was a leading member of the *Académie Royale des Sciences*. Leibniz had studied mathematics with Huygens, who was seventeen years older than him, and he considered him as his mentor. Huygens responded to Leibniz in a letter dated November 22, 1679 ([176] p. 577):

> I have examined carefully what you are asking me regarding your new characteristic, and to be frank with you, I cannot not conceive the fact that you have so much expectations from what you spread on me. Because your example of places concerns only realities which were already perfectly known, and the proposition saying that the intersection of a plane and a spherical surface makes the circumference of a circle does not follow clearly. Finally, I cannot see in what way you can apply your characteristic to which you seem you want to reduce all these different matters, like the quadratures, the invention of curves by the properties of tangents, the irrational roots of equations, Diophantus' problems, the shortest and the most beautiful constructions of the geometric problems. And what still appears to me stranger than anything else, the invention and the explanation of machines. I say it to you unsuspiciously, in my opinion this is only wishful thinking, and I need other proofs in order to believe that there could be some reality in what you present. I would nevertheless restrain myself from saying that you are mistaken, knowing the subtlety and the deepness of your mind. I only beg you that the magnificence of the things you are searching won't let you postpone from giving us those which you already found, like this Arithmetic Quadrature you discovered, concerning the roots of the equations beyond the cubical, if you are still satisfied with it.[69]

---

[68] Je n'ai qu'une remarque à ajouter, c'est que je vois qu'il est possible d'étendre la caractéristique jusqu'aux choses, qui ne sont pas sujettes à l'imagination ; mais cela est trop important et va trop loin pour que je me puisse expliquer là-dessus en peu de paroles.

[69] J'ai examiné attentivement ce que vous me demandez touchant votre nouvelle caractéristique, mais pour vous l'avouer franchement, je ne conçois pas parce que vous m'en étalez, que vous y puissiez fonder de si grandes espérances. Car votre exemple des Lieux ne regarde que des vérités qui nous étaient déjà fort connues, et la proposition de ce que l'intersection d'un plan et d'une surface sphérique fait la circonférence d'un cercle, s'y conclut assez obscurément. Enfin, je ne vois point de quel biais vous pourriez appliquer votre caractéristique à toutes ces choses différentes qu'il semble que vous y vouliez réduire, comme les quadratures, l'invention des courbes par la propriétés des tangentes, les racines irrationnelles des Équations, les problèmes de Diophante, les plus courtes et plus belles constructions des problèmes géométriques. Et ce qui me paraît encore le plus étrange, l'invention et l'explication des machines. Je vous le dis ingénument, ce ne sont là à mon avis que de beaux souhaits, et il me faudrait d'autres preuves pour croire qu'il y eût de la réalité dans ce que vous avancez. Je n'ai pourtant garde de dire que vous vous abusiez, connaissant d'ailleurs la subtilité et profondeur de votre esprit. Je vous prie seulement que la grandeur des choses que vous cherchez ne vous fasse point différer de nous donner celles que vous avez déjà trouvées, comme est

In another letter dated January 11, 1680 ([176] p. 584) Huygens writes:

> For what concerns the effects of your characteristic, I see that you insist on being persuaded of them, but as you say yourself, the examples will be more important than reasonings. This is why I am asking you much simpler examples, but capable of overcoming my incredulity, because that of the places, I confess, does not seem to me of that sort.[70]

The essay that Leibniz sent did not obtain Huygens' backing and it remained hidden among other manuscripts in Huygens' estate. It was published for the first time in 1833, and drew the attention of several nineteenth-century mathematicians, including Grassmann (1809–1877), the founder of the theory of vector spaces, who realized its importance for the new field of topology. There are two recent editions of this text, both included in doctoral dissertations, by J. Acheverría [175] (1995), in France, and by de Risi, [234] (2007), in Germany. The two dissertations contain other texts by Leibniz on the same subject.

Leibniz used several names for the new field, including *analysis situs, geometria situs, characteristica situs, characteristica geometrica, analysis geometrica, speciosa situs*, etc.

The first mathematician who worked consciously on topological questions is Euler. These questions include the definition and the invariance of the Euler characteristic of a convex polyheron, the problem known as that of the Königsberg seven bridges, another question related to the Knight's tour on the chessboard, and a musical question concerning a graph known as the *speculum musicum*. This graph was introduced in Euler's *Tentamen novae theoriae musicae ex certissimis harmoniae principiis dilucide expositae* (A attempt at a new theory of music, exposed in all clearness according to the most well-founded principles of harmony) [71]. Its vertices are the twelve notes of the chromatic scale, and the edges connect two elements which differ by a fifth or a major third with the property that one may traverse all the edges of the graph passing exactly once by each note. The article [189] is a detailed survey of the work of Euler on these questions. In the present section, we start by reviewing the work of Euler on the question of the seven bridges of Königsberg. This work shows that Euler considered himself as the direct heir of Leibniz for what concerns the field of topology. We shall then describe in detail the works of Euler and Descartes on the Euler characteristic, a question which is directly related to the topological classification of surfaces, which was one of Riemann's major achievements in topology. We recall that Euler formulated this result for a surface which is the boundary of a convex polyhedron having F faces, A edges and S vertices; the formula is then:

$$F - A + S = 2.$$

(Footnote 69 continued)
cette Quadrature Arithmétique et que vous avez découvert pour les racines des équations au-delà du cube, si vous en êtes content vous-même.

[70] Pour ce qui est des effets de votre caractéristique, je vois que vous persistez à en être persuadé, mais, comme vous dites vous-même, les exemples toucheront plus que les raisonnements. C'est pourquoi je vous en demande des plus simples, mais propres à convaincre mon incrédulité, car celui des lieux, je l'avoue, ne me paraît pas de cette sorte.

We start with the problem of the Königsberg bridges.

In the eighteenth century, the city of Königsberg[71] consisted of four quarters separated by branches of the river Pregel and related by seven bridges. The famous "problem of the seven bridges of Königsberg" asks for a path in that city that starts at a given point and returns to the same point after crossing once and only once each of the seven bridges. At the time of Euler, this was a popular question among the inhabitants of Königsberg.

Euler showed that such a path does not exist. He presented his solution to the Saint Petersburg Academy of Sciences on August 26, 1735, and in the same year he wrote a memoir on the solution of a more general problem entitled *Solutio problematis ad geometriam situs pertinentis* (Solution of a problem relative to the geometry of position) [74]. Euler learned about the problem from a letter, dated March 7, 1736, sent to him by Carl Leonhard Gottlieb Ehler, one of his friends who was the mayor of Danzig[72] and who had worked as an astronomer in Berlin. Euler solved the problem just after he received the letter. In a letter dated March 13, 1736, written to Giovanni Marioni, an Italian astronomer working at the court of Vienna, Euler declares that he became interested in this question because he realized that the problem could not be solved using geometry, algebra or combinatorics, and that therefore he wondered whether "it belonged to the 'geometry of position,' (*geometria situs*) which Leibniz has so much sought for." In the same letter, Euler announced to Marioni that after some thought, he found a proof which applies not only to that case, but to all similar problems.

In the introduction of his paper, Euler writes (translation from [29]):

> In addition to that branch of geometry which is concerned with magnitudes, and which has always received the greatest attention, there is another branch, previously almost unknown, which Leibniz first mentioned, calling it the geometry of position. This branch is concerned only with the determination of position and its properties; it does not involve measurements, nor calculations made with them. It has not yet been satisfactorily determined what kind of problems are relevant to this geometry of position, or what methods should be used in solving them. Hence, when a problem was recently mentioned, which seemed geometrical but was so constructed that it did not require the measurement of distances, nor did calculation help at all, I had no doubt that it was concerned with the geometry of position, especially as its solution involved only position, and no calculation was of any use. I have therefore decided to give here the method which I have found for solving this kind of problem, as an example of the geometry of position.

Euler's work on this problem is commented in several articles and books.

We now come to Euler's polyhedron formula, and we start with Descartes.

Long before Euler came out with his formula $F - A + S = 2$ relating the faces (F), edges (A) and vertices (S) of a convex polyhedron, Descartes obtained an equivalent result, with a geometric proof, involving the solid angles and the dihedral angles between the faces. Described in modern terms, Descartes' proof consists in computing in two different manners the total curvature of the boundary of the polyhedron.

---

[71] Today, the city of Königsberg, called Kaliningrad, is part of a Russian exclave between Poland and Lithuania on the Baltic Sea.

[72] Today, Danzig is the city of Gdansk, in Poland.

Descartes wrote that proof at the age of 25, but did not publish it. The story of Descartes' manuscript is interesting and we recall it now.

Descartes' manuscript was discovered in Hanover, among Leibniz's estate. The latter had copied Descartes' proof during a stay in Paris, in 1675 or 1676, presumably with the intention of publishing it. The original manuscript of Descartes, which carries the title *Progymnasmata de solidorum elementis* (Preparatory exercises to the elements of solids) [47] is mentioned in an inventory of papers which Descartes left in some chests in Stockholm, the city where he died. The copy, made by Leibniz, carries the same title, with the additional mention *excerpta ex manuscripto Cartesii* (Excerpt from a manuscript of Descartes). After the manuscript was discovered, a French translation was published by Foucher de Careil in 1859, in a volume of unpublished works of Descartes. This publication contained errors, because Foucher, who did it, was not a mathematician. The edition is nevertheless interesting, and in the introduction to the volume [114], Foucher recalls the story of the discovery. The story is also told by Adam in the commentaries of the volume of the Adam-Tannery edition of Descartes' works containing this theorem ([51] tome X, pp. 257–263).

In 1890, Jonquières presented to the Paris Académie des Sciences two Comptes Rendus notes entitled *Sur un point fondamental de la théorie des polyèdres* (On a fundamental property of the theory of polyhedra) [148] and *Note sur le théorème d'Euler dans la théorie des polyèdres* (Note on the theorem of Euler on the theory of polyhedra) [149], without being aware of the work of Descartes on this subject. After Jordan pointed out the existence of the work of Descartes in Foucher's edition, Jonquières published other Comptes Rendus notes on the work of Descartes, cf. [150–152]. Poincaré, in his celebrated first memoir on *Analysis situs* [200] attributes to Jonquières the generalization of Euler's theorem to non-necessarily convex polyhedral surfaces.

There is a relatively recent (1987) critical edition of Descartes' *Progymnasmata* with a French translation, with notes and commentaries, by P. Costabel [47].

Euler reported on his work on polyhedra in his correspondence with Goldbach. In a letter dated November 14, 1750, Euler informs his friend of the following two results which he refers to as Theorems 6 and 11 respectively:

> 6. In any solid enclosed by plane surfaces the sum of the number of faces and the number of solid angles is greater by 2 than the number of edges.
>
> 11. The sum if all planar angles equals four times as many right angles as the number of solid angles, decreased by 8.

The term "solid enclosed by plane surfaces" refers to a convex polyhedron. The first result is the Euler characteristic formula, and the second one is a form of the Gauss-Bonnet theorem for the sphere. Euler writes:

> I am surprised that these general properties in stereometry have not been noticed by anybody, as far as I know, but still more that the most important of them, viz., Th. 6 and Th. 11, are so hard to prove; indeed I still cannot prove them in a way that satisfies me.

In the same letter, Euler gives several examples where the two theorems are satisfied.

In his memoir [85], entitled *Demonstratio nonnullarum insignium proprietatum, quibus solida hedris planis inclusa sunt praedita* (Proof of some of the properties of solid bodies enclosed by planes) and written one year after [84], Euler gave proofs of the two results. In the introduction of [85], he declares that his polyhedron formula is part of a more general research he is conducting on polyhedra. In fact, in the letter to Goldbach we mentioned, Euler announces a result on volumes of simplices in terms of their side lengths (a three-dimensional analogue of Heron's formula for the area of triangles), which he proves later in his paper *Demonstratio nonnullarum insignium proprietatum, quibus solida hedris planis inclusa sunt praedita* [85]. Euler writes[73]:

> Although I had uncovered many properties which are common to all bodies enclosed by plane faces and which seemed to be completely analogous to those which are commonly included among the first principles of rectilinear plane figures, still, not without a great deal of surprise did I realize that the most important of those principles were so recondite that all the time and effort spent looking for a proof of them had been fruitless. Nor, when I consulted my friends, who are otherwise extremely versed in these matters and with whom I had shared those properties, were they able to shed any light from which I could derive these missing proofs. After the consideration of many types of solids I came to the point where I understood that the properties which I had perceived in them clearly extended to all solids, even if it was not possible for me to show this in a rigorous proof. Thus, I thought that those properties should be included in that class of truths which we can, at any rate, acknowledge, but which it is not possible to prove.

One advantage of Euler's proof, compared to the one of Descartes, is that it shows in a clear way the combinatorial aspect of the problem, highlighting the notion of edges and faces of the polyhedron.

The proof that Euler gives in [85] is based on an induction on the number of solid angles, reducing them by one at each step. He writes:

> These proofs are in no way inferior to those proofs used in Geometry except that here due to the nature of solids one must use more imagination, in as much as solids are being depicted on a flat surface.

At the same time, Euler was laying down the foundations of combinatorial topology. He writes (Scholion to Proposition 4):

> I admit that I have thus brought to light only the first principles of Solid Geometry, on which this science should be built as it develops further. No doubt it contains many outstanding qualities of solids of which we are so far completely ignorant. [...]

Legendre, in his *Éléments de géométrie* (Note XII) published in 1794 [173], gave a complete proof of Euler's theorem based on geometry. This proof is considered as one of the simplest, and it is repeated in more modern works, e.g. in Hopf [140].

A large number of mathematicians commented on Euler's polyhedron formula, expanding some arguments in Euler's proofs, giving new proofs, and sometimes comparing Euler's work with that of Descartes. To show the diversity of these works, we mention the papers by Andreiev [10], Bertrand [26], Bougaïev [22], Brianchon [27], Catalan [31], Cauchy [32, 33], Feil [113], Gergonne, [130], Grunert, [135], Jonquières [148–152, 154–156], Jordan [157, 158], Lebesgue [168], Lhuillier [177],

---

[73] Translation by C. Frances and D. Richeson.

Poincaré [199, 200], Poinsot [203, 204], Prouhet [208, 209], Steiner [239], Valat [249] and Thiel [244]. We shall quote some of these works below. We also mention that in 1858, the Paris *Académie des Sciences* proposed as a subject for the 1861 *Grand prix*: "To improve, in some important point, the geometric theory of polyherda." Möbius participated and presented a memoir (but did not get the prize).

In 1811, Cauchy brought out a purely combinatorial proof of that theorem. In this proof, one starts by deleting a face of the polyhedron and reduces the problem to another one concerning a planar polygon.[74] In his article *Recherches sur les polyèdres* (Researches on polyhedra) [32], published in 1813, Cauchy writes:

> Euler has determined, in the Petersburg *Mémoires*, year 1758, the relation that exists between the various elements that compose the surface of a polyhedron; and Mr. Legendre, in his *Éléments de Géométrie*, proved in a much simpler manner Euler's theorem, by considerations of spherical polygons. Having been led by some researches to a new proof of that theorem, I reached a theorem which is much more general than the one of Euler, whose statement is the following:
>
> Theorem. If we decompose[75] a polyhedron in as many others as we wish, by taking at will new vertices in the interior, and if we represent by P the number of new polyhedra thus formed, by S the total number of vertices, including those of the initial polyhedron, by F the total number of faces, and by A the total number of edges, then we will have
>
> $$S + F = A + P - 1,$$
>
> that is, the sum of the number of vertices and that of faces will overpass by one the sum of the number of edges that of polyhedra.[76]

Poinsot [203], in 1858, published a proof of Euler's formula using some of Cauchy's arguments. He writes: "This relation, which Euler was the first to prove, does not hold only for convex polyhedra, as one might think, but for polyhedra of any kind." In fact, this statement needs some explanation. We are used today to the fact that Euler's formula is valid for polyhedra which are homeomorphic to a sphere. This notion did not exist at that time, neither the word, nor the idea. One had to

---

[74] A similar proof is given by Hilbert and Cohn-Vossen [138] p. 290.

[75] Cauchy "decomposes" the polyhedron by taking new vertices in the interior of the three-dimensional polyhedron (and not on the boundary surface).

[76] Euler a déterminé, dans les *Mémoires* de Pétersbourg, année 1758, la relation qui existe entre les différents éléments qui composent la surface d'un polyèdre ; et M. Legendre, dans ses *Éléments de Géométrie*, a démontré d'une manière beaucoup plus simple le théorème d'Euler, par la considération des polygones sphériques. Ayant été conduit par quelques recherches à une nouvelle démonstration de ce théorème, je suis parvenu à un théorème plus général que celui d'Euler et dont voici l'énoncé :

Théorème. Si l'on décompose un polyèdre en tant d'autres que l'on voudra, en prenant à volonté dans l'intérieur de nouveaux sommets ; que l'on représente par P le nombre de nouveaux polyèdres ainsi formés, par S le nombre total de sommets, y compris ceux du premier polyèdre, par F le nombre total de faces, et par A le nombre total des arêtes, on aura

$$S + F = A + P - 1,$$

c'est-à-dire que la somme faite du nombre des sommets et de celui des faces surpassera d'une unité la somme faite du nombre des arêtes et de celui des polyèdres.

wait for that to the work of Jordan, who set up the precise hypotheses under which Euler's formula is valid. In his article [155], he writes that Euler's theorem is valid for polyhedra which he calls "simple," or "Eulerian," that is, polyhedra ([155] p. 35) "such that any contour drawn on the surface which does not traverse itself divides this surface into two separate regions; a category that contains as a particular case convex polyhedra."[77] A few pages later (p. 38), Jordan makes the following commentary: "It would have been easy to show that if we can draw on a polyhedron λ different contours which do not intersect each other and which do not divide the surface into separate parts, we would have $S + F = A + 2 - 2\lambda$."[78] In fact, Jordan had extracted the notion we call today "topological surface of finite type," to which the general theory applies, cf. [154] p. 86:

> A surface is said to be of type (m, n) if it is bounded by m closed contours and if furthermore we can draw on it n closed contours that do not intersect themselves nor mutually, without dividing it into two distinct regions.[79]

Then Jordan makes the relation with the polyhedra to which Euler's formula applies: "The polyherda of kind (0, 0) are nothing but those which I called *Eulerian*."[80]

It is interesting to read Lebesgues' comments on some proof of Euler's theorem, because it gives us some hints of how the subject of topology was viewed in those days. Lebesgues' comments are written in 1924 ([168] p. 319):

> I don't agree at all with those who pretend to attribute Euler's theorem to Descartes. Descartes did not state the theorem; he did not see it. Euler perceived it and he fully understood its character. For Euler, the description of the form of a polyhedron must precede the use of the measures of its elements, and this is why he set his theorem as a fundamental theorem. For him, like for us, this is a theorem of enumerative *Analysis situs*; therefore he tried to find it by considerations independent of any metrical theory, that in effect belong to what we call the field of *Analysis situs*. And this is why he left to Legendre the honor of finding a rigorous proof. None of us who had read a little bit of Euler and who were amazed by his prodigious technical masterliness will doubt, even for one second, that if Euler had thought of putting aside his theorem and deducing it from one of its metric corollaries, he would have easily succeeded. (It should be noted that Euler does not at all restrict his researches to convex polyhedra.) It seems to me, on the contrary, that the fact that Descartes passed so closely to the theorem without seeing it, emphasizes Euler's credit. (At least, this is what we believe, because Descartes employed in his notebook some algebraic characters which he used before knowing Viète's characters.) But Leibniz, who found Descartes' notebook enough interesting to copy it, who realized that Descartes' geometry does not apply to questions involving order and position relations, who dreamed of constructing the algebra of these relations and who in advance gave it the name *Analysis situs*, did not notice, in Descartes' notebook Euler's

---

[77] [...] tels que tout contour fermé tracé sur leur surface et ne se traversant pas lui-même divise cette surface en deux régions séparées ; catégorie qui renferme comme cas particulier les polyèdres convexes.

[78] Il serait aisé de démontrer que si l'on peut tracer sur un polyèdre λ contours différents, ne se coupant pas mutuellement et ne divisant pas la surface en parties séparées, on aura $S + F = A + 2 - 2\lambda$.

[79] Une surface sera dite d'espèce (m, n) si elle est limitée par m contours fermés et si l'on peut d'autre part y tracer n contours fermés ne se coupant pas eux-mêmes ni mutuellement, sans la partager en deux régions distinctes.

[80] Les polyèdres de l'espèce (0, 0) ne sont autres que ceux que j'ai appelés *eulériens*.

> theorem which is so fundamental in *Analysis situs*. This theorem really belongs to Euler. As for the proof, one could, may be with a little bit of unfairness, call it the proof of Legendre and Descartes. This proof is metrical, and it is fair to blame it for the fact that it uses notions that are foreign to *Analysis situs*. But one should not exaggerate the value of this grievance.[81]

We now give a quick review of some work of Gauss on topology, another field in which his impact on Riemann was huge.

Gauss was interested in applications of *Geometria situs* (a term he used in his writings), in particular in astronomy, geodesy and electromagnetism. In astronomy, he addressed the question of whether orbits of celestial bodies may be linked (cf. his short treatise entitled *Über die Grenzen der geocentrischen Orter der Planeten*). From his work on geodesy, we mention his letter to Schumacher, 21 Nov. 1825, (from Gauss's *Werke* vol. VIII, p. 400):

> Some time ago I started to take up again a part of my general investigations on curved surfaces, which shall become the foundation of my projected work on higher geodesy. [...] Unfortunately, I find that I will have to go very far afield [...]. One has to follow the tree down to all its root threads, and some of this costs me week-long intense thought. Much of it even belongs to geometria situs, an almost unexploited field.

From Gauss's *Nachlaß*, we know that he worked on a combinatorial theory of knot projections, during the year 1825, and again in 1844. (Gauss's *Werke*, Vol. VIII, pp. 271–286). We already mentioned at the beginning of this section, that we learn from a letter sent by Betti to Tardy that the idea of analyzing a surface by performing successive cuts was given to Riemann by Gauss, in a private conversation. Besides Riemann, Gauss had two students who worked on topology and who were certainly influenced by him: Listing and Möbius.

---

[81] Je ne suis pas du tout d'accord avec ceux qui prétendent attribuer à Descartes le théorème d'Euler. Descartes n'a pas énoncé le théorème ; il ne l'a pas vu. Euler l'a aperçu et en a bien compris le caractère. Pour Euler, la description de la forme d'un polyèdre doit précéder l'utilisation des mesures de ses éléments et c'est pourquoi il a posé son théorème comme théorème fondamental. C'est, pour lui comme pour nous, un théorème d'*Analysis situs* énumérative ; aussi a-t-il cherché à le démontrer par des considérations indépendantes de toute propriété métrique, appartenant bien à ce que nous appelons le domaine de l'*Analysis situs*. Et c'est pourquoi il a laissé à Legendre l'honneur d'en trouver la preuve rigoureuse ; aucun de ceux qui ont quelque peu lu Euler, et qui ont été stupéfaits de sa prodigieuse virtuosité technique, ne doutera un seul instant que si Euler avait pensé à faire passer son théorème au second plan et à le déduire d'un de ses corollaires métriques, il n'y eût facilement réussi. (Il convient d'ajouter qu'Euler ne restreint nullement ses recherches aux polyèdres convexes.) Que Descartes soit passé si près du théorème sans le voir me paraît au contraire souligner le mérite d'Euler. Encore peut-on dire que Descartes était jeune quand il s'occupait de ces questions. (C'est du moins ce que l'on croit, parce que Descartes a employé dans son cahier certains caractères cossiques qu'il utilisait avant de connaître les notations de Viète.) Mais Leibniz qui a trouvé le cahier de Descartes assez intéressant pour le copier, qui a reconnu que la géométrie de Descartes ne s'appliquait pas aux questions où interviennent des relations d'ordre et de position, qui a rêvé de construire l'algèbre de ces relations et l'a nommée à l'avance *Analysis situs*, n'a pas aperçu, dans le cahier de Descartes, le théorème d'Euler si fondamental en *Analysis situs*. Le théorème appartient bien à Euler ; quant à la démonstration, on pourrait, un peu injustement peut-être, la dénommer démonstration de Legendre et Descartes. Cette démonstration est métrique ; il est juste de lui reprocher de faire appel à des notions étrangères à l'*Analysis situs*. Mais il ne faudrait pas s'exagérer la valeur de ce grief.

Riemann introduced the fundamental topological notions for surfaces: connectedness, degree of connectivity, the classification of closed surfaces by their genus. He developed this theory for the purpose of using it in his work on the theory of functions of a complex variable. In his memoir on Abelian functions, he talks about *analysis situs*, referring to Leibniz:

> In the study of functions obtained by the integration of exact differentials, a few theorems of analysis situs are almost essential. Under that name, which was used by Leibniz, although may be in a slightly different sense, it is permitted to designate the theory of continuous magnitudes which studies these magnitudes, not as independent of their position and measurable with respect to each other, but by disregarding all idea of measure and studying them only for what regards their relation of position and inclusion. I intend to treat this subject later, in a way that is completely independent of any measure.

In his habilitation dissertation, Riemann mentions the possibility of working in the new field of topology, talking about the notion of "place." We quote this cryptic passage:

> Measure consists in the superposition of the magnitudes to be compared; it therefore requires a means of using one magnitude as the standard for another. In the absence of this, two magnitudes can only be compared when one is a part of the other; in which case we can only determine the more or less and not the how much. The researches which can in this case be instituted about them form a general division of the science of magnitude in which magnitudes are regarded not as existing independently of position and not as expressible in terms of a unit, but as regions in a manifoldness. Such researches have become a necessity for many parts of mathematics, e.g., for the treatment of many-valued analytical functions; and the want of them is no doubt a chief reason for which the celebrated theorem of Abel and the achievements of Lagrange, Pfaff, Jacobi for the general theory of differential equations, have so long remained unfruitful. Out of this general part of the science of extended magnitude in which nothing is assumed but what is contained in the notion of it, it will suffice for the present purpose to bring into prominence two points; the first of which relates to the construction of the notion of a multiply extended manifoldness, the second relates to the reduction of determinations of place in a given manifoldness to determinations of quantity, and will make clear the true character of an n-fold extent.

He also describes the passage from one dimension to another:

> If in the case of a notion whose specialisations form a continuous manifoldness, one passes from a certain specialisation in a definite way to another, the specialisations passed over form a simply extended manifoldness, whose true character is that in it a continuous progress from a point is possible only on two sides, forward or backwards. If one now supposes that this manifoldness in its turn passes over into another entirely different, and again in a definite way, namely so that each point passes over into a definite point of the other, then all the specialisations so obtained form a doubly extended manifoldness. In a similar manner one obtains a triply extended manifoldness, if one imagines a doubly extended one passing over in a definite way to another entirely different; and it is easy to see how this construction may be continued. If one regards the variable object instead of the determinable notion of it, this construction may be described as a composition of a variability of $n + 1$ dimensions out of a variability of n dimensions and a variability of one dimension.

Riemann's ideas on topology are explained in some sections of Klein's booklet [161]. For instance, Sect. 8 carries the title *Classification of closed surfaces according to the value of the integer p*. Let us comment on a passage from Klein's booklet concerning the classification of surfaces, as an example of his point of view on topology. We know that topology, which was an emerging subject, plays an important role in Riemann surface theory. We already mentioned that Riemann introduced several major notions on surface topology. Klein tried to make a more systematic exposition of these ideas. His book [161] contains a chapter in which the classification of closed surfaces according to genera is presented. On p. 32, he writes:

> That it is impossible to represent surfaces having different p's[82] upon one another, the correspondence being uniform, seems evident. It is not meant, however, that this kind of geometrical certainty needs no further investigation; cf. the explanations of G. Cantor (*Crelle*, t. LXXXIV. pp. 242 *et seq.*). But these investigations are meanwhile excluded from consideration in the text, since the principle there insisted upon is to base all reasonings ultimately on intuitive relations.

Klein then states the converse: between any two surfaces of the same genus it is possible to find a "uniform correspondence." He declares that this statement is more difficult to prove, and he refers to the 1866 article by Jordan [153].[83] This paper is an important milestone in the history of topology, because it contains the first attempt to classify surfaces up to homeomorphism, although there was no precise definition of homeomorphism yet.[84] Jordan's aim, in his paper, is to prove the following theorem, which he states in the introduction:

> In order for two surfaces or pieces of flexible and inextensible surfaces to be be applied onto each other without tear and duplication, the following is necessary and sufficient:
>
> (1) That the number of separated contours that respectively bound these portions of surfaces be the same. (If the surfaces considered are closed, this number is zero.)
> (2) That the maximal number of closed contours which do not self-intersect or intersect each other that we can draw on each of the two surfaces without cutting it into two separate regions be the same on both sides.

Jordan gives the following "definition" of two surfaces being "applicable onto each other." For a modern reader, this definition may seem fuzzy, but one has to remember that this paper was written in the heroic epoch of the foundations of modern topology, that the notion of homeomorphism seems extremely natural for us today, but that it was not so in those times. Jordan writes:

---

[82] We recall that p denotes the genus.

[83] Camille Jordan (1838–1922), who is mostly known for his results on the topology of surfaces and on group theory, also worked on function theory in the sense of Riemann. The title of the second part of his doctoral thesis is: "On periods of inverse functions of integrals of algebraic differentials." The subject was proposed to him by Puiseux, whom we mention in this paper concerning uniformization. Jordan is among the first who tried to study the ideas of Galois, and he is also among the first who introduced group theory in the study of differential equations.

[84] The word "homeomorphism" was introduced by Poincaré in his article [200] but with a meaning that is different from the one it has today. There is a definition of homeomorphism in the 1909 article by Hadamard [136], as being a one-to-one continuous map. This is not correct, unless Hadamard meant, by "one-to-one continuous", "one-to-one bi-continuous." We refer the reader to the paper [183] on the rise and the development of the notion of homeomorphism. This paper contains several quotes, some of which are very intriguing.

> We shall rely on the following principle, which we can consider as evident, and take it if necessary as a definition: *Two surfaces* S, S' *are applicable onto each other if we can decompose them into infinitely small elements such that to any contiguous elements of* S *correspond contiguous elements of* S'.

Besides Klein's booklet, several books and treatises explaining Riemann's ideas appeared in the decades that followed Riemann's work. We mention as examples Neumann's *Vorlesungen über Riemann's Theorie der Abel'schen Integrale* (Lectures on Riemann's theory of Abelian integrals) [186], Picard's *Traité d'Analyse* [197], Appell and Goursat's *Théorie des fonctions algébriques et de leurs intégrales* (Theory of Algebraic functions and their integrals) [16], and there are others. The last two treatises, together with several other French books on the theory of functions of a complex variable, are reviewed in Chap. 8 of the present volume [192].

Among the other important topological notions that were introduced before Riemann and that were used by him, we must mention the notion of *homotopy of paths* and its use in complex analysis (in particular, in the theory of line integrals), especially by Cauchy and Puiseux. This is discussed in detail in Chap. 7 of the present volume [191]. Cauchy published his first work on the subject in 1825 [36]. This is also a topic on which Gauss was a forerunner, but he did not publish anything about it. This is attested in his letter to Bessel, December 18, 1811, published in Volume VIII of his *Collected works* (pp. 90–92), a letter in which Gauss makes the important remark that if one defines integrals along paths in the complex plane, then the value obtained may depend on the path.

Regarding the history of Riemann's ideas on topology, we could have also commented on his predecessors regarding the notion of the discreteness and continuity of space, but this would have taken us too far. We make a few remarks on this matter in Chap. 6 of the present volume [190] .

We end this section by quoting Alexander Grothendieck, from his *Récoltes et semailles* (Harvesting and Sowing),[85] commenting on Riemann's reflections on this theme ([133] Chap. 2 Sect. 2.20):

> It must be already fifteen or twenty years ago since, skimming the modest volume that constitutes Riemann's complete works, I was struck by a remark which he made "incidentally." He observes there that it might be that the ultimate structure of space is "discrete," and that the "continuous" representations which we make of it may be an oversimplification (which may turn out to be excessive on the long run...) of a more complex reality; that for human thought, "the continuous" was easier to grasp than "the discontinuous," and that it serves us, subsequently, as an "approximation" for apprehending the discontinuous. This is an amazingly penetrating remark expressed by a mathematician, at a time where the Euclidean model of physical space was not questioned. In a strictly logical sense, it was rather the discontinuous which, traditionally, served as a technical mode of approaching the continuous.
>
> Moreover, the mathematical developments of the last decades showed an even more intimate symbiosis between continuous and discontinuous structures, which was not yet visualized

---

[85] The complete title is: *Récoltes et semailles : Réflexions et témoignage sur un passé de mathématicien* (Harvesting and Sowing: Reflections and testimony on the past of a mathematician). This is a long manuscript by Grothendieck in which he meditates on his past as a mathematician and where he presents without any compliance his vision of the mathematical milieu in which he evolved, especially the decline in morals, for what concerns intellectual honesty.

> in the first half of this century. The fact remains that finding a model which is "satisfactory" (or, if need be, a collection of such models, linked in the most possible satisfying way...), whether the latter is "continuous," "discrete," or of a "mixed" nature—such a task will surely involve a great conceptual imagination, and a consummate intuition, in order to apprehend and disclose mathematical structures of a new type. This kind of imagination and "intuition" seems to me a rare object, not only among physicists (where Einstein and Schrödinger seem to be among the rare exceptions), but even among mathematicians (and I am talking in full knowledge of the facts).
>
> To summarize, I foresee that the long-awaited renewal (if ever it comes...) will rather come from someone who has the soul of a mathematician, who is well informed about the great problems of physics, rather than from a physicist. But above all, we need a man having the "philosophical openness" required to grasp the crux of the problem. The latter is not at all of a technical nature, but it is a fundamental problem of "natural philosophy."[86]

9 Differential Geometry

In this section, we shall review some milestones in the history of differential geometry, concerning especially the study of geodesics and of curvature, from its beginning until the work of Riemann.

Differential geometry starts with the study of differentiable curves. The notion of curvature of planar curves already appears in works of Newton and of Johann I and

---

[86] Il doit y avoir déjà quinze ou vingt ans, en feuilletant le modeste volume constituant l'œuvre complète de Riemann, j'avais été frappé par une remarque de lui "en passant." Il y fait observer qu'il se pourrait bien que la structure ultime de l'espace soit "discrète", et que les représentations "continues" que nous en faisons constituent peut-être une simplification (excessive, peut-être, à la longue...) d'une réalité plus complexe; que pour l'esprit humain, "le continu" était plus aisé à saisir que "le discontinu", et qu'il nous sert, par suite, comme une "approximation" pour appréhender le discontinu. C'est là une remarque d'une pénétration surprenante dans la bouche d'un mathématicien, à un moment où le modèle euclidien de l'espace physique n'avait jamais été mis en cause; au sens strictement logique, c'est plutôt le discontinu qui, traditionnellement, a servi comme mode d'approche technique vers le continu.

Les développements en mathématique des dernières décennies ont d'ailleurs montré une symbiose bien plus intime entre structures continues et discontinues, qu'on ne l'imaginait encore dans la première moitié de ce siècle. Toujours est-il que de trouver un modèle "satisfaisant" (ou, au besoin, un ensemble de tels modèles, se "raccordant" de façon aussi satisfaisante que possible...), que celui-ci soit "continu," "discret" ou de nature "mixte"—un tel travail mettra en jeu sûrement une grande imagination conceptuelle, et un flair consommé pour appréhender et mettre à jour des structures mathématiques de type nouveau. Ce genre d'imagination ou de "flair" me semble chose rare, non seulement parmi les physiciens (où Einstein et Schrödinger semblent avoir été parmi les rares exceptions), mais même parmi les mathématiciens (et là je parle en pleine connaissance de cause).

Pour résumer, je prévois que le renouvellement attendu (s'il doit encore venir …) viendra plutôt d'un mathématicien dans l'âme, bien informé des grands problèmes de la physique, que d'un physicien. Mais surtout, il y faudra un homme ayant "l'ouverture philosophique" pour saisir le nœud du problème. Celui-ci n'est nullement de nature technique, mais bien un problème fondamental de "philosophie de la nature."

* &&

Pág. 01(09) – 182(190).
Notes on non-Euclidean Geometry

Pág. 54(62) 3.7
Projection in neutral geometry

– From Riemann to Differential Geometry and Relativity... (2017).pdf

Pág. 46(80) 8. Topology
Poincaré who is certainly the major founder of the modern field of topology

First letter by Betti, dated October 6, 1863.

Betti then writes:

"What gave Riemann the idea of the cuts was that Gauss defined them to him, talking about other matters, in a private conversation.
In his writings one finds that analysis situs, that is, the consideration of quantities independently from their measure, is "wichtig"..."

Pág. 101 (134) - 3.1 Gauss

Pág. 100 (133) →

This led Riemann to explain the difference between intrinsic and extrinsic properties of a surface.

He explained that for a sphere, which has an intrinsic geometry different from a plane, the Gaussian curvature multiplied by the area of an infinitesimal geodesic triangle is half the excess of the sum of its angles over π.

This allowed him to express the belief that the geometry of an n-dimensional manifold could be understood by understanding its sectional curvatures.

Pág. 142 (174) Theorem 2.2

If restricted to surfaces of Gauss curvature $K \leq 0$, then the surface is minimal if and only if its Gauss map is conformal.

This paper will first look at the *Hypotheses* in detail, and then consider its relation to Gauss's ideas about differential geometry. Then it will turn to the connections to physics and philosophy, and conclude by looking at the other relevant paper Riemann wrote, the *Commentatio* [27] and the the discussions that it provoked.

2 The *Hypotheses*

Riemann's essay [24] was published posthumously in 1867, and is based on the lecture he gave in 1854 as part of the process of obtaining his Habilitation, the necessary and sufficient qualification for obtaining a teaching position at a German university. As such it was given to the Philosophy Faculty at Göttingen, of which Mathematics was a Department, with Gauss as one of the examiners. These circumstances explain the unfortunate absence of formulae that would otherwise have assisted subsequent readers.

Riemann began by remarking that geometry "takes for granted the notion of space" as well as the first principles of constructions in space. The basic concepts have only nominal definitions and the crucial properties are determined from axioms, but this leaves the relationships between the axioms obscure; in fact, said Riemann, it is not even clear a priori if the relationships are possible.

This opening paragraph makes two points. First, that it is not clear what the axioms or postulates of Euclidean geometry are about; second, that it is not even clear that they are consistent.

But Riemann did not take his analysis in the direction of a more refined study of the axioms. It is likely that he saw the work of Legendre, whose name he mentions, as indicative of the poverty of such work, not only because in all the editions of his *Géométrie* [16] Legendre had failed to prove the parallel postulate, but because Riemann thought that the whole axiomatic attempt to give a geometrical account of physical space was misguided. In unpublished notes from the early 1850 s he called such enquiries "extremely unfruitful" (see Scholz ([30] p. 218)), which, as Scholz points out, makes it very unlikely that Riemann had seen any of Lobachevskii's work.

Instead, Riemann began his paper [24] by remarking that the general notion of quantity was multi-dimensional, and "it emerges from this study that a multi-dimensional object is capable of being measured in different ways and that space is only a particular example of a three-dimensional quantity." Moreover, the properties that characterise space among all three-dimensional quantities can only be determined experimentally. From this perspective the axioms of Euclid are only hypotheses, although highly probable outside "the realms of both the immeasurably great and the immeasurably small."

This opening page is one of the first places where a characteristically modern mathematical approach is taken to what many had seen as a straight-forward question. Riemann did not say that Euclidean geometry needs fixing and offer a proposal. He stepped back and asked himself: what are we studying when we study geometry? His answer was quantity, and for him that was a multi-dimensional object—the sort of thing that is described, as we shall see, with coordinates. In Riemann's opinion a geometer, even a practical one, should not concentrate on the passage from space to a mathematical theory of space (still less take one for granted) but first build up a theory of multi-dimensional quantities, and this he turned to do.

He began Part I of the paper by regretting that apart from some ideas of Gauss [14] published in his second memoir on biquadratic residues and some philosophical remarks of Herbart [15] there was little to guide him. He indicated that there are continuous and discrete manifolds depending on how the elements are determined, and here a manifold is a vague term, little more than a collection of elements. Discrete manifolds in this sense are frequently encountered, said Riemann, but continuous ones less so, and he gave examples of the location of material objects and their colours. To make sense of continuous manifolds we rely on measurement, which presupposes a measuring unit that can be freely transported. When this is not available we have only the general concept of a manifold, and this difficulty may be why the work of Lagrange, Pfaff, and Jacobi on many-valued analytic functions has been so unfruitful so far.

Riemann had already rewritten the theory of functions of a complex variable in his doctoral theses [23], published in 1851. It is likely that his ideas about surfaces were what he was alluding to here, although the great work on Abelian functions would not be published until 1857.

How then to determine position in a manifold? Riemann explained that if the manifold has dimension one then position is determined by moving forwards and backwards using some unspecified concept of length. If this one-dimensional manifold is itself then moved forwards and backwards in a different dimension then a two-dimensional manifold is obtained, and so on. The converse also holds: one can break an n-dimensional manifold down into smaller ones along which some function has a constant value, and exceptional cases aside these sets where the function takes a fixed value are $n - 1$-dimensional submanifolds.

In Part II of his paper [24] Riemann explained how to introduce metrical relations in a manifold on the assumption that lines have a length independent of their position and every line can be measured by every other line. Here he was happy to acknowledge the work of Gauss [12] on curved surfaces (which we shall look at below).

Riemann now supposed that the position of a point in perhaps some region of an n-manifold is determined by its n coordinates $(x_1, x_2, \ldots, x_n)$. He restricted his attention to continuous systems in which the coordinates can vary by amounts dx and sought an expression for the line element ds in terms of the $dx_1, dx_2, \ldots, dx_n$. He further assumed that the length of a line element is unaltered if all its points undergo the same infinitesimal displacement. If moreover distance increases as points move away from the origin and the first and second derivatives are finite then the first derivative must vanish and the second cannot be negative, so Riemann took it to be positive. He deduced that the line element ds could be "the square root of an everywhere positive quadratic form in the variables dx", as for example we take to be the case for space when we write

$$ds = \sqrt{\sum (dx)^2}.$$

Riemann noted that there are other possibilities. For example, ds could be the fourth root of a fourth power expression, but he did not see many possibilities for geometry there and he set this aside.

The quadratic form, however, did interest him. It contains $n(n+1)/2$ coefficients, of which n can be altered by a change of variables, so it depends essentially on $n(n-1)/2$ coefficients that are determined by the manifold. The example of $\sqrt{\sum (dx)^2}$ is therefore special, and Riemann proposed to call such manifolds flat.

To proceed further, Riemann considered the infinitesimal triangle with one vertex at the origin, one on a geodesic out of the origin to the point $(x_1, x_2, \ldots, x_n)$, and one on a geodesic out of the origin to the point $(dx_1, dx_2, \ldots, dx_n)$. The quotient of $\sqrt{\sum (dx)^2}$ by the area of this triangle measures the departure of this infinitesimal region from flatness, and divided by $-3/4$ is in fact the Gaussian curvature of the surface. So the curvature of an n-manifold can be understood by knowing $n(n-1)/2$ surface curvatures (the sectional curvatures, as we would say).

This led Riemann to explain the difference between intrinsic and extrinsic properties of a surface. He explained that for a sphere, which has an intrinsic geometry different from a plane, the Gaussian curvature multiplied by the area of an infinitesimal geodesic triangle is half the excess of the sum of its angles over π. This allowed him to express the belief that the geometry of an n-dimensional manifold could be understood by understanding its sectional curvatures.

Flat manifolds, he observed, have every sectional curvature zero. They are therefore a special case of the manifolds of constant curvature (that is, having the same sectional curvatures everywhere) and in these manifolds geometric figures can be freely moved around without stretching. To give examples of such manifolds he wrote down the metric

$$ds = \frac{1}{1 + \frac{\alpha}{4} \sum x^2} \sqrt{\sum (dx)^2},$$

where α is the curvature.

It was evident that the surfaces of constant positive curvature are spheres; the sphere of radius r has curvature r^{-2}. Riemann gave a complicated description of how to fit all the surfaces of constant curvature into one family. On this description, the cylinder is the example of a surface of zero curvature, and surfaces of constant negative curvature are locally like the saddle-shaped part of a torus.

In the third and final part of his paper [24] Riemann discussed how his ideas might apply to space. If all the sectional curvatures are zero, the space is Euclidean. But if we assume only that there is free mobility of bodies then space is described as a three-dimensional manifold of constant curvature, which can be determined from the knowledge of the sum of the angles in any triangle. Or, one could assume that length is independent of position but not direction.

As for empirical confirmation, the topological structures available to describe three-dimensional manifolds form a discrete set, so exact statements can be made about them even if one can never be certain of their truth. As for the metrical relations, however, these are necessarily inexact because every measurement is imprecise. This has implications for the immeasurably large and the immeasurably small.

The immeasurably large divided into spaces that are infinite and spaces that are merely unbounded, as for example a sphere. That said, Riemann regarded questions about the immeasurably large as irrelevant to the elucidation of natural phenomena, if only because existing astronomical measurements show that any non-zero sectional curvature of space can only be detected in regions vastly greater than the range of our telescopes. This seems to belong to a Göttingen tradition going back to Gauss and extending at least as far as Schwarzschild, who reported on the implications of measurements of the parallel of stars for the curvature of space in 1899 (see (Epple [7])).

Not so questions about the immeasurably small. Here "the concept of a rigid body, and the concept of a light ray, cease to be meaningful". But, Riemann concluded his lecture, these questions take us into physics "which the nature of today's occasion does not allow us to explore".

3 Influences

3.1 Gauss

It is possible to read Riemann's paper [24] in various ways. A modern mathematician can supply the missing details, or, at the other extreme, regard it as almost incoherent. It is possible for a mathematician to offer a comparably deep vision of new mathematics today, but it would be couched in a language of possible definitions, possible methods, and likely theorems that, conjecturally, resolve outstanding problems. Riemann's paper is more philosophical–in the good sense of challenging one to be clear about what is involved in an enquiry–and more speculative.

As Riemann made clear, among the few antecedents he could acknowledge were two papers by Gauss [11, 12]. The one on differential geometry is easy to appreciate. In the 1810s and 1820s Gauss had re-defined the subject in two memoirs. In his *Disquisitiones generales circa superficies curvas* of 1828 he introduced the concept of the intrinsic curvature of a surface. Gauss began his exposition by taking his readers through three definitions of a surface: in the first a surface is given by an expression of the form $z = f(x, y)$; in the second by an expression of the form $f(x, y, z) = 0$; and in the third in the parameterised form $(x(u, v), y(u, v), z(u, v))$. For each of these approaches he showed what the implications were calculating the curvatures of the principal curves at each point, which Euler [8] had showed are a good way to understand how curved a surface is at each point.

Gauss then introduced the map later known as the ‘Gauss map’. At each point P of the surface he supposed there was a vector of unit length and normal to the surface, PP', and he considered the unit vector OQ parallel to PP' that has its base point at the centre of a fixed sphere of unit radius. The image of P' on the surface under the Gauss map is the point Q on the unit sphere.

Gauss then proved that the ‘Gauss map’ has a simple effect on areas: it multiplies the infinitesimal area around a point by an amount equal to the product of the principal curvatures. This product he proposed to call the curvature of the surface, and he showed that it depends only on E, F, and G and their derivatives with respect to u and v, but not on $x(u, v)$, $y(u, v)$, and $z(u, v)$. It is therefore intrinsic to the surface—a result that surprised him so much he called it the Theorema egregium or exceptional theorem.

One reason Gauss regarded the third form of presenting a surface as not only the most general but the most important, was because it allows u and v to be used as coordinates, and because it allows for a study of maps between one surface and another. In particular, given two surfaces defined by

$$\mathbf{r} = (x(u, v), y(u, v), z(u, v)) \text{ and } \mathbf{r}' = (x'(u', v'), y'(u', v'), z'(u', v'))$$

and a map between them, one can compare the line elements

$$ds^2 = E(u, v)du^2 + 2F(u, v)dudv + G(u, v)dv^2$$

and

$$ds'^2 = E'(u', v')du'^2 + 2F'(u', v')du'dv' + G'(u', v')dv'^2,$$

where $E(u, v) = \mathbf{r}_u.\mathbf{r}_u, \ldots, G'(u', v') = \mathbf{r}'_{v'}.\mathbf{r}'_{v'}$.

For example, the map is conformal or angle preserving if

$$ds^2 = \Phi(u', v')ds'^2,$$

for some function Φ, so in particular a map between a plane and a surface with $ds^2 = E(u, v)du^2 + 2F(u, v)dudv + G(u, v)dv^2$ is conformal if and only if $E = G$ and $F = 0$. Gauss [12] had made a detailed study of maps in connection with the survey of Hannover and South Denmark on which he worked in the 1820s, and had explicitly remarked that a map between planes is conformal if and only if it is given by a complex analytic function; this was only one of several occasions when he hinted at a theory of such functions that he was never to pull together and publish. It is very likely, however, that Riemann knew some of these ideas, but it is often impossible to say if he learned of them in discussions with Gauss or only by reading Gauss's papers [11, 12] after Gauss died. However, he wrote explicitly in his doctoral paper (1851) that the conformal nature of a complex analytic map was something that he learned from Gauss's paper [12] on conformal maps, and he stressed the importance of this geometrical aspect of the maps.

So Riemann took two ideas from Gauss's work on geometry. The conformal nature of a complex analytic map (away from any branch points) surely suggested to Riemann that there was significant geometrical features of a surface as early as 1851. But the idea of the intrinsic curvature of a surface was one Riemann took far beyond what Gauss had done with it.

Gauss [12] had identified the intrinsic feature of the geometry of a surface in $\mathbb{R}^3$, but he continued to think of surfaces as lying in $\mathbb{R}^3$. The idea that a region—a surface—with two coordinates u and v and a metric

$$ds^2 = E(u, v)du^2 + 2F(u, v)dudv + G(u, v)dv^2$$

is a fit subject for geometry already, whether or not there an embedding of it in space given by functions $x(u, v)$, $y(u, v)$, and $z(u, v)$ is due entirely to Riemann. It is almost certainly what Gauss had in mind when he said to his friend Wilhelm Weber after listening to Riemann's lecture that the profundity of the ideas that Riemann had put forward had greatly astonished him (see (Dedekind [6] 581)).

But, as Riemann's paltry citations indicate, there was very little done to extend Gauss's ideas of the intrinsic geometry of surfaces in the three decades that separate Gauss's memoir [12] from Riemann's lecture. One of the few papers written on the subject was by H.F. Minding [21], who investigated surfaces of constant negative curvature in his (1839). Bonnet, and Liouville in (Monge [22]), brought Gauss's theory to France, but Riemann went only once to Paris, in 1860, and it is not clear what he knew of French work before.

Riemann had also mentioned Gauss's second memoir on biquadratic residues (Gauss [14]). This is the work in which Gauss introduced what are today called the Gaussian integers, and explained at some length what he had been less overt in his *Disquisitiones Arithmeticae*, that the complex numbers can be thought of as points in a plane. Here (see §16) he stressed the highly intuitive character of this representation, and also–which is what Riemann surely picked up on–that this illuminated the true metaphysics of imaginary quantities.[1]

Gauss went on (§20) to stress that one goes beyond the positive numbers only when what is counted has an opposite (a negative) and what is then counted is not a substance (an object thinkable in itself) but a relation between two objects. More generally one creates new objects when one has a relation that admits a concept of opposite. Then (§22) "The mathematician abstracts totally from the nature of the objects and the content of their relations; he is concerned solely with the counting and the comparison of the relations among themselves." Nonetheless, intuitive representations are helpful and once an intuitive meaning for $\sqrt{-1}$ is completely established "one needs nothing further to admit this quantity into the domain of objects of arithmetic."

By 1850 Riemann did not need to be told that complex numbers were admitted into mathematics. Like Gauss, and Cauchy, he knew that the problem was not with complex numbers but with how to define complex functions. But he may well have appreciated the metaphysics, the abstract character of mathematical objects and their relations, and the connection to intuition that, as we shall now see, was also a theme of Herbart's philosophy [15] and Riemann's physics.[2]

---

[1] See also the English translation, Ewald ([9], I, 312–313).

3.2 Herbart

The emphasis Riemann placed on Herbart's ideas came from Riemann's interest in philosophy. Herbart had been a philosopher at Göttingen from 1805 until his death in 1841, and his main book, the *Psychology as science newly founded on experience, metaphysics and mathematics* [15], appealed strongly to Riemann. But Riemann was also critical: he wrote in the philosophical passages in his collected works (1990, 539) that he could agree with almost all of Herbart's earliest research, but could not agree with his later speculations at certain essential points to do with his Naturphilosophie and psychology. He also identified himself as a Herbartian in psychology and epistemology, but not in ontology and synechology (a discipline concerned with space, time, and motion, and in particular with intelligible space, regarded as a mental construct that makes the explanation of matter possible).

In Riemann's view, natural science is the attempt to comprehend nature by precise concepts, and if concepts yield inaccurate predictions then the concepts must be modified. As a result, the more of nature we understand the more it sinks below the surface of phenomena. Riemann approved of Herbart's anti-Kantian epistemology, because Herbart [15] had argued that all our concepts arise by modifying earlier ones, and the most primitive concepts originate from attempts to understand what our senses tell us, which is why we have the possibility of forming concepts adequate for natural science. In particular they need not be a priori, as the Kantian ones are.

Herbart was a powerful source for the idea of varying quantities–ultimately, manifolds–although Herbart remained fixed on the idea that geometry was necessarily three-dimensional. But Riemann aimed at constructing coherent systems of concepts that could then be matched against the coherence of the natural world. He did not agree with Herbart's account of how our ideas of space are generated from experience, and went directly to systems of mathematical concepts. The elucidation of fundamental concepts was characteristic of Riemann's work, and it was an approach he shared with Herbart even when he did not use the same concepts himself.

3.3 Physics: Newton and Euler [8]

In a note Riemann [28] made on his work (*Werke*, 539) he wrote

> My main work consists in a new formulation of the known natural laws – expressing them in terms of other fundamental ideas – so as to make possible the use of experimental data

---

[2] This account draws on Bottazzini and Tazzioli (1995 [2]) and Scholz ([31] 1982b).

* " " <

Pag. 101 (134) - 3.1 Gauss

Pag. 100 (133) →

This led Riemann to explain the difference between intrinsic and extrinsic properties of a surface.

He explained that for a sphere which has an intrinsic geometry different from a plane the Gaussian curvature multiplied by the area of an infinitesimal geodesic triangle is half the excess of the sum of its angles over π.

This allowed him to express the belief that the geometry of an n-dimensional manifold could be understood by understanding its sectional curvatures.

Pag. 142 (174) Theorem 2.2

If restricted to surfaces of Gauss curvature $K \leq 0$ then the surface is minimal if and only if its Gauss map is conformal.

Theorem 2.2 *The Gauss map is conformal at* $P \in \Omega$ *either if the surface is umbilical at* P *or if the mean curvature vanishes. If restricted to surfaces of Gauss curvature* $K \leq 0$ *then the surface is minimal if and only if its Gauss map is conformal.*

We recall here that the stereographic projection is conformal.

Furthermore, when X, Y and Z are considered as holomorphic functions in η, Riemann writes down the pull-back metrics of the Euclidean metric on the respective complex plane as

$$dX \otimes dX' = \frac{\partial x}{\partial \eta}\frac{\partial x}{\partial \eta'} d\eta \otimes d\eta',$$

$$dY \otimes dY' = \frac{\partial y}{\partial \eta}\frac{\partial y}{\partial \eta'} d\eta \otimes d\eta',$$

and

$$dZ \otimes dZ' = \frac{\partial z}{\partial \eta}\frac{\partial z}{\partial \eta'} d\eta \otimes d\eta',$$

all of which, induce isothermal coordinates.

In §8 and §9, Riemann-Hattendorff is concerned with explicit parametrizations of the minimal surface, not in terms of η which was the Gauss map post-composed with the stereographic projection, but a new variable u which takes away the gauge freedom SO(3) arising from the position of the pole of the stereographic projection, which corresponds to the choices of oriented hyperplanes in $\mathbb{R}^3$.

Let α be a point in the yz-plane (or equivalently the η-plane) which is the image by the stereographic projection $P_S : S^2 \setminus \{S\} \to yz$-plane of a point $P_S^{-1}(\alpha) \in S^2 \setminus \{S\}$. (Recall $S = (-1, 0, 0)$.) We introduce a change of variables $\eta_\alpha := e^{i\theta}\frac{\eta-\alpha}{1+\alpha'\eta}$ which represents the transformation law which relates the image η of the stereographic projection of $S^2 \setminus \{S\}$ onto the yz-plane to the image η_α of the stereographic projection of $S^2 \setminus P_S^{-1}\{-\alpha\}$ onto the hyperplane Π_α whose unit normal vector is α. Note that the map $\eta \mapsto \eta_\alpha$ is conformal. For $P \in \Omega$, we then define $x_\alpha(P)$ to be the height of a point P over the hyperplane Π_α.

It is shown then that for $\eta_\alpha := e^{i\theta}\frac{\eta-\alpha}{1+\alpha'\eta}$,

$$(d\log\eta_\alpha)^2 \frac{\partial x_\alpha}{\partial \log\eta_\alpha} = (d\log\eta)^2 \frac{\partial x}{\partial \log\eta},$$

resulting in defining a new variable u

$$u = \int \sqrt{i\frac{\partial x}{\partial \log\eta}} d\log\eta, \tag{2.14}$$

which is independent of the gauge α.

This new variable u, a function in η, effectively contains all the information about the minimal surface Ω. It can be used to recover the three coordinate functions x, y and z as follows. By rewriting x as a function of u in Eq. (2.14) we obtain

$$x = -i \int \left(\frac{du}{d\log\eta}\right)^2 d\log\eta + i \int \left(\frac{du'}{d\log\eta'}\right)^2 d\log\eta'.$$

Recall that $\eta = \eta_0$ is the stereographic projection of $S^2 \setminus \{S\}$ from the south pole $S = (-1, 0, 0)$ so that the north pole $(1, 0, 0)$ is sent to $\alpha = (0, 0)$ in the yz-plane, or the η-plane. And x_0 is the height of the point P in Ω measured from the yz-plane. Analogously, by choosing $\alpha = (1, 0) = 1 + 0i$ in yz-plane, x_α represents the height of P measured from the xz-plane. Namely by substituting

$$\eta_\alpha = \frac{\eta - 1}{1 + \eta}$$

we obtain the representation

$$y = -\frac{i}{2} \int \left(\frac{du}{d\log\eta}\right)^2 \left(\eta - \frac{1}{\eta}\right) d\log\eta + \frac{i}{2} \int \left(\frac{du'}{d\log\eta'}\right)^2 \left(\eta' - \frac{1}{\eta'}\right) d\log\eta'.$$

Similarly, for $\alpha = (0, 1) = 0 + 1i$,

$$\eta_\alpha = \frac{\eta - i}{1 - i\eta}$$

we have

$$z = -\frac{1}{2} \int \left(\frac{du}{d\log\eta}\right)^2 \left(\eta + \frac{1}{\eta}\right) d\log\eta - \frac{1}{2} \int \left(\frac{du'}{d\log\eta'}\right)^2 \left(\eta' + \frac{1}{\eta'}\right) d\log\eta'$$

the height of (x, y, z) over the η_α-plane, which is the xy-plane.

In §10, §11 and §12, the three complex-valued functions X, Y and Z are regarded as holomorphic functions of η, and also of u, and they are locally analytically extended near a point, either in the interior, or on the boundary set of the minimal surface. As for the boundary set, it is bounded by Euclidean lines, line segments, or line segments intersecting at a rational angle $q\pi$ with $q \in \mathbb{Q}$.

In particular, Riemann makes an important geometric observation in §10, where a point P in Ω is regarded as the origin of $\mathbb{R}^3$, and the tangent plane $T_P\Omega$ is identified with the yz-plane. Then Ω is locally the graph of a function $x = x(y, z)$ with

$$x(P) = 0, \qquad \frac{\partial x}{\partial y}(P) = \frac{\partial x}{\partial z}(P) = 0$$

and the scalar function x is locally approximated by a harmonic function over the yz-plane,

$$\frac{\partial^2 x}{\partial y^2} + \frac{\partial^2 x}{\partial z^2} = 0.$$

When the second order Taylor expansion terms are nontrivial, the graph at P has nontrivial curvature, and the eigenvalues of the Hessian of $x(y, z)$ are the principal curvatures of the surface at P, of opposite signs and of the same magnitude, namely the mean curvature vanishes. When all the second order Taylor expansion terms vanish, the complex-valued function X has zeros of higher order.

Namely by supplying the harmonic conjugates $\mathfrak{x}, \mathfrak{y}, \mathfrak{z}$ to the real coordinates x, y and z, one can consider the relation among the complex valued functions X, Y and Z in η. As η can be seen as a holomorphic function in the previously introduced complex variable u, X, Y and Z are in turn holomorphic functions in u. In the consecutive sections §10, §11 and §12, functional equations of these variables are written down, at interior points, interior branch points, unramified as well as ramified boundary points, and at a boundary points where two line segments meet at a rational angle.

In §12, also considered is the situation where the minimal surface is simply connected, unbounded in $\mathbb{R}^3$ and bounded by two non-intersecting infinite lines. In commenting the setting, another geometric observation is made. When the minimal surface is bounded by a line ℓ, choose the coordinates in $\mathbb{R}^3$ so that the x-axis coincides with the line ℓ. Then the unit normal $\nu(P)$ to Ω, written as a vector in $\mathbb{R}^3$, has no x component. When the unit sphere is centered at the origin, and the north pole is on the positive x-axis, the image of the Gauss map $\nu : \ell \to S^2$ lies on the equator $\{(0, y, z) \mid y^2 + z^2 = 1\} \subset \mathbb{R}^3$, and consequently the image η of the stereographic projection $P_S : S^2 \setminus \{S\} \to yz$-plane is the equator $\{|\eta| = 1\}$ itself. Hence $\log \eta = \log |\eta| + i \arg \eta$ is purely imaginary along the boundary line of the x-axis. Then we conclude when x is seen as a function in η instead,

$$i \frac{\partial x}{\partial \log \eta}$$

is real along the boundary line. Similarly if we designate another boundary line as the $\hat{x}$-axis of another coordinate system $(\hat{x}, \hat{y}, \hat{z})$, then

$$i \frac{\partial \hat{x}}{\partial \log \eta}$$

is real along that boundary line. Now recall that the quantity

$$u = \int \sqrt{i \frac{\partial x}{\partial \log \eta}} d \log \eta$$

is independent of the coordinate system on $\mathbb{R}^3$, and hence is independent of the procedure of identifying the x-axis with the boundary line. Namely

$$du = \sqrt{i \frac{\partial x}{\partial \log \eta} d \log \eta}$$

is either real or purely imaginary, depending on the sign of the real number $i \frac{\partial x}{\partial \log \eta}$.

We recollect the construction here: when the minimal surface is bounded by a line segment, either bounded or unbounded, the neighborhood of a boundary point is realized by a conformal map from a region in the unit sphere bounded by a great circle. This fact is reflected in the observation that the function $du/d \log \eta$ is either real or purely imaginary.

Up to this point, several choices of complex parameters are presented; η, u, X, Y, Z, each of which can be used to parameterize the minimal surface Ω locally. In §13, the issue of transcribing one by another among those numbers is treated. The method is effectively what is known as the Schwarz-Christoffel mapping, where in Eq. (2.11) an explicit map $u(t)$ from the upper half t-plane $\mathbb{H}$ to the region in the u-plane, or to be exact, the region that is a ramified cover over the u-plane bounded by totally real or totally imaginary lines, is constructed. The coordinate t of the upper half plane corresponds to u by

$$t = \frac{\text{const.}}{u - b} + \text{terms holomophic in } u$$

where b is a value of u corresponding to a boundary point, and where the constant coefficient of the simple pole $u = b$ is determined by the conditions the imaginary part of t is (1) zero along the boundary, and (2) positive in the interior of the surface. By this local information, u is globally defined by the expression of Eq. (2.11) of the original paper:

$$u = \text{ const.} + \text{const.} \int \sqrt{\frac{\Pi(t-a)\Pi(t-a')\Pi(t-b)}{\Pi(t-c)}} \frac{dt}{\Pi(t-e)}$$

is obtained, where $a = \cup a_i$ denotes the set of branch points in the t-plane, $\Pi(t-a)$ denotes the product of $(t - a_i)$ over i, a' the complex conjugates of a, b the branch points on the real axis, c the pre-images of the vertices of the boundary where two line segments meet, and e stands for vertices in the unbounded sectors spanned by a pair of line segments, where each pair is designated with an angle $\alpha\pi$ between the projections of the two lines to the plane perpendicular to both lines, as well as the length of the shortest connecting line segment between the two lines. In the expression in Eq. (2.11), there appear only square roots due to the fact that Riemann considers only simple branch points, unlike the general formula of Schwarz-Christoffel's (cf [1]) where the power of $(t - c)$ could be any number in $[0, 1]$ representing the exterior angle at a vertex of the polygon.

Note that the function u above is defined over the entire t-plane. This is explained in the original article as

> In order to form the expression for $\frac{du}{dt}$ we must observe that dt is always real along the boundary, and du is either real or pure imaginary. Hence $(du/dt)^2$ is real when t is real. This function can be continuously extended over the line of real values of t by the condition that, for conjugate values t and t' of the variable, the function will also have conjugate values. Then $(du/dt)^2$ is determined for the whole t-plane and turns out to be single-valued.

Thus it is justified that du/dt is identified with the integrand of the complex line integral of Eq. (2.11). In this manner, the function u is represented by the complex variable t on the upper half-space, which then is used to reproduce the minimal surface $\Omega = \{(x(t), y(t), z(t))\}$. The by-product of the construction is the extension of the domain of u over the entire complex plane. In modern textbooks, this fact is expressed in the following important statement, often attributed to Schwarz.

Theorem 2.3 (Reflection Principle) *Let* $U(t)$ *be a minimal surface in isothermal parameters defined in a semi-disk* $D_+ = \{|t| < \varepsilon \mid \mathrm{Im}(t) > 0\}$. *Suppose there exists a line* L *in* $\mathbb{R}^3$ *such that* $U(t) \to L$ *when* $\mathrm{Im}(t) \to 0$. *Then* $U(t)$ *can be extended to a generalized minimal surface defined in the full disk* $D = \{|t| < \varepsilon\}$. *Furthermore this extended surface is symmetric in* L.

By utilizing this *universal* variable t, Riemann-Hattendorff (see also the paper [10], in addition to [9]) sets up an ansatz in the following geometric settings for the boundary of the minimal surface Ω:

- two non-intersecting non-parallel infinite lines, which would produce a part of the helicoid bounded by a pair of ruling lines;
- two lines which intersect at a point and a line on the plane whose normal vector is perpendicular to the first two lines;
- four intersecting line segments obtained by removing two edges that do not touch each other from the one skeleton of an arbitrary tetrahedron. This includes the so-called Schwarz surface under the additional symmetry;
- three infinite lines mutually skewed and nonintersecting, closely investigated in the recent work of B. Daniel [3];
- two circles on a pair of parallel planes, which would produce the celebrated "Riemann example" which is a minimal surface bounded by two parallel infinite lines, and foliated by circles. The example is explained in detail by Meeks-Pérez [6];
- two convex polygons on a pair of parallel planes, with and without symmetries, which include the so-called Schwarz P-surface and Schwarz H-surface.

A copy of the surface bounded by two circles on a pair of parallel planes can be extended across the boundary lines via the reflection, which would result in a complete periodic surface of genus zero. This example has recently received renewed attention in minimal surface theory ([6]), as the following classification [7] was demonstrated:

Theorem 2.4 *The plane, the helicoid, the catenoid and the Riemann minimal examples are the only properly embedded minimal surfaces in* $\mathbb{R}^3$ *with the topology of a planar domain.*

The Schwarz-Christoffel mapping is closely related to the theory of elliptic functions and hypergeometric functions, and those examples Riemann calculated fully make use of these theories. In this article, we will not go into the discussion of the papers [9, 10]. Interested readers are directed to Nitsche's book [11] on the subject, which records all the post-Riemannian development of the subject.

3 Representation Formulas by Riemann and Weierstrass-Enneper

In Osserman's book [12], a *generalized* minimal surface is defined as follows

Definition 3.1 A generalized minimal surface Ω in $\mathbb{R}^n$ is a non-constant map $x(p) : M \to \mathbb{R}^n$, where M is a 2-manifold with a conformal structure defined by an atlas, such that each coordinate function $x_k(p)$ is harmonic on M and furthermore

$$\sum_{k=1}^{n} \phi_k^2(\zeta) = 0 \tag{3.1}$$

where $\zeta = \xi_1 + i\xi_2$ is a complex-valued local coordinae of M, and the embedding $x(p)$ is given by

$$\phi_k(\zeta) = \frac{\partial x_k(\zeta)}{\partial \xi_1} - i\frac{\partial x_k}{\partial \xi_2}.$$

In this article, we are only interested in the case $n = 3$.

As the map $x(p)$ is non-constant, at least one of the functions x_k $(k = 1, 2, 3)$ is non-constant, which then implies that the corresponding holomorphic function $\phi_k(\zeta)$ can have at most isolated zeros. Thus the singular points of the map x satisfying

$$\sum_{k=1}^{n} |\phi_k(\zeta)|^2 = 0 \tag{3.2}$$

can exists at most at isolated points on M.

Now the following procedure to specify generalized minimal surfaces is called the Weierstrass-Enneper representation, first devised by A. Enneper (1864) and Weierstrass (1866). First we specify the triplets of holomorphic data:

Lemma 3.2 (cf. [12]) *Let D be a domain in the complex ζ plane, $\eta(\zeta)$ an arbitrary meromorphic function in D and $f(\zeta)$ an analytic function in D having the property that at each point where $\eta(\zeta)$ has a pole of order m, $f(\zeta)$ has a zero of order at least $2m$. Then the functions*

$$\phi_1 = f\eta, \quad \phi_2 = \frac{1}{2}f(1-\eta^2), \quad \phi_3 = \frac{i}{2}f(1+\eta^2) \tag{3.3}$$

Pag. 340 (367) →

Gauss's Curvature,
Riemann's manifolds,
Dedekind's ideals,
and
Grothendieck's topoi in
mathematics,
or
Einstein's spaces curved by
gravity in general relativity and
Heisenberg's matrix variables
or
Dirac's spinors in quantum
physics.

Pag. 344 (371) — In his famous extraordinary theorem (theorema egregium as he called it), Gauss proved that the curvature of a surface which he defined as well, was intrinsic to the surface (surface).

Pag. 543 (563) →

The sectional curvature $K(P)$ agrees with the Gauss curvature at p of the surface ruled by geodesic starting at p with tangent vector $X \in P$.

— Eighteen Essays in Non-Euclidean

* &

Planes of reference give rise to these formations, which define the disciplinary nature of mathematics and science, and essentially shape mathematical and scientific thinking and practice—essentially, but, I would argue, not completely or even most centrally, at least in creative mathematics and science. One might say that creative or, to adopt Thomas Kuhn's language, revolutionary, thought in philosophy and mathematics or science alike is defined by planes of immanence and creation of concepts in Deleuze and Guattari's sense, which always carry individual signatures underneath them [8, p. 50]. Just as there are Plato's ideas, Descartes's cogito, and Leibniz's monads, there are Gauss's curvature, Riemann's manifolds, Dedekind's ideals, and Grothendieck's topoi in mathematics, or Einstein's spaces curved by gravity in general relativity and Heisenberg's matrix variables or Dirac's spinors in quantum physics. This is true even though the functioning of mathematical and scientific concepts does require planes of reference, functions (or other formal mathematical entities), propositions, and so forth.

The difference between philosophy and mathematics or science may instead be, to stay with Kuhn's idiom, in the nature of *normal*, rather than *revolutionary*, practice in each domain. In philosophy, the normal practice consists primarily in understanding, interpreting, and commenting on concepts, while in mathematics and science, the normal practice consists primarily in creating, by means of planes of references, frames of reference, functions and other mathematical or scientific formations, propositions, and so forth. It is true that for Deleuze and Guattari creative, revolutionary philosophical practice is the only true philosophy. However, leaving aside an arguably too restrictive character of this view of philosophy, this is not in conflict with the view of creative mathematics or science advocated here.

Deleuze and Guattari do allow that creative mathematical and scientific thought, such as that of Riemann (one of their primary examples), could have philosophical or artistic, compositional, aspects. But they appear to associate these aspects with *philosophical* or, in Deleuze's language, *inexact* (but philosophically rigorous) thought within mathematical or scientific thought. This philosophical thought is either operative alongside mathematical and scientific thought or enters by way of interference (in the positive sense of interfering wave fronts rather than in the more negative sense of inhibition) between mathematics and philosophy [8, pp. 217–218]. Both types of association are pertinent and important, certainly in Riemann's case. My argument is different, however. I argue that creative *technical*, *exact* mathematical and scientific thought is defined by planes of immanence and multi-component mathematical or scientific concepts, the architecture of which is analogous to that of philosophical concepts in Deleuze and Guattari's sense. That is, even apart from their philosophical strata, *mathematical* and *scientific* planes of immanence and the nature of mathematical or scientific concepts are *analogous* to philosophical thought. It is not only a matter of mathematical and scientific thought becoming philosophical at certain junctures, but a matter of the mathematical and philosophical thought creating parallel homomorphic (partially corresponding to each other), although not isomorphic (fully corresponding to each other), architectures of mathematical and philosophical concepts. They are not isomorphic because of technical, exact, aspects of mathematical and scientific concepts, demanded by the disciplinary nature of mathematics and science, aspects, generally, not found in philosophical concepts.

To bring this point home, I need to say more about Deleuze and Guattari's conceptions of planes of immanence and reference. The plane of immanence, as the plane of the movement of thought, is not "a concept that is or can be thought," but is "the image of thought, the image that thought gives itself of what it means to think" [8, p. 37]. As they say: "Concepts are like multiple waves, rising and falling, but the plane of immanence is the single wave that roles them up and unrolls them" [8, p. 36]. The present argument aims to extend, rather than to juxtapose, the plane of immanence (and the relationship between it and concepts) to mathematical and scientific thought, and to join this plane with the plane of reference. For Deleuze and Guattari, mathematics or science "relinquishes the infinite, infinite speed [of thought], in order to a gain *a reference able to actualize the virtual* [of chaos]. [It] gives reference to the virtual, a reference that actualizes the virtual through functions [or other mathematical objects]" [8, p. 118; translation modified]. Thought's enactment of this process constitutes a plane of reference. Planes of reference do play a major, indeed irreducible, role in mathematical and scientific thinking, especially in the disciplinary functioning of mathematics and science. This, however, is not inconsistent with the view that planes of immanence and the creation of concepts are found in mathematics and science. Mathematical and scientific thought combines both planes (sometimes, as does philosophy, also adding planes of composition) and creates its concepts from this fusion. The processes of thought defined the plane of immanence and specifically the creation of concepts (in Deleuze and Guattari's sense) are equally found in mathematics and science, and, again, define creative thinking there most essentially, analogously to the way it happens in philosophy. Such mathematical planes of immanence and the concepts they give rise to may coexist and interact with philosophical planes of immanence and concepts, but they are not reducible to philosophical planes and concepts. This is because of the equally irreducible interaction between these mathematical planes and concepts with technical, exact aspects of mathematical and scientific thinking.

Thus, in his rethinking of spatiality and geometry, Riemann, not only laid out a philosophical plane of immanence that gives rise to philosophical concepts and architectures, as Deleuze and Guattari rightly argue [7, pp. 483–486] but also introduced *a new mathematical plane of immanence*, which gives rise to multicomponent mathematical concepts, alongside *a mathematical plane of reference*. Riemann's thought shaped the plane of immanence of modern mathematical thought arguably more than that of any other mathematician (although Newton, Gauss, and Galois before Riemann, and Poincaré and Hilbert after him offer some competition). This plane extends well beyond geometry. Riemann made major transformative contributions, especially of a conceptual nature, to many areas of modern mathematics: geometry, topology, analysis, algebra, and number theory–not the least by bringing these fields to bear on each other, and his contributions to physics or philosophy were part of this interactive thinking and practice.

The interactive heterogeneity of Riemann's practice is a crucial aspect of Riemann's mathematics, and it extends beyond mathematics and shapes its mathematical operation from this exterior. Thus, the plane of immanence of Riemann's thought also has a more strictly philosophical dimension and thus creates more strictly philosophical concepts, emphasized by Deleuze and Guattari, rather than only mathematical concepts analogous to philosophical concepts by virtue of their multi-component architecture, stressed here. Riemann's philosophical thought was uncommonly and even nearly uniquely significant for his mathematical or physical thought, although one could think of a few competing cases, such as Weyl, who might have been inspired by Riemann in this respect as well, as he was by many other aspects of Riemann's thought. So were the physical dimensions of Riemann's thought, and this, too, is shared by Weyl, or, earlier Gauss, although this is more common. This role of physics in their work is also essentially connected to the role of philosophy there, which is, again, uncommon, if one speaks of such essential connections, making one's philosophical thinking a constitutive part of one's mathematical thinking, rather than of general philosophical reflections concerning mathematics and science, on the part of mathematicians and scientists.[8] The creation of new mathematics in Riemann was enabled by all three dimensions—mathematical, physical, and philosophical—of Riemann's thought and concepts. Riemann lays out a new plane of immanence of mathematical thought by changing both how to think about geometry or space and how to pursue thinking differently mathematically, via bringing together different mathematical fields and combining them with physical and philosophical concepts. A similar claim could also be made about his thinking concerning physics.

Riemann's Habilitation lecture is a magnificent unfolding of this plane (a geometrical metaphor of "plane" is fitting here). It is a major contribution not only to mathematics but also to physics and philosophy, especially to the philosophy of mathematics and physics, but far from exclusively so, as, for example, Deleuze and Guattari's use of Riemann's concepts shows. It is difficult to overestimate the significance and impact of Riemann's thinking concerning spatiality and geometry in mathematics and physics, in shaping the planes of immanence of thought of both. In mathematics the list of even major areas of impact is nearly inexhaustible, and I shall only mention a few of, arguably, the most important ones. First of all, Riemann's rethinking of geometry in terms of manifoldness crucially expanded the idea of the multiplicity of spaces and geometries themselves. Riemann's view of geometry as, in Deleuze and Guattari's language, topology and typology of manifolds led from the late-nineteenth century on to the extraordinary (a still ongoing) progress of geometry, beginning with the work of Sophus Lie and Felix Klein, and a bit later Élie Cartan [7, p. 483]. This work also connected differential geometry of manifolds and the theory of groups, specifically Lie groups. These connections eventually proved to have a major significance for quantum theory, especially in the theory of elementary particles, which are classified by using Lie groups as symmetry groups, although in connection with the infinite-dimensional spaces. The concept of manifold was also crucial for the development of topology. Initially, it was Riemann's earlier work on Riemann surfaces that had a greater impact. Eventually, topology came to be defined by understanding, which, as I shall explain, follows Riemann, of topological spaces as composed of local neighborhoods of points and (open) subspaces of a given space. It is true that Riemann did not have a concept of topological space. My main concern, however, is the development, transformation of Riemann's concepts (such as manifold) leading to new concepts, such as topological space, which makes Riemann's concepts alive, makes them "live on" in new concepts.

---

[8] Poincaré's extensive (much more extensive than Riemann's) philosophical works (e.g. [28, 29]), while influenced, as were his mathematical works in geometry, by Riemann's Habilitation lecture, may be seen along these lines. I realize that this claim may be challenged, and make it with caution. I would, nevertheless, argue that Riemann's philosophical thinking plays a greater constitutive role in his mathematical thinking than Poincaré's philosophical thinking in his mathematical thinking. The situation is of course different when it comes to physics, which is a major part of Riemann's and Poincaré's mathematical thinking alike, and both made major contributions to physics.

3 Mathematics: Space, Geometry, and the Concept of Manifold

The significance of Riemann's lecture for mathematics, physics, and philosophy, and its impact in all three fields were immense. This impact was delayed until its publication, in 1868, two years after Riemann's death, and fourteen years after it was presented in 1854, although some of Riemann's key ideas contained there and in Riemann's related works became known and had their impact earlier. One can only surmise (a tantalizing surmise!) the consequences for the history of mathematics and physics if the lecture was published more immediately. Riemann opens as follows:

> As is well known, geometry presupposes the concept of space, as well as assuming the basic principles for construction in space. It gives only nominal definitions of these things, while their essential specification appears in the form of axioms. The relationship between these presuppositions is left in the dark; we do not see whether, or to what extent, any connection between them is necessary, or a priori whether any connection between them is even possible.
>
> From Euclid to Lagrange this darkness has been dispelled neither by the mathematicians nor the philosophers who have concerned themselves with it. The reason [ground] [*Grund*] for this is undoubtedly because the general concept of multiply extended magnitudes [*Grösse*], which includes spatial magnitudes, remains completely unexplored. I have therefore first set myself the task of constructing the concept of a multiply extended magnitude from general notions of magnitude. It will be shown that a multiply extended magnitude is susceptible of various metric relations, so that space constitutes only a special case of a triply extended magnitude. From this, however, it is a necessary consequence that the theorems of geometry cannot be deduced from general notions of magnitude, but that those properties that distinguish space from other conceivable triply extended magnitudes can only be deduced from experience. Thus arises the problem of seeking out the simplest data from which the metric relations of space can be determined, a problem that by its very nature is not completely determined, for there may be several systems of simple data that suffice to determine the metric relations of space; for the present purposes, the most important system is that laid down as a foundation of geometry by Euclid. These data are—like all data—not logically necessary, but only of empirical certainty, they are hypotheses [*Hypothesen*]; one can therefore investigate their likelihood, which is certainly very great within the bounds of observation, and afterwards decide on the legitimacy of extending them beyond the bounds of observation, both in the direction of the immeasurably large [*Unmessbargrosse*] and in the direction of the immeasurably small [*Unmessbarkleinen*]. [31, p. 23; translation modified]

* %"" %

particles, which are classified by using Lie groups as symmetry groups, although in connection with the infinite-dimensional spaces. The concept of manifold was also crucial for the development of topology. Initially, it was Riemann's earlier work on Riemann surfaces that had a greater impact. Eventually, topology came to be defined by understanding, which, as I shall explain, follows Riemann, of topological spaces as composed of local neighborhoods of points and (open) subspaces of a given space. It is true that Riemann did not have a concept of topological space. My main concern, however, is the development, transformation of Riemann's concepts (such as manifold) leading to new concepts, such as topological space, which makes Riemann's concepts alive, makes them "live on" in new concepts.

3 Mathematics: Space, Geometry, and the Concept of Manifold

The significance of Riemann's lecture for mathematics, physics, and philosophy, and its impact in all three fields were immense. This impact was delayed until its publication, in 1868, two years after Riemann's death, and fourteen years after it was presented in 1854, although some of Riemann's key ideas contained there and in Riemann's related works became known and had their impact earlier. One can only surmise (a tantalizing surmise!) the consequences for the history of mathematics and physics if the lecture was published more immediately. Riemann opens as follows:

> As is well known, geometry presupposes the concept of space, as well as assuming the basic principles for construction in space. It gives only nominal definitions of these things, while their essential specification appears in the form of axioms. The relationship between these presuppositions is left in the dark; we do not see whether, or to what extent, any connection between them is necessary, or a priori whether any connection between them is even possible.
>
> From Euclid to Lagrange this darkness has been dispelled neither by the mathematicians nor the philosophers who have concerned themselves with it. The reason [ground] [*Grund*] for this is undoubtedly because the general concept of multiply extended magnitudes [*Grösse*], which includes spatial magnitudes, remains completely unexplored. I have therefore first set myself the task of constructing the concept of a multiply extended magnitude from general notions of magnitude. It will be shown that a multiply extended magnitude is susceptible of various metric relations, so that space constitutes only a special case of a triply extended magnitude. From this, however, it is a necessary consequence that the theorems of geometry cannot be deduced from general notions of magnitude, but that those properties that distinguish space from other conceivable triply extended magnitudes can only be deduced from experience. Thus arises the problem of seeking out the simplest data from which the metric relations of space can be determined, a problem that by its very nature is not completely determined, for there may be several systems of simple data that suffice to determine the metric relations of space; for the present purposes, the most important system is that laid down as a foundation of geometry by Euclid. These data are—like all data—not logically necessary, but only of empirical certainty, they are hypotheses [*Hypothesen*]; one can therefore investigate their likelihood, which is certainly very great within the bounds of observation, and afterwards decide on the legitimacy of extending them beyond the bounds of observation, both in the direction of the immeasurably large [*Unmessbargrosse*] and in the direction of the immeasurably small [*Unmessbarkleinen*]. [31, p. 23; translation modified]

These introductory reflections are already profound and far-reaching, and Riemann develops the ideas suggested here quite a bit further in the lecture. First of all, Riemann does not appear to be interested in the axiomatic approach to geometry and is even suspicious of axioms of geometry. In contrast to most previous works on non-Euclidean geometry, the parallel postulate is not the starting point of his investigation. The reason is clearly that the concepts of space and of measuring distances in space are not adequately defined. This is what leaves "the relationships between [axioms] in the dark. We do not see whether, or to what extent, any connection between them is necessary, or a priori whether any connection between them is even possible." This questioning of the axioms of geometry was unusual at Riemann's time, and it was much deeper than customary doubts concerning the parallel postulate. The axiomatic approach to non-Euclidean geometry is, thus, abandoned and even implicitly questioned by Riemann, even though it was this approach that led to the discovery of non-Euclidean geometry by Lobachevsky and Bolyai. This may have been one of the reasons why only Gauss, rather than either Lobachevsky (highly regarded by Gauss) or Bolyai, was expressly invoked in the lecture.[9] It is true that Gauss was ultimately unable to establish a possible existence (in the sense of logical consistency) of non-Euclidean geometry, because, unlike Lobachevsky and Bolyai, he did not do this for the three-dimensional case. In spirit, however, Gauss's work on the geometry of two-dimensional surfaces and his concept of curvature as intrinsic to a given surface was much closer to Riemann. For Riemann, Gauss was pursuing a trajectory of thought better suited for the foundations of geometry and more fruitful for its conceptual development. In his famous extraordinary theorem (*theorema egrerium*, as he called it), Gauss proved that the curvature of a surface, which he defined as well, was intrinsic to the surface. More precisely, the theorem states that the Gaussian curvature of a surface does not change if one only bends the surface but does not stretch it. This means that curvature can be fully determined by measuring angles and distances on the surface itself, without considering the way it is embedded in the ambient (three-dimensional) Euclidean space or, in Riemann's terms, manifold, which makes the Gaussian curvature an intrinsic invariant of a given surface.

One nearly has here a two-dimensional conceptual architecture that suggests and even approaches that of Riemann, although major additional thinking is necessary to give this architecture the type of generality Riemann is able to do. Gauss's theorem suggests that one could see the surface as an independent curved manifold and then to generalize this concept to higher dimensions via "the concept of a multiply extended magnitude," mentioned in this passage and the concept of manifold, which is what Riemann did. These concepts also helped him to generalize to higher dimensions Gauss's concept of curvature. To do so required yet another new concept, another great invention of Riemann, the tensor of curvature, and a new form of differential calculus, tensor calculus on manifolds, a generalization of differential calculus. This calculus was fully developed later on, and it played a major role in Einstein's general relativity. The independence of the curvature or of geometry of a given manifold of any dimension from its embedding also allows one to determine, at least in principle, intrinsically whether this space is flat or curved, which is crucial for determining the nature of the physical space we inhabit. For example, as indicated earlier, space is curved in the immediate vicinity of the solar system, because of the gravity of the Sun, planets, and other material entities in the solar system (or locally around other stars), but appears to be on the overage flat on the scale of the observable Universe.[10] In any event, by virtue of inventing his concepts and by using them, Riemann provided a general rigorous grounding to all geometry, Euclidean or non-Euclidean, rather than merely establishing, as Lobachevsky and Bolyai did, the logical possibility of the non-Euclidean geometry of (constant) negative curvature in three dimensions. The absence of a concept, such as that of manifold, leaves us in "the dark," because "we do not see whether, or to what extent, any connection between [axioms] is necessary, or a priori whether any connection between them is even possible."

---

[9] There were other, more extrinsic, reasons, beginning with the fact that Gauss was Riemann's mentor and the chair of his Habilitation committee. Indeed, Gauss selected this topic among three proposed by Riemann (following the rules). The philosopher R. H. Lotze, a fervent opponent of non-Euclidean geometry, was a member of the philosophy faculty, to which Riemann's Habilitation was presented. Later on, Lotze criticized Riemann's approach anyway, as part of his general critique of non-Euclidean geometry (see [16, pp. 222–226] and [33, pp. 97–112]).

Riemann's thinking is conceptual-problematic rather than axiomatic-theorematic: he grounds mathematics, as well as physics, in hypotheses and concepts, concepts-problems, such as the concept of manifold and its subconcepts (distance, curvature, tensor, and so forth). The problem of space and geometry is now posed in terms of finding concepts that are necessary to define space and to give it geometry. A general concept of manifold was assumed to be applicable to any possible space, while specific manifolds, flat or curved, define different spaces or subspaces. In this respect, as foundational thinking, that of Riemann is different from that conforming to Hilbert's concept of "foundations" [*Grundlagen*] in the sense of axiomatic foundations, an idea that Hilbert also tried to apply to physics. One of the problems of his famous 1900 list (Problem 6) was in fact that of the development of (mathematized) axiomatic foundations, system(s) of axioms, for physics, on the model of his own *Foundations of Geometry* [13].[11] That such a system is possible is inevitably a hypothesis. It became clear subsequently that the existence of such a system is a hypothesis even in mathematics, and in view of Gödel's incompleteness theorems in 1931, ultimately a wrong hypothesis, insofar as such a system cannot be proven to be free of contradiction, once it is large enough to include arithmetic, as geometry is.

---

[10] As I qualified earlier, at least this is a workable and widely accepted view, widely but not universally. It has never been established definitively or, in any event, agreed upon whether such a determination is ever rigorously possible, as opposed to having a practically effective and possibly, within its proper limits, the best available theory or, as Poincaré would have it, "convention," without making a real claim concerning "the reality underlying space" [31, p. 33]. Einstein had his doubts too, although he was ultimately inclined to accept the possibility of such a determination, at least in principle, as, it appears, was Riemann, but not Poincaré, with whose position Einstein, nevertheless, had to contend and which he tried to accommodate within his own (e.g. [11, pp. 324–328]).

[11] For the development of Hilbert's ideas, as reflected in different editions of the *Grundlagen*, see [4]. In the first version of the book, Hilbert was closer to Riemann, and he later returned to a more Riemannian view of geometry in the wake of general relativity to which he made important contributions.

While Hilbert's foundational thinking aimed to bring physics closer to mathematics, even to make it mathematics, by giving physics an axiomatic form, that of Riemann brings mathematics closer to physics by grounding it in hypotheses and concepts, rather than in axioms.[12]

Riemann's approach and his concepts arise from the plane of immanence, at once, mathematical, physical, and philosophical, a plane defined by thinking in terms of multi-component concepts, such as that of multiply extended magnitude and manifoldness, rather than in terms of axioms and propositions, an approach that has defined most thinking concerning geometry before and even after Riemann. Riemann's thought is defined by a plane of immanence by virtue of giving rise to multicomponent concepts, rather than only creating frames of reference, functions, or propositions, from a plane of reference, although this is necessary as well. In Riemann, both planes are joined, as they must be in the creation of mathematical or scientific concepts. Thus, functions define both local neighborhoods (as infinitesimally Euclidean) and how local spaces are connected or pass into each other, and metrical relations and curvature, although curvature is ultimately defined by tensors, which are more complex entities.

The parallel postulate is, again, never mentioned in the lecture, although Euclidean geometry is invoked there as the geometry of "flat space," merely a particular and very special case of geometry, where metrical relations take an especially simple form, defined by the Pythagorean theorem. In Gray's words, "[In Riemann] geometry no longer starts with Euclidean geometry" [11, p. 52]. It is the metrical relation characterizing a given space that defines a possible geometry. In other words, the character of this relation is a hypothesis on which a geometry could be based, a hypothesis to be tested physically in order to establish whether such a space corresponds to the actual physical space. Non-Euclidean geometry (Riemann, again, does not use the term) is introduced as such a possible case of geometry, that of a curved, rather than flat, space of either negative or positive curvature, defined by Riemann by the corresponding type of quadratic form determining the metric. Riemann, thus, not only introduces a more general concept of geometry, but also gives a more rigorous conceptual grounding to non-Euclidean geometry of either negative or positive curvature. While the non-Euclidean geometry of negative curvature (hyperbolic geometry) was discovered before Riemann, Riemann's lecture introduced the three-dimensional non-Euclidean geometry of positive curvature (elliptical geometry), keeping in mind that space for Riemann meant the three-dimensional physical space. (The two-dimensional spherical geometry was considered well before Riemann.) Although eclipsed by Riemann's overall achievement in the lecture, this was a major mathematical discovery with important cosmological implications, for example, in its anticipation of the idea, later considered by Einstein, that the universe may be unbounded and yet finite.[13]

---

[12] One might challenge this argument on historical grounds because it would have been difficult, if not impossible, to present a concept such as that of manifold in axiomatic form at the time of Riemann's lecture. That may be true. My point, however, is that Riemann's alternative, conceptual-problematic rather than axiomatic-theorematic, thinking, could still be contrasted to that of Hilbert and lead to a different type of mathematical thinking. It is difficult to say how Riemann would have approached the foundations of geometry if he had means of axiomatizing his concepts. On the other hand, it is possible to argue, as I do here, *on historical grounds*, that Riemann, unlike his predecessors, Lobachevsky and Bolyai, was not pursuing an axiomatic approach to geometry.

Riemann defines the concept of space, again, understood as physical space (as against, the concept of manifold, which is mathematical), as a three-dimensional instance of the concept of continuous manifoldness, in accordance with the hypotheses that he assumed as likely given the experimental data then available. (There is still no definitive data to refute this view, unless perhaps in the very small, say, at the Planck scale.) While a manifold, as defined by Riemann, may be either discrete or continuous, the concept of continuous manifoldness has a richer and more complex architecture, and most of Riemann's lecture is devoted to it. Technically, continuous manifolds considered by Riemann were differentiable manifolds, which means that one can define differential calculus on them. Indeed, they are metrical manifolds, now called Riemannian, which allow for the concept of distance between any two points and thus for geometry.[14] I shall, however, speak of continuous manifolds, following Riemann and his juxtaposition between continuous and discrete manifoldness. In modern use, the term manifold more customarily refers to continuous (but not necessarily differentiable) manifolds, although one also refers to discrete manifolds, which have topological dimension zero, as zero-manifolds. As defined by Riemann, discrete and continuous manifolds do not appear to have that much in common, and in effect form two different concepts.[15]

Riemann's concept of continuous manifoldness was a new concept of geometrical multiplicity. It is a multiplicity of local subspaces, most specifically those, "neighborhoods," associated with each point, out of which a given space is composed. A continuous (differentiable) manifold is understood by Riemann on the model of two-dimensional surfaces, which, as explained earlier, were defined by Gauss in terms of their intrinsic geometry. Riemann defines first the concept of "n-dimensional magnitude," which allows one to determine a position in a manifold by n numerical determinations, in the same way a position is determined by coordinates in the Euclidean space of n dimensions. Riemann is rigorous to extend (scale down) the concept of manifold to one-dimensional manifolds, curves, which, however, also helps him to built up the concept n-dimensional manifold by analogy. He starts with the concept of "a simply extended [one-dimensional] manifold, whose essential characteristic is that from any point in it a continuous movement is possible in only two directions, forwards and backwards." Then, he defines a two-dimensional or "a doubly-extended manifold" by saying that "if one now imagines that this [one-dimensional manifold] passes to another, completely different one, and once again in a well-determined way, that is, so that every point passes to a well-determined point of the other, then the instances for, similarly, a double extended manifold" [31, p. 25]. In other words, one continuously "fills" a surface with curves. Then, one similarly defines a triply extended manifold by imagining a similar continuous passing of a doubly extended manifold to another, thus continuously filling a three-dimensional object with two dimensional-ones, and so forth. "This construction," Riemann says, "can be characterized as a synthesis of a variability of $n + 1$ dimensions from a variability of n dimensions and a variability of one dimension" [31, p. 25]. This construction may have been one of the reasons for his use of the term: a manifold is literally a continuous fold(er) of manifolds of lower dimensions. Conversely, one can unfold a variability of n dimensions, which allows one to determine a position in a manifold by n numerical determinations, generalizing the way a position is determined by coordinates in the Euclidean space of n dimensions.

---

[13] It would be similar to the three-dimensional sphere. As I explained, the currently dominant view or hypothesis (which appears to be confirmed by cosmological measurements) is that the universe is on average flat and is expanding.

[14] As most of his contemporaries, Riemann did not distinguish continuous and differentiable manifolds. It became eventually clear, however, that not all continuous (also called topological) manifolds are differentiable. There are topological manifolds with no differentiable structure, and some with multiple non-diffeomorphic differentiable structures. Thus, there is a continuum of non-diffeomorphic differentiable structures of $\mathbf{R}^4$.

[15] These two concepts could, especially in modern understanding, be subsumed under the same concept. This is because all zero-dimensional manifolds, which are discrete manifolds in Riemann's terms, are continuous (topological) manifolds. In fact they are also differentiable manifolds, because transition functions for them are constant functions, which are continuous and even differentiable. In modern terminology, the distinction between continuous and discrete manifolds in Riemann's lecture would be interpreted as that of zero-dimensional manifolds and positive dimensional manifolds. I am grateful to Ken'ichi Ohshika for helping to clarify this point. It is not inconceivable that Riemann's thought along similar lines, which would explain his choice of the term manifold for both discrete and continuous manifolds, although the term had a more general use at the time. (Georg Cantor, possibly influenced by Riemann's lecture, initially referred to sets as *Mannigfaltigkeiten* but eventually switched to *Mengen*.) It is, however, difficult to be certain on the basis of his lecture or his other writings. I would argue that the difference between these two types of manifolds is still crucial, both in general and for Riemann, especially for his analysis of physical space and geometry. Riemann stressed the significance of the relationships between continuity and discontinuity for mathematics, physics, and philosophy (e.g. [32, pp. 515–524]; [9, pp. 77–80]).

The most defining feature of the concept of manifold (under the assumption than one can measure the length of line-segments, straight or curved) is that it is conceived as infinitesimally Euclidean. This makes a continuous manifold into a conglomerate of local, continuously connected, small open neighborhoods around each point. The concept of neighborhood, again, assumed to be infinitesimally flat and Euclidean, is a component-concept of the concept of manifoldness. This concept of manifold as composed out of local neighborhoods is extendable to a still more general concept of topological space, in which case local neighborhood need no longer be Euclidean and can be defined with a great degree of generality. It is true that Riemann did not have a concept of topological space, in contrast to his concept of a Riemann surface, which had a more direct and immediate impact on the development of topology as an independent mathematical discipline. I would argue, however, that the conceptual architecture defining topological spaces is a generalization of that of Riemann's conceptual architecture of manifoldness as a "space" composed of neighborhoods, or generalizing it even further to other "spaces," a conception that, as will be seen presently, extends to Grothendieck's topos theory.[16] Admittedly, this architecture is presented here in the spirit of modern axiomatic thinking rather than in the spirit of Riemann's conceptual-problematic thinking, but it does, I think, inherit Riemann's conceptual architecture defining his concept of manifold. The theory of Riemann surfaces, too, came to be recast in terms of manifolds. Weyl, who, as his title *The Concept of a Riemann Surface* stated, considered a Riemann surface to be a concept, was the first to expressly define Riemann surfaces as manifolds [20, 35]. Riemann, however, undoubtedly realized that they were manifolds, and they were part of the genealogy of the concept of manifold.

In the case of Riemannian manifolds, while each neighborhood is infinitesimally flat, Euclidean, the manifold as a whole is, in general, not, except in the special case of flat, Euclidean manifolds. A manifold may be negatively or positively curved, and, which is another major innovation of Riemann, this curvature can also be variable. Riemann defined the metric form as a quadratic differential form, by the only formula in his lecture (discounting the coordinate expression for the line element), and assumed that the transition from one local coordinate system to another was differentiable. Thus, he, again, de facto, considered differentiable manifolds with positive definite metrics, Riemannian manifolds. In modern terms, such a manifold is defined by using a differentiable section of positive-definite quadratic forms on the tangent-bundle. While, however, modern technical language can bring out deeper mathematical aspects of Riemann's concepts, it can also displace how Riemann thought, mathematically, physically, and, especially, philosophically, a displacement sometimes found in twentieth-century English translations of Riemann's works, including his Habilitation lecture. One is, accordingly, always in complex negotiations between Riemann's and contemporary technical language, even though and because Riemann is so often ahead of his time, so much our contemporary.

Another important and equally future-oriented conceptual aspect of Riemann's approach is that it allows one to define a geometrical or, more generally, topological space (in modern terminology) not as a multiple, say, a set of points, but as a space that could be covered by maps (Euclidean in the case of manifolds) and in its relation to other spaces. As just explained, in part following Riemann's way of thinking, topology describes a given space not only in terms of its points, continuously connected to each other, but also and most essentially in terms of its open neighborhoods around each point. These neighborhoods are subspaces of this space, the idea that, again, underlies Riemann's concept of manifold, in this case, however, giving each neighborhood a Euclidean geometry. The approach, again, enabled Riemann to define manifolds of any dimension, even infinite-dimensional ones, in terms of its inner properties rather than in relation to the ambient Euclidean space, where a manifold could be placed, against the flat Euclidean background. It is true that, if one appeals, as is usual even in considering Riemann, to open sets, this concept of space retains the concept of set (of points) as a primitive concept.[17] Riemann's way of thinking concerning manifolds, however, also suggests a possibility of thinking of and even defining a space in terms of its relations to other spaces, which allows one to use this structure as more primordial by replacing covering a space by *open sets (of points)* with covering it by *open spaces*. A topological space, defined above in set-theoretical terms, becomes a collection of open spaces as sub-spaces with certain (algebraic) rules for the relationships between them.

---

[16] It is also worth recalling in this connection that Grothendieck's initial primary areas of mathematical research concerned topological vector spaces, which suggests yet another genealogical line in the history of the (broadly) Riemannian problematics in question here.

This way of defining space by its relation to other spaces (as opposed to their constitution as sets of points) leads all the way to Grothendieck's topos theory, inspired by Riemann's ideas of manifolds and of the so-called covering spaces, originating in Riemann's theory of Riemann surfaces. Although it extends far beyond the question of spatiality, including mathematical logic, topos theory is arguably the farthest and most abstract extension of the concept of spatiality available, if one can rigorously speak of spatiality in this case, given an essentially algebraic nature of the concept. It does, however, give important new dimensions to our understanding of spatiality, when we deal in mathematics and elsewhere with objects or concepts that are considered in spatial terms.

It would not be possible here to present topos theory in its proper abstractness and rigor, sometimes prohibitive even for those not trained in the field of algebraic geometry or mathematical logic, where the concept is used as well (e.g. [18]). The essential philosophical ideas involved may, however, be sketched, as an example of both a rich mathematical concept in its own terms and of Riemann's influence on modern mathematics.[18] First, very informally, consider the following way of endowing a space with a structure, generalizing the definition of topological space. One begins with an arbitrarily chosen space, X, potentially any given space, which may initially be left unspecified in terms of its properties and structure. What would be specified are the relationship between spaces applicable to X, such as mapping or covering one or a portion of one, by another. One calls this structure the arrow structure $Y \rightarrow X$ (X is the space under consideration), where the arrow designates the relationship(s) in question. One can also generalize the notion of neighborhood or of an open subspace of (the topology of) a topological space in this way, by defining it as a relation between a given point and space (a generalized neighborhood or open subspace) associated with it. This procedure enables one to specify a given space not in terms of its intrinsic structure (e.g., a set of points with relations among them) but sociological[ly], throughout its relationships with other spaces of the same category, say, that of Riemannian spaces as manifolds [17], p. 7]. Some among such spaces may play a special role in defining the initial space, *X*, and algebraic structures (such as homotopy and cohomology, as Riemann realized in the case of covering spaces over Riemann surfaces. Indeed, the concept of covering space was one of the main inspirations for Grothendieck's concept of topos. The so-called *étale* topos (of a scheme), one of the main motivations for the concept of topos, is directly linked to the concept of covering space, as the term *étale* suggests.

---

[17] Thus, Ferreirós's discussion of Riemann in [11, pp. 39–80] appears to me to displace Riemann's thinking into the axiomatic and set-theoretical register, dominant in the wake of Cantor, a displacement arguably due to Ferreirós's insufficient attention to the nature of Riemann's mathematical concepts, to Riemann's concept of mathematical concept. In fairness, Ferreirós does relate Riemanns view of axioms to his concept of "hypothesis" and distinguishes it from the understanding of axioms developed in the twentieth-century philosophy of mathematics and mathematical logic. It does not appear to me, however, that Riemann thinks either in terms of axioms or, especially, in terms of sets (of points), as Ferreirós contends, although it could be and subsequently has been translated into these terms (e.g. [22]). See also Note 12 above.

[18] It would be instructive on both counts, to consider, as part of this genealogy, Dedekind's and Noether's work in algebra, reflecting the impact of Riemann's work on modern algebra, and even apart from his work on the distribution of primes and his famous hypothesis concerning the ζ-function. See [19] on Noether's work in this connection.

To make this scheme more rigorous and to explain (albeit still quite informally) the concept of topos, I need to explain in my own words category theory. It was introduced in as part of the cohomology theory in algebraic topology in 1940 and later extensively used by Grothendieck in his approach to algebraic geometry, leading to the concept of topos. Category theory considers multiplicities (which need not be sets) of mathematical objects conforming to a given concept, such as the category of Riemannian manifolds, and the arrows or morphisms, the mappings between these objects that preserve this structure. Studying morphisms allows one to learn about the individual objects involved, often to learn more than we would by considering them only or primarily individually. In a certain sense, by appealing to the conceptual determination of each manifold, Riemann already thinks categorically. Thus, one does not have to, and Riemann does not, start with a Euclidean space, whether seen in terms of sets of points or otherwise. Instead the latter is just one specifiable object of a large categorical multiplicity, here that of the category of Riemannian manifolds, an object marked by a particularly simple way we can measure the distance between any two points. Categories themselves may be viewed as such objects, and in this case one speaks of "functors" rather than "morphisms." Topology relates topological or geometrical objects, such as manifolds, to algebraic ones, especially, as in the case of homotopy and cohomology theories, groups, a concept, it is true, not used by Riemann, as against Poincaré, who made it central to his geometrical and topological thinking, which established his uniquely significant role in the rise of algebraic topology. Thus, in contrast to geometry (which relates its spaces to algebraic aspects of measurement), topology, almost by its nature, deals with functors between categories of topological objects, such as manifolds, and categories of algebraic objects, such as groups.

Now, a topos in Grothendieck's sense is a category of spaces and arrows over a given space, used especially for the purpose of allowing one to define richer algebraic structures associated with this space, as explained above. There are certain additional conditions such categories must satisfy, but this is not essential at the moment. To give a simple example, for any topological space *S*, the category of sheaves on *S* is a topos. The concept of topos is, however, very general, and extends far beyond spatial or space-like mathematical objects (thus, the category of finite sets is a topos); indeed it replaces the latter with a more algebraic structure of categorical and topos-theoretical relationships between objects. On the other hand, it derives from the properties of and (arrow-like) categorical relationships between properly topological objects, such as Riemann surfaces or manifolds. The conditions, mentioned above, that categories that form topoi must satisfy have to do with these connections. The concept of topos is especially suited to deal in the way we do with standard manifolds with objects, such as certain (discrete) algebraic varieties, which are solutions of polynomial equations and are space-like, that cannot be meaningfully defined otherwise sufficiently analogously to continuous spaces, specifically in order to define nontrivial cohomology or homotopy groups for them. This had been an outstanding problem of algebraic geometry, arising from the so-called Weil conjectures for algebraic varieties over finite fields, which was solved with the help of topos theory, specifically the concept of *étale* topos, mentioned above (e.g. [1]). What both types of objects now share are analogously defined topoi associated with them and, as a result, analogously defined algebraic structures associated with them, equally enabling the necessary functoriality in both cases.

Topos theory allows for such esoteric constructions as non-trivial or non-punctual single-point "spaces" or, conversely, spaces (topoi) without points (first constructed by Pierre Deligne), sometimes slyly referred to by mathematicians as "pointless topology." Philosophically, this notion is far from pointless, especially if considered within the overall topos-theoretical framework. In particular, it amplifies a Riemannian idea that "space," especially is defined by its relation to other spaces, as a more primary object than a "point" or, again, a "set of points." Space becomes a Leibnizean, "monadological" concept, insofar as points in such a space (when it has points) may themselves be seen as a kind of monads, thus also giving a non-trivial structure to single-point spaces. These monads are certain elemental but structured entities, spaces, rather than structure-less entities (classical points), or at least as entities defined by (spatial) structures associated to and defining them [3]. Naturally, my appeal to monads here is qualified and metaphorical. Leibniz's monads are elemental souls, the atoms of soul-ness, as it were. But one might say that the space thus associated to a given point is the soul of this point, which defines its nature or structure, not unlike an infinite-dimensional Hilbert space associated with an elementary particle, such as an electron, in quantum mechanics and enabling us to predict its behavior. In other words, not all points are alike insofar as the mathematical (and possibly philosophical) nature of a given point may depend on the nature or structure of the space or topos to which it belongs or with which it is associated in the way just described. This approach also gives a much richer architecture to spaces with multiple points, such as Riemann's manifold (in which this architecture is inherent), and one might see (with caution) such spaces as analogous to Leibniz's universe composed by monads. It also allows for different (mathematical) universes associated with a given space, possibly a single-point one, in which case a monad and a universe would coincide. Grothendieck's topoi are such possible universes, possible worlds, or even com-possible worlds in Leibniz's sense, without assuming, like Leibniz (in dealing with the physical world), the existence of only one of them, the best possible one.

The outcome of Riemann's investigation into the foundations of geometry was, thus, a new mathematics of great generality, power, and potential, which involved not only new geometry, but also new topology and analysis (the tensor calculus on manifolds). Although it was Riemann's theory of continuous or, again, differentiable manifolds that had the greatest impact, the concept of discrete manifolds was important for Riemann's argument, and it is important for the modern understanding of both spatiality and geometry in mathematics, physics, and philosophy. While a discrete manifold has topological dimension zero, it may still be seen as multiply extended, if defined as forming a very fine lattice with very small intervals between points, which can be "filled," as it were, to form a continuous space of the corresponding topological dimensions. It is also possible to introduce metrical relations for discrete manifolds. This concept is important in the context of the relationships between physical, dynamical forces in nature and the nature of space or, to return to Riemann's terms, the physical "reality underlying space," although Riemann is cryptic on such metrical relations. Crucially, however, he does allow for the possibility that the physical "reality underlying space" might be "a discrete manifold" [31, p. 33]. This possibility has been entertained even before Riemann and has been even more often considered since, especially more recently. Also, mathematically, finite geometries were beginning to be developed, usually in more axiomatic ways, around Riemann's time as well, later on also under the impact of his geometrical thinking.[19] As indicated earlier, Riemann saw the relationships between continuity and discontinuity as foundationally central to mathematics, physics, and philosophy (e.g., [32, pp. 515–524]; [10, pp. 77–80]), a view confirmed by the subsequent developments in the foundations of mathematics, from Dedekind and Cantor on, and quantum physics. The latter uses continuous (technically, differential) mathematics to predict, in probabilistic terms, irreducibly discrete phenomena, that is, phenomena that are not, and that possibly cannot be, assumed to be connected to each other by a continuous physical process [27, pp. 232].

4 Physics: "The Reality Underlying Space"

As noted from the outset, Riemann's contribution to physics in his lecture was as important as his contribution to mathematics there. Riemann's terms space and geometry refer to the three-dimensional space and its geometry, in accordance with the use of these terms at the time, although this was soon to change. We now speak not only of manifolds of any dimensions, as Riemann does, but also of their geometry and refer to them as "spaces," or of discrete spaces and geometries, without necessarily assuming any connections between these objects and physical (or phenomenal) spatiality. While Riemann allows that "the reality underlying space" may prove to be discrete at a very small scale, this is not the same as extending, as was done subsequently, the terms "space" and "geometry" to finite entities defined by certain geometrical-like properties. On the other hand, closer to Riemann's view of what "the reality underlying space" could in principle be, some physical theories, dealing with such connections, suggest that the ultimate reality underlying space might be discrete or, as in superstring and brane theories, that physical space has a dimension

---

[19] For the discussion of some of these developments, see the chapter by V. Pambuccian, H. Struve, and R. Struve [21] and other chapters in the part of this volume that addresses later developments of Riemann's work.

Pag. 340(367) →

Gauss's Curvature,
Riemann's manifolds,
Dedekind's ideals,
and
Grothendieck's topoi in
mathematics,
or
Einstein's spaces curved by
gravity in general relativity and
Heisenberg's matrix variables
or
Dirac's spinors in quantum
physics.

Pag. 344(371) — In his famous extraordinary theorem (theorema egregium as he called it), Gauss proved that the curvature of a surface which he defined as well, was intrinsic to the surface (surface).

Pag. 543(563) →

The sectional curvature K(Pi) agree with the Gauss curvature at P of the surface ruled by geodesic starting at p with tangent vector $X \in J$.

— Eighteen Essays in Non-Euclidean

$$\mathrm{Curv}^g(X,Y)Z = \left(\nabla_X\nabla_Y - \nabla_Y\nabla_X - \nabla_{[X,Y]}\right)Z, \qquad X,Y,Z \in \mathfrak{X}(\mathcal{M})$$

where ∇ is the metric-compatible, torsion-free connection associated to g

$$\mathrm{Riem}^g(W,Z,X,Y) = \langle W, \mathrm{Curv}^g(X,Y)Z\rangle, \qquad W,X,Y,Z \in \mathfrak{X}(\mathcal{M})$$

where $\langle\,,\,\rangle$ denotes scalar product with g. In terms of this tensor Riemann's characterization of flat space is

Theorem 2.1 (Local characterization of flat space, Riemann [47]) *A Riemannian manifold $(\mathcal{M},g)$ is locally isometric to the flat space with the same dimension and signature if and only if $Riem^g = 0$.*

Being intended to all faculty members, Riemann's lecture contains hardly any formula (see Chap. 4 in Spivak's comprehensive treatise [73] for a detailed exposition of Riemann's lecture) so no detailed proof of this claim is given there. Not even an explicit form of the Riemann tensor in a general coordinate system appears there. Following Spivak [73], such an explicit expression, and a proof of the fact that its non-vanishing is an obstruction for a metric being locally flat, appeared first in an unpublished paper, known as the *Commentatio*, submitted in 1861 to the Paris Academy and which appeared in the second edition of the collected works of Riemann [62]. However, as emphasized in [26] (see also [34]), this unpublished paper was primarily concerned with a problem of heat conduction proposed by the Academy in 1858, so its primary concern was not in Geometry, and it is debatable whether our modern understanding of the field might be biasing the interpretation of its results (see the interesting discussion on this *Commentatio* in [34] and references therein). In any case, an explicit expression of what we now know as the Riemann tensor appeared for the first time in this unpublished paper by Riemann. The first explicit proof that the vanishing of this tensor is sufficient for a metric being (locally) flat was given by Elwin Bruno Christoffel [17]. The proof of Theorem 2.1 is so central in differential geometry that many different proofs have been given. Spivak's treatise includes seven explicit proofs, which very nicely illustrate how differential geometry has evolved over the years from a discipline where the use of coordinates was central, to a mature theory where geometric objects fully independent of coordinates are the basic entities.

Another fundamental local characterization of Riemannian manifolds involves manifolds of constant sectional curvature. Recall that the sectional curvature of a two-plane $\Pi \subset T_p\mathcal{M},\ p \in (\mathcal{M},g)$ with non-degenerate induced first fundamental form h is defined as

$$K(p;\Pi) = \frac{\mathrm{Riem}^g(X,Y,X,Y)}{h(X,X)h(Y,Y) - h(X,Y)^2}, \qquad \{X,Y\} \text{ any basis of } \Pi.$$

The sectional curvature $K(p;\Pi)$ agrees with the Gauss curvature at p of the surface ruled by geodesics starting at p with tangent vector $X \in \Pi$. A Riemannian manifold has constant curvature k if $K(p;\Pi) = k$ for all $p \in \mathcal{M}$ and all non-degenerate plane $\Pi \in T_p\mathcal{M}$. A well-known lemma by Friedrich Schur [70] asserts that any Riemannian manifold with sectional curvature $K(p, \Pi) = k(p)$ independent of Π is in fact of constant curvature.

For any $k \in \mathbb{R}$, define the manifold $\mathbb{M}_k^{(p,q)}$ as the subset $B(K) \subset \mathbb{R}^{p+q}$

$$B(K) := \{x \in \mathbb{R}^{p+q};\, 1 + \frac{k}{4}(\eta_{ij}^{(p,q)} x^i x^j) > 0\},$$

endowed with the metric g_k given by

$$\begin{aligned} g_k &:= \frac{1}{\left(1 + \frac{k}{4}(\eta_{ij} x^i x^j)\right)^2} \eta_{ij}^{(p,q)} dx^i dx^j, \\ \eta^{(p,q)} &:= \mathrm{diag}(\underbrace{-1, \ldots, -1}_{p}, \underbrace{1, \ldots, 1}_{q}) \end{aligned}$$

(it is easily seen that $B(K)$ is connected for all values of k and all signatures (p, q)). Spaces of constant curvature are all isometric to this space. This is already mentioned without proof in Riemann's inaugural lecture [47] and it was also known to Eugenio Beltrami [4].

Theorem 2.2 *Let* $(\mathcal{M}, g)$ *be a Riemannian manifold of constant sectional curvature* k *and signature* (p, q)*. Then* $(\mathcal{M}, g)$ *is locally isometric to* $\mathbb{M}_k^{(p,q)}$.

Riemannian manifolds of constant curvature are also locally characterized by the condition that the Killing algebra of $(\mathcal{M}, g)$ has maximal dimension. Recall that a Killing vector is a vector field $\xi \in \mathfrak{X}(\mathcal{M})$ satisfying $\mathcal{L}_\xi g = 0$ where $\mathcal{L}$ denotes Lie derivative. The set of Killing vectors $\mathcal{K}$ in a manifold defines a finite-dimensional algebra (with respect to the Lie bracket) called Killing algebra. If the manifold has dimension n, the Killing algebra has dimension $\dim(\mathcal{K}) \leq \frac{n(n+1)}{2}$ and whenever equality is attained, the manifold is called maximally symmetric. A classical result by Luigi Bianchi [6] asserts that maximally symmetric spaces are precisely those of constant sectional curvature.

Theorem 2.3 (Bianchi [6]) *A Riemannian manifold* $(\mathcal{M}, g)$ *is maximally symmetric if and only if it has constant sectional curvature.*

Another important class of Riemannian manifolds singled out by the properties of its isometry group is the class of Riemannian symmetric spaces. These are defined as Riemannian manifolds $(\mathcal{M}, g)$ such that, at any point $p \in \mathcal{M}$ there exists an isometry $\Psi_p : \mathcal{M} \to \mathcal{M}$, $\Psi_p^\star(g) = g$ satisfying $\Psi_p(p) = p$ and $d\Psi_p|_p = -\mathrm{Id}|_p$, i.e. this isometry leaves p invariant and reverses any vector at p. Riemannian symmetric spaces can be locally characterized in terms of the Riemann tensor, as first proved by Élie Cartan.

Theorem 2.4 (Locally symmetric spaces, Cartan [12]) *A Riemannian manifold* $(\mathcal{M}, g)$ *is locally isometric to a Riemannian symmetric space if and only if the curvature tensor satisfies*

$$\nabla Riem^g = 0.$$

The Riemann tensor can be decomposed in its trace and its trace-free parts. The trace part defines the Ricci tensor $\mathrm{Ric}^g(X, Y) = \mathrm{tr}_g \mathrm{Curv}^g(X, \cdot)Y$ from which the scalar curvature is defined as $\mathrm{Scal}^g = \mathrm{tr}_g \mathrm{Ric}^g$. Useful tensors defined in terms of the Ricci tensor are the Schouten tensor Sch^g (in dimension $n > 2$) and the Einstein tensor Ein^g:

$$\mathrm{Sch}^g := \frac{1}{n-2}\left(\mathrm{Ric}^g - \frac{1}{2(n-1)}\mathrm{Scal}^g g\right), \qquad \mathrm{Ein}^g = \mathrm{Ric}^g - \frac{1}{2}\mathrm{Scal}^g.$$

The Weyl tensor is the trace-free part of the Riemann tensor, defined as

$$\mathrm{Weyl}_g = \mathrm{Riem}^g - \mathrm{Sch}^g \odot g$$

where $\odot$ is the Kulkarni-Nomizu product on symmetric tensors

$$\begin{aligned}(A \odot B)(X_1, X_2, X_3, X_4) =& A(X_1, X_3)B(X_2, X_4) - A(X_1, X_4)B(X_2, X_3)\\ &+ A(X_2, X_4)B(X_1, X_3) - A(X_2, X_3)B(X_1.X_4).\end{aligned}$$

Hermann Weyl introduced [79] the tensor named after him in an attempt to find a theory of gravitation invariant under scale (conformally invariant). The Weyl tensor has the fundamental property that $\mathrm{Weyl}_{\Omega^2 g} = \mathrm{Weyl}_g$ where $\Omega \in C^\infty(\mathcal{M}, \mathbb{R}^+)$. Thus, it vanishes automatically in any locally conformally flat space. Schouten [66] proved that the vanishing of this tensor is also sufficient in dimension $n \geq 4$ (in dimension $n = 2$ the Weyl tensor is not defined and in $n = 3$ it vanishes identically). As is well-known, in dimension $n = 2$ any space is locally conformally flat. In dimension three, the necessary and sufficient condition for local conformal flatness was obtained by Émile Cotton [16] in terms of the vanishing of the so-called Cotton tensor, which in dimension $n > 2$ is defined as

$$\mathrm{Cott}(X, Y, Z) = (n-2)\left((\nabla_Z \mathrm{Sch})(X, Y) - (\nabla_Y \mathrm{Sch})(X, Z)\right),$$

where $X, Y, Z \in \mathfrak{X}(\mathcal{M})$. This tensor is conformally invariant in any space with vanishing Weyl tensor (hence always in dimension three). Summarizing the discussion above, we have

Theorem 2.5 (Locally conformally flat spaces, Cotton, Weyl and Schouten) *A Riemannian manifold $(\mathcal{M}^n, g)$ $n \geq 3$ of signature (p, q) is locally isometric to a conformally flat space $(\mathbb{R}^n, \Omega^2\eta^{(p,q)})$, $\Omega \in C^\infty(\mathbb{R}^n, \mathbb{R}^+)$ if and only if*

(i) If $n = 3$, the Cotton tensor vanishes.
(ii) If $n \geq 4$, the Weyl tensor vanishes.

C +,-.+ + +/ ! + %" 9 9 !3 E 9 # !
2 3 9 0 ! ! ! ! 3) 0#
* ')' !0 &

Chapter 5

Inscribing a triangle in a circle in spherical geometry

Himalaya Senapati

Contents

1. Introduction 67
2. Lagrange's solution 68
3. Constructibility on the sphere 72
4. Extension of the problem to spherical geometry 75
References 78

1 Introduction

We consider the geometrical problem of constructing a triangle using a straightedge and compass so that its three vertices lie on a given circle and its three sides when produced pass through three given points. Pappus of Alexandria formulated and solved this problem in the special case where the three given points are aligned. This is Proposition 117 of Book VII of the *Collection* [1]. Lagrange provided an algebraic solution, proving that the vertices of the triangle are constructible points. He communicated his solution to Jean de Castillon, who reproduced it in the first section of his paper "Sur une nouvelle propriété des sections coniques" (On a new property on conic sections) [2]. In the rest of his this paper, Castillon proceeded to show that the vertices of the triangle remain constructible when the circle is replaced by an ellipse or a hyperbola. Later, the solution due to Lagrange was published in his *Oeuvres complètes* under the title "Solution algébrique d'un problème de géométrie" (Algebraic solution of a geometry problem) [3]. Euler gave another solution and commented on the case where the given circle and the three given points are on a sphere [4]. Interestingly, this article is followed by one due to Nicolas Fuss, a young collaborator of Euler, who similar to Pappus, studies the special case where the three given points lie on a line [5]. Eugène-Charles Catalan also studied this problem in his book *Théorèmes et problèmes de géométrie élémentaire* (Theorems and problems of elementary geometry) [6] where he says that such a problem is known as the "Castillon problem" (see the footnote at p. 222 of [6]).

In this chapter, we first recall the algebraic solution due to Lagrange. The idea behind such a solution is to prove that the vertices of the triangle are constructible. More precisely, it is to prove that the coordinates of the vertices are roots of some quadratic polynomials with constructible coefficients. Let us recall the notion of constructibility that is relevant here. Let $\mathcal{S}$ be a given set of points in the Euclidean plane with cardinality at least 2.[1] We say that a circle is *constructible from* $\mathcal{S}$ if its center is an element of $\mathcal{S}$ and its radius is the distance between two elements of $\mathcal{S}$. We say that a line is constructible from $\mathcal{S}$ if it passes through two elements of $\mathcal{S}$. We say that a point in the Euclidean plane is *constructible in one step* from $\mathcal{S}$ if it is an intersection point of two constructible lines, one constructible line and a constructible circle, or two constructible circles. We then say that a point P of the Euclidean plane is *constructible* from $\mathcal{S}$ if there exists a finite sequence of points $P_1, P_2, \ldots, P_n = P$ such that P_1 is constructible in one step from $\mathcal{S}$ and for any i $(2 \leq i \leq n)$, the point P_i is constructible in one step from $\mathcal{S} \cup \{P_1, \ldots, P_{i-1}\}$. In particular, let us denote two points from $\mathcal{S}$ by O and I and set the distance from O to I as 1. We can now introduce the Cartesian coordinate system in the Euclidean plane where the coordinates of O and I are $(0, 0)$ and $(1, 0)$ respectively. We say that a number x is *constructible from* $\mathcal{S}$ if it is a coordinate of a constructible point from $\mathcal{S}$. It is elementary to prove that the set of constructible numbers from $\mathcal{S}$ forms a field such that the square root of any non-negative constructible number from $\mathcal{S}$ is still a constructible number from $\mathcal{S}$. In particular, a root of a quadratic polynomial with constructible (from $\mathcal{S}$) coefficients is constructible.

In the next section, we recall Lagrange's solution. We then study the problem where the given circle and the three given points are on a sphere and give a solution in the lineage of Lagrange.

2 Lagrange's solution

The problem is to construct a triangle MNP inscribed in a given circle $\mathcal{C}$ of radius r so that the three edges of the triangle NM, MP, and PN when produced pass through three given points A, B, and C respectively (see Figure 1). This construction should not require more than the given objects and a straightedge and a compass. To be more precise, the set we are dealing with consists of six points, the two points O and I defined above, the point of coordinates $(r, 0)$, and the three points A, B, and C. Let us just add that the center of the given circle can be taken as the origin O.

---

[1] We are using a modern way to formulate constructibility. Needless to say, Lagrange did not use this language of set theory or of fields.

We notice that the vertices M, N, and P are constructible if the angles $x = \widehat{AOM}$, $y = \widehat{AON}$ and $z = \widehat{AOP}$ are constructible, where an angle is said to be constructible if its sine (or equivalently cosine) is constructible. We set the lengths of segments

$$OA = a, \quad OB = b, \quad OC = c.$$

We denote the angles $\widehat{AOB}$ and $\widehat{AOC}$ by m and n. Here, 1, a, b, c, $\cos(m)$, $\cos(n)$, and r are constructible. In the triangle AON, we have

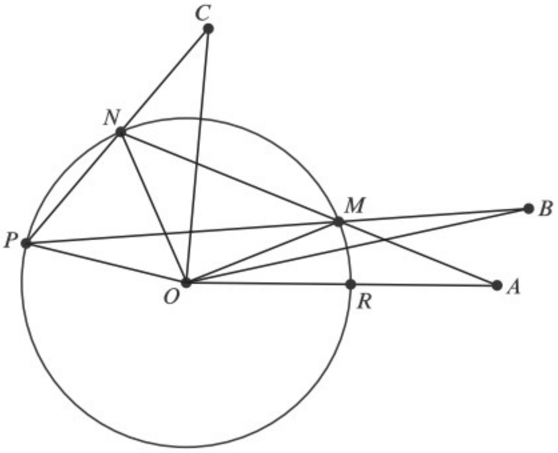

Figure 1. An inscribed triangle MNP such that NM, MP, and PN when produced pass through A, B, and C respectively

$$\widehat{ONA} = \frac{1}{2}(\pi - \widehat{NOM}) = \frac{\pi}{2} - \frac{y-x}{2}$$

and

$$\widehat{OAN} = \pi - \widehat{AON} - \widehat{ONA} = \frac{\pi}{2} - \frac{y+x}{2}.$$

Using the sine law in the triangle AON, we have

$$\frac{OA}{ON} = \frac{\sin(\widehat{ONA})}{\sin(\widehat{OAN})} \implies \frac{a}{r} = \frac{\cos\left(\frac{y-x}{2}\right)}{\cos\left(\frac{y+x}{2}\right)}.$$

This reduces to

$$\begin{aligned}
&a\Big(\cos\left(\frac{y}{2}\right)\cos\left(\frac{x}{2}\right)-\sin\left(\frac{y}{2}\right)\sin\left(\frac{x}{2}\right)\Big)\\
&\qquad=r\Big(\cos\left(\frac{y}{2}\right)\cos\left(\frac{x}{2}\right)+\sin\left(\frac{y}{2}\right)\sin\left(\frac{x}{2}\right)\Big)\\
&\implies (a-r)\cos\left(\frac{y}{2}\right)\cos\left(\frac{x}{2}\right)=(a+r)\sin\left(\frac{y}{2}\right)\sin\left(\frac{x}{2}\right)\\
&\implies \tan\left(\frac{x}{2}\right)\tan\left(\frac{y}{2}\right)=\frac{a-r}{a+r}.
\end{aligned}\tag{1}$$

In the triangle BOP, we find

$$\widehat{OPB}=\frac{\pi}{2}-\frac{z-x}{2}\quad\text{and}\quad\widehat{OBP}=\frac{\pi}{2}+m-\frac{z+x}{2}.$$

Using the sine law in the triangle BOP, we have

$$\frac{BO}{PO}=\frac{\sin(\widehat{OPB})}{\sin(\widehat{OBP})}\implies\frac{b}{r}=\frac{\cos\left(\frac{(z-m)-(x-m)}{2}\right)}{\cos\left(\frac{(z-m)+(x-m)}{2}\right)}$$

which reduces to

$$\tan\frac{x-m}{2}\tan\frac{z-m}{2}=\frac{b-r}{b+r}.\tag{2}$$

Similarly, in the triangle COP, we have

$$\widehat{OPC}=\frac{\pi}{2}-\frac{z-y}{2}\quad\text{and}\quad\widehat{OCP}=\frac{\pi}{2}+n-\frac{z+y}{2}.$$

This leads to

$$\tan\left(\frac{y-n}{2}\right)\tan\left(\frac{z-n}{2}\right)=\frac{c-r}{c+r}.\tag{3}$$

The goal is now to determine the unknowns x, y, and z from equations (1)–(3). For simplicity, let us set

$$\tan\left(\frac{x}{2}\right)=s,\quad\tan\left(\frac{y}{2}\right)=t,\quad\tan\left(\frac{z}{2}\right)=u,\quad\tan\left(\frac{m}{2}\right)=p,\quad\tan\left(\frac{n}{2}\right)=q,$$

$$\frac{a-r}{a+r}=\alpha,\quad\frac{b-r}{b+r}=\beta,\quad\frac{c-r}{c+r}=\gamma.$$

Here, α, β, γ, p, and q are constructible since a, b, c, m, n, and r are constructible. Equations (1)–(3) now become

$$s\times t=\alpha,\quad\frac{s-p}{1+ps}\times\frac{u-p}{1+pu}=\beta,\quad\frac{t-q}{1+qt}\times\frac{u-q}{1+qu}=\gamma.$$

The first and the second give

$$t = \frac{\alpha}{s} \quad \text{and} \quad u = \frac{\beta(1+ps) + p(s-p)}{(s-p) - \beta p(1+ps)}. \tag{4}$$

Substituting the first and the second in the third gives

$$\frac{\alpha - qs}{\alpha q + s} \times \frac{p(-p+q) + \beta(1+pq) + (p - q + \beta p(1+pq))s}{-(1+\beta)p + \beta q - p^2 q + (1 - \beta p^2 + (1+\beta)pq)s} = \gamma. \tag{5}$$

For conciseness, we set

$$F = p(-p+q) + \beta(1+pq) \quad \text{and} \quad G = p - q + \beta p(1+pq).$$

We also set

$$H = -(1+\beta)p + \beta q - p^2 q \quad \text{and} \quad K = 1 - \beta p^2 + (1+\beta)pq.$$

Therefore, F, G, H, and K are constructible since β, p, and q are constructible. Equation (5) now becomes

$$\frac{\alpha - qs}{\alpha q + s} \times \frac{F + Gs}{H + Ks} = \gamma$$

which implies

$$(\gamma K - Gq) \times s^2 + (\gamma H - \alpha G + \alpha q(\gamma K - F)) \times s + \alpha(\gamma H q - F) = 0. \tag{6}$$

The coefficients of this quadratic equation in s are constructible, hence s is also constructible. This implies that the right-hand sides of the two equations in (4) are constructible and thus implying that t and u are constructible. Thus

$$\cos(x) = \frac{1 - s^2}{1 + s^2}, \quad \cos(y) = \frac{1 - t^2}{1 + t^2}, \quad \cos(z) = \frac{1 - u^2}{1 + u^2}$$

are constructible. For the existence of the vertices M, N, and P, the angles $\frac{x}{2}$, $\frac{y}{2}$, and $\frac{z}{2}$ and hence the tangents s, t, and u should be real. When s is real, (4) implies $t, u \in \mathbb{R}$. Hence we have proved that when equation (6) admits real roots, a triangle exists and is constructible so that its vertices lie on a given circle and its three edges when produced pass through three given points. When equation (6) does not admit real roots, such a triangle does not exist.

Remark 2.1. We now give an example where such a triangle does not exist. Let the three given points and the centre of the given circle be collinear such that $m = n = 0$. This implies

$$p = \tan\left(\frac{m}{2}\right) = 0 \quad \text{and} \quad q = \tan\left(\frac{n}{2}\right) = 0.$$

Substituting these values, we find $F = \beta$, $G = 0$, $H = 0$, and $K = 1$. Thus, equation (6) becomes

$$\gamma \times s^2 - \alpha\beta = 0$$

with discriminant

$$4\alpha\beta\gamma = 4 \times \frac{a-r}{a+r} \times \frac{b-r}{b+r} \times \frac{c-r}{c+r}.$$

This is negative when only one of the three points is in the interior of the circle or when all three are in the interior. Thus, by continuity, there exist approximately collinear points ($m \approx n \approx 0$, $m, n \neq 0$) with either only one of the points in the interior or all three in the interior such that the discriminant is negative. In all these cases, the triangle to be inscribed does not exist.

3 Constructibility on the sphere

We define here a notion of constructibility on the sphere. This requires the use of an auxiliary Euclidean plane. Let us assume that the sphere is of radius 1. We report on the auxiliary plane two points at distance 1. Let us denote these two points by O and I. Now, let $\mathcal{S}_+$ be a set of points on the sphere. On the plane, we draw circles centered at O of radius the chord distance between two points of the given set. We set $\mathcal{S}_0$ as the set which consists of the two points O, I and the intersection points of those circles with the half-line starting at O and passing through I. We say that a circle is *constructible* from $\mathcal{S}_+$ if its pole is a point of $\mathcal{S}_+$ and the chord distance to the pole is a constructible number from $\mathcal{S}_0$. Like for the case of constructibility in the plane, we say that a point on the sphere is *constructible in one step* from $\mathcal{S}_+$ if it is an intersection point of two constructible circles. Thus, a point P on the sphere is called constructible if there exists a finite sequence of points $P_1, P_2, \ldots, P_n = P$ such that P_1 is constructible in one step from $\mathcal{S}_+$ and for $2 \leq i \leq n$, P_i is constructible in one step from $\mathcal{S}_+ \cup \{P_1, \ldots, P_{i-1}\}$.

Following our definition of constructibility on the sphere, we state some consequences which shall be useful in Section 4.

Lemma 3.1. *A great circle on the sphere is constructible if one of its poles is constructible.*

Proof. Let C be the centre of the sphere and P the constructible pole (see Figure 2). Let Q be an arbitrary point on the great circle with pole P. This implies $PC = QC = 1$ and $\widehat{PCQ} = \frac{\pi}{2}$ so that $PQ = \sqrt{2}$. Thus the chord distance of a point on the great circle from its pole is constructible on the plane implying that the great circle is constructible □

Lemma 3.2. *A geodesic through two given points on the sphere is constructible.*

Proof. Let P and Q be the two given points (see Figure 3). We construct the two great circles whose poles are P and Q respectively. Let a point of intersection be called C'. We have the chord lengths $PC' = QC' = \sqrt{2}$. Thus the great circle whose pole is C' passes through P and Q. Since C' is constructible, geodesic through P and Q is constructible. □

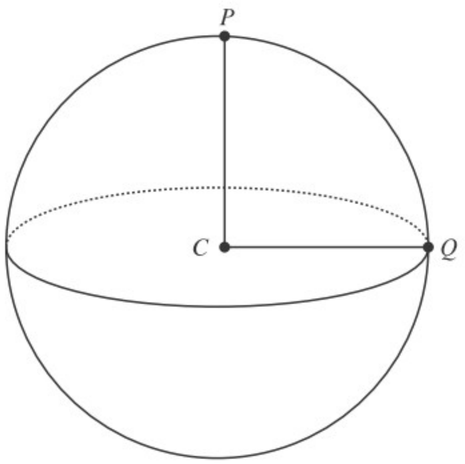

Figure 2

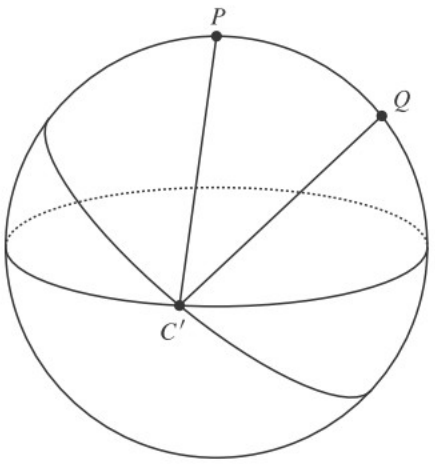

Figure 3

Lemma 3.3. *The sine of the angular separation between two constructible points on the sphere is constructible in the Euclidean plane.*

Proof. Let D and E be two given points on the sphere (see Figure 4). Construct the geodesic DE and extend it. Using a compass, find the point F on the extended geodesic so that $\mathrm{crd}(DE) = \mathrm{crd}(EF)$ where crd denotes the chord length. We have $\mathrm{crd}(DF) = 2\sin(DE)$. Since the chord distance on the sphere is constructible on the plane, $\sin(DE)$ is constructible in the Euclidean plane. ☐

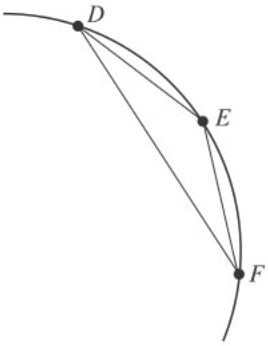

Figure 4. Determination of the sine of the angular separation between two constructible points

Lemma 3.4. *Suppose an angle θ is constructible in the Euclidean plane. Let D and E be two points on the sphere. Then, it is possible to construct a great circle arc DG such that $\widehat{EDG} = \theta$.*

Proof. Construct the great circle $\mathcal{C}_D$ with pole D. Extend the geodesic DE in the direction of E and denote the first point of intersection with this great circle by F (see Figure 5). Since θ is constructible in the Euclidean plane, $2\sin\left(\frac{\theta}{2}\right)$ is also constructible. Construct the circle $\mathcal{C}_F$ with pole F and chord distance to the pole set at $2\sin\left(\frac{\theta}{2}\right)$. Let G be a point of intersection of $\mathcal{C}_D$ and $\mathcal{C}_F$ so that $\mathrm{crd}(FG) = 2\sin\left(\frac{\theta}{2}\right)$. Thus the great circle arc FG subtends an angle θ at the centre of the sphere. Therefore, it also subtends the same angle θ at D since D is the pole of the great circle arc FG. Hence, joining DG, we have $\widehat{FDG} = \theta$. ☐

Lemma 3.5. *Suppose DE and DG are two great circle arcs on the sphere. The angle $\widehat{EDG}$ is constructible in the Euclidean plane*

Proof. The proof is very similar to that of Lemma 3.4. ☐

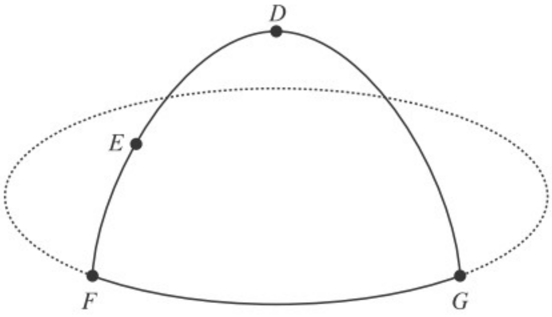

Figure 5. Given DE, constructing DG such that the angle at D is a given angle

4 Extension of the problem to spherical geometry

We now consider the extension of this construction problem to spherical geometry. Here, the problem remains to construct a triangle MNP using a straightedge and a compass whose vertices lie on a given circle $\mathcal{C}$ and whose edges when produced pass through three given points A, B, and C. However, the given circle and the three given points now lie on a sphere instead of the Euclidean plane (see Figure 6).

Let O be a pole of the circle $\mathcal{C}$. We set, as before, $OM = ON = OP = r$, $OA = a$, $OB = b$, $OC = c$, $\widehat{AOB} = m$, and $\widehat{AOC} = n$. We denote the unknown angles $\widehat{AOM}$, $\widehat{AON}$ and $\widehat{AOP}$ by x, y, and z respectively.

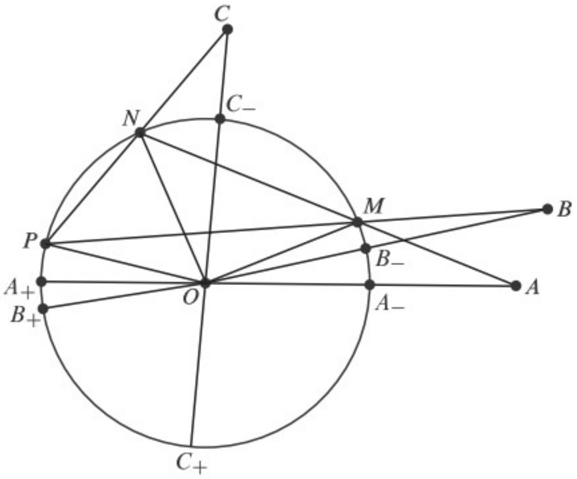

Figure 6. The circle $\mathcal{C}$ is on the surface of a sphere

By Lemma 4.2 (to be given below), the geodesic AMN intersecting the circle $\mathcal{C}$ at M and N gives the relation

$$\tan\left(\frac{x}{2}\right)\tan\left(\frac{y}{2}\right) = \frac{\sin(a-r)}{\sin(a+r)} = \alpha_+.$$

This is reminiscent of equation (1). Similarly, the geodesics BMP and CNP give identities reminiscent of equations (2) and (3):

$$\tan\left(\frac{x-m}{2}\right)\tan\left(\frac{z-m}{2}\right) = \frac{\sin(b-r)}{\sin(b+r)} = \beta_+$$

and

$$\tan\left(\frac{y-n}{2}\right)\tan\left(\frac{z-n}{2}\right) = \frac{\sin(c-r)}{\sin(c+r)} = \gamma_+.$$

As before, we set

$$\tan\left(\frac{x}{2}\right) = s, \quad \tan\left(\frac{y}{2}\right) = t, \quad \tan\left(\frac{z}{2}\right) = u, \quad \tan\left(\frac{m}{2}\right) = p, \quad \tan\left(\frac{n}{2}\right) = q.$$

Since the angles m and n are constructible on the sphere, by Lemma 3.5, they are constructible on the plane. Hence, $p = \tan\left(\frac{m}{2}\right)$ and $q = \tan\left(\frac{n}{2}\right)$ are also constructible on the plane. We also set F, G, H, and K:

$$\begin{aligned} F &= p(-p+q) + \beta_+(1+pq), \\ G &= p - q + \beta_+ p(1+pq), \\ H &= -(1+\beta_+)p + \beta_+ q - p^2 q, \\ K &= 1 - \beta_+ p^2 + (1+\beta_+)pq. \end{aligned}$$

With this notation, we obtain the equations

$$t = \frac{\alpha_+}{s}, \quad u = \frac{\beta_+(1+ps) + p(s-p)}{(s-p) - \beta_+ p(1+ps)}, \quad \frac{\alpha_+ - qs}{\alpha_+ q + s} \times \frac{F+Gs}{H+Ks} = \gamma_+. \tag{7}$$

It is clear that s, t, and u are constructible in the Euclidean plane if α_+, β_+, and γ_+ are constructible on the plane. Extending the great circle arcs OA, OB, and OC, and denoting the point of intersections with the circle by $A_\pm$, $B_\pm$, and $C_\pm$ as shown in Figure 6, we find $AA_\pm = a \pm r$, $BB_\pm = b \pm r$, and $CC_\pm = c \pm r$ that are constructible on the sphere. Hence, by Lemma 3.3, $\sin(a \pm r)$, $\sin(b \pm r)$, and $\sin(c \pm r)$ are constructible on the plane that implies α_+, β_+, and γ_+ and hence s, t, and u are constructible in the Euclidean plane. As shown before in the planar case, when $s \in \mathbb{R}$, the angles x, y, and z are real and are constructible in the Euclidean plane. By Lemma 3.4, we can construct geodesics OM, ON, and OP so that they make angles x, y, and z with OA so that the vertices M, N, and P are constructible. When $s \notin \mathbb{R}$, the required triangle does not exist. We have thus generalized the algebraic solution due to Lagrange to the spherical setting.

Remark 4.1. We give an example where the triangle to be inscribed does not exist. Consider the limiting case where the three given points A, B, and C approach O that is the pole of the given circle. In this limit, OA, OB, and OC vanish so that α_+, β_+, and γ_+ approach -1. Similarly, when A, B, and C approach the antipodal point of O, α_+, β_+, and γ_+ approach -1. In both limits,

$$-F \approx K \approx 1 + p^2 \quad \text{and} \quad G \approx H \approx -q(1 + p^2).$$

The quadratic equation in s (third equation of (7)) now becomes

$$\left(\frac{1 + qs}{q - s}\right)^2 \approx -1$$

that has no real roots. Thus, there exist A, B, and C near O (or near the antipodal point of O) so that the inscribed triangle does not exist.

It remains to prove the following lemma.

Lemma 4.2. *Let $\mathcal{C}$ be a small circle on a sphere with pole O. Let A be a point on the sphere. Let a geodesic through A intersect $\mathcal{C}$ at points M and N. Then, we have*

$$\tan\left(\frac{1}{2}\widehat{AOM}\right)\tan\left(\frac{1}{2}\widehat{AON}\right) = \frac{\sin(OA - OM)}{\sin(OA + OM)}.$$

Proof. For conciseness, let us set $OM = r$, $OA = a$, $\widehat{AOM} = x$, and $\widehat{AON} = y$ (see Figure 7). By Lemma 2.3 in [7], in the triangle AOM, we have

$$\tan(\widehat{OAM}) = \frac{2\sin(OM)}{\tan\left(\frac{1}{2}\widehat{AOM}\right)\sin(OA + OM) + \cot\left(\frac{1}{2}\widehat{AOM}\right)\sin(OA - OM)},$$

while in the triangle AON, we have

$$\tan(\widehat{OAN}) = \frac{2\sin(ON)}{\tan\left(\frac{1}{2}\widehat{AON}\right)\sin(OA + ON) + \cot\left(\frac{1}{2}\widehat{AON}\right)\sin(OA - ON)}.$$

Since $\widehat{OAN} = \widehat{OAM}$ and $OM = ON$, this implies

$$\begin{aligned}\sin(a + r)\tan\left(\frac{x}{2}\right) + \sin(a - r)\cot\left(\frac{x}{2}\right)\\ = \sin(a + r)\tan\left(\frac{y}{2}\right) + \sin(a - r)\cot\left(\frac{y}{2}\right).\end{aligned}$$

Solving for $\tan\left(\frac{x}{2}\right)$ in terms of $\tan\left(\frac{y}{2}\right)$ in this quadratic equation, we find

$$\tan\left(\frac{x}{2}\right) = \tan\left(\frac{y}{2}\right) \quad \text{or} \quad \tan\left(\frac{x}{2}\right)\tan\left(\frac{y}{2}\right) = \frac{\sin(a-r)}{\sin(a+r)}.$$

When $\tan\left(\frac{x}{2}\right) = \tan\left(\frac{y}{2}\right)$, x and y differ by an integer multiple of 2π implying that M and N coincide. Here, AM is tangent to the circle, hence $\widehat{AMO} = \frac{\pi}{2}$. Thus, we have $\cos(x) = \cot(a)\tan(r)$. This implies

$$\tan\left(\frac{x}{2}\right)\tan\left(\frac{y}{2}\right) = \tan^2\left(\frac{x}{2}\right) = \frac{1-\cos x}{1+\cos x} = \frac{\sin(a-r)}{\sin(a+r)}. \qquad \square$$

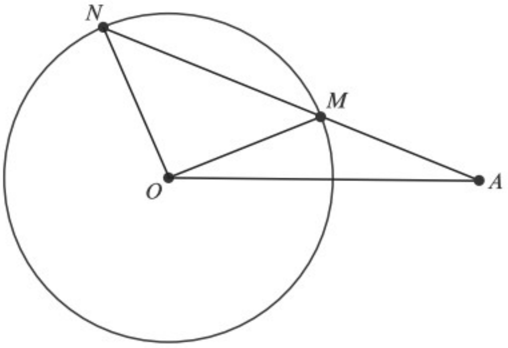

Figure 7

Acknowledgements. The author would like to thank Athanase Papadopoulos for suggesting this problem and translating the original solution due to Lagrange. The author would like to thank Vincent Alberge for critically reviewing this paper, correcting errors and suggesting significant improvements. The author acknowledges the hospitality of ICTS/Prog-ggd/2017/11 where part of this work was done.

References

[1] Pappus d'Alexandrie, *La collection mathématique,* œuvre traduite pour la premiere fois du grec en français par Paul Ver Eecke, Paris, Bruges, Desclée de Brouwer, 1933.
☞ 67

[2] J. de Castillon, “Sur une nouvelle propriété des sections coniques,” *Nouveaux mémoires de l’Académie royale des Sciences et Belles-Lettres de Berlin* 1776, 294–311.
☞ 67

[3] J.-L. Lagrange, “Solution algébrique d’un problème de géométrie,” in *Œuvres de Lagrange*, Vol. IV, Gauthier-Villars, Paris, 1868, 335–339.
☞ 67

[4] L. Euler, “Problematis cuiusdam Pappi Alexandrini constructio,” *Acta academiae scientarum imperialis Petropolitanae* 1 (1780), 91–96. Reprinted in *Opera omnia,* Series 1, Vol. XXVI, 237–242.
`http://eulerarchive.maa.org/pages/E543.html`
☞ 67

[5] N. Fuss, “Solutio problematis geometrici Pappi Alexandrini,” *Acta academiae scientarum imperialis Petropolitanae* 1 (1780), 97–104.
☞ 68

[6] E.-C. Catalan, *Théorèmes et problèmes de géométrie élémentaire,* Dunod, Paris, 1897.
☞ 68

[7] H. Senapati, “Monotonicity in spherical and hyperbolic triangles,” in *Eighteen essays in non-Euclidean geometry* (V. Alberge and A. Papadopoulos, eds.), European Mathematical Society (EMS), Zürich, 2019, Chapter 6, 81–91.
☞ 77

Chapter 8

The Gauss–Bonnet theorem and the geometry of surfaces

Son Lam Ho

Contents

1. Introduction 113
2. Geodesic curvature and geodesics 116
3. The Gauss–Bonnet theorem 120
References 123

1 Introduction

The Gauss–Bonnet theorem is one of the most beautiful equations in mathematics. It says that for an orientable closed surface S in Euclidean space $\mathbb{R}^3$, the “total curvature” is equal to 2π the Euler characteristic of S. That is,

$$\iint\limits_S K\, d\sigma = 2\pi\chi(S).$$

The left-hand side is the integral of K over the whole surface S with respect to the area form $d\sigma$; here K is the Gaussian curvature at each point on the surface. The Gaussian curvature is a geometric quantity which changes as you bend, stretch or scale S in $\mathbb{R}^3$. On the other hand, the Euler characteristic $\chi(S)$ is a topological invariant, a quantity that remains constant no matter how you deform the surface. What makes the Gauss–Bonnet formula remarkable is that it elegantly connects these two very different worlds of geometry and topology.

The above *global Gauss–Bonnet theorem* follows directly from the *local Gauss–Bonnet theorem*; we state a simplified version of it below:

Theorem 1.1 (Gauss). *Let S be a surface in $\mathbb{R}^3$ and let $T = \Delta(A, B, C)$ be a geodesic triangle in S, that is, T is a simply connected subset of S which is bounded by three geodesic segments connecting three points A, B, C. Then*

$$\iint_T K\, d\sigma = (\angle A + \angle B + \angle C) - \pi.$$

Here $\angle A, \angle B, \angle C$ are the interior angles of the triangle. Thus, the right-hand side is the angle excess of T. Indeed, since T is not a planar triangle, we should expect its interior to have non-zero total curvature. In that case the theorem proves that the sum of its angles is not π, and how different it is from π depends on curvature and area of T. Important examples are when K is constant, in which case the angle excess of a triangle T is a constant multiple of its area.

Our goal is to give the reader an informal understanding of the concepts related to the above formula without developing the machinery for a rigorous proof which can be found in many textbooks. The references [1] and [4] are good sources for a complete treatment of the subject. In Section 2, we define geodesic curvature of curves on surfaces and demonstrate its relation with angles on a piecewise geodesic curve. In Section 3, we informally explain the Gaussian curvature and provide a heuristic proof of Gauss–Bonnet formulas.

It is enlightening to know the history of how these ideas developed. Given a surface S smoothly embedded in $\mathbb{R}^3$, at each point $p \in S$ there is a tangent plane T_pS containing all tangent vectors, and there is a unit normal vector $\vec{N}_p$ perpendicular to the surface. Since S is orientable there is a coherent and continuous way to choose $\vec{N}_p$ at every point p. Pick a unit tangent vector $\vec{v} \in T_pS$. There is a unique plane $P_{\vec{v}}$ containing both $\vec{N}_p$ and $\vec{v}$, and this plane intersects S along a curve γ containing p. Locally around p we can parametrize this curve $\gamma\colon [-\epsilon, \epsilon] \to S$ such that $\gamma(0) = p$, $\gamma'(0) = \vec{v}$, and such that it is unit speed. Of course the first derivative $\gamma'(t)$ gives us tangent vectors. We can differentiate again to get the second derivative $\gamma''(t)$ of acceleration vectors. Then let us define $\kappa(\vec{v})$ to be the dot product $\vec{N}_p.\gamma''(0)$. This quantity $\kappa(\vec{v})$ will be called the curvature in the $\vec{v}$ direction at p. Since the curve γ is on the plane $P_{\vec{v}}$, it is easy to see that $\kappa(\vec{v}) = \pm\|\gamma''(0)\|$ and the sign depends on whether $\gamma''(0)$ and $\vec{N}_p$ point to the same or opposite direction. The following theorem is proven by Euler [2]; it precedes Gauss' work on surfaces.

Theorem 1.2 (Euler). *Given the above notations, if $\kappa(\vec{v})$ are not all the same as $\vec{v}$ varies in the tangent plane, then there is a unit tangent vector $\vec{v}_1$ such that $k_1 = \kappa(\vec{v}_1)$ is the minimum, and a unit tangent $\vec{v}_2$ such that $k_2 = \kappa(\vec{v}_2)$ is the maximum. Moreover, $\vec{v}_1$ and $\vec{v}_2$ are perpendicular.*

Here k_1, k_2 are called the principal curvatures of S at p, and $\vec{v}_1, \vec{v}_2$ are the corresponding principal directions.

Gauss defined surface curvature as follows [3]. Let $\mathbb{S}^2$ be the unit 2-sphere in $\mathbb{R}^3$. For a smooth surface $S \subset \mathbb{R}^3$, let $\mathcal{G}: S \to \mathbb{S}^2$ be the map taking a point $p \in S$ to its unit normal vector $\vec{N}_p$. This is called the Gauss map. We have the differential map $d\mathcal{G}_p: T_pS \to T_{\mathcal{G}(p)}\mathbb{S}^2$ where T_pS is the tangent plane to S at p and $T_{\mathcal{G}(p)}\mathbb{S}^2$ is the tangent plane to the unit sphere at $\mathcal{G}(p)$.

Definition 1.3. Let $\vec{v}_1, \vec{v}_2$ be tangent vectors at p. The *Gaussian curvature* at p is defined to be

$$K = \frac{\text{Area of the parallelogram spanned by } \mathcal{G}(\vec{v}_1), \mathcal{G}(\vec{v}_2)}{\text{Area of the parallelogram spanned by } \vec{v}_1, \vec{v}_2}.$$

Of course this is an informal version of a more rigorous definition. We can think of K as the scaling factor of the Gauss map on infinitesimal area near p.

It is then showed that the curvature K at p equals the product of the principal curvatures, that is, $K = k_1k_2$. This result gives us a geometric understanding of curvature of surfaces. At a point on S, K is positive if its principal curvatures are of the same sign, which is when the principal tangent curves bend in the same direction at p; K is negative when the principal tangent curves bend in opposite directions; and K is zero when one of the principal tangent curves is flat at p. Figure 1 demonstrates the local picture around points of positive, negative, and zero curvatures.

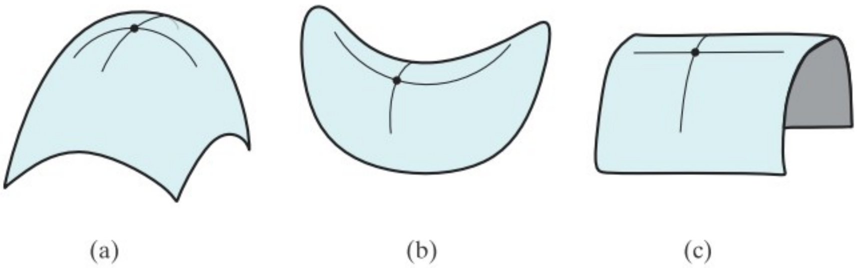
(a) (b) (c)

Figure 1

When a surface is smoothly embedded in $\mathbb{R}^3$, it inherits an infinitesimal metric which is given in coordinates by the 1st fundamental form. This metric, later known as the Riemannian metric—a notion of distance on the surface—exists without the need of embedding the surface into 3-space. Gauss formulated a more sophisticated description of K in terms of the 1st and 2nd fundamental forms, which in turn can be computed purely using the 1st fundamental form. This result is called Gauss' *Theorema egregium* [3], translated by "remarkable theorem." Indeed the theorem shows that the curvature K depends only on the Riemannian metric of the surface, hence it is an *intrinsic* property of the surface and it can be computed purely by measurements of distances and angles on the surface. This result and the works of Bernhard Riemann that follows mark the beginning of Differential Geometry, a very important and beautiful area of mathematics.

Assuming the *Theorema Egregium*, in Section 3, we will use an informal intrinsic definition of the Gaussian curvature to prove the local Gauss–Bonnet theorem.

2 Geodesic curvature and geodesics

We will now define the notion of geodesics which will be important to understand curves on surfaces. Let S be a smooth surface in 3-space. Consider a smooth curve $\gamma\colon [a,b] \to S$ of unit speed, that is, $\|\gamma'(t)\| = 1$. At the point $\gamma(t)$, the curve's tangent vector is given by $\gamma'(t)$ and its acceleration vector is $\gamma''(t)$. Since the curve is unit speed, the acceleration vector is perpendicular to the tangent vector. (One can see this by moving the base point of vector $\gamma'(t)$ to 0, so that $\gamma'(t)$ becomes a path on the unit sphere, which makes the vector $\gamma''(t)$ tangent to this path at the point $\gamma'(t)$, hence $\gamma'(t) \perp \gamma''(t)$.) Now let $(\gamma'(t))^{\perp}$ be the plane of all vectors perpendicular to $\gamma'(t)$.

We have a natural orthonormal basis for $(\gamma'(t))^{\perp}$ given by $\{\vec{N}_{\gamma(t)}, \vec{N}_{\gamma(t)} \times \gamma'(t)\}$. Recall that $\vec{N}_{\gamma(t)}$ is the unit normal vector to the surface and $\vec{N}_{\gamma(t)} \times \gamma'(t)$ is the cross product which is perpendicular to both $\vec{N}_{\gamma(t)}$ and $\gamma'(t)$. Since $\gamma''(t) \in (\gamma'(t))^{\perp} = \operatorname{span}\{\vec{N}_{\gamma(t)}, \vec{N}_{\gamma(t)} \times \gamma'(t)\}$, we have a unique decomposition

$$\gamma''(t) = k_n(t)\vec{N}_{\gamma(t)} + k_g(t)(\vec{N}_{\gamma(t)} \times \gamma'(t))$$

for $k_n(t), k_g(t) \in \mathbb{R}$.

Definition 2.1. We define the *geodesic curvature* of the curve γ at point $p = \gamma(t)$ to be $k_g(t)$.

Definition 2.2. A geodesic on S is a unit speed curve whose geodesic curvature is 0 everywhere along the curve.

The geodesic curvature is a real valued quantity measuring how much the curve is veering left or right at a given point, left and right is distinguished by the sign of $k_g(t)$. The curve is geodesic when informally speaking it is always going straight, and more precisely, it is geodesic when the acceleration vector $\gamma''(t)$ is either 0 or perpendicular to the surface.

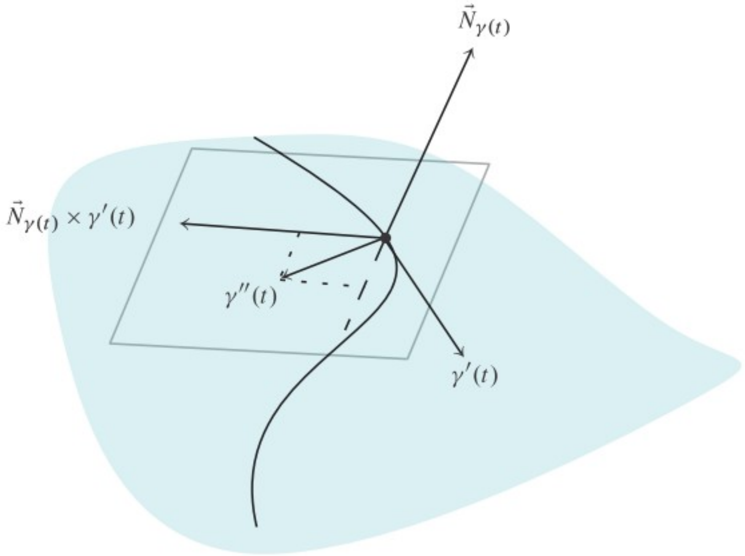

Figure 2

Using this definition one can easily see that geodesics on a plane are straight lines, and circles on a plane has constant geodesic curvature inversely proportional to the circle's radius. Indeed, suppose γ is a unit speed parametrized planar circle of radius r, so $\gamma\colon [0, 2\pi r] \to \mathbb{R}^2$ given by $\gamma(t) = \big(r\cos\big(\frac{t}{r}\big), r\sin\big(\frac{t}{r}\big)\big)$. We can check that

$$\gamma'(t) = \Big(-\sin\Big(\frac{t}{r}\Big), \cos\Big(\frac{t}{r}\Big)\Big) \quad \text{and} \quad \gamma''(t) = \frac{1}{r}\Big(-\cos\Big(\frac{t}{r}\Big), -\sin\Big(\frac{t}{r}\Big)\Big).$$

We can choose $\vec{N}_{\gamma(t)}$ so that $\vec{N}_{\gamma(t)} \times \gamma'(t) = \big(-\cos\big(\frac{t}{r}\big), -\sin\big(\frac{t}{r}\big)\big)$ thus the geodesic curvature is $k_g(t) = \frac{1}{r}$.

Note that for the case of planar circles, the acceleration vector $\gamma''(t)$ always points towards the center of the circle. Using this fact one can see that great circles on a sphere, i.e. circles which are intersection of the sphere with a plane through the center, are geodesics of the sphere.

We defined the geodesic curvature making full use of the ambient space $\mathbb{R}^3$. However, it can be defined intrinsically in a way that depends on the introduction of the unique *Levi-Civita connection* determined by the Riemannian metric. We will not develop this general machinery here.

For the remaining part of this section we will explore how geodesic curvature is related to angles. Given a smooth unit speed arc $\gamma\colon [a,b] \to S$, we can integrate the geodesic curvature with respect to arc-length:

$$\int_{\gamma} k_g \, ds = \int_{a}^{b} k_g(t) dt.$$

Think of this quantity as the *total geodesic curvature* of the arc γ. Now suppose that our surface is the $\mathbb{R}^2$ plane, we will compute the total geodesic curvature of γ as the change in direction of the tangent vector $\gamma'(t)$ as t goes from a to b. Since γ is of unit speed, we can think of $\gamma'(t)$ as a parametrized path on the unit circle $\mathbb{S}^1$. For simplicity let us suppose that $k_g(t) \geq 0$ on the whole curve γ. (For curves with k_g changing sign, we can always decompose into a concatenation of sub-arcs so that on each arc k_g is either non-positive or non-negative.) Then we have

$$\int_{a}^{b} k_g(t) dt = \int_{a}^{b} \|\gamma''(t)\| dt.$$

The right-hand side is exactly the definition of arclength for the path $\gamma'(t)$ which starts at $\gamma'(a)$ and ends at $\gamma'(b)$. Since the arc $\gamma'(t)$ is on the unit circle, its length is $\Theta + k2\pi$ where Θ is the angle between the $\gamma'(a)$ and $\gamma'(b)$ as vectors based at 0, and $k \in \mathbb{Z}$ records how many times the path $\gamma'(t)$ winds around the unit circle.

Remark 2.3. The total geodesic curvature of a planar curve γ is equal to the total amount of turning (measured by angle) of the tangent vector $\gamma'(t)$.

We also have the following claim which will be justified below.

Remark 2.4. We can think of an angle as infinitely concentrated geodesic curvature.

Consider two infinite unit speed geodesic rays $a_1, a_2\colon [0,\infty) \to S$ starting at the same point $p_0 = a_1(0) = a_2(0)$. Let $\Theta \in [0,\pi]$ be the angle between $-a_1'(0)$ and $a_2'(0)$. We let $\alpha\colon \mathbb{R} \to S$ be the piecewise geodesic curve so that $\alpha(s) = a_2(s)$ for $s \in [0,\infty)$, and $\alpha(s) = a_1(-s)$ for $s \in (-\infty, 0]$. So Θ is the *exterior angle* at the vertex $p_0 = \alpha(0)$.

One can construct a sequence of smooth curves $\alpha_1, \alpha_2, \ldots$ approximating α by smoothing out the angle as follows. Let $U_1 \supset U_2 \supset \cdots$ be a sequence of open discs all centered at p_0 such that the radius of U_i approaches 0 as $i \to \infty$. We let $\bar{U}_i = S \backslash U_i$ be their complements. We can construct smooth unit speed curves α_i so

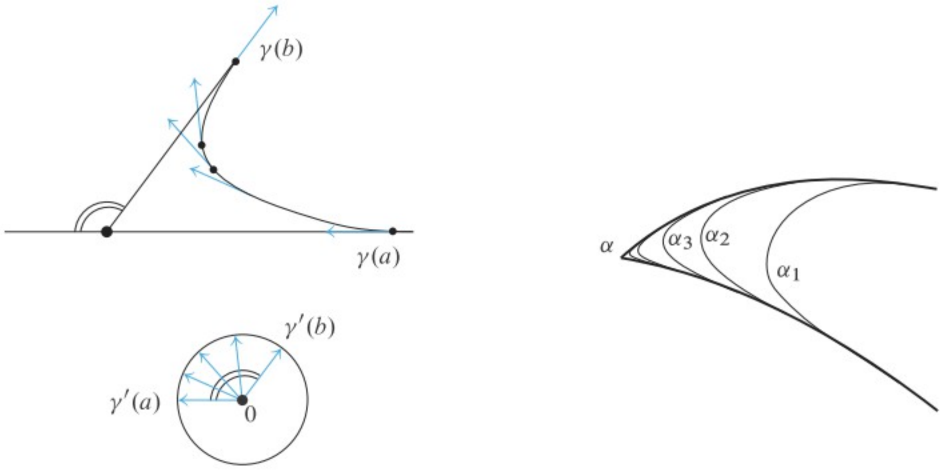

(a) (b)

Figure 3. (a) Illustration for Remark 2.3; (b) illustration for Remark 2.4

that up to a reparametrization it agrees with α in $\bar{U}_i$. In other words, α_i agrees with α everywhere except in a small disc U_i around the vertex p_0. Integrating the geodesic curvature of these approximating curves we have

$$\int\limits_{\alpha_i} k_g\, ds = \int\limits_{\alpha_i \cap U_i} k_g\, ds + \int\limits_{\alpha_i \cap \bar{U}_i} k_g\, ds = \int\limits_{\alpha_i \cap U_i} k_g\, ds$$

since $\alpha_i \cap \bar{U}_i$ is where α_i agrees with α which have 0 geodesic curvature. Now since the U_i's are getting smaller, the geometry on U_i gets closer to planar geometry where Remark 2.3 applies. Along $\alpha_i \cap U_i$, the total turning angle for the tangent vectors $\alpha_i'(t)$ approaches Θ, so $\int_{\alpha_i \cap U_i} k_g\, ds \to \Theta$, and therefore

$$\int\limits_{\alpha} k_g\, ds = \lim_{i \to \infty} \int\limits_{\alpha_i} k_g\, ds = \Theta.$$

It makes sense for α, a curve which is geodesic except at one point, to still have a total geodesic curvature measure-concentrated at p_0. Indeed there are points on $\alpha_i \cap U_i$ where the geodesic curvature approaches infinity as i increases.

The full statement of the local Gauss–Bonnet theorem is as follows.

Theorem 2.5 (Gauss). *Let S be a surface in $\mathbb{R}^3$ and let D be an open simply connected region on S whose boundary ∂D is smooth except at n vertices where the exterior angles at these vertices are $\theta_1, \ldots, \theta_n$.*

$$\iint\limits_D K\, d\sigma = 2\pi - \left(\sum_{i=1}^{n} \theta_i + \int\limits_{\partial D} k_g\, ds \right)$$

where $\int_{\partial D} k_g\, ds$ is computed as a sum of integrals over the smooth segments of ∂D.

If $n = 0$, that is, if ∂D is smooth with no vertices, then the equation reduces to $\iint_D K\, d\sigma = 2\pi - \int_{\partial D} k_g\, ds$. This simplified case can be proven by working in an orthogonal uv-coordinate patch and writing both K and k_g in terms of the first fundamental form quantities E, G and partial derivatives. The formula follows after applying the Gauss–Green theorem. This is the rigorous approach as seen in [1].

Now suppose that ∂D has $n \geq 1$ vertices. We can approximate D by a sequence of open domains D_i with smooth boundary ∂D_i agreeing with ∂D on almost all the smooth parts and replacing the angles with short arcs of large geodesic curvature, just like our discussion following Remark 2.4. So as ∂D_i approach ∂D we have the *total measure* of geodesic curvature:

$$\lim_{D_i \to D} \int\limits_{\partial D_i} k_g\, ds = \sum_{i=1}^{n} \theta_i + \int\limits_{\partial D} k_g\, ds.$$

Moreover,

$$\lim_{D_i \to D} \iint\limits_{D_i} K\, d\sigma = \iint\limits_D K\, d\sigma.$$

Thus theorem 2.5 follows from the case where ∂D is smooth.

3 The Gauss–Bonnet theorem

One thing we did not explain was how in a coordinate patch, the Gaussian curvature can be expressed as the exterior derivative of geodesic curvature—an essential step to prove the Gauss–Bonnet theorem. To explain this we will define the curvature in a different way, hoping to informally justify the Gauss–Bonnet theorem with a different perspective.

Suppose P is a geodesic polygon on S with k vertices. We only consider the case when P bounds a simply connected interior. Let $\phi_1, \dots, \phi_k$ be the interior angles of P. We can consider following quantity representing the *angle excess of* P:

$$\delta(P) := \sum_{i=1}^{k} \phi_i - (k-2)\pi = 2\pi - \sum_{i=1}^{k} (\pi - \phi_i).$$

Note that $(k-2)\pi$ is the sum of interior angles of a Euclidean k-gon, and 2π is the sum of exterior angles. If we "add" a triangle T next to P (see Figure 4), we obtain a new geodesic polygon Q of $k+1$ vertices. Let $\alpha_1, \alpha_2, \alpha_3$ be the interior angles of T and $\beta_1, \dots, \beta_{k+1}$ be the interior angles of Q.

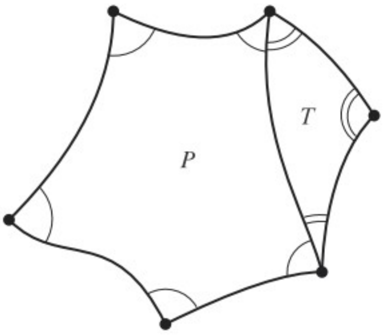

Figure 4

Remark 3.1. Angle excess is *additive*, that is, the angle excess of the new polygon Q is the sum of angle excesses of P and T:

$$\delta(Q) = \sum_{i=1}^{k+1} \beta_i - (k-1)\pi = \sum_{i=1}^{k} \phi_i - (k-2)\pi + \sum_{i=1}^{3} \alpha_i - \pi = \delta(P) + \delta(T).$$

Let p be a point on the smooth surface $S \subset \mathbb{R}^3$. Let $T = \Delta(A, B, C)$ be a geodesic triangle on S. We now define the curvature at p to be

$$K(p) = \lim_{T \to p} \frac{\delta(T)}{\text{Area}(T)} = \lim_{T \to p} \frac{(\angle A + \angle B + \angle C) - \pi}{\text{Area}(T)}.$$

The limit is taken over smaller and smaller geodesic triangles T containing p in a way such that the vertices A, B, C approach p. We will take for granted that this limit exists not depending on how the vertices approach p, and that it varies continuously as p varies in S. This is an intrinsic definition of the curvature because measurements of angles and area are taken on the surface S without refering to how S is embedded in $\mathbb{R}^3$.

Theorem 3.2 (Gauss). *Let S be a surface in $\mathbb{R}^3$ and let $T = \Delta(A, B, C)$ be a geodesic triangle in S, that is, T is a simply connected subset of S which is bounded by three geodesic segments connecting three points A, B, C. Then*

$$\iint_T K\, d\sigma = (\angle A + \angle B + \angle C) - \pi.$$

Informal proof. Since the closure of T is compact, for any $\epsilon > 0$, we can divide T into n sub-triangles $T_{\epsilon,1}, \dots, T_{\epsilon,n}$ such that each $T_{\epsilon,i}$ is of diameter less than ϵ. Here n depends on ϵ by this relation $n \approx \left(\frac{\text{diameter}(T)}{\epsilon}\right)^2$.

We choose a point $p_{\epsilon,i}$ in each $T_{\epsilon,i}$. Applying Riemann integration using the sub-triangles $T_{\epsilon,i}$ we have the following heuristic argument

$$\begin{aligned}\iint_T K\, d\sigma &= \lim_{\epsilon\to 0} \sum_{i=1}^{n} \big(K(p_{\epsilon,i}) \operatorname{Area}(T_{\epsilon,i})\big)\\ &= \lim_{\epsilon\to 0} \sum_{i=1}^{n} \frac{\delta(T_{\epsilon,i})}{\operatorname{Area}(T_{\epsilon,i})} \operatorname{Area}(T_{\epsilon,i})\\ &= \delta(T) = (\angle A + \angle B + \angle C) - \pi \qquad \text{(by Remark 3.1).}\end{aligned}$$

Therefore we obtain the local Gauss–Bonnet formula. The above argument works not only for triangles but for geodesic polygons in general. □

We have the global Gauss–Bonnet immediately after the following elegant proof.

Theorem 3.3 (Gauss). *Let S be a smooth closed surface then*

$$\iint_S K\, d\sigma = 2\pi\chi(S).$$

where $\chi(S)$ is the Euler characteristic.

Proof. We give S a triangulation by geodesic triangles T_i. Suppose that the number of vertices, edges, and faces are V, E, F respectively. We have $\chi(S) = V - E + F$ by Euler's famous formula. This is a topological invariant of the surface S not depending on the triangulation. Since each triangle has three edges, we have $3F$ equals the total number of edges double counting each edge twice, because each edge belongs to two faces. So $3F = 2E$, which implies $\chi(S) = V - E + F = V - \frac{F}{2}$.

Let $\beta_1^{(i)}, \beta_2^{(i)}, \beta_3^{(i)}$ be the interior angles of triangle T_i. Now applying Theorem 3.2 we have

$$\begin{aligned}\iint_S K\, d\sigma &= \sum_{i=1}^{F} \iint_{T_i} K\, d\sigma = \sum_{i=1}^{F} \delta(T_i)\\ &= \sum_{i=1}^{F} (\beta_1^{(i)} + \beta_2^{(i)} + \beta_3^{(i)} - \pi)\\ &= \sum_{i=1}^{F} (\beta_1^{(i)} + \beta_2^{(i)} + \beta_3^{(i)}) - F\pi\\ &= 2\pi V - F\pi\\ &= 2\pi\chi(S).\end{aligned}$$

□

Theorem 3.2 can be applied similarly to obtain a Gauss–Bonnet formula for surfaces with cone points.

Corollary 3.4. *Let S be a surface with k cone points and cone angles $\theta_1, \dots, \theta_k$, then*

$$\iint_S K\, d\sigma = 2\pi\chi(S) + \sum_{i=1}^{k} (\theta_i - 2\pi).$$

References

[1] M. do Carmo, *Differential geometry of curves and surfaces,* translated from the Portuguese, Prentice-Hall, Englewood Cliffs, N.J., 1976.
MR 0394451 Zbl 0326.53001 ☞ 114, 120

[2] L. Euler, "Recherches sur la courbure des surfaces," *Memoires de l'academie des sciences de Berlin* 16 (1767), 119–143.
☞ 114

[3] C. F. Gauss, *Disquisitiones generales circa superficies curvas,* Typis Dieterichianis, Göttingen, 1828.
`https://archive.org/details/disquisitionesg00gausgoog`
☞ 115

[4] M. Spivak, *A comprehensive introduction to differential geometry,* Vol. II, 3rd ed., Publish or Perish, Houston, TX, 1999.
MR 0532831 Zbl 1213.53001 ☞ 114

Chapter 10

Area preserving maps from the sphere to the Euclidean plane

Charalampos Charitos

Contents

1. Introduction 135
2. The distortion of area-preserving geographical projections 138
3. Proof of the main theorem 141
References 149

1 Introduction

The problem of cartography, that is, the problem of constructing the mappings from the sphere S^2 (or from a subset of S^2) to the Euclidean plane E^2 which are "best" in some sense is an ancient problem, and several prominent mathematicians have been studying it from antiquity to our days. The word "best" refers to mappings which present the least distortion with respect to distances, angles or areas, or a combination of these.

In the sequel, by a mapping from S^2 to E^2 we mean a map defined either on the whole sphere S^2 or, on a subset of S^2.

In 1777, Euler published three memoirs on cartography:

(i) "De repraesentatione superficiei sphaericae super plano" (On the representation of spherical surfaces on a plane) [2];

(ii) "De proiectione geographica superficiei sphaericae" (On the geographical projections of spherical surfaces) [3];

(iii) "De proiectione geographica Deslisliana in mappa generali imperii russici usitata" (On Delisle's geographic projection used in the general map of the Russian empire) [4].

In the first memoir, he studied the following three kinds of maps:

(1) maps where the images of all meridians are perpendicular to a given horizontal axis and all parallels are parallel to it;

(2) maps which are conformal;

(3) maps which preserve area and such that the images of all meridians intersect the images of the parallels at right angles.

All these maps considered by Euler will be referred to as *geographical projections*.

Two years later, Lagrange published two memoirs on cartography [6, 7] in which he observed that there exist infinitely many, essentially different, angle-preserving (conformal) projections from S^2 to E^2 in the following sense: any two are not equal and they remain so after a composition by a similarity of E^2. Subsequently, Lagrange dealt with the problem of finding the minimal deviation map amongst all these conformal projections. To that end, he defined at each point a *magnification ratio*, which is a deviation factor from conformality at this point.

In the nineteenth century, the French mathematician and cartographer Nicolas-Auguste Tissot made a thorough investigation of geographical projections from S^2 to E^2 and he studied several of their distortion properties. Following his ideas [11], we shall recall the definition of the distortion $D_f(p)$ of f at a point $p \in S_1$, see Definition 2.2 below, where $f : S_1 \to S_2$ is a non-conformal map of class C^1 between two Riemannian surfaces. After the work of Tissot, the term *quasiconformal map* was generated and the problem of existence and uniqueness of the best quasiconformal map between Riemann surfaces was revealed as an important and difficult problem.

For the origin and development of cartography and its relation with quasiconformal maps, we refer the reader to the paper by Papadopoulos [9] which inspired the author to write the present chapter. The paper [10], also by Papadopoulos, contains biographical material on Tissot, and a review of some of his main ideas. We note incidentally that both Grötzsch and Teichmüller, who are usually referred for the early work on quasiconformal maps, refer to Tissot; cf. the history in [9] and [10].

We now explain the results.

The sphere S^2 is parameterized by the map

$$\xi : (t,u) \in [-\pi,\pi] \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \longrightarrow (\cos(t)\cos(u), \sin(t)\cos(u), \sin(u)). \qquad (1)$$

For t fixed, the u-curves $\xi(t,\cdot)$ are the *meridians* of S^2 and are geodesics for the spherical metric. Setting $u = 0$, the curve $\xi(\cdot,0)$ is the *equator* of S^2 while the t-curves $\xi(\cdot, u)$, $u \neq 0$ are circles parallel to the equator and they are called *parallels* of S^2. The parallels are not geodesics. The circle $\xi(0, \cdot)$ will be called *prime meridian.* The pair (t, u) define the geographical coordinates of S^2, namely t is the *latitude* and u is the *longitude* of S^2. Thus, the equator and the prime meridian serve as axes of S^2 with respect to which the geographical coordinates are measured.

In what follows, we always assume that ξ is restricted to the open set $(-\pi, \pi) \times \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ because on this set ξ is one-to-one. We shall denote by S_0 the subset $\xi\left((-\pi, \pi) \times \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)\right) \subset S^2$. By means of ξ, each point $e \in S_0$ with $e = \xi(t, u)$ will be identified with the pair (t, u) if no confusion is likely.

A differentiable map $f(t, u) = (x(t, u), y(t, u))$ from S_0 to E^2 is called *regular* if $\left(\frac{\partial x}{\partial t}, \frac{\partial y}{\partial t}\right) \neq (0, 0)$ and $\left(\frac{\partial x}{\partial u}, \frac{\partial y}{\partial u}\right) \neq (0, 0)$. We denote by $\mathcal{Q}$ the class of all regular maps of class C^r, $r \geq 2$ from S_0 to E^2 which preserve area and which send all meridians and parallels to perpendicular curves of E^2. In fact, the elements of $\mathcal{Q}$ are all the geographical projections which satisfy the third hypothesis in Euler's work [2] to which we referred to above. It is not hard to see that the elements of $\mathcal{Q}$ cannot be conformal maps.

If $f(t, u) = (x(t, u), y(t, u)) \in \mathcal{Q}$, we set

$$m_f(t, u) = \sqrt{\left(\frac{\partial x}{\partial u}\right)^2 + \left(\frac{\partial y}{\partial u}\right)^2}. \tag{2}$$

One can observe that $m(t, u)$ is the length in E^2 of the tangent vector to the curve $f(t, \cdot)$. We will be writing $m(t, u)$ if there is no confusion on the map f.

Finally, there is a simple area-preserving geographical projection $f_0 \in \mathcal{Q}$ which was introduced and studied by Lambert in [8, Section 6] (see also [12], pp.76–85). This projection, known as *Lambert's cylindrical equal area projection*, was rediscovered by Euler in [2]. It is defined by the formula

$$f_0(t, u) = (t, \sin(u)).$$

The goal of this note is to prove that the distortion $D_{f_0}(t, u)$ of f_0 has remarkable extremal properties with respect to the distortion of the maps $f \in \mathcal{Q}$, namely, for every t and u fixed the value $D_{f_0}(t, u)$ appears either as a maximum or as a minimum of the values $\{D_f(t, u),\ f \in \mathcal{Q}\}$.

We have the following theorem which permits us to estimate the deviation from conformality of the mappings $f \in \mathcal{Q}$, where the conformality structure on S^2 is induced from the standard Riemannian metric $ds = \sqrt{\cos^2(u)dt^2 + du^2}$ of S^2.

Theorem 1.1. *If $f(t, u) = (x(t, u), y(t, u)) \in \mathcal{Q}$, then for each $(t, u) \in S_0$,*

(i) $D_{f_0}(t, u) \leq D_f(t, u)$ *if* $m_f(t, u) \geq 1$ *or* $m_f(t, u) \leq \cos(u)$;

(ii) $D_{f_0}(t, u) \geq D_f(t, u)$ *if* $\cos(u) \leq m_f(t, u) \leq 1$.

In the last paragraph, following Euler's method, we will find explicit maps $f \in \mathcal{Q}$ and we will show that some of them satisfy the inequality $m_f(t,u) \leq \cos(u)$ for each $(t,u) \in S_0$. Therefore, the distortion of these maps is bounded below by the distortion of f_0 for each $(t,u) \in S_0$.

Remark 1.2. Tissot in [11, Chapitre troisième] studied concrete geographical projections which preserve area and which he referred to as "projections authaliques" (see [11], № 70). He made a thorough analysis of these projections and he calculated their distortion, in the sense of Definition 2.2 below, at certain points of the sphere including the equator and the poles. Using this, Tissot compiled numerical tables containing, among other things, useful information about the distortion of these projections (see the Complement in [11]). However, since all projections considered by Tissot are contained in the set $\mathcal{Q}$, Theorem 1.1 can be used to find estimates of their distortion which become quite accurate near the poles and the equator. For instance, for any map $f \in \mathcal{Q}$, the distortion $D_f(t,u)$ goes to infinity (resp. to 1) as (t,u) approaches the poles (resp. the equator) provided that $D_f(t,u)$ is bounded below (resp. above) by $D_{f_0}(t,u)$.

2 The distortion of area-preserving geographical projections

Let A be a linear isomorphism from E^2 to E^2 which is not conformal and let us denote by $\|\cdot\|$ the Euclidean norm. Let C_r be the circle of radius $r > 0$ centered at the origin O. It is well known that each C_r is sent by A to an ellipse L_r centered at O. Furthermore, there is a unique pair v_1, v_2 of perpendicular unit vectors of E^2 such that Av_1, Av_2 are perpendicular. This result was proven by Tissot in ([11], Chapitre Premier, № 3) by a simple continuity argument. Below, we reprove it in a geometric way.

The following proposition summarizes basic and nice properties of A which are used below.

Proposition 2.1. (1) *All the ellipses L_r are homothetic with homothety center O. If a_r is the great axis of L_r and b_r its small axis, then $a_1/b_1 = a_r/b_r$ for each r.*

The ratio a_1/b_1 is called the ***distortion*** *of A.*

(2) *There is a unique pair v_1, v_2 of perpendicular unit vectors of E^2 such that Av_1, Av_2 are perpendicular. Thus, if $\|v_1\| = \|v_2\|$ and if $\|Av_1\| > \|Av_2\|$ then $a_1/b_1 = \|Av_1\|/\|Av_2\|$.*

(3) *A and A^{-1} have the same distortion.*

Proof. Let a_1, b_1 be the great and small axis of L_1 respectively and let w_1, w_2 be normal vectors in the direction of the great and small axis of L_1 respectively with $\|w_1\| = a_1$, $\|w_2\| = b_1$. Let v_1, v_2 be unit vectors such that $A(v_1) = w_1$ and $A(v_2) = w_2$. Obviously $\|w_1\| < \|A(v)\| < \|w_2\|$ for each unit vector $v \neq v_1, v_2$. From the linearity of A we have also $\|A(\lambda v_1)\| < \|A(\lambda v)\| < \|A(\lambda v_2)\|$ and $\|A(\lambda v_1)\|/\|A(\lambda v_2)\| = \|w_1\|/\|w_2\| = a_1/b_1$. Thus statement (1) follows easily.

In order to show that the vectors v_1 and v_2 are perpendicular, we consider a circle C'_a (resp. C'_b) centered at the origin and tangent to L_1 at the two points where the great axis (resp. small axis) intersects L_1. Obviously C'_b is contained in the interior of L_1 and L_1 is contained in the interior of C'_a. Let K_a (resp. K_b) be the image of C'_a (resp. C'_b) by A^{-1}. Then K_a, K_b are homothetic ellipses centered at the origin which must be tangent to C_1 at the endpoints of the vectors v_1, v_2. This implies immediately that v_1, v_2 are perpendicular and unique and thus statement (2) is proven. Finally statement (3) follows easily from (2). ☐

The above properties of linear isomorphisms of E^2 give rise to the following definition of the distortion of a diffeomorphism of class C^1 between Riemannian surfaces, i.e. surfaces equipped with Riemannian metrics.

Definition 2.2 (distorsion). Let S_1,S_2 be two Riemannian surfaces and let $f:S_1 \to S_2$ be a diffeomorphism of class C^1. At each point $p \in S_1$ the differential df_p between the tangent spaces T_pS_1 of S_1 and the tangent space $T_{f(p)}S^2$ sends the unit circle C of T_pS_1 centered at the origin to an ellipse $L = df_p(C)$. Denoting by a and b the great and the small axis respectively of L, the *distortion* $D_f(p)$ of f at p is the distortion of the linear map df_p and is defined as the ratio a/b.

In Definition 2.2, if df_p preserves the angles then $D_f(p) = 1$ since $a = b$. Thus, if f is conformal then $D_f(p) = 1$ for each $p \in S_1$.

The real function $D_f : S_1 \to \mathbb{R}$ will be referred to as the *distortion function* of f.

As a special case, let $f : \Omega \to f(\Omega)$ be a C^1-diffeomorphism between the plane domains Ω and $f(\Omega)$. Let $z = (t, u) = t + iu$ be a coordinate system on Ω and let $w = f(z) = x + iy$ be a coordinate system on $f(\Omega)$. If we consider on Ω and $f(\Omega)$ the standard Euclidean metric, the distortion of f at $z_0 \in \Omega$ may be defined as previously and will be denoted by $\bar{D}_f(z_0)$, as a reminder of the fact that it is defined with respect to the Euclidean metric.

The following well-known formula can be found in many books (see for example [5], p. 2)

$$\bar{D}_f(z_0) = \frac{|f_z(z_0)| + |f_{\bar{z}}(z_0)|}{|f_z(z_0)| - |f_{\bar{z}}(z_0)|} \geq 1. \tag{3}$$

where $f_z = 1/2(f_t - if_u)$ and $f_{\bar{z}} = 1/2(f_t + if_u)$.

Obviously, if S_1, S_2 are Riemann surfaces and f is a C^1-diffeomorphism between them, the distortion $\bar{D}_f(p)$ can be defined for each $p \in S_1$. This can be done by covering S_1 (resp. S_2) by an atlas of charts which defines the Riemann structure on S_1 (resp. on S_2).

First, we will calculate the distortion function of the map

$$\xi \colon (-\pi, \pi) \times \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \subset E^2 \longrightarrow S_0,$$

given by formula (1). The differential $d\xi_{(t,u)}$ of the map ξ sends the vectors $v_1 = (1, 0)$ and $v_2 = (0, 1)$ of the tangent space $T_{(t,u)}E^2$ to the vectors

$$r_1 = (-\sin(t)\cos(u), \cos(t)\cos(u), 0),$$

$$r_2 = (-\cos(t)\sin(u), -\sin(t)\sin(u), \cos(u))$$

respectively, of the tangent space $T_{\xi(t,u)}S^2 \subset E^3$. If we denote by $\langle \cdot, \cdot \rangle$ the Euclidean dot product, we have

$$\langle r_1, r_2 \rangle = 0, \quad \|r_1\| = \cos(u), \quad \|r_2\| = 1.$$

This implies that the unit circle C in $T_{(t,u)}E^2$ is sent to an ellipse $L = d\xi_{(t,u)}(C)$ in $T_{\xi(t,u)}S^2$ whose great axis has length 1 and the small one has length $\cos(u)$. Therefore, the distortion $D_\xi(t, u)$ of the map ξ at (t, u) is $1/\cos(u)$. Thus, if we consider the inverse of the map ξ, say $\eta = \xi^{-1} \colon S_0 \to (-\pi, \pi) \times \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$, and if $e = \xi(t, u)$, we have

$$D_\eta(e) = \frac{1}{\cos(u)} \tag{4}$$

since from Proposition 2.1 the distortion of $d\xi_{(t,u)}$ is equal to the distortion of $d\eta_e$, where $e = \xi(t, u)$.

Now let $U = (-\pi, \pi) \times \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ and $g \colon U \to f(U) \subset E^2$ be C^1-diffeomorphism such that the differential $dg_{(t,u)}$ sends the unit vectors $v_1 = (1, 0)$, $v_2 = (0, 1)$ of $T_{(t,u)}U$ to perpendicular vectors, say w_1, w_2 in $T_{g(t,u)}E^2$. Set $G = g \circ \eta \colon S_0 \to \mathbb{E}^2$.

We have the following lemma.

Lemma 2.3. *Let $e = \xi(t, u) \in S_0$. The distortion $D_G(t, u)$ of G at e is given by the product*

$$D_G(t, u) = \frac{1}{\cos(u)} \bar{D}_g(t, u). \tag{5}$$

Proof. Since $d\eta_e$ sends the vectors r_1, r_2 of T_eS^2 to v_1, $v_2 \in T_{(t,u)}U$ and $dg_{(t,u)}$ sends v_1, v_2 to the perpendicular vectors w_1, w_2 in $T_{g(t,u)}E^2$, the proof follows easily combining relation (4) and statement (2) of Proposition 2.1. □

Lemma 2.4. *Let* $f_0 \colon S_0 \to E^2$ *be the projection* $f_0(t,u) = (t, \sin(u))$. *Then*

$$D_{f_0}(t,u) = \frac{1}{\cos^2(u)}. \tag{6}$$

Proof. The differential $df_{0_{(t,u)}}$ is expressed by the matrix $A = \left(\begin{smallmatrix} 1 & 0 \\ 0 & \cos(u) \end{smallmatrix}\right)$. Since $A(1,0) = (1,0)$ and $A(0,1) = (0, \cos u)$, we have from statement (2) of Proposition 2.1 that $\bar{D}_{f_0}(t,u) = 1/\cos(u)$ and from Lemma 5 that $D_{f_0}(t,u) = 1/\cos^2(u)$. □

3 Proof of the main theorem

Let $f \colon S_0 \to E^2$, $f(t,u) = (x(t,u), y(t,u))$ be a geographical projection belonging to the set $\mathcal{Q}$. We recall that f is a regular, area-preserving map of class C^r, $r \geq 2$, which sends the families of meridians and parallels to families of perpendicular curves in E^2.

We start with the general formulas

$$dx = p\,du + q\,dt, \quad dy = r\,du + s\,dt, \tag{7}$$

where

$$p = \frac{\partial x}{\partial u}, \quad q = \frac{\partial x}{\partial t}, \quad r = \frac{\partial y}{\partial u}, \quad s = \frac{\partial y}{\partial t}.$$

Since the images of meridians and parallels are cut perpendicularly, $pq + rs = 0$ (see [1] or [2]). Therefore, there is a function $n(t,u)$ such that

$$s = -np, \quad q = nr. \tag{8}$$

Since f is regular, $n(t,u) \neq 0$ for each (t,u). Therefore, we may assume that n is positive. In fact, if n is negative we may work similarly.

From (7) and (8) we have

$$dx = p\,du + nr\,dt, \quad dy = r\,du - np\,dt. \tag{9}$$

Following an idea of Euler, we consider the points $e = (t,u)$, $e_1 = (t + dt, u)$, $e_2 = (t, u + du)$ in S_0 and we set

$$E = f(t,u), \quad E_1 = f(t + dt, u), \quad E_2 = f(t, u + du).$$

We have, as in [1],

$$\lim_{du\to 0}\frac{|e_2-e|}{|du|}=1,$$

$$\lim_{dt\to 0}\frac{|e_1-e|}{|dt|}=\cos(u),$$

$$\lim_{du\to 0}\frac{\|E_2-E\|}{|du|}=\sqrt{\left(\frac{\partial x}{\partial u}\right)^2+\left(\frac{\partial y}{\partial u}\right)^2}=\sqrt{p^2+r^2},$$

$$\lim_{dt\to 0}\frac{\|E_1-E\|}{|dt|}=\sqrt{\left(\frac{\partial x}{\partial t}\right)^2+\left(\frac{\partial y}{\partial t}\right)^2}=n\sqrt{p^2+r^2}.$$

Therefore, the projection f is area-preserving if and only if

$$n(p^2+r^2)|du||dt|=\cos(u)|du||dt|$$

or

$$n=\frac{\cos(u)}{p^2+r^2}. \tag{10}$$

From (9) and (10) we have

$$dx=pdu+\frac{r\cos(u)}{p^2+r^2}dt,\quad dy=rdu-\frac{p\cos(u)}{p^2+r^2}dt.$$

To simplify the calculations, we set

$$p=m\cos(\phi)\quad\text{and}\quad r=m\sin(\phi),$$

where

$$m=\sqrt{\left(\frac{\partial x}{\partial u}\right)^2+\left(\frac{\partial y}{\partial u}\right)^2}>0,$$

and we have

$$dx=m\cos(\phi)du+\frac{\cos(u)\sin(\phi)}{m}dt,\quad dy=m\sin(\phi)du-\frac{\cos(u)\cos(\phi)}{m}dt.$$

Setting $m(t,u) = k(t,u)\cos(u)$ we have $k(t,u) > 0$ for each (t,u) and the relations above take the form

$$dx = k\cos(u)\cos(\phi)du + \frac{\sin(\phi)}{k}dt, \quad dy = k\cos(u)\sin(\phi)du - \frac{\cos(\phi)}{k}dt. \tag{11}$$

We may verify immediately that a solution of (11) is obtained by setting $k = 1$, $\phi = \frac{\pi}{2}$. Thus we have

$$x(t,u) = t, \quad y(t,u) = \sin(u)$$

which gives Lambert's equal area projection.

Remark 3.1. Setting $k = a$, $\phi = \varphi_0$, where a, φ_0 are arbitrary constants, the formulas

$$x(t,u) = a\sin(u)\cos(\varphi_0) + \frac{t\sin(\varphi_0)}{a}, \quad y(t,u) = a\sin(u)\sin(\varphi_0) - \frac{t\cos(\varphi_0)}{a}$$

determine a more general solution of (11). This solution differs from the previous one only in that the meridians are no longer normal to the x-axis. Actually, it is not hard to show that if k is constant in the relation (11) then the angle ϕ must be constant.

Now we can prove Theorem 1.1, which is restated below.

Theorem 1.1*. *For each* $f \in \mathcal{Q}$,

(i) *if* $m_f(t,u) \geq 1$ *or* $m_f(t,u) \leq \cos(u)$, *then*

$$D_{f_0}(t,u) = \frac{1}{\cos^2(u)} \leq D_f(t,u);$$

(ii) *if* $\cos(u) \leq m_f(t,u) \leq 1$, *then*

$$D_{f_0}(t,u) = \frac{1}{\cos^2(u)} \geq D_f(t,u).$$

Proof. From (11) we have

$$\frac{\partial x}{\partial(u)} = k\cos(u)\cos(\phi), \quad \frac{\partial x}{\partial t} = \frac{\sin(\phi)}{k},$$

$$\frac{\partial y}{\partial u} = k\cos(u)\sin(\phi), \quad \frac{\partial y}{\partial t} = -\frac{\cos(\phi)}{k}.$$

Therefore, we may compute immediately the Jacobian of f and verify that it sends the tangent vectors $v_1 = (1,0)$ and $v_2 = (0,1)$ of $T_{(t,u)}E^2$ to the perpendicular vectors $w_1 = \big(\sin\big(\frac{\phi}{k}\big), -\cos\big(\frac{\phi}{k}\big)\big)$ and $w_2 = (k\cos(u)(\phi), k\cos(u)\sin(\phi))$ of $T_{f(t,u)}\mathbb{E}^2$. Thus Lemma 2.3 can be applied and we have

$$D_f(t,u) = \frac{1}{\cos(u)}\bar{D}_f(t,u).$$

In order to compute the term $\bar{D}_f(t,u)$, we use formula (3) and the formulas

$$A(t,u) = f_z = \frac{1}{2}(f_t - i f_u) = \frac{1}{2}\Big(\Big(\frac{\partial x}{\partial t} + \frac{\partial y}{\partial u}\Big) + i\Big(\frac{\partial y}{\partial t} - \frac{\partial x}{\partial u}\Big)\Big), \tag{12a}$$

$$B(t,u) = f_{\bar{z}} = \frac{1}{2}(f_t + i f_u) = \frac{1}{2}\Big(\Big(\frac{\partial x}{\partial t} - \frac{\partial y}{\partial u}\Big) + i\Big(\frac{\partial y}{\partial t} + \frac{\partial x}{\partial u}\Big)\Big). \tag{12b}$$

Therefore the equality

$$f_z = \frac{1}{2}\Big(\Big(\frac{\sin(\phi)}{k} + k\cos(u)\sin(\phi)\Big) + i\Big(-\frac{\cos(\phi)}{k} - k\cos(u)\cos(\phi)\Big)\Big)$$

implies

$$\begin{aligned}
|f_z| &= \frac{1}{2}\sqrt{\Big(\frac{\sin(\phi)}{k} + k\cos(u)\sin(\phi)\Big)^2 + \Big(-\frac{\cos(\phi)}{k} - k\cos(u)\cos(\phi)\Big)^2} \\
&= \frac{1}{2}\sqrt{\frac{1}{k^2} + k^2\cos^2(u) + 2\cos(u)} \\
&= \frac{1}{2}\sqrt{\Big(\frac{1}{k} + k\cos(u)\Big)^2} \\
&= \frac{1}{2}\Big|\frac{1}{k} + k\cos(u)\Big|.
\end{aligned}$$

So,

$$|A(t,u)| = \frac{1}{2}\Big|\frac{1}{k(t,u)} + k(t,u)\cos(u)\Big| = \frac{1}{2}\Big(\frac{1}{k(t,u)} + k(t,u)\cos(u)\Big) > 0. \tag{13a}$$

Similarly, we may prove that

$$|B(t,u)| = \frac{1}{2}\Big|\frac{1}{k(t,u)} - k(t,u)\cos(u)\Big|. \tag{13b}$$

We also have

$$1 \le 1/\sqrt{\cos(u)} \le \frac{1}{\cos(u)}.$$

These quantities separate the real line into four intervals. Therefore, we distinguish the following cases, depending on which interval the value $k(t,u)$ is located in.

CASE I: $k(t,u) \geq 1/\sqrt{\cos(u)}$. This is equivalent to

$$m(t,u) = k(t,u)\cos(u) \geq \sqrt{\cos(u)}.$$

Then, $(k(t,u))^2 \geq 1/\cos(u) \implies 1/k(t,u) - k(t,u)\cos(u) \leq 0$ and from (13) we have

$$|B(t,u)| = \frac{1}{2}\Big|\frac{1}{k(t,u)} - k(t,u)\cos(u)\Big| = \frac{1}{2}\Big(-\frac{1}{k(t,u)} + k(t,u)\cos(u)\Big). \quad (14)$$

Therefore, from (13a) and (14), we have

$$\bar{D}_f(t,u) = \frac{|A(t,u)| + |B(t,u)|}{|A(t,u)| - |B(t,u)|} = (k(t,u))^2\cos(u). \quad (15)$$

Now we distinguish the following two subcases.

SUBCASE $\mathrm{I_a}$: $k(t,u) \geq 1/\cos(u)$. We have $m(t,u) = k(t,u)\cos(u) \geq 1$, and thus

$$k(t,u) \geq \frac{1}{\cos(u)} \implies (k(t,u))^2 \geq \frac{1}{\cos^2(u)}.$$

Combining the last relation with (15),

$$\bar{D}_f(t,u) = (k(t,u))^2\cos(u) \geq \frac{1}{\cos(u)} = \bar{D}_{f_0}(t,u)$$

and Lemma 2.3 implies that

$$D_{f_0}(t,u) = \frac{1}{\cos^2(u)} \leq D_f(t,u) = (k(t,u))^2.$$

SUBCASE $\mathrm{I_b}$: $1/\sqrt{\cos(u)} \leq k(t,u) \leq 1/\cos(u)$. We have $\sqrt{\cos(u)} \leq m(t,u) \leq 1$, and therefore the equality

$$\bar{D}_f(t,u) = (k(t,u))^2\cos(u) \leq \bar{D}_{f_0}(t,u) = \frac{1}{\cos(u)}$$

implies

$$D_f(t,u) = (k(t,u))^2 \leq D_{f_0}(t,u) = \frac{1}{\cos^2(u)}.$$

CASE II: $k(t,u) \le 1/\sqrt{\cos(u)}$. This is equivalent to

$$m(t,u) = k(t,u)\cos(u) \le \sqrt{\cos(u)}.$$

First, we remark that

$$k(t,u) \le 1/\sqrt{\cos(u)} \iff 1/k(t,u) - k(t,u)\cos(u) \ge 0$$
$$\implies |B| = \frac{1}{2}\left|\frac{1}{k(t,u)} - k(t,u)\cos(u)\right| = \frac{1}{2}\Big(\frac{1}{k(t,u)} - k(t,u)\cos(u)\Big).$$

Therefore,

$$\bar{D}_f(t,u) = \frac{|A(t,u)| + |B(t,u)|}{|A(t,u)| - |B(t,u)|} = \frac{1}{(k(t,u))^2\cos(u)}. \tag{16}$$

We distinguish the following two subcases.

SUBCASE $\mathrm{II_a}$: $k(t,u) \le 1$. We have $m(t,u) \le \cos(u)$. So, from (16), the equality

$$\bar{D}_{f_0}(t,u) = \frac{1}{\cos(u)} \le \bar{D}_f = \frac{1}{(k(t,u))^2\cos(u)}$$

implies

$$D_{f_0}(t,u) = \frac{1}{\cos^2(u)} \le D_f = \frac{1}{(k(t,u))^2\cos^2 u}.$$

SUBCASE $\mathrm{II_b}$: $1 \le k(t,u) \le 1/\sqrt{\cos(u)}$. We have $m(t,u) \le \sqrt{\cos(u)}$. Thus the equality

$$\bar{D}_f(t,u) = \frac{1}{k^2\cos(u)} \le \bar{D}_{f_0}(t,u) = \frac{1}{\cos(u)}$$

implies

$$D_f(t,u) = \frac{1}{k^2\cos^2(u)} \le D_{f_0}(t,u) = \frac{1}{\cos^2(u)}. \qquad \square$$

Remark 3.2. For a geographical projection $f \in \mathcal{Q}$ we may compute the distortion $\bar{D}_f(t,u)$ at a point $(t,u) \in S_0$ in a more direct way, without using formula (3). Indeed, we have seen that the differential $df_{(t,u)}$ sends the tangent vectors $v_1 = (1,0)$ and $v_2 = (0,1)$ of $T_{(t,u)}E^2$ to the perpendicular vectors $w_1 = (\sin(\phi)/k, -\cos(\phi)/k)$ and $w_2 = (k\cos(u)\cos(\phi), k\cos(u)\sin(\phi))$ of $T_{f(t,u)}E^2$ respectively. On the other hand, we have $\|w_1\| = 1/k$ and $\|w_2\| = k\cos(u)$. Therefore, considering the cases $\mathrm{I_a}$, $\mathrm{I_b}$, $\mathrm{II_a}$, and $\mathrm{II_b}$ above, we may find which of the the values $\|w_1\|$ and $\|w_2\|$ are greater and calculate the term $\bar{D}_f(t,u)$.

Now, we will show that there exist geographical projections f from S_0 to E^2 which satisfy $m_f(t,u) \leq \cos(u)$ for each $(t,u) \in (-\pi,\pi) \times \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. In fact in [2] (see p. 36ff of the corresponding translation), Euler resolves the system of equations (11) finding particular solutions as follows.

Setting $v = \sin(u)$ in (11), we get

$$dx = k\cos(\phi)dv + \frac{\sin(\phi)}{k}dt, \quad dy = k\sin(\phi)dv - \frac{\cos(\phi)}{k}dt.$$

Assuming that k is a function only of v which is equal to V and the angle ϕ is a function only of t and equal to T,

$$dx = V\cos(T)dv + \frac{\sin(T)}{V}dt, \quad dy = V\sin(T)dv - \frac{\cos(T)}{V}dt. \tag{17}$$

From equations (17) we obtain

$$x = \cos(T)\int Vdv = \frac{1}{V}\int \sin(T)dt \tag{18a}$$

and

$$y = \sin(T)\int Vdv = -\frac{1}{V}\int \cos(T)dt. \tag{18b}$$

Therefore,

$$V\int Vdv = \frac{1}{\cos(T)}\int \sin(T)dt = a$$

and

$$-V\int Vdv = \frac{1}{\sin(T)}\int \cos(T)dt = -a.$$

where a is a constant. Differentiating the relation $\int Vdv = a/V$ with respect to the variable v, we have

$$Vdv = -\frac{adV}{V^2} \implies dv = -\frac{adV}{V^3}$$

and integrating we have

$$v + c = \frac{a}{2V^2} \implies V = \sqrt{\frac{a}{2(v+c)}}, \tag{19}$$

where c is a constant and $a/(v+c) > 0$.

Similarly, for the function T, we have

$$\int \sin(T)dt = a\cos(T) \quad \text{and} \quad \int \cos(T)dt = -a\sin(T).$$

Differentiating one of these relations we have

$$dT = -\frac{dt}{a} \implies T = -\frac{t}{a}.$$

Finally, for (19) by a simple integration, we obtain

$$\int Vdv = \sqrt{2a(v+c)},$$

hence

$$\begin{cases} x = \sqrt{2a(v+c)}\cos\left(\frac{t}{a}\right) = \sqrt{2a(\sin(u)+c)}\cos\left(\frac{t}{a}\right), \\ y = -\sqrt{2a(v+c)}\sin\left(\frac{t}{a}\right) = \sqrt{2a(\sin(u)+c)}\sin\left(\frac{t}{a}\right). \end{cases} \tag{20}$$

From (20), it follows that

$$\begin{cases} \frac{y}{x} = -\tan\left(\frac{t}{a}\right), \\ \sqrt{x^2+y^2} = \sqrt{2a(v+c)}. \end{cases} \tag{21}$$

From the first equation in (21), it is clear that all meridians are represented by straight lines emerging as radii from a single fixed point. The second formula implies that all parallels are portrayed as concentric circles. These projections are referred to as *azimuthal geographical projections*.

Finally from (20), we have

$$m^2(t,u) = \left(\frac{\partial x}{\partial u}\right)^2 + \left(\frac{\partial y}{\partial u}\right)^2 = \frac{\cos^2(u)}{2a(\sin(u)+c)}$$

and it is not difficult to see that there are constants a and c such that $m(t,u) \leq \cos(u)$ for each $(t,u) \in S_0$.

Acknowledgements. The author would like to thank Vincent Alberge and Athanase Papadopoulos for very helpful suggestions and observations which improved this work in many ways.

References

[1] C. Charitos and I. Papadoperakis, "On the non-existence of a perfect map from the 2-sphere to the Euclidean plane," in *Eighteen essays in non-Euclidean geometry* (V. Alberge and A. Papadopoulos, eds.), European Mathematical Society (EMS), Zürich, 2019, Chapter 9, 125–134.
☞ 141, 142

[2] L. Euler, "De repraesentatione superficiei sphaericae super plano," *Acta Academiae Scientarum Imperialis Petropolitinae* (1777), 107–132. Reprinted in *Opera omnia,* Series 1, Vol. 28, 248–275. Tranlation by G. Heine available on the Euler Archive.
`http://eulerarchive.maa.org/pages/E490.html`
☞ 135, 137, 141, 147

[3] L. Euler, "De proiectione geographica superficiei sphaericae," *Acta Academiae Scientarum Imperialis Petropolitinae* (1777), 133–142. Reprinted in *Opera omnia,* Series 1, Vol. 28, 276–287. Tranlation by G. Heine avaliable on the Euler Archive.
`http://eulerarchive.maa.org/pages/E491.html`
☞ 135

[4] L. Euler, "De proiectione geographica Deslisliana in mappa generali imperii russici usitata," *Acta Academiae Scientarum Imperialis Petropolitinae* (1777), 143–153. Reprinted in *Opera omnia,* Series 1, Vol. 28, 288–297. Tranlation by G. Heine avaliable on the Euler Archive.
`http://eulerarchive.maa.org/pages/E492.html`
☞ 136

[5] A. Fletcher and V. Markovic, *Quasiconformal maps and Teichmüller theory,* Oxford Graduate Texts in Mathematics 11. Oxford University Press, Oxford, 2007.
MR 2269887 Zbl 1103.30002 ☞ 139

[6] J.-L. Lagrange, "Sur la construction des cartes géographiques, Premier mémoire," *Nouveaux mémoires de l'Académie royale des Sciences et Belles-Lettres de Berlin* 1779, 161–185. Reprinted in *Œuvres de Lagrange*, Vol. 4, 637–664.
☞ 136

[7] J.-L. Lagrange, "Sur la construction des cartes géographiques, Second mémoire," *Nouveaux mémoires de l'Académie royale des Sciences et Belles-Lettres de Berlin* (1779), 186–210. Reprinted in *Œuvres de Lagrange*, Vol. 4, 664–692.
☞ 136

[8] J. H. Lambert, *Beiträge zum Gebrauche der Mathematik und deren Anwendung,* Vol. III, im Verlage des Buchladens der Realschule, Berlin 1772.
☞ 137

[9] A. Papadopoulos, "Quasiconformal mappings, from Ptolemy's geography to the work of Teichmüller," in *Uniformization, Riemann–Hilbert correspondence, Calabi–Yau manifolds, and Picard–Fuchs equations* (L. Ji and S.-T. Yau, eds.), Advanced Lectures in Mathematics 42, Higher Education Press Beijing, and International Press, Boston, 2018, 237–315.
☞ 136

[10] A. Papadopoulos, "Nicolas-Auguste Tissot: a link between cartography and quasiconformal theory," *Arch. Hist. Exact Sci.* 71 (2017), no. 4, 319–336.
MR 3665710 Zbl 1381.01007 ☞ 136

[11] N.-A. Tissot, *Mémoire sur la représentation des surfaces et les projections des cartes géographiques,* Gauthier-Villars, Paris, 1881.
☞ 136, 138

[12] John P. Snyder, *Map projections,* A Working Manual, U.S. Geological Survey Professional Paper 1395, Washington, 1987.
☞ 137

Chapter 11

Area and volume in non-Euclidean geometry

Nikolay Abrosimov and Alexander Mednykh

Contents

1. Geometry of polygons 151
2. Geometry of polyhedra 167
References 183

1 Geometry of polygons

1.1 Heron's theorem

Heron of Alexandria (c. 60 BC) is credited with the following formula that relates the area S of a triangle to its side lengths a, b, c:

$$S^2 = (s-a)(s-b)(s-c)s, \tag{1}$$

where $s = \frac{1}{2}(a+b+c)$ is the semiperimeter.

In the present section of our work we deal with the hyperbolic plane instead of the Euclidean one. The hyperbolic plane under consideration is equipped with a Riemannian metric of constant curvature $k = -1$. All the necessary definitions from hyperbolic geometry can be found in the book [8].

By definition, a *cyclic polygon* in the hyperbolic plane is a convex polygon inscribed in a curve of constant geodesic curvature (i.e. circle, horocycle or one branch of an equidistant line). Useful information about cyclic polygons can be found in [62] and [63]. In particular, it is shown in [62] that any cyclic polygon in the hyperbolic plane is uniquely determined (up to isometry) by the ordered sequence of its side lengths. In addition, among all hyperbolic polygons with fixed positive side lengths there exist polygons of maximal area. Every such a maximal polygon is cyclic (see [63]).

The following four non-Euclidean versions of Heron's formula in the hyperbolic plane are known since a long time.

Theorem 1.1 (Heron's formula in $\mathbb{H}^2$). *Let T be a hyperbolic triangle with side lengths a, b, c. Then its area $S = S(T)$ is given by each of the following formulas*:

(i) *sine of $\frac{1}{2}$ area formula,*

$$\sin^2\left(\frac{1}{2}S\right) = \frac{\sinh(s-a)\sinh(s-b)\sinh(s-c)\sinh(s)}{4\cosh^2\left(\frac{1}{2}a\right)\cosh^2\left(\frac{1}{2}b\right)\cosh^2\left(\frac{1}{2}c\right)};$$

(ii) *tangent of $\frac{1}{4}$ area formula,*

$$\tan^2\left(\frac{1}{4}S\right) = \tanh\left(\frac{1}{2}(s-a)\right)\tanh\left(\frac{1}{2}(s-b)\right)\tanh\left(\frac{1}{2}(s-c)\right)\tanh\left(\frac{1}{2}s\right);$$

(iii) *sine of $\frac{1}{4}$ area formula,*

$$\sin^2\left(\frac{1}{4}S\right) = \frac{\sinh\left(\frac{1}{2}(s-a)\right)\sinh\left(\frac{1}{2}(s-b)\right)\sinh\left(\frac{1}{2}(s-c)\right)\sinh\left(\frac{1}{2}s\right)}{\cosh\left(\frac{1}{2}a\right)\cosh\left(\frac{1}{2}b\right)\cosh\left(\frac{1}{2}c\right)};$$

(iv) *Bilinski formula,*

$$\cos\left(\frac{1}{2}S\right) = \frac{\cosh(a)+\cosh(b)+\cosh(c)+1}{4\cosh\left(\frac{1}{2}a\right)\cosh\left(\frac{1}{2}b\right)\cosh\left(\frac{1}{2}c\right)}.$$

The first two formulas are contained in the book [8], p. 66. The third formula can be obtained by square root extraction of the product of the first two. The fourth one was derived by Stanko Bilinski in [14] (see also [63]).

It should be noted that the analogous formulas in spherical space were known before the hyperbolic ones. For example, the spherical version of (i) is called Cagnoli's Theorem (see [58], Section 100), (ii) is called Lhuilier's Theorem (see [58], Section 102), formula (iv) is due to Euler (see [25]). The proofs of (iii) and (iv) can be found in [58] (Section 103).

1.2 Ptolemy's theorem and annex

We recall a few well-known facts about cyclic quadrilaterals. A convex Euclidean quadrilateral with interior angles A, B, C, D is cyclic if and only if $A+C = B+D = \pi$. A similar result for a hyperbolic quadrilateral was obtained by V. F. Petrov [46] and L. Wimmer [65]. They proved the following result.

Theorem 1.2 (Petrov, 2009, [46]; Wimmer, 2011, [65]). *A convex hyperbolic quadrilateral with interior angles A, B, C, D is cyclic if and only if $A + C = B + D$.*

Note that the sum of angles of a hyperbolic quadrilateral is less than 2π. Hence, for any cyclic hyperbolic quadrilateral we have $A + C = B + D < \pi$.

It was shown in [62] that a cyclic n-gon is uniquely determined up to isometry by the lengths of its sides. Denote the side and diagonal lengths of a quadrilateral as indicated in Figure 1. Then, in the Euclidean case, a quadrilateral is cyclic if and only if $ef = a\,c + b\,d$. This is Ptolemy's theorem. A spherical version of Ptolemy's theorem was established by Darboux and Frobenius. It can be found in the book [37] (p. 180, Proposition 4). In hyperbolic geometry, Ptolemy's identity was proposed by T. Kubota [34] and J. E. Valentine [60].

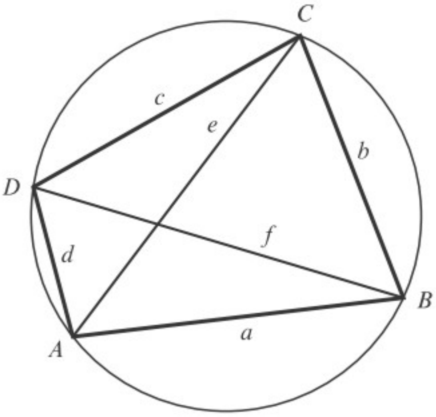

Figure 1. A cyclic quadrilateral Q with angles A, B, C, D

Theorem 1.3 (Ptolemy's identity in $\mathbb{H}^2$). *A convex hyperbolic quadrilateral with side lengths a, b, c, d and diagonal lengths e, f is cyclic if and only if*

$$\sinh\Big(\frac{1}{2}e\Big)\sinh\Big(\frac{1}{2}f\Big) = \sinh\Big(\frac{1}{2}a\Big)\sinh\Big(\frac{1}{2}c\Big) + \sinh\Big(\frac{1}{2}b\Big)\sinh\Big(\frac{1}{2}d\Big).$$

An important supplement to Ptolemy's theorem is the following property of a cyclic quadrilateral in the Euclidean plane. Its side and diagonal lengths are related by the equation

$$\frac{e}{f} = \frac{a\,d + b\,c}{a\,b + c\,d}. \tag{2}$$

Together with Ptolemy's theorem this equation allows one to express the diagonal lengths of a cyclic quadrilateral in terms of its side lengths.

It was noted in [28] that a similar relationship between the sides and diagonals of a cyclic quadrilateral is valid also in hyperbolic geometry. To make the statement true, one can replace the side length a by the quantity $s(a) = \sinh\left(\frac{1}{2}a\right)$. In particular, formula (2) can be rewritten in the following way.

Theorem 1.4. *The side lengths a, b, c, d and diagonal lengths e, f of a cyclic hyperbolic quadrilateral are related by the following equation*

$$\frac{s(e)}{s(f)} = \frac{s(a)s(d) + s(b)s(c)}{s(a)s(b) + s(c)s(d)}.$$

By making use of Theorems 1.3 and 1.4 we derive the following formulas for the diagonal lengths e, f of a cyclic hyperbolic quadrilateral. Then we have

$$s^2(e) = \frac{s(a)s(d) + s(b)s(c)}{s(a)s(b) + s(c)s(d)}(s(a)s(c) + s(b)s(d)), \tag{3}$$

$$s^2(f) = \frac{s(a)s(b) + s(c)s(d)}{s(a)s(d) + s(b)s(c)}(s(a)s(c) + s(b)s(d)). \tag{4}$$

Note that formulas (3) and (4) take place also in Euclidean and spherical geometries. In these cases, instead of the function $s(a)$ one should take the functions $s(a) = a$ and $s(a) = \sin\left(\frac{1}{2}a\right)$ respectively. See [23] and [28] for the arguments in the spherical case.

The above theorems will be used in the next section to obtain a few versions of Brahmagupta's formula for a cyclic hyperbolic quadrilateral.

1.3 Brahmagupta's formula

For polygons with more than three sides, the side lengths do not in general determine the area, but they do if the polygon is convex and cyclic (inscribed in a circle). Brahmagupta, in the seventh century, gave the following formula for a convex cyclic quadrilateral with side lengths a, b, c, d:

$$S^2 = (s-a)(s-b)(s-c)(s-d), \tag{5}$$

where $s = \frac{1}{2}(a+b+c+d)$. See [21], p. 56 for an elementary proof. An interesting consideration of the problem can be found in Möbius's paper [41]. Independently, D. P. Robbins [50] and V. V. Varfolomev [61] found a way to generalize these formulas. The main idea of both papers was to determine the squared area S^2 as a root of an algebraic equation whose coefficients are integer polynomials in the squares of the side lengths. See also [26], [18], and [51] for more detailed considerations.

One can consider Brahmagupta's formula as a generalization of Heron's formula (1).

In this section we consider the four versions of Brahmagupta's formula for a cyclic hyperbolic quadrilateral given by the second named author [38]. They are generalizations of the respective statements (i)–(iv) of Theorem 1.1.

In particular, the first statement (i) has the following analog.

Theorem 1.5 (sine of $\frac{1}{2}$ area formula). *Let Q be a cyclic hyperbolic quadrilateral with side lengths a, b, c, d. Then its area $S = S(Q)$ is given by the formula*

$$\sin^2\Big(\frac{1}{2}S\Big) = \frac{1}{4\mathfrak{N}}\sinh(s-a)\sinh(s-b)\sinh(s-c)\sinh(s-d)(1-\varepsilon),$$

where

$$\mathfrak{N} = \cosh^2\Big(\frac{1}{2}a\Big)\cosh^2\Big(\frac{1}{2}b\Big)\cosh^2\Big(\frac{1}{2}c\Big)\cosh^2\Big(\frac{1}{2}d\Big),$$

$$\varepsilon = \frac{\sinh\big(\frac{1}{2}a\big)\sinh\big(\frac{1}{2}b\big)\sinh\big(\frac{1}{2}c\big)\sinh\big(\frac{1}{2}d\big)}{\cosh\big(\frac{1}{2}(s-a)\big)\cosh\big(\frac{1}{2}(s-b)\big)\cosh\big(\frac{1}{2}(s-c)\big)\cosh\big(\frac{1}{2}(s-d)\big)}$$

and

$$s = \frac{1}{2}(a+b+c+d).$$

Note that the value ε vanishes if $d = 0$. In this case, we get formula (i) again.

Proofs of this theorem and subsequent theorems will be given later, at the end of this section.

The second statement for the case of a hyperbolic quadrilateral has the following form.

Theorem 1.6 (tangent of $\frac{1}{4}$ area formula). *Let Q be a cyclic hyperbolic quadrilateral with side lengths a, b, c, d. Then its area $S = S(Q)$ is given by the formula*

$$\tan^2\Big(\frac{1}{4}S\Big) = \frac{\sinh\big(\frac{1}{4}(-a+b+c+d)\big)\sinh\big(\frac{1}{4}(a-b+c+d)\big)}{\cosh\big(\frac{1}{4}(a+b-c-d)\big)\cosh\big(\frac{1}{4}(a-b+c-d)\big)} \frac{\sinh\big(\frac{1}{4}(a+b-c+d)\big)\sinh\big(\frac{1}{4}(a+b+c-d)\big)}{\cosh\big(\frac{1}{4}(a-b-c+d)\big)\cosh\big(\frac{1}{4}(a+b+c+d)\big)}.$$

It follows from Theorem 1.5 that for any $a, b, c, d \neq 0$ we have $1-\varepsilon > 0$ and $\varepsilon > 0$. Hence, $0 < \varepsilon < 1$. Taking into account these inequalities as an immediate consequence of Theorems 1.5 and 1.6 we obtain the following corollary.

Corollary 1.7. *For any cyclic hyperbolic quadrilateral the following inequalities take place*:

$$\sin^2\left(\frac{1}{2}S\right) < \frac{1}{4\mathfrak{N}}\sinh(s-a)\sinh(s-b)\sinh(s-c)\sinh(s-d)$$

and

$$\tan^2\left(\frac{1}{2}S\right) > \tanh\left(\frac{1}{2}(s-a)\right)\tanh\left(\frac{1}{2}(s-b)\right)\tanh\left(\frac{1}{2}(s-c)\right)\tanh\left(\frac{1}{2}(s-d)\right).$$

By squaring the product of the formulas in Theorems 1.5 and 1.6 we obtain the following result. It can be considered as a direct generalization of statement (iii) in Theorem 1.1.

Theorem 1.8 (sine of $\frac{1}{4}$ area formula). *Let Q be a cyclic hyperbolic quadrilateral with side lengths a, b, c, d. Then its area $S = S(Q)$ is given by the formula*

$$\sin^2\left(\frac{1}{4}S\right) = \frac{1}{\mathfrak{N}'}\sinh\left(\frac{1}{2}(s-a)\right)\sinh\left(\frac{1}{2}(s-b)\right)\sinh\left(\frac{1}{2}(s-c)\right)\sinh\left(\frac{1}{2}(s-d)\right),$$

where

$$\mathfrak{N}' = \cosh\left(\frac{1}{2}a\right)\cosh\left(\frac{1}{2}b\right)\cosh\left(\frac{1}{2}c\right)\cosh\left(\frac{1}{2}d\right)$$

and

$$s = \frac{1}{2}(a+b+c+d).$$

The analogous formula in spherical space can be obtained by replacing sinh and cosh by sin and cos correspondingly (see [37], p. 182, Proposition 5).

Consider a circumscribed quadrilateral Q with side lengths a, b, c, d (Figure 2). In this case we have $a+c = b+d$, hence $s-a = c,\ s-b = d,\ s-c = a,\ s-d = b$. As a result we obtain the following assertion.

Corollary 1.9 (Brahmagupta's formula for bicentric quadrilateral). *Let Q be a bicentric* (*i.e. inscribed and circumscribed*) *hyperbolic quadrilateral with side lengths a, b, c, d. Then its area $S = S(Q)$ is given by the formula*

$$\sin^2\left(\frac{1}{4}S\right) = \tanh\left(\frac{1}{2}a\right)\tanh\left(\frac{1}{2}b\right)\tanh\left(\frac{1}{2}c\right)\tanh\left(\frac{1}{2}d\right).$$

The analogous formula in spherical space can be given by replacing tanh with tan (see [37], p. 46). A Euclidean version of this result is well known (see, e.g., [31]). In that case

$$S^2 = a\,b\,c\,d.$$

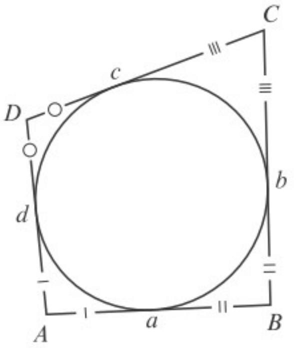

Figure 2. A circumscribed cyclic quadrilateral Q with side lengths a, b, c, d

The next theorem [38] presents a version of the Bilinski formula for a cyclic quadrilateral.

Theorem 1.10 (Bilinski formula). *Let Q be a cyclic hyperbolic quadrilateral with side lengths a, b, c, d. Then its area $S = S(Q)$ is given by the formula*

$$\cos\left(\frac{1}{2}S\right) = \frac{1}{4\mathfrak{N}'}\Big(\cosh a + \cosh b + \cosh c + \cosh d \\ - 4\sinh\left(\frac{1}{2}a\right)\sinh\left(\frac{1}{2}b\right)\sinh\left(\frac{1}{2}c\right)\sinh\left(\frac{1}{2}d\right)\Big).$$

Now we outline the proof of Theorems 1.5, 1.6, 1.8, and 1.10.

Proof. Consider a cyclic hyperbolic quadrilateral Q with side lengths a, b, c, d and interior angles A, B, C, D shown in Figure 1. By the Gauss–Bonnet formula we get the area

$$S = S(Q) = 2\pi - A - B - C - D.$$

To prove Theorem 1.5 let us find the quantities $\sin^2\left(\frac{1}{4}S\right)$ and $\cos^2\left(\frac{1}{4}S\right)$ in terms of a, b, c, d. Since $A + C = B + D$ (see Proposition 1.2) we have

$$2\sin^2\left(\frac{1}{4}S\right) = 1 - \cos\left(\frac{1}{2}S\right) = 1 - \cos(\pi - (A + C)) = 1 + \cos(A + C).$$

Hence,

$$\sin^2\left(\frac{1}{4}S\right) = \frac{1}{2}(1 + \cos(A)\cos(C) - \sin(A)\sin(C)). \tag{6}$$

Now we show that $\cos(A)$, $\cos(C)$, and the product $\sin(A)\sin(C)$ can be expressed in terms of elementary functions in a, b, c, d. To find $\cos(A)$ we use the Cosine rule for the hyperbolic triangle ABD.

$$\cos(A) = \frac{\cosh(a)\cosh(d) - \cosh(f)}{\sinh(a)\sinh(d)}. \tag{7}$$

We note that $\cosh(f) = 2s^2(f) + 1$, $\cosh(a) = 2s^2(a) + 1$, and $\cosh(d) = 2s^2(d) + 1$. Putting these identities into equations (4) and (7) we express $\cos(A)$ in terms of a, b, c, d. After a straightforward calculation we obtain

$$\begin{aligned}\cos(A) = \frac{1}{\widetilde{\mathfrak{N}}}(s^2(a) - s^2(b) - s^2(c) + s^2(d) \\ + 2s(a)s(b)s(c)s(d) + 2s^2(a)s^2(d)),\end{aligned} \tag{8}$$

where

$$\widetilde{\mathfrak{N}} = 2(s(a)s(d) + s(b)s(c))\cosh\left(\frac{1}{2}a\right)\cosh\left(\frac{1}{2}d\right).$$

In a similar way we get the formula

$$\begin{aligned}\cos(C) = \frac{1}{\widetilde{\mathfrak{N}}'}(-s^2(a) + s^2(b) + s^2(c) - s^2(d) \\ + 2s(a)s(b)s(c)s(d) + 2s^2(b)s^2(c)),\end{aligned} \tag{9}$$

where

$$\widetilde{\mathfrak{N}}' = 2(s(a)s(d) + s(b)s(c))\cosh\left(\frac{1}{2}b\right)\cosh\left(\frac{1}{2}c\right)$$

Note that $\sin(A)\sin(C) > 0$ and $\sin^2(A)\sin^2(C) = (1 - \cos^2(A))(1 - \cos^2(C))$. Then we take square roots in the latter equation, where $\cos(A)$ and $\cos(C)$ are given by formulas (8) and (9). Thus we express the product $\sin(A)\sin(C)$ in terms of a, b, c, d. Substituting $\cos(A)$, $\cos(C)$, and $\sin(A) \cdot \sin(C)$ into (6) and simplifying we get

$$\begin{aligned}\sin^2\left(\frac{1}{4}S\right) = \frac{1}{\mathfrak{N}'}\sinh\left(\frac{1}{4}(-a + b + c + d)\right)\sinh\left(\frac{1}{4}(a - b + c + d)\right) \\ \sinh\left(\frac{1}{4}(a + b - c + d)\right)\sinh\left(\frac{1}{4}(a + b + c - d)\right).\end{aligned} \tag{10}$$

This proves Theorem 1.8.

In a similar way, from the identity $\cos^2\left(\frac{1}{4}S\right) = 1 + \cos\left(\frac{1}{2}S\right) = 1 - \cos(A + C)$ we have

$$\begin{aligned}\cos^2\left(\frac{1}{4}S\right) = \frac{1}{\mathfrak{N}'}\cosh\left(\frac{1}{4}(a + b - c - d)\right)\cosh\left(\frac{1}{4}(a - b + c - d)\right) \\ \cosh\left(\frac{1}{4}(a - b - c + d)\right)\cosh\left(\frac{1}{4}(a + b + c + d)\right).\end{aligned} \tag{11}$$

The following lemma can be easily proved by straightforward calculations.

Lemma 1.11. *The expression*

$$H = \frac{\cosh\left(\frac{1}{4}(a+b-c-d)\right)\cosh\left(\frac{1}{4}(a-b+c-d)\right)}{\cosh\left(\frac{1}{4}(-a+b+c+d)\right)\cosh\left(\frac{1}{4}(a-b+c+d)\right)} \frac{\cosh\left(\frac{1}{4}(a-b-c+d)\right)\cosh\left(\frac{1}{4}(a+b+c+d)\right)}{\cosh\left(\frac{1}{4}(a+b-c+d)\right)\cosh\left(\frac{1}{4}(a+b+c-d)\right)}$$

can be rewritten in the form $H = 1 - \varepsilon$, *where*

$$\varepsilon = \frac{\sinh\left(\frac{1}{2}a\right)\sinh\left(\frac{1}{2}b\right)\sinh\left(\frac{1}{2}c\right)\sinh\left(\frac{1}{2}d\right)}{\cosh\left(\frac{1}{2}(s-a)\right)\cosh\left(\frac{1}{2}(s-b)\right)\cosh\left(\frac{1}{2}(s-c)\right)\cosh\left(\frac{1}{2}(s-d)\right)}$$

and

$$s = \frac{1}{2}(a+b+c+d).$$

Taking four times the products of equations (10) and (11) we have

$$\sin^2\left(\frac{1}{2}S\right) = \frac{1}{4\mathfrak{N}}\sinh\left(\frac{1}{2}(-a+b+c+d)\right)\sinh\left(\frac{1}{2}(a-b+c+d)\right) \sinh\left(\frac{1}{2}(a+b-c+d)\right)\sinh\left(\frac{1}{2}(a+b+c-d)\right)H, \tag{12}$$

where H is the same as in Lemma 1.11.

Then the statement of Theorem 1.5 follows from equation (12), Lemma 1.11 and the evident identity $s-a = \frac{1}{2}(-a+b+c+d)$. To prove Theorem 1.6 we divide (10) by (11).

Finally, the Bilinski formula (Theorem 1.10) follows from the identity $\cos\left(\frac{1}{2}S\right) = \cos^2\left(\frac{1}{4}S\right) - \sin^2\left(\frac{1}{4}S\right)$ and the above mentioned equations (10) and (11). □

1.4 Bretschneider's formula

In 1842, Carl Bretschneider [15] related the area of an arbitrary Euclidean quadrilateral to its side lengths and the sum of two opposite angles. The area S of a Euclidean quadrilateral with side lengths a, b, c, d and opposite angles A and C is given by the formula

$$S^2 = (s-a)(s-b)(s-c)(s-d) - a\,b\,c\,d\cos^2\left(\frac{1}{2}(A+C)\right),$$

where $s = \frac{1}{2}(a+b+c+d)$ is the semiperimeter (for a proof see e.g. [47], p. 89).

The statement of the theorem remains valid if one substitutes $A+C$ with the sum of another pair of opposite angles $B+D$. By making use of the identity

$$A+B+C+D=2\pi$$

for any Euclidean quadrilateral we can rewrite the Bretschneider's theorem in the following more symmetric way

$$S^2=(s-a)(s-b)(s-c)(s-d)-a\,b\,c\,d\,\sin^2\Big(\frac{1}{4}(A-B+C-D)\Big).$$

Precisely, this statement allows a generalization for the case of a non-Euclidean quadrilateral. In the particular case of a cyclic quadrilateral (when $A+C=B+D$) we get a Brahmagupta's formula (5). Recall that the sum of angles is not equal to 2π anymore.

The following theorem gives a hyperbolic version of Bretschneider's formula.

Theorem 1.12 (Baigonakova and Mednykh, 2012, [11]). *The area S of a hyperbolic quadrilateral with side lengths a,b,c,d and angles A,B,C,D is given by the formula*

$$\begin{aligned}&\sin^2\Big(\frac{1}{4}S\Big)\\&\quad=\frac{1}{\mathfrak{N}'}\sinh\Big(\frac{1}{2}(s-a)\Big)\sinh\Big(\frac{1}{2}(s-b)\Big)\sinh\Big(\frac{1}{2}(s-c)\Big)\sinh\Big(\frac{1}{2}(s-d)\Big)\\&\qquad-\tanh\Big(\frac{1}{2}a\Big)\tanh\Big(\frac{1}{2}b\Big)\tanh\Big(\frac{1}{2}c\Big)\tanh\Big(\frac{1}{2}d\Big)\sin^2\Big(\frac{1}{4}(A-B+C-D)\Big),\end{aligned}$$

where $s=\frac{1}{2}(a+b+c+d)$ is the semiperimeter.

Proof. Consider the *Poincaré disk model* of the hyperbolic plane. This is the unit disc $U=\{z\in\mathbb{C}\colon |z|<1\}$, provided with the metric $ds=\frac{|dz|}{1-|z|^2}$. The boundary of U consists of the infinitely distant points and called *absolute*. The geodesics in this model are circle arcs and segments orthogonal to the absolute. Let z and w be arbitrary points of the disc U, and 0 be the origin. Then the hyperbolic and Euclidean distances between these points ($\rho(z,w)$ and $|z-w|$ correspondingly) are related by the formula ([13], p. 22–123)

$$|z-w|=\frac{\sinh\big(\frac{1}{2}\rho(z,w)\big)}{\cosh\big(\frac{1}{2}\rho(0,z)\big)\cosh\big(\frac{1}{2}\rho(0,w)\big)}. \tag{13}$$

Consider a hyperbolic quadrilateral $ABCD$ (Figure 3). We suppose that the vertex A coincides with the origin 0. From (13) we find the sides of the Euclidean quadrilateral $ABCD$ with respect to the hyperbolic lengths a, b, c, d, e, f. We have

$$AB = \tanh\left(\frac{1}{2}a\right), \quad BC = \frac{\sinh\left(\frac{1}{2}b\right)}{\cosh\left(\frac{1}{2}a\right)\cosh\left(\frac{1}{2}e\right)}, \tag{14a}$$

$$AD = \tanh\left(\frac{1}{2}d\right), \quad CD = \frac{\sinh\left(\frac{1}{2}c\right)}{\cosh\left(\frac{1}{2}e\right)\cosh\left(\frac{1}{2}d\right)}. \tag{14b}$$

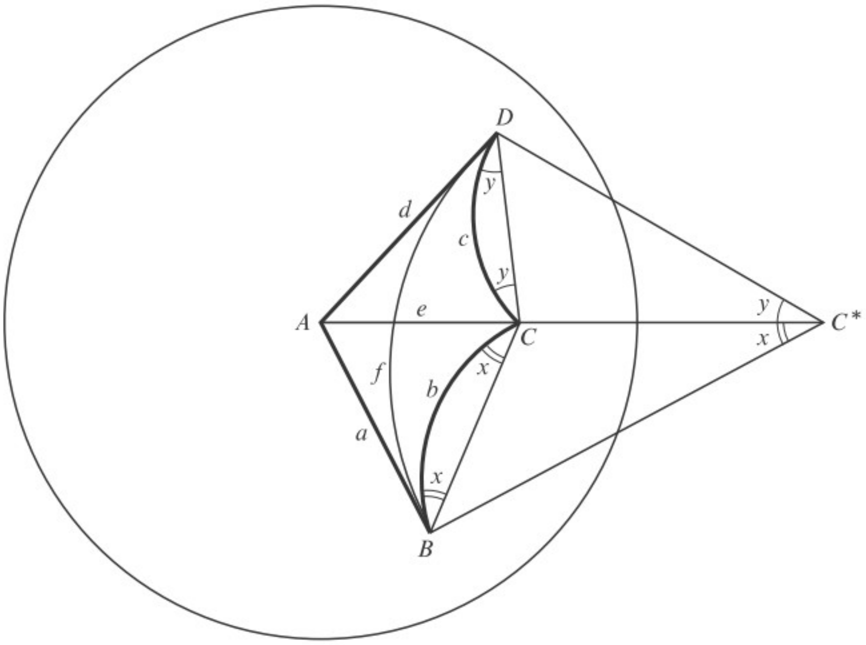

Figure 3. A hyperbolic quadrilateral $ABCD$ in the Poincaré model of $\mathbb{H}^2$

Let C^* be a point symmetric to C with respect to the absolute. To find the Euclidean distances BC^* and DC^* we use the identity $C^* = 1/C$ and the following formula from ([13], p. 123)

$$\tanh\left(\frac{1}{2}\rho(z, w)\right) = \left|\frac{z - w}{1 - z\,\overline{w}}\right|. \tag{15}$$

We get from (13) and (15)

$$BC^* = \frac{\cosh\left(\frac{1}{2}b\right)}{\cosh\left(\frac{1}{2}a\right)\sinh\left(\frac{1}{2}e\right)} \quad \text{and} \quad DC^* = \frac{\cosh\left(\frac{1}{2}c\right)}{\cosh\left(\frac{1}{2}d\right)\sinh\left(\frac{1}{2}e\right)}. \tag{16}$$

Denote the angle between the arc $\widehat{BC}$ and the chord BC by x, and the angle between the arc $\widehat{DC}$ and the chord DC by y. Then the angle measure of $\widehat{BC}$ is $2x$, while the angle measure of $\widehat{CD}$ is $2y$. Recall that the inscribed angle is equal to half of the corresponding cyclic arc. Then we obtain that the angles BC^*C and CC^*D are equal to x and y respectively. Denote by A, B, C, D the interior angles of the hyperbolic quadrilateral $ABCD$, and by $\widehat{A}, \widehat{B}, \widehat{C}, \widehat{D}$ the interior angles of the Eucledean quadrilateral $ABCD$. We have

$$\widehat{A} = A, \quad \widehat{B} = B + x, \quad \widehat{C} = C + x + y, \quad \widehat{D} = D + y.$$

Taking into account that $\widehat{A} + \widehat{B} + \widehat{C} + \widehat{D} = 2\pi$, we have

$$A + B + C + D + 2x + 2y = 2\pi.$$

Hence,

$$x + y = \frac{1}{2}(2\pi - A - B - C - D) = \frac{1}{2}S.$$

From the Euclidean triangle BC^*D we have

$$\sin^2\left(\frac{1}{2}(x + y)\right) = \sin^2\left(\frac{1}{4}S\right) = \frac{BD^2 - (BC^* - DC^*)^2}{4\,BC^* \cdot DC^*}.$$

Now using (14) and (16) we substitute the Euclidean lengths in the latter equation with the hyperbolic ones. Then

$$\begin{aligned}\sin^2\left(\frac{1}{4}S\right) = \frac{1}{4\mathfrak{N}'}\Big(\sinh\left(\frac{1}{2}e\right)\sinh\left(\frac{1}{2}f\right)\Big)^2 \\ - \Big(\cosh\left(\frac{1}{2}a\right)\cosh\left(\frac{1}{2}c\right) - \cosh\left(\frac{1}{2}b\right)\cosh\left(\frac{1}{2}d\right)\Big)^2.\end{aligned} \tag{17}$$

This formula expresses the area of a hyperbolic quadrilateral in terms of its sides and diagonals.

The next step is to find a similar formula for $\sin^2\left(\frac{1}{4}(A - B + C - D)\right)$. First, we need the following Bretschneider relation for the Euclidean quadrilateral $ABCD$ ([47], p. 85)

$$\begin{aligned}(AC \cdot BD)^2 = (AB \cdot CD)^2 + (BC \cdot AD)^2 \\ - 2AB \cdot BC \cdot CD \cdot AD \cos(\widehat{A} + \widehat{C}).\end{aligned}$$

Since

$$\widehat{A} + \widehat{C} = A + C + x + y = \pi + \frac{1}{2}(A - B + C - D),$$

we get

$$\begin{aligned}\cos\Big(\frac{1}{2}(A - B + C - D)\Big) &= -\cos(\widehat{A} + \widehat{C})\\ &= \frac{(AC \cdot BD)^2 - (AB \cdot CD)^2 - (BC \cdot AD)^2}{2\,AB \cdot BC \cdot CD \cdot AD},\end{aligned}$$

or

$$\sin^2\Big(\frac{1}{2}(A - B + C - D)\Big) = \frac{(AB \cdot CD + BC \cdot AD)^2 - (AC \cdot BD)^2}{4\,AB \cdot BC \cdot CD \cdot AD}.$$

As before, by (14) and (16) we obtain

$$\begin{aligned}&\sin^2\Big(\frac{1}{2}(A - B + C - D)\Big)\\ &\quad= \frac{\big(\sinh\big(\frac{1}{2}a\big)\sinh\big(\frac{1}{2}c\big) - \sinh\big(\frac{1}{2}b\big)\sinh\big(\frac{1}{2}d\big)\big)^2 - \big(\sinh\big(\frac{1}{2}e\big)\sinh\big(\frac{1}{2}f\big)\big)^2}{4\sinh\big(\frac{1}{2}a\big)\sinh\big(\frac{1}{2}b\big)\sinh\big(\frac{1}{2}c\big)\sinh\big(\frac{1}{2}d\big)}.\end{aligned} \tag{18}$$

Eliminating the expression $\big(\sinh\big(\frac{1}{2}e\big)\sinh\big(\frac{1}{2}f\big)\big)^2$ from (17) and (18) and making use of standard formulas we finally obtain

$$\begin{aligned}\sin^2\Big(\frac{1}{4}S\Big) = \frac{1}{\mathfrak{N}'}\sinh\Big(\frac{1}{2}(s-a)\Big)\sinh\Big(\frac{1}{2}(s-b)\Big)\sinh\Big(\frac{1}{2}(s-c)\Big)\sinh\Big(\frac{1}{2}(s-d)\Big)\\ - \tanh\Big(\frac{1}{2}a\Big)\tanh\Big(\frac{1}{2}b\Big)\tanh\Big(\frac{1}{2}c\Big)\tanh\Big(\frac{1}{2}d\Big)\\ \sin^2\Big(\frac{1}{4}(A - B + C - D)\Big).\end{aligned}$$

□

The analogous theorem holds in spherical geometry as well (for more details see the recent work [10] or the old papers [12] and [29]).

Theorem 1.13 (Baigonakova and Mednykh, 2012, [11]). *The area S of a spherical quadrilateral with side lengths a, b, c, d and angles A, B, C, D is given by the formula*

$$\begin{aligned}\sin^2\Big(\frac{1}{4}S\Big) &= \frac{\sin\big(\frac{1}{2}(s-a)\big)\sin\big(\frac{1}{2}(s-b)\big)\sin\big(\frac{1}{2}(s-c)\big)\sin\big(\frac{1}{2}(s-d)\big)}{\cos\big(\frac{1}{2}a\big)\cos\big(\frac{1}{2}b\big)\cos\big(\frac{1}{2}c\big)\cos\big(\frac{1}{2}d\big)}\\ &\quad - \tan\Big(\frac{1}{2}a\Big)\tan\Big(\frac{1}{2}b\Big)\tan\Big(\frac{1}{2}c\Big)\tan\Big(\frac{1}{2}d\Big)\sin^2\Big(\frac{1}{4}(A - B + C - D)\Big),\end{aligned}$$

where $s = \frac{1}{2}(a + b + c + d)$ is the semiperimeter.

1.5 Casey's theorem

John Casey generalized Ptolemy's theorem in the following way (see [16], p. 103).

Theorem 1.14 (Casey's theorem). *Let four circles* O_1, O_2, O_3, O_4 *on the Euclidean plane* $\mathbb{E}^2$ *be internally tangent to a circle* O *in the prescribed order. Denote by* t_{ij} *the length of the common external tangent of the circles* O_i, O_j*. Then*

$$t_{12}\, t_{34} + t_{23}\, t_{14} = t_{13}\, t_{24}\,.$$

We say that a common tangent of two circles is *external* if it does not intersect the segment joining the centers of the circles.

The following hyperbolic version of Casey's theorem is given in [7].

Theorem 1.15 (Abrosimov and Mikaiylova, 2015, [7]). *Let four circles* $O_1, O_2,$ O_3, O_4 *on the hyperbolic plane* $\mathbb{H}^2$ *be internally tangent to a circle* O *in the prescribed order. Denote by* t_{ij} *the length of the common external tangent of the circles* O_i, O_j*. Then*

$$\sinh\left(\frac{1}{2}t_{12}\right)\sinh\left(\frac{1}{2}t_{34}\right)+\sinh\left(\frac{1}{2}t_{23}\right)\sinh\left(\frac{1}{2}t_{14}\right)=\sinh\left(\frac{1}{2}t_{13}2\right)\sinh\left(\frac{1}{2}t_{24}\right).$$

Proof. Denote by O_i the centers of the circles respectively (Figure 4). Let R_i be the radii of the circles O_i and K_i the points of tangency with the circle O.

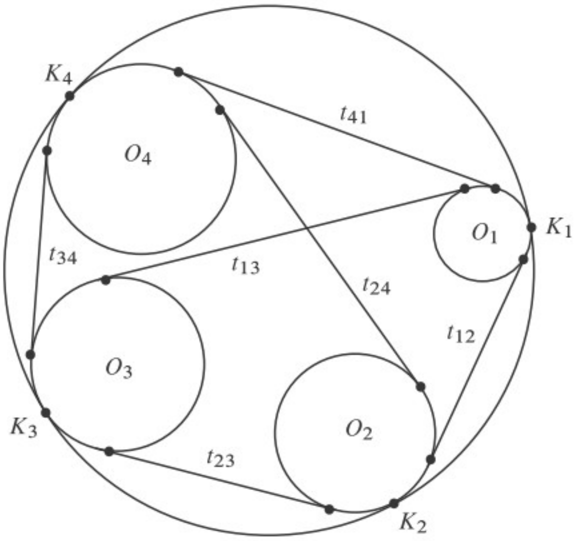

Figure 4

Consider the quadrilateral with sides O_iO_j, R_i, R_j, and t_{ij} (Figure 5). Denote by d the diagonal which divides the quadrilateral into triangles 1 and 2. Since the radii R_i and R_j are orthogonal to the common tangent t_{ij}, we can apply the hyperbolic Pythagorean theorem for triangle 1 (see, e.g., [8])

$$\cosh(d) = \cosh(R_j)\cosh(t_{ij}). \tag{19}$$

Denote the angle between the diagonal d and the tangent t_{ij} by δ. Then by the Sine rule for the hyperbolic triangle 1 we have

$$\sin\delta = \frac{\sinh(R_j)}{\sinh(d)}.$$

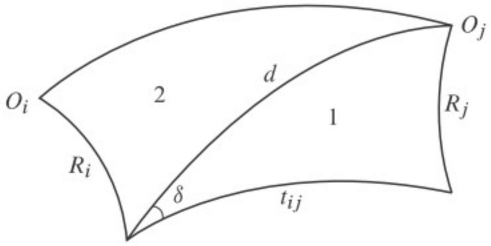

Figure 5

We use the second Cosine rule for the hyperbolic triangle 2 (see, e.g., [8])

$$\cosh(O_iO_j) = \cosh(R_i)\cosh(d) - \sinh(R_i)\sinh(R_j)\cos\left(\frac{\pi}{2} - \delta\right). \tag{20}$$

Substituting $\cosh(d)$ from (20) into (19) we obtain

$$\cosh(O_iO_j) = \cosh(t_{ij})\cosh(R_i)\cosh(R_j) - \sinh(R_i)\sinh(R_j). \tag{21}$$

We apply the first Cosine rule to the triangle OK_iK_j where $OK_i = OK_j = R$ (see Figure 6)

$$\cos(\angle K_iOK_j) = \frac{\cosh^2(R) - \cosh(K_iK_j)}{\sinh^2(R)}. \tag{22}$$

Using the first Cosine rule for the hyperbolic triangle OO_iO_j, where $OO_i = R - R_i$, $OO_j = R - R_j$, we have

$$\begin{aligned}\cosh(O_iO_j) = {} & \cosh(R - R_i)\cosh(R - R_j)\\ & - \sinh(R - R_i)\sinh(R - R_j)\cos(\angle K_iOK_j).\end{aligned}$$

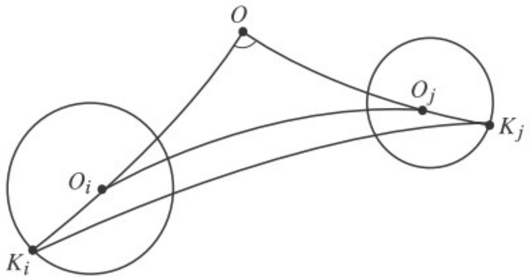

Figure 6

Combining the latter expression with (21) and (22) we obtain

$$\sinh^2\left(\frac{1}{2}K_iK_j\right) = \frac{\sinh^2(R)\cosh(R_i)\cosh(R_j)}{2\sinh(R-R_i)\sinh(R-R_j)}\sinh^2\left(\frac{1}{2}t_{ij}\right). \tag{23}$$

The points of tangency of the circles O_1, O_2, O_3, O_4 with the circle O form a cyclic quadrilateral $K_1K_2K_3K_4$ in $\mathbb{H}^2$. According to the hyperbolic Ptolemy theorem (Theorem 1.3), the segments K_iK_j satisfy the equation

$$\begin{aligned}&\sinh\left(\frac{1}{2}K_1K_3\right)\sinh\left(\frac{1}{2}K_2K_4\right)\\ &\quad= \sinh\left(\frac{1}{2}K_1K_2\right)\sinh\left(\frac{1}{2}K_3K_4\right) + \sinh\left(\frac{1}{2}K_2K_3\right)\sinh\left(\frac{1}{2}K_1K_4\right).\end{aligned} \tag{24}$$

We take the square root of (23) and substitute the expressions for $\sinh\left(\frac{1}{2}K_iK_j\right)$ in (24). Then we divide the resulting equation by

$$\sinh^2 R\sqrt{\frac{\cosh(R_1)\cosh(R_2)\cosh(R_3)\cosh(R_4)}{2\sinh(R-R_1)\sinh(R-R_2)\sinh(R-R_3)\sinh(R-R_4)}}$$

to obtain

$$\sinh\left(\frac{1}{2}t_{13}\right)\sinh\left(\frac{1}{2}t_{24}\right) = \sinh\left(\frac{1}{2}t_{12}\right)\sinh\left(\frac{1}{2}t_{34}\right) + \sinh\left(\frac{1}{2}t_{23}\right)\sinh\left(\frac{1}{2}t_{14}\right).$$

□

The spherical analog of Casey's theorem holds as well [7].

Theorem 1.16 (Abrosimov and Mikaiylova, 2015, [7]). *Let four circles O_1, O_2, O_3, O_4 on the sphere $\mathbb{S}^2$ be internally tangent to a circle O in the prescribed order. Denote by t_{ij} the length of the common external tangent of the circles O_i, O_j. Then*

$$\sin\left(\frac{1}{2}t_{12}\right)\sin\left(\frac{1}{2}t_{34}\right) + \sin\left(\frac{1}{2}t_{23}\right)\sin\left(\frac{1}{2}t_{14}\right) = \sin\left(\frac{1}{2}t_{13}\right)\sin\left(\frac{1}{2}t_{24}\right).$$

One can prove the spherical version of the theorem using the same arguments as in the hyperbolic case. The only difference is to replace hyperbolic trigonometric relations by corresponding spherical ones.

2 Geometry of polyhedra

Calculating volumes of polyhedra is a classical problem, that has been well known since Euclid. It remains relevant nowadays, partly due to the fact that the volume of a fundamental polyhedron is one of the main invariants for a three-dimensional manifold.

2.1 Volume of hyperbolic orthoscheme

Volume formulas for non-Euclidean tetrahedra in some special cases have been known since Lobachevsky, Bolyai and Schläfli. For example, Schläfli [53] found the volume of an orthoscheme in the 3-dimensional sphere $\mathbb{S}^3$. Recall that an *orthoscheme* is an n-dimensional simplex defined by a sequence of edges $(v_0, v_1), (v_1, v_2), \dots, (v_{n-1}, v_n)$ that are mutually orthogonal. In three dimensions, an orthoscheme is also called a *birectangular tetrahedron*. By definition, a 3-dimensional orthoscheme has three dihedral angles $\frac{\pi}{2}$ and three others, which we will refer as *essential* ones (see Figure 7).

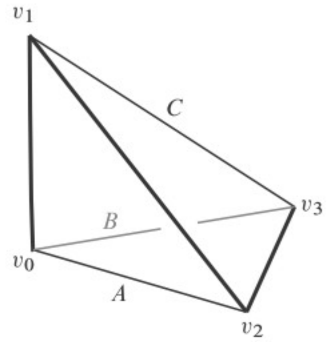

Figure 7. An orthoscheme $T = T(A, B, C)$ with essential dihedral angles A, B, C, all other angles are $\frac{\pi}{2}$

Theorem 2.1 (Schläfli, 1858, [52]). *Let T be a spherical orthoscheme with essential dihedral angles A, B, C. Then its volume $V = V(T)$ is given by the formula*

$$V = \frac{1}{4} S(A, B, C),$$

where

$$\begin{aligned} S\Big(\frac{\pi}{2} - x, y, \frac{\pi}{2} - z\Big) &\\ = \widehat{S}(x, y, z) &\\ = \sum_{m=1}^{\infty} \Big(\frac{D - \sin(x)\sin(z)}{D + \sin(x)\sin(z)}\Big)^m \frac{\cos(2mx) - \cos(2my) + \cos(2mz) - 1}{m^2} &\\ - x^2 + y^2 - z^2, & \end{aligned}$$

and $D \equiv \sqrt{\cos^2(x)\cos^2(z) - \cos^2(y)}$.

The function $S(x, y, z)$ appearing in Theorem 2.1 is called the *Schläfli function*.

The volume of a hyperbolic orthoscheme was obtained independently by Nikolai Lobachevsky [35] and János Bolyai (see [57], [56], or [64]). The following theorem was suggested by G. S. M. Coxeter [20]. It represents a result of Lobachevsky in a very simple form.

Theorem 2.2 (Lobachevsky, 1835, [35]; Coxeter, 1835, [20]). *Let T be a hyperbolic orthoscheme with dihedral angles A, B, C. Then its volume $V = V(T)$ is given by the formula $V = \frac{i}{4} S(A, B, C)$, where $S(A, B, C)$ is the Schläfli function.*

J. Bolyai found the volume of a hyperbolic orthoscheme in terms of planar angles and an edge length. This result was not published during his lifetime. It can be found in the work [57], pp. 111–118 collected on the base of original manuscripts in reliquias by János Bolyai (see also [56] and [64]).

Theorem 2.3 (Bolyai, 1832). *Consider a hyperbolic orthoscheme T with essential dihedral angles along AC, CB, BD, and all other dihedral angles equal to $\frac{\pi}{2}$ (Figure 8). T is given by its planar angles α, β, γ and edge length z. Then its volume $V = V(T)$ can be found by the formula*

$$V = \frac{\tan(\gamma)}{2\tan(\beta)} \int_0^z \frac{u \sinh(u) du}{\Big(\frac{\cosh^2(u)}{\cos^2(\alpha)} - 1\Big)\sqrt{\frac{\cosh^2(u)}{\cos^2(\gamma)} - 1}}.$$

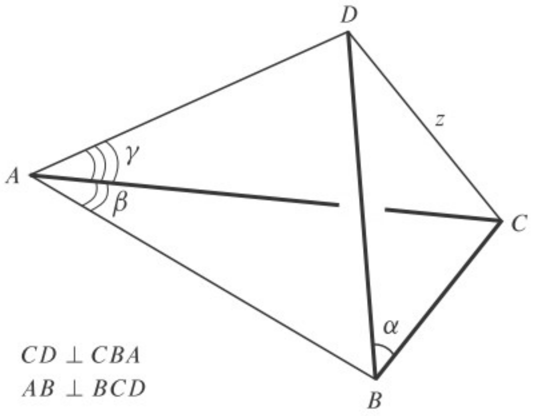
D
z
A
γ
β
C
α
B
$CD \perp CBA$
$AB \perp BCD$

Figure 8. An orthoscheme T given by planar angles α, β, γ and edge length z

2.2 Volume of ideal tetrahedra and octahedra

An *ideal tetrahedron* is a hyperbolic tetrahedron with all vertices at infinity. Opposite dihedral angles of an ideal tetrahedron are pairwise equal and the sum of dihedral angles at the edges adjacent to one vertex is $A + B + C = \pi$ (see Figure 9).

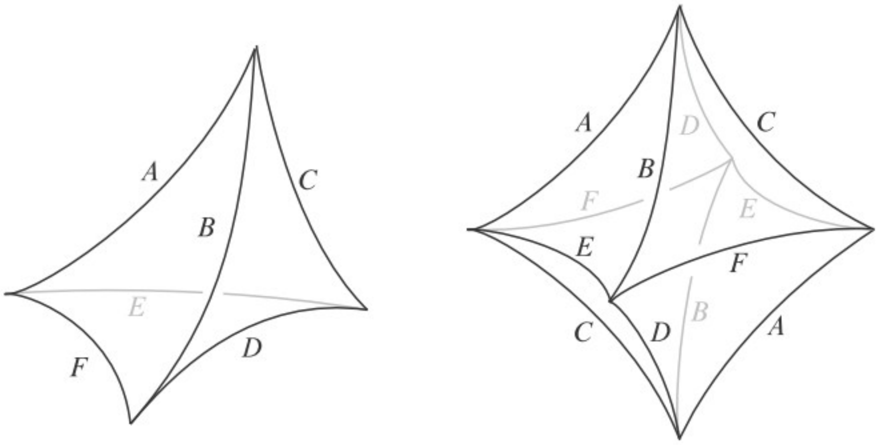
A
C
B
E
F
D
A
D
C
B
F
E
E
F
C
D
B
A

Figure 9. An ideal tetrahedron $T = T(A, B, C)$ and an ideal octahedron $O = O(A, B, C, D, E, F)$

The volume of an ideal tetrahedron has been known since Lobachevsky [35]. J. Milnor presented it in very elegant form [39].

Theorem 2.4 (Lobachevsky, 1835, [35]; Milnor, 1982, [39]). *Let T be an ideal tetrahedron with dihedral angles A, B and C. Then its volume $V = V(T)$ is given by the formula*

$$V = \Lambda(A) + \Lambda(B) + \Lambda(C),$$

where

$$\Lambda(x) = -\int_0^x \log|2\sin t|dt$$

is the Lobachevsky function.

Corollary 2.5. *The maximal volume of an ideal tetrahedron is attained in the case when the dihedral angles $A = B = C = \frac{1}{3}\pi$. It is equal to*

$$V_{\max} = 3\Lambda\Big(\frac{\pi}{3}\Big) = 1.01494\ldots.$$

A more general case of a tetrahedron with at least one vertex at infinity was investigated by E. B. Vinberg [8].

An *ideal octahedron* is a hyperbolic octahedron with all vertices at infinity. Consider a symmetric ideal octahedron whose opposite dihedral angles are pairwise equal (see Figure 9). By definition, the sum of the dihedral angles at the edges adjacent to any vertex is equal to 2π. We have

$$A + B + C + D = 2\pi, \quad A + C + E + F = 2\pi, \quad B + D + E + F = 2\pi,$$

or

$$C = \pi - A, \quad D = \pi - B, \quad F = \pi - E.$$

Hence, $A, B, C, D, E, F \in (0, \pi)$. The volume formula for such an octahedron is very similar to Milnor's formula for ideal tetrahedra and is given in [9].

Theorem 2.6 (Baigonakova and Mednykh, 2010, [9]). *Let O be a symmetric ideal octahedron with dihedral angles A, B, C, D, E, F. Then its volume $V = V(O)$ is given by the formula*

$$\begin{aligned}\frac{\mathrm{Vol}(O)}{2} &= \Lambda\Big(\frac{1}{2}(\pi + A + B + E)\Big) + \Lambda\Big(\frac{1}{2}(\pi - A - B + E)\Big)\\ &\quad + \Lambda\Big(\frac{1}{2}(\pi + A - B - E)\Big) + \Lambda\Big(\frac{1}{2}(\pi - A + B - E)\Big),\end{aligned}$$

where $\Lambda(x)$ is the Lobachevsky function.

Corollary 2.7. *The maximal volume of an ideal octahedron is attained in the case when $A = B = C = D = E = F = \frac{\pi}{2}$. It is equal to*

$$V_{\max} = 8\Lambda\Big(\frac{\pi}{4}\Big) = 3.66386\ldots.$$

2.3 Seidel's conjecture

In 1986, J. J. Seidel [54] conjectured that the volume of an ideal hyperbolic tetrahedron can be expressed as a function of the determinant and the permanent of its Gram matrix. Recall that the formula expressing the volume of such a tetrahedron in terms of dihedral angles has been known since Lobachevsky and Bolyai (Theorem 2.4). In spite of this, Seidel's problem had not been solved for a long time. Ten years later, a strengthened version of Seidel's conjecture was suggested by Igor Rivin and Feng Luo. They supposed that the volume of a non-Euclidean tetrahedron (hyperbolic or spherical one) depends only on the determinant of its Gram matrix. It was shown in [1] and [2] that the strengthened conjecture is false, while Seidel's conjecture is true within certain conditions.

Consider a non-Euclidean tetrahedron T in the sphere $\mathbb{S}^3$ or in the hyperbolic space $\mathbb{H}^3$. Denote its vertices by numbers $1, 2, 3, 4$. Let A_{ij} denote the dihedral angle at the edge joining vertices i and j. For convenience, we set $A_{ii} = \pi$ for $i = 1, 2, 3, 4$. It is well known [8] that, in the hyperbolic and spherical spaces, the tetrahedron T is uniquely (up to isometry) determined by its Gram matrix $G = \langle -\cos(A_{ij})\rangle_{i,j=1,\ldots,4}$.

Recall that the permanent of a matrix $M = \langle m_{ij}\rangle_{ij=1,2,\ldots,n}$ is defined by

$$\operatorname{per}(M) = \sum_{i=1}^{n} m_{ij}\operatorname{per}(M_{ij}), \quad \operatorname{per}(m_{ij}) = m_{ij},$$

where M_{ij} is the matrix obtained from M by removing the i-th row and j-th column. Conditions for the existence of spherical and hyperbolic tetrahedra in terms of Gram matrices are given in [36] and [59], respectively.

The solution of Seidel's problem, which was posed in [54], is given by the following theorem.

Theorem 2.8 (Abrosimov, 2010, [2]). *In each of the classes of acute or obtuse*[1] *ideal tetrahedra the volume is determined uniquely by the determinant and permanent of the Gram matrix.*

Proof. As is known (see, e.g., [39]), opposite dihedral angles of an ideal tetrahedron are pairwise equal and the sum of the dihedral angles at the edges adjacent to one vertex is $A + B + C = \pi$ (Figure 9).

---

[1] By an *obtuse tetrahedron* we mean a tetrahedron with at least one dihedral angle $> \frac{\pi}{2}$.

Thus, we can take $C = \pi - A - B$. The Gram matrix has the form

$$G = \begin{pmatrix} 1 & -\cos(A) & -\cos(B) & \cos(A+B) \\ -\cos(A) & 1 & \cos(A+B) & -\cos(B) \\ -\cos(B) & \cos(A+B) & 1 & -\cos(A) \\ \cos(A+B) & -\cos(B) & -\cos(A) & 1 \end{pmatrix}.$$

We have

$$\det(G) = -4\sin^2(A)\sin^2(B)\sin^2(A+B),$$
$$\operatorname{per}(G) = 4 + 4\cos^2(A)\cos^2(B)\cos^2(A+B).$$

To prove Theorem 2.8, we show that the dihedral angles of an ideal tetrahedron are uniquely (up to a permutation) determined by $\det G$ and $\operatorname{per} G$ in each of the cases: acute angled tetrahedron and obtuse angled tetrahedron.

Without loss of generality, we can assume that $0 < A \le B \le C = \pi - A - B$. Then the dihedral angles A, B are a priori acute, and the angle C is either acute or obtuse. In the former case, the tetrahedron under consideration is acute-angled, and in the latter case, it is obtuse-angled.

Let us introduce the new variables

$$x = \sin(A)\sin(B), \quad y = \cos(A)\cos(B)$$

and show that, for a fixed left-hand side, the solutions of the system of equations

$$-\frac{1}{4}\det(G) = x^2(1 - (y-x)^2),$$
$$\frac{1}{4}\operatorname{per}(G) = y^2(y-x)^2 + 1$$

correspond to one tetrahedron (up to isometry) in each of the two cases mentioned above.

Suppose that the system has a pair of different solutions (a, b) and (x, y). Then we have

$$a^2(1 - (b-a)^2) = x^2(1 - (y-x)^2),$$
$$b^2(b-a)^2 = y^2(y-x)^2.$$

For the angle C to be acute means that

$$\cos(A)\cos(B) - \sin(A)\sin(B) = -\cos(C) < 0,$$

i.e., both solutions satisfy the inequalities $b(b-a) < 0$ and $x(x-y) < 0$. If the angle C is obtuse, then the reverse inequality holds. This allows us to take a square root in the second equation without loosing solutions:

$$a^2(1-(b-a)^2) = x^2(1-(y-x)^2),$$

$$b\,(b-a) = y\,(y-x).$$

Expressing x from the second equation and substituting the resulting expression into the first one, we obtain a sixth-degree polynomial equation in y. Fortunately, it decomposes into the linear factors and the biquadratic polynomial

$$(b-y)\cdot(b+y)\cdot\big(y^4-(a^2+a^4+2\,ab-2a^3b-b^2+2ab^3-b^4)y^2 \\ +a^4b^2-4a^3b^3+6a^2b^4-4ab^5+b^6\big)=0.$$

Thus, all solutions can be found in radicals. Substituting the expressions for x and y in terms of dihedral angles, we see that different solutions of the system correspond to the same ideal tetrahedron $T(A,B,C)$ up to reordering the dihedral angles. □

Note that, in Theorem 2.8, the assumption that the tetrahedron is either acute-angled or obtuse-angled cannot be dispensed with. This is demonstrated by the following example.

Example 2.9. Consider a pair of ideal tetrahedra, the obtuse-angled tetrahedron $T_1(s,s,\pi-2s)$ and the acute-angled tetrahedron $T_2\big(t,\frac{1}{2}(\pi-t),\frac{1}{2}(\pi-t)\big)$, where

$$s = \arccos\left(\frac{\sqrt{2+\sqrt{4+\sqrt{170\sqrt{17}-698}}}}{2\sqrt{2}}\right),$$

$$t = \arccos\left(\frac{-1+\sqrt{17}+\sqrt{-26+10\sqrt{17}}}{8}\right).$$

The determinants and the permanents of the Gram matrices of these tetrahedra coincide. We have

$$\det(G(T_1)) = \det(G(T_2)) = \frac{107-51\sqrt{17}}{128},$$

$$\operatorname{per}(G(T_1)) = \operatorname{per}(G(T_2)) = \frac{163+85\sqrt{17}}{128}.$$

At the same time, the volumes of the tetrahedra T_1 and T_2 are different and equal to 0.847365 and 1.01483 respectively.

2.4 Volume of a general hyperbolic tetrahedron

Despite the fact that partial results on the volume of non-Euclidean tetrahedra were known, a volume formula for hyperbolic tetrahedra of general type remained unknown until recently. A general algorithm for obtaining such a formula was indicated by W.-Y. Hsiang in [30]. A complete solution was obtained more than ten years later. A general formula was proposed in paper by Y. Cho and H. Kim [17]. However, it was asymmetric with respect to permutation of angles. J. Murakami and Y. Yano [44] proposed a formula expressing the volume by dihedral angles in a symmetric way. A year later, A. Ushijima [59] presented a simple proof of the Murakami–Yano formula. He also investigated the case of a truncated hyperbolic tetrahedron.

It should be noted that in all these studies the volume is expressed as a linear combination of 16 dilogarithms or Lobachevsky functions. The arguments of these functions depend on the dihedral angles of the tetrahedron and some additional parameter, which is the root of some quadratic equation with complex coefficients in a sophisticated form.

The geometric meaning of that formula was explained by P. Doyle and G. Leibon [24] in terms of Regge symmetry. A clear description of these ideas and a complete geometric proof of the formula was given by Y. Mohanty [42]. In particular, she proved the equivalence of Regge symmetry and homogeneity (scissors congruence) in hyperbolic space [43].

In 2005, D. A. Derevnin and A. D. Mednykh [22] proposed the following integral formula for the volume of a hyperbolic tetrahedron.

Theorem 2.10 (Derevnin and Mednykh, 2005, [22]). *Let $T(A,B,C,D,E,F)$ be a compact hyperbolic tetrahedron with dihedral angles A,B,C,D,E,F. Then its volume $V = V(T)$ is given by the formula*

$$V = -\frac{1}{4}\int\limits_{z_1}^{z_2} \log \frac{\cos\left(\frac{1}{2}(A+B+C+z)\right)\cos\left(\frac{1}{2}(A+E+F+z)\right)}{\sin\left(\frac{1}{2}(A+B+D+E+z)\right)\sin\left(\frac{1}{2}(A+C+D+F+z)\right)}$$
$$\frac{\cos\left(\frac{1}{2}(B+D+F+z)\right)\cos\left(\frac{1}{2}(C+D+E+z)\right)}{\sin\left(\frac{1}{2}(B+C+E+F+z)\right)\sin\left(\frac{1}{2}z\right)}\,dz,$$

where z_1 and z_2 are the roots of the integrand satisfied the condition $0 < z_2 - z_1 \leq \pi$.

More precisely,

$$z_1 = \arctan\left(\frac{k_3}{k_4}\right) - \arctan\left(\frac{k_1}{k_2}\right), \quad z_2 = \pi - \arctan\left(\frac{k_3}{k_4}\right) - \arctan\left(\frac{k_1}{k_2}\right),$$

where

$$\begin{aligned} k_1 &= -\cos(S) - \cos(A+D) - \cos(B+E) - \cos(C+F) - \cos(D+E+F) \\ &\quad - \cos(D+B+C) - \cos(A+E+C) - \cos(A+B+F), \\ k_2 &= \sin(S) + \sin(A+D) + \sin(B+E) + \sin(C+F) + \sin(D+E+F) \\ &\quad + \sin(D+B+C) + \sin(A+E+C) + \sin(A+B+F), \\ k_3 &= 2\,(\sin A \sin D + \sin B \sin E + \sin C \sin F), \\ k_4 &= \sqrt{k_1^2 + k_2^2 - k_3^2}, \\ S &= A + B + C + D + E + F. \end{aligned}$$

2.5 Sforza's formula for non-Euclidean tetrahedra

Surprisingly, but more than 100 years ago, in 1907, Giuseppe Sforza found a fairly simple formula for the volume of a non-Euclidean tetrahedron. This fact has became widely known after a discussion between the second named author and J. M. Montesinos at the conference in El Burgo de Osma (Spain), August, 2006. Unfortunately, the outstanding work of Sforza [55] published in Italian has been completely forgotten.

The original arguments by Sforza are based on some identity given by H. W. Richmond [49] besides the Schläfli formula. Then he used the Pascal equation [45] for minors of the Gram matrix and some routine calculations. In this subsection we present another simpler proof of Sforza's formula. The idea of the proof belongs to the authors [6].

Consider a hyperbolic (or a spherical) tetrahedron T with dihedral angles A, B, C, D, E, F. Assume that A, B, C are dihedral angles at the edges adjacent to one vertex and D, E, F are opposite to them correspondingly (see Figure 10). Then the Gram matrix $G(T)$ is defined as follows:

$$G = \begin{pmatrix} 1 & -\cos(A) & -\cos(B) & -\cos(F) \\ -\cos(A) & 1 & -\cos(C) & -\cos(E) \\ -\cos(B) & -\cos(C) & 1 & -\cos(D) \\ -\cos(F) & -\cos(E) & -\cos(D) & 1 \end{pmatrix}.$$

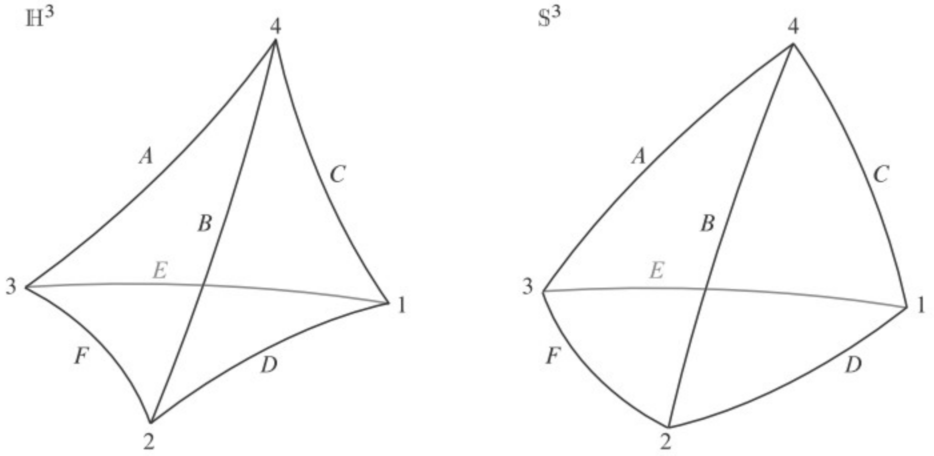

Figure 10. A tetrahedron $T = T(A, B, C, D, E, F)$ in $\mathbb{H}^3$ or $\mathbb{S}^3$

Denote by $\mathcal{C} = \langle c_{ij} \rangle_{i,j=1,2,3,4}$ a matrix of cofactors $c_{ij} = (-1)^{i+j} G_{ij}$, where G_{ij} is (i, j)-th minor of G.

In the following proposition we collect some known results about hyperbolic tetrahedra (see, e.g., [59]).

Proposition 2.11. *Let T be a compact hyperbolic tetrahedron. Then*

(i) $\det(G) < 0$;

(ii) $c_{ii} > 0$, $i \in \{1, 2, 3, 4\}$;

(iii) $c_{ij} > 0$, $i \neq j$; $i, j \in \{1, 2, 3, 4\}$;

(iv) $\cosh(\ell_{ij}) = \frac{c_{ij}}{\sqrt{c_{ii} c_{jj}}}$,

where ℓ_{ij} is the hyperbolic length of the edge joining vertices i and j.

Further, we need the following assertion due to Jacobi ([48], Theorem 2.5.1, p. 12).

Proposition 2.12 (Jacobi theorem). *Let $M = \langle m_{ij} \rangle_{i,j=1,\dots,n}$ be a matrix and $\Delta = \det(M)$ be the determinant of M. Denote by $\mathcal{C} = \langle c_{ij} \rangle_{i,j=1,\dots,n}$ the matrix of cofactors $c_{ij} = (-1)^{i+j} \det M_{ij}$, where M_{ij} is the $(n-1) \times (n-1)$ minor obtained by removing the i-th row and j-th column of the matrix M. Then for any k, $1 \leq k \leq n-1$ we have*

$$\det\langle c_{ij} \rangle_{i,j=1,\dots,k} = \Delta^{k-1} \det\langle m_{ij} \rangle_{i,j=k+1,\dots,n}.$$

One of the main tools for volume calculations in $\mathbb{H}^3$ and $\mathbb{S}^3$ is the classical Schläfli formula.

Proposition 2.13 (Schläfli's formula). *Let X^n be a space of constant curvature K. Consider a family of convex polyhedra P in X^n depending on one or more parameters in a differential manner and keeping the same combinatorial type. Then the differential of the volume $V = V(P)$ satisfies the equation*

$$(n-1)K\,dV = \sum_F V_{n-2}(F)\,d\theta(F),$$

where the sum is taken over all $(n-2)$-facets of P, $V_{n-2}(F)$ is $(n-2)$-dimensional volume of F, and $\theta(F)$ is the interior angle along F.

In the classical paper by Schläfli [52] this formula was proved for the case of a spherical n-simplex. For the hyperbolic case, it was obtained by H. Kneser [33] (for more details, see also [8] and [40]).

Now we are able to prove the following theorem (see also [6]).

Theorem 2.14 (Sforza's formula in $\mathbb{H}^3$). *Let T be a compact hyperbolic tetrahedron with Gram matrix G. Consider $G = G(A)$ as a function of the dihedral angle A. Then the volume $V = V(T)$ is given by the formula*

$$V = \frac{1}{4}\int_{A_0}^{A} \log\Big(\frac{c_{34}(A) - \sqrt{-\det G(A)}\sin(A)}{c_{34}(A) + \sqrt{-\det G(A)}\sin(A)}\Big)\,dA,$$

where A_0 is a suitable root of the equation $\det G(A) = 0$ *and $c_{34} = c_{34}(A)$ is the $(3,4)$-cofactor of the matrix $G(A)$.*

Proof. Let $\Delta = \det(G)$. By the Jacobi theorem (Proposition 2.12) applied to the Gram matrix G for $n = 4$ and $k = 2$ we obtain

$$c_{33}c_{44} - c_{34}^2 = \Delta(1 - \cos^2(A)).$$

By the Cosine rule (Proposition 2.11, (iv)) we get $\cosh(\ell_A) = \frac{c_{34}}{\sqrt{c_{33}c_{44}}}$. Hence,

$$\sinh(\ell_A) = \sqrt{\cosh^2(\ell_A) - 1} = \sqrt{\frac{c_{34}^2 - c_{33}c_{44}}{c_{33}c_{44}}} = \frac{\sqrt{-\Delta}\sin(A)}{\sqrt{c_{33}c_{44}}}.$$

Since $\exp(\pm\ell_A) = \cosh(\ell_A) \pm \sinh(\ell_A)$ we have

$$\exp(\pm\ell_A) = \frac{c_{34} \pm \sqrt{-\Delta}\,\sin(A)}{\sqrt{c_{33}c_{44}}}.$$

Hence,

$$\exp(2\,\ell_A) = \frac{\exp(\ell_A)}{\exp(-\ell_A)} = \frac{c_{34} + \sqrt{-\Delta}\,\sin(A)}{c_{34} - \sqrt{-\Delta}\,\sin(A)}$$

and

$$\ell_A = \frac{1}{2}\,\log\Big(\frac{c_{34} + \sqrt{-\Delta}\,\sin(A)}{c_{34} - \sqrt{-\Delta}\,\sin(A)}\Big).$$

By the Schläfli formula (Proposition 2.13) for $V = V(T)$ we have

$$-dV = \frac{1}{2}\sum_{\alpha} \ell_\alpha\, d\alpha, \quad \alpha \in \{A, B, C, D, E, F\}.$$

By the assumption that the angle A is variable and all the other angles are constant, we get

$$-dV = \frac{1}{2}\,\ell_A\, dA.$$

Note that $\det(G) \to 0$ as $A \to A_0$. Thus, $V \to 0$ as $A \to A_0$. Then integrating both sides of the equation we obtain

$$V = \int\limits_{A_0}^{A} \Big(-\frac{\ell_A}{2}\Big)\,dA = \frac{1}{4}\int\limits_{A_0}^{A} \log\Big(\frac{c_{34} - \sqrt{-\det G}\,\sin(A)}{c_{34} + \sqrt{-\det G}\,\sin(A)}\Big)\,dA,$$

where the lower limit A_0 is a suitable root of the equation $\det G(A) = 0$. □

In the following proposition we collect some known results about spherical tetrahedra (see, e.g., [36]).

Proposition 2.15. *Let T be a spherical tetrahedron. Then*

(i) $\det(G) > 0$;

(ii) $c_{ii} > 0$, $i = 1, \ldots, 4$;

(iii) $\cos(\ell_{ij}) = \frac{c_{ij}}{\sqrt{c_{ii}c_{jj}}}$,

where ℓ_{ij} is the length of the edge joining vertices i and j in $\mathbb{S}^3$.

The next theorem presents a spherical version of Sforza's formula.

Theorem 2.16 (Sforza's formula in $\mathbb{S}^3$). *Let T be a spherical tetrahedron with Gram matrix G. Consider $G = G(A)$ as a function of the dihedral angle A. Then the volume $V = V(T)$ is given by the formula*

$$V = \frac{1}{4i} \int\limits_{A_0}^{A} \log\Big(\frac{c_{34}(A) + i\sqrt{\det\, G(A)}\sin(A)}{c_{34}(A) - i\sqrt{\det\, G(A)}\sin(A)}\Big)\, dA,$$

where A_0 is a suitable root of the equation $\det(G(A)) = 0$ *and $c_{34} = c_{34}(A)$ is the $(3,4)$-cofactor of the matrix $G(A)$.*

The proof is similar to one given in the hyperbolic case.

2.6 Volumes of hyperbolic octahedra with symmetries

In this subsection, we consider octahedra in $\mathbb{H}^3$ and $\mathbb{S}^3$. It is known that a regular ideal octahedron is a fundamental polyhedron for the complement of the Whitehead link embedded into the three-dimensional sphere. Moreover, a hyperbolic manifold of minimum volume (the Matveev–Fomenko–Weeks manifold) and many other manifolds can be obtained by Dehn surgeries on the Whitehead link. Thus, hyperbolic octahedra can serve as fundamental polyhedra for a wide class of manifolds, including ones of small volume. The latter seems to be of particular interest when one deals with hyperbolic manifolds ordered by increasing volume.

The problem of finding a volume formula for an arbitrary non-Euclidean polyhedron is very hard. It was only solved for a general hyperbolic or spherical tetrahedron. Even for an octahedron the problem is still open. But in the case of a polyhedron with nontrivial symmetry, the calculation of the volume is simplified. This fact was first shown by Lobachevsky, who found the volume of an ideal hyperbolic tetrahedron. It follows from the definition of an ideal tetrahedron that its opposite dihedral angles are pairwise equal (i.e. it has a symmetry).

Exact formulas for the volumes of spherical octahedra that have mmm- or $2|m$-symmetry were derived in [3]. The volumes of hyperbolic octahedra with mmm- or $\bar{3}$-symmetry were calculated in [4] and [5] respectively. These studies were initiated by Galiulin, Mikhalev, and Sabitov's paper [27], in which the authors calculated the volumes of Euclidean octahedra with nontrivial symmetry. We show just two from the above mentioned results.

An octahedron is said to *have* $2|m$*-symmetry* if it admits a rotation through an angle of $\frac{2}{3}\pi$ and a mirror symmetry about a plane orthogonal to the axis of rotation (see Figure 11). The $2|m$-symmetry of the octahedron corresponds to the group C_{2h} in the Schoenflies classification, to the group $[2, 2^+]$ in the Coxeter classification, or to the group 2^* in the orbifold classification (see, e.g., [19], [32]). The volume formula for spherical octahedra with $2|m$-symmetry is given in [3].

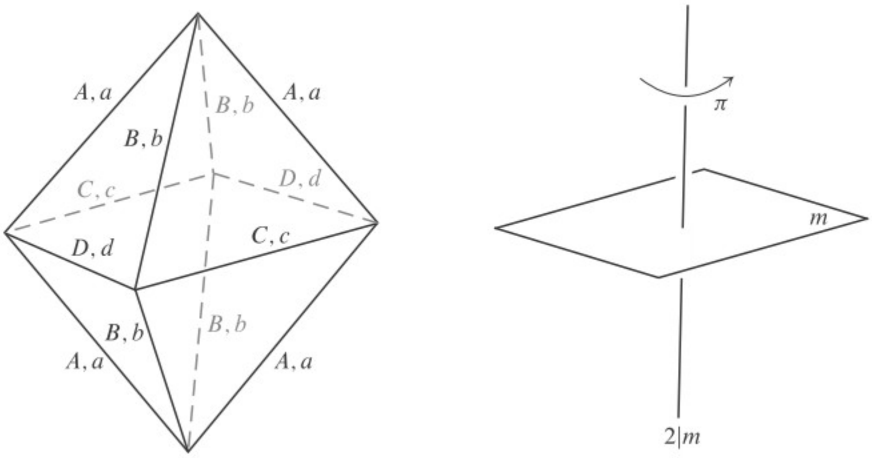

Figure 11. An octahedron $\mathcal{O} = \mathcal{O}(A, B, C, D)$ with $2|m$-symmetry

Theorem 2.17 (Abrosimov, Godoy-Molina, and Mednykh, 2008, [3]). *Let* $\mathcal{O} = \mathcal{O}(A, B, C, D)$ *be a spherical octahedron with dihedral angles* A, B, C, D *that admits a* $2|m$*-symmetry. Then its volume* $V = V(\mathcal{O})$ *is given by the formula*

$$V = 2\int\limits_{\frac{\pi}{2}}^{\theta} \big(\operatorname{arctanh}(X\cos(\tau)) + \operatorname{arctanh}(Y\cos(\tau)) + \operatorname{arctanh}(Z\cos(\tau)) + \operatorname{arctanh}(W\cos(\tau))\big)\frac{d\tau}{\cos(\tau)},$$

where $X = \cos(A)$, $Y = \cos(B)$, $Z = \cos\left(\frac{1}{2}(C + D)\right)$, $W = \cos\left(\frac{1}{2}(C - D)\right)$, *and* θ $\left(0 \le \theta \le \frac{\pi}{2}\right)$ *is a root of the equation*

$$\cos^2(\theta) + \frac{X + Y + Z + W}{XYZ + XYW + XZW + YZW} = 0.$$

Moreover, θ *is determined by the sine–tangent rule*

$$\frac{\sin(A)}{\tan(a)} = \frac{\sin(B)}{\tan(b)} = \frac{\sin\left(\frac{1}{2}(C + D)\right)}{\tan\left(\frac{1}{2}(c + d)\right)} = \frac{\sin\left(\frac{1}{2}(C - D)\right)}{\tan\left(\frac{1}{2}(c - d)\right)} = \tan(\theta).$$

An octahedron is said to *have* $\bar{3}$*-symmetry* if it admits the central symmetry and a rotation through an angle of $\frac{2}{3}\pi$ (see Figure 12). The $\bar{3}$-symmetry of the octahedron corresponds to the group S_6 in the Schoenflies classification, to the group $[2^+, 6^+]$ in the Coxeter classification, or to the group $3\times$ in the orbifold classification (see, e.g., [19] or [32], Chapter 11). The existence criterion and the volume formula for a hyperbolic octahedron with this type of symmetry are given in [5].

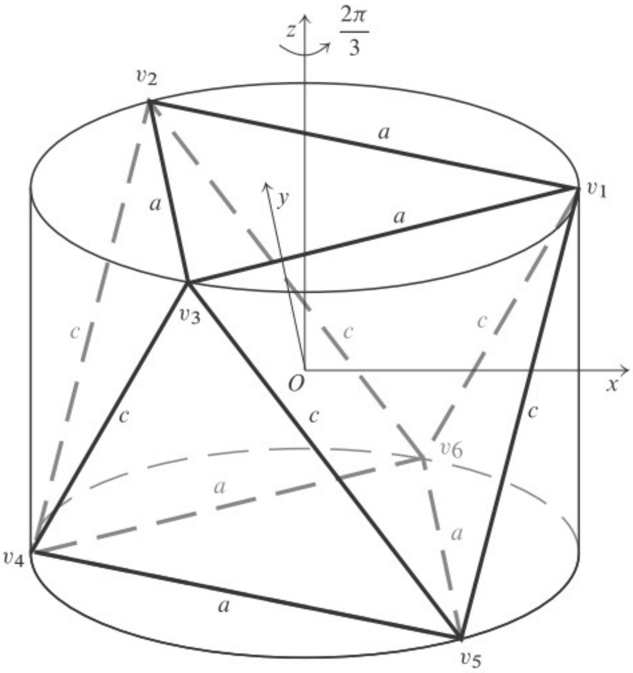

Figure 12. An octahedron $\mathcal{O} = \mathcal{O}(a, c)$ with $\bar{3}$-symmetry

Theorem 2.18 (Abrosimov, Kudina, and Mednykh, 2015, [5]). *Let* $a, c > 0$ *be some numbers. Then the following statements are equivalent:*

(i) *there exists a compact hyperbolic octahedron* $\mathcal{O}(a, c)$ *with edge lengths* a, c *that admits a* $\bar{3}$*-symmetry*;

(ii) *the inequality*

$$\frac{\sinh^2\left(\frac{1}{2}a\right)}{\sinh^2\left(\frac{1}{2}c\right)} < 3$$

holds.

Theorem 2.19 (Abrosimov, Kudina, and Mednykh, 2015, [5]). *Let* $\mathcal{O} = \mathcal{O}(a, c)$ *be a compact hyperbolic octahedron with edge lengths* a, c *that admits a* $\bar{3}$*-symmetry. Then its volume* $V = V(\mathcal{O})$ *is given by the formula*

$$V = 3 \int\limits_{0}^{a} \frac{t\,F + (1 + 2\cosh(t))c\,G}{(1 + 2\cosh(t))\Delta}\, dt,$$

where

$$\begin{aligned} F &= 1 + 2\cosh(t) + 2\cosh^2(t) + \cosh^3(t) \\ &\quad - 2\cosh(c) - 2\cosh(t)\cosh(c) - \cosh^2(c) \\ &\quad - \cosh(t)\cosh^2(c) + \cosh^2(t)\cosh(c) \\ &\quad - 4\cosh^2(t)\cosh^2(c) + 3\cosh^3(c), \\ G &= \sinh(t)\sinh(c)(-1 + 2\cosh(t)), \\ \Delta &= (\cosh(2c) - \cosh(t))\sqrt{(-2\cosh(t) + 3\cosh(c))(\cosh(t) + \cosh(c))}. \end{aligned}$$

The algorithm developed in [5] for obtaining the volume formula is rather universal. It can be also applied for other hyperbolic polyhedra with non-trivial symmetry.

The idea is the following. First, the polyhedron under consideration is modeled in Euclidean space. The coordinates of the vertices in $\mathbb{E}^3$ are calculated in some parameters according to the symmetry. Then the Euclidean polyhedron is placed into the projective Caley–Klein model of hyperbolic space. This gives rise to some additional conditions on the coordinates. Vector algebra in projective space is used to express hyperbolic edge lengths and dihedral angles in terms of coordinate parameters. Then symmetry is used again to exclude non-invariant parameters depending on the choice of coordinates. As a result we obtain a direct relation between lengths and angles of the hyperbolic polyhedron. These relations allow us to establish an existence criterion for such polyhedra in $\mathbb{H}^3$. Moreover, they allow us to find the solution of the Schläfli differential equation responsible for the volume. The latter provides an exact integral formula for the volume of the hyperbolic polyhedron.

Acknowledgement. The research of this article was funded by Russian Science Foundation (project no. 16-41-02006).

References

[1] N. V. Abrosimov, "Solving the Seidel problem on the volume of hyperbolic tetrahedron," *Sib. Èlektron. Mat. Izv.* 6 (2009), 211–218, in Russian.
MR 2586687 Zbl 1299.51007 ☞ 171

[2] N. V. Abrosimov, "Seidel's problem on the volume of a non-Euclidean tetrahedron," *Dokl. Akad. Nauk* 435 (2010), no. 1, 7–10, in Russian, English translation in *Dokl. Math.* 82 (2010), no. 3, 843–846.
MR 2790493 Zbl 1228.52007 ☞ 171

[3] N. V. Abrosimov, M. Godoy-Molina, and A. D. Mednykh, "On the volume of a spherical octahedron with symmetries," *Sovrem. Mat. Prilozh.* 60, Algebra (2008), 3–12, in Russian; English translation in *J. Math. Sci. (N.Y.)* 161 (2009), no. 1, 1–10.
MR 2676254 Zbl 1187.51021 ☞ 179, 180

[4] N. V. Abrosimov and G. A. Baigonakova, "Hyperbolic octahedron with mmm-symmetry," *Sib. Èlektron. Mat. Izv.* 10 (2013), 123–140, in Russian.
MR 3039999 Zbl 1330.52014 ☞ 179

[5] N. V. Abrosimov, E. S. Kudina, and A. D. Mednykh, "On the volume of a hyperbolic octahedron with $\bar{3} - symmetry$," *Tr. Mat. Inst. Steklova* 288 (2015), 7–15, in Russian; English translation in *Proc. Steklov Inst. Math.* 288 (2015), no. 1, 1–9.
MR 3485697 Zbl 1341.52022 ☞ 179, 181, 182

[6] N. V. Abrosimov and A. D. Mednykh, "Volumes of polytopes in constant curvature spaces," in *Rigidity and symmetry* (R. Connelly, A. I. Weiss and W. Whiteley, eds.), Fields Institute Communications, 70, Springer, New York, and Fields Institute for Research in Mathematical Sciences, Toronto, ON, 2014, 1–26.
MR 3329266 Zbl 1320.52013 ☞ 175, 177

[7] N. V. Abrosimov and L. A. Mikaiylova, "Casey's theorem in hyperbolic geometry," *Sib. Èlektron. Mat. Izv.* 12 (2015), 354–360.
MR 3493779 Zbl 06607128 ☞ 164, 166

[8] D. V. Alekseevskij, E. B. Vinberg, and A. S. Solodovnikov, "Geometry of spaces of constant curvature," in *Geometry* II (a translation of Геометрия II, Akad. Nauk SSSR, Vsesoyuz. Inst. Nauchn. i Tekhn. Inform., Moscow, 1988), translation by V. Minachin, translation edited by E. B. Vinberg, Encyclopaedia of Mathematical Sciences, 29, Springer-Verlag, Berlin, 1993, 1–138.
MR 1254932 ☞ 151, 152, 165, 170, 171, 177

[9] G. A. Baigonakova and A. D. Mednykh, “On Milnor formula for hyperbolic octahedron,” *Mat. Zamet. YAGU* 17 (2010), no. 2, 3–9.
Zbl 1274.51016 ☞ 170

[10] G. A. Baigonakova and A. D. Mednykh, “On Bretschnider formula for spherical quadrilateral,” *Mat. Zamet. YAGU* 19 (2012), no. 1, 3–11, in Russian.
Zbl 1274.51022 ☞ 163

[11] G. A. Baigonakova and A. D. Mednykh, “On Bretschnider formula for hyperbolic quadrilateral,” *Mat. Zamet. YAGU* 19 (2012), no. 1, 12–19, in Russian.
Zbl 1274.51023 ☞ 160, 163

[12] C. W. Baur, “Das Sehnenviereck in der Ebene und auf der Kugel als besonderer Fall des allgemeinen,” *Zeitschrift für Mathematik und Physik* 6 (1861), 221–234.
☞ 163

[13] A. F. Beardon, *The geometry of discrete groups,* Graduate Texts in Mathematics, 91. Springer-Verlag, New York, 1983.
MR 0698777 Zbl 0528.30001 ☞ 160, 161

[14] S. Bilinski, “Zur Begründung der elementaren Inhaltslehre in der hyperbolischen Ebene,” *Math. Ann.* 180 (1969), 256–268.
Zbl 0159.21702 ☞ 152

[15] C. A. Bretschneider, “Untersuchung der trigonometrischen Relationen des geradlinigen Viereckes,” *Archiv der Mathematik und Physik* 2, (1842), 225–261.
☞ 159

[16] J. Casey, *A sequel to the first six books of the Elements of Euclid,* containing an easy introduction to modern geometry, with numerous examples, 5th ed., Hodges, Figgis and Co., Dublin, 1888.
☞ 164

[17] Yu. Cho and H. Kim, “On the volume formula for hyperbolic tetrahedra,” *Discrete Comput. Geom.* 22 (1999), no. 3, 347–366.
MR 1706606 Zbl 0952.51013 ☞ 174

[18] R. Connelly, “Comments on generalized Heron polynomials and Robbins’ conjectures,” *Discrete Math.* 309 (2009), no. 12, 4192–4196.
MR 2537417 Zbl 1181.51021 ☞ 154

[19] J. H. Conway, H. Burgiel, and Ch. Goodman-Strass, *The symmetries of things,* A. K. Peters, Wellesley, MA, 2008.
MR 2410150 Zbl 1173.00001 ☞ 180, 181

[20] H. S. M. Coxeter, "The functions of Schläfli and Lobatschefsky," *Quart. J. Math. Oxford.* 6 (1935), no. 1, 13–29.
JFM 61.0395.02 Zbl 0011.17006 ☞ 168

[21] H. S. M. Coxeter and S. L. Greitzer, *Geometry revisited,* New Mathematical Library, 19. Random House, New York, 1967.
MR 3155265 Zbl 0166.16402 ☞ 154

[22] D. A. Derevnin and A. D. Mednykh, "A formula for the volume of a hyperbolic tetrahedron," *Uspekhi Mat. Nauk* 60 (2005), no. 2, 159–160, in Russian; English translation in *Russian Math. Surveys* 60 (2005), no. 2, 346–348.
MR 2152953 Zbl 1085.51503 ☞ 174

[23] W. Dickinson and M. Salmassi, "The right right triangle on the sphere," *College Math. J.* 39 (2008), no. 1, 24–33.
MR 2378203 Zbl 1293.97034 ☞ 154

[24] P. Doyle and G. Leibon, "23040 symmetries of hyperbolic tetrahedra," preprint, 2003.
arXiv:math/0309187v1 [math.GT]
☞ 174

[25] L. Euler, "Variae speculationes super area triangulorum sphaericorum," *Nova Acta Academiae Scientarum Imperialis Petropolitinae* 10 (1797), 47–62. Reprinted in *Opera omnia,* Series 1, Vol. 29, 253–266.
`http://eulerarchive.maa.org/pages/E698.html`
☞ 152

[26] M. Fedorchuk and I. Pak, "Rigidity and polynomial invariants of convex polytopes," *Duke Math. J.* 129 (2005), no. 2, 371–404.
MR 2165546 Zbl 1081.52012 ☞ 154

[27] R. V. Galiulin, S. N. Mikhalev, and I. Kh. Sabitov, "Some applications of the formula for the volume of an octahedron," *Mat. Zametki* 76 (2004), no. 1, 27–43, in Russian; English translation in *Math. Notes* 76 (2004), no. 1-2, 25–40.
MR 2099840 Zbl 1075.52006 ☞ 179

[28] R. Guo and N. Sönmez, "Cyclic polygons in classical geometry," *C. R. Acad. Bulgare Sci.* 64 (2011), no. 2, 185–194.
MR 2865551 Zbl 1289.51013 ☞ 154

[29] G. Hessenberg, "Elementare Beweise für eine Maximumeigenschaft des Sehnenvierecks," in *Schwarz-Festschrift,* Julius Springer, Berlin, 1914, 76–83.
☞ 163

[30] W.-Y. Hsiang, "On infinitesimal symmetrization and volume formula for spherical or hyperbolic tetrahedrons," *Quart. J. Math. Oxford Ser.* (2) 39 (1988), no. 156, 463–468.
MR 0975910 Zbl 0664.51011 ☞ 174

[31] V. F. Ivanoff, C. F. Pinzka, and J. Lipman, "Elementary problems and solutions: solutions: E1376," *Amer. Math. Monthly* 67 (1960), no. 3, 291–292.
MR 1530669 ☞ 156

[32] N. W. Johnson, *Geometries and transformations* Cambridge University Press, Cambridge, 2018.
Zbl 06706287 ☞ 180, 181

[33] H. Kneser, "Der Simplexinhalt in der nichteuklidischen Geometrie," *Deutsche Math.* 1, (1936), 337–340.
JFM 62.0653.03 Zbl 0014.36203 ☞ 177

[34] T. Kubota, "On the extended Ptolemy's theorem in hyperbolic geometry," *Science reports of the Tohoku University,* Ser. 1: Physics, chemistry, astronomy, 2 (1912), 131–156.
☞ 153

[35] N. I. Lobatschefskij (N. I. Lobachevsky), *Imaginäre Geometrie und ihre Anwendung auf einige Integrale,* German translation by H. Liebmann, Teubner, Leipzig, 1904. ☞ 168, 169, 170

[36] F. Luo, "On a problem of Fenchel," *Geom. Dedicata* 64 (1997), no. 3, 277–282.
MR 1440561 Zbl 0873.51015 ☞ 171, 178

[37] W. J. M'Clelland and T. Preston, *A treatise on spherical trigonometry, with application to spherical geometry and numerous examples,* Part II. Macmillian and Co., London, 1886.
☞ 153, 156

[38] A. D. Mednykh, "Brahmagupta formula for cyclic quadrilaterials in the hyperbolic plane," *Sib. Èlektron. Mat. Izv.* 9 (2012), 247–255.
MR 2954694 Zbl 1330.51009 ☞ 155, 157

[39] J. Milnor, "Hyperbolic geometry: the first 150 years," *Bull. Amer. Math. Soc.* (*N.S.*) 6 (1982), no. 1, 9–24.
MR 0634431 Zbl 0486.01006 ☞ 169, 170, 171

[40] J. Milnor, "How to compute volume in hyperbolic space," in *Collected papers,* 1. Geometry, Publish or Perish, Houston, TX, 1994, 189–212.
☞ 177

[41] A. F. Möbius, “Ueber die Gleichungen, mittelst welcher aus den Seiten eines in einen Kreis zu beschreibenden Vielecks der Halbmesser des Kreises und die Fläche des Vielecks gefunden werden,” *Crelle’s Journal* 35 (1828), 5–34.
☞ 154

[42] Y. Mohanty, *Hyperbolic polyhedra*: *volume and scissors congruence,* Ph.D. Thesis. University of California, San Diego, 2002.
MR 2703698 ☞ 174

[43] Y. Mohanty, “The Regge symmetry is a scissors congruence in hyperbolic space,” *Algebr. Geom. Topol.* 3 (2003), 1–31.
MR 1997312 Zbl 1026.51015 ☞ 174

[44] J. Murakami and M. Yano, “On the volume of a hyperbolic and spherical tetrahedron,” *Comm. Anal. Geom.* 13 (2005), no. 2, 379–400.
MR 2154824 Zbl 1081.51012 ☞ 174

[45] E. Pascal, *I determinanti*: *teoria ed applicazioni,* U. Hoepli, Milano, 1897.
☞ 175

[46] F. V. Petrov, “Inscribed quadrilaterals and trapezoids in absolute geometry,” *Mat. Prosveshchenie* 13 (2009), 149–154, in Russian.
☞ 152, 153

[47] Ya. P. Ponarin, *Elementary planimetria,* Vol. 1, MCCME, Moscow, 2004, in Russian.
☞ 159, 162

[48] V. V. Prasolov, *Problems and theorems in linear algebra,* translated from the Russian manuscript by D. A. Leĭtes, Translations of Mathematical Monographs 134. American Mathematical Society, Providence, R.I., 1994.
MR 1277174 Zbl 0803.15001 ☞ 176

[49] H. W. Richmond, “The volume of a tetrahedron in elliptic space,” *Quart. J. Math.* 34 (1903), 175–177.
JFM 34.0679.04 ☞ 175

[50] D. P. Robbins, “Areas of polygons inscribed in a circle,” *Discrete Comput. Geom.* 12 (1994), no. 2, 223–236.
MR 1283889 Zbl 0806.52008 ☞ 154

[51] I. Kh. Sabitov, “Solution of cyclic polyhedra,” *Mat. Prosveshchenie* 14 (2010), 83–106, in Russian.
☞ 154

[52] L. Schläfli, “On the multiple integral $\int\int\cdots\int dx dy \ldots dz$ whose limits are $p_1 = a_1 x + b_1 y + \cdots + h_1 z > 0$, $p_2 > 0, \ldots, p_n > 0$ and $x^2 + y^2 + \cdots + z^2 < 1$,” *Quart. J. Math.* 2 (1858), 269–300; ibid. 3 (1860), 54–68, 97–108.
☞ 168, 177

[53] L. Schläfli, “Theorie der vielfachen Kontinuität,” in *Gesammelte mathematishe Abhandlungen,* Birkhäuser, Basel, 1950.
☞ 167

[54] J. J. Seidel, “On the volume of a hyperbolic simplex,” *Studia Sci. Math. Hungar.* 21 (1986), no. 1-2, 243–249.
MR 0898862 Zbl 0561.52010 ☞ 171

[55] G. Sforza, “Ricerche di estensionimetria negli spazi metrico-proiettivi,” Memorie Reale Accademia di Scienze, lettere ed arti di Modena, III, VIII (Appendice) (1907), 21–66.
☞ 175

[56] P. Stäckel, “Vizsgálatok az abszolút geometria köreből Bolyai János hátrahagyott irataiban” (”Investigations in absolute geometry in the manuscripts of János Bolyai”)’ Mathematikai és Természettudományi Értesítő XX, Budapest, 1902, 175–186.
☞ 168

[57] P. Stäckel, *Geometrische Untersuchungen von Wolfgang Bolyai und Johann Bolyai* I.–II., B. G. Teubner, Leipzig, 1913.
☞ 168

[58] I. Todhunter, *Spherical trigonometry,* Macmillan and Co., London, 1886.
☞ 152

[59] A. Ushijima, “Volume formula for generalised hyperbolic tetrahedra,” in *Non-Euclidean geometries* (A. Prékopa and E. Molnár, eds.), János Bolyai memorial volume, papers from the International Conference on Hyperbolic Geometry held in Budapest, July 6–12, 2002, Mathematics and Its Applications (New York), 581. Springer, New York, 2006, 249–265.
MR 2191251 Zbl 1096.52006 ☞ 171, 174, 176

[60] J. E. Valentine, “An analogue of Ptolemy’s theorem and its converse in hyperbolic geometry,” *Pacific J. Math.* 34 (1970), 817–825.
MR 0270261 Zbl 0198.25601 ☞ 153

[61] V. V. Varfolomeev, “Inscribed polygons and Heron polynomials,” *Mat. Sb.* 194 (2003), no. 3, 3–24, in Russian; English translation in *Sb. Math.* 194 (2003), no. 3-4, 311–331.
MR 1992154 Zbl 1067.51013 ☞ 154

[62] R. Walter, “Polygons in hyperbolic geometry 1: Rigidity and inversion of the n-inequality,” preprint, 2010.
arXiv:1008.3404v1 [math.MG]
☞ 151, 153

[63] R. Walter, “Polygons in hyperbolic geometry 2: Maximality of area,” preprint, 2010.
arXiv:1008.3821v1 [math.MG]
☞ 151, 152

[64] T. Weszely, *Bolyai János matematikai munkássága* (*Mathematical works of János Bolyai*), Kriterion Könyvkiado, Bukarest, 1981.
☞ 168

[65] L. Wimmer, “Cyclic polygons in non-Euclidean geometry,” *Elem. Math.* 66 (2011), no. 2, 74–82.
MR 2796129 Zbl 1228.51023 ☞ 152, 153

Chapter 13

Contributions to non-Euclidean geometry I[*]

Eduard Study

The author's intention is to edit a series of articles under a common title concerning algebraic and differential geometry in non-Euclidean space, a subject that he has been investigating for many years although not yet arriving at any other than sketchy publications on the topic.

In the first of the present three contributions,[1] a seemingly new property of the straight line will be discussed. The second article is independent from the first. It is entitled "the notions of left and right,"[2] but it also treats other fundamental notions of non-Euclidean geometry. It is meant to be an introduction to further articles, and therefore it is kept elementary. It may perhaps also serve as an introduction to the methods that the author used in his book *Geometrie der Dynamen* (Leipzig, 1903). The third note contains a simple application to helical surfaces in non-Euclidean space of positive curvature that are considered as geodesics in a four-dimensional manifold.

I. Straight line and point as extrema

It seems not to have been noticed so far that the theorem of the straight line as the shortest path between two points has two analoga in hyperbolic geometry, where the *point* remarkably does not appear as the shortest but as the *longest path between two lines*.

If we extend the point continuum of *Lobatchevsky's* geometry to projective geometry by adding the infinitely far, or so-called ideal points in the well-known way, then one can also state this claim that obviously needs to be made more precise, in correlative form: *the straight line or rather a piece of the straight line appears as the longest path between two points.*

---

[*] "Beitrage zur Nicht-Euklidischen Geometrie. I.", *Amer. J. Math.* 29 (1907), no. 2, 101–116. Translated by Annette A'Campo-Neuen.

[1] [Editor's note] In the present volume, only the first contribution is translated, whose subtitle is *Gerade und Punkt als Extreme*, translated, as below, by *Straight line and point as extrema*.

[2] We plan to show elsewhere that a part of the concept of the notions of left and right does not need the notion of measurement but that it can be considered in the real projective geometry of space.

We first consider the plane geometry of constant curvature -1, associated to the absolute polar system

$$(xy) = x_1y_1 - x_2y_2 - x_3y_3 = 0. \tag{1}$$

Then we can uniquely define the distance (that is here always considered as *positive* ≥ 0) between two points in the interior $\{(xx) > 0, (yy) > 0\}$ of the absolute cone[3] by the formula

$$(x, y) = \operatorname{arccosh} \sqrt{\frac{(xy)^2}{(xx)(yy)}}, \tag{2}$$

where we choose the positive value for the square root and the function arccosh. The same formula provides a real "distance" for the points x, y, even in the case when they lie outside of the absolute cone such that the straight line joining them meets the absolute cone. Moreover, also in this case for three points y, x, z on a straight line, where the one in the middle x, separates the other two, the following equation holds:

$$(y, z) = (y, x) + (x, z). \tag{3}$$

Now consider a triangle x, y, z of distinct points in the exterior such that each of the lines joining two of the points meets the absolute cone. Hence

$$\begin{aligned} (yz)^2 - (yy)(zz) \geq 0, \quad (xx) < 0,\\ (zx)^2 - (zz)(xx) \geq 0, \quad (yy) < 0,\\ (xy)^2 - (xx)(yy) \geq 0, \quad (zz) < 0. \end{aligned} \tag{4}$$

Then one can distinguish two families of such triangles. If

$$(yz)(zx)(xy) < 0, \tag{5a}$$

then the longest side of the triangle x, y, z is always greater than the sum of the other two, or at least equal to their sum. And equality holds if and only if the three points are collinear.

If however

$$(yz)(zx)(xy) > 0, \tag{5b}$$

then the longest side may be greater or smaller than the sum of the other two, or it may be equal to their sum. But in this case the three points are never collinear.

---

[3] [Translator's note] The author uses the term "absoluter Kegelschnitt," i.e. absolute conic section, for the zero set of the bilinear form. Indeed the lines through zero on the cone form a conic section in the projective plane.

Let us now assume that

$$(y,z) \geq (y,x), \quad (y,z) \geq (x,z), \tag{6}$$

and ask under what assumptions we can conclude that

$$(y,z) \geq (y,x) + (x,z). \tag{7a}$$

Obviously, we can rewrite inequality (7a) in this form:

$$\cosh((y,z) - (y,x)) \geq \cosh(x,z).$$

After solving by using the addition theorem for the hyperbolic cosine on the left-hand side, the inequality takes the following form, assuming positive values wherever a square root is taken:

$$\sqrt{(yz)^2}\,\sqrt{(xy)^2} + (yy)\sqrt{(xz)^2} \geq \sqrt{(yz)^2 - (yy)(zz)}\,\sqrt{(xy)^2 - (xx)(yy)}.$$

Here, by the assumptions (4) and (6), both sides are positive. Hence the inequality is equivalent to the one that we obtain by taking squares:

$$(xx)(yy)(zz) - 2\sqrt{(yz)^2\,(zx)^2\,(xy)^2} - (xx)(yz)^2 - (yy)(zx)^2 - (zz)(xy)^2 \geq 0.$$

On the other hand,

$$(xx)(yy)(zz) + 2(yz)(zx)(xy) - (xx)(yz)^2 - (yy)(zx)^2 - (zz)(xy)^2 \geq 0,$$

since the left-hand side is equal to the square of the determinant (xyz). Therefore, if the inequality (5a) holds, then the last two inequalities coincide and (7a) holds. They are both equalities if $(xyz) = 0$. However, we cannot conclude that conversely (7a) implies (5a).

Now let us assume that (5b) holds. Then we claim that any of these relations are possible

$$(y,z) \gtreqless (y,x) + (x,z), \tag{7b}$$

or as above (assuming (6) *analogously*) the following may occur

$$(xyz)^2 - 4(yz)(zx)(xy) \gtreqless 0.$$

In fact, one can verify this easily by examples or by the argument given below. In this case (xyz) never vanishes because then in (7b) we would have equality. Therefore the product $(yz)(zx)(xy)$ would also vanish. But this is impossible by (4).

In order to obtain a clear idea of the corresponding triangular figures we want to assume that the points y, z are given. It is easy to see that then the set of all points x satisfying the inequalities (5a) or (5b) is represented by the shaded regions in Figure 1.

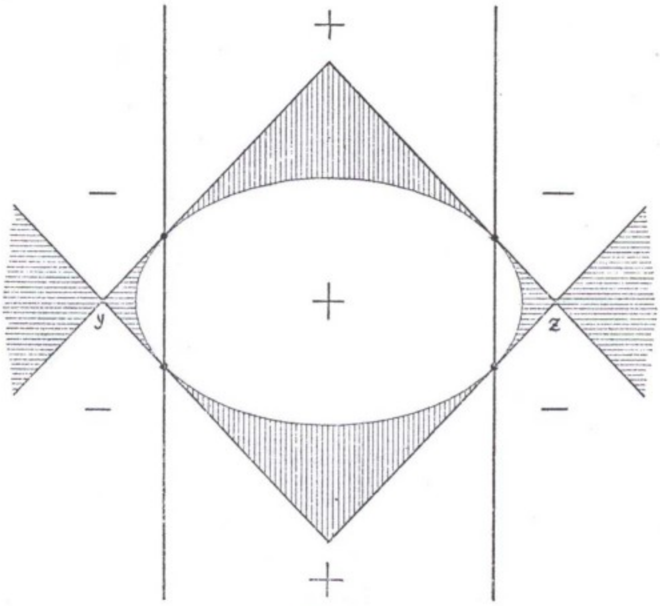

FIG. 1.

Vertical shading indicates that

$$(yz)(zx)(xy) > 0,$$

and horizontal shading indicates that

$$(yz)(zx)(xy) < 0.$$

Moreover, Figure 2 indicates with vertical and horizontal shading where

$$(y,z) - (y,x) - (x,z) \geq 0, \quad \text{or} \quad (y,z) - (y,x) - (x,z) \leq 0.$$

The figure also shows with fat lines the locus of all points x where

$$(y,z) - (y,x) - (x,z) = 0.$$

This locus consists of three completely separate curves. One of these curves is linear whereas the other two are parts of the curve of order 2

$$(yzx)^2 - 4(yz)(yx)(zx) = 0. \tag{8}$$

Firstly, this curve touches the polar lines of the points y, z, that is the lines $(yx) = 0$ and $(zx) = 0$, at their intersection points with $(yzx) = 0$. Now if we write down its equation in line coordinates using the notation

$$(ux) = x_1u_1 + u_2x_2 + u_3x_3, \quad (uv) = u_1v_1 - u_2v_2 - u_3v_3,$$

then after splitting off the non-vanishing factor $4(yz)\{(yz)^2 - (yy)(zz)\}$ from the quadratic covariant of the expression (8), one finds the equation

$$(yz)(uu) - (uy)(uz) = 0. \tag{9}$$

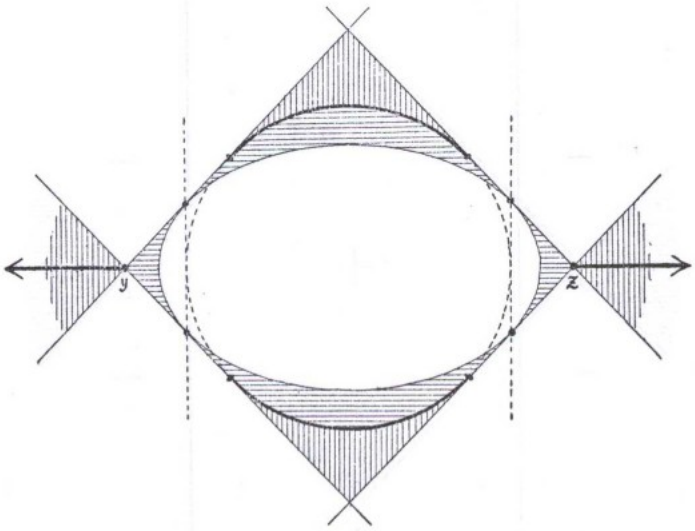

FIG. 2.

Hence one can see that the curve (8) touches the minimal lines through y, z, i.e. the tangents to the absolute cone going through y, z, of course in real points. Finally, through the two touching points of the curve (8) with the minimal lines through z there is the locus of $(y, x) = (y, z)$ which is not drawn in Figure 2, but could easily be added. The regions within the curves $(y, x) = (y, z)$ and $(z, x) = (y, z)$ belong to the inequalities $(y, z) < (y, x)$ and $(y, z) < (z, x)$, and are therefore horizontally shaded. The curve (8) however, separates the two kinds of shading wherever it lies *outside* of these regions. Now it is easy to see that the shadings are carried out correctly by letting x run through the boundary curves of the different regions.

For greater clarity, we add here a third figure—in collinear modification—showing for a given x, where the points y and z have to be if we require simultaneously

$$(yz)(zx)(xy) < 0 \quad \text{and} \quad (y, z) \geq (y, x) + (x, z). \tag{10}$$

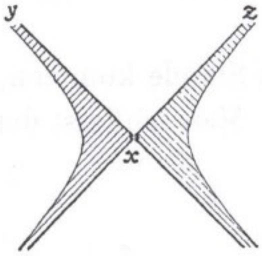

FIG. 3.

In this case, which will now be investigated further, we will say: "The point x separates the points y and z," or "The point x lies between y and z." If there is no minimal line among the lines joining the points x, y, z then the polar lines of the points x, y, z are in the relative position indicated in Figure 4a, whereas the triangles of the second family correspond to Figure 4b.

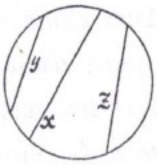

FIG. 4a.

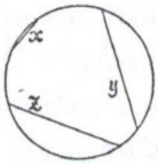

FIG. 4b.

In order to avoid lengthy arguments, for the following theorems we refer to intuition (see Figure 4a), or an obvious continuity consideration:

If the point x' lies between y and z, and the point x'' lies between (y and x' or) x' and z, then x'' also lies between y and z.

If two different points lie between y and z in such a way that the straight line joining them meets the absolute cone then all four points have a certain order y, x', x'', z or z, x'', x', y. More precisely, after suitable choice of notation, x' lies between y and x'', and x'' lies between x' and z, but not x'' between y and x', and not x' between x'' and z.

If, moreover, the point x lies between two such points x' and x'', then x also lies between y and z.[4]

---

[4] All this could be proved rigorously with purely algebraic means but as far as we can see the proof would not really be short. Unless it concerns fundamental theorems of deep theories, we may well permit ourselves to shorten the presentation if it is possible to fill the gaps easily if needed.

The two drawings in Figure 5, may sufficiently illustrate how those theorems are derived.

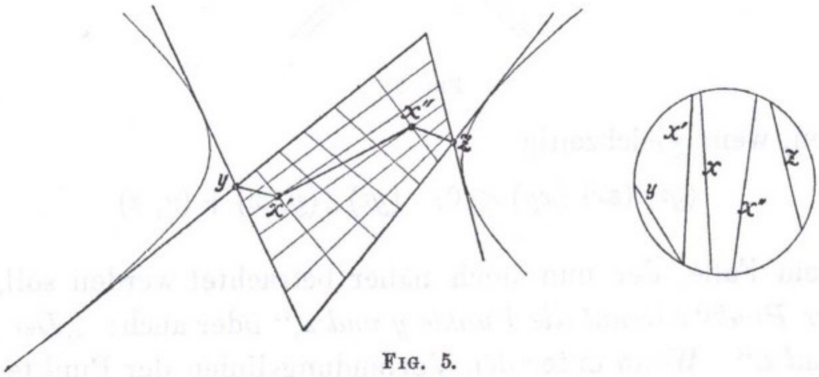

FIG. 5.

Now it is easy to conclude the following remarkable theorem:

Suppose that there is a given set of points outside of the absolute cone, forming a continuum, and such that each straight line joining any two different points of the set meets the absolute cone. Then every point of the set with at most two exceptions y_0, z_0, *has the following property: If this point* x *is removed from the set then the remaining points of the set form two again continuous subsets* (y) *and* (z) *such that* x *separates all the points* y *of the first set from all the points* z *of the second set.*

All the points of the set lying between two different points y, z *form together with* y, z *a closed, continuous curve whose points* x *can be assigned bijectively and continuously the values of a parameter* t *between* 0 *and* 1.

This curve has a well-determined length S *satisfying the inequalities*

$$0 \leq S \leq (y, z).$$

Moreover, if there are two exceptional points y_0, z_0 as mentioned above, then the entire set consists of these two points and a chain of curves joining them with a certain arc length and whose points different from y_0, z_0 all lie between y_0 and z_0. The points y, z referred to in the theorem always preserve the property of such limit points y_0, z_0 if one removes all those points of the set different from y, z that do *not* lie between y and z.

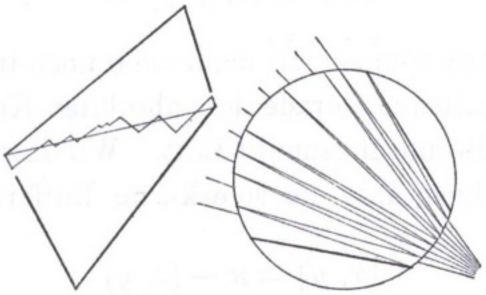

Fig. 6.

Such a chain of curves as described in the theorem will be called a *path* between y and z. Now we have the result announced in the beginning:

If y and z are two different points outside of the absolute cone, whose joining straight line meets the absolute cone, then the straight line is the longest among all paths between y and z.

Note these further useful remarks: *If the straight line joining the points y, z is a minimal line, i.e. if it touches the absolute cone, then this straight line is the unique path between y and z. In every other case, there are infinitely many paths between y and z, and arbitrarily close to any such path there are infinitely many other paths of length zero.*

Figure 6 illustrates how to construct a path that is arbitrarily close to the longest path, but itself is of length zero.

For the points of a straight line *not* intersecting the absolute cone, i.e. a line lying entirely in the ideal domain, we can define an elliptic measurement.[5] The distance of two points x, y on such a line will then only be determined up to sign and up to a multiple of a certain period, that we can assume to be π. Among the different values of this function, we choose the one with smallest absolute value, consider it as a positive value and denote it by (x, y). If moreover, we choose for all occurring roots their positive values, we obtain the equation

$$\cos(x, y) = \sqrt{\frac{(xy)^2}{(xx)(yy)}}, \tag{11}$$

and we always have

$$0 \leq \cos(x, y) \leq 1.$$

---

[5] [Translator's note] The author uses the term "Massbestimmung." It is not a metric however, since for the points in the exterior the triangle inequality does not always hold.

We can think of the same formula (11) as being true in the limiting case where the straight line containing x, y *touches* the absolute cone, but neither x nor y is the point of contact. We then set $(x, y) = 0$.

Now we want to show that the hereby defined distance (x, y) as well as its supplement

$$[x, y] = \pi - (x, y)$$

also has the property of being a *maximum* among all the lengths of certain "paths" between the two points, that still have to be defined.

Again we consider a triangle x, y, z of different ideal points but now such that the straight lines joining any two points shall *not meet or at most touch* the absolute cone:

$$\begin{aligned} (yz)^2 - (yy)(zz) &\leq 0, \quad (xx) < 0, \\ (zx)^2 - (zz)(xx) &\leq 0, \quad (yy) < 0, \\ (xy)^2 - (xx)(yy) &\leq 0, \quad (zz) < 0. \end{aligned} \tag{12}$$

Without loss of generality, we may assume that

$$\frac{\pi}{2} \geq (y, z) \geq (y, x) \geq (x, z) \geq 0. \tag{13}$$

The inequalities that we will have to assume consecutively

$$(y, z) \gtreqless (y, x) + (x, z) \tag{14}$$

may then be written in the following form:

$$\cos(x, z) - \cos(y, z)\cos(y, x) \gtreqless \sin(y, z) \cdot \sin(y, x).$$

Since here by assumption both sides are positive values, taking squares leads to equivalent inequalities

$$(xx)(yy)(zz) - 2\sqrt{(yz)^2(zx)^2(xy)^2} - (xx)(yz)^2 - (yy)(zx)^2 - (zz)(xy)^2 \gtreqless 0.$$

As before one can see that here only $>$ or equality is possible if

$$(yz)(zx)(xy) \leq 0, \tag{15}$$

whereas in the case

$$(yz)(zx)(xy) > 0 \tag{16}$$

one cannot draw a conclusion.

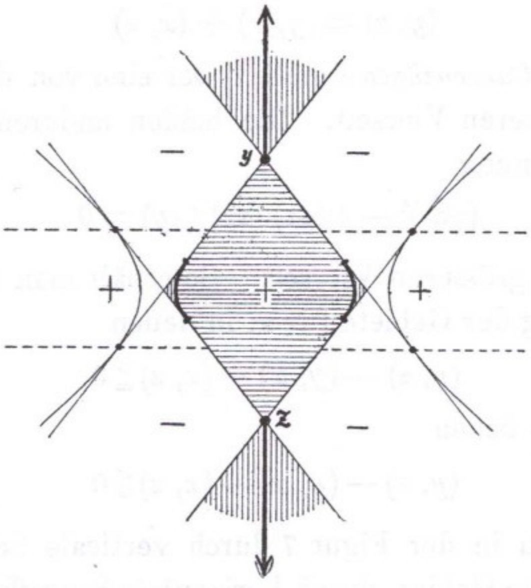

FIG. 7.

It will be sufficient if moreover we first exclude the easily treated limiting cases

$$(y, z) = \frac{\pi}{2} = [y, z] \quad \text{and} \quad (y, z) = 0, [y, z] = \pi$$

and hence assume that

$$0 < (y, z) < \frac{\pi}{2} < [y, z] < \pi.$$

Each of the two segments into which the straight line yz is decomposed by the two points y,z is now a diagonal of a simple quadrangle whose sides are pieces of minimal lines (compare Figure 7).

In the interior or on the boundary of these two quadrangles are all the points x satisfying the inequalities (12). Apart from that both quadrangles have different properties. By integrating the arc element along a straight line between y and z through the interior of one quadrangle one obtains the smaller value (y, z), and by integrating along a straight line through the other quadrangle one obtains the larger value $[y, z]$. The first “smaller" quadrangle as we call it for simplicity, is completely contained in the domain $(yz)(zx)(xy) < 0$. The other “larger" quadrangle is decomposed by the absolute polar lines of the points y, z into three parts, two outer parts where also $(yz)(zx)(xy) < 0$, and one inner part, where $(yz)(zx)(xy) > 0$. Moreover, analogously to above one finds *that the locus of all points x with*

$$(y, z) = (y, x) + (x, z) \tag{17}$$

consists of three separate curves. The first one of these is linear and lies entirely in the smaller quadrangle. The other two lie on the irreducible curve of degree 2

$$(yzx)^2 - 4(yz)(zx)(xy) = 0 \tag{18}$$

and are contained entirely in the larger quadrangle. Thus one finally obtains insight into the distribution of the regions of x where

$$(y, z) - (y, x) - (x, z) \geq 0$$

holds and the other ones with

$$(y, z) - (y, x) - (x, z) \leq 0.$$

In Figure 7, the first ones are indicated by vertical shading and the second ones or more precisely the *one* second domain is indicated by horizontal shading.

Now as above for the hyperbolic measurement, we obtain a notion of *path* between y and z lying in the smaller quadrangle, and we find the theorem that among all such paths the straight line is the longest. We further claim that in the larger quadrangle the same relations hold. That is, we claim that in the triangle x, y, z the length of the side yz is not less than the sum of the lengths of the other two sides if x is contained in the larger quadrangle and if all three lengths are defined via integration of the arc element along a straight line such that the path of integration does not leave the larger quadrangle.

One can see without difficulty that for the proof of this claim one has to distinguish three cases.

1) The point x lies in the middle part of the larger quadrangle (or on its boundary). Then $(yz)(yx)(zx) \geq 0$. The three lengths are $[y, z]$, (y, x) and (x, z). The claim is

$$[y, z] \geq (y, x) + (x, z).$$

2) and 3) The point x lies in one of the outer parts, e.g. in the part cut off by the polar line of the point y, or such that $(y, x) \geq (x, z)$. Then $(yz)(yx)(zx) \leq 0$, the three lengths are $[y, z]$, $[y, x]$, (x, z), and the claim is

$$[y, z] \geq [y, x] + (x, z) \quad \text{or} \quad (y, x) \geq (y, z) + (z, x).$$

In all these cases the proof can be carried out in the same manner as above.

Now we will state the result in a shorter form than previously and we will essentially only point out what is different from the previous theorem.

We call the used kind of measurement *pseudo-elliptic* since it coincides with a usual elliptic one in constant positive curvature $(+1)$, but differs substantially from it by the existence of paths of length zero. Using this terminology we can then say:

If y, z are two different points whose straight line joining them does not intersect the absolute cone, then there are two families of direct paths between y and z. Namely, there are two simple quadrangles bounded by minimal lines having y and z as opposite vertices. In the interior or on the boundary of each of these quadrangles there are the paths of the first or the second family respectively. Among all paths of the same family, using pseudo-elliptic measurement, the straight line appears as the longest.

Each direct path is a piece of a curve of a certain arc length whose secants either do not meet or at most touch the absolute cone. By any of its points x different from y and z, it is decomposed into two paths with the same property. If in the limiting case that we first excluded, y and z remain ideal points such that the straight line yz touches the cone then there is exactly *one* "proper" and *one* "improper" direct path between y and z. Both are linear. The first one lies entirely outside of the absolute cone; it has length zero. The length of the second path is not properly defined, but maybe set $= \pi$.

A piece of a curve can have the property that around each of its points one can draw a region where the piece of the curve is a direct path between any two of its points y', z', even though it may not have this property as a whole. For such pieces of curves, joining two given points y, z, the last theorem does not hold. On the contrary, one can obviously join two arbitrary ideal points—even those whose joining straight line intersects the absolute cone in separate points—by such curves whose length exceeds an arbitrary given number. The length of the direct path between the points y, z, if such a path exists, is hence in a much more restricted sense a maximum than the path length considered with respect to hyperbolic measurement.

Since the introduction of the so-called ideal points in Lobatchevsky's geometry is not really necessary, one might prefer a different formulation of the presented theorems avoiding that notion. Of course the translation from one language into the other can be carried out easily. For example the first result that we discussed can be rephrased as follows:

If among three straight lines y, x, z, that are pairwise either parallel or ultraparallel, the middle one separates the other two, then the length of the common normal between y and z is greater or equal to the sum of the lengths of the normals between y, x and between x, z.

Equality occurs if all three lines have a common perpendicular, or in the limiting case, if they are all parallel. If y, z are ultraparallel, the "length of the common normal between y and z" means the length of the *shortest* joining line between a point of y and a point of z, and in the limiting case of being parallel, length zero. Everything else follows immediately: The ideal point of intersection of two ultraparallel lines y, z, or its polar line, the common perpendicular of y, z, considered as the locus of the lines perpendicular to it, now appears as the *longest path* between these two lines.

Similarly, one can translate our second theorem with its corollaries. The fundamental fact in this case turns out not to be different from the well-known theorem on angles of a triangle, that can be expressed as follows:

In each ordinary triangle one of the outer angles is greater or equal to the sum of the two inner angles that occur at the side opposite to the outer angle. And equality holds if and only if the three sides go through one common point.

It is less intuitive to work with lines and their angles rather than with points and their distances. In the first case one does not have as in the second case a guiding principle in the analogous situation in Euclidean geometry. And that is the reason why up to now, one has not given this well-known result the turn on to which we focus here (if indeed the matter might be new). But if we present the straight line here as a however relative *maximum*, then this is certainly a fact that should not be missing in a complete system of non-Euclidean geometry. More generally, it may well be permitted to mention that as it seems to us, the development of non-Euclidean geometry is still very incomplete even in very elementary aspects. For example Lobatchevsky's trigonometry is certainly a very important part but always only a part: triangles with all or some ultraparallel sides, in particular the hexagons with only right angles corresponding to the first kind, have also to be considered in a system that claims some kind of completeness. The theory available at present still lacks finer perfection, the investigation of the corresponding groups, the connection to invariant theory etc., even though the tools for such investigations have long since been provided for.[6]

---

[6] The methods and formulae available in the literature are completely sufficient for those purposes. See S. W. Jacobsthal, article "Trigonometrie" in *Encyclopädie der elementaren Geometrie* (Leipzig, 1905), and E. Study, "Sphärische Trigonometrie" etc. (*Abh. d. Sächs.,* G. D. W. Leipzig, 1893), and E. Study, *Geometrie der Dynamen,* Leipzig 1903, pp. 209–212.

Now asking for an extension of the deduced results to arbitrary dimension, we find a remarkable difference among the sketched little theories.

Our first theorem on the straight line as longest path can be generalized to an arbitrary dimension, the second theorem only holds for the plane.

Namely, let the absolute polar system be given by

$$(xy) = x_1 y_1 - x_2 y_2 - \cdots - x_n y_n = 0.$$

Now first consider two points y, z outside of the absolute cone $(xx) = 0$ such that the straight line joining them meets the cone. In every plane through the line yz, we obtain the triangle figures that we have already investigated. So for hyperbolic measurement, instead of the quadrangle that we discussed above, we obtain a piece of the space of dimension $(n-1)$, bounded by pieces of two conoids emanating from y and z

$$(yx)^2 - (yy)(xx) = 0, \quad (zx)^2 - (zz)(xx) = 0.$$

Now "paths" in this piece of space have to be defined basically as before, but they do not have to lie in planes. The straight path is the longest among all such paths.

However, if the line yz does not meet the absolute cone then there are three different kinds of planar paths between y, z. Firstly, a plane through y, z might permeate the absolute cone in a real curve. For pseudo-elliptic measurement, (y, z) or $[y, z]$ appears as the maximum among all suitable piecewise linear paths. Secondly, the plane might touch the absolute cone. It is easy to see that in this case there are infinitely many equally long paths between both points. Finally, the plane may not intersect the absolute cone at all. In this case (y, z) is a minimum. *Hence the length of the straight linear path in this case is not extremal.*

But this obviously does not exclude the possibility that on a manifold of lower dimension with such a measurement, the problem of shortest (or longest) lines may get a clear content. For example in the eight-dimensional space, there is a four-dimensional manifold M_4 with the following strange property for suitably chosen pseudo-elliptic measurement:

The length of the shortest path measured on M_4 itself between two points y, z in M_4 equals the length of the straight linear path between these two points.[7]

We only want to sketch other generalizations of the matters discussed. Similar extremal properties as the ones considered here, can be found for geodesics belonging to an arc element whose square is defined by a quadratic differential form in the case that this form for line elements yields a hyperbolic measurement. Moreover, in some parts of our previous considerations, one can replace the cone by an arbitrary oval. Somewhat related to these ideas is for example the following easily proved but maybe still not completely uninteresting theorem:

---

[7] E. Study, "Kürzeste Wege im komplexen Gebiet," *Math. Ann.* 60 (1905), p. 361.

Consider a set of points in the finite part of the Euclidean plane forming a closed continuum and having the property that all secants between different points of the set have an angle φ *with a given straight line satisfying an inequality of the form*

$$|\tan(\varphi)| \leq M.$$

Then this point set is a rectifiable curve joining certain points y, z. *If* D *is the distance of the projection of the points* y, z *to the given line, then the arc length of this curve is at most equal to*

$$|D\sqrt{1+M^2}|.$$

Chapter 17

Hermitian trigonometry

Boumediene Et-Taoui

Contents

1. Introduction 413
2. Congruence theorem for triangles in $\mathbb{C}P^2$ 414
References 423

1 Introduction

The aim of this chapter is to present the trigonometry of complex projective spaces. Section 2 is devoted to a congruence theorem for triangles in $\mathbb{C}P^2$ obtained by Brehm. We prove it here by another method which we presented in [3] and in [7]. It is shown that a triangle $([x],[y],[z])$ in $\mathbb{C}P^2$ with $\|x\| = \|y\| = \|z\| = 1$ is uniquely determined up to isometry by its side lengths $a = \arccos(|\langle y,z\rangle|)$, $b = \arccos(|\langle x,z\rangle|)$ and $c = \arccos(|\langle x,y\rangle|)$ and its *shape invariant*, the real part of the complex number

$$\langle x,y\rangle\,\langle y,z\rangle\,\langle z,x\rangle = \cos(a)\cos(b)\cos(c)e^{i\omega}.$$

Section 3 consists of Hermitian trigonometry. as it is developed in [4]. Precisely, we give the two laws of sines and the law of cosines of a triangle in $\mathbb{C}P^2$. In Section 4, we define the polar triangle associated to a given triangle and derive from it the second law of sines which does not appear in [2]. On the other hand, if we consider a triangle $([x],[y],[z])$ and its polar triangle $([X],[Y],[Z])$, and if we set

$$\langle X,Y\rangle\,\langle Y,Z\rangle\,\langle Z,X\rangle = \cos(p)\cos(q)\cos(r)e^{i\Omega},$$

we show that ω and Ω are linked. Indeed, consider the complex number

$$\langle u_2,u_1\rangle\,\langle v_1,v_2\rangle\,\langle w_1,w_2\rangle = |\langle u_2,u_1\rangle\,\langle v_1,v_2\rangle\,\langle w_1,w_2\rangle|\,e^{i\alpha},$$

where $u_1, u_2; v_1, v_2; w_1, w_2$ are pairs of unit tangent vectors at the vertex $[x]$, $[y]$, $[z]$ respectively. It is displayed that

$$\Omega = -\alpha - 2\omega + (2k+1)\pi, \quad k \in \mathbb{Z}.$$

This result is new in comparison with the results in the papers quoted. In addition, we relate Brehm's invariants to those of Hsiang [10] and those of Shirokov and Rozenfeld [14]. We close this paper by some remarks on trigonometry in general symmetric spaces.

2 Congruence theorem for triangles in $\mathbb{C}P^2$

In classical geometry, Euclidean, spherical, and hyperbolic spaces are commonly characterized by the same set of congruence axioms for geodesic intervals and triangles. In the language of Riemannian geometry, these axioms are stated as follows ([10]):

1. two geodesic intervals AB and $A'B'$ are congruent if they have the same length;
2. two geodesic triangles are congruent if they have two equal angles contained by a pair of equal sides.

Let ABC be a geodesic triangle in a symmetric space M of rank 1 and $\Delta(M)$ the moduli space of congruence classes of triangles in M. In the well-known classical spaces, a triangle has six basic invariants (side lengths and angles) which map $\Delta(M)$ into $\mathbb{R}^6$. The image set is of dimension 3 since the 6 invariants satisfy three basic functional relations, the laws of sines (which involve 2 equalities) and cosines (one equality). Riemannian symmetric spaces of rank 1 are precisely the two-point homogeneous spaces [18].

In 1994, in his Ph.D. thesis [7], the author gave new congruence criteria for finite subsets in both complex projective and complex hyperbolic spaces. The case of triangles is then easily derived. Using this point of view, he derived an easy congruence theorem for triangles. This leads directly to the laws of trigonometry in $\mathbb{C}P^2$ obtained by Brehm, from which one can easily derive those of the sphere.

We consider points in $\mathbb{C}P^n$, $n \geq 2$ as one-dimensional subspaces of the vector space $\mathbb{C}^{n+1}$. For $x \in \mathbb{C}^{n+1}$, $x \neq 0$, let $[x] := \mathbb{C} \cdot x$ denote the corresponding point in $\mathbb{C}P^n$. Let $\langle x, y\rangle = \sum_{i=1}^{n+1} \overline{x_i} y_i$ be the usual Hermitian product and let d be the Fubini–Study metric on $\mathbb{C}P^n$ defined by

$$d([x],[y]) = \arccos\Big(\frac{|\langle x, y\rangle|}{\|x\|\,\|y\|}\Big).$$

It is well known that $\Phi\colon \mathbb{C}P^n \to \mathbb{C}P^n$ is an isometry of $\mathbb{C}P^n$ if and only if there exists a unitary matrix U such that either $\Phi([x]) = [Ux]$ for all $x \in \mathbb{C}^{n+1}$ or $\Phi([x]) = [\overline{Ux}]$ for all $x \in \mathbb{C}^{n+1}$.

Let $[x]$, $[y]$, $[z] \in \mathbb{C}P^n$ with $\|x\| = \|y\| = \|z\| = 1$.

Definition 2.1. A *triangle* is a triple $([x], [y], [z])$ of three different points in $\mathbb{C}P^n$ such that the distance between any two of them is strictly smaller than $\frac{\pi}{2}$.

We imagine the vertices $[x]$, $[y]$, $[z]$ of a triangle being connected by the unique shortest paths between them, called sides. We set

$$a = \arccos(|\langle y, z\rangle|), \quad b = \arccos(|\langle x, z\rangle|), \quad c = \arccos(|\langle x, y\rangle|).$$

Consider the Hermitian matrix

$$H = \begin{pmatrix} 1 & \langle x, y\rangle & \langle x, z\rangle \\ \langle y, x\rangle & 1 & \langle y, z\rangle \\ \langle z, x\rangle & \langle z, y\rangle & 1 \end{pmatrix},$$

which is positive semidefinite with 1's on the diagonal. This matrix can be brought, by multiplication of columns by unimodular complex numbers and the corresponding rows by the conjugate of those complex numbers into the following form

$$\begin{pmatrix} 1 & \cos(c) & \cos(b) \\ \cos(c) & 1 & \cos(a)e^{i\omega} \\ \cos(b) & \cos(a)e^{-i\omega} & 1 \end{pmatrix}. \tag{1}$$

The complex number $e^{i\omega}$ is called the *direct shape invariant* in [3]. It was considered by Blaschke and Terheggen. Terheggen interpreted it in spherical geometry [15]. The real number $\cos\omega$ is called by Brehm the *shape invariant.* It was shown by Brehm [4] that a triangle in $\mathbb{C}P^n$ is determined, up to isometry, by its side lengths and its shape invariant. Brehm used the invariant $\sigma = \cos(a)\cos(b)\cos(c)\cos(\omega)$. The author showed in his Ph.D. thesis that two triangles in $\mathbb{C}P^n$ are directly congruent (if we do not allow conjugation) if and only if they have equivalent associated matrices. We already saw that the form (1) is determined up to equivalence by the real numbers a, b, c, and $\cos(\omega)$ (conjugation is allowed). Now let H be any Hermitian positive semidefinite 3-matrix with 1's on the diagonal. Clearly it can be decomposed as the product T^*T where T is an upper triangular matrix $(n+1, 3)$, $n \geq 2$ with diagonal positive real numbers and T^* is the transpose of its conjugate. The columns of T are the representing vectors of the desired triangle.

More general results on finite subsets in complex projective spaces can be found in [3].

A straightforward calculation gives the following vectors which are the columns of the upper triangular matrix T,

$$\begin{aligned} x &= (1, 0, 0, \dots, 0), \\ y &= (\cos(c), \sin(c), 0, \dots, 0), \\ z &= (\cos(b), z_2 + z_2' i, z_3, 0, \dots, 0), \end{aligned}$$

where

$$z_2 = \frac{\cos(a)\cos(\omega) - \cos(c)\cos(b)}{\sin(c)}, \tag{2}$$

$$z_2' = \frac{\cos(a)\sin(\omega)}{\sin(c)}, \tag{3}$$

$$z_3 = \frac{(1 - \cos^2(a) - \cos^2(b) - \cos^2(c) + 2\cos(a)\cos(b)\cos(c)\cos(\omega))^{1/2}}{\sin(c)}. \tag{4}$$

The triple $([x], [y], [z])$ is called in [4] the normal form of the congruence class of the triangle. We have then proved the following theorem due to Brehm. The proof presented here is slightly different from Brehm's one and was obtained in the author's Ph.D. thesis [7] and published in [3].

Theorem 2.2. 1. *A triangle in* $\mathbb{C}P^n$ *is determined uniquely up to isometry by its side lengths and its shape invariant.*

2. *Let* $0 < a, b, c < \frac{\pi}{2}$ *and* $\sigma \in \mathbb{R}$. *Then there exists a triangle in* $\mathbb{C}P^n$ *with side lengths* a, b, c *and shape invariant* σ *if and only if*

$$\frac{1}{2}(\cos^2(a) + \cos^2(b) + \cos^2(c) - 1) \leq \sigma \leq |\sigma| \leq \cos(a)\cos(b)\cos(c).$$

3 Complex projective trigonometry

Let $x \in \mathbb{C}^{n+1}$, $\|x\| = 1$. The tangent space of $\mathbb{C}P^n$ at $[x]$ can be canonically identified with $(\mathbb{C} \cdot x)^{\perp}$ in $\mathbb{C}^{n+1}$. The real part $\Re(\langle , \rangle)$ of the Hermitian product is the Riemannian metric of $\mathbb{C}P^n$. We represent as in [4] the tangent unit vector $u(x, [y])$ at $[x]$ of the shortest geodesic from $[x]$ to $[y]$ as

$$u(x, [y]) := \frac{y\,\langle y, x\rangle - x\,|\langle y, x\rangle|^2}{\|y\,\langle y, x\rangle - x\,|\langle y, x\rangle|^2\,\|}.$$

Consider the normal form given above and let us set $u_1 = u(x, [y])$ and $u_2 = u(x, [z])$. By a straightforward calculation one has

$$u_1 = (0, 1, 0, \dots, 0), \quad u_2 = (0, z_2 + z_2' i, z_3, 0, \dots, 0) \sin^{-1}(b).$$

The angular invariants of [4] are defined by

$$\begin{aligned}\Lambda(A) &:= \arccos(\Re(\langle u_1, u_2 \rangle)),\\ \Phi(A) &:= \arccos(|\langle u_1, u_2 \rangle|),\\ \Psi(A) &:= \arcsin(|\Im(\langle u_1, u_2 \rangle)|),\end{aligned}$$

where A, B, C correspond to the corners $[x]$, $[y]$, $[z]$ respectively. Geometrically, $\Lambda(A)$ is the angle between the geodesics, $\Phi(A)$ is the angle between the projective lines generated by u_1 and u_2 and $\Psi(A)$ is the smallest angle between the geodesic from $[x]$ to $[z]$ and a totally geodesic $\mathbb{R}P^2$ containing $[x]$ and $[y]$. However, the three angles are dependent, they satisfy $\sin^2 \Lambda(A) = \sin^2(\Phi(A)) + \sin^2(\Psi(A))$. Clearly the pair of angular invariants of one corner together with the lengths of its adjacent sides constitute a minimal set of congruence conditions. Thus, we have $\dim \Delta(\mathbb{C}P^n) = 4$. Therefore there exist five basic functional relations among the nine invariants of a triangle.

A short computation yields

$$\cos(\Lambda(A)) = \frac{z_2}{\sin(b)}, \tag{5}$$

$$\sin(\Psi(A)) = \frac{z_2'}{\sin(b)}, \tag{6}$$

$$\sin \Phi(A) = \frac{z_3}{\sin(b)}, \tag{7}$$

Theorem 3.1. *For any triangle ABC in $\mathbb{C}P^n$,*

$$\frac{\sin(\Phi(A))}{\sin(a)} = \frac{\sin(\Phi(B))}{\sin(b)} = \frac{\sin(\Phi(C))}{\sin(c)}, \tag{8}$$

$$\frac{\sin(\Psi(A))}{\sin(2a)} = \frac{\sin(\Psi(B))}{\sin(2b)} = \frac{\sin(\Psi(C))}{\sin(2c)}, \tag{9}$$

$$\begin{aligned}&\cos^2(a) - (\cos(b)\cos(c) + \sin(b)\sin(c)\cos(\Lambda(A)))^2\\ &\quad = \sin^2(\Psi(A))\sin^2(b)\sin^2(c).\end{aligned} \tag{10}$$

Proof. Equation (8) is obtained using (4) and (7). Equation (9) is obtained from (3) and (6). Now for the law of cosines we have from (2) and (5),

$$\begin{aligned}&\cos(a)\cos(b)\cos(c)\cos(\omega)\\&\quad=\cos(b)\cos(c)\sin(b)\sin(c)\cos(\Lambda(A))+\cos^2(b)\cos^2(c),\end{aligned}\tag{11}$$

and from (4) and (7),

$$\begin{aligned}&2\cos(a)\cos(b)\cos(c)\cos(\omega)\\&\quad=\sin^2(\Phi(A))\sin^2(c)\sin^2(b)-1+\cos^2(a)+\cos^2(b)+\cos^2(c).\end{aligned}\tag{12}$$

Combining (11), (12), and the relation between Λ, Φ and Ψ we arrive easily to the law of cosines (10). □

The law of cosines (10) does not appear in this form in Hsiang's and Brehm's articles. From (10) we find again the well-known law of cosines in elliptic and spherical spaces.

Brehm's law of cosines reads as follows:

$$\begin{aligned}&\cos(2b)\cos(2c)-\cos(2a)+\cos(\Lambda(A))\sin(2b)\sin(2c)\\&\quad=2\sin^2(\Phi(A))\sin^2(b)\sin^2(c).\end{aligned}\tag{13}$$

Remark 3.2. 1. If $\Phi(A)=0$ then (13) is exactly the classical cosine law for spherical triangles, noting that $\mathbb{C}P^1$ is isometric to the sphere of radius $\frac{1}{2}$.

2. If $\Psi(A)=0$ then (10) is exactly the classical cosine law for elliptic and thus spherical triangles.

3. The congruence theorem for geodesic triangles was stated for the first time by Blaschke and Terheggen [2]. They showed that the congruence class of a geodesic triangle in a complex projective plane is determined by its side lengths and a fourth invariant $\cos\phi$, satisfiying

$$\cos(\Phi(A))\sin(\phi)=\sin(\Psi(A)).$$

Later, Terheggen interpreted in [15] the invariant ϕ as an angle in spherical geometry. In addition, he showed that all spherical trigonometry in the sense of Study [16] is contained in the geometry of the complex projective line. Although Blaschke and Terheggen were unaware that we need five relations, they obtained three. By means of the notion of polar triangle in a complex projective plane, they obtained a sine and a cosine law. However, it is seen below that the second sine law can be easily derived from their investigation. Their sine law leads easily to the well-know spherical sine law obtained by Coolidge [5]. Another approach was put forward by Shirokov, in a paper published posthumously by Rozenfeld [14]. Shirokov took two angular invariants at each vertex: the Riemannian angle $\Lambda(A)$ between the two sides and the holomorphy inclination $\gamma(A)$. It is seen that

$$\sin(\Lambda(A))\cos(\gamma(A)) = \sin(\Psi(A)).$$

4 Blaschke and Terheggen laws

In $\mathbb{C}^3$, for any vectors x and y, Blaschke and Terheggen defined the product $x \times y$ which is a vector in $\mathbb{C}^3$ with coordinates

$$x_s y_t - x_t y_s, \quad (s,t) \in \{(1,2),(2,0),(0,1)\}$$

and which satisfies

1. $\langle x \times y, \bar{x}\rangle = 0$, $\langle x \times y, \bar{y}\rangle = 0$,
2. $\langle x \times y, u \times v\rangle = \langle x, u\rangle \langle y, v\rangle - \langle x, v\rangle \langle y, u\rangle$,
3. $(x \times y) \times (u \times v) = (xuv)y - (yuv)x$.

Item 2 is Bianchi's identity and (xuv) is the determinant of the three vectors. Now let $([x],[y],[z])$ be a triangle in $\mathbb{C}P^2$ with $\|x\| = \|y\| = \|z\| = 1$, $a = \arccos(|\langle y, z\rangle|)$, $b = \arccos(|\langle x, z\rangle|)$, and $c = \arccos(|\langle x, y\rangle|)$.

Definition 4.1. The triple $([X],[Y],[Z])$ with

$$X = \frac{\bar{y} \times \bar{z}}{\sin(a)}, \quad Y = \frac{\bar{z} \times \bar{x}}{\sin(b)}, \quad Z = \frac{\bar{x} \times \bar{y}}{\sin(c)}$$

is called the *polar triangle* of the triangle $([x],[y],[z])$.

Now set $|\langle X, Y\rangle| = \cos r$, $|\langle Y, Z\rangle| = \cos p$ and $|\langle Z, X\rangle| = \cos q$. Let us also set $v_1 = u(y,[x])$, $v_2 = u(y,[z])$, $w_1 = u(z,[y])$ and $w_2 = u(z,[x])$. Now consider the complex numbers ω, α and Ω:

$$\langle x, y\rangle \langle y, z\rangle \langle z, x\rangle = \cos(a)\cos(b)\cos(c)e^{i\omega},$$

$$\langle u_2, u_1\rangle \langle v_1, v_2\rangle \langle w_1, w_2\rangle = |\langle u_2, u_1\rangle \langle v_1, v_2\rangle \langle w_1, w_2\rangle| e^{i\alpha},$$

$$\langle X, Y\rangle \langle Y, Z\rangle \langle Z, X\rangle = \cos(p)\cos(q)\cos(r)e^{i\Omega}.$$

The invariants ω, Ω, and α are related as follows.

Theorem 4.2. *For any triangle ABC in $\mathbb{C}P^2$,*

1. $\Phi(A) = p$, $\Phi(B) = q$, $\Phi(C) = r$;
2. $\Omega = -\alpha - 2\omega + (2k+1)\pi$, $k \in \mathbb{Z}$.

Proof. A direct computation leads to the following formulae:

$$\langle u_1, u_2 \rangle = \frac{\cos(a)e^{i\omega} - \cos(b)\cos(c)}{\sin(b)\sin(c)}, \tag{14}$$

and

$$\langle Y, Z \rangle = -\frac{\cos(a)e^{-i\omega}\langle u_1, u_2 \rangle}{\langle z, y \rangle}. \tag{15}$$

By a cyclic permutation of a, b, c and x, y, z and X, Y, Z we obtain

$$\langle Z, X \rangle = -\frac{\cos(b)e^{-i\omega}\langle v_2, v_1 \rangle}{\langle x, z \rangle}, \tag{16}$$

$$\langle X, Y \rangle = -\frac{\cos(c)e^{-i\omega}\langle w_2, w_1 \rangle}{\langle y, x \rangle}. \tag{17}$$

Clearly from (15), (16), and (17) we arrive to $\cos\Phi(A) = \cos p$, $\cos\Phi(B) = \cos q$, $\cos\Phi(C) = \cos r$ and

$$\langle X, Y \rangle \langle Y, Z \rangle \langle Z, X \rangle = -e^{-2i\omega}\langle u_1, u_2 \rangle \langle v_2, v_1 \rangle \langle w_2, w_1 \rangle,$$

from which assertion 2 follows. ☐

All the trigonometric laws found by Blaschke and Terheggen, Rozenfeld, Shirokov, Hsiang and Brehm are the same. Of course, the second sine law was not found by Blashke and Terheggen although they knew that a triangle in $\mathbb{C}P^2$ depends on four invariants. The second sine law is self-contained in their investigation and it is derived as follows. With our notation the angles ϕ, χ and ψ of [2] are given by

$$\sin(\phi) = \frac{\cos(a)\sin(\omega)}{\cos(p)\sin(b)\sin(c)} = \frac{\sin(\Psi(A))}{\cos(p)}, \tag{18}$$

$$\sin(\chi) = \frac{\cos(b)\sin(\omega)}{\cos(q)\sin(c)\sin(a)} = \frac{\sin(\Psi(B))}{\cos(q)}, \tag{19}$$

$$\sin(\psi) = \frac{\cos(c)\sin(\omega)}{\cos(r)\sin(b)\sin(a)} = \frac{\sin(\Psi(C))}{\cos(r)}. \tag{20}$$

It follows from the equalitie of the left hand sides of (18), (19), and (20) that

$$\frac{\cos(p)\sin(\phi)}{\sin(2a)} = \frac{\cos(q)\sin(\chi)}{\sin(2b)} = \frac{\cos(r)\sin(\psi)}{\sin(2c)}. \tag{21}$$

Formula (21) is exactly the second sine law.

In [2] or from (14) we find the formula

$$\cos p = \left| \frac{\cos(b)\cos(c)e^{-i\omega} - \cos(a)}{\sin(b)\sin(c)} \right|.$$

By a direct computation, using $p = \Phi(A)$, (2) and (5) we arrive easily to the cosine law (10).

In [2] we find

$$\frac{\sin(p)}{\sin(a)} = \frac{\sin(q)}{\sin(b)} = \frac{\sin(r)}{\sin(c)}.$$

Clearly this is the sine law (8) since $p = \Phi(A)$, $q = \Phi(B)$ and $r = \Phi(C)$. We can then consider that all the laws were found by Blaschke and Terheggen. Item 1 of the above theorem is known in [15] in case of triangles degenerating in real projective planes. On the other hand, Terheggen showed that the work of Study on spherical trigonometry [17] is contained in the trigonometry of the complex projective line [15].

Remark 4.3. Let us turn to Hsiang's laws. Brehm's angular invariants $\Lambda(A)$, $\Phi(A)$, $\Psi(A)$ correspond to the angular invariants $\Lambda(A)$, $\frac{\pi}{2} - \mu(A)$, $\frac{\pi}{2} - \theta(A)$, respectively, which have been defined by Hsiang in [10]. Of course Hsiang derived the same laws of sines but instead of (10) or (13) he obtained a more complicated equivalent equation which we do not give here.

Shirokov took two angular invariants at each vertex: the Riemannian angle $\Lambda(A)$ between the two sides and the holomorphy inclination $\gamma(A)$. By comparing his three laws, we easily see that the holomorphy inclination is linked to Brehm's angles as follows:

$$\sin(\Lambda(A))\cos(\gamma(A)) = \sin(\Psi(A)).$$

The trigonometric laws found by Rozenfeld and Shirokov are the same as (8), (9), (10), or (13) where we replace $\sin(\Psi(A))$ by $\sin(\Lambda(A))\cos(\gamma(A))$ and $\sin(\Phi(A))$ by $\sin(\Lambda(A))\sin(\gamma(A))$ and by a cyclic permutation of the letters A, B, C. This can be found in [14], [16], and [13].

5 Concluding remarks

1. A slight modification of the definitions and computation yields similar results for triangles in complex hyperbolic spaces. The results can be found in [4].

2. It follows from the normal form that each triangle in $\mathbb{K}P^n$ ($n \geq 2$, $\mathbb{K} = \mathbb{C}, \mathbb{H}$) is contained in some totally geodesic $\mathbb{C}P^2$.

3. It is known from Wolf [19] that each isometry of a totally geodesic $\mathbb{C}P^2$ in $\mathrm{Cay}(P^2)$ can be extended to the Cayley plane. Thus, it is sufficient to regard triangles of the Cayley plane only in $\mathbb{C}P^2$. The results are also valid for triangles in $\mathrm{Cay}(H^2)$.

4. According to Wang [18], $\mathbb{K}P^n$, $\mathbb{K}H^n$, $\mathbb{K} = \mathbb{R}, \mathbb{C}, \mathbb{H}$, $\mathrm{Cay}(P^2)$, $\mathrm{Cay}(H^2)$ are besides the classical spaces all the two-point homogeneous spaces.

5. Let $([x], [y], [z])$ be a triangle in $\mathbb{C}P^2$ and let S be the surface generated by the geodesic arcs issued from $[x]$ with endpoints on the geodesic arc between $[y]$ and $[z]$. Let I be the integral of the Kähler form on the surface S. Hangan and Masala showed that $\omega = -2I + 2k\pi$, for some integer k, see [8].

6. The result of [8] made more clear a result obtained in [6]. the author constructed regular tetraedra of $\mathbb{C}P^3$ which have symmetry groups isomorphic to the symmetric group S_4. It turns out that a regular tetrahedron with edge $a < \arccos \frac{\sqrt{3}}{3}$ is necessarily contained in a real projective sub-space. Indeed, by regularity and Stokes' theorem, the integral I performed on its faces vanishes and we get from (3) and (6) that $\sin \Psi(A) = 0$. However, in [6], he proved that for any a such that $a \geq \arccos \frac{\sqrt{3}}{3}$ there exists a regular tetrahedron with edge a and with zero shape invariant. In this case, $2I = \frac{\pi}{2} + k\pi$. This means that the quadruple with the geodesic arcs between the vertices and the faces generated by the geodesic arcs issued from a vertex with endpoints on the opposite edge, constitute a simplicial complex of dimension 2 whose homology class of dimension 2 generates the homology group $H_2(\mathbb{C}P^n, \mathbb{Z})$ which is isomorphic to $\mathbb{Z}$ [9].

Trigonometry in general symmetric spaces is still at its infancy.

1. The trigonometry of symmetric spaces of rank 1 was studied by Hsiang in 1989 using integration of the geodesic flow [10]. A different approach using models in projective spaces has been given by Brehm in 1990 [4]. The trigonometry of a complex projective plane was first studied by Blaschke and Terheggen in 1939 [2], Terheggen in 1940 [15] and later pursued in 1957 by Rozenfeld [14] and independently by Sirokov in 1957 [16].

2. Aslaksen in 1989 [1] developed trigonometry for triangles in the Lie group SU(3) using algebraic invariants. Leuzinger in his Ph.D. thesis in 1990 [11] developed trigonometry for arbitrary Riemannian symmetric spaces of non-compact type using the general theory of symmetric spaces.

3. Hangan [9] wrote some trigonometric formulae for Grassmann manifolds.

4. The congruence criteria for triangles in Grassmann manifolds was stated by Masala in his Ph.D. thesis in 1996 using matrices of projections [12].

References

[1] H. Aslaksen, "Laws of trigonometry on SU(3)," *Trans. Amer. Math. Soc.* 317 (1990), no. 1, 127–142.
MR 0961593 Zbl 0695.53037 ☞ 422

[2] W. Blaschke and H. Terheggen, "Trigonometria Hermitiana," *Rend. Sem. Mat. Univ. Roma Ser.* 4 (3) (1939), 153–161.
☞ 413, 418, 420, 421, 422

[3] U. Brehm and B. Et-Taoui, "Congruence criteria for finite subsets of complex projective and complex hyperbolic spaces," *Manuscripta Math.* 96 (1998), no. 1, 81–95.
MR 1624356 Zbl 0908.51001 ☞ 413, 415, 416

[4] U. Brehm, "The shape invariant of triangles and trigonometry in two point homogeneous spaces," *Geom. Dedicata* 33 (1990), no. 1, 59–76.
MR 1042625 Zbl 0695.53038 ☞ 413, 415, 416, 417, 421, 422

[5] J. C. Coolidge, "Hermitian metrics," *Ann. Math* 22 (1921), 11–28.
☞ 418

[6] B. Et-Taoui, "Sur le tétraèdre régulier dans un espace projectif complexe," *Geom. Dedicata* 50 (1994), no. 3, 235–241.
MR 1286376 ☞ 422

[7] B. Et-Taoui, *Sur les systèmes finis de points réguliers dans les espaces projectifs et hyperboliques complexes,* Ph.D. thesis, Université de Mulhouse, Mulhouse, 1994.
☞ 413, 414, 416

[8] T. Hangan and G. Masala, "A geometrical interpretation of the shape invariant for geodesic triangles in complex projective spaces," *Geom. Dedicata* 49 (1994), no. 2, 129–134.
MR 1266268 Zbl 0805.53046 ☞ 422

[9] T. Hangan, "Laws of trigonometry in Grassmann manifolds," *Rend. Sem. Mat. Uni. Pol. Torino* 50 (1992), 91–104.
☞ 422, 423

[10] W.-Y. Hsiang, "On the laws of trigonometries of two-point homogeneous spaces," *Ann. Global Anal. Geom.* 7 (1989), no. 1, 29–45.
MR 1029843 Zbl 0695.53036 ☞ 414, 421, 422

[11] E. Leuzinger, *On the trigonometry of symmetric spaces,* Ph.D. thesis, ETH, Zürich, 1990.
☞ 422

[12] G. Masala, *Trigonométrie et polyèdres dans les variétés de Grassmann* $G_2(\mathbb{R}^n)$*,* Ph.D. thesis, Université de Mulhouse, Mulhouse, 1996.
☞ 423

[13] R. Ortega and M. Santander, "Trigonometry of complex Hermitian type homogeneous symmetric spaces," *J. Phys. A* 35 (2002), no. 37, 7877–7917.
MR 1945799 Zbl 1066.53101 ☞ 421

[14] B. A. Rozenfeld, "On the theory of symmetric spaces of rank one," *Mat. Sb. N.S.* 41(83) (1957), 373–380, in Russian. *Acad. R. P. Romîne. An. Romîno-Soviet. Ser. Mat.-Fiz.* (3) 12 (1958), no. 1 (24), 19–27, in Romanian.
MR 0096271 (Russian) MR 0096272 (Romanian) ☞ 414, 419, 421, 422

[15] H. Terheggen, "Zur analytishen Geometrie auf der Geraden von Hermite als Grenzfall der Geometrie in der Hermitishen Ebene und ihr Zusammenhang mit der gewohnlichen spharichen Trigonometrie," Jareshbericht Deutschen. Math. Verein 50 (1940), 24–35.
☞ 415, 418, 421, 422

[16] P. A. Shirokov, "On a certain type of symmetric spaces," *Mat. Sb. N.S.* 41(83) (1957) 361–372, in Russian. *Acad. R. P. Romîne. An. Romîno-Soviet. Ser. Mat.-Fiz.* (3) 12 (1958), no. 1 (24), 5–18, in Romanian.
MR 0096269 (Russian) MR 0096270 (Romanian) ☞ 418, 421, 422

[17] E. Study, *Sphärische Trigonometrie, orthogonale Substitutionen and elliptische Funktionen,* Abhandlungen der math. phys. Klassen der Akademie Leipzig 20, 1893.
☞ 421

[18] H. C. Wang, "Two-point homogeneous spaces," *Ann. of Math.* (2) 55 (1952). 177–191.
MR 0047345 Zbl 0048.40503 414, 422

[19] J. Wolf, “Elliptic spaces in Grassmann manifolds,” *Illinois J. Math.* 7 (1963), 447–462.
MR 0156295 Zbl 0114.37101 422

Pág. 241 (256) – Gauss Curvature

Fermat's principle of least time saying that the path take between two points by a light ray is the path which can be traversed in the least time. (60).

↳ Letter (January 1, 1662) to M. C. de la Chambre

Employing it, he deduced the law of refraction.

Hence shortest curves on a surface are regarded as something similar to paths traced by light rays (see Sect 6.12),

and it is reasonable to say that a quantity (or a concept) attached to a surface is "intrinsic" if it is derived from distance and angle measured by using shortest curves.

Pág. 337 (352) →

Olinde Rodrigues (1795-1851)

Rodrigues is known for a formula

also dissatisfied with the state of affairs surrounding metaphysics, in contrast to the scientific model cultivated in his days, Kant strove to lay his philosophical foundation of reason and judgement on secure grounds.

He was an enthusiast of Newtonian physics framed on Euclidean geometry, and asked himself, "Are space and time real existences? Are they only determinations or relations of things as Leibniz insists?" His inquiry brought him to the conclusion that Euclidean geometry is the inevitable necessity of thought and *inherent in nature* because our space (as an "absolute" entity) is a "sacrosanct" framework for all and any experience. He also held that space (as a "relative" entity) is the "subjective constitution of the mind" (see Sect. 6.22), He stressed further that mathematical propositions are formal descriptions of the a priori structures of space and time.[56]

In the work *Allgemeine Naturgeschichte und Theorie des Himmels* (1755), Kant discussed infiniteness of the universe. He believed that the universe is infinite, because God is the "infinite being," and creates the universe in proportion to his power (recall Bruno's view). In turn, he argues that it is not philosophically possible to decide whether the universe is infinite or finite; that is, we are incapable of perceiving so large distance, because the mind is finite.[57]

As Kant emphasized in the *Kritik der reinen Vernunft* (1781), abstract speculation must be pursued without losing touch with reality grasped by intuition; otherwise the outcome could be empty. This is not least the case for scientific knowledge. Yet intuition about our space turned out to be a clinging constraint on us; it was not easy to free ourselves from it.

6.7 Gauss: Intrinsic Description of the Universe

In Christianity, God is portrayed to be omnipresent, yet His whereabouts are unknown. Nonetheless, the laity usually personify God (if not a person literally), and believe that He "dwells" outside the universe (or *empyrean*). Imagining such a deity is not altogether extravagant from the view that we will take up in investigating a model of the universe though we do not posit the existence of God. What is available in place of God is, of course, *mathematics*.[58] Thus, we pose the question, "Is it mathematically possible to tell how the universe is curved without mentioning any outside of the universe?"

---

[56] Affected by the science in his time, Kant developed further the *nebular hypothesis* proposed by E. Swedenborg in 1734 (this hypothesis that the solar system condensed from a cloud of rotating gas was discussed later by Laplace in 1796).

[57] Hegel dismissed the claim that the universe extends infinitely.

[58] Robert Grosseteste (ca. 1175–1253) in medieval Oxford says, "mathematics is the most supreme of all sciences since every natural science ultimately depends on it" (*De Luce*, 1225). He maintained, by the way, that stars and planets in a set of nested spheres around the earth were formed by crystallization of matter after the birth of the universe in an explosion.

The answer is "Yes." To explain why, let us take a look at smooth *surfaces* as 2D toy models of the universe.[59] That is, we human beings are supposed to be confined in a surface. Here space in which the surface is located is thought of as the "outside," and is called the *ambient space*. A person in the outside plays the role of God incarnate. In our story, we allow him to exist for the time being; that is, we investigate surfaces in an *extrinsic* manner. At the final stage, we are disposed to remove him (and even the outside) from the scene, and to let the inhabitants in the surface find a way to understand his universe. In other words, what we shall do is, starting from an extrinsic study of surfaces, to look for some geometric concepts with which one can talk about how surfaces are curved without the assistance of the outside. Such concepts will be called *intrinsic*, one of the most vital key terms in this essay.

A few words about intrinsicness. Imagine that an inhabitant confined in a surface wants to examine the structure of his universe. The only method he holds is, like us, measuring the distance between two points and the angle between two directions. In our universe, we use the *light ray* to this end, which in a uniform medium propagates along shortest curves (straight lines) in space; a very special case of *Fermat's principle of least time* saying that the path taken between two points by a light ray is the path which can be traversed in the least time.[60] Hence shortest curves on a surface are regarded as something similar to paths traced by light rays (see Sect. 6.12), and it is reasonable to say that a quantity (or a concept) attached to a surface is "intrinsic" if it is derived from distance and angle measured by using shortest curves.

The progenitor of the intrinsic theory of surfaces is Johann Carl Friedrich Gauss (1777–1855). In 1828, he brought out the memoir *Disquisitiones generales circa superficies curvas*, where he formulated *bona fide* intrinsic curvature, which came to be called *Gaussian curvature* later.

Just for the comparison purpose, let us primarily look at the curvature of plane curves. Subsequent to the pioneering studies by Kepler (*Opera*, vol. 2, p. 175) and Huygens (1653–4), Newton used his calculus to compute curvature.[61] His approach is to compare a curve with *uniformly curved figures*; i.e., circles or straight lines. To be exact, given a point p on a smooth curve C, he singles out the circle (or the straight line) C_0 that has the highest possible order of contact with C at p, and then defines the curvature of C at p to be the reciprocal of the radius of C_0 (when C_0 is a straight line, the curvature is defined to be 0); thus the smaller the circle C_0 is, the more curved C is. To simplify his computation, take a coordinate system (x, y) such that the x-axis is the tangent line of C at p, and express C around p as the graph associated with a smooth function $y = f(x)$ with $f(0) = f'(0) = 0$ and $f''(0) \geq 0$. Compare $f(x) = \frac{1}{2}f''(0)x^2 + \frac{1}{3!}f'''(\theta x)x^3$ $(0 < \theta < 1)$ (Taylor's theorem) with the function $f_0(x) = R - \sqrt{R^2 - x^2} = \frac{1}{2}R^{-1}x^2 + \cdots$ corresponding to the circle C_0 of radius R, tangent to the x-axis at $x = 0$. Thus C_0 has the second order of contact with C at $(0, 0)$ if and only if $R = 1/f''(0)$. Therefore the curvature is equal to $f''(0)$.

---

[59] Throughout, "smooth" means "infinite differentiable."

[60] This principle was stated in a letter (January 1, 1662) to M. C. de la Chambre. Employing it, he deduced the *law of refraction*. The first discoverer of the law is Harriot (July 1601), but he died before publishing. W. Snellius rediscovered the law (1621). Descartes derived the law in *La Dioptrique* under a wrong setup; actually he believed that light's propagation was instantaneous. The finite speed of light was demonstrated by O. Rømer in 1676.

[61] *Tractus de methodis serierum et fluxionum* (1670–1671).

Remark 7.1 For an *oriented curve* C (a curve with a consistent direction defined along the curve), we select a coordinate system (x, y) such that the direction of x-axis coincides with that of C at p. Then $f''(0)$ is possibly negative. We call $\kappa(p) := f''(0)$ the *signed curvature* of C at p. If $c : [a, b] \longrightarrow \mathbb{R}^2$ is a parameterization of C with $\|\dot{c}\| \equiv 1$, then $\ddot{c}$ is perpendicular to $\dot{c}$, and $\ddot{c}(s) = \kappa(c(s))\boldsymbol{n}(s)$, where $\boldsymbol{n}(s)$ is the unit vector obtained by the 90° counterclockwise rotation of $\dot{c}(s)$. Writing $c(s) = (x(s), y(s))$, we have $\boldsymbol{n}(s) = (-\dot{y}(s), \dot{x}(s))$, and $\kappa(c(s)) = \langle \ddot{c}(s), \boldsymbol{n}(s)\rangle = \dot{x}(s)\ddot{y}(s) - \dot{y}(s)\ddot{x}(s)$. □

In the eighteenth century, Continental mathematicians polished calculus as an effectual instrument to handle diverse problems in sciences. Leonhard Euler (1707–1783) was among those who helped to brings calculus to completion and employed it as a tool for the extrinsic geometry of surfaces. His particular interest was the curvature of the curve obtained by cutting a surface with the plane spanned by a tangent line and the normal line. This study brought him to the notion of *principal curvature*.[62] Then in 1795, Gaspard Monge (1746–1818) published the book *Application de l'analyse à la géométrie*, an early influential work on differential geometry containing a refinement of Euler's work. However neither Euler nor Monge arrived at arrived at intrinsic curvature. Here came Gauss, who found how to formulate it with absolute confidence in its significance.

What comes to mind when defining the curvature of a surface S at p is the use of a coordinate system (x, y, z) with origin p such that the xy-plane is tangent to S at p,[63] with which S is locally expressed as a graph of a smooth function $f(x, y)$ defined around $(x, y) = (0, 0)$ with $f(0, 0) = f_x(0, 0) = f_y(0, 0) = 0$, where $f_x = \partial f/\partial x$ and $f_y = \partial f/\partial y$. Then we obtain $f(x, y) = \frac{1}{2}(ax^2 + 2bxy + cy^2) + r(x, y)$, where $a = f_{xx}(0, 0)$, $b = f_{xy}(0, 0)$, $c = f_{yy}(0, 0)$, and $r(x, y)$ is of higher degree than the second. The shape of the graph (and hence the shape of S around p) is roughly determined by the quantity $ac - b^2$; it is paraboloid-like (resp. hyperboloid-like) if $ac - b^2 > 0$ (resp. $ac - b^2 < 0$) (Fig. 6.7). Given all this, we define the curvature $K_S(p)$ to be $ac - b^2$, which is independent of the choice of a coordinate system. For example, $K_S \equiv R^{-2}$ for the sphere of radius R.

To account the exact meaning of intrinsicness of the Gaussian curvature, we consider another surface S', and let Φ be a one-to-one correspondence from S to S' preserving the distance of two points and the angle between two directions (such

---

[62] *Recherches sur la courbure des surfaces*, Mém. Acad. R. Sci. Berlin, **16** (1767), 119–143.

[63] Gauss's original definition uses the map (*Gauss map*) from S to the unit sphere which associates to each point on S its oriented unit normal vector (*Disquisitiones generales*, Art. 6).

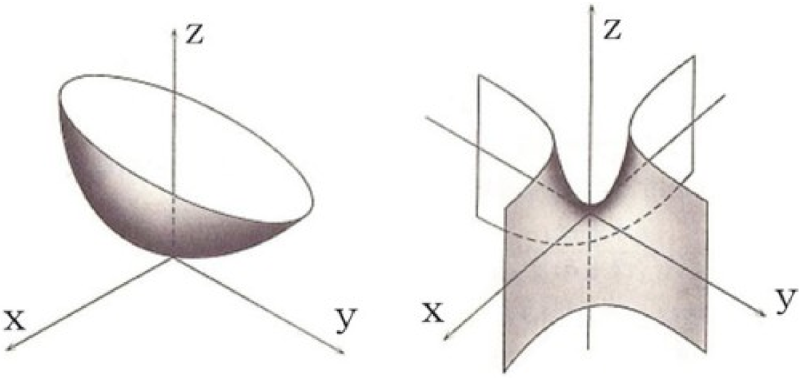

Fig. 6.7 Paraboloid and Hyperboloid

Φ is called an *isometry* in the present-day terms; see Sect. 6.20). If we would adopt the approach mentioned above on intrinsicness, Gauss's outcome could be rephrased as "$K_{S'}(\Phi(p)) = K_S(p)$." In particular, $K_S \equiv 0$ for a cylindrical surface S, since S, which is ostensibly curved, is deformed without distortion to a strip in the plane. This example articulately signifies that the Gaussian curvature could be possibly defined without the aid of the outside.

Executing lengthy computations, Gauss confirmed the intrinsicness of K_S (see Sect. 6.17).[64] Since the goal is more than a pleasing astonishment, Gauss named his outcome "Theorema Egregium (remarkable theorem)" (*Disquisitiones generales*, Art. 12). Actually, it is no exaggeration to say that this theorem is a breakthrough not only in geometry, but also in cosmology. The following is his own words (1825) unfolded in a letter to the P. A. Hansen:

This research is deeply entwined with much else, I would like to say, with metaphysics of space and I find it difficult to shake off the consequences of this, such as for instance the true metaphysics of negative or imaginary quantities (*Werke*, XII, p. 8. See [3]).[65]

Although the Theorema Egregium has a deep metaphysical nature as Gauss put it, his incipient motivation came from the practical activity pertaining to the *geodetic survey* of Hanover to link up with the existing Danish grid that started in 1818 and continued for 8 years. He traveled criss-cross in order to cover the whole country. Fueled by this arduous activity, he brought out many papers on the geometry of curves and surfaces from about 1820 onwards. The Theorema Egregium was a culmination of his intensive study in this field.

---

[64] Crucial in his argument is the commutativity of partial differentiation. Euler and A. C. Clairaut knew it, but the first correct proof was given by K. H. A. Schwarz in 1873.

[65] Even in the eighteenth century, there were a few who did not accept negative number. For example, F. Maseres said, "[negative numbers] darken the very whole doctrines of the equations and make dark of the things which are in their nature excessively obvious and simple" (1758). Although limited to the practical use, negative numbers already appeared in the *Jiuzhang Suanshu* (The Nine Chapters on the Mathematical Art) written in the period of the Han Dynasty (202 BCE–8 CE). The reason why the Chinese accepted negative numbers is, in all likelihood, that they had the concept of *yin* (negative or dark) and *yang* (positive or bright). Negative numbers were in use India at least from about the 7th century.

Remark 7.2 In her study of elasticity (1831), Sophie Germain introduced the *mean curvature* $H_S(p) := \frac{1}{2}(a + c)$, which appears in a characterization of *minimal surfaces*; that is, the surface area of S is locally minimized if and only if $H_S \equiv 0$. Putting $A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$, we have $K_S(p) = \det A$ and $H_S(p) = \frac{1}{2}\mathrm{tr}\, A$. The eigenvalues $\kappa_1(p)$, $\kappa_2(p)$ of A are called the *principal curvatures* of S. Obviously $K_S(p) = \kappa_1(p)\kappa_2(p)$ (*Disquisitiones generales*, Art. 8) and $H_S(p) = \frac{1}{2}(\kappa_1(p) + \kappa_2(p))$. □

6.8 Number Systems

It is off the subject, but while looking at the background behind Gauss's opinion about complex numbers quoted above, we shall prepare some material related to what we shall describe later (Sect. 6.21).

Gerolamo Cardano (1501–1576), who stood out above the rest at that time as an algebraist, acknowledged the existence of complex numbers in his *Ars Magna* (1545) that contains the first printed solutions to cubic and quartic equations. He observed that his formula may possibly involve complex numbers even when applied to a cubic equation possessing only real solutions.[66] Yet, he did not understand the nature of complex numbers and even thought that they are useless. Subsequently, Rafael Bombelli (1526–1572) took down the rules for multiplication of complex numbers (*binomio*) in his Book I of *L'Algebra*, while Descartes used the term "imaginary number" (*nombre imaginaire*) with the meaning of contempt. Wallis insisted that imaginary numbers are not unuseful and absurd when properly understood by using a geometric model just like negative numbers (*Algebra*, Vol. II, Chap. LXVI, 1673). Meanwhile, Euler persuaded himself that there is an advantage for the use of imaginary numbers (1751), and introduced the symbol i for the imaginary unit (1755). It was around 1740 that he found the earth-shaking formula $e^{\sqrt{-1}\theta} = \cos\theta + \sqrt{-1}\sin\theta$.

In 1799, Gauss gained his doctorate in absentia from the University of Helmstedt. The subject of his thesis is *the fundamental theorem of algebra*,[67] which, with a long history of inspiring many mathematicians, says that every polynomial equation of sand have led, by extension, to general theorems on mathematical optics [39, p. 131].

---

[66] For example, his formula applied to the cubic equation $x^3 - 15x - 4 (= (x-4)(x+2+\sqrt{3})(x+2-\sqrt{3})) = 0$ yields a solution $x = \sqrt[3]{2+\sqrt{-124}} + \sqrt[3]{2-\sqrt{-124}}$.

[67] *Demonstratio nova theorematis omnem functionem algebraicam rationalem integram unius variabilis in factores reales primi vel secundi gradus resolvi posse*. However, his proof had a gap. He gave three other proofs (1816, 1849) (see Sect. 6.21 for a "topological" proof).

Dupin later became a naval engineer, and a friend and a confident of Napoleon. He was a lover of modern Greece. He spent several years in the island of Corfu and he founded there, in 1808, together with Antoine-Marie Augoyat, a fellow student at the École polytechnique, an academy which became known as the Ionian Academy. He became a senator under Napoleon III.

To conclude with the work on Dupin, let us note that Poincaré, in 1847, published a paper titled *Démonstration nouvelle des propriétés de l'indicatrice d'une surface* (A new proof of properties, of the indicatrix of a surface) [98] in which he gives new proofs and complements to results of Dupin. Poincaré starts by giving a proof of the following result: If at each normal section at a given point on a surface we take, at the intersection with the tangent plane, a segment whose length is the square root of the radius of curvature of that section, then the result obtained is a conic. He then gives a new proof of Meusnier's theorem as well as a generalization, and a series of properties of the Dupin indicatrix, including results on conjugate tangents. Poincaré's proofs do not involve any calculation or equation, but only the synthetic theory of conics.

Olinde Rodrigues (1795–1851), who belonged to the circle of Monge,[53] wrote in 1815 and 1816 two memoirs on lines of curvature, *Sur quelques propriétés des intégrales doubles et des rayons de courbure des surfaces* (On some problems of double integrals and radii of curvature of surfaces) [104], and *Recherches sur la théorie des lignes et des rayons de courbure des surfaces, et sur la transformation d'une classe d'intégrales doubles qui ont un rapport direct avec les formules de cette théorie* (Researches on the theory of lines and of radii of curvature of surfaces, and on the transformation of a class of double integrals which have a direct relation with the formulas of this theory) [105]. Rodrigues is known for a formula on this subject carrying his name, and for developing the idea of a representation of a differentiable surface on a sphere of radius one which amounts to the Gauss map. He used such a representation to study the ratio of the area of (a piece of) the original surface to the area of its image on the sphere, arriving at a measure of curvature which in fact is the Gaussian curvature at a point of a surface.

7.7 Twentieth Century: Return to Euclid

The work of Busemann and Feller on which we shall comment below is strongly influenced by Hilbert's vision on mathematics. Both Busemann and Feller obtained their PhD in Göttingen, and although their thesis advisor was Courant, they were shaped by Hilbert's view on the foundations of geometry. Feller obtained his PhD in 1927, and Busemann in 1931. Before talking about their work, let us say a few words on some work of Hilbert related to our topic.

---

[53] Strictly speaking, Rodrigues was not a student of Monge, but according to Taton [112], Monge considered him as his student. Rodrigues became a follower of Saint-Simon (1760–1825), the famous utopian socialist whom he met in 1823, 2 years before Saint-Simon's death, and he became the main promoter of his ideas.

In 1932, together with Stephan Cohn-Vossen,[54] Hilbert published his remarkable book *Geometry and the Imagination*[55] [69] which is based on Hilbert's lectures held in 1920/21. Let us quote Hilbert from the preface, about the aim of this book:

> Here we want to look at geometry in its present condition from the point of view of the imagination. With imagination, we can approach the manifold geometric facts and questions, and moreover in many cases we can give an idea in imaginative form of the methods of investigation and proof leading to the discovery of the facts, without having to give all the details of the conceptual theories and the calculations.[56]

The point of view is indeed purely geometric, unlike the one of Euler or Monge and his students, which is based on analytic geometry. In a sense, it is a return to the methods of the Greeks. Let us quote from the chapter on differential geometry [69, p. 155] their description of the osculating circle at a point P_1 on a plane curve, a notion which we encountered several times in this essay. It says the following:

> We draw a circle through P_1 and two neighbouring points on the curve. If we let the two neighbouring points approach P_1, the corresponding circle converges to a limit position [...] This limit circle is called the osculating circle at P_1. Because of the construction described, it is usual to say that the curvature circle has three points running together in common with the curve. Similarly, one says that the tangent and the curve have two points running together in common.[57]

The osculating circle of a curve at a point P typically changes sides of the curve at the touching point P. If we consider all circles through P touching the curve at P, then in general those circles of radius greater than the radius of curvature r lie on one side of the curve whereas those of radius less than r lie on the other side of the curve. This is represented in Fig. 7.10, which is extracted from Cohn-Vossen's book (see [69, Figure 186]).

These ideas directly translate to space curves. To construct the osculating plane of a space curve at a point P, we can consider the limiting position of the

---

[54] For a comprehensive review on Cohn-Vossen's work, we refer the reader to the article [1] by Alexandrov.

[55] The original German title is: Anschauliche Geometrie.

[56] Wir wollen hier die Geometrie in ihrem gegenwärtigen Zustand von der Seite des Anschaulichen aus betrachten. An Hand der Anschauung können wir uns die mannigfachen geometrischen Tatsachen und Fragestellungen nahebringen, und darüber hinaus lassen sich in vielen Fällen auch die Untersuchungs- und Beweismethoden, die zur Erkenntnis der Tatsachen führen, in anschaulicher Form andeuten, ohne daß wir auf die Einzelheiten der begrifflichen Theorien und der Rechnung einzugehen brauchen.

[57] Wir legen einen Kreis durch P_1 und zwei benachbarte Punkte auf der Kurve. Wenn wir dann die beiden Nachbarpunkte auf der Kurve gegen P_1 rücken lassen, nähert sich der Kreis einer Grenzlage. [...] Man nennt diesen Kreis den Krümmungskreis der Kurve in P_1. [...] Der angegebenen Konstruktion wegen pflegt man zu sagen, daß der Krümmungskreis drei zusammenfallende Punkte mit der Kurve gemein hat. Ebenso sagt man, die Tangente hat mit der Kurve zwei zusammenfallende Punkte gemein.

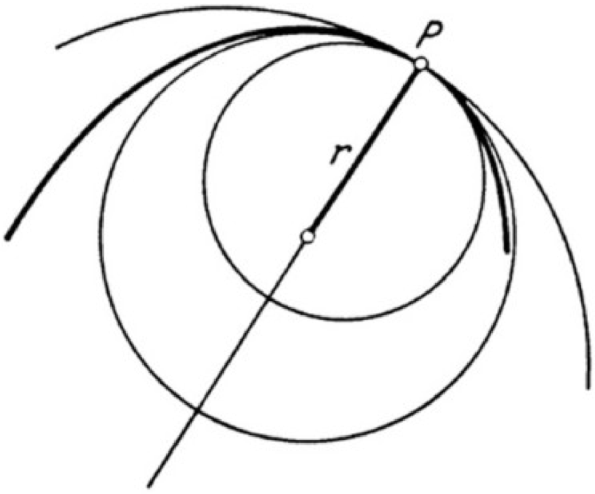

Fig. 7.10 The osculating circle at P separates two kinds of circles tangent to the curve at P: those that lie on one side of the curve and those that lie on the other side of the curve (in the neighbourhood of the point P). Figure from [69]

plane through the tangent at P and a neighbouring point that approaches P [69, p. 158]: "In the same sense as above, the osculating plane has three points that are running together in common with the curve."[58] Let us note incidentally that the term "osculating plane" can be traced back to Johann Bernoulli, specifically to his memoir *Problema: in superficie curva ducere lineam inter duo puncta brevissimam* (Problem: to draw on a curved surface a shortest line between two points) [13].[59] Bernoulli used this notion in the context of his study of geodesics on surfaces. In fact, he characterized geodesics as the curves on the surface whose osculating plane is always perpendicular to the tangent plane of the surface.

In [69], §32, Hilbert and Cohn-Vossen discuss eleven properties of the sphere, some of which characterize it and others do not. For example the third property says that the sphere is a surface of constant breadth and all the projections have constant circumference. The property of having constant breadth means that the distance between any two parallel tangent planes is constant. It is a generalization for surfaces of the property with the same name that we encountered in Euler's memoir on triangular curves [57]. As Hilbert and Cohn-Vossen explain, there are many other convex surfaces sharing this property. Figure 7.11 is extracted from Hilbert and Cohn-Vossen's book and reproduces pictures of such surfaces. In the

---

[58] Die Schmiegungsebene hat im früher erläuterten Sinn drei zusammenfallende Punkte mit der Kurve gemein.

[59] Bernoulli writes: Voco autem *planum osculans* quod transit per tria curvae quaesitae puncta infinite sibi invicem propinqua. (I also call an *osculating plane* one that passes through three points on the given curve that are infinitely close to each other).

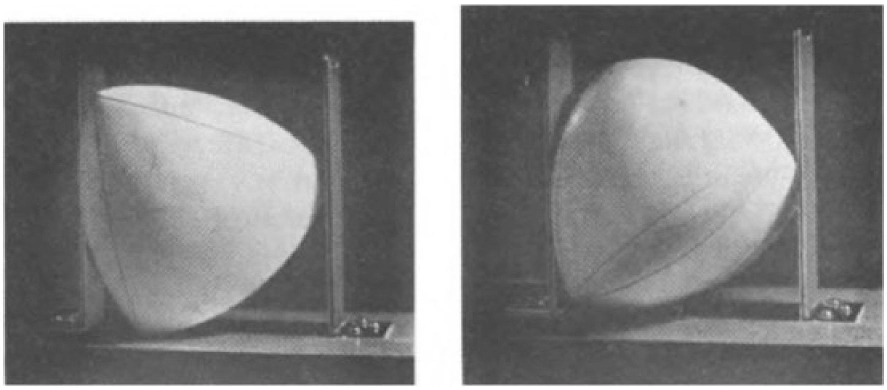

Fig. 7.11 Surfaces of constant breadth, from [69]

picture on the left hand side, one can see a surface having as a section a curve with three cusps (a triangular curve in the sense of Euler). The fourth property in §32 says that all points on a sphere are umbilical, which in fact characterizes the sphere. The fifth property concerns the focal points of a surface. Let us explain this notion. The set of centers of curvature of normal sections at a point of a surface is in general a segment contained in the normal section at that point, whose extremities are the centers of curvature of the principal directions. These extremities are called *focal points*, and we already saw that these two points coincide if and only if the given point is umbilical. The fifth property then states that the focal points of the sphere do not form a surface. Indeed the sphere is the only closed surface that has a unique focal point. But there are other surfaces where the set of focal points degenerates to a curve. These are the surfaces we mentioned, discovered by Dupin, called Dupin cyclides.

After Hilbert and Cohn-Vossen, and before going over to the work of Busemann and Feller, we briefly review the work of Bonnesen and Fenchel on convexity.

It is interesting to note that there was a strong interaction between the whole group of four mathematicians: Fenchel, Feller, Bonnesen and Busemann. Werner Fenchel obtained his PhD in 1928 in Berlin under Bieberbach. The title of his dissertation was *Über Krümmung und Windung geschlossener Raumkurven* (On curvature and torsion of closed space curves) [62]. The topic was differentiable curves in 3-space. The main results concern the image of the curve obtained by taking the tangent vectors of a given curve parametrized by arc length. This curve of tangents is considered as a curve lying on the unit sphere. Fenchel proves that (1) the total curvature of a space curve is bounded below by 2π, with equality if and only if the curve is planar and convex; (2) if the curve of tangents has a double point, then the torsion changes sign unless it is constant. The first of these two results is known as "Fenchel's theorem", and it was generalized in several ways by various people. The proofs of the results in Fenchel's thesis are very geometric in spirit. After his PhD, Fenchel moved to Göttingen, where he became Landau's assistant.

Busemann and Feller were also working in Göttingen, as young mathematicians. Fenchel remained there until 1933, when Hitler was appointed chancellor, with two interruptions between 1930 and 1932 during which he spent some time in Copenhagen, where Tommy Bonnesen was working. In 1933, Busemann, Feller and Fenchel fled to Denmark. Fenchel spent essentially the rest of his career there, whereas Busemann and Feller eventually moved to the US. The reader interested in this turbulent period of the history of Europe may refer to Feller and Busemann's biographies contained in their Selected Works editions, [61, Vol. 1] and [25, Vol. 1]. We also refer to Busemann's biography in [93].

The book [17] by Bonnesen and Fenchel, published in 1934, a year after Fenchel's move to Denmark, belongs to the Göttingen tradition of Hilbert's foundations of geometry. It is one of the founding books of modern convexity theory and, at the same time, of the synthetic point of view on curvature. §17 of the book starts with the following statement: "Far-reaching assertions can be made about the curvature relations of convex curves without making any assumption of differentiability" [17, p. 153]. Indeed, in that section and the next one, the authors attempt to develop a geometric theory of curvature of convex curves and surfaces without differentiability conditions.

Convex curves in the plane (respectively convex surfaces in 3-space) have support lines (respectively support planes) which sometimes can play the role of tangent lines (respectively planes). We recall that for any point on a convex hypersurface (that is, the boundary of a proper open convex set of maximal dimension in some Euclidean space), a support hyperplane is a hyperplane that contains that point but no point from the open convex subset it bounds. Any point on a convex surface is contained in a (not necessarily unique) support hyperplane. If the surface is differentiable at the given point, then the support hyperplane is unique and coincides with the tangent plane at that point.

Bonnesen and Fenchel call an *element* of a convex curve C in the plane a pair (P, s) consisting of a point P and a support line s at P. A *normal line* to this element is a line passing through P and perpendicular to s. An *osculating circle* at (P, s) is defined as a geometric limit of circles each of which touches s at P and which passes through a sequence of points on the curve that converges to P. The radius of such a circle is then a *radius of curvature* of the curve C at P. The values 0 and ∞ are allowed; in such cases the osculating circle is a point or a line respectively. To each element (P, s) is associated a set of osculating radii and of radii of curvature, each being an interval, the former contained in the latter. A *circle of curvature* at (P, s) is a limit of circles through P whose centers are the intersections of the normal of s at P with the normals of a sequence of elements that converge to (P, s). One also defines the *center of curvature* of C at P as the limit (if this limit exists) of the intersection of the normal line at p with the normal lines of a sequence of points on C that converge to P. The center of curvature, when it exists, may or may not coincide with the center of the osculating circle of C at P. For a given element (P, s), there may be several osculating circles and several circles of curvature. An osculating circle is a circle of curvature but the converse is not necessarily true. In the case where the osculating circle is unique, there exists also a unique circle of curvature, and both circles coincide. This occurs for almost all points of the convex curve, with respect to the Lebesgue measure of arc length on the curve; cf. Jessen [73]. In the cases where the limits do not exist (at a set of points of measure zero), one may define, by taking the smallest upper bound and the greatest lower bound, the upper and lower centers of curvature on the normal line n and then define the *upper* and *lower* curvature of the curve at the point P.

Bonnesen and Fenchel refer to the papers by Hjelmslev [71] and by Jessen [73] for these notions and their basic properties. They relate this notion of greatest lower bound of curvature radii to the question of the largest circle that can roll unhindered on a convex curve. This kind of questions remind us of questions that Aristotle addressed, and which we recalled in Sect. 7.2 of this paper. Bonnesen and Fenchel refer to papers by Bohr and Jessen [16], Mukhopadhyaya [90] and Brunn [20]. Then they discuss osculating conics to convex curves, referring to papers by Hayashi [65], Mukhopadhyaya [89] and others. In his paper *Circles incident on an oval of undefined curvature* [91] published in 1931, Mukhopadhyaya studied contact and osculation of circles relative to a convex continuous curve in the plane under the assumption that a definite tangent exists at each point but with no assumption on the existence of an osculating circle.

All this concerns curves. Regarding surfaces, Bonnesen and Fenchel state that "systematic investigations of the curvature properties of convex surfaces without differentiability assumptions are not available" [26, p. 154]. They mention preliminary tools and methods of attack due to Bouligand [19], Hadamard [64], Cohn-Vossen [29] and a few others.

In their article [26], Busemann and Feller investigate curvature properties of convex surfaces in Euclidean space. They describe their goal in the introduction thus:

> The aim of this paper is to develop the elements of differential geometry of convex surfaces without making the usual regularity assumptions. Hereby necessarily, purely geometric considerations often replace the usual analytic arguments.[60]

Consequently, they start from purely geometric definitions of tangent, tangent plane, osculating circle, radius of curvature etc. in the spirit of the book of Hilbert and Cohn-Vossen [69]. Their approach allows them to consider topological surfaces in 3-space with mild extra assumptions, and in particular convex surfaces, without further regularity conditions. Their starting point is a definition of tangent line, approximating circle, curvature and osculating plane, generalized from the setting of curves in space to that of sequences of points in space converging to a given point (Section 1 of [26]). At this point, the reader may stop and think about the meaning of these words, in the simplest situation, and without any differentiability assumption. Here is how Busemann and Feller proceed.

We start with a sequence of points $\{P_n\}$ in 3-space converging to a point P. If the lines joining P_n to P converge to a line t, this limit is called the *tangent line* of

---

[60] Cf. the English translation of the whole article in [25, Vol. 1, p. 235].

$\{P_n\}$. As in the treatise by Bonnesen and Fenchel, the pair (P, t) is said to be a *line element*. If the sequence $\{P_n\}$ has a tangent line t, then the circle in the plane (t, P_n) through P_n that is tangent to t is called the *approximating circle* of P_n with respect to the line element (P, t). If ρ_n denotes the radius of this circle and if the limit ρ of ρ_n exists, then this limit is called the *radius of curvature*, and its inverse $\frac{1}{\rho}$ the *curvature* of the sequence of points $\{P_n\}$. If the planes through P_n and t converge to a plane E, then this plane E is called the *osculating plane* of $\{P_n\}$.

Other metric notions associated with a sequence of points $\{P_n\}$ are introduced in the same paper, such as the *foot of the perpendicular*, *normal segment* and *tangent segment* from P_n to P. Formulae for curvature radii, etc. are given in terms of limits of normal segments, tangent segments, etc. generalizing the classical formulae given in terms of derivatives.

From this, Busemann and Feller introduce as follows the notions of tangent, curvature and osculating plane for a space curve:

Let κ be a continuous space curve and P a point on κ. If for any sequence P_n of points on κ converging to P the tangent and curvature exist and are independent of the choice of P_n, then these tangent and curvature are defined to be the *tangent* and *curvature of* κ at P. The osculating plane of κ at P is defined similarly.

Several other metric definitions in the same spirit are given. In particular, if the tangent and the osculating plane exist but the curvature depends on the chosen sequence, then the upper limit of all the possible curvatures is called the *upper curvature of* κ. Choosing the sequence P_n to be on the same side of P on the curve, one obtains notions of right-sided and left-sided tangent and curvature.

Busemann and Feller compare these definitions with the classical ones by taking derivatives in the special case of plane curves and they show that their concept is broader. Namely, if a curve κ is given as the graph of a continuous function f of one variable x and if at some point x_0 of the interval of definition the right second derivative of f exists, then the usual notion of right curvature of κ at x_0 coincides with the one they introduce. They show that the converse is not true: there exists a C^1-function f with right curvature at x_0 (in the sense of the authors) admitting no right second derivative at x_0.

The case of convex plane curves is of particular interest. Here, at each point of such a curve, a left and a right tangent exists. Because the slope of these one-sided tangents is monotone with respect to the point, the curve has a two-sided tangent almost everywhere (up to countably many exceptions). This monotonicity property also implies that up to a set of measure zero, f has also almost everywhere a second derivative.

In the same trend, the authors give a geometric definition of a strict tangent plane at a point of a surface [26, § 2]: given a point P on a surface Φ in the Euclidean space, a plane Π is called *strict tangent plane* of Φ at P if for every sequence of secants Q_nP_n of Φ converging to P, all the limits of the lines joining Q_n and P_n are contained in Π. Again, the case of a convex surface is a typical example: any tangent plane is automatically strict. Busemann and Feller indicate that this notion of strict tangent plane coincides with that of *paratangent* introduced by Bouligand [19].

Using these definitions, Busemann and Feller present generalizations of classical results on the differential geometry of surfaces. In particular, Euler's theorem on normal curvatures which we surveyed in Sect. 7.4 and Meusnier's formula which we mentioned in Sect. 7.6 serve as testing ground to demonstrate that the general setting is reasonable enough, that is, non-trivial results may be proven in it. More precisely, the general form of Meusnier's theorem, which holds also for non-convex surfaces, is the following: If at a point with a strict tangent plane some oblique plane section through a tangent t has curvature, then all plane sections through t have curvature as well, and the corresponding osculating circles lie on a common sphere. In particular, for a convex surface, Meusnier's formula holds almost everywhere.

For convex surfaces in 3-space, Busemann and Feller show that Euler's theorem on normal curvatures holds almost everywhere. They formulate their result in terms of Dupin's indicatrix. We already recalled in Sect. 7.6 that Dupin observed that Euler's equation relating to the normal curvatures at a point of an analytic surface define a conic (called Dupin indicatrix) in the tangent plane at the given point. In the case of a convex surface, the indicatrix can either be an ellipse or a pair of parallel lines unless all normal curvatures at the point vanish.

Busemann and Feller use the following construction as a definition of the indicatrix: Consider the tangent plane T at a point P of a convex surface where such a plane exists, and choose coordinates such that this tangent plane is not vertical and P is its origin. Take the plane parallel to T and lying above T, at distance ϵ. It intersects the surface in a curve, call it c_ϵ. Then, for almost every P, the curve c_ϵ rescaled by the factor $\frac{1}{\epsilon}$ converges, as $\epsilon \to 0$, to a curve, which the authors call the indicatrix at that point. Moreover, they show that it is either an ellipse or a pair of parallel lines. This means that for a convex surface, at almost all points normal curvatures exist and satisfy Euler's theorem. They then show that Rodrigues' formula for principal curvature directions holds in this general setting. They also obtain an umbilical theorem. We recall that in its classical form, the theorem says that an analytic surface in 3-space in which all points are umbilics (i.e. the maximal and minimal normal curvatures coincide) is either a piece of a sphere or a piece of a plane. In particular, this property characterizes the sphere among all convex closed analytic surfaces in 3-space (see also [69], §32, Property 4). Busemann and Feller generalize this result by proving that a differentiable convex surface where almost all points are umbilics and the upper normal curvatures are bounded is either a piece of a sphere or a plane. They also show, through a counterexample, that the extra condition of boundedness of the upper normal curvatures is essential. More precisely, they construct an example of a closed differentiable convex surface where almost all points are umbilics with the same curvature c and that is neither a piece of a sphere nor a plane. This example is obtained by a subtle algorithm starting with the unit ball and stepwise replacing parts by caps of curvature c and smoothing out the edges with canal surfaces of curvature greater than c.

In the last section of their paper, Busemann and Feller study geodesics on convex surfaces. They prove that a geodesic passing through (or emanating from) any point at which a tangent plane exists is in fact differentiable. Moreover, such a geodesic has a tangent line (or tangent ray) at that point, and an osculating plane which is perpendicular to the tangent plane. This generalizes a result of Johann Bernoulli which we already mentioned, from his article published in his 1728, *Problema: in superficie curva ducere lineam inter duo puncta brevissimam* [13] in which he used the term "osculating plane" for the first time.

For the proofs of all these theorems, Busemann and Feller acknowledge the strong influence of Bonnesen and Fenchel's *Theorie der konvexen Körper* [17] (1934), and this influence is also stressed in Busemann's monograph *Convex surfaces* [24] published in 1958. In the first chapter of that monograph, Busemann gives more detailed proofs of all these theorems. In the introduction, he writes that the first two chapters form a complement to Bonnesen and Fenchel's book, dealing in particular with subjects conjectured or suggested there. Zalgaller, in his paper [114], writes that Busemann took part in the preparation of the book by Bonnesen and Fenchel.

In the introduction of his monograph *Convex surfaces*, Busemann declares that his purpose is to present a subject, convex surfaces, which "during the past 25 years has experienced a striking and beautiful development, principally in Russia, but has remained largely unknown, at least in the USA." Indeed, the book is a tribute to the work of A. D. Alexandrov. As a matter of fact, this is the way Busemann presented his book project to the publisher.[61] We are not going to attempt any description of the results from Alexandrov or the school he founded, because this would lead us too far, but since our main subject matter in this article is curvature, we shall review his notion of nonpositive curvature.

We start with the definition of the length of a curve γ in a metric space (X, d), that is, of a continuous path $\gamma : [0, 1] \to X$. This is defined as

$$L(\gamma) = \sup \sum_i d(\gamma(t_i), \gamma(t_{i-1}))$$

where the supremum is taken over the set of subdivisions $0 = t_0 \leq t_1 \leq \ldots \leq t_n = 1$ of the interval $[0, 1]$. The definition is reminiscent of the notion of length of a curve as defined by Archimedes in his *On the sphere and the cylinder* [66]. If $\gamma(0) = A$ and $\gamma(1) = B$, we say that the curve joins A and B. We also say that the curve starts at A.

A metric on a surface (2-dimensional manifold) M is said to be intrinsic if the distance between two points is equal to the infimum of the lengths of curves in that surface that joins them.

A geodesic in M joining two points is a curve joining them which is the isometric image of a Euclidean interval.

---

[61] A letter written on July 23, 1956, from the editor-in-chief of Interscience Publishers Inc. (with copies to Courant, Stocker and Bers) starts with: "It is with very special pleasure indeed that we learn from Dr. Stocker of your willingness to preset in our Tracts series an account of important recent development of the work of certain Russian geometers." (A copy of the letter is kept with Courant's correspondence at the Elmer Holmes Bobst Library in New York University).

A triangle Δ in M is the union of three points in M (called the vertices of Δ) together with three minimizing geodesics (called the sides of Δ) joining them.

To each triangle Δ in M, one can associate a triangle Δ_E in the Euclidean plane E, together with a natural marking between the sides of Δ and the sides of Δ_E. The triangle Δ_E is called a comparison triangle of Δ. Now consider two curves γ_1 and γ_2 in S starting at a common point p. We wish to define the angle these curves make at p. We take a sequence of points x_n and y_n on the images of γ_1 and γ_2 respectively, both converging to p.

We then define the upper angle between γ_1 and γ_2 at p as the upper limit of the sequence of angles of the comparison triangles of p, x_n, y_n, all the angles being taken at the image of the point p in the Euclidean plane.

With this definition, each triangle in the metric surface M has a well-defined upper angle at each vertex. Let S be the sum of these three upper angles. The *angle excess* of such a triangle is then equal to $S - \pi$.

The surface M with an intrinsic metric is said to have (local) non-positive curvature in the sense of Alexandrov if (locally) each triangle has non-positive angle excess.

There are straightforward variations on this definition that lead to that of surfaces of bounded curvature (from below and from above), where, instead of taking comparison triangles in the Euclidean plane, one takes comparison triangles in one of the constant curvature geometries. Alexandrov thoroughly investigated this notion in several books and articles, see e.g. [2] and [3].

In his monograph [23] (and in other books and papers), Busemann introduced other (close but different) metric notions of curvature, and we recall the definition. He starts with the notion of a metric space *midpoint convex*, that is, for any two points x and y in that space, there exists a point z whose distances from x and from y are both equal to half the distance between x and y. Such a point z is called a *midpoint* of x and y (it is not unique). In a midpoint convex geodesic metric space which is complete, any two points may be joined by a geodesic.

Busemann defines a midpoint convex metric space (M, d) to be *non-positively curved* if for any three points a, b, c in M, if a' is a midpoint of a and c and b' a midpoint of b and c, then

$$d(a', b') \leq \frac{1}{2} d(a, b).$$

There are variations of this definition, namely, the space is called *negatively curved, non-negatively curved, or positively curved* depending on whether the above inequality is replaced by a strict inequality, or the reverse inequality or strict reverse inequality respectively. If the inequality is replaced by an equality, the space is said to be *zero curved*. These notions are used today under the name *Busemann non-positive (resp. non-negative, positive, negative, zero) curvature*.

In the same monograph, Busemann studied other notions of metric curvature, and we would like to mention three of them. The first one uses the notion of a *peakless function*. This is a continuous function f defined on an interval I of the real line such that for all $t_1 < t_2 < t_3$ in I we have

$$f(t_2) \leq \max(f(t_1), f(t_3))$$

with equality if and only if $f(t_1) = f(t_2)$. Using this, Busemann introduced the notion of *space whose distance function is peakless*, another generalization of nonpositive curvature.

Another notion of curvature, introduced by Busemann, uses the notion of *capsules* in metric spaces: If T is a segment in a metric space (M, d) and ρ a nonnegative number, a subset of the form

$$\{x \in M \mid d(x, T) \leq \rho\}$$

is said to be a *capsule* of axis T and radius ρ.

Busemann proves a local-implies-global result for convexity of capsules [23, Theorem 36.20], and he proves that a space which has Busemann non-positive (respectively negative) curvature has convex (respectively strictly convex) capsules. [23, Theorem 36.20]

In the same book, Busemann introduced a notion of curvature in terms of angle excess of triangles in spaces he calls G-spaces (G-stands for "geodesic"). These are metric spaces that are finitely compact (every bounded infinite sequence contains a convergent subsequence) and that satisfy a property of local uniqueness of extension of geodesics.

The notion of G-space, like many of the other metrical notions introduced by Busemann, is already contained in his early book *Metric Methods in Finsler Spaces and in the Foundations of Geometry* [21] (1942) and in his paper *On spaces in which points determine a geodesic* [22] (1943).

The metric notions of curvature introduced by Alexandrov and by Busemann turned out to be the bases of whole new fields in geometry, called after their names[62] and there is a multitude of theorems and research problems that rely on these definitions.

Like Busemann, Fenchel and some other twentieth-century geometers we mentioned, Alexandrov's interest abides in the most basic notions of geometry, putting isoperimetry, isoepiphany and other problems of classical mathematics at the forefront of research, with a dislike for a certain Riemannian geometry based on linear algebra and tensor calculus. We refer the reader to Alexandrov's book *Intrinsic geometry of convex surfaces* [2], in which he provided proofs of the theorems of Euler, Meusnier and Rodrigues in the general setting of convex surfaces, attributing these generalizations to Busemann and Feller. He proved in the same book [2] that any convex two-dimensional manifold is non-negatively curved in the sense he introduced it and which we recalled above, and that every non-negatively curved two-dimensional manifold homeomorphic to the sphere is isometric to some convex surface.

---

[62] There are AMS Mathematics Subject Classification codes that carry the names of Alexandrov and Busemann geometries.

As another link between the group of geometers we mentioned, let us note that in his treatise *Convex surfaces*, Busemann gives a proof of a theorem in convexity theory that was independently obtained by himself and Fenchel and Jessen [63] stating that if two convex bodies in Euclidean n-space have equal p-th order area measure (we shall not go into the definition here) for some p satisfying $1 \leq p \leq n-1$, then the two bodies are the same up to a translation.

In a tribute in memory of Alexandrov, S.S. Kutateladze, who belongs to Alexandrov's school, writes that Alexandrov's slogan in mathematics was "Retreat to Euclid" and that he used to declare that "the pathos of contemporary mathematics is the return to Ancient Greece" [76].

A geometry freed from the dominion of differential calculus kept growing since the last quarter of the twentieth century, reinvigorated in the works of Thurston whose methods constitute a return to the synthetic methods of non-Euclidean geometry, and of Gromov, closer in spirit to the Busemann and Alexandrov approaches.

References

1. A. D. Alexandrov, On the works of S. E. Cohn-Vossen, *Uspehi Matem. Nauk (N.S.)* 2, (1947), no. 3 (19), pp. 107–14. English translation by A. Iacob.
2. A. D. Alexandrov, *Intrinsic geometry of convex surfaces*, ed. S. S. Kutateladze, Y. G. Reshetnyak, transl. S. S. Kutateladze, CRC Press, 2005.
3. A. D. Alexandrov and V. A. Zalgaller, *Intrinsic geometry of surfaces*, Academy Sc. USSR, Moscow, 1962. English translation by J. M. Danskin, *AMS Translations of Mathematical Monographs*, American Mathematical Society, Providence, 1967.
4. Apollonius of Perga, *Conics, Books V to VII*, ed. G. J. Toomer, Springer-Verlag, 1990.
5. Apollonius of Perga, *Les coniques d'Apollonius de Perge* (P. Ver Eecke, ed.), Librairie Albert Blanchard, Paris, 1963.
6. Apollonius of Perga: Apollonii Pergaei Conicorum lib. V. VI. VII. paraphraste abalphato asphahanensi nunc primum editi. Additus in calce Archimedis assumptorum, ex codocibus arabicis manuscriptis. Serenissimi Magni Ducis Etruriae Abrahamus Ecchellensis Maronita. In alma Vrbe Linguar. Orient. Professor Latinos reddidit. Io. Alfonsus Borellus In Pisana Academia Matheseos Professor curam in Geometricis versioni contulit, & notas vberiores in vniuersum opus adiecit, Firenze, 1661.
7. P. Appell, Le problème géométrique des déblais et des remblais, Mémorial des sciences mathématiques, fascicule 27, Gauthier-Villars, Paris, 1928.
8. Aristotle, Metaphysics, transl. W. D. Ross, In: *The works of Aristotle translated into English*, ed. W. D. Ross and J. A. Smith, Clarendon Press, Oxford, 1908, 2nd. ed. 1928.
9. Aristotle, On the Heavens, transl. J. L. Stocks, In: *The works of Aristotle translated into English*, ed. W. D. Ross and J. A. Smith, Clarendon Press, Oxford, 1922.
10. Aristotle, On the Soul, transl. J. A. Smith, In: *The works of Aristotle translated into English*, ed. W. D. Ross and J. A. Smith, Clarendon Press, Oxford, 1927.
11. Aristotle, *Minor Works*, transl. W. S. Hett, Harvard University Press, Cambridge/London: William Heinemann, London, 1936.
12. M. Berger, *Geometry revealed: A Jacob's ladder to modern higher geometry*, Springer, Berlin, 2010.

13. J. Bernoulli, Problema: in superficie curva ducere lineam inter duo puncta brevissimam, Opera Omnia IV (reed. Georg Olms 1968), (1728), pp. 108–128.
14. J. Bertrand, Mémoire sur la théorie des courbes à double courbure, *Journal de mathématiques pures et appliquées 1re série*, tome 15 (1850), pp. 332–350.
15. W. Blaschke and G. Hessenberg, Lehrsätze über konvexe Körper. *Jber. Deutsch. Math-Vereinig*. Vol. 26 (1917), pp. 215–220.
16. H. Bohr and B. Jessen, Om Sandsynlighedsfordelinger ved Addition af konvexe Kurver. *Danske. Vid. Selsk. Skr.* (8) Vol. 12 (1929), pp. 326–405.
17. T. Bonnesen and W. Fenchel, *Theorie der konvexen Körper*, Springer, Berlin, 1934, English translation: Theory of convex bodies, BCS associates, Moscow, Idaho, 1987.
18. H J. M. Bos, Tractional motion and the legitimation of transcendental curves, *Centaurus* 31(1) (1988), pp. 9–62.
19. G. Bouligand, *Introduction à la géométrie infinitésimale directe (avec une préface d'E. Cartan)*, Vuibert, Paris, 1932.
20. H. Brunn, Über Ovale und Eiflächen, Inaugural Dissertation, München, 1887, 42 p.
21. H. Busemann, *Metric Methods in Finsler Spaces and in the Foundations of Geometry.* Annals of Mathematics Studies, no. 8. Princeton University Press, Princeton, N. J., 1942.
22. H. Busemann, On spaces in which points determine a geodesic. *Trans. Am. Math. Soc.* 54 (1943), 171–184.
23. H. Busemann, *The geometry of geodesics*, Academic Press (1955), reprinted by Dover in 2005.
24. H. Busemann, *Convex surfaces*, Interscience tracts in pure and applied mathematics, Interscience Publishers, New York and London, 1958.
25. H. Busemann, *Selected Works*, 2 volumes, ed. A. Papadopoulos, Springer International, 2018.
26. H. Busemann and W. Feller, Krümmungseigenschaften konvexer Flächen, *Acta Math.* 66 (1936), pp. 1–47. English translation by A. A'Campo Neuen: Curvature Properties of Convex Surfaces, In: Herbert Busemann, Selected works, Vol. II, Springer International (2018), pp. 235–270.
27. M. Chasles, Aperçu historique sur l'origine et le développement des méthodes en géométrie. Mémoires couronnés par l'Académie royale des sciences et belles-lettres de Bruxelles, tome XI, Hayez, Bruxelles, 1837.
28. N. G. Chebotarev, On the breadth of contours and bodies (Russian), *Ber. Wiss. Forschgsinst. Odessa* Vol. 1 (1924), pp. 29–41.
29. S. Cohn-Vossen, Singularitäten konvexer Flächen. *Math. Ann.* Vol. 97 (1927), pp. 377–386.
30. A.-C. Clairaut, *Recherches sur les courbes à double courbure*, Nyon, Didot et Quillau, Paris, 1731.
31. A. Clairaut, *Théorie de la figure de la terre, tirée des principes de l'hydrostatique*, David Fils, Paris, 1743.
32. N. de Condorcet, Éloge d'Huyghens, In : Œuvres de Condorcet, publiées par A. Condorcet O'Connor et M. F. Arago, tome II, Firmin Didot Frères, Paris (1847–1849), pp. 54–72.
33. J. L. Coolidge, The unsatisfactory story of curvature, *American Mathematical Monthly* 59 (1952), pp. 375–379.
34. G. Darboux, Notice historique sur le général Meusnier, membre de l'Ancienne Académie des Sciences, lue dans la séance publique annuelle du 20 décembre 1909, Paris, Gauthier-Villars, 1910, and Mém. Acad. Sciences, Vol. 51, 2e série (1910), pp. 1–40.
35. J. Dieudonné, *History of algebraic geometry*, Wadsworth Inc., Monterey, California, 1985.
36. Ch. Dupin, Le problème géométrique des déblais et remblais, Correspondance de l'École polytechnique, VIIe cahier (1807), pp. 218–225.
37. Ch. Dupin, Développements de géométrie, avec des applications à la stabilité des vaisseaux, aux déblais et remblais, au défilement, à l'optique, etc. Paris, Coursier, 1813.
38. Ch. Dupin, Exercices de géométrie et d'algèbre, 1818.
39. Ch. Dupin, *Essai historique sur les services et les travaux scientifiques de Gaspard Monge*, Bachelier, Paris, 1819.
40. Ch. Dupin, *Applications de la géométrie à la mécanique*, Bachelier, Paris, 1822.

41. Euclid (Th. L. Heath, ed.) *The thirteen books of Euclid's Elements*, 3 volumes, Cambridge University Press, 1908, Reprint, Dover.
42. L. Euler, Constructio linearum isochronarum in medio quocunque resistente. *Acta Eruditorum* (1726), pp. 361–363, Opera Omnia: Series 2, Volume 6, pp. 1–3. English translation by I. Bruce available on the internet.
43. L. Euler, Meditationes super problemate nautico, quod illustrissima regia Parisiensis Academia scientarum proposuit, Pièce qui a remporté le prix de l'Academie Royale des sciences, 1727 (1728), pp. 1–48, Opera Omnia, Series 2, Volume 20, pp. 1–35.
44. E. Euler, De linea brevissima in superficie quacunque duo quaelibet puncta jungente, *Commentarii academiae scientiarum Petropolitanae* 3 (1732), 110–124, Opera Omnia, Series 1, Volume 25, pp. 1–12.
45. L. Euler, De innumerabilibus curvis tautochronis in vacuo, *Commentarii academiae scientiarum Petropolitanae* 4 (1735), pp. 49–67, Opera Omnia: Series 2, Volume 6, pp. 15–31. English translation by I. Bruce available on the internet.
46. L. Euler, Curva tautochrona in fluido resistentiam faciente secundum quadrata celeritatum, *Commentarii academiae scientiarum Petropolitanae* 4 (1735), pp. 67–89, Opera Omnia: Series 2, Volume 6, pp. 32–50. English translation by I. Bruce available on the internet.
47. L. Euler, *Mechanica*, 2 volumes, published as a book in 1736, Opera Omnia: Series 2, Volumes 1 and 2.
48. L. Euler, De linea celerrimi descensus in medio quocunque resistente, *Commentarii academiae scientiarum Petropolitanae* 7 (1740), pp. 135–149, Opera Omnia: Series 1, Volume 25, pp. 41–53
49. L. Euler, De fractionibus continuis dissertatio Commentarii academiae scientiarum Petropolitanae 9 (1744), pp. 98–137, Opera Omnia: Series 1, Volume 14, pp. 187–216.
50. L. Euler, Methodus inveniendi lineas curvas maximi minimive proprietate gaudentes, sive solutio problematis isoperimetrici lattissimo sensu accepti, Lausanne and Geneva, 1744, Opera Omnia, Series 1, vol. 24
51. L. Euler, Introductio in analysin infinitorum, First edition: Lausannae: Apud Marcum-Michaelem Bousquet & socios., 1748. Opera omnia, Series 1, Vol. VIII. English translation by J. T. Blanton, 2 vol., Springer-Verlag, New York, 1988, 1990.
52. L. Euler, *Introduction to Analysis of the Infinite*, English translation of [51] by Blanton, 2 vol., Springer-Verlag, New York, 1988.
53. L. Euler, *Scientia navalis seu tractatus de construendis ac dirigendis navibus*, 2 volumes, Saint Petersburg, 1749. Opera Omnia, Series 2, Vol. 18 and 19, Zürich and Basel, 1967 and 1972.
54. L. Euler, Investigatio curvarum quae evolutae sui similes producunt, *Commentarii academiae scientiarum Petropolitanae* 12 (1750), pp. 3–52, Opera Omnia: Series 1, Volume 27, pp. 130–180.
55. L. Euler, Recherches sur la courbure des surfaces, *Mémoires de l'académie des sciences de Berlin* 16 (1767), pp. 119–143, Opera Omnia, Series 1, Volume 28, pp. 1–22.
56. L. Euler, De solidis quorum superficiem in planum explicare licet, *Novi Commentarii academiae scientiarum Petropolitanae* 16 (1772), pp. 3–34. Opera Omnia Series 1, Volume 28, pp. 161–186.
57. L. Euler, De curvis triangularibus, *Acta Academiae Scientarum Imperialis Petropolitinae* (1778), pp. 3–30, Opera Omnia, Series 1, Volume 28, pp. 298–321.
58. L. Euler, Methodus facilis omnia symptomata linearum curvarum non in eodem plano sitarum investigandi, *Acta Academiae Scientarum Imperialis Petropolitinae* (1782), pp. 19–57, Opera Omnia, Series 1, Volume 28, pp. 348–381.
59. L. Euler, De plurimis quantitatibus transcendentibus, quas nullo modo per formulas integrales exprimere licet, *Acta Academiae Scientarum Imperialis Petropolitinae* 4, 1784, pp. 31–37, Opera Omnia: Series 1, Volume 15, pp. 522–527.
60. L. Euler, Investigatio curvarum quae similes sint suis evolutis vel primis vel secundis vel tertiis vel adeo ordinis cuiuscunque, *Nova Acta Academiae Scientarum Imperialis Petropolitinae* 1 (1787), pp. 75–116 Opera Omnia: Series 1, Volume 29, pp. 72–111.

61. W. Feller, *Selected Works*, 2 volumes, ed. R. L. Schilling Z. Vondraček and W. A. Woyczyński, Springer Verlag International, 2015.
62. W. Fenchel, Über Krümmung und Windung geschlossener Raumkurven, *Math. Ann.*, 101 (1929), pp. 238–252.
63. W. Fenchel and B. Jessen, Mengenfunktionen und konvexe Körper, *Danske Vid. Selskab Mat.-Fys.Medd.* 16, 3 (1938), pp. 1–31.
64. J. Hadamard, Sur certaines propriétés des trajectoires en dynamique. *J. Math. Pures et Appl.* 5 (1897), pp. 331–387.
65. T. Hayashi, The least number of the sextatic points of a central oval, *Sci. Rep. Tohoku Uni.* Vol. 12 (1924) pp. 393–395.
66. L. Heath (ed.) *The works of Archimedes*, Cambridge University Press, 1897, Reprint, Dover, 2002.
67. T. L. Heath, *A history of Greek Mathematics*, Cambridge University Press, Cambridge, 1921, reprint, Dover.
68. Heron of Alexandria, Metrica. In: *Codex Constantinopolitanus Palatii Veteris*, ed. E. M. Bruins, Brill, 1964.
69. D. Hilbert and S. Cohn-Vossen, *Anschauliche Geometrie*, Springer, Berlin and Heidelberg, 1932.
70. Ph. de La Hire, Gnomonique ou l'art de tracer des cadrans ou horloges solaires sur toutes les surfaces, par différentes pratiques, avec les démonstrations géométriques de toutes les opérations, Paris Estienne Michallet, 1682.
71. J. Hjelmslev, Über die Grundlagen der kinematischen Geometrie, *Acta Math.* 47 (1926), pp. 143–188.
72. C. Huygens, *Œuvres complètes*, 22 volumes, Société Hollandaise des Sciences, La Haye, 1888–1950.
73. B. Jessen, Om konvexe Kurvers Krumning. *Mat. Tidssr. B* (1929) pp. 50–62.
74. A. R. Jones, The roofed spherical sundial and the Greek geometry of curves, In: *Studies in the ancient exact sciences in honor of Lis Brack-Bernsen*, ed. J. Steele and M. Ossendrijver, Berlin studies of the ancient world, Topoi, Berlin, 2017.
75. J. Kepler, *Mysterium cosmographicum* (1596). English translation: The secret of the universe, transl. A. M. Duncan, Abaris, New York, 1981.
76. S. S. Kutateladze, Alexandrov of Ancient Hellas, *Siberian Electronic Mathematical Reports*, 9 (2012), pp. A6-A11.
77. J.-L. de Lagrange, Théorie des fonctions analytiques contenant les principes du calcul différentiel dégagés de toute considération d'infiniment petits ou d'évanouissans de limites ou de fluxions, Paris, Imprimerie de la République, 1797.
78. M.-A. Lancret, Mémoire sur les courbes à double courbure, Académie des Sciences de Paris, Mémoire des savants étrangers, Vol. I (1806), pp. 416–454.
79. M.-A. Lancret, Mémoire sur les développoïdes des courbes planes, des courbes à double courbure, et des surfaces développables, Mémoire des savants étrangers, Vol. I (1811), pp. 1–79.
80. G. W. Leibniz, Meditatio nova de natura anguli contactus et osculi: horumque usu in practica mathesi ad figuras faciliores succedaneas difficilioribus substituendas, Acta Eruditorum, 1686, apud J. Grossium & J. F. Gletitschium, 1687.
81. G. W. Leibniz, Generalia de natura linearum, anguloque contactus et osculi, provolutionibus, aliisque cognatis, et eorum usibus nunnullis, Acta Eruditorum, 1692.
82. J.-B. Meusnier, Mémoire sur la courbure des surfaces, Mém. div. sav. Paris, Vol. X (1785), pp. 477–510.
83. J.-F.-Th. Maurice, article sur Christiaan Huygens, Biographie Universelle, Paris, Michaud, 1818.
84. G. Monge, Mémoire sur les propriétés de plusieurs genres de surfaces courbes et particulièrement sur celles des surfaces développables avec une application à la théorie générale des ombres et des pénombres, Mém. Div. savants 9 (1780), pp. 382–440.
85. G. Monge, Mémoire sur la théorie des déblais et remblais, Mém. Div. sav., 1781.

86. G. Monge, Mémoire sur les développées, les rayons de courbure et les différents genres d'inflexions des courbes à double courbure, Imprimerie Royale, Paris, 1785.
87. G. Monge, Sur les lignes de courbure de la surface de l'ellipsoïde, *J. Éc. Polytechnique*, 2e cahier (1796), pp. 145–165.
88. G. Monge, Applications de l'analyse à la géométrie, Paris, 1807.
89. S. Mukhopadhyaya, Note on T. Hayashi's paper on the osculating ellipse of a plane curve, *Rend. Circ. palermo* vol. 51 (1927), pp. 394.
90. S. Mukhopadhyaya, Extended minimum number theorems of cyclic and sextatic points on a plane oval, *Math. Z.* 33 (1931) pp. 648–662.
91. S. Mukhopadhyaya, Circles incident on an oval of undefined curvature, *Tohoku vo.* 34 (1931) pp. 115–129.
92. O. Neugebauer, The astronomical origin of the theory of conic sections, *Proc. Am. Phil. Soc.* 92, No. 3 (1948), pp. 136–138.
93. A. Papadopoulos, Herbert Busemann, *Notices of the American Mathematical Society*, 2018, Vol. 65, No. 3, March 2018, pp. 341–343.
94. Pappus, the Collection, French translation by P. Ver Eecke, Fondation universitaire de Belgique, 1932. Reprint, Blanchard, Paris, 1982.
95. F. C. Penrose, *An Investigation of the Principles of Athenian Architecture; or, The Results of a Survey Conducted Chiefly with Reference to the Optical Refinements Exhibited in the Construction of the Ancient Buildings at Athens*, Macmillan, London, 1888.
96. H. Pemberton, *A View of Sir Isaac Newton's Philosophy*, printed by S. Palmer, London, 1728.
97. H. Pitot, Sur la quadrature de la moitié d'une courbe qui est la compagne des arcs, appelée la compagne de la cycloïde, *Histoire de l'Académie des sciences* (1976), pp. 107–113.
98. H. Poincaré, Démonstration nouvelle des propriétés de l'indicatrice d'une surface. *Nouvelles annales de mathématiques*, 2e série, Vol. 13 (1874) pp. 449–456.
99. Proclus, *A commentary on the first book of Euclid's Elements*, Princeton University Press, 1970.
100. R. Rashed, Apollonius : Les Coniques, tome 1.1 : Livre I, commentaire historique et mathématique, édition et traduction du texte arabe, de Gruyter, 2008.
101. R. Rashed, Descartes et l'infiniment petit, *Boll. Stor. Sci. Mat.* 33, No. 1, 151–169 (2013).
102. K. Reidemeister, Über Körper konstanten Durchmessers, *Math. Z.* Vol. 10 (1921), pp. 214–216.
103. E. Rinner, Ancient Greek sundials and the theory of conics reconsidered, In: *Studies in the ancient exact sciences in honor of Lis Brack-Bernsen*, ed. J. Steele and M. Ossendrijver, Berlin studies of the ancient world, Topoi, Berlin, 2017.
104. O. Rodrigues, Sur quelques propriétés des intégrales doubles et des rayons de courbure des surfaces, *Bulletin scientifique de la société philomatique de Paris* (1815), pp. 34–36.
105. O. Rodrigues, Recherches sur la théorie des lignes et des rayons de courbure des surfaces, et sur la transformation d'une classe d'intégrales doubles qui ont un rapport direct avec les formules de cette théorie, *Correspondance de l'Ecole Polytechnique*, 3 (1815), pp. 162–182.
106. P. Roitman and H. Le Ferrand, The strange case of Paul Appell's last memoir on Monge's problem: "sur les déblais et remblais", *Historia Mathematica*, Vol. 43, Issue 3, August 2016, pp. 288–309.
107. I. M. Serrano and B. D. Suceava, A Medieval Mystery: Nicole Oresme's Concept of Curvitas. *Notices of the AMS*, 62 (October 2015), pp. 1030–1034.
108. D. J. Struik, *Outline of a history of differential geometry*, Saint Catherine Press Limited, 1933.
109. D. J. Struik, *Lectures on classical differential geometry*, Addison Wesley, Reading, Mass. 1961.
110. P. Tannery, Pour l'histoire des lignes et des surfaces courbes dans l'antiquité, *Bulletin des Mathématiques et astronomiques*, 8 (1884), 25–27.
111. J.-A. Taschereau, Notice sur le général Meusnier d'après, des notes biographiques de Monge et d'autres documents également inédits, Revue rétrospective, ou Bibliothèque historique, contenant des mémoires et documents authentiques (2) 4 (1835), pp. 82–91.

112. R. Taton, *L'œuvre scientifique de Monge*, Presses Universitaires de France, Paris, 1951.
113. R. Taton, La première note mathématique de Gaspard Monge (juin 1769), *Revue d'histoire des sciences et de leurs applications*, tome 19, no. 2 (1966), pp. 143–149.
114. V. A. Zalgaller, Memoirs of A. D. Alexandrov and his Leningrad geometry seminar, *Siberian Electronic Mathematical Reports*, 9 (2012), pp. A26-A80.

*   &)

Pág. 241(256) – Gauss Curvature

Fermat's principle of least time saying that the path take between two points by a light ray is the path which can be traversed in the least time. (60).

↳Letter (January 1, 1662) to M. C. de la Chambre

Employing it, he deduced the law of refraction.

Hence shortest curves on a surface are regarded as something similar to paths traced by light rays (see Sect 6.12),

and it is reasonable to say that a quantity (or a concept) attached to a surface is "intrinsic" if it is derived from distance and angle measured by using shortest curves.

Pág. 337(352) →

Olinde Rodrigues (1795-1851)

Rodrigues is known for a formula

on this subject carrying his
name, and for
developing the idea of a
representation of a differentiable
surface on a sphere of radius
one which amounts to the
Gauss map.

He used such a (fal) ✡
representation to study the ratio of
the area of (a piece of)
the original surface to the
area of its image on the
sphere.)

⟸

arriving
at a measure of curvature
which in fact is the
Gaussian Curvature
at a point of a surface

Pag. 448(461) →

Husserl was confusing
intrinsic and extrinsic
natures of space.
He did not see to pay attention
to the fact that spheres and Euclidean
planes can be realised even in hyperbolic
space.

* "" "' ?

in particular think of Frege, Husserl and Russell as those who studied seriously epistemological problem of space and time. All of them had an academic training in mathematics. Frege, who is often regarded as a precursor of logical positivism, studied the foundation of mathematics, as in [8], which inspired both Wittgenstein and the famous work of Russell-Whitehead, *Principia Mathematica*. Frege also had good knowledge on the work of Riemann. Still, his attitude toward the non-Euclidean geometries and epistemological problem of space can be regarded as staying within a Kantian framework: Frege seemed to believe that our space of intuition should be the Euclidean three-dimensional space, as can be seen in his correspondence with Hilbert and his review on Hilbert's book, both of which can be found in [9] (see also Shipley [45] for a more detailed account on Frege's epistemological view on geometry and space).

Husserl, who is now regarded as a precursor of phenomenology, also worked for philosophical aspects of mathematics. He gave a course on Riemann's theory of geometry at University of Halle in 1889 [16], where he dealt with Riemann's notion of metric critically. In particular, after criticising Riemann's definition of curvature, Husserl claimed that since Euclidean space is necessary to define Gaussian curvature, Euclidean space is a special entity which is given to us by pure intuition in Kantian sense. If we scrutinise his argument from the viewpoint of modern geometry, we notice that Husserl was confusing intrinsic and extrinsic natures of space. He did not seem to pay attention to the fact that spheres and Euclidean planes can be realised even in hyperbolic space. (Recall that this is why Lobachevsky coined the term "Pangeometry" as we noted in §4.)

Russell, who belongs to a younger generation than the other two, published in his twenties a book entitled *An essay on foundation of geometry* [42]. In this book, he criticised Kantian apriorism in geometry, and observed that there are two elements in geometry, what is a priori and what is empirical. Russell rejected the a-priority of Euclidean geometry, and claimed that experience can decide which of Euclidean or non-Euclidean geometry is valid. He further insisted that their common ground, which is projective geometry, should be a priori. This means that Russell embraced some part of conventionalism, but retained a weak version of Kantian apriorism. This position of Russell was harshly criticised by Poincaré in his review of this book.

Thus we have seen that even these three famous philosophers who were familiar with Riemann's work could not be as radical as Poincaré in abandoning Kantianism, and this seems to be caused by the fact that without experience of research on geometry and topology, they could not reach the level of Poincaré's understanding of space. In the latter half of the twentieth century, there appeared philosophers such as Merleau-Ponty, who analysed how our perception leads to our recognition of space, but in Poincaré's day no professional philosopher could approach the epistemology of space as deeply as Poincaré. According to Althusser, as we have explained above, the role of philosophers would be to draw a line of demarcation between science and ideology in spontaneous philosophy by scientists. However for this to be possible, philosophers need to have a deep understanding on what constitutes the core of spontaneous philosophy. In Poincaré's case, the core is nothing but his geometric conventionalism, which in turn depends on his research in geometry and topology. In his time, it was very difficult for other philosophers to reach this level for playing the role to draw such a line of demarcation in Poincaré's epistemology.

References

1. L. Althusser, *Philosophie et philosophie spontanée des savants*, François Maspero, Paris, 1974.
2. L. Althusser, *Lenine et la philosophie*, François Maspero, Paris, 1972.
3. E. C. Banks, Kant, Herbart and Riemann, Kant-Studien, *Philosophische Zitschrift*, 96 (2005) 208–234.
4. E. Beltrami, Teoria fondamentale degli spazii di curvatura costante, *Annali di Matematica Pura ed Applicata*, 2 (1868) 232–255.
5. V. N. Berestovskii, Poincaré conjecture and related statements, in *Geometry in History*, edited by S.G. Dani, A. Papadopoulos, Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-13609-3_17
6. E. Betti, Sopra gli spazi de un numero qualunque di dimensioni, *Annali di Matematica* (2) 4, 140–158.
7. M. Freedman, The topology of four-dimensional manifolds, *J. Differential Geom.* 17 (1982), 357–453.
8. G. Frege, *Die Grundlagen der Arithmetik, eine logisch mathematische Untersuchung über den Begriff der Zahl*, Koebner, Breslau, 1884.
9. G. Frege, Kleine Schriften ; herausgegeben von Ignacio Angelell, Georg Olms, Hildesheim, 1967.
10. J. F. Fries, *Neue oder anthropologische Kritik der Vernunft*, Bande 1–3, 1828–1831, Winter, Heidelberg.
11. W. Goldman, Flat affine, projective and conformal structures on manifolds: a historical perspective, in *Geometry in History*, edited by S.G. Dani, A. Papadopoulos, Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-13609-3_14
12. J. Gray, *Plato's ghost: the modernist transformation of mathematics*, Princeton University Press, Princeton NJ, 2008.
13. F. Hausdorff, *Grundzüge der Mengenlehre*, Veit, Leipzig, 1914.
14. P. Heegaard, Forstudier til en topologisk teori for de algebraiske Fladers Sammenhäng, Thesis; Traduction française: Sur l'Analysis Situs, *Bull. Soc. Math. France*, 44 (1916), 161–242.
15. J. F. Herbart, *Allgemeine Metaphysik, nebst den Anfängen der philosophischen Naturlehre* (2 Teile), A.W. Unzer, Königsberg, T.1 1828, T. 2, 1829.
16. E. Husserl, Husserliana : gesammelte Werke Bd. 21, Studien zur Arithmetik und Geometrie : Texte aus dem Nachlass (1886–1901), Nijhoff, Hague, 1983.
17. I. Kant, *Critik der reinen Vernunft*, Johann Friedrich Hartknoch, Riga 1781, 2 Aufl 1787.
18. F. Klein, Ueber die sogenannte Nicht-Euklidische Geometrie, *Math. Ann.*, 4, 573–625 (1871).
19. V. Lénine, *Matérialisme et empirio-criticisme*, Editions sociales Paris et Editions du progrès Moscou, (traduit du russe).
20. N. Lobatchevsky, *Pangeometry*, edited by A. Papadopoulos, EMS Heritage of European Mathematics, 4, 2010.
21. J. Morgan and G. Tian, *Ricci Flow and the Poincare Conjecture*, Clay Mathematics Monographs, 3. American Mathematical Society, Providence, RI; Clay Mathematics Institute, Cambridge, MA, 2007. xlii+521.
22. M. Morse, *Calculus of Variations in the Large*. Colloquium Publications Volume: 18; 1934; 368 pp; American Mathematical Society, Providence, R.I..

23. K. Ohshika, The origin of the notion of manifold: from Riemann's Habilitationsvortrag onward, in "*From Riemann to differential geometry and relativity*" edited by L. Ji, A. Papadopoulos and S. Yamada, Springer 2017.
24. A. Papadopoulos, On the work of Euler and his followers on spherical geometry, *Gaṇita Bhāratī* 36 (2014), no. 1, 53–108.
25. A. Papadopoulos, Euler, la géométrie sphérique et le calcul des variations, in "*Leonhard Euler : Mathématicien, physicien et théoricien de la musique*", CNRS Éditions, Paris, 349–392, 2015.
26. E. Picard, Sur une extension aux fonctions de deux variables du problème de Riemann relatif aux fonctions hypergéométriques. *Annales scientifiques de l'École Normale Supérieure*, Sér. 2, 10 (1881), pp. 305–322.
27. V. Poénaru, A glimpse into the problems of the fourth dimension, in this volume.
28. H. Poincaré, Théorie des groupes Fuchsiens, *Acta Math.*, 1, 1–62, (1882).
29. H. Poincaré, Sur l'Analysis Situs, *Comptes rendus hebdomadaires de l'Académie des sciences de Paris*, 115 (1892), 189–192.
30. H. Poincaré, Analysis situs, *Journal de l'École Polytechnique*, t. 1, pp.1–121 (1895).
31. H. Poincaré, Complément à l'Analysis situs, premier complément, *Rendiconti del Circolo Matematico di Palermo*, 13, 285–343 (1899).
32. H. Poincaré, Second complément à l'Analysis situs, *Proceedings of the London Mathematical Society*, 32, 277–308 (1900).
33. H. Poincaré, Sur certaines surfaces algébriques ; troisième complément à l'Analysis Situs, *Bulletin de la Société Mathématique de France*, 30, 49–70 (1902).
34. H. Poincaré, Sur les cycles des surfaces algébriques. Quatrième complément à l'Analysis Situs, *Journal de Mathématiques*, 8, 169–214 (1902).
35. H. Poincaré, Cinquième complément à l'Analysis Situs, *Rendiconti del Circolo matematico di Palermo*, 18, 45–110 (1904).
36. H. Poincaré, *La science et l'hypothèse*, Flammarion, Paris, 1902.
37. H. Poincaré, *La valeur de la science*, Flammarion, Paris,1905.
38. H. Poincaré, *Science et méthode*, Flammarion, Paris, 1908
39. H. Poincaré, *Dernières pensées*, Flammarion, Paris, 1913.
40. B. Riemann, Über die Hypothesen, welche der Geometrie zu Grunde liegen, *Abhandlungen der Königlichen Gesellschaft der Wissenschaften zu Göttingen*, 13, (1867).
41. L. Rollet, Henri Poincaré, des Mathématiques à la philosophie, Étude du parcours intellectuel, social et politique d' un mathématicien au début du siècle, Thèse, 1999, Université de Nancy.
42. B. Russell, *An essay on the foundations of geometry*, Cambridge University Press, Cambridge 1897.
43. K. S. Sarkaria, The topological work of Henri Poincaré, in "*History of Topology*", Edited by I.M. James, North-Holland, 123–167.
44. E. Scholtz, The Concept of Manifold, 1850–1950, in "*History of Topology*", Edited by I.M. James, North-Holland, 25–64.
45. J. Shipley, Frege on the foundation of geometry in intuition, *Journal for the History of Analytical Philosophy*, 3–6, (2015).
46. S. Smale, Generalized Poincaré's conjecture in dimensions greater than four, *Ann. of Math.* (2) 74 1961 391–406.
47. W. Thurston, Three-dimensional manifolds, Kleinian groups and hyperbolic geometry, *Bull. Amer. Math. Soc.* (N.S.) 6 (1982), no. 3, 357–381.

= $ $) %
R7 (%& * ' !# " # R7 ! 4 6
! & #

! " # !# $ $ % & "' (!
) # !# * #
! +,-.+ + +/ ! + "' & 0 !# 0
1 2 ! # 0 3 2 3 3#
( )) * *

4.4 Post-Thurston Era

In this century, there were two miracles that independently resolved these extremely wild conjectures of 1982. We will briefly describe them in this section.

4.4.1 Geometrization Conjecture

There had been many attempts to partially answer or approach the geometrization conjecture for the case of infinite π_1 since it was formulated. However the case of finite π_1 had stayed completely outside of the development in the last century except for the orbifold theorem and Hamilton–Perelman's attempts.

Perelman put three papers [30–32] on arXiv in 2002 and 2003 and declared the resolution of the geometrization conjecture in the affirmative. The proof was based on Hamilton's Ricci flow, which is a flow on the space of Riemannian metrics on a manifold satisfying an evolutional PDE,

$$\frac{d}{dt} g = -2\,\mathrm{Ric}_g$$

where Ric represents the Ricci curvature tensor. This equation can be regarded as a variational equation that averages metrics in the long term. After Hamilton proved a sort of geometrization for 3-manifolds with positive Ricci curvature in [11], Yau suggested to Hamilton that the solution of the Ricci flow might converge to the metric that the geometrization conjecture expects. Hamilton had been seriously involved in Yau's suggestion and has been establishing a program to resolve the geometrization conjecture through several deep studies on the Ricci flow in a series of papers [12–14]. A survey by Chow [7] discusses Hamilton's work along these lines.

Perelman essentially followed Hamilton's program. However, he had to fill in conceptual and technical details and needed to modify Hamilton's set up. His breakthrough would be to establish a locally non-collapsing theorem which guarantees that the rescaling limit of the solution of the Ricci flow is controllable. This is presented in the first preprint [30]. Then following [13], Perelman introduced Ricci flow with surgery in [31] and claimed the resolution of the geometrization conjecture.

Theorem 6 (Perelman) *The geometrization conjecture is true.*

This theorem was officially recognized at the ICM 2006 in Madrid. Excellent expositions of this innovation were presented by Milnor [25] in its very early stage and by McMullen [23] in its mature stage. The third preprint [32] presents a bypath argument independent from a half of the argument in [31] to prove for instance the Poincaré conjecture, when we start with a connected sum of manifolds with finite π_1.

Perelman's proof of the geometrization conjecture has a similarity with the proof of the orbifold theorem. Both arguments are based on the deformation of metrics. The difference is that Thurston used hyperbolic metrics with cone singularities, while Perelman used nonsingular Riemaniann metrics. Thus Thurston's argument is geometric, while Perelman's involves more analysis. Perelman's arguments brought a big surprise not only to geometers but also to analysts world-wide. His arguments certainly led to a new direction for the study of singular solutions for some PDE. In fact, Perelman's work has promoted the popularity of the term "blow up" in contemporary PDE theory.

4.4.2 *Virtual Fiber Conjecture*

Once again, the virtual Haken conjecture by Waldhausen states that every irreducible 3-manifold with infinite π_1 is virtually Haken, namely, finitely covered by a Haken manifold. After Perelman's resolution of the geometrization conjecture, this conjecture remained unsolved only for hyperbolic manifolds. Thus, the virtual fiber conjecture came to include the virtual Haken conjecture, and this motivated a significant number of people. There were actually several serious studies by experts including Thurston, but most of them were rather illustrative or experimental before Agol [1]. In this paper, Agol presented a fairly generic criterion for virtual fibering that he called residually finite rational solvability. It is an effective refinement of separability of a subgroup in a group. Agol showed, in particular, that any subgroup of a right-angled Artin group has this useful property.

There was a strong reason for Agol to reach to this definition. Early this century, Wise and his many collaborators developed the theory of special groups. They started with groups acting effectively on non-positively curved (NPC for short) cube complexes. One of their most notable contributions was the discovery of the specialness of a group or an action defined in [15]. Specialness can be defined by a somewhat technical combinatorial structure of an NPC cube complex. Surprisingly, Haglund–Wise [15] showed that specialness can be characterized in a purely group theoretic fashion: that a special group is a subgroup of a right angled Artin group.

An epoch making result was provided by Kahn–Markovic [18] who showed that there are many almost geodesic quasi-Fuchsian surfaces in any closed hyperbolic 3-manifold. Their result, which was presented at the 21st Nevanlinna Colloquium held at Kyoto University in 2009, gave a completely new tool to discuss the virtual fiber conjecture. In fact, Bergeron–Wise [3] immediately noticed that the result could be used to build a wall system of Sageev, as in [36], in the 3-dimensional hyperbolic space invariant under the action of a cocompact Kleinian group. Then the recipe by Sageev [35, 36] leads to a construction of a proper group action on an NPC cube complex by the same group.

Agol then succeeded in showing that the action constructed in this way is special, and concluded that

Theorem 7 (Agol) *The virtual fiber conjecture is true, and so is the virtual Haken conjecture.*

What Agol actually proved in [2] was much stronger. He showed that every compact cube complex with a hyperbolic fundamental group is virtually special. In other words, a hyperbolic group that acts on an NPC cube complex cocompactly is virtually special. The resolution of the virtual fiber conjecture is one of its many consequences when combined with previous results by Sageev, Wise and his collaborators, Kahn–Markovic and Bergeron–Wise. Thus the theorem by Agol pertains not only to the theory of 3-manifolds. However, we would like to note that the idea of the proof was connected to the techniques developed in the theory of 3-manifold, in the Thurston Era, especially hyperbolic Dehn surgery theory in the context of group theory.

We would like to refer for more details of this innovative result to the articles by Klarreich [20] for general audiences, and Bestvina [4] and Friedle [9] for experts.

4.5 Epilogue

Thurston's vision of 3-manifold theory, as presented for instance in the problem set in [40], has been completed by Perelman and Agol. However, it is more important to mention that Thurston presented us simultaneously with many new directions in mathematics even within the theory of 3-manifolds. In fact, we found fine geometric nature of 3-manifolds and many interesting questions. The connection between quantum topology and hyperbolic geometry in 3-manifolds could be one such fascinating topics. For example, there is the volume conjecture, which expects to precisely identify the hyperbolic volume of a knot complement with a certain asymptote of special values of quantum invariants of the knot. This conjecture has motivated and involved many researchers including people in mathematical physics. There are many other germs to create new links from 3-manifold theory within topology, with the other mathematical fields including geometry and science in general. For instance, geometric group theory is now expanding in many new directions, Teichmüller theory is finding interest in a higher dimensional analogue, and so on.

What Thurston did and left to us was extremely incredible and more than dramatic indeed.

References

1. I. Agol, Criteria for virtual fibering. J. Topol. **1**, 269–284 (2008)
2. I. Agol, The virtual Haken conjecture, with appendix by I. Agol, D. Groves and J. Manning. Doc. Math. **18**, 1045–1087 (2013)
3. N. Bergeron, D. Wise, A boundary criterion for cubulation. Am. J. Math. **134**, 843–859 (2012)
4. M. Bestvina, Geometric group theory and 3-manifolds hand in hand: the fulfillment of Thurston's vision. Bull. Am. Math. Soc. **51**, 53–70 (2014)
5. M. Boileau, J. Porti, Geometrization of 3-dimensional orbifolds. Ann. Math. **162**, 195–290 (2005)
6. A. Casson, D. Jungreis, Convergence groups and Seifert fibered 3-manifolds. Invent. Math. **118**, 441–456 (1994)
7. B. Chow, A survey on Hamilton's program for the Ricci flow on 3-manifolds. AMS Contemp. Math. **367**, 63–78 (2005)
8. D. Cooper, C. Hodgson, S. Kerckhoff, The orbifold theorem. MSJ Mem. **5** (2000)
9. S. Friedle, Thurston's vision and the virtual fibering theorem for 3-manifolds, Jahresber. Deutsch. Math.-Verein. **116**, 223–241 (2014)
10. D. Gabai, Convergence groups are Fuchsian groups. Ann. Math. **136**, 447–510 (1992)
11. R. Hamilton, Three-manifolds with positive Ricci curvature. J. Differ. Geom. **17**, 255–306 (1982)
12. R. Hamilton, The formation of singularities in the Ricci flow. Surv. Differ. Geom. **2**, 7–136 (1995)
13. R. Hamilton, Four-manifolds with positive isotoropic curvature. Comm. Anal. Geom. **5**, 1–92 (1997)
14. R. Hamilton, Non-singular solution of the Ricci flow on 3-manifolds. Comm. Anal. Geom. **7**, 695–729 (1999)
15. F. Haglund, D. Wise, Special cube complexes. Geom. Funct. Anal. **17**, 1551–1620 (2008)
16. W. Jaco ad P. Shalen, Seifert fibered spaces in 3-manifolds. Mem. Amer. Math. Soc. **21**(2), 220 (1979)
17. K. Johannson, *Homotopy Equivalence of 3-manifolds with Boundaries*. Lecture Notes in Mathematics, no. 7761 (Springer, Berlin, 1979)
18. J. Kahn, V. Markovic, Immersing almost geodesic surfaces in a closed hyperbolic 3-manifold, Ann. Math. **175**, 1127–1190 (2012)
19. M. Kapovich, Hyperbolic manifolds and discrete groups (Modern Birkhäuser Classics, Birkhäuser, 2009)
20. E. Klarreich, Getting into shapes: from hyperbolic geometry to cube complexes and back. Quanta Magazine. Sci. Am. (2012). https://scientificamerican.com/article/getting-into-shapes/
21. H. Kneser, Geschlossen Flächen in dreidimensionalen Mannigfaltigkeiten. Jber. Deutsch. Math. Verein. **38**, 248–260 (1929)
22. C. McMullen, Renormalization and 3-manifolds which fiber over the circle. Ann. Math. Study **142** (1996)
23. C. McMullen, The evolution of geometric structures on 3-manifolds, Bull. Amer. Math. Soc. **48**, 259–274 (2011)
24. J. Milnor, A unique decomposition theorem for 3-manifolds, Amer. J. Math. **84**, 1–7 (1962)
25. J. Milnor, Towards the Poincaré conjecture and the classification of 3-manifolds. Notices Amer. Math. Soc. **50**, 1226–1233 (2003)
26. J. Morgan, H. Bass, *The Smith Conjecture*. Pure and Applied Mathematics, vol. 112 (Academic Press, New York, 1984)
27. G. Mostow, Quasi-conformal mappings in n-space and the rigidity of the hyperbolic space forms. Publ. IHES **34**, 53–104 (1968)
28. J.P. Otal, Le théorème d'hyperbolisation pour les variétés fibrées de dimension 3. Astérisque **235** (1986)

29. J.P. Otal, Thurston's hyperbolization of Haken manifolds, in *Surveys in Differential Geometry*, vol. III, ed. by C.C. Hsiung, S.T. Yau (International Press, Boston, 1998), pp. 77–194
30. G. Perelman, The entropy formula for the Ricci flow and its geometric applications (2002). arxiv:math/0211159
31. G. Perelman, Ricci flow with surgery on three-manifolds (2003). arXiv:math/0303109
32. G. Perelman, Finite extinction time for the solutions to the Ricci flow on certain three-manifolds (2003). arXiv:math/0309021
33. H. Poincaré, Cinquième élément à l'analysis situs. Rend. Circa. Mat. Palema **18**, 45–110 (1904)
34. G. Prasad, Strong rigidity of $\mathbb{Q}$-rank 1 lattices. Inventiones Math. **21**, 255–286 (1973)
35. M. Sageev, Codimension-1 subgroups and splittings of groups. J. Algebra **189**, 377–389 (1997)
36. M. Sageev, CAT(0) Cube Complexes and Groups. PCMI Lecture Notes **21**, 7–54 (2014)
37. P. Scott, The geometry of 3-manifolds. Bull. London Math. Soc. **15**, 401–487 (1983)
38. D. Sullivan, Travaux Thurston sur les groups quasi-Fuchsiens et les variétés hyperboliques de dimension 3 firbrees sur S^1. Seminaire Bourbaki, 1979/80 no. 554
39. W. Thurston, *The Geometry and Topology of* 3-*manifolds*. Lecture Notes (Princeton University, Princeton, 1979). Available at http://library.msri.org/books/gt3m/
40. W. Thurston, Three dimensional manifolds, Kleinian groups and hyperbolic geometry. Bull. Amer. Math. Soc. **6**, 357–382 (1982)
41. W. Thurston, Hyperbolic structures on 3-manifolds, I: Deformations of acylinderical manifolds. Ann. Math. **124**, 203–246 (1986)
42. W. Thurston, On the geometry and dynamics of diffeomorphisms of surfaces. Bull. Amer. Math. Soc. **19**, 417–431 (1988)
43. W. Thurston, Hyperbolic structures on 3-manifolds III: Deformations of 3-manifolds with incompressible boundary (1998). arXiv:math/9801058
44. W. Thurston, Hyperbolic structures on 3-manifolds II: Surface groups and 3-manifolds which fiber over the circle (1999). arXiv:math/9801049
45. F. Waldhausen, On irreducible 3-manifolds which are sufficiently large. Ann. Math. **87**, 56–88 (1968)

20210915 (cont.)

Books-018-2021/20210915/
/Handbook of mathematics/
- In the tradition of Thurston -
Geometry and Topology (2021).pdf
Pág. 167 (186) 4.4
Post-Thurston Era
4.4.1 Geometrization Conjecture
Perelman (three papers)
The proof was based on
Hamilton's Ricci flow,
which is a flow on the space of
Riemannian metrics on a
manifold satisfying an
evolutional PDE,

$$\frac{d}{dt} g = -2 \, Ric \, g$$

where Ric represents the RICCI
CURVATURE TENSOR.

Pag. 174(192) → Richard Hamilton

—

— Geometries and transformations (2018).pdf

Pag. 57(75) Chapter 3 Circular Geometries

Pag. 58(76) → A. Inversive geometry.

A projective hyperplane $\check{U}$, with coordinates $\llbracket u \rrbracket$, meets Φ in an inversive $(n-1)$-sphere, i.e., a hypersphere of I^n, which is real if $\llbracket u\,u \rrbracket > 0$, degenerate if $\llbracket u\,u \rrbracket = 0$, imaginary if $\llbracket u\,u \rrbracket < 0$, where $\llbracket u\,u \rrbracket$ is the quadratic form associated with the bilinear form

$$\llbracket u\,v \rrbracket = -u_0v_0 + u_1v_1 + \dots + u_{n+1}v_{n+1}.$$

—

Chapter 5
Combinatorics Encoding Geometry: The Legacy of Bill Thurston in the Story of One Theorem

Philip L. Bowers

Abstract This chapter presents a whirlwind tour of some results surrounding the *Koebe–Andre'ev–Thurston Theorem*, Bill Thurston's seminal circle packing theorem that appears in Chapter 13 of *The Geometry and Topology of Three-Manifolds*.

Keywords Circle packing · Uniformization · Circle domain · Discrete conformal geometry · Conformal type · Inversive distance · Polyhedron

AMS Codes 52C26, 57M20

5.1 Introduction

Bill Thurston was the most original and influential topologist of the last half-century. His impact on the discipline of geometric topology during that time is unsurpassed, and his insights in the topology and geometry of three-manifolds led finally to the resolution of the most celebrated problem of topology over the last century—The Poincaré Conjecture. He made fundamental contributions to many sub-disciplines within geometric topology, from the theory of foliations on manifolds to the combinatorial structure of rational maps on the two-sphere, and from geometric and automatic group theory to classical polyhedral geometry. Of course his foundational work on three-manifolds, first laid out in his courses at Princeton in the late 1970s, compiled initially as a Princeton paper-back monograph inscribed by Bill Floyd and available upon request as *The Geometry and Topology of Three-Manifolds* (GTTM) [73], and maturing as the famous *Thurston Geometrization Conjecture* of the early 1980s, was the driving force behind the development of geometric topology for the next thirty years. The final confirmation of the Geometrization Conjecture

---

P. L. Bowers (✉)
Department of Mathematics, Florida State University, Tallahassee, FL, USA
e-mail: bowers@math.fsu.edu

© Springer Nature Switzerland AG 2020

K. Ohshika, A. Papadopoulos (eds.), *In the Tradition of Thurston*,
https://doi.org/10.1007/978-3-030-55928-1_5

by Giorgi Perelman using the flow of Ricci curvature, following a program that had been introduced by Richard Hamilton, is one of the crown jewels of twentieth century mathematics.

Thurston marks a watershed in the short history of topology,[1] a signpost, demarcating topology before Thurston, and topology after Thurston. This is evidenced not only in the fabulous results he proved, explained, and inspired, but even more so in how he taught us to do mathematics. Topology before Thurston was dominated by the general and the abstract, entrapped in the rarified heights that captured the mathematical world in general, and topology in particular, in the period from the 1930s until the 1970s. Topology after Thurston is dominated by the particular and the geometric, a throwback to the nineteenth-century, having much in common with the highly geometric landscape that inspired Felix Klein and Max Dehn, who walked around and within Riemann surfaces, knew them intimately, and understood them in their particularity. Thurston's vision gave a generation of topologists permission to get their collective hands dirty by examining in great depth specific structures on specific examples.

One of the organizing principles that lies behind Thurston's vision is that geometry informs topology, and that the non-Euclidean geometry of Lobachevski, Bolyai, and Beltrami in particular is systemic to the study of topology. Hyperbolic geometry permeates topology after Thurston, and it is hyperbolic geometry that becomes the common thread of the present chapter. This will be seen in the interrelated studies presented here. All to varying degrees are due to the direct influence of Bill Thurston and his generalization of the earlier results of Koebe and Andre'ev. All involve hyperbolic geometry in some form or influence, and even further all illustrate how combinatorics encodes geometry, another of the principles that underlies Thurston's vision. To my mind, the proposition that *combinatorics encodes geometry, which in turn informs topology* has become a fundamental guiding motif for topology after Thurston. I offer this chapter as a celebration of Bill Thurston's vision and his immense influence over our discipline.

5.1.1 An Introductory Overview

The Koebe–Andre'ev–Thurston Theorem represents a rediscovery and broad generalization of a curiosity of Paul Koebe's from 1936, and has an interpretation that recovers a characterization of certain three-dimensional hyperbolic polyhedra due to E.M. Andre'ev in two papers from 1970. This theorem is the foundation stone of the discipline that has been dubbed as *discrete conformal geometry*, which itself has been developed extensively by many mathematicians in many different directions over the last thirty years. Discrete conformal geometry in its purest form is geometry born of combinatorics, but it has theoretical and practical applications.

---

[1] I will use the term *topology* henceforth to mean *geometric topology*. By dropping the adjective *geometric* I certainly mean no slight of general, set-theoretic, or algebraic topology.

In the theoretical realm, it produces a discrete analytic function theory that is faithful to its continuous cousin, a quantum theory of complex analysis from which the classical theory emerges in the limit of large scales. In the realm of applications, it has been developed in a variety of directions, for practical applications in areas as diverse as biomedical imaging and 3D print head guidance. This rather large body of work flows from simple insights that Thurston presented in his lecture at Purdue University in 1985 on how to use the most elementary case of his circle packing theorem to provide a practical algorithm for approximating the Riemann mapping from a proper, simply-connected planar domain to the unit disk. A personal accounting of this development can be found in the author's own review [14] of the bible of circle packing theory, Ken Stephenson's *Introduction to Circle Packing: The Theory of Discrete Analytic Functions* [71].

A perusal of the section and subsection headings of this chapter will give the reader a clue as to where I am going in this survey. I primarily stick with the theoretical results for which there are fairly direct lines from the Koebe–Andre'ev–Thurston Theorem to those results. This means in particular that I almost totally ignore the really vast array of practical applications that circle packing has found, especially in the last two decades as discrete differential geometry has become of primary importance in so many applications among computer scientists and computational geometers. A survey of applications will have to wait as space constraints preclude a discussion that does justice to the topic.

5.1.2 Dedication and Appreciation

This chapter is dedicated to the memory of Bill Thurston and his student Oded Schramm, and to an appreciation of Jim Cannon and Ken Stephenson. I have spoken already of Bill Thurston's legacy. Oded Schramm was one of the first to press Thurston's ideas on circle packings to a high level of development and application, and his great originality in approaching these problems has bequeathed to us a treasure trove of beautiful gems of mathematics. Most of Oded's work on circle packing and discrete geometry was accomplished in the decade of the 1990s. As Bill is a demarcation point in the history of topology, Oded is one in the history of probability theory. In the late nineties, Oded became interested in some classical open problems in probability theory generated by physicists, in percolation theory and in random planar triangulations in particular. Physicists had much theoretical and computational evidence for the veracity of their conjectures, but little mathematical proof, or even mathematical tools to approach their verifications. In Oded's hands these venerable conjectures and problems began to yield to mathematical proof, using ingenious tools developed or refined by Oded and his collaborators, chief among which are SLE_κ, originally Stochastic-Loewner Evolution, now renamed as Schramm–Loewner Evolution, and UIPT's, or Uniform Infinite Planar Triangulations. For a wonderful biographical commentary on Oded's contributions to mathematics, see Steffen Rohde's article *Oded Schramm: From Circle Packing to SLE* in [58].

The two individuals who have had the greatest impact on my mathematical work are Jim Cannon and Ken Stephenson, the one a mathematical hero of mine, the other my stalwart collaborator for three decades. Jim's work has influenced mine significantly, and I greatly admire his mathematical tastes and contributions. Pre-Thurston, Jim had made a name for himself in geometric topology in the flavor of Bing and Milnor, having solved the famous double suspension problem and having made seminal contributions to cell-like decomposition theory and the characterization of manifolds. In the beginning of the Thurston era, his influential paper *The combinatorial structure of cocompact discrete hyperbolic groups* [24] anticipated many of the later developments of geometric group theory, presaging Gromov's thin triangle condition and, à la Dehn, the importance of negative curvature in solving the classical word and conjugacy problems of combinatorial group theory. He with Thurston invented automatic group theory and then Jim settled upon the conjecture that bears his name as the work that for three decades has consumed his attention. Ken has been a joy with whom to collaborate over the past three decades. He was inspired upon attending Thurston's Purdue lecture in 1985 to change his mathematical attentions from a successful career as a complex function theorist, to a geometer exploring this new idea of circle packing using both traditional mathematical proof and the power of computations for mathematical experimentation. I began my foray into Thurston-style geometry and topology by answering in [17] a question of Ken and Alan Beardon from one of the first research papers [7] to appear on circle packings after Rodin and Sullivan's 1987 paper [62] confirming the conjecture of Thurston from the Purdue lecture. Ken and I are co-authors on a number of research articles and his down-to-earth approach to the understanding of mathematics has been a constant check on my tendency toward flights of fancy. I have learned from him how to tell a good story of a mathematical topic. For Ken's warm friendship and collaboration I am grateful.

5.2 The Koebe–Andre'ev–Thurston Theorem, Part I

5.2.1 Koebe Uniformization and Circle Packing

In the early years of the twentieth century, rigorous proofs of the Riemann Mapping Theorem and the more general Uniformization Theorem were given by such eminent mathematicians as Osgood, Carathéodory, Poincaré, and Koebe, and refinements and re-workings would continue to be made by others, even up to the present.[2] The generalization of the Riemann Mapping Theorem to multiply-

---

[2] The author recommends rather highly the article *On the history of the Riemann mapping theorem* [38] by Jeremy Gray and the monograph *Uniformization of Riemann Surfaces: Revisiting a Hundred-Year-Old Theorem* [32]. These two works give insightful historical accountings of the discovery, articulation, understanding, and finally rigorous proofs of the Riemann Mapping

4 +,-.+ + +/ ! + "' 1 ! # ! !
! 4 # 9 8 2 3#
(" " !3 '

3

CIRCULAR GEOMETRIES

If projective geometry is about incidences of points, lines, and planes, inversive geometry deals with point pairs, circles, and spheres. In a manner that closely parallels how a linear metric space can be extracted from projective or affine geometry, a circular metric space can be derived from inversive geometry by fixing a certain *central inversion* (Johnson 1981; cf. Alekseevsky, Vinberg & Solodovnikov 1988, chap. 2, § 2). Depending on the nature of the fixed inversion, the resulting geometry may be either *spherical* or *pseudospherical*.

When represented in Euclidean space, inversive geometry also provides a way to construct conformal (angle-preserving) models of elliptic and hyperbolic spaces. The two non-Euclidean metrics will be seen to have a kind of "pseudo-symmetry" that is evident in formulas relating the angles and sides of a triangle and in measurements of circular arcs and sectors.

3.1 INVERSIVE AND SPHERICAL GEOMETRY

We begin with a hyperbolic polarity in real projective $(n+1)$-space. Let the points X and hyperplanes $\check{U}$ of P^{n+1} have homogeneous coordinates

$$((x)) = (x_0, x_1, \dots, x_{n+1}) \quad \text{and} \quad [[u]] = [u_0, u_1, \dots, u_{n+1}],$$

and let an oval n-quadric Φ have the locus or envelope equation

$$x_1{}^2 + \cdots + x_{n+1}{}^2 = x_0{}^2 \quad \text{or} \quad u_1{}^2 + \cdots + u_{n+1}{}^2 = u_0{}^2,$$

corresponding to the matrix $\backslash -1,\ 1, \ldots,\ 1\backslash$. We may regard Φ as the real inversive, or *Möbius*, n-sphere I^n (Veblen & Young 1918, pp. 262–264). The case $n = 2$ is of particular interest; a fuller description of the properties of the real inversive sphere I^2 will be given in Chapter 9.

A. *Inversive geometry.* A projective hyperplane $\check{U}$, with coordinates $[[u]]$, meets Φ in an inversive $(n-1)$-sphere $\mathring{u}$, a *hypersphere* of I^n, which is

$$\begin{array}{lll} \textit{real} & \text{if} & [[u\,u]] > 0, \\ \textit{degenerate} & \text{if} & [[u\,u]] = 0, \\ \textit{imaginary} & \text{if} & [[u\,u]] < 0, \end{array}$$

where $[[u\,u]]$ is the quadratic form associated with the bilinear form

$$[[u\,v]] = -u_0v_0 + u_1v_1 + \cdots + u_{n+1}v_{n+1}.$$

A point X lies on I^n if its coordinates $((x))$ satisfy the locus equation for Φ and on the hypersphere $\mathring{u}$ with hyperplanar coordinates $[[u]]$ if $((x))[[u]] = 0$. The equation $((x))[[u]] = 0$ has at least two, just one, or no real solutions for $((x))$ (on I^n) according as $\mathring{u}$ is real, degenerate, or imaginary; in the last case, the equation has two or more imaginary solutions.

If the discriminant of two real or degenerate inversive hyperspheres $\mathring{u}$ and $\mathring{v}$ is defined by

$$||u\,v|| = [[u\,u]]\,[[v\,v]] - [[u\,v]]^2,$$

then the hyperspheres are

$$\begin{array}{lll} \textit{separating} & \text{if} & ||u\,v|| > 0, \\ \textit{tangent} & \text{if} & ||u\,v|| = 0, \\ \textit{separated} & \text{if} & ||u\,v|| < 0. \end{array}$$

Applied to two point-pairs $\ddot{u} = \{U_1,\ U_2\}$ and $\ddot{v} = \{V_1,\ V_2\}$ on a circle I^1, separation has the usual meaning associated with cyclic order (Coxeter 1966a, p. 218). That is, $\ddot{u}$ and $\ddot{v}$ are separating if $U_1U_2 \,/\!/\, V_1V_2$, tangent if $\ddot{u}$ and $\ddot{v}$ have a point in common, and separated otherwise. For $n \geq 2$, two hyperspheres of I^n are separating, tangent, or separated according as they intersect in a real, degenerate, or imaginary $(n-2)$-sphere.

The angle between two separating inversive hyperspheres $\mathring{u}$ and $\mathring{v}$ is given by

$$(\mathring{u}\ \mathring{v}) = \cos^{-1}\frac{|[\![u\ v]\!]|}{\sqrt{[\![u\ u]\!]}\sqrt{[\![v\ v]\!]}}. \tag{3.1.1}$$

Two separated inversive hyperspheres $\mathring{u}$ and $\mathring{v}$ have an analogous *inversive distance* (Coxeter 1966b), given by

$$)\mathring{u}\ \mathring{v}(= \cosh^{-1}\frac{|[\![u\ v]\!]|}{\sqrt{[\![u\ u]\!]}\sqrt{[\![v\ v]\!]}}. \tag{3.1.2}$$

Another inversive invariant of note is the cross ratio of four points (Wilker 1981, pp. 384–386; Beardon 1983, pp. 32–33). The distance between two points is not defined, and an inversive hypersphere has no center or radius.

B. *Spherical geometry.* A projective collineation of P^{n+1} that commutes with the polarity that defines the n-quadric Φ induces a circle-preserving transformation, or *circularity*, on Φ as the Möbius n-sphere I^n. In particular, a harmonic homology whose center O and median hyperplane $\check{O}$ are pole and polar with respect to Φ induces an *inversion* on I^n. When O and $\check{O}$ have the respective coordinates

$$(1,\ 0, \ldots,\ 0,\ 0) \quad \text{and} \quad [1,\ 0, \ldots,\ 0,\ 0],$$

this is an *elliptic* inversion $Z : \mathrm{I}^n \to \mathrm{I}^n$, i.e., the inversion in the imaginary hypersphere in which $\check{O}$ meets Φ, leaving no real points fixed. If we restrict ourselves to those circularities that commute with Z, we convert the Möbius n-sphere I^n into the *elliptic* n-sphere S^n ("spherical n-space"), with Z becoming the familiar *central inversion* that interchanges antipodal points. (It is a happy accident that two different, seemingly unrelated, meanings of "inversion"—in a hypersphere and in a point—are united in the same transformation.)

For every point X on S^n, $x_0 \neq 0$, and normalized coordinates for X are the last $n+1$ projective coordinates divided by x_0, i.e.,

$$\mathbf{(}x\mathbf{)} = (x_1, x_2, \ldots, x_{n+1}),$$

where ${x_1}^2+{x_2}^2+\cdots+{x_{n+1}}^2 = 1$. The points $\mathbf{(}x\mathbf{)}$ and $-\mathbf{(}x\mathbf{)}$ are antipodal. A hyperplane $\check{U}$ passing through the point O, so that $u_0 = 0$, intersects S^n in a *great hypersphere* $\check{U}$, with homogeneous coordinates

$$\mathbf{[}u\mathbf{]} = [u_1, u_2, \ldots, u_{n+1}].$$

We have $X \Diamond \check{U}$ if and only if $\mathbf{(}x\mathbf{)[}u\mathbf{]} = 0$. Other hyperplanes intersect S^n in real, degenerate, or imaginary "small" hyperspheres.

Two great hyperspheres of S^n are always separating, with any two great circles on an elliptic sphere S^2 meeting in a pair of antipodal points.

Whereas the geometry of I^n is only conformal, the elliptic n-sphere S^n is a full-fledged metric space. In terms of the bilinear forms

$$\mathbf{(}x\ y\mathbf{)} = x_1y_1 + \cdots + x_{n+1}y_{n+1} \quad \text{and} \quad \mathbf{[}u\ v\mathbf{]} = u_1v_1 + \cdots + u_{n+1}v_{n+1},$$

with $\mathbf{(}x\ x\mathbf{)} = \mathbf{(}y\ y\mathbf{)} = 1$, the distance (from 0 to π) between two points X and Y on S^n is given by

$$[\![X\ Y]\!] = \cos^{-1} \mathbf{(}x\ y\mathbf{)}, \tag{3.1.3}$$

and the angle (from 0 to $\pi/2$) between two great hyperspheres $\check{U}$ and $\check{V}$ by

$$((\check{U}\check{V})) = \cos^{-1} \frac{|\mathbf{[}u\ v\mathbf{]}|}{\sqrt{\mathbf{[}u\ u\mathbf{]}}\sqrt{\mathbf{[}v\ v\mathbf{]}}}. \tag{3.1.4}$$

The distance from a point X to a great hypersphere $\check{U}$ is given by

$$[\![X\check{U})\!) = \sin^{-1}\frac{|(x)[u]|}{\sqrt{[u\,u]}}. \tag{3.1.5}$$

An n-sphere S^n can be embedded in other metric spaces. For instance, the point O and the hyperplane $\check{O}$ may be taken as the origin and the hyperplane at infinity of Euclidean $(n+1)$-space E^{n+1}, and S^n is then the unit hypersphere $(x\,x) = 1$. If antipodal points are identified, as happens under projection from O to $\check{O}$, the elliptic n-sphere S^n is transformed into elliptic n-space $\mathrm{eP}^n = \mathrm{PS}^n = \mathrm{S}^n/\langle Z\rangle$. Great hyperspheres of S^n project into hyperplanes of eP^n.

EXERCISES 3.1

1. Show that the circles $[1, 1, 0, 1]$ and $[1, 0, -1, 2]$ on the inversive sphere I^2 are separating and find the angle between them.

2. Show that the circles $[1, 0, -1, 1]$ and $[1, 1, 2, 0]$ on the inversive sphere I^2 are separated and find the distance between them.

3. Find the distance from the point $(\tfrac{1}{3}, -\tfrac{2}{3}, \tfrac{2}{3})$ to the great circle $[4, 0, 3]$ on the elliptic sphere S^2.

3.2 PSEUDOSPHERICAL GEOMETRY

Returning to the Möbius n-sphere as an oval n-quadric Φ in P^{n+1} defined by

$$x_1{}^2 + \cdots + x_{n+1}{}^2 = x_0{}^2 \quad \text{or} \quad u_1{}^2 + \cdots + u_{n+1}{}^2 = u_0{}^2,$$

a harmonic homology whose center O and median hyperplane $\check{O}$ have the respective coordinates

$$(0, 0, \ldots, 0, 1) \quad \text{and} \quad [0, 0, \ldots, 0, 1]$$

induces a *hyperbolic* inversion $W : \mathrm{I}^n \to \mathrm{I}^n$—i.e., the inversion in the real inversive hypersphere $\mathring{\omega}$ in which $\check{O}$ meets Φ, leaving every point of $\mathring{\omega}$ fixed. If W is taken as the central inversion, interchanging antipodal points, then the Möbius n-sphere I^n is converted into the *hyperbolic* n-sphere $\ddot{\mathrm{S}}^n$, separated by $\mathring{\omega}$ into two n-hemispheres. We call $\mathring{\omega}$ the *equatorial hypersphere* of $\ddot{\mathrm{S}}^n$; all other hyperspheres and all points not on $\mathring{\omega}$ are *ordinary*. Up to a scalar factor, the hyperbolic n-sphere has the same metric as an n-pseudosphere (equidistant hypersurface) in H^{n+1}, and $\ddot{\mathrm{S}}^n$ may be called "pseudospherical n-space."

For an ordinary point X on $\ddot{\mathrm{S}}^n$, $x_{n+1} \neq 0$, and normalized coordinates for X are the first $n+1$ projective coordinates (rearranged) divided by x_{n+1}, i.e.,

$$\mathbf{(}x\mathbf{)} = (x_1, \ldots, x_n \,|\, x_0),$$

with ${x_1}^2 + \cdots + {x_n}^2 - {x_0}^2 = -1$. Two ordinary points X and Y with coordinates $\mathbf{(}x\mathbf{)}$ and $\mathbf{(}y\mathbf{)}$ lie in the same n-hemisphere if $x_0 y_0 > 0$. The points $\mathbf{(}x\mathbf{)}$ and $-\mathbf{(}x\mathbf{)}$ are antipodal, and points on the equatorial hypersphere are self-antipodal. A hyperplane $\breve{U}$ passing through O, so that $u_{n+1} = 0$, and having a real intersection with $\ddot{\mathrm{S}}^n$, so that $[\![u\ u]\!] > 0$, meets $\ddot{\mathrm{S}}^n$ in a *great hypersphere* $\breve{U}$, having homogeneous coordinates

$$\mathbf{[}u\mathbf{]} = [u_1, \ldots, u_n \,|\, u_0],$$

with ${u_1}^2 + \cdots + {u_n}^2 - {u_0}^2 > 0$. As before, $X \Diamond \breve{U}$ if and only if $\mathbf{(}x\mathbf{)}\mathbf{[}u\mathbf{]} = 0$. Other hyperplanes meet $\ddot{\mathrm{S}}^n$ in various real, degenerate, or imaginary "small" hyperspheres.

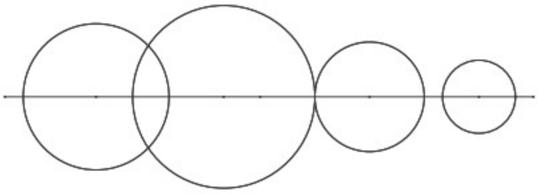

Figure 3.2a Separating, tangent, and separated great circles

Two great hyperspheres of $\ddot{\mathrm{S}}^n$ may be separating, tangent, or separated. (To distinguish the cases, apply the discriminant for inversive hyperspheres, assuming a final hyperplanar coordinate of zero.) Any two separating great circles on a hyperbolic sphere $\ddot{\mathrm{S}}^2$ meet in a pair of antipodal points; tangent great circles have one common equatorial point; separated great circles have no real points in common (see Figure 3.2a). An ordinary small circle on $\ddot{\mathrm{S}}^2$ is a *pseudocycle*, a *horocycle*, or a *pericycle* according as it and the equatorial circle are separating, tangent, or separated (see Figure 3.2b).

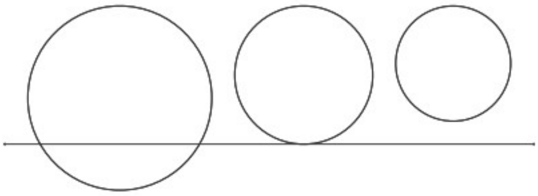

Figure 3.2b Pseudocycle, horocycle, and pericycle

In terms of the bilinear forms

$$(x \ddot{\ } y) = x_1y_1 + \cdots + x_ny_n - x_0y_0 \quad \text{and} \quad [u \ddot{\ } v] = u_1v_1 + \cdots + u_nv_n - u_0v_0,$$

with $(x \ddot{\ } x) = (y \ddot{\ } y) = -1$ (cf. Ratcliffe 1994/2006, § 3.2), the distance between two points X and Y in the same n-hemisphere of $\ddot{\mathrm{S}}^n$ is given by

$$]]XY[[= \cosh^{-1} |(x \ddot{\ } y)|, \tag{3.2.1}$$

the angle (from 0 to $\pi/2$) between two separating or tangent great hyperspheres $\breve{U}$ and $\breve{V}$ by

$$((\breve{U}\breve{V})) = \cos^{-1} \frac{|[u \ddot{\ } v]|}{\sqrt{[u \ddot{\ } u]}\sqrt{[v \ddot{\ } v]}}, \tag{3.2.2}$$

and the minimum distance between two separated great hyperspheres $\breve{U}$ and $\breve{V}$ by

$$]]\breve{U}\breve{V}[[= \cosh^{-1} \frac{|[u \ddot{\ } v]|}{\sqrt{[u \ddot{\ } u]}\sqrt{[v \ddot{\ } v]}}. \tag{3.2.3}$$

The distance from a point X to a great hypersphere $\breve{U}$ is given by

$$]]X\breve{U}((= \sinh^{-1} \frac{|\mathbf{(}x\mathbf{)[}u\mathbf{]}|}{\sqrt{\mathbf{[}u\,\ddot{}\,u\mathbf{]}}}. \tag{3.2.4}$$

The point O and the hyperplane $\breve{O}$ can be taken as the origin and the hyperplane at infinity of *Lorentzian* (or "pseudo-Euclidean") $(n+1)$-space $\mathrm{E}^{n,1}$, in which $\ddot{\mathrm{S}}^n$ appears as an n-hyperboloid of two sheets (a hyperbola if $n = 1$), asymptotic to the "light cone" $\mathbf{(}x\,\ddot{}\,x\mathbf{)} = 0$. Lorentzian space is homogeneous but not isotropic. The (not necessarily real) distance between two points X and Y with respective Cartesian coordinates $\mathbf{(}x\mathbf{)}$ and $\mathbf{(}y\mathbf{)}$ is given by

$$|X\,\ddot{}\,Y| = \sqrt{\mathbf{(}y-x\,\ddot{}\,y-x\mathbf{)}} = [(y_1-x_1)^2+\cdots+(y_n-x_n)^2-(y_0-x_0)^2]^{1/2}, \tag{3.2.5}$$

so that metrically $\ddot{\mathrm{S}}^n$ is a hypersphere with center O and imaginary radius i. The properties of such an "imaginary-spherical" geometry were anticipated by Johann Lambert (1728–1777) and extensively explored by Franz Taurinus (1794–1874), but it was Lobachevsky (1835, 1836) who first recognized it as a valid alternative to Euclidean and spherical geometry.

Just as the elliptic n-sphere S^n can be embedded in hyperbolic $(n+1)$-space as an n-perisphere with center O (and ideal median hyperplane $\breve{O}$), so the hyperbolic n-sphere $\ddot{\mathrm{S}}^n$ can be embedded in H^{n+1} as an n-pseudosphere with median hyperplane $\breve{O}$ (and ideal center O).

Either of the two n-hemispheres into which $\ddot{\mathrm{S}}^n$ is separated by the equatorial hypersphere $\mathring{\omega}$ (i.e., either sheet of the n-hyperboloid in $\mathrm{E}^{n,1}$ or the n-pseudosphere in H^{n+1}) will serve as an isometric model for hyperbolic n-space. When $n = 2$, the Lorentzian realization is the *hyperboloid model* of the hyperbolic plane (Ratcliffe 1994/2006, § 3.2; cf. Reynolds 1993). If antipodal points are identified, as by projecting from O to $\breve{O}$, the hyperbolic n-sphere $\ddot{\mathrm{S}}^n$ is transformed into extended hyperbolic n-space $\bar{\mathrm{H}}^n = \mathrm{P}\ddot{\mathrm{S}}^n = \ddot{\mathrm{S}}^n/\langle W\rangle$. Great hyperspheres of $\ddot{\mathrm{S}}^n$

project into hyperplanes of H^n, and the equatorial hypersphere $\mathring{\omega}$ of $\ddot{\mathrm{S}}^n$ becomes the absolute hypersphere ω of $\bar{\mathrm{H}}^n$.

Still another way to represent $\ddot{\mathrm{S}}^n$ is by Euclidean n-space E^n augmented by a single "exceptional" point that lies on every line, with some hyperplane $\check{O}$ taken to represent the equatorial hypersphere $\mathring{\omega}$. In this conformal "double half-space" model, the points and the k-planes or k-spheres of E^n lying in $\check{O}$ are the equatorial points and k-spheres of $\ddot{\mathrm{S}}^n$. Points that are mirror images with respect to $\check{O}$ are antipodal points of $\ddot{\mathrm{S}}^n$. Euclidean k-planes or k-spheres orthogonal to $\check{O}$ are great k-spheres of $\ddot{\mathrm{S}}^n$.

EXERCISES 3.2

1. Find the angle between the separating great circles $[-1, 1 \mid 1]$ and $[1, 2 \mid 1]$ on the hyperbolic sphere $\ddot{\mathrm{S}}^2$.
2. Find the distance between the separated great circles $[2, 0 \mid 1]$ and $[-2, \sqrt{3} \mid 2]$ on the hyperbolic sphere $\ddot{\mathrm{S}}^2$.

3.3 CONFORMAL MODELS

The elliptic and hyperbolic n-spheres can be projected into Euclidean n-space to provide inversive models for elliptic and hyperbolic n-space. Such mappings preserve circles and angular measure.

A. *Elliptic geometry.* Elliptic n-space eP^n can be conformally represented by a closed ball in E^n, i.e., a hypersphere Ω together with its interior. The k-planes of eP^n are diametral k-disks or caps of k-spheres intersecting Ω in great $(k-1)$-spheres, with diametrically opposite points of Ω being identified. When $n = 2$, this is Klein's "conformal disk" model of the elliptic plane.

To obtain explicit metrics for this and other Euclidean models, let the points of E^n have Cartesian coordinates $(\dot{x}) = (\dot{x}_1, \ldots, \dot{x}_n)$, and let

$(\dot{x}\,\dot{x}) = \dot{x}_1{}^2 + \cdots + \dot{x}_n{}^2$. In the elliptic "conformal ball" model, points of eP^n are represented by Euclidean points $(\dot{x})$ with $(\dot{x}\,\dot{x}) \leq 1$, where $(\dot{x})$ and $-(\dot{x})$ are identified if $(\dot{x}\,\dot{x}) = 1$. Each such point X also has normalized spherical coordinates

$$(x) = (x_1, \ldots, x_n, x_{n+1}),$$

where

$$x_i = \frac{2\dot{x}_i}{1 + (\dot{x}\,\dot{x})}\ (1 \leq i \leq n) \quad \text{and} \quad x_{n+1} = \frac{1 - (\dot{x}\,\dot{x})}{1 + (\dot{x}\,\dot{x})}. \tag{3.3.1}$$

(Optionally, the sign of the last coordinate can be reversed.) Each hyperplane $\check{U}$ has homogeneous spherical coordinates

$$[u] = [u_1, \ldots, u_n, u_{n+1}],$$

with $X \lozenge \check{U}$ if and only if $(x)[u] = 0$.

In terms of the bilinear form

$$(x\,y) = x_1y_1 + \cdots + x_ny_n + x_{n+1}y_{n+1},$$

with $(x\,x) = (y\,y) = 1$, the elliptic distance between two points X and Y is given by

$$[XY] = \cos^{-1} |(x\,y)|. \tag{3.3.2}$$

The angle $(\check{U}\check{V})$ between two hyperplanes $\check{U}$ and $\check{V}$ and the distance $[X\check{U})$ from a point X to a hyperplane $\check{U}$ are respectively equal to the spherical angle $((\check{U}\check{V}))$ and distance $[[X\check{U}))$ defined by (3.1.4) and (3.1.5).

For example (cf. Figure 3.3a), selected points of E^2 with Cartesian coordinates $(\dot{x}_1,\ \dot{x}_2) = (x,\ y)$ correspond to points of eP^2 with coordinates (x_1, x_2, x_3) as follows:

$$(0,\ 0) \mapsto (0,\ 0,\ 1), \quad (1,\ 0) \mapsto (1,\ 0,\ 0), \quad (-1,\ 0) \mapsto (-1,\ 0,\ 0),$$

$$(\tfrac{1}{2}, \tfrac{1}{2}) \mapsto (\tfrac{2}{3}, \tfrac{2}{3}, \tfrac{1}{3}), \quad (-\tfrac{1}{2}, -\tfrac{1}{2}) \mapsto (-\tfrac{2}{3}, -\tfrac{2}{3}, \tfrac{1}{3}),$$

$$(0, \sqrt{2} - 1) \mapsto (0, \tfrac{1}{2}\sqrt{2}, \tfrac{1}{2}\sqrt{2}).$$

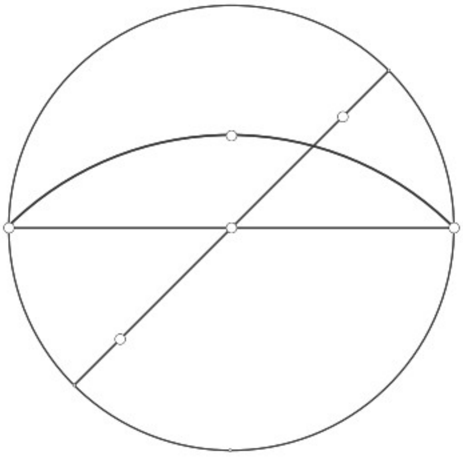

Figure 3.3a Conformal-disk model of eP^2

Coordinates $(1, 0, 0)$ and $(-1, 0, 0)$ represent the same point. Lines through distinct points of eP^2 have coordinates $[u_1, u_2, u_3]$ as follows:

$$(0,\ 0) \vee (1,\ 0) = [0,\ 1,\ 0], \quad (0,\ 0) \vee (\tfrac{1}{2}, \tfrac{1}{2}) = [1,\ -1,\ 0],$$
$$(1,\ 0) \vee (0, \sqrt{2} - 1) = [0,\ 1,\ -1].$$

Using Formula (3.3.2), we see that the elliptic distance between $(0, 0)$ and either $(½, ½)$ or $(-½, -½)$ is $\cos^{-1} ⅓$ and that the distance between $(½, ½)$ and $(-½, -½)$ is $\cos^{-1} 7/9$. The angle between $[0, 1, 0]$ and either $[1, -1, 0]$ or $[0, 1, -1]$ is $\pi/4$, and the angle between $[1, -1, 0]$ and $[0, 1, -1]$ is $\pi/3$.

B. *Hyperbolic geometry.* Hyperbolic n-space H^n can be conformally represented by an open ball in Euclidean n-space, i.e., the interior of a hypersphere Ω. The k-planes of H^n are diametral k-disks or caps of k-spheres orthogonal to Ω. When $n = 2$, this is the "conformal disk" model of the hyperbolic plane. Equivalently, either of the half-spaces into which E^n is separated by a hyperplane $\check{O}$ provides another conformal model. In this case, the k-planes of H^n are represented by Euclidean k-half-planes or k-hemispheres orthogonal to $\check{O}$. When $n = 2$, this is the *Poincaré model*, also called the "upper half-plane" model, of the hyperbolic plane (Poincaré 1882).

Like the projective model developed by Klein, these conformal representations of hyperbolic n-space were first described by Beltrami (1868b; cf. Milnor 1982, pp. 10–11; Stillwell 1996, pp. 35–38). The inversive and projective models are closely related: each arc orthogonal to the boundary circle in E^2 can be transformed into a corresponding chord, converting a disk model of H^2 that preserves angles into one that preserves collinearity.

In the hyperbolic "conformal ball" model (cf. Yaglom 1956, supp. 2; Ratcliffe 1994/2006, § 4.5), points of H^n are represented by Euclidean points $(\dot{x})$ with $(\dot{x}\,\dot{x}) < 1$. Each such point X also has normalized pseudospherical coordinates

$$(x) = (x_1, \ldots, x_n \,|\, x_0),$$

where

$$x_0 = \frac{1 + (\dot{x}\,\dot{x})}{1 - (\dot{x}\,\dot{x})} \quad \text{and} \quad x_i = \frac{2\dot{x}_i}{1 - (\dot{x}\,\dot{x})}\ (1 \le i \le n). \tag{3.3.3}$$

In the "upper half-space" model (cf. Ratcliffe 1994/2006, § 4.6), points of H^n are represented by Euclidean points $(\dot{x})$ with $\dot{x}_n > 0$. Each such point X also has normalized pseudospherical coordinates (x), where

$$x_0 = \frac{1 + (\dot{x}\,\dot{x})}{2\dot{x}_n}, \quad x_i = \frac{2\dot{x}_i}{2\dot{x}_n}\ (1 \le i \le n-1), \quad x_n = \frac{1 - (\dot{x}\,\dot{x})}{2\dot{x}_n}. \tag{3.3.4}$$

In either case, each hyperplane $\check{U}$ has homogeneous pseudospherical coordinates

$$[u] = [u_1, \ldots, u_n \,|\, u_0],$$

with ${u_1}^2 + \cdots + {u_n}^2 - {u_0}^2 > 0$. The point X and the hyperplane $\check{U}$ are incident if and only if $(x)[u] = 0$.

In terms of the bilinear form

$$(x\,\ddot{}\,y) = x_1y_1 + \cdots + x_ny_n - x_0y_0,$$

with $(x\,\ddot{}\,x) = (y\,\ddot{}\,y) = -1$, the hyperbolic distance between two points X and Y is given by

$$]XY[= \cosh^{-1} |(x\,\ddot{}\,y)|. \tag{3.3.5}$$

The angle $(\check{U}\check{V})$ between two intersecting or parallel hyperplanes $\check{U}$ and $\check{V}$, the minimum distance $)\check{U}\check{V}($ between two diverging hyperplanes $\check{U}$ and $\check{V}$, and the distance $]X\check{U}($ from a point X to a hyperplane $\check{U}$ are respectively equal to the pseudospherical angle $((\check{U}\check{V}))$ and distances $))\check{U}\check{V}(($ and $]]X\check{U}(($ defined by (3.2.2), (3.2.3), and (3.2.4).

For example (cf. Figure 3.3b), by (3.3.3) with $n = 2$, in the "conformal disk" model of the hyperbolic plane, points of E^2 with Cartesian coordinates (x, y) correspond to points of H^2 with coordinates $(x_1, x_2 \,|\, x_0)$ as follows:

$$(0,0) \mapsto (0,0\,|\,1), \quad (\tfrac{1}{2},0) \mapsto (\tfrac{4}{3},0\,|\,\tfrac{5}{3}), \quad (-\tfrac{1}{2},0) \mapsto (-\tfrac{4}{3},0\,|\,\tfrac{5}{3}),$$

$$(\tfrac{1}{2},\tfrac{1}{2}) \mapsto (2,2\,|\,3), \quad (-\tfrac{1}{2},-\tfrac{1}{2}) \mapsto (-2,-2\,|\,3),$$

$$(\tfrac{1}{5},\tfrac{2}{5}) \mapsto (\tfrac{1}{2},1\,|\,\tfrac{3}{2}), \quad (\tfrac{2}{5},\tfrac{1}{5}) \mapsto (1,\tfrac{1}{2}\,|\,\tfrac{3}{2}),$$

$$(-\tfrac{1}{5},-\tfrac{2}{5}) \mapsto (-\tfrac{1}{2},-1\,|\,\tfrac{3}{2}), \quad (-\tfrac{2}{5},-\tfrac{1}{5}) \mapsto (-1,-\tfrac{1}{2}\,|\,\tfrac{3}{2}).$$

Hyperbolic lines through certain pairs of points have coordinates $[u_1, u_2 \,|\, u_0]$ as follows:

$$(0,0) \vee (\tfrac{1}{2},0) = [0,1\,|\,0], \quad (0,0) \vee (\tfrac{1}{2},\tfrac{1}{2}) = [1,-1\,|\,0],$$

$$(\tfrac{1}{5},\tfrac{2}{5}) \vee (\tfrac{2}{5},\tfrac{1}{5}) = [1,1\,|\,-1], \quad (-\tfrac{1}{5},-\tfrac{2}{5}) \vee (-\tfrac{2}{5},-\tfrac{1}{5}) = [1,1\,|\,1].$$

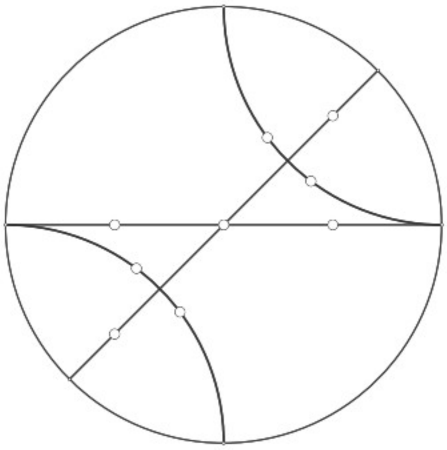

Figure 3.3b Conformal-disk model of H^2

Using Formula (3.3.5), we see that the hyperbolic distance between $(0, 0)$ and either (½, ½) or (−½, −½) is $\cosh^{-1} 3$ and that the distance between (½, ½) and (−½, −½) is $\cosh^{-1} 17$. The angle between $[0, 1 \,|\, 0]$ and $[1, -1 \,|\, 0]$ is $\pi/4$, and the angle between $[1, -1 \,|\, 0]$ and either $[1, 1 \,|\, -1]$ or $[1, 1 \,|\, 1]$ is $\pi/2$. The line $[0, 1 \,|\, 0]$ is parallel (in opposite directions) to both $[1, 1 \,|\, -1]$ and $[1, 1 \,|\, 1]$, and the angle between it and either of these lines is therefore 0. The lines $[1, 1 \,|\, -1]$ and $[1, 1 \,|\, 1]$ themselves are diverging, and the distance between them is $\cosh^{-1} 3$.

Inversion in the circle $x^2 + (y + 1)^2 = 2$ converts the "conformal disk" representation of H^2 into the "upper half-plane" model. Points (x, y) with $x^2 + y^2 < 1$ are interchanged with points (x', y') with $y' > 0$, where

$$x' = \frac{2x}{x^2 + (y + 1)^2}, \quad y' = \frac{2(y + 1)}{x^2 + (y + 1)^2} - 1. \tag{3.3.6}$$

The correspondence between Cartesian and hyperbolic point coordinates is now given by (3.3.4) with $n = 2$.

For example, the inversion given by (3.3.6) interchanges particular points in the unit disk with points in the upper half-plane as follows:

$$(0,0) \leftrightarrow (0,1), \quad (\tfrac{1}{2},0) \leftrightarrow (\tfrac{4}{5},\tfrac{3}{5}), \quad (-\tfrac{1}{2},0) \leftrightarrow (-\tfrac{4}{5},\tfrac{3}{5}),$$

$$(\tfrac{1}{2},\tfrac{1}{2}) \leftrightarrow (\tfrac{2}{5},\tfrac{1}{5}), \quad (-\tfrac{1}{2},-\tfrac{1}{2}) \leftrightarrow (-2,1),$$

$$(\tfrac{1}{5},\tfrac{2}{5}) \leftrightarrow (\tfrac{1}{5},\tfrac{2}{5}), \quad (-\tfrac{1}{5},\tfrac{2}{5}) \leftrightarrow (-\tfrac{1}{5},\tfrac{2}{5}),$$

$$(-\tfrac{1}{5},-\tfrac{2}{5}) \leftrightarrow (-1,2), \quad (-\tfrac{2}{5},-\tfrac{1}{5}) \leftrightarrow (-1,1).$$

The inversive-disk model of the hyperbolic plane is converted into the projective-disk (Beltrami-Klein) model by the mapping $(x,\ y) \mapsto (\xi,\ \eta)$, where

$$\xi = \frac{2x}{x^2 + y^2 + 1}, \quad \eta = \frac{2y}{x^2 + y^2 + 1}. \tag{3.3.7}$$

Each circular arc orthogonal to the unit circle $x^2 + y^2 = 1$ is mapped onto the chord with the same endpoints; each diameter is mapped onto itself. If we write $x = r\cos\theta$ and $y = r\sin\theta$, then the mapping (3.3.7) can be expressed in polar form as $(r,\ \theta) \mapsto (\rho,\ \theta)$, where $\rho = 2r/(r^2+1)$.

C. *Spherical and pseudospherical geometry.* Spherical n-space S^n can be conformally represented by a hypersphere of radius r in Euclidean $(n+1)$-space, and pseudospherical n-space $\ddot{\mathrm{S}}^n$ by a hypersphere of radius $r\mathrm{i}$ in Lorentzian $(n+1)$-space. For $n \geq 2$, the (Gaussian) *curvature* of S^n or $\ddot{\mathrm{S}}^n$, or of the derived elliptic or hyperbolic n-space, is the reciprocal of the square of the radius of this n-sphere. Thus S^n and eP^n have constant positive curvature $1/r^2$, while $\ddot{\mathrm{S}}^n$ and H^n have constant negative curvature $-1/r^2$; Euclidean n-space E^n is *flat*, with curvature 0.

For $n \geq 2$, the only simply connected n-manifolds of constant curvature are S^n, E^n, and H^n. Elliptic n-space eP^n is not simply connected, and pseudospherical n-space $\ddot{\mathrm{S}}^n$ is not connected.

For simplicity, all non-Euclidean distance formulas given in this section or elsewhere in the book assume a spatial curvature of ± 1. In spherical or elliptic n-space of curvature 1, a k-sphere of radius r or coradius $\check{r} = \pi/2 - r$ has the same curvature $\csc^2 r = \sec^2 \check{r}$ as a Euclidean k-sphere of radius $\sin r$. In pseudospherical or hyperbolic n-space of curvature -1, a k-perisphere of radius r has the same curvature $\operatorname{csch}^2 r$ as a Euclidean k-sphere of radius $\sinh r$; a k-horosphere has curvature 0; a k-pseudosphere of coradius $\check{r}$ has curvature $-\operatorname{sech}^2 \check{r}$.

Points on the elliptic sphere S^2, as the unit sphere of Euclidean 3-space, have normalized coordinates (x, y, z) with $x^2+y^2+z^2 = 1$. With the terrestrial (or celestial) sphere as a model, the antipodal points $(0, 0, 1)$ and $(0, 0, -1)$ are the "north pole" and "south pole" of S^2. Any great semicircle joining the poles is a "meridian," the one whose points have $y = 0$ and $x > 0$ being the "prime" meridian.

A point on S^2 can be given polar (or "geographic") coordinates (ϕ, θ), with $0 \leq \phi \leq \pi$ and $-\pi < \theta \leq \pi$ (see Figure 3.3c). The

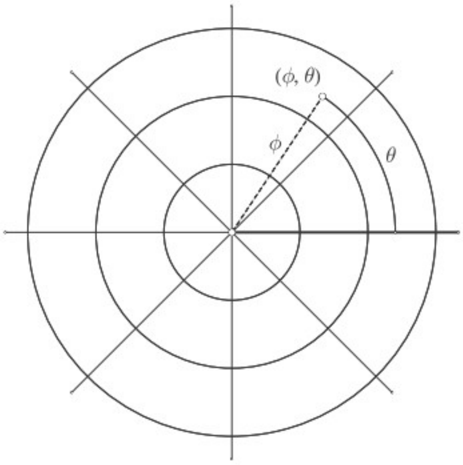

Figure 3.3c Polar coordinates

colatitude ϕ is the distance of the point from the north pole, while the *longitude* θ of a nonpolar point is the directed angle from the prime meridian to the meridian of the point. The longitude of either pole is arbitrary. The *latitude* of any point is the directed distance $\pi/2-\phi$ from the point to the "equator" $[0, 0, 1]$. A point on the elliptic sphere with polar coordinates (ϕ, θ) has normalized spherical coordinates (x, y, z), where

$$x = \sin\phi\cos\theta, \quad y = \sin\phi\sin\theta, \quad z = \cos\phi. \tag{3.3.8}$$

Similarly, ordinary points on the hyperbolic sphere $\ddot{\mathrm{S}}^2$ have normalized coordinates $(x, y \,|\, t)$ with $x^2 + y^2 - t^2 = -1$. The antipodal points $(0, 0 \,|\, 1)$ and $(0, 0 \,|\, {-1})$ are the "north pole" and "south pole" of $\ddot{\mathrm{S}}^2$. As with the elliptic sphere, a "meridian" is a great semicircle joining the poles, with points on the "prime" meridian having $y = 0$ and $x > 0$.

A point in either hemisphere of $\ddot{\mathrm{S}}^2$ can be given polar coordinates (p, θ), with $p \geq 0$ and $-\pi < \theta \leq \pi$. The *polaritude* p is the distance of the point from the pole in the same hemisphere, while the *longitude* θ of a nonpolar point is the directed angle from the prime meridian to the meridian of the point. The longitude of a pole is arbitrary. Antipodal points have the same polaritude and the same longitude. A point on the hyperbolic sphere with polar coordinates $(p,\ \theta)$ has normalized pseudospherical coordinates $(x, y \mid t)$, where

$$t = \pm\cosh p, \quad x = \sinh p\, \cos\theta, \quad y = \sinh p\, \sin\theta, \tag{3.3.9}$$

the sign of t depending on which hemisphere the point lies in. Since either hemisphere of $\ddot{\mathrm{S}}^2$ is an isometric model for H^2, polar coordinates $(p,\ \theta)$ can also be used for points in the hyperbolic plane.

EXERCISES 3.3

1. Show that in the "conformal disk" model of the elliptic plane, a line $[u_1,\ u_2,\ u_3]$ corresponds either to a circle with center $(-u_1/u_3,\ -u_2/u_3)$ and radius $\sqrt{u_1{}^2 + u_2{}^2 + u_3{}^2}/|u_3|$ or (if $u_3 = 0$) to the disk diameter with endpoints $(\pm u_2\sqrt{u_1{}^2 + u_2{}^2}, \mp u_1\sqrt{u_1{}^2 + u_2{}^2})$.

2. Show that in the "conformal disk" model of the hyperbolic plane, a line $[u_1,\ u_2 | u_0]$ corresponds either to a circle with center $(-u_1/u_0,\ -u_2/u_0)$ and radius $\sqrt{u_1{}^2 + u_2{}^2 - u_0{}^2}/|u_0|$ or (if $u_0 = 0$) to the disk diameter with endpoints , $(\pm u_2\sqrt{u_1{}^2 + u_2{}^2}, \mp u_1\sqrt{u_1{}^2 + u_2{}^2})$.

3. Find the point of H^2 corresponding to $(0, \sqrt{2} - 1)$ in the "conformal disk" model.

4. Verify the correspondences given in the text between Euclidean points $(x,\ y)$ and $(x',\ y')$ in the "conformal disk" and "upper half-plane" models of the hyperbolic plane.

5. Show that the mapping $(x,\ y) \mapsto (\xi,\ \eta)$ defined by (3.3.7) maps both the interior and the exterior of the unit circle onto the interior.

6. Find the points of E^2 in the projective-disk model for H^2 that correspond to the points $(0, 0)$, $(1/3, 0)$, $(1/2, 0)$, $(1/2, 1/2)$, and $(1/5, 2/5)$ in the inversive-disk model.

7. $\vdash$ The projective-disk model for H^2 is converted into the inversive-disk model by the mapping $(\xi,\ \eta) \mapsto (x,\ y)$, where

$$x = \frac{(1-\sqrt{1-\xi^2-\eta^2})\xi}{\xi^2+\eta^2}, \quad y = \frac{(1-\sqrt{1-\xi^2-\eta^2})\eta}{\xi^2+\eta^2}$$

if ξ and η are not both zero, with $(0,\ 0) \mapsto (0,\ 0)$.

8. $\vdash$ The distance between two points on S^2 with polar coordinates $(\phi_1,\ \theta_1)$ and $(\phi_2,\ \theta_2)$ is $\cos^{-1}[\cos\phi_1\ \cos\phi_2 + \sin\phi_1\ \sin\phi_2 \cos(\theta_2-\theta_1)]$.

3.4 TRIANGLES AND TRIGONOMETRY

The properties of Euclidean and spherical triangles have been studied since antiquity. Together with their elliptic and hyperbolic counterparts, these elementary geometric figures play a key role in the theory of metric spaces.

A. *Line segments and great arcs.* Any two points A and B in the affine plane A^2 or in the Euclidean or hyperbolic plane E^2 or H^2 lie on a unique line $\cdot AB\cdot$ and, by virtue of the relation of *intermediacy* (or "betweenness") introduced by Moritz Pasch in 1882, separate it into two infinite *rays* $\cdot A/B$ ("A away from B") and $\cdot B/A$ ("B away from A") and a finite *segment* AB (Coxeter 1942, p. 161), so that

$$\cdot AB\cdot = \cdot A/B \cup \{A\} \cup AB \cup \{B\} \cup \cdot B/A.$$

A ray is the same as an *open half-line*; the union $\cdot\overline{A}/B$ of a ray $\cdot A/B$ and its endpoint A is a *closed half-line*. A segment is the same as an *open interval*; the union $\overline{AB}$ of a segment AB and its endpoints A and B is a *closed interval*.

Two nonantipodal points A and B on the elliptic sphere S^2 lie on a unique great circle $\breve{C}$. The points A and B and their antipodes $\bar{A}$ and $\bar{B}$ separate $\breve{C}$ into four *proper* great arcs AB, $B\bar{A}$, $\bar{A}\bar{B}$, $\bar{B}A$—each shorter than a semicircle—with AB congruent to $\bar{A}\bar{B}$, $B\bar{A}$ congruent to $\bar{B}A$, and other pairs of supplementary lengths. The elliptic plane eP^2 is obtained by identifying antipodal points of S^2. Any two points A and B in eP^2 lie on a unique line $\cdot AB\cdot$, separating it into two segments AB and BA of supplementary lengths.

Any two ordinary points A and B in the same hemisphere of the hyperbolic sphere $\ddot{\mathrm{S}}^2$ lie on a unique great circle $\breve{C}$. The points A and B, together with their antipodes $\bar{A}$ and $\bar{B}$, separate $\breve{C}$ into two *proper* great arcs AB and $\bar{A}\bar{B}$, each lying entirely in one hemisphere, and two *improper* great arcs $A\bar{A}$ and $B\bar{B}$ crossing the equator. The hyperbolic plane H^2 is obtained by identifying antipodal points of $\ddot{\mathrm{S}}^2$, with antipodal pairs of proper great arcs projecting into segments and improper great arcs with antipodal ends projecting into rays.

B. *Metric triangles.* Any three noncollinear points A, B, and C in the affine plane A^2 or in the Euclidean or hyperbolic plane E^2 or H^2 define a unique *triangle* $\langle ABC\rangle$. With more structure than a projective triangle, which is just a configuration of three points and three lines, this is a set

$$\langle ABC\rangle = \{\varnothing; A, B, C; BC, CA, AB; ABC\},$$

partially ordered by dimensionality and incidence, consisting of eight geometric "entities": the *nullity* (the empty set $\varnothing$); three *vertices* (the points A, B, and C); three *sides* (the segments BC, CA, and AB); and the triangle's *body* (the interior region ABC).

Any three points A, B, and C on the elliptic sphere S^2 that do not lie on a great circle are the vertices of a unique triangle $\langle ABC\rangle$, each of whose sides BC, CA, and AB is a proper great arc and whose body ABC is contained within a hemisphere. The three points, together with their antipodes $\bar{A}$, $\bar{B}$, and $\bar{C}$, are the vertices of a tessellation of S^2 by eight spherical triangles, consisting of four antipodal pairs of congruent triangles, namely,

$$\langle ABC\rangle = \{\varnothing; A, B, C; BC, CA, AB; ABC\} \text{ and}$$
$$\langle \bar{A}\bar{B}\bar{C}\rangle = \{\varnothing; \bar{A}, \bar{B}, \bar{C}; \bar{B}\bar{C}, \bar{C}\bar{A}, \bar{A}\bar{B}; \bar{A}\bar{B}\bar{C}\},$$

$$\langle \bar{A}CB\rangle = \{\varnothing; \bar{A}, C, B; BC, B\bar{A}, \bar{A}C; \bar{A}CB\} \text{ and}$$

$$\langle A\bar{C}\bar{B}\rangle = \{\varnothing; A, \bar{C}, \bar{B}; \bar{B}\bar{C}, \bar{B}A, A\bar{C}; A\bar{C}\bar{B}\},$$

$$\langle C\bar{B}A\rangle = \{\varnothing; C, \bar{B}, A; \bar{B}A, CA, C\bar{B}; C\bar{B}A\} \text{ and}$$

$$\langle \bar{C}B\bar{A}\rangle = \{\varnothing; \bar{C}, B, \bar{A}; B\bar{A}, \bar{C}\bar{A}, \bar{C}B; \bar{C}B\bar{A}\},$$

$$\langle BA\bar{C}\rangle = \{\varnothing; B, A, \bar{C}; A\bar{C}, \bar{C}B, AB; BA\bar{C}\} \text{ and}$$

$$\langle \bar{B}\bar{A}C\rangle = \{\varnothing; \bar{B}, \bar{A}, C; \bar{A}C, C\bar{B}, \bar{A}\bar{B}; \bar{B}\bar{A}C\}.$$

Two triangles that have two vertices and one side in common, the other pair of vertices being antipodal and the other pairs of sides being supplementary—e.g., $\langle ABC\rangle$ and each of the three triangles $\langle \bar{A}BC\rangle$, $\langle C\bar{B}A\rangle$, $\langle BA\bar{C}\rangle$—are said to be a *colunar* pair. The region bounded by two great semicircles that is the union of the bodies and the common side of such a pair of triangles is called a *lune* (or "biangle").

In all cases, the body of a triangle is a bounded convex region, and in any of the metric spaces S^2, E^2, or H^2, each interior angle of a triangle has a radian measure between 0 and π. Triangles in the elliptic plane eP^2 also have these properties. Any three noncollinear points A, B, and C determine a tessellation of eP^2 by four triangles, any two of which are colunar (except that antipodal vertices are now coincident). All four triangles have the same vertices A, B, and C. If the sides of one triangle are labeled BC, CA, and AB and its body denoted by ABC—with the supplementary segments being CB, AC, and BA and the colunar regions being ACB, CBA, and BAC—then we have the triangles

$$\langle ABC\rangle = \{\varnothing; A, B, C; BC, CA, AB; ABC\},$$

$$\langle ACB\rangle = \{\varnothing; A, C, B; BC, BA, AC; ACB\},$$

$$\langle CBA\rangle = \{\varnothing; C, B, A; BA, CA, CB; CBA\},$$

$$\langle BAC\rangle = \{\varnothing; B, A, C; AC, CB, AB; BAC\}.$$

Three points A, B, and C in the same hemisphere of the hyperbolic sphere $\ddot{\mathrm{S}}^2$ that do not lie on a great circle and the three antipodal points $\bar{A}$, $\bar{B}$, and $\bar{C}$ are the vertices of an antipodal pair of triangles

$$\langle ABC\rangle = \{\varnothing; A, B, C; BC, CA, AB; ABC\} \text{ and}$$
$$\langle \bar{A}\bar{B}\bar{C}\rangle = \{\varnothing; \bar{A}, \bar{B}, \bar{C}; \bar{B}\bar{C}, \bar{C}\bar{A}, \bar{A}\bar{B}; \bar{A}\bar{B}\bar{C}\}.$$

Each side of either triangle is a proper great arc, and each of their bodies is a bounded convex region.

Because S^2, E^2, and H^2 are simply connected, any line or great circle separates the Euclidean or hyperbolic plane or the elliptic sphere into two half-planes or hemispheres. If A, B, and C are the vertices of a triangle, the (open) half-plane or hemisphere bounded by the line or great circle joining A and B that contains C may be denoted by $\cdot AB\cdot \uparrow C$ and the one that does not by $\cdot AB\cdot /C$. The body (interior) of $\langle ABC\rangle$ is then the intersection of $\cdot BC\cdot \uparrow A$, $\cdot CA\cdot \uparrow B$, and $\cdot AB\cdot \uparrow C$, and its exterior is the union of $\cdot BC\cdot /A$, $\cdot CA\cdot /B$, and $\cdot AB\cdot /C$.

The *area* of a triangle is the area of its body. If the interior angles at the vertices A, B, and C of a metric triangle $\langle ABC\rangle$ have the respective measures α, β, and γ, then the sum $\alpha + \beta + \gamma$ lies in the range 0 to 3π, with

$$\begin{aligned} 3\pi > \alpha + \beta + \gamma > \pi \quad & \text{for } \langle ABC\rangle \text{ in } \mathrm{S}^2 \text{ or } \mathrm{eP}^2;\\ \alpha + \beta + \gamma = \pi \quad & \text{for } \langle ABC\rangle \text{ in } \mathrm{E}^2;\\ 0 < \alpha + \beta + \gamma < \pi \quad & \text{for } \langle ABC\rangle \text{ in } \mathrm{H}^2 \text{ or } \ddot{\mathrm{S}}^2. \end{aligned} \tag{3.4.1}$$

The amount by which the angle sum $\alpha+\beta+\gamma$ exceeds or falls short of π for a non-Euclidean triangle $\langle ABC\rangle$ is its *spherical excess* or *hyperbolic defect* and (assuming a spatial curvature of ± 1) is equal to its area.

The area of a triangle is also a function of the lengths of its sides, which depend on the chosen unit of distance. Some writers treat non-Euclidean lengths and areas as if they were independent of each other and consequently arrive at mutually inconsistent formulas.

C. *The Triangle Inequality.* If the sides BC, CA, and AB of a metric triangle $\langle ABC\rangle$ have the respective lengths a, b, and c, then because lines in a plane or great circles on a sphere are geodesics and because ABC is a bounded convex region, it follows that for both Euclidean and non-Euclidean triangles we have the strict inequalities

$$b + c > a, \quad c + a > b, \quad a + b > c. \tag{3.4.2}$$

Moreover, most of the familiar Euclidean theorems about congruent triangles—in particular side-angle-side, angle-side-angle, and side-side-side—hold for non-Euclidean triangles as well. (However, angle-angle-side congruence does not generally hold either in the elliptic plane or on the elliptic sphere.) Since the interior angles of a non-Euclidean triangle determine not just its shape but its size, there is also an angle-angle-angle congruence theorem.

In the Euclidean and hyperbolic planes E^2 and H^2 and on the elliptic and hyperbolic spheres S^2 and $\ddot{\mathrm{S}}^2$, the length of a side AB of a triangle $\langle ABC\rangle$ is the same as the distance between the points A and B. This makes it easy to extend the Triangle Inequality

$$|AB| \leq |BC| + |CA|, \tag{3.4.3}$$

which holds for any three points A, B, and C in Euclidean n-space, to hyperbolic n-space and to the elliptic and hyperbolic n-spheres (in $\ddot{\mathrm{S}}^n$ the points must be restricted to the same n-hemisphere). If A, B, and C do not lie on a line or a great circle, they are the vertices of a unique triangle $\langle ABC\rangle$ and the conclusion follows from (3.4.2). Otherwise, we still have strict inequality except when C lies in the segment or proper great arc AB or, in the case of S^n, when A and B are antipodal, the only situations in which one distance equals the sum of the other two.

Because the total length of an elliptic line is π, the length of a side AB of a triangle in the elliptic plane eP^2 may be greater than the distance $[AB]$ between the points A and B (no distance can exceed $\pi/2$).

Consequently, the validity of the Triangle Inequality for three noncollinear points A, B, and C in eP^n does not immediately follow from the inequalities (3.4.2). Nevertheless, by considering triangles $\langle ABC\rangle$ with sides of various lengths less than, equal to, or greater than $\pi/2$, one can verify that $[AB] < [BC] + [CA]$ in every case. (I leave the details to the reader.) For three collinear points A, B, and C, either the sum of the three distances is π or else one of the distances is the sum of the other two; in either case, we have $[AB] \leq [BC] + [CA]$. Combining the two results, we find that the Triangle Inequality holds for any three points in elliptic n-space.* Thus distance in each of the geometries S^n, E^n, H^n, and eP^n has the essential properties that define a metric space.

D. *Trigonometry.* How the sides and the angles of a triangle are related depends on the curvature of the underlying space. For example, in a Euclidean triangle $\langle ABC\rangle$, the side lengths a, b, c and the corresponding angles α, β, γ satisfy the Law of Sines

$$\frac{\sin\alpha}{a} = \frac{\sin\beta}{b} = \frac{\sin\gamma}{c}. \tag{3.4.4}$$

On the elliptic sphere S^2 or in the elliptic plane eP^2, the formula is

$$\frac{\sin\alpha}{\sin a} = \frac{\sin\beta}{\sin b} = \frac{\sin\gamma}{\sin c}, \tag{3.4.5}$$

while on the hyperbolic sphere $\ddot{\mathrm{S}}^2$ or in the hyperbolic plane H^2 it is

$$\frac{\sin\alpha}{\sinh a} = \frac{\sin\beta}{\sinh b} = \frac{\sin\gamma}{\sinh c}. \tag{3.4.6}$$

---

* Two elliptic triangles whose corresponding sides have equal lengths are congruent. However, two triangles having equal distances between corresponding pairs of vertices need not be. Moreover, if A, B, and C are three points on a line such that none of the three distances $[BC]$, $[CA]$, and $[AB]$ is equal to the sum of the other two, then the three segments into which they separate the line can have the same lengths as the sides of a triangle, and the three distances can be equal to the distances between pairs of vertices of a triangle.

Two sides and the included angle of a Euclidean triangle $\langle ABC\rangle$ are related to the third side by the Law of Cosines

$$c^2 = a^2 + b^2 - 2ab\cos\gamma. \tag{3.4.7}$$

When the angle at C is a right angle, this reduces to the Pythagorean Theorem. On the elliptic sphere S^2 or in the elliptic plane eP^2, we have the dual formulas

$$\begin{aligned} \cos c &= \quad \cos a\,\cos b + \sin a\,\sin b\,\cos\gamma, \\ \cos\gamma &= -\cos\alpha\,\cos\beta + \sin\alpha\,\sin\beta\,\cos c. \end{aligned} \tag{3.4.8}$$

On the hyperbolic sphere $\ddot{\mathrm{S}}^2$ or in the hyperbolic plane H^2, the formulas are

$$\begin{aligned} \cosh c &= \cosh a\,\cosh b - \sinh a\,\sinh b\,\cos\gamma, \\ \cos\gamma &= -\cos\alpha\,\cos\beta + \sin\alpha\,\sin\beta\,\cosh c. \end{aligned} \tag{3.4.9}$$

In the hyperbolic plane we can also consider *asymptotic* triangles, with one or more absolute vertices, at which the interior angle is zero. Setting $\beta = \pi/2$ and $\gamma = 0$ in the last formula, we obtain the expression $\sin^{-1}\operatorname{sech} c$ for the angle of parallelism $\Pi(c) = \alpha$. The area of an asymptotic triangle is still finite and equal to the hyperbolic defect, so that a "tricuspal" triangle, with three absolute vertices, has an area of π.

Two Euclidean triangles whose corresponding angles are equal have the lengths of their sides proportional: the triangles are similar but not necessarily congruent. In the elliptic or hyperbolic plane, however, equality of angles implies equality of side lengths. Consequently, it is sometimes said that scale models are impossible in non-Euclidean geometry, but this is not entirely true. A spherical triangle with shorter sides can have the same angles as a plane triangle in elliptic 3-space, and a pseudospherical triangle with longer sides can have the same angles as a plane triangle in hyperbolic 3-space.

The n-dimensional analogue of a triangle, a polytope with $n + 1$ vertices and $n + 1$ sides (or "facets"), is called a *simplex*. A 3-simplex is a *tetrahedron*. The properties of simplexes and other polytopes will be discussed further in later chapters.

EXERCISES 3.4

1. The intermediacy relation "point X lies between points A and B" may be abbreviated as $/AXB/$. Extending the affine (or Euclidean) line $\cdot AB\cdot$ by the point at infinity P_∞, define $/AXB/$ in terms of separating pairs of points. Express the line segment AB and the rays $\cdot A/B$ and $\cdot B/A$ as segments of the circuit $\cdot AB\cdot \cup \{P_\infty\}$.

2. $\vdash$ The area of a spherical lune of angle θ is 2θ.

3. Show that the inequality $[AB] < [BC] + [CA]$ holds for any three noncollinear points A, B, and C in eP^2. (Consider triangles with sides of lengths less than, equal to, or greater than $\pi/2$.)

4. Show that the inequality $[AB] \leq [BC]+[CA]$ holds for any three collinear points A, B, and C in eP^2.

5. The area Δ of a Euclidean triangle with sides a and b and included angle γ is given by $\Delta = \frac{1}{2}ab \sin \gamma$. Combine this formula with the Law of Cosines (3.4.7) to obtain Heron's Formula $\Delta = \sqrt{s(s-a)(s-b)(s-c)}$, where $s = \frac{1}{2}(a+b+c)$ is the triangle's semiperimeter.

6. Find the area of a hyperbolic triangle having two angles of $\pi/4$ adjacent to a side of length $\cosh^{-1} 2$.

7. Prove the identity $\sin^{-1} \operatorname{sech} x = \cos^{-1} \tanh x$.

3.5 NON-EUCLIDEAN CIRCLES

The methods of integral calculus can be used to determine the arc lengths of portions of smooth curves and the areas of regions with nonrectilinear boundaries. In this section we describe some of the metric properties of circles in different geometries.

In the Euclidean plane, a circle with center (h, k) and radius r has equation $(x-h)^2 + (y-k)^2 = r^2$. Given a circle of radius r, an arc subtending a central angle of radian measure θ has length $s = r\theta$, so that the circle's circumference is $2\pi r$. The corresponding circular sector has area $\frac{1}{2}rs$, making the area of the whole circle equal to πr^2.

In the elliptic plane eP^2, a circle with center $\mathbf{(}h\mathbf{)} = (h_1,\ h_2,\ h_3)$ and radius r or, equivalently, with axis $\mathbf{[}h\mathbf{]} = [h_1,\ h_2,\ h_3]$ and coradius $\check{r} = \pi/2 - r$, has equation

$$|\mathbf{(}x\ h\mathbf{)}| = \cos r \quad \text{or} \quad |\mathbf{(}x\mathbf{)[}h\mathbf{]}| = \sin \check{r}, \tag{3.5.1}$$

where $\mathbf{(}x\ h\mathbf{)} = \mathbf{(}x\mathbf{)[}h\mathbf{]} = x_1h_1+x_2h_2+x_3h_3$ and where $\mathbf{(}x\ x\mathbf{)} = \mathbf{(}h\ h\mathbf{)} = 1$. The length of an arc of a circle of radius r subtending a central angle or axial distance of θ is $s = \theta \sin r = \theta \cos \check{r}$; the circumference is thus seen to be $2\pi \sin r = 2\pi \cos \check{r}$. A circular sector $V(r,\ s)$ of radius r and arc length s has area

$$|V(r,s)| = s \tan \tfrac{1}{2}r \quad (0 < r < \pi/2). \tag{3.5.2}$$

The area of the whole circle is $4\pi \sin^2 \tfrac{1}{2}r$. Complementary to such a sector is a *cosector* $U(\check{r},\ s)$ of coradius $\check{r}$ and arc length s, whose area is

$$|U(\check{r},s)| = s \tan \check{r} = \theta \sin \check{r} \quad (0 < \check{r} < \pi/2). \tag{3.5.3}$$

The sector and cosector together form a triangle with two right angles, a *great sector* of radius $\pi/2$ (see Figure 3.5a). The combined area is equal to the central angle or axial distance θ. Thus when $\theta = 1$, the

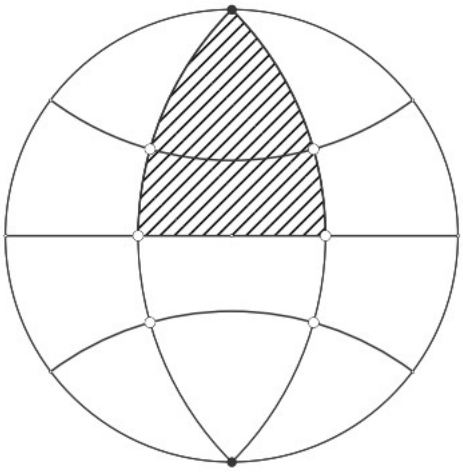

Figure 3.5a Circular sector and cosector in eP^2

area is 1. The above formulas also apply to small circles on the elliptic sphere S^2.

In the hyperbolic plane H^2, a pericycle—i.e., a proper circle—with center $\mathbf{(}h\mathbf{)} = (h_1,\ h_2 \,|\, h_0)$ and radius r has equation

$$|\mathbf{(}x\ddot{\ }h\mathbf{)}| = \cosh r, \tag{3.5.4}$$

where $\mathbf{(}x\ddot{\ }h\mathbf{)} = x_1h_1 + x_2h_2 - x_0h_0$, with coordinates normalized so that $\mathbf{(}x\ddot{\ }x\mathbf{)} = \mathbf{(}h\ddot{\ }h\mathbf{)} = -1$. An arc of such a pericycle subtending a central angle of θ has length $s = \theta\ \sinh r$, and the circumference is therefore $2\pi\ \sinh r$. A pericyclic sector $\ddot{V}(r,s)$ of radius r and arc length s (see Figure 3.5b) has area

$$|\ddot{V}(r,s)| = s\ \tanh \tfrac{1}{2}r \quad (r > 0). \tag{3.5.5}$$

The area of the whole pericycle is $4\pi\ \sinh^2 \tfrac{1}{2}r$.

A pseudocycle (equidistant curve) with axis $\mathbf{[}m\mathbf{]} = [m_1,\ m_2 \,|\, m_0]$ and coradius $\check{r}$ has equation

$$|\mathbf{(}x\mathbf{)[}m\mathbf{]}| = \sinh \check{r}, \tag{3.5.6}$$

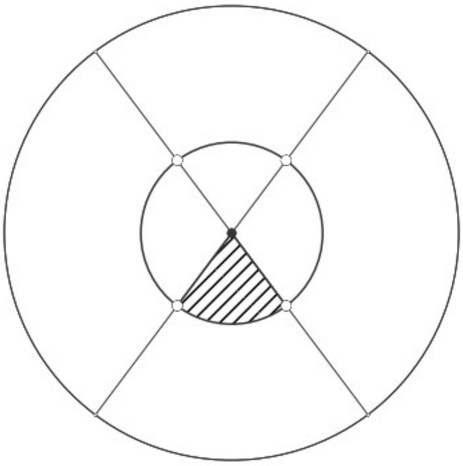

Figure 3.5b Pericyclic sector in H^2

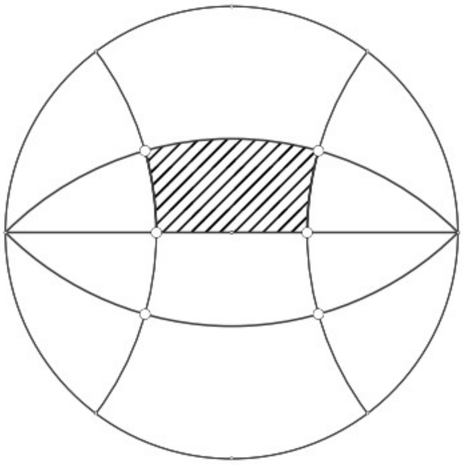

Figure 3.5c Pseudocyclic cosector in H^2

where $\mathbf{(}x\mathbf{)[}m\mathbf{]} = x_0m_0 + x_1m_1 + x_2m_2$, with coordinates normalized so that $\mathbf{(}x\ddot{\ }x\mathbf{)} = -1$ and $\mathbf{[}m\ddot{\ }m\mathbf{]} = 1$. An arc of such a pseudocycle subtending an axial distance of t has length $s = t \cosh \check{r}$. A pseudocyclic cosector $\ddot{U}(\check{r}, s)$ of coradius $\check{r}$ and arc length s (Figure 3.5c) has area

$$|\ddot{U}(\check{r}, s)| = s \tanh \check{r} = t \sinh \check{r} \quad (\check{r} > 0). \tag{3.5.7}$$

When the arc of a pseudocyclic cosector is replaced by the chord with the same endpoints, the resulting figure is the kind of quadrilateral used by Gerolamo Saccheri (1667–1733) in his attempted proof of Euclid's Fifth Postulate.

Whereas a pericycle has an ordinary point for its center and a pseudocycle has an ordinary line for its axis, the center of a horocycle (limiting curve) is an absolute point, and its axis is the absolute line through this point. Thus a horocycle cannot have an equation analogous to (3.5.4) or (3.5.6). However, all horocycles are congruent, so that, following Coxeter (1942, p. 250), we can deduce their metric properties from a conveniently chosen particular case.

The three points

$$P_\infty = (0,\ 1\,|\,1), \quad P_0 = (0,\ 0\,|\,1), \quad \text{and} \quad P_1 = (1, \tfrac{1}{2}\,|\,\tfrac{3}{2})$$

determine a horocycle whose center is the absolute point P_∞ and whose ordinary points are, for each real number x,

$$P_x = \left(x, \tfrac{1}{2}x^2\,\middle|\,\tfrac{1}{2}x^2 + 1\right).$$

The radial line joining P_x to the center and the tangent line at P_x are respectively

$$\check{R}_x = \left[1, x\,\middle|\,-x\right] \quad \text{and} \quad \check{T}_x = \left[x, \tfrac{1}{2}x^2 - 1\,\middle|\,-\tfrac{1}{2}x^2\right].$$

The horocyclic arc with endpoints P_x and P_y has length $s = |y - x|$, and a horocyclic sector $\ddot{W}(s)$ of arc length s (see Figure 3.5d) has area s. Thus the arc from P_0 to P_1 has length 1 and the corresponding sector has area 1. Moreover, the radial line at one end of a horocyclic arc of unit length is parallel to the tangent line at the other end. Thus $\check{R}_0$ is parallel to $\check{T}_1$ and $\check{R}_1$ is parallel to $\check{T}_0$.

The foregoing results for pericycles, pseudocycles, and horocycles also hold for small circles on the hyperbolic sphere $\ddot{\mathrm{S}}^2$.

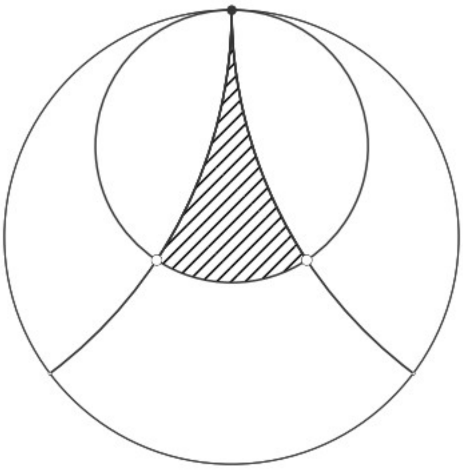

Figure 3.5d Horocyclic sector in H^2

EXERCISES 3.5

1. $\vdash$ The area of the elliptic plane is 2π.
2. Find the length of the chord of a pericycle of radius $\sinh^{-1} 2$ joining the ends of an arc of length π.
3. $\vdash$ The tangent lines at the ends of a horocyclic arc of unit length intersect at an angle of $\pi/3$.
4. Show that the isosceles asymptotic triangle $\langle P_0 P_1 P_\infty \rangle$ has area $\pi - 2 \tan^{-1} 2$.

3.6 SUMMARY OF REAL SPACES

The following diagram (Figure 3.6a) exhibits the relationships among the various real linear and circular geometries. A line joining two geometries indicates that the one below can be derived from the one above by selecting a quadric; by fixing an inversion, a polarity, or a hyperplane; or (for geometries joined by a double arrow) by identifying antipodal points.

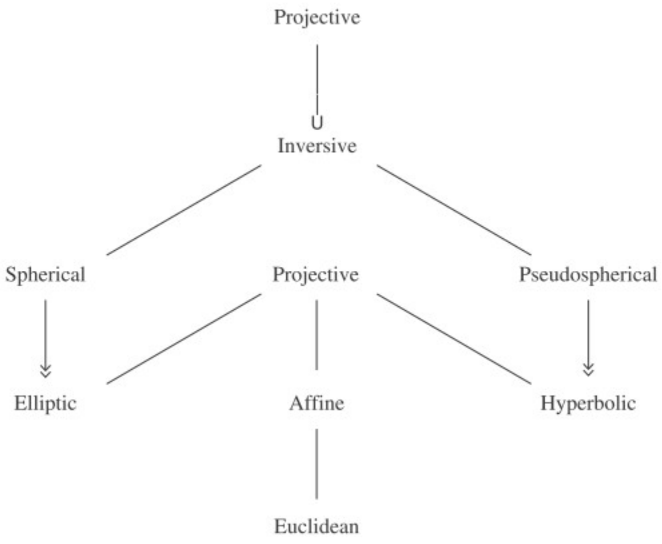
Projective
∪
Inversive
Spherical
Projective
Pseudospherical
Elliptic
Affine
Hyperbolic
Euclidean

Figure 3.6a Linear and circular geometries

Pág. 174(192) → Richard Hamilton

—

— Geometries and transformations (2018).pdf

Pág. 57(75) Chapter 3 Circular Geometries

Pág. 58(76) → A. Inversive geometry.

A projective hyperplane $\check{U}$, with coordinates $[\![u]\!]$, meets Φ in an inversive $(n-1)$-sphere, i.e., a hypersphere of I^n, which is real if $[\![u\,u]\!] > 0$, degenerate if $[\![u\,u]\!] = 0$, imaginary if $[\![u\,u]\!] < 0$, where $[\![u\,u]\!]$ is the quadratic form associated with the bilinear form

$$[\![u\,v]\!] = -u_0v_0 + u_1v_1 + \ldots + u_{n+1}v_{n+1}.$$

—

Pág. 71 (89) - Gaussian curvature
→ Spherical and pseudospherical geometry.

→ hypersphere of radius ri in Lorentzian (n+1)-space.

Books - 018 - 2021 / Importantíssimo /
- SuperPrinciple Relativity_Part_1.pdf
- SuperPrinciple Relativity_Part_2.pdf
Pág. 6 - Ricci Curvature
"Time-Like" - "Inverse Square Distance" Equation

Look: $\left(\frac{8\pi \cdot G}{c^4_{light}}\right)$

adequar.

- SuperPrinciple Relativity_Part_3.pdf
Pág. 16 1.5
The Gravitational Field
Metric Coefficient
($\Sigma g_{\mu\nu}$),

+,-.+ + +/ ! + "' 3 3 - ! 8 2: ! : 3#

+,-.+ + +/ ! + "' 3 3 - ! 8 2: ! : 3#

' +,-.+ + +/ ! + "' 3 3 - ! 8 2: ! :' 3#

/4 ( ) - 8!

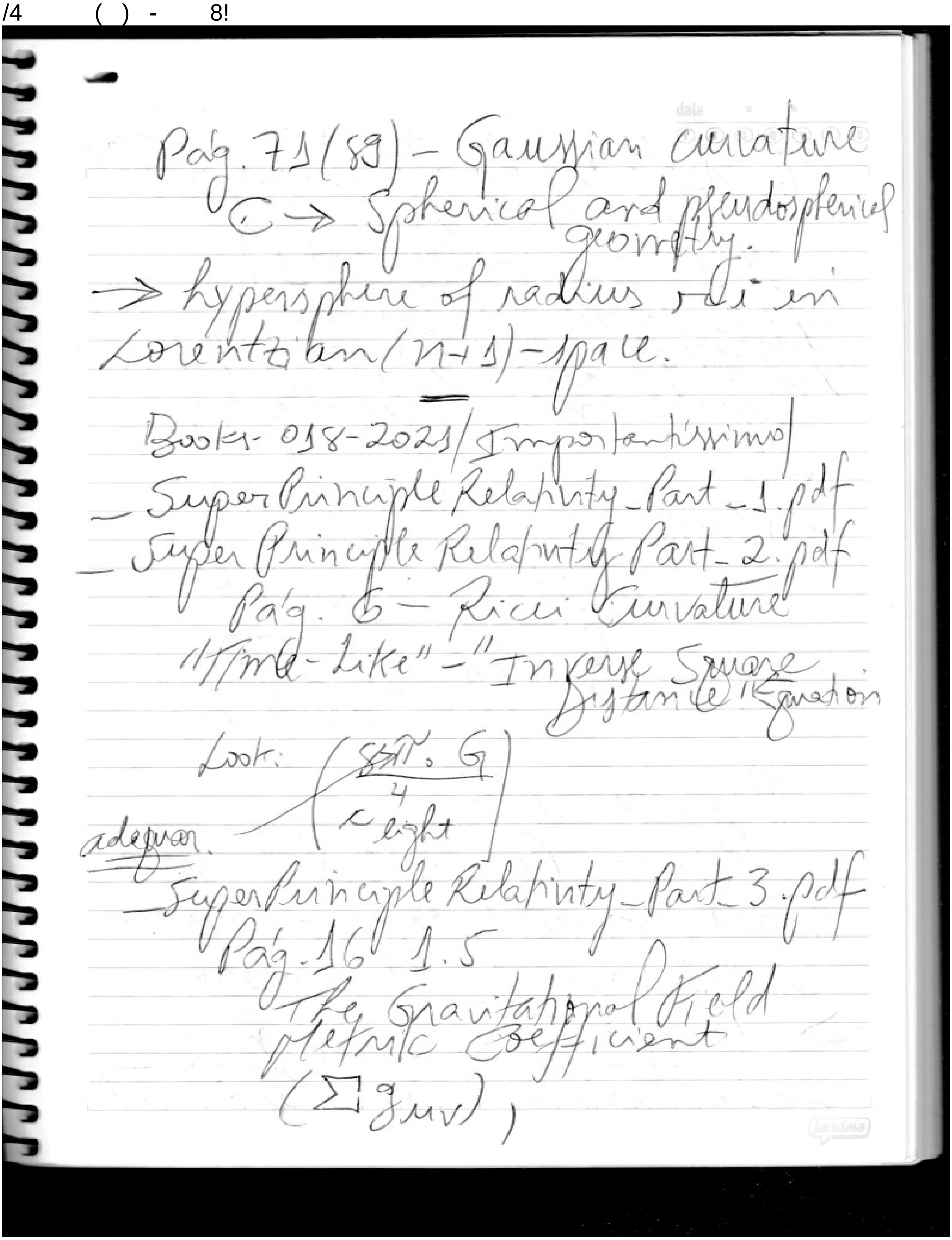

Pag. 71 (89) – Gaussian curvature
→ Spherical and pseudospherical geometry.
→ hypersphere of radius r i in Lorentzian (n+1)-space.

Books - 018 - 2021 / Importantíssimo /
– SuperPrinciple Relativity_Part_1.pdf
– SuperPrinciple Relativity_Part_2.pdf
Pág. 6 – Ricci Curvature
"Time-Like" – "Inverse Square Distance" Equation

Look: $\left(\frac{8\pi \cdot G}{c^4_{light}}\right)$

adequar.

– SuperPrinciple Relativity_Part_3.pdf
Pág. 16 1.5
The Gravitational Field Metric Coefficient
$(\sum g_{\mu\nu})$,

( )*

is a unit-less quantity that can be described as a determinant of a series of independent "metric coefficients" which describes the coordinates or the dimensions of a frame of reference in a localized space, space-time, or gravitational field.

$$\sum g_{\mu\nu} = \begin{pmatrix} g_{11} & g_{12} & g_{13} & g_{14} \\ g_{21} & g_{22} & g_{23} & g_{24} \\ g_{31} & g_{32} & g_{33} & g_{34} \\ g_{41} & g_{42} & g_{43} & g_{44} \end{pmatrix}$$

Pág. 28 1.8

Super Principle Relativity Theory - Hydrodynamic - Einstein Gravitational Field Equations

⇒ 1.78 → Einstein "Dynamic" Gravitational Field Equatio — Local Volume Expansions of

Space-Time Equation –
"Inverse Square Distance" Equation
(1.78)

$$\left(\frac{8\overset{\downarrow}{\pi}\cdot G}{c_{light}^{4}}\right)\cdot T(A,t)\,\psi \rightarrow 1/m^2$$

20210916$_{III}$

Google Workspace Enterprise Standard

---

math2.fr/geometrie-projective/

quicklatex.com

www.imo.universite-paris-saclay.fr/~perrin/Livre_de_geometrie_projective.html

---

(.SVG Scalable Vectorial Graphics

Convertio.co
or
svgtopng.com

') ( ))
! +,-.+ + +/ ! + ') ; 85 1 2 < = & 3#
! +,-.+ + +/ ! + ') ; 85 1 2 < = && 3#
/4
! (% 0 ! >
! (' % 0
! ( ) > ! # !

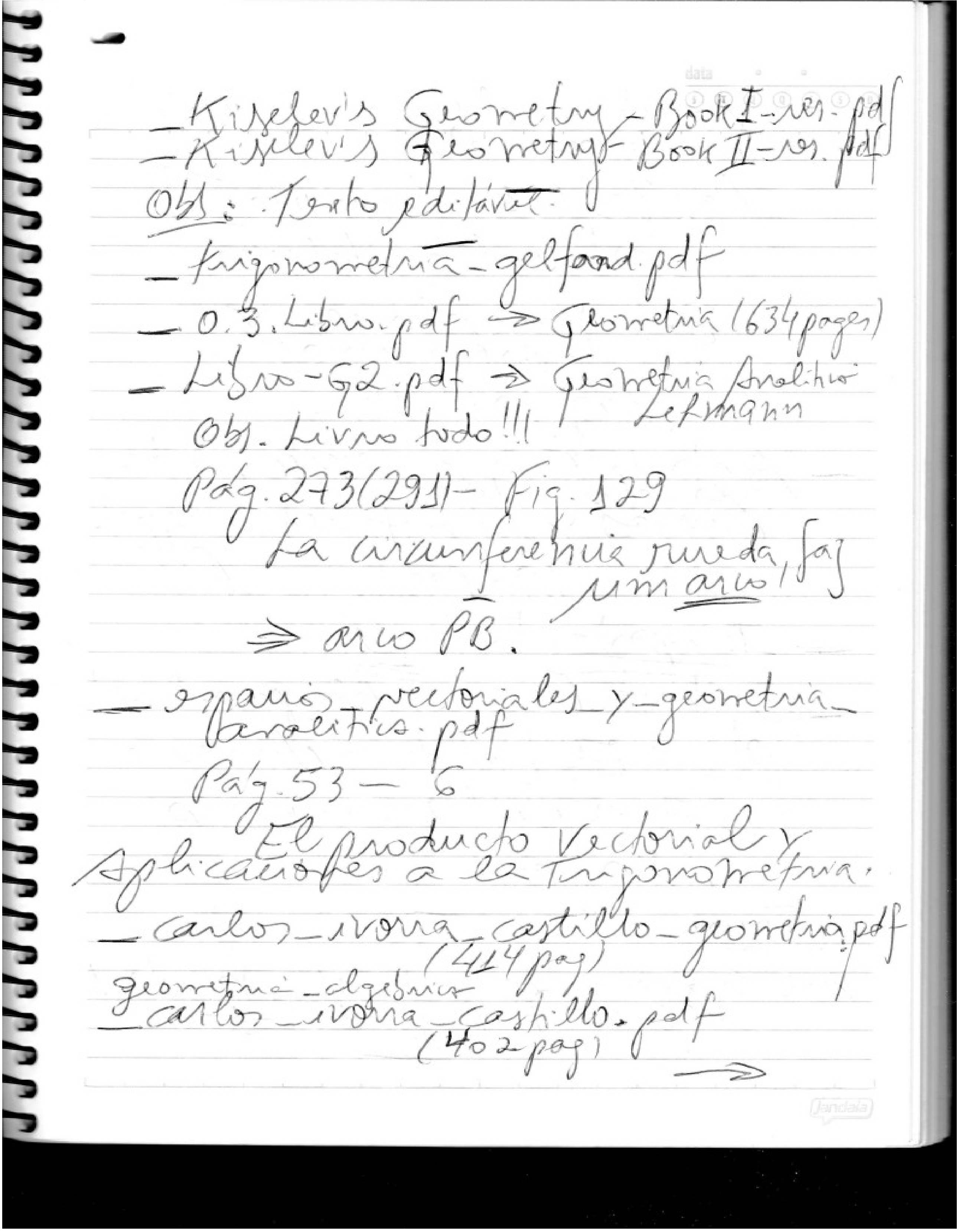

- Kiselev's Geometry - Book I - res. pdf
- Kiselev's Geometry - Book II - res. pdf

Obs: Texto editável.

- trigonometria_gelfand.pdf
- 0.3.Libro.pdf ⇒ Geometria (634 pages)
- Libro-G2.pdf ⇒ Geometria Analítica Lehmann

Obs. Livro todo!!!

Pág. 273(291) - Fig. 129

La circunferencia rueda, faz um arco!

⇒ arco PB.

- espacios_vectoriales_y_geometria_analiticas.pdf

Pág. 53 - 6

El producto vectorial y Aplicaciones a la Trigonometría.

- carlos_ivorra_castillo_geometria.pdf (414 pag)
- geometria_algebraica_carlos_ivorra_castillo.pdf (402 pag)

→

( )
4 +,-.+ + +/ ! + ') !: ! # 3#
+,-.+ + +/ ! + ') ' > 4 3#
+,-.+ + +/ ! + ') > 4 1 3#
(' % ? %

y las circunferencias concéntricas de radios a y b se llaman, respectivamente, *círculo principal* y *círculo menor* de la elipse.

Una representación paramétrica sencilla de la hipérbola puede obtenerse como sigue. Tracemos dos circunferencias concéntricas que tengan su centro común en el origen y que sus radios sean $OA = a$ y $OB = b$, en que $a > b$, como se ve en la figura 128. A partir de O tracemos una recta cualquiera l que forme un ángulo θ con la parte positiva del eje X, y sea C el punto de intersección con la circunferencia de radio a. En C tracemos la tangente a la circunferencia; designemos por D el punto en que esta tangente corta al eje X. En B tracemos una perpendicular al eje X y sea E su punto de intersección con l. Por D y E tracemos rectas paralelas a los ejes Y y X, respectivamente; designemos por P el punto de intersección de estas rectas. Ahora vamos a obtener las ecuaciones paramétricas del lugar geométrico de $P(x, y)$, usando θ como parámetro. De los triángulos rectángulos OCD y OBE, tenemos

$$x = \overline{OD} = \overline{OC} \sec \theta = a \sec \theta$$

y

$$y = \overline{DP} = \overline{BE} = \overline{OB} \operatorname{tg} \theta = b \operatorname{tg} \theta .$$

Por tanto, las ecuaciones paramétricas del lugar geométrico de P son

$$x = a \sec \theta , \quad y = b \operatorname{tg} \theta , \tag{6}$$

y la ecuación rectangular puede hallarse fácilmente y es (véase el ejemplo 1 del Artículo 90)

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1 . \tag{7}$$

Por tanto, las ecuaciones (6) son una representación paramétrica de la hipérbola (7). El parámetro θ se llama *ángulo excéntrico* del punto P, y el círculo de radio a se llama *círculo auxiliar* de la hipérbola.

93. La cicloide. Sea P un punto cuya posición sea fija con relación a una curva C. Si la curva C rueda, sin resbalar, sobre una curva fija C', el lugar geométrico descrito por el punto P se llama *ruleta*.

Un caso importante de ruleta es la curva llamada *cicloide*. Una cicloide es el lugar geométrico descrito por cualquier punto fijo de una circunferencia que rueda, sin resbalar, sobre una recta fija. Deduciremos las ecuaciones paramétricas de la cicloide tomando la recta fija como eje X y una de las posiciones del punto móvil sobre el eje X como origen. Sea $P(x, y)$ un punto cualquiera del lugar geométrico, a el radio y C el centro de la circunferencia que rueda, como se indica en la figura 129. Tomaremos como parámetro el ángulo θ que gira la circunferencia al rodar partiendo de su posición inicial en el origen. Sean A y B, respectivamente, los pies de las perpendiculares bajadas de P y C al eje X. Tracemos PD perpendicular a BC. Como la circunferencia rueda, sin resbalar, desde O hasta B, tenemos

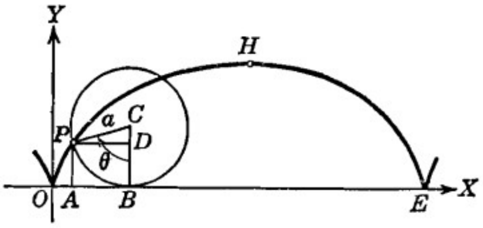

Fig. 129

$$\overline{OB} = \text{arco } \overline{PB}.$$

Si θ se mide en radianes, tenemos (Apéndice IC, 4)

$$\text{arco } \overline{PB} = a\theta .$$

Por tanto, de la figura 129,

$$x = \overline{OA} = \overline{OB} - \overline{AB} = a\theta - \overline{PD} = a\theta - a \operatorname{sen} \theta ,$$
$$y = \overline{AP} = \overline{BD} = \overline{BC} - \overline{DC} = a - a \cos \theta ,$$

de manera que las ecuaciones paramétricas de la cicloide son

$$x = a(\theta - \operatorname{sen} \theta), \quad y = a(1 - \cos \theta). \qquad (1)$$

Por el método empleado en el ejemplo del Artículo 91, podemos demostrar que la ecuación rectangular de la cicloide (1) es

$$x = a \text{ arc cos } \frac{a-y}{a} \mp \sqrt{2ay - y^2}, \qquad (2)$$

en donde debe tomarse el signo positivo o el negativo según que θ sea menor o mayor que π radianes en el arco comprendido entre $\theta = 0$ y $\theta = 2\pi$.

El punto medio H de cualquier arco de la cicloide se llama *vértice* del arco. Aquella porción OE de la recta fija comprendida entre los puntos extremos de un arco se llama ***base*** del arco; su longitud es, evidentemente, igual a $2\pi a$, que es la longitud de la circunferencia generatriz. Cada extremo de un arco, tal como O y E, se llama *pico* o *cúspide*.

A la cicloide también se le da a veces el nombre de *braquistocrona* o curva del más rápido descenso, porque, si se invierte la curva de la figura 129, se puede demostrar que es el recorrido descrito por una partícula que cae desde un punto dado a otro en el intervalo de tiempo *mínimo*. Además, si se sueltan dos partículas simultáneamente desde dos puntos cualesquiera del arco invertido de una cicloide, llegarán ambas al punto más bajo (el vértice) al *mismo* tiempo.

La cicloide es un caso especial de la ruleta conocida con el nombre de *trocoide*, que es el lugar geométrico descrito por un punto de un radio fijo de una circunferencia que rueda, sin resbalar, sobre una recta. Si el punto generador $P(x, y)$ está a una distancia b del centro del círculo rodante de radio a, si una posición del radio fijo es a lo largo del eje Y, y si la recta fija se toma como el eje X, puede demostrarse que las ecuaciones paramétricas de la trocoide son

$$x = a\theta - b \operatorname{sen} \theta, \quad y = a - b \cos \theta. \tag{3}$$

Se dice de la trocoide que es una *cicloide acortada* o *alargada* según que

$$b < a \text{ o } b > a.$$

Para $b = a$, las ecuaciones (3) se reducen a las ecuaciones paramétricas (1) de la cicloide.

94. Epicicloide e hipocicloide. Ahora consideremos dos tipos de ruletas que difieren de la cicloide en que la curva fija es una circunferencia en vez de una recta. Estas curvas, llamadas epicicloide e hipocicloide, son importantes en el diseño de dientes de engranajes.

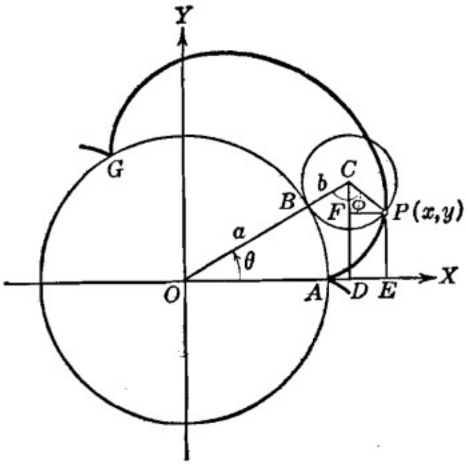

Fig. 130

Una *epicicloide* es el lugar geométrico descrito por un punto fijo cualquiera de una circunferencia que rueda exteriormente, sin resbalar, sobre una circunferencia fija. Deduciremos las ecuaciones paramétricas de la epicicloide en el caso en que la circunferencia fija tenga su centro en el origen y una posición del punto que describe la curva está sobre la parte positiva del eje X y sobre la circunferencia fija. Sea $P(x, y)$ un punto cualquiera del lugar geométrico; sean a y b, respectivamente, los radios de las circunferencias fija y rodante, y sea C el centro de la circunferencia rodante o generatriz, como se ve en la figura 130. Tomaremos como parámetro el ángulo θ que forma la recta de los centros OC con la parte positiva del eje X. Sea A el punto sobre el eje X que representa la posición inicial del punto P que describe la curva, y sea B el punto de tangencia de las dos circunferencias. Desde C y P bajemos las perpendiculares CD y PE, respectivamente, al eje X, y tracemos PF perpendicular a CD. Llamemos ϕ al ángulo OCP y β al ángulo PCF. Consideraremos ambos ángulos ϕ y θ medidos en radianes.

Como la circunferencia generatriz rueda, sin resbalar, de A a B, tenemos

$$\text{arco } AB = \text{arco } PB,$$

o sea,

$$a\theta = b\phi.$$

Por tanto, $\phi = \frac{a}{b}\theta$ y $\theta + \phi = \theta + \frac{a}{b}\theta = \frac{a+b}{b}\theta$. Tenemos, también,

$$\beta = \phi - \text{ángulo } OCD = \phi - \left(\frac{\pi}{2} - \theta\right) = \theta + \phi - \frac{\pi}{2}.$$

Por tanto,

$$\begin{aligned}\operatorname{sen}\beta = \operatorname{sen}\left(\theta + \phi - \frac{\pi}{2}\right) &= -\operatorname{sen}\left(\frac{\pi}{2} - [\theta + \phi]\right)\\ &= -\cos(\theta + \phi) = -\cos\frac{a+b}{b}\theta,\end{aligned}$$

y

$$\begin{aligned}\cos\beta = \cos\left(\theta + \phi - \frac{\pi}{2}\right) &= \cos\left(\frac{\pi}{2} - [\theta + \phi]\right)\\ &= \operatorname{sen}(\theta + \phi) = \operatorname{sen}\frac{a+b}{b}\theta.\end{aligned}$$

Para las coordenadas (x, y) del punto P, tenemos:

$$\begin{aligned} x &= \overline{OE} = \overline{OD} + \overline{DE} = \overline{OD} + \overline{FP} = \overline{OC} \cos\theta + \overline{CP} \operatorname{sen}\beta \\ &= (a+b)\cos\theta - b\cos\frac{a+b}{b}\theta, \\ y &= \overline{EP} = \overline{DF} = \overline{DC} - \overline{FC} = \overline{OC} \operatorname{sen}\theta - \overline{CP}\cos\beta \\ &= (a+b)\operatorname{sen}\theta - b\operatorname{sen}\frac{a+b}{b}\theta, \end{aligned}$$

de manera que las ecuaciones paramétricas de la epicicloide son

$$\left.\begin{aligned} x &= (a+b)\cos\theta - b\cos\frac{a+b}{b}\theta, \\ y &= (a+b)\operatorname{sen}\theta - b\operatorname{sen}\frac{a+b}{b}\theta. \end{aligned}\right\} \qquad (1)$$

Cada punto de la epicicloide que está sobre la circunferencia fija, tales como A y G, es un *pico;* la porción de curva comprendida entre dos picos sucesivos se llama *arco*. El número de picos y arcos depende de las magnitudes relativas de los radios a y b. Sea r la razón de a a b, de manera que $a = rb$. Si r es un número entero, la epicicloide será, evidentemente, una curva cerrada que tiene exactamente r picos y r arcos; se dice entonces que la curva es una *epicicloide de* r *picos*. Si r no es un número entero pero es racional, el punto trazador P dará la vuelta en torno de la circunferencia fija dos o más veces antes de regresar al punto de partida A; en este caso, los arcos de la curva de diferentes circuitos se cortarán. Si r es irracional, el punto trazador no regresa exactamente al punto de partida.

Cuando $a = b$, de manera que $r = 1$, tenemos la epicicloide de un pico o *cardioide* (véase el ejemplo 1 del Artículo 82 y el ejercicio 21 del grupo 41, Artículo 88). De las ecuaciones (1) se deducen las siguientes ecuaciones paramétricas de la cardioide:

$$x = 2a\cos\theta - a\cos 2\theta, \quad y = 2a\operatorname{sen}\theta - a\operatorname{sen} 2\theta. \qquad (2)$$

Una *hipocicloide* es el lugar geométrico de un punto fijo cualquiera de una circunferencia que rueda *interiormente*, sin resbalar, sobre otra circunferencia fija. Por un procedimiento semejante al empleado para la epicicloide, podemos demostrar que las ecuaciones paramétricas de la hipocicloide son

$$\left.\begin{aligned} x &= (a-b)\cos\theta + b\cos\frac{a-b}{b}\theta, \\ y &= (a-b)\operatorname{sen}\theta - b\operatorname{sen}\frac{a-b}{b}\theta, \end{aligned}\right\} \qquad (3)$$

en donde a y b son, respectivamente, los radios de las circunferencias fija y rodante, y el parámetro θ es el ángulo que la recta de los centros OC forma con la parte positiva del eje X, tal como puede verse en la figura 131. El lector debe observar que las ecuaciones paramétricas (3) de la hipocicloide pueden obtenerse reemplazando b por $-b$ en las ecuaciones paramétricas (1) de la epicicloide.

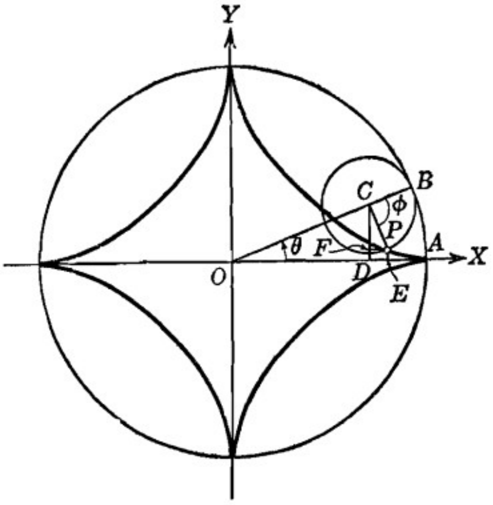

Fig. 131

Sea r la razón de a a b, de modo que $a = rb$. Si r es un número entero, tenemos una *hipocicloide de* r *picos*. La *hipocicloide de cuatro picos* está representada en la figura 131; esta curva se llama también *astroide*. Las ecuaciones paramétricas de la astroide pueden simplificarse de manera que tomen una forma muy simple. Así, para $b = \frac{a}{4}$, las ecuaciones paramétricas (3) se convierten en

$$\left.\begin{aligned} x &= \frac{3a}{4}\cos\theta + \frac{a}{4}\cos 3\theta, \\ y &= \frac{3a}{4}\operatorname{sen}\theta - \frac{a}{4}\operatorname{sen} 3\theta. \end{aligned}\right\}$$

Si en estas ecuaciones sustituímos los valores de $\cos 3\theta$ y $\operatorname{sen} 3\theta$ dados por las identidades trigonométricas

$$\cos 3\theta = 4\cos^3\theta - 3\cos\theta,$$

$$\operatorname{sen} 3\theta = 3\operatorname{sen}\theta - 4\operatorname{sen}^3\theta,$$

/4 & 3 ! @@@ A ! (!# !
3 $$ $ $)$ B ! !3 ! ==) *

obtenemos la forma simplificada de las ecuaciones paramétricas de la astroide,

$$x = a \cos^3 \theta , \quad y = a \operatorname{sen}^3 \theta . \tag{4}$$

Si tomamos la potencia dos tercios de ambos miembros de cada una de las ecuaciones (4) y sumamos, obtenemos como ecuación rectangular de la hipocicloide de cuatro picos

$$x^{2/3} + y^{2/3} = a^{2/3} . \tag{5}$$

EJERCICIOS. Grupo 43

Dibujar una figura para cada ejercicio.

1. De las ecuaciones paramétricas (2) del Artículo 92, demostrar que el tiempo en el cual alcanza el proyectil su altura máxima está dado por

$$t = \frac{v_0 \operatorname{sen} \alpha}{g}.$$

2. Si se conocen los ejes mayor y menor de una elipse, hallar un método para construir cualquier punto P de la elipse conociendo su ángulo excéntrico.

3. Dados el centro y el eje mayor de una elipse, hallar un procedimiento para construir el ángulo excéntrico de cualquier punto dado P de la elipse.

4. Sean P_1 y P_2 puntos extremos de dos diámetros conjugados de una elipse (véase el ejercicio 25 del grupo 29, Art. 63). Demostrar que los ángulos excéntricos de P_1 y P_2 difieren en 90° ó 270°.

5. Obtener las ecuaciones paramétricas (6) del Artículo 92 para una hipérbola, empleando una construcción en que $b > a$.

6. Sea l una recta dirigida hacia arriba, y sean α y β, respectivamente, los ángulos formados por l y las partes positivas de los ejes X y Y (ver el ejercicio 19 del grupo 14, Art. 37). Si l no es paralela a ninguno de los ejes coordenados y contiene al punto fijo $P_1(x_1, y_1)$, puede demostrarse que (ver el ejercicio 21 del grupo 14, Art. 37) la ecuación de l puede escribirse en la forma

$$\frac{x - x_1}{\cos \alpha} = \frac{y - y_1}{\cos \beta}.$$

De aquí, demostrar que una representación paramétrica de la recta l está dada por

$$x = x_1 + t \cos \alpha, \quad y = y_1 + t \cos \beta,$$

en donde el parámetro t representa la distancia variable del punto fijo $P_1(x_1, y_1)$ a cualquier punto $P(x, y)$ sobre l.

7. Discutir la recta cuyas ecuaciones paramétricas son

$$x = 3 + \frac{3}{5} t, \quad y = 4 + \frac{4}{5} t,$$

en donde el parámetro t tiene el significado establecido en el ejercicio 6.

8. Una recta cuya pendiente es $-\frac{5}{12}$ pasa por el punto $(2, -1)$. Hallar sus ecuaciones paramétricas en la forma dada en el ejercicio 6.

9. Demostrar la ecuación rectangular (2) de la cicloide dada en el Artículo 93.

10. Si los ejes coordenados son trasladados de tal manera que el nuevo origen sea el vértice H de la cicloide de la figura 129 del Artículo 93, demuéstrese que las ecuaciones paramétricas de la cicloide con respecto a los nuevos ejes están dadas por

$$x = a(\theta - \pi - \operatorname{sen} \theta), \quad y = -a(1 + \cos \theta).$$

11. Trazar la cicloide del ejercicio 10 cuando $a = 2$.

12. Deducir las ecuaciones paramétricas (3) de la trocoide dadas en el Artículo 93.

13. Obtener la ecuación rectangular de la trocoide a partir de las ecuaciones paramétricas (3) del Artículo 93.

14. Trazar la trocoide del ejercicio 12 cuando $a = 2$ y $b = 3$.

15. Trazar la epicicloide a partir de sus ecuaciones paramétricas (1) del Artículo 94 cuando $a = 3b$.

16. Deducir las ecuaciones paramétricas (2) de la cardioide, dadas en el Artículo 94, directamente a partir de una figura.

17. Deducir las ecuaciones paramétricas (3) de la hipocicloide, directamente de la figura 131.

18. Trazar la hipocicloide a partir de sus ecuaciones paramétricas (3) del Artículo 94 cuando $a = 3b$.

19. Demostrar, analíticamente, que cuando $a = 2b$ la hipocicloide (3) del Artículo 94 representa un diámetro de la circunferencia fija.

20. Si un hilo enrollado alrededor de una circunferencia fija se desenrolla manteniéndolo tirante en el plano de la circunferencia, cualquier punto fijo del hilo traza una curva llamada *evolvente de la circunferencia*. Hallar las ecuaciones paramétricas de la evolvente de la circunferencia $x^2 + y^2 = a^2$ bajo las siguientes condiciones: Si P es un punto cualquiera del lugar geométrico, sea el punto $A(a, 0)$ su posición inicial, y para cualquiera otra posición, sea T el punto de contacto de la tangente PT a la circunferencia. Tómese el ángulo $AOT = \theta$ como parámetro.

95. Resolución de problemas de lugares geométricos por el método paramétrico. Para ciertos lugares geométricos del tipo de curvas llamadas ruletas, hallamos que su representación paramétrica es preferible a su representación rectangular. Para muchas curvas, sin embargo, la ecuación rectangular es más deseable, pero esta ecuación puede determinarse a veces más convenientemente obteniendo primero las ecuaciones paramétricas a partir de las condiciones que el lugar geométrico debe satisfacer. Esto requiere la introducción de un parámetro, o posiblemente de dos o más parámetros, que deben eliminarse posteriormente. A este respecto, los parámetros son incidentales en la determinación de la ecuación rectangular y por esto se llaman a veces *variables auxiliares*. El lector debe notar que si se introducen n parámetros, es necesario tener $n+1$ ecuaciones para efectuar su eliminación y obtener la ecuación rectangular buscada. Si la ecuación rectangular de un lugar geométrico se obtiene mediante la introducción de uno o más parámetros, se suele decir que la resolución se ha efectuado por el *método paramétrico.*

Ejemplo 1. Hallar la ecuación del lugar geométrico del punto de intersección de dos rectas perpendiculares cualesquiera tangentes ambas a la elipse

$$b^2x^2 + a^2y^2 = a^2b^2.$$

Solución. Supongamos que el punto $P(x, y)$ (fig. 132) representa un punto cualquiera del lugar geométrico. Como las rectas son perpendiculares entre sí, podemos representar sus pendientes por m y $-\frac{1}{m}$, siendo la variable m el parámetro. Por el teorema 5 del Artículo 63 las ecuaciones de las tangentes son

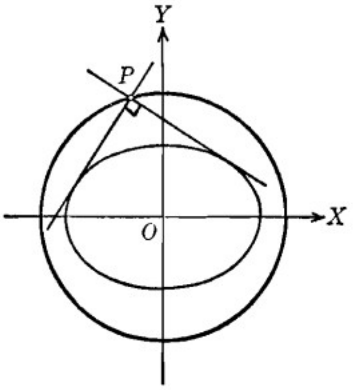

Fig. 132

$$y = mx \pm \sqrt{a^2m^2 + b^2}$$

y

$$y = -\frac{1}{m}x \pm \sqrt{\frac{a^2}{m^2} + b^2}.$$

Para obtener la ecuación rectangular requerida del lugar geométrico de P, debemos eliminar el parámetro m entre estas dos ecuaciones. Para esto, las escribiremos en las formas

$$y - mx = \pm\sqrt{a^2m^2 + b^2},$$

$$my + x = \pm\sqrt{a^2 + b^2m^2}.$$

Elevando al cuadrado ambos miembros de cada una de estas ecuaciones, y sumando, obtenemos

$$y^2 + m^2x^2 + m^2y^2 + x^2 = a^2m^2 + b^2 + a^2 + b^2m^2,$$

de donde

$$(m^2+1)(x^2+y^2) = (m^2+1)(a^2+b^2).$$

Si sustituímos estos valores de x', y' y x'' en la ecuación (4), obtenemos

$$\frac{2y}{\frac{15-4y}{3}} = \frac{6}{\frac{6x+4y-9}{3}}$$

la cual, después de simplificarla, nos da la ecuación buscada

$$6xy + 4y^2 + 3y - 45 = 0,$$

que representa una hipérbola. El estudiante debe trazar la gráfica correspondiente de este lugar geométrico.

Un tipo interesante de curvas, cuya ecuación se obtiene más fácilmente mediante el método paramétrico, son las llamadas *podarias* o *curvas pedales*, definidas de la siguiente manera: si desde un punto fijo Q se trazan perpendiculares a las tangentes a una curva C, el lugar geométrico de los pies de las perpendiculares es otra curva llamada *podaria de la curva* C *con respecto al punto* Q.

Ejemplo 3. Hallar la ecuación de la podaria de una parábola con respecto al vértice.

Solución. El problema no pierde generalidad si tomamos la forma canónica de la ecuación de la parábola, $y^2 = 4px$. Sea $P(x, y)$ (fig. 134) un punto cualquiera del lugar geométrico. Por el teorema 5 del Artículo 57, la ecuación de la tangente de pendiente m a la parábola $y^2 = 4px$ es

$$y = mx + \frac{p}{m}, \quad m \neq 0. \qquad (5)$$

Por ser OP perpendicular a la tangente (5), su ecuación es

$$y = -\frac{1}{m}x. \qquad (6)$$

La ecuación rectangular de la podaria se obtiene eliminando el parámetro m entre las ecuaciones (5) y (6). Para ello, de la ecuación (6) se obtiene $m = -\frac{x}{y}$, valor que sustituído en la ecuación (5) nos da:

$$y = -\frac{x^2}{y} - \frac{py}{x}.$$

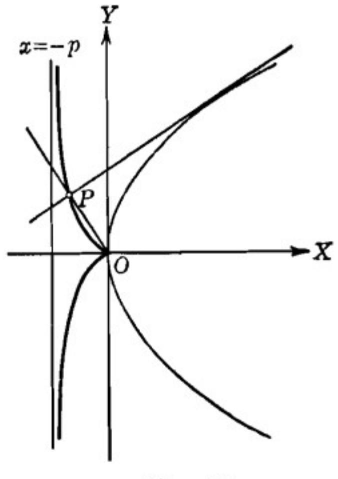

Fig. 134

Despejando y^2 obtenemos la ecuación rectangular buscada

$$y^2 = -\frac{x^3}{x+p},$$

que representa una cisoide con asíntota $x = -p$. (Véase el ejemplo 1 del Artículo 19 y el ejercicio 6 del grupo 41, Art. 88.)

+,-.+ + +/ ! + ')
3! :8 ! :2: !:! ! ! 3#
("') A 3 # 8 ! ,3 ! ! ! 0 !

6

EL PRODUCTO VECTORIAL Y APLICACIONES A LA TRIGONOMETRIA

A. EL CONCEPTO DE ANGULO

Hemos definido el ángulo entre dos vectores de un espacio vectorial euclidiano por su coseno, según la fórmula [13] del capítulo 2. Supuesto el ángulo comprendido entre 0° y 180°, su coseno lo determina sin ambigüedad. Sin embargo, conviene extender el dominio de los valores de los ángulos, por ejemplo, a valores negativos. Para ello es necesario definir una "orientación" en el espacio vectorial .

Consideremos un espacio vectorial de dimensión dos.

En la sección D del capítulo 3 vimos que dado un vector no nulo **U** existen siempre dos únicos versores ortogonales al mismo, los cuales son opuestos (o sea, difieren en el factor -1). Como ahora nos van a interesar únicamente las direcciones, o sea, las clases por la equivalencia "producto por un escalar", podemos suponer que **U** es un versor. Supongamos que se ha dado un criterio para asignar a todo versor **U** un único versor ortogonal, que indicaremos por $\mathbf{U}^\circ$, o sea, un criterio para elegir uno de los dos versores que existen ortogonales a **U**, cualquiera que sea este versor. Diremos que **U** y $\mathbf{U}^\circ$ son asociados.

Puesto que **U** y $\mathbf{U}^\circ$ forman una base, todo otro versor **V** será de la forma

$$\mathbf{V} = u\,\mathbf{U} + v\,\mathbf{U}^\circ, \text{ con } u^2 + v^2 = 1. \qquad [1]$$

Si

$$\mathbf{V}^\circ = u^\circ\,\mathbf{U} + v^\circ\,\mathbf{U}^\circ, \text{ con } u^{\circ 2} + v^{\circ 2} = 1 \qquad [2]$$

es el versor asociado a **V**, será

$$u\,u^\circ + v\,v^\circ = 0. \qquad [3]$$

De [1], [2], [3] se deduce

$$\begin{vmatrix} u & v \\ u^\circ & v^\circ \end{vmatrix}^2 = (uv^\circ - vu^\circ)^2 = (u^2+v^2)(u^{\circ 2}+v^{\circ 2}) - (uu^\circ + vv^\circ)^2 = 1$$

y por tanto

$$\begin{vmatrix} u & v \\ u^\circ & v^\circ \end{vmatrix} = \pm 1. \qquad [4]$$

Esto permite establecer la siguiente definición.

Definición. Dado un espacio vectorial de dimensión dos, se dice que el mismo está orientado si a cada versor **U** del espacio le está asignado un único versor asociado $\mathbf{U}^\circ$, ortogonal a **U**, y además si $(\mathbf{U}, \mathbf{U}^\circ)$ y $(\mathbf{V}, \mathbf{V}^\circ)$ son dos pares de versores asociados ligados por las relaciones [1] y [2], se cumple la igualdad

$$\begin{vmatrix} u & v \\ u^\circ & v^\circ \end{vmatrix} = 1. \qquad [5]$$

Según [13], capítulo 2, el ángulo α de los versores **U** y **V** (igual al ángulo de dos vectores cualesquiera proporcionales a **U** y **V**) está dado por la fórmula

$$\cos\alpha = \mathbf{U}\cdot\mathbf{V}. \qquad [6]$$

Si el espacio vectorial está orientado, definimos la función sen α por

$$\operatorname{sen}\alpha = \mathbf{U}^\circ\cdot\mathbf{V}. \qquad [7]$$

Multiplicando escalarmente ambos miembros de [1] primero por **U** y luego por $\mathbf{U}^\circ$, resulta $u = \cos\alpha$, $v = \operatorname{sen}\alpha$, de donde

$$\mathbf{V} = \cos\alpha\ \mathbf{U} + \operatorname{sen}\alpha\ \mathbf{U}^\circ. \qquad [8]$$

De aquí, elevando al cuadrado y teniendo en cuenta que por hipótesis **V** es un versor, resulta la relación fundamental

$$\cos^2\alpha + \operatorname{sen}^2\alpha = 1. \qquad [9]$$

Obsérvese que si hubieramos definido sen α a partir del cos α por esta relación [9] hubiera quedado indeterminado el signo, lo que no ocurre con la definición [7].

La condición [5] se escribe

$$v^\circ \cos\alpha - u^\circ \operatorname{sen}\alpha = 1$$

que junto con [2] obliga a que sea

$$u^\circ = -\operatorname{sen}\alpha, \quad v^\circ = \cos\alpha$$

y por tanto

$$\mathbf{V}^\circ = -\operatorname{sen}\alpha\ \mathbf{U} + \cos\alpha\ \mathbf{U}^\circ. \qquad [10]$$

Con todo esto podemos ya pasar a definir los ángulos negativos. Para ello se aceptan como *postulados* las *relaciones de Chasles*:

$$\left.\begin{aligned} \operatorname{ang}(\mathbf{U},\mathbf{V}) + \operatorname{ang}(\mathbf{V},\mathbf{U}) &= 0 \\ \operatorname{ang}(\mathbf{U},\mathbf{V}) + \operatorname{ang}(\mathbf{V},\mathbf{W}) + \operatorname{ang}(\mathbf{W},\mathbf{U}) &= 0 \end{aligned}\right\} \qquad [11]$$

donde **U**, **V** y **W** son tres versores cualesquiera del espacio vectorial de dimensión dos que estamos considerando.

De la primera relación [11] se deduce que si α es el ángulo entre **U** y **V**, el ángulo $-\alpha$ será el que forman **V** y **U**. Según [6] y [7] es

$$\cos(-\alpha) = \mathbf{V}\cdot\mathbf{U} = \mathbf{U}\cdot\mathbf{V} = \cos\alpha, \quad \operatorname{sen}(-\alpha) = \mathbf{V}^\circ\cdot\mathbf{U} = -\operatorname{sen}\alpha$$

donde la última igualdad resulta de [10] al multiplicar escalarmente por **U**.

Pongamos $\alpha = \operatorname{ang}(\mathbf{U},\mathbf{V})$ y $\beta = \operatorname{ang}(\mathbf{U},\mathbf{W})$, con lo cual el versor **W** será de la forma [8] aplicada a **W**, o sea

$$\mathbf{W} = \cos\beta\ \mathbf{U} + \operatorname{sen}\beta\ \mathbf{U}^\circ.$$

Según las relaciones [11] será $\operatorname{ang}(\mathbf{V},\mathbf{W}) = \beta - \alpha$ y por tanto aplicando las definiciones [6] y [7] y las fórmulas [8], [12] y [10] resulta

$$\cos(\beta - \alpha) = \mathbf{V}\cdot\mathbf{W} = \cos\alpha\ \cos\beta + \operatorname{sen}\alpha\ \operatorname{sen}\beta$$

$$\operatorname{sen}(\beta - \alpha) = \mathbf{V}^\circ\cdot\mathbf{W} = \operatorname{sen}\beta\ \cos\alpha - \cos\beta\ \operatorname{sen}\alpha$$

Para las fórmulas de adición basta substituir α por $-\alpha$.

B. TRIGONOMETRIA PLANA

Las fórmulas trigonométricas referentes a los triángulos del plano resultan muy fácilmente como relaciones vectoriales.

Sea el triángulo ABC y llamemos a, b y c a las longitudes de sus lados, o sea, a los módulos de los vectores B-C, C-A y A-B, respectivamente. Sean α, β y γ los ángulos correspondientes a los vértices A, B y C.

Elevando al cuadrado ambos miembros de la identidad

$$B - A = (B - C) - (A - C)$$

y puesto que $(B-C)\cdot(A-C) = a\, b \cos\gamma$, resulta

$$c^2 = a^2 + b^2 - 2\, a\, b \cos\gamma$$

que es la llamada fórmula del coseno.

Supongamos ahora que el espacio vectorial asociado al plano euclidiano está orientado. Sea $\mathbf{C}$ el versor asociado correspondiente al vector B - A, o sea, $\mathbf{C}\cdot(B - A) = 0$.

Según [7] y siendo $|C - A| = b$, se tiene

$$b \operatorname{sen}\alpha = \mathbf{C}\cdot(C - A) = \mathbf{C}\cdot(C - B + B - A) = \mathbf{C}\cdot(C - B) = a \operatorname{sen}\beta$$

de donde

$$\frac{a}{\operatorname{sen}\alpha} = \frac{b}{\operatorname{sen}\beta}$$

De igual manera se obtiene que estos cocientes son iguales a $c/\operatorname{sen}\gamma$. Por tanto

$$\frac{a}{\operatorname{sen}\alpha} = \frac{b}{\operatorname{sen}\beta} = \frac{c}{\operatorname{sen}\gamma}$$

que es la llamada fórmula del seno.

C. EL PRODUCTO VECTORIAL

En el espacio euclidiano de tres dimensiones, además del producto escalar de vectores, es útil considerar el llamado producto vectorial.

Sean

$$\mathbf{A} = a_1\,\mathbf{I} + a_2\,\mathbf{J} + a_3\,\mathbf{K}, \quad \mathbf{B} = b_1\,\mathbf{I} + b_2\,\mathbf{J} + b_3\,\mathbf{K}$$

dos vectores del espacio, expresados mediante la base **I, J, K.** Todo vector $\mathbf{C} = c_1\,\mathbf{I} + c_2\,\mathbf{J} + c_3\,\mathbf{K}$ que sea ortogonal a **A** y **B** debe cumplir las condiciones

$$\mathbf{C}\cdot\mathbf{A} = c_1\,a_1 + c_2\,a_2 + c_3\,a_3 = 0,$$

$$\mathbf{C}\cdot\mathbf{B} = c_1\,b_1 + c_2\,b_2 + c_3\,b_3 = 0. \qquad [12]$$

De estas ecuaciones se deduce

$$c_1 = \frac{a_2 b_3 - a_3 b_2}{a_1 b_2 - a_2 b_1}\,c_3, \quad c_2 = \frac{a_3 b_1 - a_1 b_3}{a_1 b_2 - a_2 b_1}\,c_3$$

o sea

$$\frac{c_1}{a_2\,b_3 - a_3\,b_2} = \frac{c_2}{a_3\,b_1 - a_1\,b_3} = \frac{c_3}{a_1\,b_2 - a_2\,b_1}$$

Consideremos, además, la identidad

$$(a_2\,b_3 - a_3\,b_2)^2 + (a_3\,b_1 - a_1\,b_3)^2 + (a_1\,b_2 - a_2\,b_1)^2$$

$$= (a_1^2 + a_2^2 + a_3^2)(b_1^2 + b_2^2 + b_3^2) - (a_1\,b_1 + a_2\,b_2 + a_3\,b_3)^2. \qquad [13]$$

El vector de componentes

$$a_2\,b_3 - a_3\,b_2, \quad a_3\,b_1 - a_1\,b_3, \quad a_1\,b_2 - a_2\,b_1 \qquad [14]$$

según [12] es ortogonal a **A** y **B**, o sea, es ortogonal al plano que ellos determinan (supuestos llevados a partir de un mismo origen) y según [13] tiene por módulo $(a^2b^2 - a^2b^2\cos^2\alpha)^{1/2} = ab\,|\mathrm{sen}\,\alpha|$, siendo a y b los módulos de **A**, **B** y α el ángulo que estos vectores forman entre sí. Esto nos dice que el vector [14] a pesar de haber sido definido mediante las componentes **A** y **B** respecto a cierta base **I, J, K** tiene la dirección y el módulo independientes de esta base, o sea, está vinculado intrínsecamente a los vectores dados **A** y **B**. No ocurre lo mismo con el sentido. Cambiando la base **I, J, K** y tomando las mismas expresiones [14] con las componentes de **A** y **B** en la nueva base, puede ocurrir que el vector obtenido tenga el sentido opuesto al anterior, o sea, que difiera en un factor -1. Por esto se dice que las expresiones [14] son componentes de un pseudo vector, el cual se llama el producto vectorial de **A** y **B** y se representa $\mathbf{A}\wedge\mathbf{B}$. La regla práctica para acordarse de las componentes [14] es considerarlas como resultado del desarrollo del determinante

$$\mathbf{A} \wedge \mathbf{B} = \begin{vmatrix} \mathbf{I} & \mathbf{J} & \mathbf{K} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} . \qquad [15]$$

Mientras mantengamos una base **I**, **J**, **K** fija, el producto vectorial actúa como un vector bien definido por [15] y por consiguiente se opera con él lo mismo que con los vectores. Es únicamente al cambiar de base que puede ocurrir que cambie el sentido, o sea el signo del producto vectorial, pero como su dirección y su módulo no cambian resulta un producto muy útil desde el punto de vista de las aplicaciones geométricas.

D. PRODUCTO MIXTO DE TRES VECTORES Y DOBLE PRODUCTO VECTORIAL

Se llama producto mixto de tres vectores **A**, **B** y **C** al escalar

$$(\mathbf{A}\ \mathbf{B}\ \mathbf{C}) = (\mathbf{A} \wedge \mathbf{B})\cdot\mathbf{C} = \begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix} .$$

De aquí se deduce, recordando la propiedad de los determinantes de cambiar de signo al permutar dos filas entre sí,

$$(\mathbf{A}\ \mathbf{B}\ \mathbf{C}) = (\mathbf{B}\ \mathbf{C}\ \mathbf{A}) = (\mathbf{C}\ \mathbf{A}\ \mathbf{B}) = -(\mathbf{B}\ \mathbf{A}\ \mathbf{C}) = -(\mathbf{A}\ \mathbf{C}\ \mathbf{B}) = -(\mathbf{C}\ \mathbf{B}\ \mathbf{A}).$$

Pasando a las componentes respecto de una determinada base, se pueden comprobar las siguientes identidades

$$(\mathbf{A} \wedge \mathbf{B}) \wedge \mathbf{C} = (\mathbf{A}\cdot\mathbf{C})\ \mathbf{B} - (\mathbf{B}\cdot\mathbf{C})\ \mathbf{A},$$

$$(\mathbf{A} \wedge \mathbf{B})\cdot(\mathbf{C} \wedge \mathbf{D}) = (\mathbf{A}\cdot\mathbf{C})(\mathbf{B}\cdot\mathbf{D}) - (\mathbf{B}\cdot\mathbf{C})(\mathbf{A}\cdot\mathbf{D}) \text{ (identidad de Lagrange)}$$

$$(\mathbf{A} \wedge \mathbf{B}) \wedge (\mathbf{C} \wedge \mathbf{D}) = (\mathbf{A}\ \mathbf{B}\ \mathbf{D})\ \mathbf{C} - (\mathbf{A}\ \mathbf{B}\ \mathbf{C})\ \mathbf{D}.$$

Mediante artificios adecuados estas identidades pueden demostrarse de manera más simple, como puede verse en cualquier texto clásico de cálculo vectorial.

E. TRIGONOMETRIA ESFERICA

Vamos a ver como las relaciones anteriores permiten obtener las fórmulas clásicas de la trigonometría esférica.

Fórmula del coseno. Consideremos los versores **A**, **B** y **C**. Supuestos llevados a partir de un mismo origen, sus extremos serán los vértices de un triángulo esférico A B C; llamemos a, b y c a los lados de este triángulo y α, β y γ a sus ángulos, de manera que será

$$\mathbf{A}\cdot\mathbf{B} = \cos c, \quad \mathbf{A}\cdot\mathbf{C} = \cos b, \quad \mathbf{C}\cdot\mathbf{B} = \cos a. \qquad [16]$$

En la identidad de Lagrange

$$(\mathbf{A} \wedge \mathbf{C})\cdot(\mathbf{B} \wedge \mathbf{C}) = (\mathbf{A}\cdot\mathbf{B})\,(\mathbf{C}\cdot\mathbf{C}) - (\mathbf{A}\cdot\mathbf{C})(\mathbf{C}\cdot\mathbf{B})$$

el primer miembro es el producto de $|\mathbf{A} \wedge \mathbf{C}| = \operatorname{sen} b$ por $|\mathbf{B} \wedge \mathbf{C}| = \operatorname{sen} a$, por el coseno del ángulo que forman dos vectores ortogonales a los planos **A, C** y **B, C**, es decir, por cos γ. Los factores del segundo miembro están dados por [16]. Sustituyendo, resulta

$$\cos c = \cos a \;\; \cos b + \operatorname{sen} a \;\; \operatorname{sen} b \;\; \cos \gamma$$

que es la fórmula del coseno de la trigonometría esférica.

Fórmula del seno. Apliquemos ahora la fórmula

$$(\mathbf{A} \wedge \mathbf{C}) \wedge (\mathbf{B} \wedge \mathbf{C}) = (\mathbf{A}\ \mathbf{C}\ \mathbf{C})\ \mathbf{B} - (\mathbf{A}\ \mathbf{C}\ \mathbf{B})\ \mathbf{C}$$

que por ser $(\mathbf{A}\ \mathbf{C}\ \mathbf{C}) = 0$ se reduce a

$$(\mathbf{A} \wedge \mathbf{C}) \wedge (\mathbf{B} \wedge \mathbf{C}) = (\mathbf{A}\ \mathbf{B}\ \mathbf{C})\ \mathbf{C}.$$

Igualando los módulos de los dos miembros de esta igualdad resulta

$$\operatorname{sen} a \operatorname{sen} b \operatorname{sen} \gamma = |(\mathbf{A}\ \mathbf{B}\ \mathbf{C})|.$$

Por permutación circular de los versores **A**, **B** y **C** puesto que

$$|(\mathbf{ABC})| = |(\mathbf{BCA})| = |(\mathbf{CAB})|, \text{ resulta}$$

$$\operatorname{sen} a \;\; \operatorname{sen} b \;\; \operatorname{sen} \gamma = \operatorname{sen} b \;\; \operatorname{sen} c \;\; \operatorname{sen} \alpha = \operatorname{sen} c \;\; \operatorname{sen} a \;\; \operatorname{sen} \beta$$

que puede escribirse

$$\frac{\text{sen } a}{\text{sen } \alpha} = \frac{\text{sen } b}{\text{sen } \beta} = \frac{\text{sen } c}{\text{sen } \gamma}$$

que es la fórmula del seno de la trigonometría esférica.

* * * * *

+,-.+ + +/ ! + ') ! : 8 !: ! : ! 3#
+,-.+ + +/ ! + ') ! ! 4 ! ! # !
8 ! 3#
(' ' !3C &&& & ! !4 ! !

Capítulo VIII

Integrales abelianas

A finales del siglo XVII, Jakob Bernoulli introdujo el término "lemniscata" para referirse a las curvas con forma de 8 (el lemnisco era la cinta que adornaba las coronas con que se distinguía a los atletas en la antigua Grecia). Entre ellas se encuentra la que ahora se conoce como lemniscata de Bernoulli.

Del mismo modo que una elipse puede definirse como el conjunto de los puntos del plano tales que la suma de sus distancias a dos focos es constante, podemos considerar el conjunto de los puntos del plano tales que el producto de sus distancias a dos focos sea constante. Eligiendo adecuadamente el sistema de referencia, no perdemos generalidad si suponemos que los focos tienen coordenadas $(\pm c, 0)$. Si $(x, y) \in \mathbb{R}^2$ es un punto de módulo $\rho = \sqrt{x^2 + y^2}$, su distancia a cada foco es

$$\begin{aligned} r_1^2 &= (x+c)^2 + y^2 = \rho^2 + 2cx + c^2, \\ r_2^2 &= (x-c)^2 + y^2 = \rho^2 - 2cx + c^2, \end{aligned}$$

luego el punto está en en el conjunto si cumple

$$r_1^2 r_2^2 = (\rho^2 + c^2)^2 - 4c^2x^2 = \rho^4 + 2c^2\rho^2 - 4c^2x^2 + c^4 = d^4,$$

para una constante $d > 0$. Si $d < c$, estos puntos forman una curva con dos componentes conexas, cada una de las cuales rodea a uno de los focos. Si $d > c$ obtenemos una curva conexa que rodea a ambos focos. En cambio, si $d = c$ obtenemos una figura en forma de 8, es decir, una lemniscata, que es precisamente la *lemniscata de Bernoulli.* En este caso la ecuación se reduce a

$$\rho^4 + 2c^2\rho^2 - 4c^2x^2 = 0,$$

o, equivalentemente,

$$(x^2 + y^2)^2 = 2c^2(y^2 - x^2).$$

El valor de c es irrelevante, en el sentido de que, aplicando una homotecia, podemos hacer que $c = 1/\sqrt{2}$, con lo que $2c^2 = 1$, con lo que la ecuación se reduce a

$$\rho^4 + \rho^2 - 2x^2 = 0, \qquad \text{o} \qquad (x^2 + y^2)^2 = y^2 - x^2.$$

Así pues, salvo un cambio de escala, toda lemniscata de Bernoulli tiene el aspecto que muestra la figura siguiente:

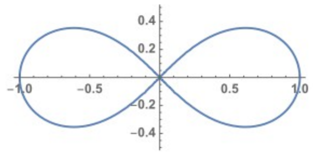

Es fácil ver que una parametrización de la lemniscata es:

$$x = \pm\frac{1}{\sqrt{2}}\sqrt{\rho^2+\rho^4}, \qquad y = \pm\frac{1}{\sqrt{2}}\sqrt{\rho^2-\rho^4}.$$

Un simple cálculo nos da que el elemento de longitud es

$$ds = \frac{d\rho}{\sqrt{1-\rho^4}}.$$

Así pues, la longitud del arco de lemniscata comprendido entre el origen y un punto situado a distancia $0 < r < 1$ (unidos por el camino más corto) es

$$s(r) = \int_0^r \frac{d\rho}{\sqrt{1-\rho^4}}.$$

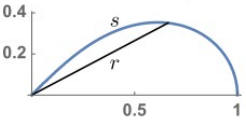

En 1718, el conde Fagnano descubrió una curiosa propiedad sobre el arco de una lemniscata. Probablemente, trató de explotar la analogía entre esta integral y la bien conocida

$$\operatorname{arcsen} r = \int_0^r \frac{d\rho}{\sqrt{1-\rho^2}}.$$

Esta integral se racionaliza con el cambio de variable

$$\rho = \frac{2\tau}{1+\tau^2},$$

por lo que parece razonable aplicar a nuestra integral el cambio

$$\rho^2 = \frac{2\tau^2}{1+\tau^4}.$$

Un simple cálculo nos da que

$$s(r) = \int_0^r \frac{d\rho}{\sqrt{1-\rho^4}} = \sqrt{2}\int_0^t \frac{d\tau}{\sqrt{1+\tau^4}}, \qquad r^2 = \frac{2t^2}{1+t^4}.$$

En lo que se refiere al cálculo de la integral no hemos ganado nada, pero Fagnano se dio cuenta de

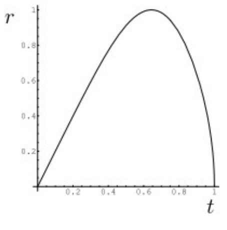

que si ahora aplicamos el mismo cambio pero con un signo negativo en el denominador, obtenemos

$$s(r) = \int_0^r \frac{d\rho}{\sqrt{1-\rho^4}} = 2\int_0^t \frac{d\tau}{\sqrt{1-\tau^4}} = 2s(t),$$

donde

$$r^2 = \frac{2\frac{2t^2}{1-t^4}}{1+\left(\frac{2t^2}{1-t^4}\right)^2} = \frac{4t^2(1-t^4)}{(1+t^4)^2}.$$

Esto tiene una interpretación geométrica: si señalamos un punto de la lemniscata a una distancia $0 < t < 1$ del origen, entonces el punto situado a la distancia r dada por la fórmula anterior determina un arco de doble longitud. (En realidad es necesario suponer que t no exceda del punto donde $r(t)$ deja de ser inyectiva, lo cual equivale a que al duplicar el arco no sobrepasemos el cuadrante). La sencilla relación entre r y t permite, por ejemplo, duplicar un arco de lemniscata con regla y compás. También podemos expresar t en función de r, lo que nos da un método para bisecar un arco de lemniscata.

Así, en la figura, el arco s_1, correspondiente a $t = 0.5$, tiene la misma longitud que s_2, cuyo extremo derecho corresponde a $r = 4\sqrt{15}/17$, dado por la relación precedente.

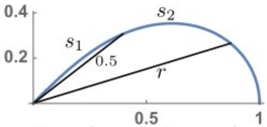

En 1751, Euler llegó más lejos que Fagnano. En primer lugar, vio que la analogía con el arco seno podía aprovecharse más aún. La relación

$$\operatorname{sen}(x+y) = \operatorname{sen} x \cos y + \cos x \operatorname{sen} y$$

puede expresarse en la forma

$$u\sqrt{1-v^2} + v\sqrt{1-u^2} = \operatorname{sen}(x+y), \qquad u = \operatorname{sen} x, \quad v = \operatorname{sen} y,$$

o también,

$$\int_0^u \frac{d\rho}{\sqrt{1-\rho^2}} + \int_0^v \frac{d\rho}{\sqrt{1-\rho^2}} = \int_0^r \frac{d\rho}{\sqrt{1-\rho^2}}, \qquad r = u\sqrt{1-v^2} + v\sqrt{1-u^2}.$$

Aquí hay que entender que u y v son números positivos suficientemente pequeños. Euler consiguió probar una fórmula análoga para la lemniscata, a saber:

$$\int_0^u \frac{d\rho}{\sqrt{1-\rho^4}} + \int_0^v \frac{d\rho}{\sqrt{1-\rho^4}} = \int_0^r \frac{d\rho}{\sqrt{1-\rho^4}}, \quad r = \frac{u\sqrt{1-v^4} + v\sqrt{1-u^4}}{1+u^2v^2}.$$

Esta fórmula se particulariza a la de Fagnano cuando $u = v$, y permite sumar fácilmente dos arcos de lemniscata suficientemente pequeños como para que la suma no exceda un cuadrante.

Más aún, Euler generalizó su resultado probando que sigue siendo cierto si cambiamos el polinomio $P(u) = 1 - u^4$ por $P(u) = 1 + au^2 - u^4$, para cualquier valor de a.

Los resultados de Euler fueron notablemente generalizados por Abel, quien estudió el comportamiento de las integrales de la forma

$$\int R(x, y)\, dx,$$

donde $R(X, Y)$ es una función racional e $y(x)$ es una función algebraica determinada por una relación polinómica $P(x, y) = 0$. Estas integrales se conocen como integrales abelianas. En términos más modernos, podemos definir las integrales abelianas como las integrales curvilíneas de formas diferenciales meromorfas sobre superficies de Riemann. Por ejemplo, para interpretar como integral abeliana la integral que da la longitud de arco de la lemniscata consideramos la curva proyectiva S dada por $Y^2 = 1 - X^4$ y, en ella, la forma $\omega = dx/y$. Así,

$$s(r) = \int_\sigma \omega, \qquad \text{donde } \sigma(\rho) = (\rho, \sqrt{1 - \rho^4}).$$

Dedicamos este capítulo a estudiar las integrales abelianas. Ello nos llevará a unos resultados generales de los cuales no sólo se deducen los hechos que acabamos de comentar, sino también resultados de gran interés teórico. Por ejemplo, vamos a determinar la estructura del grupo de clases de grado 0 de una curva proyectiva.

8.1 Homología y cohomología

Vamos a recordar y generalizar al caso complejo algunos resultados sobre topología algebraica y geometría diferencial que vamos a necesitar. Antes de entrar en materia, recordemos el comportamiento de las integrales curvilíneas sobre un abierto $U \subset \mathbb{C}$ simplemente conexo. Si $f : U \longrightarrow \mathbb{C}$ es una función holomorfa, el teorema de Cauchy [VC 1.34] afirma que la integral de f a lo largo de un arco cerrado es nula. Ello se debe esencialmente a que los arcos cerrados son homotópicos a un punto y las integrales no se alteran al transformar homotópicamente los arcos. Como consecuencia, fijado $P \in U$, la integral

$$\int_P^z f(\zeta)\, d\zeta \tag{8.1}$$

no depende del arco sobre el que se calcula y determina una función holomorfa en U, una primitiva de f, de hecho (véase la prueba de [VC 3.8]).

Si f es una función meromorfa, entonces ya no podemos aplicar el teorema de Cauchy, sino que en su lugar tenemos el teorema de los residuos [VC 3.14], que dice que la integral a lo largo de un arco cerrado será una combinación lineal entera de los residuos de f multiplicados por $2\pi i$. Pese a ello podemos hablar aún de la integral (8.1), si bien ahora se trata de una integral multiforme meromorfa, en el sentido de que en un entorno (simplemente conexo) de cada punto z_0 que no sea un polo de f podemos determinar ramas uniformes meromorfas de la integral (la rama dependerá del camino elegido para unir P con z_0). Un caso típico es la integral

$$\int_1^z \frac{d\zeta}{\zeta},$$

que determina la función logaritmo en $\mathbb{C} \setminus \{0\}$.

Si pasamos a considerar integrales de formas diferenciales en una superficie de Riemann, en primer lugar nos encontramos con que, salvo en el caso trivial de la esfera, ya no estamos en un espacio simplemente conexo, por lo que el teorema de Cauchy no es válido ni siquiera para integrales de formas holomorfas. Por ejemplo, si integramos una forma holomorfa en un toro a lo largo de un arco que una dos puntos, el resultado dependerá del número de vueltas que demos al toro. De hecho hay dos clases de vueltas distintas: las que damos transversalmente, alrededor del "tubo" y las que damos longitudinalmente, a lo largo del "tubo". Los valores que toma la integral sobre dos arcos cerrados que den una única vuelta transversal y longitudinal respectivamente se llaman periodos de la integral, y de nuevo sucede que la integral entre dos puntos está bien definida módulo estos dos periodos.

Si el integrando es una forma meromorfa, además de los periodos en el sentido anterior tenemos también periodos polares, como en el caso plano (salvo que la forma tenga residuos nulos, es decir, salvo que sea de segunda clase).

Homología y cohomología compleja Vamos a precisar las ideas precedentes, para lo cual consideramos en principio una variedad diferencial S de dimensión $2n$, en las mismas condiciones que en la sección A.5 de [VC]. De acuerdo con lo expuesto allí, para cada $P \in S$, podemos considerar el espacio tangente real $T_P(S)$, que es un $\mathbb{R}$-espacio vectorial de dimensión $2n$, y también el espacio $T_P(S, \mathbb{C}) \cong \mathbb{C} \otimes_{\mathbb{R}} T_P(S)$, que es un $\mathbb{C}$-espacio vectorial de dimensión $2n$.

Consideramos ahora [GD 3.24] los espacios $\Lambda^p(S) = \Lambda^p(S, \mathbb{R})$ de las p-formas diferenciales en S. Definimos

$$\Lambda^p(S, \mathbb{C}) = \mathbb{C} \otimes_{\mathbb{R}} \Lambda^p(S, \mathbb{R}) = \{\omega_1 + i\omega_2 \mid \omega_1, \omega_2 \in \Lambda^p(S, \mathbb{R})\}.$$

En particular, $\Lambda^0(S, \mathbb{C})$ es el espacio de todas las funciones diferenciables $f : S \longrightarrow \mathbb{C}$. Similarmente, cada $\omega \in \Lambda^1(S, \mathbb{C})$ puede verse como una función que a cada $P \in S$ le asigna una 1-forma $\omega_P \in \mathbb{C} \otimes_{\mathbb{R}} T_P(S)^* \cong T_P(S, \mathbb{C})^*$.

Nota Nos va a bastar con este hecho, pero, en general, es claro que, si V es un $\mathbb{R}$-espacio vectorial de dimensión finita, todo tensor covariante $\alpha \in \mathcal{T}_s^0(\mathbb{C} \otimes_R V)$ [GD A.1], es decir, toda aplicación $\mathbb{C}$-multilineal

$$\alpha : (\mathbb{C} \otimes_{\mathbb{R}} V) \times \cdots \times (\mathbb{C} \otimes_{\mathbb{R}} V) \longrightarrow \mathbb{C}$$

está determinada por su acción sobre los elementos de B^s, donde B es una base de V, por lo que cada $\alpha \in \mathcal{T}_s^0(V)$ tiene una extensión única $\bar{\alpha} \in \mathcal{T}_s^0(\mathbb{C} \otimes_{\mathbb{R}} V)$, lo que nos da un isomorfismo natural $\mathbb{C} \otimes_{\mathbb{R}} \mathcal{T}_s^0(V) \cong \mathcal{T}_s^0(\mathbb{C} \otimes_{\mathbb{R}} V)$ dado por $z \otimes \alpha \mapsto z\bar{\alpha}$.

Este isomorfismo se restringe a un isomorfismo $\mathbb{C}\otimes_{\mathbb{R}}\Lambda^s(V)\cong\Lambda^s(\mathbb{C}\otimes_{\mathbb{R}}V)$. Si aplicamos esto al caso en que $V=T_P(S)$, vemos que cada $\omega\in\Lambda^p(S,\mathbb{C})$ puede identificarse con una aplicación que a cada $P\in S$ le asigna una forma constante $\omega_P\in\Lambda^p(T_P(S,\mathbb{C}))$. ■

Podemos definir de forma natural el producto de una función $f\in\Lambda^0(S,\mathbb{C})$ por una forma $\omega\in\Lambda^p(S,\mathbb{C})$.

Definimos $d:\Lambda^p(S,\mathbb{C})\longrightarrow\Lambda^{p+1}(S,\mathbb{C})$ mediante $d(\omega_1+i\omega_2)=d\omega_1+id\omega_2$.

Notemos que si $f\in\Lambda^0(S,\mathbb{C})$, la diferencial que acabamos de definir es la misma que ya teníamos definida en la sección A.5 de [VC]. Allí se muestra que si S es el dominio de una carta de coordenadas $z_1,\ldots,z_n$, entonces, para cada punto $P\in S$, las diferenciales $dz_i|_P$, $d\bar{z}_i|_P$ forman una base de $T_P(S,\mathbb{C})^*$, concretamente, la base dual de los vectores tangentes $\partial_{z_i}|_P$, $\partial_{\bar{z}_i}|_P$. En particular, si $f\in\Lambda^0(S,\mathbb{C})$, se cumple que

$$df=\sum_{i=1}^n\frac{\partial f}{\partial z_i}\,dz_i+\sum_{i=1}^n\frac{\partial f}{\partial\bar{z}_i}\,d\bar{z}_i.$$

El operador d cumple trivialmente $d\circ d=0$, luego define grupos de cohomología $H^p(S,\mathbb{C})$ análogamente al caso real [GD 6.11], es decir, como el cociente del espacio de las p-formas cerradas (con diferencial nula) sobre el subespacio de las p-formas exactas. Es claro que la asignación $[\omega_1]+i[\omega_2]\mapsto[\omega_1+i\omega_2]$ determina un isomorfismo natural

$$\mathbb{C}\otimes_{\mathbb{R}}H^p(S,\mathbb{R})\cong H^p(S,\mathbb{C}).$$

Por otra parte, si fijamos un anillo conmutativo y unitario $\mathbb{A}$ (que en la práctica será siempre $\mathbb{Z}$, $\mathbb{R}$ o $\mathbb{C}$), podemos considerar [TA 9.6] los $\mathbb{A}$-módulos $C_p(S,\mathbb{A})$ de cadenas de símplices singulares diferenciables (para la diferenciabilidad véase el final de la sección 9.2 de [TA]), con los que podemos definir los grupos de homología singular $H_p(S,\mathbb{A})$. Según [TA 12.21], existe un isomorfismo natural

$$H_p(S,\mathbb{C})\cong\mathbb{C}\otimes_{\mathbb{R}}H_p(S,\mathbb{R}).$$

Definimos la integral de una forma $\omega=\omega_1+i\omega_2\in\Lambda^p(S,\mathbb{C})$ sobre un símplice $\sigma\in C_p(S,\mathbb{C})$ mediante

$$\int_\sigma\omega=\int_\sigma\omega_1+i\int_\sigma\omega_2.$$

(La integral de una forma diferencial real sobre un símplice está definida antes de [TA 13.12].) Por linealidad, la integral se extiende a una forma bilineal

$$\int:C_p(S,\mathbb{C})\times\Lambda^p(S,\mathbb{C})\longrightarrow\mathbb{C}.$$

La versión compleja del teorema de Stokes [TA 13.13] se sigue inmediatamente de la versión real, es decir, si $c\in C_{p+1}(S,\mathbb{C})$ y $\omega\in\Lambda^p(S,\mathbb{C})$, entonces

$$\int_c d\omega=\int_{\partial c}\omega.$$

Esto nos permite definir una forma bilineal

$$\int^* : H_p(S, \mathbb{C}) \times H^p(S, \mathbb{C}) \longrightarrow \mathbb{C}.$$

Si $\phi : S \longrightarrow T$ es una aplicación diferenciable, tenemos definidas las retracciones $\phi^* : \Lambda^p(T, \mathbb{R}) \longrightarrow \Lambda^p(S, \mathbb{R})$ (véase el final de la sección 6.2 de GD]), que nos permiten definir aplicaciones lineales $\phi^* : \Lambda^p(T, \mathbb{C}) \longrightarrow \Lambda^p(S, \mathbb{C})$ mediante

$$\phi^*(\omega_1 + i\omega_2) = \phi^*(\omega_1) + i\phi^*(\omega_2).$$

El teorema [GD 3.48] se traduce inmediatamente a su análogo complejo, es decir, a que $\phi^* \circ d = d \circ \phi^*$, por lo que ϕ^* induce una aplicación lineal $\phi^* : H^p(T, \mathbb{C}) \longrightarrow H^p(S, \mathbb{C})$.

Por otra parte, ϕ induce aplicaciones lineales $\phi^\sharp : C_1(S, \mathbb{C}) \longrightarrow C_1(T, \mathbb{C})$ (véase [TA 9.6]), que a su vez inducen aplicaciones $\phi_* : H_1(S, \mathbb{C}) \longrightarrow H_1(T, \mathbb{C})$ (véase la sección 9.3 de [TA]).

El teorema [TA 13.12] se generaliza trivialmente al caso complejo, de modo que, para todo $c \in C_p(S, \mathbb{C})$ y toda $\omega \in \Lambda^p(S, \mathbb{C})$, se cumple

$$\int_{\phi^\sharp(c)} \omega = \int_c \phi^*(\omega)$$

y, a su vez,

$$\int^*_{\phi_*([c])} [\omega] = \int^*_{[c]} \phi^*([\omega]).$$

También es fácil probar las relaciones

$$\phi^*(f\omega) = (\phi \circ f)\phi^*(\omega), \qquad \phi^*(\omega)_P(v) = \omega_{\phi(P)}(d_P\phi(v)), \quad \omega \in \Lambda^1(S, \mathbb{C}).$$

Formas holomorfas A partir de este momento suponemos que S es una variedad analítica compacta de dimensión 1 (con lo que en particular sigue siendo una variedad diferencial de dimensión 2, como hasta ahora con $n = 1$). En tal caso, según [VC A.38], podemos descomponer $T_P(S, \mathbb{C})$ como suma directa de un espacio tangente holomorfo y otro antiholomorfo, ambos de dimensión compleja 1:

$$T_P(S, \mathbb{C}) = T_P^h(S) \oplus T_P^a(S).$$

Además, según [VC A.39], podemos identificar $T_p(S)$ con $T_P^h(S)$ de forma natural.

Si $\omega \in \Lambda^1(S, \mathbb{C})$ y $U \subset S$ es el dominio de una carta z, tenemos que $\omega|_U = f\,dz + g\,d\bar{z}$, para ciertas funciones f, $g \in \Lambda^0(U)$ unívocamente determinadas.

Llamaremos *formas diferenciales holomorfas* en S a las que cumplan que $g = 0$ y $f \in \mathcal{H}(U)$ para todo abierto coordenado U, es decir, a las que cumplen $\omega|_U = f\,dz$, para cierta función f holomorfa en U. Es fácil ver que, para que esto suceda, basta con que se cumpla para un atlas analítico de S. Llamaremos $\Lambda^1(S)$ al espacio de todas las formas diferenciales holomorfas en S.

Así, si $\omega \in \Lambda^1(S)$, para cada $P \in S$ se cumple que ω_P se anula sobre los vectores tangentes antiholomorfos (pues una base de $T_P^a(S)$ es $\partial_{\bar{z}}$, donde z es una carta alrededor de P).

Se comprueba fácilmente que si $\phi : S \longrightarrow T$ es una aplicación holomorfa entre variedades analíticas, entonces ϕ^* se restringe a una aplicación lineal

$$\phi^* : \Lambda^1(T) \longrightarrow \Lambda^1(S),$$

con lo que Λ^1 es un funtor contravariante.

Observemos ahora que si $\sigma : [0,1] \longrightarrow U \subset \mathbb{C}$ es un 1-símplice en un abierto de $\mathbb{C}$, entonces la integral $\int_\sigma f\,dz$ coincide con la integral curvilínea usual en variable compleja.

En efecto, teniendo en cuenta que $\sigma = \sigma^\sharp(I)$, donde I es la identidad en $[0,1]$ y el análogo complejo a [TA 13.12], tenemos que

$$\int_\sigma f\,dz = \int_I \sigma^*(f\,dz) = \int_I (\sigma \circ f)\,d(\sigma \circ z) = \int_I (\sigma \circ f)\,d\sigma,$$

donde hemos usado que z no es sino la identidad en U. Si t es la identidad en $[0,1]$ (vista ahora como carta de $[0,1]$), entonces $d\sigma = d\sigma(\partial_t)\,dt = \sigma'(t)\,dt$, luego, por la definición de integral sobre un 1-símplice:

$$\int_\sigma f\,dz = \int_0^1 f(\sigma(t))\,\sigma'(t)\,dt.$$

Todas las formas holomorfas son cerradas. En efecto, si $\omega \in \Lambda^1(S)$ y $P \in S$, tomamos una carta z en un entorno U de P, de modo que

$$\omega|_U = f\,dz = \operatorname{Re} f\,dx - \operatorname{Im} f\,dy + i(\operatorname{Im} f\,dx + \operatorname{Re} f\,dy),$$

para una cierta función homomorfa f. Teniendo en cuenta que la diferencial es un operador local, vemos que

$$d\omega|_U = \left(-\frac{\partial \operatorname{Re} f}{\partial y} - \frac{\partial \operatorname{Im} f}{\partial x}\right) dx \wedge dy + i\left(-\frac{\partial \operatorname{Im} f}{\partial y} + \frac{\partial \operatorname{Re} f}{\partial x}\right) dx \wedge dy.$$

Como f satisface las ecuaciones de Cauchy-Riemann, concluimos que $d\omega = 0$.

Por otra parte, si una forma holomorfa ω cumple $\omega = dg$, para cierta función $g \in \Lambda^0(S, \mathbb{C})$, de hecho ha de ser $g \in \mathcal{H}(S)$.

En efecto, en un entorno coordenado U de un punto arbitrario tenemos que

$$\omega|_U = dg|_U = \frac{\partial \operatorname{Re} g}{\partial x}\,dx + \frac{\partial \operatorname{Re} g}{\partial y}\,dy + i\frac{\partial \operatorname{Im} g}{\partial x}\,dx + i\frac{\partial \operatorname{Im} g}{\partial y}\,dy$$

y también

$$\omega|_U = f\,dz = \operatorname{Re} f\,dx - \operatorname{Im} f\,dy + i(\operatorname{Im} f\,dx + \operatorname{Re} f\,dy),$$

para una cierta función holomorfa f. Comparando ambas expresiones concluimos que g cumple las ecuaciones de Cauchy-Riemann, luego es una función holomorfa.

Esto implica que si definimos $\mathcal{H}^1(S)$ como el cociente del espacio $\Lambda^1(S)$ de las formas holomorfas en S sobre el subespacio de las diferenciales de funciones holomorfas, la aplicación $[\omega] \mapsto [\omega]$ es un monomorfismo de $\mathcal{H}^1(S)$ en $H^1(S, \mathbb{C})$. En general no es un isomorfismo.

También es claro que si $\phi : S \longrightarrow T$ es una aplicación holomorfa entre variedades analíticas, entonces $\phi^* : \Lambda^1(T) \longrightarrow \Lambda^1(S)$ induce una aplicación lineal $\phi^* : \mathcal{H}^1(T) \longrightarrow \mathcal{H}^1(S)$, con lo que $\mathcal{H}^1$ es también un funtor contravariante.

Notemos que estos hechos generalizan el teorema de Cauchy sobre integrales curvilíneas. En efecto, si c es una 2-cadena en una superficie analítica S y ω es una forma diferencial holomorfa en S, entonces

$$\int_{\partial c} \omega = \int_c d\omega = \int_c 0 = 0.$$

El teorema de Cauchy clásico se sigue de aquí porque si S es simplemente conexa entonces $H_1(S) = 0$, es decir, todo ciclo es una frontera.

Supongamos ahora que S puede cubrirse por una carta z y que $\omega = f\,dz$ es una forma diferencial meromorfa en S (que, por consiguiente, es una forma holomorfa en S menos un número finito de puntos). Sea $\tilde{f} = z^{-1} \circ f$ la lectura de f en la carta. Es claro entonces que $\omega = z^*(\tilde{f}\,dz)$, donde en el segundo miembro z es la identidad en $\mathbb{C}$. Para cada $x \in S$, el desarrollo de f en potencias de $z - z(x)$ es el mismo que el de $\tilde{f}$, luego, en particular, $\operatorname{Res}_x \omega = \operatorname{Res}_{z(x)} \tilde{f}$. Esto nos permite traducir el teorema de los residuos a integrales en variedades:

Supongamos que $\sigma : \Delta_2 \longrightarrow S$ es un 2-símplice positivamente orientado que se extienda a un difeomorfismo y cuya frontera no contenga polos de ω. Entonces

$$\int_{\partial\sigma} \omega = \int_{\partial\sigma} z^\sharp(\tilde{f}\,dz) = \int_{z_\sharp(\partial\sigma)} \tilde{f}\,dz = \int_{\partial(\sigma\circ z)} \tilde{f}\,dz = 2\pi i \sum_x \operatorname{Res}_x \omega,$$

donde x recorre los polos de ω contenidos en la imagen de σ (que se corresponden con los polos de $\tilde{f}$ contenidos en la imagen de $\sigma \circ z$).

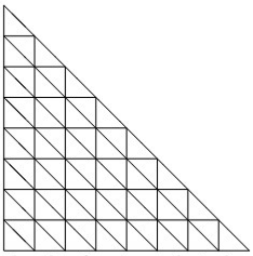

Un poco más laxamente: si $\sigma : \Delta_2 \longrightarrow S$ es un 2-símplice arbitrario y ω es una forma meromorfa que no tenga polos en la imagen de c, por el lema del cubrimiento de Lebesgue [TA 1.1], podemos subdividir Δ_2 en triángulos suficientemente pequeños como para que sus imágenes por σ estén contenidas en dominios de cartas de S. Además podemos retocar levemente la subdivisión para que ningún polo de ω se encuentre sobre los lados de los triángulos obtenidos. Entonces la integral de ω sobre $\partial\sigma$ es igual a la suma de las integrales de ω sobre las fronteras de todos los triángulos pequeños, pues las integrales sobre los lados interiores se cancelan dos a dos y las de los lados exteriores se suman hasta formar $\partial\sigma$.

Al igual que hemos visto antes (aunque sin suponer que los triángulos sean biyectivos) la integral de ω sobre cada uno de estos triángulos equivale a la integral de otra función $\tilde{f}$ sobre un ciclo en un abierto de $\mathbb{C}$, cuyos residuos se corresponden con (algunos de) los de ω. El teorema de los residuos nos da que la integral

$$\int_{\partial\sigma} \omega$$

es combinación lineal entera de los *periodos polares* $2\pi i \operatorname{Res}_P \omega$ de ω. Por linealidad lo mismo vale para integrales sobre fronteras arbitrarias.

Superficies de Riemann Supongamos ahora que S es una superficie de Riemann compacta (en tendiendo que las superficies de Riemann son conexas por defiinición). En virtud del teorema de clasificación de las superficies compactas, toda superficie de Riemann, como espacio topológico, puede obtenerse como cociente de un polígono regular $\mathcal{P}$ de $4g$ lados identificándolos dos a dos en la forma

$$a_1b_1a_1^{-1}b_1^{-1}a_2b_2a_2^{-1}b_2^{-1}\cdots a_gb_ga_g^{-1}b_g^{-1}.$$

(Véase la discusión previa a [VC A.25].) Sea $\phi : \mathcal{P} \longrightarrow S$ la aplicación continua y suprayectiva (inyectiva en el interior de $\mathcal{P}$) que identifica los lados de $\mathcal{P}$ en la forma indicada. Por ejemplo, la figura muestra el polígono cuyo cociente es la superficie de género 2:

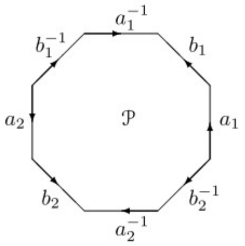

Cada arco a_k (recorrido en el sentido de la flecha) se corresponde en S con el mismo arco que a_k^{-1}, e igualmente con los b_k. En general, todos los vértices de $\mathcal{P}$ se identifican con un mismo punto de S, y así, en S tenemos $2g$ arcos cerrados $a_1, b_1, \ldots, a_g, b_g$.

En la prueba del teorema de clasificación se ve que el polígono $\mathcal{P}$ se construye a partir de una triangulación de S. Pero en la prueba de la fórmula de Hurwitz [VC A.25] (véase la nota posterior) se ve que si partimos de una triangulación diferenciable de una esfera podemos construir una triangulación diferenciable de cualquier superficie de Riemann, es decir, una triangulación formada por triángulos $\sigma : \Delta_2 \longrightarrow S$ que se extienden a difeomorfismos de un entorno simplemente conexo de Δ_2 en un entorno de la imagen.

Añadiendo este hecho a la prueba del teorema de clasificación se ve que la aplicación $\phi : \mathcal{P} \longrightarrow S$ se puede tomar diferenciable a trozos, es decir, que se puede triangular $\mathcal{P}$ de modo que ϕ coincida con un difeomorfismo en cada triángulo.

Si además elegimos los triángulos de $\mathcal{P}$ de forma que conserven la orientación, tenemos que los lados interiores (compartidos por dos triángulos contiguos) se recorren necesariamente en sentidos opuestos, por lo que constituyen 1-cadenas homólogas en $\mathcal{P}$. Esto nos permite ver a la triangulación de $\mathcal{P}$ como una cadena $\zeta_{\mathcal{P}} \in C_2(\mathcal{P}, \mathbb{Z})$ cuya frontera es homóloga al ciclo

$$a_1 + b_1 - a_1^{-1} - b_1^{-1} + a_2 + \cdots$$

El hecho de que ϕ sea diferenciable a trozos hace que $\zeta_S = \phi^\sharp(\zeta_{\mathcal{P}})$ (que no es sino la triangulación de S) sea también una cadena en S, cuya frontera es ahora

$$a_1 + b_1 - a_1 - b_1 + \cdots = 0.$$

Se demuestra que las clases de los ciclos $a_1, b_1, \ldots, a_g, b_g$ forman una base de $H_1(S, \mathbb{Z})$. En lo sucesivo, salvo que indiquemos lo contrario, cuando hablemos de una base de $H_1(S)$ se entenderá que es una base obtenida de esta forma a partir de un polígono.

La cohomología de las formas holomorfas es especialmente simple: observemos que las únicas funciones holomorfas en S son las constantes, luego no existen formas diferenciales holomorfas exactas (salvo la forma nula). Esto significa que $\mathcal{H}^1(S) = \Lambda^1(S)$ (sin ninguna identificación).

8.2 Integración de formas meromorfas

En 7.24 hemos distinguido tres clases de diferenciales en un cuerpo de funciones algebraicas K. En particular, si K es el cuerpo de funciones meromorfas de una superficie de Riemann S, entonces los divisores primos de K se identifican con los puntos de S y las formas diferenciales de K son las formas diferenciales meromorfas en S. Las diferenciales que hemos llamado de primera clase son simplemente las diferenciales holomorfas, mientras que las diferenciales de segunda clase son las que tienen integral nula sobre las fronteras en S pues, según hemos visto en la sección anterior, la integral de una forma meromorfa ω sobre una frontera es una combinación lineal entera de sus periodos polares, es decir, de los números $2\pi i \operatorname{Res}_P \omega$.

Sea ω una forma diferencial en una superficie de Riemann S y sea $a_1, \ldots, a_g$, $b_1, \ldots, b_g$ una base de $H_1(S)$ en las condiciones de la sección anterior, es decir, obtenida a partir de una identificación $\phi : \mathcal{P} \longrightarrow S$, donde $\mathcal{P}$ es un polígono de $4g$ lados. Podemos suponer que los ciclos a_k, b_k no pasan por polos de ω (por ejemplo, porque existe un difeomorfismo de S en sí misma que transforma cualquier conjunto finito de puntos prefijado en cualquier otro [GD 1.25]).

Llamaremos *periodos cíclicos* (o simplemente periodos) de ω (respecto a la base fijada) a los números complejos

$$A_k = \int_{a_k} \omega, \qquad B_k = \int_{b_k} \omega, \qquad k = 1, \ldots, g.$$

Si z es un ciclo en S, tenemos que su clase de homología se expresa como

$$[z] = \sum_{k=1}^{g} (m_k[a_k] + n_k[b_k]), \tag{8.2}$$

para ciertos $m_k, n_k \in \mathbb{Z}$, luego

$$z = \sum_{k=1}^{g} (m_k a_k + n_k b_k) + \partial c,$$

para cierta 2-cadena c en S. Consecuentemente,

$$\int_z \omega = \sum_{k=1}^{g} (m_k A_k + n_k B_k) + \int_{\partial c} \omega.$$

Por el teorema de los residuos, la última integral es combinación lineal entera de los periodos polares de ω. En resumen:

Teorema 8.1 *Si S es una superficie de Riemann de género $g \geq 1$ y ω es una forma diferencial meromorfa en S, entonces las integrales de ω sobre ciclos en S recorren el grupo generado por los periodos (polares y cíclicos) de ω.*

Notemos que ciertamente se recorre todo el grupo porque integrando sobre una pequeña circunferencia alrededor de un polo obtenemos el correspondiente periodo polar. Este teorema muestra que, aunque los periodos de ω dependen de la elección de la base de homología respecto a la que se calculan, el grupo de periodos es independiente de ella.

Fijado un punto $O \in S$, la integral

$$\int_O^P \omega$$

define una función holomorfa multiforme en S menos los polos de ω. Dos valores de la integral en un mismo punto P (calculados con arcos distintos de extremos O y P) se diferencian en un elemento del grupo de periodos de ω.

En el estudio de las integrales abelianas será fundamental una fórmula que vamos a deducir a continuación. Fijemos una forma diferencial ω de primera clase y una forma η arbitraria. Tomamos una base de homología a_k, b_k construida a partir de una identificación $\phi : \mathcal{P} \longrightarrow S$, donde $\mathcal{P}$ es un polígono en forma canónica. Podemos exigir que los ciclos no pasen por polos de η. Llamemos

$$A_k = \int_{a_k} \omega, \quad B_k = \int_{B_k} \omega, \quad A_k' = \int_{a_k} \eta, \quad B_k' = \int_{b_k} \eta.$$

Sea $D \subset S$ la imagen por ϕ del interior de $\mathcal{P}$, que es un abierto simplemente conexo en S. Si P_1 y P_2 son puntos de D, podemos definir

$$\int_{P_1}^{P_2} \omega$$

como la integral sobre cualquier arco en D que una P_1 con P_2. Fijado un punto $O \in D$, la función $g : D \longrightarrow \mathbb{C}$ dada por

$$g(P) = \int_O^P \omega$$

es holomorfa (uniforme) en D.

Sea ζ_S una triangulación (orientada) de S diferenciable a trozos (vista como 2-cadena en S). Podemos suponer que los lados de los triángulos no pasan por polos de η. Sabemos que $\partial\zeta_S = 0$, lo cual significa que cada arista es compartida por dos triángulos, pero recorrida en sentidos opuestos. Observemos ahora que si $\sigma : \Delta_2 \longrightarrow S$ es uno de los triángulos, entonces la restricción de g al interior de su imagen se extiende a una función holomorfa en un entorno de $\sigma[\Delta_2]$. En efecto, basta considerar un abierto simplemente conexo U que contenga a $\sigma[\Delta_2]$. Fijado un punto P_1 en el interior de $\sigma[\Delta_2]$, tenemos que, para cualquier otro punto P en dicho interior,

$$g(P) = \int_O^{P_1} \omega + \int_{P_1}^P \omega,$$

donde la segunda integral se calcula respecto de cualquier arco contenido en U que una P_1 con P, pero esta integral define una función holomorfa en U, que nos permite definir la extensión holomorfa que buscábamos. Obviamente esta extensión es única.

La construcción muestra que si $P \in \sigma[\Delta_2]$, entonces

$$g(P) = \int_O^P \omega,$$

donde la integral se calcula sobre cualquier arco que una O con P y que esté contenido en D salvo a lo sumo en su extremo final.

Consideremos ahora un punto P situado sobre la arista común de dos triángulos σ_1 y σ_2 y vamos a calcular la relación entre los valores $g_1(P)$ y $g_2(P)$. Si $P \in D$ tenemos $g_1(P) = g_2(P) = g(P)$. En caso contrario P es la imagen por ϕ de dos puntos P_1 y P_2 situados en la frontera de $\mathcal{P}$, tal y como muestra la figura.

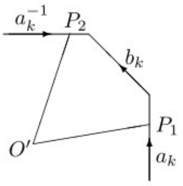

(En realidad la figura ilustra una posibilidad. La otra es que P_1 y P_2 estén sobre dos lados b_k y b_k^{-1}, en cuyo caso la flecha del lado intermedio tendría sentido opuesto.) Llamamos O' al punto de $\mathcal{P}$ que se corresponde con O.

La poligonal γ que va de O' a P_1, de P_1 a P_2 sobre la frontera de $\mathcal{P}$ y de P_2 a O' es un ciclo en $\mathcal{P}$, y como éste es simplemente conexo, γ es una frontera. Por consiguiente, su imagen $\phi^\sharp(\gamma)$ es una frontera en S, luego

$$0 = \int_{\phi^\sharp(\gamma)} \omega = g_1(P) + B_k - g_2(P),$$

donde hemos usado que las integrales sobre los dos fragmentos de a_k y a_k^{-1} se cancelan porque, en S, son integrales de la misma forma ω sobre el mismo arco recorrido en sentidos opuestos. Así pues, $g_1(P) - g_2(P) = -B_k$.

Similarmente, si P_1 está en b_k y P_2 en b_k^{-1}, resulta $g_1(P) - g_2(P) = A_k$.

Sea $\zeta_S = \sigma_1 + \cdots + \sigma_r$ la triangulación que hemos tomado. Definimos

$$\Sigma = \sum_{j=1}^{r} \int_{\partial\sigma_j} g_j\eta, \tag{8.3}$$

donde g_j es la extensión de g al triángulo σ_j.

Cada lado interior a D es recorrido dos veces en sentidos opuestos y, sobre ellos, la función g_j es la misma en ambos casos (pues coincide con la función g definida sobre D), luego las integrales se cancelan, y sólo quedan los términos correspondientes a lados contenidos en los arcos a_k y b_k. Éstos no se cancelan, porque para cada punto $P = \phi(P_1) = \phi(P_2)$ situado sobre uno de estos lados, con P_1 sobre a_k (o b_k) y P_2 sobre a_k^{-1} (o b_k^{-1}), la función g_j toma valores diferentes, digamos $g^+(P)$ cuando recorremos a_k o b_k positivamente y $g^-(P)$ cuando lo hacemos negativamente. El resultado es

$$\begin{aligned} S &= \sum_{k=1}^{g} \left(\int_{a_k} g^+\eta + \int_{b_k} g^+\eta - \int_{a_k} g^-\eta - \int_{b_k} g^-\eta \right) \\ &= \sum_{k=1}^{g} \left(\int_{b_k} (g^+ - g^-)\eta + \int_{a_k} (g^+ - g^-)\eta \right). \end{aligned}$$

Pero hemos visto que $g^+(P) - g^-(P) = A_k$ sobre b_k y $g^+(P) - g^-(P) = -B_k$ sobre a_k. Por consiguiente,

$$\Sigma = \sum_{k=1}^{g} (A_k B_k' - B_k A_k').$$

Por otra parte, podemos evaluar Σ calculando la integral sobre cada triángulo. Observemos que $g_j\eta$ es una forma diferencial meromorfa en un entorno (simplemente conexo) de cada uno de ellos. Subdividiendo la triangulación si es preciso, no perdemos generalidad si suponemos que dicho entorno es el dominio de una carta. Así, aplicando el teorema de los residuos vemos que

$$\Sigma = 2\pi i \sum_P \operatorname{Res}_P(g\eta),$$

donde P recorre los polos de η y g es ahora la función holomorfa (fija) definida en D. Si suponemos que η es una diferencial de tercera clase, entonces se cumple $\operatorname{Res}_P(g\eta) = g(P)\operatorname{Res}_P \eta$, pues si $g(P) = 0$ entonces $g\eta$ es holomorfa en P y su residuo es nulo.

En resumen, si ω es una diferencial de primera clase y η de tercera clase, sus periodos cíclicos A_k, B_k, A'_k, B'_k y los periodos polares de η (es decir, los números $2\pi i \operatorname{Res}_P \eta$) satisfacen la relación:

$$\sum_{k=1}^{g} (A_k B'_k - B_k A'_k) = 2\pi i \sum_P g(P) \operatorname{Res}_P \eta. \tag{8.4}$$

Si η es también una diferencial de primera clase, el segundo miembro es nulo, con lo que obtenemos el siguiente caso particular:

Teorema 8.2 (Primeras relaciones de Riemann) *Sean ω y η dos diferenciales de primera clase en una superficie de Riemann S de género $g \geq 1$, sean A_k, B_k los periodos de ω y A'_k, B'_k los periodos de η. Entonces*

$$\sum_{k=1}^{g} (A_k B'_k - B_k A'_k) = 0.$$

Más importante es la relación siguiente, pues de ella se deduce que si dos diferenciales de primera clase tienen los mismos A-periodos (o los mismos B-periodos) entonces son iguales:

Teorema 8.3 (Segundas relaciones de Riemann) *Si A_k y B_k son los periodos de una diferencial de primera clase no nula ω en una superficie de Riemann S de género $g \geq 1$, entonces*

$$i \sum_{k=1}^{g} (A_k \bar{B}_k - B_k \bar{A}_k) > 0.$$

Demostración: Consideramos

$$\Sigma' = \sum_{j=1}^{r} \int_{\partial\sigma_j} \bar{g}_j \omega,$$

donde g_j y σ_j son los mismos que en (8.3). El mismo razonamiento que hemos aplicado a Σ nos da ahora que

$$\Sigma' = \sum_{k=1}^{g} (\bar{A}_k B_k - \bar{B}_k A_k),$$

luego basta probar que, para cada j,

$$-i \int_{\partial\sigma_j} \bar{g}_j \omega > 0.$$

Recordemos que, salvo un sumando constante, la función g_j se obtiene integrando ω desde un punto fijo de σ_j, por lo que $\omega|_{\sigma_j[\Delta_2]} = dg_j$. Sin embargo, ahora no podemos decir que la integral sobre $\partial\sigma_j$ es nula porque la forma $\bar{g}_j\omega$ no es holomorfa. Vamos a calcularla explícitamente.

Sea $g_j = u + iv$, de modo que $\omega = dg_j = du + idv$. Aplicamos el teorema de Stokes:

$$-i\int_{\partial\sigma_j} \bar{g}_j\omega = -i\int_{\partial\sigma_j}((u\,du + v\,dv) + i(u\,dv - v\,du)) = \int_{\sigma_j} du \wedge dv.$$

Podemos suponer que $\sigma_j[\Delta_2]$ está contenido en el dominio de una carta z, con lo que

$$\begin{aligned} du \wedge dv &= \left(\frac{\partial\operatorname{Re} g_j}{\partial x}\,dx + \frac{\partial\operatorname{Re} g_j}{\partial y}\,dy\right) \wedge \left(\frac{\partial\operatorname{Im} g_j}{\partial x}\,dx + \frac{\partial\operatorname{Im} g_j}{\partial y}\,dy\right) \\ &= \left(\frac{\partial\operatorname{Re} g_j}{\partial x}\frac{\partial\operatorname{Im} g_j}{\partial y} - \frac{\partial\operatorname{Re} g_j}{\partial y}\frac{\partial\operatorname{Im} g_j}{\partial x}\right) dx \wedge dy \\ &= \left(\left(\frac{\partial\operatorname{Re} g_j}{\partial x}\right)^2 + \left(\frac{\partial\operatorname{Re} g_j}{\partial y}\right)^2\right) dx \wedge dy. \end{aligned}$$

Así pues, al transportar la integral mediante la carta, obtenemos

$$\int_{\sigma_j} du \wedge dv = \int_{z_\sharp(\sigma_j)} \left(\left(z^{-1} \circ \frac{\partial\operatorname{Re} g_j}{\partial x}\right)^2 + \left(z^{-1} \circ \frac{\partial\operatorname{Re} g_j}{\partial y}\right)^2\right) dxdy,$$

que es un número real estrictamente positivo, ya que si fuera cero entonces $\operatorname{Re} g_j$ sería constante en un abierto de S, al igual que $\operatorname{Im} g_j$ por las ecuaciones de Cauchy-Riemann, pero entonces g_j sería constante y ω sería nula. ■

Como ya hemos dicho, este teorema implica que si una diferencial de primera clase ω tiene todos sus A-periodos (o todos sus B-periodos) nulos, entonces es la forma nula. Más aún, comparando los dos teoremas anteriores vemos que si ω tiene todos sus periodos reales entonces ha de ser la forma nula.

Las relaciones de Riemann pueden expresarse matricialmente:

Teorema 8.4 (Relaciones de Riemann) *Sea $\Omega = (\omega_1, \ldots, \omega_g)$ una base del espacio de diferenciales de primera clase en una superficie de Riemann de género $g \geq 1$ y sean A_{jk}, B_{jk} los periodos de ω_j, que forman sendas matrices cuadradas A y B. Entonces $AB^t = BA^t$ y la matriz $i(A\bar{B}^t - B\bar{A}^t)$ es definida positiva.*

DEMOSTRACIÓN: No sólo vamos a probar lo que afirma el teorema, sino, de hecho, que este enunciado es equivalente a los enunciados de los dos teoremas precedentes.

Si $\omega = \alpha\Omega$ y $\eta = \beta\Omega$ son dos diferenciales de primera clase arbitrarias (donde α, $\beta \in \mathbb{C}^g$), los vectores de periodos de ω son αA, αB, y los de η son βA, βB. Las primeras relaciones de Riemann afirman que $\alpha AB^t\beta^t - \alpha BA^t\beta^t = 0$ o, equivalentemente, que $\alpha(AB^t - BA^t)\beta^t = 0$ para todos los α, $\beta \in \mathbb{C}^g$. Esto se reduce a la igualdad $AB^t = BA^t$.

Las segundas relaciones de Riemann afirman que si $\alpha \neq 0$ entonces

$$i(\alpha A\bar{B}^t\bar{\alpha}^t - \alpha B\bar{A}^t\bar{\alpha}^t) > 0$$

o, lo que es lo mismo, que $\alpha(i(A\bar{B}^t - B\bar{A}^t))\bar{\alpha}^t > 0$ para todo $\alpha \in \mathbb{C}^g$ no nulo. Esto es, por definición, que la matriz $i(A\bar{B}^t - B\bar{A}^t)$ es definida positiva. ■

Como consecuencia, la matriz A ha de ser regular, pues en caso contrario existiría un $\alpha \in \mathbb{C}^g$ no nulo tal que $\alpha A = 0$, lo que a su vez implicaría que $\alpha(i(A\bar{B}^t - B\bar{A}^t))\bar{\alpha}^t = 0$, en contradicción con el teorema anterior.

Observemos ahora que si M es una matriz regular y consideramos la base $\Omega' = \Omega M$, entonces $\omega'_i = \sum\limits_k m_{ki}\omega_k$, y el periodo A'_j de ω'_i es $A'_j = \sum\limits_k m_{ki}A_{kj}$, luego la matriz de A-periodos de Ω' es M^tA. En particular, tomando la matriz $M = (A^{-1})^t$, obtenemos una base cuya matriz de A-periodos es la identidad.

Definición 8.5 Sea S una superficie de Riemann de género $g \geq 1$. Una *base canónica* del espacio de diferenciales de primera clase de S (respecto de una base prefijada de $H_1(S)$) es una base tal que su matriz de A-periodos es la identidad.

Acabamos de demostrar que existen bases canónicas. Respecto a tales bases, las relaciones de Riemann se expresan de forma especialmente simple: las primeras relaciones afirman que la matriz B es simétrica, mientras que las segundas se reducen a que la matriz $i(\bar{B} - B) = 2\,\mathrm{Im}\,B$ es definida positiva. Puesto que se trata de una matriz real, esto implica que es definida positiva como matriz real, es decir, que $\alpha(\mathrm{Im}\,B)\alpha^t > 0$ para todo $\alpha \in \mathbb{R}^g$ no nulo.

A partir de una base canónica podemos formar diferenciales de primera clase con cualquier vector de A-periodos prefijado. Como consecuencia podemos dar una condición de unicidad en la descomposición de una forma diferencial:

Teorema 8.6 *Si ω es una forma diferencial en una superficie de Riemann de género $g \geq 1$, entonces ω se descompone de forma única como $\omega = \omega_1 + \omega_2 + \omega_3$, donde los sumandos son, respectivamente, una diferencial de primera clase, una diferencial de segunda clase con A-periodos nulos y una diferencial de tercera clase con A-periodos nulos (respecto de una base de homología prefijada cuyos ciclos no pasen por los polos de ω).*

Demostración: El teorema 7.26 nos da una diferencial de tercera clase ω'_3 con los mismos polos de orden 1 que ω y con los mismos residuos, luego $\omega'_2 = \omega - \omega_3$ es una diferencial de segunda clase. Sean ω'_1 y ω''_1 diferenciales de primera clase con los mismos A-periodos que ω'_2 y ω'_3 respectivamente. Entonces $\omega_2 = \omega'_2 - \omega'_1$ y $\omega_3 = \omega'_3 - \omega''_1$ son diferenciales de segunda y tercera clase con A-periodos nulos y, llamando $\omega_1 = \omega'_1 + \omega''_1$ tenemos $\omega = \omega_1 + \omega_2 + \omega_3$.

La descomposición es única, pues si $\omega = \omega'_1 + \omega'_2 + \omega'_3$ es otra descomposición en las mismas condiciones, entonces ω_1 y ω'_1 son diferenciales de primera clase con los mismos A-periodos (los de ω), luego $\omega_1 - \omega'_1$ tiene A-periodos nulos y es, por consiguiente, la forma nula.

Similarmente, ω_3 y ω'_3 tienen los mismos polos simples que ω con los mismos residuos, luego su diferencia no tiene polos, luego se trata también de una diferencial de primera clase con A-periodos nulos y por lo tanto $\omega_3 = \omega'_3$. Necesariamente entonces $\omega_2 = \omega'_2$. ■

El grupo de periodos de una forma diferencial es en general denso en $\mathbb{C}$, por lo que calcular una integral salvo periodos es hacer poco. La situación es distinta si trabajamos vectorialmente, como vamos a ver a continuación:

Sea $\Omega = (\omega_1, \ldots, \omega_g)$ una base del espacio de diferenciales de primera clase en una superficie de Riemann S. Definimos

$$A_k = \int_{a_k} \Omega = \left(\int_{a_k} \omega_1, \ldots, \int_{a_k} \omega_g \right) \in \mathbb{C}^g,$$

$$B_k = \int_{b_k} \Omega = \left(\int_{b_k} \omega_1, \ldots, \int_{b_k} \omega_g \right) \in \mathbb{C}^g.$$

A estos vectores los llamaremos *periodos* de Ω. Si z es un ciclo en S, la relación (8.2) implica que

$$\int_z \Omega = \left(\int_z \omega_1, \ldots, \int_z \omega_g \right)$$

es una combinación lineal entera de los periodos de Ω, por lo que si P, $Q \in S$, la integral

$$\int_P^Q \Omega = \left(\int_P^Q \omega_1, \ldots, \int_P^Q \omega_g \right),$$

(donde todas las integrales se calculan sobre un mismo arco que una P con Q) está definida salvo combinaciones enteras de los periodos de Ω. La diferencia es que, según se deduce del teorema siguiente, el grupo generado por los periodos es ahora un subespacio discreto de $\mathbb{C}^g$.

Teorema 8.7 *Sea S una superficie de Riemann de género $g \geq 1$ y sea Ω una base de las diferenciales de primera clase. Entonces, los periodos A_k, B_k de Ω (respecto a una base de $H_1(S)$) son linealmente independientes sobre $\mathbb{R}$.*

DEMOSTRACIÓN: Al construir las bases canónicas hemos visto que los periodos de dos bases distintas se corresponden por un automorfismo de $\mathbb{C}^g$, luego podemos suponer que la base Ω es canónica. Si vemos la matriz de periodos (A, B) como una matriz real de dimensión $2g \times 2g$, es decir, si desdoblamos cada fila en dos filas correspondientes a la parte real y la parte imaginaria, obtenemos una matriz de la forma

$$\left(\begin{array}{c|c} I & \operatorname{Re} B \\ \hline 0 & \operatorname{Im} B \end{array} \right)$$

Sabemos que $\operatorname{Im} B$ es definida positiva, luego en particular es regular, luego la matriz tiene determinante no nulo y sus columnas son linealmente independientes. ■

Por consiguiente, el grupo generado por los periodos de una base Ω es un retículo R_Ω en $\mathbb{C}^g$. Observemos que no depende de la elección de la base de homología, ya que

$$R_\Omega = \left\{ \int_\gamma \Omega \;\Big|\; \gamma \in Z_1(S) \right\}.$$

Por otra parte, los retículos asociados a dos bases del espacio de diferenciales de primera clase se corresponden a través de un automorfismo de $\mathbb{C}^g$, lo que implica que el toro complejo $\mathbb{C}^g/J_\Omega$ está determinado por S salvo un isomorfismo inducido por un automorfismo de $\mathbb{C}^g$, que es claramente una aplicación biholomorfa.

Definición 8.8 Si S es una superficie de Riemann de género $g \geq 1$, definimos la *variedad jacobiana* de S como el toro complejo $J(S) = \mathbb{C}^g/J_\Omega$, donde Ω es una base del espacio de diferenciales de primera clase de S y J_Ω es el retículo generado por sus periodos.

En las próximas secciones estudiaremos la relación entre una superficie S y su variedad jacobiana. Ahora vamos a probar que las variedades jacobianas cumplen la hipótesis del teorema de Lefschetz 4.41, por lo que son proyectivas:

Teorema 8.9 *Sea S una superficie de Riemann de género $g \geq 1$, sea Ω una base de las diferenciales de primera clase y sea $R_\Omega \subset \mathbb{C}^g$ el retículo generado por los periodos A_k, B_k de Ω respecto a una base de $H_1(S)$. Entonces existe una forma de Riemann en $\mathbb{C}^g$ respecto de R_Ω. Por consiguiente, la variedad jacobiana $J(S)$ es proyectiva.*

DEMOSTRACIÓN: Como ya hemos observado, no perdemos generalidad si tomamos como Ω una base canónica. Consideremos los periodos A_k, B_k como una $\mathbb{R}$-base de $V = \mathbb{C}^g = \mathbb{R}^{2g}$ y sea x_k, y_k su base dual. Definimos la forma bilineal alternada $E : V \times V \longrightarrow \mathbb{R}$ mediante

$$E(u, v) = \sum_{j=1}^{g} (y_j(u)x_j(v) - x_j(u)y_j(v)).$$

Es obvio que E toma valores enteros sobre R_Ω. Para probar que E es una forma de Riemann sólo falta ver que la forma $S(u, v) = E(iu, v)$ es simétrica y definida positiva.

Dado $u \in V$, sean (α, β) sus coordenadas en la base de periodos, es decir, $u = \alpha + \beta B$. Por otra parte, descompongamos $u = a + ib$, donde a, $b \in \mathbb{R}^g$ y, del mismo modo, $B = P + iQ$, donde las matrices reales P y Q son simétricas y Q es definida positiva. En estos términos, $a + ib = u = \alpha + \beta P + i\beta Q$, luego $a = \alpha + \beta P$, $b = \beta Q$, luego

$$y_j(u) = \beta_j = (bQ^{-1})_j, \qquad x_j(u) = \alpha_j = (a - bQ^{-1}P)_j.$$

Por consiguiente, $y_j(iu) = (aQ^{-1})_j$, $x_j(iu) = -(b + aQ^{-1}P)_j$. Ahora tomamos dos vectores $u = a + ib$, $v = a' + ib'$ y calculamos:

$$S(u, v) = E(iu, v) = \sum_{j=1}^{g} (y_j(iu)x_j(v) - x_j(iu)y_j(v))$$

$$= \sum_{j=1}^{g} ((aQ^{-1})_j(a' - b'Q^{-1}P)_j + (b + aQ^{-1}P)_j(b'Q^{-1})_j)$$

$$= (aQ^{-1})(a' - b'Q^{-1}P) + (b + aQ^{-1}P)(b'Q^{-1}) = aQ^{-1}a'^t + bQ^{-1}b'^t,$$

donde hemos usado que todas las matrices son simétricas. Ahora es inmediato que la forma S es simétrica. Además,

$$S(u,u) = aQ^{-1}a^t + bQ^{-1}b^t.$$

Ahora basta tener en cuenta que si Q es definida positiva Q^{-1} también lo es.[1]

■

8.3 El teorema de Abel

El teorema de Abel nos da una caracterización en términos de integrales de los divisores principales de una superficie de Riemann S. Equivalentemente, nos da una condición necesaria y suficiente para que exista una función meromorfa en S con una distribución dada de ceros y polos. Si Ω es una base del espacio de las diferenciales de primera clase de S, en la sección anterior hemos visto que la integral

$$\int_P^Q \Omega$$

está bien definida módulo los periodos de Ω, es decir, como elemento de la variedad jacobiana $J(S)$, con independencia del arco de extremos P y Q con que la calculemos.

Si identificamos los puntos de S con los divisores primos del cuerpo $\mathcal{M}(S)$ de las funciones meromorfas en S, entonces la integral (con un origen fijo $O \in S$) se extiende por linealidad a un homomorfismo sobre todo el grupo de divisores de S, de modo que

$$\int_O^{\mathfrak{a}} \Omega = \sum_Q v_Q(\mathfrak{a}) \int_O^Q \Omega \in J.$$

La restricción de este homomorfismo al grupo de los divisores de grado 0 es independiente de la elección de O, pues

$$\int_O^{\mathfrak{a}} \Omega - \int_{O'}^{\mathfrak{a}} \Omega = \sum_Q v_Q(\mathfrak{a}) \int_O^{O'} \Omega = 0.$$

Tenemos así un homomorfismo natural del grupo de divisores de grado 0 en la variedad jacobiana J. El teorema de Abel dice esencialmente que el núcleo de este homomorfismo es el subgrupo de los divisores principales. Para probarlo usaremos la versión vectorial de la relación (8.4), que se prueba sin más que aplicar dicha relación componente a componente:

Teorema 8.10 *Sea S una superficie de Riemann de género $g \geq 1$, sea Ω una base del espacio de diferenciales de primera clase en S y η una diferencial de*

---

[1] Existe una matriz regular H tal que $Q' = HQH^t$ es diagonal [TAl 6.3]. Entonces Q' es definida positiva, lo que equivale a que todos los coeficientes de su diagonal sean positivos. Lo mismo le sucede a $Q'^{-1} = H^{-1t}Q^{-1}H^{-1}$, luego Q^{-1} es definida positiva.

geometria-algebraica-de-carlos-
ivorra.pdf

Pág. 301 (311) Capitulo VIII

Integrales abelianos

8.1 Homología y cohomología

(Pág. 304 (314))

- Antonio F. Costa - ejercicios de geometría diferencial de curvas y superficies... (2005).pdf
(804 pag)
- Important: Pi
- 3. Hist. Form. y Alg. para. PI.pdf

obs.: El Duelo de los Números
Javier Cilleruelo Mateo
(20 pág)

Source:
La Gaceta de la RSME, Vol. 10.1 (2007), Págs. 159-178

+,-.+ + +/ ! + ') , ? ! A7 #
C! # ! # 8! 2 3 ! 6 2 0 " 3#
/4 > 8 ! @@@

+,-.+ + +/ ! + ') ' D ? 2 , 3! ! & 3#
/4 & ! @@@

La Gaceta de la RSME, Vol. 10.1 (2007), Págs. 159–178

El diablo de los números

Sección a cargo de

Javier Cilleruelo Mateo

Historia de las fórmulas y algoritmos para π

por

Jesús Guillera Goyanes

A lo largo de más de dos milenios se han obtenido muchas fórmulas, algunas de gran belleza, para el número π. Entre ellas: series, productos infinitos, desarrollos en fracciones continuas y desarrollos con radicales. También se han conseguido evaluar algunas expresiones que, ¡oh maravilla!, estaban relacionadas con π. Además ha habido una continua batalla por conseguir records de cálculo de las cifras de este número, que se han ido estableciendo con series de convergencia rápida, algoritmos ultrarápidos y otros de cálculo aislado de cifras realmente sorprendentes. En estos logros de cálculo el desarrollo de ordenadores cada vez más potentes ha jugado un papel esencial.

1 . La primera fórmula: El algoritmo de Arquímedes

El algoritmo para el cálculo de π que Arquímedes demostró unos 250 años A.C. constituyó durante un periodo de aproximadamente 1800 años la forma más eficiente para el cálculo de π. La idea consistía en considerar un círculo de diámetro unidad y polígonos circunscritos e inscritos de $3 \cdot 2^n$ lados. Denotando sus perímetros con a_n y b_n respectivamente, Arquímedes, con razonamientos puramente geométricos, demostró las relaciones

$$a_1 = 2\sqrt{3}, \quad b_1 = 3, \qquad a_{n+1} = \frac{2a_n b_n}{a_n + b_n}, \qquad b_{n+1} = \sqrt{a_{n+1} b_n}. \qquad (1)$$

Obviamente $b_n < \pi < a_n$ y las sucesiones a_n y b_n convergen hacia el número π. Se trata de un algoritmo que hoy en día podemos demostrar fácilmente usando trigonometría elemental. En efecto, llamando $k_n = 3 \cdot 2^n$ se tienen las relaciones

$$a_n = k_n \tan \frac{\pi}{k_n}, \qquad b_n = k_n \mathrm{sen} \frac{\pi}{k_n}.$$

Por lo tanto

$$\frac{2a_n b_n}{a_n + b_n} = 2k_n \frac{\tan \frac{\pi}{k_n} \mathrm{sen} \frac{\pi}{k_n}}{\tan \frac{\pi}{k_n} + \mathrm{sen} \frac{\pi}{k_n}} = 2k_n \tan \frac{\pi}{2k_n} = a_{n+1}.$$

Por otra parte

$$\sqrt{a_{n+1} b_n} = \sqrt{2k_n \tan \frac{\pi}{2k_n} \cdot k_n \mathrm{sen} \frac{\pi}{k_n}} = 2k_n \sqrt{\tan \frac{\pi}{2k_n} \mathrm{sen} \frac{\pi}{2k_n} \cos \frac{\pi}{2k_n}} = b_{n+1}.$$

La convergencia del algoritmo es lineal, cada 5 iteraciones se obtienen 3 cifras de π. Para verlo utilizamos las desigualdades $\tan x < x$ y $1 - \cos x < \frac{x^2}{2}$. Se tiene entonces

$$a_n - b_n = k_n \tan \frac{\pi}{k_n} \left(1 - \cos \frac{\pi}{k_n}\right) < \frac{\pi^3}{2k_n^2} = \frac{\pi^3}{18 \cdot 2^{2n}} < \frac{\pi^3}{18 \cdot 1000^{n/5}}.$$

Para $n = 7$ este algoritmo proporciona para π la acotación $3{,}1415 < \pi < 3{,}1417$.

Es instructivo señalar que en la época de Arquímedes no se conocía la notación decimal de los números ni ninguna otra notación posicional, por lo que el resultado que obtuvo con polígonos de 96 lados ($n = 5$) lo expreso con fracciones:

$$3 + \frac{10}{71} < \pi < 3 + \frac{1}{7}.$$

Esta acotación determina dos cifras decimales de π.

2 . SIN GRANDES NOVEDADES HASTA 1800 AÑOS DESPUÉS

Durante un periodo de aproximadamente 1800 años todas las fórmulas que se obtuvieron para π se basaron en la idea de aproximar el círculo mediante polígonos regulares y tienen la forma de algoritmos iterativos. En el año 1593 Viète obtiene una expresión diferente mediante un bonito desarrollo con infinitos radicales anidados:

$$\frac{2}{\pi} = \sqrt{\frac{1}{2}} \cdot \sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2}}} \cdot \sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2}}}} \cdots . \tag{2}$$

El método usado por Viète se basa, como todos los anteriores, en la geometría y consiste en considerar un círculo de diámetro 2 e ir aproximando su área, es decir π, sucesivamente por las áreas de un cuadrado inscrito, de un octógono inscrito, de un polígono inscrito de 16 lados, etc. De esta manera Viète obtiene π con 9 cifras decimales: 3,141592653.

Hacia el año 1650 los métodos analíticos de Descartes y Fermat, precursores del cálculo infinitesimal que pronto sería desarrollado, comenzaron a sustituir a los métodos geométricos. Uno de los resultados para π correspondiente a esta etapa es el famoso producto infinito de Wallis (1655):

$$\frac{2}{\pi} = \frac{1\cdot 3}{2\cdot 2}\cdot\frac{3\cdot 5}{4\cdot 4}\cdot\frac{5\cdot 7}{6\cdot 6}\cdot\frac{7\cdot 9}{8\cdot 8}\cdots \tag{3}$$

En 1999 Osler demostró una fórmula que incluye como casos particulares tanto la fórmula de Viète como el producto infinito de Wallis. La veremos en la sección 10.

3 . COMIENZOS DEL CÁLCULO INFINITESIMAL: NUEVAS FÓRMULAS PARA π

Entre 1665 y 1680 Newton y Leibniz desarrollaron de forma independiente los principios básicos del cálculo infinitesimal que proporcionaron una poderosa herramienta de cálculo no sólo para las matemáticas sino también para la física y otras ciencias. Estos métodos, como no podía ser de otra manera, tienen también su repercusión en la obtención de nuevas fórmulas para π como por ejemplo la obtenida por Newton

$$\pi = \frac{3\sqrt{3}}{4} + 2 - \frac{3}{4}\sum_{n=0}^{\infty}\frac{\binom{2n}{n}}{(n+1)(2n+5)16^n}, \tag{4}$$

con la cual calculó 15 decimales de π en el año 1665. Newton llega a esta fórmula observando que el área limitada entre la circunferencia de ecuación

$$\left(x-\frac{1}{2}\right)^2 + y^2 = \frac{1}{4}$$

y las rectas $x = \frac{1}{4}$ e $y = 0$ es igual a la diferencia entre el área de un sector circular que abarca un ángulo de 60^o y su triángulo correspondiente y que por lo tanto

$$\int_0^{1/4} y\,dx = \int_0^{1/4}\sqrt{x-x^2}dx = \int_0^{1/4} x^{1/2}(1-x)^{1/2}dx = \frac{\pi}{24} - \frac{\sqrt{3}}{32}.$$

La fórmula de Newton se obtiene a partir de aquí desarrollando $(1-x)^{1/2}$ por medio de la fórmula del binomio, que él mismo había demostrado, e integrando a continuación término a término. Fórmulas como la (4) se conocen actualmente como sumas binomiales. Una serie similar a la de Newton pero más sencilla es

$$\frac{\pi}{3} = \sum_{n=1}^{\infty} \frac{\binom{2n}{n}}{(2n+1)16^n}, \tag{5}$$

que se obtiene sustituyendo $x = \frac{1}{4}$ en el desarrollo en serie (ver [3], pág. 386)

$$\frac{\operatorname{arc\,sen} 2x}{2x} = \sum_{n=1}^{\infty} \frac{\binom{2n}{n}}{2n+1} x^{2n}.$$

Otro ejemplo interesante de suma binomial es la obtenida por Comtet en 1974

$$\pi^4 = \frac{3240}{17} \sum_{n=1}^{\infty} \frac{1}{n^4 \binom{2n}{n}} \tag{6}$$

que se demuestra a partir de la identidad [5]

$$\sum_{n=1}^{\infty} \frac{1}{n^4 \binom{2n}{n}} = 2 \int_0^{\pi/3} \theta \log^2 \left(2\text{sen} \frac{\theta}{2} \right) d\theta.$$

Gregory en 1671 integra término a término la serie geométrica

$$1 - x^2 + x^4 - x^6 + \cdots = \frac{1}{1+x^2}, \qquad |x| < 1$$

y obtiene la fórmula (que lleva su nombre)

$$\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots, \qquad |x| \leq 1,$$

que es lo que luego se conocería como desarrollo en serie de potencias de x de la función $\arctan x$. Sustituyendo $x = 1$, Leibniz obtiene en 1674 la serie para π

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots. \tag{7}$$

Es extraño que Gregory no diera él mismo este paso, seguramente no lo hizo porque no estaba interesado en series de convergencia lenta. Sharp sustituye $x = \sqrt{3}/3$ en la fórmula de Gregory y obtiene la serie

$$\frac{\pi}{6} = \frac{\sqrt{3}}{3} \left(1 - \frac{1}{3 \cdot 3} + \frac{1}{5 \cdot 3^2} - \frac{1}{7 \cdot 3^3} + \cdots \right) \tag{8}$$

que le sirve para conseguir en 1699 un record de 71 cifras decimales de π.

En 1665 Brounker halla el bonito desarrollo en fracción continua

$$\frac{4}{\pi} = 1 + \cfrac{1^2}{2 + \cfrac{3^2}{2 + \cfrac{5^2}{2 + \cfrac{7^2}{2 + \cfrac{9^2}{2 + \ddots}}}}}. \tag{9}$$

Unos 100 años más tarde Euler la demuestra de forma ingeniosa: Comienza con la siguiente igualdad

$$a_1 + a_1a_2 + a_1a_2a_3 + a_1a_2a_3a_4 \cdots = \cfrac{a_1}{1 - \cfrac{a_2}{1 + a_2 - \cfrac{a_3}{1 + a_3 - \cfrac{a_4}{1 + a_4 - \ddots}}}},$$

lo que permite representar la serie de Gregory escrita en la forma

$$\arctan x = x + x\left(\frac{-x^2}{3}\right) + x\left(\frac{-x^2}{3}\right)\left(\frac{-3x^2}{5}\right) + x\left(\frac{-x^2}{3}\right)\left(\frac{-3x^2}{5}\right)\left(\frac{-5x^2}{7}\right) + \cdots$$

mediante una fracción continua de la que obtiene la fórmula de Brounker sustituyendo $x = 1$.

Un ejemplo reciente de desarrollo de π en fracción continua es [16]

$$\pi = 3 + \cfrac{1^2}{6 + \cfrac{3^2}{6 + \cfrac{5^2}{6 + \cfrac{7^2}{6 + \cfrac{9^2}{6 + \ddots}}}}}, \tag{10}$$

obtenido en 1999 por Lange.

4 . SERIES RÁPIDAS DE TIPO MACHIN

En 1706 John Machin demostró la fórmula que lleva su nombre

$$\frac{\pi}{4} = 4 \arctan \frac{1}{5} - \arctan \frac{1}{239} \tag{11}$$

y utilizando la fórmula de Gregory la escribió en forma de serie y consiguió con ella calcular los 100 primeros decimales de π. Las siguientes identidades son del mismo tipo pero conducen a series de convergencia más lenta:

$$\frac{\pi}{4} = \arctan\frac{1}{2} + \arctan\frac{1}{3}, \qquad \frac{\pi}{4} = 2\arctan\frac{1}{3} + \arctan\frac{1}{7}. \tag{12}$$

Demostrar una fórmula de tipo Machin, una vez conocida, se reduce a la comprobación rutinaria de una identidad entre operaciones elementales con números complejos. Así, la identidad

$$\frac{(5+i)^4}{239+i} = 2+2i$$

encapsula la demostración de la fórmula de Machin –basta con igualar los argumentos. Gauss también participó en la búsqueda de series para π de convergencia rápida y en 1863 demostró la siguiente identidad

$$\frac{\pi}{4} = 12\arctan\frac{1}{18} + 8\arctan\frac{1}{57} - 5\arctan\frac{1}{239}, \tag{13}$$

que es como las anteriores pero que involucra un término más. En el año 2002 Kanada y su equipo, empleando ciertos trucos para hacer más eficaces los cálculos por ordenador, consiguieron un record de cálculo de π por medio de las series:

$$\frac{\pi}{4} = 12\arctan\frac{1}{49} + 32\arctan\frac{1}{57} - 5\arctan\frac{1}{239} + 12\arctan\frac{1}{110443}, \tag{14}$$

$$\frac{\pi}{4} = 44\arctan\frac{1}{57} + 7\arctan\frac{1}{239} - 12\arctan\frac{1}{682} + 24\arctan\frac{1}{12943}, \tag{15}$$

la primera obtenida en 1982 por Takano (profesor de secundaria y compositor de canciones) y la segunda en 1896 por Störmer. El record consistió en más de un billón de cifras decimales de π.

5 . LAS FÓRMULAS DEL GENIAL MATEMÁTICO SUIZO L. EULER

Grandes matemáticos, entre ellos Leibniz y los hermanos Jacob (1654-1705) and Johann Bernoulli (1667-1748) habían intentado sin lograrlo evaluar la suma infinita

$$\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \cdots . \tag{16}$$

Finalmente Euler, en 1736, demostró que su valor era $\frac{\pi^2}{6}$. Para ello, partió del desarrollo en serie de la función $\frac{\operatorname{sen} x}{x}$

$$\frac{\operatorname{sen} x}{x} = 1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \frac{x^6}{7!} + \cdots$$

y consideró la ecuación

$$1-\frac{x^2}{3!}+\frac{x^4}{5!}-\frac{x^6}{7!}+\cdots=0.$$

Las soluciones de esta ecuación tenían que ser las soluciones de sen$x = 0$, excepto lógicamente $x = 0$, es decir: $\pm\pi$, $\pm 2\pi$, $\pm 3\pi$, $\cdots$. Imaginó la serie infinita como si se tratase de un polinomio y procedió a descomponerlo en factores

$$1-\frac{x^2}{3!}+\frac{x^4}{5!}-\frac{x^6}{7!}+\cdots=\left(1-\frac{x^2}{\pi^2}\right)\left(1-\frac{x^2}{4\pi^2}\right)\left(1-\frac{x^2}{9\pi^2}\right)\cdots.$$

Igualando los coeficientes de x^2 obtuvo finalmente la demostración. Además, demostró muchas fórmulas parecidas, entre ellas

$$\sum_{n=1}^{\infty}\frac{1}{(2n+1)^2}=\frac{\pi^2}{8},\qquad \sum_{n=1}^{\infty}\frac{1}{n^4}=\frac{\pi^4}{90},\qquad \sum_{n=0}^{\infty}\frac{(-1)^n}{(2n+1)^3}=\frac{\pi^3}{32} \tag{17}$$

y la no tan parecida

$$\sum_{n=1}^{\infty}\frac{1+\frac{1}{2}+\frac{1}{3}+\cdots\frac{1}{n}}{n^3}=\frac{\pi^4}{72}. \tag{18}$$

Fórmulas similares a esta última se conocen hoy en día como sumas de Euler [10]. Algunas de ellas sólo se han conseguido sumar muy recientemente, como por ejemplo

$$\sum_{n=1}^{\infty}\frac{\left(1+\frac{1}{2}+\frac{1}{3}+\cdots\frac{1}{n}\right)^2}{n^2}=\frac{17\pi^4}{360} \tag{19}$$

cuya demostración se obtuvo en 1991.

Euler, en su estudio de la divergencia de los inversos de los números primos, consideró progresiones geométricas de la forma:

$$1+\frac{1}{p^2}+\frac{1}{p^4}+\frac{1}{p^6}+\frac{1}{p^8}+\cdots=\frac{1}{1-\frac{1}{p^2}},$$

y observó que multiplicando estos desarrollos para todos los primos p se obtiene

$$\sum_{n=1}^{\infty}\frac{1}{n^2}=\frac{\pi^2}{6}=\prod_{p}\left(1-\frac{1}{p^2}\right)^{-1}, \tag{20}$$

que relaciona el número π con los números primos.

6 . UN EXTRAORDINARIO ALGORITMO OLVIDADO: FÓRMULA DE C. F. GAUSS

Con tan sólo 14 años de edad, Gauss (uno de los más grandes matemáticos de todos los tiempos) considera el siguiente proceso iterativo

$$a_0 = a, \quad b_0 = b, \quad a_{n+1} = \frac{a_n + b_n}{2}, \qquad b_{n+1} = \sqrt{a_n b_n}$$

y define la media aritmético-geométrica de a y b como el límite común

$$M(a, b) = \lim a_n = \lim b_n.$$

En el año 1800 sus investigaciones sobre la media aritmético geométrica le conducen a la siguiente maravillosa fórmula (ver [1], pág. 94-102) en la cual los valores iniciales son $a_0 = 1$ y $b_0 = 1/\sqrt{2}$:

$$\sum_{k=1}^{\infty} 2^k (a_k^2 - b_k^2) = \frac{1}{2} - \frac{2}{\pi} \left[M \left(1, \frac{\sqrt{2}}{2} \right) \right]^2. \tag{21}$$

Por razones no muy claras Gauss no utilizó su fórmula para calcular π y la fórmula se olvidó hasta que en 1970 fue redescubierta de forma independiente por Eugene Salamin y Richard Brent [20] que la expresan en forma apropiada para el cálculo iterativo:

$$\lim \frac{2a_n^2}{\frac{1}{2} - \sum_{k=1}^{n} 2^k (a_k^2 - b_k^2)} = \pi.$$

Se puede demostrar que la convergencia hacia π es cuadrática, es decir al pasar de n a $n+1$ se dobla el número de decimales correctos. Brent y Salamin programan el algoritmo en un ordenador, lo que les permite conseguir en 1976 un record al calcular π con 3 millones de cifras decimales, cuando el record hasta entonces era de 1 millón de cifras decimales y se había logrado con fórmulas de tipo Machin.

7 . SERIES PARA $1/\pi$ DE UN GENIO MATEMÁTICO INCREÍBLE: S. RAMANUJAN

Gran parte de la obra del genio matemático hindú Ramanujan se relaciona con la teoría de las funciones modulares, que tiene su origen en el estudio de las integrales elípticas de primera y segunda especie

$$K(k) = \int_0^{\pi/2} \frac{1}{\sqrt{1 - k^2 \mathrm{sen}^2 t}}\, dt, \qquad E(k) = \int_0^{\pi/2} \sqrt{1 - k^2 \mathrm{sen}^2 t}\, dt.$$

En 1914 estas investigaciones le llevan a la obtención de extraordinarias series para $1/\pi$ totalmente diferentes de todas las conocidas hasta entonces y que son del tipo siguiente

$$\sum_{n=0}^{\infty} B_n z^n (bn+a) = \frac{1}{\pi},$$

siendo $b, a > 0$ y $-1 < z < 1$ números algebraicos, y que quedan clasificadas en 4 familias de acuerdo a las siguientes posibilidades de B_n

$$\frac{(2n)!^3}{n!^6}, \qquad \frac{(4n)!}{n!^4}, \qquad \frac{(2n)!(3n)!}{n!^5}, \qquad \frac{(6n)!}{(3n)!n!^3}.$$

Demuestra parcialmente fórmulas que permiten obtener las series de la primera familia [19] y da 3 ejemplos, uno de los cuales es

$$\sum_{n=0}^{\infty} \frac{(2n)!^3}{n!^6} \frac{1}{2^{12n}} (42n+5) = \frac{16}{\pi}. \tag{22}$$

Continúa diciendo que existen teorías alternativas (pero casi no da ningún detalle de ellas) con las que pueden obtenerse series pertenecientes a las otras tres familias y da 14 ejemplos [19] entre los cuales están los siguientes

$$\sum_{n=0}^{\infty} \frac{(4n)!}{n!^4} \frac{1}{396^{4n}} (26390n+1103) = \frac{9801\sqrt{2}}{4\pi}, \tag{23}$$

$$\sum_{n=0}^{\infty} \frac{(2n)!(3n)!}{n!^5} \frac{1}{1458^n} (15n+2) = \frac{27}{4\pi}, \tag{24}$$

$$\sum_{n=0}^{\infty} \frac{(6n)!}{(3n)!n!^3} \frac{1}{54000^n} (11n+1) = \frac{5\sqrt{15}}{6\pi}. \tag{25}$$

La serie (23) es increíblemente rápida ya que con ella se obtienen unos 8 decimales de π por término. En 1985 B. Gosper programa la fórmula en un ordenador, antes de que se conociera la demostración, y consigue con ella un record de 17 millones de cifras de π.

Los hermanos Jonathan and Peter Borwein son los primeros en demostrar (1985) fórmulas que permiten obtener las dos primeras familias de fórmulas de Ramanujan [3] y consiguen demostrar la identidad (23). Los hermanos David y Gregory Chudnovsky contribuyen hallando la cuarta familia [9] y como caso particular consiguen la impresionante serie, no descubierta por Ramanujan:

$$\sum_{n=0}^{\infty} \frac{(6n)!}{(3n)!n!^3} \frac{(-1)^n}{640320^{3n}} (545140134n+13591409) = \frac{\sqrt{640320^3}}{12\pi} \tag{26}$$

que aporta 15 decimales de π por término y con la cual baten records de cálculo de π en los años 1989, 1991 y 1994, este último de 4044 millones de cifras de π. Finalmente, H.H. Chan, W. C. Liaw y V. Tan hallan las fórmulas necesarias que determinan la tercera familia [7] y a partir de ellas demuestran la fórmula, tampoco descubierta por Ramanujan:

$$\sum_{n=0}^{\infty} \frac{(3n)!(2n)!}{n!^5} \frac{(-1)^n}{300^{3n}} (14151n + 827) = \frac{1500\sqrt{3}}{\pi}. \tag{27}$$

La demostración de las fórmulas correspondientes a familias de series para $1/\pi$ consiste en hallar ciertas funciones $z(q)$, $b(q)$ y $a(q)$, relacionadas con las funciones modulares elípticas, tales que

$$\sum_{n=0}^{\infty} B_n \left\{ b(q)n + a(q) \right\} z(q)^n = \frac{1}{\pi}.$$

Recientemente el autor ha conjeturado (ver [14] y [15]) que si definimos las funciones

$$S(z) = \sum_{n=0}^{\infty} B_n z^n, \qquad W(z) = \sum_{n=0}^{\infty} B_n^{'} z^n,$$

donde $B_n^{'}$ se obtiene derivando B_n respecto de n como si se tratase de una variable continua, entonces $z(q)$ es la solución de la ecuación funcional

$$q = z \exp \frac{W(z)}{S(z)}$$

y que las funciones $b(q)$ y $a(q)$ vienen dadas por

$$b = \sqrt{N} \frac{q}{zS} \frac{dz}{dq}, \qquad a = \frac{1}{S} \left[\frac{1}{\pi} - \frac{q\sqrt{N}}{S} \frac{dS}{dq} \right].$$

Esta conjetura parece estar en perfecta armonía con los resultados ya establecidos y permitiría obtener los desarrollos en series de potencias de q para $z(q)$, $S(q)$ que a su vez determinan $b(q)$ y $a(q)$. Para N racional y $q = \pm e^{-\pi\sqrt{N}}$ los valores que se obtienen para $z(q)$, $b(q)$, $a(q)$ son números algebraicos que coinciden con los ya conocidos. Por ejemplo, para

$$S(q) = \sum_{n=0}^{\infty} \frac{(6n)!}{(3n)!n!^3} z(q)^n$$

se obtienen los desarrollos

$$z = q - 744q^2 + 356652q^3 - 140361152q^4 + 49336682190q^5 - 16114625669088q^6 + \cdots$$

$$S = 1 + 120q - 6120q^2 + 737760q^3 - 107249640q^4 + 17385063120q^5 - 3014720249760q^6 + \cdots$$

que son precisamente los desarrollos de las funciones

$$z(q) = J(q)^{-1}, \qquad S(q) = \theta_3^4(q)\sqrt{1 + \mu(-q)},$$

donde

$$J(q) = \frac{16}{\mu(q)}\left\{1 + 16\mu(q)\right\}^3, \quad \mu(q) = \frac{\theta_2^4(q)\theta_3^4(q)}{\theta_4^8(q)}$$

siendo θ_2, θ_3, θ_4, las funciones elípticas θ de Jacobi, definidas mediante

$$\theta_2(q) = \sum_{n=-\infty}^{n=\infty} q^{(n+1/2)^2}, \quad \theta_3(q) = \sum_{n=-\infty}^{n=\infty} q^{n^2}, \quad \theta_4(q) = \sum_{n=-\infty}^{n=\infty} (-1)^n q^{n^2}.$$

Estas fórmulas son equivalentes a las ya obtenidas por los hermanos David and Gregory Chudnovsky que, tomando $q = -e^{-\pi\sqrt{163}}$, obtienen la serie (26) y demuestran que es la serie de tipo Ramanujan con parámetro z racional cuya velocidad de convergencia es la más rápida posible.

8 . El algoritmo cuártico de los hermanos Borwein

Las funciones

$$s(q) = \frac{\theta_2(q)}{\theta_3(q)}, \quad t(q) = \frac{1}{\pi} \cdot \frac{1}{\theta_3^4(q)}\left[1 + (4q\ln q)\frac{\theta_4'(q)}{\theta_4(q)}\right].$$

verifican las siguientes magníficas propiedades difíciles de demostrar:

$$s(q^4) = \frac{1 - [1 - s^4(q)]^{1/4}}{1 + [1 - s^4(q)]^{1/4}},$$

$$t(q^4) = [1 + s(q^4)]^4 t(q) - \frac{4\ln q}{\pi} s(q^4)[1 + s(q^4) + s(q^4)^2],$$

que relacionan valores de las funciones en q con valores en q^4, lo que nos va a permitir obtener un algoritmo cuártico [4]. En efecto, llamando

$$q = e^{-2\pi\cdot 4^n}, \quad s_n = s(q), \quad t_n = t(q)$$

y teniendo en cuenta que los valores de s_0 y t_0 son conocidos, obtenemos el siguiente algoritmo cuártico

$$s_0 = \sqrt{2} - 1, \quad t_0 = 6 - 4\sqrt{2},$$

$$s_{n+1} = \frac{1-(1-s_n^4)^{1/4}}{1+(1-s_n^4)^{1/4}}, \tag{28}$$

$$t_{n+1} = t_n(1+s_{n+1})^4 - 2^{2n+3}s_{n+1}(1+s_{n+1}+s_{n+1}^2)$$

en el cual t_n tiende a $1/\pi$, cuadruplicándose en cada iteración el número de cifras exactas. Este asombroso algoritmo fue descubierto y demostrado por los hermanos Borwein en 1987.

Podría pensarse que con los algoritmos cuadráticos, cuárticos, etc. se tendrían que lograr siempre los records de cálculo de π. Sin embargo no es así pues estos algoritmos no auto-corrigen los errores y obligan a realizar todas las operaciones con la precisión final deseada. En cambio para las series se ha encontrado la forma de ir aumentando la precisión de cálculo sin afectar el resultado final, lo que les permite batir records aún teniendo la desventaja de ser lineales.

9 . EL CÁLCULO DESORDENADO DE LAS CIFRAS DE π

En 1995 Peter Borwein y Simon Plouffe se dieron cuenta de que la serie

$$\ln 2 = \frac{1}{1\cdot 2} + \frac{1}{2\cdot 2^2} + \frac{1}{3\cdot 2^3} + \frac{1}{4\cdot 2^4} + \cdots = \sum_{n=1}^{\infty} \frac{1}{n2^n}$$

podía ser utilizada para hallar el valor de una de las cifras en binario de $\ln 2$ sin necesidad de hallar las cifras que le preceden, mediante el siguiente ingenioso esquema en el cual $\{x\}$ significa la parte fraccionaria de x:

$$\left\{2^k \ln 2\right\} = \left\{\left\{\sum_{n=0}^{k} \frac{2^{k-n}}{n}\right\} + \sum_{n=k+1}^{\infty} \frac{1}{n2^{n-k}}\right\} = \left\{\left\{\sum_{n=0}^{k} \frac{2^{k-n} \bmod n}{n}\right\} + \sum_{n=k+1}^{\infty} \frac{1}{n2^{n-k}}\right\},$$

que permite obtener de forma eficiente las cifras binarias de $\ln 2$ a partir de la posición $k+1$. La principal observación para efectuar los cálculos implicados es que el numerador de la primera suma, es decir $2^{k-n} \bmod n$, se puede evaluar con extraordinaria rapidez. Por ejemplo, el cálculo 2^{65} no requiere 65 multiplicaciones sino sólo 7 si lo realizamos en la forma $((((((2^2)^2)^2)^2)^2)^2)\cdot 2$. Teniendo en cuenta la explicación anterior podemos evaluar las operaciones de aritmética modular aplicando un algoritmo muy eficiente que para el caso particular $2^{65} \bmod 9$ consiste en hacer una reducción módulo 9 después de cada multiplicación de las 7 que se requieren en lugar de hacer todas las multiplicaciones seguidas.

En 1989 los hermanos Borwein afirmaron que obtener una cifra binaria de π no podía ser más sencillo que calcular todas las cifras que le preceden. Sin embargo, después de este descubrimiento para $\ln 2$, ya no lo tenían tan claro y el equipo formado por D. Bailey, P. Borwein y S. Plouffe se puso a buscar, con la ayuda del algoritmo PSLQ, una serie para π que permitiera el cálculo aislado de sus cifras en base 2 (ver [6], capítulo 8).

El algoritmo PSLQ es un algoritmo numérico que opera con una precisión prefijada asociando a una entrada de n elementos $x_1, x_2, \cdots, x_n$ una salida de n números enteros $a_1, a_2, \cdots, a_n$, no todos nulos, de forma que, para la precisión prefijada, se verifica

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = 0.$$

El éxito llegó al considerar los valores

$$x_j = \sum_{n=0}^{\infty} \frac{1}{(8n+j)16^n}, \qquad j = 1, 2, \cdots, 7$$

y buscar con PSLQ una relación lineal con coeficientes enteros entre las entradas π, x_1, x_2, x_3, x_4, x_5, x_6, x_7. La salida obtenida fue la sucesión de enteros: -1, 4, 0, 0, -2, -1, -1, 0. Es decir habían encontrado una serie [2] que llamaron de tipo BBP (iniciales de los nombres del equipo)

$$\pi = \sum_{n=0}^{\infty} \frac{1}{16^n}\left(\frac{4}{8n+1} - \frac{2}{8n+4} - \frac{1}{8n+5} - \frac{1}{8n+6}\right), \tag{29}$$

que permitía el cálculo desordenado de las cifras de π en hexadecimal y por lo tanto también en binario. La demostración de la fórmula se consiguió poco después a partir del sencillo resultado

$$\int_0^{1/\sqrt{2}} \frac{x^{k-1}}{1-x^8}dx = \int_0^{1/\sqrt{2}} \sum_{n=0}^{\infty} x^{k-1+8n}dx = \frac{1}{2^{k/2}} \sum_{n=0}^{\infty} \frac{1}{16^n(8n+k)}$$

que permite escribir

$$\frac{1}{16^n}\left(\frac{4}{8n+1} - \frac{2}{8n+4} - \frac{1}{8n+5} - \frac{1}{8n+6}\right) = \int_0^{1/\sqrt{2}} \frac{4\sqrt{2} - 8x^3 - 4\sqrt{2}x^4 - 8x^5}{1-x^8}dx,$$

integral que se puede evaluar fácilmente con el cambio de variable $y = x\sqrt{2}$ de la manera siguiente

$$16\int_0^1 \frac{4-2y^3-y^4-y^5}{16-y^8}dy = \int_0^1 \frac{16y-16}{(y^2-2)(y^2-2y+2)}dy = \pi.$$

Nada más anunciar la sorprendente fórmula (29) otros investigadores se lanzaron también a la caza de otras bellas fórmulas de tipo BBP. La más sencilla para π fue encontrada y demostrada por Viktor Adamchik y Stan Wagon (ver [23] pág. 66 y 67)

$$\pi = \sum_{n=0}^{\infty} \frac{(-1)^n}{4^n} \left(\frac{2}{4n+1} + \frac{2}{4n+2} + \frac{1}{4n+3} \right), \tag{30}$$

que se puede demostrar de forma similar a la anterior partiendo de la identidad

$$\int_0^{1/\sqrt{2}} \frac{x^{k-1}}{1+x^4} dx = \int_0^{1/\sqrt{2}} \sum_{n=0}^{\infty} (-1)^n x^{k-1+4n} dx = \frac{1}{2^{k/2}} \sum_{n=0}^{\infty} \frac{(-1)^n}{4^n(4n+k)}.$$

En 1997 Fabrice Bellard consigue un record de cálculo con su fórmula

$$\pi = 4 \sum_{n=0}^{\infty} \frac{(-1)^n}{4^n(2n+1)} - \frac{1}{64} \sum_{n=0}^{\infty} \frac{(-1)^n}{1024^n} \left(\frac{32}{4n+1} + \frac{8}{4n+2} + \frac{1}{4n+3} \right). \tag{31}$$

El último record es de C. Percival que ha obtenido que la cifra binaria de π que ocupa la posición 10000 billones es 0.

También se conocen fórmulas BBP para otros valores relacionados con π, por ejemplo para $\pi\sqrt{3}$

$$\pi\sqrt{3} = \frac{9}{32} \sum_{n=0}^{\infty} \frac{1}{64^n} \left(\frac{16}{6n+1} + \frac{8}{6n+2} - \frac{2}{6n+4} - \frac{1}{6n+5} \right) \tag{32}$$

y para π^2

$$\pi^2 = \frac{9}{8} \sum_{n=0}^{\infty} \frac{1}{64^n} \left(\frac{16}{(6n+1)^2} - \frac{24}{(6n+2)^2} - \frac{8}{(6n+3)^2} - \frac{6}{(6n+4)^2} + \frac{1}{(6n+5)^2} \right). \tag{33}$$

10 . BONITOS DESARROLLOS INFINITOS PARA π CON PRODUCTOS Y RADICALES

Osler, en 1999, demuestra la siguiente fórmula [17]:

$$\frac{2}{\pi} = \prod_{n=1}^{p} \sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2} + \overset{n}{\cdots} + \frac{1}{2}\sqrt{\frac{1}{2}}}}} \cdot \prod_{n=1}^{\infty} \frac{2^{p+1}n-1}{2^{p+1}n} \cdot \frac{2^{p+1}n+1}{2^{p+1}n}, \tag{34}$$

donde la n en el interior de las raíces denota el número de sumandos. Esta fórmula engloba la fórmula de Viète (2) ($p = \infty$); el producto de Wallis (3) ($p = 0$) y también fórmulas nuevas. Por ejemplo tomando $p = 2$ se tiene

$$\frac{2}{\pi} = \sqrt{\frac{1}{2}}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}}} \cdot \frac{7\cdot 9}{8\cdot 8} \cdot \frac{15\cdot 17}{16\cdot 16} \cdot \frac{23\cdot 25}{24\cdot 24} \cdot \frac{31\cdot 33}{32\cdot 32} \cdots . \tag{35}$$

Para llegar a su expresión para π, Osler, partiendo del seno de x, aplica p veces, de forma reiterada, la fórmula del seno del ángulo doble:

$$\operatorname{sen} x = 2\cos\frac{x}{2}\operatorname{sen}\frac{x}{2} = 2^2\cos\frac{x}{2}\cos\frac{x}{2^2}\operatorname{sen}\frac{x}{2^2} = 2^p\cos\frac{x}{2}\cos\frac{x}{2^2}\cdots\cos\frac{x}{2^p}\operatorname{sen}\frac{x}{2^p}$$

y sustituye $t = \frac{x}{2^p}$ en la fórmula de Euler

$$\frac{\operatorname{sen} t}{t} = \prod_{n=1}^{\infty}\left(1-\frac{t^2}{\pi^2 n^2}\right) = \prod_{n=1}^{\infty}\left(\frac{n\pi - t}{n\pi}\cdot\frac{n\pi + t}{n\pi}\right)$$

lo que le permite combinar las dos fórmulas anteriores para llegar a

$$\frac{\operatorname{sen} x}{x} = \cos\frac{x}{2}\cos\frac{x}{2^2}\cdots\cos\frac{x}{2^p}\cdot\prod_{n=1}^{\infty}\left(\frac{2^p n\pi - x}{2^p n\pi}\cdot\frac{2^p n\pi + x}{2^p n\pi}\right).$$

Para terminar utiliza las siguientes relaciones de trigonometría elemental

$$\cos\frac{x}{2} = \sqrt{\frac{1}{2}+\frac{1}{2}\cos x}, \qquad \cos\frac{x}{2^2} = \sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}+\frac{1}{2}\cos x}}, \qquad \cdots$$

y sustituye $x = \frac{\pi}{2}$.

J. Sondow obtiene en 2002 otra bonita expresión con productos y radicales [21]:

$$\frac{\pi}{2} = \prod_{n=0}^{\infty}\left[1^{(-1)^1\binom{n}{0}}\cdot 2^{(-1)^2\binom{n}{1}}\cdots(n+1)^{(-1)^{n+1}\binom{n}{n}}\right]^{\frac{1}{2^n}} \tag{36}$$

Para ello comienza tomando logaritmos en el producto de Wallis:

$$\ln\frac{\pi}{2} = \sum_{k=1}^{\infty}(-1)^{k-1}\ln\frac{k+1}{k}.$$

Teniendo en cuenta que $\sum_{n\geq k}\frac{\binom{n}{k}}{2^{n+1}} = 1$, podemos continuar de la siguiente manera

$$\ln\frac{\pi}{2} = \sum_{k\geq 1}(-1)^{k-1}\ln\frac{k+1}{k}\sum_{n\geq k}\frac{\binom{n}{k}}{2^{n+1}} = \sum_{n=0}^{\infty}\frac{1}{2^{n+1}}\sum_{k=0}^{n}(-1)^k\binom{n}{k}\ln\frac{k+2}{k+1}.$$

En la última expresión reemplaza n con $n-1$, escribe el logaritmo como $\ln(k+2)-\ln(k+1)$ y la suma sobre k como la diferencia de dos sumas, en la primera de las cuales reemplaza k con $k-1$. Aplicando la relación binomial $\binom{n-1}{k-1}+\binom{n-1}{k}=\binom{n}{k}$ llega al resultado mediante exponenciación.

11 . RESULTADOS MUY RECIENTES: SERIES PARA $1/\pi$, $1/\pi^2$ Y $1/\pi^3$

T. Sato, en una conferencia en Japón en 2002, presenta la serie

$$\sum_{n=0}^{\infty} u_n \left(\frac{\sqrt{5}-1}{2}\right)^{12n} (20n+10-3\sqrt{5}) = \frac{20\sqrt{3}+9\sqrt{15}}{6\pi}, \tag{37}$$

siendo u_n los números de Apéry, definidos mediante

$$u_n = \sum_{k=0}^{n} \binom{n}{k}^2 \binom{n+k}{k}^2.$$

Poco después se encuentran series de este estilo con otros tipos de números [8], [22], por ejemplo con los números de Domb; H. H. Chan y Y. Yang desarrollan independientemente una teoría explicativa de este tipo de fórmulas que incluye también como caso particular a las de tipo Ramanujan y les dan el nombre de fórmulas de tipo Ramanujan-Sato.

Una explicación radicalmente diferente para algunas fórmulas de tipo Ramanujan, que no hace uso de la teoría de funciones modulares, se basa en el método WZ de Herb Wilf y Doron Zeilberger [18] que permite demostrar automáticamente identidades de la forma

$$\sum_{n=0}^{\infty} G(n,k) = \text{Constante}$$

en el caso en que la función $G(n,k)$ sea hipergeométrica en sus dos variables, es decir si los cocientes $G(n+1,k)/G(n,k)$ y $G(n,k+1)/G(n,k)$ son funciones racionales. El programa EKHAD escrito por D. Zeilberger realiza el trabajo duro encontrando a la función $G(n,k)$ una compañera $F(n,k)$ tal que $F(0,k)=0$ y con la cual forma un par de Wilf y Zeilberger (WZ), es decir, un par que se caracteriza por la propiedad

$$F(n+1,k)-F(n,k)=G(n,k+1)-G(n,k).$$

Además Zeilberger ha demostrado que si definimos $H(n,k)=F(n+1,n+k)+G(n,n+k)$, entonces obtenemos una identidad asociada a la de partida

$$\sum_{n=0}^{\infty} H(n,k) = \sum_{n=0}^{\infty} G(n,k) = \text{Constante}.$$

Es importante señalar que el método WZ es un método simbólico y no numérico como el algoritmo PSLQ (ver sección 9), por lo que las demostraciones que se obtienen son completamente rigurosas. El precio a pagar es la dificultad de encontrar funciones interesantes que sean una componente de un par.

Un tipo de identidades descubiertas por el autor [13], entre las que se encuentran

$$\sum_{n=0}^{\infty} \frac{1}{2^{8n}2^{4k}} \frac{\binom{2n}{n}^3 \binom{2k}{k}^2}{\binom{n+k}{k}^2} (6n+4k+1) = \frac{4}{\pi}, \tag{38}$$

$$\sum_{n=0}^{\infty} \frac{(-1)^n}{2^{12n}2^{8k}} \frac{\binom{2n}{n}^5 \binom{2k}{k}^4}{\binom{n+k}{k}^4} (20n^2+8n+1+24kn+8k^2+4k) = \frac{8}{\pi^2} \tag{39}$$

son aptas para la demostración automática con EKHAD. Para ver en que consiste el método WZ, explicamos a continuación con detalle la demostración de la identidad (38) basada en dicho método: A la función

$$G(n,k) = \frac{1}{2^{8n}2^{4k}} \frac{\binom{2n}{n}^3 \binom{2k}{k}^2}{\binom{n+k}{k}^2} (6n+4k+1),$$

EKHAD le encuentra una compañera

$$F(n,k) = \frac{1}{2^{8n}2^{4k}} \frac{\binom{2n}{n}^3 \binom{2k}{k}^2}{\binom{n+k}{k}^2} 8n$$

con la que forma un par WZ, es decir que verifica

$$F(n+1,k) - F(n,k) = G(n,k+1) - G(n,k),$$

propiedad qué tiene la siguiente consecuencia

$$\sum_{n=0}^{\infty} [G(n,k+1) - G(n,k)] = \sum_{n=0}^{\infty} [F(n+1,k) - F(n,k)] = -F(0,k) = 0.$$

Por lo tanto

$$\sum_{n=0}^{\infty} G(n,0) = \sum_{n=0}^{\infty} G(n,1) = \sum_{n=0}^{\infty} G(n,2) = \cdots .$$

Pero

$$\mathop{\mathrm{lím}}_{k\to\infty} \sum_{n=0}^{\infty} G(n,k) = \mathop{\mathrm{lím}}_{k\to\infty} G(0,k) + \mathop{\mathrm{lím}}_{k\to\infty} \sum_{n=1}^{\infty} G(n,k) = \mathop{\mathrm{lím}}_{k\to\infty} \frac{\binom{2k}{k}^2}{2^{4k}} (4k+1) = \frac{4}{\pi}.$$

Así que

$$\sum_{n=0}^{\infty} G(n,0) = \sum_{n=0}^{\infty} G(n,1) = \sum_{n=0}^{\infty} G(n,2) = \cdots = \frac{4}{\pi}.$$

Más aún, aplicando un teorema de Carlson se deduce que

$$\sum_{n=0}^{\infty} G(n,k) = \frac{4}{\pi}$$

incluso para valores no enteros de k. Sustituyendo $k=0$ en la identidad (38) y en su asociada, obtenemos las series de Ramanujan

$$\sum_{n=0}^{\infty} \frac{(2n)!^3}{n!^6} \frac{1}{2^{8n}} (6n+1) = \frac{1}{4} \sum_{n=0}^{\infty} \frac{(2n)!^3}{n!^6} \frac{1}{2^{12n}} (42n+5) = \frac{4}{\pi}. \tag{40}$$

Para la función $G(n,k)$ en la serie (39), EKHAD encuentra su compañera

$$F(n,k) = \frac{(-1)^n}{2^{12n} 2^{8k}} \frac{\binom{2n}{n}^5 \binom{2k}{k}^4}{\binom{n+k}{k}^4} \cdot 8n(2n+4k+1),$$

y podemos decir que la demostración de (39) está encapsulada en dicho par $F(n,k)$, $G(n,k)$. Sustituyendo $k=0$ en la identidad (39) y en su asociada, obtenemos las nuevas fórmulas [11], [13]:

$$\sum_{n=0}^{\infty} \frac{(2n)!^5}{n!^{10}} \frac{(-1)^n}{2^{12n}} (20n^2+8n+1) = \frac{1}{16} \sum_{n=0}^{\infty} \frac{(2n)!^5}{n!^{10}} \frac{(-1)^n}{2^{20n}} (820n^2+180n+13) = \frac{8}{\pi^2}. \tag{41}$$

Con el algoritmo PSLQ el autor consigue encontrar más series de este estilo para $1/\pi^2$ [12], por ejemplo

$$\sum_{n=0}^{\infty} \frac{(6n)!}{n!^6} \frac{(-1)^n}{2880^{3n}} (5418n^2 + 693n + 29) = \frac{128\sqrt{5}}{\pi^2}, \tag{42}$$

para las que todavía no se conoce la demostración. En [14] se encuentran de nuevo las mismas series a partir de una conjetura.

B. Gourevitch, haciendo uso del algoritmo PSLQ, encuentra una serie para $1/\pi^3$ [12]

$$\sum_{n=0}^{\infty} \frac{(2n)!^7}{n!^{14}} \frac{1}{2^{20n}} (168n^3 + 76n^2 + 14n + 1) = \frac{32}{\pi^3} \tag{43}$$

que tampoco ha podido ser demostrada todavía.

W. Zudilin, mediante transformaciones cuadráticas de las identidades (41) demuestra las nuevas identidades [24]

$$\sum_{n=0}^{\infty} w_n \frac{(4n)!}{n!^2(2n)!}(18n^2-10n-3)\frac{1}{(2^8 5^2)^n} = \frac{10\sqrt{5}}{\pi^2}, \tag{44}$$

$$\sum_{n=0}^{\infty} w_n \frac{(4n)!}{n!^2(2n)!}(1046529n^2+227104n+16032)\frac{1}{(5^4 41^2)^n} = \frac{5^4 41\sqrt{41}}{\pi^2}, \tag{45}$$

siendo w_n los números definidos por

$$w_n = \sum_{k=0}^{n} \binom{2k}{k}^3 \binom{2n-2k}{n-k} 2^{4(n-k)}$$

y con estas fórmulas damos por concluido nuestro recorrido histórico.

AGRADECIMIENTO al profesor D. Javier Cilleruelo por sus valiosas sugerencias y comentarios.

REFERENCIAS

[1] J. ARNDT, C. HAENEL, *π unleashed*, Springer-Verlag, 2001.

[2] D. BAILEY, P. BORWEIN, S. PLOUFFE, On the rapid computation of various polylogarithmic constants. *Mathematics of Computation* **66** (1997) 903–913.

[3] J. BORWEIN, P. BORWEIN, *Pi and the AGM*, Wiley Interscience, (1987).

[4] J. BORWEIN, P. BORWEIN, D. BAILEY, Ramanujan, Modular Equations, and Approximations to Pi or How to compute One Billion Digits of Pi. *American Mathematical Monthly* **96** (1989).

[5] J. BORWEIN, D. BROADHURST, J. KAMNITZER, Central Binomal Sums, Multiple Clausen Values and Zeta Values. *Experimental Mathematics* **10** (2001).

[6] J. P. DELAHAYE, *Le fascinant nombre π*, Bibliothèque pour la science, (1997).

[7] HENG HUAT CHAN, WEN-CHIN LIAW, VICTOR TAN, Ramanujan's class invariant λ_n and a new class of series for $1/\pi$. *Journal of the London Mathematical Society* **64** (2001) 93–106.

[8] HENG HUAT CHAN, SONG HENG CHAN, ZHIGUO LIU, Domb's numbers and Ramanujan-Sato type series for $1/\pi$. *Advances in Mathematics*, **186** (2004) 396–410.

[9] D. V. CHUDNOVSKY, G.V. CHUDNOVSKY, Approximations and complex multiplication according to Ramanujan. En *Ramanujan revisted*. Academic Press, 1988, pp. 375–472.

[10] P. Flajolet, B. Salvy, Euler sums and contour integral representations. *Experimental Mathematics* **7** (1998) 15–35.

[11] J. Guillera, Some binomial series obtained by the WZ-method. *Advances in Applied Mathematics* **29** (2002) 599–603.

[12] J. Guillera, About a new kind of Ramanujan type series. *Experimental Mathematics*, **12** (2003) 4, 507–510.

[13] J. Guillera, Generators of Some Ramanujan Formulas. *The Ramanujan Journal* **11** (2006), 41–48.

[14] J. Guillera, A new method to obtain series for $1/\pi$ and $1/\pi^2$. *Experimental Mathematics* **15** (2006) 83–89.

[15] J. Guillera, A class of conjectured series representations for $1/\pi$. *Experimental Mathematics* **15** (2006) 409–414.

[16] L. J. Lange, An elegant new continued fraction for π. *American Mathematical Monthly* **106** (1999) 456–458.

[17] T. J. Osler, The united Vieta's and Wallis's products for π. *American Mathematical Monthly* **106** 774–776. (1999).

[18] M. Petkovšek, H. S. Wilf, D. Zeilberger, *A=B*, A.K. Peters Ltd., Appendix A, (1996).

[19] S. Ramanujan, Modular equations and approximations to π. *Quarterly Journal of Mathematics* **45** (1914) 350–372.

[20] E. Salamin, Computation of π Using Arithmetic-Geometric Mean. *Mathematics of Computation* **30** (1976) 565–570.

[21] J. Sondow, A faster product for pi and a new integral for ln pi/2. *American Mathematical Monthly* **112** (2003) 729–734.

[22] Y. Yang, On differential equations satisfied by modular forms. *Mathematische Zeitschrift* **246** (2004) 1–19.

[23] A. V. Zhúkov, *El omnipresente número π*, Serie de divulgación científica *Matemática* 11. Traducido de la edición rusa: Editorial URSS, Moscú, (2004).

[24] W. Zudilin, *Quadratic transformations and Guillera's formulae for* $1/\pi^2$. *Matematicheskie Zametki* **81** (2007) 335–340 (versión en ruso). También aparecerá en *Mathematical Notes* **87** (2001).

Jesús Guillera Goyanes
Zaragoza
Correo electrónico:`jguillera@able.es`

geometria-algebrica-de-carlos-ivorra.pdf

Pág. 301 (311) Capítulo VIII

Integrales abelianos

8.1 Homología y cohomología

(Pág. 304 (314))

- Antonio F. Costa – ejercicios de geometría diferencial de curvas y superficies... (2005).pdf (804 pag)
- Important: Pi
- 3. Hist. Form. y Alg. para PI.pdf

obs.: El Duelo de los Números

Javier Cilleruelo Mateo

(20 pág)

Source:

La Gaceta de la RSME, Vol. 10.1 (2007), Pags. 159-178

7 +,-.+ + +/ ! + ')) " *% * 3#
/4 A #!# 2 !
(' " * '

24.3 Aplicación de la ley de Gauss a varias distribuciones de carga

Como se mencionó, la ley de Gauss es útil para determinar campos eléctricos cuando la distribución de carga está caracterizada por un alto grado de simetría. Los ejemplos siguientes muestran cómo escoger la superficie gaussiana que permita simplificar la integral de superficie dad por en la ecuación 24.6 y determinar el campo eléctrico. Al seleccionar la superficie, siempre debe aprovechar la simetría de la distribución de la carga de manera que retire a E de la integral y la resuelva. El objetivo en este tipo de cálculo es encontrar una superficie para la que cada parte de la superficie satisfaga una o más de las condiciones siguientes:

1. Demostrar por simetría que el valor del campo eléctrico es constante sobre la porción de superficie.
2. Que el producto punto de la ecuación 24.6 se expresa como un producto algebraico simple $\vec{\mathbf{E}}$, ya que $d\vec{\mathbf{A}}$ son paralelos entre sí.
3. Que el producto punto de la ecuación 24.6 sea cero, ya que $\vec{\mathbf{E}}$ y $d\vec{\mathbf{A}}$ son perpendiculares entre sí.
4. Que el campo eléctrico es igual a cero sobre la porción de superficie.

Diferentes porciones de la superficie gaussiana puedan satisfacer varias condiciones en tanto que cada porción satisfaga al menos una condición. Estas cuatro condiciones son utilizadas en los ejemplos del resto de este capítulo y serán identificadas por su número. Si la distribución de carga no tiene simetría suficiente para que una superficie gaussiana que satisfaga estas condiciones se pueda encontrar, la ley de Gauss sigue siendo cierta, pero no es útil para determinar el campo eléctrico para esta distribución de carga.

Prevención de riesgos ocultos 24.2

Las superficies gaussianas no son reales La superficie gaussiana es una superficie imaginaria que se elige para satisfacer las condiciones mencionadas en este caso. No tiene que coincidir con una superficie física en una situación determinada.

Ejemplo 24.3 Distribución de carga para una simetría esférica

Una esfera sólida aislante con radio a tiene una densidad de carga volumétrica uniforme ρ y una carga positiva total Q (figura 24.10).

(A) Calcule la magnitud del campo eléctrico en un punto fuera de la esfera.

SOLUCIÓN

Conceptualizar Observe cómo difiere este problema de la explicación anterior de la ley de Gauss. El campo eléctrico debido a cargas puntuales se explicó en la sección 24.2. Ahora se considera el campo eléctrico debido a una distribución de carga. En el capítulo 23 encontró el campo para varias distribuciones de carga, al integrar sobre la distribución. Este ejemplo muestra una diferencia con nuestra discusión en el capítulo 23. En este capítulo debe encontrar el campo eléctrico mediante la ley de Gauss.

Categorizar Puesto que la distribución de la carga es uniforme en toda la esfera, la distribución de carga tiene simetría esférica y se puede aplicar la ley de Gauss para hallar el campo eléctrico.

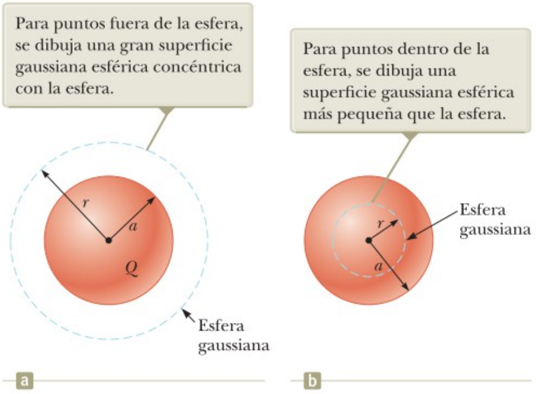
Para puntos fuera de la esfera, se dibuja una gran superficie gaussiana esférica concéntrica con la esfera.
r
a
Q
Esfera gaussiana
a
Para puntos dentro de la esfera, se dibuja una superficie gaussiana esférica más pequeña que la esfera.
r
a
Esfera gaussiana
b

Figura 24.10 (Ejemplo 24.3) Esfera aislante con carga uniforme de radio a y carga total Q. En diagramas como éste, la línea discontinua representa la intersección de la superficie gaussiana con el plano de la página.

Analizar Para reflejar la simetría esférica, elija una superficie gaussiana esférica de radio r, concéntrica con la esfera, como se muestra en la figura 24.10a. Para esta elección, la condición (2) se satisface en cualquier parte sobre la superficie y $\vec{\mathbf{E}} \cdot d\vec{\mathbf{A}} = E\,dA$.

continúa

24.3 continuación

Sustituya $\vec{\mathbf{E}} \cdot d\vec{\mathbf{A}}$ en la ley de Gauss con $E\,dA$:

$$\Phi_E = \oint \vec{\mathbf{E}} \cdot d\vec{\mathbf{A}} = \oint E\,dA = \frac{Q}{\epsilon_0}$$

Por simetría, E es constante en todas partes sobre la superficie, lo que satisface la condición (1), de modo que se puede retirar E de la integral:

$$\oint E\,dA = E\oint dA = E(4\pi r^2) = \frac{Q}{\epsilon_0}$$

Al resolver para E:

$$(1)\quad E = \frac{Q}{4\pi\epsilon_0 r^2} = \boxed{k_e\frac{Q}{r^2}\quad (\text{para } r > a)}$$

Finalizar Este campo es idéntico al de una carga puntual. Por lo tanto, **el campo eléctrico debido a una esfera con carga uniforme en la región externa a la esfera es *equivalente* a una carga puntual ubicada en el centro de la esfera.**

(B) Encuentre la magnitud del campo eléctrico en un punto dentro de la esfera.

SOLUCIÓN

Analizar En este caso, elija una superficie gaussiana esférica que tenga radio $r < a$, concéntrica con la esfera aislante (figura 24.10b). Sea V' el volumen de esta esfera más pequeña. Para aplicar la ley de Gauss en esta situación, considere que la carga q_{int} dentro de la superficie gaussiana de volumen V' es menor que Q.

Calcule q_{int} usando $q_{\text{int}} = \rho V'$:

$$q_{\text{int}} = \rho V' = \rho\left(\tfrac{4}{3}\pi r^3\right)$$

Observe que las condiciones (1) y (2) se satisfacen en todas partes sobre la superficie gaussiana en la figura 24.10b. Aplique la ley de Gauss en la región $r < a$:

$$\oint E\,dA = E\oint dA = E(4\pi r^2) = \frac{q_{\text{in}}}{\epsilon_0}$$

Resuelva para E y sustituya para q_{int}:

$$E = \frac{q_{\text{int}}}{4\pi\epsilon_0 r^2} = \frac{\rho\left(\frac{4}{3}\pi r^3\right)}{4\pi\epsilon_0 r^2} = \frac{\rho}{3\epsilon_0}r$$

Sustituya $\rho = Q/\frac{4}{3}\pi a^3$ y $\epsilon_0 = 1/4\pi k_e$:

$$(2)\quad E = \frac{Q/\frac{4}{3}\pi a^3}{3(1/4\pi k_e)}r = \boxed{k_e\frac{Q}{a^3}r\quad (\text{para } r < a)}$$

Finalizar Este resultado para E difiere del obtenido en el inciso (A). Muestra que $E \to 0$ cuando $r \to 0$. Por lo tanto, el resultado elimina el problema que existiría en $r = 0$ si E variara como $1/r^2$ dentro de la esfera como lo hace afuera de la esfera. Es decir: si $E \propto 1/r^2$ para $r < a$, el campo sería infinito en $r = 0$, lo que es físicamente imposible.

¿QUÉ PASARÍA SI? Suponga que la posición radial $r = a$ se alcanza desde adentro de la esfera y desde el exterior. ¿Se obtiene el mismo valor del campo eléctrico desde ambas direcciones?

Respuesta La ecuación (1) muestra que el campo eléctrico se aproxima a un valor desde el exterior dado por

$$E = \lim_{r \to a}\left(k_e\frac{Q}{r^2}\right) = k_e\frac{Q}{a^2}$$

Desde el interior, la ecuación (2) da

$$E = \lim_{r \to a}\left(k_e\frac{Q}{a^3}r\right) = k_e\frac{Q}{a^3}a = k_e\frac{Q}{a^2}$$

Por lo tanto, el valor del campo es el mismo, aunque se llegue a la superficie desde ambas direcciones. En la figura 24.11 se muestra una gráfica de E en función de r. Observe que la magnitud del campo es continua.

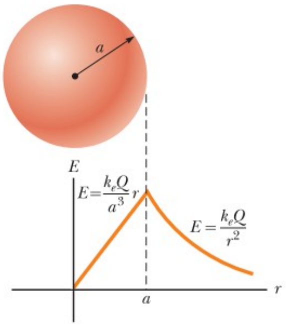

Figura 24.11 (Ejemplo 24.3) Gráfica de E en función de r para una esfera aislante con carga uniforme. El campo eléctrico dentro de la esfera ($r < a$) varía linealmente con r. El campo afuera de la esfera ($r > a$) es el mismo que el de una carga puntual Q ubicada en $r = 0$.

Ejemplo 24.4 Distribución de carga con simetría cilíndrica

Encuentre el campo eléctrico a una distancia r de una línea de carga positiva de longitud infinita y carga constante por unidad de longitud λ (figura 24.12a).

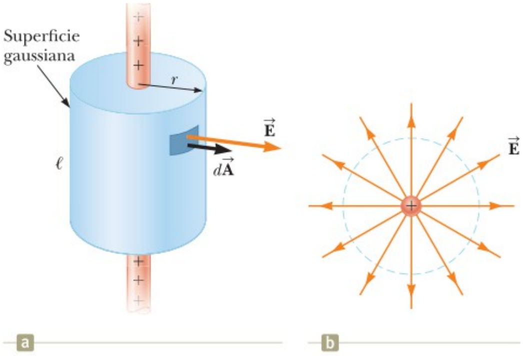

Figura 24.12 (Ejemplo 24.4) (a) Línea infinita de carga rodeada por una superficie gaussiana cilíndrica concéntrica con la línea. (b) Una vista lateral muestra que el campo eléctrico en la superficie cilíndrica es constante en magnitud y perpendicular a la superficie.

SOLUCIÓN

Conceptualizar La línea de carga es *infinitamente* larga. Por consiguiente, el campo es el mismo en todos los puntos equidistantes de la línea, sin importar la posición vertical del punto en la figura 24.12a. Esperamos que el campo sea más débil a medida que nos alejamos de la línea de carga.

Categorizar Ya que la carga está distribuida uniformemente a lo largo de la línea, la distribución de carga tiene simetría cilíndrica y se puede aplicar la ley de Gauss para encontrar el campo eléctrico.

Analizar La simetría de la distribución de carga requiere que $\vec{\mathbf{E}}$ sea perpendicular a la línea de carga y dirigida hacia fuera, como se muestra en las figuras 24.12b. Para reflejar la simetría de la distribución de carga, elija una superficie gaussiana cilíndrica de radio r y longitud ℓ que sea coaxial con la línea de carga. Para la parte curva de esta superficie, $\vec{\mathbf{E}}$ es constante en magnitud y perpendicular a la superficie en cada punto, lo que satisface las condiciones (1) y (2). Además, el flujo a través de los extremos del cilindro gaussiano es cero porque $\vec{\mathbf{E}}$ es paralelo a estas superficies. Ésta es la primera aplicación de la condición (3).

Debe tomar la integral de superficie en la ley de Gauss sobre toda la superficie gaussiana. Sin embargo, ya que $\vec{\mathbf{E}} \cdot d\vec{\mathbf{A}}$ es cero para los extremos planos del cilindro, restrinja la atención sólo a la superficie curva del cilindro.

Aplique la ley de Gauss y las condiciones (1) y (2) para la superficie curva, y note que la carga total dentro de la superficie gaussiana es $\lambda\ell$:

$$\Phi_E = \oint \vec{\mathbf{E}} \cdot d\vec{\mathbf{A}} = E \oint dA = EA = \frac{q_{\text{int}}}{\epsilon_0} = \frac{\lambda \ell}{\epsilon_0}$$

Sustituya el área $A = 2\pi r\ell$ de la superficie curva:

$$E(2\pi r \ell) = \frac{\lambda \ell}{\epsilon_0}$$

Resuelva para la magnitud del campo eléctrico:

$$E = \frac{\lambda}{2\pi\epsilon_0 r} = \boxed{2k_e \frac{\lambda}{r}} \qquad \textbf{(24.7)}$$

Finalizar Este resultado muestra que el campo eléctrico debido a una distribución de carga con simetría cilíndrica varía como $1/r$, mientras que el campo externo a una distribución de carga con simetría esférica varía como $1/r^2$. La ecuación 24.7 también se puede deducir mediante integración directa sobre la distribución de carga. (Véase el problema 44 del capítulo 23.)

¿QUÉ PASARÍA SI? ¿Y si el segmento de línea en este ejemplo no fuese infinitamente largo?

Respuesta Si la línea de carga de este ejemplo tuviese longitud finita, el campo eléctrico no se conocería por la ecuación 24.7. Una línea de carga finita no tiene suficiente simetría para utilizar la ley de Gauss, porque la magnitud del campo eléctrico ya no es constante sobre la superficie del cilindro gaussiano: el campo cerca de los extremos de la línea sería diferente de los que se encuentren lejos de los extremos. Por lo tanto, la condición (1) no se satisfaría en esta situación. Además, $\vec{\mathbf{E}}$ no es perpendicular a la superficie cilíndrica en todos los puntos: los vectores de campo cerca de los extremos tendrían una componente paralela a la línea. Por consiguiente, la condición (2) no se satisfaría. Para puntos cerca de una línea de carga finita y lejos de los extremos, la ecuación 24.7 da una buena aproximación del valor del campo.

Se le deja demostrar (véase el problema 33) que el campo eléctrico dentro de una barra de radio finito y longitud infinita, con carga uniforme, es proporcional a r.

Ejemplo 24.5 Un plano de carga

Encuentre el campo eléctrico debido a un plano infinito de carga positiva con densidad de carga superficial uniforme σ.

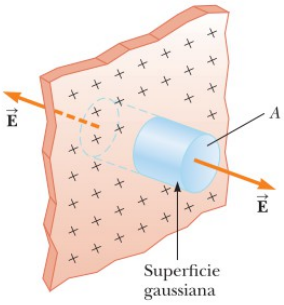

Figura 24.13 (Ejemplo 24.5) Superficie gaussiana cilíndrica que penetra un plano infinito de carga. El flujo es EA a través de cada extremo de la superficie gaussiana y cero a través de su superficie curva.

SOLUCIÓN

Conceptualizar Observe que el plano de carga es *infinitamente* largo. Por lo tanto, el campo eléctrico debe ser el mismo en todos los puntos equidistantes al plano. ¿Cómo esperaría que el campo eléctrico dependa de la distancia al plano?

Categorizar Ya que la carga se distribuye de manera uniforme sobre el plano, la distribución de carga es simétrica; en consecuencia, se puede usar la ley de Gauss para encontrar el campo eléctrico.

Analizar Por simetría, $\vec{\mathbf{E}}$ debe ser perpendicular al plano en todos los puntos. La dirección de $\vec{\mathbf{E}}$ se aleja de las cargas positivas, lo que indica que la dirección de $\vec{\mathbf{E}}$ en un lado del plano debe ser opuesta a su dirección sobre el otro lado, como se muestra en la figura 24.13. Una superficie gaussiana que refleja la simetría es un pequeño cilindro cuyo eje es perpendicular al plano y cada uno de sus extremos tiene un área A y son equidistantes del plano. Ya que $\vec{\mathbf{E}}$ es paralelo a la superficie curva, y debido a eso perpendicular a $d\vec{\mathbf{A}}$ en todos los puntos sobre la superficie, la condición (3) se satisface y no hay aportación a la integral de superficie por parte de esta superficie. Para los extremos planos del cilindro, se satisfacen las condiciones (1) y (2). El flujo a través de cada extremo del cilindro es EA; por esto, el flujo total a través de toda la superficie gaussiana es justo el que atraviesa los extremos, $\Phi_E = 2EA$.

Escriba la ley de Gauss para esta superficie y observe que la carga encerrada es $q_{\text{int}} = \sigma A$:

$$\Phi_E = 2EA = \frac{q_{\text{int}}}{\epsilon_0} = \frac{\sigma A}{\epsilon_0}$$

Resuelva para E:

$$E = \boxed{\frac{\sigma}{2\epsilon_0}} \qquad \textbf{(24.8)}$$

Finalizar Ya que la distancia desde cada extremo plano del cilindro al plano no aparece en la ecuación 24.8, se concluye que $E = \sigma/2\epsilon_0$ en *cualquier* distancia desde el plano. Es decir: el campo es uniforme en todas partes. La figura 24.14 muestra este campo eléctrico uniforme debido a un plano infinito de carga, visto de perfil.

¿QUÉ PASARÍA SI? Suponga que dos planos infinitos de carga son mutuamente paralelos, uno con carga positiva y el otro con carga negativa. Ambos planos tienen la misma densidad de carga superficial. ¿En esta situación a qué se parece el campo eléctrico?

Respuesta En primer lugar, abordamos esta configuración en la sección **¿Qué pasaría si?** del ejemplo 23.9. Los campos eléctricos debidos a los dos planos se suman en la región entre los planos, lo que resulta en un campo uniforme de magnitud σ/ϵ_0, y se cancela en otra parte para dar un campo de cero. La figura 24.15 muestra las líneas de campo para tal configuración. Este método es una forma práctica de lograr campos eléctricos uniformes con planos de tamaño finito colocados uno cerca del otro.

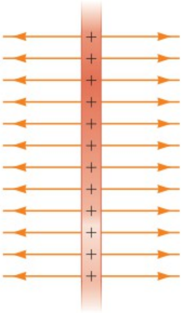

Figura 24.14 (Ejemplo 24.5) Líneas de campo eléctrico debidas a un plano infinito de carga positiva.

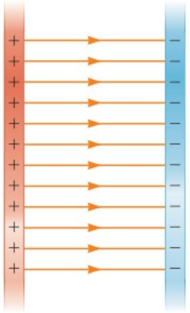

Figura 24.15 (Ejemplo 24.5) Líneas de campo eléctrico entre dos planos infinitos de carga, uno positivo y otro negativo. En la práctica, las líneas de campo cerca de los bordes de las hojas de tamaño finito de carga se curvan hacia fuera.

Ejemplo conceptual 24.6 ¡En este caso no utilice la ley de Gauss!

Explique por qué no es posible utilizar la ley de Gauss para calcular el campo eléctrico cercano a un dipolo eléctrico, a un disco con carga o a un triángulo con una carga puntual en cada vértice.

24.6 continuación

SOLUCIÓN

Las distribuciones de carga de todas estas configuraciones no tienen una simetría suficiente que haga práctico el uso de la ley gaussiana. En este caso no es posible encontrar una superficie cerrada que rodee a cualquiera de las distribuciones y satisfaga una o más de las cuatro condiciones mencionadas al principio de esta sección.

24.4 Conductores en equilibrio electrostático

Como aprendió en la sección 23.2, un buen conductor eléctrico contiene cargas (electrones) que no se encuentran unidas a ningún átomo y debido a eso tienen la libertad de moverse en el interior del material. Cuando dentro de un conductor no existe ningún movimiento neto de carga, el conductor está en **equilibrio electrostático.** Un conductor en equilibrio electrostático tiene las siguientes propiedades:

Propiedades de un conductor en equilibrio electrostático.

1. En el interior del conductor el campo eléctrico es cero, si el conductor es sólido o hueco.
2. Si un conductor aislado tiene carga, ésta reside en su superficie.
3. El campo eléctrico justo fuera de un conductor con carga es perpendicular a la superficie del conductor y tiene una magnitud σ/ϵ_0, donde σ es la densidad de carga superficial en ese punto.
4. En un conductor de forma irregular, la densidad de carga superficial es máxima en aquellos puntos donde el radio de curvatura de la superficie es menor.

En el análisis siguiente se comprueban las primeras tres propiedades. Se presenta la cuarta propiedad para tener una lista completa de las mismas para conductores en equilibrio electrostático (pero esta última no podrá ser comprobada sino hasta el capítulo 25.)

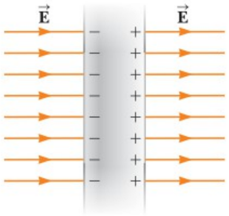

Figura 24.16 Placa conductora en un campo eléctrico externo $\vec{\mathbf{E}}$. Las cargas inducidas sobre las dos superficies de la placa producen un campo eléctrico que se opone al campo externo, lo que produce en el interior de la placa un campo eléctrico resultante igual a cero.

Es posible comprender la primera propiedad considerando una placa conductora colocada en un campo eléctrico externo $\vec{\mathbf{E}}$ (figura 24.16). El campo eléctrico en el interior del conductor *debe ser* igual a cero bajo la hipótesis de que existe equilibrio electrostático. En caso de que el campo no sea cero, los electrones libres en el interior del conductor experimentarían una fuerza eléctrica ($\vec{\mathbf{F}} = q\vec{\mathbf{E}}$) y, debido a ella, se acelerarían. Sin embargo, este movimiento de electrones significaría que el conductor no está en equilibrio electrostático. Por lo tanto, la existencia de un equilibrio electrostático es consistente solamente cuando se tiene un campo cero dentro del conductor.

Vea cómo se da este campo cero. Antes de aplicar el campo externo, en todo el volumen del conductor están distribuidos electrones libres de manera uniforme. Al aplicar el campo externo, los electrones libres se aceleran hacia la izquierda, como se ve en la figura 24.16, lo que genera un plano con carga negativa que se acumula en la superficie izquierda. El movimiento de los electrones hacia la izquierda causa un plano de carga positiva en la superficie derecha. Estos planos de carga crean un campo eléctrico adicional en el interior del conductor que se opone al campo externo. Conforme se mueven los electrones, las densidades de carga superficial de las superficies izquierda y derecha se incrementan hasta que la magnitud del campo interno es igual a la magnitud del externo, lo que resulta en un campo eléctrico cero en el interior del conductor. El tiempo que necesita un buen conductor para alcanzar el equilibrio es del orden de 10^{-16} s, lo que para la mayor parte de los efectos se considera instantáneo.

Si el conductor es hueco, el campo eléctrico dentro del conductor también es cero, ya sea que considere puntos en el conductor o en la cavidad dentro del conductor. El valor cero del campo eléctrico en la cavidad es más fácil de argumentar con el concepto de potencial eléctrico, así que este tema se abordará en la sección 25.6.

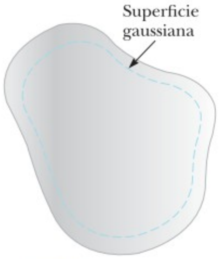

Figura 24.17 Conductor con forma arbitraria. La línea discontinua representa una superficie gaussiana que puede estar justo adentro de la superficie del conductor.

Es posible utilizar la ley de Gauss para verificar la segunda propiedad de un conductor en equilibrio electrostático. La figura 24.17 muestra un conductor de forma arbitraria. Se ha dibujado una superficie gaussiana en el interior del conductor, que quede

( )%

- Books-019-2021/20210916/
Fisica-Eletricidad Y Magnetismo/
6075224904.pdf

Pág. 731(51) 24.3

Pag. 725(45) Capitulo 24
Ley de Gauss.

---

milling - fresamento.

Books-019-2021/Important/
Geometry Illuminated... pdf

Pág. 115(132) - Chapter 12
Parallel Projection

Pág. 116(133) - Theorem 12.3
A parallel projection is both one-to-one and onto.

Obs: Ver livro todo!!!

Pag. 149(166) - Chapter 15
Circumference

Figure 1. Shaving a corner off of a polygon.

7 +,-.+ + +/ ! + ') 1 2 & ! # , & ! #
& # A # ! ! # D23 4 ! 1 2 3#
(" ' !3 ! ! 7

12

Parallel Projection

In the course of this chapter, we will need to use a few properties of parallelograms. It is rather common to confuse the properties of a parallelogram with its definition, so first, recall the definition of a parallelogram.

Definition 12.1 A *parallelogram* is a simple quadrilateral whose opposite sides are parallel.

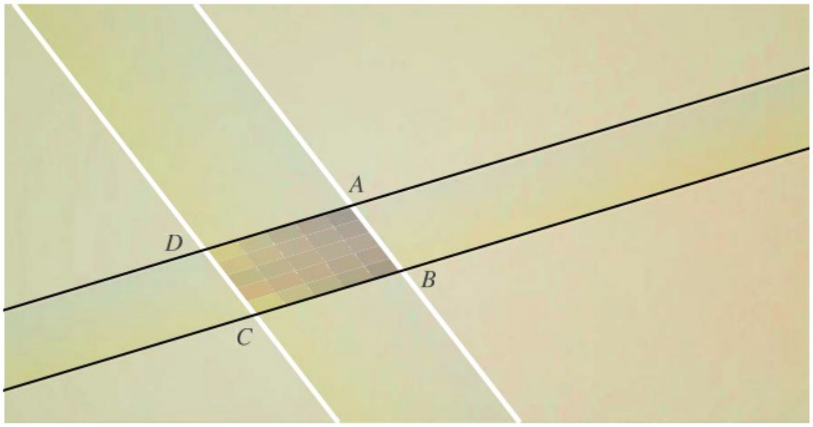

Figure 1. A parallelogram.

Now on to the properties of parallelograms that we will need for this chapter.

Exercise 12.1 Prove that in a parallelogram, the two pairs of opposite sides are congruent and the two pairs of opposite angles are congruent.

Exercise 12.2 Prove that if a convex quadrilateral has one pair of opposite sides that are both parallel and congruent, then it is a parallelogram.

Exercise 12.3 Let $\square ABB'A'$ be a simple quadrilateral. Verify that if AA' and BB' are parallel, but AB and $A'B'$ are not, then AA' and BB' cannot be congruent.

12.1 Parallel projection

The purpose of this chapter is to introduce a mechanism called *parallel projection*, a particular kind of mapping from the points on one line to the points on another. Parallel projection is the piece of machinery that you have to have in place to understand similarity, which is in turn essential for much of what we will be doing in the next chapters. The primary goal at this point is to understand how distances between points may be distorted by the parallel projection mapping. Once that is figured out, we will be able to turn our attention to the geometry of similarity.

Definition 12.2 A *parallel projection* from one line ℓ to another ℓ' is a map $\Phi : \ell \to \ell'$ that assigns to each point P on ℓ a point $\Phi(P)$ on ℓ' so that all the lines connecting a point and its image are parallel to one another.

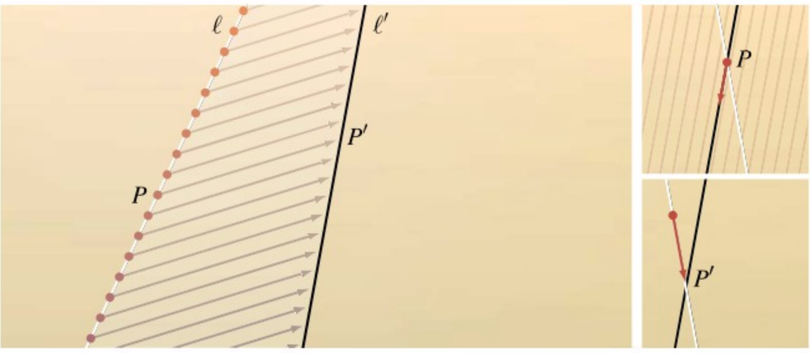

Figure 2. The path from a point P on ℓ to a point P' on ℓ' defines a parallel projection as long as neither P nor P' is the intersection of ℓ and ℓ' (as shown at right).

It is easy to construct parallel projections. Any point P on ℓ and its image $\Phi(P)$ on ℓ' completely determines the projection: for any other point Q on ℓ there is a unique line that passes through Q and is parallel to $\overline{P\,\Phi(P)}$. Wherever the line intersects ℓ' will have to be $\Phi(Q)$. There are only two cases where this construction will not work: if P is the intersection of ℓ and ℓ', then the lines of projection run parallel to ℓ' and so fail to provide a point of intersection; and if $\Phi(P)$ is the intersection of ℓ and ℓ', then the lines of projection actually coincide rather than being parallel.

Theorem 12.3 *A parallel projection is both one-to-one and onto.*

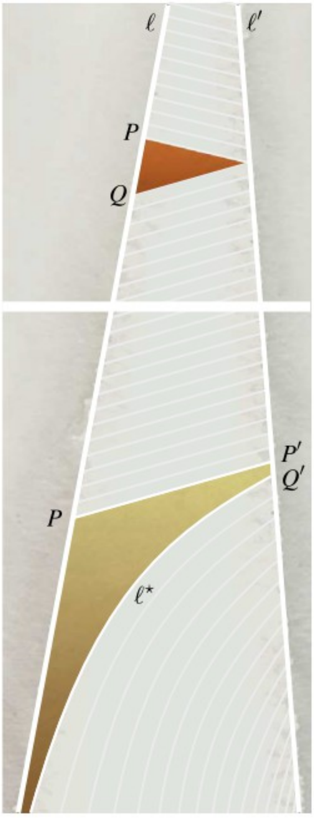

Figure 3. Using proofs by contradiction, we can show that parallel projection is both one-to-one (top) and onto (bottom).

Proof Let $\Phi : \ell \to \ell'$ be a parallel projection. First let's see why Φ is one-to-one. Suppose that it is not. That is, suppose that P and Q are two distinct points on ℓ but that $\Phi(P) = \Phi(Q)$. Then the two projecting lines $\overline{P\,\Phi(P)}$ and $\overline{Q\,\Phi(Q)}$, which ought to be parallel, actually share a point. This can't happen.

Now let's see why Φ is onto. Suppose that Q' is a point on ℓ' that is not in the image of Φ. To get a sense of how Φ is casting points from ℓ to ℓ', consider a point P on ℓ and its image $P' = \Phi(P)$ on ℓ'. By Playfair's axiom, there is a unique line $\ell^\star$ that passes through Q' and is parallel to $\overline{PP'}$. If $\ell^\star$ intersects ℓ at a point Q, then $\Phi(Q) = Q'$, contradicting the initial supposition that Q' is not in the image of Φ. If $\ell^\star$ does not intersect ℓ, then both ℓ and $\overline{PP'}$ are lines that pass through P and are parallel to $\ell^\star$, in violation of Playfair's axiom. In either case, we arrive at a contradiction. Therefore, an arbitrary point Q' of ℓ must be in the image of Φ, and so Φ is onto. $\square$

Now that we have established that parallel projection is a bijection, it helps to use a naming convention to makes things more readable. I will use a prime mark $'$ to indicate the parallel projection of a point, so $\Phi(P) = P'$, $\Phi(Q) = Q'$, and so on.

12.2 Parallel projection, order, and congruence

So far we have seen that parallel projection establishes a correspondence between the points of one line and the points of another. What about the order of the points? Can points get shuffled up in the process of a parallel projection? No.

Theorem 12.4 *Let $\Phi : \ell \to \ell'$ be a parallel projection. If A, B, and C are points on ℓ and $A * B * C$, then $A' * B' * C'$.*

Proof Because B is between A and C, A and C must be on opposite sides of $\overline{BB'}$. But, $\overline{AA'}$ does not intersect $\overline{BB'}$, so A' has to be on the same side of $\overline{BB'}$ as A; and $\overline{CC'}$ does not

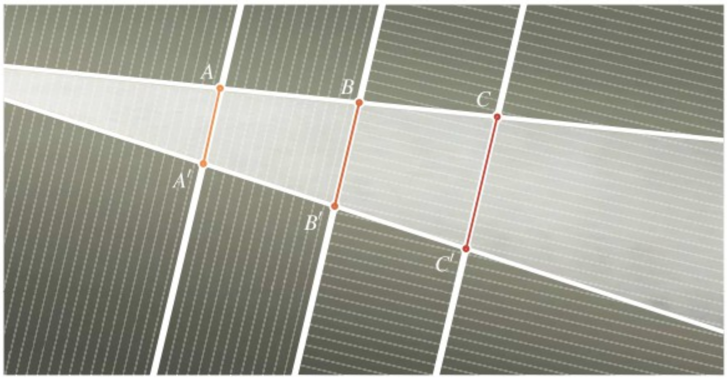

Figure 4. Parallel projection keeps points in order.

intersect $\overline{BB'}$, so C' has to be on the same side of $\overline{BB'}$ as C. Therefore A' and C' have to be on opposite sides of $\overline{BB'}$, and so the intersection of $\overline{BB'}$ and $A'C'$, which is B', must be between A' and C'. □

Now what about congruence?

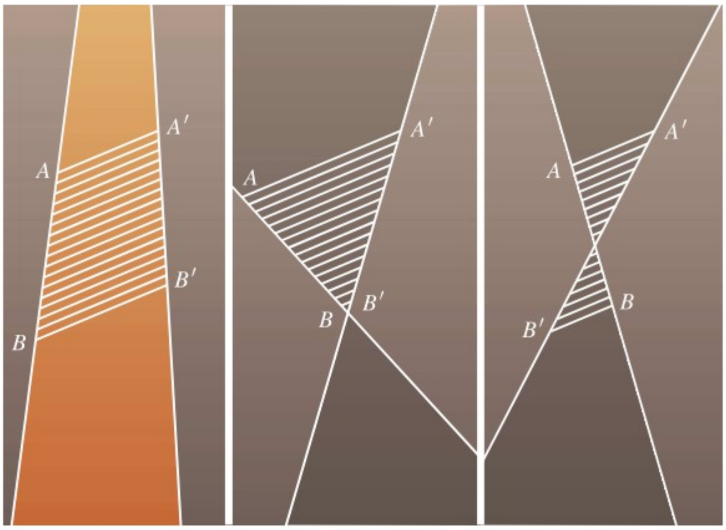

Figure 5. There are three positions for A and B relative to the image line: both on the same side, one on the image line, or one on each side. Likewise, there are three positions for a and b. There are nine cases in all.

Theorem 12.5 *Let $\Phi : \ell \to \ell'$ be a parallel projection. If a, b, A, and B are points on ℓ and if $ab \simeq AB$, then $a'b' \simeq A'B'$.*

Proof There are several cases, depending on the positions of ab and AB relative to ℓ'. They could lie on the same side of ℓ', they could lie on opposite sides of ℓ', one or both could straddle

ℓ', or one or both could have an endpoint on ℓ'. Each case must be handled differently. We will look at the case where both segments are on the same side of ℓ'. The others are left as exercises. There are still two cases to consider, depending on whether or not ℓ and ℓ' are parallel.

Case 1: when ℓ and ℓ' are parallel (see figure 6). This is the simpler case, but it helps illuminate the more general case. Because aa' is parallel to bb' and ab is parallel to $a'b'$, $\square aa'b'b$ is a parallelogram; AA' is parallel to BB' and AB is parallel to $A'B'$ so $\square AA'B'B$ is also a parallelogram. Because the opposite sides of a parallelogram are congruent (exercise 12.1), $a'b' \simeq ab$ and $AB \simeq A'B'$. Since $ab \simeq AB$, $a'b' \simeq A'B'$.

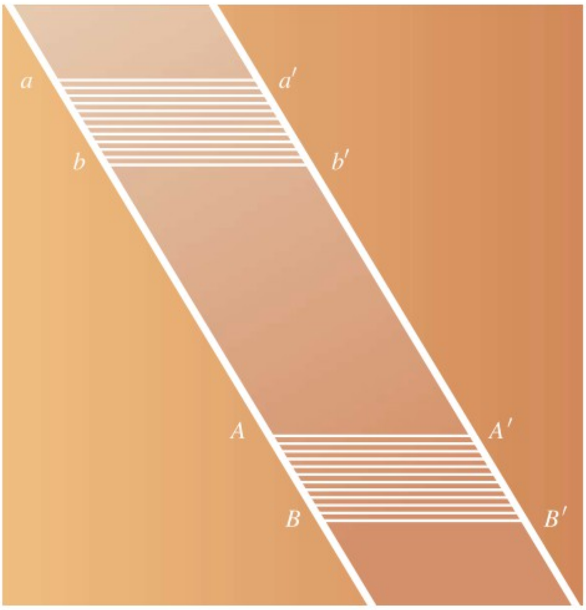

Figure 6. Case 1, when the two lines are parallel.

Case 2: when ℓ and ℓ' are not parallel (see figure 7). This is the more likely case, in which $\square aa'b'b$ and $\square AA'B'B$ will not be parallelograms. We can takes a similar approach to case 1, though, by building some parallelograms into the problem. Because ℓ and ℓ' are not parallel, the segments aa' and bb' cannot have the same length (exercise 12.3), and the segments AA' and BB' cannot have the same length. Let's assume that aa' is shorter than bb' and that AA' is shorter than BB'. If this is not the case, then it is just a matter of switching some labels to make it so. Then

- there is a point c between b and b' so that $bc \simeq aa'$, and
- there is a point C between B and B' so that $BC \simeq AA'$.

This creates four shapes of interest: the quadrilaterals $\square a'abc$ and $\square A'ABC$ that are parallelograms (exercise 12.2), and the triangles $\triangle a'b'c$ and $\triangle A'B'C$. The key here is to prove that $\triangle a'b'c \simeq \triangle A'B'C$.

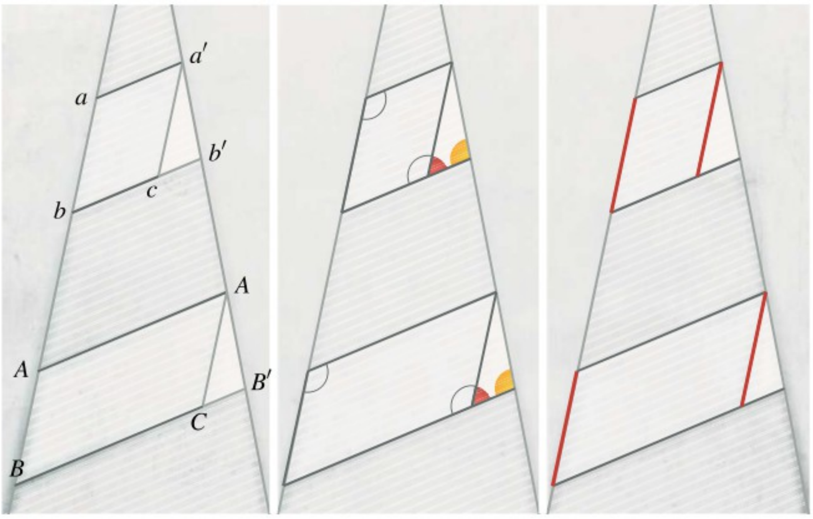

Figure 7. The non-parallel case. (l) First construct parallelograms. (c) The two angles and (r) the sides needed for AAS to show that $\triangle a'b'c' \simeq \triangle A'B'C'$.

- **A:** $\angle b' \simeq \angle B'$. The lines $\overline{cb'}$ and $\overline{CB'}$ are parallel (they are two of the projecting lines) and they are crossed by the transversal ℓ'. By the converse of the alternate interior angle theorem, $\angle a'b'c$ and $\angle A'B'C$ are congruent.
- **A:** $\angle c \simeq \angle C$. The opposite angles of the two parallelograms are congruent. Therefore $\angle a'cb \simeq \angle a'ab$ and $\angle A'AB \simeq \angle A'CB$. But $\overline{aa'}$ and $\overline{AA'}$ are parallel lines cut by the transversal ℓ, so $\angle a'ab \simeq \angle A'AB$, so $\angle a'cb \simeq \angle A'CB$, and so their supplements $\angle a'cb'$ and $\angle A'CB'$ are also congruent.
- **S:** $a'c \simeq A'C$. The opposite sides of the two parallelograms are congruent too. Therefore $a'c \simeq ab$ and $AB \simeq A'C$, and since $ab \simeq AB$, we get $a'c \simeq A'C$.

By AAS, then, $\triangle a'b'c \simeq \triangle A'B'C$. The corresponding sides $a'b'$ and $A'B'$ have to be congruent. □

Exercise 12.4 Fill in some gaps:

(1) Prove the theorem in the case when both a and A lie on ℓ' (at the intersection of ℓ and ℓ').
(2) Prove the theorem in the case when a and A lie on opposite sides of ℓ' from b and B.

12.3 Parallel projection and distance

That brings us to the question at the very heart of parallel projection. If Φ is a parallel projection and A and B are two points on ℓ, how do the lengths $|AB|$ and $|A'B'|$ compare? In case 1 of the last proof, AB and $A'B'$ ended up being congruent, but that was because ℓ and ℓ' were parallel. In general, AB and $A'B'$ do not have to be congruent. But it turns out that in the process of parallel projecting from one line to another, all distances are scaled by a constant multiple.

Theorem 12.6 *If $\Phi : \ell \to \ell'$ is a parallel projection, then there is a constant k such that*

$$|A'B'| = k|AB|$$

for all points A and B on ℓ.

It is helpful to get a little perspective on this before the proof. First note that the previous theorem gives us a way to narrow the scope of the problem. Fix a point O on ℓ and let r be one of the two rays along ℓ with O as its endpoint. The segment construction axiom says that every segment AB on ℓ is congruent to a segment OP where P is some point on r. We have just seen that parallel projection maps congruent segments to congruent segments. So if Φ scales all segments of the form OP by a factor of k, then it must scale all the segments of ℓ by the same factor.

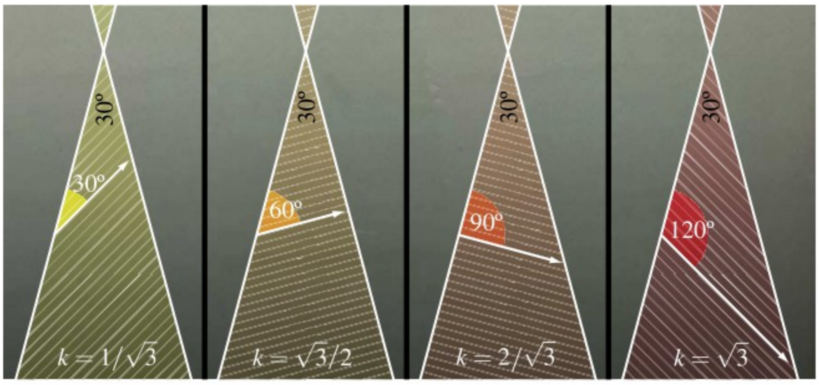

Figure 8. Some parallel projections and their scaling constants.

Second, we can extend that by considering what happens when we parallel project end-to-end congruent copies of a segment. For this, let me introduce another convenient notation convention: for the rest of this argument, let P_d denote the point on r that is a distance d from O. Now, pick a positive real value x, and let

$$k = |O'P_x'|/|OP_x|,$$

so that Φ scales the segment OP_x by a factor of k. Eventually we will have to show that Φ scales all segments by that factor, but for now let's restrict our attention to the segments OP_{nx}, where n is a positive integer. Between O and P_{nx} are P_x, P_{2x}, $\ldots$ $P_{(n-1)x}$ in order:

$$O * P_x * P_{2x} * \cdots * P_{(n-1)x} * P_{nx}.$$

We have seen that parallel projection preserves the order of points, so

$$O' * P_x' * P_{2x}' * \cdots * P_{(n-1)x}' * P_{nx}'.$$

Each segment $P_{ix}P_{(i+1)x}$ is congruent to OP_x and consequently each parallel projection $P'_{ix}P'_{(i+1)x}$ is congruent to $O'P'_x$. Add their lengths:

$$\begin{aligned}|O'P'_{nx}| &= |O'P'_x| + |P'_xP'_{2x}| + |P'_{2x}P'_{3x}| + \cdots + |P'_{(n-1)x}P'_{nx}| \\ &= kx + kx + kx + \cdots + kx \quad (n \text{ times}) \\ &= k \cdot nx\end{aligned}$$

and so Φ scales OP_{nx} by a factor of k.

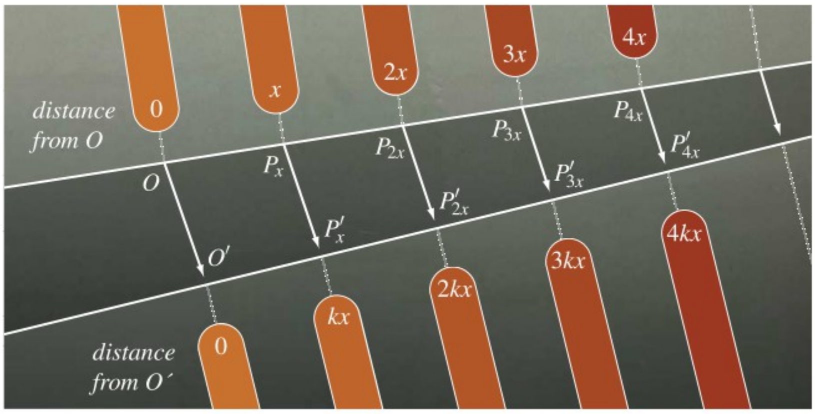

Figure 9. Scaling of segments of length kx, where k is an integer.

Sadly, no matter what x is, the points P_{nx} account for an inconsequential portion of the set of all points of r. However, if OP_x and OP_y were to have two different scaling factors we could use this end-to-end copying to magnify the difference between them. Third, then, consider an example that demonstrates how this works, and which shows how this ultimately prevents there being two different scaling factors. In this example, let's suppose that $|O'P'_1| = 2$, so all integer length segments on ℓ are scaled by a factor of 2. Let's take a look at what this means for $P_{3.45}$. Let k be the scaling factor for $OP_{3.45}$. Figure 10 shows how the first few end-to-end copies of $OP_{3.45}$ begin to limit the possible values of k.

Fourth, note that the floor function, $f(x) = \lfloor x \rfloor$, assigns to each real number x the largest integer that is less than or equal to it. The ceiling function, $f(x) = \lceil x \rceil$, assigns to each real number x the smallest integer that is greater than or equal to it.

It is finally time to prove that parallel projection scales distance.

Proof Let $k = |O'P'_1|$ so that k is the scaling factor for the segment of length 1 (and consequently for all integer length segments). Now take a point P_x on ℓ and let k' be the scaling factor for the segment OP_x. We want to show that $k' = k$ and we can do that by following the same strategy as in the example above, by capturing k' in an increasingly narrow band around k by looking at the parallel projection of P_{nx} as n increases. Start by working back and forth between

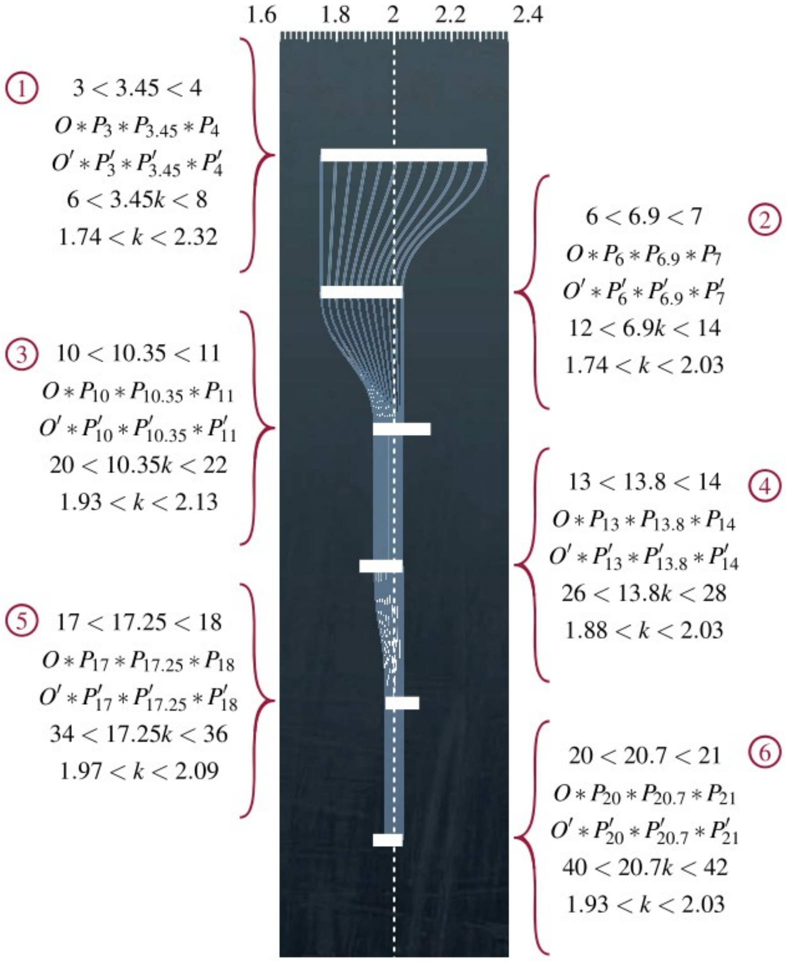
1.6
1.8
2
2.2
2.4
①
$3 < 3.45 < 4$
$O * P_3 * P_{3.45} * P_4$
$O' * P'_3 * P'_{3.45} * P'_4$
$6 < 3.45k < 8$
$1.74 < k < 2.32$
②
$6 < 6.9 < 7$
$O * P_6 * P_{6.9} * P_7$
$O' * P'_6 * P'_{6.9} * P'_7$
$12 < 6.9k < 14$
$1.74 < k < 2.03$
③
$10 < 10.35 < 11$
$O * P_{10} * P_{10.35} * P_{11}$
$O' * P'_{10} * P'_{10.35} * P'_{11}$
$20 < 10.35k < 22$
$1.93 < k < 2.13$
④
$13 < 13.8 < 14$
$O * P_{13} * P_{13.8} * P_{14}$
$O' * P'_{13} * P'_{13.8} * P'_{14}$
$26 < 13.8k < 28$
$1.88 < k < 2.03$
⑤
$17 < 17.25 < 18$
$O * P_{17} * P_{17.25} * P_{18}$
$O' * P'_{17} * P'_{17.25} * P'_{18}$
$34 < 17.25k < 36$
$1.97 < k < 2.09$
⑥
$20 < 20.7 < 21$
$O * P_{20} * P_{20.7} * P_{21}$
$O' * P'_{20} * P'_{20.7} * P'_{21}$
$40 < 20.7k < 42$
$1.93 < k < 2.03$

Figure 10. An example. Projecting end-to-end congruent copies of a segment of length 3.45 by a parallel projection that scales all integer length segments by a factor of two.

inequalities of real numbers and orders of points:

$$
\begin{gathered}
\lfloor nx \rfloor \le nx \le \lceil nx \rceil \\
O * P_{\lfloor nx \rfloor} * P_{nx} * P_{\lceil nx \rceil} \\
O' * P'_{\lfloor nx \rfloor} * P'_{nx} * P'_{\lceil nx \rceil} \\
k\lfloor nx \rfloor \le k'nx \le k\lceil nx \rceil .
\end{gathered}
$$

(The notation here is a bit casual—if nx is an integer, then $\lfloor nx \rfloor = nx$, so $P_{\lfloor nx \rfloor}$ is not actually between O and P_{nx}. This does not cause any problems for the proof.) Because $(nx - 1) \le \lfloor nx \rfloor$

and $\lceil nx \rceil \leq (nx + 1)$ we can move to a less precise but more manageable set of inequalities:

$$k(nx - 1) \leq k'nx \leq k(nx + 1).$$

Divide by nx to get

$$k\left(\frac{nx - 1}{nx}\right) \leq k' \leq k\left(\frac{nx + 1}{nx}\right).$$

As n increases, the ratios $(nx - 1)/(nx)$ and $(nx + 1)/(nx)$ both approach 1. In the limit as n goes to infinity, they are 1. Since the inequalities have to be true for all n, the only possible value for k' , then, is k. □

12.4 Exercises

12.5. Suppose that Φ is a parallel projection from ℓ to ℓ'. If ℓ and ℓ' intersect at P, prove that $\Phi(P) = P$.

12.6. Prove that if ℓ and ℓ' are parallel, then the scaling factor of a parallel projection between them must be one, but that if ℓ and ℓ' are not parallel, then there is a parallel projection with every possible scaling factor k where $0 < k < \infty$.

12.7. Let $\Phi : \ell \to \ell'$ be a parallel projection. Show that the inverse map $\Phi^{-1} : \ell' \to \ell$ is also a parallel projection.

12.8. Let $\Phi_1 : \ell \to \ell'$ and $\Phi_2 : \ell' \to \ell''$ be parallel projections. Is the composition $\Phi_2 \circ \Phi_1 : \ell \to \ell''$ necessarily a parallel projection?

12.9. In chapter 6, we constructed a distance function, and one of the keys to the construction was locating the points on a ray that were a distance of $m/2^n$ from its endpoint. In Euclidean geometry, there is a construction that locates all the points on a ray that are any rational distance m/n from its endpoint. Take two non-opposite rays r and r' with a common endpoint O. Along r, lay out m congruent copies of a segment of length 1, ending at the point P_m. Along r', lay out n congruent copies of a segment of length 1, ending at the point Q_n. Mark the point Q_1 on r' that is a distance 1 from O. Verify that the line that passes through Q_1 and is parallel to $P_m Q_n$ intersects r a distance of m/n from O.

( )%

- Books-019-2021/20210916/
Fisica-Eletricidad Y Magnetismo/
6075224904.pdf

Pág. 731(51) 24.3

Pag. 725(45) Capitulo 24
Ley de Gauss.

—

milling - fresamento.

Books-019-2021/Important/
Geometry Illuminated... pdf

Pág. 115(132) - Chapter 12
Parallel Projection

Pág. 116(133) - Theorem 12.3
A parallel projection is both one-to-one and onto.

Obs: Ver livro todo!!!

—

Pag. 149(166) - Chapter 15
Circumference

Figure 1. Shaving a corner off of a polygon.

15

Circumference

15.1 A theorem on perimeters

In chapter 9 on polygons, we defined the perimeter of a polygon $\mathcal{P} = P_1 \cdots P_n$ as

$$|\mathcal{P}| = \sum_{i=1}^{n} |P_i P_{i+1}|,$$

but then left it at that. In this lesson we are going to use perimeters of cyclic polygons to find the circumference of the circle. Along the way, we will need to use the following result that compares the perimeters of two convex polygons when one is contained in the other. We will prove the theorem by a process of "shaving down" the sides of the larger polygon until we reach the smaller one.

Theorem 15.1 *If $\mathcal{P}$ and $\mathcal{Q}$ are convex polygons and all the points of $\mathcal{P}$ are on or inside $\mathcal{Q}$, then $|\mathcal{P}| \leq |\mathcal{Q}|$.*

Proof Some of the edges of $\mathcal{P}$ may run along the edges of $\mathcal{Q}$, but unless $\mathcal{P} = \mathcal{Q}$, at least one edge of $\mathcal{P}$ must pass through the interior of $\mathcal{Q}$. Let s be one of those interior edges. The line containing s intersects $\mathcal{Q}$ twice—call the intersections a and b—dividing $\mathcal{Q}$ into two smaller polygons that share the side ab, one on the same side of s as $\mathcal{P}$, the other on the opposite side. Define a new polygon $\mathcal{Q}_1$ that consists of

- points of $\mathcal{Q}$ on the same side of s as $\mathcal{P}$, and
- points on the segment ab.

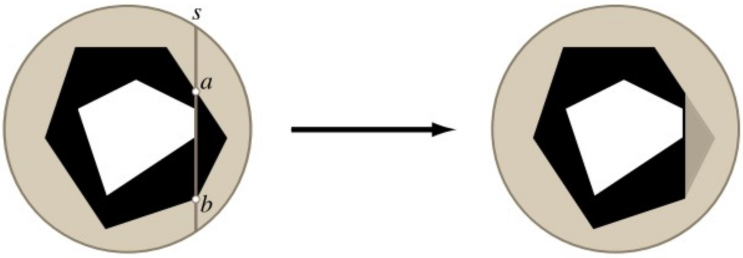

Figure 1. Shaving a corner off of a polygon.

There are two things to notice about $\mathcal{Q}_1$. First, $\mathcal{Q}_1$ and $\mathcal{P}$ have one more coincident side (the side s) than $\mathcal{Q}$ and $\mathcal{P}$ had. Second, the portions of $\mathcal{Q}$ and $\mathcal{Q}_1$ on the side of s with $\mathcal{P}$ are identical, so the segments making up that part contribute the same amount to their perimeters. On the other side, though, the path that $\mathcal{Q}$ takes from a to b is longer than the direct route along the segment ab of $\mathcal{Q}_1$ (because of the triangle inequality). Combining the two parts, we get $|\mathcal{Q}_1| \leq |\mathcal{Q}|$.

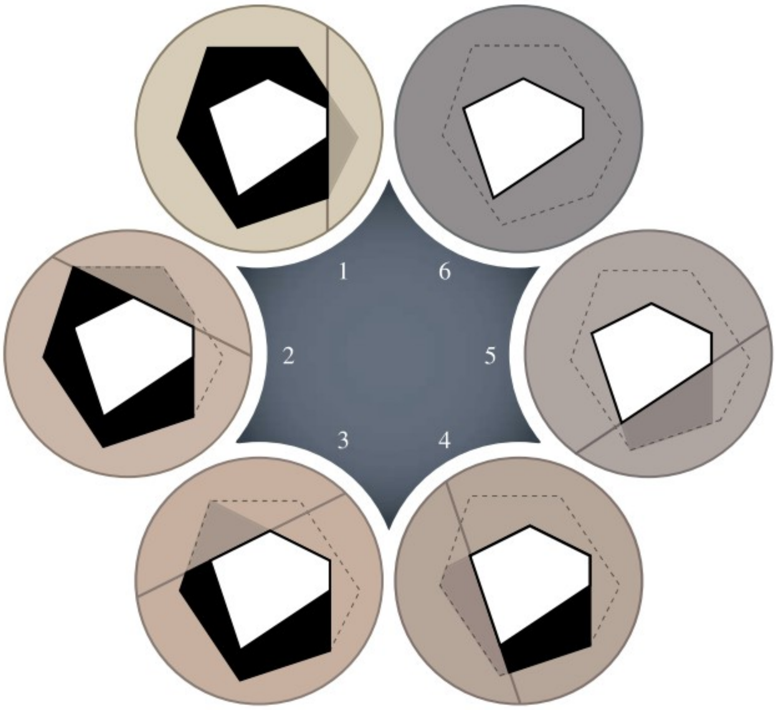

Figure 2. One at a time, shave the edges of the outer polygon down to the inner one.

Now we can repeat this process with $\mathcal{P}$ and $\mathcal{Q}_1$, generating $\mathcal{Q}_2$ with even smaller perimeter than $\mathcal{Q}_1$ and another coincident side with $\mathcal{P}$. And again, to get $\mathcal{Q}_3$. Eventually, though, after say m steps, we run out of sides that pass through the interior, at which point $\mathcal{P} = \mathcal{Q}_m$. Then

$$|\mathcal{P}| = |\mathcal{Q}_m| \leq |\mathcal{Q}_{m-1}| \leq \cdots \leq |\mathcal{Q}_2| \leq |\mathcal{Q}_1| \leq |\mathcal{Q}|. \qquad \square$$

15.2 Circumference

The purpose of this chapter is to understand the connection between the radius of a circle, and the distance around it, its circumference. The final result, the formula $C = 2\pi r$, ranks with the Pythagorean theorem in celebrity, although I think it is an often misunderstood celebrity. Let's be clear about what this equation is not. It is not an equation comparing two known quantities C and $2\pi r$. Instead, it is saying that the ratio $C/(2r)$ is a constant. The constant is what we have come to call π.

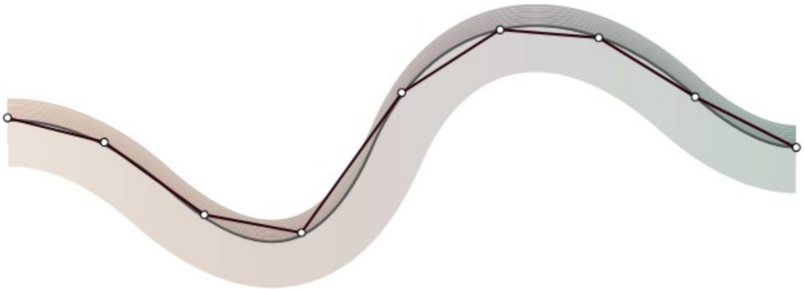

Figure 3. An approximation of an arc by segments.

To define the circumference of a circle, let's borrow an idea from calculus, the idea of approximating a curve by straight line segments, and then refining the approximation by increasing the number of segments. For a circle $\mathcal{C}$, the approximating line segments will be the edges of a simple cyclic polygon $\mathcal{P}$ inscribed in the circle. We will want the circumference of $\mathcal{C}$ to be bigger than the perimeter of $\mathcal{P}$. We should also expect that by adding in additional vertices to $\mathcal{P}$, we should be able to get the perimeter of $\mathcal{P}$ as close as we want to the circumference of $\mathcal{C}$. All this suggests that to get the circumference of $\mathcal{C}$, we need to find out how large the perimeters of inscribed polygons can be.

Definition 15.2 The circumference of a circle $\mathcal{C}$, written $|\mathcal{C}|$, is

$$|\mathcal{C}| = \sup \left\{ |\mathcal{P}| \middle| \mathcal{P} \text{ is a simple cyclic polygon inscribed in } \mathcal{C} \right\}.$$

There is nothing in the definition to guarantee that the supremum exists. It is conceivable that the lengths of the approximating perimeters might grow with bound. An example is the *Koch snowflake*. It is constructed as follows. Take an equilateral triangle with sides of length 1. The perimeter of this triangle is 3. Divide each side into thirds. On each middle third, build an equilateral triangle by adding two more sides then remove the original side. You have made a shape with $3 \cdot 4$ sides, each with length 1/3, for a perimeter of 4. Now iterate: divide the sides into thirds; build equilateral triangles on each middle third, and remove the base. That will make $3 \cdot 16$ sides of length 1/9, for a perimeter of 16/3. After the next step, there are $3 \cdot 64$ sides of length 1/27 for a perimeter of $64/9$. Generally, after n iterations, there are $3 \cdot 4^n$ sides of length $1/3^n$ for a total perimeter of $4^n/3^{n-1}$, and

$$\lim_{n\to\infty} \frac{4^n}{3^{n-1}} = \lim_{n\to\infty} 3\left(\frac{4}{3}\right)^n = \infty.$$

The Koch snowflake, which is the limiting shape in this process, has infinite perimeter! The first thing we need to do, then, is to make sure that circles are better behaved than this.

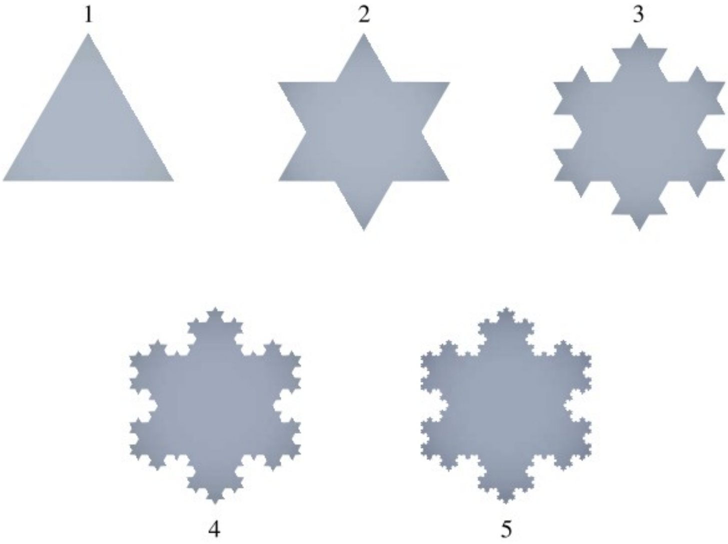

Figure 4. The first few steps in the construction of the Koch snowflake.

Theorem 15.3 *If* $\mathcal{C}$ *is a circle of radius* r*, then* $|\mathcal{C}| \leq 8r$.

Proof The first step is to build a circumscribing square around $\mathcal{C}$, the smallest square that still contains $\mathcal{C}$. Begin by choosing two perpendicular diameters d_1 and d_2. Each will intersect $\mathcal{C}$ twice, for a total of four intersections, P_1, P_2, P_3, and P_4. For each i between 1 and 4, let t_i be the tangent line to $\mathcal{C}$ at P_i. The tangents intersect to form the circumscribing square. The length of each side of the square is equal to the diameter of $\mathcal{C}$, so the perimeter of the square is $4 \cdot 2r = 8r$.

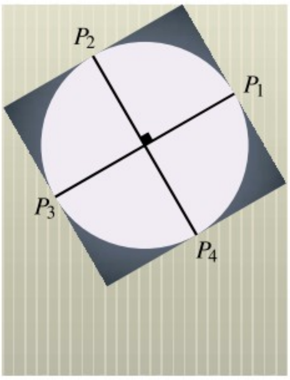

Figure 5. The circumscribing square.

Now we turn to the theorem we proved to start this chapter. Each simple cyclic polygon inscribed in $\mathcal{C}$ is a convex polygon contained in the circumscribing square. Therefore the perimeter of any such approximating polygon is bounded above by $8r$. Remember that we have defined $|\mathcal{C}|$ to be the supremum of all the approximating perimeters, so it cannot exceed $8r$ either. □

Now that we know that a circle has a circumference, the next step is to find a way to calculate it. The key to that is the next theorem.

Theorem 15.4 *The ratio of the circumference of a circle to its radius is a constant.*

Proof Let's suppose that the ratio is not a constant, so that there are two circles $\mathcal{C}_1$ and $\mathcal{C}_2$ with centers O_1 and O_2 and radii r_1 and r_2, but with unequal ratios

$$|\mathcal{C}_1|/r_1 > |\mathcal{C}_2|/r_2.$$

As we have defined circumference, there are approximating cyclic polygons to $\mathcal{C}_1$ whose perimeters are arbitrarily close to its circumference. In particular, there is an approximating cyclic polygon $\mathcal{P} = P_1 P_2 \dots P_n$ for $\mathcal{C}_1$ so that

$$|\mathcal{P}|/r_1 > |\mathcal{C}_2|/r_2.$$

We can construct a cyclic polygon $\mathcal{Q}$ on $\mathcal{C}_2$ that is similar to $\mathcal{P}$. Intuitively, we just need to scale $\mathcal{P}$ so that it fits in the circle $\mathcal{C}_2$. The formal construction is:

- Begin by placing a point Q_1 on circle $\mathcal{C}_2$.
- Locate Q_2 on $\mathcal{C}_2$ so that $\angle P_1 O_1 P_2$ is congruent to $\angle Q_1 O_2 Q_2$ (there are two choices for Q_2).
- Locate Q_3 on $\mathcal{C}_2$ and on the opposite side of $O_2 Q_2$ from Q_1 so that $\angle P_2 O_1 P_3$ is congruent to $\angle Q_2 O_2 Q_3$.
- Continue placing points on $\mathcal{C}_2$ in this fashion until Q_n has been placed to form the polygon $\mathcal{Q} = Q_1 Q_2 \dots Q_n$. At the end of the process $\angle P_n O_1 P_1$ is congruent to $\angle Q_n O_2 Q_1$ by angle subtraction.

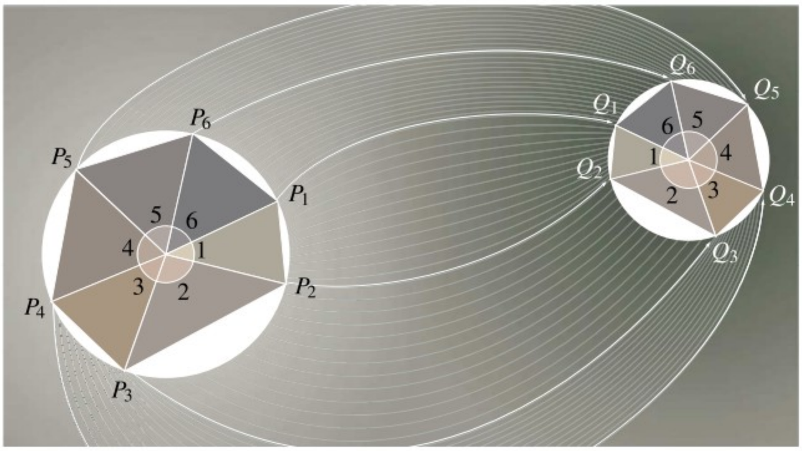

Figure 6. The ratio $|Q_i Q_{i+1}|/|P_i P_{i+1}|$ is a constant.

Then

$$\frac{|O_2Q_i|}{|O_1P_i|} = \frac{r_2}{r_1} = \frac{|O_2Q_{i+1}|}{|O_1P_{i+1}|} \text{ and } \angle Q_iO_2Q_{i+1} \simeq \angle P_iO_1P_{i+1},$$

so by SAS similarity, $\triangle Q_iO_2Q_{i+1}$ is similar to $\triangle P_iO_1P_{i+1}$. That gives us the ratio of the third sides of the triangle as

$$\frac{|Q_iQ_{i+1}|}{|P_iP_{i+1}|} = \frac{r_2}{r_1}.$$

Thus we can describe the perimeter of $\mathcal{Q}$ as

$$|\mathcal{Q}| = \sum_{i=1}^{n} |Q_iQ_{i+1}| = \sum_{i=1}^{n} \frac{r_2}{r_1}|P_iP_{i+1}| = \frac{r_2}{r_1}\sum_{i=1}^{n} |P_iP_{i+1}| = \frac{r_2}{r_1}|\mathcal{P}|.$$

That would mean that

$$\frac{|\mathcal{Q}|}{r_2} = \frac{|\mathcal{P}|}{r_1} > \frac{|\mathcal{C}_2|}{r_2}.$$

That is, $|\mathcal{Q}| > |\mathcal{C}_2|$, when in fact the circumference of $\mathcal{C}_2$ is supposed to be greater than the perimeter of any of its approximating cyclic polygons. □

Definition 15.5 The constant π is the ratio of the circumference of a circle to its diameter: $\pi = |\mathcal{C}|/(2r)$.

It is more convenient to express this definition in terms of a limit rather than a supremum. Fortunately, the supremum can be reached as a limit of perimeters of a sequence of regular polygons as follows. Arrange n angles each measuring $360^\circ/n$ around the center of a circle $\mathcal{C}$. The rays of the angles intersect $\mathcal{C}$ n times, and the points P_i are the vertices of a regular n-gon, $\mathcal{P}_n = P_1P_2\ldots P_n$. The tangent lines to $\mathcal{C}$ at the neighboring points P_i and P_{i+1} intersect at a point Q_i. Taken together, the n points are the vertices of another regular n-gon $\mathcal{Q}_n = Q_1Q_2\ldots Q_n$. The polygon $\mathcal{P}_n$ is just one of the many cyclic polygons inscribed in $\mathcal{C}$ so $|\mathcal{P}_n| \le |\mathcal{C}|$. The polygon $\mathcal{Q}_n$ circumscribes $\mathcal{C}$, and every cyclic polygon inscribed on $\mathcal{C}$ lies inside $\mathcal{Q}_n$, so $|\mathcal{Q}_n| \ge |\mathcal{C}|$.

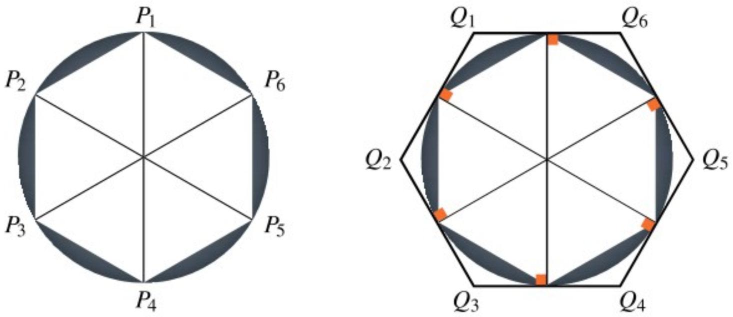

Figure 7. Regular inscribed (l) and circumscribed (r) hexagons.

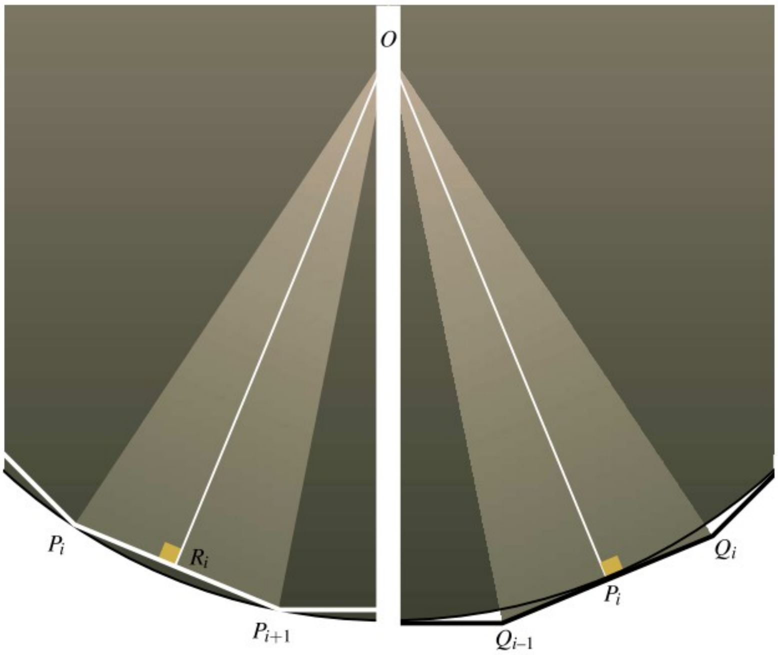

Figure 8. Approximation by inscribed and circumscribed polgons.

The lower bound prescribed by $\mathcal{P}_n$. $\overrightarrow{OQ_i}$ is a perpendicular bisector of P_iP_{i+1}, intersecting it at a point R_i and dividing $\triangle OP_iP_{i+1}$ in two. By the HL congruence theorem for right triangles, the two parts, $\triangle OR_iP_i$ and $\triangle OR_iP_{i+1}$, are congruent. Then $\mathcal{P}_n$ is built from $2n$ segments of length $|P_iR_i|$. Now for a little trigonometry:

$$\sin(360^\circ/2n) = \frac{|P_iR_i|}{r} \implies |P_iR_i| = r\sin(360^\circ/2n)$$

so $|\mathcal{P}_n| = 2nr\sin(360^\circ/2n)$.

The upper bound prescribed by $\mathcal{Q}_n$. $\overrightarrow{OP_i}$ is a perpendicular bisector of $Q_{i-1}Q_i$, intersecting it at P_i and dividing $\triangle OQ_{i-1}Q_i$ in two. By SAS, the two parts, $\triangle OP_iQ_{i-1}$ and $\triangle OP_iQ_i$, are congruent. Then $\mathcal{Q}_n$ is built from $2n$ segments of length $|P_iQ_i|$. Because

$$\tan(360^\circ/2n) = |P_iQ_i|/r \implies |P_iQ_i| = r\tan(360^\circ/2n),$$

$|\mathcal{Q}_n| = 2nr\tan(360^\circ/2n)$.

Let's compare $|\mathcal{P}_n|$ and $|\mathcal{Q}_n|$ as n increases (the key to the calculation is that as x approaches 0, $\cos(x)$ approaches 1):

$$\begin{aligned}\lim_{n\to\infty}\frac{|\mathcal{P}_n|}{|\mathcal{Q}_n|} &= \lim_{n\to\infty}\frac{2nr\tan(360^\circ/2n)}{2nr\sin(360^\circ/2n)}\\ &= \lim_{n\to\infty}\frac{\sin(360^\circ/2n)}{\cos(360^\circ/2n)}\cdot\frac{1}{\sin(360^\circ/2n)}\\ &= \lim_{n\to\infty}\frac{1}{\cos(360^\circ/2n)}\\ &= 1.\end{aligned}$$

Therefore $\lim_{n\to\infty} |\mathcal{P}_n| = \lim_{n\to\infty} |\mathcal{Q}_n|$. Since $|\mathcal{C}|$ is trapped between $|\mathcal{P}_n|$ and $|\mathcal{Q}_n|$ for all n, and since they approach the same number as n goes to infinity, $|\mathcal{C}|$ also approaches this number. That gives a more comfortable equation for circumference as

$$|\mathcal{C}| = \lim_{n\to\infty} 2nr \sin(360^\circ/2n),$$

and since $|\mathcal{C}| = 2\pi r$, we can make a definition of π as

$$\pi = \lim_{n\to\infty} n \sin(360^\circ/2n).$$

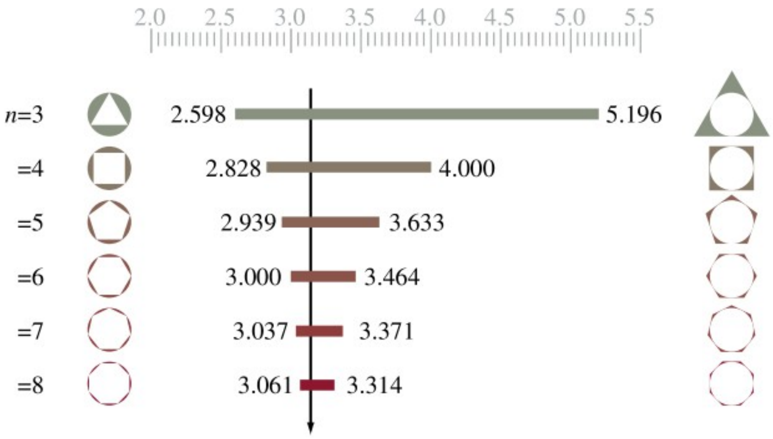

Figure 9. Using perimeters of regular inscribed and circumscribed polygons to determine upper and lower bounds for π.

15.3 Lengths of arcs and radians

It doesn't take much modification to get a formula for the length of an arc. The 360° in the formula for $|\mathcal{C}|$ is the measure of the central angle corresponding to an arc that goes completely around the circle. To get the measure of any other arc, we can replace the 360° with the measure of the corresponding central angle.

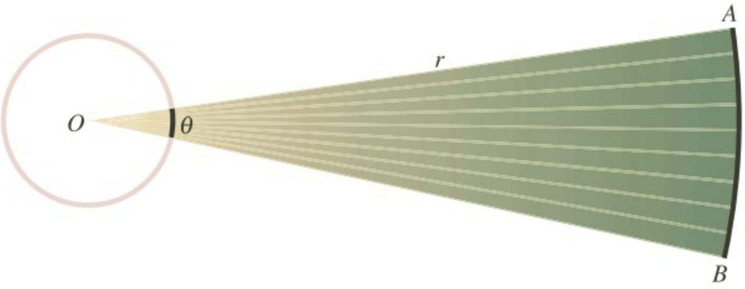

Figure 10. An arc.

Theorem 15.6 *If $\widehat{AB}$ is the arc of a circle with radius r, and if θ is the measure of the central angle $\angle AOB$, then*

$$|\widehat{AB}| = \frac{\pi}{180^\circ}\theta \cdot r.$$

Proof To start, replace the 360° in the circumference formula with θ:

$$|\widehat{AB}| = \lim_{n\to\infty} 2nr \sin(\theta/2n) = 2r \cdot \lim_{n\to\infty} n \sin(\theta/2n).$$

This limit is clearly related to the one that defines π. We can absorb the difference between the two into the variable with the substitution $n = m \cdot \theta/360^\circ$. As n approaches infinity, m will as well, so

$$\begin{aligned} |\widehat{AB}| &= 2r \cdot \lim_{m\to\infty} \frac{m \cdot \theta}{360^\circ} \sin\left(\frac{\theta}{2m \cdot \theta/360^\circ}\right) \\ &= \frac{2r\theta}{360^\circ} \cdot \lim_{m\to\infty} m \sin(360^\circ/2m) \\ &= \frac{\theta}{180^\circ} r\pi. \end{aligned}$$

□

There is one more thing to notice before the end of this chapter. The arc length formula provides a direct connection between angle measure (of the central angle) and distance (along the arc). And yet, the $\frac{\pi}{180^\circ}$ factor in that formula suggests that distance and the degree measurement system are a little out of sync with one another. This can be fixed by modernizing our method of angle measurement. The preferred angle measurement system, and the one that we will use, is *radian* measurement.

Definition 15.7 One radian is $180^\circ/\pi$.

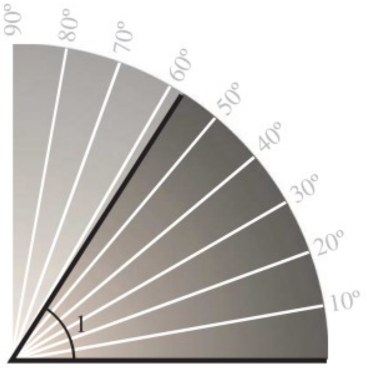

Figure 11. One radian is approximately 57.296°.

Then the measure of a straight angle is π radians. The measure of a right angle is $\pi/2$ radians. One complete turn of the circle is 2π radians. If $\theta = (\angle AOB)$ is measured in radians, then $|\widehat{AB}| = r \cdot \theta$.

15.4 Exercises

15.1. Let A and B be points on a circle $\mathcal{C}$ with radius r. Let θ be the measure of the central angle corresponding to the minor arc (or semicircle) $\widehat{AB}$. What is the relationship (in the form of an equation) between θ, r, and $|AB|$?

15.2. Let A and B be points on a circle $\mathcal{C}$ with center O and radius r. If $(\angle BAO) = \theta$, what is $|\widehat{AB}|$ (in terms of r and θ)?

15.3. Let AB be a diameter of a circle $\mathcal{C}$, and let P be a point on AB. Let $\mathcal{C}_1$ be the circle with diameter AP and let $\mathcal{C}_2$ be the circle with diameter BP. Show that the sum of the circumferences of $\mathcal{C}_1$ and $\mathcal{C}_2$ is equal to the circumference of $\mathcal{C}$ (the shape formed by the three semicircles on one side of AB is called an *arbelos*).

15.4. What is the radius of the circle that passes through the three vertices of a equilateral triangle with sides of length 1?

15.5. Let $\mathcal{C}_1$ be the circle inscribed in an equilateral triangle $\triangle ABC$ with sides of length 1. There is a line that is parallel to AB and tangent to $\mathcal{C}_1$ that divides $\triangle ABC$ into a trapezoid and a smaller equilateral triangle. Let $\mathcal{C}_2$ be the circle inscribed in the smaller equilateral triangle. Continuing in this way, we can form a stack of smaller and smaller circles. Is the sum of their circumferences finite, and if so, what is it?

15.6. In the construction of the Koch snowflake, the middle third of each segment is replaced with two-thirds of an equilateral triangle. Suppose, instead, that middle third was replaced with three of the four sides of a square. What is the perimeter of the n-th stage of this operation? Would the limiting perimeter still be infinite?

15.7. This problem deals with the possibility of angle measurement systems other than degrees or radians. Let $\mathscr{A}$ be the set of angles in the plane. Consider a function

$$\star : \mathscr{A} \to (0, \infty) : \angle A \to (\angle A)^\star$$

that satisfies

- if $\angle A \simeq \angle B$, then $(\angle A)^\star = (\angle B)^\star$
- if D is in the interior of $\angle ABC$, then

$$(\angle ABC)^\star = (\angle ABD)^\star + (\angle DBC)^\star.$$

Prove that the $\star$ measurement system is a constant multiple of the degree measurement system (or, for that matter, the radian measurement system). That is, prove that there is a $k > 0$ such that for all $\angle A \in \mathscr{A}$,

$$(\angle A)^\star = k \cdot (\angle A).$$

15.5 References

The Koch snowflake is an example of a fractal. Gerald Edgar's book *Measure, Topology, and Fractal Geometry* [Edg90], deals with these objects and their measures.

( )%

- Books-019-2021/20210916/
Fisica-Eletricidad Y Magnetismo/
6075224904.pdf

Pág. 731(51) 24.3

Pag. 725(45) Capitulo 24

Ley de Gauss.

---

milling - fresamento.

Books-019-2021/Important/
Geometry Illuminated... pdf

Pág. 115(132) - Chapter 12
Parallel Projection

Pág. 116(133) - Theorem 12.3
A parallel projection is both one-to-one and onto.

Obs: Ver livro todo!!!

Pag. 149(166) - Chapter 15
Circumference

Figure 1. Shaving a corner off of a polygon.

Applications d'Analyse et de Géométrie
Poncelet (2 vol)

—

www.archive.org/details/aperuhistoriq00chas

—

Lobachevski
https://archive.org/details/recherchesgom00lobauoft

—

Géométrie imaginaire
https://archive.org/details/gomtrieimagi00lobauoft/page/584/mode/2up

—

https://archive.org/details/ljs320/

—

- https://archive.org/details/questioniriguard00enri

—

' (
! 3 $$! 8 $# ! $!3 B !

+,-.+ + +/ ! + ' !3 B ! 3#

4 3 $$! 8 $# ! $ 4!

+,-.+ + +/ ! + ' 4! 3#
/4 > 4! =2

3 $$! 8 $# ! $! 4!

+,-.+ + +/ ! + ' ! 4! 3#
/4
1E F & ! G > 4! =
& 3 !
3 $$! 8 $# ! $!4 "' # $3! $*$ # $ 3
G ! > 4! =

+,-.+ + +/ ! + ' !4 "' # 3#

3 $$! 8 $# ! $B !

+,-.+ + +/ ! + ' B ! # 3#

+,-.+ + +/ ! + ' ! !: !: 4! 8 = 3#

+,-.+ + +/ ! + ' H!8 # A , < #
! !# I J &K ! J K /3 4 % " 3#

+,-.+ + +/ ! + ' H!8 # A , < #
! !# I J &&K ! J K /3 4 % " 3#
(0 0 ! # . !# ! , !

that I am quite wrong." I supposed, when I heard the story—but it would not have been seemly to say it—that Lord A. exhaled candor and sense, which infected those who came within reach: he would have done so, if anybody.

THE TRISECTION AND QUADRATURE AGAIN.

A method to trisect a series of angles having relation to each other; also another to trisect any given angle. By James Sabben. 1848 (two quarto pages).

"The consequence of years of intense thought": very likely, and very sad.

1848. The following was sent to me in manuscript. I give the whole of it:

"*Quadrature of the Circle.*—A quadrant is a curvilinear angle traversing round and at an equal distance from a given point, called a center, no two points in the curve being at the same angle, but irreptitiously graduating from 90 to 60. It is therefore a mean angle of 90 and 60, which is 75, because it is more than 60, and less than 90, approximately from 60 to 90, and from 90 to 60, with equal generation in each irreptitious approximation, therefore meeting in 75, and which is the mean angle of the quadrant.

"Or suppose a line drawn from a given point at 90, and from the same point at 60. Let each of these lines revolve on this point toward each other at an equal ratio. They will become one line at 75, and bisect the curve, which is one-sixth of the entire circle. The result, taking 16 as a diameter, gives an area of 201.072400, and a circumference of 50.2681.

"The original conception, its natural harmony, and the result, to my own mind is a demonstrative truth, which I presume it right to make known, though perhaps at the hazard of unpleasant if not uncourteous remarks."

I have added punctuation: the handwriting and spelling are those of an educated person; the word *irreptitious* is indubitable. The whole is a natural curiosity.

The quadrature and exact area of the circle demonstrated. By Wm. Peters. 8vo. *n. d.* (circa 1848).[1]

Suggestions as to the necessity for a revolution in philosophy; and prospectus for the establishment of a new quarterly, to be called the *Physical Philosopher and Heterodox Review.* By Q. E. D. 8vo. 1848.

These works are by one author, who also published, as appears by advertisement,

"Newton rescued from the precipitancy of his followers through a century and a half,"[2] and "Dangers along a coast by correcting (as it is called) a ship's reckoning by bearings of the land at night fall, or in a fog, nearly out of print. Subscriptions are requested for a new edition."

The area of a circle is made four-fifths of the circumscribed square: proved on an assumption which it is purposed to explain in a longer essay.[3] The author, as Q. E. D., was in controversy with the *Athenæum* journal, and criticised a correspondent, D., who wrote against a certain class of discoverers. He believed the common theories of hydrostatics to be wrong, and one of his questions was:

"Have you ever taken into account anent gravity and gravitation the fact that a five grain cube of cork will of itself half sink in the water, whilst it will take 20 grains of brass, which will sink of itself, to pull under the other half? Fit this if you can, friend D., to your notions of gravity and specific gravity, as applied to the construction of a universal law of gravitation."

This the *Athenæum* published—but without some Italics, for which the editor was sharply reproved, as a sufficient specimen of the *quod erat* D. *monstrandum*: on which the author remarks—"D,—Wherefore the e caret? is it D apostrophe? D', D'M, D'Mo, D'Monstrandum; we cannot find the *wit* of it." This I conjecture to contain an illusion to the name of the supposed author; but whether De Mocritus, De Mosthenes, or De Moivre was intended, I am not willing to decide.

---

[1] This was published in London in 1851 instead of 1848.

[2] This appeared in 1846.

[3] This was done in *The Circle Squared*, published at Brighton in 1865.

The Scriptural Calendar and Chronological Reformer, for the statute year 1849. Including a review of recent publications on the Sabbath question. London, 1849, 12mo.[4]

This is the almanac of a sect of Christians who keep the Jewish Sabbath, having a chapel at Mill Yard, Goodman's Fields. They wrote controversial works, and perhaps do so still; but I never chanced to see one.

Geometry *versus* Algebra; or the trisection of an angle geometrically solved. By W. Upton, B. A.[5] Bath (circa 1849). 8vo.

The author published two tracts under this title, containing different alleged proofs: but neither gives any notice of the change. Both contain the same preface, complaining of the British Association for refusing to examine the production. I suppose that the author, finding his first proof wrong, invented the second, of which the Association never had the offer; and, feeling sure that they would have equally refused to examine the second, thought it justifiable to present that second as the one which they had refused. Mr. Upton has discovered that the common way of finding the circumference is wrong, would set it right if he had leisure, and, in the mean time, has solved the problem of the duplication of the cube.

---

[4] It first appeared in 1847, under the title, *The Scriptural Calendar and Chronological Reformer, 1848. Including a review of tracts by Dr. Wardlaw and others on the Sabbath question. By W. H. Black.* The one above mentioned, for 1849, was printed in 1848, and was also by Black (1808-1872). He was pastor of the Seventh Day Baptists and was interested in archeology and in books. He catalogued the manuscripts of the Ashmolean Museum at Oxford.

[5] William Upton, a Trinity College man, Dublin. He also wrote *Upton's Physioglyphics,* London, 1844; *Pars prima. Geometria vindicata; antiquorumque Problematum, ad hoc tempus desperatorum, Trisectionis Anguli, Circulique Quadraturae, Solutio, per Eucliden effecta, London* (printed at Southampton), 1847; *The Uptonian Trisection,* London, 1866; and *The Circle Squared,* London, 1872.

The trisector of an angle, if he demand attention from any mathematician, is bound to produce, from his construction, an expression for the sine or cosine of the third part of any angle, in terms of the sine or cosine of the angle itself, obtained by help of no higher than the square root. The mathematician knows that such a thing cannot be; but the trisector virtually says it can be, and is bound to produce it, to save time. This is the misfortune of most of the solvers of the celebrated problems, that they have not knowledge enough to present those consequences of their results by which they can be easily judged. Sometimes they have the knowledge and quibble out of the use of it. In many cases a person makes an honest beginning and presents what he is sure is a solution. By conference with others he at last feels uneasy, fears the light, and puts self-love in the way of it. Dishonesty sometimes follows. The speculators are, as a class, very apt to imagine that the mathematicians are in fraudulent confederacy against them: I ought rather to say that each one of them consents to the mode in which the rest are treated, and fancies conspiracy against himself. The mania of conspiracy is a very curious subject. I do not mean these remarks to apply to the author before me.

One of Mr. Upton's trisections, if true, would prove the truth of the following equation:

$$3\cos(\theta/3) = 1 + \surd(4 - \sin^2\theta)$$

which is certainly false.[6]

---

[6] For example, if $\theta = 90°$ we should have $3\cos 30° = 1 + \sqrt{4 - \sin^2 90°}$, or $3 . \frac{1}{2}\sqrt{3} = 1 + \sqrt{3}$, or $\frac{1}{2}\sqrt{3} = 1$.

In 1852 I examined a terrific construction, at the request of the late Dr. Wallich,[7] who was anxious to persuade a poor countryman of his, that trisection of the angle was waste of time. One of the principles was, that "magnitude and direction determine each other." The construction was equivalent to the assertion that, θ being any angle, the cosine of its third part is

$$\sin 3\theta . \cos(5\theta/2) + \sin^2\theta \sin(5\theta/2)$$

divided by the square root of

$$\sin^2 3\theta . \cos^2(5\theta/2) + \sin^4\theta + \sin 3\theta . \sin 5\theta . \sin^2\theta .$$

This is from my rough notes, and I believe it is correct.[8] It is so nearly true, unless the angle be very obtuse, that common drawing, applied to the construction, will not detect the error. There are many formulae of this kind: and I have several times found a speculator who has discovered the corresponding construction, has seen the approximate success of his drawing—often as great as absolute truth could give in graphical practice,—and has then set about his demonstration, in which he always succeeds to his own content.

There is a trisection of which I have lost both cutting and reference: I think it is in the *United Service Journal.* I could not detect any error in it, though certain there must

---

[7] Nathaniel Wallich (1786-1854) was surgeon at the Danish settlement at Serampore when the East India Company took over the control in 1807. He entered the British medical service and was invalided to England in 1828. His *Plantae Asiaticae Rariores* (3 vols., London, 1830-1832) was recognized as a standard. He became vice-president of the Linnean Society, F. R. S., and fellow of the Royal Asiatic Society.

[8] But if $\theta = 90°$ this asserts that

$$\cos 30° = \frac{\sin 270° . \cos 225° + \sin^2 90° . \sin 225°}{\sqrt{\sin^2 270° . \cos^2 225° + \sin^4 90° + \sin 270° . \sin 450° . \sin^2 90°}},$$

or that

$$\tfrac{1}{2}\sqrt{3} = \frac{-1 . (-1/\sqrt{2}) + 1 . (-1/\sqrt{2})}{\sqrt{1 . \tfrac{1}{2} + 1 - 1 . 1 . 1}} = 0/\sqrt{\tfrac{1}{2}},$$

so that De Morgan must have made some error in copying.

be one. At least I discovered that two parts of the diagram were incompatible unless a certain point lay in line with two others, by which the angle to be trisected—and which was trisected—was bound to be either 0° or 180°.

Aug. 22, 1866. Mr. Upton sticks to his subject. He has just published "The Uptonian Trisection. Respectfully dedicated to the schoolmasters of the United Kingdom." It seems to be a new attempt. He takes no notice of the sentence I have put in italics: nor does he mention my notice of him, unless he means to include me among those by whom he has been "ridiculed and sneered at" or "branded as a brainless heretic." I did neither one nor the other: I thought Mr. Upton a paradoxer to whom it was likely to be worth while to propound the definite assertion now in italics; and Mr. Upton does not find it convenient to take issue on the point. He prefers general assertions about algebra. So long as he cannot meet algebra on the above question, he may issue as many "respectful challenges" to the mathematicians as he can find paper to write: he will meet with no attention.

There is one trisection which is of more importance than that of the angle. It is easy to get half the paper on which you write for margin; or a quarter; but very troublesome to get a third. Show us how, easily and certainly, to fold the paper into three, and you will be a real benefactor to society.

Early in the century there was a Turkish trisector of the angle, Hussein Effendi, who published two methods. He was the father of Ameen Bey, who was well known in England thirty years ago as a most amiable and cultivated gentleman and an excellent mathematician. He was then a student at Cambridge; and he died, years ago, in command of the army in Syria. Hussein Effendi was instructed in mathematics by Ingliz Selim Effendi, who translated a work of Bonnycastle[9] into Turkish.[10] This Englishman was Richard Baily, brother of Francis Baily[11] the astronomer, who emigrated to Turkey in his youth, and adopted the manners of the Turks, but whether their religion also I never heard, though I should suppose he did.

I now give the letters from the agricultural laborer and his friend, described on page 12, Vol. I. They are curiosities; and the history of the quadrature can never be well written without some specimens of this kind:

"Doctor Morgan, Sir. Permit me to address you

"Brute Creation may perhaps enjoy the faculty of beholding visible things with a more penitrating eye than ourselves. But Spiritual objects are as far out of their reach as though they had no being

"Nearest therefore to the brute Creation are those men who Suppose themselves to be so far governed by external objects as to believe nothing but what they See and feel And Can accomedate to their Shallow understanding and Imaginations

"My Dear Sir Let us all Consult ourselves by the wise proverb.

"I believe that evry man[s] merit & ability aught to be appreciated and valued In proportion to its worth & utility

"In whatever State or Circumstances they may fortunately or unfortunately be placed

"And happy it is for evry man to know his worth and place

"When a Gentleman of your Standing in Society Clad with those honors Can not understand or Solve a problem That is explicitly explained by words and Letters and mathematically operated by figuers He had best consult the wise proverd

---

[9] John Bonnycastle (died in 1821) was professor of mathematics at Woolwich. His edition of Bossut's *History of Mathematics* (1803), and his works on elementary mathematics were well known.

[10] The bibliographies give Husain Rifki as the translator, a practical geometry as the work, and 1802 as the date.

[11] See Vol. I, page 309, note 2.

"Do that which thou Canst understand and Comprehend for thy good.

"I would recommend that Such Gentleman Change his business

"And appropriate his time and attention to a Sunday School to Learn what he Could and keep the Litle Children form durting their Close

"With Sincere feelings of Gratitude for your weakness and Inability I am

"Sir your Superior in Mathematics ——"

"1849 June th29."

"Dor Morgin Sir

"I wrote and Sent my work to Professor —— of —— State of —— United States

"I am now in the possession of the facts that he highly approves of my work. And Says he will Insure me Reward in the States

"I write this that you may understand that I have knowledge of the unfair way that I am treated In my own nati County

"I am told and have reasons to believe that it is the Clergy that treat me so unjust.

"I am not Desirous of heaping Disonors upon my own nation. But if I have to Leave this kingdom without my Just dues. The world Shall know how I am and have been treated.

"I am Sir Desirous of my

"Just dues ——"

"1849 July 3."

"July 7th, 1849.

"Sir, I have been given to understand that a friend of mine one whom I shall never be ashamed to acknowledge as such tho' lowly his origine; nay not only not ashamed but proud of doing so for I am one of those who esteem and respect a man according to his ability and probity, deeming with Dr. Watts 'that the mind is the standard of the man,'[12] has laid before you and asked your opinion of his extraordinary performance, viz. the quadrature of the circle, he did this with the firmest belief that you would not only treat the matter in a straightforward manner but with the conviction that from your known or supposed knowledge of mathematicks would have given an upright and honorable decision upon the subject; but the question is have you done so? Could I say yes I would with the greatest of pleasure and have congratulated you upon your decision whatever it might have been but I am sorry to say that I cannot your letter is a paltry evasion, you say 'that it is a great pity that you (Mr. ——) should have attempted this (the quadrature of the circle) for your mathematical knowledge is not sufficient to make you know in what the problem consists,' you don't say in what it does consist *according to your ideas,* oh! no nothing of the sort, you enter into no disquisition upon the subject in order to show where you think Mr.—— is wrong and why you have not is simply—*because you cannot*—you know that he has done it and what is if I am not wrongly informed *you have been heard to say so.* He has done what you nor any other mathematician as those who call themselves such have done. And what is the reason that you will not candidly acknowledge to him as you have to others that he has squared the circle shall I tell you? it is because he has performed the feat to obtain the glory of which mathematicians have battled from time immemorial that they might encircle their brows with a wreath of laurels far more glorious than ever conqueror won it is simply this that it is a poor man a humble artisan who has gained that victory that you don't like to acknowledge it you don't like to be beaten and worse to acknowledge that you have miscalculated, you have in short too small a soul to acknowledge that he is right.

---

[12] Probably in *The Improvement of the Mind* which Isaac Watts (1674-1748) published in 1741. His *Horae Lyricae* appeared in 1706, and the *Hymns,* by which he is still well known, in 1707.

"I was asked my opinion and *I* gave it unhesitatingly in the affirmative and I am backed in my opinion not only by Mr. —— a mathematician and watchmaker residing in the boro of Southwark but by no less an authority than the Professor of mathematics of —— College —— —— United States Mr.—— and I presume that he at least is your equal as an authority and Mr. —— says that the government of the U. S. will recompense M. D. for the discovery he has made if so what a reflection upon Old england the boasted land of freedom the nursery of arts and sciences that her sons are obliged to go to a foreign country to obtain that recompense to which they are justly entitled

"In conclusion I had to contradict an assertion you made to the effect that 'there is not nor ever was any reward offered by the government of this country for the discovery of the quadrature of the circle.' I beg to inform you that there *was* but that it having been deemed an impossibility the government has withdrawn it. I do this upon no less an authority than the Marquis of Northampton.[13]

"I am, sir, yours ——"

"Dr. Morgan."

THE MOON'S ROTATION.

Notes on the Kinematic Effects of Revolution and Rotation, with reference to the Motions of the Moon and of the earth. By Henry Perigal, Jun. Esq. London, 1846-1849, 8vo.

On the misuse of technical terms. Ambiguity of the terms *Rotation* and *Revolution,* owing to the double meaning improperly

---

[13] Spencer Joshua Alwyne Compton, second Marquis of Northampton (1790-1851), was a poet, a scientist, and a statesman. He was president of the Royal Society from 1838 to 1849.

& ' (
. !# ! # !

! 3 $$! 8 $# ! $ #

+,-.+ + +/ ! + ' # 3#

4 3 $$! 8 $# ! $ #

+,-.+ + +/ ! + ' # 3#

3 $$! 8 $# ! $!33

+,-.+ + +/ ! + ' !33 3#

+,-.+ + +/ ! + '
D :# : : : !:J K !:L ! M :43)=) "7 3#

+,-.+ + +/ ! + '
D :# : : : !:J K !:L ! M :43)=%)* ' 3#

! # N ! ! 4

+,-.+ + +/ ! + '
. !# ! :# : :3! : :J K< ! : ! :43)= * %*8 3#

3 $$! 8 $# ! $! B ! !4

+,-.+ + +/ ! + ' ! B ! !4 3#

3 $$! 8 $# ! $4 4 !3 2 3
/4 0 4

+,-.+ + +/ ! + ' 4 4 !3 2 3 3#

3 $$! 8 $# ! $ ' % *)' ' *

+,-.+ + +/ ! + ' ' % *)' ' * 3#

3 $$! 8 $# ! $!# 4 # 8

+,-.+ + +/ ! + ' !# 4 # 8 3#

7 3 $$! 8 $# ! $!# 4 # 8

+,-.+ + +/ ! + ' !# 4 # 8 3#

= 3 $$! 8 $# ! $ B

+,-.+ + +/ ! + ' B 3#

3 $$! 8 $# ! $B !# !

+,-.+ + +/ ! + ' B !# ! 3#

3 $$! 8 $# ! $ ' % * "%*

+,-.+ + +/ ! + ' ' % * "%* 3#

3 $$! 8 $# ! $ B ! #

+,-.+ + +/ ! + ' % " * 3#

/4 1! ! . !# ! ? !

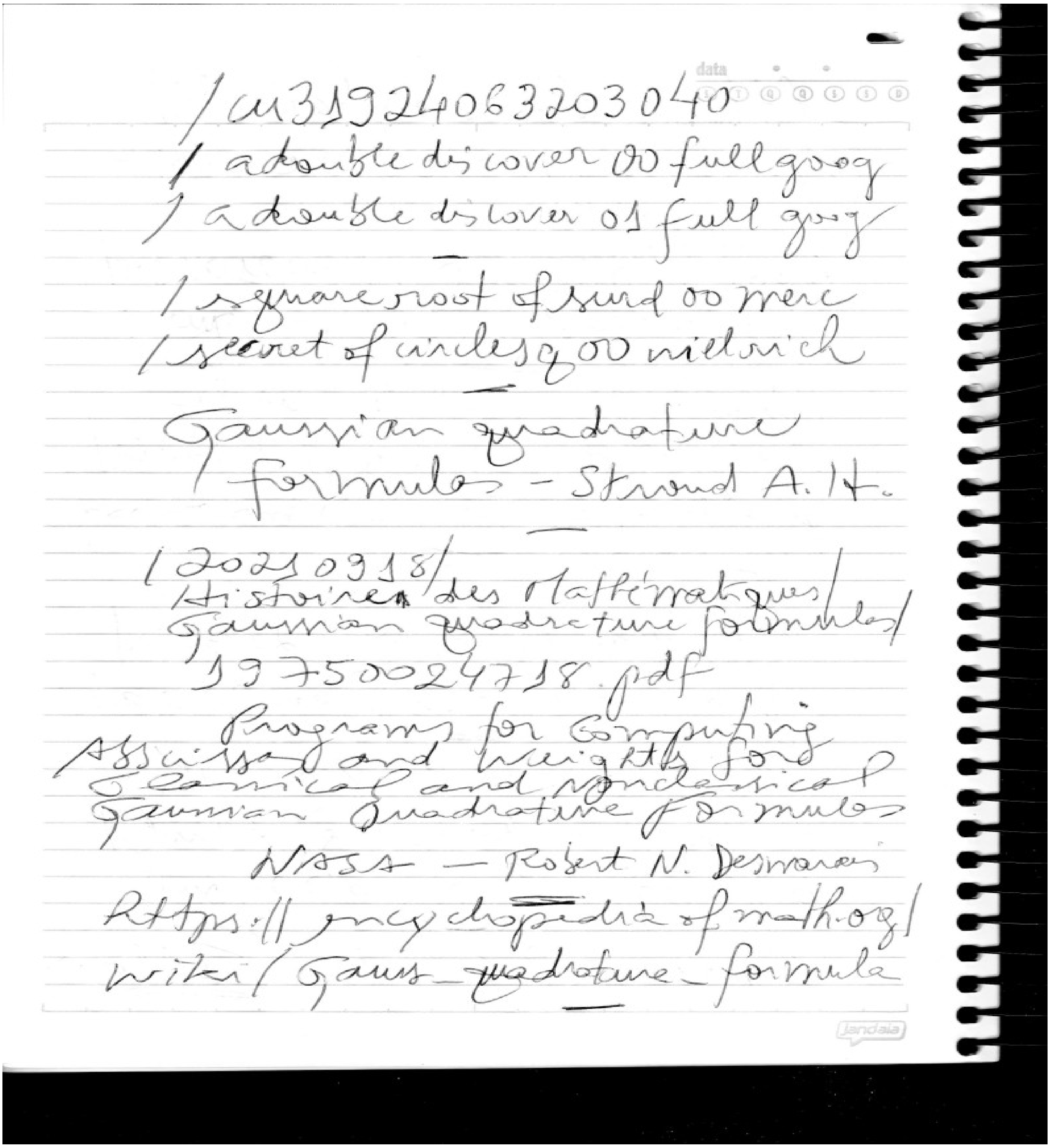

/m31924063203040
/ a double discover 00 full goog
/ a double discover 01 full goog

—

/ square root of surd 00 merc
/ secret of circles q 00 nielrich

—

Gaussian quadrature
/ formulas - Stroud A. H.

—

/20210918/
Histoires des Mathématiques/
Gaussian quadrature formulas/
197500024718.pdf
Programs for Computing
Abscissas and Weights for
Classical and nonclassical
Gaussian Quadrature Formulas
NASA — Robert N. Desmarais
https://encyclopediaofmath.org/
wiki/Gauss_quadrature_formula

—

https://ia804507.us.archive.org/
BookReader/BookReaderImages.
_tu00stro_0019.jp2

/24/items/

https://ia904507.us.archive.org/
24/items/gaussianquadratu00stro/

or

view-source: https://ia...

Disquisitiones Generales
circa
Superficies curvas

Carolo Friderico Gauss

Méthode Vectorielle de
Grassmann

Géométrie Infinitésimale

Points et de Surfaces
algébriques
Darboux.

( )

archive.org/details/
sureneclasserema00darbuoft-

recherchesgn00gaus

Lecons sur la théorie générale
des surfaces et les applications
géométriques du calcul
infinitésimal
4 volumes

https://archive.org/details/
leconssurlath01darbuoft
02

p3 leconssurla00darbuoft
p4

leconsurlagendes01darbüch
02
03
04

leconssurlathorie 01 darb.pdf
02
03
04

surlatheogensu 01 darbrich
04

Um lugar chamado Notting Hill

Books-019-2021/ 20210918/
/Histoire des Mathématiques/

coll. - Histoire de problèmes,
histoire des mathématiques...pdf

others.

Books-019-2021/ Important/

- coll. - Histoire de problemes,
  histoire des mathématiques...pdf
- Commissions inter-IREM -
  Histoires des mathématiques...
  (1993).djvu

Books-019-2021/
20210918/
Ellipse-Editora/

- Michel Carral - Géométrie...(1998).pdf
- Topology... (1998).pdf
- Des fleurs pour Schröndiger
  ... (2009).pdf
- Calcul différentiel ...(1997).pdf
- Convexité dans les espaces

'% ( )
! 3 $$! 8 $# ! $! ! #! 4

+,-.+ + +/ ! + '% ! ! #! 4 3#

4 3 $$! 8 $# ! $! ! #! 4

+,-.+ + +/ ! + '% ! ! #! 4 3#

3 $$! 8 $# ! $! !

+,-.+ + +/ ! + '% ! ! 3#

3 $$! 8 $# ! $! 8 !') ")

+,-.+ + +/ ! + '% ') ") 3# 3#

' 3 $$! 8 $# ! $4 4: 4:4O ,,,,, ,,L

+,-.+ + +/ ! + '% 4 4: 4:4O ,,,,, ,,L 3#

3 $$! 8 $# ! $! #! 4

+,-.+ + +/ ! + '% ! #! 4 3#

3 $$! 8 $# ! $! #! 4

+,-.+ + +/ ! + '% ! #! 4 3#

#' 3 $$! 8 $# ! $3* ! #! 4

+,-.+ + +/ ! + '% 3* ! #! 4 3#

#* 3 $$! 8 $# ! $! #! 4

+,-.+ + +/ ! + '% ! #! 4 3#

#" 3 $$! 8 $# ! $! #! 4

+,-.+ + +/ ! + '% ! #! 4 3#

#) 3 $$! 8 $# ! $! '#! 4

+,-.+ + +/ ! + '% ! '#! 4 3#

3 $$! 8 $# ! $! *#! 4

+,-.+ + +/ ! + '% ! *#! 4 3#

3 $$! 8 $# ! $! # #! 4

+,-.+ + +/ ! + '% ! # #! 4 3#

#% 3 $$! 8 $# ! $! # #! 4

+,-.+ + +/ ! + '% ! # #! 4 3#

3 $$! 8 $# ! $! # '#! 4

+,-.+ + +/ ! + '% ! # '#! 4 3#

3 $$! 8 $# ! $! # *#! 4

+,-.+ + +/ ! + '% ! # *#! 4 3#

3 $$! 8 $# ! $ 2 #! 4

+,-.+ + +/ ! + '% 2 #! 4 3#

' 3 $$! 8 $# ! $3 3 # #! 4

+,-.+ + +/ ! + '% 3 3 # #! 4 3#

'% (
! +,-.+ + +/ ! + '% D # 3 4 P
! ! B A 3 %%' 3#

4 +,-.+ + +/ ! + '% &-A D
3 4 F D # ! E ! B 3 %%' #78

4 +,-.+ + +/ ! + '% &-A D
3 4 F D # ! E ! B 3 %%' 3#

+,-.+ + +/ ! + '% ! ! 1E E
3 # ' I !8 A 3 ! = %% 3#

+,-.+ + +/ ! + '% 0 3 ! ! 2
QRSTUVWXVYZQX[QXV\V]^XQV_`aVXbXTQZQXUVXcVdaVeTRfYghXUVQRTTZijU\kYYZeUXUVI

+,-.+ + +/ ! + '% H 3 s# >!
! 8 E #5E !33 ! A 3 ! = % 3#

+,-.+ + +/ ! + '% ! # E I
E A 3 ! = %% 3#

+,-.+ + +/ ! + '% 8 I E #! 3!
0 E ! 3! I 3 A 3 ! = * 3#

fonctionnels ... (2004).pdf

- Alain Guichardet - Intégration-Analyse Hilbertienne ...(1908).pdf
- Espaces vectoriels ...(2000).pdf
- Ivan Gozard - Théorie de Galois-...(1997).pdf
- Dominique Bourn - Traité de Géométrie Affine ...(2012).pdf
- Géométrie Affine, Projective, Euclidienne et Anallagmatique ...(2003).pdf
- Aux origines du calcul infinitésimal... (1999).pdf
- La notion d'infini ...(2001).pdf

Books-019-2021/ Books-98-2020/

- Importants/
  - Dissertação Pappus.pdf
  - Tabak_J._geometry_the_language_of_space.pdf
- Pappus-20200822/
- Riemann/
- Supergravity/

+,-.+ + +/ ! + '% , ! 1 ! # & E !
, ! 2 D 4 A 3 ! = %% 3#

+,-.+ + +/ ! + '% A 3! 8 P !33 !
E! P ! # ! ! ! ! ! A 3 ! = 3#

7 +,-.+ + +/ ! + '% t8! 1 6! # 0 E # 1!
A 3 ! = %% 3#

= +,-.+ + +/ ! + '% H B < 0 ! E #
1E , A 3 ! = 3#

+,-.+ + +/ ! + '% 1E E , P 7 8 P
A # , ! ! ! B A 3 ! = ' 3#

+,-.+ + +/ ! + '% , I # ! E !
A 3 %%% 3#

+,-.+ + +/ ! + '% >! u #v >v&
! E ! B 2 F ! A 3 3#

+,-.+ + +/ ! + '% H !wx : ! C - /
! 3#

3 +,-.+ + +/ ! + '%
1 : 2!: ! ! ! :# 8 2 3#
/4 > 8 ! @@@

B +,-.+ + +/ ! + '% !33 3#

+,-.+ + +/ ! + '%
!4!=:7: 2: : ! ! : : 3! :! #: 3#

/

! <//; 6+<//; +< = % +< = % + !33
) +

4 +y& '+<//; 6+<//; +< = % +< = %
+- ! +

+y& '+<//; 6+<//; +< = % +< = %
+ 3 !8 2+

(%

/Supergravity/

- Supergravity 2012...(1).pdf
  Pág. 12(30) - 1.2.3
  Spacetime symmetries — the Lorentz and Poincaré groups

/Infinitesimal/

- Applications_d'analyse_et_de_géométrie... Poncelet...
  ...2139.pdf (VOL. I)
- Traitdesprop01poncuoft.pdf
- Traitdesprop02poncuoft.pdf
- A039a423.pdf }
- A039a424.pdf } →

Traité de mécanique industrielle
2 TOMOS
Poncelet

Books-019-2021/20210918/
Advanced Analytics with Excel 2019/
Advanced Analytics...pdf

(%
! +,-.+ + +/ ! + '% 3 !8 2 P ! 4 #
9 8 2 3#
(' ' 3! 2

It follows that finite group transformations compose as

$$\phi \xrightarrow{\theta_2} \phi' = U(\Theta_2)\phi \xrightarrow{\theta_1} \phi'' = U(\Theta_2)U(\Theta_1)\phi. \tag{1.23}$$

We show in Sec. 1.5 that this agrees with the composition of the unitary transformations which implement the symmetry in the quantum theory.

1.2.3 Spacetime symmetries – the Lorentz and Poincaré groups

The Lorentz group is defined as the group of homogeneous linear transformations of coordinates in D-dimensional Minkowski spacetime that preserve the Minkowski norm of any vector. We write

$$x^\mu = \Lambda^\mu{}_\nu x'^\nu \qquad \text{or} \qquad x'^\mu = \Lambda^{-1\mu}{}_\nu x^\nu\,, \tag{1.24}$$

and require that

$$x^\mu \eta_{\mu\nu} x^\nu = x'^\mu \eta_{\mu\nu} x'^\nu. \tag{1.25}$$

The Poincaré group is defined by adjoining global translations and considering

$$x'^\mu = \Lambda^{-1\mu}{}_\nu \left(x^\nu - a^\nu\right). \tag{1.26}$$

In this section we review the properties of these groups, their Lie algebras and the group action on fields such as $\phi^i(x)$.

If (1.25) holds for any vector x^μ, it follows that

$$\Lambda^\mu{}_\rho \eta_{\mu\nu} \Lambda^\nu{}_\sigma = \eta_{\rho\sigma}\,. \tag{1.27}$$

This is the defining property of the Λ matrices. If the Minkowski metric were replaced by the Kronecker delta, $\delta_{\mu\nu}$, these conditions would define the orthogonal group $\mathrm{O}(D)$, but here we have the pseudo-orthogonal group $\mathrm{O}(D-1,1)$. For most purposes we need only the connected component of this group, which we call the proper Lorentz group.

The metric tensor (and its inverse) are used to lower (or raise) vector indices. Thus one has, for example, $x_\mu = \eta_{\mu\rho} x^\rho$ and $\Lambda_{\mu\nu} = \eta_{\mu\rho}\Lambda^\rho{}_\nu$. Upper or lower indices are called contravariant or covariant, respectively.

Exercise 1.3 *Show that (1.24) and (1.27) imply*

$$\begin{aligned} \Lambda_{\mu\nu} &= (\Lambda^{-1})_{\nu\mu}\,, \qquad \Lambda^\mu{}_\nu = (\Lambda^{-1})_\nu{}^\mu\,, \\ x'_\mu &= (\Lambda^{-1})_\mu{}^\nu x_\nu = x_\nu \Lambda^\nu{}_\mu. \end{aligned} \tag{1.28}$$

The first relation of the exercise resembles the standard matrix orthogonality property, but it holds for Lorentz when both indices are down (or both up), which is not the correct position for their action as linear transformations. Indeed, Lorentz matrices must be multiplied with indices in up–down position, viz. $\Lambda^\mu{}_\rho \Lambda'^\rho{}_\nu$.

We now introduce the Lie algebra of the Lorentz group, proceeding in parallel[7] to the discussion of the group SO(n) in Sec. 1.2.1. For a small parameter ϵ, we write the infinitesimal transformation

$$\Lambda^{\mu}{}_{\nu} = \delta^{\mu}_{\nu} + \epsilon m^{\mu}{}_{\nu} + \cdots . \tag{1.29}$$

It is straightforward to see that (1.29) satisfies (1.27) to first order in ϵ as long as the generator (m with two lower indices) is antisymmetric, viz.

$$m_{\mu\nu} \equiv \eta_{\mu\rho} m^{\rho}{}_{\nu} = -m_{\nu\mu} . \tag{1.30}$$

The *Lie algebra* is the real linear space spanned by the $\frac{1}{2}D(D-1)$ independent generators, with the commutator product $[m, m'] = m\,m' - m'm$. These matrices must also be multiplied as $m^{\mu}{}_{\rho} m'^{\rho}{}_{\nu}$, but the forms with both indices down, as in (1.30), (or both up) are often convenient.

A useful basis for the Lie algebra is to choose generators that act in each of the $\frac{1}{2}D(D-1)$ coordinate 2-planes. For the 2-plane in the directions ρ, σ this generator is given by

$$m_{[\rho\sigma]}{}^{\mu}{}_{\nu} \equiv \delta^{\mu}_{\rho}\eta_{\nu\sigma} - \delta^{\mu}_{\sigma}\eta_{\rho\nu} \;= -m_{[\sigma\rho]}{}^{\mu}{}_{\nu} . \tag{1.31}$$

Note the distinction between the coordinate plane labels in brackets and the row and column indices. In this basis, a finite proper Lorentz transformation is specified by a set of $\frac{1}{2}D(D-1)$ real parameters $\lambda^{\rho\sigma} = -\lambda^{\sigma\rho}$ and takes exponential form

$$\Lambda = \mathrm{e}^{\frac{1}{2}\lambda^{\rho\sigma} m_{[\rho\sigma]}} . \tag{1.32}$$

When matrix indices are restored, we have, with the representation (1.31), the series[8]

$$\Lambda^{\mu}{}_{\nu} = \delta^{\mu}{}_{\nu} + \lambda^{\mu}{}_{\nu} + \tfrac{1}{2}\lambda^{\mu}{}_{\rho}\lambda^{\rho}{}_{\nu} + \ldots . \tag{1.33}$$

The commutators of the generators defined in (1.31) are

$$[m_{[\mu\nu]}, m_{[\rho\sigma]}] = \eta_{\nu\rho} m_{[\mu\sigma]} - \eta_{\mu\rho} m_{[\nu\sigma]} - \eta_{\nu\sigma} m_{[\mu\rho]} + \eta_{\mu\sigma} m_{[\nu\rho]} . \tag{1.34}$$

These equations specify the structure constants of the Lie algebra, which may be written as

$$f_{[\mu\nu][\rho\sigma]}{}^{[\kappa\tau]} = 8\eta_{[\rho[\nu}\delta^{[\kappa}_{\mu]}\delta^{\tau]}_{\sigma]} . \tag{1.35}$$

Note that antisymmetrization is always done with 'weight 1', see (A.8), such that the right-hand side can be written as eight terms with coefficients ±1.

Exercise 1.4 *Check that (1.34) leads to (1.35). To compare with (1.13), you have to replace each of the indices A, B, C by antisymmetric combinations, e.g. $A \to [\mu\nu]$. Moreover, you have to insert a factor $\frac{1}{2}$ each time that you sum over such a combined index to avoid double counting, as e.g. the factor $\frac{1}{2}$ in (1.32). Therefore we rewrite (1.34) as*

---

[7] Specifically, it is Λ^{-1} which is the analogue of R in (1.12) and of $U(\theta)$ in (1.15).

[8] Thus, we now replace in (1.29) $\epsilon m^{\mu}{}_{\nu}$ by $\frac{1}{2}\lambda^{\rho\sigma} m_{[\rho\sigma]}{}^{\mu}{}_{\nu}$.

$$[m_A, m_B] = f_{AB}{}^C m_C \to [m_{[\mu\nu]}, m_{[\rho\sigma]}] = \tfrac{1}{2} f_{[\mu\nu][\rho\sigma]}{}^{[\kappa\tau]} m_{[\kappa\tau]}. \tag{1.36}$$

Under a symmetry, each group element is mapped to a transformation of the configuration space of the dynamical fields. This map must give a group homomorphism. For the Lorentz matrix Λ, the transformation of the scalar fields is defined as

$$\phi^i(x) \xrightarrow{\Lambda} \phi'^i(x) \equiv \phi^i(\Lambda x). \tag{1.37}$$

Using (1.24), we find that $\phi'^i(x') = \phi^i(x)$.

Exercise 1.5 *Show that the action (1.4) is invariant under the transformation (1.37).*

We now define differential operators which implement the coordinate change due to an infinitesimal transformation. A transformation in the $[\rho\sigma]$ 2-plane is generated by

$$L_{[\rho\sigma]} \equiv x_\rho \partial_\sigma - x_\sigma \partial_\rho. \tag{1.38}$$

The commutator algebra of these operators is isomorphic to (1.34), and we thus have a realization of the Lie algebra acting as differential operators on functions.

Exercise 1.6 *Compute the commutators $[L_{[\mu\nu]}, L_{[\rho\sigma]}]$ and show that they agree with that of (1.34) for matrix generators. Show that to first order in $\lambda^{\rho\sigma}$*

$$\phi^i(x^\mu) - \tfrac{1}{2}\lambda^{\rho\sigma} L_{[\rho\sigma]}\phi^i(x^\mu) = \phi^i(x^\mu + \lambda^{\mu\nu}x_\nu). \tag{1.39}$$

We then define the differential operator

$$U(\Lambda) \equiv \mathrm{e}^{-\frac{1}{2}\lambda^{\rho\sigma} L_{[\rho\sigma]}}. \tag{1.40}$$

Using this operator, the mapping (1.37) which defines the action of finite Lorentz transformations on scalar fields can then be written as

$$\phi^i(x) \to \phi'^i(x) = U(\Lambda)\phi^i(x) = \phi^i(\Lambda x). \tag{1.41}$$

Box 1.1 **Symmetries**

Symmetries are implemented by transformations that act on fields.

We see that both Lorentz and internal symmetries (see (1.16)) are implemented by linear operators acting on the classical fields, a differential operator for Lorentz and a matrix operator for internal. Both operators depend on group parameters in the same way.

By expanding the exponential, this defines the infinitesimal Lorentz transformation of the scalars:

$$\delta(\lambda)\phi^i(x) = \phi'^i(x) - \phi^i(x) = \phi^i(\Lambda x) - \phi^i(x) = \lambda^\mu{}_\nu x^\nu \partial_\mu \phi^i(x) = -\tfrac{1}{2}\lambda^{\mu\nu} L_{[\mu\nu]}\phi^i(x). \tag{1.42}$$

This definition does give a homomorphism of the group; namely the product of maps, first with Λ_2, then with Λ_1, produces the compound map associated with the product $\Lambda_1\Lambda_2$. This can be seen from the sequence of steps

$$\begin{aligned}\phi^i(x) \xrightarrow{\Lambda_2} \phi'^i(x) = U(\Lambda_2)\phi^i(x) \xrightarrow{\Lambda_1} \phi''^i(x) &= U(\Lambda_2)U(\Lambda_1)\phi^i(x)\\ &= U(\Lambda_2)\phi^i(\Lambda_1 x)\\ &= \phi^i(\Lambda_1\Lambda_2 x).\end{aligned} \tag{1.43}$$

As mentioned above, a symmetry transformation acts *directly on the fields*. This convention determines the order of operations in the first line. In the second and third lines, we use the action of the differential operators to arrive at a transformation with matrix product $\Lambda_1\Lambda_2$. This is exactly the same for internal symmetries in (1.23), where the action of U is obtained by matrix multiplication.

Exercise 1.7 *It is instructive to check (1.43) for Lorentz transformations which are close to the identity. Specifically, use the definition (1.40) to show that the order $\lambda_1\lambda_2$ terms in the product $U(\Lambda_2)U(\Lambda_1)\phi^i$ of differential operators acting on ϕ^i agrees with terms of the same order in $\phi^i(\Lambda_1\Lambda_2 x)$.*

Calculate the infinitesimal commutator $[U(\Lambda_2), U(\Lambda_1)]$ to order $\lambda_1\lambda_2$. Show that the commutator is a Lorentz transformation with (matrix) parameters $[\lambda_1, \lambda_2]$. Note that the second transformation acts on $\phi^i(x)$, and not on the x-dependent factor in (1.42).

It is important to extend the Lorentz transformation rules to covariant and contravariant vector fields, which are, respectively, sections of the cotangent and tangent bundles of Minkowski spacetime. The transformation of a general covariant vector field $W_\mu(x)$ can be modeled on that of the gradient of a scalar $\partial_\mu\phi(x)$. From (1.37) we find

$$\partial_\mu\phi(x) \to \partial_\mu\phi'(x) = \frac{\partial}{\partial x^\mu}\phi(\Lambda x) = \Lambda^{-1}{}_\mu{}^\nu(\partial_\nu\phi)(\Lambda x)\,, \tag{1.44}$$

where we have used the chain rule and (1.28) in the last step. Thus we define the transformation of a general covariant field as

$$W_\mu(x) \to W'_\mu(x) \equiv \Lambda^{-1}{}_\mu{}^\nu W_\nu(\Lambda x). \tag{1.45}$$

For contravariant vectors we assume a transformation of the form

$$V^\mu(x) \to B^\mu{}_\sigma V^\sigma(\Lambda x). \tag{1.46}$$

The matrix can determined by requiring that the inner product $V^\mu(x)W_\mu(x)$ transforms as a scalar. This fixes $B^\mu{}_\sigma = \Lambda^{-1\mu}{}_\sigma$, and we have the transformation rule

$$V^\mu(x) \to V'^\mu(x) \equiv \Lambda^{-1\mu}{}_\nu V^\nu(\Lambda x). \tag{1.47}$$

Next we define generators of Lorentz transformations appropriate to the covariant and contravariant vector representations. They are combined differential/matrix operators given by

$$\begin{aligned} J_{[\rho\sigma]}V^\mu(x) &\equiv (L_{[\rho\sigma]}\delta^\mu{}_\nu + m_{[\rho\sigma]}{}^\mu{}_\nu)V^\nu(x)\,, \\ J_{[\rho\sigma]}W_\nu(x) &\equiv (L_{[\rho\sigma]}\delta_\nu{}^\mu + m_{[\rho\sigma]\nu}{}^\mu)W_\mu(x). \end{aligned} \tag{1.48}$$

We avoid an overly decorative notation by suppressing row and column indices on $J_{[\rho\sigma]}$. For each case a finite Lorentz transformation is implemented by the operator

$$U(\Lambda) = \mathrm{e}^{-\frac{1}{2}\lambda^{\rho\sigma}J_{[\rho\sigma]}}. \tag{1.49}$$

Exercise 1.8 *Verify that* $J_{[\rho\sigma]}(V^\mu W_\mu) = L_{[\rho\sigma]}(V^\mu W_\mu)$.

The Lorentz group has many representations, both higher rank tensor and spinor representations (to be discussed in Ch. 2) and combinations thereof. Let $\psi^i(x)$ denote a set of fields where i is now an index of the components of a general representation. There is a corresponding Lie algebra representation with matrices $m_{[\rho\sigma]}$ which act on the indices and differential/matrix generators

$$J_{[\rho\sigma]} = L_{[\rho\sigma]}\mathbb{1} + m_{[\rho\sigma]}\,, \tag{1.50}$$

in which $\mathbb{1}^i{}_j = \delta^i_j$ is the unit matrix. A finite Lorentz transformation is then the mapping (suppressing the index i for simplicity of the notation)

$$\psi(x) \to \psi'(x) = U(\Lambda)\psi(x) = \mathrm{e}^{-\frac{1}{2}\lambda^{\rho\sigma}m_{[\rho\sigma]}}\psi(\Lambda x). \tag{1.51}$$

The precise forms of the operators $m_{[\rho\sigma]}$, $J_{[\rho\sigma]}$ and $U(\Lambda)$ depend on the representation under study.

Spacetime translations $x^\mu \to x'^\mu = x^\mu - a^\mu$ are much simpler because they are implemented in the same way in all representations of the Lorentz group, namely by the mapping

$$\psi^i(x) \to \psi'^i(x) = \psi^i(x+a) = U(a)\psi^i(x)\,, \tag{1.52}$$

$$U(a) = \mathrm{e}^{a^\mu P_\mu}\,, \tag{1.53}$$

$$P_\mu = \partial_\mu = \frac{\partial}{\partial x^\mu}. \tag{1.54}$$

In (1.52) we have defined the generator P_μ and the finite translation operator $U(a)$ which are differential operators. Finite transformations of the Poincaré group are implemented by the operator $U(a, \Lambda) \equiv U(\Lambda)U(a)$, which acts as follows:

$$\begin{aligned}\psi(x) \;\rightarrow\; \psi'(x) &\equiv U(a,\Lambda)\psi(x) = U(\Lambda)U(a)\psi(x)\\ &= \mathrm{e}^{-\frac{1}{2}\lambda^{\rho\sigma}m_{[\rho\sigma]}}\psi(\Lambda x + a).\end{aligned} \tag{1.55}$$

Exercise 1.9 *Prove that $U((\Lambda)^{-1}a)U(\Lambda) = U(\Lambda)U(a)$. Verify for operators which are close to the identity that*

$$\begin{aligned}U(a)\phi(\Lambda' x + b) &= \phi(\Lambda' x + \Lambda' a + b)\,,\\ U(\Lambda)\phi(\Lambda' x + b) &= \phi(\Lambda'\Lambda x + b).\end{aligned} \tag{1.56}$$

The Lie algebra of the Poincaré group contains the $D(D+1)/2$ generators $J_{[\mu\nu]}$, P_μ. The following commutation rules complete the specification of the Lie algebra:

$$\begin{aligned}[][J_{[\mu\nu]}, J_{[\rho\sigma]}] &= \eta_{\nu\rho}J_{[\mu\sigma]} - \eta_{\mu\rho}J_{[\nu\sigma]} - \eta_{\nu\sigma}J_{[\mu\rho]} + \eta_{\mu\sigma}J_{[\nu\rho]}\,,\\ [J_{[\rho\sigma]}, P_\mu] &= P_\rho\eta_{\sigma\mu} - P_\sigma\eta_{\rho\mu}\,,\\ [P_\mu, P_\nu] &= 0.\end{aligned} \tag{1.57}$$

Exercise 1.10 *Verify (1.57).*

We now discuss the implementation of this Lie algebra on fields. The treatment is parallel to the discussion of internal symmetry at the end of Sec. 1.2.2. The infinitesimal variation of the fields $\psi^i(x)$ is defined as the first order truncation of the exponential in (1.55):

$$\delta\psi = [a^\mu P_\mu - \tfrac{1}{2}\lambda^{\rho\sigma}J_{[\rho\sigma]}]\psi. \tag{1.58}$$

The $J_{[\rho\sigma]}$ operator appropriate to the representation is used, and representation indices are suppressed for simplicity. Consider now the transformation δ_2 with group parameters $(a_2,\,\lambda_2)$ followed by transformation δ_1 with parameters $(a_1,\,\lambda_1)$. The result is defined as

$$\begin{aligned}\delta_1\delta_2\psi &= \delta_1[a_2^\mu P_\mu - \tfrac{1}{2}\lambda_2^{\rho\sigma}J_{[\rho\sigma]}]\psi\\ &= [a_2^\mu P_\mu - \tfrac{1}{2}\lambda_2^{\rho\sigma}J_{[\rho\sigma]}][a_1^\nu P_\nu - \tfrac{1}{2}\lambda_1^{\kappa\tau}J_{[\kappa\tau]}]\psi.\end{aligned} \tag{1.59}$$

The second transformation acts only on the field variable. With some care one can calculate the commutator of two variations, which yields a third transformation with parameters $(a_3,\,\lambda_3)$, that is

$$\begin{aligned}[][\delta_1,\,\delta_2]\psi &= \delta_3\psi\,,\\ \lambda_3^{\rho\sigma} &= [\lambda_1,\,\lambda_2]^{\rho\sigma} = \tfrac{1}{4}f_{[\mu\nu][\kappa\tau]}{}^{[\rho\sigma]}\lambda_1^{\mu\nu}\lambda_2^{\kappa\tau}\,,\\ a_3^\mu &= a_{1\rho}\lambda_2^{\rho\mu} - a_{2\rho}\lambda_1^{\rho\mu} = \tfrac{1}{2}f_{\nu[\rho\sigma]}{}^{\mu}\left(a_1^\nu\lambda_2^{\rho\sigma} - a_2^\nu\lambda_1^{\rho\sigma}\right)\end{aligned} \tag{1.60}$$

(remember Ex. 1.4). Note that the parameters of infinitesimal transformations compose with the structure constants of the Lie algebra, exactly as in the case of internal symmetry.

Exercise 1.11 *Verify the rule of combination of parameters in (1.60).*

As in the discussion of internal symmetry at the end of Sec. 1.2.2, the product of two finite transformations of the Poincaré group, the first with parameters $(a_2,\ \lambda_2)$ followed by the second with parameters $(a_1,\ \lambda_1)$, is given by

$$\psi(x) \to \psi''(x) = U(a_2,\ \lambda_2)U(a_1,\ \lambda_1)\psi(x). \tag{1.61}$$

The compound map is a representation, as we discuss at the end of Sec. 1.5 below.

1.3 Noether currents and charges

It is well known that, using the Noether method, one can construct a conserved current for every continuous global symmetry of the action of a classical field theory. Integrals of the time component of the currents are conserved charges, and an infinitesimal symmetry transformation is implemented on fields by the Poisson bracket of charge and field. Poisson brackets and the correspondence principle provide a useful bridge to the quantum theory in which finite group transformations are implemented through unitary transformations. In this section and the next we review this formalism in order to show that the conventions for symmetries used are compatible with readers' previous study of symmetries in quantum field theory. The Noether formalism has other important applications in supersymmetry and supergravity which enter in later chapters of this book.

We assume that we are dealing with a system of scalar fields $\phi^i(x)$, $i = 1, \ldots, n$, whose Lagrangian density is a function of the fields and their first derivatives, as described by the action (1.17). The Euler–Lagrange equations of motion are

$$\frac{\partial}{\partial x^\mu}\frac{\delta \mathcal{L}}{\delta \partial_\mu \phi^i(x)} - \frac{\delta \mathcal{L}}{\delta \phi^i(x)} = 0. \tag{1.62}$$

We define a generic infinitesimal symmetry variation of the fields by

$$\delta\phi^i(x) \equiv \epsilon^A \Delta_A \phi^i(x)\,, \tag{1.63}$$

in which the ϵ^A are constant parameters. This formula includes as special cases the various internal and spacetime symmetries discussed in Sec. 1.2. In these cases each $\Delta_A\phi^i(x)$ is linear in ϕ^i and given by a matrix or differential operator applied to ϕ^i. Specifically,

$$\text{internal} \quad \epsilon^A \Delta_A \phi^i \to -\theta^A (t_A)^i{}_j \phi^j\,, \tag{1.64}$$

$$\text{spacetime} \quad \epsilon^A \Delta_A \phi^i \to \left[a^\mu \partial_\mu - \tfrac{1}{2}\lambda^{\rho\sigma}(x_\rho \partial_\sigma - x_\sigma \partial_\rho)\right]\phi^i. \tag{1.65}$$

If (1.63) is a symmetry of the theory, then the action is invariant, and the variation of the Lagrangian density is an explicit total derivative, i.e. $\delta\mathcal{L} = \epsilon^A \partial_\mu K^\mu_A$. This must hold for *all* field configurations, not merely those which satisfy the equations of motion (1.62). In detail, the variation of the Lagrangian density is

4 <//; 6+<//; +< = % +< = %
+& ! +

4 +,-.+ + +/ ! + '%
,33 ! :#5! ! 2 : :# : E E :J K :L ! :43)=%))) 3#

4 +,-.+ + +/ ! + '%
,33 ! :#5! ! 2 : :# : E E :J K :L ! :43)=% 'B 3#

3 $$! ! 4 $! = $ * $43)=)") "' G
/4 AI # 1E E

+,-.+ + +/ ! + '%
:# : ! E ! B :E E ! :AI :J K:43)=)") "' 3#

3 $$! ! 4 $! = $ * $43)=)*"*"%"' G

+,-.+ + +/ ! + '%
AI :# : E E ::43)=)*"*"%"' 3#

+,-.+ + +/ ! + '%
,#8! #:, ! 2 : :AI : %: :H! !:, ! 2 :9 :AI : : : 3 ! :?
! 3#

(%

/Supergravity/
- Supergravity 2012...(1).pdf
Pág. 12(30) - 1.2.3
Spacetime symmetries -
the Lorentz and Poincaré
groups

/Infinitesimal/
- Applications_d'analyse_et_de_
géométrie ... Poncelet...
... 2139.pdf (HP.I)
- Traitdesprop01ponc uoft.pdf
- Traitdesprop02poncuoft.pdf
- A039a423.pdf }
- A039a424.pdf } →
Traité de mécanique
industrielle
2 TOMOS
Poncelet

Books-019-2021/20210918/
Advanced Analytics with Excel 2019/
Advanced Analytics...pdf

* (
! +,-.+ + +/ ! + * u # ! 1 2 A 8 &
%)* 3#
(" ! !

CHAPTER **3**

From the History of the Parallel Postulate

The history of the axiom of parallels affords the best approach to noneuclidean geometry. Familiarity with the Hilbert system of axioms will make it easier for us to understand the problem. However, we must begin with Euclid. Euclid's fifth postulate states:

(P) If a straight line falling on two straight lines makes the interior angles on the same side together less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which the angles are together less than two right angles."

The word "parallel" is defined by Euclid as follows (def. 23):

> Parallel straight lines are straight lines which, being in the same plane and being produced indefinitely in both directions, do not meet one another in either direction.

For centuries mathematicians were convinced that the fifth postulate could be deduced from the others. If so, it should not, strictly speaking, be included among the postulates.

There were arguments which supported this point of view. Thus it was well known that the converse of the parallel postulate was a provable proposition. This converse can be formulated as follows:

(C) Consider a line falling on two intersecting lines. Then the sum of exactly one pair of interior angles on the same side is less than two right angles.

Proposition (C), i.e., the converse of the parallel postulate, is equivalent to Theorem 8 of the preceding chapter. Indeed, if Theorem 8 were false, then we could have $\alpha_1 \equiv \beta$ (cf. Fig. 14) or, what amounts to the same thing, $\alpha + \beta = 2R$ and g_1 could intersect g_2. But this would contradict (C) which asserts that if g_1 and g_2 intersect, then $\alpha + \beta < 2R$. Thus (C) implies Theorem 8. It is easy to prove that Theorem 8 implies (C).

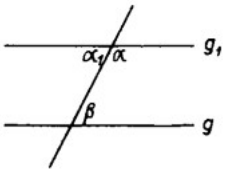

FIG. 14.

While mathematicians of past ages were not familiar with Hilbert's system, they could deduce Theorem 8 or (C) from Euclid's axioms and postulates without the use of the parallel postulate. (In proving Theorem 8 we did not make use of Axiom V.)

It was thought impossible for a geometric theorem not to be provable if its converse was provable. Indeed, in the case of most theorems of elementary geometry proving the converse of a theorem does not require the use of additional axioms. For example, the theorem on the congruence of base angles in an isosceles triangle and its converse (if two angles in a triangle are congruent then the sides opposite to these angles are congruent) are both provable by means of the congruence axioms. It is therefore understandable why mathematicians tried so eagerly for 2000 years to prove the axiom of parallels. In this they failed.

In many cases the "proofs" involved the use of an intuitive fact equivalent to the parallel postulate. For example, it is easy to prove the parallel postulate if we take it for granted that the sum of the angles in a triangle is equal to two right angles.

Other attempts at proof made use of the assertion that the length of the perpendicular from a point on a line to a parallel line is independent of that point. However, parallel lines are defined as nonintersecting lines (cf. above) and this does not automatically mean that the lengths in question are equal.

Of course, it is possible to prove this last assertion if we make use of the theorem on the alternate interior angles for parallel lines. But we would then be guilty of circular reasoning, since the theorem just mentioned cannot be proved without the use of the parallel postulate.

In 1763 G. S. Kluegel, a student of Kestner, wrote a dissertation [3] in which he brought together and criticized all significant attempts to prove the parallel postulate contributed by mathematicians over the 2000 years between the publication of the "Elements" and his own time. He found, correctly, that all 28 "proofs" were false.

Most of these attempts are of historical interest only. Some, however, pointed the way to new insights.

The first significant contribution is due to the Jesuit priest Saccheri (1667–1733). In his extensive work on parallels Saccheri studied a quadrilateral *ABCD* in which $AC \equiv BD$ and the angles at *A* and *B* are right angles (Fig. 15).

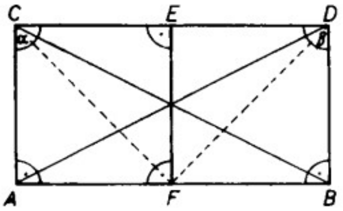

FIG. 15.

If we do not wish to make use of the parallel postulate, then it is not possible to prove that the angles *ACD* and *BCD* are right angles, i.e., that *ABDC* is a rectangle. However, we can use congruence considerations to prove that $\angle ACD \equiv \angle BDC$; namely,

$$\triangle ABC \equiv \triangle ABD \qquad \text{(sas)},$$

so that

$$AD \equiv BC.$$

Congruence of the diagonals implies

$$\triangle ACD \equiv \triangle BCD \qquad \text{(sss)}.$$

Hence

$$\alpha \equiv \beta.$$

Now we ask: Are the congruent angles α and β in the Saccheri quadrilateral acute, obtuse, or right angles?

First we establish a connection between this question and the sum of the angles in a triangle.

Let ABC (Fig. 16) be a triangle; let D, E be the midpoints of AC and BC; let F, G, H be the feet of the perpendiculars from C, A, B on to the line DE. Then

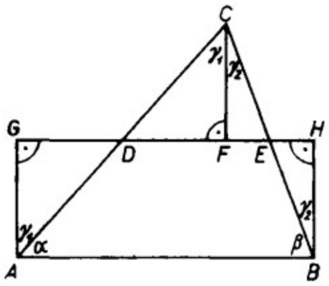

FIG. 16.

$$FC \equiv AG \equiv BH.$$

This follows from the congruence of the triangles DFC and ADG (saa) and the congruence of the triangles EFC and BEH.

The quadrilateral $ABHG$ is thus a Saccheri quadrilateral with right angles at G and H.

It follows that the the angles at A and B are congruent:

$$\alpha + \gamma_1 \equiv \beta + \gamma_2.$$

Since

$$\alpha + \beta + (\gamma_1 + \gamma_2) \equiv W$$

is the sum of the angles in the triangle ABC, it follows that each of the angles at A and B is equal to $W/2$.

"The hypothesis of the acute angle"—$W/2 < R$—implies that the sum of the angles in a triangle is $<2R$.

"The hypothesis of the obtuse angle"—$W/2 > R$—implies that the sum of the angles in a triangle is $>2R$.

"The hypothesis of the right angle"—$W/2 \equiv R$—implies that the sum of the angles in a triangle is $\equiv 2R$.

Saccheri attempted to eliminate the first two hypotheses by showing that they led to contradictions. If proved, this would imply that the sum of the angles in a triangle is equal to two right angles. But this would imply (cf. Theorem 12, below) the parallel postulate.

Saccheri succeeded in eliminating "the hypothesis of the obtuse angle." His proof is clumsy but correct. On the other hand, his proof of the inadmissibility of "the hypothesis of the acute angle" is not acceptable.

To sum up: Saccheri succeeded in showing that the sum of the angles in a triangle does not exceed two right angles.

Saccheri failed to show that the assumption

$$\alpha + \beta + \gamma < 2R$$

leads to a contradiction.

We shall see that there are deep reasons for these outcomes.

Instead of reproducing the clumsy argument of Saccheri we present Legendre's (1752–1833) version of the theorem on the sum of the angles in a triangle.

Thus, let α, β, γ be the angles of a given triangle ABC. Lay off the segment AB n times in succession on a line (Fig. 17) and construct over each of these segments a triangle $A_vA_{v+1}B_v$ congruent to the given triangle (all of the contructed triangles are to lie in the same plane).

We claim that $\gamma_1 \geqq \gamma$. If we have $\gamma_1 < \gamma$, then

$$A_v A_{v+1} > B_v B_{v+1}. \tag{1}$$

This conclusion can easily be deduced from the axioms of congruence. Further, the polygonal line $A_1B_1 + B_1B_2 + \cdots + B_nA_{n+1}$ is greater than the segment A_1A_{n+1}:

$$\begin{aligned} A_1B_1 + n \cdot B_1B_2 + B_nA_{n+1} &> n \cdot A_1A_2, \\ A_1B_1 + B_1A_2 &> n \cdot (A_1A_2 - B_1B_2). \end{aligned} \tag{2}$$

In view of (1) the difference $A_1A_2 - B_1B_2$ is equal to some segment ε. Hence (2) can be written in the form

$$n \cdot \varepsilon < A_1B_1 + B_1A_2. \tag{3}$$

Since n can be taken arbitrarily large, (3) contradicts Axiom IV, 1 (axiom of Archimedes).

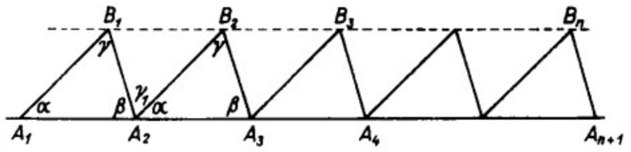

FIG. 17.

The assumption $\gamma_1 < \gamma$ is thus ruled out. But $\gamma_1 \geqslant \gamma$ implies $\alpha + \beta + \gamma \leqslant \alpha + \beta + \gamma_1 \equiv 2R$. This proves:

Theorem 11. The sum of the angles in a triangle is at most equal to two right angles.

Note that Theorem 11 was proved using the axioms in group I–III and the first continuity axiom (axiom of Archimedes) in group IV.

There is a very close connection between the sum of the angles in a triangle and the axiom of parallels. As noted before one can prove that:

Theorem 12. If the sum of the angles in a triangle is equal to two right angles, then the axiom of parallels holds.

We give a brief proof of this assertion.

Let A be the given point and g the given line (Fig. 18); B the foot of the perpendicular from A to g; C another point on g; h the perpendicular to AB at A. By Theorem 5, g and h cannot intersect. We shall show presently that h is the only parallel to g at A.

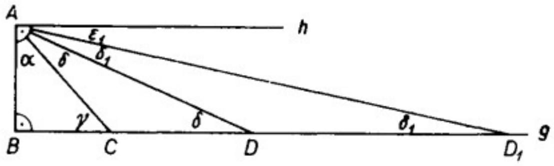

FIG. 18.

Thus let α and γ be the angles in the triangle ABC with vertices A and C. Then, in view of our assumption,†

$$\alpha + \gamma = R.$$

Choose D on g so that $AC \equiv CD$ and so that C lies between B and D. Then $\triangle ACD$ is isosceles and

$$\delta = \angle ADC = \gamma/2.$$

Again, let D_1 be a point on the extension of BD beyond D such that $AD \equiv DD_1$. Then

$$\delta_1 = \angle AD_1B = \delta/2 = \gamma/4.$$

Since $\angle D_1AB = R - \delta_1$, it follows that $\varepsilon_1 = \angle(h, AD_1) = \delta_1 = \gamma/4$.

By continuing this procedure we obtain segments AD_2, AD_3, ... which form with D_2B, D_3B, ... angles $\gamma/8$, $\gamma/16$, ... The corresponding angles $\varepsilon_2, \varepsilon_3, \ldots$ are likewise equal to $\gamma/8$, $\gamma/16$, ...

Now assume that $k \neq h$ is a ray emanating from A which does not intersect g. Let φ be the angle between k and h. If n is large enough, then the angle $\varepsilon_n = \gamma/2^{n+1}$ between the ray AD_n

---

† We shall sometimes refer to congruent angles as "equal". Equal angles have the same radian measure. Thus $\varphi = \psi$ is an equality between numbers.

and h will be smaller than φ. But then the ray k will lie in the interior of the angle BAD_n (Fig. 19) and, by the "triangle axiom" II, 4, will intersect BD_n and therefore the line g. A similar argument holds for the halfplane determined by AB and not containing the point C. Thus h is seen to be the only line through A parallel to g. This proves the parallel postulate under the assumption that the sum of the angles in a triangle is equal to two right angles.

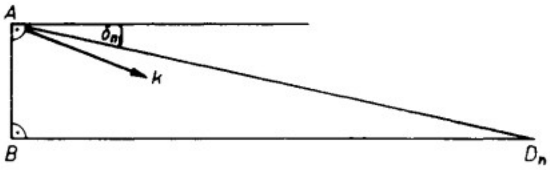

FIG. 19.

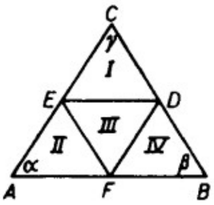

FIG. 20.

What follows is a quick "proof" that the sum of the angles in a triangle is equal to two right angles.

Thus, let D, E, F in Fig. 20 be the midpoints of the sides of a triangle, and I, II, III, IV the triangles determined by the segments DE, EF, DF. Let W be the (unknown) sum of the angles in a triangle. Then it is easy to see that

$$4W = (\alpha + \beta + \gamma) + 3\cdot 2R.$$

It follows that

$$3W = 6R, \qquad W = 2R.$$

According to Theorem 12 this implies the parallel postulate. Where is the error?

The careful reader will note that the only thing which has been proved is the following interesting assertion:

Theorem 13. If the sum of the angles in a triangle is the same for all triangles, then it is equal to two right angles.

To resume our narrative. As we saw, Legendre proved that the sum of the angles in a triangle could not exceed two right angles. It is understandable that he thought he was close to the solution of a problem which had kept mathematicians busy for more than 2000 years.

In 1823 Legendre thought that he had proved the parallel postulate. He reasoned as follows:

If $\alpha + \beta + \gamma < 2R$, then there exists an angle δ such that

$$\alpha + \beta + \gamma = 2R - \delta.$$

δ is called the "defect" of the triangle. The following is clear (Fig. 21): If we divide a triangle by means of a transversal through a vertex into two triangles ADC and BDC, then the defect of the whole triangle is the sum of the defects of its parts. Indeed, by consulting Fig. 21 we see that

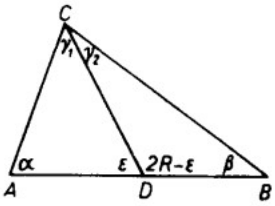

FIG. 21.

$$2R - (\alpha + \beta + \gamma) = 2R - (\alpha + \varepsilon + \gamma_1) + 2R - \\ - (\beta + \gamma_2 + 2R - \varepsilon).$$

This result remains true if the triangle is divided into more than two "subtriangles."

The additivity of the defect is at the heart of Legendre's argument. He reflects the triangle ABC about AB and draws through the image D of C a line which meets the rays CA and CB in E and F. The defect of the triangle CEF is at least equal to the sum 2δ of the defects of the triangles ACB and BAD (Fig. 22). Another reflection applied to $\triangle CEF$ yields a triangle with defect $\geqslant 4\delta$, etc. After n steps we obtain a triangle with defect $2^n \cdot \delta$. For n large enough $2^n \cdot \delta$ would exceed $2R$, which is impossible. Hence $\delta = 0$ and, in view of Theorem 12, the parallel postulate follows.

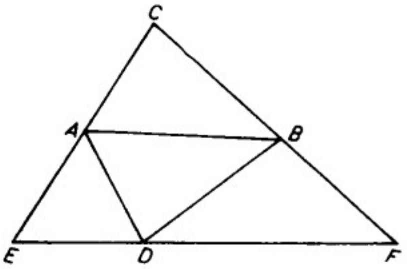

FIG. 22.

The reader is invited to study the argument carefully and detect the flaw in it.

The proof uses as an "axiom" the proposition that it is always possible to draw a line through a point (D) in the interior of an angle which intersects both sides of the angle. Legendre was fully aware of the fact that he was making use of this proposition as an "axiom."

One is likely to regard this proposition as being intuitively obvious. However, in trying to decide whether the available axioms imply this "obvious" result, we get into difficulties.

It is certainly not immediately apparent that this result is equivalent to the parallel postulate and we can hardly blame Legendre for thinking that it was independent of the parallel postulate.

To see that Legendre was mistaken we need only consult our pseudogeometry of Example 1. It is only fair to point out that this technique of testing independence of axioms was unknown to Legendre.

It is obvious that in the case of the point S, say (Fig. 2), there exists no "pseudoline" which intersects *both* sides of the angle α.

Legendre's axiom is seen to be another equivalent of the parallel postulate. Thus we end up with an interesting fact and an incorrect proof of the parallel postulate.

We cannot give here a more detailed history of the attempts to prove the parallel postulate and must leave many such interesting attempts unmentioned.

After 2000 years of vain effort many mathematicians began to show signs of skepticism. They began to doubt that a proof was at all possible. It should be noted that the persistence displayed by so many mathematicians was due to the fact that they viewed the unproved parallel postulate as a shocking flaw in geometry.

This statement is well illustrated by the following excerpt from a letter which the Hungarian mathematician Wolfgang Bolyai [21] wrote to his son, Johann:

> "It is unbelievable that this stubborn darkness, this eternal eclipse, this flaw in geometry, this eternal cloud on virgin truth can be endured."

At the same time the father is horrified by the thought that his son is attracted by the problem of parallels. Wolfgang Bolyai writes:

> "You must not attempt this approach to parallels. I know this way to its very end. I have traversed this bottomless night, which extinguished all light and joy of my life. I entreat you, leave the science of parallels alone ... . I thought I would sacrifice myself for the sake of the truth. I was ready to become a martyr who would remove the flaw from geometry and return it purified to mankind. I accomplished monstrous, enormous labors; my creations are far better than those of others and yet I have not achieved complete satisfaction. For here it is true that si paullum a summo discessit, vergit ad imum. I turned back when I saw that no man can reach the bottom of this night. I turned back unconsoled, pitying myself and all mankind."

And yet again:

> "I admit that I expect little from the deviation of your lines. It seems to me that I have been in these regions; that I have traveled past all reefs of this infernal Dead Sea and have always come back with broken mast and torn sail. The ruin of my disposition and my fall date back to this time. I thoughtlessly risked my life and happiness—aut Caesar aut nihil."

But this resignation of experienced mathematicians was not to be the last chapter in the exciting story of our geometric theorem. Owing in no small measure to the work of the younger Bolyai (1831) the issue took a surprising turn.

Most attempts to prove the parallel postulate employed the indirect method of proof: One argued that if the parallel theorem is false, then the sum of the angles in a triangle is less than two right angles. Then the triangle has a "defect." Then ... .

This mode of argument was expected to produce a contradiction. But none could be found. The thought suggested itself that it might after all be possible to construct a geometry in which the negation of the parallel theorem rather than the parallel theorem holds. After all, it is conceivable that the sum of the angles in a triangle is only approximately equal to two right angles.

Gauss followed up this thought. He used a transit to measure the sum of the angles of a huge triangle and found no significant deviation from two right angles. It was nevertheless possible to conceive of a geometry in which the parallel postulate did not hold and in which the sum of the angles in a triangle was less than two right angles.

Gauss was perhaps the first to have a clear conception of such a "noneuclidean" geometry. The thought struck him as being so revolutionary that he would not make it public. In 1829 he wrote to Bessel:

> "It may take very long before I make public my investigations on this issue; in fact this may not

> happen in my lifetime for I fear the scream of the dullards if I made my views explicit."

In a few places in his letters Gauss expresses the conviction that there exists a closed and consistent geometry in which the euclidean parallel postulate does not hold.

The first published accounts dealing with such a geometry are due to the Hungarian Johann Bolyai (1802–1860) and the Russian Lobachevski (1793–1856).

In 1823 Johann Bolyai could tell his father, who had tried so hard to make him give up his interest in the problem, that he was succeeding:

> "I am resolved to publish a work on parallels as soon as I can put it in order, complete it, and the opportunity arises. I have not yet made the discovery but the path which I have followed is almost certain to lead me to my goal, provided this goal is possible. I do not yet have it but I have found things so magnificent that I was astounded. It would be an eternal pity if these things were lost as you, my dear father, are bound to admit when you see them. All I can say now is that I have created a new and different world out of nothing. All that I have sent you thus far is like a house of cards compared with a tower."

His father advised him to publish his results as soon as possible. Johann Bolyai's comment follows:

> "He advised me that, if I was really successful, I should speedily make a public announcement and that for two reasons. One reason is that the idea might easily pass to someone else who would then publish it. Another reason—and one that seems valid enough—is that when the time is ripe for certain things, these things appear in different places in the manner of violets coming to light in early spring. And since scientific striving is like a war of which one does not know when it will be replaced by peace one

> must, if possible, win; for here preeminence comes to him who is first."

The advice was sound for, when Johann Bolyai published his work in 1831, he had been anticipated by Lobachevski who had presented his treatise to the physical-mathematical division of the University of Kazan on February 2, 1826.

Priority arguments are not important. The time was ripe for this particular insight and it is safe to say that Gauss, Bolyai, and Lobachevski independently drew the same conclusion from the impossibility of proving the parallel postulate.

Their contemporaries paid almost no attention to these new ideas. The view that euclidean geometry was the only possible geometry was so firmly rooted in thought that the new notion was hardly recognized. Although its discoverers developed the new geometry to the point of working out its trigonometry, this geometry lacked intuitive appeal and so was almost incomprehensible. A few decades passed before significant numbers of mathematicians took notice of the works of Bolyai and Lobachevski. The new geometry gained intuitive, appeal as a result of the "models" constructed by Felix Klein and Poincaré.

In our introduction we shall make use of the easy-to-follow Poincaré model.

We precede our exposition with a chapter containing results of elementary geometry which may not be known to some of our readers.

4 +,-.+ + +/ ! + * ! P ! 2 A
0 u A / ! ! &: , & 4 : D !
%) 3#

4 +,-.+ + +/ ! + * ! P ! 2 A
0 u A ! ! && , 4 ! ! # 1 2 && z! H 1 2
%) 3#

4' +,-.+ + +/ ! + * ! P ! 2 A
0 u A ! ! &&&P ! : ! ! ! ! ! &&&
4 : D ! %) 3#
/4 +,-.+ + +/ ! + * ! P ! 2 A
0 u A ! ! && , 4 ! ! # 1 2 && z! H 1 2
%) 3#
(' " % !3 & 7

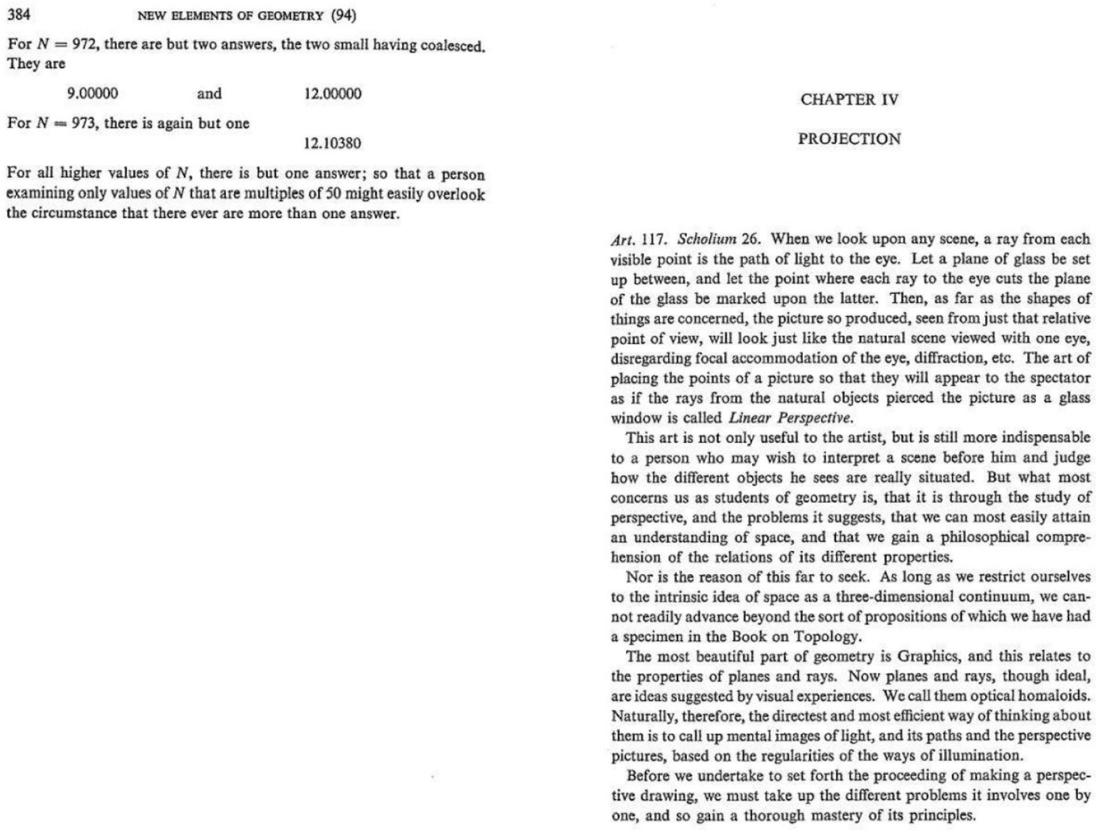

For $N = 972$, there are but two answers, the two small having coalesced. They are

9.00000 and 12.00000

For $N = 973$, there is again but one

12.10380

For all higher values of N, there is but one answer; so that a person examining only values of N that are multiples of 50 might easily overlook the circumstance that there ever are more than one answer.

CHAPTER IV

PROJECTION

Art. 117. *Scholium* 26. When we look upon any scene, a ray from each visible point is the path of light to the eye. Let a plane of glass be set up between, and let the point where each ray to the eye cuts the plane of the glass be marked upon the latter. Then, as far as the shapes of things are concerned, the picture so produced, seen from just that relative point of view, will look just like the natural scene viewed with one eye, disregarding focal accommodation of the eye, diffraction, etc. The art of placing the points of a picture so that they will appear to the spectator as if the rays from the natural objects pierced the picture as a glass window is called *Linear Perspective.*

This art is not only useful to the artist, but is still more indispensable to a person who may wish to interpret a scene before him and judge how the different objects he sees are really situated. But what most concerns us as students of geometry is, that it is through the study of perspective, and the problems it suggests, that we can most easily attain an understanding of space, and that we gain a philosophical comprehension of the relations of its different properties.

Nor is the reason of this far to seek. As long as we restrict ourselves to the intrinsic idea of space as a three-dimensional continuum, we cannot readily advance beyond the sort of propositions of which we have had a specimen in the Book on Topology.

The most beautiful part of geometry is Graphics, and this relates to the properties of planes and rays. Now planes and rays, though ideal, are ideas suggested by visual experiences. We call them optical homaloids. Naturally, therefore, the directest and most efficient way of thinking about them is to call up mental images of light, and its paths and the perspective pictures, based on the regularities of the ways of illumination.

Before we undertake to set forth the proceeding of making a perspective drawing, we must take up the different problems it involves one by one, and so gain a thorough mastery of its principles.

Art. 118. *Definition* 164. A *correspondence* is a system of relationship between two sets of objects which connects all the objects of the first set each with the same number of objects of the second set; and all the objects of the second set each with the same number of objects of the first set. If the two numbers are *M* and *N*, the correspondence is called an *M* to *N* correspondence. Thus, if in a school every pupil has 3 teachers and every teacher 20 pupils, there is a 20 to 3 correspondence between pupils and teachers.

The word *projection* is used in mathematics in several senses. A system by which the points on the surface of the earth, or other surface, are made to correspond one to one to the points of a map, or other plane, and the lines of the former surface are made to correspond one to one to the lines of the plane, is called, by cartographers, a *projection*.[73] Moreover, the operation or result of constructing a figure on a surface which represents a figure in space as if rays were drawn to the eye from all the points (or all the visible points) of the spacious figure and were then cut by the surface of the constructed figure and their points of intersection with it were marked, is called a *central projection*, or, since this is the only kind of projection with which Projective Geometry deals, a *projection* simply. Sig. Professor Cremona, a celebrated Italian geometer, whose beautiful treatise on Projective Geometry has been translated into our barbarous tongue, has used the word *projection* for the first part of the whole operation of central projection. But this is better called production.

Definition 165. From every point of a figure, called the *original* figure (whether plane or spacious) let rays be drawn to one point, called the *eye* of the production or projection, so that from every point is produced a ray and from every ray a plane. Then this operation is called a *production to* the eye *from* the original figure *of* the new spacious figure.

Definition 166. The operation of *plane section* (called section, for short) consists in intercepting a solid, or spacious, figure by a plane, called the plane of section or of delineation, so that every plane is reduced to a ray, and every ray to a point.

---

[73] Peirce was well acquainted with such problems having invented a quincuncial map in 1876 which has proved useful in recent times. See "Charles S. Peirce and the Problem of Map-Projection" by the editor in the *Proceedings of the American Philosophical Society*, vol. 107, 4, August 1963.

Definition 167. The operation of *perspective delineation* consists in the production from the original figure to the eye of a spacious figure, followed by the plane section of the latter by the plane of delineation.

Scholium 27. We here use the traditional phraseology which speaks as if lines and surfaces were something we make. This is not strictly correct. The lines and surfaces are *places* which are *there*, whether we think of them or not. They are quite ideal, it is true. But they are there, in this sense, that we can think of them as being there without being drawn into any absurdity. The ordinary diction however which represents us as *drawing*, — or technical language "describing," — lines and surfaces, and which talks of a moving line as *generating* a surface, with other such expressions, is rather convenient, and can hardly mislead.

Art. 119. *Definition* 168. In perspective the eye of production is generally called the *station point*. The plane of the section is called the *plane of delineation*. The original figures are called *natural*, or *original*.

Art. 120. *Theorem 17.* Two rays in one plane have a point in common.

Demonstration. Let {*b*} and {*c*} be two different rays in one plane (*bc*). Then, I say, they have a common point [*bc*].

For take any point [A] not in (*bc*). Then there is a plane (A*b*) and a plane (A*c*); and these two are different planes. For were they the same, the ray

{(A*b*)(*bc*)}

would be the same as the ray

{(A*c*)(*bc*)}.

But one of these is {*b*}, while the other {*c*}, which are distinct rays. Hence the planes (A*b*) and (A*c*) have in common a ray

{(A*b*)(A*c*)};

and this ray is not in (*bc*), for if it were, it would be common to (A*b*) and (*bc*), and therefore the only ray they have in common. It would be {(A*b*)(*bc*)}; and it would also be common to (A*c*) and (*bc*), and therefore the only ray they have in common. It would be {(A*c*)(*bc*)}. And thus {(A*b*)(*bc*)} and {(A*c*)(*bc*)} would be the same ray. Since then {(A*b*)(A*c*)} is not in (*bc*) it has a point in common with (*bc*).

But [(*bc*){(A*b*)(A*c*)}] is in (A*b*) since it is in the ray {(A*b*)(A*c*)} which

is in this plane. It is in (A*c*) for the same reason. It is in (*bc*). Hence it is in the common ray of (A*b*) and (*bc*) or {(A*b*)(*bc*)}, which is {*b*}, and in the common ray of (A*c*) and (*bc*), or {(A*c*)(*bc*)}, which is {*c*}. It is therefore a point common to {*b*} and {*c*}. Consequently there is such a point. Q.E.D. The student can prove that if two rays have a common point they lie in one plane.

Art. 121. *Theorem 18. The delineation of a straight line is a straight line.* For the original straight line and the station-point lie in one plane (as any point and ray do), and this plane is cut by the plane of delineation in a straight line (as any plane is cut by another in a ray).

Theorem 19. The delineation of a point is a point. For the original point and the station point lie in one ray (as any two points lie), and this ray pierces the plane of delineation in one point (as any ray cuts a plane).

Art. 122. *Scholium* 28. When a spectator stands upon the sea-shore with his eye about 12 feet above the sea-level, the line between sea and sky, owing to the combined effects of the earth's curvature and the refraction of light, appears lower than his eye by the breadth of a line $\frac{1}{100}$ inch thick viewed at a distance of ten inches. And this which is called the *dip of the horizon* will be more, if the spectator's eye is higher, in the proportion of the square root of that height.

If there were no refraction and the earth's surface, instead of being curved, were plane, there would practically be a similar line between sea and sky, only a trifle higher. Whether if vision were perfect, and no objects intervened, such a line would exist, we will not here inquire. Suffice it that for pictorial purposes it may be assumed that there would be such a line at the height of the eye.

Corollary 25. The delineation of three lines which meet in one point is three lines meeting at one point.

Art. 123. *Theorem 20. The delineation of a ray cuts the original ray.*

Demonstration. The delineation lies in the plane of the station point and the original ray. But two rays in one plane have a point in common.

Art. 124. *Theorem 21. The points at which the delineations of three rays which intersect in three points respectively cut their originals lie on one ray (are collinear).*

Demonstration. The three points lie in one plane; and the three original

rays lie in that plane. Their delineations lie in the plane of delineation. Hence, the intersections of the original rays with their delineations lie both in the plane of the three points and in the plane of delineation. But two planes have only a ray in common. Hence, those three intersections lie in the ray in which those two planes intersect.

Art. 125. *Scholium* 29. In order to draw a straight line, the student provides himself with a straight edge. And looks along the edge to make sure it coincides with a ray of light.

In order to draw a line which meets a given line at the firmament, we make use of a ruler along which slides one side of a triangle, another side being a straight edge. Keeping the ruler fixed we slide the triangle into a new position for every new line, and the lines so drawn may be assumed to meet at the firmament, for when continued to a very great distance they will appear to meet there [Fig. 60].

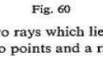

Fig. 60

Art. 126. *Definition* 169. Two rays which lie in one plane, or four points which lie in one plane, or two points and a ray which lie in one plane are said to be *coplanar*.

Definition 170. Two rays through a point, or four planes through one point, or two planes and a ray having a common point are said to be *concurrent*, or *copunctual*.

Definition 171. Three points on one ray, or three planes through one ray, or two points with a plane through them, or two planes with a point on their intersection are said to be *collinear*, or better, *coradial*.

Art. 127. *Theorem 22. If six points lie two by two on three concurrent rays, and six other rays join each point to one of the points on each of the concurrent rays on which that point does not itself lie, then the three intersections each of two rays joining points on one pair of the three concurrent rays are coradial.*

That is (the capital letters standing for points) if the following triplets are collinear OCG, OIJ, OUV, then (pairs of letters standing for rays through the points denoted by those letters) the intersections of {CI} and {GJ}, of {CU} and {GV}, of {IU} and {JV} are coradial.

Demonstration. First, suppose the three concurrent rays are not coplanar. Then the plane of the three points (CIU) contains the rays {CI}, {CU}, and {IU} and the plane of the three points (GJV) contains the rays {GJ} {GV} {JV}. Consequently, the intersections of {CI} and {GJ}, of {CU} and {GV}, of {IU} and {JV} lie at once in the plane (CIU) and in the plane (GJV). But two planes have only one ray in common. Hence, those three intersections lie in that ray.

Second, suppose the concurrent rays are coplanar. Then the three concurrent rays are a perspective delineation of three concurrent rays not coplanar. Namely, we can assume the plane (OCI) as the plane of delineation, taking for the station-point, which we may denote by [O′], any point not in that plane. Then, in the plane (O′OU), we may assume any line through [O] of which {OU} becomes the delineation. We draw the ray {O′U} and denote the point where it cuts the new ray through [O] by [U′]. We draw the line {O′V} and denote the point where it cuts the new ray through [O] by [V′]. Then [U] is the delineation of [U′] and [V] of [V′]; while [C], [G], [I], [J] are delineations from the same centre of themselves. But we have just seen that the intersections of {CI} and {GJ}, of {CU′} and {GV′}, of {IU′} and {JV′} will lie on one ray. The delineation of this ray will be a ray. Hence, the intersections of {CI} and {GJ}, of {CU} and {GV}, of {IU} and {JV} will also lie on one ray, since these are the delineations of the other set of intersections.

Scholium 30. *Geometrical Nets.* The last part of the above demonstration, though perfectly conclusive, is unsatisfactory, or rather betrays an unsatisfactoriness in the statement of the properties of space, inasmuch as it proves something about a plane, having nothing to do with a third dimension, by means of a perspective delineation involving a third dimension. This is unavoidable, because we have not stated what the course of a ray is in a plane, in such a way that anything can be deduced from it directly. There are various ways, none of them all we could desire, for remedying this defect.

One way is to say that when we draw a figure by first assuming certain points and rays arbitrarily, and afterwards marking certain *junctions* (or rays containing two points already marked) and *intersections* (or points contained in two rays already marked), all the arbitrarily taken elements may be considered and treated as if they were in as many dimensions as their number would admit of their extending into. For example, our three collinear lines may be reasoned about exactly as if they were in three dimensions. After them, the six points are taken on them and can therefore not be in additional dimensions. This is a way of perfectly describing the course of a ray.

Corollary 26. The above theorem gives by the principle of duality the following proposition. Through each of three coplanar rays, $\{a\}$, $\{b\}$, $\{c\}$ (which may or may not belong to one aster), we take two planes, planes (α) and (α') in $\{a\}$, planes (β) and (β') in $\{b\}$, planes (γ) and (γ') in $\{c\}$. The intersection of (α) and (β) will lie in one plane with that of (α') and (β'). The same will be true of $\{\beta\gamma\}$ and $\{\beta'\gamma'\}$, of $\{\alpha\gamma\}$ and $\{\alpha'\gamma'\}$. The three planes of those pairs of intersections are coradial.

Corollary 27. The converse of theorem 22 also holds, namely that if the first three rays are not concurrent, the last three intersections are not collinear. In the case of the plane figure, this follows by the principle of plane duality (which interchanges points and rays). In the case of the solid it is evident; for if from three points in the range of the hinge-ray of two planes be drawn from each two rays in the two planes respectively, it follows from [Cor. 15] that the three pairs of rays lie in three planes; and these intersect in one point.

Corollary 28. The converse of the first corollary is also true, being derivable from the last corollary by the principle of duality. The proposition is that if through a ray we take three planes (α), (β), (γ), and draw three rays $\{a\}$ in (α), $\{b\}$ in (β), $\{c\}$ in (γ), so that $\{a\}$, $\{b\}$, $\{c\}$, are concurrent, and if we draw three other concurrent rays $\{a'\}$ in (α), $\{b'\}$ in (β), $\{c'\}$ in (γ), (whether these are concurrent with all the other rays or not), then the three planes of (ab), (ac), (bc), will respectively intersect the planes of $(a'b')$, $(a'c')$, $(b'c')$, and the three rays of intersection will be coplanar.[74]

Art. 128. *Scholium* 31. Suppose a spectator to stand upon a boundless plane, — in geometrical strictness [Fig. 61]. Then, according to the idea of space adopted in graphics every ray through his eye will cut that plane. But graphics is non-committal as to the distances at which these rays cut. The question whether there be any parts of the plane at an infinite distance is beyond its purview.

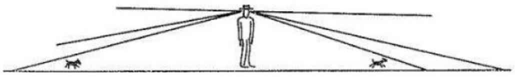

Fig. 61

---

[74] P. 235 of Peirce's MS. is missing.

Perhaps, for ought graphics can say, a dog running off along a ray would after a number of days reappear from behind, having completed the circuit of space. Perhaps, on the contrary, the observer standing on the plane would see a circle lower than his eye which was at an infinite distance so that time would not be long enough to carry light any further. Perhaps, as is generally assumed, all parts of the plane below his eye are at a finite distance, but the ray on a level with his eye would strike the "horizon" or limit between earth and sky. In this case, the parts of space at an infinite distance form the limit of planes, and it is perfectly proper and true to say they are a plane at an infinite distance. Now this may not be quite true: We shall discuss this somewhat under Metrics: but it is certainly excessively close to the truth. It would be quite nonsensical to draw a perspective picture on any other supposition.

Definition 172. The plane at infinity is called the *Absolute*, or the *firmament*.

Definition 173. The perspective delineation of the intersection of a plane with the firmament is called the *vanishing-line* of that plane. The perspective delineation of the intersection of a ray with the firmament is called the *vanishing-point* of that ray.

Art. 129. *Scholium* 32. The art of linear perspective is not merely useful in drawing, for which purpose a perfect theoretical knowledge of it is not indispensable to an artist of keenly accurate perceptions and vivid imagination. It is also valuable in enabling us accurately to interpret a perspective picture or a scene before us, and from the appearance to understand how the different objects are really situated. But what renders it most important to the student of geometry is that in it lies the best key to the science of space. There is no other way of learning geometry comparable to that by the road of projections for facility and for philosophical insight into the true relations of different theorems.

Scholium 33. Two great faults of perspective are ordinarily committed by artists. The first is that of exaggeration. The treatises on the subject inculcate this. A perspective picture is drawn to be looked at with one eye (for binocular perspective is impossible without a stereoscope) from a particular stand-point. If the picture is otherwise viewed, its perspective is falsified. Now, artists have the habit of constructing their linear perspective as if the picture were to be viewed from a distance of a foot or less, while the whole treatment is such that it requires to be seen from a distance five times as great. Whether this is justifiable by artistic principles, or not, it is not for the mere geometrician to say. Perhaps the effect is to partially compensate for the absence of the binocular perspective. But certainly *true* perspective should first be learned.

In the second place, artists usually paint landscapes with figures so that to be viewed rightly they should be placed almost on the floor, while the spectator squats like a tailor on a low stool. For the horizon line which must be very nearly on the level of the observer's eye (being lower by only about $\frac{1}{100}$ of the distance of the plane of delineation from the point of view, even if the spectator be supposed to be 1200 feet above the level of the horizon) is often represented as cutting figures standing on the level ground of the spectator at about their waists.

Art. 130. *Scholium* 34. To draw a picture in perspective, the paper should be set up in the position in which it is to be viewed and the most important planes should first be treated, two lines being drawn for each, first, the *intersecting line*, or ray in which each cuts the plane of delineation, and second, the *vanishing line*, or the intersecting line of its vanishing plane, that is, the line in which the plane through the eye that cuts the original plane in a ray of the firmament cuts the plane of delineation.

The most important plane is the plane on which the observer stands and which is level under his feet. The intersecting line of this plane cannot be drawn on the paper, but it should be drawn on the floor. The vanishing line of the ground plane is the horizon and is a level line about five feet above the floor.

We next mark upon the horizon the *centre of the picture*, which is the point of the plane of delineation nearest the observer's eye. From the centre of the picture, by means of a plummet we draw a plumb line to the floor. It will now be convenient to divide the ground into squares. For this purpose, we may proceed as follows. We mark on the floor the point plumb beneath the station point. Then, on the intersecting line of the ground and plane of delineation we lay off to the left $\frac{3}{4}$ of the distance of the eye from the picture, and to the right $1\frac{1}{3}$ of the same distance. From these two points we draw lines to the point plumb under the eye. The left hand line will be $1\frac{1}{4}$ times as long as the distance of the eye from the picture, while the right hand one will be $1\frac{2}{3}$ as long as the same distance [Fig. 62].[75]

---

[75] This section is incomplete. For Peirce writes a note to himself to "finish later."

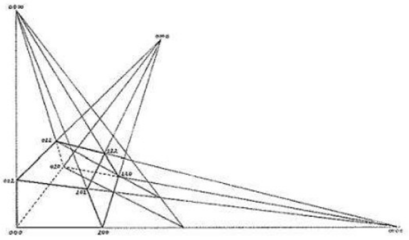

Fig. 62

Art. 131. *Theorem 23. Let any point be designated as* 000. *Let three rays be drawn through this point, and two points be taken on each and marked, those on one* 200 *and* ∞00, *those on another* 020 *and* 0∞0, *and those on the third* 002 *and* 00∞. *Let other rays be drawn and their intersections marked, but only when their marks can be determined by the rule that any point marked on a ray drawn between two points must have all its three numbers weighted means of the corresponding numbers of the other points, the weights being the same for all the numbers. Then, every triplet of numbers will belong to one point and one only, and if the first three rays are not coplanar every point of space is designated by one triplet of numbers and one only.*

Explanation. Some points in this statement require explanation.

Not every intersection will receive any mark. It is easy to find these in practice, so that the algebraic condition that an intersection shall be marked is hardly required. But the condition is that if the numbers attached to the points on one of the intersecting rays are x_1 y_1 z_1 and x_2 y_2 z_2 and those of the points on the other one of the intersecting rays are X_1 Y_1 Z_1 and X_2 Y_2 Z_2, then the intersection cannot be marked unless

$$\begin{aligned}&(x_1y_2-x_2y_1)(Z_1-Z_2) + (y_1z_2-y_2z_1)(X_1-X_2)\\ &+(z_1x_2-z_2x_1)(Y_1-Y_2)+(x_1-x_2)(Y_1Z_2-Y_2Z_1)\\ &+(y_1-y_2)(Z_1X_2-X_2X_1)+(z_1-z_2)(X_1Y_2-X_2Y_1)=0.\end{aligned}$$

Denoting the numbers of the intersection by x, y, z,

$$x = \frac{(X_2-X_1)(x_1y_2-x_2y_1)+(x_2-x_1)(X_2Y_1-X_1Y_2)}{(x_1-x_2)(Y_2-Y_1)-(y_1-y_2)(X_2-X_1)}$$

$$y = \frac{(Y_2-Y_1)(y_1z_2-y_2z_1)+(y_2-y_1)(Y_2Z_1-Y_1Z_2)}{(y_1-y_2)(Z_2-Z_1)-(z_1-z_2)(Y_2-Y_1)}$$

$$z = \frac{(Z_2-Z_1)(z_1x_2-z_2x_1)+(z_2-z_1)(Z_2X_1-Z_1X_2)}{(z_1-z_2)(X_2-X_1)-(x_1-x_2)(Z_2-Z_1)}$$

But these cumbrous formulae are seldom of any use; for the calculation is perfectly simple.

When one of the numbers is ∞, the corresponding number of the other point has to be changed alone.

By the "weighted mean" of two numbers is meant their sum after multiplying them respectively by two fractions whose sum is 1, or their difference after multiplying them respectively by two numbers whose difference is 1. It must be understood that ∞ multiplied by 0 may give any number as product.

After the first seven points have been laid down, namely, 000, 200, ∞00, 020, 0∞0, 002, 00∞, we next lay down 220 which is the intersection of the ray joining 020 and ∞00 ($220 = 1.020 + 0.\infty 00$) and the ray joining 200 and 0∞0; we lay down 202 which is the intersection of the ray joining 002 and ∞00 and the ray joining 200 and 00∞; and we lay down 022 which is the intersection of the ray joining 020 and 00∞ and the ray joining 002 and 0∞0. Also 222 which is the intersection of the ray joining 022 and ∞00 and the ray joining 202 and 0∞0. It also should by the rule be on the ray joining 220 and 00∞; and so it is found to be. This is an example of how the theorem contains a statement of a truth about space, and is not merely a system of designating points, for which it might be mistaken. At the intersection of the ray joining 202 and 220 and the ray joining 222 and 200 is the point 211. At the intersection of the ray joining 211 and 00∞ and of the ray joining 202 and 222 is the point 212. At the intersection of the ray joining 211 and 00∞ and of the ray joining 200 and 220 is the point 210. At the intersection of the ray joining 212 and 022 and of the ray joining 002 and 202 is the point 302. At the intersection of the ray joining 210 and 020 and of the ray joining 000 and 200 is the point 300. The point 300 should also lie on the ray joining 302 and 00∞; and so it is found to do. In that way we can with the greatest ease go on adding points and designating them without end. The meaning of the theorem having thus been rendered clear, it would be in order to demonstrate its truth.

Scholium 35. But the strict demonstration is difficult just now, while after the barycentric notation is mastered, the student will be able to make it out for himself. The object of enunciating the theorem here is to show how perspective may be used to obtain truths of this sort.

It is necessary here to use a proposition, which will also not be strictly proved, called the *fundamental theorem of axonometry*. This proposition is substantially that a cube, or die, may be so represented in perspective that three edges meeting in one corner have for their delineations any three lines meeting at a point, no matter how inclined to one another, or of what lengths, they may be. The fact that a solid body has six degrees of freedom while six numbers are sufficient and necessary to define the directions and lengths of the three lines suggests that this is probably true, as it is in fact.

But regarding the figure drawn on the paper as the perspective representation of a cube, we see that the above theorem is equivalent to saying that space can be cut up into indefinitely small cubes, a familiar fact, though a proposition not of graphics but of metrics. The theorem itself, however, is purely graphical.

CHAPTER V

NOTATIONS FOR GRAPHICS

Art. 132. We shall use square brackets, [ ], to denote the *point* anyway determined by the places denoted by the letters written within the brackets. Thus

$$[\alpha\beta\gamma]$$

will denote the point contained at once in the planes α, β, γ

$$[\beta c]$$

will denote the point where the ray, c, cuts the plane β.

We shall use parentheses, ( ), to denote a plane. Thus

$$(\mathrm{B}c)$$

will denote the plane containing the point B and the ray c.

We shall use braces, { }, to denote rays. Thus,

$$\{\alpha\beta\}$$

will be the ray contained in the two planes α and β. So

$$\{\mathrm{AB}\}$$

will be the ray through the two points A and B.

Theorem 24. If two rays a and b are concurrent with two others, c and d, then

$$\{[ac][bd]\} = \{(ad)(bc)\}.$$

Demonstration.

For $[ac]$ write P for $[bd]$ write Q;
For $[ad]$ write R for $[bc]$ write S.

Then $\{[ac][bd]\} = \{\mathrm{PQ}\}$. Since a contains P and also R,

$$a = \{\mathrm{PR}\};$$

(*%) !3 & 0

The notation represents these triplets to be *coradial* (in one ray) for taking the first, twice the first added to the third gives the second. Thus

$$\begin{array}{llll} 2\mathrm{O} & +8\mathrm{X} & -6\mathrm{Y} & \\ 7\mathrm{O} & & +6\mathrm{Y} & +5\mathrm{Z} \\ 9\mathrm{O} & +8\mathrm{X} & & +5\mathrm{Z} \end{array}$$

CHAPTER VI

THE PHIMUS

Art. 140. *Intersections of Rays.*

Theorem 25. Space contains a quadruply continuous multitude of rays.

Demonstration. Every ray cuts every plane, except the book of planes hinged about it. Take two planes and a point in each; then just one ray exists through each such pair of points (except that there is only one ray for all the pairs of points which are both in both planes, and only one ray for all the pairs of points of which one is any one point in both planes), and these are all the rays of space. Hence, since there is a doubly continuous multitude of points in each plane, there is a quadruply continuous multitude of rays in space.

Art. 141. *Theorem 26. Cutting every ray there is a triply continuous multitude of rays.*

Demonstration. Through every ray there is a singly continuous multitude of planes. In each of those planes there is a doubly continuous multitude of rays, all of which are different and all cut the hinge ray, except one in each plane which coincides with it. The double continuity of the rays in one plane, together with the single continuity of planes, makes up a triple continuity.

Art. 142. *Theorem 27. Cutting every pair of rays is a doubly continuous multitude of rays.*

Demonstration. Each ray of a pair of rays contains a singly continuous multitude of points, and through any pair of points, one in each of the two rays, there is just one ray.

Remark that it makes no difference whether the first pair of rays cut one another or not.

Art. 143. *Theorem 28. Cutting every trio of rays is a singly continuous*

multitude of rays and a straight fibre successively occupying those positions *generates a ruled surface.*

Demonstration. For consider a flat lamina through one of the fixed rays. This flat lamina is cut by the other two fixed rays in two points. Through these two points is a ray, which ray cuts every other ray in the same plane, of which the first fixed ray is one. As the flat lamina turns round the first ray, the straight fibre cutting the three fixed rays moves; and a moving fibre generates a surface.

Definition 175. A surface generated by a moving straight fibre is called a *regulus*, or *ruled surface*. If the successive generators cut one another, it has an edge of regression and is called a *developable*, because a lamina of that form may be bent flat without stretching. Otherwise, it is called a *torus*.

Definition 176. A torus generated by a straight fibre which constantly cuts three fixed rays is called a *phimus* [Fig. 63].

Phimus (φίμος) is the Greek for dice-box. That surface resembles a somewhat flattened dice-box extending to infinity.

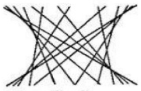

Fig. 63

Art. 144. *Theorem 29. If all of three rays are cut by each of three other rays and if a seventh cuts the first three and an eighth the second three, then the seventh and eighth intersect.*

Demonstration. Let i, j, k, be the first three rays, and p, q, r, the second three, s the seventh and l the eighth.

Use the following notation

$$[ip] = \mathrm{A}, \qquad [ir] = \mathrm{B},$$
$$[kp] = \mathrm{C}, \qquad [kr] = \mathrm{D}.$$

We are now free to call any point we like $[\mathrm{A}+\mathrm{B}+\mathrm{C}+\mathrm{D}]$. We will apply that designation to $[jq]$.

It follows that

$$[iq] = [\mathrm{A}+\mathrm{B}], \qquad [jp] = [\mathrm{A}+\mathrm{C}],$$
$$[kq] = [\mathrm{C}+\mathrm{D}], \qquad [jr] = [\mathrm{B}+\mathrm{D}].$$

Suppose the seventh ray which cuts $j = \{[\mathrm{A}+\mathrm{C}][\mathrm{B}+\mathrm{D}]\}$ cuts it in the point $[p[\mathrm{A}+\mathrm{C}]+(1-p)[\mathrm{B}+\mathrm{D}]]$
and that the eighth ray, which cuts $q = \{[\mathrm{A}+\mathrm{B}][\mathrm{C}+\mathrm{D}]\}$, cuts it in the point $[q[\mathrm{A}+\mathrm{B}]+(1-q)[\mathrm{C}+\mathrm{D}]]$.
Then the seventh ray cuts $i = \{\mathrm{AB}\}$ in

$$[p\mathrm{A}+\overline{1-p}\,\mathrm{B}],$$

and cuts $k = \{\mathrm{CD}\}$ in

$$[p\mathrm{C}+\overline{1-p}\,\mathrm{D}];$$

and the eighth ray cuts $p = \{\mathrm{AC}\}$ in

$$[q\mathrm{A}+\overline{1-q}\,\mathrm{C}],$$

and cuts $r = \{\mathrm{BD}\}$ in

$$[q\mathrm{B}+\overline{1-q}\,\mathrm{D}].$$

The general point of the seventh ray is got by varying x in the syzygy

$$[xp\,\mathrm{A}+x\overline{1-p}\cdot\mathrm{B}+\overline{1-x}\cdot p\,\mathrm{C}+\overline{1-x}\cdot\overline{1-p}\cdot\mathrm{D}].$$

The general point of the eighth ray is got by varying y in the syzygy

$$[qy\,\mathrm{A}+q\cdot\overline{1-y}\cdot\mathrm{B}+\overline{1-q}\cdot y\,\mathrm{C}+\overline{1-q}\cdot\overline{1-y}\cdot\mathrm{D}].$$

If we put $x = q$ and $y = p$, we get for a point on both rays

$$[qp\,\mathrm{A}+q\cdot\overline{1-p}\cdot\mathrm{B}+\overline{1-q}\cdot p\,\mathrm{C}+\overline{1-q}\cdot\overline{1-p}\cdot\mathrm{D}].$$

Thus the seventh and eighth rays have a point in common.

Art. 145. *Theorem 30. Every phimus has two sets of generators.*

Demonstration. Let i, j, k be the three fixed rays which every generator of the phimus intersects. Let p, q, r be three generators. Then there is a phimus through these generators, of which i, j, k are generators. Let l be another generator of it and s any generator of the first phimus. Then l and s intersect. Thus, every generator of either is cut by a generator of the other. That is, all the phimi through three generators of the first phimus are identical. But the first phimus is itself such a phimus. Hence, there is no other through the three rays i, j, k. This phimus has, therefore, two sets of generators, all of the first cutting i, j, k, and all of the second cutting p, q, r.

Art. 146. *Definition* 177. A curve [which is] the section of a phimus by a plane is called a *conic*.

Theorem 31. (*Pascal's Hexagram.*) If A, B, C, D, E, F are six points on a conic the points

$$[\{AB\}\{DE\}],$$
$$[\{BC\}\{EF\}],$$
$$[\{CD\}\{FA\}],$$

lie in one ray.

Demonstration. Let i, j, k be three generators of the phimus of one set through A, C, E. Let p, q, r be three generators of the phimus of the other set through B, D, F. Then $\{AB\}$ is in the plane (ip) and $\{DE\}$ is in the plane (kq). Hence $[\{AB\}\{DE\}]$ is in the ray $\{(ip)(kq)\} = \{[iq][kp]\}$.

In like manner $\{BC\}$ is in (jp), and $\{EF\}$ is in (kr) and hence $[\{BC\}\{EF\}]$ is in $\{(jp)(kr)\} = \{[jr][kp]\}$. And in like manner $\{CD\}$ is in (jq) and $\{FA\}$ is in (ir), and hence $[\{CD\}\{FA\}]$ is in $\{(jq)(ir)\} = \{[iq][jr]\}$. Now the three rays

$$\{[iq][kp]\},$$
$$\{[jr][kp]\},$$
$$\{[iq][jr]\},$$

are all in the plane $([iq][jr][kp])$. Hence the intersections of these planes with the plane of the conic which are

$$[\{AB\}\{DE\}],$$
$$[\{BC\}\{EF\}],$$
$$[\{CD\}\{FA\}],$$

lie in the intersection of those two planes, which is a ray. Q.E.D.

Corollary 29. We may join the 6 points in any order we please. Thus taking any one, this may be directly joined to any of the ten pairs among the remaining five, and the other three may be taken in 6 orders, so that there are 60 Pascal lines, as they are called, for each hexagram.

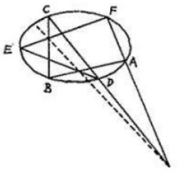

Fig. 64

Corollary 30. The Pascal hexagram, when five points on the conic are given, enables us to find as many more as we please [Fig. 64]. Thus, suppose F is unknown. We draw $\{AF\}$ in any direction we please. We shall thus take

$$[\{CD\}\{FA\}]$$

wherever we like upon $\{CD\}$. The point

$$[\{AB\}\{DE\}]$$

will be given, and through these we shall draw the Pascal line. The point where this cuts $\{BC\}$ will be

$$[\{BC\}\{EF\}]$$

We shall now draw $\{EF\}$ and this will cut $\{AF\}$ in F, which will thus be ascertained.

Corollary 31. One conic, and one only, passes through any one set of five points in a plane, no four of them coradial.

Art. 147. Imagine the three *directrices*, or fixed rays defining a phimus, to move, so that the phimus changes shape, in the following way [Fig. 65]. Let a point, V, be selected not in either of the three directrices. Let

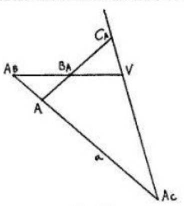

Fig. 65

the three directrices be a, b, c. Consider the plane (aV). Denote by B_A and C_A the points

$$B_A = [b(aV)]$$
$$C_A = [c(aV)]$$

Denote by A_B and A_C the points

$$A_B = [a(bV)]$$
$$A_C = [a(cV)]$$

Denote by A the point

$$A = [a\{B_A C_A\}].$$

Now a is moving to continue in the plane (aV), b in the plane (bV), c in the plane (cV). The directrix a is to continue to pass through A as a fixed point; and b and c are to continue to pass through their corresponding points and the relative speed of motion is to be such that B_A and C_A in their motion continue to be in one ray through A. And so with A_B and C_B, and with A_C and B_C.

Let us see what sort of change this will make in the shape of the phimus. The ray $\{B_A C_A\}$ is plainly the generator which is in the plane (aV). Thus the two rays a and $\{B_A C_A\}$ lie in the plane (aV). Then it follows from the explanation of contact given in Book II that where these rays intersect is the point where the plane (aV) is in contact with the *phimus*. As this is not to move, and is to continue the intersection of the directrix and generator, it follows that the same three planes are to continue to be in contact with the phimus at the same points.

The three planes will come together at [V], and the motion can be made in such sense that the phimus will become more and more [contracted] toward V; and as the limit of the endless series of shapes, the three directrices will pass through V.

Definition 178. A surface generated by a straight fibre which in moving constantly passes through a fixed point, V, and always touches a fixed phimus, is called a *quadric cone*. The point, V, is called the *vertex* of the cone.

Corollary 32. A pair of concurrent rays is a limiting case of a conic. For a pair of rays is the section of a quadric cone by a plane through the vertex. Now a quadric cone is the limit of an endless series of phimi; and the plane section of a phimus is a conic.

Definition 179. A limiting case of any regular place which suddenly becomes two or more regular places of simpler nature is called a *degenerate* place.

Corollary 33. A pair of concurrent rays is a degenerate conic.

Corollary 34. Pascal's hexagram holds good of a pair of concurrent rays [Fig. 66].

Corollary 35. Every section of a cone is a conic or limit of conics, because the cone is the limit of an endless series of phimi; and a conic is by definition the plane section of a phimus.

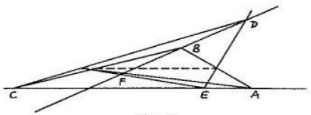

Fig. 66

It has hitherto been the practice to *define* a conic as the plane section of a cone. But the definition here used is more convenient and accurate, and facilitates the development of the subject.

Corollary 36. By applying the principle of duality to the proposition about Pascal's hexagram we get the following. Any six planes tangent to a quadric cone, as $\alpha, \beta, \gamma, \delta, \varepsilon, \zeta$, giving a cycle of six rays of intersection $\{\alpha\beta\}$, $\{\beta\gamma\}$, $\{\gamma\delta\}$, $\{\delta\varepsilon\}$, $\{\varepsilon\zeta\}$, $\{\zeta\alpha\}$, the planes through the three most distant pairs, that is,

$$(\{\alpha\beta\}\{\delta\varepsilon\}),$$
$$(\{\beta\gamma\}\{\varepsilon\zeta\}),$$
$$(\{\gamma\delta\}\{\zeta\alpha\}),$$

contain one ray in common.

Art. 148. *Corollary* 37. Since a section of a cone is a conic, by cutting the cone of the last corollary by a plane, we get the following, called *Brianchon's theorem*. The intersections of the three most distant pairs of any six tangents to one conic meet in one point. "Most distant," here, means most distant in any cyclical order in which the tangents may be taken [Fig. 67].

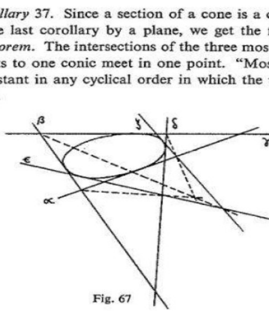

Fig. 67

Corollary 38. It follows that a duality exists between the properties of points on and tangents to a conic.

Corollary 39. In the case of Pascal's Theorem if A and B are the limit of points indefinitely near and the same is true of D and E, the proposition becomes that [of] taking any two tangents to the conic, and joining the points of tangency A and D with any two other points C and F of the conic [Fig. 68]. The intersection of the tangent is coradial with

[{AC}{DF}]
and [{AF}{DC}].

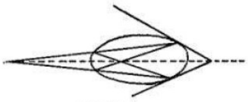

Fig. 68

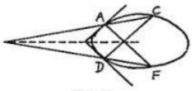

Fig. 69

Corollary 40. It is very important in geometry to take care that propositions are completely stated and considered. The last corollary may be otherwise stated as follows. If A, C, D, F are four points on a conic then

[{AC}{DF}]
and [{AF}{DC}]

are coradial with the intersection of the tangents at A and D. But this must be equally true of the tangents at C and F, for the lettering of A and C, D and F might be interchanged [Fig. 69].

Art. 149. Take any two points upon a phimus. Let the generators of one set through those points be *a* and *b*, and those of the other set *p* and *q*. Then the points will be

[*ap*] and [*bq*]

joined by the ray

{[*ap*][*bq*]}.

But *a* also cuts *q* and *b* cuts *p*. So that there is another ray

{[*bp*][*aq*]}.

These rays are *conjugate*, that is, each has the same relation to the other.[77]

---

[77] Peirce probably intended to add more material on conics at this point.

BOOK IV

METRICS

- Herbert Meschkowski/
- Noneuclidean Geometry...(1964).pdf

Chapter 3 (Pag.21(25)

From the History of the Parallel Postulate

- Algebra and Geometry Peirce/
- Charles... New Elements of Mathematics Volum II...(1976).pdf

Algebra e Geometry

Pag. 385 (209) → Chapter IV

Projection

Art. 131 - Theorem 23

Pag. 419(226) - Chapter VI

The PHIMUS.

Theorem 25 Space contains a quadruply continuous multitude of rays.

Pág. 420(227) - Definition 176

A torus generated by a straight fibre which constantly cuts three fixed rays is called a phimus (Fig. 63)

- Charles... New Elements of Mathematics Volume III ... 1976.pdf

Pág. 495 (268) 9

Map Projection

Pág. 497 (269) Appendix No. 15

A Quincuncial Projection of the Sphere

Orthomorphic, or conform projection formed by transforming the stereographic projection, with a pole at infinity, by means of an elliptic function.

For that purpose, l being the latitude, and θ the longitude, we put —

$$\cos^2\varphi = \frac{\sqrt{1-\cos^2 l\cos^2\theta} - \sin l}{1+\sqrt{1-\cos^2 l\cos^2\theta}},$$

and then $\frac{1}{2}F\varphi$ is the value of one the rectangular co-ordinates of the point on the new projection.

This is the same as taking —

$\cos \mathrm{am}(x+y\sqrt{-1})$ (angle of mod. = 45°) = $\tan\frac{p}{2}(\cos\theta+\sin\theta\sqrt{-1})$, where

+,-.+ + +/ !+* ! P ! 2 A 0
u A ! ! &&&P ! : ! ! ! ! ! &&&
4 :D ! %) 3#
(*%") % !3 7

is constant. Then number the meridians and also the parallels. Then draw curves through the intersections of meridians with meridians, the sums of numbers of the intersecting meridians being constant on any one curve. Also do the same thing for the parallels. Then these curves will represent the meridians and parallels of a new projection having north poles and south poles wherever the component projections had such poles.

Functions may, of course, be classified according to the pattern of the projection produced by such a transformation of the stereographic projection with a pole at the tangent points. Thus we shall have —

1. Functions with a finite number of zeroes and infinites (algebraic functions).
2. Striped functions (trigonometric functions). In these the stripes may be equal, or may vary progressively or periodically. The stripes may be simple, or themselves compounded of stripes. Thus, sin (a sin z) will be composed of stripes each consisting of a bundle of parallel stripes (infinite in number) folded over onto itself.
3. Chequered functions (elliptic functions).
4. Functions whose patterns are central or spiral.

I. *Table of rectangular co-ordinates for construction of the "quincuncial projection."*

x (for longitudes in upper line). y (for longitudes in lower line).

Lat.	0° 90	5° 85	10° 80	15° 75	20° 70	25° 65	30° 60	35° 55	40° 50	45° 45	50° 40	55° 35	60° 30	65° 25	70° 20	75° 15	80° 10	85° 5	Lat.
°																			°
85	.033	.033	.033	.032	.031	.030	.029	.027	.025	.024	.021	.019	.017	.014	.011	.009	.006	.003	85
80	.067	.066	.066	.064	.063	.061	.058	.055	.051	.047	.043	.038	.033	.028	.023	.017	.012	.006	80
75	.100	.100	.099	.097	.094	.091	.087	.082	.077	.071	.065	.058	.050	.042	.034	.026	.017	.009	75
70	.135	.134	.133	.130	.127	.122	.117	.110	.103	.095	.087	.077	.067	.057	.046	.035	.023	.012	70
65	.169	.169	.167	.163	.159	.154	.147	.139	.130	.120	.109	.097	.085	.072	.058	.044	.029	.015	65
60	.205	.204	.201	.198	.192	.185	.177	.168	.157	.145	.131	.117	.102	.086	.070	.053	.036	.018	60
55	.241	.240	.237	.232	.226	.218	.208	.197	.184	.170	.154	.138	.120	.102	.082	.062	.042	.021	55
50	.278	.277	.274	.269	.261	.251	.240	.227	.212	.196	.178	.159	.139	.117	.095	.072	.048	.024	50
45	.317	.316	.312	.306	.297	.286	.273	.258	.241	.223	.202	.181	.158	.134	.109	.083	.055	.028	45
40	.357	.356	.351	.344	.334	.321	.307	.290	.270	.250	.228	.204	.179	.151	.123	.094	.063	.032	40
35	.400	.398	.393	.384	.373	.358	.341	.322	.301	.279	.254	.228	.200	.170	.139	.106	.071	.036	35
30	.446	.443	.437	.427	.413	.396	.377	.356	.332	.308	.281	.253	.222	.190	.155	.119	.081	.041	30
25	.495	.492	.484	.471	.455	.435	.414	.391	.365	.338	.309	.279	.246	.211	.174	.134	.091	.046	25
20	.548	.545	.534	.518	.498	.476	.452	.426	.398	.369	.339	.307	.272	.235	.195	.151	.104	.053	20
15	.609	.604	.589	.568	.544	.517	.490	.461	.432	.401	.369	.336	.300	.262	.219	.173	.121	.062	15
10	.681	.672	.649	.620	.590	.559	.528	.497	.466	.434	.401	.367	.330	.291	.248	.200	.143	.076	10
5	.775	.752	.713	.673	.635	.600	.566	.532	.500	.467	.433	.399	.363	.324	.282	.234	.177	.102	5
0	1.000	.841	.774	.723	.679	.639	.602	.567	.533	.500	.467	.433	.398	.361	.321	.277	.226	.159	0

II. *Preceding table enlarged for the spaces surrounding infinite points.*[2]

x (for longitudes in upper line). y (for longitudes in lower line).

Lat.	0° 90	1° 89	2° 88	3° 87	4° 86	5° 85	6° 84	8° 82	10° 80	12½° 77½	15° 75	75° 15	77½° 12½	80° 10	82° 8	84° 6	85° 5	86° 4	87° 3	88° 2	89° 1	Lat.
°																						°
15	.609	.609	.608	.607	.606	.604	.602	.596	.589	.579	.568	.173	.147	.121	.098	.074	.062	.050	.038	.025	.013	15
12½	.643	.643	.642	.641	.639	.636	.634	.627	.618	.606	.594	.185	.159	.131	.107	.082	.069	.055	.042	.028	.014	12½
10	.681	.681	.680	.678	.675	.672	.668	.659	.649	.635	.620	.200	.173	.143	.118	.091	.076	.062	.047	.031	.016	10
8	.715	.714	.713	.710	.706	.702	.697	.686	.674	.658	.641	.213	.185	.155	.129	.100	.085	.069	.052	.035	.018	8
6	.753	.752	.750	.746	.741	.735	.728	.714	.700	.681	.662	.227	.199	.169	.142	.112	.095	.078	.060	.040	.020	6
5	.775	.774	.770	.765	.759	.752	.745	.729	.713	.692	.673	.234	.207	.177	.150	.119	.102	.084	.065	.044	.022	5
4	.798	.797	.793	.786	.779	.770	.761	.743	.725	.704	.683	.242	.215	.185	.158	.128	.110	.092	.071	.049	.025	4
3	.825	.823	.817	.808	.798	.788	.778	.757	.738	.715	.693	.250	.224	.194	.168	.137	.120	.101	.079	.053	.029	3
2	.857	.853	.843	.831	.819	.806	.794	.772	.750	.726	.703	.259	.233	.204	.178	.148	.131	.112	.090	.065	.035	2
1	.899	.889	.872	.854	.839	.824	.810	.785	.763	.737	.713	.268	.243	.215	.190	.161	.144	.126	.105	.079	.046	1
0	1.000	.929	.899	.877	.857	.841	.825	.798	.774	.747	.723	.277	.253	.226	.202	.175	.159	.143	.123	.101	.071	0

[2] The map is found as the frontispiece of Volume 3. Among so-called "fragments" in the Peirce Collection at the Houghton Library is a quincuncially-ruled chart on which Peirce plotted details of the celestial sphere. The constellation *Virgo* is sketched in quite distinctly, wings and all.

B. P.S. OF A LETTER TO J. M. PEIRCE (L 339: 5 APRIL 1894)

My chart is a skew Mercator, that is, a Mercator with the poles of an oblique great circle at infinity. Its value is great. The formulae are pretty. Start with a stereographic centred at the pole of that circle. Then the equation of every circle (meridian or parallel) is in polar coordinates

$$(r\cos\theta - M)^2 + r^2\sin^2\theta = N^2$$

or

$$2M\cos\theta = r + (M^2 - N^2)\,r^{-1}$$

But if x and y are the coördinates on the skew mercator $x = \theta$, $y = \log r$.

Hence

$$\cos x = \frac{\operatorname{Cosh}(y - A)}{B}$$

To take account of the compression *exactly* is a problem I have not quite solved. First, to make the angles preserved, for r a complicated function has to be substituted. But then the geodesic line does not return into itself and consequently is not a circle on the stereographic and it is necessary to take such a function of $r\mathfrak{S}^{\theta i}$ as to render it a circle. This is possible, and the approximate solution is easy. The exact solution would require some days of calm leisure. As it would evidently involve elliptic functions, the whole earth would be represented ∞^2 times but probably not in rectangular checkers. Nor is it quite clear to me that the scale of the map would be uniform over the whole of the geodesic taken as central and reduced to a straight line.

My notion about Father's geometry would be to make an introduction of three chapters. The first on the nature of mathematical reasoning and how to perform it. The second on projective geometry, especially perspective. The third on the metrical properties of space and the justification of the idea that an angle is a difference of directions.

I don't know that the body of the book would require any correction.

The thing would be to preserve its leading characteristic, which is to lay hold of and render definite our instinctive or natural ways of considering space, especially in reference to moving about in it.

I think, that done, the work would be a master-piece and would be much admired. Its sale would be good, too. Plimpton is anxious to have us do it. You and I should do it together, n'est-ce pas? What say you?

I don't know that I would not reform and enlarge the part relating to plane problems and append a chapter on Descriptive Geometry, and perhaps an appendix on Modern Geometry.

Rambling back to my chart again, the formulae are like this. For the meridians, I take Θ and H and η subsidiary quantities and have

$$\tan\Theta = \cos 51^\circ \tan \text{longitude}$$
$$H = \sin\Theta \tan 51^\circ$$
$$\sin(x - \Theta) = H\tan\eta = H \operatorname{Sinh} y$$
$$y = \frac{180^\circ}{M\pi}\log\tan\left(45^\circ + \frac{\eta}{2}\right)$$

For the parallels I have the subsidiary quantity ♈

$$♈ = -\frac{1}{M}\log(\cos 51^\circ + \sin \text{latitude})$$

$$\operatorname{Cosh}(y - ♈) = \sin 51^\circ \cos x = \sec\eta$$

$$y = \frac{180^\circ}{M\pi}\left\{\log(\cos 51^\circ + \sin \text{lat}) + \log\tan\left(45^\circ + \frac{\eta}{2}\right)\right\}$$

The approximate solution taking account of compression. In order to preserve the angles, all the books say that on the polar stereographic you must multiply r by

$$\left(\frac{1 + \varepsilon\sin\varphi}{1 - \varepsilon\sin\varphi}\right)^{\varepsilon/2}$$

where ε is the eccentricity of a meridian $= 0.08227$. If we make

$$\sin\varphi' - \varepsilon\sin\varphi$$

this is to say that in place of

$$r = \tan\left(45^\circ - \frac{\varphi}{2}\right)$$

we have

$$r = \tan^{\varepsilon}\left(45^\circ + \frac{\varphi'}{2}\right)\tan\left(45^\circ - \frac{\varphi}{2}\right)$$

latitude south, not while it traverses 180° of longitude, but before it has quite got round so far, so that if the geodetic be indefinitely prolonged it will from a grating of its own cycles over all that part of the globe intermediate between its extreme latitudes, or more properly speaking a reticulation.

"5. In the Skew Mercator one oblique geodetic, which we may designate as A, is represented by a straight line, say a, on the plane, and is so represented that to equal distances on A correspond equal distances on a. Moreover, letting M represent any geodetic line on the ellipsoid (or other surface) which cuts A at right angles, this geodetic (and every one like it) is to be represented on the plane by a straight line, m, cutting a at right angles at the point that represents the point at which M cuts A at right angles. And the scale of the map is at every point the same in all directions."

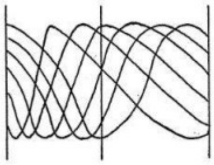

Fig. 1

D. FORMULAE AND TABLES FOR CONSTRUCTING TWO DIFFERENT CONFORMAL MAP-PROJECTIONS SUITABLE TO THE EXHIBITION OF ALL THE TERRITORY AND POSSESSIONS OF THE UNITED STATES OF AMERICA (1350)

EXPLANATION OF THE PROJECTIONS

A map is to be conceived as showing a part on the whole of a vertical projection of the Earth's surface as projected vertically (i.e. above the plane of gravity) upon the mean sea-level.

A "Conformal," or "Orthomorphic" map-projection is usually defined as one that shows all angles in their true magnitudes at their vertices; but perhaps a clearer definition is that one that has the same scale in all directions. This is not exactly true of any other kind of map nor even approximately so if the region represented covers as considerable part of the Earth's surface, as the possessions of the United States.

The form of the mean surface of the sea must, indisputably, always be regarded, by the cartographer, as ellipsoidal. For though departures from such a form may be made out by geodesists with more or less plausibility they will certainly be so minute, so irregular, and so liable to revision, that any correction of the distance between two points on such a basis will be far preferably made by means [of] an independent table of double entry rather than by a modification of the map-projection.

Everybody who has ever taken a course in Coördinatal Geometry (less precisely called *Analytical*, or *Algebraic Geometry*) may recall that the equation of an ellipsoid is

$$\left(\frac{x}{X}\right)^2 + \left(\frac{y}{Y}\right)^2 + \left(\frac{z}{Z}\right)^2 = 1,$$

where X, Y, Z are X the largest, Z smallest semidiameters, and Y is the semidiameter perpendicular to both the others; while x, y, z are the rectangular coordinates of any one point of the surface measured parallel respectively to the semi-axes X, Y, Z, from the centre of the ellipsoid.

In the case of the mean surface of the sea Z will be the distance from the centre of the earth to either pole (often called its polar *semi-axis*). As to X and Y, the only proper assumption in the present state of geodetical measurement is that they are equal. For the only attempt that I know of to determine their longitudes and relative length, that of Colonel Clarke, in 1878, being, I presume, deemed premature. Anyway he made their difference of length to be only 1524 feet, or say $\frac{2}{7}$ of a mile, each being about 4000 miles in length. If their difference can be neglected in geodetic calculations, it certainly ought to be neglected for any map on paper, which is liable, however great be the care taken of it, to shrink and swell unequally in different directions; while to take account of variations of the radius of the equator would multiply the number of oblique spherical triangles that have to be solved in making the table of coördinates for the projection, — would require this number of calculations, I say, to be multiplied by the number of meridians laid down in 45° of longitude. The labor (other complications being considered) would be from ten to twenty times as great, to save an error of one part in fourteen thousand. Now, who is so simple as to attempt to make measurement on paper accurate to one part in fourteen thousand?

E. [MATHEMATICAL NOTES ON THE SHAPE OF THE EARTH] (1071)

It is generally assumed that the earth's mean sea surface is an ellipsoid of revolution, and it has been generally regarded as proved that, neglecting tidal effects, such a surface might belong to a body in stable equilibrium. Poincaré, however, seems to have proved, or nearly proved, that other forms of stable equilibrium were possible. There is an interesting popular account of the matter in Sir George Howard Darwin's book entitled "*The Tides*," 1898 (Houghton, Mifflin, & Co.). In connection with this Sir G. H. Darwin has put forth the theory that our large moon, which is unparalleled in the solar system, originated by separation from the earth before the latter had acquired a stable equilibrium. This seems a reasonable conjecture, and if it be true, the earth probably must, after the separation, have gradually assumed a form approximately that of an ellipsoid with three unequal axes. Col. A. R. Clarke, an eminent geodesist, had undertaken (rather prematurely, perhaps, I think) to compute in 1878 what would be the ellipsoid, not necessarily one of revolution, that would best satisfy the geodetic observations. His result was that the difference between the greatest and least semi-diameters of the equator was only $\frac{1}{13721.2}$ part of the former being only about a quarter of a mile, and this supposed form did not satisfy the observations very much better than an ellipsoid of revolution would.

If the three semiaxes of an ellipsoid be denoted by X, Y, Z, and if x, y, z, be the sides of a rectangular parallelopipedon respectively parallel to those axes and having one vertex at the centre and the opposite vertex at any point on the ellipsoidal [at the sea-level] surface, these quantities will satisfy the equation

$$(1) \qquad \left(\frac{x}{X}\right)^2 + \left(\frac{y}{Y}\right)^2 + \left(\frac{z}{Z}\right)^2 = 1$$

which equation may be regarded as a definition of what is meant by an

data

- Charles... New Elements of Mathematics Volume III ... 1976.pdf

Pág. 495 (268) 9

Map Projection

Pág. 497 (269) Appendix No. 15

A Quincuncial Projection of the Sphere

Orthomorphic, or conform projection formed by transforming the stereographic projection, with a pole at infinity, by means of an elliptic function.

For that purpose, l being the latitude, and θ the longitude, we put —

$$\cos^2\varphi = \frac{\sqrt{1-\cos^2 l\cos^2\theta} - \sin l}{1+\sqrt{1-\cos^2 l\cos^2\theta}},$$

and then $\frac{1}{2}F\varphi$ is the value of one the rectangular co-ordinates of the point on the new projection.

This is the same as taking —

$\cos \operatorname{am}(x+y\sqrt{-1})$ (angle of mod. $= 45^\circ$) $=$ $\tan\frac{p}{2}(\cos\theta + \sin\theta\sqrt{-1})$, where

Jandaia

where x and y are the coordinates on the new projection, p is the north polar distance!.

Pág. 507(274) E.

Mathematical Notes on the Shape of the Earth (1071).

Equations...

Pág. 515(278) — Varying scale of Skew Mercator.
(enviesamento Mercator)

Fig. 1

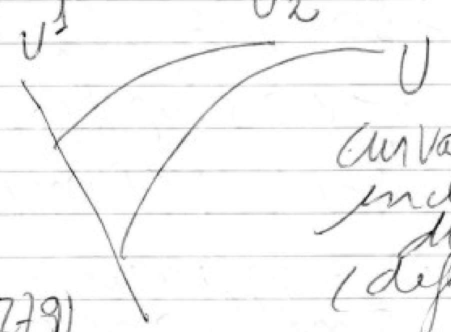

curvas com inclinações diferentes (deformações)

Leste:

Pág. 516(279)

Hence if rectangular geodesics are drawn between great circles whose northernmost (cujo norte extremo) points are constant, the difference of the squares of the sines of the reduced latitudes of their extremities is constant, or

(" "

In the case of the mean surface of the sea Z will be the distance from the centre of the earth to either pole (often called its polar *semi-axis*). As to X and Y, the only proper assumption in the present state of geodetical measurement is that they are equal. For the only attempt that I know of to determine their longitudes and relative length, that of Colonel Clarke, in 1878, being, I presume, deemed premature. Anyway he made their difference of length to be only 1524 feet, or say $\frac{2}{7}$ of a mile, each being about 4000 miles in length. If their difference can be neglected in geodetic calculations, it certainly ought to be neglected for any map on paper, which is liable, however great be the care taken of it, to shrink and swell unequally in different directions; while to take account of variations of the radius of the equator would multiply the number of oblique spherical triangles that have to be solved in making the table of coördinates for the projection, — would require this number of calculations, I say, to be multiplied by the number of meridians laid down in 45° of longitude. The labor (other complications being considered) would be from ten to twenty times as great, to save an error of one part in fourteen thousand. Now, who is so simple as to attempt to make measurement on paper accurate to one part in fourteen thousand?

E. [MATHEMATICAL NOTES ON THE SHAPE OF THE EARTH] (1071)

It is generally assumed that the earth's mean sea surface is an ellipsoid of revolution, and it has been generally regarded as proved that, neglecting tidal effects, such a surface might belong to a body in stable equilibrium. Poincaré, however, seems to have proved, or nearly proved, that other forms of stable equilibrium were possible. There is an interesting popular account of the matter in Sir George Howard Darwin's book entitled "*The Tides*," 1898 (Houghton, Mifflin, & Co.). In connection with this Sir G. H. Darwin has put forth the theory that our large moon, which is unparalleled in the solar system, originated by separation from the earth before the latter had acquired a stable equilibrium. This seems a reasonable conjecture, and if it be true, the earth probably must, after the separation, have gradually assumed a form approximately that of an ellipsoid with three unequal axes. Col. A. R. Clarke, an eminent geodesist, had undertaken (rather prematurely, perhaps, I think) to compute in 1878 what would be the ellipsoid, not necessarily one of revolution, that would best satisfy the geodetic observations. His result was that the difference between the greatest and least semi-diameters of the equator was only $\frac{1}{13721.2}$ part of the former being only about a quarter of a mile, and this supposed form did not satisfy the observations very much better than an ellipsoid of revolution would.

If the three semiaxes of an ellipsoid be denoted by X, Y, Z, and if x, y, z, be the sides of a rectangular parallelopipedon respectively parallel to those axes and having one vertex at the centre and the opposite vertex at any point on the ellipsoidal [at the sea-level] surface, these quantities will satisfy the equation

$$(1) \qquad \left(\frac{x}{X}\right)^2 + \left(\frac{y}{Y}\right)^2 + \left(\frac{z}{Z}\right)^2 = 1$$

which equation may be regarded as a definition of what is meant by an

ellipsoid. If two of the semi-axes, say X and Y are equal, we may replace them by the single letter, R, when the equation will become

$$(2) \qquad \frac{x^2+y^2}{R^2} + \frac{z^2}{Z^2} = 1.$$

If we are to regard this as representing the mean surface of the sea, the arduous and triumphant investigation of Mr. J. F. Hayford, which has rendered all other attempts to define the form of the mean sea-level obsolete (as far, at least, as the part of it covered by the United States is concerned), and which defines not the mean sea surface itself but a surface [with]

(3) R, the semidiameter of the equator

and

(4) (Z being the polar semiaxis) $R = \left(1 + \frac{1}{294}\right) Z.$

But instead of z, the distance of any point of the surface from the equator, it will be more convenient to use a differently shaped zed, $\mathscr{Z}$, which is measured not from the plane of the equator, but from a parallel plane touching the ellipsoid at the North Pole, so that

$$(5) \qquad z = \mathscr{Z} - Z$$

while instead of $x^2 + y^2$ I will write r, the distance of any point of the surface from the polar axis. With these substitutions, equation (2) becomes

$$(6) \qquad r^2 = \left(1 + \frac{1}{294}\right)^2 \quad (2Z\mathscr{Z} - \mathscr{Z}^2)$$

$$(7) \qquad r = \left(1 + \frac{1}{294}\right) \sqrt{2Z\mathscr{Z} - \mathscr{Z}^2}.$$

Differentiating (7) we get

$$(8) \qquad dr = \left(1 + \frac{1}{294}\right) \frac{Z - \mathscr{Z}}{\sqrt{2Z\mathscr{Z} - \mathscr{Z}^2}} (d\mathscr{Z})$$

whence

$$(9) \qquad (dr)^2 = \left(1 + \frac{1}{146\frac{2}{3}}\right)\left(\frac{Z^2}{2Z\mathscr{Z} - \mathscr{Z}^2} - 1\right)(d\mathscr{Z})^2$$

and adding $(d\mathscr{Z})^2$

$$(10) \qquad (dr)^2 + (d\mathscr{Z})^2 = \frac{1}{146\frac{2}{3}}\left(147\tfrac{2}{3}\,\frac{Z^2}{2Z\mathscr{Z} - \mathscr{Z}^2} - 1\right)(d\mathscr{Z})^2.$$

[Another version of this problem follows]

$$(1) \qquad \left(\frac{x}{X}\right)^2 + \left(\frac{y}{Y}\right)^2 + \left(\frac{z}{Z}\right)^2 = 1.$$

(2) If $X = Y, \dfrac{(x^2+y^2)}{X^2} + \dfrac{z^2}{Z^2} = 1.$

If we denote $x^2 + y^2$ by r^2 and adopt Hayford's value of the compression of the earth

(2) becomes

$$(3) \qquad (0.99323 \pm V)\, r^2 + z^2 = Z^2$$

where $V = 10^{-5}$

Changing the origin to the N pole, by putting $z = \mathscr{Z} - Z$

$$(4) \qquad (0.99323 \pm V) r^2 = 2Z\mathscr{Z} - \mathscr{Z}^2$$

or

$$(5) \qquad (0.99661 \pm \tfrac{1}{2}V) r = \sqrt{2\mathscr{Z}Z - \mathscr{Z}^2}$$

or

$$(6) \qquad r = (1.003401 \pm .6V) \sqrt{2\mathscr{Z}Z - \mathscr{Z}^2}$$

Differentiating, we get

$$(7) \qquad dr = (1.003401 \pm .6V) \frac{\mathscr{Z}(Z - \mathscr{Z})}{\sqrt{2\mathscr{Z}Z - \mathscr{Z}^2}} d\mathscr{Z}$$

squaring

$$(8) \qquad (dr)^2 = (1.006814 \pm 1.1V) \left(\frac{Z^2}{2\mathscr{Z}Z - \mathscr{Z}^2} - 1\right)(d\mathscr{Z})^2$$

$$\begin{aligned}(9) \qquad (d\mathscr{Z})^2 + (dr)^2 &= (1.006814 \pm 1.1V) \frac{Z^2}{2\mathscr{Z}Z - \mathscr{Z}^2} (d\mathscr{Z})^2 - \\ &\qquad - (0.006814 \pm 1.1V) \\ &= \left[\frac{Z^2}{2\mathscr{Z}Z - \mathscr{Z}^2} + (0.006814 \pm 1.1V) \frac{(Z - \mathscr{Z})^2}{2\mathscr{Z}Z - \mathscr{Z}^2}\right](d\mathscr{Z})^2 \\ &= \frac{Z^2}{2\mathscr{Z}Z - \mathscr{Z}^2}\left[1 + 0.006814 \pm 1.1V)\left(1 - \frac{\mathscr{Z}}{Z}\right)^2\right](d\mathscr{Z})^2\end{aligned}$$

which will give nearly

$$(10) \qquad \sqrt{(d\mathscr{Z})^2 + (dr)^2} = \frac{Z}{\sqrt{2\mathscr{Z}Z - \mathscr{Z}^2}}\left[1 + .003407\left(1 - \frac{\mathscr{Z}}{Z}\right)^2\right] d\mathscr{Z}$$

F. NEW THEORY OF THE SKEW MERCATOR (1353)

1894 March 27

Suppose a geodesic line to be represented by a straight line and equal arcs by equal lengths, and that the projection is conform. The straight lines perpendicular to this central line will represent geodesics. For if from neighboring points on a geodesic, geodesic perpendiculars of equal length be erected (on the same side of the first) the junction of their extremities will by a known theorem be perpendicular to them. Therefore the angles will be preserved in making these perpendicular straight lines represent geodesics.

Therefore the poles will not lie on any of them except those that are meridians.[1]

THE MATHEMATICAL THEORY

§0. It is a comfort to me, who am reading mathematics constantly, to get a subject treated *ab ovo* provided the elementary part is abridged, even if I am quite familiar with it. Hence, I hope the navigators for whom this exposition is intended will not object to my going into elementary explanations.

§1. A map is a representation of a round surface upon a plane, by which we mean that,

1st. In a general way, *points* upon the round surface are represented by *points* on the map-plane, and *vice versa*,

2nd. *Continuous series* of points on the round surface are represented by *continuous series* of points on the map-plane.

Let us first imagine that

(1) Every point on the round surface is represented on the map-plane;

(2) Every point on the map-plane represents a point on the round surface;

(3) No point on the round surface is represented by more than a single point on the map-plane;

(4) No point on the map-plane represents more than one point on the round surface.

These things being assumed, imagine that round any point on the map-plane is drawn an oval, by which I mean a curve returning into itself without cutting itself. This oval will represent an oval on the round surface. Let the oval on the map-plane be itself enclosed in another oval. This second oval will represent an oval on the round surface not intersecting the first oval and not on the same [side] of the latter as the [origin point] represented by the point originally taken on the map-plane. Imagine an endless series [of] ovals to be drawn on the map-plane each enclosing the previous one and not approximating closer and closer to any finite oval. These will represent an endless series of ovals on the round surface enclosing or enclosed by one another not approximating to any oval. Then, these ovals on the round must either approximate to a point or a limited line. This point, or limited line, is represented by the infinitely distant parts of the plane. We may hold ourselves at liberty to assume anything we please about them, provided it involves no contradictory result. Hence we must either assume for our purpose that the parts of the plane infinitely distant are one point, or that they are a limited line. I shall not examine the latter assumption, but shall assume that for our purpose the infinitely distant parts of the plane are to be regarded as a single point.

[ADDITIONAL] MATHEMATICAL THEORY

§1. The position of a point may be expressed in rectangular coördinates, x and y, or polar coördinates r and θ having the same origin. But I reckon θ as on a clock from vertically up round to the right.

For the purposes of the conform projections, or those which preserve the angles, it is convenient to make one imaginary expression define the position of a point. Namely

$$z = re^{\theta i} = r\cos\theta + r\sin\theta . i \ = y + xi.$$

§2. If a projection have these two properties

1st, that for each place, there is some right angle in nature (I mean, the surface represented, earth, heavens or what) having its vertex there which is represented by a right angle on the map,

2nd, that along the two legs of such [a] right angle the scales are the same,

then the projection preserves all the angles. For let XOY be such a right angle in nature, represented by the right angle $X'O'Y'$ on the map. Imagine the portion seen to be infinitesimal, so that there shall be no curves. Then, completing the rectangle [to a fourth vertex Z and Z'], if $\frac{OY}{O'Y'} = \frac{OX}{O'X'}$ that is to say, if the scale of the map is the same along the two legs, it is plain that the angles $ZOY = Z'O'Y'$. Hence, obviously, all angles having their vertices at O are represented by equal angles.

---

[1] In the worksheets on the geodesic Peirce makes use of Hoüel (Recueil, p. lxiii).

THE SKEW MERCATOR PROJECTION

It is well known that the polar stereographic projection of the sphere is produced by making

$$r = R \tan\left(45° - \frac{\varphi}{2}\right)$$

where r is the radius vector on the map from the pole as origin, and φ is the latitude [λ is the longitude so reckoned].

Upon this projection, every circle is projected into a circle.

It is also known that in order to preserve the angles taking account of the compression of the earth we may put

$$r = \left(\frac{1 - \varepsilon \sin \varphi}{1 + \varepsilon \sin \varphi}\right)^{\frac{\varepsilon}{2}} \tan\left(45° - \frac{\varphi}{2}\right)$$

where ε is the eccentricity of a meridian $= 0.08227$.* (*For (Clarke's *Geodesy*, p. 104) the radius of curvature of a meridian is $\frac{a(1-\varepsilon^2)}{(1-\varepsilon^2 \sin^2\varphi)^{3/2}}$ and that of a normal section perpendicular to the meridian is $\frac{a}{(1-\varepsilon^2 \sin^2 \varphi)^{\frac{1}{2}}}$.)

But upon this projection geodesic lines no longer appear as circles. The reason is that the difference of longitude between the intersection of a geodesic line with the equator and its northernmost point is not 90°, but is

Maximum Latitude	Diff Long
18°	90°-17′
36°	90°-15′
54°	90°-12′
72°	90°- 6′

A change $\delta\lambda$ in longitude, in passing along the outline of the last section makes a rotation of the normal equal to $\cos\varphi \cdot \delta\lambda$.[2] Hence, if $\frac{\delta s}{\delta \lambda}$ is constant for all latitudes (as for a regular Mercator's chart)

$$\frac{\delta s}{\delta \varphi} = \frac{1}{\cos\varphi} \frac{1-\varepsilon^2}{1-\varepsilon^2 \sin^2\varphi}$$

$$\frac{\delta s}{\delta \varphi} - \frac{1}{\cos\varphi} = \frac{1}{\cos\varphi}\left(\frac{1-\varepsilon^2}{1-\varepsilon^2\sin^2\varphi} - 1\right)$$

$$= \frac{1}{\cos\varphi}\left(\frac{-\varepsilon^2 \cos^2\varphi}{1-\varepsilon^2\sin^2\varphi}\right)$$

$$= \frac{-\varepsilon^2\cos\varphi}{1-\varepsilon^2\sin^2\varphi}$$

$$\frac{\delta s}{\delta\varphi} = \frac{1}{\cos\varphi} - \varepsilon^2 \frac{\cos\varphi}{1-\varepsilon^2\sin^2\varphi}$$

Now $\quad D\varphi \log \frac{1-\varepsilon \sin\varphi}{1+\varepsilon\sin\varphi} = \frac{-\varepsilon\cos\varphi}{1-\varepsilon\sin\varphi} - \frac{\varepsilon\cos\varphi}{1+\varepsilon\sin\varphi} =$

$$= -\varepsilon\cos\varphi \frac{2}{1-\varepsilon^2\sin^2\varphi}$$

And $\quad D\varphi \log\tan\left(45° + \frac{\varphi}{2}\right)$

$$= \frac{1}{2\tan\left(45° + \frac{\varphi}{2}\right)} \frac{1}{\cos^2\left(45° + \frac{\varphi}{2}\right)} = \frac{1}{\sin(90° + \varphi)} = \frac{1}{\cos\varphi}$$

Hence

$$s = \log\tan\left(45° + \frac{\varphi}{2}\right) + \frac{\varepsilon}{2}\log\frac{1-\varepsilon\sin\varphi}{1+\varepsilon\sin\varphi}$$

(This agrees with Craig's Projections, p. 65.)[3]

Hence we must have an angular motion along that section, equal to $\cos\varphi \cdot \delta\lambda$, represented in the map by a line whose ratio to $d\varphi$ is $\frac{1-\varepsilon^2}{1-\varepsilon^2\sin^2\varphi}$.

---

[2] On another fragment Peirce writes, "Hence we must have an angular motion along that section equal to cos φ.dλ, represented on the map by a line whose ratio to dφ is $\frac{1-\varepsilon^2}{1-\varepsilon^2 \sin^2\varphi}$."

[3] Craig was a member of the Mathematics Department at the Johns Hopkins University during Peirce's association with that institution and was later Editor-in-Chief of the *American Journal of Mathematics*. He had also served in the Coast and Geodetic Survey.

THE SKEW MERCATOR

On the Stereographic, the equation of a meridian is

$$(\xi - A)^2 + (\eta - B)^2 = C^2$$

When $\xi = 0$

$$\eta_1 = B + \sqrt{C^2 - A^2} = S \tan 25^\circ \cdot 5$$
$$\eta_2 = B - \sqrt{C^2 - A^2} = S \cot 25^\circ \cdot 5$$
$$(\xi - A)\,d\xi + (\eta - B)\,d\eta = 0$$

When $\xi = 0$ $\quad \frac{d\xi}{d\eta} = \frac{\sqrt{C^2 - A^2}}{A} = \tan \text{long}$

$$B = -S \cot 51^\circ$$
$$\sqrt{C^2 - A^2} = S \operatorname{cosec} 51^\circ$$
$$A = S \operatorname{cosec} 51^\circ \cot \text{long} \qquad C = S \operatorname{cosec} 51^\circ \operatorname{cosec} \text{long}$$

Putting $\xi = r \sin\theta \qquad \eta = r\cos\theta$
the equation is

$$(r \sin\theta - S \operatorname{cosec} 51^\circ \cot \text{long})^2 + (r\cos\theta + S \cot 51^\circ)^2 = S^2 \operatorname{cosec}^2 51^\circ \operatorname{cosec}^2 \text{long}$$
$$S^{-1}r - Sr^{-1} = 2(\sin\theta \operatorname{cosec} 51^\circ \cot \text{long} - \cos\theta \cot 51^\circ)$$
$$= 2\,\frac{\sin\theta \cot \text{long} - \cos\theta \cos 51^\circ}{\sin 51^\circ}$$

Putting $\tan\Theta = \cos 51^\circ \tan \text{long}$ and $E = \frac{\sqrt{\cot^2 \text{long} + \cos^2 51^\circ}}{\sin 51^\circ}$

$$S^{-1}r - Sr^{-1} = 2E \sin(\theta - \Theta)$$

Now $r = \epsilon^y \qquad \theta = x$

Put $\qquad S = \epsilon^Y$

$$\frac{r}{S} = \epsilon^{y-Y} = \operatorname{Cosh}(y - Y) + \operatorname{Sinh}(y - Y)$$
$$-\frac{S}{r} = \epsilon^{Y-y} = -\operatorname{Cosh}(y - Y) + \operatorname{Sinh}(y - Y)$$
$$\operatorname{Sinh}(y - Y) = E \sin(\theta - \Theta)$$

On the Stereographic the equation of a parallel is

$$\xi^2 + (\eta - \beta)^2 = \gamma^2$$

When $\xi = 0$

$$\eta_1 = \beta + \gamma = S \tan\frac{90^\circ - \text{lat} + 51^\circ}{2}$$
$$\eta_2 = \beta - \gamma = -S \tan\frac{90^\circ - \text{lat} - 51^\circ}{2} = S \tan\frac{51^\circ - 90^\circ + \text{lat}}{2}$$

$$\beta = \frac{S \sin 51^\circ}{\cos 51^\circ + \sin \text{lat}} \qquad \gamma = \frac{S \cos \text{lat}}{\cos 51^\circ + \sin \text{lat}}$$
$$r^2 \sin^2\theta + \left(r\cos\theta - S\,\frac{\sin 51^\circ}{\cos 51^\circ + \sin \text{lat}}\right)^2 = S^2 \frac{\cos^2 \text{lat}}{(\cos 51^\circ + \sin \text{lat})^2}$$

[Incomplete]

VARYING SCALE OF SKEW MERCATOR

a Azimuth of geodesic
u Reduced lat
U ditto of northernmost points of geodesic
ε Eccentricity of meridian

$$\sin\alpha = \frac{\cos U}{\cos u}$$
$$\alpha' = 90^\circ - \alpha$$
$$\sin\alpha' = \sqrt{1 - \frac{\cos^2 U}{\cos^2 u}}$$
$$\cos U' = \sqrt{\cos^2 u - \cos^2 U}$$
$$\sin\alpha_2' = \frac{\sqrt{\cos^2 u - \cos^2 U}}{\cos u_2}$$
$$\alpha_2 = 90^\circ - \alpha_2'$$
$$\sin\alpha_2 = \frac{\sqrt{\cos^2 u_2 - \cos^2 u + \cos^2 U}}{\cos u_2}$$
$$\cos U'' = \sqrt{\cos^2 u_2 - \cos^2 u + \cos^2 U}$$
$$\therefore \sin^2 U'' = \sin^2 U - \sin^2 u + \sin^2 u_2 \quad \text{[Fig. 1]}$$

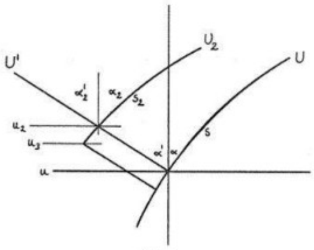

Fig. 1

Hence if rectangular geodesics are drawn between great circles whose northernmost points are constant, the difference of the squares of the sines of the reduced latitudes of their extremities is constant, or

$$\sin u \cdot d \sin u = \sin u_2 \, d \sin u_2$$

Let two such be drawn infinitely near, and compare the infinitestimal segments

$$ds = a\frac{\sqrt{1-\varepsilon^2\cos^2 u}}{\sqrt{1-\dfrac{\cos^2 U}{\cos^2 u}}}\,du = a\frac{\sqrt{1-\varepsilon^2\cos^2 u}}{\sqrt{\sin^2 U-\sin^2 u}}\,d\sin u$$

$$= a\frac{\sqrt{1-\varepsilon^2+\varepsilon^2\sin^2 u}}{\sqrt{\sin^2 U-\sin^2 u}}\,d\sin u$$

$$ds_2 = a\frac{\sqrt{1-\varepsilon^2+\varepsilon^2\sin^2 u_2}}{\sqrt{\sin^2 U_2-\sin^2 u_2}}\,d\sin u_2$$

$$= a\frac{\sqrt{1-\varepsilon^2+\varepsilon^2\sin^2 u_2}}{\sqrt{\sin^2 U-\sin^2 u}}\,d\sin u_2$$

$$= a\frac{\sqrt{1-\varepsilon^2+\varepsilon^2\sin^2 u_2}}{\sqrt{\sin^2 U-\sin^2 u}}\,\frac{\sin u}{\sin u_2}\,d\sin u$$

$$\frac{ds_2}{ds} = \frac{\sqrt{1-\varepsilon^2+\varepsilon^2\sin^2 u}}{\sin u_2}\cdot\frac{\sin u}{\sqrt{1-\varepsilon^2+\varepsilon^2\sin^2 u}}$$

$$= \frac{\sin \mathrm{lat}}{\sin \mathrm{lat}_2}$$

Though this is not what was wanted, it is a curious result. It is also puzzling.

[C. S. PEIRCE'S SKEW MERCATOR I]

C. S. Peirce's "Skew Mercator" is a projection *of the sphere* (i.e. not corrected for the compression of the Earth, since in measuring distances on a map of any considerable fraction of the earth's surface it will always be necessary to use numerical computation, either to assure oneself that the number read from the scale used sufficiently represents what one ought to mean for the special purpose in view, by the "distance" between two points on the earth's surface, or else to deduce this "distance" from the measure on the map) upon the plane of the great circle passing through latitude 60°, this projection being of the same kind as that of the ordinary Mercator's Projection, which ought to be a conform projection *of the sphere*.

But a map of a large fraction of the Earth's surface ought not to be regarded as intended to measure upon closer than to the first significant figure of our customary numerical notation, but rather to be viewed with the unaided eye. Now if any figure is drawn upon a plane surface held square before the eyes, and this figure is then turned in its own plane through a right angle, so that what was, before the turning, its height becomes its width right and left and *vice versa*, its *apparent* shape to a spectator not specially trained, will undergo a change ascribed to the greater effort required to turn our eyes up and down than to turn them right or left. But this change of apparent shape is not found to be a great inconvenience. Now that apparent change is just about double the exaggeration of [the] linear scale of Mercator's chart for a small area at 23° of latitude compared with a like object on the equator. Thus, to use the phrases of conscientious parlance, the accuracy of Mercator's chart within say 30° of the equator is truly astonishing, as compared with any visual aspect of it from without.

Now it happens that all parts and possessions of the United States of America, from Puerto Rico to Alaska and straight on to the Philippines lie within that distance of a Great Circle excepting Hawaii on one side and the Panama Canal on the other; and these are not very far outside of 30° from that central circle.

[C. S. PEIRCE'S SKEW MERCATOR II]

... If the sum of the orders of North poles is not equal to that of South poles, the point at infinity must make up the deficiency of the lesser sum. We, then, find a function which has these zeros and infinities; and finally we are free to multiply by a constant which serves to fix the latitude and longitude of any other selected point on the map. In this way, formulae can be obtained with great facility. For instance, suppose we wish the formulae for the stereographic projection when the plane of projection is tangent to a certain place. We calculate r and θ for that place, and denote them by r_0 and θ_0. Then we wish a function of $re^{\theta i}$ which shall vanish for $re^{\theta i} = r_0 e^{\theta_0 i}$ and shall become infinite for $re^{\theta i} = \frac{1}{r_0} e^{-\theta_0 i}$. This gives ($r_1$ and θ_1 being the coördinates on the new map)

$$r_1 e^{\theta_1 i} = C \frac{re^{\theta i} - r_0 e^{\theta_0 i}}{re^{\theta i} - \frac{1}{r_0} e^{-\theta_0 i}}$$

In the case of Mercator's projection the North pole is at $+\infty$ and the South pole at $-\infty$. Obviously, therefore,

$$x_1 + y_1 i = \log (\mathrm{re}^{\theta i})$$

or

$$x_1 = \log r$$
$$y_1 = \theta.$$

A map may in the first place have a finite number of North and South poles. In this case, it is connected with an algebraic function. If the number of poles is infinite, they cannot be specified singly and only the law of their arrangement can be described. The map may show the whole earth repeated over and over in stripes. These stripes may be all alike or they may be broader and broader indefinitely in one direction. Each stripe may show the whole earth but once, or several times, or even an infinite number of times. An example of the last case is the map resulting from the function sin · sin $a\varphi$, of which a sketch is given.[1] The map instead of being striped may be checkered. In this case, we have to do with a doubly periodic function. The pattern whether striped or checkered may be disturbed by irregularities, as in the case of the sum of a periodic function and an algebraic one. Finally, the map may have a pattern neither striped nor checkered. Thus, the whole earth might be represented in concentric rings, as by the function

$$\prod_{m=1}^{m=\infty} \quad \frac{m-x}{m+x}$$

---

[1] No sketch accompanies this fragment. In another fragment Peirce writes: "The idea of the [Skew Mercator Map] is that since Mercator's Chart preserves the shapes and sizes of the globe within 40° of the equator sufficiently for comparison by the unaided and not excessively critical eye and since all the possessions of the United States (except Hawaii) lie within that angular distance of a certain great circle (namely one inclined 60° to the equator and cutting it in 35° West Longitude), it follows that a map which represents that great circle by a straight line, and preserves all angles unchanged, will give the best possible eye-view of those possessions."

G. [NOTES FOR] THE SKEW MERCATOR ON THE SPHERE (1584)

§1. Property of the Stereographic.

If a point in space X, Y, Z be projected from the origin upon the plane of $Z = 1$, its projection x, y will be

$$x = \frac{X}{Z} \qquad y = \frac{Y}{Z}$$

If we imagine the sphere through the origin

$$X^2 + Y^2 + (Z-1)^2 = 1$$

to be cut by the plane

$$\frac{X}{A} + \frac{Y}{B} + \frac{Z}{C} = 1$$

we have

$$\frac{x^2}{Z^2} + \frac{y^2}{Z^2} = -1 + \frac{2}{Z}$$
$$\frac{X}{AZ} + \frac{Y}{BZ} = \frac{1}{Z} - \frac{1}{C}$$

or eliminating Z

$$x^2 + y^2 + 1 = \frac{2}{A}x + \frac{2}{B}y + \frac{2}{C}$$

which is the equation of a circle.

If $C = \frac{1}{2}$, the plane passes through the centre of the sphere.

That the angles are preserved is evident from this: that the tangent plane to the sphere is as much inclined to the radius of projection as is the plane of the map.

In the case of the polar projection it is plain that

$$\frac{\sqrt{x^2 + y^2}}{Z} = \tan\left(45^\circ - \frac{\varphi}{2}\right)$$

where φ = latitude.

§2. The Skew Mercator

We transpose to another pole on the sphere, distant we will say by an angle a from the a [Fig. 1]. The first pole we call B the second C.

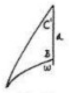

Fig. 1

Then the latitude $= 90° - C$
Longitude $= 180° - B$

reckoning longitude from the line of junction of the poles.

We might solve spherical triangle given $B = 180° - a$

$$a$$
$$c = 90° - \varphi$$

to find C and c.

We have two sides and included angle and could apply formulae of spherical trigonometry.

Namely,

$$\tan\tfrac{b}{2} = \sqrt{\frac{\dfrac{\tan\frac{a}{2}}{\tan\frac{c}{2}} + \dfrac{\tan\frac{c}{2}}{\tan\frac{a}{2}} + \frac{1}{2}\cos\omega}{\dfrac{1}{\tan\frac{a}{2}\tan\frac{c}{2}} + \tan\frac{a}{2}\tan\frac{c}{2} + \frac{1}{2}\cos\omega}}$$

$$\tan C = \frac{\sin\omega}{\tan\varphi\sin a + \cos a\cos\omega}$$

More simply

$$r'\,\mathfrak{S}^{\theta' i} = r\,\mathfrak{S}^{\theta i} + R = \sqrt{R^2 + r^2}\left(\frac{r(\cos\theta + R)}{\sqrt{R^2 + r}} + \frac{r\sin\theta}{\sqrt{R^2 + r^2}}\,i\right)$$

$$\log r' + \theta' i = \log\sqrt{R^2 + r^2} + \text{arc}\tan\frac{\tan\theta}{1 + \dfrac{R}{r}\sec\theta}$$

Geodetic Line
ω Longitude
q Shortest distance station to axis ret.
s arc of geodesic
a azimuth
$q^2 d\omega = C ds$

$C = q\sin a$
φ Latitude
u Reduced latitude
a Semiaxis major

$$\sin\varphi = \frac{\sin u}{\sqrt{1 - e^2\cos^2 u}}$$

$q = a\cos u$
e is eccentricity of meridian .08227
Compressed Earth

$$\frac{q^2}{a^2} + \frac{z^2}{b^2} = 1$$

To preserve angles on polar projection

$$\begin{aligned} r &= q + (b - z)\tan\left(45° + \frac{\varphi}{2}\right) \\ &= a\cos u + b(1 - \sin u)\tan\left(45° + \frac{\varphi}{2}\right) \\ &= \frac{a\cos\varphi}{\sqrt{1 - e^2\sin^2\varphi}} + b\frac{\sqrt{1 - e^2\sin^2\varphi} - \sqrt{1 - e^2}\sin\varphi}{\sqrt{1 - e^2\sin^2\varphi}} \end{aligned}$$

$$\tan\left[45° + \frac{\varphi}{2}\right] = \frac{\cos\varphi}{1 - \sin\varphi}$$

where x and y are the coordinates on the new projection, p is the north polar distance'.

Pág. 507(274) E.

Mathematical Notes on the Shape of the Earth (1071).

Equations...

Pág. 515(278) — Varying scale of Skew Mercator. (enviesamento Mercator)

Fig. 1

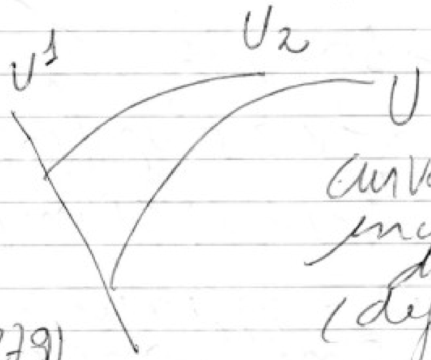

Curvas com inclinações diferentes (deformações)

Corte:

Pág. 516(279)

Hence if rectangular geodesics are drawn between great circles whose northernmost (mais norte extremo) points are constant, the difference of the squares of the sines of the reduced latitudes of their extremities is constant, or

$\sin u \cdot d \sin u =$

$\sin u_2 \, d \sin u_2$

Pág. 519 ($\bar{2}80$) G

The Skew Mercator on the Sphere (1584).

Pág. 687 ($\bar{3}64$) —

Non-Euclidean Geometry

Obs: Analisar tudo!!!

Pág. 710 (376) F.

On two Map-Projection of the Lobatschewskian Plane

Source: manuscript of the Smithsonian Institution.

Pág. 712 (377)

A Conformed Projection

Look: Fig. 2

The Projection is thus established (Hyperbolic stereographic projection).

We have an *apriori* or natural idea of space, which by some kind of evolution has come to be very closely in accord with observations. But we find in regard to our natural ideas, in general, that while they do accord in some measure with fact, they by no means do so to such a point that we can dispense with correcting them by comparison with observations.

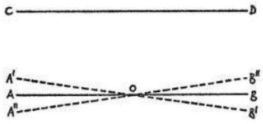

Fig. 1

Given a line *CD* and a point *O* [Fig. 1]. Our natural (Euclidean) notion is that

1st there is a line *AB* through *O* in the plane *OCD* which will not meet *CD* at any finite distance from *O*.

2nd that if any line *A'B'* or *A''B''* through *O* in the plane *OCD* be inclined by any finite angle, however small, to *AB*, it will meet *CD* at some finite distance from *O*.

Is this natural notion *exactly* true?

A. This is not certain.

B. We have no probable reason to believe it so.

C. We *never* can have positive evidence lending it any degree of likelihood. It *may* be disproved in the future.

D. It may be true, perhaps. But since the chance of this is as $1:\infty$ or $\frac{0}{1}$, the logical presumption is, and must ever remain, that it is not true.

E. If there is some influence in evolution tending to adapt the mind to nature, it would probably not be completed yet. And we find other natural ideas require correction. Why not this, too? Thus, there is some reason to think this natural idea is *not* exact.

F. I have a theory which fits all the facts as far as I can compare them, which would explain how the natural notion came to be so closely approximate as it is, and how space came to have the properties we find it has. According to this theory, this natural notion would not be exact.

To give room for the non-Euclidean geometry, it is sufficient to admit the first of these propositions.

Either the first or the second of the two natural propositions on page 1 of this article may be denied, giving two corresponding kinds of non-Euclidean geometry. Though neither of these is quite so easy as ordinary geometry, they can be made intelligible. For this purpose, it will suffice to consider plane geometry. The plane in which the figures lie must be regarded in perspective.

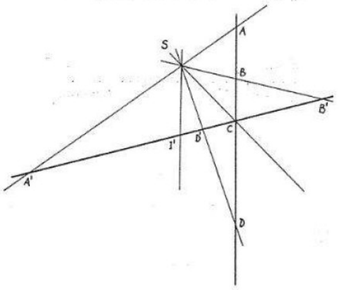

Fig. 2

Let *ABCD* be this plane, which I call the natural plane,* (* merely because so-called by writers on Perspective. Nothing to do with the "natural assumptions" of page 1) seen edgewise [Fig. 2]. Let *S* be the eye, or point of view, or centre of projection. From every point of the natural plane, rays, or straight lines proceed to the point of view, and are continued beyond it if necessary. If three points in the natural plane lie in one straight line, the rays from them through the point of view will lie in one plane. Let *A'B'* be the plane of the delineation or picture seen edgewise. It cuts all the rays through *S* in points, and so many of these rays as lie in one plane it cuts in a straight line; for the intersection of two planes is a straight line. The points in which this plane cuts those rays are the perspective delineations of the natural points, i.e. the corresponding points in the natural plane. We extend this to cases in which the point of view is between the natural and the delineated points.

It is readily seen that the delineation of a *point* is a *point*, and that to every point in the picture corresponds a point in the natural plane. And to a straight line in the natural plane corresponds a straight line in the picture. For the first straight line and the point of view lie in one plane, and the intersection of this plane by the plane of the picture is a straight line.

All this is just as true for the non-Euclidean as for the Euclidean geometry.

But now let us consider the parts of the natural plane which lie at an infinite distance and see how they look in perspective.

First suppose the natural, Euclidean, or parabolic geometry to be true. Then all the rays through *S* from infinitely distant parts of the plane, themselves lie in one plane. For let *SI'* be such a ray, then if *SI'* be turned about *S* the least bit out of the plane parallel to the natural plane it will cut the latter at a finite distance.

These rays through *S* from the infinitely distant parts of the natural plane, since they lie in one plane, will cut the plane of the picture in a straight line, called the vanishing line of the natural plane.

The delineations of any two parallel lines will cross one another on this vanishing line.

Next, suppose that the 2nd natural proposition on p. 1 [of this article] is false and that *SI'* may be tilted through a finite angle without cutting the natural plane at a finite distance.

Then, I will *state*, what there is no difficulty in proving, that the delineation of the infinitely distant parts of the natural plane will occupy a space on the picture bounded by a conic section.

The picture looks exactly as before, *only* that the *real distances* of certain parts which on the first assumption were finite, now become infinite.

The following two figures [Fig. 3 and Fig. 4] show the straight line which on the first assumption alone represents parts *really* infinitely distant, as well as the conic which on the second assumption bounds

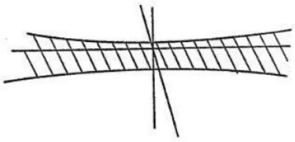

Fig. 3

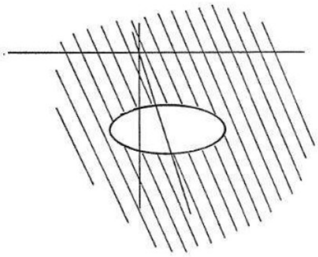

Fig. 4

the delineation of the really infinitely distant parts. With two lines parallel on the first assumption.

In this case, space is *limited*. But though *limited*, it is *immeasurable*. Imagine, for instance, that every kind of unit of linear measure, should shrink up as it was removed from a fixed centre, and perhaps differently according as it was radially or peripherally placed, so that it never could in any finite number of repetitions, get beyond a certain spherical surface. Then that surface would be at an infinite distance and no moving body could ever traverse it, for the distance moved over in a unit of time would be a unit of distance. But unless this linear unit were to shrink according to a peculiar law, the result would be that different parts of space would be unlike. That is to say, for example, if we were to draw the plane figure *ABCD*, by means of given lengths *AB*, *AC*, *AD*, *BC*, we should find the resulting length of *CD* to be different in different parts of space [see Fig. 5]. Now, geometers assume, perhaps with little reason, that

the fact that rigid bodies move about readily in space without change of proportions, shows that all the parts of space are alike.

With this condition, I will *state*, what is again readily proved that the law of the representation of equal distances and equal angles is in this second or hyperbolic geometry for the distances [see Fig. 6]. Let the conic be the "absolute," or the perspective representation boundary of infinitely distant regions. Let *AB* be a unit of distance. Through *AB* draw the line *IJ*, let *I'J'* be any other line. Draw straight lines through their intersections *II'* and *JJ'* with the absolute. Let *O* be the point where these lines intersect; from *O* draw rays to *A* and *B*. Then the distance *A'B'* cut off by these rays on the line *I'J'* is equal to *AB*.

The rule for angles is this. Let the conic be the absolute [Fig. 7]. Let *ACB* be a given angle. From *C* draw *CI* and *CJ* tangent to the absolute. Take any other vertex *C'*, and draw tangents *C'I* and *C'J* to the absolute. Through *I* the intersection of *CI* and *C'I* and *J* the intersection of *CJ* and *C'J* draw the straight line *IJ* cutting *AC* and *BC* in *A* and *B*. Then *AC'B* = *ACB*.

The third kind of geometry, which denies the 1st natural assumption supposes space to be *unlimited* but *finite*, that is, measurable. It is as if the linear unit so expanded in departing from a centre as to enable us to pass through what is naturally supposed to be at an infinite distance.

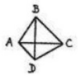

Fig. 5

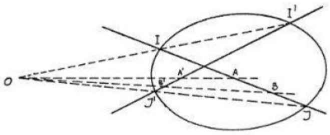

Fig. 6

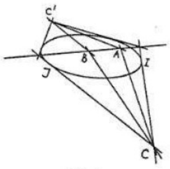

Fig. 7

B. REFLECTIONS ON NON-EUCLIDEAN GEOMETRY (118)

Although the ideas of imaginaries and of the absolute are the creation of algebra, yet I think it quite in the spirit of modern algebra, the ideas once gained, to treat them as synthetically as possible.

The non-Euclidean geometry treats exclusively of the projective relations of a certain individual surface, the absolute.

This surface is a quadric, so that every line either lies wholly within this surface, all its points and planes being elements of the surface, or else it has but two points and two planes in common with the surface. If either pair coalesces, both coalesce. Taking a plane and point incident the one on the other, in that plane through that point there are but two lines tangent to the surface, unless all such lines are so, and in that case, there are but two which lie wholly within the surface. Either of the pairs of lines spoken of may coalesce.

No definition of a plane is possible except that it is a linear surface. All transportation of a rigid body is representable by the result of a series of projections. These are subject only to the condition that they must not move the absolute away from itself or bring any other surface into tangency with it.

A pair of points or planes may have 9 singular positions in reference to the absolute.

1. It may coalesce on the absolute.
2. It may be distinct, but on a generator of the absolute.
3. It may be on the absolute but not on one generator.

4. One terminal only may lie on the absolute, and the other on a tangent line having that same point or plane of tangency.
5. One terminal may lie on the absolute, but the pair not on a tangent line.
6. Neither terminal lying on the absolute, the line may be a tangent line.
7. The pair may be harmonically situated with reference to the absolute.
8. One terminal and one common point of the line of the pair and the absolute may be harmonically situated with reference to the other two.
9. The pair may be equianharmonic relatively to the absolute.

In reference to the principles of measurement.

1. A spread which can be transported so as more than to cover a former place is infinite.
2. A spread which can be transported so as partly to fall short or partly extend beyond a former place but cannot more than cover a former place is finite.

Commensurable ratios of magnitude are determined by counting successive superpositions.

Incommensurable ratios depend on approximations which require us to take notice that of magnitudes not coincident one is more nearly so than another.

The section of the absolute by a plane, as well as any segment of it, is infinite. A sector is also infinite. But the area of a triangle inscribed in the absolute is finite. (In fact, a triangle cannot be infinite.) For it may partly fall short of and partly overlap a former position but cannot more than cover it. Moreover, all such triangles are plainly equal; since any one may be superposed on any other [Fig. 1]. All the angles of such triangles are zero. If such a triangle is separated into two by a

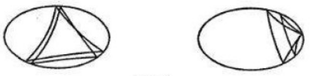

Fig. 1

line from the vertex to an opposite side, their angles are supplements. The areas of such triangles are determined as follows. Let α and β be respectively the angles of two triangles whose other vertices are on the absolute. Place the triangles side by side with their finite angles at [a] coincident vertex. From the quadrilateral so formed we may subtract a triangle having three vertices on the absolute or else we may subtract the quadrilateral from the triangle. The result will be a triangle with two vertices on the absolute and its finite angle equal to $\alpha + \beta$, in the first case or to $360° - (\alpha + \beta)$ in the second. It follows that the areas of triangles having two vertices on the absolute are proportionate to the difference of the sum of the angles from 180°. The difference between one such triangular area and another is equal to the area of a triangle having one vertex on the absolute, and the difference of the areas of two such triangles is equal to the area of a triangle with all its vertices.

So far we have assumed the triangle inside the absolute, the angles measuring round and round through 360°. In like manner, we might have set out from a triangle circumscribing the absolute having all its angles infinite, and all its sides zero [Fig. 2]. We easily prove that area is the sum of the three sides subtracted from the total length of a line.

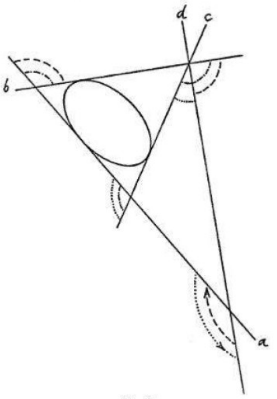

Fig. 2

C. NON-EUCLIDEAN GEOMETRY (122)

Suppose the parts of a line are displaced along that line in any way so long as the continuity of points (including those that are imaginary) is not broken and so that the new position of each point shall coincide with the old position of just one point, and *vice versa*. Then, let x and y be the homogeneous coordinates of a point on the line, these quantities being connected by a linear equation, but otherwise distributed over the line in any continuous and [ordinal] manner. If then (x', y') is a point which in its new position coincides with the old position of the point (x, y), we have a lineolinear equation between the two pairs of coordinates which may be written

$$axx' + exy' + iyx' + oyy' = 0.$$

Putting $x = x'$ and $y = y'$ we get the quadratic

$$ax^2 + (e + i)xy + oy^2 = 0,$$

which defines a pair of points on the line which remain unmoved by the displacement.

The manifoldness of displacements is evidently ∞^3, because there are 3 independent constants in the equation. There are ∞^2 positions of the unmoved points and ∞ displacements for each. But considering imaginaries, all the exponents must be doubled. All the displacements having any one position of the unmoved points are iterations of one infinitesimal displacement, and may collectively be called an area of displacement.

Let the equation of a displacement be written

$$xx' - \tfrac{1}{2}(\alpha + \beta + \tau)x - \tfrac{1}{2}(\alpha + \beta - \tau)x' + \alpha\beta = 0$$

or, writing Fx for x',

$$Fx = \frac{\tfrac{1}{2}(\alpha + \beta + \tau)x - \alpha\beta}{x - \tfrac{1}{2}(\alpha + \beta - \tau)}$$

This gives the equation of finite differences

$$F^{n-1}\cdot F^{n}-\tfrac{1}{2}(\alpha+\beta+\tau)F^{n-1}-\tfrac{1}{2}(\alpha+\beta-\tau)F^{n}+\alpha\beta=0,$$

the general solution of which (that in Boole's Finite Differences is wrong) is (α and β being unequal)

$$F^{n}=\frac{C_1\alpha(\tau+\alpha-\beta)^{n}+C_2\beta(\tau-\alpha+\beta)^{n}}{C_1(\tau+\alpha-\beta)^{n}+C_2(\tau-\alpha+\beta)^{n}}$$

To determine the ratio of the arbitrary constants, we have

$$F^{\circ}x=x=\frac{C_1\alpha+C_2\beta}{C_1+C_2}$$

and so finally

$$F^{n}=\frac{\alpha(x-\beta)(\tau+\alpha-\beta)^{n}+\beta(\alpha-x)(\tau-\alpha+\beta)^{n}}{(x-\beta)(\tau+\alpha-\beta)^{n}+(\alpha-x)(\tau-\alpha+\beta)^{n}}$$

If the moduli of $\tau+\alpha-\beta$ and $\tau-\alpha+\beta$ are not equal assume the former to be the larger. Then, $F^{\infty}x=\alpha$, $F^{-\infty}x=\beta$.

If the moduli in question are equal, then for some real value of ω, and in any case for some imaginary value

$$F^{\omega}x=x.$$

Considering the plane of imaginary quantity, every displacement of a given area of displacements may be conceived of composed of two different kinds. The displacements of the first kind are those for which $\tau+\alpha-\beta$ and $\tau-\alpha+\beta$ have the same argument, so that $x-\beta$ and $\alpha-x$ retain the same argument, and these displace all the points along circumferences through the unmoved points. The displacements of the second kind are those which displace points along circumferences orthogonal to the last, $\tau+\alpha-\beta$ and $\tau-\alpha+\beta$ having the same modulus.

The only general mathematical condition to which the rigid displacement of a line along itself is subject is, that a positive or negative, integral or fractional, but real number of repetitions of any such displacement will carry any real point to some real point and not to any imaginary point. Such a path of displacement must therefore appear as a straight line upon the plane of imaginary quantity. But there are only two straight lines among the system of circles just considered. Either therefore α and β are conjugate imaginaries, and a finite number of rigid displacements brings all the points of the line back to their original positions, or α and β are real and then they are absolute limits to rigid displacement.

SKETCH OF A SYNTHETIC TREATMENT

Chapter I. Geometry of points on a line

1. Taking any three points A, B, C, any fourth point X is either 1st on the way from A to B not passing C, 2nd on the way from B to C not passing A, or 3rd on the way from C to A not passing B.

2. In each of these divisions the multitudes of points is such that they cannot in any way be placed in one-to-one correspondence with the integral numbers.

This does not say that there is a continuous succession; but I do not see that the idea of continuity is possible without a second dimension.

Chapter II. Correspondences between 2 lines of points

Only a few topics of this vast subject to be noticed.

A *consecutive* correspondence is one in which every four points on one line have the same order of succession as the corresponding four points on the other.

A two-to-one consecutive correspondence is either pursuing or meeting. The pursuing correspondence has nothing particular, except that going once round one line corresponds to going twice round the other. But in the meeting correspondence, as a point moves part way round one line the corresponding point moves all the way round the other and then as the first point finishes the circuit of the first line the corresponding point returns and makes its circuit in the opposite direction; so that these ought to be self-evident, and indubitable apart from experience; and then many attempts were made to replace Euclid's assumption by one which should seem undeniable. General Perronet Thomson enumerated thirty of these, which the reader will find in an appendix to this treatise in an article by DeMorgan from the *English Cyclopædia*. Perhaps the two most successful of these attempts are those of Playfair. The first adopted in his edition of Euclid assumes that of two intersecting straight lines one or other must cut every third straight line in the same plane. The second is that a figure turned completely round in a plane, although not always turning round the same point must be rotated through four right angles; and this is substantially the assumption in my father's geometry. But all such premises were felt, after all, to be assumptions which might or might not be exactly true; and finally Lobachevski produced a geometry in which they were assumed to be false in some unknown measure. The falsity might, as a special case of his general admission, be nothing.

But all this time another demonstration by Euclid which is completely fallacious, had been admitted as sound reasoning. It still appears in the best treatises [Euclid's *Elements*. Book I. Prop. XVI]. ...

ON THE NON-EUCLIDEAN GEOMETRY

In this geometry there is a fixed quadric surface (it might be called a sphere, or rather the exemplar of a sphere) called the absolute. Every superposition or rigid displacement is by projection and section which does not move the absolute, unless in its own surface.

Thus the distance *AB* becomes in this way superposed upon *AC* (Fig. 1). In like manner the angles *ab* and *ac* are superposed (Fig. 2).

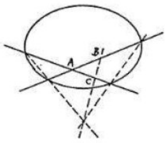

Fig. 1

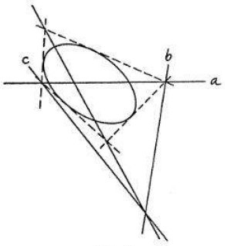

Fig. 2

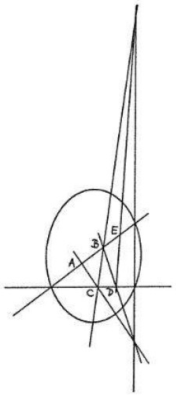

Fig. 3

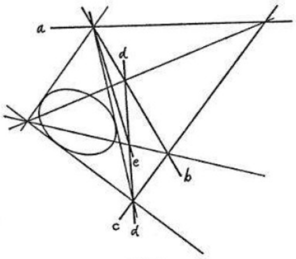

Fig. 4

So also (Fig. 3) *AB* is superposed first on *CD* and then on *BE*. And in like manner the angle *ab* is superposed first on *cd* and then on *be*.

THE PRINCIPLES OF NON-EUCLIDEAN GEOMETRY

This geometry treats of the projective relations of a certain individual quadric surface called the absolute.

This quadric surface, like every other, has in common with every line two points and two tangent planes, and has in every plane through every point in it two tangent lines. Only, in every tangent plane, intersecting at the point of tangency are two lines lying wholly in the surface.

Every transformation of this geometry is composed of displacements each of which carries any two points into new positions such that all four lie in one plane and carries any two planes into new positions so that all four pass through one point; and no point, line, or plane is carried to or from the absolute. In such displacement, all points at first in one plane continue in one plane, and all planes at first through one point continue to meet in one point.

There are six singular positions which a pair of points or a pair of planes may have, as follows:

1st, they may be in coincidence.

2nd, their line may be a tangent line. In this case a displacement may bring them into coincidence. They are then called "parallel."

3rd, one of the pair may belong to the absolute.

4th, they may be harmonically situated each on the polar or pole of the other.

5th, one of them may be harmonically situated towards one of the absolute points or planes of their line with reference to the other and the other absolute point or plane of their line. [In Fig. 1], *ABC* are harmonically situated.

6th, they may be equianharmonically situated with reference to the absolute.

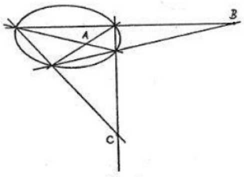

Fig. 1

The mode of superposition is as follows: Given the point *O*. Draw two lines through it, and through the intersections two lines meeting at two points *P*, *Q*, harmonic to *O*. Then by projection from *Q*, *OA* is superposed upon *OB*, by projection from *P*, *OB* is superposed on *OC*, and by projection again *OA* or *OC* is superposed upon *OD* [see Fig. 2].

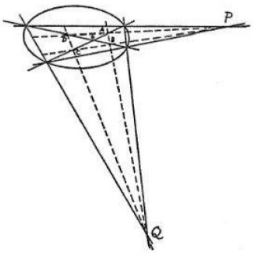

Fig. 2

D. THE ELEMENTS OF NON-EUCLIDEAN GEOMETRY. PREFACE (120)

The attempted proof by Euclid that the sum of the angles of a triangle cannot be less than two right angles was never regarded as satisfactory, not even, it is safe to say, by himself. For he baldly begs the question in his fifth postulate (often, owing to a bad text, called the 11th axiom) which is as follows:

"If two straight lines in a plane are crossed by a third, so as to make the sum of the inner angles on one side of the last less than two right angles, then those two lines will meet, if sufficiently produced, on that side of the third on which the sum of those angles is less than two right angles."

But this, I say, is begging the question; for if anybody thinks that the sum of the angles of a triangle can be less than two right angles, he will naturally hold this to be case for all triangles, and in fact it can be proved that the proposition must be true of all or of none. He will therefore hold that on a given base there will be a maximum value which the sum of two angles cannot exceed.

Euclid calls his premise a postulate, that is, an assumption; and he seems purposely to have stated it in a form in which the importance of it could be recognized. He did not seek to entrap his pupils. But the idea arose that the principles of geometry should be self evident, and indubitable apart from experience; and accordingly many attempts were made to replace Euclid's assumption by one which should seem undeniable. We reprint in an appendix a valuable article by DeMorgan from the *English Cyclopædia* enumerating some 30 of these. The most successful was that of Playfair that two straight lines which intersect must one or other cut every third straight line in the same plane.

E. [FROM A PAPER FOR THOMAS S. FISKE] (121)

... had told him had made it clear to him that the *strict geometrical truth* is that the sum of the three angles of a triangle is 180° *minus* a hypothetical correction.[1] (He should have said a constant proportion of the area.) He adds that the Euclidean geometry is practically true, at least for terrestrial figures. The limitation is significant, considering that the business of Bessel's life was with non-terrestrial figures. Gauss, after fourteen months interval, expresses his delight at learning that Bessel had come to that conclusion. The certainty of arithmetic and algebra, he adds, is absolute; but then they do not relate to anything but our own creations (unseres Geistes Product). But space has a reality without us (the Newtonian, anti-Leibnizian, anti-Kantian doctrine), and its laws cannot be known a priori. In other words, Gauss divides Geometry into Physical Geometry, a branch of physics which inquires what the properties of real space are, and all whose conclusions are affected with a probable error, and Mathematical Geometry, which in consequence of the suggestions of physical geometry, develops certain ideas of space. I do not merely say to trace out the consequences of hypotheses; for certainly Gauss would have held that it was part of the mathematician's function to invent the n-dimensional and other varieties of the space-idea.

It is singular that two fallacies in the reasoning of the first book of Euclid should have remained unnoticed to our day. One of these, in the 16th proposition, is that in the figure, where $AE\Gamma$ are coradial (a word I prefer to collinear), BEZ are coradial, $B\Gamma\Delta$ are coradial, $AE = E\Gamma$, and $BE = EZ$. Euclid assumes

$$\angle E\Gamma\Delta > \angle E\Gamma Z \quad \text{[Fig. 1].}$$

---

[1] Professor Fiske's name is written in C.S.P.'s hand on the back of the last page of manuscript, which had been folded as if once sent to him. Peirce had read a paper entitled "Rough Notes on Geometry, Constitution of Real Space" at a meeting of the American Mathematical Society on 24 November 1894. It is highly probable that this paper was the basis of his remarks at that time.

The missing first page apparently referred to a letter to Gauss from Bessel.

D. THE ELEMENTS OF NON-EUCLIDEAN GEOMETRY. PREFACE (120)

The attempted proof by Euclid that the sum of the angles of a triangle cannot be less than two right angles was never regarded as satisfactory, not even, it is safe to say, by himself. For he baldly begs the question in his fifth postulate (often, owing to a bad text, called the 11th axiom) which is as follows:

"If two straight lines in a plane are crossed by a third, so as to make the sum of the inner angles on one side of the last less than two right angles, then those two lines will meet, if sufficiently produced, on that side of the third on which the sum of those angles is less than two right angles."

But this, I say, is begging the question; for if anybody thinks that the sum of the angles of a triangle can be less than two right angles, he will naturally hold this to be case for all triangles, and in fact it can be proved that the proposition must be true of all or of none. He will therefore hold that on a given base there will be a maximum value which the sum of two angles cannot exceed.

Euclid calls his premise a postulate, that is, an assumption; and he seems purposely to have stated it in a form in which the importance of it could be recognized. He did not seek to entrap his pupils. But the idea arose that the principles of geometry should be self evident, and indubitable apart from experience; and accordingly many attempts were made to replace Euclid's assumption by one which should seem undeniable. We reprint in an appendix a valuable article by DeMorgan from the *English Cyclopædia* enumerating some 30 of these. The most successful was that of Playfair that two straight lines which intersect must one or other cut every third straight line in the same plane.

E. [FROM A PAPER FOR THOMAS S. FISKE] (121)

... had told him had made it clear to him that the *strict geometrical truth* is that the sum of the three angles of a triangle is 180° *minus* a hypothetical correction.[1] (He should have said a constant proportion of the area.) He adds that the Euclidean geometry is practically true, at least for terrestrial figures. The limitation is significant, considering that the business of Bessel's life was with non-terrestrial figures. Gauss, after fourteen months interval, expresses his delight at learning that Bessel had come to that conclusion. The certainty of arithmetic and algebra, he adds, is absolute; but then they do not relate to anything but our own creations (unseres Geistes Product). But space has a reality without us (the Newtonian, anti-Leibnizian, anti-Kantian doctrine), and its laws cannot be known a priori. In other words, Gauss divides Geometry into Physical Geometry, a branch of physics which inquires what the properties of real space are, and all whose conclusions are affected with a probable error, and Mathematical Geometry, which in consequence of the suggestions of physical geometry, develops certain ideas of space. I do not merely say to trace out the consequences of hypotheses; for certainly Gauss would have held that it was part of the mathematician's function to invent the *n*-dimensional and other varieties of the space-idea.

It is singular that two fallacies in the reasoning of the first book of Euclid should have remained unnoticed to our day. One of these, in the 16th proposition, is that in the figure, where $AE\Gamma$ are coradial (a word I prefer to collinear), BEZ are coradial, $B\Gamma\Delta$ are coradial, $AE = E\Gamma$, and $BE = EZ$. Euclid assumes

$$\angle E\Gamma\Delta > \angle E\Gamma Z \quad \text{[Fig. 1]}.$$

---

[1] Professor Fiske's name is written in C.S.P.'s hand on the back of the last page of manuscript, which had been folded as if once sent to him. Peirce had read a paper entitled "Rough Notes on Geometry, Constitution of Real Space" at a meeting of the American Mathematical Society on 24 November 1894. It is highly probable that this paper was the basis of his remarks at that time.

The missing first page apparently referred to a letter to Gauss from Bessel.

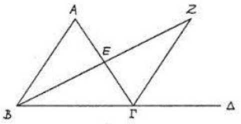

Fig. 1

That this does not follow was overlooked by Saccheri, and is overlooked by the Society for the Improvement of Geometrical Teaching, and by many other authors; but it is pointed out by Stäckel. The other, in the 7th proposition, where the construction is $A\Gamma = A\Delta$, $B\Gamma = B\Delta$, consists in inferring that $\angle A\Delta\Gamma > \angle\Delta\Gamma B$, because $\angle A\Gamma\Delta = \angle A\Delta\Gamma$ [Fig. 2]. This is overlooked even by Stäckel. Yet it is not true, even in Euclidean geometry, when Γ and Δ are at infinity; and it is not generally true in that variety of non-Euclidean geometry in which the intersection of the plane with the absolute is a point. It is true, that simple observations refute that geometry; but they are not put forward by Euclid.

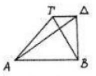

Fig. 2

Both of these fallacies arise from Euclid's confidence in his 8th Axiom, that the whole is greater than its part. These are the only places in the first book in which he makes use of that axiom, except in 18, 20, 24, and 26, which, properly understood, are equally fallacious. It has been abundantly shown by Cauchy, G. Cantor, and others, that the whole is only necessarily greater than its part when that part is finite and positive. Were the logic of relatives, either with or without any of the algebras I have proposed to facilitate its study, made use of, such fallacious reasoning would be impossible.

I maintain that Euclid was himself a non-Euclidean geometer. I do not mean, in the complete, Gaussian and Besselian sense, but more so than Saccheri and Lambert. I proceed to state the evidence of this. In the first place, the effort of all writers on the theory parallels, all defenders of the Euclidean geometry, was to produce some proposition in place of the 5th Postulate which should be more readily acceptable. This was very easy. It was so easy that Euclid could not but have known it.

For instance, he must manifestly have known that it was sufficient to say that if of three rays in one plane one was perpendicular to a second while the third was not perpendicular to that second, then the first and third must meet on the side of the acute interior angle. But Euclid disdains every effort to gloss over his assumption, and prefers to state it in the form in which its hypothetical character is most apparent. The only reasonable explanation is that he was fully alive to the hypothetical character of the proposition, and thought truth demanded its being presented in that light. Besides, consider the order of propositions in the first book. The first six are valid for all varieties of non-Euclidean geometry. The next eight (that is, to I.14, inclusive) are false in case the intersection of the plane with the absolute is a point, and in no other case. Proposition 15, that vertical angles are equal seems to the modern geometer no theorem at all; but Euclid's treatment of it is as if he conceived it to be of the same class as the preceding eight. All the next twelve propositions, in the sense in which Euclid, no doubt, would understand them, are false of the elliptic geometry alone. Owing to the confusion between internal and external angles, some of these would now be considered true in all cases. Finally, all the remaining propositions of the book are false both of elliptical and of hyperbolical geometry. It seems, therefore, that the propositions of the first book, which is, of course, the book of Fundamenta, are arranged in reference to the different kinds of non-Euclidean geometry. It seems to me that the most probable, and I might almost say the only rational, explanation of this arrangement is to suppose that Euclid had studied the non-Euclidean geometries.

In this connection let me ask: is it possible that Euclid, the author of the Data and Porisms, was unaware that metrical geometry is not the real basis of geometry?

The following count of the different kinds of propositions of the elements may be useful.

Book I, on Fundamentals, contains 23 definitions, 5 postulates, 9 (5 genuine) axioms, and 48 demonstrated propositions, of which 1, 2, 3, 9, 10, 11, 12, 22, 23, 31, 40 (11 in all) are problems.

Book II, on Areas, contains 2 definitions, and 14 demonstrated propositions, of which the 14th is a problem.

Book III, on Circles, contains 11 definitions, and 37 demonstrated propositions, of which 1, 17, 30, 33, 34 (5 in all) are problems, and 1, 16, 31 have corollaries.

Book IV, on Polygons, has 7 definitions, and 16 demonstrated propositions, all problems, of which 5 and 15 have corollaries.

Book V, on Proportions, has 18 definitions and 25 demonstrated propositions, all theorems, of which the 7th has a corollary.

Book VI, has 5 definitions and 33 demonstrated propositions, of which 9, 10, 11, 12, 13, 18, 25, 28, 29, 30, 31 (11 in all) are problems, and the 8th, 19th, 20th have corollaries. There is a lemma before the 23rd.

Book VII, on Arithmetic, has 22 definitions and 39 demonstrated propositions, all theorems, of which the 2nd has a corollary.

Book VIII, on Arithmetic, has 27 demonstrated propositions, all theorems, of which the 2nd has a corollary.

Book IX, has 36 demonstrated propositions, all theorems, of which the 11th has a corollary.

Book X, has 4 definitions before the first proposition, 6 more before the 48th, and 6 more before the 85th, 115 regular propositions all theorems, of which the 3rd, 4th, 6th, 9th, 23rd, 111th, 114th have corollaries. There are lemmas before 10, 14, 17, 19, two before 29, before 23, 42, 54, 60. There is a scholium, or remark, before the 72nd.

Book XI, on Stereometry, has 28 definitions, 39 regular propositions, of which 11, 12, 23, 26, 27 (5 in all) are problems, and 33 and 35 have corollaries. There is a lemmatic problem before the 24th.

Book XI has 18 regular propositions of which 16 and 17 (two in all) are problems, and 7, 8, 17 have corollaries. There are lemmas before the 3rd and 5th.

Book XIII has 18 regular propositions, all theorems, of which 16 and 17 have corollaries. There are so-called Lemmas before the 3rd and 14th and at the end. There is also a scholium, or remark, at the end.

Total: 132 definitions, 465 regular propositions, 27 corollaries, 17 "lemmas," 2 remarks, 5 postulates, and 9 axioms.

It is worth notice that Ptolemy's theorem (which lies at the foundation of the Brocard geometry) holds good for every form of non-Euclidean geometry.

The properties of Space may be stated (more fully, I believe, than has hitherto been done) as follows:

I. *Intrinsic Properties of Space.*

1. Space is continuous.

2. Space is tridimensional.
3. Space has no topical singularities, that is, no surfaces, lines, or points from which it extends in more or less multiple ways than from neighboring surfaces, lines, or points.
4. Space is perissid; that is, there are innumerable multitudes of unbounded surfaces of which any three have an odd number of points in common.
5. The chorisis (number of separate point-regions), cyclosis, periphraxis, and immensity of space are as small as the fourth property allows, namely, each 1.

II. *Property of the Image of Space, or the Optical Property of Space.*

6. Through any three points there is just one plane surface; and the place common to two planes is a ray, or free path of vision; and any plane and ray not lying in that plane have just one point in common.* (* The point of intersection of two rays in one plane is, in one sense, a real point, even though such a system of measurement be adopted that its distances from other real points of those rays be imaginary. We conceive of vision as making the complete circuit of the ray, irrespective of the length.)

III. *Property of the Displacement of Rigid Bodies in Euclidean Space, or the Metrical Property of Euclidean Space.*

7. When Space is measured by a rigid body, there is a plane firmament, or surface which no particle can either leave or arrive at.

It follows from what Bessel implies, in the letter above referred to, that the physical geometry of celestial triangles needs examination, in order to ascertain whether the constant of space may not have a sensible magnitude. I have undertaken such an examination. I began by forming a list of all possible methods of determining this quantity by means of the following observations: 1st, the parallaxes of stars; 2nd, the numbers of stars of each parallax; 3rd, the proper motions of stars; 4th, the numbers of stars of different proper motions; 5th, the spectroscopic determinations of the motions of stars in the line of sight; 6th, the magnitudes of stars; 7th, the numbers of stars of each magnitude. My list of possible methods was long. All of them, it is true, involved some hypothetical element; but that is true of any research, whatever, into the value of a physical quantity; and it is possible so to modify the methods that the hypotheses that appear the most dangerous may probably be eliminated. I applied several methods: they seemed to indicate a hyperbolic space with a constant far from insignificant. The best method seemed to be by proper motions. We know that our ordinary dynamics gives a definite value to absolute, or nonrelative velocity of rotation. Thus, Foucault's pendulum experiment shows that the earth has an absolute motion of rotation, and not merely a rotation relatively to the sphere of the heavens. All German attempts to escape this conclusion are metaphysical word-spinning. In the Euclidean geometry, translation is an infinitesimal rotation about an axis lying in the firmament. In the non-Euclidean geometry there is no plane firmament and, consequently, no translation. In every rotation, every particle lying in the ray which is the polar conjugate with respect to the absolute of the axis of rotation moves along that ray. All other particles describe circles. Consequently, there is an absolute velocity in all motion. The same apparent effect cannot be produced by a motion of the stars and an opposite motion of the solar system. The proper motion due to the absolute motion of a star varies inversely as the sine of the product of the distance by the square root of a constant. The proper motion due to the absolute motion of the sun varies inversely as the tangent of the product of the distance by the square root of the same constant. If, therefore, we had all the proper motions plotted upon a globe, and if we were first to throw out all those which are excessively large, as probably belonging to stars so near the sun that they may be supposed sensibly to partake of the system of motion to which the sun's motion belongs, — whether they move in the same direction or with the same velocity or not, — we should expect to find the remaining proper motions presenting one or other of two appearances. Namely, if space is elliptical (that is, finite), and by consequence the square root by which the distance is multiplied be real, the proper motions due to the sun's motion would for very distant stars become insensible (because divided by indefinitely great tangents) and therefore those stars whose proper motions were very small would be moving about in every direction quite irrespective (or, at least, relatively irrespective) of the motion of the sun. If, however, space is hyperbolical, and that square root aforesaid imaginary, we should expect to find that those stars whose proper motions were very small would (except for the effect of errors of observation) move more directly away from the apex of the sun's motion than those whose proper motions were relatively large. I made the necessary computations for a selection of stars from Newcomb's catalogue. Only, instead of judging which were the most distant stars by the proper motions alone, I formed my judgment by combining the evidence from proper motion and from magnitude.

The result was most markedly in favor of the hyperbolical geometry. A rough calculation of probabilities showed that it could not be with probability attributed to chance. I was in hopes sufficient interest would be excited by this result to enable me to obtain the funds to employ a computer for a few months to extend the calculations to all the proper motions of Auwers. But as the question has no immediate relation either to business or comfort, this has not hitherto been the case. Its decision must strongly influence the scientific conception of the universe; and might thus lead to practical results sooner than one might, at first sight, suppose possible. I have now laid out a new plan of the work; so that for $150, a rough graphical result may be obtained; and if this should be of a convincing character, $500 more may be expended in computations, without the determination of any connected constants. It might ultimately be thought worth while to work in a more fundamental way. It will be remarked, that if a constant value results from such calculations, as the work I have already done gives me no room to doubt, then that constant must mean *something*. Even if I am wrong in thinking that it can only be attributed to a property of space, it still must have some significance for the visible universe. I shall be happy to receive subscriptions payable when enough have been received to cover the expense of the graphical solution.[1]

C.S. Peirce

---

[1] At this point the reader is referred to pp. 421-424 of "The Charles S. Peirce–Simon Newcomb Correspondence" by the editor in *The Proceedings of the American Philosophical Society*, 101:5 (October 1957), 409-433.

F. ON TWO MAP-PROJECTIONS OF THE LOBATSCHEWSKIAN PLANE[1]

The proper motions of the stars show very strong indications that our space is really hyperbolic, or, what comes to the same thing, that the law of dynamics, or kinematics, is such that if two stars move in the same plane with the same uniform velocity, without being acted upon by any forces, and each appears at one moment as seen from the other to be abreast with it, then as time goes on each will fall behind the other. Other indirect arguments tend to confirm me in this opinion; and thus, entertaining for this species of geometry something closely approaching belief, I have found it convenient to give it intuitional shapes by map-projections.

I use two, both very simple. To avoid confusion, I will write all formulae applying to the noneuclidean space in *red* (in printing, Old English can be used), all formulae applying to the representation of that space in Euclidean space in *blue* (in printing, Roman type), all formulae relating to the connection between the two in *brown* (in printing, heavy-faced type; on the blackboard, yellow chalk), and all formulae applicable everywhere since they merely relate to algebra in *black* (in printing, Italics; on the blackboard, white chalk.) I also use the sign $\backsim$ to mean "represents," or "corresponds to."

The formulae of Lobatchewskian trigonometry can be obtained by the following system of correspondences.

$$(1)\quad \begin{cases} \mathbf{A}\text{ (an angle in spherical trig.) } \backsim \mathbf{A}\text{ (in Lobat. plane trig.)} \\ \sin \mathbf{a}\text{ (a being a side in sph. trig.) } \backsim -\mathbf{i}\operatorname{Sinh} \mathbf{a}\text{ (in Lob. pl. tr.)} \\ \cos \mathbf{a}\text{ (a being a side in sph. trig.) } \backsim \operatorname{Cosh} \mathbf{a}\text{ (in Lob. plane trig.)} \end{cases}$$

Although only a few of the following formulae will be needed, the others are given for convenience.

Right Triangles

$$\operatorname{Cosh} \mathfrak{h} = \operatorname{Cosh} \mathfrak{a}.\ \operatorname{Cosh} \mathfrak{b}$$

$$(2)\qquad \operatorname{Cosh} \mathfrak{h} = \cot \mathfrak{A}.\ \cot \mathfrak{B}$$

$$(3)\qquad \cos \mathfrak{A} = \operatorname{Cosh} \mathfrak{a}.\ \sin \mathfrak{B}$$

$$\operatorname{Sinh} \mathfrak{h} = \frac{\operatorname{Sinh} \mathfrak{a}}{\sin \mathfrak{A}}$$

$$(4)\qquad \operatorname{Tanh} \mathfrak{h} = \frac{\operatorname{Tanh} \mathfrak{a}}{\cos \mathfrak{B}}$$

$$\operatorname{Sinh} \mathfrak{b} = \frac{\operatorname{Tanh} \mathfrak{a}}{\tan \mathfrak{A}}$$

Oblique Triangles

$$\frac{\operatorname{Sinh} \mathfrak{a}}{\sin \mathfrak{A}} = \frac{\operatorname{Sinh} \mathfrak{b}}{\sin \mathfrak{B}} = \frac{\operatorname{Sinh} \mathfrak{c}}{\sin \mathfrak{C}}$$

$$(5)\qquad \operatorname{Cosh} \mathfrak{c} = \operatorname{Cosh} \mathfrak{a}.\ \operatorname{Cosh} \mathfrak{b} - \operatorname{Sinh} \mathfrak{a}.\ \operatorname{Sinh} \mathfrak{b}.\ \cos \mathfrak{C}$$

$$\cos \mathfrak{C} = -\cos \mathfrak{A} \cos \mathfrak{B} + \sin \mathfrak{A} \sin \mathfrak{B} \operatorname{Cosh} \mathfrak{c}$$

$$\operatorname{Sinh} \mathfrak{a}.\ \operatorname{Coth} \mathfrak{b} = \cot \mathfrak{B}.\ \sin \mathfrak{C} + \operatorname{Cosh} \mathfrak{a}.\ \cos \mathfrak{C}$$

A few formulae of hyperbolic functions may also be added, as follows.

$$(6)\qquad \operatorname{Tanh} x = \frac{\operatorname{Sinh} x}{\operatorname{Cosh} x} \qquad \operatorname{Coth} x = \frac{\operatorname{Cosh} x}{\operatorname{Sinh} x}$$

$$(7)\qquad \operatorname{Sech} x = \frac{1}{\operatorname{Cosh} x} \qquad \operatorname{Cosech} x = \frac{1}{\operatorname{Sinh} x}$$

$$(8)\qquad \begin{cases} \operatorname{Sinh} 0 = 0 \\ \operatorname{Sinh} \infty = \infty \end{cases} \qquad \begin{matrix} \operatorname{Cosh} 0 = 1 \\ \operatorname{Cosh} \infty = \infty \end{matrix} \qquad \begin{matrix} \operatorname{Tanh} 0 = 0 \\ \operatorname{Tanh} \infty = \infty \end{matrix}$$

$$\operatorname{Sinh} x = \frac{\mathfrak{e}^{x} - \mathfrak{e}^{-x}}{2} \qquad \operatorname{Cosh} x = \frac{\mathfrak{e}^{x} + \mathfrak{e}^{-x}}{2}$$

$$\operatorname{Cosh}^2 x = 1 + \operatorname{Sinh}^2 x$$

$$(9)\qquad \operatorname{Tanh}^2 x + \operatorname{Sech}^2 x = 1$$

---

[1] This manuscript is to be found in the archives of the Smithsonian Institution. For details on Peirce's association with that Institution when Samuel P. Langley was the Secretary see "The Scientist-Philosopher C. S. Peirce at the Smithsonian" by the editor in the *Journal of the History of Ideas* XVIII:4 (October 1957). Filed with the manuscript at the Smithsonian are two letters. Peirce's covering letter to Langley dated 18 April 1901 reads: "If the Academy is in session when this reaches you I should like to have it presented. I fear it is too trifling to be printed as a Memoir, though I know not where else to print it ...." It was received at the Smithsonian on April 20.

Langley's reply was written on 22 April 1901 and reads: "Your paper 'On two Map-Projections of the Lobatschewskian Plane' only reached me the day after the Academy had completed its session. What would you like to have me do with it?"

MS. 114 at the Houghton Library is a partial draft of this paper. It is entitled "On Hyperbolic Geometry."

(10) $\mathrm{Coth}^2 x = \mathrm{Cosech}^2 x + 1$

$\mathrm{Sinh}\,(x+y) = \mathrm{Sinh}\,x.\ \ \mathrm{Cosh}\,y + \mathrm{Cosh}\,x.\ \ \mathrm{Sinh}\,y$

$\mathrm{Cosh}\,(x+y) = \mathrm{Cosh}\,x.\ \ \mathrm{Cosh}\,y + \mathrm{Sinh}\,x.\ \ \mathrm{Sinh}\,y$

(11) $\mathrm{Coth}\,x + \mathrm{Cosech}\,x = \mathrm{Coth}\,\frac{x}{2}$

(12) $\mathrm{Coth}\,x - \mathrm{Cosech}\,x = \mathrm{Tanh}\,\frac{x}{2}$

(13) $\mathrm{Tanh}\,x = \dfrac{2\,\mathrm{Tanh}\,\frac{x}{2}}{1+\mathrm{Tanh}^2\frac{x}{2}} = \dfrac{2\,\mathrm{Coth}\,\frac{x}{2}}{1+\mathrm{Coth}^2\frac{x}{2}}$

(14) $\mathrm{Coth}\,x.\ \mathrm{Coth}\,\frac{x}{2} = \frac{1}{2}(1+\mathrm{Coth}^2\frac{x}{2})$

(15) $\mathrm{Coth}\,x.\ \mathrm{Tanh}\,\frac{x}{2} = \frac{1}{2}(1+\mathrm{Tanh}^2\frac{x}{2})$

(16) $\mathrm{Cosech}\,x.\ \mathrm{Coth}\,\frac{x}{2} = \dfrac{1}{2\,\mathrm{Sinh}^2\frac{x}{2}}$

(17) $\mathrm{Cosech}\,x.\ \mathrm{Tanh}\,\frac{x}{2} = \dfrac{1}{2\,\mathrm{Cosh}^2\frac{x}{2}}$

A CONFORMED PROJECTION.

Let us now proceed to construct a projection, fulfilling the following conditions:

1st, all angles to be preserved unchanged;

2nd, there is to be a centre of projection all straight lines through which are represented by straight lines [Fig. 1].

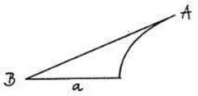

Fig. 1

Let B be at the centre of projection, and denote the angle between two straight lines by the same letter. Upon one of these lines, at a distance a from the centre, erect a perpendicular, intersecting the other line in an internal angle, A. Then (3) applies.

Upon the map, let the point corresponding to B be taken as an origin of polar coordinates. Let the other extremity of a be represented by a point whose coordinates are:

(18) $\boldsymbol{\theta = 0} \qquad \mathbf{r = Fa = l}$

Then, by (3),

(19) $\cos A = \mathrm{Cosh}\,a \cdot \sin\theta$

Upon the map, we shall have a plane triangle having, at the origin, an angle $d\theta$, included between two sides whose lengths are r and $r+dr$, while one of the other angles is A. Then, by plane trigonometry, we shall have

$$\tan A = \frac{r\cdot\sin d\theta}{r+dr-r\cdot\cos d\theta} = r\frac{d\theta}{dr}$$

whence

(20) $$\frac{d\log r}{d\theta} = \cot A = \frac{\cos A}{\sqrt{1-\cos^2 A}} = \frac{\mathrm{Cosh\,a}\cdot\sin\theta}{\sqrt{1-\mathrm{Cosh}^2\mathrm{a}\cdot\sin^2\theta}} = \frac{\sin\theta}{\sqrt{\cos^2\theta-\mathrm{Tanh}^2\mathrm{a}}}$$

[see (9)]. Multiply both sides by $\dfrac{d\theta}{d\cos\theta} = -\dfrac{1}{\sin\theta}$; and we have

(21) $$\frac{d\log r}{d\cos\theta} = -\frac{1}{\sqrt{\cos^2\theta-\mathrm{Tanh}^2\mathrm{a}}}$$

The integral of this is

(22) $$\frac{\mathrm{C}}{r} = \frac{\cos\theta}{\mathrm{Tanh\,a}} + \sqrt{\frac{\cos^2\theta}{\mathrm{Tanh}^2\mathrm{a}}-1}$$

It is obvious that the two values resulting from the ambiguous sign of the square root are reciprocals of one another; for

$$(x+\sqrt{x^2-1})(x-\sqrt{x^2-1}) = x^2-(x^2-1) = 1.$$

Multiplying (22) by r, and putting

(23) $r\cos\theta = x \quad r^2 = x^2+y^2,$

we get

(24) $$\left(x-\frac{\mathrm{C}}{\mathrm{Tanh\,a}}\right)^2 + y^2 = (\mathrm{C\,Cosech\,a})^2;$$

so that the equation represents a circle.

In order to determine the value of C, we use (18) and find from (22)

(25) $$\frac{\mathrm{C}}{l} = \mathrm{Coth\,a} \pm \sqrt{\mathrm{Coth}^2\mathrm{a}-1} = \begin{cases}\mathrm{Coth\,a} + \mathrm{Cosec\,a} = \mathrm{Coth}\,\frac{a}{2}\\ \mathrm{Coth\,a} - \mathrm{Cosec\,a} = \mathrm{Tanh}\,\frac{a}{2}\end{cases}$$

[see (10), (11). (12)]. Substituting this in (24), we get, by (14), (15), (16), (17),

(26) $$[x - \tfrac{1}{2}l(1 + \mathrm{Coth}^2\tfrac{a}{2})]^2 + y^2 = \left(\frac{l}{2\,\mathrm{Sinh}^2\frac{a}{2}}\right)^2$$

$$[x - \tfrac{1}{2}l(1 + \mathrm{Tanh}^2\tfrac{a}{2})]^2 + y^2 = \left(\frac{l}{2\,\mathrm{Cosh}^2\frac{a}{2}}\right)^2$$

The difficulty of conceiving a straight line to be represented by a circle which is topically untransformable into a straight line will be considered below, and shown not to be serious.

Since the leg b of the triangle is represented by a circle, it follows that, as B increases, the angle A must ultimately become zero. In fact, by (2), when $A = 0$, $\mathrm{Cosh}\, h = \infty$, $h = \infty$. But the circle representing the leg b is a finite one, and thus the locus of points infinitely distant from the centre must be represented by a finite circle. Let the equation of the circle representing the absolute be

(28) $$r = L$$

Then, every circle orthogonal to this circle on the map will represent a straight line in the Lobatchewskian plane. The general equation of such a circle is

(29) $$\left(x - \frac{L^2 + l^2}{2l}\right)^2 + y^2 = \left(\frac{L^2 - l^2}{2l}\right)^2$$

The same equation precisely would hold if l, instead of representing the minimum distance of the circle representing leg b from the origin, represented the maximum distance. If (29) be compared with (26) and with (27), each of these last two equations affords two ways of determining the value of l. From (26), we have

(30) $$\frac{L^2}{l^2} = \mathrm{Coth}^2\tfrac{a}{2}, l = L\,\mathrm{Tanh}\tfrac{a}{2}; \text{and } \frac{L^2}{l^2} - 1 = \mathrm{Cosech}^2\tfrac{a}{2}, l = L\,\mathrm{Tanh}\tfrac{a}{2}$$

while from (27) we have

(31) $$\frac{L^2}{l^2} = \mathrm{Tanh}^2\tfrac{a}{2}, \mathrm{l} = L\,\mathrm{Coth}\tfrac{a}{2}; \text{and } \frac{L^2}{l^2} - 1 = -\mathrm{Sech}^2\tfrac{a}{2}, l = L\,\mathrm{Coth}\tfrac{a}{2}.$$

In order to make the last right, since a hyperbolic tangent of a real quantity is essentially less than 1, we have to assume that l is the maximum distance of the circle representing the leg b, from the origin. The circle on the map is cut twice by the radius vector, although the straight line it represents is cut but once by the hypotheneuse represented by the radius vector. In short, one point on the Lobatchewskian plane is represented twice over on the map; and generally, all that part of the map outside

the circle representing the absolute only represents again what is represented inside. Equations (26) and (30) refer to the inside map; equations (27) and (31) to the outside map. On comparing (25) with (30) and (31), we get

(32) $$C = L.$$

It is now proved that (30) gives the rule for the projection, if there is any such projection; but it is perhaps not proved that that rule does not lead to contradictory results. At any rate, it will be worthwhile to test it by applying (4). Substituting (13) in this, we get

(33) $$\mathrm{Tanh}\frac{\mathfrak{h}}{2} + \mathrm{Coth}\frac{\mathfrak{h}}{2} = \left(\mathrm{Tanh}\frac{\mathfrak{a}}{2} + \mathrm{Coth}\frac{\mathfrak{a}}{2}\right)\cos\mathfrak{B}$$

In the figure [2], the distance between 0, the origin, and C, the centre of the circle representing the leg b is $\frac{L^2 + l^2}{2l}$ and the radius of that circle is $\frac{L^2 - l^2}{2l}$. Hence, by plane trigonometry,

(34) $$\left(\frac{L^2 - l^2}{2l}\right)^2 = \left(\frac{L^2 + l^2}{2l}\right)^2 + r^2 - 2\frac{L^2 + l^2}{2l} r\cos\theta$$

which reduces at once to

(35) $$\frac{L}{r} + \frac{r}{L} = \left(\frac{L}{l} + \frac{l}{L}\right)\cos B$$

Now since we know, by (30), that

(30) $$\frac{\mathbf{l}}{\mathbf{L}} = \mathrm{Tanh}\,\frac{\mathbf{a}}{2}$$

it follows from the comparison of (33) and (35) that

(36) $$\frac{\mathbf{r}}{\mathbf{L}} = \mathrm{Tanh}\,\frac{\mathbf{h}}{2}$$

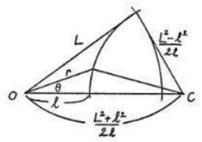

Fig. 2

The projection is thus established. It may be very appropriately called the hyperbolic stereographic projection, since its rule of construction only differs from that of the stereographic projection of the elliptic plane in that the radii vectores are proportional to $\text{Tanh}\,\frac{h}{2}$ instead of to $\tanh\frac{h}{2}$, where h is the distance of the point to be represented from the centre. It also agrees with the elliptic stereographic projection in preserving the angles. The other important property of the latter is that every circle represents a circle (or straight line), and conversely. In order to ascertain whether the hyperbolic stereographic projection has this property or not, we assume the equation of a circle, which by plane trigonometry is

$$R^2 = r^2 + Q^2 - 2rQ\cos\theta \tag{37}$$

Now, of course, the circle cannot represent a real circle unless

$$L > Q + R \tag{38}$$

so that $\sqrt{(L+Q+R)(L-Q-R)(Q-R-L)(R-L-Q)}$ is real. Call this square root V for the sake of brevity, $\frac{L+Q-R}{V}$ and $\frac{L-Q+R}{V}$ will, in case of reality, be positive quantities greater than 1 and, therefore, of the nature of hyperbolic cosines. We may assume, then,

$$\text{Cosh}\,\mathbf{b} = \frac{\mathbf{L}+\mathbf{Q}-\mathbf{R}}{V} \tag{39}$$

$$\text{Cosh}\,\mathbf{c} = \frac{\mathbf{L}-\mathbf{Q}+\mathbf{R}}{V} \tag{40}$$

Solving these equations for Q and R, we find

$$\mathbf{Q} = \frac{\mathbf{L}\,\text{Sinh}\,\mathbf{b}}{\text{Cosh}\,\mathbf{b} + \text{Cosh}\,\mathbf{c}}$$

$$\mathbf{R} = \frac{\mathbf{L}\,\text{Sinh}\,\mathbf{c}}{\text{Cosh}\,\mathbf{b} + \text{Cosh}\,\mathbf{c}}$$

Putting also

$$\boldsymbol{\theta} = \mathbf{C}$$
$$\mathbf{r} = \mathbf{L}\,\text{Tanh}\,\tfrac{\mathbf{a}}{2}$$

the equation (37) of the circle becomes

$$\begin{aligned}\text{Sinh}^2\,\mathfrak{c} = {} & \text{Tanh}^2\tfrac{\mathfrak{a}}{2}\,(\text{Cosh}\,\mathfrak{b} + \text{Cosh}\,\mathfrak{c})^2 + \text{Sinh}^2\mathfrak{b}\\ & -2\text{Tanh}\tfrac{\mathfrak{a}}{2}\cdot\text{Sinh}\,\mathfrak{b}\,(\text{Cosh}\,\mathfrak{b} + \text{Cosh}\,\mathfrak{c})\cos\mathfrak{C}\end{aligned}$$

or, dividing by $\text{Cosh}\,b + \text{Cosh}\,c$, and observing that

$$\text{Sinh}^2 c - \text{Sinh}^2 b = (\text{Cosh}\,c - \text{Cosh}\,b)(\text{Cosh}\,c + \text{Cosh}\,b)$$

we get

$$\text{Cosh}\,\mathfrak{c} - \text{Cosh}\,\mathfrak{b} = \text{Tanh}^2\tfrac{\mathfrak{a}}{2}(\text{Cosh}\,\mathfrak{b} + \text{Cosh}\,\mathfrak{c}) - 2\text{Tanh}\tfrac{\mathfrak{a}}{2}\cdot\text{Sinh}\,\mathfrak{b}\cdot\cos\mathfrak{C}$$

Dividing by $1 - \text{Tanh}^2\frac{a}{2}$, and observing that

$$\text{Cosh}\,a = \frac{1 + \text{Tanh}^2\frac{a}{2}}{1 - \text{Tanh}^2\frac{a}{2}}$$

$$\text{Sinh}\,a = \frac{2\,\text{Tanh}\frac{a}{2}}{1 - \text{Tanh}^2\frac{a}{2}}$$

we get

$$\text{Cosh}\,\mathfrak{c} = \text{Cosh}\,\mathfrak{a}\cdot\text{Cosh}\,\mathfrak{b} - \text{Sin}\,\mathfrak{a}\cdot\text{Sin}\,\mathfrak{b}\cdot\cos\mathfrak{C}$$

which, when a and C are variable is, by (5), the equation of a circle. Hence, this projection represents circles (and straight lines) by circles, and conversely. Summing up its properties:

1st. All angles are preserved.

2nd. Representing any point on the hyperbolic plane by the origin on the map, all straight lines through that centre are represented by straight lines.

3rd. A point distant by h from the centre is represented by a point distant by L $\text{Tanh}\,\frac{h}{2}$ from the original.

4th. Every straight line in the hyperbolic plane is represented by a circle orthogonal to the circle of radius L with its centre at the origin.

5th. Every circle lying wholly within that absolute circle on the map represents a real circle of the hyperbolic plane with a real radius, and conversely. Straight lines not through the origin and circles cutting the absolute circle represent real circles of imaginary radius, a sort of locus peculiar to hyperbolic geometry.

Practically, this projection is distinctly more convenient even than the ordinary stereographic; but theoretically it presents greater inconveniences. As the elliptic stereographic represents every point twice over, so also does the hyperbolic stereographic projection. This arises from the fact that it is conformal. For conformal projections are exclusively between artiad surfaces, that is, surfaces of even cyclosis, while the projective plane, or plane upon which unbounded lines can intersect in an odd number of points, cannot by any distortion be brought into coincidence with an artiad surface. In the hyperbolic plane, we have to distinguish between imaginary points and points at imaginary distances from the origin. That every pair of real rays in a plane intersect in a real point is a proposition of projective geometry, which no mode of measurement, that is, no system of designating the points, can falsify. The elliptic stereographic divides the plane along the whole extent of a ray and there joins it to a similarly cut copy of itself turned upside down;

so that every part of the elliptic plane is twice represented in relatively perverted images. The hyperbolic stereographic cuts a disk out of the plane, along the absolute conic, and joins that to a perverted image of itself, leaving all real points at imaginary distances entirely unrepresented. It obviously cannot represent those points conformedly since the angles are there imaginary. This is a grave inconvenience for many purposes.

The projection is subject to another grave inconvenience which manifests itself, for example, if we attempt to apply the projective definition of a circle as a conic having double contact with the absolute. Since the real part of the absolute circle is shown in the projection, we might naturally expect to find some traces of the double contact of circles with the absolute, although the plane of a conformal projection has no imaginary parts, being, on the contrary, itself the representation of a single imaginary variable. But Cayley's theory of distance, beautiful as it is, takes but a narrow view of the subject. The system of distances is the expression of the law which defines the possibilities of displacement of a rigid body. In our universe, a part of that law is that rigid bodies can receive no imaginary displacements. Nor need imaginaries be called in to define the law in general terms, as I intend to show in another paper.

A CENTRAL PROJECTION OF THE LOBATCHEWSKIAN PLANE

The Lobatchewskian plane can also be so represented, that every ray, α, is represented by a ray cutting at right angles that radius vector from the origin which represents the ray through a fixed centre that the ray α cuts at right angles; and so that all angles about that centre are represented by equal angles about the origin [Fig. 3. △B, red; △O blue].

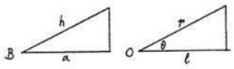

Fig. 3

A right triangle having one of its angles, B, at the centre will be represented by a right triangle having the corresponding angle at the origin. Then we shall have

$$\mathbf{l} \backsim \mathbf{a}$$
$$\mathbf{r} \backsim \mathbf{h}$$
$$\boldsymbol{\theta} = \mathbf{B}$$

$$\mathbf{l} = \mathbf{Fa}$$
$$\mathbf{r} = \mathbf{Fh}$$
$$\mathbf{r} = \mathbf{l}\sec\boldsymbol{\theta}$$
$$\mathrm{Tanh}\,\mathfrak{h} = \mathrm{Tanh}\,\mathfrak{a} \cdot \sec\mathfrak{B}$$

Hence the projection of any point will be determined by the equations

$$\boldsymbol{\theta} = \mathbf{B}$$
$$\mathbf{r} = \mathrm{L}\,\mathrm{Tanh}\,\mathbf{h}$$

where L is the radius of the circle which represents the circle at infinity.

Distances are constructed with the greatest ease. Suppose, for example, that having given two points, C and H, on the map, we wish to know where the ellipse representing a circle about C as a centre and having the radius represented by CH will cut another ray through C. The figure [4] almost explains itself. The line CH is produced to PR, its intersection with the absolute, and the other line is produced to QS, its intersections with the same circle. The quadrangle $PQRS$ is completed so that A is collinear with P and Q and with R and S, while B is collinear with P and S and with Q and R. Then construct I at the inter-

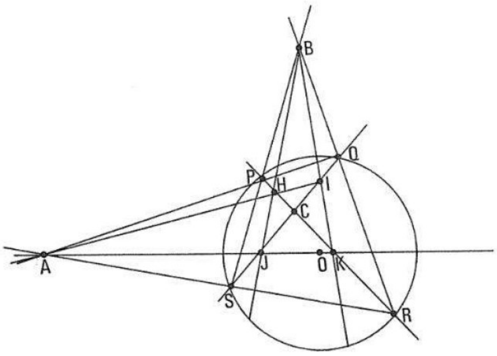

Fig. 4

section of AH and $[QS]$, J at the intersection of BH and $[QS]$, and K at the intersection of AJ and BI. [Then] CH, CI, CJ, CK represent equal distances. By drawing a diameter through C and O, the centre of the circle representing the absolute and laying down four points situated symmetrically to H, I, J, K, on the other side of this diameter, we have eight points on the ellipse representing the circle. There are other obvious modes of constructing equal distances. [Fig. 4 follows Peirse's instructions.]

I do not know any simple geometrical construction for angles, or any which is purely projective. There probably exists such a real projection. Triangles can, however, readily be solved by means of the following

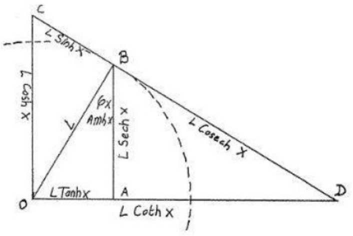

Fig. 5

figure [5]. Let O be the centre of the circle of radius L representing the absolute. Let OA represent the distance x, so that

$$OA = L \operatorname{Tanh} x$$

At A erect a perpendicular to OA cutting the circle of the absolute in B, so that $OB = L$. Then,

$$AB = L \operatorname{Sech} x$$
$$\angle OBA = 2 \arctan e^{x} - \frac{\pi}{2}$$

Through B draw a perpendicular to OB cutting OA produced in D. Then

$$OD = L \operatorname{Coth} x$$
$$BD = L \operatorname{Cosech} x$$

At O raise a perpendicular to OA cutting BD produced in C. Then

$$OC = L \operatorname{Cosh} x$$
$$BC = L \operatorname{Sinh} x.$$

On these principles, angles can be brought to the centre; but with some trouble [see Fig. 6].

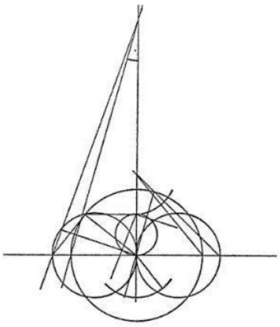

Fig. 6

(*

$\sin u \cdot d \sin u =$

$\sin u_2 \, d \sin u_2$

Pag. 519 (280) G

The Skew Mercator on the Sphere (1584).

Pag. 687 (364) —

Non-Euclidean Geometry

Obs: Analizar tudo!!!

Pag. 710 (376) F.

On two Map-Projection of the Lobatschewskian Plane

Source: manuscript of the Smithsonian Institution.

Pag. 712 (377)

A Conformed Projection

Look: Fig. 2

The projection is thus established

(hyperbolic stereographic projection).

Pág. 718(380) A Central Projection
of the Lobatchewskian Plane
Fig. 3 and Fig. 4
Pág. 720(381) - Fig. 5 and Fig. 6.

Books-019-2021/20210920/
Pythagorean-hodograph.../
— Treatise of plane geometry
through Geometric Algebra...
Calvet(2001).pdf

Look. Pág. 1

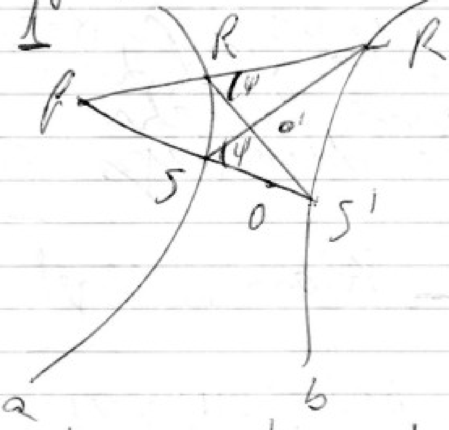

Analizar
a curvatura
que nos
é espelhada
a e b,
nos observamos
o mesmo ψ. Interessante, utilizá-lo
conjuntamente com Pappus (3 points)
R O' S' → 1ª reta.
S O' R' → 2ª reta.

("
+,-.+ + +/ !+* 0 ! 3 ! 2
! 4 ! H -! 1 6(6 ! 8 3#

Pag. 718(380) A Central Projection of the Lobatchewskian Plane
Fig. 3 and Fig. 4

Pag. 720(381) - Fig. 5 and Fig. 6.

Books-019-2021/20210920/
Pythagorean-hodograph.../
- Treatise of plane geometry through Geometric Algebra...
Calvet(2001).pdf

Look. Pag. 1

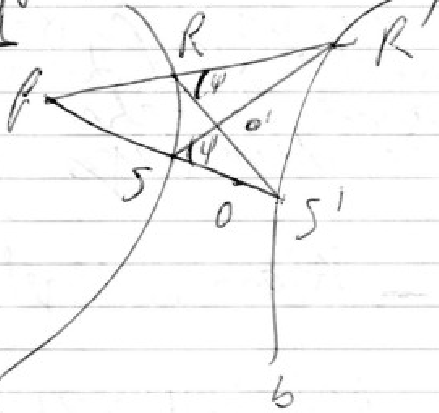

Analizar a curvatura que nos é espelhada a e b, nos observamos o mesmo φ. Interessante, utilizá-lo conjuntamente com Pappus (3 points)

R O' S' → 1ª reta.
S O' R' → 2ª reta.

(

TREATISE OF PLANE GEOMETRY THROUGH GEOMETRIC ALGEBRA

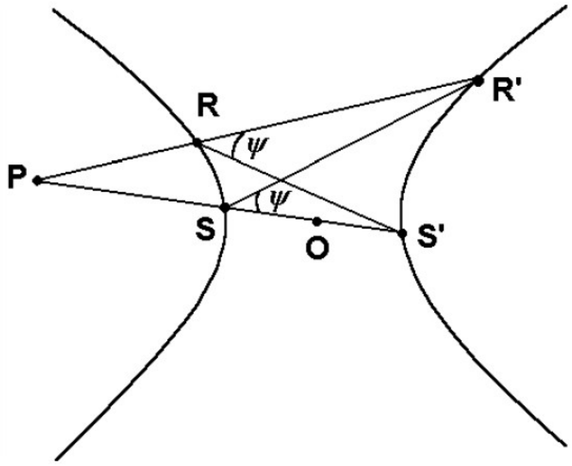
R
R'
ψ
P
ψ
S
O
S'

Ramon González Calvet

( ))

additional rotation with the same centre. Then, a composite homothety is equivalent to a simple homothety followed by a rotation. Direct similarities and homotheties are different names for the same transformations (exercise 7.5). Also homotheties with $|k|=1$ and displacements are equivalent.

The *single ratio of three points A*, *B*, and *C* is defined as:

$$(A\ \ B\ \ C) = AB\, AC^{-1}$$

The single ratio is a real number when the three points are aligned, and a complex number in the other case. Under a homothety, the single ratio of any three points remains invariant. Let us prove this. Because any vector is dilated and rotated by a factor *k*, we have:

$$A'B' = OB' - OA' = OB\,k - OA\,k = AB\,k$$

$$A'C' = AC\ k \qquad \Rightarrow \qquad A'C'^{\,-1} = k^{-1}\,AC^{-1}$$

$$(A'\ \ B'\ \ C') = AB\,k\,k^{-1}\,AC^{-1} = AB\,AC^{-1} = (A\ \ B\ \ C)$$

If the single ratio is invariant for a geometric transformation, then it transforms triangles into directly similar triangles and hence also polygons into similar polygons:

$$(A\ \ B\ \ C) = (A'\ \ B'\ \ C') \qquad \Rightarrow \quad AB\,AC^{-1} = A'B'\,A'C'^{\,-1} \quad \Rightarrow AC^{-1}\,A'C' = AB^{-1}\,A'B'$$

Therefore, the homothety always transforms triangles into directly similar triangles as shown in the figure 7.6. It follows immediately that the homothety preserves the angles between lines. It is the simplest case of *conformal* transformations, the geometric transformations which preserve the angles between lines. Since a line may also be transformed into a curve, and for the general case a curve into a curve, the conformal transformations are those which preserve the angle between any pair of curves, that is, the angle between the tangent lines to both curves at the intersection point. A transformation is *directly* or *oppositely conformal* whether it preserves or changes the sense of the angle between two curves (figure 7.7). This condition is equivalent to the conservation of the single ratio of any three points at the limit of accumulation:

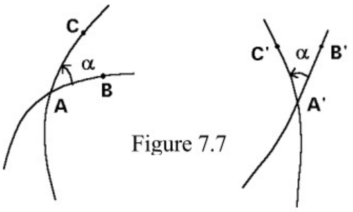

Figure 7.7

$$\lim_{B,C\to A} (A\,B\,C) = \lim_{B',C'\to A'} (A'\,B'\,C') \qquad \text{directly conformal}$$

$$\lim_{B,C\to A} (A\,B\,C) = \lim_{B',C'\to A'} (A'\,B'\,C')^{*} \qquad \text{oppositely conformal}$$

( )

$$N = \frac{1}{2}\left(\frac{P+H}{2}\right) + \frac{1}{2}\left(\frac{Q+R}{2}\right)$$

it follows that (*P*+*H*)/2 and (*Q*+*R*)/2 are opposite extremes of the same diameter. Then the vectors *J*–(*P*+*H*)/2 and *J*–(*Q*+*R*)/2, which are orthogonal:

$$\left(J - \frac{P+H}{2}\right) \cdot \left(J - \frac{Q+R}{2}\right) = 0$$

are the sides of a right angle which intercepts a half circumference of the nine point circle. Therefore, this angle is inscribed and its vertex *J* also lies on the nine-point circle. Also, we may go through the algebraic way to arrive at the same conclusion. Developing the former inner product:

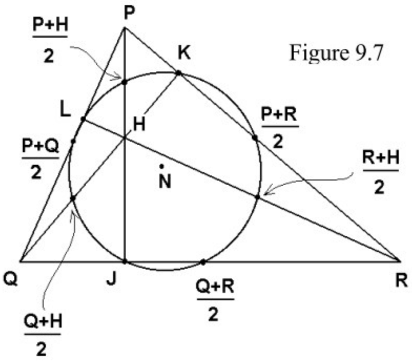

Figure 9.7

$$J^2 - J \cdot \frac{Q+R}{2} - \frac{P+H}{2} \cdot J + \frac{P+H}{2} \cdot \frac{Q+R}{2} = 0$$

Introducing the centre *N* of the nine-point circle, we have:

$$J^2 - 2J \cdot N + 2N \cdot \frac{Q+R}{2} - \left(\frac{Q+R}{2}\right)^2 = 0$$

and after adding and subtracting N^2 we arrive at:

$$(J-N)^2 - \left(\frac{Q+R}{2} - N\right)^2 = 0$$

which shows that the length of *JN* is the radius of the nine-point circle.

From the formulas for the centroid and orthocentre one obtains the centre *N* of the nine-point circle:

$$N = \frac{3G+H}{4} = (P\,QR\,P + Q\,RP\,Q + R\,PQ\,R)(4PQ \wedge QR)^{-1}$$

that is, the point *N* also lies on the Euler's line. The relative distances between the circumcentre *O*, the centroid *G*, the nine-point circle centre *N* and the orthocentre *H* are shown in the figure 9.8.

Figure 9.8

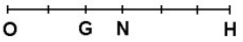

have:

$$A'C'\,A'D'^{-1} = (OC' - OA')\,(OD' - OA')^{-1} = (OC^{-1} - OA^{-1})\,(OD^{-1} - OA^{-1})^{-1}$$

$$= OC^{-1}\,(\,OA - OC\,)\,OA^{-1}\,[\,OD^{-1}\,(\,OA - OD\,)\,OA^{-1}\,]^{-1}$$

$$= OC^{-1}\,CA\,OA^{-1}\,[\,OD^{-1}\,DA\,OA^{-1}\,]^{-1} = OC^{-1}\,CA\,OA^{-1}\,OA\,DA^{-1}\,OD$$

$$= OC^{-1}\,AC\,AD^{-1}\,OD$$

Analogously we have:

$$B'D'\,B'C'^{-1} = OD^{-1}\,BD\,BC^{-1}\,OC$$

The product of both expressions is the cross ratio:

$$(\,A'\,B'\,C'\,D'\,) = A'C'\,A'D'^{-1}\,B'D'\,B'C'^{-1}$$

$$= OC^{-1}\,AC\,AD^{-1}\,BD\,BC^{-1}\,OC$$

$$= OC^{-1}\,(\,A\,B\,C\,D\,)\,OC = (\,A\,B\,C\,D\,)^*$$

Figure 10.4

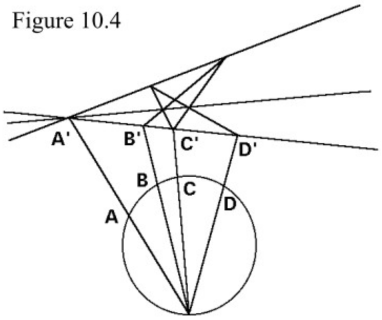

So, the cross ratio of four points on a line or a circle, which is real, is preserved, otherwise it becomes conjugated.

Finally let us see that the single ratio is the limit of the cross ratio for one point at infinity:

$$(\,\infty\,B\,C\,D\,) = \lim_{A \to \infty} AC\;AD^{-1}\;BD\;BC^{-1} = BD\;BC^{-1} = (\,B\,D\,C\,)$$

Analogously $(\,A\,\infty\,C\,D\,) = (\,A\,C\,D\,)$, etc.

The homography (Möbius transformation)

The *homography* is defined as the geometric transformation that preserves the complex cross ratio of four points on the plane. In order to specify a homography one must gives only the images A', B' and C' of three points A, B, and C because the fourth image D' is determined by the conservation of their cross ratio:

$$(\,A'\,B'\,C'\,D'\,) = (\,A\,B\,C\,D\,)$$

$$A'C'\,A'D'^{-1}\,B'D'\,B'C'^{-1} = AC\,AD^{-1}\,BD\,BC^{-1}$$

Arranging vectors we write:

$$A'D'^{-1}\,B'D' = A'C'^{-1}\,AC\,AD^{-1}\,BD\,BC^{-1}\,B'C' = A'C'^{-1}\,AC\;\;BC^{-1}\,B'C'\,AD^{-1}\,BD$$

After conjugation we find:

11. CONICS

Conic sections

The *conic sections* (or *conics*) are the intersections of a conic surface with any transversal plane (figure 11.1). The *proper conics* are the curves obtained with a plane not passing through the vertex of the cone. If the plane contains the vertex we have improper conics, which reduce to a pair of straight lines or a point. In general, I shall regard only proper conics, taking into account that the other case can be usually obtained as a limit case.

Let us consider the two spheres inscribed in the cone which are tangent to the plane of the conic. The spheres touch this plane at the points F and F', the *foci* of the conic section. On the other hand, both spheres touch the cone surface at two tangency circles lying in planes perpendicular to the cone axis. The *directrices* of the conic are defined as the intersections of these planes with the plane of the conic.

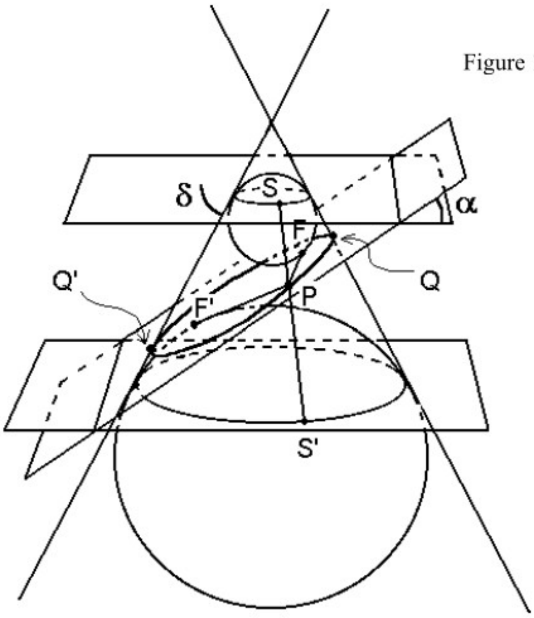
Figure 11.1

Let P be a point on the conic. Since it lies on the plane of the conic, the segment going from P to the focus F is tangent to the sphere. Since it belongs to the cone surface, the generatrix passing through P is also tangent to the sphere. PF and PS are tangent to the same sphere so that their lengths are equal: $|PF| = |PS|$. Also PF' and PS' are tangent to the other sphere so that their lengths are also equal: $|PF'| = |PS'|$. On the other hand the segment SS' of any directrix has constant length, what implies that the addition of distances from any point P on the conic to both foci is constant:

$$|PF|+|PF'|=|PS|+|PS'|=|SS'|=\text{constant}$$

This discussion slightly changes for a plane which intercepts the upper and lower cones. In this case, the distances from *P* to the upper sphere and to the upper branch of the curve must be considered negative[1]. If the plane is parallel to a generatrix of the cone these distances become infinite. This unified point of view for all conics using negative or infinite distances when needed will be the guide of this chapter. I shall not separate equations but only specify cases for distinct conics.

A lateral view of figure 11.1 is given in the figure 11.2. In this figure the point *P* has been drawn in both extremes of the conic (and *S* and *S'* also twice), although we must consider any point on the conic section. As indicated previously $|PF|=|PS|$. But $|PS|$ is proportional to the distance from *P* to the upper plane containing *S* and this distance is also proportional to the distance from *P* to the directrix *r* (figure 11.2):

$$|PF|=|PS|=\frac{\sin\alpha}{\sin\delta}\,\mathrm{d}(P,r)$$

The quotient of the sines of both angles is called the *eccentricity* e of the conic:

$$\mathrm{e}=\frac{\sin\alpha}{\sin\delta}>0$$

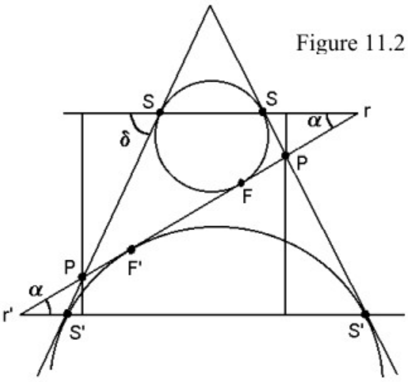

Figure 11.2

Then a conic can be also defined as the geometric locus of the points *P* whose distance from the focus *F* is proportional to the distance from the directrix *r* (figure 11.3).

$$|FP|=\mathrm{e}\,\mathrm{d}(P,r)$$

$|FP|$ is called the *focal radius*. Let *T* be the point of the directrix closest to the focus (figure 11.3). Then the *main axis of symmetry* of the conic is the line which is perpendicular to the directrix and passes through the focus. Under a reflection with respect the main axis of symmetry, the conic is preserved. The oriented distance from any point *P* to the directrix *r* is expressed, using the vector *FT*, as:

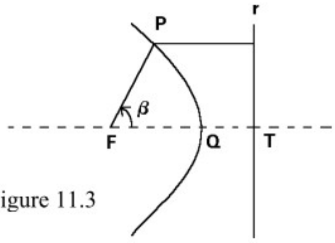

Figure 11.3

---

[1] To consider sensed distances is not a trouble but an advantage. For example, the distance from a point to a line can be also taken as an oriented distance, whose sign indicates the half-plane where the point lies.

$$\mathrm{d}(P,r)=\frac{FT\cdot PT}{|FT|}$$

Defined in this way, the distance from P to r is positive when P is placed on the half-plane that contains the focus of both in which the directrix divides the plane, and negative when P lies on the other half-plane. The equation of the conic becomes:

$$|FP|=\mathrm{e}\,\mathrm{d}(P,r)=\mathrm{e}\frac{FT\cdot PT}{|FT|}$$
$$|FP|\;|FT|=\mathrm{e}\,FT\cdot PT=\mathrm{e}\,FT\cdot(FT-FP)=\mathrm{e}\left(FT^2-FT\cdot FP\right)$$

Let us denote by β the angle between the main axis FT and the focal vector FP:

$$|FP|\;|FT|=\mathrm{e}\left(FT^2-|FT|\;|FP|\cos\beta\right)$$

$$|FP|\;|FT|\,(1+\mathrm{e}\;\cos\beta)=\mathrm{e}\,FT^2$$

From here one obtains the *polar equation* of a conic:

$$|FP|=\frac{|FT|}{1+\mathrm{e}\;\cos\beta}$$

According to the value of the eccentricity there are three types of proper conics:

1) For $0<\mathrm{e}<1$, the inclination α of the plane of the conic is lower than the inclination δ of the cone generatrix and the denominator is always positive. Then the focal radius is always positive and the points P form a closed curve on the focal half-plane, an *ellipse*.
2) For $\mathrm{e}=1$, $\alpha=\delta$, the plane of the conic is parallel to a cone generatrix. Since the denominator vanishes for $\beta=\pi$, the points P form an open curve called *parabola*. Except for this value of β, the radius is always positive and the curve lies on the focal half-plane.
3) For $\mathrm{e}>1$, $\alpha>\delta$, the plane of the conic intercepts both upper and lower cones. The denominator vanishes twice so that it is an open curve with two branches called *hyperbola* (figure 11.4). The lowest positive angle that makes the denominator zero is the asymptote angle β_a:

Figure 11.4

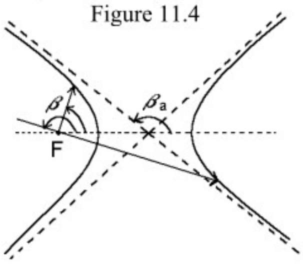

$$\beta_a=\arccos\left(-\frac{1}{\mathrm{e}}\right)$$

The maximum eccentricity of the hyperbolas as cone sections is obtained with $\alpha=\pi/2$:

$$1 < \mathrm{e} \le \frac{1}{\sin\delta}$$

The focal radius is positive (P lies on the focal branch) for the ranges of β:

$$0 < \beta < \beta_a \qquad \text{and} \qquad 2\pi - \beta_a < \beta < 2\pi$$

and negative (P lies on the non focal branch) for the range:

$$\beta_a < \beta < 2\pi - \beta_a$$

A negative focal radius means that we must take the sense opposite that determined by the angle (figure 11.4). For example, a radius –2 with an angle $7\pi/6$ is equivalent to a radius 2 with an angle of $\pi/6$. So the vertex T' of the non focal branch has an angle π (not zero as it would appear) and a negative focal radius.

Two foci and two directrices

Figure 11.5

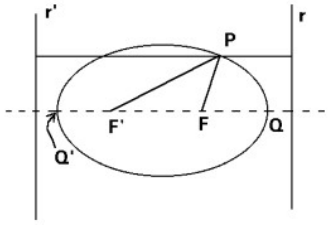

Figure 11.6

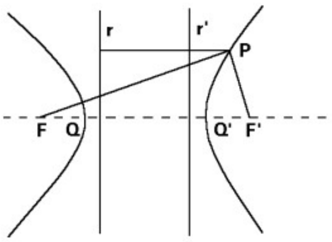

There are usually two spheres touching the cone surface and the conic plane, and hence there are also two foci. For the case of the ellipse, both tangent spheres are located in the same cone. For the case of the hyperbola, each sphere is placed in each cone so that each focus is placed next to each branch. In the case of the parabola, one sphere and the corresponding focus is placed at the infinity. For the improper conics, the tangent spheres have null radius and the foci are coincident with the vertex of the cone, which is the crossing point of both lines.

As the figure 11.1 shows for a proper conic, both directrices r and r' are parallel, and both foci F and F' are located on the axis of symmetry of the conic. Above we have already seen that the addition of oriented distances from both foci to any point P is constant. Let us see the relation with the distance between the directrices. If P is any point on the conic then:

$$|FP| + |F'P| = \mathrm{e}\,\mathrm{d}(P,r) + \mathrm{e}\,\mathrm{d}(P,r') = \mathrm{e}\,\mathrm{d}(r,r')$$

The sum of the oriented distances from any point P to any two parallel lines (the directrices in our case) is constant[2]. Then the addition of both focal radii of any point P

---

[2] Note that this statement is only right for oriented distances but not for positive distances.

on the conic is the product of the eccentricity multiplied by the oriented distance between both directrices. P can be any point, for example Q, the extreme of the conic:

$$|FP|+|F'P|=|FQ|+|F'Q|=|F'Q'|+|F'Q|=|QQ'|$$

whence it follows that the addition of both focal radii is $|QQ'|$, the *major diameter*. Specifying:

1) For the case of the ellipse (figure 11.5), both focal radii are positive and also the major diameter.
2) For the case of the parabola, one directrix is the line at the infinity and one focus is a point at the infinity on the main axis of symmetry. Hence the major diameter has an infinite value.
3) For the case of the hyperbola (figure 11.6), the focal radius of a point on the non focal branch is negative, its absolute value being higher than the other focal radius, which is positive, yielding a negative major diameter.

Vectorial equation

The polar equation of a conic may be written having as parameter the *focal distance* $|FQ|$ instead of the distance from the focus to the directrix $|FT|$:

$$|FP|=\frac{1+\mathrm{e}}{1+\mathrm{e}\ \cos\beta}|FQ|$$

Let Q' be the vertex of the conic closest to the focus F'. The distance from the focus F to Q' is found for $\beta=\pi$:

$$|FQ'|=\frac{1+\mathrm{e}}{1-\mathrm{e}}|FQ|$$

For the ellipse $|FQ'|$ is a positive distance; for the hyperbola it is a negative distance and for the parabola Q' is at infinity. Then the distance between both foci is:

$$|FF'|=|FQ'|-|F'Q'|=|FQ'|-|FQ|=\frac{2\ \mathrm{e}}{1-\mathrm{e}}|FQ|$$

On the other hand the major diameter is:

$$|QQ'|=|FQ|+|FQ'|=\frac{2|FQ|}{1-\mathrm{e}}$$

By dividing both equations, an alternative definition of the eccentricity is obtained:

$$\mathrm{e}=\frac{|FF'|}{|QQ'|}$$

The eccentricity is the ratio of the distance between both foci divided by the major diameter. For the case of ellipse, both distances are positive. For the case of the hyperbola, both distances are negative, so the eccentricity is always positive. When e = 0 both foci are coincident in the centre of a circle. Note that the circle is obtained when we cut the cone with a horizontal plane. In this case, the directrices are the line at the infinity.

The *vectorial equation* of a conic is obtained from the polar equation and contains the radius vector FP. Since FP forms with FQ an angle β (figure 11.3), FP is obtained from the unitary vector of FQ via multiplication by the exponential with argument β and by the modulus of FP yielding:

$$FP = \frac{1+\mathrm{e}}{1+\mathrm{e}\cos\beta} FQ\left(\cos\beta + e_{12}\sin\beta\right)$$

On the other hand, from the directrix property, one easily finds the following equation for a conic:

$$FP^2\, FT^2 = \mathrm{e}^2\,(\, FT^2 - FT \cdot FP\,)^2$$

F, T and e are parameters of the conic, and $P = (x, y)$ is the mobile point. Therefore from this equation we will also obtain a Cartesian equation of second degree. For example, let us calculate the Cartesian equation of an ellipse with eccentricity ½ and focus at the point (3, 4) and vertex at (4,5):

$$\mathrm{e} = 1/2 \qquad F = (\,3, 4\,) \qquad Q = (\,4, 5\,) \qquad P = (\,x, y\,)$$

$$FT = \frac{1+\mathrm{e}}{\mathrm{e}} FQ = 3\,e_1 + 3\,e_2 \qquad T = F + FT = (\,6, 7\,)$$

$$FP = (\,x-3\,)\,e_1 + (\,y-4\,)\,e_2$$

Then the equation of the conic is:

$$[\,(\,x-3\,)^2 + (\,y-4\,)^2\,]\,18 = \frac{1}{4}[\,18 - (\,3\,(x-3) + 3\,(\,y-4\,)\,)\,]^2$$

After simplification:

$$7\,x^2 - 2\,x\,y + 7\,y^2 - 22\,x - 38\,y + 31 = 0$$

The Chasles' theorem

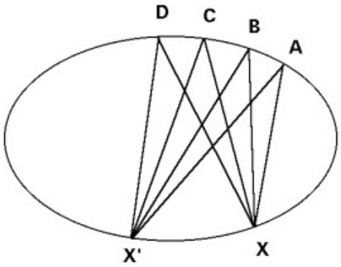

Figure 11.7

According to this theorem[3], the projective cross ratio of any four given points A, B, C and D on a conic with respect to a point X also on this conic is constant independently of the choice of

---

[3] Michel Chasles, *Traité des sections coniques*, Gauthier-Villars, Paris, 1865, p. 3.

("

sector, the modulus of the hyperbolic radius is imaginary. The abscissa x and ordinate y divided by r, the modulus of the radius vector, is also imaginary so the angle is a real quantity plus a right angle and the hyperbolic sine and cosine are exchanged. The values of the angles consistent with the analytic continuation are displayed in the figure 13.1.

Congruence of segments and angles

Two segments (vectors) are *congruent* (have *equal length*) if their determinants are equal, in other words, when one segment can be obtained from the other through an isometry. Segments having equal length always lie on the same sector.

Two hyperbolic angles are said to be *congruent* (or *equal*) if they have the same moduli, that is, if they intercept arcs of hyperbola with the same hyperbolic length.

A triangle is said to be *isosceles* if it has two congruent sides. In colloquial language we talk about "sides with equal length", even "equal sides". Let us see the *isosceles triangle theorem*: the angles adjacent to the base of an isosceles triangle are equal. The proof uses the outer product. The triangle area is the half of the outer product of any two sides, as proved below in the section on the area. Suppose that the sides a and b have the same modulus. Then:

$$|a|\,|c|\,|\sinh\beta| = |b|\,|c|\,|\sinh\alpha| \qquad \Rightarrow \qquad |\beta| = |\alpha|$$

Isometries

In the Euclidean plane, an isometry is a geometric transformation that preserves the modulus of vectors and complex numbers. Now I give a more general definition: a geometric transformation is an *isometry* if it preserves the determinant of every element of the geometric algebra. In fact, the isometries are the inner automorphism of matrices:

$$\mathbf{A'} = \mathbf{B}^{-1}\,\mathbf{A}\,\mathbf{B} \Rightarrow \qquad \det \mathbf{A'} = \det \mathbf{A}$$

When **B** represents a Euclidean vector, the isometry is a reflection in the direction of **B**. If a complex number of argument α is represented by **B**, the transformation is a rotation of angle 2α.

We wish now to obtain the isometries for hyperbolic vectors. The hyperbolic rotation of a hyperbolic vector (figure 13.2) is obtained through the product by an exponential having unity determinant:

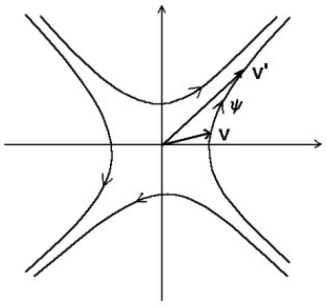

Figure 13.2

$$v'_2 e_2 + v'_{21}\, e_{21} = (v_2 e_2 + v_{21}\, e_{21})(\cosh\psi + \sinh\psi\, e_1)$$

Writing the components, we have:

$$\begin{cases} v'_2 = v_2 \cosh\psi + v_{21}\sinh\psi \\ v'_{21} = v_2 \sinh\psi + v_{21}\cosh\psi \end{cases}$$

which is the Lorentz transformation[3] of the relativity. When a vector is turned through a positive angle ψ, its extreme follows the hyperbola in the direction shown by the arrowheads in the figure 13.2, as is deduced from the components. So, they indicate the geometric positive sense of hyperbolic angles (a not trivial question). Since the points of intersections with the axis corresponds to $\psi = 0$ plus multiples of $\pi\, e_{12}/2$, the signs of the hyperbolic angles are determined: positive in the first and third quadrant, and negative in the second and fourth quadrant.

We can also write the hyperbolic rotation as an inner automorphism of matrices by using the half argument identity for hyperbolic trigonometric functions:

$$\cosh\psi + e_1 \sinh\psi \equiv \left(\cosh\frac{\psi}{2} + e_1 \sinh\frac{\psi}{2} \right)^2$$

and the permutative property:

$$v'_2 e_2 + v'_{21}\, e_{21} = \left(\cosh\frac{\psi}{2} - \sinh\frac{\psi}{2} e_1 \right)(v_2 e_2 + v_{21}\, e_{21})\left(\cosh\frac{\psi}{2} + \sinh\frac{\psi}{2} e_1 \right)$$

If $z = \exp(e_1 \psi/2)$, then the rotation is written as:

$$v' = z^{-1}\, v\, z$$

Now it is not needed that z be unitary and can have any modulus. Also, we see that the hyperbolic rotation leaves the hyperbolic numbers invariant.

Analogously to reflections in the Euclidean plane, the hyperbolic reflection of a hyperbolic vector v with respect to a direction given by the vector u is:

$$v' = u^{-1}\, v\, u = u^{-1}\, (\, v_{\|} + v_{\perp})\, u = v_{\|} - v_{\perp}$$

because the proportional vectors commute and those which are orthogonal anticommute (also in the hyperbolic case, because the inner product is zero). Every pair of orthogonal directions (for example $u_{\|}$ and $u_{\perp}$ in figure 13.3) are always seen by us as having symmetry with respect to the quadrant bisectors. Since the determinant is preserved, a vector and its reflection always belong to the same sector, the hyperbolic angle between each vector and the direction of reflection being equal with opposite sign.

Figure 13.3

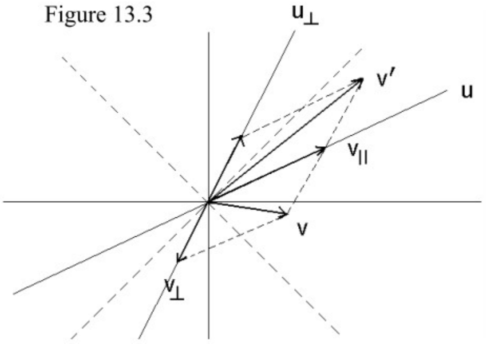

---

[3] The hyperbolic vector in the space-time is $ct\, e_0 + x\, e_1$ where x is the space coordinate, t the time and c the light celerity. The argument of the hyperbolic rotation ψ is related with the relative velocity V of both frames of reference through $\psi = \text{arg tgh } V/c$.

Finally, note that the extremes of a vector and its transformed vector under a hyperbolic reflection (or any isometry) lie on an equilateral hyperbola.

Theorems about angles

The sum of the oriented angles of a hyperbolic triangle is minus two right circular angles. To prove this theorem, draw a line parallel to a side and apply the Z-theorem (figure 13.4): they will form an angle between opposite directions, which has the value $-\pi\ e_{12}$. The minus sign is caused by the fact that, following the positive orientation of angles, a right angle is subtracted each time an asymptote is trespassed.

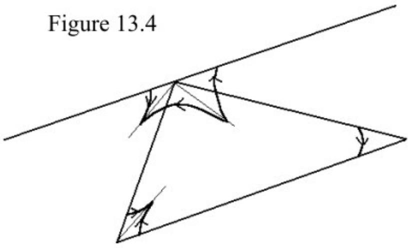

Figure 13.4

Look at the figure 13.5 where an angle inscribed in an equilateral hyperbola is drawn. A diameter divides the inscribed angle into the angles α and β. Then we draw two radius from the origin to both extremes of the angle. All the radius have the same length, so that the upper and lowest triangle are isosceles and have two equal angles. Since the sum of the angles of each hyperbolic triangle is $-\pi\ e_{12}$, the third angle is $-\pi\ e_{12}-2\alpha$ and $-\pi\ e_{12}-2\beta$ respectively. So the supplementary angles are 2α and 2β, and hence the theorem follows: the inscribed angle on a hyperbola $x^2 - y^2 = r^2$ is the half of the central angle (intercepted arc) so that it is constant and independent of the location of its vertex.

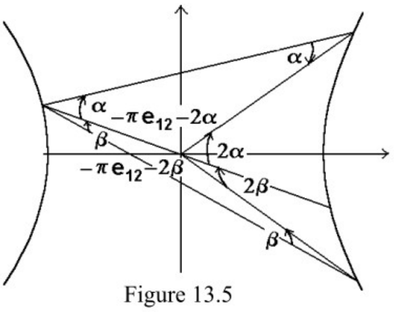

Figure 13.5

For example, take the points (5, 3) and (5, −3) as extremes of an inscribed angle with the vertex placed at the negative branch, for example at (−4, 0) or (−5, −3). Even you can take the vertex at the positive branch, e. g. at (4, 0). Anyway its value is log 2. Now calculate the central angle or intercepted arc (with vertex at the origin) and see that its value is log 4, twice the inscribed angle.

Distance between points

The *distance between two points* on the hyperbolic plane is defined as the modulus of the vector going from one point to the another:

$$d(P,Q) = |\,PQ\,| = \sqrt{(x_Q - x_P)^2 - (y_Q - y_P)^2}$$

In the hyperbolic plane, the distance is a real or imaginary positive number (depending on the sector), which has the following properties:

1) $d(P,Q) = d(Q,P)$
2) If the vectors *PQ*, *QR* and *PR* lie on the positive half sector then they fulfil the triangular inequality:

$$d(P,R) \geq d(P,Q) + d(Q,R)$$

The prove is obtained by means of the inner product:

$$PR^2 = (PQ + QR)^2 = PQ^2 + QR^2 + 2\, PQ \cdot QR$$

$$= PQ^2 + QR^2 + 2\,|\,PQ\,|\,|\,QR\,|\cosh\psi \geq PQ^2 + QR^2 + 2\,|\,PQ\,|\,|\,QR\,|$$

The extraction of the square root yields the triangular inequality:

$$|\,PR\,| \geq |\,PQ\,| + |\,QR\,|$$

So the sum of two sides of a triangle is smaller than the third side whenever they are taken in the positive half sector. For example, let us apply this statement to the triangle with vertices *A* = (5, 3), *B* = (1, 0) and *C* = (10, 1). Taking the modulus of the sides positive, we have:

$$|\,BC\,| = \sqrt{80} \qquad |\,BA\,| = \sqrt{7} \qquad |\,AC\,| = \sqrt{21}$$

$$|\,BC\,| \geq |\,BA\,| + |\,AC\,| \qquad \Rightarrow \qquad \sqrt{80} \geq \sqrt{7} + \sqrt{21}$$

This result must be commented with more detail. Firstly, we see that the straight path has the highest length among the possible paths between two given points. In the Minkowski's space-time this fact causes the *twin paradox*. Since the pseudo-Euclidean length of the path is the proper time for each person, the brother who followed the straight path -going in an inertial frame- has aged more than the brother who followed another path -subjected to accelerations-.

Area on the hyperbolic plane

Now I give a general geometric definition of area valid for both Euclidean and hyperbolic planes. If a parallelogram has orthogonal sides, then the modulus of its *area* is equal to the product of the lengths of the orthogonal sides. In Euclidean geometry a parallelogram with orthogonal sides is called a *rectangle*. In the hyperbolic plane, two sides are orthogonal if their directions are seen by us as symmetric with respect to the direction of the quadrant bisector. So, I have preferred to avoid the word *rectangle* in this case, while the term *parallelogram* is still valid within this context.

Which is the suitable algebraic expression to calculate the area of any parallelogram? Suppose that it is the outer product. If a and b are the sides of the parallelogram then:

$$|A| = |a \wedge b| = |a||b||\sinh\psi(a,b)|$$

Let us see the consistence of this expression. First at all, recall that the angle between orthogonal hyperbolic vectors is a right circular angle:

$$\psi_{\perp} = \pm\frac{\pi}{2}e_{12}$$

so that the area of a parallelogram with orthogonal sides is the product of their lengths as expected:

$$|A| = |a||b|\left|\sinh\left(\pm\frac{\pi}{2}e_{12}\right)\right| = |a||b|\sin\frac{\pi}{2} = |a||b|$$

For any parallelogram not having orthogonal sides, the area is the product of the *base* (one side) for the *altitude* (the projection of the other side onto the direction orthogonal to the base) and this product is only given by the outer product because its anticommutativity removes the proportional projection:

$$|A| = |a \wedge b| = |(a_{\parallel} + a_{\perp}) \wedge b| = |a_{\perp}||b|$$

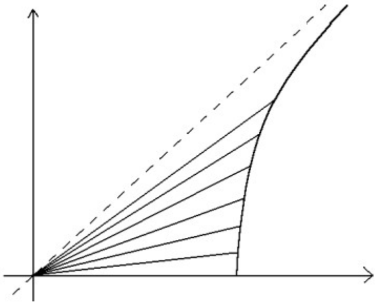

Figure 13.6

The expression for the area in Cartesian components is equal to that for the Euclidean plane. It means that we can calculate areas graphically in the usual way, fact that has allowed till now to define the hyperbolic trigonometric functions from the area scanned by the hyperbola radius in the Euclidean plane, in spite of not being their proper plane. However, as I have explained in the footnote at the page 142, we must perceive the pseudo-Euclidean nature of the area in the hyperbolic plane. The figure 13.6 displays how are the radii perpendicular to the hyperbola because the bisector of the radius and the tangent vector is parallel to the quadrant bisector. Also it may be proved analytically by means of differentiation of the equation $x^2 - y^2 = r^2$:

$$2\,x\,dx - 2\,y\,dy = 0 \quad \Leftrightarrow \quad r \cdot dr = 0$$

showing that the radius vector and its differential are orthogonal for the hyperbola with constant radius. On the other hand, observe that the triangles drawn in the figure 13.6 are isosceles and have two equal angles approaching the right angle value at the limit of null area.

Diameters of the hyperbola and Apollonius' theorem

At the page 129, I postponed an interpretation of the central equation for the hyperbola analogous to the cylindrical angles for the ellipse because it is properly of pseudo-Euclidean nature. Here I develop this interpretation.

If *P* is any point on a hyperbola with major and minor half-axis *OQ* and *OS* (figure 11.17 reproduced below) then its central equation is:

$$OP = \pm \frac{OQ \cosh\chi + OS \dfrac{\sinh\chi}{\sqrt{e^2-1}}}{\sqrt{\cosh^2\chi - \dfrac{\sinh^2\chi}{e^2-1}}}$$

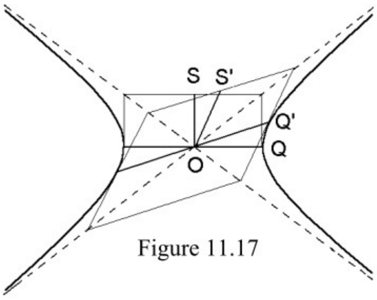

Figure 11.17

Since *OQ* and *OS* are orthogonal, the square of this equation fulfils:

$$\frac{OP^2}{OQ^2}\cosh^2\chi + \frac{OP^2}{OS^2}\sinh^2\chi = 1$$

Both half-axis are related with the eccentricity through:

$$OS^2 = \left(1-e^2\right)OQ^2$$

Note that $OS^2<0$ since it is the square of an ordinate on the hyperbolic plane.

Introducing the hyperbolic angle ψ in the following way:

$$\cosh\psi = \frac{\cosh\chi}{\sqrt{\cosh^2\chi - \dfrac{\sinh^2\chi}{e^2-1}}}$$

now the equation of the hyperbola becomes:

$$OP = \pm\left(OQ\cosh\psi + OS\sinh\psi\right)$$

Figure 13.7

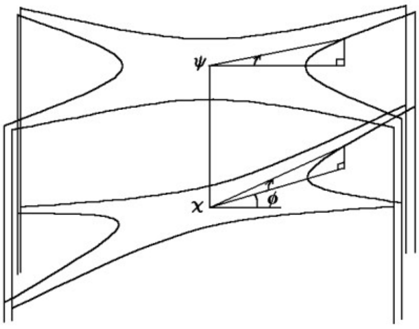

Observe in the figure 13.7 that any hyperbola can be obtained as an intersection of a transverse plane with the equilateral hyperbolic prism. The plane of the acute hyperbola (with $1<e<\sqrt{2}$) forms an angle ϕ with respect to the horizontal, and hence:

$$\cosh\psi = \frac{\cosh\chi\cos\phi}{\sqrt{\cosh^2\chi\cos^2\phi - \sinh^2\chi}}$$

So the eccentricity e of the hyperbola is related with the obliquity ϕ of the transverse plane through the relationship:

$$\cos\phi = \sqrt{\mathrm{e}^2 - 1} \quad \text{with} \quad 1 < \mathrm{e} < \sqrt{2}$$

The Apollonius' conjugate diameters of any hyperbola are the intersections of the transverse plane with a pair of orthogonal axial planes; in other words, two radii are conjugate (figure 11.17) if their projections onto the horizontal plane are turned through the same hyperbolic angle φ:

$$OQ' = OQ\cosh\varphi + OS\sinh\varphi$$

$$OS' = OQ\sinh\varphi + OS\cosh\varphi$$

Our Euclidean eyes see the horizontal projections as symmetric lines with respect to the quadrant bisector. However, they are actually orthogonal because:

$$OQ'^2 - OS'^2 = OQ^2 - OS^2$$

and, therefore, can be taken as a new system of orthogonal coordinates. Even we can draw a new picture with the new diameters on the Cartesian axis.

The central equation of the hyperbola using the rotated axis is:

$$OP = \pm\left(OQ'\cosh(\psi - \varphi) + OS'\sinh(\psi - \varphi)\right)$$

which shows that a hyperbolic rotation of the coordinate axis has been made with respect to the principal diameter of the hyperbola.

The law of sines and cosines

Since the modulus of the area is identical on the Euclidean and hyperbolic planes, a parallelogram is divided by its diagonal in two triangles of equal area. This statement is somewhat subtle since the Euclidean congruence of triangles is not valid in the hyperbolic plane. I shall return to this question later. Now we only need to know that the area of a hyperbolic triangle is the half of the outer product of any two sides.

Following the perimeter of a triangle, let *a*, *b*, and *c* be its sides respectively opposite to the angles α, β and γ. Then the angles formed by the oriented sides are supplementary of the angles of the triangle and:

$$a \wedge b = b \wedge c = c \wedge a \;\Rightarrow\; -|\,a\,|\,|\,b\,|\sinh\gamma = -|\,b\,|\,|\,c\,|\sinh\alpha = -|\,c\,|\,|\,a\,|\sinh\beta$$

$$\frac{|\,a\,|}{\sinh\alpha} = \frac{|\,b\,|}{\sinh\beta} = \frac{|\,c\,|}{\sinh\gamma}$$

which is the *law of sines*.

From $a + b + c = 0$, we have:

$$a^2 = (-b-c)^2 = b^2 + c^2 + 2\, b\cdot c \qquad \Rightarrow \qquad a^2 = b^2 + c^2 - 2\,|\,b\,|\,|\,c\,|\cosh\alpha$$

which is the *law of cosines*. And also:

$$b^2 = a^2 + c^2 - 2\,|\,a\,|\,|\,c\,|\cosh\beta$$

$$c^2 = a^2 + b^2 - 2\,|\,a\,|\,|\,b\,|\cosh\gamma$$

When applying both theorems, we must take care with the sides having imaginary length and the signs of the angles and trigonometric functions.

As an application of the law of sines and cosines, consider the hyperbolic triangle with vertices *A*=(5, 3), *B*=(1, 0), *C*=(10, 1), whose sides belong to the real sector (figure 13.8):

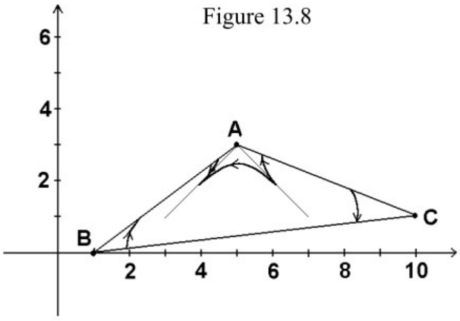

Figure 13.8

$$AB = B - A = -4\,e_2 - 3\,e_{21} \qquad |\,AB\,| = \sqrt{(-4)^2 - (-3)^2} = \sqrt{7}$$

$$BC = C - B = 9\,e_2 + e_{21} \qquad |\,BC\,| = \sqrt{9^2 - 1^2} = \sqrt{80}$$

$$CA = A - C = -5\,e_2 + 2\,e_{21} \qquad |\,CA\,| = \sqrt{(-5)^2 - 2^2} = \sqrt{21}$$

$$BC^2 = CA^2 + AB^2 - 2\,|\,CA\,|\,|\,AB\,|\cosh\alpha \qquad \cosh\alpha = -\frac{26}{7\sqrt{3}}$$

$$CA^2 = AB^2 + BC^2 - 2\,|\,AB\,|\,|\,BC\,|\cosh\beta \qquad \cosh\beta = \frac{33}{4\sqrt{35}}$$

$$AB^2 = BC^2 + CA^2 - 2\,|\,BC\,|\,|\,CA\,|\cosh\gamma \qquad \cosh\gamma = \frac{47}{4\sqrt{105}}$$

From where it follows that:

$$\alpha = -1.3966\ldots - \pi\, e_{12} \qquad \beta = 0.8614\ldots \qquad \gamma = 0.5352\ldots$$

I have obtained their signs from the definition of the angles $\alpha = BAC$, $\beta = CBA$, $\gamma = ACB$ and the geometric plot (figure 13.8). Note that $\alpha + \beta + \gamma = -\pi\, e_{12}$ and they fulfil the law of sines:

$$\frac{|\,BC\,|}{\sinh\alpha} = \frac{|\,CA\,|}{\sinh\beta} = \frac{|\,AB\,|}{\sinh\gamma}$$

because $\sinh\alpha = \sinh(-1.3966+\pi e_{12}) = -\sinh(-1.3966) = \sinh 1.3966$.

Consider now another triangle $A = (2, 4)$, $B = (1,0)$ and $C = (6, 1)$, having sides on real and imaginary sectors (figure 13.9):

$$AB = B - A = -e_2 - 4\, e_{21}$$

$$|\,AB\,| = \sqrt{(-1)^2 - (-4)^2} = \sqrt{15}\; e_{12}$$

$$BC = C - B = 5\, e_2 + e_{21}$$

$$|\,BC\,| = \sqrt{5^2 - 1^2} = \sqrt{24}$$

$$CA = A - C = -4\, e_2 + 3\, e_{21}$$

$$|\,CA\,| = \sqrt{(-4)^2 - 3^2} = \sqrt{7}$$

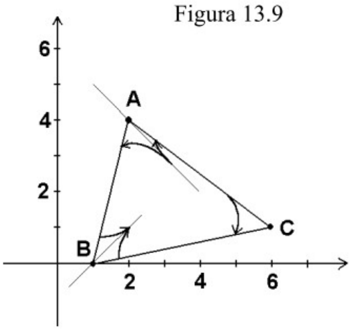

Figura 13.9

$$BC^2 = CA^2 + AB^2 - 2\,|\,CA\,|\,|\,AB\,|\cosh\alpha \qquad \cosh\alpha = \frac{16\, e_{12}}{\sqrt{105}}$$

$$CA^2 = AB^2 + BC^2 - 2\,|\,AB\,|\,|\,BC\,|\cosh\beta \qquad \cosh\beta = \frac{-\,e_{12}}{6\sqrt{10}}$$

$$AB^2 = BC^2 + CA^2 - 2\,|\,BC\,|\,|\,CA\,|\cosh\gamma \qquad \cosh\gamma = \frac{23}{2\sqrt{42}}$$

From the last hyperbolic cosine, which is real, we find the hyperbolic sine for γ, which is a positive angle as shown by the plot:

$$\sinh\gamma = \frac{19}{2\sqrt{42}}$$

which we may use in the law of sines:

$$\frac{\sqrt{24}}{\sinh\alpha} = \frac{\sqrt{7}}{\sinh\beta} = \frac{\sqrt{15}\, e_{12}}{19/2\sqrt{42}} \qquad \Rightarrow \qquad \sinh\alpha = -\frac{19\, e_{12}}{\sqrt{105}} \qquad \sinh\beta = -\frac{19\, e_{12}}{6\sqrt{10}}$$

Recalling that for ψ real:

$$\sinh\left(\psi \pm e_{12}\,\frac{\pi}{2}\right) \equiv \pm e_{12}\cosh\psi \qquad\qquad \cosh\left(\psi \pm e_{12}\,\frac{\pi}{2}\right) \equiv \pm e_{12}\sinh\psi$$

the angles α, β and γ follow:

$$\alpha = -1.2284... - \frac{\pi}{2} e_{12} \qquad \beta = 0.0527... - \frac{\pi}{2} e_{12} \qquad \gamma = 1.1757$$

Observe that the addition of the three angles is $-\pi\ e_{12}$, as expected. According to the definition $\alpha = BAC$, $\beta = CBA$, $\gamma = ACB$, let you see the consistence with the geometric plot. The angle γ has positive sign as shown by the figure 13.9. The bisector of α parallel to the quadrant bisector divides it in two angles, one real and the other complex (with imaginary part $\pi\ e_{12}\ /2$). The algebraic addition of both angles is α.. Taking into account that the second has opposite orientation and must be subtracted and predominates over the first, the negative value of α is explained.

Hyperbolic similarity

Two triangles ABC and $A'B'C'$ are said to be *directly similar*[4] and their vertices and sides denoted with the same letters are *homologous* if:

$$AB\ BC^{-1} = A'B'\ B'C'^{\ -1} \qquad \Rightarrow \qquad AB^{-1}\ A'B' = BC^{-1}\ B'C'$$

One can prove easily that the third quotient of homologous sides also coincides with the other quotients:

$$AB^{-1}\ A'B' = BC^{-1}\ B'C' = CA^{-1}\ C'A' = r$$

The *similarity ratio* r is defined as the quotient of every pair of homologous sides, which is a hyperbolic number. The modulus of the similarity ratio is the size ratio and the argument is the angle of rotation of the triangle $A'B'C'$ with respect to the triangle ABC.

$$r = \frac{|\,A'B'\,|}{|\,AB\,|} \exp\left[\alpha\left(AB, A'B'\right) e_1\right]$$

The definition of similarity is generalised to any pair of polygons in the following way. The polygons $ABC...Z$ and $A'B'C'...Z'$ are said to be directly similar with similarity ratio r and the sides denoted with the same letters to be homologous if:

$$r = AB^{-1}\ A'B' = BC^{-1}\ B'C' = CD^{-1}\ C'D' = ... = YZ^{-1}\ Y'Z' = ZA^{-1}\ Z'A'$$

Here also, the modulus of r is the size ratio of both polygons and the argument is the angle of rotation. The fact that the homologous exterior and interior angles are equal for directly similar polygons is trivial because:

$$B'A'\ B'C'^{\ -1} = BA\ BC^{-1} \qquad \Rightarrow \qquad \text{angle } A'B'C' = \text{angle } ABC$$

$$C'B'\ C'D'^{-1} = CB\ CD^{-1} \qquad \Rightarrow \qquad \text{angle } B'C'D' = \text{angle } BCD \qquad \text{etc.}$$

---

[4] A direct similarity is also called a *similitude* and an opposite similarity sometimes an *antisimilitude*.

The direct similarity is an equivalence relation since it has the reflexive, symmetric and transitive properties. This means that there are classes of equivalence with directly similar figures.

A similitude with $|r|=1$ is a *displacement*, since both polygons have the same size and orientation.

Two triangles *ABC* and *A'B'C'* are *oppositely similar* and the sides denoted with the same letters are *homologous* if:

$$AB\,BC^{-1} = (\,A'B'\,B'C'^{-1}\,)^{*} = B'C'^{-1}\,A'B'$$

where the asterisk denotes the hyperbolic conjugate. The former equality cannot be arranged into quotients of pairs of homologous sides as done before. Because of this, the similarity ratio cannot be defined for the opposite similarity but only the size ratio, which is the quotient of the lengths of any two homologous sides. An opposite similarity is always the composition of a reflection in any line and a direct similarity.

$$AB\,BC^{-1} = v^{-1}\,A'B'\,B'C'^{-1}\,v \quad\Leftrightarrow\quad BC^{-1}\,v^{-1}\,B'C' = AB^{-1}\,v^{-1}\,A'B' \quad\Leftrightarrow$$

$$BC^{-1}\,(\,v^{-1}\,B'C'\,v\,) = AB^{-1}\,(\,v^{-1}\,A'B'^{-1}\,v\,) = r$$

where *r* is the ratio of a direct similarity whose argument is not defined but depends on the direction vector *v* of the reflection axis. Notwithstanding, this expression allows to define the opposite similarity of two polygons. So two polygons *ABC...Z* and *A'B'C'...Z'* are oppositely similar and the sides denoted with the same letters are homologous if for any hyperbolic vector *v* the following equalities are fulfilled:

$$AB^{-1}\,(\,v^{-1}\,A'B'^{-1}\,v\,) = BC^{-1}\,(\,v^{-1}\,B'C'\,v\,) = \dots. = ZA^{-1}\,(\,v^{-1}\,Z'A'\,v\,)$$

that is, if after a reflection one polygon is directly similar to the other. The opposite similarity is not reflexive nor transitive and there are not classes of oppositely similar figures.

An opposite similarity with $|r|=1$ is called a *reversal*, since both polygons have the same size but opposite orientations.

Obviously the plot of similar figures on the hyperbolic plane breaks our Euclidean intuition about figures with the same form. The algebraic definition is very rigorous and clear, but we must change our visual illusions. So I recommend the exercise 13.6.

This chapter is necessarily unfinished because the geometry of the hyperbolic plane can be developed and studied with the same extension and profundity as for the Euclidean plane. Then, it is obvious that many theorems should follow in the same way as Euclidean geometry. As a last example I will define the power of a point with respect to a hyperbola.

Power of a point with respect to a hyperbola with constant radius

The locus of the points placed at a fixed distance *r* from a given point *O* is an equilateral hyperbola centred at this point with equation:

$$|OP| = r \qquad (x - x_O)^2 - (y - y_O)^2 = r^2$$

The *power of a point* with respect to this hyperbola is the product of both oriented distances from P to the intersections R and R' of a line passing through P with the hyperbola. The power of a point is constant for every line of the pencil of lines of P. In the proof the inscribed angle theorem is used (figure 13.10): the angles $S'RR'$ and $S'SR'$ are equal so the triangles SPR' and RPS' are oppositely similar. Then:

$$PR'\,PS^{-1} = PR^{-1}PS' \;\Rightarrow\; PR\,PR' = PS'\,PS$$

Figure 13.10

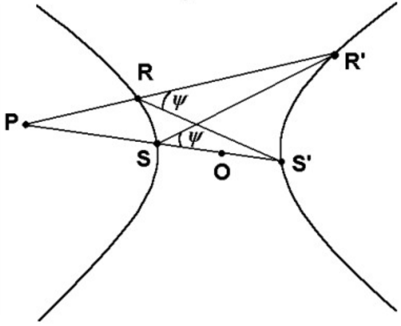

Developing the product of distances on the line passing through the centre of the hyperbola we find:

$$PS\,PS' = (PO + OS)(PO + OS') = PO^2 + OS\,OS' = (x_P - x_O)^2 - (y_P - y_O)^2 - r^2$$

that is, the power of a point is obtained by substitution of the coordinates of P on the hyperbola equation.

Exercises

13.1 Let $A = (2, 2)$, $B = (1, 0)$ and $C = (5, 3)$ be the vertices of a hyperbolic triangle. Calculate all the sides and angles and also the area.

13.2 Turn the vector $2\,e_2 + e_{21}$ through an angle $\psi = \log 2$. Do a reflection in the direction $3\,e_2 - e_{21}$. Make also an inversion with radius 3.

13.3 Find the direction and normal vectors of the line $y = 2\,x + 1$ and calculate the perpendicular line passing through the point (3, 1).

13.4 Calculate the power of the point $P = (-7, 3)$ with respect to the hyperbola $x^2 - y^2 = 16$ using the intersections of the lines $y = 3$, $y = -3x - 18$ and that passing through the centre of the hyperbola $y = -3x / 7$. See that in all cases the power of P is equal to the value found by substitution in the Cartesian equation.

13.5 In this chapter I have deduced the law of sines and cosines. Therefore a law of tangents should be expected. Find and prove it.

13.6 Check that the triangle $A = (0, 0)$, $B = (5, 0)$ and $C = (5, 3)$ is directly similar to the triangle $A' = (0, 0)$, $B' = (25, -15)$, $C' = (16, 0)$. Find the similarity ratio and the rotation and dilation of the corresponding homothety. Draw the triangles and you will astonish.

FOURTH PART: PLANE PROJECTIONS OF TRIDIMENSIONAL SPACES

The complete study of the geometric algebra of the tridimensional spaces falls out the scope of this book. However, due to the importance of the Earth charts and of the Lobachevsky's geometry, the first one being more practical and the second one more theoretical, I have written this last section. In order to make the explanations clearer, the tridimensional geometric algebra has been reduced to the minimal concepts, enhancing the plane projections.

The geometric quality of being Euclidean or pseudo-Euclidean is not the signature + or − of a coordinate, but the fact that two coordinates have the same or different signature, in other words, it is a characteristic of a plane. For instance, a plane with signatures + + is equivalent, from a geometric point of view, to another with − − . Therefore, only two kinds of three-dimensional spaces exist: the room space where all the planes are Euclidean (signatures + + + or − − −), and the pseudo-Euclidean space, which has one Euclidean plane and two orthogonal pseudo-Euclidean planes (signatures + − − or + + −).

14. SPHERICAL GEOMETRY IN THE EUCLIDEAN SPACE

The geometric algebra of the Euclidean space

A *vector* of the Euclidean space is an oriented segment in this space with direction and sense, although it can represent other physical magnitudes such as forces, velocities, etc. The set of all the segments (geometric vectors) together with their addition (parallelogram rule) and the product by real numbers (dilation of vectors) has a structure of vector space, symbolised with V_3. Every vector in V_3 is of the form:

$$v = v_1\, e_1 + v_2\, e_2 + v_3\, e_3$$

where e_i are three unitary perpendicular vectors, which form the base of the space. If we define an associative product (*geometric* or *Clifford product*) being a generalisation of that defined for the plane in the first chapter of this book, we will arrive to:

$$e_i^2 = 1 \qquad \text{and} \qquad e_i\, e_j = -e_j\, e_i \ \text{ for } i \neq j$$

In general, the square of a vector is the square of its modulus and perpendicular vectors anticommute whereas proportional vectors commute.

From the base vectors one deduces that the geometric algebra generated by the space V_3 has eight components:

$$Cl(V_3) = Cl_{3,\,0} = \langle 1, e_1, e_2, e_3, e_{23}, e_{31}, e_{12}, e_{123} \rangle$$

Let us see with more detail the product of two vectors:

$$v\, w = (v_1\, e_1 + v_2\, e_2 + v_3\, e_3)(w_1\, e_1 + w_2\, e_2 + w_3\, e_3) = v_1\, w_1 + v_2\, w_2 + v_3\, w_3$$
$$+ (v_2\, w_3 - v_3\, w_2)\, e_{23} + (v_3\, w_1 - v_1\, w_3)\, e_{31} + (v_1\, w_2 - v_2\, w_1)\, e_{12}$$

The product (or quotient) of two vectors is said a *quaternion*[1]. The quaternions are the even subalgebra of $Cl_{3,\,0}$ that generalise the complex numbers to the space. Splitting a quaternion in the real and bivector parts, we obtain the *inner* (or *scalar*) product and the *outer* (or *exterior*) product respectively:

$$v \cdot w = v_1\, w_1 + v_2\, w_2 + v_3\, w_3$$

$$v \wedge w = \left(v_2\, w_3 - v_3\, w_2\right) e_{23} + \left(v_3\, w_1 - v_1\, w_3\right) e_{31} + \left(v_1\, w_2 - v_2\, w_1\right) e_{12}$$

The *bivectors* are oriented plane surfaces and indicate the direction of planes in the space. Who be acquainted with the vector analysis will say that both vectors and bivectors are the same thing. This confusion was originated by Hamilton[2] himself, and continued by the founders of vector analysis, Gibbs and Heaviside. However, vectors and bivectors are different things just as physicists have experienced and know long time ago. The proper vectors are called usually "polar vectors" while the pseudo-vectors that actually are bivectors are usually called "axial vectors". The following magnitudes are vectors: of course a geometric segment, but also a velocity, an electric field, the momentum, etc. On the other hand, the oriented area is, of course, a bivector, but also the angular momentum, the angular velocity and the magnetic field. As a criterion to distinguish both kind of magnitudes one uses the reversal of coordinates, which changes the sense of vectors while leaves bivectors invariant.

The product of two bivectors yields a real number plus a bivector. Both parts can be separated as the symmetric and antisymmetric product. The symmetric product is a real number and its negative value will be denoted here with the symbol • while the antisymmetric product is also a bivector and will be noted here with the symbol × :

$$v \bullet w = -\frac{1}{2}\left(v\, w + w\, v\right) = v_{23}\, w_{23} + v_{31}\, w_{31} + v_{12}\, w_{12}$$

$$v \times w = -\frac{1}{2}\left(v\, w - w\, v\right)$$

$$= \left(v_{31}\, w_{12} - v_{12}\, w_{31}\right) e_{23} + \left(v_{12}\, w_{23} - v_{23}\, w_{12}\right) e_{31} + \left(v_{23}\, w_{31} - v_{31}\, w_{23}\right) e_{12}$$

$$v\, w = -v \bullet w - v \times w$$

Let us see what happens with the outer product of three vectors. According to the extension theory of Grassmann, the product $u \wedge v \wedge w$ is the oriented volume generated by the surface represented by the bivector $u \wedge v$ when it is translated parallelly along the segment w:

---

[1] Hamilton discovered the quaternions in October 16th 1843 and defined them as quotients of two vectors. From this definition he deduced the properties of the product of quaternions. I recommend you the reading of the initial chapters of the *Elements of Quaternions* because of its pedagogic importance.

[2] This confusion is due to the fact that vectors and bivectors are dual spaces of the algebra $Cl_{3,\,0}$. However, this duality does not exist at higher dimensions, although there is also duality among other spaces.

$$u \wedge v \wedge w = \begin{vmatrix} u_x & v_x & w_x \\ u_y & v_y & w_y \\ u_z & v_z & w_z \end{vmatrix} e_{123}$$

Finally, let us see how is the product of three vectors u, v and w. The vector v can be resolved into a component coplanar with u and v and another component perpendicular to the plane u-v:

$$u\, v\, w = u\, v_{\|}\, w + u\, v_{\perp}\, w$$

Now let us analyse the permutative property. In the plane we found $u\, v\, w - w\, v\, u = 0$. In the space the permutative property becomes[3]:

$$u\, v\, w - w\, v\, u = u\, v_{\perp}\, w - w\, v_{\perp}\, u = v_{\perp}\left(-u\, w + w\, u\right)$$

$$= -2\, v_{\perp}\, u \wedge w = -2\, v \wedge u \wedge w = 2\, u \wedge v \wedge w$$

I take the same algebraic hierarchies as in chapter 1: all the abridged products must be operated before the geometric product, convention which is coherent with the fact that in many algebraic situations, the abridged products must be developed in sums of geometric products.

Spherical trigonometry

In this section the relations for the sides and angles of the spherical triangles are deduced. I will take for convenience the sphere having unity radius, although the trigonometric identities are equally valid for a sphere of any radius.

Let us consider any three points A, B and C on the sphere with unity radius centred at the origin (figure 14.1). Then $|A| = |B| = |C| = 1$. The angles formed by each pair of sides will be denoted by α, β and γ, and the sides respectively opposite to these angles will be symbolised by a, b and c respectively. Then a is the arc of the great circle passing through the points B and C, that is, the angle between these vectors:

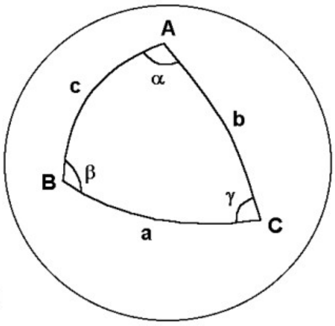

Figure 14.1

$$\sin a = |\, B \wedge C \,|$$

---

[3] From this result it follows that $u \wedge v \wedge w = (u\, v\, w - u\, w\, v + v\, w\, u - v\, u\, w + w\, u\, v - w\, v\, u) / 6$. In other words, the outer product is a fully antisymmetric product. However although being beautiful, it is not so useful for geometric algebra as the permutative property $u \wedge v \wedge w = (u\, v\, w - w\, v\, u)/2$.

Note that in this equality the sine is positive and therefore $a \leq \pi$. So, the spherical trigonometry is deduced for *strict triangles*, those having a, b, $c \leq \pi$.

Also, α is the angle between the sides b and c of the triangle, that is, the angle between the planes passing through the origin, A and B, and the origin, A and C. Since the direction of a plane is given by its bivector, which can be obtained through the outer product, we have:

$$\sin\alpha = \frac{\left|(A\wedge B)\times(A\wedge C)\right|}{|A\wedge B|\,|A\wedge C|}$$

Now we write the products of the numerator using the geometric product:

$$\sin\alpha = \frac{\left|-(A\,B - B\,A)(A\,C - C\,A) + (A\,C - C\,A)(A\,B - B\,A)\right|}{8\,|A\wedge B|\,|A\wedge C|}$$

$$\sin\alpha = \frac{\left|-A\,B\,A\,C + A\,B\,C\,A + B\,A^2\,C - B\,A\,C\,A + A\,C\,A\,B - A\,C\,B\,A - C\,A^2\,B + C\,A\,B\,A\right|}{8\,|A\wedge B|\,|A\wedge C|}$$

We extract the vector A as common factor at the left, but without writing it because $|A| = 1$:

$$\sin\alpha = \frac{\left|-B\,A\,C + B\,C\,A + A\,B\,C - A^{-1}\,B\,A\,C\,A + C\,A\,B - C\,B\,A - A\,C\,B + A^{-1}\,C\,A\,B\,A\right|}{8\,|A\wedge B|\,|A\wedge C|}$$

Applying the permutative property to the suitable pairs of products, we have:

$$\sin\alpha = \frac{\left|6\,A\wedge B\wedge C + 2\,A^{-1}\,A\wedge B\wedge C\,A\right|}{8\,|A\wedge B|\,|A\wedge C|} = \frac{|A\wedge B\wedge C|}{|A\wedge B|\,|A\wedge C|}$$

since the volume $A \wedge B \wedge C$ is a pseudoscalar, which commutes with all the elements of the algebra. Now the *law of sines for spherical triangles* follows:

$$\frac{\sin\alpha}{\sin a} = \frac{\sin\beta}{\sin b} = \frac{\sin\gamma}{\sin c} = \frac{|A\wedge B\wedge C|}{|A\wedge B|\,|B\wedge C|\,|C\wedge A|} \qquad \text{(I)}$$

Let us see the law of cosines. Since $e_{23}^2 = e_{31}^2 = e_{12}^2 = -1$ then:

$$\cos\alpha = \frac{(A\wedge B)\bullet(A\wedge C)}{|A\wedge B|\,|A\wedge C|} = \frac{(A\wedge B)\bullet(A\wedge C)}{\sin c\,\sin b}$$

$$\sin b\,\sin c\,\cos\alpha = -\frac{1}{8}\left[(A\,B - B\,A)(A\,C - C\,A) + (A\,C - C\,A)(A\,B - B\,A)\right]$$

$$= -\frac{1}{8}\left(A\,B\,A\,C - A\,B\,C\,A - B\,A^2\,C + B\,A\,C\,A + A\,C\,A\,B - A\,C\,B\,A - C\,A^2\,B + C\,A\,B\,A\right)$$

Now taking into account that:

$$4\,A\cdot B\;A\cdot C = (A\,B + B\,A)(A\,C + C\,A) = A\,B\,A\,C + A\,B\,C\,A + B\,A^2\,C + B\,A\,C\,A$$

But also:

$$4\,A\cdot B\;A\cdot C = (A\,C + C\,A)(A\,B + B\,A) = A\,C\,A\,B + A\,C\,B\,A + C\,A^2\,B + C\,A\,B\,A$$

and adding the needed terms, we find:

$$\sin b\sin c\cos\alpha = -\frac{1}{8}\left(8\,A\cdot B\;A\cdot C - 2\,A\,B\,C\,A - 2\,B\,A^2\,C - 2\,A\,C\,B\,A - 2\,C\,A^2\,B\right)$$

Extracting common factors and using $A^2 = B^2 = 1$, we may write:

$$\sin b\sin c\cos\alpha = -\frac{1}{8}\left(8\,A\cdot B\;A\cdot C - 2\,A\,(B\,C + C\,B)\,A - 2\,(B\,C + C\,B)\right) = -A\cdot B\;A\cdot C + B\cdot C$$

$$\sin b\sin c\cos\alpha = -\cos c\cos b + \cos a$$

$$\cos a = \cos b\cos c + \sin b\sin c\cos\alpha \qquad \text{(II)}$$

which is the *law of cosines for sides*. The substitution of cos c by means of the law of cosines gives:

$$\cos a = \cos b\,(\cos a\cos b + \sin a\sin b\cos\gamma) + \sin b\sin c\cos\alpha$$

$$\cos a\,(1 - \cos^2 b) = \cos b\sin a\sin b\cos\gamma + \sin b\sin c\cos\alpha$$

and the simplification of sin b:

$$\cos a\sin b = \cos b\sin a\cos\gamma + \sin c\cos\alpha$$

The substitution of sin c = sin a sin γ / sin α yields:

$$\cos a\sin b = \cos b\sin a\cos\gamma + \frac{\sin a\sin\gamma\cos\alpha}{\sin\alpha}$$

Dividing by sin a :

$$\cot a\sin b = \cos b\cos\gamma + \sin\gamma\cot\alpha \qquad \text{(III)}$$

The dual spherical triangle

The perpendicular to the plane containing each side of a spherical triangle cuts the spherical surface in a point. The three points A', B' and C' obtained in this geometric way form the *dual triangle*. The algebraic way to calculate them is the duality operation, which maps bivectors into the perpendicular vectors. The dual of any element is obtained as the product by the pseudoscalar unity $-e_{123}$, which commutes with all the elements of the algebra:

$$A' = -e_{123}\,\frac{B \wedge C}{|B \wedge C|} \qquad B' = -e_{123}\,\frac{C \wedge A}{|C \wedge A|} \qquad C' = -e_{123}\,\frac{A \wedge B}{|A \wedge B|}$$

The inner product of two vectors yields:

$$B' \cdot C' = \frac{(C \wedge A) \bullet (A \wedge B)}{|C \wedge A|\,|A \wedge B|} = -\frac{(A \wedge B) \bullet (A \wedge C)}{|A \wedge B|\,|A \wedge C|} \quad \Leftrightarrow \quad \cos a' = -\cos\alpha$$

which shows that the angle a' and α are supplementary, and so also the other angles:

$$a' = \pi - \alpha \qquad b' = \pi - \beta \qquad c' = \pi - \gamma$$

It is trivial that the dual of the dual triangle is the first triangle, and hence:

$$\alpha' = \pi - a \qquad \beta' = \pi - b \qquad \gamma' = \pi - c$$

We may apply the laws of sines and cosines to the dual triangle. The law of sines is self-dual and may be written in a more symmetric form:

$$\frac{\sin\alpha}{\sin a} = \frac{\sin\beta}{\sin b} = \frac{\sin\gamma}{\sin c} = \frac{|A \wedge B \wedge C|}{|A' \wedge B' \wedge C'|} \qquad \text{(I)}$$

On the other hand, the law of cosines yields a new result when applying duality:

$$-\cos\alpha' = \cos\gamma' \cos\beta' - \sin\beta' \sin\gamma' \cos a'$$

Removing the marks, since this law must be valid for any spherical triangle, we find the *law of cosines for angles*:

$$\cos\alpha = -\cos\gamma \cos\beta + \sin\beta \sin\gamma \cos a \qquad \text{(IV)}$$

Also, applying the equality (III) to the dual triangle we find:

$$\cot\alpha \sin\beta = -\cos\beta \cos c + \sin c \cot a \qquad \text{(V)}$$

The five Bessel's equalities (I to V) allow to solve every spherical triangle knowing any three of its six elements.

Right spherical triangles and Napier’s rule

For the case of a right angle spherical triangle, the five Bessel’s formulas are reduced to a simpler form, and then they may be remembered with the help of the Napier’s pentagon rule (figure 14.2). Draw the angles and the sides of the triangle following the order of the perimeter removing the right angle and writing instead of the legs (the sides adjacent to the right angle) the complementary arcs. Then follow the Pentagon rules:

1) The cosine of every element is equal to the product of the cotangents of the adjacent elements.
2) The cosine of every element is equal to the product of the sines of the nonadjacent elements.

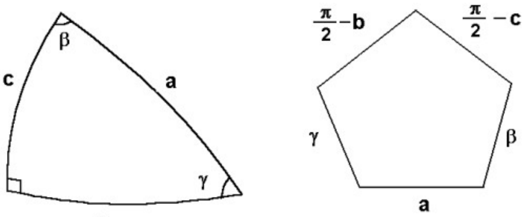

Figure 14.2

For example:

$$\cos a = \sin\left(\frac{\pi}{2} - b\right)\sin\left(\frac{\pi}{2} - c\right) = \cos b \cos c$$

also: $$\cos\left(\frac{\pi}{2} - b\right) = \cot\gamma \cot\left(\frac{\pi}{2} - c\right) \qquad \Rightarrow \qquad \sin b = \cot\gamma \operatorname{tg} c$$

This rule is applied to the right side triangles in the same way: remove the right side and write the complementary of the adjacent angles.

Area of a spherical triangle

A *lune* is a two-sided polygon on the sphere defined by two great circles. The area of a lune is proportional to the angle α between both great circles. For an angle π/2 its area is π, therefore the area of a lune with angle α is 2α. Now let us consider in the sphere shown by figure 14.3 the three lunes having the angles α, β and γ. Then:

$$\begin{cases} s + t = 2\,\alpha \\ s + u' = 2\,\beta \\ s + v = 2\,\gamma \end{cases}$$

where u' is the area of the antipodal triangle of u. Both triangles u and u' have the same angles and area and the system can be rewritten:

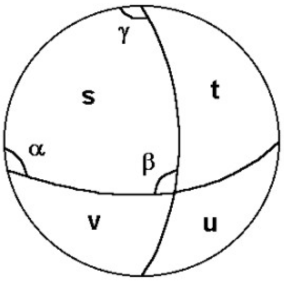

Figure 14.3

$$\begin{cases} s + t = 2\,\alpha \\ s + u = 2\,\beta \\ s + v = 2\,\gamma \end{cases}$$

Adding the three equations, we find:

$$3\,s + t + u + v = 2\left(\alpha + \beta + \gamma\right)$$

The four triangles *s*, *t*, *u* and *v* fill an hemisphere:

$$s + t + u + v = 2\,\pi$$

so the area of the triangle *s* is the *spherical excess*, that is, the addition of the three angles minus π:

$$s = \alpha + \beta + \gamma - \pi$$

Properties of the projections of the spherical surface

No chart of the spherical surface preserving the scale of distances everywhere exists, that is, we cannot depict any map with distances proportional to those measured on the sphere. The *distortion* is the variation of the scale and the angles. Usually there is a line of zero distortion, where the scale is constant.

Since the scale of distances is never preserved for all the points, projections with other interesting properties have been searched in cartography. A projection is said to be *conformal* if it preserves in the map the angles between great circles on the sphere. A projection is said to be *equivalent* if it preserves the area, that is if the figures on the sphere are projected into figures on the map having the same area. A projection is said to be *equidistant* if the scale of distances is preserved, not everywhere but on the line perpendicular to the line of zero distortion, or radially outwards from a point of zero distortion.

The more general concept of projection is any one-to-one mapping of any point (*x*, *y*, *z*) on the sphere into a point (*u*, *v*) on the plane, although the main types of projections are azimuthal, cylindrical and conic. An *azimuthal* projection is a standard projection into a plane, which may be considered touching the sphere. In the tangency point there is zero distortion and the bearings or azimuths from this point are correctly shown. A *cylindrical* projection is a projection into a cylindrical surface around the sphere that will be unrolled. A *conic* projection is a projection into a conical surface tangent to the sphere that will also be unrolled.

Now, I review the main and more used projections beginning with the azimuthal projections.

The central or gnomonic projection

Let us consider the sphere with unity radius (figure 14.4) centred at the origin. Any point on the sphere has the coordinates (*x*, *y*, *z*) fulfilling:

$$x^2+y^2+z^2=1$$

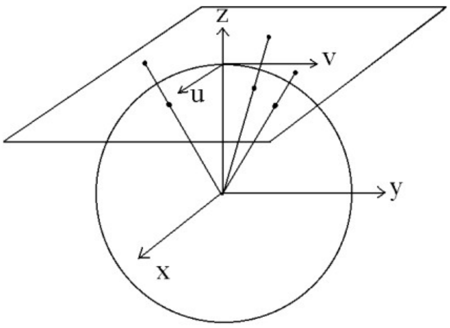

Figure 14.4

Every point on the upper hemisphere is projected into another point on the plane $z = 1$ using as centre of projection the centre of the sphere. Let u and v be the Cartesian coordinates on the projection plane. Taking similar triangles the following relations are found:

$$u=\frac{x}{z} \qquad v=\frac{y}{z}$$

from where we obtain:

$$x=\frac{u}{\sqrt{u^2+v^2+1}} \qquad y=\frac{v}{\sqrt{u^2+v^2+1}} \qquad z=\frac{1}{\sqrt{u^2+v^2+1}}$$

The differential of the arc length is obtained through the differentiation of the above relationships:

$$ds=dx\,e_1+dy\,e_2+dz\,e_3$$

$$ds^2=dx^2+dy^2+dz^2=\frac{\left(1+v^2\right)du^2-2\,u\,v\,du\,dv+\left(1+u^2\right)dv^2}{\left(1+u^2+v^2\right)^2}$$

The geodesics of the sphere are the great circles, which are the intersections with planes passing through its centre. These planes cut the projection plane in straight lines, which are the projections of the geodesics. In other words, any great circle is projected into a line on the projection plane. Taking as equation of the line:

$$v=k\,u+l$$

with k and l constant, the substitution in ds gives:

$$ds=\frac{\sqrt{1+k^2+l^2}}{u^2\left(1+k^2\right)+2\,k\,l\,u+1+l^2}\,du$$

By integration we arrive to the following primitive:

$$s=\operatorname{arctg}\frac{u\left(1+k^2\right)+k\,l}{\sqrt{1+k^2+l^2}}+\text{const}$$

The arc length between two points on this great circle is the difference of this primitive between both points. However it is more advantageous to write it using cosines instead of tangents by means of the trigonometric identity:

$$\cos(s_B - s_A) \equiv \frac{1 + \operatorname{tg} s_A \operatorname{tg} s_B}{\sqrt{1 + \operatorname{tg}^2 s_A}\sqrt{1 + \operatorname{tg}^2 s_B}}$$

After removing *k* and *l* using the equation of the line, we arrive to:

$$\cos(s_B - s_A) = \frac{1 + u_A u_B + v_A v_B}{\sqrt{1 + u_A{}^2 + v_A{}^2}\sqrt{1 + u_B{}^2 + v_B{}^2}}$$

$$= \frac{(u_A e_1 + v_A e_2 + e_3)\cdot(u_B e_1 + v_B e_2 + e_3)}{|u_A e_1 + v_A e_2 + e_3| \; |u_B e_1 + v_B e_2 + e_3|}$$

a trivial result because the arc length is the angle between the position vectors of both points, and also of the proportional vectors going to the projection plane. However, the interest of this result is its analogy with the result found for the hyperboloidal surface.

From this value of the cosine, we may obtain the sine of the arc:

$$\sin(s_B - s_A) = \sqrt{1 - \cos^2(s_B - s_A)} = \frac{\sqrt{(u_A v_B - u_B v_A)^2 + (u_A - v_A)^2 + (u_B - v_B)^2}}{\sqrt{1 + u_A{}^2 + v_A{}^2}\sqrt{1 + u_B{}^2 + v_B{}^2}}$$

$$= \frac{|(u_A e_1 + v_A e_2 + e_3)\wedge(u_B e_1 + v_B e_2 + e_3)|}{|u_A e_1 + v_A e_2 + e_3| \; |u_B e_1 + v_B e_2 + e_3|}$$

which is also a trivial result, since the sine of the angle is proportional to the modulus of the outer product.

Let us see the area function. The differential of the area is easily obtained taking into account that it is a bivector and using the outer product of the differentials of the coordinates:

$$dA = \sqrt{(dx \wedge dy)^2 + (dy \wedge dz)^2 + (dz \wedge dx)^2} = \frac{du \wedge dv}{\left(1 + u^2 + v^2\right)^{3/2}}$$

This result shows that the central projection is not *equivalent* and the distortion increases with the distance to the origin.

Let us consider a plane passing through the centre of the sphere, which cuts its surface in the great circle determined by the equation system:

$$\begin{cases} x^2 + y^2 + z^2 = 1 \\ a\,x + b\,y + z = 0 \quad a, b \quad \text{real} \end{cases}$$

Then the angle between two great circles is the angle between both central planes:

$$\cos\alpha = \frac{(a\, e_{23} + b\, e_{31} + e_{12}) \bullet (a'\, e_{23} + b'\, e_{31} + e_{12})}{|\, a\, e_{23} + b\, e_{31} + e_{12} \,|\; |\, a'\, e_{23} + b'\, e_{31} + e_{12} \,|} = \frac{a\, a' + b\, b' + 1}{\sqrt{a^2 + b^2 + 1}\,\sqrt{a'^2 + b'^2 + 1}}$$

The sides of a spherical triangle are great circles; therefore the central projection of a spherical triangle is a triangle whose sides are straight lines.

Stereographic projection

In the stereographic projection the point of view is placed on the spherical surface. As before, every point (x, y, z) on the sphere with unity radius centred at the origin fulfils the equation:

$$x^2 + y^2 + z^2 = 1$$

We project the spherical surface into the plane $z = 0$ locating the centre of projection at the pole $(0, 0, -1)$ (figure 14.5). The upper hemisphere is projected inside the circle of unity radius while the lowest hemisphere is projected outside. If u and v are the Cartesian coordinates on the projection plane, we have by similar triangles:

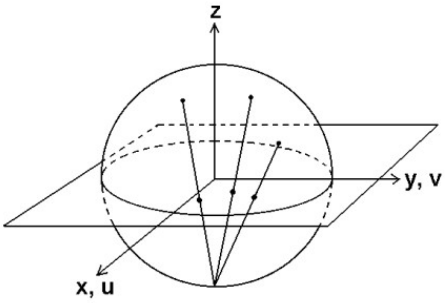

Figure 14.5

$$\frac{x}{u} = z + 1 \qquad \frac{y}{v} = z + 1$$

Using the equation of the sphere one arrives to:

$$x = \frac{2\,u}{1 + u^2 + v^2} \qquad y = \frac{2\,v}{1 + u^2 + v^2} \qquad z = \frac{2}{1 + u^2 + v^2} - 1$$

from where the differential of the arc length is obtained:

$$ds^2 = dx^2 + dy^2 + dz^2 = \frac{4\left(du^2 + dv^2\right)}{\left(1 + u^2 + v^2\right)^2}$$

Now we see that this projection is not equidistant and the distortion increases with the distance to the origin. The factor 4 indicates that the lengths at the origin of coordinates on the plane are the half of the lengths on the sphere. Taking instead of the plane $z = 0$, the plane $z = 1$ this factor becomes 1. Then, we can state correctly the scale of the chart (for example, a polar chart).

The geodesic lines (great circles) are intersection of the sphere surface with planes passing through the origin, which have the equation:

$$z = a\, x + b\, y$$

The substitution by the stereographic coordinates yields:

$$(u+a)^2+(v+b)^2=a^2+b^2+1$$

which is the equation of a circle centred at $(-a, -b)$ with radius $r=\sqrt{a^2+b^2+1}$. Observe that this radius is the hypotenuse of a right angle triangle having as legs the distance to the origin and the unity. That is, the great circles on the sphere are shown in the stereographic projection as circles that intersect the circle $x^2+y^2=1$ in extremes of diameters (figure 14.6). Usually only the projection of the upper hemisphere is used, so that the great circles are represented as circle arcs inside the circle $x^2+y^2\le 1$. The angle α between two of these circles, as explained in the page 89, is obtained from their radii r and r' and centres O and O' through:

Figure 14.6

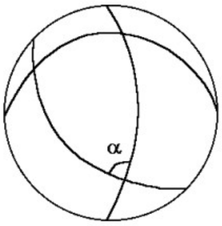

$$\cos\alpha=\frac{r^2+r'^2-(O-O')^2}{2\,r\,r'}=\frac{a\,a'+b\,b'+1}{\sqrt{a^2+b^2+1}\sqrt{a'^2+b'^2+1}}$$

Just this is the angle between the planes $a\,x+b\,y-z=0$ and $a'\,x+b'\,y-z=0$, that is, the angle between the two great circles represented by the projected circles. Therefore, the stereographic projection is a conformal projection of the spherical surface.

If we calculate the differential of area we find:

$$dA=\frac{4\,du\wedge dv}{\left(1+u^2+v^2\right)^2}$$

Now we see that this projection is not equivalent since distortion of areas increases with the distance from the origin (as commented above, the factor 4 becomes 1 projecting into the plane $z=1$).

Orthographic projection

The orthographic projection of the sphere is a parallel projection (the point of view is placed at the infinity). If we make a projection parallel to the z-axis upon the plane $z=0$, the Cartesian coordinates on the map are identical to x, y and we have:

$$z=\sqrt{1-x^2-y^2} \qquad dz=-\frac{x\,dx+y\,dy}{\sqrt{1-x^2-y^2}}$$

$$ds^2=dx^2+dy^2+dz^2=\frac{dx^2+dy^2+2\,x\,y\,dx\,dy}{1-x^2-y^2}$$

$$dA = \sqrt{\frac{1 + x^2 + y^2}{1 - x^2 - y^2}}\, dx \wedge dy$$

Introducing the distance r to the origin of coordinates and the angle φ with respect to the x-axis, we have:

$$x = r\cos\varphi \qquad y = r\sin\varphi$$

$$ds^2 = \frac{dr^2 + r^2 d\varphi^2 + 2\, r\, dr\, d\varphi}{1 - r^2}$$

An example of orthographic projection is the Earth image that appears in many TV news.

Spherical coordinates and cylindrical equidistant (Plate Carrée) projection

For a sphere with unity radius, the spherical coordinates[4] are related with the Cartesian coordinates by means of:

$$x = \sin\theta\cos\varphi$$

$$y = \sin\theta\sin\varphi$$

$$z = \cos\theta$$

Then the differential of arc length is:

$$ds = (\cos\theta\cos\varphi\, d\theta - \sin\theta\sin\varphi\, d\varphi)\, e_1 + (\cos\theta\sin\varphi\, d\theta + \sin\theta\cos\varphi\, d\varphi)\, e_2 - \sin\theta\, d\theta\, e_3$$

Introducing the unitary vectors e_θ and e_φ as:

$$e_\theta = \cos\theta\cos\varphi\, e_1 + \cos\theta\sin\varphi\, e_2 - \sin\theta\, e_3$$

$$e_\varphi = -\sin\varphi\, e_1 + \cos\varphi\, e_2$$

the differential of arc length in spherical coordinates becomes:

$$ds = d\theta\, e_\theta + \sin\theta\, d\varphi\, e_\varphi$$

$$ds^2 = d\theta^2 + \sin^2\theta\, d\varphi^2$$

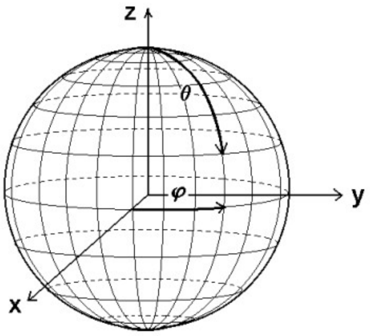

Figure 14.7

---

[4] For geographical coordinates, θ is the colatitude and φ the longitude.

Note that e_θ and e_φ are orthogonal vectors, since their inner product is zero. At $\theta = 0$, *ds* depends only on $d\theta$ and there is a pole. Then θ is the arc length from the pole to the given point (figure 14.7), while φ is the arc length over the equator ($\theta = \pi/2$, $\sin\theta = 1$). The meridians ($\varphi =$ constant) are geodesics, but the parallels ($\theta =$ constant) are not. Exceptionally, the equator is also a geodesic.

In the Plate Carrée projection $u = \varphi$ and $v = \pi/2 - \theta$ (latitude) so that the meridians and parallels are shown in a squared graticule. This projection is equidistant for any meridian whereas the distortion in the parallels increases, as well as for every cylindrical projection, as we separate from the equator. Let us see the area:

$$dA = \sin\theta \; d\theta \wedge d\varphi = \cos v \; du \wedge dv$$

The ratio of the real area with respect to the represented area on the chart is equal to the cosine of the latitude and therefore the projection is not equivalent.

Mercator's projection

The Mercator's projection[5] is defined as the cylindrical conformal projection. If we wish to preserve the angles between curves, we must enlarge the meridians by the same amount as the parallels are enlarged in a cylindrical projection, that is, by the factor 1/sin θ (the secant of the geographical latitude):

$$dv = \frac{-d\theta}{\sin\theta} \qquad \Rightarrow \qquad v = -\log \operatorname{tg}\frac{\theta}{2}$$

The differential of arc length is:

$$ds^2 = d\theta^2 + \sin^2\theta \, d\varphi^2 = \sin^2\theta \left(du^2 + dv^2\right) = \frac{4\exp(2v)}{\left(\exp(2v)+1\right)^2}\left(du^2 + dv^2\right)$$

$$ds = \frac{2\exp(v)}{\exp(2v)+1}\sqrt{du^2 + dv^2}$$

where the distances are increased in an amount independent of the direction and proportional to the inverse of the sine of θ. The differential of area is:

$$dA = \sin\theta \; d\theta \wedge d\varphi = \frac{4\exp(2v)}{\left(\exp(2v)+1\right)^2} du \wedge dv$$

---

[5] The difference between the U.T.M. (Universal Transverse Mercator) projection and the Mercator planisphere is not geometric but geographic: in the U.T.M. the cylinder of projection is tangent to a meridian instead of the equator. All the Earth has been divided in zones of 6º of longitude, where the cylinder of projection is tangent to the central meridian (3º, 9º, 15º ...).

Peters' projection

The Peters' projection is the cylindrical equivalent projection. If we wish to preserve area, we must shorten the meridians in the same amount as the parallels are enlarged in the cylindrical projection, what yields:

$$dv = -\sin\theta\, d\theta \qquad \Rightarrow \qquad v = \cos\theta$$

$$ds^2 = d\theta^2 + \sin^2\theta\, d\varphi^2 = \left(1 - v^2\right) du^2 + \frac{dv^2}{1 - v^2}$$

$$dA = \sin\theta\, d\theta \wedge d\varphi = du \wedge dv$$

which displays clearly the equivalence of the projection. Observe that $v = \cos\theta$ means the sphere is projected following planes perpendicular to the cylinder of projection (figure 14.8).

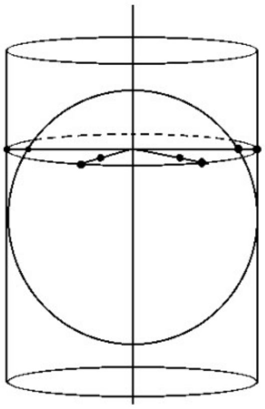

Figure 14.8

Conic projections

These projections are made into a cone surface tangent to the sphere (figure 14.9). Because the cone surface unrolled is a plane circular sector, they are often used to display middle latitudes, while the azimuthal projections are mainly used for poles. In a conic projection a small circle (a parallel) is shown as a circle with zero distortion. The characteristic parameter of a conic projection is the constant of the cone $n = \cos\theta_0$, being θ_0 the angle of inclination of the generatrix of the cone and also the angle from the axis of the cone to any point of tangency with the sphere. Since the graticule of the conic projections is radial, to use the radius r and the angle χ is more convenient:

$$dr = f(\theta)\, d\theta \qquad d\chi = n\, d\varphi$$

The differentials of arc length and area for a conic projection are:

$$ds^2 = d\theta^2 + \sin^2\theta\, d\varphi^2 = \frac{dr^2}{[f(\theta)]^2} + \frac{\sin^2\theta}{n^2}\, d\chi^2$$

$$dA = \sin\theta\, d\theta \wedge d\varphi = \frac{\sin\theta}{n\, f(\theta)}\, dr \wedge d\chi$$

Let us see as before the three special cases: equidistant, conformal and equivalent projections. The differential of area for polar coordinates r, χ is $dA = r\, dr \wedge d\chi$. If the projection is equivalent, we must identify both dA to find :

$$\frac{d}{d\theta}\left[\frac{\sin\theta}{n\,f(\theta)}\right]=f(\theta) \qquad \Rightarrow \qquad f(\theta)=\frac{1}{\sqrt{n}}\cos\frac{\theta}{2} \quad \text{and} \quad r=\frac{2}{\sqrt{n}}\sin\frac{\theta}{2}$$

$$ds^2=\frac{n}{1-\frac{n\,r^2}{4}}dr^2+\frac{1-\frac{r^2\,n}{4}}{n}r^2d\chi^2$$

Figure 14.9

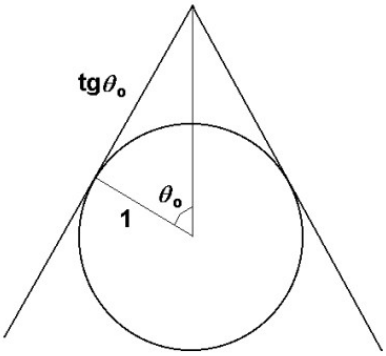

If the projection is equidistant, the meridians have zero distortion so $d\theta=dr/n$ and:

$$ds^2=\frac{1}{n^2}\left(dr^2+\sin^2\left(\frac{r}{n}\right)d\chi^2\right)$$

If the projection is conformal then $ds^2\propto dr^2+r^2d\chi^2$ so:

$$ds^2=\frac{\sin^2\theta}{n^2r^2}\left(dr^2+r^2\,d\chi^2\right)$$

Solving the differential equation:

$$d\theta=\frac{\sin\theta}{n\,r}dr$$

with the boundary condition $\operatorname{tg}\theta_0=r_0$ as shown by the figure 14.9 we find the Lambert's conformal projection:

$$r=\operatorname{tg}\theta_0\left(\frac{\operatorname{tg}\frac{\theta}{2}}{\operatorname{tg}\frac{\theta_0}{2}}\right)^n$$

Exercises

14.1 We have seen the gnomonic projection of great circles being always straight lines. Then, why does the shadow of a gnomon of a sundial follow a hyperbola on a plane surface instead of a line during a day? Why does this hyperbola become a straight line the March 21 and September 23?

14.2 Three points on the sphere are projected by means of the stereographic projection into a circle of unity radius with the coordinates A = (–0.5, 0.5), B = (0, –2/3) and C = (2/3, 0). Calculate the sides, angles and area of the triangle which they form.

14.3 Built at Belfast, the Titanic begun its first and last travel in Southampton the April 10th 1912. After visiting Cherbourg the Titanic weighed anchor in the Cork harbour with bearing New York the 11th April. Calculate the length of the shortest trajectory from Fastnet (Ireland) at 51º30'N 9º35'W to Sandy Hook (New York) at 40º30'N 74ºW. From January 15th to July 15th the ships had to follow the orthodrome (great circle) from Fastnet to the point 42º30'N 47ºW and from this point to Sandy Hook passing at twenty miles from the floating lighthouse of Nantucket. Calculate the length of the route the Titanic should have followed. Take as an averaged radius of the Earth 6366 km. The Titanic sank the 15th April at the position 41º46'N 50º14'W. Is this point on the obliged track?

Figure 14.10

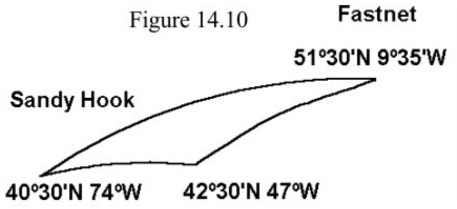

14.4 The figure 14.11 is a photograph of the Hale-Bopp comet. The orientation of the camera is unknown, but the W of Cassiopeia constellation appears in the photograph. The declination *D* of a star is the angular distance from the celestial equator to the star (measured on the great circle passing through the celestial pole and the star). The right ascension *A* is the arc of celestial equator measured eastward from the vernal equinox (one of the intersections of the celestial equator with the ecliptic, also called Aries point) to the foot of the great circle passing through the star and the pole. The right ascension and declination of the stars are constant while those of the comets, planets, the Sun and the Moon are variable. The data of Cassiopeia constellation are:

star	Magnitude	*A*	*D*
α Cassiopeia (Schedir)	2.2	0h 40min 6.4s = 10.0267º	56º 29' 57" N
β Cassiopeia (Caph)	2.3	0h 8min 48.1s = 2.2004º	59º 6' 40" N
γ Cassiopeia (Tsih)	2.4	0h 56min 16.9s = 14.0704º	60º 40' 44" N
δ Cassiopeia (Ruchbah)	2.7	1h 25min 21.2s = 21.3383º	60º 11' 57" N

In your system of coordinates, take the *x*-axis as the Aries point and the *z*-axis as the north pole. Then $D = \pi/2-\theta$ and $A=\varphi$.

a) Calculate the focal distance of the photograph, that is, the distance from the point of view to the plane of the photograph. Knowing that the negative was universal (24×36 mm), calculate the focal distance of the camera.
b) Calculate the orientation (right ascension and declination) of the photographic camera.
c) Calculate the coordinates of the Hale-Bopp comet.

("

Pag. 718(380) A Central Projection of the Lobatchewskian Plane
Fig. 3 and Fig. 4

Pag. 720(381) - Fig. 5 and Fig. 6.

Books-019-2021/20210920/
Pythagorean-hodograph.../
- Treatise of plane geometry through Geometric Algebra...
Calvet(2001).pdf

Look. Pag. 1

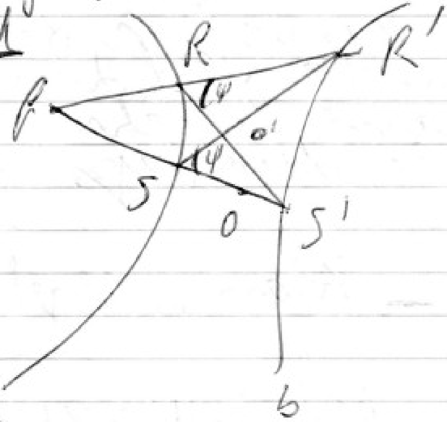

Analizar a curvatura que nos é espelhada a e b, nos observamos o mesmo φ. Interessante, utilizá-lo conjuntamente com Pappus (3 points)

R O' S' → 1ª reta.
S O' R' → 2ª reta.

Notwithstanding = a despeito de

Pag. 48(62) - Fig 5.12

The hypothesis states that the lines AA', BB' and CC' intersect at the same point O.

Obs: Interessante!!!

Utilizei em conjunto com a observação anterior.

Pag. 49(63)

- Desargues theorem

We will prove the Desargues theorem in the other direction following the same algebraic way but applying the duality. (Fig 5.13)

— Alan ... Algebra and Geometry ... (2005).pdf

Pag. 74(88) - 5.1

Spherical distance.

$S=\{x \in \mathbb{R}^3 : ||x||=1\}$

Pág. 89 (103) - 6.1 Quaternions and isometries

Isometries of Euclidean space

$\mathbb{R}^3$ is an isometry if it preserves distances, that is, if for all x and y,

$$\|f(x) - f(y)\| = \|x - y\|.$$

Important:

Quaternions ler em conjunto com a pág. 02 do item anterior - Treatise of plane geometry .... Calvet (2001).pdf

/Hamilton's Elements of Quaternions.

Maxwell himself already applied to the electromagnetic field.

Pág. 307 (320) Chapter 15

/Hyperbolic geometry

- Linear Algebra and Projective Geometry.pdf

Pág. 66 (75) III.3

The Second Fundamental Theorem of Projective Geometry

Pág. 69 (78) Appendix II

jandaia

The Theorem of Pappus

In the Second Fundamental theorem of Projective Geometry, we have met a first instance where the commutativity of the field F (of coordinates) appears important.

Obs: In the present appendix, we give Hilbert's celebrated characterization of commutativity of F by means geometric configurations.

Fig. 4 — ten coplanar points

Pág. 71(80) - III. 4.

The Projective Geometry of a Line in Space; Cross Ratios.

Look: Fig. 5 — Pág. 76(85)

Pág. 85(94) Proposition 4

Suppose that σ is one-to-one correspondence mapping the points on the line L upon the totality of points on the line L^σ; and suppose that the center of F contains at least three elements

then σ preserves all the cross ratios
if, and only if,
the σ-admissible permutations of
the field F are
semi-automorphisms which
leave invariant every class of
conjugate elements in F

Proof: Pag. 86(95)

Look: Pág. 91(100) Fig. 6

Pág. 131(140) Appendix II
Projective Relations between
Lines Induced by Polarities

Pág. 266(275) VII. 3.
The Theorem of Desargues.
(important)

Proposition 1: Assume that the three lines L, L', L'' pass through the point P but are not contained in a plane.

— Books-019-2021/20210920/
Pythagorean-Hodograph curves.../

*   (   )
! +,-.+   +   +/ ! + * , ! ? < ! # , 4 ! ! #
1 2 ! 4 # 9 8 2   " 3#
(* "

5

Spherical geometry

5.1 Spherical distance

This chapter is devoted to *spherical geometry*; that is, to geometry on the surface of the sphere

$$\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^3 : ||\mathbf{x}|| = 1\}.$$

Later in this chapter we shall use spherical trigonometry to derive Euler's formula for polyhedral surfaces, and this will lead (eventually) to a discussion of the symmetry groups of regular polyhedra.

We take for granted the fact that one can measure the length of a smooth curve, and the area of a (reasonably simple) set, on $\mathcal{S}$. A *great circle*, of length 2π, is the intersection of $\mathcal{S}$ and a plane that passes through the origin. Every other circle on $\mathcal{S}$ is a plane section of $\mathcal{S}$ of length less than 2π. If σ is an arc of a great circle, then the length of σ is the angle (in radians) subtended by σ at 0. Given any two points **a** and **b** on $\mathcal{S}$ there is a unique great circle, say C, that contains them (and C lies in the plane through **0**, **a** and **b**). Also, **a** and **b** divide C into two arcs which have different lengths unless $\mathbf{b} = -\mathbf{a}$.

Definition 5.1.1 Let **a** and **b** be two points on $\mathcal{S}$. Then the *spherical distance* $\delta(\mathbf{a}, \mathbf{b})$ between **a** and **b** is the length of the shorter of the two arcs of the (unique) great circle through **a** and **b**. Clearly,

$$\delta(\mathbf{a}, \mathbf{b}) = \cos^{-1}(\mathbf{a} \cdot \mathbf{b}), \tag{5.1.1}$$

where $\cos^{-1}(\mathbf{a} \cdot \mathbf{b})$ is chosen in the range $[0, \pi]$. □

As an application, consider the Earth to be a perfect sphere of radius R whose centre lies at the origin **0** in $\mathbb{R}^3$. We may suppose that the **i**-axis meets the surface of the earth at the point with zero latitude and longitude, and that the positive **k**-axis passes through the north pole. Thus the point on the Earth's surface with latitude α (positive in the northern hemisphere, and negative in the southern hemisphere), and longitude β is given by the vector

$$R\big(\cos\alpha\cos\beta\,\mathbf{i}+\cos\alpha\sin\beta\,\mathbf{j}+\sin\alpha\,\mathbf{k}\big)$$

(the reader should draw a diagram). Suppose, now that $\mathbf{x}_1$ and $\mathbf{x}_2$ are two points on the Earth's surface, and write these as

$$\begin{aligned}\mathbf{x}_1 &= R\big(\cos\alpha_1\cos\beta_1\,\mathbf{i}+\cos\alpha_1\sin\beta_1\,\mathbf{j}+\sin\alpha_1\,\mathbf{k}\big),\\ \mathbf{x}_2 &= R\big(\cos\alpha_2\cos\beta_2\,\mathbf{i}+\cos\alpha_2\sin\beta_2\,\mathbf{j}+\sin\alpha_2\,\mathbf{k}\big).\end{aligned}$$

Then, from (5.1.1),

$$\delta(\mathbf{x}_1,\mathbf{x}_2)=R\cos^{-1}\big[\cos\alpha_1\cos\alpha_2\cos(\beta_1-\beta_2)+\sin\alpha_1\sin\alpha_2\big]. \qquad (5.1.2)$$

This formula gives us the distance (measured on the surface of the Earth) between the two points with latitude α_i and longitude β_i, $i=1,2$.

Exercise 5.1

1. Verify that any point with latitude α is a spherical distance $R(\pi/2-\alpha)$ from the north pole.
2. Assume that the Earth is a sphere of radius 4000 miles. Show that the spherical distance between London (latitude 51° north, longitude 0°) and Sydney (latitude 34° south, longitude 151° east) is approximately 10 500 miles.
3. Suppose that an aircraft flies on the shortest route from London (latitude 51° north, longitude 0°) to Los Angeles (latitude 34° north, longitude 151° east). How close does the aircraft get to the north pole?
4. Let $\mathbf{x}$ and $\mathbf{y}$ be two points on the sphere. Show that the normal to the plane determined by the great circle through $\mathbf{x}$ and $\mathbf{y}$ intersects the sphere at the points $\pm\mathbf{z}$, where $\mathbf{z}=(\mathbf{x}\times\mathbf{y})/||\mathbf{x}\times\mathbf{y}||$. Suppose that $\mathbf{w}$ lies on the same side of the plane as $\mathbf{z}$. Show that $\cos\delta(\mathbf{w},\mathbf{z})=[\mathbf{w},\mathbf{x},\mathbf{y}]$.

5.2 Spherical trigonometry

We begin our discussion of spherical trigonometry with the the spherical version of Pythagoras' theorem. Consider a triangle on $\mathcal{S}$, by which we mean three points $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ of $\mathcal{S}$ which do not lie on a great circle, where the sides of the triangle are the arcs of great circles that join these points in pairs. We assume that the angle in the triangle at $\mathbf{c}$ is $\pi/2$, and we may position this triangle so

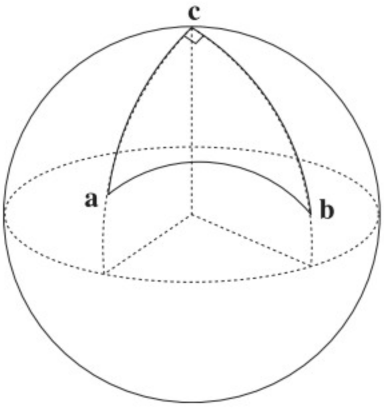

Figure 5.2.1

that $\mathbf{c} = \mathbf{k}$, $\mathbf{a}$ is in the $(\mathbf{i}, \mathbf{k})$-plane, and that $\mathbf{b}$ is in the $(\mathbf{j}, \mathbf{k})$-plane (if necessary we may interchange the labels on **a** and **b**). With this, we have

$$\mathbf{c} = \mathbf{k}, \quad \mathbf{a} = \cos\alpha_1\,\mathbf{i} + \sin\alpha_1\,\mathbf{k}, \quad \mathbf{b} = \cos\alpha_2\,\mathbf{j} + \sin\alpha_2\,\mathbf{k} \qquad (5.2.1)$$

for some α_1 and α_2 (see Figure 5.2.1) so that $(\mathbf{a}\cdot\mathbf{b}) = (\mathbf{a}\cdot\mathbf{c})(\mathbf{b}\cdot\mathbf{c})$. If we now apply (5.1.1) we immediately obtain the following result.

Pythagoras' theorem *Let* **a**, **b** *and* **c** *be the vertices of a spherical triangle on* $\mathcal{S}$*, with the sides making an angle of* $\pi/2$ *at* **c**. *Then*

$$\cos\delta(\mathbf{a}, \mathbf{b}) = \cos\delta(\mathbf{a}, \mathbf{c})\cos\delta(\mathbf{b}, \mathbf{c}). \qquad (5.2.2)$$

Consider (informally) an 'infinitesimal' right-angled triangle. As this triangle is nearly flat we would expect (5.2.2) to look roughly like the usual Euclidean version $a^2 + b^2 = c^2$ of Pythagoras' theorem. This is indeed the case because for small θ, $\cos\theta$ is approximately $1 - \theta^2/2$.

Next, we consider a spherical triangle (as illustrated in Figure 5.2.1) with vertices **a**, **b** and **c**. As is customary in Euclidean geometry, we let the angles at **a**, **b** and **c** be α, β and γ, respectively, and we denote the lengths of the sides by a, b and c; thus $a = \delta(\mathbf{b}, \mathbf{c})$, $b = \delta(\mathbf{c}, \mathbf{a})$ and $c = \delta(\mathbf{a}, \mathbf{b})$. We now prove the following identity; note that the right-hand side here involves *two sides* and *one angle*.

Theorem 5.2.1 *In any spherical triangle with vertices* **a**, **b** *and* **c** *we have*

$$[\mathbf{a}, \mathbf{b}, \mathbf{c}] = \sin a\ \sin b\ \sin\gamma. \qquad (5.2.3)$$

Proof We may choose our axes so that $\mathbf{c} = \mathbf{k}$, and that **a** lies in the $(\mathbf{i}, \mathbf{k})$-plane. As $\delta(\mathbf{a}, \mathbf{c}) = b$, we see that $\mathbf{a} = (\sin b, 0, \cos b)$. Similarly, **b** has latitude

$\pi/2 - a$ and longitude γ; thus

$$\mathbf{b} = (\sin a\ \cos\gamma,\ \sin a\ \sin\gamma,\ \cos a),$$

and (5.2.3) follows directly from these expressions. □

We end this section with the sine rule, and the cosine rule, for spherical geometry.

The sine and cosine rules *In a spherical triangle with vertices* **a**, **b** *and* **c**, *we have the sine rule:*

$$\frac{\sin\alpha}{\sin a} = \frac{\sin\beta}{\sin b} = \frac{\sin\gamma}{\sin c},$$

and the cosine rule: $\cos c = \cos a\ \cos b + \sin a\ \sin b\ \cos\gamma$.

Proof The sine rule follows immediately from (5.2.3) and the fact that $[\mathbf{a}, \mathbf{b}, \mathbf{c}] = [\mathbf{c}, \mathbf{a}, \mathbf{b}] = [\mathbf{b}, \mathbf{c}, \mathbf{a}]$. To prove the cosine rule we may choose **a**, **b** and **c** as in the proof of Theorem 5.2.1. Then

$$\cos c = \cos\ \delta(\mathbf{a}, \mathbf{b}) = \mathbf{a}\cdot\mathbf{b} = \cos a\ \cos b + \sin a\ \sin b\ \cos\gamma,$$

as required. □

Exercise 5.2

1. Derive Pythagoras' Theorem from the cosine rule.
2. Show that if an equilateral spherical triangle has sides of length a and interior angles α, then $\cos(a/2)\ \sin(\alpha/2) = 1/2$. Deduce that $\alpha > \pi/3$ (so that the angle sum of the triangle exceeds π).
3. Calculate the perimeter of a spherical triangle all of whose angles are $\pi/2$.

5.3 Area on the sphere

Let us now find the formula for the area of a spherical triangle on $\mathcal{S}$. We denote the spherical area (that is, the area on $\mathcal{S}$) of a set E by $\mu(E)$, and as the surface area of a sphere of radius r is $4\pi r^2$, we see that $\mu(\mathcal{S}) = 4\pi$. Two (distinct) great circles meet at diametrically opposite points, and they divide the sphere into four regions, called *lunes*. The angle of a lune is the angle (in the lune) at which the circles meet, and it is clear that *the area of a lune of angle* α *is* 2α, for it is obviously proportional to α, and equal to 4π when $\alpha = 2\pi$.

Unlike Euclidean geometry, the area of a spherical triangle is completely determined by its angles (there are no similarity maps in spherical geometry), and the following formula was first found by A. Girard in 1625.

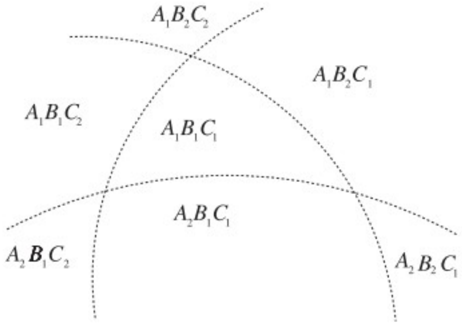

Figure 5.3.1

Theorem 5.3.1 *Let T be a spherical triangle with angles α, β and γ. Then $\mu(T) = \alpha + \beta + \gamma - \pi$.*

Proof The triangle T is formed from sides that lie on three great circles which we denote by $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$. The great circle $\mathcal{A}$ subdivides the sphere into two hemispheres which we denote by A_1 and A_2. We define B_1, B_2, C_1 and C_2 similarly, and these may be chosen so that $T = A_1 \cap B_1 \cap C_1$, with T having angles α, β and γ at its vertices on $\mathcal{B} \cap \mathcal{C}$, $\mathcal{C} \cap \mathcal{A}$ and $\mathcal{A} \cap \mathcal{B}$, respectively. Now $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ divide the sphere into the eight triangles $A_i \cap B_j \cap C_k$, where $i, j, k = 1, 2$, and which, for brevity, we write as $A_i B_j C_k$. The triangle T $(= A_1 B_1 C_1)$) and its six neighbours are illustrated (symbolically) in Figure 5.3.1; the only triangle that does not appear in this illustration is $A_2 B_2 C_2$.

Now (for example) $A_1 B_1 C_1$ and $A_2 B_1 C_1$ together comprise a lune of angle α determined by the great circles $\mathcal{B}$ and $\mathcal{C}$. Thus we find that

$$\begin{aligned}
\mu(A_1 B_1 C_1) + \mu(A_2 B_1 C_1) &= 2\alpha; \\
\mu(A_1 B_1 C_1) + \mu(A_1 B_2 C_1) &= 2\beta; \\
\mu(A_1 B_1 C_1) + \mu(A_1 B_1 C_2) &= 2\gamma; \\
\mu(A_2 B_2 C_2) + \mu(A_1 B_2 C_2) &= 2\alpha; \\
\mu(A_2 B_2 C_2) + \mu(A_2 B_1 C_2) &= 2\beta; \\
\mu(A_2 B_2 C_2) + \mu(A_2 B_2 C_1) &= 2\gamma.
\end{aligned}$$

Adding each side of these six equations, and noting that

$$\mu(A_1 B_1 C_1) = \mu(A_2 B_2 C_2), \qquad \sum_{i,j,k=1}^{2} \mu(A_i B_j C_k) = 4\pi,$$

we see that $\mu(A_1 B_1 C_1) + \pi = \alpha + \beta + \gamma$ as required. □

The formula for the area of a triangle extends readily to the area of a spherical polygon.

Theorem 5.3.2 *Let P be a polygon on the sphere (with each of its n sides being an arc of a great circle), and let the interior angles of the polygon be $\theta_1, \dots, \theta_n$. Then the area $\mu(P)$ of the polygon is given by*

$$\mu(P) = \theta_1 + \cdots + \theta_n - (n-2)\pi. \tag{5.3.1}$$

The proof of (5.3.1) in the case of a *convex polygon* on the sphere is easy. Suppose that a polygon P has the property that there is some point $\mathbf{x}$ in P that can be joined to each vertex $\mathbf{v}_j$ by an arc of a great circle, with these arcs being non-intersecting (except at $\mathbf{x}$) and lying in P. Then these arcs divide P into n triangles, and applying Theorem 5.3.1 to each, and summing the results, we obtain (5.3.1). The proof of (5.3.1) for a non-convex polygon will be given in the next section. □

Exercise 5.3

1. Calculate the area of a spherical triangle all of whose angles are $\pi/2$, and also the area of a spherical triangle all of whose angles are $3\pi/2$.
2. For which values of θ is it possible to construct an equilateral spherical triangle with each angle equal to θ?
3. Prove the famous result of Archimedes that the area of the part of $\mathcal{S}$ that lies between the two parallel planes given, say, by $x_3 = a$ and $x_3 = b$, is the same as the area of the part of the circumscribing cylinder (given by $x_1^2 + x_2^2 = 1$) that lies between these two planes. Hence find the area of the 'polar cap' $\{\mathbf{x} \in \mathcal{S} : \delta(\mathbf{x}, \mathbf{k}) < r\}$.

5.4 Euler's formula

A spherical triangle T is a region of $\mathcal{S}$ bounded by three arcs σ_1, σ_2 and σ_3 of great circles. The arcs σ_j are the *edges* of T, and the three points $\sigma_i \cap \sigma_j$ are the *vertices* of T. A *triangulation* of the sphere $\mathcal{S}$ is a partitioning of $\mathcal{S}$ into a finite number of non-overlapping spherical triangles T_j such that the intersection of any two of the T_j is either empty, or a common edge, or a common vertex, of the two triangles. The *edges* of the triangulation are the edges of all of the T_j; the *vertices* of the triangulation are all of the vertices of the T_j.

The simplest example of a triangulation on the sphere (say, the surface of the Earth) is found by drawing the equator and n lines of longitude. In this case the triangulation contains $2n$ triangles, $n + 2$ vertices (n on the equator, and one at each pole), and $3n$ edges. If we denote the numbers of triangles (which we now call *faces*), edges and vertices by F, E and V, respectively, we find that $F - E + V = 2$; thus the expression $F - E + V$ *does not depend on the choice of* n. It is even more remarkable that the formula $F - E + V = 2$ holds for *all* triangulations of the sphere. This famous result is due to the Swiss mathematician Leonard Euler (1707–83) who was one of the most productive mathematicians of all time.

Euler's theorem *Suppose that a triangulation of* $\mathcal{S}$ *has* F *triangles,* E *edges and* V *vertices. Then* $F - E + V = 2$.

We shall give Legendre's beautiful proof that is based on the area of a spherical triangle.

Proof The area of a spherical triangle Δ with angles θ_1, θ_2 and θ_3 is $\theta_1 + \theta_2 + \theta_3 - \pi$. Suppose that there are F triangles, E edges and V vertices in a triangulation of the sphere. Then, summing over all angles in all triangles, the total angle sum is $2\pi V$ (for all of the angles occur at a vertex without overlap, and the angle sum at any one of these V vertices is exactly 2π). Also, the sum of the areas of the triangles is the area of the sphere; thus $2\pi V - F\pi = 4\pi$, or $2V = F + 4$. Now (by counting the edges of each triangle, and noting that this counts each edge twice), we obtain $3F = 2E$; thus

$$F - E + V = F - 3F/2 + (F + 4)/2 = 2.$$

□

There are various important extensions of Euler's theorem which we shall now discuss briefly and informally. A *spherical polygon* is a region bounded by a finite number of arcs of great circles in such a way that the arcs form a 'closed' curve on $\mathcal{S}$ that divides $\mathcal{S}$ into exactly two regions. A spherical polygon P is *convex* if any two points in P can be joined by an arc (of a great circle) that lies entirely within P. Now partition $\mathcal{S}$ into a finite number of non-overlapping convex spherical polygons (which now need not be spherical triangles), and suppose that there are F polygons (or faces), E edges (each edge counted only once), and V vertices (each vertex counted only once); then, again, $F - E + V = 2$. To prove this consider one face of the triangulation, and suppose that this is bounded by a polygonal closed curve that comprises m edges and m vertices, say $\mathbf{v}_j$. We take any point $\mathbf{x}$ in the face and join this to each $\mathbf{v}_j$ by a segment of a great circle, thus producing a triangulation of this face that includes, among its edges and vertices, all of the original edges and vertices of the face in the original partition. The contribution of this face (without its boundary) to the count $F - E + V$ from the new triangulation is 1, for it contributes m triangles, m edges and one vertex (at $\mathbf{x}$). As its contribution to $F - E + V$ in the original partition of $\mathcal{S}$ is also 1 (it contributes one to F and zero to E and V), it makes no difference to the count of $F - E + V$ whether we subdivide the face into triangles or not. If we carry out this subdivision for all faces, and then use Euler's theorem, we find that $F - E + V = 2$.

Next, suppose that we partition the sphere into convex polygons and that we then deform the partition continuously in such a way that the numbers F, E and V do not change during the deformation; then (obviously) after the deformation we still have $F - E + V = 2$. In particular, this will still be true when the edges are not necessarily arcs of great circles (or even arcs of any circles) providing that the given 'curvilinear partition' can be 'deformed' into a partition of the prescribed type in such a way that the numbers of faces, edges and vertices remains constant during the 'deformation'. Of course, all this is obvious; the difficulty (if one is to be rigorous) lies in the deformations.

Finally, we can extend this idea and now allow deformations of the sphere itself (together with any given partition of it). For example, as we can deform the sphere to a cube, the formula $F - E + V = 2$ will still hold for a cube. Indeed, if we partition the cube in the natural way into its six square faces, we see that $F = 6$, $E = 12$ and $V = 8$, so $F - E + V = 2$. If we wish, we can now divide each square face into two triangles by drawing one diagonal across each face, and then regard these diagonals together with the 'natural' edges of the cube as the edges of a new triangulation. This triangulation has twelve triangles, eighteen edges and eight vertices and again, $12 - 18 + 8 = 2$. Collectively, these ideas are described by saying that Euler's formula is a *topological invariant*, but we shall not discuss this any further.

Roughly speaking, a *polyhedron* is a 'closed' surface (that is, a surface with no bounding edges) made up by 'joining' polygons together along edges (of the same length), with the resulting surface being capable of being deformed into a sphere. We can also define a *convex polyhedron* to be the (non-empty) intersection of a finite number of half-spaces (a *half-space* is one 'side' of a plane); for example, a cube is easily seen to be the intersection of exactly six half-spaces. A convex polyhedron can be easily deformed into a sphere (by choosing a sphere which contains the polyhedron, and whose centre lies inside the polyhedron, and then projecting the polyhedron radially from the centre onto the sphere); thus, if a convex polyhedron has F faces, E edges and V vertices, then $F - E + V = 2$.

We complete this informal discussion by considering Euler's formula for a plane polygon. Suppose that a closed polygonal curve C (lying in the plane) divides the plane into two regions. Exactly one of these regions will be bounded, and we denote this by P. Suppose now that we triangulate P in the manner considered above. It seems clear that we can deform the polygon P until it lies on the 'southern hemisphere' of $\mathcal{S}$, in such a way that the boundary curve of P lies in the 'equator'. Then there will be, say, m vertices and m edges lying on the equator (for each edge ends at a vertex which is the starting point of the next edge). Let us now include the 'northern hemisphere' N, say, as one of the polygons on the sphere, with the m edges and m vertices being the edges and vertices of N regarded as a the polygon. Then we have constructed a partioning of $\mathcal{S}$ and, for this partition, $F - E + V = 2$. As we have simply added one face to the original picture, it follows that for the triangulation of the plane polygon P, we must have $F - E + V = 1$.

We end this section by completing the proof of Theorem 5.3.2.

The proof of Theorem 5.3.2 Let P be any polygon on the sphere. We extend each side of P to the great circle that contains it, and in this way we subdivide the sphere into a finite number of convex polygons. As each of these polygons either lies within P or outside of P, this shows that *we can subdivide any spherical polygon into a finite number of convex polygons*, and hence also *into a finite number of spherical triangles*.

Now take any polygon P on $\mathcal{S}$, and divide it into triangles T_j as described above. Suppose that the triangulation of P into these triangles T_j has F triangles, E edges and V vertices; then by Euler's formula (as described in the previous paragraph), $F - E + V = 1$. In order to describe the next step of the proof, let us denote the original polygonal curve by C, and its 'interior' (which we have just divided into triangles) by P_0. Suppose now that E_0 of the edges of this new triangulation lie on some side of C, so that $E - E_0$ edges lie in P_0. As each of the F triangles T_j has three sides, we see that $3F = E_0 + 2(E - E_0)$; thus $3F + E_0 = 2E$. We now compute areas. We suppose that the original polygon has n vertices with internal angles $\theta_1, \dots, \theta_n$ at these vertices. These vertices lie on C. Now E_0 of the V vertices in the new triangulation lie on C, and of these, n occur as original vertices of P (on C), while the remaining $E_0 - n$ occur as 'new' vertices lying on C (and interior to an original side of P). The angle sum at each of these $E_0 - n$ vertices is π. The remaining $V - E_0$ vertices are in P_0, and the angle sum at each of these vertices is 2π. Thus

$$\begin{aligned}\mu(P) &= \sum_{j=1}^{F} \mu(T_j)\\ &= (\theta_1 + \cdots + \theta_n) + (E_0 - n)\pi + (V - E_0)2\pi - \pi F\\ &= (\theta_1 + \cdots + \theta_n) - n\pi + (2V - F - E_0)\pi\\ &= (\theta_1 + \cdots + \theta_n) - (n-2)\pi,\end{aligned}$$

because $F - E + V = 1$ and $3F + E_0 = 2E$. □

Exercise 5.4

1. Verify Euler's formula for a 'pyramid' that has an n-gon as a base.
2. What is Euler's formula for a plane polygon from which two polygonal 'holes' have been removed?

5.5 Regular polyhedra

A *regular polyhedron* is a polyhdron that is made by joining together a finite number of congruent, regular polygons, each with p sides, and with exactly q polygons meeting at each vertex, and we say that this a regular polyhedron of *type* (p, q). The *faces* of the polyhedron are the regular p-gons, the *edges* are the segments where two faces are joined, and the *vertices* are the points at the ends of the edges. Such solids, which have a high degree of symmetry, are also known as *Platonic solids* (after Plato who associated them with earth, water, fire, air and the cosmos).

It is clear that if we are to construct a regular polyhedron of type (p, q), then we must be able to places q of the polygons together, in the plane and without overlapping, each having one vertex at the origin. Indeed, the sum of their angles at the origin must be strictly less than 2π, so that p and q must satisfy $q(\pi - 2\pi/p) < 2\pi$. Now this inequality is equivalent to

$$(p-2)(q-2) < 4 \tag{5.5.1}$$

and the only solutions (p, q) of this that are consistent with the obvious geometric constraints $p \geq 3$ and $q \geq 3$ are (3, 3), (3, 4), (3, 5), (4, 3) and (5, 3). In particular, there are at most five such polyhedra (up to scaling).

Now consider a regular polyhedron of type (p, q) and suppose that this has F faces, E edges and V. Then, by Euler's formula, $F - E + V = 2$. As each edge is the edge of exactly two faces, and each face has p edges, we see that $2E = pF$ (we simply count each 'side' of each edge in two different ways). Similarly, as each edge has two 'ends', and each vertex is the endpoint of q 'ends', we see that $2E = qV$. We now have three simultaneous equations, namely $F - E + V = 2$, $pF = 2E$ and $qV = 2E$, and if we solve these for F, E and V in terms of p and q, we find that

$$F = \frac{4q}{2p + 2q - pq}, \quad E = \frac{2pq}{2p + 2q - pq}, \quad V = \frac{4p}{2p + 2q - pq}, \tag{5.5.2}$$

Notice, from (5.5.1), that $2p + 2q - pq > 0$. If we now substitute the five distinct possibilities for (p, q) in the formulae in (5.5.2) we obtain the following

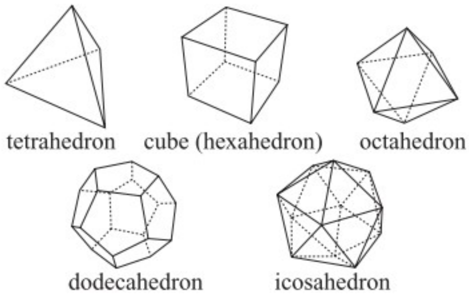

Figure 5.5.1

complete list of regular polyhedra:

polyhedron	*faces*	*edges*	*vertices*
tetrahedron	4	6	4
cube	6	12	8
octahedron	8	12	6
icosahedron	20	30	12
dodecahedron	12	30	20

Of course, this does not show that these solids do actually exist but they do (see Figure 5.5.1), and we shall accept this for the moment.

Finally, we remark that to obtain the Platonic solids it is not enough to merely assume that the faces of the polyhedron are congruent regular polygons. For example, we can attach five equilateral triangles together to form a pyramid with a regular pentagonal base and then join two of these pyramids together across their base; the resulting polyhedron has congruent faces that are equilateral triangles, but it is not a Platonic solid.

Exercise 5.5

1. Establish the formulae (5.5.2). These arise from three simultaneous equations; those who know about matrices should write these equations in matrix form and solve them by finding the inverse matrix. What is the determinant of the matrix?
2. The deficiency of a polyhedron is $2\pi V - \Theta$, where Θ is the sum of all the interior angles of all of its faces. Show that each of the five Platonic solids has deficiency 4π.
3. Show that the mid-points of the faces of a cube form the vertices of a regular octahedron.

4. Let A, B, C and D be the vertices of a regular tetrahedron. Show that the midpoints of the sides AB, BC, CD and DA are coplanar, and form the vertices of a square.

5.6 General polyhedra

In this section we shall examine several results that are closely related to Euler's formula. We have seen that the condition $F - E + V = 2$ is a *necessary* condition for the existence of a polyhedron (that can be deformed into a sphere) with F faces, E edges and V vertices. However, this is not the only necessary condition. Suppose that we 'cut' each edge of the polyhedron at its midpoint and then count the segments that are formed by this cutting. As each edge produces two segments, this count is clearly $2E$. On the other hand, every vertex has at least three segments ending there, and each segment ends at only one vertex; thus the count is at least $3V$. We conclude that $2E \geq 3V$. In a similar way, imagine that we have 'separated' the polyhedron into its faces, and that we now count the edges of all of these faces. Clearly this count is $2E$. On the other hand, each face has at least three sides so the count must be at least $3F$. We deduce that $2E \geq 3F$. Thus

$$F - E + V = 2, \quad 2E \geq 3V, \quad 2E \geq 3F. \tag{5.6.1}$$

are all necessary conditions for the existence of a polyhedron (that can be deformed into a sphere), with F faces, E edges and V vertices. These show that Euler's relation $F - E + V = 2$ is not by itself sufficient for the existence of a polyhedron; for example, there is no polyhedron with $F = 4$, $E = 7$ and $V = 5$ because for these values, $2E < 3V$.

Theorem 5.6.1 *There exists a convex polyhedron with F faces, E edges and V vertices if and only if* (5.6.1) *holds.*

Proof We know that (5.6.1) are necessary conditions for the existence of a polyhedron. To prove their sufficiency; we take positive integers F, E and V satisfying (5.6.1), and we divide the proof into two cases, namely (i) $V \geq F$ and (ii) $V < F$.

(i) Suppose that a convex polyhedron has f faces, e edges and v vertices, and that it has at least one vertex of valency three (that is, exactly three faces meet at the vertex). Then we can truncate the polyhedron at that vertex (that is, we can 'slice' the vertex from the polyhedron) and the resulting polyhedron has $f + 1$ faces, $e + 3$ edges and $v + 2$ vertices. As the resulting polyhedron also has vertices of valency three, we can repeat this operation as often as we wish, say t times, and so obtain a convex polyhedron with $f+t$ faces, $e+3t$ edges and $v+2t$ vertices. If we apply this to the polyhedron that is a pyramid P_m whose base is an m-gon (this has $m+1$ faces, $2m$ edges and $m+1$ vertices), the resulting polyhedron will have $m+1+t$ faces, $2m+3t$ edges and $m+1+2t$ vertices; thus it suffices to show that the equations

$$\begin{aligned} m+1+t &= F, \\ 2m+3t &= E, \\ m+1+2t &= V \end{aligned}$$

have a solution in integers m and t with $m \geq 3$ and $t \geq 0$. These equations are consistent (because $F-E+V=2$), and the solution is $t=V-F$ and $m=2F-V-1$. Finally, $t \geq 0$ by (i), and $m=2F-V-1=(2E-3V)+3 \geq 3$.

(ii) Suppose that a convex polyhedron has f faces, e edges and v vertices, and that one of the faces is triangular. Then we can glue a tetrahedron onto that face, and if the tetrahedron is sufficiently 'flat' the resulting polyhedron will be convex, with a triangular face, and it will have $f+2$ faces, $e+3$ edges and $v+1$ vertices. If we do this t times we obtain a convex polyhedron with $f+2t$ faces, $e+3t$ edges and $v+t$ vertices. If the original polyhedron is the polyhedron P_m used above, the resulting polyhedron will have $m+1+2t$ faces, $2m+3t$ edges and $m+1+t$ vertices, and we now have to solve the equations

$$\begin{aligned} m+1+2t &= F, \\ 2m+3t &= E, \\ m+1+t &= V, \end{aligned}$$

again with $t \geq 0$ and $m \geq 3$. These equations are consistent, with solution $t=F-V$ and $m=2V-F-1$. By (ii), $t>0$, and $m=2V-F-1=(2E-3F)+3 \geq 3$. □

Our next result was known to Euler. Consider a polyhedron with F faces, E edges and V vertices. To examine the geometry in greater detail, let F_n be the number of faces with exactly n edges (so F_3 is the number of triangular faces, and so on), and let V_m be the number of vertices which have exactly m edges ending at that vertex (equivalently, exactly m faces containing that vertex). Clearly,

$$F=F_3+F_4+\cdots, \quad V=V_3+V_4+\cdots. \tag{5.6.2}$$

Theorem 5.6.2

(1) *For any polyhedron,* $3F_3 + 2F_4 + F_5 \geq 12$. *In particular, among all of the faces of a polyhedron, there are at least four faces each of which has at most five edges.*
(2) *For any polyhedron,* $F_3 + V_3 \geq 8$. *In particular, any polyhedron has either a triangular face, or a vertex with exactly three edges meeting at that vertex (or both).*
(3) *The integers* $F_3 + F_5 + F_7 + \cdots$ *and* $V_3 + V_5 + V_7 + \cdots$ *are even.*

Remark There is equality in (1) and (2) in the case of the tetrahedron and the cube, so both inequalities are best possible. Note that (1) implies that one cannot make a polyhedron (however large and complicated) out of a 'random' collection of, say, hexagons. This is not obvious!

Proof As every edge bounds exactly two faces, we have

$$2E = 3F_3 + 4F_4 + 5F_5 + \cdots,$$

and as every edge contains exactly two vertices,

$$2E = 3V_3 + 4V_4 + 5V_5 + \cdots.$$

This proves (3). To prove (2), we note that (summing over $k = 3, 4 \ldots$)

$$\begin{aligned}\sum(4-k)F_k + \sum(4-k)V_k &= 4\sum F_k + 4\sum V_k - \sum kF_k - \sum kV_k \\ &= 4F + 4V - 2E - 2E \\ &= 8.\end{aligned}$$

This shows that

$$F_3 + V_3 = 8 + \sum_{k\geq 5}(k-4)(F_k + V_k) \geq 8, \tag{5.6.3}$$

which is stronger than (2). Finally, Euler's relation in the form

$$6\sum F_k + 6\sum V_k = 12 + \sum kF_k + 2\sum kV_k$$

leads to the identity

$$3F_3 + 2F_4 + F_5 = 12 + \sum_{k\geq 6}(k-6)F_k + \sum_{k\geq 3}(2k-6)V_k \geq 12,$$

and this gives (1). □

We end this chapter with a result, due to Descartes, that is equivalent to Euler's formula. The *deficiency of a vertex* v of a polyhedron P is $2\pi - \sum_j \theta_j$, where the θ_j are the angles at v of the faces that meet at v. The *total deficiency* of P is the sum of the deficiencies of each of its vertices. For example, if P is a cube, then the deficiency of each vertex is $\pi/2$, and so the total deficiency of a cube is 4π. This is no accident.

Descartes' theorem *The total deficiency of any polyhedron is* 4π.

Proof We suppose that a polyhedron has F faces, E edges and V vertices. We have seen that $F = \sum_{m\geq 3} F_m$ and $2E = \sum_{m\geq 3} mF_m$. Thus

$$\begin{aligned} 2(F - E + V) &= 2V - (2E - 2F) \\ &= 2V - \sum_{m\geq 3}(m-2)F_m. \end{aligned}$$

As the sum of the interior angle of an m-gon is $(m-2)\pi$, the total deficiency of the polyhedron is D, where

$$D = 2\pi V - \sum_m (m-2)F_m\pi = 2(F - E + V)\pi.$$

It follows that $D = 4\pi$ *if and only if* $F - E + V = 2$; thus Descartes' theorem is equivalent to Euler's theorem. □

Exercise 5.6

1. Prove (in the notation of the text) that $3V_3 + 2V_4 + V_5 \geq 12$.
2. We have seen that $F_3 + V_3 \geq 8$. Suppose that $F_3 + V_3 = 8$. Use (5.6.3) to show that F_3 and V_3 are even, so that the possible values of (F_3, V_3) are $(0, 8)$, $(2, 6)$, $(4, 4)$, $(6, 2)$ and $(8, 0)$. Show that each of these pairs is attained by some polyhedron.
3. A *deltahedron* is a polyhdedron (that can be deformed into a sphere) whose faces are congruent equilateral triangles. Show that every vertex of a convex deltahedron has valency at most five. Deduce that for any convex deltahedron, (F, E, V) must be of the form $(2k, 3k, k+2)$, where $k = 2, 3, \ldots, 10$. In fact, convex deltahedra exist for all of these k except 9. There exists such a polyhedron with eighteen triangular 'faces', but some of these 'faces' are coplanar.

Pág. 89 (103) - 6.1 Quaternions and
isometries
Isometries of Euclidean space

$\mathbb{R}^3$ is an isometry if it preserves distances, that is, if for all x and y, $\|f(x) - f(y)\| = \|x - y\|$.

Important:

Quaternions ler em conjunto com a pág. 02 do item anterior - Treatise of plane geometry .... Caret (2001).pdf

/Hamilton's Elements of Quaternions

Maxwell himself already applied to the electromagnetic field.

Pág. 307 (320) Chapter 15
/Hyperbolic geometry

- Linear Algebra and Projective Geometry.pdf

Pág. 66 (75) III.3

The Second Fundamental Theorem of Projective Geometry

Pág. 69 (78) Appendix II

Pág. 89 (103) - 6.1 Quaternions and
Isometries of Euclidean space
isometries

$\mathbb{R}^3$ is an isometry if it preserves distances, that is, if for all x and y,
$\|f(x) - f(y)\| = \|x - y\|$.

Important:

Quaternions ler em conjunto com a pág. 02 do item anterior - Treatise of plane geometry.... Cauet (2001).pdf

/Hamilton's Elements of Quaternions.

Maxwell himself already applied to the electromagnetic field.

Pág. 307 (320) Chapter 15
/Hyperbolic geometry

- Linear Algebra and Projective Geometry.pdf

Pág. 66 (75) III.3
The Second Fundamental Theorem of Projective Geometry

Pág. 69 (78) Appendix II

6

Quaternions and isometries

6.1 Isometries of Euclidean space

Our first objective is to understand isometries of $\mathbb{R}^3$.

Definition 6.1.1 A map $f : \mathbb{R}^3 \to \mathbb{R}^3$ is an *isometry* if it preserves distances; that is, if for all $\mathbf{x}$ and $\mathbf{y}$, $||f(\mathbf{x}) - f(\mathbf{y})|| = ||\mathbf{x} - \mathbf{y}||$.

Each reflection across a plane is an isometry, and we shall see later that every isometry is a composition of reflections. Consider the plane Π given by $\mathbf{x}{\cdot}\mathbf{n} = d$, where $||\mathbf{n}|| = 1$, and let R be the reflection across Π. As $\mathbf{n}$ is the normal to Π, we see that $R(\mathbf{x}) = \mathbf{x} + 2t\mathbf{n}$, where t is chosen so that the midpoint, $\mathbf{x} + t\mathbf{n}$, of $\mathbf{x}$ and $R(\mathbf{x})$ lies on Π. This last condition gives $d = \mathbf{x}{\cdot}\mathbf{n} + t$, so that

$$R(\mathbf{x}) = \mathbf{x} + 2(d - \mathbf{x}{\cdot}\mathbf{n})\mathbf{n}. \tag{6.1.1}$$

It is geometrically obvious that (a) $R(\mathbf{x}) = \mathbf{x}$ if and only if $\mathbf{x} \in \Pi$, (b) for all $\mathbf{x}$, $R(R(\mathbf{x})) = \mathbf{x}$, and (c) R is an isometry. These properties can be verified algebraically from (6.1.1), as can the next result.

Theorem 6.1.2 *A reflection across a plane* Π *is a linear map if and only if* $0 \in \Pi$.

Now consider two parallel planes, say $\mathbf{x}{\cdot}\mathbf{n} = d_1$ and $\mathbf{x}{\cdot}\mathbf{n} = d_2$, and let R_1 and R_2 denote the reflections in these planes. It is easy to see that $R_1R_2(\mathbf{x}) = \mathbf{x} + 2(d_1 - d_2)\mathbf{n}$, so that *the composition of two reflections in parallel planes is a translation* (and conversely). Clearly, the translation is by $2d\mathbf{n}$, where $\mathbf{n}$ is the common normal of the two planes, and d is their distance apart.

Next, we consider the composition of reflections R_j in two distinct intersecting planes Π_1 and Π_2. The planes intersect in a line L, and each R_j fixes every point of L. In addition, $R_j(\Pi) = \Pi$ for every plane Π that is orthogonal to L. The action of R_2R_1 on Π is a reflection across the line $\Pi \cap \Pi_1$ followed by a reflection across the line $\Pi \cap \Pi_2$; thus we see that *$R_2 R_1$ is a rotation of $\mathbb{R}^3$ about the axis L of an angle equal to twice the angle between the planes* Π_j. This leads us to the following definition of a rotation.

Definition 6.1.3 A *rotation* of $\mathbb{R}^3$ is the composition of reflections across two distinct non-parallel planes. The line of intersection of the planes is the *axis* of the rotation.

Notice that the fact that every rotation has an axis is part of our definition of a rotation. Later we will give an alternative (but equivalent) definition of a rotation in terms of matrices, but then we will need to *prove* that every rotation has an axis. As a rotation of $\mathbb{R}^2$ has one fixed point and no axis, the proof cannot be entirely trivial (in fact, it depends on the fact that every real cubic polynomial has a real root). Each rotation of an odd-dimensional space has an axis, but a rotation of an even-dimensional space need not have an axis (because, as we shall see later, every real polynomial of odd degree has a real root, whereas a real polynomial of even degree need not have any real roots). As each reflection is an isometry, so is any composition of reflections. The converse is also true, and in the following stronger form.

Theorem 6.1.4 *Every isometry of $\mathbb{R}^3$ is the composition of at most four reflections. In particular, every isometry is a bijection of $\mathbb{R}^3$ onto itself. Every isometry that fixes $\mathbf{0}$ is the composition of at most three reflections in planes that contain $\mathbf{0}$.*

A reflection R across a plane is a permutation of $\mathbb{R}^3$ whose square is the identity map. Thus, in some sense, Theorem 6.1.4 is analogous to the result that a permutation of a finite set can be expressed as a product of transpositions. The common feature of these two results (and other results too) is that a given map f is being expressed as a composition of simpler maps f_j for which $f_j^2 = I$. Our proof of Theorem 6.1.4 is based on the following three simple results.

Lemma 6.1.5 *Suppose that f is an isometry with $f(\mathbf{0}) = \mathbf{0}$. Then for all $\mathbf{x}$ and $\mathbf{y}$, $||f(\mathbf{x})|| = ||\mathbf{x}||$ and $f(\mathbf{x}){\cdot}f(\mathbf{y}) = \mathbf{x}{\cdot}\mathbf{y}$ (that is, f preserves norms and scalar products).*

Proof First, $||f(\mathbf{x})|| = ||f(\mathbf{x}) - f(\mathbf{0})|| = ||\mathbf{x} - \mathbf{0}|| = ||\mathbf{x}||$, so that f preserves norms. Next,

$$\begin{aligned} ||f(\mathbf{x})||^2 + ||f(\mathbf{y})||^2 - 2\big(f(\mathbf{x}){\cdot}f(\mathbf{y})\big) &= ||f(\mathbf{x}) - f(\mathbf{y})||^2 \\ &= ||\mathbf{x} - \mathbf{y}||^2 \\ &= ||\mathbf{x}||^2 + ||\mathbf{y}||^2 - 2(\mathbf{x}{\cdot}\mathbf{y}), \end{aligned}$$

so that f also preserves scalar products. □

Lemma 6.1.6 *If an isometry f fixes* $\mathbf{0}$, $\mathbf{i}$, $\mathbf{j}$ *and* $\mathbf{k}$ *then* $f = I$.

Proof Let $f(\mathbf{x}) = \mathbf{y}$. Then $||\mathbf{y} - \mathbf{i}|| = ||f(\mathbf{x}) - f(\mathbf{i})|| = ||\mathbf{x} - \mathbf{i}||$, and as $||\mathbf{y}|| = ||\mathbf{x}||$, we have $\mathbf{y}{\cdot}\mathbf{i} = \mathbf{x}{\cdot}\mathbf{i}$. The same holds for $\mathbf{j}$ and $\mathbf{k}$, so that $\mathbf{y} - \mathbf{x}$ is orthogonal to $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$. Thus $\mathbf{y} = \mathbf{x}$. □

Lemma 6.1.7 *Suppose that* $||\mathbf{a}|| = ||\mathbf{b}|| \neq 0$. *Then there is a reflection R across a plane Π through* $\mathbf{0}$ *such that* $R(\mathbf{a}) = \mathbf{b}$ *and* $R(\mathbf{b}) = \mathbf{a}$.

Proof We may suppose that $\mathbf{a} \neq \mathbf{b}$, since if $\mathbf{a} = \mathbf{b}$ we can take any plane Π through $\mathbf{0}$ and $\mathbf{a}$. Our geometric intuition tells us that Π should be $\mathbf{x}{\cdot}\mathbf{n} = 0$, where $\mathbf{n} = (\mathbf{a} - \mathbf{b})/(||\mathbf{a} - \mathbf{b}||$, so let R be the reflection in this plane. Then $R(\mathbf{x}) = \mathbf{x} - 2(\mathbf{x}{\cdot}\mathbf{n})\mathbf{n}$, and a computation shows that $R(\mathbf{a}) = \mathbf{b}$. Note the use of the identity $2\mathbf{a}{\cdot}(\mathbf{a} - \mathbf{b}) = ||\mathbf{a} - \mathbf{b}||^2$ in this computation. □

The proof of Theorem 6.1.4 Let f be any isometry of $\mathbb{R}^3$. If $f(\mathbf{0}) \neq \mathbf{0}$ then then there is a reflection R_1 that interchanges $\mathbf{0}$ and $f(\mathbf{0})$. If f fixes $\mathbf{0}$, we let $R_1 = I$. Thus $R_1 f$ is an isometry that fixes $\mathbf{0}$, and we let $f_1 = R_1 f$. As $||f_1(\mathbf{k})|| = ||\mathbf{k}||$ there is a reflection R_2 in some plane through the origin that interchanges $f_1(\mathbf{k})$ and $\mathbf{k}$; thus $R_2 f_1$ is an isometry that fixes $\mathbf{0}$ and $\mathbf{k}$. Let $f_2 = R_2 f_1$, and note that by Lemma 6.1.5, $f(\mathbf{i})$ and $f(\mathbf{j})$ are orthogonal unit vectors in the plane $\mathbf{x}{\cdot}\mathbf{k} = 0$. There is a reflection R_3 in some vertical plane through the origin that maps $f(\mathbf{j})$ to $\mathbf{j}$ (and fixes $\mathbf{0}$ and $\mathbf{k}$) so that now, $R_3 f_2$ fixes $\mathbf{0}$, $\mathbf{j}$ and $\mathbf{k}$. Let $f_3 = R_3 f_2$; then f_3 maps $\mathbf{i}$ to $\pm\mathbf{i}$. If f_3 fixes $\mathbf{i}$, let $R_4 = I$. If not, let R_4 be the reflection in $\mathbf{x}{\cdot}\mathbf{i} = 0$. Then Lemma 6.1.6 implies that $R_4 f_3 = I$, so that $f = R_1 R_2 R_3 R_4$. □

A slight modification of the proof just given yields the next result.

Theorem 6.1.8 *The most general isometry f is of the form* $f(\mathbf{x}) = A(\mathbf{x}) + f(\mathbf{0})$, *where A is a linear map.*

Proof Suppose that f is an isometry. Then $A(\mathbf{x}) = f(\mathbf{x}) - f(\mathbf{0})$ is an isometry that fixes $\mathbf{0}$, and the proof of Theorem 6.1.4 shows that g is a composition of (at most three) reflections in planes through $\mathbf{0}$. As each such reflection is a linear map, so is A. □

It is clear that the composition of two isometries is an isometry, and the composition of functions is always associative. Trivially, the identity map is an isometry. If we write an isometry f as a composition, say $R_1 \cdots R_p$, of reflections R_j, then $f^{-1} = R_p \cdots R_1$ and this is an isometry. This proves the next result.

Theorem 6.1.9 *The set of isometries of* $\mathbb{R}^3$ *is a group with respect to composition.*

Although it may seem obvious that the inverse of an isometry f is an isometry, one has to prove (somewhere) that f^{-1} exists, and also that it is defined on $\mathbb{R}^3$.

We have seen that every isometry can be expressed as the composition of at most four reflections, and we want to give an example of an isometry that cannot be expressed as a composition of fewer than four reflections. We need some preliminary remarks. The reflection R given in (6.1.1) can be written as $R = T\tilde{R}$, where $\tilde{R}$ is the reflection in the plane $\mathbf{x}{\cdot}\mathbf{n} = 0$, and $T(\mathbf{x}) = \mathbf{x} + 2d\mathbf{n}$. Thus every reflection is a reflection in a plane through $\mathbf{0}$ followed by a translation. Suppose we have (with the obvious notation) $R_1(\mathbf{x}) = \tilde{R}_1(\mathbf{x}) + \mathbf{a}_1$, and similarly for R_2. Then, as $\tilde{R}_1$ is linear, we have

$$R_1R_2(\mathbf{x}) = \tilde{R}_1\big(\tilde{R}_2(\mathbf{x}) + \mathbf{a_2}\big) + \mathbf{a}_1 = \tilde{R}_1\tilde{R}_2(\mathbf{x}) + \tilde{R}_1(\mathbf{a_2}) + \mathbf{a}_1.$$

A similar argument applies to any composition of reflections; thus *given reflections $R_1, \dots, R_n$ there are reflections $\tilde{R}_1, \dots, \tilde{R}_n$ across planes through the origin, and a vector* $\mathbf{b}$*, such that* $R_1 \cdots R_n(\mathbf{x}) = \tilde{R}_1 \cdots \tilde{R}_n(\mathbf{x}) + \mathbf{b}$. We can now give our example.

Example 6.1.10 Let f be the isometry defined as the rotation of angle π about the $\mathbf{k}$-axis followed by the translation $\mathbf{x} \mapsto \mathbf{x} + \mathbf{k}$; thus f is a 'screw motion' which is given explicitly by $(x_1, x_2, x_3) \mapsto (-x_1, -x_2, x_3 + 1)$. Let A_1 and A_2 be the reflections in the planes $\mathbf{x}{\cdot}\mathbf{i} = 0$ and $\mathbf{x}{\cdot}\mathbf{j} = 0$, respectively; then $f(\mathbf{x}) = A_1A_2(\mathbf{x}) + \mathbf{k}$. Now suppose that f can be written as a composition of p reflections and, as above, write $f(\mathbf{x}) = \tilde{R}_1 \cdots \tilde{R}_p(\mathbf{x}) + \mathbf{b}$, where each $\tilde{R}_j$ is a reflection in a plane through $\mathbf{0}$. As $\mathbf{b} = f(\mathbf{0}) = \mathbf{k}$, we find that $A_2A_1 = \tilde{R}_1 \cdots \tilde{R}_p$. We shall prove below that if Q is any reflection in a plane though the origin then, for all vectors $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$, $[Q(\mathbf{x}), Q(\mathbf{y}), Q(\mathbf{z})] = -[\mathbf{x}, \mathbf{y}, \mathbf{z}]$ (this is because Q reverses the orientation of the vectors). Repeated applications of this rule applied to A_1A_2 and to $\tilde{R}_1 \cdots \tilde{R}_p$ give $(-1)^2 = (-1)^p$ so that p is even. It is clear that f is not the composition of exactly two reflections (for then f would be a translation, or would have a line of fixed points); thus f cannot be expressed as the composition of fewer than four reflections. □

The next result (which we have used in Example 6.1.10) shows how scalar and vector products behave under a reflection in a plane through the origin. In particular, such a reflection reverses the orientation of any three vectors, and this fact was used in Example 6.1.10.

Theorem 6.1.11 *Let R be the reflection in a plane Π that contains the origin. Then for all vectors* $\mathbf{x}$, $\mathbf{y}$ *and* $\mathbf{z}$,

$$R(\mathbf{x})\cdot R(\mathbf{y}) = \mathbf{x}\cdot\mathbf{y}, \tag{6.1.2}$$

$$R(\mathbf{x}) \times R(\mathbf{y}) = -R(\mathbf{x}\times\mathbf{y}), \tag{6.1.3}$$

$$[R(\mathbf{x}), R(\mathbf{y}), R(\mathbf{z})] = -[\mathbf{x}, \mathbf{y}, \mathbf{z}]. \tag{6.1.4}$$

Proof We may assume that Π is given by $\mathbf{x}\cdot\mathbf{n} = 0$, where $||\mathbf{n}|| = 1$, so that $R(\mathbf{x}) = \mathbf{x} - 2(\mathbf{x}\cdot\mathbf{n})\mathbf{n}$, and (6.1.2) follows immediately. Next, (6.1.3) follows from two applications of the formula (4.5.1) for the vector triple product, for

$$\begin{aligned} R(\mathbf{x}) \times R(\mathbf{y}) &= (\mathbf{x}\times\mathbf{y}) - 2(\mathbf{x}\cdot\mathbf{n})(\mathbf{n}\times\mathbf{y}) - 2(\mathbf{y}\cdot\mathbf{n})(\mathbf{x}\times\mathbf{n}) \\ &= (\mathbf{x}\times\mathbf{y}) - 2\mathbf{n}\times\big[(\mathbf{x}\cdot\mathbf{n})\mathbf{y} - (\mathbf{y}\cdot\mathbf{n})\mathbf{x}\big] \\ &= (\mathbf{x}\times\mathbf{y}) - 2\mathbf{n}\times\big[\mathbf{n}\times(\mathbf{y}\times\mathbf{x})\big] \\ &= (\mathbf{x}\times\mathbf{y}) - 2\big[\big(\mathbf{n}\cdot(\mathbf{y}\times\mathbf{x})\big)\mathbf{n} - (\mathbf{y}\times\mathbf{x})\big] \\ &= -R(\mathbf{x}\times\mathbf{y}). \end{aligned}$$

This proves (6.1.3). Finally, (6.1.4) holds because

$$\begin{aligned} [R(\mathbf{x}), R(\mathbf{y}), R(\mathbf{z})] &= R(\mathbf{x})\cdot\big(R(\mathbf{y})\times R(\mathbf{z})\big) \\ &= -R(\mathbf{x})\cdot\big(R(\mathbf{y}\times\mathbf{z})\big) \\ &= -\mathbf{x}\cdot(\mathbf{y}\times\mathbf{z}) \\ &= -[\mathbf{x}, \mathbf{y}, \mathbf{z}]. \end{aligned}$$

□

It is now clear how scalar and vector products behave under rotations that fix the origin; *if A is a rotation about the origin, then*

$$A(\mathbf{x})\cdot A(\mathbf{y}) = \mathbf{x}\cdot\mathbf{y}, \quad A(\mathbf{x})\times A(\mathbf{y}) = A(\mathbf{x}\times\mathbf{y}), \tag{6.1.5}$$

and

$$[A(\mathbf{x}), A(\mathbf{y}), A(\mathbf{z})] = [\mathbf{x}, \mathbf{y}, \mathbf{z}].$$

Note that rotations about the origin preserve the orientation of vectors, and this leads to our final result (which is proved without reference to matrices).

Theorem 6.1.12 *The set of rotations of $\mathbb{R}^3$ that fix $\mathbf{0}$ is a group with respect to composition.*

Proof We need only show that the composition of two rotations is again a rotation, for the rest of the requirements for a group are easily verified. Let

$R_3 = R_2R_1$, where R_1 and R_2 are rotations that fix $\mathbf{0}$. As we have remarked above, each rotation, and therefore R_3 also, preserves the orientation of vectors, and this shows that if we express R_3 as a composition of, say, p reflections in planes through the origin, then p is even. However, by Theorem 6.1.4, R_3 can be expressed as a composition of at most three reflections; thus we must be able to express R_3 as a composition of exactly two reflections, and hence R_3 is a rotation. □

Let f be an isometry, say $f(\mathbf{x}) = \mathbf{x}_0 + A(\mathbf{x})$, where A is the identity map I, or a composition of k reflections in planes through the origin, where k is one, two or three. We say that A is a *direct* isometry if $A = I$ or $k = 2$; otherwise we say that f is *indirect*. Direct isometries preserve orientation; indirect isometries reverse orientation.

Definition 6.1.13 A *screw-motion* in $\mathbb{R}^3$ is a rotation about some line in $\mathbb{R}^3$, followed by a translation in the direction of that that line.

We include the possibilities that the rotation is trivial (so the screw-motion is a translation) and that the translation is trivial (so that the screw-motion is a rotation). It is clear that every screw-motion is a direct isometry of $\mathbb{R}^3$; it is less clear that the converse is true.

Theorem 6.1.14 *Every direct isometry of $\mathbb{R}^3$ is a screw-motion.*

Proof As every translation is a screw-motion, we may confine our attention to a direct isometry f that is of the form $f(\mathbf{x}) = \mathbf{x}_0 + A(\mathbf{x})$, where A is a rotation (so $k = 2$). We choose a vector, say $\mathbf{a}$, lying along the axis of A. If $\mathbf{x}_0$ is a scalar multiple of $\mathbf{a}$, there is nothing to prove, but, in general, this will not be so.

A rotation of the same angle as A, but whose axis is a line in the direction $\mathbf{a}$, and passing through a point $\mathbf{w}$, is given by $\mathbf{x} \mapsto \mathbf{w} + A(\mathbf{x} - \mathbf{w})$; thus we need to be able to show that f can be written in the form $f(\mathbf{x}) = \mathbf{w} + A(\mathbf{x} - \mathbf{w}) + \lambda\mathbf{a}$ for some vector $\mathbf{w}$ and some real λ. Equivalently, because A is linear (so that $A(\mathbf{x} - \mathbf{w}) = A(\mathbf{x}) - A(\mathbf{w})$), we need to be able to find $\mathbf{w}$ and λ such that $\mathbf{x}_0 = \mathbf{w} - A(\mathbf{w}) + \lambda\mathbf{a}$. Now write $\mathbf{x}_0$ in the form $\mathbf{x}_1 + \mu\mathbf{a}$, where $\mathbf{x}_1 \perp \mathbf{a}$; then we need $\mathbf{w}$ and λ to satisfy $\mathbf{w} - A(\mathbf{w}) + \lambda\mathbf{a} = \mathbf{x}_1 + \mu\mathbf{a}$. We take $\lambda = \mu$, and then seek a solution $\mathbf{w}$ of $\mathbf{w} - A(\mathbf{w}) = \mathbf{x}_1$. If W is the plane given by $\mathbf{x}{\cdot}\mathbf{a} = 0$, it is apparent (by elementary geometry, or by using complex numbers) that there is some $\mathbf{w}$ in W such that $\mathbf{w} - A(\mathbf{w}) = \mathbf{x}_1$ and the proof is complete. □

Finally, we consider the following natural problem. Given rotations R_1 and R_2, how can we find the axis and angle of rotation of the rotation R_2R_1? One way is to let Π_0 be the plane containing the axes (which we may assume are distinct) of R_1 and R_2, and let α_0 be the reflection across Π_0. Then choose planes

Π_1 and Π_2 (both containing $\mathbf{0}$) so that if α_1 and α_2 denote reflections across these planes, then $R_2 = \alpha_2\alpha_0$ and $R_1 = \alpha_0\alpha_1$. Then $R_2R_1 = \alpha_2\alpha_0\alpha_0\alpha_1 = \alpha_2\alpha_1$. Further, if we examine the intersection of these planes with the unit sphere, then the calculations of angles and distances involved are a matter of spherical (rather than Euclidean) geometry. We also give an algebraic solution to this problem in terms of quaternions (in Section 6.3), and in terms of matrices (in Chapter 11).

Exercise 6.1

1. Find the formula for the reflection in the plane $x_2 = 0$ followed by the reflection in the plane $x_3 = 0$. Find the fixed points of this transformation, and identify it geometrically.
2. Given unit vectors **a** and **b**, let $R_{\mathbf{a}}$ be the reflection across the plane $\mathbf{x}{\cdot}\mathbf{a}$, and similarly for $R_{\mathbf{b}}$. Show that $R_{\mathbf{a}}R_{\mathbf{b}} = R_{\mathbf{b}}R_{\mathbf{a}}$ if and only if $\mathbf{a} \perp \mathbf{b}$.
3. Let $R_{\mathbf{a}}$ and $R_{\mathbf{b}}$ be the reflections in $\mathbf{x}{\cdot}\mathbf{a} = 0$ and $\mathbf{x}{\cdot}\mathbf{b} = 0$, respectively, where $||\mathbf{a}|| = 1$ and $||\mathbf{b}|| = 1$. Find a formula for the composition $R_{\mathbf{a}}(R_{\mathbf{b}}(\mathbf{x}))$. This composition of reflections is a rotation whose axis is the line of intersection of the two planes. Explain why this axis is the set of scalar multiples of $\mathbf{a} \times \mathbf{b}$, and verify analytically that both $R_{\mathbf{a}}$ and $R_{\mathbf{b}}$ fix every scalar multiple of $\mathbf{a} \times \mathbf{b}$.
4. Let L_1, L_2 and L_3 be three lines in the complex plane $\mathbb{C}$, each containing the origin, and let R_j be the reflection (of $\mathbb{C}$ into itself) across L_j. Show that $R_1R_2R_3$ is a reflection across some line in $\mathbb{C}$. Show, however, that if the L_j do not have a common point then $R_1R_2R_3$ need not be a reflection.
5. Suppose that $R_1, \dots, R_p, S_1, \dots, S_q$ are all reflections across planes that contain $\mathbf{0}$. Show that if $R_1 \cdots R_p = S_1 \cdots S_q$ then $(-1)^p = (-1)^q$. Compare this result with the definition of the signature of a permutation.

6.2 Quaternions

In 1843 Hamilton introduced quaternions as a way to generalize the algebra of complex numbers to higher dimensions. We shall use them to represent reflections and rotations in $\mathbb{R}^3$ algebraically. There are various equivalent ways of describing quaternions but fundamentally they are points in Euclidean four-dimensional space $\mathbb{R}^4$. Before we consider the various representations of quaternions it is convenient to introduce the idea of a Cartesian product of two sets. Given any two sets A and B, the Cartesian product set, which we denote by $A \times B$, is the set of all ordered pairs (a, b), where $a \in A$ and $b \in B$. In as far as there is a natural identification of the objects

$$(w, x, y, z), \quad \big((w, x), (y, z)\big), \quad \big(w, (x, y, z)\big),$$

in $\mathbb{R}^4$, $\mathbb{R}^2 \times \mathbb{R}^2$ and $\mathbb{R} \times \mathbb{R}^3$, respectively, we can identify each of these spaces with the others. Formally, a complex number is an ordered pair of real numbers so that $\mathbb{C} = \mathbb{R}^2$ (this really is equality here, and no identification is necessary), so we can add $\mathbb{C} \times \mathbb{C}$ to this list. Notice that a point in $\mathbb{R} \times \mathbb{R}^3$ is an ordered pair $(a, \mathbf{x})$, where $a \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^3$, and we can write this as $a + x_1\mathbf{i} + x_2\mathbf{j} + x_3\mathbf{k}$.

Definition 6.2.1 A *quaternion* is an expression $a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$, where a, b, c and d are real numbers. The set of quaternions is denoted by $\mathbb{H}$.

We now define an addition and a multiplication on the set $\mathbb{H}$ of quaternions, and it is sufficient to do this (and to carry out any subsequent calculations) in any of the models given above since a statement about one model can be converted on demand into a statement about another model. The addition of quaternions is just the natural vector addition in $\mathbb{R}^4$, namely

$$(x_1, x_2, x_3, x_4) + (y_1, y_2, y_3, y_4) = (x_1 + y_1, x_2 + y_2, x_3 + y_3, x_4 + y_4),$$

and clearly $\mathbb{R}^4$ is an abelian group with respect to addition. In terms of the representation in $\mathbb{R} \times \mathbb{R}^3$ this becomes $(a, \mathbf{x}) + (b, \mathbf{y}) = (a + b, \mathbf{x} + \mathbf{y})$, and in terms of $\mathbb{C} \times \mathbb{C}$ it is $(z_1, z_2) + (w_1, w_2) = (z_1 + w_1, z_2 + w_2)$.

The multiplication of quaternions is rather special. We recall how in Section 3.1 we motivated the multiplication of complex numbers by considering the formal product

$$(x + iy)(u + iv) = ux + i(xv + yu) + yvi^2,$$

and then imposing the condition that $i^2 = -1$ to return us to something of the form $p + iq$. We shall now adopt a similar strategy for quaternions. We write the general quaternion as $a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$ and then, using the distributive laws, we take a formal product of two such expressions. We now need to specify what we mean by the products of any two of $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$, and these are specified as follows:

(1) $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -1$, and
(2) $\mathbf{ij} = \mathbf{k} = -\mathbf{ji}, \quad \mathbf{jk} = \mathbf{i} = -\mathbf{kj}, \quad \mathbf{ki} = \mathbf{j} = -\mathbf{ik}$.

Notice that the product of one of these vectors with itself is a *scalar*; otherwise the product is a vector. Note also how (2) mimics the vector product of these vectors, and (1) mimics the formula $i^2 = -1$. We shall also assume that any real number commutes with each of the vectors $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$; for example, $3\mathbf{j} = \mathbf{j}3$. Of course, we could rewrite this whole discussion formally in terms of four-tuples of real numbers, but to do so would not contribute to an understanding of the underlying ideas. Indeed, the following example is probably more useful:

$$\begin{aligned}(2+3\mathbf{i}-4\mathbf{k})(\mathbf{i}-2\mathbf{j}) &= 2(\mathbf{i}-2\mathbf{j})+3\mathbf{i}(\mathbf{i}-2\mathbf{j})-4\mathbf{k}(\mathbf{i}-2\mathbf{j})\\ &= 2\mathbf{i}-4\mathbf{j}+3\mathbf{i}^2-6\mathbf{ij}-4\mathbf{ki}+8\mathbf{kj}\\ &= 2\mathbf{i}-4\mathbf{j}-3-6\mathbf{k}-4\mathbf{j}-8\mathbf{i}\\ &= -3-6\mathbf{i}-8\mathbf{j}-6\mathbf{k}.\end{aligned}$$

We note that the set of quaternions is closed under multiplication, but that multiplication is *not* commutative (for example, $\mathbf{ij} \neq \mathbf{ji}$); the reader should constantly be aware of this fact. It is an elementary (though tedious) matter to check that the associative law holds for multiplication, and that the distributive laws also hold. If we rewrite the definition of multiplication in terms of quaternions in the form $(a, \mathbf{x})$ we see a striking link with the vector algebra on $\mathbb{R}^3$; this is given in the next result.

Theorem 6.2.2 *The product of quaternions is given by the formula*

$$(a,\mathbf{x})(b,\mathbf{y}) = \big(ab-\mathbf{x}\cdot\mathbf{y},\, a\mathbf{y}+b\mathbf{x}+(\mathbf{x}\times\mathbf{y})\big), \tag{6.2.1}$$

where $\mathbf{x}\cdot\mathbf{y}$ *and* $\mathbf{x}\times\mathbf{y}$ *are the scalar and vectors products, respectively.*

The proof is by pure computation and we omit the details, but the reader should certainly verify this when $a = b = 0$. Note that multiplication by the 'real number' $(a, \mathbf{0})$ gives $(a,\mathbf{0})(b,\mathbf{y}) = (ab, a\mathbf{y}) = (b,\mathbf{y})(a,\mathbf{0})$. This shows that the quaternion $(1, \mathbf{0})$ acts as the identity element with respect to multiplication. Theorem 6.2.2 is of special interest when the two quaternions are pure quaternions, which we now define (these play the role of the purely imaginary numbers in $\mathbb{C}$).

Definition 6.2.3 A *pure quaternion* is a quaternion of the form $(0, \mathbf{x})$ where $\mathbf{x} \in \mathbb{R}^3$. The set of pure quaternions is denoted by $\mathbb{H}_0$.

When applied to pure quaternions, Theorem 6.2.2 yields the important formula

$$(0,\mathbf{x})(0,\mathbf{y}) = \big(-\mathbf{x}\cdot\mathbf{y},\, \mathbf{x}\times\mathbf{y}\big) \tag{6.2.2}$$

in which the quaternion product is expressed only in terms of the scalar and vector products. This shows that if q is the pure quaternion $(0, \mathbf{x})$, and $||\mathbf{x}|| = 1$, then $q^2 = (-1, \mathbf{0})$, and this extends the identities $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -1$ to *all* pure quaternions of unit length. It also shows that some quaternions, for example -1, have *infinitely many square roots*.

Let us return to the notation given in Definition 6.1.1 so that we now write $a + \mathbf{x}$ instead of $(a, \mathbf{x})$. The conjugate of a quaternion is defined by analogy with the conjugate of a complex number.

Definition 6.2.4 The *conjugate* of the quaternion $a + \mathbf{x}$ is $a - \mathbf{x}$. We denote the conjugate of a quaternion q by $\bar{q}$.

Theorem 6.2.5 *For quaternions p and q, we have $\overline{(pq)} = \bar{q}\,\bar{p}$.*

This follows directly from Theorem 6.2.2, but note that the order of terms here is important. Finally, we note an important identity: if $q = a + \mathbf{x}$ then, from (6.2.1),

$$q\bar{q} = (a + \mathbf{x})(a - \mathbf{x}) = a^2 + ||\mathbf{x}||^2.$$

Of course,

$$(a^2 + ||\mathbf{x}||^2)^{1/2} = \sqrt{a^2 + x_1^2 + x_2^2 + x_3^2},$$

and as this is the distance between the origin and the point $a + \mathbf{x}$ viewed as a point of $\mathbb{R}^4$, we write this as $||a + \mathbf{x}||$. This is the *norm* of the quaternion, and it obeys the important rule that for any quaternions p and q,

$$||pq|| = ||p||\,||q||.$$

The proof is easy for, using the associative law and Theorem 6.2.5,

$$||pq||^2 = (pq)\overline{(pq)} = p\,q\,\bar{q}\,\bar{p} = p\,||q||^2\,\bar{p} = ||q||^2\,p\bar{p} = ||q||^2\,||p||^2 \tag{6.2.3}$$

as required. □

Finally, we have just seen that if q is not the zero quaternion, and if $q' = ||q||^{-2}\bar{q}$, then $qq' = 1 = q'q$. Thus each non-zero quaternion has a multiplicative inverse, and we have proved the following result.

Theorem 6.2.6 *The set of non-zero quaternions is a non-abelian group with respect to the multiplication defined above.*

In conclusion, the set $\mathbb{H}$ of quaternions, with the addition and multiplication defined above, satisfies all of the axioms for a field *except* that multiplication is not commutative. Accordingly, the quaternions are sometimes referred to as a *skew field*.

Exercise 6.2

1. Simplify the sequence of quaternions $\mathbf{i}$, $\mathbf{ij}$, $\mathbf{ijk}$, $\mathbf{ijki}$, $\mathbf{ijkij}$, ...
2. Let $q = a + \mathbf{x}$. Express a and $\mathbf{x}$ in terms of q and $\bar{q}$.
3. Let q be a pure quaternion. Compute q^2, and hence show that $q^{-1} = -||q||^{-2}q$.

4. Show that $\{1, -1, \mathbf{i}, -\mathbf{i}, \mathbf{j}, -\mathbf{j}, \mathbf{k}, -\mathbf{k}\}$ is a group under multiplication.
5. Prove (6.2.1).
6. Prove Theorem 6.2.5.
7. Show that a quaternion q is a pure quaternion if and only if q^2 is real and not positive.
8. Let p and q be pure quaternions. Show that (in the obvious sense) pq is the vector product $p \times q$ if and only if $p \perp q$ (in $\mathbb{R}^3$). In the same sense, show that $pq = qp$ if and only if $p \times q = 0$.

6.3 Reflections and rotations

The formula (6.1.1) for a reflection across a plane is simple; however the formula for a rotation, obtained by combining two reflections, is not. In this section we shall see how quaternions provide an alternative algebraic way to express reflections and rotations. Quaternions are points of $\mathbb{R}^4$, but we shall pay special attention to the space $\mathbb{H}_0$ of pure quaternions which we identify with $\mathbb{R}^3$.

Consider the map $\theta : \mathbb{H} \to \mathbb{H}$ given by $\theta(y) = -qyq^{-1}$, where q is a non-zero pure quaternion. It is clear that θ is a linear map; that is, $\theta(\lambda_1 y_1 + \lambda_2 y_2) = \lambda_1\theta(y_1) + \lambda_2\theta(y_2)$ and, trivially, $\theta(0) = 0$ and $\theta(q) = -q$. These facts suggest that θ might be related to the reflection across the plane with normal q, and we shall now show that this is so.

Theorem 6.3.2 *The map $\theta : \mathbb{H} \to \mathbb{H}$ given by $\theta(y) = -qyq^{-1}$, where q is a non-zero pure quaternion, maps the set $\mathbb{H}_0$ of pure quaternions into itself. Further, when $\mathbb{H}_0$ is identified with $\mathbb{R}^3$, and q with $\mathbf{q}$, θ is the reflection across the plane $\mathbf{x}\cdot\mathbf{q} = 0$. If $\mathbf{q}$ is a pure quaternion of unit length, then $\theta(y) = qyq$.*

Proof We begin by looking at the fixed points of θ in $\mathbb{H}_0$. A general quaternion y is fixed by θ if and only if $-qy = yq$, and if we write $q = (0, \mathbf{q})$ and $y = (0, \mathbf{y})$, and use (6.2.2), we see that a pure quaternion $\mathbf{y}$ is fixed by θ if and only if $\mathbf{y}\cdot\mathbf{q} = 0$. Now let Π be the plane in $\mathbb{R}^3$ given by $\mathbf{x}\cdot\mathbf{q} = 0$; then θ fixes each point of Π. Now take any pure quaternion $\mathbf{y}$. We can write $\mathbf{y} = \mathbf{p} + \lambda\mathbf{q}$, where λ is real and $\mathbf{p} \in \Pi$. Thus, as θ is linear,

$$\theta(\mathbf{y}) = \theta(\mathbf{p}) + \lambda\theta(\mathbf{q}) = \mathbf{p} - \lambda\mathbf{q}.$$

This shows that $\theta(\mathbf{y})$ is a pure quaternion, and also that it is the reflection of $\mathbf{y}$ in Π. Finally, if $\mathbf{q}$ is a pure quaternion of unit length, then $q^2 = -1$ so that $q^{-1} = -q$ and $\theta(y) = qyq$. □

Before moving on to discuss rotations, we give a simple example.

Example 6.3.3 The reflection across the plane $\mathbf{x}\cdot\mathbf{k} = 0$ is given by $(x_1, x_2, x_3) \mapsto (x_1, x_2, -x_3)$ or, in terms of quaternions, $\mathbf{x} \mapsto \mathbf{kxk}$. We also note (in preparation for what follows) that as $\mathbf{k}^{-1} = -\mathbf{k}$, the map $\mathbf{x} \mapsto \mathbf{kxk}^{-1}$ is given by $(x_1, x_2, x_3) \mapsto (-x_1, -x_2, x_3)$, and this is a rotation of angle π about the **k**-axis. □

We now use quaternions to describe rotations that fix the origin. Consider the rotation R obtained by the reflection across the plane with normal unit **p** followed by the reflection across the plane with unit normal **q**. According to Theorem 6.3.2,

$$R(\mathbf{x}) = q(p\mathbf{x}p)q = (qp)\mathbf{x}(pq), \tag{6.3.1}$$

where $p = (0, \mathbf{p})$ and $q = (0, \mathbf{q})$. This observation yields the following result.

Theorem 6.3.4 *Let $r = (\cos\frac{1}{2}\theta, \sin\frac{1}{2}\theta\,\mathbf{n})$, where **n** is a unit vector. Then r is a unit quaternion, and the map $\mathbf{x} \mapsto r\mathbf{x}r^{-1} = r\mathbf{x}\bar{r}$ is a clockwise rotation by an angle θ about the axis in the direction **n**.*

Proof Choose unit vectors **p** and **q** such that $\mathbf{p}\cdot\mathbf{q} = \cos\frac{1}{2}\theta$ and $\mathbf{p} \times \mathbf{q} = \sin\frac{1}{2}\theta\,\mathbf{n}$. Then the rotation described (geometrically) in the theorem is reflection in $\mathbf{x}\cdot\mathbf{p} = 0$ followed by reflection in $\mathbf{x}\cdot\mathbf{q} = 0$; thus it is given by R in (6.3.1). Now a simple calculation using (6.2.2) shows that $qp = -r$ and $pq = -\bar{r}$. Thus $R(\mathbf{x}) = (-r)\mathbf{x}(-\bar{r}) = r\mathbf{x}\bar{r}$. Finally, as $||r|| = 1$ we see that $\bar{r} = r^{-1}$. □

Theorem 6.3.4 provides a way to find the geometric description of the composition of two given rotations. Let R_r and R_s be the rotations associated with the quaternions

$$r = \cos\tfrac{1}{2}\theta + \sin\tfrac{1}{2}\theta\,\mathbf{n}, \quad s = \cos\tfrac{1}{2}\varphi + \sin\tfrac{1}{2}\varphi\,\mathbf{m}.$$

Then the composition $R_s R_r$ is the map $\mathbf{x} \mapsto s(r\mathbf{x}\bar{r})\bar{s} = (sr)\mathbf{x}(\overline{sr})$, and as sr is also a unit quaternion (which is necessarily of the form $\cos\frac{1}{2}\psi + \sin\frac{1}{2}\psi\,\mathbf{h}$ for some unit vector **h**), we see that $R_s R_r = R_{sr}$. By computing sr as a quaternion product, we can find the axis and angle of rotation of the composition $R_s R_r$. We urge the reader to experiment with some simple examples to see this idea in action.

Exercise 6.3

1. Let r be a unit quaternion ($||\mathbf{r}|| = 1$). Show that we can write $r = \cos\theta + \sin\theta\,\mathbf{p}$, where **p** is a unit pure quaternion.

2. Use Theorem 6.2.2 to show that if p and q are pure quaternions ($p = \mathbf{p}$ and $q = \mathbf{q}$), and if $\mathbf{p} \perp \mathbf{q}$, then $-qpq^{-1} = p$.
3. Use quaternions to find the image of $\mathbf{x}$ under a rotation about the $\mathbf{k}$-axis of angle $\pi/6$. Now verify your result by elementary geometry.
4. Let Π be the plane given by $\mathbf{x}{\cdot}\mathbf{n} = 0$, where $\mathbf{n} = (1, -1, 0)$, and let R be the reflection across Π. Write $\mathbf{x} = (a, b, c)$ and $\mathbf{y} = R(\mathbf{x})$. Verify, both using elementary geometry and quaternion algebra, that $\mathbf{y} = (b, a, c)$.

(' ' !3 " D23 4 2

15

Hyperbolic geometry

15.1 The hyperbolic plane

In the earlier chapters we have discussed both Euclidean geometry and spherical (non-Euclidean) geometry, and in this last chapter we discuss a second type of non-Euclidean geometry, namely hyperbolic geometry. Gauss introduced the term *non-Euclidean geometry* to describe a geometry which does *not* satisfy Euclid's axiom of parallels, namely that if a point P is not on a line L, then there is exactly one line through P that does not meet L. In spherical geometry, the 'lines' are the great circles, and in this case any two lines meet. Hyperbolic geometry is a geometry in which there are infinitely many lines through the point P that do not meet the line L, and it was developed independently by Gauss (in Germany), Bolyai (in Hungary) and Lobatschewsky (in Russia) around 1820.

We begin by describing the points and lines of hyperbolic geometry without any reference to distance. We shall take the hyperbolic plane to be the upper half-plane $\mathcal{H} = \{x + iy : y > 0\}$ in $\mathbb{C}$. Notice that the real axis $\mathbb{R}$ is not part of $\mathcal{H}$. A *hyperbolic line* (that is, a line in the hyperbolic geometry) is a semicircle in $\mathcal{H}$ whose centre lies on $\mathbb{R}$; such semi-circles are orthogonal to $\mathbb{R}$. However, as our concept of circles includes 'straight lines' (see Chapter 14), we must also regard those straight lines that are orthogonal to $\mathbb{R}$ as hyperbolic lines. Figure 15.1.1 illustrates the hyperbolic lines in $\mathcal{H}$, and we remark that the two 'different' types of line are only different because we are viewing them from a Euclidean perspective.

We notice immediately that Euclid's Parallel Axiom fails; indeed, the two semi-circles have a common point P that does not lie on the line L; moreover, it is easy to see that there are infinitely many hyperbolic lines through P and not meeting L. It is clear, however, that *any two hyperbolic lines meet in at most one point*, and that *there is is a unique hyperbolic line through any two distinct points in $\mathcal{H}$.*

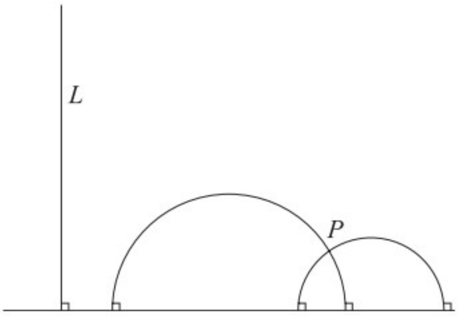

Figure 15.1.1

Now let

$$\Gamma = \left\{ z \mapsto \frac{az+b}{cz+d} : a, b, c, d \in \mathbb{R},\ ad - bc > 0 \right\}.$$

First, we leave the reader to check that Γ is a group. Next, we note that if g is in Γ then it maps $\mathcal{H}$ into itself. Indeed, if we write $g(z) = (az + b)/(cz + d)$, and $z = x + iy$, then

$$\begin{aligned} \operatorname{Im}[g(x+iy)] &= \frac{(az+b)\overline{(cz+d)}}{(cz+d)\overline{(cz+d)}} \\ &= \frac{(ad-bc)\,y}{|cz+d|^2} > 0. \end{aligned} \tag{15.1.1}$$

Exactly the same reason shows that g also maps the lower half-plane (given by $y < 0$) into itself; thus g must also map the circle $\mathbb{R} \cup \{\infty\}$ into itself. As Γ is a group, the same holds for g^{-1}; thus

$$g\bigl(\mathbb{R} \cup \{\infty\}\bigr) = \mathbb{R} \cup \{\infty\} = g^{-1}\bigl(\mathbb{R} \cup \{\infty\}\bigr).$$

This implies that the coefficients a, b, c and d in g may be chosen to be real (note that we cannot assert that they *are* real, for they are only determined to within a complex scalar multiple). The case $c = 0$ is easy, so we may assume that $c \neq 0$. Then, by scaling the coefficients by the factor $1/c$, we can choose these coefficients so that, in effect, $c = 1$. Then, as $g(\infty) = a$, and $g^{-1}(\infty) = -d$, we see that a and d are real. Finally, if $a = 0$ then $-b = ad - bc > 0$ so that b is real. If, however, $a \neq 0$, then $-b/a = g^{-1}(0)$ so that b is again real. To summarize: if $g(z) = (az + b)/(cz + d)$, and $g \in \Gamma$, then we may assume that a, b, c and d are real. This implies that, for each z,

$$\overline{g(z)} = \overline{\left(\frac{az+b}{cz+d}\right)} = \frac{\bar{a}\bar{z}+\bar{b}}{\bar{c}\bar{z}+\bar{d}} = \frac{a\bar{z}+b}{c\bar{z}+d} = g(\bar{z}). \tag{15.1.2}$$

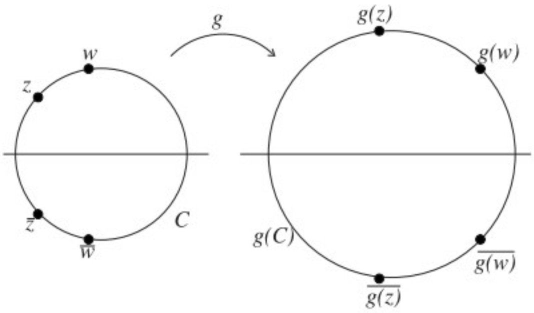

Figure 15.1.2

Let us now see how g in Γ acts on hyperbolic lines. Let z and w be distinct points of $\mathcal{H}$, and let us consider the hyperbolic line L through z and w. Now L is part of the (unique) Euclidean circle C that passes through z, w, $\bar{z}$ and $\bar{w}$, so that $g(C)$ passes through $g(z)$, $g(w)$, $g(\bar{z})$ and $g(\bar{w})$. However, (15.2.1) now implies that $g(C)$ passes through $g(z)$, $\overline{g(z)}$, $g(w)$ and $\overline{g(w)}$, so that $g(C)$ is orthogonal to $\mathbb{R}$ (see Figure 15.1.2). It follows that $g(L) = \mathcal{H} \cap g(C)$, and this proves the following result.

Theorem 15.1.1 *If $g \in \Gamma$ and L is a hyperbolic line, then $g(L)$ is a hyperbolic line.*

Theorem 15.1.1 suggests that the elements of Γ might be regarded as the rigid motions of hyperbolic geometry. This suggestion is strengthened by the fact (which will not be proved here) that any bijective map of $\mathbb{C}_\infty$ onto itself that maps circles to circles is a Möbius map of z or of $\bar{z}$; this is a type of converse of Theorem 13.3.2. Further, any Möbius map that preserves $\mathcal{H}$ must be in Γ (Exercise 15.1.1). In the next section we shall introduce a distance in $\mathcal{H}$, and we shall then see that the elements of Γ are indeed the isometries of $\mathcal{H}$.

There is a second model of the hyperbolic plane which is useful, and often preferable to the model $\mathcal{H}$. In this model the hyperbolic plane is the unit disc $\mathbb{D}$, namely $\{z : |z| < 1\}$ (see Figure 15.1.3). The Möbius map $g(z) = (z - i)/(z + i)$ maps $\mathcal{H}$ onto $\mathbb{D}$ (because $\mathcal{H}$ is given by $|z - i| < |z + i|$), and so we may take the hyperbolic lines in the model $\mathbb{D}$ to be the images under g of the hyperbolic lines in $\mathcal{H}$. Thus the hyperbolic lines in $\mathbb{D}$ are the arcs of circles in $\mathbb{D}$ whose endpoints lie on the cirle $|z| = 1$ and which are orthogonal to this circle at their endpoints. The two models $\mathcal{H}$ and $\mathbb{D}$ may be used interchangeably, and any result about one may be transferred to the other by any Möbius map that maps $\mathcal{H}$ to $\mathbb{D}$, or $\mathbb{D}$ to $\mathcal{H}$.

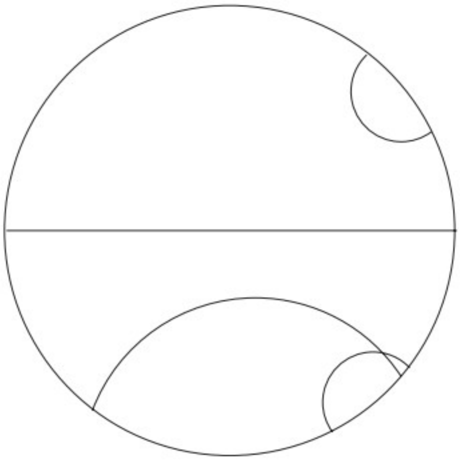

Figure 15.1.3

Exercise 15.1

1. Show that if f is a Möbius map which maps $\mathbb{R} \cup \{\infty\}$ onto itself, then f can be written in the form $f(z) = (az + b)/(cz + d)$, where a, b, c, d are real and $ad - bc \neq 0$. Show further that $f(\mathcal{H}) = \mathcal{H}$ if and only if $ad - bc > 0$.
2. Let z_1 and z_2 be distinct points in the hyperbolic plane. Show that there is a unique hyperbolic line that passes through z_1 and z_2.
3. Suppose that w_1 and w_2 are in $\mathcal{H}$. Show that there is some g in Γ such that $g(w_1) = w_2$. This shows that *the stabilizer of any point in $\mathcal{H}$ is conjugate to the stabilizer of any other point in $\mathcal{H}$.*
4. Verify the steps in the following argument. Let $g(z) = (z - i)/(z + i)$; then g maps $\mathcal{H}$ onto $\mathbb{D}$, where $\mathbb{D} = \{z : |z| < 1\}$. Note that $g(i) = 0$ and $g(-i) = \infty$. Now suppose that f maps $\mathcal{H}$ onto itself and fixes i; then f also fixes $-i$. It follows that gfg^{-1} maps $\mathbb{D}$ onto itself, and fixes 0 and ∞. Thus gfg^{-1} is a Euclidean rotation about the origin, and hence *the group of hyperbolic isometries that fix a given point w is isomorphic to the group of Euclidean rotations that fix the origin.*

15.2 The hyperbolic distance

We shall now introduce a distance in $\mathcal{H}$, and then show that the elements of Γ are isometries for this distance (in fact, they are the only orientation-preserving isometries). There are two ways to define this *hyperbolic distance*, and we shall

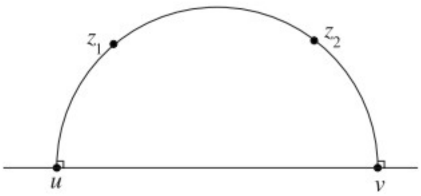

Figure 15.2.1

start with the more elementary way. Consider distinct points z_1 and z_2 in $\mathcal{H}$, and let L be the hyperbolic line through these points. Then L has endpoints u and v, say, chosen so that u, z_1, z_2 and v occur in this order along L (see Figure 15.2.1). We can find a Möbius map g in Γ such that $g(u) = 0$ and $g(v) = \infty$ (see Exercise 15.2.1); then $g(z_1) = ia$ and $g(z_2) = ib$, say, where $0 < a < b$. If we now recall that cross-ratios are invariant under Möbius maps, we have

$$[u, z_1, z_2, v] = [0, ia, ib, \infty] = b/a > 1.$$

This allows us to make the following definition.

Definition 15.2.1 The *hyperbolic distance* between z_1 and z_2 in $\mathcal{H}$ is $\log [u, z_1, z_2, v]$ when $z_1 \neq z_2$, and zero otherwise. We denote this distance by $\rho(z_1, z_2)$. □

Theorem 15.2.2 *The elements of* Γ *preserve the hyperbolic distance between two points in* $\mathcal{H}$.

Proof This is immediate because the hyperbolic distance is defined as a cross-ratio, cross-ratios are invariant under Möbius maps, and each g in Γ is a Möbius map. □

Notice that if, in Definition 15.2.1, we have $z_1 = ia$ and $z_2 = ib$, where $0 < a < b$, then

$$\rho(ia, ib) = \log[0, ia, ib, \infty] = \log b/a$$

(see Definition 13.4.1). This leads to the following more general result.

Theorem 15.2.3 *The hyperbolic distance is additive along hyperbolic lines.*

Proof Suppose that z_1, z_2, z_3 lie on a hyperbolic line L with end-points u and v such that u, z_1, z_2, z_3, v occur in this order along L. We can find some g in Γ such that $g(u) = 0$ and $g(v) = \infty$, and then $g(z_j) = ia_j$, where $0 < a_1 < a_2 < a_3$.

As $\rho(z_i, z_j) = \log a_j/a_i$ when $i < j$ we see that

$$\begin{aligned}\rho(z_1, z_3) &= \log a_3/a_1\\ &= \log a_3/a_2 + \log a_2/a_1\\ &= \rho(z_1, z_2) + \rho(z_2, z_3).\end{aligned}$$

□

We are now in a position to give explicit formulae for the hyperbolic distance (see Exercise 15.2.2).

Theorem 15.2.4 *For z and w in $\mathcal{H}$,*

$$\sinh^2 \frac{1}{2}\rho(z, w) = \frac{|z - w|^2}{4\,\mathrm{Im}[z]\,\mathrm{Im}[w]}, \qquad (15.2.1)$$

$$\cosh^2 \frac{1}{2}\rho(z, w) = \frac{|z - \bar{w}|^2}{4\,\mathrm{Im}[z]\,\mathrm{Im}[w]}. \qquad (15.2.2)$$

Proof First, choose g in Γ (as above) so that $g(z) = ia$ and $g(w) = ib$, where $0 < a < b$. By applying the map $z \mapsto z/a$ (which is in Γ), we may assume that $a = 1$. Then $\rho(z, w) = \rho(i, ib) = \log b$, so that

$$\sinh^2 \frac{1}{2}\rho(z, w) = \sinh^2\big(\log\sqrt{b}\big) = \frac{(b-1)^2}{4b}. \qquad (15.2.3)$$

Next, let

$$F(z, w) = \frac{|z - w|^2}{4\,\mathrm{Im}[z]\,\mathrm{Im}[w]}.$$

Then, from (13.1.2) and (15.1.1), we see that F is invariant under any g in Γ; that is,

$$F\big(g(z), g(w)\big) = F(z, w).$$

Thus

$$F(z, w) = F(i, ib) = \frac{(b-1)^2}{4b}, \qquad (15.2.4)$$

and this together with (15.2.3) gives (15.2.1). The second formula (15.2.2) follows from the fact that, for all z, $\cosh^2 z = 1 + \sinh^2 z$. □

We remark that, as Theorem 15.2.4 suggests, in calculations involving the hyperbolic distance it is almost always advantageous to use the functions sinh or cosh of $\rho(z, w)$ or $\frac{1}{2}\rho(z, w)$; *only rarely is $\rho(z, w)$ used by itself.*

We end with a brief discussion of an alternative (but equivalent) way to define distance. First, we define the *hyperbolic length* of a curve γ in $\mathcal{H}$ to be the line integral

$$\int_\gamma \frac{|dz|}{y},$$

where, as usual $z = x + iy$. Now let L be the hyperbolic line through two points z and w in $\mathcal{H}$, and let σ be the arc of L that lies between z and w. It can be shown that σ has hyperbolic length $\rho(z, w)$ and, moreover, that any other curve joining z to w has a greater hyperbolic length than σ. Thus the hyperbolic line through two points does indeed give the shortest path between these points.

Finally, if g is in Γ, and $g(z) = (az + b)/(cz + d)$, then, from (13.1.2), we see that

$$|g'(z)| = \frac{|ad - bc|}{|cz + d|^2},$$

where $g'(z)$is the usual derivative of g. In conjunction with (15.1.1) this gives

$$\frac{|g'(z)|}{\mathrm{Im}[g(z)]} = \frac{1}{\mathrm{Im}[z]}.$$

This (together with the formula for a change of variable in a line integral) shows that for each g in Γ, and each curve γ, $g(\gamma)$ has the same hyperbolic length as γ.

Exercise 15.2

1. Show that for any u and v in $\mathbb{R} \cup \{\infty\}$ with $u \neq v$, there is a g in Γ with $g(u) = 0$ and $g(v) = \infty$. [Hint: apply $z \mapsto -1/(z - v)$ and then a translation.]
2. The functions sinh and cosh are defined by $\sinh z = (e^z - e^{-z})/2$ and $\cosh z = (e^z + e^{-z})/2$. Show that (a) $\cosh^2 z - \sinh^2 z = 1$, and (b) $\cosh 2z = 2\cosh^2 z - 1 = 1 + 2\sinh^2 z$.
3. Find the hyperbolic distance between the points $1 + iy$ and $-1 + iy$ as a function of y. Show that for a given positive t there is a value of y such that this distance is t.
4. Let L be the Euclidean line given by $\mathrm{Im}[z] = 2$. Show that $2i$ is the point on L that is closest (as measured by the hyperbolic distance) to the point i.

15.3 Hyperbolic circles

Suppose that $w \in \mathcal{H}$, and $r > 0$. The *hyperbolic circle* with *hyperbolic centre* w and *hyperbolic radius* r is the set $\{z \in \mathcal{H} : \rho(z, w) = r\}$.

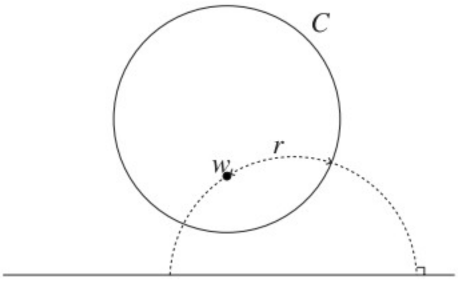

Figure 15.3.1

Theorem 15.3.1 *Each hyperbolic circle is a Euclidean circle in* $\mathcal{H}$.

Proof Let C be the hyperbolic circle with centre w and radius r. There is a map g in Γ with $g(w) = i$ (see Exercise 15.1.2), so that $g(C)$ is the hyperbolic circle with centre i and radius r. Now by Theorem 15.2.4, $z \in g(C)$ if and only if $|z - i|^2/4y = \sinh^2 \frac{1}{2}r$, where $z = x + iy$. This equation simplifies to give $x^2 + (y - \cosh r)^2 = \sinh^2 r$, so that $g(C)$ is a Euclidean circle in $\mathcal{H}$. As g^{-1} maps circles to circles, and $\mathcal{H}$ to itself, we see that $g^{-1}\big(g(C)\big)$, namely C, is a Euclidean circle in $\mathcal{H}$. □

Notice that the hyperbolic centre of a hyperbolic circle is *not* the same as its Euclidean centre (and similarly for the radii); indeed, the hyperbolic circle $g(C)$ in the proof of Theorem 15.3.1 has hyperbolic centre i, and Euclidean centre $i \cosh r$ (and $\cosh r > 1$). A hyperbolic circle with centre w and hyperbolic radius r is illustrated in Figure 15.3.1. Finally, it can be shown that the length of a hyperbolic circle of hyperbolic radius r is $2\pi \sinh r$, and that its hyperbolic area (which we have not defined) is $4\pi \sinh^2(\frac{1}{2}r)$. Notice that the hyperbolic radius of a hyperbolic circle of radius r grows roughly like πe^r; in the Euclidean case, it is $2\pi r$. Finally, we mention (but do not prove) the hyperbolic counterpart of the fact that the area of a spherical triangle is π less than its angle sum.

Theorem 15.3.2 *The area of a hyperbolic triangle with angles* α, β *and* γ *is* $\pi - (\alpha + \beta + \gamma)$. *In particular, this area cannot exceed* π.

Exercise 15.3

1. Find the equation of the hyperbolic circle with centre $2i$ and radius e^2. Suppose that this circle meets the imaginary axis at ia and ib, where $0 < a < 2 < b$. Find a and b, and verify directly that $\rho(ia, 2i) = e^2 = \rho(2i, ib)$.

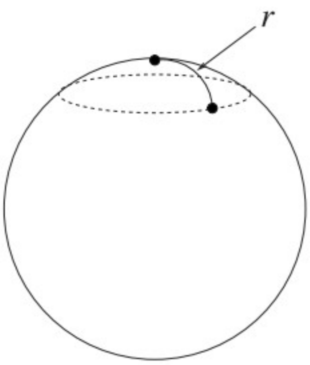

Figure 15.3.2

2. Consider the unit sphere in $\mathbb{R}^3$ as a model of spherical geometry in which distances are measured on the surface of the sphere. What is the circumference of the circle whose centre is at the ‘north pole’ (see Figure 15.3.2) and whose radius is r? Now compare the circumference of a circle of radius r in Euclidean geometry, spherical geometry and hyperbolic geometry.

15.4 Hyperbolic trigonometry

A hyperbolic triangle is a triangle whose sides are arcs of hyperbolic lines. We begin with the hyperbolic version of Pythagoras’ theorem (see Figure 15.4.1).

Theorem 15.4.1 *Suppose that a hyperbolic triangle has sides of hyperbolic lengths a, b and c, and that the two sides of lengths a and b are orthogonal. Then* $\cosh c = \cosh a \; \cosh b$.

In our proof of this result we shall need to use the fact that a Möbius map is *conformal*; that is, it preserves the angles between circles. In particular, this implies that if two circles C and C' are orthogonal, and if g is any Möbius map, then $g(C)$ and $g(C')$ are orthogonal. We shall not give a proof of this (although the proof is not difficult).

Proof Let the vertices of the hyperbolic triangle be v_a, v_b and v_c, where v_a is opposite the side of length a, and so on. There is some g in Γ such that $v_c = i$ and $v_b = ik$, where $k > 1$. As g preserves the orthogonality of circles we see that g maps v_a to some point $s + it$, where $s^2 + t^2 = 1$ (see Figure 15.4.2). As g preserves hyperbolic distances, this means that we may assume that $v_a = s + it$,

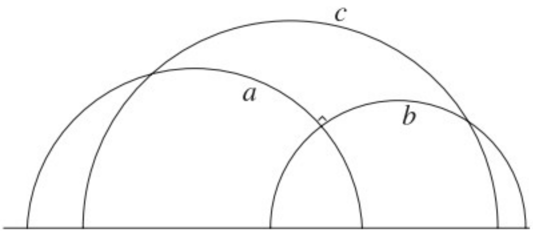

Figure 15.4.1

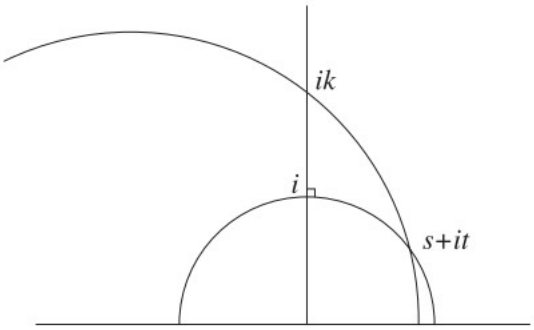

Figure 15.4.2

$v_b = ik$ and $v_c = i$. Now $\rho(i, ik) = a$, $\rho(i, s + it) = b$ and $\rho(ik, s + it) = c$. As

$$\cosh a = \frac{k^2 + 1}{2k}, \quad \cosh b = \frac{1}{t}, \quad \cosh c = \frac{k^2 + 1}{2tk},$$

the given formula follows. □

It is interesting to examine Pythagoras' theorem for small triangles, and for large triangles. As $\cosh z = 1 + z^2/2\,! + z^4/4\,! + \cdots$, we see that when a, b and c are very small, the formula is, up to the second-order terms, $c^2 = a^2 + b^2$. Thus, infinitesimally, the hyperbolic version of Pythagoras' theorem agrees with the Euclidean version. This is because the hyperbolic distance is obtained form the Euclidean distance by applying a 'local scaling factor' of $1/y$ at z. As this scaling factor is essentially constant on an infinitesimal neighbourhood of a point, the 'infinitesimal hyperbolic geometry is just a scaled version of the Euclidean geometry. However, as the scaling factor varies considerably over large distances, the global hyperbolic geometry is very different from the Euclidean geometry. For example, if in Pythagoras' theorem, a, b and c are all very large, then, as $\cosh x$ is approximately $e^x/2$ when x is large, we have (approximately) $4e^c = e^a e^b$ so that $c = a + b - \log 2$. In other words, in a 'large' hyperbolic right-angled triangle, the length of the hypotenuse is almost the sum of the lengths of the other two sides! If this were the case in Euclidean geometry, then the triangle would be very 'flat', but this is not so in hyperbolic geometry.

Finally, we remark that hyperbolic trigonometry is as rich and well understood as Euclidean trigonometry (and spherical trigonometry) is. For example, there is a sine rule, and cosine rules in hyperbolic geometry. In most applications it is the hyperbolic trigonometry that is important, and hyperbolic geometry by itself has relatively few applications.

Exercise 15.4

1. Suppose that a, b and c are the sides of a right-angled hyperbolic triangle with the right-angle opposite the side of length c. Prove that $c \leq a + b$; this is a special case of the triangle inequality.
2. Consider a right-angled hyperbolic triangle with both sides ending at the right angle having length a. Let the height of this triangle be h (the distance from the right angle to the third side). Find h as a function of a. What is the limiting behaviour of h as $a \to +\infty$?

15.5 Hyperbolic three-dimensional space

We end this text with a very brief description of three-dimensional hyperbolic geometry, and a sketch of the proof of Theorem 14.3.2. These are given in this and the next section, and they combine many of the ideas that have been introduced in this text. We take *hyperbolic space* to be the upper-half of $\mathbb{R}^3$, namely

$$\mathcal{H}^3 = \{(x, y, t) \in \mathbb{R}^3 : t > 0\}.$$

It is convenient to identify the point (x, y, t) with the quaternion $x + y\mathbf{i} + t\mathbf{j}$, and also to identify the quaternion $\mathbf{i}$ with the complex number i. Thus we can write (x, y, t) as $z + t\mathbf{j}$, where z is the complex number $x + iy$. Note that in this notation we have the convenient formula

$$z\mathbf{j} = (x + y\mathbf{i})\mathbf{j} = x\mathbf{j} + y\mathbf{k} = x\mathbf{j} - y\mathbf{j}\mathbf{i} = \mathbf{j}\bar{z}. \tag{15.5.1}$$

Suppose now that $g(z) = (az + b)/(cz + d)$, where $ad - bc \neq 0$. We can now let g act on hyperbolic space $\mathcal{H}^3$ by the rule

$$g : z + t\mathbf{j} \mapsto \left[a(z + t\mathbf{j}) + b\right]\left[c(z + t\mathbf{j}) + d\right]^{-1}, \tag{15.5.2}$$

where this computation is to be carried out in the algebra of quaternions. This lengthy (but elementary) exercise shows that

$$g(z + t\mathbf{j}) = \frac{(az + b)(\bar{c}\bar{z} + \bar{d}) + a\bar{c}t^2 + |ad - bc|t\mathbf{j}}{|cz + d|^2 + |c|^2t^2}. \qquad (15.5.3)$$

Notice that as quaternions are not commutative, we have to choose (and then be consistent about) which side we shall write the inverse in (15.5.2). However, in (15.5.3), the denominator is real (and positive), and as every real number commutes with every quaternion, we can write it in the usual form for a fraction without any ambiguity. Notice also that if we put $t = 0$ in (15.5.3), we recapture the correct formula for the action of g on $\mathbb{C}$.

The consequences of (15.5.3) are far-reaching. First, if we consider g to be a translation, say $g(z) = z + b$, then we find that $g(z + t\mathbf{j}) = (z + b) + t\mathbf{j}$; thus g is just the 'horizontal' translation by b. If $g(z) = az$, then we find that

$$g(z + t\mathbf{j}) = az + |a|t\mathbf{j}.$$

If $|a| = 1$, so that g is a rotation of the complex plane, then g acts on $\mathcal{H}^3$ as a rotation about the vertical axis through the origin. If $a > 0$, so that g acts as a 'stretching' from the origin by a factor a in $\mathbb{C}$, then g also acts as a stretching (from the origin, and by the same factor) in $\mathcal{H}^3$. Of course, the more interesting case is when $g(z) = 1/z$; here

$$g(z + t\mathbf{j}) = \frac{\bar{z} + t\mathbf{j}}{|z|^2 + t^2}.$$

We define the lines in $\mathcal{H}$ to be the 'vertical' semi-circles, and the 'vertical' rays (exactly as in the two-dimensional case; see Figure 15.5.1), and we can define the hyperbolic distance between two points again by a cross-ratio (thinking of the vertical plane through the two points as the complex plane), or by integrating $|dx|/x_3$, where $x = (x_1, x_2, x_3)$, over curves. When all this has been done, we arrive at the following beautiful result.

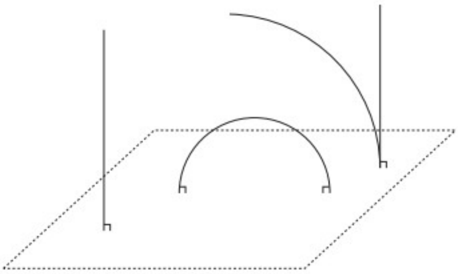

Figure 15.5.1

Theorem 15.5.1 *Every Möbius map acts on hyperbolic space $\mathcal{H}^3$ as a hyperbolic isometry, and every isometry that preserves orientation is a Möbius map.*

15.6 Finite Möbius groups

Finally, we give only the briefest sketch of the ideas behind a proof of Theorem 14.3.2. The aim of this sketch is to give the reader a glimpse of some beautiful interaction between algebra and geometry, and it is far from being complete. First, the Möbius maps (that act on $\mathbb{C}_\infty$) can be extended (either as a composition of reflections and inversions, or in terms of the quaternion algebra) to act on all of $\mathbb{R}^3$. The upper-half $\mathcal{H}^3$ of $\mathbb{R}^3$ with the hyperbolic metric $ds = |dx|/x_3$ is a model of three-dimensional hyperbolic geometry, and the Möbius group is the group of orientation-preserving isometries of this space.

Now let G be a finite Möbius group; then G may be regarded as a finite group of isometries of $\mathcal{H}^3$, so that each point in $\mathcal{H}^3$ has a finite orbit. Take any orbit and let B be the smallest hyperbolic ball that contains the orbit. Analytic arguments show that B is unique, and as the chosen orbit is invariant under G, so is B, and hence (finally) so too is the hyperbolic centre of B. This argument proves that the elements of the finite group G have a common fixed point ζ in $\mathcal{H}^3$. There is now a Möbius map (which acts on all of $\mathbb{R}^3 \cup \{\infty\}$) that converts the upper-half space model of three-dimesional hyperbolic space into the unit ball model (much as there is a Möbius map that takes the upper half-plane to the unit disc). This can be chosen so that ζ is carried to the origin; thus the finite Möbius G is conjugate to a Möbius group G' of hyperbolic isometries that act on the unit ball in $\mathbb{R}^3$ with the extra property that every element of G' fixes 0. It is not difficult to show that every such isometry is a Euclidean rotation of $\mathbb{R}^3$ and the sketch of the proof is complete. □

4 +,-.+ + +/ !+* > ! , 4!!# 7 8
1 2 3#
( )) " &&& '0 #? #! !0 7 8 1 2

Consequently there exist linear transformations $\sigma_1, \cdots, \sigma_n$ each of which induces a perspectivity such that $\sigma\sigma_1 \cdots \sigma_n$ leaves invariant every element in S. The projectivity ρ induced by $\sigma_n^{-1} \cdots \sigma_1^{-1}$ is consequently a product of [at most] n perspectivities and satisfies $X^{\rho} = X^{\sigma}$ for every subspace X of S.

Theorem 1: *A projectivity is a collineation if, and only if, it is a product of* [*at most* 3] *projectivities each of which possesses a full line of fixed points.*

PROOF: The sufficiency of our condition is an immediate consequence of Lemma 2. If conversely σ is a collineation, and if L is a line, then we deduce from Proposition 1 the existence of perspectivities σ_1 and σ_2 such that $P^{\sigma} = P^{\sigma_1\sigma_2}$ for every point P on L. Then $\sigma = \rho\sigma_1\sigma_2$ where the projectivity ρ leaves invariant every point on the line L and where the perspectivities σ_i have at least a line of fixed points each.

Theorem 2: *A projectivity of a linear manifold of finite rank is a collineation if, and only if, it is a product of perspectivities.*

This is an immediate consequence of Lemma 2 and Proposition 1.

For linear manifolds of infinite rank we may give to Proposition 1 a similar interpretation by saying that a projectivity is a collineation if, and only if, it may be "approximated" by products of perspectivities. We leave the details to the topologically interested reader.

III.3. The Second Fundamental Theorem of Projective Geometry

We assume in the course of the present section that A is an F-space of finite rank $r(A) = n$. We shall assume that $1 < n$, since otherwise our investigation would not have any content [and some assertions would cease to be valid]. It is, however, not neccessary to assume that $2 < n$, since we are going to investigate collineations which are *defined* by the fact that they are induced by linear transformations.

It will be convenient to term *simplex* a set of $r(A) + 1$ points no $r(A)$ of which are contained in a hyperplane of A. This is equivalent to saying that any $r(A)$ of the $r(A) + 1$ points of a simplex are independent. If, for instance, A is a plane, then a simplex consists of four points no three of which are collinear.

Lemma 1: *If the points P_i [$i = 0, 1, \cdots, r(A)$] form a simplex, then there exists a basis $b_1, \cdots, b_n$ of A such that $P_i = Fb_i$ for $i = 1, \cdots, n$ and*

$$P_0 = F\sum_{j=1}^{n} b_j.$$

Proof: Let b_0 be any element such that $P_0 = Fb_0$. Since the points $P_1, \cdots, P_n$ are independent, and since n is the finite rank of A, we have $A = \sum_{j=1}^{n} P_j$. Consequently there exist elements b_j in P_j such that $b_0 = b_1 + \cdots + b_n$. If b_i were 0, then b_0 and hence P_0 would be part of the hyperplane $\sum_{i \neq j} P_j$ which is impossible. Consequently $P_i = Fb_i$; and from the independence of the points $P_1, \cdots, P_n$ it follows now that $b_1, \cdots, b_n$ is a basis of A.

Proposition 1: *If P_i' and P_i'' are simplices, then there exists a collineation σ such that $P_i'' = P_n'^{\sigma}$ for $i = 0, 1, \cdots, n$.*

Proof: We infer from Lemma 1 the existence of bases $p_1', \cdots, p_n'$ and $p_1'', \cdots, p_n''$ of A such that $P_i' = Fp_i'$, $P_i'' = Fp_i''$ for $i = 1, \cdots, n$ and $P_0' = F\sum_{i=1}^{n} p_i'$, $P_0'' = F\sum_{i=1}^{n} p_i''$. There exists furthermore one and only one linear transformation σ of A such that $p_i'^{\sigma} = p_i''$ for $i = 1, \cdots, n$; and it is clear that the collineation induced by σ meets all the requirements.

Proposition 2: *The identity is the only collineation leaving invariant every point of a simplex if, and only if, F is a commutative field.*

Proof: Denote by P_i some simplex of A. Then there exists, by Lemma 1, a basis $b_1, \cdots, b_n$ of A such that $P_i = Fb_i$ for $i = 1, \cdots, n$ and $P_0 = F\sum_{i=1}^{n} b_i$.

Assume now that the identity is the only collineation which leaves invariant every point P_i of our simplex. Suppose that $f \neq 0$ is a number in F. Then there exists one and only one linear transformation σ of A such that $b_i^{\sigma} = fb_i$ for $i = 1, \cdots, n$. This linear transformation σ induces a collineation which satisfies $P_i^{\sigma} = P_i$ for $i = 0, 1, \cdots, n$. Hence σ induces the identity projectivity. Consider now some element z in F. Then it follows that

$$F(b_1 + zb_2) = [F(b_1 + zb_2)]^{\sigma} = F(fb_1 + zfb_2) = F(b_1 + f^{-1}zfb_2).$$

Because b_1 and b_2 are independent, it follows now that $z = f^{-1}zf$ or $fz = zf$; and this is true for every f and every z. Hence F is commutative.

Assume now conversely that the field F be commutative. Consider some collineation σ which leaves invariant each of the points P_i of our simplex. This collineation σ is induced by a linear transformation which we denote by σ too. From $P_i^{\sigma} = P_i$ for $i = 0, 1, \cdots, n$ we deduce the existence of numbers $f_i \neq 0$ in F such that

$$b_i^\sigma = f_i b_i \text{ for } i = 1, \cdots, n, \quad \left[\sum_{i=1}^{n} b_i\right]^\sigma = f_0 \sum_{i=1}^{n} b_i.$$

But this implies

$$\sum_{i=1}^{n} f_0 b_i = \sum_{i=1}^{n} b_i^\sigma = \sum_{i=1}^{n} f_i b_i;$$

and we infer from the independence of the b_i that $f_0 = f_1 = \cdots = f_n$. If a is any element in A, then

$$a^\sigma = \left[\sum_{i=1}^{n} a_i b_i\right]^\sigma = \sum_{i=1}^{n} a_i f_0 b_i = f_0 \sum_{i=1}^{n} a_i b_i = f_0 a$$

since F is commutative. Hence σ leaves invariant every point in A so that σ is the identity projectivity.

Remark 1: Mapping the point (x_0,x_1,x_2) of the complex plane upon its conjugate complex point $(\bar{x}_0,\bar{x}_1,\bar{x}_2)$ we obtain a semi-linear transformation which leaves invariant every point of the simplex (1,1,1), (1,0,0), (0,1,0), (0,0,1); but which does not leave invariant the point represented by $(i,1,0)$ where $i^2 = -1$. Thus the Proposition 2 ceases to be true, if we omit the restriction to collineations.

Remark 2: If the rank of A is at least 3, then we may consider the group of all projectivities which leave invariant every point of some given simplex. This group may be seen [by application of the First Fundamental Theorem of Projective Geometry of III.1] to be isomorphic to the group of automorphisms of the field F. See III.2, Survey of the groups of a linear manifold.

Remark 3: If F is not commutative, then there exist collineations, not the identity, which leave invariant every point in a given simplex [Proposition 2]. It follows from III.2, Theorem 2 that this collineation is a product of perspectivities. Thus a product of perspectivities may leave invariant every point in a simplex without being the identity.

The Second Fundamental Theorem of Projective Geometry may now be stated as follows:

Every simplex may be mapped upon every other simplex by one and only one collineation if, and only if, the field of coordinates is commutative.

The proof is an immediate consequence of Propositions 1 and 2.

Remark 4: It is possible to extend the results of this section to linear manifolds of infinite rank. If one wants to do this, one has to investigate the effect of a linear transformation on the subspaces of finite rank. Furthermore one has to find a substitute for the concept of simplex which in its present form looses its meaning in the case of infinite rank. We leave this discussion to the interested reader.

APPENDIX II

The Theorem of Pappus

In the Second Fundamental Theorem of Projective Geometry [§ 3] we have met a first instance where the commutativity of the field F [of coordinates] appears important. In the present appendix we give Hilbert's celebrated characterization of commutativity of F by means of geometric configurations.

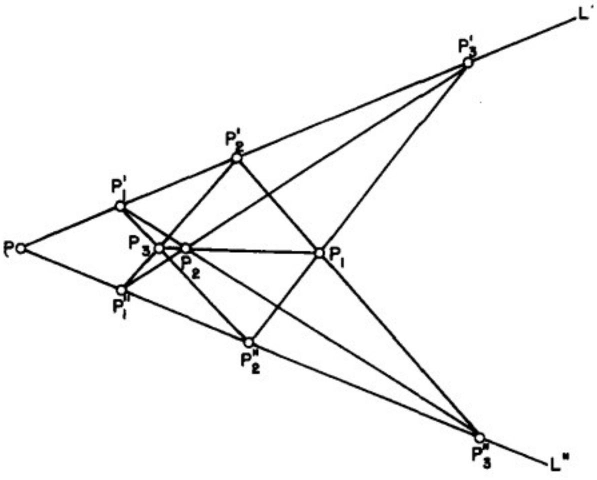

Fig. 4

We are considering configurations of ten coplanar points which are obtained as follows. There are given two distinct lines L' and L'' meeting in a point P and on each of the lines three distinct points P'_i on L' and and P''_i on L'' such that $P \neq P'_i$ and $P \neq P''_i$ for $i = 1,2,3$. One verifies easily that the three intersections

$$P_i = (P'_j + P''_k) \cap (P''_j + P'_k)$$

where i,j,k is a permutation of 1,2,3 are well determined points; and we say that *this configuration* $[P_i,P'_i,P''_i]$ *has the Property of Pappus, if* P_1,P_2,P_3 *are collinear points.*

We select at random an element a such that $P = Fa$. From

$$P \leqslant L' = P_1' + P_2'$$

we infer the existence of elements b and b' in P_1' and P_2' respectively such that $a = b' - b$; and b and b' are different from 0, since P is different from P_i'. Hence $P_1' = Fb$ and $P_2' = F(a + b)$. Since P_3' is on $L' = P + P_1'$, we have $P_3' = F(a + d'b)$ with $d' \neq 0, 1$, if we only note that P_3' is different from P, P_1', P_2'. Likewise we find an element c in A and an element $d'' \neq 0, 1$ in F such that $P_1'' = Fc$, $P_2'' = F(a + c)$, $P_3'' = F(a + d''c)$.

The configuration $[P_i, P_i', P_i'']$ *has the Property of Pappus if, and only if,* $d'd'' = d''d'$.

PROOF: We note first that a,b,c are three independent elements. Thus $F(a + b + c)$ is a point; and it is clear that this point is on the lines $P_2' + P_1'' = F(a + b) + Fc$ and $P_2'' + P_1' = F(a + c) + Fb$. But these lines meet in P_3; and thus it follows that $P_3 = F(a + b + c)$. Similarly one sees that $P_2 = F(a + d'b + d''c)$. Finally we note that

$$(d' - 1)^{-1}d' + (d'' - 1)^{-1} = (d' - 1)^{-1} + (d'' - 1)^{-1}d''$$

and that therefore

$$\begin{aligned}(d' - 1)^{-1}d'(a + b) &+ (d'' - 1)^{-1}(a + d''c)\\ &= (d' - 1)^{-1}(a + d'b) + (d'' - 1)^{-1}d''(a + c)\end{aligned}$$

is the intersection P_1 of the lines $P_2' + P_3'' = F(a + b) + F(a + d''c)$ and $P_2'' + P_3' = F(a + c) + F(a + d'b)$. Since P_2 and P_3 are certainly different points, it follows that our configuration has the Property of Pappus if, and only if, $P_1 \leqslant P_2 + P_3$; and this is equivalent to the property:

There exist numbers h,k in F such that

$$\begin{aligned}(d' - 1)^{-1}d'(a + b) &+ (d''- 1)^{-1}(a + d''c)\\ &= h(a + d'b + d''c) + k(a + b + c).\end{aligned} \tag{1}$$

From the independence of a,b,c it follows that the numbers h,k satisfy equation (1) if, and only if, they satisfy the following equations:

$$\left\{\begin{aligned} h + k &= (d' - 1)^{-1}d' + (d'' - 1)^{-1},\\ hd' + k &= (d' - 1)^{-1}d',\\ hd'' + k &= (d''- 1)^{-1}d''.\end{aligned}\right. \tag{2}$$

Thus the existence of solutions h,k of (2) is necessary and sufficient for the validity of the Property of Pappus.

If there exist numbers h,k which satisfy (2), then subtraction of the first equation (2) from the other two equations (2) gives:

$$h(d' - 1) = - (d'' - 1)^{-1}, \qquad h(d'' - 1) = 1 - (d' - 1)^{-1}d' = - (d' - 1)^{-1}.$$

Consequently

$$-(d''-1)^{-1}(d'-1)^{-1} = h = -(d'-1)^{-1}(d''-1)^{-1}$$

or $(d'-1)(d''-1) = (d''-1)(d'-1)$; and this implies $d'd'' = d''d'$.

If conversely $d'd'' = d''d'$, then one sees easily that

$$h = -(d''-1)^{-1}(d'-1)^{-1}, \qquad k = (d'-1)^{-1}d''(d''-1)^{-1}$$

are solutions of the equations (2). Thus we have shown that $d'd'' = d''d'$ is necessary and sufficient for the validity of the Property of Pappus, as we claimed.

Finally we say that the *Postulate of Pappus* is satisfied by the linear manifold (F,A) [of rank at least 3], if every configuration $[P_i,P_i',P_i'']$ has the Property of Pappus. If a,b,c are three independent elements, and if neither of the numbers d',d'' is 0 or 1, then the points $P_1' = Fb$, $P_2' = F(a+b)$, $P_3' = F(a+d'b)$ and $P_1'' = Fc$, $P_2'' = F(a+c)$, $P_3'' = F(a+d''c)$ lead as before to a configuration of the type considered here. Consequently we have the fundamental result:

The Postulate of Pappus is satisfied by the linear manifold (F,A) of rank not less than 3 if, and only if, the field F is commutative.

III.4. The Projective Geometry of a Line in Space; Cross Ratios

A line [= linear manifold of rank 2] has no geometrical structure, if considered as an isolated or absolute phenomenon, since then it is nothing but a set of points with the number of points on the line as the only invariant [see III.1, Remark 1 and the Projective Structure Theorem]. But the line receives a definite geometrical structure, if imbedded into a linear manifold of higher rank.

If (F,A) is a linear manifold of rank not less than 3, and if L' and L'' are lines in (F,A), then geometrical equivalence of a configuration on L' with one on L'' in the absolute sense is defined by any one-to-one correspondence between the points on L' and those on L'' whereas geometrical equivalence relative to the containing manifold (F,A) is defined by those correspondences which are induced by auto-projectivities of (F,A). Thus there arises the problem of characterizing those correspondences which are induced by projectivities. To do this we have to introduce the concept of cross ratio; and we emphasize that this concept, so central in classical discussions of projective geometry, is needed only for a treatment of the rather special problem under consideration.

DEFINITION 1: *Suppose that P,Q,R,S are four distinct points on the line*

L. Then the number s in F belongs to the cross ratio $\begin{bmatrix} P & Q \\ S & R \end{bmatrix}$ *if, and only if, there exist elements p,q in A such that*

$$P = Fp, \qquad Q = Fq, \qquad R = F(p + q), \qquad S = F(p + sq).$$

The justification for the "square" symbol indicating the cross ratio [which is due to F. Levi] will be found in the symmetry rules that we are going to derive later. The justification for defining the cross ratio not as a number but as set of numbers will be found in the following proposition.

Proposition 1: *The cross ratio of four distinct collinear points is a complete class of conjugate numbers.*

Here we term as usual the numbers h and k in F *conjugate*, if there exists a number $c \neq 0$ in F such that $h = c^{-1}kc$; and we shall designate the totality of numbers conjugate to the number k in F by the symbol $\langle k \rangle$.

PROOF: Consider four distinct collinear points P,Q,R,S. Then

$$P + Q = P + R = P + S = Q + R = Q + S = R + S = L$$

is a line in our linear manifold (F,A). We have to show first that there exists at least one number in the cross ratio $\begin{bmatrix} P & Q \\ S & R \end{bmatrix}$.

To do this consider any element r such that $R = Fr$. Since r belongs to $P \oplus Q$, there exist uniquely determined elements p,q in P and Q respectively such that $r = p + q$; and p and q are different from 0, since R is different from Q and P. Hence $P = Fp$, $Q = Fq$, $R = F(p + q)$. Since p is on $Q + S = P + Q$, there exist elements q' and w in Q and S respectively such that $p = q' + w$; and q' and w are different from 0, since P is different from S and Q. Hence $Q = Fq'$ and $S = Fw$. From $Fq = Q = Fq'$ we deduce the existence of a number $s \neq 0$ such that $q' = - sq$; and now it follows that $S = Fw = F(p + sq)$. Hence s belongs to the cross ratio $\begin{bmatrix} P & Q \\ S & R \end{bmatrix}$; and we note that $s \neq 0,\ 1$.

Suppose now that s is some definite number in the cross ratio. Then there exist elements p,q in A such that

$$P = Fp, \qquad Q = Fq, \qquad R = F(p + q), \qquad S = F(p + sq).$$

Consider now a number s' in F conjugate to s. Then there exists a number $t \neq 0$ in F such that $s = t^{-1}s't$; and we find

$$P = F(tp), \qquad Q = F(tq), \qquad R = F(tp + tq), \qquad S = F(tp + s'(tq))$$

so that s' belongs to the cross ratio $\begin{bmatrix} P & Q \\ S & R \end{bmatrix}$. Assume conversely that s' belongs to the cross ratio. Then there exist elements p',q' in A such that

$$P = Fp', \qquad Q = Fq', \qquad R = F(p' + q'), \qquad S = F(p' + s'q').$$

There exist numbers h,k,m,n in F neither of which is 0 such that

$$p' = hp, \qquad q' = kq, \qquad p' + q' = m(p + q), \qquad p' + s'q' = n(p + sq).$$

Hence $m(p + q) = hp + kq$ and $n(p + sq) = hp + s'kq$. From the independence of p and q we deduce now the validity of $m = h = k$ and $n = h$, $ns = s'k$ or $s' = nsn^{-1}$ so that s' is conjugate to s. Thus the cross ratio $\begin{bmatrix} P & Q \\ S & R \end{bmatrix}$ is the totality of numbers conjugate to s, as we intended to prove.

Remark 1: We have shown incidentally the obvious fact that 0, 1 do not belong to the cross ratio $\begin{bmatrix} P & Q \\ S & R \end{bmatrix}$, if we insist that the points P,Q,R,S are all distinct. If we would permit $S = P$, then our definition would give the value 0 to the cross ratio; and if we permit $S = R$, then our definition would give the value 1 to the cross ratio. These two possibilities we shall consequently allow occasionally. But $S = Q$ cannot be admitted unless we want to give the value ∞ to the cross ratio; and the points P,Q,R have to be considered distinct in any case. By and large it will prove most convenient to insist that P,Q,R,S are four distinct collinear points.

Remark 2: From $P = Fp$, $Q = Fq$, $R = F(p + q)$, $S = F(p + sq)$ we deduce also $S = F(q + s^{-1}p)$, provided $s \neq 0$; and this shows that $\begin{bmatrix} Q & P \\ S & R \end{bmatrix}$ is the totality of inverses of elements in $\begin{bmatrix} P & Q \\ S & R \end{bmatrix}$. This fact we may express in the form

$$\begin{bmatrix} P & Q \\ S & R \end{bmatrix}^{-1} = \begin{bmatrix} Q & P \\ S & R \end{bmatrix};$$

and one verifies similarly that

$$\begin{bmatrix} P & Q \\ S & R \end{bmatrix}^{-1} = \begin{bmatrix} P & Q \\ R & S \end{bmatrix}.$$

Next we note that $1 - t^{-1}st = t^{-1}(1 - s)t$. Thus $1 - s$ ranges with s over a whole class of conjugate numbers; and with every class C of conjugate numbers $1 - C$ is a well defined class of conjugate numbers in F. If we note now that

$$S = F(p + sq), \qquad P = F(-p), \qquad Q = F[(p + sq) + (-p)],$$
$$R = F[(p + sq) + (1 - s)(-p)],$$

then we find that

$$\begin{bmatrix} S & P \\ R & Q \end{bmatrix} = 1 - \begin{bmatrix} P & Q \\ S & R \end{bmatrix}.$$

The values of the cross ratios which we obtain by all the 24 possible permutations of the four points P,Q,R,S are now easily computed. In particular one checks the validity of *the symmetry formulas*

$$\begin{bmatrix} P & Q \\ S & R \end{bmatrix} = \begin{bmatrix} S & R \\ P & Q \end{bmatrix} = \begin{bmatrix} Q & P \\ R & S \end{bmatrix} = \begin{bmatrix} R & S \\ Q & P \end{bmatrix}.$$

REMARK 3: One verifies without difficulty that every class of conjugate numbers in F, not 0 or 1, is the cross ratio of some quadruplet of collinear points.

The following formulas will prove useful in our discussion; and they may also remind the reader of known expressions for cross ratios in classical instances.

Lemma 1: *Suppose that p and q are two independent elements in A.*

(*a*) *If u,v,w are three distinct numbers in F, then*

$$\begin{bmatrix} Fq & F(p+uq) \\ F(p+wq) & F(p+vq) \end{bmatrix} = \langle (v-u)(w-u)^{-1} \rangle.$$

(*b*) *If h,k,m,n are four distinct numbers in F, then*

$$\begin{bmatrix} F(p+hq) & F(p+kq) \\ F(p+nq) & F(p+mq) \end{bmatrix} = \langle (h-n)(k-n)^{-1}(k-m)(h-m)^{-1} \rangle.$$

PROOF: Formula (*a*) one verifies by realising that

$$\begin{aligned} F(p+vq) &= F[(v-u)q + (p+uq)], \\ F(p+wq) &= F[(v-u)q + (v-u)(w-u)^{-1}(p+uq)]. \end{aligned}$$

To prove (*b*) we verify first the following equality.

(*c*) If x,y,z are distinct numbers in F, then

$$F(p+xq) = F[-(p+yq) + (y-x)(z-x)^{-1}(p+zq)].$$

This identity follows from the following identities:

$$\begin{aligned} &(y-x)(z-x)^{-1} - 1 = (y-x-z+x)(z-x)^{-1} = (y-z)(z-x)^{-1}, \\ &(y-x)(z-x)^{-1}z - y = (y-x)(z-x)^{-1}z - z - (y-z) \\ &\quad = [(y-x)(z-x)^{-1} - 1]z - (y-z) \\ &\quad = (y-z)(z-x)^{-1}z - (y-z) = (y-z)(z-x)^{-1}[z-(z-x)] \\ &\quad = (y-z)(z-x)^{-1}x. \end{aligned}$$

From (*c*) one deduces (*b*) by remarking that

$$F(p+mq) = F[-(p+hq) + (h-m)(k-m)^{-1}(p+kq)]$$

and

$$\begin{aligned} F(p+nq) &= F[-(p+hq) + (h-n)(k-n)^{-1}(p+kq)] \\ &= F[-(p+hq) + (h-n)(k-n)^{-1}(k-m)(h-m)^{-1}(h-m)(k-m)^{-1}(p+kq)]. \end{aligned}$$

So far our discussion of cross ratios has been very similar to the classical discussion of this concept, in spite of the fact that the cross ratio is a class of conjugate numbers. We come now to a result which is rather different from classical results. If P,Q,R are three distinct collinear points, and if $\langle s \rangle$ is a class of conjugate numbers in F, then we may always find a point S on the line $P+Q$ such that $\begin{bmatrix} P & Q \\ S & R \end{bmatrix} = \langle s \rangle$, since we can always find elements p,q in A such that $P = Fp$, $Q = Fq$, $R = F(p+q)$, and since we may then choose $S = F(p+sq)$. But the point S is, in general, not uniquely determined by P,Q,R and the cross ratio, as may be seen from the following proposition.

Proposition 2: *The following properties of the four distinct points P,Q,R,S on the line L are equivalent.*

(I) *If X is a point on L such that $\begin{bmatrix} P & Q \\ S & R \end{bmatrix} = \begin{bmatrix} P & Q \\ X & R \end{bmatrix}$, then $S = X$.*

(II) $\begin{bmatrix} P & Q \\ S & R \end{bmatrix}$ *is a one element set.*

(III) $\begin{bmatrix} P & Q \\ S & R \end{bmatrix}$ *is part of the center of F.*

Proof: The equivalence of properties (III) and (II) is evident once one remembers that an element in F belongs to the center of F if, and only if, it commutes with every element in F. To prove the equivalence of (I) and (II) we remark first that there exist elements p,q in A and an element s in F such that

$$P = Fp, \quad Q = Fq, \quad R = F(p+q), \quad S = F(p+sq)\,; \begin{bmatrix} P & Q \\ S & R \end{bmatrix} = \langle s \rangle.$$

Assume now the validity of (I) and consider some element $f \neq 0$ in F. If we let $X = F(p + f^{-1}sfq)$, then

$$\begin{bmatrix} P & Q \\ X & R \end{bmatrix} = \langle s \rangle = \begin{bmatrix} P & Q \\ S & R \end{bmatrix};$$

and it follows from (I) that $F(p + f^{-1}sfq) = X = S = F(p+sq)$; and this implies $f^{-1}sf = s$ because of the independence of the elements p,q. Thus $\langle s \rangle$ is a one element set; and we have shown that (II) is a consequence of (I). If conversely (II) is satisfied, and if X is a point on L such that $\begin{bmatrix} P & Q \\ S & R \end{bmatrix} = \begin{bmatrix} P & Q \\ X & R \end{bmatrix}$, then $X = F(p+xq)$, since X is different from Q. Hence $\langle s \rangle = \langle x \rangle$; and this implies $s = x$ by (II) so that $S = X$. Thus (I) is a consequence of (II); and this completes the proof.

Remark 4: It is an immediate consequence of Proposition 2 that commutativity of the field F is necessary and sufficient for the fact that three collinear points and the cross ratio determine uniquely the fourth point; see Theorem 2 below.

Remark 5: If F does not have characteristic 2, then -1 is a center element different from 0 and 1. It is customary to term the four distinct collinear points P,Q,R,S *a harmonic set,* if $\begin{bmatrix} P & Q \\ S & R \end{bmatrix} = -1$. In this case the four points may be represented in the form:

$$P = Fp, \qquad Q = Fq, \qquad R = F(p+q), \qquad S = F(p-q).$$

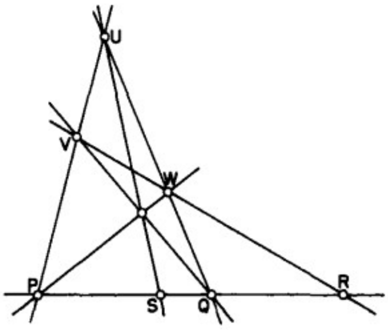

Fig. 5

It is now easy to verify that the following property is necessary and sufficient for the four points P,Q,R,S to form a harmonic set:

If U is a point not on the line $P+Q$, if V and W are points such that

$$\begin{aligned} U+V &= V+P = P+U, \\ U+W &= W+Q = Q+U, \\ V+W &= W+R = R+V, \end{aligned}$$

then the three lines $U+S$, $W+P$, $V+Q$ meet in a point.

Since we do not want to exclude fields of characteristic 2 from our discussion, it will be impossible [and usually unnecessary] to give harmonic sets the preferential treatment so often accorded to them.

Theorem 1: *Suppose that P,Q,R,S and P',Q',R',S' are quadruples of distinct collinear points.*

(*a*) *There exists a projectivity π of the linear manifold (F,A) such that $P^\pi = P'$, $Q^\pi = Q'$, $R^\pi = R'$, $S^\pi = S'$ if, and only if, there exists an automorphism α of the field F such that* $\begin{bmatrix} P & Q \\ S & R \end{bmatrix}^\alpha = \begin{bmatrix} P' & Q' \\ S' & R' \end{bmatrix}$.

(*b*) *There exists a collineation σ of (F,A) such that $P^\sigma = P'$, $Q^\sigma = Q'$, $R^\sigma = R'$, $S^\sigma = S'$ if, and only if,* $\begin{bmatrix} P & Q \\ S & R \end{bmatrix} = \begin{bmatrix} P' & Q' \\ S' & R' \end{bmatrix}$.

Proof: We note first that we may represent the points of the first quadruple in the form:

$$P = Fp, \quad Q = Fq, \quad R = F(p+q), \quad S = F(p+sq), \quad \begin{bmatrix} P & Q \\ S & R \end{bmatrix} = \langle s \rangle.$$

It follows from the First Fundamental Theorem of Projective Geometry

[III.1] that every projectivity π is induced by a semi-linear transformation α. But then

$$P^\pi = Fp^\alpha, \qquad Q^\pi = Fq^\alpha, \qquad R^\pi = F(p^\alpha + q^\alpha), \qquad S^\pi = F(p^\alpha + s^\alpha q^\alpha),$$
$$\begin{bmatrix} P^\pi & Q^\pi \\ S^\pi & R^\pi \end{bmatrix} = \langle s^\alpha \rangle = \langle s \rangle^\alpha = \begin{bmatrix} P & Q \\ S & R \end{bmatrix}^\alpha;$$

and this proves the necessity of our condition.

If there exists an automorphism α such that $\begin{bmatrix} P' & Q' \\ S' & R' \end{bmatrix} = \begin{bmatrix} P & Q \\ S & R \end{bmatrix}^\alpha$, then $\begin{bmatrix} P' & Q' \\ S' & R' \end{bmatrix} = \langle s^\alpha \rangle$. Consequently there exist elements p',q' in A such that

$$P' = Fp',\ Q' = Fq',\ R' = F(p' + q'),\ S' = F(p' + s^\alpha q').$$

Since p,q as well as p',q' are pairs of independent elements, one constructs easily a semi-linear transformation (α,β) of (F,A) such that $p^\beta = p'$, $q^\beta = q'$, $(xy)^\beta = x^\alpha y^\beta$ for x in F and y in A [use the fact that p,q as well as p',q' are part of some basis and that any two bases contain the same number of elements; see Chapter II]. This semi-linear transformation induces a projectivity π which maps P,Q,R,S upon P',Q',R',S' respectively. This completes the proof of (*a*). The proof of (*b*) is quite similar to the proof of (*a*), once one remembers that collineations are induced by linear transformations [$\alpha = 1$ in the above notation].

The preceding theorem shows that cross ratios characterize in a way the projective structure of quadruples of collinear points. Our next objective is the characterization of those mappings which preserve some or all cross ratios. It is a fairly obvious consequence of Theorem 1 that it will suffice to consider permutations of the points on some line which possess at least three fixed points.

Suppose now that σ is a permutation of the points on the line L and that the three distinct points U, V, W are fixed points of σ. Then we show as usual the existence of [independent] elements p,q such that $U = Fp$, $V = Fq$, $W = F(p + q)$. Every point, not V, on L may be represented in one and only one way in the form $F(p + xq)$ for x in F so that the elements p,q define a one-to-one correspondence between the points, not V, on L and the elements in F. Consequently there exists a well determined permutation σ' of the numbers in F such that

$$[F(p + xq)]^\sigma = F(p + x^{\sigma'} q).$$

It is clear that $0 = 0^{\sigma'}$, $1 = 1^{\sigma'}$; and that every permutation σ' of the numbers in F which leaves invariant 0 and 1 comes from a permutation of the points on L with fixed points U, V, W.

Proposition 3: *Assume that p, q are two independent elements in A, that σ is a permutation of the points on the line $Fp + Fq = L$ with fixed points Fp, Fq, $F(p + q)$ and that the corresponding permutation of the numbers in F is denoted by σ too so that*

$$[F(p + xq)]^\sigma = F(p + x^\sigma q) \text{ for every } x \text{ in } F.$$

If z is a number neither 0 nor 1 in the center of F, then the following properties of z and σ are equivalent.

(I) *If P, Q, R, S are four distinct points on L such that $\begin{bmatrix} P & Q \\ S & R \end{bmatrix} = z$, then $\begin{bmatrix} P^\sigma & Q^\sigma \\ S^\sigma & R^\sigma \end{bmatrix} = z$.*

(II) *The permutation σ of F satisfies $z = z^\sigma$ and $x^\sigma + y^\sigma = (x + y)^\sigma$, $(xyx)^\sigma = x^\sigma y^\sigma x^\sigma$ for every x, y in F.*

(III) *σ is an automorphism or anti-automorphism of F which leaves invariant z.*

We note that the center element z is the only element in its class of conjugate elements and that we may consequently identify z with its class $\langle z \rangle$. Furthermore we remind the reader that an *anti-automorphism* of the field F is a permutation of the elements in F which preserves addition, but inverts the order of multiplication $[(xy)^\sigma = y^\sigma x^\sigma]$.

PROOF: Assume first the validity of (I). Then $z^\sigma = z$, since

$$\begin{bmatrix} Fp & Fq \\ F(p + zq) & F(p + q) \end{bmatrix} = z,$$

and since therefore

$$\langle z^\sigma \rangle = \begin{bmatrix} Fp & Fq \\ F(p + z^\sigma q) & F(p + q) \end{bmatrix} = \begin{bmatrix} [Fp]^\sigma & [Fq]^\sigma \\ [F(p + zq)]^\sigma & [F(p + q)]^\sigma \end{bmatrix} = z.$$

Next we deduce from Lemma 1, (a) that we have for every $x \neq 0$ in F

$$\begin{bmatrix} Fq & Fp \\ F(p + xq) & F(p + zxq) \end{bmatrix} = z;$$

and consequently it follows from (I) and Lemma 1, (a) as before that

$$z = \begin{bmatrix} Fq & Fp \\ F(p + x^\sigma q) & F(p + zx)^\sigma q) \end{bmatrix} = \langle (zx)^\sigma (x^\sigma)^{-1} \rangle.$$

But this implies clearly

$$(zx)^\sigma = zx^\sigma \text{ for every } x \text{ in } F. \tag{I.1}$$

[Note that our proof holds for $x \neq 0$ only; but our formula is true for $x = 0$ too, since $0 = 0^\sigma$.]

If u and v are two different numbers in F, then we let $w = z^{-1}(v - u) + u$ which is possible, since $z \neq 0$. Since $z \neq 1$, w is different from u and v. Since z is in the center of F, we deduce from Lemma 1, (*a*) that

$$\begin{bmatrix} Fq & F(p + uq) \\ F(p + wq) & F(p + vq) \end{bmatrix} = \langle (v - u)(w - u)^{-1} \rangle = z.$$

As before we deduce from (i) and Lemma 1, (*a*) [which may be applied, since the permutation σ maps distinct elements upon distinct elements] that

$$z = \begin{bmatrix} Fq & F(p + u^\sigma q) \\ F(p + w^\sigma q) & F(p + v^\sigma q) \end{bmatrix} = \langle (v^\sigma - u^\sigma)(w^\sigma - u^\sigma)^{-1} \rangle.$$

Consequently

$$z(w^\sigma - u^\sigma) = v^\sigma - u^\sigma \qquad \text{or} \qquad zw^\sigma = v^\sigma - u^\sigma + zu^\sigma.$$

From (i,1) and the definition of w we deduce that

$$zw^\sigma = (zw)^\sigma = (v - u + zu)^\sigma;$$

and thus we have shown the validity of

(i.2) $\qquad (v - u + zu)^\sigma = v^\sigma - u^\sigma + zu^\sigma$ for every u,v in F.

[Actually we have verified this formula for $u \neq v$ only; but in case $u = v$, our formula is equivalent with (i.1).]

If x and y are numbers in F, then we may form $u = (z - 1)^{-1}y$, since $z \neq 1$. It follows from (i.2) that

$$(x + y)^\sigma = (x + zu - u)^\sigma = x^\sigma - u^\sigma + zu^\sigma = x^\sigma + (z - 1)u^\sigma.$$

Letting $v = 0$ in (i.2) we obtain $[(z - 1)u]^\sigma = (z - 1)u^\sigma$; and now it follows that

$$x^\sigma + (z - 1)u^\sigma = x^\sigma + [(z - 1)u]^\sigma = x^\sigma + y^\sigma.$$

Thus we have shown:

(i.3) $\qquad (x + y)^\sigma = x^\sigma + y^\sigma$ for every x and y in F.

Next we prove that

(i.4) $\qquad [k(zk - h)^{-1}h]^\sigma = k^\sigma(zk^\sigma - h^\sigma)^{-1}h^\sigma, \qquad$ provided $h \neq zk$.

This formula is trivially true, if k or $h = 0$. Thus we assume that neither h nor k is 0. If $h = k$, then (i.4) is a consequence of $(z - 1)u^\sigma = [(z - 1)u]^\sigma$ which equation is a consequence of (i.1) and (i.3). Thus we assume that $h \neq k$. Now we let $m = (z - 1)k(zk - h)^{-1}h$. If h were equal to m, then we would have $(z-1)k = zk - h$ or $k = h$ which is impossible; and $k \neq m$ is shown likewise. Now we infer from Lemma 1 (*b*) that

$$\begin{bmatrix} F(p+hq) & F(p+kq) \\ Fp & F(p+mq) \end{bmatrix} = \langle hk^{-1}(k-m)(h-m)^{-1} \rangle.$$

But $k^{-1}(k-m) = 1 - k^{-1}m = 1 - (z-1)(zk-h)^{-1}h$

$$= (zk-h)^{-1}[zk - h - (z-1)h] = (zk-h)^{-1}z(k-h)$$

and consequently

$$\begin{aligned} z(h-m) &= z[1-(z-1)k(zk-h)^{-1}]h = z[(zk-h)-(z-1)k](zk-h)^{-1}h \\ &= z[k-h](zk-h)^{-1}h = (zk-h+h-zh)(zk-h)^{-1}h \\ &= h + (1-z)h(zk-h)^{-1}h = h[1+(1-z)(zk-h)^{-1}h] \\ &= h(zk-h)^{-1}[zk-h+(1-z)h] = h(zk-h)^{-1}z(k-h) \\ &= hk^{-1}(k-m) \end{aligned}$$

or

$$\begin{bmatrix} F(p+hq) & F(p+kq) \\ Fp & F(p+mq) \end{bmatrix} = z.$$

Now we apply (I) and Lemma 1, (*b*) to find that

$$h^\sigma(k^\sigma)^{-1}(k^\sigma - m^\sigma)(h^\sigma - m^\sigma)^{-1} = z \quad \text{or} \quad h^\sigma(k^\sigma)^{-1}(k^\sigma - m^\sigma) = z(h^\sigma - m^\sigma).$$

By a direct computation it follows that $m^\sigma = (z-1)k^\sigma(zk^\sigma - h^\sigma)^{-1}h^\sigma$; and now (I.4) is an immediate consequence of the formula $[(z-1)u]^\sigma = (z-1)u^\sigma$ which we have used before and which is contained in (I.1) and (I.3).

(I.5) $\qquad (ab^{-1}a)^\sigma = a^\sigma(b^\sigma)^{-1}a^\sigma$ for $b \neq 0$ and a in F.

We let $k = a$ and $h = zk - b$. Then $h \neq zk$, since $b \neq 0$. Thus we may apply (I.4) and find that [$zk - h = b$ and therefore]

$$\begin{aligned} [ab^{-1}(za-b)]^\sigma &= [k(zk-h)^{-1}h]^\sigma = k^\sigma(zk^\sigma - h^\sigma)^{-1}h^\sigma \\ &= k^\sigma[(zk-h)^\sigma]^{-1}h^\sigma = a^\sigma(b^\sigma)^{-1}(za-b)^\sigma \end{aligned} \qquad \text{by (I.1) and (I.3).}$$

From (I.3) we deduce now $(ab^{-1}za)^\sigma = a^\sigma(b^\sigma)^{-1}(za)^\sigma$; and (I.5) is a consequence of (I.1).

If we let in (I.5) in particular $a = 1$ and remember that $1^\sigma = 1$, then we find that $(b^{-1})^\sigma = (b^\sigma)^{-1}$; and now one deduces from (I.5) the validity of $(aba)^\sigma = a^\sigma b^\sigma a^\sigma$. Thus (II) is a consequence of (I).

Assume now the validity of (II). Considering a and $b = a^{-1}$ we find that $a^\sigma = a^\sigma(a^{-1})^\sigma a^\sigma$ or $(a^{-1})^\sigma = (a^\sigma)^{-1}$. Letting $b = 1$ and remembering $1^\sigma = 1$ we find that $(a^2)^\sigma = (a^\sigma)^2$; and letting $a = x + y$ we deduce from this last formula:

$$\begin{aligned} (a^\sigma)^2 + a^\sigma b^\sigma + b^\sigma a^\sigma + (b^\sigma)^2 &= [(a+b)^\sigma]^2 = [(a+b)^2]^\sigma \\ &= [a^2\, b + ab + a + b^2]^\sigma = (a^2)^\sigma + (ab)^\sigma + (ba)^\sigma + (b^2)^\sigma \\ &= (a^\sigma)^2 + (ab)^\sigma + (ba)^\sigma + (b^\sigma)^2 \end{aligned}$$

or

$$a^\sigma b^\sigma + b^\sigma a^\sigma = (ab)^\sigma + (ba)^\sigma.$$

Next we note that

$$\begin{aligned}(ba)^\sigma &= [(ab)(ab)^{-1}(ba)]^\sigma = [a(b(ab)^{-1}b)a]^\sigma = a^\sigma[b(a)b^{-1}b]^\sigma a^\sigma \\ &= a^\sigma b^\sigma[(ab)^{-1}]^\sigma b^\sigma a^\sigma = a^\sigma b^\sigma[(ab)^\sigma]^{-1}b^\sigma a^\sigma.\end{aligned}$$

Hence

$$a^\sigma b^\sigma + b^\sigma a^\sigma = (ab)^\sigma + (ba)^\sigma = (ab)^\sigma + a^\sigma b^\sigma[(ab)^\sigma]^{-1}b^\sigma a^\sigma$$

and consequently

$$[(ab)^\sigma - a^\sigma b^\sigma][(ab)^\sigma]^{-1}[(ab)^\sigma - b^\sigma a^\sigma] = 0$$

so that we have shown:

(II.1) If a,b are elements in F, then $(ab)^\sigma = \begin{cases} a^\sigma b^\sigma \\ b^\sigma a^\sigma \end{cases}$.

Next we prove:

(II.2) If a,b,c are elements in F, then

(*a*) either $(ab)^\sigma = a^\sigma b^\sigma$ and $(ac)^\sigma = a^\sigma c^\sigma$ or else $(ab)^\sigma = b^\sigma a^\sigma$ and $(ac)^\sigma = c^\sigma a^\sigma$; and

(*b*) either $(ba)^\sigma = b^\sigma a^\sigma$ and $(ca)^\sigma = c^\sigma a^\sigma$ or else $(ba)^\sigma = a^\sigma b^\sigma$ and $(ca)^\sigma = a^\sigma c^\sigma$.

Suppose (*a*) were false. Then we may assume without loss in generality that $(ab)^\sigma \neq a^\sigma b^\sigma$. It follows from (II.1) that $b^\sigma a^\sigma = (ab)^\sigma \neq a^\sigma b^\sigma$; and this implies because of the invalidity of (*a*) that $(ac)^\sigma \neq c^\sigma a^\sigma$. Hence $a^\sigma c^\sigma = (ac)^\sigma$ by (II.1). Likewise

$$\begin{aligned}b^\sigma a^\sigma + a^\sigma c^\sigma &= (ab)^\sigma + (ac)^\sigma = (ab + ac)^\sigma = [a(b + c)]^\sigma \\ &= \begin{cases} a^\sigma(b + c)^\sigma = a^\sigma b^\sigma + a^\sigma c^\sigma \\ (b + c)^\sigma a^\sigma = b^\sigma a^\sigma + c^\sigma a^\sigma \end{cases}\end{aligned}$$

and both these possibilities lead to a contradiction. Hence (*a*) is true; and the validity of (*b*) is shown likewise.

If σ were neither an automorphism nor an anti-automorphism of F, then there would exist elements d,e,f,g in F such that $(de)^\sigma \neq d^\sigma e^\sigma$ and $(fg)^\sigma \neq g^\sigma f^\sigma$. We infer from (II.1) and (II.2) that

$$\begin{aligned}&(de)^\sigma = e^\sigma d^\sigma \text{ and } (dx)^\sigma = x^\sigma d^\sigma,\ (xe)^\sigma = e^\sigma x^\sigma \text{ for every } x \text{ in } F,\\ &(fg)^\sigma = f^\sigma g^\sigma \text{ and } (fy)^\sigma = f^\sigma y^\sigma,\ (yg)^\sigma = y^\sigma g^\sigma \text{ for every } y \text{ in } F.\end{aligned}$$

In particular we have therefore $(dg)^\sigma = d^\sigma g^\sigma = g^\sigma d^\sigma$ and $(fe)^\sigma = e^\sigma f^\sigma = f^\sigma e^\sigma$. From (II.1) we deduce now that

$$\begin{aligned}e^\sigma d^\sigma + (fe)^\sigma + (dg)^\sigma + f^\sigma g^\sigma &= (de)^\sigma + (fe)^\sigma + (dg)^\sigma + (fg)^\sigma \\ &= [(d + f)(e + g)]^\sigma \\ &= \begin{cases} (d + f)^\sigma(e + g)^\sigma = d^\sigma e^\sigma + (fe)^\sigma + (dg)^\sigma + f^\sigma g^\sigma \\ (e + g)^\sigma(d + f)^\sigma = e^\sigma d^\sigma + (fe)^\sigma + (dg)^\sigma + g^\sigma f^\sigma. \end{cases}\end{aligned}$$

But both these possibilities lead to a contradiction. Thus it is impossible that σ is neither an automorphism nor an anti-automorphism. Hence (III) is a consequence of (II).

Assume finally the validity of (III). Then σ certainly satisfies the following conditions:

$$(x+y)^\sigma = x^\sigma + y^\sigma, \qquad (zx)^\sigma = zx^\sigma \text{ for } x,y \text{ in } F,$$

since z is a center element of F such that $z = z^\sigma$. Consider now four distinct points P,Q,R,S on the line L such that $\begin{bmatrix} P & Q \\ S & R \end{bmatrix} = z$. We distinguish two cases.

Case 1: One of the points P,Q,R,S is the point Fq.

In Remark 2 we have shown that the cross ratio does not change its value, if the four points undergo a certain transitive group of permutations. Hence we may assume without loss in generality that

$$P = Fq, \qquad Q = F(p+uq), \qquad R = F(p+vq), \qquad S = F(p+wq).$$

It follows from Lemma 1, (*a*) that

$$(v-u)(w-u)^{-1} = z \qquad \text{or} \qquad v-u = z(w-u).$$

From the properties of σ already stated it follows that

$$v^\sigma - u^\sigma = (v-u)^\sigma = [z(w-u)]^\sigma = z(w-u)^\sigma = z(w^\sigma - u^\sigma)$$

or $(v^\sigma - u^\sigma)(w^\sigma - u^\sigma)^{-1} = z$; and now it follows from Lemma 1, (*a*) that

$$\begin{bmatrix} P^\sigma & Q^\sigma \\ S^\sigma & R^\sigma \end{bmatrix} = \begin{bmatrix} Fq & F(p+u^\sigma q) \\ F(p+w^\sigma q) & F(p+v^\sigma q) \end{bmatrix} = \langle (v^\sigma - u^\sigma)(w^\sigma - u^\sigma)^{-1} \rangle = z.$$

Case 2: None of the points P,Q,R,S is the point Fq.

Then our points have the form $P = F(p+hq)$, $Q = F(p+kq)$, $R = F(p+mq)$, $S = F(p+nq)$; and it follows from Lemma 1, (*b*) that

$$(*) \qquad (h-n)(k-n)^{-1}(k-m)(h-m)^{-1} = z;$$

and that

$$(**) \qquad \begin{bmatrix} P^\sigma & Q^\sigma \\ S^\sigma & R^\sigma \end{bmatrix} = \langle (h^\sigma - n^\sigma)(k^\sigma - n^\sigma)^{-1}(k^\sigma - m^\sigma)(h^\sigma - m^\sigma)^{-1} \rangle.$$

If σ happens to be an automorphism of the field F, then we deduce from (*), (**) and the invariance of z immediately that

$$\begin{bmatrix} P^\sigma & Q^\sigma \\ S^\sigma & R^\sigma \end{bmatrix} = z.$$

Thus we assume finally that σ is an anti-automorphism of F. We note first that

$$(+)\left\{\begin{aligned} &1-[(h-n)(k-n)^{-1}-1][(h-n)(h-m)^{-1}-1]\\ &\quad=(h-n)(k-n)^{-1}+(h-n)(h-m)^{-1}-(h-n)(k-n)^{-1}(h-n)(h-m)^{-1}\\ &\quad=(h-n)[(k-n)^{-1}(h-m)+1-(k-n)^{-1}(h-n)](h-m)^{-1}\\ &\quad=(h-n)(k-n)^{-1}[(h-m)+(k-n)-(h-n)](h-m)^{-1}\\ &\quad=(h-n)(k-n)^{-1}(k-m)(h-m)^{-1}. \end{aligned}\right.$$

Since this expression is by (*) the center element z, and since from $z=1-uv$ we may deduce $z=u^{-1}[1-uv]u=1-vu$, we find that

$$\begin{aligned} z&=(h-n)(k-n)^{-1}(k-m)(h-m)^{-1}\\ &=1-[(h-n)(k-n)^{-1}-1][(h-n)(h-m)^{-1}-1]\\ &=1-[(h-n)(h-m)^{-1}-1][(h-n)(k-n)^{-1}-1]\\ &=1-[(n-h)(m-h)^{-1}-1][(n-h)(n-k)^{-1}-1] \qquad [\text{by } (+)]\\ &=(n-h)(m-h)^{-1}(m-k)(n-k)^{-1}\\ &=(h-n)(h-m)^{-1}(k-m)(k-n)^{-1}\\ &=(h-m)^{-1}(k-m)(k-n)^{-1}(h-n) \end{aligned}$$

since z is in the center and since therefore transformation with $(h-n)$ does not change the value of z. Now it follows from (**) and the invariance of z under σ that

$$\begin{bmatrix} P^\sigma & Q^\sigma \\ S^\sigma & R^\sigma \end{bmatrix} = \langle [(h-m)^{-1}(k-m)(k-n)^{-1}(h-n)]^\sigma \rangle = z;$$

and thus we have shown that (I) is a consequence of (III). This completes the proof.

REMARK 6: When proving that (III) is a consequence of (II), we actually showed the following fact:

A permutation σ of the field F is an automorphism or anti-automorphism of F if, and only if, $1=1^\sigma$, $(x+y)^\sigma=x^\sigma+y^\sigma$, $(xyx)^\sigma=x^\sigma y^\sigma x^\sigma$.

These permutations σ of F shall be termed *semi-automorphisms of F.*

The proof we gave is a reproduction of a proof due to Hua [see L.-K. Hua: On the Automorphism of a Sfield, *Proceedings of the National Acad. of Sciences*, vol. 35 (1949), 386-389]. For a more recent investigation of the general problem of semi-automorphisms the reader is advised to consult N. Jacobson–C. E. Rickart: Jordan Homomorphisms of Rings, *Trans. Amer. Math. Soc.*, vol. 69 (1950), pp. 479–502.

REMARK 7: If one considers in particular the case $z=-1$ [which is impossible if, and only if, F has characteristic 2], then one investigates the permutations of the points of a line which preserve harmonic sets [Remark 5]; and one aims at a generalization of what is known as von

Staudt's Theorem; see in this context Proposition 4 below and G. Ancochea: Le théorème de von Staudt en géométrie projective quaternionienne, *Journal für die reine und angewandte Mathematik*, vol. 184 (1942), pp. 193-198, where further references may be found.

For a convenient enunciation of further results we have to introduce a concept which generalizes the relation between point permutations and number permutations appearing in Proposition 3. Consider a one-to-one correspondence σ mapping the totality of points on the line L upon the totality of points on the line L^σ. Then we term the permutation τ of the numbers in the field F a *σ-admissible permutation of F*, if there exist independent elements p,q in L and independent elements p',q' in L^σ such that

$$(Fp)^\sigma = Fp',\ [F(p+q)]^\sigma = F(p'+q'),\ (Fq)^\sigma = Fq',$$
$$[F(p+xq)]^\sigma = F(p'+x^\tau q') \qquad \text{for every } x \text{ in } F.$$

We note that $0 = 0^\tau$, $1 = 1^\tau$ is an immediate consequence of the preceding equations; and that the field permutation σ appearing in Proposition 3 is admissible with respect to the point permutation σ. It is easy to see that to every one-to-one mapping σ of the points on L upon the points on L^σ there exists at least one σ-admissible permutation of the field F; and that permutations of the field F which leave invariant 0 and 1 lead [in many ways] to one-to-one mappings of the points on one line onto the points on some other line.

Lemma 2: *If σ is a one-to-one correspondence mapping the points on the line L upon the totality of points on the line L^σ, and if the semi-automorphism α of the field F is σ-admissible, then*

$$\begin{bmatrix} P^\sigma & Q^\sigma \\ S^\sigma & R^\sigma \end{bmatrix} = \begin{bmatrix} P & Q \\ S & R \end{bmatrix}^\alpha \text{ for any four points } P,Q,R,S \text{ on } L.$$

Proof: Since σ is α-admissible, there exist elements p,q,p',q' such that

$$L = Fp + Fq, \qquad L^\sigma = Fp' + Fq',$$
$$(Fp)^\sigma = Fp', \qquad [F(p+q)]^\sigma = F(p'+q'), \qquad (Fq)^\sigma = Fq',$$
$$[F(p+xq)]^\sigma = F(p'+x^\alpha q') \text{ for every } x \text{ in } F.$$

From now on we are going to use arguments quite similar to those used already when deriving (I) from (III) in the proof of Proposition 3.

Consider four distinct points P,Q,R,S on L. We distinguish two cases.

Case 1: One of the four points is Fq.

We may assume without loss in generality that

$$P = Fq \text{ and } Q = F(p+uq),\ R = F(p+vq),\ S = F(p+wq).$$

Then u,v,w are three distinct numbers in F; and we deduce from Lemma 1, (*a*) that

$$\begin{bmatrix} P & Q \\ S & R \end{bmatrix} = \langle (v-u)(w-u)^{-1} \rangle, \qquad \begin{bmatrix} P^\sigma & Q^\sigma \\ S^\sigma & R^\sigma \end{bmatrix} = \langle (v^\alpha - u^\alpha)(w^\alpha - u^\alpha)^{-1} \rangle.$$

But α is an automorphism or anti-automorphism of F; and the numbers xy^{-1} and $y^{-1}x = y^{-1}(xy^{-1})y$ are conjugate in F. Hence

$$\langle (v^\alpha - u^\alpha)(w^\alpha - u^\alpha)^{-1} \rangle = \langle [(v - u)(w - u)^{-1}]^\alpha \rangle$$

or

$$\begin{bmatrix} P^\sigma & Q^\sigma \\ S^\sigma & R^\sigma \end{bmatrix} = \begin{bmatrix} P & Q \\ S & R \end{bmatrix}^\alpha$$

Case 2: None of the four points P,Q,R,S equals Fq.

In this case we have $P = F(p + hq)$, $Q = F(p + kq)$, $R = F(p + mq)$, $S = F(p + nq)$ where h,k,m,n are four different numbers in F. It follows from Lemma 1, (*b*) that

$$\begin{bmatrix} P & Q \\ S & R \end{bmatrix} = \langle (h - n)(k - n)^{-1}(k - m)(h - m)^{-1} \rangle,$$

$$\begin{bmatrix} P^\sigma & Q^\sigma \\ S^\sigma & R^\sigma \end{bmatrix} = \langle (h^\alpha - n^\alpha)(k^\alpha - n^\alpha)^{-1}(k^\alpha - m^\alpha)(h^\alpha - m^\alpha)^{-1} \rangle.$$

During the last part of the proof of Proposition 3 we verified an identity:

$$(+) \qquad \begin{aligned} &(h - n)(k - n)^{-1}(k - m)(h - m)^{-1} \\ &\quad = 1 - [(h - n)(k - n)^{-1} - 1][(h - n)(h - m)^{-1} - 1] = t; \end{aligned}$$

and this number t is conjugate in F to

$$\begin{aligned} t' &= 1 - [(h - n)(h - m)^{-1} - 1][(h - n)(k - n)^{-1} - 1] \\ &= 1 - [(n - h)(m - h)^{-1} - 1][(n - h)(n - k)^{-1} - 1] \\ &= (n - h)(m - h)^{-1}(m - k)(n - k)^{-1} \end{aligned}$$

by a second application of the identity $(+)$; and this number t' is conjugate in F to

$$t'' = (m - h)^{-1}(m - k)(n - k)^{-1}(n - h) = (h - m)^{-1}(k - m)(k - n)^{-1}(h - n)$$

which is obtained from t by inversion of the order of the factors. We may state this result as follows:

$$(++) \quad \langle (h - n)(k - n)^{-1}(k - m)(h - m)^{-1} \rangle = \langle (h - m)^{-1}(k - m)(k - n)^{-1}(h - n) \rangle.$$

If α is an automorphism, then we may apply it on the first expression in $(++)$; and if α is an anti-automorphism, then we apply it on the second expression in $(++)$. In this way we see that

$$\begin{bmatrix} P & Q \\ S & R \end{bmatrix}^\alpha = \langle (h^\alpha - n^\alpha)(k^\alpha - n^\alpha)^{-1}(k^\alpha - m^\alpha)(h^\alpha - m^\alpha)^{-1} \rangle = \begin{bmatrix} P^\sigma & Q^\sigma \\ S^\sigma & R^\sigma \end{bmatrix};$$

and this completes the proof of the lemma.

Proposition 4: *Suppose that* σ *is a one-to-one correspondence mapping the points on the line* L *upon the totality of points on the line* L^σ*; and suppose that the center of* F *contains at least three elements. Then* σ *preserves all the* *cross ratios if, and only if, the σ-admissible permutations of the field F are semi-automorphisms which leave invariant every class of conjugate elements in F.*

PROOF: The sufficiency of our condition is an immediate consequence of Lemma 2. Assume conversely that σ preserves all the cross ratios and consider a σ-admissible permutation τ of the elements in the field F. Then there exist pairs of independent elements p,q and p',q' in L and L^{σ} respectively such that

$$(Fp)^{\sigma} = Fp', \quad [F(p+q)]^{\sigma} = F(p'+q'), \quad (Fq)^{\sigma} = Fq',$$
$$[F(p+xq)]^{\sigma} = F(p'+x^{\tau}q') \text{ for every } x \text{ in } F.$$

There exists clearly a linear transformation η of the F-space L^{σ} upon the F-space L such that $(xp'+yq')^{\eta} = xp+yq$ for x,y in F. This linear transformation η induces a one-to-one correspondence σ' of the totality of points on L^{σ} upon the totality of points on L; and it is easy to see [either by application of Lemma 1 or by application of Lemma 2] that σ' preserves cross ratios. Consequently $\sigma\sigma'$ is a cross ratio preserving permutation of the points on the line L; and this permutation has clearly the following further properties:

$$(Fp)^{\sigma\sigma'} = Fp, \quad [F(p+q)]^{\sigma\sigma'} = F(p+q), \quad (Fq)^{\sigma\sigma'} = Fq,$$
$$[F(p+xq)]^{\sigma\sigma'} = F(p+x^{\tau}q) \text{ for } x \text{ in } F.$$

By hypothesis there exists in the center of F a number $z \neq 0, 1$. Since the permutation $\sigma\sigma'$ preserves all cross ratios, it satisfies Proposition 3, (I); and consequently it follows from Proposition 3 that τ is a semi-automorphism. That this semi-automorphism τ leaves invariant every class of conjugate elements in F, is an immediate consequence of the fact that

$$\langle x\rangle = \begin{bmatrix} Fp & Fq \\ F(p+xq) & F(p+q) \end{bmatrix} = \begin{bmatrix} Fp & Fq \\ F(p+x^{\tau}q) & F(p+q) \end{bmatrix} = \langle x^{\tau}\rangle$$

holds for every x in F. This completes the proof.

REMARK 8: If σ is a semi-automorphism of the field F which leaves invariant every class of conjugate elements in F, then σ leaves invariant every element in the center Z of F. If σ is an automorphism of the field F which leaves invariant every element in the center Z of F, and if F is finite over Z, then it is a celebrated theorem that σ is an inner automorphism of F [see, for instance, Artin-Nesbitt-Thrall (1), p. 66, Corollary 7.2D]. But if F is infinite over Z, then there exist automorphisms of F which leave invariant every class of conjugate elements in F and which are *not* inner automorphisms of F. [An example may be found in G. Köthe: Schiefkörper unendlichen Ranges über dem Zentrum, *Math.*

Ann., vol. 105 (1931), pp. 15-39.] If F is finite over Z, and if σ is an anti-automorphism of F, then σ leaves invariant every element in Z whereas the elements x and x^σ, for x in F, always satisfy the same equations over Z. It follows that x and x^σ are conjugate elements in F [see, for instance, Artin-Nesbitt-Thrall (1), p. 67, Theorem 7.2 E]. An example of such an anti-automorphism may be obtained by mapping every real quaternion $x_0 + x_1 i + x_2 j + x_3 k$ upon its "conjugate" $x_0 - x_1 i - x_2 j - x_3 k$. This shows that semi-automorphisms which leave invariant every class of conjugate elements may be inner automorphisms, automorphisms which are not inner and anti-automorphisms which are not automorphisms; and it is clear that cross ratios do not suffice for distinguishing between these various kinds of mappings of the points on a line.

Theorem 2: *The field F is commutative if, and only if, the identity is the only permutation of the points on a line which preserves cross ratios and possesses three fixed points.*

This theorem is a kind of complement to the Second Fundamental Theorem of Projective Geometry [III.3]; and it is also an improvement upon Remark 4.

PROOF: Assume first that the identity is the only permutation of the points on a line which preserves cross ratios and possesses three fixed points. If $g \neq 0$ is a number in F, and if p,q are two independent elements in A, then we may define a permutation σ of the points on the line $L = Fp + Fq$ by the rule:

$$(Fq)^\sigma = Fq, \qquad [F(p + xq)]^\sigma = F(p + g^{-1}xgq) \text{ for } x \text{ in } F.$$

It is clear that σ possesses the three distinct fixed points, Fp, Fq, $F(p + q)$; and it follows from Lemma 2 that σ preserves cross ratios, since the mapping $x \rightarrow g^{-1}xg$ is an inner automorphism of the field F. It follows from our hypothesis that $\sigma = 1$. Hence $p + g^{-1}xgq$ belongs, for every x in F, to $F(p + xq)$. This implies $x = g^{-1}xg$ for every x in F, since p and q are independent elements. Thus g belongs to the center of F. Since we have shown that every element in F belongs to the center of F, we have shown the commutativity of F.

Assume conversely the commutativity of the field F. If F consists of the two elements 0 and 1 only, then lines carry three points only; and the only permutation of the points on a line with three fixed points is naturally the identity. Thus we may assume that F contains at least three elements. Consider now a cross ratio preserving permutation σ of the points on the line L which possesses three fixed points. These fixed points may be assumed to have the form Fp, $F(p + q)$, Fq. Then there exists a σ-admissible permutation τ of the numbers in F such that

$$[F(p + \ )]xq^\sigma = F(p + x^\tau q)$$

for every x in F. It follows from Proposition 4 that τ is a semi-automorphism of F which leaves invariant every class of conjugate elements in F. But F is a commutative field so that $\tau = 1$. Hence $\sigma = 1$, as we wanted to show.

Remark 9: It is easy to see the equivalence of the conditions of Theorem 2 with the following property:

The one-to-one correspondences σ' and σ'' of the totality of points on the line L upon the totality of points on the line L' satisfy $\sigma' = \sigma''$ if, and only if,

(*a*) There exist three distinct points P_i $[i = 1,2,3]$ on L such that $P_i^{\sigma'} = P_i^{\sigma''}$ for $i = 1,2,3$; and

(*b*) $\begin{bmatrix} P^{\sigma'} & Q^{\sigma'} \\ S^{\sigma'} & R^{\sigma'} \end{bmatrix} = \begin{bmatrix} P^{\sigma''} & Q^{\sigma''} \\ S^{\sigma''} & P^{\sigma''} \end{bmatrix}$ for any distinct points P,Q,R,S on L.

Proposition 5: *Suppose that the characteristic of the field F is not 2 [so that $+1 \neq -1$]. Then the following properties of the one-to-one mapping σ of the points on the line L onto the points on the line L^σ are equivalent.*

(I) $\begin{bmatrix} P & Q \\ S & R \end{bmatrix} = \begin{bmatrix} P' & Q' \\ S' & R' \end{bmatrix}$, *if, and only if,* $\begin{bmatrix} P^\sigma & Q^\sigma \\ S^\sigma & R^\sigma \end{bmatrix} = \begin{bmatrix} P'^\sigma & Q'^\sigma \\ S'^\sigma & R'^\sigma \end{bmatrix}$.

(II) *Suppose that P,Q,R,S is a harmonic set. Then P',Q',R',S' is a harmonic set if, and only if,*

$$\begin{bmatrix} P^\sigma & Q^\sigma \\ S^\sigma & R^\sigma \end{bmatrix} = \begin{bmatrix} P'^\sigma & Q'^\sigma \\ S'^\sigma & R'^\sigma \end{bmatrix}$$

(III) *σ-admissible permutations of F are semi-automorphisms.*

Proof: It is trivial that (I) implies (II), since harmonic sets have cross ratio -1. Assume next the validity of (II). Consider any harmonic set P,Q,R,S; and suppose that W is a point on L^σ such that

$$\begin{bmatrix} P^\sigma & Q^\sigma \\ S^\sigma & R^\sigma \end{bmatrix} = \begin{bmatrix} P^\sigma & Q^\sigma \\ W & R^\sigma \end{bmatrix}$$

Then there exists one and only one point V on the line L such that $W = V^\sigma$; and it follows from (II) that P,Q,R,V is also a harmonic set. It follows from Proposition 2 that $V = S$; and this implies $W = S^\sigma$. Applying Proposition 2 again it follows now that the cross ratio $\begin{bmatrix} P^\sigma & Q^\sigma \\ S^\sigma & R^\sigma \end{bmatrix}$ is a center element z of F. With P,Q,R,S also P,Q,S,R is a harmonic set; and it follows from (II) that

$$z = \begin{bmatrix} P^\sigma & Q^\sigma \\ S^\sigma & R^\sigma \end{bmatrix} = \begin{bmatrix} P^\sigma & Q^\sigma \\ R^\sigma & S^\sigma \end{bmatrix} = z^{-1} \text{ or } z = \pm 1.$$

Since σ maps the four distinct points P,Q,R,S upon four distinct points, it is impossible that $z = 1$; and thus we have shown that σ maps harmonic sets upon harmonic sets. By an argument quite similar to the one used in the proof of Proposition 4 we may deduce now from Proposition 3 that σ-admissible permutations of F are semi-automorphisms; and thus (III) is a consequence of (II). That (I) is a consequence of (III), is a fairly immediate consequence of Lemma 2.

Lemma 3: *Suppose that σ is a one-to-one mapping of the points on the line L upon the points on the line L^σ.*

(*a*) *σ is induced in L by an auto-projectivity of the F-space A if, and only if, the σ-admissible permutations of F are automorphisms.*

(*b*) *σ is induced in L by a collineation of the F-space A if, and only if, the σ-admissible permutations of F are inner automorphisms.*

The proof is easily constructed once one remembers that projectivities are induced by semi-linear transformations [III.1, First Fundamental Theorem of Projective Geometry] and that a semi-linear transformation induces a collineation if, and only if, the component automorphism of F is inner [III.2, Lemma 1].

If the σ-admissible permutations of the field F are anti-automorphisms but not automorphisms, then σ cannot be induced by an auto-projectivity. It is, however, possible to deduce these mappings in some fashion from auto-dualities by means of the theory developed in the next chapter. We leave this to the interested reader.

It is clear from the discussion of Remark 8 that cross ratios will not suffice to characterize those mappings of the points on a line which are induced by auto-projectivities. Thus we need a certain generalization of the concept of cross ratio.

If P,Q,R are three distinct points on the line L, then we may represent them in the form: $P = Fp$, $Q = Fq$, $R = Fr$ subject to the condition that $p + q + r = 0$. If X,Y,Z are points on L none of which equals Q, then these points may be represented in one and only one way in the form $X = F(p + xq)$, $Y = F(p + yq)$, $Z = F(p + zq)$. If we select another admissible representation: $P = Fp'$, $Q = Fq'$, $R = Fr'$ subject to the condition $p' + q' + r' = 0$, then there exists a number $g \neq 0$ in F such that $p' = gp$, $q' = gq$, $r' = gr$; and we find that

$$X = F(p' + gxg^{-1}q'),\ Y = F(p' + gyg^{-1}q'),\ Z = F(p' + gzg^{-1}q').$$

Hence the second representation leads us to the conjugate triplet

$$g(x,y,z)g^{-1} = (gxg^{-1}, gyg^{-1}, gzg^{-1})$$

of numbers in F. It is easy to see that by changing the admissible representations we obtain a full class of conjugate triplets of numbers in F. Using the notation $\langle (x,y,x) \rangle$ for the totality of triplets $g(x,y,z)g^{-1}$ with $g \neq 0$ in F, we may now define

DEFINITION 2 : $\begin{bmatrix} Fp & Fq & F(p+q) \\ F(p+xq) & F(p+yq) & F(p+zq) \end{bmatrix} = \langle (x,y,z) \rangle.$

We note that $\begin{bmatrix} P & Q & R \\ X & Y & Z \end{bmatrix}$ is defined whenever the points P,Q,R,X,Y,Z are collinear, the points P,Q,R are different from each other and none of the points X,Y,Z equals Q. If furthermore

$$\begin{bmatrix} P & Q & R \\ X & Y & Z \end{bmatrix} = \langle (x,y,z) \rangle,$$

then

$$\begin{bmatrix} P & Q \\ X & R \end{bmatrix} = \langle x \rangle, \begin{bmatrix} P & Q \\ Y & R \end{bmatrix} = \langle y \rangle, \begin{bmatrix} P & Q \\ Z & R \end{bmatrix} = \langle z \rangle;$$

but the converse will not always be true except if x,y,z belong all three to the center of F.

Proposition 6: *Suppose that the center of the field F contains at least three elements, and that σ is a one-to-one mapping of the points on the line L upon the points on the line L^{σ}. Then σ is induced by an auto-projectivity of the F-space A if, and only if, there exists an automorphism α of the field F such that*

$$\begin{bmatrix} P^{\sigma} & Q^{\sigma} & R^{\sigma} \\ X^{\sigma} & Y^{\sigma} & Z^{\sigma} \end{bmatrix} = \begin{bmatrix} P & Q & R \\ X & Y & Z \end{bmatrix}^{\alpha} \textit{for all [admissible] } P,Q,R,X,Y,Z \textit{ on } L.$$

PROOF: It suffices to prove the sufficiency of the condition. If the condition is satisfied, then there exists a σ-admissible permutation τ of F which satisfies

$$\langle (x^{\tau},y^{\tau},z^{\tau}) \rangle = \langle (x^{\alpha},y^{\alpha},z^{\alpha}) \rangle \text{ for } x,y,z \text{ in } A.$$

One notes furthermore easily that $\tau\alpha^{-1}$ is a σ'-admissible permutation of F where σ' is an [easily constructed] permutation of the points on L which preserves cross ratios and has at least three fixed points. It follows from Proposition 3 that $\tau\alpha^{-1}$ is a semi-automorphism β of the field F such that (x,y,z) and $(x^{\beta},y^{\beta},z^{\beta})$ are conjugate in F. Consequently there exists to every pair x,y of elements in F an element $g \neq 0$ in F such that

$$x^{\beta} = g^{-1}xg, \; y^{\beta} = g^{-1}yg, \; (xy)^{\beta} = g^{-1}xyg.$$

But this implies

$$(xy)^{\beta} = g^{-1}xg\, g^{-1}yg = x^{\beta}y^{\beta}.$$

Hence β is an automorphism of F. Since α is an automorphism of F, τ is an automorphism of F; and now it follows from Lemma 3 that σ is induced by an auto-projectivity of the F-space A.

The discussion of Remark 8 makes it evident that this new concept, though sufficient to characterize the mappings induced by auto-projectivities, will not suffice to characterize those induced by collineations, again with the exception of certain special situations. What distinguishes the mapping induced by collineations from the others, is the fact that the former admit of geometrical construction; and this we are going to show now.

Suppose that the lines L' and L'' belong to the same plane of the F-space A; and that P is some point in this plane which is neither on L' nor on

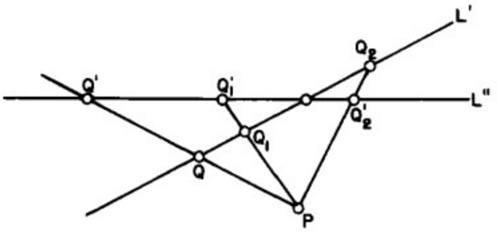

Fig. 6

L''. Then $L' \oplus P = L'' \oplus P$ is a plane carrying the lines L' and L''. If Q' is a point on L', then $Q'' = L'' \cap (P + Q')$ is a point on L'' [as follows from II.2, Rank Formula]. Likewise $L' \cap (P + Q'')$ is a point on L' for every point Q'' on L''; and these two operations are reciprocals of each other, since one deduces from Dedekind's Law that

$$\begin{aligned} L' \cap (P + [L'' \cap (P + Q')]) &= L' \cap (P + L'') \cap (P + Q') \\ &= L' \cap (P + Q') = Q', \end{aligned}$$

if we only note that $L' \leqslant P + L''$. This implies in particular that the mapping of the point Q' on L' upon the point $L'' \cap (P + Q')$ on L'' is one-to-one and exhaustive; and this mapping we shall call *the projection of the line L' from the point P upon the line L''*. [Note that this projection is the identity mapping, if $L' = L''$.]

Suppose now that $L_0, L_1, \cdots, L_k$ are lines in the F-space A and that L_i, L_{i+1} are always coplanar lines. If P_i is some point neither on L_i nor on L_{i+1}, such that $L_i + P_i = P_i + L_{i+1}$, then the projection σ_i of the line L_i from the point P_i upon the line L_{i+1} is well defined. We may form the product $\sigma = \sigma_0\sigma_1 \cdots \sigma_{k+1}$ which is a one-to-one mapping of the totality of points on L_0 upon the totality of points on L_k. Any such mapping σ of the points on L_0 upon the points on L_k shall be called *a linear mapping of the line L_0 upon the line L_k*. [Various names have been given in the literature to what we called here, for purely algebraical reasons, a linear mapping.]

Proposition 7: *The mapping σ of the points on the line L' upon the points on the line L'' is a linear mapping if, and only if, σ is induced by a collineation of the F-space A.*

Proof: We want to show first that linear mappings are induced by collineations. Since every linear mapping is a product of projections, and since every product of collineations is itself a collineation, it suffices to prove that every projection is induced by a collineation. Consider therefore two different coplanar lines L' and L'' and a point P on the plane

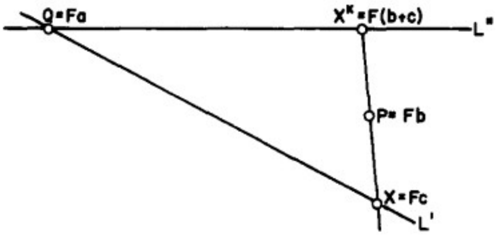

Fig. 7

$L' + L''$ which is neither on L' nor on L''. Then $L' + L'' = L'' + P = P + L'$ is a plane; and the lines L' and L'' meet in a point Q. Consider some point X on L' such that $X \neq Q$; and denote by $\varkappa$ the projection of L' from P onto L''. Then $X,P,X^{\varkappa}$ are three distinct collinear points; and as usual we can find elements b,c such that $P = Fb$, $X = Fc$, $X^{\varkappa} = F(b + c)$; and we let a be some element such that $Q = Fa$. Then there exists a linear transformation τ of the F-space A such that $a = a^{\tau}$, $b = b^{\tau}$, $c^{\tau} = b + c$ which leaves invariant every element in some space complementary to $L' + L''$. Every point on L' has the form $F(ua + vc)$. From

$$(ua + vc)^{\tau} = ua + v(b + c) = (ua + vc) + vb$$

we infer that $[F(ua + vc)]^{\tau}$ is the intersection $[F(ua + vc)]^{\varkappa}$ of the lines $L'' = Fa + F(b + c)$ and $F(ua + vc) + P$. Thus $\varkappa$ is induced by τ; and this shows, as we remarked before, that every linear mapping is induced by a collineation of the F-space A.

Suppose next that the mapping σ of the line L' upon the line L'' is induced by a collineation of the F-space A. Every collineation is induced by a linear transformation; and thus we may assume that σ is induced by the linear transformation η of the F-space A.

Case 1: $L' \neq L''$.

Since L' carries at least three points, there exist two different points Fu and Fv on L' such that neither $(Fu)^{\sigma} = Fu^{\eta}$ nor $(Fv)^{\sigma} = Fv^{\eta}$ is on L'. At most one of the points Fu and Fv is on L''; and thus we may assume without loss in generality that Fv is not on L''. If we let $u^{\eta} = u'$, $v^{\eta} = v'$, then it follows from our choice of u and v that the triplets u,v,u' and v,u',v' are independent. Now we define linear transformations ν,ω of the planes $Fu + Fv + Fu'$ and $Fv + Fu' + Fv'$ respectively by the rules

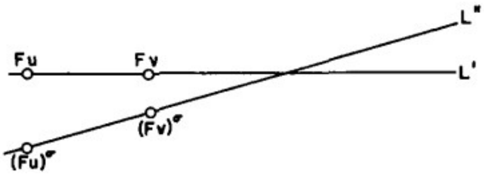

Fig. 8

$$u^{\nu} = u',\ v^{\nu} = v,\ u'^{\nu} = 2u' - u; \qquad v^{\omega} = v',\ u'^{\omega} = u',\ v'^{\omega} = 2v' - v.$$

One verifies that ν induces a projection of the line $Fu + Fv = L'$ from the point $F(u' - u)$ upon the line $Fu' + Fv$; that ω induces a projection of the line $Fu' + Fv$ from the point $F(v' - v)$ upon the line $Fu' + Fv' = L''$; and that $\nu\omega$ induces a linear transformation of L' upon L'' which coincides with the linear transformation η. Thus we have represented σ as a product of two projections.

Case 2: $L' = L''$.

Since the F-space A has rank not less than 3, there exists a line L different from $L' = L''$. There exists a projection σ' of L' upon L; and we have shown in the first part of the proof that σ' is induced by a collineation and therefore by a linear transformation η' of the F-space A. The linear transformation $\eta\eta'$ of the F-space A induces the mapping $\sigma\sigma'$ of the line L' upon the line $L \neq L'$, since σ maps L' upon $L'' = L'$. From what we have shown in Case 1 it follows that $\sigma\sigma'$ is the product of two projections. Since the inverse of a projection is likewise a projection, we have shown that σ is the product of three projections. This completes the proof.

APPENDIX III

Projective Ordering of a Space

Suppose that an algebraical ordering of the field F is given. By this we mean that *a domain of positivity* P has been distinguished in F. Such a subset P meets the following requirements:

(*a*) 0 is not in P.
(*b*) If x and y are in P, then $x + y$ and xy are in P.
(*c*) If $x \neq 0$ is not in P, then $-x$ is in P.

The numbers in P are referred to as positive numbers. [The reader should verify the fact that the definition

$$x < y \text{ if, and only if, } y - x \text{ is in } P$$

leads to an algebraical ordering of the field F which has the properties we expect an algebraical ordering to have.]

One proves:

(*d*) If $x \neq 0$, then x^2 is in P and either x, x^{-1} are both in P or x, x^{-1} are both not in P.

We leave the simple verification to the reader. From (*d*) one deduces easily.

(*e*) If y is in P, then $x^{-1}yx$ is in P for every $x \neq 0$ in F.

Consider now four distinct points P, Q, R, S on a line L in the F-space A. Then their cross ratio $\begin{bmatrix} P & Q \\ S & R \end{bmatrix}$ is a complete class of conjugate numbers in F none of which is 0 [or 1]. Thus it follows from (*e*) that either all the numbers in the cross ratio belong to P or none of them is in P. In the first case we term the cross ratio *positive* and in the second case *negative*. In a similar fashion we may see that either all the numbers in the cross ratio are *smaller than* 1 or all of them are *greater than* 1. [For $f^{-1}xf < 1$ if, and only if, $1 - f^{-1}xf = f^{-1}(1 - x)f$ belongs to P; and this is equivalent to saying that $1 - x$ is in P or that $x < 1$.] Now we may define:

The pair P, Q separates the pair R, S if, and only if, the cross ratio $\begin{bmatrix} P & Q \\ S & R \end{bmatrix}$ is both positive and smaller than 1.

That this soc. cyclical ordering of the points on L has all the properties we expect of it, the reader will easily verify using the symmetry properties of cross ratios. [For an account of the basic properties of cyclical order see, for instance, Coxeter (1), p. 22-24.] This ordering is furthermore invariant under all projections and collineations as follows from III.4, Proposition 7 and III.4, Lemma 2.

REMARK: The theory of commutative ordered fields is well developed [Artin-Schreier Theory; see, for instance, van der Waerden: Moderne Algebra, vol. 1]. But as regards non-commutative ordered fields there is known not much more than their existence [see, for instance, Reidemeister (1)] and the equivalence of orderability and formal reality recently established by I. Szele.

Another approach to the problem of order in projective spaces is given by E. Sperner: Beziehungen zwischen geometrischer und algebraischer Anordnung, *Sitzungsberichte der Heidelberger Akademie der Wissenschaften*; Math.-nat. Kl., 1949 (10), and E. Sperner: Die Ordnungsfunktionen einer Geometrie, *Math. Ann.*, vol. 121 (1949), pp. 107 ss.

Pág. 89 (103) - 6.1 Quaternions and
isometries
Isometries of Euclidean space

$\mathbb{R}^3$ is an isometry if it preserves distances, that is, if for all x and y,
$\|f(x) - f(y)\| = \|x - y\|$.

Important:

Quaternions ler em conjunto com a pág. 02 do item anterior - Treatise of plane geometry .... Cavet (2001).pdf

Hamilton's Elements of Quaternions.

Maxwell himself already applied to the electromagnetic field.

Pág. 307 (320) Chapter 15
Hyperbolic geometry

- Linear Algebra and Projective Geometry.pdf

Pág. 66 (75) III.3
The Second Fundamental Theorem of Projective Geometry

Pág. 69 (78) Appendix II

The Theorem of Pappus

In the Second Fundamental theorem of Projective Geometry, we have met a first instance where the commutativity of the field F (of coordinates) appears important.

Obs: In the present appendix, we give Hilbert's celebrated characterization of commutativity of F by means geometric configurations.

Fig. 4 — ten coplanar points

Pág. 71(80) – III. 4.

The Projective Geometry of a Line in Space; Cross Ratios-

Look: Fig. 5 — Pág. 76(85)

Pág. 85(94) Proposition 4

Suppose that σ is one-to-one correspondence mapping the points on the line L upon the totality of points on the line L^σ; and suppose that the center of F contains at least three elements

then σ preserves all the cross ratios
if, and only if,
the σ-admissible permutations of
the field F are
semi-automorphisms which
leave invariant every class of
conjugate elements in F

Proof: Pag. 86(95)

Look: Pág. 91(100) Fig. 6

Pág. 131(140) Appendix II
Projective Relations between
Lines Induced by Polarities

Pág. 266(275) VII. 3.
the Theorem of Desargues.
(important)

Proposition 1: Assume that the three lines L, L', L'' pass through the point P but are not contained in a plane.

— Books-019-2021/20210920/
Pythagorean-Hodograph curves.../

(' * ,33 # I &&

lutorial anti-automorphism of F, such that R consists exactly of those numbers in F which are left invariant by α, then a discussion quite similar to the one under Application 1 leads now to the following normal form of f:

$$f(x,y) = \sum_{i=1}^{p} x_i y_i^{\alpha} - \sum_{i=p+1}^{r(A)} x_i y_i^{\alpha}$$

where p is again the number of positive points in maximal O-sets.

Remark 6: In the preceding examples the index of inertia is actually equal to the index $i(\sigma)$ introduced in section IV.4. In general one can only prove that $i(\sigma)$ does not exceed the index of inertia, but it is fairly easy to construct examples where the index of inertia actually exceeds the index $i(\sigma)$.

APPENDIX II

Projective Relations between Lines Induced by Polarities

If σ is a polarity of the linear manifold (F,A), and if L is a line, then A/L^{σ} has rank 2 [since the only proper subspaces of L are points, the only subspaces properly between L^{σ} and A are hyperplanes; and this is equivalent to saying that $r(A/L^{\sigma}) = 2$]. If L' is some line satisfying $L' \cap L^{\sigma} = 0$, then it follows from $r(A/L^{\sigma}) = 2$ that $A = L' \oplus L^{\sigma}$. But this is equivalent to $A = L'^{\sigma} \oplus L$ so that we have shown:

The following properties of the lines L and L' are equivalent:

$$L' \cap L^{\sigma} = 0,\ A = L' \oplus L^{\sigma},\ A = L \oplus L'^{\sigma},\ 0 = L \cap L'^{\sigma}.$$

Consider now a pair of lines L and L' meeting these equivalent requirements. If P is a point on L, then $L' \cap P^{\sigma} = P'$ is a point, since the hyperplane P^{σ} contains L^{σ} and satisfies

$$P^{\sigma} = L^{\sigma} \oplus (P^{\sigma} \cap L') = L^{\sigma} \oplus P'.$$

Furthermore we deduce from Dedekind's Law that

$$L \cap P'^{\sigma} = L \cap (L' \cap P^{\sigma})^{\sigma} = L \cap (L'^{\sigma} + P) = P + (L \cap L'^{\sigma}) = P.$$

Thus we see that mapping the point P on L onto the point $P' = L' \cap P^{\sigma}$ on L' and mapping the point Q on L' onto the point $L \cap Q^{\sigma}$ on L are reciprocal mappings; and this shows in particular that these two mappings are both one-to-one and exhaustive. It will be convenient to term the mapping of the point $P \leqslant L$ onto the point $P^{\sigma} \cap L'$ *the σ-mapping of the line L upon the line L'*; and we note that we can speak of this σ-mapping if, and only if, $A = L' \oplus L^\sigma$. Our objective in the present appendix is an investigation of these mappings.

Lemma 1: *Assume that the polarity σ is represented by the symmetrical α-form $f(x,y)$, that $L = Fp + Fq$ is a line in the linear manifold (F,A), and that the lines L and L' satisfy $A = L' \oplus L^\sigma$. Then there exist elements p',q' such that*

$$L' = Fp' + Fq',\ Fq' = L' \cap (Fq)^\sigma,\ F(p' + x^\alpha q') = L' \cap [F(p + xq)]^\sigma$$

for every number x in F.

Proof: We let $P' = L' \cap P^\sigma$ for every point P on L. Then mapping P onto P' is the σ-mapping of L upon L'; and this mapping is one-to-one and exhaustive, since $A = L' \oplus L^\sigma$, as we have pointed out before. The points $(Fp)'$, $(Fq)'$ and $[F(p + q)]'$ are consequently three distinct points on L'; and we may as usual show the existence of elements p'', q'' such that $(Fp)' = Fp''$, $(Fq)' = Fq''$, $[F(p + q)]' = F(p'' + q'')$. From $Fp'' \leqslant (Fp)^\sigma$ we deduce $f(p,p'') = f(p'',p) = 0$, since σ is represented by f. If $f(p'',q)$ were 0 too, then it would follow that $f(p'',L) = 0$ so that Fp'' would be on L^σ. But $L' \cap L^\sigma = 0$ by hypothesis and Fp'' is a point on L'. Hence $t = f(p'',q) \neq 0$; and we may let $p' = t^{-1}p''$, $q' = t^{-1}q''$. Then we have naturally

$$(Fp)' = Fp',\ (Fq)' = Fq',\ [F(p + q)]' = F(p' + q'),\ 1 = f(p',q)$$

so that in particular $L' = Fp' + Fq'$. As before we note that $f(p,p') = f(p',p) = 0$, $f(q,q') = f(q',q) = 0$; and we find likewise

$$0 = f(p + q,p' + q') = f(p,p') + f(p,q') + f(q,p') + f(q,q') = f(p,q') + 1$$

or $f(p,q') = f(q',p) = -1$, since $f(x,y) = f(y,x)^\alpha$ because of the assumed symmetry of the form f. Now we verify that

$$\begin{aligned} f(p + xq,p' + x^\alpha q') &= f(p,p') + f(p,q')x + xf(q,p') + xf(q,q')x \\ &= -x + x = 0, \end{aligned}$$

since α is involutorial. Hence $F(p' + x^\alpha q')$ is on $[F(p + xq)]^\sigma$ as well as on L'; and now our Lemma is an immediate consequence of the fact that the σ-mapping of L upon L' is a one-to-one and exhaustive mapping of the points on L onto the points on L'.

Proposition 1: *Suppose that the linear manifold (F,A) has rank not less than 3 and that the polarity σ of (F,A) is not a null system.*

(*a*) *The σ-mappings of lines are induced by auto-projectivities of (F,A) if, and only if, F is commutative.*

(*b*) *The σ-mapping of lines are induced by collineations of (F,A) if, and only if, σ may be represented by bilinear forms.*

Proof: It is a consequence of IV.3, Theorem 1 that σ may be represented by a symmetrical α-form f where α is an involutorial anti-automorphism of the field F. If L is a line, then there exists a subspace T of A such that $A = L \oplus T$. Naturally $L' = T^\sigma$ is a line such that $A = L \oplus L'^\sigma$; and thus we have shown that there exist σ-mappings of all lines.

It follows from Lemma 1 and III.4, Lemma 3 that the σ-mappings of lines are induced by auto-projectivities if, and only if, α is an automorphism of F; and that the σ-mappings of lines are induced by collineations if, and only if, α is an inner automorphism of A. But the anti-automorphism α of F is an automorphism if, and only if, F is commutative; and α is inner if, and only if, $\alpha = 1$; and this completes the proof of our proposition.

Remark 1: If the lines L and L' satisfy $A = L \oplus L'^\sigma$, then it follows from Lemma 1 and the definition of cross ratios [in III.4] that

$$\begin{bmatrix} L' \cap P^\sigma & L' \cap Q^\sigma \\ L' \cap S^\sigma & L' \cap R^\sigma \end{bmatrix} = \begin{bmatrix} P & Q \\ S & R \end{bmatrix}^\alpha$$

holds for any four distinct points P,Q,R,S on L, provided σ is represented by a symmetrical α-form. Thus apart from auto-projectivities also the polarities provide a means of constructing mappings of the points on a line which conserve equality of cross ratios.

Proposition 2: *Assume that L is a line and σ a polarity of the linear manifold (F,A).*

(*a*) *If $L \cap L^\sigma$ is a point, then $L \cap L^\sigma$ is the one and only one N-point on L.*

(*b*) *If σ is represented by a semi-bilinear form, if the characteristic of F is not 2, and if L carries one and only one N-point, then $L \cap L^\sigma$ is a point.*

Proof: If $P = L \cap L^\sigma$ is a point, then $0 < L^\sigma$ and $L < A$ so that $3 \leqslant r(A)$. Hence σ may be represented by a semi-bilinear form [IV.1, Proposition 1]. From $P \leqslant L \leqslant L + L^\sigma = (L^\sigma \cap L)^\sigma = P^\sigma$ we infer that P is an N-point on L. Assume that Q is a second N-point on L. Then $L = P + Q$, $Q \leqslant L \leqslant P^\sigma$; and it follows from IV.4, Lemma 3 that L is an N-line. Consequently we deduce from IV.4, Corollary 1 that $L \cap L^\sigma = 0$ or L, contradicting our hypotheses that $L \cap L^\sigma$ is a point. Thus P is the only N-point on L. Hence (*a*) is true.

Assume now that σ may be represented by some semi-bilinear form, that the characteristic of F is not 2, and that L carries one and only one N-point P. The line L carries some point $P' \neq P$. Then P' is on L, but not on P'^σ; and this implies in particular that L is not part of the hyperplane P'^σ. Hence $P'' = L \cap P'^\sigma$ is a point which is different from P', since otherwise P' would be an N-point. Either $P'' = P$ or else there exist, as usual, elements p', p'' such that $P' = Fp'$, $P = F(p' + p'')$, $P'' = Fp''$. There exists an α-form $f(x,y)$ which represents σ and which satisfies

$1 = f(p',p')$, since P' is not an N-point; and it follows from IV.3, Proposition 1 that $f(x,y)$ is α-symmetrical and $\alpha^2 = 1$. Since P'' is on P'^{σ} we have $0 = f(p'',p') = f(p',p'')$; and since P is an N-point, we have

$$0 = f(p' + p'',p' + p'') = f(p',p') + f(p'',p'') = 1 + f(p'',p'')$$

or $f(p'',p'') = -1$. Since the characteristic of F is not 2, the point $F(p' - p'')$ is different from P; and this point is clearly an N-point, since $0 = f(p' - p'',p' - p'')$. Thus we have been led to a contradiction by assuming that $P \neq P''$. Hence $P = P'' = L \cap P'^{\sigma}$. Consequently

$$P \leqslant L \cap P^{\sigma} \cap P'^{\sigma} = L \cap (P + P')^{\sigma} = L \cap L^{\sigma},$$

since P is an N-point on L. It is impossible that $L = L \cap L^{\sigma}$, since then every point on L would be an N-point. Since L is a line, this implies $P = L \cap L^{\sigma}$, completing the proof.

Remark 2: If A is a plane and the characteristic of F is different from 2, then the point L^{σ} is the pole of the line L; and Proposition 2 asserts that the pole of the line L is on L if, and only if, L carries one and only one N-point.

Remark 3: If $A = L$ is a line, then $L^{\sigma} = 0$, and every permutation of the points on L which possesses one and only one fixed point provides an example of a polarity for which (*b*) is not valid. Thus the hypothesis that σ be represented by a semi-bilinear form is indispensable. We note that this hypothesis is satisfied whenever $3 \leqslant r(A)$ as follows from IV. 1, Proposition 1.

Remark 4: If the characteristic of the field F is 2, then (*b*) ceases to be valid, as follows from the following often used example. Denote by F any commutative field of characteristic 2 and by (F,A) the plane over F. If we represent the elements in A, as we may, by triplets $x = (x_1, x_2, x_3)$ of numbers x_i in F, then we may consider the bilinear form

$$f(x,y) = \sum_{i=1}^{3} x_i y_i$$

[= scalar product]. The point $F(x_1,x_2,x_3)$ is an N-point if, and only if, it is on the line $\sum_{i=1}^{3} x_i = 0$, since $f(x,x) = \sum_{i=1}^{3} x_i^2 = \left[\sum_{i=1}^{3} x_i\right]^2$ as the characteristic of F is 2. This N-line L is met by every other line in the plane so that every line not L carries one and only one N-point. But the line $x_1 = 0$ has the pole $F(1,0,0)$ which is not on its polar so that this line carries one and only one N-point, though it does not carry its pole.

Lemma 2: *The involutorial anti-automorphism α of the field F is the identity-automorphism if, and only if,*

(*a*) $rr^\alpha = 1$ *implies* $r = \pm 1$ *and*

(*b*) $xx^\alpha = x^\alpha x$ *for every* x *in* F *whenever the characteristic of* F *is two.*

REMARK: It seems to be an open question whether condition (*b*) is a consequence of condition (*a*).

PROOF: The necessity of our conditions is fairly obvious. Assume now the validity of condition (*a*). Then we prove first that

(*a*.1) $xx^\alpha = x^\alpha x$ *implies* $x^\alpha = \pm x$.

Since this is certainly true, if $x = 0$, we may assume that $x \neq 0$; and then we may form $r = x^{-1}x^\alpha$. Since x and x^α commute, we find that

$$rr^\alpha = x^{-1}x^\alpha x x^{-\alpha} = 1.$$

It follows from (*a*) that $r = \pm 1$; and this is clearly equivalent to $x^\alpha = \pm x$.

If the field F has characteristic 2, then $\alpha = 1$ is an immediate consequence of (*a*.1) and (*b*) and therefore of (*a*) and (*b*).

Assume next that the characteristic of F is different from 2; and assume the validity of (*a*) and (*a*.1). If t is any element in F, then the element $s = (t^\alpha - t) - 1$ certainly commutes with the element

$$s^\alpha = t - t^\alpha - 1 = -(t^\alpha - t) - 1 = -s - 2.$$

Hence it follows from (*a*.1) that

$$t - t^\alpha - 1 = s^\alpha = \pm s = \pm (t^\alpha - t - 1).$$

The minus sign is impossible in this equation, since this would imply $-1 = +1$ contradicting our assumption that the characteristic of F is not two. Hence we have $t - t^\alpha - 1 = t^\alpha - t - 1$ or $2t^\alpha = 2t$; and this implies $t^\alpha = t$, since we assumed the characteristic of F to be different from 2. Thus we have shown that $\alpha = 1$; and this completes the proof.

Proposition 3: *Suppose that the polarity σ of the linear manifold (F,A) possesses N-points and that $3 \leqslant r(A)$. Then σ may be represented by ordinary bilinear forms if, and only if,*

(*a*) *lines containing more than two N-points are N-lines and*

(*b*) F *is commutative.*

If the characteristic of F is different from 2, then condition (b) is a consequence of condition (a).

PROOF: Since $3 \leqslant r(A)$, it is possible to represent σ by a semi-bilinear form [IV.1, Proposition 1]; and it should be noted that this will be the only application of the hypothesis $3 \leqslant r(A)$. It follows from IV.1, Proposition 3 that every form representing σ is an ordinary bilinear form, if one of the representing forms has this property.

Assume now that it is possible to represent σ by ordinary bilinear forms. Then the identity automorphism is an anti-automorphism of F so that F is a commutative field. Suppose now that the line L carries the three distinct N-points P,Q,R. Then there exist elements p,q such that $P = Fp$, $Q = Fq$, $R = F(p + q)$. If f is some ordinary bilinear form representing σ, then we have

$$0 = f(p,p) = f(q,q) = f(p,q) + f(q,p);$$

and it follows from the commutativity of F and the bilinearity of f that

$$\begin{aligned} f(xp + yq, xp + yq) &= xf(p,p)x + xf(p,q)y + yf(q,p)x + yf(q,q)y \\ &= xy[f(p,q) + f(q,p)] = 0. \end{aligned}$$

Hence L is an N-line. This completes the proof of the necessity of conditions (*a*) and (*b*). The reader ought to investigate the connection between condition (*a*), Proposition 1 and the fact that a linear transformation of a line with three fixed points is the identity.

We assume now the validity of (*a*). There exists at least one N-point P. We distinguish two cases.

Case 1: Every point, not on P^σ, is an N-point.

If Q is a point on P^σ, but $Q \neq P$, then $Q^\sigma \neq P^\sigma$ so that the hyperplane Q^σ carries at least one point Q' not on P^σ. Then the line $L = Q + Q'$ is not on P^σ and meets therefore P^σ in the point Q. If $Q = Fq$, $Q' = Fq'$ and $R = F(q + q')$, then Q' and R are certainly not on P^σ and are therefore N-points. If f is any semi-bilinear form representing σ, then we have consequently

$$\begin{aligned} &0 = f(q,q') = f(q',q) = f(q',q') \text{ and} \\ &0 = f(q + q', q + q') = f(q,q) \end{aligned}$$

so that Q itself is an N-point. Thus we have shown that every point in A is an N-point. Hence σ is a null system; and it follows from IV.2, Proposition 1 that f is an ordinary bilinear form.

Case 2: There exists a point W which is neither an N-point nor on P^σ.

Let $W = Fw$. Then it is possible to select an α-form f which represents σ and satisfies $f(w,w) = 1$; and we may infer from IV.3, Proposition 1 that α is involutorial and f symmetrical. We have $f(w,P) \neq 0$, since W is not on P^σ; and consequently we may select an element p such that $P = Fp$ and $1 = f(w,p)$. Because of the symmetry of f we have $f(p,w) = 1$; and $f(p,p) = 0$, since P is an N-point. Every point on the line $L = P + W$ with the exception of W has the form $F(p + xw)$; and the correspondence between the numbers x in F and the points different from W on L is a one-to-one correspondence. The point $F(p + xw)$ is an N-point if, and only if,

$$0 = f(p + xw, p + xw) = x + x^\alpha + xx^\alpha = (x + 1)(x + 1)^\alpha - 1.$$

The number of N-points on L is consequently exactly the number of solutions x of the equation $(x+1)(x+1)^{\alpha}=1$, if we remember that W is not an N-point. As this point W is on L, the line L is not an N-line; and it follows from condition (a) that L carries at most two N-points. Consequently we have shown that:

(a^{*}) *The equation* $(x+1)(x+1)^{\alpha}=1$ *has at most two solutions* x *in* F.

If we let $y=x+1$, then we see that (a^{*}) is equivalent with the assertion that the equation $yy^{\alpha}=1$ has at most two solutions y in F. This equation has certainly the solutions $y=\pm 1$. If it possessed a solution $z\neq\pm 1$, then $z^{-1}\neq z$ would be a further solution which is different from ± 1; and thus we would have at least three solutions of $yy^{\alpha}=1$ [at least four if $+1\neq -1$]. Since this is impossible, we have shown that:

(a^{**}) $jj^{\alpha}=1$ *implies* $j=\pm 1$.

Thus we have seen that our present condition (a) implies the validity of Lemma 2, (a); and Lemma 2, (b) is certainly a consequence of our present condition (b) which we need apply only in case the characteristic of F is two. Thus we may deduce $\alpha=1$ from Lemma 2; and we have shown again that σ may be represented by ordinary bilinear forms. This completes the proof.

Remark 5: The question whether our condition (b) is a consequence of our condition (a) is equivalent to the corresponding question concerning Lemma 2, since (a^{**}) has been shown to be a consequence of (a) alone. On the other hand it is completely impossible to omit the hypothesis assuring the existence of N-points, as may be seen from easily constructed examples of "positive definite semi-bilinear forms."

Remark 6: Suppose that the polarity σ of the linear manifold (F,A) possesses N-points, that lines carrying more than two N-points are N-lines, that $3\leqslant r(A)$ and that the characteristic of F is two. Then we show:

(c) *Lines carrying at least two N-points are N-lines.*

Suppose namely that P and Q are different N-points and that f is some semi-bilinear form representing σ. As the characteristic of F is two, we may assume without loss in generality that f is symmetrical. If firstly P is on Q^{σ}, then we have

$$0=f(P,P)=f(P,Q)=f(Q,P)=f(Q,Q);$$

and this implies clearly that $0=f(P+Q,P+Q)$ so that $P+Q$ is an N-line. If secondly P is not on Q^{σ}, then $f(P,Q)\neq 0$; and we may select elements p and q such that $P=Fp$, $Q=Fq$ and $1=f(p,q)=f(q,p)$. Since we have $0=f(p,p)=f(q,q)$, it follows that

$$f(p+q,p+q)=f(p,q)+f(q,p)=2=0;$$

and we have shown the existence of a third N-point on $P + Q$. It follows from our hypothesis that $P + Q$ is an N-line; and this completes the proof of (c).

From (c) one deduces without difficulty that:

(d) *The totality of N-points is a subspace.*

Some information on such a subspace may be found in IV.4.

Proposition 4: *If each of the lines L' and L'' carries at least two N-points, though neither of them is an N-line, then L' and L'' carry the same number of N-points.*

Proof: This proposition is vacuously true, if either $r(A) < 3$ or our polarity σ is a null system. Thus we assume that σ is not a null system and that $3 \leqslant r(A)$. Then it follows from IV.3, Theorem 1 that σ may be represented by a symmetrical α-form f with $\alpha^2 = 1$. Consider now a line L which is not an N-line, but carries the two distinct N-points P and Q. Then P is not on Q^σ, since otherwise L would be an N-line [IV.4, Lemma 3]. Let $Q = Fq$. Then $f(q,q) = 0$, but $f(P,q) \neq 0$. Consequently there exists p such that $P = Fp$ and $f(p,q) = 1$. Since f is symmetrical, we have $f(q,p) = 1$; and since P is an N-point, we have $f(p,p) = 0$. To every point X on L different from P and Q, there exists one and only one number $x \neq 0$ in F such that $X = F(p + xq)$. This point is an N-point if, and only if,

$$0 = f(p + xq, p + xq) = x + x^\alpha.$$

If we denote now by C the cardinal number of elements x in F such that $x^\alpha = -x$ [the skew-symmetric elements in F], then we have found that:

The number of N-points on L is exactly $C + 1$.

This last assertion is actually slightly sharper than our proposition.

APPENDIX III

The Theorem of Pascal

Our discussion of the celebrated Theorem of Pascal is rather similar in intent to our previous discussion of the Theorem of Pappus [III, Appendix II]. Whereas this theorem is of universal validity in the classical theory of conics, it will be shown that from our more general point of view it must be considered to be valid only under quite exceptional circumstances. The reader may consult Coxeter [1] or Levi [1] for a discussion of Pascal's Theorem in its classical framework.

It will be convenient to assume throughout this appendix that the linear manifold (F,A) is a plane $[r(A) = 3]$. This will simplify our dis-

( )) " && ' 0 0 H !

We remark finally that the name *hyperplane* will be used for elements H such that A/H is a point. If A has finite rank, then $r(A) = r(H) + 1$; and if A has infinite rank, then $r(A) = r(H)$. Note that the rank condition defines hyperplanes in the finite case only.

Whenever possible, we shall make use of the appropriate geometrical terminology. Thus we shall speak of lines passing through a point, carrying a point, meeting in a point etc.

VII.3. The Theorem of Desargues

The considerations of the present section are almost identical with those customary in ordinary three dimensional projective geometry. Thus we shall omit some of the details.

Proposition 1: *Assume that the three lines L, L', L'' pass through the point P, but are not contained in a plane. Assume furthermore that Q, R are points on L, that Q', R' are points on L' and that Q'', R'' are points on L'' and that none of these points equals P. Assume finally that $Q + Q' \neq R + R'$, $Q' + Q'' \neq R' + R''$ and $Q'' + Q \neq R'' + R$. Then $(Q + Q') \cap (R + R')$, $(Q' + Q'') \cap (R' + R'')$, $(Q'' + Q) \cap (R'' + R)$ are collinear points.*

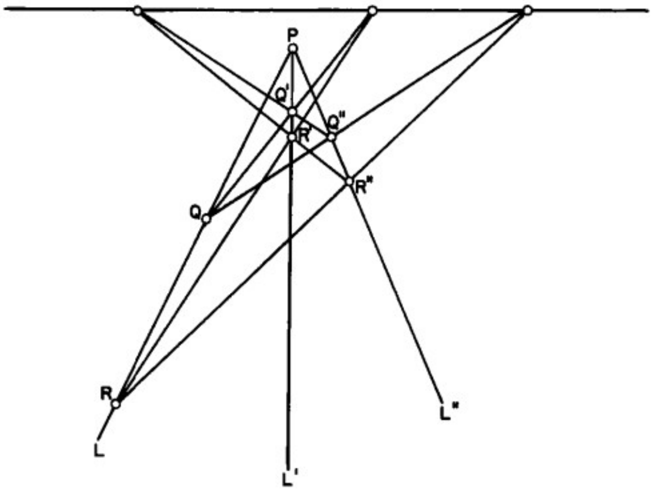

Fig. 13

Proof: We note first that $L + L'$, $L' + L''$ and $L'' + L$ are three distinct planes. Next we note that $Q + Q' + Q''$ and $R + R' + R''$ are two distinct planes. Finally we remark that $Q + Q'$ and $R + R'$ are two distinct lines meeting in the point $(Q + Q') \cap (R + R')$; and likewise we see that $(Q' + Q'') \cap (R' + R'')$ and $(Q'' + Q) \cap (R'' + R)$ are points. But all these points are on the intersection of the distinct planes $Q + Q' + Q''$ and $R + R' + R''$. But their intersection is clearly a line, since they are contained in the element $L + L' + L''$ of rank 4. This completes the proof.

Proposition 2: *Assume that* $3 < r(A)$. *Suppose that the three distinct lines* L, L', L'' *pass through the point* P, *that the points* Q, R *are on* L, *the points* Q', R' *on* L' *and the points* Q'', R'' *on* L'', *and that none of these points equals* P. *If finally* $Q + Q' \neq R + R'$, $Q' + Q'' \neq R' + R''$ *and* $Q'' + Q \neq R'' + R$, *then* $(Q + Q') \cap (R + R')$, $(Q' + Q'') \cap (R' + R'')$ *and* $(Q'' + Q) \cap (R'' + R)$ *are collinear points.*

The reduction of Proposition 2 to Proposition 1 may be effected in the usual fashion. [See e.g. Veblen-Young [1], p. 41.]

If we omit in Proposition 2 the hypothesis that $3 < r(A)$ or in Proposition 1 the hypothesis that L, L', L'' are not coplanar, then these Propositions cease to be true. At this point begins the investigation of the Theory of Projective Planes [which may or may not be Desarguesian]; and this theory is completely outside the scope of our investigation. It is, however, necessary for our purposes that we are assured of the validity of the Theorem of Desargues; and thus we impose the following further postulate.

IX. *If the three distinct lines* L, L', L'' *pass through the point* P, *if the points* Q, R *are on* L, *the points* Q', R' *on* L' *and the points* Q'', R'' *on* L'', *if none of these points equals* P, *if* $Q + Q' \neq R + R'$, $Q' + Q'' \neq R' + R''$ *and* $Q'' + Q = R'' + R$, *then* $(Q + Q') \cap (R + R')$, $(Q' + Q'') \cap (R' + R'')$, $(Q'' + Q) \cap (R'' + R)$ *are collinear points.*

VERIFICATION OF POSTULATE IX IN $S(F,A)$:

If the F-space A has rank not less than 4, then the validity of IX is a consequence of Proposition 2. If the F-space A has rank less than 4, then we form $A \oplus A$ which more comprehensive F-space has rank greater than 3, provided $r(A) = 3$ [if $r(A) < 3$, then IX is trivially satisfied]. Hence IX holds in $A \oplus A$ and consequently in the subspace A of $A \oplus A$. It would not have been difficult to verify the validity of IX by direct computation, a useful exercise.

Proposition 3: *Assume the validity of Postulates* I *to* IX. *If* P, Q, R *are three different, but collinear points, if* $P', P'', Q', Q'', R', R''$ *are points such that* $P' + Q'$ *and* $P'' + Q''$ *are different lines meeting in* R, $Q' + R'$ *and* $Q'' + R''$ *are different lines meeting in* P, $R' + P'$ *and* $R'' + P''$ *are different lines meeting in* Q, *and if* $P' + P''$, $Q' + Q''$ *and* $R' + R''$ *are three distinct lines, then these lines* $P' + P''$, $Q' + Q''$ *and* $R' + R''$ *are copunctual* [*have a point in common*].

This dualized theorem of Desargues may be verified as follows. If $3 < r(A)$, then the considerations used in proving Proposition 2 may be used [mutatis mutandis]; and if $3 = r(A)$, then the principle of duality may be applied to Postulate IX. We leave the details to the reader.

VII.4. The Imbedding Theorem

The objective of the present section is the proof of the following

Imbedding Theorem: *If the partially ordered set S satisfies Postulates* I *to* VIII, *and if the maximal element A in S has rank not less than* 3, *then the following condition is necessary and sufficient for the validity of the Theorem of Desargues [Postulate* IX] *in S:*

IX*. *S is contained in a partially ordered set T which is different from S and which satisfies Postulates* I *to* VIII.

If in particular the maximal element A has rank greater than 3, then the validity of IX is a consequence of VII.3, Proposition 2; and thus it follows from the Imbedding Theorem that S has also the Imbedding Property IX*. This is the principal reason why we have to prove the comparatively deep Imbedding Theorem instead of substituting for Postulate IX the Postulate IX*.

The Imbedding Theorem will be exploited in the next section. The following remark may indicate how this can be done. If S is properly contained in T, then there exists in T an element B such that A is a hyperplane with respect to B or B is a point over A. One may then construct the homothetic group of B over A; and this group contains in a way the additive group and the field of a linear manifold whose system of subspaces is projectively equivalent with S [see III, Appendix I].

That the Theorem of Desargues [Postulate IX] is a consequence of IX*, may be deduced from VII.3, Proposition 2, since the maximal element B of the containing system T certainly satisfies $A < B$ so that $3 < r(B)$.

Throughout the remainder of this section [which may be omitted in a first reading] we shall be concerned with the proof of the sufficiency of IX. Thus we assume throughout that Postulates I to IX are satisfied by S and that its maximal element A has rank not less than 3. We begin by constructing

THE BASIC QUADRANGLE:

It consists of four lines L_o, L_r, L_s, L_t in S which are coplanar, though no three of these lines are copunctual.

The existence of this basic quadrangle is easily deduced from the fact that there exist at least three independent points $[3 \leqslant r(A)]$ and that every line carries at least three distinct points [VIII].

then σ preserves all the cross ratios
if, and only if,
the σ-admissible permutations of
the field F are
semi-automorphisms which
leave invariant every class of
conjugate elements in F

Proof: Pág. 86(95)

Look: Pág. 91(100) Fig. 6
Pág. 131(140) Appendix II
Projective Relations between
Lines Induced by Polarities

Pág. 266(275) VII.3.
The Theorem of Desargues.
(important)

Proposition 1: Assume that the three
lines L, L', L'' pass through the point
P but are not contained in a
plane.

— Books-019-2021| 20210920|
Pythagorean-Hodograph curves...|

+,-.+ + +/ ! + * ! ! # 1
,33 ! -! ! 2 4 3#
(

The Rationality of the Moduli Space of Curves of Genus 3 after P. Katsylo

Christian Böhning

Mathematisches Institut
Bunsenstrasse 3-5
37073 Göttingen, Germany
`boehning@uni-math.gwdg.de`

Summary. This article is a survey of P. Katsylo's proof that the moduli space $\mathfrak{M}_3$ of smooth projective complex curves of genus 3 is rational. We hope to make the argument more comprehensible and transparent by emphasizing the underlying geometry in the proof and its key structural features.

Key words: Rationality, moduli spaces of curves

2000 Mathematics Subject Classification codes: 14E08, 14H10, 14H45

1 Introduction

The question whether or not $\mathfrak{M}_3$ is a rational variety had been open for a long time until an affirmative answer was finally given by P. Katsylo in 1996. There is a well-known transition in the behavior of the moduli spaces $\mathfrak{M}_g$ of smooth projective complex curves of genus g from unirational for small g to general type for larger values of g; the moral reason that $\mathfrak{M}_3$ should have a good chance to be rational is that it is birational to a quotient of a projective space by a *connected* linear algebraic group. No variety of this form has been proved to be irrational up to now. More precisely, $\mathfrak{M}_3$ is birational to the moduli space of plane quartic curves for $\mathrm{PGL}_3\,\mathbb{C}$-equivalence. All the moduli spaces $C(d)$ of plane curves of given degree d are conjectured to be rational (see [Dol2, p.162]; in fact, there it is conjectured that all the moduli spaces of hypersurfaces of given degree d in $\mathbb{P}^n$ for the $\mathrm{PGL}_{n+1}\,\mathbb{C}$-action are rational).

There are heuristic reasons that the spaces $C(d)$ should be rational at least for all large enough values for d. It should not be completely out of reach to prove this rigorously and we hope to return to this problem in the future. In any case, one might guess that irregular behavior of $C(d)$ is most likely to be found for small values of d, and showing rationality for $C(4)$ turned out to be exceptionally hard.

F. Bogomolov, Y. Tschinkel (eds.), *Cohomological and Geometric Approaches to Rationality Problems,* Progress in Mathematics 282, DOI 10.1007/978-0-8176-4934-0_2,
© Birkhäuser Boston, a part of Springer Science+Business Media, LLC 2010

/4 , # @@@

+,-.+ + +/ ! + * , 4 !P , P ! # 1 2
& <! % 3#
# +,-.+ + +/ ! + * , 4 !P , P ! # 1 2
&& % 3#
/4
( )% *

Analytic Curves in Algebraic Varieties over Number Fields

Jean-Benoît Bost[1] and Antoine Chambert-Loir[2]

[1] Département de mathématiques, Université de Paris-Sud (Orsay), Bâtiment 425, F-91405 Orsay Cedex, France. `Jean-Benoit.Bost@math.u-psud.fr`

[2] Université de Rennes 1, Institut de recherche mathématique de Rennes, Campus de Beaulieu, F-35042 Rennes Cedex, France.
`antoine.chambert-loir@univ-rennes1.fr`

To Yuri Ivanovich Manin on the occasion of his 70th birthday

Summary. We establish algebraicity criteria for formal germs of curves in algebraic varieties over number fields and apply them to derive a rationality criterion for formal germs of functions on algebraic curves that extends the classical rationality theorems of Borel–ŨDwork and Polya–ŨBertrandias, valid over the projective line, to arbitrary algebraic curves over a number field. The formulation and the proof of these criteria involve some basic notions in Arakelov geometry, combined with complex and rigid analytic geometry (notably, potential theory over complex and p-adic curves). We also discuss geometric analogues, pertaining to the algebraic geometry of projective surfaces of these arithmetic criteria.

Key words: Arakelov geometry, capacity theory, BorelŨ–Dwork rationality criterion, algebricity of formal subschemes, rigid analytic geometry, slope inequality.

2000 Mathematics Subject Classifications: 14G40 (Primary); 14G22, 31A15, 14B20 (Secondary)

1 Introduction

The purpose of this article is to establish *algebraicity criteria* for *formal germs of curves* in algebraic varieties over number fields and to apply them to derive a *rationality criterion* for *formal germs of functions*, which extends the classical rationality theorems of Borel–Dwork [6], [22] and Pólya–Bertrandias [1, Chapter 5], [43], (see also [16]), valid over the projective line, to arbitrary algebraic curves over a number field.

Our algebraicity criteria improve on those in [12] and [13], which themselves were inspired by the papers [19] and [20] of D. V. and G. V. Chudnovsky

Y. Tschinkel and Y. Zarhin (eds.), *Algebra, Arithmetic, and Geometry*,
Progress in Mathematics 269, DOI 10.1007/978-0-8176-4745-2_3,
© Springer Science+Business Media, LLC 2009

and by the subsequent works by André [2] and Graftieaux [26, 27]. As in [12] and [13], our results will be proved by means of a geometric version of "transcendence techniques," which avoids the traditional constructions of "auxiliary polynomials" and the explicit use of Siegel's lemma, replacing them by a few basic concepts of Arakelov geometry. In the proofs, our main objects of interest will be some evaluation maps, defined on the spaces of global sections of powers of an ample line bundle on a projective variety by restricting these sections to formal subschemes or to subschemes of finite lengths. Arakelov geometry enters through the estimates satisfied by the heights of the evaluation maps, and the slopes and Arakelov degrees of the hermitian vector bundles defined by spaces of sections (see [17] and [14] for more details and references on this approach).

Our main motivation in investigating the algebraicity and rationality criteria presented in this article has been the desire to obtain theorems respecting the classical principle of number theory that "all places of number fields should appear on an equal footing"—which actually is not the case in "classical" Arakelov geometry and in its applications in [12]. A closely related aim has been to establish arithmetic theorems whose geometric counterparts (obtained through the analogy between number fields and function fields) have simple formulations and proofs. These concerns led us to two technical developments in this paper: the use of (rigid) analytic geometry over p-adic fields to define and estimate local invariants of formal curves over number fields,[3] and the derivation of a rationality criterion from an algebraicity criterion by means of the Hodge index theorem on (algebraic or arithmetic) surfaces.

Let us describe the contents of this article in more detail.

In Section 2, we discuss geometric analogues of our arithmetic theorems. Actually, these are classical results in algebraic geometry, going back to Hartshorne [32] and Hironaka–Matsumura [35]. For instance, our algebraicity result admits as analogue the following fact. *Let* X *be a quasiprojective variety over a field* k*, and let* Y *be a smooth projective integral curve in* X*; let* $\widehat{S}$ *be a smooth formal germ of surface through* Y (that is, a smooth formal subscheme of dimension 2, containing Y, of the completion $\widehat{X}_Y$). *If the degree* $\deg_Y N_Y\widehat{S}$ *of the normal bundle to* Y *in* $\widehat{S}$ *is positive, then* $\widehat{S}$ *is algebraic.*

Our point is that, transposed to a geometric setting, the arguments leading to our algebraicity and rationality criteria in the arithmetic setting—which rely on the consideration of suitable evaluation maps and on the Hodge index theorem—provide simple proofs of these nontrivial algebro-geometric results, in which the geometric punch line of the arguments appears more clearly.

In Section 3, we introduce the notion of *A-analytic curve* in an algebraic variety X over a number field K. By definition, this will be a smooth formal curve $\widehat{C}$ through a rational point P in $X(K)$—that is, a smooth formal subscheme of dimension 1 in the completion $\widehat{X}_P$—which, firstly, is analytic at every place of K, finite or infinite. Namely, if v denotes any such place and K_v the corresponding completion of K, the formal curve $\widehat{C}_{K_v}$ in X_{K_v} deduced from $\widehat{C}$ by the extension of scalars $K \hookrightarrow K_v$ is the formal curve defined by a K_v-analytic curve in $X(K_v)$. Moreover the v-adic radius r_v (in $]0,1]$) of the open ball in $X(K_v)$ in which $\widehat{C}_{K_v}$ "analytically extends" is required to "stay close to 1 when v varies," in the sense that the series $\sum_v \log r_v^{-1}$ has to be convergent. The precise formulation of this condition relies on the notion of *size* of a smooth analytic germ in an algebraic variety over a p-adic field. This notion was introduced in [12, 3.1]; we review it in Section 3.A, adding some complements.

---

[3] Since the first version of this paper was written, the relevance of rigid analytic geometry à la Berkovich to develop a non-archimedean potential theory on p-adic curves, and consequently a "modern" version of Arakelov geometry of arithmetic surfaces satisfying the above principle of "equality of places," has been largely demonstrated by A. Thuillier in his thesis [51].

With the above notation, if $\mathscr{X}$ is a model of X over the ring of integers $\mathscr{O}_K$ of K, and if P extends to an integral point $\mathscr{P}$ in $\mathscr{X}(\mathscr{O}_K)$, then a formal curve $\widehat{C}$ through P is A-analytic if it is analytic at each archimedean place of K and extends to a smooth formal surface $\widehat{\mathscr{C}}$ in $\widehat{\mathscr{X}}_{\mathscr{P}}$. For a general formal curve $\widehat{C}$ that is analytic at archimedean places, being an A-analytic germ may be seen as a weakened form of the existence of such a smooth extension $\widehat{\mathscr{C}}$ of $\widehat{C}$ along $\mathscr{P}$. In this way, an A-analytic curve through the point P appears as an arithmetic counterpart of the smooth formal surface $\widehat{S}$ along the curve Y in the geometric algebraicity criterion above.

The tools needed to formulate the arithmetic counterpart of the positivity condition $\deg_Y N_Y \widehat{S} > 0$ are developed in Sections 4 and 5. We first show in Section 4 how, for any germ of analytic curve $\widehat{C}$ through a rational point P in some algebraic variety X over a local field K, one is led to introduce the so-called *canonical seminorm* $\|\cdot\|^{\mathrm{can}}_{X,\widehat{C}}$ on the K-line $T_P\widehat{C}$ through the consideration of the metric properties of the evaluation maps involved in our geometric version of the method of auxiliary polynomials. This extends a definition introduced in [13] when $K = \mathbf{C}$. In Section 5, we discuss the construction of Green functions and capacities on rigid analytic curves over p-adic fields. We then extend the comparison of "canonical seminorms" and "capacitary metrics" in [13], 3.4, to the non-archimedean setting.

In Section 6, we apply these notions to formulate and establish our algebraicity results. If $\widehat{C}$ is an A-analytic curve through a rational point P in an algebraic variety X over some number field K, then the K-line $T_P\widehat{C}$ may be equipped with a "K_v-adic semi-norm" for every place v by the above construction—namely, the seminorm $\|\cdot\|^{\mathrm{can}}_{X_{K_v},\widehat{C}_{K_v}}$ on

$$T_P\widehat{C}_{K_v} \simeq T_P\widehat{C} \otimes_K K_v.$$

The so-defined metrized K-line $\overline{T_P\widehat{C}}$ has a well-defined Arakelov degree in $]-\infty,+\infty]$, and our main algebraicity criterion asserts that $\widehat{C}$ *is algebraic if the Arakelov degree* $\widehat{\deg}\, \overline{T_P\widehat{C}}$ *is positive.* Actually, the converse implication also holds: when $\widehat{C}$ is algebraic, the canonical seminorms $\|.\|^{\mathrm{can}}_{X_{K_v},\widehat{C}_{K_v}}$ all vanish, and $\widehat{\deg}\,\overline{T_P\widehat{C}} = +\infty$.

Finally, in Section 7, we derive an extension of the classical theorems of Borel, Dwork, Pólya, and Bertrandias, which gives a criterion for the rationality of a formal germ of function φ on some algebraic curve Y over a number field. By considering the graph of φ—a formal curve $\widehat{C}$ in the surface $X := Y \times \mathbf{A}^1$—we easily obtain the algebraicity of φ as a corollary of our previous algebraicity criterion. In this way, we are reduced to establishing a rationality criterion for an algebraic formal germ. Actually, rationality results for algebraic functions on the projective line have been investigated by Harbater [30], and used by Ihara [36] to study the fundamental group of some arithmetic surfaces. Ihara's results have been extended in [11] using Arakelov geometry on arithmetic surfaces. Our rationality argument in Section 7, based on the Hodge index theorem on arithmetic surfaces of Faltings–Hriljac, is a variation on the proof of the Lefschetz theorem on arithmetic surfaces in [11].

It is a pleasure to thank A. Ducros for his helpful advice on rigid analytic geometry during the preparation of this article.

Some of the results below were presented, in a preliminary form, during the "Arithmetic Geometry and Number Theory" conference in honor of N. Katz, in Princeton, December 2003, and were announced in [14].

During the preparation of this article, the authors benefitted from the support of the Institut Universitaire de France.

It would be difficult to acknowledge fairly the multifaceted influences of Yuri Ivanovich Manin on our work. We hope that this article will appear as a tribute, not only to his multiple contributions to algebraic geometry and number theory, but also to his global vision of mathematics, emphasizing geometric insights and analogies. The presentation of this vision in his 25th-Arbeitstagung report *New Dimensions in Geometry* [38] has been, since it was written, a source of wonder and inspiration to one of the authors, and we allowed ourselves to borrow the terminology "*A*-analytic" from the "*A*-geometry" programmatically discussed in [38]. It is an honor for us to dedicate this article to Yuri Ivanovich Manin.

2 Preliminary: the geometric case

The theorems we want to prove in this paper are analogues in arithmetic geometry of classical algebro-geometric results going back—at least in an implicit form—to Hartshorne, Hironaka, and Matsumura ([31],[34],[35]). Conversely, in this section we give short proofs of algebraic analogues of our main arithmetic theorems.

Proposition 2.1. *Let $\mathscr{X}$ be a quasi-projective scheme over a field k and let $\mathscr{P}$ be a projective connected subscheme of dimension 1 in $\mathscr{X}$. Let $\widehat{\mathscr{C}}$ be a formal*

+,-.+ + +/ ! + * , 4 !P , P ! # 1 2
&& % 3#
(' '''

Cyclic Homology with Coefficients

D. Kaledin*

Steklov Math Institute, Moscow, Russia
`kaledin@mccme.ru`

To Yu. I. Manin, the founder, on the occasion of his 70th birthday

Summary. We propose a category which can serve as the category of coefficients for the cyclic homology $HC_*(A)$ of an associative algebra A over a field k. The construction is categorical in nature, and essentially uses only the tensor category A-bimod of A–*bimodules*; objects of our category are A–bimodules with an additional structure. We also generalize the construction to a more general tensor k-linear tensor category C instead of A-bimod. We add some remarks about Getzler's version of the Gauss–Manin connection for the periodic cyclic homology $HP_*(A)$.

Key words: cyclic homology, Gauss–Manin connection, tensor categories

2000 Mathematics Subject Classifications: 19D55, 18E25, 18G10

Introduction

Ever since it was discovered in 1982 by A. Connes [C1] and B. Tsygan [Ts], cyclic homology has occupied a strange place in the realm of homological algebra. Normally in homological algebra problems, one expects to start from some data, for instance a topological space X, then construct some abelian category, such as the category of sheaves on X, and then define the cohomology of X by computing the derived functors of some natural functor, such as the global sections functor $\Gamma(X,-)$. Admittedly, this is a modern formulation, but it was certainly current already in 1982. Cyclic homology starts with an associative algebra A, and defines its homology groups $HC_\bullet(A)$, but there are absolutely no derived functors in sight. Originally, groups $HC_\bullet(A)$ were defined as the homology of an explicit complex, which anyone trained to use triangulated categories cannot help but take as an insult. Later, A.

---

[*] Partially supported by CRDF grant RUM1-2694-MO05.

Y. Tschinkel and Y. Zarhin (eds.), *Algebra, Arithmetic, and Geometry*,
Progress in Mathematics 270, DOI 10.1007/978-0-8176-4747-6_2,

© Springer Science+Business Media, LLC 2009

Connes [C2] improved on the definition by introducing the abelian category of so-called cyclic vector spaces. However, the passage from A to its associated cyclic vector space $A_\#$ is still done by an explicit *ad hoc* formula. It is as if we were to know the bar-complex that computes the homology of a group, without knowing the definition of the homology of a group.

This situation undoubtedly irked many people over the years, but to the best of my knowledge, no satisfactory solution has been proposed, and it may not exist; indeed, many relations to the de Rham homology notwithstanding, it is not clear whether cyclic homology properly forms a part of homological algebra at all (to the point that for instance in [FT], the word "homology" is not used at all for $HC_\bullet(A)$, and it is called instead *additive K-theory* of A). In the great codification of homological algebra done in [GM1], cyclic homology appears only in the exercises. This is not surprising, since the main unifying idea of [GM1] is the ideology of "linearization": homological algebra linearizes geometry, just as functional analysis used to do 50 years ago; triangulated categories and adjoint functors are modern-day versions of Banach spaces and adjoint linear operators. This has been an immensely successful and clarifying point of view in general, but $HC_\bullet(A)$ sticks out on a complete tangent: there is simply no natural place for it in this framework.

This paper arose as one more attempt to propose a solution to the difficulty: to find a natural triangulated category where $HC_\bullet(-)$ would be able to live with a certain level of comfort (and with all the standard corollaries such as the notion of cyclic homology with coefficients, the ability to compute cyclic homology by whatever resolution is convenient, not just the bar resolution, and so on).

In a sense, our attempt has been successful: we define a triangulated category that can serve as the natural "category of coefficients" for cyclic homology of an algebra A, and we prove the comparison theorem that shows that when the coefficients are trivial, the new definition of cyclic homology is equivalent to the old one. In fact, the algebra A enters into the construction only through the category $A\text{-}\mathsf{bimod}$ of A-bimodules; we also show how to generalize the construction so that $A\text{-}\mathsf{bimod}$ is replaced with a more general tensor abelian category $\mathcal{C}$.

From a different point of view, though, out attempt failed miserably: the correspondence $A \mapsto A_\#$, being thrown out of the window, immediately returns through the door in a new and "higher-level" disguise: it is now applied not to the algebra A, but to the tensor category $\mathcal{C} = A\text{-}\mathsf{bimod}$. Then in practice, the freedom to choose an arbitrary resolution to compute the derived functors leads, in our approach to $HC_\bullet(-)$, to complexes that are even larger than the original complex, and at some point the whole exercise starts to look pointless.

Still, we believe that all said and done, some point can be found, and some things are clarified in our approach; one such thing is, for instance, the version of Gauss–Manin connection for cyclic homology discovered by E. Getzler [Ge]. Moreover, we do propose a definition of cyclic homology that – makes sense for a general tensor category; and in some particular questions, even the computations can be simplified. As for the presence of the $A_{\#}$-construction, this might be in the nature of things, after all: not a bug in the theory, but a necessary feature. However, we leave it to the reader to be the judge.

The paper is organized as follows. In Section 1 we recall A. Connes's second definition of cyclic homology, which uses the cyclic category Λ; we also recall some facts about homology of small categories that we will need. We have tried to give only the absolute minimum; the reader not familiar with the material will have to consult the references. In Section 2 we introduce our main object: the notion of a *cyclic bimodule* over an associative algebra A, and the derived category of such bimodules. We also introduce cyclic homology $HC_{\bullet}(A, M)$ with coefficients in a cyclic bimodule M. In Section 3 we give a very short derivation of the Gauss–Manin connection; strictly speaking, the language of cyclic bimodules is not needed for this, but we believe that it shows more clearly what is really going on. In Section 4, we show how to replace the category $A\text{-}\mathsf{bimod}$ everywhere with a more general tensor abelian category $\mathcal{C}$. Section 5 is a postface, or a "discussion" (as is done in medical journals); we discuss some of the further things one might (and should) do with cyclic bimodules, and how to correct some deficiencies of the theory developed in Sections 2 and 4.

Acknowledgments.

In the course of this work, I have benefited greatly from discussions with A. Beilinson, E. Getzler, V. Ginzburg, A. Kuznetsov, N. Markarian, D. Tamarkin, and B. Tsygan. I am grateful to Northwestern Univeristy, where part of this work was done, and where some of the results were presented in seminars, with great indulgence from the audience toward the unfinished state of the work. And last but not least, it is a great pleasure and a great opportunity to dedicate this paper to Yuri Ivanovich Manin on his birthday. Besides all the usual things, I would like to stress that it is the book [GM1], and [GM2] to a lesser extent, that shaped the way we look at homological algebra today, at least "we" of my generation and of the Moscow school. Without Manin's decisive influence, this paper certainly would not have appeared (as in fact at least half of the papers I have written).

1 Recollection on cyclic homology.

We start by recalling, extremely briefly, A. Connes' approach to cyclic homology, which was originally introduced in [C2] (for detailed overviews, see, e.g., [L, Section 6] or [FT, Appendix]; a brief but complete exposition using the same language and notation as in this paper can be found in [Ka, Section 1]).

+,-.+ + +/ ! + * 1 2 /8 u # ? #
3 & ! ! 4 3#

+,-.+ + +/ ! + * < ! ! 1 2P -! ! 8 P
! # , 3 ! u t = ' 3#
(" * 8 > H ?! !

Curves of Low Degrees on Fano Varieties

Olivier Debarre

Mathematics Subject Classification codes (2010): 14C05, 14C30, 14C34, 14D20, 14E05, 14E08, 14E20, 14H10, 14J10, 14J30, 14J35, 14J45, 14J60, 14J70, 14M20, 14M22, 14N25

1 Introduction

We work over the field of complex numbers. If $X \subset \mathbf{P}^N$ is a smooth projective variety, we let $\mathscr{C}_d^g(X)$ be the (quasi-projective) subscheme of the Hilbert scheme of X corresponding to smooth, irreducible, genus-g, degree-d curves on X.

These schemes, and their compactifications $\overline{\mathscr{C}}_d^g(X)$, particularly for low d and g, have proved very useful in the study of the geometry and the period maps of some Fano varieties of low degrees.

In this informal survey article, we explain, mainly without proofs, what is known (or conjectured) in the following cases: cubic threefolds, Fano threefolds of degree 14 and index 1, cubic fourfolds, Fano varieties of degree 10, coindex 3, and dimensions 3, 4, or 5.

In the case of a smooth cubic threefold $X \subset \mathbf{P}^4$ (Sect. 2), Hilbert schemes of curves have long been used to parametrize useful subvarieties of the intermediate Jacobian $J(X)$. For example, the (smooth projective) surface $\mathscr{C}_1^0(X)$ parametrizing lines contained in X was an essential ingredient in the proofs by Clemens and Griffiths of the non-rationality of X and of the Torelli theorem (that $J(X)$ determines X). The Abel–Jacobi maps $\mathrm{aj} : \overline{\mathscr{C}}_d^g(X) \to J(X)$ have been since proved in some cases to be mrc fibrations (see Sect. 2 for definitions and results).

O. Debarre (✉)
Département de Mathématiques et Applications – CNRS UMR 8553, École Normale Supérieure, 45 rue d'Ulm, 75230 Paris cedex 05, France
e-mail: Olivier.Debarre@ens.fr

F. Bogomolov et al. (eds.), *Birational Geometry, Rational Curves, and Arithmetic*, DOI 10.1007/978-1-4614-6482-2_6,
© Springer Science+Business Media New York 2013

Curves on Fano threefolds can also be used to construct interesting birational correspondences. This is illustrated in Sect. 2.9, where it is explained that Fano threefolds of degree 14 and index 1 are birational to cubic threefolds. The situation is very rich and gives rise to a beautiful description of the period maps for these Fano threefolds.

In Sect. 3, we move on to smooth cubic fourfolds $Y \subset \mathbf{P}^5$, for which there are no intermediate Jacobians. Nevertheless, the mrc fibrations of $\overline{\mathscr{C}}^g_d(Y)$ should be substitutes for the Abel–Jacobi maps and should provide interesting information. In another direction, ideas of Mukai show how to construct *symplectic forms* on some moduli spaces of sheaves (Sect. 3.1). This can be used to prove that the space of lines $\mathscr{C}^0_1(Y)$ is a symplectic fourfold (Beauville and Donagi), whose Hodge structure is closely related to that of Y. This was the basis of Voisin's proof of the injectivity of the period map for smooth cubic fourfolds. This period map is now completely described by work of Looijenga and Laza and identifies the moduli space of smooth cubic fourfolds with an explicit open subset of the Baily–Borel compactification of the period domain (Sect. 3.2). Analogous results are available for the moduli space of smooth cubic threefolds (Sect. 3.3).

Mukai's construction was also shown to apply to Fano fourfolds X^4_{10} of degree 10 and index 2 (Example 3.5). In this case, the scheme $\overline{\mathscr{C}}^0_2(X^4_{10})$ leads to the construction of a symplectic fourfold called a *double EPW sextic* (Iliev and Manivel, O'Grady).

In Sect. 4, we study in some details the period maps of these Fano varieties of degree 10, in dimensions 3, 4, and 5. There is beautiful geometry at work here, with intriguing interplay with that of EPW sextics (work in progress with Iliev and Manivel).

2 Cubic Threefolds

The *maximal rationally connected fibration* (mrc fibration for short) of a smooth proper (complex) variety Y is a rational dominant map $\rho : Y \dashrightarrow R(Y)$ such that for $z \in R(Y)$ very general, the fiber $\rho^{-1}(z)$ is proper and rationally connected (two general points can be joined by a rational curve), and any rational curve in Y that meets $\rho^{-1}(z)$ is contained in $\rho^{-1}(z)$. The mrc fibration exists and is unique up to birational equivalence. The variety $R(Y)$ is called the reduction of Y.

Let $X \subset \mathbf{P}^4$ be a general (although some results are known for *any smooth* X) cubic hypersurface. The intermediate Jacobian

$$J(X) := H^{2,1}(X)^\vee / H_3(X, \mathbf{Z})$$

is a 5-dimensional principally polarized abelian variety. We let $\mathscr{X}_3$ be the (irreducible, 10-dimensional) moduli space for smooth cubic threefolds, and we let $\mathscr{A}_n$ be the moduli space for principally polarized abelian varieties of dimension n. Taking X to $J(X)$ defines the *period map* $\mathscr{X}_3 \to \mathscr{A}_5$. We have:

- For $g = 0$ or $d \leq 5$, $\mathscr{C}_d^g(X)$ is integral of dimension $2d$ ([7], Theorem 1.1);
- For $d \leq 5$, the *Abel–Jacobi map* $\mathrm{aj} : \overline{\mathscr{C}}{}_d^g(X) \to J(X)$ induces the mrc fibration ([6], Theorem 1.1);
- For $d \geq 4$, $\mathrm{aj} : \overline{\mathscr{C}}{}_d^0(X) \to J(X)$ is dominant with irreducible general fibers.

It is natural to ask whether the Abel–Jacobi map $\mathrm{aj} : \overline{\mathscr{C}}{}_d^0(X) \to J(X)$ is the mrc fibration for all $d \geq 4$ (i.e., are general fibers rationally connected?).

2.1 Lines

- $\mathscr{C}_1^0(X)$ (which parametrizes lines on X) is a smooth projective irreducible surface of general type;
- The image of $\mathrm{aj} : \mathscr{C}_1^0(X) \to J(X)$ is a surface S with minimal class $[\Theta]^3/3!$ and $S - S$ is a theta divisor Θ;
- aj induces an isomorphism between the Albanese variety of $\mathscr{C}_1^0(X)$ and $J(X)$ ([2]).

The second item yields a proof of Torelli: *the period map* $\mathscr{X}_3 \to \mathscr{A}_5$ *is injective.*

2.2 Conics

- $\overline{\mathscr{C}}{}_2^0(X)$ is a smooth projective irreducible fourfold;
- $\mathrm{aj} : \overline{\mathscr{C}}{}_2^0(X) \to J(X)$ is a $\mathbf{P}^2$-bundle over the surface S of Sect. 2.1 (a conic is uniquely determined by the plane that it spans and the residual line in X). The mrc fibration is therefore $\overline{\mathscr{C}}{}_2^0(X) \to S$.

2.3 Plane Cubics

- $\overline{\mathscr{C}}{}_3^1(X)$ is isomorphic to the Grassmannian $G(3,5)$;
- The Abel–Jacobi map is therefore constant.

2.4 Twisted Cubics

- $\mathrm{aj} : \mathscr{C}_3^0(X) \to J(X)$ is birational to a $\mathbf{P}^2$-bundle over a theta divisor Θ ([6], Sect. 4). The mrc fibration is therefore $\overline{\mathscr{C}}{}_3^0(X) \to \Theta$.

2.5 Elliptic Quartics

- $\mathrm{aj} : \mathscr{C}_4^1(X) \to J(X)$ is birational to a $\mathbf{P}^6$-bundle over the surface S of Sect. 2.1 ([6], Sect. 4.1). The mrc fibration is therefore $\overline{\mathscr{C}}{}_4^1(X) \to S$.

2.6 Normal Rational Quartics

- $\mathrm{aj} : \overline{\mathscr{C}}{}_4^0(X) \to J(X)$ is dominant and the general fiber is birational to X ([9], Theorem 5.2), hence unirational. It is therefore the mrc fibration.

2.7 Normal Elliptic Quintics

- $\mathscr{C}_5^1(X)$ is an irreducible tenfold;
- There is a factorization ([14], Theorem 5.6; [9], Theorem 3.2)

$$\mathrm{aj} : \overline{\mathscr{C}}{}_5^1(X) \xrightarrow{\alpha} M_X(2;0,2) \xrightarrow{\beta} J(X),$$

  where

  - $M_X(2;0,2)$ is some component of the moduli space of rank-2 stable vector bundles on X with Chern classes $c_1 = 0$ and $c_2 = 2$;
  - α is a $\mathbf{P}^5$-bundle over a dense open subset of $M_X(2;0,2)$;
  - β is birational (this is proved in [9] via ingenious geometrical constructions).

  In particular, $\overline{\mathscr{C}}{}_5^1(X) \to J(X)$ is therefore the mrc fibration.

This is seen as follows. The map α is obtained via the Serre construction: to $C \in \mathscr{C}_5^1(X)$, one associates a stable rank-2 vector bundle $\mathscr{E}_C$ on X with Chern classes $c_1 = 0$ and $c_2 = 2$ such that C is the zero-locus of a section of $\mathscr{E}_C(1)$.

The fibers of α are $\mathbf{P}(H^0(X, \mathscr{E}_C(1))) \simeq \mathbf{P}^5$; hence the Abel–Jacobi map factors through α.

According to Murre, the Chow group of algebraic 1-cycles of fixed degree on X modulo rational equivalence is canonically isomorphic to $J(X)$. The map β can then be defined directly as $\mathscr{E} \mapsto c_2(\mathscr{E})$.

2.8 Normal Elliptic Sextics

- $\mathscr{C}_6^1(X)$ is an irreducible 12-fold;
- $\mathrm{aj} : \overline{\mathscr{C}}{}_6^1(X) \to J(X)$ is the mrc fibration ([21], Theorem 2.1).

2.9 Fano Threefolds of Degree 14 and Index 1

There is a very interesting relationship between cubic threefolds and Fano threefolds of degree 14 and index 1 [9, 10, 14]. The latter are obtained as linear sections of $G(2,6) \subset \mathbf{P}^{14}$ by a $\mathbf{P}^9$.

Let $X \subset \mathbf{P}^4$ be a smooth cubic threefold, let C be general in $\mathscr{C}_5^1(X)$, and let $\pi : \tilde{X} \to X$ be its blow-up, with exceptional divisor E. We have

$$-K_{\tilde{X}} \underset{\text{lin}}{\equiv} -\pi^* K_X - E \underset{\text{lin}}{\equiv} 2\pi^* H - E.$$

This linear system induces a morphism $\tilde{X} \to \mathbf{P}^4$ which induces a small contraction φ onto the normalization $\bar{X}$ of its image. Its non-trivial fibers are the strict transforms of the 25 lines bisecant to C: the divisor E is φ-ample hence there is a flop

$$\chi : \tilde{X} \xrightarrow{\varphi} \bar{X} \xleftarrow{\varphi'} \tilde{X}',$$

where $\tilde{X}'$ is smooth projective and $\chi(E)$ is φ'-antiample.

We have $\rho(\tilde{X}') = 2$. Since the extremal ray generated by the classes of curves contracted by φ' has $K_{\tilde{X}'}$-degree 0 and $K_{\tilde{X}'}$ is not nef ($-K_{\tilde{X}'} \underset{\text{lin}}{\equiv} \varphi'^* \bar{H}$), the other extremal ray is $K_{\tilde{X}'}$-negative and defines a contraction $\pi' : \tilde{X}' \to X'$. One checks that:

- X' is a smooth Fano threefold of degree 14 and index 1, with Picard group generated by $H' := -K_{X'}$;
- π' is the blowup of a smooth elliptic quintic curve $C' \subset X'$, with exceptional divisor $E' \underset{\text{lin}}{\equiv} 5\varphi'^* \bar{H} - 3\chi_*(E)$ and $\chi^* \pi'^* H' \underset{\text{lin}}{\equiv} 7H - 4E$.

Conversely, given a general elliptic quintic curve C' in a smooth Fano threefold X' of degree 14 and index 1, one can reverse the construction above and obtain a cubic threefold X with a quintic curve $C \subset X$. In other words, if $\mathscr{C}_5^1(\mathscr{X}_3)$ denotes the moduli space of all pairs (X, C) as above, and $\mathscr{C}_5^1(\mathscr{X}_{14})$ the moduli space of all pairs (X', C'), we have a birational isomorphism

$$\mathscr{C}_5^1(\mathscr{X}_3) \dashrightarrow \mathscr{C}_5^1(\mathscr{X}_{14})$$

between (irreducible, 20-dimensional) varieties. We have the following two properties.

- *The intermediate Jacobians of X and X' are isomorphic* (this holds because $J(X) \times J(C) \simeq J(X') \times J(C')$, $J(X')$ is not a product, and $J(X)$ and $J(X')$ have the same dimension). Since we have Torelli for cubic threefolds (Sect. 2.1), the cubic obtained from a pair (X', C') only depends on X'.
- *The variety X' only depends on the vector bundle $\mathscr{E}_C$* defined in Sect. 2.7 (Kuznetsov proves in [10] that if $\mathscr{S}$ is the rank-2 tautological vector bundle on $G(2,6)$, the variety $\mathbf{P}(\mathscr{S}|_{X'})$ is obtained by a flop of $\mathbf{P}(\mathscr{E}_C)$).

So, if $\mathscr{X}_{14}$ is the moduli space for smooth Fano threefolds of degree 14 and index 1, and if $\mathscr{M}_{\mathscr{X}_3}(2;0,2)$ is the moduli space of pairs $(X,\mathscr{E})$, with $[\mathscr{E}] \in M_X(2;0,2)$ (see Sect. 2.7), we get a commutative diagram:

$$\begin{tikzcd}
\text{dim. 20} & \mathscr{C}^1_5(\mathscr{X}_3) \arrow[rr, dashed, leftrightarrow] \arrow[d, "\mathbf{P}^5-\text{bundle}"'] & & \mathscr{C}^1_5(\mathscr{X}_{14}) \arrow[d] & \text{dim. 20} \\
\text{dim. 15} & \mathscr{M}_{\mathscr{X}_3}(2;0,2) \arrow[rr, "\gamma", "\text{isom.}"'] \arrow[d] & & \mathscr{X}_{14} \arrow[lld, "\delta"] \arrow[ld, "\text{period map}"] & \text{dim. 15} \\
\text{dim. 10} & \mathscr{X}_3 \arrow[rd, hook, "\text{period map}"'] & & & \\
 & & \mathscr{A}_5. & &
\end{tikzcd}$$

The map γ is actually an isomorphism of stacks and the fiber of δ (between stacks) at $[X]$ is the (quasi-projective) moduli space of stable rank-2 vector bundles $\mathscr{E}$ such that $c_1(\mathscr{E}) = 0$ and $H^1(X,\mathscr{E}(-1)) = 0$ (instanton bundles), an open subset of $J(X)$ ([10], Theorem 2.9).

3 Cubic Fourfolds

Let $Y \subset \mathbf{P}^5$ be a general cubic hypersurface.

The variety $\overline{\mathscr{C}}^0_d(Y)$ is integral of dimension $3d+1$ ([5], Proposition 2.4). To study these varieties, one could use, instead of the Abel–Jacobi map, their mrc fibration [5]. Here is what is known, or conjectured:

- $\mathscr{C}^0_1(Y)$ is a symplectic fourfold [1] hence is its own reduction;
- The map $\overline{\mathscr{C}}^0_2(Y) \dashrightarrow \mathscr{C}^0_1(Y)$ given by residuation is a $\mathbf{P}^3$-bundle so this is the mrc fibration;
- $\overline{\mathscr{C}}^0_3(Y)$ is uniruled: a general cubic curve lies on a unique cubic surface and moves in a 2-dimensional linear system on it; so its reduction has dimension ≤ 8;
- There is a map $\overline{\mathscr{C}}^0_4(Y) \dashrightarrow \mathscr{J}(Y)$ (where $\mathscr{J}(Y) \to (\mathbf{P}^5)^\vee$ is the relative intermediate Jacobian of smooth hyperplane sections of Y) whose general fibers are these cubic threefolds (Sect. 2.6) hence are unirational; moreover, $\mathscr{J}(Y)$ should be non-uniruled (see also Example 3.4 below), so this should be the mrc fibration;
- For $d \geq 5$ odd, there is ([5], Theorem 1.2; [11]) a holomorphic 2-form on (a smooth non-singular model of) $\mathscr{C}^0_d(Y)$ which is *non-degenerate at a general point.* In particular, $\overline{\mathscr{C}}^0_d(Y)$ is not uniruled.

3.1 Constructing Symplectic Forms on Moduli Spaces

Mukai proved in 1984 that the moduli space $\mathscr{M}$ of simple sheaves on a K3 or abelian surface carries a closed non-degenerate holomorphic 2-form: the tangent space to $\mathscr{M}$ at a point $[\mathscr{F}]$ representing a simple sheaf $\mathscr{F}$ on a smooth projective variety Z is isomorphic to $\mathrm{Ext}^1(\mathscr{F},\mathscr{F})$. The Yoneda coupling

$$\mathrm{Ext}^1(\mathscr{F},\mathscr{F}) \times \mathrm{Ext}^1(\mathscr{F},\mathscr{F}) \longrightarrow \mathrm{Ext}^2(\mathscr{F},\mathscr{F})$$

is skew-symmetric whenever $[\mathscr{F}]$ is a *smooth* point of $\mathscr{M}$. When Z is a symplectic surface S with a symplectic holomorphic form $\omega \in H^0(S,\Omega_S^2)$, Mukai composes the Yoneda coupling with the map

$$\mathrm{Ext}^2(\mathscr{F},\mathscr{F}) \xrightarrow{\mathrm{Tr}} H^2(S,\mathscr{O}_S) \xrightarrow{\cup\omega} H^2(S,\Omega_S^2) = \mathbf{C},$$

and this defines the symplectic structure on the smooth locus of $\mathscr{M}$.

Over an n-dimensional variety Z such that $h^{q,q+2}(Z) \neq 0$ for some integer q, we:

- Pick a non-zero element $\omega \in H^{n-q-2}(Z,\Omega_Z^{n-q})$;
- Use the exterior power $\mathrm{At}(\mathscr{F})^{\wedge q} \in \mathrm{Ext}^q(\mathscr{F},\mathscr{F}\otimes\Omega_Z^q)$ of the Atiyah class[1] $\mathrm{At}(\mathscr{F}) \in \mathrm{Ext}^1(\mathscr{F},\mathscr{F}\otimes\Omega_Z^1)$.

We then define

$$\mathrm{Ext}^2(\mathscr{F},\mathscr{F}) \xrightarrow{\mathrm{At}(\mathscr{F})^{\wedge q}\circ\bullet} \mathrm{Ext}^{q+2}(\mathscr{F},\mathscr{F}\otimes\Omega_Z^q) \xrightarrow{\mathrm{Tr}}$$
$$H^{q+2}(Z,\Omega_Z^q) \xrightarrow{\cup\omega} H^n(Z,\Omega_Z^n) \simeq \mathbf{C}.$$

Composing the Yoneda coupling with this map provides a closed (possibly degenerate) 2-form on the smooth locus of the moduli space.

Example 3.3. Let $Y \subset \mathbf{P}^5$ be a smooth cubic fourfold. Since $h^{1,3}(Y) = 1$, the construction provides a (unique) 2-form on the smooth loci of moduli spaces of sheaves on Y. Kuznetsov and Markushevich use this construction in a roundabout

---

[1] Let $\Delta : Z \to Z\times Z$ be the diagonal embedding and let $\Delta(Z)^{(2)} \subset Z\times Z$ be the closed subscheme defined by the sheaf of ideals $\mathscr{I}^2_{\Delta(Z)}$. Since $\mathscr{I}_{\Delta(Z)}/\mathscr{I}^2_{\Delta(Z)} \simeq \Omega_Z$; we have an exact sequence

$$0 \to \Delta_*\Omega_Z \to \mathscr{O}_{\Delta(Z)^{(2)}} \to \Delta_*\mathscr{O}_Z \to 0.$$

If $\mathscr{F}$ is a locally free sheaf on Z, we obtain an exact sequence

$$0 \to \mathscr{F}\otimes\Omega_Z \to p_{1*}(p_2^*(\mathscr{F}\otimes\mathscr{O}_{\Delta(Z)^{(2)}})) \to \mathscr{F} \to 0,$$

hence an extension class $\mathrm{At}_{\mathscr{F}} \in \mathrm{Ext}^1(\mathscr{F},\mathscr{F}\otimes\Omega_Z)$. The same construction can be extended to any coherent sheaf on Z by working in the derived category (Illusie).

way to produce a symplectic structure on the (smooth) fourfold $\mathscr{C}_1^0(Y)$ of lines $L \subset Y$ (originally constructed by Beauville and Donagi by a deformation argument; note that the simple-minded idea to look at sheaves of the form $\mathscr{O}_L$ does not work).

Example 3.4. Let $Y \subset \mathbf{P}^5$ be a smooth cubic fourfold. Let $\mathscr{N}_Y$ be the (quasi-projective) moduli space of sheaves on Y of the form $i_*\mathscr{E}$, where $i : X \to Y$ is a non-singular hyperplane section of Y and $[\mathscr{E}] \in M_X(2;0,2)$. By Sect. 2.7, $\mathscr{N}_Y$ is a torsor under the (symplectic) relative intermediate Jacobian $\mathscr{J}(Y)$ of smooth hyperplane sections of Y. The Donagi–Markman symplectic structure ([4], 8.5.2) on $\mathscr{J}(Y)$ induces a symplectic structure on $\mathscr{N}_Y$ which should be the same as the Kuznetsov–Markushevich structure ([11, 15], Theorem 7.3 and Remark 7.5). Note that since we do not know whether the Donagi–Markman symplectic form on $\mathscr{J}(Y)$ extends to a smooth compactification, we cannot deduce that $\mathscr{J}(Y)$ is not uniruled.

Example 3.5. Let X_{10}^4 be a smooth Fano fourfold obtained by intersecting the Grassmannian $G(2,5)$ in its Plücker embedding with a general hyperplane and a general quadric (see Sect. 4). We have $h^{1,3}(X_{10}^4) = 1$. The Hilbert scheme $\overline{\mathscr{C}}_2^0(X_{10}^4)$ of (possibly degenerate) conics in X_{10}^4 is smooth [8]; hence it is endowed, by the construction above, with a canonical global holomorphic 2-form. Since $\overline{\mathscr{C}}_2^0(X_{10}^4)$ has dimension 5, this form must be degenerate.

There is a naturally defined morphism

$$\overline{\mathscr{C}}_2^0(X_{10}^4) \to \mathbf{P}(\mathrm{I}_{X_{10}^4}(2)) \simeq \mathbf{P}^5$$

whose image is an *Eisenbud–Popescu–Walter* (EPW for short) *sextic hypersurface* $Z_{X_{10}^4} \subset \mathbf{P}^5$ (see [16] for the definition of these sextics). In the Stein factorization

$$\overline{\mathscr{C}}_2^0(X_{10}^4) \to Y_{X_{10}^4} \to Z_{X_{10}^4},$$

the projective variety $Y_{X_{10}^4}$ is a smooth fourfold over which $\overline{\mathscr{C}}_2^0(X_{10}^4)$ is (essentially) a smooth fibration in projective lines. The 2-form on $\overline{\mathscr{C}}_2^0(X_{10}^4)$ thus descends to $Y_{X_{10}^4}$ and makes $Y_{X_{10}^4}$ into a holomorphic symplectic fourfold [8, 16] called a *double EPW sextic.*

3.2 Periods for Cubic Fourfolds

Let $Y \subset \mathbf{P}^5$ be a smooth cubic fourfold. Since Y has even dimension, it has no intermediate Jacobian but still an interesting Hodge structure “of K3-type”:

$$\begin{array}{lccccc} H^4(Y,\mathbf{C})_{\mathrm{prim}} = & H^{1,3}(Y) & \oplus & H^{2,2}(Y)_{\mathrm{prim}} & \oplus & H^{3,1}(Y) \\ \text{dimensions:} & 1 & & 20 & & 1 \end{array}$$

with period domain

$$\mathscr{D}^{20} = \{[\omega] \in \mathbf{P}^{21} \mid Q(\omega,\omega) = 0,\ Q(\omega,\bar{\omega}) > 0\},$$

a 20-dimensional bounded symmetric domain of type IV (where Q is the (non-degenerate) intersection form on $H^4(Y,\mathbf{C})$). Let $\mathscr{Y}_3$ be the (irreducible, 20-dimensional) moduli space for smooth cubic fourfolds. We get a period map

$$\mathscr{Y}_3 \to \mathscr{D}^{20}/\Gamma,$$

where Γ is an explicit discrete arithmetic group. Voisin proved that it is injective and its (open) image was determined in [12].

3.3 Periods for Cubic Threefolds

One can construct another period map for cubic threefolds (compare with Sect. 2): to such a cubic $X \subset \mathbf{P}^4$, we associate the cyclic triple cover $Y_X \to \mathbf{P}^4$ branched along X. It is a cubic fourfold; hence this construction defines a map $\mathscr{X}_3 \to \mathscr{Y}_3 \to \mathscr{D}^{20}/\Gamma$.

Because of the presence of an automorphism of Y_X of order 3, we can restrict the image and define a period map (Allcock–Carlson–Toledo)

$$\mathscr{X}_3 \to \mathscr{D}^{10}/\Gamma',$$

where

$$\mathscr{D}^{10} := \{\omega \in \mathbf{P}^{10} \mid Q(\omega,\bar{\omega}) < 0\} \simeq \mathbf{B}^{10}$$

and Γ' is a discrete arithmetic group. Again, it is an isomorphism onto an explicitly described open subset of $\mathscr{D}^{10}/\Gamma'$.

4 Fano Varieties of Degree 10

Let V_5 be a 5-dimensional complex vector space. We define, for $k \in \{3,4,5\}$, a smooth, degree-10, coindex-3, Fano k-fold by

$$X_{10}^k := G(2,V_5) \cap \mathbf{P}^{k+4} \cap \Omega \subset \mathbf{P}(\wedge^2 V_5),$$

where $\mathbf{P}^{k+4}$ is a general $(k+4)$-plane and Ω a general quadric. Let $\mathscr{X}_{10}^k$ be the moduli stack for smooth varieties of this type.

Enriques proved that all (smooth) X_{10}^3, X_{10}^4, and X_{10}^5 are unirational. A general X_{10}^3 is not rational, whereas all (smooth) X_{10}^5 are rational [19]. Some smooth X_{10}^4 are rational ([17, 18]), but the rationality of a general X_{10}^4 is an open question.

We have

$$I_{X_{10}^k}(2) \simeq \mathbf{C}\Omega \oplus V_5,$$

where V_5 corresponds to the rank-6 *Plücker quadrics* $\omega \mapsto \omega \wedge \omega \wedge v$ ($v \in V_5$). In the 5-plane $\mathbf{P}(I_{X_{10}^k}(2))$, the degree-$(k+5)$ hypersurface corresponding to singular quadrics decomposes as

$$(k-1)\mathbf{P}(V_5) + Z^\vee_{X_{10}^k},$$

where $Z^\vee_{X_{10}^k} \subset \mathbf{P}(I_{X_{10}^k}(2))$ is an *EPW sextic* ([8], Sect. 2.2; this is indeed the projective dual of the sextic defined in Example 3.5 when $k=4$).

If $\mathscr{E}\mathscr{P}\mathscr{W}$ is the (irreducible, 20-dimensional) moduli space of EPW sextics, we get morphisms

$$\mathrm{epw}^k : \mathscr{X}_{10}^k \to \mathscr{E}\mathscr{P}\mathscr{W},$$

which are *dominant* ([8], Corollary 4.17). Note that

$$\begin{aligned}
\dim(\mathscr{X}_{10}^5) &= 25,\\
\dim(\mathscr{X}_{10}^4) &= 24,\\
\dim(\mathscr{X}_{10}^3) &= 22.
\end{aligned}$$

Proposition 4.1 (Debarre–Iliev–Manivel). *For a general EPW sextic Z, with projective dual $Z^\vee$, the fiber $(\mathrm{epw}^3)^{-1}([Z])$ is isomorphic to the smooth surface* $\mathrm{Sing}(Z^\vee)$.

4.1 Gushel Degenerations

The link between these Fano varieties of various dimensions can be made through a construction of Gushel analogous to what we did with cubics in Sect. 3.3.

Let $CG \subset \mathbf{P}(\mathbf{C} \oplus \wedge^2 V_5)$ be the cone, with vertex $v = \mathbf{P}(\mathbf{C})$, over the Grassmannian $G(2,V_5)$. Intersect CG with a general quadric $\Omega \subset \mathbf{P}(\mathbf{C} \oplus \wedge^2 V_5)$ and a linear space $\mathbf{P}^{k+4}$ to get a Fano variety T^k of dimension k. There are two cases:

- Either $v \notin \mathbf{P}^{k+4}$, in which case T^k is isomorphic to the intersection X_{10}^k of $G(2,V_5)$ with the projection of $\mathbf{P}^{k+4}$ to $\mathbf{P}(\wedge^2 V_5)$ and a quadric
- Or $v \in \mathbf{P}^{k+4}$, in which case $\mathbf{P}^{k+4}$ is a cone over a $\mathbf{P}^{k+3} \subset \mathbf{P}(\wedge^2 V_5)$ and T^k is a double cover X_G^k of $G(2,V_5) \cap \mathbf{P}^{k+3}$ branched along its intersection X_{10}^{k-1} with a quadric

The second case is a specialization of the first, and the EPW sextics $Z_{X_{10}^k}$ from the first case degenerate to the sextics $Z_{X_G^k}$ from the second case. Moreover, in the second case, *the sextics $Z_{X_G^k}$ and $Z_{X_{10}^{k-1}}$ are the same.*

Let $\mathscr{X}^k_G$ be the moduli stack for smooth varieties of type X^k_{10} and their Gushel degenerations. The Gushel constructions therefore yield morphisms $\mathscr{X}^k_{10} \to \mathscr{X}^{k+1}_G$ such that the diagrams

$$\begin{tikzcd}
\mathscr{X}^k_{10} \arrow[r] \arrow[dr, "\mathrm{epw}^k"'] & \mathscr{X}^{k+1}_G \arrow[d, "\mathrm{epw}^{k+1}_G"] \\
& \mathscr{EPW}
\end{tikzcd}$$

commute.

We can perform a "double Gushel construction" as follows. Let $CCG \subset \mathbf{P}(\mathbf{C}^2 \oplus \wedge^2 V_5)$ be the cone, with vertex $L = \mathbf{P}(\mathbf{C}^2)$, over the Grassmannian $G(2,V_5)$. Intersect CCG with a general quadric $\Omega \subset \mathbf{P}(\mathbf{C}^2 \oplus \wedge^2 V_5)$ and a codimension-2 linear space $\mathbf{P}^9$ to get a Fano variety T of dimension 5.

- If $L \cap \mathbf{P}^9 = \varnothing$, the variety T is smooth of type X^5_{10}.
- If $L \subset \mathbf{P}^9$, in which case $\mathbf{P}^9$ is a cone over a $\mathbf{P}^7 \subset \mathbf{P}(\wedge^2 V_5)$, the corresponding variety T_0 meets L at two points p and q, and the projection from L induces a rational conic bundle $T_0 \dashrightarrow W^4_5 := G(2,V_5) \cap \mathbf{P}^7$, undefined at p and q, whose discriminant locus is a threefold X^3_{10}. Blowing up these two points, we obtain a conic bundle $\hat{T}_0 \to W^4_5$ with two disjoint sections corresponding to the two exceptional divisors. These two sections trivialize the canonical double étale cover $\tilde{X}^3_{10} \to X^3_{10}$ of the discriminant.

We call the singular variety T_0 a double-Gushel degeneration. If we let $\overline{\mathscr{X}}^5_{GG}$ be the moduli stack for smooth varieties of type X^5_{10} and their Gushel and double-Gushel degenerations, we obtain a commutative diagram:

$$\begin{tikzcd}
\mathscr{X}^3_{10} \arrow[r] \arrow[dr, "\mathrm{epw}^3"'] & \mathscr{X}^4_G \arrow[r] \arrow[d, "\mathrm{epw}^4_G"] & \mathscr{X}^5_{GG} \arrow[dl, "\mathrm{epw}^5_{GG}"] \\
& \mathscr{EPW}. &
\end{tikzcd}$$

4.2 Period Maps

Both $H^3(X^3_{10},\mathbf{Q})$ and $H^5(X^5_{10},\mathbf{Q})$ have dimension 20 and carry Hodge structures of weight 1. This gives rise to period maps

$$\wp^3 : \mathscr{X}^3_{10} \to \mathscr{A}_{10} \quad \text{and} \quad \wp^5 : \mathscr{X}^5_{10} \to \mathscr{A}_{10}.$$

By [3], the general fibers of $\wp^3$ are unions of smooth proper surfaces that come in pairs:

- $\mathscr{F}_{X^3_{10}}$, isomorphic to the quotient of $\overline{\mathscr{C}}^0_2(X^3_{10})$ by a fixed-point-free involution, and also to $\mathrm{Sing}(Z^\vee_{X^3_{10}})$.
- $\mathscr{F}^\star_{X^3_{10}}$, the analogous surface for any "line-transform" of X^3_{10}.

In particular, by Proposition 4.1, there is a commutative diagram:

$$\begin{tikzcd}
\mathscr{X}^3 \arrow[r, "\mathrm{epw}^3"] \arrow[dr, "\wp^3"'] & \mathscr{E}\mathscr{P}\mathscr{W} \arrow[r] \arrow[d, "\wp"] & \mathscr{E}\mathscr{P}\mathscr{W}/\text{duality} \arrow[dl, "\bar{\wp}"] \\
& \mathscr{A}_{10}, &
\end{tikzcd}$$

where the map $\bar{\wp}$ is generically finite (presumably birational) onto its image. Since $J(X^3_{10})$ is isomorphic to the Albanese variety of the surface $\overline{\mathscr{C}}^0_2(X^3_{10})$ [13], the map $\wp$ is defined by sending the class of a general EPW sextic Z to the Albanese variety of the canonical double cover of its singular locus (the resulting principally polarized abelian varieties are isomorphic for Z and $Z^\vee$).

Theorem 4.2. *The image of* $\wp^3$ *and the image of* $\wp^5$ *have same closures.*

Proof. This is proved using a double Gushel degeneration (see Sect. 4.1): with the notation above, one proves that intermediate Jacobian $J(\hat{T}_0)$ is still a 10-dimensional principally polarized abelian variety which is a limit of intermediate Jacobians of Fano fivefolds of type X^5_{10}, hence belongs to the closure of $\mathrm{Im}(\wp^5)$.

Since, by Lefschetz theorem, we have $H^5(W^4_5,\mathbf{Q}) = H^3(W^4_5,\mathbf{Q}) = 0$, the following lemma implies that the intermediate Jacobians $J(\hat{T}_0)$ and $J(X^3_{10})$ are isomorphic. This proves already that the image of $\wp^3$ is contained in the closure of the image of $\wp^5$.

Lemma 4.3. *Let* T *and* W *be smooth projective varieties. Assume*

$$H^k(W,\mathbf{Q}) = H^{k-2}(W,\mathbf{Q}) = 0.$$

Let $\pi : T \to W$ *be a conic bundle with smooth irreducible discriminant divisor* $X \subset W$. *Assume further that* $\pi^{-1}(X)$ *is reducible. There is an isomorphism of polarized Hodge structures*

$$H^k(T,\mathbf{Z})/\mathrm{tors} \simeq H^{k-2}(X,\mathbf{Z})/\mathrm{tors}.$$

To finish the proof of the theorem, we prove, by computing the kernel of its differential, that the fibers of $\wp^5$ have dimension at least 5. ☐

Question 4.4. Is $\wp^5$ equal to $\wp \circ \mathrm{epw}^5$?

Question 4.5. Can one extend $\wp$ to the GIT compactification of $\mathscr{E}\mathscr{P}\mathscr{W}$ studied in [20], with values in a suitable compactification of $\mathscr{A}_{10}$?

Acknowledgements O. Debarre is part of the project VSHMOD-2009 ANR-09-BLAN-0104-01. These are notes from a talk given at the conference "Geometry Over Non-Closed Fields" funded by the Simons Foundation, whose support is gratefully acknowledged.

References

1. BEAUVILLE, A. AND DONAGI, R., *La variété des droites d'une hypersurface cubique de dimension 4*, C. R. Acad. Sci. Paris Sér. I Math. **301**, 703–706, (1985).
2. CLEMENS, C. H. AND GRIFFITHS, P. A., *The intermediate Jacobian of the cubic threefold*, Ann. of Math. **95**, 281–356, (1972).
3. DEBARRE, O., ILIEV, A., AND MANIVEL, L., *On the period map for prime Fano threefolds of degree 10*, J. Algebraic Geom. **21**, 21–59, (2012).
4. DONAGI, R. AND MARKMAN, E., *Spectral covers, algebraically completely integrable Hamiltonian systems, and moduli of bundles*, In: Francaviglia, M. (ed.) et al., *Integrable systems and quantum groups,* CIME Lectures, Italy, June 14–22, 1993. Lect. Notes Math. **1620**, 1–119, Springer-Verlag, Berlin, 1996.
5. DE JONG, A. J. AND STARR, J., *Cubic fourfolds and spaces of rational curves*, Illinois J. Math. **48**, 415–450, (2004).
6. HARRIS, J., ROTH, M., AND STARR, J., *Abel-Jacobi maps associated to smooth cubic threefolds*, `arXiv:math/0202080`, (2002).
7. HARRIS, J., ROTH, M., AND STARR, J., *Rational curves on hypersurfaces of low degree*, J. Reine Angew. Math. **571**, 73–106, (2004).
8. ILIEV, A. AND MANIVEL, L., *Fano manifolds of degree 10 and EPW sextics*, Ann. Sci. École Norm. Sup. **44** 393–426, (2011).
9. ILIEV, A. AND MARKUSHEVICH, D., *The Abel-Jacobi map for a cubic threefold and periods of Fano threefolds of degree 14*, Doc. Math. **5** 23–47, (2000).
10. KUZNETSOV, A., *Derived category of a cubic threefold and the variety V_{14}*, (in Russian) Tr. Mat. Inst. Steklova **246** (2004), Algebr. Geom. Metody, Svyazi i Prilozh., 183–207; English translation in Proc. Steklov Inst. Math. **246**, 171–194, (2004).
11. KUZNETSOV, A. AND MARKUSHEVICH, D., *Symplectic structures on moduli spaces of sheaves via the Atiyah class*, J. Geom. Phys. **59**, 843–860, (2009).
12. LAZA, R., *The moduli space of cubic fourfolds via the period map*, Ann. of Math. **172**, 673–711, (2010).
13. LOGACHEV, D., *Fano threefolds of genus 6*, Asian J. Math. 16, 515–560, (2012).
14. MARKUSHEVICH, D. AND TIKHOMIROV, A., *The Abel-Jacobi map of a moduli component of vector bundles on the cubic threefold*, J. Algebraic Geom. **10**, 37–62, (2001).
15. MARKUSHEVICH, D. AND TIKHOMIROV, A., *Symplectic structure on a moduli space of sheaves on the cubic fourfold*, Izv. Ross. Akad. Nauk Ser. Mat. **67**, 131–158, (2003).
16. O'GRADY, K., *Irreducible symplectic 4-folds and Eisenbud-Popescu-Walter sextics*, Duke Math. J. **134**, 99–137, (2006).
17. PROKHOROV, YU., *Rationality constructions of some Fano fourfolds of index 2*, Moscow University Mathematics Bulletin **48**, 32–35, (1993).
18. ROTH, L., *Algebraic varieties with canonical curve sections*, Ann. Mat. Pura Appl. (4) **29**, 91–97, (1949).
19. SEMPLE, J. *On quadric representations of the lines of four-dimensional space*, Proc. London Math. Soc. **30**, 500–512, (1930).
20. O'GRADY, K., *Moduli of double EPW-sextics*, `arXiv:1111.1395,` (2011).
21. VOISIN, C., *Abel-Jacobi map, integral Hodge classes and decomposition of the diagonal*, J. Algebraic Geom. **22**, 141–174, (2013).

(%

- Algebra, Arithmetic, and Geometry - Volume II... (2009).pdf
  Pág. 23(35)
  Cyclic Homology with Coefficients
  Pág. 33(45) 3
  Gauss-Manin connection.
- Geometry over Nonclosed Field (2017).pdf
- Birational Geometry, Rational Curves, and Arithmetic... (2013).pdf
  Pág. 133(141)
  Curves of Low Degrees on
  Fano Varieties

Comparar:

Geometry over nonclosed field...pdf
Pág. 174(180) 4
Del Pezzo Surfaces

Pezzo surfaces with a view towards
Brauer-Severi varieties.

Look: Pág. 187(193)

8. Del Pezzo Surfaces of Small Degree.

(8.1) Birationality to Brauer-Severi Surfaces
(Severi)

↳ Comparar com o livro:

Birational Geometry, Rational Curves, and Arithmetic.(2013).pdf

Pág. 93(101)

Birational Geometry via Moduli Spaces

and

Pag. 133(141)

Curves of Low Degrees on Fano Varieties.

3! !
+,-.+ + +/ ! + * 1 2 /8 u # ? #
3 & ! ! 4 3#
(* * H 66 !

4 Del Pezzo Surfaces

For the remainder of this article, we study del Pezzo surfaces with a view towards Brauer–Severi varieties. Most, if not all, results of these sections are known in some form or another to the experts. However, our more geometric approach, as well as some of the proofs, are new.

Let us first recall some classical results about del Pezzo surfaces, and refer the reader to [35, Chap. IV] or the surveys [7, 41, 47] for details, proofs, and references. For more results about the classification of geometrically rational surfaces, see [25, 34].

Definition 4.1 A *del Pezzo surface* is a smooth and proper variety X of dimension two over a field k such that ω_X^{-1} is ample. The *degree* of a del Pezzo surface is the self-intersection number of ω_X.

In arbitrary dimension, smooth and proper varieties X over k with ample ω_X^{-1} are called *Fano varieties*. As discussed in Example 3.7, Fano varieties of dimension one over k are the same as Brauer–Severi curves over k.

4.1 Geometry

The degree d of a del Pezzo surface X over a field k satisfies $1 \leq d \leq 9$. Set $\overline{X} := X_{\overline{k}}$. We will say that X is *of product type* if

$$\overline{X} \cong \mathbb{P}^1_{\overline{k}} \times \mathbb{P}^1_{\overline{k}},$$

in which case we have $d = 8$. If X is not of product type, then there exists a birational morphism

$$\overline{f} : \overline{X} \to \mathbb{P}^2_{\overline{k}}$$

that is a blow-up of $(9 - d)$ closed points $P_1, \ldots, P_{9-d}$ in general position, i.e., no 3 of them lie on a line, no 6 of them lie on a conic, and there is no cubic through all these points having a double point in one of them. In particular, if $d = 9$, then $\overline{f}$ is an isomorphism and X is a Brauer–Severi surface over k.

4.2 Arithmetic

By the previous discussion and Lemma 2.13, the *Néron–Severi rank* of a del Pezzo surface X of degree d over k satisfies

$$1 \;\leq\; \rho(X) \;:=\; \text{rank}\,\text{Pic}(X) \;=\; \text{rank}\,\text{Pic}_{(X/k)(\text{fppf})}(k) \;\leq\; 10 - d,$$

and $\rho(X_{\overline{k}}) = 10 - d$.

The following result about geometrically rational surfaces allows using methods from Galois theory even if the ground field k is not perfect. This result is particularly useful in proofs, see also the discussion in [47, Sect. 1.4]. In particular, it applies to del Pezzo surfaces.

Theorem 4.2 (Coombes+ε) *Let X be a smooth and proper variety over k such that $X_{\overline{k}}$ is birational to $\mathbb{P}^2_{\overline{k}}$. Then,*

(1) $X_{k^{\mathrm{sep}}}$ is birationally equivalent to $\mathbb{P}^2_{k^{\mathrm{sep}}}$ via a sequence of blow-ups in points in k^{sep}-rational points and their inverses.
(2) The natural map $\mathrm{Pic}_X(k^{\mathrm{sep}}) \to \mathrm{Pic}_X(\overline{k})$ is an isomorphism.

Proof Assertion (1) is the main result of [11]. Clearly, assertion (2) holds for projective space over any field. Next, let Y be a variety that is smooth and proper over k^{sep}, $\widetilde{Y} \to Y$ be the blow-up of a k^{sep}-rational point, and let $E \subset \widetilde{Y}$ be the exceptional divisor. Then, $\mathrm{Pic}_{\widetilde{Y}}(K) = \mathrm{Pic}_Y(K) \oplus \mathbb{Z} \cdot E$ for $K = k^{\mathrm{sep}}$, as well as for $K = \overline{k}$. Using (1) and these two observations, assertion (2) follows. □

We will also need the following useful observation, due to Lang [33] and Nishimura [39], which implies that having a k-rational point is a birational invariant of smooth and proper varieties over k. We refer to [47, Sect. 1.2] for details and proof.

Lemma 4.3 (Lang–Nishimura) *Let $X \dashrightarrow Y$ be a rational map of varieties over k, such that X is smooth over k, and such that Y is proper over k. If X has a k-rational point, then so has Y.* □

Moreover, we have already seen that a Brauer–Severi variety P over k is isomorphic to projective space over k if and only if P has a k-rational point, and we refer the interested reader to [14] for an algorithm to decide whether a Brauer–Severi surface has a k-rational point. In Definition 2.8, we defined the Amitsur group and showed its birational invariance in Proposition 2.10. Using Iskovskih's classification [25] of geometrically rational surfaces, we obtain the following list and refer to [10, Proposition 5.2] for details and proof.

Theorem 4.4 (Colliot-Thélène–Karpenko–Merkurjev) *Let X be a smooth and proper variety over a perfect field k such that $X_{\overline{k}}$ is birationally equivalent to $\mathbb{P}^2_{\overline{k}}$. Then, $\mathrm{Am}(X)$ is one of the following groups*

$$0, \quad \mathbb{Z}/2\mathbb{Z}, \quad (\mathbb{Z}/2\mathbb{Z})^2, \quad \textit{and} \quad \mathbb{Z}/3\mathbb{Z}.$$

We will see explicit examples of all these groups arising as Amitsur groups of del Pezzo surfaces in the next sections.

We now introduce another important invariant. Namely, if G_k denotes the absolute Galois group of k, and $H \subseteq G_k$ is a closed subgroup, then we consider for a smooth and projective variety X over k the group cohomology

$$H^1\left(H, \mathrm{Pic}_{X/k}(k^{\mathrm{sep}})\right),$$

which is an abelian torsion group. If $b_1(X) = 0$, then $\mathrm{Pic}_{X/k}(k^{\mathrm{sep}})$ is finitely generated by Lemma 2.13 and then, $H^1(H, \mathrm{Pic}_{X/k}(k^{\mathrm{sep}}))$ is a finite abelian group. Moreover, if $X_{k^{\mathrm{sep}}}$ is a rational surface, then $\mathrm{Br}'(X_{k^{\mathrm{sep}}}) = 0$ (see, for example, [35, Theorem 42.8] or [36]) and an appropriate Hochschild–Serre spectral sequence yields an exact sequence

$$0 \to \mathrm{Br}'(X)/\mathrm{Br}(k) \xrightarrow{\alpha} H^1\left(G_k, \mathrm{Pic}_{X/k}(k^{\mathrm{sep}})\right) \to H^3(G_k, (k^{\mathrm{sep}})^{\times}).$$

Moreover, if k is a global field, then the term on the right is zero by a theorem of Tate (see, for example, [38, Chap. VIII.3]), thus, α is an isomorphism, and we obtain an interpretation of this cohomology group in terms of Brauer groups, see [47, Sect. 3.4].

Lemma 4.5 *If P is a Brauer–Severi variety over k, then*

$$H^1\left(H, \mathrm{Pic}_{P/k}(k^{\mathrm{sep}})\right) = 0$$

for all closed subgroups $H \subseteq G_k$.

Proof Since $\mathrm{Pic}_{P/k}(k^{\mathrm{sep}}) \cong \mathbb{Z} \cdot \mathcal{O}_P(1)$ and since G_k acts trivially on the class $\mathcal{O}_P(1)$, the desired H^1 is isomorphic to $\mathrm{Hom}(H, \mathbb{Z})$, see [4, Chap. III.1, Exercise 2], for example. This is zero since H is a profinite group and the homomorphisms to $\mathbb{Z}$ are required to be continuous. □

In Proposition 2.10, we established birational invariance of $\mathrm{Am}(X)$. The following result of Manin [35, Sect. 1 of the Appendix] shows that also the above group cohomology groups are a birational invariants.

Theorem 4.6 (Manin) *For every closed subgroup $H \subseteq G_k$, the group*

$$H^1\left(H, \mathrm{Pic}_{X/k}(k^{\mathrm{sep}})\right)$$

is a birational invariant of smooth and projective varieties over k. □

Remark 4.7 Every birational map between smooth and projective surfaces can be factored into a sequence of blow-ups in closed points, see [35, Chap. III]. Using this, one can give very explicit proofs of Proposition 2.10 and Theorem 4.6 in dimension 2. (For such a proof of Theorem 4.6 in dimension 2, see the proof of [35, Theorem 29.1].)

4.3 Hasse Principle and Weak Approximation

For a global field K, i.e., a finite extension of $\mathbb{Q}$ or of $\mathbb{F}_p(t)$, we denote by $\mathbf{\Omega}_K$ the set of its places, including the infinite ones if K is of characteristic zero. A class $\mathcal{C}$ of varieties over K satisfies

(1) the *Hasse principle*, if for every $X \in \mathcal{C}$ we have $X(K) \neq \emptyset$ if and only if $X(K_\nu) \neq \emptyset$ for all $\nu \in \mathbf{\Omega}_K$. Moreover, $\mathcal{C}$ satisfies

(2) *weak approximation*, if the diagonal embedding

$$X(K) \to \prod_{\nu\in\Omega_K} X(K_\nu)$$

is dense for the product of the ν-adic topologies.

If $\mathcal{C}$ satisfies weak approximation, then it obviously also satisfies the Hasse principle, but the converse need not hold. For example, Brauer–Severi varieties over K satisfy the Hasse principle by a theorem of Châtelet [5], as well as weak approximation. However, both properties may fail for del Pezzo surfaces over K, and we refer to [47] for an introduction to this topic. We end this section by noting that the obstruction to a class $\mathrm{Pic}_{(X/K)(\mathrm{fppf})}(K)$ coming from $\mathrm{Pic}_X(K)$ satisfies the Hasse principle.

Lemma 4.8 *Let X a proper variety over a global field K and let $\mathcal{L} \in \mathrm{Pic}_{(X/K)(\mathrm{fppf})}(K)$. Then, the following are equivalent*

(1) $0 = \delta(\mathcal{L}) \in \mathrm{Br}(K)$, *and*
(2) $0 = \delta(\mathcal{L} \otimes_K K_\nu) \in \mathrm{Br}(K_\nu)$ *for all* $\nu \in \Omega_K$.

PROOF A class in $\mathrm{Br}(K)$ is zero if and only if its image in $\mathrm{Br}(K_\nu)$ is zero for all $\nu \in \Omega_K$ by the Hasse principle for the Brauer group. From this, and functoriality of the exact sequence from Proposition 2.5, the assertion follows. □

For example, if $X(K_\nu) \neq \emptyset$ for all $\nu \in \Omega_X$, then δ is the zero map by Proposition 2.5 and this lemma. In this case, every class in $\mathrm{Pic}_{(X/K)(\mathrm{fppf})}(K)$ comes from an invertible sheaf on X.

5 Del Pezzo Surfaces of Product Type

In this section, we classify degree 8 del Pezzo surfaces of product type over k, i.e., surfaces X over k with $X_{\overline{k}} \cong \mathbb{P}^1_{\overline{k}} \times \mathbb{P}^1_{\overline{k}}$, in terms of Brauer–Severi varieties.

First, for $\mathbb{P}^1_k \times \mathbb{P}^1_k$, the anti-canonical embedding can be written as composition of Veronese- and Segre-maps as follows

$$|-K_{\mathbb{P}^1_k\times\mathbb{P}^1_k}| : \mathbb{P}^1_k \times \mathbb{P}^1_k \xrightarrow{\nu_2\times\nu_2} \mathbb{P}^2_k \times \mathbb{P}^2_k \xrightarrow{\sigma} \mathbb{P}^8_k .$$

Next, the invertible sheaf $\omega^{-1}_{\mathbb{P}^1_k\times\mathbb{P}^1_k}$ is uniquely 2-divisible in the Picard group, and we obtain an embedding as a smooth quadric

$$|-\tfrac{1}{2}K_{\mathbb{P}^1_k\times\mathbb{P}^1_k}| : \mathbb{P}^1_k \times \mathbb{P}^1_k \xrightarrow{\sigma} \mathbb{P}^3_k .$$

Now, let X be a degree 8 del Pezzo surface of product type over k. Then, the anti-canonical linear system yields an embedding of X as a surface of degree 8 into $\mathbb{P}^8_k$.

However, the "half-anti-canonical linear system" exists in general only as a morphism to a Brauer–Severi threefold as the following result shows.

Theorem 5.1 *Let X be a degree 8 del Pezzo surface of product type over a field k. Then, there exist a unique class $\mathcal{L} \in \mathrm{Pic}_{(X/k)(\mathrm{fppf})}(k)$ with $\mathcal{L}^{\otimes 2} \cong \omega_X^{-1}$ and an embedding*

$$|\mathcal{L}| : X \hookrightarrow P$$

into a Brauer–Severi threefold P over k with Brauer class

$$\delta(\mathcal{L}) = [P] \in \mathrm{Br}(k),$$

and such that $X_{\overline{k}}$ is a smooth quadric in $P_{\overline{k}} \cong \mathbb{P}^3_{\overline{k}}$. Moreover, X is rational if and only if X has a k-rational point. In this case, we have $P \cong \mathbb{P}^3_k$.

PROOF To simplify notation, set $L := k^{\mathrm{sep}}$. We have $X(L) \neq \emptyset$, for example, by [18, Proposition A.1.1], as well as $\mathrm{Pic}(X_L) \cong \mathrm{Pic}(X_{\overline{k}}) \cong \mathbb{Z}^2$ by Theorem 4.2. The classes $(1, 0)$ and $(0, 1)$ of $\mathrm{Pic}(X_L)$ give rise to two morphisms $X_L \to \mathbb{P}^1_L$, and we obtain an isomorphism $X_L \cong \mathbb{P}^1_L \times \mathbb{P}^1_L$. By abuse of notation, we re-define $\overline{X}$ to be X_L. Next, the absolute Galois group G_k acts trivially on the canonical class $(-2, -2)$, and thus, the G_k-action on $\mathbb{Z}(1, 1) \subset \mathbb{Z}^2$ is trivial. By Proposition 2.7, we have $\mathrm{Pic}_{X/k}(K)^{G_k} \cong \mathrm{Pic}_{(X/k)(\mathrm{fppf})}(k)$, and, since $(1, 1) \in \mathbb{Z}^2$ is G_k-invariant, the unique invertible sheaf $\mathcal{L}$ on $\overline{X}$ with $\mathcal{L}^{\otimes 2} \cong \omega_{\overline{X}}^{-1}$ descends to a class in $\mathrm{Pic}_{(X/k)(\mathrm{fppf})}(k)$. Over L, the class $\mathcal{L}$ is very ample and defines an embedding of $\overline{X}$ as smooth quadric surface into $\mathbb{P}^3_L$. Thus, by Theorem 3.4, we obtain an embedding $|\mathcal{L}| : X \hookrightarrow P$, where P is a Brauer–Severi threefold over k with $\delta(\mathcal{L}) = [P] \in \mathrm{Br}(k)$.

Finally, if X is rational, then it has a k-rational point, and then, also P has a k-rational point, i.e., $P \cong \mathbb{P}^3_k$. Conversely, if there exists a k-rational point $x \in X$, then X is a quadric in $\mathbb{P}^3_k$, and projection away from x induces a birational map $X \dashrightarrow \mathbb{P}^2_k$. ☐

Next, we establish an explicit classification of degree 8 del Pezzo surfaces of product type in terms of the Néron–Severi rank ρ and Brauer–Severi curves. To simplify notation in the sequel, let us recall the definition of contracted products. If a finite group G acts on a scheme X from the right and it acts on a scheme Y from the left and all schemes and actions are over $\mathrm{Spec}\, k$ for some field k, then we denote the quotient of $X \times_{\mathrm{Spec}\, k} Y$ by the diagonal G-action defined by $(x, y) \mapsto (xg, g^{-1}y)$ for all $g \in G$ by

$$X \wedge^G Y := (X \times_{\mathrm{Spec}\, k} Y)/G.$$

We refer to [19, Chap. III.1.3] for details and applications.

Proposition 5.2 *Let X and $X \subset P$ be as in Theorem 5.1.*

(1) if $\rho(X) = 2$, then

$$X \cong P' \times P'',$$

where P' and P'' are Brauer–Severi curves over k, whose Brauer classes satisfy $[P] = [P'] + [P''] \in \mathrm{Br}(k)$. In particular, $P \cong \mathbb{P}^3_k$ if and only if $P' \cong P''$.

(2) If $\rho(X) = 1$, then there exist a Brauer–Severi curve P' over k and a finite Galois extension K/k with Galois group $H := \mathbb{Z}/2\mathbb{Z}$, such that X arises as twisted self-product

$$X \cong (P' \times P')_K/H = \operatorname{Spec} K \wedge^H (P' \times P'),$$

where the H-action permutes the factors of $P'_K \times P'_K$. Moreover, $P \cong \mathbb{P}^3_k$ and P' is a hyperplane section of $X \subset \mathbb{P}^3_k$.

Proof We keep the notations and assumptions from the proof of Theorem 5.1. The G_k-action fixes the class $(1, 1)$. Since the G_k-action preserves the intersection pairing on $\mathrm{Pic}_{X/k}(k^{\mathrm{sep}})$, it follows that G_k acts on $\mathbb{Z}(1, -1)$ either trivially, or by sign changes. We have $\rho(X) = 2$ in the first case, and $\rho(X) = 1$ in the latter.

First, assume that $\rho(X) = 2$. By Theorem 3.4, the classes $(1, 0)$ and $(0, 1)$ give rise to morphisms to Brauer–Severi curves $X \to P'$ and $X \to P''$ of class $[P'] = \delta((1, 0))$ and $[P''] = \delta((0, 1))$ in $\mathrm{Br}(k)$, respectively. Thus, we obtain a morphism $X \to P' \times P''$, which is an isomorphism because it is an isomorphism over k^{sep}. Since δ is a homomorphism, we find $[P] = \delta(\mathcal{L}) = \delta((1, 1)) = \delta((1, 0)) + \delta((0, 1)) = [P'] + [P'']$. Using that P' and P'' are of period 2, we find that $P \cong \mathbb{P}^3_k$ if and only if $[P] = 0$, i.e., if and only if $[P'] = [P'']$. By Corollary 2.16, the latter is equivalent to $P' \cong P''$.

Second, assume that $\rho(X) = 1$. Then, the G_k-action permutes $(0, 1)$ and $(1, 0)$, i.e., it permutes the factors of $\mathbb{P}^1_{k^{\mathrm{sep}}} \times \mathbb{P}^1_{k^{\mathrm{sep}}}$. Thus, there exists a unique quadratic Galois extension K/k, such that $\mathrm{Gal}(k^{\mathrm{sep}}/K)$ acts trivially on $\mathrm{Pic}_{X/k}(k^{\mathrm{sep}})$ and by the previous analysis we have $X_K := Q'' \times Q'''$ for two Brauer–Severi curves Q'', Q''' over K. Using these and the $H := \mathrm{Gal}(K/k)$-action, we obtain a H-stable diagonal embedding $Q' \subset X_K$ of a Brauer–Severi curve over K, and then, the two projections induce isomorphisms $Q' \cong Q''$ and $Q' \cong Q'''$ over K. Taking the quotient by H, we obtain a Brauer–Severi curve $P' := Q'/H \subset X$ over k. Clearly, $P'_K \cong Q'$ and we obtain the description of X as twisted self-product. On X, the curve P' is a section of the class $(1, 1)$, which implies that this class comes from an invertible sheaf, and thus, $0 = \delta((1, 1)) \in \mathrm{Br}(k)$ by Proposition 2.5. Since $\delta((1, 1)) = [P]$, we conclude $P \cong \mathbb{P}^3_k$. □

Remark 5.3 In the case of quadrics in $\mathbb{P}^3$, similar results were already established in [9]. A related, but somewhat different view on degree 8 del Pezzo surfaces of product type was taken in (the proof of) [10, Proposition 5.2]: If X is such a surface, then there exists a quadratic Galois extension K/k and a Brauer–Severi curve C over K, such that $X \cong \mathrm{Res}_{K/k} C$, where $\mathrm{Res}_{K/k}$ denotes Weil restriction, see also [41].

Corollary 5.4 *Let X be as in Theorem 5.1. Then,*

$$H^1\left(H, \mathrm{Pic}_{X/k}(k^{\mathrm{sep}})\right) = 0$$

for all closed subgroups $H \subseteq G_k$, and

$$\mathrm{Am}(X) \cong \begin{cases} 0 & \textit{if } \rho = 1 \textit{ or if } X \cong \mathbb{P}^1_k \times \mathbb{P}^1_k, \\ (\mathbb{Z}/2\mathbb{Z})^2 & \textit{if } \rho = 2 \textit{ and } \mathbb{P}^1_k \not\cong P' \not\cong P'' \not\cong \mathbb{P}^1_k, \\ (\mathbb{Z}/2\mathbb{Z}) & \textit{in the remaining } \rho = 2\textit{-cases.} \end{cases}$$

PROOF Set $H^1(H) := H^1(H, \mathrm{Pic}_{X/k}(k^{\mathrm{sep}}))$. If $\rho = 2$, then the G_k-action on $\mathrm{Pic}_{X/k}(k^{\mathrm{sep}})$ is trivial, and we find $H^1(H) = 0$ as in the proof of Lemma 4.5. Moreover, $\mathrm{Am}(X)$ is generated by $\delta((0,1)$ and $\delta((1,0))$, i.e., by $[P']$ and $[P'']$ in $\mathrm{Br}(k)$. From this, the assertions on $\mathrm{Am}(X)$ follow in case $\rho = 2$.

If $\rho = 1$, then there exists an isomorphism $\mathrm{Pic}_{X/k}(k^{\mathrm{sep}}) \cong \mathbb{Z}^2$, such that the G_k-action factors through a surjective homomorphism $G_k \to \mathbb{Z}/2\mathbb{Z}$ and acts on $\mathbb{Z}^2$ via $(a,b) \mapsto (b,a)$. In particular, we find $H^1(\mathbb{Z}/2\mathbb{Z}, \mathbb{Z}^2) = 0$ with respect to this action, see, for example, [4, Chap. III.1, Example 2]. From this, we deduce $H^1(H) = 0$ using inflation maps. Moreover, $\mathrm{Am}(X)$ is generated by $\delta((1,1))$, which is zero, since $(1,1)$ is the class of an invertible sheaf. □

Corollary 5.5 *If X is as in Theorem 5.1, then the following are equivalent*

(1) X is birationally equivalent to a Brauer–Severi surface,
(2) X is rational,
(3) X has a k-rational point, and
(4) X is isomorphic to

$$X \cong \mathbb{P}^1_k \times \mathbb{P}^1_k \quad \textit{or to} \quad X \cong \mathrm{Spec}\, K \wedge (\mathbb{P}^1_k \times \mathbb{P}^1_k).$$

PROOF The implications (2) ⇒ (1) and (2) ⇒ (3) are trivial, and we established (3) ⇒ (2) in Theorem 5.1. Moreover, if X is birationally equivalent to a Brauer–Severi surface P, then $\mathrm{Am}(P) = \mathrm{Am}(X)$ is cyclic of order 1 or 3 by Lemma 4.5 and Theorem 4.6. Together with Corollary 5.4, we conclude $\mathrm{Am}(P) = \mathrm{Am}(X) = 0$, i.e., $P \cong \mathbb{P}^2_k$, which establishes (1) ⇒ (2).

Since (4) ⇒ (3) is trivial, it remains to establish (3) ⇒ (4). Thus, we assume $X(k) \neq \emptyset$. If $\rho = 2$, then $X \cong P' \times P''$ and both Brauer–Severi curves P' and P'' have k-rational points, i.e., $X \cong \mathbb{P}^1_k \times \mathbb{P}^1_k$. If $\rho = 1$, we have an embedding $X \subset \mathbb{P}^3_k$ and $X \cong \mathrm{Spec}\, K \wedge (P' \times P')$. Since $X(k) \neq \emptyset$, we have $X(K) \neq \emptyset$, which yields $P'(K) \neq \emptyset$, and thus $P'_K \cong \mathbb{P}^1_K$. A k-rational point on X gives rise to a K-rational and $\mathrm{Gal}(K/k)$-stable point on $X_K \cong \mathbb{P}^1_K \times \mathbb{P}^1_K$. In particular, this point lies on some diagonal $\mathbb{P}^1_K \subset X_K$, and thus, lies on some diagonal $P'' \subseteq X$ with $X \cong \mathrm{Spec}\, K \wedge (P'' \times P'')$. Since $P''(k) \neq \emptyset$, we find $P'' \cong \mathbb{P}^1_k$. □

We refer to Sect. 6.1 for more applications of these results to the arithmetic and geometry of these surfaces.

6 Del Pezzo Surfaces of Large Degree

Let X be a del Pezzo surface of degree d over a field k that is not of product type. Then, there exists a birational morphism

$$\overline{f} : \overline{X} \to \mathbb{P}^2_{\overline{k}}$$

that is a blow-up in $(9-d)$ closed points $P_1, \ldots, P_{9-d}$ in general position. We set $H := \overline{f}^{\,*}\mathcal{O}_{\mathbb{P}^2_{\overline{k}}}(1)$ and let $E_i := \overline{f}^{\,-1}(P_i)$ be the exceptional divisors of $\overline{f}$. Then, there exists an isomorphism of abelian groups

$$\mathrm{Pic}(\overline{X}) \cong \mathbb{Z}H \,\oplus\, \bigoplus_{i=1}^{9-d} \mathbb{Z}E_i .$$

The (-1)-curves of $\overline{X}$ consist of the E_i, of preimages under $\overline{f}$ of lines through two distinct points P_i, of preimages under $\overline{f}$ of quadrics through five distinct points P_i, etc., and we refer to [35, Theorem 26.2] for details. Let $K_{\overline{X}}$ be the canonical divisor class of $\overline{X}$, and let $\widetilde{E}$ be the sum of all (-1)-curves on $\overline{X}$. We leave it to the reader to verify the following table.

d	class of $\widetilde{E}$ in $\mathrm{Pic}(\overline{X})$	relations
9	0	$3H = -K_{\overline{X}}$
8	E_1	$3H = -K_{\overline{X}} + \widetilde{E}$
7	H	$H = \widetilde{E}$
6	$3H - \sum_{i=1}^{3} E_i$	$0 = -K_{\overline{X}} - \widetilde{E}$
5	$6H - 2\sum_{i=1}^{4} E_i$	$0 = -2K_{\overline{X}} - \widetilde{E}$
4	$12H - 4\sum_{i=1}^{5} E_i$	$0 = -4K_{\overline{X}} - \widetilde{E}$
3	$27H - 9\sum_{i=1}^{6} E_i$	$0 = -9K_{\overline{X}} - \widetilde{E}$
2	$84H - 28\sum_{i=1}^{7} E_i$	$0 = -28K_{\overline{X}} - \widetilde{E}$
1	$720H - 240\sum_{i=1}^{8} E_i$	$0 = -240K_{\overline{X}} - \widetilde{E}$

Together with Theorem 3.4, we obtain the following result.

Theorem 6.1 *Let X be a del Pezzo surface of degree $d \geq 7$ over a field k that is not of product type. Then, $\overline{f}$ descends to a birational morphism*

$$f : X \to P$$

to a Brauer–Severi surface P over k, where

$$\delta(H) = [P] \in \mathrm{Br}(k) \quad \textit{and} \quad \mathrm{Am}(X) \cong \mathbb{Z}/\mathrm{per}(P)\mathbb{Z}.$$

Moreover, X is rational if and only if $P \cong \mathbb{P}^2_k$. This is equivalent to X having a k-rational point.

Proof By Theorem 4.2, the invertible sheaf H on $X_{\overline{k}}$ defining $\overline{f}$ already lies in $\mathrm{Pic}_X(k^{\mathrm{sep}})$, i.e., $\overline{f}$ descends to k^{sep}, and by abuse of notation, we re-define $\overline{X}$ to be $X_{k^{\mathrm{sep}}}$. Clearly, the canonical divisor class $K_{\overline{X}}$ is G_k-invariant, and since G_k permutes the (-1)-curves of $\overline{X}$, also the class of $\widetilde{E}$ is G_k-invariant. In particular, $K_{\overline{X}}$ and $\widetilde{E}$ define classes in $\mathrm{Pic}_{X/k}(k^{\mathrm{sep}})^{G_k} \cong \mathrm{Pic}_{(X/k)(\mathrm{fppf})}(k)$. If $d \geq 7$, then the above table shows that there exist positive multiples of H that are integral linear combinations of $K_{\overline{X}}$ and $\widetilde{E}$. Thus, $H \in \mathrm{Pic}_{X/k}(k^{\mathrm{sep}})$ descends to a class in $\mathrm{Pic}_{(X/k)(\mathrm{fppf})}(k)$. By Theorem 3.4, $\overline{f}$ descends to a birational morphism $f : X \to P$, where P is a Brauer–Severi surface of class $\delta(H) \in \mathrm{Br}(k)$. The assertion on $\mathrm{Am}(X)$ follows from Proposition 2.10 and Theorem 2.19.

If X has a k-rational point, then so has P, and then $P \cong \mathbb{P}^2_k$. Since f is a birational morphism, $P \cong \mathbb{P}^2_k$ implies that X is rational. And if X is rational, then it has a k-rational point by Lemma 4.3. □

As an immediate consequence, we obtain rationality and the existence of k-rational points in some cases.

Corollary 6.2 *Let X be as in Theorem 6.1. If $d \in \{7, 8\}$, then X has a k-rational point and $\overline{f}$ descends to a birational morphism $f : X \to \mathbb{P}^2_k$.*

Proof By Theorem 6.1, there exists a birational morphism $X \to P$ that is a blow-up in a closed subscheme $Z \subset P$ of length $(9 - d)$. By Corollary 2.21, we have $P \cong \mathbb{P}^2_k$ if 3 and $(9 - d)$ are coprime. In particular, we have $X(k) \neq \emptyset$ in these cases by Theorem 6.1 and Lemma 4.3. □

Since a del Pezzo surface of degree 9 is a Brauer–Severi surface, it has rational points if and only if it is trivial. In particular, Corollary 6.2 does not hold for $d = 9$.

6.1 Applications to Arithmetic Geometry

We now give a couple of applications of the just established results. Again, we stress that most if not all of these applications are well-known, and merely illustrate the usefulness of studying varieties via Brauer–Severi varieties.

Corollary 6.3 *If X is a del Pezzo surface of degree ≥ 7 over k, then*

$$H^1\left(H,\ \mathrm{Pic}_{X/k}(k^{\mathrm{sep}})\right) = 0.$$

for all closed subgroups $H \subseteq G_k$

PROOF If X is not of product type, then it is birationally equivalent to a Brauer–Severi surface P by Theorem 6.1, and then the statement follows from Theorem 4.6 and Lemma 4.5. If X is of product type, then this is Corollary 5.4. □

For the next application, let us recall that a surface is called *rational* if it is birationally equivalent to $\mathbb{P}^2_k$, and that it is called *unirational* if there exists a dominant and rational map from $\mathbb{P}^2_k$ onto it. The following result is a special case of [35, Theorem 29.4].

Corollary 6.4 *Let X be a del Pezzo surface of degree ≥ 7 over a field k. Then, the following are equivalent:*

(1) X is rational,
(2) X is unirational, and
(3) X has a k-rational point.

PROOF Clearly, we have (1) $\Rightarrow$ (2) $\Rightarrow$ (3), whereas (3) $\Rightarrow$ (1) follows from Corollary 5.5 and Theorem 6.1. □

This leads us to the question whether a del Pezzo surface necessarily has a k-rational point. Over finite fields, this is true and follows from the Weil conjectures, which we will recall in Theorem 8.1 below. By a theorem of Wedderburn, finite fields have trivial Brauer groups, and thus, the following corollary gives existence of k-rational points for more general fields.

Corollary 6.5 *Let X be a del Pezzo surface of degree ≥ 7 over a field k with* $\mathrm{Br}(k) = 0$. *Then, X has a k-rational point, and thus, is rational.*

Proof If X is not of product type, then there exists a birational morphism $f : X \to P$ to a Brauer–Severi surface by Theorem 6.1. Since $\mathrm{Br}(k) = 0$, we have $P \cong \mathbb{P}^2_k$, and Theorem 6.1 gives $X(k) \neq \emptyset$.

Thus, let X be of product type. By Proposition 5.2, X is a product of Brauer–Severi curves ($\rho = 2$), or contains at least a Brauer–Severi curve ($\rho = 1$). Since $\mathrm{Br}(k) = 0$, all Brauer–Severi curves are isomorphic to $\mathbb{P}^1_k$, and thus, contain k-rational points. In particular, we find $X(k) \neq \emptyset$. □

In Sect. 4.3, we discussed the Hasse principle and weak approximation for varieties over global fields. Here, we establish the following.

Corollary 6.6 *Del Pezzo surfaces of degree ≥ 7 over global fields satisfy weak approximation and the Hasse principle.*

Proof If X is not of product type, then it is birationally equivalent to a Brauer–Severi surface by Theorem 6.1, and since the two claimed properties are preserved under birational maps and hold for Brauer–Severi varieties, the assertion follows in this case.

If X is of product type, then there are two cases by Proposition 5.2. If $\rho = 2$, then X is a product of two Brauer–Severi curves, and we conclude as before.

Thus, we may assume $\rho = 1$. Let us first establish the Hasse principle: there exists a quadratic Galois extension L/K, such that $\rho(X_L) = 2$. From $X(K_\nu) \neq \emptyset$ for all $\nu \in \Omega_K$, we find $X_{L_\mu} \cong \mathbb{P}^1_{L_\mu} \times \mathbb{P}^1_{L_\mu}$ for all $\mu \in \Omega_L$, and thus, $X_L \cong \mathbb{P}^1_L \times \mathbb{P}^1_L$ by the Hasse principle for Brauer–Severi curves. As in the proof of Corollary 5.5, we exhibit X as twisted self-product of $\mathbb{P}^1_k$, which has a k-rational point and establishes the Hasse principle. Thus, to establish weak approximation, we may assume that X has a k-rational point. But then, X is rational by Corollary 5.5, and since weak approximation is a birational invariant, the assertion follows. □

7 Del Pezzo Surfaces of Degree 6

In the previous sections, we have seen a close connection between Brauer–Severi varieties and del Pezzo surfaces of degree ≥ 7. In this section, we discuss del Pezzo surfaces of degree 6, which are not so directly linked to Brauer–Severi varieties.

For the geometry and the arithmetic of these surfaces, we refer the interested reader to [6, 35], and the survey [47, Sect. 2.4]. We keep the notation introduced in Sect. 6: If X is a degree 6 del Pezzo surface over a field k, then there exists a blow-up $f_{\overline{k}} : \overline{X} \to \mathbb{P}^2_{\overline{k}}$ in three points in general position with exceptional (-1)-curves E_1, E_2, and E_3. Then, there are six (-1)-curves on X, namely the three exceptional curves E_i, $i = 1, 2, 3$ of $\overline{f}$, as well as the three curves $E_i' := H - E_j - E_k$, $i = 1, 2, 3$ where $\{j, k\} = \{1, 2, 3\} \backslash \{i\}$ and where $H = \overline{f}^* \mathcal{O}_{\mathbb{P}^2}(1)$ as in Sect. 6. These curves intersect in a hexagon as follows.

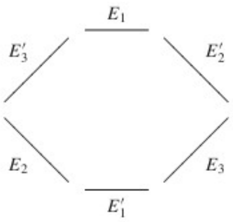

The absolute Galois group G_k acts on these six (-1)-curves on $X_{k^{\mathrm{sep}}}$, and associated to this action, we have following field extensions of k.

(1) Since G_k acts on the two sets $\{E_1, E_2, E_3\}$ and $\{E_1', E_2', E_3'\}$, there is a group homomorphism

$$\varphi_1 \,:\, G_k \,\to\, S_2 \,\cong\, \mathbb{Z}/2\mathbb{Z}.$$

The fixed field of either of the two sets is a finite separable extension $k \subseteq K$ with $[K : k]|2$, and $k \neq K$ if and only if φ_1 is surjective.

(2) Since G_k acts on the three sets $\{E_i, E_i'\}$, $i = 1, 2, 3$, there is a group homomorphism

$$\varphi_2 \,:\, G_k \,\to\, S_3.$$

There exists a finite separable extension $k \subseteq L$ with $[L : k]|3$, unique up to conjugation in k^{sep}, over which at least one of these three sets is defined. We have $k \neq L$ if and only if 3 divides the order of $\varphi_2(G_k)$. Next, there exists a finite and separable extension $L \subseteq M$ with $[M : L]|2$, over which all three sets are defined.

Combining φ_1 and φ_2, we obtain a group homomorphism

$$G_k \xrightarrow{\varphi_1 \times \varphi_2} \mathbb{Z}/2\mathbb{Z} \times S_3 \,\cong\, D_{2\cdot 6},$$

where $D_{2\cdot 6}$ denotes the dihedral group of order 12, i.e., the automorphism group of the hexagon. Using these field extensions, we obtain the following classification, which uses and slightly extends a classical result of Manin from [35] in case (3).

Theorem 7.1 *Let X be a del Pezzo surface of degree* 6 *over a field k.*

(1) The morphism $\overline{f}$ descends to a birational morphism

$$f \,:\, X \,\to\, P$$

to a Brauer–Severi surface P if and only if $k = K$. In this case, $\rho(X) \geq 2$ and $\mathrm{Am}(X) = \mathrm{Am}(P)$.

(2) There exists a birational morphism $X \to Y$ onto a degree 8 *del Pezzo surface Y of product type if and only if $k = L$. In this case,*

	$\rho(X)$	Y
$k \neq M$	3	$\mathrm{Spec}\, M \wedge (\mathbb{P}^1_k \times \mathbb{P}^1_k)$
$k = M$	4	$\mathbb{P}^1_k \times \mathbb{P}^1_k$

X has a k-rational point, and $\mathrm{Am}(X) = 0$.

(3) If $k \neq K$ and $k \neq L$, then $\rho(X) = 1$, $\mathrm{Am}(X) = 0$, *and the following are equivalent.*

(a) X is birationally equivalent to a Brauer–Severi surface,
(b) X is birationally equivalent to a product of two Brauer–Severi curves,
(c) X is rational, and
(d) X has a k-rational point.

Proof Let us first show (1). If $k = K$, then $F := E_1 + E_2 + E_3$ descends to a class in $\mathrm{Pic}(X_{k^{\mathrm{sep}}})^{G_k} = \mathrm{Pic}_{(X/k)(\mathrm{fppf})}(k)$ and we find $\rho(X) \geq 2$. Thus, also $H = \frac{1}{3}(-K_X + F)$ descends to a class in $\mathrm{Pic}_{(X/k)(\mathrm{fppf})}(k)$, and by Theorem 3.4, we obtain a birational morphism $|H| : X \to P$ to a Brauer–Severi surface, which coincides with $\overline{f}$ over $\overline{k}$. Conversely, if $\overline{f}$ descends to a birational morphism $f : X \to P$, then the exceptional divisor of f is of class F or $E'_1 + E'_2 + E'_3$, and we find $k = K$. Moreover, we have $\mathrm{Am}(X) = \mathrm{Am}(P)$ by Theorem 4.6.

If $k = L$, then, say $E_1 + E'_1$, descends to a class in $\mathrm{Pic}(X_{k^{\mathrm{sep}}})^{G_k}$. Moreover, we find that the classes $\frac{1}{2}(-K_X + E_1 + E'_1) = 2H - E_2 - E_3$ as well as $\frac{1}{2}(-K_X - E_1 - E'_1) = H - E_1$, and thus, the classes H, E_1, and $E'_1 = H - E_2 - E_3$ lie in $\mathrm{Pic}(X_{k^{\mathrm{sep}}})^{G_k}$. The G_k-action is trivial on H and E_1, whereas it is either trivial on the set $\{E_2, E_3\}$ (if $k = M$) or permutes the two (if $k \neq M$). Since the class of E_1 is G_k-invariant and there is a unique effective divisor in this linear system, we find that $\mathbb{P}^1_k \cong E_1 \subset X$. In particular, X has a k-rational point and $\mathrm{Am}(X) = 0$. Using Theorem 3.4 and the fact that X has a k-rational point, we obtain a birational morphism

$$|\frac{1}{2}(-K_X + E_1 + E'_1)| \,:\, X \to Y \subset \mathbb{P}^3_k$$

onto a smooth quadric Y with a k-rational point. In particular, Y is a degree 8 del Pezzo surface of product type. Over k^{sep}, this morphism contracts E_1 and E_1' and thus, we find

$$\mathrm{Pic}(Y_{k^{\mathrm{sep}}}) \cong \left(\mathbb{Z}H \oplus \bigoplus_{i=1}^{3} \mathbb{Z}E_i \right) / \langle E_1, E_1' \rangle \cong \mathbb{Z}\overline{E}_2 \oplus \mathbb{Z}\overline{E}_3.$$

The G_k-action on it is either trivial ($k = M$) or permutes the two summands ($k \neq M$). Using $Y(k) \neq \emptyset$ and Corollary 5.5, we find $\rho(X) = 4$ and $Y \cong \mathbb{P}^1_k \times \mathbb{P}^1_k$ in the first case, and $\rho(X) = 3$ and $Y \cong \mathrm{Spec}\, M \wedge (\mathbb{P}^1_k \times \mathbb{P}^1_k)$ in the latter. Conversely, if there exists a birational morphism $X \to Y$ onto a degree 8 del Pezzo surface of product type, then the exceptional divisor is of class $E_i + E_i'$ for some i, and thus, $k = L$. This establishes (2).

Finally, assume that $k \neq K$ and $k \neq L$. Then, φ_1 is surjective, and $\varphi_2(G_k)$ contains all 3-cycles of S_3. From this, it is not difficult to see that $\mathrm{Pic}(X_{\overline{k}})^{G_k}$ is of rank 1 and generated by the class of K_X. Since this latter class is an invertible sheaf, we find $\mathrm{Am}(X) = 0$. Thus, if X is birationally equivalent to a Brauer–Severi surface P, then $\mathrm{Am}(X) = 0$ together with Lemma 4.5 and Theorem 4.6 implies that $P \cong \mathbb{P}^2_k$. Similarly, if X is birationally equivalent to the product $P' \times P''$ of two Brauer–Severi curves, then $P' \cong P'' \cong \mathbb{P}^1_k$. From this, we obtain the implications $(a) \Leftrightarrow (b) \Leftrightarrow (c) \Rightarrow (d)$. The implication $(d) \Rightarrow (c)$ is due to Manin [35, Theorem 29.4]. □

Remark 7.2 In case (1) of the above theorem it is important to note that P need not be unique, but that $\mathrm{Am}(P)$ is well-defined. More precisely, if we set $F := E_1 + E_2 + E_3$ and $F' = E_1' + E_2' + E_3'$, then Theorem 3.4 provides us with two morphisms to Brauer–Severi surfaces P_1 and P_2

$$\begin{array}{rcl} |H| & = |\tfrac{1}{3}(-K_X + F)| : X & \to P_1 \\ |H'| & := |\tfrac{1}{3}(-K_X + F')| : X & \to P_2 \end{array}$$

Since $H + H' = -K_X$ and $\delta(K_X) = 0$, we find

$$[P_1] = \delta(H) = \delta(-K_X - H') = -\delta(H') = -[P_2] \in \mathrm{Br}(k),$$

and thus, $P_1 \cong P_2$ if and only if both are isomorphic to $\mathbb{P}^2_k$. On the other hand, P_1 and P_2 are birationally equivalent, since we have birational morphisms

$$P_1 \xleftarrow{|H|} X \xrightarrow{|H'|} P_2 \,.$$

Over $\overline{k}$, this becomes the blow-up of three closed points Z followed by the blow-down of the three (-1)-curves that are the strict transforms of lines through any two of the points in Z. This is an example of a *Cremona transformation*.

We remark that a surface of case (3) and without k-rational points is neither birationally equivalent to a Brauer–Severi surface nor to the product of two Brauer–Severi curves. For finer and more detailed classification results for degree 6 del Pezzo surfaces, we refer the interested reader to [2, 10, 13]. Finally, the sum $\widetilde{E}$ of all (-1)-curves on $X_{k^{\mathrm{sep}}}$ is a G_k-invariant divisor, and thus, descends to a curve on X. By [35, Theorem 30.3.1], the complement $X\backslash\widetilde{E}$ is isomorphic to a torsor under a two-dimensional torus over k, which can be used to study the arithmetic and geometry of these surfaces, see also [43].

8 Del Pezzo Surfaces of Small Degree

For the remainder of this article, our results will be less complete and less self-contained. We will circle around questions of birationality of a del Pezzo surface X of degree ≤ 5 to Brauer–Severi surfaces, and about descending the morphism $\overline{f}:\overline{X}\to\mathbb{P}^2_{\overline{k}}$ to k.

8.1 Birationality to Brauer–Severi Surfaces

Let $k=\mathbb{F}_q$ be a finite field of characteristic p, and let X be a smooth and projective surface over k such that $X_{\overline{k}}$ is birationally equivalent to $\mathbb{P}^2$. Then, it follows from the Weil conjectures (in this case already a theorem of Weil himself) that the number of k-rational points is congruent to 1 modulo q, see [35, Chap. IV.27]. In particular, we obtain that

Theorem 8.1 (Weil) *If X is a del Pezzo surface over a finite field $\mathbb{F}_q$, then X has a $\mathbb{F}_q$-rational point.*

Since $\mathrm{Br}(\mathbb{F}_q)=0$ by a theorem of Wedderburn, there are no non-trivial Brauer–Severi varieties over $\mathbb{F}_q$.

Remark 8.2 Let X be a del Pezzo surface of degree ≥ 5 over a field k. Manin [35, Theorem 29.4] showed that X is rational if and only if it contains a k-rational point. Even if X has no k-rational point, Manin [35, Theorem 29.3] showed that

$$H^1\left(H,\ \mathrm{Pic}_{(X/k)(\mathrm{fppf})}(k^{\mathrm{sep}})\right) = 0$$

for all closed subgroups $H\subseteq G_k$. We refer to [8, Théorème 2.B.1] for a general principle explaining this vanishing of cohomology.

In this section, we give a partial generalization to birational maps to Brauer–Severi surfaces.

Lemma 8.3 *Let X be a degree d del Pezzo surface over k. Then,*

(1) There exists an effective zero-cycle Z of degree d on X. If $d \neq 2$ or if $\mathrm{char}(k) \neq 2$, then there exists such a zero-cycle Z, whose closed points have residue fields that are separable over k.
(2) The abelian group $\mathrm{Am}(X)$ is finite and every element has an order dividing d.

PROOF If $d \geq 3$, then ω_X^{-1} is very ample, and $|\omega_X^{-1}|$ embeds X as a surface of degree d into $\mathbb{P}_k^d$. Intersecting X with a linear subspace of codimension 2, we obtain an effective zero-cycle Z of degree d on X. The closed points of Z have automatically separable residue fields if k is finite. Otherwise, k is infinite, and then, the intersection with a generic linear subspace of codimension 2 yields a Z that is smooth over k by [28, Théorème I.6.3]. Thus, in any case, we obtain a Z, whose closed points have residue fields that are separable over k. If $d = 2$, then $|\omega_X^{-1}|$ defines a double cover $X \to \mathbb{P}_k^2$, and the pre-image of a k-rational point yields an effective zero-cycle Z of degree 2 on X. If $\mathrm{char}(k) \neq 2$, then residue fields of closed points of Z are separable over k. If $d = 1$, then $|-K_X|$ has a unique-base point, and in particular, X has a k-rational point. This establishes (1). Since $b_1(X) = 0$, the group $\mathrm{Am}(X)$ is finite by Lemma 2.13. Then, assertion (2) follows from Lemma 2.9. □

Corollary 8.4 *Let X be a del Pezzo surface of degree d over a field k.*

(1) If $d \in \{1, 2, 4, 5, 7, 8\}$ and X is birationally equivalent to a Brauer–Severi surface P, then $P \cong \mathbb{P}_k^2$ and X has a k-rational point.
(2) If $d \in \{1, 3, 5, 7, 9\}$ and X is birationally equivalent to a product $P' \times P''$ of two Brauer–Severi curves, then $P' \cong P'' \cong \mathbb{P}_k^1$ and X has a k-rational point.

PROOF Let X and d be as in (1). Then, every element of $\mathrm{Am}(X)$ is of order dividing d by Lemma 8.3, but also of order dividing 3 by Theorems 2.18 and 4.6. By our assumptions on d, we find $\mathrm{Am}(P) = 0$, and thus, $P \cong \mathbb{P}_k^2$. Since the latter has a k-rational point, so has X by Lemma 4.3. This shows (1). The proof of (2) is similar and we leave it to the reader. □

Combining this with a result of Coray [12], we obtain the following.

Theorem 8.5 *Let X be a del Pezzo surface of degree $d \in \{5, 7, 8\}$ over a perfect field k. Then, the following are equivalent*

(1) There exists a dominant and rational map $P \dashrightarrow X$ from a Brauer–Severi surface P over k,
(2) X is birationally equivalent to a Brauer–Severi surface,
(3) X is rational, and
(4) X has a k-rational point.

PROOF The implications (3) $\Rightarrow$ (2) $\Rightarrow$ (1) are trivial.

Let $\varphi : P \dashrightarrow X$ be as in (1). By Lemma 8.3, there exists a zero-cycle of degree 9 on P, and another one of degree d on X. Using φ, we obtain a zero-cycle of degree dividing 9 on X. By assumption, d is coprime to 9, and thus, there exists a zero-cycle of degree 1 on X. By [12], this implies that X has a k-rational point and establishes (1) $\Rightarrow$ (4).

The implication (4) $\Rightarrow$ (3) is a result of Manin [35, Theorem 29.4]. □

Now, if a del Pezzo surface X over a field k is birationally equivalent to a Brauer–Severi surface, then $H^1(H, \mathrm{Pic}_{X/k}(k^{\mathrm{sep}})) = 0$ for all closed subgroups $H \subseteq G_k$ by Theorem 4.6. Moreover, this vanishing holds for all del Pezzo surfaces of degree ≥ 5, see Remark 8.2. However, for del Pezzo surfaces of degree ≤ 4, these cohomology groups may be non-zero, see [35, Sect. 31], [32, 45, 46]. In particular, del Pezzo surfaces of degree ≤ 4 are in general *not* birationally equivalent to Brauer–Severi surfaces.

For further information concerning geometrically rational surfaces, unirationality, central simple algebras, and connections with cohomological dimension, we refer the interested reader to [10].

8.2 Del Pezzo Surfaces of Degree 5

In order to decide whether a birational map $f_{\overline{k}} : X_{\overline{k}} \to \mathbb{P}^2_{\overline{k}}$ as in Sect. 6 descends to k for a degree 5 del Pezzo surface X over k, we introduce the following notion.

Definition 8.6 Let X be a del Pezzo surface over a field k. A *conic* on X is a geometrically integral curve C on X with $C^2 = 0$ and $-K_X \cdot C = 2$. An element $\mathcal{L} \in \mathrm{Pic}_{(X/k)(\mathrm{fppf})}(k)$ is called a *conic class* if $\mathcal{L} \otimes_k \overline{k} \cong \mathcal{O}_{X_{\overline{k}}}(\overline{C})$ for some conic $\overline{C}$ on $X_{\overline{k}}$.

The following is an analogue of Theorem 6.1 for degree 5 del Pezzo surfaces.

Theorem 8.7 *Let X be a del Pezzo surface of degree 5 over a field k. Then, the following are equivalent:*

(1) There exists a birational morphism $f : X \to P$ to a Brauer–Severi surface, such that $f_{\overline{k}}$ is the blow-up of 4 points in general position.
(2) There exists a birational morphism $f : X \to \mathbb{P}^2_k$, such that $f_{\overline{k}}$ is the blow-up of 4 points in general position.
(3) There exists a class $F \in \mathrm{Pic}_{(X/k)(\mathrm{fppf})}(k)$ such that

$$F_{\overline{k}} \cong \mathcal{O}_{\overline{X}}(E_1 + E_2 + E_3 + E_4),$$

where the E_i are disjoint (-1)-curves on $\overline{X}$.
(4) There exists a conic class in $\mathrm{Pic}_{(X/k)(\mathrm{fppf})}(k)$.

If these equivalent conditions hold, then X has a k-rational point.

PROOF If f is as in (1), then X has a k-rational point by Corollary 8.4. Thus, $P \cong \mathbb{P}^2_k$, and we obtain (1) $\Rightarrow$ (2).

If f is as in (2), then the exceptional divisor of f is a class F as stated in (3), and we obtain (2) $\Rightarrow$ (3).

If f is as in (3), then, using Theorem 3.4, there exists a birational morphism $|\frac{1}{3}(-K_X - F)| : X \to P$ to a Brauer–Severi surface P as in (1), which establishes (3) $\Rightarrow$ (1).

If f is as in (2), let $Z \subset \mathbb{P}^2_k$ be the degree 4 cycle blown up by f. Then $f^*(\mathcal{O}_{\mathbb{P}^2_k}(2)(-Z))$, i.e., the pullback of the pencil of conics through Z, is a conic class on X and establishes (2) $\Rightarrow$ (4).

Finally, if C is a conic class on X, then, using Theorem 3.4, there exists a birational morphism $| - K_X + C| : X \to P$ to a Brauer–Severi surface P as in (1), which establishes (4) $\Rightarrow$ (1). □

Remark 8.8 By theorems of Enriques, Swinnerton-Dyer, Skorobogatov, Shepherd-Barron, Kollár, and Hassett (see [47, Theorem 2.5] for precise references and overview), a degree 5 del Pezzo X over a field k always has a k-rational point. Thus, X is rational by [35, Theorem 29.4], and we have

$$\mathrm{Am}(X) = 0, \quad \text{as well as} \quad H^1(H, \mathrm{Pic}_{X/k}(k^{\mathrm{sep}})) = 0$$

for every closed subgroup $H \subseteq G_k$ by Corollary 2.12, Theorem 4.6, and Lemma 4.5.

8.3 Del Pezzo Surfaces of Degree 4

A classical theorem of Manin [35, Theorem 29.4] states that a del Pezzo surface of degree 4 over a sufficiently large field k is unirational if and only if it contains a k-rational point. Here, we have the following analogue in our setting.

Proposition 8.9 *Let X be a del Pezzo surface of degree 4 over a perfect field k. Then, the following are equivalent*

(1) There exists a dominant rational map $P \dashrightarrow X$ from a Brauer–Severi surface P over k.
(2) X is unirational,
(3) X has a k-rational point,

PROOF The implications (2) $\Rightarrow$ (1) is trivial and (2) $\Rightarrow$ (3) is Lemma 4.3.

The implication (3) $\Rightarrow$ (2) is shown in [35, Theorem 29.4] and [35, Theorem 30.1] if k has at least 23 elements and in [31, Theorem 2.1] and [40, Proposition 5.19] in the remaining cases.

To show (1) $\Rightarrow$ (3), we argue as in the proof of the implication (1) $\Rightarrow$ (4) of Theorem 8.5 by first exhibiting a degree 1 zero-cycle on X, and then, using [12] to deduce the existence of a k-rational point on X. We leave the details to the reader. □

If a field k is finite or perfect of characteristic 2, then a degree 4 del Pezzo surface over k always has a k-rational point, see [35, Theorem 27.1] and [16]. In this case, we also have $\mathrm{Am}(X) = 0$. From Lemma 8.3, we infer that $\mathrm{Am}(X)$ is at most 4-torsion for degree 4 del Pezzo surfaces. For the possibilities of $H^1(G_k, \mathrm{Pic}_{X/k}(k^{\mathrm{sep}}))$, see [45].

The following is an analog of Theorem 6.1 for degree 4 del Pezzo surfaces.

Theorem 8.10 *Let X be a del Pezzo surface of degree 4 over a field k. Then, the following are equivalent:*

(1) There exists a birational morphism $f : X \to P$ to a Brauer–Severi surface, such that $f_{\overline{k}}$ is the blow-up of 5 points in general position.
(2) There exists a birational morphism $f : X \to \mathbb{P}^2_k$, such that $f_{\overline{k}}$ is the blow-up of 5 points in general position.
(3) There exists a curve $\mathbb{P}^1_k \cong E \subset X$ with $E^2 = -1$.
(4) There exists a class $E \in \mathrm{Pic}_{(X/k)(\mathrm{fppf})}(k)$ with $E^2 = K_X \cdot E = -1$.

If these equivalent conditions hold, then X has a k-rational point.

PROOF The implication (2) ⇒ (1) is trivial. If f is as in (1), then X has a k-rational point by Corollary 8.4. Thus, $P \cong \mathbb{P}^2_k$, and we obtain (1) ⇒ (2).

If f is as in (2), let $Z \subset \mathbb{P}^2_k$ be the degree 5 cycle blown up by f. Then $f^*(\mathcal{O}_{\mathbb{P}^2_k}(2)(-Z))$, i.e., the pullback of the class of the unique conic through Z, is a class E as stated in (4) on X and establishes (2) ⇒ (4).

If E is a class as in (4), then, using Theorem 3.4, there exists a birational morphism $|-K_X - E| : X \to P$ to a Brauer–Severi surface P as in (1), which establishes (4) ⇒ (1).

The implication (3) ⇒ (4) is trivial, and if E is a class as in (4), then there exists a unique section of the associated invertible sheaf on k^{sep}. This is necessarily G_k-invariant, thus, descends to a curve on X, and establishes (4) ⇒ (3). □

Remark 8.11 In [44], Skorobogatov called del Pezzo surfaces of degree 4 that satisfy condition (3) above *quasi-split*.

Before proceeding, let us recall a couple of classical results on the geometry of degree 4 del Pezzo surfaces, and refer the interested reader to [44] and [15, Chap. 8.6] for details. The anti-canonical linear system embeds X as a complete intersection of two quadrics in $\mathbb{P}^4_k$, i.e., X is given by $Q_0 = Q_1 = 0$, where Q_0 and Q_1 are two quadratic forms in five variables over k. The *degeneracy locus* of this pencil of quadrics

$$\mathrm{Deg}_X \;:=\; \{\det(t_0 Q_0 + t_1 Q_1) \;=\; 0\} \;\subset\; \mathbb{P}^1_k \;=\; \mathrm{Proj}\, k[t_0, t_1]$$

is a zero-dimensional subscheme, which is étale and of length 5 over k. Over $\overline{k}$, its points correspond to the singular quadrics containing X, all of which are cones over smooth quadric surfaces. Let $\nu_2 : \mathbb{P}^1_k \to \mathbb{P}^2_k$ be the 2-uple Veronese embedding and set

$$Z := \nu_2(\mathrm{Deg}_X) \;\subset\; C := \nu_2(\mathbb{P}^1_k) \;\subset\; \mathbb{P}^2_k.$$

If X contains a k-rational (-1)-curve, i.e., if X is quasi-split, then X is the blow-up of $\mathbb{P}^2_k$ in Z, see Theorem 8.10 and [44, Theorem 2.3].

Proposition 8.12 *Let X be a del Pezzo surface of degree 4 over a field k of characteristic $\neq 2$ with at least 5 elements. Then, the following are equivalent:*

(1) The degeneracy scheme Deg_X has a k-rational point.
(2) There exists a finite morphism $\psi : X \to S$ of degree 2, where S is a del Pezzo surface of degree 8 of product type.

Moreover, if ψ is as in (2), then S is isomorphic to a quadric in $\mathbb{P}^3_k$.

PROOF To show (1) $\Rightarrow$ (2), assume that Deg_X has a k-rational point. Thus, there exists degenerate quadric Q with $X \subset Q \subset \mathbb{P}^4_k$. As explained in the proof of [15, Theorem 8.6.8], Q is a cone over a smooth quadric surface, and the projection away from its vertex $\mathbb{P}^4_k \dashrightarrow \mathbb{P}^3_k$ induces a morphism $X \to \mathbb{P}^3_k$ that is finite of degree 2 onto a smooth quadric surface S. In particular, S is a del Pezzo surface of degree 8 of product type.

To show (2) $\Rightarrow$ (1), let $\psi : X \to S$ be as in the statement. Then, we have a short exact sequence (which even splits since $\mathrm{char}(k) \neq 2$)

$$0 \to \mathcal{O}_S \to \psi_*\mathcal{O}_X \to \mathcal{L}^{-1} \to 0,$$

where $\mathcal{L}$ is an invertible sheaf on S, which is of type $(1, 1)$ on $S_{\overline{k}} \cong \mathbb{P}^1_{\overline{k}} \times \mathbb{P}^1_{\overline{k}}$. In particular, $|\mathcal{L}|$ defines an embedding $\iota : S \to \mathbb{P}^3_k$ as a quadric, and establishes the final assertion. Now, $\iota \circ \psi$ arises from a 4-dimensional subspace V inside the linear system $(\iota \circ \psi)^*\mathcal{O}_{\mathbb{P}^3_k}(1) \cong \omega_X^{-1}$. Thus, $\iota \circ \psi$ is the composition of the anti-canonical embedding $X \to \mathbb{P}^4_k$, followed by a projection $\mathbb{P}^4_k \dashrightarrow \mathbb{P}^3_k$. As explained in the proof of [15, Theorem 8.6.8], such a projection induces a degree 2 morphism onto a quadric if and only if the point of projection is the vertex of a singular quadric in $\mathbb{P}^4_k$ containing X. In particular, this vertex and the corresponding quadric are defined over k, giving rise to a k-rational point of Deg_X. □

In order to refine Proposition 8.12, we will use conic classes as introduced in Definition 8.6.

Proposition 8.13 *Let X be a del Pezzo surface of degree 4 over a field k. Then, the following are equivalent:*

(1) There exists a conic class in $\mathrm{Pic}_{(X/k)(\mathrm{fppf})}(k)$.
(2) There exists a finite morphism $\psi : X \to P' \times P''$ of degree 2, where P' and P'' are a Brauer–Severi curves over k.

Moreover, if ψ is as in (2), then $P' \cong P''$.

PROOF Let $\mathcal{L} \in \mathrm{Pic}_{(X/k)(\mathrm{fppf})}(k)$ be a conic class. By Theorem 3.4, there exist morphisms $|\mathcal{L}| : X \to P'$ and $|\omega_X^{-1} \otimes \mathcal{L}^{-1}| : X \to P''$, where P' and P'' are Brauer–Severi curves over k. Combining them, we obtain a finite morphism $X \to P' \times P''$

of degree 2. As in the proof of (2) ⇒ (1) of Proposition 8.12 we find that $P' \times P''$ embeds into $\mathbb{P}^3$, and thus, $0 = [\mathbb{P}^3_k] = [P'] + [P''] \in \mathrm{Br}(k)$ by Proposition 5.2. This implies $[P'] = [P'']$ since these classes are 2-torsion, and thus, $P' \cong P''$ by Corollary 2.16. This establishes (1) ⇒ (2).

Conversely, let $\psi : X \to P' \times P''$ be as in (2). Then, $\psi^*(\mathcal{O}_{P'}(1) \boxtimes \mathcal{O}_{P''}(1))$ is a conic class, and (1) follows. □

8.4 Del Pezzo Surfaces of Degree 3

For these surfaces, we have the following analogue of Theorem 6.1.

Theorem 8.14 *Let X be a del Pezzo surface of degree 3 over a field k. Then, the following are equivalent:*

(1) There exists a birational morphism $f : X \to P$ to a Brauer–Severi surface, such that $f_{\overline{k}}$ is the blow-up of 6 points in general position.
(2) There exists a class $F \in \mathrm{Pic}_{(X/k)(\mathrm{fppf})}(k)$ such that

$$F_{\overline{k}} \cong \mathcal{O}_{\overline{X}}(E_1 + E_2 + E_3 + E_4 + E_5 + E_6),$$

where the E_i are disjoint (-1)-curves on $\overline{X}$.

Proof The proof is analogous to that of Theorem 8.7, and we leave the details to the reader. □

Note that if the equivalent conditions of this theorem are fulfilled, then X is not minimal. But the converse does not hold in general: If Y is a unirational, but not rational del Pezzo surface of degree 4 over k, and $y \in Y$ is a k-rational point not lying on an exceptional curve, then the blow-up $X \to Y$ in y is a non-minimal degree 3 del Pezzo surface over k with k-rational points that is not birationally equivalent to a Brauer–Severi surface over k.

By [35, Theorem 28.1], a degree 3 del Pezzo surface X is minimal if and only if $\rho(X) = 1$, i.e., $\mathrm{Pic}_{(X/k)(\mathrm{fppf})}(k) = \mathbb{Z} \cdot \omega_X$. In this case, we have $\mathrm{Am}(X) = 0$. In particular, if such a surface is birationally equivalent to a Brauer–Severi surface P, then $P \cong \mathbb{P}^2_k$ by Proposition 2.10 and Theorem 2.19. In particular, X is rational and has a k-rational point in this case.

8.5 Del Pezzo Surfaces of Degree 2

Arguing as in the proof of Theorem 8.5, it follows that if there exists a dominant and rational map $P \dashrightarrow X$ from a Brauer–Severi surface P onto a degree 2 del Pezzo surface over a perfect field k, then X has a k-rational point, and thus $\mathrm{Am}(X) = 0$.

In particular, if X is birationally equivalent to a Brauer–Severi surface, then it is rational, see also Corollary 8.4.

By work of Manin [35, Theorem 29.4], a degree 2 del Pezzo surface over a field k is unirational if it has a k-rational point not lying on an exceptional curve. Together with non-trivial refinements of [17, 42], such surfaces over finite fields are always unirational.

By Lemma 8.3, we have that $\mathrm{Am}(X)$ is at most 2-torsion for degree 2 del Pezzo surfaces. For the possibilities of $H^1(G_k, \mathrm{Pic}_{X/k}(k^{\mathrm{sep}}))$, as well as further information concerning arithmetic questions, we refer to [32].

8.6 Del Pezzo Surfaces of Degree 1

If X is a del Pezzo surface of degree 1, then it has a k-rational point, namely the unique base point of $|-K_X|$. Thus, we have $\mathrm{Am}(X) = 0$, and there are no morphisms or rational maps to non-trivial Brauer–Severi varieties.

Acknowledgements It is a pleasure for me to thank Jörg Jahnel, Andrew Kresch, Raphael Riedl, Ronald van Luijk, and Anthony Várilly-Alvarado for comments and discussions. I especially thank Jean-Louis Colliot-Thélène and Alexei Skorobogatov for providing me with references, discussions, and pointing out mistakes, as well as correcting some of my too naive ideas. Last, but not least, I thank the referee for careful proof-reading and the many useful suggestions.

References

1. S.A. Amitsur, Generic splitting fields of central simple algebras. Ann. Math. (2) **62**, 8–43 (1955)
2. M. Blunk, Del Pezzo surfaces of degree 6 over an arbitrary field. J. Algebra **323**, 42–58 (2010)
3. S. Bosch, W. Lütkebohmert, M. Raynaud, Néron models, *Ergebnisse der Mathematik und ihrer Grenzgebiete (3)*, vol. 21 (Springer, New York, 1990)
4. K.S. Brown, *Cohomology of Groups*, vol. 87, Graduate Texts in Mathematics (Springer, New York, 1982)
5. F. Châtelet, Variations sur un thème de H. Poincaré. Ann. Sci. École Norm. Sup. (3) **61**, 249–300 (1944)
6. J.-L. Colliot-Thélène, Surfaces de Del Pezzo de degré 6. C. R. Acad. Sci. Paris Sér. A-B **275**, A109–A111 (1972)
7. J.-L. Colliot-Thélène, Points rationnels sur les variétés non de type général, Course de J.-L. Colliot-Thélène á Orsay/IHP (1999), available from the author's webpage
8. J.-L. Colliot-Thélène, J.-J. Sansuc, La descente sur les variétés rationnelles. II. Duke Math. J. **54**(2), 375–492 (1987)
9. J.-L. Colliot-Thélène, A. Skorobogatov, Groupe de Chow des zéro-cycles sur les fibrés en quadriques. K-Theory **7**(5), 477–500 (1993)
10. J.-L. Colliot-Thélène, N.A. Karpenko, A.S. Merkurjev, Rational surfaces and the canonical dimension of the group PGL_6, Algebra i Analiz **19**(5), 159–178 (2007), translation in St. Petersburg Math. J. **19**(5), 793–804 (2008)
11. K.R. Coombes, Every rational surface is separably split. Comment. Math. Helv. **63**(2), 305–311 (1988)

12. D. Coray, Points algébriques sur les surfaces de del Pezzo. C. R. Acad. Sci. Paris Sér. A-B **284**, 1531–1534 (1977)
13. P. Corn, Del Pezzo surfaces of degree 6. Math. Res. Lett. **12**(1), 75–84 (2005)
14. W.A. de Graaf, M. Harrison, J. Pílniková, J. Schicho, A Lie algebra method for rational parametrization of Severi-Brauer surfaces. J. Algebra **303**(2), 514–529 (2006)
15. I. Dolgachev, *Classical Algebraic Geometry. A Modern View* (Cambridge University Press, Cambridge, 2012)
16. I. Dolgachev, A. Duncan, Regular pairs of quadratic forms on odd-dimensional spaces in characteristic 2 (2015). arXiv:1510.06803
17. D. Festi, R. van Luijk, Unirationality of del Pezzo surfaces of degree two over finite fields. Bull. Lond. Math. Soc. **48**, 135–140 (2016)
18. P. Gille, T. Szamuely, *Central Simple Algebras and Galois Cohomology*, vol. 101, Cambridge Studies in Advanced Mathematics (Cambridge University Press, Cambridge, 2006)
19. J. Giraud, Cohomologie non abélienne, *Grundlehren der mathematischen Wissenschaften*, vol. 179 (Springer, New York, 1971)
20. A. Grothendieck, Le groupe de Brauer. I. Algèbres d'Azumaya et interprétations diverses, *Dix Exposés sur la Cohomologie des Schémas* (North-Holland, Amsterdam, 1968), pp. 46–66
21. A. Grothendieck, Le groupe de Brauer. III. Exemples et compléments, *Dix Exposés sur la Cohomologie des Schémas* (North-Holland, Amsterdam, 1968), pp. 88–188
22. A. Grothendieck, Technique de descente et théorèmes d'existence en géométrie algébrique. V. Les schémas de Picard: théorèmes d'existence, Séminaire Bourbaki 7, Exp. No. 232, 143–161, Soc. Math. France (1995)
23. A. Grothendieck, Technique de descente et théorèmes d'existence en géométrie algébrique. VI. Les schémas de Picard: propriétés générales, Séminaire Bourbaki 7, Exp. No. 236, 221–243, Soc. Math. France (1995)
24. R. Hartshorne, *Algebraic Geometry*, vol. 52, GTM (Springer, New York, 1977)
25. V.A. Iskovskih, Minimal models of rational surfaces over arbitrary fields. Izv. Akad. Nauk SSSR Ser. Mat. **43**(1), 19–43 (1979)
26. N. Jacobson, *Finite-Dimensional Division Algebras Over Fields* (Springer, New York, 1996)
27. J. Jahnel, The Brauer–Severi variety associated with a central simple algebra: a survey, Linear Algebraic Groups and Related Structures, vol. 52 (2000), pp. 1–60, available from the author's webpage
28. J.-P. Jouanolou, Théorèmes de Bertini et applications, *Progress in Mathematics*, vol. 42 (Birkhäuser, Boston, 1983)
29. M.C. Kang, Constructions of Brauer-Severi varieties and norm hypersurfaces. Can. J. Math. **42**(2), 230–238 (1990)
30. S.L. Kleiman, The Picard scheme, *Fundamental Algebraic Geometry*, vol. 123, Mathematical Surveys and Monographs (AMS, Providence, 2005), pp. 235–321
31. A. Knecht, Degree of unirationality for del Pezzo surfaces over finite fields (2013). arXiv:1307.3215
32. A. Kresch, Y. Tschinkel, On the arithmetic of del Pezzo surfaces of degree 2. Proc. Lond. Math. Soc. (3) **89**(3), 545–569 (2004)
33. S. Lang, Some applications of the local uniformization theorem. Am. J. Math. **76**, 362–374 (1954)
34. Y.I. Manin, Rational surfaces over perfect fields. Inst. Hautes Études Sci. Publ. Math. **30**, 55–113 (1966)
35. Y.I. Manin, *Cubic Forms. Algebra, Geometry, Arithmetic*, vol. 4, 2nd edn., North-Holland Mathematical Library (North-Holland Publishing, Amsterdam, 1986)
36. J.S. Milne, The Brauer group of a rational surface. Invent. Math. **11**, 304–307 (1970)
37. J.S. Milne, *Étale Cohomology*, vol. 33, Princeton Mathematical Series (Princeton University Press, Princeton, 1980)
38. J. Neukirch, A. Schmidt, K. Wingberg, Cohomology of number fields, *Grundlehren der mathematischen Wissenschaften*, vol. 323, 2nd edn. (Springer, New York, 2008)

39. H. Nishimura, Some remarks on rational points. Mem. Coll. Sci. Univ. Kyoto. Ser. A. Math. **29**, 189–192 (1955)
40. M. Pieropan, On the unirationality of del Pezzo surface over an arbitrary field, Master thesis (2012)
41. B. Poonen, Rational points on varieties, available from the author's webpage
42. C. Salgado, D. Testa, A. Várilly-Alvarado, On the unirationality of del Pezzo surfaces of degree 2. J. Lond. Math. Soc. (2) **90**(1), 121–139 (2014)
43. A. Skorobogatov, *Torsors and Rational Points*, vol. 144, Cambridge Tracts in Mathematics (Cambridge University Press, Cambridge, 2001)
44. A. Skorobogatov, Del Pezzo surfaces of degree 4 and their relation to Kummer surfaces. Enseign. Math. (2) **56**, 73–85 (2010)
45. P. Swinnerton-Dyer, The Brauer group of cubic surfaces. Math. Proc. Camb. Philos. Soc. **113**, 449–460 (1993)
46. A. Várilly-Alvarado, Weak approximation on del Pezzo surfaces of degree 1. Adv. Math. **219**(6), 2123–2145 (2008)
47. A. Várilly-Alvarado, Arithmetic of del Pezzo surfaces, *Birational Geometry, Rational Curves, and Arithmetic* (Springer, New York, 2013), pp. 293–319

(%

- Algebra, Arithmetic, and Geometry-
Volume II... (2009).pdf
Pág. 23(35)
Cyclic Homology with Coefficients
Pág. 33(45) 3
Gauss-Manin connection.
- Geometry over Nonclosed Field (2017).pdf
- Birational Geometry, Rational Curves, and Arithmetic... (2013).pdf
Pág. 133(141)
Curves of Low Degrees on
Fano Varieties

Comparar:
Geometry over nonclosed field...pdf
Pág. 174(180) 4
Del Pezzo Surfaces

Pezzo surfaces with a view towards
Brauer-Severi varieties.

Look: Pág. 187(193)

8. Del Pezzo Surfaces of
Small Degree.

(8.1) Birationality to
Brauer-Severi Surfaces
(Severi)

↳ Comparar
com o livro:

Birational Geometry, Rational
Curves, and Arithmetic.(2013).pdf

Pág. 93(101)
Birational Geometry via
Moduli Spaces

and

Pág. 133(141)
Curves of Low Degrees
on
Fano Varieties.

+,-.+ + +/ ! + * < ! ! 1 2P -! ! 8 P
! # , 3 ! u t = ' 3#
(%'

Birational Geometry via Moduli Spaces

Ivan Cheltsov, Ludmil Katzarkov, and Victor Przyjalkowski

Mathematics Subject Classification codes (2000): 14M25, 14H10, 14Q15, 14D07

1 Introduction

1.1 Moduli Approach to Birational Geometry

In recent years, we have witnessed many significant developments in the Minimal Model Program (MMP) (see, e.g., [10]), based on major advances in the study of singularities of pairs. An alternative approach to MMP, relying on the correspondence between Mori fibrations (MF) and semiorthogonal decompositions (SOD), was proposed by Kawamata [67]; it does not involve the study of discrepancies or the effective cone.

In the meantime, a new epoch, the epoch of wall-crossing, has emerged. At present, after papers of Seiberg–Witten [91], Gaiotto–Moore–Neitzke [40], and Cecotti–Vafa [20] and seminal works by Donaldson–Thomas [33], Joyce–Song [62],

---

I. Cheltsov
University of Edinburgh, Edinburgh, EH1 2QL, UK
e-mail: I.Cheltsov@ed.ac.uk

L. Katzarkov
University of Miami, Coral Gables, FL, USA

University of Vienna, Vienna, Austria
e-mail: lkatzark@math.uci.edu

V. Przyjalkowski (✉)
Steklov Mathematical Institute, Moscow, Russia
e-mail: victorprz@mi.ras.ru

F. Bogomolov et al. (eds.), *Birational Geometry, Rational Curves, and Arithmetic*, DOI 10.1007/978-1-4614-6482-2_5,
© Springer Science+Business Media New York 2013

Maulik–Nekrasov–Okounkov–Pandharipande [79, 80], Douglas [35], Bridgeland [13], and Kontsevich–Soibelman [72, 73], the situation with the wall-crossing phenomenon is quite similar to the situation with the theory of Higgs Bundles after Higgs and Hitchin: it was clear that a general "Hodge type" theory existed and that it needed to be developed. This led to major mathematical applications, for example, in uniformization and in the Langlands program, to mention a few. Similarly, in the framework of wall-crossing, it is also apparent that a "Hodge type" theory needs to be developed in order to reap mathematical benefits, e.g., solve some long-standing problems in algebraic geometry.

Already, several attempts have been made to connect homological mirror symmetry (HMS) to birational geometry. For instance, the paper [1] was motivated by the idea that a proof of HMS is likely to involve a detailed study of birational transformations (including, noncommutative) on the A and B sides of HMS. Later, new ideas were introduced in [4, 27, 28, 45, 68, 70]. These can be summarized as follows:

1. The moduli spaces of stability conditions of Fukaya–Seidel categories can be included in a one-parametric family with the moduli space of Landau–Ginzburg (LG) models as a central fiber.
2. The moduli space of LG models determines the birational geometry; this was proved in the toric case in [27].

In this paper, we simultaneously consider all Mori fiber spaces, all Sarkisov links and all relations between Sarkisov links, and explore these in terms of the geometry of moduli space of LG models and the moving schemes involved.

We introduce the following main ideas:

1. All Fano varieties are connected, via their degenerations, by simple basic links—projections of a special kind. We show this for Fano threefolds of Picard rank 1 (see Table 2). This agrees with their toric Landau–Ginzburg models (see Theorem 3.12).
2. All Fano manifolds can be considered together, i.e., there exists a big moduli space of LG models, which includes mirrors of all Fano's. We demonstrate this partially for two- and three-dimensional Fano's.
3. We introduce an analog of the canonical divisor measure for a minimal model. For us, this is the local geometry of singularities and a fiber at infinity of the LG model. The last one affects the geometry of moduli spaces of LG models, the stability conditions. We propose local models for these moduli spaces (stacks). The correspondence between the usual and the categorical approach to birational geometry is displayed in Table 1.
4. Following [28] and the pioneering work [12], we develop the notion of phantom category, emphasizing its connection with the notion of a moving scheme, defined in this paper. Moving schemes determine the geometry of the moduli space of LG models and thus the geometry of the initial manifold. We conjecture:

Table 1 Extended Kawamata program

Classical	Derived
Mori fibrations	SOD or circuits
$(K_X + \Delta)$ log differentials	Moving schemes of the fiber at ∞ in Landau–Ginzburg model described by differentials
Sarkisov links	Two-dimensional faces
Relations between Sarkisov links	Three-dimensional faces

A. For classical surfaces of general type (Campedelli, Godeaux, Burniat, Dolgachev surfaces, quotients of products of higher-genus curves and fake $\mathbb{P}^2$s) there should exist *quasi-phantoms* in their SOD, i.e., nontrivial categories with trivial Hochschild homology. These surfaces are clearly not rational as they have nontrivial fundamental groups, but also since, conjecturally, they have quasi-phantom subcategories in their SOD. On the Landau–Ginzburg side, these quasi-phantoms are described by the moving scheme. The deformation of the Landau–Ginzburg models is determined by the moving scheme so the quasi-phantoms factor in the geometry of the Landau–Ginzburg models. On the mirror side, this translates to the fact that the Ext groups between the quasi-phantom and the rest of the SOD determine the moduli space.

B. For surfaces of general type with trivial fundamental groups (e.g., the Barlow surface, see [28]), we conjecture the existence of a nontrivial *phantom category*, a category with a trivial K^0 group. The deformation of the Landau–Ginzburg models is determined by the moving scheme, so the phantom factors in the geometry of the Landau–Ginzburg models. Similarly, on the mirror side this translates to the fact that the Ext groups between the phantom and the rest of the SOD determine the moduli space.

C. We relate the existence of such phantom categories with nonrationality questions. For surfaces, it is clear that phantoms lead to nonrationality. For threefolds, we exhibit examples (of Sarkisov type) of nonrational threefolds where phantoms conjecturally imply nonrationality. We introduce "higher" categorical invariants detecting nonrationality: gaps of spectra. Conjecturally, these are not present in the Sarkisov examples. These ideas are a natural continuation of [50].

5. We introduce conjectural invariants associated to our moduli spaces—gaps and local differentials. We suggest that these numbers (changed drastically via "wall-crossing") produce strong birational invariants. We relate these invariants to the "Hassett–Kuznetsov–Tschinkel" program (see [46,77])—a program for studying rationality of the four-dimensional cubic and its "relatives."

This paper is organized as follows. In Sects. 2 and 3 we relate degenerations of Fano manifolds via projections. Using mirror symmetry in Sect. 3 we transfer

/4 ?! ! # 8!3 7

these connections to the side of Landau–Ginzburg model. Based on that, in Sect. 4 we suggest a generalization of Kawamata's categorical approach to birational geometry, enhancing it via the geometry of moduli spaces of Landau–Ginzburg models. We give several applications, most notably a conjectural application to the Hassett–Kuznetsov–Tschinkel program. Our approach is based on two categorical nonrationality invariants—phantoms and gaps. Full details will appear in a future paper.

Notations. Smooth del Pezzo threefolds (smooth Fano threefolds of index 2) are denoted by V_n, where n is its degree with respect to a Picard group generator, except for the quadric denoted by Q. Fano threefolds of Picard rank 1, index 1, and degree n are denoted by X_n. The remaining Fano threefolds are denoted by $X_{k.m}$, where k is a Picard rank of a variety and m is its number according to [57]. The Laurent polynomials from the k-th line of Table 2 are denoted by f_k. A toric variety whose fan polytope is a Newton polytope of f_k is denoted by F_k or just a variety number **k**. The Landau–Ginzburg model for a variety X is denoted by $LG(X)$.

2 "Classical" Birational Geometry

In this section we recall some facts from classical birational geometry of three-dimensional Fano varieties. Our presentation of this geometry is adopted to HMS.

2.1 The Importance of Being Gorenstein

Among singular Fano varieties, those with canonical Gorenstein singularities are of special importance. They arise in many different geometrical problems: degeneration of smooth Fano varieties with a special regard to mirror symmetry (see [6,7]), classification of reflexive polytopes (see [74,75]), midpoints of Sarkisov links and bad Sarkisov links (see [22,24]), compactification of certain moduli spaces (see [81]), etc. Historically, Fano varieties with canonical Gorenstein singularities are the original Fano varieties. Indeed, the name *Fano varieties* originated in the works of Iskovskikh (see [54, 55]) that filled the gaps in old results by Fano who studied in [36, 37] anticanonically embedded Fano threefolds with canonical Gorenstein singularities without naming them so (cf. [15]).

In dimension two, canonical singularities are always Gorenstein, so being Gorenstein is a vacuous condition. Surprisingly, the classification of del Pezzo surfaces with canonical singularities is simpler than the classification of smooth del Pezzo surfaces (see [26, 84]). Fano threefolds with canonical Gorenstein singularities are not yet classified, but first steps in this directions are already have been made by Mukai, Jahnke, Radloff, Cheltsov, Shramov, Przyjalkowski, Prokhorov, and Karzhemanov (see [17, 60, 65, 66, 81, 85]).

Curves of Low Degrees on Fano Varieties

Olivier Debarre

Mathematics Subject Classification codes (2010): 14C05, 14C30, 14C34, 14D20, 14E05, 14E08, 14E20, 14H10, 14J10, 14J30, 14J35, 14J45, 14J60, 14J70, 14M20, 14M22, 14N25

1 Introduction

We work over the field of complex numbers. If $X \subset \mathbf{P}^N$ is a smooth projective variety, we let $\mathscr{C}_d^g(X)$ be the (quasi-projective) subscheme of the Hilbert scheme of X corresponding to smooth, irreducible, genus-g, degree-d curves on X.

These schemes, and their compactifications $\overline{\mathscr{C}}_d^g(X)$, particularly for low d and g, have proved very useful in the study of the geometry and the period maps of some Fano varieties of low degrees.

In this informal survey article, we explain, mainly without proofs, what is known (or conjectured) in the following cases: cubic threefolds, Fano threefolds of degree 14 and index 1, cubic fourfolds, Fano varieties of degree 10, coindex 3, and dimensions 3, 4, or 5.

In the case of a smooth cubic threefold $X \subset \mathbf{P}^4$ (Sect. 2), Hilbert schemes of curves have long been used to parametrize useful subvarieties of the intermediate Jacobian $J(X)$. For example, the (smooth projective) surface $\mathscr{C}_1^0(X)$ parametrizing lines contained in X was an essential ingredient in the proofs by Clemens and Griffiths of the non-rationality of X and of the Torelli theorem (that $J(X)$ determines X). The Abel–Jacobi maps $\mathrm{aj} : \overline{\mathscr{C}}_d^g(X) \to J(X)$ have been since proved in some cases to be mrc fibrations (see Sect. 2 for definitions and results).

O. Debarre (✉)
Département de Mathématiques et Applications – CNRS UMR 8553, École Normale Supérieure, 45 rue d'Ulm, 75230 Paris cedex 05, France
e-mail: Olivier.Debarre@ens.fr

F. Bogomolov et al. (eds.), *Birational Geometry, Rational Curves, and Arithmetic*, DOI 10.1007/978-1-4614-6482-2_6,
© Springer Science+Business Media New York 2013

Curves on Fano threefolds can also be used to construct interesting birational correspondences. This is illustrated in Sect. 2.9, where it is explained that Fano threefolds of degree 14 and index 1 are birational to cubic threefolds. The situation is very rich and gives rise to a beautiful description of the period maps for these Fano threefolds.

In Sect. 3, we move on to smooth cubic fourfolds $Y \subset \mathbf{P}^5$, for which there are no intermediate Jacobians. Nevertheless, the mrc fibrations of $\overline{\mathscr{C}}^g_d(Y)$ should be substitutes for the Abel–Jacobi maps and should provide interesting information. In another direction, ideas of Mukai show how to construct *symplectic forms* on some moduli spaces of sheaves (Sect. 3.1). This can be used to prove that the space of lines $\mathscr{C}^0_1(Y)$ is a symplectic fourfold (Beauville and Donagi), whose Hodge structure is closely related to that of Y. This was the basis of Voisin's proof of the injectivity of the period map for smooth cubic fourfolds. This period map is now completely described by work of Looijenga and Laza and identifies the moduli space of smooth cubic fourfolds with an explicit open subset of the Baily–Borel compactification of the period domain (Sect. 3.2). Analogous results are available for the moduli space of smooth cubic threefolds (Sect. 3.3).

Mukai's construction was also shown to apply to Fano fourfolds X^4_{10} of degree 10 and index 2 (Example 3.5). In this case, the scheme $\overline{\mathscr{C}}^0_2(X^4_{10})$ leads to the construction of a symplectic fourfold called a *double EPW sextic* (Iliev and Manivel, O'Grady).

In Sect. 4, we study in some details the period maps of these Fano varieties of degree 10, in dimensions 3, 4, and 5. There is beautiful geometry at work here, with intriguing interplay with that of EPW sextics (work in progress with Iliev and Manivel).

2 Cubic Threefolds

The *maximal rationally connected fibration* (mrc fibration for short) of a smooth proper (complex) variety Y is a rational dominant map $\rho : Y \dashrightarrow R(Y)$ such that for $z \in R(Y)$ very general, the fiber $\rho^{-1}(z)$ is proper and rationally connected (two general points can be joined by a rational curve), and any rational curve in Y that meets $\rho^{-1}(z)$ is contained in $\rho^{-1}(z)$. The mrc fibration exists and is unique up to birational equivalence. The variety $R(Y)$ is called the reduction of Y.

Let $X \subset \mathbf{P}^4$ be a general (although some results are known for *any smooth* X) cubic hypersurface. The intermediate Jacobian

$$J(X) := H^{2,1}(X)^\vee / H_3(X, \mathbf{Z})$$

is a 5-dimensional principally polarized abelian variety. We let $\mathscr{X}_3$ be the (irreducible, 10-dimensional) moduli space for smooth cubic threefolds, and we let $\mathscr{A}_n$ be the moduli space for principally polarized abelian varieties of dimension n. Taking X to $J(X)$ defines the *period map* $\mathscr{X}_3 \to \mathscr{A}_5$. We have:

- For $g = 0$ or $d \leq 5$, $\mathscr{C}^g_d(X)$ is integral of dimension $2d$ ([7], Theorem 1.1);
- For $d \leq 5$, the *Abel–Jacobi map* $\mathrm{aj} : \overline{\mathscr{C}}^g_d(X) \to J(X)$ induces the mrc fibration ([6], Theorem 1.1);
- For $d \geq 4$, $\mathrm{aj} : \overline{\mathscr{C}}^0_d(X) \to J(X)$ is dominant with irreducible general fibers.

It is natural to ask whether the Abel–Jacobi map $\mathrm{aj} : \overline{\mathscr{C}}^0_d(X) \to J(X)$ is the mrc fibration for all $d \geq 4$ (i.e., are general fibers rationally connected?).

2.1 Lines

- $\mathscr{C}^0_1(X)$ (which parametrizes lines on X) is a smooth projective irreducible surface of general type;
- The image of $\mathrm{aj} : \mathscr{C}^0_1(X) \to J(X)$ is a surface S with minimal class $[\Theta]^3/3!$ and $S - S$ is a theta divisor Θ;
- aj induces an isomorphism between the Albanese variety of $\mathscr{C}^0_1(X)$ and $J(X)$ ([2]).

The second item yields a proof of Torelli: *the period map* $\mathscr{X}_3 \to \mathscr{A}_5$ *is injective.*

2.2 Conics

- $\overline{\mathscr{C}}^0_2(X)$ is a smooth projective irreducible fourfold;
- $\mathrm{aj} : \overline{\mathscr{C}}^0_2(X) \to J(X)$ is a $\mathbf{P}^2$-bundle over the surface S of Sect. 2.1 (a conic is uniquely determined by the plane that it spans and the residual line in X). The mrc fibration is therefore $\overline{\mathscr{C}}^0_2(X) \to S$.

2.3 Plane Cubics

- $\overline{\mathscr{C}}^1_3(X)$ is isomorphic to the Grassmannian $G(3,5)$;
- The Abel–Jacobi map is therefore constant.

2.4 Twisted Cubics

- $\mathrm{aj} : \mathscr{C}^0_3(X) \to J(X)$ is birational to a $\mathbf{P}^2$-bundle over a theta divisor Θ ([6], Sect. 4). The mrc fibration is therefore $\overline{\mathscr{C}}^0_3(X) \to \Theta$.

(%

Pezzo surfaces with a view towards
Brauer-Severi varieties.

Look: Pág. 187(193)
8. Del Pezzo Surfaces of
Small Degree.
(8.1) Birationality to
Brauer-Seweri Surfaces
(Severi)
↳ Comparar
com o livro:
Birational Geometry, Rational
Curves, and Arithmetic. (2013).pdf
Pág. 93(101)
Birational Geometry via
Moduli Spaces
and
Pág. 133(141)
Curves of Low Degrees
on
Fano Varieties.

— A Geometry of Approximation.. (2008).pdf

Obs: Usar a estrutura de pesquisa do livro para dirimir flutuações incongruências geométricas da minha pesquisa.

— Geometry and Algebra of Multidimensional Three-Webs ... (1992).pdf

Pág. XIII (11) Preface

Wilhelm Blaschke e T. Thomsen

New topic in differential geometry — WEB GEOMETRY (Web Geometry)

→ Applications of ideas de Felix Klein to differential geometry.

Klein's "Erlangen program" (1872).

According to this program, every geometry studies properties of geometric figure that are invariant under transformations composing a certain group.

From this point of view, the

foundations of affine differential geometry and conformal differential geometry were presented in Blaschke's books, published in the 1920's.

Blaschke extended the same group-theoretic approach to web geometry.

→ pseudogroup

All differentiable transformations of a manifold, where a web is given, as a fundamental group, and studied local web invariants relative to transformation of this group.

Em minha pesquisa, haverá a projeção ortogonal (one-to-one) sobre um plano.

Figure 0.1

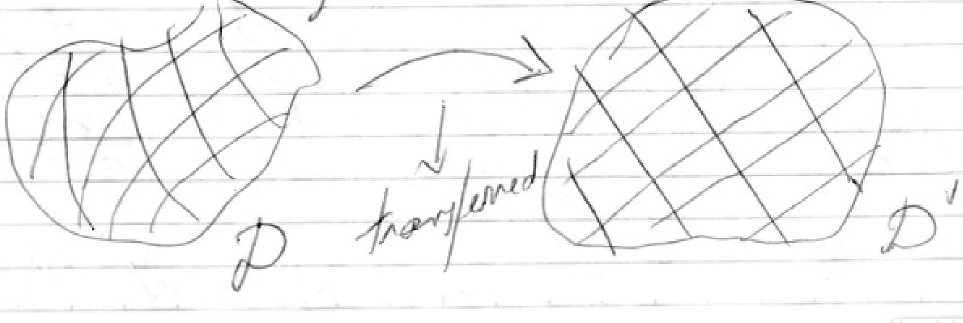

+,-.+ + +/ ! + * , 1 2 ,33 I !
- 0 2 > P , 4 ! ! # 0 3 2 3 ! ! 3 3#

+,-.+ + +/ ! + * 1 2 ! # , 4 !
! 0 z 4 3 u ! # %% 3#
(O&&& !

Preface

1. In 1927–1928 the papers of Wilhelm Blaschke, who was a well-known geometer by that time, and his student T. Thomsen appeared (see [Bl 28] and [T 27]). The papers were devoted to a new topic in differential geometry – web geometry. Many of Blaschke's previous papers were connected with applications of ideas of Felix Klein to differential geometry. These ideas were formulated in Klein's famous "Erlangen program" (1872). According to this program, every geometry studies properties of geometric figures that are invariant under transformations composing a certain group. From this point of view, the foundations of affine differential geometry and conformal differential geometry were presented in Blaschke's books, published in the 1920's.

Blaschke extended the same group-theoretic approach to web geometry. He took the group (more precisely, pseudogroup) of all differentiable transformations of a manifold, where a web is given, as a fundamental group, and studied local web invariants relative to transformations of this group.

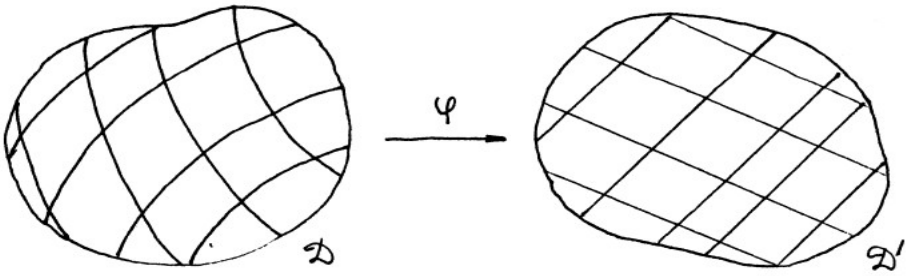

Figure 0.1

Let us explain this approach to web geometry in more detail. Consider a regular family of smooth curves in a two-dimensional domain D. By means of an appropriate differentiable transformation, the domain D can be transferred into a domain D' of an affine plane A^2 in such a way that the family of lines given in D will be transferred into a family of parallel lines of D'. This shows that a family of smooth curves in D does not have local invariants. Two regular families of smooth curves, that are in general position in D, also do not have local invariants since one can always find a diffeomorphism that transfers them into two families of parallel lines of a domain D' of an affine plane A^2 (see Figure 0.1). Thus, the structures, defined in D by one or two families of curves, are locally trivial.

Consider now three regular families of smooth curves in D, that are in general position in D, i.e. form a three-web in the domain D. Such a structure is no longer locally trivial. In fact, in a neighborhood of each point p of the domain D, where the web is given, one can construct a family of hexagonal figures as shown in Figure 0.2.

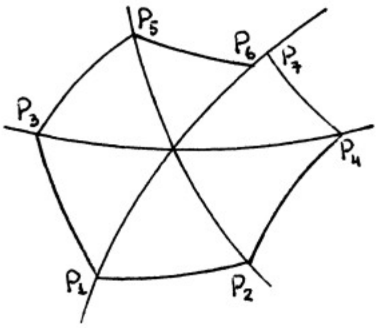

Figure 0.2

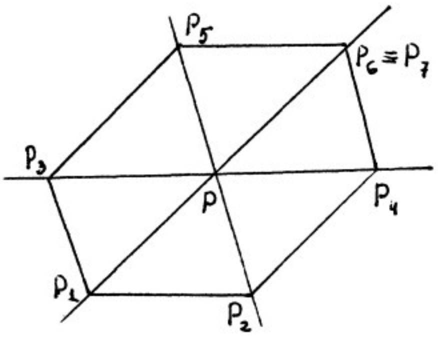

Figure 0.3

In this figure the lines p_1p_6, p_3p_5, p_2p_4 belong to the first family, the lines p_2p_5, p_1p_3, p_4p_7 belong to the second family, and the lines p_3p_4, p_1p_2, p_5p_6 belong to the third family. In the general case, the points p_6 and p_7 of these figures do not coincide, i.e. hexagonal figures are not closed. However, there exist three-webs, on which all hexagonal figures are closed, i.e. on these webs the points p_6 and p_7 coincide. Such three-webs are called hexagonal. For example, the web, formed by three families of parallel lines in an affine plane A_2, is hexagonal (see Figure 0.3). It is remarkable that the converse theorem also holds: any hexagonal three-web in a plane admits a differentiable mapping on a three-web, formed by three families of parallel lines, i.e. such a web is parallelizable. This theorem was first proved by Thomsen in the above mentioned paper [T 27] (see also the book [Bl 55]).

Since the fundamental group in web geometry is the group of arbitrary differentiable mappings, Blaschke refers the theory of webs as one of "topological questions of differential geometry" (Topologische Fragen der Differentialgeometrie). Numerous papers of Blaschke, his students and co-workers as well as the well-known monograph *Web Geometry* (*Geometrie der Gewebe*) [BB 38], written jointly by Blaschke and Bol and published in 1938, were devoted to this subject.

Blaschke noticed the connections of the web theory with many branches of geometry, in particular, with projective geometry. In a projective plane P^2 it is natural to consider a three-web formed by three one-parameter families of curves, each enveloping a curve X_α, $\alpha = 1, 2, 3$. In this case, one straight line from each family passes through a point p of a domain $D \subset P^2$ (Figure 0.4). By a correlative transformation of the plane P^2, these families of straight lines will be transferred into three curves X^*_α in a dual plane P^{2*}, and to each point p there corresponds a straight line p^* (see Figure 0.5). To straight lines of the original rectilinear three-web there correspond pencils in the plane P^{2*} with centers on the curves X^*_α. For this configuration the following remarkable theorem X_α in a plane P^2 is hexagonal if and only if the triple of curves X^*_α in the plane P^{2*}, which is dual to this web, belongs to a cubic curve. This theorem is called the Graf–Sauer theorem and follows from the Chasles' theorem on plane cubic curves. Its proof can be found in the book [Bl 55].

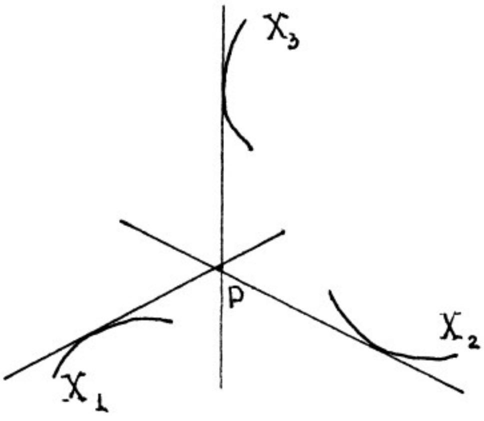

Figure 0.4

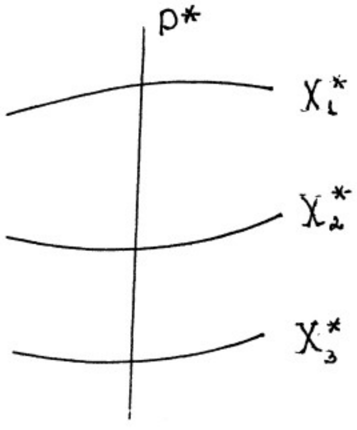

Figure 0.5

At the very beginning of the development of web theory its connection with algebra was discovered. In 1928 and 1932 the papers of K. Reidemeister [R 28] and H. Kneser [Kn 32] were published, in which it was established that a structure that has many of the properties of a three-web, can be constructed from an arbitrary group G. Namely, consider a cartesian square $X = G \times G$, whose elements are pairs (x, y) of elements of G. Call these pairs elements of the considered "three-web". Consider subsets $(x, b), (a, y), \{(x, y) | x \cdot y = c\}$ of the set X, where a, b, c are constants and x, y are variable elements of the group G, and call them lines of this "three-web". In this case, through each point of the web, there pass three lines from different families, and two lines from any two different families intersect at a point. The last two properties of this three-web allow us to consider it as an incidental structure consisting of points and lines. Because of this, the theory of such abstract webs is developed in the same way as the abstract theory of projective planes [Pi 75]. The only difference is that instead of the Pappus and Desargues configurations, which are considered in projective planes, in web theory the Reidemeister and Thomsen figures arise. We will consider these configurations in detail in Chapter 2.

Later on, Bol in his paper [B 37] observed that for construction of the abstract three-web of this type it is possible to use a groupoid of more general kind than a group. At present this kind of groupoid is called a quasigroup. Bol introduced more general closure figures on an abstract three-web, which are called Bol figures. He found that these figures are connected with certain weakened associativity conditions in a quasigroup generating a three-web.

However, it turned out that all these closure conditions cannot be geometrically realized by means of two-dimensional geometric webs, which were considered in the papers of Blaschke and Thomsen: on such webs the hexagonality condition – the weakest among the classical closure conditions – implies the web parallelizability and consequently the realization of all other closure conditions on the web.

Nevertheless all closure conditions indicated above can be realized on multidimensional geometric webs. This was a stimulus for their study. The first works in theory of multidimensional webs were the Bol paper [B 35], where a three-web on a four-dimensional manifold was considered, and the Chern paper [C 36a], where the theory of invariants of a three-web of dimension $2r$ was constructed. These two papers were published in Germany in the 1930's, and after them, for a long time, there were no publications devoted to theory of multidimensional three-webs.

Only at the end of 1960's, was the study of multidimensional three-webs continued. In 1969 the first author of this book published the paper [A 69b], and this paper was followed by an extensive series of his papers as well as his students' papers. At approximately the same time M.A. Akivis' geometry seminar at Moscow Insitute of Steel and Alloys began. Until this day, all of the most important problems of web theory are presented and discussed in this seminar.

Due to the effort of the second author, since 1981 Kalinin State University has regularly published collections of papers under the title *Webs and Quasigroups*. As a result, extensive material in the theory of multidimensional three-webs has been accumulated. The main problems of this theory as well as its connections to related problems of geometry and algebra are presented in the current monograph.

However, because of the size restriction, we cannot include everything we wanted to discuss in this book. Some of the results are given in the book without proofs. Other results are listed only in bibliographical notes and in the Bibliography.

2. We will make a few general remarks for readers of this book. First of all, note that the book is intended for graduate students whose field is differential geometry, and for mathematicians and teachers conducting research in this subject. This book can also be used for a few special courses for graduate students in Mathematics.

In our presentation of material we use the tensorial methods in combination with the methods of exterior differential forms and moving frames of Élie Cartan. The reader is assumed to be familiar with these methods as well as with the basics of modern differential geometry. Many notions of differential geometry are explained briefly in the text and some are given without any explanation. As references, the books [KN 63], [St 83], [CaH 67] and [V 87] are recommended.

All functions, vector and tensor fields and differential forms are considered differentiable sufficiently many times. All variables are assumed to be real, although in many cases this does not exclude a possibility of considering complex variables. For this reason we use the notation $\mathbf{GL}(n)$ here instead of $\mathbf{GL}(n, \mathbf{R})$ or $\mathbf{GL}(n, \mathbf{C})$ as often done.

The book consists of eight chapters. Sections, formulas and figures are numbered within each chapter. Each chapter is accompanied by a set of problems and notes with remarks of historical and bibliographical nature. A sufficiently complete bibliography, a list of notations and index are given at the end of the book.

We asked E.V. Ferapontov to write an appendix devoted to applications of web geometry to some problems of Mathematical Physics. We hope that this supplement enlargens the circle of the readers of our monograph and attracts physicists to the problems considered in the book.

In conclusion, we express our sincere gratitude to V.V. Goldberg, with whom we discussed different aspects of this book in the process of our work on the book, to V.F. Kirichenko and V.V. Timoshenko, who carefully read the manuscript and made many useful suggestions, and to L.V. Goldstein and V.B. Lazareva for their invaluable assistance in preparing the manuscript for publication.

Moscow, C.I.S. *Maks A. Akivis*

Tver, C.I.S. *Alexander M. Shelekhov*

— A Geometry of Approximation… (2008).pdf

Obs: Usar a estrutura de pesquisa do livro para diminuir flutuações incongruências geométricas da minha pesquisa.

— Geometry and Algebra of Multidimensional three-Webs … (1992).pdf

Pág. XIII (11) Preface

Wilhelm Blaschke e T. Thomsen

New topic in differential geometry — WEB GEOMETRY (Web Geometry)

→ Applications of ideas de Felix Klein to differential geometry.

Klein's "Erlangen program" (1872).

According to this program, every geometry studies properties of geometric figure that are invariant under transformations composing a certain group.

From this point of view, the

foundations of affine differential geometry and conformal differential geometry were presented in Blaschke's books, published in the 1920's.

Blaschke extende the same group-theoretic approach to web geometry.

→ pseudogroup

All differentiable transformations of a manifold, where a web is given, as a fundamental group, and studied local web invariants relative to transformation of this group.

Em minha pesquisa, haverá a projeção ortogonal (one-to-one) sobre um plano.

Figure 0.1

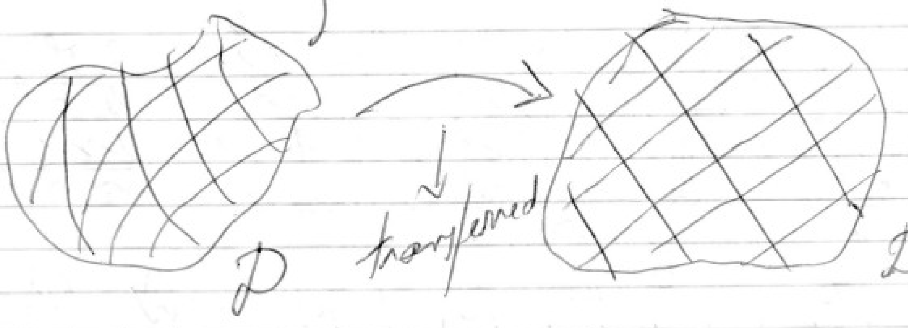

(%"

- Introductory Calculus...(1972).pdf
- Space-Time structure - Algebra and Geometry...(2007).pdf

1820 - The Arithmometer (Calculadora)
Patenteada por Thomas de Colmar
Fabricada entre 1851-1915

Iridium melting point of
4,435°F (2,466°C).

Books:

Books-019-2021/Herbert Meschkowski/

- Ungelöste und unlösbare Probleme der Geometrie...(1960).pdf

Pág. 92(99) VII.

Die Zerlegung von Rechtecken in inkongruente Quadrate
(A decomposição de retângulos em incongruentes quadrados).

- Geometry of time-spaces non-commutative algebraic geometry...pdf

+,-.+ + +/ ! + * & # 2 ! z , ! 2
1 2 ! # > ! , 4 ! A 8 & P , !# & % 3#

7 +,-.+ + +/ ! + * 3! , 4 ! ! #
2 3#

= +,-.+ + +/ ! + * 9 s # s 4! 4
1 {0 4 ! %) 3#
(% %% &&

VII. Die Zerlegung von Rechtecken in inkongruente Quadrate

1. Das Problem

Im 19. und 20. Jahrhundert ist es der Forschung gelungen, die Unlösbarkeit einiger Probleme nachzuweisen, mit denen sich die Mathematiker viele Jahrhunderte hindurch beschäftigt hatten. Dazu gehört die Quadratur des Zirkels (Kap. X), das Parallelenproblem und die Dreiteilung des Winkels (Kap. XI). Auf den Gedanken, daß das vorliegende Problem vielleicht unlösbar sei, kam man erst, nachdem Generationen von Mathematikern sich vergebens um die Lösung bemüht hatten.

Manchmal kommt man gerade auf dem umgekehrten Weg zu neuen Einsichten. Als die Frage auftauchte, ob es möglich sei, ein Quadrat in lauter inkongruente Teilquadrate zu zerlegen, waren viele Mathematiker der Ansicht, daß diese Aufgabe nicht lösbar sei. So heißt es noch im Jahre 1930 bei *Kraitschik* [VII 2] (S. 272):

> „... il n'est pas possible de décomposer un carré donné en un nombre fini de carrés inégaux deux à deux. Cette proposition quoique non démontrée paraît être vraie. Elle nous a été communiquée par M. Lusin, professeur à Moscou."

Auch *R. Sprague* beschäftigte sich mit dem Problem zunächst in der Absicht, seine Unlösbarkeit zu beweisen. Er fand[1]) aber dann eine Zerlegung des Quadrats in 55 verschiedene Teilquadrate [VII 4]. Ein Jahr später wurde eine umfangreiche Arbeit von *R. L. Brooks*, *C. A. B. Smith*, *A. H. Stone* und *W. T. Tutte* veröffentlicht, die u. a. die Zerlegung in nur 26 inkongruente Teilquadrate brachte [VII 7].

In den mathematischen Zeitschriften werden Zerlegungsprobleme der hier besprochenen Art gewöhnlich als „Unterhaltungsmathematik" registriert. Es ist aber beachtenswert, daß unser Problem eine technische Bedeutung hat, die auf den ersten Blick gar nicht ersichtlich ist. In den Arbeiten [VII 7] und [VII 10] wird das mathematische Problem sofort elektrotechnisch gedeutet, und in den Beweisen wechseln mathematische und physikalische Argumentationen. Wir wollen in diesem Kapitel unser Problem zunächst rein mathematisch angreifen und erst am Schluß auf die Möglichkeit einer technischen Deutung eingehen.

---

[1]) In der Literatur wird manchmal berichtet (z. B. [VII 10]), daß die Arbeit [VII 7] die erste Zerlegung eines Quadrats in inkongruente Teilquadrate brachte. Das ist aber nicht richtig. — Nach [VII 10] ist die erste Zerlegung eines *Rechtecks* in inkongruente Teilquadrate im Jahre 1925 gefunden worden. Das Buch [VII 2] von *Kraitschik* vom Jahre 1930 enthält als Beispiel nur eine solche Zerlegung, in der Paare gleicher Quadrate und auch zwei gleiche Quadrate mit einer gemeinsamen Seite auftreten.

Beginnen wir mit einigen Verabredungen über eine zweckmäßige Bezeichnung! Abb. 30 zeigt die Zerlegung eines Rechtecks mit dem Seitenverhältnis 33 : 32 in 9, Abb. 31 die eines Quadrats in 24 inkongruente

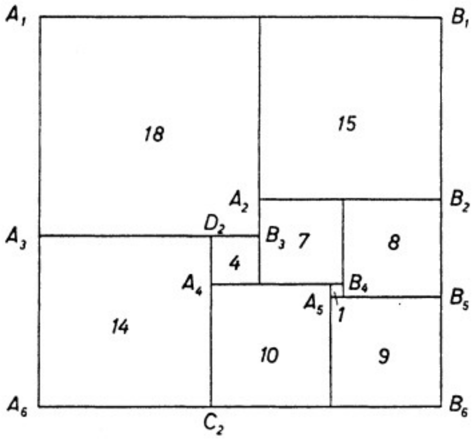

Abb. 30

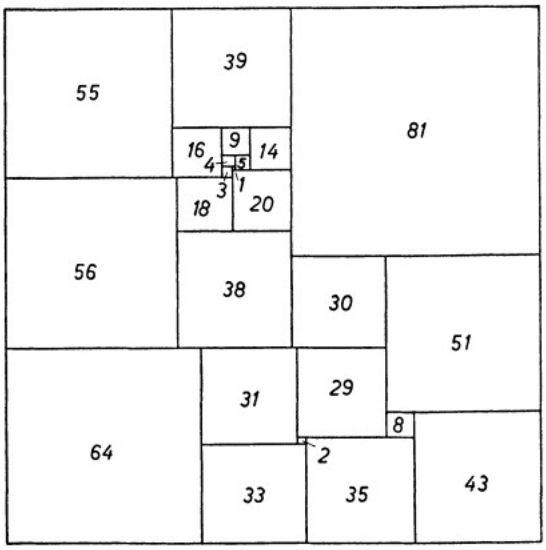

Abb. 31

Quadrate. Da man nicht gut für alle zu diskutierenden Zerlegungen Zeichnungen drucken kann, ist eine einfache und einheitliche Beschreibung einer vorgenommenen Zerlegung wünschenswert. Nach dem Vorschlag von *C. J. Bouwcamp* [VII 8] kann das so geschehen: Rechtecke werden so orientiert, daß die größere Seite waagerecht liegt und in der linken oberen Ecke sich das größte der 4 Eckquadrate befindet (Abb. 30). Wir bezeichnen dann die in einer waagerechten Geraden liegenden aus Quadratseiten zusammengesetzten Strecken der Zerlegungsfigur als *waagerechte Abschnitte*. In Abb. 34 sind das die Strecken $A_\nu B_\nu$ ($\nu = 1, 2, \ldots, 6$). Wir notieren nun die Zerlegung so, daß wir die Seitenlängen aller Teilquadrate aufschreiben. Dabei beginnen wir mit den an dem ersten waagerechten Abschnitt hängenden Quadraten. Sie werden von links nach rechts laufend registriert. Es folgen dann (immer in Klammern zusammengefaßt) die an den folgenden Abschnitten hängenden Quadrate. Unser Code für die Abb. 30 sieht dann so aus:

$$(18{,}15)\quad(7{,}8)\quad(14{,}4)\quad(10{,}1)\quad(9)\,. \tag{1}$$

Die Abb. 31 wäre so zu verschlüsseln:

$$\begin{gathered}(55{,}39{,}81)\quad(16{,}9{,}14)\quad(4{,}5)\quad(3{,}1)\quad(20)\quad(56{,}18)\\(38)\quad(64{,}31{,}29)\quad(8{,}43)\quad(2{,}35)\quad(33)\,.\end{gathered} \tag{2}$$

Der folgende Code (3) gibt dann die Verschlüsselung einer von *R. L. Brooks* [VII 9] stammenden Zerlegung des Quadrats in 38 verschiedene Teilquadrate an:

$$\begin{gathered}(2132{,}1440{,}1348)\quad(256{,}1092)\quad(692{,}584{,}164)\quad(420)\\(108{,}281{,}615)\quad(840{,}758{,}534{,}627{,}173)\quad(454)\quad(199{,}893)\\(120{,}694)\quad(244{,}310)\quad(217{,}984)\quad(82{,}900)\quad(922)\quad(527)\\(1587)\quad(287{,}240)\quad(7{,}1177)\quad(104{,}1130)\quad(1026)\,.\end{gathered} \tag{3}$$

Die Zerlegungen der Abbildungen 30 und 31 weisen einen bemerkenswerten Unterschied auf: In Abb. 31 kann man eine Teilmenge der Teilquadrate zu einem Rechteck zusammenfassen. In Abb. 30 ist das nicht möglich. Das Quadrat der Abb. 31 enthält (links oben) ein Rechteck mit den Seiten $(55 + 39) \cdot (55 + 56)$, das in 13 Teilquadrate zerlegt ist. *Eine Zerlegung in Quadrate, bei der eine Zusammenfassung einer Teilmenge der Quadrate zu einem Rechteck nicht möglich ist, heißt „einfach“.* Die erste von *R. Sprague* angegebene Zerlegung in 55 Teilquadrate war nicht einfach, auch nicht die in [VII 7] angegebene in 26 Teilquadrate. Die kleinste Zahl n, für die bisher eine einfache Zerlegung bekannt ist, ist 38. Die Verschlüsselung dieser Zerlegung ist oben unter (3) aufgeführt. Zerlegungen von Rechtecken in *lauter verschiedene* Quadrate heißen auch „perfekte“ Zerlegungen. Wir haben bisher keine anderen in Betracht gezogen. Es hat aber durchaus Sinn, nach *einfachen imperfekten* Zerlegungen zu fragen. Abb. 32 zeigt eine solche *nicht perfekte aber einfache* Zerlegung eines Quadrats in 13 Teilquadrate.

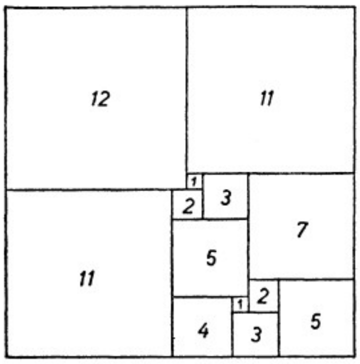

Abb. 32

Es liegen nun folgende Fragen nahe, mit denen wir uns im folgenden beschäftigen wollen:

(6) *Welches ist die kleinste Anzahl n (n', n'') von Quadraten, die für*
(6a) *eine perfekte (einfache, einfache und perfekte) Zerlegung eines Qua-*
(6b) *drats gebraucht werden?*

(6c) *Welches ist die kleinste Anzahl n_R von Teilquadraten bei einer perfekten Zerlegung von Rechtecken?*

(6d) *Für welche Rechtecke ist eine perfekte Zerlegung in Teilquadrate möglich?*

Die Anzahl n (Problem **(6)**) ist noch unbekannt, ebenso n''. Dagegen kennt man n'. Diese Zahl ist 13: Die in Abb. 32 gezeichnete einfache, aber imperfekte Zerlegung kann nicht „verbessert" werden. Auch die Fragen **(6d)** und **(6c)** sind gelöst. Die Antworten auf diese Fragen geben wir am Schluß dieses Artikels.

2. Definition eines Graphen

Eine Reihe von Ergebnissen aus unserem Problemkreis sind offensichtlich durch unermüdliches Probieren gefunden worden. Es liegt aber nahe, nach einem Verfahren zu suchen, durch das man systematisch alle möglichen Rechteckzerlegungen herausfinden kann. Als einen ersten Schritt in dieser Richtung kann man versuchen, die Zahl der zu untersuchenden Möglichkeiten einzuschränken.

Zu diesem Zweck ist eine graphische Darstellung der Zerlegung angebracht, die durch das auf S. 94 beschriebene Code-Verfahren nahegelegt wird. Dort wurden die Quadrate durch Klammern zusammengefaßt, die am gleichen waagerechten Abschnitt hingen. Ersetzt man nun die waagerechten Abschnitte AB durch einen Punkt P_ν (z. B. den Mittelpunkt eines jeden waagerechten Abschnitts), jedes Quadrat durch eine diese Punkte verbindende Strecke, so wird man auf einen Graphen geführt, wie er für die Zerlegung von Abb. 30 in Abb. 33a schematisch

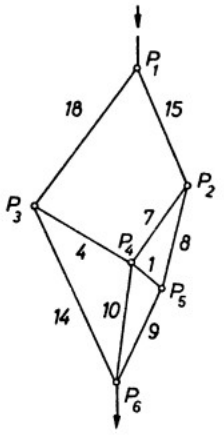

Abb. 33a

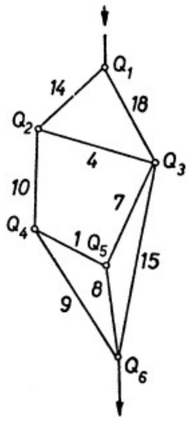

Abb. 33b

dargestellt wird. Die den Strecken angeschriebenen Zahlen sind die Seitenlängen der entsprechenden Quadrate. Die Länge der Strecken des Graphen ist für unsere Zwecke unwesentlich. Es kommt nur auf die topologische Struktur (und die notierten Längen der Quadratseiten) an. Allerdings ist es zweckmäßig, die Punkte des Graphen so zu zeichnen, daß (bei geeigneter Orientierung der Zeichnung) P_ν „höher“ liegt als $P_{\nu+1}$.

Für die spätere systematische Suche nach Zerlegungen des Rechtecks ist es nun wesentlich, daß in einem Graphen mit E Eckpunkten nur der erste und der letzte Punkt (P_1 und P_E) vom Grade 2 sein können. Ein innerer Punkt P_m $(1 < m < E)$ vom Grade 2 wäre ja nur möglich, wenn zwei gleiche Quadrate eine gemeinsame Seite haben. Das ist aber ausdrücklich ausgeschlossen. Deshalb haben wir für die Ecken und Kanten unseres Graphen die Ungleichung

$$3\,E \leqq 2\,K + 2\,. \tag{4}$$

Nach dem Eulerschen Satz für Netze (Satz 1 S. 9)

$$E - K + F = 1$$

folgt daraus für die Zahl F der Flächen („Maschen"):

$$F \geqq \frac{K+1}{3}. \tag{5}$$

Nun kann man in Analogie zu den waagerechten auch die senkrechten Abschnitte einer Rechteckzerlegung einführen. Das sind die aus den in einer Geraden liegenden senkrechten Quadratseiten zusammengesetzten Strecken. In Abb. 30 ist nur der eine senkrechte Abschnitt C_2D_2 als Beispiel eingezeichnet. Jeder senkrechte Abschnitt verbindet zwei Punkte von zwei waagerechten Abschnitten $A_\varrho B_\varrho$ und $A_\sigma B_\sigma$, und die dem senkrechten Abschnitt links und rechts anliegenden Quadrate entsprechen im Graphen solchen Streckenzügen, die P_ϱ und P_σ verbinden. Damit ist eine eindeutige Zuordnung hergestellt zwischen den *inneren* senkrechten Abschnitten und den Flächen des Graphen. Die in (4) und (5) auftretende Zahl F ist zugleich die Zahl der *inneren* senkrechten Abschnitte. Nehmen wir noch die beiden senkrechten Seiten des gegebenen Rechtecks dazu, so haben wir insgesamt $s = F + 2$ senkrechte Abschnitte.

Nun kann man natürlich dem unserer Verabredung entsprechend orientierten Rechteck noch einen zweiten Graphen zuordnen, bei dem die Punkte den senkrechten Abschnitten der Zerlegung entsprechen. Die Strecken dieses dem ersten *dualen* Graphen entsprechen wieder den Teilquadraten, die Maschen des Netzes aber hier den waagerechten Abschnitten. Abb. 33b zeigt den zu Abb. 33a dualen Graphen.

Bezeichnen wir die Ecken, Kanten und Flächen des zweiten (zum ersten dualen) Graphen mit E_1, F_1, K_1, so haben wir analog zu (4) und (5):

$$3\,E_1 \leqq 2\,K + 2\,, \tag{4'}$$

$$F_1 \geqq \frac{K+1}{3} \tag{5'}$$

Dabei ist schon berücksichtigt, daß die Zahl der Kanten in beiden Graphen gleich ist: $K = K_1$. Weiter ist die Zahl der Ecken E_1 im zweiten Graphen gleich der Zahl der senkrechten Abschnitte, also

$$E_1 = F + 2 = s\,. \tag{6}$$

Analog haben wir für E:

$$E = F_1 + 2 = w\,, \tag{7}$$

wobei w die Zahl der waagerechten Abschnitte in der gegebenen Zerlegung ist.

Aus (4), (4′), (5), (5′), (6) und (7) gewinnt man schließlich für die Maschenzahl der beiden Graphen die Abschätzung

$$\frac{K+1}{3} \leqq \begin{Bmatrix} F \\ F_1 \end{Bmatrix} \leqq \frac{2K-4}{3} \tag{8}$$

Diese Ungleichung gibt uns zusammen mit der Eulerschen Gleichung (Satz 1) die Möglichkeit, für die Zerlegung von Rechtecken in eine gegebene Zahl $K = K_1$ von Quadraten die möglichen Zahlen für die Flächen („Maschen") und Ecken des Graphen zu bestimmen. Da nur ganzzahlige Werte in Frage kommen, haben wir für $K = 5, 6, \ldots, 13, 14$ die in der folgenden Tabelle zusammengestellten Möglichkeiten zu berücksichtigen:

K	F	E	K	F	E
5	2	4	10	4 ; 5	7 ; 6
6	—	—	11	4 ; 5 ; 6	8 ; 7 ; 6
7	3	5	12	5 ; 6	8 ; 7
8	3 ; 4	6 ; 5	13	5 ; 6 ; 7	9 ; 8 ; 7
9	4	6	14	5 ; 6 ; 7 ; 8	10 ; 9 ; 8 ; 7

3. Berechnung der Zerlegungen

Wir müssen nun prüfen, welche der nach der Tabelle möglichen Graphen wirklich bei einer Rechteckzerlegung auftreten. Dazu beachten wir die Gesetze, denen die Seitenzahlen der Quadrate unterliegen. Diese Zahlen werden ja den Strecken des Graphen immer beigeschrieben (Abb. 33a und 33b). Die einem waagerechten Abschnitt „aufsitzenden" Quadrate liefern dieselbe Summe der Seitenlängen wie die einem solchen Abschnitt „angehängten". Die entsprechende Aussage gilt für die einem senkrechten Abschnitt rechts und links angelagerten Quadrate. Da die waagerechten Abschnitte den Punkten und die (inneren) senkrechten Abschnitte den Maschen des Graphen entsprechen, ergeben sich daraus Gesetze für die Seitenlängen der Quadrate.

Um diese Gesetze bequem formulieren zu können, wollen wir verabreden, daß einer Strecke $P_\varrho P_\sigma$ eines Graphen eine *negative* Seitenzahl $x_{\varrho\sigma}$ zuzuordnen ist, wenn $\varrho > \sigma$ ist, eine positive, wenn $\varrho < \sigma$. Also: Der Strecke P_3P_4 in Abb. 37a haben wir die Zahl 4 angeschrieben; für P_4P_3 wäre danach -4 die zugeordnete Seitenzahl. Wir wollen weiter sagen, daß zu einem inneren Punkt P_ϱ eines Graphen die Seitenzahlen gehören, die den sämtlichen in P_ϱ zusammenlaufenden Strecken $P_\varrho P_{\sigma_\nu}$ des

Graphen zugeordnet sind. Einer Masche $P_{n_1} P_{n_2} \ldots P_{n_\tau} P_{n_1}$ des Graphen ordnen wir entsprechend die sämtlichen den Strecken $P_{n_\nu} P_{n_\nu + 1}$ angeschriebenen Zahlen $x_{n_\nu\, n_\nu + 1}$ zu, immer unter Beachtung der Vorzeichenregel. Dann folgt sofort aus der Definition der Seitenzahlen und der Vorzeichenregel:

Die algebraische Summe der einem inneren Punkt oder einer Masche zugeordneten Seitenzahlen ist gleich 0:

$$\sum_{\nu=1}^{g} x_{\varrho\sigma_\nu} = 0\,, \qquad \sum_{\nu=1}^{h} x_{n_\nu n_\nu + 1} = 0, \quad n_{h+1} = n_1 . \tag{9}$$

Dabei ist g der Grad des Punktes, h die Eckenzahl der Masche.

Die Suche nach Zerlegungen von Rechtecken (oder womöglich von Quadraten) kann nun so erfolgen: Man zeichnet die nach der Tabelle von S. 98 topologisch möglichen Graphen und prüft, ob für sie eine brauchbare Lösung der diophantischen Gleichungen (9) existiert. Nicht „brauchbar“ sind z. B. solche Lösungen, bei denen ein $x_{ik} = 0$ ist. Auch Lösungen mit $x_{ik} = x_{i\,k+1}$ sind auszuschließen, weil sie ja zu einer Zerlegung mit zwei gleichen benachbarten Quadraten führen würde. Es zeigt sich dann sehr rasch, daß die meisten der topologisch möglichen Graphen keine brauchbaren Zerlegungen liefern.

Nehmen wir als ein Beispiel den ersten Fall unserer Tabelle $K = 5$ (Abb. 34). Daß diese Figur nicht der Graph einer Rechteckzerlegung sein kann, sieht man leicht unmittelbar ein. Wir wollen aber doch die Begründung aus unserer Aussage über die algebraische Summe der Seitenzahlen herleiten, um zu zeigen, wie man auch in komplizierteren Fällen zum Ziel kommen kann.

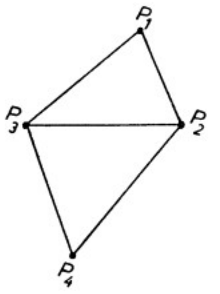

Abb. 34

Es sei (entsprechend unserer Verabredung über die Orientierung der Rechtecke) $x_{13} > x_{12}$ angenommen. Dann müßte sein

$$x_{21} + x_{23} + x_{24} = 0\,, \quad x_{31} + x_{32} + x_{34} = 0\,, \qquad (10)$$
$$x_{13} + x_{34} + x_{42} + x_{21} = 0\,.$$

Daraus folgt nun

$$x_{12} = - x_{21} = x_{23} + x_{24} > x_{24}\,, \quad x_{13} = x_{34} - x_{23} < x_{34}\,,$$

wegen $x_{13} > x_{12}$ also

$$x_{12} + x_{24} < 2\,x_{12} < 2\,x_{13}\,x_{13} + x_{34}$$

im Gegensatz zur 3. Gleichung von (10).

In ähnlicher Weise kann man nun die weiteren Graphen untersuchen. Die systematische Prüfung der Graphen für $K = 6, 7, \ldots$ ist eine nicht besonders schwierige, aber doch recht langwierige Arbeit. *C. J. Bouwkamp* hat diese Untersuchung für $K = 5, 6, \ldots, 13$ vollständig durchgeführt. Es zeigt sich dabei, daß erst bei $K = 9$ brauchbare Lösungen möglich sind. Es treten also bei Zerlegungen von Rechtecken in Quadrate verschiedener Seitenlänge stets mindestens 9 solcher Quadrate auf. Die Abb. 30 zeigt ein Beispiel einer solchen Zerlegung mit $K = 9$.

Mit wachsender Zahl K ergeben sich immer mehr Möglichkeiten. Für $K = 13$ nimmt in der Arbeit von *Bouwkamp* allein die Aufzählung der möglichen Lösungen nach seinem Code-Verfahren mehrere Druckseiten ein. Es gibt ([VII 8] S. 1299) 585 Zerlegungen von Rechtecken in 9, 10, 11, 12 oder 13 inkongruente Quadrate. Unter diesen Rechtecken befindet sich *kein einziges Quadrat*. Dagegen befindet sich unter den 72 für $9 \leqq K \leqq 13$ möglichen imperfekten (aber nicht trivialen) Zerlegungen auch die einfachste „einfache“ (aber nicht perfekte) Zerlegung in 13 Quadrate, die in Abb. 32 dargestellt ist.

Man kann perfekte Quadratzerlegungen aus einfacheren Zerlegungen von Rechtecken aufbauen. Das Verfahren ist schon von *R. Sprague* bei der Konstruktion seines ersten Beispiels benutzt worden [VII 4]. Der Grundgedanke ist folgender: Man zerlege ein Quadrat (Abb. 35) in zwei

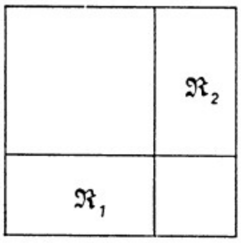

Abb. 35

ungleiche Quadrate $\mathfrak{Q}_1$ und $\mathfrak{Q}_2$ und zwei Rechtecke $\mathfrak{R}_1$ und $\mathfrak{R}_2$. Wenn es gelingt, die Rechtecke $\mathfrak{R}_1$ und $\mathfrak{R}_2$ so in lauter verschiedene Quadrate zu zerlegen, daß keins der Quadrate von $\mathfrak{R}_1$ einem Quadrat von $\mathfrak{R}_2$ gleich ist, wenn ferner alle Teilquadrate von $\mathfrak{R}_\nu$ $(\nu = 1, 2)$ von $\mathfrak{Q}_1$ und $\mathfrak{Q}_2$ verschieden sind, so ist damit eine perfekte (aber nicht einfache!) Zerlegung des ganzen Quadrats erreicht. Durch zwei verschiedene Zerlegungen des Rechtecks mit dem Seitenverhältnis 13 : 16 kam *R. Sprague* auf diese Weise zu einer Zerlegung des Quadrats in 55 Teilquadrate. Mit den Ergebnissen von *Bouwkamp* kann man nach diesem Verfahren auch eine Zerlegung in 28 inkongruente Teilquadrate gewinnen. Es existieren nämlich zwei verschiedene Zerlegungen von Rechtecken mit dem Seitenverhältnis 593 : 422 in je 13 verschiedene Teilquadrate:

$$\begin{array}{ll}
\mathfrak{R}_1: & (222,164,207)\quad (80,41,43)\quad (39,2)\quad (37,215)\quad (200,22)\\
 & \qquad\qquad\qquad\qquad\quad (78)\\
\mathfrak{R}_2: & (247,154,192)\quad (116,38)\quad (230)\quad (175,72)\\
 & \qquad\quad (49,67)\quad (103,18)\quad (85)
\end{array} \tag{11}$$

Es gibt noch einige andere Methoden ([VII 8] S. 74), um aus Rechteckzerlegungen eine perfekte Aufteilung des Quadrats zu gewinnen. Alle so gefundenen Zerlegungen sind nicht einfach. In der Arbeit [VII 7] ist ein Verfahren angegeben, um einfache und perfekte Quadratzerlegungen herauszufinden. Das Verfahren ist freilich recht kompliziert und läßt die Frage offen, welches die kleinste Zahl n'' von Teilquadraten ist, die bei einer einfachen und perfekten Zerlegung auftreten.

4. Die elektrotechnische Deutung

Auf S. 93 wurde schon erwähnt, daß die Zerlegung von Rechtecken in inkongruente Quadrate für ein elektrotechnisches Problem von Bedeutung ist. Wir haben die technische Fragestellung aus der mathematischen Überlegung herausgehalten, wollen aber doch die interessante Deutung unseres Problems nicht übergehen.

An einer homogenen Metallplatte von rechteckiger Gestalt seien an zwei gegenüberliegenden (horizontal orientierten) Seiten Elektroden angebracht. Von einer Elektrode zur andern möge ein konstanter elektrischer Strom fließen. Die Einheit der Stromstärke und der Spannung denken wir uns so gewählt, daß die horizontale Rechteckseite gerade gleich der Stromstärke, die vertikale gleich der Spannung ist. Die Stromlinien verlaufen dann vertikal, die Potentiallinien waagerecht. Es ändert deshalb nichts am Verlauf des Stromes, wenn wir senkrechte Schnitte in der Platte anbringen und in einer waagerechten Strecke einen Leiter vom Widerstand 0 in die Platte einfügen.

Nehmen wir nun an, daß unser Rechteck in lauter inkongruente Quadrate zerlegt sei. Wir schneiden die Platte nun an den senkrechten Seiten der Teilquadrate auf und prägen der Platte an den waagerechten Seiten Elektroden von unendlich hoher Leitfähigkeit ein. Dadurch ändert sich nichts am Strom. Das Verhältnis $V : J$ (Spannung : Stromstärke) ist nach unserer Wahl der Einheiten gerade gleich 1 in *jedem* der Teilquadrate. Jedes Teilquadrat repräsentiert also einen Widerstand von der Größe 1. Man kann nun noch die sämtlichen waagerechten Elektroden in einem Punkt zusammenziehen und die Teilquadrate zu einem Draht zusammenrollen. Dann entsteht aus der rechteckigen Platte ein elektrisches Netzwerk, wie wir es schematisch schon in den Graphen dargestellt haben. Jede nicht triviale Zerlegung eines Rechtecks entspricht dann einem Netzwerk aus Einheitswiderständen, in dem keine einfachen Parallel- oder Reihenschaltungen vorkommen. Ist das zu zerlegenden Rechteck speziell ein Quadrat, so gewinnt man auf diese Weise ein elektrisches Netzwerk aus Widerständen von x Ohm, das als ganzes ebenfalls den Widerstand x Ohm hat. Die Regeln (9) können dann als Kirchhoffsche Gesetze für Stromverzweigungen gedeutet werden.

Für die elektrotechnische Fragestellung ist es unwichtig, ob die auftretenden Teilquadrate alle verschiedene Seitenlängen haben oder nicht. Es kommt nur darauf an, daß die Zerlegung nicht trivial ist, als keine Quadrate mit *gemeinsamer* Seite hat. Der zu der Zerlegung in Abb. 32 gehörende Graph stellt danach das einfachste Netz aus Widerständen von der Größe 1 dar, das selbst den Widerstand 1 hat.

5. Zerlegung von Rechtecken mit kommensurablen Seiten

Schon in der ersten Arbeit über die Zerlegung von Rechtecken hat *Dehn* [VII 1] gezeigt, daß nur für Rechtecke mit kommensurablen Seiten die Zerlegung in lauter ungleiche Quadrate möglich ist. Es dauerte recht lange, bis der Beweis gelang, daß jedes Rechteck mit kommensurablen Seiten auch tatsächlich in der vorgeschriebenen Weise zerlegbar ist [VII 5].

Wir wollen uns darauf beschränken, den Grundgedanken eines solchen Zerlegungsverfahrens zu skizzieren. Man zeigt zuerst, daß das Einheitsquadrat auf unendlich viele verschiedene Arten in inkongruente Quadrate zerlegt werden kann. Dabei kommt es darauf an, daß in einer solchen Folge

$$\mathfrak{Z}_1, \mathfrak{Z}_2, \mathfrak{Z}_3, \ldots \tag{12}$$

von Zerlegungen jedes Quadrat von $\mathfrak{Z}_\nu$ von jedem von $\mathfrak{Z}_\mu$ verschieden ist. Solche Folge (12) von Zerlegungen kann man z. B. erhalten, indem

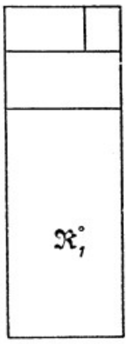

Abb. 36

man von einem Rechteck $\mathfrak{R}$ ausgeht, das auf zwei verschiedene Weisen in inkongruente Quadrate zerlegt werden kann. Diese mit der Quadratzerlegung versehenen Rechtecke wollen wir $\mathfrak{R}_1{}^0$ und $\mathfrak{R}_2{}^0$ nennen.[1])

Man erhält nun eine Folge von Rechteckpaaren $\mathfrak{R}_1^\nu$ und $\mathfrak{R}_2^\nu$, indem man an eine der kleineren Seiten von $\mathfrak{R}_1^{\nu-1}$ bzw. $\mathfrak{R}_2^{\nu-1}$ ein ähnliches Bild dieser Rechtecke anlegt (Abb. 36).[2]).

Dabei soll in jedes der Rechtecke die Zerlegung in Teilquadrate eingetragen sein. Aus den Rechteckpaaren werden nun Quadrate $\mathfrak{Q}_\nu$ $(\nu = 1, 2, 3, \ldots)$ nach dem Muster von Abb. 35 aufgebaut, und durch eine geeignete Ähnlichkeitstransformation gewinnt man daraus eine Folge (12) von Zerlegungen des Einheitsquadrats. Es muß nur bewiesen werden, daß die in den verschiedenen Zerlegungen $\mathfrak{Z}_\nu$ auftretenden Teilquadrate voneinander verschieden sind[3]).

Auf diese Weise hat man eine Möglichkeit gewonnen, jedes Rechteck mit kommensurablen Seiten in lauter verschiedene Quadrate zu zerlegen. In der Tat: Man braucht ja nur das gegebene Rechteck durch äquidistante Parallelen in lauter gleiche Quadrate zu zerlegen und auf jedes der so entstehenden Quadrate q_ϱ $(\varrho = 1, 2, \ldots, m)$ die Zerlegung $\mathfrak{Z}_\varrho$ $(\varrho = 1, 2, 3, \ldots, m)$ anzuwenden. Auf diese Weise hat man eine perfekte, aber natürlich nicht einfache Zerlegung des gegebenen Rechtecks gewonnen.

Von den auf S. 95 formulierten Fragen **(6)**, **(6a)**, **(6b)**, **(6c)** und **(6d)** sind **(6)** und **(6b)** noch offen, die anderen sind durch die Ausführungen der Abschnitte VII 3 und VII 5 beantwortet. Für die Zahlen n, n' und n'' (S. 95) haben wir die Aussagen:

$$13 < n \leqq 24\,, \quad n' = 13\,, \quad 13 < n'' \leqq 38\,.$$

---

[1]) *R. Sprague* hat in seiner Arbeit [VII 5] ein Rechteckpaar benutzt, das ein in 55 Teilquadrate zerlegtes Quadrat liefert. Bei Benutzung des Beispiels (11) kommt man auf ein Quadrat, das in 28 inkongruente Teilquadrate zerlegt ist.

[2]) Auch andere Verfahren zum Aufbau perfekter Quadrate aus zerlegten Rechtecken können zum Aufbau einer Folge (12) benutzt werden, z. B. die von *Bouwkamp* ([VII 8] S. 74) angegebenen.

[3]) Wir übergeben diesen Beweis. Siehe dazu [VII 5].

(%"

- Introductory Calculus..(1972).pdf
- Space-Time structure - Algebra and Geometry...(2007).pdf

1820 - The Arithmometer (Calculara)
Patenteada por Thomas de Colmar
Fabricada entre 1851-1915

Iridium melting point of
4,435°F (2,466°C).

Books:

Books-019-2021/Herbert Meschkowski/
- Ungelöste und unlösbare Probleme der Geometrie...(1960).pdf
Pág. 92(99) VII.
Die Zerlegung von Rechtecken in inkongruente Quadrate
(A decomposição de retângulos em incongruentes quadrados).

- Geometry of time-spaces non-commutative algebraic geometry...pdf

Life of Pi

In the early nineteenth century, news of boy wonder George Parker Bidder, the son of a Devonshire stonemason, reached the ears of Queen Charlotte. She had a question for him:

> ‘From the Land’s-end, Cornwall, to Farret’s-head, in Scotland, is found by measurement to be 838 miles; how long would a snail be creeping that distance, at the rate of 8 feet per day?’

The exchange and the answer – 553,080 days – is mentioned in a popular book of the time, *A short Account of George Bidder, the celebrated Mental Calculator: with a Variety of the most difficult Questions, Proposed to him at the principal Towns in the Kingdom, and his surprising rapid Answers!* The pages list the child’s greatest calculations, including such classics as ‘What is the square root of 119,550,669,121?’ (345,761, answered in half a minute) and ‘How many pounds weight of sugar are there in 232 hogsheads, each weighing 12cwt. 1qr. 22lbs?’ (323,408lbs, also answered in half a minute.)

Arabic numerals made doing sums easier for everyone, but an unexpected consequence was the discovery that certain people were blessed with truly astonishing arithmetical skills. Often, these prodigies excelled in no other way than their facility with numbers. One of the earliest-known examples was a Derbyshire farmhand, Jedediah Buxton, who amazed locals with his abilities in multiplication despite being barely able to read. He could, for example, calculate the value of a farthing when doubled 140 times. (The answer is 39 digits long, with 2 shillings 8 pence left over.) In 1754, curiosity about Buxton’s talent led to him being invited to visit London, where he was examined by members of the Royal Society. He seems to have had some of the symptoms of high-functioning autism, for when he was taken to see Shakespeare’s *Richard III* he was left nonplussed by the experience, although he notified his hosts that the actors had taken 5202 steps and spoken 14,445 words.

In the nineteenth century ‘lightning calculators’ were international stage stars. Some showed aptitude at an extraordinarily young age. Zerah Colburn, from Vermont, was five when he gave his first public demonstration and eight when he sailed to England with dreams of big-time success. (Colburn was born with hexadactyly, but it is not known if his extra fingers gave him an advantage when learning to count.) A contemporary of Colburn's was the Devonshire lad George Parker Bidder. The two prodigies crossed paths in 1818, when Colburn was 14 and Bidder 12, and the encounter, in a London pub, inevitably led to a maths duel.

Colburn was asked how long it would take a balloon to circumnavigate the globe if the balloon were travelling at 3878 feet per minute and the world were 24,912 miles around. It was a suitably international question for the unofficial title of smartest alec on Earth. But after deliberating for nine minutes, Colburn failed to give an answer. A London newspaper gushed that his opponent, on the other hand, took only two minutes before giving the correct reply, '23 days, 13 hours and 18 minutes, [which] was received with marks of great applause. Many other questions were proposed to the American boy, all of which he refused answering; while young Bidder replied to all.' In his charming autobiography, *A memoir of Zerah Colburn, written by himself*, the American gives a different version of the contest. '[Bidder] displayed great strength and power of mind in the higher branches of arithmetic,' he said, before adding dismissively, 'but he was unable to extract the roots, and find the factors of numbers.' The mpionship was left undecided. Edinburgh University subsequently offered to take over Bidder's education. He went on to become an important engineer, at first in railroads and eventually supervising construction of London's Victoria Docks. Colburn, on the other hand, returned to America, became a preacher and died aged 35.

The ability to calculate rapidly has no great correlation with mathematical insight or creativity. Only a few great mathematicians have demonstrated lightning-calculator skills, and many mathematicians have surprisingly poor arithmetic. Alexander Craig Aitken was a well-known lightning calculator in the first half of the twentieth century, unusual in that he was also a professor of mathematics at Edinburgh University. In 1954 Aitken gave a lecture to the Society of Engineers in London, in which he explained some of the methods in his repertoire, such as algebraic shortcuts and – crucially – the importance of memory. To prove his point he rattled off the decimal expansion for $\frac{1}{97}$, which repeats only after 96 digits.

Aitken ended his talk with the rueful comment that when he acquired his first desk calculator his abilities began to deteriorate. ‘Mental calculators may, like the Tasmanian or the Moriori, be doomed to extinction,’ he predicted. ‘Therefore…you may be able to feel an almost anthropological interest in surveying a curious specimen, and some of my auditors here may be able to say in the year 2000 ce, “Yes, I knew one such.”’

This was one calculation, however, that Aitken got wrong.

‘Neurons! On the ready! Go!’

With an impatient snap and swoosh, the contestants in the multiplication round at the Mental Calculation World Cup turned over their papers. The room at Leipzig University was silent as the 17 men and two women contemplated the first question: 29,513,736× 92,842,033.

Arithmetic is back in vogue. Thirty years after the first cheap electronic calculators precipitated a widespread demise in mental calculation skills, a backlash is under way. Newspapers offer up daily maths brainteasers, popular computer games with arithmetic puzzles sharpen our minds and – at the high end – lightning calculators compete in regular international tournaments. The Mental Calculation World Cup was founded in 2004 by German computer scientist Ralf Laue, and takes place every two years. It was the inevitable culmination of Laue’s two hobbies: mental arithmetic and collecting unusual records (such as the Most Grapes Thrown over a Distance of 15ft and Caught in the Mouth in One Minute, which is 55 of them). The internet helped, enabling him to meet kindred spirits – mental arithmeticians are not, in general, extroverts. The global community of human calculators, or ‘mathletes’, was well represented in Leipzig, with contestants from countries as diverse as Peru, Iran, Algeria and Australia.

How do you measure calculating skills? Laue adopted the categories already chosen by *Guinness World Records* – the multiplication of two eight-digit numbers, the addition of ten ten-digit numbers, extracting the square root of six-figure numbers to eight significant figures, and finding the day of the week of any date between 1600 and 2100. The latter is known as a calendar calculation, and is a flashback to the golden age of lightning calculation, when performers would ask a member of the audience their birth date and then instantly name the day of the week it fell on.

Regulation, and a spirit of competitiveness, have come at the expense of theatrics. The youngest contestant at the World Cup, an 11-year-old boy from India, performed the ‘air abacus’ – his hands were jerking wildly around rearranging imaginary beads, while all the others were quiet and still, occasionally jotting down their answers. (The rules say that only the final answer can be written down.) After 8 minutes and 25 seconds, Alberto Coto of Spain stuck his hand up like an excited schoolboy. The 38-year-old had completed ten multiplications of two eight-digit numbers in that time, smashing the world record. It was evidently an awesome accomplishment, yet watching him was as compelling as invigilating an exam.

Conspicuously absent from proceedings in Leipzig, however, was perhaps the world’s most famous mathlete, the French student Alexis Lemaire, who prefers another yardstick to measure computational power. In 2007 Lemaire, aged 27, made international headlines when, at the Science Museum in London, he took just 70.2 seconds to calculate the thirteenth root of:

**85,877,066,894,718,045,602,549,144,850,158,599,202
748,960,878,023,151,390,314,284,284,465,842,798,37
242,826,571,823,153,045,030,300,932,591,615,405,92
640,895,967,991,430,381,763,526,613,357,308,674,59
521,841,103,664,923,661,204,223**

Lemaire’s achievement was undoubtedly the more spectacular. The number has 200 digits, which can barely be pronounced in 70.2 seconds. But did his feat mean that, as he claims, he is the greatest lightning calculator of all time? This is a matter of deep controversy in the calculation milieu, mirroring the battle almost 200 years ago between Zerah Colburn and George Bidder, both exceptional at their own type of sum.

The term ‘thirteenth root of *a*’ refers to the number that when multiplied by itself 13 times equals *a*. Only a fixed amount of numbers when multiplied by themselves 13 times equal a 200-digit number. (It is a large fixed amount. The answer is limited to about 400 trillion possibilities, all 16 digits long and beginning with 2.) Because 13 is prime and considered unlucky, Lemaire’s calculation was vested with an extra aura of mystery. In fact, 13 brings with it some advantages. For instance when 2 is multiplied by itself 13 times, the answer ends in 2. When 3 is multiplied by itself 13 times the answer ends in 3. The same is true for 4, 5, 6, 7, 8 and 9. In other words, the last digit of the thirteenth root of a number is the same as the last digit of the original number. We get this number for free, without having to do any calculation.

Lemaire has worked out algorithms, which he has not divulged, to calculate the other 14 digits in the final answer. Purists, possibly unfairly, say that his skill is less a feat of calculation and more one of memorizing huge strings of numbers. And they point out that Lemaire cannot find the thirteenth root of *any* 200-digit number given to him. At the Science Museum in London he was presented with several hundred numbers and allowed to select the one that he would do the calculation for.

Still, Lemaire's performance was more in keeping with the tradition of the stage calculators of old. Audiences want to feel the 'wow', rather than understand the process. By contrast, at the Mental Calculation World Cup, Coto had no choice about the problem to be solved and used no hidden techniques when he multiplied 29,513,736×92,842,033. He simply used his 1-to 9-times tables. The fastest way to multiply eight digits by eight digits is using the Vedic sutra *Vertically and Cross-Wise*, which breaks the sum into 64 multiplications of single-digit figures. He managed to get the right answer in an average of less than 51 seconds. Knowing what he was doing made it less dazzling, even though it was obviously still a formidable feat.

As I talked with competitors in Leipzig, I discovered that many of them had fallen in love with speed arithmetic thanks to Wim Klein, a Dutch lightning calculator who was a celebrity in the 1970s. Klein was already a veteran of circuses and music halls when, in 1958, he was given a job by Europe's top physics institute – the European Organization for Nuclear Research (CERN) in Geneva, providing calculations for the physicists. He was probably the last human calculator to have been employed as one. As computers developed, his skills became redundant, and in retirement he returned to showbiz, appearing frequently on TV. (Klein, in fact, was one of the first to promote thirteenth-root calculations.)

A century before Klein another human calculator, Johann Zacharias Dase, was also employed by the scientific establishment to do their sums for them. Dase was born in Hamburg and started performing as a lightning calculator in his teens, when he was taken under the wing of two eminent mathematicians. In the days before electronic or mechanical calculators, scientists relied on logarithm tables to do complicated multiplications and divisions. As I will explain in more detail later, every number has its own logarithm, which can be calculated using a laborious procedure of adding fractions. Dase calculated the natural logarithms of the first 1,005,000 numbers, each to seven decimal places. It took him three years, and he said he enjoyed the task. Then, on the recommendation of the mathematician Carl Friedrich Gauss, Dase embarked on another enormous project: compiling a table of factors of all the numbers between 7,000,000 and 10,000,000. This means he looked at every number in the range and calculated its factors, which are the whole numbers that divide that number. For example, 7,877,433 has only two factors: 3 and 2,625,811. When Dase died, aged 37, he had completed a substantial part of it.

Yet it was another calculation for which Dase is best remembered. When still a teenager he calculated pi to 200 places, a record for the time.

Circles are everywhere in the natural world – you see them in the full moon, in the eyes of humans and animals and in the cross-section of an egg. Tie a dog to a post and the path it patrols when the lead is taut is a circle. The circle is the simplest two-dimensional geometrical shape. An Egyptian farmer counting how much of a crop to plant in a round field, or a Roman mechanic measuring the length of wood for a wheel would have needed to make calculations involving circles.

The ancient civilizations realized that the ratio of a circle's circumference to its diameter was always the same no matter how big or small you made the circle. (The circumference is the distance around a circle, and the diameter is the distance across it.) The ratio is known as pi, or p, and it works out as just over three. So, if you take the diameter of a circle and curve it around the circumference, you will find that it fits just over three times.

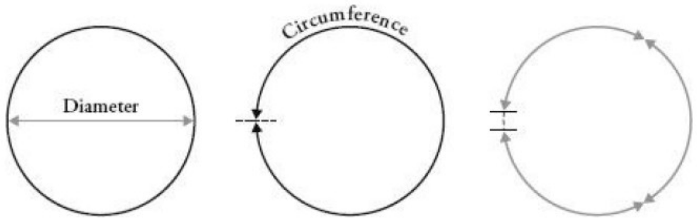

Even though pi is a simple ratio between the basic properties of a circle, the task of finding its exact value has proved to be far from simple. This elusiveness has made pi an object of fascination for thousands of years. It is the only number that is both the name of a song by Kate Bush and a fragrance by Givenchy, whose PR department sent me the following text:

π– PI
BEYOND INFINITY

Four thousand years have passed and the mystery remains.

Although every schoolchild studies, the familiar symbol still manages to hide an abyss of great complexity.

Why choose to symbolise the eternal masculine?

It’s a matter of signs and directions. If is the story of the long struggle to achieve the unattainable, it is also a portrait of the fabled conqueror in search of knowledge.

Pi speaks of men, of all men, of their scientific genius, their taste for adventure, their willingness to act, and of their passions to the extreme.

The earliest approximations for pi came from the Babylonians, who used a value of 3 $\frac{1}{8}$, and the Egyptians, who used 4($\frac{8}{9}$, which translate, respectively, into decimals as 3.125 and 3.160. A line in the Bible reveals a situation in which pi is taken as 3: ‘Also he made a molten sea of ten cubits from brim to brim, round in compass, and five cubits the height thereof; and a line of thirty cubits did compass it round about’ (I Kings 7:23).

If the shape of the sea is a circle with a circumference of 30 cubits and a diameter of 10, then pi is $\frac{30}{10}$, or 3. Many excuses have been given for the Bible’s inaccurate value, such as the claim that the sea was in a circular vessel with a thick rim. In this case the quoted 10-cubit diameter covers the sea and the rim (making the true diameter of the sea a little less than 10 cubits), while the circumference of the sea is taken as the inside of the rim. A mystical explanation is much more enticing: due to the peculiarities of Hebrew pronunciation and spelling, the word ‘line’, or *qwh*, is pronounced *qw*. Totting up the numerological values of the letters gives 111 for *qwh* and 106 for *qw*. Multiplying three by $\frac{111}{106}$ gives 3.1415, which is pi correct to five significant figures.

The first genius whose extreme passion for discovery about pi does justice to the aspirations of Givenchy’s aftershave is equally the man who took the most famous bath in the history of science. Archimedes slipped into the tub and noticed that the volume of water he displaced was equal to the volume of his own body under the water. He instantly realized that he could therefore find the volume of any object by submersing it, in particular the crown of the King of Syracuse, and so would be able to ascertain if that piece of royal bling was made of pure gold or not by working out its density. (It wasn’t.) As a result, he ran naked into the streets shouting ‘Eureka! [I have found it!]’, thus displaying – for the citizens of Syracuse, at least – the eternal masculine. Archimedes loved to grapple with problems in the real world, unlike Euclid, who dealt uniquely in abstractions. His many inventions were said to have included a giant catapult and a system of huge mirrors that reflected the sun’s rays with such intensity that they set Roman ships ablaze during the Siege of Syracuse. He was also the first person to come up with an apparatus to capture pi.

To do so he first drew a circle, and then he constructed two hexagons – one that he fitted inside the circle, and one that he put outside it, as in the diagrams below. This already tells us that pi must be somewhere between 3 and 3.46, which is determined by calculating the perimeters of the hexagons. If we let the diameter of the circle be 1, then the perimeter of the inner hexagon is 3, which is less than the circumference of the circle, which is pi, which is less than the perimeter of the outer one, which is $3\sqrt{2}$, or 3.46 to two decimal places. (The way that Archimedes calculated this value was by using a method that was essentially a fiddly precursor to trigonometry, and which is too complicated to go into here.)

So, $3 < \text{pi} < 3.46$.

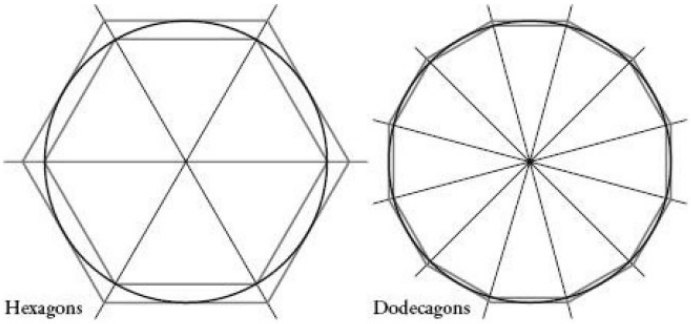

Now if you were to repeat the calculation using two regular polygons with more than six sides, you would get a narrower bound for pi. This is because the more sides that the polygons have, the closer their perimeters are to the circumference, as we can see in the diagram above that uses a 12-sided polygon. The polygons act like walls closing in on pi, squeezing it from above and from below, between narrower and narrower limits. Archimedes started with a hexagon and eventually constructed polygons of 96 sides, allowing him to calculate pi as follows: $3\frac{10}{71} < \text{pi} < 3\frac{1}{7}$

This translates into $3.14084 < \text{pi} < 3.14289$, an accuracy of two decimal places.

But pi hunters weren’t about to stop there. In order to get closer to the number’s true value, all that was required was to create polygons with more sides. Liu Hui, in third-century China, employed a similar method, using the area of a polygon of 3072 sides to pin pi to five decimal places: 3.14159. Two centuries later Tsu ChungChih and his son Tsu Keng-Chih went one digit further, to 3.141592, with a polygon of 12,288 sides.

The Greeks and the Chinese were hampered by cumbersome notation. When mathematicians were eventually able to use Arabic numerals, the record tumbled. In 1596 the Dutch fencing master Ludolph van Ceulen used a souped-up polygon of 60×2^{29} sides to find pi to 20 decimal places. The pamphlet on which he printed his result ended with the words: ‘whoever wants to, can come closer’, and no one felt the urge as strongly as he did. He went on to calculate pi to 32 and then 35 decimal places – which were engraved on his tombstone. In Germany, *die Ludolphsche Zahl*, Ludolph’s number, is still understood as a term for pi.

For 2000 years the only way to pinpoint pi with accuracy was by using polygons. But, in the seventeenth century, Gottfried Leibniz and John Gregory ushered in a new age of pi appreciation with the formula:

$$\frac{\text{Pi}}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} \cdots$$

In other words, a quarter of pi is equal to one minus a third plus a fifth minus a seventh plus a ninth and so on, alternating the addition and subtraction of unit fractions of the odd numbers as they head to infinity. Before this point scientists were only aware of the scatter-gun randomness of pi’s decimal expansion. Yet here was one of the most elegant and uncomplicated equations in maths. Pi, the poster boy of disorder, it turned out, had some kind of order in his DNA.

Leibniz had devised the formula using ‘the calculus’, a powerful type of mathematics he had discovered, in which a new understanding of infinitesimal amounts was used to calculate areas, curves and gradients. Isaac Newton had also come up with calculus, independently, and the men spent a good deal of time bickering about who had got there first. (For years, Newton was considered to have won the argument, based on the dates of his unpublished manuscripts, but it now appears that a version of calculus was actually first invented in the fourteenth century by the Indian mathematician Madhava.)

The formula that Leibniz found for pi is what is known as an *infinite series*, a sum that goes on and on for ever, and it provides a way to calculate pi. First, we need to multiply both sides of the formula by 4 to get:

$$\text{Pi} = 4 - \frac{4}{3} + \frac{4}{5} - \frac{4}{7} + \frac{4}{9} \cdots$$

Starting with the first term and adding successive terms produces the following progression (converted into decimals):

$$4 \rightarrow 2.667 \rightarrow 3.467 \rightarrow 2.896 \rightarrow 3.340 \rightarrow \ldots$$

The total gets closer and closer to pi in smaller and smaller leaps. Yet this method requires more than 300 terms to get an answer for pi accurate to two decimal places, so it was impracticable for those wanting to use it to find more digits in the decimal expansion.

Eventually, calculus provided other infinite series for pi that were less pretty but more effective for number-crunching. In 1705 the astronomer Abraham Sharp used one to calculate pi to 72 decimal places, smashing van Ceulen’s century-old record of 35. While this was quite an achievement, it was also a useless one. There is no practical reason to know pi to 72 digits, or to 35 digits for that matter. Four decimal places is enough for the engineers of precision instruments. Ten decimal places is sufficient to calculate the circumference of the Earth to within a fraction of a centimetre. With 39 decimal places, it is possible to compute the circumference of a circle surrounding the known universe to within an accuracy of a radius of a hydrogen atom. But practicality wasn’t the point. Application was not a concern to the pi men of the Enlightenment; digit-hunting was an end in itself, a romantic challenge. A year after Sharp’s effort, John Machin reached 100 digits and in 1717 the Frenchman Thomas de Lagny added a further 27. By the turn of the century, the Slovenian Jurij Vega was in the lead with 140.

Zacharias Dase, the German lightning calculator, pushed the record for pi to 200 decimal places in 1844 in an intense two-month burst. Dase used the following series, which looks more convoluted than the pi formula above but is in fact far more user-friendly. This is because, first, it bears down on pi at a respectable rate. An accuracy of two decimal places is reached after the first nine terms. Second, the $\frac{1}{2}$, $\frac{1}{5}$ and $\frac{1}{8}$ that reappear every third term are very convenient to manipulate. If $\frac{1}{5}$ is rewritten as $\frac{2}{10}$ and $\frac{1}{8}$ as $\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}$, all the multiplications involving those terms can be reduced to combinations of doubling and halving. Dase would have written out a reference table of doubles to help him with his calculations, starting 2, 4, 8, 16, 32, and carried on for as far as he needed – which, since he was calculating pi to 200 places, would be when the final double is 200 digits long. This happens after 667 successive doublings.

Dase used this series:

$$
\begin{aligned}
\frac{\pi}{4} &= \frac{1}{2} - \frac{(\tfrac{1}{2})^3}{3} + \frac{(\tfrac{1}{2})^5}{5} - \dots \\
&\quad + \frac{1}{5} - \frac{(\tfrac{1}{5})^3}{3} + \frac{(\tfrac{1}{5})^5}{5} - \dots \\
&\quad + \frac{1}{8} - \frac{(\tfrac{1}{8})^3}{3} + \frac{(\tfrac{1}{8})^5}{5} - \dots \\
&= \frac{1}{2} + \frac{1}{5} + \frac{1}{8} - \left[\frac{(\tfrac{1}{2})^3}{3} + \frac{(\tfrac{1}{5})^3}{3} + \frac{(\tfrac{1}{8})^3}{3}\right] \\
&\quad + \left[\frac{(\tfrac{1}{2})^5}{5} + \frac{(\tfrac{1}{5})^5}{5} + \frac{(\tfrac{1}{8})^5}{5}\right] - \dots
\end{aligned}
$$

So $\pi = 4\ (0.825 - 0.0449842 + 0.00632 - \dots$

After one term this is 3.3

After two terms this is 3.1200

After three terms this is 3.1452

Dase barely had time to rest on his laurels before the Brits decided to covet his achievement, and within a decade William Rutherford had calculated pi to 440 places. He encouraged his protégé William Shanks, an amateur mathematician who ran a boarding school in County Durham, to go even further. In 1853 he reached 607 digits, and by 1874 he was at 707. His record held for 70 years until D.F. Ferguson, of the Royal Naval College in Chester, found a mistake in Shanks's calculation. He had made an error at the 527th place so all the subsequent numbers were wrong. Ferguson spent the last year of the Second World War calculating pi longhand. One can only assume that he thought the war was already won. By May 1945 he had 530 places, and by July 1946 he reached 620, and no one has ever calculated further using only paper and pen.

Ferguson was the last of the manual digit-hunters and the first of the mechanical. Using a desk calculator he added almost another 200 places in just over a year, so in September 1947 pi was known to 808 decimal places. Computers then changed the race. The first of them to battle pi was the Electronic Numerical Integrator and Computer, or ENIAC, which was built in the final years of the Second World War at the US Army's Ballistic Research Laboratory in Maryland. It was the size of a small house. In September 1949 the ENIAC took 70 hours to calculate pi to 2037 digits – smashing the record by more than a thousand decimal places.

As more and more digits were found in pi, one thing seemed pretty clear: the numbers obeyed no obvious pattern. Yet it was only in 1767 that mathematicians were able to prove that the higgledy-piggledy sequence of digits would never repeat itself. The discovery followed from considering what type of number pi might be.

The most familiar *type* of number is the *natural numbers*. They are the counting numbers starting at one:

1, 2, 3, 4, 5, 6…

The natural numbers, however, are limited in scope because they expand in only one direction. More ul are the *integers,* which are the natural numbers together with zero and the negatives of the natural numbers:

… –4, –3, –2, –1, 0, 1, 2, 3, 4…

The integers cover every positive or negative whole number from minus infinity to plus infinity. If there were a hotel with an unlimited amount of floors and an unlimited amount of lower and lower basements, the buttons in the elevator would be the integers.

Another basic type of number is the *fraction*, which are the numbers written when *a* and *b* are integers but *b* is not 0. The top number in a fraction is the *numerator* and the bottom number is the *denominator*. If we have several fractions, the *lowest common denominator* is the lowest number that can be divided by all the denominators without leaving a remainder. So, if we have $\frac{1}{2}$ and $\frac{3}{10}$, the lowest common denominator is 10, since both 2 and 10 divide into 10. What about the lowest common denominator of $\frac{1}{3}$,$\frac{3}{4}$,$\frac{2}{9}$ and $\frac{7}{13}$? In other words, what is the smallest number that 3, 4, 9 and 13 divide into? The answer is surprisingly big: 468! I mention this to make a semantic rather than a mathematical point. The phrase ‘lowest common denominator’ is often used to describe something basic or unsophisticated. It sounds evocative, but misrepresents the arithmetic. Lowest common denominators can often be big and unconventional: 468 is quite an impressive number! A more arithmetically meaningful phrase for something mainstream and cheap is *highest common* *factor* – which is the largest number that can be divided into every one of a group of numbers. The highest common factor of 3, 4, 9 and 13, for example, is 1, and you can’t get much lower or more unsophisticated than that.

Since fractions are equivalent to ratios between integers, they are also called *rational numbers*, and there is an infinite amount of them. In fact, there is an infinite number of rational numbers between 0 and 1. For example, let’s take every fraction where the numerator is 1 and the denominator is a natural number bigger than or equal to 2. This is the set composed of:

$$\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6} \; \ldots$$

We can go further and prove that there is an infinite number of rational numbers between *any two* rational numbers. Let *c* and *d* be any two rational numbers, with *c* less than *d*. The point halfway between *c* and *d* is a rational number: it is $\frac{(c+d)}{2}$. Call this point *e*. We can now find a point halfway between *c* and *e*. It is $\frac{(c+e)}{2}$. It is rational and also between *c* and *d*. We can carry on ad infinitum, always splitting the distance between *c* and *d* into smaller and smaller parts. No matter how tiny the distance between *c* and *d* is in the first place, there will always n infinite amount of rational numbers in between them.

Since we can always find an infinite number of rational numbers between any two rational numbers, it might be thought that the rational numbers cover every number. Certainly, this is what Pythagoras had hoped. His metaphysics was based on the belief that the world was made up of numbers and the harmonic proportions between them. The existence of a number that could not be described as a ratio diminished his position, at the very least, if it did not contradict it outright. Yet unfortunately for Pythagoras, there *are* numbers that cannot be expressed in terms of fractions, and – rather embarrassingly for him – it is his own theorem that leads us to one. If you have a square where each side has length 1, then the length of the diagonal is the square root of two, which cannot be written as a fraction. (I have included a proof as an appendix.)

Numbers that cannot be written as fractions are called *irrational*. According to legend, their existence was first proved by the Pythagorean disciple Hippasus, which did not endear him to the Brotherhood; he was declared a heretic and was drowned at sea.

When a rational number is written out as a decimal fraction, either it has a finite amount of digits, in the way that $\frac{1}{2}$ can be written 0.5, or the expansion will end up repeating itself, just as $\frac{1}{3}$ is 0.3333…where the 3s go on for ever. Sometimes the recurring loop is more than one digit, as is the case of $\frac{1}{11}$, which is 0.090909…where the digits 09 repeat for ever, or $\frac{1}{19}$, which is 0.0526315789473684210…where 052631578947368421 repeats for ever. By contrast, and this is the crucial point, when a number is irrational its decimal expansion never repeats.

In 1767 the Swiss mathematician Johann Heinrich Lambert proved that pi was indeed irrational. The early pi-men may have hoped after the initial chaos of 3.14159…that the noise would calm and a pattern ensue. Lambert's discovery confirmed that this was impossible. Pi's decimal expansion cavorts towards infinity in a predestined yet apparently indiscriminate way.

Mathematicians interested in the irrationals wanted to categorize them further. In the eighteenth century they started to speculate about a special type of irrational called *transcendental numbers*. These were numbers so mysterious and evasive that finite mathematics could not capture them. The square root of two, $\sqrt{2}$, for example, is irrational but can be described as the solution to the equation $x^2 = 2$. A transcendental number is an irrational that cannot be described by an equation with a finite amount of terms. When the concept of transcendental numbers was first mooted, no one knew if they even existed.

They did exist, although it took about a hundred years before Joseph Liouville, a French mathematician, came up with a few examples. Pi was not among them. Only after another 40 years did the German Ferdinand von Lindemann prove that pi was indeed transcendental. The number existed beyond the realm of finite algebra.

Lindemann's discovery was a milestone for number theory. It also settled, once and for all, what was probably the most celebrated unsolved problem in mathematics: whether or not it was ple to square the circle. In order to explain how it did this, however, I need to introduce the formula that says that the area of a circle is πr^2, where r is the radius. (The radius is the distance from the centre to the side, or half the diameter.) A visual proof of why this is true is an instance where a pie is the best metaphor for pi. Imagine you have two same-sized circular pies, a white one and a grey one, as below in A. The circumference of each pie is pi times the diameter, or pi times twice the radius, or $2pr$. When sliced into equal segments the pieces can be rearranged, as in B with quarter segments, or as in C with ten segments. In both cases the length of the side remains $2pr$. If we keep on slicing smaller and smaller segments, then the shape would eventually become a rectangle, as in D, with sides r and $2\pi r$. The area of the rectangle – which is the area of the two pies – is therefore $2\pi r^2$, so the area of one pie is πr^2.

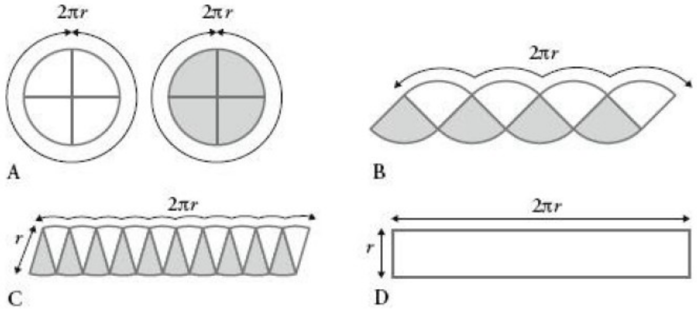

Proof that the area of a circle = πr^2.

To square a circle we must construct (using just a compass and straightedge) a square that has the same area as a given circle. Now we know that a line with length r is the radius of a circle with area πr^2, and we also know that a square with area πr^2 must have a side length r $r\sqrt{\pi}$(since ($(r\sqrt{\pi})^2 = r^2(\sqrt{\pi})^2 = r^2\pi = \pi r^2$)). So squaring the circle can be reduced to the challenge of constructing the length $r\pi$ from the length r. Or, taking r as 1 for convenience, the length π from 1.

Using coordinate geometry, which I will cover later, it is possible to express the process of the construction of a line algebraically, as a finite equation. It can be shown that so long as x is the solution to a finite equation, then starting with a line of length 1 we can construct a line with length x. But if x is not the solution to a finite equation – in other words, if x is transcendental – it is impossible to construct a line with length x. Now, the fact that p is transcendental means that π is also transcendental. (You’re going to have to trust me on that.) So it is impossible to construct the length π. The transcendence of pi proves that the circle cannot be squared.

Lindemann’s proof of the transcendence of pi dashed what had been the dream, for thousands of years, of countless mathematicians. Perhaps the most eminent figure ever to have declared that he had squared the circle was Thomas Hobbes, the seventeenth-century English thinker whose book *Leviathan* founded political philosophy. Having become a keen amateur geometer in later life, Hobbes published his solution when he was 67. Even though circle-squaring was still an open question at the time, his proof was received with bemusement by the scientific community. John Wallis, professor at Oxford and the finest British mathematician before Isaac Newton, exposed Hobbes’s errors in a pamphlet, thus setting in motion one of the most entertaining – and pointless – feuds in the history of British intellectual life. Hobbes replied to Wallis’s comments with an addendum to his book entitled *Six Lessons to the Professors of Mathematics*. Wallis countered with *Due Correction for Mr Hobbes in School Discipline for not saying his Lessons right*. Hobbes followed this with *Marks of the Absurd Geometry, Rural Language, Scottish Church Politics and Barbarisms of John Wallis*. This led to Wallis’s *Hobbiani Puncti Dispunctio! or the Undoing of Mr Hobbes’s Points*. The quarrelling lasted almost a quarter of a century, until Hobbes’s death in 1679. Wallis rather enjoyed the hostilities, since it was a way of casting aspersions on Hobbes’s political and religious views, which he despised. And, of course, he was right. In many disputes truths are shared between both sides. Not in Hobbes vs Wallis. Hobbes could not square the circle because it is impossible to do so.

Proof that you cannot square the circle has not put people off trying. In 1897 the Indiana state legislature famously considered a bill containing a proof of squaring the circle by E.J. Goodwin, a country doctor, who offered it ‘as a gift to the State of Indiana’. He was, of course, misguided. Since Ferdinand von Lindemann in 1882, the phrase circle-squarer has been mathspeak for crank.

Pi’s enigmatic attributes were, in the eighteenth and nineteenth centuries, revealed not only to be at the heart of ancient geometrical problems but also deeply rooted in new fields of science that were not always obviously related to circles. ‘This mysterious 3.141592…which comes in at every door and window, and down every chimney,’ wrote the British mathematician Augustus De Morgan. For example, the time it takes a pendulum to swing is dependent on pi. The distribution of deaths in a population is a function of pi. If you toss a coin $2n$ times, the probability when n is very large of getting exactly 50 percent heads and 50 percent tails is $\frac{1}{\sqrt{(n\pi)}}$.

The man whose name is most associated with offbeat occurrences of pi was the French polymath Georges-Louis Leclerc, the Comte de Buffon (1707–88). Of Buffon's many colourful scientific endeavours, perhaps his most ambitious was the construction of a working version of Archimedes' weapon of mirrors, with which he was said to have set fire to ships. Buffon's contraption was made up of 168 flat mirrors, each six by eight inches, and it was able to ignite a wooden plank at a distance of 150 feet, a good effort, though on a different scale from setting a Roman fleet ablaze.

With regard to pi, Buffon is remembered for having devised an equation that led to a new method for calculating pi, though Buffon did not himself make the connection. Buffon arrived at his equation by studying an eighteenth-century gambling game called 'clean tile', in which you throw a coin on to a tiled surface and bet on whether it will touch the cracks between tiles or rest cleanly. Buffon came up with the following alternative scenario: imagine that a floor is marked with parallel lines spaced evenly apart and that a needle is thrown on it. He then correctly calculated that if the length of the needle is l and the distance between lines is d, then the following equation holds:

$$\text{Probability of the needle touching the line} = \frac{2l}{\pi d}$$

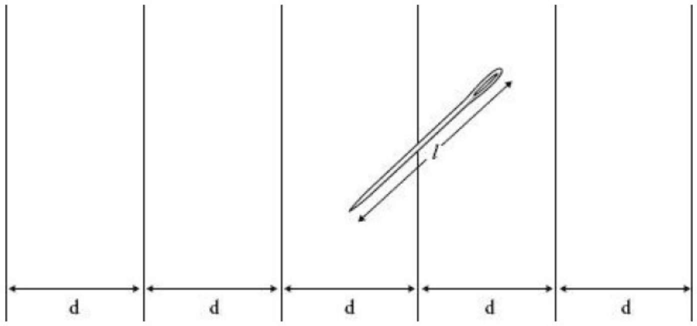

A few years after Buffon died, Pierre Simon Laplace realized that this equation could be used to estimate a value for pi. If you throw lots and lots of needles on the floor, then the ratio of the number of times that the needle hits the line to the total number of throws will be approximately equal to the mathematical probability of the needle touching the line. In other words, after many throws

$$\frac{\text{number of times needle touches a line}}{\text{number of total throws}} \approx \frac{2l}{\pi d}$$

or:

$$\pi \approx \frac{2l\,(\text{number of total throws})}{d\,(\text{number of times needle touches a line})}$$

(The symbol ≈ means ‘is approximately equal to’.)

Even though Laplace was the first to write about how pi could be estimated this way, his work followed from Buffon’s equation, so Buffon is the person remembered for it. His achievement put him in esteemed company as a member of the club of mathematicians, including Archimedes and Leibniz, who each found a new way to calculate pi.

The more throws of the needle that are taken, the better the approximation, and aiming needles at boards has become a standard diversion for mathematicians unable to think of more creative ways to pass the time. You need, however, to keep on going a fair while before any interesting result is achieved. An early adopter is said to have been a certain Captain Fox in the American Civil War, who, while recovering from battle wounds, threw a piece of wire eleven hundred times on a board of parallel lines and managed to derive pi to 2 decimal places.

Pi’s mathematical properties have made it a celebrity among numbers, and also a more general cultural icon. Because pi’s digits never repeat, they are perfect for feats of memorization. If remembering numbers is your thing, the *ne plus ultra* of digits is the digits in pi. This has been a pastime since at least 1838, when *The Scotsman* reported that a 12-year-old Dutch boy recited all the 155 digits that were known at the time to an audience of scientists and royals. Akira Haraguchi, a 60-year-old retired engineer, holds the current world record. In 2006 he was filmed in a public hall near Tokyo reciting pi to 100,000 decimal places. The performance took him 16 hours and 28 minutes, including five-minute breaks every two hours to eat rice balls. He explained to a journalist that pi symbolized life since its digits never repeated and followed no pattern. Memorizing pi, he added, was ‘the religion of the universe’.

Pi memorization gets a little dull, but pi memorization while juggling, now there’s a competitive sport! The record is held by Mats Bergsten, an actuary in Sweden in his late fifties, who has recited 9778 digits while juggling with three balls. He told me, however, that he is proudest of his achievements in the ‘erest test’, in which the first 10,000 digits of pi’s expansion are divided into 2000 groups of five, beginning with 14159. In the test 50 groups are randomly read out, and the contestant has to say from memory which five numbers precede and succeed each of them. Mats Bergsten is one of only four people in the world who can do this with no errors, and his time, 17min 39secs, is the fastest. It is much more mentally straining, he told me, to recall 10,000 digits randomly than merely remembering them in order.

When Akira Haraguchi recited 100,000 digits of pi by heart, he used a mnemonic technique, assigning syllables to each number from 0 to 9 and then translating pi’s decimals into words, which in turn formed sentences. The first fifteen digits sounded like: ‘the wife and children have gone abroad; the husband is not scared.’ Using words to remember the digits in pi this way is used by schoolchildren in cultures all over the world, but usually this is done not by assigning syllables, but by creating a phrase in which the number of letters in each word represents each consecutive digit in the decimal expansion of pi. A well-known English one is credited to the astrophysicist Sir James Jeans: *How I need a drink, alcoholic in nature, after the heavy lectures involving quantum mechanics. All of thy geometry, Herr Planck, is fairly hard*. ‘How’ has 3 letters, ‘I’ has 1, ‘need’ has 4, and so on.

Among numbers, only pi has inspired this type of fandom. No one wants to memorize the square root of two, which is just as challenging. Pi is also the only number to have inspired its own literary subgenre. Constrained writing is a technique in which some condition is adopted that imposes a pattern or forbids certain things in the text. Entire poems – or ‘piems’ – have been written under the constraint that the number of letters per word is determined by pi, usually with the convention that a 0 in the expansion requires a ten-letter word. The most ambitious piem is the *Cadaeic Cadenza* by Mike Keith, which follows pi for 3835 digits. It begins as a pastiche of Edgar Allan Poe:

> One; A poem
>
> A Raven
>
> Midnights so dreary, tired and weary,
>
> Silently pondering volumes extolling all by-now obsolete lore. During my rather long nap – the weirdest tap!
>
> An ominous vibrating sound disturbing my chamber’s antedoor. ‘This,’ I whispered quietly, ‘I ignore.’

Keith says that writing with a difficult constraint is an exercise both in discipline and discovery. Since the digits in pi are random, the constraint is, he said, ‘like bringing order out of chaos’. When I asked him ‘Why pi?’ he replied that pi was ‘a metaphor for all things infinite, or inscrutable, or unpredictable, or full of endless wonder’.

Pi has gone by this name only since 1706, when the Welshman William Jones introduced the symbol p in his book, the snappily titled *A New Introduction to the Mathematics, for the Use of some Friends who have neither Leisure, Convenience, nor, perhaps, Patience, to search into so many different Authors, and turn over so many tedious Volumes, as is unavoidably required to make but tolerable progress in the Mathematics*. The Greek letter, which was probably an abbreviation for the word periphery, did not immediately catch on, howeve, becoming standard notation for pi only 30 years later when Leonhard Euler adopted it.

Euler was the most prolific mathematician of all time (he published 886 books), and he is possibly the one who contributed most to an understanding of pi. It was his improved formulae for pi that enabled the eighteenth-and nineteenth-century digit-hunters to peel back more and more decimal places. In the beginning of the twentieth century the Indian mathematician Srinivasa Ramanujan devised many more Euler-style infinite series for pi.

Ramanujan was a largely self-taught mathematician who worked as a clerk in Madras before writing a letter to Cambridge university professor G.H. Hardy. Hardy was flabbergasted to see that Ramanujan had rediscovered results that had taken centuries to achieve, and invited him to England, where the men collaborated before Ramanujan died aged 32. His work showed an extraordinary intuition about the properties of numbers, including pi, and his most famous formula is the following:

$$\frac{1}{\pi} = \frac{2\sqrt{2}}{9801} \sum_{n=0}^{\infty} \frac{(4n)!(1103 + 26390n)}{(n!)^4 396^{4n}}$$

The $\sum_{n=0}^{\infty}$ symbol indicates a series of values all added up, starting with the value when *n* equals zero, added to the value when *n* equals one, and so on to infinity. Even without understanding the notation, however, one can appreciate the drama of such an equation. The Ramanujan formula races towards pi with remarkable speed. From the very start, when *n* is 0 the formula has one term and gives a value of pi accurate to six decimal places. For each increase in the value of *n*, the formula adds roughly eight new digits to pi. It is an industrial-strength pi-making machine.

Inspired by Ramanujan, in the 1980s the Ukrainian-born mathematicians Gregory and David Chudnovsky devised an even more ferocious formula. Each new term adds roughly 15 digits.

$$\frac{1}{\pi} = \sum_{n=0}^{\infty} (-1)^n \times \frac{(6n)!}{(3n)!\, n!^3} \times \frac{163096908 + 6541681608n}{(262537412640768000)^{n+½}}$$

The first time I saw the Chudnovsky formula I was standing on it. Gregory and David are brothers and they share an office at the Polytechnic University in Brooklyn. It consists of an open-plan space with a sofa in the corner, a couple of chairs and a blue floor decorated with dozens of formulae for pi. ‘We wanted to put something on the floor and what else can you put on the floor other than stuff which relates to mathematics?' explained Gregory.

In fact, the pi floor pattern was their second choice. The original plan was to lay down a giant reproduction of *Melencolia I* by Albrecht Dürer (reproduced on chapter 6). The sixteenth-century woodcut is beloved of mathematicians since it is full of playful references to numbers, geometry and perspective.

'One night, when there was nothing on the surface, we printed 2000 or more pages of [*Melencolia I*] and we laid it on the floor,' said David. 'But if you walked around it, you wanted to throw up! Because your point of view changes extremely abruptly.' David began to study the floors of the cathedrals and castles of Europe in order to work out how he could decorate the office without inducing nausea in anyone walking through. 'I discovered they are mostly laid out in a –'

'Simple geometric style,' interrupted Gregory.

'Black, white, black, white squares…#8217; said David.

'You see, if you really have a complex picture and you try to walk on it, the angle changes so abruptly that your eyes don't like it,' added Gregory. 'So the only way you can do something like that is to –'

'Hang from the ceiling!' David shouted in my left ear, and both men lost themselves in guffaws.

Talking to the Chudnovskys was like wearing stereo headphones with an erratically alternating connection to both ears. They sat me on their sofa and sat on either side of me. Constantly interrupting each other, they finished each other's sentences, speaking in a highly melodic English thick with Slavic tones. The brothers were born in Kiev, when it was in the Soviet Republic of Ukraine, although they have lived in the United States since the late 1970s and are American citizens. They have collaborated on so many papers and books together that they encourage you to think of them as one mathematician, not two.

For all their genetic, conversational and professional homogeneity, however, the men look very different. This is mostly because Gregory, who is 56, suffers from *myasthenia gravis*, an auto-immune disorder of the muscles. He is so thin and frail that he spends most of his time lying down. I never saw him get up off the sofa. Still, the energy his limbs lacked was compensated for by a brilliantly expressive face that burst into life as soon as he talked about maths. He has pointed features, large brown eyes, a white beard and wispy unkempt hair. David, who has blue eyes, is five years older, rounder of body and fuller of face.

He was clean-shaven and his short hair was hidden under an olive-green baseball cap.

The Chudnovskys are arguably the mathematicians who have done the most to popularize pi in recent years. In the early 1990s they built a supercomputer in Gregory's Manhattan apartment out of mail-order parts that, using their own formula, calculated the number to more than two billion decimal places – a record at the time.

This amazing achievement was chronicled in a *New Yorker* article, which in turn inspired the 1998 film *Pi*. The main protagonist was an unruly-haired maths genius looking for hidden patterns in stock-market data on a homemade supercomputer. I was curious to discover if the Chudnovskys had seen the film, which garnered favourable reviews and has become a reference for low-budget, black-and-white psycho-mathematical thrillers. 'No, no, we haven't seen it,' said Gregory.

'You have to realize, usually movie-makers repeat their internal state,' added David sarcastically.

I tell them I thought they might have been flattered by the attention.

'No, no,' grinned Gregory.

'Let me tell you another thing,' David cut in. 'Two years ago I came back from France. A couple of days before I left, there was a huge book fair. I stopped at a stand where there was a book that had a detective story on it. It was written by an engineer. It was a murder mystery, you know. A lot of dead bodies, mostly mistress in hotel, and the source that determined everything he did was pi.'

Gregory was smiling from ear to ear and said under his breath: 'OK, I am not going to read this book, *zat's* for sure.'

David carried on: 'So I talked to the guy. He is a very eucated man.' He paused, shrugged his shoulders and raised his pitch by an octave: 'As I say, I bear no responsibility!'

David said that he was taken aback the first time that he saw billboards advertising the Givenchy perfume. 'All the way down the street was pi…pi…pi…' He was wailing now: '*Pi…pi…pi!* Do I bear any responsibility?'

Gregory glanced at me and said: 'For some reason, the general public is fascinated with this stuff. They get kind of the wrong inference.' There are many professional mathematicians, he said, who study pi. He added wryly: 'Usually these people are not allowed to see the light of day.’

In the 1950s and 1960s advances in computer technology were reflected in the number of new digits found in pi. By the end of the 1970s the record had been broken nine times and stood at just over a million decimal places. In the 1980s, however, a combination of even faster computers and brand-new algorithms led to a frenetic new era of digit-hunting. Yasumasa Kanada, a young computer scientist at Tokyo University, was first off the block in what became a two-way pi race between Japan and the United States. In 1981 he used an NEC computer to calculate pi to 2 million digits in 137 hours. Three years later he was up to 16 million. William Gosper, a mathematician in California, then nudged into the lead with 17.5 million, before David H. Bailey, at NASA, bettered him with 29 million. In 1986 Kanada overtook them both with 33 million and broke his own record three times in the next two years to reach 201 million with a new machine, the S-820, which did the calculation in just under six hours.

Away from the digit-hunting spotlight, the Chudnovskys were also beavering away on pi. Using a new method of communication called the internet, Gregory connected the computer at his bedside to two IBM supercomputers at different sites in the US. The brothers then devised a program to calculate pi based on the new superfast pi formula they had discovered. They were allowed on to the computers only when no one else was using them, at nights and weekends.

‘It was a great thing,’ remembered Gregory nostalgically. In those days there was no computer capacity to store the numbers that the brothers were calculating. ‘They kept the pi on magnetic tape,’ he said.

‘Mini-tape. And you had to call the guy and ask…’ added David.

‘And say tape number such and such,’ continued Gregory. ‘And sometimes if somebody else is more important your tapes are dismounted in the middle of computation.’ His eyes rolled as if to throw his hands in the air.

Despite the obstacles, the Chudnovskys kept on going, pushing beyond a billion digits. Kanada then nudged ahead of them briefly, before the Chudnovskys retook the lead with 1.13 billion. David and Gregory then decided that if they were serious about calculating pi they needed their own machine.

The Chudnovskian supercomputer lived in a room in Gregory’s apartment.

Made up of processors linked by cables, the whole thing cost, according to their estimates, about $70,000. It was a steal, compared to the millions of dollars that would have bought a machine of similar capacity; although it came with its own complications to their lifestyles. The computer, which they called *m zero*, was switched on at all times, just in case switching it off was irreversible, and keeded 25 fans in the room to keep it cool. The brothers were careful not to switch on too many lights in the apartment just in case the added demand blew the wiring.

In 1991 David and Gregory's homemade contraption calculated pi to more than two billion places. Then they got distracted by other problems. By 1995 Kanada was ahead once again, and he reached 1.2 trillion digits in 2002, a record that lasted only until 2008, when compatriots at the University of Tsukuba revealed 2.6 trillion. In December 2009 the Frenchman Fabrice Bellard claimed a new record using the Chudnovsky formula: almost 2.7 trillion places. The calculation had taken 131 days on his desktop PC.

If you wrote a trillion digits in small type, the distance would cover from here to the sun. If you put 5000 digits on a page (which is very small type) and stacked the pages on top of each other, the pi in the sky would be 10km high. What is the point of calculating pi to such absurd lengths? One reason is very human: records exist to be broken.

But there is another, more important motivation. Finding new digits in pi is ideal for testing the processing capacity and reliability of computers. 'I have no interest as a hobby for extending the known value of pi itself,' Kanada once said. 'I have a major interest for improving the performance of the computation.' Pi calculation is now essential for quality-testing supercomputers because it is a 'high-duty job which requires large main memory, operates huge number-crunching and gives [an] easy [way] to check [the] correct answer. Mathematical constants like the square root of two, *e*[*] [and] gamma are some of the candidates, but pi is most effective.'

The story of pi has wonderful circularity. It is the simplest and most ancient ratio in maths, which has been reinvented as a massively important tool on the frontline of computer technology.

In fact, the Chudnovskys' interest in pi came primarily from their desire to build supercomputers, a passion that still burns brightly. The brothers are currently designing a chip that they claim will be the fastest in the world, only 2.7cm wide but containing 160,000 smaller chips and 1.75km of wire.

On discussion of their new chip, Gregory became very high-spirited:

‘Computers double their power every 18 months, not because they are faster but because they can pack more stuff in. But there is a catch,’ he said. The mathematical challenge was how to partition the smaller pieces so that they can talk to each other in the most efficient way. His laptop showed the chip’s circuitry. ‘I’d say the problem with this chip is that it is a capitalist chip!’ he exclaimed. ‘The problem is that most of the stuff here is not doing anything. There are not too much proles here.’ He pointed at one section. ‘This is just management of the stores inside the chip,’ he lamented. ‘The majority of these guys just do warehousing and accounting. This is awful! Where is the manufacturing sector?’

In Carl Sagan’s bestselling book *Contact*, an extraterrestrial informs a woman on Earth that after a certain amount of digits the randomness in pi stops and there is a message written in 0s and 1s. This message occurs after 10^{20} decimal places – which is the number described by 1 followed by 20 zeros. Since we currently know pi to ‘only’ 2.7 trillion places (27 followed by 11 zeros), we have a little way to goe s to check that he was making it up. Actually, we have further to go since the message is apparently written in base 11.

The idea that there is a pattern in pi is an exhilarating one. Mathematicians have been looking for signs of order in the decimal expansion of pi for as long as there have been decimal expansions. The irrationality of pi means that the numbers keep on spewing out with no repeated pattern, but this does not eliminate the possibility of patches of order – such as a message written in 0s and 1s. So far, however, no one has found anything significant. The number, does, though, have its quirks. The first 0 comes in position 32, which is much later than expected if the digits are randomly distributed. The first time a digit is repeated six times consecutively is 999999 at the 762nd decimal place. The likelihood of six 9s occurring so early if they occur randomly is less than 0.1 percent. This sequence is known as the Feynman point, since physicist Richard Feynman once remarked that he wanted to memorize pi until that point and finish by saying ‘nine, nine, nine, nine, nine, nine, and so on’. The next time pi throws up six consecutive identical digits happens in position 193,034 and they are also 9s. Is this a message from beyond, and if so what is it saying?

A number is considered *normal* if each of the digits 0 to 9 occurs equally often in the number’s decimal expansion. Is pi normal? Kanada looked at the first 200 billion digits of pi and found that the digits occurred with the following frequencies:

0	20,000,030,841	5	19,999,917,053
1	19,999,914,711	6	19,999,881,515
2	20,000,136,978	7	19,999,967,594
3	20,000,069,393	8	20,000,291,044
4	19,999,921,691	9	19,999,869,180

Only the digit 8 seems a little overabundant, yet the difference is statistically insignificant. It would appear that pi is normal, yet no one has been able to prove it. And neither has anyone been able to prove that such a proof is impossible. There is a chance, therefore, that pi is *not* normal. Maybe after 10^{20} there really are just 0s and 1s?

A different, but related issue, is the positioning of the numbers. Are they distributed randomly? Stan Wagon analysed the first ten million digits of pi with a 'poker test': take five consecutive digits and consider them as if they are a poker hand.

Type of hand	**Actual occurrence**	**Expected occurrence**
All digits different	604,976	604,800
One pair, three different	1,007,151	1,008,000
Two pairs	216,520	216,000
Three of a kind	144,375	144,000
Full house	17,891	18,000
Four of a kind	8887	9000
Five of a kind	200	200

The right column is how many times we would expect to see the poker hands if pi was normal and each decimal place had an equal chance of being occupied by any digit. The results are well within the boundaries of what we would expect. Each pattern of numbers seems to appear with the frequency it would had each decimal place been randomly generated.

There are websites that can find you the first occurrence in pi of the numbers of your birthday. The first time the sequence 0123456789 occurs is at the 17,387,594,880th place – which was discovered only when Kanada got that far in 1997.

I asked Gregory if he ever believed an order would be found in pi. ‘There is no order,’ he replied dismissively. ‘And if there would be an order, it would be weird and not right. So there is no point wasting the time.’

Instead of focusing on patterns in pi, some see its very randomness as a tremendous expression of mathematical beauty. Pi is predetermined but it seems to mimic randomness extraordinarily well.

‘It is a very good random number,’ agreed Gregory.

Shortly after the Chudnovskys started computing pi they got a call from the United States government. David squealed an impersonation of the voice at the end of the line: ‘Could you please send pi?’

Random numbers are needed in industry and commerce. For example, just say a market-research company needs to poll a representative sample of a thousand people from a population of a million. The company will use a random-number generator to select the sample group. The better the generator is at providing random numbers, the more representative the sample will be – and the poll will be more accurate. Likewise, streams of random numbers are needed to simulate unpredictable scenarios when testing computer models. The more random the numbers, the more robust the testing will be. In fact, projects can fail if the random numbers used to test them are not random enough. ‘You are only as good as your random number,’ remarked David. ‘If you use terrible random numbers, you will end up in a terrible condition,’ concluded Gregory. Of all sets of available random numbers, the decimal expansion of pi is the best.

Yet there is a philosophical paradox. Pi is self-evidently not random. Its digits may behave like random digits, but they are fixed. For example, if the digits in pi were truly random then there would be only a 10 percent chance that the first digit after the decimal point is a 1, yet we know that it is 1 with absolute certainty. Pi exhibits randomness non-randomly – which is fascinating, and weird.

Pi is a mathematical concept that has been studied for thousands of years, yet it holds many secrets. There have been no great advances in understanding its nature since transcendence was proved almost a century and a half ago.

‘We actually don’t know about most of this stuff,’ said Gregory.

I asked if there would ever be a new advance in understanding pi.

‘Of course, of course,’ said Gregory, ‘There are always advances. Mathematics moves forward.’

‘It will be more miraculous but it won’t be nice,’ David said.

Nineteen sixty-eight was a year of counter-cultural uprisings around the world, and Britain was not immune to such generational upheaval. In May the Treasury announced the introduction of a revolutionary new coin.

The 50p piece was designed to replace the old ten-shilling note as part of the switch from imperial to decimal currency. Yet what set the coin apart was not its denomination, but its unorthodox shape.

‘It’s no ordinary coin,’ exhorted the *Daily Mirror*. ‘Why, the Decimal Currency Board even go so far as to call it a “multilateral-curve heptagon”.’ Never before had a country introduced a seven-sided coin. And never before had a nation been so outraged over the aesthetics of a geometric shape. Leading the foment was retired army colonel Essex Moorcroft, of Rosset, Derbyshire, who formed the Anti-Heptagonists. ‘We take our motto from Cromwell’s heartfelt cry, “Take away this bauble”. I have founded the society because I believe our Queen is insulted by this heptagonal monstrosity,’ he said. ‘It is an ugly coin and an insult to our Sovereign, whose image it bears.’

Nonetheless, the 50p entered circulation in October 1969, and Colonel Moorcroft did not take to the barricaes. Indeed, by January 1970 *The Times* reported that ‘the curvaceous heptagon seems to have won for herself some affection’. Today the 50p piece is considered a distinctive and cherished part of British heritage. When a 20p piece was introduced in 1982, it too was heptagonal.

The 50p and 20p pieces are, in fact, design classics. Their seven-sided shape means that they are easily distinguished from circular coins, helping in particular the blind and partially sighted. They are also the most thought-provoking coins in circulation. The circle is not the only interesting round shape in mathematics.

A circle can be defined as a curve for which every point is equidistant from a fixed point, the centre. This property has many practical applications. The wheel – generally trumpeted as humanity’s first great invention – is the most obvious. An axle attached to the centre of a wheel will stay at a fixed point above the ground when the wheel rotates smoothly along a surface, which is why carts, cars and trains run smoothly without bobbing up and down.

For the transportation of very heavy loads, however, an axle might not take the weight. One alternative is to use rollers. A roller is a long tube with a circular cross-section, laid out on the ground. If a heavy object with a flat base (such as a giant cuboidal piece of stone intended for a pyramid) is put on several rollers, then it can be pushed smoothly over the rollers, with new rollers being put in front as it inches forward.

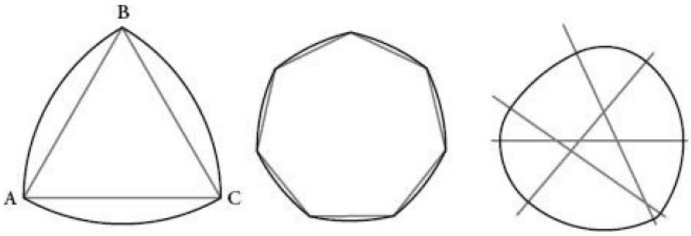

Curves of constant width include the Reuleaux triangle (left) and the multilateral-curve heptagon, better known as the 50p piece (centre).

The crucial feature of a roller is that the distance between the ground and the top of the roller is always the same. This is obviously the case with a circular cross-section, since the width of a circle (the diameter) is always the same.

Do all rollers have to have a circular cross-section? Are there any other shapes that will work? It may appear counter-intuitive, but there are, in fact, a variety of shapes that will make perfect rollers. An example is the 50p piece.

If you welded coins together to form rollers that had a 50p piece as a cross-section and then placed this book on these rollers, the book would not wobble up and down as you pushed it along. It would ride as smoothly as if on circular cylinders.

The reason this happens is because the 50p piece is a *curve of constant width*. Wherever you measure it around its perimeter, the 50p has the same breadth. So, when a 50p rolls along the floor, the distance from the floor to the top of the coin is always equal. A book lying atop a set of 50p rollers therefore stays at the same height.

Surprisingly, there are many, many curves of constant width. The simplest is the Reuleaux triangle. It is constructed from an equilateral triangle by putting the point of a compass at each vertex and drawing the arc that goes between the other two vertices. In the diagram above, put the compass at A and move the pencil from B to C, and then repeat with the other vertices. The multilateral-curve heptagon is constructed the same way. Curves of constant width do not need to be symmetrical. It is possible to construct them from any number of lines crossing, as shown above right. The sections of perimeter are always arcs of a circle centred in the opposing vertex.

The curvilinear triangle owes its name to Franz Reuleaux, a German engineer who first wrote about its applications in his 1876 book *Kinematics of Machinery*. This was read many years later by H.G. Conway, a former president of the Institution of Mechanical Engineers, who sat on the British Treasury's Decimal Currency Board. Conway suggested a non-circular curve of constant breadth for the 50p piece since this property made it suitable for use in coin-operated machines. Machines distinguish coins by measuring diameter, and the 50p piece has the same width whichever position it is in. (A square coin, even with rounded sides, can never have constant width, which is why there are no four-sided coins.) Seven sides were chosen since it was considered the most aesthetically pleasing.

While the Reuleaux triangle reinvented the roller, it didn't reinvent the wheel. Wheels cannot be made of Reuleaux triangles because non-circular curves of constant width do not have a 'centre' – a fixed point that is equidistant from every point on the perimeter. If you put an axle on a Reuleaux triangle and rolled it, the height of the shape would stay the same but the axle would judder around.

One useful property of a Reuleaux triangle is that it can be rotated inside a square so that it touches all four sides of the square at all times. This property was exploited by Harry James Watts, an English engineer living in Pennsylvania, in 1914, when he designed one of the most bizarre tools in existence: a drill that can drill square holes. (The corners are rounded, rather than sharp, so, strictly speaking, the hole is a modified square.)

The cross-section of Watts's invention is simply a Reuleaux triangle, with three portions removed to make a cutting edge. It comes with a special chuck to compensate for the wobble of the drill bit's centre as it rotates. The Watts square-hole drill is still used today.

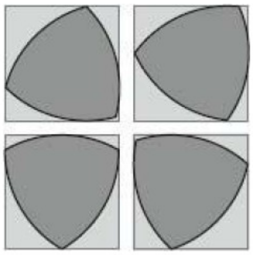

Reuleaux triangle rotating in square.

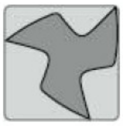

Cross-section of Watts's square drill.

(%)

20210920/ Space-Time/

- Alex's Adventures in Numberland.pdf

Pag. 154 - Life of Pi
until
Pag. 188

Obs: Pag. 184

Important:

Use este gráfico conjuntamente com Pappus' theorem

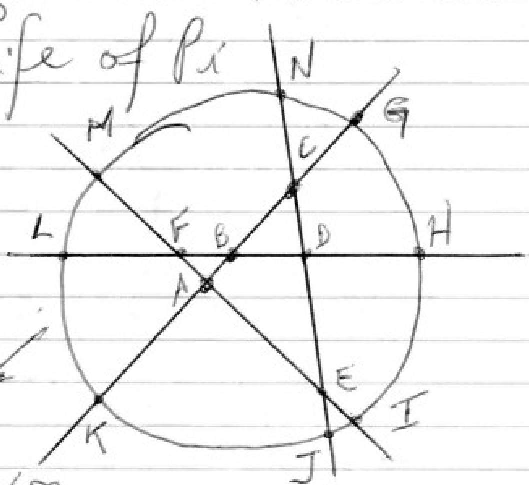

Retas → $\overline{ABC}$

$\overline{CDE}$

$\overline{EAF}$

$\overline{FBD}$

Obs: 5 pontos (reta com 5 pontos)

$\overline{KABEG}$

$\overline{NCDEJ}$

$\overline{HDBFL}$

$\overline{IEAFM}$

- The Math Book from Pythagoras to the 57th Dimension ...pdf

- The Physics Books from the Big Bang to Quantum Resurrection ... pdf

+,-.+ + +/ !+* 0 ! < =? 2 ! !
" H P " D 2 ! ! 3#

+,-.+ + +/ !+* 0 2 < =? <
<! .! - P " D 2 2 3#

+,-.+ + +/ !+* 0 2 -! 2 3! P 0 P
! P 3#
( ) " '

reentry as shown in the figure above. The recursion corresponding to the Koch eigenform is illustrated in the next figure (Fig. 8). Here we see the sequence of approximations leading to the infinite self-reflecting eigenform that is known as the Koch snowflake fractal.

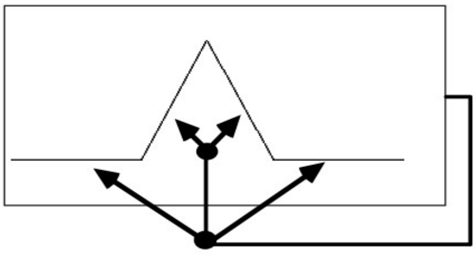

K = K { K K } K

Figure 7.

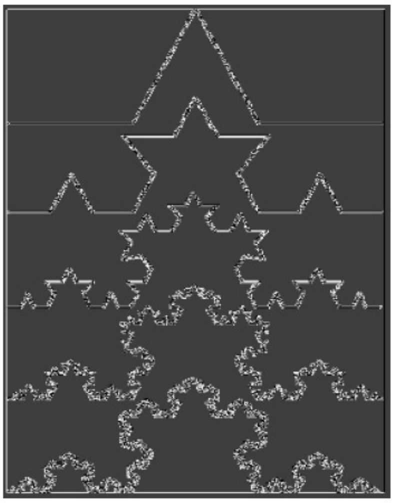

Figure 8.

Five stages of recursion are shown. To the eye, the last stage vividly illustrates how the ideal fractal form contains four copies of itself, each one-third the size of the whole. The abstract schema

K = K { K K } K

for this fractal can itself be iterated to produce a skeleton of the geometric recursion:

K = K { K K } K

= K { K K } K { K { K K } K K { K K } K } K { K K } K

= ...

We have only performed one line of this skeletal recursion. There are sixteen **K**'s in this second expression just as there are sixteen line segments in the second stage of the geometric recursion. Comparison with this symbolic recursion shows how geometry aids the intuition. The interaction of eigenforms with the geometry of physical, mental, symbolic and spiritual landscapes is an entire subject that is in need of deep exploration.

It is usually thought that the recognition of an object arises in some simple way from the assumed existence of the object and the action of our perceiving systems. This is a fine tuning to the point where the action of the perceiver and the perception of the object are indistinguishable. Such tuning requires an intermixing of the perceiver and the perceived that goes beyond description. Yet in the mathematical levels, such as number or fractal pattern, part of the process is slowed down to the point where we can begin to apprehend the process. There is a stability in the comparison, in the correspondence that is a process happening at once in the present time. The closed loop of perception occurs in the eternity of present individual time. Each such process depends upon linked and ongoing eigenbehaviors and yet is seen as simple by the perceiving mind. The perceiving mind is itself an eigenform.

3. The Square Root of Minus One and the Algebra of Iterants

The purpose of this section is to place **i**, square root of minus one, and its algebra in a process context. We then see that the square root of minus one is intimately related to time and the mathematical expression of time. The main point is that **i** can be interpreted as a primitive dynamical system that undergoes oscillation and nevertheless has a significant algebraic form associated with this iteration.

Traditionally **i**, the square root of minus one, arises from the fact that it is suggested by the solutions to quadratic equations. The simplest instance is the equation $\mathbf{x^2 + 1 = 0}$. A solution to this equation must have square equal to **-1**, but there are no real numbers whose square is **-1**. The square of a negative number is positive and the square of a positive number is positive and the square of zero is zero. Thus mathematicians (reluctantly at first) introduced an ideal number **i** such that $\mathbf{i^2 = -1}$. This number can be used to find solutions to other quadratic equations. For example **1+i** and **1-i** are the roots of $\mathbf{x^2 - 2x + 2 = 0}$.

The use of such complex numbers became indispensable when it was realized that they could be used in the solution of cubic equations in such a way that the real solution of a cubic was expressed naturally as a combination of two complex numbers such that the imaginary parts canceled out and all that was left was the real solution. Mathematicians began to seriously use complex numbers in the 1500's. It was not until around 1800 that Gauss and Argand discovered a geometric interpretation that opened up the subject to genuinely deep explorations.

The geometric interpretation of the complex numbers is well-known and we shall not repeat it here except to remark that **i** seen as a vector of unit length in the plane, and perpendicular to the horizontal axis. The horizontal axis is taken to be the real numbers. Multiplication by **i** is interpreted geometrically as rotation by **90** degrees. Thus **ii** is rotation by **180** degrees. Indeed the **180** degree rotate of **+1** is **-1**, and so **ii = -1**.

We now describe a process point of view for complex numbers. Think of the oscillatory process generated by

$$\mathbf{R(x) = -1/x.}$$

The fixed point would satisfy $\mathbf{i = -1/I}$ and multiplying, we get that $\mathbf{ii = -1}$. On the other hand the iteration of **R** on **1** yields $\mathbf{+1, R(1) = -1, R(R(1)) = +1, -1, +1, ...}$

Thus there must be a linkage between this ideal number **i** whose square is **-1** and the recursion that leads to an oscillation. The square root of minus one is a perfect example of an eigenform that occurs in a new and wider domain than the original context in which its recursive process arose. The process has no fixed point in the original domain.

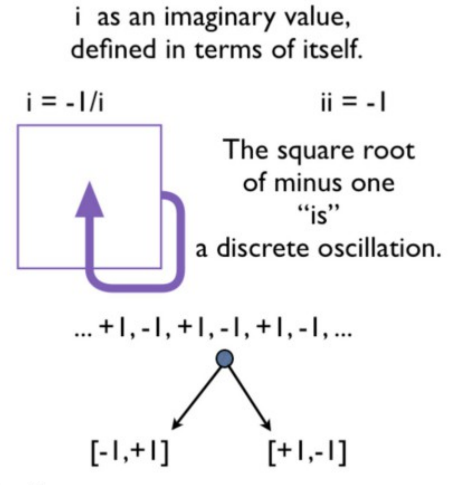

Figure 9.

Looking at the oscillation between **+1** and **-1**, we see that there are naturally two viewpoints that we denote by **[+1,-1]** and **[-1,+1]**. These viewpoints correspond to whether one regards the oscillation at time zero as starting with **+1** or with **-1**. We shall let **I{+1,-1}** stand for an undisclosed alternation or ambiguity between **+1** and **-1** and call **I{+1,-1}** an *iterant*. There are two *iterant views*: **[+1,-1]** and **[-1,+1]**.

Given an iterant **[a,b]**, we can think of **[b,a]** as the same process with a shift of one time step. These two iterant views, seen as points of view of an alternating process, will become the square roots of negative unity, **i** and **-i.** We introduce a *temporal shift operator* η such that

$$\mathbf{[a,b]}\eta = \eta\,\mathbf{[b,a]} \text{ and } \eta\,\eta = \mathbf{1}$$

for any iterant **[a,b]**, so that concatenated observations can include a time step of one-half period of the process **...abababab...** .

We combine iterant views term-by-term as in

$$\mathbf{[a,b][c,d] = [ac,bd]}.$$

We now define i by the equation, $\mathbf{i = [1,-1]}\eta$.

This makes **i** both a view of the iterant process and an operator that takes into account a step in time.

We calculate

$$\mathbf{ii = [1,-1]}\eta\,\mathbf{[1,-1]}\eta = \mathbf{[1,-1][-1,1]}\eta\,\eta = \mathbf{[-1,-1] = -1.}$$

Thus we have constructed the square root of minus one by using an iterant viewpoint. The key to rethinking the complex numbers in these terms is to understand that **i** represents a discrete oscillating temporal process and it is an eigenform participating in the algebraic structure of the complex numbers.

The Temporal Nexus

*We take as a matter of principle that the usual real variable **t** for time is better represented as **it** so that time is seen to be a process, an observation and a magnitude all at once.* This principle of imaginary time is justified by the eigenform approach to the structure of time and the structure of the square root of minus one. An example of the use of the Temporal Nexus, consider the expression $\mathbf{x^2 + y^2 + z^2 + t^2}$, the square of the Euclidean distance of a point **(x,y,z,t)** from the origin in Euclidean four-dimensional space. Now replace **t** by **it**, and find

$$\mathbf{x^2 + y^2 + z^2 + (it)^2 = x^2 + y^2 + z^2 - t^2,}$$

the squared distance in hyperbolic metric for special relativity. By replacing **t** by its process operator value **it** we make the transition to the physical mathematics of special relativity. We will discuss this theme further and relate the Temporal Nexus, the role of complex numbers in quantum mechanics and the role of temporal shift operators in discrete physics.

4. Quantum Physics, Eigenvalue, and Eigenform

Quantum is inextricably related to the complex numbers and in particular to the square root of minus one. We wish to show that the process view of the square root of minus one and the Temporal Nexus inform the epistemology of quantum theory. The reader should recall the Temporal Nexus from the previous section. *We take as a matter of principle that the usual real variable* ***t*** *for time is better represented as* ***it*** *so that time is seen to be a process, an observation and a magnitude all at once.* This principle of imaginary time is justified by our approach to the structure of time and the structure of the square root of minus one.

In quantum physics [11], the state of a physical system is modeled by a vector in a high-dimensional space, called a Hilbert space. As time goes on the vector rotates in this high dimensional space. Observable quantities correspond to Hermitian operators H on this vector space. Linearity is usually a simplifying assumption in mathematical models, but it is an essential feature of quantum mechanics. *In quantum mechanics observation is founded on the production of eigenvectors,* ***v*** *with* ***Gv = v*** *where* ***v*** *is a vector in a Hilbert space and* ***G*** *is a linear operator on that space.*

Many of the strange and fascinating properties of quantum mechanics emanate directly from this model of observation. In order to observe a quantum state, its vector is projected into an eigenvector for that particular mode of observation. By projecting the vector into that mode and not another, one manages to make the observation, but at the cost of losing information about the other possibilities inherent in the vector. This is the source, in the mathematical model, of the complementarities that allow exact determination of the position of a particle at the expense of nearly complete uncertainty about its momentum (or vice versa the determination of momentum at the expense of knowledge of the position).

Observation and quantum evolution (the determinate rotation of the state vector in the high dimensional Hilbert space) are interlocked. Each observation discontinuously projects the state vector to an eigenvector. The intervals between observations allow the continuous evolution of the state vector. This tapestry of interaction of the continuous and the discrete is the basis for the quantum mechanical description of the world. Is the quantum model, in its details, a consequence of general principles about systems? This is an exploration that needs to be made. We can only ask the question here. But the mysteries of the interpretation of quantum mechanics all hinge on an assumption of a world external to the quantum language. Thinking in terms of eigenform we can begin to look at how the physics of objects emerges from the model itself.

Where are the eigenforms in quantum physics? They are in the mathematics itself. For example, we have the simplest wave-function

$$\psi(x,t) = e^{i(kx - wt)}.$$

Since we know that the function $E(x) = e^x$ is an eigenform for operation of differentiation with respect to x, $\psi(x,t)$ is a special multiple eigenform from which the energy can be extracted by temporal differentiation, and the momentum can be extracted by spatial differentiation. We see in $\psi(x,t)$ the complexity of an individual who presents many possible sides to the world. $\psi(x,t)$ is an eigenform for more than one operator and it is a composition of e^x and **i**, the square root on minus one, a primordial eigenvalue related to time. It is this internal complexity that is mirrored in the uncertainty relations of Heisenberg and the complementarity of Bohr. The eigenforms themselves, as wavefunctions, are *inside* the mathematical model, on the *other side* of that which can be observed by the physicist.

We have seen eigenforms as the constructs of the observer, and in that sense they are on the side of the observer, even if the process that generates them is outside the realm of his perception. This suggests that we think again about the nature of the wave function in quantum mechanics. Is it also a construct of the observer? To see quantum mechanics and the world in terms of eigenforms requires a turning around, a shift of perception where indeed we shall find that the distinction between model and reality has disappeared into the world of appearance.

This is a reversal of epistemology, a complete turning of the world upside down. Eigenform has tricked us into considering the world of our experience and we find that it is our world, generated by our actions. The world has become objective through the self-generated stabilities of those actions.

A Quick Review of Quantum Mechanics

De Broglie hypothesized two fundamental relationships: between energy and frequency, and between momentum and wave number. These relationships are summarized in the equations

$$E = \hbar\omega,$$

$$p = \hbar k,$$

where **E** denotes the energy associated with a wave and **p** denotes the momentum associated with the wave**.** Here $\hbar = h/2\pi$ where **h** is Planck's constant. Schrödinger answered the question: *Where is the wave equation for De Broglie's waves?* Writing an elementary wave in complex form

$$\psi = \psi(x,t) = \exp(i(kx - wt)),$$

we see that we can extract De Broglie's energy and momentum by differentiating:

$$i\hbar\partial\psi/\partial t = E\psi \quad \text{and} \quad -i\hbar\partial\psi/\partial x = p\psi\,.$$

This led Schrödinger to postulate *the identification of dynamical variables with operators* so that the first equation, $i\hbar\partial\psi/\partial t = E\psi$, is promoted to the status of an equation of motion while the second equation becomes the definition of momentum as an operator:

$$p = -ih\partial/\partial x\,.$$

Once **p** is identified as an operator, the numerical value of momentum is associated with an eigenvalue of this operator, just as in the example above. In our example $py = \hbar k\psi$**.** In this formulation, the position operator is just multiplication by **x** itself. Once we have fixed specific operators for position and momentum, the operators for other physical quantities can be expressed in terms of them. We obtain the energy operator by substitution of the momentum operator in the classical formula for the energy:

$$E = (1/2)mv^2 + V$$
$$E = p^2/2m + V$$
$$E = -(\hbar^2/2m)\partial^2/\partial x^2 + V.$$

Here **V** is the potential energy, and its corresponding operator depends upon the details of the application. With this operator identification for **E**, Schrödinger's equation

$$i\hbar\partial\psi/\partial t = -(\hbar^2/2m)\partial^2\psi/\partial x^2 + V\psi$$

or equivalently,

$$\partial y/\partial t = i(\hbar/2m)\partial^2\psi/\partial x^2 - (i/\hbar)V\psi$$

is an equation in the first derivatives of time and in second derivatives of space. In this form of the theory one considers general solutions to the differential equation and this in turn leads to excellent results in a myriad of applications. It is useful to point out that Schrödinger's equation without the potential term has the form

$$\partial y/\partial t = i(\hbar/2m)\partial^2\psi/\partial x^2.$$

It is in this form that we shall compare it with discrete processes in the next section. There we shall use the Temporal Nexus to obtain new insight into the role of **i** in this equation.

In quantum theory, observation is modeled by the concept of eigenvalues for corresponding operators. *The quantum model of an observation is a projection of the wave function into an eigenstate.* An energy spectrum, **{Ek}** corresponds to wave functions, **y** satisfying the Schrödinger equation, such that there are constants Ek with $E\psi = Ek\psi$. An *observable* (such as energy) **E** is a Hermitian operator on a Hilbert space of wavefunctions. Since Hermitian operators have real eigenvalues, this provides the link with measurement for the quantum theory.

It is important to notice that there is no mechanism postulated in this theory for how a wave function is sent into an eigenstate by an observable. Just as mathematical logic need not demand causality behind an implication between propositions, the logic of quantum mechanics does not demand a specified cause behind an observation. This absence of an assumption of causality in logic does not obviate the possibility of causality in the world. Similarly, the absence of causality in quantum observation does not obviate causality in the physical world. Nevertheless, the debate over the interpretation of quantum theory has often led its participants into asserting that causality has been demolished in physics.

Note that the operators for position and momentum satisfy the equation $xp - px = \hbar i$. This corresponds directly to the equation obtained by Heisenberg, on other grounds, that dynamical variables can no longer necessarily commute with one another. In this way, the points of view of De Broglie, Schrödinger and Heisenberg came together, and quantum mechanics was born. In the course of this development, interpretations varied widely. Eventually, physicists came to regard the wave function not as a generalized wave packet, but as a carrier of information about possible observations. In this way of thinking y^*y (y^* denotes the complex conjugate of **y**) represents the probability of finding the particle (A particle is an observable with local spatial characteristics.) at a given point in spacetime. Strictly speaking, it is the spatial integral of $\psi^*\psi$ that is interpreted as a total probability with $\psi^* \psi$ the probability density. This way of thinking is supported by the fact that the total spatial integral is time-invariant as a consequence of Schrödinger's equation!

5. Iterants, Complex Numbers, and Quantum Mechanics

We have seen that there are indeed eigenforms in quantum mechanics. *The eigenforms in quantum mechanics are the mathematical functions such as* e^x *that are invariant under operators such as* $D = d/dx$.

But we wish to examine the relationship between recursion, reflexive spaces and the properties of the quantum world. The hint we have received from quantum theory is that we should begin with mathematics which is replete with eigenforms. In fact, this hint seems very rich when we consider that **i**, the square root of minus one, is a key eigenform in our panoply of eigenforms and it is a key ingredient in quantum mechanics intimately related to the role of time.

Revisiting the Temporal Nexus

If we define

$$R(A) = -1/A$$

then **R(i) = i** since $i^2 = -1$ is equivalent to **i = -1/i.** Using the infinite recursion we would then write

$$i = -1/-1/-1/-1/-1/... ,$$

making **i** an infinite reentry form for the operator **R.** We will write

$$i = [-1/*]$$

where * denotes the reentry of the whole form into that place in the right-hand part of the expression. Similarly, if **F(x) = 1 + 1/x,** then the eigenform would be **[1 + 1/*]** and we could write $(1+\sqrt{5})/2 = [1 + 1/*]$. With this in place we can now consider wave functions in quantum mechanics such as

$$y(x,t) = \exp(i(kx - wt)) = \exp([-1/*] (kx - wt))$$

and we can consider classical formulas in mathematics such as *Euler's formula*

$$\exp([-1/*]j) = \cos(j) + [-1/*] \sin(j)$$

in this light. We start here with Euler's formula, for this formula is the key relation between complex numbers, **i**, waves and periodicity.

We return to the finite nature of **[-1/*]**. This eigenform is an oscillator between **-1** and **+1**. It is only **i** in its idealization. In its appropriate synchronization it has the property that **i = -1/i**. As a real oscillator, the equation **R(i) = -1/i** tells us that when **i** is **1**, then i is transformed to **-1** and when **i** is **-1** then **i** is transformed to **+1**. There is no fixed point for **R** in the real domain. The eigenform is achieved by leaving the real domain for a new and larger domain.

We know that this larger domain can be conceptualized as the plane with Euclidean rotational geometry, but here we explore the larger domain in terms of eigenforms. *We find that* **i** *itself is a fundamental discrete process, and it is in the microworld of such discrete physical processes that not only quantum mechanics, but also classical mechanics is born.*

Iterants and Iterant Views

In order to think about i, consider an infinite oscillation between **+1** and **-1**:

... -1,+1,-1,+1,-1,+1,-1,+1,...

This oscillation can be seen in two distinct ways. It can be seen as **[-1,+1]** (a repetition in this order) or as **[+1,-1]** (a repetition in the opposite order). This suggests regarding an infinite alternation such as

... a,b,a,b,a,b,a,b,a,b,a,b,a,b,...

as an entity that can be seen in two possible ways, indicated by the ordered pairs **[a,b]** and **[c,d]**. We shall call the infinite alternation of **a** and **b** the *iterant* of **a** and **b** and denote it by **I{a,b}**. Just as with a set **{a,b}**, the iterant is independent of the order of **a** and **b**.

We have **I{a,b} = I{b,a}**, but there are two distinct views of any iterant and these are denoted by **[a,b]** and **[b,a]**. The key to iterants is that two representatives of an iterant can *by themselves* appear identical, but *taken together* are seen to be different. For example, consider

... a,b,a,b,a,b,a,b,a,b,a,b,a,b,...

and also consider

...b,a,b,a,b,a,b,a,b,a,b,a,b,a,...

There is no way to tell the difference between these two iterants except by a direct comparison as shown below

... a,b,a,b,a,b,a,b,a,b,a,b,a,b,...
... b,a,b,a,b,a,b,a,b,a,b,a,b,a,...

In the direct comparison we see that if one of them is **[a,b]**, then the other one should be **[b,a]**. Still, there

(%

- The Physics of Reality Space, Time, Matter, Cosmos... pdf

  Pág. 06(33) 3.

  The Square Root of Minus One and the Algebra of Iterants.

- Symmetries and Group Theory in Particle Physics... pdf

  Pág. 61(76) - Chapter 4

  The Poincaré transformations

→ We examine the general properties of the Poincaré group, i.e., the group of inhomogeneous Lorentz transformations, which are space-time coordinate transformations between any two inertial frames of reference.

Then we consider those unitary irreducible representations of this group which are the most suitable for the description of the quantum mechanical states of one or more particles.

—

- Astronautics the Physics of Space Flight... pdf
- Basics of Space Flight... pdf

—

+,-.+ + +/ ! + * 2 ! # 1 3 0 2
! 2 , & # 3! 0 ! # & ! 2 3#
( )) !3 * 0 ! E 0 ! !

4

The Poincaré transformations

In this Chapter, we examine the general properties of the Poincaré group, i.e. the group of inhomogeneous Lorentz transformations, which are space-time coordinate transformations between any two inertial frames of reference. Then we consider those unitary irreducible representations of this group which are the most suitable for the description of the quantum mechanical states of one or more particles.

It is well known from special relativity that all inertial frames of reference are completely equivalent for the description of the physical phenomena; in fact, the physical laws are the same in all inertial frame or, in other words, they are *invariant* under the Poincaré transformations. The importance of the Poincaré group is related to this invariance principle. We refer to Section 9.2 for a brief discussion on *invariance principles*, symmetries and conservation laws.

4.1 Group properties

The inhomogeneous Lorentz transformations, or Poincaré transformations, connect the space-time coordinates of any two frames of reference whose relative velocity is constant.

A general Poincaré transformation can be written, as a generalization of a homogeneous Lorentz transformation (3.1), in the form

$$x'^{\mu} = \Lambda^{\mu}{}_{\nu} x^{\nu} + a^{\mu} \,, \tag{4.1}$$

where a^{μ} $(\mu = 0, 1, 2, 3)$ stand for the components of a vector in $\mathcal{R}^4$. We shall denote the above transformation by (a, Λ).

It is easy to see that the transformations (a, Λ) form a group: the so-called *Poincaré group*, denoted in the following by $\mathcal{P}$. In particular, the application of two successive transformations (a_1, Λ_1), (a_2, Λ_2) gives

$$x' = \Lambda_1 x + a_1 \,, \qquad x'' = \Lambda_2 x' + a_2 = \Lambda_2 \Lambda_1 x + \Lambda_2 a_1 + a_2 \,, \tag{4.2}$$

G. Costa and G. Fogli, *Symmetries and Group Theory in Particle Physics*,
Lecture Notes in Physics 823, DOI: 10.1007/978-3-642-15482-9_4,

© Springer-Verlag Berlin Heidelberg 2012

so that the product of the two transformations is

$$(a_2, \Lambda_2)(a_1 \Lambda_1) = (a_2 + \Lambda_2 a_1, \Lambda_2 \Lambda_1) \ . \tag{4.3}$$

Clearly the unit element is $(0, I)$; then from Eq. (4.3) one gets the inverse element

$$(a, \Lambda)^{-1} = (-\Lambda^{-1} a, \Lambda^{-1}) \ . \tag{4.4}$$

The group $\mathcal{P}$ is characterized by 10 parameters: the 6 parameters of the homogeneous Lorentz group plus the 4 parameters a^μ.

Obviously, the Poincaré group contains, as a subgroup, the homogeneous Lorentz group $\mathcal{L}$ of the transformations $(0, \Lambda)$; it contains also the subgroup $\mathcal{S}$ of the four dimensional translations(a, I), which is an Abelian invariant subgroup of $\mathcal{P}$. Any Poincaré transformation can be written in a unique way as the product of a pure translation and a homogeneous Lorentz transformation

$$(a, \Lambda) = (a, I)(0, \Lambda) \ , \tag{4.5}$$

$$(a, \Lambda) = (0, \Lambda)(\Lambda^{-1} a, I) \ . \tag{4.6}$$

This shows that the Poincaré group is the semi-direct product of the two subgroups:

$$\mathcal{P} = \mathcal{S} \ \circledS \ \mathcal{L} \ . \tag{4.7}$$

We note that the order of the transformations in Eqs. (4.5), (4.6) is important, since the translations do not commute with the Lorentz transformations. From the above relations one gets easily the useful result

$$(0, \Lambda^{-1})(a, I)(0, \Lambda) = (\Lambda^{-1} a, I) \ . \tag{4.8}$$

We know that the group $\mathcal{L}$ consists of four disjoint components $\mathcal{L}_+^\uparrow$, $\mathcal{L}_-^\uparrow$, $\mathcal{L}_+^\downarrow$, $\mathcal{L}_-^\downarrow$, corresponding to the possible choices of the signs of $\det \Lambda$ and $\Lambda^0{}_0$ listed in Table 3.1. Similarly, the group $\mathcal{P}$ consists of four disjoint components $\mathcal{P}_+^\uparrow$, $\mathcal{P}_-^\uparrow$, $\mathcal{P}_+^\downarrow$, $\mathcal{P}_-^\downarrow$, each of which contains the corresponding component of the subgroup $\mathcal{L}$.

In this Chapter, we shall limit ourselves to the transformations of $\mathcal{P}_+^\uparrow = \mathcal{S} \ \circledS \ \mathcal{L}_+^\uparrow$, which form the proper orthochronous inhomogeneous Lorentz group. We briefly mention the topological properties of $\mathcal{P}_+^\uparrow$, which are easily derived from those of its subgroups:

a) The group $\mathcal{P}_+^\uparrow$ is *non-compact*; in fact, it contains the subgroups $\mathcal{L}_+^\uparrow$ and $\mathcal{S}$ which are both non-compact (the translation group $\mathcal{S}$ is clearly non-compact, since $\mathcal{R}^4$ is not compact);
b) The group $\mathcal{P}_+^\uparrow$ is *doubly-connected*: it has the same connectedness of its subgroup $\mathcal{L}_+^\uparrow$ (specifically of $SO(3)$). Its *universal covering group* is defined by the transformations (a, A), where A is an element of $SL(2, C)$;
c) The group $\mathcal{P}_+^\uparrow$ is neither simple nor semi-simple, in fact, the translation group $\mathcal{S}$ is an invariant subgroup of $\mathcal{P}_+^\uparrow$.

Next we consider the Lie algebra of the group $\mathcal{P}_+^\uparrow$; its basic elements are the infinitesimal generators of $\mathcal{L}_+^\uparrow$ and $\mathcal{S}$.

In general, an infinitesimal translation is given by

$$(\delta a, I) = I - i\delta a_\mu P^\mu \ , \tag{4.9}$$

introducing four operators P^μ, which are the infinitesimal generators of the translations. A finite translation is given by exponentiation, as

$$(a, I) = e^{-ia_\mu P^\mu} \ . \tag{4.10}$$

The commutation relations of the infinitesimal generators are easily obtained giving a specific representation of the Lie algebra of $\mathcal{P}_+^\uparrow$. For this purpose it is convenient to write (a, Λ) as a 5×5 matrix

$$\begin{pmatrix} \Lambda & a \\ 0 & 1 \end{pmatrix} , \tag{4.11}$$

where the four vector a is written as a column matrix. Eq. (4.2) is then obtained by matrix product, and Eq. (4.1) can be considered as the transformation of a vector

$$y = \begin{pmatrix} x \\ 1 \end{pmatrix} \tag{4.12}$$

in a five-dimensional space. In fact, it can be re-written as

$$\begin{pmatrix} x' \\ 1 \end{pmatrix} = \begin{pmatrix} \Lambda & a \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix} = \begin{pmatrix} \Lambda x + a \\ 1 \end{pmatrix} . \tag{4.13}$$

Expressing in similar way the infinitesimal transformation (4.9), one can obtain explicitly

$$P_0 = \begin{pmatrix} 0 & 0 & 0 & 0 & i \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} , \qquad P_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & i \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} , \tag{4.14}$$

and similarly for P_2, P_3.

In the same representation, the generators J_i and K_i are replaced by 5×5 matrices, obtained from (3.50) and (3.54) by adding a fifth row and a fifth column of zeros.

One can check that the following commutation relations hold:

$$\begin{aligned} [P_\mu, P_\nu] &= 0 \ , \\ [J_i \ , P_0] &= 0 \ , \\ [J_i \ , P_i\,] &= 0 \ , \\ [J_i \ , P_j] &= i\epsilon_{ijk} P_k \ , \\ [K_i, P_0] &= -iP_i \ , \\ [K_i, P_i\,] &= -iP_0 \ , \\ [K_i, P_j] &= 0 \ . \end{aligned} \tag{4.15}$$

In terms of the antisymmetric tensor $M_{\mu\nu}$ given by (3.59), the above commutators become:

$$\begin{aligned} [P_\mu, P_\nu] &= 0 \,, \qquad (4.16)\\ [M_{\mu\nu}, P_\rho] &= -i(g_{\mu\rho}P_\nu - g_{\nu\rho}P_\mu) \,, \qquad (4.17) \end{aligned}$$

and, together with the commutators

$$[M_{\lambda\rho}, M_{\mu\nu}] = -i(g_{\lambda\mu}M_{\rho\nu} + g_{\rho\nu}M_{\lambda\mu} - g_{\lambda\nu}M_{\rho\mu} - g_{\rho\mu}M_{\lambda\nu}) \,, \qquad (3.60)$$

they define the Lie algebra of $\mathcal{P}_+^\uparrow$.

One can check that the operator

$$P^2 = P_\mu P^\mu \qquad (4.18)$$

commutes with all the generators P_μ, $M_{\mu\nu}$ of the Lie algebra; hence P^2 is an invariant operator under the Poincaré transformations.

Let us introduce the operator

$$W_\mu = \epsilon_{\mu\nu\sigma\tau} M^{\nu\sigma} P^\tau \,; \qquad (4.19)$$

it follows immediately that

$$W_\mu P^\mu = 0 \,. \qquad (4.20)$$

Moreover, making use of Eqs. (3.60), (4.17), one can obtain the following commutators

$$\begin{aligned} [P_\mu, W_\nu] &= 0 \,, \\ [M_{\mu\nu}, W_\sigma] &= -i(g_{\nu\sigma}W_\mu - g_{\mu\sigma}W_\nu) \,, \qquad (4.21)\\ [W_\mu, W_\nu] &= i\epsilon_{\mu\nu\sigma\tau} W^\sigma P^\tau \,. \end{aligned}$$

Finally, one can prove that the operator

$$W^2 = W_\mu W^\mu \qquad (4.22)$$

commutes with all the P_μ and $M_{\mu\nu}$ and hence it is an invariant operator.

The eigenvalues of the invariant operators P^2 and W^2 are used to label the IR's of the group; for some classes of IR's one can have additional invariants.

4.2 Unitary representations of the proper orthochronous Poincaré group

In the previous Section, we have exhibited a 5-dimensional representation of the proper orthochronous Poincaré group $\mathcal{P}_+^\uparrow$, which is not unitary, since both the generators P_μ and K_i are expressed by non-hermitian matrices. In fact, as pointed out in Subsection 1.2.1, since the group $\mathcal{P}$ and its subgroup $\mathcal{P}_+^\uparrow$ are not compact, no finite-dimensional unitary IR's exists.

We recall from quantum mechanics[1] the well known fact that the infinitesimal generators of translations P_μ can be identified with the *energy-momentum operators*. Moreover, the infinitesimal generators $M_{\mu\nu}$ can be identified with the components of the *angular momentum tensor*. For physical applications, we are interested in those IR's in which the operators P_μ and $M_{\mu\nu}$ are *hermitian*, since they correspond to dynamical variables, i.e. in the *unitary*, and hence infinite-dimensional, IR's of the Poincaré group $\mathcal{P}$. Here we are mainly concerned with the IR's of the restricted group $\mathcal{P}_+^\uparrow$.

Before going to the IR's of $\mathcal{P}_+^\uparrow$, we want to review briefly some general properties of the invariance principles for a quantum mechanical system. Each state of the system is represented by a vector $|\Phi>$ in a Hilbert space. The state vectors are normalized to unity, to allow the probability interpretation; two vectors that differ only by a phase represent the same physical state. In fact, the results of experiments are expressed in terms of transition probabilities

$$|<\Psi|\Phi>|^2 \tag{4.23}$$

between the states $|\Phi>$ and $|\Psi>$, which depend only on the absolute value of the scalar product between the two states. This shows that each vector in the set $e^{i\alpha}|\Phi>$, where α is a real number, corresponds to the same physical state. The collection of all the state vectors of the form $e^{i\alpha}|\Phi>$ is called a *ray*, which we denote simply by Φ.

An *invariance principle*, or *symmetry operation*, of a physical system is a one-to-one correspondence between two rays Φ and Φ' (and similarly between Ψ and Ψ'), representing physically realizable states, such that the transition probabilities remain unchanged:

$$|<\Psi'|\Phi'>|^2 = |<\Psi|\Phi>|^2 \ . \tag{4.24}$$

An important theorem due to Wigner[2] states that, if $\Phi \to \Phi'$ is a symmetry operation of a physical theory, there exists a *linear unitary* or *antilinear unitary*[3] (antiunitary) operator U, determined up to a phase, such that:

$$\Phi' = U\Phi \ . \tag{4.25}$$

---

[1] See e.g. S. Weinberg, *The Quantum Theory of Fields, Vol. I, Foundations*, Cambridge University Press (1995).

[2] See e.g. F.R. Halpern, *Special Relativity and Quantum Mechanics*, Prentice-Hall (1968).

[3] An antilinear unitary operator U is defined by

$$\begin{aligned} <U\Psi|U\Phi> &= <\Psi|\Phi>^*, \\ U|\alpha\Psi + \beta\Phi> &= \alpha^* U|\Psi> + \beta^* U|\Phi> \ . \end{aligned}$$

In particular, if a physical system is invariant under the transformations of the Poincaré group, to each element (a, Λ) of $\mathcal{P}$ there corresponds a linear operator $U(a, \Lambda)$ which transforms the state Φ into

$$\Phi' = U(a, \Lambda)\,\Phi\ , \tag{4.26}$$

leaving Eq. (4.24) invariant. The operator $U(a, \Lambda)$ is determined up to a phase; however, for those Poincaré transformations which are connected with the identity, such as the transformations of the subgroup $\mathcal{P}_+^\uparrow$, it can be shown that it is possible to choose a phase convention such that the product of any two elements of $\mathcal{P}_+^\uparrow$ is always given by[4]:

$$U(a_1, \Lambda_1)U(a_2, \Lambda_2) = \pm U(a_3, \Lambda_3)\ . \tag{4.27}$$

The sign ambiguity is related to the fact that the group $\mathcal{P}$ is doubly connected, and it can have double-valued representations.

We stress the fact that the operators $U(a, \Lambda)$ of $\mathcal{P}_+^\uparrow$ are *unitary*. In fact, for any element (a, Λ) of $\mathcal{P}_+^\uparrow$ one can find an element (a', Λ') such that

$$(a, \Lambda) = (a', \Lambda')(a', \Lambda')\ . \tag{4.28}$$

Then, from Eq. (4.27):

$$U(a, \Lambda) = \pm U(a', \Lambda')U(a', \Lambda')\ . \tag{4.29}$$

According to Wigner's theorem, $U(a', \Lambda')$ could be either unitary or antiunitary, but in both cases, since the square of an antiunitary operator is unitary, $U(a, \Lambda)$ can only be unitary. Since $U(a, \Lambda)$ is an arbitrary element of $\mathcal{P}_+^\uparrow$, it follows that all the linear operator $U(a, \Lambda)$ are unitary.

We have thus obtained a representation of the group $\mathcal{P}_+^\uparrow$ in the Hilbert space of the state vectors $|\Phi>$ in terms of the unitary operators $U(a, \Lambda)$.

The unitary representation realized by the operators $U(a, \Lambda)$ is, of course, *infinite-dimensional*, since the Hilbert space of the state vectors $|\Phi>$ is infinite-dimensional, but it is, in general, reducible. To obtain the *unitary irreducible representations*, one has to find the invariant subspaces, i.e. invariant under the transformations (a, Λ).

To make this point clear, let us consider a given set of vectors $|\Phi>$ in the Hilbert space. Since the P_μ commute among themselves (see eq. (4.16)), it would be convenient to start from the vectors $|p, \zeta>$, labelled by the eigenvalues p_μ of P_μ and by other, for the moment unspecified, quantum number ζ:

$$P_\mu|p, \zeta> = p_\mu|p, \zeta>\ . \tag{4.30}$$

Now we are facing the following problem: the vectors $|p, \zeta>$ correponding to particular points p_μ in the momentum space are not normalizable, so that they do not belong to the Hilbert space. One can obtain normalizable vectors by convenient linear superposition of $|p,\zeta>$[5]. However, for the sake of simplicity, we shall adopt the normalization in terms of the Dirac delta function, widely used in the physics literature[6] and, with the inclusion of the convenient factor $2p_0$, define:

---

[4] F.R. Halpern, quoted ref.

$$<p',\zeta'|p,\zeta> = 2p_0\,\delta^3(\mathbf{p}'-\mathbf{p})\delta_{\zeta'\zeta}\;. \tag{4.31}$$

For a pure translation, from Eq. (4.10), we get

$$U(a,I)\,|p,\zeta> = e^{-ia_\mu p^\mu}|p,\zeta>\;, \tag{4.32}$$

while, for a homogeneous Lorentz transformation $(0,\Lambda)$, we obtain

$$U(0,\Lambda)\,|p,\zeta> = \sum_{\zeta'} Q_{\zeta'\zeta}|p',\zeta'>\;, \tag{4.33}$$

where Q is a unitary matrix and

$$p'^\mu = \Lambda^\mu{}_\nu p^\nu\;. \tag{4.34}$$

Eq. (4.34) is obtained from eq. (4.30) and from the relation

$$U^{-1}(0,\Lambda)P^\mu U(0,\Lambda) = \Lambda^\mu{}_\nu P^\nu\;, \tag{4.35}$$

which can be derived from Eqs. (4.8) and (4.9). This shows that the four operators P^μ transform as the components of a four-vector.

Once more we see that $P^2 = P_\mu P^\mu$ is invariant; in fact, we know that P^2 is one of the Casimir operators of the Poincaré group. Its eigenvalues p^2, when $p^2 = m^2 > 0$, correspond to the squared total energy in the c. m. system. If the physical system is a particle, m is the *mass of the particle*.

The subspace spanned by the vectors $|p,\zeta>$ with a given value of p^2 is then *invariant* under the transformations (a,Λ), so that the representation $U(0,\Lambda)$ given by (4.33) is, in general, reducible and can be decomposed into an (infinite) direct sum of representations corresponding to different eigenvalues m^2.

To reduce further the representation, we need to consider also the operator W_μ (Eq. (4.19)), whose square W^2 is the other Casimir operator of the Poincaré group. Since the components W_μ commute with P_μ but not among themselves (Eq. (4.21)), we can diagonalize W^2 and *one* component, say W_3. Among the quantum numbers ζ labelling the eigenstates $|p,\zeta>$ one can then single out the eigenvalues w_3 of W_3. For the sake of simplicity, we shall neglect in the following the other unspecified quantum numbers which should be added to make the set of observables complete.

---

[5] See, e.g., A. Messiah, *Quantum Mechanics*, North Holland (1962).

[6] See e.g. M.E. Peskin, D.V. Schroeder, *An Introduction to Quantum Field Theory*, Addison-Wesley (1995); S. Weinberg, *The Quantum Theory of Fields, Vol. I, Foundations*, Cambridge University Press (1995)

Let us first consider those representations for which $p^2 = m^2 > 0$. In that case $p_0/|p_0|$, i.e. the sign of the energy, is also an invariant of the group $\mathcal{P}_+^\uparrow$. In the four-momentum space, the eigenstates of $p^2 = m^2 > 0$ with $p_0 > 0$, which correspond to physical states, are represented by the points in the upper branch of the hyperboloid shown in Fig. 4.1. Under a transformation of $\mathcal{P}_+^\uparrow$ the representative point moves on the *same* branch of the hyperboloid.

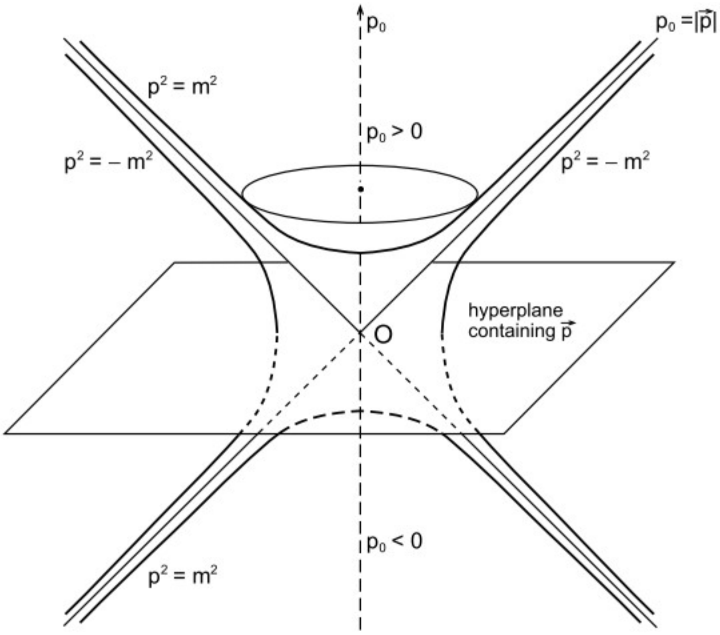

Fig. 4.1. *Hyperboloid $p^2 = m^2$ in the four-momentum space.*

An irreducible representation is characterized by the eigenvalues of P^2, W^2 and the sign of $p_0/|p_0|$; an IR with $p_0 > 0$ is then spanned by the vectors $|p, w_3>$ with $p^2 = m^2 = \text{const.}$, and $p_0 = +\sqrt{\mathbf{p}^2 + m^2}$.

The physical meaning of W^μ is made clear by considering explicitly its components. Using the definitions (3.59), one gets from (4.19)

$$W_0 = \mathbf{P} \cdot \mathbf{J}\,, \tag{4.36}$$

$$\mathbf{W} = P_0\, \mathbf{J} - \mathbf{P} \times \mathbf{K}\,. \tag{4.37}$$

It is convenient to go to the rest frame, in which $\mathbf{p} = 0$. In this frame Eqs. (4.36), (4.37) are equivalent to

$$W^\mu = m(0, J^1, J^2, J^3)\,, \tag{4.38}$$

i.e. W^μ reduces essentially to the components of the total angular momentum $\mathbf{J}$, which is the *spin* in the case of a particle. A rest frame can then be specified by $|\tilde{p}, j_3\rangle$, where $\tilde{p} = (m, \mathbf{0})$. According to Eq. (4.38), one gets

$$\begin{aligned} W_3|\tilde{p}, j_3\rangle &= mj_3|\tilde{p}, j_3\rangle \ , \\ (W)^2|\tilde{p}, j_3\rangle &= -m^2 j(j+1)|\tilde{p}, j_3\rangle \ . \end{aligned} \tag{4.39}$$

If $p^2 = 0$, one cannot apply the above arguments, since one cannot go to the rest frame. Since we are interested in the physical case of a massless particle, we take also $w^2 = 0$ ($w^2 \neq 0$ would represent continuous spin values, as we will see in Section 5.3). However, these two null quantum numbers are not sufficient to characterize an IR. Taking into account also Eq. (4.20) we have: $p^2 = 0$, $w^2 = 0$ and $w_\mu p^\mu = 0$. These conditions are satisfied only if the two four-vectors w_μ and p_μ are parallel: $w_\mu = \lambda p_\mu$. In terms of operators we can write[7]:

$$W_\mu = \lambda P_\mu \ , \tag{4.40}$$

where λ is an *invariant*. The physical interpretation of λ becomes evident making use of Eq. (4.36) and of the relation $\mathbf{p}^2 = p_0^2$:

$$\lambda = \frac{W_0}{P_0} = \frac{\mathbf{P} \cdot \mathbf{J}}{P_0} = \frac{\mathbf{P} \cdot \mathbf{J}}{|\mathbf{P}|} \ . \tag{4.41}$$

This shows that λ represents the component of the spin along the direction of motion, which is usually called *helicity*. In the case $p^2 = 0$, we label the state vector with the eigenvalues of λ; i.e. by $|p, \lambda\rangle$.

The above examples show that the operator W_μ describes the polarization states of both massive and massless particles; it represent the covariant generalization of the spin, and it is called, in fact, *covariant spin*.

We have considered only two kinds of IR's of $\mathcal{P}_+^\uparrow$ of physical interest. To summarize and complete the preceding discussion, we list in the following all the IR's of $\mathcal{P}_+^\uparrow$[8]:

a) $p^2 = m^2 > 0$.
For each pair of values of p^2 and w^2, there are two IR's, one for each value of $p_0/|p_0|$. The vector basis of an IR is given by $|p, w_3\rangle$; in the rest frame it reduces to the $(2j+1)$-dimensional basis of the IR of the rotation group, which corresponds to the $(2j+1)$ states of polarization.

b) $p^2 = 0$, $p_\mu \neq 0$, $w^2 = 0$.
Each IR is characterized by the sign of p_0 and by the value (integer or half-integer) of the helicity λ. For a given momentum p_μ and given $\lambda \neq 0$, there exist two independent states (they belong to two independent IR's), which correspond to the two different helicities $\pm\lambda$.

---

[7] Usually we denote operators by capital letters (P_μ, W_μ...) and the corresponding eigenvalues by small letters (p_μ, w_μ...).

[8] I.M. Gel'fand, R.A. Minlos, Z.Ya. Shapiro, *Representations of the Rotations and Lorentz Groups and Their Applications*, Pergamon Press (1963); M.A. Naimark, *Linear Representations of the Lorentz Group*, Pergamon Press (1964).

c) $p^2 = 0$, $p_\mu \neq 0$, $w^2 < 0$.
The IR's are infinite-dimensional also in the spin variable and would correspond to particles with continuous spin.

d) $p^2 = 0$, $p_\mu = 0$, $w^2 < 0$.
The unitary IR's coincide with those of the homogeneous Lorentz group, and they will not be considered further (they provide a powerful tool for the study of the analytic properties of the forward scattering amplitude in the angular momentum variable).

e) $p^2 < 0$.
These IR's correspond to space-like vectors p_μ, so they cannot be used in the description of physical states (they can be useful, however, in the study of the analytic properties of the scattering amplitude).

Problems

4.1. Show that the Poincaré transformations (a, Λ) form a group and give the explicit proof that the translation group $\mathcal{S}$ is an invariant subgroup.

4.2. Prove that P^2 and W^2 are invariant operators of the group $\mathcal{P}_+^\uparrow$.

4.3. Derive the transformation properties of the generators P_μ under $\mathcal{L}_+^\uparrow$ and the commutation relations (4.17), making use of the general form

$$U(a, \Lambda) = e^{-ia_\mu P^\mu} e^{-\frac{1}{2}i\omega_{\mu\nu}M^{\mu\nu}}$$

of the unitary IR's of $\mathcal{P}_+^\uparrow$.

4.4. Derive the transformation properties of the generators $M^{\mu\nu}$ under $\mathcal{L}_+^\uparrow$ and the commutation relations (3.60) starting from the general form

$$U(0, \Lambda) = e^{-\frac{1}{2}i\omega_{\mu\nu}M^{\mu\nu}}$$

of the unitary IR's of $\mathcal{L}_+^\uparrow$.

4.5. Making use of the Lorentz transformation which brings the state $|p, \zeta>$ at rest, show that the commutators $[W_\mu, W_\nu] = i\epsilon_{\mu\nu\sigma\tau}W^\sigma P^\tau$ reduce to $[J_i, J_j] = i\epsilon_{ijk}J_k$.

4.6. Show that the Lorentz transformations of the form

$$U(0, \Lambda) = e^{-in_\mu W^\mu} ,$$

where n is an arbitrary four-vector, leave the eigenvalues p_μ of P_μ invariant.

(%

- The Physics of Reality Space, Time, Matter, Cosmos... pdf

  Pág. 06(33) 3.

  The Square Root of Minus One and the Algebra of Iterants.

- Symmetries and Group Theory in Particle Physics... pdf

  Pág. 61(76) - Chapter 4

  The Poincaré transformations

→ We examine the general properties of the Poincaré group, i.e., the group of inhomogeneous Lorentz transformations, which are space-time coordinate transformations between any two inertial frames of reference.

Then we consider those unitary irreducible representations of this group which are the most suitable for the description of the quantum mechanical states of one or more particles.

- Astronautics the Physics of Space Flight... pdf
- Basics of Space Flight... pdf

3 +,-.+ + +/ ! + * , ! 0 2 3!
? 3#
B +,-.+ + +/ ! + * <! 3! ? 3#
/4
+,-.+ + +/ ! + * , ! 0 2 3!
? 3#
(% ' % *

8.1.4 Tangent Plane Maneuvers

A general maneuver changes the shape of an orbit, a, e, its orientation ω within the orbital plane, and the orientation of the plane, i, Ω. Practically, the orientation ω of the orbital plane is largely irrelevant, because most of the spacecraft are in a circular orbit having no ω. In addition, for orbit transitions a Hohmann transfer altering a, e jointly with a plane change i, Ω is essential. We read from Eq. (8.1.1) that to solely change a, e and i, Ω only $\delta v_{||}$ and $\delta v_{\perp\perp}$ kick-burns are relevant. This implies that $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ span a plane, tangent to the orbit at the kick-burn location. This important class of orbit maneuvers is therefore called *tangent plane maneuvers*. Figure 8.4 shows a tangent change maneuver with $v_1 = v_2 =: v$, that is, without a, e changes. Such a particular tangent plane maneuver is hence called *genuine plane change maneuver* (see next section).

Transition Angle in Terms of Orbital Elements
The important property of a tangent plane maneuver is that because the plane spaned by $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ is vertical to the initial and final plane, the maneuver causes a tilt of the initial (index 1) to the final (index 2) orbit about a mutual nodal line, for which $\phi = \angle(\boldsymbol{v}_1, \boldsymbol{v}_2) = \angle(\boldsymbol{I}_1, \boldsymbol{I}_2)$ holds. Here $\boldsymbol{I}$ is the three-dimensional inclination vector (see Eq. (7.3.25)) given as

$$\boldsymbol{I} = \begin{pmatrix} \sin\Omega \sin i \\ -\cos\Omega \sin i \\ \cos i \end{pmatrix}_{IJK}$$

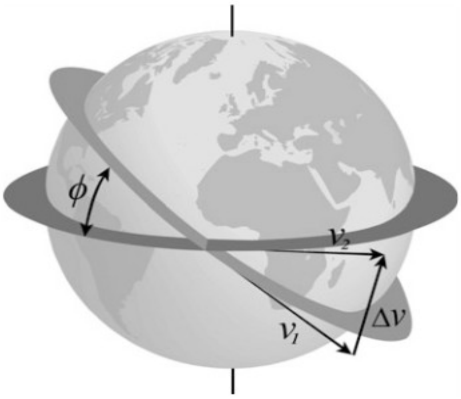

Fig. 8.4 A tangent plane maneuver with $v_1 = v_2$ (a.k.a. genuine plane change maneuver) in a circular orbit. The point where the maneuver takes place is the node of the mutual nodal line between the initial and final orbit plane

With this the transition angle is determined by

$$\cos\phi = \hat{\boldsymbol{v}}_1\hat{\boldsymbol{v}}_2 = \boldsymbol{I}_1\boldsymbol{I}_2 = \sin i_1 \sin i_2 \cos\Delta\Omega + \cos i_1 \cos i_2$$

or after some trigonometric modifications alternatively

$$\sin^2\frac{\phi}{2} = \sin^2\frac{\Delta i}{2} + \sin i_1 \sin i_2 \sin^2\frac{\Delta\Omega}{2} \tag{8.1.18}$$

where $\Delta i := i_2 - i_1$ and $\Delta\Omega := \Omega_2 - \Omega_1$. Obviously, the vector $\boldsymbol{N}$ of the nodal line between the initial and final orbit is given by

$$\boldsymbol{N} = \boldsymbol{I}_1 \times \boldsymbol{I}_2 \tag{8.1.19}$$

Maneuvers with either Inclination Change or RAAN Change

There are two special plane change maneuvers, which are of practical interest. One, in which the RAAN is kept constant, $\Delta\Omega = 0$. We thus obtain from Eq. (8.1.18)

$$\begin{aligned} &\phi = \Delta i \\ &\Delta v = \sqrt{v_1^2 + v_2^2 - 2v_1v_2\cos\Delta i} \end{aligned} \quad @\ \Delta\Omega = 0 \quad \textbf{nodal transfer} \tag{8.1.20}$$

According to Table 8.1 this plane tilt is achieved for a $\Delta\boldsymbol{v}$ that lies in either plane tangent to the two nodes of the initial orbit with the reference plane. Such transfers are therefore called *nodal transfers* and are the most frequently used plane change maneuvers.

The other special case is achieved if the inclination is kept constant, $\Delta i = 0$. We thus have from Eq. (8.1.18)

$$\begin{aligned} &\sin(\phi/2) = \sin i \cdot |\sin(\Delta\Omega/2)| \\ &\Delta v = \sqrt{(v_1 - v_2)^2 + 4v_1v_2\sin^2 i\,\sin^2(\Delta\Omega/2)} \end{aligned} \quad @\ \Delta i = 0 \tag{8.1.21}$$

According to Table 8.1 $\Delta i = 0$ is achieved for a $\Delta\boldsymbol{v}$ that lies in either plane tangent to the two points at 90° abeam from the nodes of the initial orbit with the reference plane.

8.1.5 Genuine Plane Change Maneuvers

We finally study just plane change maneuvers without a change in orbital velocity, $v_1 = v_2 =: v$ as depicted in Fig. 8.4. Because we now apply only a $\delta v_{\perp\perp}$ kick-burn, genuine plane change maneuvers are a subset of tangent plane maneuvers, which according to Eq. (8.1.1) only change i, Ω, ω. From Eq. (8.1.17) we derive for $v = v_1 = v_2$ quite generally

$$\boxed{\Delta v = 2v \cdot \left|\sin\frac{\phi}{2}\right|} \quad @ \quad v := v_1 = v_2 \tag{8.1.22}$$

Because $\Delta v \propto v$ we find the general rule of thumb

Plane Change Rule
An orbit plane change should be performed at the smallest possible orbital velocity, i.e. at best at the apoapsis.

For this reason it might be even preferable to bring the vehicle first into a higher orbit, change the plane there at apoapsis at low speed and low budget, and finally take it back to its original orbit. It is this principle, that the bi-elliptic transfer (Sect. 8.4.2) and the super-synchronous transfer (Sect. 8.4.3) make use of.

Genuine Inclination Change Maneuver
For $\Omega_1 = \Omega_2$ we have from Eq. (8.1.20) $\phi = \Delta i$. We hence obtain from Eq. (8.1.22)

$$\boxed{\Delta v = 2v \cdot \left|\sin\frac{\Delta i}{2}\right|} \quad @ \; v := v_1 = v_2,\; \Delta\Omega = 0 \tag{8.1.23}$$

According to Table 8.1 this tilt is achieved for a $\Delta \boldsymbol{v}$ that lies in either plane tangent to the two nodes of the initial orbit with the reference plane and is orthogonal to the plane bisecting the old and new orbit plane.

Genuine RAAN Change Maneuver
For $\Delta i = 0$ we have from Eq. (8.1.21)

$$\boxed{\Delta v = 2v \cdot \sin i \cdot \left|\sin\frac{\Delta\Omega}{2}\right|} \quad @ \; v := v_1 = v_2, \Delta i = 0 \tag{8.1.24}$$

This is the effort to rotate the inclination vector by $\Delta\Omega$ about the *K*-axis in the geocentric equatorial reference frame. According to Table 8.1 this rotation is achieved for a $\Delta \boldsymbol{v}$ that lies in either plane tangent to the two points at 90° abeam from the nodes of the initial orbit with the reference plane. Obviously, a RAAN change by $\Delta\Omega = 180°$ is identical to an inclination change by $\Delta i = 2i$.

Because $\Delta v \propto v \sin i$ we find the general rule of thumb.

RAAN Change Rule
Adjust RAAN at the lowest inclination and/or orbital velocity possible.

8.1.6 Tangent Maneuver

Another quite special case of a tangent plane maneuver is $\phi = 0°$, i.e., the burn is along the direction of motion. We call this a *tangent maneuver* for which from Eq. (8.1.17) follows

$$\Delta v_{||} = |v_2 - v_1| \quad @ \quad \phi = 0° \tag{8.1.25}$$

From Eq. (8.1.1) we see that such a maneuver changes a, e, ω. As we will see in Sect. 8.3.1 this is the most efficient maneuver to increase the size of an orbit and hence to transfer to an outer orbit.

Let us assume we are in an initial orbit with a_1, e_1, ω_1 and perform at $r := r_1 = r_2$ a tangent maneuver $r, v_1, \theta_1 \to r, v_2, \theta_2$ with

$$v_2 = v_1 \pm \Delta v_{||} \tag{8.1.26}$$

We want to know what is a_2, e_2, ω_2? From the vis-viva Eq. (7.2.15) we immediately find

$$\frac{1}{a_2} = \frac{2}{r} - \frac{v_2^2}{\mu} \tag{8.1.27}$$

To determine the other modified orbital elements we make use of the fact that a tangent maneuver obviously does not change the flight path angle $\gamma := \gamma_1 = \gamma_2$ (see Sect. 7.3.3). From Eq. (7.3.16) we derive the relationships

$$\begin{aligned} v \cos \gamma &= \frac{h}{r} \\ \frac{1}{2} v^2 \sin 2\gamma &= \frac{\mu}{r} e \sin \theta \end{aligned}$$

Applying these equations to the orbits before and after the tangent maneuver we find

$$\frac{v_2}{v_1} = \frac{h_2}{h_1} = \frac{a_2\left(1 - e_2^2\right)}{a_1\left(1 - e_1^2\right)}$$

$$\frac{v_2^2}{v_1^2} = \frac{e_2 \sin \theta_2}{e_1 \sin \theta_1}$$

from which we derive

$$e_2 = \sqrt{1 - \frac{a_1 v_2}{a_2 v_1}\left(1 - e_1^2\right)} \tag{8.1.28}$$

$$\sin\,\theta_2 = \frac{v_2^2\,e_1}{v_1^2\,e_2}\sin\,\theta_1 \tag{8.1.29}$$

To finally determine ω_2 we make use of the fact that a tangent maneuver also does not change the argument of latitude $u = \omega + \theta$ (see Sect. 7.3.5). We hence obtain

$$\omega_2 = \omega_1 + \theta_1 - \theta_2 \tag{8.1.30}$$

Example (GEO Positioning)

Let us assume that a communication satellite is launched with an ATLAS rocket from Cape Canaveral with orbital inclination $i_L = 28^\circ$ after launch and needs to be positioned in GEO (see Berlin (2005)). As we will see in Sect. 12.6.1, due to lunisolar perturbations an optimum initial state in GEO would be $i_{GEO} = 3^\circ,\ \Omega_{GEO} = 280^\circ$. After launch and at the descending node the rocket will make an injection burn at $r_{LEO} = 6578.14\,\text{km}$ into the GTO with apoapsis at $r_{GEO} = 42{,}166\,\text{km}$. Because the descending node happens to coincide with the Greenwich meridian we have $\Omega_{GTO} = 180^\circ$.

To place the satellite in its final orbit position we have the choice of many different maneuver sequences for all of which are based on the following velocities before (−) and after (+) the kick-burns in LEO and GEO (see Eqs. (7.4.3) and (7.4.10)):

$$\begin{aligned} v_{LEO-} &= 7.784\,\text{km s}^{-1} \\ v_{LEO+} &= 10.239\,\text{km s}^{-1} \end{aligned} \quad \text{and} \quad \begin{aligned} v_{GEO-} &= 1.597\,\text{km s}^{-1} \\ v_{GEO+} &= 3.075\,\text{km s}^{-1} \end{aligned}$$

We now study exemplarily four different maneuvering sequences with the following total delta-v:

Sequence 1

- At the descending node after launch an inclination change maneuver ((Eq. (8.1.20) into Eq. (8.1.17)) is made that brings the satellite into a GTO with $i_{GEO} = 3^\circ,\ \Omega_{GTO} = 180^\circ$.
- At apogee a genuine RAAN change maneuver (see Eq. (8.1.24)) changes the orbit plane to $\Omega_{GEO} = 280^\circ$.
- Finally, a tangent maneuver (see Eq. (8.1.25)) brings the satellite into a circular GEO.

$$
\begin{aligned}
\Delta v = &\sqrt{v_{LEO-}^2 + v_{LEO+}^2 - 2v_{LEO-}v_{LEO+}\cos(28^\circ - 3^\circ)} \\
&\quad + 2v_{GEO-}\sin 3^\circ \cdot \sin[(280^\circ - 180^\circ)/2] + (v_{GEO+} - v_{GEO-}) \\
= &\, 4.578 + 0.128 + 1.478 = 6.184\,\text{km s}^{-1}
\end{aligned}
$$

Sequence 2

- At the descending node after launch a tangent maneuver transfers the satellite into a GTO, i.e., $i_{GTO} = 28^\circ,\ \Omega_{GTO} = 180^\circ$.
- At apogee a genuine RAAN change maneuver (see Eq. (8.1.24)) changes the orbit plane to $\Omega_{GEO} = 280^\circ$.
- Finally, a inclination change maneuver (Eq. (8.1.20) into Eq. (8.1.17)) is performed that circularizes the satellite into GEO at $i_{GEO} = 3^\circ$.

$$
\begin{aligned}
\Delta v = &(v_{LEO+} - v_{LEO-}) + 2v_{GEO-}\sin 28^\circ \cdot \sin[(280^\circ - 180^\circ)/2] \\
&\quad + \sqrt{v_{GEO-}^2 + v_{GEO+}^2 - 2v_{GEO-}v_{GEO+}\cos(28^\circ - 3^\circ)} \\
= &\, 2.455 + 1.149 + 1.762 = 5.366\,\text{km s}^{-1}
\end{aligned}
$$

Sequence 3

1. At the descending node after launch a tangent maneuver transfers the satellite into a GTO, i.e., $i_{GTO} = 28^\circ,\ \Omega_{GTO} = 180^\circ$.
2. At apogee a genuine inclination change maneuver (see Eq. (8.1.23)) changes the inclination to $i_{GEO} = 3^\circ$.
3. Still at apogee it is immediately followed by a genuine RAAN change maneuver (see Eq. (8.1.24)) that changes the orbit plane to $\Omega_{GEO} = 280^\circ$.
4. Finally, the orbit is circularized with an apogee kick-burn.

$$
\begin{aligned}
\Delta v = &(v_{LEO+} - v_{LEO-}) + 2v_{GEO-} \cdot \sin 12.5^\circ \\
&\quad + 2v_{GEO-}\sin 3^\circ \cdot \sin[(280^\circ - 180^\circ)/2] + (v_{GEO+} - v_{GEO-}) \\
= &\, 2.455 + 0.691 + 0.128 + 1.478 = 4.752\,\text{km s}^{-1}
\end{aligned}
$$

Sequence 4

- At the descending node the satellite is transferred into a GTO, i.e., $i_{GTO} = 28^\circ, \Omega_{GTO} = 180^\circ$.
- At apogee, an inclination-change maneuver (Eq. (8.1.20) into Eq. (8.1.17)) is performed such that the orbit is circularized at $i_{GEO} = 3^\circ$.
- Finally, a genuine RAAN change maneuver (see Eq. (8.1.24)) positions the satellite at $\Omega_{GEO} = 280^\circ$.

$$\begin{aligned}\Delta v &= (v_{LEO+} - v_{LEO-}) + \sqrt{v_{GEO-}^2 + v_{GEO+}^2 - 2v_{GEO-}v_{GEO+}\cos(28^\circ - 3^\circ)} \\ &\quad + 2v_{GEO+}\sin 3^\circ \cdot \sin[(280^\circ - 180^\circ)/2] \\ &= 2.455 + 1.762 + 0.247 = 4.464\,\text{km s}^{-1}\end{aligned}$$

The essence of comparing the four sequences is that for GEO positioning an inclination change should be combined with the orbit circularization at apoapsis and RAAN should be adjusted at the lowest possible inclination. This is how GEO positioning is actually performed.

This example corroborates the following rule of thumb, which is based on the fact (see Problem 8.1) that a tangent plane maneuver is always more efficient than a tangent maneuver and plane change maneuver performed sequentially at the same point in space:

All-in-One Rule
If possible combine all orbital changes that need to be done into one kick-burn at one orbital position rather than making single burns at successive positions.

8.2 Lambert Transfer

Having explored the characteristics of one-impulse maneuvers between intersecting orbits we are now set to perform orbital transfers between two conic orbits that do not intersect and therefore require a transfer orbit with two orbital transition maneuvers at each end. The most general case is the so-called Lambert transfer, which we will investigate first. It played an important role in the 1960s for Gemini and Apollo orbital rendezvous. Today, Hohmann transfers are more common, which will be studied in Sect. 8.3.

8.2.1 Orbital Boundary Value Problem

Lambert Transfer
Let us assume that we have a S/C at point P_1 on an conic Earth orbit that needs to transfer to point P_2 on a coplanar conic target orbit, both of which do not intersect. (If they intersect we can perform an one-impulse maneuver at the intersection point, as described in Sect. 8.1, and move from there on the target orbit to point P_2.) Because the orbits and hence also the transfer orbit are coplanar they are fully characterized by their form and orientation to each other, i.e., by three orbital elements, namely $a_i, \boldsymbol{e}_i$ for the initial Earth orbit, $a, \boldsymbol{e}$ for the transfer orbit, and $a_f, \boldsymbol{e}_f$

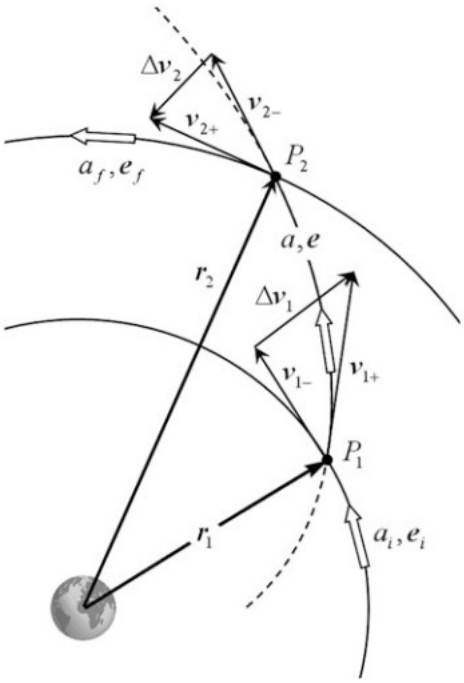
$\Delta \boldsymbol{v}_2$
$\boldsymbol{v}_{2-}$
$\boldsymbol{v}_{2+}$
P_2
$a_f, \boldsymbol{e}_f$
$a, \boldsymbol{e}$
$\boldsymbol{r}_2$
$\Delta \boldsymbol{v}_1$
$\boldsymbol{v}_{1-}$
$\boldsymbol{v}_{1+}$
P_1
$\boldsymbol{r}_1$
$a_i, \boldsymbol{e}_i$

Fig. 8.5 A Lambert transfer (see text for details)

for the final target orbit. The most general description of this problem is that under the influence of a common central gravitational potential a S/C in the initial orbit at P_1 possesses the state vector $(\boldsymbol{r}_1, \boldsymbol{v}_{1-})$ and wants to transfer to P_2 with state vector $(\boldsymbol{r}_2, \boldsymbol{v}_{2+})$. This is a general two-impulse transfer, which we call a *Lambert transfer*, embracing a first impulse maneuver $\Delta \boldsymbol{v}_1$ at P_1 resulting in an unpowered conic coplanar Lambert transfer orbit with orbital elements $a, \boldsymbol{e}$ that ends at P_2 where the second impulse maneuver $\Delta \boldsymbol{v}_2$ takes place.

This can be shown as follows (see Fig. 8.5):

$$(a_i, \boldsymbol{e}_i)\, \underbrace{[\Delta \boldsymbol{v}_1]}_{P_1} \xrightarrow{a,\boldsymbol{e}} \underbrace{[\Delta \boldsymbol{v}_2]}_{P_2} (a_f, \boldsymbol{e}_f)$$

Lambert transfer

$$(\boldsymbol{r}_1, \boldsymbol{v}_{1-})\, \underbrace{{}_{-}[\Delta \boldsymbol{v}_1]_{+}}_{P_1} \xrightarrow{a,\boldsymbol{e}} \underbrace{{}_{-}[\Delta \boldsymbol{v}_2]_{+}}_{P_2} (\boldsymbol{r}_2, \boldsymbol{v}_{2+})$$

where the $+$ and $-$ signs indicate the situation just before or after an impulse maneuver. Since only the initial and final orbits and P_1 and P_2 are specified, while $\Delta \boldsymbol{v}_1$ and $\Delta \boldsymbol{v}_2$ remain unspecified, there in general exist many different Lambert transfer orbits connecting the two points P_1 and P_2.

Orbital Boundary Value Problem

The problem of finding a transfer orbit connecting two given points P_1 and P_2 is called the *orbital boundary value problem*, which is a special case of the general mathematical *two-point boundary value problem*.

Let us study the orbital boundary value problem by first assuming that just P_1 and P_2 are given through the position vectors $\boldsymbol{r}_1, \boldsymbol{r}_2$. With these we get from the orbit Eqs. (7.3.4) and (7.3.6) the two conditional equations for the three orbital elements $a, \boldsymbol{e}$

$$\begin{aligned} r_1 + \boldsymbol{e}\boldsymbol{r}_1 &= a\left(1 - e^2\right) \\ r_2 + \boldsymbol{e}\boldsymbol{r}_2 &= a\left(1 - e^2\right) \end{aligned}$$

Therefore, two orbital points are not sufficient to unequivocally define a Keplerian transfer orbit, but leave one orbital element undefined. We need a third position vector or an additional scalar specification to concretize the transfer orbit. Solving the orbital boundary value problem by three position vectors is accomplished by Gibbs' method, which will not be investigated here, but see e.g., Vallado (2001) or Curtis (2005). An additional scalar specification might be for instance the transfer time. This famous orbital boundary value problem is called *Lambert's problem* and will be discussed in Sect. 8.2.3.

As we will see in Sect. 8.2.2, even an additional scalar specification generally still leaves the alternative between two so-called *conjugate orbits*: a long path and a short path transfer orbit. Therefore the general orbital boundary value problem for a Lambert transfer to find the orbital elements $a, \boldsymbol{e}$ of the transfer orbit is uniquely defined only by providing the following boundary conditions:

$\boldsymbol{r}_1,\ \boldsymbol{r}_2$
+ an additional scalar specification
+ a specification for a long or short path transfer orbit

boundary conditions (8.2.1)

Given these boundary conditions we are able derive a unique solution $a, \boldsymbol{e}$ to the orbital boundary value problem in the following way. *Suppose we have found the unique transfer orbit by providing p.* According to Battin (1987), the terminal velocities at P_1 and P_2 are then determined as

$$\begin{aligned} \boldsymbol{v}_{1+} &= \frac{\sqrt{\mu p}}{r_1 r_2 \sin(\Delta\theta)}\left[(\boldsymbol{r}_2 - \boldsymbol{r}_1) + \frac{r_2}{p}(1 - \cos\,\Delta\theta)\boldsymbol{r}_1\right] \\ \boldsymbol{v}_{2-} &= \frac{\sqrt{\mu p}}{r_1 r_2 \sin(\Delta\theta)}\left[(\boldsymbol{r}_2 - \boldsymbol{r}_1) - \frac{r_1}{p}(1 - \cos\,\Delta\theta)\boldsymbol{r}_2\right] \end{aligned} \tag{8.2.2}$$

where $\Delta\theta = \angle(\boldsymbol{r}_1, \boldsymbol{r}_2)$ is the so-called *transfer angle*. Having thus found the state vectors $(\boldsymbol{r}_1, \boldsymbol{v}_{1+})$ at point P_1 and $(\boldsymbol{r}_2, \boldsymbol{v}_{2-})$ at point P_2 of the transfer orbit we derive from Eq. (7.3.3) the orientation of the conic transfer orbit as

$$\boldsymbol{e} = \left(\frac{1}{r_1} - \frac{1}{a}\right)\boldsymbol{r}_1 - \frac{1}{\mu}(\boldsymbol{r}_1\boldsymbol{v}_{1+})\boldsymbol{v}_{1+} = \left(\frac{1}{r_2} - \frac{1}{a}\right)\boldsymbol{r}_2 - \frac{1}{\mu}(\boldsymbol{r}_2\boldsymbol{v}_{2-})\boldsymbol{v}_{2-} \tag{8.2.3}$$

Having thus determined the relevant orbital elements $a, \boldsymbol{e}$ of the Lambert transfer orbit, we have found the solution to the orbital boundary value problem. Note that with such methods the orbital boundary value problem is also used to preliminarily determine an unknown orbit. Preliminary orbit determination will be discussed in Chap. 14.

With the state vectors $(\boldsymbol{r}_1, \boldsymbol{v}_{1+})$ and $(\boldsymbol{r}_2, \boldsymbol{v}_{2-})$ we also know the required orbital maneuvers and the total delta-v

$$\begin{aligned} \Delta\boldsymbol{v}_1 &= \boldsymbol{v}_{1+} - \boldsymbol{v}_{1-} \\ \Delta\boldsymbol{v}_2 &= \boldsymbol{v}_{2+} - \boldsymbol{v}_{2-} \\ \Delta v &= |\Delta\boldsymbol{v}_1| + |\Delta\boldsymbol{v}_2| \end{aligned} \tag{8.2.4}$$

Having thus fully determined the Lambert transfer we recognize that everything hinges on the problem of finding the semi-latus rectum p of the Lambert transfer orbit under the given boundary conditions Eq. (8.2.1) to which we turn now.

8.2.2 Lambert Transfer Orbits

We will first study the characteristics of the solutions of the orbital boundary value problem if only the position vectors $\boldsymbol{r}_1$, $\boldsymbol{r}_2$ are given. In this case there exists an infinite number of orbits with different a and e connecting P_1 and P_2. Let us first assume that with an additional specification we have picked a particular but otherwise arbitrary a.

Conjugate Orbits: One *a*, But Two Transfer Orbits

It is a basic property of the orbital boundary value problem that for any particular semi-major axis a in general there exist two different transfer orbits, so-called *conjugate orbits*, with different eccentricity and hence semi-latus rectum, which we will label p and $\tilde{p}$.

From geometrical reasoning, it is enlightening to see why. Suppose there is a transfer orbit with a given a. We define the distances from its empty focus F' to the points P_1 and P_2 as r'_1 and r'_2, respectively. Then from the definition of an ellipse we have

$$\begin{aligned} r_1 + r'_1 &= 2a \\ r_2 + r'_2 &= 2a \end{aligned}$$

Since the location of the focal point F is given by the central mass, we can establish the transfer ellipses by determining their empty focal points $F'_{1,2}$. We do so by drawing a circle with radius $r'_1 = 2a - r_1$ around P_1 and a circle with radius

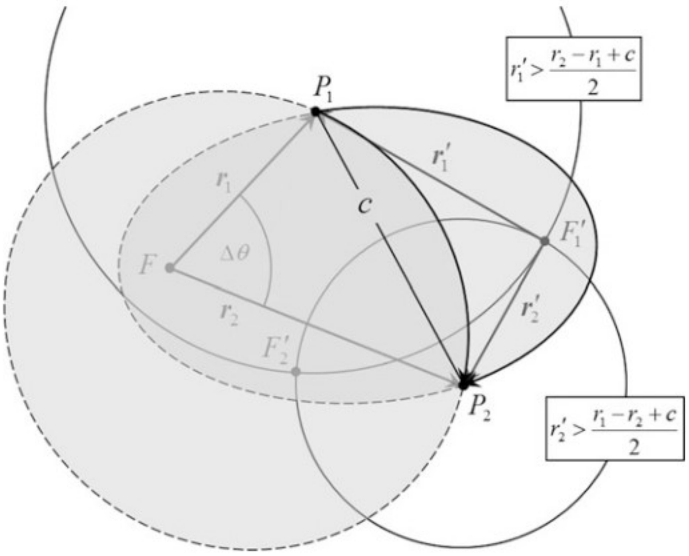
$r_1' > \frac{r_2 - r_1 + c}{2}$
P_1
r_1'
r_1
c
F
$\Delta\theta$
F_1'
r_2
r_2'
F_2'
P_2
$r_2' > \frac{r_1 - r_2 + c}{2}$

Fig. 8.6 Circles, with radii larger than some critical value, drawn around the points P_1 and P_2 intersect at two points defining the empty focal points F' of the two different transfer ellipses

$r_2' = 2a - r_2$ around P_2. If a is large enough ("large enough" will be shown in a moment to be $a \geq (r_1 + r_2)/2$), the two circles will intersect at two empty focal points on opposite sides of the chord connecting P_1 and P_2 as shown in Fig. 8.6, which define the two transfer ellipses.

A transfer orbit is called a *short path ellipse* if it has the shorter path between P_1 and P_2, and hence a smaller eccentricity, and a shorter transfer time; while the other is called a *long path ellipse*.

Note that their major-axes are indeed of identical length, only their eccentricities and their orientations are different.

When a is known, it can be shown (see Battin (1964, 1987), or Kemble (2006)) that the semi-latus rectum p of the two paths are given as

$$\begin{aligned} p &= a\,\frac{r_1 r_2}{c^2}\,(1 - \cos\Delta\theta)[1 - \cos(\alpha + \beta)] \\ &= 2a\,\frac{r_1 r_2}{c^2}(1 - \cos\Delta\theta)\sin^2\frac{\alpha+\beta}{2} \end{aligned} \quad \text{@ short path} \qquad (8.2.5\text{a})$$

$$\begin{aligned} \tilde{p} &= \frac{r_1 r_2}{a}\,\frac{1-\cos\Delta\theta}{1-\cos(\alpha + \beta)} \\ &= 2a\,\frac{r_1 r_2}{c^2}(1 - \cos\Delta\theta)\sin^2\frac{\alpha-\beta}{2} \end{aligned} \quad \text{@ long path} \qquad (8.2.5\text{b})$$

where the chord length c, which is the separation distance between P_1 and P_2, is given from trigonometry by

$$c = \sqrt{r_1^2 + r_2^2 - 2r_1 r_2 \cos \Delta\theta} \tag{8.2.6}$$

and

$$\sin\frac{\alpha}{2} := \sqrt{\frac{r_1 + r_2 + c}{4a}}, \quad \sin\frac{\beta}{2} := \sqrt{\frac{r_1 + r_2 - c}{4a}}$$

From Eqs. (8.2.5a) and (8.2.5b) we see immediately that

$$p \cdot \tilde{p} = \left[\frac{r_1 r_2}{c}(1 - \cos \Delta\theta)\right]^2 = p_{\min}^2$$

where $p_{\min}$ is the semi-latus rectum of the minimum energy transfer orbit discussed in the following (see Eq. (8.2.9)). We also see that owing to $0 \le \beta \le \alpha \le \pi$

$$c\tilde{p} < r_1 r_2 (1 - \cos \Delta\theta) < cp \quad \textbf{long/short path condition} \tag{8.2.7}$$

Minimum Energy Transfer Orbit, But Maximum Transfer Time

Because Eq. (7.3.18) directly relates the magnitude of the semi-major axis to the orbital energy by $\varepsilon = -\mu/2a$, this condition states that among all transfer orbits that

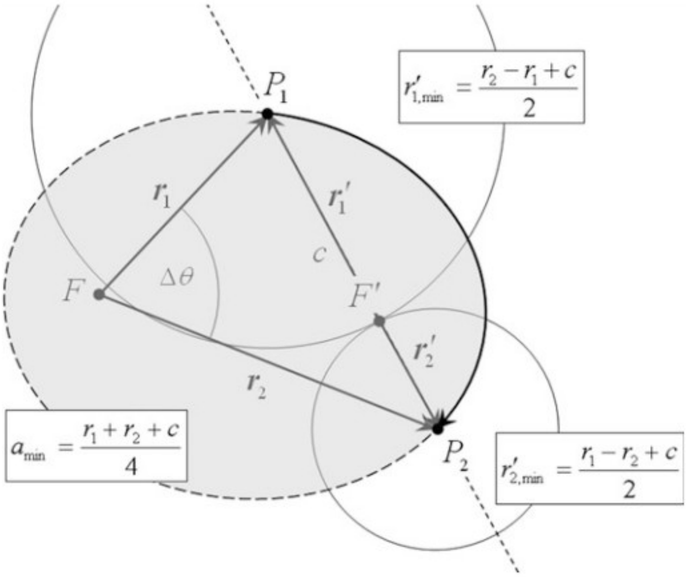
P_1
$r'_{1,\min} = \frac{r_2 - r_1 + c}{2}$
r_1
r_1'
c
F
$\Delta\theta$
F'
r_2'
r_2
$a_{\min} = \frac{r_1 + r_2 + c}{4}$
P_2
$r'_{2,\min} = \frac{r_1 - r_2 + c}{2}$

Fig. 8.7 The minimum energy transfer orbit occurs when the empty focal point lies on the chord c connecting point P_1 and P_2

go through P_1 and P_2 there is an orbit with minimum orbital energy. To figure out what its a is, we consider again $r_1 + r'_1 = 2a$ and, $r_2 + r'_2 = 2a$ from which follows

$$4a = r_1 + r_2 + r'_1 + r'_2$$

With $r_1 + r_2$ as a given quantity, a is minimum if $r'_1 + r'_2$ becomes minimal, which from Fig. 8.7 is the case if $F'_{1,2}$ lies on the chord c joining P_1 and P_2. Therefore the minimum energy transfer orbit has

$$a_{\min} = \frac{r_1 + r_2 + c}{4} \tag{8.2.8}$$

We recognize that in this limiting situation the two circles around P_1 and P_2 do not intersect, but just touch each other leading to just one empty focal point and therefore to one, and only one, transfer ellipse. If a is smaller than the minimum value the circles will not intersect and no transfer solution exists—the minimum energy orbit is its own conjugate. On the other hand, because $c \le r_1 + r_2$ always holds, a transfer orbit with $a \ge (r_1 + r_2)/2$ will always be large enough to ensure intersecting circles and hence link the points P_1 and P_2. It can be shown (see, e.g., Battin (1987) or Vallado (2001)) that $p_{\min}$ for the minimum energy orbit, is given by

$$p_{\min} = \frac{r_1 r_2}{c}(1 - \cos\Delta\theta) = \frac{c}{2}\left[1 - \left(\frac{r_1 - r_2}{c}\right)^2\right] \tag{8.2.9}$$

We summarize by stating that

> The **minimum energy ellipse** has a minimum semi-latus rectum given by Eq. (8.2.9) and its empty focal point lies on the chord connecting P_1 and P_2.

Notice that with the additional specification "minimum energy transfer" we have rounded out all boundary conditions and hence settled on a specific solution with an unique $a = a_{\min}$ and $e = \sqrt{1 - p_{\min}/a_{\min}}$.

Note that the occurrence of the minimum energy transfer orbit does not necessarily mean that the total delta-v for the Lambert transfer also becomes a minimum. As seen from Eqs. (8.2.2) and (8.2.4) the total delta-v rather depends in a complex way on the transition from the initial velocity to the transfer velocity at P_1 and back to the final orbital velocity at P_2. But, in general it can be said that any transfer close to a minimum energy transfer is a "good" transfer. Observe that for a minimum energy transfer Δt no longer is a variable that can be freely chosen. The free variable then is the proper configuration of the objects at P_1 and P_2 in their orbits around F that complies with a minimum energy transfer. The search for such an optimal configuration, the so-called orbit phasing, is the objective of Sect. 9.3 for interplanetary transfers.

It can be shown (Problem 8.8) that the minimum energy transfer orbit goes hand in hand with a maximum transfer time. The semi-latus rectum for a minimum energy transfer is given above as $4a_{\min} = r_1 + r_2 + c$. With the transformations introduced for Lambert's problem in Sect. 8.2.3, we have

$$\sin\frac{\alpha_{\min}}{2} = \sqrt{\frac{r_1 + r_2 + c}{4a_{\min}}} = 1,$$
$$\sin\frac{\beta_{\min}}{2} = \sqrt{\frac{r_1 + r_2 - c}{r_1 + r_2 + c}}$$

and hence $\alpha_{\min} = \pi$. With Lambert's equation (see Eq. (8.2.13)) the maximum transfer time of the minimum energy transfer orbit then reads

$$\begin{aligned}\Delta t_{\max} &= \sqrt{\frac{a_{\min}^3}{\mu}}[\pi - \beta_{\min} + \sin\beta_{\min}] \\ &= \sqrt{\frac{a_{\min}^3}{\mu}}\left[\pi - \arccos\left(\frac{c}{a_{\min}} - 1\right) + \sqrt{\frac{c}{a_{\min}}\left(2 - \frac{c}{a_{\min}}\right)}\right]\end{aligned} \tag{8.2.10}$$

The Fundamental Ellipse

There are many transfer orbits with metric orbital elements a, e that connect P_1 and P_2. However, as there is just one transfer orbit with a characteristic a—the minimum energy orbit—there exists just one transfer orbit with a specified characteristic eccentricity: the *fundamental ellipse* with a minimum eccentricity.

To find the orbit with minimum e we first investigate the eccentricity vector as given by the orbit equation Eq. (7.3.4) in vectorial form and apply it to the points P_1 and P_2, obtaining

$$\boldsymbol{e}\boldsymbol{r}_1 = p - r_1$$
$$\boldsymbol{e}\boldsymbol{r}_2 = p - r_2$$

Subtracting these from each other yields

$$\boldsymbol{e}(\boldsymbol{r}_1 - \boldsymbol{r}_2) = r_1 - r_2$$

Since $\boldsymbol{r}_1 - \boldsymbol{r}_2 = c\boldsymbol{u}_c$ is the cord with unit vector $\boldsymbol{u}_c$ and length c, we get

$$\boldsymbol{e}\boldsymbol{u}_c = \frac{r_1 - r_2}{c}$$

This equation states that the eccentricity vector of any transfer orbit between P_1 and P_2 has the same projection on the chord. The fundamental ellipse now is that one of those with the minimum eccentricity. Because the minimum eccentricity is achieved if $\boldsymbol{e}||\boldsymbol{u}_c$, we have for the fundamental ellipse

$$e_F = \frac{r_1 - r_2}{c} = \min \tag{8.2.11}$$

and because $\boldsymbol{e}||\boldsymbol{u}_c$ implicates $2a = r'_1 + r'_2 = r_1 + r_2$ we have

$$p_F = a\left(1 - e_F^2\right) = \frac{r_1 r_2}{c^2}(r_1 + r_2)(1 - \cos\,\Delta\theta) \tag{8.2.12}$$

Because the fundamental ellipse has a particular p_F it has a particular flight time, which is identical with that of its conjugate orbit having the same semi-latus rectum but different eccentricity.

Note that though the minimum energy transfer ellipse and the fundamental ellipse share the same semi-latus rectum $a_{\min} = a_F = (r_1 + r_2)/2$, they do not share the same position of the empty focal point F'. While F' for the minimum energy orbit lies on the chord, the one for the fundamental ellipse lies on a line through F that runs parallel to the chord (see the "low-e" ellipse in Fig. 8.6 that is close to the fundamental ellipse). Therefore,

The **fundamental ellipse** has the lowest eccentricity and shares the minimum semi-latus rectum given by Eq. (8.2.12) of all elliptic orbits that lie on P_1 and P_2. Its line of apsides runs parallel to the chord line.

8.2.3 Lambert's Problem

Apart from the minimum energy or minimum eccentricity specifications there might be other specifications that settle on a transfer orbit between P_1 and P_2 with a particular a. From a practical and historical point of view the most important, however, is the specification of a transfer time. The problem to find the according transfer orbit is called *Lambert's problem* (Lambert 1761). It can be stated as follows: "the determination of an orbit, having a specified transfer time and connecting two position vectors", or in other words "to solve for the trajectory connecting two position vectors with a given time of flight".

Because in Lambert's problem both transfer time and orbital anomalies are involved, for a solution we start out with Kepler's Eq. (7.4.15) for an elliptic transition and Eq. (7.4.30) for a hyperbolic orbit

$$n \cdot (t_1 - t_2) = n \cdot \Delta t = E_1 - E_2 - e(\sin E_1 - \sin E_2) \qquad \text{@ elliptic orbit}$$
$$n \cdot (t_1 - t_2) = n \cdot \Delta t = e(\sinh F_1 - \sin F_2) - (F_1 - F_2) \qquad \text{@ hyperbolic orbit}$$

It can be shown (for a proof see, e.g., Battin (1964, 1987), or Kemble (2006)) that with the substitutions

$$\sin\frac{\alpha - \beta}{2} := \frac{E_2 - E_1}{2}, \cos\frac{\alpha + \beta}{2} := e\,\cos\frac{E_2 + E_1}{2} \qquad \text{@ elliptic}$$

$$\sinh\frac{\gamma - \delta}{2} := \frac{F_2 - F_1}{2}, \cosh\frac{\gamma + \delta}{2} := e\,\cosh\frac{F_2 + F_1}{2} \qquad \text{@ hyperbolic}$$

and with the quantities

$$c = \sqrt{r_1^2 + r_2^2 - 2r_1 r_2 \cos\,\Delta\theta} \qquad @\, 0 \le \Delta\theta \le \pi$$

$$s_\pm := \frac{1}{2}\sqrt{\frac{r_1 + r_2}{c} \pm 1}$$

Kepler's equations can be rewritten to yield *for the short path and hence short transition times and for small transfer angles* $0 \le \Delta\theta \le \pi$ the Lambert's equations

Elliptic Orbit (short path, $0 \le \Delta\theta \le \pi$)

$$\boxed{\Delta t = \sqrt{\tfrac{a^3}{\mu}}[\alpha - \beta - (\sin\,\alpha - \sin\,\beta)]} \quad \textbf{Lambert's equation} \qquad (8.2.13)$$

with

$$\sin\frac{\alpha}{2} = \sqrt{\frac{r_1 + r_2 + c}{4a}} = \sqrt{\frac{c}{a}s_+} > 0$$

$$\sin\frac{\beta}{2} = \sqrt{\frac{r_1 + r_2 - c}{4a}} = \sqrt{\frac{c}{a}s_-} > 0 \qquad @ \quad 0 \le \beta \le \alpha \le \pi$$

$$p = a\tfrac{r_1 r_2}{c^2}(1 - \cos\Delta\theta)[1 - \cos(\alpha + \beta)]$$

Hyperbolic Orbit (short path, $0 \le \Delta\theta \le \pi$)

$$\boxed{\Delta t = \sqrt{\tfrac{-a^3}{\mu}}[\sinh\gamma - \sinh\delta - (\gamma - \delta)]} \quad \textbf{Lambert's equation} \qquad (8.2.14)$$

with

$$\sinh\frac{\gamma}{2} = \sqrt{-\frac{r_1 + r_2 + c}{4a}} = \sqrt{\frac{c}{-a}s_+} > 0$$

$$\sinh\frac{\delta}{2} = \sqrt{-\frac{r_1 + r_2 - c}{4a}} = \sqrt{\frac{c}{-a}s_-} > 0 \qquad @\, 0 \le \delta \le \gamma$$

$$p = a\tfrac{r_1 r_2}{c^2}(1 - \cos\Delta\theta)[1 - \cosh(\gamma + \delta)]$$

where c again is the chord length, i.e., the straight distance between the points P_1 and P_2, and $\Delta\theta = \angle(\boldsymbol{r}_1, \boldsymbol{r}_2)$ is again the so-called *transfer angle*. Obviously, $r_1 + r_2 + c \ge r_1 + r_2 - c$ implies that $0 \le \beta \le \alpha \le \pi$ must hold.

Parabolic and Near-Parabolic Trajectories

If $|a| \to \infty$, i.e., if we have a near-parabolic transfer path, it can be shown (see Problem 8.9) that a can be provided analytically as

$$\frac{c}{a} = \frac{5}{2} \frac{\Delta t \sqrt{\mu / c^3} - \frac{4}{3}\left(s_+^3 - s_-^3\right)}{s_+^5 - s_-^5} \quad \text{@ near-parabolic path} \tag{8.2.15}$$

Equation (8.2.15) provides also the condition at which $|a| \to \infty$, namely $\frac{3}{4} \Delta t \sqrt{\mu / c^3} \approx s_+^3 - s_-^3$. To determine the semi-latus rectum p for $c/a > 0$ Eq. (8.2.13), and for $c/a < 0$ Eq. (8.2.14) need to be applied.

If

$$\Delta t = \frac{4}{3} \sqrt{\frac{c^3}{\mu}} \left(s_+^3 - s_-^3\right) \quad \text{@ parabolic path}$$

then $a = \infty$ and we have a parabolic transfer path with a single p given by

$$\begin{aligned} p &= 2 \frac{r_1 r_2}{c} (1 - \cos \Delta\theta)(s_+ + s_-)^2 \\ &= 2 \frac{r_1 r_2}{c^2} \sin^2 \frac{\Delta\theta}{2} \left(r_1 + r_2 + 2\sqrt{r_1 r_2} \cos \frac{\Delta\theta}{2} \right) \end{aligned} \quad \text{@ parabolic path} \tag{8.2.16}$$

With this the orbit orientation $\boldsymbol{e}$ can be deduced via Eqs. (8.2.2) and (8.2.3).

Lambert's Theorem

From these results Lambert's theorem can be derived.

> **Lambert's theorem**
>
> *The transfer time of a body moving between two points on a conic trajectory is a function only of the sum of the distances of the two points from the origin of force, the length c of the chord joining these two positions, and the semi-major axis of the conic:* $\Delta t = f(r_1 + r_2, c, a)$

Alternatively, Lambert's theorem can be stated as

$$\Delta t = f(r_1, r_2, a, \Delta\theta)$$

This more explicitly shows that the transfer time does not depend on the individual locations of the starting or arrival point or their true anomalies, but just on their separation angle $\Delta\theta$. This is for instance useful in finding a solution to the orbital rendezvous problem (see Sect. 8.6). It is remarkable that transfer time does not depend on the eccentricity of the transfer orbit. Remarkably, this is the third orbital quantity, besides the specific orbital energy (Eq. (7.3.19)) and the orbital period of an ellipse (Eq. (7.4.2)), that does not depend on the eccentricity. It is just this independency of e that makes the fundamental ellipse dispensable for practical purposes.

Universal Formulation of Lambert's Equation

Given Lambert's equation we are set to provide a universal formulation of Lambert's equation. With P_1 and P_2 given, we know $\boldsymbol{r}_1, \boldsymbol{r}_2$, $\Delta\theta$, and hence chord length c. For a given Δt we then can derive from Lambert's equation (8.2.13) the semi-major axis a of the corresponding short transfer path numerically, for instance with Newton's method (see Sect. 7.4.2).

In the following we want to provide a numerically robust algorithm for any Keplerian orbit. For that we first define the dimensionless numbers

$$\begin{aligned}
&x := c/a \\
&\Delta\tau := \Delta t\sqrt{\frac{\mu}{c^3}} \\
&\boldsymbol{\rho}_1 = \boldsymbol{r}_1/c,\ \boldsymbol{\rho}_2 = \boldsymbol{r}_2/c \\
&s_\pm = \frac{1}{2}\sqrt{\rho_1 + \rho_2 \pm 1} \\
&c_\pm(x) := \sqrt{1 - x s_\pm^2}
\end{aligned}$$

With this and the condition $0 \le \beta \le \alpha \le \pi$ we find for an elliptic orbit

$$\begin{aligned}
\sin\,\alpha - \sin\,\beta &= 2\left(\sin\frac{\alpha}{2}\cos\frac{\alpha}{2} - \sin\frac{\beta}{2}\cos\frac{\beta}{2}\right) \\
&= 2(s_+ c_+ - s_- c_-)\sqrt{x} > 0 \\
\alpha - \beta &= 2\left(\arcsin s_+\sqrt{x} - \arcsin s_-\sqrt{x}\right) \\
&= 2\arcsin\left[(s_+ c_- - s_- c_+)\sqrt{x}\right] > 0
\end{aligned}$$

and for a hyperbolic orbit with $0 \le \beta \le \alpha$

$$\begin{aligned}
\sinh\,\gamma - \sinh\,\delta &= 2\left(\sinh\frac{\gamma}{2}\cosh\frac{\gamma}{2} - \sinh\frac{\delta}{2}\cosh\frac{\delta}{2}\right) \\
&= 2(s_+ c_+ - s_- c_-)\sqrt{-x} > 0 \\
\gamma - \delta &= 2\left(\mathrm{arsinh} s_+\sqrt{-x} - \arcsin s_-\sqrt{-x}\right) \\
&= 2\mathrm{arsinh}\left[(s_+ c_- - s_- c_+)\sqrt{-x}\right] > 0
\end{aligned}$$

so that we can rewrite and merge Lambert's equations for both elliptical and hyperbolic orbits into one equation

$$\frac{\Delta\tau}{2}x = \frac{g\left[(s_+ c_- - s_- c_+)\sqrt{|x|}\right]}{\sqrt{|x|}} - (s_+ c_+ - s_- c_-) \quad @\ \text{Keplerian orbits}$$

where

$$g\left(\sqrt{|u|}\right) := \begin{cases} \arcsin(\sqrt{u}) & @\ 0 < u \leq 1 \\ \operatorname{arsinh}(\sqrt{-u}) & @\ u < 0 \end{cases}$$

The similarity of the power series expansion in both cases gives rise to an universal **L-function**

$$\begin{aligned} L(u) &= 1 + \frac{1}{2 \cdot 3} u + \frac{1 \cdot 3}{2 \cdot 4 \cdot 5} u^2 + \frac{1 \cdot 3 \cdot 5}{2 \cdot 4 \cdot 6 \cdot 7} u^3 \dots \\ &\quad + \frac{(2n-1)!!}{2^n n! (2n+1)} u^n + \ \dots \qquad\qquad @ -\infty < u \leq 1 \\ &= \begin{cases} \dfrac{\operatorname{arsin}\sqrt{u}}{\sqrt{u}} & @\ 0 \leq u \leq 1 \\[2ex] \dfrac{\operatorname{arsinh}\sqrt{-u}}{\sqrt{-u}} = \dfrac{\ln\left(\sqrt{-u} + \sqrt{1-u}\right)}{\sqrt{-u}} & @\ u < 0 \end{cases} \end{aligned}$$

(For the mathematically inclined reader: Note the relationship between the L-function and Stumpff function $c_2(u)$, see Sect. 7.4.8). We therefore finally obtain

$$\boxed{\frac{\Delta\tau}{2} x = (s_+ c_- - s_- c_+) \cdot L(u) - (s_+ c_+ - s_- c_-)} \quad @\ \text{Keplerian orbits} \qquad (8.2.17)$$

with $u = (s_+ c_- - s_- c_+)^2 x$. Note that for $x = 0$ due to $L(0) = 1$ and $c_\pm(0) = 1$ the right hand side vanishes and so does the left hand side.

Owing to the above dimensionless formulation of Lambert's equation, Lambert's theorem can be reduced to $\Delta\tau = f\left(\frac{c}{r_1 + r_2}, \frac{c}{a}\right)$ or $\frac{c}{a} = g\left(\Delta\tau; \frac{c}{r_1 + r_2}\right)$, respectively, as displayed in Fig. 8.8. Because $0 \leq |r_2 - r_1| \leq c \leq r_2 + r_1$ the parameter $c/(r_1 + r_2)$ is restricted to

$$0 \leq \frac{c}{r_1 + r_2} \leq 1$$

and according to Eq. (8.2.8)

$$4a \geq r_1 + r_2 + c$$

Universal Semi-Latus Rectum

To determine the semi-latus rectum we further define the dimensionless numbers

$$\begin{aligned} \rho &:= (r_1 r_2 - \boldsymbol{r}_1 \boldsymbol{r}_2)/c^2 = \frac{1}{2}\left[1 - \left(\frac{r_1 - r_2}{c}\right)^2\right] \\ \lambda &:= 2(s_+ c_- + s_- c_+)^2 \end{aligned}$$

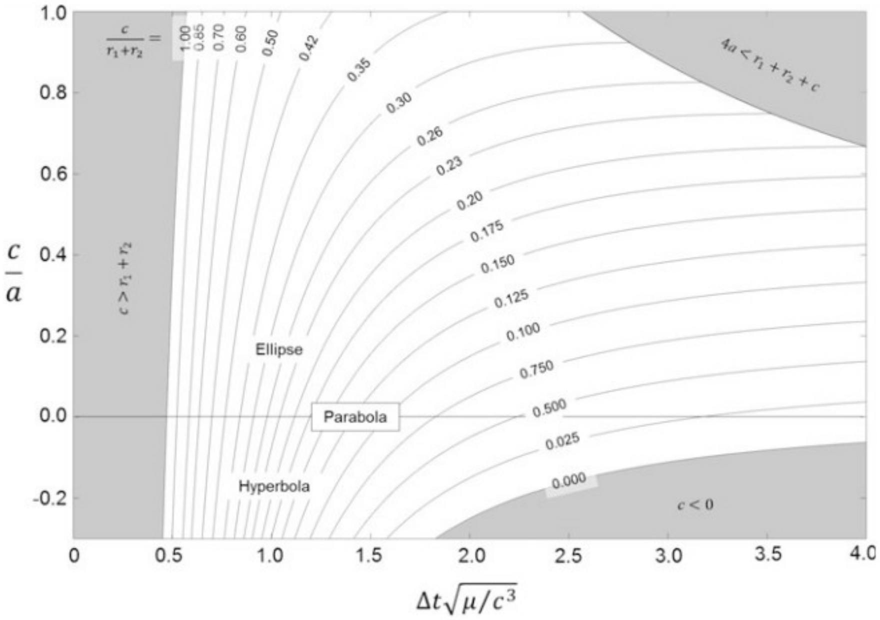

Fig. 8.8 Solution of Lambert's Problem in graphical form

which yields

$$r_1 r_2(1-\cos\Delta\theta) = r_1 r_2 - r_1 r_2 \cos\Delta\theta = r_1 r_2 - \boldsymbol{r}_1 \boldsymbol{r}_2 = c^2\rho$$

$$\begin{aligned} a[1-\cos(\alpha+\beta)] &= 2a\sin^2\frac{\alpha+\beta}{2} = 2a\left(\sin\frac{\alpha}{2}\cos\frac{\beta}{2} + \cos\frac{\alpha}{2}\sin\frac{\beta}{2}\right)^2 \\ &= 2c(s_+c_- + s_-c_+)^2 = c\lambda \qquad @ \quad x \geq 0 \end{aligned}$$

$$\begin{aligned} a[1-\cosh(\gamma+\delta)] &= -2a\sinh^2\frac{\lambda+\delta}{2} = -2a\left(\sinh\frac{\gamma}{2}\cosh\frac{\delta}{2} + \cosh\frac{\gamma}{2}\sinh\frac{\delta}{2}\right)^2 \\ &= 2c(s_+c_- + s_-c_+)^2 = c\lambda \qquad @ \quad x \leq 0 \end{aligned}$$

We therefore have from Eqs. (8.2.13), (8.2.14), and (8.2.16) for the short path of any Keplerian orbit

$$p = c\rho\lambda \qquad @\text{ Keplerian orbit} \quad \textbf{semi-latus rectum} \tag{8.2.18}$$

With this and some trigonometric expansions we find the following calculation scheme:

/4 A #! ! 8! !@@@

Calculation Scheme for an Universal Solution to Lambert's Problem

0. Given $\boldsymbol{r}_1, \boldsymbol{r}_2$, Δt, and μ

1. With $c := \sqrt{r_1^2 + r_2^2 - 2\boldsymbol{r}_1\boldsymbol{r}_2}$ calculate the dimensionless numbers

$$\Delta\tau = \Delta t\sqrt{\frac{\mu}{c^3}}$$

$$s_\pm = \frac{1}{2}\sqrt{\frac{r_1 + r_2}{c} \pm 1}$$

2. With $c_\pm := \sqrt{1 - xs_\pm^2}$ and $u = (s_+c_- - s_-c_+)^2 x$ solve the equation

$$\frac{\Delta\tau}{2} \cdot x + s_+c_+ - s_-c_- = (s_+c_- - s_-c_+)L(u)$$

for x with Newton's method or for difficult hyperbolic orbits $x < 0$ with the bisection method. (If $x = c/a < 0$ hyperbola, if $x = 0$ parabola, if $x > 0$ ellipse)

3. Calculate the dimensionless number and vectors

$$\lambda = 2(s_+c_- + s_-c_+)^2$$

$$\boldsymbol{\rho}_1 = \boldsymbol{r}_1/c,\ \boldsymbol{\rho}_2 = \boldsymbol{r}_2/c$$

$$\rho = \rho_1\rho_2 - \boldsymbol{\rho}_1\boldsymbol{\rho}_2 = \frac{1}{2}\left[1 - \left(\frac{r_1 - r_2}{c}\right)^2\right]$$

With this the eccentricity vector (see Eq. (8.2.3)) and the dimensionless terminal velocities (see Eq. (8.2.2)) and transfer path eccentricity (see Eq. (8.2.3)) are determined as

$$\boldsymbol{v}_{1+}\sqrt{\frac{c}{\mu}} = \frac{\sqrt{\rho\lambda}}{|\boldsymbol{\rho}_1 \times \boldsymbol{\rho}_2|}\left[\boldsymbol{\rho}_2 - \boldsymbol{\rho}_1 + \frac{1}{\lambda}\hat{\boldsymbol{\rho}}_1\right]$$
$$\boldsymbol{v}_{2-}\sqrt{\frac{c}{\mu}} = \frac{\sqrt{\rho\lambda}}{|\boldsymbol{\rho}_1 \times \boldsymbol{\rho}_2|}\left[\boldsymbol{\rho}_2 - \boldsymbol{\rho}_1 - \frac{1}{\lambda}\hat{\boldsymbol{\rho}}_2\right]$$
$$\boldsymbol{e} = (1 - x\rho_1)\hat{\boldsymbol{\rho}}_1 - \frac{c}{\mu}(\boldsymbol{\rho}_1\boldsymbol{v}_{1+})\boldsymbol{v}_{1+}$$

8.2.4 *Minimum Effort Lambert Transfer*

Having analyzed the Lambert transfer orbit we will now finally turn to the most important problem from a practical point of view: Given two transfer points P_1 and P_2 with state vectors $(\boldsymbol{r}_1, \boldsymbol{v}_{1-})$ and $(\boldsymbol{r}_2, \boldsymbol{v}_{2+})$ what is the Lambert transfer orbit with the least propulsion effort? So the additional scalar specification here is the least effort equaling a minimum total $\Delta\boldsymbol{v}$ as given in Eq. (8.2.4). In order to determine that transfer orbit, we recall the two orbital equations at P_1 and P_2

$$\begin{aligned} r_1 + \boldsymbol{e}\boldsymbol{r}_1 &= a\left(1 - e^2\right) \\ r_2 + \boldsymbol{e}\boldsymbol{r}_2 &= a\left(1 - e^2\right) \end{aligned}$$

These conditional equations leave one of the three orbital elements undetermined. It is this undetermined element that provides us with the freedom to optimize the transfer orbit. We choose as the free element the orientation of the conic orbit, position vector $\boldsymbol{r}_1$. We denote this angle as

$$\varphi = \angle(\boldsymbol{e}, \boldsymbol{r}_1)$$

We therefore have $\hat{\boldsymbol{e}}\hat{\boldsymbol{r}}_1 = \cos\varphi$ and for geometrical reasons $\hat{\boldsymbol{e}}\hat{\boldsymbol{r}}_2 = \cos(\Delta\theta - \varphi)$. From the first equation we then have

$$p = a\left(1 - e^2\right) = r_1(1 + e\cos\varphi) \tag{8.2.19}$$

Subtracting from this equation the second one from above yields

$$r_1 - r_2 + e r_1 \cos\varphi - e r_2 \cos(\Delta\theta - \varphi) = 0$$

from which we derive

$$e = \frac{r_2 - r_1}{r_1\cos\varphi - r_2\cos(\Delta\theta - \varphi)} \tag{8.2.20}$$

Hence, we have derived the orbital elements a, e and p as a function of the orientation of the transfer orbit $\varphi = \angle(\boldsymbol{e}, \boldsymbol{r}_1)$. We insert this result into Eq. (8.2.2) and find with Eq. (8.2.4) the total Δv as a function of φ. Employing an optimization algorithm that does "minimize $\Delta v(\varphi)$", one finds the optimal φ and from the above equations also the optimal orbital elements.

8.3 Hohmann Transfer

We have already seen that a minimum energy Lambert orbit does not necessarily imply a minimum energy Lambert transfer because for that we also have to take into account the delta-v of the two impulse maneuvers into and out of the transfer orbit. So we are now seeking for a minimum energy Lambert transfer. This is the Hohmann transfer.

The general problem statement is this: What is the transfer orbit between two Keplerian orbits with the least propulsion demand? Because this general problem is quite complex, we will restrict ourselves to the following conditions:

- The two orbits are elliptic and coplanar.
- The orbital bodies revolve around the same central body with the same orientation (co-revolving), so their angular momentum vectors are collinear, only their magnitudes are different.
- The orbital elements are such that the two orbits nowhere touch or cross each other.

Due to the last constraint we can define an inner and an outer orbit, which we will denote by the symbols • and O.

Throughout this section we assume in addition that the transfers are achieved with impulsive thrust maneuvers, so-called *boosts* or *kick-burns* with $F_* \gg F_{ext}$ (see Sect. 2.3) so that the created Δv is determined solely by the thrust characteristics. This assumption is in general valid for today's chemical propulsion engines for orbit control.

8.3.1 The Minimum Principle

As an example for an orbit transfer, let us consider the situation where we are in an elliptic LEO, and we want to get from any point on this orbit into an elliptic geosynchronous orbit, GSO (a GSO is in sync with Earth's rotation only on the average of an orbit) (see Fig. 8.9). A kick-burn in LEO will first take us into an elliptic transfer orbit, the so-called GSO transfer orbit (GTO). This transfer orbit has to cross or touch the GSO at some point. Once we are at the crossing point, a second kick-burn would bring us into GSO.

Fig. 8.9 The GTO as a Hohmann transfer orbit from an elliptic LEO to an elliptic GSO (not to scale)

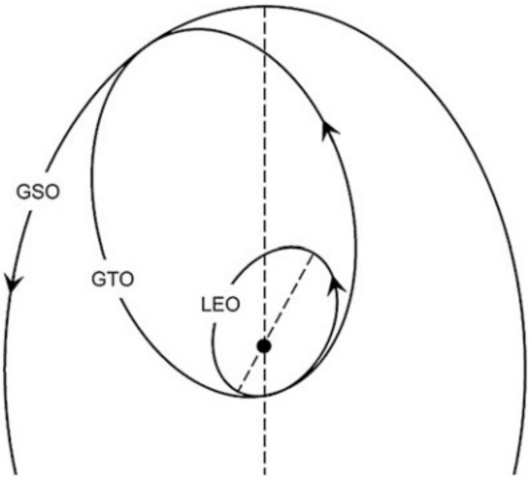

The general problem statement is this: What is the transfer orbit between two Keplerian orbits with the least propulsion demand? Because this general problem is quite complex, we will restrict ourselves to the following conditions:

- The two orbits are elliptic and coplanar.
- The orbital bodies revolve around the same central body with the same orientation (co-revolving), so their angular momentum vectors are collinear, only their magnitudes are different.
- The orbital elements are such that the two orbits nowhere touch or cross each other.

Due to the last constraint we can define an inner and an outer orbit, which we will denote by the symbols • and O.

Throughout this section we assume in addition that the transfers are achieved with impulsive thrust maneuvers, so-called *boosts* or *kick-burns* with $F_* \gg F_{ext}$ (see Sect. 2.3) so that the created Δv is determined solely by the thrust characteristics. This assumption is in general valid for today's chemical propulsion engines for orbit control.

8.3.1 The Minimum Principle

As an example for an orbit transfer, let us consider the situation where we are in an elliptic LEO, and we want to get from any point on this orbit into an elliptic geosynchronous orbit, GSO (a GSO is in sync with Earth's rotation only on the average of an orbit) (see Fig. 8.9). A kick-burn in LEO will first take us into an elliptic transfer orbit, the so-called GSO transfer orbit (GTO). This transfer orbit has to cross or touch the GSO at some point. Once we are at the crossing point, a second kick-burn would bring us into GSO.

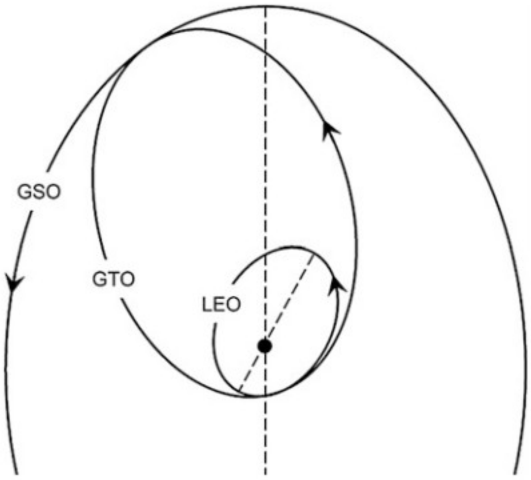

Fig. 8.9 The GTO as a Hohmann transfer orbit from an elliptic LEO to an elliptic GSO (not to scale)

Such a Hohmann transfer orbit is shown in Fig. 8.9. A Hohmann transfer of course also works the other way round, i.e., for a transfer from an outer to an inner elliptic orbit. So we have arrived at an answer to our optimization problem for a given starting point on the initial orbit. This leaves open the answer to the following question: At which point on the initial orbit should we perform the kick-burn to optimize the overall transfer between two elliptic orbits. One could presume that due to $\Delta\varepsilon \propto \boldsymbol{v} \cdot \Delta\boldsymbol{v}$ a first kick-burn at the highest orbital velocity, i.e., at the periapsis, would always be a good choice. But this neglects the second kick-burn at low speeds at the apoapsis of the transfer ellipse. A survey of the transfer at different orbital elements shows that the ratio of the eccentricities of the inner and the outer orbit is important. In general one can say that if the inner orbit has a higher eccentricity than the outer orbit, the optimal transfer varies, but

> If the inner orbit has a lower eccentricity than the outer orbit, the transfer to or from the apoapsis of the outer orbit requires the least delta-v budget.

The Earth–Mars transfer is a nice example for this rule, because here $e_{earth} = 0.0167$ and $e_{mars} = 0.0934$. Figure 8.10 shows the delta-v budget of the transfer as a function of the orbit angle of the transition point on the Mars orbit.

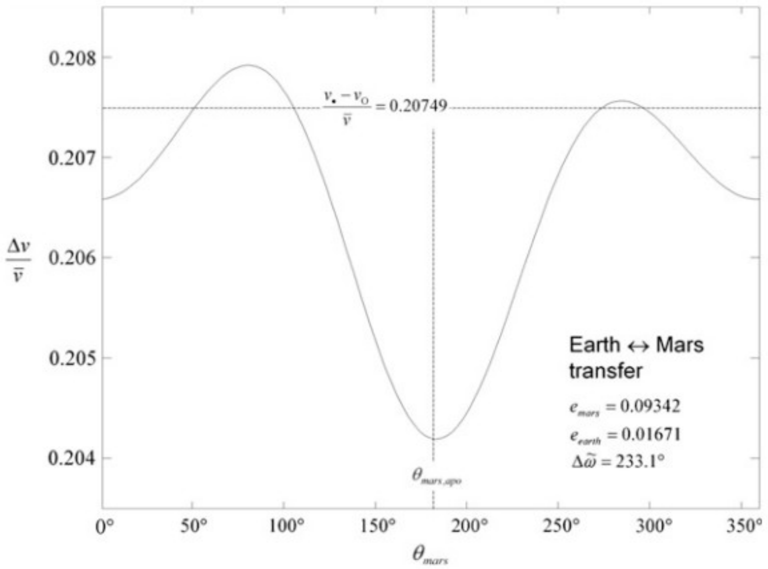

Fig. 8.10 Normalized delta-v budget for Earth ↔ Mars Hohmann transfers as a function of the orbit angle of the intersection with Mars orbit. The angle between their lines of apsis is 233.1°. The horizontal dashed line is the limiting value if both orbits were circular. Note the stretch of the *y*-scale

The transfer optimization problem gets even more intricate if arbitrary angles between the lines of apsides, or non-coplanar, or crossing orbits are considered. Because these cases become too complex to lay them out in a textbook, we skip them. In the following, we rather focus on the transfer between circular and near-circular orbits, which are of great practical value.

8.3.2 *Transfer Between Circular Orbits*

Hohmann transfers are specifically interesting between two circular orbits. Planetary orbits are mostly circular because at a given orbital energy they minimize atmospheric drag and provide steady orbit conditions, or because in the geostationary orbit the satellite has to rotate exactly in accord with the Earth. For circular orbits $r_\bullet = a_\bullet = const$ and $r_\bigcirc = a_\bigcirc = const$. It is easy to find out the metric orbital elements of a Hohmann transfer orbit between circular orbits. For the semi-major axis of this Hohmann transfer orbit, the following is obviously applicable:

$$a_H = \frac{1}{2}(a_\bullet + a_\bigcirc) \tag{8.3.2}$$

The transfer time is exactly half a period of the transfer ellipse, so according to Eq. (7.4.12)

$$t_H = \pi\sqrt{\frac{a_H^3}{\mu}} \tag{8.3.3}$$

The transfer ellipse with its two degrees of freedom a_H and e_H is completely and unambiguously determined by the boundary condition $r_{H,per} = a_H(1 - e_H)$ (see Eq. (7.4.6)) and by $r_{H,per} = a_\bullet$ and $r_{H,apo} = a_\bigcirc$. As for the semi-minor axis $b_H = a_H\sqrt{1 - e_H^2}$ applies, and the other metric orbital elements can be easily derived:

$$e_H = \frac{a_\bigcirc - a_\bullet}{a_\bigcirc + a_\bullet} \tag{8.3.4}$$

$$b_H = \sqrt{a_\bullet a_\bigcirc} \tag{8.3.5}$$

Let us determine the Δv demand for a Hohmann transfer between two circular orbits. According to Eq. (2.4.1), this is calculated for two Hohmann transfer kick-burns as the sum of the individual amounts (see Fig. 8.11):

$$\begin{aligned} \Delta v &= \left\{ \begin{matrix} |\Delta v_{\bullet\to H}| + |\Delta v_{H\to\bigcirc}| & @ \;\; \bullet \to \bigcirc \\ |\Delta v_{\bigcirc\to H}| + |\Delta v_{H\to\bullet}| & @ \;\; \bigcirc \to \bullet \end{matrix} \right\} \\ &= (v_{H\bullet} - v_\bullet) + (v_\bigcirc - v_{H\bigcirc}) \;@\;\; \bullet \leftrightarrow \bigcirc \end{aligned}$$

Fig. 8.11 Hohmann transfer between an inner and an outer circular orbit

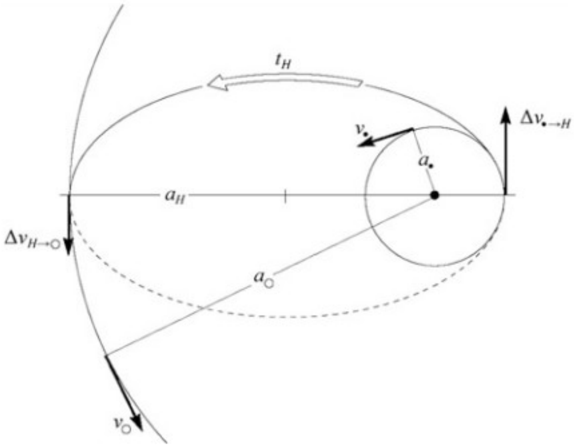

The latter holds because $|\Delta v_{\bullet\to H}| = |\Delta v_{H\to\bullet}| = v_{H\bullet} - v_\bullet$ and $|\Delta v_{O\to H}| = |\Delta v_{H\to O}| = v_O - v_{HO}$. According to Eq. (8.3.4) $(1+e_H)/(1-e_H) = a_O/a_\bullet$ and therefore from Eq. (7.4.10) follows $v_{H\bullet} = v_{per} = v_\bullet\sqrt{a_O/a_H}$ and $v_{HO} = v_{apo} = v_O\sqrt{a_\bullet/a_H}$. Inserting this into the above equation yields the two contributions

$$\begin{aligned} |\Delta v_{\bullet\leftrightarrow H}| &= v_\bullet\left(\sqrt{\frac{a_O}{a_H}} - 1\right) \\ |\Delta v_{O\leftrightarrow H}| &= v_O\left(1 - \sqrt{\frac{a_\bullet}{a_H}}\right) \end{aligned} \tag{8.3.6}$$

which are shown in Fig. 8.12. So, in total we get

$$\Delta v = |\Delta v_{\bullet\leftrightarrow H}| + |\Delta v_{H\leftrightarrow O}| = v_\bullet\left(\sqrt{\frac{a_O}{a_H}} - 1\right) + v_O\left(1 - \sqrt{\frac{a_\bullet}{a_H}}\right) \tag{8.3.7}$$

If the terms in Eq. (8.3.7) are arranged skillfully and extended by $\sqrt{\mu} = v_\bullet\sqrt{a_\bullet} = v_O\sqrt{a_O}$, we get

$$\boxed{\Delta v = (v_\bullet - v_O)\left(\frac{\sqrt{a_\bullet} + \sqrt{a_O}}{\sqrt{a_H}} - 1\right) < v_\bullet - v_O} \quad @\,\bullet \leftrightarrow O \tag{8.3.8}$$

Equation (8.3.8) is valid for both transfer directions, i.e., • → O and O → •.

Why is There a Bump?

The two contributions exhibit quite different shapes. Why? And why does $\Delta v_{O\leftrightarrow H}$ have a bump? The first (transfer injection) boost increases continuously with increasing distance between the orbits. This is quite easy to understand, as the semi-major axis a_H of the transfer orbit is determined by this distance, and the

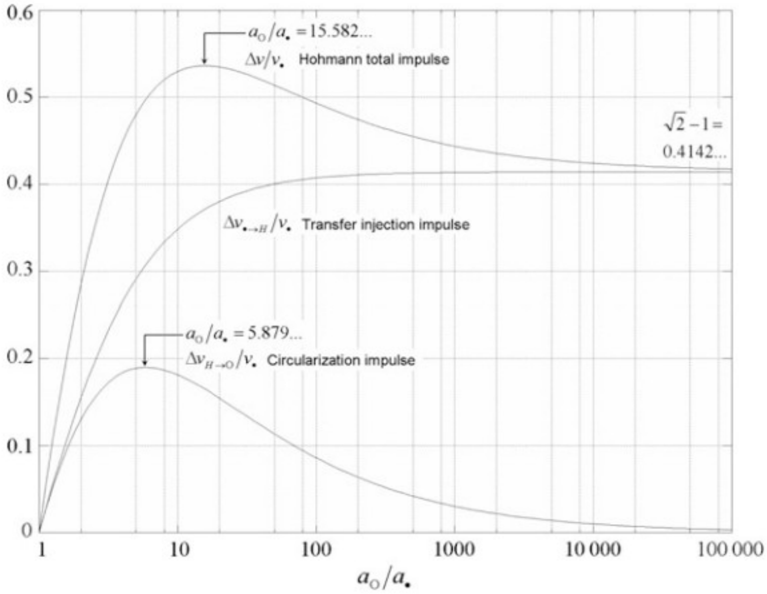

Fig. 8.12 Propulsion demand for Hohmann transfers from an inner to an outer orbit

transfer orbital energy also increases non-linearly with $\varepsilon = -\mu/2a$ according to Eq. (7.3.19). The second (circularization) boost, however, has a maximum at $a_O/a_\bullet = 5.879362...$ (exercise, Problem 8.4a). The reason is as follows. At the one extreme, if both orbits are in close vicinity to each other, i.e., $\Delta a = a_O - a_\bullet \to 0$, we havewith

$$a_H = \frac{1}{2}(a_\bullet + a_O) = a_\bullet\left(1 + \frac{1}{2}\frac{\Delta a}{a_\bullet}\right)$$

and according to Eq. (8.3.6)

$$|\Delta v_{H\leftrightarrow O}| \approx v_O\left(1 - 1\bigg/\sqrt{1 + \frac{1}{2}\frac{\Delta a}{a_\bullet}}\right) \approx \frac{1}{4}\frac{\Delta a}{a_\bullet}v_0$$

So, for $\Delta a = 0$ the second boost of course is zero and increases roughly linearly with increasing distance Δa. At the other extreme, when $a_O \to \infty$, then $a_H, \Delta a \to \infty$ as well, and we have for the second boost at infinite distance

$$|\Delta v_{H\leftrightarrow O}| = v_O\left(1 - \sqrt{\frac{a_\bullet}{a_H}}\right) \approx v_0\left(1 - \sqrt{\frac{2a_\bullet}{\Delta a}}\right) \approx v_0 \approx 0$$

So, because at infinity the Hohmann and the target orbital velocity is zero the transfer boost between the two vanishes as well. Because the second boost is positive between these limiting cases it must exhibit a maximum somewhere in between. Because of this maximum of the second boost, the total Hohmann transfer also has a maximum $\Delta v_{\max} \approx 0.536258 \cdot v_\bullet$ at $a_O / a_\bullet = 15.58172\ldots$ (exercise, Problem 8.4b).

Adjacent Circular Orbits

For a Hohmann transfer between adjacent circular orbits, $a_O \approx a_\bullet$, we can approximate (exercise, Problem 8.2)

$$\frac{\sqrt{a_\bullet} + \sqrt{a_O}}{\sqrt{a_H}} - 1 \;\approx 1 - \frac{1}{16}\left(\frac{a_O - a_\bullet}{a_\bullet}\right)^2 \to 1 \tag{8.3.9}$$

hence

$$\boxed{\Delta v \approx v_\bullet - v_O} \quad @\, a_O \approx a_\bullet \tag{8.3.10}$$

> *Example*
>
> For $a_O / a_\bullet \leq 2.5$, i.e., for orbits with altitudes up to 10,000 km above Earth's surface, the error due to Eq. (8.3.10) is smaller than 5.1%.

Distant Circular Orbits

For circular orbits that are distant from each other, we rewrite Eq. (8.3.8)

$$\frac{\Delta v}{v_\bullet} = \left(1 - \sqrt{\frac{a_\bullet}{a_O}}\right)\left(\sqrt{2}\frac{1 + \sqrt{a_\bullet / a_O}}{\sqrt{1 + a_\bullet / a_O}} - 1\right) = \sqrt{2}\frac{1 - a_\bullet / a_O}{\sqrt{1 + a_\bullet / a_O}} - 1 + \sqrt{\frac{a_\bullet}{a_O}}$$

For these distant orbits $a_\bullet / a_O \ll 1$ and we can approximate

$$\frac{\Delta v}{v_\bullet} = \sqrt{2} - 1 + \sqrt{\frac{a_\bullet}{a_O}} - \frac{3}{\sqrt{2}}\frac{a_\bullet}{a_O} + \cdots \tag{8.3.11}$$

Using $v = \sqrt{\mu / a}$ we therefore obtain in the limit of infinite distances

$$\Delta v = \left(\sqrt{2} - 1\right) v_\bullet + v_O \quad @\, a_O / a_\bullet \to \infty \tag{8.3.12}$$

Between these two limiting cases the expression $\left(\sqrt{a_\bullet} + \sqrt{a_O}\right) / \sqrt{a_H} - 1$ is strictly monotonously decreasing (see Fig. 8.13), and that is why in Eq. (8.3.8) the inequality strictly holds.

Hohmann Transfer as a Special Case of the Minimum Energy Lambert Transfer

Let us consider the case of a Lambert transfer orbit (see Sect. 8.2) between two circular orbits without specified transition points. We examine the minimum energy transfer orbit of Sect. 8.2.2, pick any P_1 on the initial circle, and, as shown in Fig. 8.14, move P_2, because it is arbitrary, until $\Delta\theta = \pi$. In this special case $c = r_1 + r_2$ and both the occupied and empty focal points lie on the chord. This is

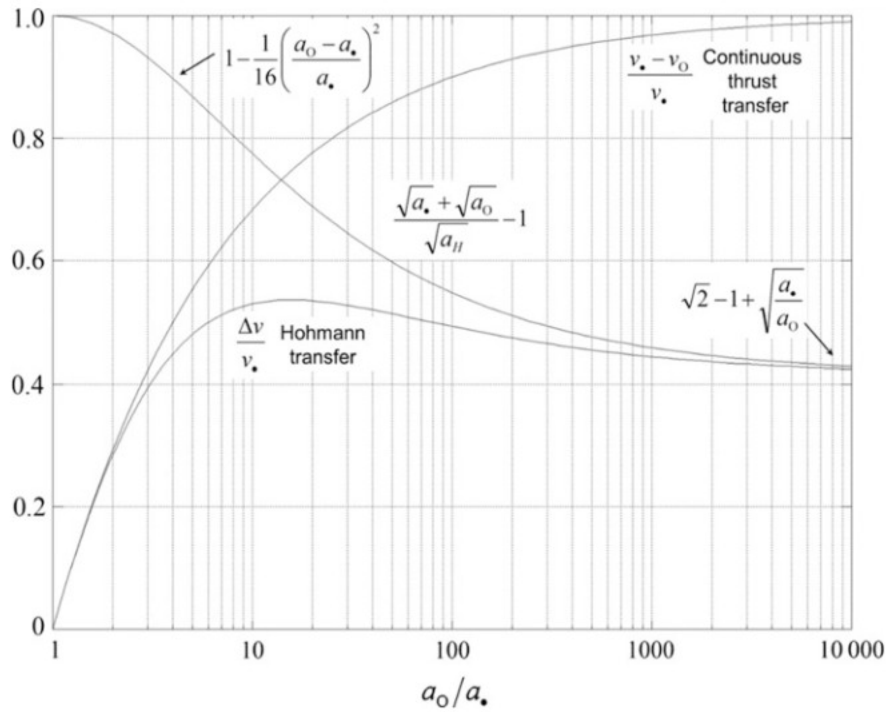

Fig. 8.13 The terms of the Eq. (8.3.8) and their limiting values. The factor $(v_\bullet - v_O)/v_\bullet$ is the propulsion demand of a continuous thrust transfer (see Sect. 8.4.5)

the condition of a minimum energy transfer orbit and it obviously is a Hohmann transfer orbit. By this procedure, we in a different way have verified that the Hohmann transfer is a minimum energy transfer with maximum transfer time.

Note that for this special Lambert transfer the eccentricity is no longer defined. In fact, whatever e might be, the transfer time is independent of e! For the Hohmann transfer, this additional degree of freedom is taken to adjust the tangents of the transfer ellipse to those of the circles at the touching points so that delta-v becomes minimal.

8.3.3 *Transfer Between Near-Circular Orbits*

We now consider Hohmann transfers between two coplanar, co-revolving, non-crossing ellipses, with small eccentricities $e_\bullet, e_O \ll 1$. The detailed analytical approximations are quite convoluted due to $e_\bullet e_O$ cross-terms, but the general upshot is that the dependency of delta-v as a function of $e_\bullet, e_O$ is very weak. This can be observed, for instance, in Fig. 8.15 where for Hohmann transfers between two apsides points on two orbits with $a_O/a_\bullet = 1.52365$ (Earth–Mars) the normalized delta-v is depicted as a function of $e_\bullet = e_O$ numerically.

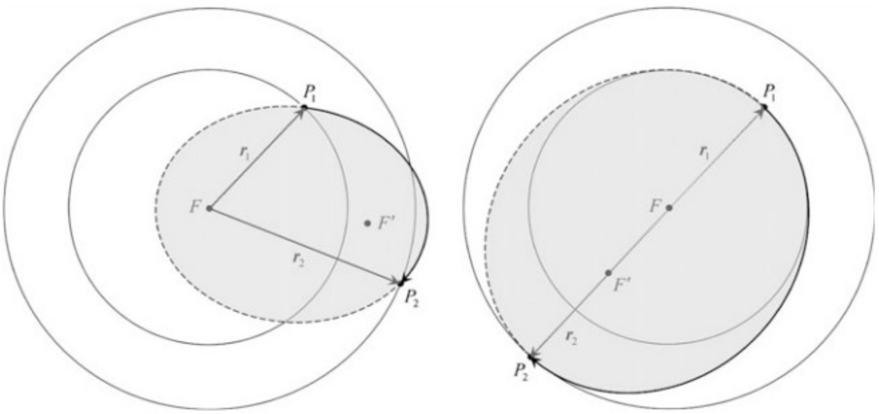

Fig. 8.14 A Lambert transfer between arbitrary points on two circular orbits (left) and the crossover to the Hohmann transfer for $\Delta\theta = \pi$ (right)

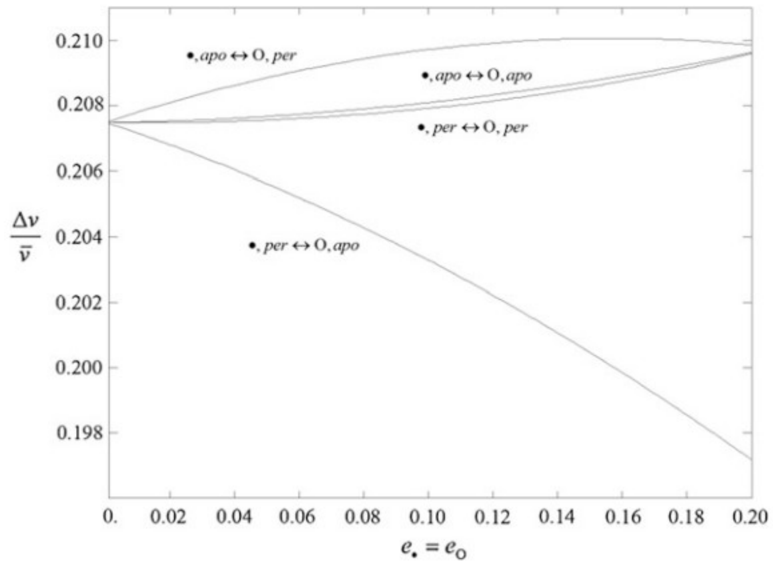

Fig. 8.15 The delta-v budget as a function of the common eccentricities. The dependencies are given for orbits with $a_O/a_\bullet = 1.52365$ (Earth–Mars). Note the stretched Δv-scale

For $e \leq 0.1$ the variations are smaller than 2% of their corresponding circular orbits value $(v_\bullet - v_O)/\bar{v} = 0.20749$. As we can see from Fig. 8.10 this is in accordance with the Earth ↔ Mars transfer, $e_{Mars} = 0.09342$ and $e_{earth} = 0.01671$, even if the transfers would be performed from any position of the orbit. So for all practical purposes, we get the same result as for a transition between two circular orbits.

$$\Delta v \approx (v_\bullet - v_\mathrm{O})\left(\frac{\sqrt{a_\mathrm{O}} + \sqrt{a_\bullet}}{\sqrt{(a_\mathrm{O} + a_\bullet)/2}} - 1\right) \qquad @\ e_\bullet, e_\mathrm{O} \leq 0.1 \tag{8.3.13}$$

Nevertheless, owing to the Oberth effect as discussed in Sect. 8.1.3, the general transfer rule is

> The most delta-v efficient Hohmann transfer between a near-circular or even elliptic inner orbit and any outer orbit is always from or to the periapsis of the inner orbit.

8.3.4 Sensitivity Analysis

From an energy point of view Hohmann transfers are the most favorable two-impulse transfer orbits. But they also have some disadvantages. They can be very sensitive to small inaccuracies of the transfer injection impulse. Let us have a closer look at this dependence for transfers between circular orbits.

Let r_{per} and r_{apo} be the periapsis and the apoapsis radius of the transfer orbit respectively (for convenience, we drop the index H to indicate the Hohmann transfer orbit). The periapsis is determined by the initial orbit radius. From the vis-viva Eq. (7.2.15) we get for the velocity on the Hohmann transfer orbit at the periapsis

$$v_{per}^2 = \frac{2\mu}{r_{per}} - \frac{\mu}{a} = \frac{\mu_\oplus}{a}\frac{1+e}{1-e}$$

If the transfer injection burn has a small error dv_{per} at the fixed periapsis $r_{per} = const$ we get

$$2v_{per}dv_{per} = \frac{\mu}{a^2}da$$

from which follows

$$\frac{da}{a} = 2\frac{av^2}{\mu}\frac{dv_{per}}{v_{per}} = 2\frac{1+e}{1-e}\frac{dv_{per}}{v_{per}} \tag{8.3.14}$$

So, a thrust error generates a certain variation of the semi-major axis. Now we want to know how this affects the position of the apoapsis. We start with Eq. (7.4.7)

$$r_{apo} = a(1+e)$$

Its change in position is determined by differentiation

$$\frac{dr_{apo}}{r_{apo}} = \frac{da}{a} + \frac{de}{1+e} \tag{8.3.15}$$

Both parameters a and e are not independent of each other, but linked by the constancy of the periapsis of the initial orbit:

$$const = r_{per} = a(1-e)$$

After differentiating this equation, we see how their changes depend on each other

$$de = (1-e)\frac{da}{a} \tag{8.3.16}$$

So together with Eq. (8.3.15), we get

$$\frac{dr_{apo}}{r_{apo}} = \frac{2}{1+e}\frac{da}{a}$$

and finally with Eq. (8.3.14), we get

$$\boxed{\frac{dr_{apo}}{r_{apo}} = \frac{4}{1-e}\frac{dv}{v} = \frac{4a}{r_{per}}\frac{dv_{per}}{v_{per}} = 2\left(1+\frac{r_{apo}}{r_{per}}\right)\frac{dv_{per}}{v_{per}}} \tag{8.3.17}$$

That is, for a given injection burn error dv_{per}, the relative target point accuracy decreases with increasing transfer distances.

> *Example 1*
>
> Let us examine the Hohmann transfer from an initial LEO orbit (h = 400 km) to GEO. Because of $r_{GEO}/r_{LEO} = 6.232$, we get
>
> $$\frac{dr_{apo}}{r_{apo}} = 14.46\frac{dv_{per}}{v_{per}}$$
>
> A relatively small burn error of just 0.5% would lead to an inaccuracy in the target distance of 7.2%. That is, 3000 km deviation from the GEO orbit!

> *Example 2*
>
> Let us have a look at the Hohmann transfer from an initial LEO orbit (h = 400 km) to the Moon. Because of $r_{moon}/r_{earth} = 56.654$, we get
>
> $$\frac{dr_{apo}}{r_{apo}} = 115.3\frac{dv_{per}}{v_{per}}$$
>
> This means that the same small burn error of just 0.5% would lead to an inaccuracy in the target distance of 58%. That would just bring us to nirvana!

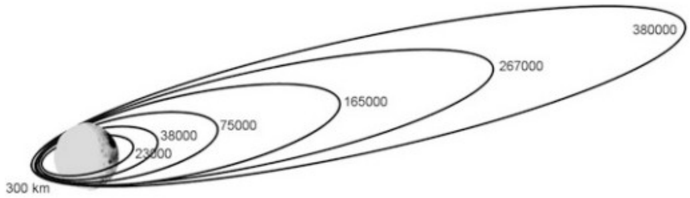

Fig. 8.16 The transfer trajectories of Chandrayaan-1 from Earth to Moon. The apogee (given in km) was increased step by step by successive kick-burns at 300 km perigee

Summary
Plain Hohmann transfers may be the most favorable transfer orbits from an energy point of view, but they implicate long transfer times and the large possible target errors.

However, already small deviations from an ideal Hohmann transfer and in-flight corrections strongly reduce these drawbacks as discussed in Sect. 9.3.2 for near-Hohmann transfers to Mars. In addition, if transfer time is not an issue, one can get a grip on the high sensitivity of highly elliptic transfer orbits by gradually raising the apogee of the transfer orbit to the final Hohmann transfer orbit. Thereby after each raising maneuver the next maneuver can be adjusted to the actual apogee. Figure 8.16 shows the transfer trajectory of the Indian Chandrayaan-1 Moon mission, which was launched on October 22, 2008 and arrived at the Moon 13 days, later where it was captured by Moon's gravitation into a highly elliptic Moon orbit.

8.4 Other Transfers

In this section we study some orbit transfers which are non-Hohmann, but nevertheless of practical value.

8.4.1 Parabolic Escape Transfer

In Sect. 7.4.4. we have studied parabolic orbits. They have the special property that the body reaches infinity at zero velocity. Therefore the simplest escape orbit is a parabolic orbit. We also saw that at a given orbital radius the escape velocity v_{esc} enables a spacecraft to go to infinity independent of the direction in which the final velocity vector points. Yet, the propulsion effect to achieve the escape velocity strongly depends on this direction (see Sect. 8.1.3). In Sect. 8.3.1 we saw that at a given orbital position with $\boldsymbol{r}_0, \boldsymbol{v}_0$ the least effort to achieve a predetermined $\boldsymbol{v}_{esc}$ is by the impulse maneuver $\Delta\boldsymbol{v}||\boldsymbol{v}_0$, i.e. a boost into the instantaneous flight direction in the initial orbit. From Eq. (7.3.16) this implies that the flight angle λ of an initial elliptic orbit with a, e and final parabolic orbit with p_p then is identical

$$\cos\gamma = \sqrt{\frac{a(1-e^2)}{r_0(2-r_0/a)}} = \sqrt{\frac{p_p}{2r_0}}$$

From this it follows that the semi-latus rectum of the parabolic escape orbit is

$$p_p = 2\frac{a(1-e^2)}{2-r_0/a} \quad @\ \Delta\boldsymbol{v}||\boldsymbol{v}_0 \tag{8.4.1}$$

The orientation of the escape orbit, i.e. its line of apsides is given from the parabolic orbit equation as

$$\angle(\boldsymbol{e}_p, \boldsymbol{r}_0) = \cos\,\theta_{p,0} = \frac{p_p}{r_0} - 1 \tag{8.4.2}$$

The initial elliptic orbit and the final parabolic orbit are coplanar.

Finally it would be interesting to know at which point on the ellipse an escape would be most efficient. We calculate Δv from the vis-viva equation to be

$$\Delta v = \sqrt{\mu}\left(\sqrt{\frac{2}{r_0}} - \sqrt{\frac{2}{r_0} - \frac{1}{a}}\right) \tag{8.4.3}$$

From this equation it is obvious that the required Δv lessens with smaller orbit radius. Therefore it is minimal at its periapsis and maximal at its apoapsis, although the differences are only slight for $e < 0.1$. Nevertheless, this result is a good example of the Oberth effect decribed in Sect. 8.1.3.

But, most importantly, the boost position solely determines the asymptotic direction of motion. So, the escape procedure is to first figure out the direction of asymptotic motion and measure it as an orbit angle relative to the line of apsides of the initial elliptic orbit. Call this angle θ_∞. Then $\theta_0 = \theta_\infty + \theta_{p,0}$ or

$$\cos(\theta_0 - \theta_\infty) = \cos\,\theta_{p,0} = \frac{p_p}{r_0} - 1 = \frac{2(1+e\,\cos\,\theta_0)^2}{1+2e\,\cos\,\theta_0 + e^2} - 1 \tag{8.4.4}$$

So, for a given θ_∞ the orbit angle θ_0 of the boost position is the root of this equation.

> *Example*
>
> We assume a circular orbit with $r = r_0 = a,\ e = 0$. At any point in orbit we want to transfer into an escape parabola with the least Δv effort. What then is the escape parabola?
>
> From Eq. (8.4.1) follows $p = 2r_0$ and from Eq. (7.3.16) $\sin\gamma = \sin 0° = 0 = \sin\theta \rightarrow \theta = 0°$. This means that the boost location constitutes the periapsis of the escape parabola.

8.4.2 Bi-elliptic Transfer

The maximum of the total delta-v for Hohmann transfers (see Fig. 8.12) at $a_O/a_\bullet = 15.582$ occurs because $\Delta v_{H\to O}$ achieves its maximum at $a_O/a_\bullet \approx 5.88$ for turning into the target orbit. This second delta-v contribution however vanishes for $r \to \infty$. This gives rise to the assumption that it might be possible to save propulsion demand with a total of three impulses by first escaping from the initial orbit to far out or even to infinity, and then turning back again to the coplanar target orbit (see Fig. 8.17).

We want to determine the propulsion demand for such a three-impulse transfer (a.k.a. *bi-elliptic transfer*). According to Eq. (8.3.12) the delta-v budget to a remote circular orbit with orbit radius $a_\times \gg a_O > a_\bullet$ and orbital velocity $v_\times = \sqrt{\mu/a_\times}$ is

$$|\Delta v_{\bullet\to\times}| \approx \left(\sqrt{2}-1\right)v_\bullet + v_\times$$

The delta-v budget for a kick-burn that brings us back into the target orbit in turn amounts to

$$|\Delta v_{\times\to O}| \approx \left(\sqrt{2}-1\right)v_O + v_\times$$

So, in total we have

$$\Delta v = |\Delta v_{\bullet\to\times}| + |\Delta v_{\times\to O}| \approx \left(\sqrt{2}-1\right)(v_\bullet + v_O) + 2v_\times \quad @\, a_\times \gg a_O \qquad (8.4.5)$$

Fig. 8.17 Schematic of a bi-elliptic transfer with an infinite intermediate orbit

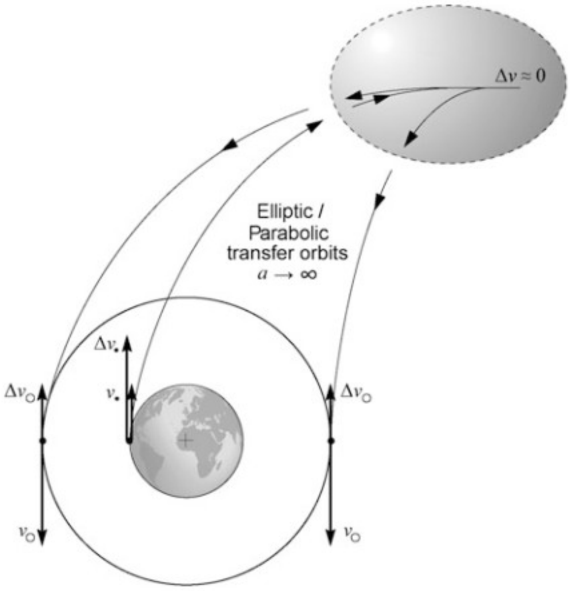

For a bi-parabolic transfer, when the intermediate orbit is infinitely far away, then $v_\times = 0$ and the total delta-v budget becomes minimal

$$\Delta v = \left(\sqrt{2} - 1\right)(v_\bullet + v_\mathrm{O}) \quad @\, a_\times / a_\mathrm{O} \to \infty \tag{8.4.6}$$

Equation (8.4.6) is illustrated in Fig. 8.18. For $a_\mathrm{O}/a_\bullet > 11.938765\ldots$ the propulsion demand is indeed more favorable with bi-parabolic transfers than with Hohmann transfers (exercise, Problem 8.5). A practical and hence important benefit of a bi-parabolic transfer is that at the outermost position, where $v \approx 0$, a change of the orbital plane, flight direction, or even the direction of rotation of the orbital curve may be done virtually without any propulsion demand. The serious drawback is that it takes increasingly longer to get farther out.

Break-Even Point with Hohmann Transfer

To ease this problem, it would be interesting to know at which intermediate orbit distance the bi-elliptic transfer starts to be advantageous to the Hohmann transfer. From Eqs. (8.4.5) and (8.3.11) we get the conditional equation $\left(\text{note that } O(v_\times) = O(v_\mathrm{O}^2/v_\bullet)\right)$

$$\Delta v_{bi} \approx \left(\sqrt{2} - 1\right)(v_\bullet + v_\mathrm{O}) + 2v_\times < \left(\sqrt{2} - 1\right)v_\bullet + v_\mathrm{O} - \frac{3}{\sqrt{2}}\frac{v_\mathrm{O}^2}{v_\bullet} \approx \Delta v_H$$

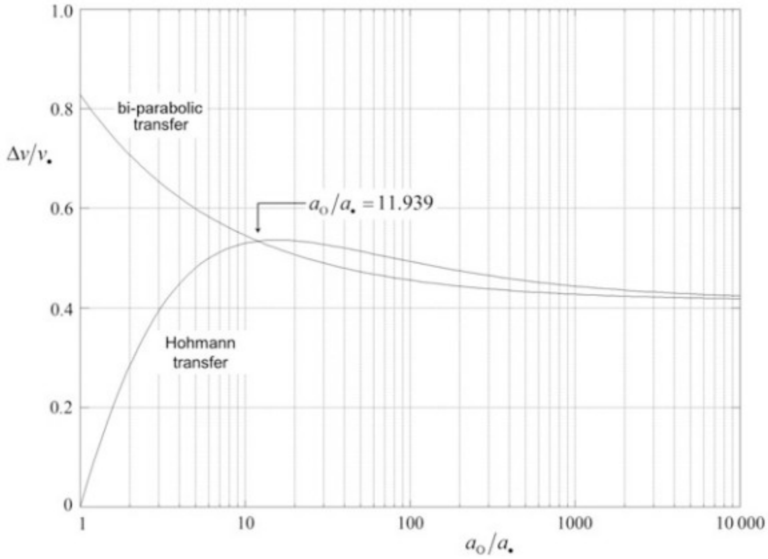

Fig. 8.18 Comparison between an infinite bi-parabolic transfer and a Hohmann transfer

This implies

$$\frac{v_\times}{v_\bigcirc} < 1 - \frac{1}{\sqrt{2}} - \frac{3}{2\sqrt{2}}\frac{v_\bigcirc}{v_\bullet}$$

and therefore

$$\frac{a_\times}{a_\bigcirc} > \frac{8}{\left[2\left(\sqrt{2}-1\right) - 3\sqrt{a_\bullet/a_\bigcirc}\right]^2} \tag{8.4.7}$$

For example let us assume three ratios $a_\bigcirc/a_\bullet = 20,\ 100,\ 500$. For this we find $a_\times > (322,\ 28.6,\ 16.60) \cdot a_\bigcirc = (6441,\ 2860,\ 8299) \cdot a_\bullet$.

Finally, one might question at which $a_\bigcirc/a_\bullet$ the bi-elliptic transfer is most advantageous to a Hohmann transfer, i.e.: When does $v_\times/v_\bigcirc$ become maximal, or when does the relation $a_\times/a_\bullet$ become minimal? The answer is easily calculated from Eq. (8.4.7) to be $a_\bigcirc/a_\bullet|_{opt} = \left[3/\left(\sqrt{2}-1\right)\right]^2 = 52.456$ and $a_\times/a_\bullet|_{\min} = 2445.8$. This optimum case corresponds roughly to a transition from an Earth parking orbit at altitude $h = 250\ km$ to the Moon, $a_{Moon}/a_{park} = 57.3$. Even at this near optimal situation one would need to travel to $a_\times = 2455 \cdot a_{park} = 42.8 \cdot a_{moon}$ to break even with a Hohmann transfer, not to talk about the travel time to get there and back to the Moon. This impractical situation, providing only humble delta-v advantage, is why bi-elliptic transfers are academic cases that are rarely employed.

8.4.3 Super-Synchronous Transfer Orbits

This section is about some special types of transfer orbits to GEO. A classical transfer orbit is the GEO transfer orbit (GTO), which is Hohmann transfer from a low Earth orbit to GEO. In practice, though, there is no LEO. The launch vehicle injects directly from its launch ascent trajectory into an elliptic orbit with apogee at GEO.

There are cases where other transfer orbits, so-called *super-synchronous (GEO) transfer orbit* (SSTO), a.k.a. Super-GTO, or GTO+, are superior to a regular GTO. A SSTO generally is an orbit that rather than a GEO synchronous orbit (GSO), which is any elliptic orbit (incl. GEO) with period of a sideral day (see Sect. 13.2), T_{GSO}, features $T_{SSTO} > T_{GSO}$. According to Eq. (7.4.12) this implies $a_{SSTO} > a_{GSO}$. There exist the following two types of SSTOs.

Bi-Elliptic Transfer

The delta-v benefit of a bi-elliptic transfer increases with the inclination to be changed at the apoapsis. This benefit is decisive for launches from launch sites at high latitude β, which equals the launch orbit inclination i (see Sect. 8.6.1), to GEO. For such a bi-elliptic transfer the upper stage of the rocket injects the satellite into an SSTO at the descending node of the launch trajectory, i.e. when crossing the equator. This injection burn includes the so-called minimum residual shutdown strategy (MRS) where any excess propellant is expended to reduce the inclination as far as possible, usually by 2–3°. The apogee of the SSTO usually is chosen to be twice to three times the GEO radius (see Fig. 8.19). At SSTO apogee an apogee boost places the spacecraft into a so-called interim transfer orbit (ITO) to GEO and at the same time zeroing the residual inclination. When arriving at GEO, a final burn brings the spacecraft into a so-called circular drift orbit, which is just little smaller than GEO to let the S/C drift from the burn position to its final longitude position. Having arrived there, tiny adjustment burns bring the S/C into its operational orbit state.

The key point with such SSTOs is that, even though the total Δv of all three burns may be larger, the satellite's propellant spending for the SSTO apogee boost and the final burns are lower than for a standard GTO. So, as long as the additional Δv of the rocket upper stage for SSTO is within its propellant budget, any reduction of the satellite's Δv is welcome to invest the saved propellant into later station-keeping and thus extend the satellite's orbital life time. This is why bi-elliptic SSTOs for transfering satellites into GEO are quite frequent for high latitude launch sites such as Cape Canaveral (Atlas) and Baikonur (Proton).

Common to all such SSTOs is the fact that the SSTO injection burn is performed at the descending node of the launch trajectory, which implies $\omega = 180°$ for the SSTO. Hence the tangent plane maneuver (see Sect. 8.1.4) at SSTO apogee into a ITO is a nodal transfer. The RAAN of the ITO is irrelevant because the inclination is usually nullified at SSTO apogee.

Hence both Δv_1 and Δv_2 are nodal transfers as determined by Eq. (8.1.20). The delta-v for the three burns are therefore as follows

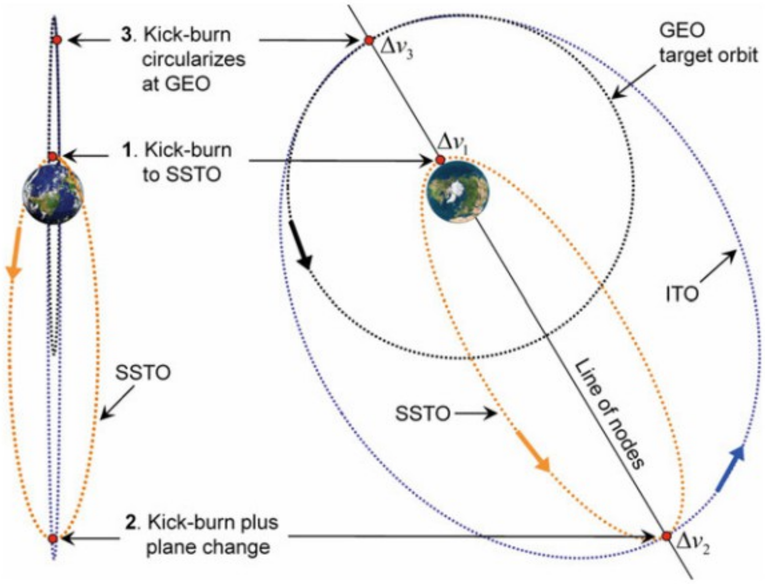

Fig. 8.19 A bi-elliptical super-synchronous transfer with its three kick-burns. Views are in Earth's equatorial plane (left) and normal to it (right)

$$\begin{aligned}\Delta v_1 &= \sqrt{v_{LEO-}^2 + v_{LEO+}^2 - 2v_{LEO-}v_{LEO+}\cos\Delta i_1}\\ \Delta v_2 &= \sqrt{v_{ITO-}^2 + v_{ITO+}^2 - 2v_{ITO-}v_{ITO+}\cos\Delta i_2}\\ \Delta v_3 &= v_{GEO-} - v_{GEO+}\end{aligned} \tag{8.4.8}$$

where the subscripts − and + indicate the values before and after the burn, respectively.

Case Study: The Super-Synchronous Transfer Orbit Mission of *Orion 1*

The Telstar 11 telecommunications satellite, operated by Loral Skynet, provided video for broadcast and business television, high-speed internet access and multimedia services, broadband and data networking to Europe, the United States to the Rocky Mountains and parts of Canada and Mexico.

It was launched as Orion 1 (a.k.a. Orion F1) on November 29, 1994, on an Atlas IIA rocket from Cape Canaveral (latitude 28.5°). The standard Atlas IIA GTO foresees a nodal transfer at the descending node, i.e. a GTO injection burn at 185 km altitude including an inclination change of about 2.8° (depending on the wind conditions during ascent and hence on the decisive propellant excess just before main engines cut off, MECO). A transfer burn at apogee, which is GTOs ascending node and a circularization burn including an inclination change of $\Delta i = 25°$ at the point where GTO touches the GEO, delivers the satellite into a so-called drift orbit. There the satellite drifts with a rate of 0.3°W per day into its final station-keeping box at 37.5°W. This earlier and standard Atlas IIA LEO→GTO→GEO sequence requires a GTO-injection boost of $\Delta v_{rocket} = 2.497\,\text{km/s}$ for the rocket's last stage and a $\Delta v_{sat} = 1.764\,\text{km/s}$ for the spacecraft. The latter would imply a lifetime in GEO to be less than 7 years, much less than required.

It was therefore decided that at the descending node the launcher should inject the spacecraft into a SSTO, in order to significantly reduce the later spacecraft's $\Delta v_{sat} = \Delta v_2 + \Delta v_3$ at the expense of a relatively small Δv_1 increase of the launcher. The SSTO chosen had the properties

$$\begin{aligned} r_{apo} &= 129{,}885\,\text{km}\\ r_{per} &= 6563.1\,\text{km}\\ i &= 25.7°\\ \Omega &= 173.6°\\ \omega &= 179.98°\end{aligned}$$

To bring the spacecraft from SSTO apogee into GEO, one apogee burn would change the inclination $\Delta i = 25^\circ$ and bring the S/C into the interim transfer orbit (ITO) to GEO, while a final burn at ITO perigee would circularize it into GEO. For safety reasons each burn was split into two, which in view of delta-v, however, are equal to the said two burns. According to Eq. (8.4.8) the delta-v for the three burns are as follows

$$\begin{aligned} v_{LEO-} &= 7.793\,\mathrm{km/s}, \\ v_{LEO+} &= 10.753\,\mathrm{km/s} \\ \Delta v_1 &= \sqrt{v_{LEO-}^2 + v_{LEO+}^2 - 2v_{LEO-}v_{LEO+}\cos(28.5^\circ - 25.7^\circ)} \\ &= 2.993\ \mathrm{km/s} \end{aligned}$$

$$v_{ITO-} = 0.5434\ \mathrm{km/s}, \quad v_{ITO+} = 1.227\ \mathrm{km/s}$$

$$\begin{aligned} \Delta v_2 &= \sqrt{v_{ITO-}^2 + v_{ITO+}^2 - 2v_{ITO-}v_{ITO+}\cos(25.7^\circ - 0.6^\circ)} \\ &= 0.770\ \mathrm{km/s} \end{aligned}$$

and

$$\begin{aligned} v_{GEO-} = 3.778\ \mathrm{km/s}, \quad v_{GEO+} = 3.075\ \mathrm{km/s} \\ \Delta v_3 = v_{GEO-} - v_{GEO+} = 0.703\ \mathrm{km/s} \end{aligned}$$

So, the Atlas rocket had to provide a $\Delta v_1 = 2.993$ km/s and hence an additional

$$\Delta v_{rocket+} = 2.993 - 2.497\ \mathrm{km/s} = 0.496\ \mathrm{km/s}$$

which was within the fuel budget. On the other hand, the total delta-v for the Orion satellite was $\Delta v_1 + \Delta v_2 = 0.770 + 0.703 = 1.473\,\mathrm{km/s}$ thus providing a benefit of

$$\Delta v_{sat+} = 1.473\,\mathrm{km/s} - 1.764\,\mathrm{km/s} = -291\,\mathrm{m/s}$$

This delta-v benefit extended Orion's lifetime for an additional 5 years, sufficient to meet the mission requirements.

Because any reduction of the satellite's total delta-v, $\Delta v_{sat} = \Delta v_2 + \Delta v_3$, is essential for the satellite's lifetime, Fig. 8.20 plots $\Delta v_2 + \Delta v_3$ for $\Delta i_1 = 0$ and different $\Delta i \equiv \Delta i_2$ as a function of SSTO apogee distance.

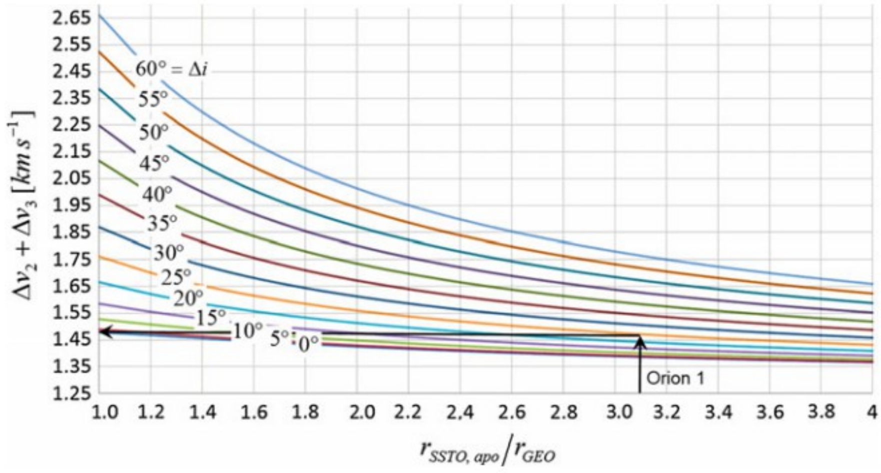

Fig. 8.20 The total delta-v of the satellite for a LEO → SSTO → ITO → GEO transfer as a function of relative SSTO apogee distance and for different inclination changes at the second kick-burn (SSTO apogee) and vanishing inclination change at the first and third kick-burn

Continuous Thrust Transfer

With today's trend to all-electric propulsion platforms (cf. Sect. 5.2.2, Subsection *Comparsion with other Thrusters*) there is the need for an efficient electric propulsion transfer from a SSTO, as provided by the launch vehicle, to GEO.

Early transfers were based on GTOs using a thrust arc strategy around apogee, similar to the chemical apogee-thrusts. Such a transfer requires about 150 days. In 1995, A. Spitzer proposed a thrust arc strategy based on SSTO (so-called Spitzer's scheme) with 16% less delta-v and a significantly reduced transfer time of about 110 days. Immediately thereafter it was found by C.R. Koppel that a continuous thrust strategy (CTS, see Sect. 8.4.5), i.e. continuously thrusting over all orbits, was even more practical. Although CTS requires a 17% larger delta-v than *Spitzer's scheme*, it reduces the transfer time to about 90 days, with an equivalent reduction of orbits passing the radiation of the Van Allen belts. This is why CTS for all-electric propulsion S/C today is the means of choice.

For a CTS the launch vehicle injects the S/C into a SSTO with typically 200 km × 60,000 km altitude and with a launch-specific non-zero RAAN and inclination. The electric thrust vector then is inertially oriented (star mode) with some out-of-plane component, such that in course of the transfer the in-plane component gradually removes the eccentricity (see Eq. (8.1.1b)) and the out-of plane component decreases the inclination to zero (see Eq. (8.1.1c), RAAN hence becomes irrelevant). The orbit thus gradually crosses over to the target GEO (see Fig. 8.21).

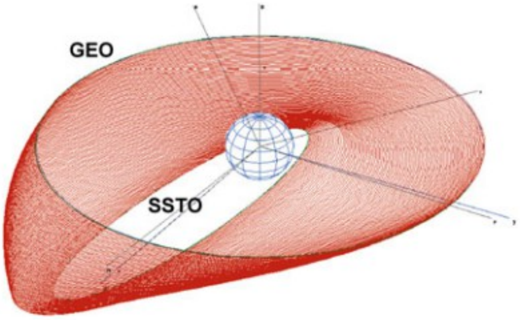

Fig. 8.21 Continuous thrust transfer from a SSTO (185 km × 6000 km, $i = 28^\circ$) to GEO (*Credit* C.R Koppel, 1999)

8.4.4 n-Impulse Transfers

Very generally it can be shown that for a transition between *any* two elliptic orbits the total delta-v budget for a three-impulse transfer (not necessarily bi-elliptic) might be smaller, but does not have to be, than that with the two-impulse Hohmann transfer. The above bi-elliptic transfer for $a_O/a_\bullet > 11.94$ is an example for this. In addition, it can also be shown that the total delta-v budget minimized by Hohmann or a three-impulse transfer cannot be further minimized by maneuvers with more than three impulses. So the Hohmann transfer or the minimum three-impulse transfer represents the absolute minimum for the propulsion demand in a two-body problem. But note that things are different in the three-body problem (see Note in Sect. 8.3.1).

8.4.5 Continuous Thrust Transfer

If we have electrical propulsion engines such as ion thrusters, thrust is low and continuous, so impulse transfers are impossible. An ion engine would rather have to fire permanently in the direction of motion (recall: $\Delta\varepsilon \propto \boldsymbol{v} \cdot \Delta\boldsymbol{v}$) to optimally but slowly spiral the satellite into higher and higher orbits. We now want to calculate the transfer orbit and the delta-v for a continuous tangential thrust maneuver for an initially circular orbit. How does an infinitesimal small tangential thrust $dv := \delta v_{||}$ alter a circular orbit? The answer is given by Eq. (8.1.14):

$$\frac{dr}{r} = 2\sqrt{\frac{r}{\mu}}\,dv$$

To find the total delta-v, we have to integrate the differentially small velocity increase between an initial $r_\bullet$ and final orbit radius r_O

$$\Delta v = \int dv = \frac{\sqrt{\mu}}{2}\int_{r_\bullet}^{r_O} \frac{dr}{r^{3/2}} = -\left.\sqrt{\frac{\mu}{r}}\right|_{r_\bullet}^{r_O} = v_\bullet - v_O \tag{8.4.9}$$

Of course it does not make a difference for the propulsion demand Δv whether we spiral up from $r_\bullet$ to r_O or down from r_O to $r_\bullet$. If we compare this result with Eq. (8.3.8), we see from Fig. 8.13 that the Hohmann transfer is always more favorable than a continuous thrust transfer. But with ion propulsions and their very tiny thrusts there is no alternative to that.

To calculate the transfer time between an initial circular orbit with orbit radius $r_\bullet$ and the instantaneous circular orbit with radius r, we need the explicit trajectory $r = r(t)$. As the circular condition $v = \sqrt{\mu/r}$ is valid for each point of the orbital curve, it is sufficient to find $v = v(t)$. We find it with the help of the thrust equation $F_* = m \cdot \dot{v} = \dot{m}v_*$. To be able to apply it, we have to consider the mass reduction $\dot{m}$ due to the propellant consumption. We assume that the vehicle with the mass m is accelerated by a constant thrust F with constant mass flow rate $\dot{m}_p = -\dot{m} = const$, so: $m = m_0 + \dot{m}t$. With this relation, we integrate the corresponding equation $dv = \dot{m}v_*/m \cdot dt$

$$\int_{v_\bullet}^{v} dv' = \int_0^t \frac{\dot{m}v_*}{m}\, dt'$$

to get

$$v - v_\bullet = v_* \int_0^t \frac{dt'}{m_0/\dot{m} + t'} = v_* \ln\left(\frac{m_0/\dot{m} + t}{m_0/\dot{m}}\right) = v_* \ln\left(1 - \frac{\dot{m}_p}{m_0} t\right) \tag{8.4.10}$$

Because of $v = \sqrt{\mu/r}$, this results in the following spiral trajectory (see Fig. 8.22)

$$r(t) = \mu\left[v_\bullet + v_* \ln\left(1 - \frac{\dot{m}_p}{m_0} t\right)\right]^{-2} \quad @\, \dot{m}_p = const \tag{8.4.11}$$

By solving for t we get the following from Eq. (8.4.10) for the transfer time

$$t_{CT} = \frac{m_0}{\dot{m}_p}\left[1 - \exp\left(\frac{v - v_\bullet}{v_*}\right)\right] = \frac{m_0 v_*}{F_*}\left\{1 - \exp\left[\frac{\sqrt{\mu}}{v_*}\left(\frac{1}{\sqrt{r}} - \frac{1}{\sqrt{r_\bullet}}\right)\right]\right\} \tag{8.4.12}$$

Example

A satellite is released from the Space Shuttle payload bay at an altitude of 300 km, and it is supposed to spiral with an ion thruster ($F_*/m_0 = 100\,\mu\mathrm{g} = 10^{-3}\mathrm{m\ s^{-2}}$ and $v_* = 10\,000\ \mathrm{m\,s^{-1}}$) to GEO. Because $v_\bullet = 7.72\,\mathrm{km\ s^{-1}}$ and $v_{GEO} = v_\mathrm{O} = 3.07\,\mathrm{km\ s^{-1}}$, the transfer time is calculated to be $t_{CT} = 43$ days.

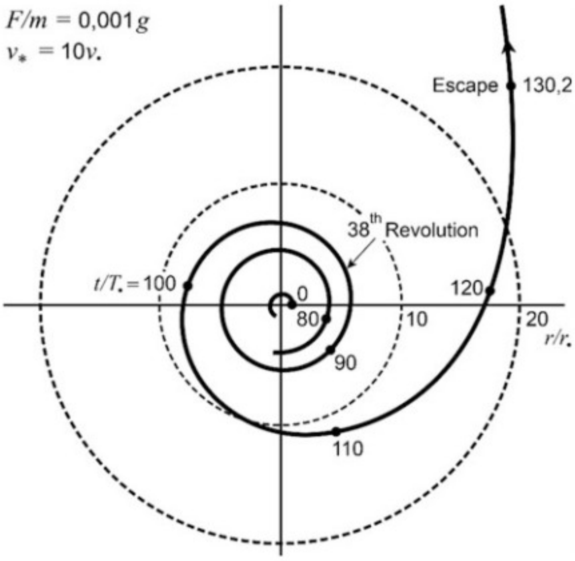

Fig. 8.22 Continuous thrust trajectory as a result of a continuous tangential thrust. Index • refers to the initial orbit. Because the spiral is very narrow at the beginning, only the last two revolutions are shown

8.5 Relative Orbits

Having explored the conditions for a general transfer between orbits, the more delicate problem is to meet a target point in the final orbit, a so-called orbital rendezvous. In this and the next section we will focus on orbital rendezvous in Earth orbits though the described methods apply to any planetary orbit. For interplanetary orbital rendezvous we refer the reader to Sect. 9.3.

In orbital rendezvous, there is generally a passive target object to rendezvous with an interceptor (a.k.a. *chaser*) as the active part that performs the rendezvous maneuvers. Rather than describe the required rendezvous maneuvers in an inertial reference frame such as a geocentric reference frame, it is very convenient to describe them relative to the target. It is convenient because if the interceptor moves in the vicinity of the target we can linearize the equations of motion, which will enable us to solve them analytically. On the other hand, the target in a conical motion about the center of gravity constitutes a non-inertial reference system, a fact that will complicate our equation of motion somewhat. Nevertheless, the description of relative motion about a target object, which is done in the following, is a prerequisite for analyzing rendezvous maneuvers, which is the objective of Sect. 8.6.

8.5.1 General Equation of Motion

Suppose there is a target object in a planetary orbit and a S/C (usually dubbed chaser) wants to navigate in its vicinity and with respect to it. Good examples are the approach maneuvers of a Soyuz vehicle to the ISS or the re-docking of the ascending Lunar Module to the Command/Service Module in low Lunar orbit during the Apollo moon missions. For this mission we need to know the relative motion to the target and how to maneuver to reach the target. This is the subject of the next two sections.

Let us assume the target (for instance, we assume to be the ISS) is in an arbitrary terrestrial conic orbit with position vector $\boldsymbol{R}$ in the geocentric reference frame. For this orbit Newton's gravitational EoM (7.1.19) must hold

$$\ddot{\boldsymbol{R}} = -\frac{\mu}{R^3}\boldsymbol{R} \tag{8.5.1}$$

Of course, Newton's gravitational EoM must equally hold for the chaser's orbit with position vector $\boldsymbol{\rho}$ in the geocentric reference frame, i.e.

$$\ddot{\boldsymbol{\rho}} = -\frac{\mu\boldsymbol{\rho}}{\rho^3}$$

We start deriving the chaser's equation of motion relative to the ISS by defining the relative vector $\boldsymbol{r} = \boldsymbol{\rho} - \boldsymbol{R}$ as shown in Fig. 8.23. The specific reference frame is the ISS-centric coordinate system RSW as described in Sect. 13.1.4, Fig. 13.4, with unit vectors $\boldsymbol{u}_x, \boldsymbol{u}_y, \boldsymbol{u}_z = \boldsymbol{S}, \boldsymbol{W}, \boldsymbol{R}$: The z-axis points along the radial vector while the x-axis points along the moving direction of the target. The y-axis completes the right-handed reference system. Our goal is to find the equation of motion of the relative vector

$$\boldsymbol{r} = \boldsymbol{\rho} - \boldsymbol{R} = x\boldsymbol{u}_x + y\boldsymbol{u}_y + z\boldsymbol{u}_z =: (x, y, z)$$

expressed in the co-moving coordinate system $(\boldsymbol{u}_x, \boldsymbol{u}_y, \boldsymbol{u}_z)$. Since we seek for a differential equation for $\boldsymbol{r}$, we first have to express $\boldsymbol{\rho}$ in terms of $\boldsymbol{R}$ and $\boldsymbol{r}$. Because $r \ll R$ we can expand the term $1/\rho^3$ as follows:

$$\begin{aligned}\frac{1}{\rho^3} &= \frac{1}{\left[x^2 + y^2 + (z+R)^2\right]^{3/2}} = \frac{1}{\left[r^2 + 2zR + R^2\right]^{3/2}} \approx \frac{1}{R^3}\frac{1}{(1+z/R)^3} \\ &\approx \frac{1}{R^3}\left(1 - 3\frac{z}{R}\right)\end{aligned}$$

and hence

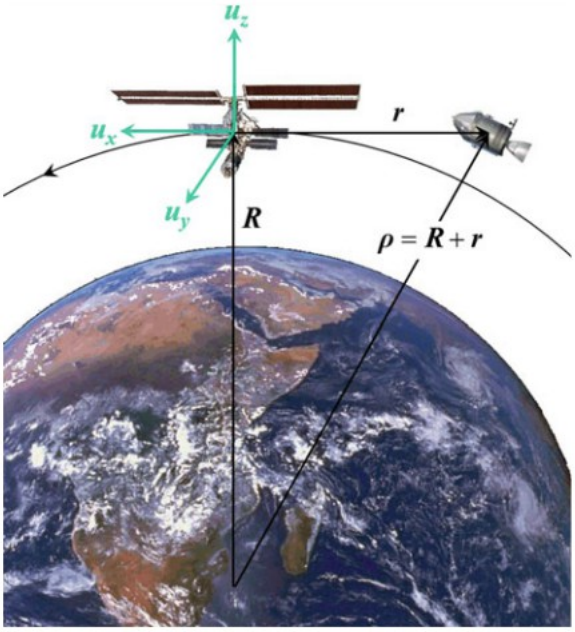

Fig. 8.23 Definition of relative vectors and RSW coordinate system in the target object (ISS), that moves on a circular orbit. The *z*-axis points along the radial vector while the *x*-axis points along the moving direction of the target object. The *y*-axis completes the right-handed reference system

$$\ddot{\boldsymbol{\rho}} = -\frac{\mu\boldsymbol{\rho}}{\rho^3} \approx -\frac{\mu}{R^3}(\boldsymbol{R}+\boldsymbol{r})\left(1-3\frac{z}{R}\right) \approx -\frac{\mu}{R^3}(\boldsymbol{R}+\boldsymbol{r}-3z\boldsymbol{u}_z)$$

Because from the definition of $\boldsymbol{\rho}$ and from Eq. (8.5.1) also $\ddot{\boldsymbol{\rho}} = \ddot{\boldsymbol{R}} + \ddot{\boldsymbol{r}} = -\mu\boldsymbol{R}/R^3 + \ddot{\boldsymbol{r}}$ holds, we get

$$-\frac{\mu}{R^3}\boldsymbol{R} + \ddot{\boldsymbol{r}} \approx -\frac{\mu}{R^3}(\boldsymbol{R}+\boldsymbol{r}-3z\boldsymbol{u}_z)$$

and finally

$$\boxed{\ddot{\boldsymbol{r}} = -\frac{\mu}{R^3}(\boldsymbol{r}-3z\boldsymbol{u}_z)} \quad @\, x,y,z \ll R \tag{8.5.2}$$

We want to solve this vector equation in the non-inertial co-moving target RSW coordinate system $(\boldsymbol{u}_x, \boldsymbol{u}_y, \boldsymbol{u}_z)$. Now, as $\boldsymbol{r} = x\boldsymbol{u}_x(t) + y\boldsymbol{u}_y + z\boldsymbol{u}_z(t)$ is already given in RSW, how is $\ddot{\boldsymbol{r}}$ expressed in it?

The answer is provided by Eq. (7.2.2) for $\boldsymbol{a} \equiv \boldsymbol{r} = x\boldsymbol{u}_x(t) + y\boldsymbol{u}_y + z\boldsymbol{u}_z(t) =: (x,y,z)_{xyz}$ and for the angular velocity of the target $\boldsymbol{\omega} = \omega(0,1,0)_{xyz}$ as

$$
\begin{aligned}
\ddot{\boldsymbol{r}} &= \ddot{\boldsymbol{r}}_{xyz} + 2\boldsymbol{\omega} \times \dot{\boldsymbol{r}}_{xyz} + \boldsymbol{\omega} \times (\boldsymbol{\omega} \times \boldsymbol{r}) + \dot{\boldsymbol{\omega}} \times \boldsymbol{r} \\
&= \left(\ddot{x} + 2\omega\dot{z} - \omega^2 x + \dot{\omega}z\right)\boldsymbol{u}_x + \ddot{y}\boldsymbol{u}_y + \left(\ddot{z} - 2\omega\dot{x} - \omega^2 z - \dot{\omega}x\right)\boldsymbol{u}_z \\
&= \begin{pmatrix} \ddot{x} + 2\omega\dot{z} - \omega^2 x + \dot{\omega}z \\ \ddot{y} \\ \ddot{z} - 2\omega\dot{x} - \omega^2 z - \dot{\omega}x \end{pmatrix}_{xyz}
\end{aligned}
$$

where

$$\dot{\boldsymbol{r}}_{xyz} := (\dot{x}, \dot{y}, \dot{z})_{xyz},\ \ddot{\boldsymbol{r}}_{xyz} := (\ddot{x}, \ddot{y}, \ddot{z})_{xyz}$$

If we insert this into Eq. (8.5.2), we finally obtain the equation of motion

$$\ddot{\boldsymbol{r}} = \begin{pmatrix} \ddot{x} + 2\omega\dot{z} - \omega^2 x + \dot{\omega}z \\ \ddot{y} \\ \ddot{z} - 2\omega\dot{x} - \omega^2 z - \dot{\omega}x \end{pmatrix}_{xyz} = -\frac{\mu}{R^3}\begin{pmatrix} x \\ y \\ z - 3z \end{pmatrix}_{xyz}$$

Because this vectorial equation must hold for each orthogonal component separately, we finally find the three differential equations

$$
\begin{aligned}
&\ddot{x} + 2\omega\dot{z} - \left(\omega^2 - \tfrac{\mu}{R^3}\right)x + \dot{\omega}z = 0 \\
&\ddot{y} \qquad\qquad + \tfrac{\mu}{R^3}y \qquad\quad = 0 \\
&\ddot{z} - 2\omega\dot{x} - \left(\omega^2 + 2\tfrac{\mu}{R^3}\right)z - \dot{\omega}x = 0
\end{aligned}
\quad @x, y, z \ll R \qquad \textbf{equations of motion} \qquad (8.5.3)
$$

This are the equations of motion of a chaser, not being subject to external forces, relative to a target with coordinates (x, y, z) in the target's topocentric RSW coordinate system, which quite generally is in a conic orbit about a center of gravity and, due to its conical trajectory, rotates with angular velocity $\omega(t)$. Note that from Eqs. (7.2.7), (7.3.7) and (7.4.13) follows

$$\frac{\mu}{R^3} = \sqrt{n_e}\omega^{3/2}$$

with

$$n_e := \frac{n}{\left(1 - e^2\right)^{3/2}} = \sqrt{\frac{\mu}{a^3\left(1 - e^2\right)^3}}$$

Observe that only the first and third equations are coupled. If initially $y(0) = 0$ and $\dot{y}(0) = 0$, i.e., if the chaser initially is in the orbital plane, then $\ddot{y} = 0$ and hence $y(t) = 0$, which means that it will always stay in that plane. Otherwise and if $R \approx const$ it will oscillate about this plane with frequency $\sqrt{\mu/R^3}$.

Elliptic Target Orbit

To get more practical and to find a solution to Eq. (8.5.3) we now assume an elliptic orbit with orbital elements a, e about a planet. For this, we get according to Eq. (7.2.7)

$$\omega = \frac{\sqrt{\mu a(1-e^2)}}{R^2}$$

and with Eqs. (7.3.15a) and (7.4.14c)

$$\dot{\omega} = -2\frac{\mu e \sin\theta}{R^3} = -\frac{2\mu a e}{R^4}\sqrt{1-e^2}\sin E$$

with

$$R(t) = a[1 - e\cos E(t)]$$

where the eccentric anomaly $E(t)$ is the solution of Kepler's Eq. (7.4.15)

$$n \cdot t = t\sqrt{\frac{\mu}{a^3}} = E(t) - e\sin E(t)$$

This transcendental equation can be solved by Newton's iteration method (Eq. (7.4.16))

$$E_{i+1} = E_i - \frac{E_i - e\sin E_i - n \cdot t}{1 - e\cos E_i}$$

Of course, similar equations can easily be derived for a hyperbolic trajectory (cf. Sect. 7.4.3).

Given these time dependent quantities, the differential Eq. (8.5.3) for a rendezvous on an elliptic target orbit can be solved numerically for instance by a Runge–Kutta method.

8.5.2 *Circular Orbits*

We now come to the most common situation where the target is in a circular orbit. Circular orbits are so common because for a given minimum altitude they have the lowest orbital energy, entailing the least delta-v to achieve. Therefore, nearly every target in a low Earth orbit will attain a circular or quasi-circular orbit. For instance, the ISS has an eccentricity of typically $e = 0.0001 - 0.001$, that can be shown to be induced by the anisotropies of Earth's gravitational potential (see Sect. 12.3.2).

For a circular orbit we have $\ddot{\boldsymbol{R}} = -\mu \boldsymbol{R}/R^3 = -n^2\boldsymbol{R}$ and $n = \sqrt{\mu/R^3} = 2\pi/T = const$. If be insert this into the general equation of motion Eq. (8.5.3) we obtain

$$\boxed{\begin{array}{l}\ddot{x}+2n\dot{z}=0\\ \ddot{y}+n^2y=0\\ \ddot{z}-2n\dot{x}-3n^2z=0\end{array}} \quad @\, x,y,z \ll R \quad \textbf{Hill's equations} \tag{8.5.4}$$

These are the famous *Hill's equations* (a.k.a. *Clohessy–Wiltshire equations*) governing the motion of a S/C, not being subject to external forces, with coordinates (x, y, z) in the topocentric RSW coordinate system of a target object circling a planet with constant orbital velocity $\omega = n$.

μg-Forces on Space Stations

A modest application of Hill's equation in a circular orbit is the residual forces acting on parts of a satellite or a space station. Such residual forces are important for experiments in space, which contribute some unwanted artificial gravity to the otherwise weightlessness in space. To determine their level, let us assume we have an experimental drawer in a science rack in one of the ISS laboratories located at the distance $\boldsymbol{r} = (x, y, z)$ from the center of mass (CM) of the ISS, which orbits at orbital frequency $n = \sqrt{\mu/R^3}$. Since the drawer is fixed relative to CM, we have $\dot{\boldsymbol{r}} = (\dot{x}, \dot{y}, \dot{z}) = 0$. If the sample in the drawer is further out in the z-direction (see Fig. 8.24) than CM, it experiences an enhanced centrifugal force, but smaller gravitational force. If it is further toward Earth it experiences the opposite effect: a smaller centrifugal force but a larger gravitational force. In total, it thus experiences some residual force, a so-called *tidal force*. We can determine this tidal force if we interpret the terms $\ddot{x}, \ddot{y}, \ddot{z}$ in Eq. (8.5.4) as the residual accelerations making up the tidal force. In doing so we end up with the tidal acceleration in terms of *Earth's mean gravitational acceleration at its surface* (see Sect. 7.1.2), $g_0 := \mu/R_\oplus^2 \approx \mu/R^2 = 9.80\,\mathrm{m\,s^{-2}}$, as

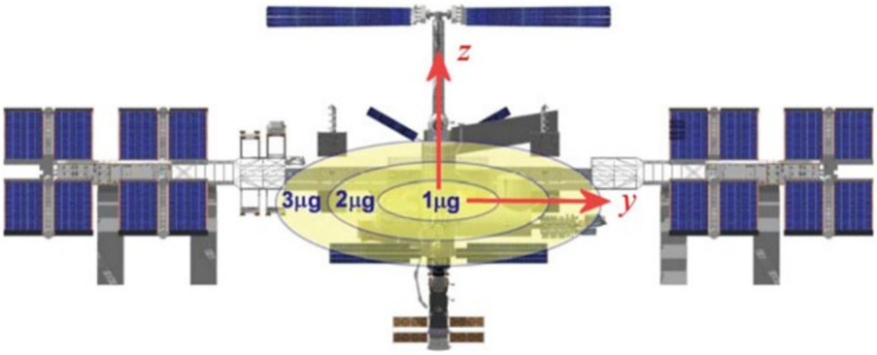

Fig. 8.24 The levels of residual μg-forces on the International Space Station at positions off the center of mass

$$\frac{\ddot{x}}{g_0} = 0, \quad \frac{\ddot{y}}{g_0} = -\frac{y}{R}, \quad \frac{\ddot{z}}{g_0} = 3\frac{z}{R}$$

So, depending on the distances y, z from CM we get different levels of tidal forces. By analyzing Eq. (8.5.3) we realize that in the z-direction there are two gravitational tide contributions and just one centrifugal tide contribution in the overall total of 3. In the y-direction there is just one contribution from the gravitational (lateral) tide. In general the tidal forces are of order $10^{-6}g = 1\mu g$. This is why experiments on space stations are commonly called μg-experiments. For instance, on the ISS, having a typical altitude of 350 km, a level of $1\mu g$ is obtained at a horizontal distance of $y = 10^{-6} \times 6.73 \times 10^6\,\text{m} = 6.73\,\text{m}$ and at a vertical distance of $z = \frac{1}{3} \times 10^{-6} \times 6.73 \times 10^6\,\text{m} = 2.24\,\text{m}$ from CM. Overall, we have μg-ellipses as depicted in Fig. 8.24.

Solution of Hill's Equations

To describe the explicit motion $\boldsymbol{r}(t) = (x(t), y(t), z(t))$ of a chaser, we have to solve Hill's equations. This can be done straightforwardly with only little effort. To do so efficiently, we first replace time by the dimensionless mean anomaly and define both it and its derivatives.

$$\begin{aligned} &M = nt \\ &(.)' := \frac{d}{dM} = \frac{1}{n}\frac{d}{dt}(.) \\ &(.)'' := \frac{d^2}{dM^2} = \frac{1}{n^2}\frac{d^2}{dt^2}(.) \end{aligned} \tag{8.5.5}$$

and initial conditions at $t = 0$ as

$$\begin{aligned} x(0) &= x_0, \quad x'(0) = x'_0 \\ y(0) &= y_0, \quad y'(0) = y'_0 \\ z(0) &= z_0, \quad z'(0) = z'_0 \end{aligned} \quad \textbf{initial conditions} \tag{8.5.6}$$

With this Hill's equations read

$$\begin{aligned} x'' + 2z' &= 0 \\ y'' + y &= 0 \\ z'' - 2x' - 3z &= 0 \end{aligned} \tag{8.5.7}$$

First observe that the second Hill equation is decoupled from all others. It has the form of a harmonic oscillator. We therefore find its solution as

$$y = y_0 \cos M + y'_0 \sin M$$

We now integrate the first Hill equation directly. By taking the initial conditions into account we get

$$x' = -2z + 2z_0 + x'_0$$

We insert this result into the third Hill equation getting

$$z'' = 2x' + 3z = -\left[z - \left(4z_0 + 2x'_0\right)\right]$$

This, too, is the equation of a harmonic oscillator for the expression $z - \left(4z_0 + 2x'_0\right)$ yielding

$$z - \left(4z_0 + 2x'_0\right) = z'_0 \sin M - \left(3z_0 + 2x'_0\right) \cos M$$

under consideration of the initial conditions $z(0) = z_0, z'(0) = z'_0$. We solve for z and find

$$z = 4z_0 + 2x'_0 + z'_0 \sin M - \left(3z_0 + 2x'_0\right) \cos M$$

Inserting this into the above equation $x' = -2z + 2nz_0 + x'_0$ we get

$$x' = -\left(6z_0 + 3x'_0\right) - 2z'_0 \sin M + \left(6z_0 + 4x'_0\right) \cos M$$

This can be integrated directly to deliver with initial condition $x(0) = x_0$

$$x = x_0 - 2z'_0 - \left(6z_0 + 3x'_0\right)M + 2z'_0 \cos M + \left(6z_0 + 4x'_0\right) \sin M$$

We summarize the results as

$$\begin{pmatrix} x \\ z \\ x' \\ z' \end{pmatrix} = \begin{bmatrix} 1 & 6\sin M - 6M & 4\sin M - 3M & 2\cos M - 2 \\ 0 & -3\cos M + 4 & -2\cos M + 2 & \sin M \\ 0 & 6\cos M - 6 & 4\cos M - 3 & -2\sin M \\ 0 & 3\sin M & 2\sin M & \cos M \end{bmatrix} \begin{pmatrix} x_0 \\ z_0 \\ x'_0 \\ z'_0 \end{pmatrix} \tag{8.5.8}$$

and in convenient terms

$$\boxed{\underbrace{\begin{pmatrix} x \\ z \\ \dot{x} \\ \dot{z} \end{pmatrix}}_{\substack{\textit{final} \\ \textit{state vector}}} = \underbrace{\begin{bmatrix} 1 & 6(\sin nt - nt) & (4\sin nt)/n - 3t & 2(\cos nt - 1)/n \\ 0 & 4 - 3\cos nt & 2(1 - \cos nt)/n & (\sin nt)/n \\ 0 & 6n(\cos nt - 1) & 4\cos nt - 3 & -2\sin nt \\ 0 & 3n \sin nt & 2\sin nt & \cos nt \end{bmatrix}}_{\textit{time-evolution matrix}} \underbrace{\begin{pmatrix} x_0 \\ z_0 \\ \dot{x}_0 \\ \dot{z}_0 \end{pmatrix}}_{\substack{\textit{initial} \\ \textit{state vector}}}} \tag{8.5.9}$$

and

$$\boxed{\begin{pmatrix} y \\ \dot{y} \end{pmatrix} = \begin{bmatrix} \cos nt & (\sin nt)/n \\ -n \sin nt & \cos nt \end{bmatrix} \begin{pmatrix} y_0 \\ \dot{y}_0 \end{pmatrix}} \tag{8.5.10}$$

These are the basic equations for orbital motion in the vicinity of the reference object and with respect to its co-moving topocentric coordinate system RSW.

Application to Orbital Rendezvous

A typical application of Eq. (8.5.9) is the design of an orbital rendezvous. If at time $t_0 = 0$ a S/C is at an initial point $\boldsymbol{r}_0 = (x_0, y_0, z_0)$ the problem to solve is: What should be the initial velocity $\boldsymbol{v}_0 = (\dot{x}_0, \dot{y}_0, \dot{z}_0)$ to meet after time t a given target point $\boldsymbol{r} = (x, y, z) = (0, 0, 0)$? The solution can be derived from the equations for x, y, and z of Eqs. (8.5.9) and (8.5.10) by setting $x = y = z = 0$ and solving for $\dot{x}_0, \dot{y}_0, \dot{z}_0$. If this is done, one straightforwardly obtains

$$\begin{aligned} \frac{\dot{x}_0}{n} &= \frac{2(\cos nt - 1)x_0 + (4 \sin nt - 3nt \cos nt)z_0}{3nt \sin nt + 8(\cos nt - 1)} \\ \frac{\dot{z}_0}{n} &= \frac{\sin nt \cdot x_0 - [6nt \sin nt + 14(\cos nt - 1)]z_0}{3nt \sin nt + 8(\cos nt - 1)} \\ \frac{\dot{y}_0}{n} &= -\frac{y_0}{\tan nt} \end{aligned} \tag{8.5.11}$$

or in matrix expression

$$\begin{pmatrix} \dot{x}_0 \\ \dot{z}_0 \end{pmatrix} = \frac{n}{3nt \sin nt + 8 \cos nt - 8} \begin{pmatrix} 2 \cos nt - 2 & 4 \sin nt - 3nt \cos nt \\ \sin nt & 14 - 14 \cos nt - 6nt \sin nt \end{pmatrix} \begin{pmatrix} x_0 \\ z_0 \end{pmatrix}$$

$$\dot{y}_0 = -\frac{n}{\tan nt} y_0$$

The required delta-v for these maneuvers then is given by

$$\Delta v = \sqrt{(\dot{x}_0 - \dot{x}_{0-})^2 + (\dot{y}_0 - \dot{y}_{0-})^2 + (\dot{z}_0 - \dot{z}_{0-})^2} \tag{8.5.12}$$

where $\boldsymbol{v}_{0-} = (\dot{x}_{0-}, \dot{y}_{0-}, \dot{z}_{0-})$ is the velocity of the S/C just before the impulse maneuver at (x_0, y_0, z_0).

It is an easy exercise to show that in the limit $t \to 0$ Eq. (8.5.11) pass over into

$$\begin{aligned} \dot{x}_0 &= -\frac{x_0}{t} - n z_0 - \frac{1}{6} n^2 x_0 t + O(n^3 z_0 t^2) \\ \dot{z}_0 &= -\frac{z_0}{t} + n x_0 - \frac{5}{6} n^2 z_0 t + O(n^3 x_0 t^2) \qquad @\, t \to 0 \\ \dot{y}_0 &= -\frac{y_0}{t} + \frac{1}{3} n^2 y_0 t + O(n^4 y_0 t^3) \end{aligned}$$

Therefore, for $(nt)^2 \ll 1$, which implies roughly $t < 0.02 \cdot T$, we can write

$$\boxed{\begin{aligned} \dot{x}_0 &= -\tfrac{x_0}{t} - n\,z_0 \\ \dot{z}_0 &= -\tfrac{z_0}{t} + n\,x_0 \\ \dot{y}_0 &= -\tfrac{y_0}{t} \end{aligned}} \quad @\ (nt)^2 \ll 1 \tag{8.5.13}$$

These expressions will later be useful to explore maneuvers in the vicinity of the target object.

8.5.3 Flyaround Trajectories

In order to fly in a controlled way in the vicinity of a target object we need to know: What does a flyaround trajectory of a S/C look like? We will now explore the two different types of trajectories because they are essential for understanding the general behavior of such a S/C.

The Prolate Cycloid

On November 18, 2008, the astronaut Heidemarie Stefanyshyn-Piper lost her toolbox in space when carrying out repair work outside the International Space Station. She accidentally and gently touched the toolbox (which should have been secured by a line either to her or the ISS, but was not) giving the box a push such that it slowly drifted away from the station (see Fig. 8.25).

This accident is a good example to study the trajectory of slowly moving objects near a reference system, which for our purposes is the ISS at an average altitude of 350 km. So, what was the trajectory of the toolbox? It is important to know in which direction the push was relative to the flight direction of the ISS. Because we do not know, we will assume two different situations. First, Heidemarie pushed the toolbox along the flight direction, with an initial velocity of, say, $\boldsymbol{v}_0 = (\dot{x}_0, \dot{y}_0, \dot{z}_0) = (v_0, 0, 0)$.

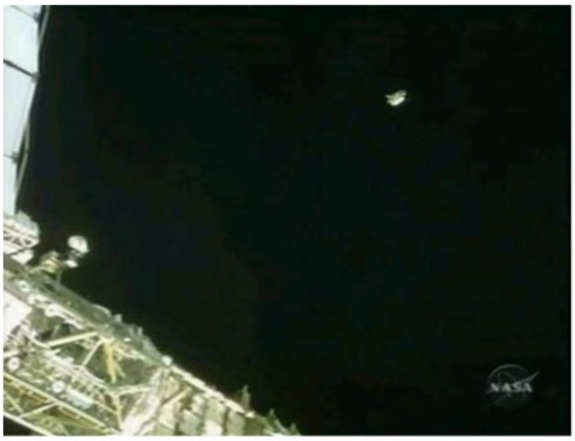

Fig. 8.25 The toolbox (top right) of astronaut Stefanyshyn-Piper slowly drifting away from the ISS truss structure (bottom left). *Credit* NASA

We study the trajectory in the RSW coordinate system of Heidemarie, so $\boldsymbol{r}_0 = (x_0, y_0, z_0) = (0, 0, 0)$. Under these initial assumptions, we read from Eq. (8.5.9)

$$\begin{aligned} x(t) &= \frac{v_0}{n}(4\sin nt - 3nt) \approx v_0 t \\ \dot{x}(t) &= v_0(4\cos nt - 3) \\ z(t) &= 2\frac{v_0}{n}(1 - \cos nt) \approx n v_0 t^2 \\ \dot{z}(t) &= 2v_0 \sin nt \end{aligned} \tag{8.5.14}$$

where the approximations are for early times on the trajectory, $nt \ll 1$. If we solve the approximation for x and insert it in the approximation for z, we obtain a parabolic trajectory

$$z = \frac{nx^2}{v_0}$$

To be more concrete, let us assume that Heidemarie pushed the toolbox backward, against flight direction, with an initial velocity of, say, $v_0 = -0.1\,\mathrm{m\,s^{-1}}$. Because v_0 is negative we get a down curved parabola (see Fig. 8.26): The toolbox dives below the orbit of the ISS, which first seems to be quite surprising. Interestingly the toolbox comes to a halt in –x-direction when $\cos nt = 3/4$ and thereafter moves back toward the ISS, yet at increasingly lower altitudes. After half an orbit, it reaches the lowest point (perigee), and then it climbs up to reach the initial altitude of 350 km after a full revolution of the ISS, however at a distance of 1.65 km in front of the ISS. The toolbox will repeat this loopy motion indefinitely thereby moving away from the ISS—the toolbox will be lost. Such a winding trajectory is mathematically called a *prolate cycloid*.

What is the reason for such a weird trajectory? When the toolbox is kicked off backward it has a lower velocity as it should have for a circular orbit at this altitude.

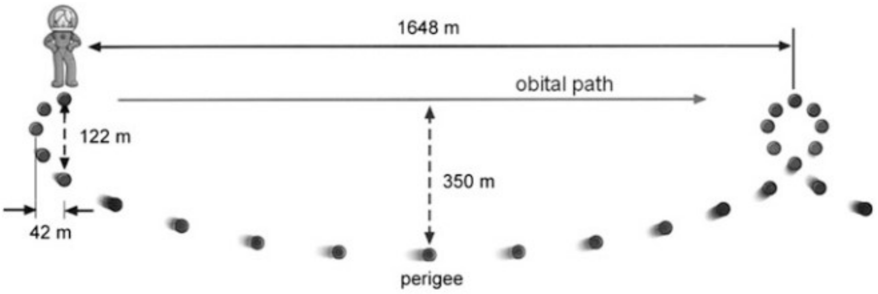

Fig. 8.26 The prolate cycloid trajectory of the toolbox that Heidemarie lost, if its initial motion, $v_0 = -0.1\,\mathrm{m\,s^{-1}}$, was reverse to the flight direction of the ISS. ISS altitude is $h = 350\,\mathrm{km}$

So the centrifugal force $F_c = mv^2/r$ is lower than the gravitational force, which rapidly pulls the box down. But upon dropping it gains speed, which brings it to a lateral halt at 42 m away from the ISS. With increasing speed the centrifugal force increases that after half a revolution at perigee brings the decline to a halt at 350 m below and 824 m ahead of the ISS. Since the speed now is too large for this altitude, the box dives up and decelerates until it "kisses" the orbit to iterate its dive.

The size of the cycloid depends linearly on the initial velocity while its orientation reverses when v_0 is with or against the flight direction of the ISS. Figure 8.27 gives an overview of this behavior for different initial x_0.

The reason for this weird motion can be understood when seen in a geocentric reference system (see Fig. 8.28). An object initially slower than the ISS at point 0 will have less orbital energy and hence move on an elliptic trajectory with smaller major-axis than the ISS. This implies that the object will drop below the ISS and will have a smaller orbital period, so showing up in front of the ISS after one revolution at the initial point 0. An object with a larger initial velocity will move on a larger ellipse with reversed behavior.

This view in an inertial reference frame makes clear that the prolate cycloid is the trajectory of a Hohmann transfer as observed in the target's reference frame. This should have been clear from the very beginning, because the initial kick-burn is tangential to the initial orbit. This is the attribute of a Hohmann transfer.

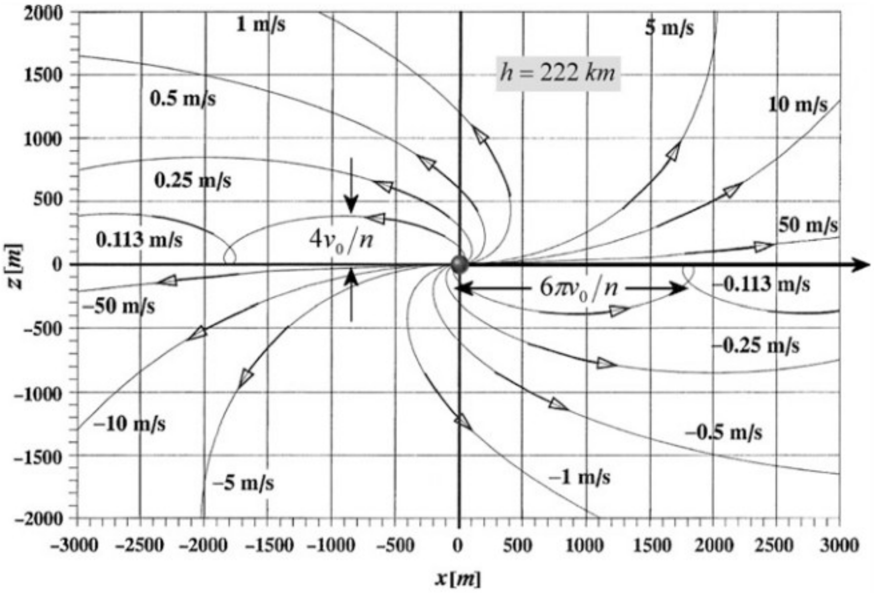

Fig. 8.27 Shown are for different v_0 in x-direction the trajectories (prolate cycloid) of the object moving relative to a reference point [center dot, which itself moves on an orbit at altitude $h =$ 222 km to the right (bold arrow)]. *Credit* Vallado (2007)

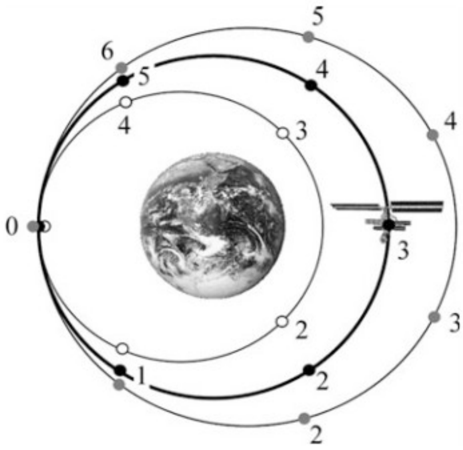

Fig. 8.28 A schematic sketch of the prolate cycloid motion as viewed from the inertial reference frame of Earth. The smaller ellipse is for an object with smaller velocity and the larger ellipse with a larger velocity at point 0. The numbered points give the positions on each orbit after constant time intervals

The Ellipse

We now assume that Heidemarie pushed the toolbox vertically to the flight direction, with initial velocity $\boldsymbol{v}_0 = (\dot{x}_0, \dot{y}_0, \dot{z}_0) = (0, 0, v_0)$. Interestingly, the trajectory of the toolbox would be totally different. From Eq. (8.5.9) we derive the trajectory to be

$$\begin{aligned} x &= 2\frac{v_0}{n}(\cos nt - 1) \approx -nv_0t^2 \\ z &= \frac{v_0}{n}\sin nt \approx v_0 t \end{aligned} \tag{8.5.15}$$

again with approximations for early times on the trajectory, $nt \ll 1$. By removing the time parameter for $nt \ll 1$, we again find a parabola, that, however, opens to the rear side of the ISS.

$$z = \sqrt{-\frac{v_0}{n}x}$$

To derive the full trajectory we recognize that we can solve both equations for $\cos\, nt$ and $\sin\, nt$ and via the trigonometric equation $\sin^2 x + \cos^2 x = 1$, we find

$$\frac{(x + 2v_0/n)^2}{(2v_0/n)^2} + \frac{z^2}{(v_0/n)^2} = 1$$

which is the equation of an ellipse with semi-major axis $a = 2v_0/n$ along the x-axis and semi-minor axis $b = v_0/n$ along the z-axis, and therefore eccentricity $e = \sqrt{1 - b^2/a^2} = \sqrt{3/4} = 0.8660\ldots$. Details of such ellipses with $v_0 = 0.1\,\mathrm{m\,s^{-1}}$ are shown in Fig. 8.29 and for varying initial velocities in Fig. 8.30.

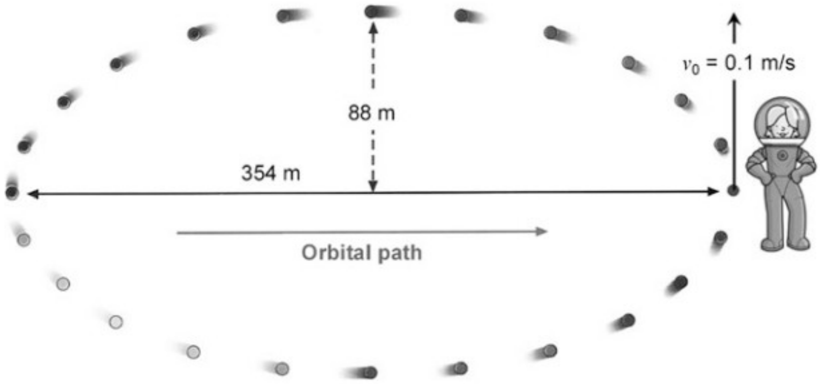

Fig. 8.29 The path of Heidemarie's toolbox if pushed with $v_0 = 0.1\ \mathrm{m\ s^{-1}}$ along the positive z-direction. The box follows a closed elliptic path and returns to Heidemarie after one orbital revolution of the ISS. ISS altitude is $h = 350$ km

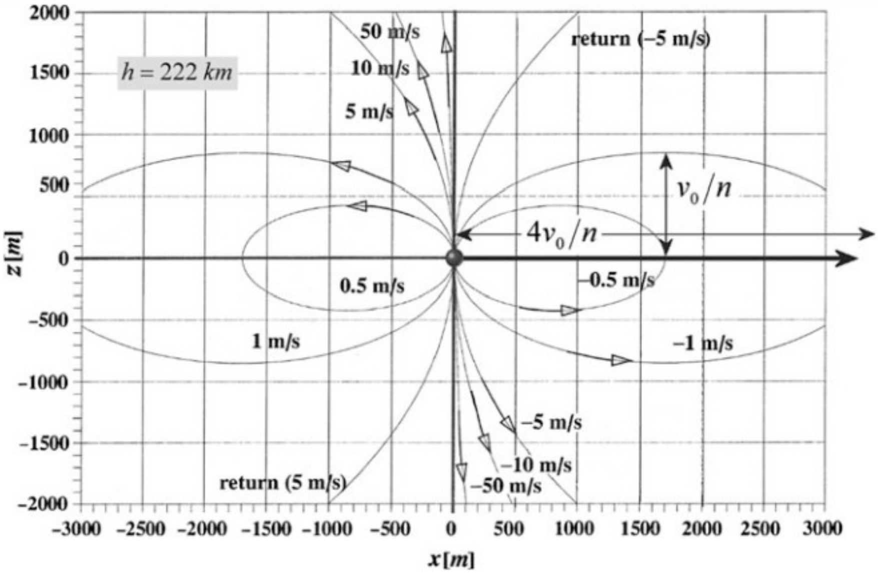

Fig. 8.30 Trajectories (ellipses) of an object moving relative to a reference point [center dot, which itself moves on an orbit at altitude $h = 222$ km to the right (bold arrow)] for different v_0 in z-direction. *Credit* Vallado (2007)

We can summarize the behavior of the toolbox as follows: When initially moving up, the orbital velocity remains constant. However, the angular velocity $\omega = v/r$ and the centrifugal force $F_c = mv^2/r$ decrease. Owing to this, the toolbox falls behind the ISS and also loses vertical speed, because the gravitation pull becomes bigger than the centrifugal force. After a quarter of an orbital revolution of the ISS, the toolbox drops down and crosses the orbit after half an orbital revolution, but now with downward velocity. The inverse behavior thereafter would bring the toolbox back to Heidemarie.

If a S/C relative to a target object starts its trajectory from the orbital path of the target object and initially moves along the x- or z-direction, the ellipse and the prolate cycloid are the two possible forms of this trajectory. The trajectories become more intricate if $\boldsymbol{v}_0$ is not pointing into the x- or z-direction, or the initial location is displaced from the orbital path, or both. In any of these cases the main features of such convoluted trajectories are the ellipse and the prolate cycloid. The reason is, whatever the initial conditions of the S/C might be, it will move on an ellipse around the Earth. If its semi-major axis a is the same as that of the target object it will just fly around the target object periodically. If its a is different, its orbital period will be different and therefore the periodic motion will be superimposed by a motion that let the S/C and the target object drift relative to each other. These two features are extensively used for applied orbital rendezvous maneuvers to which we will turn now.

8.5.4 Near-Circular Orbits

Differential Equations

We now want to convey these results to near-circular orbits, i.e. orbits with $e \ll 1$. For small eccentricities it is easy to show (exercise, Problem (8.9)) that Eq. (8.5.3) pass over into

$$\begin{aligned} \ddot{x} + 2n\dot{z} &= \quad en(nx \cos nt + 2nz \sin nt - 4\dot{z} \cos nt) \\ \ddot{y} + n^2 y &= -3en^2 y \cos nt \\ \ddot{z} - 2n\dot{x} - 3n^2 z &= 2en(-nx \sin nt + 5nz \cos nt + 2\dot{x} \cos nt) \end{aligned} \qquad @\, e \ll 1 \quad (8.5.16)$$

or in our dimensionless notation (see Sect. 8.5.2)

$$M = nt, \quad (.)' := \frac{d}{dM} = \frac{1}{n}\frac{d}{dt}, \quad (.)'' = \frac{d^2}{dM^2} = \frac{1}{n^2}\frac{d^2}{dt^2}$$

they read

$$\begin{aligned} x'' + 2z' &= \quad e(x \cos M + 2z \sin M - 4z' \cos M) \\ y'' + y &= 3e(-y \cos M) \\ z'' - 2x' - 3z &= 2e(-x \sin M + 5z \cos M + 2x' \cos M) \end{aligned} \qquad @\, e \ll 1 \quad (8.5.17)$$

Here M is measured relative to the periapsis of the elliptical target orbit.

Solution to Differential Equations

It is quite intricate to solve these differential equations. We therefore leave this as an exercise to the reader (see Problem 8.10) and provide here only the exact results:

$$x(M) = \left[\begin{pmatrix} 1 \\ 6\sin M - 6M \\ 4\sin M - 3M \\ 2\cos M - 2 \end{pmatrix}^T + e \begin{pmatrix} -\cos M + 1 \\ \sin M \cos M + 12 \sin M - 6M \cos M - 15M \\ 6\sin M \cos M - 3M \cos M - 3M \\ 3 \cos^2 M - 2 \cos M - 1 \end{pmatrix}^T \right] \begin{pmatrix} x_0 \\ z_0 \\ x'_0 \\ z'_0 \end{pmatrix}$$

$$y(M) = \left[\begin{pmatrix} \cos M \\ \sin M \end{pmatrix}^T + e \begin{pmatrix} \cos^2 M + \cos M - 2 \\ \sin M \cos M - \sin M \end{pmatrix}^T \right] \begin{pmatrix} y_0 \\ y'_0 \end{pmatrix} \tag{8.5.18}$$

$$z(M) = - \left[\begin{pmatrix} 0 \\ 3\cos M - 4 \\ 2\cos qM - 2 \\ -\sin M \end{pmatrix}^T + e \begin{pmatrix} 0 \\ 6\cos^2 M + 10\cos M + 6M \sin M - 16 \\ 4\cos^2 M + 2\cos M + 3M \sin M - 6 \\ -2\sin M \cos M + 2\sin M \end{pmatrix}^T \right] \begin{pmatrix} x_0 \\ z_0 \\ x'_0 \\ z'_0 \end{pmatrix}$$

We recall that M is measured relative to the periapsis of the elliptical target orbit. Observe that these solutions satisfy our standard initial conditions

$$\begin{aligned} x(0) &= x_0, \quad x'(0) = x'_0 \\ y(0) &= y_0, \quad y'(0) = y'_0 \\ z(0) &= z_0, \quad z'(0) = z'_0 \end{aligned}$$

Flyaround Trajectories

At the periapsis let us first place a S/C at $(x_0, 0, 0)$ behind the ISS with initial zero relative velocity, $\boldsymbol{v}_0 = (0, 0, 0)$, and see what happens. From Eq. (8.5.18) we obtain

$$\begin{aligned} x(t) &= [1 + e(1 - \cos nt)]x_0 \\ y(t) &= 0 \\ z(t) &= 0 \end{aligned}$$

If the reference orbit were circular we would get a steady distance, $x(t) = x_0 = \mathit{const}$ to the target because the orbit velocity is always constant. For an elliptical target the distance slightly changes over one orbit in the interval $x_0 \le x(t) \le (1 + 2e) \cdot x_0$. Where $x = (1 + 2e) \cdot x_0$ happens at the apoapsis (Fig. 8.31).

The Prolate Cycloid

We now start with an initial longitudinal velocity $\boldsymbol{v}_0 = (\dot{x}_0, \dot{y}_0, \dot{z}_0) = (v_0, 0, 0)$ and again study the trajectory in the RSW coordinate system of the ISS, so $\boldsymbol{r}_0 = (x_0, y_0, z_0) = (0, 0, 0)$. Under these initial assumptions, we read from Eq. (8.5.18) after some minor modification

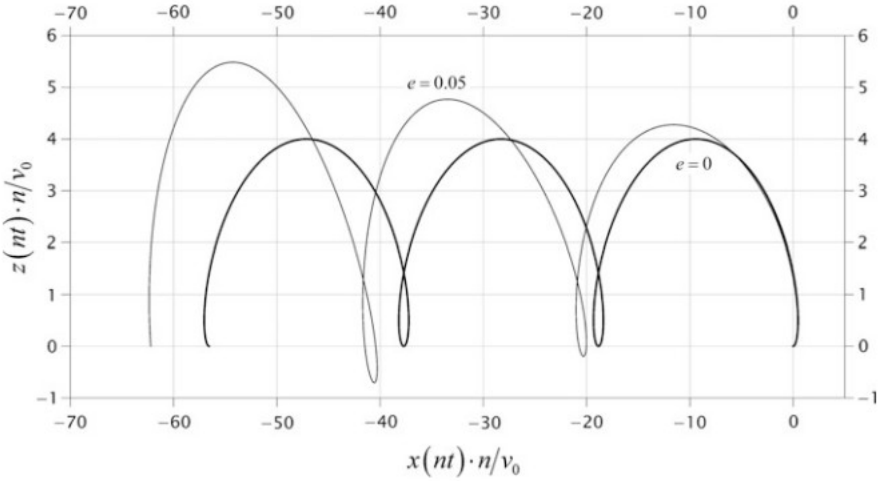

Fig. 8.31 The prolate cycloid as given by Eq. (8.5.19) with $e = 0.05$ and with $e = 0$ for reference

$$\begin{aligned} x(t) &= \frac{v_0}{n}[4\sin nt - 3nt + 3e(2\sin nt\cos nt - nt(1+\cos nt))] \\ \dot{x}(t) &= v_0[4\cos nt - 3 + 3e(2\cos 2nt - \cos nt + nt\sin nt - 1)] \end{aligned} \tag{8.5.19}$$

and

$$\begin{aligned} z(t) &= -\frac{v_0}{n}\left[2\cos nt - 2 + e\left(4\cos^2 nt + 2\cos nt + 3nt\sin nt - 6\right)\right]\dot{z}(t) \\ &= v_0[2\sin nt + e(4\sin 2nt - \sin nt - 3nt\cos nt)] \end{aligned}$$

The Ellipse

As in Sect. 8.5.2 we finally examine the flyaround trajectory with an initial transversal velocity $\boldsymbol{v}_0 = (\dot{x}_0, \dot{y}_0, \dot{z}_0) = (0, 0, v_0)$. In this case we find

$$\begin{aligned} x(t) &= \frac{v_0}{n}\left[2\cos nt - 2 + e\left(3\cos^2 nt - 2\cos nt - 1\right)\right] \\ z(t) &= \frac{v_0}{n}\sin nt[1 + 2e(\cos nt - 1)] \end{aligned} \tag{8.5.20}$$

We eliminate the time dependency by using $\sin^2 nt + \cos^2 nt = 1$ and find the trajectory

$$\left[\frac{x + 2v_0/n[1 - \frac{e}{2}(3\cos^2 nt - 2\cos nt - 1)]}{2v_0/n}\right]^2 + \left[\frac{z - 2ev_0\sin nt(\cos nt - 1)/n}{v_0/n}\right]^2 = 1$$

This is also an ellipse, with an center offset by $2e$ as shown in Fig. 8.32.

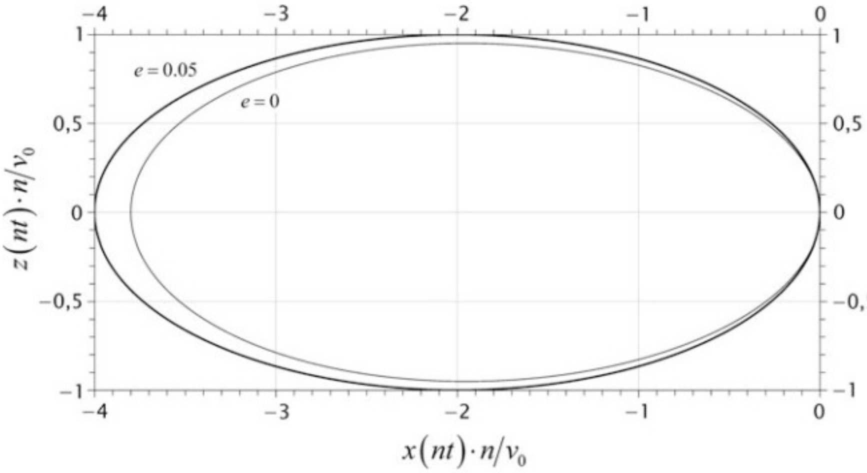

Fig. 8.32 The ellipse as given by Eq. (8.5.20) with $e = 0.05$ and $e = 0$ for reference

8.6 Orbital Rendezvous

Rendezvous and docking (R&D) form some of the most complex maneuvers to be carried out by a spacecraft in spaceflight. To accomplish successful rendezvous and mating of two spacecraft, absolute and relative navigation on orbit, sensing of the target object, precise attitude determination and control, maneuver planning, and the use of highly complex mechanisms must be mastered. The development of these skills was the purpose of the US Gemini program and the early Soviet Soyuz program, and they were further refined in the era of the Space Shuttle and the ISS. In recent years, the desire for autonomous robotic on-orbit servicing systems and plans for more ambitious human and robotic exploration of the solar system create new challenges for technology and mission designers.

The purpose of any R&D mission is to establish physical contact between two or more spacecraft to establish electrical, material, and crew exchange. Therefore R&D is a prerequisite for the construction and maintenance of space stations such as the ISS, as well as any servicing mission to satellites in Earth orbit. Such missions can be manned (e.g., the Hubble servicing missions) or unmanned (e.g., future On-Orbit Servicing). R&D is also a mission-enabling skill in human and robotic exploration of the solar system. Without successful R&D, the Apollo missions would not have been possible.

As the established term *rendezvous and docking* implies, the involved operations are divided into two distinct parts, each with a particular set of goals.

Rendezvous: During the rendezvous part of the mission, the involved spacecraft are guided to meet in the same volume of space, at the same time. In most applications, the target object (also often referred to as *resident space object*) is inert, and the interceptor (a.k.a. chaser) performs all maneuvers to meet the position and time requirements. However, as an exception to this rule, in the so-called *control box rendezvous*, it is the target spacecraft that executes a number of maneuvers to meet the interceptor after it was launched. This reduces interceptor vehicle propellant consumption, naturally at a cost to the target. It therefore can only be performed with targets having orbit maneuvering capabilities, which usually rules out space stations and a large number of satellites. Nonetheless, this type of rendezvous was performed on some Space Shuttle missions (e.g., STS-49 to service Intelsat VI) and was also planned for the contingency rescue mission STS-400 to Atlantis' Hubble Servicing Mission 4.

Docking: The goal of docking in a mission is to establish physical contact between the involved spacecraft. Although commonly the term *docking* is used, there actually exist two distinct cases: *docking*, and *capture and berthing*. In

(1) *Docking*. The interceptor approaches the target with non-zero relative velocity, brings its docking tool into alignment with the target's counterpart, and establishes a firm structural connection by using its own momentum. Docking therefore relies only on maneuvering capabilities of the two spacecraft and on properly functional docking tools. This approach was used during Gemini and Apollo and still is in use with Soyuz/Progress, the Space Shuttle, and ESA's Automated Transfer Vehicle (ATV) missions to ISS.
(2) *Capture and berthing*. The interceptor is maneuvered into close proximity of the target and an initial mechanical connection between both is established by a robotic manipulator. This manipulator can be either situated on the interceptor (as is the case with the Shuttle Robotic Manipulator System (RMS) used for capturing the Hubble Space Telescope), or on the target vehicle. This is the approach taken with the Space Station Remote Manipulator System (SSRMS) on ISS capturing the Japanese H-II Transfer Vehicle (HTV). After capture, the captured spacecraft is then moved by the manipulator to a berthing position, which is a device similar to a docking port.

The choice of the manipulator's location is, on the one hand, dependent on the sizes and masses of the spacecraft. The attitude of a heavier spacecraft is less influenced by the disturbance torques caused by the movement of the manipulator. On the other hand, a space manipulator is a very complex and hence an expensive mechanism. It will therefore be mounted on the spacecraft with the longer lifetime and/or reentry capability, and not on the disposable spacecraft like HTV.

The combination of both rendezvous and docking is not an end to itself, but serves to fulfill the purpose of a mission. It therefore must always be carefully planned and designed within the larger mission context. This influences not only design decisions such as launch windows, approach trajectories, and the selection of sensors, but also the general approach modes, be it operator-in-the-loop or autonomous robotic. To understand the challenges, choices, and trade-offs involved, the following section provides details about R&D mission design.

Mission Phases

Generally, R&D missions involve both an interceptor spacecraft, which begins the mission on the launch pad, and a target spacecraft, which is usually already in orbit by the time of the interceptor's launch. The following sections will discuss the mission phases of typical R&D missions in Low Earth Orbit (LEO), which are in the following order:

1. Launch
2. Phasing
3. Homing
4. Closing
5. Final Approach
6. Docking/Capture

This sequence of mission steps results in a typical R&D mission profile as shown in Fig. 8.33.

This R&D mission profile typically takes two days to accomplish. This is why it is sometimes called a *2-day rendezvous profile*. It is a standard profile used for R&D missions in particular for ISS R&D by Soyuz, Shuttle or ESA ATV.

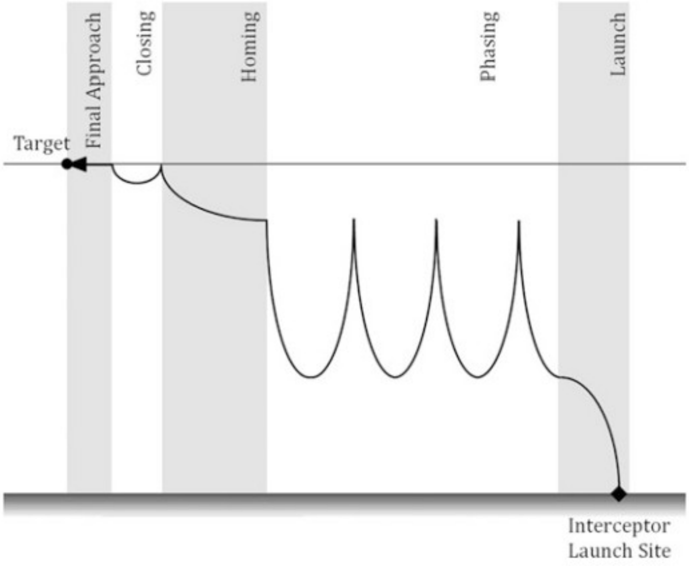

Fig. 8.33 Sketch of a typical R&D mission profile consisting of launch phase, phasing maneuvers, homing, and close-range rendezvous (closing and final approach) including docking

Mission Phases

Generally, R&D missions involve both an interceptor spacecraft, which begins the mission on the launch pad, and a target spacecraft, which is usually already in orbit by the time of the interceptor's launch. The following sections will discuss the mission phases of typical R&D missions in Low Earth Orbit (LEO), which are in the following order:

1. Launch
2. Phasing
3. Homing
4. Closing
5. Final Approach
6. Docking/Capture

This sequence of mission steps results in a typical R&D mission profile as shown in Fig. 8.33.

This R&D mission profile typically takes two days to accomplish. This is why it is sometimes called a *2-day rendezvous profile*. It is a standard profile used for R&D missions in particular for ISS R&D by Soyuz, Shuttle or ESA ATV.

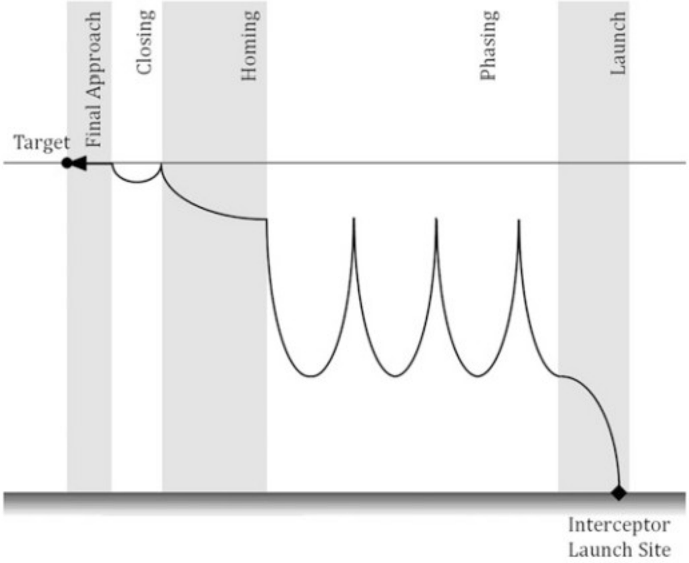

Fig. 8.33 Sketch of a typical R&D mission profile consisting of launch phase, phasing maneuvers, homing, and close-range rendezvous (closing and final approach) including docking

From $\cos i = \cos\beta \cdot \sin\varphi$ we can read that because $|\sin\varphi| \leq 1$, we have $\cos i \leq \cos\beta$ implying $i \geq \beta$. So there exists no launch azimuth to achieve an $i < \beta$, or in other words: Orbits with $i < \beta$ cannot be reached directly. This seemingly paradoxical situation is elucidated in Fig. 8.35. For example, launchers from Kennedy Space Center with $\beta = 28.47°$ cannot reach orbits with $i < 28.47°$. Only in the limiting case $\varphi = 90°,\ 270°$, i.e., if the launch azimuth is East or West (into a retrograde orbit), we obtain $i = \beta$. If orbits $i < \beta$ need to be reached, the launcher is first launched with $i = \beta$. When its orbit intersects the desired target orbit having $i < \beta$, a plane change maneuver (a.k.a. dogleg maneuver) is performed that, however, is associated with large propellant consumption and thus reduced payload for a given target orbit.

What is the right time to launch? The Universal Time of launch, T_{UT}, defines the moment in time when the launch site is in the plane of the International Space Station's orbit, the so-called *in-plane time*. At this launch time the interceptor reaches the targeted orbital plane with the least effort (delta-v). It is given by

$$\boxed{T_{UT} = \frac{\theta_{GMST} - \theta_{GMST0}}{\omega_{\oplus}}} \quad \textbf{in-plane time (launch time)} \qquad (8.6.2)$$

where $\omega_{\oplus} = 7.2921150 \times 10^{-5}\,\mathrm{s}^{-1}$ is Earth's sidereal rotation rate $\omega_{\oplus}$. The hour angle θ_{GMST}, representing the Greenwich Mean Sidereal Time (GMST) for the launch to occur at, is calculated as

$$\theta_{GMST} = \Omega + \lambda_u - \lambda$$

where the auxiliary angle λ_u (see Fig. 8.34), which is the difference between the launch site longitude λ and the target orbit is Right Ascension of Ascending Node (RAAN) Ω, is calculated from Napier's rules as

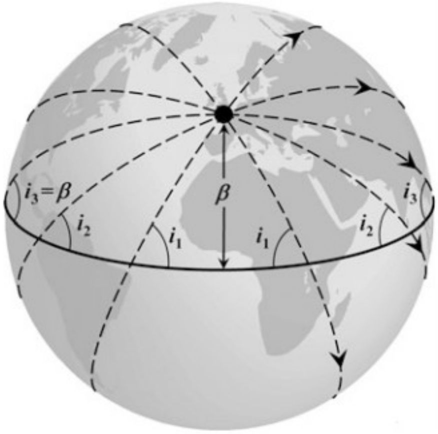

Fig. 8.35 Whatever the launch azimuth is, the accessible orbit inclination i is always larger than the latitude of the launch site β. If the launch azimuth is $\varphi = 90°$ then $i = \beta$ is achieved

$$\lambda_u = \arccos \frac{\cos \varphi_{1,2}}{\sin i} = \pm \arcsin \frac{\sqrt{1 - (\cos i / \cos \beta)^2}}{\sin i}$$

The required hour angle θ_{GMST0} of a particular launch day at hour 00:00 h can be obtained from the Astronomical Almanac of the given year or from

$$\begin{aligned}\theta_{GMST0} =& 100.4606184^\circ + 3600.77005361^\circ \, \mathrm{s}^{-1} \cdot T_{UT1} \\ &+ 0.00038793^\circ \, \mathrm{s}^{-2} \cdot T_{UT1}^2 - 2.6 \times 10^{-8\,\circ} \, \mathrm{s}^{-3} \cdot T_{UT1}^3\end{aligned}$$

The term T_{UT1} denotes the number of *Julian centuries* of the launch day at 00:00:00 h elapsed since the *standard epoch* J2000, which is given as

$$T_{UT1} = \frac{JD - 2\,451\,545.0}{36\,525}$$

where in turn the Julian date *JD* of the given launch day at 00:00:00 h, which should be provided in terms of years $\langle yyyy \rangle$, months $\langle mm \rangle$, and days $\langle dd \rangle$, is computed by

$$\begin{aligned}JD = 367 \cdot \langle yyyy \rangle - INT\left\{1.75 \cdot \left[\langle yyyy \rangle + INT\left(\frac{\langle mm \rangle + 9}{12}\right)\right]\right\} \\ + INT\left(\frac{275}{9} \langle mm \rangle\right) + \langle dd \rangle + 1\,721\,013.5\end{aligned}$$

where the function $INT(x)$ truncates the real number x to the next lower integer number.

Launch Windows

Any deviation from the in-plane launch time would implicate a dogleg maneuver and hence an additional propulsion demand for a plane-change maneuver (see Eq. (8.1.22) and Fig. 8.4 for details). On the other hand, since this precision is impractical to achieve both for organizational and technical reasons, one allows a small amount of time for the launch of the interceptor on both sides of the ideal launch time T_{UT}. This is the so-called **planar launch window**. The width of the planar launch window depends on the launch azimuth. If the launch azimuth is $\varphi \approx 90^\circ$, i.e., if the inclination matches the launch site latitude, the launch window is typically up to 1 h because no later plane adjustment would be necessary. With decreasing launch azimuth any deviation of the launch time from the in-plane time will cause increasing plane differences and hence an increasing propulsion demand to correct them. For Shuttle launches to the ISS $(i = 51.6^\circ)$ the planar launch window decreases to only 10 min. Targets above 57° inclinations have planar launch windows of less than 5 min. In missions to ISS, in practice the Shuttle launch time is appointed to the opening of the launch window, i.e., 5 min before the in-plane time. In case of a launch delay due to a possible countdown problem there remain 10 min to fix the problem. If no problem occurs, the Shuttle is put on hold for 5 min and is launched on in-plane time.

Apart from orbital mechanics there are other restrictions defining other kinds of launch windows on different time scales. An important consideration is the *sun angle*, which is the angle between the direction to the Sun and the targeted orbital plane. The sun angle is important for visibility conditions during final approach (cf. Fig. 8.45 where the Shuttle needs to see the ISS during daytime conditions, in particular during final approach) and for solar power generation of the docked spacecraft. For the Space Shuttle another concern is the ability to monitor its ascent and to visually check the external tank for damages during launch and ascent.

When launching toward ISS, traffic conditions also impact launch window planning, since multiple spacecraft such as Soyuz, Progress, ATV, HTV, or Shuttle want to approach the station, during beneficial lighting conditions. All these factors must be considered in mission planning, which leads to the small number of launch windows available to ISS or Hubble per year. This is the reason why the failure to launch a mission during the originally intended window can cause launch delays of months, instead of hours or days as dictated by orbital mechanics. Also the so-called *phase window* (see next section) restricts the launch opportunities.

For a given launch site the range of permitted launch azimuths is usually restricted due to safety concerns of flying a launch vehicle over densely populated areas. For instance, the launch azimuth restrictions at Kennedy Space Center are $35^\circ \leq \varphi \leq 120^\circ$, where the lower bound is due to the US West Coast and the upper bound to the Caribbean islands. This restricts the Shuttle to have only one launch opportunity per day to the ISS.

8.6.2 Phasing

After successful completion of the launch phase, the interceptor spacecraft achieves a stable orbit within the same plane as the target. The two orbits are thus coplanar and typically near circular. (Alternatively, the interceptor is in a plane from which the target plane can be reached within the capabilities of the orbital maneuvering system of the interceptor.)

However, the target might be anywhere on its orbit. Therefore, the first part of the target rendezvous, the so-called *far range rendezvous*, requires first a reduction in the distance to the target, until it can be acquired by the sensors of the interceptor, and then a transfer to a stable holding point on the trailing side of the target.

This first part of far range rendezvous phase is called *phasing* because it is to reduce the so-called *(orbital) phase angle* ϑ , which is the difference in true anomaly as measured in the flight direction from the target to the interceptor. Phasing is typically conducted in absolute navigation, i.e., with reference to an inertial reference frame, and guided by ground control. As an example, Space Shuttle phasing maneuvers are planned by mission control using orbit determination data obtained by ground radar and *Tracking and Data Relay Satellite System*

(TDRSS) Doppler measurements. By processing both the target and interceptor tracking data, the orbital phase angle is determined.

Usually, the interceptor needs to chase the target, so only negative initial phase angles are permitted. This is why the interceptor is often called *chaser*. In addition, only certain initial phase angles are permitted. Owing to performance limitations and constrained crew activities this so-called **phase window** (window accounting for phase angle makeup capability) varies for the Shuttle between $40^\circ \le -\vartheta_i \le 360^\circ$. The relationship of the phase window to the planar launch window changes each day and depends on the target's orbital period and inclination.

Owing to a given initial phase angle to be made up, the interceptor will have finalized its launch trajectory on a generally slightly elliptic or circular orbit with its semi-major axis a_I smaller than the a_T of the target orbit. Whether the initial apogee coincides with the target orbit depends on the particular rendezvous strategy, but usually the interceptor orbit nowhere crosses the target orbit. Because $a_I < a_T$ the orbital frequency (mean motion) is larger than that of the target, which implies that the orbital phase angle is reduced continuously—the goal of phasing.

How much is the phase angle reduced in course of one orbital revolution? Assuming small differences in the semi-major axes, which is a quite good approximation for LEO and since $n = \sqrt{\mu/a^3} = 2\pi/T$ of the target the phase reduction after one orbital revolution is

$$\Delta\vartheta = -\Delta n \cdot T = -\left(\frac{dn}{da} \cdot \Delta a\right) \cdot T = \frac{3}{2}\frac{n}{a} \cdot \Delta a \cdot T$$

Therefore

$$\Delta\vartheta = 3\pi \frac{\Delta a}{a} \text{ per orbit} \tag{8.6.3}$$

So, the phase angle from the interceptor (behind the target) to the target decreases if $a_I < a_T$, i.e., if $\Delta a = a_I - a_T < 0$. This phase angle reduction relates to a closing distance of

$$\Delta x = \Delta\vartheta \cdot a = -3\pi \cdot \Delta a \quad \text{per orbit}$$

To give an example, the ISS orbits Earth at an altitude of about $h = 350\,\text{km}$. Owing to drag becoming too excessive, the lowest chaser altitude is limited to about $h = 250\,\text{km}$. This implies $\Delta a \ge -100\,\text{km}$.

So, for instance the Soyuz is able to approach the ISS typically with $\Delta\vartheta_{\max} = 7^\circ$ per orbit equaling $\Delta x_{\max} = 820\,\text{km}$ per orbit. To limit the physical and fatigue stress to astronauts in the small S/C one typically sets a limit of two days to get to ISS docking. Accounting also for the first checkout orbit, the homing, closing, and final approach phases, the number of Soyuz phasing orbits are typically limited to 21 orbits equaling a maximum initial phase angle of $\Delta\vartheta_i \le 150^\circ$. This of course has a grave impact on the possible launch windows. To ease this problem and to even further reduce the stress to astronauts the trajectory department of RSC Energia in March 2013 implemented a new rendezvous profile for Soyuz TMA-08M and for later Soyuz missions, where the phasing orbits were reduced to 5 orbits or less (typically 4), which has become known as the *fast rendezvous profile* a.k.a. *four-orbit rendezvous profile*. This fast profile was tested first on the unmanned ISS supply mission Progress M-16M in August 2012, while it was first applied to the manned mission Soyuz TMA-08M in March 2013 and to all manned Soyuz mission since then. This new 4-orbit rendezvous profile implies and initial phase angle of $\Delta\vartheta_i 15^\circ - 20^\circ$, which is achieved by adjusting the ISS orbit raising maneuvers months before launch day such that the ISS falls within this initial phase angle range.

8.6.3 Homing Phase

The interceptor is now on a phasing orbit about 50 km away from the target drifting slowly toward the target. The objective of the upcoming homing transfer, the second part of the far range rendezvous, is to transfer the interceptor to a stable holding and aiming point in the vicinity of the target (see Figs. 8.33 and 8.36). A prerequisite of the transfer is that the target must be acquired by the relative navigation sensors of the interceptor. For the Space Shuttle, mission control hands over rendezvous guidance to the orbiter's crew at 74 km from the target. At this point a target like the ISS can be tracked using star trackers or radar.

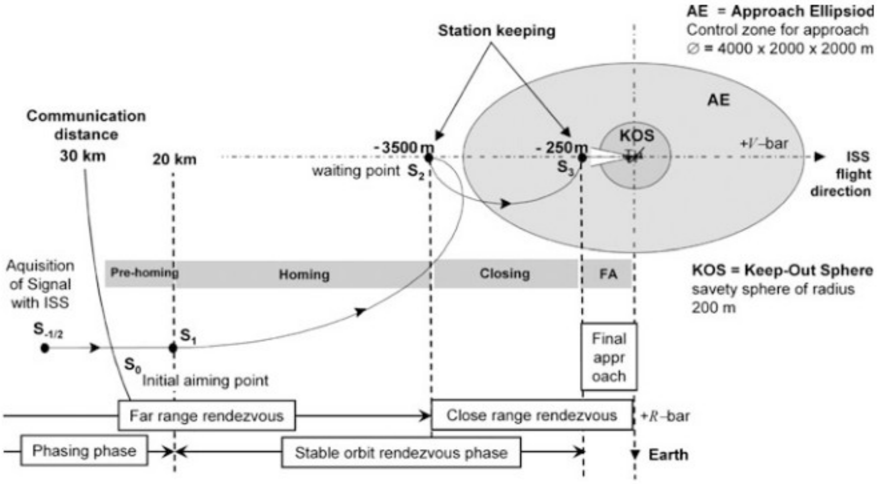
AE = Approach Ellipsiod
Control zone for approach
∅ = 4000 x 2000 x 2000 m
Station keeping
AE
Communication distance
30 km
20 km
-3500 m
waiting point S_2
- 250 m
S_3
KOS
$+V$-bar
ISS flight direction
Aquisition of Signal with ISS
Pre-homing
Homing
Closing
FA
KOS = Keep-Out Sphere
savety sphere of radius 200 m
$S_{-1/2}$
S_1
S_0
Initial aiming point
Final appr oach
Far range rendezvous
Close range rendezvous
$+R$-bar
Phasing phase
Stable orbit rendezvous phase
Earth

Fig. 8.36 Homing, closing, and final approach profile and phases for ISS rendezvous. ISS safe-approach procedures require station-keeping points S on the V-bar, an approach corridor, and a keep-out sphere around ISS that approaching spacecraft must use. *Credit* Wartenberg and Amadieu/ESA

With the homing maneuver also the relative approach velocity must be reduced to a safe level. In addition, the dispersions in position, orientation, and angular rate must be reduced to meet the conditions required for the upcoming close range rendezvous. This includes the synchronization of the motion timeline of the two spacecraft.

LVLH Reference Frame

For the discussion of the now following rendezvous approaches, the *Local Vertical Local Horizontal* (LVLH) reference frame is defined (see Fig. 8.37). The origin of LVLH is located at the center of mass of the target. Its +*x*-axis, also called the +V-bar, points along the target's velocity vector. The −*z*-axis, referred to as +R-bar, points antiparallel to the target's radial vector. The −*y*-axis, also called +H-bar, completes the right-handed system and thus points along orbit normal. In the following a "+V-bar approach" means that the interceptor approaches the target on the target's +V-bar (into −V-bar direction). Accordingly, a "−R-bar approach" is on the target's −R-bar (into +R-bar direction), etc.

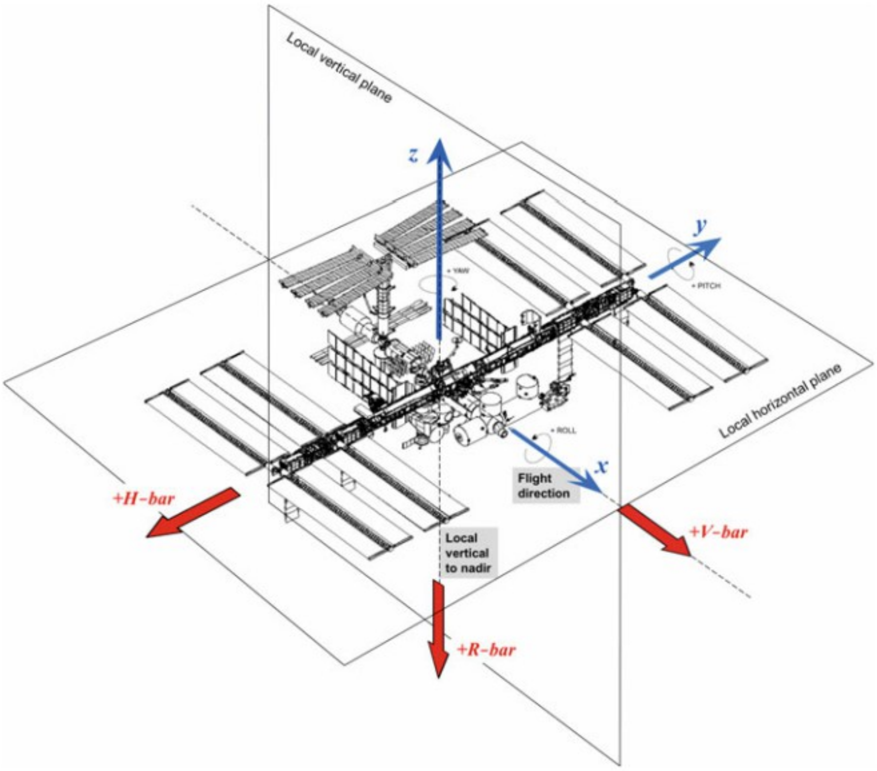

Fig. 8.37 Local vertical local horizontal (LVLH) reference frame: +V-bar (+*x*-axis) is in the direction of the spacecraft's velocity vector, +R-bar (+*z*-axis) is in the direction of the negative radius vector, and +H-bar (−*y*-axis) completes the right-handed system

Homing Transfer

The homing transfer, which commences at point S1 in Fig. 8.36, is a classical Hohmann transfer as described in Sect. 8.3. The principle situation is shown in Fig. 8.38, where the orbits, however, are not to scale because at the end of the phasing phase, the two orbits with $a \approx 6750\,\text{km}$ in LEO have a radial distance of typically only about 10 km. Owing to this, the phase angle at the beginning of the Hohmann maneuver is practically zero while its complement to 180°, the so-called *lead angle*, is $\alpha_L = 180^\circ - \vartheta \approx 180^\circ$.

Let ϑ_i be the initial phase angle and let ϑ_f be the final phase angle at S2 behind the target. The key question is: At a given a, Δa, and ϑ_f, what is the right ϑ_i and the right tangential kick-burn Δv to get to S2? To find an answer we apply from Sect. 8.3.2 the essential results of a Hohmann transfer to adjacent circular orbits as for rendezvous orbits

$$a_H = \frac{1}{2}(a_I + a) = a\left(1 + \frac{\Delta a}{2a}\right)$$

$$t_H = \pi\sqrt{\frac{a_H^3}{\mu}}$$

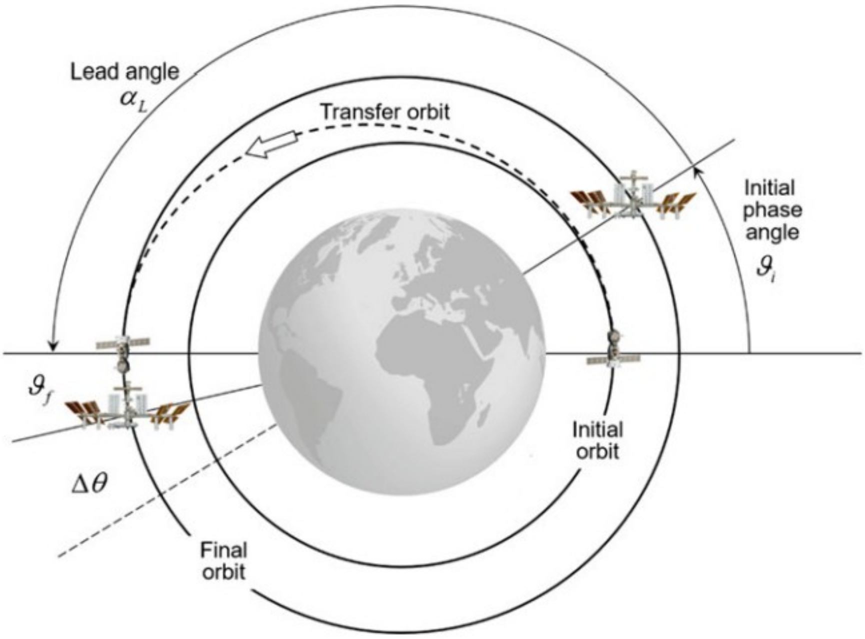

Fig. 8.38 General homing transfer of an interceptor (Soyuz) to meet a target point behind a target (ISS) in a circular coplanar orbit. Initially the required phase angle is ϑ and the interceptor begins the phasing maneuver by applying a thrust. After a half revolution on the transfer orbit, the interceptor reaches the target point, where a thrust maneuver brings the interceptor to a steady relative position behind the target and thus finalizes the Hohmann transfer

Now, while the interceptor transits over an orbit angle of 180° on the Hohmann transfer orbit, the target with $n = \sqrt{\mu/a^3}$ covers the orbit angle (see Fig. 8.38)

$$180^\circ - \Delta\theta = n \cdot t_H = 180^\circ \cdot \left(\frac{a_H}{a}\right)^{3/2} = 180^\circ \cdot \left(1 + \frac{\Delta a}{2a}\right)^{3/2} \approx 180^\circ \cdot \left(1 + \frac{3}{4}\frac{\Delta a}{a}\right)$$

Because in LEO $\Delta a/a \approx 10^{-3}$, we could safely neglect terms of higher order. Therefore, we have for the initial phase angle (see Fig. 8.38)

$$\vartheta_i = \vartheta_f + \Delta\theta = \vartheta_f - 180^\circ \cdot \frac{3}{4}\frac{\Delta a}{a}$$

or

$$\boxed{\vartheta_i = \vartheta_f - 135^\circ \frac{\Delta a}{a}} \tag{8.6.4}$$

For the required delta-v, we find with a circular orbital velocity $v = \sqrt{\mu/a}$

$$\Delta v = v_I - v_T = v_T\left(\sqrt{\frac{a}{a + \Delta a}} - 1\right) = v_T\left(\sqrt{1 - \frac{\Delta a}{a}} - 1\right)$$

and hence

$$\boxed{\Delta v = -\frac{1}{2}\frac{\Delta a}{a}\sqrt{\frac{\mu}{a}}} \tag{8.6.5}$$

> *Example*
>
> The Shuttle shall perform a homing maneuver from $\Delta a = -10\,\text{km}$ to a waiting point $S2 = 3\,\text{km}$ behind the ISS at $h = 350\,\text{km}$.
>
> We have $\vartheta_f = (180^\circ/\pi) \cdot (10/6728) = 0.085^\circ$ and $v_T = 7.697\ \text{km s}^{-1}$. With this we get $\vartheta_i = \vartheta_f + 0.201^\circ = 0.286^\circ$. This means the Shuttle has to perform a burn with $\Delta v = 5.72\ \text{m s}^{-1}$ into flight direction at a position 33.6 km behind the ISS. Because the Shuttle initially is 10 km below the ISS, the viewing distance to the ISS at burn is $\sqrt{10^2 + 33.6^2} = 35.0\,\text{km}$

NASA's Space Rendezvous History

At NASA there exist two general approaches for homing and closing maneuvers: The historical *coelliptic rendezvous* and today's *stable orbit rendezvous* (SOR).

Coelliptic Rendezvous

Coelliptic orbits are coplanar elliptic (including circular) orbits with a common occupied focus (see Fig. 8.39). The arguments of perigee ω are equal, meaning that the lines of apsides of the orbits are congruent. In addition to this, the differences in perigee and apogee radii are equal. In a spacecraft-fixed reference frame, coelliptic orbits appear as two parallel lines. These particular orbits allow for easy, intuitive,

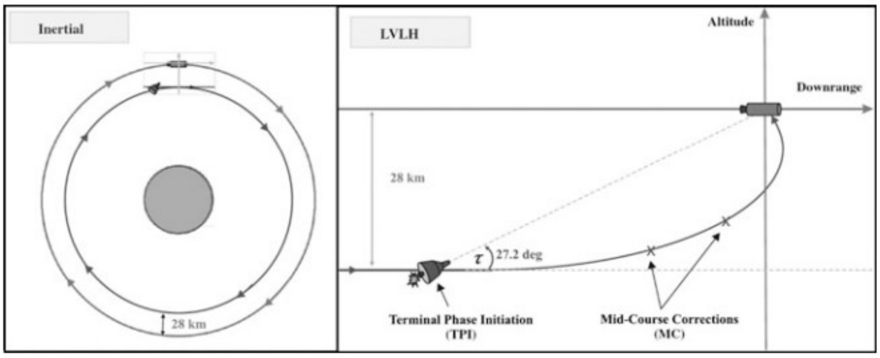

Fig. 8.39 Gemini coelliptic rendezvous in inertial and spacecraft reference frames; τ denotes the trigger angle between the direction to the target and the local horizontal plane at which the transfer maneuver is triggered. *Credit* Woffinden and Geller (2007)

and robust maneuver planning by means of so-called *trigger angle targeting*. This technique was developed during the Gemini program. It allows astronaut pilots to reliably achieve rendezvous by pointing the interceptor spacecraft at the target at a certain trigger angle τ (see Fig. 8.39) relative to the direction of flight and then engaging the orbital maneuvering thrusters, the so-called *Terminal Phase Initiation* (TPI). The trigger angle can be measured with simple cueing devices similar to a sextant. During the transfer, the target will move along its orbit within the so-called *transfer angle* α (the orbit angle covered during transfer). If rendezvous is achieved within a single revolution of the target, thus with $\alpha < 360°$, it is called a *direct rendezvous*. Any case with multiple target revolutions is referred to as *indirect rendezvous*. The point aimed at by the interceptor, in front of or behind the target, is referred to as the *downrange targeting location*.

It can be shown (see Woffinden (2007)), and this is a key property of coelliptic rendezvous, that for direct rendezvous, the elevation trigger angle τ is independent of the angular velocity ω and the relative altitude of the coelliptic trajectory. Hence, the same trigger angle applies for all orbits, regardless of the coelliptic height differential. For indirect rendezvous, the trigger angles show dependency on the ratio between the downrange targeting location and the relative altitude of the coelliptic trajectory. In recent numerical simulations it was determined that the optimal trigger angle for a minimum-Δv intercept maneuver is 27.0°, with a transfer angle of 163.1°. Interestingly, this is not equal to a Hohmann transfer with a trigger angle of 0° and a transfer angle of 180°. In reality, this optimal angle will not be perfectly achievable. A range between 26.8° and 27.3° was therefore identified as providing optimal combinations of required Δv, Line-of-Sight (LOS) approach rates, and positioning accuracy.

Interestingly, without running numerical computer simulations, NASA selected a trigger angle of 27° with a transfer angle of 130° for Gemini/Apollo rendezvous operations. For these manned rendezvous missions trigger angle targeting was particularly attractive. It allowed the use of the astronauts' eyes and simple elevation cueing for maneuver triggering in case of the failure of the rendezvous radar system. After applying the initial Δv along LOS, the pilot performed one or two mid-course correction maneuvers, before finally approaching the target for docking. During final approach, the pilot benefited from a low inertial LOS approach rate during final breaking and approach, as well as from good visibility of the target against the star background. This was important since LOS closing rates are difficult to judge visually without ambient references.

Coelliptic rendezvous remained NASA's rendezvous approach of choice throughout Gemini and Apollo. Its major strength was the backup capability to perform the TPI burn manually. It was then modified into a *dual coelliptic rendezvous* profile for the Skylab missions. In this profile, the interceptor flew a coelliptic transfer onto a holding orbit below the target. From there another coelliptic maneuver finalized rendezvous. The modifications were applied to improve the final approach lighting conditions for manual piloting, as well as the quality of long-range optical tracking using reflected sunlight. This dual coelliptic profile was then also the baseline for Shuttle R&D missions. Given the characteristics of typical Shuttle rendezvous targets, there existed concerns regarding the usability and quality of optical tracking of small target objects using reflected sunlight in the presence of Earth's illuminated surface and bright celestial objects. Another issue was the depletion of the Shuttle's *Reaction Control System* (RCS) propellant due to high relative approach velocities. This initially led to the adoption of a so-called *tuned coelliptic rendezvous* (TCR) profile.

Stable Orbit Rendezvous (SOR)

For current space station operations all these coelliptic rendezvous approaches were replaced by the so-called *stable orbit rendezvous*. Such a SOR was first flown on Gemini XI and was later suggested to address the concerns over target tracking and propellant consumption for Space Shuttle R&D. The reason is that it supports inertial approaches with lower relative velocity than the inertial approaches from the Apollo legacy coelliptic profile. In addition, a stable orbit profile desensitizes the mission timeline from trajectory considerations, as the interceptor could theoretically remain at the waiting point for indefinite periods of time. Stable orbit station-keeping at multiple kilometers of distance to the target (15 km for Space Shuttle ISS approaches) was also preferable to the close-range (at distances of tens of meters) station-keeping associated with coelliptic approaches. In such close proximity, continuous crew monitoring and frequent correction maneuver are needed, resulting in high propellant expenditure. Therefore, the advantages of SOR profiles over coelliptic approaches are lower propellant consumption and stable station-keeping points on V-bar, leading to less demand on crew position monitoring and correction. Hence, the SOR has become the standard for ISS operations for Space Shuttle, Soyuz and ATV, as well as for other rendezvous operations, such as with the Hubble Space Telescope.

For more details on coelliptic and Space Shuttle R&D missions refer, e.g., to Goodman (2006).

8.6.4 *Closing Phase*

Once on the target orbit at S2 (see Fig. 8.36) jets are fired to bring the interceptor to a hold at a safe distance about 3 km behind the target. The target now is within range of the interceptor sensors and thus relative navigation can commence. This station-keeping point S2 is essential to assess the situation and plan the upcoming closing maneuver.

The closing maneuver depends on the type of final approach: For a final ±R-bar approach the interceptor needs to get to the ±R-bar below/above the target. For a final +V-bar approach the interceptor needs to proceed further on the −V-bar closer to the target to the station-keeping point S3, and for a +V-bar approach the interceptor has to fly around the target to approach it from the leading end. Anyway, with the closing maneuver we ingress the *Approach Ellipsoid* of the ISS. All operations inside the Approach Ellipsoid are “combined operations” involving the mission control authorities in Houston and Moscow. From here on safety as not to collide with the ISS has the highest priority.

−V-Bar Approach

Let us assume that the docking port is on the trailing end of the target and therefore an approach further on the −V-bar is favorable. What are the options to carry it out? In Sect. 8.5.3 we have seen that we may approach the target on the V-bar either via a prolate cycloid (which is a Hohmann trajectory; see end of section “Ellipse” in Sect. 8.5.3) or via an ellipse. The decision is based on safety versus efficiency: If in course of the prolate cycloid we would lose control over the interceptor we would drift away infinitely and if it is a “flat” cycloid we might even hit the target. However, if safety is not paramount then one could traverse the distance from S2 to S3 by one or several cycles keeping the momentum along the V-bar and hence save fuel. At each reversal point one could even stop the approach, assess the situation, and fine-tune the further approach. This option is chosen by NASA for the Shuttle closing phase (see Fig. 8.45). If safety is top priority then the ellipse trajectory is the choice, because if control over the interceptor would be lost, it would automatically return to S2. However, this method requires to fire the jet at every intersection with the V-bar, thus coming to a halt and thereafter repeat the whole cycle procedure. “Safety-first” requirement of course comes at the expense of a higher propulsion demand.

So, for +V-bar approach of a “safety-first” rendezvous with the ISS as shown in Figs. 8.36 and 8.40 the ellipse maneuver is used. As shown in Sect. 8.5.3, in particular Figs. 8.29 and 8.30, a vertical (radial) burn v_0 will bring us on an ellipse to the next waiting point S3 that lies by $\Delta x = S2 - S3 = 4v_0/n < 0$, with $S2 < 0,\ S3 < 0$, closer to the target. When arriving at S3 a reverse burn must be fired to bring the interceptor to a halt. For the required delta-v for such a step of width Δx we find from Fig. 8.30

$$\boxed{\Delta v = \pm 2 \times \frac{\Delta x}{4}\sqrt{\frac{\mu}{a^3}}} \quad \text{@ elliptic trajectory} \qquad (8.6.6)$$

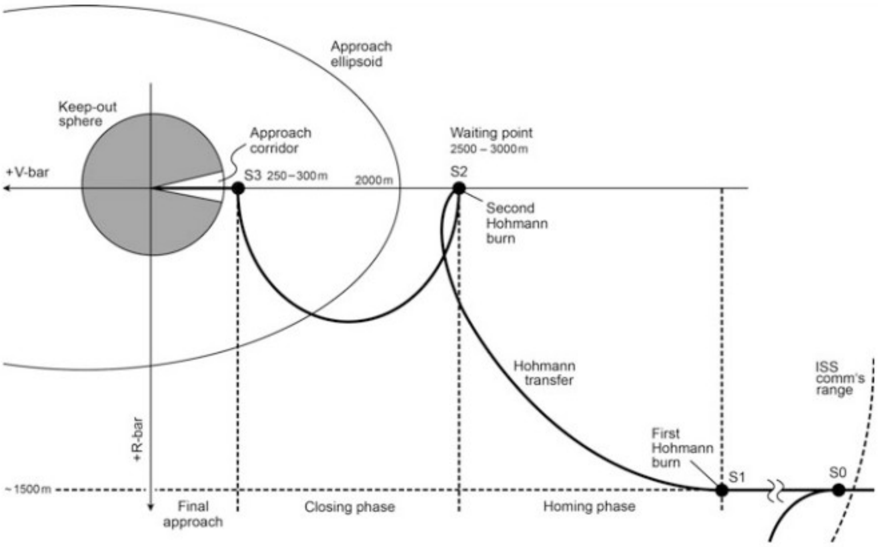
Approach
ellipsoid
Keep-out
sphere
Approach
corridor
Waiting point
2500 – 3000m
+V-bar
S3 250–300m
2000m
S2
Second
Hohmann
burn
Hohmann
transfer
ISS
comm's
range
+R-bar
First
Hohmann
burn
S1
S0
~1500m
Final
approach
Closing phase
Homing phase

Fig. 8.40 ISS stable orbit approach on −V-bar as typically adopted by an ATV rendezvous. First, a Hohmann transfer brings the interceptor to the waiting point S2. Then, it approaches the ISS on an elliptic trajectory to waiting point S3. *Credit* Fehse (2003) and U. Walter

where the factor 2 indicates that we need two burns for the entire approach maneuver and the ± sign that the elliptic trajectory might be on either side of the V-bar. This maneuver can be performed at any step size and as often as wanted to get to S3. Observe that it does not make any difference for the total delta-v if a given distance is covered with more or less steps because $\Delta v \propto \Delta x$. How many increments are to be chosen is just a matter of safety and time (every step lasts one orbital period). So, if Δx is the total distance between S2 and S3, Eq. (8.6.6) provides the delta-v for the entire closing transfer independent of the number of steps. To approach the ISS, every incremental distance of 1 km requires $\Delta v = 2 \cdot 0.286\,\mathrm{m\,s^{-1}}$.

If the cycloidal approach would be chosen, we have from Fig. 8.27 $\Delta x = 6\pi v_0/n$ and hence

$$\boxed{\Delta v = \pm 2 \times \frac{\Delta x}{6\pi \cdot k}\sqrt{\frac{\mu}{a^3}}} \quad \text{@ cycloidal trajectory} \tag{8.6.7}$$

where Δx is the total distance between S2 and S3 and k is the number of cycloidal revolutions to traverse this distance. Here again the factor 2 indicates that we need a initiation burn at S2 plus a stop burn of equal absolute value at S3. We therefore see that the cycloidal approach is by a factor $6\pi \cdot k/4 = 4.7 \cdot k$ more efficient that the elliptic approach—at the expense of safety. We recall that every cycloidal or elliptic cycle takes a full orbital period.

Approach to +V-Bar

If one wants to approach on the –V-bar from S2 to the waiting point S3 on the +V-bar, i.e., to the other side of the target, one has to apply the same maneuvers (elliptic or cycloidal trajectories) as above, the only difference being that for $\Delta x = \mathrm{S2} - \mathrm{S3} < 0$ holds $\mathrm{S2} < 0,\ \mathrm{S3} > 0$.

Approach to ±R-Bar

If the final approach is on +R-bar or −R-bar the closing phase must bring the interceptor in a loop from −V-bar to ±R-bar (see Fig. 8.41) at below/above the target. Here again both the cycloidal and elliptic trajectory would do, but here both approaches are equally safe. This is because if control is lost for a cycloidal trajectory it would just pass by the target at the aimed distance and after that drift away from the target. In the elliptic trajectory case the interceptor would just orbit the target on an ellipse. Because we have seen that the cycloidal trajectory is much more efficient it is the preferred approach method.

The cycloidal loop is performed in two (or one) steps. First, a tangential burn Δv brings the interceptor on a prolate cycloid (Hohmann trajectory) to the transition point $\mathrm{S3}_-$ where a burn Δv is fired against the direction of movement. This brings the interceptor into a circular Earth orbit below/above the ISS. On this orbit it slowly drifts to the point S3 below the ISS from where the final approach commences. Note that neither $\mathrm{S3}_-$ nor S3 are stable points with respect to the ISS. If the first burn is accurate enough it can be set such that $\mathrm{S3}_-$ coincides with S3.

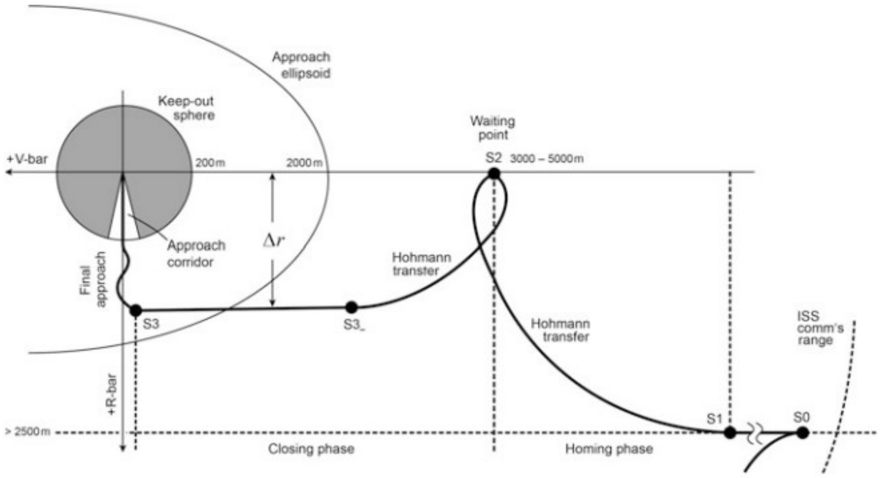
Approach
ellipsoid
Keep-out
sphere
+V-bar
200m
2000m
Waiting
point
S2 3000 – 5000m
Δr
Approach
corridor
Final
approach
Hohmann
transter
S3
S3_
Hohmann
transfer
ISS
comm's
range
+R-bar
> 2500m
S1
S0
Closing phase
Homing phase

Fig. 8.41 From −V-bar to +R-bar approach to ISS. A Hohmann transfer brings the interceptor first to the intermediate point $\mathrm{S3}_-$ where it crosses over into a circular orbit on which it drifts to the final point S3. *Credit* Fehse (2003) and U. Walter

What needs to be determined is: Given the initial point S2 and a radial distance Δr (counted positively outward along the z-axis) the interceptor shall dive below $(\Delta r < 0)$ or above $(\Delta r > 0)$ the ISS, what is the delta-v to perform the entire maneuver? From Fig. 8.27 we derive that the diving distance is $\Delta r \equiv \Delta z = 4v_0/n$. We therefore find for the initial delta-v

$$\boxed{\Delta v_i = \frac{\Delta r}{4}\sqrt{\frac{\mu}{a^3}}} \quad \text{@ initial burn} \tag{8.6.8}$$

Upon diving down/up, the interceptor speeds up and according to Eq. (8.5.14) achieves at the lowest/highest point of the trajectory, which is S3$_-$, the relative velocity $\dot{x} = v_0(4\cos\pi - 3) = -7v_0$. This corresponds in an Earth centered system to

$$v_- = \sqrt{\frac{\mu}{a}} - 7v_0$$

At $r = a + \Delta r$, the anticipated circular orbit has an orbital velocity of

$$v_+ = \sqrt{\frac{\mu}{a + \Delta r}} = \sqrt{\frac{\mu}{a}}\left(1 - \frac{1}{2}\frac{\Delta r}{a}\right) = \sqrt{\frac{\mu}{a}} - 2v_0$$

where the latter follows by applying Eq. (8.6.8). For the delta-v of the braking burn at S3$_-$, we therefore obtain

$$\boxed{\Delta v_f = v_- - v_+ = -5v_0} \tag{8.6.9}$$

So the total delta-v is

$$\Delta v = |\Delta v_i| + |\Delta v_f| = 6 \times \frac{\Delta r}{4}\sqrt{\frac{\mu}{a^3}}$$

Thus, while for the −V-bar approach our effort to get to S3 is maximally $\Delta v = 2|v_0|$ (elliptic trajectory), it is much bigger, namely $\Delta v = 6|v_0|$, in the case of an approach to ±R-bar.

8.6.5 Final Approach

The interceptor is now on the waiting point S3 just outside of the Keep-Out Sphere, about 200 m away from the ISS (see Fig. 8.42). This is where the *final approach* (a.k.a. *proximity operations or terminal phase*) begins. Other than during the stable orbit rendezvous where the interceptor approached the target via stable orbits (i.e., cycloidal trajectories that can be stopped at the reversal points or elliptic trajectories that even back off from the target if rendezvous control is lost and therefore are "safe" orbits), the trajectories now are more or less straight to directly intercept the

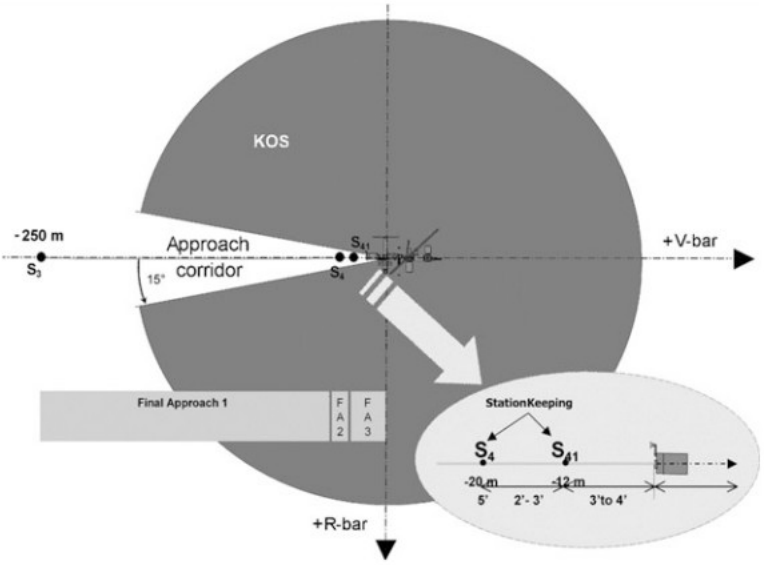

Fig. 8.42 ISS approach corridor for V-bar approach. *Credit* Wartenberg and Amadieu/ESA

target and therefore are on "collision course" with it. The final approach ends at a distance of a few meters upfront the ISS, either when docking is imminent or when the target is within capture distance of the manipulator. During this phase the spacecraft usually maneuvers autonomously, i.e., without intervention by ground control. The spacecraft control loop must therefore be closed locally, either by the crew or by the automatic controllers. Final approach can thus be considered the most critical part of the R&D mission. During this phase, minor errors can cause accidents.

±V-Bar approach

First, we consider the approach on +V-bar or on −V-bar (see definition of "±V-bar approach" in the section *LVLH Reference Frame* above) from waiting point S3 through the conical approach corridor with half angle 15° to the next waiting point S4 (the initial waiting point for docking approach). To penetrate the cone one could continue to apply the ellipse maneuvers. However, every incremental step takes one orbital period of 91.5 min in case of the ISS, which is far too much.

One therefore switches to another approach mode called *forced translation.* Figure 8.43 shows its principle. Let us assume a burn is performed aiming directly at the target. According to Fig. 8.27, the higher speed would force the interceptor to drift upward (centrifugal force is bigger than the gravitational force at this circular altitude), violate the approach cone, and never encounter S4 (see Fig. 8.43). However, we can counteract the updrift by providing in addition to the forward translation an initial small downward force. If this is done properly, it forces the

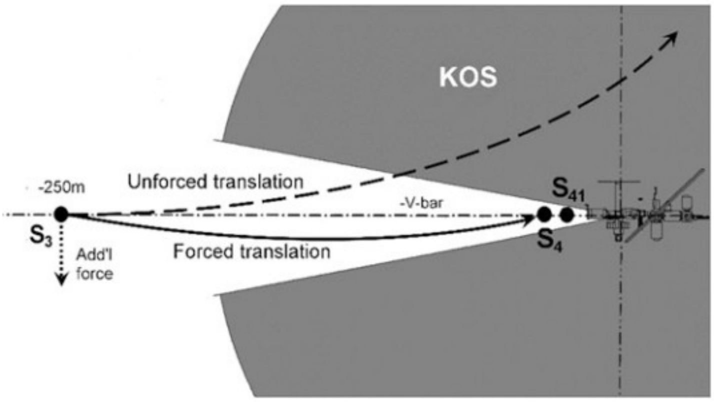

Fig. 8.43 A straight trajectory would turn up and leave the approach corridor. A downward forced translation enforces the trajectory to stay in the corridor and hit the waiting point

initial trajectory slope down just that the trajectory at its end hits S4. The forward translation with this extra little downward force is called *forced translation*.

It needs to be determined how big the additional delta-v, equivalent to the downward force, is. We therefore revisit Sect. 8.5.2 where we have shown that for translation times $(nt)^2 \ll 1$ Eq. (8.5.13) holds. These equations give the answer to the question what the initial velocity $\boldsymbol{v}_0 = (\dot{x}_0, \dot{y}_0, \dot{z}_0)$ at the initial point (x_0, y_0, z_0) should be in order to meet after time *t* a given target point at the origin $(0, 0, 0)$. In our case the vector from the target point to the initial distance of the interceptor is $(x_0, y_0, z_0) = (\pm\Delta x, 0, 0)$ for a ± V-bar approach. Therefore, Eq. (8.5.13) reduces to

$$\boldsymbol{v}_0 = \left(v_{0x},\ v_{0y},\ v_{0z}\right) = \left(\mp\frac{\Delta x}{t},\ 0,\ \pm n \cdot \Delta x\right) \quad @ \pm \text{V-bar approach} \qquad (8.6.10)$$

$$\Delta\boldsymbol{v} = \boldsymbol{v}_0 - \boldsymbol{v}_i$$

where $n = \sqrt{\mu/a^3}$, $\Delta\boldsymbol{v}$ is the initiation burn vector, and $\boldsymbol{v}_i$ is the interceptor's velocity incident of the starting point S3 (if the starting point was a waiting point, $\boldsymbol{v}_i = 0$). Of course, for braking the forced translation at S4 the same amount of Δv is required, however, for v_{0x} into the opposite direction. The initial speed in *x*-direction, v_{0x}, is easy to grasp: The required speed is distance divided by flight time. The delta-v in negative *z*-direction (radial thrust downward) is just the forced part. Assuming that the delta-v in each direction is generated by separate thrusters (as usually the case), the absolute value of the total approach delta-v is

$$\Delta v = 2\left(\frac{\Delta x}{t} + n \cdot \Delta x\right) = 2 \cdot \Delta x \cdot \left(\frac{1}{t} + n\right)$$

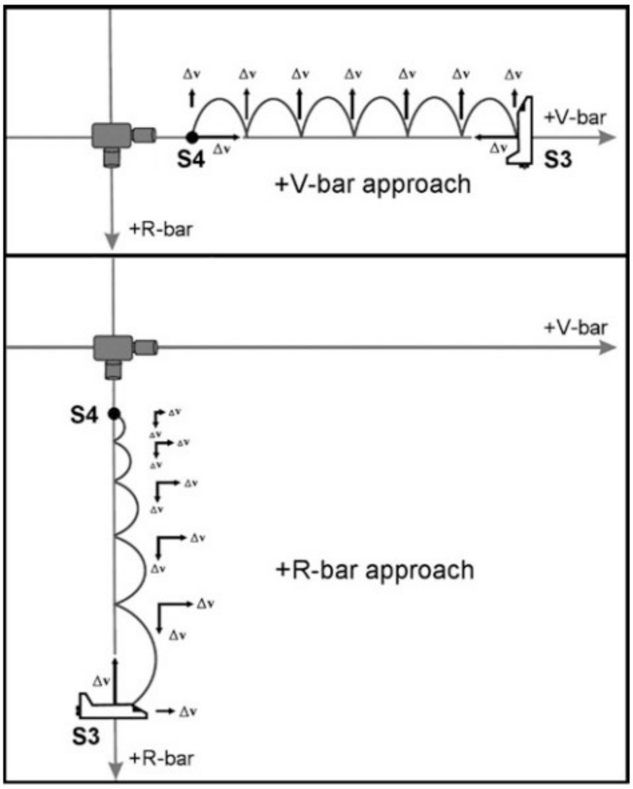

Fig. 8.44 Final approach hopping profiles of a Shuttle final approach along the +V-bar (above) and the +R-bar (below, including deceleration burns). *Credit* Woffinden and Geller (2007)

It would not be wise to cover the final approach distance in one move. Initial point and/or firing errors would jeopardize encountering exactly the target point. One rather splits the distance into two or more parts and whenever the trajectory meets the V-bar a new forced delta-v of $\pm 2 \times n \cdot \Delta x$ is applied starting the forced translation anew. This leads to a hopping approach as depicted in Fig. 8.44 (top) for a Shuttle +V-bar approach. This is why it is called *hopping trajectory*. Because the total delta-v has the linear dependency $\Delta v \propto \Delta x$, it does not make any difference in terms of delta-v effort to split the final approach distance into k shorter hops of length ε, $\Delta x = k \cdot \varepsilon$, but still cover the total distance in the same time t, or not. However, owing to boost errors, it is preferable to make more shorter hops.

±R-Bar Approach

Forced translation can also be applied on the +R-bar or −R-bar, which for instance is performed by a Soyuz or Progress docking to the radial port (Pirs docking compartment) of the ISS. This approach was originally designed for docking the Shuttle with *MIR*. It was also used on Hubble servicing mission STS-82. To determine the initiating delta-v for the forced translation from S3 to S4 we have from Eq. (8.5.13) with the initial distance vector $(0, 0, \mp\Delta z)$ for ±R-bar approach

$$\boldsymbol{v}_0 = \left(v_{0x},\ v_{0y},\ v_{0z}\right) = \left(\pm n \cdot \Delta z,\ 0,\ \pm\frac{\Delta z}{t}\right) \quad @ \pm \text{R-bar approach} \qquad (8.6.11)$$

$$\Delta\boldsymbol{v} = \boldsymbol{v}_0 - \boldsymbol{v}_i$$

where $n = \sqrt{\mu/a^3}$, $\Delta\boldsymbol{v}$ the initiation burn vector, and $\boldsymbol{v}_i$ is the interceptor's velocity incident to the starting point S3. The total approach delta-v is by the same token as above

$$\Delta v = 2 \cdot \Delta z\left(\frac{1}{t} + n\right)$$

And also in this case owing to the linear dependency $\Delta v \propto \Delta z$ it does not make any difference in terms of delta-v effort to split the distance into more shorter hops. Thus, we might have also hopping trajectories as shown in Fig. 8.44 (bottom).

Proximity Operations

From an orbital mechanics point of view the last few meters starting out from S4, the docking approach in the proximity of the target, is the most easy part of the rendezvous. This is because for very short distances $x,\ z,\ \tau \to 0$ and hence

$$\begin{aligned} v_{0x} &= -\frac{x_0}{t} - nz_0 & v_{0x} &= -\frac{x_0}{t} \\ v_{0z} &= -\frac{z_0}{t} + nx_0 \ \rightarrow & v_{0z} &= -\frac{z_0}{t} \\ v_{0y} &= -\frac{y_0}{t} & v_{0y} &= -\frac{y_0}{t} \end{aligned} \qquad (8.6.12)$$

$$\Delta\boldsymbol{v} = \left(v_{0x},\ v_{0y},\ v_{0z}\right) - \boldsymbol{v}_i$$

where $\boldsymbol{v}_i$ is the interceptor's velocity incident to the starting point S4. Therefore, if on this docking approach the commander or pilot navigates the interceptor, steering becomes intuitively easy because the required momentary speed is just distance per time. However, in order not to crash into the docking port the approach speed has to be continuously reduced (see Fig. 8.44, bottom).

Docking/Capture Phase

At the end of the final approach phase, the interceptor is in position in front of the target's docking port or capture interface and all thrusting has ceased. The interceptor's relative velocity is either zero for capture & berthing, or slightly above zero for docking. On one hand, this approach rate must be great enough to prevent the vehicles from bouncing off each other without capture being achieved. On the other hand, it must be low enough to prevent structural damage or loss of control and/or the ability to attenuate the momentum. The exact position, velocity, orientation, and angular rate tolerances depend on the specific docking or capture tools being used.

The docking/capture phase is the conclusion of a R&D mission. It encompasses the following activities:

- docking/capture of the target by the interceptor (or vice versa)
- establishment of a rigid structural connection
- connection of fluid, gas, electrical, propellant and communication lines
- establishment of a pressurized passageway, if crew transfer is part of the mission goals.

Docking means that the active spacecraft positions itself and establishes the physical connection using its own momentum. In capture and berthing, either the target or interceptor is captured, positioned and connected by a robotic manipulator to a berthing mechanism. Berthing thus allows contact to be made at a near-zero cloture rate, which means a higher level of control for the operator and avoids the process of one vehicle basically flying into the other. It is therefore the generally preferable approach but comes at the cost of requiring a complex, heavy and expensive *Remote Manipulator System* (RMS).

8.6.6 Shuttle-ISS Rendezvous

The Shuttle's close range rendezvous with the ISS is somewhat different, though, because the docking port is on the leading side (+V-bar) of the ISS. Therefore, beginning at S2 the Shuttle flies in two cycloidal steps and within two orbits a homing approach from S2 to S3 such that it goes up just below the ISS (see Fig. 8.45). Note that the launch window was chosen such that daytime (i.e., sunlight) condition is at those parts of the trajectory, including the final approach and docking (see Figs. 8.46 and 8.47), where the ISS needs to be seen from the Shuttle and vice versa. Just before arriving at S3 the commander of the Shuttle takes over manual control for the remainder of the approach and docking. He will stop the Shuttle at S3 some 180 m below the ISS and will maneuver the Shuttle through a 9 min, 360° backflip (a.k.a. *Rendezvous Pitch Maneuver*, RPM) that allows the station crew to take pictures of the Shuttle's heat shield to see whether it was damaged during launch. The Commander then will move the Shuttle at a so-called TORVA maneuver from the +R-bar to the +V-bar in a position about 120 m directly in front of the station in preparation for the final approach to the pressurized mating adapter PMA-2 located at the leading end of the US utility hub *Harmony*.

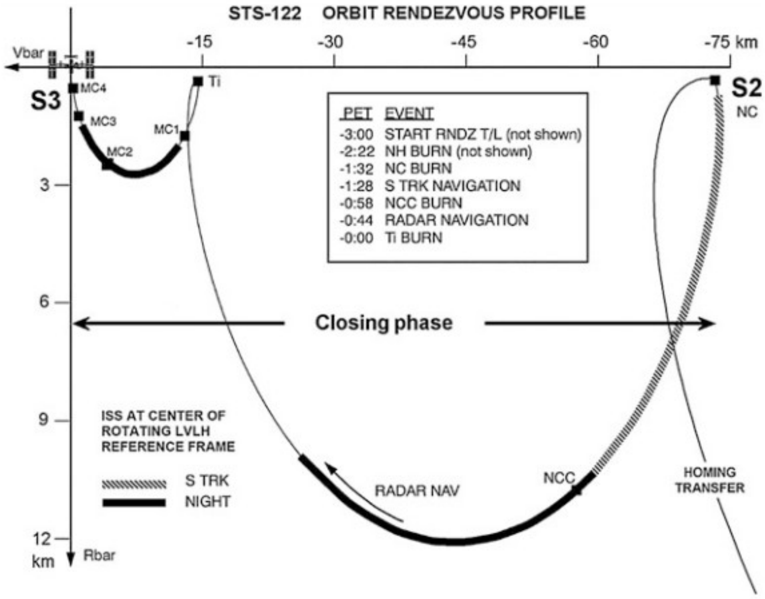
STS-122 ORBIT RENDEZVOUS PROFILE
Vbar
-15
-30
-45
-60
-75 km
S3
MC4
MC3
MC2
MC1
Ti
S2
NC
PET EVENT
-3:00 START RNDZ T/L (not shown)
-2:22 NH BURN (not shown)
-1:32 NC BURN
-1:28 S TRK NAVIGATION
-0:58 NCC BURN
-0:44 RADAR NAVIGATION
-0:00 Ti BURN
3
6
9
12 km
Rbar
Closing phase
ISS AT CENTER OF ROTATING LVLH REFERENCE FRAME
S TRK
NIGHT
RADAR NAV
NCC
HOMING TRANSFER

Fig. 8.45 The cycloidal trajectory in the closing phase of a Shuttle STS-122 rendezvous with the ISS. *Credit* NASA

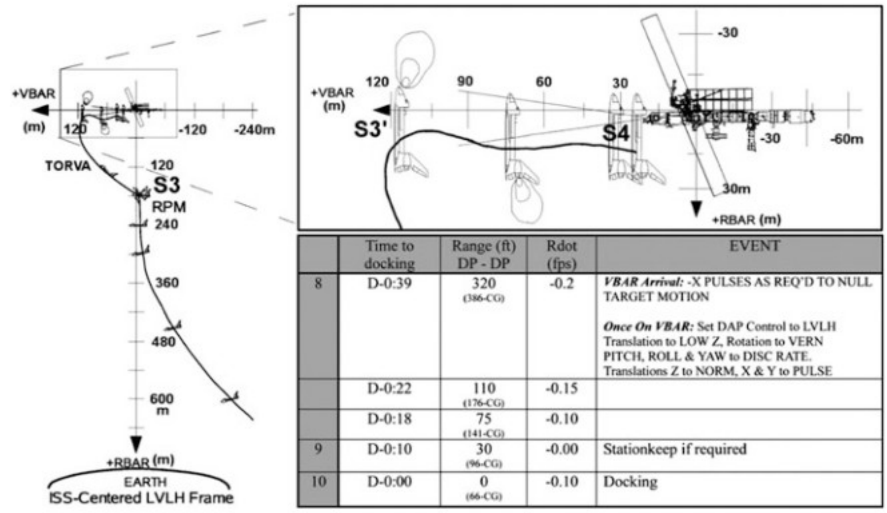
+VBAR (m)
120
-120
-240m
TORVA
120
S3
RPM
240
360
480
600 m
+RBAR (m)
EARTH
ISS-Centered LVLH Frame
+VBAR (m)
120
90
60
30
-30
S3'
S4
-30
-60m
30m
+RBAR (m)

	Time to docking	Range (ft) DP - DP	Rdot (fps)	EVENT
8	D-0:39	320 (386-CG)	-0.2	***VBAR Arrival:*** -X PULSES AS REQ'D TO NULL TARGET MOTION ***Once On VBAR:*** Set DAP Control to LVLH Translation to LOW Z, Rotation to VERN PITCH, ROLL & YAW to DISC RATE. Translations Z to NORM, X & Y to PULSE
	D-0:22	110 (176-CG)	-0.15	
	D-0:18	75 (141-CG)	-0.10	
9	D-0:10	30 (96-CG)	-0.00	Stationkeep if required
10	D-0:00	0 (66-CG)	-0.10	Docking

Fig. 8.46 Final approach trajectory of the Shuttle STS-122 approaching the ISS. *Credit* NASA

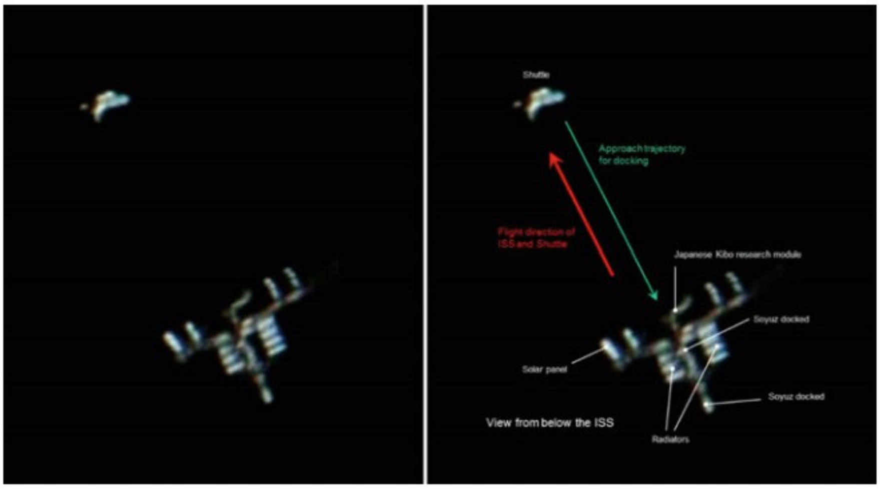
Shuttle
Approach trajectory for docking
Flight direction of ISS and Shuttle
Japanese Kibo research module
Soyuz docked
Solar panel
Soyuz docked
View from below the ISS
Radiators

Fig. 8.47 Final approach trajectory of Shuttle STS-133 approaching ISS (left raw, right annotat ed picture). Picture taken with ISO 800, 1/800 s, through a 8.5″ Newtonian on February 26, 2011, 18:35:54 h GMT, over UK. *Credit* Rob Bullen

Acronyms for NASA Rendezvous Maneuvers

NC	(Phasing correction burn) Performed to hit a range relative to the target at future times
NH	(Height adjust burn) Performed to hit a delta-height relative to the target at future times
NPC	(Plane change burn) Performed to remove planar errors relative to the target at future times
NCC	(Corrective combination burn) First on-board targeted burn in the rendezvous sequence to reduce phasing and height errors relative to the target at Ti
Ti	(Terminal intercept burn) Second on-board targeted burn in the rendezvous sequence to place the orbiter on a trajectory to intercept the target in one orbit
MC-1,2,3,4	(Midcourse correction burns) On-board targeted burns to correct the post Ti trajectory in preparation of the final approach phase
RPM	(Rendezvous Pitch Maneuver) A 360° backflip that allows the station crew to take pictures from the Zvezda Service Module of the Shuttle's heat shield
TORVA	(Twice Orbital Rate R-bar to V-bar Approach) This manually performed maneuver brings the orbiter from the +R-bar to the +V-bar

8.6.7 Plume Impingement

The term *plume impingement* covers all effects exerted on the target object if it is impacted by the exhaust gases of the interceptor's reaction control system (RCS, the maneuvering control system of an interceptor) thrusters. One of these effects is the plume pressure force acting on the target and causing position and attitude disturbances. Another is the heat load placed on the target's structure by the hot gases. This can lead to overheating of parts of the surface and the underlying structure. The third effect is the contamination of the target's surfaces by combustion products and unburned propellant components. This can cause contamination of not only sensitive elements on the target's surfaces, particularly optical elements such as camera lenses, solar arrays or docking sensors, but also of sealing elements of the docking mechanism. This risk of contamination must be considered in particular during orbit servicing missions such as Hubble servicing, where these considerations impacted the design of final approach trajectories.

Therefore, plume impingement is one of mission planners' major concerns during proximity operations, apart from collision avoidance and maneuver precision. It can only be avoided if thruster activity near the target is minimized. This in turn means that the interceptor's relative velocity must diminish below a threshold value as it approaches the target. During Gemini and Apollo, plume impingement never became a significant issue due to the thrust magnitude, the position and canting of the RCS nozzles, as well as the roughly equal sizes of interceptor and target and the absence of large appendages such as solar arrays. This changed during the Skylab missions. During Skylab 2, the Apollo *Command and Service Module* (CSM) was maneuvered within close proximity so that a crewman standing in the hatch could reach the stuck solar array with a deployment tool. The CSM thrusting to null the closing velocity triggered Skylab AOCS to fire its jets in order to maintain its attitude. This resulted in an opening rate between the two vehicles. On the later Apollo-Soyuz Test Mission, four of the CSM's thrusters were inhibited 2 s prior to docking contact in order to prevent plume loading of the Soyuz solar arrays.

Space Shuttle

These lessons were carried into the Space Shuttle design process. The massive orbiter is designed to assemble and maintain large space stations and service comparatively small and light satellites. These are equipped with large solar arrays and antennas or sensitive optics. Plume impingement therefore is a prime concern. The size of the Shuttle was predetermined by the payload it was designed to carry and the location of RCS thrusters by its shape, which in turn was determined by the requirements of re-entry and atmospheric flight. Plume impingement concerns could therefore be addressed only by careful design of R&D approaches.

The underlying assumptions were as follows: The target spacecraft could not be designed with features preventing contamination (e.g., movable sensor covers as

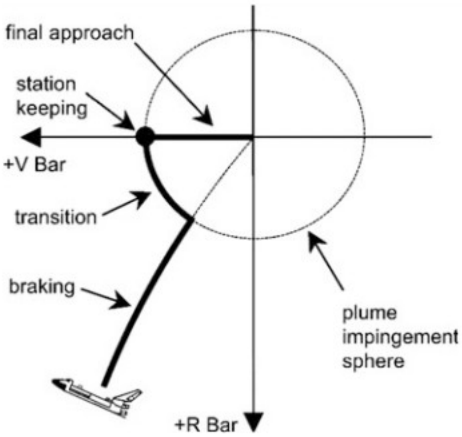

Fig. 8.48 Approach profile to avoid plume impingement on target. When the approaching interceptor reaches the border of a plume impingement sphere, it performs a transition maneuver to a station-keeping point on the V-bar following the borders of the plume impingement sphere. After the station-keeping point is reached, the interceptor flies a forced translation final approach. *Credit* Goodmann (2006)

found on Hubble), and the control of the target attitude could not prevent contamination. Therefore, on each mission a target-dependent minimum range existed, at which the thrusters could still be fired in the direction of the target without contamination concern. At the minimum range, the orbiter was to transition from a direct approach trajectory to a station-keeping point on the V-bar (see Fig. 8.48). From this point on the final approach would be flown in forced translation. A number of such approaches were planned for the Shuttle's *Long Duration Exposure Facility* (LDEF) mission. Simulations showed that an Apollo-type inertial approach and braking technique would cause LDEF to tumble. In addition, plume impingement induced dynamics at grapple ranges that could make both LDEF deployment and retrieval difficult.

Another countermeasure for plume impingement issues specific to the Space Shuttle was the development of the *Low-Z* approach. In this approach, all forward firing RCS jets are inhibited, with all thrust thus acting primarily along the spacecraft's longitudinal x-axis (see Fig. 8.49). All braking thrust in the z-direction therefore results from the canting of the longitudinal thrusters. This provides minimal RCS braking capability while minimizing RCS plume impingement. It is also expensive in terms of propellant use. Notwithstanding its limitations, *Low-Z* mode has been employed on satellite servicing missions, including Hubble servicing, and the missions to MIR and ISS.

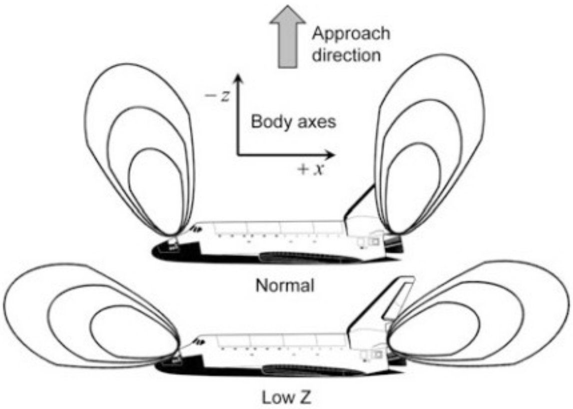

Fig. 8.49 Space shuttle low-Z RCS mode. The orbiter's thrusters are fired only in the Shuttle's longitudinal axis (x-axis). This significantly reduces plume impingement. On the other hand, owing to the canting of the thrusters, this provides minimal braking capability in the forward z-direction

8.7 Problems

Problem 8.1 *Efficiency of Combined Orbital Maneuvers*
Prove the statement in Sect. 8.1.3 that in terms of delta-v an orbit transition with $\boldsymbol{v}_1 || \boldsymbol{v}_2$ combined with a change in orbit orientation $v_1 = v_2,\ \phi \neq 0$ is always more efficient than performing the two changes sequentially at the same point in space.

Problem 8.2 *Adjacent Circular Orbit Approximation*
Prove Eq. (8.3.9)

$$\frac{\sqrt{a_\bullet} + \sqrt{a_\mathrm{O}}}{\sqrt{a_H}} - 1 \ \approx 1 - \frac{1}{16}\left(\frac{a_\mathrm{O} - a_\bullet}{a_\bullet}\right)^2 \quad @\, a_\mathrm{O} \to a_\bullet$$

Problem 8.3 *Transfer between Aligned Ellipses*
Consider a Hohmann transfer between two coplanar and coaxial ellipses. Show that the propulsion demand for the transition between the periapsis of the inner ellipse and the apoapsis of the outer ellipse is

$$\begin{aligned}\Delta v &= \sqrt{\mu}\left(\frac{1}{\sqrt{r_{\bullet,per}}} - \frac{1}{\sqrt{r_{\mathrm{O},apo}}}\right)\left(\frac{\sqrt{r_{\mathrm{O},apo}} + \sqrt{r_{\bullet,per}}}{\sqrt{a_T}} - \frac{\sqrt{r_{\mathrm{O},apo}}\sqrt{1+e_\bullet} - \sqrt{r_{\bullet,per}}\sqrt{1-e_\mathrm{O}}}{\sqrt{r_{\mathrm{O},apo}} - \sqrt{r_{\bullet,per}}}\right)\\ &\leq \sqrt{\mu}\left(\frac{1}{\sqrt{r_{\bullet,per}}} - \frac{1}{\sqrt{r_{\mathrm{O},apo}}}\right)\end{aligned}$$

Remark *For circular orbits this expression passes over to Eq. (8.3.8).*

Problem 8.4 *Hohmann Transfer Maxima*

(a) Prove that the maximum of the circularization impulse $\Delta v_{HO \to O}$ (see Fig. 8.12) is the root of the equation $x^3 - 5x^2 - 5x - 1 = 0$. Find the root $x = r_O / r_\bullet = 5.879362\ldots$ by Newton's method.

(b) By the same token, prove that the propulsion demand of a Hohmann transfer achieves a maximum at $x = a_O / a_\bullet = 15.58172\ldots$, which is the root of $x^3 - 15x^2 - 9x - 1 = 0$.

Problem 8.5 *Hohmann Versus Bi-Elliptic Transfer*

Prove that for $x = r_O / r_\bullet > 11.9387654724\ldots$ the bi-elliptic transfer has a lower propulsion demand than the Hohmann transfer. Show that $x = r_O / r_\bullet$ is the root of the equation $x - 1 = \sqrt{1+x}(\sqrt{x} - \sqrt{2} + 1)$.

Problem 8.6 *Variations of Orbital Elements by Kick-Burns*

(a) Prove Eq. (8.1.1) as described in the text.

(b) Prove Table 8.1 by applying Eq. (8.1.1) at the given positions.

Problem 8.7 *Orbit Phasing*

Suppose two satellites are flying in a close formation on the same orbit at relative distance s and orbital period T. Show that, if s needs to be corrected, a kick-burn $\delta v_{||}$ at the periapsis will cause a position shift of

$$\delta s_{per} = 3T \frac{1+e}{1-e} \delta v_{||}$$

at the periapsis after one orbit, while a kick-burn $\delta v_{||}$ at the apoapsis will cause a position shift of

$$\delta s_{apo} = 3T \frac{1-e}{1+e} \delta v_{||}$$

at the apoapsis after one orbit.

Problem 8.8 *Maximum Transfer Time for Minimum Energy Transfer*

Given the minimum energy transfer orbit between two points P_1 and P_2 having $4a_{\min} = r_1 + r_2 + c$. Show that for any slightly larger transfer orbit with $a = a_{\min} + \Delta a$ the transfer time is

$$\Delta t \approx \Delta t_{\max} - 4a_{\min} \sqrt{\frac{\Delta a}{\mu}} \quad @\, \beta < \pi$$

$$\Delta t \approx \Delta t_{\max} - a_{\min} \sqrt{\frac{\Delta a}{\mu}} \quad @\, \beta \approx \pi$$

and therefore the minimum energy transfer orbit indeed has a maximum transfer time. We will explore the characteristic square root behavior $\delta(\Delta t) \propto -\sqrt{\Delta a}$ near the maximum transfer time for a Hohmann transfer a bit further in Sect. 9.3.2 before Eq. (9.3.15).

Problem 8.9 *Lambert's Problem for $a \to \infty$*

Prove that from Lambert's Eq. (8.2.13) for $a \to \infty$ follows

$$a = \frac{3}{40} \cdot \frac{\Delta_+^{5/2} - \Delta_-^{5/2}}{6\Delta t\sqrt{\mu} - \left(\Delta_+^{3/2} - \Delta_-^{3/2}\right)}$$

with $\Delta_\pm = r_1 + r_2 \pm \sqrt{r_1^2 + r_2^2 - 2r_1 r_2 \cos \Delta\theta}$. Since $\sqrt{\Delta_\pm} = 2s\sqrt{c}$ this delivers Eq. (8.2.15).

Problem 8.10 *Solution of the Differential Equations for Elliptic Orbits*
It is the objective of this exercise to solve the differential Eq. (8.5.17). According to the theory of differential equations the general solution is the sum of the solution of the homogeneous differential equations given by Eq. (8.5.8) and a special solution to the inhomogeneous differential equations. We therefore just have to seek for this special solution.

(a) Before doing so we make use of the fact that the right hand side of the differential solution is small, namely of order e. We therefore can insert the homogeneous solutions on the right hand side. Show that the differential equations then read

$$x'' + 2z' = e \begin{pmatrix} \cos M \\ -6M\cos M + 8\sin M - 6\sin 2M \\ -3M\cos M + 4\sin M - 4\sin 2M \\ -2\cos M - 2\cos 2M \end{pmatrix}^T \begin{pmatrix} x_0 \\ z_0 \\ x_0' \\ z_0' \end{pmatrix}$$

$$y'' + y = -3e \begin{pmatrix} \cos^2 M \\ \sin M \cos M \end{pmatrix}^T \begin{pmatrix} y_0 \\ y_0' \end{pmatrix}$$

$$z'' - 3z - 2x' = 2e \begin{pmatrix} -\sin M \\ 6M\sin M + 8\cos M + \frac{3}{2}\cos 2M - \frac{9}{2} \\ 3M\sin M + 4\cos M + \cos 2M - 3 \\ +2\sin M - \frac{1}{2}\sin 2M \end{pmatrix}^T \begin{pmatrix} x_0 \\ z_0 \\ x_0' \\ z_0' \end{pmatrix}$$

(b) Define the second equation as

$$y'' + y = f(M)$$

The homogeneous differential equation is known to have the fundamental system of solutions $y_{1,2}(M) = (\sin M, \cos M)$. Now, according to the theory of differential equations a special solution is given as

$$y^*(M) = -\sin M \int_{M_0}^{M} \frac{f(\mu)\cos\mu}{W(\mu)} \cdot d\mu + \cos M \int_{M_0}^{M} \frac{f(\mu)\sin\mu}{W(\mu)} \cdot d\mu$$

where $W(M)$ is the Wronskian of the fundamental system of solutions. Show that this delivers

$$y^*(M) = -e \begin{pmatrix} \cos^2 M + \cos M - 2 \\ \sin M \cos M - \sin M \end{pmatrix}^T \begin{pmatrix} y_0 \\ y'_0 \end{pmatrix}$$

(c) By the same token as in Sect. 8.5.2 derive the differential equation $z'' + z = g(M)$. Show that

$$u'' + u = e \begin{pmatrix} 0 \\ -12\cos M + 9\cos 2M \\ -6\cos M + 6\cos 2M \\ -3\sin 2M \end{pmatrix}^T \begin{pmatrix} x_0 \\ z_0 \\ x'_0 \\ z'_0 \end{pmatrix} =: g(M)$$

with $u = z - (4 + 13e)z_0 - (2 + 4e)x'_0$. Then apply the above integral expression for the special solution $z^*(M)$. Finally integrate the differential equation $x' + 2z = h(M)$ directly to derive $x(M)$.

Invariant Angular Elements

The classical way of doing this is using the geocentric equatorial coordinate system *IJK* (see Sect. 13.1.4), in which the Cartesian coordinate system $(\boldsymbol{u}_I,\ \boldsymbol{u}_J,\ \boldsymbol{u}_K)$ has the following orientation (see Fig. 13.2): The unit vector $\boldsymbol{u}_I$ points to the vernal point, $\boldsymbol{u}_K$ to Earth's north pole, and $\boldsymbol{u}_J = \boldsymbol{u}_K \times \boldsymbol{u}_I$. With this we establish the unit vector

$$\boldsymbol{u}_n = \frac{\boldsymbol{u}_K \times \hat{\boldsymbol{h}}}{\left|\boldsymbol{u}_K \times \hat{\boldsymbol{h}}\right|} \quad \text{unit vector to the ascending node}$$

These unit vectors determine the following angular orbital elements (see Fig. 7.10)

$$i = \arccos\left(\boldsymbol{u}_K \hat{\boldsymbol{h}}\right) \qquad 0 \le i < 180^\circ \tag{7.3.21}$$

$$\Omega = \begin{cases} \arccos(\boldsymbol{u}_I \boldsymbol{u}_n) & @ \ \boldsymbol{u}_J \boldsymbol{u}_n \ge 0 \\ 2\pi - \arccos(\boldsymbol{u}_I \boldsymbol{u}_n) & @ \ \boldsymbol{u}_J \boldsymbol{u}_n < 0 \end{cases} \qquad 0 \le \Omega < 360^\circ \tag{7.3.22}$$

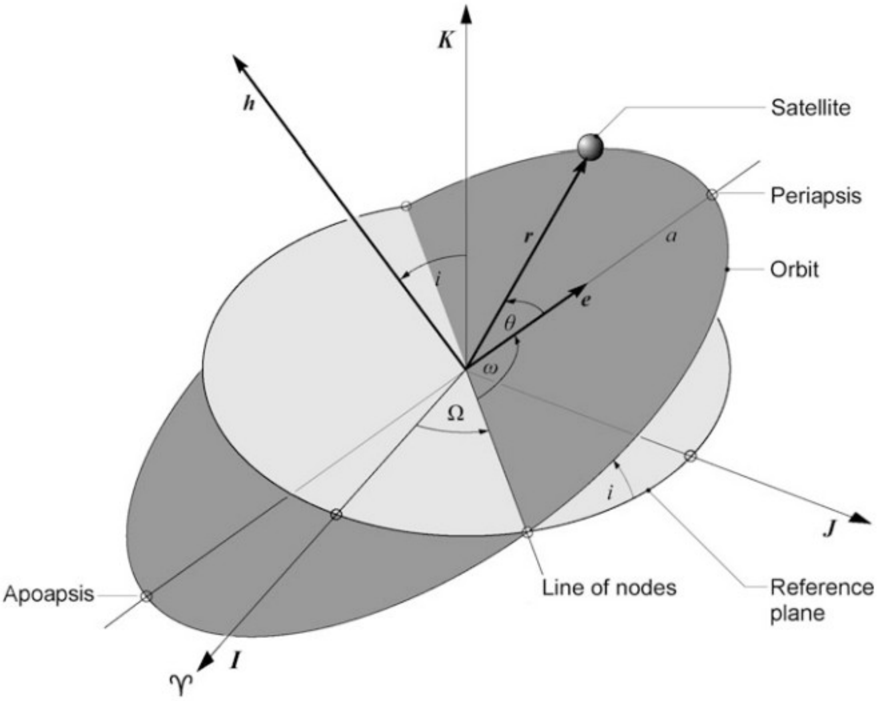

Fig. 7.10 The graphical representation of orbital elements of an elliptic satellite orbit in a geocentric equatorial coordinate system *IJK* (see Sect. 13.1.3), where the plane spanned by *I* and *J* with the *I*-axis oriented toward the vernal point ♈ is the reference plane. The orbital plane and reference plane intersect in the so-called line of nodes. The line of nodes in turn intersects the orbit in the so-called ascending node (where the satellite moves toward the upper side with respect to ***h*** of the reference plane) and in the descending node (vice versa)

$$\omega = \begin{cases} \arccos(\boldsymbol{u}_n \hat{\boldsymbol{e}}) & @ \ \boldsymbol{u}_K \hat{\boldsymbol{e}} \geq 0 \\ 2\pi - \arccos(\boldsymbol{u}_n \hat{\boldsymbol{e}}) & @ \ \boldsymbol{u}_K \hat{\boldsymbol{e}} < 0 \end{cases} \quad 0 \leq \omega < 360° \tag{7.3.23}$$

The following classification of orbits according to their inclination is common:

$i = 0°$	*Equatorial Orbit*. Prominent examples are the geostationary orbits.
$0° < i < 90°$	*Prograde Orbit*. This orbit is used by most satellites in Low Earth Orbits.
$i \approx 90°$	*Polar Orbit*. Orbits used for sun-synchronous Earth observation satellites.
$i > 90°$	*Retrograde Orbit*. Very few satellites have been put into this type of orbit.

Invariant Orbit Vectors

In terms of these angular elements, the eccentricity vector can be represented in the *IJK* system as follows (see also Sect. 13.1.5)

$$\boldsymbol{e} = e \begin{pmatrix} \cos\Omega\,\cos\omega - \sin\Omega\,\sin\omega\,\cos i \\ \sin\Omega\,\cos\omega + \cos\Omega\,\sin\omega\,\cos i \\ \sin\omega\,\sin i \end{pmatrix}_{IJK} \quad \textbf{eccentricity vector} \tag{7.3.24}$$

Equivalently, the angular momentum unity vector $\hat{\boldsymbol{h}}$, also known as the *three-dimensional inclination vector* (a.k.a. *orbit pole vector*), given as (see Eq. (13.1.5)) as

$$\boldsymbol{I} \equiv \hat{\boldsymbol{h}} = \begin{pmatrix} \sin\Omega\,\sin i \\ -\cos\Omega\,\sin i \\ \cos i \end{pmatrix}_{IJK} \quad \textbf{three-dimensional inclination vector} \tag{7.3.25}$$

For instance, in Fig. 12.28 this 3D-inclination vector depicts the evolution of i, Ω under lunisolar perturbations. Its projection onto the *IJ* reference plane is known as

$$\boldsymbol{i} = \sin i \begin{pmatrix} \sin\Omega \\ -\cos\Omega \end{pmatrix}_{IJ} \quad \text{(two-dimensional) } \textbf{inclination vector} \tag{7.3.26}$$

Note that for $i \to 0$ conveniently

$$\boldsymbol{i} = i \begin{pmatrix} \sin\Omega \\ -\cos\Omega \end{pmatrix}_{IJ} \quad @ \ i \to 0$$

We now rotate the inclination vector in the *IJ* reference plane by 90° counter-clockwise into the direction of the ascending node and thus get the *ascending node vector*

$$\boldsymbol{n} = \sin i \begin{pmatrix} \cos\Omega \\ \sin\Omega \end{pmatrix}_{IJ} = \sin i \begin{pmatrix} \cos\Omega \\ \sin\Omega \end{pmatrix}_{EQ} \quad \textbf{ascending node vector} \qquad (7.3.27)$$

All the above invariant quantities are orbital elements. This gives rise to the question of whether there is a minimum number of orbital elements that uniquely specify an orbit and how many of them are constant.

State Vector

We investigate the above question by going back to the basics of orbital motion. Depending on the initial conditions, orbital motion takes place in 6-dimensional phase space $\boldsymbol{r}$, $\boldsymbol{v}$, a space in which all possible states of a system are represented, with each possible state corresponding to one unique point in the phase space. Once an initial state $(\boldsymbol{r}_0, \boldsymbol{v}_0)$ is given, the two first-order differential vector equations

$$\dot{\boldsymbol{r}} = \boldsymbol{v}, \quad \dot{\boldsymbol{v}} = -\frac{\mu}{r^3}\boldsymbol{r}$$

derived from Newton's gravitation equation of motion (7.1.19) determine the future motion in phase space. Obviously, phase space has 6 dimensions and therefore the 6 elements of an initial state vector are the minimum number of orbital elements plus the independent parameter *t*. This is the significance of the state vector $(\boldsymbol{r}, \boldsymbol{v})$ as introduced in Sect. 7.2.2. It also has the advantage that it is the direct outcome of ranging and range rate measurements as part of orbit determination procedures (see Sect. 14.2.1). On the other hand it has the major drawback that $\boldsymbol{r}$ and $\boldsymbol{v}$ continuously change with time.

Number of Constant Orbital Elements

So, what is the maximum number of constant orbital elements? We have seen in Sect. 7.1.3 that linear momentum, angular momentum, and total energy always need to be conserved. Because the motion of a 2-body system in a central force field, such as the gravitational field and as shown in Sect. 7.2.2, always takes place in just 2 dimensions, we expect that any orbital motion exhibits $2 + 2 + 1 = 5$ constant (integrals) of motion and hence 5 constant orbital elements. This leaves one degree of freedom to the orbital motion, which is the 1-dimensional line of orbit. We therefore expect one additional orbital element to determine the moving position of the body on its orbit. Hence, in total there must be 5 constant orbital elements and 1 time-dependent orbital element. Obviously $\boldsymbol{h}$, $\boldsymbol{e}$ and $\theta(t)$ lend themselves as such orbital elements. Although they seem to make up $3 + 3 + 1 = 7$ elements, one has to consider the constraint $\boldsymbol{h} \cdot \boldsymbol{e} = 0$, and therefore the set $(\boldsymbol{h}, \boldsymbol{e}, \theta(t))$ actually represents the set of 6 independent elements and hence a minimum complete set of orbital elements.

We have seen that we may choose many different sets of 5 constant orbital elements, and any set is equivalent to any other. Next we to explore some common sets of orbital elements.

Common Sets of Orbital Elements

The most vivid orbital elements are the angular orbital elements i, Ω, ω (see Fig. 7.10) as given in Eq. (7.3.21), (7.3.22), and (7.3.23). In addition and from Eq. (7.3.7) we have from $\boldsymbol{h}$, $\boldsymbol{e}$

$$e = |\boldsymbol{e}|$$

$$a = \frac{\boldsymbol{h}^2}{\mu(1-e^2)}$$

We thus obtain the highly important and so-called six *classical orbital elements* (see Fig. 7.10)

Classical Orbital Elements $\left(a, e, i, \Omega, \omega \,/\, t - t_p \leftrightarrow M\right)$

Two elements, so-called *metric elements*, describe the dimension of the orbit:

a **Semi-major axis** axis, $-\infty < a < \infty$.
It specifies the size of the orbit. Hyperbolic orbits have negative values.

e **Eccentricity**, $0 \le e < \infty$.
It specifies the shape and jointly with a the type of the orbit.

The next three elements are called *angular elements*. Two elements describe the orientation of the orbital plane:

i **Inclination**, $0 \le i < 180°$
Angle between the angular momentum vector h and the z-direction of the reference frame = angle between the corresponding planes perpendicular to these vectors.

Ω **Right ascension of ascending node** (abbreviated: RAAN), $0 \le \Omega < 360°$
Longitude of the ascending node = angle between the line from the origin O of the reference frame to the vernal point and from O to the ascending node.

One element *determines* the orientation of the orbit in the orbital plane:

$\boldsymbol{\omega}$ **Argument of periapsis**, $0 \le \omega < 360°$
Angle in the orbital plane between line of nodes and the periapsis measured in the direction of the motion.

The sixth parameter is the orbital state sequencer, it determines the location on the orbit. The following two sequencers can be used alternatively:

$\boldsymbol{t - t_p}$ **Time after periapsis passage**
$\boldsymbol{t_p}$ is the time, when the satellite was at periapsis (a.k.a. *epoch*).

or

M Mean anomaly, $M > 0$
$M := n \cdot (t - t_p)$, dimensionless (see Sect. 7.3.2).

Although M formally is an angle (however not having an geometrical interpretation) with dimension [rad], it commonly is treated as dimensionless and hence in modern orbital mechanics is a more convenient variant of time. Both sequencers have the disadvantage that their relationship to r and θ (see Sect. 7.4.2) is transcendental.

There is an alternative and more traditional set of orbital elements, called *Keplerian elements*

Keplerian Elements $(a, e, i, \Omega, \omega \,/\, \theta \leftrightarrow E,\ F, G)$
The Keplerian elements share the metric and angular classical orbital elements. The following two sequencers can be used alternatively:

θ True anomaly (a.k.a. *orbit angle*), $0 \le \theta < 360°$
Angle between the direction to the periapsis and the current position vector $\boldsymbol{r}$.

or

E Eccentric anomaly for elliptic orbits (see Sect. 7.4.2), $0 \le E < 2\pi$
F Hyperbolic anomaly for hyperbolic orbits (see Sect. 7.4.3), $0 \le F < \infty$
G Universal anomaly for all orbits (see Sect. 7.4.6), $0 \le G < \infty$

The advantage of this set is that the determination of $r(\theta)$, or equivalently $r(E, F, G)$, is algebraic and hence straightforward. The disadvantage is that θ is not a convenient orbital state sequencer, but time t. The inclusion of θ, which Leonhard Euler called "angle of elements", into the set of elements dates back to Johann Bernoulli.

Sometimes the compound angle

$$u = \omega + \theta \qquad \textbf{argument of latitude}$$

i.e., the position of the orbit measured relative to the line of nodes, is a more valuable parameter for orbit analysis. Note that, though $\boldsymbol{h},\ \boldsymbol{e},\ \theta(t)$ usually are defined in the geocentric equatorial coordinate system, of course any inertial reference system will do and only the angular elements will change by a different choice of reference system.

Degenerate Orbits

There exist two particular cases where some of the angular elements become undefined:

1. ω and θ (or M) will become undefined for $e = 0$.
2. Ω and ω will become undefined for $i = 0$.

Both happens, for example, for satellites in geostationary orbit. These degenerate cases, however, pose no problem.

In the first case, $e = 0$, where ω becomes irrelevant the argument of latitude $u = \theta$ is used as a substitute to determine the orbital position, i.e. the true anomaly is measured relative the line of nodes.

In the second case, the compound angles

$\varpi = \Omega + \omega$	**longitude of periapsis**	$@\ i = 0$
$L = \Omega + \omega + \theta = \varpi + \theta = \Omega + u$	**orbital longitude**	$@\ i, e = 0$
$l = \Omega + \omega + M = \varpi + M$	**mean longitude**	$@\ i, e = 0$

measured relative to the vernal point Υ are used as substitutes for ω and θ or M (Ω is irrelevant for $i = 0$). The line of nodes and the vernal point then become the respective common reference directions in the orbital plane. Note that Ω is measured in the reference plane, while ω and θ are measured in the orbital plane. So, for $i > 0$, the compound angles ϖ, L, and l would be "doglegged" and no longer be true angles, but nevertheless they can be defined as mathematical quantities and treated like that. These doglegged compound angles are frequently used in the literature, but will not be used in this book.

Equinoctial Elements and Coordinate System

One problem of orbit degeneracy with angular elements arises because Ω and ω are measured relative to the ascending line of nodes, which vanishes for $i = 0$. This problem can be removed by rotating the *IJ* reference plane along the line of nodes by the inclination angle into the orbital plane and renaming the axes $K \rightarrow W,\ I \rightarrow E,\ J \rightarrow Q$. This reference frame is called *equinoctial coordinate system EQW* (see Fig. 7.11). By construction, for $i \rightarrow 0$ it crosses over into the *IJK*

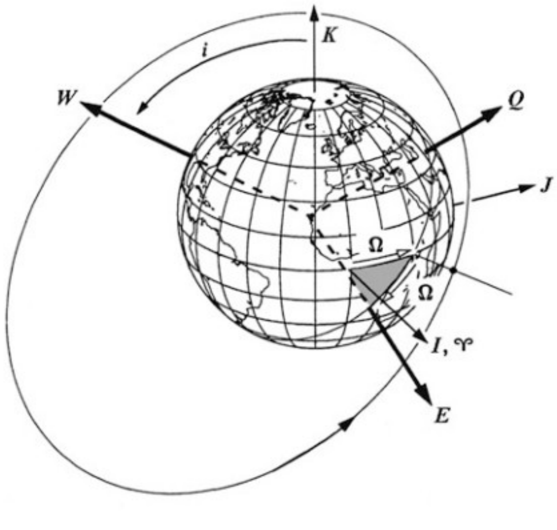

Fig. 7.11 The equinoctial coordinate system EQW. This system is formed from the IJK system by rotating IJK by the inclination angle about the line of nodes. *Credit* Vallado (2007)

coordinate system. The other orbit degeneracy problem, namely that ω and θ (or M) will become undefined for $e = 0$, can be removed by making the eccentricity vector part of the orbital elements. Both strategies are adopted to form a nonsingular set of orbital elements, the so-called **standard equinoctial orbital elements**, which read as follows:

(Standard) Equinoctial Elements $(a,\ h,\ k,\ p,\ q,\ l)$

a Semi-major axis, $-\infty < a < \infty$

Two components of the eccentricity vector:

h $= e\ \sin(\omega + \Omega)$
k $= e\ \cos(\omega + \Omega)$

Two components of the rescaled ascending node vector:

p $= \tan\frac{i}{2} \sin\Omega$
q $= \tan\frac{i}{2} \cos\Omega$

The sixth parameter again determines the location of the periapsis in time:

l $= \Omega + \omega + M(t)$, mean longitude, $0 \leq l < 360°$

Note that the RAAN angle Ω now lies in the orbital plane and therefore the mean longitude $l = \Omega + \omega + M$ is no longer doglegged, but a true angle in the orbital plane. Since the equinoctial elements exhibit no singularities they are frequently employed in perturbational equations for equatorial orbits. For more details of equinoctial elements see Battin (1987), and for coordinate transformations to and from the equinoctial coordinate system see Vallado (2007).

In terms of the invariant orbit vectors we can identify

$$\boldsymbol{e} = \begin{pmatrix} h \\ k \end{pmatrix}_{EQ} = e \begin{pmatrix} \sin(\omega + \Omega) \\ \cos(\omega + \Omega) \end{pmatrix}_{EQ} \tag{7.3.28}$$

and the vector (q, p) actually is a rescaled ascending node vector

$$(q, p) = \frac{1}{1 + \cos i} \boldsymbol{n} = \tan\frac{i}{2} \cdot \begin{pmatrix} \cos\Omega \\ \sin\Omega \end{pmatrix}_{EQ}$$

The reason for scaling by $1/(1 + \cos i)$ lies in the better adaptation to Lagrange's planetary equations where the equinoctial elements are frequently used for GEO orbits.

The introduction of the singular free parameters (h, k) and (q, p) is most probably due to Joseph-Louis Lagrange. Today, there are some variants of the standard equinoctial orbital elements, most importantly the so-called **modified equinoctial orbital elements** (Walker et al. 1985) with the only modification $a \rightarrow p = a(1 - e^2)$. But also the replacements $\tan\frac{i}{2} \rightarrow \tan i$, $\sin i$, or $\sin\frac{i}{2}$ are quite common.

Conversion: State Vector → Orbital Elements

A frequent problem encountered in astrodynamics, in particular with orbit determination, is the conversion of the state vector $(\boldsymbol{r},\ \boldsymbol{v})$ into (osculating) Keplerian elements and vice versa. In the following, we will have a look at this problem.

First, we want to convert a state vector $(\boldsymbol{r},\ \boldsymbol{v})$ to Keplerian elements. To start with, we compute

$$\boldsymbol{h} = \boldsymbol{r} \times \boldsymbol{v}$$

$$\hat{\boldsymbol{h}} = \boldsymbol{h}/h, \quad \hat{\boldsymbol{r}} = \boldsymbol{r}/r, \quad \hat{\boldsymbol{v}} = \boldsymbol{v}/v$$

The semi-major axis then is determined from the vis-viva Eq. (7.2.14) as

$$a = \frac{\mu r}{2\mu - r v^2} \tag{7.3.29}$$

Observe that the semi-major axis and thus the orbital period $T = 2\pi\sqrt{a^3/\mu}$ can be determined merely from the absolute values of the position vector and orbital velocity. For the eccentricity we have from Eq. (7.3.3) and with Eq. (7.3.6)

$$e = |\boldsymbol{e}| = \sqrt{1 - \frac{h^2}{\mu a}}$$

$$\hat{\boldsymbol{e}} = \boldsymbol{e}/e$$

$$\theta = \begin{cases} \arccos(\hat{\boldsymbol{r}}\hat{\boldsymbol{e}})@ \ \ \hat{\boldsymbol{r}}\hat{\boldsymbol{v}} \geq 0 \\ 2\pi - \arccos(\hat{\boldsymbol{r}}\hat{\boldsymbol{e}})@ \ \ \hat{\boldsymbol{r}}\hat{\boldsymbol{v}} < 0 \end{cases} \tag{7.3.30}$$

The angular orbital elements i, Ω, ω are given in the geocentric equatorial reference system *IJK* by Eqs. (7.3.21), (7.3.22), and (7.3.23). To determine the mean anomaly M, we derive from Eq. (7.4.13a)

$$\begin{array}{ll} E = \arccos[(1 - r/a)/e] & @ \ \ |\hat{\boldsymbol{r}} \times \hat{\boldsymbol{e}}| > 0 \\ E = 2\pi - \arccos[(1 - r/a)/e] & @ \ \ |\hat{\boldsymbol{r}} \times \hat{\boldsymbol{e}}| < 0 \end{array} \tag{7.3.31}$$

Hence from the Kepler's Eq. (7.4.15) we derive

$$M = E - e\ \sin E$$

and from Eq. (7.4.14d) we have alternatively

$$\theta = 2\arctan\left(\sqrt{\frac{1+e}{1-e}}\tan\frac{E}{2}\right)$$

Conversion: Orbital Elements → State Vector
Let the orbital elements $a, e, i, \omega, \Omega, \theta$ or $a, e, i, \omega, \Omega, M$ be given. The orbit radius then is determined as

$$r = \frac{a(1-e^2)}{1+e\,\cos\theta} = a(1-e\,\cos E)$$

where E is the solution of $M = E - e\,\sin E$ (see Newton's method in Sect. 7.4.2). As shown in Sect. 13.1.5, the radial vector and the orbital velocity vector in the *IJK* system (see Sect. 13.1.3) are then determined as

$$\boldsymbol{r} = r\begin{pmatrix} \cos\Omega\,\cos u - \sin\Omega\,\sin u\,\cos i \\ \sin\Omega\,\cos u + \cos\Omega\,\sin u\,\cos i \\ \sin u\,\sin i \end{pmatrix}_{IJK} \tag{7.3.32}$$

$$\boldsymbol{v} = \sqrt{\frac{\mu}{a(1-e^2)}}\begin{pmatrix} -\cos\Omega(\sin u + e\,\sin\omega) - \sin\Omega(\cos u + e\,\cos\omega)\cos i \\ -\sin\Omega(\sin u + e\,\sin\omega) + \cos\Omega(\cos u + e\,\cos\omega)\cos i \\ (\cos u + e\,\cos\omega)\,\sin i \end{pmatrix}_{IJK} \tag{7.3.33}$$

with $u = \omega + \theta$ the argument of latitude. For equatorial orbits Ω is undefined and can be set to zero. The same holds for ω in circular orbits, and hence $u = \theta$.

7.4 Keplerian Orbits

In this chapter we study the detailed properties of the different Keplerian orbits (a.k.a. *conic orbits*) and derive and present their basic results. This includes, in particular, the analytic solutions to $t = t(\theta)$ and the numeric algorithms to calculate $\theta = \theta(t)$ for elliptic and hyperbolic orbits.

7.4.1 Circular Orbit

The most common orbits of spacecraft around celestial bodies are circular or near-circular in shape, since they take on the highest minimum orbital altitude (implies least atmospheric drag) at a given orbital energy (cf. Fig. 7.7) and provide steady orbit conditions. According to Eq. (7.3.18) they also feature maximum angular momentum at a given orbital energy. For $e = 0$, Eq. (7.3.9) can be solved immediately and according to Eq. (7.3.5) we get

$$\begin{aligned} r &= a = p = const \\ \theta &= \sqrt{\frac{\mu}{a^3}}\cdot\left(t - t_p\right) = n\left(t - t_p\right) \end{aligned} \tag{7.4.1}$$

(%

- Astronautics the Physics of Space Flight...pdf

obs: Ver livro todo!!!

Pág. 292(319) 8.1.4

Tangent Plane Maneuvers

I is the three-dimensional inclination vector

(see Equation 7.3.25)

Pág. 200(228) - Invariant Angular Elements

Fig 7.10 - the graphical representation of orbital elements of an elliptic satellite orbit in a geocentric equatorial coordinate system IJK

(see Sect 13.1.3)

↳ Pág. 666(691)

the Equatorial Coordinate System (IJK).

Pág. 201(229)

the angular momentum unity vector h, also known as the three-dimensional inclination vector (a.k.a, orbit pole vector), give as (see Eq. 13.1.5)

↳ Pág. 670(695) —

3 +,-.+ + +/ ! + * L! - =! , 3 ! ?
H2 ! ! # , ! ? ! H,- 3 ! ' 3#

3 +,-.+ + +/ ! + * L! - =! , 3 ! ?
H2 ! ! # , ! ? 3! & H,- 3 ! 3#

B +,-.+ + +/ ! + * <! = 2 - ! ! !
0 2 - = ? < # #78
+,-.+ + +/ ! + * <! = 2 - ! ! ! 0 2
- = ? < # 3#

+,-.+ + +/ ! + * = ? H2 ! 3 ,
> ! 2 ,33 ! , ! !4 2 ! # < D ! 3#

+,-.+ + +/ ! + * , I! # < = u - =
3! >! ! # ?) 3#

? $ $ % ") *
|I & * (%% !# * # |I ! B 8
!) #

! " # !# $ $ % % & % ' !
((# !# % #
!)*+,)))- !) % . (/0122 3#
-4 2 5 ! # * ! 5
' (! . ! # * 4 6 !

_ Airplane Flight Dynamics...(2003).pdf

_ Mathematical Theory of Rocket Flight ...(2008).pdf

= Flight Dynamics Principles...(2007).pdf

= Non-Rocket Space Launch and Flight...(2006).pdf

20210921///

Books:

Books-019-2021/20210921/

_ Geometry Vector Analysis/
B095Z3GG8M.pdf

Geometry and Vector Analysis

Pág. G-4 (08) →

3 Relation between Cartesian and Polar Coordinates

4. Distance between Two Points

Pág. 09 - 5. Polar Equation of a Straight Line.

Pág. 10 7. Polar Equation of a

Straight Line through Two Points.

Pág. 84 3. General Equation of a Plane

Pág. 115 The Straight Line

Obs: Ver livro todo!!!

- 20210921| Differential Calculus...|
B095Z6555B.pdf
→ Differential Calculus
Pág. 187 Curvature.

- 20210921| Complex Analysis...|
B095Z551R2.pdf
⇒ Complex Analysis.

- 20210921| Several Variables and Partial Differential Equations|
B095YBYMDH.pdf

- 20210921| Linear Algebra...|
B095Y9BGCB.pdf
⇒ Linear Algebra

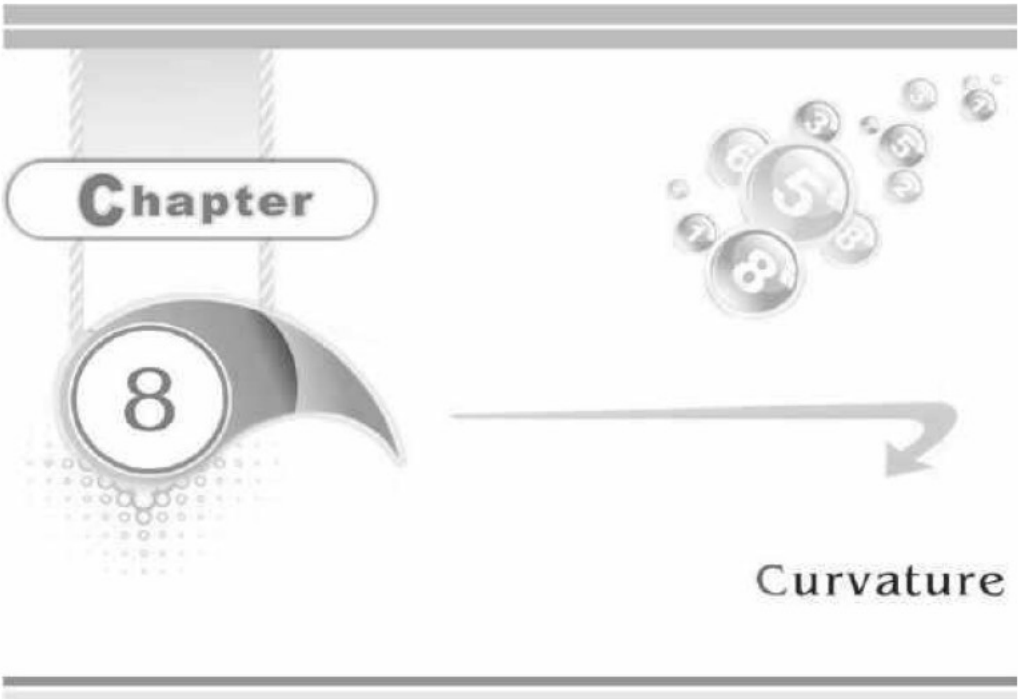

Chapter 8

Curvature

1 Meaning of Curvature

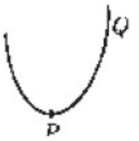

In the adjoining figure we see that the curve bends more sharply at the point P than at the point Q. We express this feeling by saying that the curve has a greater curvature at P than at Q. However, in order to get a quantitative estimate of curvature, we should give a mathematical definition of curvature which should be in agreement with our intuitive notion of curvature.

2 Definition of Curvature **(Purvanchal 2010)**

Let P and Q be two neighbouring points on a curve, ψ and $\psi + \delta\psi$ the angles which the tangents at P and Q make with the x-axis.

Let A be any fixed point on the curve.

Let arc $AP = s$, arc $AQ = s + \delta s$, so that arc $PQ = \delta s$.

The symbol, $\delta\psi$ denotes the angle through which the tangent turns as a point moves along the curve from P to Q through a distance δs. The angle $\delta\psi$ is called the *contingence* of the arc PQ. Obviously, $\delta\psi$ will be large or small, as compared with δs, depending on the degree of sharpness of

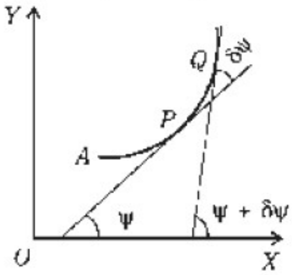

the bend of the arc PQ. This suggests us to make the following definitions :

(*i*) $\delta\psi$ *is defined to be* **total curvature** *of the arc PQ ;*

(*ii*) *the ratio* $\dfrac{\delta\psi}{\delta s}$ *is defined to be the* **average curvature** *of the arc PQ ;*

(*iii*) *the* **curvature** *of the curve at P is defined to be* $\lim\limits_{Q\to P}\dfrac{\delta\psi}{\delta s}$ *i.e.,* $\dfrac{d\psi}{ds}$.

Thus $\dfrac{d\psi}{ds}$ *is a mathematical measure for the curvature of curve at any point P.*

3 Radius of Curvature **(Meerut 2010)**

Let P be a given point on a given curve, and Q any other point on it in the neighbourhood of P. Let N be the point of intersection of the normals at P and Q. Suppose N tends to definite position C as Q tends to P, whether from the right or from the left.

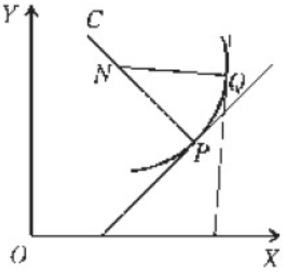

Then C is called the **centre of curvature** of the curve at P.

The distance CP is called the **radius of curvature** of the curve at P and is usually denoted by the Greek letter ρ.

The circle with its centre at C and radius CP is called the **circle of curvature** at P. Any chord drawn through P, of the circle of curvature at P, is called a **chord of curvature at** P.

4 Intrinsic Formula for the Radius of Curvature **(Kashi 2011)**

The relation between s and ψ for any curve is called its **intrinsic equation**. Let P be a given point on the curve $s = f(\psi)$, and Q a point on it in the neighbourhood of P. Let ψ and $\psi + \delta\psi$ be the angles which tangents at P and Q make with the x-axis. Let A be any fixed point on the curve. Let

arc $AP = s$, arc $AQ = s + \delta s$, so that arc $PQ = \delta s$.

Let R be the point of intersection of the tangents at P and Q and N be the point of intersection of the normals at these two points. Suppose $N \to C$ as $Q \to P$.

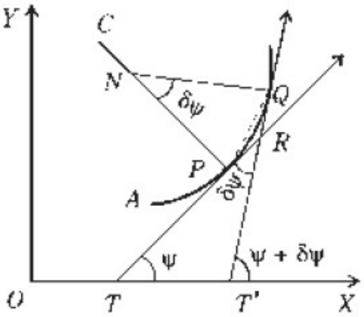

Then the radius of curvature at $P = \rho = \lim\limits_{Q\to P} PN$.

We have $\angle PNQ = \angle TRT' = \delta\psi$.

Now from the triangle PNQ, we have

$$\frac{PN}{\sin NQP} = \frac{\text{chord } PQ}{\sin PNQ} = \frac{\text{chord } PQ}{\sin \delta\psi}.$$

$$\therefore \qquad PN = \frac{\text{chord } PQ}{\sin \delta\psi}\sin NQP = \frac{\text{chord } PQ}{\delta s}\cdot\frac{\delta s}{\delta\psi}\cdot\frac{\delta\psi}{\sin\delta\psi}\cdot \sin NQP.$$

Now as $Q \to P$, we have $\delta\psi \to 0$, $\delta s \to 0$, chord $PQ \to$ tangent at P,

$QN \to$ normal at P and consequently $\angle NQP \to \dfrac{\pi}{2}$.

Therefore $\rho = \lim\limits_{Q \to P} PN$

$$= \left(\lim_{Q \to P} \frac{\text{chord}\, PQ}{\text{arc}\, PQ} \right) \cdot \left(\lim_{\delta\psi \to 0} \frac{\delta s}{\delta \psi} \right) \cdot \left(\lim_{\delta\psi \to 0} \frac{\delta \psi}{\sin \delta\psi} \right) \cdot \left(\lim_{Q \to P} \sin NQP \right)$$

$$= 1 \cdot \left(\frac{ds}{d\psi} \right) \cdot 1.\sin \frac{\pi}{2} = \frac{ds}{d\psi}.$$

Hence, $\boldsymbol{\rho = \frac{ds}{d\psi}}$.

Corollary : The curvature of the curve at any point P is by definition, equal to $\frac{d\psi}{ds}$. Hence *the curvature of the curve at any point is equal to the reciprocal of the radius of curvature at that point i.e., curvature =* $\frac{1}{\rho}$.

Example : *Find the radius of curvature for the curve whose intrinsic equation is*

$$s = a \log \tan \left(\frac{\pi}{4} + \frac{\psi}{2} \right).$$

Solution : We have $\rho = \frac{ds}{d\psi} = a \frac{1}{\tan\left(\frac{\pi}{4} + \frac{\psi}{2}\right)} \sec^2 \left(\frac{\pi}{4} + \frac{\psi}{2} \right) \cdot \frac{1}{2}$

$$= \frac{a}{2 \sin\left(\frac{\pi}{4} + \frac{\psi}{2}\right) \cos\left(\frac{\pi}{4} + \frac{\psi}{2}\right)} = \frac{a}{\sin\left(\frac{\pi}{2} + \psi\right)} = \frac{a}{\cos\psi} = a \sec\psi.$$

5 Cartesian Formula for Radius of Curvature (Gorakhpur 2005)

Let the equation of the curve be $y = f(x)$.

We know that $\frac{dy}{dx} = \tan\psi$.

Differentiating with respect to x, we get

$$\frac{d^2 y}{dx^2} = \sec^2 \psi \cdot \frac{d\psi}{dx} = \sec^2 \psi \frac{d\psi}{ds} \cdot \frac{ds}{dx}.$$

$$\therefore \quad \frac{d\psi}{ds} = \frac{\frac{d^2 y}{dx^2}}{\sec^2 \psi \frac{ds}{dx}}. \qquad \therefore \quad \frac{ds}{d\psi} = \frac{(1 + \tan^2 \psi) \cdot \frac{ds}{dx}}{\frac{d^2 y}{dx^2}}.$$

But we have $\frac{ds}{dx} = \sqrt{\left\{ 1 + \left(\frac{dy}{dx} \right)^2 \right\}}$.

Hence $$\boldsymbol{\rho = \frac{ds}{d\psi} = \frac{\left\{ 1 + \left(\frac{dy}{dx} \right)^2 \right\}^{3/2}}{\frac{d^2 y}{dx^2}} = \frac{(1 + y_1^2)^{3/2}}{y_2}}.$$ **(Bundelkhand 2008)**

Note 1 : The radius of curvature ρ can come out to be positive or negative. If in the relation

$$\frac{ds}{dx}=\sqrt{\left\{1+\left(\frac{dy}{dx}\right)^2\right\}},$$

we take the positive sign before the radical, the value of ρ is positive or negative according as $\frac{d^2y}{dx^2}$ is positive or negative. However, if we define the radius of curvature in such a way that it is to be always positive, then we should ignore the sign whenever we get a negative value for ρ.

Note 2 : Since the radius of curvature is a length therefore its value is independent of the choice of x-axis and y-axis. Hence interchanging x and y, we obtain

$$\rho=\frac{\left\{1+\left(\frac{dx}{dy}\right)^2\right\}^{3/2}}{\frac{d^2x}{dy^2}}.$$

This formula is specially useful when (dy/dx) is infinite *i.e.*, when the tangent is perpendicular to x-axis.

Illustrative Examples

Example 1 : *Find the curvature at the point* $(3a/2, 3a/2)$ *of the curve* $x^3+y^3=3axy.$ **(Meerut 2010; Agra 05; Kumaun 12; Kashi 14; Avadh 14)**

Solution : The curve is $x^3+y^3=3axy.$ ...(1)

Differentiating with respect to x, we get

$$3x^2+3y^2\frac{dy}{dx}=3ay+3ax\frac{dy}{dx}$$

or $$x^2+y^2\frac{dy}{dx}=ay+ax\frac{dy}{dx}. \qquad \ldots(2)$$

$$\therefore \quad \frac{dy}{dx}=\frac{x^2-ay}{ax-y^2}. \qquad \therefore \quad \left[\frac{dy}{dx}\right]_{\left(\frac{3}{2}a,\,\frac{3}{2}a\right)}=-1.$$

Again, differentiating (2), with respect to x, we get

$$2x+2y\left[\frac{dy}{dx}\right]^2+y^2\frac{d^2y}{dx^2}=a\frac{dy}{dx}+a\frac{dy}{dx}+ax\frac{d^2y}{dx^2}$$

or $$(ax-y^2)\frac{d^2y}{dx^2}=2x+2y\left(\frac{dy}{dx}\right)^2-2a\frac{dy}{dx}. \qquad \ldots(3)$$

Putting $x=\frac{3a}{2}, y=\frac{3a}{2}$ and $\left[\frac{dy}{dx}\right]_{(3a/2,\,3a/2)}=-1$ in (3), we get

$$\left[\frac{d^2y}{dx^2}\right]_{(3a/2,\,3a/2)}=-\frac{32}{3}\cdot\frac{1}{a}.$$

Hence the radius of curvature ρ at $\left(\frac{3a}{2},\frac{3a}{2}\right)$

$$= \left[\frac{\left\{1+\left(\frac{dy}{dx}\right)^2\right\}^{3/2}}{\frac{d^2y}{dx^2}}\right]_{(3a/2,\,3a/2)} \qquad = \frac{(1+1)^{3/2}}{-\frac{32}{3}\cdot\frac{1}{a}} = -\frac{3a}{8\sqrt{2}}.$$

$\therefore$ Curvature at $\left(\frac{3a}{2}, \frac{3a}{2}\right) = \frac{1}{\rho} = -\frac{8\sqrt{2}}{3a}$.

If we ignore the negative sign, the value of curvature at $\left(\frac{3a}{2}, \frac{3a}{2}\right) = \frac{8\sqrt{2}}{3a}$.

Example 2 : *If a curve is defined by the equations* $x = f(t)$ *and* $y = \phi(t)$, *prove that the radius of curvature* ρ *is equal to* $\frac{(x'^2 + y'^2)^{3/2}}{x'y'' - y'x''}$, *where accents (i.e., dashes) denote differentiation with respect to t.* **(Kumaun 2007)**

Solution : We have $\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{\phi'(t)}{f'(t)} = \frac{y'}{x'}$.

Also $\frac{d^2y}{dx^2} = \frac{d}{dx}\left(\frac{y'}{x'}\right) = \left\{\frac{d}{dt}\left(\frac{y'}{x'}\right)\right\}\cdot\frac{dt}{dx}$.

$$= \frac{y''x' - x''y'}{x'^2}\cdot\frac{1}{x'} \qquad \left[\because \frac{dx}{dt} = x' \text{ and } \frac{dt}{dx} = \frac{1}{x'}\right]$$

$$= \frac{y''x' - x''y'}{x'^3}.$$

Hence $\rho = \frac{\left\{1+\left(\frac{dy}{dx}\right)^2\right\}^{3/2}}{\frac{d^2y}{dx^2}} = \frac{\left\{1+\frac{y'^2}{x'^2}\right\}^{3/2}}{\frac{y''x' - x''y'}{x'^3}} = \frac{(x'^2 + y'^2)^{3/2}}{y''x' - x''y'}.$

Example 3 : *In the cycloid* $x = a(t + \sin t)$, $y = a(1 - \cos t)$, *prove that* $\rho = 4a\cos\frac{1}{2}t$. **(Lucknow 2005, 09, 10; Bundelkhand 07, 12, 14; Kumaun 07, 10)**

Solution : Here $\frac{dx}{dt} = a(1 + \cos t)$ and $\frac{dy}{dt} = a\sin t$.

$$\therefore \quad \frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{a\sin t}{a(1+\cos t)} = \frac{2\sin t/2\cos t/2}{2\cos^2 t/2} = \tan\frac{t}{2}.$$

Also $\frac{d^2y}{dx^2} = \frac{d}{dx}\left(\frac{dy}{dx}\right) = \frac{d}{dx}\left(\tan\frac{t}{2}\right) = \frac{1}{2}\cdot\sec^2\frac{t}{2}\cdot\frac{dt}{dx}$

$$= \frac{1}{2}\sec^2\frac{t}{2}\cdot\frac{1}{a(1+\cos t)} = \frac{1}{2}\sec^2\frac{t}{2}\cdot\frac{1}{2a\cos^2 t/2} = \frac{1}{4a}\sec^4\frac{t}{2}.$$

Hence $\rho = \frac{\left\{1+\tan^2\frac{t}{2}\right\}^{3/2}}{\frac{1}{4a}\sec^4\frac{t}{2}} = \frac{4a\sec^3\frac{t}{2}}{\sec^4\frac{t}{2}} = 4a\cos\frac{1}{2}t\cdot$

Example 4 : *If CP, CD be a pair of conjugate semi-diameters of an ellipse, prove that the radius of curvature at P is CD^3/ab, a and b being the lengths of the semi-axes of the ellipse.* **(Meerut 2001, 05B; Rohilkhand 11)**

Solution : (Note. Two perpendicular diameters are called **conjugate diameters**)

Let CP and CD be a pair of conjugate semi-diameters of the ellipse $\dfrac{x^2}{a^2}+\dfrac{y^2}{b^2}=1$, the centre C is origin.

Let 't' be the eccentric angle of the point P. Then the co-ordinates of P are $x = a\cos t$, $y = b\sin t$.

The eccentric angle of D will be $t+\frac{1}{2}\pi$, so that the co-ordinates of D are

$$\left[a\cos\left(\tfrac{1}{2}\pi+t\right), b\sin\left(\tfrac{1}{2}\pi+t\right)\right] \text{ i.e., } (-a\sin t, b\cos t).$$

Now for the point P we have $x = a\cos t$ and $y = b\sin t$.

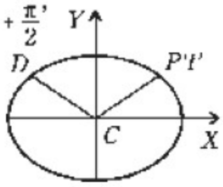

$$\therefore \quad \frac{dx}{dt} = -a\sin t,\ \frac{dy}{dt} = b\cos t.$$

$$\text{Hence} \quad \frac{dy}{dx} = -\frac{b}{a}\cot t.$$

$$\text{Also} \quad \frac{d^2y}{dx^2} = \frac{d}{dx}\left(\frac{dy}{dx}\right) = \frac{d}{dx}\left(-\frac{b}{a}\cot t\right)$$

$$= \left\{\frac{d}{dt}\left(-\frac{b}{a}\cot t\right)\right\}\cdot\frac{dt}{dx}$$

$$= \left(\frac{b}{a}\operatorname{cosec}^2 t\right)\cdot\left(-\frac{1}{a}\operatorname{cosec} t\right) = -\frac{b}{a^2}\operatorname{cosec}^3 t\cdot$$

$\therefore$ Radius of curvature of the point 't'

$$= \rho = \frac{\left\{1+\dfrac{b^2\cos^2 t}{a^2\sin^2 t}\right\}^{3/2}}{-\dfrac{b}{a^2}\operatorname{cosec}^3 t} = -\frac{(a^2\sin^2 t+b^2\cos^2 t)^{3/2}}{ab}.$$

Neglecting the negative sign, we have $\rho = \dfrac{(a^2\sin^2 t+b^2\cos^2 t)^{3/2}}{ab}$.

$$\text{Now} \quad CD = \sqrt{\{(-a\sin t-0)^2+(b\cos t-0)^2\}} = (a^2\sin^2 t+b^2\cos^2 t)^{1/2}.$$

$$\therefore \quad \frac{CD^3}{ab} = \frac{(a^2\sin^2 t+b^2\cos^2 t)^{3/2}}{ab} = \rho.$$

Example 5 : *For the curve* $y=\dfrac{ax}{a+x}$, *if* ρ *is the radius of curvature at any point* (x, y), *show that* $(2\rho/a)^{2/3} = (y/x)^2+(x/y)^2$.

(Kumaun 2008, 09; Rohilkhand 10B, 13, 14; Avadh 10, 13)

Solution : We have $y=\dfrac{ax}{a+x}$ ...(1)

$$\therefore \quad \frac{dy}{dx} = a\frac{(a+x)-x}{(a+x)^2} = \frac{a^2}{(a+x)^2} = a^2(a+x)^{-2},$$

and $$\frac{d^2y}{dx^2}=\frac{d}{dx}\left(\frac{dy}{dx}\right)=-2a^2(a+x)^{-3}=\frac{-2a^2}{(a+x)^3}=\frac{-2a^2}{(ax/y)^3}=-\frac{2y^3}{ax^3}.$$

Now $$1+\left(\frac{dy}{dx}\right)^2=1+\frac{a^4}{(a+x)^4}=1+\frac{a^4}{(ax/y)^4}=1+\frac{y^4}{x^4}=\frac{x^4+y^4}{x^4}.$$

$$\therefore \quad \rho=\frac{[1+(dy/dx)^2]^{3/2}}{d^2y/dx^2}=\frac{[(x^4+y^4)/x^4]^{3/2}}{(2y^3/ax^3)},$$

(negative sign being neglected)

$$=\frac{a(x^4+y^4)^{3/2}}{2x^6(y^3/x^3)}=\frac{a}{2}\frac{(x^4+y^4)^{3/2}}{x^3y^3}.$$

Hence $$\left(\frac{2\rho}{a}\right)^{2/3}=\frac{x^4+y^4}{x^2y^2}=\frac{x^4}{x^2y^2}+\frac{y^4}{x^2y^2}=\frac{x^2}{y^2}+\frac{y^2}{x^2}=\left(\frac{x}{y}\right)^2+\left(\frac{y}{x}\right)^2.$$

6 Radius of Curvature at the Origin (Another Method)

Radius of curvature at the origin can be found by substituting $x=0, y=0$ in the value of ρ obtained from the formula of article 5. Here we shall give an alternative method.

Since the curve passes through the origin, therefore $(y)_0=0$ *i.e.*, the value of y at $x=0$ is 0.

Let $$\left(\frac{dy}{dx}\right)_{(0,0)}=(y_1)_0=p \quad \text{and} \quad \left(\frac{d^2y}{dx^2}\right)_{(0,0)}=(y_2)_0=q.$$

Then ρ (at origin) $=\dfrac{(1+p^2)^{3/2}}{q}$. ...(1)

To get the values of p and q, we know by Maclaurin's theorem that

$$y=(y)_0+(y_1)_0x+\frac{(y_2)_0}{2!}x^2+\dots \qquad \dots(2)$$

Since the curve passes through origin, therefore (2) becomes

$$y=px+\tfrac{1}{2}qx^2+\dots\dots$$

Thus to get the values of p and q, we should get from the equation of the curve an expansion for y in ascending powers of x by algebraic or trigonometric methods. The coefficient of x in this expansion will be equal to p and the coefficient of x^2 will be equal to $\frac{1}{2}q$. Putting the values of p and q in (1), we shall get ρ at origin.

7 Newton's Method for Radius of Curvature at the Origin

Suppose a curve passes through the origin and the x-axis is tangent to the curve at origin.

Then $(y)_0=0$ and $\left(\dfrac{dy}{dx}\right)_{(0,0)}$ *i.e,* $(y_1)_0=0$.

Therefore in this case by Maclaurin's expansion, we have

$$y = 0 + 0 . x + \frac{q}{2} . x^2 + \frac{(y_3)_0}{3\,!} x^3 + \ldots \qquad \ldots(1)$$

where $(y_2)_0 = q$.

Multiplying (1) by $\frac{2}{x^2}$, we get $\frac{2y}{x^2} = q + \frac{2}{3!} (y_3)_0 \, x + \ldots \qquad \ldots(2)$

Taking limit as $x \to 0$ of both sides of (2), we get $\lim\limits_{x \to 0} \frac{2y}{x^2} = q$.

Also in this case ρ at origin $= \frac{(1 + 0)^{3/2}}{q} = \frac{1}{q} = \lim\limits_{x \to 0} \frac{x^2}{2y}$.

Therefore *when x-axis is tangent to the curve at the origin,*

$$\rho \text{ (at origin)} = \lim_{x \to 0} \left(\frac{x^2}{2y} \right).$$

Similarly it can be shown that if y-axis is tangent to the curve at the origin, then

$$\rho \text{ (at origin)} = \lim_{x \to 0} \frac{y^2}{2x}.$$

These two formulae are known as **Newton's formulae.**

Illustrative Examples

Example 1 : *Find the radius of curvature at origin for the curve*

$$x^3 + y^3 - 2x^2 + 6y = 0.$$

Solution : The curve passes through origin. Equating to zero the lowest degree terms we get $y = 0$, *i.e.,* x-axis as tangent to the curve at origin

$\therefore$ By Newton's method ρ (at origin) $= \lim\limits_{x \to 0} \frac{x^2}{2y}$.

Dividing by $2y$, the equation of the curve can be written as

$$x \cdot \frac{x^2}{2y} + \frac{1}{2} y^2 - 2 . \frac{x^2}{2y} + 3 = 0.$$

Taking limit as $x \to 0$, $y \to 0$ and $\lim\limits_{x \to 0} \frac{x^2}{2y} = \rho$, we get

$$0\rho + 0 - 2\rho + 3 = 0 \ \ i.e., \ \ \rho = 3/2.$$

Example 2 : *Show that the radii of curvature of the curve* $y^2 = x^2 (a + x)/(a - x)$ *at the origin are* $\pm\, a\sqrt{2}$. **(Gorakhpur 2006; Kumaun 13)**

Solution : The curve passes through the origin and the tangents at origin are $y^2 = x^2$ *i.e.,* $y = \pm\, x$. Thus neither of the coordinate axes is tangent at the origin. Therefore we cannot apply Newton's method. But the equation of the curve can be written as

$$y = \frac{\pm\, x (a + x)^{1/2}}{(a - x)^{1/2}} \quad \text{or} \quad y = \pm\, x \left(1 + \frac{x}{a} \right)^{1/2} \left(1 - \frac{x}{a} \right)^{-1/2}$$

or $$y = \pm\, x \left\{ 1 + \frac{1}{2} \frac{x}{a} + \ldots\ldots \right\} \left\{ 1 + \frac{1}{2} \frac{x}{a} + \ldots\ldots \right\}$$

expanding by Binomial Theorem

or $$y = \pm\, x \left[1 + \frac{x}{a} + \ldots\ldots \right].$$

Comparing this equation with the equation

$$y = px + q\frac{x^2}{2} + \ldots, \text{ we get } p = 1, q = \frac{2}{a} \quad \text{or} \quad p = -1, q = -\frac{2}{a}.$$

But ρ at origin $= \dfrac{(1+p^2)^{3/2}}{q}$.

$\therefore$ When $p = 1, q = \dfrac{2}{a}$, ρ at origin $= \dfrac{(1+1)^{3/2}}{\dfrac{2}{a}} = a\sqrt{2}$.

Also when $p = -1, q = -\dfrac{2}{a}$, ρ at origin $= \dfrac{(1+1)^{3/2}}{-2/a} = -a\sqrt{2}$.

Example 3 : *Find the radii of curvature at the origin for the curve*

$$y^2 - 3xy - 4x^2 + x^3 + x^4 y + y^5 = 0.$$

Solution : The curve passes through the origin and the tangents at origin are $y^2 - 3xy - 4x^2 = 0$. Thus neither of the co-ordinate axes is tangent at the origin. Therefore Newton's method cannot be applied. Also we cannot put the equation of the curve in the form $y = px + \dfrac{qx^2}{2} + \ldots\ldots$

Hence substituting $px + \dfrac{qx^2}{2} + \ldots$ for y in the equation of the curve, we get the identity,

$$\left(px + \frac{qx^2}{2} + \ldots\ldots\right)^2 - 3x\left(px + \frac{qx^2}{2} + \ldots\ldots\right)$$
$$- 4x^2 + x^3 + x^4\left(px + \frac{qx^2}{2} + \ldots\ldots\right) + \ldots = 0.$$

Equating to zero the coefficients of x^2 and x^3, we get

$$p^2 - 3p - 4 = 0 \quad \text{and} \quad pq - \frac{3q}{2} + 1 = 0.$$

Solving these we get $p = 4, -1$.

When $p = 4, q = -2/5$ and when $p = -1, q = 2/5$.

Now ρ (at origin) $= \dfrac{(1+p^2)^{3/2}}{q}$.

$\therefore$ When $p = 4, q = 2/5$, ρ at origin $= \dfrac{(1+16)^{3/2}}{-2/5} = -\dfrac{85\sqrt{17}}{2}$

and when $p = -1, q = 2/5$, ρ at origin $= \dfrac{(1+1)^{3/2}}{2/5} = 5\sqrt{2}$.

Comprehensive Exercise 1

1. Find the radius of curvature at the point (s, ψ) on the following curves :

(i) $s = c \tan \psi$ (Catenary)

(ii) $s = 8a \sin^2 \frac{1}{6}\psi$ (Cardioid)

(iii) $s = 4a \sin \psi$ (Cycloid) **(Bundelkhand 2001; Rohilkhand 08; Kashi 11)**

(iv) $s = c \log \sec \psi$ (Tractrix). **(Kashi 2012)**

2. Find the radius of curvature at the point (x, y) on the following curves :

(i) $a^2 y = x^3 - a^3$ (ii) $y^2 = 4ax$

(iii) $xy = c^2$ (iv) $ay^2 = x^3$

(v) $y = \frac{1}{2} a \left(e^{x/a} + e^{-x/a}\right)$ **(Agra 2007)**

(vi) $y = c \log \sec (x/c)$ **(Kanpur 2007; Purvanchal 09)**

(vii) $x^{1/2} + y^{1/2} = a^{1/2}$.

(viii) $x^{2/3} + y^{2/3} = a^{2/3}$ **(Rohilkhand 2009B; Kashi 12)**

(ix) $x^m + y^m = 1$.

3. (i) Find the radius of curvature of the curve $y = e^x$, at the point where it crosses the y-axis. **(Agra 2014)**

(ii) Find the radius of curvature of the curve $\sqrt{x} + \sqrt{y} = 1$ at the point $\left(\frac{1}{4}, \frac{1}{4}\right)$.

4. (i) Prove that at the point $x = \frac{1}{2}\pi$ of the curve $y = 4 \sin x - \sin 2x$, $\rho = \frac{5\sqrt{5}}{4}$.

(ii) Prove that for the curve $s = a \log \cot \left(\frac{\pi}{4} - \frac{\psi}{2}\right) + a \sin \psi \sec^2 \psi$, $\rho = 2a \sec^3 \psi$;

and hence that $\frac{d^2 y}{dx^2} = \frac{1}{2a}$. **(Lucknow 2008)**

5. In the curve $y = ae^{x/a}$, prove that

$$\rho = a \sec^2 \theta \operatorname{cosec} \theta, \text{ where } \theta = \tan^{-1}(y/a).$$

6. Show that the radius of curvature at a point $(a \cos^3 \theta, a \sin^3 \theta)$ on the curve

$x^{2/3} + y^{2/3} = a^{2/3}$ is $3a \sin \theta \cos \theta$. **(Meerut 2000, 05; Kashi 13)**

7. Prove that for the ellipse $x^2/a^2 + y^2/b^2 = 1$,

$\rho = \frac{a^2 b^2}{p^3}$, p being the perpendicular from the centre upon the tangent at (x, y). **(Meerut 2002, 04B, 07; Avadh 05, 09)**

8. In the ellipse $x^2/a^2 + y^2/b^2 = 1$, show that the radius of curvature at an end of the major axis is equal to the semi-latus rectum of the ellipse.

9. If ρ and ρ' be the radii of curvature at the extremities of two conjugate diameters of an ellipse, prove that $(\rho^{2/3} + \rho'^{2/3})(ab)^{2/3} = a^2 + b^2$.

(Meerut 2001, 03, 04, 06, 11; Bundelkhand 06; Kanpur 11; Lucknow 07; Rohilkhand 13B; Kashi 14)

10. Prove that if ρ be the radius of curvature at any point P on the parabola $y^2 = 4ax$ and S be its focus, then ρ^2 varies as $(SP)^3$.

11. If the co-ordinates of a point on a curve be given by the equations

$$x = c \sin 2\theta (1 + \cos 2\theta), y = c \cos 2\theta (1 - \cos 2\theta),$$

show that the radius of curvature at the point is $4c \cos 3\theta$.

12. If the co-ordinates of a point on a curve be given by the equations

$$x = a \sin t - b \sin (at/b), y = a \cos t - b \cos (at/b),$$

show that the radius of curvature at the point is $\frac{4ab}{a + b} \sin \frac{a - b}{2b} t$.

13. Prove that for the curve $s = ae^{x/a}$, $a\rho = s(s^2 - a^2)^{1/2}$.

Show that for the curve $s^2 = 8ay$, $\rho = 4a\sqrt{\left(1 - \frac{y}{2a}\right)}$. **(Kanpur 2009)**

14. Find the radius of curvature at the origin of the following curves :

(i) $y = x^4 - 4x^3 - 18x^2$. (ii) $y = x^3 + 5x^2 + 6x$.

15. Show that the radii of curvature of the curve $a(y^2 - x^2) = x^3$ at the origin are $\pm 2a\sqrt{2}$.

Answers 1

1. (i) $c \sec^2 \psi$ (ii) $\frac{4}{3} a \sin \frac{1}{3} \psi$ (iii) $4a \cos \psi$ (iv) $c \tan \psi$.

2. (i) $\frac{(a^4 + 9x^4)^{3/2}}{6a^4 x}$ (ii) $\left(\frac{2}{\sqrt{a}}\right)(x + a)^{3/2}$ (iii) $\frac{(x^2 + y^2)^{3/2}}{2c^2}$

(iv) $\frac{1}{6a}(4a + 9x)^{3/2} x^{1/2}$ (v) y^2/a (vi) $c \sec(x/c)$

(vii) $\frac{2(x + y)^{3/2}}{\sqrt{a}}$ (viii) $3a^{1/3} x^{1/3} y^{1/3}$ (ix) $\frac{(x^{2m-2} + y^{2m-2})^{3/2}}{(1 - m)x^{m-2} y^{m-2}}$.

3. (i) $\sqrt{8}$. (ii) $\frac{1}{\sqrt{2}}$. **14.** (i) $1/36$, (ii) $37\sqrt{(37)}/10$.

8 Pedal Formula for Radius of Curvature

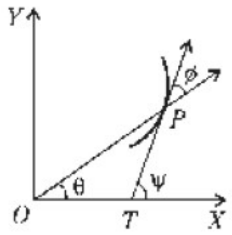

We have the relation $\psi = \theta + \phi$, ...(1)

as is obvious from the adjoining figure.

Differentiating (1) w.r.t. s, we get

$$\frac{d\psi}{ds} = \frac{d\theta}{ds} + \frac{d\phi}{ds} \quad \text{or} \quad \frac{d\psi}{ds} = \frac{d\theta}{ds} + \frac{d\phi}{dr} \cdot \frac{dr}{ds}$$

or $$\frac{1}{\rho} = \frac{1}{r} \sin\phi + \cos\phi \frac{d\phi}{dr}$$

$$\left[\because \rho = \frac{ds}{d\psi}, \sin\phi = r\frac{d\theta}{ds} \text{ and } \cos\phi = \frac{dr}{ds}\right]$$

or $$\frac{1}{\rho} = \frac{1}{r}\left(\sin\phi + r\cos\phi \frac{d\phi}{dr}\right) = \frac{1}{r}\frac{d}{dr}(r\sin\phi) = \frac{1}{r}\frac{dp}{dr}. \qquad [\because p = r\sin\phi]$$

Hence $\boldsymbol{\rho = r\frac{dr}{dp}}$. **(Meerut 2003, 07B)**

9 Polar Formula for Radius of Curvature

We know that $\frac{1}{p^2} = \frac{1}{r^2} + \frac{1}{r^4}\left(\frac{dr}{d\theta}\right)^2$. ...(1)

Differentiating (1) with respect to r, we get

$$-\frac{2}{p^3}\frac{dp}{dr} = -\frac{2}{r^3} - \frac{4}{r^5}\left(\frac{dr}{d\theta}\right)^2 + \frac{1}{r^4}\left\{\frac{d}{dr}\left(\frac{dr}{d\theta}\right)^2\right\}$$

$$= -\frac{2}{r^3} - \frac{4}{r^5}\left(\frac{dr}{d\theta}\right)^2 + \frac{1}{r^4}\left\{\frac{d}{d\theta}\left(\frac{dr}{d\theta}\right)^2\right\} \cdot \frac{d\theta}{dr}$$

$$= -\frac{2}{r^3} - \frac{4}{r^5}\left(\frac{dr}{d\theta}\right)^2 + \frac{1}{r^4} \cdot 2\left(\frac{dr}{d\theta}\right) \cdot \frac{d^2 r}{d\theta^2} \cdot \frac{d\theta}{dr}$$

$$= -\frac{2}{r^3} - \frac{4}{r^5}\left(\frac{dr}{d\theta}\right)^2 + \frac{2}{r^4}\frac{d^2 r}{d\theta^2}$$

$$\therefore \qquad \frac{1}{p^3}\frac{dp}{dr} = \frac{1}{r^5}\left\{r^2 + 2\left(\frac{dr}{d\theta}\right)^2 - r\frac{d^2 r}{d\theta^2}\right\}$$

Therefore $$\rho = r\frac{dr}{dp} = \frac{r \cdot \dfrac{1}{p^3}}{\dfrac{1}{r^5}\left\{r^2 + 2\left(\dfrac{dr}{d\theta}\right)^2 - r\dfrac{d^2 r}{d\theta^2}\right\}}.$$

But from (i), $$\frac{1}{p^3} = \left\{\frac{1}{r^2} + \frac{1}{r^4}\left(\frac{dr}{d\theta}\right)^2\right\}^{3/2} = \frac{1}{r^6}\left\{r^2 + \left(\frac{dr}{d\theta}\right)^2\right\}^{3/2}.$$

Hence $$\rho = \frac{r^6 \cdot \dfrac{1}{r^6}\left\{r^2 + \left(\dfrac{dr}{d\theta}\right)^2\right\}^{3/2}}{r^2 + 2\left(\dfrac{dr}{d\theta}\right)^2 - r\dfrac{d^2 r}{d\theta^2}}.$$

Therefore, $$\boldsymbol{\rho = \frac{\left\{r^2 + \left(\dfrac{dr}{d\theta}\right)^2\right\}^{3/2}}{r^2 + 2\left(\dfrac{dr}{d\theta}\right)^2 - r\dfrac{d^2 r}{d\theta^2}}}.$$ **(Kumaun 2011)**

Corollary : If we put $u = \dfrac{1}{r}$ or $r = \dfrac{1}{u}$, then

$$\frac{dr}{d\theta} = -\frac{1}{u^2}\frac{du}{d\theta} \quad \text{and} \quad \frac{d^2 r}{d\theta^2} = \frac{2}{u^3}\left(\frac{du}{d\theta}\right)^2 - \frac{1}{u^2}\frac{d^2 u}{d\theta^2}.$$

Putting these values in the polar formula for ρ, we get

$$\rho = \frac{\left\{\dfrac{1}{u^2} + \dfrac{u'^2}{u^4}\right\}^{3/2}}{\dfrac{1}{u^2} + \dfrac{2u'^2}{u^4} - \dfrac{2u'^2}{u^4} + \dfrac{u''}{u^3}} = \boldsymbol{\frac{(u^2 + u'^2)^{3/2}}{u^3(u + u'')}},$$

where dashes denote differentiation with respect to θ.

Note : We see that the pedal formula for ρ is simpler than the polar formula. Therefore in case the equation of the curve is given in polar form, it is often convenient to change it first to pedal equation and then to find ρ with the help of the pedal formula.

Illustrative Examples

Example 1 : *Show that for the cardioid* $r = a(1 + \cos\theta)$, $\rho = \frac{2}{3}\sqrt{(2ar)}$.

(Purvanchal 2006, 11; Rohilkhand 09B, 10; Agra 14)

Solution : The curve is $r = a(1 + \cos\theta)$.

$$\therefore \qquad \frac{dr}{d\theta} = -a\sin\theta \quad \text{and} \quad \frac{d^2r}{d\theta^2} = -a\cos\theta.$$

$$\text{Now} \quad \rho = \frac{\{r^2 + (dr/d\theta)^2\}^{3/2}}{r^2 + 2(dr/d\theta)^2 - r(d^2r/d\theta^2)}$$

$$= \frac{\{a^2(1 + \cos\theta)^2 + a^2\sin^2\theta\}^{3/2}}{a^2(1 + \cos\theta)^2 + 2(-a\sin\theta)^2 - a(1 + \cos\theta)(-a\cos\theta)}$$

$$= \frac{\left(4a^2\cos^4\frac{1}{2}\theta + 4a^2\cos^2\frac{1}{2}\theta\sin^2\frac{1}{2}\theta\right)^{3/2}}{a^2 + 2a^2(\cos^2\theta + \sin^2\theta) + 3a^2\cos\theta}$$

$$= \frac{(4a^2\cos^2\frac{1}{2}\theta)^{3/2}[\cos^2\frac{1}{2}\theta + \sin^2\frac{1}{2}\theta]^{3/2}}{3a^2(1 + \cos\theta)} = \frac{8a^3\cos^3\frac{1}{2}\theta}{6a^2\cos^2\frac{1}{2}\theta} = \left[\frac{4a}{3}\right]\cos\frac{1}{2}\theta.$$

$$\text{But} \qquad r = a(1 + \cos\theta) = 2a\cos^2\frac{1}{2}\theta.$$

$$\therefore \qquad \cos\frac{1}{2}\theta = \sqrt{(r/2a)}.$$

$$\text{Hence} \quad \rho = \frac{4a}{3}\sqrt{\left[\frac{r}{2a}\right]} = (2/3)\sqrt{(2ar)}.$$

Note : We could have solved this problem more easily by changing the equation of the curve to pedal form.

Example 2 : *Show that in the rectangular hyperbola* $r^2\cos 2\theta = a^2$, $\rho = r^3/a^2$.

Solution : The curve is $r^2\cos 2\theta = a^2$. …(1)

Taking logarithm of both sides of (1), we get $2\log r + \log\cos 2\theta = 2\log a$.

Differentiating with respect to θ, we get

$$\frac{2}{r}\frac{dr}{d\theta} + \frac{1}{\cos 2\theta}(-2\sin 2\theta) = 0$$

$$\text{or} \qquad \frac{1}{r}\frac{dr}{d\theta} = \cot\phi = \tan 2\theta = \cot\left[\frac{\pi}{2} - 2\theta\right].$$

$$\therefore \qquad \phi = \frac{\pi}{2} - 2\theta.$$

$$\text{Now} \quad p = r\sin\phi = r\sin\left[\frac{\pi}{2} - 2\theta\right] = r\cos 2\theta. \qquad \text{But} \ \cos 2\theta = \frac{a^2}{r^2}\cdot$$

Hence the pedal equation of the curve is

$$p = r\,.\,\frac{a^2}{r^2} \quad \text{or} \quad p = \frac{a^2}{r}\cdot$$

$$\therefore \qquad \frac{dp}{dr} = -\frac{a^2}{r^2}\cdot$$

Hence $\rho = r\dfrac{dr}{dp} = -\dfrac{r^3}{a^2}$.

Neglecting the negative sign, we have $\rho = \dfrac{r^3}{a^2}$.

Example 3 : *Find the radius of curvature at the point* (p, r) *on the ellipse*

$$\frac{1}{p^2} = \frac{1}{a^2} + \frac{1}{b^2} - \frac{r^2}{a^2 b^2}.$$

Solution : Differentiating the given equation with respect to r, we get

$$-\frac{2}{p^3}\frac{dp}{dr} = -\frac{2r}{a^2 b^2}. \qquad \therefore \quad \frac{dp}{dr} = \frac{rp^3}{a^2 b^2}.$$

Hence $\rho = r\dfrac{dr}{dp} = r \cdot \dfrac{a^2 b^2}{rp^3} = \dfrac{a^2 b^2}{p^3}$.

10 Tangential Polar Formula for Radius of Curvature

A relation between p and ψ holding for every point of a curve, is called tangential polar equation of the curve. Thus the tangential polar equation of the curve is of the form $p = f(\psi)$.

We have $\dfrac{dp}{d\psi} = \dfrac{dp}{dr} \cdot \dfrac{dr}{ds} \cdot \dfrac{ds}{d\psi} = \dfrac{dp}{dr}\cos\phi \,.\, \rho$

$$= \frac{dp}{dr}\cos\phi \,.\, r\frac{dr}{dp} \qquad \left[\because \frac{dr}{ds} = \cos\phi \text{ and } \rho = \frac{ds}{d\psi} = r\frac{dr}{dp}\right]$$

$= r\cos\phi$.

Also $p = r\sin\phi$.

$$\therefore \quad p^2 + \left(\frac{dp}{d\psi}\right)^2 = r^2(\sin^2\phi + \cos^2\phi) \text{ or } p^2 + \left(\frac{dp}{d\psi}\right)^2 = r^2. \qquad \ldots(1)$$

Differentiating (1) with respect to p, we get

$$2p + 2\left(\frac{dp}{d\psi}\right) \cdot \frac{d^2p}{d\psi^2} \cdot \frac{d\psi}{dp} = 2r\frac{dr}{dp} \qquad \text{or} \qquad r\frac{dr}{dp} = p + \frac{d^2p}{d\psi^2}.$$

Hence $\rho = \boldsymbol{p + \dfrac{d^2 p}{d\psi^2}}$.

Illustrative Examples

Example 1 : *Show that for the epi-cycloid* $p = a\sin b\psi$, ρ *varies as* p.

Solution : We have $\dfrac{dp}{d\psi} = ab\cos b\psi$ and $\dfrac{d^2p}{d\psi^2} = -ab^2\sin b\psi = -b^2 p$.

$$\therefore \quad \rho = p + \frac{d^2p}{d\psi^2} = p - b^2 p = (1 - b^2)\,p.$$

Hence ρ varies as p.

Comprehensive Exercise 2

1. Find the radius of curvature at the point (p, r) on the following curves :

(i) $p^2 = ar$ (parabola). (ii) $2ap^2 = r^3$ (cardioid).

(iii) $a^2 p = r^3$ (Lemniscate). (iv) $p^2 = \dfrac{r^4}{(r^2 + a^2)}$.

2. Prove that for any curve $\dfrac{r}{\rho} = \sin\phi\left(1 + \dfrac{d\phi}{d\theta}\right)$, where ρ is the radius of curvature and $\tan\phi = r\dfrac{d\theta}{dr}$. **(Gorakhpur 2005; Lucknow 10)**

3. In the curve $p = r^{n+1}/a^n$, show that the radius of curvature varies inversely as the $(n-1)^{th}$ power of the radius vector.

4. Find the radius of curvature at the point (r, θ) on each of the following curves :

(i) $r = a\cos\theta$. **(Kanpur 2006)**

(ii) $r(1 + \cos\theta) = 2a$.

(iii) $r^n = a^n \cos n\theta$. **(Rohilkhand 2005; Kumaun 15)**

(iv) $r^n = a^n \sin n\theta$ **(Agra 2006; Rohilkhand 12; Avadh 12)**

(v) $r = a(1 - \cos\theta)$. **(Avadh 2010; Kumaun 14)**

(vi) $r^2 = a^2\cos 2\theta$.

5. Forming the pedal equation of the curve

$$\theta = a^{-1}(r^2 - a^2)^{1/2} - \cos^{-1}\left(\frac{a}{r}\right), \text{ show that } \rho = \sqrt{(r^2 - a^2)}.$$

(Meerut 2006B, 08; Rohilkhand 06; Kashi 11)

6. For the rectangular hyperbola $xy = c^2$, prove that $\rho = \frac{1}{2} r^3/c^2$, r being the central radius vector of the point considered.

7. Show that at any point on the equiangular spiral $r = ae^{\theta\cot\alpha}$, $\rho = r\operatorname{cosec}\alpha$, and that it subtends a right angle at the pole.

8. If ρ_1, ρ_2 be radii of curvature at the extremities of any chord of the cardioid $r = a(1 + \cos\theta)$, which passes through the pole, then show that

$\rho_1{}^2 + \rho_2{}^2 = 16a^2/9$. **(Kanpur 2008)**

9. Show that the radius of curvature at any point on the curve $r = a(1 \pm \cos\theta)$ varies as square root of the radius vector.

10. Find the radius of curvature of the cardioid $r = a(1 - \cos\theta)$ at the pole (origin).

Answers 2

1. (i) $\dfrac{2p^3}{a^2}$ (ii) $\frac{2}{3}\sqrt{(2ar)}$ (iii) $a^2/3r$

(iv) $(r^2 + a^2)^{3/2}/(r^2 + 2a^2)$. **2.** $\dfrac{a^2 b^2}{p^3}$.

4. (i) $a/2$ (ii) $2\sqrt{(r^3/a)}$ (iii) $\dfrac{a^n}{(n+1)\, r^{n-1}}$

(iv) $\dfrac{a^n}{(n+1)\,r^{n-1}}$ (v) $\frac{2}{3}\sqrt{(2ar)}$ (vi) $\dfrac{a^2}{3r}$.

10. 0.

11 Co-ordinates of Centre of Curvature (Meerut 2008)

Let the equation of the curve be $y = f(x)$.

Let P be the given point (x, y) on this curve and Q the point $(x + \delta x, y + \delta y)$ in the neighbourhood of P. (See the fig. of article 11.3). Let N be the point of intersection of the normals at P and Q. As $Q \to P$, suppose $N \to C$. Then C is the centre of curvature of P.

Suppose co-ordinates of C are (α, β).

From the equation of the curve, we have $\dfrac{dy}{dx} = f'(x) = \phi(x)$, say.

The equation of normal at P is

$$(Y - y)\,\phi(x) + (X - x) = 0. \qquad \text{...(1)}$$

The equation of normal at Q is

$$\{Y - (y + \delta y)\}\,\phi(x + \delta x) + \{X - (x + \delta x)\} = 0. \qquad \text{...(2)}$$

Subtracting (1) from (2), we get

$$(Y - y)\,\{\phi(x + \delta x) - \phi(x)\} - \phi(x + \delta x)\,\delta y - \delta x = 0.$$

Dividing by δx, we get

$$(Y - y)\left\{\frac{\phi(x + \delta x) - \phi(x)}{\delta x}\right\} - \phi(x + \delta x)\,\frac{\delta y}{\delta x} - 1 = 0. \qquad \text{...(3)}$$

The value of Y obtained from this equation will give us the y co-ordinate of the point of intersection of (1) and (2).

Now as $Q \to P$, $\delta x \to 0$ and Y obtained from (3)$\to \beta$.

Therefore taking limit of (3) as $\delta x \to 0$, we get

$$(\beta - y)\lim_{\delta x \to 0}\left\{\frac{\phi(x + \delta x) - \phi(x)}{\delta x}\right\} - \lim_{\delta x \to 0}\phi(x + \delta x)\,.\,\lim_{\delta x \to 0}\frac{\delta y}{\delta x} - 1 = 0$$

or $$(\beta - y)\,\frac{d}{dx}\,\phi(x) - \phi(x)\cdot\frac{dy}{dx} - 1 = 0$$

or $$(\beta - y)\,\frac{d}{dx}\left(\frac{dy}{dx}\right) - \frac{dy}{dx}\cdot\frac{dy}{dx} - 1 = 0. \qquad \left[\because\ \phi(x) = \frac{dy}{dx}\right]$$

or $$(\beta - y)\,\frac{d^2 y}{dx^2} - \left\{\left(\frac{dy}{dx}\right)^2 + 1\right\} = 0.$$

$$\therefore \qquad \boldsymbol{\beta = y + \frac{1 + \left(\dfrac{dy}{dx}\right)^2}{\dfrac{d^2 y}{dx^2}}}.$$

Also (α, β) lies on (1). Therefore, we get

$$(\beta - y)\frac{dy}{dx} + (\alpha - x) = 0$$

i.e.,
$$(\alpha - x) = -(\beta - y)\frac{dy}{dx} = -\frac{dy}{dx}\frac{1 + \left(\frac{dy}{dx}\right)^2}{\frac{d^2y}{dx^2}}.$$

$$\therefore \qquad \boldsymbol{\alpha = x - \frac{\frac{dy}{dx}\left\{1 + \left(\frac{dy}{dx}\right)^2\right\}}{\frac{d^2y}{dx^2}}}.$$

12 Evolute of a Curve

The locus of the centre of curvature of a curve is called its evolute.

13 Equation of the Circle of Curvature

If (α, β) be the co-ordinates of the centre of curvature and ρ the radius of curvature at any point (x, y) on a curve, then the equation of the circle of curvature at that point is $(X - \alpha)^2 + (Y - \beta)^2 = \rho^2$.

Illustrative Examples

Example 1 : *Find the co-ordinates of the centre of curvature for the point* (x, y) *on the parabola* $y^2 = 4ax$. **(Lucknow 2008)**

Also find the equation of the evolute of the parabola.

Solution : Here $2y\frac{dy}{dx} = 4a$, *i.e.,* $\frac{dy}{dx} = \frac{2a}{y} = a^{1/2}x^{-1/2} = \sqrt{\left(\frac{a}{x}\right)}$.

$$\therefore \qquad \frac{d^2y}{dx^2} = -\frac{1}{2}a^{1/2}x^{-3/2}.$$

If (α, β) be the centre of curvature for the point (x, y), then

$$\alpha = x - \frac{\frac{dy}{dx}\left\{1 + \left(\frac{dy}{dx}\right)^2\right\}}{\frac{d^2y}{dx^2}} = x - \frac{\sqrt{\left(\frac{a}{x}\right)}\left\{1 + \frac{a}{x}\right\}}{-\frac{1}{2}\frac{1}{x}\sqrt{\left(\frac{a}{x}\right)}} = x + 2x(1 + a/x)\cdot$$

$$\therefore \qquad \alpha = 3x + 2a, \qquad \text{...(1)}$$

and
$$\beta = y + \frac{1 + \left(\frac{dy}{dx}\right)^2}{\frac{d^2y}{dx^2}} = y + \frac{1 + \left(\frac{a}{x}\right)}{-\frac{1}{2}\frac{1}{x}\sqrt{\left(\frac{a}{x}\right)}}$$

$$= 2a^{1/2}x^{1/2} - 2a^{-1/2}x^{3/2}(1 + a/x)$$
$$= 2a^{1/2}x^{1/2} - 2a^{-1/2}x^{3/2} - 2a^{1/2}x^{1/2}$$

$$\therefore \qquad \beta = -2x\sqrt{\left(\frac{x}{a}\right)}. \qquad \ldots(2)$$

Therefore the required centre of curvature is $\left((3x + 2a), -2x\sqrt{\left(\frac{x}{a}\right)}\right)$.

Evolute of the parabola : Let us eliminate x between (1) and (2). From (2), we get

$$\beta^2 = \frac{4x^3}{a} \quad \text{or} \quad x^3 = \frac{a\beta^2}{4}.$$

From (i), we get $x = \dfrac{\alpha - 2a}{3}$.

$$\therefore \qquad \left(\frac{\alpha - 2a}{3}\right)^3 = \frac{a\beta^2}{4} \quad \text{or} \quad 27\,a\beta^2 = 4(\alpha - 2a)^3.$$

Hence the locus of (α, β) is $27\,ay^2 = 4(x - 2a)^3$, which is the evolute of parabola.

Example 2 : *Prove that the co-ordinates* (α, β) *of the centre of curvature at any point* (x, y) *can be expressed in the form*

$$\alpha = x - \frac{dy}{d\psi} \text{ and } \beta = y + \frac{dx}{d\psi}.$$

Solution : Let $C(\alpha, \beta)$ be the centre of curvature of the point $P(x, y)$ on the curve $y = f(x)$. The line TP is tangent to the curve at P and obviously PC is normal at P to the curve, since the centre of curvature is a point on the normal.

ρ = radius of curvature at $P = PC$ and CN is the ordinate of C. Draw PM perpendicular to CN. Obviously

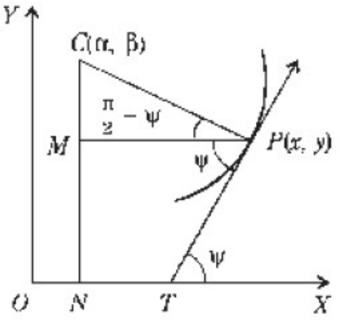

$$\angle CPM = \tfrac{1}{2}\pi - \psi.$$

$$\text{Then} \qquad \alpha = x - PM = x - \rho\sin\psi$$
$$= x - \frac{dy}{ds}\cdot\frac{ds}{d\psi} \qquad \left[\because\ \rho = \frac{ds}{d\psi} \text{ and } \frac{dy}{ds} = \sin\psi\right]$$
$$= x - \frac{dy}{d\psi}.$$

$$\text{Also} \qquad \beta = y + CM = y + \rho\cos\psi$$
$$= y + \frac{dx}{ds}\cdot\frac{ds}{d\psi} \qquad \left[\because\ \cos\psi = \frac{dx}{ds}\right]$$
$$= y + \frac{dx}{d\psi}.$$

14 Chord of Curvature through the Origin (Pole)

Let C be the centre of curvature at the point P on any given curve.

Then $CP = \rho$ = radius of curvature at P. O is the pole. Join OP to meet the circle of curvature in E. Then PE is the chord of curvature through the origin.

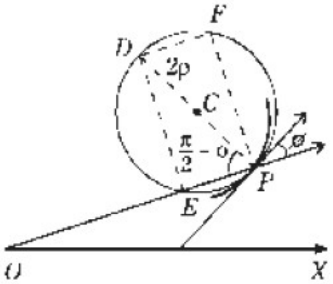

PD is the diameter of the circle of curvature. We have $PD = 2\rho$

and $\angle PED = 90^{\circ}$, being an angle in a semi-circle.

Also *PD* is normal to the curve at *P*.

$\therefore \quad \angle EPD = \frac{1}{2}\pi - \phi.$

Hence from the right-angled triangle *PED*, we have

$$PE = PD \cos\left(\frac{1}{2}\pi - \phi\right) = 2\rho \sin \phi.$$

$\therefore$ *chord of curvature through pole* = $\mathbf{2\rho \sin \phi}$.

Corollary : Chord of curvature perpendicular to the Radius Vector :

Suppose a line through *P*, perpendicular to the radius vector *OP*, meets the circle of curvature in *F*.

Then *PF* is the chord of curvature perpendicular to the radius vector.

We have $PF = ED = 2\rho \sin\left(\frac{1}{2}\pi - \phi\right) = 2\rho \cos \phi.$

$\therefore$ *Chord of curvature perpendicular to the radius vector* = $2\rho\,\mathbf{\cos}\,\phi$.

15 Chord of Curvature Parallel to the Axes

(i) Chord of Curvature Parallel to the *x*-axis :

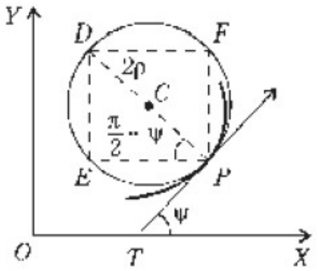

Let *C* be the centre of curvature at the point *P* on any given curve. Suppose a line through *P*, drawn parallel to the axis of *x*, meets the circle of curvature in *E*. Then *PE* is the chord of curvature parallel to *x*-axis.

PD is the diameter of the circle of curvature. We have $PC = 2\rho$ and $\angle PED = 90^{\circ}$.

Obviously $\angle EPD = \frac{1}{2}\pi - \psi.$

Hence from the right angled triangle *PED*, we have

$$PE = 2\rho \cos\left(\frac{1}{2}\pi - \psi\right) = 2\rho \sin \psi.$$

$\therefore$ *Chord of curvature parallel to x-axis* = $\mathbf{2\rho \sin \psi}$.

(ii) Chord of Curvature Parallel to *y*-axis :

Draw a line through *P*, parallel to *y*-axis, to meet the circle of curvature in *F*.

Then PF = chord of curvature parallel to *y*-axis

$$= ED = 2\rho \sin\left(\frac{\pi}{2} - \psi\right) = \mathbf{2\rho \cos \psi}.$$

Illustrative Examples

Example 1 : *Show that the chord of curvature through the pole of the curve* $r^n = a^n \cos n\theta$ *is* $2r/(n+1)$. **(Lucknow 2005; Purvanchal 14)**

Solution : The curve is $r^n = a^n \cos n\theta$. ...(1)

Taking logarithm of both sides, we get $n \log r = n \log a + \log \cos n\theta$.

Differentiating with respect to θ, we get $\frac{n}{r}\frac{dr}{d\theta} = -\frac{n}{\cos n\theta}\sin n\,\theta$

i.e., $$\cot\phi = -\tan n\,\theta = \cot\left(\tfrac{1}{2}\pi + n\,\theta\right)\cdot$$

$$\therefore \quad \phi = \tfrac{1}{2}\pi + n\,\theta.$$

Now $p = r\sin\phi = r\sin\left(\frac{1}{2}\pi + n\,\theta\right) = r\cos n\,\theta.$

Therefore the pedal equation of the given curve is $p = r^{n+1}/a^n$.

$$\therefore \quad \frac{dp}{dr} = \frac{(n+1)\,r^n}{a^n}.$$

Also $$\rho = r\frac{dr}{dp} = \frac{a^n}{(n+1)\,r^{n-1}}.$$

Hence the chord of curvature through the pole

$$= 2\rho\sin\phi = 2\rho\sin\left(\frac{1}{2}\pi + n\,\theta\right) = 2\rho\cos n\,\theta$$

$$= 2\cdot\frac{a^n}{(n+1)\,r^{n-1}}\cdot\frac{r^n}{a^n} = \frac{2r}{(n+1)}\cdot$$

Example 2 : *In the curve $y = a\log\sec(x/a)$, prove that the chord of curvature parallel to the axis of y is of constant length.* **(Rohilkhand 2009; Lucknow 11)**

Solution : Differentiating the equation of the curve with respect to x, we get

$$\frac{dy}{dx} = a\,.\,\frac{1}{\sec(x/a)}\,.\,\sec(x/a)\tan(x/a)\,(1/a) = \tan(x/a).$$

$$\therefore \quad \frac{d^2y}{dx^2} = (1/a)\sec^2(x/a).$$

Chord of curvature parallel to y-axis

$$= 2\rho\cos\psi = \frac{2\rho}{\sec\psi} = \frac{2\rho}{\sqrt{\left\{1 + \left(\frac{dy}{dx}\right)^2\right\}}}$$

$$= 2\cdot\frac{\left\{1 + \left(\frac{dy}{dx}\right)^2\right\}^{3/2}}{\frac{d^2y}{dx^2}}\cdot\frac{1}{\sqrt{\left\{1 + \left(\frac{dy}{dx}\right)^2\right\}}}$$

$$= 2\cdot\frac{\left\{1 + \left(\frac{dy}{dx}\right)^2\right\}}{\frac{d^2y}{dx^2}}$$

$$= 2\cdot\frac{\left(1 + \tan^2\frac{x}{a}\right)}{\frac{1}{a}\sec^2\left(\frac{x}{a}\right)} = 2a, \text{ which is constant.}$$

Comprehensive Exercise 3

1. In the parabola $x^2 = 4ay$, prove that the co-ordinates of the centre of curvature are $\left(-\dfrac{x^3}{4a^2}, 2a + \dfrac{3x^2}{4a}\right)$.

2. In the catenary $y = c \cosh(x/c)$, show that the centre of curvature (α, β) is given by $\alpha = x - y\{(y^2/c^2) - 1)\}^{1/2}, \beta = 2y$.

3. For the curve $a^2 y = x^3$, show that the centre of curvature (α, β) is given by

$$\alpha = \frac{x}{2}\left\{1 - \frac{9x^4}{a^4}\right\}, \quad \beta = \frac{5x^3}{2a^2} + \frac{a^2}{6x}.$$

4. Show that the centre of curvature (α, β) at the point determined by t on the ellipse $x = a\cos t, y = b \sin t$, is given by $\alpha = \dfrac{a^2 - b^2}{a}\cos^3 t, \beta = -\dfrac{a^2 - b^2}{b}\sin^3 t$.

(Lucknow 2007, 10)

Also show that the evolute of the ellipse is $(ax)^{2/3} + (by)^{2/3} = (a^2 - b^2)^{2/3}$.

5. Prove that the centre of curvature (α, β) for the curve $x = 3t, y = t^2 - 6$ is $\alpha = -\frac{4}{3}t^3, \beta = 3t^2 - \frac{3}{2}$.

6. Show that in any curve the chord of curvature perpendicular to the radius vector is $2\rho\sqrt{(r^2 - p^2)}/r$.

7. Show that the chord of curvature through the pole of the equiangular spiral $r = ae^{m\theta}$ is $2r$.

8. Show that the chord of curvature, through the pole, for the cardioid

$$r = a(1 + \cos\theta) \text{ is } \frac{4}{3}r.$$

9. Show that the circle of curvature at the point $(am^2, 2am)$ of the parabola $y^2 = 4ax$ has for its equation

$$x^2 + y^2 - 6am^2 x - 4ax + 4am^3 y - 3a^2 m^4 = 0.$$

10. If C_x and C_y be the chords of curvature parallel to the axes at any point of the curve $y = ae^{x/a}$, prove that $\dfrac{1}{C_x^{\,2}} + \dfrac{1}{C_y^{\,2}} = \dfrac{1}{2aC_x}$.

(Agra 2007; Rohilkhand 07; Purvanchal 07)

Objective Type Questions

Fill in the Blanks:

Fill in the blanks "…… ", so that the following statements are complete and correct.

1. The relation between s and ψ for any curve is called its …… equation.

2. By definition the curvature of the curve at any point P is equal to …… .

3. For a curve $y = f(x)$, we have $\rho =$ …… .

4. Intrinsic formula for the radius of curvature is …… .

5. The radius of curvature at any point of the cycloid

$x = a(t + \sin t), y = a(1 - \cos t)$ is …… .

6. For a curve defined by the equations $x = f(t)$ and $y = \phi(t)$ the radius of curvature is …… .

Multiple Choice Questions:

Indicate the correct answer for each question by writing the corresponding letter from (a), (b), (c) *and* (d).

7. If y-axis is the tangent to the given curve at the origin, then radius of curvature at the origin is equal to

(a) $\lim\limits_{x \to 0} \dfrac{x^2}{2y}$ (b) $\lim\limits_{x \to 0} \dfrac{y^2}{2x}$ (c) $\lim\limits_{x \to 0} \dfrac{x^2}{y}$ (d) $\lim\limits_{x \to 0} \dfrac{y^2}{x}$

(Kumaun 2013)

8. Pedal formula for radius of curvature is

(a) $\dfrac{1}{r}\dfrac{dr}{dp}$ (b) $r\dfrac{dr}{dp}$ (c) $\dfrac{1}{r}\dfrac{dp}{dr}$ (d) $r\dfrac{dp}{dr}$

9. Chord of curvature parallel to y-axis is

(a) $2\rho \sin \phi$ (b) $2\rho \cos \phi$ (c) $2\rho \sin \psi$ (d) $2\rho \cos \psi$

10. Radius of curvature of the curve $p^2 = ar$ is :

(a) p^2/a^2 (b) $2p/a^2$ (c) $2p^3/a^3$ (d) $p^3/2a^2$

(Kumaun 2015)

True or False:

Write 'T' for true and 'F' for false statement.

11. There is no difference between curvature of the circle and circle of curvature. **(Meerut 2003)**

12. The curvature of the curve at any point is equal to the reciprocal of the radius of curvature at that point.

13. The polar formula for radius of curvature is : $\rho = \dfrac{\left\{r^2 + \left(\dfrac{dr}{d\theta}\right)^2\right\}^{3/2}}{r^2 + 2\left(\dfrac{dr}{d\theta}\right)^2 - r\dfrac{d^2 r}{d\theta^2}}.$

14. The tangential polar formula for radius of curvature is $\rho = p + \dfrac{d^2 p}{d\psi^2}.$

15. Pedal formula for radius of curvature is $\rho = r\dfrac{dr}{dp}.$ **(Agra 2006)**

Answers

1. intrinsic. **2.** $\dfrac{d\psi}{ds}.$ **3.** $\dfrac{\left[1 + \left(\dfrac{dy}{dx}\right)^2\right]^{3/2}}{\dfrac{d^2 y}{dx^2}}.$ **4.** $\dfrac{ds}{d\psi}.$

5. $4a \cos \frac{1}{2} t.$ **6.** $\dfrac{(x'^2 + y'^2)^{3/2}}{x'y'' - y'x''}$, where $x'y'' - y'x'' \neq 0$.

7. (b). **8.** (b). **9.** (d). **10.** (c) **11.** *F.*

12. *T.* **13.** *T.* **14.** *T.* **15.** *T.*

)*+,)))- !) % . (/0 + 3#
-4 3 9 * ! 5

)*+,)))- !) % . (/:.: 7; 3#
-4 8 ! ! !4

)*+,)))- !) % . (/:(.2 . 3#
-4 < ! * 4 !

)*+,)))- !) % . (/:(: 0 3#
-4 *#8! # * 4 !

)*+,)))- !) % 2 # = 8 #
2> ? ># @ + ! # A B6 4 ! ((3#

)*+,)))- !) % ! ! # ! !
! 2 5 ! # & *33 ! 3 ! C : D (3#
' 1 %

2

Geometry Background

"Let no one ignorant of Geometry enter" — At entrance to Plato's Academy

2.1 Preliminaries

2.1.1 In the Beginning ...

The Nile's flooding in Egypt used to destroy landmarks and position markers, raising the need for reliable measurements and originating the development of geometry *Geo*-$\Gamma\epsilon\omega$ (for land) *metry* (for measure-$\mu\epsilon\tau\rho\iota\alpha$).[10] The Rhind Papyrus, which is in the British Museum, was written by an Egyptian priest named Ahmes about 1,800 B.C.E. It treats the measurement of planar regions and solids. It contains the remarkable approximation of $\pi \approx (\frac{16}{9})^2 = 3.1604$. The contemporary Babylonians, as well as Chinese, used 3 for the value of π. The value of 3 for π is also mentioned in the Hebrew scriptures (see I Kings 7:23). A rabbi called Nehemia (p. 5 [74]) proposed the value of $\frac{22}{7}$ for π, but apparently it was not accepted. By the way, only in 1789 was it first proved by J. Lambert that π is an irrational number and in 1882, F. Lindemann proved that π is transcendental (i.e., it is not the root of any algebraic equation with rational coefficients). For more enjoyable folklore on π the reader is referred to [12].

It is fair to consider the Egyptians more as surveyors, who used a collection of geometrical formulas, mostly approximations, rather than geometers. For proper attribution, *deductive* geometry was accomplished by the Greeks, probably beginning with Thales of Miletus ($\approx$ 640–545 B.C.E.), who was familiar with Egyptian

---

[10] This story was reported by Proclus in the fifth century C.E., who wrote the early history of geometry.

A. Inselberg, *Parallel Coordinates*, DOI 10.1007/978-0-387-68628-8_2,
© Springer Science+Business Media, LLC 2009

and Babylonian mathematics. Pythagoras, Hippocrates (not the physician of the same name), Menaechmus, Archimedes, Apollonius, Plato, Archytas, Eudoxus and Theaetetus continued the development until it was codified by Euclid in his *Elements* $\Sigma\tau o\iota\chi\epsilon\iota\alpha$ ($\approx$ 300 B.C.E.). Entirely novel, and characteristic of Greek mathematics, is the *axiomatic method* whereby theorems are established by proof. Particularly relevant to computer science are the *constructions* of Euclidean geometry. They consist of a finite set of well-defined operations and provide an unambiguous procedure that terminates for all inputs! Truly the first *algorithms ... and with proofs*. The constructions of Euclidean geometry involved ruler and compass. A remarkable discovery was made in 1672 by the Danish geometer G. Mohr (1640–1697) and independently in 1797 by the Italian L. Mascheroni (1750–1800).[11] They showed that *all* of these constructions can be accomplished by *compass alone!* [38]. The proofs were, in the true spirit of computer science, by simulation. It was demonstrated that any operation involving a ruler can be replaced by a finite number of compass operations; a real equivalence proof like those in automata theory. Alternatively, J. Steiner and J.V. Poncelet showed that all constructions in Euclidean geometry can be accomplished with a straightedge when in addition, a single circle and its center are also given. Continuing in this vein, it is natural to ask what is left if the compass is discarded and only the ruler is left. What remains is *projective geometry*, an elegant branch of geometry simpler than Euclid's but rich and interesting in its own right. It deals with points, lines, planes, and spaces where distance between two points or angles between two lines are *not* measured and there are *no parallel lines* in the plane! For a modern treatment of projective geometry see [154].

Exercises

1. To get a feel for the Mohr–Mascheroni result, using compass *only* write (and prove) an algorithm for:
    (a) constructing 3 collinear points,
    (b) given a segment $\overline{OA}$ construct a point B such that $\overline{OB}$ is twice as long as $\overline{OA}$,
    (c) bisecting a segment,
    (d) trisecting a segment,
    (e) given a circle C and two points determining a line l, construct the point of intersection $C \cap l$ (be careful here, the line l is *not given*, but only two points on the line l are. This is what makes the problem *hard*).
    (f) constructing the point of intersection between two lines (*hard*).

---

[11] For a short exposition on the *geometry of compasses* see [145, (pp. 23–25)] and [36, (pp. 145–151)], quoted in [38].

2. Given an angle A, provide a construction for finding an angle = 1/2 A using the Steiner–Poncelet constraint, i.e., with a straightedge, a single circle, and its center.

   To recall and rejuvenate some Euclidean geometry skills try some of the following[12]:

3. Provide a Euclidean geometry construction (i.e., using circle and ruler only) that transforms a rectangle into a square of equal area.
4. One of the three famous unsolvable problems in Euclidean geometry is "squaring the circle." Given a square with area **A** construct, using ruler and compass only, a square having area **2A**. What other shapes can this be generalized to?
5. Prove that a triangle ABC is isosceles if the bisectors AD and BE of the two angles A and B are equal. Here D and E are the points of intersection of the angle bisectors with the sides BC and AC respectively. This is a deceptively *hard* problem.

2.2 ♣ *FT-1* Why Projective Geometry?

We begin the formal aspects with an excursion into projective geometry, but why? There are two important concepts and tools that we will need and that are also widely used in many other fields and applications. They are:

- *duality* and
- *homogeneous coordinates* (see [146] for applications of homogeneous coordinates to computer graphics).

Both of these are projective concepts and can be properly understood only in this context. The experience is quite broadening and will get us "unglued" from the narrow mindset of Euclidean geometry. For example, entirely new notions, such as finite geometries, arise naturally. Projective geometry was initially motivated by the fine arts. The Italian architect Brunelleschi in 1425 first proposed a geometrical theory of perspective that was formalized by Alberti a few years later. It included the idea of "points at infinity" and influenced art in a profound way, as shown by contrasting prior "flat" paintings such as Martini's *Annunciation* with those where "depth" is perceived, such as Leonardo's *Last Supper* and Raphael's *Athens*, shown in the chapter on projective geometry by Morris Kline [140, vol. **1**, 622–641]. In the absence of photography, realism in drawings had not only artistic but also great practical value.

---

[12] See also the delightful short paper by Hudson [90].

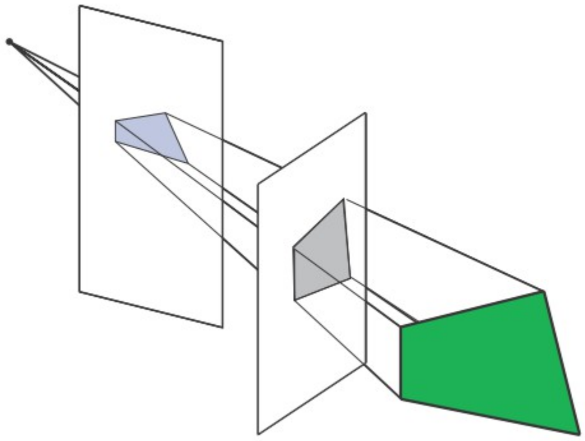

Figure 2.1. Projections and Sections.

2.2.1 Projective Plane Model

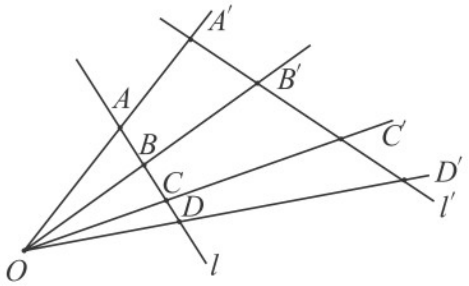

Figure 2.2. Points and lines in Perspective.

Projective geometry deals with properties that are invariant under projections and sections, see Fig. 2.1, such as "incidence" and "collinearity." A beautiful projective theorem, discovered by the French architect Desargue in 1640, states that two triangles, ABC and $A'B'C'$, are in perspective from a point O if and only if they are in perspective from a line *TSR*; see Fig. 2.4 (left) in the Euclidean plane and Fig. 2.4 (right), where O is a "point at infinity" in the vertical direction. Note the role of *duality* in the interchange between the words *point* and *line*. Of course, we have not explained what is meant by "perspective"; see Fig. 2.2 for an indication, but we are on the way.

It will help our intuition to have a model for visualizing the projective space, and only afterward do we present the more formal axiomatic treatment. Our "building blocks," otherwise called *primitives*, are concepts and objects with which we are already familiar. The notation commonly used in geometry appears in the list below; points and lines are denoted by capital and lowercase Latin letters (preferably in *italics*) respectively, while planes are denoted by lowercase Greek letters.

Primitives

- *points* - P, Q, A
- *lines* - ℓ, *r*, *s*
- *planes* - π, α, β
- *incidence* relation - P is *incident* to ℓ, or P is *on* ℓ.

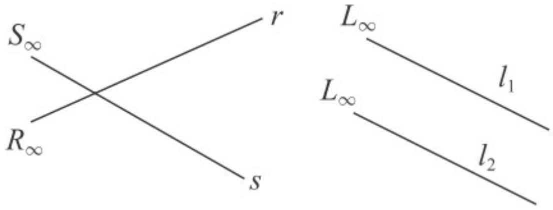

Figure 2.3. Ideal points and augmented lines.

Of course, incidence and containment are not equivalent, since P is *incident* to ℓ if and only if ℓ is *incident* to P; $P \in \ell$ is true and $\ell \in P$ is not.

We start with the Euclidean space $\mathbb{R}^3$ and add something new. Let $\pi \subset \mathbb{R}^3$ be a plane. To each line $\ell \subset \pi$ we associate an *ideal point* ("point at infinity") L_∞ such that distinct intersecting lines have distinct ideal points while parallel lines have the *same* ideal point. Effectively, then, ideal points signify direction. Since no

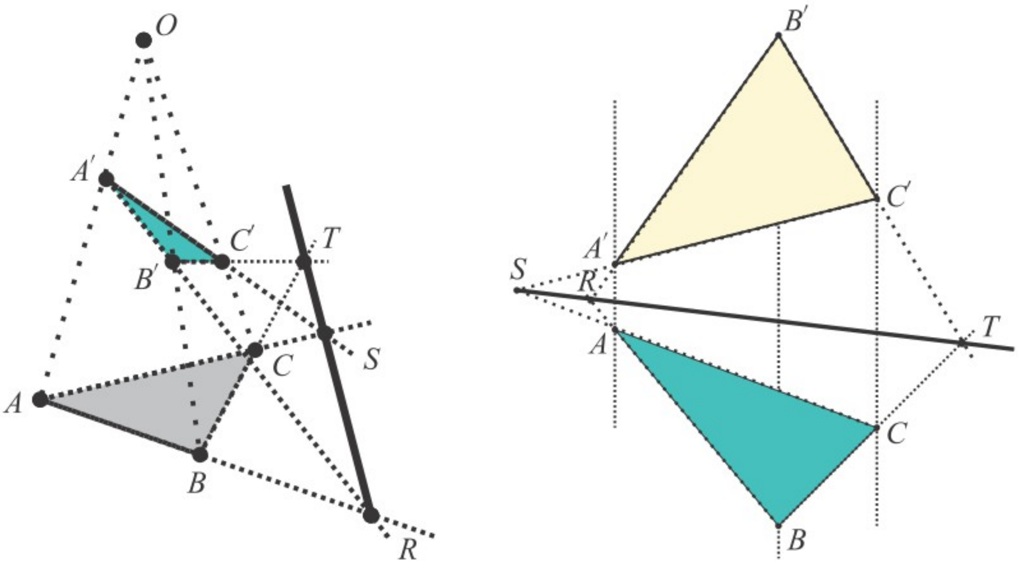

Figure 2.4. Desargue's theorem. The three points T, S, R, $T = B'C' \cap BC$, $S = A'C' \cap AC$, $R = A'B' \cap AB$, are always collinear for when the triangles are in perspective with respect to an ordinary point (left) or an ideal point ("point at infinity") (right).

distinction is made between the "+" and "−" directions for each line ℓ, one and not two L_∞ need to be appended.

On occasion, the notation P_m^∞ is used for ideal points, with m being the slope of the corresponding direction. The points of $\mathbb{R}^3$ are called *regular* to distinguish them from the "*ideal*" ones. Proceeding, we call the pair (ℓ, L_∞) an *augmented line*. Then for $r, s \subset \pi$,

$$(r, R_\infty) \cap (s, S_\infty) = \begin{cases} P \text{ ordinary point} & \Longleftrightarrow \text{r not} \parallel \text{s}, \\ R_\infty & \Longleftrightarrow \text{r} \parallel \text{s}. \end{cases}$$

For a given plane $\pi \subset \mathbb{R}^3$, let $p_\infty = \{L_\infty | \ell \subset \pi, (\ell, L_\infty)\}$ be the collection of all ideal points (directions) on π. We will see that p_∞ satisfies the axioms for a line, and for this reason it is called an *ideal line*. The pair (π, p_∞) is called an *augmented plane* for, as will be seen, it satisfies the axioms for a plane. Proceeding, the ideal lines for each plane $\pi \subset \mathbb{R}^3$ are collected and denoted by $\pi_\infty = \{p_\infty | \pi \subset \mathbb{R}^3, (\pi, p_\infty)\}$; we refer to π_∞ as the *ideal plane* and the pair $(\mathbb{R}^3, \pi_\infty)$ as the *augmented space*. All in all, then, "points" are either ordinary or ideal, "lines" are either augmented or ideal, and "planes" are either augmented planes or the ideal plane.

Hilbert [85] proposed an elegant model of the projective plane whose construction (actually a modified version) is shown in Fig. 2.6. A hemisphere S together with the disk D capping it is placed in a Euclidean plane π, with its south pole tangent to the origin. A *stereographic* projection is established such that a point $P \in \pi$ is connected to the center C of S. The point $P' = CP \cap S$ is the image of P. All points of π (i.e., "ordinary points") are mapped uniquely onto the surface points of S. As the point $P = (a_1, a_2)$ goes farther and farther away from the origin in the constant direction having slope $m = a_2/a_1$, it approaches the ideal point[13] P_m^∞. Then the line CP approaches the line parallel to π that includes the diameter d_m of the disk D having slope m. This establishes a one-to-one correspondence $d_m \longleftrightarrow P_m^\infty$. The points on the hemisphere's surface are the images of the ordinary points, while the diameters of D are the images of the ideal points providing the image of all of $\mathbb{P}^2$ as described above.[14] It is a very helpful visual model. For example, a line $\ell \subset \pi$ is mapped into a great semicircle (see exercise below), and the augmented line $(l, \mathbb{P}_m^\infty)$ includes the diameter d_m. Parallel augmented lines, as shown in Fig. 2.7, "intersect" at their common ideal point. Two points on a projective line specify *two*

---

[13] The subscript m, in the notation of an ideal point, indicates the corresponding direction, and it is included only when the direction is relevant.

[14] We use homogeneous coordinates only with real values so there is no need to use $\mathbb{RP}^2$ as seen elsewhere, the $\mathbb{R}$ distinguishing the *real* projective plane from others.

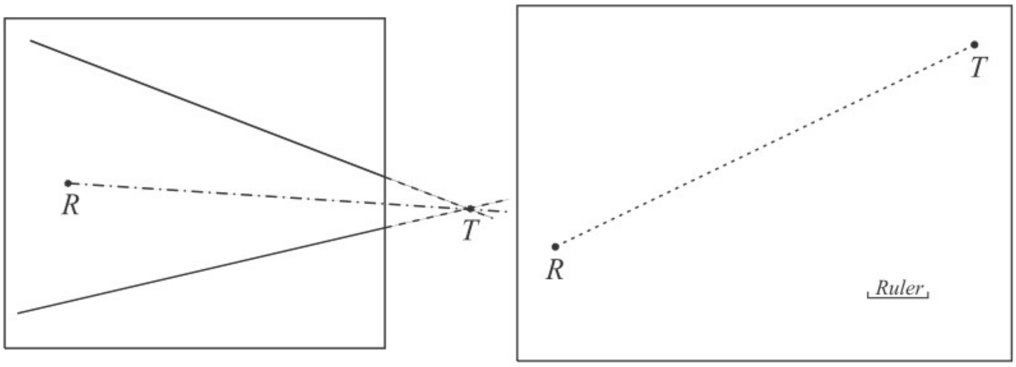

Figure 2.5. On the left, figure for Exercise 4; and on the right, figure for Exercise 5.

intervals; one containing the ideal point and its complement being the "ordinary" interval of the Euclidean line.

A very basic though often unfamiliar property of $\mathbb{P}^2$ is that it is not *orientable*. That is, it is not possible to assign a consistent sense of left and right ([18] p. 372). This matter arises in various topics in our development, and we shall learn how to cope with it when the need arises.

Exercises

1. Show that lines are mapped into great semicircles by stereographic projection.
2. Show that stereographic projection preserves angles (i.e., is isogonal)
3. How are circles mapped by stereographic projection?
4. On a piece of paper, as shown in Fig. 2.5 (left), there is a point R and portions of two lines intersecting at some point T outside the paper. Using only a straightedge (ruler), draw the line joining the points R and T (Hint: Use Desargues's theorem). From [13, vol. I, p. 119].
5. Given two points R and T and a short ruler, as shown in Fig. 2.5 (right), not long enough to reach both points, draw the line joining R and T (Hints: Use the result of the previous exercise and slide the ruler along a line segment.) From [13, vol. I, p. 119].

2.2.2 Axioms for Projective Geometry

The time has come to describe projective 3-space more formally in terms of the axioms, i.e., the properties of the primitives: points, lines, planes, and the *incidence* (or "*on*") relation.

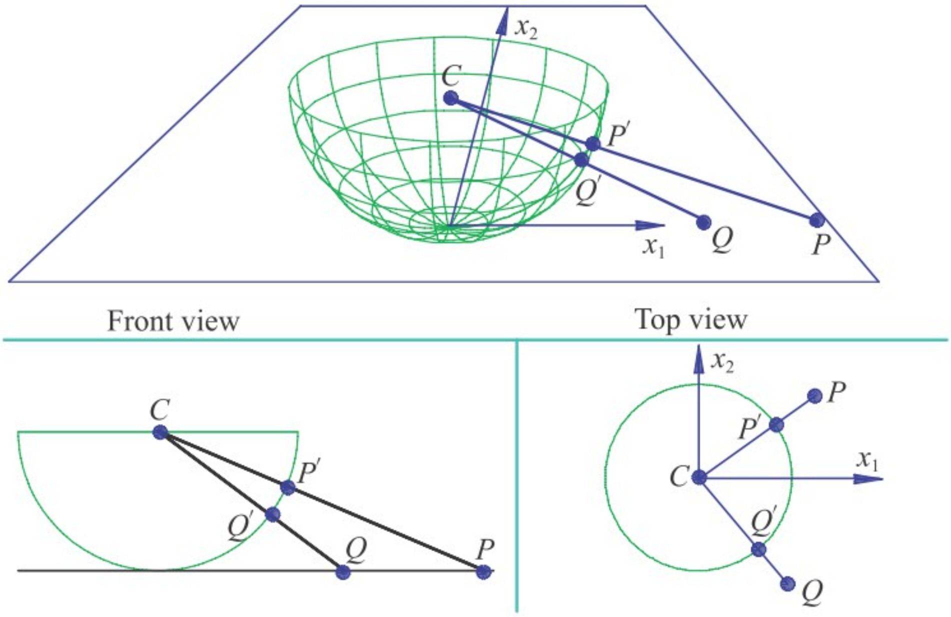
x_2
C
P'
Q'
x_1
Q
P
Front view
Top view
C
P'
Q'
Q
P
x_2
P
P'
C
x_1
Q'
Q

Figure 2.6. Model of the projective plane.

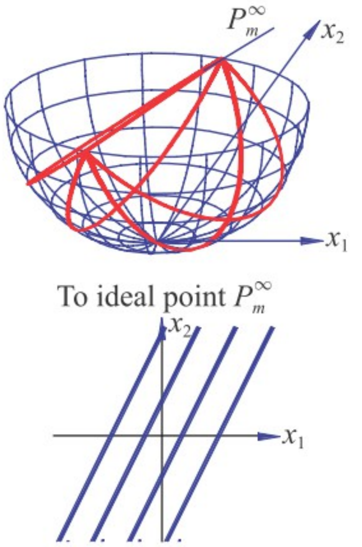
P_m^∞
x_2
x_1
To ideal point P_m^∞
x_2
x_1

Figure 2.7. Parallel lines on the projective plane model.

Axioms for 2-D and 3-D

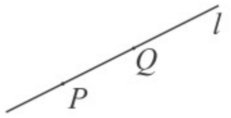

1. (2-D) Two distinct points are *on* (or incident to) a unique line. That is, if $P \neq Q$ then $P, Q \in \ell_1$ and $P, Q \in \ell_2 \Rightarrow \ell_1 = \ell_2$.

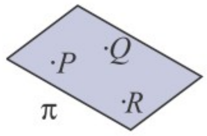

2. (3-D) Three distinct noncollinear points are *on* a unique plane. That is, if $P \neq Q, Q \neq R, P \neq R$, and there is no line $\ell \ni \{P, Q, R\} \subset \ell$, then $\{P, Q, R\} \subset \pi_i$, $i = 1, 2 \Rightarrow \pi_1 = \pi_2$.

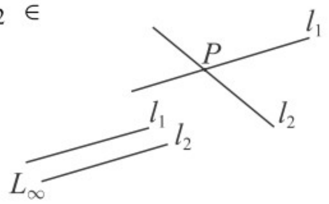

3. (* 2-D) Two distinct coplanar lines are *on* (i.e., intersect or determine) a unique point. With the "ideal" elements we always have $\ell_1 \cap l_2 \neq \emptyset$ and $P_1, P_2 \in \ell_1 \cap \ell_2 \Rightarrow P_1 = P_2$.

This axiom and the remaining axioms are marked with an asterisk $(*)$, for they differ from their corresponding Euclidean cousins. Specifically, this axiom corresponds to Euclid's famous fifth postulate, which postulated the existence of parallel lines.

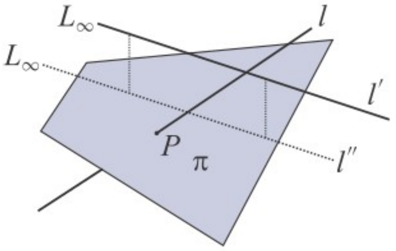

4. (* 3-D) A plane π and a line ℓ not contained in π are *on* (i.e., intersect at) a unique point $\ell \cap \pi \neq \emptyset$ always and $P_1, P_2 \in \ell \cap \pi \Rightarrow P_1 = P_2$.

5. (* 3-D) Two distinct planes are *on* (i.e., intersect in) a unique line $\pi_1 \cap \pi_2 \neq \emptyset$ and $\pi_1 \cap \pi_2 = \ell$, where ℓ is unique.

It is not difficult to prove that the Euclidean plane together with the ideal and augmented elements satisfies the projective plane axioms. Let us assume that ideal points and augmented lines satisfy the axioms (see exercises) and study an example.

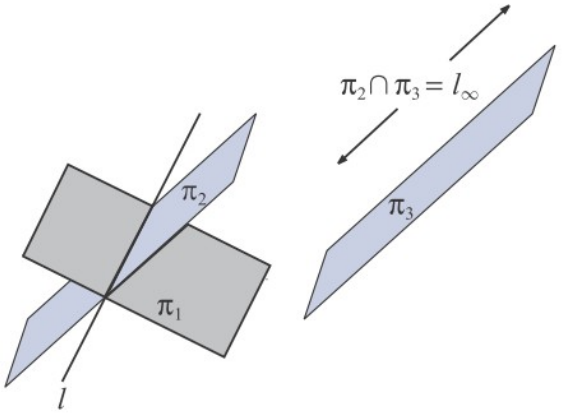

Theorem 2.2.1. *The ideal line p_∞ of an augmented plane (π, p_∞) is a line.*

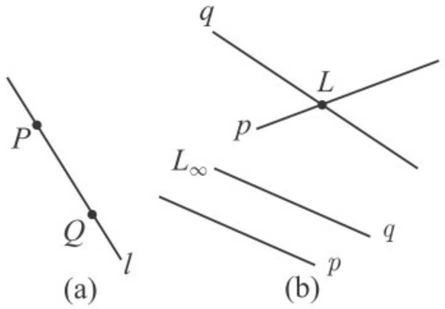

(a) (b)

Figure 2.8. Axiom 1 is illustrated in (a) and its dual axiom (3) in (b).

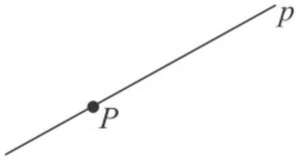

Figure 2.9. Self-dual: *point on line* and *line on point*.

Proof. We need to show that p_∞ satisfies the axioms for lines, namely axioms 3 and 4.

Let (ℓ, L_∞) be any augmented line on (π, p_∞). Since p_∞ is the collection of all ideal points on the augmented plane, $(\ell, L_\infty) \cap p_\infty = L_\infty$, so axiom 3 is satisfied.

Let (π', p'_∞) be an augmented plane not containing p_∞; therefore π' is not $\|$ to π. Hence $\pi \cap \pi' = \ell \Rightarrow (\pi, p_\infty) \cap (\pi', p'_\infty) = (\ell, L_\infty)$, where L_∞ is the ideal point associated with ℓ. ■

Two excellent references for projective geometry are [7] and [39].

Exercises

1. Show that the ideal points satisfy the axioms of points.
2. Show that augmented lines satisfy the axioms of lines.
3. Show that augmented planes satisfy the axioms of planes.
4. Show that the ideal plane π_∞ satisfies the axioms of a plane.

2.2.3 Principle of Duality

The axioms yield a principle not found in Euclidean geometry. In 2-D there is a point ↔ line duality stemming from (axiom 1) two *points* are on a *line* ℓ and (axiom 3) two *lines* are on a *point* L. As a result, planar figures can be "dualized" by switching the words point and line. For example, three distinct noncollinear (i.e., not on the same *line*) *points P, Q, and R* determine a triangle. Dually, three distinct nonintersecting (i.e., not on the same *point*) *lines p, q, r* also determine a triangle. That is, triangles are self-dual. Note also the projective cases in Fig. 2.11: (left)

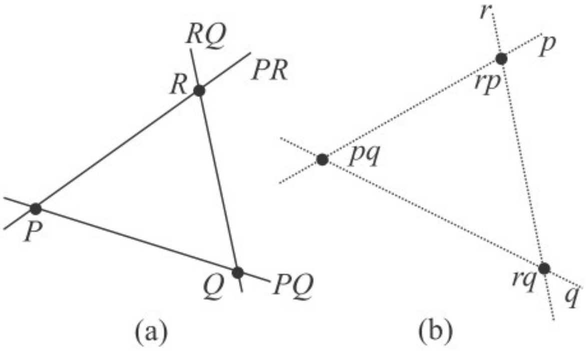

Figure 2.10. Self-dual 2: Triangle (a) and "triline" in (b).

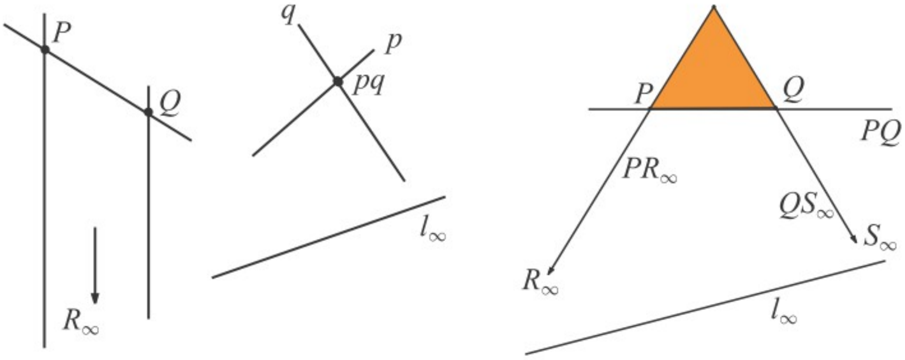

Figure 2.11. Projective triangles (left) and quadrangle (right).

triangles and (right) the four points P, Q, R_∞, S_∞, which form a quadrilateral ("quadrangle") with edges $PQ, PR_\infty, QS_\infty, \ell_\infty$.

The 3-D duality is point ↔ plane and results from axiom 2, whose dual is that three noncollinear *planes* π_1, π_2, π_3 are on a point P; Fig. 2.12. By the way, this is the reason why point ↔ line dualities do not generalize to higher dimensions. Other examples of the point ↔ plane duality are that two *planes* π_1, π_2 are on a line, whose dual is axiom 1, and two *points* P_1, P_2 are on a line. Also axiom 4: a *plane* π and a line l not on π are on (determine) a *point* is dualized to a *point* P and a line l not on P are on (determine) a *plane*.

The implications of duality are far deeper than those just indicated. For if a (projective) theorem involves points and lines in 2-D or points and planes in 3-D, its dual is also true. We illustrate this with the earliest known projective theorem,

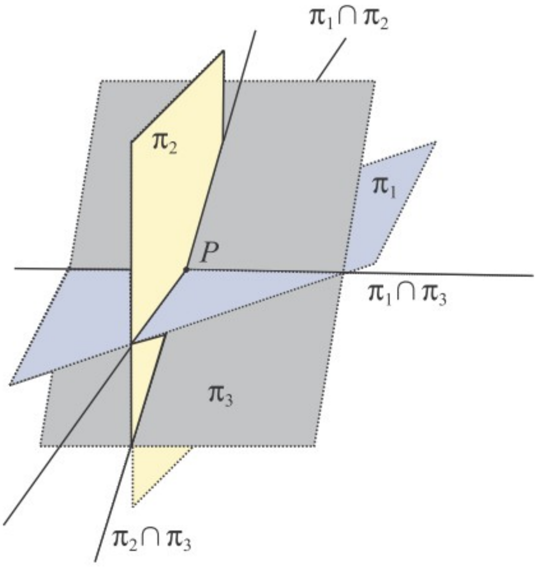
$\pi_1 \cap \pi_2$
π_2
π_1
P
$\pi_1 \cap \pi_3$
π_3
$\pi_2 \cap \pi_3$

Figure 2.12. Three noncollinear *planes on* a *point*; dual of axiom 2.

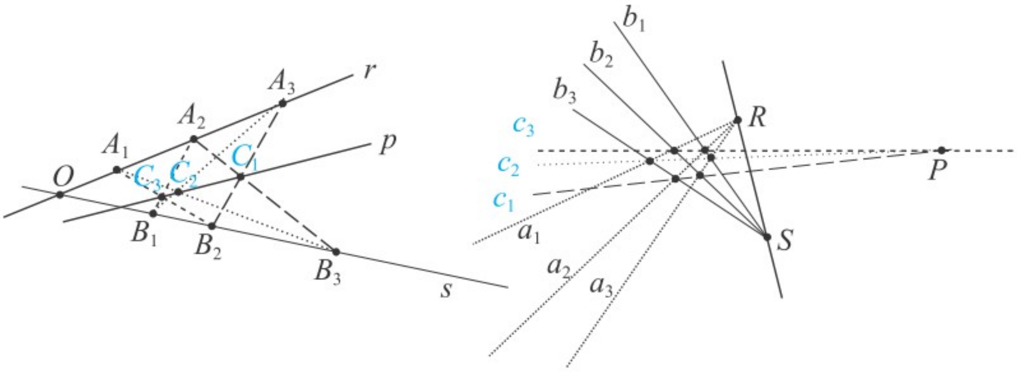
r
A_3
A_2
A_1
O
C_3
C_2
C_1
p
B_1
B_2
B_3
s
b_1
b_2
b_3
c_3
c_2
c_1
R
P
a_1
a_2
a_3
S

Figure 2.13. Pappus's theorem (left) and its dual (right).

proved by Pappus in the third century C.E. Using the notation $\cdot$, instead of $\cap$, for intersection Pappus's theorem states that for any distinct *points* A_i on a *line* r and any distinct *points* B_i on a *line* s, where $i = 1, 2, 3$, the *points* C_i, where $C_i = A_j B_k \cdot A_k B_j$ for $j \neq k \neq i$, are always on a *line* p. The dual theorem is obtained simply by interchanging the *words* point ↔ line. In our notation it is equivalent to switching lowercase ↔ capital letters as follows: for any distinct *lines* a_i on a *point* R and any distinct *lines* b_i on a *point* S, where $i = 1, 2, 3$, the *lines* c_i, where $c_i = a_j b_k \cdot a_k b_j$ for $j \neq k \neq i$, are always on a *point* P. No further proof is needed!

Fun and Games with the ILM

Open the ILM: set *Chapter = 2D Point Line duality*.
The display shows two sets of axes.
On the right, click to draw a point and see the corresponding line on the left
Move the point and watch the line.

Next set *Subject = Point ↔ Line*.
Draw a point on the left and see the corresponding line on the right.

2.3 ** Finite Geometries

Let a *finite* set of points and a nonempty collection of point subsets called *lines* satisfy the following axioms:

1. Any two distinct points A and B are *on* a unique line ℓ.
2. If p and q are distinct lines, there is a unique point *on* both.
3. There are at least three distinct points *on* any line.
4. Not all points are *on* the same line.

Such a set is called a *finite geometry*. It exemplifies the stark contrast with Euclidean geometry. *Point, line*, and the *on* relation are primitives specified by the above axioms only. Additional properties can be introduced to enrich the geometry.

2.3.1 Seven-Point Geometry

From the third axiom there are $n \geq 3$ points on a line. The smallest geometry, with $n = 3$ points on each line and a total of $n^2 - n + 1 = 7$ (see Exercise 3 below) points or lines is illustrated in Fig. 2.14. The lines are $a = [A, D, B]$, $b = [A, E, C]$, $c = [C, F, B]$, $d = [A, G, F]$, $e = [B, G, E]$, $f = [C, G, D]$, $g = [D, E, F]$. Observe that the projective plane axioms are satisfied (i.e., two distinct points are on a unique line and vice versa) for ∀ lines $x \in \{a, \dots, g\}$

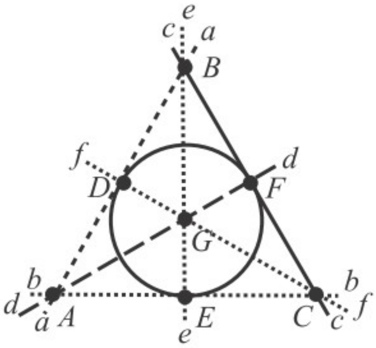
e
c
a
B
f
d
D
F
G
b
b
d
a
A
E
C
c
f
e

Figure 2.14. A "picture" of the 7-point geometry.

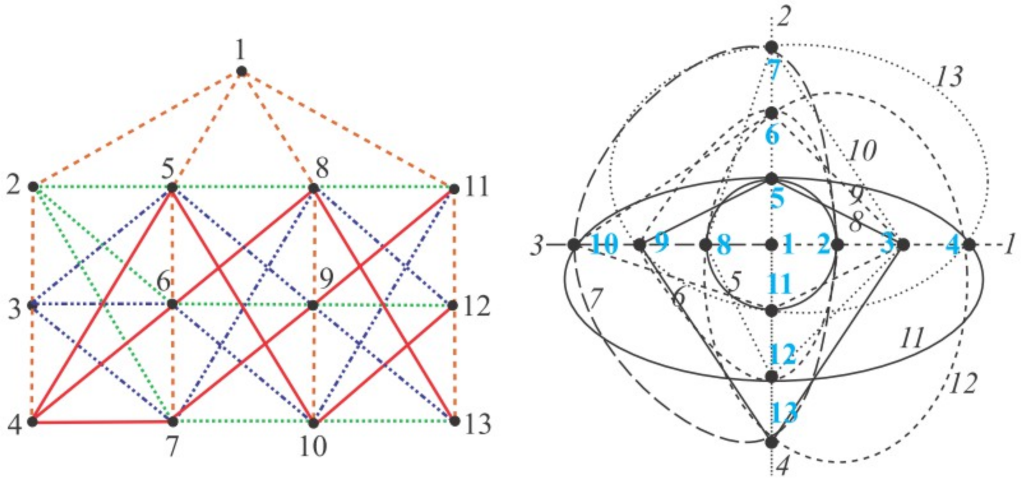
1
2
5
8
11
3
6
9
12
4
7
10
13
2
13
7
6
10
5
9
8
3
10
9
8
1
2
3
4
1
5
11
7
6
11
12
12
13
4

Figure 2.15. "Pictures" for 13-point geometry. There are four "lines" on each point and vice versa. On the right, the points are indicated in **bold numbers** and lines in *italics*.

and ∀ points $X \in \{A, \dots G\}$. The lines/curves joining the • in the picture have no significance other than showing the association of the points belonging to a particular line.

↓ *lines*	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	← **points**
1	**1**	**2**	**3**	**4**										
2	**1**				**5**	**6**	**7**							
3	**1**							**8**	**9**	**10**				
4	**1**										**11**	**12**	**13**	
5		**2**			**5**			**8**			**11**			
6		**2**				**6**			**9**			**12**		
7		**2**					**7**			**10**			**13**	
8			**3**		**5**				**9**				**13**	
9			**3**			**6**				**10**	**11**			
10			**3**				**7**	**8**				**12**		
11				**4**	**5**					**10**		**12**		
12				**4**		**6**		**8**					**13**	
13				**4**			**7**		**9**		**11**			

Figure 2.16. Incidence matrix for the 13-point geometry.

2.3.2 Finite Geometry with 13 Points

Finite geometries can be pictured in different ways, as in Figure 2.15, and specified by an incidence matrix for points and lines as for the 13-point geometry. Here, it is more convenient to denote the points and lines by numbers rather than letters. Those interested in pursuing the fascinating subject of finite geometries are referred to [116] and [117].

Exercises

1. Prove the 3-D dual shown in Fig. 2.12, i.e., that three noncollinear planes are on a point, using only the projective 3-D space axioms.
2. The second axiom may be weakened to 2a: "If p and q are distinct lines, there is *at least* one point on *on* both lines." Prove that with axiom 2 replaced by 2a above, the resulting set of points and lines also satisfies the projective plane axioms.
3. Prove that in a finite geometry with n points on any line there are $n^2 - n + 1$ points (or lines) altogether.
4. Construct diagrams for the geometries with $n = 5$ and 6 points and try to generalize. This is hard: you may want to look up a topic in combinatorics called "Latin squares."

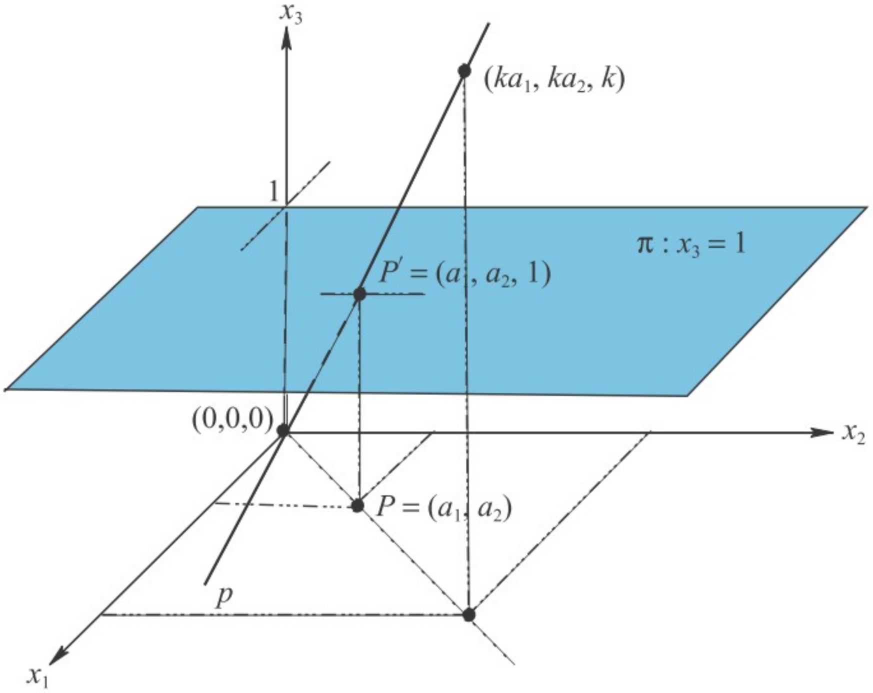

Figure 2.17. Homogeneous coordinates. Correspondence between points P in the x_1x_2-plane and lines p through the origin in 3-D.

2.4 ♣ *FT-2* Analytic Projective Geometry

Even though there is no notion of distance in projective space, it is possible to introduce a very useful coordinate system. The geometric development here is for the plane and can be generalized to projective N-space $\mathbb{P}^N$.

We consider the x_1x_2 plane in $\mathbb{R}^3$ with Cartesian coordinates as shown in Fig. 2.17 and the plane $\pi : x_3 = 1$. The point $P = (a_1, a_2)$ in the x_1x_2-plane is identified with the point $P' = (a_1, a_2, 1)$ in the plane π. The line p through the points $(0, 0, 0)$ and P' is constructed and is described by the equations

$$p : \begin{cases} x_2 = \frac{a_2}{a_1} x_1, \\ x_3 = \frac{1}{a_2} x_2. \end{cases}$$

In this way, the point-to-line correspondence

$$
\begin{aligned}
&P = (a_1, a_2) \in x_1x_2\text{-}plane \\
&\quad \longleftrightarrow\ p \subset x_1x_2x_3\text{-}space,\ p\ \ on\ \ (a_1, a_2, 1), (0, 0, 0),
\end{aligned}
$$

is established. Since p can be identified by any one of its points (ka_1, ka_2, k) for $k \neq 0$, the correspondence can be alternatively written as

$$P : (a_1, a_2) \longleftrightarrow (ka_1, ka_2, k),\ \ 0 \neq k \in \mathbb{R}.$$

That is, (k_1a_1, k_1a_2, k_1) and any other multiple (k_2a_1, k_2a_2, k_2) for $k_1 \neq k_2$ represent the *same* point, since they both lie on the *same line* p. These triples are the *point coordinates* of P.

Let us return to Fig. 2.17 and imagine the point P moving farther away from the origin but always in the direction with slope $m = \frac{a_2}{a_1}$ approaching the ideal point in this direction. The point coordinates of P are $(a_1, a_2, 1)$ or $(1, m, \frac{1}{a_1})$, and as $a_1 \to \infty$ the ideal point is represented by the triples $(ka_1, ka_2, 0)$ for $k \neq 0$. The important realization is that the 0 in the third component identifies ideal points. This system is called *homogeneous coordinates*, allowing the representation both ordinary *and* ideal points, something that cannot be done with Cartesian coordinates; it has other nice properties as well.

Appealing to duality, a *line* x can be represented as an ordered triple of real numbers that is distinguished from a triple representing *points* using capitals for the coordinates themselves and "[ ]" instead of "( )" as

$$x : [X_1, X_2, X_3] \neq [0, 0, 0],$$

again with the convention that $[kX_1, kX_2, kX_3]$, $\forall k \neq 0$, represents the same line. These triples are called *line coordinates*. So $(1, 2, 3)$ represents a point and $[1, 2, 3]$ a line. Actually, this is quite reasonable, for consider a line described in Cartesian coordinates by

$$A_1x_1 + A_2x_2 + A_3 = 0. \tag{2.1}$$

For $A_3 \neq 0$ the same line is also given by

$$\frac{A_1}{A_3}x_1 + \frac{A_2}{A_3}x_2 + 1 = 0. \tag{2.2}$$

In either case, the three constants A_1, A_2, A_3 completely specify the line. So taking the triple $[A_1, A_2, A_3]$ with the understanding that A_i is the coefficient of x_i for

$i = 1, 2$ and A_3 is the constant completely describes the line in (2.1). Equivalently, the triple $[\frac{A_1}{A_3}, \frac{A_2}{A_3}, 1]$ describes the line as given by (2.2), and hence these two triples and their multiples identify the same line.

There remains to define the *on* relation. We say that a point $X = (x_1, x_2, x_3)$ is *on* the line $x = [X_1, X_2, X_3]$, and dually the line x is *on* the point if and only if

$$X \cdot x = x \cdot X = x_1 X_1 + x_2 X_2 + x_3 X_3 = 0.$$

Consider now the three points $Y = (y_1, y_2, y_3), Z = (z_1, z_2, z_3), W = (w_1, w_2, w_3)$, all on the line $x = [X_1, X_2, X_3]$. Then

$$\begin{aligned} y_1 X_1 + y_2 X_2 + y_3 X_3 &= 0, \\ z_1 X_1 + z_2 X_2 + z_3 X_3 &= 0, \\ w_1 X_1 + w_2 X_2 + w_3 X_3 &= 0. \end{aligned}$$

This system of linear homogeneous equations has a nontrivial solution $[X_1, X_2, X_3]$ if and only if the deteterminant of the coefficients is zero. That is, Y, Z, W are on a line x if and only if

$$\begin{vmatrix} y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \\ w_1 & w_2 & w_3 \end{vmatrix} = 0. \tag{2.3}$$

This condition provides the line coordinates of x as

$$\left[\begin{vmatrix} y_2 & y_3 \\ z_2 & z_3 \end{vmatrix}, \begin{vmatrix} y_3 & y_1 \\ z_3 & z_1 \end{vmatrix}, \begin{vmatrix} y_1 & y_2 \\ z_1 & z_2 \end{vmatrix} \right]. \tag{2.4}$$

A direct consequence is that the points X, Y, and $\lambda X + \mu Y$ for $\lambda, \mu \in \mathbb{R}$ are collinear.

What are the line coordinates of the ideal line? We can find out by substituting two distinct ideal points in (2.4), i.e., $z_3 = y_3 = 0$. This yields $[0, 0, y_1 z_2 - y_2 z_1]$ or $[0, 0, 1]$. So far, it is all falling into place: we have a coordinate system that provides coordinates for every regular point (from which its Cartesian coordinates can be recovered) and *also* coordinates for the ideal elements (points and line). In addition, homogeneous coordinates can be extended to projective N-space and finite geometries. ♣ *FT-2e*

Consider now the 7-point geometry (i.e., $n = 3$) described earlier and take the algebra J_2, the integers modulo 2, with elements $\{1, 0\}$. The binary operations $+$ and $\times$ are defined by

$$\begin{array}{c|cc} + & 0 & 1 \\ \hline 0 & 0 & 1 \\ 1 & 1 & 0 \end{array} \qquad \begin{array}{c|cc} \times & 0 & 1 \\ \hline 0 & 0 & 0 \\ 1 & 0 & 1 \end{array}.$$

Now we collect all triples (x_1, x_2, x_3) except $(0, 0, 0)$ and label them:

$$A = (0, 0, 1), \quad B = (0, 1, 0), \quad C = (1, 0, 0),$$
$$D = (0, 1, 1), \quad E = (1, 0, 1), \quad F = (1, 1, 0), \quad G = (1, 1, 1).$$

A pair of points is chosen specifying a line $x = [X_1, X_2, X_3]$ whose coordinates are computed from (2.4) for $X_i \in J_2$, yielding

$$a = [1, 0, 0] \quad b = [0, 1, 0] \quad c = [0, 0, 1]$$
$$d = [1, 1, 0] \quad e = [1, 0, 1] \quad f = [0, 1, 1] \quad g = [1, 1, 1].$$

Note that each point is on two lines; for example A is *on* the lines having $X_3 = 0$. Similarly, other finite geometries can be coordinatized using the appropriate algebra J_p of integers modulo p, that is, the set of integers $\{1, 2, \dots, p-1\}$ with addition and multiplication analogously defined (see a book on modern algebra). To clarify, let us use J_3 to coordinatize the 13-point geometry. Excluding $(0, 0, 0)$, there are 26 triples (a_1, a_2, a_3), $a_i \in J_3$, with the multiples $k(a_1, a_2, a_3)$, $k \neq 0 \in J_3$, also representing the same point. The only multiple of interest is $2(a_1, a_2, a_3)$, which equals (a_1, a_2, a_3) for $a_i \in J_3$, yielding 13 distinct triples. Similarly, with J_4 there 63 nonzero triples whose multiples by 2 or 3 are the same in that algebra, yielding 21 distinct triples, which can therefore serve as the homogeneous coordinates of the 21-point geometry (Exercise 3 below).

2.4.1 Homogeneous Coordinates on a Line

Let us specialize homogeneous coordinates to a line o, which for convenience we take as being on the two points

$$O = (0, 1, 1), \quad U = (1, 1, 1).$$

The coordinates of a point $X = (x_1, x_2, x_2) \in o$ together with those of O and U satisfy the collinearity condition of (2.3); therefore $x_2 = x_3$, so $X = (x_1/x_2, 1, 1)$, with $X = (x_1, 0, 0)$ being the line's ideal point if $x_2 = 0$. Since there is no variation in the third component, the coordinates of regular points can be simplified to $O = (0, 1)$, $U = (1, 1)$ and in general, $X = (x_1, x_2)$, as indicated in Fig. 2.18. We can write

$$X : (x_1, x_2) = a_2(1, 1) + a_1(0, 1) = a_2 U + a_1 O$$

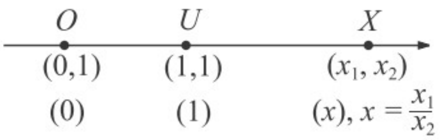

Figure 2.18. Homogeneous coordinates on the line.

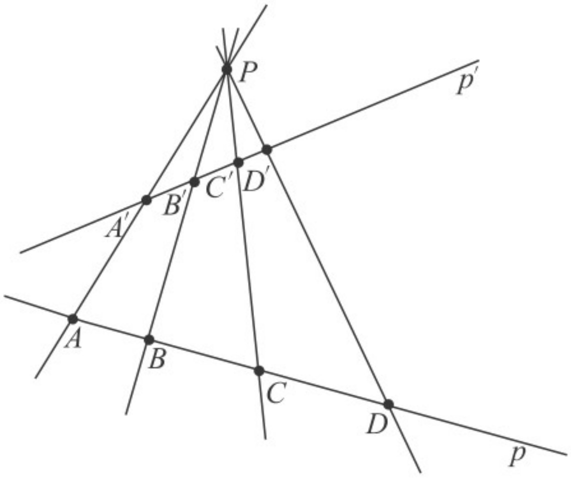

Figure 2.19. Two lines p, p' in perspective from the point P.

for $x_1 = a_2$, $x_2 = a_1 + a_2$, and O, U acting as a basis. The point O can be thought of as the "origin," and U establishes the "unit"; recall that we do not have a metric in the projective plane. There is a further simplification obtained by dividing by x_2 to represent the point X by the single value $x = x_1/x_2$.

2.5 The Fundamental Transformations of Projective Geometry

A short diversion into the basic transformations of projective geometry is needed in order to prove a fundamental property of homogeneous coordinates. Consider two *directed* (in the Euclidean plane) lines p, p' and a point P not on the lines, as shown in Fig. 2.19. A line on P intersecting p at the point A and p' at A' establishes a one-to-one correspondence called *perspectivity* between the points of p and p'. It is denoted by

$$p(A, B, C, D, \dots) \overset{P}{\underset{\wedge}{=}} p'(A', B', C', D', \dots).$$

The two lines are said to be in perspective with respect to the point P, which may be indicated above the $\underset{\wedge}{=}$ symbol. Of course, there is the dual concept, shown in Fig. 2.20, of two points P, P' being in perspective with respect to a line p denoted by

$$P(a, b, c, d, \dots) \underset{\wedge}{=} P'(a', b', c', d', \dots).$$

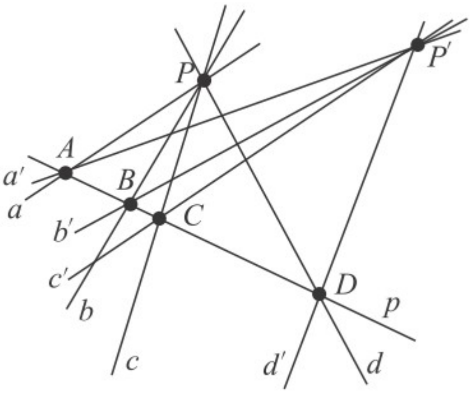

Figure 2.20. Two points P, P' in perspective from the line p.

Consider now a perspectivity $p(A, B, C, D, \dots) \underset{\wedge}{=} p'(A', B', C', D', \dots)$ as shown in Fig. 2.21 and the *cross-ratio*

$$(A, B; C, D) = \frac{AC/BC}{AD/BD}.$$

Theorem 2.5.1. *The cross-ratio is invariant under perspectivity.*

Proof. We give an elementary proof in the Euclidean plane, leaving the projective aspects for the exercises. With reference to Fig. 2.21,

$$(A, B; C, D) = \frac{\frac{IP \cdot AC}{IP \cdot BC}}{\frac{IP \cdot AD}{IP \cdot BD}} = \frac{\frac{\text{area}\triangle APC}{\text{area}\triangle BPC}}{\frac{\text{area}\triangle APD}{\text{area}\triangle BPD}} = \frac{\frac{AP \cdot CP \sin(APC)}{BP \cdot CP \sin(BPC)}}{\frac{AP \cdot DP \sin(APD)}{BP \cdot DP \sin(BPD)}},$$

or

$$(A, B; C, D) = \pm \frac{\sin(APC) \cdot \sin(BPD)}{\sin(BPC) \cdot \sin(APD)}. \tag{2.5}$$

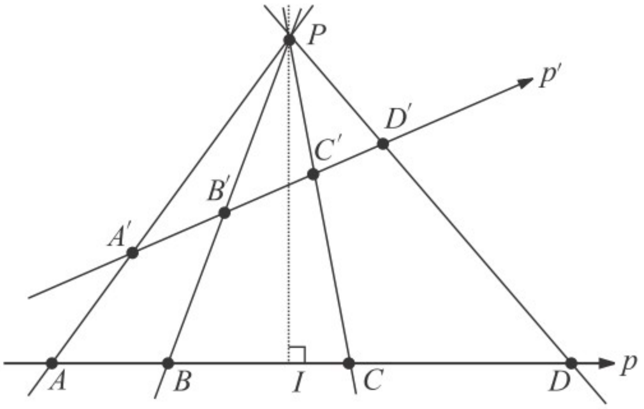

Figure 2.21. Diagram for the cross-ratio property.

Similarly, it can be shown that $(A', B'; C', D')$ also equals the right-hand-side of (2.5). There remains to show that the signs match (see exercises below). As stated, this definition of the cross-ratio applies only to the Euclidean plane, but it can be appropriately generalized to the augmented plane (π, p_∞), the cross-section invariance being a *projective* property. ■

The set of all points on a line is called a *pencil of points*, and dually, the set of all lines on a point is called a *pencil of lines*. A pencil of points may undergo successive perspectivities, as in

$$p_1 \overset{P_1}{\doublebarwedge} p_2 \overset{P_2}{\doublebarwedge} \cdots \overset{P_{n-1}}{\doublebarwedge} p_n.$$

A sequence of perspectivities is called a *projectivity* and is denoted by $p \barwedge p_n$. In projective geometry we study the properties that remain invariant under projectivities. We state without proof a couple of important results.

Theorem 2.5.2. *A projectivity is determined by the correspondence between two triples of distinct points. That is, if the triples* (A, B, C) *and* (A', B', C') *correspond under a projectivity* $p \barwedge p_n$*, then the correspondence between any* $X \in p$ *and* $X' \in p'$ *is uniquely determined (i.e., projectivity is determined by just three points).*

This does not determine the actual sequence of perspectivities from which the projectivity resulted. Rather, it just determines the correspondence between the original and final pencils of points.

Theorem 2.5.3. *A projectivity can be expressed by at most two perspectivities.*

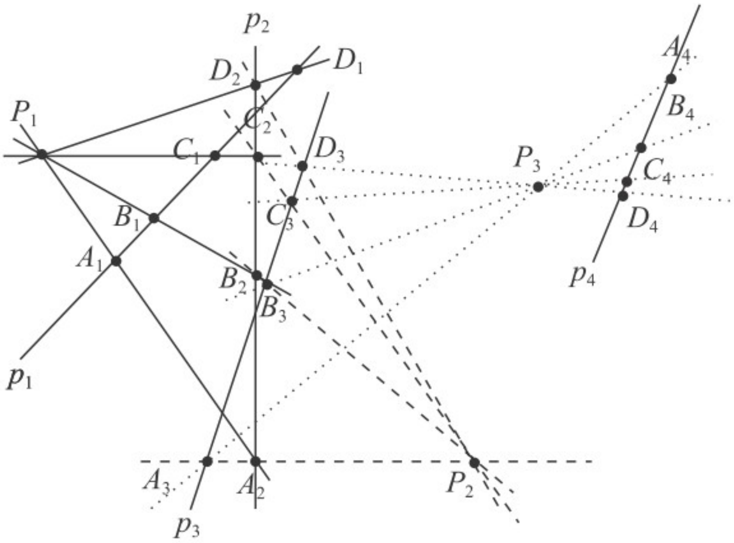

Figure 2.22. A projectivity composed of four perspectivities.

Exercises

1. Show that the 24 distinct permutations of four symbols A, B, C, and D result in exactly six different values of the cross-ratios as formed from these four symbols (corresponding to four points on a line)
2. Prove Theorem 2.5.1 when the perspectivity is with respect to an ideal point.
3. Prove Theorem 2.5.2 in the Euclidean plane and also with a perspectivity with respect to an ideal point.
4. Prove Theorem 2.5.3 in the Euclidean plane and also with a perspectivity with respect to an ideal point.

2.6 More on Homogeneous Coordinates

2.6.1 Linearizability of Projective Transformations

It is easy to see that the cross-ratio remains invariant under a projectivity and in fact, completely characterizes (determines) the projectivity. This invariance leads to a very important result and the reason homogeneous coordinates are so widely used in fields such as computer graphics. It is proved first for the plane $\mathbb{P}^2$ and then generalized to higher dimensions. For a projectivity $s \barwedge s'$ between two directed lines

s and s' (as in Fig. 2.21, see lines p and p'), four points and their corresponding single homogeneous coordinates indicated by (), $P : (p)$, $Q : (q)$, $R : (r)$, $X : (x)$ are chosen to form the cross-ratio

$$(P, Q; R, X) = \frac{(p - r)(q - x)}{(p - x)(q - r)}.$$

Similarly, the cross-ratio

$$(P', Q'; R', X') = \frac{(p' - r')(q' - x')}{(p' - x')(q' - r')}$$

is formed with the corresponding points $P' : (p')$, $Q' : (q')$, $R' : (r')$, $X' : (x')$ on s'. By the cross-ratio invariance, under a projectivity, we have

$$(P, Q; R, X) = (P', Q'; R', X'), \tag{2.6}$$

yielding

$$x' = \frac{ax + b}{cx + d}. \tag{2.7}$$

The a, b, c, d are given in terms of the coordinates p, q, r, p', q', and r', and their actual values are immaterial to the development. This is a *linear fractional transformation*, also called a *Möbius transformation*, completely specifying the projectivity $s \,\overline{\wedge}\, s'$. Reverting back to the ratio obtained from the homogeneous coordinates, i.e., $x = x_1/x_2$ and $x' = x'_1/x'_2$, we obtain from (2.7),

$$\frac{x'_1}{x'_2} = \frac{ax_1 + bx_2}{cx_1 + dx_2}. \tag{2.8}$$

Then rewriting the homogeneous coordinates of a point on a line in terms of two components as $X = (x_1, x_2)$, see Fig. 2.18, we get

$$k\left(\frac{x'_1}{x'_2}, 1\right) = \left(\frac{ax_1 + bx_2}{cx_1 + dx_2}, 1\right).$$

Here k is an arbitrary proportionality constant, which is needed, since *multiples* of a triple (the 1 in the third position is not shown; see Fig. 2.18) represent the same point. Eliminating the fractions yields

$$k(x'_1, x'_2) = (ax_1 + bx, cx_1 + dx_2),$$

from which we obtain the *linear* relations

$$kx'_1 = ax_1 + bx_2, kx'_2 = cx_1 + dx_2 \tag{2.9}$$

between the coordinates on the directed lines s and s'. The transformation (2.8) is called *projective*, and this is the most general class of *nonlinear* transformations that can be linearized by homogeneous coordinates. It is worth examining this in greater detail. Rewrite the transformation as

$$T(x) = \frac{ax+b}{cx+d} = \left(\frac{bc-ad}{c^2}\right) \cdot \left(\frac{1}{x+d/c}\right) + \left(\frac{a}{c}\right),$$

which provides the decomposition of T into

$$T = T_4 \circ T_3 \circ T_2 \circ T_1,$$

where $\circ$ denotes the transformation composition and

$$T_1(x) = x + \frac{d}{c}, \quad T_2(x) = \frac{1}{x}, \quad T_3(x) = \frac{bc-ad}{c^2}x, \quad T_4(x) = x + \frac{a}{c}.$$

That is, T is composed of two translations T_1, T_4, a scaling T_3, and an inversion T_2. Inversions and translations, being nonlinear, are linearized, and this property makes homogeneous coordinates especially useful in applications, since the *projective* transformations, translation, perspective, rotation, and scaling, can all be described in terms of matrices. Hence, a sequence of such transformations can be expressed as a matrix product and computed in a convenient and efficient way (e.g., see [146]).

For n-dimensional projective space $\mathbb{P}^N$ and projectivities there, the linear fractional transformation (2.7) generalizes to

$$x_i'' = \frac{a_{1i}x_1' + a_{2i}x_2' + \cdots + a_{N,i}x_N' + a_{(N+1),i}}{b_{1i}x_1' + b_{2i}x_2' + \cdots + b_{N,i}x_N' + b_{(N+1),i}}, \quad i = 1, 2, \ldots, n, \qquad (2.10)$$

where we have reverted to homogeneous coordinates by

$$x_i'' = \frac{x_i'}{x_{(N+1)}'}, \quad x_j' = \frac{x_j}{x_{(N+1)}}, \quad j = 1, 2, \ldots, (n+1).$$

Here the x_i are Cartesian coordinates, while the x_i' and x_i'' are homogeneous coordinates.

2.7 ** The Linearized Projective Transformations

To recap, the projective transformation

$$T(x) = ax \qquad (2.11)$$

is linear and is called *scaling* when $a > 0, a \neq 1$; then it is a *contraction* for $a < 1$ and a *stretching*, or *magnification*, when $a > 1$. For complex $a = r(\cos\theta + i\sin\theta)$, the result is a *rotation* of x by an angle θ about the origin as well as a scaling by r. By contrast, *translation*

$$T(x) = x + a \tag{2.12}$$

and *inversion* or *perspective*

$$T(x) = \frac{1}{x} \tag{2.13}$$

are not linear, and for these the linearization, accomplished via homogeneous coordinates, enables the transformations to be expressed in matrix form. This is of importance in such applications as *computer graphics* and *geometric modeling*.

2.7.1 Projective Transformations in the Plane $\mathbb{P}^2$

For notational convenience, a point $P = (x_1, x_2, x_3)$ and its transform $P' = (x_1', x_2', x_3')$, expressed in homogeneous coordinates, are taken as row vectors, while the transformations are denoted by two capital letters. We take $x_3' = c'$ and $x_3 = c$, such that for ordinary points, $c, c' \neq 0$, and $c, c' = 0$ for ideal points. Many references in computer graphics, geometric modeling, robotics, and other fields consider only the case $x_3 = 1$, which unnecessarily excludes the ideal points.

The scaling transformation $P' = PS(s_1, s_2)$, where $x_i' = s_i x_i$ and $x_3' = x_3 = c$, in matrix form is

$$PS: \quad (x_1', x_2', c) = (x_1, x_2, c)\begin{pmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{2.14}$$

The rotation $P' = PR(\theta)$ by an angle θ about the origin is

$$PR: \quad (x_1', x_2', c) = (x_1, x_2, c)\begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{2.15}$$

For the translation $P' = PT(t_1, t_2)$, with $x_i' = x_i + t_i$, we can set $c = 1$, since translation of ideal points is not meaningful, to obtain

$$PT: \quad (x_1', x_2', 1) = (x_1, x_2, 1)\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ t_1 & t_2 & 1 \end{pmatrix}. \tag{2.16}$$

Successive scaling or rotation transformations result in successive multiplications of the corresponding matrices, as they should, since these are linear transformations. The fun and beauty is that it also works with translations. For $P' = PT_1(t_1^1, t_2^1)$ and $P'' = P'T_2(t_1^2, t_2^2)$, the result is $P'' = PT_1T_2 = PT_2T_1$ so that $x_i'' = x_i + t_i^1 + t_i^2$. The matrices commute with respect to multiplication, as they should, since the result of successive translations does not depend on their order.

The composition of scalings, rotations, and translations is obtained by the multiplication of the corresponding matrices. In addition, the matrices are sparse (i.e., they have many zeros), and the computational advantage is tremendous.

2.7.2 Projective Transformations in $\mathbb{P}^3$

The generalizations of these transformations to 3-D are straightforward, with the previous remarks regarding $x_3 = c$ now pertaining to $x_4 = c$. For scaling, $P' = PS(s_1, s_2, s_3)$,

$$PS: \quad (x_1', x_2', x_3', c) = (x_1, x_2, x_3, c)\begin{pmatrix} s_1 & 0 & 0 & 0 \\ 0 & s_2 & 0 & 0 \\ 0 & 0 & s_3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{2.17}$$

Let the matrices for the rotation about the x_i-axes by an angle θ be denoted by $PR_i(\theta)$, $i = 1, 2, 3$. The rotation about x_3 is given by

$$PR_3: \quad (x_1', x_2', x_3', c) = (x_1, x_2, x_3, c)\begin{pmatrix} \cos\theta & \sin\theta & 0 & 0 \\ -\sin\theta & \cos\theta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{2.18}$$

and is a direct extension of the 2-D rotation. The two other rotation matrices are

$$\begin{gathered} PR_1(\theta) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta & \sin\theta & 0 \\ 0 & -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \\ PR_2(\theta) = \begin{pmatrix} \cos\theta & 0 & -\sin\theta & 0 \\ 0 & 1 & 0 & 0 \\ \sin\theta & 0 & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \end{gathered} \tag{2.19}$$

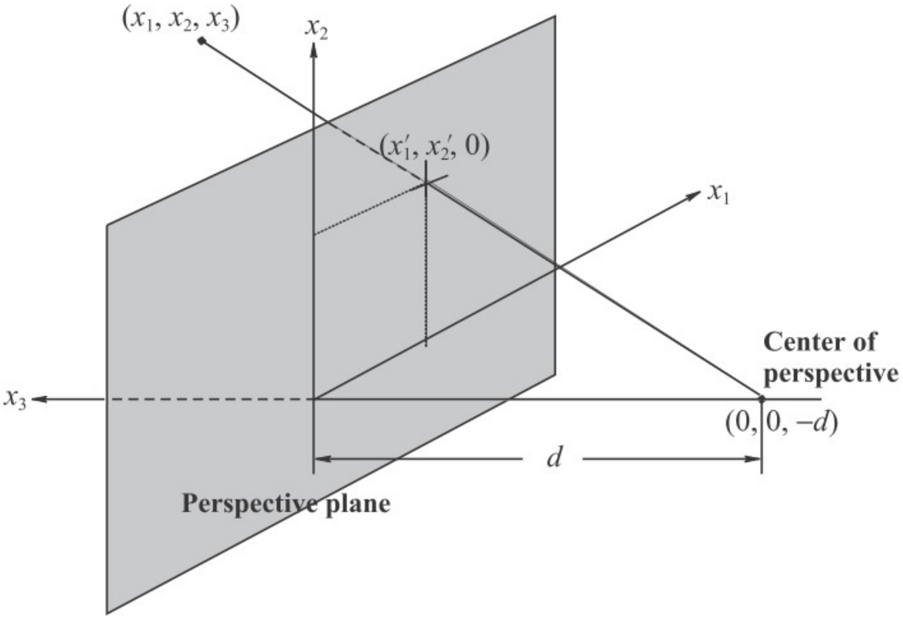

Figure 2.23. Perspective with respect to the x_3-axis.

For 3-D translations $P' = PT(t_1, t_2, t_3)$, $c = 1$ as in 2-D, and

$$PT: \quad (x_1', x_2', x_3', 1) = (x_1, x_2, x_3, 1)\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ t_1 & t_2 & t_3 & 1 \end{pmatrix}. \tag{2.20}$$

In 3-D, the perspective transformation is important, for it provides images with a sense of depth and realism. It is convenient to choose the center of the perspective transformation on the negative x_3-axis at a distance d behind the perspective plane, as shown in Fig. 2.23. The transformation equations, obtained by similar triangles, are

$$x_1' = \frac{x_1}{x_3/d + 1}, \quad x_2' = \frac{x_2}{x_3/d + 1}, \quad x_3' = 0. \tag{2.21}$$

Denote the perspective transformation matrix by $V[\pi_{per}, C_{per}]$, where π_{per} and C_{per} are the perspective plane and center respectively. For the setup in Fig. 2.23, $P' = PV[x_3 = 0, (0, 0, -d)]$, and explicitly,

$$(x'_1, x'_2, 0, w) = (x_1, x_2, x_3, c)\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1/d \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{2.22}$$

with the relation between c and w emerging next. The perspective image of an ideal point $P = (1, x_2/x_1, x_3/x_1, 0)$, called a *vanishing point*, is the *ordinary point* $P' = (x'_1, x'_2, 0, x_3/d) = (dx_1/x_3, dx_2/x_3, 0, 1)$. Hence

$$w = \frac{x_3}{d} + c, \quad x'_1 = \frac{x_1}{w}, \quad x'_2 = \frac{x_2}{w}, \quad x'_3 = 0.$$

All projective transformations can now be performed successively by matrix multiplication. The matrix of the perspective transformation, unlike those of other projective transformations, is singular, confirming that an object cannot be uniquely retrieved from its perspective image.

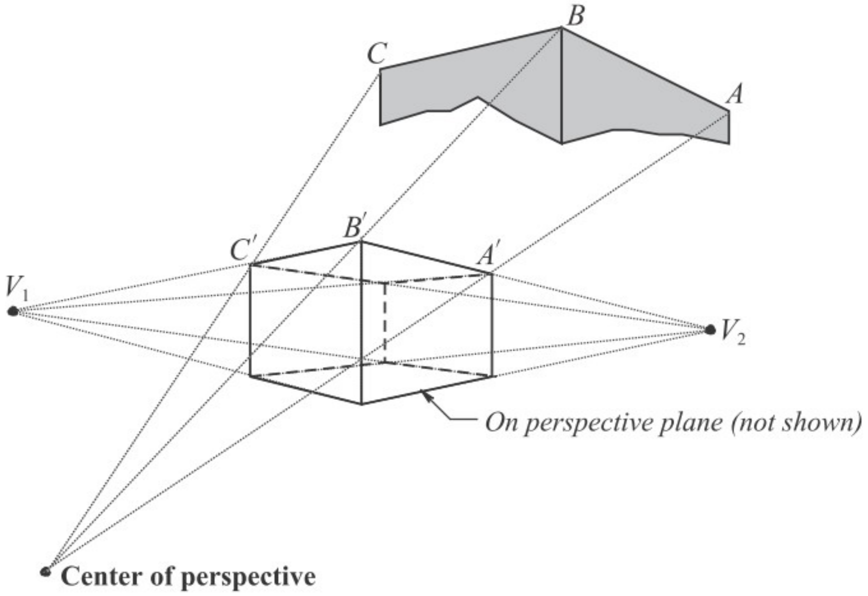

Figure 2.24. The perspective image of an object (partially shown).
In the figure, the perspective transformation of three vertices A, B, C is traced, showing their images A', B', C'. This is sometimes called a *2-point perspective*, having two *vanishing points* V_1 and V_2, the images of ideal points, giving the image "3-dimensionality." Actually, the ideal point in the vertical direction is the "third vanishing point." It would have been a *3-point perspective* had the perspective plane not been parallel to the front (vertical) face of the object. All lines joining the object's points with their images intersect at the *center of perspective*.

2.8 ** Coordinate System Rotations

This section, complementing the discussion of rotations about one axis above, is adapted from an old gem [161, pp. 240–241] (the textbook from which I learned analytic geometry). In $\mathbb{R}^3$ with a positive direction assigned to a line ℓ, translate the coordinate system to any point $Q \in \ell$, denoting by α, β, γ the *direction angles* between the positive direction of ℓ and the translates of the axes x_1, x_2, x_3 respectively. Their cosines, called *direction cosines*, have the property

$$\cos^2\alpha + \cos^2\beta + \cos^2\gamma = 1. \tag{2.23}$$

Denoting by $\hat{\mathbf{e}}_\mathbf{i}$ the unit vectors in the x_i directions, we obtain that

$$\hat{\ell} = \hat{\mathbf{e}}_\mathbf{1}\cos\alpha + \hat{\mathbf{e}}_\mathbf{2}\cos\beta + \hat{\mathbf{e}}_\mathbf{3}\cos\gamma \tag{2.24}$$

is the unit vector in the positive direction of ℓ, with (2.23) equivalent to $\hat{\ell}\cdot\hat{\ell} = |\hat{\ell}|^2$. The cosine of the angle ϕ between two directed lines ℓ_1, ℓ_2 is given in terms of their corresponding unit vectors $\hat{\ell}_\mathbf{1}, \hat{\ell}_\mathbf{2}$, and their corresponding direction angles, subscripted by 1 and 2, respectively, are given by

$$\cos\phi = \hat{\ell}_\mathbf{1}\cdot\hat{\ell}_\mathbf{2} = \cos\alpha_1\cos\alpha_2 + \cos\beta_1\cos\beta_2 + \cos\gamma_1\cos\gamma_2, \tag{2.25}$$

with $\cos\phi = 0$ when the lines are perpendicular.

For a plane π, let n be the directed line through the origin O and perpendicular to π with direction angles α, β, γ. The normal form of the equation of π is

$$\pi : x_1\cos\alpha + x_2\cos\beta + x_3\cos\gamma = d, \tag{2.26}$$

with d the directed distance of π from O. It is an easy matter to verify that the directed distance p of π from the point $P = (p_1, p_2, p_3)$ is given by

$$p = p_1\cos\alpha + p_2\cos\beta + p_3\cos\gamma - d. \tag{2.27}$$

With reference to Fig. 2.25, the coordinate system x_1, x_2, x_3 is rotated about the origin O to the system x_1', x_2', x_3', all points remaining stationary with a point P having two sets of coordinates (x_1, x_2, x_3), (x_1', x_2', x_3'). Our task is to find the relationship between them. Letting the direction cosines of the x_i' axes with respect to x_i be $\cos\alpha_i, \cos\beta_i, \cos\gamma_i$ for $i = 1, 2, 3$ respectively, we obtain

$$\begin{aligned}&\cos^2\alpha_i + \cos^2\beta_i + \cos^2\gamma_i = 1,\\ &\cos\alpha_j\cos\alpha_k + \cos\beta_j\cos\beta_k + \cos\gamma_j\cos\gamma_k = 0,\end{aligned} \tag{2.28}$$

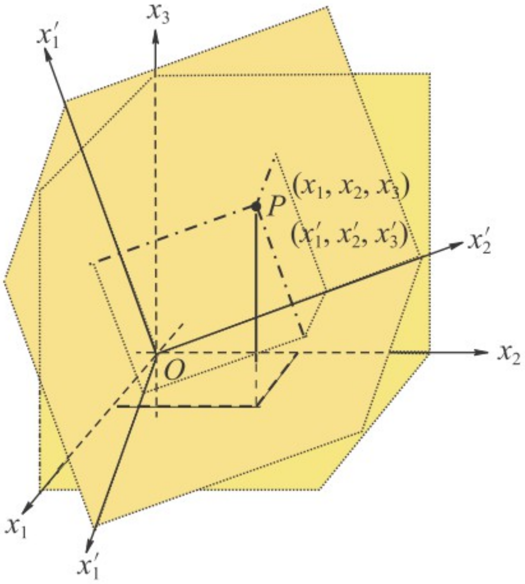

Figure 2.25. Rotation of the coordinate system $x_1x_2x_3$ about the origin.

where $j \neq k\, i,\, j, k = 1, 2, 3$, the first part corresponding to (2.23) for the x_i' and the second due to the orthogonality between the pairs x_j', x_k' of axes. The plane $x_2'x_3'$, which passes through the origin O, has equation

$$x_2'x_3' : x_1 \cos\alpha_1 + x_2 \cos\beta_1 + x_3 \cos\gamma_1 = 0,$$

with coefficients the direction cosines of the axis x_1', which is orthogonal to it. Similarly, for all the 2-planes,

$$x_j'x_k' : x_1 \cos\alpha_i + x_2 \cos\beta_i + x_3 \cos\gamma_i = 0, \quad i \neq j \neq k. \tag{2.29}$$

The directed distance of the point P from the plane $x_j'x_k'$ is its x_i' coordinate, $i \neq j \neq k$; hence by (2.27), (2.29),

$$x_i' = x_1 \cos\alpha_i + x_2 \cos\beta_i + x_3 \cos\gamma_i. \tag{2.30}$$

These describe the coordinate transformation $\{x_1, x_2, x_3\} \to \{x_1', x_2', x_3'\}$. Reversing the roles of x_i and x_i' in the previous reasoning, we obtain the inverse transformation $\{x_1', x_2', x_3'\} \to \{x_1, x_2, x_3\}$:

$$x_i = x_1 \cos\alpha_i + x_2 \cos\beta_j + x_3 \cos\gamma_k, \quad i \neq j \neq k. \tag{2.31}$$

For consistency with Section 2.7, (2.30) and (2.31) are rewritten in homogeneous coordinates and in matrix form:

$$(x_1', x_2', x_3', 1) = (x_1, x_2, x_3, 1)\begin{pmatrix} \cos\alpha_1 & \cos\beta_1 & \cos\gamma_1 & 0 \\ \cos\alpha_2 & \cos\beta_2 & \cos\gamma_2 & 0 \\ \cos\alpha_3 & \cos\beta_3 & \cos\gamma_3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{2.32}$$

$$(x_1, x_2, x_3, 1) = (x_1', x_2', x_3', 1)\begin{pmatrix} \cos\alpha_1 & \cos\alpha_2 & \cos\alpha_3 & 0 \\ \cos\beta_1 & \cos\beta_2 & \cos\beta_3 & 0 \\ \cos\gamma_1 & \cos\gamma_2 & \cos\gamma_3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{2.33}$$

The matrix A of the transformation in (2.30) has determinant $|A| = 1$, and the matrix of the inverse transformation (2.31) is the transpose A^T. These, and the matrices corresponding to the rotations about one axis in Section 2.7, are examples of *orthogonal* matrices (transformations), which also appear in the previous section. As a reminder, the above equations *together* with the six conditions of (9.47) specify the rotation of the coordinate system about the origin. By the way, such a rotation about an arbitrary point can be obtained by translating the origin first and then rotating, or equivalently, multiplying the matrix A in (9.48) by that in (2.20), with resulting matrix describing the combined transformation. That is,

$$(x_1', x_2', x_3', 1) = (x_1, x_2, x_3, 1)\begin{pmatrix} \cos\alpha_1 & \cos\beta_1 & \cos\gamma_1 & 0 \\ \cos\alpha_2 & \cos\beta_2 & \cos\gamma_2 & 0 \\ \cos\alpha_3 & \cos\beta_3 & \cos\gamma_3 & 0 \\ t_1 & t_2 & t_3 & 1 \end{pmatrix}. \tag{2.34}$$

From the vector and matrix form, the generalization to $\mathbb{R}^N$ is direct, with a slight change in notation denoting the direction angle between the x_i axis and a directed line $\ell \subset \mathbb{R}^N$ by α_i and the unit vector in the direction of ℓ:

$$\hat{\ell} = \sum_{i=1}^{N} \hat{\mathbf{e}}_{\mathbf{i}} \cos\alpha_i, \quad \sum_{i=1}^{N} \cos^2\alpha_i = 1. \tag{2.35}$$

Proceeding as above, the corresponding $(N+1)\times(N+1)$ transformation matrix is obtained *together* with the $(N^2+N)/2$ additional conditions analogous to (9.47): N conditions from the above equation and $N(N-1)/2$ more due to the orthogonality of all pairs of axes.

2.9 ** Analytic Proofs

$\mathcal{D}$esargues's theorem will be helpful in subsequent chapters. This is an oppurtune time to prove it and learn yet another nice way to utilize homogeneous coordinates.

2.9.1 Computational Proof

Three noncollinear points P_i, specified in homogeneous coordinates by triples $A_i, i = 1, 2, 3$, are chosen on the projective plane $\mathbb{P}^2$. Another point P, with triple A, is selected such that $A = c_1 A_1 + c_2 A_2 + c_3 A_3$ with all $c_i \neq 0$. Each point P_i is also represented by the triple $c_i A_i = A_i^*$, which are fixed from now on, leaving $A = A_1^* + A_2^* + A_3^*$. Any other point Q can be described as $Q = a_1 A_1^* + a_2 A_2^* + a_3 A_3^*$ with the A_i^* of the multiples now fixed. The coefficients can be used as a new set of coordinates, i.e., $Q = (a_1, a_2, a_3)$ and $P_1 = (1, 0, 0), P_2 = (0, 1, 0), P_3 = (0, 0, 1), P = (1, 1, 1)$. In this coordinate system it is easy to see that $Q = (a_1, a_2, a_3)$ and $Q' = k(a_1, a_2, a_3), k \neq 0$, represent the same point, allowing us to write $Q = Q'$. This is similar to homogeneous coordinates with 0 in the third position no longer implying that the point is ideal. This turns out to be a very useful tool.

To illustrate, we provide an analytic proof of Desargues's theorem. It is convenient to take the triangle's vertices A, B, C as the basis triples and $P = A + B + C$ or $P = (1, 1, 1)$ as the point of perspectivity, as shown in Fig. 2.26. The three lines from P through the vertices of the triangle are denoted by a', b', and c'. With $A = (1, 0, 0)$ and $P = (1, 1, 1)$ on a', from (2.4) we find that

$$a' = \left[\begin{vmatrix} 1 & 1 \\ 0 & 0 \end{vmatrix}, \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix}, \begin{vmatrix} 1 & 1 \\ 1 & 0 \end{vmatrix} \right] = [0, 1, -1].$$

For a point $A' = (a, a_1, a_2) \in a'$, where $A \neq A'$, the collinearity of P, A, A' and (2.3) yield

$$\begin{vmatrix} a & a_1 & a_2 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{vmatrix} = 0,$$

which shows that $a_1 = a_2$. Since these values are arbitrary, we can take $A' = (a, 1, 1)$. Similarly, we obtain the line coordinates $b' = [1, 0, -1]$, $c' = [1, -1, 0]$ and the point coordinates $B' = (1, b, 1)$, $C' = (1, 1, c)$. Let r be the line determined by A and B. Then

$$r = \left[\begin{vmatrix} 0 & 0 \\ 1 & 0 \end{vmatrix}, \begin{vmatrix} 0 & 1 \\ 0 & 0 \end{vmatrix}, \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} \right] = [0, 0, 1].$$

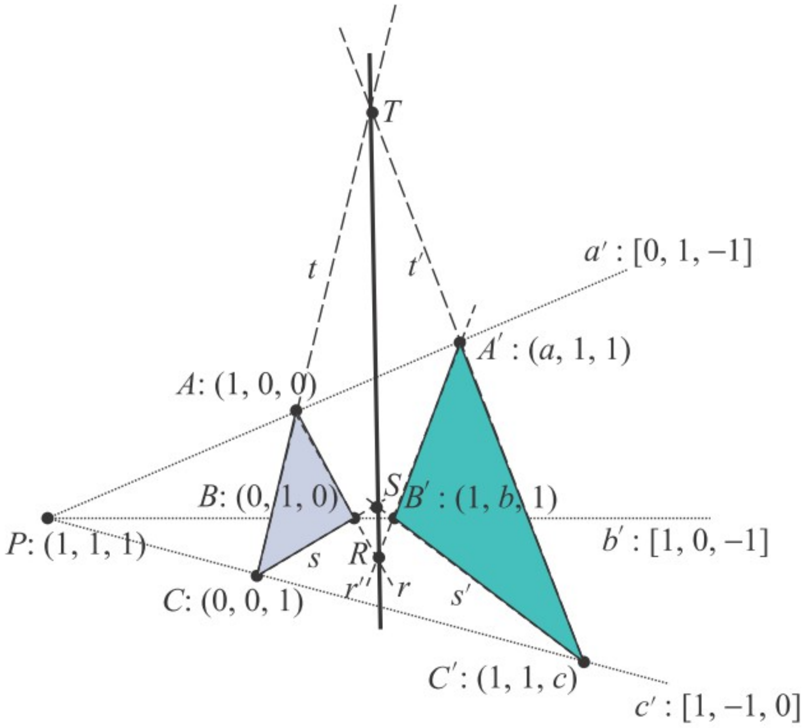

Figure 2.26. Diagram for analytic proof of Desargue's theorem.

Similarly, for the line r' determined by A' and B',

$$r' = \left[\begin{vmatrix} 1 & 1 \\ b & 1 \end{vmatrix}, \begin{vmatrix} a & 1 \\ 1 & 1 \end{vmatrix}, \begin{vmatrix} a & 1 \\ 1 & b \end{vmatrix} \right] = [1-b, a-1, ab-1].$$

To find the intersection $r \cap r' = R$, we can use the *dual* of (2.4). That is, in its original form, (2.4) yields the coordinates of the *line* determined by the coordinates of two of its points. In the dual, the same equation yields the coordinates of the *point* on the two lines from their coordinates. Namely,

$$R = \left(\begin{vmatrix} 0 & 1 \\ a-1 & ab-1 \end{vmatrix}, \begin{vmatrix} 1 & 0 \\ ab-1 & 1-b \end{vmatrix}, \begin{vmatrix} 0 & 0 \\ 1-b & a-1 \end{vmatrix} \right)$$

$$= (1-a, b-1, 0).$$

Continuing in the same way, we obtain

$$S = BC \cdot B'C' = s \cap s' = (0, 1 - b, c - 1),$$

$$T = AC \cdot A'C' = t \cap t' = (a - 1, 0, 1 - c).$$

Applying again (2.3) with the coordinates found for R, S, T yields

$$\begin{vmatrix} 1-a & b-1 & 0 \\ 0 & 1-b & c-1 \\ a-1 & 0 & 1-c \end{vmatrix} = 0,$$

showing that R, S, and T are always collinear, proving Desargues's theorem in the *projective plane*.

2.9.2 Algebraic Proof

When the coordinates of the points and lines are not explicitly needed, their computation can be avoided with an algebraic proof. In Fig. 2.26, let all points be identified by their homogeneous coordinate triples. By the collinearity of P, A, A' there exist nonzero constants c_1, c_2 such that

$$P = c_1 A + c_2 A'. \tag{2.36}$$

The multiple $c_1 A$ identifies the *same* point as A, and similarly, $c_2 A'$ identifies $-A'$. Without loss of generality we can take the triples $c_1 A, c_2 A'$ for the representation of the points $A, -A'$, simplifying (2.37) to

$$P = A - A'. \tag{2.37}$$

The new triples for A, A' then remain *fixe* as in the previous section. Similarly, there exist constant multiples of B, B', C, C' such that

$$P = B - B', \quad P = C - C'. \tag{2.38}$$

Then

$$A - B = A' - B'. \tag{2.39}$$

The point $A - B$ is on the line AB, and the point $A' - B'$ is on $A'B'$. Therefore,

$$A - B = A' - B' = R = AB \cap A'B'. \tag{2.40}$$

Continuing, we obtain

$$C - A = C' - A' = T = AC \cap A'C', \quad B - C = B' - C' = S = BC \cap B'C'. \tag{2.41}$$

Altogether,

$$R + S + T = A - B + C - A + B - C = 0, \tag{2.42}$$

showing that the intersection points R, S, and T are collinear.

It is worth contrasting these proofs with the synthetic proofs of Desargues's theorem, as illustrated in Fig. 2.27.

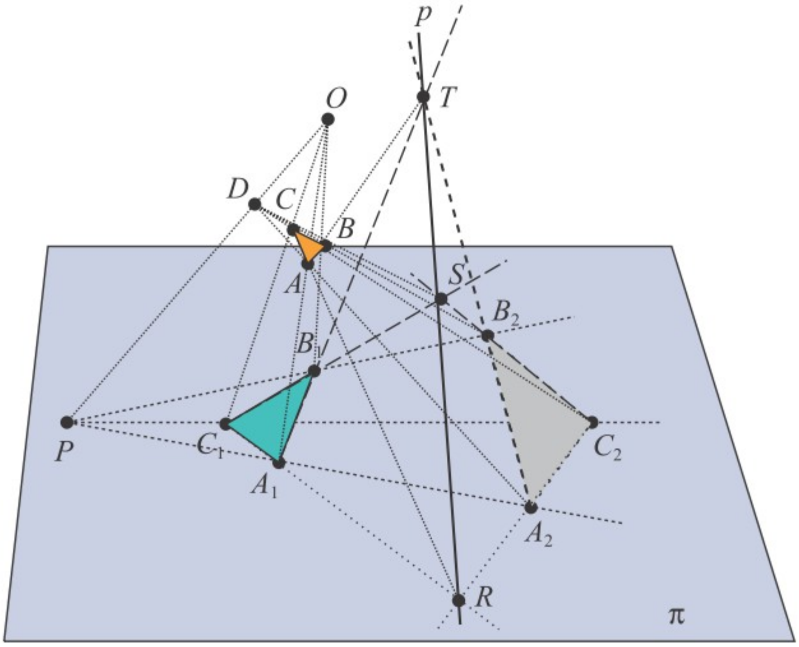

Figure 2.27. Diagram for the synthetic proof of Desargues theorem.
This proof involves embedding the 2-D perspectivity in 3-D. The theorem statement is that *two triangles in perspective from a point* P *are in perspective from a line* p.

Fun and Games with the ILM

Open the ILM: set *Chapter* = *Geometry* and click on *"Lock Guides."*
The display shows **Desargues's theorem** with the three intersection points R, S, T collinear.
Click and move any of the **red** triangle vertices along the lines.
The three points R, S, T remain collinear.

Next set *Example* = *Pappus* and click on *"Lock Guides."*
The display shows **Pappus's theorem**.
Again click on and move any of the six points A_i, B_i along the lines.
The three intersection points C_i remain collinear.

Exercises

1. Provide an analytic proof of Pappus's theorem.
2. Generalize and find the representation of the ideal elements of $(\mathbb{R}^3, p_\infty)$ in terms of ordered quadruples.
3. Generalize the representation of the ideal elements of $\mathbb{P}^N$ in terms of ordered $(N+1)$-tuples.

2.10 ** Conics on the Projective Plane

Curves can be considered as a collection of points and dually as a collection of lines whose *envelope* forms a curve, (see Fig. 2.28.) To distinguish them, they are referred to as *point-curves* and *line-curves* respectively. The duality of the projective plane enables us to treat them equivalently and leads to some nice results, illustrated here with conic point- and line-curves.

2.10.1 In the Projective Plane All Conics Are Ellipses

That all conics are ellipses in the projective plane can be observed on the right of Fig. 2.28, which relates to the position of an ellipse with respect to the ideal line l_∞. That is, an ordinary (Euclidean) ellipse has no ideal points, and a Euclidean parabola corresponds to an ellipse tangent to the ideal line and hence having one ideal point corresponding to the direction in which the parabola "opens toward

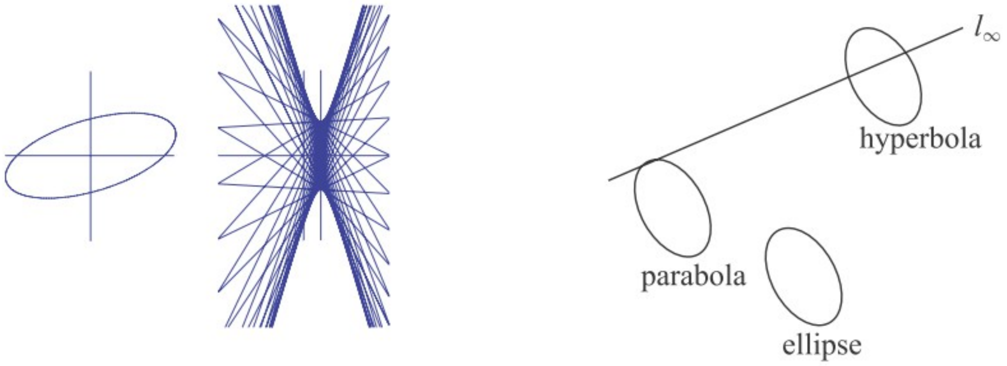

Figure 2.28. Ellipse (point-curve) and hyperbola (line-curve). Projective conics are ellipses.

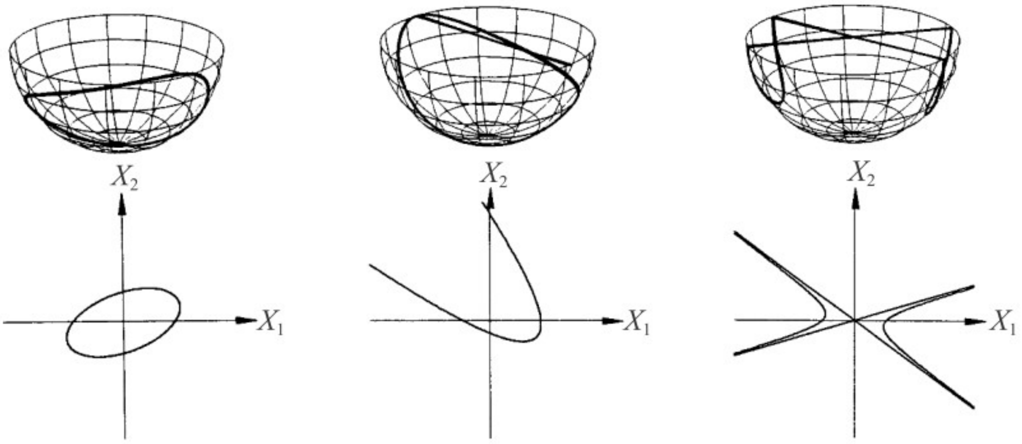

Figure 2.29. Ellipse (left), parabola (center), and hyperbola (right) in the projective plane model.

infinity." A Euclidean hyperbola is a projective ellipse intersecting l_∞ at two ideal points corresponding to the directions of its asymptotes. The three cases in Fig. 2.29 are shown in the projective plane model.

2.10.2 Generalizing Pappus's Theorem

In 1640, Pascal generalized Pappus's theorem to conics.

Theorem 2.10.1 **(Pascal)** *If a simple hexagon* $A_1A_2A_3B_1B_2B_3$ *is inscribed in a conic, the intersections* $C_i = A_jB_k \cdot A_kB_j$, $i \neq j \neq k$, *are collinear.*

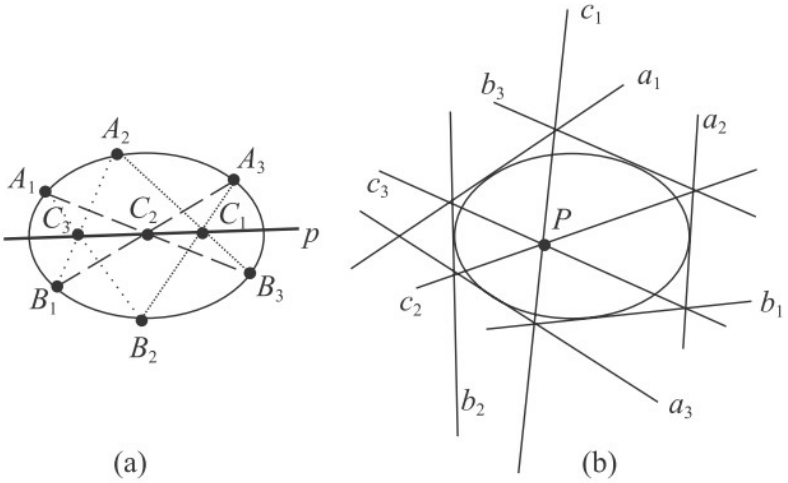

Figure 2.30. Pascal's theorem (a) and its dual, Briachnon's theorem (b).

Briachnon proved the result depicted in Fig. 2.30 in 1806, which turned out to be the *dual* of Pascal's theorem.

2.11 ** Geometries and Invariants

As already mentioned in connection with (2.10), homogeneous coordinates can be extended to higher dimensions. Specifically, a point $P \in \mathbb{P}^N$ is represented by an (N+1)-tuple, the *point* coordinates

$$P = (x_1, x_2, \dots, x_N, x_{N+1}) \text{ or } (kx_1, kx_2, \dots, kx_N, kx_{N+1}), k \neq 0,$$

where not all components are zero. Dually, a hyperplane π is represented by the $(N+1)$-tuple

$$\pi = [X_1, X_2, \dots, X_N, X_{N+1}], \text{ or } [kX_1, kX_2, \dots, kX_N, kX_{N+1}],$$

where again not all components are zero, these being the *hyperplane* coordinates of π. The *on* relation as before is defined by the inner product of the coordinates, i.e.,

$$P \text{ } on \text{ } \pi \iff \sum_{i=1}^{N+1} x_i X_i = 0.$$

A projective transformation in $\mathbb{P}^N$ is expressed by (2.10) and can be linearized by homogeneous coordinates. Projective transformations leave the cross-ratio invariant, and they are the most general transformations employed in projective geometry.

Various geometries were developed during the nineteenth century in addition to the Euclidean: affine, projective, spherical and inversive. Felix Klein proposed a way of classifying geometries that became known as the Erlangen Program, since he was a professor at the University of Erlangen. A geometry is regarded as a *space* together with a *group of transformations* acting on it. Geometrical properties are those remaining *invariant* under the group of transformations. As a reminder, a group is a set of elements G and a binary operation $\circ$ with

$a \circ b = c \in G \ \forall a, b \in G$ (closure property),
$(a \circ b) \circ c = a \circ (b \circ c)$ (associativity)
$\exists\, e \in G \ni a \circ e = e \circ a = a, \forall a \in G$ (identity)
$\forall a \in G \ \exists a^{-1} \ni a^{-1} \circ a = a \circ a^{-1} = e$ (inverse).

Exercises

Show that the set of all linear fractional transformations (i.e., the projective transformations) with the composition operation form a group.

The group of projective transformations has important special cases that form subgroups of transformations, i.e., subsets that are groups. With slight changes in the notation, consider the case of *translations*, for which the b's in the denominator of (2.10) are zero except for the last one, i.e.,

$$\mathbf{x}' = \mathbf{x}A + \mathbf{c},$$

with row vectors denoted in **bold**, e.g., $\mathbf{x}$, and A is an $N \times N$ matrix. In general, a mapping

$$L : D \to D$$

is a *linear transformation* if it preserves linear combinations, i.e.,

$$L(a\mathbf{x} + b\mathbf{y}) = aL(\mathbf{x}) + bL(\mathbf{y}).$$

A general *affine* transformation can be written as

$$T(\mathbf{u}) = L(\mathbf{u}) + \mathbf{c},$$

where $\mathbf{c}$ is a constant vector and L is a linear transformation. In this case,

$$T(a\mathbf{x}+b\mathbf{y}) = aL(\mathbf{x}) + bL(\mathbf{y}) + \mathbf{c} \neq aT(\mathbf{x}) + bT(\mathbf{y}) = aL(\mathbf{x}) + bL(\mathbf{y}) + 2\mathbf{c}.$$

In short, translations are not linear transformations. Affine transformations leave *convex* combinations $\left(\sum a_i x_i,\ \text{where } \sum a_1 = 1\right)$ invariant and also preserve *incidence* (i.e., the *on* relation).

Let us return to the representation

$$\mathbf{x}' = \mathbf{x}A + \mathbf{c}$$

and consider the case where the matrix has the property $AA^T = \lambda^2 I$, where λ is real and ^{T} denotes the transpose of a matrix. This is the *similarity* group. First take the case $\mathbf{c} = \mathbf{0}$. The square of the length, $|\mathbf{x}|$, of $\mathbf{x}$ is given by $\mathbf{x}\cdot\mathbf{x} = \mathbf{x}\mathbf{x}^T$ and is called the *norm* of $\mathbf{x}$. For the similarity transformation $\mathbf{x}' = \mathbf{x}A$, we have

$$\mathbf{x}'\cdot\mathbf{x}' = (\mathbf{x}A)\cdot(\mathbf{x}A) = \lambda^2\mathbf{x}\cdot\mathbf{x}.$$

Effectively, $|\mathbf{x}|$ has been amplified by $\pm\lambda$. For two vectors $\mathbf{x}$, $\mathbf{y}$, under similarity we obtain

$$\mathbf{x}'\cdot\mathbf{y}' = \lambda^2\mathbf{x}\cdot\mathbf{y} \quad \mathbf{x}\cdot\mathbf{y} = |\mathbf{x}|\cdot|\mathbf{y}|\cos(\mathbf{x},\mathbf{y}),$$

where $(\mathbf{x},\mathbf{y})$ is the angle between the two vectors. So

$$|\mathbf{x}'|\cdot|\mathbf{y}'|\cos(\mathbf{x}',\mathbf{y}') = \lambda^2|\mathbf{x}|\cdot|\mathbf{y}|\cos(\mathbf{x},\mathbf{y}).$$

Therefore, $\mathbf{x}\perp\mathbf{y} \iff \mathbf{x}'\perp\mathbf{y}'$, and in general, $\cos(\mathbf{x}',\mathbf{y}') = \cos(\mathbf{x},\mathbf{y})$. So under similarity, the ratio of distances and angles between vectors is preserved, or

$$\frac{|\mathbf{x}'|}{|\mathbf{y}'|} = \frac{|\mathbf{x}|}{|\mathbf{y}|}.$$

Specializing further to *orthogonal* matrices with $|A| = \pm 1$ or $|A||A^T| = |A|^2 = 1$, where $|A|$ denotes the determinant of A, we obtain the important group of orthogonal transformations.

The Euclidean distance between two vectors $\mathbf{x}, \mathbf{y}$ is given by $d(\mathbf{x},\mathbf{y}) = [(\mathbf{x}-\mathbf{y})\cdot(\mathbf{x}-\mathbf{y})^T]^{1/2}$:

$$\begin{aligned} d(\mathbf{x}',\mathbf{y}') &= [(\mathbf{x}'-\mathbf{y}')\cdot(\mathbf{x}'-\mathbf{y}')^T]^{1/2} = [(\mathbf{x}-\mathbf{y})AA^T(\mathbf{x}-\mathbf{y})^T]^{1/2} \\ &= [(\mathbf{x}-\mathbf{y})\cdot(\mathbf{x}-\mathbf{y})^T]^{1/2} = d(\mathbf{x},\mathbf{y}), \end{aligned}$$

showing that distance is also invariant under orthogonal transformations. Altogether then, areas, angles, and orthogonality are also preserved. All these together with the “pure” translation

$$\mathbf{x}' = \mathbf{x} + \mathbf{c}$$

form the *rigid motions*, which are the invariants of Euclidean geometry.

To recapitulate, as pointed out by Klein, the above geometries are projective geometries with additional conditions. For example, affine geometry is obtained by selecting any line in a projective geometry and considering only the transformations leaving that line invariant. This line plays the role of the ideal line and can be safely ignored. The reader is referred to [163], [157] for background and history of N-dimensional geometry, and to [152], [18] for learning more about the various geometries. Aspects of differential geometry, starting with a short chapter on the theory of envelopes, are covered later in the study of curves and surfaces.

3

♣ FT-1 The Plane with Parallel Coordinates

3.1 The Fundamental Duality

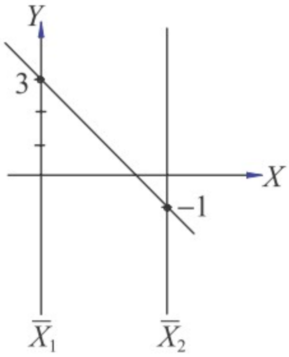

Figure 3.1. Points on the plane are represented by lines.

A point in the x_1x_2 plane is represented in parallel coordinates (abbreviated ∥-coords) by a line in the xy-plane (Fig. 3.1). To find out what a line "looks like" in ∥-coords, a set of collinear points is selected (Fig. 3.2, top right), and the lines representing them intersect (top left) at a point! Remarkably, this is true in general, and for the line

$$\ell : x_2 = mx_1 + b, \tag{3.1}$$

the corresponding point is

$$\bar{\ell} : \left(\frac{d}{1-m}, \frac{b}{1-m} \right), \quad m \neq 1, \tag{3.2}$$

where d is the distance between the parallel axes. As will be appreciated later, the interaxis distance is *directed*, so that interchanging the axes makes the sign of d negative. The point $\bar{\ell}$ is said to *represent* the line ℓ. It is evident that a point ↔ line duality[15] is at work here, which must be properly considered in the *projective* $\mathbb{P}^2$ rather than the Euclidean $\mathbb{R}^2$ plane. In (3.2), as

---

[15] See section on *duality* in Chapter 2.

A. Inselberg, *Parallel Coordinates*, DOI 10.1007/978-0-387-68628-8_3,
© Springer Science+Business Media, LLC 2009

- 20210921/Advanced Algebra.../
B035Y9Y7ZV.pdf
- 20210921/Robert Osserman/
Geometrie des Universums...pdf
- 20210921/Parallel Coordinates.../
Parallel Coordinates-Visual...pdf

Obs: Todo o livro!!!

Pág. 13(40) 2.2.2
Axioms for Projective Geometry
Pág. 18(45)
Figure 2.13
Pappus's theorem (left) and its dual (right).

Obs: Look

Pág. 20(47) Figure 2.14
A "picture" of the 7-point geometry

Obs: Utilizar Pappus's theorem (3 retas com 3 pontos cada).

$\overline{AGF}, \overline{DGC}, \overline{BGE}$

use Pappus's theorem

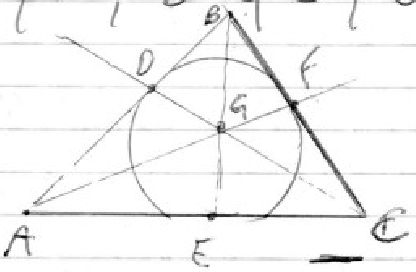

- Books-019-2021/20210921/
Pappus-Theorem/

- Revolutions of Geometry...pdf
Pag. 08(24) -

$\pi \approx 3\frac{13}{81} = 3.16049\ldots$

Obs: Livro todo!!!

- Projective Geometry...pdf
(coxeter).

Pag. 26 (38) Chapter 3
The Desargues configuration

Pag. 33 (45) Chapter 4
The Fundamental theorem and
Pappus's Theorem

Pag. 102 (114) Chapter 11 - Parallelism

.--? @).--?). D ( ) ( ) !33 6)

F)*+,)))- !) % + 8 5 3#
' %

degree. This was the case with Rhind's discovery. The scroll was 1 foot wide and 18 feet long but missing its middle portions. The content of the scroll that remained was seen to be written in the ancient Egyptian writing called **hieratic**. Beginning as a simple modification of hieroglyphics, hieratic later became a language in which each syllable is represented as a symbol and each symbol represents a particular idea. Because of this connection to hieroglyphics, the Rosetta Stone made possible the translation of the scroll that we know as the **Rhind Mathematical Papyrus** (Kline 1972, 15–16).

About four years after Rhind found his papyrus, the American Edwin Smith was studying in Egypt when he purchased what he believed to be a scroll on ancient Egyptian medicine. However, it was a fraud. Someone had taken pieces from a papyrus and created what looked like a scroll, but only the outside layers were ancient Egyptian. The inside was fake. Smith was an expert on Egypt, and it seems that it would be unlikely for someone in his position to be fooled by a forgery. We must remember, however, that the condition of any found papyri will be very fragile. It takes time and care to open an ancient scroll. Fortunately for posterity, Smith kept his find, and it went with the rest of his collection to the New York Historical Society after his death. In 1922 experts at the museum determined that the fraudulent scroll was assembled using the lost pieces of Rhind's papyrus! The pieces were shipped to the British Museum, where experts were able to fit together most of the missing section and complete its translation (Burton 1985, 36).

Other papyri discovered are the Golenischev Papyrus, usually called the **Moscow Papyrus** because it is now a part of the collection of the Museum of Fine Arts in Moscow, and a second purchase of Rhind's, which we know as the **Egyptian Mathematical Leather Scroll**. All of these finds have greatly aided our understanding of ancient Egyptian mathematics. The ancient Egyptian's wrote their mathematics in the form of questions followed by answers. The goal was to calculate values associated with government and commerce, specifically the computation of areas and volumes. They used specific numbers to solve problems, and they did not generalize their results. Any formulas that they did use were not deduced from first principles but were discovered empirically by trial and error. Because of this, some of their formulas were not accurate but only approximations, which, considering their needs, often sufficed. The greatest mystery regarding the mathematics of the ancient Egyptians is that for about two millennia they made very little progress beyond their initial work (Burton 1985, 36, 52, 57, 64).

Approximations

The mathematical scrolls show that the Egyptians did not distinguish between algebra and geometry. They viewed mathematics simply as a tool by which to obtain needed values. They had formulas for basic areas and volumes. This included the volumes of a cube and a cylinder. There is some debate in academic circles as to whether the Egyptians ever gave any justifications for any of the steps in their solutions. None have been discovered, but it is clear that the Egyptians did not consider geometry to be a logical system (Kline 1972, 19–20).

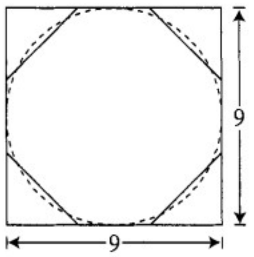

Figure 1.1 Using a 9 × 9 square to estimate the area of a circle.

The Rhind Mathematical Papyrus provides an example of an ancient Egyptian formula that did satisfactorily approximate its intended value. Take a square with sides of length 9. Trisect each of the sides, and then by joining consecutive points at the corners, create a regular octagon. This is illustrated in Figure 1.1, but we must note that the diagram found in the papyrus was a simple sketch of this diagram without the circle. The scribe of the papyrus considered the octagon a good estimate for the area of the circle. To find the octagon's area, we simply remove the four equilateral triangles from the square. Therefore, its area is

$$9^2 - 4 \cdot \frac{9}{2} = 63.$$

It is conjectured that this provided the ancient Egyptians a reason to believe that for a circle of diameter d,

$$A \approx \left(\frac{8d}{9}\right)^2. \tag{1.1}$$

The Egyptians could have relied on the octagonal estimate in support of Equation 1.1 because it yields an area of 64 square units when $d = 9$, which is close to the Rhind approximation. Furthermore, notice what Equation 1.1 implies about the Egyptian value of π. Since $A = \pi d^2/4$, we substitute Equation 1.1 to find

$$\pi \approx 3\frac{13}{81} = 3.16049\ldots. \tag{1.2}$$

This is very close to the common approximation of $3\frac{1}{7}$ (Burton 1985, 58).

An example of an equation from ancient Egypt that did not provide very good estimates is one that was found at the temple of Horus at Edfu. The priests of the temple received gifts in the form of patches of land. Each plot was in the form of a quadrilateral but was not necessarily rectangular. If we let a, b, c, and d represent the lengths of consecutive sides of the quadrilateral, their formula is

$$A = \frac{1}{4}(a + c)(b + d).$$

Notice that this amounts to multiplying the average of opposite sides together. It is not a bad estimate for regions that are nearly rectangular, but for most quadrilaterals it is an example of an instance when the Egyptians did try to generalize a result but did it poorly (Burton 1985, 57).

Pyramids

Herodotus recounts an encounter that he had with one of the priests (Burton 1985, 62–63). He claims that he was told that the Great Pyramid was designed so that the area of a square with side equal to the height of the pyramid is equal to the area of one of the faces of the pyramid. Using Figure 1.2, this is equivalent to writing

$$h^2 = \frac{1}{2}(2ba) = ab,$$

where $2b$ is the length of one of the sides of its square base, a is the altitude of one of the faces, and h is the height of the pyramid. From this we can calculate the ratio b/a by first noting that since

$$h^2 + b^2 = a^2,$$

we have

$$a^2 - b^2 = ab.$$

Dividing through by a^2 and a little rearranging yields

$$\left(\frac{b}{a}\right)^2 + \frac{b}{a} - 1 = 0.$$

This is a quadratic equation with b/a serving as the unknown. Thus, since both a and b are positive,

$$\frac{b}{a} = \frac{\sqrt{5} - 1}{2}. \tag{1.3}$$

Geometrically, a **pyramid** is a solid consisting of a collection of polygons joined at their edges. Recall that a **polygon** is a closed plane figure consisting of line segments called **sides**. An endpoint of a side is called a **vertex**. A polygon with n sides is often called an ***n*-gon**. If all the sides of the polygon are of the same magnitude and if the same can be said of the interior angles, the polygon is **regular**. In a pyramid one

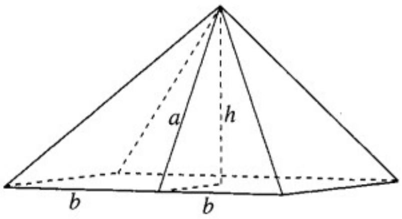

Figure 1.2 The Great Pyramid.

D)*+,)))- !) % F 8 2 5 *
& # -9 # ;! #4 D 3#

D)*+,)))- !) % F 8 2 5G #
A# 3#
' 1 / !33 ! # 7 !

EXERCISE

Dualize Theorem 4.31 and Figure 4.3B.

4.4 Pappus and Desargues

We are now ready to prove one of the oldest of all projective theorems. Pappus of Alexandria, living in the fourth century A.D., used a laborious development of Euclid's methods (see Reference **21**, pp. 286–290, and **8a**, pp. 66–69).

4.41 PAPPUS'S THEOREM: *If the six vertices of a hexagon lie alternately on two lines, the three pairs of opposite sides meet in collinear points.*

PROOF. Let the hexagon be $AB'CA'BC'$, as Figure 4.4A (which shows two of the many different ways in which it can be drawn). Since alternate vertices are collinear, there is a projectivity $ABC \barwedge A'B'C'$. The pairs of opposite sides, namely

$$B'C, BC'; \qquad C'A, CA'; \qquad A'B, AB',$$

are just the cross-joins of the pairs of corresponding points

$$BB', CC'; \qquad CC', AA'; \qquad AA', BB'.$$

By Theorem 4.31, their points of intersection

$$L = B'C \cdot BC', \qquad M = C'A \cdot CA', \qquad N = A'B \cdot AB'$$

all lie on the axis of the projectivity.

Alternatively, using further points

$$J = AB' \cdot CA', \qquad E = AB \cdot A'B', \qquad K = AC' \cdot CB',$$

we have

$$ANJB' \overset{A'}{\barwedge} ABCE \overset{C'}{\barwedge} KLCB'.$$

Thus B' is an invariant point of the projectivity $ANJ \barwedge KLC$. By Theorem 4.22, this projectivity is a perspectivity, namely

$$ANJ \overset{M}{\barwedge} KLC.$$

Thus M lies on NL; that is, L, M, N are collinear.

Similarly, since Theorem 4.22 is a consequence of the five axioms 2.13, 3.11, 3.12, 2.17, 2.18, we could have proved 2.32 (Desargues's theorem) as follows.

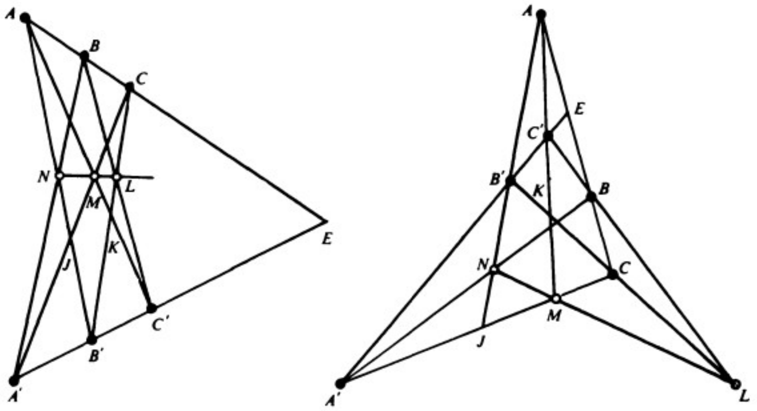

FIGURE 4.4A

Assuming that the lines PP', QQ', RR' all pass through O, as in Figure 2.3A, we define $D = QR \cdot Q'R'$, $E = RP \cdot R'P'$, $F = PQ \cdot P'Q'$ and three further points

$$A = OP \cdot DE, \qquad B = OQ \cdot DE, \qquad C = OR \cdot DE.$$

Then

$$OPAP' \overset{E}{\barwedge} ORCR' \overset{D}{\barwedge} OQBQ',$$

so that O is an invariant point of the projectivity $PAP' \barwedge QBQ'$. By 4.22, this projectivity is a perspectivity. Its center, F, lies on AB, which is DE; that is, D, E, F are collinear.

This procedure (see Reference **20**, p. 16) has the advantage of allowing us to omit 2.32 from the list of two-dimensional axioms proposed in Section 3.1. (The list is thus seen to be redundant: only five of the six are really needed.) It has the disadvantage of making Theorem 3.51 depend on Axiom 2.18, so that the theorem can no longer be used to make the axiom plausible!

We saw, in Section 3.2, that the figure for Desargues's theorem is a self-dual configuration 10_3. Somewhat analogously, the figure for Pappus's theorem is a self-dual configuration 9_3: nine points and nine lines, with three points on each line and three lines through each point. This fact becomes evident as soon as we have made the notation more symmetrical by calling the nine points

$$A_1 = A,\ \ B_1 = B,\ \ C_1 = C,\ \ A_2 = A',\ \ B_2 = B',\ \ C_2 = C',$$
$$A_3 = L,\ \ B_3 = M,\ \ C_3 = N$$

and the nine lines

$$a_1 = BL,\quad b_1 = AM,\quad c_1 = A'B',\quad a_2 = CM,\quad b_2 = CL,\quad c_2 = AB,$$
$$a_3 = AN,\quad b_3 = BN,\quad c_3 = LM.$$

The three triangles

$$A_1B_1C_2,\ A_2B_2C_3,\ A_3B_3C_1 \quad \text{or} \quad a_1b_1c_2,\ a_3b_3c_1,\ a_2b_2c_3$$

provide an instance of *Graves triangles:* a cycle of three triangles, each inscribed in the next. (See page 130 of Graves's paper mentioned in Exercise 2 of Section 3.2. This aspect of the Pappus configuration was rediscovered by G. Hessenberg.)

EXERCISES

1. Using the "symmetrical" notation for the Pappus configuration, prepare a table showing which are the three points on each line and which are the three lines through each point. Observe that the three points A_i, B_j, C_k are collinear (and the three lines a_i, b_j, c_k are concurrent) whenever $i + j + k$ is a multiple of three. (Reference **12a**, p. 108.)
2. Given a triangle $A_1A_2A_3$ and two points B_1, B_2, locate a point B_3 such that the three lines A_1B_1, A_2B_3, A_3B_2 are concurrent while also the three lines A_1B_3, A_2B_2, A_3B_1 are concurrent. Then also the three lines A_1B_2, A_2B_1, A_3B_3 are concurrent; in other words, *if two triangles are doubly perspective, they are triply perspective.* (Reference **19**, p. 100.)
3. What can be said about the three "diagonals" of the hexagon $A_1B_3A_2B_2A_3B_1$?
4. State the dual of Pappus's theorem and name the sides of the hexagon in the notation of Exercise 1.
5. Consider your solution to Exercise 2 of Section 4.2. Could this be developed into a proof of Pappus's theorem?
6. If one triangle is inscribed in another, any point on a side of the former can be used as a vertex of a third triangle which completes a cycle of Graves triangles (each inscribed in the next).
7. Assign the digits 0, 1, . . . , 8 to the nine points of the Pappus configuration in such a way that 801, 234, 567 are three triads of collinear points while 012, 345, 678 is a cycle of Graves triangles. (E. S. Bainbridge.*)

---

* In his answer to an examination question.

data

$\overline{AGF}, \overline{DGC}, \overline{BGE}$

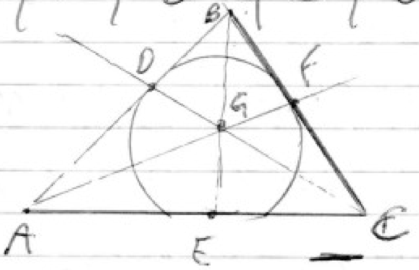

use Pappus's theorem

- Books-019-2021/20210921/
Pappus-Theorem/
- Revolutions of Geometry...pdf
Pág. 08(24) -
$\pi \approx 3\frac{13}{81} = 3.16049\ldots$
Obs: Livro todo!!!
- Projective Geometry...pdf
(coxeter).
Pág. 26 (38) Chapter 3
The Desargues configuration
Pág. 33 (45) Chapter 4
The Fundamental theorem and
Pappus's Theorem
Pág. 102 (114) Chapter 11 - Parallelism

CHAPTER ELEVEN

Parallelism

> Projective geometry was historically developed as a part of Euclidean geometry. . . . Von Staudt, for instance, still needed the parallel axiom in his study of the foundations of projective geometry. Klein, in his work on non-Euclidean geometry, established the independence of projective geometry from the theory of parallels (in 1871). This opened the possibility of an independent foundation for projective geometry. The pioneer work was done by Moritz Pasch (in 1882).
>
> *D. J. Struik* (1894–)
> (Reference **17**, p. 72)

11.1 Is the Circle a Conic?

The attentive reader must have noticed that most of the conics appearing in our figures look like something that has been familiar ever since he first saw the full moon: they look like *circles*. This observation raises the important question: *Is the circle a conic*? We can answer *Yes* as soon as we have found a characteristic property of a conic that is also a property of the familiar circle. One way to do this is to give a Euclidean definition for the pole-and-polar relation with respect to a circle, and carry the discussion far enough to find that this relation satisfies the projective definition for a polarity. A quicker way is to give a Euclidean proof for the Braikenridge-MacLaurin construction (Figure 9.2B). More precisely, we select five points on a circle and prove that the unique conic that can be drawn through these points coincides with the circle. For convenience we shall take the five points A, P, B, Q, C

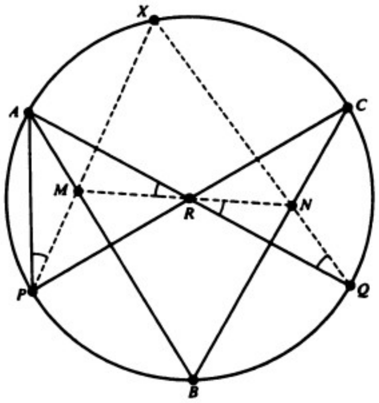

FIGURE 11.1A

(Figure 11.1A) to be five of the six vertices of a regular hexagon inscribed in the circle.

The following simple proof is due to S. L. Greitzer. Let AQ and CP meet in R (which is, of course, the center of the circle), let a variable diameter meet AB in M, BC in N, and let PM meet QN in X. Since AB, being a median of the equilateral triangle APR, is the perpendicular bisector of PR, and similarly BC is the perpendicular bisector of QR, we have

$$\angle XPA = \angle MPA = \angle ARM = \angle QRN = \angle NQR = \angle XQA.$$

By Euclid III.21 and 22, the locus of X is the circle APQ. By the Braikenridge-MacLaurin construction, the locus of X is the conic $APBQC$. Hence the conic coincides with the circle.

EXERCISES

1. What happens when the diameter is parallel to AB?
2. How does the proof have to be modified if A, P, B, Q, C are arbitrarily placed on the circle?

11.2 Affine Space

The time has come for us to investigate the connection between projective space, as determined by our axioms, and *affine* space, that is, the ordinary space of elementary solid geometry, in which two coplanar lines, or a line and a plane, or two planes, are said to be *parallel* if they do not meet. Instead of deliberately modifying the customary notions of space as we did in Section 1.1, let us now exploit the powerful theory of parallelism as Euclid did in his eleventh book. The idea of parallel lines leads immediately to that of a parallelogram, and thence to ratios of distances along parallel lines, or on one line (Reference **7**, pp. 2, 115–128; Reference **8**, pp. 175, 202, 222).

I. M. Yaglom, in the third volume of his *Geometric Transformations* (Reference **23**, pp. 10, 21), shows two sketches of a house. The former depicts the living-room on a sunny day. The window panes cast shadows on the floor. The rectangles are distorted into parallelograms, but those that are equal remain equal. Affine geometry deals with properties that are maintained by such "parallel projection." His other sketch is a night scene outside the house. A lamp inside casts shadows of the same window panes on the lawn. The rectangles are distorted into quadrangles of various sizes, but their sides are still straight. Projective geometry deals with properties that are maintained by such "central projection."

One way of expressing the connection is to regard affine space as *part of* projective space, namely, projective space minus one plane. For this purpose, we specialize any one plane of the projective space and call it "the plane at infinity." (It is still, of course, a *projective* plane.) Two lines, or a line and a plane, or two planes, are then said to be parallel if they meet on this special plane. We soon see that parallelism, so defined, has all its familiar properties.

Another way is to begin with the affine space, regarding its properties as known, and seeking certain figures which behave like the projective points and lines. Such a representation of the projective *plane* is easy: we can use the lines and planes through a fixed point O in the affine space. For instance, to verify 2.22 we merely have to observe that any two distinct planes α and β, through O, have just one common line, namely the line $\alpha \cdot \beta$. For projective *space*, the appropriate figures are bundles and axial pencils, defined as follows.

A *bundle* is the set of all lines and planes through a point.

An *axial pencil* is the set of all planes through a line.

When there is any possibility of confusion, the other kind of pencil (the set of all lines that lie in a plane and pass through a point) is called a *flat* pencil. In the terminology of Section 1.5, an axial pencil, like a flat pencil, is a "one-dimensional form." But a bundle is a combination of two two-dimensional forms: the set of *lines* through a point (which is the space-dual of the set of *lines* in a plane) and the set of *planes* through the same point (which is the space-dual of the set of *points* in a plane). What makes these forms useful in the present connection is that their description is precisely the same in affine space as in projective space.

Clearly, a range is determined by any two of its points, and these may be any two distinct points; a flat pencil is determined by any two of its lines, and these may be any two lines that meet; an axial pencil is determined by any two of its planes, and these may be any two planes that meet; finally, a bundle (like a flat pencil) is determined by any two of its lines.

EXERCISE

A range, a flat pencil, and an axial pencil are three kinds of one-dimensional form. Which two are space-duals of each other? Which one is its own dual (that is, which one is "self-dual")?

11.3 How Two Coplanar Lines Determine a Flat Pencil and a Bundle

If (in ordinary space) we are given a point P, and two coplanar lines a and b whose point of intersection O is inconveniently far away, how can we construct the line through P of the pencil or bundle determined by a and b? If O were available we could simply draw OP, but can we still locate this line without using O? If P is not in the plane ab, we merely have to draw the planes Pa and Pb; these meet in a line p through P, which is the desired line of the bundle (see Figure 11.3A). In a single symbol, the member through P (outside the plane ab) is the line

$$p = Pa \cdot Pb.$$

On the other hand, if P is in the plane ab, we can use an auxiliary point Q outside the plane, locate the member q through Q (which is the line OQ), and consider (as in Figure 11.3B) the line of intersection of the planes ab and Pq. In other words, the line through P (of the bundle or pencil) is now

$$p = ab \cdot Pq \quad \text{where} \quad q = Qa \cdot Qb.$$

This construction owes its importance to the fact that it remains valid when a and b are parallel, so that O does not exist! The lines a, b, q, p, which originally passed through O, are now all parallel.

A practical model for this new situation is easily made by dividing a rectangular card into four unequal strips (of widths roughly proportional to

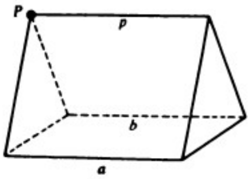

FIGURE 11.3A

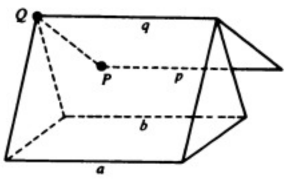

FIGURE 11.3B

4 : 6 : 7 : 5, as in Figure 11.3C) by three cuts made half-way through the thickness of the card. When suitably folded, this makes a solid version of Figure 11.3B.

Since we can derive p without inquiring whether a and b are parallel or not, we now feel justified in extending the meaning of the words *pencil* and *bundle* so as to allow the determining lines a and b to be *any two coplanar lines*. If a and b happen to be parallel, the bundle consists of all the lines and planes parallel to them, and the pencil consists of all the lines parallel to them in their own plane. Accordingly, we speak of a *bundle of parallels* and a (flat) *pencil of parallels*.

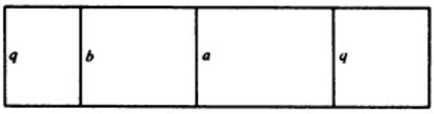

FIGURE 11.3C

It must be remembered that two planes may be parallel to a line without being parallel to each other. (For instance, in Figure 11.3B, the intersecting planes Qa and Qb are both parallel to the line p.) Thus a bundle of parallels contains a lot of lines, all parallel to one another, and a lot of planes, not all parallel to one another but each containing two (and therefore infinitely many) of the lines.

A familiar instance of a bundle of parallels is the set of all "vertical" lines and planes. If we take a cosmic standpoint and insist that two vertical lines are not strictly parallel but meet in the center of the earth, then we have an ordinary bundle instead of a bundle of parallels.

EXERCISE

Two lines parallel to the same line are parallel to each other. Does this remain true

(i) when the word "lines" is replaced by "planes,"
(ii) when the word "line" is replaced by "plane,"
(iii) when both substitutions are made simultaneously?

11.4 How Two Planes Determine an Axial Pencil

If we are given a point P, and two planes α and β whose line of intersection is far away, how can we construct the member through P of the axial pencil determined by α and β? This can be done by means of two applications of

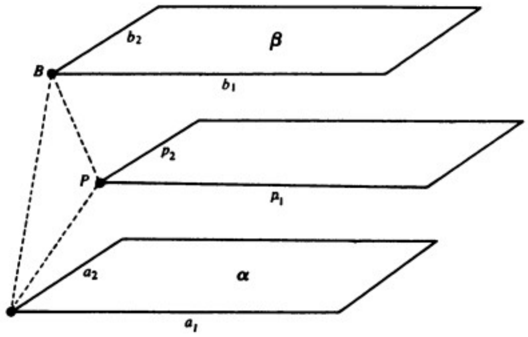

FIGURE 11.4A

the previous construction. We take any two intersecting lines a_1 and a_2 in α, and a point B in β (but not in either of the planes Pa_1, Pa_2), and draw the lines

$$b_1 = Ba_1 \cdot \beta, \qquad b_2 = Ba_2 \cdot \beta,$$
$$p_1 = Pa_1 \cdot Pb_1, \qquad p_2 = Pa_2 \cdot Pb_2,$$

as in Figure 11.4A. Then the desired plane through P is p_1p_2. For, if α and β meet in a line o, we may assume a_1 and a_2 to be chosen so as to meet o in two distinct points O_1 and O_2. Since $O_1 = o \cdot a_1$ lies in both the planes Ba_1 and β, it lies on their common line b_1. Since O_1 lies in both the planes Pa_1 and Pb_1, it lies on their common line p_1. Similarly O_2 lies on p_2. Therefore the join $o = O_1O_2$ lies in the plane p_1p_2.

If, on the other hand, α and β are parallel planes (conveyed by parallelograms in Figure 11.4A), the construction makes the lines b_1 and p_1 parallel to a_1, and the lines b_2 and p_2 parallel to a_2; therefore the plane p_1p_2 is parallel to α and β. Allowing P to take various positions, we thus obtain a *pencil of parallel planes*, consisting of all the planes parallel to a given plane.

A familiar instance is the set of all "horizontal"planes. If we insist that two horizontal planes are not strictly parallel but intersect in a line called the "horizon," then we have an ordinary axial pencil instead of a pencil of parallel planes.

EXERCISE

Can two lines be parallel to the same plane without being coplanar?

11.5 The Language of Pencils and Bundles

We have seen that a bundle can be derived from two of its lines, and an axial pencil from two of its planes, by constructions that remain valid when the two lines or planes are parallel.

Since an ordinary bundle consists of all the lines and planes through a point, and an ordinary axial pencil consists of all the planes through a line, any simple statement about points and lines can be "translated" into a corresponding statement about bundles and axial pencils. For instance, the statement

Any two distinct points lie on a line

becomes:

The common planes of any two distinct bundles form an axial pencil.

It is significant that the latter statement remains true when one of the bundles is replaced by a bundle of parallels, and again when *both* are replaced by bundles of parallels. In fact, the common planes of an ordinary bundle and a bundle of parallels form the ordinary axial pencil whose axis is the common line of the two bundles. (For instance, the bundle of lines and planes through any point O shares with the bundle of vertical lines and planes the vertical line through O and all the planes through this line.) Again, the common planes of two bundles of parallels form a pencil of parallel planes. (For instance, the bundle whose lines are horizontal in the north-south direction and the bundle whose lines are horizontal in the east-west direction have in common all the horizontal planes.)

EXERCISE

Translate the following statement into the language of pencils and bundles: *If two distinct lines have a common point they lie in a plane.*

11.6 The Plane at Infinity

These considerations serve to justify a convenient extension of space by the invention of an "ideal" plane whose points and lines represent the bundles of parallels and pencils of parallel planes, respectively. Remembering that an ordinary bundle consists of all the lines and planes through an ordinary point, we regard a bundle of parallels as consisting of all the lines and planes through an ideal point. Similarly, we regard a pencil of parallel planes as consisting of all the planes through an ideal line, and we say that an ideal point lies on an ideal line if the bundle contains the pencil. We can still assert that any two distinct points lie on a line. If one of the points is ordinary, so is the line; but if both are ideal, the line is ideal.

Since an ordinary bundle contains no pair of parallel planes, an ordinary point cannot lie on an ideal line; that is, all the "points" on an ideal line are ideal points. On the other hand, since a bundle of parallels contains ordinary axial pencils as well as pencils of parallel planes, an ideal point lies on some ordinary lines as well as on some ideal lines. Since any ordinary line belongs to just one bundle of parallels (consisting of all the lines and planes parallel to it), it contains just one ideal point, which we call its *point at infinity*. Thus we regard any two parallel lines as meeting in an ideal point: their common point at infinity. Since any plane belongs to just one pencil of parallel planes (consisting of all the planes parallel to it), it contains just one ideal line, which we call its *line at infinity*. Thus we regard any two parallel planes as meeting in an ideal line: their common line at infinity. In a given plane, each point on the line at infinity is the "center"of a pencil of parallel lines.

Since any two pencils of parallel planes belong to a bundle of parallels,

Any two ideal lines meet in an ideal point.

It follows that, if a and b are any two ideal lines, every other ideal line meets both a and b. This state of affairs resembles what happens in a plane. For, if a and b are two ordinary intersecting lines, every point in the plane ab lies on a line that meets both a and b. Accordingly, it is appropriate to regard the set of all ideal points and ideal lines as forming an ideal plane: *the plane at infinity*. This makes it possible to assert that any two intersecting (or parallel) lines determine a plane through both of them. If one of the lines is ordinary this is an ordinary plane; if both are ideal it is the plane at infinity.

Since each point (or line) at infinity is joined to an ordinary point O by an ordinary line (or plane), the points and lines of the projective *plane* may simply be regarded as a "new language" for the lines and planes (respectively) through O. In other words,

The projective plane is faithfully represented by a bundle.

Historically, points at infinity, lines at infinity, and the plane at infinity were first thought of by Kepler, Desargues, and Poncelet, respectively.

EXERCISE

Examine all the Axioms 2.11–2.18 for projective space, verifying that each is satisfied in our "extended" space, that is, in the ordinary affine space plus the plane at infinity with all its points and lines.

11.7 Euclidean Space

Although elementary solid geometry operates in affine space (Section 11.2), we must not imagine that affine geometry is merely another name for

Euclidean geometry! Affine geometry is the part of Euclidean geometry in which distances are compared only on the same line or on parallel lines.* Affine geometry becomes Euclidean geometry as soon as we have said what we mean by *perpendicular*. (For, right angles lead to circles and spheres, and thus enable us to compare distances.)

We have already referred to the set of all vertical lines as a familiar instance of a bundle of parallels, and to the set of all horizontal planes as a familiar instance of a pencil of parallel planes. More generally, every bundle of parallels in Euclidean space determines a unique axial pencil (of parallel planes), whose planes are perpendicular to the lines and planes of the bundle; and, conversely, every pencil of parallel planes determines a perpendicular bundle (of parallels). In the language of the plane at infinity, we thus have a special one-to-one correspondence between points at infinity and lines at infinity. As we have already remarked, the plane at infinity is a projective plane. Accordingly, we are not surprised to find that this correspondence between its points and lines is a polarity (called the *absolute polarity*). A line and a plane are perpendicular if the point at infinity on the line is the pole of the line at infinity in the plane. Two lines (or two planes) are perpendicular if their sections by the plane at infinity are conjugate points (or lines). Since no line or plane is perpendicular to itself, the polarity is elliptic. In fact, just as affine space can be derived from projective space by singling out a plane ("at infinity") and using it to define parallelism, so Euclidean space can be derived from affine space by singling out an elliptic polarity in the plane at infinity and using it to define perpendicularity.

EXERCISE

Let P and Q be two distinct points in Euclidean space. What is the locus of the point of intersection of a variable line through P and the perpendicular plane through Q?

---

* Melvin Hausner, "The Center of Mass and Affine Geometry," Am. Math. Monthly, **69** (1962), p. 730.

Pág. 24(36) - Chapter 3
The Principle of Duality

Obs: Pág. 30(42) - Exercises
2 → Locate the trilinear pole of S with respect to the triangle ABC.
(Figure 3.4A).

Look: Pág. 32(44) - Figure 3.5A
A' B' C' e D' = linha com 4 pontos
(Pappus's theorem)

Obs: Em seguida li todo o capítulo IV. →
The Fundamental Theorem and Pappus's Theorem

—

- 20210921| Pappus-Theorem|
A Modern View of Geometry... pdf
Pág. 97(97) - Chapter V
Coordinates in an Affine Plane with Desargues and Pappus Properties →

)*+,)))- !) % * # 2 5 3#
' (!3

V

Coordinates in an Affine Plane with Desargues and Pappus Properties

Foreword (The First Desargues Property)

For a reason that will be made clear later, an affine plane is said to have *the first Desargues property,* provided (whenever lines *AB*, *A′B′*, *A″B″* of the plane are pairwise mutually parallel) that when line *AA′* is parallel to line *BB′* and line *A′A″* is parallel to line *B′B″*, then line *A A″* is parallel to line *BB″*. Girard Desargues (1593–1662), the French architect-mathematician, was one of the first to study conic sections as projections of circles.

The property enunciated above is a special case of his beautiful theorem concerning perspective triangles to which we shall often refer. The example of an affine plane given in Section IV.11 shows that not every affine plane has the first Desargues property. Since this property is essential in order that the binary relation for vectors we are establishing be an equivalence relation, we restrict our attention from now on to those planes that, in addition to satisfying Postulates 1, 2, and 3, Section IV.1, have the first Desargues property. In these planes the equivalence relation for three non-null vectors with pairwise distinct carriers is transitive. The ordinary euclidean plane has the first Desargues property.

V.I. Completion of the Equivalence

Definition for Vectors

The definition of the desired relation is now completed by defining it for distinct vectors with the same carrier. The criterion is embodied in the following rule.

RULE 4. *If* $\overrightarrow{AB}$, $\overrightarrow{EF}$ *are distinct vectors with the same carrier,* $\overrightarrow{AB} = \overrightarrow{EF}$ *if and only if a vector* $\overrightarrow{CD}$ *exists, with carrier distinct from that of* $\overrightarrow{AB}$, *such that (under Rule 3)* $\overrightarrow{AB} = \overrightarrow{CD}$ *and* $\overrightarrow{CD} = \overrightarrow{EF}$.

Clearly, we wish the equivalence of the vectors $\overrightarrow{AB}$, $\overrightarrow{EF}$, with the same carrier, to be *independent* of the choice of the auxiliary vector $\overrightarrow{CD}$, so Rule 4 needs to be justified by showing that this is indeed the case.

Let $\overrightarrow{C'D'}$ be another choice of a vector, with carrier distinct from that of $\overrightarrow{AB}$, such that $\overrightarrow{AB} = \overrightarrow{C'D'}$ and $\overrightarrow{C'D'} = \overrightarrow{EG}$. Suppose $\overrightarrow{CD}$ and $\overrightarrow{C'D'}$ have different carriers (see Figure 17).

Since the carriers of vectors $\overrightarrow{AB}$, $\overrightarrow{CD}$, $\overrightarrow{C'D'}$ are pairwise distinct, and the plane has the first Desargues property, $\overrightarrow{AB} = \overrightarrow{C'D'}$, $\overrightarrow{AB} = \overrightarrow{CD}$ imply $\overrightarrow{CD} = \overrightarrow{C'D'}$. This, together with $\overrightarrow{CD} = \overrightarrow{EF}$, yields (for the same reason) $\overrightarrow{C'D'} = \overrightarrow{EF}$, and, consequently, line *D'F* is parallel to line *C'E*. But $\overrightarrow{C'D'} = \overrightarrow{EG}$ implies line *D'G* is parallel to line *C'E*. Hence line *D'F* is identical to line *D'G*, so *F* = *G*.

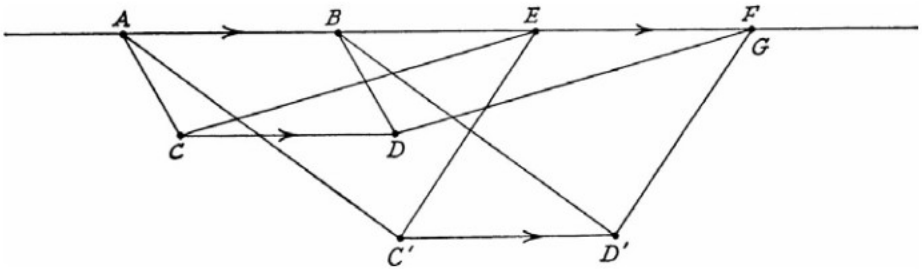

Figure 17

Thus, if $\overrightarrow{AB} = \overrightarrow{EF}$ follows from Rule 4, by use of a vector $\overrightarrow{CD}$ whose carrier is distinct from that of $\overrightarrow{AB}$, $\overrightarrow{AB} = \overrightarrow{EF}$ follows also, by use of *any* vector $\overrightarrow{C'D'}$ whose carrier is distinct from that of $\overrightarrow{CD}$ and such that $\overrightarrow{AB} = \overrightarrow{C'D'}$, $\overrightarrow{EF} = \overrightarrow{C'D'}$. How would the reader proceed in the event a vector $\overrightarrow{C'D'}$ has the same carrier as $\overrightarrow{CD}$?

● EXERCISES

1. Show that the binary relation defined by Rules 1–4 in the set of vectors of plane Π, with the first Desargues property is an equivalence relation (that is, is reflexive, symmetric, and transitive).

2. If $\overrightarrow{AA'} = \overrightarrow{BB'}$, $\overrightarrow{A'A''} = \overrightarrow{B'B''}$ (no two vectors having the same carriers), prove $\overrightarrow{AA''} = \overrightarrow{BB''}$. How is this property related to the first Desargues property?

3. Prove $\overrightarrow{AB} = \overrightarrow{AC}$ if and only if $B = C$.

V.2. Addition of Vectors

If $\overrightarrow{AB}$ is any vector and P is any point, there is exactly one point Q such that $\overrightarrow{AB} = \overrightarrow{PQ}$. There is at most one such point Q, since if $\overrightarrow{AB} = \overrightarrow{PQ_1}$ and $\overrightarrow{AB} = \overrightarrow{PQ_2}$, then $\overrightarrow{PQ_1} = \overrightarrow{PQ_2}$, by transitivity of vector equivalence, and $Q_1 = Q_2$ (Exercise 3, Section V.1) If $\overrightarrow{AB}$ is a null vector, $Q = P$ is the desired point (Rule 1, Section IV.10). If $\overrightarrow{AB}$ is not a null vector, suppose, first, P is not on the carrier of $\overrightarrow{AB}$. The line on P, parallel to line AB, intersects the line on B, parallel to line AP, in the desired point Q ($\overrightarrow{AB} = \overrightarrow{PQ}$, by Rule 3, Section IV.10). Finally, if P is on the line AB, let $\overrightarrow{CD}$ be any vector such that $\overrightarrow{AB} = \overrightarrow{CD}$ and line AB is distinct from line CD. The line on D, parallel to line CP, meets line AB in the desired point Q ($\overrightarrow{AB} = \overrightarrow{PQ}$, by Rule 4, Section V.1).

DEFINITION (VECTOR ADDITION). *Let* $\overrightarrow{AB}$, $\overrightarrow{CD}$ *be any two vectors. If* $B = C$, *then* $\overrightarrow{AB} + \overrightarrow{CD} = \overrightarrow{AD}$. *If* $B \neq C$, *and* Q *denotes the unique point such that* $\overrightarrow{BQ} = \overrightarrow{CD}$, *then* $\overrightarrow{AB} + \overrightarrow{CD} =$
$\overrightarrow{AB} + \overrightarrow{CD} = \overrightarrow{AQ}$.

Remark 1. If $\overrightarrow{CD} = \overrightarrow{C'D'}$, $\overrightarrow{AB} + \overrightarrow{CD} = \overrightarrow{AB} + \overrightarrow{C'D'}$. For, by definition of vector addition, $\overrightarrow{AB} + \overrightarrow{CD} = \overrightarrow{AQ} = \overrightarrow{AB} + \overrightarrow{BQ}$, where $\overrightarrow{BQ} = \overrightarrow{CD}$. Then $\overrightarrow{BQ} = \overrightarrow{C'D'}$, and $\overrightarrow{AB} + \overrightarrow{C'D'} = \overrightarrow{AB} + \overrightarrow{BQ} = \overrightarrow{AQ}$.

Remark 2. If $\overrightarrow{CD} = \overrightarrow{C'D'}$, $\overrightarrow{CD} + \overrightarrow{AB} = \overrightarrow{C'D'} + \overrightarrow{AB}$. For $\overrightarrow{CD} + \overrightarrow{AB} = \overrightarrow{CD} + \overrightarrow{DQ} = \overrightarrow{CQ}$, where $\overrightarrow{DQ} = \overrightarrow{AB}$ and $\overrightarrow{C'D'} + \overrightarrow{AB} = \overrightarrow{C'D'} + \overrightarrow{D'Q'} = \overrightarrow{C'Q'}$, where $\overrightarrow{D'Q'} = \overrightarrow{AB}$. Hence $\overrightarrow{DQ} =$

$\overrightarrow{D'Q'}$, which, together with $\overrightarrow{CD} = \overrightarrow{C'D'}$, gives $\overrightarrow{CQ} = \overrightarrow{C'Q'}$. This is justified by Exercise 2, Section V.1, if the carriers of $\overrightarrow{CD}$, $\overrightarrow{C'D'}$, $\overrightarrow{DQ}$, $\overrightarrow{D'Q'}$ are pairwise distinct. Let the reader check the other possibility.

Remark 3. If $\overrightarrow{AB} = \overrightarrow{A'B'}$ and $\overrightarrow{CD} = \overrightarrow{C'D'}$, $\overrightarrow{AB} + \overrightarrow{CD} = \overrightarrow{A'B'} + \overrightarrow{C'D'}$. For, by Remark 2, $\overrightarrow{AB} + \overrightarrow{CD} = \overrightarrow{A'B'} + \overrightarrow{CD}$, and, by Remark 1, $\overrightarrow{A'B'} + \overrightarrow{CD} = \overrightarrow{A'B'} + \overrightarrow{C'D'}$. Transitivity of vector equivalence gives $\overrightarrow{AB} + \overrightarrow{CD} = \overrightarrow{A'B'} + \overrightarrow{C'D'}$.

THEOREM V.2.1. *Vector addition is associative.*

Proof. If $\overrightarrow{AB}$, $\overrightarrow{CD}$, $\overrightarrow{EF}$ are any three vectors, consider the unique vectors $\overrightarrow{BP}$, $\overrightarrow{PQ}$ such that $\overrightarrow{BP} = \overrightarrow{CD}$ and $\overrightarrow{PQ} = \overrightarrow{EF}$. Then

$$\begin{aligned}(\overrightarrow{AB} + \overrightarrow{CD}) + \overrightarrow{EF} &= (\overrightarrow{AB} + \overrightarrow{BP}) + \overrightarrow{PQ} \qquad \text{(Remark 3)}\\ &= \overrightarrow{AP} + \overrightarrow{PQ} \qquad \text{(definition of vector addition)}\\ &= \overrightarrow{AQ} \qquad \text{(definition of vector addition)},\end{aligned}$$

$$\begin{aligned}\overrightarrow{AB} + (\overrightarrow{CD} + \overrightarrow{EF}) &= \overrightarrow{AB} + (\overrightarrow{BP} + \overrightarrow{PQ})\\ &= \overrightarrow{AB} + \overrightarrow{BQ}\\ &= \overrightarrow{AQ},\end{aligned}$$

and, consequently, $(\overrightarrow{AB} + \overrightarrow{CD}) + \overrightarrow{EF} = \overrightarrow{AB} + (\overrightarrow{CD} + \overrightarrow{EF})$.

THEOREM V.2.2. *Vector addition is commutative.*

Proof. Let $\overrightarrow{AB}$, $\overrightarrow{CD}$ be any two vectors, and let $\overrightarrow{BQ} = \overrightarrow{CD}$. Then $\overrightarrow{AB} + \overrightarrow{CD} = \overrightarrow{AQ}$. Suppose $\overrightarrow{AB}$, $\overrightarrow{BQ}$ have distinct carriers, and let the line on *A*, parallel to line *BQ*, meet the line on *Q*, parallel to line *AB*, at *P*. Then $\overrightarrow{CD} = \overrightarrow{BQ} = \overrightarrow{AP}$, and $\overrightarrow{AB} = \overrightarrow{PQ}$ yields $\overrightarrow{CD} + \overrightarrow{AB} = \overrightarrow{AP} + \overrightarrow{PQ} = \overrightarrow{AQ}$.

If, on the other hand, $\overrightarrow{AB}$, $\overrightarrow{BQ}$ have the same carrier, let *S* be a point of Π not on line *AB* (Postulate 3, Section IV.1). Then

$$\begin{aligned}
\overrightarrow{CD} + \overrightarrow{AB} &= \overrightarrow{BQ} + \overrightarrow{AB} = (\overrightarrow{BS} + \overrightarrow{SQ}) + \overrightarrow{AB} \\
&= \overrightarrow{BS} + (\overrightarrow{SQ} + \overrightarrow{AB}) \quad \text{(Theorem V.2.1)} \\
&= \overrightarrow{BS} + (\overrightarrow{AB} + \overrightarrow{SQ}) \quad \text{(since } \overrightarrow{AB}, \overrightarrow{SQ} \text{ have different} \\
&\qquad\qquad\qquad\qquad\qquad \text{carriers)} \\
&= (\overrightarrow{BS} + \overrightarrow{AB}) + \overrightarrow{SQ} \\
&= (\overrightarrow{AB} + \overrightarrow{BS}) + \overrightarrow{SQ} = \overrightarrow{AS} + \overrightarrow{SQ} = \overrightarrow{AQ},
\end{aligned}$$

and the proof is complete.

These two theorems have important consequences in regard to the algebraic structure of the ternary ring [Γ, *T*]. A one-to-one correspondence between the set of vectors $\overrightarrow{OX}$ and the points *X* of the unit line is obtained by associating with each such vector its terminal point *X* (the origin *O* is associated with the null vector $\overrightarrow{OO}$). This one-to-one correspondence, together with the one-to-one correspondence *γ* between the points of the unit line and the elements of the coordinate set Γ (see Section IV.3) establishes a one-to-one correspondence between the vectors $\overrightarrow{OX}$, *X* ∈ line *OI*, and the elements of Γ, in which $\overrightarrow{OA}$ corresponds to *a*, where $a = \gamma(A) \in \Gamma$.

If *A*(*a*, *a*), *B*(*b*, *b*) ∈ line *OI*, and *C* = *A* + *B* (see Figure 11), $\overrightarrow{OA} = \overrightarrow{UP} = \overrightarrow{BC}$, so $\overrightarrow{OA} + \overrightarrow{OB} = \overrightarrow{BC} + \overrightarrow{OB} = \overrightarrow{OB} + \overrightarrow{BC} = \overrightarrow{OC}$. Hence if *A*, *B* ∈ line *OI* that correspond, respectively, to vectors $\overrightarrow{OA}$, $\overrightarrow{OB}$, their sum *A* + *B* corresponds to the vector sum $\overrightarrow{OA} + \overrightarrow{OB}$, and, consequently, if *a*, *b* ∈ Γ that correspond, respectively, to points A, B (and hence to vectors $\overrightarrow{OA}$, $\overrightarrow{OB}$), their sum $a + b$ corresponds to $C = A + B$, and hence to $\overrightarrow{OC} = \overrightarrow{OA} + \overrightarrow{OB}$. Thus, the vector that corresponds to the sum of a and b is the sum of the vector corresponding to a and the vector corresponding to b. We conclude that this correspondence between vectors $\overrightarrow{OX}$, $X \in$ line OI, and elements x of Γ is an *isomorphism with respect to addition*. Since vector addition is associative and commutative, it follows that *addition in* $[\Gamma, T]$ *is associative and commutative also*. Suppose $a, b \in \Gamma$ and $a \sim \overrightarrow{OA}$, $b \sim \overrightarrow{OB}$ in the correspondence defined above (where the symbol ~ is read "corresponds to"). Then $a + b \sim \overrightarrow{OA} + \overrightarrow{OB}$, and $b + a \sim \overrightarrow{OB} + \overrightarrow{OA}$, and, since $\overrightarrow{OA} + \overrightarrow{OB}$ and $\overrightarrow{OB} + \overrightarrow{OA}$ are the *same* vector $\overrightarrow{OC}$, and the correspondence is one-to-one, $a + b = b + a$. In a similar manner addition in $[\Gamma, T]$ is shown to be associative.

We have proved that with respect to ordered addition (Remark 2, Section IV.8) $[\Gamma, T]$ is a loop. Imposing the first Desargues property on plane Π results in associativity of addition in $[\Gamma, T]$, hence $[\Gamma, +]$ is a *group* (a loop with associativity of the operation). Since addition is, moreover, commutative, the group $[\Gamma, +]$ is an *abelian* (that is, commutative) group.

We have now established the following important result.

THEOREM V.2.3. *If a rudimentary affine plane* Π *has the first Desargues property, every planar ternary ring* $[\Gamma, T]$ *defined over* Π *is an abelian group with respect to addition.*

V.3. Linearity of the Ternary Operator

Without assuming the affine plane Π to have the first Desargues property, we have seen that certain classes of lines of Π are represented by equations having the same forms as in ordinary analytic geometry. Thus lines in the parallel class of the y-line have equations $x = a$, $a \in \Gamma$; those in the parallel class of the x-line have equations $y = b$, $b \in \Gamma$; those in the parallel class of the unit line have equations $y = x + c$, $c \in \Gamma$; and those lines on O with slope m, $m \in \Gamma$, have equations $y = xm$ (where the dot indicating multiplication has been omitted). Now it will be established that, as a consequence of assuming the first Desargues property valid in II, each line of Π with slope m and y-intercept b has an equation $y = xm + b$, as in ordinary analytic geometry.

Consider the line with equation $y = T(x, m, b)$, where $m \neq 0, 1$ and $b \neq 0$ (Figure 18).

The two triples A^*, A', A'' and B^*, B', B'' of Figure 18 are such that: (1) the six points are pairwise distinct, (2) lines A^*B^*, $A'B'$, $A''B''$ are pairwise mutually parallel, and (3) line A^*A' is parallel to line B^*B', and line A^*A'' is parallel to line B^*B''. It follows from the first Desargues property that line $A'A''$ is parallel to line $B'B''$, and hence line $B'B''$ is in the parallel class of the x-line. Consequently, points B', B' have the same ordinates, so $T(a, m, b) = am + b$, $a \neq 0$. By Section IV.6, $T(0, m, b) = b = T(a, 0, b)$; by Section IV.8, $T(a, 1, b) = a + b$; and by Section IV.9, $T(a, m, 0) = am$. Hence for every triple a, m, $b \in \Gamma$, $T(a, m, b) = am + b$, and each line with slope m and y-intercept b has equation $y = T(x, m, b) = xm + b$. Thus $T(x, m, b)$ is linear.

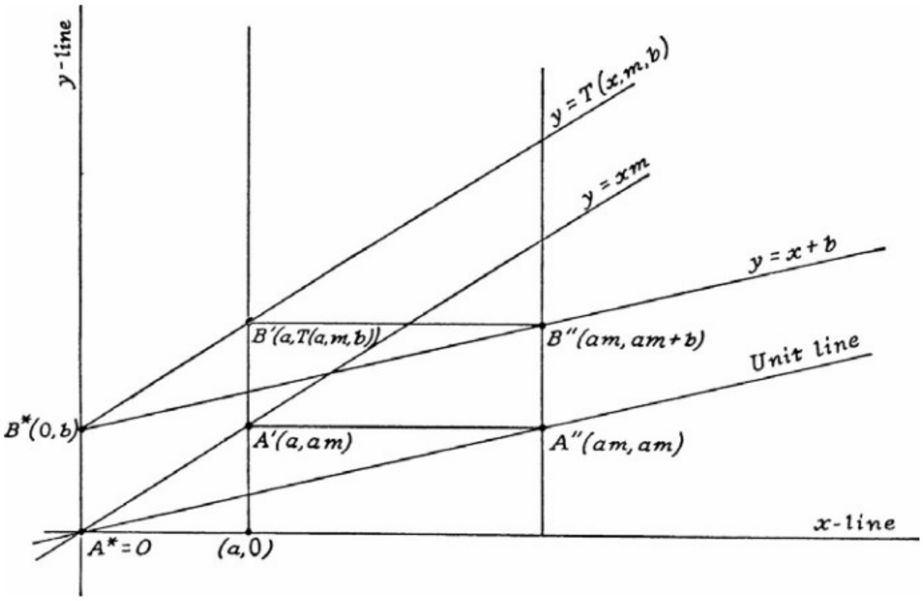

Figure 18

Since $x \cdot y = T(x, y, 0)$, $t + z = T(t, 1, z)$ $(x, y, z, t \in \Gamma)$, it follows that $T(x, y, z)=xy + z= T(xy, 1, z) = T[T(x, y, 0), 1, z]$.

We recall that a coordinate system was established in the affine plane Π by selecting an arbitrary point O of Π for the origin and by choosing three pairwise distinct lines on O and designating them the x-line, the y-line, and the unit line. On the unit line any point distinct from O was chosen and labeled I, and an arbitrary abstract set Γ was selected (subject to the sole requirement that there is a one-to-one correspondence γ between the points of the line OI and the elements of Γ), and $\gamma(O)$, $\gamma(I)$ were labeled 0, 1, respectively. Then a procedure was defined that established a one-to-one correspondence between the points of Π and the elements of the set of all ordered pairs of elements of Γ. The ordered pair of elements of Γ corresponding to a point P of Π are the coordinates of P. A ternary operation $T(a, m, b)$ was defined in Γ, and the system $[\Gamma, T]$ was called a *planar ternary ring*.

Now if a different point O' of Π were chosen for the origin, and three lines on O' selected as the x'-line, y'-line, and unit line (with point I' selected on it), the same set Γ could serve as the coordinate set, since there exists a one-to-one correspondence between the elements of Γ and the points of the new unit line [each line of Π contains the same number (finite or transfinite) of points]. But the ternary operation T', defined by the new lines and new origin, would, in general, be quite different from the ternary operation T defined by the original coordinate system (though the *domain* of the two operators is the same set Γ), so a different planar ternary ring $[\Gamma, T']$ would result. If the first Desargues property is assumed for Π, both operators T, T' are linear; that is $T(x, y, z) = xy + z$, $T'(x, y, z) = [x \odot y] \oplus z$, where the addition $\oplus$ and the multiplication $\odot$ are, in general, different from the addition $+$ and multiplication $\cdot$ defined by using the original coordinate system. We have established the following result.

THEOREM V.3.1. *If the first Desargues property is valid in the affine plane* Π, *every* (*planar*) *ternary ring defined in* Π *is linear*.

Is the converse of this theorem valid; that is, if every (planar) ternary ring defined in the affine plane Π is linear, does it follow that the first Desargues property is valid in Π? The answer is affirmative, as we shall now show.

THEOREM V.3.2. *If every* (*planar*) *ternary ring defined in an affine plane* Π *is linear,* Π *has the first Desargues property*.

Proof. Let A^*, A', A'' and B^* B', B'' be any two triples of points of Π (each two points distinct) such that lines A^*B^*, $A'B'$, $A''B''$ are pairwise mutually parallel and lines A^*A', $A'A''$ are parallel, respectively, to lines B^*B', $B'B''$. We wish to prove that line A^*A'' is parallel to line B^*B''.

We introduce in Π a coordinate system in the following manner. Select A' as the origin, line $A'B'$ as the y-line, the line on A' parallel to line A^*A'' as the x-line, and line $A'A''$ as the unit line (Figure 19).

Then line $A'A^*$ has equation $y = xm$, and line $B'B^*$ has equation $y = T(x, m, b)$ for some element b of Γ (since it is parallel to line $A'A^*$). Since line $B'B''$ is parallel to the unit line, it has equation $y = x + b$. Point A^* has coordinates (a, am) for some element a of Γ, so point A'' has coordinates (am, am), and the coordinates of B^* and B'' are $(a, T(a, m, b))$ and $(am, am + b)$, respectively. But since every ternary ring defined in Π is linear, $T(a, m, b) = am + b$, and, consequently, points B^* and B'' have the same ordinates. It follows that line B^*B'' is in the parallel class of the x-line, and since line A^*A'' is also in that parallel class, and is distinct from line B^*B'', it follows that line A^*A'' is parallel to line B^*B''.

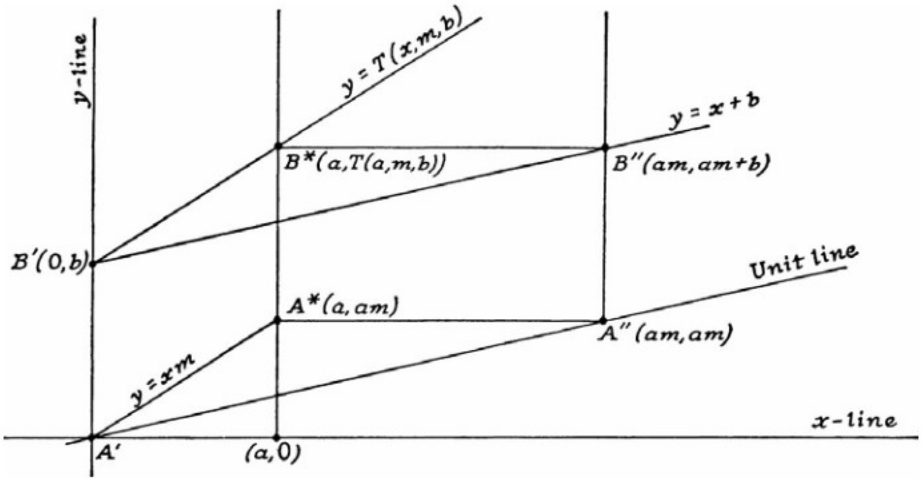

Figure 19

Combining the last two theorems yields the following important result.

THEOREM V.3.3. *The first Desargues property is valid in an affine plane Π if and only if every (planar) ternary ring defined in Π is linear.*

The (geometric) first Desargues property has, then, the algebraic counterpart or *equivalent* that every ternary ring of Π is

Pág. 24(36) - Chapter 3
The Principle of Duality

Obs: Pág. 30(42) - Exercises
2 → Locate the trilinear pole of 5 with respect to the triangle ABC.
(Figure 3.4A).

Look: Pág. 32(44) - Figure 3.5A
A' B' C' e D' = linha com 4 pontos
(Pappus's theorem)

Obs: Em seguida leia todo o capítulo IV. →
The Fundamental Theorem and Pappus's Theorem

—

- 20210921| Pappus_Theorem|
A Modern View of Geometry...pdf
Pág. 97(97) - Chapter V
Coordinates in an Affine Plane with Desargues and Pappus Properties →

Affine Plane (incidence geometry)

Satisfy the following axioms:

- Any two distinct points lie on a unique line.
- Each line has at least two points.
- Given any line and any point not on that line, there is a unique line which contains the point and does not meet the given line. (Playfair's axiom)
- There exist three non-collinear points (points not on a single line)

In an affine plane, two lines are called parallel if they are equal or disjoint. (disjoint set)

Using this definition, Playfair's axiom above can be replaced by:

- Given a point and a line, there is a unique line which contains the point and is parallel to the line.

Pág. 97 an affine plane is said to have the first Desargues property → parallel.

- Pappus-Theorem/
Affine and Projective Geometry...pdf
Pág. 47(64) - Chapter 3
Desarguesian Affine Plane
Obs: Pappus's Theorem — theorem 12 of Chapter 1.

Obs: Livro todo!!!

- Pappus-Theorem/
Perspectives on Projective Geometry...pdf
(Jürgen Richter-Gebert)

Obs: Important!!!
Livro todo!

- Pappus-Theorem/
Projective Geometry....pdf
(Casse)

Obs: Livro todo!!!

Pág. 22(35) - Chapter 3
Introduction to axiomatic geometry.

→

)*+,)))- !)% * !# F 8 2 53#
' % "% !3 1 7 ! ! * !

CHAPTER 3

Desarguesian Affine Planes

The Greek geometers introduced "geometric algebra" by adding and multiplying lengths of segments in the Euclidean plane. Descartes formalized geometric algebra when he represented the synthetic Euclidean plane as the real coordinate plane. In particular, the points on the x- and y-axes are in one-one correspondence with the real numbers; these points can be added and multiplied geometrically. Addition is done by parallel displacement, while multiplication uses proportional segments. In this chapter we will define and investigate the properties of addition and multiplication in affine planes.

3.1 THE FUNDAMENTAL THEOREM

For a well-defined addition to satisfy the associative and commutative laws, and for a well-defined multiplication to satisfy the associative law, Desargues' Theorem (Theorems 13 and 14 of Chapter 1) is needed. For multiplication to be commutative, Pappus' Theorem (Theorem 12 of Chapter 1) is needed. The statements of the theorems of Desargues and Pappus are meaningful in an arbitrary affine plane because they only involve parallelism. The proofs of these results, as given in Chapter 1, however, rely on concepts (e.g., congruence and similar triangles) relevant to Euclidean geometry but meaningless in arbitrary affine planes.

Thus this chapter concerns affine planes in which Desargues' Theorem is true. Its aim is to define addition and multiplication of points on a given line, relative to fixed units O and I, and to prove the following theorem.

■ **THEOREM 1. The fundamental theorem of synthetic affine planes.** Relative to two fixed points, O and I, any line in a Desarguesian affine plane is a *division ring* in that its addition and multiplication satisfy the following properties:

Addition	**Multiplication**
Associative	Associative
O is the identity	I is the identity
All elements have inverses	All nonzero elements have inverses
Commutative	

Multiplication distributes over addition

A division ring in which multiplication is commutative is a *field*, as defined in Chapter 1.

3.2 ADDITION ON LINES

An addition can be defined on the points of any line through any fixed point O, called the *origin*, and is defined *relative to this* O. In Chapter 4 the effect of selecting different origins will be investigated, and comparisons of sums of points relative to distinct origins will be made.

Definition of addition. In any affine plane $(\mathscr{P}, \mathscr{L})$ select and fix a line $\ell \in \mathscr{L}$, and a point $O \in \ell$. (See Fig. 3.1.) For points A and C in ℓ, to add A and C (relative to O):

1. Select B not in ℓ.
2. Let m be the line through B and parallel to ℓ.
3. Let k be the line through A parallel or equal to $\ell(O, B)$.
4. Let $D = m \cap k$.
5. Let n be the line through D parallel or equal to $\ell(B, C)$.
6. $A + C$ is defined to be $n \cap \ell$.

It will be shown in Section 3.4 that the addition of points on ℓ is independent of the choice of the fixed point B.

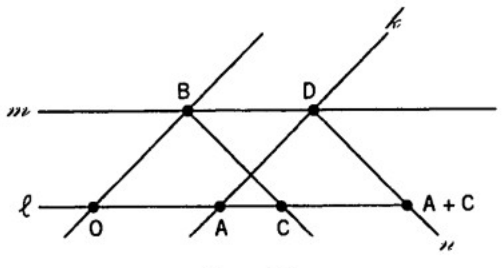

Figure 3.1

If A happens to be the special point O, then line k, as defined in step 3 is $\ell(\mathrm{O}, \mathrm{B})$. The point D defined in step 4 always exists. Otherwise, m and k would be parallel, giving two lines (ℓ and k) through A and parallel to m, whereby $\ell = k$. But ℓ and k are distinct, for k is parallel or equal to $\ell(\mathrm{O}, \mathrm{B})$, and ℓ is neither parallel nor equal to $\ell(\mathrm{O}, \mathrm{B})$. Similarly n is not parallel to ℓ. Thus A + C as defined in step 6 exists.

Parallelograms

Parallelograms, familiar in Euclidean geometry, can also be defined in a general affine plane. A *parallelogram* is an ordered quadruple (A, B, C, D) of points such that $\ell(\mathrm{A}, \mathrm{B}) \| \ell(\mathrm{C}, \mathrm{D})$ and $\ell(\mathrm{B}, \mathrm{C}) \| \ell(\mathrm{A}, \mathrm{D})$. The lines $\ell(\mathrm{A}, \mathrm{C})$ and $\ell(\mathrm{B}, \mathrm{D})$ are called its *diagonals*. Two parallelograms arise in the preceding definition of addition of points, namely (O, A, D, B) and (C, A + C, D, B) which share the common vertices D and B. The definition of addition given here does, in fact, generalize addition on the (real) number line. In that case the *length* of the segment OA is added to the length of OC by moving OA by *parallel displacement*, first to BD and then to C(A + C).

■ **THEOREM 2.** O + C = C for any C in ℓ.

Proof: In the definition above A = O. Therefore, since $k = \ell(\mathrm{O}, \mathrm{B})$ and B = D, we have $n = \ell(\mathrm{B}, \mathrm{C})$ and O + C = C. ■

■ **THEOREM 3.** A + O = A for any A in ℓ.

Proof: Here $\ell(\mathrm{B}, \mathrm{C}) = \ell(\mathrm{B}, \mathrm{O})$, so $n = k$. Since $\mathrm{A} \in k$, $k \cap \ell = \mathrm{A}$, and A + O = A. ■

Either Theorem 2 or 3 implies the following result.

Corollary. O + O = O.

Clearly the fixed origin O acts as a "zero" or additive identity. However, in order that addition satisfy various other familiar properties such as commutativity and associativity, we need to assume *as an axiom* that Desargues' Theorem holds. This is the subject of the next section.

EXERCISES

1. Let $(\mathcal{P}, \mathcal{L})$ be the affine plane of order 3 defined in Example 7 of Chapter 2. On line {GHI} set G = O, and add H + I, using B as the point B was used in the definition of addition. Now compute H + I using whatever point you please for B.

2. Let $(\mathscr{P}, \mathscr{L})$ be the affine plane of order 3 as in Exercise 1. Make an addition table for {AEI}, setting A = O.

+	O	E	I
O			
E			
I			

3. Suppose that A, B, and C are noncollinear points in an affine plane $(\mathscr{P}, \mathscr{L})$. Prove that there is a unique point D in $\mathscr{P}$ such that (A, B, C, D) is a parallelogram.
4. Let $(\mathscr{P}, \mathscr{L})$ be the real Euclidean plane described in Example 1 of Chapter 2. Prove that $((a, b), (c, d), (e, f), (g, h))$ is a parallelogram if and only if $(a + e, b + f) = (c + g, d + h)$.
5. Make an addition table for the line {OAJH} in the plane of order 4, given in Example 8 of Chapter 2.

3.3 DESARGUES' THEOREM

In 1902 F. R. Moulton gave a simple example of an affine plane $(\mathscr{P}, \mathscr{L})$ called the *Moulton Plane*, in which Desargues' Theorem fails to hold. In this plane $\mathscr{P}$ is the set of ordered pairs of real numbers (i.e., the points of the real coordinate plane). $\mathscr{L}$ consists of all Euclidean horizontal lines, vertical lines, and lines with negative slope, together with all "broken"

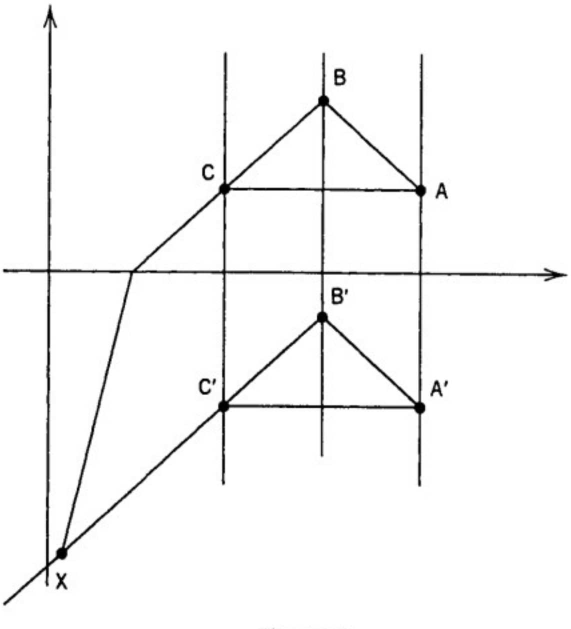

Figure 3.2

Euclidean lines that have positive slope m above the x-axis and slope $2m$ below the axis. It is an exercise in analytic geometry to show that any two points of the Moulton plane lie on one and only one (Moulton) line by finding its equation if it is a Euclidean line, or the point of intersection of the line with the x-axis if it is a broken line. Horizontal lines are parallel, vertical lines are parallel, and lines of the same negative slope are parallel. Broken lines whose slopes agree both above and below the x-axis are parallel as well, so it is apparent that the parallel axiom holds in the Moulton plane. Axiom $A3$ clearly holds. The easiest way to see that the Moulton plane does not satisfy Desargues' Theorem is to look at the example illustrated in Fig. 3.2 where the hypotheses of Desargues' Theorem are satisfied, but $\ell(\mathrm{B},\mathrm{C}) \cap \ell(\mathrm{B}',\mathrm{C}') = \mathrm{X}$.

Thanks to Moulton's example and others, it is clear that if Desargues' Theorem is to be used in affine plane geometry, it must be assumed as an additional axiom. Thus we make the following definition

Definition. A *Desarguesian affine* plane is an affine plane in which the following axiom holds (see Fig. 3.3):

$A4$. Desargues' Theorem.

(I) Suppose that $\ell(\mathrm{A},\mathrm{A}') \| \ell(\mathrm{B},\mathrm{B}') \| \ell(\mathrm{C},\mathrm{C}')$. If $\ell(\mathrm{A},\mathrm{B}) \| \ell(\mathrm{A}',\mathrm{B}')$, and $\ell(\mathrm{A},\mathrm{C}) \| \ell(\mathrm{A}',\mathrm{C}')$, then $\ell(\mathrm{B},\mathrm{C}) \| \ell(\mathrm{B}',\mathrm{C}')$.

(II) Suppose $\ell(\mathrm{A},\mathrm{A}') \cap \ell(\mathrm{B},\mathrm{B}') \cap \ell(\mathrm{C},\mathrm{C}') = \mathrm{P}$. If $\ell(\mathrm{A},\mathrm{B}) \| \ell(\mathrm{A}',\mathrm{B}')$ and $\ell(\mathrm{A},\mathrm{C}) \| \ell(\mathrm{A}',\mathrm{C}')$, then $\ell(\mathrm{B},\mathrm{C}) \| \ell(\mathrm{B}',\mathrm{C}')$.

The rest of this chapter treats Desarguesian planes, while in Chapter 5 we will give some examples of finite non-Desarguesian planes.

Desargues' Theorem is named for the French mathematician and architect Girard Desargues (1593–1662), who first proved it. A contemporary and friend of Descartes, he also designed part of the current city hall in Lyon, France. As an affine axiom, Desargues' Theorem differs from

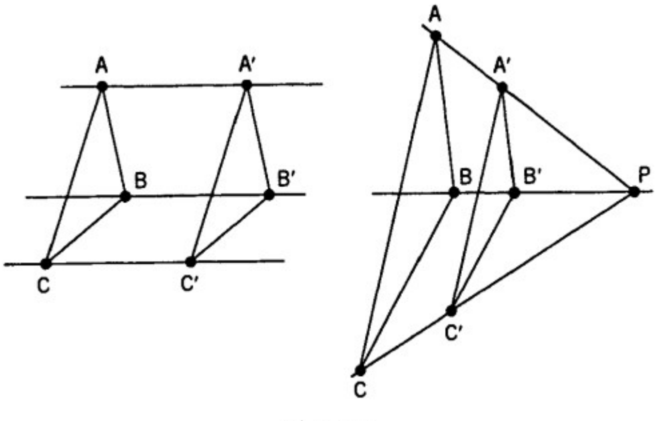

Figure 3.3

the first three in several ways. Clearly its statement is more complicated than the statements of the first three axioms, and in fact has two parts. Additionally it is equivalent to its own converse, which will be shown in Section 3.5. In Chapter 1 Desargues' Theorem was proved synthetically in the Euclidean plane; the following theorem shows that it holds in the coordinate real and rational planes.

■ **THEOREM 4.** The real and rational coordinate planes are Desarguesian.

Proof of (I): First note that A, A′, B, for example are vectors (a_1, a_2), (a'_1, a'_2), (b_1, b_2). (See Fig. 3.4): Since (A, C, C′, A′) is a parallelogram, C′ = C − A + A′. Since (A, B, B′, A′) is a parallelogram, B′ = B − A + A′. Thus C′ = C − B + B′, so (C, B, B′, C′) is a parallelogram, with $\ell(B, C) \| \ell(B', C')$.

Proof of (II): If P = (0, 0), then for some nonzero number m, A′ = mA, B′ = mB, and C′ = mC. (See Fig. 3.4). Thus B′ − C′ = m(B − C), so $\ell(B, C) \| \ell(B', C')$. If P = (p_1, p_2), then subtracting the vector P from each of the other six vectors translates the configuration to one in which the point of intersection of the three lines is the origin. Thus $\ell(B - P, C - P) \| \ell(B' - P, C' - P)$. Adding back the vector P, we have $\ell(B, C) \| \ell(B', C')$. ■

The two statements of Desargues' Theorem for affine planes reduce to just one statement for projective planes. If in an affine plane $(\mathscr{P}, \mathscr{L})$ we have $\ell(A, A') \| \ell(B, B') \| \ell(C, C')$ or $\ell(A, A') \cap \ell(B, B') \cap \ell(C, C') = P$, after adjoining a line at infinity to form the projective plane $(\mathscr{P}', \mathscr{L}')$ the three lines meet at a point, possibly a point at infinity. Thus Axiom $P4$,

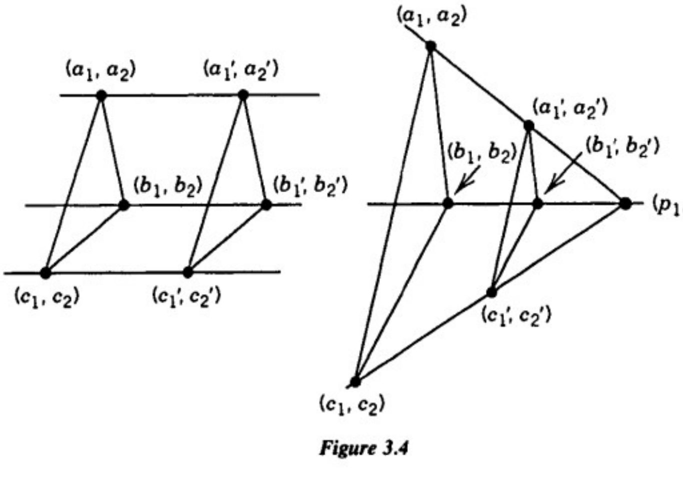

Figure 3.4

which follows, is the statement of Desargues' Theorem for projective planes.

Definition. A projective plane is said to be *Desarguesian* if it satisfies the following axiom (see Fig. 3.5):

> *P*4. Suppose that $\ell(A, A') \cap \ell(B, B') \cap \ell(C, C') = P$. Let $C'' = \ell(A, B) \cap \ell(A', B')$, $B'' = \ell(A, C) \cap \ell(A', C')$, and $A'' = \ell(B, C) \cap \ell(B', C')$. Then A'', B'', and C'' are collinear.

Taking $\ell(A'', B'')$ as the line at infinity, gives Desargues' Theorem in the resulting affine plane, and the two affine statements of Desargues' Theorem correspond to the cases where P either is, or is not, on $\ell(A'', B'')$. Thus, the statement of Desargues' Theorem for the projective plane is an

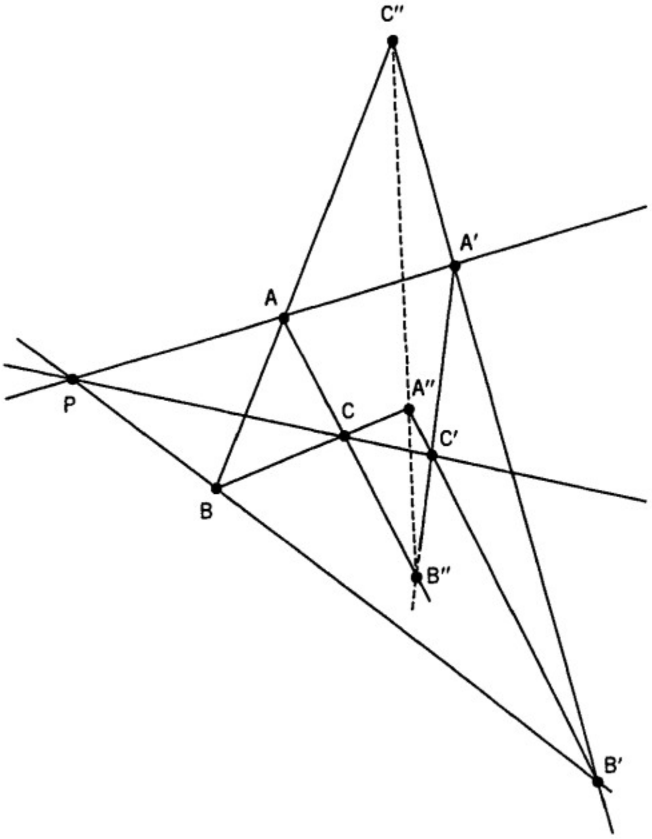

Figure 3.5

example of a single statement in projective geometry being equivalent to more than one affine statement.

Definition. A *triangle* (A, B, C) in a projective plane is a set of three noncollinear points called *vertices*, and the lines containing them called *sides*. The triangles (A, B, C) and (A′, B′, C′) are called *perspective from a point* if the lines joining their vertices are *concurrent*, namely if $\ell(\mathrm{A}, \mathrm{A}') \cap \ell(\mathrm{B}, \mathrm{B}') \cap \ell(\mathrm{C}, \mathrm{C}') = \mathrm{P}$ for some point; they are *perspective from a line* if their corresponding sides meet in collinear points. Thus Desargues' Theorem for the projective plane can be stated as follows:

> *P*4′. If two triangles are perspective from a point, they are perspective from a line.

Note that "triangle" is a self-dual concept in a projective plane, while "perspective from a point" and "perspective from a line" are dual concepts. Therefore

■ **THEOREM 5.** A projective plane is Desarguesian if and only if it satisfies the converse of Desargues' Theorem.

EXERCISES

1. Find the point where the Moulton line ℓ containing (2, 3) and (−1, −4) crosses the *x*-axis. Find the point where the line containing (3, 3) and parallel to ℓ crosses the *x*-axis.
2. Find an example of Desargues' configuration in the affine plane of order 3 (Example 7 of Chapter 2). Do the same for the affine plane of order 4 (Example 8 of Chapter 2).
3. State the converse of Desargues' Theorem as given in Axiom *P*4, and prove that it holds in any Desarguesian projective plane.

*4. Write Desargues' Theorem for affine planes as a single statement. State its converse.

5. Make sketches of the following configurations:
   a. ℓ and m are parallel lines with X, Y, Z, in ℓ and E, F, G, in m.
   b. Modify the sketch from (a) so that

$$\ell(\mathrm{Y}, \mathrm{E}) \| \ell(\mathrm{Z}, \mathrm{F}) \text{ and } \ell(\mathrm{Y}, \mathrm{G}) \| \ell(\mathrm{X}, \mathrm{F}).$$

6. Make sketches of the following configurations:
   a. (A, B, C, D) is a quadrilateral with no pair of its sides parallel.
   b. In the configuration from (a), insert A′ so that (A′, B, C, D) is a parallelogram.

c. In the configuration from (b) insert B′ so that (A, B′, C, D) is a parallelogram.

d. Similarly insert C′ and D′.

*7. Let $(\mathbb{R}, +, \circ)$ be the real numbers with their usual addition, and let k be a fixed positive real number different from 1. Define $r \circ s = rs$ if either r or s is positive or zero, and $r \circ s = krs$ if r and s are both negative. Let $(\mathscr{P}, \mathscr{L})$ be given by $\mathscr{P} = \{(x, y): x, y \in \mathbb{R}\}$ and $\ell \in \mathscr{L}$ if and only if $\ell = \{(x, y): a \circ x + b \circ y = c\}$ where a and b are not both zero.

a. Show that $(\mathscr{P}, \mathscr{L})$ is an affine plane.

b. For $k = 2$, sketch the line $1 \circ x + (-1) \circ y = 0$.

c. For $k = 1/3$, sketch the line $1 \circ x + (-1) \circ y = 0$.

d. For which value of k is $(\mathscr{P}, \mathscr{L})$ the Moulton Plane?

3.4 PROPERTIES OF ADDITION IN AFFINE PLANES

To define addition of coordinates on an affine line, two distinguished (or "special") points were selected: the origin O on ℓ, and a point B (which will be called an *auxiliary point*) not on ℓ. The effect of changing the O will be considered later; now, using Desargues' Theorem, it can be shown that regardless of the choice of B, the same sum (relative to O) will be obtained for any fixed A and C in ℓ.

■ **THEOREM 6.** The definition of A + C is independent of the choice of B.

Proof: Add A and C, using two auxiliary points B_1 and B_2, obtaining $(A + C)_1$ and $(A + C)_2$, as shown in Fig. 3.6. Note that once B_1 and B_2 are chosen, D_1 and D_2 are determined.)

Case 1. B_1, B_2, and O are collinear. The parallel axiom ensures that D_1, D_2, and A are also collinear, on the line containing A and parallel to $\ell(O, B_1)$. Now use Desargues' Theorem on triangles B_1, B_2, C and D_1,

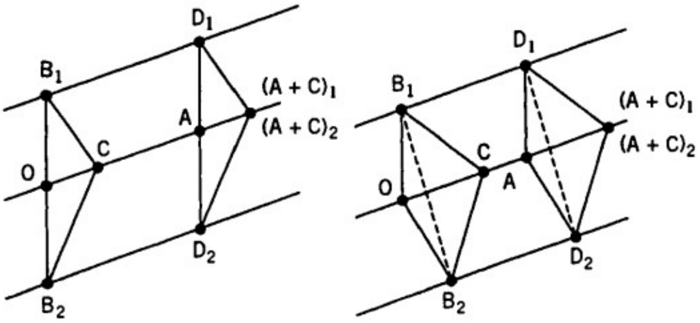

Figure 3.6

D_2, $(A + C)_1$, which gives $\ell(B_2, C) \| \ell(D_2, (A + C)_1)$. But we know that $\ell(B_2, C) \| \ell(D_2, (A + C)_2)$; this implies (by the parallel axiom and the fact that two lines intersect in at most one point) that $(A + C)_1 = (A + C)_2$.

Case 2. $B_2 \notin \ell(O, B_1)$. Desargues' Theorem on triangles B_1, O, B_2 and D_1, A, D_2, gives $\ell(B_1, B_2) \| \ell(D_1, D_2)$. Then using Desargues' Theorem again, this time on triangles B_1, B_2, C and $D_1, D_2, (A + C)_1$, we obtain $\ell(B_2, C) \| \ell(D_2, (A + C)_1)$. As in Case 1, this leads to the conclusion that $(A + C)_1 = (A + C)_2$. ■

Observe the use of the phrase "use Desargues' Theorem on triangles . . ." in the proof above. This is followed by two *ordered* triples of points, representing *corresponding* vertices. In the example above, the listing is B_1, B_2, C and $D_1, D_2, (A + C)_1$, where $\ell(B_1, D_1)$, $\ell(B_2, D_2)$, and $\ell(C, (A + C)_1)$ are parallel lines, while $\ell(B_1, B_2) \| \ell(D_1, D_2)$ and $\ell(B_1, C) \| \ell(D_1, (A + C)_1)$. It becomes much easier to use Desargues' Theorem if the six points involved are listed in the order given here.

In the sequel the phrase, "add A + C using B (and D)" will also be used. This is almost self-explanatory. By putting "(and D)" in parentheses, one emphasizes the fact that choosing an *auxiliary* point B determines which point will play the role of D (the second auxiliary point) in the definition of addition.

Pursuing, as far as possible, the aim of imitating Descartes' introduction of *arithmetic* into geometry, it can now be shown that each point on ℓ has an *additive inverse* on ℓ.

■ **THEOREM 7.** For each point $A \in \ell$, there is a point $(-A) \in \ell$ such that $A + (-A) = (-A) + A = O$.

Proof: Select $B \notin \ell$, and let D be the point such that $\ell(B, D) \| \ell$ and $\ell(O, B) \| \ell(A, D)$ as in the definition of addition. Now let $(-A)$ be the point of intersection of ℓ with the line containing B and parallel to $\ell(O, D)$. (Check that this point exists!) Adding $A + (-A)$ using B (and D), we get $A + (-A) = O$. Adding $(-A) + A$ using D (and B), $(-A) + A = O$ as well. See Fig. 3.7 where this proof is illustrated. ■

The proof that addition of points on a line is *associative* uses the fact that one has complete freedom in choosing "auxiliary" points for addition.

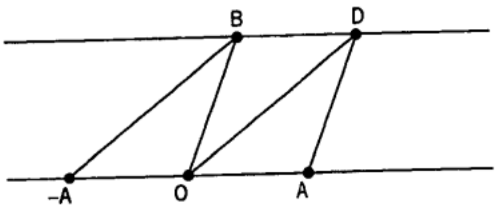

Figure 3.7

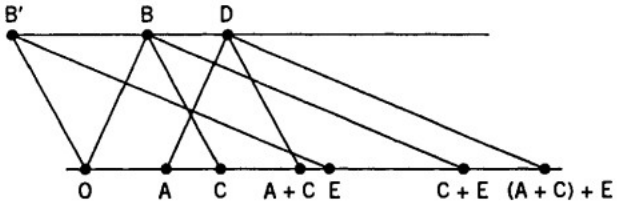

Figure 3.8

■ **THEOREM 8.** Given A, C, and E on ℓ, $(A + C) + E = A + (C + E)$.

Proof: First add $(A + C)$ using B (and D). Then add $(A + C) + E$ using B′ (chosen on $\ell(B, D)$ so that $\ell(O, B')\|\ell(B, C)$). This will force D to be the second auxiliary point, and we obtain $\ell(B', E)\|\ell(D, (A + C) + E)$). (see Fig. 3.8)

Now add $(C + E)$ using B′ (and B), and finally add $A + (C + E)$ using B (and D). This gives $\ell(B, (C + E))\|\ell(D, A + (C + E))$. Since $\ell(B', E)\|\ell(B, C + E)$, the parallel axiom implies that the points $(A + C) + E$ and $A + (C + E)$ are the same. ■

The last main theorem on addition states that addition is *commutative*. For this the *converse* of Desargues' Theorem for affine planes is needed. This, together with a proof that Desargues' Theorem (I) is implied by Desargues' Theorem (II) is the subject of the next section.

EXERCISES

1. Construct an addition table for the line {AEI} in the affine plane of order 3 (Example 7 in Chapter 2) with A = O. Use this table to find the inverse of the point E. Now find this inverse geometrically as in Theorem 7, and observe that they are the same.

   Exercises 2–4 refer to the affine plane of order 4 whose points and lines are repeated below. $\mathcal{P} = \{A, B, \dots, P\}$, and the lines are grouped into pencils:

{OIDF}	{HBCD}	{HFKP}	{OAJH}	{HEIM}
{ECPA}	{EFGJ}	{BEOL}	{GIPB}	{BFAN}
{KMJB}	{IAKL}	{JNCI}	{FCML}	{CGKO}
{GHLN}	{MNOP}	{AGDM}	{EDKN}	{DJLP}

2. a. Find the line containing K and parallel to {BEOL}. Whree does this line intersect {MNOP}?
   b. On {MNOP} add M + N, using B as auxiliary point. What is the second auxiliary point?

3. Complete the addition table for {OIDF}. What is the inverse of D?

+	O	I	D	F
O				
I				
D				
F				

4. Find a parallelogram in the plane. What can you say about is diagonals?
5. Suppose that A, B, and C are noncollinear points in a Desarguesian affine plane $(\mathscr{P}, \mathscr{L})$. Let O be the unique point such that (A, B, C, O) is a parallelogram, and let D be the unique point such that (A, C, B, D) is a parallelogram. Relative to O, show that A + A = D.
6. Suppose that $(\mathscr{P}, \mathscr{L})$ is a Desarguesian affine plane in which the diagonals of every parallelogram are parallel. Suppose that A is any point in $\mathscr{P}$. Prove that relative to any O, A + A = O.
7. Find five orthogonal partitions of the numbers {1, . . . , 16}.
8. Find three orthogonal Latin squares of order 4.

3.5 THE CONVERSE OF DESARGUES' THEOREM

As was stated above, Desargues' Theorem is equivalent to its own converse in the affine as well as the projective plane. By this is meant that the converse of Desargues' Theorem can be proved using Axioms *A*1, *A*2, *A*3, and *A*4, but also by assuming *A*1, *A*2, *A*3, and the converse of Desargues' Theorem, Desargues' Theorem can be proved. In fact more can be done. We will first use *A*1, *A*2, *A*3, and Desargues' Theorem (II) to prove the converse. Then we will use *A*1, *A*2, *A*3, and the converse to prove Desargues' Theorem. This will imply that, instead of assuming Desargues' Theorem, it would have been enough to assume only Desargues' Theorem (II).

■ **THEOREM 9. The converse of Desargues' Theorem for affine planes.** Let A, B, C and A′, B′, C′ be two triples of noncollinear points in a Desarguesian affine plane. Suppose that

$$\ell(A, B) \| \ell(A', B'),\ \ell(A, C) \| \ell(A', C'),\ \text{and}\ \ell(B, C) \| \ell(B', C').$$

Then either $\ell(A, A') \| \ell(B, B') \| \ell(C, C')$ or $\ell(A, A') \cap \ell(B, B') \cap \ell(C, C') = P$ for some P.

Proof (using A1, A2, A3, and Desargues' Theorem (II); see Fig. 3.9): Assume that $\ell(A, B) \| \ell(A', B')$; $\ell(A, C) \| \ell(A', C')$; $\ell(B, C) \| \ell(B', C')$. If it

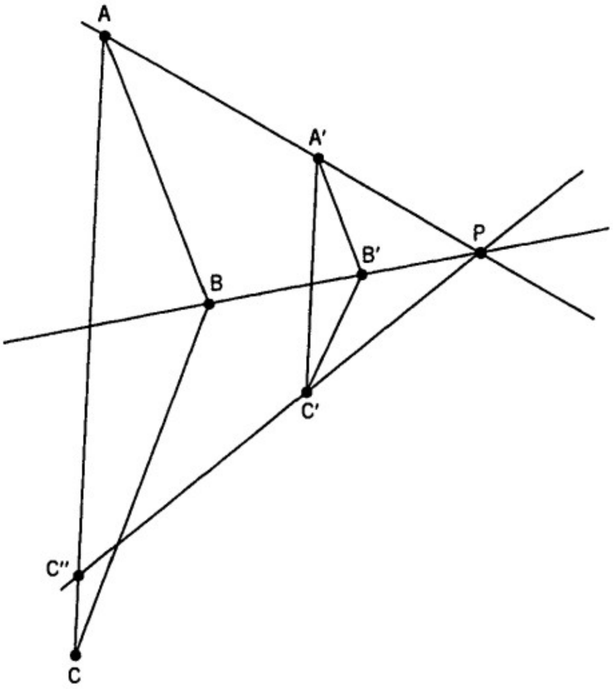

Figure 3.9

is not the case that the lines $\ell(A, A')$, $\ell(B, B')$, and $\ell(C, C')$ are all parallel, two of them must meet. Suppose that $\ell(A, A') \cap \ell(B, B') = P$. Let $\ell(C', P) \cap \ell(A, C) = C''$. Use Desargues' Theorem (II) on A, B, C'' and A', B', C' to obtain $\ell(B, C'') \| \ell(B', C')$. But $\ell(B, C)$ is the unique line through B and parallel to $\ell(B', C')$. Thus $\ell(B, C) = \ell(B, C'')$. This line intersects $\ell(A, C)$ in at most one point; hence $C = C''$, and $P \in \ell(C, C')$.

■

■ **THEOREM 10.** Desargues' Theorem is true in any affine plane in which its converse holds.

Proof: (I) Suppose that A, B, C and A′, B′, C′ are triples of noncollinear points. Suppose that $\ell(A, A') \| \ell(B, B') \| \ell(C, C')$, and let $\ell(A, B) \| \ell(A', B')$ with $\ell(A, C) \| \ell(A', C')$. It must be shown that $\ell(B, C) \| \ell(B', C')$.

Suppose not. Then let C'' be the point on $\ell(A, C)$ such that $\ell(B, C'') \| \ell(B', C')$. If $C \neq C''$, then $\ell(C'', C') \cap \ell(B, B') = P$ for some point P (they cannot be parallel since $\ell(C, C')$ is the unique line containing C′

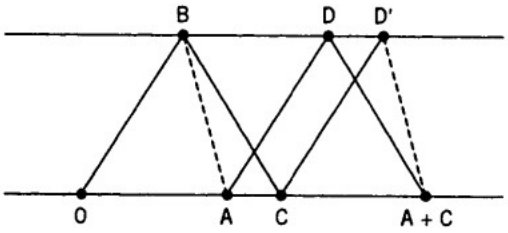

Figure 3.10

parallel to $\ell(B, B')$). Use the converse of Desargues' Theorem on A, B, C'' and A', B', C' to get $\ell(A, A')$, $\ell(B, B')$ and $\ell(C'', C')$ all parallel or all meeting at one point. But the first two are assumed parallel, and the latter two intersect at P, a contradiction.

(II) Let A, B, C and A', B', C' be triples of noncollinear points. Suppose $\ell(A, A') \cap \ell(B, B') \cap \ell(C, C') = P$, and let $\ell(A, B) \| \ell(A', B')$ with $\ell(A, C) \| \ell(A', C')$. It must be shown that $\ell(B, C) \| \ell(B', C')$.

Suppose not. Then let C'' be the point on $\ell(A, C)$ such that $\ell(B, C'') \| \ell(B', C')$. If $C \neq C''$, then P is not contained in $\ell(C'', C')$. Use the converse of Desargues' Theorem on A, B, C'' and A', B', C' to get $\ell(A, A')$, $\ell(B, B')$, and $\ell(C'', C')$ all parallel or all meeting at one point. But the first two are assumed to meet at P, which is not in the third line, a contradiction. ■

Corollary. In any affine space in which Desargues' Theorem (II) holds, Desargues' Theorem holds as well.

■ **THEOREM 11.** Addition of points on a line in any Desarguesian affine plane is commutative.

Proof: Compute $A + C$ using any point B (and D). Now find $C + A$ using the same point B (and D'). Since $\ell(B, A) \| \ell(D', C + A)$, $C + A$ is the same point as $A + C$ if $\ell(B, A) \| \ell(D', A + C)$. (See Fig. 3.10)

First use the converse of Desargues' Theorem on $A, D, A + C$, and D', C, B to show that the lines $\ell(A, D')$, $\ell(D, C)$, and $\ell(A + C, B)$ are all parallel or meet at a single point. Now use Desargues' Theorem on A, D, B, and $D', C, A + C$ to obtain $\ell(A, B) \| \ell(D', A + C)$, which completes the proof. ■

EXERCISES

1. Suppose that ℓ and m are parallel lines in a Desarguesian affine plane $(\mathscr{P}, \mathscr{L})$. Suppose that P, Q, and R are distinct points on ℓ, with S, T, and U distinct points on m as shown in Fig. 3.11. Assume that $\ell(P, S) \| \ell(R, U)$ and $\ell(P, T) \| \ell(Q, U)$. Prove that $\ell(Q, S) \| \ell(R, T)$.

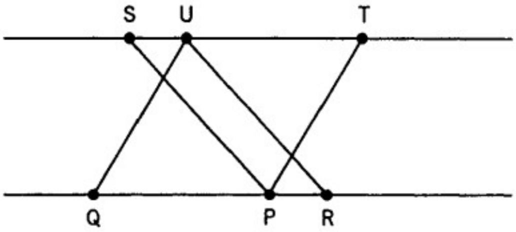

Figure 3.11

2. The statement in Exercise 1 is a theorem in Desarguesian affine planes. State the corresponding theorem for Desarguesian projective planes.
3. Let (A, B, C, D) be a quadrilateral in a Desarguesian affine plane, none of whose sides are parallel. Define A′ to be the unique point such that (A′, B, C, D) is a parallelogram, and B′ the unique point such that (A, B′, C, D) is a parallelogram as shown in Fig. 3.12. Prove that $\ell(A, A') \| \ell(B, B')$.

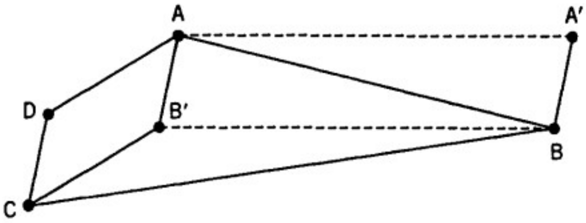

Figure 3.12

4. Let $(\mathscr{P}, \mathscr{L})$ be a Desarguesian affine plane containing the distinct points A, B, C, D, E, F, G. Assume that $\ell(A, D) \| \ell(B, G) \| \ell(E, F)$; $\ell(A, F) \| \ell(D, E)$;

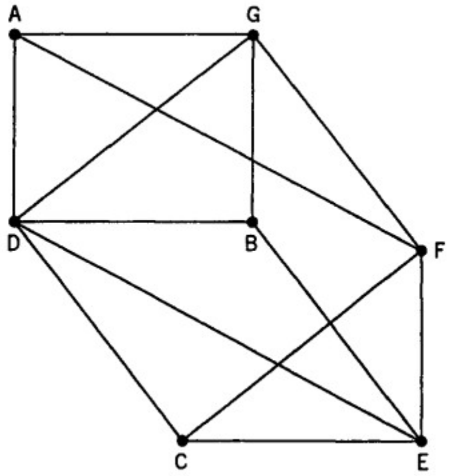

Figure 3.13

$\ell(A, G) \| \ell(D, B) \| \ell(C, E)$; $\ell(D, C) \| \ell(B, E)$ where all lines listed above are distinct. (See Fig. 3.13.) Prove:

a. $\ell(D, C) \| \ell(G, F)$.

b. $\ell(D, G) \| \ell(C, F)$.

3.6 MULTIPLICATION ON LINES OF AN AFFINE PLANE

In the preceding sections it was shown that addition is associative and commutative, O acts as an identity for addition, and every point has an additive inverse. These properties are discussed in Appendix A, together with an appropriate concept of *multiplication*, which can now be introduced for points on a line in an affine plane. Starting with a line ℓ, through a fixed origin O, as above, a fixed point I, *different* from O is chosen (recall that Axiom $A3$ guarantees at least two points on each line). Having these points, multiplication can be defined with respect to the selected O and I.

Definition of multiplication. Again $(\mathscr{P}, \mathscr{L})$ is a Desarguesian affine plane, with ℓ a line and O and I distinguished points of ℓ. (See Fig. 3.14) For points A and C in ℓ, to multiply A and C (relative to O and I):

1. Select B not in ℓ.
2. Let m be the line containing A, and parallel or equal to $\ell(I, B)$.
3. Let $D = m \cap \ell(O, B)$.
4. Let k be the line containing D and parallel or equal to $\ell(B, C)$.
5. AC is defined to be $k \cap \ell$.

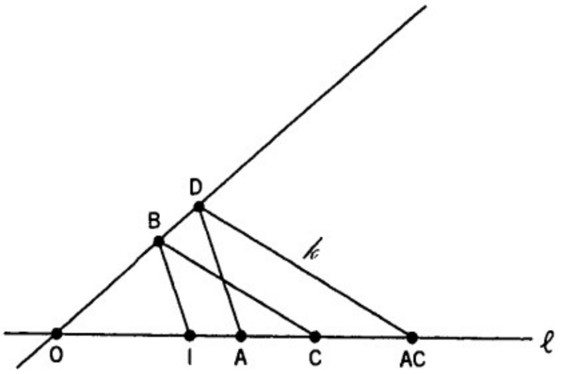

Figure 3.14

It will be shown in Theorem 15 that the multiplication of points on ℓ does not depend on the choice of the fixed point B.

Following the pattern established for addition, it is first shown that the suggestively named "I" acts as a multiplicative identity. Existence of nonzero inverses and associativity will follow as well; however multiplication is *not* in general commutative.

■ **THEOREM 12.** For any A in ℓ, AI = IA = A.

Proof: Select B (and D) as described above. Since $\ell(B, I) \| \ell(D, A)$, it follows that AI = A. In computing IA, note that line m described in step 2 is equal to $\ell(I, B)$, whence D = B and $k = \ell(B, A)$. Thus IA ($= \ell \cap k$) = A. ■

■ **THEOREM 13.** AO = OA = O for any A in ℓ.

Proof: In computing AO, the line k in step 4 is equal to $\ell(B, O)$, whence AO = O. In computing OA, the auxiliary point D is O, thus OA = O. ■

■ **THEOREM 14.** If A ≠ O, there is a point (A^{-1}) in ℓ such that $A(A^{-1}) = (A^{-1})A = I$.

Proof: Select B and D so that $\ell(I, B) \| \ell(A, D)$ with D on $\ell(O, B)$. (See Fig. 3.15.) Let n be the line containing B and parallel to $\ell(D, I)$, and let $n \cap \ell = (A^{-1})$. Multiplying $A(A^{-1})$ using B and D gives I, and multiplying $(A^{-1})A$ using D and B also yields I. ■

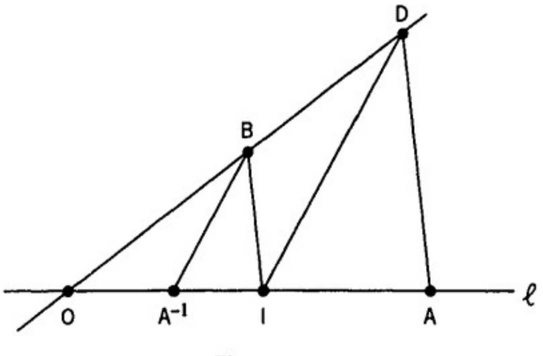

Figure 3.15

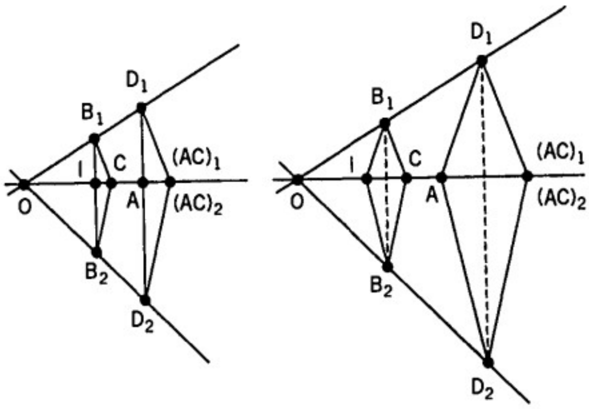

Figure 3.16

In Theorem 6 Desargues' Theorem (I) was used to prove that the sum of two points on a line (relative to a fixed origin O) does not depend on the choice of the auxiliary points. An analogous result (this time depending on Desargues' Theorem (II)) holds for multiplication. Observe also that as was the case with addition, the first auxiliary point determines the second one uniquely.

■ **THEOREM 15.** The definition of AC is independent of the choice of B above.

Proof: Obtain the product $(AC)_1$ by using B_1 (and D_1); then obtain $(AC)_2$ using B_2 (and D_2) as in Fig. 3.16.

CASE 1. I, B_1, and B_2 are collinear. Use Desargues' Theorem on B_1, B_2, C and D_1, D_2, $(AC)_1$ to obtain $\ell(B_2, C) \| \ell(D_2, (AC)_1)$, which forces $(AC)_1 = (AC)_2$.

CASE 2. The proof of Case 2 is left as an exercise. ■

As was the case with addition, Theorem 15 can be used to show that multiplication of points on a line is associative.

■ **THEOREM 16.** For A, C, and E on ℓ, $(AC)E = A(CE)$.

Proof: (See Fig. 3.17, and compare this with the proof that addition is associative.) First multiply AC using B (and D). Now multiply (AC)E using B' (and D), where $\ell(I, B') \| \ell(AC, D)$. Multiply CE using B' (and B), and finally multiply A(CE) using B (and D) to obtain $A(CE) = (AC)E$. ■

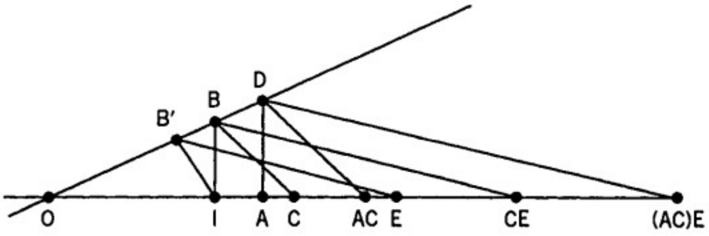

Figure 3.17

EXERCISES

1. In the affine plane of order 3 given in Example 7 of Chapter 2, fix ℓ as {CFI}, and let C = O. Make a multiplication table for the line (which is now {OIF}).
2. In the affine plane of order 4 given with the exercises of Section 3.4, p. 57, make a multiplication table for the line {OIDF}.
   a. What is the multiplicative inverse of D? of F?
   b. Compute (DD)F and D(DF) from the table, and check that they are the same.
   c. Compute D(I + F) and DI + DF from the table and check that they are the same.
3. Prove Case 2 of Theorem 15.
4. Let O, I, and A be distinct collinear points in a Desarguesian affine plane $(\mathscr{P}, \mathscr{L})$. Define the point 2I to be I + I and 2A to be the point A + A, the addition done relative to O. Prove geometrically that (2I)A = A + A assuming that the points O, I, 2I, A, and 2A are all distinct. [*Hint:* Compute I + I using B (and D) as auxiliary points as shown in Fig. 3.18. (2I)A is computed using B (and D′). Compute A + A using B and E.]

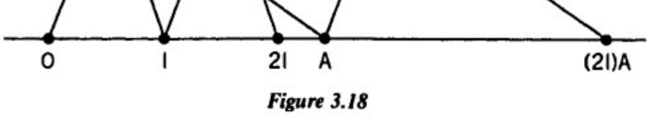

Figure 3.18

3.7 PAPPUS' THEOREM AND FURTHER PROPERTIES

Multiplication in a Desarguesian affine plane is not necessarily commutative. To present an example of a plane with noncommutative multiplication, some algebraic results are necessary, so the example is deferred until Chapter 4. However, the geometric counterpart of commutativity of multiplication is *Pappus' Theorem*, which was stated as a *Euclidean* theorem in Chapter 1, Theorem 12.

■ **PAPPUS' THEOREM.** Let $\ell \cap m = O$, with P, Q, and R in ℓ; S, T, U in m; $\ell(P, T) \| \ell(Q, U)$, and $\ell(Q, S) \| \ell(R, T)$. Then $\ell(P, S) \| \ell(R, U)$.

■ **THEOREM 17.** Pappus' Theorem is true in a Desarguesian affine plane if and only if multiplication on the points of its lines is commutative.

Proof: Choosing I in ℓ so that $\ell(I, S) \| \ell(P, T)$, then PQ = R (multiplied using S (and T)). If Pappus' Theorem is true, then QP = R as well (multiplied using S (and U)). (See Fig. 3.19) Conversely, if Pappus' Theorem fails, then, relative to O and I, PQ = R but QP ≠ R, so multiplication is not commutative. ■

Pappus and Desargues

In all the examples of Desarguesian affine planes presented so far, Pappus' Theorem holds. In fact algebraic methods can be used to show that Pappus' Theorem is true in any *finite* Desarguesian affine plane, but there are (infinite) Desarguesian affine planes in which Pappus' Theorem fails. On the other hand, it has been shown that Pappus' Theorem implies that of Desargues (Hessenberg 1905).

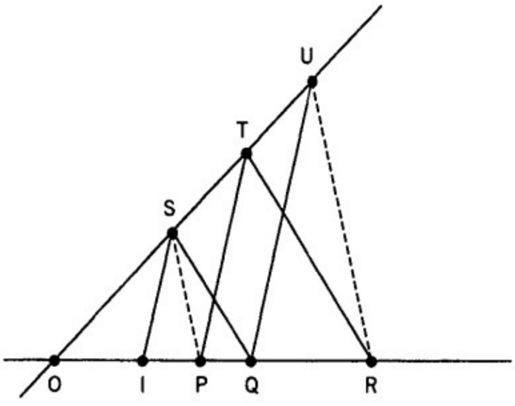

Figure 3.19

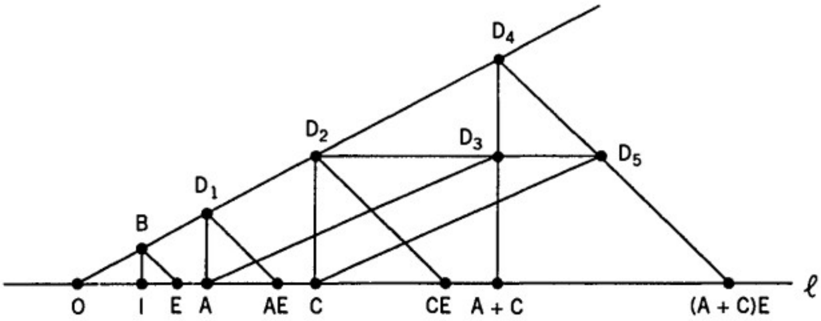

Figure 3.20

The Distributive Laws

Besides their individual properties, addition and multiplication are related by the *distributive laws*. Because multiplication is not necessarily commutative, two distributive laws are needed. Furthermore, after laws $(A + C)E = AE + CE$ and $E(I + C) = E + EC$ are proved *geometrically*, an algebraic proof shows that, for an arbitrary A, $E(A + C) = EA + EC$.

■ **THEOREM 18.** For A, C, and E in ℓ, $(A + C)E = AE + CE$.

Proof: Select $B \notin \ell$. (See Fig. 3.20.)
Compute:

AE using B (and D_1).
CE using B (and D_2).
$A + C$ using D_2 (and D_3).
$(A + C)E$ using B (and D_4).

Let D_5 be in $\ell(D_4, (A + C)E)$, so that $\ell(A, D_3) \| \ell(AE, D_5)$. Use Desargues' Theorem on D_4, D_3, D_5 and D_1, A, AE to get $\ell(D_3, D_5) \| \ell(A, AE)$, which shows that $D_5 \in \ell(D_2, D_3)$. Now add $AE + CE$ using D_2 (and D_5), and note that $\ell(D_2, CE) \| \ell(D_5, AE + CE)$. But this means that $AE + CE = (A + C)E$. ■

■ **THEOREM 19.** For C and E in ℓ, $E(I + C) = E + EC$.

Proof: Select $B \notin \ell$. (See Fig. 3.21.)
Compute:

EC using B (and D).
$I + C$ using B (and D_1).
$E + EC$ using D_1 (and D_2).

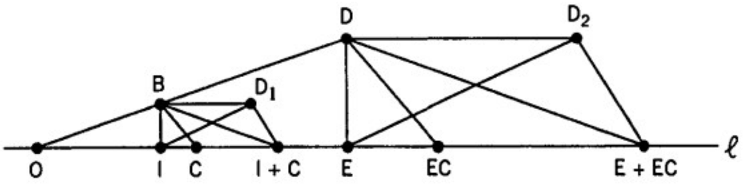

Figure 3.21

Use the converse of Desargues' Theorem on D_1, B, I and D_2, D, E which implies that $O \in \ell(D_1, D_2)$. Now use Desargues' Theorem on D_1, D_2, $E + EC$, and $B, D_1, I + C$, which gives $\ell(D, E + EC) \| \ell(B, I + C)$, whence $E + EC = E(I + C)$. ■

Corollary. For A, C, and E in ℓ, $E(A + C) = EA + EC$.

Proof: If $A = O$, the result is immediate. If $A \neq O$, then

$$\begin{aligned}(E(A + C) - EA - EC)A^{-1} &= E((A + C)A^{-1} - E - ECA^{-1} \\ &= E(I + CA^{-1}) - E - ECA^{-1} \\ &= E + ECA^{-1} - E - ECA^{-1} \\ &= O.\end{aligned}$$

Thus $E(A + C) - EA - EC = O$, and

$$E(A + C) = EA + EC. \qquad ■$$

The fundamental theorem of synthetic affine geometry, stated as Theorem 1 of this chapter and repeated below, has now been proved.

■ **THEOREM 20. The fundamental theorem of synthetic affine geometry.** Relative to fixed coordinates O and I, any line in a Desarguesian affine plane is a *division ring* in that its addition and multiplication are associative, and have identites, every element has an additive inverse, and every nonzero element has a multiplicative inverse. Addition is commutative, and multiplication distributes over addition.

In the next chapter we will prove that for any division ring $\mathbb{D}$, the elements of $\mathbb{D} \times \mathbb{D}$ are the points of a Desarguesian affine plane whose lines satisfy linear equations of the form $xa + yb = c$. Conversely, any Desarguesian affine plane can be so represented over (or *coordinatized by*) a division ring.

This chapter concludes with another example of an affine plane of order 5. It will be shown in Chapter 4 that this plane can be coordinatized by the field $\mathbb{Z}_5$.

□**EXAMPLE 1: An affine plane of order 5.** $\mathscr{P} = \{A, B, C, \ldots, Y\}$, and the lines are

{OIABC}
{DEFGH}
{JKLMN}
{PQRST}
{UVWXY}
{OHMRV}
{IDNSW}
{AEJTX}
{BFKPY}
{CGLQU}

{ODJPU}
{IEKQV}
{AFLRW}
{BGMSX}
{CHNTY}
{OGKTW}
{IHLPX}
{ADMQY}
{BENRU}
{CFJSV}

{OELSY}
{CDKRX}
{BHJQW}
{AGNPV}
{IFMTU}
{OFNQX}
{IGJRY}
{AHKSU}
{BDLTV}
{CEMPW}

□

EXERCISES

The following exercises pertain to the plane of order 5 given above.

1. Make addition and multiplication tables for {OIABC}.
2. Make addition and multiplication tables for $\mathbb{Z}_5$. Show that {OIABC} is *isomorphic* to $\mathbb{Z}_5$ by finding a bijection from {OIABC} to $\mathbb{Z}_5$ that preserves addition and multiplication. [*Hint:* Map O to 0 and I to 1.]
3. Find a parallelogram in $(\mathscr{P}, \mathscr{L})$ and find the point of intersection of its diagonals.
4. Find an instance of the configurations for Desargues' Theorem in $(\mathscr{P}, \mathscr{L})$.
5. Find an instance of the configuration for Pappus' Theorem in $(\mathscr{P}, \mathscr{L})$.
6. Use $(\mathscr{P}, \mathscr{L})$ to define a projective plane of order 5.

*7. Let $(\mathscr{P}', \mathscr{L}')$ be a projective plane in which Pappus' Theorem holds, and prove Desargues' Theorem as follows: Let $\ell(A, A') \cap \ell(B, B') \cap \ell(C, C') = O$, and let $A'' = \ell(B, C) \cap \ell(B', C')$, $B'' = \ell(A, C) \cap \ell(A', C')$, and $C'' = \ell(A, B) \cap \ell(A', B')$. Define $E = \ell(A, C) \cap \ell(B', C')$, $F = \ell(A, B') \cap \ell(O, C)$, $G = \ell(A, B) \cap \ell(O, E)$, and $H = \ell(A', B') \cap \ell(O, E)$. Prove that A'', B'', and C'' are collinear by using Pappus' Theorem on O, B, B' and A, E, C; then on O, A, A' and B', C', E, and finally, on A, B', F and H, G, E. [*Hint:* In the first use of Pappus' Theorem, since $\ell(B, C) \cap \ell(B', E) = A''$, $\ell(A, B') \cap \ell(O, C) = F$ and $\ell(O, E) \cap \ell(A, B) = G$, we obtain A'', F, and G collinear.]

8. Find six orthogonal partitions of the set $\{1, 2, \ldots, 25\}$.

9. Find four orthogonal Latin squares of order 5.

SUGGESTED READING

G. Hessenberg, Beweis des Desarguesschen Satzes aus dem Pascalschen, *Math. Ann.* **61** (1905), 161–172.

David Hilbert, *Foundations of Geometry*, La Salle, IL: Open Court, 1971.

F. R. Moulton, A simple non-Desarguesian plane geometry, *Trans. AMS* **3** (1902), 192–195.

- Pappus-Theorem/
  Affine and Projective Geometry...pdf
  Pág. 47(64) - Chapter 3
  Desarguesian Affine Plane
  Obs: Pappus's Theorem — Theorem 12 of Chapter 1.

Obs: Livro todo!!!

- Pappus-Theorem/
  Perspectives on Projective Geometry...pdf
  (Jürgen Richter-Gebert)

Obs: Important!!!
Livro todo!

- Pappus-Theorem/
  Projective Geometry....pdf
  (Casse)

Obs: Livro todo!!!

Pág. 22(35) - Chapter 3
Introduction to axiomatic geometry.

⟹

- Pappus-Theorem/

  Affine and Projective Geometry...pdf

  Pág. 47(64) - Chapter 3

  Desarguesian Affine Plane

  Obs: Pappus's Theorem — Theorem 12 of Chapter 1.

Obs: Livro todo!!!

- Pappus-Theorem/

  Perspectives on Projective Geometry...pdf

  (Jürgen Richter-Gebert)

Obs: Important!!!

Livro todo!

- Pappus-Theorem/

  Projective Geometry....pdf

  (Casse)

Obs: Livro todo!!!

  Pág. 22(35) - Chapter 3

  Introduction to axiomatic geometry.

⟶

Return:

Pág. 20(33) - Theorem 2.11

The Principle of Duality

Example 2.12

Desargues's Theorem + Dual

Pág. 21 (34) ⇒ image

By the Principle of Duality, if Desargues's Theorem is valid in the EEP (Extended Euclidean Plane), so is its dual.

Pág. 99 (112) Chapter 5

Coordinatising a projective plane.

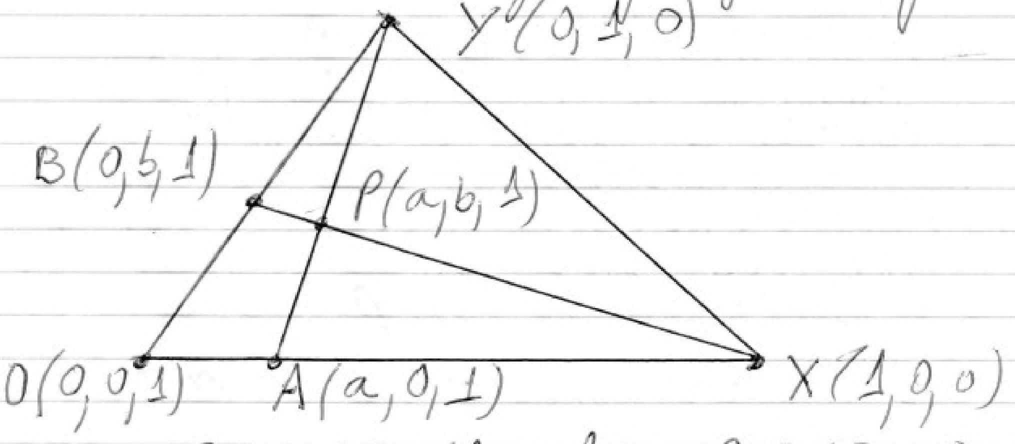

Consider the plane PG(2,F).

Jandaia

)*+,)))- !) % ! ! 2 5 A # ! G
6 ! ! ! G & 8 8 G ! # F 8 3#
-4 < 8 # HHH

3)*+,)))- !) %
?IJK D L. DLM 2 ! 7 8 & C 4 6 3#

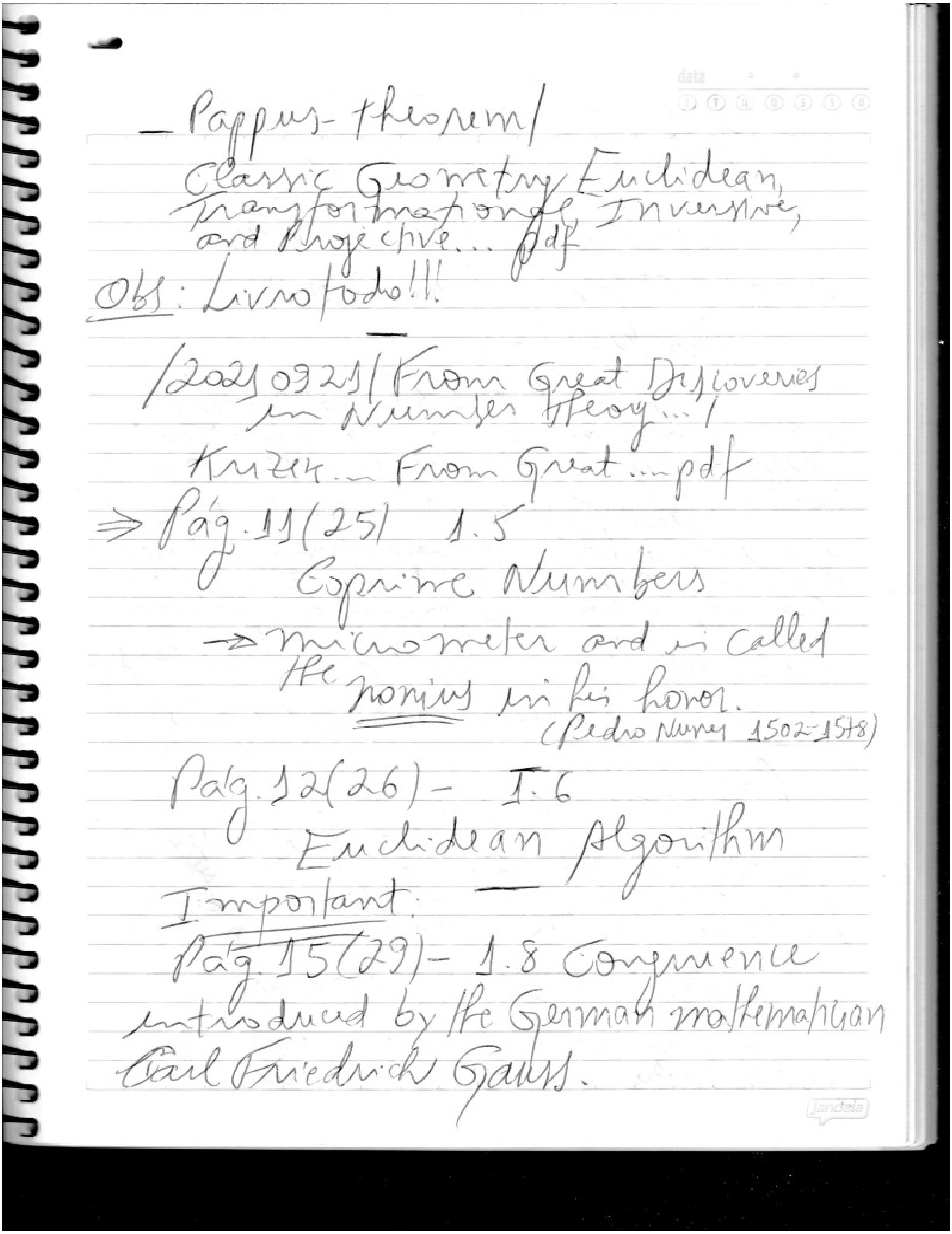

data

— Pappus-theorem/

Classic Geometry Euclidean, Transformational, Inversive, and Projective... pdf

Obs: Livro todo!!!

—

/2021 09 21/ From Great Discoveries in Number Theory.../

Krizek... From Great ...pdf

⇒ Pág. 11(25) 1.5

Coprime Numbers

→ micrometer and is called the nonius in his honor.
(Pedro Nunes 1502-1578)

Pág. 12(26) - 1.6

Euclidean Algorithm

Important: —

Pág. 15(29) - 1.8 Congruence
introduced by the German mathematician Carl Friedrich Gauss.

data
S T Q Q S S D

Gauss → Pag. 53(67) →
The celebrated **Law of Quadratic Reciprocity** was first proved by Gauss.

Pag. 90(103) –

Until 1796, Fermat numbers were more or less a mathematical curiosity. Interest in them grew when Carl Friedrich Gauss discovered an incredible connection between Fermat primes and the Euclidean construction of regular polygons (i.e., using a compass and straightedge)

Look: Theorem 4.12 (Gauss)

Pag. 119(132)

The distribution of the Eisenstein primes in the Gaussian Complex Plane (see Fig 4.11) has an interesting hexagonal symmetry.

Pag. 186(197) Theorem 8.5

Every positive integer is the sum of four squares. 1796 - Gauss.

Jandaia

Pag. 188(199) Fig. 8.3

Partition of a unit square into two triangles and of a unit cube into 6 tetrahedra.

Pag. 251(260) 10.8

What Mathematics is Hidden behind the main clock?

· Clock ⇒ as a consequence of Gauss Theorem (4.12)

↓ Pag. 90(103)

Theorem 4.12 (Gauss)

There exists a Euclidean construction of a regular n-gon if and only if the number of its sides is equal to

$$n = 2^i F_{m_1} F_{m_2} \dots F_{m_j},$$

where $i \geq 0$, $j \geq 0$, $n \geq 3$ are integers and $F_{m_1}, F_{m_2}, \dots, F_{m_j}$ are distinct Fermat primes.

For a proof based on Galois theory see [199, Chapter 16].

↳ 17 Lectures on Fermat Numbers: From Number

theory to Geometry.
Cont.

Gauss theorem 4.12, we get:

theorem 10.16

Let p be an odd prime.
Then there exists a Euclidean construction by ruler and compass of the regular p-gon if and only if p is of the form $p = 2^n + 1$.

⇘ Obs: Usar isto para muito conjuntamente
Coprime Numbers
(Pag. 11(25)).

20210921/
From Great Discoveries.../
82058304.pdf
→ Gauss and the Definition of the Plane Concept in Euclidean Elementary Geometry.
July 28, 1797.

Book: Gauss's contribution to
Euclidean geometry Krizek
Pág 101-107 (1998)

Source: Nº 183
Pág. 317(324) do livro
Krizek...From Great Discoveries...pdf

— 20768551_lese-1.pdf
Gauss and Non-Euclidean
Geometry

20210921 (20210921)
cont.

/20210921/Desargues/
= Euclidean Geometry and its
Subgeometries...pdf
Pág. 265(280) Chapter 12
Isometries of a Euclidean
plane (ISM).

Obs.: Livro todo!!!
Return: Pág. 37(56) - Chapter 2

)*+,)))- !) % " // L L 3#
-4 2! ! # C A # ! 2 5

.--? @).--?). D ( ) ( )7 !)

)*+,)))- !) % A # ! 2 5 ! #
4 7M7 8 3#
' "/ !3 & ! A # ! ! &

'

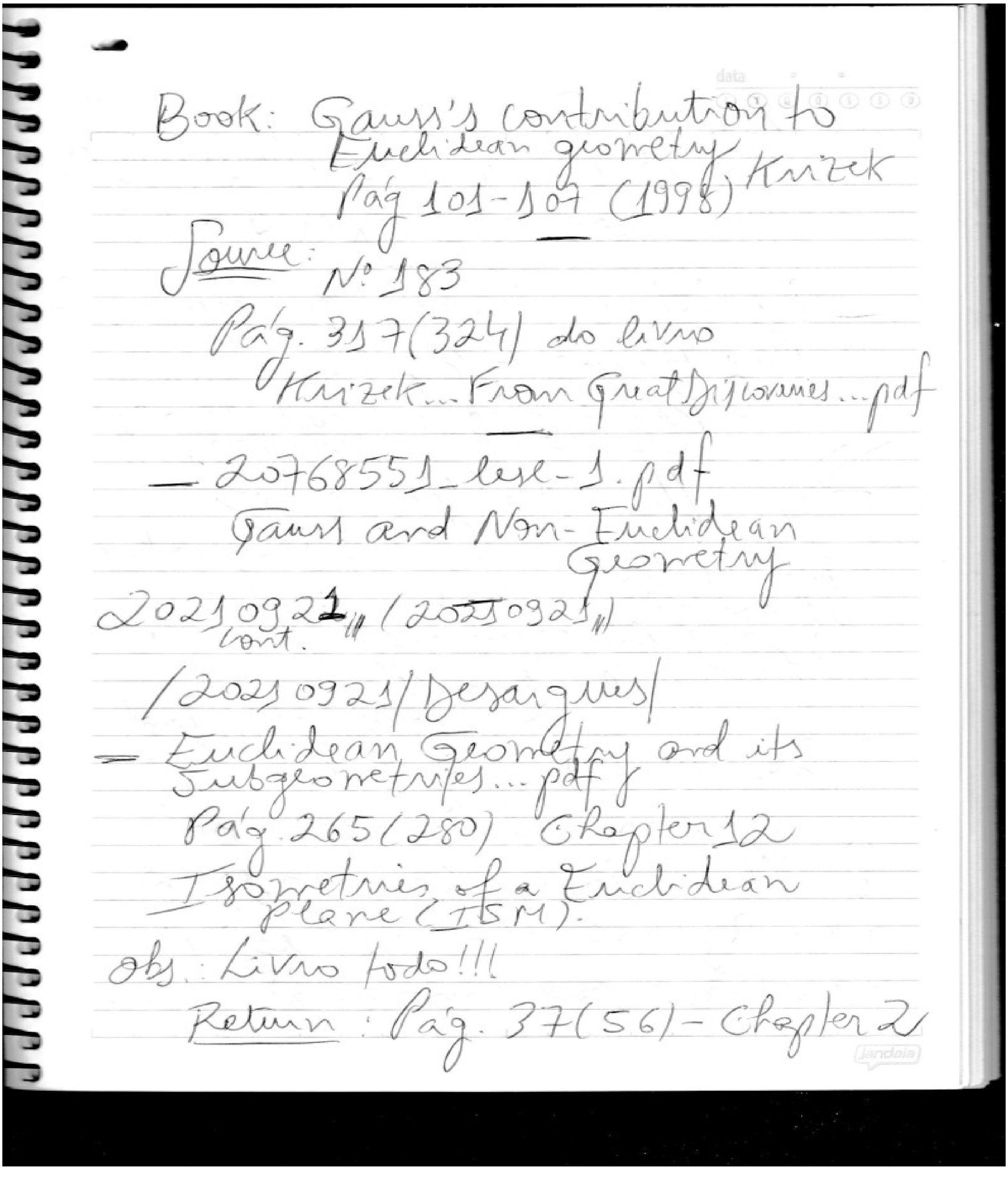

Book: Gauss's contribution to
Euclidean geometry Krizek
Pág 101-107 (1998)

Source: Nº 183
Pág. 317(324) do livro
Krizek... From Great Discoveries... pdf

— 20768551_lese-1.pdf
Gauss and Non-Euclidean
Geometry

20210921// (20210921//)
cont.

/20210921/Desargues/

— Euclidean Geometry and its
Subgeometries... pdf
Pág. 265(280) Chapter 12
Isometries of a Euclidean
plane (ISM).

obs.: Livro todo!!!
Return: Pág. 37(56) - Chapter 2

Affine Geometry:
Incidence with Parallelism (IP).
Obs: Look at Chapter 2 and 3

– Euclidean distance geometry...pdf
Pag. 48(60) – Fig. 5.5
Reflecting (analisar o que é melhor, reflecting or projection orthogonal!)

↘ ⇓

Reflecting through p^y first and p^z second is equivalent to reflecting through p^z first and (the reflection of p^y through p^z) second.

Pag. 49(61) Fig. 5.6
Action of discretization group $G_D(G) = \langle g_3, g_4, g_5 \rangle$ on a realization $x^1 = (x^1_1, \dots x^1_5)$ is transitive and commutative: ...

)*+,)))- !) % A # ! # ! 5 !
3#
' / " M / / +

5.2.2 The discretization group

We generalize the above discussion to general K. Consider a ^{K}DMDGP graph $G = (V, E)$ with as few edges as possible, and edge weight function d. The fact that E has minimum cardinality, while G still has a contiguous $(K - 1)$-lateration order, implies that G is just a chain of $(K + 1)$-cliques, with each consecutive pair sharing a K-clique, see Fig. 5.4. In fact, the edges in E are all discretization edges, according to the edge classification introduced at the end of Sect. 4.4.4.

Let X be the set of all realizations of G in $\mathbb{R}^K$ up to rotations and translations. For any $x \in X$ and $v \in V$ with $v > K$, we define the *partial reflection* operator $g_v : \mathbb{R}^{Kn} \to \mathbb{R}^{Kn}$ given by:

$$g_v(x_1, \ldots, x_n) = (x_1, \ldots, x_{v-1}, R_x^v(x_v), \ldots, R_x^v(x_n)), \tag{5.1}$$

where $R_x^v : \mathbb{R}^K \to \mathbb{R}^K$ is a reflection operator that maps each point in $\mathbb{R}^K$ to its reflection with respect to the hyperplane defined by $\{x_{v-1}, \ldots, x_{v-K}\}$.

We remark that for the R_x^v to be reflections in $\mathbb{R}^K$, we need the edge weight function d to be such that the assumption on the rank of A being full holds (Sect. 3.3.6). This assumption will stand for the rest of this chapter.

Obviously, if you apply any reflection operator twice, you reflect points back to their original positions. Since g_v is simply a composition of $v - 1$ identities and $n - v + 1$ reflections, it enjoys the same property: thus, every g_v is idempotent.

Also, for $x \in X$ and $K < u < v$, we have:

$$\begin{aligned}
g_u g_v(x) &= g_u(x_1, \ldots, x_{v-1}, R_x^v(x_v), \ldots, R_x^v(x_n)) \\
&= (x_1 \ldots, x_{u-1}, R^u_{g_v(x)}(x_u), \ldots, R^u_{g_v(x)} R_x^v(x_v), \ldots, R^u_{g_v(x)} R_x^v(x_n)) \\
&= (x_1 \ldots, x_{u-1}, R_x^u(x_u), \ldots, R^v_{g_u(x)} R_x^u(x_v), \ldots, R^v_{g_u(x)} R_x^u(x_n)) \\
&= g_v(x_1, \ldots, x_{u-1}, R_x^u(x_u), \ldots, R_x^u(x_n)) \\
&= g_v g_u(x).
\end{aligned}$$

To infer the third line from the second, we used the fact that $R^u_{g_v(x)} R_x^v(x_w) = R^v_{g_u(x)} R_x^u(x_w)$ for each $w \geq v$. A proof of this is given in [85, Lemma 1]; we only provide here a graphical sketch of the proof in 2D (Fig. 5.5), where, to simplify notation, we let $R_x^v = \rho^y$, where y is the normal to the line through x_{v-1}, x_{v-2}.

Let $\mathscr{G}_D(G) = \langle g_v \mid v > K \rangle$, i.e., the symmetry group generated by the action of all g_v's on the realizations of G, be the *discretization group* of G. Since its elements commute, as shown above, $\mathscr{G}_D$ is an abelian group. Moreover, all of the elements are idempotent. Thus, $\mathscr{G}_D$ is isomorphic to C_2^{n-K}, and the Cartesian product of $n - K$ copies of the cyclic group C_2.

Notice that if $v \in V$ such that $v > K$ and H is the subgraph of G induced by $\{1, \ldots, v\}$, then the discretization group of H is isomorphic to a subgroup of the discretization group of G: this follows

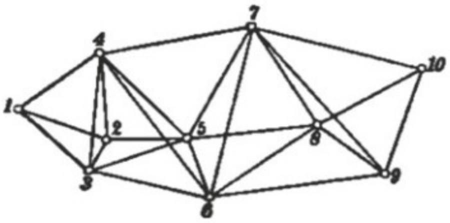

Fig. 5.4 A chain of 4-cliques in $\mathbb{R}^3$ [58].

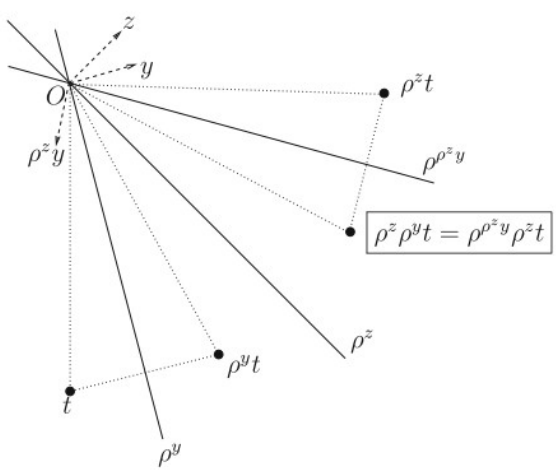

Fig. 5.5 Reflecting through ρ^y first and ρ^z second is equivalent to reflecting through ρ^z first and (the reflection of ρ^y through ρ^z) second.

by mapping $g_u \in \mathscr{G}_D(H)$ to $\bar{g}_u \in \mathscr{G}_D(G)$ (for each $u \leq v$) such that for $\bar{x}$ a realization of G and $x = (\bar{x}_1, \ldots, \bar{x}_v)$ the induced realization of H,

$$\bar{g}_u(\bar{x}) = (g_u(x), R_x^u(x_{v+1}), \ldots, R_x^u(x_n)).$$

The point of this remark is to emphasize that we can apply partial reflections in $\mathscr{G}_D(H)$ to realizations of G.

The most interesting feature of the discretization group is that its action on X is transitive: given a single realization $x^1 \in X$, for every realization $x \in X$ there is $g \in \mathscr{G}_D(G)$ such that $g(x^1) = x$. In other words, $\mathscr{G}_D(G)x^1 = X$. This can be shown by induction on n.

The induction starts because when $n = K + 1$ we have $X = \{x^1, x^2\}$, where x^2 is a reflection of x^1 with respect to $\{x_1, \ldots, x_K\}$ (see Sect. 3.3.4): so $g_{K+1}(x^2) = x^1$ and vice versa.

Now let $H = G[\{1, \ldots, n-1\}]$, $X_H = \mathsf{RealizeDMDGP}(H, d, K)$, and assume that the action of $\mathscr{G}_D(H)$ on X_H is transitive. Pick any two realizations x^1, x^2 of G in X, and consider the partial realizations consisting of their first $n-1$ vectors: let $x^1 = (\bar{x}^1, y)$ and $x^2 = (\bar{x}^2, z)$, where $\bar{x}^1, \bar{x}^2 \in \mathbb{R}^{K(n-1)}$ and $y, z \in \mathbb{R}^K$. Note that RealizeDMDGP computes X by extending partial realizations in X_H using Next, so $\bar{x}^1, \bar{x}^2$ must be in X_H by definition of Next.

By the induction hypothesis, since $\bar{x}^1, \bar{x}^2 \in X_H$, and $\mathscr{G}_D(H)$ is assumed to be transitive, there is $\bar{g} \in \mathscr{G}_D(H)$ such that $\bar{g}(\bar{x}^2) = \bar{x}^1$. Since every hyperplane W, given e.g. as K points in $\mathbb{R}^K$, splits $\mathbb{R}^K$ into two (open) half-spaces, which we can label as positive and negative, we can assign $+, -$ labels to both y and z, according to which side of the reflecting hyperplane W they are on (this labelling is known as an *orientation*). Since a partial reflection g_v will swap the orientation of every realization vector from the v-th onwards, either $\bar{g}$ or $g_n\bar{g}$ maps x^2 to x^1 and vice versa, which concludes the proof.

Limited to graphs with no pruning edges, this suggests an alternative implementation of RealizeDMDGP: whenever $S = \mathsf{Next}_d^K(W)$ is computed (see Alg. 5), arbitrarily dismiss one of the two elements, but record g_v at each level. At level n, when a realization x is found, simply compute X by applying all possible compositions of the group elements $g_{K+1}, \ldots, g_n$ to x. In this setting, x can be found in polytime, since there is no branching, but it still takes exponential time to list all possible subsets of the generators of $\mathscr{G}_D(G)$, see Fig. 5.6.

Lastly, remark that, since $\mathscr{G}_D(G)x = X$ for all $x \in X$, then $\mathscr{G}_D(G)X = \{gx \mid g \in \mathscr{G}_D(G) \wedge x \in X\} = X$: so X is *invariant* with respect to $\mathscr{G}_D(G)$, i.e., X does not change, as a set, under the action of the discretization group.

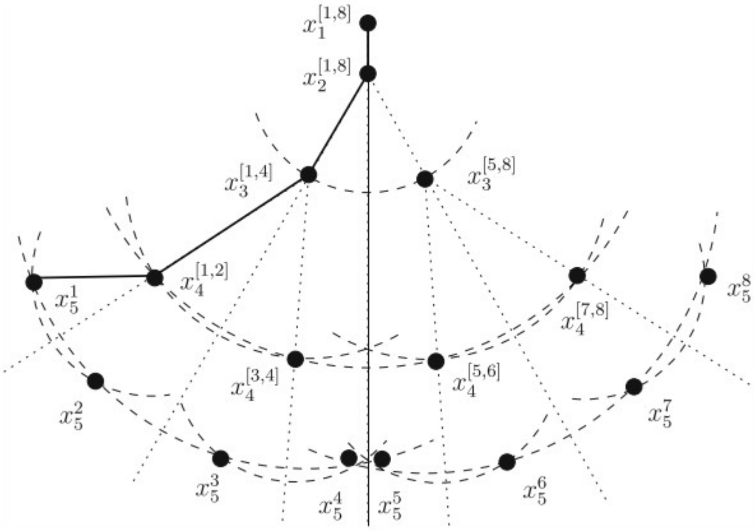

Fig. 5.6 Action of the discretization group $\mathscr{G}_D(G) = \langle g_3, g_4, g_5 \rangle$ on a realization $x^1 = (x_1^1, \ldots, x_5^1)$ is transitive and commutative: we have $g_5(x^1) = x^2, g_4(x^1) = x^4, g_5 g_4(x^1) = g_4 g_5(x^1) = x^3, g_3(x^1) = x^8, g_5 g_3(x^1) = g_3 g_5(x^1) = x^7, g_4 g_3(x^1) = g_3 g_4(x^1) = x^5, g_5 g_4 g_3(x^1) = g_5 g_3 g_4(x^1) = g_4 g_5 g_3(x^1) = g_4 g_3 g_5(x^1) = g_3 g_5 g_4(x^1) = g_3 g_4 g_5(x^1) = x^6$. The notation $x_j^{[i,\ell]}$ indicates that the neighboring dot represents the position of the (equal) point vectors $x_j^i, x_j^{i+1}, \ldots, x_j^{\ell} \in \mathbb{R}^K$, for $j \in \{1, \ldots, 5\}$ and $x^i \in X$ for all $i \leq 8$.

5.2.3 The pruning group

The results of Sect. 5.2.2 rely on G being a chain of $(K+1)$-cliques. But our motivation for studying $^{\mathsf{K}}$DMDGP graphs is that they provide a good model for protein backbones. The issue is that protein graphs have more edges besides those necessary for the existence of their contiguous trilateration order: in other words, their pruning edge set is nonempty. These edges, and their associated distance values, give RealizeDMDGP the information that is necessary in order to prune realizations that are not consistent with the measured distances; therefore, it would be unwise to disregard them. Below, we study how partial reflection symmetries behave in the presence of pruning edges.

Let $G = (V, E)$ be a weighted $^{\mathsf{K}}$DMDGP graph and d an edge weight function. We partition E into discretization edges E_D and pruning edges E_P. The graph $G[E_D]$ induced by the discretization edges is a minimal $^{\mathsf{K}}$DMDGP graph, and hence, by the results in Sect. 5.2.2, its realization set $\bar{X}$ is invariant with respect to its discretization group. We define $\mathscr{G}_D(G) = \mathscr{G}_D(G[E_D])$.

Figure 5.7 shows the partial planar realizations computed by RealizeDMDGP in the first five BP tree levels of an instance where $K = 2$. The positions of vertices at level j arise as the intersection of two circles (dashed arcs) centered at vertex levels $j-1$ and $j-2$, with respective radii $d_{j-1,j}, d_{j-2,j}$. The partial reflection symmetry at each level is defined across the dotted lines through the circle centers at the preceding two levels. The solid concentric circles at levels $1, \ldots, 5$ are all centered at $x_1^{[1,8]}$, and their radii are the distance values on pruning edges which are compatible with the given discretization edges. The crucial observation is the symmetric placement of the points x_j^i on the solid circles, for $i \in \{1, \ldots, 8\}$ and $j \in \{3, 4, 5\}$. More precisely, pairs like $x_4^{[1,2]}, x_4^{[7,8]}$ or $x_4^{[3,4]}, x_4^{[5,6]}$, which are reflections along the line defined by x_1, x_2, are at the same distance from $x_1^{[1,8]}$. So, if $X \neq \varnothing$ and we had a pruning edge $\{1, 4\}$, the distance d_{14} would either prune $x_4^{[1,2]}, x_4^{[7,8]}$ or $x_4^{[3,4]}, x_4^{[5,6]}$; it could

Affine Geometry:
Incidence with Parallelism (IP).
Obs: Look at Chapter 2 and 3

– Euclidean distance geometry...pdf
Pag. 48(60) – Fig. 5.5
Reflecting (analisar o que é melhor, reflecting or projection orthogonal.)

↘ ⇓

Reflecting through ρ^y first and ρ^z second is equivalent to reflecting through ρ^z first and (the reflection of ρ^y through ρ^z) second.

Pag. 49(61) Fig. 5.6
Action of discretization group $\mathcal{G}_D(G) = \langle g_3, g_4, g_5 \rangle$ on a realization $x^1 = (x^1_1, \dots x^1_5)$ is transitive and commutative: ...

Usar esta forma de representação de um plano que posteriormente, será curvo e projetado ou refletido sobre uma superfície plana.

Obs: Usar os pontos pretos para fazer triangulação com 3 pontos (Pappus's Theorem)

↓ usar os conceitos de curvas e manter a linha reta estabelecido por Pappus's Theorem.

Look book: /20210921/
Pappus-Theorem/
Perspectives on Projective
Geometry...pdf

– 20210921/
Desargues/
Two-dimensional Spaces, Volume 3...pdf

Pág. 55(68) - Chapter 3
Gravity As Curvature

Pág. 57(70) - Chapter 4
Curvature by Polyhedral Approximation

)*+,)))- !) % 6 # ! 3! G 1
C # ! 2 5 ! # 8! 3#
' // " !3 1

CHAPTER 3

Gravity As Curvature

We now turn to the description of curvature.

3.1. Einstein Identified Gravity with the Curvature of Space-Time

Albert Einstein identified gravity with the curvature of space-time: A massive object distorts, or curves, space-time; and other objects respond gravitationally by following (locally) shortest paths in that curved space, paths that deserve to be called straight. Misner, Thorne, and Wheeler, in their book "Gravitation" [**97**], compare the path followed by an object acting under the influence of gravity to a straight cut in the surface of an apple where this cut, near the stem of the apple, dips toward the stem as if that stem were a source of gravitational force.

Two problems must be solved in application of Einstein's ideas: (1) We usually recognize the curvature of an object by observing how it curves in some larger space (= extrinsic curvature). Since we live *within* space-time, how can we recognize and describe its curvature from within space-time itself (= intrinsic curvature)? (2) How do we recognize and describe the paths in a curved object that are locally shortest, hence deserve to be called *straight* (= geodesic)?

Gauss showed us how these problems can be solved for 2-dimensional surfaces. Riemann generalized Gaussian curvature to higher dimensions by considering the Gaussian curvature of each 2-dimensional slice in higher dimensional space.

In this introduction to curvature, we will restrict ourselves, as Gauss did, to 2-dimensional surfaces that reside in 3-dimensional Euclidean space. But we will not follow Gauss's path. For both questions, we will first approximate the surface by a polyhedral surface where it is easy to answer both questions: How flat is a polyhedral surface? How straight is an edge path in a polyhedral surface? The corresponding concepts in smooth surfaces will then follow by a limiting process.

CHAPTER 4

Curvature by Polyhedral Approximation

4.1. Approximating Smooth Surfaces by Polyhedra

Gauss described a smooth surface in $\mathbb{R}^3$ locally as the image of a map $S : \mathbb{R}^2 \to \mathbb{R}^3 : (x, y) \mapsto (x, y, z)$, where z is a many-times differentiable function of x and y. It is an easy matter to triangulate such a surface.

Key Idea 4.1. *Every smooth surface S in $\mathbb{R}^3$ can be closely approximated locally by polyhedral surfaces S''.*

We restrict our attention to piecewise smooth disks in the surface S.

Definition 4.2. A disk D in a smooth surface S is called *piecewise smooth* if there are finitely many points p_0, p_1, …, $p_n = p_0$ in the boundary ∂D of D that divide ∂D into arcs p_0p_1, p_1p_2, …, $p_{n-1}p_n$ each of which is smooth and such that there is a well-defined angle of turning in ∂D at each of the points p_1, p_2, …, p_n.

Theorem 4.3. *If $D \subset \mathbb{R}^2$ is a disk that is mapped homeomorphically by $S : \mathbb{R}^2 \to \mathbb{R}^3$ to a piecewise smooth disk $S(D) \subset S$, then $S(D)$ can be closely approximated by polyhedral disks $D' \subset \mathbb{R}^3$.*

Proof. Choose points q_0, q_1, …, $q_m = q_0$ in cyclic order on ∂D that divide ∂D into tiny arcs. Since D is piecewise smooth, the points q_i will be the vertices of a polygonal simple closed curve J in $\mathbb{R}^2$ that closely approximates ∂D. The curve J bounds a disk $D' \subset \mathbb{R}^2$ that approximates D. The disk D' can be triangulated with very small triangles T_j. If A, B, and C are the vertices of T_j, then S maps them to vertices $A' = S(A)$, $B' = S(B)$, and $C' = S(C)$. The vertices A', B', and C' are the vertices of a triangle $T'_j \subset \mathbb{R}^3$. The triangles T'_j are the triangles of a polyhedral disk D' that closely approximates $S(D)$. ☐

4.2. The Curvature of a Polyhedral Disk

In this chapter we define the curvature of a polyhedral disk in $\mathbb{R}^3$. We shall see that curvature is the average obstruction to flattening that disk into the plane locally without distorting lengths or areas.

4.3. How Flat Is a Disk?

We consider a polyhedral disk D, not assumed to be flat, triangulated by finitely many Euclidean triangles. Combinatorially, we can picture the disk as lying in the flat plane as in Figure 1, but to do this we may have to change angles and side lengths of the individual triangles.

If we fix our intention on one of the interior vertices v of our disk, then the triangles containing that vertex form a polyhedral cone with that vertex as tip of the cone. This polyhedral cone is called the *star* of the vertex v in D and is denoted

↘ Usar esta forma de representação
de um plano que posteriormente,
será curvo e projetado ou
refletido sobre uma superfície
plana.

Obs: Usar os pontos pretos para fazer
triangulação com 3 pontos
(Pappus's Theorem)

↓ usar os
conceitos de curvas e manter a
linha reta estabelecido por Pappus's
Theorem.

Look book: /20210921/
Pappus-Theorem/
Perspectives on Projective
Geometry...pdf

— 20210921/
Desargues/
Two-dimensional Spaces, Volume 3...pdf

→ Pág. 55(68) - Chapter 3
Gravity As Curvature

Pág. 57(70) - Chapter 4
Curvature by Polyhedral Approximation

Key Idea 4.1

Every smooth surface S in $\mathbb{R}^3$ can be closely approximated locally by polyhedral surfaces S''.

Return:

Pág. 55 (68) – Chapter 3

extrinsic curvature

intrinsic curvature

straight (= geodesic)

GAUSS showed us how these problems can be solved for 2-dimensional surfaces.

RIEMANN generalized Gaussian curvature to higher dimensions by considering the Gaussian curvature of each 2-dimensional slice in higher dimensional space.

Question:

How flat is a polyhedral surface?

How straight is an edge path in a polyhedral surface?

Pág. 69 (82) - Chapter 5
Curvature as a length derivative.
Pág. 78 (91) - 5.4
Evaluating the Double Limit
Figure 11 Calculating geodesic curvature.

- 20210921/ Desargues/
Hexagon Theorem of Pappus...pdf
→ Theorems of Desargues and Pappus.

- 20210921/ Desargues/
Geometría Moderna...pdf
Pág. 235 (250) - Capítulo 10
Teorema de Pappus, Pascal y Brianchon.
Teorema 10.1 (Carnot para rectas).

- 20210921/ Desargues/
ch1.pdf → Pappus's theorem:

)*+,)))- !) % ; 9! 6
!33 ! ! 3#

8)*+,)))- !) % 2 J! # ! 3#
' 1/ / !3J 6 ! # !33 G ! ! 5 . !

Capítulo 10

Teorema de Pappus, Pascal y Brianchon

Usaremos los resultados anteriores para demostrar los teoremas de Pappus, Pascal y Brianchon. Demostraremos los siguientes resultados debidos a Carnot.

Teorema 10.1 (Carnot para rectas) *Si la recta l intersecta los lados BC, CA y AB del triángulo $\triangle ABC$ en los puntos P, Q y R (pero no uno de los vértices) y la recta l' intersecta los lados BC, CA y AB en los puntos P', Q' y R' (pero no uno de los vértices), entonces*

$$\frac{AR}{RB}\frac{BP}{PC}\frac{CQ}{QA}\frac{AR'}{R'B}\frac{BP'}{P'C}\frac{CQ'}{Q'A} = 1$$

Demostración: Usando el teorema de Menelao con las rectas l y l'.

$$\begin{aligned}\frac{AR}{RB}\frac{BP}{PC}\frac{CQ}{QA} &= -1 \\ \frac{AR'}{R'B}\frac{BP'}{P'C}\frac{CQ'}{Q'A} &= -1\end{aligned}$$

que multiplicadas nos dan la igualdad de Carnot. ■

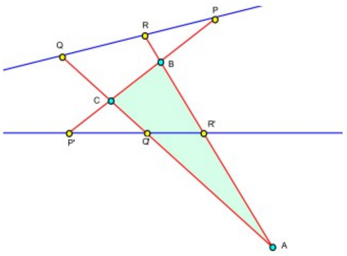

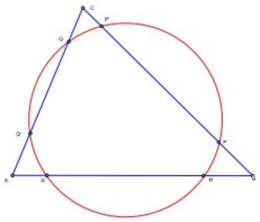

Teorema 10.2 (Carnot) *Si un círculo intersecta los lados BC, CA y AB del triángulo $\triangle ABC$ en los puntos P, P'; Q, Q'; R, R'; respectivamente (pero no uno de los vértices), entonces*

$$\frac{AR}{RB}\frac{BP}{PC}\frac{CQ}{QA}\frac{AR'}{R'B}\frac{BP'}{P'C}\frac{CQ'}{Q'A}=1$$

Demostración: Las potencias de los puntos A, B y C respecto al círculo son:

$$\begin{aligned} AR\cdot AR' &= AQ\cdot AQ' \\ BP\cdot BP' &= BR\cdot BR' \\ CQ\cdot CQ' &= CP\cdot CP' \end{aligned}$$

que multiplicadas nos dan la igualdad de Carnot ■

Los teoremas anteriores se generalizan al caso de una cónica, recordando que un círculo y dos rectas son casos especiales de secciones planas de un cono. Este resultado se puede demostrar usando el corolario 3.75, pag. 114.

Teorema 10.3 (Carnot) *Si una cónica intersecta los lados BC, CA y AB del triángulo $\triangle ABC$ en los puntos P, P'; Q, Q'; R, R'; respectivamente (pero no uno de los vértices), entonces*

$$\frac{AR}{RB}\frac{BP}{PC}\frac{CQ}{QA}\frac{AR'}{R'B}\frac{BP'}{P'C}\frac{CQ'}{Q'A}=1$$

Usando los teoremas de Carnot podemos demostrar ahora los siguientes importantes teoremas. La demostración es la misma para ambos teoremas.

Teorema 10.4 (Pappus) *Si A, B, C son puntos de una recta y D, E y F son puntos de otra recta, entonces los puntos $R = AE\cap BD$, $P = BF\cap CE$ y $Q = AE\cap BD$ están alineados.*

Teorema 10.5 (Pascal) *Si A, B, C, D, E y F son puntos en un círculo (o una cónica), entonces los puntos $R = AE\cap BD$, $P = BF\cap CE$ y $Q = AE\cap BD$ están alineados.*

Demostración: Si $S = AE\cap DC$, $T = BF\cap CD$ y $U = AE\cap BF$ y usamos el teorema de Menelao en el triángulo $\triangle STU$ con las rectas AF, BD y CE obtenemos

$$\begin{aligned} \frac{SQ}{QT}\frac{TF}{FU}\frac{UA}{AS} &= -1 \\ \frac{SD}{DT}\frac{TB}{BU}\frac{UR}{RS} &= -1 \\ \frac{SC}{CT}\frac{TP}{PU}\frac{UE}{ES} &= -1 \end{aligned}$$

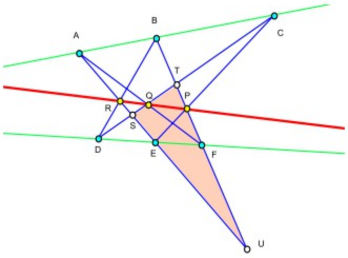

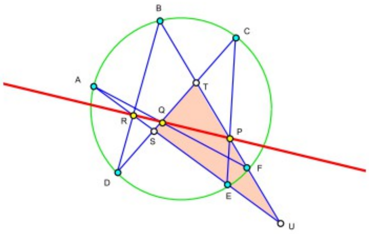

y del teorema de Carnot con el triángulo $\triangle STU$

$$\frac{SC}{CT}\frac{TB}{BU}\frac{UA}{AS}\frac{SD}{DT}\frac{TF}{FU}\frac{UE}{ES}=1$$

si multiplicamos las tres primeras igualdades y el resultado lo dividimos entre la última obtenemos

$$\frac{SQ}{QT}\frac{TP}{PU}\frac{UR}{RS}=-1$$

y usando el inverso del teorema de Menelao obtenemos el resultado. ■

Con la notación de los teoremas anteriores denotaremos por $\left\{\begin{smallmatrix} ABC \\ DEF \end{smallmatrix}\right\}$ la recta de Pappus o Pascal respectivamente.

Problemas

Problema 1 *Usar el problema 18, pág. 128,demostrar el teorema de Pappus. (Ayuda: Usar los triángulos de lados AE, BF y CD; y BD, CE y AR).*

Problema 2 **Dado un triángulo ΔABC y un punto P; sean X_a, X_b, P_a y P_b las proyecciones de P en las rectas AP, BP, AC y BC respectivamente. Entonces las rectas X_aP_a, X_bP_b y AB son concurrentes.*

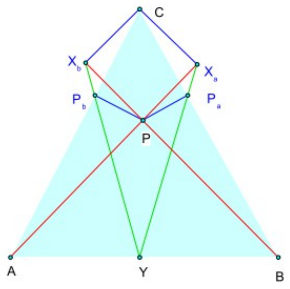

Figura 10-1

Problema 1 **Si A, B, C, D y E son puntos en un círculo $\mathcal{C}$ y F es otro punto tales que los puntos $R = AE \cap BD$, $P = BF \cap CE$ y $Q = AE \cap BD$ están alineados. Entonces $F \in \mathcal{C}$. (Este es un inverso parcial del teorema de Pascal)*

Problema 2 **Dados un triángulo ΔABC, un punto P y una recta p que pasa por P; sean $X_a = p \cap a$, $X_b = p \cap b$, $X_c = p \cap c$, Y_A, Y_B y Y_C las otras intersecciones de AP, BP y CP con el circuncírculo $\mathcal{C}$ del triángulo. Entonces X_aY_A, X_bY_B y X_cY_C son concurrentes.*

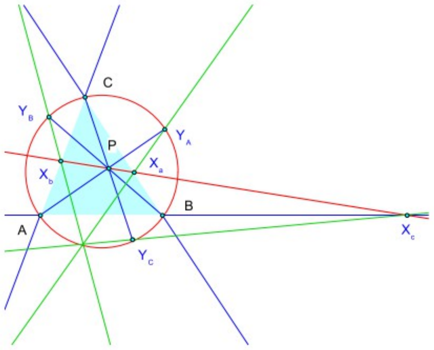

Figura 10-2

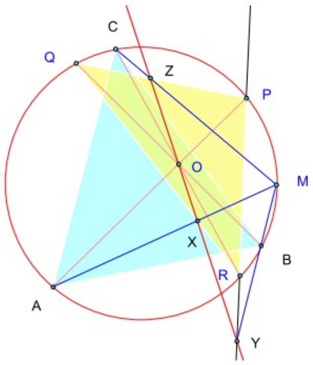

Figura 10-3

Problema 3 *Dados dos triángulos* ΔABC *y* ΔPQR *en perspectiva desde un punto* O *e inscritos en un círculo* $\mathcal{C}$. *Entonces para todo punto* $M \in \mathcal{C}$ *los puntos* O, $X = AM \cap p$, $Y = BM \cap q$ *y* $Z = CM \cap r$ *están alineados.*

Problema 4 **Dado un triángulo* ΔABC *y dos puntos* P *y* Q *conjugados isogonales con triángulos pedales* $\Delta P_aP_bP_c$ *y* $\Delta Q_aQ_bQ_c$ *respectivamente. Entonces los puntos* $X = P_bQ_c \cap P_cQ_b$, $Y = P_cQ_a \cap P_aQ_c$ *y* $Z = P_aQ_b \cap P_bQ_a$ *están en la recta* PQ.

Rectas de Pappus*

Entre las rectas Pappus hay varias relaciones geométricas, como lo muestra el siguiente teoremas.

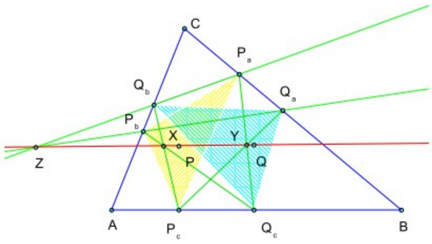

Figura 10-4

Teorema 10.6 (Steiner) *Si A, B, C son puntos de una recta y D, E y F son puntos de otra recta, entonces las rectas $\left\{ {ABC \atop DEF} \right\}$, $\left\{ {ABC \atop EFD} \right\}$ y $\left\{ {ABC \atop FDE} \right\}$ son concurrentes en un punto S^*; también las rectas $\left\{ {ABC \atop EDF} \right\}$, $\left\{ {ABC \atop FED} \right\}$ y $\left\{ {ABC \atop DFE} \right\}$ son concurrentes en un punto S.*

Demostración: Si definimos los triángulos:

- Triángulo $\triangle_1 = \triangle P_1P_2P_3$ de lados AE, CD y BF.
- Triángulo $\triangle_2 = \triangle Q_1Q_2Q_3$ de lados BD, AF y CE.
- Triángulo $\triangle_3 = \triangle R_1R_2R_3$ de lados CF, BE y AD.

Tenemos los siguientes resultados:

Los puntos P_1, Q_1 y R_1 (resp. P_2, Q_2 y R_2; resp. P_3, Q_3 y R_3) definen las rectas de Pappus $c_1 = \left\{ {ABC \atop EDF} \right\}$, $c_2 = \left\{ {ABC \atop FED} \right\}$ y $c_3 = \left\{ {ABC \atop DFE} \right\}$. Los lados correspondientes de los triángulos $\triangle_1$ y $\triangle_2$ se cortan el las rectas de Pappus $x_1 = \left\{ {ABC \atop DEF} \right\}$. Por lo tanto $\triangle_1$ y $\triangle_2$ están en perspectiva y por lo tanto las rectas $\left\{ {ABC \atop DEF} \right\}$, $\left\{ {ABC \atop EFD} \right\}$ y $\left\{ {ABC \atop FDE} \right\}$ son concurrentes, esto es que los triángulos $\triangle_1$, $\triangle_2$ y $\triangle_3$ tienen un centro común de perspectiva.

Análogamente lados correspondientes de los triángulos $\triangle_2$ y $\triangle_3$ (resp. $\triangle_3$ y $\triangle_1$) se cortan el las rectas de Pappus $x_2 = \left\{ {ABC \atop EFD} \right\}$ y $x_3 = \left\{ {ABC \atop FDE} \right\}$. El problema 4, pág. 92 nos dice que los ejes de perspectiva son concurrentes, esto es que las rectas $\left\{ {ABC \atop EDF} \right\}$, $\left\{ {ABC \atop FED} \right\}$ y $\left\{ {ABC \atop DFE} \right\}$ son concurrentes en un punto S. ■

Definición 10.7 *Si A, B, C son puntos de una recta y D, E y F son puntos de otra recta, entonces los puntos S^* y S donde se intersectan las rectas $\left\{ {ABC \atop DEF} \right\}$, $\left\{ {ABC \atop EFD} \right\}$ y $\left\{ {ABC \atop FDE} \right\}$; y $\left\{ {ABC \atop EDF} \right\}$, $\left\{ {ABC \atop FED} \right\}$ y $\left\{ {ABC \atop DFE} \right\}$ respectivamente, se llaman los puntos de Steiner.*

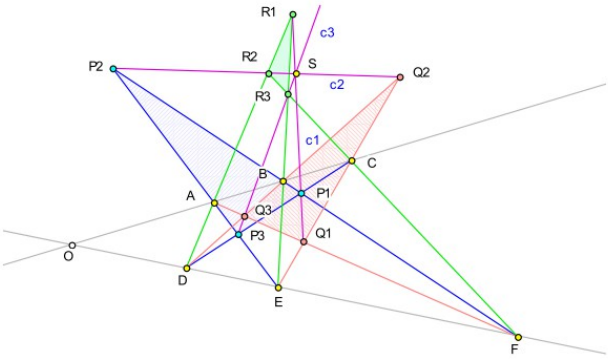

Figura 10-5

Rectas de Pascal*

Corolario 10.8 (Pascal) *Si A, B, C, D, E y F son puntos en una cónica entonces los puntos $R = AE \cap BD$, $P = BF \cap CE$ y $Q = AE \cap BD$ están alineados.*

Definición 10.9 *Si A, B, C, D, E y F son puntos en un cónica, entonces la recta que une los puntos $R = AE \cap BD$, $P = BF \cap CE$ y $Q = AE \cap BD$ se llama la recta de Pascal de los puntos dados. Usaremos la notación: $\left\{\begin{smallmatrix}ABC\\DEF\end{smallmatrix}\right\}$ para esta recta.*

El enunciado clasico de estos teoremas es el siguiente: Dado un exágono $ABCDEF$ inscrito en un cónica, entonces los lados opuestos del exágono se intersectan en tres puntos alineados. Sin embargo como los resultados más interesantes usan distintos ordenes en los puntos dados la notación anterior es más adecuada. Hay que notar que si $p = \left\{\begin{smallmatrix}P_1P_2P_3\\Q_1Q_2Q_3\end{smallmatrix}\right\}$ etonces: $p = \left\{\begin{smallmatrix}P_{i_1}P_{i_2}P_{i_3}\\Q_{i_1}Q_{i_2}Q_{i_3}\end{smallmatrix}\right\}$ y $p = \left\{\begin{smallmatrix}Q_{i_1}Q_{i_2}Q_{i_3}\\P_{i_1}P_{i_2}P_{i_3}\end{smallmatrix}\right\}$ para toda permutación $(i_1i_2i_3)$ de (123). Por lo tanto hay 6 rectas de Pappus asociadas a seis puntos en dos rectas. Análogamente hay $60 = \frac{6!}{3!\cdot 2}$ rectas de Pascal asociadas a seis puntos de una cónica.

Entre estas rectas hay varias relaciones geométricas, como lo muestran los siguientes teoremas.

Claramente la misma demostración que la del teorema 10.6; pag. 239, nos da el siguiente teorema.

Teorema 10.10 (Steiner) *Si A, B, C, D, E y F son puntos en una cónica entonces las rectas $\left\{\begin{smallmatrix}ABC\\DEF\end{smallmatrix}\right\}$, $\left\{\begin{smallmatrix}ABC\\EFD\end{smallmatrix}\right\}$ y $\left\{\begin{smallmatrix}ABC\\FDE\end{smallmatrix}\right\}$ son concurrentes en un punto S; también las rectas $\left\{\begin{smallmatrix}ABC\\EDF\end{smallmatrix}\right\}$, $\left\{\begin{smallmatrix}ABC\\FED\end{smallmatrix}\right\}$ y $\left\{\begin{smallmatrix}ABC\\DFE\end{smallmatrix}\right\}$ son concurrentes en un punto S^*.* ■

Podemos enunciar el teorema de la siguiente forma: Dados A, B, C, D, E y F son puntos en una cónica, entonces para cada partición de los puntos dados en subconjuntos de tres puntos P_1, P_2 y P_3; Q_1, Q_2 y Q_3, entonces las rectas $\left\{\begin{smallmatrix}P_1P_2P_3\\Q_1Q_2Q_3\end{smallmatrix}\right\}$, $\left\{\begin{smallmatrix}P_1P_2P_3\\Q_2Q_3Q_1\end{smallmatrix}\right\}$ y $\left\{\begin{smallmatrix}P_1P_2P_3\\Q_3Q_1Q_2\end{smallmatrix}\right\}$ son concurrentes en un punto S; también las rectas $\left\{\begin{smallmatrix}P_1P_2P_3\\Q_1Q_3Q_2\end{smallmatrix}\right\}$, $\left\{\begin{smallmatrix}P_1P_2P_3\\Q_3Q_2Q_1\end{smallmatrix}\right\}$ y $\left\{\begin{smallmatrix}P_1P_2P_3\\Q_2Q_2Q_3\end{smallmatrix}\right\}$ son concurrentes en un punto

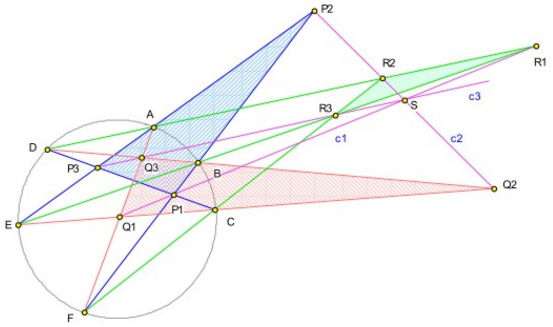

Figura 10-6

S^*. Si suponemos que P_1, P_2 y P_3 y Q_1, Q_2 y Q_3 están en el mismo orden lexicográfico se usará la notación $S = S_{P_1P_2P_3}$ y $S^* = S^*_{P_1P_2P_3}$. Los puntos S y S^* se llaman los puntos de Steiner de $P_1P_2P_3$. Tenemos que $S = S_{Q_1Q_2Q_3}$ y $S^* = S^*_{Q_1Q_2Q_3}$. Por lo tanto podemos suponer que $P_1 = A$ y que $P_1P_2P_3$ está ordenado lexicográfico.

Dados A, B, C, D, E y F puntos en una cónica entonces definimos los siguientes elementos:

A) Los puntos A, B, C, D, E y F los llamaremos los puntos P.

B) Las 15 rectas AB,..., EF las llamamos las rectas C.

C) Cada una de las rectas C, por ejemplo AB, es cortada por las otras en A, B y otros 6 puntos, que llamaremos puntos c. Hay $45 = \frac{6\cdot 15}{2}$ puntos c. Estos puntos son los puntos $AE \cap BD$,..., $CD \cap EF$.

D) Por cada punto c, $AE \cap BD$ por ejemplo, pasan 4 rectas de Pascal $\left\{\begin{smallmatrix}ABC\\DEF\end{smallmatrix}\right\}$, $\left\{\begin{smallmatrix}ABF\\DEC\end{smallmatrix}\right\}$, $\left\{\begin{smallmatrix}ADC\\BEF\end{smallmatrix}\right\}$ y $\left\{\begin{smallmatrix}ADF\\BEC\end{smallmatrix}\right\}$. Por lo tanto hay $60 = \frac{45\cdot 4}{3}$ rectas de Pascal.

E) A los puntos de Steiner los llamaremos puntos g. En cada recta de Pascal hay un solo punto de Steiner. Por lo tanto hay $20 = \frac{60}{3}$ puntos g. Para cada tercia de puntos, ABC por ejemplo, hay dos puntos de Steiner que llamaremos conjugados; los denotaremos por S_{ABC} y S^*_{ABC}.

Los siguientes teoremas nos permitiran definir más elementos asociados a seis puntos de una cónica.

Teorema 10.11 (Kirkman) *Si A, B, C, D, E y F son puntos en una cónica, entonces las rectas de Pascal $\left\{\begin{smallmatrix}ABF\\DEC\end{smallmatrix}\right\}$, $\left\{\begin{smallmatrix}ADC\\FBE\end{smallmatrix}\right\}$ y $\left\{\begin{smallmatrix}EBC\\AFD\end{smallmatrix}\right\}$ son concurrentes.*

Demostración: Sean los triángulos:

- Triángulo $\triangle_1 = \triangle P_1P_2P_3$ de lados AE, BD y CF.
- Triángulo $\triangle_2 = \triangle Q_1Q_2Q_3$ de lados las rectas de Pascal $\left\{\begin{smallmatrix}CFA\\EBD\end{smallmatrix}\right\}$, $\left\{\begin{smallmatrix}CFD\\EBA\end{smallmatrix}\right\}$ y $\left\{\begin{smallmatrix}EFA\\CBD\end{smallmatrix}\right\}$.

Estos triángulos tienen las siguientes propiedades:

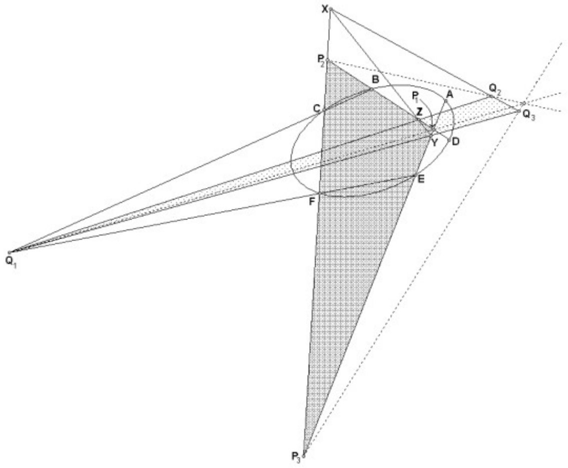

Figura 10-7

Los vértices de $\triangle_1$ son $BD \cap CF$, $CF \cap AE$ y $AE \cap BD$; Los vértices de $\triangle_2$ son $AC \cap DE$, $AB \cap DF$ y $BC \cap EF$. Por lo tanto los vértices correspondientes de los triángulos están en en las rectas de Pascal $\left\{\begin{smallmatrix}EBC\\AFD\end{smallmatrix}\right\}$, $\left\{\begin{smallmatrix}ADC\\FBE\end{smallmatrix}\right\}$ y $\left\{\begin{smallmatrix}ABF\\DEC\end{smallmatrix}\right\}$.

Las intersecciones de los lados correspondientes de los triángulos $\triangle_1$ y $\triangle_2$ son $AE \cap CD$, $BD \cap AF$ y $CF \cap BE$ que definen la recta de Pascal $\left\{\begin{smallmatrix}ABC\\DFE\end{smallmatrix}\right\}$, lo que implica que los triángulos $\triangle_1$, $\triangle_2$ están en perspectiva. Por lo tanto las rectas de Pascal $\left\{\begin{smallmatrix}EBC\\AFD\end{smallmatrix}\right\}$, $\left\{\begin{smallmatrix}ADC\\FBE\end{smallmatrix}\right\}$ y $\left\{\begin{smallmatrix}ABF\\DEC\end{smallmatrix}\right\}$ son concurrentes. ■

Nota: Llamaremos al punto $\left[\begin{smallmatrix}ABC\\DEF\end{smallmatrix}\right] = \left\{\begin{smallmatrix}ABF\\DEC\end{smallmatrix}\right\} \cap \left\{\begin{smallmatrix}ADC\\FBE\end{smallmatrix}\right\} \cap \left\{\begin{smallmatrix}EBC\\AFD\end{smallmatrix}\right\}$ un punto de Kirkman. En general el punto de Kirkman $\left[\begin{smallmatrix}P_1P_2P_3\\Q_1Q_2Q_3\end{smallmatrix}\right]$ es la intersección $\left\{\begin{smallmatrix}P_1P_2Q_3\\Q_1Q_2P_3\end{smallmatrix}\right\} \cap \left\{\begin{smallmatrix}P_1Q_1P_3\\Q_3P_2Q_2\end{smallmatrix}\right\} \cap \left\{\begin{smallmatrix}Q_2P_2P_3\\P_1Q_3Q_1\end{smallmatrix}\right\}$.

Teorema 10.12 *Si A, B, C, D, E y F son puntos en una cónica, entonces el punto de Steiner S_{ABC} y los puntos de Kirkman $\left[\begin{smallmatrix}ABC\\DEF\end{smallmatrix}\right]$, $\left[\begin{smallmatrix}ABC\\EFD\end{smallmatrix}\right]$ y $\left[\begin{smallmatrix}ABC\\FDE\end{smallmatrix}\right]$ están alineados.*

Demostración: Sean los triángulos:

- Triángulo $\triangle_1 = \triangle P_1P_2P_3$ de lados BD, AF y CE.
- Triángulo $\triangle_2 = \triangle Q_1Q_2Q_3$ de lados las rectas de Pascal $\left\{\begin{smallmatrix}ABF\\DEC\end{smallmatrix}\right\}$, $\left\{\begin{smallmatrix}AEC\\DBF\end{smallmatrix}\right\}$ y $\left\{\begin{smallmatrix}DBC\\AEF\end{smallmatrix}\right\}$.
- Triángulo $\triangle_3 = \triangle R_1R_2R_3$ de lados las rectas de Pascal $\left\{\begin{smallmatrix}EBC\\AFD\end{smallmatrix}\right\}$, $\left\{\begin{smallmatrix}ABD\\EFC\end{smallmatrix}\right\}$ y $\left\{\begin{smallmatrix}AFC\\EBD\end{smallmatrix}\right\}$.

Estos triángulos tienen las siguientes propiedades:

Los vértices de $\triangle_1$ son $AF \cap CE$, $BD \cap CE$ y $BD \cap AF$; Los vértices de $\triangle_2$ son $AB \cap DE$, $AC \cap DF$ y $BC \cap EF$; Los vértices de $\triangle_3$ son $BC \cap DF$, $AB \cap EF$ y $AC \cap DE$. Por lo tanto los vértices correspondientes de los triángulos están en en las rectas de Pascal $\left\{\begin{smallmatrix}FEB\\CAD\end{smallmatrix}\right\}$, $\left\{\begin{smallmatrix}EDA\\BCF\end{smallmatrix}\right\}$ y $\left\{\begin{smallmatrix}DFC\\ABE\end{smallmatrix}\right\}$.

Las intersecciones de los lados correspondientes de los triángulos $\triangle_1$ y $\triangle_2$ son $BD \cap AE$, $AF \cap CD$ y $CE \cap BF$ que están en la recta de Pascal $\left\{{ABC \atop DEF}\right\}$, lo que implica que los triángulos $\triangle_1$, $\triangle_2$ y $\triangle_3$ tienen un centro común de perspectiva. Las intersecciones de los lados correspondientes de los triángulos $\triangle_1$ y $\triangle_3$ son $BD \cap CF$, $AF \cap BE$ y $CE \cap AD$ que están en la recta de Pascal $\left\{{ABC \atop EFD}\right\}$. Por lo tanto los ejes de perspectiva de los triángulos $\triangle_1$ y $\triangle_2$ y de los triángulos $\triangle_1$ y $\triangle_3$ se intersectan en el punto de Steiner S_{ABC}.

Por lo tanto el eje de perspectiva de los triángulos $\triangle_2$ y $\triangle_3$ pasa por S_{ACE}, pero este eje esta definido por los puntos de Kirkman $\left[{ABC \atop DEF}\right]$, $\left[{ABC \atop EFD}\right]$ y $\left[{ABC \atop FDE}\right]$. ■

Por lo tanto tenemos las siguientes propiedades asociadas a seis puntos en una cónica

F) Claramente la recta de Pascal $\left\{{P_1P_2P_3 \atop Q_1Q_2Q_3}\right\}$ contiene los puntos de Kirkman siguientes: $\left[{P_1P_2P_3 \atop Q_1Q_2Q_3}\right]$, $\left[{P_3P_1P_2 \atop Q_3Q_1Q_2}\right]$ y $\left[{P_2P_3P_1 \atop Q_2Q_3Q_1}\right]$. Y por lo tanto hay 60 puntos de Kirkman.

G) Los 60 puntos de Kirkman están en 20 rectas una por cada punto de Steiner, las llamamos las rectas G.

Teorema 10.13 *Los puntos de Steiner S^*_{ABC}, S^*_{ABD}, S^*_{AED} y S^*_{AEC} están alineados.*

Demostración: Sean las rectas de Pascal siguientes: $x_1 = \left\{{ABF \atop DEC}\right\}$, $x_2 = \left\{{ACD \atop FEB}\right\}$, $x_3 = \left\{{ADF \atop ECB}\right\}$, $p_1 = \left\{{AEC \atop DFB}\right\}$, $p_2 = \left\{{AED \atop BCF}\right\}$ y $p_3 = \left\{{ADB \atop CEF}\right\}$. Entonces tenemos los puntos de Kirman siguientes:

$$\begin{aligned}
O &= \begin{bmatrix} ABC \\ DEF \end{bmatrix} = x_1 \cap x_2 \cap x_3 \\
P_1 &= \begin{bmatrix} EAF \\ CBD \end{bmatrix} = x_1 \cap p_2 \cap p_3 \\
P_2 &= \begin{bmatrix} FDC \\ EAB \end{bmatrix} = p_1 \cap x_2 \cap p_3 \\
P_3 &= \begin{bmatrix} EDB \\ CFA \end{bmatrix} = p_1 \cap p_2 \cap x_3
\end{aligned}$$

Definimos ahora el triángulo $\triangle Q_1Q_2Q_3$, donde $Q_1 = DF \cap AC$, $Q_2 = ED \cap CB$ y $Q_3 = EF \cap AB$. Entonces $Q_i \in x_i$, $i = 1, 2, 3$, y además los lados del triángulo $\triangle Q_1Q_2Q_3$ son las rectas de Pascal $q_1 = \left\{{AEC \atop FBD}\right\}$, $q_2 = \left\{{AED \atop FBC}\right\}$ y $q_3 = \left\{{ADB \atop FCE}\right\}$.

Por lo tanto tenemos O es el centro de perspectiva de los triángulos $\triangle P_1P_2P_3$ y $\triangle Q_1Q_2Q_3$. Y por lo tanto las interseccións $p_i \cap q_i$, $i = 1, 2, 3$, están alineadas. Pero estas intersecciones son:

$$\begin{aligned}
p_1 \cap q_1 &= \begin{Bmatrix} AEC \\ DFB \end{Bmatrix} \cap \begin{Bmatrix} AEC \\ FBD \end{Bmatrix} = S^*_{AEC} \\
p_2 \cap q_2 &= \begin{Bmatrix} AED \\ BCF \end{Bmatrix} \cap \begin{Bmatrix} AED \\ FBC \end{Bmatrix} = S^*_{AED} \\
p_3 \cap q_3 &= \begin{Bmatrix} ADB \\ CEF \end{Bmatrix} \cap \begin{Bmatrix} ABD \\ FEC \end{Bmatrix} = S^*_{ABD}
\end{aligned}$$

Si intercambiamos los puntos B y E obtenemos que los puntos

$$\begin{aligned} p_1 \cap q_1 &= \left\{ \begin{matrix} ABC \\ DFE \end{matrix} \right\} \cap \left\{ \begin{matrix} ABC \\ FED \end{matrix} \right\} = S^*_{ABC} \\ p_2 \cap q_2 &= \left\{ \begin{matrix} ABD \\ ECF \end{matrix} \right\} \cap \left\{ \begin{matrix} ABD \\ FEC \end{matrix} \right\} = S^*_{ABD} \\ p_3 \cap q_3 &= \left\{ \begin{matrix} ADE \\ CBF \end{matrix} \right\} \cap \left\{ \begin{matrix} AED \\ FBC \end{matrix} \right\} = S^*_{AED} \end{aligned}$$

están alineados. Por lo tanto S^*_{ABC}, S^*_{ABD}, S^*_{AED} y S^*_{AEC} están alineados. ■

Las tercias de puntos ABC, ABD, ADE y ACE que definen los puntos de Steiner en el teorema anterior son las tercias tales que no tiene puntos comunes de la tercia las parejas $\{AF, BE, CD\}$, como hay 15 tercias de parejas distintas y cada una define un recta que contiene a 4 puntos de Steiner.

Más aun si $\{PX, QY, RZ\}$ es una tercia de parejas de puntos, entonces los puntos de Steiner que definen son los que están en las rectas de Pascal $\left\{ {PQR \atop XYZ} \right\}$, $\left\{ {PQZ \atop XYR} \right\}$, $\left\{ {PYR \atop XQZ} \right\}$ y $\left\{ {PYZ \atop XQR} \right\}$. De esta presentación se ve que si PQR y XYZ están en el mismo orden lexicográfico tenemos el punto S_{PQR} en caso contrario a S^*_{PQR}, igualmente con los otros tres puntos. Se deja como ejercicio demostrar que el número de puntos conjugados S^* es igual a 0, 2 o 4.

Teorema de Brianchon

Demostraremos ahora el teorema de Brianchon, que es el teorema dual al teorema de Pascal. La demostración que dio Brianchon usa la teoría de polos y polares, ver teorema 13.17, pag. 274, sin embargo por mucho tiempo se busco una demostración que usara únicamente herramientas euclideanas. No es hasta 1961 que Smogorzhevskil, ver [38], dio una demostración de este estilo. Usaremos el concepto de centro radical para demostrar el teorema de Brianchon. Para esto necesitamos el siguiente lema.

Lema 10.14 *Si en un círculo se tienen dos tangentes* t_A *y* t_B *en los puntos* A *y* B *y se dan dos puntos* A' *y* B' *en* t_A *y* t_B *respectivamente tales que* $AA' = BB'$*, entonces existe un círculo tangente a* t_A *y* t_B *en los puntos* A' *y* B'*.*

Demostración: Si $O = t_A \cap t_B$, entonces la homotecia con centro O que lleva A en A', lleva B en B'. Y por lo tanto lleva el círculos en el círculo buscado. ■

Teorema 10.15 (Brianchon) *Si* $ABCDEF$ *es un exágono circunscrito a un círculo* $\mathcal{C}$*, entonces las rectas* AD*,* BE *y* CF *que unen los vértices opuestos son concurrentes.*

Demostración: Sean P, Q, R, S, T y U los puntos de tangencia de los lados AB, BC, CD, DE, EF y FA respectivamente. Tomemos puntos P_B, Q_B, R_D, S_D, U_F y T_F en los lados AB, BC, CD, DE, EF y FA de tal forma que el punto X_Y esta en la tangente el el punto X y del mismo lado que el punto Y, además

$$PP_B = QQ_B = RR_D = SS_D = TT_F = UU_F$$

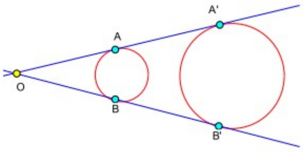

Figura 10-8

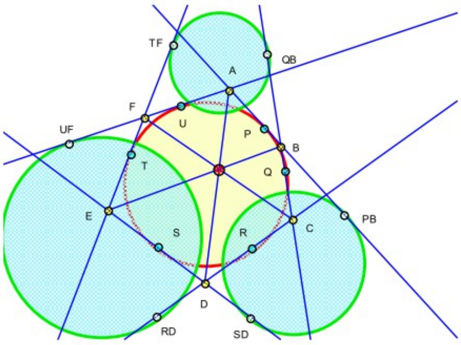

Figura 10-9

por el lema anterior tenemos circulos $\mathcal{C}_1$, $\mathcal{C}_2$ y $\mathcal{C}_3$ tangentes a los lados del exágono en los puntos T_F y Q_B; P_B, S_D; R_D y U_F; respectivamente. Entonces

$$\begin{aligned} BQ_B &= QQ_B - BQ = PP_B - PB = BP_B \\ ET_F &= TT_F + TE = SS_D + SE = ES_D \end{aligned}$$

por lo tanto B y E tienen la misma potencia respecto a los círculos $\mathcal{C}_2$ y $\mathcal{C}_3$ y por lo tanto BE es su eje radical. Análogamente AB y CF son los ejes radical de los círculos $\mathcal{C}_1$ y $\mathcal{C}_2$ y $\mathcal{C}_1$, $\mathcal{C}_3$ respectivamente. Por lo tanto estos tres ejes radicales son concurrentes. ■

Pág. 69(82) - Chapter 5
Curvature as a length derivative.
Pág. 78(91) - 5.4
Evaluating the Double Limit
Figure 11 Calculating geodesic curvature.

- 20210921/ Desargues/
Hexagon Theorem of Pappus...pdf
→ Theorems of Desargues and Pappus.

- 20210921/ Desargues/
Geometría Moderna...pdf
Pág. 235(250) - Capítulo 10
Teorema de Pappus, Pascal y Brianchon.
Teorema 10.1 (Carnot para rectas).

- 20210921/ Desargues/
ch1.pdf → Pappus's Theorem:

9)*+,)))- !) % 3#
-4 !33 E 6 C 3 ! # 8! !

)*+,)))- !) % 3#
)*+,)))- !) % 3#
)*+,)))- !) % 1 3#
)*+,)))- !) % % 3#
)*+,)))- !) % / 3#
)*+,)))- !) % " 3#
)*+,)))- !) % 3#
)*+,)))- !) % 3#
)*+,)))- !) % (3#
)*+,)))- !) % 3#
)*+,)))- !) % 3#
)*+,)))- !) % 3#
)*+,)))- !) % 1 3#

' "

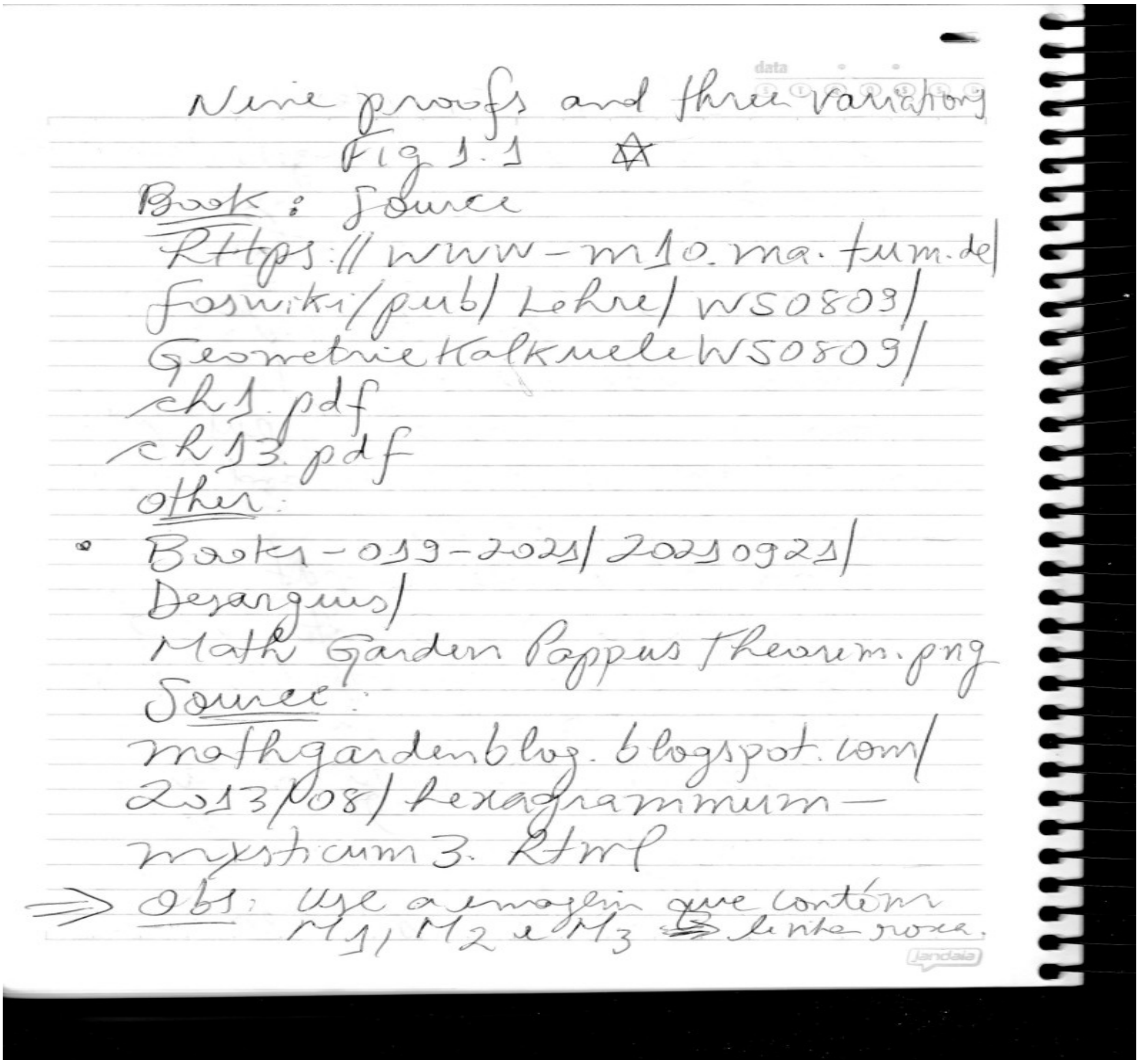

Nine proofs and three variations
Fig 1.1 ✡
Book: Source
https://www-m10.ma.tum.de/
foswiki/pub/Lehre/WS0809/
GeometrieKalkuele WS0809/
ch1.pdf
ch13.pdf
other:
• Books-019-2021/20210921/
Desargues/
Math Garden Pappus Theorem.png
Source:
mathgardenblog.blogspot.com/
2013/08/hexagrammum-
mysticum3.html
⇒ Obs: use a imagem que contém
M1, M2 e M3 ⇒ linha roxa.

%1 '

As imagens estão no diretório:
20210921/Desargues/images/

20210922/
Books-019-2021/20210922/
Machine Design with Cad...|
— Machine Design with Cad....pdf
Pag. 101(133) 4.1.2

- Homogeneous Coordinates Versus Cartesian Coordinates

Pag. 205(237) 5.5.4.3
Gradient Projection Method
Figure 5.24
Projecting the gradient vector ∇f onto the constraint tangent to get ~~—~~ the gradient projection ∇f_p.

obs: livro todo!! olhar!!!

— 20210922/ Krishna's Text on Calculus/

B0967YD4TJ.pdf
(T)

!)*+,)))- !) %1 ! 7 *7 ! #
-3 @! 3#
' 11 %

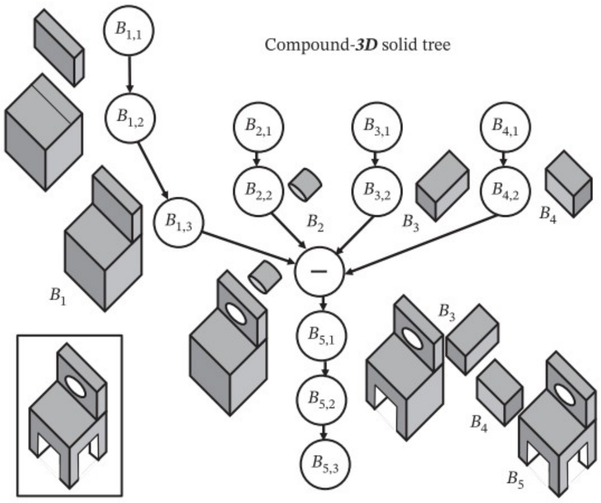

Figure 4.9 Constructing compound ***3D*** solid generated from several developed primitive solids.

operations is utilizing solids $B_{1,1}$ and $B_{1,2}$ to produce a single solid body B_1 through the $B_{1,3}$ *union* process. Similarly solid bodies B_2, B_3, and B_4 are created. Solid B_5 is then generated from solids B_1, B_2, B_3, and B_4 using a *difference Boolean* operation. After the *Boolean* operation, several edges of solid B_5 may be changed to fillet surfaces using a rounding operation (not shown in Figure 4.9). Other procedures are possible to produce a compound solid object. These procedures depend on the proficiency of the user in building ***3D*** models in different software packages. The chair-like object in Figure 4.9 can be produced in fewer steps such as a seat, a back, and four legs. The seat and the back can also be a one free-form surface with a certain thickness.

Trees representing the solid design process are usually displayed on a side screen to enable the user to document and edit the process. A single solid tree is representing the design process of a single solid (Figure 4.8). A compound solid tree is representing the relationships of solid bodies, which are combined using *Boolean* or other operations; see Figure 4.9. Using such trees, a designer can enter undo or redo commands to regenerate a previous solid if it is allowed. To generate or modify a solid, the user gives commands to the geometric modeling module. The *tree processor* stores the operations of all trees in some geometric modeling packages.

4.1.2 Homogeneous Coordinates Versus Cartesian Coordinates

Homogeneous coordinates have been devised to provide means to solve several problems such as solving sets of simultaneous *homogeneous* equations. They are also described or defined as *affine transformation*. Figure 4.10 defines the homogeneous coordinates in the two-dimensional (***2D***) Cartesian space. The homogeneous coordinates introduce an extra coordinate h_H. In that, the *homogeneous* coordinate of a *Cartesian* point of (x, y) is defined as (x, y, h_H) with the extra h_H added to the definition. However, we usually *normalize* the coordinates to have all coordinates lie on the homogenous plane of $(x, y, 1)$. All points (x, y, h_H) in the homogeneous coordinates will map into $(x/h_H, y/h_H, 1)$ as apparent in Figure 4.10. Homogenization operation in R^3 affirms that point (x, y, h_H) is homogenized to the point $(x/h_H, y/h_H, 1)$. In ***3D***, the *normalized homogeneous coordinate* $(x, y, z, 1)$ represents the *Cartesian* point (x, y, z). The homogeneous coordinates of (x, y, z, h_H) represent the *Cartesian* coordinates of $(x/h_H, y/h_H, z/h_H)$. Homogeneous coordinates of $(h_H x, h_H y, h_H z, h_H)$ should represent the *Cartesian* coordinates of

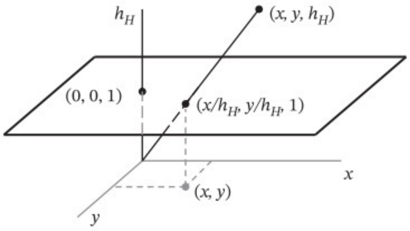

Figure 4.10 Definition of the homogeneous coordinates in the ***2D*** *Cartesian* space.

(x, y, z). In homogeneous coordinates, each (unique) *Cartesian* coordinate represents an unlimited or (∞) numbers of *homogeneous* coordinates, i.e. homogeneous coordinates are redundant. The value of h_H is an arbitrary constant.

4.1.2.1 Point in Space

In *Cartesian* coordinates a point is defined by its Cartesian coordinate values $\boldsymbol{P}_{C1}$ (x_1, y_1, z_1) and can also take the form of a *position vector* from the origin to the point.

$$\boldsymbol{P}_{C1} = [x_1 \ \ y_1 \ \ z_1]^T \tag{4.3}$$

The vectors are usually defined as *column arrays* but may also be defined as *row arrays*.

In *homogeneous* coordinates,

$$\boldsymbol{P}_{H1} = \begin{bmatrix} x_1 & y_1 & z_1 & 1 \end{bmatrix} \tag{4.4}$$

In this text, it was observed that points in homogeneous coordinates are better defined as *rows*. This allows the definition of body geometry by its main characteristic points, vertices, or corners as a set in a column array of these distinctive points. It can also be assumed that one might use $\boldsymbol{P}_1 = \boldsymbol{P}_{H1}$.

In *geometric modeling*, points $\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots, \boldsymbol{P}_n$ may also be referred to as vertices $\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_n$.

4.1.2.2 Vectors

In *Cartesian* coordinates a vector $\boldsymbol{V}_{C1}$ is defined by its coordinate components. It is usually set in a *column* array form.

$$\boldsymbol{V}_{C1} = \begin{bmatrix} \boldsymbol{V}_{1x} & \boldsymbol{V}_{1y} & \boldsymbol{V}_{1z} \end{bmatrix}^T \tag{4.5}$$

In *homogeneous* coordinates, a vector can be defined as the subtraction of two position arrays.

$$\boldsymbol{V}_{H1} = \begin{bmatrix} \boldsymbol{V}_{1x} & \boldsymbol{V}_{1y} & \boldsymbol{V}_{1z} & 0 \end{bmatrix}^T \tag{4.6}$$

It should be noted that the *0* in the fourth location should be carefully handled. The components of the vector are given the names $\boldsymbol{V}_{1x}$, $\boldsymbol{V}_{1y}$, and $\boldsymbol{V}_{1z}$ rather than x_1, y_1, and z_1 for location of point "*1*."

4.1.2.3 Lines

In homogeneous coordinates, a line $\boldsymbol{L}_H$ between two points $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ is given by

$$\boldsymbol{L}_H = \begin{bmatrix} \boldsymbol{P}_1 \\ \boldsymbol{P}_2 \end{bmatrix} = \begin{bmatrix} x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \end{bmatrix}, (2 \times j),\ j = 4 \tag{4.7}$$

The number of rows is 2, which is the number of end points $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ defining the line. The number of the columns *j* is the number of homogeneous coordinates, which is 4. Note that $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are used to represent $\boldsymbol{P}_{H1}$ and $\boldsymbol{P}_{H2}$.

4.1.2.4 Body Geometry and Vertices

Define the geometry of the body $\boldsymbol{B}_H$ in homogeneous coordinates by the matrix $[\boldsymbol{P}_B]$ representing the object *vertices* or *points* $\boldsymbol{P}_i$, each of which is a row $(x_i, y_i, z_i, 1)$ in this matrix. The body $\boldsymbol{B}_H$ can then have the following definition:

$$\boldsymbol{B}_H = \begin{bmatrix} \boldsymbol{P}_1 \\ \boldsymbol{P}_2 \\ \cdots \\ \boldsymbol{P}_n \end{bmatrix} = \begin{bmatrix} x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ \cdots & \cdots & \cdots & \cdots \\ x_n & y_n & z_n & 1 \end{bmatrix}, (i \times j), \quad i = n, \quad j = 4 \tag{4.8}$$

Note that $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_n$ are used to represent $\boldsymbol{P}_{H1}, \ldots, \boldsymbol{P}_{Hn}$.

4.1.3 Body Transformation

The body $\boldsymbol{B}_H$ can be transformed in the homogeneous space by *translation, rotation, scaling, zooming, skewing*, or *perspective viewing*. A simple original ***2D*** shape of a body in a plane is shown in Figure 4.11a, and each of its particular transformation is indicated as in Figure 4.11b for translation, in Figure 4.11c for scaling, in Figure 4.11d for rotation, in Figure 4.11e for skewing, and in Figure 4.11f for perspective. The transformed body matrix $\boldsymbol{B}_H{}'$ is obtained collectively or by any one of the transformations in any other order by the following matrix multiplication:

$$\begin{aligned} \boldsymbol{B}_H{}' &= \boldsymbol{B}_H \ \ [\boldsymbol{T}_{xyz}][\boldsymbol{R}_{xyz}][\boldsymbol{S}_H][\boldsymbol{Z}_H][\boldsymbol{K}_H] \ldots [\boldsymbol{P}_H] \\ &= \boldsymbol{B}_H[\boldsymbol{T}_{xyz}\, \boldsymbol{R}_{xyz}\, \boldsymbol{S}_H\, \boldsymbol{Z}_H\, \boldsymbol{K}_H \ldots \boldsymbol{P}_H] \\ &= \boldsymbol{B}_H \ \ \boldsymbol{T}_{\mathrm{GH}} \end{aligned} \tag{4.9}$$

where $\boldsymbol{T}_{\mathrm{xyz}}$ is the translation matrix; $\boldsymbol{R}_{xyz}$ is the rotation matrix about x, y, or z; $\boldsymbol{S}_H$ is the scaling matrix; $\boldsymbol{Z}_H$ is the zooming matrix; $\boldsymbol{K}_H$ is the skewing matrix; $\boldsymbol{P}_H$ is the perspective matrix; and $\boldsymbol{T}_{\mathrm{GH}}$ is the general total homogeneous

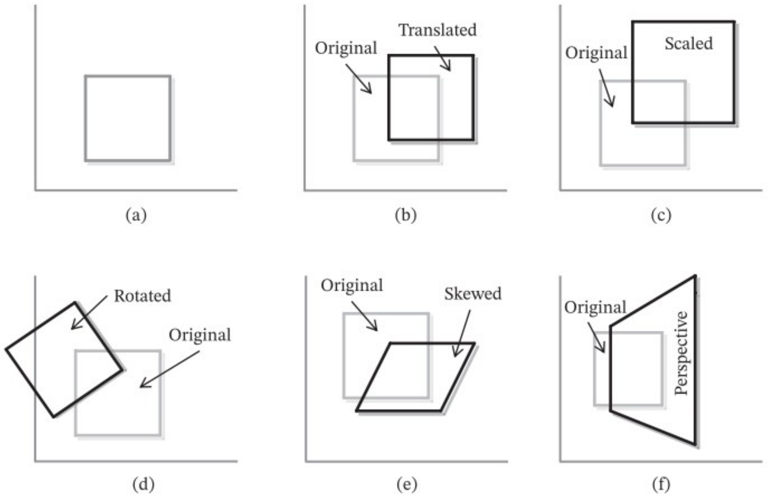

(a) (b) (c)

(d) (e) (f)

Figure 4.11 A simple original shape of a body (a) in a plane and each of its transformation is indicated in (b) translation, (c) scaling, (d) rotation, (e) skewing, and (f) perspective.

transformation matrix. The order of multiplication is important in defining the intended "motion" or transformation of the body in space from its original position to its final intended location and defined shape. Note that matrix multiplication is *associative*, but not *commutative*. The matrices will be defined later on, but their final multiplication form represents the general transformation matrix, which is post-multiplied by the body matrix $\boldsymbol{B}_H$ to get the shape of the transformed body matrix $\boldsymbol{B}_H{}'$. The general homogeneous transformation matrix $\boldsymbol{T}_{\mathrm{GH}}$ can then have the following form:

$$\boldsymbol{T}_{\mathrm{GH}} = [\boldsymbol{T}_{\mathrm{GH}}] = \begin{bmatrix} a & k & n & h \\ d & b & p & g \\ e & m & c & f \\ t & u & v & z \end{bmatrix} \tag{4.10}$$

The components of the matrix are representing different transformations. Components a, b, and c are the *scaling* amounts in x, y, and z, respectively. The components t, u, and v are the *translation* amounts in x, y, and z, respectively. The components h, g, and f are the *perspective* amounts in x, y, and z, respectively. The components d, e, k, m, n, and p are the *skewing* amounts in x, y, and z. The component z is the *inverse zooming* amount in all x, y, and z. Each of the respective transformation matrixes is defined separately as follows:

4.1.3.1 Translation

The *translation matrix* can have the form

$$\boldsymbol{T}_{xyx} = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ t_x & t_y & t_z & 1 \end{bmatrix} \tag{4.11}$$

The components t_x, t_y, and t_z are the translation amounts in x, y, and z, respectively. All empty locations in the $\boldsymbol{T}_T$ matrix are zeros.

4.1.3.2 Rotation

The *rotation matrix* about z, which can be checked with that defined in mechanics subject, has the following form:

$$\boldsymbol{R}_z = \begin{bmatrix} \cos(\theta_z) & \sin(\theta_z) & 0 & 0 \\ -\sin(\theta_z) & \cos(\theta_z) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{4.12}$$

where θ_z is the rotation angle about z.

The *rotation matrix* about y has the form

$$\boldsymbol{R}_y = \begin{bmatrix} \cos(\theta_y) & 0 & -\sin(\theta_y) & 0 \\ 0 & 1 & 0 & 0 \\ \sin(\theta_y) & 0 & \cos(\theta_y) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{4.13}$$

where θ_y is the rotation angle about y.

The *rotation matrix* about x has the form

$$\boldsymbol{R}_x = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \cos(\theta_x) & \sin(\theta_x) & 0 \\ 0 & -\sin(\theta_x) & \cos(\theta_x) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{4.14}$$

where θ_x is the rotation angle about x.

4.1.3.3 Scaling

The *scaling matrix* is given by the following form:

$$\boldsymbol{S}_H = \begin{bmatrix} S_x & 0 & 0 & 0 \\ 0 & S_y & 0 & 0 \\ 0 & 0 & S_z & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{4.15}$$

The components S_x, S_y, S_z are the scales in x, y, and z, respectively.

4.1.3.4 Zooming

The *zooming matrix* is given by the following form:

$$\boldsymbol{Z}_H = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1/z_f \end{bmatrix} \tag{4.16}$$

where z_f is the zooming factor and all empty locations are zeros.

4.1.3.5 Skewing

The *skewing matrix* is given by the following form:

$$\boldsymbol{K}_H = \begin{bmatrix} 1 & k & n & \\ d & 1 & p & \\ e & m & 1 & \\ & & & 1 \end{bmatrix} \tag{4.17}$$

where d, e, k, m, n, and p are the skewing amounts in x, y, and z, where all empty locations in the matrix $\boldsymbol{K}_H$ are zeros.

4.1.3.6 Perspective

The *perspective matrix* is defined on the condition that the eye of the observer is on the z-coordinate axis and at a distance of f_z from the body.

$$\boldsymbol{P}_H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1/f_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{4.18}$$

where f_z is the distance from the eye to the projection plane when the *eye* is on the z-axis as previously required by this procedure. Usually the projection plane is set to be the x–y plane (or z plane). The distance f_z may be called the *focal distance* along the z-axis. This is usually performed after rotating the body and the eye to set the looking *eye* on the z direction at a f_z distance off the body location.

4.1.3.7 Orthographic Projection

After all transformations and considering the z-axis to be normal to the plane of the screen (the +ve behind the screen), the simplest method of projection is to ignore the z-coordinate of every point in space. This would render the ***3D*** points (x, y, z) to be only ***2D*** points (x, y).

4.1.3.8 Body Transformation Systems

Note that since matrix multiplication is *associative*, but not *commutative*, the order of multiplication is important.

If the body matrix $\boldsymbol{B_C}$ representing the object vertices or points each of which as a ***column*** vector $[x_i \quad y_i \quad z_i \quad 1]^T$ in this matrix, the transformed body matrix $\boldsymbol{B}_C{}'$ is obtained by any one of the transformations or collectively in any other proper order by

$$\begin{aligned} \boldsymbol{B}_C' &= [\boldsymbol{T}^T_{xyz}][\boldsymbol{R}^T_{xyz}][\boldsymbol{S}^T_H][\boldsymbol{Z}^T_H][\boldsymbol{K}^T_H]\cdots[\boldsymbol{P}^T_H]\ \boldsymbol{B}_C \\ &= \boldsymbol{T}^T_{\mathrm{GH}}\ \boldsymbol{B}_C \end{aligned} \tag{4.19}$$

This process is used by many computer graphics texts and by most *OpenGL* systems. For the computer programming, however, it is usually easier to represent and save a body n points or vertices in an $(n \times 4)$ matrix instead of the $(4 \times n)$ matrix. Several graphics texts are also using the formulation procedure adopted herein rather than the usual *OpenGL* Systems. Note that using the transpose will cause the multiplication to be in a *post-multiplication* rather a *pre-multiplication* way. In providing *OpenGL* systems with vertices, this is not an issue since the order of the internal calculations is not observed or be concerned with.

Example 4.1 A line is defined by its two ***3D*** end points $\boldsymbol{P}_1$ (1, 0, 0) and $\boldsymbol{P}_2$ (2, 2, 0) in *Cartesian* coordinates. It is required to translate the line 2 units in the x direction and 1 unit in the y direction simultaneously. After translation, it is required to rotate the line by −10° about the z-axis.

Solution

The line $\boldsymbol{L}$ can be identified in *homogeneous* coordinate by applying Eq. (4.7) such that

$$\boldsymbol{L} = \boldsymbol{L}_H = \begin{bmatrix} \boldsymbol{P}_1 \\ \boldsymbol{P}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 2 & 2 & 0 & 1 \end{bmatrix} \tag{a}$$

To translate in x and y by 2 and 1, respectively, the transformation matrix for translation is

$$\boldsymbol{T}_{xyz} = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ t_x & t_y & t_z & 1 \end{bmatrix} = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ 2 & 1 & 0 & 1 \end{bmatrix} \tag{b}$$

Considering the line to be our ***3D*** body, the new translated line $\boldsymbol{L}_1$ is obtained by applying Eq. (4.9) to get the following:

$$\boldsymbol{L}_1 = \boldsymbol{L}_H\ \boldsymbol{T}_{xyz} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 2 & 2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ 2 & 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 1 & 0 & 1 \\ 4 & 3 & 0 & 1 \end{bmatrix} \tag{c}$$

The solution is simply the multiplication of the two matrices. Figure 4.12 shows the original line $\boldsymbol{L}$ and the translated line $\boldsymbol{L}_1$.

The following rotation is −10° about the z-axis with the rotation matrix defined by Eq. (4.11):

$$\boldsymbol{R}_z = \begin{bmatrix} \cos(\theta_z) & \sin(\theta_z) & 0 & 0 \\ -\sin(\theta_z) & \cos(\theta_z) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \cos(-10) & \sin(-10) & 0 & 0 \\ -\sin(-10) & \cos(-10) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0.9848 & -0.1736 & 0 & 0 \\ -(-0.1736) & 0.9848 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{d}$$

$$\boldsymbol{L}_2 = \boldsymbol{L}_1\ \boldsymbol{R}_z = \begin{bmatrix} 3 & 1 & 0 & 1 \\ 4 & 3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0.985 & -0.174 & 0 & 0 \\ 0.174 & 0.985 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 3.129 & 0.463 & 0 & 1 \\ 4.462 & 2.259 & 0 & 1 \end{bmatrix} \tag{e}$$

Figure 4.12 shows the original line $\boldsymbol{L}$, the translated line $\boldsymbol{L}_1$, and the rotated line $\boldsymbol{L}_2$.

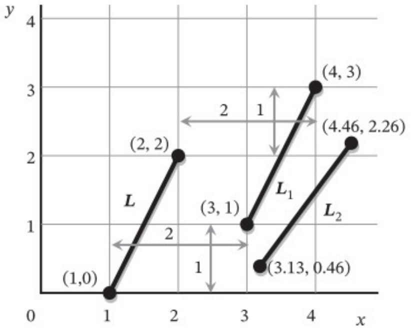

Figure 4.12 A line $\boldsymbol{L}$ defined by its two end points $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ in *Cartesian* coordinates. It is translated to line $\boldsymbol{L}_1$ by *2* units in the x direction and *1* unit in the y direction. It is then rotated to line $\boldsymbol{L}_2$ by −10°.

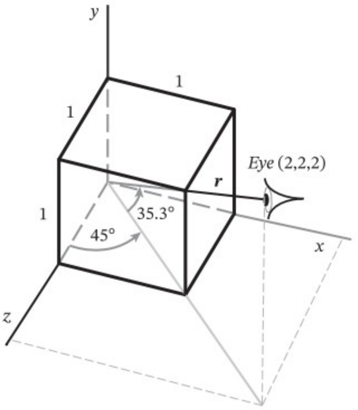

Figure 4.13 A unit cube body attached to the origin of the coordinate system *xyz*. The observer eye is located at the position (2, 2, 2) and is looking along the cube diagonal toward the origin.

Example 4.2 A unit cube body has its bottom corner attached to the origin of the coordinate system *xyz* as shown in Figure 4.13. The observer eye is located at the position (2, 2, 2) and is looking along the cube diagonal toward the origin. It is required to find the projection of the cube on the normal plane to the line of sight and passing by the origin. Does this projection produce the perceived ***perspective*** of the cube by the observer?

Solution

The body matrix $\boldsymbol{B}_H$ is defined by the following:

$$\boldsymbol{B}_H = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}^T \tag{a}$$

Note that the cube corners, vertices, or points $\boldsymbol{P}_i$ are considered as row vectors. The *eye (2, 2, 2)*, or in general location vector $\boldsymbol{r}$ *(a_E, b_E, c_E)*, makes the following angles to transform the *eye* to be located on the z-axis in preparation for applying the perspective transformation. The radial distance r_{eye} from the origin to the eye is then

$$r_{\text{eye}} = (a_E{}^2 + b_E{}^2 + c_E{}^2)^{1/2} \quad \text{Or} \quad r_{2,2,2} = (12)^{1/2} = 3.46 = f_z \tag{b}$$

The angle θ_y between the projected $\boldsymbol{r}$ vector on the x–z plane and the y–z plane is

$$\theta_y = \tan^{-1} a_E/c_E = \tan^{-1} 1 = 45^\circ \tag{c}$$

Then the *eye* and the body should be rotated in the opposite direction, i.e. rotated by $-\theta_y$ so that the *eye* would be located in the y–z plane. Therefore, the rotation should be $\theta_y = -45^\circ$.

The angle between the *eye* – while on the y–z plane after θ_y rotation – and the x–z plane is given by

$$\beta_x = -\sin^{-1}\left(b_E/(a_E{}^2 + b_E{}^2 + c_E{}^2)^{1/2}\right) = -\sin^{-1}(2/3.46) = -35.3^\circ \tag{d}$$

Here, the *eye* and the body should be rotated in the opposite direction, i.e. by $-\beta_x$ so that the eye would be located on the z-axis. Therefore, the rotation should be $\beta_x = 35.3^\circ$.

For the rotation matrices, one calculates $\cos\theta_y = 0.707$, $\sin\theta_y = -0.707$, $\cos\beta_x = 0.816$, and $\sin\beta_x = 0.578$. The rotation matrices are then

$$\boldsymbol{R}_y = \begin{bmatrix} \cos(\theta_y) & 0 & -\sin(\theta_y) & 0 \\ 0 & 1 & 0 & 0 \\ \sin(\theta_y) & 0 & \cos(\theta_y) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0.707 & 0 & 0.707 & 0 \\ 0 & 1 & 0 & 0 \\ -0.707 & 0 & 0.707 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{e}$$

$$\boldsymbol{R}_x = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \cos(\beta_x) & \sin(\beta_x) & 0 \\ 0 & -\sin(\beta_x) & \cos(\beta_x) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0.816 & 0.578 & 0 \\ 0 & -0.578 & 0.816 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{f}$$

The order of rotation is important and should be $\boldsymbol{R}_y$ then $\boldsymbol{R}_x$, i.e.,

$$\boldsymbol{R}_{xyz} = \boldsymbol{R}_y \boldsymbol{R}_x = \begin{bmatrix} 0.707 & 0 & 0.707 & 0 \\ 0 & 1 & 0 & 0 \\ -0.707 & 0 & 0.707 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0.816 & 0.578 & 0 \\ 0 & -0.578 & 0.816 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0.707 & -0.409 & 0.577 & 0 \\ 0 & 0.816 & 0.578 & 0 \\ -0.707 & -0.409 & 0.577 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{g}$$

Transformation of rotation then *perspective* $\boldsymbol{P}_H$, where $\boldsymbol{P}_H$ is given by

$$\boldsymbol{P}_H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1/3.464 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -0.289 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{h}$$

The total transformation matrix $\boldsymbol{T}_{\text{GH}}$ is then (after multiplication)

$$\boldsymbol{T}_{\text{GH}} = \boldsymbol{R}_{xyz}\ \boldsymbol{P}_H = \begin{bmatrix} 0.707 & -0.409 & 0.577 & 0 \\ 0 & 0.816 & 0.578 & 0 \\ -0.707 & -0.409 & 0.577 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -0.289 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0.707 & -0.409 & 0.577 & -0.167 \\ 0 & 0.816 & 0.578 & -0.167 \\ -0.707 & -0.409 & 0.577 & -0.167 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{i}$$

Note that the first three columns are exactly the same as the total rotation matrix $\boldsymbol{R}_{xyz}$.

The body after total transformation is then given by (after multiplication)

$$\boldsymbol{B}'_H = \boldsymbol{B}_H \;\; \boldsymbol{T}_{\mathrm{GH}} = \begin{bmatrix} 0 & 0 & 0 & 1.0000 \\ 0.7071 & -0.4082 & 0.5774 & 0.8333 \\ 0.7071 & 0.4082 & 1.547 & 0.6667 \\ 0 & 0.8165 & 0.5774 & 0.8333 \\ -0.7071 & -0.4082 & 0.5774 & 0.8333 \\ 0.0000 & -0.8165 & 1.1547 & 0.6667 \\ 0.0000 & 0.0000 & 1.7321 & 0.5000 \\ -0.7071 & 0.4082 & 1.1547 & 0.6667 \end{bmatrix} \tag{j}$$

To project the body, a *homogeneous normalization* is necessary. It means that the fourth column must have the value of one for each point or row. This can be implemented by dividing each component of a row by the fourth column value of that row, which normalizes the fourth component to the value of *1*. Performing the process for all the rows causes the fourth column to a unity. After *homogeneous normalization*, the perspective projection body is then

$$\boldsymbol{B}'_H = \begin{bmatrix} 0 & 0 & 0 & 1.0000 \\ 0.8486 & -0.4899 & 0.6929 & 1.0000 \\ 1.0606 & 0.6123 & 1.7320 & 1.0000 \\ 0 & 0.9798 & 0.6929 & 1.0000 \\ -0.8486 & -0.4899 & 0.6929 & 1.0000 \\ 0 & -1.2247 & 1.7320 & 1.0000 \\ 0 & 0 & 3.4642 & 1.0000 \\ -1.0606 & 0.6123 & 1.7320 & 1.0000 \end{bmatrix} \tag{k}$$

The plotting of the body on x–y plane is shown in Figure 4.14. This projection produces the perceived *perspective* of the cube by the observer. If all lines are extended, they will meet at the three *vanishing points* known in regular graphics or drawing subjects.

4.1.4 Stereo Viewing

To produce a stereo view for a body, we need to generate two views of the body, one for each eye of the viewer. The two views are looked at simultaneously by each eye to generate the ***3D*** stereo of the body. The right eye should only see the right body view, and the left eye should only see the left view of the body. This is similar to the ***3D*** movies where one would use eyeglasses with right and left piece of different polarization planes. Usually one piece is of a vertical polarization, and the other is of a horizontal polarization. The movie is produced with two cameras for the right and left eyes and projects both simultaneously through the two polarized planes onto the theater screen. For LCD or LED screens, the process is different so far. It requires that the eyeglasses should synchronize with the projection of the right and left views; see also Section 3.4 about virtual reality (VR).

Figure 4.14 Plotting of the body onto the projection x–y plane. The projection produces the perceived perspective of the cube by the observer.

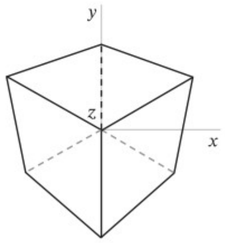

It should be of interest to note that by comparing Eq. (5.139) to Eq. (5.143), one can get the *inverse* of the *non-square matrix* as follows:

$$\boldsymbol{G}^{-1} = [\boldsymbol{G}^T\boldsymbol{G}]^{-1}\boldsymbol{G}^T \tag{5.144}$$

This is called the ***pseudo inverse*** or *minimum norm inverse*. MATLAB provide a *pseudo inverse function* "`pinv`" to carry such an inverse.

5.5.4.3 Gradient Projection Method

The method of ***GP*** is concerned about moving along the constraint boundary. If we hit the constraints, we may like to move on the constraints by projecting the gradient vector onto the constraints tangent as shown in Figure 5.24. The projected gradient vector consists of the gradient vector itself minus the component u_i normal to the tangent (i.e. parallel to $\nabla\mathbf{g}$). This is translated to the following relation for r constraints *in effect* at the point in question:

$$\nabla f_p = \nabla f - u_1\nabla g_1 - u_2\nabla g_2 \ldots\ \ u_r\nabla g_r \tag{5.145}$$

Note again that only few "r" of the "m" constraints are active. Equation (5.145) can be rewritten as

$$\nabla\boldsymbol{f}_p = \nabla\boldsymbol{f} - \begin{bmatrix}\nabla\mathbf{g}_1 & \nabla\mathbf{g}_2 & \ldots. & \nabla\mathbf{g}_r\end{bmatrix}\begin{bmatrix}u_1\\ u_2\\ \ldots\\ u_r\end{bmatrix} = \nabla\boldsymbol{f} - \boldsymbol{G}_r\boldsymbol{u} \tag{5.146}$$

It should be noted that, generally, $\boldsymbol{G}_r \neq \boldsymbol{G}$. But the rejection vector must be *orthogonal* to the constraints gradient, which requires that

$$\nabla\mathbf{g}_1^T\nabla\boldsymbol{f}_p = \nabla\mathbf{g}_2^T\nabla\boldsymbol{f}_p = \cdots = \nabla\mathbf{g}_r^T\nabla\boldsymbol{f}_p = 0 \tag{5.147}$$

This is rewritten as

$$\boldsymbol{G}_r^T\nabla\boldsymbol{f}_p = \mathbf{0} \tag{5.148}$$

Substituting Eq. (5.148) into Eq. (5.145) by the through multiplication by G_r^T, one obtains the following:

$$\begin{aligned}&\boldsymbol{G}_r^T[\nabla\boldsymbol{f} - \boldsymbol{G}_r\boldsymbol{u}] = \mathbf{0}\\ &\boldsymbol{G}_r^T\nabla\boldsymbol{f} = \boldsymbol{G}_r^T\boldsymbol{G}_r\boldsymbol{u}\end{aligned} \tag{5.149}$$

Or

$$\boldsymbol{u} = [\boldsymbol{G}_r^T\boldsymbol{G}_r]^{-1}\boldsymbol{G}_r^T\nabla\boldsymbol{f} \tag{5.150}$$

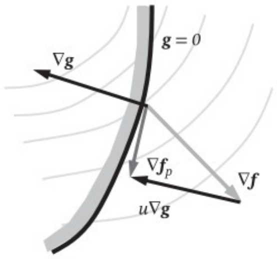

Figure 5.24 Projecting the gradient vector $\nabla\boldsymbol{f}$ onto the constraints tangent to get the gradient projection $\nabla\boldsymbol{f}_p$.

This relation is in a similar form to the *Lagrange multipliers* in Eq. (5.143), but without the negative sign. Substituting Eq. (5.150) into Eq. (5.145) gives the projection vector.

$$\nabla \boldsymbol{f}_p = \nabla \boldsymbol{f} - \boldsymbol{G}_r\,[\boldsymbol{G}_r^T \boldsymbol{G}_r]^{-1} \boldsymbol{G}_r^T \nabla \boldsymbol{f}$$

$$\nabla \boldsymbol{f}_p = \{I - \boldsymbol{G}_r\,[\boldsymbol{G}_r^T \boldsymbol{G}_r]^{-1} \boldsymbol{G}_r^T \nabla \boldsymbol{f}\} \nabla \boldsymbol{f} \tag{5.151}$$

Therefore, for minimization we can proceed in the $-\nabla \boldsymbol{f}_p$ direction. In that situation, what might happen? Three cases are expected to occur; see Figure 5.25 for a hypothetical constraint that has *convex*, *linear*, and *concave* sections:

- At point (a) on the *convex constraint* section, obviously we go into the infeasible region. If we want to continue, we have to step in the $-\nabla \boldsymbol{g}$ direction to converge. Thus, method is not appropriate for convex constraints.
- For *concave constraint* section at point (b), we move into the feasible region, and we can continue in the normal search procedure.
- For *linear constraint* section at (c), we move in just the right direction. The method in most useful and suitable. This is one way of handling constraints in *linear programming*.

5.5.4.4 *Heuristic Gradient Projection Method (HGP)*

As indicated in the previous ***GP*** section and for convex constraints, the search direction should depart the constraint boundary into the infeasible region if these constraints were active. To continue, one must step into the $-\nabla \boldsymbol{g}_i$ to converge. If the component of $-\nabla \boldsymbol{g}_i$ is included in the new combined ***GP*** search direction $-\nabla \boldsymbol{g}_i{}^T \nabla \boldsymbol{f}$, the process should be automatic. Therefore, this method is suitable for convex constraints. For concave constraints, the search direction also steps into the feasible region. To continue, one may need to step into the $\nabla \boldsymbol{g}_i$ to converge. If the component of $\nabla \boldsymbol{g}_i$ is also included in the new combined search direction $\nabla \boldsymbol{g}_i{}^{\boldsymbol{T}} \nabla \boldsymbol{f}$, the process should also be automatic. Figure 5.26 shows the projection of the gradient vector ∇f onto the constraints tangent to get the *heuristic gradient projection* (***HGP***) ∇f_{pH}, which is away from the constraint. In many cases, however, the constraint gradient may be in the same direction as the gradient of the objective function. The search can thus continue in the normal procedure; see, e.g. Abd El Malek (2005) and Metwalli (2012).

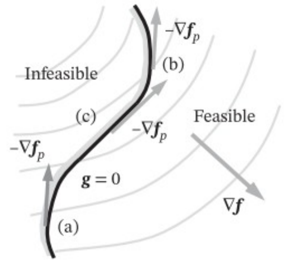
$-\nabla \boldsymbol{f}_p$
(b)
Infeasible
(c)
$-\nabla \boldsymbol{f}_p$
Feasible
$-\nabla \boldsymbol{f}_p$
$\boldsymbol{g} = 0$
$\nabla \boldsymbol{f}$
(a)

Figure 5.25 A hypothetical constraint having convex, linear, and concave sections.

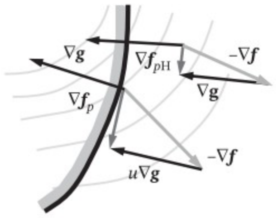

Figure 5.26 Projecting the gradient vector $\nabla \boldsymbol{f}$ onto the constraints tangent to get the heuristic gradient projection $\nabla \boldsymbol{f}_{\mathrm{pH}}$ away from the constraint.

The ***HGP*** method is most suitable to apply for problems that are expected to have a constrained minimum with defined formulation for both objective function and constraints. The optimization of ***3D*** space frames is a problem where the objective function is defined as the total volume and the constraints are restricted to satisfy allowable stress values. The optimization of ***3D*** space frames is then most suitable to demonstrate the ***HGP*** method. The objective function for ***3D*** space frames is the total volume of the n members of circular cross section each of a diameter d_i and a length l_i. The total volume is therefore the objective function defined as

$$f(d) = \sum_{i=1}^{n} \pi {d_i}^2 l_i/4 \tag{5.152}$$

Without any loss of generality, if the ***3D*** frame will not be *topologically optimized* except by removing trivially loaded members, the objective function will take the form

$$f(d) = \sum_{i=1}^{n} a_i {d_i}^2 \tag{5.153}$$

where a_i is a constant that can represent $\pi l_i/4$. In the local coordinates, the stress in each member is evaluated numerically (Abd El Malek et al. 2005). Without key loss of generality, the stress constraints may be initially defined by the following expression:

$$g_i(d) \cong b_i {d_i}^{-3} - c_i \leq 0 \tag{5.154}$$

where b_i is a constant that includes the *stress* constants and c_i is a constant representing the *allowable strength*. It is assumed initially that the coupling between members and its feedback can be automatically accounted for during iterations. If, however, the members were mainly subjected to axial loading, Eq. (5.154) would instead be mainly in (d^{-2}). This is one other alternative, which can be used. Other options have also been used; see, e.g. Senousy et al. (2005), Hanafy and Metwalli (2009), and Abd El-Aziz et al. (2016).

The components of the gradient vectors for the objective function and the constraints are obtained from Eqs. (5.153) and (5.154) as follows:

$$\begin{aligned} \nabla f(d) &= \alpha_i d_i \\ \nabla g_i(d) &= -\beta_i {d_i}^{-4} \end{aligned} \tag{5.155}$$

The components of *projected gradient vector* can be obtained as follows:

$$\nabla {g_i}^T \nabla f = -\gamma_i \, {d_i}^{-3} \tag{5.156}$$

To implement the procedure and considering the *slackness variables* along the ***HGP*** direction, we use the recursive heuristic direction components for constraints and minimize the objective function at the same time such that

$$s_{\mathrm{pi}}(d_i) = \sigma_{i,\mathrm{equiv}} \; (d_{i,\mathrm{new}})^{-3} - S_{i,\mathrm{const}} \; (d_i)^{-3} \tag{5.157}$$

where d_i is the *member diameter*, $d_{i,\mathrm{new}}$ is the *new diameter* of the member, s_{pi} is the component of slackness on the projection vector, $\sigma_{i,\mathrm{equiv}}$ is the *equivalent stress* (e.g. *von Mises*), and $S_{i,\mathrm{const}}$ is the *allowable strength* constraint.

This can be implemented by *recursive* use of the following relation:

$$d_{i,\mathrm{new}}^3 = d_i^3 \frac{\sigma_{i,\mathrm{equiv}}}{S_{i,\mathrm{const}}} \tag{5.158}$$

The relation defined in Eq. (5.158) satisfies the constraint; in addition, it reduces the objective function. Therefore, these steps chase the constraint boundary while reducing the objective function. This is tantamount to the process of ***GP*** method, but does not require following the constraint tangents. This is the reason of applying the name ***HGP*** on this method.

5.5.4.5 Constrained Optimization Sample

This section presents a simple constrained problem to clarify concepts and demonstrate constraint optimization handling.

Example 5.14 Consider the simple objective function defined by

$$\boldsymbol{f} = -(x_1 + x_2) \tag{a}$$

The constraints for this problem are

$$\begin{aligned} \mathbf{g}_1 &= -x_1 \leq 0 \\ \mathbf{g}_2 &= x_1^2 + x_2^2 - 1 \leq 0 \end{aligned} \tag{b}$$

Start at $\boldsymbol{X}_o = (0, -1/2)$, use the first direction of any first-order method (*gradient*) to reach the next point. Check if the point $x_1 = \sqrt{2}/2,\ x_2 = \sqrt{2}/2$ is a constraint optimum.

Solution

Data: $\boldsymbol{X}_0 = (0, -1/2)$ and check *K–T* conditions at (0.707, 0.707). The value of the objective point at $\boldsymbol{X}_0$ is $\boldsymbol{f}_0 = 1/2$.

The plot of the function contours and constraints is given in Figure 5.27 showing the infeasible and feasible regions. The objective function gradient is

$$\nabla f = \begin{bmatrix} \dfrac{\partial f}{\partial x_1} \\ \dfrac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \tag{c}$$

First search direction $\boldsymbol{S}$ of any first-order method is in the negative direction of the gradient, *i.e.* $-\nabla \boldsymbol{f}$. Get the intersection of $\boldsymbol{S}$ line with the constraint such as

$$\begin{aligned} x_2 &= x_1 - 0.5, \\ x_1^2 + (x_1 - 0.5)^2 - 1 &= 0 \end{aligned} \tag{d}$$

which gives a point at $x_1 = 0.9114$ and -0.4114. At $x_1 = 0.9114$, the next point is then $\boldsymbol{X}_1 = (0.9114, 0.4115)$. At $x_1 = -0.4114$, the point is in the infeasible region and thus will be discarded. The required next point $\boldsymbol{X}_1 = (0.9114, 0.4115)$ is then shown in Figure 5.13.

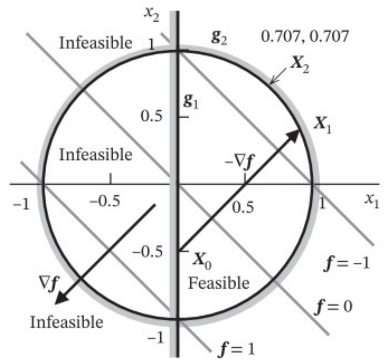

Figure 5.27 The plot of the function contours and constraints of Example 5.14 showing the infeasible and feasible regions.

Pág D-109(114) 7. Curvature

Pág D-179(185) 10. Singular Points:
Curve Tracing

Pág I-121(346) 6. Areas of Curves

Quadrature - The process of finding the area of any bounded portion of a curve is called quadrature.

Pág. I-145(370) 7. Rectification

The process of finding the length of an arc of a curve between two given points.

Obs: Livro todo!!!

_ 20210922/Krishna's Text...

Krishna's Complex Analysis...pdf

Pag C-3(8) 1. Complex Numbers
and their Geometrical
Representation.

Pag C-101(106) 3. Conformal
Mapping

Obs:
Pag C-102(107) → correspondence
is one to one.
The corresponding sets of points in the two planes are called images of each other.

' 7 (% !3 8!
' 7 (/ !3 ! 8 6 !
' & 1%" !3 " * ! 8
' & %/ 1 !3 + !
'

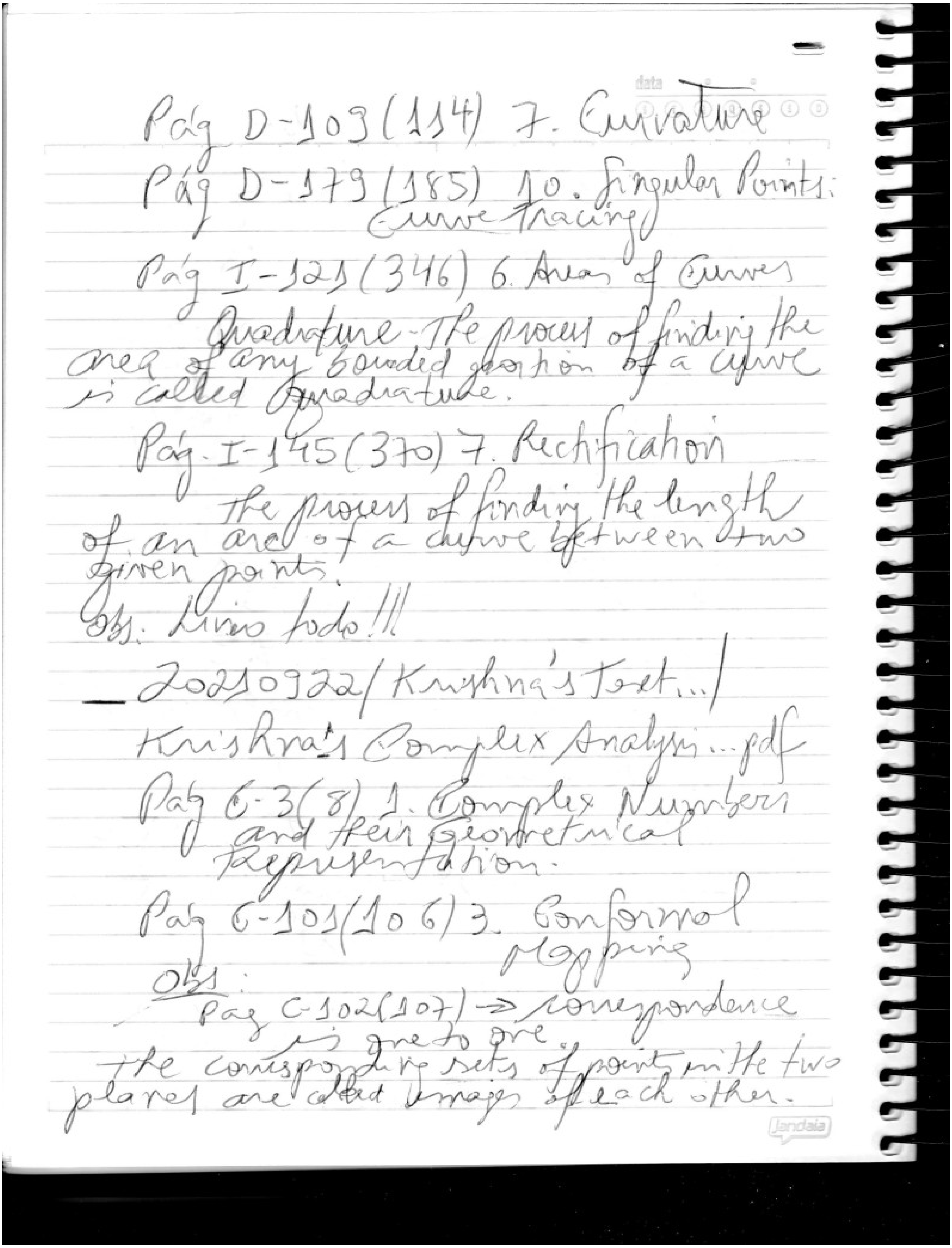

Pág D-103(114) 7. Curvature

Pág D-179(185) 10. Singular Points: Curve Tracing

Pág I-121(346) 6. Areas of Curves

Quadrature - The process of finding the area of any bounded portion of a curve is called quadrature.

Pág. I-145(370) 7. Rectification

The process of finding the length of an arc of a curve between two given points.

Obs: Livro todo!!!

— 20210922/Krishna's Text.../

Krishna's Complex Analysis...pdf

Pág C-3(8) 1. Complex Numbers and their Geometrical Representation.

Pág C-101(106) 3. Conformal Mapping

Obs:

Pág C-102(107) → correspondence is one to one.

The corresponding sets of points in the two planes are called images of each other.

Pág D-109(114) 7. Curvature

Pág D-179(185) 10. Singular Points: Curve Tracing

Pág I-121(346) 6. Areas of Curves

Quadrature - The process of finding the area of any bounded portion of a curve is called quadrature.

Pág. I-145(370) 7. Rectification

The process of finding the length of an arc of a curve between two given points.

Obs: Livro todo!!!

_ 20210922/Krishna's Text.../

Krishna's Complex Analysis...pdf

Pág C-3(8) 1. Complex Numbers and their Geometrical Representation.

Pág C-101(106) 3. Conformal Mapping

Obs:

Pág C-102(107) → correspondence is one to one.

The corresponding sets of points in the two planes are called images of each other.

data

_ 20210922/Krishna's.../
Krishnas_TB_Differential_Calculus.pdf
Pág. D-219(224) 11. Curvature (erros na grafia)
obs: Comparar com o outro capitulo do livro anterior (B0967Y04TJ.pdf).
↳ o texto está sem erros de grafia.

_ 20210922/The Mathematical Career of Pierre de Fermat.../
0691036667.pdf
Obs. Livro todo!!! Contents não está bem detalhado.

_ 20210922/Landscape of 21 st.../
303080626X.pdf
Pág. 207(217) - Chapter 5 Geometry and Topology

_ 20210922/Infinite Dimensional Analysis.../
3540295860.pdf (717 pág)
Pág. 21 (40) - Chapter 2 Topology
Pág. 69 (87) - Chapter 3 Metrizable Spaces

jandaia

)*+,)))- !) %1 "(1""" 3#
-4 M !

)*+,)))- !) %1 1 1 " "Q 3#
' !3 / 2 5 ! # 6 3 5

)*+,)))- !) %1 1/% (/ " 3#
-4 & 7 ! * ! 5
' % !3 6 3 5
' "(!3 1 @!4 3!

' (

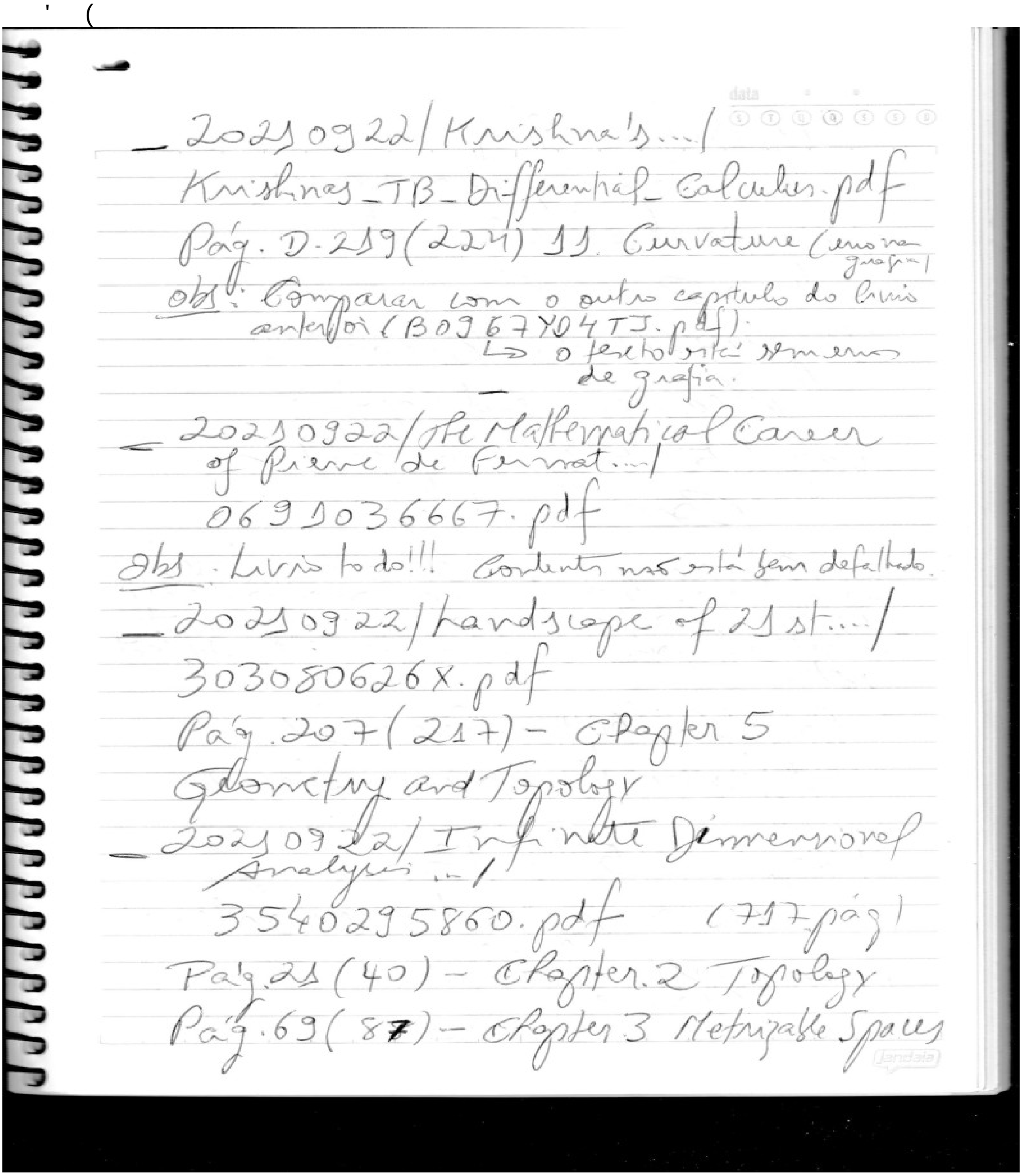

_ 20210922/Krishna's.../
Krishnas_TB_Differential_Calculus.pdf
Pág. D-219(224) 11. Curvature (erros na grafia)
obs: Comparar com o outro capítulo do livro anterior (B0967Y04TJ.pdf).
↳ o texto está sem erros de grafia.

_ 20210922/The Mathematical Career of Pierre de Fermat.../
0691036667.pdf
obs. Livro todo!!! Contents não está bem detalhado.

_ 20210922/Landscape of 21 st.../
303080626X.pdf
Pág. 207(217) - Chapter 5 Geometry and Topology

_ 20210922/Infinite Dimensional Analysis.../
3540295860.pdf (717 pág)
Pág. 21(40) - Chapter 2 Topology
Pág. 69(87) - Chapter 3 Metrizable Spaces

Pág. 163(180) - Chapter 5
Topological vector spaces
Pág. 251(268) Chapter 7
Convexity
Pág. 433(448) Chapter 12
Measures and Topology
Pág. 555(568) Chapter 17
Correspondences

A correspondence is a set-value function. That is, a correspondence assigns to each point in one set a set of points in a possibly different set.

Pág. 591(604) Chapter 18
Measurable correspondences

Pág. 571(584) 17.6
Vector-valued correspondences

⊙ – convex hull
(convex)

_20210921/ Infinite Dimensional...|
Basic Analysis III...pdf
Pág. 13(24) Chapter 2
Metric Spaces.

Obs. No mesmo diretório há toda a
coleção Basic Analysis I ao V.
✓

br1lib.org

_20210922/ Infinite Dimension...|
_An Introduction to Metric Spaces...pdf

_Metric Spaces by Satish Shirali...pdf
Pág. 64 (71) Chapter 2
Topology of a Metric Space

_Topology of Metric Spaces!...pdf
Pág. 02(15) Figure 1.1
Triangle Inequality.

We refer to d as the absolute value metric.

)*+,)))- !) %1 .! * ! 5 &&& !33
& 7 ! 3! !3 ! ! # ;! + 3#
' 1 % !3 3!

Table of Contents

I Introduction **1**

1 Introduction **3**
- 1.1 The Analysis Courses 3
  - 1.1.1 Senior Level Analysis 4
  - 1.1.2 The Graduate Analysis Courses 5
  - 1.1.3 More Advanced Courses 7
- 1.2 Table of Contents 8

II Metric Spaces **11**

2 Metric Spaces **13**
- 2.1 The Construction of the Reals 13
  - 2.1.1 Totally Ordering of the Rationals 14
  - 2.1.2 The Construction of the Reals 15
  - 2.1.3 The Equivalence Classes of Q are Totally Ordered 16
  - 2.1.4 The Ordering and the Operations are Compatible 18
  - 2.1.5 Cauchy Sequences of Equivalence Classes Converge 19
- 2.2 The Idea of a Metric 21
- 2.3 Examples 23
  - 2.3.1 Homework 24
- 2.4 More on Symbol Sequences 25
  - 2.4.1 Homework 26
- 2.5 Symbol Spaces over Two Symbols 27
  - 2.5.1 Homework 28
- 2.6 Periodic Points of a Map 28
- 2.7 Completeness 31
  - 2.7.1 Homework 35
- 2.8 Function Metric Spaces 36
- 2.9 Sequences and Series of Complex Numbers 38
- 2.10 Sequence Metric Spaces 41
  - 2.10.1 Sequence Space Metrics 48
- 2.11 Hölder's and Minkowski's Inequality in Function Spaces 49
  - 2.11.1 Function Space Metrics 53
- 2.12 More Completeness Results 55
- 2.13 More on Separability 61

3 Completing a Metric Space **65**
- 3.1 The Completion of a Metric Space 65
- 3.2 Completing the Integrable Functions 68

)*+,)))- !) %1 * & # 3! 3#

F)*+,)))- !) %1 3! 45 ! ! G
;! D ! < ! # 8! @ 4 3#

D)*+,)))- !) %1 6 3 5 3! 45
? ! ! @ 4 3#

-4
F)*+,)))- !) %1 3! 45 ! ! G
;! D ! < ! # 8! @ 4 3#
' "% !3 6 3 5 ! 3!

D)*+,)))- !) %1 6 3 5 3! 45
? ! ! @ 4 3# P ' / M
'

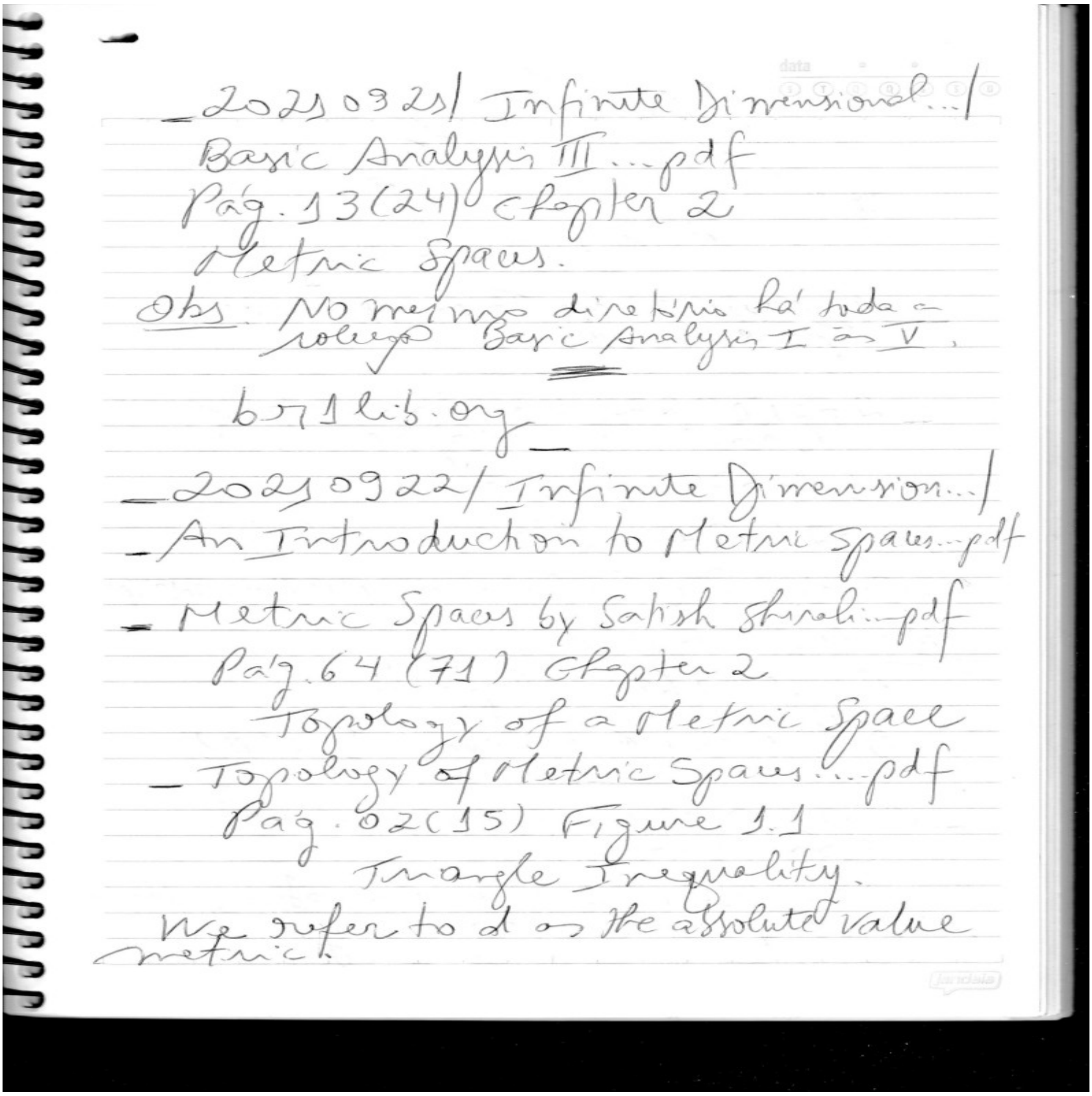

_20210921/ Infinite Dimensional.../
Basic Analysis III ...pdf
Pág. 13(24) Chapter 2
Metric Spaces.

Obs: No mesmo diretório há toda a coleção Basic Analysis I ao V.

br1lib.org

_20210922/ Infinite Dimension.../
_An Introduction to Metric Spaces...pdf
_Metric Spaces by Satish Shirali...pdf
Pág. 64 (71) Chapter 2
Topology of a Metric Space
_Topology of Metric Spaces....pdf
Pág. 02(15) Figure 1.1
Triangle Inequality.

We refer to d as the absolute value metric.

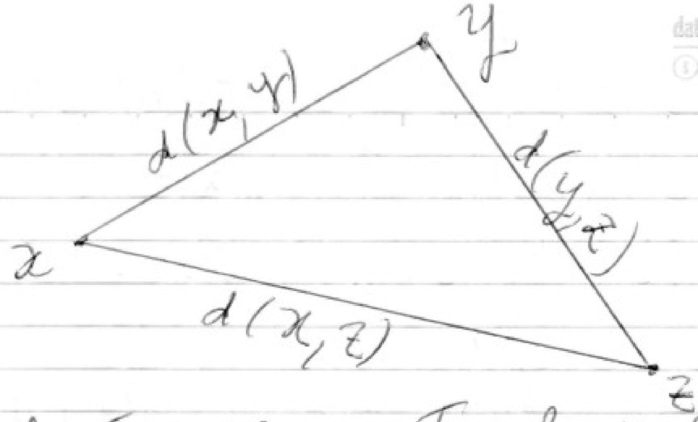

- A Course on Topological Vector Spaces... pdf

- 20210922/ Random Walk, Brownian Motion.../

  3030789373.pdf

- 20210922/ Celestial Calculations.../

• Celestial Calculations... MIT Press (2019).pdf

Obs: Ver livro todo!!!

Pág. 53(68) Chapter 4

Orbits and Coordinate Systems.

- 20210922/ Homogeneous Coordinates/

  The Role of Nonassociative Algebra in Projective Geometry...pdf

)*+,)))- !) %1 * 6 3 !
3! 45 O @ 4 3#

)*+,)))- !) %1 1 1 (1 1 3#

)*+,)))- !) %1 ! ! ! * 2
& # 3 ! ! * 5 &6 (3#
' /1 " !3 % - 4 ! # # ! 5

)*+,)))- !) %1 6 + C ! ! 8 * 4 !
F 8 2 5 3#

'

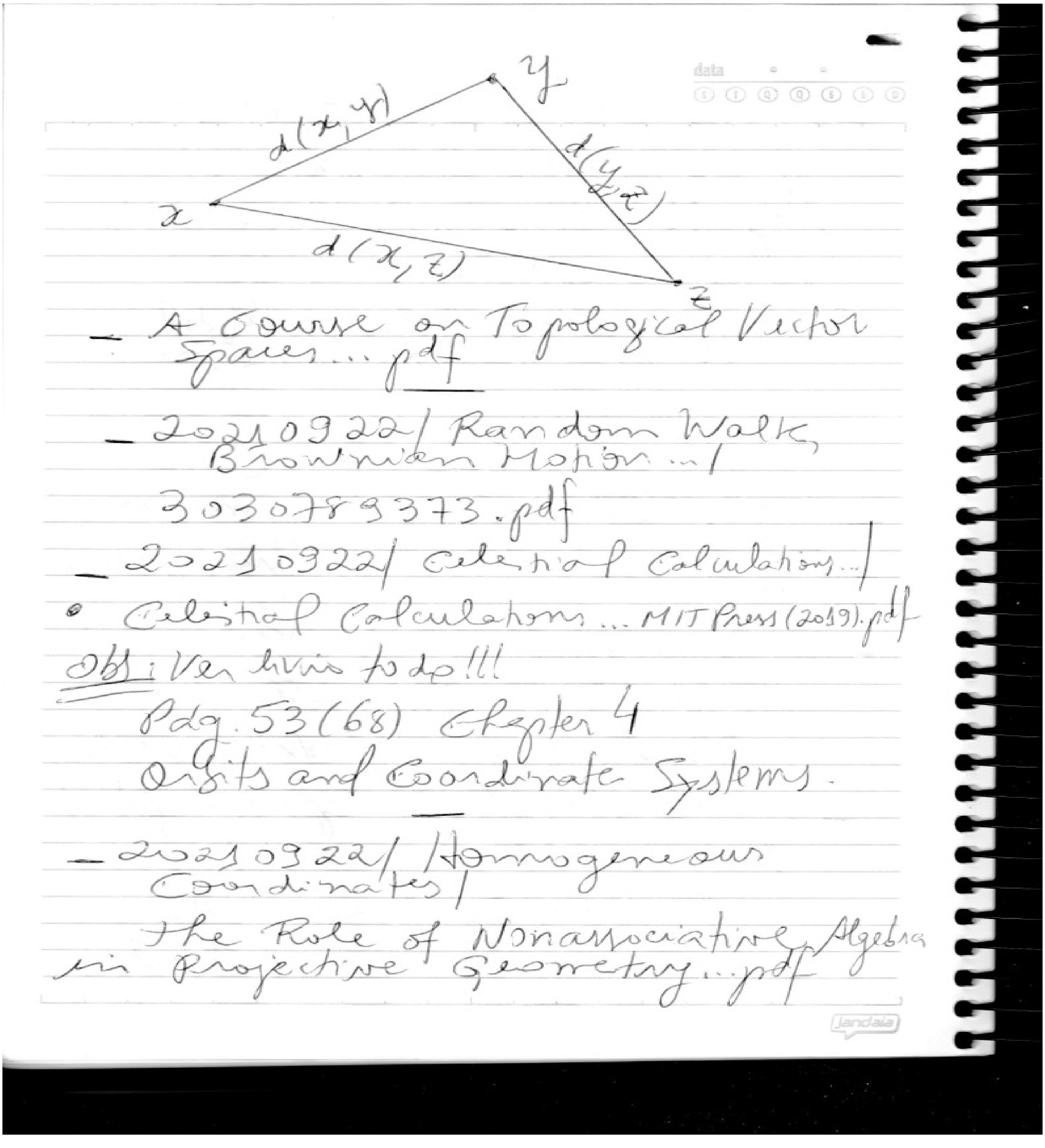
y
d(x, y)
d(y, z)
x
d(x, z)
z

- A Course on Topological Vector Spaces... pdf

- 20210922/ Random Walk, Brownian Motion.../
3030789373.pdf
- 20210922/ Celestial Calculations.../
• Celestial Calculations... MIT Press (2019).pdf

Obs: Ver livro todo!!!
Pág. 53(68) Chapter 4
Orbits and Coordinate Systems.

- 20210922/ Homogeneous Coordinates/
The Role of Nonassociative Algebra in Projective Geometry...pdf

Pág. 01 (17) Chapter 1
Affine and Projective Plane
Pág. 08 (24) Affine projection. (image)
1.5 Duality.
Pág. 23 (39) Chapter 3
Coordinates for Projective Planes
Pág. 99 (115) Chapter 7
Projective Geometries
Pág. 131 (147) Chapter 10
Homogeneous Maps
Pág. 181 (197) Chapter 13
Projective Remoteness Planes

_ 20210922/ Homogeneous Coordinates/
Methods of Algebraic Geometry in Control Theory. Part I... pdf
Pág. 65 (74) 1.2. Affine Algebraic Geometry: Products, Graphs and Projections.

_ Methods of Algebraic Geometry... Part II... pdf

N)*+,)))- !) %1 # * 4 ! 2 5
6 5 ! & ! ! < ! 5 ! # * * 4 ! 2 5 3#
' "/ % * * 4 ! 2 5 # G 2 !3 ! # F

N)*+,)))- !) %1 # * 4 ! 2 5
6 5 ! && 8! !4 < ! 5 ! # F 8 * 4 ! 2 5 3#
' / / F 8 3!

' %

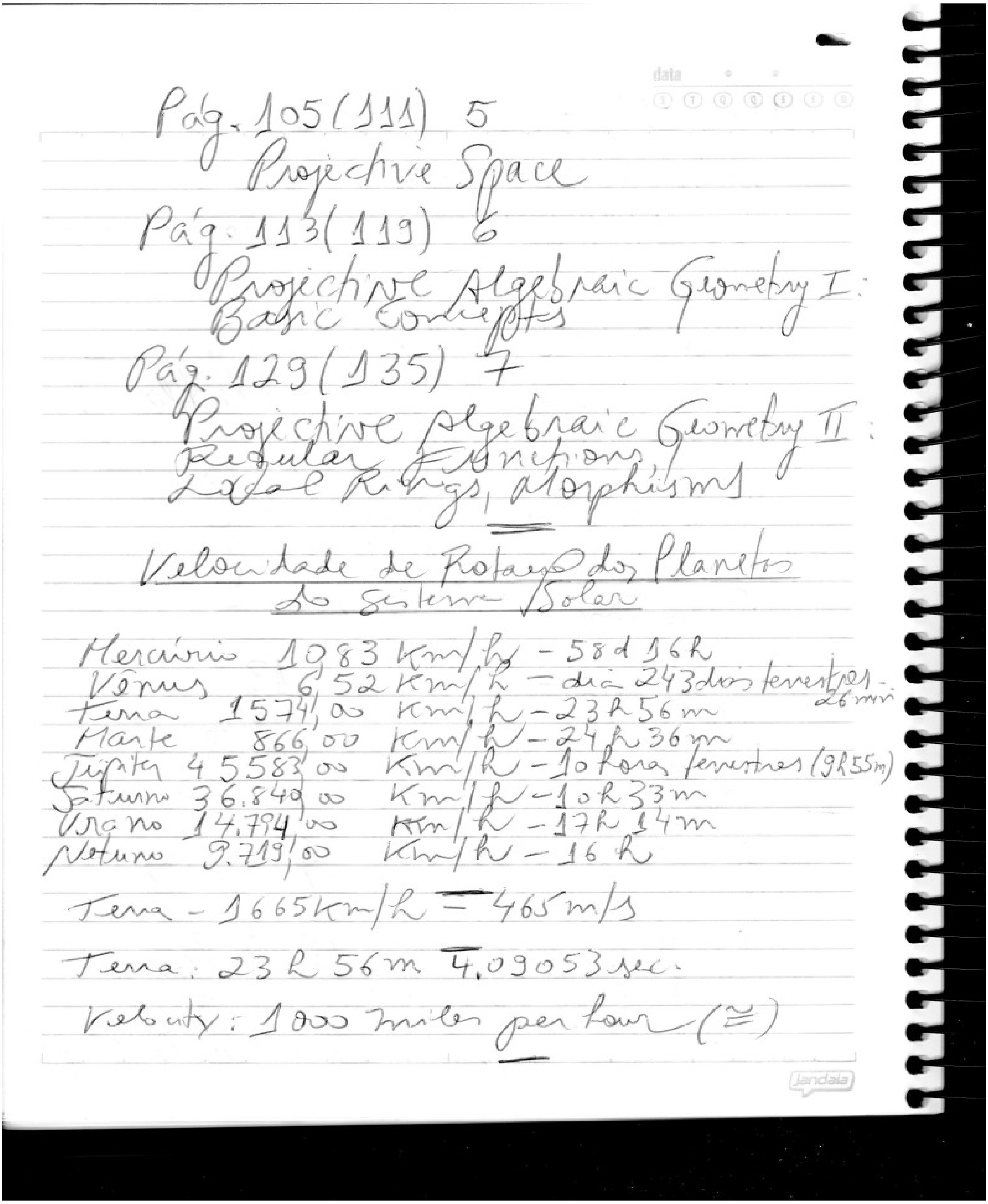

Pág. 105 (111) 5
Projective Space
Pág. 113 (119) 6
Projective Algebraic Geometry I: Basic Concepts
Pág. 129 (135) 7
Projective Algebraic Geometry II: Regular Functions, Local Rings, Morphisms

Velocidade de Rotação dos Planetas do Sistema Solar

Mercúrio 10,83 Km/h - 58d 16h
Vênus 6,52 Km/h - dia 243 dias terrestres 26 min
Terra 1574,00 Km/h - 23h 56m
Marte 866,00 Km/h - 24h 36m
Júpiter 45583,00 Km/h - 10 horas terrestres (9h55m)
Saturno 36.840,00 Km/h - 10h 33m
Urano 14.794,00 Km/h - 17h 14m
Netuno 9.719,00 Km/h - 16 h

Terra - 1665 Km/h ≅ 465 m/s

Terra: 23h 56m 4,09053 sec.

Velocity: 1000 miles per hour (≅)

Circumference of the Earth - 40.070km
is 1670 Km/h (1037 miles/h).

cosine of 45° is 0.707, so the spin speed at 45 degrees is roughly

0.707 x 1037 = 733 mph (1.180Km/h)

Orbital speed 107.000 Km/h
(107.000)

distance from Earth to the Sun —
92,955,807 miles (149,597,870 km)

circumference of a circle $2\pi r$
940.000.000 Km

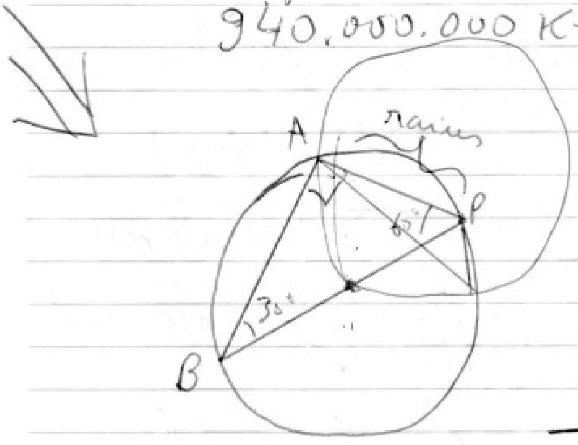

Construct Equilateral Triangle

Livros: Data-NovoVivo-019-2021/ Books-019-2021/
Important/

data

20210923/// 2412-5333.

algoritmo ganancioso.

Books:

Data-NovoVivo-019-2021/
Books-019-2021/
20210923/
Math

- 20210923/Analytical Geometry.../

B095PDCGXP-Analytical-
Geometry... pdf

- 20210923/Math/

475961.pdf

→ Advanced Geometry of
Plane Curves and their
applications.
(Zwikker)

- 20210923/Elements of Mathematics.../

Geometry-Intuitive, Discrete, and
Convex... pdf

Pág. 25-2. An Atom for the
Set of Parallelohedra

.--? @).--?). D ( ) (1) !)
)*+,)))- !) %1 . (/ 7 2Q * ! 5 !
2 5 ! 3#

)*+,)))- !) %1 % /(" 3#
P 6 *#8! # 2 5 ! 8 ! # 6 *33 ! 0 DD

)*+,)))- !) %1 2 5 & 8 G 7 G ! #
8 9 * 6 4 <' @ R M F 6R 3 3#
' / * * ! ! # !

' "

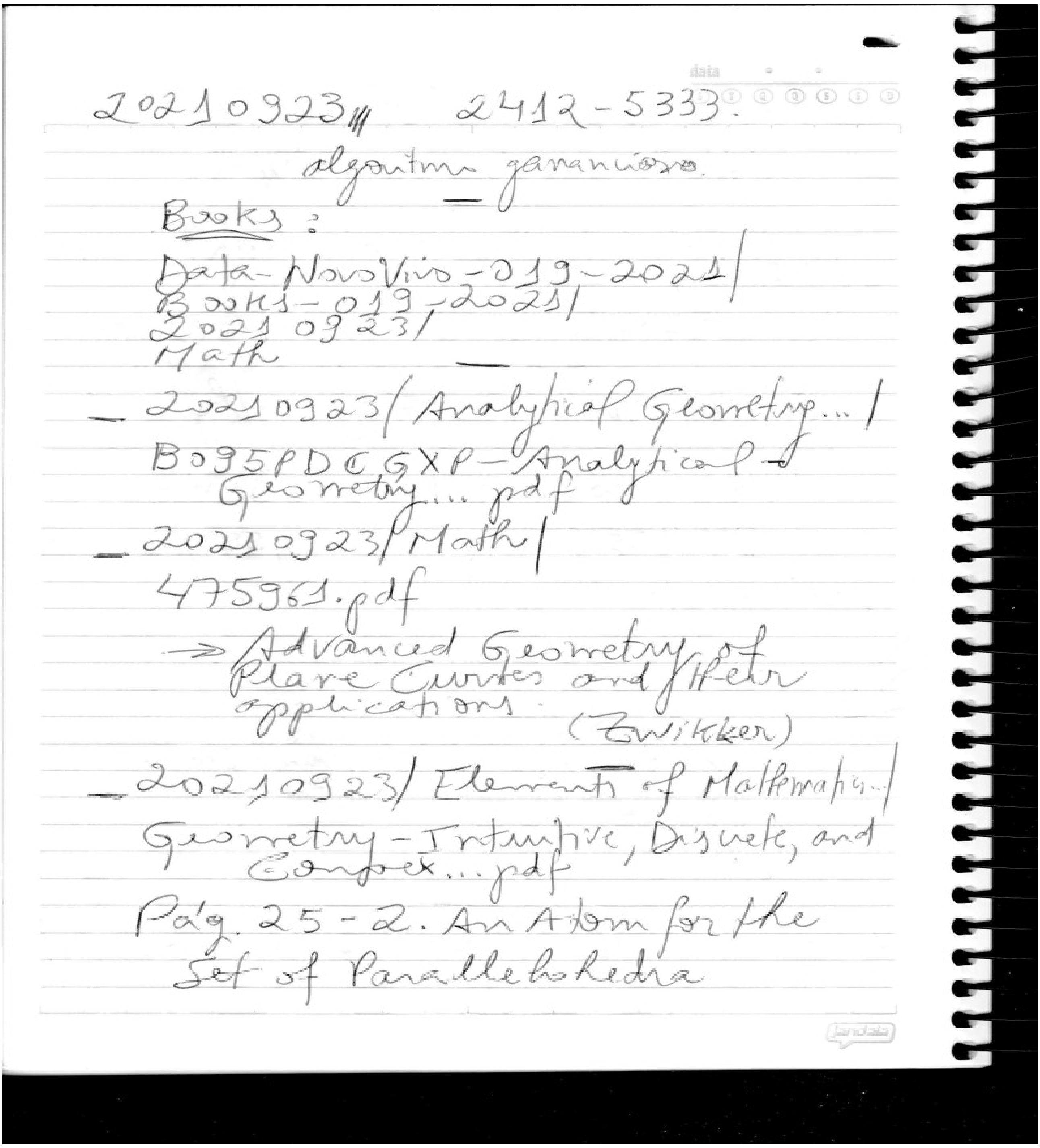

20210923/// 2412-5333.

algoritmo ganancioso.

Books:

Data-NovoVivo-019-2021/
Books-019-2021/
20210923/
Math

- 20210923/Analytical Geometry.../
B095PDCGXP-Analytical-Geometry....pdf
- 20210923/Math/
475961.pdf
→ Advanced Geometry of Plane Curves and their applications. (Zwikker)
- 20210923/Elements of Mathematics.../
Geometry-Intuitive, Discrete, and Convex....pdf
Pág. 25 - 2. An Axiom for the Set of Parallelohedra

Fig. 2 (b) Developments of Pentadra

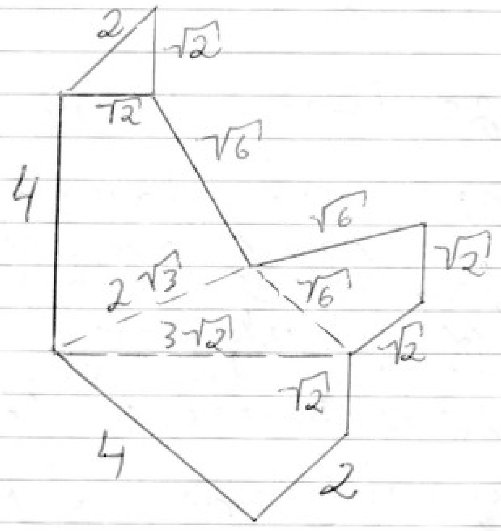

2
$\sqrt{2}$
$\sqrt{2}$
$\sqrt{6}$
4
$\sqrt{6}$
$\sqrt{2}$
$2\sqrt{3}$
$\sqrt{6}$
$3\sqrt{2}$
$\sqrt{2}$
$\sqrt{2}$
4
2

(b)

_ 20210923/ Elements of Mathematics/
New Trends in Intuitive Geometry...
(2018).pdf

Pág. 219(225) —
The Twelve spheres Problem

_ 20210923/ Elements of Mathematics.../
Treks into Intuitive Geometry...
(2016).pdf.

_ 20210923/ Understanding Gravitational
Waves/ 3030742067.pdf

denote the five families of parallelohedra, namely, parallelepiped, rhombic dodecahedron, hexagonal prism, elongated rhombic dodecahedron, and truncated octahedron, respectively. For each of these five families F_i ($i = 1, \ldots, 5$), choose a polyhedron $P_i \in F_i$ to be a **representative** in any manner. The resulting set $\Sigma = \{P_1, \ldots, P_5\}$ consists of five parallelohedra, and we call it a **representative set of parallelohedra.**

Let $\Sigma = \{P_1, \ldots, P_5\}$ be a representative set. Given any parallelohedron $P \in \Pi$, there exists an affine-stretching transformation $A : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ such that $A(P) \in \Sigma$. Hence, in order to determine an atom for the set Π, it suffices to find a polyhedron σ and a representative set $\Sigma = \{P_1, \ldots, P_5\}$ such that each parallelohedron $P_i \in \Sigma$ is the union of a finite number of copies of σ. Therefore, in this paper we consider the **representative set** that consists of the cube, the rhombic dodecahedron, the skewed hexagonal prism, the elongated rhombic dodecahedron, and the truncated octahedron.

2. AN ATOM FOR THE SET OF PARALLELOHEDRA

Throughout the paper, we denote the set of all parallelohedra by Π, and we use Σ to denote the **representative set** of parallelohedra that consists of the cube, the rhombic dodecahedron, the skewed hexagonal prism, the elongated rhombic dodecahedron, and the truncated octahedron. In the following theorem, we will show that a special kind of pentahedron, which we call a **pentadron**, is an **atom** for the set Π. A **pentadron** is either one of the pentahedra shown in Figure 2(a), with developments shown in Figure 2(b).

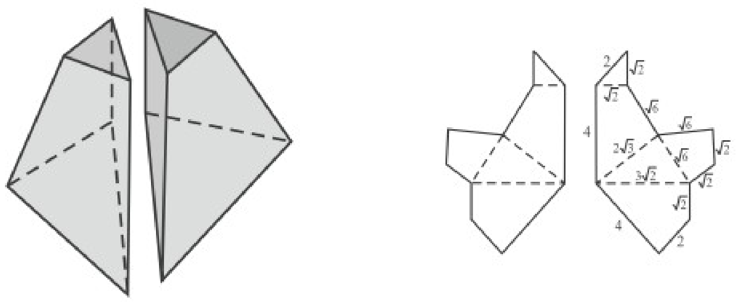

(a) Two Congruent Pentadra (b) Developments of Pentadra

Fig. 2. Pentahedra as an Atom

)*+,)))- !) %1 C # 8 5
3 3#
' (/ 6 6 8 3 4

'

Fig. 2 (b) Developments of Pentadra

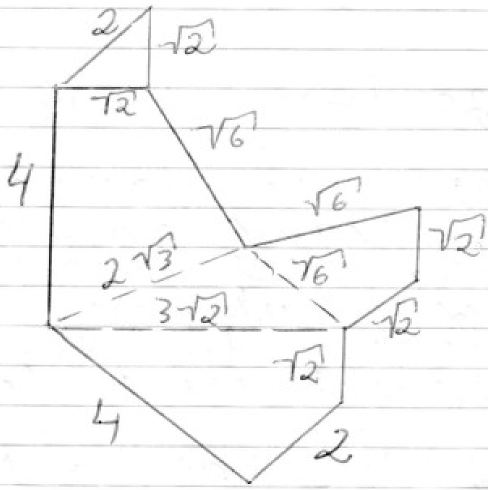

(b)

_ 20210923/ Elements of Mathematics/
New Trends in Intuitive Geometry...
(2018).pdf

Pág. 219(225) —
The Twelve Spheres Problem

_ 20210923/ Elements of Mathematics.../
Treks into Intuitive Geometry...
(2016).pdf.

_ 20210923/ Understanding Gravitational
waves/ 3030742067.pdf

8)*+,)))- !) %1 6 D & 8 2 5 6
S # 5 ! # 5 # ! 3 " 3#

9)*+,)))- !) %1 1 1 % " 3#
' 1(6 7 ! # 6 C !

Rotating asymmetrical objects will emit at their rotational frequencies, but only rapidly rotating neutron stars and black holes are likely to be detectable for a long time to come, so their frequencies of interest range from about 10 Hz to 10 kHz. Asymmetrical supernova explosions and the like will probably emit between 100 mHz and 10 Hz.

{90} GW170817: Three Detectors and Two Neutron Stars

We have seen (Chapters 1 and 4) that the first gravitational wave event, **{10}** GW150914, arose from the merger of two black holes. Furthermore, as just considered above, several types of compact object mergers and binary systems may potentially produce observable gravitational waves. All currently known gravitational wave events are reviewed in Chapter 6 and the vast majority of those also arise from black hole mergers. Let us here, therefore, look in more detail at a gravitational wave event which resulted from one of the other types of sources discussed above and which is different in many ways from the black hole mergers.

This event, **{90}** GW170817, originated in the merger of two neutron stars and is of great significance in its own right. It was also only the second event clearly detected by three gravitational wave detectors; the two AdvLIGO instruments and AdvVirgo (Chapters 6 and 8), which enabled a fairly accurate position in the sky to be found for it. This position then enabled **{90}** GW170817 to be identified with a gamma ray burst and radio and optical transient events so that it was the first and, so far, only gravitational wave event to be unequivocally detected via its electromagnetic emissions as well.

Two well-separated detectors enable the position of the source of the gravitational waves to be constrained to lie within an annulus of the sky. However, that annulus still covers a large area of the sky (Figure 5.3), so any attempts to locate other emissions from the event, such as gamma rays, light, neutrinos, etc., are almost certainly doomed to failure; even if black hole mergers led to significant such emissions which, by and large, they do not.

During AdvLIGO's second observing run (O2), however, it was joined on the 1st of August 2017 by the AdvVirgo gravitational wave detector beginning its science operations (Chapter 8). With three detectors, the problem of identifying the source of a gravitational wave event is dramatically changed. By taking the detectors in pairs, we get three position annuli on the sky and, in exactly the same way that navigators use triangulation of three sightings to determine their positions at sea, so the position of the source may be identified to lie within a much smaller area of the sky by triangulation of the position annuli (Figure 5.3).

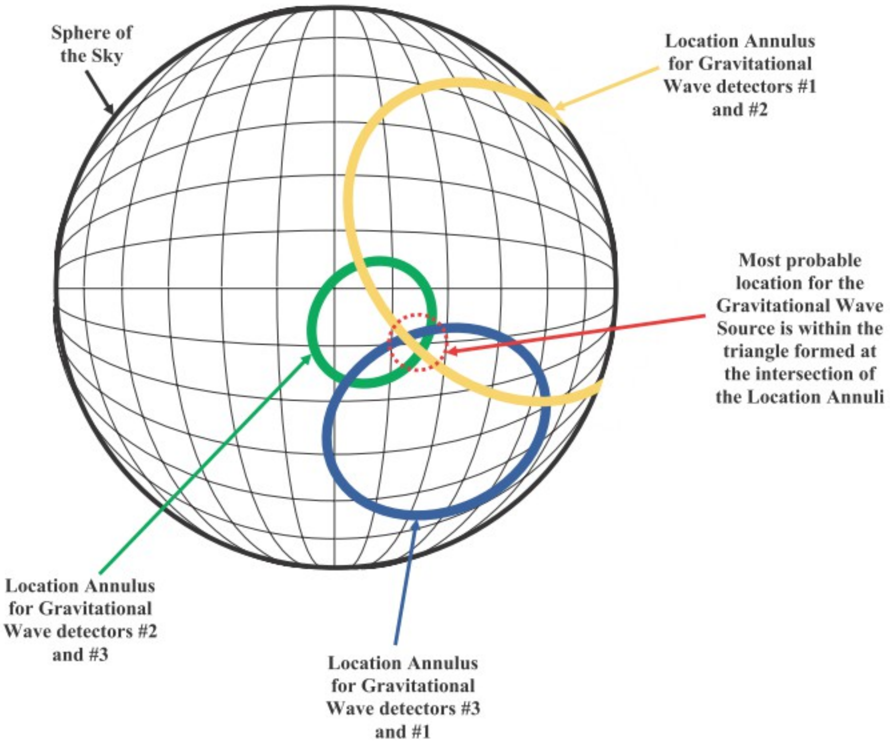
Sphere of the Sky
Location Annulus for Gravitational Wave detectors #1 and #2
Most probable location for the Gravitational Wave Source is within the triangle formed at the intersection of the Location Annuli
Location Annulus for Gravitational Wave detectors #2 and #3
Location Annulus for Gravitational Wave detectors #3 and #1

Figure 5.3 The narrowing down of the position of a gravitational wave source when three detectors are in use, by triangulation. The huge reduction in the necessary search area is clear to see

The AdvLIGO and AdvVirgo operators were very lucky—within two weeks a black hole merger was detected by all three detectors. **{80}** GW170814—detected on the 14th of August 2017—was around 580 Mpc away and involved black holes with masses of 30.7 $M_{\odot}$ and 25.3 $M_{\odot}$ (Chapter 6). The triangulation from the three detectors limited its position to within about 60 square degrees; ~10% of the smallest possible search area for **{10}** GW150914 and in an area of Eridanus between Horologium and Phoenix. As already mentioned, though, the energy released during black hole mergers goes mostly into the gravitational wave emissions, so even with the small location area to be searched, no electromagnetic or other emissions could be detected. The precise identification of the whereabouts of the source, that is, the galaxy within which it occurred, therefore, still remains unknown.

Just three days later, though; the 17th of August 2017, the next event was very different. **{90}** GW170817 involved two neutron stars merging, not black holes—though they may have formed a black hole after the merger (below). Neutron stars (Chapters 5 and 6), unlike black holes, are not reticent about emitting electromagnetic radiation across the whole spectrum and the resulting fireworks were observed from gamma rays to radio waves, as well, of course, as the gravitational wave event itself.

The three gravitational wave detectors triangulated the search area down to a lenticular zone close to gamma Hydra and covering about 16 square degrees (Figure 5.4). Ironically, AdvVirgo did not appear to have detected the event, but its negative data was still used to narrow the search area for the source—by utilising the area of the sky that was AdvVirgo's blind spot. Later work showed the AdvVirgo had made a detection, but that it was very faint through the source being located within that blind spot (Figure 5.6).

Sixteen square degrees (initially 28 square degrees) is still a large(ish) area of the sky—about 80 times the area of the full Moon—but it is small enough for radio and optical astronomers to search with some hope of successfully finding the source from its electromagnetic emissions. Those searches started within a few hours and had their first result within less than half a day.

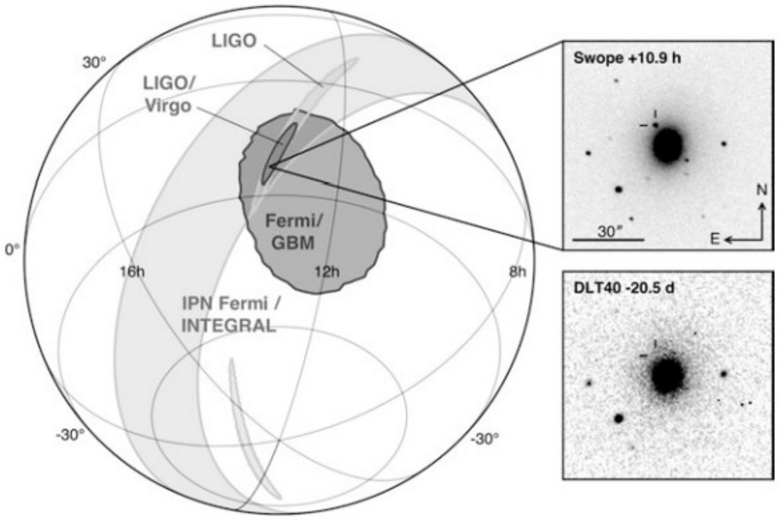

Figure 5.4 Triangulation of the search area for **{90}** GW170817 from the two AdvLIGO detectors and the AdvVirgo detector (indicated as 'LIGO/Virgo'). Two narrow lenticular areas based upon the two AdvLIGO detectors alone (the top one indicated as 'LIGO') are also shown. Plus, the possible source location areas for the gamma ray burst GRB170817A from the Fermi gamma ray Burst Monitor ('Fermi/GBM') and from timing measurements by Fermi and INTEGRAL spacecraft detections processed by the InterPlanetary Network ('IPN Fermi/INTEGRAL') are indicated. An image of the optical transient thought to be identified with both **{90}** GW170817 and GRB170917A in NGC 4993 (shown by the two dashes) obtained 10.9 hours after GRB170817A by the 1 m Swope telescope in Chile is shown at the upper right, while a pre-discovery image of NGC4993 obtained by the 0.4 m PROMPT telescope in Chile (operated by the DLT Supernova Search Collaboration and which also detected the optical counterpart of the **{90}** GW170817 event) is shown at the bottom right for comparison. (Reproduced by courtesy of B. P. Abbott *et al*. 'Multi-messenger Observations of a Binary Neutron Star Merger'. ApJ Lett., L12, **848**, 2017, DOI:https://doi.org/10.3847/2041-8213/aa91c9)

Before then, however, and independently of the gravitational wave detection, two gamma ray space observatories, Fermi and INTEGRAL, had detected a gamma ray burst just 1.74 seconds later than AdvLIGO's and AdvVirgo's receptions of the gravitational wave (Figure 5.5). GRB170817A was registered as occurring at 12:41:06 UTC, and its detection automatically generated an electronic alert to other observatories just 14 seconds later. **{90}** GW170817 occurred at 12:41:04 UTC; being received in the order: AdvVirgo → 22 milli-seconds → Livingston → 3 milli-seconds → Hanford, but the first, low priority, alert did not go out for 26 minutes. The first alert that gave a reasonably precise and accurate search area for the gravitational wave event was sent out just over 12 hours after it had occurred—so the initial searches were actually looking for the source of the gamma ray burst, not the gravitational wave event.

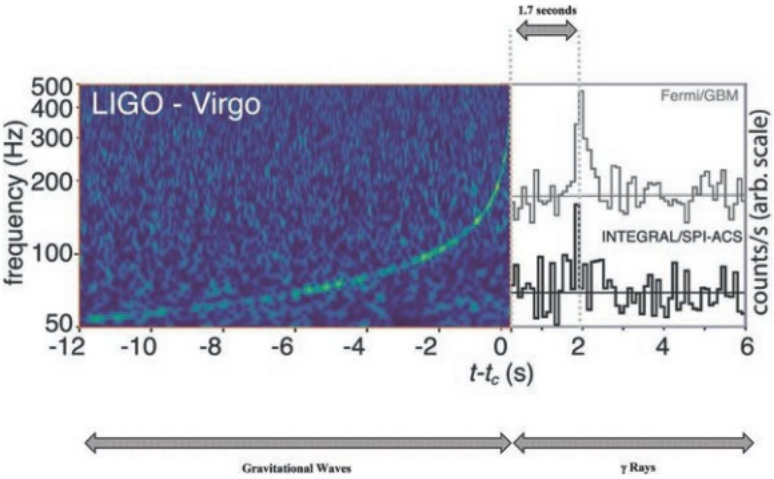

Figure 5.5 The times of reception of the **{90}** GW170817 and GRB170817A signals at AdvLIGO Livingston/Hanford and Fermi/INTEGRAL. (Reproduced by courtesy of B. P. Abbott *et al*. 'Multi-messenger Observations of a Binary Neutron Star Merger'. ApJ Lett., L12, **848**, 2017, https://doi.org/10.3847/2041-8213/aa91c9)

Of course, it could be that the gravitational wave emission, the gamma ray burst and the optical transient were three separate and independent events. However, the luminosity distance to **{90}** GW170817 is 40 Mpc (±10 Mpc) and those to the optical transient and NGC4993 from their cosmological redshifts are both 42.9 Mpc (±3.2 Mpc); so the three distances are essentially the same. Adding the excellent overlap of the search areas (Figure 5.4) to the distance estimates, then a false alarm; that is, just coincidence of three independent events, would occur only about once in 80,000 years. Hence **{90}** GW170817, GRB170817A and the optical transient in NGC4993 are now unanimously taken to have arisen from the same binary neutron star merger.

The distance of **{90}** GW170817 and its associated events of about 43 Mpc is equal to about 140 million light years. That means that although we saw the events in the year 2017, they actually occurred some 140,000,000 years earlier; the look-back time. This would have placed it in the middle of the Mesozoic era, when the first mammals were starting to evolve and dinosaurs were at their peak. For comparison, the look-back time to **{10}** GW150914 is about ten times larger—during the Earth's Proterozoic era—when multicellular life was still struggling to emerge.

The masses of the two neutron stars involved in **{90}** GW170817 were relatively small—the primary probably being between 1.36 and 1.60 $M_{\odot}$ and the secondary between 1.17 and 1.36 $M_{\odot}$, with a total mass close to 2.74 $M_{\odot}$. These lower masses than those of the merging black hole binaries (Chapters 4 and 6) meant that the merger event happened more slowly. The signals for **{90}** GW170817 thus lasted for nearly 30 seconds (Figures 5.5 and 5.6)—compared with the 200 milli-seconds for **{10}** GW150914 (Figure 3.1). The mass lost during the merger and so also the total amount of energy radiated away was about 1% that of **{10}** GW150914, but the distance being only about a tenth that of **{10}** GW150914, meant that the apparent gravitational brightnesses of the two events were similar. **{90}** GW170817, though, radiated much more electromagnetic energy than **{10}** GW150914.

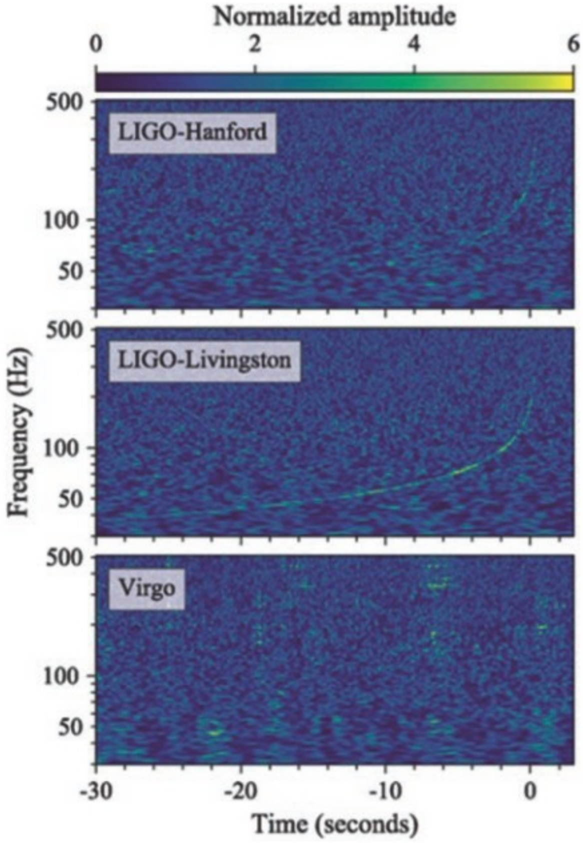

Figure 5.6 {90} GW170817—The three gravitational wave signals for AdvLIGO Hanford, AdvLIGO Livingston and AdvVirgo. Note the absence of a noticeable signal from AdvVirgo, arising from {90} GW170817's source being within AdvVirgo's blind spot. (Reproduced by courtesy of B. P. Abbott *et al*. 'GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral'. Phys. Rev. Lett., 161101-1, **119**, 2017. DOI:https://doi.org/10.1103/PhysRevLett.119.161101)

The merger of the neutron stars resulted in a short gamma ray burst (sGRB—GRB170817A), as already mentioned. Some tens of sGRBs have now been observed and, of those whose distances can be estimated, GRB170817 at about 40 Mpc, is the closest to us so far. The merger of two neutron stars leaves a large amount of debris surrounding the final merged object. The electromagnetic emissions are largely coming from this debris. The energy for those emissions is derived from radioactive decay of elements produced by nucleosynthesis during the merger. For {90} GW170817 about 0.05 $M_\odot$ or about 20,000 Earth masses, of such radioactive elements are estimated to have been synthesised, including about ten Earth masses of platinum and gold—which latter detail is what made the popular-press headlines.

The radioactive elements are built up from lighter nuclei during the merger through their rapid capture of free neutrons (the r-process). In this they resemble 'ordinary' novae, but the total energy emitted in that way is up to a thousand times that of a nova. The processes leading up to the electromagnetic events are hence sometimes given the name of 'Kilonova' (also Macronova).

The merged object in the case of **{90}** GW170817 was initially thought to be a small(ish) black hole, but the possibility of it being a very massive magnetar; a neutron star possessing a very intense magnetic field—(Chapter 5) is now thought to be the more probable. The evidence for this is a five second chirp, descending in frequency from about 1 kHz to 49 Hz, occurring immediately after the initial ascending frequency chirp and observed by the two AdvLIGO detectors. This descending frequency chirp is interpreted as arising from the simple slowing of the rotation or spin down of the merged object as a neutron star rather than the ringdown of a merged black hole as in **{10}** GW150914 (Chapters 4 and 7). More recent observations have, at least partially, re-raised the possibility that the merged object has become a black hole—so that matter is still being debated at the time of writing.

In the months following the outburst, the radio emissions from the electromagnetic transient increased in intensity steadily, but after half a year or so, they started to fade rapidly. Modelling of this behaviour suggests that material ejected during the outburst was initially expanding outwards relatively slowly and cocooning the central remnant. Later a narrow high-speed jet developed that punched its way through the cocoon and is now taking away most of the energy of the outflow. As of the time of writing (Dec 2020), gamma, X-ray and radio emissions are still being received from the debris, though declining in their intensities steadily. The optical emissions, though, are now too faint to detect.

Scientifically, though, the most significant observation to come from **{90}** GW170817 was probably the 1.7 second delay between the gravitational and gamma ray signals. Even if the peak of the gravitational wave emission and that of the gamma ray emission exactly coincided—and they probably did not, the 1.7 second difference in the arrival times after travelling 43 Mpc means that the speeds of gravity and light are almost identical. As we have seen earlier (Chapter 1), the speed of gravity being equal to the speed of light is a fundamental assumption underlying General Relativity. So, this result provides yet more strong evidence for the validity of General Relativity over alternative theories of gravity—some of which would have predicted delays between the two signals measured in years.

If we assume that the gravitational and gamma ray signals were emitted simultaneously, then the delay implies that the speed of gravity lies between about 9×10^{-7} m s^{-1} slower than the speed of light and 2×10^{-7} m s^{-1} faster than the speed of light—for comparison, an oak tree grows in height at a rate of about 2×10^{-8} m s^{-1}. However, of course, it is quite likely that the two signals were not emitted simultaneously—they originate from different physical processes, after all—and so the speeds of gravity and light could well be exactly equal. Confirmation of this equality, based upon the gravitational and gamma ray signal timings, has come independently from the transit times of the gravitational waves alone between the detectors for eight of the 11 O1 and O2 events.

The 1.7 second delay also provides a second confirmation of the probable validity of General Relativity via testing the equivalence principle. The Shapiro effect (Chapter 1) should act on both gravitons, if these are massless (Chapter 3), and photons equally. The magnitudes of those effects depend upon the mass distribution all along the path from their source to the Earth and that is not well known. However, an approximate estimate of that mass distribution plus the 1.7 second delay leads to the result that the Shapiro effect for gravitons differs from that for photons by less than one part in a million—again a striking confirmation of General Relativity. The result also sets an upper limit on the rest mass of the graviton of $\leq\sim 1.4 \times 10^{-58}$ kg; that is, at least 10^{28} gravitons would be required to add up to the mass of one electron—the Chandra spacecraft though has already set a limit on the graviton mass that is a million or more times smaller than this value (Chapter 3).

From the estimated number of neutron star binaries in the Universe and their estimated merger rate, events like **{90}** GW170817 may occur within a spherical volume, 1,000 Mpc in radius and centred on the Earth, at a rate of between 1,000 and 20,000 per year. **{90}** GW170817 was 'only' about 40 Mpc, that is about 400 times the size of the optically visible part of our Milky Galaxy, away. With the levels of sensitivity which AdvLIGO and AdvVirgo achieved during the O2 observing run, it provided a relatively strong signal—probably, it could have been detected out to some 100 Mpc or so.

With the O3 observing run (Chapter 6), sensitivities were improved by about 50–100% over the O2 run. Although the data analysis is still in progress at the time of writing, an additional five binary neutron star merger events were detected with reasonable certainty during run O3, with another four events being possibilities—though none of these events has had associated gamma ray or other non-gravitational transient events detected. The actual discovery rate of neutron star binary mergers was thus about one per month, compared with nearly six mergers per month involving black holes or black hole and neutron stars.

Some Final Questions and Thoughts About Gravitational Wave Sources

> I think the problem, (said the Deep Thought computer—Author's note) to be quite honest with you, is that you've never actually known what the question is. (Douglas Adams. 'Hitchhiker's Guide to the Galaxy'. BBC Radio 4, 1978).

Does the Earth Generate Gravitational Waves?

Yes, of course.

It is a member of a complex multiple system of mutually interacting objects, that is, the Solar System, with the Earth–Moon and Sun–Earth binary systems being the main, but not the only, contributors to the Earth's gravitational emissions. The Earth is also rotating and is asymmetric—Everest and the Marianas trench immediately spring to mind, but the Earth is also squashed along its polar axis; an oblate spheroid, so that its polar diameter is 42 km shorter than its equatorial diameter. On the intermediate size scales, there are very large regions of the Earth that are higher or lower than the average by tens of metres. Much of Europe and the North Atlantic, for example, is raised by around 50 m, whilst most of the Indian ocean floor is lower than average by 80 or more metres.

More significantly, the Earth's mantle is not uniform and large masses move slowly around so that the Earth's pattern of asymmetry changes. N.B.—it is the instantaneous pattern of the Earth's asymmetry combined with its rotation that produces gravitational waves of even remotely possible interest. While the motions of masses of material within the mantle doubtless will also produce gravitational waves, it seems unlikely that they could ever be detectable.

Pág. 127(139) - GW170817:
Three Detectors and Two Neutron Stars

Pág. 128(140) - Figure 5.3
The narrowing down of the position of a gravitational wave source when three detectors are in use, by triangulation.
The huge reduction in the necessary search are is clear to see.

- 20210923/ Elements of Mathematics../

Toth2021_Book_Elements of Mathematics.pdf

Pág. 207(218) Chapter 5
Real Analytic Plane Geometry

Pág. 224(235) Babylonian Method for $\sqrt{2}$ starting with $q_0 = 1$:

5 $\frac{886731088897}{627013566048}$

Pág. 225(236) 5.5
Lines and Circles

5)*+,)))- !) %1
6 L. DLA - ! ! 3#
' !3 / + ! * ! 5 ! 2 5
' % 1/ .!45 !

Note that, by the substitution above, if a, b, c are the side lengths of a triangle, the inequality of Example 5.4.3 above gives[14]

$$\sqrt{a+b-c}+\sqrt{b+c-a}+\sqrt{c+a-b} \leq \sqrt{a}+\sqrt{b}+\sqrt{c}.$$

Example 5.4.4 Show that, for $0 \leq x_1, \dots, x_n \in \mathbb{R}$, $2 \leq n \in \mathbb{N}$, we have

$$\sum_{1 \leq i < j \leq n} x_i x_j \cdot \sum_{k=1}^{n} x_k^2 \leq \frac{1}{8} \left(\sum_{k=1}^{n} x_k \right)^4 .$$

Indeed, using the AM-GM inequality as in Example 5.4.1, we calculate

$$\begin{aligned} \left(\sum_{k=1}^{n} x_k \right)^4 &= \left(\sum_{k=1}^{n} x_k^2 + 2 \sum_{1 \leq i < j \leq n} x_i x_j \right)^2 \\ &\geq 4 \left(\sum_{k=1}^{n} x_k^2 \right) \left(2 \sum_{1 \leq i < j \leq n} x_i x_j \right) = 8 \sum_{1 \leq i < j \leq n} x_i x_j \cdot \sum_{k=1}^{n} x_k^2 \end{aligned}$$

The inequality now follows.

We complete this cadre of examples by one that shows how the AM-GM inequality can sometimes be used to solve a system of non-linear equations.

Example 5.4.5 Solve the following system of equations for $x, y, z \in \mathbb{R}$:

$$x + y = 2, \quad xy - z^2 = 1.$$

First, $xy = 1 + z^2 \geq 1$, so that, by the first equality, we obviously have $x, y > 0$. Now, the AM-GM inequality gives $2 = x + y \geq 2\sqrt{xy}$. Hence $xy \leq 1$. Combining this with the previous inequality, we obtain $xy = 1$, and consequently $x = y = 1$. Finally, the second equality gives $z = 0$.

We finish this section by another application of the AM-GM inequality: The **Babylonian Method** on how to approximate the square root of a natural number $a \in \mathbb{N}$ by rational numbers.

We assume that $a \in \mathbb{N}$ is not a perfect square. We let $0 < q_0 \in \mathbb{Q}$ and define the sequence $(q_n)_{n \in \mathbb{N}_0}$ inductively by

$$q_{n+1} = \frac{1}{2} \left(q_n + \frac{a}{q_n} \right), \quad n \geq 0.$$

---

[14] This was a problem in the Asian Pacific Mathematical Competition, 1996.

Remark We will give a geometric interpretation of this formula in Section 8.2. This can be substantially generalized to what is known as **Newton's Method**, and it goes far beyond our rational approximations of the square root of a natural number. This particular case was also known to the ancient Babylonians, and this is why it carries the name "Babylonian Method."

Clearly, $(q_n)_{n \in \mathbb{N}_0}$ is a sequence of rational numbers. We claim that it is decreasing from the **first** term q_1 onward, and $\lim_{n\to\infty} q_n = \sqrt{a}$.

By our initial choice, we have $q_0 > 0$. Assuming $q_n > 0$, $n \in \mathbb{N}_0$, a quick look at the inductive formula above gives $q_{n+1} > 0$. It now follows from Peano's Principle of Induction that $q_n > 0$ for all $n \in \mathbb{N}_0$.
We can actually say more. The AM-GM inequality gives

$$0 < a = q_n \cdot \frac{a}{q_n} < \frac{1}{4}\left(q_n + \frac{a}{q_n}\right)^2 = q_{n+1}^2.$$

Note the sharp inequality in the middle as $q_n \neq a/q_n$ (since otherwise we would have $q_n^2 = a$, and a would be a perfect square). We thus obtain $q_{n+1}^2 > a$ for all $n \in \mathbb{N}_0$.

Rearranging this last inequality, we have

$$q_{n+1} > \frac{1}{2}\left(q_{n+1} + \frac{a}{q_{n+1}}\right) = q_{n+2}, \quad n \in \mathbb{N}_0.$$

We see that the sequence $(q_n)_{n\in\mathbb{N}}$ is strictly decreasing and bounded from below. By (Cauchy) completeness of $\mathbb{R}$, we have $\lim_{n\to\infty} q_n = r \in \mathbb{R}$, and, in addition, we also have $r^2 \geq a$. Letting $n \to \infty$ in the inductive definition of the sequence $(q_n)_{n\in\mathbb{N}}$ above, we obtain

$$r = \frac{1}{2}\left(r + \frac{a}{r}\right).$$

Rearranging, we obtain $r^2 = a$. This finally gives[15] $r = \sqrt{a}$.

To find the rate of convergence, we introduce the **relative error**

$$\delta_n = \frac{q_n}{\sqrt{a}} - 1 > -1, \quad n \in \mathbb{N}_0.$$

Clearly, $\delta_n \neq 0$ since $\sqrt{a}$ is irrational. We rewrite this as $q_n = \sqrt{a}(\delta_n + 1)$, $n \in \mathbb{N}_0$. We now claim that the following inductive relation holds for the relative error:

$$\delta_{n+1} = \frac{\delta_n^2}{2(\delta_n + 1)}, \quad n \in \mathbb{N}_0.$$

---

[15] Note that this can also serve as a definition of $\sqrt{a}$ as the equivalence class of the rational Cauchy sequence $(q_n)_{n\in\mathbb{N}}$.

Note that this implies $\delta_n > 0$ for $n \in \mathbb{N}$. (The initial δ_0 may be negative.) We now calculate

$$\delta_{n+1} = \frac{q_{n+1}}{\sqrt{a}} - 1 = \frac{q_n + a/q_n}{2\sqrt{a}} - 1 = \frac{q_n^2 + a}{2\sqrt{a}q_n} - 1$$
$$= \frac{a(1+\delta_n)^2 + a}{2\sqrt{a}\sqrt{a}(1+\delta_n^2)} - 1 = \frac{(1+\delta_n)^2 + 1}{2(\delta_n + 1)} - 1 = \frac{\delta_n^2}{2(\delta_n + 1)}.$$

The inductive formula for the relative error above follows.

Since $\delta_n > 0$ for $n \in \mathbb{N}$, this implies

$$\delta_{n+1} = \frac{\delta_n}{\delta_n + 1} \cdot \frac{\delta_n}{2} \leq \frac{\delta_n}{2} \quad \text{and} \quad \delta_{n+1} = \frac{1}{\delta_n + 1} \cdot \frac{\delta_n^2}{2} \leq \frac{\delta_n^2}{2}.$$

(The first estimate is better than the second for $\delta_n > 1$, which may happen for some initial values of the indices $n \in \mathbb{N}$.) Putting these together, we obtain

$$0 < \delta_{n+1} \leq \min\left(\frac{\delta_n}{2}, \frac{\delta_n^2}{2}\right), \quad n \in \mathbb{N}.$$

The first estimate implies $\delta_2 \leq \delta_1/2$ ($n = 1$), $\delta_3 \leq \delta_2/2 \leq \delta_1/2^2$ ($n = 2$), $\delta_4 \leq \delta_3/2 \leq \delta_1/2^3$ ($n = 3$), etc. In general, we have $0 < \delta_{n+1} \leq \delta_1/2^n$, $n \in \mathbb{N}$. In particular, we obtain $\lim_{n\to\infty} \delta_n = 0$. The Babylonian approximation method is established.

The following table depicts the first 5 iterates of the Babylonian Method for $\sqrt{2}$ starting with $q_0 = 1$:

n	q_n	δ_n
0	1	$-2.92893218 \cdot 10^{-1}$
1	$\frac{3}{2} = 1.5$	$6.06601778 \cdot 10^{-2}$
2	$\frac{17}{12} = 1.41\overline{6}\ldots$	$1.73460668 \cdot 10^{-3}$
3	$\frac{577}{408} = 1.414\overline{2156862745098039}\ldots$	$1.50182509 \cdot 10^{-6}$
4	$\frac{665857}{470832} \approx 1.41421356237468991062629{6}$	$1.12773761 \cdot 10^{-12}$
5	$\frac{886731088897}{627013566048} \approx 1.414213562373095048802$	$6.35896059 \cdot 10^{-25}$

Exercise

5.4.1. A triangle with side lengths that form three consecutive terms in a geometric sequence exists if and only if the ratio q of the geometric sequence satisfies $1/\tau < q < \tau$, where τ is the golden number.

5.5 Lines and Circles

We now return to our Cartesian distance d and study its invariance properties under some transformations (self-maps) of the plane $\mathbb{R}^2$.

For $W \in \mathbb{R}^2$, we define the **translation** by W as the map $T_W : \mathbb{R}^2 \to \mathbb{R}^2$ given by $T_W(P) = P + W$, $P \in \mathbb{R}^2$. In coordinates, if $W = (u, v) \in \mathbb{R}^2$, then we have

$$T_W(P) = P + W = (x + u, y + v), \quad P = (x, y) \in \mathbb{R}^2.$$

The Cartesian distance d is invariant (unchanged) under translations; that is, for $W \in \mathbb{R}^2$, we have

$$d(T_W(P_0), T_W(P_1)) = d(P_0, P_1), \quad P_0, P_1 \in \mathbb{R}^2.$$

Indeed, this is obvious since, in the definition of the Cartesian distance, we take differences of the respective coordinates of P_0 and P_1, and thus the coordinates u and v of W cancel.

Another type of transformation of the plane $\mathbb{R}^2$ that we will utilize is the (positive) **quarter-turn**. We define the (positive) quarter-turn $S_0 : \mathbb{R}^2 \to \mathbb{R}^2$ about the origin by $S_0(P) = S_0(x, y) = (-y, x)$, $P = (x, y) \in \mathbb{R}^2$. Its square $S_0^2 = S_0 \circ S_0 : \mathbb{R}^2 \to \mathbb{R}^2$ (by composition) is the negative of the identity $-\mathrm{id}_{\mathbb{R}^2}$, the **half-turn** about the origin, given by $S_0^2(x, y) = (-x, -y)$, $P = (x, y) \in \mathbb{R}^2$. The (positive) quarter-turn about any point $O \in \mathbb{R}^2$ is defined by the composition $S_O = T_O \circ S_0 \circ T_{-O}$. Once again, the square S_O^2 is the half-turn about the point O. We call O the **center** of S_O.

Once again, it follows easily that the Cartesian distance d is invariant under a quarter-turn about any center.

These transformations are **affine** in the sense that they map lines to lines preserving the respective affine coordinate functions. (This follows immediately from the strict triangle inequality or by direct computation.) In particular, for $P_0, P_1 \in \mathbb{R}^2$, we have $T_W([P_0, P_1]) = [T_W(P_0), T_W(P_1)]$, $W \in \mathbb{R}^2$, and $S_O([P_0, P_1]) = [S_O(P_0), S_O(P_1)]$, $O \in \mathbb{R}^2$. In addition, these transformations preserve the relation being parallel, that is, they send pencils of parallel lines to pencils of parallel lines.

A translation sends a line to a **parallel** line. Indeed, if a line ℓ is given by the equation $ax - by = c$, $a^2 + b^2 > 0$, then a translation T_W with $W = (u, v) \in \mathbb{R}^2$ sends ℓ into a line with equation $a(x - u) - b(y - v) = c$, that is, $ax - by = c + au - bv$.

A half-turn sends a line to a **parallel** line. Indeed, since translations do the same (by the above), it is enough to show this for the half-turn about the origin, the negative of the identity map. Now, if a line ℓ is given by the equation $ax - by = c$, $a^2 + b^2 > 0$, then the half-turn about the origin sends ℓ to the line with equation $-ax + by = c$, that is, $ax - by = -c$.

Finally note that these transformations preserve ω, and therefore they are **orientation preserving** in the sense that if (P_0, P_1, P_2) is a positively oriented triple, then so are the corresponding transformed triples.

Remark The adjective "positive" for the quarter-turn S_O about a point $O \in \mathbb{R}^2$ is due to the fact that, for any $P \in \mathbb{R}^2$, $P \neq O$, the triple $(O, P, S_O(P))$ is **positively** oriented. Indeed, since translations are orientation preserving, it is enough to show this for the quarter-turn about the origin. Now, if $P = (x, y)$, $x^2 + y^2 > 0$, then we have $\omega((0, 0), (x, y), (-y, x)) = x^2 + y^2 > 0$.

We say that two lines are **perpendicular** if one is obtained from the other by a quarter-turn. As simple computation shows, if a line ℓ is given by the equation $ax - by = c$, $a^2 + b^2 > 0$, then, for $O = (u, v) \in \mathbb{R}^2$, the transformed perpendicular line $S_O(\ell)$ is given by the equation $bx + ay = c + a(v - u) + b(v + u)$.

On the other hand, as noted above, a pencil of parallel lines is given by the equation $ax - by = c$, where the constant $c \in \mathbb{R}$ varies over all real numbers. It follows that the relation of being perpendicular depends only on the pencils of parallel lines that each of the two perpendicular lines is participating in. In other words, if two lines are perpendicular, then so are any two lines in the respective pencils of parallel lines.

Since every pencil of parallel lines contains a unique representative through the origin, and $S_0^2 = -\,\mathrm{id}$, it also follows that the relation being perpendicular is symmetric.

Finally, if two lines ℓ and ℓ' are intersected by another line perpendicular to both, then ℓ and ℓ' are parallel.

Example 5.5.1 (Perpendicular Bisector) Determine the set of points that are **equidistant** from two given distinct points P_0 and P_1 on the plane $\mathbb{R}^2$.

I. Algebraic Solution. Let $P_0 = (x_0, y_0)$ and $P_1 = (x_1, y_1)$, $P_0 \neq P_1$. A variable point $P = (x, y)$ is equidistant from P_0 and P_1 if and only if $d(P, P_0) = d(P, P_1)$. Using the distance formula, after squaring, we calculate

$$\begin{aligned}
(x - x_0)^2 + (y - y_0)^2 &= (x - x_1)^2 + (y - y_1)^2 \\
2(x_0 - x_1)x + 2(y_0 - y_1)y &= x_0^2 - x_1^2 + y_0^2 - y_1^2 \\
2(x_0 - x_1)x + 2(y_0 - y_1)y &= (x_0 - x_1)(x_0 + x_1) + (y_0 - y_1)(y_0 + y_1).
\end{aligned}$$

A final rearrangement and grouping the multiples of $(x_0 - x_1)$ and $(y_0 - y_1)$ give the symmetric form

$$(x_1 - x_0)\left(x - \frac{x_0 + x_1}{2}\right) + (y_1 - y_0)\left(y - \frac{y_0 + y_1}{2}\right) = 0.$$

First, this equation is linear (since $(x_0 - x_1)^2 + (y_0 - y_1)^2 > 0$), and hence it must represent a line. Second, it is also clear (by substituting $x = (x_0 + x_1)/2$ and $y = (y_0 + y_1)/2$) that this line contains the **midpoint** M of the line segment $[P_0, P_1]$ (Example 5.3.2) since

$$M = \left(\frac{x_0 + x_1}{2}, \frac{y_0 + y_1}{2} \right).$$

Finally, as shown above, an equation of the line containing the points $P_0 = (x_0, y_0)$ and $P_1 = (x_1, y_1)$ is given by

$$(y_1 - y_0)x - (x_1 - x_0)y = x_0 y_1 - x_1 y_0.$$

Comparing the equations of these lines, we see that they are perpendicular.

We conclude that the set of points equidistant from two distinct points P_0 and P_1 is the **perpendicular bisector** of the line segment $[P_0, P_1]$ that contains their **midpoint** M. As a byproduct, we also recover the **midpoint formula** for M above (Example 5.3.2 again).

II. Geometric Solution After Euclid.[16] Let P be a point such that $d(P, P_0) = d(P, P_1)$. Then the triangle $\triangle[P, P_0, P_1]$ is isosceles. Thus, by the Theorem on Isosceles Triangles in Euclid's *Elements* (Book I, Proposition 5), the angles at P_0 and P_1 are congruent. Consider the midpoint M of the line segment $[P_0, P_1]$. Then the sub-triangles $\triangle[P, P_0, M]$ and $\triangle[P, P_1, M]$ are congruent.[17] Thus, these sub-triangles must have right angle at M. We obtain that P is on the perpendicular bisector of the line segment $[P_0, P_1]$.

History

In the discussion above we used the **Theorem on Isosceles Triangles**: The angles opposite to the equal sides of an isosceles triangle are equal. It is also called the **pons asinorum**, "bridge of donkeys" in Latin. It is either a somewhat derogatory phrase pointing to and challenging the reader to tackle this first non-trivial proposition in the *Elements* and pass this bridge to get to harder ones, or the isosceles triangle depicts an actual pointed bridge that only a brave and sure-footed donkey can pass.

We now introduce another fundamental concept of Greek geometry, the concept of a circle. Given a point $O \in \mathbb{R}^2$ and a positive real number $0 < r \in \mathbb{R}$, we define the **circle** of radius r and center at O as the set[18]

$$\mathbb{S}_{O,r} = \{P \in \mathbb{R}^2 \mid d(P, O) = r\}.$$

Letting $O = (u, v)$ and $P = (x, y)$, the Cartesian distance formula gives the equation of the circle $\mathbb{S}_{O,r}$ as

$$(x - u)^2 + (y - v)^2 = r^2,$$

---

[16] This geometric solution can be reworded to become a simple consequence of Birkhoff's Postulates of Angle Measure and Similarity. The validity of these postulates in our model will be proved in Section 5.7. Hence, for a change, we give here a proof based on Euclid's *Elements*.

[17] Warning: Congruence of the triangles $\triangle[P, P_0, M]$ and $\triangle[P, P_1, M]$ also follows from the observation that the lengths of the three pairs of sides of these triangles are equal, but in the *Elements*, this occurs after Proposition 5 of Book I.

[18] Compare this with primitive #15 in the *Elements* as stated at the beginning of this chapter.

where we squared both sides to eliminate the square root. If the center of the circle is the origin 0, then we write $\mathbb{S}_r = \mathbb{S}_{0,r}$. Similarly, if the circle has a unit radius ($r = 1$), then we write $\mathbb{S}_O = \mathbb{S}_{O,1}$. Finally, $\mathbb{S}$ denotes the unit radius circle with center at the origin.

As a simple application of the perpendicular bisector and the concept of a circle, consider three distinct points $A, B, C \in \mathbb{R}^2$ on the plane. We ask the following question: Is there any point on the plane which is equidistant from all these three points?

If A, B, C are **collinear**, then the answer is no. Indeed, the set of points equidistant from A and B is the perpendicular bisector of the line segment $[A, B]$, and the set of points equidistant from B and C is the perpendicular bisector of $[B, C]$. These bisectors are distinct and perpendicular to the line that passes through A, B, C. Thus, they are parallel and have no common intersection.

Assume now that A, B, C are **not collinear**. Consider the perpendicular bisectors ℓ_a, ℓ_b, ℓ_c of the line segments $[B, C], [C, A], [A, B]$. By the previous step, each pair of bisectors intersects in a point. We claim that these (three) points are the same. Indeed if O is the common intersection of ℓ_a and ℓ_b, then $d(O, B) = d(O, C)$ and $d(O, C) = d(O, A)$. Therefore, $d(O, A) = d(O, B)$ so that O is equidistant from A and B, and hence it is on the bisector ℓ_c. We obtain that if A, B, C are not collinear, then there is a **unique** point equidistant from all these three points.

This conclusion can be put into a familiar framework if we consider the points A, B, C, the vertices of the triangle $\triangle[A, B, C]$. Since $d(O, A) = d(O, B) = d(O, C) = R$, say, we see that the circle with center O and radius R contains the points A, B, C. This is the unique circle **circumscribed** about the triangle. We call this the **circumcircle** and its radius the **circumradius** R of the triangle.

We now return to the main line and study the possible configurations of a circle and a line. We claim that there are three possibilities; namely, the circle and the line may be disjoint, meet at one point, or meet at two points. When a circle meets a line at exactly one point, we say that the line is **tangent** to the circle. A line is **secant** to a circle if they intersect at exactly two points.[19]

Discarding the case when the circle and the line are disjoint, we assume that the circle $\mathbb{S}_{O,r}$ and the line ℓ intersect at least in one point $P_0 \in \mathbb{S}_{O,r} \cap \ell$. For simplicity, we translate the entire configuration such that the center of the circle is at the origin. The equation of the circle $\mathbb{S}_r$ above reduces to

$$x^2 + y^2 = r^2.$$

We let $P_0 = (x_0, y_0)$, so that $x_0^2 + y_0^2 = r^2$. As usual, we write the equation of the line ℓ through P_0 as

$$ax - by = ax_0 - by_0, \quad a^2 + b^2 > 0.$$

---

[19] The words "tangent" and "secant" are derived from "tangere" and "secare," respectively, which in Latin mean "to touch" and "to cut."

After simplification and factoring, we obtain that any intersection point $P = (x, y) \in \mathbb{S}_r \cap \ell$ satisfies the system of equations

$$(x - x_0)(x + x_0) + (y - y_0)(y + y_0) = 0$$
$$a(x - x_0) = b(y - y_0).$$

Assuming $P \neq P_0$, that is, discarding the solution $x = x_0$ and $y = y_0$, we obtain $b(x + x_0) + a(y + y_0) = 0$. A simple computation now gives

$$x = x_0 - 2b\frac{bx_0 + ay_0}{a^2 + b^2} \quad \text{and} \quad y = y_0 - 2a\frac{bx_0 + ay_0}{a^2 + b^2}.$$

This is a solution different from $x = x_0$ and $y = y_0$ (which has been discarded) if and only if $bx_0 + ay_0 \neq 0$.

Turning the question around, we see that the circle $\mathbb{S}_r$ and the line ℓ have a **unique** intersection point $P_0 = (x_0, y_0)$ if and only if $bx_0 + ay_0 = 0$. By the discussion at the beginning of this section, $bx + ay = 0$ is an equation of a line **perpendicular** to the tangent line ℓ, and the latter has the equation $ax - by = ax_0 - by_0$.

As a final note, we claim that the entire tangent line ℓ with the exception of the point of tangency P_0 lies in the **exterior** of the circle $\mathbb{S}_r$.

Indeed, letting $P_t = (x_0 + tb, y_0 + at)$, $t \in \mathbb{R}$, the equation of the tangent line ℓ in the form $a(x - x_0) = b(y - y_0)$ above clearly shows that $\ell = \{P_t \mid t \in \mathbb{R}\}$. We now calculate the distance as

$$\begin{aligned} d(P_t, P_t)^2 &= (x_0 + tb)^2 + (y_0 + ta)^2 = x_0^2 + y_0^2 + 2t(bx_0 + ay_0) + t^2(a^2 + b^2) \\ &= r^2 + t^2(a^2 + b^2) \geq r^2 \end{aligned}$$

with equality if and only if $t = 0$, that is, at the point of tangency P_0. The claim follows.

Summarizing, we obtain that through any point P_0 of a circle $\mathbb{S}_{O,r}$ (with center O), there is a unique line ℓ that is **tangent** to $\mathbb{S}_{O,r}$, and it is characterized by the property that it is perpendicular to the radial line containing P_0 and the center O of the circle. Any other line through P_0 is a secant to $\mathbb{S}_{O,r}$, that is, it intersects the circle in two distinct points. Finally, the entire tangent line lies in the exterior of the circle $\mathbb{S}_{O,r}$, except the point of tangency P_0.

History

The ancient Greek mathematicians elevated the study of geometric configurations to the discipline of **Geometry**. These include lines, polygons, circles, parabolas, ellipses, hyperbolas, their metric properties and mutual relationships, such as tangents, secants, intersections, etc. As noted previously, the ancient Greeks also created **Geometric Algebra**, which associated algebraic terms with geometric objects, such as length, perimeter, area, etc. **Algebra** as a discipline separate from Geometry (and Arithmetic) was established by Muhammad ibn Mūsā al-Khwārizmī. To a large extent it was an early theory of equations that studied solutions of linear and quadratic equations. This theory had different developments by Diophantus of Alexandria and the Indian mathematician Brahmagupta. Throughout the Middle Ages Arabic scholars raised this discipline to new heights. The modern symbolic representation of variables and constants, introduced by the French mathematician François Viète (Latin Vieta) (1540–1603) and subsequently brought to perfection by Descartes, put algebra on a solid foundation. But it was the introduction and systematic use of coordinate systems that revolutionized mathematics by establishing a bridge between geometry and algebra.

As an application, we introduce the concept of **distance of a point from a line**. Let O be a point and ℓ a line. By definition, the distance of O from ℓ is

$$d(O, \ell) = \inf\{d(O, P) \mid P \in \ell\}.$$

If $O \in \ell$, then $d(O, \ell) = 0$. We may therefore assume that $O \notin \ell$.

We claim that there is a unique circle with center at O and with ℓ as a tangent line to the circle.

Indeed, by what we proved above, this circle is obtained by taking the line ℓ' through O perpendicular to ℓ, and the radius of the circle is the distance of the intersection point $P_0 = \ell \cap \ell'$ from the center O.

Since ℓ is tangent to the circle, all the points on ℓ except P_0 are in the exterior of the circle. Thus, the radius of the circle realizes the infimum above, and therefore it is the distance $d(O, \ell)$.

To obtain an explicit formula, let ℓ be given by the equation $ax - by = c$, $a^2 + b^2 > 0$, and let $O = (u, v) \in \mathbb{R}^2$. The equation of ℓ' through O and perpendicular to ℓ is given by $bx + ay = bu + av$. Putting these two equations together, a short computation gives the intersection point $P_0 = \ell \cap \ell'$ as

$$P_0 = \left(\frac{b(bu + av) + ac}{a^2 + b^2}, \frac{a(bu + av) - bc}{a^2 + b^2} \right).$$

By a short computation, we arrive at the distance of P_0 from $O = (u, v)$ as

$$d(O, \ell) = \frac{|au - bv - c|}{\sqrt{a^2 + b^2}}.$$

Example 5.5.2 (Angular Bisector) Determine the set of points that are **equidistant** from two given distinct lines ℓ_0 and ℓ_1 on the plane $\mathbb{R}^2$.

If ℓ_0 and ℓ_1 are parallel, then the set of points equidistant from both lines is the parallel line midway between ℓ_0 and ℓ_1.

Assume now that ℓ_0 and ℓ_1 intersect in a point C. The intersecting lines ℓ_0 and ℓ_1 split the plane into four angular sectors. Let P be a point such that $d(P, \ell_0) = d(P, \ell_1)$. We may assume that $P \neq C$ (since C is clearly equidistant from both lines with zero distance) and also that P is not on any of these two lines. Thus, P is contained in one of the open angular sectors. Let $P_0 \in \ell_0$ and $P_1 \in \ell_1$ be such that $d(P, P_0) = d(P, \ell_0) = d(P, \ell_1) = d(P, P_1)$. The two **right** triangles $\triangle[P, P_0, C]$ and $\triangle[P, P_1, C]$ are congruent since they have two equal sides, and they also have a

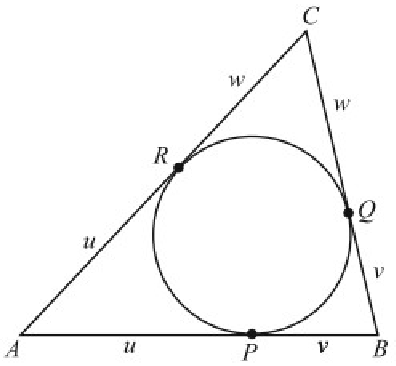

Fig. 5.1 The incircle of a triangle.

common side.[20] Thus their angles at C are equal. We obtain that P is on the bisector of the angular sector that contains P. We conclude that the set of points equidistant from two intersecting lines is the pair of perpendicular angular bisectors.

As an application, consider three lines ℓ_0, ℓ_1, and ℓ_2 that are the extensions of the sides of a (non-degenerate) triangle $\triangle[P_0, P_1, P_2]$. This triangle is the intersection of three angular sectors, one from each vertex (see Figure 5.1). The bisectors corresponding to each angular sector intersect in a point C. This is the unique point equidistant from all the three lines. We obtain that C is the center of the unique **inscribed** circle touching each line at the points where the distance of C and the lines are realized. We call this the **incircle** and its radius the **inradius** of the triangle.

Remark If $\triangle[A, B, C]$ is a (non-degenerate) triangle, then its incircle touches each side of the triangle at a specific point, $P \in [A, B]$, $Q \in [B, C]$, $R \in [C, A]$, say. Clearly, we have $d(A, P) = d(A, R)$, $d(B, P) = d(B, Q)$, and $d(C, Q) = d(C, R)$. Denoting these distances by u, v, and w, we obtain the substitution

$$a = v + w, \quad b = w + u, \quad c = u + v.$$

This gives the geometric interpretation of the substitution at the beginning of the previous section.

We now turn to study **secant** lines. Let $P_0, P_1 \in \mathbb{S}_{O,r}$ be two distinct points on the circle. Once again we translate the entire configuration so that the center of the circle $\mathbb{S}_r$ is at the origin, and thereby it has the equation $x^2 + y^2 = r^2$. By the strict triangle inequality, we have $d(P_0, P_1) \leq d(0, P_0) + d(0, P_1) = 2r$ with equality if and only if $0 \in [P_0, P_1]$.

---

[20] In this example we assume Birkhoff's Postulates of Angle Measure and Similarity and, consequently, the Pythagorean Theorem, whose validity, in our model, will be proved in Section 5.7. This is pedagogically justified since this example is a perfect fit for our present line of argument. Alternatively, one can also refer here to Euclid's *Elements*.

Letting $P_0 = (x_0, y_0)$, $x_0^2 + y_0^2 = r^2$, and $P_1 = (x_1, y_1)$, $x_1^2 + y_1^2 = r^2$, we calculate

$$\begin{aligned} d(P_0, P_1)^2 = (x_0 - x_1)^2 + (y_0 - y_1)^2 &= x_0^2 + y_0^2 + x_1^2 + y_1^2 - 2x_0x_1 - 2y_0y_1 \\ &= 2r^2 - 2(x_0x_1 + y_0y_1). \end{aligned}$$

Now recall the affine coordinate function that parametrizes the line through P_0 and P_1 under which the point $P_t = (1-t)P_0 + tP_1$ corresponds to the parameter $t \in \mathbb{R}$.

Using the result of the previous computation, we have

$$\begin{aligned} d(0, P_t)^2 &= ((1-t)x_0 + tx_1)^2 + ((1-t)y_0 + ty_1)^2 \\ &= (1-t)^2r^2 + t^2r^2 + 2t(1-t)(x_0x_1 - y_0y_1) \\ &= (1-t)^2r^2 + t^2r^2 + t(1-t)(2r^2 - d(P_0, P_1)^2) \\ &= r^2 - t(1-t)d(P_0, P_1)^2. \end{aligned}$$

In particular, we see that, for $t \in [0, 1]$, we have $d(0, P_t) \leq r$ with equality if and only if $t = 0, 1$.

Summarizing, given two distinct points P_0 and P_1 on a circle $\mathbb{S}_{O,r}$, for $t \in [0, 1]$, we have $d(P_t, O) \leq r$ (with equality if and only if $t = 0, 1$), that is, the line segment $[P_0, P_1]$ is contained in the interior of the circle $\mathbb{S}_{O,r}$. Similarly, for $t \notin [0, 1]$, we have $d(P_t, O) > r$.

Exercises

5.5.1. Calculate the length of the hypotenuse of the right triangle $\triangle[A, B, C]$ with right angle at C, where $d(A, C) = d(A, M) = 1$ and M is the midpoint of the hypotenuse $[A, B]$.

5.5.2. Use the pons asinorum to prove **Thales' Theorem**: If A, B, C are distinct points on a circle $\mathbb{S}_{O,r}$ and $O \in [A, B]$ (that is, $[A, B]$ is a diameter), then $\angle ACB$ is a right angle. Generalize this to the case when $[A, B]$ is a chord of the circle, $O \notin [A, B]$; the **Central Angle Theorem**: If C is on the longer circular arc of $\mathbb{S}_{O,r}$ with end-points A, B, then $\mu(\angle AOB) = 2\mu(\angle ACB)$; and if C is on the shorter circular arc of $\mathbb{S}_{O,r}$ with end-points A, B, then $\mu(\angle AOB) = 2(\pi - \mu(\angle BCA))$.

5.5.3. The **power** $\mathfrak{p}_S(P)$ of a point $P \in \mathbb{R}^2$ with respect to a circle $S = \mathbb{S}_{O,r}$ is defined as $\mathfrak{p}_S(P) = d(P, O)^2 - r^2$. (Note that the power is zero for points on the circle, negative for points inside the circle, and positive for points outside the circle.) (a) Prove the **Intersecting Chords Theorem**: Let P be outside S. Show that any line through P that meets S in the points $A, B \in S$;

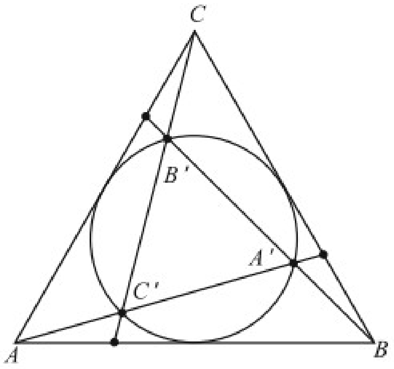

Fig. 5.2 An occurrence of the golden number τ.

we have $\mathfrak{p}_S(P) = d(P, A) \cdot d(P, B)$. (Note that, for a line tangent to S, the point of tangency is $A = B$ so that we have $\mathfrak{p}_S(P) = d(P, A)^2$, and the circle with center at P and radius $d(P, A)$ is orthogonal to S.) (b) Extend (a) to the case when P is inside S. (c) Let $S_1 = \mathbb{S}_{O_1,r_1}$ and $S_2 = \mathbb{S}_{O_2,r_2}$ be two disjoint circles. Show that the set of points in $P \in \mathbb{R}^2$ that have the same power with respect to S_1 and S_2 is a line, the so-called **radical line**. (d) Generalize the radical line to the case of intersecting circles or two circles with one inside the other. (e) Prove **Monge's Theorem**: Given three disjoint circles (with non-parallel radical axes), there is a circle orthogonal to all three.

5.5.4. Let $\triangle[A, B, C]$ be an equilateral triangle and S the incircle. Let $\triangle[A', B', C']$ be an equilateral triangle inscribed in S, that is, $A', B', C' \in S$, such that the line extensions of the sides of this triangle pass through the vertices of $\triangle[A, B, C]$; that is, $A' \in [B, B']$, $B' \in [C, C']$, $C' \in [A, A']$. Show that the ratio $d(A', C')/d(A, C')$ is the golden number τ (see Figure 5.2).

5.5.5. Given a line segment $[A, B]$ on the plane $\mathbb{R}^2$, determine the set of points $C \in \mathbb{R}^2$ such that the non-degenerate triangle $\triangle[A, B, C]$ has obtuse angle at C.

5.5.6. Let A and B be two points on the plane unit distance apart, and $0 < q < 1$, $q \in \mathbb{R}$. Show that the set $\{P \in \mathbb{R}^2 \mid d(P, A) = q \cdot d(P, B)\}$ is a circle, and determine its center and radius (in terms of q).

5.5.7. In a triangle $\triangle[A, B, C]$, let the angular bisector of the (interior) angle at the vertex A intersect the opposite side at the point $D \in [B, C]$. Prove the **Angle Bisector Theorem** $d(A, B)/d(A, C) = d(B, D)/d(C, D)$.

5.5.8. Consider two parallel chords of a circle S with lengths a and b which are d distance apart.[21] Let a third parallel chord of length c be in the midway of the first two. Express c in terms of a, b, d.

---

[21] Generalization of a problem in the American High School Mathematics Examination, 1995.

5.5.9. Three circles of common radius $0 < r \in \mathbb{R}$ are mutually and externally tangent. (a) If they are also internally tangent to a larger circle of radius $0 < R \in \mathbb{R}$, then find R/r. (b) Find the perimeter of the triangle whose sides are tangent to each pair of the three circles.

5.5.10. The vertices of a square of side length 2 are the centers of four circles of radius 1. Find the radius of the smaller circle externally tangent to these four circles whose center is the center of the square.

5.5.11. Two tangents are drawn from a point A to a circle of radius $0 < r \in \mathbb{R}$ and center O. Another tangent to the circle meets the two tangent lines at points B and C such that the triangle $\triangle[A, B, C]$ is disjoint from the circle. Find the perimeter of the triangle in terms of $d(A, O)$ and r.

5.5.12. A circle touches all four sides of an isosceles trapezoid. Find the radius r of the circle in terms of the parallel side lengths (bases) a and c of the trapezoid.

5.6 Arc Length on the Unit Circle

In this section we make a detailed and rigorous study of the **arc length** of circular arcs of a **unit** radius circle $\mathbb{S}_O$ with center $O \in \mathbb{R}^2$. The material presented here is technically demanding, and the readers who have only marginal interest in axiomatic developments may skip it. The results of this section will only be used for the establishment of the Birkhoff angle measure in the next section and for the proof of the Law of Cosines in trigonometry discussed in the last chapter of this book.

Let $P_0, P_1 \in \mathbb{S}_O$ be distinct points. The secant line through P_0 and P_1 divides the circle $\mathbb{S}_O$ into two **circular arcs** with end-points P_0 and P_1. These two circular arcs can be obtained from the equation of the line through the points $P_0 = (x_0, y_0)$ and $P_1 = (x_1, y_1)$ by intersecting the circle $\mathbb{S}_O$ with the two half-planes

$$(y_1 - y_0)x - (x_1 - x_0)y \gtreqless x_0 y_1 - x_1 y_0$$

defined by the equation of the (common boundary) line.

If $d(P_0, P_1) = 2$, then, by the strict triangle inequality, $O \in [P_0, P_1]$, and the two circular arcs are called **semi-circles**.[22] They are congruent via the half-turn about the center O. A line through O and perpendicular to the line extension of $[P_0, P_1]$ splits these two semi-circles into four **quarter-circles**. These quarter-circles are permuted (cyclically) by the quarter-turn S_O.

Assume now that $d(P_0, P_1) < 2$ so that, by the strict triangle inequality, $O \notin [P_0, P_1]$. Unless stated otherwise, we will always denote by $\mathcal{C} \subset \mathbb{S}_O$ the circular arc,[23] which is in the opposite side of the line extension of $[P_0, P_1]$ to the center O.

---

[22] Compare this with definition #18 in the *Elements* at the beginning of this chapter.

[23] As we will see below, $\mathcal{C}$ is the **shorter** (arc length) circular arc with end-points P_0 and P_1.

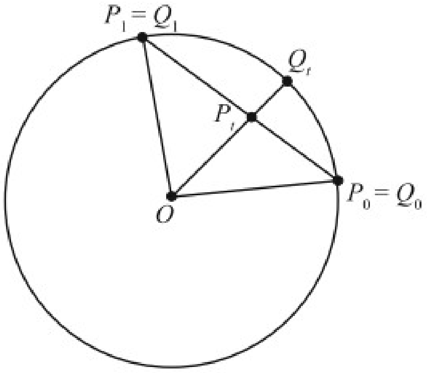

Fig. 5.3 Parametrization of a circular arc.

The other circular arc[24] will be denoted by $\mathcal{C}^c$ and will be called the **complement** of $\mathcal{C}$. For uniformity, for $d(P_0, P_1) = 2$, $\mathcal{C}$ (and $\mathcal{C}^c$) will denote either of the semicircles with end-points P_0 and P_1.

Once again, recall the affine coordinate function that parametrizes the line through P_0 and P_1 under which the point $P_t = (1-t)P_0 + tP_1$ corresponds to the parameter $t \in \mathbb{R}$.

By the computations at the end of the previous section, we have

$$d(O, P_t)^2 = 1 - t(1-t)d(P_0, P_1)^2.$$

For $t \in [0, 1]$, we define[25]

$$Q_t = \frac{1}{d(O, P_t)} P_t + \left(1 - \frac{1}{d(O, P_t)}\right) O, \quad Q_0 = P_0, \ Q_1 = P_1.$$

Clearly, we have $d(O, Q_t) = d(O, P_t)/d(O, P_t) = 1$, or equivalently, $Q_t \in \mathbb{S}_O$, $t \in [0, 1]$ (see Figure 5.3). By the first formula defining the half-planes above, we see that the points Q_t, $t \in [0, 1]$, are in the half-plane that does not contain the point O. This gives

$$\mathcal{C} = \{Q_t \,|\, 0 \leq t \leq 1\}.$$

(If $Q \in \mathcal{C}$, then $[O, Q]$ and $[P_0, P_1]$ must intersect in a point P_t, $t \in [0, 1]$, say, so that $Q = Q_t$ holds.)

With this preparation, we now define the arc length of $\mathcal{C}$.

A **partition** of the interval $[0, 1]$ is a finite strictly increasing sequence:

$$(t_0, t_1, \ldots, t_{n-1}, t_n), \quad 0 = t_0 < t_1 < \cdots < t_{n-1} < t_n = 1, \quad n \in \mathbb{N}.$$

---

[24] The circular arc $\mathcal{C}^c$ is **not** the set-theoretic complement of $\mathcal{C}$ with respect to the whole circle $\mathbb{S}_O$ because $\mathcal{C}$ and $\mathcal{C}^c$ overlap in the two common end-points P_0 and P_1.

[25] Geometrically, the point $Q_t \in \mathcal{C}$ is obtained from P_t by radial projection from the center O.

We denote by Π the set of all partitions of $[0, 1]$ (for all $n \in \mathbb{N}$). A partition $(t_0, t_1, \ldots, t_{n-1}, t_n) \in \Pi$ of $[0, 1]$ defines an open **polygonal path** of $\mathcal{C}$ connecting $Q_0(= P_0)$ and $Q_1(= P_1)$ with consecutive vertices $Q_0 = Q_{t_0}, Q_{t_1}, \ldots, Q_{t_{n-1}}, Q_{t_n} = Q_1 \in \mathcal{C}$ with union $\bigcup_{i=1}^n [Q_{t_{i-1}}, Q_{t_i}]$. The **length** of a polygonal path is $\sum_{i=1}^n d(Q_{t_{i-1}}, Q_{t_i})$, the sum of the lengths of the participating line segments.

We now define the arc length of the circular arc $\mathcal{C} \subset \mathbb{S}_O$ by

$$\mathcal{L}_{\mathcal{C}} = \sup \left\{ \sum_{i=1}^n d(Q_{t_{i-1}}, Q_{t_i}) \,\middle|\, (t_0, t_1, \ldots, t_{n-1}, t_n) \in \Pi \right\}.$$

We need to show that the arc length $\mathcal{L}_{\mathcal{C}}$ is a (finite) real number; that is, the set of lengths of all polygonal paths of $\mathcal{C}$ is bounded above.

Let ℓ_0, resp. ℓ_1, denote the half-line with end-point O and containing P_0, resp. P_1 (For the notations introduced here and below, refer to Figure 5.4.). Let m_0, resp. m_1, be the tangent line to $\mathcal{C}$ through the point $Q_0 = P_0$, resp. $Q_1 = P_1$. By the results of the previous section, m_0, resp. m_1, is perpendicular to the line extension of ℓ_0, resp. ℓ_1. The lines m_0 and m_1 cannot be parallel since $O \notin [P_0, P_1]$. Let M be the intersection point of m_0 and m_1.

We claim that

$$\mathcal{L}_{\mathcal{C}} \le d(Q_0, M) + d(Q_1, M).$$

Consider a polygonal path $\bigcup_{i=1}^n [Q_{t_{i-1}}, Q_{t_i}]$ with vertices $Q_0 = Q_{t_0}, Q_{t_1}, \ldots, Q_{t_{n-1}}, Q_{t_n} = Q_1 \in \mathcal{C}$ corresponding to a partition $(t_0, t_1, \ldots, t_{n-1}, t_n) \in \Pi$ of $[0, 1]$ (see again Figure 5.4). For $i = 0, 1, 2, \ldots, n$, let h_i, resp. k_i, be the line through Q_{t_i} and parallel to m_1, resp. m_0. Let h_i and m_0 meet at the point R_i, $i = 0, \ldots, n$; in particular, $R_0 = Q_0$ and $R_n = M$. Similarly, let k_i and m_1 meet at the point S_i, $i = 0, \ldots, n$; in particular, $S_0 = M$ and $S_n = Q_1$. Finally, for $i = 1, \ldots, n$, let h_i and k_{i-1} meet at the point T_i.

With these notations, by the triangle inequality, we have

$$d(Q_{t_{i-1}}, Q_{t_i}) \le d(Q_{t_{i-1}}, T_i) + d(Q_{t_i}, T_i), \quad i = 1, \ldots, n.$$

On the other hand, $d(Q_{t_{i-1}}, T_i) = d(R_{i-1}, R_i)$ and $d(Q_{t_i}, T_i) = d(S_{i-1}, S_i)$, $i = 1, \ldots, n$, as the respective points are vertices of parallelograms.[26]

The next lemma will imply that $R_{i-1} * R_i * R_{i+1}$ and $S_{i-1} * S_i * S_{i+1}$, $i = 1, \ldots, n-1$. Once this is proved, it will follow that $\sum_{i=1}^n d(R_{i-1}, R_i) = d(R_0, R_n) = d(Q_0, M)$ and $\sum_{i=1}^n d(S_{i-1}, S_i) = d(S_0, S_n) = d(Q_1, M)$, so that the stated upper bound above holds, and the supremum defining the arc length is finite.

---

[26] The opposite sides of a parallelogram have equal lengths. This follows from translation invariance of the distance as shown at the beginning of Section 5.5.

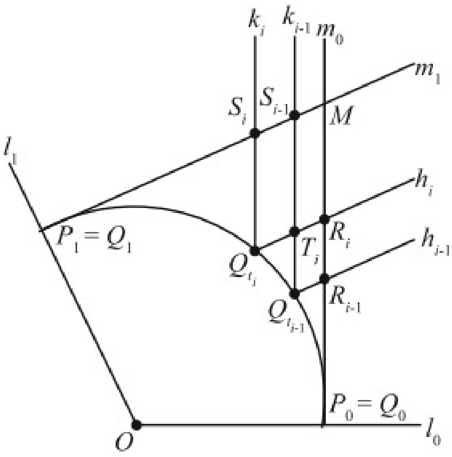

Fig. 5.4 Upper bound for the arc length.

In the lemma, without loss of generality, we assume that the center of the ambient circle is at the origin.

Lemma *Let* $\mathbb{S}$ *be the unit circle with center at the origin and* $P_0, P_1 \in \mathbb{S}$ *two distinct points with* $d(P_0, P_1) < 2$. *For* $t \in [0, 1]$, *let* $Q_t = P_t/d(0, P_t) \in \mathbb{S}$, $P_t = (1-t)P_0 + tP_1$; *and let* sP_0, $s \in \mathbb{R}$, *be the intersection of the radial line extension of* $[0, P_0]$ *and the line through* Q_t *perpendicular to this radial line. Then we have*

$$s = \frac{2 - td^2}{2\sqrt{t^2d^2 - td^2 + 1}}, \quad d = d(P_0, P_1).$$

In particular, s, *as a function of* $t \in [0, 1]$, *is strictly decreasing for* $0 \leq t \leq \min(2/d^2, 1)$ *and strictly increasing*[27] *for* $\min(2/d^2, 1) \leq t \leq 1$ *(see Figure 5.5).*

Proof We let $P_0 = (x_0, y_0)$, $x_0^2 + y_0^2 = 1$, and $P_1 = (x_1, y_1)$, $x_1^2 + y_1^2 = 1$. Setting $d = d(P_0, P_1)$, the computations at the end of the previous section give $d^2 = 2 - 2(x_0x_1 - y_0y_1)$ and $d(0, P_t)^2 = t^2d^2 - td^2 + 1$.

Since $d < 2$, for $t \in [0, 1]$, we have

$$d(0, P_t)^2 = t^2d^2 - td^2 + 1 = (t - 1/2)^2d^2 + 1 - d^2/4 > 0.$$

This gives

$$d(0, P_t) = \sqrt{t^2d^2 - td^2 + 1}, \quad t \in [0, 1].$$

---

[27] Monotonicity changes only if $d^2 > 2$, and then it does across $s = 0$, that is, when the sign of s changes from positive to negative; $s = 0$ corresponds to Q_{2/d^2} and P_{2/d^2} being perpendicular to P_0.

Fig. 5.5 Illustration for the monotonicity lemma.

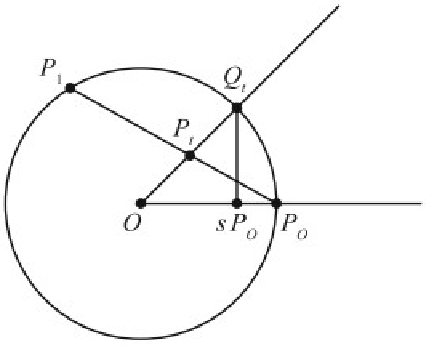

The equation of the line extension of the radial line segment $[0, P_0]$ is of the form $y_0x - x_0y = 0$. The pencil of parallel lines perpendicular to this are described by the equations $x_0x + y_0y = c$, where $c \in \mathbb{R}$. Now, the value of the constant c is determined by the constraint that the perpendicular line must pass through the point $Q_t = P_t/d(0, P_t)$. This gives

$$\begin{aligned} c &= \frac{x_0((1-t)x_0 + tx_1) + y_0((1-t)y_0 + ty_1)}{\sqrt{t^2d^2 - td^2 + 1}} = \frac{1 - t + t(x_0x_1 + y_0y_1)}{\sqrt{t^2d^2 - td^2 + 1}} \\ &= \frac{1 - t + t(1 - d^2/2)}{\sqrt{t^2d^2 - td^2 + 1}} = \frac{2 - td^2}{2\sqrt{t^2d^2 - td^2 + 1}}. \end{aligned}$$

On the other hand, by definition, the point $sP_0 = (sx_0, sy_0)$ must be contained in this perpendicular line. This gives $s = s(x_0^2 + y_0^2) = c$. Putting everything together, we obtain

$$s = \frac{2 - td^2}{2\sqrt{t^2d^2 - td^2 + 1}}.$$

It remains to show the last statement of the lemma, the monotonicity properties of s with respect to t.

First, let $0 \le t \le \min(2/d^2, 1)$. Since in this range $s \ge 0$, it is enough to show that s^2 is strictly decreasing. Now, a simple computation gives

$$s^2 = \frac{d^2}{4} + \left(1 - \frac{d^2}{4}\right) \frac{1 - td^2}{t^2d^2 - td^2 + 1}.$$

Since $d < 2$, we need to show that, for $0 \le t < t' \le \min(2/d^2, 1)$, we have

$$\frac{1 - td^2}{t^2d^2 - td^2 + 1} > \frac{1 - t'd^2}{t'^2d^2 - t'd^2 + 1}.$$

Eliminating the denominators, simplifying and factoring, this becomes

$$(t' - t)(td^2 + t'd^2 - (td^2)(t'd^2)) > 0.$$

This, however, clearly follows[28] since $0 \leq td^2 < t'd^2 \leq 2$. The claimed monotonicity is proved.

Second, for $\min(2/d^2, 1) \leq t \leq 1$, we have $s \leq 0$. Squaring again, the same argument gives the opposite monotonicity property. The lemma follows.

To establish the existence of (an upper bound of) the arc length for a circular arc $\mathcal{C} \subset \mathbb{S}_O$, we followed a **geometric method**. We now describe another essentially **analytic method** to obtain the same result.

First, we derive a geometric formula that will be used several times in the future. Consider a (non-degenerate) triangle $\triangle[A, B, C]$ with (non-collinear) vertices A, B, C. We let

$$A_0 = \frac{1}{d(C, A)} A + \left(1 - \frac{1}{d(C, A)}\right) C$$
$$B_0 = \frac{1}{d(C, B)} B + \left(1 - \frac{1}{d(C, B)}\right) C$$

In other words, A_0, and respectively B_0, is the point at unit distance from C on the half-line with end-point C and containing A, and respectively B. We then have the following important formula:[29]

$$d(A_0, B_0)^2 = 2 + \frac{d(A, B)^2 - d(C, A)^2 - d(C, B)^2}{d(C, A)d(C, B)}.$$

Indeed, using $A_0 - C = (A - C)/d(C, A)$ and $B_0 - C = (B - C)/d(C, B)$, and translation invariance of the distance multiple times, we calculate[30]

$$d(A_0, B_0)^2 = d(A_0 - C, B_0 - C)^2 = d\left(\frac{A - C}{d(A, C)}, \frac{B - C}{d(B, C)}\right)^2$$
$$= d\left(\frac{A - C}{d(A, C)} - \frac{B - C}{d(B, C)}, 0\right)^2 = d\left(\frac{A - C}{d(A, C)}, 0\right)^2 + \left(\frac{B - C}{d(B, C)}, 0\right)^2$$

---

[28] $0 < a, b \leq 2$ implies $(a + b)/(ab) = 1/a + 1/b \geq 1/2 + 1/2 = 1$.

[29] In different (non-axiomatic) developments, this formula is equivalent to the so-called Law of Cosines.

[30] It is customary to set $|P| = d(P, 0)$, the distance of a point P from the origin. Algebraically, $|P|^2$ is then the sum of squares of the two coordinates of P. Translation invariance then gives $d(P, Q)^2 = d(P - Q, 0)^2 = |P - Q|^2$. Using the fact that this is a quadratic form in the coordinates of the points P and Q, the computations above become more familiar.

$$+\frac{1}{d(A,C)d(B,C)}\left(d(A-B,0)^2-d(A-C,0)^2-d(B-C,0)^2\right)$$
$$=d(C,A_0)^2+d(C,B_0)^2+\frac{d(A,B)^2-d(C,A)^2-d(C,B)^2}{d(C,A)d(C,B)}.$$

Since $d(C,A_0)=d(C,B_0)=1$, the formula follows.

Returning to the main line, recall that we use the affine parametrization $P_t=(1-t)P_0+tP_1$, $t\in[0,1]$, for the line segment $[P_0,P_1]$, and the parametrization Q_t, $t\in[0,1]$, for the circular arc $\mathcal{C}$. For $t,t'\in[0,1]$, letting $P_t=A$, $P_{t'}=B$, $O=C$, so that $Q_t=A_0$, $Q_{t'}=B_0$, the formula above is rewritten as

$$d(Q_t,Q_{t'})^2=2+\frac{d(P_t,P_{t'})^2-d(O,P_t)^2-d(O,P_{t'})^2}{d(O,P_t)d(O,P_{t'})}$$

For future reference, we include here a useful equivalent form of this as

$$d(Q_t,Q_{t'})^2=2-\frac{2-(t+t'-2tt')d^2}{\sqrt{t^2d^2-td^2+1}\sqrt{t'^2d^2-t'd^2+1}},\quad d=d(P_0,P_1),$$

where we used $d(P_t,P_{t'})^2=(t-t')^2d^2$, $d(O,P_t)=\sqrt{t^2d^2-td^2+1}$ and $d(O,P_{t'})=\sqrt{t'^2d^2-t'd^2+1}$.

We rewrite the original formula and estimate

$$d(Q_t,Q_{t'})^2=\frac{d(P_t,P_{t'})^2-(d(O,P_t)-d(O,P_{t'}))^2}{d(O,P_t)d(O,P_{t'})}\leq\frac{d(P_t,P_{t'})^2}{d(O,P_t)d(O,P_{t'})}$$

Setting, as usual, $d=d(P_0,P_1)<2$, for the denominator, we have

$$d(O,P_t)^2=t^2d^2-td^2+1=\left(t-\frac{1}{2}\right)^2d^2+1-\frac{d^2}{4}\geq 1-\frac{d^2}{4}.$$

With this, we arrive at

$$d(Q_t,Q_{t'})\leq\frac{d}{\sqrt{1-d^2/4}}|t-t'|,\quad t,t'\in[0,1],\ d=d(P_0,P_1)<2.$$

We express this by saying that the map $Q:[0,1]\to\mathbb{R}^2$, $Q(t)=Q_t$, $t\in[0,1]$, satisfies the **Lipschitz condition** with **Lipschitz constant** $d/\sqrt{1-d^2/4}$.

Applying this to a partition $(t_0,t_1,\dots,t_{n-1},t_n)\in\Pi$ of $[0,1]$, we see that the length of the corresponding polygonal path

$$\sum_{i=1}^n d(Q_{t_{i-1}},Q_{t_i})\leq\frac{d}{\sqrt{1-d^2/4}}\sum_{i=1}^n(t_i-t_{i-1})=\frac{d}{\sqrt{1-d^2/4}}.$$

Taking the supremum for all polygonal paths, we finally obtain

$$\mathcal{L}_{\mathcal{C}} \leq \frac{d}{\sqrt{1-d^2/4}}, \quad d = d(P_0, P_1) < 2.$$

(Note that, as a simple computation shows, this upper estimate is the same as the one we obtained above by our geometric method).

We now explore the properties of the arc length of circular arcs in $\mathbb{S}_O$. We say that **additivity** holds in a circular arc if whenever it is split into two circular arcs $\mathcal{C}_1$ and $\mathcal{C}_2$ (by a common end-point), then we have

$$\mathcal{L}_{\mathcal{C}_1 \cup \mathcal{C}_2} = \mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_2}.$$

We first claim that additivity holds in a circular arc $\mathcal{C} \subset \mathbb{S}_O$ with end-points P_0 and P_1 satisfying $d(P_0, P_1) < 2$ (and $\mathcal{C}$ and O are in opposite sides of the line extension of $[P_0, P_1]$).

Let $Q \in \mathcal{C}$, $P_0 \neq Q \neq P_1$. Let $\mathcal{C}_1 \subset \mathcal{C}$, resp. $\mathcal{C}_2 \subset \mathcal{C}$, be the circular arc with end-points P_0 and Q, resp. P_1 and Q. We have $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2$. We claim that

$$\mathcal{L}_{\mathcal{C}} = \mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_2}.$$

Indeed, since a partition of $[P_0, Q]$ and a partition of $[P_1, Q]$ can be united to define a partition of $[P_0, P_1]$, taking suprema, we see that $\mathcal{L}_{\mathcal{C}} \geq \mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_2}$. On the other hand, letting $0 < \epsilon \in \mathbb{R}$, we can choose a polygonal path $\bigcup_{i=1}^n [Q_{t_{i-1}}, Q_{t_i}]$ such that

$$\mathcal{L}_{\mathcal{C}} - \epsilon \leq \sum_{i=1}^n d(Q_{t_{i-1}}, Q_{t_i}).$$

By the triangle inequality, adding $Q \in \mathcal{C}$ to a polygonal path for $\mathcal{C}$ increases its length. If Q participates in a polygonal path for $\mathcal{C}$, then this path can be split into the union of two polygonal paths, one for $\mathcal{C}_1$ and the other for $\mathcal{C}_2$. Once again, taking suprema, we obtain $\mathcal{L}_{\mathcal{C}} - \epsilon \leq \mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_2}$. Since this is true for all $0 < \epsilon \in \mathbb{R}$, we obtain $\mathcal{L}_{\mathcal{C}} \leq \mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_2}$. Additivity in the circular arc $\mathcal{C}$ follows.

Remark The arc length $\mathcal{L}_{\mathcal{C}}$ of a circular arc $\mathcal{C}$ with end-points P_0 and P_1 depends only on the distance $d(P_0, P_1) < 2$. This means that, if $\mathcal{C}'$ is another circular arc with end-points P_0' and P_1' such that $d(P_0', P_1') = d(P_0, P_1) < 2$, then we have $\mathcal{L}_{\mathcal{C}} = \mathcal{L}_{\mathcal{C}'}$. Indeed, a partition of the interval $[0, 1]$ induces a partition of $[P_0, P_1]$ and $[P_0', P_1']$, and the associated polygonal paths have the same lengths because, by the formula above, the distance $d(Q_t, Q_{t'})$ depends only on $d(P_0', P_1') = d(P_0, P_1)$ and the parameters $t, t' \in [0, 1]$. These equal lengths contribute the same amount to the suprema that define the arc lengths $\mathcal{L}_{\mathcal{C}}$ and $\mathcal{L}_{\mathcal{C}'}$, which thereby must be equal.

Finally, note an important special case. Let $\mathcal{C}$ be a circular arc with the end-points P_0 and P_1. We claim that $d(P_0, P_1) = \sqrt{2}$ if and only if the line extensions of the half-lines ℓ_0 and ℓ_1 (from the center O to the respective end-points) are perpendicular.

Indeed, assuming (for simplicity) that the center is at the origin, the equation of the line extension of the half-line ℓ_0 through P_0 is $y_0 x - x_0 y = 0$. The quarter-turn S_0 sends this line to the perpendicular line with the equation $x_0 x + y_0 y = 0$. This line contains the point $P_1 = (x_1, y_1)$ if and only if $x_0 x_1 + y_0 y_1 = 0$. Since $d(P_0, P_1)^2 = 2 - 2(x_0 x_1 + y_0 y_1)$ (see the end of the previous section), this is equivalent to $d(P_0, P_1)^2 = 2$. The claim follows.

With this, we can introduce the positive real number $\pi \in \mathbb{R}$ such that, for $d(P_0, P_1) = \sqrt{2}$, we have $\mathcal{L}_{\mathcal{C}} = \pi/2$. By the remark above, the arc length of a circular arc depends only on the distance between its end-points, so that π is well-defined.

We now **extend** the definition of the arc length to any circular arc of $\mathbb{S}_O$ using additivity.

If $d(P_0, P_1) < 2$, then the arc length of the circular arc $\mathcal{C}$ with end-points P_0 and P_1 has been defined above as the supremum of the lengths of its polygonal paths. In particular, for any quarter-circle $\mathcal{C}$ ($d(P_0, P_1) = \sqrt{2}$), we have $\mathcal{L}_{\mathcal{C}} = \pi/2$.

If $d(P_0, P_1) = 2$, then, by the sharp triangle inequality, the points P_0 and P_1 are collinear, and the center O is at the midpoint of the line segment $[P_0, P_1]$. We define $\mathcal{L}_{\mathcal{C}} = \mathcal{L}_{\mathcal{C}^c} = \pi$ for either of the semi-circles $\mathcal{C}$ or $\mathcal{C}^c$ with end-points P_0 and P_1. (Note that they are congruent via the half-turn S_O^2.)

If $d(P_0, P_1) < 2$, then we define the arc length of the circular arc $\mathcal{C}^c$ **complementary** to $\mathcal{C}$ by $\mathcal{L}_{\mathcal{C}^c} = 2\pi - \mathcal{L}_{\mathcal{C}}$.

Finally, we define the arc length of the entire circle $\mathbb{S}_O$ to be 2π.

We note that the arc length, being defined in terms of the Cartesian distance, is preserved under translations and half-turns.

We now claim that **additivity** holds in any circular arc of $\mathbb{S}_O$.

First, we show additivity in a semi-circle $\mathcal{C}$. Let $\mathcal{C}$ have end-points P_0 and P_1, and $Q \in \mathcal{C}$ with $P_0 \neq Q \neq P_1$. Then Q splits $\mathcal{C}$ into two circular arcs: $\mathcal{C}_1$ with end-points P_0 and Q and $\mathcal{C}_2$ with end-points P_1 and Q. We claim that $\mathcal{L}_{\mathcal{C}} = \mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_2} = \pi$.

Indeed, either $\mathcal{C}_1$ or $\mathcal{C}_2$ contains a quarter-circle. Assume, without loss of generality, that the first does. If $\mathcal{C}_1$ is itself a quarter-circle, then so is $\mathcal{C}_2$ and the statement holds. Otherwise, split $\mathcal{C}_1$ into a quarter circle with one end-point at P_0 and a circular arc $\mathcal{C}_1'$ with one end-point at Q. Then, by additivity in $\mathcal{C}_1$ already shown, we have $\mathcal{L}_{\mathcal{C}_1} = \pi/2 + \mathcal{L}_{\mathcal{C}_1'}$. Since $\mathcal{C}_1'$ and $\mathcal{C}_2$ join to form another quarter-circle, again by additivity in quarter-circles already shown, we also have $\mathcal{L}_{\mathcal{C}_1'} + \mathcal{L}_{\mathcal{C}_2} = \pi/2$. Putting these together, we obtain $\mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_2} = (\pi/2 + \mathcal{L}_{\mathcal{C}_1'}) + (\pi/2 - \mathcal{L}_{\mathcal{C}_1'}) = \pi = \mathcal{L}_{\mathcal{C}}$. Additivity in semi-circles follows.

Next, we need to show additivity in a circular arc of the form $\mathcal{C}^c$ complementary to the circular arc $\mathcal{C}$ with end-points P_0 and P_1 such that $d(P_0, P_1) < 2$. Let $Q \in \mathcal{C}^c$ distinct from P_0 and P_1. Then Q splits $\mathcal{C}^c$ into two circular arcs: $\mathcal{C}_1$ with one end-point at P_0 and another circular arc $\mathcal{C}_2$ with one end-point at P_1. We need to show $\mathcal{L}_{\mathcal{C}^c} = \mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_2}$.

First, assume that either $\mathcal{C}_1$ or $\mathcal{C}_2$ contains a half-circle. Without loss of generality, we may assume that $\mathcal{C}_1$ does. We have $\mathcal{L}_{\mathcal{C}_1} = \mathcal{L}_{(\mathcal{C}\cup\mathcal{C}_2)^c} = 2\pi - \mathcal{L}_{\mathcal{C}\cup\mathcal{C}_2} = 2\pi - (\mathcal{L}_{\mathcal{C}} + \mathcal{L}_{\mathcal{C}_2})$, where we used additivity in $\mathcal{C} \cup \mathcal{C}_2$ (including the case when $\mathcal{C} \cup \mathcal{C}_2$ is a semi-circle). On the other hand, we have $\mathcal{L}_{\mathcal{C}^c} = 2\pi - \mathcal{L}_{\mathcal{C}} = 2\pi - (2\pi - (\mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_2})) = \mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_2}$. Additivity follows in this case.

Second, assume neither $\mathcal{C}_1$ nor $\mathcal{C}_2$ contains a semi-circle. Let $P_2 \in \mathbb{S}_O$ be the opposite to P_0 with respect to O. Since neither $\mathcal{C}$ nor $\mathcal{C}_1$ contains semi-circles, we have $P_2 \in \mathcal{C}_2$, $P_1 \neq P_2 \neq Q$. Then P_2 splits $\mathcal{C}_2$ into two circular arcs: $\mathcal{C}_2'$ with one end-point at P_1 and another $\mathcal{C}_2''$ with one end-point at Q. Clearly, $\mathcal{C}_1 \cup \mathcal{C}_2''$ is a semi-circle, so that, by the previous case, we have $\mathcal{L}_{\mathcal{C}^c} = \mathcal{L}_{\mathcal{C}_1\cup\mathcal{C}_2''} + \mathcal{L}_{\mathcal{C}_2'}$. On the other hand, by additivity in semi-circles, we also have $\mathcal{L}_{\mathcal{C}_1\cup\mathcal{C}_2''} = \mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_1''}$. Putting these together, we obtain $\mathcal{L}_{\mathcal{C}^c} = \mathcal{L}_{\mathcal{C}_1\cup\mathcal{C}_2''} + \mathcal{L}_{\mathcal{C}_2'} = \mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_1''} + \mathcal{L}_{\mathcal{C}_2'} = \mathcal{L}_{\mathcal{C}_1} + \mathcal{L}_{\mathcal{C}_2}$. This finishes the second case. Additivity in complementary circular arcs follows.

Finally, note that additivity in the entire circle $\mathbb{S}_O$ follows from the definitions as any split consists of complementary pairs of circular arcs.

The proof of the additivity of the arc length in general is now complete.

Remark In view of the additivity and the forthcoming discussion, it is convenient to define the arc length of a single point on $\mathbb{S}_O$ to be zero.

As the final task in this section, we claim that any given number $0 \leq r \leq 2\pi$ arises as the arc length of a circular arc $\mathcal{C} \subset \mathbb{S}_O$, that is, we have $\mathcal{L}_{\mathcal{C}} = r$. By additivity, it is enough to show this for $0 < r < \pi/2$. Let ℓ_0 and ℓ_1 be perpendicular half-lines with common end-point O such that, for $P_0 \in \ell_0$ and $P_1 \in \ell_1$, $d(O, P_0) = d(O, P_1) = 1$, and hence $d(P_0, P_1) = \sqrt{2}$. The unit interval $[0, 1]$ parametrizes the line segment $[P_0, P_1]$ by the affine coordinate function $P_t = (1-t)P_0 + tP_1$, $t \in [0, 1]$, and the quarter-circle with end-points P_0 and P_1 by $Q_t = P_t/d(O, P_t) \in \mathbb{S}_O$, $t \in [0, 1]$.

For $t \in [0, 1]$, we denote by $\mathcal{C}_t \subset \mathbb{S}_O$ the circular arc with end-points $Q_0 (= P_0)$ and Q_t. The quarter-circle itself is then equal to $\mathcal{C}_1$, and we have

$$\mathcal{L}_{\mathcal{C}_0} = 0 \quad \text{and} \quad \mathcal{L}_{\mathcal{C}_1} = \pi/2.$$

We first study the properties of the arc length $\mathcal{L}_{\mathcal{C}_t}$ as a function of $t \in [0, 1]$. Since $d = d(P_0, P_1) = \sqrt{2}$, we have $d/\sqrt{1 - d^2/4} = 2$, so that the previous Lipschitz estimate gives

$$d(Q_t, Q_{t'}) \leq 2|t - t'|, \quad t, t' \in [0, 1].$$

On the other hand, the explicit formula for this distance specializes to

$$\begin{aligned} d(Q_t, Q_{t'})^2 &= 2 - \frac{2 - 2(t + t' - 2tt')}{\sqrt{2t^2 - 2t + 1}\sqrt{2t'^2 - 2t' + 1}} \\ &= 2 - \frac{2(1-t)(1-t')}{\sqrt{2t^2 - 2t + 1}\sqrt{2t'^2 - 2t' + 1}}, \quad t, t' \in [0, 1]. \end{aligned}$$

This gives

$$\max_{t,t' \in [0,1]} d(Q_t, Q_{t'}) = \sqrt{2}.$$

We now let $t, t' \in [0, 1]$ and consider the circular arc $\mathcal{C}_{t,t'} \subset \mathbb{S}_O$ with end-points Q_t and $Q_{t'}$. The general upper estimate derived earlier for the arc length gives

$$\mathcal{L}_{\mathcal{C}_{t,t'}} \leq \frac{d(Q_t, Q_{t'})}{\sqrt{1 - d(Q_t, Q_{t'})^2/4}} \leq 2\sqrt{2}|t - t'|, \quad t, t' \in [0, 1],$$

where we used the Lipschitz estimate and the maximum above. Finally, by additivity, the arc length of the path $\mathcal{C}_{t,t'}$ is equal to $\mathcal{L}_{\mathcal{C}_{t,t'}} = |\mathcal{L}_{\mathcal{C}_t} - \mathcal{L}_{\mathcal{C}_{t'}}|$. Putting these together, we obtain

$$|\mathcal{L}_{\mathcal{C}_t} - \mathcal{L}_{\mathcal{C}_{t'}}| \leq 2\sqrt{2}|t - t'|, \quad t, t' \in [0, 1].$$

This means that the arc length $\mathcal{L}_{\mathcal{C}_t}$ as a function of $t \in [0, 1]$ satisfies the Lipschitz property with Lipschitz constant $2\sqrt{2}$.

We now note the simple fact that the Lipschitz property above implies **continuity** of the function $t \mapsto \mathcal{L}_{\mathcal{C}_t}$, $t \in [0, 1]$. (Indeed, for any $0 < \epsilon \in \mathbb{R}$, we can choose $\delta = \epsilon/(2\sqrt{2})$, universally.) Hence, by the Intermediate Value Theorem (Section 4.2), for any given $0 \leq r \leq 2\pi$, there exists $t \in [0, 1]$ such that $\mathcal{L}_{\mathcal{C}_t} = r$. The claim follows.

Remark The arc length $\mathcal{L}_{\mathcal{C}}$ of a circular arc $\mathcal{C}$ with end-points P_0 and P_1 depends only on the distance $d(P_0, P_1) < 2$. We now make the additional claim that the correspondence that associates with the arc length $\mathcal{L}_{\mathcal{C}}$ of a circular arc $\mathcal{C}$ with end-points P_0 and P_1 and the distance $d(P_0, P_1)$ is **strictly increasing** in the sense that if $\mathcal{C}'$ and $\mathcal{C}''$ are circular arcs, then $\mathcal{L}_{\mathcal{C}'} < \mathcal{L}_{\mathcal{C}''}$ if and only if $d(P_0', P_1') < d(P_0'', P_1'')$ for the corresponding end-points.

Indeed, let $\mathcal{C}$ be a circular arc with end-points P_0 and P_1, $d(P_0, P_1) < 2$, such that $\max(\mathcal{L}_{\mathcal{C}'}, \mathcal{L}_{\mathcal{C}''}) < \mathcal{L}_{\mathcal{C}}$. Then, by the above, there exist $\mathcal{C}_0' \subset \mathcal{C}$ with one end-point at P_0 and congruent to $\mathcal{C}'$ and $\mathcal{C}_0'' \subset \mathcal{C}$ with one end-point at P_0 and congruent to $\mathcal{C}''$. Now the claim is equivalent to monotonicity of the distance $d(Q_t, Q_0)$ in $t \in [0, 1]$. As for this, first note that, as a special case of the general formula derived earlier, we have

$$d(Q_t, Q_0) = 2 - \frac{2 - td^2}{\sqrt{t^2d^2 + td^2 + 1}}, \quad d = d(P_0, P_1).$$

Now, monotonicity follows from the proof of the lemma above.

Exercise

5.6.1. Let $\triangle[A, B, C]$ be a triangle with vertices $A, B, C \in \mathbb{R}^2$ and side lengths $a, b, c \in \mathbb{R}$. Let $0 < r \in \mathbb{R}$ such that $2r < \min(a, b, c)$. Consider the configuration of three circles with centers A, B, C and radius r. What is the shortest length of a band that stretches around the outside of the three circles?

5.7 The Birkhoff Angle Measure

We now turn to the concept of **angle measure** μ for angles in the **Birkhoff Postulate of Angle Measure** (Section 5.1) in our model $\mathbb{R}^2$.

We first define the **angle measure** for angles in our model $\mathbb{R}^2$. Let $\angle AOB$ be an angle formed by the ordered triple (O, A, B), $A \neq O \neq B$. Let $\mathbb{S}_O$ be the unit radius circle with center O.

If (O, A, B) is a positively oriented triple, then we define the angle measure as the arc length $\mu(\angle AOB) = \mathcal{L}_{\mathcal{C}} \pmod{2\pi}$, where $\mathcal{C} \subset \mathbb{S}_O$ is the circular arc with end-points A_0 and B_0, the points at unit distance from O on the half-line with end-point O and containing A and B.[31]

If A, O, B are collinear, then we define $\mu(\angle AOB) = 0 \pmod{2\pi}$ if O is not between A and B, and $\mu(\angle AOB) = \pi \pmod{2\pi}$ if O is between A and B.

If (O, A, B) is a negatively oriented triple, then we define the angle measure as $\mu(\angle AOB) = -\mathcal{L}_{\mathcal{C}} \pmod{2\pi}$, where $\mathcal{C} \subset \mathbb{S}_O$ is the circular arc with end-points A_0 and B_0 as above.

Note that the angle measure $\mu(\angle AOB)$ depends only on the half-lines ℓ_0 and ℓ_1 with end-point O containing A and B, respectively. Thus, we can write $\mu(\angle \ell_0 O \ell_1) = \mu(\angle AOB)$.

To derive the Birkhoff Postulate of Angle Measure, for $O \in \mathbb{R}^2$, we need to define the function α_O on the set of all half-lines with end-point O to the real numbers modulo 2π. For O at the origin 0, we let $\alpha_0(\ell) = \mu(\angle \ell_+ 0 \ell)$, where ℓ is any half-line with end-point 0 and ℓ_+ is the positive first axis. With this, we define α_O by translating all the geometric entities from O to the origin 0.

---

[31] Note that the triple (O, A_0, B_0) is also positively oriented. Recall also that, according to our conventions, $\mathcal{C}$ is the shorter arc length circular arc with end-points A_0 and B_0.

Pág. 127(139) - GW170817:

Three Detectors and Two Neutron Stars

Pág. 128(140) - Figure 5.3

The narrowing down of the position of a gravitational wave source when three detectors are in use, by triangulation.

The huge reduction in the necessary search are is clear to see.

- 2021 09 23 | Elements of Mathematics.. |

Toth2021_Book_Elements of Mathematics.pdf

Pág. 207(218) Chapter 5

Real Analytic Plane Geometry

Pág. 224(235) Babylonian Method for $\sqrt{2}$ starting with $q_0 = 1$:

5 $\frac{886731088897}{627013566048}$

Pág. 225(236) 5.5

Lines and Circles

translation by W as the map

$T_W : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$T_W(P) = P + W$, $P \in \mathbb{R}^2$.

In coordinates, if $W = (u, v) \in \mathbb{R}^2$, then we have

$T_W(P) = P + W = (x+u, y+v)$, $P = (x, y) \in \mathbb{R}^2$.

Pag. 230 (241) →

The two right triangles $\Delta[P, P_0, C]$ and $\Delta[P, P_1, C]$ are congruent since they have two sides, and they also have a common side.[20]

In this example, we assume Birkhoff's Postulates of Angle Measure and Similarity and, consequently, the Pythagorean Theorem, whose validity in our model will be proved in Section 5.7.

Pag. 245 (256) 5.7 The Birkhoff Angle Measure.

Pag. 233 (244) Fig. 5.2
An occurrence of the golden number τ.

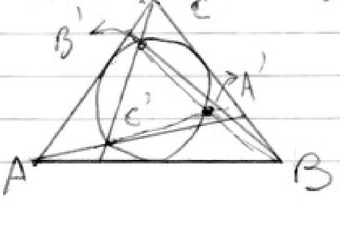

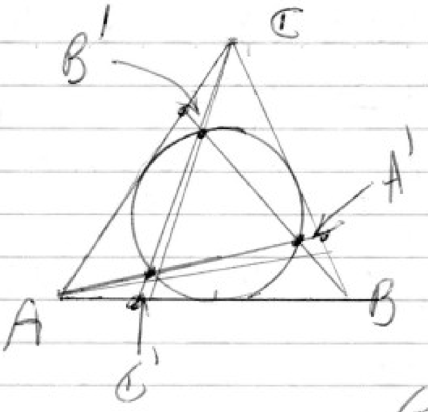

Obs:

Use esta configuração conjuntamente com Pappus's Theorem (three points)

orthogonal to $\underline{S}$.

(e) Prove Monge's theorem: Given three disjoint circles (with non-parallel radical axes), there is a circle orthogonal to all three.

Pag. 235 (246) Fig. 5.3
Parametrization of a circular arc.

Pag. 237 (248) Fig. 5.4
Upper bound for the arch length.

Obs: Utilizar com Pappus's theorem (three points)!!

m_j, h_i h_{i-1} (paralelos).

three points Q_{t_i}, T_i, R_i

Pag. 238(249) Fig. 5.5

Illustration for the monotonicity lemma.

→ use three points (Pappus's theorem)

⇒ P_1 P_t P_0.

obs Use esta reta determinada pelos três pontos quando você for fazer a deformação da imagem.

Pag. 239(250)

→ distance from C on the half-line (half-line)
with end-point C and containing A, and respectively B. We then have the following important formula: 29

In different (non-axiomatic) development, this formula is equivalent to the so-called Law of Cosines.

Pag. 240(251) – Lipschitz condition with Lipschitz constant

$d/\sqrt{1-d^2/4}$!

IMPORTANT:

Fig. 5.6

The sum of angles in a triangle is π.

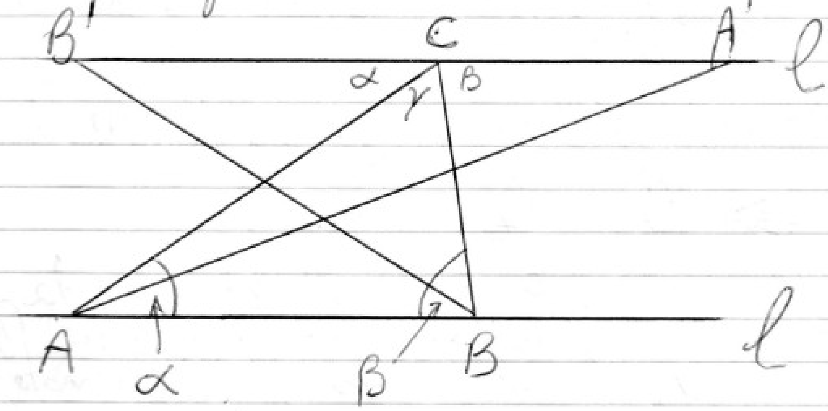

surjective
injective

$\alpha_0(l_1) - \alpha_0(l_0) = \mu(\angle l_0 O l_1) \pmod{2\pi}$.

Pág. 246(257) → interessante página.

$$\alpha + \beta + \gamma = \pi$$

Pág. 247(258) Fig. 5.7
Proof of Birkhoff's Postulate of Similarity.

Pág. 252(263) – 5.8
The Principle of Shortest Distance

Pág. 253(264) – rectifiable curve

if there exists a one-to-one 35

35 the property of being simple, that is, one-to-one, excludes "self-intersections".
As we consider here only open curves and minimize the arc length, imposing this does not restrict the generality.

Pag. 257(268) - 5.9

π According the Archimedes*

$3\frac{10}{71} < \pi \; 3\frac{1}{7}$ ///

20210923/// (20210923)

_ 20210923/ Birkhoff/

Groups, Matrices, and Vector Spaces...pdf

Pag. 211(221) - 7.3

Isometries and Orthogonal Mapping

The term orthogonal comes from the fact that an orthogonal mapping preserves the orthogonal relationship between orthogonal pairs of vectors.

@)*+,)))- !) %1 2 3 G ! G ! # 3!
* 2 3 6 *33 ! < ! * 4 ! 3#
' 1 & ! # - ! !33

7.3 Isometries and Orthogonal Mappings

Throughout this section, V will denote a real finite-dimensional inner product space. Recall that the inner product on V defines a distance function $d(\mathbf{a}, \mathbf{b}) = |\mathbf{a} - \mathbf{b}|$. Linear mappings $T : V \to V$ that preserve distances are called *isometries*. (Actually, as we will see below, a distance-preserving map is automatically linear, so the definition can be simplified.) The isometries of V form an important group, called $O(V)$. When $V = \mathbb{R}^n$, $O(V)$ is the matrix group $O(n, \mathbb{R})$ of all orthogonal $n \times n$ matrices over $\mathbb{R}$. We will first consider isometries in general and then specialize to rotations and reflections. Finally, we will consider the isometries of $\mathbb{R}^2$. Here we prove the $O(2, \mathbb{R})$-dichotomy: every isometry of $\mathbb{R}^2$ is either a rotation or a reflection. We will also show that the dihedral group $D(m)$ can be realized as the group of all isometries of a regular m-gon in $\mathbb{R}^2$. This will give another proof that the order of $D(m)$ is $2m$.

7.3.1 Isometries and orthogonal linear mappings

A mapping $S : V \to V$ is said to be *orthogonal* if

$$(S(\mathbf{x}), S(\mathbf{y})) = (\mathbf{x}, \mathbf{y}) \tag{7.5}$$

for all $\mathbf{x}, \mathbf{y} \in V$. The term orthogonal comes from the fact that an orthogonal mapping preserves the orthogonal relationship between orthogonal pairs of vectors. Since orthogonal mappings preserve inner products, they preserve lengths of vectors, distances between vectors, and angles between vectors, since the angle θ between $\mathbf{v}$ and $\mathbf{w}$ is found by the identity

$$\mathbf{v} \cdot \mathbf{w} = |\mathbf{v}||\mathbf{w}| \cos \theta.$$

It turns out that all orthogonal mappings are linear.

Proposition 7.7. *Let $S : V \to V$ be orthogonal. Then S is linear. In fact, the orthogonal mappings are exactly the isometries. In particular, all distance-preserving mappings on V are linear. Conversely, every isometry is orthogonal.*

Proof. To show that S is linear, we first show that for all $\mathbf{a}, \mathbf{b} \in V$, $|S(\mathbf{a} + \mathbf{b}) - S(\mathbf{a}) - S(\mathbf{b})|^2 = 0$. But

$$|S(\mathbf{a}+\mathbf{b}) - S(\mathbf{a}) - S(\mathbf{b})|^2 = (S(\mathbf{a}+\mathbf{b}) - S(\mathbf{a}) - S(\mathbf{b}), S(\mathbf{a}+\mathbf{b}) - S(\mathbf{a}) - S(\mathbf{b}))$$

$$= (S(\mathbf{a} + \mathbf{b}), S(\mathbf{a} + \mathbf{b})) + (S(\mathbf{a}), S(\mathbf{a})) + (S(\mathbf{b}), S(\mathbf{b})) - 2(S(\mathbf{a} + \mathbf{b}), S(\mathbf{a}))$$

$$-2(S(\mathbf{a}+\mathbf{b}),S(\mathbf{b}))+2(S(\mathbf{a}),S(\mathbf{b})).$$

Using the fact that S is orthogonal, it follows that

$$|S(\mathbf{a}+\mathbf{b})-S(\mathbf{a})-S(\mathbf{b})|^2=(\mathbf{a}+\mathbf{b},\mathbf{a}+\mathbf{b})+(\mathbf{a},\mathbf{a})+(\mathbf{b},\mathbf{b})$$
$$-2(\mathbf{a}+\mathbf{b},\mathbf{a})-2(\mathbf{a}+\mathbf{b},\mathbf{b})+2(\mathbf{a},\mathbf{b}).$$

Expanding further, one sees that the right-hand side is zero. Thus, $S(\mathbf{a}+\mathbf{b}) = S(\mathbf{a})+S(\mathbf{b})$. The proof that $S(r\mathbf{a}) = rS(\mathbf{a})$ for all $r \in \mathbb{F}$ is similar. Therefore, every orthogonal mapping is linear. To see that S is an isometry, note that

$$\begin{aligned}|S(\mathbf{b})-S(\mathbf{a})|^2 &= |S(\mathbf{b}-\mathbf{a})|^2\\ &= (S(\mathbf{a}-\mathbf{b}),S(\mathbf{a}-\mathbf{b}))\\ &= (\mathbf{b}-\mathbf{a},\mathbf{b}-\mathbf{a})\\ &= |\mathbf{b}-\mathbf{a}|^2.\end{aligned}$$

This completes the proof that orthogonal mappings are isometries. We will leave the rest of the proof as an exercise. □

More generally, every mapping $F : V \to V$ that preserves distances turns out to be orthogonal, and hence is an isometry (see Exercise 7.3.3).

7.3.2 Orthogonal linear mappings on $\mathbb{R}^n$

We are now going to show that the matrix of an isometry is orthogonal, and conversely, that every orthogonal matrix defines an isometry.

Proposition 7.8. *Every isometry* $T : \mathbb{R}^n \to \mathbb{R}^n$ *is the matrix linear mapping associated with a unique orthogonal matrix. Conversely, every orthogonal matrix defines a unique isometry.*

Proof. Suppose $T : \mathbb{R}^n \to \mathbb{R}^n$ is an isometry. Since T is linear, $T = T_Q$ for a unique $Q \in \mathbb{R}^{n\times n}$. But T is orthogonal, so $T_Q(\mathbf{x}) \cdot T_Q(\mathbf{y}) = Q\mathbf{x} \cdot Q\mathbf{y} = \mathbf{x} \cdot \mathbf{y}$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Thus

$$(Q\mathbf{x})^T(Q\mathbf{y}) = (\mathbf{x}^TQ^T)(Q\mathbf{y}) = \mathbf{x}^T(Q^TQ)\mathbf{y} = \mathbf{x}^T\mathbf{y} \tag{7.6}$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Setting $Q\mathbf{e}_i = \mathbf{q}_i$, (7.6) implies $\mathbf{q}_i^T\mathbf{q}_j = \mathbf{e}_i^T\mathbf{e}_j$. Hence $Q^TQ = I_n$, so Q is orthogonal. We leave the converse to the reader. □

Recall that the set of orthogonal $n \times n$ matrices is the orthogonal group $O(n,\mathbb{R})$, so $O(n,\mathbb{R})$ is also the group of isometries of $\mathbb{R}^n$.

7.3.3 Projections

Let $\mathbf{a} \in \mathbb{R}^2$ be nonzero, and consider the line $\mathbb{R}\mathbf{a}$ spanned by $\mathbf{a}$. In Section 6.6.2, we called the mapping

$$P_{\mathbf{a}}(\mathbf{x}) = \Big(\frac{\mathbf{a} \cdot \mathbf{x}}{\mathbf{a} \cdot \mathbf{a}}\Big)\mathbf{a}$$

the *projection* onto $\mathbb{R}\mathbf{a}$. We leave it as an exercise to show that the projection $P_{\mathbf{a}}$ is linear. Note that if $\mathbf{u}$ is the unit vector determined by $\mathbf{a}$, then since $\mathbf{a} \cdot \mathbf{a} = |\mathbf{a}|^2$, it follows that

$$P_{\mathbf{u}}(\mathbf{x}) = P_{\mathbf{a}}(\mathbf{x}) = (\mathbf{u}^T\mathbf{x})\mathbf{u}.$$

Thus,

$$P_{\mathbf{u}}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = (u_1x_1 + u_2x_2)\begin{pmatrix} u_1 \\ u_2 \end{pmatrix},$$

so

$$P_{\mathbf{u}}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} u_1^2 & u_1u_2 \\ u_1u_2 & u_2^2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Hence the matrix of $P_{\mathbf{u}}$ is

$$\begin{pmatrix} u_1^2 & u_1u_2 \\ u_1u_2 & u_2^2 \end{pmatrix}. \tag{7.7}$$

Note that a projection matrix is symmetric. The image of the projection $P_{\mathbf{u}}$ is the line $\mathbb{R}\mathbf{u}$, while its kernel is the line orthogonal to $\mathbb{R}\mathbf{u}$. A reflection $P_{\mathbf{u}}$ has the property that $P_{\mathbf{u}} \circ P_{\mathbf{u}} = P_{\mathbf{u}}$, since

$$P_{\mathbf{u}} \circ P_{\mathbf{u}}(\mathbf{v}) = P_{\mathbf{u}}((\mathbf{v} \cdot \mathbf{u})\mathbf{u})) = (\mathbf{v} \cdot \mathbf{u})P_{\mathbf{u}}(\mathbf{u}) = (\mathbf{v} \cdot \mathbf{u})\mathbf{u} = P_{\mathbf{u}}(\mathbf{v}).$$

We will next apply projections to find a general formula for the reflection $\mathbb{R}^n$ through a hyperplane.

7.3.4 Reflections

We will first find an expression for a plane reflection and use that to suggest how to define a reflection of $\mathbb{R}^n$ through a hyperplane. Let ℓ be a line through the origin of $\mathbb{R}^2$. We want to consider how to reflect $\mathbb{R}^2$ through ℓ. The sought-after mapping $H : \mathbb{R}^2 \to \mathbb{R}^2$ fixes every vector on ℓ and sends every vector $\mathbf{v}$ orthogonal to ℓ to $-\mathbf{v}$. (If a reflection is linear, this is enough information to determine H.) Suppose $\mathbf{v} \in \mathbb{R}^2$ is on neither ℓ nor $\ell^\perp$. Then $\mathbf{v}$ and its reflection $H(\mathbf{v})$ form an isosceles triangle with equal sides $\mathbf{v}$ and $H(\mathbf{v})$, which H swaps.

Choosing a unit vector $\mathbf{u}$ on $\ell^{\perp}$, let us write

$$\mathbf{v} = P_{\mathbf{u}}(\mathbf{v}) + \mathbf{c},$$

where $\mathbf{c} \in \ell$. This is the orthogonal decomposition of $\mathbf{v}$ with respect to ℓ and $\ell^{\perp}$. Hence the component on ℓ is $\mathbf{c} = \mathbf{v} - P_{\mathbf{u}}(\mathbf{v})$. By Euclidean geometry, $H(\mathbf{v}) = -P_{\mathbf{u}}(\mathbf{v}) + \mathbf{c}$. Replacing $\mathbf{c}$ by $\mathbf{v} - P_{\mathbf{u}}(\mathbf{v})$, we get the formula

$$H(\mathbf{v}) = -P_{\mathbf{u}}(\mathbf{v}) + (\mathbf{v} - P_{\mathbf{u}}(\mathbf{v})) = \mathbf{v} - 2P_{\mathbf{u}}(\mathbf{v}).$$

Therefore,

$$H(\mathbf{v}) = \mathbf{v} - 2P_{\mathbf{u}}(\mathbf{v}) = \mathbf{v} - 2(\mathbf{u} \cdot \mathbf{v})\mathbf{u} = \mathbf{v} - 2(\mathbf{u}^T\mathbf{v})\mathbf{u}. \tag{7.8}$$

This expression immediately establishes the following result.

Proposition 7.9. *The reflection H of $\mathbb{R}^2$ through a line ℓ passing through* $\mathbf{0}$ *is a linear mapping.*

Example 7.7. Let us find the reflection H through the line ℓ given by $x = -y$. Now $\mathbf{u} = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^T$ is a unit vector on $\ell^{\perp}$. Thus,

$$\begin{aligned} H\begin{pmatrix} a \\ b \end{pmatrix} &= \begin{pmatrix} a \\ b \end{pmatrix} - 2(\begin{pmatrix} a \\ b \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}) \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \\ &= \begin{pmatrix} a - (a+b) \\ b - (a+b) \end{pmatrix} \\ &= \begin{pmatrix} -b \\ -a \end{pmatrix}. \end{aligned}$$

Thus the matrix of H is

$$\begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}.$$

⊠

Now let Q denote the matrix of the reflection through the line orthogonal to a unit vector $\mathbf{u} \in \mathbb{R}^2$. Applying (7.7) and (7.8), one sees that

$$Q = \begin{pmatrix} 1 - 2u_1^2 & -2u_1u_2 \\ -2u_1u_2 & 1 - 2u_2^2 \end{pmatrix} = \begin{pmatrix} u_2^2 - u_1^2 & -2u_1u_2 \\ -2u_1u_2 & u_1^2 - u_2^2 \end{pmatrix}. \tag{7.9}$$

Thus, every 2×2 reflection matrix has the form

$$Q = \begin{pmatrix} a & b \\ b & -a \end{pmatrix}, \tag{7.10}$$

where $a^2 + b^2 = 1$. Conversely, as we will show below, a matrix of the form $Q = \begin{pmatrix} a & b \\ b & -a \end{pmatrix}$, where $a^2 + b^2 = 1$, is a reflection. Hence, the symmetric 2×2 orthogonal matrices are exactly the matrices of reflections.

Let us now ask what the linear mapping H defined by (7.8) does in the case of $\mathbb{R}^n$. The equation $\mathbf{v} \cdot \mathbf{u} = 0$ defines an $(n-1)$-dimensional subspace W of $\mathbb{R}^n$, namely the hyperplane through the origin orthogonal to $\mathbf{u}$. If $\mathbf{v} \in W$, then $H(\mathbf{v}) = \mathbf{v}$. On the other hand, if $\mathbf{v} = r\mathbf{u}$, then $H(\mathbf{v}) = rH(\mathbf{u}) = r(\mathbf{u} - 2\mathbf{u}) = -r\mathbf{u} = -\mathbf{v}$. In particular, H leaves the hyperplane W pointwise fixed and reverses vectors on the line $W^\perp$ orthogonal to this hyperplane.

Definition 7.4. Let $\mathbf{u} \in \mathbb{R}^n$ be a unit vector, and let W be the hyperplane in $\mathbb{R}^n$ orthogonal to the line $\mathbb{R}\mathbf{u}$. Then the linear mapping $H : \mathbb{R}^n \to \mathbb{R}^n$ defined by

$$H(\mathbf{v}) = \mathbf{v} - 2(\mathbf{v} \cdot \mathbf{u})\mathbf{u} = \mathbf{v} - 2(\mathbf{u}^T\mathbf{v})\mathbf{u} \tag{7.11}$$

is called the *reflection of* $\mathbb{R}^n$ *through* W.

Since reflecting $\mathbf{v} \in \mathbb{R}^n$ twice through a hyperplane W returns $\mathbf{v}$ to itself, a reflection H has the property that $H \circ H = I_{\mathbb{R}^n}$. This can be checked directly by matrix multiplication. Let P denote the matrix of the projection $P_{\mathbf{u}}$. Then the matrix Q of H is $Q = I_n - 2P$. Thus,

$$Q^2 = (I_n 2 - 2P)(I_n - 2P) = I_n - 4P + 4P^2 = I_n,$$

since $P^2 = P$.

Since Q is symmetric and $Q^2 = I_n$, Q is by definition orthogonal. Thus we get the following.

Proposition 7.10. *The matrix of a reflection is orthogonal. Hence, reflections are orthogonal linear mappings.*

7.3.5 Projections on a general subspace

As a final example of a linear mapping, let us work out the projection of finite-dimensional inner product space V onto an arbitrary subspace W. As we saw in Proposition 6.38, every $\mathbf{x} \in V$ admits a unique orthogonal decomposition

$$\mathbf{x} = \mathbf{w} + \mathbf{y},$$

where $\mathbf{w} \in W$ and $\mathbf{y} \in W^\perp$. The *projection of* V *onto* W is the mapping $P_W : V \to V$ defined by $P_W(\mathbf{x}) = \mathbf{w}$. The following proposition justifies calling P_W a projection.

Proposition 7.11. *The mapping* $P_W : V \to V$ *has the following properties:*

(*i*) P_W *is linear,*

(*ii*) $P_W(\mathbf{w}) = \mathbf{w}$ *if* $\mathbf{w} \in W$*,*

(*iii*) $P_W(W^\perp) = \{\mathbf{0}\}$*, and finally,*

(*iv*) $P_W + P_{W^\perp} = I$*.*

Proof. Since W is a finite-dimensional inner product space, we have shown that it has an orthonormal basis, say $\mathbf{u}_1, \dots, \mathbf{u}_k$. Assume that $\mathbf{w} \in W$. By Proposition 6.36,

$$\begin{aligned}\mathbf{w} &= \sum_{i=1}^{k}(\mathbf{w}, \mathbf{u}_i)\mathbf{u}_i \\ &= \sum_{i=1}^{k}(\mathbf{x} - \mathbf{y}, \mathbf{u}_i)\mathbf{u}_i \\ &= \sum_{i=1}^{k}(\mathbf{x}, \mathbf{u}_i)\mathbf{u}_i - \sum_{i=1}^{k}(\mathbf{y}, \mathbf{u}_i)\mathbf{u}_i \\ &= \sum_{i=1}^{k}(\mathbf{x}, \mathbf{u}_i)\mathbf{u}_i.\end{aligned}$$

The last identity holds, since $\mathbf{y} \in W^\perp$ and all the $\mathbf{u}_i$ are in in W. Hence,

$$P_W(\mathbf{x}) = \sum_{i=1}^{k}(\mathbf{x}, \mathbf{u}_i)\mathbf{u}_i.$$

This shows that P_W is linear and $P_W(\mathbf{w}) = \mathbf{w}$ if $\mathbf{w} \in W$. Thus, (*i*) and (*ii*) hold. Statement (*iii*) follows from the fact that $(\mathbf{y}, \mathbf{u}_i) = 0$ for all $\mathbf{y} \in W^\perp$, and (*iv*) is a consequence of the decomposition $\mathbf{x} = \mathbf{w} + \mathbf{y}$, since $P_{W^\perp}(\mathbf{x}) = \mathbf{y}$. □

When $V = \mathbb{R}^n$, Exercise 7.3.9 below gives an interesting alternative expression for P_W that doesn't require knowing an orthonormal basis of W.

7.3.6 Dimension two and the $O(2, \mathbb{R})$-dichotomy

The orthogonal group $O(2, \mathbb{R})$ of all isometries of $\mathbb{R}^2$ contains the group $\mathrm{Rot}(2)$ of all rotations of $\mathbb{R}^2$ as a normal subgroup. Recall that every rotation matrix has the form $R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$. We are now going to show that $O(2, \mathbb{R})$ has surprising decomposition.

if there exists a one-to-one ^{35}

^{35} the property of being simple, that is, one-to-one, excludes "self-intersections".
As we consider here only open curves and minimize the arc length, imposing this does not restrict the generality.

Pag. 257(268) - 5.9

π According the Archimedes*

$3\frac{10}{71} < \pi \; 3\frac{1}{7}$ ///

20210923/// (20210923)

_ 20210923/ Birkhoff/

Groups, Matrices, and Vector Spaces...pdf

Pag. 211(221) - 7.3

Isometries and Orthogonal Mapping

The term orthogonal comes from the fact that an orthogonal mapping preserves the orthogonal relationship between orthogonal pairs of vectors.

Since orthogonal mapping preserve inner products, they preserve the lengths of vectors, distance between vectors, and angles between vectors, since the angle θ between

v and w is found by the identity

$V \cdot W = |V| \, |W| \cos \theta$.

It turns out that all orthogonal mapping are linear.

Pág. 135(145) Chapter 6
Vector Spaces

Pág. 207(217) Chapter 7
Linear Mapping

20210923/// (20210923)

_ 20210923/ Mathematical thought from Ancient to Modern times - 3 volumes/

Morris Kline - Mathematical...pdf

_ 20210923/ A History of Geometrical Methods/
Julian... A History of Geometrical... djvu

! .--? @).--?). D ( ) (1) ! ! !
6 * # 6 1)
)*+,)))- !) %1 ? ! ! !
6 * # 6 8 B B1 -9 # = 8 5 (3#

!)*+,)))- !) %1 O ! < # * !
! 4 ! 3 ! 8 C : DG 7 8 4 ! (/(3#

!)*+,)))- !) %1 O ! < # 6
A C A # ! 2 5 = 8 5 < 4 ! 5 ((3#

! 1)*+,)))- !) %1 O ! < # 6
2 5 3 9 7 ! (% 3#

M 1 1 "
R9 & %% ' 1/ !# % # R9 ! N 8
! #

! " !" # # $% & " (!
%$ " !" ' "
!)**+ ,-)**+ -) . / - / '-+ !0
1 ! 2 3 ! " 1! 43 ! -

! -256- - -* ! - ") 7 1!
43 ! 1"

! -256- - -* ! - ") 87 '9 1 !
2 3 ! " 1! 43 ! 1"

!% -256- - -* ! - ") 6:6 4
; 7 !1 9 3 1"

!' -256- - -* ! - ") /$ <5$94 4 !
=> ! ; ?!1 ! 9 ! 1"

!$ -256- - -* ! - ") /$:$8<@ 2"A! "
! 1"

!B -256- - -* ! - ") /$ /56'C ? ! 2 D ! 1"
(? B !1 '

Chapter 4

Quotient Space and Direct Sum of Vector Spaces

1 Homomorphism of Vector Spaces or Linear Transformation

Definition: *Let U (F) and V (F) be two vector spaces. Then a mapping*

$$f : U \rightarrow V$$

is called a homomorphism or a linear transformation of U **into** *V, if*

(i) $f(\alpha + \beta) = f(\alpha) + f(\beta), \forall \alpha, \beta \in U$

and (ii) $f(a\alpha) = af(\alpha) \forall a \in F, \quad \forall \alpha \in U.$

The conditions (i) and (ii) can be combined into a single condition

$$f(a\alpha + b\beta) = af(\alpha) + bf(\beta) \forall a, b \in F \text{ and } \forall \alpha, \beta \in U.$$

If f is a homomorphism of U **onto** V, then V is called a homomorphic image of U.

Theorem: *If f is a homomorphism of U (F) into V (F), then*

(i) $f(\mathbf{0}) = \mathbf{0}'$ *where* $\mathbf{0}$ *and* $\mathbf{0}'$ *are the zero vectors of U and V respectively.*

(ii) $f(-\alpha) = -f(\alpha) \forall \alpha \in U.$

Proof: (i) Let $\alpha \in U$. Then $f(\alpha) \in V$. Since $\mathbf{0}'$ is the zero vector of V, therefore

$$f(\alpha) + \mathbf{0}' = f(\alpha) = f(\alpha + \mathbf{0}) = f(\alpha) + f(\mathbf{0}).$$

Now V is an abelian group with respect to addition of vectors.

$\therefore$ $\quad f(\alpha) + \mathbf{0}' = f(\alpha) + f(\mathbf{0})$

$\Rightarrow$ $\quad \mathbf{0}' = f(\mathbf{0})$ by left cancellation law.

(ii) If $\alpha \in U$, then $-\alpha \in U$. Also we have

$$\mathbf{0}' = f(\mathbf{0}) = f[\alpha + (-\alpha)] = f(\alpha) + f(-\alpha).$$

Now $f(\alpha) + f(-\alpha) = \mathbf{0}' \Rightarrow f(-\alpha) =$ the additive inverse of $f(\alpha)$

$\Rightarrow$ $\quad f(-\alpha) = -f(\alpha).$

2 Isomorphism of Vector Spaces

(Garhwal 2010B)

Definition: *Let $U(F)$ and $V(F)$ be two Vector spaces. Then a mapping*

$$f : U \to V$$

is called an isomorphism of U onto V, if

(i) *f is one-one,*

(ii) *f is onto,*

(iii) $f(a\alpha + b\beta) = a f(\alpha) + b f(\beta) \;\forall\; a, b \in F, \alpha, \beta \in U.$

Also then the two vector spaces U and V are said to be isomorphic and symbolically we write

$$U(F) \cong V(F).$$

The vector space $V(F)$ is also called the isomorphic image of the vector space $U(F)$.

If f is a homomorphism of $U(F)$ into $V(F)$, then f will become an isomorphism of U into V if f is one-one. Also in addition if f is onto V, then f will become an isomorphism of U onto V.

3 Isomorphism of Finite Dimensional Vector Spaces

Theorem 1: *Two finite dimensional vector spaces over the same field are isomorphic if and only if they are of the same dimension.* **(Garhwal 2006, 10B; Kumaun 12)**

Proof: First suppose that $U(F)$ and $V(F)$ are two finite dimensional vector spaces each of dimension n. Then to prove that

$$U(F) \cong V(F).$$

Let the sets of vectors

$$\{\alpha_1, \alpha_2, \ldots, \alpha_n\} \quad \text{and} \quad \{\beta_1, \beta_2, \ldots, \beta_n\}$$

be the bases of U and V respectively.

Any vector $\alpha \in U$ can be uniquely expressed as

$$\alpha = a_1\alpha_1 + a_2\alpha_2 + \ldots + a_n\alpha_n.$$

Let $f : U \in V$ be defined by

$$f(\alpha) = a_1\beta_1 + a_2\beta_2 + \ldots + a_n\beta_n.$$

(%$

20210924!!!

Books-019-2021 | 20210924 /
Krishna's Spherical Astronomy and Space Dynamics /

- B08... Space-Dynamics.pdf
- B08... Spherical Astronomy... pdf
and Space Dynamics fro pdf

estudar conjuntamente!!!

- B08... Discrete Structures Graph Theory.pdf
- B08... Differential Equations
(B09) Laplace Transforms.pdf
- B09... Advanced Calculus.pdf
- B09... Linear Algebra.pdf

Pag. L71(76) Chapter 4

Quotient Space and Direct Sum of Vector Spaces

1. Homomorphism of Vector Spaces or Linear Transformations

Let U(F) and V(F) be two vector spaces. Then a mapping f: U→V is

jandaia

(%B

called a homomorphism or a linear transformation of U into V,

if (i) $f(\alpha+\beta) = f(\alpha)+f(\beta), \forall \alpha, \beta \in U$

and

(ii) $f(a\alpha) = af(\alpha) \; \forall a \in F, \forall \alpha \in U$.

- 2021 09/24

Hilbert and Banach Space.../

Hilbert and Banach Space....pdf

Pág. 79 (92) 3.2 Orthogonally scattered dilations

Pág. 92 (105) 3.3 Gramian orthogonally scattered dilations

- 2021 09 24/

Finite Element Analysis Applications/

0128099526.pdf

Pág. $\underline{\text{XIII}}$ (3)

Finite element analysis (FEA)

Pág. 32(36) Chapter 2

Decomposition

Functional Requirements → FRs

D -256- - -* ! - " : D 2 ")! ! 1! A! "
! 1"
(/ / % * ! 3 ! " " !

$$
\begin{aligned}
&= \sup\left\{\frac{\pi^2\|\phi\|_H^2\|f\|_2^2}{6} : \|\phi\|_H \le 1\right\}\\
&= \frac{\pi^2\|f\|_2^2}{6} < \infty.
\end{aligned}
$$

If we take $\phi = \frac{\sqrt{6}}{\pi}\sum_{k=1}^{\infty}\frac{1}{k}\phi_k$, then actually the equality holds in (1.16) and hence $|m_\xi|_w(\mathbb{R},\mathbb{R}) = \frac{\pi^2\|f\|_2^2}{6}$. But $\|\xi\|_s(\mathbb{R}) = \infty$ was shown in (3).

(6) $\|\xi\|_s(\mathbb{R})^2 < |M_\xi|(\mathbb{R},\mathbb{R}) = \infty$: Let ξ be as in (2). Then we see that

$$
\begin{aligned}
\|\xi\|_s(\mathbb{R})^2 &= \sup\left\{\left\|\sum_{k=1}^{\infty}\alpha_k\Big\langle\xi\big(\{k\}\big),\phi_k\Big\rangle_H\right\|_2^2 : \alpha_k = 0,1,\ \|\phi_k\|_H \le 1,\ k \in \mathbb{N}\right\}\\
&= \sup\left\{\left\|\sum_{k=1}^{\infty}\alpha_k\big\langle g_k\phi,\phi_k\big\rangle_H\right\|_2^2 : \alpha_k = 0,1,\ \|\phi_k\|_H \le 1,\ k \in \mathbb{N}\right\}\\
&= \sup\left\{\left\|\sum_{k=1}^{\infty}\alpha_k g_k\right\|_2^2\|\phi\|_H^2 : -1 \le \alpha_k \le 1,\ k \in \mathbb{N}\right\}\\
&= \sup\left\{\sum_{j,k=1}^{\infty}\alpha_j\alpha_k(g_j,g_k)_2\|\phi\|_H^2 : -1 \le \alpha_k \le 1,\ k \in \mathbb{N}\right\}\\
&\le \alpha\|\phi\|_H^2 < \infty
\end{aligned}
$$

(cf. (1.15)). Clearly $|M_\xi|(\mathbb{R},\mathbb{R}) \ge |m_\xi|(\mathbb{R},\mathbb{R}) = \infty$ by (2).

Remark 25. As will be seen in Section 3, $\|\xi\|_s(A) < \infty$ if and only if $\|\xi\|_o(A) < \infty$ (cf. Theorem 3.15). Hence Example 24 (6) serves as one for which $\|\xi\|_o(\mathbb{R})^2 < |M_\xi|(\mathbb{R},\mathbb{R}) = \infty$. But it is possible to construct a $\xi \in ca(\mathfrak{B}, X)$ such that $\|\xi\|_s(\mathbb{R}) < \|\xi\|_o(\mathbb{R}) < \infty$.

3.2. Orthogonally scattered dilations

In this section we discuss the orthogonally scattered dilation of Hilbert space-valued measures. In doing this, we use the Grothendieck inequality and the notion of absolutely 2-summing operators. As in the previous section, let $(\Theta, \mathfrak{A})$ be a measurable space.

We begin with the integration of scalar-valued functions with respect to a Banach space-valued measure. By a **$\mathbb{C}$-valued $\mathfrak{A}$-simple** (or, merely, **simple**) **function** we mean a function f of the form $f = \sum_{k=1}^{n}\alpha_k 1_{A_k}$ for $\alpha_1, \dots, \alpha_n \in \mathbb{C}$ and $\{A_1, \dots, A_n\} \in \Pi(\Theta)$, where 1_A denotes the **indicator function** of A. $L^0(\Theta)$

denotes the set of all $\mathbb{C}$-valued $\mathfrak{A}$-simple functions on Θ. If f is a $\mathbb{C}$-valued function on Θ, then the **sup-norm,** $\|f\|_\infty$, of f is defined by $\|f\|_\infty = \sup_{t\in\Theta} |f(t)|$.

Definition 1. Let $\mathfrak{X}$ be a Banach space and $\xi \in ca(\mathfrak{A}, \mathfrak{X})$. $A \in \mathfrak{A}$ is said to be **ξ-null** if $\|\xi\|(A) = 0$. Thus **ξ-a.e.** (= **ξ-almost everywhere**) refers to the complement of a ξ-null set. The **ξ-essential sup-norm** $\|f\|_{\infty,\xi}$ of a $\mathbb{C}$-valued $\mathfrak{A}$-measurable function f on Θ is defined by

$$\|f\|_{\infty,\xi} = \inf\Big\{\alpha > 0 : \big[|f| \ge \alpha\big] \text{ is } \xi\text{-null}\Big\},$$

where $[|f| \ge \alpha] = \{t \in \Theta : |f(t)| \ge \alpha\}$. It is easy to check that $\|\cdot\|_{\infty,\xi}$ is actually a norm, where we identify two functions if they agree ξ-*a.e.* For a simple function $f = \sum_{k=1}^n \alpha_k 1_{A_k} \in L^0(\Theta)$, the **integral** of f over $A \in \mathfrak{A}$ with respect to ξ is defined by

$$\int_A f\,d\xi = \sum_{k=1}^n \alpha_k \xi(A_k \cap A). \tag{2.1}$$

In this case we have

$$\left\| \int_A f\,d\xi \right\|_{\mathfrak{X}} \le \|f\|_{\infty,\xi}\|\xi\|(A). \tag{2.2}$$

$L^\infty(\xi) = L^\infty(\xi\,;\mathbb{C})$ denotes the set of all $\mathbb{C}$-valued functions f on Θ such that $\|f_n - f\|_{\infty,\xi} \to 0$ for some sequence $\{f_n\}_{n=1}^\infty \subseteq L^0(\Theta)$. It is not hard to see that $L^\infty(\xi)$ is a Banach space with the norm $\|\cdot\|_{\infty,\xi}$. For $f \in L^\infty(\xi)$ let $\{f_n\}_{n=1}^\infty \subseteq L^0(\Theta)$ be as above. Then, $\{\int_A f_n\,d\xi\}_{n=1}^\infty \subseteq \mathfrak{X}$ is a Cauchy sequence in view of the inequality (2.2) since $\{f_n\}_{n=1}^\infty$ is Cauchy in $L^\infty(\xi)$. Thus we can define the **integral** of f over A with respect to ξ by

$$\int_A f\,d\xi = \lim_{n\to\infty} \int_A f_n\,d\xi.$$

A more sophisticated definition of the integral of $\mathbb{C}$-valued functions with respect to a Banach space-valued measure is given as follows.

Definition 2. Let $\mathfrak{X}$ be a Banach space and $\xi \in ca(\mathfrak{A}, \mathfrak{X})$. A $\mathbb{C}$-valued $\mathfrak{A}$-measurable function f on Θ is said to be **ξ-integrable in the DS-sense** if there is a sequence $\{f_n\}_{n=1}^\infty \subseteq L^0(\Theta)$ of simple functions such that

a) $f_n(t) \to f(t)$ ξ-*a.e.* t;

b) $\{\int_A f_n\,d\xi\}_{n=1}^\infty \subseteq \mathfrak{X}$ is a Cauchy sequence for every $A \in \mathfrak{A}$.

In this case the **integral** of f over A with respect to ξ is defined by

$$\int_A f\,d\xi = \lim_{n\to\infty} \int_A f_n\,d\xi, \tag{2.3}$$

which is also termed the **DS-integral.** The sequence $\{f_n\}_{n=1}^{\infty}$ is called a **determining sequence** for f. $L^1_{DS}(\xi) = L^1_{DS}(\xi\,;\mathbb{C})$ denotes the set of all $\mathbb{C}$-valued $\mathfrak{A}$-measurable functions on Θ which are ξ-integrable in the DS-sense.

Remark 3. Let $\xi \in ca(\mathfrak{A}, \mathfrak{X})$, $\mathfrak{X}$ being a Banach space. As was shown in Dunford and Schwartz [1, IV.10.7] we have the following.

(1) The DS-integral (2.3) is well-defined, i.e., it is independent of the determining sequences of simple functions.

(2) For $f \in L^1_{DS}(\xi)$ and $A \in \mathfrak{A}$ it holds that

$$\left\| \int_A f\,d\xi \right\|_{\mathfrak{X}} \le \|f\|_{\infty,\xi} \|\xi\|(A).$$

(3) $L^1_{DS}(\xi)$ is a linear space and the DS-integral is linear.

(4) $L^\infty(\xi)$ is a subspace of $L^1_{DS}(\xi)$.

(5) If $f \in L^1_{DS}(\xi)$ and $\xi_f(A) = \int_A f\,d\xi$ for $A \in \mathfrak{A}$, then $\xi_f \in ca(\mathfrak{A}, \mathfrak{X})$.

Moreover, the following are true.

(6) For $A \in \mathfrak{A}$ it holds that

$$\begin{aligned}\|\xi\|(A) &= \sup\left\{ \left\| \int_A f\,d\xi \right\|_{\mathfrak{X}} : f \in L^0(\Theta),\ \|f\|_{\infty,\xi} \le 1 \right\} \\ &= \sup\left\{ \left\| \int_A f\,d\xi \right\|_{\mathfrak{X}} : f \in L^1_{DS}(\xi),\ \|f\|_{\infty,\xi} \le 1 \right\}.\end{aligned}$$

This follows from Definition 1.1 and (2) above.

(7) For $f \in L^1_{DS}(\xi)$, $\|\xi_f\|(\Theta) = 0$ if and only if $f = 0$ ξ-*a.e.* Hence, if we identify $f, g \in L^1_{DS}(\xi)$ when $f = g$ ξ-*a.e.*, then

$$\|f\|_{s,\xi} = \|\xi_f\|(\Theta)$$

(*"s"* for semivariation) defines a norm in $L^1_{DS}(\xi)$, which may be termed the "semivariation norm" of f. In fact, assume that $\|\xi_f\|(\Theta) = 0$. Put

$$A_n = \left[0 < |f| < \frac{1}{n}\right] = \left\{ t \in \Theta : 0 < |f(t)| < \frac{1}{n} \right\}$$

for $n \in \mathbb{N}$. Then, $\|\xi\|(A_n) \to 0$ by Remark 1.2 (2) since $A_n \downarrow \emptyset$. If we set $B = [f \ne 0]$, then it holds that for $n \in \mathbb{N}$

$$\|\xi\|(B\backslash A_n) = \sup\left\{ \left\| \int_{B\backslash A_n} g\,d\xi \right\|_{\mathfrak{X}} : g \in L^1_{DS}(\xi),\ \|g\|_{\infty,\xi} \le 1 \right\}$$

$$\begin{aligned}&\le \sup\left\{\left\|\int_{B\setminus A_n} gnf\,d\xi\right\|_{\mathfrak{X}} : g \in L^1_{DS}(\xi),\ \|g\|_{\infty,\xi} \le 1\right\}\\ &\le n\|\xi_f\|(B\setminus A_n) \le n\|\xi_f\|(\Theta) = 0.\end{aligned}$$

Thus, $\|\xi\|(B) \le \|\xi\|(B\setminus A_n) + \|\xi\|(A_n) = \|\xi\|(A_n) \to 0$, so that $\|\xi\|(B) = 0$. Therefore $f = 0$ *ξ-a.e.*

Theorem 4. *If* $\mathfrak{X}$ *is a Banach space and* $\xi \in ca(\mathfrak{A}, \mathfrak{X})$, *then* $(L^1_{DS}(\xi), \|\cdot\|_{s,\xi})$ *is a Banach space and the set of all simple functions* $L^0(\Theta)$ *is dense in it. Moreover, the set* $\{\xi_f : f \in L^1_{DS}(\xi)\}$ *is a closed subspace of* $ca(\mathfrak{A}, \mathfrak{X})$.

Proof. Clearly $(L^1_{DS}(\xi), \|\cdot\|_{s,\xi})$ is a normed linear space. We first prove that $L^0(\Theta)$ is dense in $L^1_{DS}(\xi)$. Let $f \in L^1_{DS}(\xi)$ and $\varepsilon > 0$ be given. Define $A_n = [|f| > n]$ for $n \in \mathbb{N}$. Then $\|\xi_f\|(A_n) \to 0$ by Remark 1.2 (2) since $A_n \downarrow \emptyset$. Choose $n_0 \in \mathbb{N}$ such that $\|\xi_f\|(A_{n_0}) < \frac{\varepsilon}{2}$ and choose $g \in L^0(\Theta)$ such that $g = 0$ on A_{n_0} and $|g - f| < \frac{\varepsilon}{2\|\xi\|(A^c_{n_0})}$ on $A^c_{n_0}$. Then it holds that

$$\begin{aligned}\|g - f\|_{s,\xi} &= \|\xi_g - \xi_f\|(\Theta) \le \|\xi_{g-f}\|(A^c_{n_0}) + \|\xi_f\|(A_{n_0})\\ &\le \frac{\varepsilon}{2\|\xi\|(A^c_{n_0})}\|\xi\|(A^c_{n_0}) + \frac{\varepsilon}{2} = \varepsilon.\end{aligned}$$

Thus $L^0(\Theta)$ is dense in $L^1_{DS}(\xi)$.

To see that $L^1_{DS}(\xi)$ is a Banach space we only have to show the completeness of $L^1_{DS}(\xi)$. So let $\{f_n\}_{n=1}^\infty \subseteq L^1_{DS}(\xi)$ be a Cauchy sequence. Since $L^0(\Theta)$ is dense in $L^1_{DS}(\xi)$ we can find a sequence $\{g_n\}_{n=1}^\infty \subseteq L^0(\Theta)$ such that $\|g_n - f_n\|_{s,\xi} \to 0$ as $n \to \infty$. Note that $\{g_n\}_{n=1}^\infty$ is also a Cauchy sequence in $L^1_{DS}(\xi)$. We can assume that

$$\sum_{k=1}^\infty \|g_{k+1} - g_k\|_{s,\xi} \equiv \alpha < \infty,$$

since, otherwise we can consider a suitable subsequence of $\{g_n\}_{n=1}^\infty$. Put

$$B = \left\{t \in \Theta : \sum_{k=1}^\infty \left|g_{k+1}(t) - g_k(t)\right| = \infty\right\}$$

and we shall show that $\|\xi\|(B) = 0$. So suppose that $\|\xi\|(B) = \varepsilon > 0$. Define

$$B_n = \left\{t \in \Theta : \sum_{k=1}^n \left|g_{k+1}(t) - g_k(t)\right| > \frac{2\alpha}{\varepsilon}\right\}, \qquad n \in \mathbb{N}.$$

Letting $B_0 = \bigcup_{n=1}^\infty B_n$, we see that $B_n \uparrow B_0$ and $\|\xi\|(B_0\setminus B_n) \to 0$. Hence we can find $n_0 \in \mathbb{N}$ such that $\|\xi\|(B_{n_0}) > \frac{\varepsilon}{2}$ since $\|\xi\|(B_0) \ge \|\xi\|(B) = \varepsilon$. Thus we have

$$\alpha = \sum_{k=1}^\infty \|g_{k+1} - g_k\|_{s,\xi} \ge \sum_{k=1}^{n_0} \left\|\xi_{(g_{k+1}-g_k)}\right\|(B_{n_0})$$

$$\geq \big\| \xi_{\Sigma_{k=1}^{n_0} |g_{k+1} - g_k|} \big\| (B_{n_0}) \geq \frac{2\alpha}{\varepsilon} \|\xi\|(B_{n_0}) > \alpha,$$

a contradiction. Consequently $\|\xi\|(B) = 0$.

Define a function f on Θ by

$$f(t) = \begin{cases} g_1(t) + \sum\limits_{k=1}^{\infty} \big(g_{k+1}(t) - g_k(t)\big), & \text{if } t \in B^c \\ 0, & \text{if } t \in B \end{cases}$$

and observe that $g_n \to f$ ξ-*a.e.* and, for $A \in \mathfrak{A}$, $\{\int_A g_n \, d\xi\}_{n=1}^{\infty}$ is a Cauchy sequence in $\mathfrak{X}$, since for $n, m \in \mathbb{N}$

$$\begin{aligned} \left\| \int_A g_{n+m} \, d\xi - \int_A g_n \, d\xi \right\|_{\mathfrak{X}} &\leq \|\xi_{(g_{n+m} - g_n)}\|(\Theta) \\ &\leq \sum_{k=n}^{\infty} \|g_{k+1} - g_k\|_{s,\xi} \to 0 \qquad (n \to \infty), \end{aligned}$$

which implies that $f \in L^1_{DS}(\xi)$. Since $ca(\mathfrak{A}, \mathfrak{X})$ is complete (cf. Remark 1.2 (3)) there exists an $\eta \in ca(\mathfrak{A}, \mathfrak{X})$ such that $\|\xi_{g_n} - \eta\|(\Theta) \to 0$. But since $\xi_{g_n}(A) \to \xi_f(A)$ for $A \in \mathfrak{A}$, we see that $\eta = \xi_f$. Therefore $\|g_n - f\|_{s,\xi} \to 0$ and hence $\|f_n - f\|_{s,\xi} \to 0$. The last assertion is obvious. □

Let us now consider H-valued measures. If $\xi \in caos(\mathfrak{A}, H)$, then $\nu_\xi(\cdot) \equiv \|\xi(\cdot)\|_H^2 \in ca(\mathfrak{A}, \mathbb{R}^+)$. Note that ν_ξ is a control measure of ξ (cf. Remark 1.2 (2)). For $f \in L^0(\Theta)$ we can define the **integral** of f over $A \in \mathfrak{A}$ with respect to ξ by (2.1). Observe that for $f = \sum_{k=1}^n \alpha_k 1_{A_k} \in L^0(\Theta)$

$$\begin{aligned} \left\| \int_A f \, d\xi \right\|_H^2 &= \left(\sum_{j=1}^n \alpha_j \xi(A_j \cap A), \sum_{k=1}^n \alpha_k \xi(A_k \cap A) \right)_H \\ &= \sum_{j,k=1}^n \alpha_j \overline{\alpha}_k \big(\xi(A_j \cap A), \xi(A_k \cap A)\big)_H \\ &= \sum_{k=1}^n |\alpha_k|^2 \nu_\xi(A_k \cap A) \\ &= \int_A |f|^2 \, d\nu_\xi = \|1_A f\|_{2,\nu_\xi}^2, \end{aligned} \tag{2.4}$$

where $\|\cdot\|_{2,\nu_\xi}$ is the norm in $L^2(\nu_\xi) = L^2(\Theta, \nu_\xi)$. For $f \in L^2(\nu_\xi)$ we can find a sequence $\{f_n\}_{n=1}^{\infty} \subseteq L^0(\Theta)$ such that $\|f_n - f\|_{2,\nu_\xi} \to 0$ and $f_n \to f$ ξ-*a.e.*, since

ξ-*a.e.* and ν_ξ-*a.e.* are equivalent because of $\|\xi\|(A)^2 = \|\xi(A)\|_H^2 = \nu_\xi(A)$ for $A \in \mathfrak{A}$ by Theorem 1.5 (4). By (2.4) we have for $A \in \mathfrak{A}$

$$\begin{aligned}\left\| \int_A f_n \, d\xi - \int_A f_m \, d\xi \right\|_H &= \big\| 1_A (f_n - f_m) \big\|_{2,\nu_\xi} \\ &\le \|f_n - f_m\|_{2,\nu_\xi} \to 0 \qquad (n, m \to \infty),\end{aligned}$$

i.e., $\{\int_A f_n \, d\xi\}_{n=1}^\infty$ is a Cauchy sequence in H. Thus f is ξ-integrable in the DS-sense. Moreover, (2.4) holds for $f \in L^2(\nu_\xi)$ and $A \in \mathfrak{A}$. Consequently $L^2(\nu_\xi) \subseteq L^1_{DS}(\xi)$.

Actually we see that the equality $L^2(\nu_\xi) = L^1_{DS}(\xi)$ is valid. In fact, for any $f \in L^1_{DS}(\xi)$ we have that by (2.4)

$$\|f\|_{s,\xi} = \|\xi_f\|(\Theta) = \big\|\xi_f(\Theta)\big\|_H = \left\| \int_\Theta f \, d\xi \right\|_H = \|f\|_{2,\nu_\xi},$$

since $\xi_f \in caos(\mathfrak{A}, H)$, which can be easily verified, and we can apply Theorem 1.5 (4) again. Thus the closure of $L^0(\Theta)$ with respect to $\|\cdot\|_{s,\xi}$ and $\|\cdot\|_{2,\nu_\xi}$ are the same. Therefore we have the following.

Proposition 5. *If* $\xi \in caos(\mathfrak{A}, H)$, *then* $(L^2(\nu_\xi), \|\cdot\|_{2,\nu_\xi}) = (L^1_{DS}(\xi), \|\cdot\|_{s,\xi})$, *where* $\nu_\xi(\cdot) = \|\xi(\cdot)\|_H^2$.

Definition 6. $\xi \in ca(\mathfrak{A}, H)$ is said to have an **orthogonally scattered dilation** if there exist a Hilbert space $\mathfrak{H}$ containing H as a closed subspace and an $\eta \in caos(\mathfrak{A}, \mathfrak{H})$ such that $\xi = J\eta$, where $J : \mathfrak{H} \to H$ is the orthogonal projection. The triple $\{\eta, \mathfrak{H}, J\}$ is also called an **orthogonally scattered dilation** of ξ.

ξ is said to have a **spectral dilation** if there exist a Hilbert space $\mathfrak{H}$, a countably additive spectral measure $E(\cdot)$ in $\mathfrak{H}$, an operator $S \in B(\mathfrak{H}, H)$ and a vector $\psi \in \mathfrak{H}$ such that

$$\xi(\cdot) = SE(\cdot)\psi.$$

Now the question is: Does every measure $\xi \in ca(\mathfrak{A}, H)$ have an orthogonally scattered dilation? An affirmative answer will be given after some considerations.

Definition 7. Let $\mathfrak{X}$ and $\mathfrak{Y}$ be two normed linear spaces and $p > 0$. A linear operator $T : \mathfrak{X} \to \mathfrak{Y}$ is said to be **absolutely** p**-summing** if $\pi_p(T) < \infty$, where $\pi_p(T) = \inf\{C > 0 : (2.5) \text{ holds}\}$:

$$\left(\sum_{j=1}^n \|Tx_j\|_{\mathfrak{Y}}^p \right)^{\frac{1}{p}} \le C \sup \left\{ \left(\sum_{j=1}^n \big| x^*(x_j) \big|^p \right)^{\frac{1}{p}} : x^* \in \mathfrak{X}^*, \ \|x^*\|_{\mathfrak{X}^*} \le 1 \right\} \tag{2.5}$$

for any $n \in \mathbb{N}$ and $x_1, \dots, x_n \in \mathfrak{X}$. $\pi_p(T)$ is called the **absolutely p-summing norm** of T. We denote by $\Pi_p(\mathfrak{X}, \mathfrak{Y})$ the set of all absolutely p-summing operators from $\mathfrak{X}$ into $\mathfrak{Y}$.

Remark 8. Let $\mathfrak{W}, \mathfrak{X}, \mathfrak{Y}$ and $\mathfrak{Z}$ be Banach spaces.

(1) $(\Pi_p(\mathfrak{X}, \mathfrak{Y}), \pi_p(\cdot))$ is a Banach space for $1 \le p \le \infty$. In particular, $\Pi_\infty(\mathfrak{X}, \mathfrak{Y}) = B(\mathfrak{X}, \mathfrak{Y})$.

(2) Let $1 \le p < \infty$. If $T \in \Pi_p(\mathfrak{X}, \mathfrak{Y})$, $V \in B(\mathfrak{W}, \mathfrak{X})$ and $U \in B(\mathfrak{Y}, \mathfrak{Z})$, then $UTV \in \Pi_p(\mathfrak{W}, \mathfrak{Z})$ and $\pi_p(UTV) \le \|U\|\pi_p(T)\|V\|$.

(3) If $\mathfrak{H}$ and $\mathfrak{K}$ are Hilbert spaces, then $\Pi_2(\mathfrak{H}, \mathfrak{K}) = S(\mathfrak{H}, \mathfrak{K})$ and $\pi_2(T) = \|T\|_\sigma$ for $T \in S(\mathfrak{H}, \mathfrak{K})$ (cf. Wojtaszczyk [1, pp. 243–244] and Diestel, Jarchow and Tonge [1] (1995)).

Note that $(L^0(\Theta), \|\cdot\|_\infty)$ is a normed linear space and $(L^0(\Theta), \|\cdot\|_\infty)^* = fa(\mathfrak{A}, \mathbb{C})$, where $fa(\mathfrak{A}, \mathbb{C})$ is the Banach space of all $\mathbb{C}$-valued finitely additive measures on $\mathfrak{A}$ with the total variation norm $|\cdot|(\Theta)$ and the isometric isomorphism $U : fa(\mathfrak{A}, \mathbb{C}) \to L^0(\Theta)^*$ is given by

$$(U\nu)(f) = \int_\Theta f \, d\nu, \qquad f \in L^0(\Theta), \ \nu \in fa(\mathfrak{A}, \mathbb{C})$$

(cf. Lemma 3.3 in the next section). Here, the variation and the semivariation of a (scalar- or vector-valued) finitely additive measure are defined in the same manner as in Definition 1.1, and the integral with respect to a finitely additive measure is in Dunford and Schwartz [1, Chapter III].

We need one technical lemma.

Lemma 9. *Let $\nu_0 \in fa(\mathfrak{A}, \mathbb{R}^+)$ and define ν on $\mathfrak{A}$ by*

$$\nu(A) = \inf\left\{ \sum_{k=1}^{\infty} \nu_0(A_k) : \{A_k\}_{k=1}^{\infty} \text{ is a countable } \mathfrak{A}\text{-measurable partition of } A \right\}. \tag{2.6}$$

Then $\nu \in ca(\mathfrak{A}, \mathbb{R}^+)$ and $\nu(A) \le \nu_0(A)$ for every $A \in \mathfrak{A}$.

Proof. For $A \in \mathfrak{A}$ denote by $\Pi^c(A)$ the set of all countable $\mathfrak{A}$-measurable partitions of A. ν is finitely additive. In fact, let $A, B \in \mathfrak{A}$ be disjoint. Then we have

$$\begin{aligned}\nu(A \cup B) &= \inf\left\{ \sum_{k=1}^{\infty} \nu_0(C_k) : \{C_k\}_{k=1}^{\infty} \in \Pi^c(A \cup B) \right\} \\ &= \inf\left\{ \sum_{k=1}^{\infty} \bigl\{\nu_0(A \cap C_k) + \nu_0(B \cap C_k)\bigr\} : \{C_k\}_{k=1}^{\infty} \in \Pi^c(A \cup B) \right\}\end{aligned}$$

$$\geq \nu(A) + \nu(B).$$

On the other hand, let $\varepsilon > 0$ be given and choose $\{A_k\}_{k=1}^\infty \in \Pi^c(A)$ and $\{B_k\}_{k=1}^\infty \in \Pi^c(B)$ such that

$$\nu(A) > \sum_{k=1}^\infty \nu_0(A_k) - \frac{\varepsilon}{2}, \qquad \nu(B) > \sum_{k=1}^\infty \nu_0(B_k) - \frac{\varepsilon}{2}.$$

Then we have that

$$\begin{aligned} \nu(A) + \nu(B) &> \sum_{k=1}^\infty \big\{\nu_0(A_k) + \nu_0(B_k)\big\} - \varepsilon \\ &= \sum_{k=1}^\infty \nu_0(A_k \cup B_k) - \varepsilon \geq \nu(A \cup B) - \varepsilon. \end{aligned}$$

Thus $\nu(A) + \nu(B) \geq \nu(A \cup B)$. Consequently ν is finitely additive.

To show that ν is countably additive, let $\{A_n\}_{n=1}^\infty \subseteq \mathfrak{A}$ and $A_n \downarrow \emptyset$. For each $n \in \mathbb{N}$, $A_n = \bigcup_{k=n}^\infty (A_k \backslash A_{k+1})$ and $\{A_k \backslash A_{k+1}\}_{k=n}^\infty \in \Pi^c(A_n)$. Hence

$$\sum_{k=1}^\infty \nu_0(A_k \backslash A_{k+1}) \leq \nu_0(A_1) < \infty.$$

Thus $0 \leq \nu(A_n) \leq \sum_{k=n}^\infty \nu_0(A_k \backslash A_{k+1}) \to 0$ as $n \to \infty$. Therefore ν is countably additive. □

Now we can give some necessary and sufficient conditions for orthogonally scattered dilation.

Theorem 10. *Let $\xi \in ca(\mathfrak{A}, H)$. Then the following conditions are equivalent.*

(1) *ξ has an orthogonally scattered dilation.*

(2) *ξ has a spectral dilation.*

(3) *There exists a constant $C > 0$ such that for any $n \in \mathbb{N}$ and $f_1, \dots, f_n \in L^0(\Theta)$ it holds that*

$$\sum_{j=1}^n \bigg\| \int_\Theta f_j \, d\xi \bigg\|_H^2 \leq C \bigg\| \sum_{j=1}^n |f_j|^2 \bigg\|_\infty. \tag{2.7}$$

(4) *The operator $S_\xi : (L^0(\Theta), \|\cdot\|_\infty) \to H$ is absolutely 2-summing, where S_ξ is defined by $S_\xi f = \int_\Theta f \, d\xi$ for $f \in L^0(\Theta)$.*

(5) *There exists a positive finite measure $\nu \in ca(\mathfrak{A}, \mathbb{R}^+)$ such that*

$$\bigg\| \int_\Theta f \, d\xi \bigg\|_H^2 \leq \int_\Theta |f|^2 \, d\nu, \qquad f \in L^0(\Theta). \tag{2.8}$$

In this case ν *is called a* **2-majorant** *of* ξ.

Proof. (1) $\Rightarrow$ (2). Suppose that ξ has an orthogonally scattered dilation $\{\eta, \mathfrak{H}, J\}$, so that $\xi = J\eta$ where $J : \mathfrak{H} \to H$ is the orthogonal projection. We can assume that

$$\mathfrak{H} = \mathfrak{S}_0\big\{\eta(A) : A \in \mathfrak{A}\big\},$$

the closed subspace of $\mathfrak{H}$ generated by the set $\{\eta(A) : A \in \mathfrak{A}\}$. For $A \in \mathfrak{A}$ let $E(A)$ be the orthogonal projection of $\mathfrak{H}$ onto $\mathfrak{H}(A) = \mathfrak{S}_0\{\eta(A \cap B) : B \in \mathfrak{A}\}$. Then it is not hard to check that $E(\cdot)$ is a countably additive spectral measure in $\mathfrak{H}$ such that

$$\eta(A) = E(A)\eta(\Theta), \qquad A \in \mathfrak{A}.$$

Therefore, $\xi(\cdot) = JE(\cdot)\eta(\Theta)$, i.e., ξ has a spectral dilation.

(2) $\Rightarrow$ (3). Suppose that (2) holds, so that $\xi(\cdot) = SE(\cdot)\psi$ as stated in Definition 6. Note that $\eta(\cdot) = E(\cdot)\psi \in caos(\mathfrak{A}, \mathfrak{H})$ and $\nu(\cdot) = \|\eta(\cdot)\|_{\mathfrak{H}}^2 \in ca(\mathfrak{A}, \mathbb{R}^+)$ with the total mass $\nu(\Theta) = \|\psi\|_{\mathfrak{H}}^2$. Now for any $n \in \mathbb{N}$ and $f_1, \dots, f_n \in L^0(\Theta)$ we have

$$\begin{aligned}\sum_{j=1}^{n} \left\| \int_{\Theta} f_j \, d\xi \right\|_H^2 &= \sum_{j=1}^{n} \left\| S \int_{\Theta} f_j \, d\eta \right\|_H^2 \\ &\le \|S\|^2 \sum_{j=1}^{n} \int_{\Theta} |f_j|^2 \, d\nu \le \|S\|^2 \|\psi\|_{\mathfrak{H}}^2 \left\| \sum_{j=1}^{n} |f_j|^2 \right\|_{\infty}.\end{aligned}$$

Thus (2.7) holds with $C = \|S\|^2 \|\psi\|_{\mathfrak{H}}^2$.

(3) $\Leftrightarrow$ (4). Since $L^0(\Theta)^* = fa(\mathfrak{A}, \mathbb{C})$ it is sufficient to show that for any $n \in \mathbb{N}$ and $f_1, \dots, f_n \in L^0(\Theta)$

$$\left\| \sum_{j=1}^{n} |f_j|^2 \right\|_{\infty} = \sup \left\{ \sum_{j=1}^{n} \left| \int_{\Theta} f_j \, d\nu \right|^2 : \nu \in fa(\mathfrak{A}, \mathbb{C}),\ |\nu|(\Theta) \le 1 \right\}. \tag{2.9}$$

The inequality "$\le$" is seen from

$$\begin{aligned}\text{LHS of (2.9)} &= \sup_{t \in \Theta} \sum_{j=1}^{n} \big| f_j(t) \big|^2 \\ &= \sup_{t \in \Theta} \int_{\Theta} \sum_{j=1}^{n} |f_j|^2 \, d\delta_t \le \text{RHS of (2.9)},\end{aligned}$$

where δ_t is the point mass at $t \in \Theta$, i.e., $\delta_t(A) = 1$ if $t \in A$ and $= 0$ otherwise. To see the converse inequality "$\ge$" let $\nu \in fa(\mathfrak{A}, \mathbb{C})$ be such that $|\nu|(\Theta) \le 1$ and $f_1, \dots, f_n \in L^0(\Theta)$ be arbitrary. We can write them as

$$f_j = \sum_{k=1}^{q} \alpha_{j,k} 1_{A_k}, \qquad 1 \le j \le n,$$

where $\{A_1, \dots, A_q\} \in \Pi(\Theta)$. Then we see that

$$\begin{aligned}\sum_{j=1}^{n} \left| \int_{\Theta} f_j \, d\nu \right|^2 &= \sum_{j=1}^{n} \sum_{k,\ell=1}^{q} \nu(A_k)\overline{\nu(A_\ell)}\alpha_{j,k}\overline{\alpha}_{j,\ell} = \sum_{k,\ell=1}^{q} \nu(A_k)\overline{\nu(A_\ell)}(\mathbf{x}_k, \mathbf{x}_\ell)_{\mathbb{C}^n} \\ &= \left\| \sum_{k=1}^{q} \nu(A_k)\mathbf{x}_k \right\|_{\mathbb{C}^n}^2 \le \left(\sum_{k=1}^{q} \big|\nu(A_k)\big| \|\mathbf{x}_k\|_{\mathbb{C}^n} \right)^2 \\ &\le \max_{1\le k\le q} \|\mathbf{x}_k\|_{\mathbb{C}^n}^2 \left(\sum_{k=1}^{q} |\nu(A_k)| \right)^2 \le \max_{1\le k\le q} \sum_{j=1}^{n} |\alpha_{j,k}|^2 |\nu|(\Theta)^2 \\ &\le \left\| \sum_{j=1}^{n} |f_j|^2 \right\|_{\infty},\end{aligned}$$

where $\mathbf{x}_k = (\alpha_{j,k})_{j=1}^{q} \in \mathbb{C}^n$ for $1 \le k \le q$, and $(\cdot,\cdot)_{\mathbb{C}^n}$ and $\|\cdot\|_{\mathbb{C}^n}$ are the inner product and the norm in $\mathbb{C}^n$, respectively.

(3) ⇒ (5). Assume that (3) holds. Let $L^0_{\mathbb{R}}(\Theta) = \{f \in L^0(\Theta) : f \text{ is } \mathbb{R}\text{-valued}\}$ and define $\rho_0(f)$ for $f \in L^0_{\mathbb{R}}(\Theta)$ by

$$\rho_0(f) = \inf \left\{ \sup_{t\in\Theta} \left\{ Cf(t) + C\sum_{k=1}^{n} |f_k(t)|^2 - \sum_{k=1}^{n} \left\| \int_{\Theta} f_k \, d\xi \right\|_H^2 \right\} : \right. \\ \left. f_1, \dots, f_n \in L^0_{\mathbb{R}}(\Theta),\, n \in \mathbb{N} \right\}.$$

Then we see that ρ_0 is a **sublinear** (i.e., positive homogeneous and subadditive) functional on $L^0_{\mathbb{R}}(\Theta)$ such that

$$C \inf_{t\in\Theta} f(t) \le \rho_0(f) \le C \sup_{t\in\Theta} f(t).$$

It follows from the Hahn-Banach Theorem that there exists a linear functional ρ on $L^0_{\mathbb{R}}(\Theta)$ such that $\rho(f) \le \rho_0(f)$ for $f \in L^0_{\mathbb{R}}(\Theta)$. Hence we have that

$$C \inf_{t\in\Theta} f(t) \le -\rho_0(-f) \le \rho(f) \le \rho_0(f) \le C \sup_{t\in\Theta} f(t).$$

Moreover, (2.7) implies that for $f \in L^0(\Theta)$

$$\rho_0(-|f|^2) \le \sup_{t\in\Theta} \left\{ -C|f(t)|^2 + C|f(t)|^2 - \left\| \int_{\Theta} f \, d\xi \right\|_H^2 \right\} = -\left\| \int_{\Theta} f \, d\xi \right\|_H^2$$

and hence by $-\rho(|f|^2) = \rho(-|f|^2) \le \rho_0(-|f|^2)$

$$\left\| \int_{\Theta} f \, d\xi \right\|_H^2 \le \rho(|f|^2). \tag{2.10}$$

Define ν_0 on $\mathfrak{A}$ by $\nu_0(A) = \rho(1_A)$ for $A \in \mathfrak{A}$. Then it is easily verified that $\nu_0 \in fa(\mathfrak{A}, \mathbb{R}^+)$ since $\nu_0(\Theta) = \rho(1_\Theta) = C$ and ρ is linear. Furthermore, define ν on $\mathfrak{A}$ by (2.6). Then, Lemma 9 implies that ν is countably additive and $\nu(A) \le \nu_0(A)$ for $A \in \mathfrak{A}$.

Now we shall show that (2.8) holds. So let $f = \sum_{j=1}^n \alpha_j 1_{A_j} \in L^0(\Theta)$. For $j = 1, \dots, n$ and $q \in \mathbb{N}$ let $\{A_{j,k}^q\}_{k=1}^\infty \subseteq \mathfrak{A}$ be a countable partition of A_j such that $\lim_{q\to\infty} \sum_{k=1}^\infty \nu_0(A_{j,k}^q) = \nu(A_j)$, which can be a nonincreasing limit. Define for $q, N \in \mathbb{N}$

$$f_N^q = \sum_{j=1}^n \sum_{k=1}^N \alpha_j 1_{A_{j,k}^q}$$

and observe that $f_N^q \to f$, pointwise, as $N \to \infty$ for $q \in \mathbb{N}$ and that by (2.10)

$$\left\| \int_\Theta f_N^q \, d\xi \right\|_H^2 \le \sum_{j=1}^n |\alpha_j|^2 \sum_{k=1}^N \nu_0(A_{j,k}^q), \qquad N, q \in \mathbb{N}.$$

Since for $q \in \mathbb{N}$ we have that

$$\int_\Theta f_N^q \, d\xi = \sum_{j=1}^n \alpha_j \xi\Big(\bigcup_{k=1}^N A_{j,k}^q \Big) \to \sum_{j=1}^n \alpha_j \xi(A_j) = \int_\Theta f \, d\xi \qquad (N \to \infty),$$

it follows that

$$\left\| \int_\Theta f \, d\xi \right\|_H^2 \le \sum_{j=1}^n |\alpha_j|^2 \sum_{k=1}^\infty \nu_0(A_{j,k}^q), \qquad q \in \mathbb{N}.$$

Letting $q \to \infty$, we see that the RHS of the above inequality converges in an nonincreasing manner to $\sum_{j=1}^n |\alpha_j|^2 \nu(A_j) = \int_\Theta |f|^2 \, d\nu$ and the desired inequality (2.8) is obtained.

(5) ⇒ (1). Suppose that (5) holds. Define $m : \mathfrak{A} \times \mathfrak{A} \to \mathbb{C}$ by

$$m(A, B) = \nu(A \cap B) - m_\xi(A, B), \qquad A, B \in \mathfrak{A},$$

where $m_\xi(A, B) = (\xi(A), \xi(B))_H$. To see that m is a positive definite kernel let $n \in \mathbb{N}$, $A_1, \dots, A_n \in \mathfrak{A}$ and $\alpha_1, \dots, \alpha_n \in \mathbb{C}$. If $f = \sum_{i=1}^n \alpha_i 1_{A_i} \in L^0(\Theta)$, then (2.8) implies that

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \overline{\alpha}_j m_\xi(A_i, A_j) &= \left\| \sum_{i=1}^n \alpha_i \xi(A_i) \right\|_H^2 = \left\| \int_\Theta f \, d\xi \right\|_H^2 \\ &\le \int_\Theta |f|^2 \, d\nu = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \overline{\alpha}_j \nu(A_i \cap A_j). \end{aligned}$$

Hence, it follows that

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\overline{\alpha}_j m(A_i,A_j)=\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\overline{\alpha}_j\big[\nu(A_i\cap A_j)-m_\xi(A_i,A_j)\big]\geq 0.$$

Thus, m is a positive definite kernel and hence there exist the RKHS $\mathfrak{H}_m$ of m and a $\zeta\in ca(\mathfrak{A},\mathfrak{H}_m)$ such that $m=m_\zeta$, i.e., $m(A,B)=(\zeta(A),\zeta(B))_{\mathfrak{H}_m}$ for $A,B\in\mathfrak{A}$. Put $\mathfrak{H}=H\oplus\mathfrak{H}_m$, the direct sum, and let $J:\mathfrak{H}\to H$ be the orthogonal projection, where H is identified with $H\oplus\{0\}\subset\mathfrak{H}$. Now define $\eta\in ca(\mathfrak{A},\mathfrak{H})$ by

$$\eta(A)=\big(\xi(A),0\big)+\big(0,\zeta(A)\big)=\big(\xi(A),\zeta(A)\big),\qquad A\in\mathfrak{A}.$$

Then we see that η is orthogonally scattered since for disjoint $A,B\in\mathfrak{A}$ it holds that

$$\begin{aligned}\big(\eta(A),\eta(B)\big)_{\mathfrak{H}}&=\big((\xi(A),\zeta(A)),(\xi(B),\zeta(B))\big)_{\mathfrak{H}}\\&=\big(\xi(A),\xi(B)\big)_H+\big(\zeta(A),\zeta(B)\big)_{\mathfrak{H}_m}\\&=m_\xi(A,B)+m(A,B)\\&=\nu(A\cap B)=0.\end{aligned}$$

Moreover, it holds that $\xi=J\eta$. Therefore $\{\eta,\mathfrak{H},J\}$ is an orthogonally scattered dilation of ξ. □

The condition (2.7) is of particular interest and is closely related to the following Grothendieck inequality (cf. Grothendieck [1, p. 59]). The proof may be found in several books such as Diestel [1, pp. 174–180], Jameson [1, pp. 104–108], Lindenstrauss and Tzafriri [1, pp. 68–69] and Wojtaszczyk [1, pp. 210–211].

Theorem 11 (Grothendieck inequality). *Let $n\in\mathbb{N}$ be arbitrary and $\mathfrak{H}$ be any Hilbert space. If $(a_{jk})_{j,k=1}^n$ is an $n\times n$ matrix such that*

$$\left|\sum_{j,k=1}^{n}a_{jk}s_jt_k\right|\leq\alpha\max_{1\leq j,k\leq n}|s_j||t_k|\tag{2.11}$$

for every $s_1,\dots,s_n,t_1,\dots,t_n\in\mathbb{C}$, then, for every $x_1,\dots,x_n,y_1,\dots,y_n\in\mathfrak{H}$ it holds that

$$\left|\sum_{j,k=1}^{n}a_{jk}(x_j,y_k)_{\mathfrak{H}}\right|\leq K_G\,\alpha\max_{1\leq j,k\leq n}\|x_j\|_{\mathfrak{H}}\|y_k\|_{\mathfrak{H}},\tag{2.12}$$

where $K_G>0$ is the **Grothendieck universal constant**, *independent of n and $\mathfrak{H}$, and $(\cdot,\cdot)_{\mathfrak{H}}$ and $\|\cdot\|_{\mathfrak{H}}$ are the inner product and the norm in $\mathfrak{H}$, respectively.*

Using the above inequality we can derive the condition (2.7) as follows.

Lemma 12. *Let* $\xi \in ca(\mathfrak{A}, H)$*. Then there exists a* $C > 0$ *such that for any* $q \in \mathbb{N}$ *and* $f_1, \dots, f_q \in L^0(\Theta)$

$$\sum_{\ell=1}^{q} \left\| \int_\Theta f_\ell \, d\xi \right\|_H^2 \leq C \left\| \sum_{\ell=1}^{q} |f_\ell|^2 \right\|_\infty . \tag{2.13}$$

Proof. Let $f = \sum_{j=1}^n \alpha_j 1_{A_j},\ g = \sum_{k=1}^n \beta_k 1_{A_k} \in L^0(\Theta)$. Then we have that

$$\begin{aligned}\Big| \sum_{j,k=1}^{n} \alpha_j \overline{\beta}_k \big(\xi(A_j), \xi(A_k)\big)_H \Big| &= \left| \left(\int_\Theta f \, d\xi, \int_\Theta g \, d\xi \right)_H \right| \\ &\leq \left\| \int_\Theta f \, d\xi \right\|_H \left\| \int_\Theta g \, d\xi \right\|_H \\ &\leq \|\xi\|(\Theta)^2 \|f\|_\infty \|g\|_\infty \\ &= \|\xi\|(\Theta)^2 \max_{1 \leq j,k \leq n} |\alpha_j||\beta_k|. \end{aligned} \tag{2.14}$$

Let $f_1, \dots, f_q \in L^0(\Theta)$ and write them as

$$f_\ell = \sum_{j=1}^{n} \alpha_{\ell,j} 1_{A_j}, \qquad 1 \leq \ell \leq q,$$

where $\{A_1, \dots, A_n\} \in \Pi(\Theta)$. Then it holds that

$$\begin{aligned}\sum_{\ell=1}^{q} \left\| \int_\Theta f_\ell \, d\xi \right\|_H^2 &= \sum_{\ell=1}^{q} \sum_{j,k=1}^{n} \alpha_{\ell,j} \overline{\alpha}_{\ell,k} \big(\xi(A_j), \xi(A_k)\big)_H \\ &= \sum_{j,k=1}^{n} \left(\sum_{\ell=1}^{q} \alpha_{\ell,j} \overline{\alpha}_{\ell,k} \right) \big(\xi(A_j), \xi(A_k)\big)_H. \end{aligned} \tag{2.15}$$

Considering the $n \times n$ matrix $((\xi(A_j), \xi(A_k))_H)_{j,k=1}^n$ and the vectors $\mathbf{x}_j = (\alpha_{\ell,j})_{\ell=1}^q \in \mathbb{C}^q,\ 1 \leq j \leq n$ and noting that $(\mathbf{x}_j, \mathbf{x}_k)_{\mathbb{C}^q} = \sum_{\ell=1}^q \alpha_{\ell,j} \overline{\alpha}_{\ell,k}$, we can apply Theorem 11, so that we have by (2.11), (2.12), (2.14) and (2.15)

$$\begin{aligned}\sum_{\ell=1}^{q} \left\| \int_\Theta f_\ell \, d\xi \right\|_H^2 &\leq K_G \|\xi\|(\Theta)^2 \max_{1 \leq j \leq n} \|\mathbf{x}_j\|_{\mathbb{C}^q}^2 \\ &= K_G \|\xi\|(\Theta)^2 \max_{1 \leq j \leq n} \sum_{\ell=1}^{q} |\alpha_{\ell,j}|^2 \end{aligned}$$

$$= K_G \|\xi\|(\Theta)^2 \left\| \sum_{\ell=1}^{q} |f_\ell|^2 \right\|_\infty,$$

which implies that (2.13) holds with $C = K_G\|\xi\|(\Theta)^2$. □

It follows from Theorem 10 and Lemma 12 that:

Theorem 13. *Every* $\xi \in ca(\mathfrak{A}, H)$ *has an orthogonally scattered dilation.*

In this chapter we assumed that H is separable. However, Theorem 13 is true for arbitrary Hilbert space H. It follows from Lemma 12 and the proof of Theorem 10 that, if $\xi \in ca(\mathfrak{A}, H)$ has a spectral dilation $\xi = SE\psi$, then we can take S and ψ to satisfy $\|S\|\|\psi\|_{\mathfrak{H}} \le \sqrt{K_G}\|\xi\|(\Theta)$.

3.3. Gramian orthogonally scattered dilations

In this section we consider the gramian orthogonally scattered dilation of $X = L_0^2(\Omega\,; H)$-valued measures. Although every Hilbert space-valued (bounded) measure has an orthogonally scattered dilation, not every X-valued measure has a gramian orthogonally scattered dilation. So we shall give some necessary and sufficient conditions for that. As before let $(\Theta, \mathfrak{A})$ be a measurable space.

Definition 1. An X-valued measure $\xi \in ca(\mathfrak{A}, X)$ is said to have a **gramian orthogonally scattered dilation** if there exist a normal Hilbert $B(H)$-module Y containing X as a closed submodule and an $\eta \in cagos(\mathfrak{A}, Y)$ such that $\xi = P\eta$, where $P : Y \to X$ is the gramian orthogonal projection. The triple $\{\eta, Y, P\}$ is also called a **gramian orthogonally scattered dilation** of ξ.

Let $\xi \in ca(\mathfrak{A}, X)$ and consider $L_0^2(\Omega)$-valued measures $\xi_\phi \in ca(\mathfrak{A}, L_0^2(\Omega))$ for $\phi \in H$, i.e., $\xi_\phi(\cdot) = \langle \xi(\cdot), \phi\rangle_H$ (cf. (1.2) and Definition 1.4 (3)). It follows from Theorem 2.13 that, for each $\phi \in H$, ξ_ϕ has an orthogonally scattered dilation $\{\eta_\phi, \mathfrak{K}_\phi, J_\phi\}$, i.e., there exist a Hilbert space $\mathfrak{K}_\phi$ and $\eta_\phi \in caos(\mathfrak{A}, \mathfrak{K}_\phi)$ such that $\xi_\phi = J_\phi \eta_\phi$, where $J_\phi : \mathfrak{K}_\phi \to L_0^2(\Omega)$ is the orthogonal projection. If $\dim H = q < \infty$, then, letting $\{\phi_1, \dots, \phi_q\}$ be an orthonormal basis in H, we put $\mathfrak{K} = \bigoplus_{j=1}^q \mathfrak{K}_{\phi_j}$, $\eta = (\eta_{\phi_j})_{j=1}^q$ and $J = \bigoplus_{j=1}^q J_{\phi_j}$. So we see that $\{\eta, \mathfrak{K}, J\}$ is a gramian orthogonally scattered dilation of ξ. If $\dim H = \infty$, we have to work a bit more.

Let $\xi \in ca(\mathfrak{A}, X)$. $L^0(\Theta\,; B(H))$ denotes the set of all $B(H)$-valued $\mathfrak{A}$-simple functions on Θ. The **sup-norm** $\|\Phi\|_\infty$ of $\Phi \in L^0(\Theta\,; B(H))$ is defined by $\|\Phi\|_\infty = \sup_{t\in\Theta} \|\Phi(t)\|$. The **integral** of $\Phi \in L^0(\Theta\,; B(H))$ with respect to ξ over $A \in \mathfrak{A}$ is

-256- - -* ! - " //$ B 1"
E < = 2 ! 3 211 !
(% %B !1 4 1

(%B

called a homomorphism or a linear transformation of U into V,

if (i) $f(\alpha+\beta) = f(\alpha)+f(\beta), \forall \alpha, \beta \in U$

and

(ii) $f(a\alpha) = af(\alpha) \; \forall a \in F, \forall \alpha \in V$.

- 2021 09/24

Hilbert and Banach Space.../

Hilbert and Banach Space....pdf

Pág. 79 (92) 3.2 Orthogonally scattered dilations

Pág. 92 (105) 3.3 Gramian orthogonally scattered dilations

- 2021 09 24/

Finite Element Analysis Applications/

0128099526.pdf

Pág. XIII (3)

Finite element analysis (FEA)

Pág. 32(36) Chapter 2

Decomposition

Functional Requirements → FRs

(%

Pág. 33(37) Figure 2.2
Complex of a finite element analysis model.

→ Geometric shapes,
Physics behaviors,
Dynamics,
Uncertainties,
Assemblies.

Pág. 39(42) 2.3.5 Geometric [illegible] Shapes.

Pág. 51(55) Decomposition
Table 2.2 Classification of Meshes

Pág. 53(57) Table 2.3
Definition of Aspect Ratios of Different Element Types.

Pág. 54(58) Figure 2.19
Example of Jacobian check in quadratic-triangular element

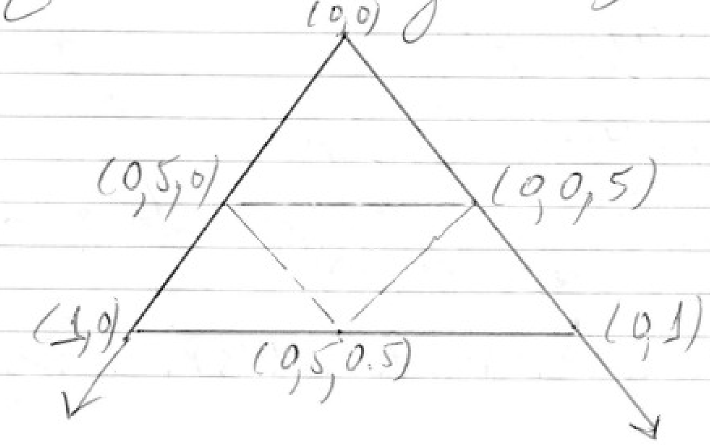

(%

Important:

Pág. 41(45) 2.5 Axiomatic Design Theory. (ADT)

ADT uses two basic axioms as several other corollaries and theorems, to support the design theory.

First axiom is to maintain the independence of FRs and the

Second axiom is to minimize the information content of the resulting design solution.

The act of design consists of the mapping of FRs into design parameters (DPs), and eventually to arrive at a solution, which satisfies these two axiomas (axioms).

_20210924/Physics of Nuclear Reactors/ 012822411X.pdf

Pág. 34(44) Fig 1.21

⇒ Parallelogram representation of the velocity vectors.

transformations using the

E :G & *< 8H ?=25 5=2 9*5
(%' " <

Since the total kinetic energy is conserved, taking the velocities obtained above for the neutron and for the target, before and after the elastic collision, the energy conservation equation for the CM system is given by,

$$\frac{1}{2}(1)\left(\frac{Av_L}{A+1}\right)^2+\frac{1}{2}(A)\left(\frac{v_L}{A+1}\right)^2=\frac{1}{2}(1)\left(v_C'\right)^2+\frac{1}{2}(A)\left(V_C'\right)^2 \tag{1.32}$$

Using the relation $v_C' = AV_C'$ in the above, the velocities in the CM system after the collision can be obtained in terms of the neuron velocity in the L system before collision, that is,

$$v_C' = \frac{Av_L}{A+1} \tag{1.33}$$

and

$$V_C' = \frac{v_L}{A+1} \tag{1.34}$$

Comparing the velocities after the collision with those before the collision in the CM system, both in terms of the initial neutron velocity in the LAB system, it is seen that the magnitudes of the CM system velocities do not change on account of the collision, though there is a change in the direction, indicated as φ in Fig. 1.20.

When observed from the LAB system, the CM system is moving with a velocity V_{CM}. The final velocities of the neutron in the LAB and CM systems, namely, v_L' and v_C' may therefore be obtained by the transformation using the parallelogram law of vectors, as given in the Fig. 1.21.

By vector addition we get

$$v_L'^2 = V_{CM}^2 + v_C'^2 + 2(V_{CM})(v_C')\cos\varphi \tag{1.35}$$

$$v_L'^2 = \left(\frac{v_L}{A+1}\right)^2 + \left(\frac{Av_L}{A+1}\right)^2 + 2\left(\frac{v_L}{A+1}\right)\left(\frac{Av_L}{A+1}\right)\cos\varphi \tag{1.36}$$

$$v_L'^2 = v_L^2\frac{A^2+2A\cos\varphi+1}{(A+1)^2} \tag{1.37}$$

The above equation connects the velocity of the neutron after the collision to that before the collision in the LAB system. The appearance of squares of the velocities obviously indicates the relationship in kinetic energies after and before the collision. Thus

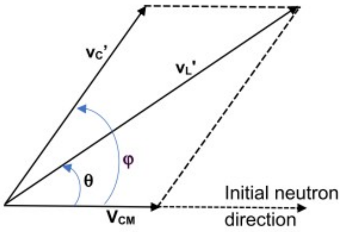

FIG. 1.21 Parallelogram representation of the velocity vectors.

$$E_L' = E_L \frac{A^2 + 2A\cos\varphi + 1}{(A+1)^2} \tag{1.38}$$

The maximum of the kinetic energy $E_{\max}$ after the collision occurs when $\cos\varphi = 1$, that is, when the scattering angle is zero. Thus

$$E_{\max} = E_L \tag{1.39}$$

The minimum energy $E_{\min}$ occurs when $\cos\varphi = -1$, that is, when the scattering angle $\varphi = \pi$ (=180°). Thus

$$E_{\min} = E_L \frac{A^2 - 2A + 1}{(A+1)^2} = \frac{(A-1)^2}{(A+1)^2} E_L = \alpha E_L \tag{1.40}$$

where

$$\alpha = \frac{(A-1)^2}{(A+1)^2} \tag{1.41}$$

Eq. (1.39) indicates the obvious fact that the projectile cannot gain energy by colliding with a stationary target, and at most, it can scatter without an energy loss. The minimum energy $E_{\min}$ depends on α, which is a property of the target through its mass. It is also clear that, only when $A = 1$, that is, when the target is ^{1}H, $E_{\min} = 0$, which means the neutron can lose all its energy in the elastic collision. For elastic collision with any other nuclide, the neutron has no chance to lose its entire energy. The energy lost in the collision is given by $E_L - E_L'$. The maximum energy loss occurs when the scattered neutron comes out with the minimum energy $= \alpha E_L$, which is given by

$$E_L - E_{\min} = E_L - \alpha E_L = (1-\alpha)E_L \tag{1.42}$$

As A increases, the value of α increases toward unity, indicating lower energy loss. Thus elastic scattering with a heavy nuclide (say ^{238}U) will result in little energy loss, and substantial energy loss is possible only with light nuclides.

In general, if the incident neutron energy is E, then the probability that the elastically scattered neutron has its energy in the range E' and $E' + \mathrm{d}E'$ is given by

$$P_{\mathrm{el}}(E \to E')\mathrm{d}E' = \frac{\mathrm{d}E'}{(E - \alpha E)} \tag{1.43}$$

This is based on the assumption that the elastic scattering is isotropic in the CM system, which is practically true except for very light nuclides. The following relationships, not derived herein, are also important results from the elastic scattering kinematics:

Referring to the velocity vectors in the LAB and the CM systems, it is easy to see that

$$\cos\theta = \frac{A\cos\varphi + 1}{\sqrt{(A^2 + 2A\cos\varphi + 1)}} \tag{1.44}$$

is the cosine of the scattering angle in the LAB system. The average value of this quantity is approximately given by

$$\overline{\mu_L} = \overline{\cos\theta} = \frac{2}{3A} \tag{1.45}$$

The energy of a scattered neutron with initial energy E_1, will be in the range between E_1 and αE_1, as seen above. If the energy after scattering is E_2. The energy decrement is E_1–E_2. The decrement in the logarithms of the two energies, namely, $\ln E_1 - \ln E_2$ or $\ln(E_1/E_2)$ is called the "logarithmic energy decrement." The average of this is an important quantity, called the "average logarithmic energy decrement" of elastic scattering, given as

$$\xi = 1 + \frac{(A-1)^2}{2\mathrm{A}} \ln\left(\frac{A-1}{A+1}\right) = 1 + \frac{\alpha}{1-\alpha} \ln\alpha \tag{1.46}$$

It is interesting to note that the average logarithmic energy decrement of a neutron in an elastic scattering collision with a target of mass A (i.e., A times as heavy as the neutron) depends on A, but not on its initial energy. In other words, the ratio of the initial to the final energies, that is, E_1/E_2, of the neutron in an elastic scattering, on an average, is always the same.

1.12.2 The moderator

All the quantities, called the "scattering parameters," derived above, are important, specifically, in the selection of materials for a "moderator," whose function in a reactor is to slow down fast neutrons to very low speeds. As seen in Eqs. (1.40)–(1.43), only a light nuclide with $\alpha << 1$ appears suitable for the purpose. Further, the average number of elastic collisions a neutron needs for a nuclide, such that its energy is reduced from a given initial energy E_1 to a specified final energy E_2 is given by

$$\text{Average number of collisions} = \frac{\ln\left(\frac{E_1}{E_2}\right)}{\xi}. \tag{1.47}$$

TABLE 1.4 Number of elastic collisions to thermalize a 2 MeV neutron.

Material	α	ξ	Number of collisions to thermalize
H	0.0	1.0	18
D	0.11	0.72	25
Be	0.64	0.21	85
C	0.72	0.16	115
O	0.78	0.12	150
U-238	0.98	0.008	2172

(%

Important:

Pág. 41(45) 2.5 Axiomatic Design Theory. (ADT)

ADT uses two basic axioms as several other corollaries and theorems, to support the design theory.

First axiom is to maintain the independence of FRs and the

Second axiom is to minimize the information content of the resulting design solution.

The act of design consists of the mapping of FRs to design parameters (DPs), and eventually to arrive at a solution, which satisfies these two axioms (axioms).

_20210924/Physics of Nuclear Reactors/
01282241X.pdf

Pág. 34(44) Fig 1.21

⇒ Parallelogram representation of the velocity vectors.

transformations using the

parallelogram law of vectors,
as given in the Fig. 1 2 1

By vector addition we get

$$v_L'^2 = V_{CM}^2 + v_C'^2 + 2(V_{CM})(v_C')\cos\varphi$$

$$v_L'^2 = \left(\frac{v_L}{A+1}\right)^2 + \left(\frac{A v_L}{A+1}\right)^2 + 2\left(\frac{v_L}{A+1}\right)\left(\frac{A v_L}{A+1}\right)\cos\varphi$$

$$v_L'^2 = v_L^2 \, \frac{A^2 + 2A\cos\varphi + 1}{(A+1)^2}$$

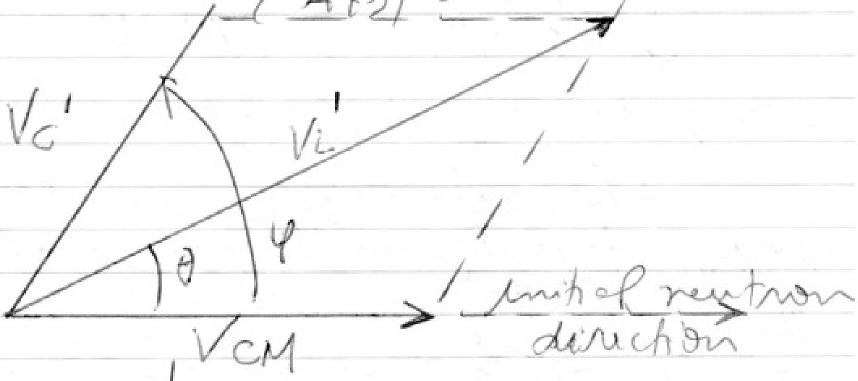
V_C'
V_L'
θ
φ
V_{CM}
initial neutron direction

20210924 | Quantum Mechanics in Potential.... |
Quantum Mechanics in Potential....pdf
Pag. 29(50) Chapter 3
Nuclear Forces ↙

→ the Interactions between Nucleons.

-256- - -* ! - " 6 ! ! & !
5 1 ! 2 " 211 ! 1"
(/ $!1 % 8 ! <

Chapter 3

Nuclear Forces

3.1 The Interactions between Nucleons

Gluons and quarks are confined in a limited space, and the forces in nuclei are produced due to the interchange of mesons whose field is similar to the electromagnetic field, yet is of a much shorter range. Keeping in mind the above fact, we will calculate the interaction radius of nucleons in nuclei. For the $^{197}_{79}Au$ nucleus, both neutrons and protons are diffusing in the region $R_d = 2R$ of physical vacuum defined by radius R in the coordinates of the center of mass. The formula A[21] for calculating the distance of interaction of an external proton with bound energy 7.73 MeV can be given as follows:

$$R_d = 2R = \frac{\hbar}{\sqrt{2m|E_B|}} = 1.638 \cdot 10^{-15}\,\text{m} = 1.638\,\text{fm}. \tag{3.1}$$

Yukawa declared that the interaction with a short interaction radius was generated for field quanta with negligible rest mass. The electromagnetic field is occupied by photons of zero rest mass that satisfy the wave equation

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) U(r,t) = 0, \tag{3.2}$$

which can be written as

$$-p^2 + \frac{1}{c^2}E^2 = 0. \tag{3.3}$$

The corresponding relativistic equation for particles having the rest mass is

$$-p^2 - m^2c^2 + \frac{1}{c^2}E^2 = 0. \tag{3.4}$$

After the introduction of the quantum mechanics operators

$$\vec{p} = -i\hbar\vec{\nabla}, \quad E = i\hbar\frac{\partial}{\partial t}, \tag{3.5}$$

we can transform it into the below Klein–Gordon equation for a spinless relativistic free particle

$$\left(\nabla^2 - \frac{m^2c^2}{\hbar^2} - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)\phi = 0. \tag{3.6}$$

This wave equation for the time-independent meson field

$$(\nabla^2 - \mu^2)\phi = 0, \quad \mu = \frac{mc}{\hbar} \tag{3.7}$$

was solved similar to the following the spherically symmetric solution

$$\phi = g\frac{e^{-\mu\cdot r}}{r}, \tag{3.8}$$

where g is an undetermined constant which plays the same role as the charge in the electrostatic case. In this case, it depends on the source at the origin but in a different way, where the mass and binding energy of the nucleon must be included (3.1). The potential energy between two electrical charges is $V_C = qU = q^2/4\pi \cdot r$. Then, [1] can be written similarly for the nucleon interaction potential of strength g

$$V(r) = g\frac{e^{-\mu\cdot r}}{r}, \tag{3.9}$$

arising from the continuous transfer of virtual mesons of rest mass m between two nucleons.

The interactions can be classified according to the value of a characteristic dimensionless constant related through a coupling constant. The strong interactions increase the interaction cross section and reduce the interaction time [2, 3]. The interaction strength gives a dimensionless constant for the strong interactions, i.e., $g^2/\hbar c \approx 1, \dots, 10$. For the electromagnetic interaction, $e^2/\hbar c = 1/137$. The weak interaction has the potential shown in (3.9) and

$g_w^2 \approx 10^{-5}$. The gravitation interaction between two protons can be represented by the constant [3] $Gm_p/\hbar c \approx 6.0 \cdot 10^{-39}$, which is so small that it can be usually neglected in particle physics. The duration of expedition of virtual mesons from nucleon to nucleon can be found from the uncertainty principle

$$\Delta t \Delta E \approx \hbar, \quad \Delta E \approx mc^2, \quad \Delta r \approx c\Delta t \approx \frac{\hbar}{mc}, \tag{3.10}$$

and the range of force Δr is equal to the Compton wave length. Calculating the range of potentials given in (3.9) arising from the different types of meson exchange and considering

$$\mu \cdot r_0 = 1, \quad r_0 = \frac{1}{\mu} = \frac{\hbar}{mc}, \tag{3.11}$$

we obtain a result similar to Δr in (3.10). Considering the mass [2] as $135\,\mathrm{MeV/c^2}$ for the lightest pi meson π^0, from (3.11) we obtain the range for the force between nucleons as 1.429 fm. The calculations (3.1) using mass and binding energies of nucleons, diffusing in excited physical vacuum, give a similar result, i.e., 1.638 fm. However, the result in formula (3.1) depends on the quantum states of sources of virtual mesons that are in the nuclei. More massive virtual mesons [3] $(\pi, \sigma, \rho, \omega)$ with spins and isospins (J^π, T) are described by the different ranges and kinds of potential, which can also play an important part in the interactions that occur between nucleons. The term virtual indicates that the meson cannot be released from the nucleus when the energy is less than m_0c^2. Emission of virtual mesons and re-capture must occur under the violation of conservation of energy within a short time interval Δt. The resulting potential between two nucleons [2] depends on the spin, i.e.,

$$V(\vec{r}) = V_C(r) + V_{C\cdot S}(r)\frac{\vec{S}_1 \cdot \vec{S}_2}{\hbar^2} + V_T(r)\frac{\hat{S}_{12}}{\hbar^2} + V_{SL}(r)(\vec{L} \cdot \vec{S}). \tag{3.12}$$

The potential presented is invariant with regard to the rotations, mirror reflections, and interchange of the spins of the interacting nucleons. For this case, the states of interacting nucleons can be classified using quantum numbers $\vec{J} = \vec{L} + \vec{S}$ representing the squares of the sum of orbital momentum $\vec{L}$ and total spin $\vec{S} = \vec{s}_1 + \vec{s}_2$ moments, their projection M_J, the sum of spins S and the parity $(-1)^L$. The total orbital quantum number L can be used when the tensor of non-central forces is not included for singlet states of ${}^{2S+1}L_j$, $S = 0$.

Here, we have $V_C(r)$, the spin-independent central potential; $V_{CS}(r)$, the spin-dependent central potential, $V_T(r)$, tensor potential, V_{LS}, the two-body spin–orbit potential, and $\vec{s} = \frac{1}{2}\hat{\sigma} \cdot \hbar$, where $\vec{\sigma}$ is the Pauli spin matrix. The non-central interaction represents the tensor potential which depends on the angles between the spin vectors of the neutron and the proton and the radius vector. Tensor forces are proportional to the below non-central tensor operator

$$\hat{S}_{12} = \frac{3}{r^2}(\vec{\sigma}_1 \cdot \vec{r})(\vec{\sigma}_2 \cdot \vec{r}) - \vec{\sigma}_1 \cdot \vec{\sigma}_2, \tag{3.13}$$

which is equal to zero for the singlet state $S = 0$. The tensor potential $\hat{S}_{12}$ can explain the electric quadruple moment of the deuteron described by mixed triplet states, $a \cdot \psi({}^3S_1) + b \cdot \psi({}^3D_1)$. For the normalized function of triple states [1] $a^2 + b^2 = 1$, $b^2 = 0.004$, influence of state ${}^3D_1$ is small, about 4%. The spin–orbit potential V_{SL} can be expressed as follows:

$$\begin{aligned} V_{SL} &= V_{SL}(r)(\vec{L} \cdot \vec{S}), (\vec{L} \cdot \vec{S}) \\ &= \frac{1}{2}[J(J+1) - L(L+1) - S(S+1)]. \end{aligned} \tag{3.14}$$

This last expression is equal to zero when $S = 0$, $(L = J)$ or $L = 0$, $(S = J)$. Using the total spin dependence and tensor terms, (3.12) can be rewritten [2] as follows:

$$(\vec{s}_1 \cdot \vec{s}_2) = \frac{1}{4}[2S(S+1) - 3], \tag{3.15}$$

$$\hat{S}_{12} = 6(\vec{n}\vec{S})^2 - 2\vec{S}^2. \tag{3.16}$$

From the above formulas we see that the spin-dependent expectation values of $(\vec{s}_1 \cdot \vec{s}_2)$ and the tensor force operators are different for singlet $S = 0$ and triplet $S = 1$ states. The operator $\hat{S}_{12}$ depends on the angle between the unit vector $\vec{n}$ directed from one nucleon to another and $\vec{S}$. The wave function Y_{lSJM} that depends on the eigenfunctions of the orbital moment $Y_{lm}(\vartheta, \varphi)$ and the spin moment $\chi_{S\mu}$ as follows:

$$Y_{lSJM} = \sum_{m+\mu=M} (lmS\mu|JM) Y_{lm} \chi_{S\mu} \tag{3.17}$$

is expressed by Clebsch–Gordan coefficients. Interaction of the potential (3.12) and energy of the two potentials depend on the quantum states ISJ. The wave function of the deuteron representing the bound states of both neutron and proton is a linear combination of the following functions [2] that are expressed by orthonormal functions Y_{lSJM}.

$$\psi(^3S_1) = \frac{u(r)}{r} Y_{011M}, \quad \psi(^3D_1) = \frac{\omega(r)}{r} Y_{211M} \tag{3.18}$$

Then the required normalization of the deuteron function

$$\int |\psi_0|^2 d\vec{r} = 1 \tag{3.19}$$

can be obtained by using

$$\int_0^\infty (u^2 + \omega^2) dr = 1 \tag{3.20}$$

and the probabilities of deuterons in states $^3S_1$ and $^3D_1$, i.e.,

$$P(^3S_1) = \int_0^\infty u^2 dr, \;\; P(^3D_1) = \int_0^\infty \omega^2 dr, \;\; P(^3S_1) + P(^3D_1) = 1. \tag{3.21}$$

The Schrödinger equation for the neutron–proton bounded system or the deuteron [2]

$$[-(\hbar^2/M)\Delta + V_C + V_T S_{12} - E]\psi = 0 \tag{3.22}$$

was solved for the function $\psi_d = a \cdot \psi(^3S_1) + b \cdot \psi(^3D_1)$ only at large distances for the spin function $r > r_0$, when the total spin coincides with the projection of the total moment of the deuteron. The asymptotes of solutions are expressed by the binding energy E, i.e.,

$$u(r) = C_U \exp[-\alpha \cdot r], \tag{3.23}$$

$$\begin{gathered}\omega(r) = C_W \left[1 + \frac{3}{\alpha r} + \frac{3}{(\alpha r)^2}\right] \exp[-\alpha \cdot r], \\ \alpha = \sqrt{2ME}/\hbar, \quad E = 2.226\,\mathrm{MeV}.\end{gathered} \tag{3.24}$$

The range of central (3.22) and tensor forces can be approximately expressed by the radius of the quantum diffusion A[21] region of the nucleus, $R = \hbar/\sqrt{2ME}$ which is defined as $R = 2.158 \cdot 10^{-15}$ fm for deuterons or as 2.158 fm for the interaction energy E of nucleons.

At small distances, the asymptotic solutions are $u \cong r$ and $\omega \cong r^3$ and also r^2 in Ref. A[19] for $l = 0$, when the relativistic correction for kinetic energy is included. The component of the quadruple moment operator for the deuteron in the case $M = J = 1$ is given as

$$Q = \frac{1}{4} e(3z^2 - r^2) \tag{3.25}$$

and can be used for the calculation of the expectation value, i.e.,

$$\langle Q \rangle = \int_0^\infty \psi_d{}^* Q \psi_d dr, \tag{3.26}$$

which for $P(^3D_1) = 0.04$ gives [1, 2] $\langle Q \rangle = 2.82 \cdot 10^{-27} \mathrm{e} \cdot \mathrm{cm}^2$.

3.2 The Shell Model and Mean Field Potentials

The simplest shell model of hydrogen atom was proposed by Niels Bohr by introducing the quantization of angular momentum

$$mv\ r = l\hbar, \quad l = 1, 2, 3, \ldots \tag{3.27}$$

for an electron moving with velocity v at a distance r from the positive proton or any nucleus with charge eZ. For a circular orbit the centrifugal forces are equal to the attractive forces as shown below:

$$\frac{mv^2}{r} = \frac{Ze^2}{r^2}. \tag{3.28}$$

(%/

parallelogram law of vectors, as given in the Fig. 1 2 1

By vector addition we get

$$v_L'^2 = V_{CM}^2 + v_C'^2 + 2(V_{CM})(v_C')\cos\varphi$$

$$v_L'^2 = \left(\frac{v_L}{A+1}\right)^2 + \left(\frac{A v_L}{A+1}\right)^2 + 2\left(\frac{v_L}{A+1}\right)\left(\frac{A v_L}{A+1}\right)\cos\varphi$$

$$v_L'^2 = v_L^2 \, \frac{A^2+2A\cos\varphi+1}{(A+1)^2}$$

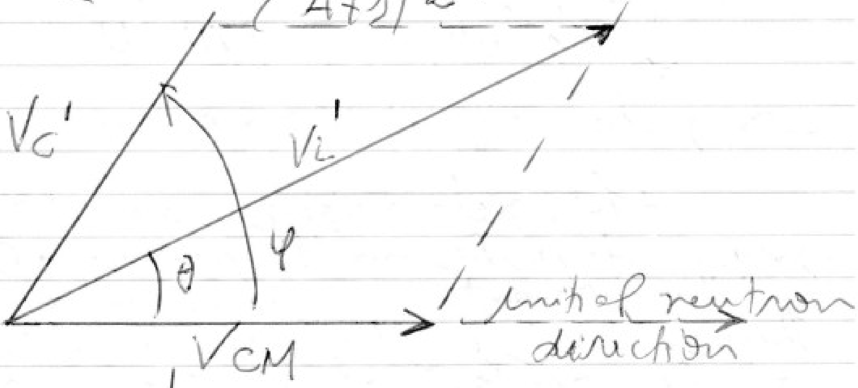
V_C'
V_L'
φ
θ
V_{CM}
initial neutron direction

-20210924|
Quantum Mechanics in Potential... |
Quantum Mechanics in Potential....pdf
Pag. 29(50) Chapter 3
Nuclear Forces

→ the Interactions between Nucleons.

Gluons and quarks are confined in a limited space and the forces in nuclei are produced due to the interchange of mesons whose field is similar to the electromagnetic field, yet is of a much shorter range.

✡ ⇑ gravity.

(geometria dessas forças).

Pág. 183 (204) Chapter 19

Stability of Nuclei

→ relativistic corrections

The relativistic corrections of the mass depende on the principal, and the orbital quantum numbers and do not have any influence on the definition of the constant of the spin-orbit interaction.

— 20210924/

Solving Linear PDE Spectra/

Solving Linear PDE Spectra.pdf

Pág. 349 (366) – Chapter 11

Quantum Mechanics =

-256- - -* ! - "
A l? ! l ! ! l4 ! l=> ! l 1 ! 1"
(%'/ %BB !1 6 ! !

Chapter 11

Quantum Mechanics

In this chapter we shall show how the mathematical theory developed in the book can be applied to questions in quantum mechanics. We shall begin with simpler situations and work up to more general systems.

11.1 The Schrödinger operator for a particle

The energy operator corresponding to a single particle subjected to a static potential is of the form

$$H = -\frac{h^2}{8\pi^2 m}\Delta + q(x) = \frac{h^2}{8\pi^2 m}(\mathrm{D}_1^2 + \mathrm{D}_2^2 + \mathrm{D}_3^2) + q(x) \tag{11.1.1}$$

in E^3, where h is Planck's constant, m is the mass of the particle and $q(x)$ is a real valued function describing the potential energy. Here Δ represents the Laplacian in E^3 (recall that $\mathrm{D}_k = \partial/\mathrm{i}\partial x_k$).

Theorem 11.1.1. *Suppose that $q(x)$ is locally integrable, and define*

$$\rho(x)^2 = \max[q(x), 0],\ \sigma(x)^2 = \rho(x)^2 - q(x). \tag{11.1.2}$$

Assume that $\sigma \in N_1(E^3)$ and that

$$N_1'(\sigma) = 0. \tag{11.1.3}$$

Then H on $\mathrm{C}_0^\infty \cap D(q)$ has a regularly accretive self-adjoint extension in L^2. If

$$N_{1,R}(\sigma) \to 0 \text{ as } R \to \infty, \tag{11.1.4}$$

then it has such an extension B satisfying

$$\sigma_e(B) \subset [0,\infty). \tag{11.1.5}$$

If, in addition, $\rho \in N_2(E^3)$ *and*

$$N_{2,R}(\rho) \to 0 \text{ as } R \to \infty, \tag{11.1.6}$$

then

$$\sigma_e(B) = [0,\infty). \tag{11.1.7}$$

If $q \in L^2_{loc}$*, then* $D(q) \supset C_0^\infty$ *and* H *on* C_0^∞ *is essentially self-adjoint. Thus* B *is the closure in* L^2 *of* H *on* C_0^∞.

Proof. The first statement follows from Theorem 9.1.6. The second and third follow from Theorems 9.2.7 and 9.2.3, respectively. The last statement follows from Theorem 10.2.1. □

We consider several examples.

(a) *Free particle.* In this case $q(x) = 0$. H is a constant coefficient operator with real coefficients. Its closure H_0 in L^2 is self-adjoint (Corollary 5.2.3) and

$$\sigma_e(H_0) = \sigma(H_0) = [0,\infty) \tag{11.1.8}$$

(Corollary 5.3.3 and Lemma 10.4.5).

(b) *Coulomb potential.* In this case $q(x) = -e^2/|x|$, where e is a constant (charge). Here $\rho = 0$ and all of the hypotheses of Theorem 11.1.1 are satisfied. Thus H on C_0^∞ is essentially self-adjoint and its closure in L^2 satisfies

$$\sigma_e(H_0) = [0,\infty). \tag{11.1.9}$$

This of course is classical. The spectrum of H_0 represents the energy states of the hydrogen atom. We know that the spectrum of this operator consists of the non-negative real axis and the negative eigenvalues

$$\lambda_n = -2m\pi^2 e^4/h^2 n^2, \quad n = 1, 2 \ldots, \tag{11.1.10}$$

with multiplicities n^2. Note that the lowest eigenvalue is

$$\lambda_1 = -2m\pi^2 e^4/h^2$$

(ground state) and the only limit point is 0 (escape energy).

(c) Consider the potential

$$q(x) = \sum_1^N c_k \mathrm{e}^{-a_k|x|}/|x|^{b_k}, \tag{11.1.11}$$

where the a_k, b_k, c_k are constants and $a_k \geqslant 0, b_k < 3$ for each k. Thus $q(x)$ is locally integrable. If $b_k < 2$ for each k such that $c_k < 0$ and $a_k = 0$, then (11.1.3) and (11.1.4) hold. In this case H on $C_0^\infty \cap D(q)$ has a self-adjoint extension B satisfying (11.1.5). If, in addition, $b_k > 0$ whenever $a_k = 0$, then (11.1.6) is also satisfied. In this case B satisfies (11.1.7). If, in addition, $b_k < 2$ for each k, then B is the closure in L^2 of H on C_0^∞.

Examples of such potentials arising in practice are e^{-ar}/r and $\mathrm{d}(\mathrm{e}^{-ar}/r)\mathrm{d}r$ where $r = |x|$.

(d) *Hydrogen molecular ion.* Suppose we have two stationary positive charges $+e$ a distance R apart and one negative charge $-e$ in motion. We may assume that the first positive charge is at the origin and that the second is at a point r with $|r| = R$. The Hamiltonian (energy) operator for this system is

$$H = -\frac{h^2}{8\pi^2 m}\Delta + \frac{e^2}{R} - \frac{e^2}{|x|} - \frac{e^2}{|x-r|}. \tag{11.1.12}$$

We write this in the form

$$H = P(\mathrm{D}) + q(x), \tag{11.1.13}$$

where

$$P(\mathrm{D}) = -\frac{h^2}{8\pi^2}\Delta + \frac{e^2}{R} \tag{11.1.14}$$

and

$$q(x) = -\frac{e^2}{|x|} - \frac{e^2}{|x-r|}. \tag{11.1.15}$$

Since $P(\mathrm{D})$ is a constant coefficient operator, we know that

$$\sigma(P_0) = \left[\frac{e^2}{R}, \infty\right) \tag{11.1.16}$$

(Corollary 5.3.4). Moreover q is P_0-compact by Corollary 7.4.2. Hence H on C_0^∞ is essentially self-adjoint and its closure B satisfies

$$\sigma_{\mathrm{e}}(B) = \left[\frac{e^2}{R}, \infty\right). \tag{11.1.17}$$

11.2 Two particle systems

Consider a nucleus fixed at the origin in E^3 with a positive charge $2e$ and two moving particles each with negative charge $-e$ subject to Coulomb potentials. The corresponding energy operator is

$$\begin{aligned} H = {} & \frac{h^2}{8\pi^2 m_1}(\mathrm{D}_1^2 + \mathrm{D}_2^2 + \mathrm{D}_3^2) + \frac{h^2}{8\pi^2 m_2}(\mathrm{D}_4^2 + \mathrm{D}_5^2 + \mathrm{D}_6^2) \\ & - \frac{2e^2}{|x^{(1)}|} - \frac{2e^2}{|x^{(2)}|} + \frac{e^2}{|x^{(1)} - x^{(2)}|} \end{aligned} \tag{11.2.1}$$

where $x^{(1)} = (x_1, x_2, x_3)$ represents the coordinates of the first particle and $x^{(2)} = (x_4, x_5, x_6)$ those of the second. The masses of the particles are m_1 and m_2, respectively. Here H is an operator on $C_0^\infty(E^6)$ with E^6 represented by $x = (x^{(1)}, x^{(2)}) = (x_1, \dots, x_6)$.

Theorem 11.2.1. *The operator H given by (11.2.1) is essentially self-adjoint on C_0^∞ and its closure H_0 satisfies*

$$\sigma_{\mathrm{e}}(H_0) = \left[-\frac{8m\pi^2 e^4}{h^2}, \infty\right), \tag{11.2.2}$$

where $m = \max(m_1, m_2)$.

Proof. Set

$$H_1 = \frac{h^2}{8\pi^2 m_1}(\mathrm{D}_1^2 + \mathrm{D}_2^2 + \mathrm{D}_3^2) - \frac{2e^2}{|x^{(1)}|} \tag{11.2.3}$$

$$H_2 = \frac{h^2}{8\pi^2 m_2}(\mathrm{D}_4^2 + \mathrm{D}_5^2 + \mathrm{D}_6^2) - \frac{2e^2}{|x^{(2)}|}. \tag{11.2.4}$$

According to (11.1.10) the lowest eigenvalue of H_k on $C_0^\infty(E^3)$ is

$$-8m_k\pi^2 e^4/h^2, \quad k = 1, 2. \tag{11.2.5}$$

We now apply Lemma 10.8.1 and Theorem 10.8.2. We take $T_i = H_i$, $i = 1$, 2 and

$$U_i = V = \frac{e^2}{|x^{(1)} - x^{(2)}|}, \quad i = 1, 2. \tag{11.2.6}$$

Note that $V_i = Q_i = S_i = 0$ and

$$q_i(x^{(i)}) = -\frac{2e^2}{|x^{(i)}|}, \quad i = 1, 2. \tag{11.2.7}$$

In order to apply the theorem we need

$$\int_{|y_{(i)}|<1} |q_i(x^{(i)} - y^{(i)})|\ |y^{(i)}|^{-1} \mathrm{d}y^{(i)} \to 0 \text{ as } |x^{(i)}| \to \infty, \tag{11.2.8}$$

which is clearly satisfied. By (11.2.5)

$$\varkappa = \min_{i=1,2} \min_{\lambda \in \sigma(H_i)} \lambda = -8m\pi^2 e^4/h^2. \tag{11.2.9}$$

Hence we may conclude that

$$\sigma_{\mathrm{e}}(H_0) \subset [\varkappa, \infty). \tag{11.2.10}$$

In order to obtain the other conclusion, we apply Theorem 10.8.3. Let $\{\xi_{(m)}\}$ be any sequence of vectors in E^3 such that $|\xi_{(m)}| \to \infty$. Then

$$\int_{|x-y|<1} [q_i(x^{(i)} + \xi_{(m)})]^2 \mathrm{d}x \to 0 \text{ as } m \to \infty \tag{11.2.11}$$

for each $y \in E^6$. This is all that is needed to conclude that

$$\sigma_{\mathrm{e}}(H_0) \supset [\varkappa, \infty). \tag{11.2.12}$$

The proof is complete. □

Let us now consider a slightly different situation. Suppose that the second particle is free, i.e., has no interaction with either the nucleus or the first particle. The Hamiltonian in this case is

$$H = \frac{h^2}{8\pi^2 m_1}(\mathrm{D}_1^2 + \mathrm{D}_2^2 + \mathrm{D}_3^2) + \frac{h^2}{8\pi^2 m_2}(\mathrm{D}_4^2 + \mathrm{D}_5^2 + \mathrm{D}_6^2) - \frac{2e^2}{|x^{(1)}|}. \tag{11.2.13}$$

Theorem 11.2.2. *The operator H given by* (2.13) *is essentially self-adjoint on C_0^∞ and its closure H_0 satisfies*

$$\sigma(H_0) = [-8m_1\pi^2 e^4/h^2, \infty). \tag{11.2.14}$$

Proof. Set

$$P(\mathrm{D}) = \frac{h^2}{8\pi^2 m_2}(\mathrm{D}_4^2 + \mathrm{D}_5^2 + \mathrm{D}_6^2).$$

The spectrum of its closure P_0 satisfies (11.1.8). Since the lowest eigenvalue of H_1 given by (11.2.3) is $\varkappa_1 = -8m_1\pi^2 e^4/h^2$, we know that

$$\sigma(H_0) \supset [\varkappa_1, \infty)$$

by Lemma 10.7.1. On the other hand for $\varphi \in C_0^\infty(E^6)$

$$\begin{aligned}(H\varphi, \varphi) &= (H_1\varphi, \varphi) \geqslant (P(\mathrm{D})\varphi, \varphi)\\ &= (H_1\varphi, \varphi) + \frac{h^2}{8\pi^2 m_2}(||\mathrm{D}_4\varphi||^2 + ||\mathrm{D}_5\varphi||^2 + ||\mathrm{D}_6\varphi||^2)\\ &\geqslant \varkappa_1 ||\varphi||^2\end{aligned}$$

(see Lemma 2.7.17). This gives

$$\sigma(H_0) \subset [\varkappa_1, \infty),$$

and the proof is complete. □

As a contrast, consider the case of two moving particles, the first with positive charge e and the second with negative charge $-e$, attracting each other with Coulomb forces. The Hamiltonian is

$$\begin{aligned}H = \frac{h^2}{8\pi^2 m_1}(\mathrm{D}_1^2 + \mathrm{D}_2^2 + \mathrm{D}_3^2) + \frac{h^2}{8\pi^2 m_2}(\mathrm{D}_4^2 + \mathrm{D}_5^2 + \mathrm{D}_6^2)\\ - \frac{e^2}{|x^{(1)} - x^{(2)}|}.\end{aligned} \tag{11.2.15}$$

Theorem 11.2.3. *The operator H given by (11.2.15) is essentially self-adjoint on C_0^∞ and its closure H_0 satisfies*

$$\sigma(H_0) = \left[-\frac{2m_1 m_2 \pi^2 e^4}{(m_1 + m_2)h^2}, \infty\right). \tag{11.2.16}$$

Proof. Let z be a non-vanishing vector in E^3 and set $\xi = (z, z)$. Since $x^{(1)} + tz - (x^{(2)} + tz) = x^{(1)} - x^{(2)}$, we see that

$$q(x + t\xi) = q(x), \quad x \in E^6, t \text{ real},$$

where

$$q(x) = \frac{-e^2}{|x^{(1)} - x^{(2)}|}.$$

Thus we may apply Theorem 10.7.3 to conclude that $\sigma(H_0)$ consists of an interval of the form $[\varkappa, \infty)$. It remains to compute $\varkappa$. To this end make the following coordinate transformation:

$$\tilde{x}^{(1)} = x^{(1)} - x^{(2)}, \quad \tilde{x}^{(2)} = m_1 x^{(1)} + m_2 x^{(2)}. \tag{11.2.17}$$

A simple calculation shows that H transforms into

$$\begin{aligned} H = {} & \frac{h^2}{8\pi^2 m}(\tilde{\mathrm{D}}_1^2 + \tilde{\mathrm{D}}_2^2 + \tilde{\mathrm{D}}_3^2) \\ & + (m_1 + m_2)\frac{h^2}{8\pi^2}(\mathrm{D}_4^2 + \mathrm{D}_5^2 + \mathrm{D}_6^2) \\ & - \frac{e^2}{|\tilde{x}^{(1)}|} \end{aligned} \tag{11.2.18}$$

in the new coordinates, where $m = m_1 m_2/(m_1 + m_2)$. This is of the form (11.2.13) with m_1 replaced by m, m_2 replaced by $1/(m_1 + m_2)$ and e replaced by $e/\sqrt{2}$. Thus our result follows from Theorem 11.2.2. □

11.3 The existence of bound states

If H is the Hamiltonian corresponding to a system of particles and its closure H_0 satisfies

$$\sigma_{\mathrm{e}}(H_0) = [\nu, \infty), \tag{11.3.1}$$

this interval is said to represent *free states* because they correspond to the situation when at least one of the particles is free, i.e., it has the energy to move far from the others. Any $\lambda \in \sigma(H_0)$ below ν must be an eigenvalue of finite multiplicity. The corresponding eigenfunctions are called *bound* or *stationary states.* The eigenfunction corresponding to the lowest eigenvalue is called the *ground state.* Equation (11.3.1) does not tell us whether or not there exist bound states. Since the existence of bound states (and consequently of a ground state) is important in quantum theory (see section 6), we shall give a simple method of demonstrating their existence and estimating the corresponding eigenvalues.

The theory is very simple. Suppose we can find a function u in $D(H_0)$ such that

$$(H_0 u, u) = \mu ||u||^2, \tag{11.3.2}$$

where $\mu < \nu$. Then we know that the spectrum of H_0 extends below ν at least down to μ (Theorem 2.7.17). Since every point of $\sigma(H_0)$ below ν must be an eigenvalue of finite multiplicity, the existence of bound states with energies $\leqslant \mu$ is established. In particular, there is a ground state with energy $\leqslant \mu$.

Let us illustrate the method for the two particle system described at the beginning of section 2. We write

$$H = H_1 + H_2 + W, \tag{11.3.3}$$

where

$$H_1 = \frac{h^2}{8\pi^2 m_1}(\mathrm{D}_1^2 + \mathrm{D}_2^2 + \mathrm{D}_3^2) - \frac{2e^2}{|x^{(1)}|} \tag{11.3.4}$$

$$H_2 = \frac{h^2}{8\pi^2 m_2}(\mathrm{D}_4^2 + \mathrm{D}_5^2 + \mathrm{D}_6^2) - \frac{e^2}{|x^{(2)}|} \tag{11.3.5}$$

and

$$W = \frac{e^2}{|x^{(1)} - x^{(2)}|} - \frac{e^2}{|x^{(2)}|}. \tag{11.3.6}$$

From the discussion of section 1 we know that the operators H_1 and H_2 in E^3 have ground states with energies

$$\lambda_1 = -8m_1\pi^2 e^4/h^2, \quad \lambda_2 = -2m_2\pi e^4/h^2, \tag{11.3.7}$$

respectively. We note also that the corresponding normalized eigenfunctions (states) are

$$\varphi_1(x^{(1)}) = \frac{2^{\frac{9}{2}}\pi^{\frac{5}{2}} m_1^{\frac{3}{2}} e^3}{h^3} \exp\left\{-\frac{8\pi^2 m_1 e^2}{h^2}|x^{(1)}|\right\} \tag{11.3.8}$$

and

$$\varphi_2(x^{(2)}) = \frac{8\pi^{\frac{5}{2}} m_2^{\frac{3}{2}} e^3}{h^3} \exp\left\{-\frac{4\pi^2 m_2 e^2}{h^2}|x^{(2)}|\right\}, \tag{11.3.9}$$

respectively. Actually, all we need to know is that $\varphi_1(x^{(1)})$ depends only on $|x^{(1)}|$. Set

$$\varphi(x) = \varphi_1(x^{(1)})\varphi_2(x^{(2)}). \tag{11.3.10}$$

Then $\varphi(x)$ is in S and consequently in the domain of H_0. Furthermore

$$(H_1\varphi, \varphi) = (H_1\varphi_1, \varphi_1)_{E^3}||\varphi_2||^2_{E^3} = \lambda_1 \tag{11.3.11}$$

and

$$(H_2\varphi, \varphi) = (H_2\varphi_2, \varphi_2)_{E^3} = \lambda_2. \tag{11.3.12}$$

We also have

$$\begin{aligned}(W\varphi, \varphi) = e^2 \int |\varphi_2(x^{(2)})|^2 \mathrm{d}x^{(2)} \int |\varphi_1(x^{(1)})|^2 \\ \times \left\{\frac{1}{|x^{(1)} - x^{(2)}|} - \frac{1}{|x^{(1)}|}\right\} \mathrm{d}x^{(1)}.\end{aligned} \tag{11.3.13}$$

Let R be any positive number, and set

$$f(x^{(2)}) = \int_{|x^{(1)}|=R} \frac{\mathrm{d}S}{|x^{(1)} - x^{(2)}|}. \tag{11.3.14}$$

If $|x^{(2)}| > R$, then $1/|x^{(1)} - x^{(2)}|$ is harmonic in $|x^{(1)}| \leqslant R$ and consequently

$$f(x^{(2)}) = \frac{1}{|x^{(2)}|} \int_{|x^{(1)}|=R} \mathrm{d}S = \frac{4\pi R^2}{|x^{(2)}|}, \quad |x^{(2)}| > R, \tag{11.3.15}$$

by the mean value theorem. If $|x^{(2)}| < R$, set

$$h(x^{(1)}) = \frac{1}{R} - \frac{1}{|x^{(1)}|}.$$

Since h is harmonic in $\varepsilon < |x^{(1)}| < R$ for $0 < \varepsilon < |x^{(2)}|$, we have

$$\begin{aligned} 4\pi h(x^{(2)}) = &\int_{|x^{(1)}|=\varepsilon} \frac{\partial h}{\partial |x^{(1)}|} \frac{\mathrm{d}S}{|x^{(1)} - x^{(2)}|} \\ &- \int_{|x^{(1)}|=\varepsilon} \frac{\partial h}{\partial |x^{(1)}|} \frac{\mathrm{d}S}{|x^{(1)} - x^{(2)}|} \\ &+ \left(\frac{1}{R} - \frac{1}{\varepsilon}\right) \int_{|x^{(1)}|=\varepsilon} \frac{\partial}{\partial |x^{(1)}|} \left(\frac{1}{|x^{(1)} - x^{(2)}|}\right) \mathrm{d}S, \end{aligned} \tag{11.3.16}$$

where we used the fact that $h(x^{(1)}) = 0$ for $|x^{(1)}| = R$. Since $1/|x^{(1)} - x^{(2)}|$ is harmonic in $|x^{(1)}| < \varepsilon$, we have

$$\int_{|x^{(1)}|=\varepsilon} \frac{\partial}{\partial |x^{(1)}|} \left(\frac{1}{|x^{(1)} - x^{(2)}|}\right) \mathrm{d}S = 0.$$

Thus (11.3.16) yields

$$4\pi \left(\frac{1}{R} - \frac{1}{|x^{(2)}|}\right) = \frac{1}{R^2} \int_{|x^{(1)}|=R} \frac{\mathrm{d}S}{|x^{(1)} - x^{(2)}|} - \frac{4\pi}{|x^{(2)}|}$$

by (11.3.15). This gives

$$f(x^{(2)}) = 4\pi R, \quad |x^{(2)}| < R. \tag{11.3.17}$$

On the other hand, we always have

$$\int_{|x^{(1)}|=R} \left\{ \frac{1}{|x^{(1)} - x^{(2)}|} - \frac{1}{|x^{(1)}|} \right\} \mathrm{d}S \leqslant 0. \tag{11.3.18}$$

We now use the fact that $\varphi_1(x^{(1)})$ depends only on $|x^{(1)}|$. Thus if we evaluate the integral in (11.3.13) by first integrating over $|x^{(1)}| = R$ and apply (11.3.18), we see that $(W\varphi, \varphi) \leqslant 0$. This in conjunction with (11.3.11) and (11.3.12) shows that

$$(H\varphi, \varphi) \leqslant \lambda_1 + \lambda_2.$$

By Theorem 11.2.1, the smallest value of $\sigma_{\mathrm{e}}(H_0)$ is λ_1 (note that $m_1 = \max(m_1, m_2)$). Since $\lambda_2 < 0$, we see that there must be bound states. In particular we have shown the existence of bound states for the helium atom.

20210924 (Conti.)//

20210924/ A History of Greek Mathematics.../

- Heath... A History of Greek mathematics vol. I..pdf
- Heath... A History of Greek mathematics vol. II..pdf

20210924/ A History of Greek.../

- the Shaping of Deduction in Greek Mathematics...pdf

Pág. 24(43) Figure 1.5

Appolonius's Conics I. II.

Utilizá-los conjuntamente!!!

Cod. aax-vsak-bsx (meet-google)
(17h22min) Care Munch

20210925//

Kurt Gödel - Obras Completas
R$150,00

Book:

- 20210925/ Shape the Hidden Geometry.../

-256- - -* ! - " :! J 9 ! ? 2 3
7 . ! ! & 9 ! = " / 1"

-256- - -* ! - " :! J 9 ! ? 2 3
7 . ! ! && 2 ! 4 1 ! / 1"

-256- - -* ! - " 9:= :2 &87 *< 4=4H 9&*8
&8 75==+ 29:= 29& 2 "3 A : 3 % 1"
(' '% < $ 211 0 & &&

which are left ambiguous by the text – which the text does not specify fully, *given the mathematical purposes*.

Now to the results.[33] In Euclid's *Elements* XIII, about 47% of the letters are completely specified, about 8% are underspecified, about 19% are completely unspecified, and about 25% begin as completely unspecified or underspecified, and get increased specification later. In Apollonius' *Conics* I, about 42% are completely specified, about 37% are underspecified, about 4% are completely unspecified, and about 16% begin as completely unspecified or underspecified, and get increased specification later. The total number of letters in both surveys is 838.

Very often – most often – letters are not completely specified. So how do we know what they stand for? Very simple: we see this in the diagram.

In fact the difficult thing is to 'unsee' the diagram, to teach oneself to disregard it and to imagine that the only information there is is that supplied by the text. Visual information is compelling itself in an unobtrusive way. Here the confessional mode may help to convert my readers. It took me a long time to realise how ubiquitous lack of specification is. The following example came to me as a shock. It is, in fact, a very typical case.

Look at Apollonius' *Conics* I.11 (fig. 1.5). The letter Λ is specified at 38.26, where it is asserted to be on a parallel to ΔΕ, which passes through Κ. Λ is thus on a definite line. But as far as the text is concerned, there is no way of knowing that Λ is a very specific point on that line, the one intersecting with the line ΖΗ. But I had never even thought about this insufficiency of the text: I always read the diagram into the text. This moment of shock started me on this survey. Having completed the survey, its implications should be considered.

First, why are there so many cases falling short of full specification? To begin to answer this question, it must be made clear that my results have little quantitative significance. It is clear that the way in which letters in Apollonius fail to get full specification is different from that in Euclid. I expect that there is a strong variability between works by the same author. The *way* in which letters are not fully specified depends upon mathematical situations. Euclid, for instance, in book XIII, may construct a circle, e.g. ΑΒΓΔΕ and then construct a pentagon within

---

[33] The complete tables, with a more technical analysis of the semantics of specification, are to appear in Netz (forthcoming).

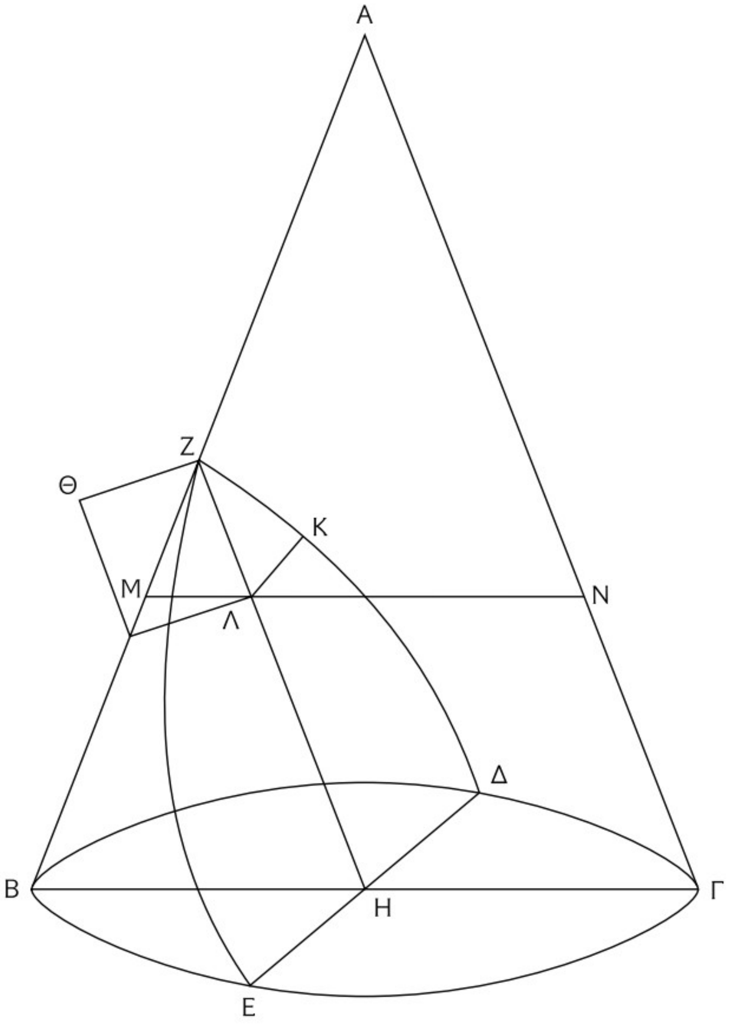

Figure 1.5. Apollonius' *Conics* 1.11.

the same circle, such that its vertices are the very same ΑΒ ΓΔΕThis is moving from underspecification to complete specification, and is demanded by the subject matter dealt with in his book. In the *Conics*, parallel lines and ordinates are the common constructions, and letters on them are often underspecified (basically, they are similar to 'BC' in figs. 1.2a, 1.2b above).

What seems to be more stable is the percentage of fully specified letters. Less than half the letters are fully specified – but not much less than half. It is as if the authors were indifferent to the question of whether a letter were specified or not, full specification being left as a random result.

This, I claim, is the case. Nowhere in Greek mathematics do we find a moment of specification *per se*, a moment whose purpose is to make sure that the attribution of letters in the text is fixed. Such moments are very common in modern mathematics, at least since Descartes.[34] But specifications in Greek mathematics are done, literally, *ambulando*. The essence of the 'imperative' element in Greek mathematics – 'let a line be drawn . . .', etc. – is to do some job upon the geometric space, to get things moving there. When a line is drawn from one point to another, the letters corresponding to the start and end positions of movement ought to be mentioned. But they need not be carefully differentiated; one need not know precisely which is the start and which is the end – both would do the same job, produce the same line (hence underspecification); and points traversed through this movement may be left unmentioned (hence complete unspecification).

What we see, in short, is that while the text is being worked through, the diagram is assumed to exist. The text takes the diagram for granted. This reflects the material implementation discussed above. This, in fact, is the simple explanation for the use of *perfect* imperatives in the references to the setting out – 'let the point *A* have been taken'. It reflects nothing more than the fact that, by the time one comes to discuss the diagram, it has already been drawn.[35]

The next point is that, conversely, the text is not recoverable from the diagram. Of course, the diagram does not tell us what the proposition asserts. It could do so, theoretically, by the aid of some symbolic apparatus; it does not. Further, the diagram does not specify all the objects on its own. For one thing, at least in the case of sphaerics, it does not even look like its object. When the diagram is 'dense', saturated in detail, even the attribution of letters to points may not be obvious from the diagram, and modern readers, at least, reading modern diagrams, use the text, to some extent, in order to elucidate the diagram. The stress of this section is on *inter*-dependence. I have not merely tried to upset the traditional balance between text and diagram; I have tried to show that they cannot be taken apart, that neither makes sense in the absence of the other.

---

[34] In Descartes, the same thing is both geometric and algebraic: it is a line (called *AB*), and it is an algebraic variable (called *a*). When the geometrical configuration is being discussed, '*AB*' will be used; when the algebraic relation is being supplied, '*a*' is used. The square on the line is 'the square on *AB*' (if we look at it geometrically) or a^2 (if we look at it algebraically). To make this double-accounting system workable, Descartes must introduce explicit, *per se* specifications, identifying *symbols*. This happens first in Descartes (1637) 300. This may well be the first *per se* moment of specification in the history of mathematics.

[35] The suggestion of Lachterman (1989) 65–7, that past imperatives reflect a certain *horror operandi*, is therefore unmotivated, besides resting on the very unsound methodology of deducing a detailed philosophy, presumably shared by each and every ancient mathematician, from linguistic practices. The methodology adopted in my work is to explain shared linguistic practices by shared situations of communication.

2.1.2 The role of text and diagram for derivations

In general, assertions may be derived from the text alone, from the diagram alone, or from a combination of the two. In chapter 5, I shall discuss grounds for assertions in more detail. What is offered here is an introduction.

First, some assertions do derive from the text alone. For instance, take the following:[36]

> As BE is to EΔ, so are four times the rectangle contained by BE, EA to four times the rectangle contained by AE, EΔ.

One brings to bear here all sorts of facts, for instance the relations between rectangles and sides, and indeed some basic arithmetic. One hardly brings to bear the diagram, for, in fact, 'rectangles' of this type often involve lines which do not stand at right angles to each other; the lines often do not actually have any point in common.[37]

So this is one type of assertion: assertions which may be viewed as verbal and *not* visual.[38] Another class is that of assertions which are based on the visual alone. To say that such assertions exist means that the text hides implicit assumptions that are contained in the diagram.

That such cases occur in Greek mathematics is of course at the heart of the Hilbertian geometric programme. Hilbert, one of the greatest mathematicians of the twentieth century, who repeatedly returned to foundational issues, attempted, in Hilbert (1899), to rewrite geometry without any unarticulated assumptions. Whatever the text assumes in Hilbert (1899), it either proves or explicitly sets as an axiom. This was never done before Hilbert, mainly because much information was taken from the diagram. As is well known, the very first proposition

---

[36] Apollonius' *Conics* I.33, 100.7–8. The Greek text is more elliptic than my translation.

[37] Here the lines mentioned do share a point, but they are not at right angles to each other. See, for instance, *Conics* I.34, 104.3, the rectangle contained by KB, AN – lines which do not share a point.

[38] This class is not exhausted by examples such as the above (so-called 'geometrical algebra'). For instance, any calculation, as e.g. in Aristarchus' *On Sizes and Distances*, owes nothing to the diagram. It should be noted that even 'geometrical algebra' is still 'geometrical': the text does not speak about multiplications, but about rectangles. This of course testifies to the primacy of the visual over the verbal. In general, see Unguru (1975, 1979), Unguru and Rowe (1981–2), Unguru and Fried (forthcoming), Hoyrup (1990a), for a detailed criticism of any interpretation of 'geometrical algebra' which misses its visual motivation. The term itself is misleading, but helps to identify a well-recognised group of propositions, and I therefore use it, quotation marks and all.

20210924 (Conti.)//

20210924/ A History of Greek Mathematics.../

_Heath... A History of Greek mathematics vol. I..pdf

_Heath... A History of Greek mathematics vol. II..pdf

20210924/ A History of Greek.../

_the Shaping of Deduction in Greek Mathematics...pdf

Pág. 24(43) Figure 1.5

Appolonius's Conics I. II.

Utilizá-los conjuntamente!!!

Cod. aax-vsak-bsx (meet-google)
(17h22min) Care Munch

20210925//

Kurt Gödel-Obras Completas

R$150,00

Book:

_20210925/ Shape the Hidden Geometry.../

5hope.... ~~pdf~~ .epub

Pag. 173 – Cremona transformation is "a birational automorphism of projective space".

What is the ratio of the area of a size-zero square to its perimeter?

If the square has a side of length 1, that radio is 1/4 or 0.25.

When the side length shrinks to 1/2, the area is 1/4 and the perimeter is 2, so the ratio has gone down to 1/8.

side length 0.1, ratio is 0.01/0.4 or 0.025

... 0/0 = 0

– 20210925/the Theory of Complex Angular Momenta/

0521818346.pdf

Pag. 11(25) 1.2.1

the Mandelstam plane

We use here the well know geometrical fact that the sum of the distances from

If I were to tell you that a Cremona transformation is "a birational automorphism of projective space," I would just be flinging phonemes at you, so let me put it a different way: What is 0/0? You have probably learned at some point that you're supposed to say "undefined," and that is, in a way, correct, but it's also the coward's way out. It really depends which zeroes you're dividing! What is the ratio of the area of a size-zero square to its perimeter? Sure, you could say it's undefined, but why not be bold and define it? If the square has a side of length 1, that ratio is 1/4 or 0.25. When the side length shrinks to 1/2, the area is 1/4 and the perimeter is 2, so the ratio has gone down to 1/8. If the side length is 0.1, the ratio is 0.01/0.4 or 0.025. The ratio gets smaller and smaller as the square does, which means there's only one good answer for what happens when the square shrinks to a point: in *this* case, 0/0 = 0. On the other hand, what if we ask about the ratio of a line segment measured in centimeters to its length measured in inches? That ratio is 2.54 for a long segment and it's 2.54 for a short segment, and if the segment shrinks down to a point, *that* 0/0 should be 2.54.

Geometrically, you can follow Descartes's example and think of a pair of numbers as a point in the plane. The point (1,2) is the point 1 unit to the right and 2 units up from the center. And the ratio 2/1 is the slope of the line joining the center to (1,2). When the point is (0,0), the center itself, there *is* no line segment, so there's no slope. The very simplest type of Cremona transformation swaps out the plane for a very similar geometry, in which (0,0) is replaced with multiple points—infinitely many, in fact! Each one remembers not just its location at (0,0) but a slope as well, as if you were keeping track not just of where you were but the direction of the path you took to get there.fn1 This kind of transformation, in which one point explodes into infinitely many, is called a *blow-up*. The higher-dimensional Cremona transformations Hudson studied are decidedly more involved; you might call them a general geometric theory of assigning values to those "undefined" ratios a more timid calculator would back away from.

Although the amplitude of the $2 \to 2$ process depends evidently on two independent variables, that is the energy of the incoming particles and the scattering angle, it is more convenient to consider A as a function of three Mandelstam variables

$$s = (p_1 + p_2)^2, \quad t = (p_1 - p_3)^2, \quad u = (p_1 - p_4)^2 .$$

They are related to each other by

$$s + t + u = \sum_{i=1}^{4} m_i^2$$

where the sum runs over the masses of all particles participating in the collision.

For the sake of simplicity, in what follows we restrict ourselves to the case of equal particle masses, $m_i = \mu$.

The Mandelstam variables have a simple physical meaning. For instance, in the centre-of-mass system (cms) of the reaction $a + b \to c + d$ (the so-called s-channel), s is the square of the total energy of the colliding particles and $t = -(\boldsymbol{p}_1 - \boldsymbol{p}_3)^2$ is the square of the momentum transfer from a to c. In the cms of the reaction $a + \bar{c} \to \bar{b} + d$ (t-channel), t plays the role of the total energy squared, and s is the square of momentum transfer. The variables u and t, respectively, play similar rôles in the u-channel reaction $a + \bar{d} \to \bar{b} + c$.

1.2.1 The Mandelstam plane

It is convenient, following Landau, to represent the kinematics of the three reactions graphically on the Mandelstam plane. We use here the well known geometrical fact that the sum of the distances from a point on the plane to the sides of an equilateral triangle does not depend on the position of the point. Therefore, taking into account the condition $s + t + u = 4\mu^2$, let us measure s, t and u as the distances to the sides of the triangle.

It is easy then to represent the physical region of any reaction on such a plane. For instance, the physical region of the reaction $a + \bar{c} \to \bar{b} + d$ corresponds to $t \geq 4\mu^2$, $s \leq 0$, $u \leq 0$ and it is shown on Fig. 1.2 as the upper shaded area. The physical regions of the other reactions can be identified in a similar manner.

In the case of the scattering of *identical* neutral particles the amplitude in each physical region is the same and it satisfies the unitarity condition separately in each region.

Examining the Mandelstam plane Fig. 1.2 we notice an interesting feature: as we move from positive to negative values of s (from the physical

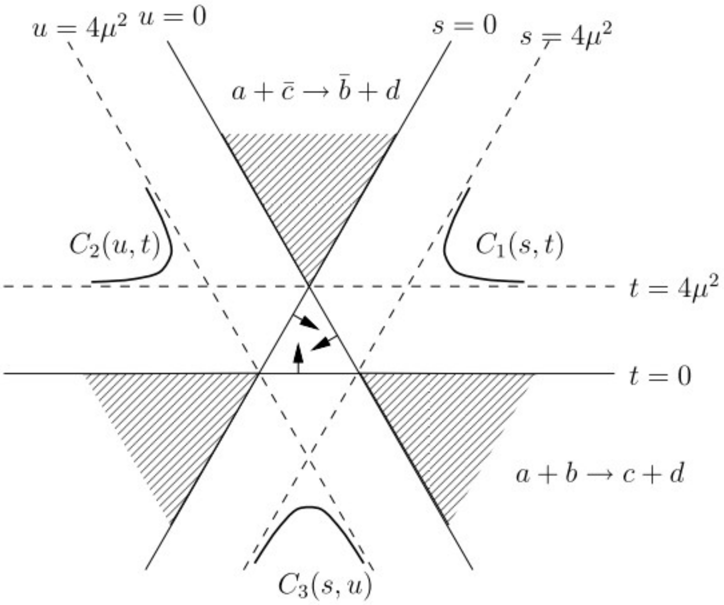

Fig. 1.2. Crossing reactions on the Mandelstam plane

region of the s-channel to the u-channel), the energy dependence of the scattering amplitude turns into the angular dependence.

1.2.2 Threshold singularities on the Mandelstam plane

Let us discuss now singularities of the amplitude. As an illustration, we consider elastic scattering of neutral pions: $\pi^0 + \pi^0 \to \pi^0 + \pi^0$. We will assume that (in accordance with experiment) pions are the lightest stable hadrons and that there is no bound state of two neutral pions. Then, the amplitude has no singularities at $s < 4\mu^2$. The first threshold lies at $s = (2\mu)^2$. It corresponds to the two-particle intermediate state. The next, three-particle threshold could have appeared at $s = (3\mu)^2$. In reality, the second threshold in the pion scattering amplitude is situated at $s = (4\mu)^2$ – the four-particle state, since the transition of two pions into three is forbidden by G-parity conservation.

Similar singularities in *energy* are known to appear in quantum mechanics, for instance the threshold singularity at $s \to 4\mu^2$.

There is however a principal difference between relativistic and nonrelativistic theories in the interpretation of the singularities in *momentum transfer*.

In quantum mechanics such singularities are determined by the potential. For instance, the Yukawa potential

$$V(r) \propto \frac{\exp(-\alpha r)}{r}$$

corresponds to a pole of the scattering amplitude in the plane of the squared momentum transfer k:

$$A(k^2) \propto \frac{1}{k^2+\alpha^2}.$$

In the relativistic theory the rôle of the potential is played by energy singularities in the t-channel, thresholds at $t = 4\mu^2$, $16\mu^2$ and so on.

Let us illustrate this statement by considering the box diagram

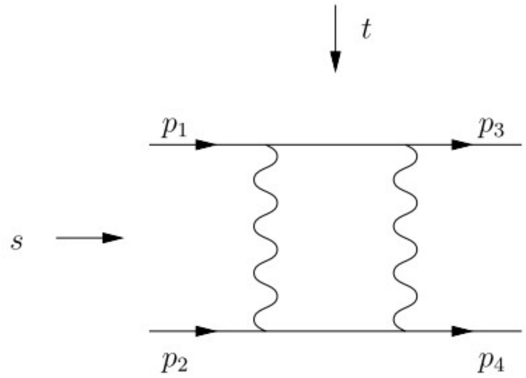

whose contribution we may interpret as defining the potential in the *next-to-Born* approximation. It is easy to see that the radius of this potential is $r = 1/2\mu$.

Thus, the assumption that all the singularities of the scattering amplitude are determined by the masses of real particles implies that there are no potentials with an infinite radius (since all hadrons have non-zero masses).

1.3 Partial wave expansion and unitarity

In order to obtain more concrete results, we must exploit analyticity and unitarity of the S-matrix.

Due to conservation of angular momentum, the unitarity condition for scattering amplitudes with given angular momentum ℓ becomes diagonal. It is convenient, therefore, to expand the s-channel amplitude into partial waves:

$$A(s,t) = \sum_{\ell=0}^{\infty} f_\ell(s)(2\ell+1)P_\ell(z), \tag{1.3a}$$

Shape.... ~~pdf~~ .epub

Pag. 173 - Cremona transformation is "a birational automorphism of projective space".

What is the ratio of the area of a size-zero square to its perimeter?

If the square has a side of length 1, that radio is 1/4 or 0.25.

When the side length shrinks to 1/2, the area is 1/4 and the perimeter is 2, so the ratio has gone down to 1/8.

side length 0.1, ratio is 0.01/0.4 or
0.025

... 0/0=0

—

- 20210925/The Theory of Complex Angular Momenta/

0521818346.pdf

Pag. 11(25) 1.2.1

the Mandelstam plane

We use here the well know geometrical fact that the sum of the distances from

('%

a point on the plane to the sides of any equilateral plane does not depend on the position of the point.

$1 + t + u = 4\mu^2$

↳ distances to the sides of the triangle.

_20210925/ Real Analysis...|

B095...Real-Analysis.pdf

Pág. R-341(346) - Chapter 7

The Riemann Integral

Pág. R-389(394) - Chapter 8

The Riemann-Stieltjes Integral

_20210925/ Complex Analysis...|

B095-...Complex Analysis Calculus of Variations.pdf

Pág. C-3(8) - Chapter 1

Analytic Functions

⇒ 1. Curves in the Argand Plane

$z = z(t) = x(t) + iy(t)$ where

$z = x + iy$.

. -256- - -* ! - ") /$ 4)8 : 5 ! 2 ! 3 1"
(5 %' %'B !1 9 5 ! & !

Chapter 7

The Riemann Integral

1 Introduction

In elementary treatments, the process of integration is generally introduced as the inverse of differentiation. If $F'(x) = f(x)$ for all x belonging to the domain of the function f, F is called an integral of the given function f.

Historically, however, the subject of integral arose in connection with the problem of finding areas of plane regions in which the area of a plane region is calculated as the limit of a sum. This notion of integral as summation is based on geometrical concepts.

A German mathematician G.F.B. Riemann gave the first rigorous arithmetic treatment of definite integral free from geometrical concepts. Riemann's definition covered only bounded functions. It was Cauchy who extended this definition to unbounded functions. Later on early in the twentieth century Lebesgue introduced the integral on a firm foundation with many refinements and generalisations.

In the present chapter we shall study the Riemann integral of real valued, bounded functions defined on some closed interval.

2 Partitions and Riemann Sums

Definition 1: *Let* $I = [a, b]$ *be a closed and bounded interval. Then by a* **partition** *(or a* **dissection** *or a* **net***) of* I *we mean a finite set of real numbers* $P = \{x_0, x_1, \dots, x_{n-1}, x_n\}$ *having the property that* $a = x_0 < x_1 < \dots < x_{n-1} < x_n = b$.

The closed sub-intervals

$$I_1 = [x_0, x_1], I_2 = [x_1, x_2], \dots, I_n = [x_{n-1}, x_n]$$

determined by P *constitute the* **segments** *of the partition.*

By changing the set of points, the partition can be changed. Thus there can be an infinite number of partitions of the interval $[a, b]$. The family of all partitions of $[a, b]$ shall be denoted by $\mathbf{P}[a, b]$.

We write $\Delta x_r = x_r - x_{r-1}$ for $r = 1, 2, \dots, n$ so that Δx_r is the length of the segment $[x_{r-1}, x_r]$. The **norm** (or **mesh**) of a partition P is the greatest of the lengths of the segments of a partition P and it is denoted by $||P||$. Thus

$$||P|| = \text{max.} (\Delta x_r : r = 1, 2, \dots, n).$$

Sometimes the norm of a partition P is also denoted by $\mu(P)$.

Definition 2: *A partition* P^* *is called a* **refinement** *of another partition* P *or we say that* P^* *is* **finer than** P *iff* $P^* \supset P$, *i.e., if every point of* P *is used in the construction of* P^*.

If $P^* = P_1 \cup P_2$, *then* P^* *is called the* **common refinement** *of the given two partitions* P_1 *and* P_2.

Definition 3: *Let* f *be a bounded function defined on a bounded interval* $[a, b]$. *Also let* $P = \{a = x_0, x_1, \dots, x_n = b\}$ *be any partition of* $[a, b]$. *Also let* m_r *and* M_r *be the infimum and supremum respectively of the function* f *on* I_r, *for* $r = 1, 2, \dots, n$ *i.e.,*

$$m_r = inf \{ f(x) : x_{r-1} \le x \le x_r \}$$

and $$M_r = sup \{ f(x) : x_{r-1} \le x \le x_r \}.$$

Let us now form two sums

$$L(P, f) = \sum_{r=1}^{n} m_r \Delta x_r \text{ and } U(P, f) = \sum_{r=1}^{n} M_r \Delta x_r .$$

Then $L(P, f)$ *is called the* **lower Riemann sum** (*or* **lower Darboux sum**) *of* f *on* $[a, b]$ *with respect to the partition* P *and* $U(P, f)$ *is called the* **upper Riemann sum** (*or* **upper Darboux sum**) *of* f *on* $[a, b]$ *with respect to the partition* P.

In brief, we shall refer to these sums as the *lower and upper R-sums* of f with respect to P. Obviously $L(P, f) \le U(P, f)$.

Theorem 1: *Let* f *be a bounded function defined on* $[a, b]$ *and let* m *and* M *be the infimum and supremum of* $f(x)$ *in* $[a, b]$. *Then for any partition* $\mathbf{P}$ *of* $[a, b]$, *we have*

$$m(b - a) \le L(P, f) \le U(P, f) \le M(b - a).$$ **(Garhwal 2008)**

Proof: Let $P = \{a = x_0, x_1, \dots, x_n = b\}$ be any partition of $[a, b]$. Then $I_r = [x_{r-1}, x_r], r = 1, 2, \dots, n$ are the subintervals of $[a, b]$. Let m_r and M_r be the infimum and supremum of $f(x)$ in $[x_{r-1}, x_r]$. Then for every value of r, we have

$$m \le m_r \le M_r \le M$$

$\Rightarrow$ $$m\,\Delta x_r \le m_r\,\Delta x_r \le M_r\,\Delta x_r \le M\,\Delta x_r \qquad [\because \Delta x_r > 0]$$

$\Rightarrow$ $$\sum_{r=1}^{n} m\,\Delta x_r \le \sum_{r=1}^{n} m_r\,\Delta x_r \le \sum_{r=1}^{n} M_r\,\Delta x_r \le \sum_{r=1}^{n} M\,\Delta x_r\,. \qquad \text{...(1)}$$

Now $$\sum_{r=1}^{n} m\,\Delta x_r = m \sum_{r=1}^{n} \Delta x_r = m \sum_{r=1}^{n} (x_r - x_{r-1})$$
$$= m\,(x_1 - x_0 + x_2 - x_1 + \ldots + x_n - x_{n-1})$$
$$= m\,(x_n - x_0) = m\,(b-a).$$

Similarly $$\sum_{r=1}^{n} M\Delta\,x_r = M\,(b-a).$$

Also $$\sum_{r=1}^{n} m_r\,\Delta x_r = L\,(P, f\,) \text{ and } \sum_{r=1}^{n} M_r\,\Delta x_r = U\,(P, f\,).$$

Hence, from (1), we conclude that

$$m\,(b-a) \le L\,(P, f\,) \le U\,(P, f\,) \le M\,(b-a)\ \forall\, P \in \boldsymbol{P}\,[a, b].$$

It follows that the sets of upper sums and lower sums are bounded.

Theorem 2: *If* $f : [a, b] \to \mathrm{R}$ *is a bounded function, then*

$$U\,(P, -f\,) = -L\,(P, f\,) \text{ and } L\,(P, -f\,) = -U\,(P, f\,). \qquad \textbf{(Meerut 2012)}$$

Proof: Let $P = \{a = x_0, x_1, \ldots, x_n = b\}$ be any partition of $[a, b]$. Let M_r and m_r be the supremum and infimum of f in I_r.

Now f is bounded on $[a, b] \Rightarrow -f$ is bounded on $[a, b]$.

Again M_r, m_r are supremum and infimum of f in I_r

$\Rightarrow$ $-m_r, -M_r$ are supremum and infimum of $-f$ in I_r.

We have $$U\,(P, -f\,) = \sum_{r=1}^{n} (-m_r)\,\Delta x_r\,, \text{by definition of upper } R\text{-sum}$$
$$= -\sum_{r=1}^{n} m_r\,\Delta x_r = -L\,(P, f\,).$$

Also $$L\,(P, -f\,) = \sum_{r=1}^{n} (-M_r)\,\Delta x_r\,, \text{by definition of lower } R\text{-sum}$$
$$= -\sum_{r=1}^{n} M_r\,\Delta x_r = -U\,(P, f\,).$$

Theorem 3: *Let f be a bounded function defined on $[a, b]$ and let P be a partition of $[a, b]$. If P^* is a refinement of P, then*

$$L\,(P^*, f\,) \ge L\,(P, f\,) \textit{ and } U\,(P^*, f\,) \le U\,(P, f\,).$$

Proof: Let $P = \{a = x_0, x_1, x_2, \ldots, x_{r-1}, x_r, \ldots, x_n = b\}$

and $P^* = \{a = x_0, x_1, x_2, \ldots, x_{r-1}, y_1, x_r, \ldots, x_n = b\}$,

so that P^* has one more partition point y_1 than P.

Let m_r and M_r be the infimum and supremum of f in $[x_{r-1}, x_r]$. Let M_r', M_r'' be the suprema of f in $[x_{r-1}, y_1], [y_1, x_r]$ and m_r', m_r'' be the infima of f in $[x_{r-1}, y_1], [y_1, x_r]$ respectively.

Then $M_r \ge M_r', M_r''$ and $m_r \le m_r', m_r''$.

Since the rth subinterval only of P is split into two more subintervals of P^* and the remaining subintervals are identical in P and P^*, therefore, we have

$$U(P, f) - U(P^*, f)$$
$$= M_r(x_r - x_{r-1}) - \{M_r'(y_1 - x_{r-1}) + M_r''(x_r - y_1)\}.$$

But $\quad M_r'(y_1 - x_{r-1}) + M_r''(x_r - y_1) \leq M_r(y_1 - x_{r-1}) + M_r(x_r - y_1)$
$$= M_r(x_r - x_{r-1}).$$

It gives $\quad U(P, f) - U(P^*, f) \geq M_r(x_r - x_{r-1}) - M_r(x_r - x_{r-1}) = 0$

i.e., $\quad U(P, f) \geq U(P^*, f)$ or $U(P^*, f) \leq U(P, f)$.

If P^* contains p more partition points than P then repeating p times the above argument we can show that

$$U(P^*, f) \leq U(P, f). \qquad \text{...(1)}$$

In a similar manner we can show that

$$L(P^*, f) \geq L(P, f). \qquad \text{...(2)}$$

Note: We know that for any partition P^*,

$$L(P^*, f) \leq U(P^*, f). \qquad \text{...(3)}$$

Thus from (1), (2) and (3), we get

$$L(P, f) \leq L(P^*, f) \leq U(P^*, f) \leq U(P, f).$$

Theorem 4: *If P_1 and P_2 be any two partitions of $[a, b]$ then*

$$U(P_1, f) \geq L(P_2, f).$$

Proof: Let $P = P_1 \cup P_2$. Then P is the common refinement of both the partitions P_1 and P_2. Therefore by the theorem 3 above, we get

$$L(P_2, f) \leq L(P, f) \text{ and } U(P, f) \leq U(P_1, f).$$

But we have $\quad L(P, f) \leq U(P, f)$.

Thus $\quad L(P_2, f) \leq L(P, f) \leq U(P, f) \leq U(P_1, f)$.

It follows that $L(P_2, f) \leq U(P_1, f)$ or $U(P_1, f) \geq L(P_2, f)$, *that is, every upper sum for f is greater than or equal to every lower sum for f.*

Theorem 5: *Let f, g be bounded functions defined on $[a, b]$ and let P be any partition of $[a, b]$. Then*

$$L(P, f + g) \geq L(P, f) + L(P, g)$$

and $$U(P, f + g) \leq U(P, f) + U(P, g).$$

Proof: Let $P = \{a = x_0, x_1, x_2, \ldots, x_n = b\}$ be any partition of $[a, b]$. Since f and g are bounded functions on $[a, b]$, $f + g$ is also bounded on $[a, b]$.

Let m_r', M_r' be the infimum and supremum of f on I_r,

m_r'', M_r'' be the infimum and supremum of g on I_r

and m_r, M_r be the infimum and supremum of $f + g$ on I_r.

By definition of infimum we find that

$$f(x) \geq m_r', \; g(x) \geq m_r'' \;\forall\, x \in I_r$$

$$\Rightarrow \quad f(x) + g(x) \geq m_r' + m_r'' \;\forall\, x \in I_r$$

$$\Rightarrow \qquad (f+g)(x) \geq m_r' + m_r'' \ \forall\, x \in I_r$$

$$\Rightarrow \qquad m_r' + m_r'' \text{ is a lower bound of } f+g \text{ on } I_r .$$

But m_r is the greatest lower bound of $f + g$ on I_r .

$$\therefore \qquad m_r \geq m_r' + m_r'' \Rightarrow m_r\, \Delta x_r \geq m_r'\, \Delta x_r + m_r''\, \Delta x_r$$

$$\Rightarrow \qquad \sum_{r=1}^{n} m_r\, \Delta x_r \geq \sum_{r=1}^{n} m_r'\, \Delta x_r + \sum_{r=1}^{n} m_r''\, \Delta x_r$$

$$\Rightarrow \qquad L(P, f+g) \geq L(P, f) + L(P, g).$$

Similarly we can prove the other result.

3 Lower and Upper Riemann Integrals

(Purvanchal 2009, 12)

Let f be a real valued bounded function defined on $[a, b]$. We know that the set of all numbers $L(P, f)$ with respect to all possible partitions P of $[a, b]$ is bounded above by $M(b-a)$ and hence there exists a supremum of $L(P, f)$. The **lower Riemann integral (lower *R*-integral)** of f over $[a, b]$ is the supremum of $L(P, f)$ over all partitions $P \in \mathbf{P}[a, b]$. It is denoted by

$$\underline{\int_a^b} f(x)\, dx.$$

Similarly the set of numbers $U(P, f)$ is bounded below by $m(b-a)$ and so it possesses an infimum. The **upper Riemann integral** (**upper *R*-integral**) of f over $[a, b]$ is the infimum of $U(P, f)$ over all partitions $P \in \mathbf{P}[a, b]$. It is denoted by

$$\overline{\int_a^b} f(x)\, dx.$$

Thus
$$\underline{\int_a^b} f(x)\, dx = \sup \{L(P, f) : P \text{ is a partition of } [a, b]\}$$

and
$$\overline{\int_a^b} f(x)\, dx = \inf \{U(P, f) : P \text{ is a partition of } [a, b]\}.$$

We denote the lower and upper integrals of f simply by

$$\underline{\int_a^b} f \qquad \text{and} \qquad \overline{\int_a^b} f .$$

Since $L(P, -f) = -U(P, f)$ and $U(P, -f) = -L(P, f)$, it gives that

$$\underline{\int_a^b} (-f) = -\overline{\int_a^b} f \quad \text{and} \quad \overline{\int_a^b} (-f) = -\underline{\int_a^b} f.$$

Theorem 1: *The lower R-integral cannot exceed the upper R-integral, i.e.,*

$$\underline{\int_a^b} f \leq \overline{\int_a^b} f.$$

(Purvanchal 2007, 09; Rohilkhand 11)

Proof: If P_1 and P_2 are any two partitions of $[a, b]$, then by theorem 4 of article 2, we have

$$L(P_1, f) \leq U(P_2, f). \qquad \text{...(1)}$$

First, keeping P_2 fixed and taking the supremum over all partitions P_1, (1) gives

Chapter 8

The Riemann-Stieltjes Integral

1 A Generalization of the Riemann Integral

In the preceding chapter we have seen that the definitions of Riemann integral are founded on finite sums of the form

$$\sum_{r=1}^{n} m_r \, \Delta \, x_r\,, \quad \sum_{r=1}^{n} M_r \, \Delta \, x_r \quad \text{or} \quad \sum_{r=1}^{n} f\,(\xi_r)\, \Delta \, x_r.$$

If in the above sums we introduce $\Delta g_r = g\,(x_r) - g\,(x_{r-1})$ in place of $\Delta \, x_r$, where g is a more or less an arbitrary function, then a generalization of the Riemann integral is achieved which reduces to the Riemann case in the particular case $g\,(x) = x$. It was Thomas Jan Stieltjes who gave a generalization of Riemann integral. In the present chapter we shall not discuss the Stieltjes integrals as such but only a sort of special case thereof known as Riemann-Stieltjes integral in some detail.

2 Some Definitions

For convenience we repeat some definitions again here although we have given them in the preceding chapter.

Partition: **Definition.** *By a partition of a closed interval* $I = [a, b]$ *we mean a finite set of real numbers*

$$P = \{x_0, x_1, \ldots, x_{n-1}, x_n\} \text{ having the property that}$$

$$a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b.$$

The closed sub-intervals

$$I_1 = [x_0, x_1], I_2 = [x_1, x_2], \ldots, I_n = [x_{n-1}, x_n]$$

determined by P constitute the **segments** of the partition.

We write $\Delta x_r = x_r - x_{r-1}$ for $r = 1, 2, \ldots, n$, so that Δx_r is the **length** of the segment $[x_{r-1}, x_r]$. The **norm** of a partition P is the greatest of the lengths of the segments of a partition P and it is denoted by $\|P\|$. Thus

$$\|P\| = \max. (\Delta x_r : r = 1, 2, \ldots, n).$$

A partition P^* *is called a* ***refinement*** *of another partition* P *or we say that* P^* *is* ***finer than*** P *iff* $P^* \supset P$, *i.e., if every point of* P *is used in the construction of* P^*.

If $P^* = P_1 \cup P_2$, *then* P^* *is called the* ***common refinement*** *of the given two partitions* P_1 *and* P_2.

By an ***intermediate partition*** *of* P, *we mean a set* $Q = \{\xi_1, \xi_2, \ldots, \xi_n\}$ *of at most* n *points having the property that* $x_{r-1} \le \xi_r \le x_r$ *for* $r = 1, 2, \ldots, n$.

According to this definition, two consecutive points of Q can be identical. Thus, it may happen that $\xi_{r-1} = \xi_r = x_r$ for some value of r. But three or more points cannot coincide.

3 Lower and Upper Riemann-Stieltjes Sums

Let f be a bounded real valued function defined on $[a, b]$ and let g be a monotonically non-decreasing (real valued) function on $[a, b]$. We shall write

$$\Delta g_r = g(x_r) - g(x_{r-1})$$

corresponding to each partition $P = \{x_0, x_1, \ldots, x_n\}$ of $[a, b]$.

Note that $\Delta g_r \ge 0$ since g is monotonically non-decreasing on $[a, b]$. Also, we define

$$m_r = \inf \{f(x) : x_{r-1} \le x \le x_r\},$$

$$M_r = \sup \{f(x) : x_{r-1} \le x \le x_r\},$$

$$m = \inf \{f(x) : a \le x \le b\}, M = \sup \{f(x) : a \le x \le b\}.$$

Let us now form the two sums

$$L(P, f, g) = \sum_{r=1}^{n} m_r \Delta g_r \text{ and } U(P, f, g) = \sum_{r=1}^{n} M_r \Delta g_r.$$

These two numbers are called the **lower** and **upper Riemann Stieltjes Sums** (or in brief lower and upper RS sums) respectively, of f with respect to g and corresponding to the partition P.

In case $g(x) = x$, these reduce to the lower and upper Riemann sums defined in the preceding chapter.

Obviously $L(P, f, g) \le U(P, f, g)$, for any partition P.

If P^* is a refinement of P, then

$$L(P, f, g) \le L(P^*, f, g) \qquad \text{...(1)}$$

and $$U(P^*, f, g) \le U(P, f, g). \qquad \text{...(2)}$$

For any partitions P_1, P_2 of $[a, b]$, we have

$$L(P_1, f, g) \le U(P_2, f, g). \qquad \text{...(3)}$$

The results (1), (2) and (3) can be proved similarly as proved in the preceding chapter with x's replaced by $g(x)$'s and Δx_r by Δg_r etc.

4 The Lower and Upper Riemann-Stieltjes Integrals

(Gorakhpur 2010)

Definition: *Let f be a bounded function and g a non-decreasing function defined on $[a, b]$. The **lower Riemann-Stieltjes integral** of f relative to g over $[a, b]$ is the supremum of $L(P, f, g)$ over all partitions P of $[a, b]$.* It is denoted by

$$\underline{\int_a^b} f\,dg \quad or \quad \underline{\int_a^b} f(x)\,dg(x).$$

*The **upper Riemann-Stieltjes integral** of f relative to g over $[a, b]$ is the infimum of $U(P, f, g)$ over all partitions P of $[a, b]$.* It is denoted by

$$\overline{\int_a^b} f\,dg \quad or \quad \overline{\int_a^b} f(x)\,dg(x).$$

Theorem: *Let f be a bounded function and g a non-decreasing function on $[a, b]$. Then the lower RS-integral of f relative to g cannot exceed the upper RS-integral*

i.e., $$\underline{\int_a^b} f\,dg \le \overline{\int_a^b} f\,dg.$$

(Gorakhpur 2010)

Proof: If P_1 and P_2 are any two partitions of $[a, b]$ then by (3) of article 3, we get

$$L(P_1, f, g) \le U(P_2, f, g). \qquad \text{...(1)}$$

First, keeping P_2 fixed and taking the supremum over all partitions P_1, (1) gives

$$\underline{\int_a^b} f\,dg \le U(P_2, f, g). \qquad \text{...(2)}$$

Now taking infimum over all partitions P_2, (2) gives

$$\underline{\int_a^b} f\,dg \le \overline{\int_a^b} f\,dg.$$

5 The Riemann-Stieltjes Integral

Definition: *Let f be a bounded function and g a non-decreasing function on $[a, b]$. Then f is called* ***Riemann-Stieltjes integrable*** *(or simply RS-integrable) on $[a, b]$ relative to g iff*

$$\underline{\int_a^b} f\,dg = \overline{\int_a^b} f\,dg$$

and their common value is called the RS-integral of f on $[a, b]$ relative to g and is denoted by

$$\int_a^b f\,dg \quad or \quad \int_a^b f(x)\,dg(x).$$

f is called the ***Integrand*** *and g the* ***Integrator.*** **(Gorakhpur 2012)**

The class of all RS-integrable functions relative to g over $[a, b]$ is denoted by $RS([a, b], g)$ or simply by $RS(g)$.

From the above definition, it is obvious that quite apart from conditions which might be necessary to the existence of an RS-integral, the integrand f must be bounded and the integrator g non-decreasing.

As in the case of Riemann integral, we define

$$\int_a^b f\,dg = -\int_b^a f\,dg \text{ for } b < a \text{ and } \int_a^a f\,dg = 0.$$

If $g(x) = x$, then the lower and upper RS-integrals reduce respectively to the lower and upper R-integrals. Hence the RS-integral of f relative to x is the same as the R-integral of f.

Note: The integral depends on f, g, a and b and not on the so-called variable of integration

i.e.,
$$\int_a^b f(x)\,dg(x) = \int_a^b f(y)\,dg(y).$$

So omitting the variable of integration we shall prefer to write

$$\int_a^b f\,dg \text{ in place of } \int_a^b f(x)\,dg(x).$$

6 Necessary and Sufficient Condition for RS-Integrability

Theorem: *Let f be a bounded and g a non-decreasing function on $[a, b]$. Then $f \in RS(g)$ if and only if for every $\varepsilon > 0$ there exists a partition P of $[a, b]$ such that*

$$U(P, f, g) - L(P, f, g) < \varepsilon.$$ **(Gorakhpur 2012)**

Proof: The condition is necessary (Only if part).

Let $f \in RS(g)$ so that

$$\underline{\int_a^b} f\,dg = \overline{\int_a^b} f\,dg. \qquad \text{...(1)}$$

Since $\underline{\int_a^b} f\,dg$ is the supremum of $L(P, f, g)$ over all partitions P, there exists a partition P_1 such that

$$\underline{\int_a^b} f\,dg < L(P_1, f, g) + \frac{\varepsilon}{2}\cdot$$

Similarly since $\overline{\int_a^b} f\,dg$ is the infimum of $U(P, f, g)$ over all partitions P, there exists a partition P_2 such that

$$U(P_2, f, g) < \overline{\int_a^b} f\,dg + \frac{\varepsilon}{2}\cdot$$

If $P = P_1 \cup P_2$, then P is the common refinement of P_1 and P_2. It follows from (1) and (2) of article 3 that

$$\underline{\int_a^b} f\,dg < L(P, f, g) + \frac{\varepsilon}{2} \qquad \text{...(2)}$$

and $$U(P, f, g) < \int_a^b f\,dg + \frac{\varepsilon}{2}\cdot \qquad \text{...(3)}$$

Adding the inequalities (2) and (3), we get

$$\underline{\int_a^b} f\,dg + U(P, f, g) < L(P, f, g) + \overline{\int_a^b} f\,dg + \varepsilon\,.$$

In view of (1), this gives

$$U(P, f, g) < L(P, f, g) + \varepsilon \text{ i.e., } U(P, f, g) - L(P, f, g) < \varepsilon.$$

Hence the condition is necessary.

The condition is sufficient (if part).

Let for every $\varepsilon > 0$, there exists a partition P of $[a, b]$ such that

$$U(P, f, g) - L(P, f, g) < \varepsilon. \qquad \text{...(4)}$$

By the definition of upper and lower RS-integrals, we have

$$\overline{\int_a^b} f\,dg \le U(P, f, g) \text{ and } \underline{\int_a^b} f\,dg \ge L(P, f, g)$$

i.e., $$L(P, f, g) \le \underline{\int_a^b} f\,dg \le \overline{\int_a^b} f\,dg \le U(P, f, g). \qquad \text{...(5)}$$

It follows from (4) and (5) that

$$\overline{\int_a^b} f\,dg - \underline{\int_a^b} f\,dg \le U(P, f, g) - L(P, f, g) < \varepsilon$$

or $$\overline{\int_a^b} f\,dg - \underline{\int_a^b} f\,dg < \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we have $\overline{\int_a^b} f\, dg = \underline{\int_a^b} f\, dg$ *i.e.*, the function f is *RS*-integrable with respect to g.

Note: It is worth while to observe that the details of the proof of this theorem are exactly similar to those of the corresponding theorem for *R*-integrals given in the preceding chapter. In fact most of the elementary properties of *R*-integrals remain valid for *RS*-integrals with their corresponding proofs exactly similar except for the notation used for the two types of integrals.

Illustrative Examples

Example 1: *Let f be a constant function on $[a, b]$ defined by $f(x) = k$ and g a monotonically non-decreasing function on $[a, b]$. Prove that $\int_a^b f\, dg$ exists and $\int_a^b f\, dg = k\,[g(b) - g(a)]$.*

(Meerut 2004)

Solution: Obviously the given function is bounded over $[a, b]$. Let $P = \{a = x_0, x_1, x_2, \ldots, x_n = b\}$ be any partition of $[a, b]$. Then for any sub-interval $[x_{r-1}, x_r]$, we have

$$m_r = k,\ M_r = k.$$

Now
$$U(P, f, g) = \sum_{r=1}^{n} k\, \Delta g_r = k \sum_{r=1}^{n} [g(x_r) - g(x_{r-1})]$$

$$= k\,[g(x_1) - g(x_0) + g(x_2) - g(x_1) + \ldots + g(x_n) - g(x_{n-1})]$$

$$= k\,[g(x_n) - g(x_0)] = k\,[g(b) - g(a)],$$

and
$$L(P, f, g) = \sum_{r=1}^{n} k\, \Delta g_r = k\,[g(b) - g(a)].$$

$\therefore$
$$\overline{\int_a^b} f\, dg = \inf U(P, f, g) = \inf k\,[g(b) - g(a)]$$

$$= k\,[g(b) - g(a)]$$

and
$$\underline{\int_a^b} f\, dg = \sup L(P, f, g) = \sup k\,[g(b) - g(a)] = k\,[g(b) - g(a)].$$

Since $\underline{\int_a^b} f\, dg = \overline{\int_a^b} f\, dg$, the integral $\int_a^b f\, dg$ exists

and
$$\int_a^b f\, dg = k\,[g(b) - g(a)].$$

Example 2: *If f is non-negative and integrable function with respect to α on $[a, b]$ and if α is an increasing function on $[a, b]$, then show that*

$$\int_a^b f\, d\alpha \geq 0.$$

-256- - -* ! - ") /$ % 9' 1 L 2 ! 3 ;
! ! ! 1"
(% !1

Chapter 1

Analytic Functions

1 Curves in the Argand Plane

We know that the equations of the type $x = x(t)$, $y = y(t)$, where t is the parameter, give the parametric representation of a curve in the plane. Using the complex variable z, these equations can be written as a single equation

$$z = z(t) = x(t) + i\,y(t) \text{ where } z = x + iy.$$

Definitions: (a) *In the Argand plane, a continuous complex valued function* $z(t) = x(t) + iy(t)$, *where* $x(t)$ *and* $y(t)$ *are real valued continuous functions of a real variable t, defined in the range* $\alpha \le t \le \beta$ *where* $\alpha < \beta$ *is called a continuous* ***arc*** *or a* ***curve***. We call $z(\alpha)$ and $z(\beta)$ the end points of the curve, $z(\alpha)$ is the initial point and $z(\beta)$ the terminal point of the curve. If $z(\alpha) = z(\beta)$ *i.e.,* if the initial and terminal points of a curve coincide, the curve is said to be a **closed curve.**

A point z_1 is a multiple point of the curve if the equation $z_1 = x(t) + iy(t)$ is satisfied by more than one value of t in the given range. In particular the multiple point is called a double point if the above equation is satisfied by two values of t in the given range.

(b) A curve Γ given by, $z(t) = x(t) + iy(t), \alpha \le t \le \beta$ is called a **Jordan arc** or a **simple**

curve if $t_1 \neq t_2$ implies $z\ (t_1) \neq z\ (t_2)$ *i.e.*, $z\ (t)$ is one-one. A Jordan arc is a curve without multiple points.

(c) A closed curve Γ given by $z\ (t) = x\ (t) + iy\ (t), \alpha \leq t \leq \beta$, is called **simple** if $t_1 < t_2$ and $z\ (t_1) = z\ (t_2)$ imply $t_1 = \alpha$ and $t_2 = \beta$.

We usually refer to such curves as **simple closed Jordan curves.**

Example: The circle $z = \cos t + i \sin t, 0 \leq t \leq 2\pi$ is a simple closed Jordan curve since the values of $z\ (t)$ coincide only at the end points $t = 0$ and $t = 2\pi$.

The Jordan curve theorem: The theorem states that *a simple closed Jordan curve divides the Argand plane into two open domains which have the curve as common boundary.*

Of these two domains one is bounded and it is called the **interior domain** ; the other is unbounded and is called the **exterior domain.** For example, the circle $|z| = r$ divides the Argand plane into two open domains given by $|z| < r$ and $|z| > r$. Out of these $|z| < r$ is bounded and is the interior of the circle ; the other $|z| > r$ is unbounded and is the exterior of the circle, $|z| = r$. The circle is the common boundary of the two domains.

2 Functions of a Complex Variable

If certain rules be given by means of which it is possible to find one or more complex numbers w for every value of z in a certain domain D, w is said to be a function of z defined on the domain D and we write

$$w = f\ (z).$$

Since $z = x + iy$, $f\ (z)$ will be of the form $u + iv$ where u and v are functions of two real variables x and y. Thus we can write $w = u\ (x, y) + i\, v\ (x, y)$, x, y are real.

Single valued and multiple valued functions: If w takes only one value for each value of z in the region D, w is said to be a *uniform* or *single valued function* of z. If there correspond two or more values of w for some or all values of z in the region D, w is called *multiple valued function* of z.

3 Neighbourhood of a Point

A neighbourhood of a point z_0 in the Argand plane is the set of all points z such that $|z - z_0| < \delta$, where δ is an arbitrary small positive number. The number δ is called the radius of this neighbourhood.

Deleted neighbourhood: If from a neighbourhood of a point z_0, the point z_0 itself is deleted or excluded, we get a *deleted neighbourhood* of z_0.

4 Limit and Continuity

Let $f\ (z)$ be any function of the complex variable z defined in a bounded and closed domain D. Then l is said to be the limit of $f\ (z)$ as z approaches a along any path in D if for any arbitrarily chosen *positive number ε, however small but not zero, there exists a corresponding number δ greater than zero such that*

$$| f(z) - l | < \varepsilon,$$

for all values of z for which $0 < | z - a | < \delta$.

In symbols, we write $\lim\limits_{z \to a} f(z) = l$.

Continuity: *A function $f(z)$ of a complex variable z defined in the closed and bounded domain D is said to be continuous at $a \in D$ if and only if for any arbitrarily chosen positive number ε, however small but not zero, there exists a corresponding number $\delta > 0$ such that*

$$| f(z) - f(a) | < \varepsilon \text{ whenever } | z - a | < \delta.$$

It follows from the definitions of limit and continuity that $f(z)$ is continuous at

$$z = a \text{ iff } \lim_{z \to a} f(z) = f(a).$$

We say that $f(z)$ is a continuous function in a domain D if it is continuous at every $z \in D$.

We can easily show that $f(z) = u(x, y) + i\,v(x, y)$ is a continuous function of z iff u and v are continuous functions of x and y.

Uniform continuity: *A function $f(z)$ defined in a domain D is said to be uniformly continuous in D if given $\varepsilon > 0$, there exists a $\delta > 0$ such that $| f(z_1) - f(z_2) | < \varepsilon$ whenever $| z_1 - z_2 | < \delta$, where z_1 and z_2 are points in D.*

It should be noted carefully that uniform continuity is a property associated with a domain and not with a single point of it.

If $f(z)$ is uniformly continuous in a domain D, it is continuous in D.

A function which is continuous in a closed and bounded domain D is uniformly continuous in D, whereas a function continuous in an open domain D' may fail to be uniformly continuous in the domain D'.

5 Differentiability

Since the mode of definitions of continuity is the same both in case of the functions of the real and complex variables therefore definition of differentiability of a complex function is identical with that of the real function.

Let $w = f(z)$ be a function of a complex variable z defined in a domain D. Then $f(z)$ is said to be differentiable at a point z_0 of D iff

$$\lim_{\Delta z \to 0} \frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z} \quad \text{or} \quad \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0} \text{ exists uniquely and finitely}$$

and this limit, if it exists finitely, *is called the* **differential coefficient** or **derivative** of f with respect to z at $z = z_0$.

It is denoted by $f'(z_0)$ or by $Df(z_0)$.

('%

a point on the plane to the sides of an equilateral plane does not depend on the position of the point.

$s + t + u = 4\mu^2$

↳ distances to the sides of the triangle.

_20210925/ Real Analysis.../

B095... Real-Analysis.pdf

Pág. R-341(346) - Chapter 7

The Riemann Integral

Pág. R-389(394) _ Chapter 8

The Riemann-Stieltjes Integral

_20210925/ Complex Analysis.../

B095-... Complex Analysis
Calculus of Variations.pdf

Pág. C-3(8) - Chapter 1

Analytic Functions

⇒ 1. Curves in the Argand Plane

$z = z(t) = x(t) + iy(t)$ where

$z = x + iy$.

("

data

_20210925/ Quantum Atom...|

11 08838596-Quantum-Atom....pdf

Pag.100(115) - Fig 6.1

d sen θ.

the diffraction of light by a grating.

—

malavasi

—

0,01/0,4=0,025

—

20210926m

squeaky = estridente

haste make waste =

pressa faz desperdício

—

Books:

20210926/Modern Physics...|

Modern_physics_6e.pdf

Pag.17(35) 1.3

-256- - -* ! - " % $/B 6 ! 2
*1 1"
($ < B

6

Diffraction of Atoms Using Standing Wave Light

6.1 Introduction

Diffraction of waves – like sound waves, water waves, and light waves – is characterized by the spreading or "flaring" of waves after passing through an aperture. For diffraction to occur, the size of the aperture must be comparable to the wavelength of the incident wave. A classic diffraction experiment was performed by Thomas Young in 1801, where a monochromatic beam of light is coherently split into two distinct beams after passing through two slits. In the region beyond the aperture, the split beams diffract and produce interference fringes on a screen placed a large distance away in comparison with the distance between the slits. As with other waves mentioned above, matter waves like electrons and neutrons also diffract after passing through an aperture. For instance, in electron beam diffraction, highly energetic electrons that are incident on a periodic crystalline structure diffract in a predictable manner prescribed by Bragg's law.

Beyond particles such as photons, electrons, and neutrons, Young's double slit experiment has been performed with atoms. In 1991, a highly excited helium atom was interfered after passing through two slits in gold foil. In contrast to the gold foil technique, most modern experiments use apertures made from light crystals to diffract atoms. This will be the focus of this chapter. We begin our discussion in Section 6.2 by reviewing the diffraction of light and matter waves. Next, we describe the interaction with detuned off-resonant light and solve the resulting dynamics under Bragg conditions and Raman transitions. The results of this chapter are used in the subsequent two chapters discussing metrology. For experiments that use the results we refer the reader to references found at the end of this chapter.

6.2 Theory of Diffraction

Diffraction gratings consist of many slits that are spaced at a constant interval of distance d between the slits. Gratings are often used to study different light sources. When light passes through a slit, it diffracts in accordance with Huygens's wave theory, where each slit acts as a source of light waves that produces a secondary wavefront. The wavefronts spread as they advance and give rise to diffraction. Thus, it is possible for a light wave from one slit to interfere with a light wave from another slit. The final interference pattern depends on the directional angle measured from the horizontal at which the rays are examined. To see this, consider a grating illuminated by a light wave of wavelength λ as shown in Fig. 6.1. Each slit on the grating is a

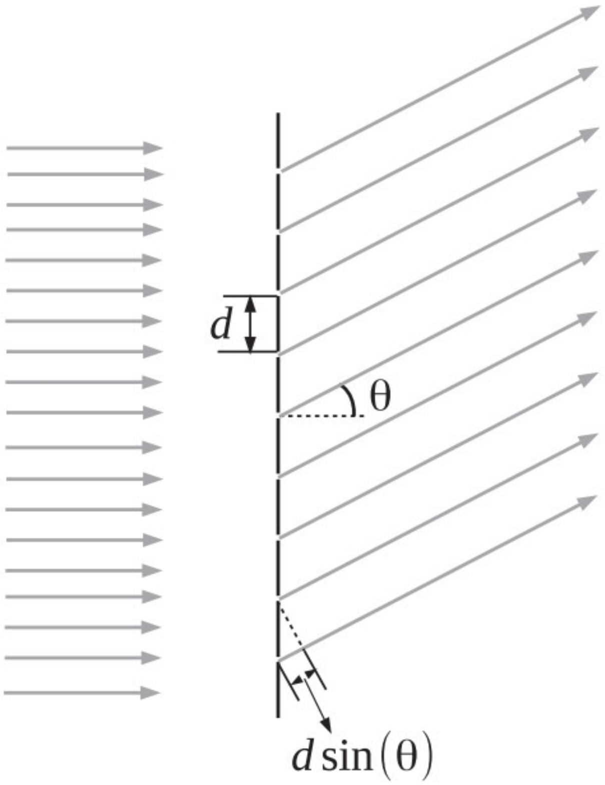

Fig. 6.1 The diffraction of light by a grating. Each slit on the grating is a source of light waves, and the slits are at equal distances *d* from one another. The wave from each slit is nearly parallel to the others, and the path difference of the light waves from any adjacent slits is $d \sin \theta$.

source of a light wave that emits rays of light that are mutually parallel to each other. They are also in phase as they leave the slit. As the rays propagate further toward a point on a screen, it is observed that the path difference between the rays depends on the angle a ray makes with the horizontal. For any two adjacent rays, the path difference between them is $d \sin \theta$. On the screen placed a large distance from the slit, bright bands are observed when all the waves arriving at the point on the screen are in phase. This is realized if the path difference is equal to a wavelength or integral multiples of a wavelength. Therefore, the maximum in a bright band is observed if the path difference is

$$d \sin \theta = m\lambda, \tag{6.1}$$

where $m = 0, \pm 1, \pm 2, \pm 3, \ldots$ is the diffraction order. To observe diffraction from the grating, the separation distance d between the slits should be large in comparison with the wavelength λ of light. Otherwise, the slits would appear to be the same source and no diffraction pattern would be observed. Because of this requirement, a diffraction grating made for one wave of a particular wavelength may not be suitable in the study of another wave of a very different wavelength.

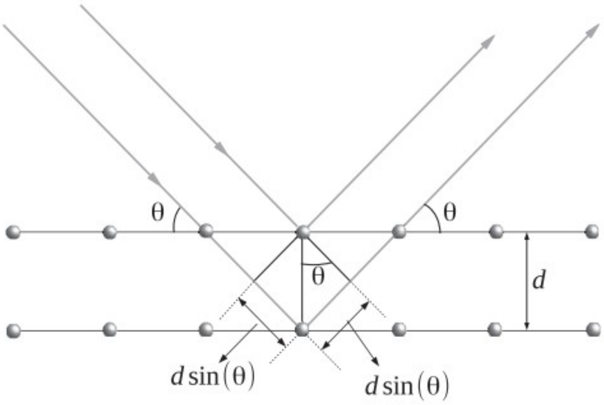

Fig. 6.2 The diffraction of a matter wave by a crystal. Each plane consists of atoms arranged in a regular fashion, and the planes are separated by a distance d. For waves incident on the crystal at angle θ to the plane, waves from the bottom plane travel an extra distance of $2d \sin\theta$ relative to those from the upper plane. Color version of this figure available at cambridge.org/quantumatomoptics.

Perhaps the simplest way to realize matter wave interference is to observe the diffraction of electrons in a crystal lattice. The slits in the crystal – NaCl, for instance – are made of atoms, which are arranged in a specific order. When matter waves are incident on a crystal, the atoms forming the crystal reflect the matter wave at regular intervals. The intensity of the emerging matter wave is very strong in a particular direction, which corresponds to constructive interference. For example, suppose that the incident matter wave makes an angle θ with the planes of the crystal, as shown in Fig. 6.2. Each plane is at a constant spacing d, and we consider two planes as shown in Fig. 6.2. Matter waves reflected from the upper plane travel a smaller distance compared with those reflected from the bottom plane. Waves from the two planes would reinforce at a point on the screen if the path difference $2d \sin\theta$ traveled by the waves is an integer multiple of the wavelength λ of the matter waves. Thus, a maximum exists within a constructive interference if

$$2d \sin\theta = m\lambda, \tag{6.2}$$

where $m = 1, 2, 3, \ldots$.

6.3 Ultra-cold Atom Interaction with Standing Light Wave

Material gratings, such as the crystals described in Section 6.2, work well to perform diffraction of particles such as electrons and neutrons but are less suited to diffraction of atoms. This is partly because atoms have a large mass compared to that of electrons and neutrons, which results in a much smaller de Broglie wavelength for a given velocity. Also, due to the charge neutrality of atoms, producing an energetic atom with high velocity was almost impossible until high-energy helium was produced via high-impact electron excitation. But even at that, the challenge still remained on how to diffract a stationary atom or an atom with negligible velocity, as found in ultra-cold

_20210925/ Quantum Atom...|

1108838596-Quantum-Atom....pdf

Pag. 100(115) - Fig 6.1

d sen θ.

the diffraction of light by a grating.

—

malavasi

—

0,01/0,4 = 0,025

—

20210926m

squeaky = estridente

haste make waste =

pressa faz desperdício

—

Books:

20210926/Modern Physics...|

Modern_physics_6e.pdf

Pag. 17(35) 1.3

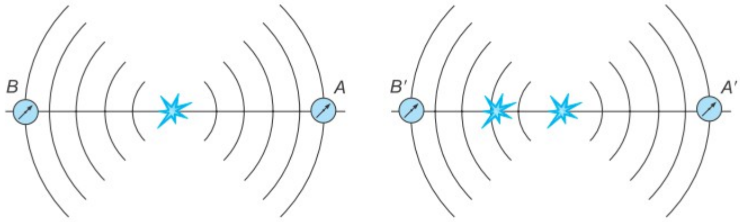

(*a*) Earth view of Earth clocks

(*b*) Spaceship view of spaceship clocks

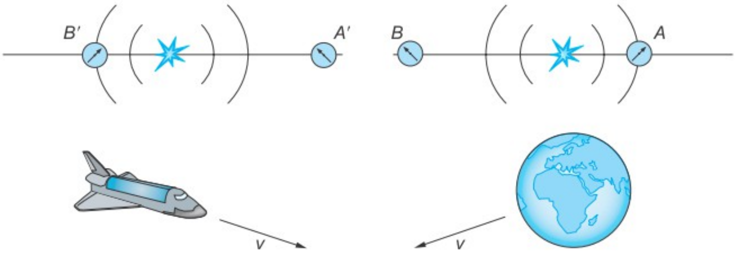

(*c*) Earth view of spaceship clocks

(*d*) Spaceship view of Earth clocks

FIGURE 1-16 A light flash occurs on Earth midway between two Earth clocks. At the instant of the flash the midpoint of a passing spaceship coincides with the light source. (*a*) The Earth clocks record the lights' arrival simultaneously and are thus synchronized. (*b*) Clocks at both ends of the spaceship also record the lights' arrival simultaneously (Einstein's second postulate) and they, too, are synchronized. (*c*) However, the Earth observer sees the light reach the clock at B' *before* the light reaches the clock at A'. Since the spaceship clocks read the same time when the light arrives, the Earth observer concludes that the clocks at A' and B' are not synchronized. (*d*) Illustrates that the spaceship observer similarly concludes that the Earth clocks are not synchronized.

more detail in the next section. Figure 1-16 illustrates the relativity of simultaneity from a different perspective.

Questions

3. In addition to the method described above, what would be another possible method of synchronizing all of the clocks in an inertial reference system?
4. Using Figure 1-16*d*, explain how the spaceship observer concludes that the Earth clocks are not synchronized.

1-3 The Lorentz Transformation

We now consider a very important consequence of Einstein's postulates, the general relation between the spacetime coordinates x, y, z and t of an event as seen in reference frame S and the coordinates x', y', z' and t' of the same event as seen in reference frame S', which is moving with uniform velocity relative to S. For simplicity we will consider only the special case in which the origins of the two coordinate systems are coincident at time $t = t' = 0$ and S' is moving, relative to S, with speed v along the x (or x') axis and with the y' and z' axes parallel, respectively, to the y and z axes as shown in Figure 1-17. As we discussed earlier (Equation 1-2), the classical Galilean coordinate transformation is

$$x' = x - vt \qquad y' = y \qquad z' = z \qquad t' = t \tag{1-2}$$

which expresses coordinate measurements made by an observer in S' in terms of those measured by an observer in S. The inverse transformation is

$$x = x' + vt' \qquad y = y' \qquad z = z' \qquad t = t'$$

and simply reflects the fact that the sign of the relative velocity of the reference frames is different for the two observers. The corresponding classical velocity transformation was given in Equation 1-3, and the acceleration, as we saw earlier, is invariant under a Galilean transformation. (For the rest of the discussion we will ignore the equations for y and z, which do not change in this special case of motion along the x and x' axes.) These equations are consistent with experiment as long as v is much less than c.

It should be clear that the classical velocity transformation is not consistent with the Einstein postulates of special relativity. If light moves along the x axis with speed c in S, Equation 1-3 implies that the speed in S' is $u'_x = c - v$ rather than $u'_x = c$. The Galilean transformation equations must therefore be modified to be consistent with Einstein's postulates, but the result must reduce to the classical equations when v is much less than c. We will give a brief outline of one method of obtaining the relativistic transformation that is called the *Lorentz transformation*, so named because of its original discovery by H. A. Lorentz.[12] We assume the equation for x' to be of the form

$$x' = \gamma(x - vt) \tag{1-11}$$

where γ is a constant that can depend on v and c but not on the coordinates. If this equation is to reduce to the classical one, γ must approach 1 as v/c approaches 0. The inverse transformation must look the same except for the sign of the velocity:

$$x = \gamma(x' + vt') \tag{1-12}$$

With the arrangement of the axes in Figure 1-17, there is no relative motion of the frames in the y and z directions; hence $y' = y$ and $z' = z$. However, insertion of the as yet unknown multiplier γ modifies the classical transformation of time, $t' = t$. To see this, we substitute x' from Equation 1-11 into Equation 1-12 and solve for t'. The result is

$$t' = \gamma\left[t + \frac{(1 - \gamma^2)}{\gamma^2}\frac{x}{v}\right] \tag{1-13}$$

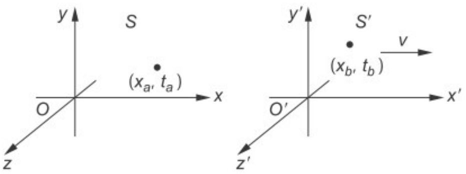

FIGURE 1-17 Two inertial frames S and S' with the latter moving at speed v in the $+x$ direction of system S. Each set of axes shown is simply the coordinate axes of a lattice like that in Figure 1-13. Remember, there is a clock at each intersection. A short time before the times represented by this diagram, O and O' were coincident and the lattices of S and S' were intermeshed.

Now let a flash of light start from the origin of S at $t = 0$. Since we have assumed that the origins coincide at $t = t' = 0$, the flash also starts at the origin of S' at $t' = 0$. The flash expands from *both* origins as a spherical wave. The equation for the wave front according to an observer in S is

$$x^2 + y^2 + z^2 = c^2t^2 \tag{1-14}$$

and according to an observer in S' it is

$$x'^2 + y'^2 + z'^2 = c^2t'^2 \tag{1-15}$$

where both equations are consistent with the second postulate. Consistency with the first postulate means that the relativistic transformation that we seek must transform Equation 1-14 into Equation 1-15, and vice versa. For example, substituting Equations 1-11 and 1-13 into 1-15 results in Equation 1-14 *if*

$$\gamma = \frac{1}{\sqrt{1 - \dfrac{v^2}{c^2}}} = \frac{1}{\sqrt{1 - \beta^2}} \tag{1-16}$$

where $\beta = v/c$. Notice that $\gamma = 1$ for $v = 0$ and $\gamma \to \infty$ for $v = c$. How this is done is illustrated in Example 1-2 below.

EXAMPLE 1-2 Relativistic Transformation Multiplier γ Show that γ must be given by Equation 1-16 if Equation 1-15 is to be transformed into Equation 1-14 consistent with Einstein's first postulate.

SOLUTION

Substituting Equations 1-11 and 1-13 into 1-15 and noting that $y' = y$ and $z' = z$ in this case yields

$$\gamma^2(x - vt)^2 + y^2 + z^2 = c^2\gamma^2\left[t + \frac{1 - \gamma^2}{\gamma^2}\frac{x}{v}\right]^2 \tag{1-17}$$

To be consistent with the first postulate, Equation 1-15 must be identical to Equation 1-14. This requires that the coefficient of the x^2 term in Equation 1-17 be equal to 1, that of the t^2 term be equal to c^2, and that of the xt term be equal to 0. Any of those conditions can be used to determine γ, and all yield the same result. Using, for example, the coefficient of x^2, we have from Equation 1-17 that

$$\gamma^2 - c^2\gamma^2\frac{(1 - \gamma^2)^2}{\gamma^4 v^2} = 1$$

which can be rearranged to

$$-c^2\frac{(1 - \gamma^2)^2}{\gamma^2 v^2} = (1 - \gamma^2)$$

Canceling $1 - \gamma^2$ on both sides and solving for γ yields

$$\gamma = \frac{1}{\sqrt{1 - \dfrac{v^2}{c^2}}}$$

With the value for γ found in Example 1-2, Equation 1-13 can be written in a somewhat simpler form and with it the complete Lorentz transformation becomes

$$x' = \gamma(x - vt) \qquad y' = y$$
$$t' = \gamma\left(t - \frac{vx}{c^2}\right) \qquad z' = z \tag{1-18}$$

and the inverse

$$x = \gamma(x' + vt') \qquad y = y'$$
$$t = \gamma\left(t' + \frac{vx'}{c^2}\right) \qquad z = z' \tag{1-19}$$

with

$$\gamma = \frac{1}{\sqrt{1 - \beta^2}}$$

EXAMPLE 1-3 Transformation of Time Intervals The arrivals of two cosmic-ray μ mesons (muons) are recorded by detectors in the laboratory, one at time t_a at location x_a and the second at time t_b at location x_b in the laboratory reference frame, S in Figure 1-17. What is the time interval between those two events in system S', which moves relative to S at speed v?

SOLUTION

Applying the time coordinate transformation from Equation 1-18,

$$t'_b - t'_a = \gamma\left(t_b - \frac{vx_b}{c^2}\right) - \gamma\left(t_a - \frac{vx_a}{c^2}\right)$$
$$t'_b - t'_a = \gamma(t_b - t_a) - \frac{\gamma v}{c^2}(x_b - x_a) \tag{1-20}$$

We see that the time interval measured in S' depends not just on the corresponding time interval in S, but also on the spatial separation of the clocks in S that measured the interval. This result should not come as a total surprise since we have already discovered that, although the clocks in S are synchronized with each other, they are not, in general, synchronized for observers in other inertial frames.

Special Case 1

If it should happen that the two events occur at the same location in S, that is, $x_a = x_b$, then $(t_b - t_a)$, the time interval measured on a clock located at the events, is called the *proper time interval.* Notice that, since $\gamma > 1$ for all frames moving relative to S, the proper time interval is the *minimum* time interval that can be measured between those events.

Special Case 2

Does there exist an inertial frame for which the events described above would be measured to be simultaneous? Since the question has been asked, you probably suspect that the answer is yes, and you are right. The two events will be simultaneous in a system S'' for which $t''_b - t''_a = 0$, that is, when

$$\gamma(t_b - t_a) = \frac{\gamma v}{c^2}(x_b - x_a)$$

or when

$$\beta = \frac{v}{c} = \left(\frac{t_b - t_a}{x_b - x_a}\right)c \qquad \textbf{1-21}$$

Notice that $(x_b - x_a)/c =$ time for a light beam to travel from x_a to x_b; thus we can characterize S'' as being that system whose speed relative to S is that fraction of c given by the time interval between the events divided by the travel time of light between them. (Note, too, that $c(t_b - t_a) > (x_b - x_a)$ implies that $\beta > 1$, a nonphysical situation that we will discuss in Section 1-4.)

While it is possible for us to get along in special relativity without the Lorentz transformation, it has an application that is quite valuable: it enables the spacetime coordinates of events measured by the measuring rods and clocks in the reference frame of one observer to be translated into the corresponding coordinates determined by the measuring rods and clocks of an observer in another inertial frame. As we will see in Section 1-4, such transformations lead to some startling results.

Relativistic Velocity Transformations

The transformation for velocities in special relativity can be obtained by differentiation of the Lorentz transformation, keeping in mind the definition of the velocity. Suppose a particle moves in S with velocity $\mathbf{u}$ whose components are $u_x = dx/dt$, $u_y = dy/dt$, and $u_z = dz/dt$. An observer in S' would measure the components $u'_x = dx'/dt'$, $u'_y = dy'/dt'$, and $u'_z = dz'/dt'$. Using the transformation equations, we obtain

$$dx' = \gamma(dx - vdt) \qquad dy' = dy$$

$$dt' = \gamma\left(dt - \frac{vdx}{c^2}\right) \qquad dz' = dz$$

from which we see that u'_x is given by

$$u'_x = \frac{dx'}{dt'} = \frac{\gamma(dx - vdt)}{\gamma\left(dt - \dfrac{vdx}{c^2}\right)} = \frac{(dx/dt - v)}{1 - \dfrac{v}{c^2}\dfrac{dx}{dt}}$$

or

$$u'_x = \frac{u_x - v}{1 - \dfrac{vu_x}{c^2}} \qquad \textbf{1-22}$$

and, if a particle has velocity components in the y and z directions, it is not difficult to find the components in S' in a similar manner.

$$u'_y = \frac{u_y}{\gamma\left(1 - \dfrac{vu_x}{c^2}\right)} \qquad u'_z = \frac{u_z}{\gamma\left(1 - \dfrac{vu_x}{c^2}\right)}$$

Remember that this form of the velocity transformation is specific to the arrangement of the coordinate axes in Figure 1-17. Note, too, that when $v \ll c$, that is, when $\beta = v/c \approx 0$ the relativistic velocity transforms reduce to the classical velocity addition of Equation 1-3. Likewise, the inverse velocity transformation is

$$u_x = \frac{u'_x + v}{\left(1 + \frac{vu'_x}{c^2}\right)} \qquad u_y = \frac{u'_y}{\gamma\left(1 + \frac{vu'_x}{c^2}\right)} \qquad u_z = \frac{u'_z}{\gamma\left(1 + \frac{vu'_x}{c^2}\right)} \qquad \textbf{1-23}$$

EXAMPLE 1-4 Relative Speeds of Cosmic Rays Suppose that two cosmic-ray protons approach Earth from opposite directions as shown in Figure 1-18*a*. The speeds relative to Earth are measured to be $v_1 = 0.6c$ and $v_2 = -0.8c$. What is Earth's velocity relative to each proton, and what is the velocity of each proton relative to the other?

SOLUTION

Consider each particle and Earth to be inertial reference frames S', S'', and S, with their respective x axes parallel as in Figure 1-18*b*. With this arrangement $v_1 = u_{1x} = 0.6c$ and $v_2 = u_{2x} = -0.8c$. Thus, the speed of Earth measured in S' is $v'_{Ex} = -0.6c$ and the speed of Earth measured in S'' is $v''_{Ex} = 0.8c$.

To find the speed of proton 2 with respect to proton 1, we apply Equation 1-22 to compute u'_{2x}, that is, the speed of particle 2 in S'. Its speed in S has been measured to be $u_{2x} = -0.8c$, where the S' system has relative speed $v_1 = 0.6c$ with respect to S. Thus, substituting into Equation 1-22, we obtain

$$u'_{2x} = \frac{-0.8c - (0.6c)}{1 - (0.6c)(-0.8c)/c^2} = \frac{-1.4c}{1.48} = -0.95c$$

and the first proton measures the second to be approaching (moving in the $-x'$ direction) at $0.95c$.

The observer in S'' must of course make a consistent measurement, that is, find the speed of proton 1 to be $0.95c$ in the $+x''$ direction. This can be readily shown by a second application of Equation 1-22 to compute u''_{1x}.

$$u''_{1x} = \frac{0.6c - (-0.8c)}{1 - (0.6c)(-0.8c)/c^2} = \frac{1.4c}{1.48} = 0.95c$$

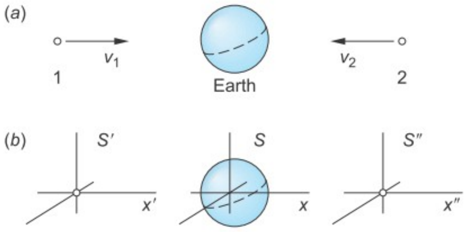

FIGURE 1-18 (*a*) Two cosmic-ray protons approach Earth from opposite directions at speeds v_1 and v_2 with respect to Earth. (*b*) Attaching an inertial frame to each particle and Earth enables one to visualize the several relative speeds involved and apply the velocity transformation correctly.

Questions

5. The Lorentz transformation for y and z is the same as the classical result: $y = y'$ and $z = z'$. Yet the relativistic velocity transformation does not give the classical result $u_y = u'_y$ and $u_z = u'_z$. Explain.
6. Since the velocity components of a moving particle are different in relatively moving frames, the *directions* of the velocity vectors are also different, in general. Explain why the fact that observers in S and S' measure different directions for a particle's motion is not an inconsistency in their observations.

Spacetime Diagrams

The relativistic discovery that time intervals between events are not the same for observers in different inertial reference frames underscores the four-dimensional character of spacetime. With the diagrams that we have used thus far, it is difficult to depict and visualize on the two-dimensional page events that occur at different times since each diagram is equivalent to a snapshot of spacetime at a particular instant. Showing events as a function of time typically requires a series of diagrams, such as Figures 1-14, 1-15, and 1-16, but even then our attention tends to be drawn to the space coordinate systems rather than the events, whereas it is the *events* that are fundamental. This difficulty is removed in special relativity with a simple, yet powerful graphing method called the *spacetime diagram*. (This is just a new name given to the t vs. x graphs that you first began to use when you discussed motion in introductory physics.) On the spacetime diagram we can graph both the space and time coordinates of many events in one or more inertial frames, albeit with one limitation. Since the page offers only two dimensions for graphing, we suppress, or ignore for now, two of the space dimensions, in particular y and z. With our choice of the relative motion of inertial frames along the x axis, $y' = y$ and $z' = z$ anyhow. (This is one of the reasons we made that convenient choice a few pages back, the other reason being mathematical simplicity.) This means that for the time being, we are limiting our attention to one space dimension and to time, that is, to events that occur, regardless of when, along one line in space. Should we need the other two dimensions, for example, in a consideration of velocity vector transformations, we can always use the Lorentz transformation equations.

In a spacetime diagram the space location of each event is plotted along the x axis horizontally and the time is plotted vertically. From the three-dimensional array of measuring rods and clocks in Figure 1-13, we will use only those located on the x axis as in Figure 1-19. (See, things are simpler already!) Since events that exhibit relativistic effects generally occur at high speeds, it will be convenient to multiply the time scale by the speed of light (a constant), which enables us to use the same units and scale on both the space and time axes, for example, meters of distance and meters of light travel time.[13] The time axis is, therefore, c times the time t in seconds, that is, ct. As we will see shortly, this choice prevents events from clustering about the axes and makes possible the straightforward addition of other inertial frames into the diagram.

As time advances, notice that in Figure 1-19 each clock in the array moves vertically upward along the dotted lines. Thus, as events A, B, C, and D occur in spacetime, one of the clocks of the array is at (or very near) each event when it happens. Remembering that the clocks in the reference frame are synchronized, you will see that the difference in the readings of clocks located at each event records the *proper time interval* between the events (see Example 1-3). In the figure, events A and D

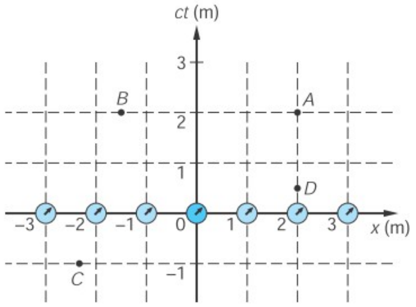

FIGURE 1-19 Spacetime diagram for an inertial reference frame S. Two of the space dimensions (y and z) are suppressed. The units on both the space and time axes are the same, meters. A meter of time means the time required for light to travel one meter, that is, 3.3×10^{-9} s.

occur at the same place ($x = 2$ m), but at different times. The time interval between them measured on clock 2 is the proper time interval since clock 2 is located at *both* events. Events A and B occur at different locations, but at the same time (i.e., simultaneously in this frame). Event C occurred before the present since $ct = -1$ m. For this discussion we will consider the time that the coordinate origins coincide, $ct = ct' = 0$, to be the present.

Worldlines in Spacetime Particles moving in space trace out a line in the spacetime diagram called the *worldline* of the particle. The worldline is the "trajectory" of the particle on a ct versus x graph. To illustrate, consider four particles moving in space (not spacetime) as shown in Figure 1-20*a*, which shows the array of synchronized clocks on the x axis and the space trajectories of four particles, each starting at $x = 0$ and moving at some constant speed, during 3 m of time. Figure 1-20*b* shows the worldline for each of the particles in spacetime. Notice that constant speed means that the worldline has constant slope; that is, it is a straight line (slope $= \Delta t/\Delta x = 1/(\Delta x/\Delta t) = 1/\text{speed}$). That was also the case when you first encountered elapsed time versus displacement graphs in introductory physics. Even then, you were plotting spacetime graphs and drawing worldlines! If the particle is accelerating—either speeding up like particle 5 in Figure 1-20*c* or slowing down like particle 6—the worldlines are curved. Thus, the worldline is the record of the particle's travel through spacetime, giving its speed (= 1/slope) and acceleration (= 1/rate at which the slope changes) at every instant.

EXAMPLE 1-5 Computing Speeds in Spacetime Find the speed u of particle 3 in Figure 1-20.

SOLUTION

The speed $u = \Delta x/\Delta t = 1/\text{slope}$ where we have $\Delta x = 1.5 - 0 = 1.5$ m and $c\Delta = c \cdot \Delta t = 3.0 - 0 = 3.0$ m (from Figure 1-20). Thus, $\Delta t = (3.0/c) = (3.0/3.0 \times 10^8) = 10^{-8}$ s and $u = 1.5 \text{ m}/10^{-8} \text{ s} = 1.5 \times 10^8 \text{ m/s} = 0.5c$.

The speed of particle 4, computed as shown in Example 1-5, turns out to be c, the speed of light. (Particle 4 is a light pulse.) The slope of its worldline $\Delta(ct)/\Delta x = 3 \text{ m}/3 \text{ m} = 1$. Similarly, the slope of the worldline of a light pulse moving in the

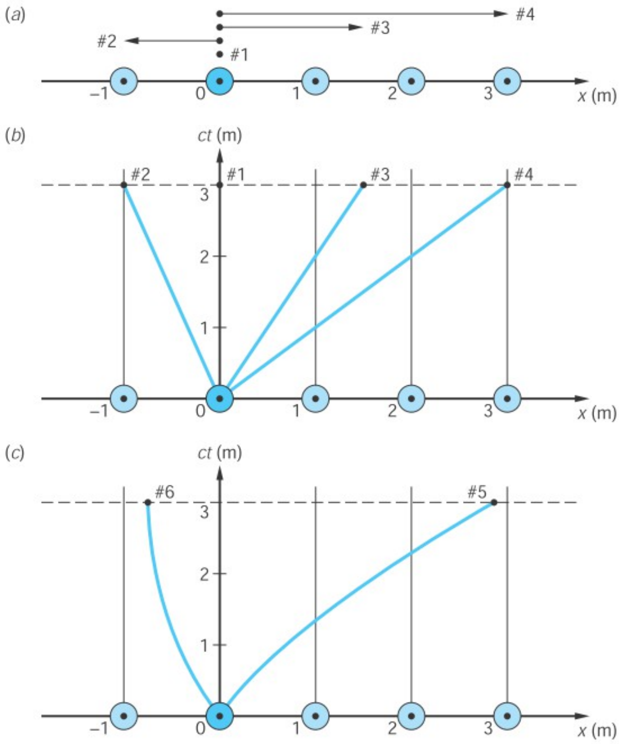

FIGURE 1-20 (*a*) The space trajectories of four particles with various constant speeds. Note that particle 1 has a speed of zero and particle 2 moves in the $-x$ direction. The worldlines of the particles are straight lines. (*b*) The worldline of particle 1 is also the ct axis since that particle remains at $x = 0$. The constant slopes are a consequence of the constant speeds. (*c*) For accelerating particles 5 and 6 [not shown in (*a*)], the worldlines are curved, the slope at any point yielding the instantaneous speed.

$-x$ direction is -1. Since relativity limits the speed of particles with mass to less than c, as we will see in Chapter 2, the slopes of worldlines for particles that move through $x = 0$ at $ct = 0$ are limited to the larger shaded triangle in Figure 1-21. The same limits to the slope apply at every point along a particle's worldline, such as point A on the curved spacetime trajectory in Figure 1-21. This means that the particle's possible worldlines for times greater than $ct = 2$ m must lie within the heavily shaded triangle.

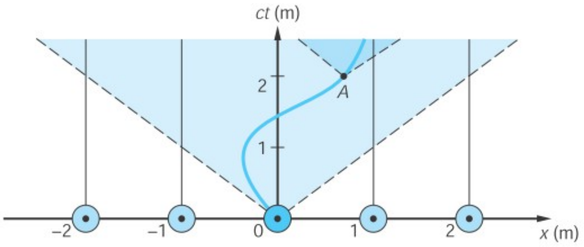

FIGURE 1-21 The speed-of-light limit to the speeds of particles limits the slopes of worldlines for particles that move through $x = 0$ at $ct = 0$ to the shaded area of spacetime, that is, to slopes < -1 and $> +1$. The dashed lines are worldlines of light flashes moving in the $-x$ and $+x$ directions. The curved worldline of the particle shown has the same limits at every instant. Notice that the particle's speed = 1/slope.

Event Analysis Using Worldlines Analyzing events and motion in inertial systems that are in relative motion can now be accomplished more easily than with diagrams such as Figures 1-14 through 1-18. Suppose we have two inertial frames S and S' with S' moving in the $+x$ direction of S at speed v as in those figures. The clocks in both systems are started at $t = t' = 0$ (the present) as the two origins $x = 0$ and $x' = 0$ coincide, and, as before, observers in each system have synchronized the clocks in their respective systems. The spacetime diagram for S is, of course, like that in Figure 1-19, but how does S' appear in that diagram, that is, with respect to an observer in S? Consider that, as the origin of S' (i.e., the point where $x' = 0$) moves in S, its worldline is the ct' axis since the ct' axis is the locus of all points with $x' = 0$ (just as the ct axis is the locus of points with $x = 0$). Thus, the slope of the ct' axis as seen by an observer in S can be found from Equation 1-18, the Lorentz transformation, as follows

$$x' = \gamma(x - vt) = 0 \quad \text{for} \qquad x' = 0$$

or

$$x = vt = (v/c)(ct) = \beta ct$$

and

$$ct = (1/\beta)x$$

which says that the slope (in S) of the worldline of the point $x' = 0$, the ct' axis, is $1/\beta$ (see Figure 1-22*a*).

In the same manner, the x' axis can be located using the fact that it is the locus of points for which $ct' = 0$. The Lorentz transformation once again provides the slope:

$$t' = \gamma\left(t - \frac{vx}{c^2}\right) = 0$$

or

$$t = \frac{vx}{c^2} \qquad \text{and} \qquad ct = \frac{v}{c}x = \beta x$$

Thus, the slope of the x' axis as measured by an observer in S is β, as shown in Figure 1-22*a*. Don't be confused by the fact that the x axes don't look parallel anymore. They are still parallel in *space*, but this is a *spacetime* diagram. It shows motion in both space and time. For example, the clock at $x' = 1$ m in Figure 1-22*b* passed the point $x = 0$ at about $ct = -1.5$ m as the x' axis of S' moved both upward and to the right in S. Remember, as time advances, the array of synchronized clocks and measuring rods that are the x axis also moves upward, so that, for example, when $ct = 1$, the origin of $S'(x' = 0, ct' = 0)$ has moved $vt = (v/c)ct = \beta ct$ to the right along the x axis.

Question

7. Explain how the spacetime diagram in Figure 1-22*b* would appear drawn by an observer in S'.

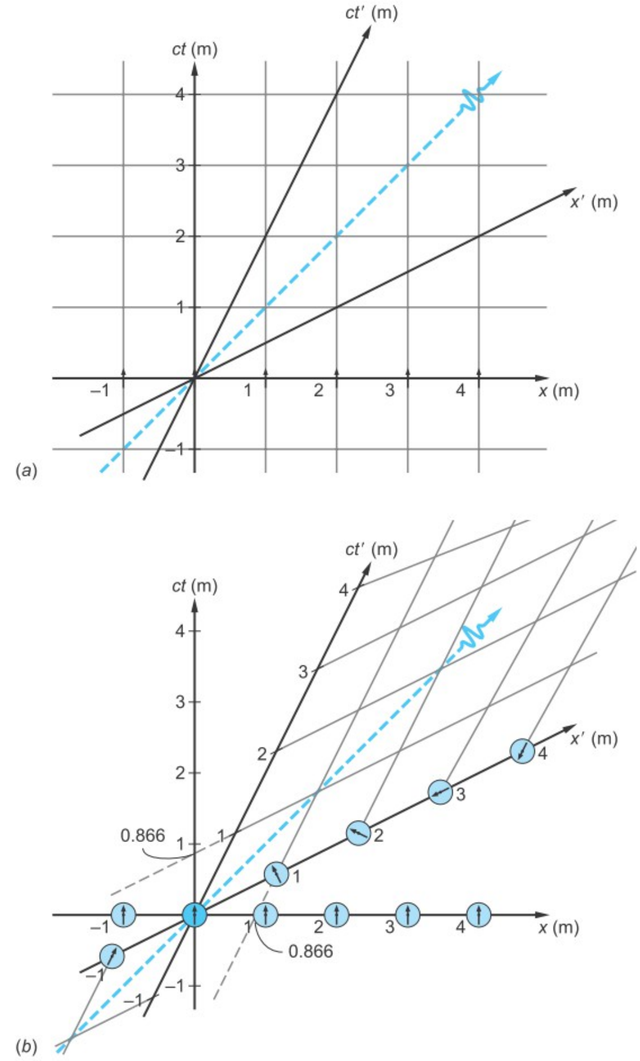

FIGURE 1-22 Spacetime diagram of S showing S' moving at speed $v = 0.5c$ in the $+x$ direction. The diagram is drawn with $t = t' = 0$ when the origins of S and S' coincided. The dashed line shows the worldline of a light flash that passed through the point $x = 0$ at $t = 0$ heading in the $+x$ direction. Its slope equals 1 in both S and S'. The ct' and x' axes of S' have slopes of $1/\beta = 2$ and $\beta = 0.5$, respectively. (*a*) Calibrating the axes of S' as described in Exploring (pages 26–27) allows the grid of coordinates to be drawn on S'. Interpretation is facilitated by remembering that (*b*) shows the system S' as it is observed in the spacetime diagram of S.

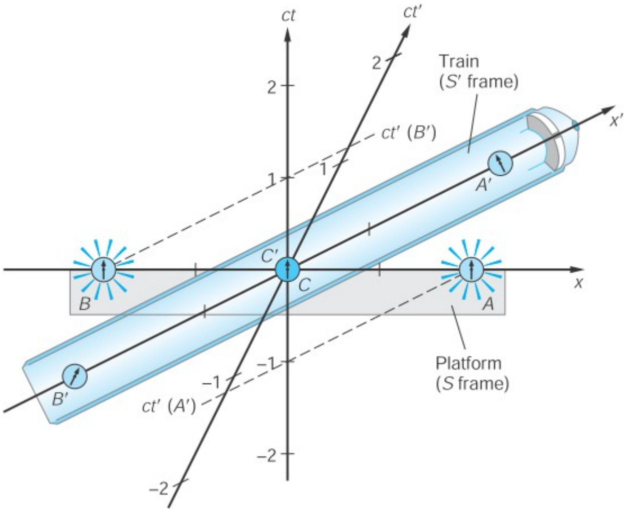

FIGURE 1-23 Spacetime equivalent of Figure 1-15, showing the spacetime diagram for the system *S* in which the platform is at rest. Measurements made by observers in *S*′ are read from the primed axes.

EXAMPLE 1-6 **Simultaneity in Spacetime** Use the train-platform example of Figure 1-15 and a suitable spacetime diagram to show that events simultaneous in one frame are not simultaneous in a frame moving relative to the first. (This is the corollary to the relativity of simultaneity that we first demonstrated in the previous section using Figure 1-15.)

SOLUTION

Suppose a train is passing a station platform at speed v and an observer C at the midpoint of the platform, system S, announces that light flashes will be emitted at clocks A and B located at opposite ends of the platform at $t = 0$. Let the train, system S', be a rocket train with $v = 0.5c$. As in the earlier discussion, clocks at C and C' both read 0 as C' passes C. Figure 1-23 shows this situation. It is the spacetime equivalent of Figure 1-15.

Two events occur; the light flashes. The flashes are simultaneous in S since both occur at $ct = 0$. In S', however, the event at A occurred at $ct'(A')$ (see Figure 1-23), about 1.2 ct' units *before* $ct' = 0$, and the event at B occurred at $ct'(B')$, about 1.2 ct' units *after* $ct' = 0$. Thus, the flashes are not simultaneous in S' and A occurs before B, as we also saw in Figure 1-15.

EXPLORING

Calibrating the Spacetime Axes

By calibrating the coordinate axes of S' consistent with the Lorentz transformation, we will be able to read the coordinates of events and calculate space and time intervals between events as measured in both S and S' directly from the diagram, in addition to calculating them from Equations 1-18 and 1-19. The calibration of the S' axes is straightforward and is accomplished as follows. The locus of points, for example, with

$x' = 1$ m, is a line parallel to the ct' axis through the point $x' = 1$ m, $ct' = 0$, just as we saw earlier that the ct' axis was the locus of those points with $x' = 0$ through the point $x' = 0$, $ct' = 0$. Substituting these values into the Lorentz transformation for x', we see that the line through $x' = 1$ m intercepts the x axis, that is, the line where $ct = 0$ at

$$x' = \gamma(x - vt) = \gamma(x - \beta ct) \qquad \textbf{1-24}$$
$$1 = \gamma x \qquad \text{or} \qquad x = 1/\gamma = \sqrt{1 - \beta^2}$$

or, in general,

$$x = x'\sqrt{1 - \beta^2}$$

In Figure 1-22*b*, where $\beta = 0.5$, the line $x' = 1$ m intercepts the x axis at $x = 0.866$ m. Similarly, if $x' = 2$ m, $x = 1.73$ m; if $x' = 3$ m, $x = 2.60$ m, and so on.

The ct' axis is calibrated in a precisely equivalent manner. The locus of points with $ct' = 1$ m is a line parallel to the x' axis through the point $ct' = 1$ m, $x' = 0$. Using the Lorentz transformation, the intercept of that line with the ct axis (where $x = 0$) is found as follows.

$$t' = \gamma(t - vx/c^2)$$

which can also be written as

$$ct' = \gamma(ct - \beta x) \qquad \textbf{1-25}$$

or $ct' = \gamma ct$ for $x = 0$. Thus, for $ct' = 1$ m, we have $1 = \gamma ct$ or $ct = (1 - \beta^2)^{1/2}$ and, again, in general, $ct = ct'(1 - \beta^2)^{1/2}$. The $x' \cdot ct'$ coordinate grid is shown in Figure 1-22*b*.

Notice in Figure 1-22*b* that the clocks located in S' are *not* found to be synchronized by observers in S, even though they are synchronized in S'. This is exactly the conclusion that we arrived at in the discussion of the lightning striking the train and platform. In addition, those with positive x' coordinates are behind the S' reference clock and those with negative x' coordinates are ahead, the differences being greatest for those clocks farthest away. This is a direct consequence of the Lorentz transformation of the time coordinate—that is, when $ct = 0$ in Equation 1-25, $ct' = -\gamma\beta x$. Note, too, that the slope of the worldline of the light beam equals 1 in S', as well as in S, as required by the second postulate.

1-4 Time Dilation and Length Contraction

The results of correct measurements of the time and space intervals between events do not depend on the kind of apparatus used for the measurements or on the events themselves. We are free therefore to choose any events and measuring apparatus that will help us understand the application of the Einstein postulates to the results of measurements. As you have already seen from previous examples, convenient events in relativity are those that produce light flashes. A convenient, simple such clock is a *light clock*, pictured schematically in Figure 1-24. A photocell detects the light pulse and sends a voltage pulse to an oscilloscope, which produces a vertical deflection of the oscilloscope's trace. The phosphorescent material on the face of the oscilloscope tube gives a persistent light that can be observed visually, photographed, or recorded electronically. The time between two light flashes is determined by measuring the distance between pulses on the scope and knowing the sweep speed. Such clocks can easily be calibrated and compared with other types of clocks. Although not drawn as in Figure 1-24, the clocks used in explanations in this section may be thought of as light clocks.

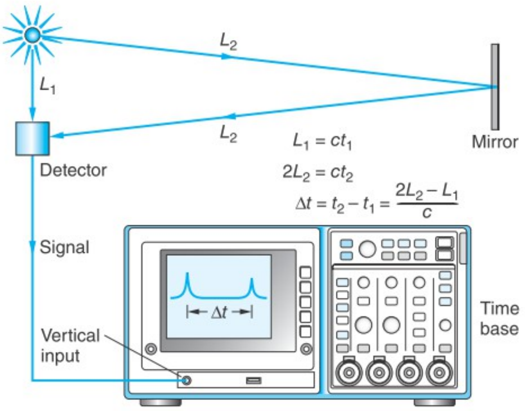

FIGURE 1-24 Light clock for measuring time intervals. The time is measured by reading the distance between pulses on the oscilloscope after calibrating the sweep speed.

Time Dilation (or Time Stretching)

We first consider an observer A' at rest in frame S' a distance D from a mirror, also in S', as shown in Figure 1-25*a*. He triggers a flash gun and measures the time interval $\Delta t'$ between the original flash and the return flash from the mirror. Since light travels with speed c, this time is $\Delta t' = (2D)/c$.

We now consider these same two events, the original flash of light and the returning flash, as observed in reference frame S, with respect to which S' is moving to the right with speed v. The events happen at two different places, x_1 and x_2, in frame S because between the original flash and the return flash observer A' has moved a horizontal distance $v\Delta t$, where Δt is the time interval between the events measured in S. In Figure 1-25*b*, a space diagram, we see that the path traveled by the light is longer in S than in S'. However, by Einstein's postulates, light travels with the same speed c in frame S as is does in frame S'. Since it travels farther in S at the same speed, it takes longer in S to reach the mirror and return. The time interval between flashes in S is thus longer than it is in S'. We can easily calculate Δt in terms of $\Delta t'$. From the triangle in Figure 1-25*c*, we see that

$$\left(\frac{c\Delta t}{2}\right)^2 = D^2 + \left(\frac{v\Delta t}{2}\right)^2$$

or

$$\Delta t = \frac{2D}{\sqrt{c^2 - v^2}} = \frac{2D}{c}\frac{1}{\sqrt{1 - v^2/c^2}}$$

Using $\Delta t' = 2D/c$, we have

$$\Delta t = \frac{\Delta t'}{\sqrt{1 - v^2/c^2}} = \gamma\Delta t' = \gamma\tau \qquad \mathbf{1\text{-}26}$$

where $\tau = \Delta t'$ is the *proper time interval* that we first encountered in Example 1-3. Equation 1-26 describes *time dilation*; that is, it tells us that the observer in frame S always measures the time interval between two events to be longer (since $\gamma > 1$) than the corresponding interval measured on the clock located at both events in the frame where they occur at the same location. Thus, observers in S conclude that the clock

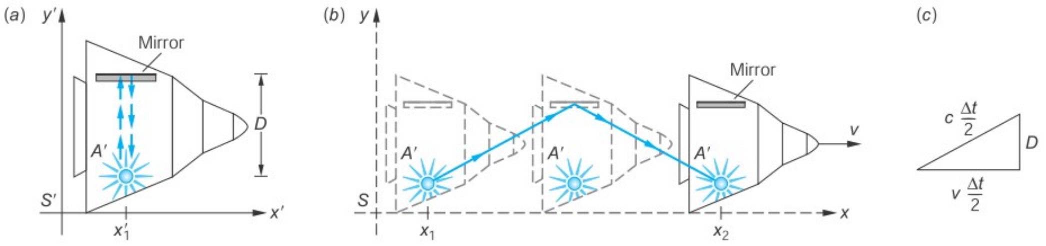

FIGURE 1-25 (*a*) Observer A' and the mirror are in a spaceship at rest in frame S'. The time it takes for the light pulse to reach the mirror and return is measured by A' to be $2D/c$. (*b*) In frame *S*, the spaceship is moving to the right with speed *v*. If the speed of light is the same in both frames, the time it takes for the light to reach the mirror and return is longer than $2D/c$ in *S* because the distance traveled is greater than $2D$. (*c*) A right triangle for computing the time Δt in frame *S*.

at A' in S' runs slow since that clock measures a smaller time interval between the two events. Notice that the faster S' moves with respect to *S*, the larger is γ, and the slower the S' clocks will tick. It appears to the *S* observer that time is being stretched out in S'.

Be careful! The *same* clock must be located at each event for $\Delta t'$ to be the proper time interval τ. We can see why this is true by noting that Equation 1-26 can be obtained directly from the inverse Lorentz transformation for *t*. Referring again to Figure 1-25 and calling the emission of the flash event 1 and its return event 2, we have that

$$\Delta t = t_2 - t_1 = \gamma\left(t_2' + \frac{vx_2'}{c^2}\right) - \gamma\left(t_1' + \frac{vx_1'}{c^2}\right)$$

$$\Delta t = \gamma(t_2' - t_1') + \frac{\gamma v}{c^2}(x_2' - x_1')$$

or

$$\Delta t = \gamma\Delta t' + \frac{\gamma v}{c^2}\Delta x' \qquad \textbf{1-27}$$

If the clock that records t_2' and t_1' is located at the events, then $\Delta x' = 0$. If that is not the case, however, $\Delta x' \neq 0$ and $\Delta t'$, though certainly a valid measurement, is not a proper time interval. Only a clock located *at* an event *when* it occurs can record a proper time interval.

EXAMPLE 1-7 **Spatial Separation of Events** Two events occur at the same point x_0' at times t_1' and t_2' in S', which moves with speed *v* relative to *S*. What is the spatial separation of these events measured in *S*?

SOLUTION

1. The location of the events in *S* is given by the Lorentz inverse transformation, Equation 1-19: $x = \gamma(x' + vt')$
2. The spatial separation of the two events $\Delta x = x_2 - x_1$ is then $\Delta x = \gamma(x_0' + vt_2') - \gamma(x_0' + vt_1')$

3. The $\gamma x'_0$ terms cancel: $\Delta x = \gamma v(t'_2 - t'_1) = \gamma v \Delta t'$
4. Since $\Delta t'$ is the proper time interval τ, Equation 1-26 yields $\Delta x = v\gamma\tau = v\Delta t$
5. Using the situation in Figure 1-26 as a numerical example, where $\beta = 0.5$ and $\gamma = 1.15$, we have
$$\Delta x = \gamma \frac{v}{c} \Delta(ct') = (1.15)(0.5)(2) = 1.15 \text{ m}$$

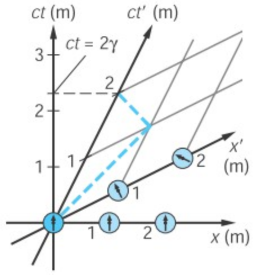

FIGURE 1-26 Spacetime diagram illustrating time dilation. The dashed line is the worldline of a light flash emitted at $x' = 0$ and reflected back to that point by a mirror at $x' = 1$ m. $\beta = 0.5$.

EXAMPLE 1-8 The Pregnant Elephant[14] Elephants have a gestation period of 21 months. Suppose that a freshly impregnated elephant is placed on a spaceship and sent toward a distant space jungle at $v = 0.75c$. If we monitor radio transmissions from the spaceship, how long after launch might we expect to hear the first squealing trumpet from the newborn calf?

SOLUTION

1. In S', the rest frame of the elephant, the time interval from launch to birth is $\tau = 21$ months. In the Earth frame S the time interval is Δt_1 given by Equation 1-26:
$$\Delta t_1 = \gamma\tau = \frac{1}{\sqrt{1-\beta^2}}\tau = \frac{1}{\sqrt{1-(0.75)^2}}(21 \text{ months}) = 31.7 \text{ months}$$
2. At that time the radio signal announcing the happy event starts toward Earth at speed c, but from where? Using the result of Example 1-7, since launch the spaceship has moved Δx in S given by
$$\Delta x = \gamma v\tau = \gamma\beta c\tau = (1.51)(0.75)(21\ c \cdot \text{months}) = 23.8c \cdot \text{months}$$
where $c \cdot$ month is the distance light travels in one month.
3. Notice that there is no need to convert Δx into meters since our interest is in how long it will take the radio signal to travel this distance in S. That time is Δt_2 given by
$$\Delta t_2 = \Delta x/c = 23.8c \cdot \text{months}/c = 23.8 \text{ months}$$
4. Thus, the good news will arrive at Earth at time Δt after launch where
$$\Delta t = \Delta t_1 + \Delta t_2 = 31.7 + 23.8 = 55.5 \text{ months}$$

Remarks: *This result, too, is readily obtained from a spacetime diagram. Figure 1-27 illustrates the general appearance of the spacetime diagram for this example, showing the elephant's worldline and the worldline of the radio signal.*

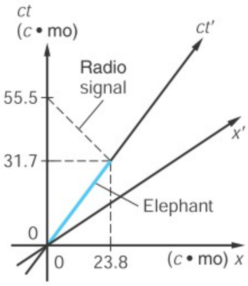

FIGURE 1-27 Sketch of the spacetime diagram for Example 1-8. $\beta = 0.75$. The colored line is the worldline of the pregnant elephant. The worldline of the radio signal is the dashed line at 45° toward the upper left.

Question

8. You are standing on a corner and a friend is driving past in an automobile. Both of you note the times when the car passes two different intersections and determine from your watch readings the time that elapses between the two events. Which of you has determined the proper time interval?

The time dilation of Equation 1-26 is easy to see in a spacetime diagram such as Figure 1-26, using the same round trip for a light pulse used above. Let the light flash leave $x' = 0$ at $ct' = 0$, when the S and S' origins coincided. The flash travels to $x' = 1$ m, reflects from a mirror located there, and returns to $x' = 0$. Let $\beta = 0.5$. The dotted line shows the worldline of the light beam, reflecting at ($x' = 1$, $ct' = 1$) and returning to $x' = 0$ at $ct' = 2$ m. Note that the S observer records the latter event at $ct > 2$ m; that is, the observer in S sees the S' clock running slow.

Experimental tests of the time dilation prediction have been performed using macroscopic clocks, in particular, accurate atomic clocks. In 1975, C. O. Alley conducted a test of both general and special relativity in which a set of atomic clocks were carried by a U.S. Navy antisubmarine patrol aircraft while it flew back and forth over the same path for 15 hours at altitudes between 8000 m and 10,000 m over Chesapeake Bay. The clocks in the plane were compared by laser pulses with an identical group of clocks on the ground. (See Figure 1-13 for one way such a comparison might be done.) Since the experiment was primarily intended to test the gravitational effect on clocks predicted by general relativity (see Section 2-5), the aircraft was deliberately flown at the rather sedate average speed of 270 knots (140 m/s) $= 4.7 \times 10^{-7}c$ so as to minimize the time dilation due to the relative speeds of the clocks. Even so, after deducting the effect of gravitation as predicted by general relativity, the airborne clocks lost an average of 5.6×10^{-9} s due to the relative speed during the 15-hour flight. This result agrees with the prediction of special relativity, 5.7×10^{-9} s to within 2 percent, even at this low relative speed. More recently, in 2010 J. C.-W. Chou and his coworkers at the National Institute of Science and Technology (NIST) used precision optical clocks to detect the minuscule time dilation at a speed of only 10 m/s, about the speed of a collegiate track sprinter. These and other experimental results leave little basis for further debate as to whether traveling clocks of all kinds lose time. They do.

Length Contraction

A phenomenon closely related to time dilation is *length contraction.* The length of an object measured in the reference frame in which the object is at rest is called its *proper length* L_p. In a reference frame in which the object is moving, the measured length parallel to the direction of motion is shorter than its proper length. Consider a rod at rest in the frame S' with one end at x'_2 and the other end at x'_1 as illustrated in Figure 1-28. The length of the rod in this frame is its proper length $L_p = x'_2 - x'_1$. Some care must be taken to find the length of the rod in frame S. In this frame, the rod is moving to the right with speed v, the speed of frame S'. The length of the rod in frame S is *defined* as $L = x_2 - x_1$, where x_2 is the position of one end at some time t_2 and x_1 is the position of the other end *at the same time* $t_1 = t_2$ as measured in frame S. Since the rod is at rest in S', t'_2 need not equal t'_1. Equation 1-18 is convenient to use to calculate $x_2 - x_1$ at some time t because it relates x, x', and t, whereas Equation 1-19 is not convenient because it relates x, x', and t':

$$x'_2 = \gamma(x_2 - vt_2) \quad \text{and} \quad x'_1 = \gamma(x_1 - vt_1)$$

Since $t_2 = t_1$, we obtain

$$x'_2 - x'_1 = \gamma(x_2 - x_1)$$

$$x_2 - x_1 = \frac{1}{\gamma}(x'_2 - x'_1) = \sqrt{1 - \frac{v^2}{c^2}}(x'_2 - x'_1)$$

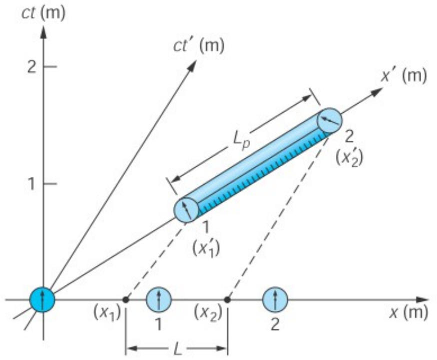

FIGURE 1-28 A measuring rod, a meter stick in this case, lies at rest in S' between $x'_2 = 2$ m and $x'_1 = 1$ m. System S' moves with $\beta = 0.79$ relative to S. Since the rod is in motion, S must measure the locations of the ends of the rod x_2 and x_1 simultaneously in order to have made a valid length measurement. L is obviously shorter than L_p. By direct measurement from the diagram (use a millimeter scale) $L/L_p = 0.62 = 1/\gamma$.

or

$$L = \frac{1}{\gamma}L_p = \sqrt{1 - \frac{v^2}{c^2}}L_p \qquad \textbf{1-28}$$

Thus, the length of a rod is smaller when it is measured in a frame with respect to which it is moving. Before Einstein's paper was published, Lorentz and FitzGerald had independently shown that the null result of the Michelson-Morley experiment could be explained by assuming that the lengths in the direction of the interferometer's motion contracted by the amount given in Equation 1-28. For that reason, the length contraction is often called the *Lorentz-FitzGerald contraction.*

EXAMPLE 1-9 Speed of S' A stick that has a proper length of 1 m moves in a direction parallel to its length with speed v relative to you. The length of the stick as measured by you is 0.914 meter. What is the speed v?

SOLUTION

1. The length of the stick measured in a frame relative to which it is moving with speed v is related to its proper length by Equation 1-28:

$$L = \frac{L_p}{\gamma}$$

2. Rearranging to solve for γ:

$$\gamma = \frac{L_p}{L}$$

3. Substituting the values of L_p and L:

$$\gamma = \frac{1\text{ m}}{0.914\text{ m}} = \frac{1}{\sqrt{1 - v^2/c^2}}$$

4. Solving for v:

$$\sqrt{1 - v^2/c^2} = 0.914$$

$$1 - v^2/c^2 = (0.914)^2 = 0.835$$

$$v^2/c^2 = 1 - 0.835 = 0.165$$

$$v^2 = 0.165c^2$$

$$v = 0.406c$$

It is important to remember that the relativistic contraction of moving lengths occurs only parallel to the relative motion of the reference frames. In particular, observers in relatively moving systems measure the same values for lengths in the y and y' and in the z and z' directions perpendicular to their relative motion. The result is that observers measure different shapes and angles for two- and three-dimensional objects (see Example 1-10 and Figures 1-29 and 1-30).

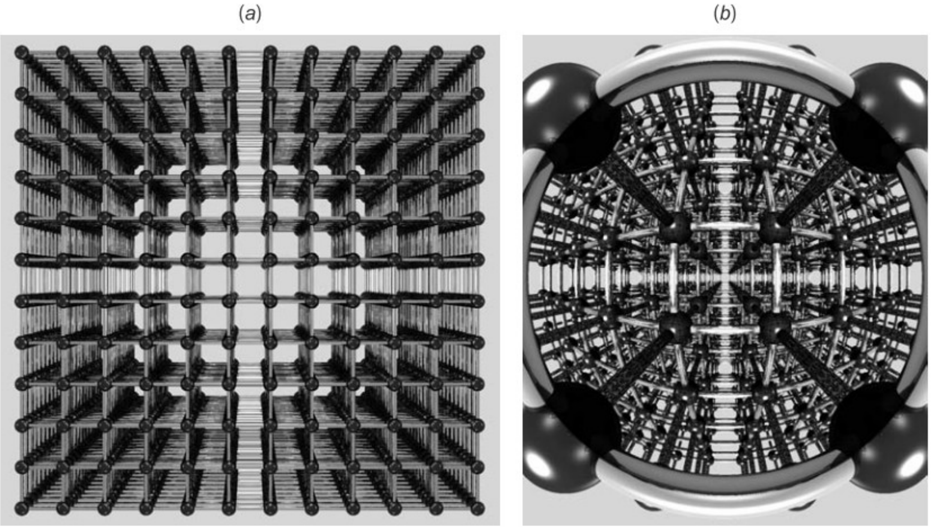

FIGURE 1-29 The appearance of rapidly moving objects depends on both length contraction in the direction of motion and the time when the observed light left the object. (*a*) The array of clocks and measuring rods that represents S' as viewed by an observer in S with $\beta = 0$. (*b*) When S' approaches the S observer with $\beta = 0.9$, the distortion of the lattice becomes apparent. This is what an observer on a cosmic-ray proton might see as it passes into the lattice of a face-centered-cubic crystal such as NaCl. [*P.-K. Hsiung, R. Dunn, and C. Cox. Courtesy of C. Cox, Adobe Systems, Inc., San Jose, CA.*]

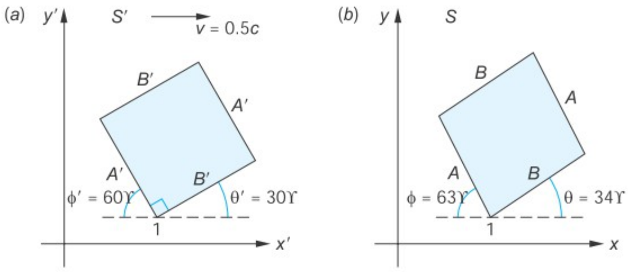

FIGURE 1-30 Length contraction distorts the shape and orientation of two- and three-dimensional objects. The observer in S measures the square shown in S' as a rotated parallelogram.

EXAMPLE 1-10 **The Shape of a Moving Square** Consider the square in the $x'y'$ plane of S' with one side making a 30° angle with the x' axis as in Figure 1-30*a*. If S' moves with $\beta = 0.5$ relative to S, what is the shape and orientation of the figure in S?

SOLUTION

The S observer measures the x components of each side to be shorter by a factor $1/\gamma$ than those measured in S'. Thus, S measures

$$A = [\cos^2 30 + \sin^2 30/\gamma^2]^{1/2} A' = 0.968A'$$

$$B = [\sin^2 30 + \cos^2 30/\gamma^2]^{1/2} B' = 0.901B'$$

Since the figure is a square in S', $A' = B'$. In addition, the angles between B and the x axis and between A and the x axis are given by, respectively,

$$\theta = \tan^{-1}\left[\frac{B'\sin 30}{B'\cos 30/\gamma}\right] = \tan^{-1}\left[\gamma\frac{\sin 30}{\cos 30}\right] = 33.7°$$

$$\varphi = \tan^{-1}\left[\frac{A'\cos 30}{A'\sin 30/\gamma}\right] = \tan^{-1}\left[\gamma\frac{\cos 30}{\sin 30}\right] = 63.4°$$

Thus, S concludes from geometry that the interior angle at vertex 1 is not 90°, but $180° - (63.4° + 33.7°) = 82.9°$—that is, the figure is not a square, but a parallelogram whose shorter sides make 33.7° angles with the x axis! Its shape and orientation in S are shown in Figure 1-30*b*.

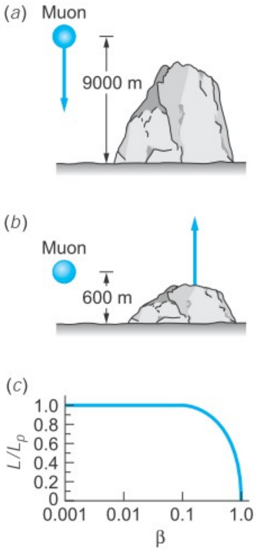

FIGURE 1-31 Although muons are created high above Earth and their mean lifetime is only about 2 μs when at rest, many appear at Earth's surface. (*a*) In Earth's reference frame, a typical muon moving at $0.998c$ has a mean lifetime of 30 μs and travels 9000 m in this time. (*b*) In the reference frame of the muon, the distance traveled by Earth is only 600 m in the muon's lifetime of 2 μs. (*c*) L varies only slightly from L_p until v is of the order of $0.1c$. $L \to 0$ as $v \to c$.

Muon Decay

An interesting example of both time dilation and length contraction is afforded by the appearance of muons as secondary radiation from cosmic rays. Muons decay according to the statistical law of radioactivity:

$$N(t) = N_0 e^{(-t/\tau)} \qquad \textbf{1-29}$$

where N_0 is the original number of muons at time $t = 0$, $N(t)$ is the number remaining at time t, and τ is the mean lifetime (a proper time interval), which is about 2 μs for muons. Since muons are created (from the decay of pions) high in the atmosphere, usually several thousand meters above sea level, few muons should reach sea level. A typical muon moving with speed $0.998c$ would travel only about 600 m in 2 μs. However, the lifetime of the muon measured in Earth's reference frame is increased according to time dilation (Equation 1-26) by the factor $1/(1 - v^2/c^2)^{1/2}$, which is 15 for this particular speed. The mean lifetime measured in Earth's reference frame is therefore 30 μs, and a muon with speed $0.998c$ travels about 9000 m in this time. From the muon's point of view, it lives only 2 μs, but the atmosphere is rushing past it with a speed of $0.998c$. The distance of 9000 m in Earth's frame is thus contracted to only 600 m in the muon's frame as indicated in Figure 1-31.

It is easy to distinguish experimentally between the classical and relativistic predictions of the observations of muons at sea level. Suppose that we observe 10^8 muons at an altitude of 9000 m in some time interval with a muon detector. How many would we expect to observe at sea level in the same time interval? According to the nonrelativistic prediction, the time it takes for these muons to travel 9000 m is (9000 m)/$0.998c \approx 30$ μs, which is 15 lifetimes. Substituting $N_0 = 10^8$ and $t = 15\tau$ into Equation 1-29, we obtain

$$N = 10^8 e^{-15} = 30.6$$

We would thus expect all but about 31 of the original 100 million muons to decay before reaching sea level.

According to the relativistic prediction, Earth must travel only the contracted distance of 600 m in the rest frame of the muon. This takes only $2\ \mu s = 1\tau$. Therefore, the number of muons expected at sea level is

$$N = 10^8 e^{-1} = 3.68 \times 10^7$$

Thus, relativity predicts that we would observe 36.8 million muons in the same time interval. Experiments of this type have confirmed the relativistic predictions.

The Spacetime Interval

We have seen earlier in this section that time intervals and lengths (= space intervals), quantities that were absolutes, or invariants, for relatively moving observers using the classical Galilean coordinate transformation, are not invariants in special relativity. The Lorentz transformation and the relativity of simultaneity lead observers in inertial frames to conclude that lengths moving relative to them are contracted and time intervals are stretched, both by the factor γ. The question naturally arises: Is there *any* quantity involving the space and time coordinates that is invariant under a Lorentz transformation? The answer to that question is yes, and as it happens, we have already dealt with a special case of that invariant quantity when we first obtained the correct form of the Lorentz transformation. It is called the *spacetime interval*, or usually just the *interval*, Δs, and is given by

Experiments with muons moving near the speed of light are performed at many accelerator laboratories throughout the world despite their short mean life. Time dilation results in much longer mean lives relative to the laboratory, providing plenty of time to do experiments.

$$(\Delta s)^2 = (c\Delta t)^2 - [\Delta x^2 + \Delta y^2 + \Delta z^2] \qquad \textbf{1-30}$$

or, specializing it to the one-space-dimensional systems that we have been discussing,

$$(\Delta s)^2 = (c\Delta t)^2 - (\Delta x)^2 \qquad \textbf{1-31}$$

It may help to think of Equations 1-30 and 1-31 like this:

$$[\text{interval}]^2 = [\text{separation in time}]^2 - [\text{separation in space}]^2$$

The interval Δs is the only measurable quantity describing pairs of events in spacetime for which observers in all inertial frames will obtain the same numerical value. The negative sign in Equations 1-30 and 1-31 implies that $(\Delta s)^2$ may be positive, negative, or zero depending on the relative sizes of the time and space separations. With the sign of $(\Delta s)^2$, nature is telling us about the causal relation between the two events. Notice that whichever of the three possibilities characterizes a pair for one observer, it does so for all observers since Δs is invariant. The interval is called *timelike* if the time separation is the larger and *spacelike* if the space separation predominates. If the two terms are equal, so that $\Delta s = 0$, then it is called *lightlike*.

Timelike Interval Consider a material particle[15] or object, such as, the elephant in Figure 1-27, that moves relative to *S*. Since no material particle has ever been measured traveling faster than light, particles always travel less than 1 m of distance in 1 m of light travel time. We saw that to be the case in Example 1-8, where the time interval between launch and birth of the baby elephant was 31.7 months on the *S* clock, during which time the mother elephant had moved a distance of $23.8c \cdot$ months. Equation 1-31 then yields $(c\Delta t)^2 - (\Delta x)^2 = (31.7c)^2 - (23.8c)^2 = (21.0c)^2 = (\Delta s)^2$ and the interval in *S* is $\Delta s = 21.0c \cdot$ months. The time interval term being the larger, Δs is a timelike interval and we say that material particles have *timelike worldlines*. Such

('$

The Lorentz Transformation

Pag. 21(39) Relativistic Velocity Transformation

Pag. 23(41) Spacetime Diagrams

Pag. 27(45) Figure 1-22

Pag. 29(47) 1.4

Time Dilation and Length Contraction

Pag. 97(115) 2.5 General Relativity

→ Pag. 103(121) Deflection of Light in a Gravitational Field.

Pag. 104(122) Figure 2-20

Deflection of a beam of starlight due to the gravitational attraction of the SUN.

$\alpha = 1.75$ arc second 18 Einstein Eddington

Pag. 156(174) 4-2

Rydberg-Ritz formula

$$\frac{1}{\lambda_{mn}} = R\left(\frac{1}{m^2} - \frac{1}{n^2}\right) \text{ for } n > m$$

in 1971 by Shapiro and coworkers using radar signals reflected from several planets. Two of these experimental tests of relativity's predictions, bending of light and gravitational redshift, are discussed in the Exploring sections that follow. The perihelion of Mercury's orbit and the delay of light are discussed in More sections on the web page. Many other predictions of general relativity are subjects of active current research. Two of these, black holes and gravity waves, are discussed briefly in the concluding paragraphs of this chapter.

EXPLORING

Deflection of Light in a Gravitational Field

With the advent of special relativity, several features of the Newtonian law of gravitation, $F_G = GMm/r^2$, became conceptually troublesome. One of these was the implication from the relativistic concept of mass-energy equivalence that even particles with zero rest mass should exhibit properties such as weight and inertia, thought of classically as masslike; classical theory does not include such particles. According to the equivalence principle, however, light, too, would experience the gravitational force. Indeed, the deflection of a light beam passing through the gravitational field near a large mass was one of the first consequences of the equivalence principle to be tested experimentally.

This relativistic effect results in gravitational lenses in the cosmos that focus light from extremely distant galaxies, greatly improving their visibility in telescopes, both on Earth and in orbit.

To see why a deflection of light would be expected, consider Figure 2-19, which shows a beam of light entering an accelerating compartment. Successive positions of the compartment are shown at equal time intervals. Because the compartment is accelerating, the distance it moves in each time interval increases with time. The path of the beam of light, as observed from inside the compartment, is therefore a parabola. But according to the equivalence principle, there is no way to distinguish between an accelerating compartment and one with uniform velocity in a uniform gravitational field. We conclude, therefore, that a beam of light will accelerate in a gravitational field as do objects with rest mass. For example, near the surface of Earth light will fall with acceleration 9.8 m/s^2. This is difficult to observe because of the enormous speed of light. For example, in a distance of 3000 km, which takes about 0.01 second to cover, a beam of light should fall about 0.5 mm. Einstein pointed out that the deflection of a light beam in a gravitational field might be observed when light from a distant star passes close to

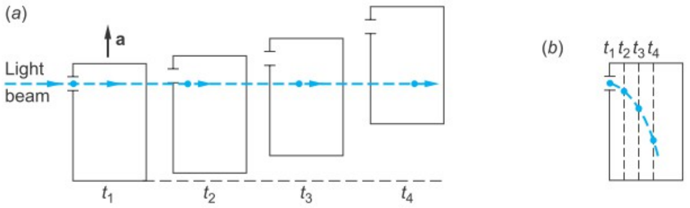

FIGURE 2-19 (*a*) Light beam moving in a straight line through a compartment that is undergoing uniform acceleration. The position of the light beam is shown at equally spaced times t_1, t_2, t_3, t_4. (*b*) In the reference frame of the compartment, the light travels in a parabolic path, as would a ball were it projected horizontally. Note that in both (*a*) and (*b*) the vertical displacements are greatly exaggerated for emphasis.

the Sun.[16] The deflection, or bending, is computed as follows. Rewriting the spacetime interval (Equation 2-42) in polar coordinates,

$$ds^2 = c^2\,dt^2 - (dr^2 + r^2\,d\theta^2 + r^2\sin^2\theta\,d\varphi^2) \tag{2-47}$$

Since the deflection of the light beam occurs in a plane, the two-dimensional version of Equation 2-47 is

$$ds^2 = c^2\,dt^2 - (dr^2 + r^2\,d\theta^2) \tag{2-48}$$

Einstein showed that Equation 2-48 is modified in the presence of a (spherical, nonrotating) mass M to become

$$ds^2 = \gamma(r)^2 c^2\,dt^2 - dr^2/\gamma(r)^2 - r^2\,d\theta^2 \tag{2-49}$$

where $\gamma(r) = (1 - 2GM/c^2r)^{1/2}$, with G = universal gravitational constant and r = distance from the center of mass M. The factor $\gamma(r)$ is roughly analogous to the γ of special relativity. In the following Exploring section on gravitational redshift, we will describe how $\gamma(r)$ arises. For now, $\gamma(r)$ can be thought of as correcting for *gravitational time dilation* (the first term on the right of Equation 2-49) and *gravitational length contraction* (the second term).

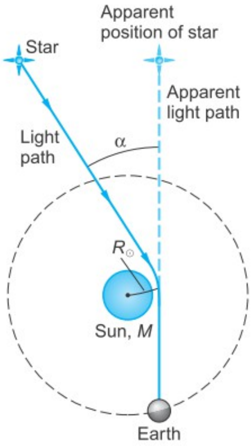

FIGURE 2-20 Deflection (greatly exaggerated) of a beam of starlight due to the gravitational attraction of the Sun.

This situation is illustrated in Figure 2-20, which shows the light from a distant star just grazing the edge of the Sun. The gravitational deflection of light (with mass $\gamma m = E/c^2$) can be treated as a refraction of the light. The speed of light is reduced to $\gamma(r)c$ in the vicinity of the mass M since $\gamma(r) < 1$ (see Equation 2-49), thus bending the wave fronts, and hence the beam, toward M. This is analogous to the deflection of starlight toward Earth's surface as a result of the changing density—hence index of refraction—of the atmosphere. By integrating Equation 2-49 over the entire trajectory of the light beam (recall that $ds = 0$ for light) as it passes by M, the total deflection α is found to be[17]

$$\alpha = 4GM/c^2R \tag{2-50}$$

where R = distance of closest approach of the beam to the center of M. For a beam just grazing the Sun, $R = R_\odot$ = solar radius = 6.96×10^8 m. Substituting the values for G and the solar mass ($M = 1.99 \times 10^{30}$ kg) yields $\alpha = 1.75$ arc second.[18]

Ordinarily, of course, the brightness of the Sun prevents astronomers (or anyone else) from seeing stars close to the limbs (edges) of the Sun, except during a total eclipse. Einstein completed the calculation of α in 1915, and in 1919 expeditions were organized by Eddington[19] at two points along the line of totality of a solar eclipse, both of which were successful in making measurements of α for several stars and testing the predicted $1/R_\odot$ dependence of α. The measured values of α for grazing beams at the two sites were

At Sobral (South America): $\alpha = 1.98 \pm 0.12$ arc seconds
At Principe Island (Africa): $\alpha = 1.61 \pm 0.30$ arc seconds

their average agreeing with the general relativistic prediction to within about 2 percent. Figure 2-21 illustrates the agreement of the $1/R_\odot$ dependence with Equation 2-50. (Einstein learned of the successful measurements via a telegram from H. A. Lorentz.) Since 1919, many measurements of α have been made during eclipses. Since the development of radio telescopes, which are not blinded by sunlight and hence don't require a total eclipse, many more measurements have been made. The latest data agree with the deflection predicted by general relativity to within about 0.1 percent.

The gravitational deflection of light is being put to use by modern astronomers via the phenomenon of *gravitational lensing* to help in the study of galaxies and other large masses in space. Light from very distant galaxies passing near or through other galaxies or clusters of galaxies between the source and Earth can be bent so as to reach Earth in much the same way that light from an object on a bench in the laboratory can be refracted by a glass lens and thus reach the eye of an observer. An intervening galaxy or cluster of galaxies can thus produce images of the distant source, even ones magnified and distorted, just as the glass lens can. Figure 2-22*a* will serve as a reminder of a refracting lens in the laboratory, while Figure 2-22*b* illustrates the corresponding action of a gravitational lens. The accompanying photograph shows the images of several distant galaxies drawn out into arcs by the lens effect of the cluster of galaxies in the center. The first confirmed discovery of images formed by a gravitational lens, the double image of the quasar QSO 0957, was made in 1979 by D. Walsh and his coworkers. Since then astronomers have found many such images. Their discovery and interpretation is currently an active area of research (see Chapter 13).

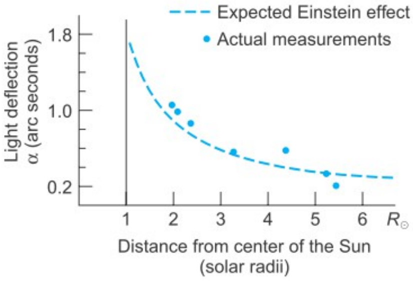

FIGURE 2-21 The deflection angle α depends on the distance of closest approach $R_\odot$ according to Equation 2-50. Shown here is a sample of the data for 7 of the 13 stars measured by the Eddington expeditions. The agreement with the relativistic prediction is apparent.

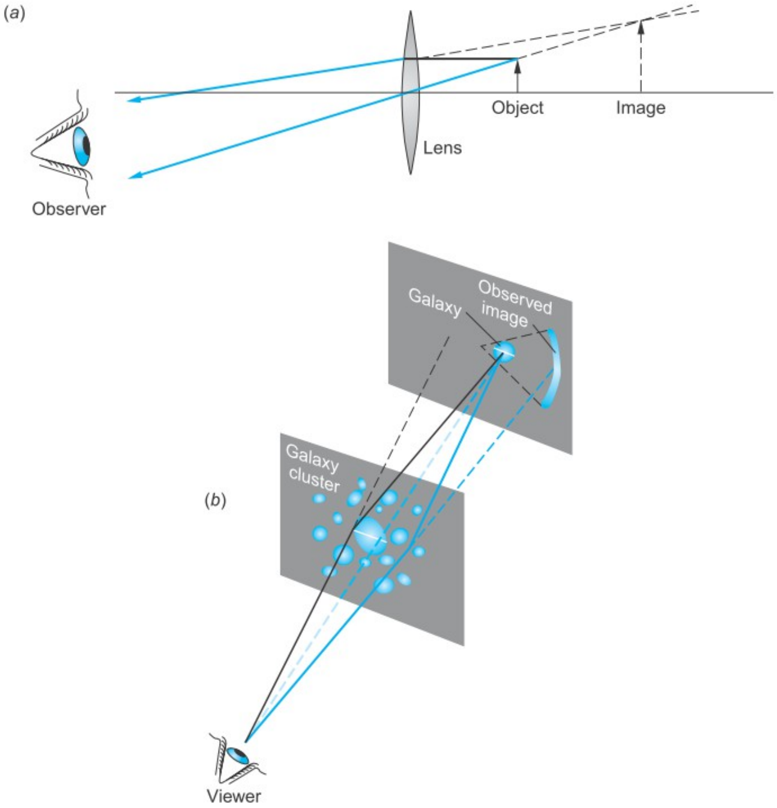

FIGURE 2-22 (*a*) Ordinary refracting lens bends light, causing many rays that would not otherwise have reached the observer's eye to do so. Their apparent origin is the image formed by the lens. Notice that the image is not the same size as the object (magnification) and, although not shown here, the shape of the lens can cause the image shape to be different from that of the object. (*b*) Gravitational lens has the same effects on the light from distant galaxies seen at Earth.

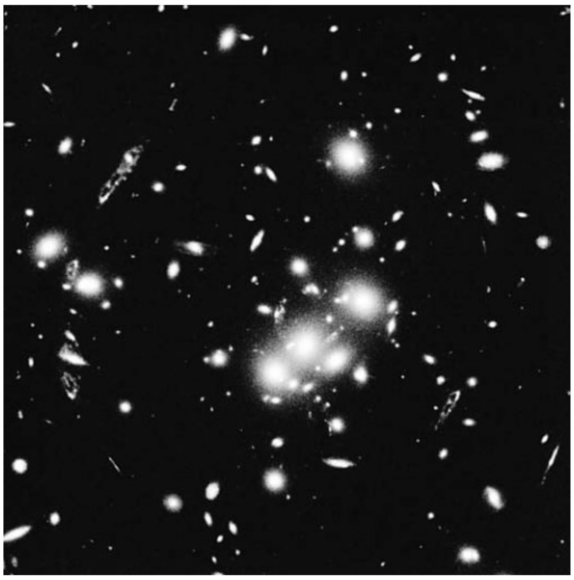

Images of distant galaxies are drawn out into arcs by the massive cluster of galaxies Abell 2218, whose enormous gravitational field acts as a lens to magnify, brighten, and distort the images. Abell 2218 is about 2 billion $c \cdot$ y from Earth. The arcs in this January 2000 Hubble Space Telescope photograph are images of galaxies 10 to 20 billion $c \cdot$ y away. [*NASA, A. Fruchter; ERO Team.*]

EXPLORING
Gravitational Redshift

A second prediction of general relativity concerns the rates of clocks and the frequencies of light in a gravitational field. As a specific case that illustrates the gravitational redshift as a direct consequence of the equivalence principle, suppose we consider two identical light sources (A and A') and detectors (B and B') located in identical spaceships (S and S') as illustrated in Figure 2-23. The spaceship S' in Figure 2-23*b* is located far from any mass. At time $t = 0$, S' begins to accelerate, and simultaneously an atom in the source A' emits a light pulse of its characteristic frequency f_0. During the time $t(= h/c)$ for the light to travel from A' to B', B' acquires a speed $v = at = gh/c$, and the detector B', receding from the original location of A', measures the frequency of the incoming light to be f *redshifted* by a fractional amount $(f_0 - f)/f_0 \approx \beta$ for $v = c$ (see Section 1-5). Thus,

$$(f_0 - f)/f_0 = \Delta f/f \approx \beta = v/c = gh/c^2 \qquad \textbf{2-51}$$

Notice that the right side of Equation 2-51 is equal to the gravitational potential (i.e., the gravitational potential energy per unit mass) $\Delta\phi = gh$ between A and B, divided by c^2. According to the equivalence principle, the detector at B in S must also measure the frequency of the arriving light to be f, even though S is at rest on the planet and, therefore, the shift cannot be due to the Doppler effect! Since the vibrating atom that produced the light pulse at A can be considered to be a clock, and since no "cycles" of the vibration

('$

The Lorentz Transformation

Pag. 21(39) Relativistic Velocity Transformation

Pag. 23(41) Spacetime Diagrams

Pag. 27(45) Figure 1-22

Pag. 29(47) 1.4

Time Dilation and Length Contraction

Pag. 97(115) 2.5 General Relativity

→ Pag. 103(121) Deflection of Light in a Gravitational Field.

Pag. 104(122) Figure 2-20

Deflection of a beam of starlight due to the gravitational attraction of the SUN.

$\alpha = 1.75$ arc second 18 Einstein Eddington

Pag. 156(174) 4-2

Rydberg-Ritz formula

$$\frac{1}{\lambda_{mn}} = R\left(\frac{1}{m^2} - \frac{1}{n^2}\right) \text{ for } n > m$$

('B

data

- 20210926/ Numerical Geometry.../
3030767973.pdf.
Pág. 13(29) - Manifolds of Triangulations.
Pág. 37(53) 1. The GAUSS integral for the Linking Numbers of Curves. /
Pág. 39(55) 2 - the outline of the Invariant-Based Formula and Consequences
theorem 1 (Properties of the Linking Number).
Fig. 1 → interessante.

—

Stephen Gray → descobriu a condutividade da eletricidade.
(1666-1736)

—

Madras → hoje Chennai (Índia)
Suástica no Tibete = "yungdrung".
Ernst Schafer 1939.
Livro: Vaibhav Purandare (author).

D -256- - -* ! - '$ % % B / % 1"
(% / ! " 9 ! !
(% $% 9 72H ! ? . 8 D A

-256- - -* ! - '$ 4 ! => ! 2 "
211 ! @!A !1 ! ! " :! l 5 1"

" -256- - -* ! - '$ 5 A 3 1"
(B 2 ! ! "

Area and Circumference

Any proof that relies on Lemma 6.2.2 is said to use the **method of exhaustion**. Eudoxus used this technique to prove many theorems. These include that the volumes of two spheres are to each other as the cubes of their radii, the volume of a pyramid equals one third the volume of a prism of the same base and the same height, and the volume of a cone is one third the volume of a cylinder with the same base and of the same height (Kline 1972, 50). He also used it to prove the next theorem.

■ **THEOREM 6.2.3 [*Elements* XII.2]**

The areas of two circles are to one another as the squares of their diameters.

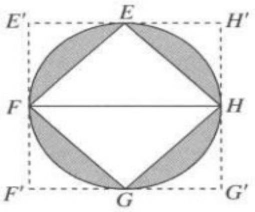

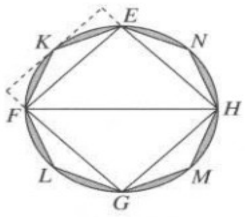

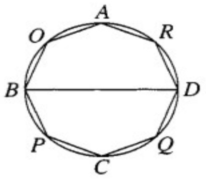

PROOF

Take two circles and call them $\odot ABCD$ and $\odot EFGH$, where $\overline{BD}$ is a diameter of the first circle and $\overline{FH}$ is a diameter of the second. We need to show

$$\frac{\text{area}(\odot ABCD)}{\text{area}(\odot EFGH)} = \frac{BD^2}{FH^2}.$$

Suppose not, and let s be an area that is either less than or greater than the area of $\odot EFGH$, so that

$$\frac{\text{area}(\odot ABCD)}{s} = \frac{BD^2}{FH^2}. \qquad (6.8)$$

We will show that the first case leads to a contradiction and leave the second to Exercise 6. So assume that

$$s < \text{area}(\odot EFGH).$$

We will apply the procedure of Lemma 6.2.2 to the magnitudes

$$\text{area}(\odot EFGH)$$

and

$$\text{area}(\odot EFGH) - s.$$

To do this, first construct $\square EFGH$ inscribed within $\odot EFGH$ and $\square E'F'G'H'$ circumscribed about it. Then appealing to Exercise 5, we have

$$2\text{area}(\square EFGH) = \text{area}(\square E'F'G'H') > \text{area}(\odot EFGH), \qquad (6.9)$$

so

$$\text{area}(\square EFGH) > \frac{1}{2}\text{area}(\odot EFGH).$$

Take $\square EFGH$ and remove it from $\odot EFGH$, resulting in the shaded area in the first figure. We next construct isosceles $\triangle FKE$ with K on the circle. We observe that

$$\text{area}(\triangle FKE) = \frac{1}{2}\text{area}(\sqsubset FKE) > \frac{1}{2} \text{ the area enclosed by } \widehat{FKE}.$$

This is because if a tangent is drawn at K and a rectangle is constructed on the tangent with base $\overline{FE}$, the area of $\triangle EKF$ is half that of the rectangle. Repeat this with the isosceles $\triangle ENH$, $\triangle HMG$, and $\triangle GLF$. Letting a represent the sum of the areas of these four triangles, we conclude that

$$a > \frac{1}{2}\text{area(shading on the right)},$$

and we remove a from the shading in the first figure resulting in the shading of the second. We continue this process of constructing isosceles triangles so that at each step an area is removed that is more than half of the area that remains from the previous step. Lemma 6.2.2 tells us that eventually the remaining area will be less than the area of $\odot EFGH$ minus s. Suppose that the process stops as with the second diagram on page 186 so that

$$\text{area(shading on the right)} < \text{area}(\odot EFGH) - s.$$

Since the area of polygon $EKFLGMHNE$ is equal to the area of $\odot EFGH$ less the shading, we conclude that

$$\text{area}(EKFLGMHN) > s. \qquad (6.10)$$

We now need to include $\odot ABCD$ so that we can set up the ratio. Construct the polygon $AOBPCQDR$ so that it is similar to the polygon $EKFLGMHN$ and inscribed within $\odot ABCD$. This yields

$$\frac{BD^2}{FH^2} = \frac{\text{area}(AOBPCQDR)}{\text{area}(EKFLGMHN)} \qquad (6.11)$$

by Exercise 5, and then by Equation 6.8 we have

$$\frac{\text{area}(\odot ABCD)}{s} = \frac{\text{area}(AOBPCQDR)}{\text{area}(EKFLGMHN)}.$$

Since the area of $\odot ABCD$ is greater than the area of $AOBPCQDR$, s must be greater than the area of $EKFLGMHN$, a contradiction of Inequality 6.10. Therefore, s cannot be less than the area of $\odot EFGH$. This combined with Exercise 6 proves that $s = \text{area}(\odot EFGH)$. ■

The theorem can be used to derive the formula for the area of a circle. Let A be the area of a randomly selected circle and r its radius. The theorem shows that the ratio $A/4r^2$ is the same number no matter which circle is chosen. Call this number $\pi/4$ and write

$$\frac{A}{4r^2} = \frac{\pi}{4},$$

from which we quickly derive

$$A = \pi r^2. \tag{6.12}$$

Now to find the circumference. Archimedes provided the calculation in his work *Measurement of a Circle* (Archimedes 1897, 91–98).

■ **THEOREM 6.2.4** [***Measurement of a Circle* 1**]

The area of a circle is equal to the area of a right triangle in which the length of one leg equals the radius of the circle and the other leg equals the circumference of the circle.

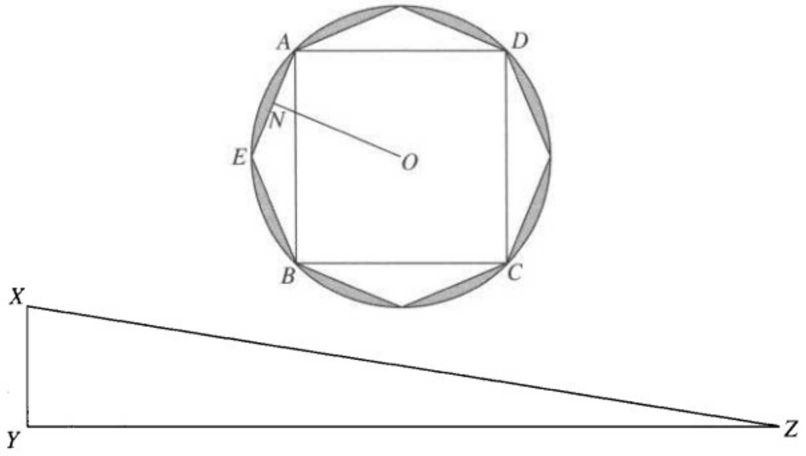

PROOF

The proof uses the method of exhaustion. Let $\odot ABCD$ be the given circle and let $\triangle XYZ$ be such that XY equals the radius of the circle and YZ equals its circumference. To obtain a contradiction, first suppose that the area of the circle is greater than the area of the triangle. Inscribe $\square ABCD$ within $\odot ABCD$. Bisect the arcs cut by $\overline{AB}$, $\overline{BC}$, $\overline{CD}$, and $\overline{DA}$. Join the resulting midpoints to form a regular octagon. Continue this process until we find a regular polygon inscribed within the circle such that its area is a and

$$\text{area}(\odot ABCD) - a < \text{area}(\odot ABCD) - \text{area}(\triangle XYZ)$$

by Lemma 6.2.2. Thus, we have

$$a > \text{area}(\triangle XYZ).$$

Suppose that the polygon with area a is the polygon in the diagram. Let $\overline{AE}$ be a side of the polygon. Find N on $\overline{AE}$ such that $\overline{ON}$ is perpendicular to $\overline{AE}$. Then ON is less than the radius of $\odot ABCD$ and thus less than XY. Furthermore, the perimeter of the polygon is less than the circumference, so it is also less than YZ. Therefore, since the area of a regular polygon equals half the product of its apothem and its perimeter,

$$a < \text{area}(\triangle XYZ),$$

a contradiction. The case when the area of the circle is less than the area of the triangle is similarly brought to a contradiction by using circumscribed polygons (Exercise 7). ■

In modern notation the theorem states that if A is the area of a circle, C its circumference, and r is its radius,

$$A = \frac{Cr}{2}.$$

By Equation 6.12 we conclude that $\pi r^2 = Cr/2$. Hence,

$$C = 2\pi r,$$

and we see further that π represents the ratio of a circle's circumference to its diameter.

Estimating Pi

Now that we have formulas for the area and circumference of a circle, it is natural to try to determine the value of π. Archimedes succeeded in finding an estimate in his *Measurement of a Circle*. His used the strategy found in the method of exhaustion.

■ **THEOREM 6.2.5** [*Measurement of a Circle* 3]

The circumference of a circle is greater than $3\frac{10}{71}$ times its diameter but less than $3\frac{1}{7}$ times its diameter.

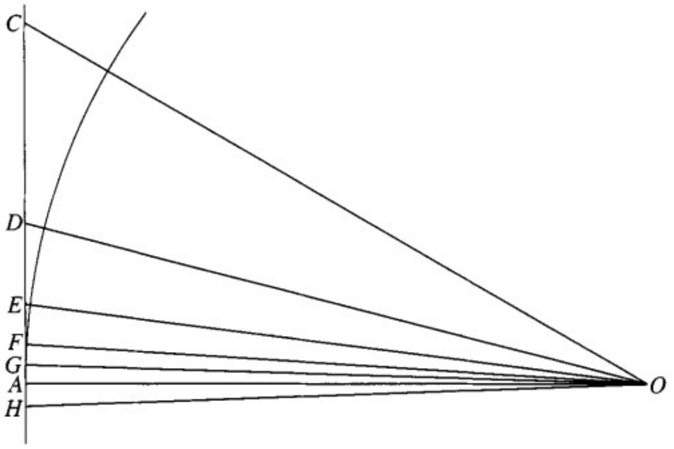

PROOF

We are given $\odot O$ with radius $\overline{OA}$. The goal is to both inscribe and circumscribe a regular 96-gon about the circle and to use those polygons to estimate π. Find C so that $\overline{AC}$ is tangent to $\odot O$ and $\sphericalangle COA$ is a third of a right angle. We may view $\overline{AC}$ as half of one side of a regular dodecagon circumscribed about the circle. Now construct a sequence of angle bisectors:

- Bisect $\sphericalangle COA$ with $\overline{OD}$.
- Bisect $\sphericalangle DOA$ with $\overline{OE}$.
- Bisect $\sphericalangle EOA$ with $\overline{OF}$.
- Bisect $\sphericalangle FOA$ with $\overline{OG}$.

Assume that D, E, F, and G are on $\overline{AC}$. Therefore, $m\sphericalangle GOA$ is $1/48$ that of a right angle. Extend $\overline{GA}$ to H so that $\overline{GA} \cong \overline{AH}$ and $\sphericalangle GOA \cong \sphericalangle HOA$. We see that $m\sphericalangle GOH$ is $1/24$ that of a right angle, so $\overline{GH}$ can be considered one side of a regular 96-gon circumscribed about $\odot O$. Its perimeter is $96GH$. We therefore conclude that

$$\pi = \frac{\text{circumference of } \odot O}{\text{diameter of } \odot O} < \frac{96GH}{2OA} = \frac{96GA}{OA}. \tag{6.13}$$

We need to find an estimate for GA/OA. In order to accomplish this, begin by noticing that the angles constructed above yield a sequence of ratios,

$$\frac{OA}{CA}, \frac{OA}{DA}, \frac{OA}{EA}, \frac{OA}{FA}, \frac{OA}{GA},$$

where each numerator is the radius of the circle and each denominator is half the length of a side of a regular polygon that has been circumscribed about the circle. Note that since $\triangle COA$ is a 30-60-90 triangle,

$$\frac{OC}{CA} = 2 \text{ and } \frac{OA}{CA} = \sqrt{3}.$$

Hence,

$$\frac{OC}{CA} = \frac{306}{153} \tag{6.14}$$

and

$$\frac{OA}{CA} > \frac{265}{153}. \tag{6.15}$$

The fraction $265/153$ is chosen to approximate $\sqrt{3}$, and $306/153$ is chosen to have the same denominator as the first fraction, yet equal 2.

For the second ratio of our sequence, because $\overline{OD}$ bisects $\sphericalangle COA$, we may appeal to Exercise 19 of Section 5.4 and conclude that

$$\frac{OC}{OA} = \frac{CD}{DA}.$$

Therefore,

$$\frac{OC + OA}{OA} = \frac{CD + DA}{DA} = \frac{CA}{DA},$$

and we have *permutando*,

$$\frac{OA}{DA} = \frac{OC + OA}{CA} = \frac{OC}{CA} + \frac{OA}{CA} > \frac{306}{153} + \frac{265}{153}.$$

The last inequality follows from Equation 6.14 and Inequality 6.15. Therefore, we have

$$\frac{OA}{DA} > \frac{571}{153}$$

as a lower bound for the second term of the sequence. By the Pythagorean Theorem we know that $OD^2 = OA^2 + DA^2$. Hence,

$$\frac{OD^2}{DA^2} = \frac{OA^2 + DA^2}{DA^2} > \frac{571^2 + 153^2}{153^2} = \frac{349,450}{23,409},$$

so we have

$$\frac{OD}{DA} > \frac{591\frac{1}{8}}{153}.$$

This allows us to find a lower bound for the third term of the sequence because we may repeat the argument above to find that

$$\frac{OA}{EA} = \frac{OD}{DA} + \frac{OA}{DA},$$

and then

$$\frac{OA}{EA} > \frac{1162\frac{1}{8}}{153} \text{ and } \frac{OE}{EA} > \frac{1172\frac{1}{8}}{153}. \tag{6.16}$$

Similarly,

$$\frac{OA}{FA} > \frac{2334\frac{1}{4}}{153} \text{ and } \frac{OF}{FA} > \frac{2339\frac{1}{4}}{153}, \tag{6.17}$$

and finally,

$$\frac{OA}{GA} > \frac{4673\frac{1}{2}}{153}. \tag{6.18}$$

From Inequalities 6.13 and 6.18, we conclude that

$$\pi < \frac{96GA}{OA} < \frac{96 \cdot 153}{4673\frac{1}{2}} = 3 + \frac{667\frac{1}{2}}{4673\frac{1}{2}} < 3 + \frac{667\frac{1}{2}}{4672\frac{1}{2}} = 3\frac{1}{7} \tag{6.19}$$

A similar argument using inscribed polygons will yield the lower limit. ■

Without decimal notation, calculators, or computers, Archimedes found upper and lower bounds on π equivalent to

$$3.140845070 < \pi < 3.142857143.$$

Furthermore, the upper bound of π is the famous Archimedean approximation:

$$\pi \approx \frac{22}{7}.$$

Exercises

1. Theodorus showed that $\sqrt{n}$ is irrational when the radicand n is not a perfect square and $2 \leqslant n \leqslant 17$. Use the diagram below to make a conjecture on why Theodorus stopped at 17.

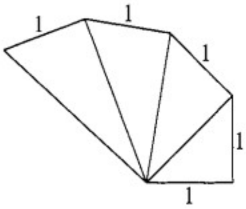

2. Use Antiphon's method of finding the area of a circle to estimate π accurate to five decimal places. What is the minimum number of sides of the inscribed polygon that is necessary to accomplish this?

3. Show that the proof of Lemma 6.2.2 still works if $\overline{DE}$ is divided into an even number of segments.

4. Identify the unstated assumptions in Lemma 6.2.2.

5. Prove Equations 6.9 and 6.11 of the proof of Theorem 6.2.3.

6. Prove that the assumption $s >$ area($\odot EFGH$) also leads to a contradiction in the proof of Theorem 6.2.3.

7. In Archimedes' proof of Theorem 6.2.4, prove that the case when the area of the circle is less than the area of the triangle yields a contradiction using circumscribed polygons.

8. Burton (1985, 126) observes that Archimedes used but never proved that

$$\frac{1351}{780} > \sqrt{3} > \frac{265}{153}, \tag{6.20}$$

and he conjectured that Archimedes would have demonstrated the left-hand inequality by relying on the following two inequalities. Show both.

(a) $26 - \frac{1}{52} = \sqrt{26^2 - 1 + \left(\frac{1}{52}\right)^2} > \sqrt{26^2 - 1}.$

(b) $\frac{1351}{780} = \frac{1}{15}\left(26 - \frac{1}{52}\right) > \frac{\sqrt{26^2 - 1}}{15} = \sqrt{3}.$

9. Prove the right-hand side of Inequality 6.20 by replacing $1/52$ with $1/51$ in the inequalities of Exercise 8.

10. Given a circle with unit radius:
    (a) Construct an octagon inscribed within the circle.
    (b) Construct an octagon circumscribed about the circle.
    (c) Use the two octagons to approximate the area of the circle and then use this result to approximate the value of π.
    (d) What size n-gon is required so that the difference between the estimate for π and its true value is less than 0.001?

11. Use calculus to prove that $A = \pi r^2$ and $C = 2\pi r$.

12. Check that Inequalities 6.16, 6.17, 6.18, and 6.19 on page 191 hold.

13. Show that the ratio of the area of a circle to the square of its diameter is approximately $11/14$ without a calculator.

14. The volume of a cone is one third the volume of a cylinder with the same base and of the same height.
    (a) Write an algebraic equation for this geometric expression.
    (b) Use calculus to prove it.

15. The volumes of two spheres are to one another as the cubes of their radii.
    (a) Write an algebraic equation for this geometric expression.
    (b) Use calculus to prove it.

6.3 THE METHOD

Take a two-by-four and place it across a sawhorse. What we have is a crude **lever** with the sawhorse as the **fulcrum**. Put a brick on one end of the two-by-four and three bricks at the other end. We know that for this to balance, we must move the lumber so that the distance from the fulcrum to the lone brick is longer than that to the three bricks. More precisely, if m_1 is the mass of the single brick and d_1 is its distance to the fulcrum, and if m_2 is the mass of the three bricks with d_2 the distance from the bricks to the fulcrum, then for the lever to balance we need

$$\frac{m_1}{m_2} = \frac{d_2}{d_1}. \tag{6.21}$$

Now make the situation abstract. Represent the two-by-four as a line segment and the sawhorse as a point on the segment. Treat the masses as if they are concentrated in a point. We say that this **system** is in **equilibrium** if the lever balances, and it will balance if and only if Equation 6.21 holds. This proposition is known as the **Law of the Lever**, and Archimedes gave a nearly complete proof of it in his *On the Equilibrium of Planes* (Archimedes 1897, 189–220).

The Law of the Lever was known before the time of Archimedes. The significance of *On the Equilibrium of Planes* was not the novelty of the concept but how it was presented. Archimedes was the first to derive the Law of the Lever from a set of postulates. Each of these assumptions was chosen by Archimedes because he considered each to be clearly true based on his own experience. In so doing, Archimedes became the first Greek to treat what happened on the earth mathematically. Plato viewed the universe as an imperfect representation of the realm of forms, and mathematics was the way to escape this world and learn about the true reality. Aristotle believed that one learns truth by observation, but although his *Physics* dealt with natural phenomena, it was based on nonmathematical intuition. Archimedes followed neither one. Instead, he believed that natural phenomena can be studied using the method of the *Elements*. From agreed upon and simple first principles that reflect nature, laws of the universe can be deduced mathematically because mathematics perfectly represents the world in which we live (Boyer and Merzbach 1991, 121–122).

Center of Gravity

The point at which we find the fulcrum of a system of weights that is in equilibrium is called the **center of gravity** of the system. We assume that the total weight of the system is concentrated at this point. We now generalize this concept of balance to other situations. We first deal with a single object. Take either a plane or solid figure. If it is only two-dimensional, view it as having a nearly infinitesimal thickness and uniform density. Such an object is called a **lamina**. The ancient Greeks knew of three possible interpretations for what it meant to balance such an object (Figure 6.7). The first two have the object being suspended in some way. The **horizontal suspension** point is the location where an object must be attached to a wire so that it is perfectly balanced in the horizontal direction. Now take the object and hang it vertically by

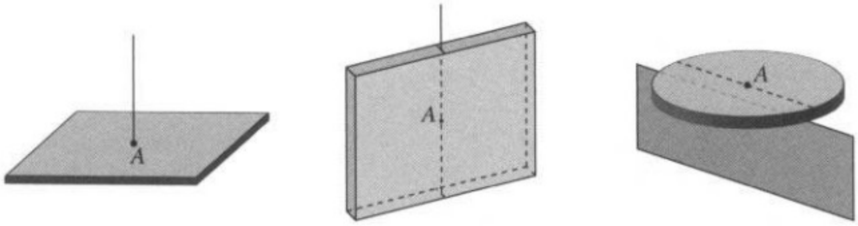

(a) Horizontal suspension. (b) Vertical suspension. (c) Balance.

Figure 6.7 Centers of gravity.

the wire. From the point connected to the wire draw a vertical line. Remove the object, rotate it a few degrees, and then hang again vertically so that it balances. When the second vertical line is drawn it will intersect the first in the **vertical suspension** point. Finally, imagine an extremely thin metal sheet standing upright. The intersection of all lines on which the object will balance on that sheet we call the **balance** point. Although experiment would confirm our guess, we assume that these three interpretations result in the same point that we define to be the center of gravity of the object (Stein 1999, 15–17). We also make the following two assumptions:

■ **POSTULATES 6.3.1 [*On the Equilibrium of Planes* I, Postulates]**

1. The centers of gravity of similar figures are in the same relative positions.
2. The center of gravity of a convex polygon must lie within the figure.

We now follow Archimedes to generalize the center of gravity to any number of weights. He began by proving the following special case. See Exercise 2.

■ **THEOREM 6.3.2 [*On the Equilibrium of Planes* I.4, 5]**

- If the centers of gravity of two equal weights are not at the same point, the weights balance at the midpoint of the segment joining their centers of gravity.
- If three equal weights have their centers of gravity on a straight line at equal distances, the center of gravity of the system will coincide with the center of gravity of the middle weight.

Theorem 6.3.2 can be generalized to any number of weights. We deal with the even number case here and leave the odd case to Exercise 4. Since the two-weight result is the first part of Theorem 6.3.2, to prove the result for an even number of weights, we start with four.

■ **THEOREM 6.3.3 [*On the Equilibrium of Planes* I.5 Corollary]**

If an even number of equal weights have their centers of gravity on a straight line such that the distances between any pair of consecutive centers of gravity are the same, the midpoint of the segment joining the middle two centers of gravity is the center of gravity of the entire system.

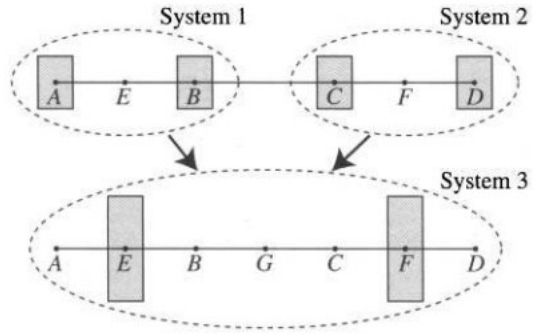

PROOF

Let weights of equal magnitude be placed on a line segment so that their centers of gravity are at points A, B, C, and D. Let E be the midpoint of $\overline{AB}$, and let F be the midpoint of $\overline{CD}$. By Theorem 6.3.2, the total mass of System 1 is concentrated at E, and the total mass of System 2 is at F. Now the lever can be viewed as in System 3, with G being the midpoint of $\overline{EF}$. Since the mass at E equals the mass at F, the system is in equilibrium with center of gravity G. Since $\overline{EB} \cong \overline{CF}$ and $\overline{EG} \cong \overline{GF}$, we have $\overline{BG} = \overline{GC}$. Hence, G is the midpoint of $\overline{BC}$. Generalize this procedure to complete the proof. ■

As an example of how Archimedes worked with centers of gravity, we shall find the center of gravity of a parallelogram.

■ **THEOREM 6.3.4 [*On the Equilibrium of Planes* I.9]**

The center of gravity of a parallelogram lies on the segment joining the midpoints of opposite sides.

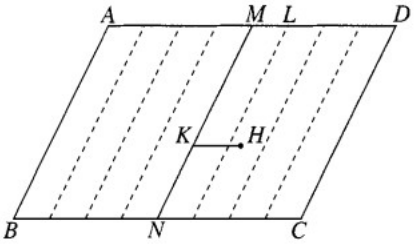

PROOF

Let $ABCD$ be a parallelogram. Assume that M is the midpoint of $\overline{AD}$ and N is the midpoint of $\overline{BC}$. Join $\overline{MN}$. Designate the center of gravity of the parallelogram by H. We must prove that H is on $\overline{MN}$, so suppose it lies elsewhere inside the parallelogram (Postulate 6.3.1.2). Find K on $\overline{MN}$ so that $\overline{KH}$ is parallel to $\overline{AD}$. Find the midpoint of $\overline{MD}$. This divides the segment into two segments. Find the midpoint of one of these resulting segments and divide again. By continually finding midpoints in this fashion, we can locate L on $\overline{MD}$ such that $ML < KH$ (Lemma 6.2.2). Using this length, divide both $\overline{AM}$ and $\overline{MD}$ into consecutive congruent line segments of length ML. At the endpoints of each of these congruent segments, draw lines parallel to $\overline{MN}$. The resulting parallelograms are congruent. Therefore, their centers of gravity are in the same relative position (Postulate 6.3.1.1). This means that we have an even number of magnitudes of equal weight with collinear centers of gravity. Thus, by Theorem 6.3.3 the center of gravity of the system of parallelograms is on the midpoint of the line segment joining the centers of gravity of the middle two parallelograms, but this is impossible because H does not lie within either of these parallelograms. We are forced to conclude that H lies on $\overline{MN}$. ■

The last proposition quickly yields the next, so its proof is left to Exercise 5.

■ **THEOREM 6.3.5**

The center of gravity of a parallelogram lies at the intersection of its diagonals.

The Letter to Eratosthenes

Nearly all of the extant works of the ancient Greek mathematicians came down to us in their finished, polished form. By our own experience, work in mathematics does not flow directly from theorem to finished proof in one sitting. It may take hours, days, weeks, and in some cases years to write a proof. There are times of progress and times of stagnation. Often, a promising argument leads to a dead end, while an idea out of nowhere may be exactly what is needed. It would be fascinating and highly instructive if we could find a book of notes written by someone like Eudoxus that would illuminate the process by which he came by his results. In the case of Archimedes, *The Method* is that example. As with many of the works of Archimedes, *The Method* is actually a letter. It was written to his good friend Eratosthenes. More important, Archimedes knew that Eratosthenes was the perfect person to which to send this.

> Seeing moreover in you, as I say, an earnest student, a man of considerable eminence in philosophy, and an admirer [of mathematical inquiry], I thought fit to write out for you and explain in detail in the same book the peculiarity of a certain method, by which it will be possible for you to get a start to enable you to investigate some of the problems in mathematics by means of mechanics. (Archimedes 1897, *The Method*, Introduction, 13).

Eratosthenes was born in Cyrene around 284 B.C. Beside mathematics he studied philosophy and poetry. He combined his interests and wrote a summary of the mathematics needed for the study of Platonic philosophy. It included the tale of the Delians with their problem of doubling the volume of their altar (Section 4.3). His *On Means* was included in the *Treasury of Analysis* of Pappus. He studied geography and wrote *On the Measurement of the Earth*. Because of his expertise, he was invited to Egypt by Ptolemy Evergetes around the year 244 B.C. to tutor his son. He later became the librarian at Alexandria. Although he was expert in many fields, he was never viewed as the leading scholar in any of them. Mathematics is a case in point, where he was second to Archimedes. Because he always seemed to come in second, he gained the nickname *Beta* (Heath 1921b, 104–107).

Eratosthenes is famous for two main contributions that are still taught today. The first is in the field of number theory. The ancient Greeks did study numbers, including the prime numbers (page 47). Eratosthenes' contribution is an algorithm designed to identify the primes in a list of given numbers. It is known as the **sieve of Eratosthenes** (Figure 6.8). The sieve works like this. Make a list of consecutive numbers starting with 2 and ending with n. Circle 2 but cross out all other multiples of two. Next, circle 3 (the next number that is neither circled nor crossed-out) and cross out all other multiples of three. Repeat this process until the next available number is greater than $\sqrt{n}$. At his point, as seen in Figure 6.8, all circled and nonmarked numbers are prime.

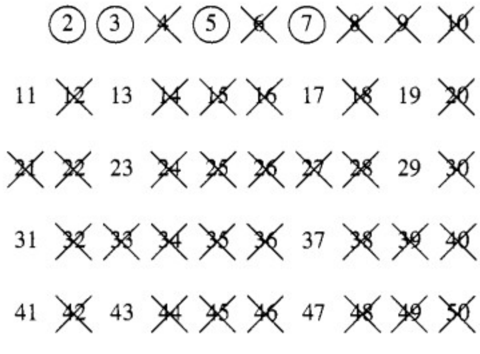

Figure 6.8 The sieve of Eratosthenes.

The second of Eratosthenes' contributions is usually taught in trigonometry or high school geometry courses. It was known that on the summer solstice the sun would be directly above Syene (now Aswan) in Egypt. This was confirmed by noting that the sun could be seen at the bottom of a well. Also on the summer solstice, at Alexandria the sun formed an angle equivalent to one fiftieth of a circle. This is the angle α in Figure 6.9. If we suppose that the rays of the sun form parallel lines, we may conclude that the measure of the arc between the two cities is also α. Estimate the distance between Syene and Alexandria to be 5000 stadia. This is based on how long it took a caravan with camels to travel between those two cities (Kline 1972, 160–161). Since the cities were approximately on the same longitudinal line and α represents one fiftieth of a circle, the circumference of the globe can be calculated to be approximately 250,000 stadia. Assuming that a stadia is equivalent to 517 feet, Eratosthenes' measurement equals 24,479 miles. The earth's true circumference is approximately 24,900 miles (24,902 miles at the equator and 24,860 miles pole to pole), so his measurement is off by 2%, the best among the ancient Greeks.

We see then that Eratosthenes had the intellectual résumé to appreciate and use the method that Archimedes was about to show him. Archimedes continues:

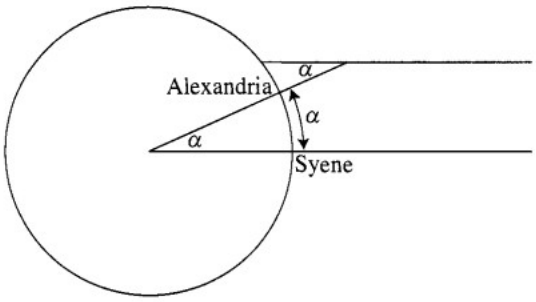

Figure 6.9 Eratosthenes and the circumference of the earth.

> This procedure, is I am persuaded, no less useful even for the proof of the theorems themselves; for certain things first became clear to me by a mechanical method, although they had to be demonstrated by geometry afterwards because their investigation by the said method did not furnish an actual demonstration. But it is of course easier, when we have previously acquired, by the method, some knowledge of the questions, to supply the proof than it is to find it without any previous knowledge. This is a reason why, in the case of the theorems the proof of which Eudoxus was the first to discover, namely that the cone is a third part of the cylinder, and the pyramid of the prism, having the same base and equal height, we should give no small share of the credit to Democritus, who was the first to make the assertion with regard to the said figure though he did not prove it. (Archimedes 1897, *The Method*, Introduction, 13)

Archimedes told Eratosthenes that the first theorem that he discovered using the method is the formula on page 177 for the area of a parabola. By repeating the method for different questions he convinced himself of the truth of many other results that he outlined for Eratosthenes. We will examine one of these by following Archimedes' line of thought as he uses his method to convince himself of the truth of his formula regarding the volume of a sphere, namely that

> the volume of a sphere is four times the volume of a cone with base area equal to that of a great circle of the sphere and height equal to its radius.

We start with the preliminary constructions. Let $\odot ABCD$ be a great circle of a sphere with perpendicular diameters $\overline{AC}$ and $\overline{BD}$ (Figure 6.10). Construct a circle with diameter $\overline{BD}$ and perpendicular to $\overline{AC}$. With this circle as its base, draw a cone with vertex A. Extend this cone to a new base that is a circle perpendicular to $\overline{AC}$ and with diameter $\overline{ECF}$. On this circle, construct a cylinder with height equal to AC. Finally, extend $\overline{AC}$ beyond A to H such that $\overline{HA} \cong \overline{AC}$. We shall view $\overline{HC}$ as a lever with fulcrum A.

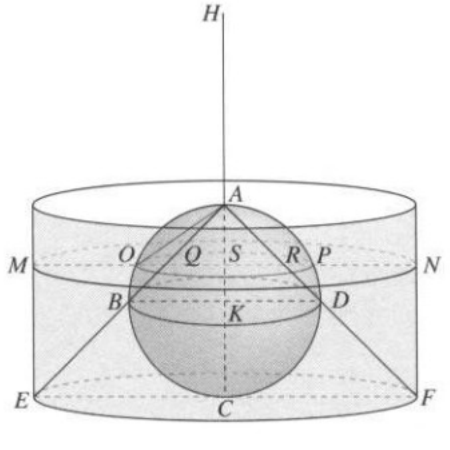

Figure 6.10 The sphere and its cone in a cylinder.

Next take a plane parallel to $\odot BD$ and intersect the cylinder. This forms a circle with diameter $\overline{MN}$. Call the points where $\overline{MN}$ intersects the sphere O and P. Further, $\overline{MN}$ intersects $\overline{AC}$ at S, $\overline{AE}$ at Q, and $\overline{AF}$ at R. Join $\overline{AO}$. This plane also intersects the sphere, forming a circle with diameter $\overline{OP}$, and it intersects the cone, forming a circle with diameter $\overline{QR}$.

Now that the setup is complete,

$$\overline{MS} \cong \overline{AC} \text{ and } \overline{QS} \cong \overline{AS}. \tag{6.22}$$

Therefore,

$$MS \cdot QS = AC \cdot AS = AO^2 = OS^2 + QS^2.$$

The last equality follows by the Pythagorean Theorem, and the second-to-last equality holds because $\triangle OAC \sim \triangle SAO$. Then, since $\overline{HA} \cong \overline{AC}$,

$$\frac{HA}{AS} = \frac{AC}{AS} = \frac{MS}{QS} = \frac{MS \cdot MS}{MS \cdot QS} = \frac{MS^2}{OS^2 + QS^2} = \frac{MN^2}{OP^2 + QR^2}.$$

Therefore,

$$\frac{HA}{AS} = \frac{\text{area(circle with diameter } \overline{MN})}{\text{area(circle with diameter } \overline{OP}) + \text{area(circle with diamter } \overline{QR})},$$

This means that if we take lever $\overline{HS}$ and place $\odot MN$ at S and both $\odot OP$ and $\odot QR$ at H, it balances at fulcrum A (Figure 6.11). The above can be repeated for any plane that cuts the cylinder parallel to its base. Archimedes now argues that since all of the slices of the cylinder at each S balance with the corresponding slices of the sphere and cone when placed at H with fulcrum A, the same lever will balance when the

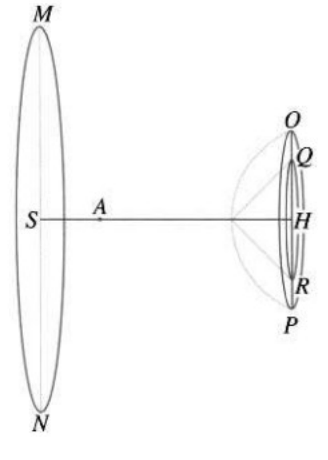

Figure 6.11 Circles formed by the plane in balance.

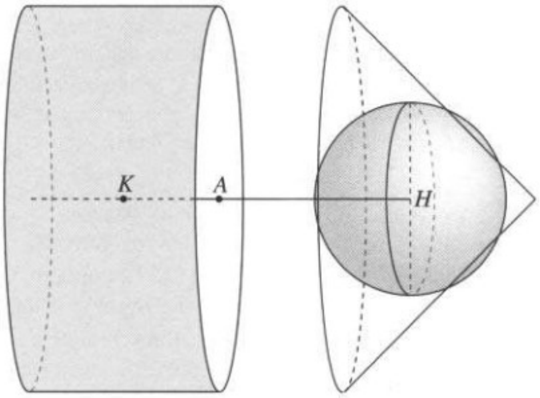

Figure 6.12 The cylinder balancing the sphere and cone.

entire cylinder is placed so that its center of gravity is at K while the sphere and cone are placed with their centers of gravity at H (Figure 6.12). By the Law of the Lever,

$$\frac{HA}{AK} = \frac{\text{volume(cylinder)}}{\text{volume(sphere)} + \text{volume(cone } AEF)}.$$

Since $HA = 2AK$,

$$\text{volume(cylinder)} = 2\text{volume(sphere)} + 2\text{volume(cone } AEF).$$

By Exercise 14 of Section 6.2,

$$\text{volume(cylinder)} = 3\text{volume(cone } AEF).$$

Thus,

$$\text{volume(cone } AEF) = 2\text{volume(sphere)},$$

but since $EF = 2BD$,

$$\text{volume(cone } AEF) = 8\text{volume(cone } ABD).$$

We may now conclude that

$$\text{volume(sphere)} = 4\text{volume(cone } ABD),$$

which is the formula.

At this point Archimedes observed that this method, which brought him to the formula for the volume of a sphere, also allowed him to derive the formula for the surface area of a sphere. He wrote:

> [f]rom this theorem, to the effect that a sphere is four times as great as the cone with a great circle of the sphere as base and with height equal to the radius of

> the sphere, I conceived the notion that the surface of any sphere is four times as great as a great circle in it; for, judging from the fact that any circle is equal to a triangle with base equal to the circumference and height equal to the radius of the circle, I apprehended that, in like manner, any sphere is equal to a cone with base equal to the surface of a sphere and height equal to its radius. (Archimedes 1897, *The Method*, 2, 20–21)

It seems that Archimedes' technique was quite productive.

The Method of Archimedes would have remained lost save by happy chance. In 1899, John L. Heiberg, a Danish scholar in ancient Greek mathematics and specialist on Archimedes, learned of a manuscript held at the library of the Holy Sepulchre at Jerusalem. The document was a **palimpsest**, a parchment manuscript that has had its original writing washed or scraped off and then reused for another purpose. In this case, the work was an Eastern Orthodox prayer book dating to about the thirteenth century. Fortunately, the original writing, dating to the tenth century, was not perfectly removed. Because of this, some of the original was translated and was seen to be a copy of a work of Archimedes. Hearing this, Heiberg had to see it, and he did in 1906 and 1908 at Constantinople. Heiberg translated what he could of the old text and discovered that the palimpsest contained copies of many Archimedean works. It includes *On the Sphere and Cylinder*, most of *On Spirals*, and parts of both *Measurement of a Circle* and *On the Equilibrium of Planes*, not to mention the only known Greek copy of *On Floating Bodies*. More important, it contains nearly all of *The Method*. This work of Archimedes had been cited by various writers, including some quotations, but the exact content had been unknown. The good fortune of the Heiberg discovery enables us to see some of the ideas behind Archimedes' great work. For further details see Netz and Noel (2007), Dijksterhuis (1987), and Archimedes (1897, The *Method* of Archimedes).

Exercises

1. Masses of a and b are placed at the ends of lever $\overline{AB}$ with fulcrum F.

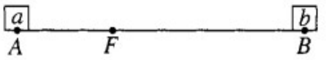

   (a) If $a = 3$, $b = 7$, and $AF = 10$, what is FB?
   (b) If a is three times as massive as b when $AF = 20$, what is FB?
   (c) If the lever balances in its current configuration and then a mass b is added to the left-hand side, how far to the left must F be moved to make the lever balance?

2. Prove Theorem 6.3.2.

3. Rewrite the proof of Theorem 6.3.3 for the case of six equally spaced weights on a lever.

4. Prove that if five weights have their centers of gravity on a straight line such that the distances between any pair of consecutive centers of gravity are the same, the center of gravity of the middle segment is the center of gravity of the entire system.

5. Prove Theorem 6.3.5.

6. Show using any method that the center of gravity of:
   (a) a circle is its center.
   (b) a triangle is the intersection of two segments each joining a vertex with the midpoint of the opposite side.

7. Construct the center of gravity of a rectangle, a circle, and a triangle.

8. Use the sieve of Eratosthenes to find all prime numbers between 51 and 150.

9. The circumference of the earth's equator is 24,902 miles and the circumference of the great circle that has its diameter as the segment with its endpoints at the poles is 24,860 miles. Find the approximate volume of the earth in cubic miles.

10. Put a coordinate system on Figure 6.10 so that the center of the sphere K is the origin, D is $(0,\ a,\ 0)$, and A is $(0,\ 0,\ a)$. Find equations for:
   (a) $\overline{AF}$
   (b) $\odot OP$ where S is $(0,\ 0,\ s)$
   (c) $\odot QR$
   (d) $\odot MN$

11. Continue with Exercise 10.
   (a) Find the areas of $\odot OP$, $\odot QR$, and $\odot MN$
   (b) Show that HA/AS equals

$$\frac{\text{area(circle with diameter } \overline{MN})}{\text{area(circle with diameter } \overline{OP}) + \text{area(circle with diamter } \overline{QR})},$$

   (c) Compute the volume of the sphere $ABCD$, the cone AEF, and the cylinder with height AC and base $\odot EF$ by integration.
   (d) Use the volume just found to show that

$$\frac{HA}{AK} = \frac{\text{volume(cylinder)}}{\text{volume(sphere)} + \text{volume(cone } AEF)}.$$

12. Use Archimedes' method to show the following:
   (a) The volume of a cylinder with base equal to a great circle of a sphere and height equal to its diameter is $3/2$ times the volume of the sphere.
   (b) The volume of a right-angled conoid between its vertex and a plane perpendicular to its axis is $3/2$ times the volume of a cone that has the same base and the same height as the segment of the conoid.

13. Find the centers of gravity of the given solids using any method.
   (a) A right-angled conoid.
   (b) A hemisphere.
   (c) An obtuse-angled conoid.
   (d) A spheroid (oblong or flat).

7.2 THREE PROBLEMS

Viète made other contributions to mathematics besides his analytic art. In the field of geometry he worked issues related to the three famous Greek construction problems. The first is the Delian problem. The other two are known as **squaring the circle** and **trisecting an angle**. To square a circle means to find a square with the area of a given circle, while to trisect an angle means to divide an arbitrary angle into three congruent angles. These three problems cannot be constructed with a compass and straightedge alone, so they require other methods.

Square the Circle

Viète used his analysis to give the first equation for π involving an infinite sequence of algebraic operations. If he could find its value, he could square the circle by computing $\sqrt{\pi}$. Up to his day there were many estimates for π. Simply use Archimedes' method with regular polygons with more and more sides inscribed in and circumscribed about a circle. Viète's approach found this equation:

$$\frac{2}{\pi} = \sqrt{\frac{1}{2}}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}}}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}}}}\cdots.$$

To represent this infinite product, Viète wrote the first four terms (without the square root sign but with an l for *latus*) and then noted, "*in infinitum observata uniformi methodo*" (Cajori 1928, vol. II, 78, 185). Boyer and Merzbach (1991, 321) note that this equation is closely linked to a curve called the **quadratrix**. The curve was discovered by Hippias around 475 B.C. Hippias was born in Elis in southern Greece, and he was a Sophist. Plato included him in two of his dialogues, but his portrayal was not very flattering (Plato 1961, *Hippias Major* and *Hippias Minor*).

The quadratrix was the first locus to be described by a Greek mathematician beyond the simple line and circle. It works like this: Take $\square ABCD$ and let $\overset{\frown}{BD}$ be the first quadrant of $\odot A$ with radius $\overline{AB}$. If $\overline{BC}$ moves uniformly to $\overline{AD}$ in the same amount of time as radius $\overline{AB}$ turns to $\overline{AD}$, the result is curve BHG in Figure 7.2. Letting HF represent the height of $\overline{BC}$ at a given point in time, due to the uniform motion, we see that

$$\frac{HF}{AB} = \frac{\theta}{\pi/2}.$$

Since $HF = AH\sin\theta$,

$$\frac{AH\sin\theta}{\theta} = \frac{2AB}{\pi}.$$

Letting $a = AB$ and $r = AH$, this becomes

$$r = \frac{2a\theta}{\pi\sin\theta} \tag{7.1}$$

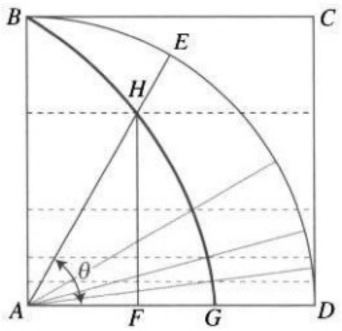

Figure 7.2 The quadratrix of Hippias.

in polar form. Therefore, because

$$\lim_{\theta \to 0^+} \frac{\theta}{\sin \theta} = 1,$$

there exists a point G on $\overline{AD}$ such that H lands on G when $\theta = 0$ and

$$AG = \frac{2a}{\pi}. \tag{7.2}$$

This equation means that the quadratrix has given us a segment that we can now use to construct a segment of length π with only a straightedge and compass. Namely,

$$\pi = \frac{2a}{AG},$$

and from this we can construct a segment of length $\sqrt{\pi}$. This is what we need to square the circle (Heath 1921a, 226–227).

Back to Viète's formula. Define the sequence r_n using Equation 7.1 with $a = 1$ and $\theta = \pi/2^n$,

$$r_n = \left(2^{n-1} \sin \frac{\pi}{2^n}\right)^{-1}$$

This equation amounts to taking a right angle, continually bisecting it, and finding the corresponding value on the quadratrix. Furthermore, observe that $r_1 = 1$ and by a double-angle formula,

$$\sin \frac{\pi}{2^{n-1}} = 2 \sin \frac{\pi}{2^n} \cos \frac{\pi}{2^n}.$$

This gives

$$\left(2 \sin \frac{\pi}{2^n}\right)^{-1} \left(\sin \frac{\pi}{2^{n-1}}\right) = \cos \frac{\pi}{2^n}.$$

Therefore,

$$\frac{r_n}{r_{n-1}} = \left(2^{n-1}\sin\frac{\pi}{2^n}\right)^{-1}\left(2^{n-2}\sin\frac{\pi}{2^{n-1}}\right) = \cos\frac{\pi}{2^n}. \tag{7.3}$$

We now proceed by examining the infinite product:

$$\frac{r_2}{r_1}\cdot\frac{r_3}{r_2}\cdot\frac{r_4}{r_3}\cdots = \lim_{n\to\infty}\left(r_2\cdot\frac{r_3}{r_2}\cdot\frac{r_4}{r_3}\cdots\frac{r_n}{r_{n-1}}\right) = \lim_{n\to\infty} r_n = \frac{2}{\pi}.$$

The last equality holds due to the limit implied by Equation 7.2. Substituting with Equation 7.3 gives

$$\cos\frac{\pi}{4}\cdot\cos\frac{\pi}{8}\cdot\cos\frac{\pi}{16}\cdots = \frac{2}{\pi},$$

and we then arrive at Viète's formula because

$$\begin{aligned}
\cos\frac{\pi}{4} &= \sqrt{\frac{1+\cos\pi/2}{2}} = \sqrt{\frac{1}{2}} \\
\cos\frac{\pi}{8} &= \sqrt{\frac{1+\cos\pi/4}{2}} = \sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}}} \\
\cos\frac{\pi}{16} &= \sqrt{\frac{1+\cos\pi/8}{2}} = \sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}}}} \\
&\vdots
\end{aligned}$$

Duplicate the Cube

Viète solved the problem of duplicating the cube by again finding inspiration from the seventh book of the *Mathematical Collection* of Pappus. There he found problems that dealt with ***neusis***, a term that means "verging" or "inclination." According to Pappus, if a line passes through a point when produced, the line makes a neusis on that point. The general problem is that given two lines, place between them a segment of a given magnitude so that the segment intersects a particular point. The *neusis* often assumed that the segment had to be shifted from some other position so that it would include the desired point. Many Greek geometers used a *neusis* without explicitly stating it as a postulate or proving it as a theorem. For example, Archimedes did this in his work *On Spirals* (Pappus 1986, *Mathematical Collection*, Introduction, 112; Boyer 1956, 143; Archimedes 1897, c-ci).

To solve a *neusis* problem, one needs a tool besides a straightedge and compass. The ancient Greek geometer Nicomedes, who lived around 200 B.C., provided such a tool. Start with a horizontal **ruler** labeled $\overline{AB}$ (Figure 7.3). At its midpoint F attach another ruler perpendicular to the first. Call it $\overline{FE}$. On this second ruler fix a small peg at point C called the **pole**, and on C attach a third ruler that itself has a peg at D and at the other end a slot in which the peg at C can run. The peg of the third

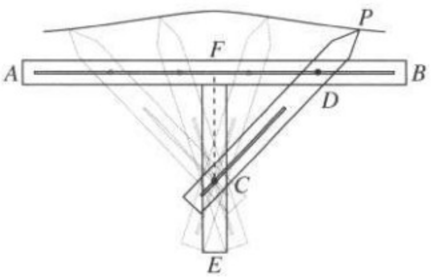

Figure 7.3 The first conchoid of Nicomedes.

ruler slides in a groove in ruler $\overline{AB}$. As the peg D slides along $\overline{AB}$, the endpoint P of the third ruler describes a locus of points that is known as the **first conchoid** of Nicomedes. If we let $DP = a$ (the **distance**) and $FC = b$ (Figure 7.4), then using polar coordinates the curve is given by the equation

$$r = a + b\csc\theta. \tag{7.4}$$

The derivation is left for Exercise 7. Proclus noted that there were other conchoids, but the details of the work of Nicomedes are no longer extant (Heath 1921a, 238–240).

Related to the conchoid is a curve known as a **cissoid**. The credit of the discovery of this curve is due to Diocles, who around 100 B.C. wrote *On Burning-glasses* in which he gave a solution to the duplication of the cube problem using this curve. The details of his solution are left to Exercise 11. To define the curve, take a circle with radius $\overline{AB}$ and construct $\overline{CD}$ to be the perpendicular bisector of the diameter (Figure 7.5). Find G and H on $\overline{AB}$ so that $\overline{GC} \cong \overline{CH}$ and erect perpendiculars at G and H that meet the circumference of the circle at E and F, respectively. Join $\overline{EB}$ and let P be the intersection of $\overline{EB}$ with $\overline{HF}$. The locus of all points P formed in this manner as different points G and H are found on $\overline{AB}$ with $\overline{GC} \cong \overline{CH}$ is the cissoid (Heath 1921a, 264–265).

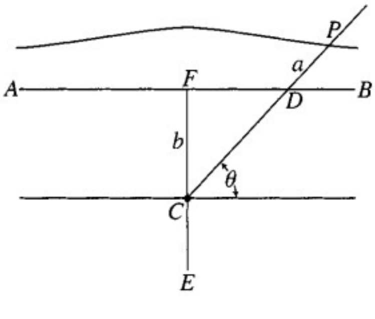

Figure 7.4 Deriving the polar equation for the first conchoid.

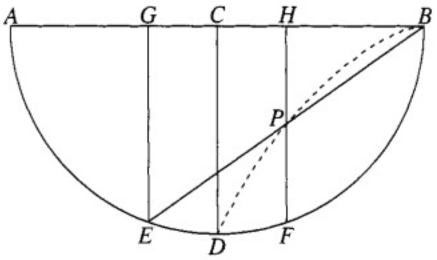

Figure 7.5 The cissoid of Diocles.

When the cissoid of Diocles is generalized, we see the the conchoid is a type of cissoid in this new sense. Let c and c' be two curves and fix a point B that we will call the **pole**. Take any point E on c and find the line l that joins B and E. Let E' be the intersection of l and c' and take P on l such that $BP = EE'$. The locus of all points P is the **generalized cissoid** (Figure 7.6). If c is a circle, B is a point on c, and c' is the tangent line to c that is opposite B, the resulting curve is the cissoid of Diocles. The conchoid of Nicomedes is a generalized cissoid when c is a line, c' a circle, and B the center of the circle. See Exercise 12.

Inspired by the *neusis* problems, Viète took what he learned from Pappus and expanded Euclid's postulate set to include the following three propositions. These and the proofs that follow are from Viète's *A Supplement to Geometry* (Viète 1983, 388–398).

■ POSTULATE 7.2.1

Through a point not on two given lines, there exists a third line that intersects the first two lines, forming a segment of any possible length.

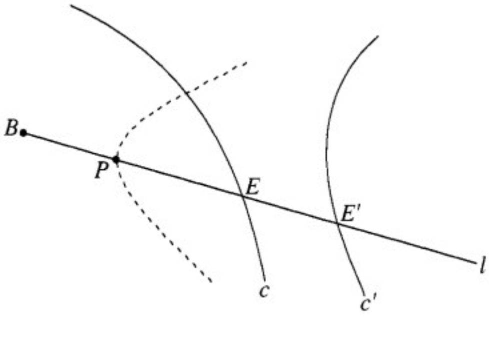

Figure 7.6 The generalized cissoid.

POSTULATE 7.2.2

Given a point on the exterior of an angle such that there exists lines through the point intersecting both sides of the angle, there exists a line through the point that intersects both sides of the angle, forming a segment of any possible length.

POSTULATE 7.2.3

There exists a line from any given point within or on a circle to any line that intersects the circle, forming a segment of any possible length.

The conchoid provided justification for Viète that these postulates were reasonable. For example, we can use it to confirm Postulate 7.2.1. Let the pole C be the given point and the distance a the given length. Suppose that one of the lines is $\overline{AB}$ and the other line l intersects the generated conchoid. If we move the ruler so that P is at the intersection of the conchoid and l, the line formed by the ruler is the desired line that passes through C and intersects the two lines forming a segment of length a.

Nicomedes used his curve to duplicate the cube and trisect an angle (Exercises 9 and 10). Viète used his *neusis* postulates to do the same. First, he duplicated the cube by finding two mean proportionals. We start with a lemma.

LEMMA 7.2.4 [*A Supplement to Geometry*, IV]

Let B be on the exterior of $\odot O$. If $\overline{BD}$ intersects $\odot O$ at C and D and $\overline{BF}$ intersects $\odot O$ at E and F such that

$$BC \cdot BE = CD \cdot EF,$$

$\overline{BC}$ and $\overline{BE}$ will be two continued mean proportionals between $\overline{CD}$ and $\overline{EF}$.

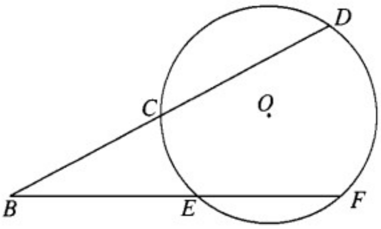

PROOF

Let $\overline{BD}$ intersect $\odot O$ at C and D and $\overline{BF}$ intersect the circle at E and F. By hypothesis we have

$$\frac{CD}{BE} = \frac{BC}{EF}.$$

Since $BC = BD - CD$ and $EF = BF - BE$, we have

$$\frac{CD}{BE} = \frac{BD}{BF}.$$

By the Secant Theorem (5.5.4),

$$\frac{BD}{BF} = \frac{BE}{BC}.$$

Hence,

$$\frac{CD}{BE} = \frac{BE}{BC} = \frac{BC}{EF}. \blacksquare$$

Viète uses the lemma to prove that one can construct (using his additional postulates) two mean proportionals between two given magnitudes, and thus one can duplicate the cube.

■ **CONSTRUCTION 7.2.5 [*A Supplement to Geometry*, V]**

Find the two mean continued proportionals between two given magnitudes.

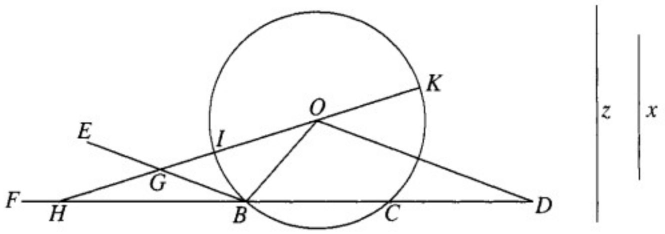

PROOF

We are given line segments of length x and z. Draw $\odot O$ with radius $\overline{OB}$ such that $OB = z/2$. Next draw chord $\overline{BC}$ such that $BC = x$ by Postulate 7.2.3. Extend $\overline{BC}$ to D so that C is the midpoint of $\overline{BD}$. Join $\overline{OD}$. Let $\overline{BE}$ be parallel to $\overline{OD}$ and extend $\overline{BD}$ indefinitely to F beyond B. By Postulate 7.2.2, find H on $\overline{BF}$ such that $\overline{OH}$ meets $\overline{BE}$ at G with $\overline{HG} \cong \overline{OB}$. Finally, let $\overline{OH}$ be extended to meet $\odot O$ at K and let I be the intersection of $\overline{OH}$ with $\odot O$. Notice that we have

$$\frac{HG}{HB} = \frac{OG}{BD} \tag{7.5}$$

since $\overline{OD}$ is parallel to $\overline{BG}$ (Exercise 6), and because $HG = OB$ we have

$$\frac{HG}{IK} = \frac{BC}{BD}. \tag{7.6}$$

Because of Equations 7.5 and 7.6, we see that

$$HB \cdot OG = IK \cdot BC.$$

Because $HG = OI$, we have $OG = HI$. Hence,

$$HB \cdot HI = IK \cdot BC.$$

By the last equation we see that the conditions to Lemma 7.2.4 are satisfied. Therefore, HB and HI are the two mean continued proportionals desired because $IK = z$ and $BC = x$. ■

Trisect an Angle

The third postulate allows Viète to trisect an arbitrary angle. The proof of this theorem is similar to that of Archimedes (1897, *Book of Lemmas*, Proposition 8, 309).

■ **CONSTRUCTION 7.2.6 [*A Supplement to Geometry*, IX]**

Trisect a given angle.

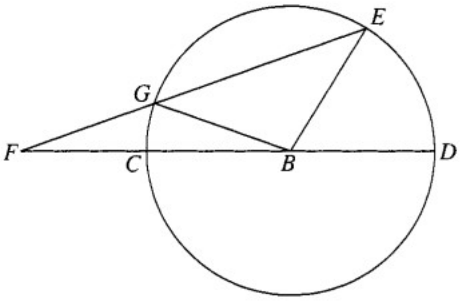

PROOF

Let $\measuredangle A$ be the given angle. If $\measuredangle A$ is a right angle, it is a theorem of Euclid's geometry that $\measuredangle A$ can be trisected. Suppose, then, that $\measuredangle A$ is acute. The obtuse case will be similar to what follows.

Construct $\odot B$ with diameter $\overline{CD}$. Find E on the circumference such that $\measuredangle EBD \cong \measuredangle A$. Find F on an extension of $\overline{CD}$ beyond C so that $\overline{EF}$ intersects the circle at G and $\overline{FG} \cong \overline{BC}$ (Postulate 7.2.3). To trisect $\measuredangle EBD$, first join $\overline{BG}$. This yields the isosceles $\triangle FGB$. The details that follow are left unstated by Viète. They are due to Witmer (Viète 1983, 398). Now $\measuredangle BGE$ is an exterior angle of $\triangle BGF$. Hence,

$$\mathrm{m}\measuredangle BGE = \mathrm{m}\measuredangle BFG + \mathrm{m}\measuredangle GBF = 2\mathrm{m}\measuredangle GBF$$

since the base angles are congruent. However, $\measuredangle BGE \cong \measuredangle BEG$ because $\triangle BEG$ is isosceles. This gives

$$\begin{aligned}\mathrm{m}\measuredangle GBE &= \text{two right angles} - \mathrm{m}\measuredangle BGE - \mathrm{m}\measuredangle BEG\\ &= \text{two right angles} - 4\mathrm{m}\measuredangle GBF.\end{aligned}$$

Furthermore,

$$\mathrm{m}\measuredangle GBD = \text{two right angles} - \mathrm{m}\measuredangle GBF.$$

Thus,

$$\begin{aligned}\mathrm{m}\measuredangle EBD &= \text{two right angles} - \mathrm{m}\measuredangle GBE - \mathrm{m}\measuredangle GBF\\ &= \text{two right angles} - (\text{two right angles} - 4\mathrm{m}\measuredangle GBF) - \mathrm{m}\measuredangle GBF\\ &= 3\mathrm{m}\measuredangle GBF.\end{aligned}$$

Hence, $\measuredangle A$ is trisected. ■

*D 9 ! 2

Exercises

1. Suppose that for all circles, there exists a rectangle with area equal to the circle. From this prove that every circle can be squared.

2. Use the spiral of Archimedes to square the circle and trisect an angle.

3. Find the equation in rectangular coordinates for the quadratrix and graph it.

4. The quadratrix can be used to trisect an angle. Take $\sphericalangle GAH$ and let H lie on the quadratrix. Find M on $\overline{BA}$ and N on $\overline{CD}$ such that $\overline{MN}$ is parallel to $\overline{AD}$. Trisect both $\overline{MA}$ at points M' and M''. Do the same for $\overline{ND}$ at N' and N''. Let H' be the intersection of the quadratrix with $\overline{M'N'}$ and H'' be the intersection between the quadratrix and $\overline{M''N''}$.

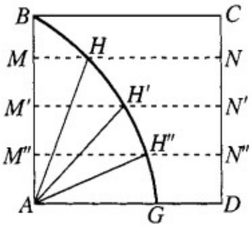

   (a) Prove that $\sphericalangle GAH'' \cong \sphericalangle H''AH' \cong \sphericalangle H'AH$.
   (b) Perform the construction to trisect a randomly selected acute angle.

5. Use the quadratrix to divide an angle that measures half a right angle into five congruent angles.

6. Confirm Equation 7.5 in the proof for Construction 7.2.5.

7. Derive Equation 7.4. This is illustrated in Figure 7.4.

8. The conchoid of Nicomedes can be described by the equation

$$(x^2 + y^2)(x - b)^2 - a^2x^2 = 0.$$

   (a) Show how this equation is derived from Equation 7.4.
   (b) Graph this equation.
   (c) Find the area of the loop when $b/a = 1/2$.
   (d) Generalize the argument on page 239 to generate the loop.

9. Nicomedes used the conchoid to duplicate the cube (Heath 1921a, 260–262). To do this it suffices to find two continued mean proportional between two segments $\overline{AB}$ and $\overline{BC}$. Draw them at right angles. Find J so that $AJCB$ is a rectangle. Let D be the midpoint of $\overline{AB}$ and E the midpoint of $\overline{BC}$. Join $\overline{DJ}$ and produce it beyond D to meet $\overline{BC}$ produced at G. Construct $\overline{EF}$ such that $\overline{FC} \cong \overline{AD}$. Join $\overline{GF}$. Construct $\overline{CH}$ parallel to $\overline{GF}$.

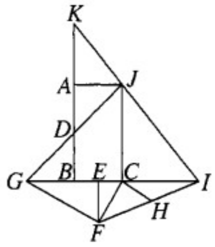

(a) Draw $\overline{FI}$ and assume that H is on it so that I is on $\overline{BC}$ produced and $\overline{HI} \cong \overline{FC}$. An appropriate conchoid will enable us to find I. Give the pole, ruler, and distance of this conchoid.

(b) Join $\overline{JI}$ and extend it to meet $\overline{AB}$ produced at K. Prove that CI and AK are two continued mean proportionals between AB and BC.

10. The conchoid can be used to trisect an angle (Heath 1921a, 236, 238). Given $\measuredangle ABC$, draw $\overline{AC}$ perpendicular to $\overline{BC}$. Find D so that $DACB$ is a rectangle. Let E be on $\overline{DA}$ produced. To find the exact location of E, draw the conchoid with pole B, ruler $\overline{AC}$, and distance $2AB$. Join $\overline{BE}$ and let G be the intersection of $\overline{BE}$ with $\overline{AC}$. Let F be the midpoint of $\overline{DE}$. Prove that $m\measuredangle ABC = 3m\measuredangle GBC$.

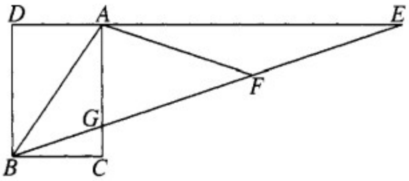

11. The cissoid of Diocles is illustrated in Figure 7.5. The point P is a point on the curve. Let $r = PB$, $\theta = m\measuredangle ABP$, and $AB = 1$.

(a) Show that the equation in polar coordinates for the cissoid of Diocles is $r = \sec\theta - \cos\theta$.

(b) Prove that the equation in part (a) is equivalent to $r = \tan\theta \sec\theta$.

(c) Show that the equation for this cissoid is $x(x^2 + y^2) - y^2 = 0$.

(d) Solve the Delian problem using the cissoid by showing that HF and HB are two continued mean proportionals between AH and HP.

12. Show that both the conchoid of Nicomedes and the cissoid of Diocles are generalized cissoids.

13. Given a right triangle, find another right triangle with the same height such that the product of the difference between their bases and the difference between their hypotenuses is equal to the area of a given rectangle.

14. If there exists an isosceles triangle and a line from an endpoint of the base to the opposite leg of length equal to that leg, prove that the exterior angle formed by the base and the line has measure equal to three times one of the base angles.

data

Canon 85 mm f1.2 FD SSC
Aspherical US$ 23,000.00

otani.camera.tokyo

55 13 97425 1172 Caiyá
mudou endereço

Books:

20210927/Advanced Engineering Mathematics/

→ todo diretório !!!

— Differential Equations... (2020).pdf

20210927/// (20210927).cont

20210927/Riemann-Geometry/
Revolutions of Geometry...pdf
Pag. 186(202) Area and Circumference
Pag. 236(252) 7.2 three Problems
Square the Circle ($\sqrt{\pi}$)
Pag. 238(254) Duplicate the Cube
Viète solved the problems of

duplicating the cube by again finding inspiration from seventh book the Mathematical Collection of Pappus.

Pag. 243(259)

Trisect an Angle

The third postulate allows Viète to trisect an arbitrary angle.

20210928///

www.reschio.com/
luxury-villas-umbria

Fulvio Ricci

fricci@sns.it

Books: (Estes dois livros conjuntamente).

20210928/ Geometric Aspects of Harmonic Analysis/

2021_Book_GeometricAspects...pdf

Pag. 12(24) 5. Trace Hardy and Hardy Inequalities

-256- - -* ! - '$
l) .l7 2 1 * :! 2 ! 1"
(' $ 9 ! :! "3 ! " :! "3 & > !

which has been proved in Cowling–Haagerup [6] (see also [5]). Here $\mathcal{L}_s$ is the conformally invariant fractional power of the sublaplacian which is defined by the relation $\widehat{\mathcal{L}_s f}(\lambda) = \hat{f}(\lambda) M_s(\lambda)$ where $M_s(\lambda)$ is the same family of operators used in the definition of G_s and $\hat{f}$ stands for the operator valued group Fourier transform of f on $\mathbb{H}^n$. From (9) we obtain

$$\pi(\Phi_{-s,\rho}) G_s f(\xi, w) = (2\pi)^{2s} \rho^{2s} \pi(\Phi_{s,\rho}) f(\xi, w) = (2\pi)^{2s} u(\xi, w, \rho). \qquad (10)$$

In [14] the authors have calculated that $-\rho^{1-2s}\partial_\rho \varphi_{-s,\rho} = \rho^{-2n-2}\psi_{s,1}(\rho^{-1}z, \rho^{-2}w)$ for an explicit function $\psi_{s,1}$ and constant $C_2(n,s)$, where $\varphi_{s,\rho}(z,w) = ((\rho^2 + |z|^2)^2 + 16w^2)^{-\frac{n+s+1}{2}}$. Thus differentiating both sides of (10) by ρ and multiplying by $-\rho^{1-2s}$, we obtain

$$C_1(n, -s)\pi(\psi_{s,p}) G_s f(\xi, w) = -(2\pi)^{2s} \rho^{1-2s} \partial_\rho u(\xi, w, \rho).$$

This proves our claim with $C_{n,s} = (2\pi)^{-2s} C_1(n, -s) = 2^{n-3s-1}\pi^{-n-s-2}\Gamma(\frac{n+1-s}{2})^2$. Finally we calculate B_s using the value of $C_2(n,s)$ calculated in [14]:

$$B_s = \left(2^{n-3s-1}\pi^{-n-s-2}\Gamma(\frac{n+1-s}{2})^2\right) \times \left(2^{-n+s}\pi^{n+1}\Gamma(1-s)/\Gamma(\frac{n+1-s}{2})^2\right)$$
$$= 2^{-1-2s}\pi^{-1-s}\Gamma(1-s).$$

The proof is complete. □

5 Trace Hardy and Hardy Inequalities

Consider the vector fields $X_j = \xi_j \frac{\partial}{\partial w}$, $Y_j = \frac{\partial}{\partial \xi_j}$ and $T = \frac{1}{2}\rho\frac{\partial}{\partial w}$ on $\mathbb{R}^{n+1} \times \mathbb{R}^+$. Let X be one of these vector fields. For real valued functions u, v defined on $\mathbb{R}^{n+1} \times \mathbb{R}^+$ consider

$$\int_0^\infty \int_{\mathbb{R}^{n+1}} \left(Xu - \frac{u}{v}Xv\right)^2 \rho^{1-2s}\, d\xi\, dw d\rho.$$

Using integration by parts and assuming that u and v are such that $\frac{u^2}{v}Xv$ vanishes at infinity, we have

$$\int_{\mathbb{R}^{n+1}} \frac{u}{v} Xu\, Xv\, d\xi dw = -\int_{\mathbb{R}^{n+1}} \frac{u}{v} Xu\, Xv\, d\xi dw - \int_{\mathbb{R}^{n+1}} u^2 X(\frac{1}{v}Xv)\, d\xi dw.$$

Simplifying, we get

$$\int_{\mathbb{R}^{n+1}} \frac{u^2}{v^2}(Xv)^2\, d\xi - 2\int_{\mathbb{R}^{n+1}} \frac{u}{v} Xu\, Xv\, d\xi = \int_{\mathbb{R}^{n+1}} \frac{u^2}{v} X^2 v\, d\xi. \tag{11}$$

Consequently,

$$\begin{aligned}&\int_0^\infty \int_{\mathbb{R}^{n+1}} \Big(Xu - \frac{u}{v}Xv\Big)^2 \rho^{1-2s}\, d\xi\, dw d\rho \\ &\quad = \int_0^\infty \int_{\mathbb{R}^{n+1}} (Xu)^2 \rho^{1-2s}\, d\xi\, dw d\rho + \int_0^\infty \int_{\mathbb{R}^{n+1}} \frac{u^2}{v}(X^2 v)\rho^{1-2s}\, d\xi\, dw d\rho.\end{aligned}$$

In a similar way, using integration by parts, we can check that

$$\begin{aligned}&\int_0^\infty \frac{u^2}{v^2}(\partial_\rho v)^2 \rho^{1-2s}\, d\rho - 2\int_0^\infty \frac{u}{v}\partial_\rho u\, \partial_\rho v\, \rho^{1-2s}\, d\rho \\ &\quad = \int_0^\infty \frac{u^2}{v}\partial_\rho(\rho^{1-2s}\partial_\rho v)\, d\rho + \lim_{\rho\to 0}\left(\frac{u^2}{v}\rho^{1-2s}\partial_\rho v\right)\end{aligned}$$

which leads to the equation

$$\begin{aligned}&\int_0^\infty \int_{\mathbb{R}^{n+1}} \Big(\partial_\rho u - \frac{u}{v}\partial_\rho v\Big)^2 \rho^{1-2s}\, d\xi\, d\rho \\ &\quad = \int_0^\infty \int_{\mathbb{R}^{n+1}} (\partial_\rho u)^2 \rho^{1-2s}\, d\xi\, d\rho + \int_0^\infty \int_{\mathbb{R}^{n+1}} \frac{u^2}{v}\partial_\rho(\rho^{1-2s}\partial_\rho v)\rho^{1-2s}\, d\xi\, d\rho \\ &\qquad + \int_{\mathbb{R}^{n+1}} \frac{u^2(\xi,0)}{v(\xi,0)} \lim_{\rho\to 0}(\rho^{1-2s}\partial_\rho v)(\xi,\rho)\, d\xi.\end{aligned}$$

Let us now consider the gradient

$$\nabla u = (X_1 u, \cdots, X_n u, Y_1 u, \cdots, Y_n u, \frac{1}{2}\rho\, \partial_w u, \partial_\rho u). \tag{12}$$

Adding the above equations we obtain the identity

$$\begin{aligned}&\int_0^\infty \int_{\mathbb{R}^{n+1}} |\nabla u - \frac{u}{v}\nabla v|^2 \rho^{1-2s}\, d\xi\, dw d\rho = \int_0^\infty \int_{\mathbb{R}^{n+1}} |\nabla u|^2 \rho^{1-2s}\, d\xi\, dw d\rho \\ &\qquad + \int_{\mathbb{R}^{n+1}} \frac{u^2(\xi,0)}{v(\xi,0)} \lim_{\rho\to 0}(\rho^{1-2s}\partial_\rho v)(\xi,\rho)\, d\xi\, dw d\rho \\ &\qquad\quad + \int_0^\infty \int_{\mathbb{R}^{n+1}} \frac{u^2}{v}(-G + \frac{1}{4}\rho^2\partial_w^2 + \partial_\rho)(\rho^{1-2s}\partial_\rho v) d\xi\, dw d\rho.\end{aligned}$$

If v solves the extension problem

$$\left(-G+\partial_\rho^2+\frac{1-2s}{\rho}\partial_\rho+\frac{1}{4}\rho^2\partial_w^2\right)v(\xi,\rho)=0,$$

then the last term in the above vanishes. As the left hand side is non-negative, this leads to the following inequality known as trace Hardy inequality in the literature.

Proposition 7 *Let u be a real valued compactly supported continuous function on $\mathbb{R}^{n+1}\times[0,\infty)$ which is smooth for $\rho>0$. Let v be a real valued function which solves the extension problem for the Grushin operator. Then*

$$\int_0^\infty\int_{\mathbb{R}^{n+1}}|\nabla u(\xi,w,\rho)|^2\rho^{1-2s}\,d\xi\,dwd\rho \geqslant -\int_{\mathbb{R}^{n+1}}\frac{u^2(\xi,w,0)}{v(\xi,w,0)}\lim_{\rho\to 0}(\rho^{1-2s}\partial_\rho v)(\xi,w,\rho)\,d\xi dw.$$

When v is the solution of the extension problem for the Grushin operator G with initial condition $\varphi\in L^2(\mathbb{R}^{n+1})$ defined by (6) then we have proved that $\lim_{\rho\to 0}\rho^{1-2s}\partial_\rho v=-B_sG_s\varphi$ where B_s is given in Theorem 4.1. Thus the inequality in the above proposition takes the form

$$\int_0^\infty\int_{\mathbb{R}^{n+1}}|\nabla u(\xi,w,\rho)|^2\rho^{1-2s}\,d\xi\,dwd\rho\geqslant c_s\int_{\mathbb{R}^{n+1}}\frac{u^2(\xi,w,0)}{\varphi(\xi,w)}G_s\varphi(\xi,w,\rho)\,d\xi dw$$

where $c_s=2^{1-2s}\frac{\Gamma(1-s)}{\Gamma(s)}$. In the case of the sublaplacian on H-type groups, there are explicit functions $\varphi_{s,\delta}$ such that $\mathcal{L}_s\varphi_{-s,\delta}=\delta^sC_{n,s}\varphi_{s,\delta}$ with explicit constant $C_{n,s}$ which have allowed the authors in [14] to simplify the quotient $\frac{\mathcal{L}_s\varphi_{-s,\delta}}{\varphi_{-s,\delta}}$ to get a sharp trace Hardy inequality. Unfortunately in our context, though we can find analogues of $\varphi_{s,\delta}$ the quotient $\frac{G_s\varphi_{-s,\delta}}{\varphi_{-s,\delta}}$ does not seem to simplify. But things are not so bad if we slightly modify the definition of the fractional power G_s.

Recall that G_s is defined in terms of the multiplier

$$M_s(\lambda)=(2|\lambda|)^s\sum_{k=0}^\infty\frac{\Gamma(\frac{2k+n+1+s}{2})}{\Gamma(\frac{2k+n+1-s}{2})}P_k(\lambda).$$

Instead of this we use the slightly different multiplier

$$\tilde{M}_s(\lambda)=(2|\lambda|)^s\sum_{k=0}^\infty\frac{\Gamma(\frac{2k+n+2+2s}{4})}{\Gamma(\frac{2k+n+2-2s}{4})}P_k(\lambda)$$

in defining the modified fractional power $\tilde{G}_s$. Note that $\tilde{G}_s$ is nothing but $(\frac{1}{2}G)_s$. The operators G_s and $\tilde{G}_s$ are comparable. For $0<s<1$ let $C_1(s)$ and $C_2(s)$ be defined by

$$C_1(s) = \inf_k \frac{\Gamma(\frac{2k+n+1+s}{2})}{\Gamma(\frac{2k+n+1-s}{2})} \frac{\Gamma(\frac{2k+n+2-2s}{4})}{\Gamma(\frac{2k+n+2+2s}{4})}, \; C_2(s)^{-1} = \inf_k \frac{\Gamma(\frac{2k+n+1-s}{2})}{\Gamma(\frac{2k+n+1+s}{2})} \frac{\Gamma(\frac{2k+n+2+2s}{4})}{\Gamma(\frac{2k+n+2-2s}{4})}.$$

In view of Stirling's formula for the gamma function, these constants are positive and finite. Then it follows that

$$C_1(s)\langle \tilde{G}_s f, f\rangle \leq \langle G_s f, f\rangle \leq C_2(s)\langle \tilde{G}_s f, f\rangle. \tag{13}$$

The operator $\tilde{G}_s$ is better behaved as we can see from the following proposition.

For any $s \in \mathbb{R}$ and $\delta > 0$ let $u_{s,\delta}(x, w) = ((\delta + |\xi|^2)^2 + w^2)^{-\frac{n+2+2s}{4}}$ defined on $\mathbb{R}^{n+1}$. We have an explicit expression for the action of $\tilde{G}_s$ on $u_{-s,\delta}$.

Proposition 8 *For any $\delta > 0$ and $0 < s < 1$, we have $\tilde{G}_s u_{-s,\delta}(\xi, w) = C_{n,s}\delta^s u_{s,\delta}(\xi, w)$, where $C_{n,s} = \frac{2^{2s}\Gamma(\frac{n+2+2s}{4})^2}{\Gamma(\frac{n+2-2s}{4})^2}$.*

Proof Since $u_{-s,\delta}$ is radial in ξ the action of G on $u_{-s,\delta}$ is the same as that of the generalised sublaplacian $\mathcal{L}(n/2-1) = -\partial_r^2 - \frac{(n-1)}{r}\partial_r - r^2\partial_w^2$. Therefore, the result follows from Theorem 3.11 in [5]. This can be proved by expanding the function $u^{\lambda}_{-s,\delta}(r)$ in terms of Laguerre functions of type $(n/2 - 1)$. We leave the details to the reader. □

Using the above proposition it is easy to prove a Hardy inequality for the modified fractional power $\tilde{G}_s$. If we let T_s to stand for the operator

$$T_s f(\xi, w) = ((\delta + |\xi|^2)^2 + w^2)^{\frac{n+2}{4}} \tilde{G}_s\big(((\delta + |\xi|^2)^2 + w^2)^{-\frac{n+2}{4}} f\big)(\xi, w),$$

then it follows that

$$T_s((\delta + |\xi|^2)^2 + w^2)^{\frac{s}{2}} = C_{n,s}\delta^s((\delta + |\xi|^2)^2 + w^2)^{-\frac{s}{2}}.$$

Therefore, using Schur test we can prove the following inequality (see Section 5.1 in [13]).

Theorem 9 *Let $f \in L^2(\mathbb{R}^{n+1})$ be real valued and assume that $G_s f \in L^2(\mathbb{R}^{n+1})$. Then for any $\delta > 0$ we have the inequality*

$$\langle \tilde{G}_s f, f\rangle \geqslant A_1(n, s)\delta^s \int_{\mathbb{R}^{n+1}} \frac{(f(\xi, w))^2}{((\delta + |\xi|^2)^2 + w^2)^s}\, d\xi\, dw$$

where $A_1(n, s) = 4^s \frac{\Gamma(\frac{n+2+2s}{4})^2}{\Gamma(\frac{n+2-2s}{4})^2}$. The inequality is sharp and equality holds when $f = u_{-s,\delta}$.

Remark 10 In view of (13) we also have the Hardy inequality for G_s, namely

$$\langle G_s f, f\rangle \geqslant C_1(s)A_1(n,s)\delta^s \int_{\mathbb{R}^{n+1}} \frac{(f(\xi,w))^2}{((\delta+|\xi|^2)^2+w^2)^s}\,d\xi\,dw.$$

Finally, using the above Hardy inequality we can prove the following trace Hardy inequality for the Grushin operator G.

Theorem 11 *Let $\delta > 0$ and $0 < s < 1$. For any real valued compactly supported continuous function u on $\mathbb{R}^{n+1} \times [0,\infty)$ which is smooth for $\rho > 0$, we have the inequality*

$$\int_0^\infty \int_{\mathbb{R}^{n+1}} |\nabla u(\xi,w,\rho)|^2 \rho^{1-2s}\,d\xi\,dw d\rho$$
$$\geqslant C_1(s)B_s A_1(n,s)\delta^s \int_{\mathbb{R}^{n+1}} \frac{u(\xi,w,0)^2}{((\delta+|\xi|^2)^2+w^2)^s}\,d\xi\,dw.$$

In view of Hardy's inequality for G_s all we have to do is to prove the following energy estimate for the Grushin operator. The following result is the analogue of Theorem 1.2 in [8] proved in the context of Heisenberg groups.

Theorem 12 *Let $\delta > 0$ and $0 < s < 1$. For any real valued compactly supported continuous function u on $\mathbb{R}^{n+1} \times [0,\infty)$ which is smooth for $\rho > 0$, we have the inequality*

$$\int_0^\infty \int_{\mathbb{R}^{n+1}} |\nabla u(\xi,w,\rho)|^2 \rho^{1-2s}\,d\xi\,dw d\rho \geqslant B_s\langle G_s f, f\rangle$$

where $f(\xi,w) = u(\xi,w,0)$ and $B_s = 2^{-1-2s}\pi^{-1-s}\Gamma(1-s)$.

Proof In proving this theorem we closely follow the proof of Theorem 1.2 in [8]. We therefore give only a sketch of the proof referring to [8] for details. In what follows we assume that u is smooth. The general case can be dealt with using an approximation argument.

Let $H^s(\mathbb{R}^{n+1})$ be the completion of $C_0^\infty(\mathbb{R}^{n+1})$ with respect to the norm $\|f\|_{(s)}^2 = \langle G_s f, f\rangle$. It can be verified that the dual of $H^s(\mathbb{R}^{n+1})$ is $H^{-s}(\mathbb{R}^{n+1})$. If $g \in H^{-s}(\mathbb{R}^{n+1})$, it follows that $h = G_s^{-1}g = G_{-s}g \in H^s(\mathbb{R}^{n+1})$. Let $H(\xi,w,\rho)$ be the solution of extension problem with initial condition h defined as in (6). In view of Theorem 2 and Theorem 6, $H(\xi,w,\rho)$ converges to $C_s h$ with $C_s = \frac{1}{4}\Gamma(s)\pi^{-s-1}$ and $-\rho^{1-2s}\partial_\rho H(\xi,w,\rho)$ converges to $B_s G_s h$ with $B_s = 2^{-1-2s}\pi^{-1-s}\Gamma(1-s)$. If we let $W(\xi,w,q) = H(2^{-1/2}\xi, 2^{-1}w,\rho)$ with $q = \rho^2/2$, then W satisfies the equation

$$2\big(q\partial_q^2 + (1-s)\partial_q + q\partial_w^2 - G\big)W(\xi,w,q) = 0. \tag{14}$$

Moreover, as $q \to 0$, we have $W(\xi, w, q) \to C_s h(\xi, w)$ in $H^s(\mathbb{R}^{n+1})$ and $-q^{1-s}\partial_q W(\xi, w, q) \to 2^{s-1}B_s G_s h(\xi, w)$ in $H^{-s}(\mathbb{R}^{n+1})$. We define $U(\xi, w, q) = u(\xi, w, \rho)$ with $q = \rho^2/2$ and proceed as in the proof of Theorem 1.2 in [8]. We leave the details to the reader. □

6 An Isometry Property of the Solution Operator Associated to the Extension Problem

In this section we will prove an isometry property of the solution operator associated to the extension problem for the Grushin operator. Such a property has been already proved in the context of $\mathbb{R}^n$ and $\mathbb{H}^n$ in [11], see also [14]. Let $u(\xi, w, \rho)$ be the solution of the extension problem given by (6). For $s > 0$, for which G_s makes sense (e.g. $0 < s < (n+1)$), recall that the Sobolev space $H^s(\mathbb{R}^{n+1})$ is defined as the completion of $C_0^\infty(\mathbb{R}^{n+1})$ under the norm

$$\|f\|_{(s)}^2 = \langle G_s f, f\rangle = \|G_s^{1/2} f\|_{L^2(\mathbb{R}^{n+1})}^2.$$

Let $\Phi_\alpha^\lambda,\ \alpha \in \mathbb{N}^n$ be the scaled Hermite functions which are eigenfunctions of scaled Hermite operator $H(\lambda) = -\Delta + \lambda^2|\xi|^2$. In view of the spectral decomposition of G_s we have that

$$\|f\|_{(s)}^2 = (2\pi)^{-1}\int_{-\infty}^{\infty}(2|\lambda|)^s\left(\sum_{\alpha\in\mathbb{N}^n}\frac{\Gamma(\frac{2|\alpha|+n+1+s}{2})}{\Gamma(\frac{2|\alpha|+n+1-s}{2})}|\langle f^\lambda, \Phi_\alpha^\lambda\rangle|^2\right)d\lambda.$$

We think of the solution $u(\xi, w, \rho)$ as a function on $\mathbb{R}^{n+3}$ which is radial in the third variable. Thus $U(\xi, w, \zeta) = u(\xi, w, |\zeta|)$ is a function on $\mathbb{R}^{n+3}$. For $(\alpha, \beta) \in \mathbb{N}^n \times \mathbb{N}^2$, let $\Phi_{\alpha,\beta}^\lambda(\xi, \zeta) = \Phi_\alpha^\lambda(\xi)\Phi_\beta^{\lambda/2}(\zeta)$, where $\Phi_\alpha^\lambda(\xi)$ and $\Phi_\beta^{\lambda/2}(\zeta)$ are Hermite functions on $\mathbb{R}^n$ and $\mathbb{R}^2$ respectively. Now we define the Sobolev space $\widetilde{H}^{s+1}(\mathbb{R}^{n+3})$ in terms of $\Phi_{\alpha,\beta}^\lambda(\xi, \zeta)$ as the space of all functions $U \in L^2(\mathbb{R}^{n+3})$ for which the following norm is finite.

$$\begin{aligned}&\|U\|_{\widetilde{H}^{s+1}(\mathbb{R}^{n+3})}^2\\ &= (2\pi)^{-1}\int_{-\infty}^{\infty}(2|\lambda|)^{s+1}\left(\sum_{(\alpha,\beta)\in\mathbb{N}^n\times\mathbb{N}^2}\frac{\Gamma(\frac{2|\alpha|+|\beta|+n+1+1+s+1}{2})}{\Gamma(\frac{2|\alpha|+|\beta|+n+1+1-s-1}{2})}|\langle U^\lambda, \Phi_{\alpha,\beta}^\lambda\rangle|^2\right)d\lambda\end{aligned}$$

where U^λ is defined as usual by

$$U^\lambda(\xi;\zeta) = \int_{-\infty}^{\infty} U(\xi, w, \zeta)e^{i\lambda w}d\lambda.$$

duplicating the cube by again finding inspiration from seventh book the Mathematical Collection of Pappus.

Pag. 243 (259)

Trisect an Angle

The third postulate allows Viete to trisect an arbitrary angle.

20210928///

www.reschio.com/luxury-villas-umbria

Fulvio Ricci

fricci@sns.it

Books: (Estes dois livros conjuntamente).

20210928/Geometric Aspects of Harmonic Analysis/

2021_Book_GeometricAspects...pdf

Pag. 12(24) 5. Trace Hardy and Hardy Inequalities vector fields

Pág. 467(476)

~~On Regularity and Irregularity of Certain Holomorphic Singular Integral Operators.~~

Cauchy-Szegõ projection

20210928/Fulvio-Ricci/

hardy.pdf

Pág. 3 Chapter I

Hardy Spaces on the Unit Disc

Pág. 26 6. The Cauchy Projection

Pág. 38 8. Dual Spaces

the Cauchy projection

Pág. 56 4. The Cauchy Projection and the Hilbert Transform.

↓

Since the orthogonal projection of $L^2(\mathbb{R})$ onto L^2_+ is obviously the map $g \mapsto g\chi_{\mathbb{R}_+}$, ...

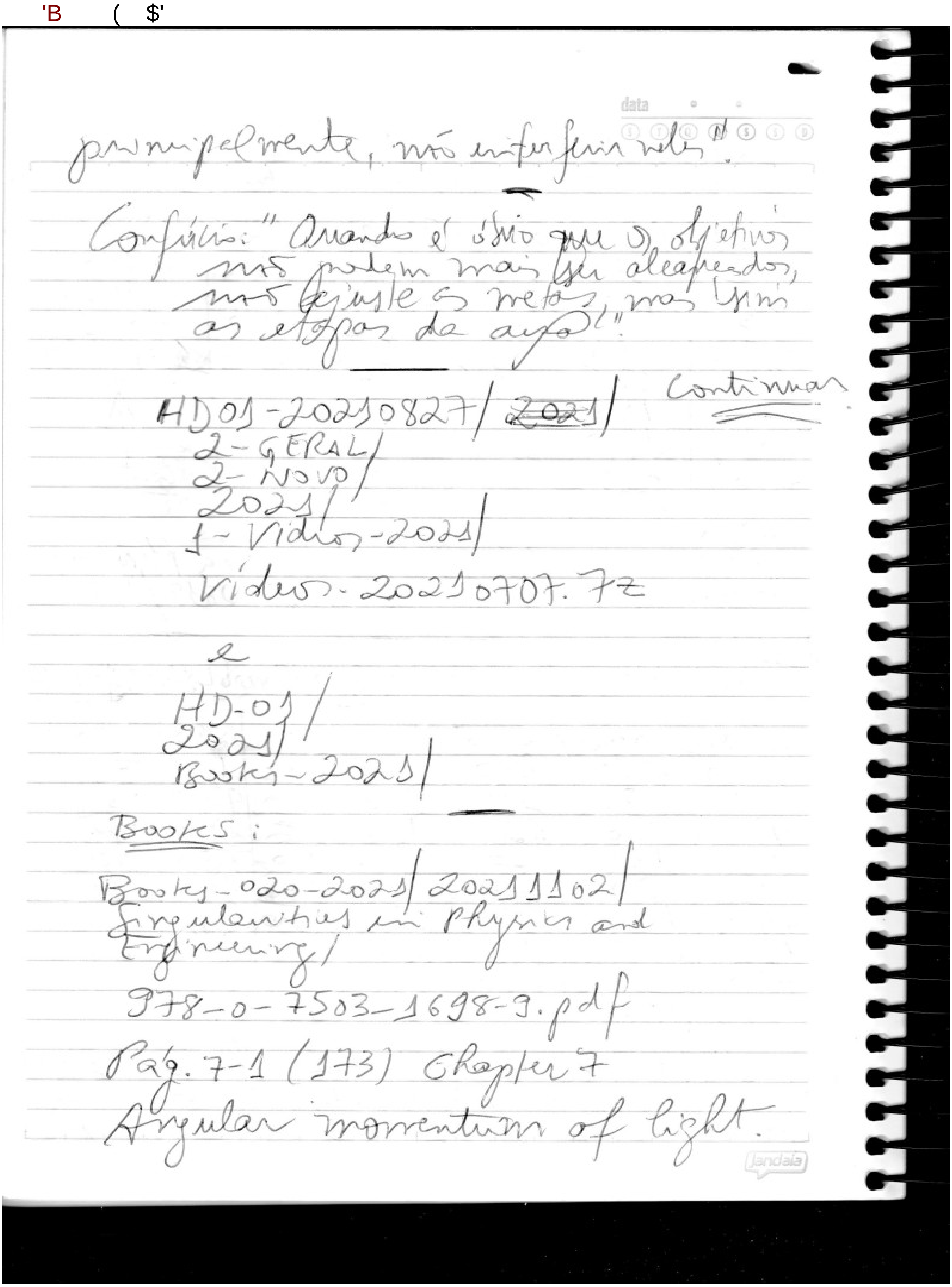

principalmente, não interferir nelas".

Confúcio: "Quando é óbvio que os objetivos
não podem mais ser alcançados,
não ajuste as metas, mas sim
as etapas da ação!"

HD01-20210827/ ~~2021~~/ Continuar
2-GERAL/
2-NOVO/
2021/
1-Videos-2021/
Videos-20210707.7z

e

HD-01/
2021/
Books-2021/

Books:

Books-020-2021/20211102/
Singularities in Physics and
Engineering/
978-0-7503-1698-9.pdf
Pag. 7-1 (173) Chapter 7
Angular momentum of light.

! -256- - -* ! - 'B / $ % B/ / 1"
(% !1 2 !

IOP Publishing

Singularities in Physics and Engineering
Properties, methods, and applications
Paramasivam Senthilkumaran

Chapter 7

Angular momentum of light

7.1 Introduction

Electromagnetic waves carry energy and momentum. The energy flux is given by the Poynting vector $\vec{S}$ and the momentum density is given by $\vec{\mathscr{P}} = \epsilon_0\mu_0\vec{S}$. Johannes Kepler used the concept of radiation pressure way back in the seventeenth century, to explain why the tail of a comet around the Sun's orbit always points away from the Sun. This radiation pressure is due to the momentum carried by light and it can exert a pressure on any object exposed to light. Maxwell published a paper on radiation pressure in 1862 and it was experimentally demonstrated by Lebedev in 1900.

7.2 Linear momentum

The linear momentum carried by electromagnetic waves is given by $\vec{p} = \hbar\vec{k} = \frac{E}{c}\hat{k}$ for a linearly polarized light. This is the momentum per photon. Here E is the energy per photon, c speed of light and $\vec{k}$ is the propagation vector. The linear momentum density $\vec{\mathscr{P}}$ averaged over a cycle, of a light beam can be calculated from its electric and magnetic fields

$$\vec{\mathscr{P}} = \epsilon_0\langle\vec{E} \times \vec{B}\rangle \tag{7.1}$$

where $\vec{E}$ and $\vec{B}$ are electric and magnetic fields respectively and ϵ_0 is the permittivity of free space.

Consider a monochromatic plane wave traveling in the z direction. The electric and magnetic fields of the wave are

$$\begin{aligned}\vec{E} &= E_0\cos(kz - \omega t)\hat{x}\\ \vec{B} &= \frac{E_0}{c}\cos(kz - \omega t)\hat{y}\end{aligned} \tag{7.2}$$

doi:10.1088/978-0-7503-1698-9ch7

© IOP Publishing Ltd 2018

Hence the momentum density is given by

$$\overrightarrow{\mathscr{P}} = \frac{1}{c}\epsilon_0 {E_0}^2 \cos^2(kz - \omega t)\hat{z} = \frac{1}{c}\epsilon_0 E^2 \hat{z} \tag{7.3}$$

The energy per unit volume $\mathscr{E}$ stored in EM fields is

$$\frac{1}{2}\left(\epsilon_0 E^2 + \frac{1}{\mu_0}B^2\right) = \epsilon_0 E^2 = \mathscr{E} \tag{7.4}$$

7.3 Angular momentum

Poynting using the analogy with the wave motion predicted that the circularly polarized photons carry $\hbar$ units of angular momentum [1]. A circularly polarized light passing through a half wave plate can transfer its angular momentum to the plate to become a circularly polarized light with opposite handedness and this momentum transfer was detected by Beth in his experiments with suspended quarter wave plate [2]. This angular momentum carried by circularly polarized light is called spin angular momentum (SAM) of light. Apart from that, the angular momentum carried by the phase of the light is called orbital angular momentum (OAM).

In 1992, Allen *et al* recognized that light beams with an azimuthal phase dependence of $\exp(il\phi)$ carry an orbital angular momentum (OAM) that can be many times greater than the spin angular momentum, where ϕ is the azimuthal coordinate in the beam's cross section, and l can take any integer value, positive or negative [3, 4]. The magnitude of orbital angular momentum carried by the beam is related to the topological charge of the beam and is $\pm l\hbar$ per photon. The key point was that this OAM was a natural property of all helically phased beams, and hence could be readily realizable in the laboratory. There is also great interest in these types of beams with OAM due to diverse possible applications [5–9]. Orbital angular momentum arises whenever a beam's phase fronts are not perpendicular to the general propagation direction and hence transverse components of propagation vectors are present. Termed as internal flows, the angular momentum and internal energy carried by such beams were described in chapter 5. The most common form of helically phased beam is the so-called Laguerre–Gaussian (LG) laser hybrid mode. Other beams with helical wavefronts such as Bessel beams [10], Mathieu beams [11], and Ince–Gaussian beams [12] can also carry orbital angular momentum. In chapter 9, it is shown that plane wavefronts with characteristic inhomogeneous polarization distribution can also carry orbital angular momentum. In the same way that the suspended half-wave plate acquires an angular momentum of $2\hbar$ when a right circularly polarized photon passes through it (figure 7.1), the cylindrical mode converter can acquire an angular momentum of $2l\hbar$ when a photon with orbital angular momentum $l\hbar$ passes through it.

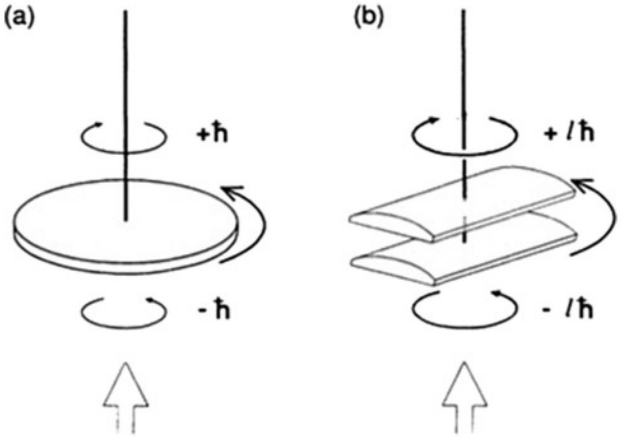

Figure 7.1. (a) A suspended half-wave plate undergoes torque in transforming right-handed into left-handed circularly polarized light, Beth's experiment. (b) Suspended cylindrical lenses undergo torque in transforming a Laguerre–Gaussian mode of orbital angular momentum $-l\hbar$ per photon, into one with $+l\hbar$ per photon [3]. Reprinted with permission from [3]. Copyright (1992) by the American Physical Society.

The local wavefront normal indicates the direction of energy and momentum flow in the fields. Since the vortex beam has helical shaped wavefronts the local normal to the wavefront also spirals as the beam advances. This gives rise to transverse components of propagation vector distribution which has non-zero curl. This suggests that the singular beam carries OAM. The photon trajectories for a Bessel beam with $l = 2$ are shown in figure 7.2. These trajectories are reconstructed from measurements of the transverse Poynting-vector distribution, with a constant longitudinal component added.

The angular momentum density, $\vec{\mathscr{J}}$, of a light beam can be calculated from

$$\vec{\mathscr{J}} = \epsilon_0(\vec{r} \times \langle \vec{E} \times \vec{B} \rangle) = \vec{r} \times \vec{\mathscr{P}} \tag{7.5}$$

This equation encompasses both the spin and orbital angular momentum density of a light beam. The local value of the linear momentum density [16] under paraxial approximation is given by

$$\vec{\mathscr{P}} = i\omega\frac{\epsilon_0}{2}(u^*\nabla u - u\nabla u^*) + \omega k\epsilon_0|u|^2\hat{z} + \omega\sigma\frac{\epsilon_0}{2}\frac{\partial|u|^2}{\partial r}\hat{\phi} \tag{7.6}$$

where $u = u(r, \phi, z)$ is the complex scalar function describing the distribution of the field amplitude. Here σ describes the degree of polarization of the light; $\sigma = \pm 1$ for right- and left-hand circularly polarized light, respectively, and $\sigma = 0$ for linearly polarized light. The cross product of this linear momentum density with the radius vector $\vec{r}(r, 0, z)$ yields an angular momentum density. The angular momentum density in the z direction can come from the first or the third term of the linear momentum density equation (7.6). As one can see, the first term depends on the phase gradient of the complex field and the third term depends on the polarization of

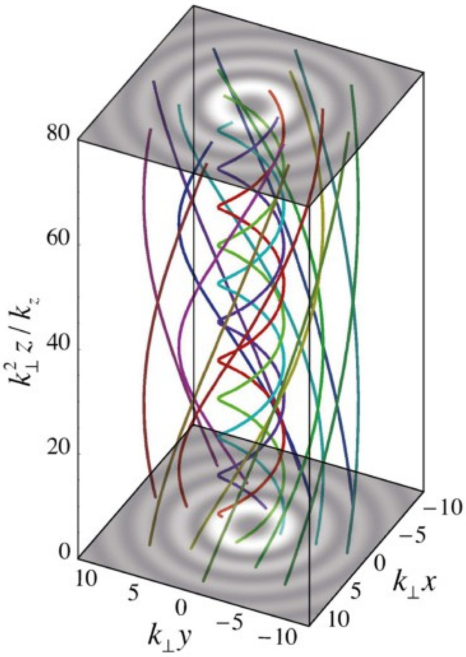

Figure 7.2. Photon trajectories are helical in a Bessel beam with a vortex of charge 2. This suggests that photon momentum is helical and hence singular beam carries OAM. Reproduced from [18] © IOP Publishing Ltd and Deutsche Physikalische Gesellschaft. CC BY-NC-SA 3.0.

the complex field. The angular momentum arising from the cross product of radius vector $\overrightarrow{r}$ with the phase gradient term is called the orbital angular momentum and the cross product with the third term is called the spin angular momentum of light.

7.4 Orbital and spin angular momentum of light

7.4.1 Angular momentum due to circular polarization

The spin angular momentum arising due to polarization of light is given by $\sigma_z\hbar$ per photon with $-1 \leqslant \sigma_z \leqslant 1$, where the extrema $\sigma_z = \pm 1$ corresponds to pure circular polarization.

From the azimuthal part of the linear momentum density given in equation (7.6), the z-component of the angular momentum density is given by

$$\mathscr{J}_z = |\overrightarrow{r} \times \overrightarrow{\mathscr{P}_\phi}| = r\mathscr{P}_\phi = r\omega\sigma\frac{\epsilon_0}{2}\frac{\partial |u|^2}{\partial r} \tag{7.7}$$

The angular momentum density in equation (7.7) integrated over the beam, gives spin angular momentum (SAM) and is given by $\sigma\hbar$.

7.4.2 Angular momentum due to azimuthal phase dependence in the beam

The z-component of the angular momentum density arising from the phase gradient terms in equation (7.6) is given by

$$\mathscr{J}_z = \overrightarrow{r} \times \left\{ i\omega\frac{\epsilon_0}{2}(u^*\nabla u - u\nabla u^*) \right\} \tag{7.8}$$

The angular momentum density in equation (7.8) integrated over the beam gives OAM and is given by $l\hbar$. Note that in chapter 5, we have seen that the optical current in terms of phase gradient is given by

$$\frac{i}{2}(u^*\nabla u - u\nabla u^*) = Im(u^*\nabla u) = I\nabla\phi \tag{7.9}$$

The orbital angular momentum is due to the azimuthal phase dependence in the beam and is equivalent to $l\hbar$ per photon. The analogy between paraxial wave equation with the Schrodinger wave equation suggests the identification of an orbital angular momentum for light beams [13]. The operator corresponding to the z-component of orbital angular momentum is $L_z = -i\hbar\frac{\partial}{\partial\phi}$ and applying this to beams having azimuthal phase dependence like $\exp\{il\phi\}$ gives an orbital angular momentum component to phase singular beams. This analogy between paraxial optics and quantum mechanics has been exploited to introduce an eigenfunction description of laser beams, and quantum harmonic oscillator modes [14]. Laguerre–Gaussian modes are the laser mode analog of the angular momentum eigenstates of the isotropic two-dimensional harmonic oscillator.

In the case of speckle patterns, the overall OAM is zero as there is an equal number of positive and negative charge optical vortices. The angular momentum density $\overrightarrow{\mathscr{J}}$, is related to the linear momentum density $\overrightarrow{\mathscr{P}} = \epsilon_0\langle\overrightarrow{E} \times \overrightarrow{B}\rangle$ by the relation $\overrightarrow{\mathscr{J}} = \overrightarrow{r} \times \overrightarrow{\mathscr{P}}$ where ϵ_0 is the free space permittivity and $\overrightarrow{E}$ and $\overrightarrow{B}$ are the electric and magnetic fields, respectively. From this relation, for any angular momentum component directed along the z direction, there must be a transverse component of linear momentum, i.e. in the xy-plane. This means that to have a z-component of angular momentum there must be a transverse component, i.e. a z-component of the electric and/or magnetic field. Further, such a transverse component of momentum should be circulating. The center of the vortex beam corresponds to the position of zero intensity and hence it carries neither linear nor angular momentum. Indeed, the angular momentum is associated with regions of high intensity, which is a bright annular ring for a donut vortex beam. The azimuthal component of the Poynting vector is parallel to that of the transverse linear momentum density in a helical phase front [7]. Spin angular momentum (SAM) has only two independent states corresponding to left- and right-handed circular polarization, while orbital angular momentum has an unlimited number of possible states, corresponding to all integer values of l.

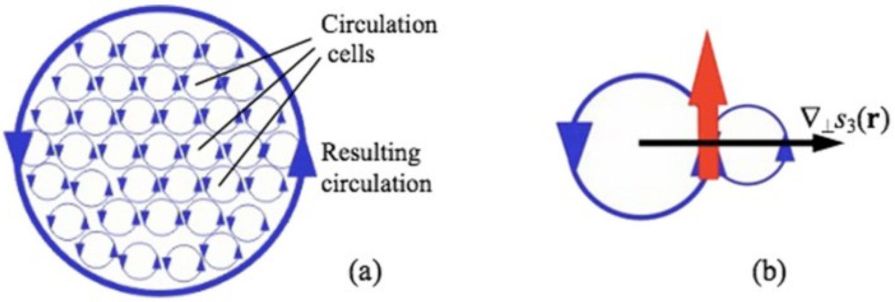

Figure 7.3. Orbital angular momentum due to spatially varying spin angular momentum. (a) In a uniform distribution of spin angular momentum according to Stokes theorem, the internal circulations cancel out and only circulation at the boundary remains. (b) In an non-uniform distribution of spin angular momentum, there can be a volume spin current, as there is incomplete internal cancellations of circulations. Reproduced with permission from the corresponding author [15]. © IOP Publishing. Reproduced with permission. All rights reserved.

7.4.3 Angular momentum due to spatially varying circular polarization

There is a possibility that a spatially varying circular polarization can result in orbital angular momentum. This can result in optical current flow across the beam cross section. Consider the fundamental theorem of curls which states that

$$\int\int(\nabla \times \vec{A}) \cdot \vec{da} = \oint \vec{A} \cdot \vec{dl} \tag{7.10}$$

All the elementary circulations, which are of the same magnitude, can result in a single large circulation around the boundary enclosing the region. This is due to the cancellation of circulating components between adjacent circulating cells. This is explained in figure 7.3(a). But when there is a spatially varying spin angular momentum distribution, there can be a volume spin current due to incomplete cancellation of circulations. As a result any larger circulation arising due to the spatially varying circular polarization is not origin-dependent as it arises due to individual spins which are origin-independent. In figure 7.3(b), the spin current arising between two unequal adjacent circular polarization components is depicted. The gradient of Stokes' S_3 parameter in the figure indicates spatially varying circular polarization. The Stokes' parameter $S_3 = \pm 1$ for circular states. Any state of polarization can be decomposed into right and left circularly polarized components and ∇S_3 in the figure represents spatial variation of SAM component.

7.5 Intrinsic and extrinsic angular momenta

There are three forms of angular momenta namely intrinsic spin angular momentum, intrinsic orbital angular momentum and extrinsic orbital angular momentum [16] abbreviated as EOAM. The angular momentum that is independent of the choice of the axis is intrinsic. The spin angular momentum is therefore intrinsic and the orbital angular momentum carried by phase singularity with its core along the optical axis of the system is intrinsic. Any vortex which is situated away from the center of the beam axis gives rise to extrinsic orbital angular momentum.

These three forms of the AM of light beams correspond to different degrees of freedom of light and to different structural constituents. In free space propagation these different forms of AM are conserved. Any transverse shift of the field centroid (trajectory), leads to the conversion from intrinsic OAM to the extrinsic OAM.

The angular momentum which arises for any light beam from the product of the z-component of linear momentum about a radius vector, may be said to be extrinsic because its value depends upon the choice of calculation axis.

Berry showed [17] that the orbital angular momentum of a light (J) beam does not depend upon the lateral position of the axis and can therefore also be said to be intrinsic, provided the direction of the axis is chosen so that the transverse momentum is zero. When integrated over the whole beam the angular momentum in the z direction is

$$\vec{J_z} = \epsilon_0 \int \int dxdy(\vec{r} \times \langle \vec{E} \times \vec{B} \rangle) \tag{7.11}$$

If the axis is laterally displaced by $\vec{r_0}(r_{0x}, r_{0y})$ it is easy to show that the change in the z-component of angular momentum is given by

$$\begin{aligned} \Delta J_z &= (r_{0x} \times P_y) + (r_{0y} \times P_x) \\ &= r_{0x}\epsilon_0 \int \int dxdy\langle \vec{E} \times \vec{B} \rangle_y + r_{0y}\epsilon_0 \int \int dxdy\langle \vec{E} \times \vec{B} \rangle_x \end{aligned} \tag{7.12}$$

The angular momentum is intrinsic only if ΔJ_z equals zero. Any beam with a helical phase front apertured symmetrically about the beam axis has zero ΔJ_z and, consequently, an orbital angular momentum of $l\hbar$ per photon, independent of the axis of calculation [16]. The orbital angular momentum of the light beam may therefore be described as intrinsic. However, when the beam is passed through an off-axis aperture, its transverse momentum is non-zero and the orbital angular momentum depends upon the choice of calculation axis and so must be described as extrinsic.

The OAM is intrinsic only if the vortex core coincides with the beam axis. In this case the OAM will turn out to be independent of the choice of the calculation axis. For apertured beams even if the vortex is positioned at the center of the aperture and the beam axis does not coincide with the center of the aperture, the computed OAM will yield different results and is dependent on the choice of the coordinate system used. Therefore, the OAM in this case is extrinsic. This means that vortices positioned away from the beam axis (also called the centroid of the beam) have extrinsic OAM. Therefore, it does not follow that the angular momentum of the apertured beam is intrinsic as the result does depend upon the choice of calculation axis. Figure 7.4 depicts the numerically calculated local spin and orbital angular momentum densities in the direction of propagation for a circularly polarized LG beam. It shows how (1) the axis of the beam, (2) the position about which angular momentum is calculated and (3) aperturing, affect the calculated OAM. Further,

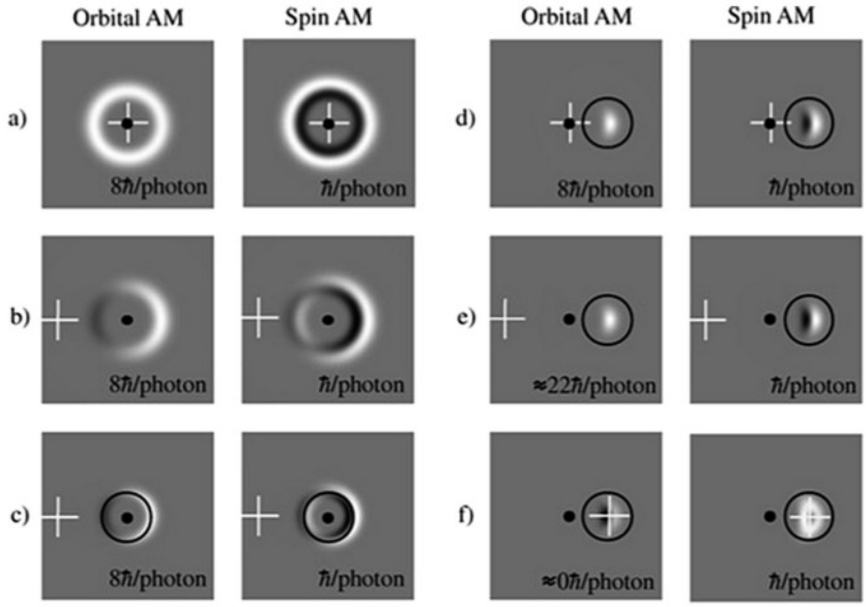

Figure 7.4. Numerically calculated local spin and orbital angular momentum densities in the direction of propagation for a $l = 8$ and $\sigma = 1$ Laguerre–Gaussian mode [16]. In the gray scale, white represents the positive and black the negative contributions. The axis of the beam is marked by a black spot while the axis about which the angular momenta are calculated is shown by a white cross. The black circle marks the position of a soft edged aperture. Note that the spin angular momentum is equivalent to $\sigma\hbar$ per photon irrespective of the choice of aperture or calculation axis, whereas the orbital angular momentum is only $l\hbar$ per photon if the aperture or calculation axis coincide with the axis of the original beam. Reprinted with permission from [16], Copyright (2002) by the American Physical Society.

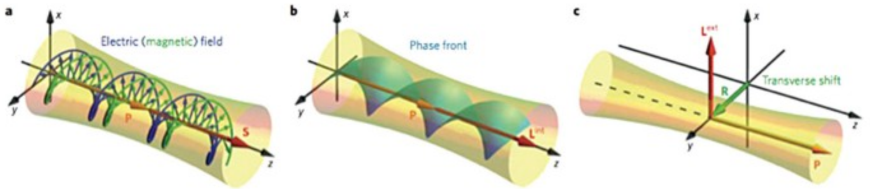

Figure 7.5. Angular momenta of paraxial optical beams [19]. (a) SAM for a right-hand circularly polarized beam with $\sigma = 1$. The instantaneous electric and magnetic field vectors are shown. (b) IOAM in a vortex beam with $l = 2$. The instantaneous surface of a constant phase is shown. (c) EOAM due to the propagation of the beam at a distance R from the coordinate origin. Reprinted by permission from [19]. Copyright (2015) Springer Nature.

figure 7.5 shows the EOAM due to the shift in the centroid of the beam from the optical axis.

References

[1] Poynting J H 1909 The wave motion of a revolving shaft, and a suggestion as to the angular momentum in a beam of circularly polarised light *Proc. R. Soc. Lond. Ser.* A **82** 560–7

[2] Beth R A 1936 Mechanical detection and measurement of the angular momentum of light *Phys. Rev.* **50** 115–27

[3] Allen L, Beijersbergen M W, Spreeuw R J and Woerdman J P 1992 Orbital angular momentum of light and the transformation of Laguerre–Gaussian laser modes *Phys. Rev.* A **45** 8185–9

[4] Allen L, Padgett M J and Babikar M 1999 The orbital angular momentum of light *Progress in Optics* vol 39 (Amsterdam: Elsevier)

[5] Padgett M J 2017 Orbital angular momentum 25 years on *Opt. Express* **25** 11265–74

[6] Simpson B N, Dholakia K, Allen L and Padgett M J 1997 Mechanical equivalence of spin and orbital angular momentum of light: An optical spanner *Opt. Lett.* **22** 52–4

[7] Torres J P and Torner L 2011 *Twisted Photons: Applications of Light with Orbital Angular Momentum* (New York: Wiley)

[8] Willner A E *et al* 2015 Optical communications using orbital angular momentum beams *Adv. Opt. Photon* **7** 66–106

[9] Yao A M and Padgett M J 2011 Orbital angular momentum: origins, behavior and applications *Adv. Opt. Photon* **3** 161–204

[10] McGloin D and Dholakia K 2005 Bessel beams: diffraction in a new light *Contemp. Phys.* **46** 15–28

[11] Gutiérrez Vega J C, Iturbe-Castillo M and Chav́ez Cerda S 2000 Alternative formulation for invariant optical fields: Mathieu beams *Opt. Lett.* **25** 1493–5

[12] Bandres M A and Gutiérrez Vega J 2004 Ince–Gaussian beams *Opt. Lett.* **29** 144–6

[13] van Enk S J and Nienhuis G 1992 Eigenfunction description of laser beams and orbital angular momentum of light *Opt. Commun.* **94** 147–58

[14] Nienhuis G and Allen L 1993 Paraxial wave optics and harmonic oscillators *Phys. Rev.* A **48** 656–65

[15] Bekshaev A, Bliokh K Y and Soskin M 2011 Internal flows and energy circulation in light beams *J. Opt.* **13** 053001

[16] O'Neil A T, MacVicar I, Allen L and Padgett M J 2002 Intrinsic and extrinsic nature of the orbital angular momentum of a light beam *Phys. Rev. Lett.* **88** 053601

[17] Berry M V 1998 *Int. Conf. on Singular Optics* vol 3487 (Bellingham, WA: SPIE)

[18] Bliokh K Y, Bekshaev A Y, Kofman A G and Nori F 2013 Photon trajectories anomalous velocities and weak measurements: a classical interpretation *New J. Phys.* **15** 073022

[19] Bliokh K Y, Rodrigues-fortuno F J, Nori F and Zayats A V 2015 Spin-orbit interactions of light *Nat. Photon.* **9** 796–808

D -256- - -* ! - 'B
& " 3l6 ! l 3 l2 "l5 ! A 3l "l=" 1"
(!1

Chapter 10

Fine Structure of the Hydrogen Atom

Now that we have covered the two revolutionary theorems of quantum physics and relativity, we can start to think about how they mesh together. After all, a complete description of the universe should encompass both theories. We will study relativistic quantum mechanics in detail in Chapter 11 but, for now, consider the case of the hydrogen atom, which is an ideal system for starting to understand the connection. One of the reasons for this is that the speed of an electron in the ground state of the hydrogen atom is (see (2.49)) $v = \alpha c$, i.e. about $c/137$. This is significantly smaller than the speed of light, so relativistic effects normally can be ignored. However, in the very precise realm of atomic spectroscopy, these effects are measurable and need to be accounted for. The good news is that, since the corrections are small, we can use perturbation theory to study them and so do not need a fully relativistic quantum theory just yet.

10.1. Relativistic correction to the kinetic energy

When we considered the spectrum of the hydrogen atom in Section 5.4, we calculated the energy levels by using the nonrelativistic Hamiltonian. This was found by adding the kinetic energy and Coulomb potential energy of the electron, and converting this total energy to an operator by use of the correspondence principle. However, the form of the kinetic energy is different if we take into account special relativity. In this section, we will see how this effect changes the energy levels of hydrogen.

The relativistic form of the kinetic energy can be written as

$$T = \sqrt{p^2c^2 + M^2c^4} - Mc^2, \tag{10.1}$$

where the first term is the total relativistic energy and the second term is the rest energy of the electron. Expanding the square-root, we get[1]

$$\begin{aligned} T &= Mc^2 \left[\sqrt{1 + \frac{p^2}{M^2c^2}} - 1 \right] \\ &= Mc^2 \left[-1 + 1 + \frac{1}{2}\left(\frac{p^2}{M^2c^2}\right) - \frac{1}{8}\left(\frac{p^2}{M^2c^2}\right)^2 + \cdots \right] \\ &\approx \frac{p^2}{2M} - \frac{p^4}{8M^3c^2}. \end{aligned} \tag{10.2}$$

We can truncate the expansion since $p \ll Mc$. The first term in (10.2) is just the nonrelativistic kinetic energy and so the second term is the first-order correction to the kinetic energy due to relativistic effects. If the total Hamiltonian is $H = H_0 + H_1$, where H_0 is the nonrelativistic Hamiltonian, the perturbation term is

$$H_1 = -\frac{p^4}{8M^3c^2} = -\frac{E_k^2}{2Mc^2},$$

where $E_k = p^2/(2M)$ is the nonrelativistic form of the kinetic energy. Using perturbation theory, the first-order correction to a general (unperturbed) state Ψ_{nlm} of the hydrogen atom is, therefore,

$$\begin{aligned} \Delta E_1 &= -\frac{1}{2Mc^2}\langle \Psi_{nlm}|E_k^2|\Psi_{nlm}\rangle \\ &= -\frac{1}{2Mc^2}\langle (E_n - V(r))^2 \rangle \\ &= -\frac{1}{2Mc^2}\left(\langle E_n^2\rangle - 2\langle E_n\, V(r)\rangle + \langle V^2(r)\rangle\right). \end{aligned} \tag{10.3}$$

In the second line, we have used $E_k = E_n - V(r)$, where $E_n \approx -13.6\,\mathrm{eV}/n^2$ is the total energy, $V(r) = -e^2/(4\pi\epsilon_0 r)$ is the potential energy, and we have

---

[1] Using Newton's generalised binomial theorem:

$$(x+y)^r = x^r + rx^{r-1}y + \frac{r(r-1)}{2!}x^{r-2}y^2 + \cdots.$$

dropped the state we are averaging over for notational simplicity. Now, since the wave functions ψ_{nlm} are eigenfunctions of E_n, we have

$$\begin{aligned}\Delta E_1 &= -\frac{1}{2Mc^2}\left(E_n^2 - 2E_n\langle V\rangle + \langle V^2\rangle\right) \\ &= -\frac{1}{2Mc^2}\left(E_n^2 - 2E_n\langle\frac{-e^2}{4\pi\epsilon_0 r}\rangle + \langle\frac{e^4}{(4\pi\epsilon_0)^2 r^2}\rangle\right). \qquad (10.4)\end{aligned}$$

As an example, let's calculate the energy shift for the 1s state of hydrogen. We know from Section 5.4 that the wave function for this state is

$$\Psi_{100} = \frac{1}{\sqrt{\pi a_0^3}}\,\mathrm{e}^{-r/a_0}, \qquad (10.5)$$

and $E_1 = -\alpha^2 Mc^2/2$ (see (2.52)), where α is the fine-structure constant as given in (2.50). We can now evaluate $\langle V\rangle$ and $\langle V^2\rangle$ directly,[2]

$$\begin{aligned}\langle V\rangle &= 4\pi\int_0^\infty r^2\left(\frac{-e^2}{4\pi\epsilon_0 r}\right)\frac{1}{\pi a_0^3}\,\mathrm{e}^{-2r/a_0}\,dr \\ &= \frac{-e^2}{4\pi\epsilon_0 a_0} = -\alpha^2 Mc^2 \qquad (10.6) \\ \langle V^2\rangle &= 4\pi\int_0^\infty r^2\left(\frac{-e^2}{4\pi\epsilon_0 r}\right)^2\frac{1}{\pi a_0^3}\,\mathrm{e}^{-2r/a_0}\,dr \\ &= 2\left(\frac{e^2}{4\pi\epsilon_0 a_0}\right)^2 = 2\alpha^4 M^2 c^4. \qquad (10.7)\end{aligned}$$

Substituting these results into (10.4), we obtain

$$\Delta E_1 = -\frac{5}{8}Mc^2\alpha^4, \qquad (10.8)$$

which is the first-order relativistic correction for the ground state of hydrogen. You can probably see now why α is called the fine-structure constant.

---

[2] Using the integral identity

$$\int_0^\infty r^n\,\mathrm{e}^{-\beta r}\,dr = n!\,\beta^{-(n+1)}.$$

A more general calculation[3] gives

$$\Delta E_1 = -\frac{1}{2}Mc^2\frac{(Z\alpha)^4}{n^4}\left(\frac{n}{l+1/2}-\frac{3}{4}\right). \tag{10.9}$$

We see that the energy now depends on l as well as n, i.e. the $2s$ and $2p$ states will have different energies. This is significant as it is the first departure we have seen from Bohr's theory, which says that the energy levels depend only on the quantum number n.

10.2. Addition of angular momenta

In an atom, there are different angular momenta associated with the particles, such as the orbit of the electrons and the spin of the electron or the nucleus. It is therefore important that we understand how to combine angular momenta in this setting. In classical physics, it is straightforward. Angular momentum is a continuous vector quantity and thus, to find the total angular momentum we just vectorially sum all the contributions

$$\vec{L} = \sum_i \vec{L}_i. \tag{10.10}$$

In quantum physics, things are more complicated since angular momentum is quantised not just in the values it can take, but also in the directions that it can point in. So the question is: how do we combine angular momenta in quantum physics? Let's start by considering the addition of two angular momenta since, once we know how to add two, we can add any number by applying the rules sequentially. Suppose we have two angular momenta $\vec{L}_1$ and $\vec{L}_2$ with corresponding quantum numbers l_1 and l_2. We know that the magnitude of the two vectors will be

$$|\vec{L}_1| = \hbar\sqrt{l_1(l_1+1)} \quad \text{and} \quad |\vec{L}_2| = \hbar\sqrt{l_2(l_2+1)} \tag{10.11}$$

and their components along the z axis will be $L_{1z} = m_1\hbar$ and $L_{2z} = m_2\hbar$, where m_1 and m_2 are given by $-l_1 \leq m_1 \leq l_1$ and $-l_2 \leq m_2 \leq l_2$. The two vectors will be able to add up in different (quantised) ways — as shown in Fig. 10.1 — that we want to understand. Firstly, it is clear that the component of the overall vector, $\vec{L} = \vec{L}_1 + \vec{L}_2$ in the z-direction, will simply

---

[3] Given as a exercise (with solution) at the end of the chapter.

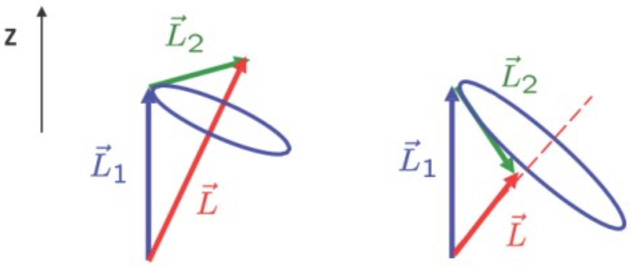

Figure 10.1: Addition of angular momenta. Two vectors $\vec{L}_1$ and $\vec{L}_2$ with fixed magitude are added and, owing to the different projections they can have on the z axis, result in different vectors $\vec{L}$. The allowed magnitudes of the resulting vector are given by (10.14). As described in the text, the good quantum numbers can be visualised by considering rotations about the $\vec{L}$ direction.

be the sum of the z-components of the other vectors, i.e.

$$L_z = m\hbar = L_{1z} + L_{2z} = m_1\hbar + m_2\hbar. \tag{10.12}$$

So we have $m = m_1 + m_2$. Since the overall vector is an angular momentum, we know that its length can be written in terms of a quantum number l in the usual form

$$|\vec{L}| = \hbar\sqrt{l(l+1)}. \tag{10.13}$$

The rule for finding l in terms of l_1 and l_2 is

$$|l_1 - l_2| \leq l \leq l_1 + l_2, \tag{10.14}$$

where l increases in integer steps. This is an important and general rule for adding angular momenta in quantum mechanics.

We can know the lengths of $\vec{L}$, $\vec{L}_1$ and $\vec{L}_2$ simultaneously; their operators all commute and so they are compatible observables. Their corresponding quantum numbers are all good quantum numbers, that is, they can all be specified simultaneously. As we have seen above, the z-component of $\vec{L}$ is also compatible with the other observables. However, m_1 and m_2 are not good quantum numbers. While their sum $m = m_1 + m_2$ is well-defined, we cannot specify m_1 and m_2 separately. A good way of visualising this is with Fig. 10.1. If we imagine that everything is rotating about the vector $\vec{L}$, then we can see that $\vec{L}$ has a fixed length and direction (i.e. l and m can be specified). We also see that $\vec{L}_1$ and $\vec{L}_2$ have fixed lengths, meaning that l_1 and l_2 can be specified. However, the projections of $\vec{L}_1$ and $\vec{L}_2$ onto the z axis change with time, meaning that m_1 and m_2 are not good quantum numbers.

Let's conclude this section with an example of how to apply the angular momentum addition rule given by (10.14). Suppose we have two particles, each with quantum numbers $l_1 = l_2 = 1$, e.g two electrons in a p subshell. The possible combined values of l are

$$0 \leq l \leq 2, \tag{10.15}$$

i.e. $l = 0, 1, 2$. This is an odd number of values and it is worth noting that we always get an odd number of possible l values whenever the individual values that are combined (l_1 and l_2 in this case), are integers. For each value of l, the possible values of m are given by the usual rule, $-l \leq m \leq l$. So for $l = 2$, we would have $m = -2, -1, 0, 1, 2$.

10.3. Spin-orbit coupling

In the early days of spectroscopy, scientists were puzzled by the appearance of closely spaced lines, called doublets, in the spectra of alkali metals such as lithium, sodium and potassium.[4] The lines were too close together to be able to be explained by Bohr's theory, where the energies depend only on n. However, as we saw in Section 10.1, relativistic effects mean that atomic spectra depend on l as well as n, so perhaps the doublet could be explained in terms of the addition on orbital angular momenta. Unfortunately, this was not so. Since the quantum numbers l can only take integer values, we have seen that they combine to always give an odd number of different values, so we should expect to see an odd number of closely spaced lines rather than the two we see in a doublet. This is all reminiscent of the Stern-Gerlach experiment that we saw in Section 6.1 and the resolution is, of course, the same — that the electron has a spin value of one-half. The precise details of the doublet can be explained in terms of coupling between the electron's orbital angular momentum and its spin — the so-called spin-orbit coupling. In order to understand spin-orbit coupling, it is helpful to first have an operator for magnetic moment.

We can find the magnetic moment operator by using classical physics and then applying the correspondence principle. Consider the classical

---

[4] The reader may be familiar with the sodium doublet from the orange street lamps often seen in cities. The light that these emit are from the two sodium doublet lines at around 589 nm. These lamps are particularly useful where light pollution is an issue, such as near astronomical observatories, since they emit light at two dominant frequencies and so have minimal interference with the other frequencies that we may be interested in seeing.

D ($$

20111002/ Introductory Quantum Physics...|

Introductory_Quantum_Physics_and_Relativity_1.pdf

Obs: Livro todo!!!

Pág. 187(200)- Chapter 10

Fine Structure of the Hydrogen Atom.

Aplicam a deformação do tempo-espaço

Obs: Look at equation: (10.7)
(10.6)

(10.6)

$$\langle V \rangle = 4\pi \int_0^\infty r^2 \left(\frac{-e^2}{4\pi\epsilon_0 r} \right) \frac{1}{\pi a_0^3} e^{-2r/a_0} \, dr$$

etc

O rio que esquece onde nasce, seca.

University of Rijeka
Department of Mathematics
Ulica Radmile Matejčić 2, Rijeka
Rijeka, Primorje-Gorski County, Croatia

20211003 (cont)

3-Aberto (zippando)

Books	45
Data	130
SV	
Videos	
Zip	
75	
120	
45	

Copiar 3-Aberto → p/
(amanhã) 3-Aberto-copy

/2021/DATA/

Cantina do Marinheiro

Mem de Sá 125 Moça

Infinite Potential film (2020)

Quantum Theory Bohm

The Quantum Dissidents

Cosmologia Quântica - um ramo da Física que tenta combinar Teoria Quântica com Gravitação, ou seja, Teoria Quântica com Teoria da Relatividade.

Efeitos Quânticos (Bohm) →

($

um deles Efeito Aharonov-Bohm

Emaranhamento Quântico e
Efeito Aharonov-Bohm

Einstein → Emaranhamento, ele chamava
"spooky action at a distance"

—

Bohm → outros Conceitos

a) Potencial Quântico
b) Totalidade
c) Ordem Implicada.

a) "Trata-se de uma função matemática
que conecta instantaneamente
partículas muito distantes!"

Tanto a Mecânica Quântica
quanto o Potencial Quântico
estavam sugerindo uma ideia
de Totalidade.

Livro: Totalidade e o Universo
Indiviso
(Bohm e Basil Hiley)

Niels Bohr estava certo,
você não pode separar o
observador do observado.

David Bohm introduziu o

($/

Conceito de Ordem –

a/ Ordem Implicada ou Implícita
b/ Ordem Explicada ou Explícita.

David Bohm e Jiddu Krishnamurti

Livro: ↙

"O observador é o observado"!

esposa de Bohm encontrou este livro.
Em uma livraria.

---

3 – Aberto/2021/Data (copy)

20211004

ZIPPAR: Videos-Geral

40	1150	1100	100	√SV
110	50	150	100	√Ungar
120	100	150	100	√Videos
40	300	150	100	√Videos-Import
310	130	550	100	√Videos-Math
230	50	100	500	√Videos-others
540	480	650		√Videos-por-data
60	150			
600	630			
100	50			
700	680			

' (B

Books:

① Books-020-2021/20211004/
Thomas' Calculus.../
Thomas_Calculus....pdf
~~Pág. 740 (765) Chapter 15~~
~~Differentiation~~
* Pág. 388 (389) Chapter 7
Transcendental Function

② 20211004/ Mathematics for
Engineering.../
Mathematics for Engineers...pdf
Pág. 740 (765) Chapter 15
Differentiation

③ 20211004/ Homotopy Theory.../
3030789764.pdf
→ (Homotopy theory and Arithmetic
Geometry – Motivic and Diophantine
Aspects.

! -256- - -* ! - '
9 ! I ! I I &IH I ' I " 1"
(% % / !1 9 ! ! <

D -256- - -* ! - ' ! ! = J $
=" 9 4< 1"
(' B$!1 $

-256- - -* ! - ' % % / B' 1"
*D : 13 9 3 ! " 2 7 3 A ! " 4 1 ! 2 1

" -256- - -* ! - '
:! "D .I I ! I= I " 1"

-256- - -* ! - ' I/ % % % $ 1"
($ 7! ! 8 7 !

-256- - -* ! - ' = = ! ! "
@!A 1"
(' ' =L! 1 $ % A " D! < $ B

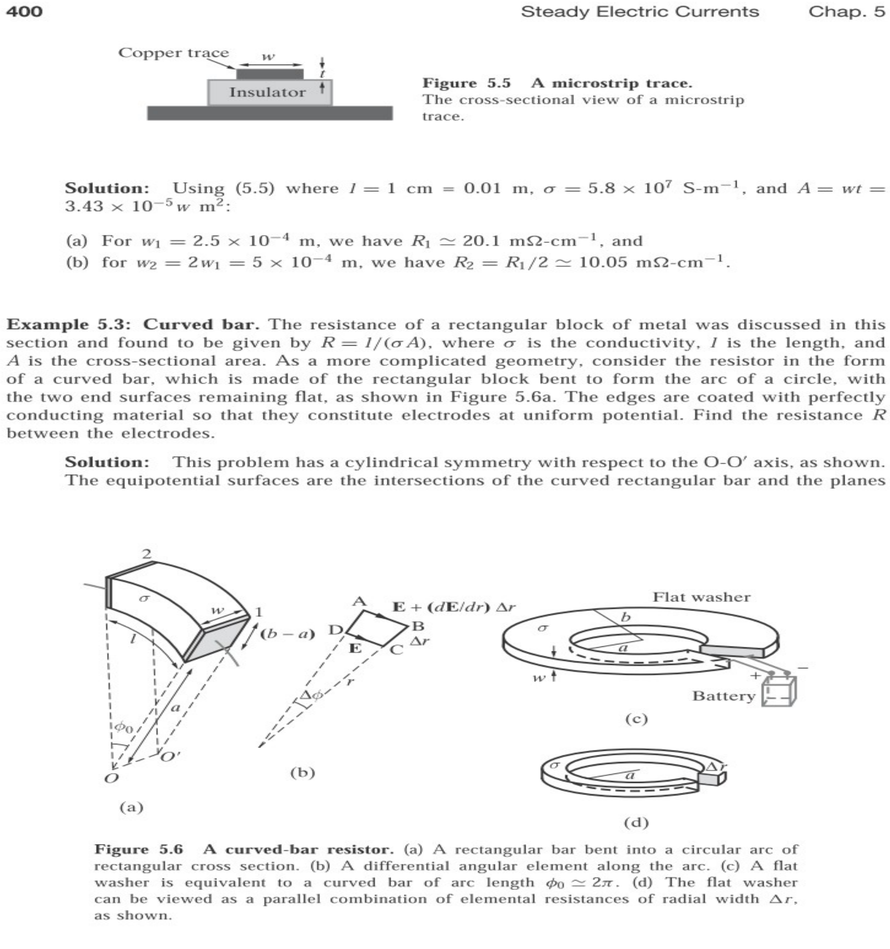

Copper trace
w
t
Insulator

Figure 5.5 A microstrip trace. The cross-sectional view of a microstrip trace.

Solution: Using (5.5) where $l = 1$ cm $= 0.01$ m, $\sigma = 5.8 \times 10^7$ S-m^{-1}, and $A = wt = 3.43 \times 10^{-5} w$ m^2:

(a) For $w_1 = 2.5 \times 10^{-4}$ m, we have $R_1 \simeq 20.1$ mΩ-cm^{-1}, and
(b) for $w_2 = 2w_1 = 5 \times 10^{-4}$ m, we have $R_2 = R_1/2 \simeq 10.05$ mΩ-cm^{-1}.

Example 5.3: Curved bar. The resistance of a rectangular block of metal was discussed in this section and found to be given by $R = l/(\sigma A)$, where σ is the conductivity, l is the length, and A is the cross-sectional area. As a more complicated geometry, consider the resistor in the form of a curved bar, which is made of the rectangular block bent to form the arc of a circle, with the two end surfaces remaining flat, as shown in Figure 5.6a. The edges are coated with perfectly conducting material so that they constitute electrodes at uniform potential. Find the resistance R between the electrodes.

Solution: This problem has a cylindrical symmetry with respect to the O-O′ axis, as shown. The equipotential surfaces are the intersections of the curved rectangular bar and the planes

Figure 5.6 A curved-bar resistor. (a) A rectangular bar bent into a circular arc of rectangular cross section. (b) A differential angular element along the arc. (c) A flat washer is equivalent to a curved bar of arc length $\phi_0 \simeq 2\pi$. (d) The flat washer can be viewed as a parallel combination of elemental resistances of radial width Δr, as shown.

passing through the O-O′ axis. The current flow lines **J** are circular arcs centered on the O-O′ axis (i.e., **J** is in the $\hat{\boldsymbol{\phi}}$ direction). We may suppose from symmetry that the **E** vector is the same at all points inside the bar that are equidistant from the O-O′ axis. However, the electric field must vary with radial distance from the axis (r) across the curved bar, since the inward portions of the end faces are closer to each other than the outward portions are (i.e., **E** must be a function of r, or $\mathbf{E} = \mathbf{J}/\sigma = \hat{\boldsymbol{\phi}} E_\phi(r)$). To determine how the electric field varies with radial distance, we can use the fact that $\oint_C \mathbf{E} \cdot d\mathbf{l} = 0$ (equation (4.20)). Consider a closed contour ABCD, as shown in Figure 5.6b. We have

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = \left(E_\phi + \frac{dE_\phi}{dr} \Delta r \right) (r + \Delta r) \Delta\phi - E_\phi r \Delta\phi = 0$$

where the first term is the contribution from path segment AB and the second term is that from path segment CD. We thus have

$$E_\phi \Delta r + r \frac{dE_\phi}{dr} \Delta r + \frac{dE_\phi}{dr} (\Delta r)^2 = 0 \qquad \rightarrow \qquad \frac{dE_\phi}{E_\phi} = -\frac{dr}{r}$$

where the last term drops out as $\Delta r \to 0$. Therefore

$$\ln E_\phi = -\ln r + K_1 \qquad \rightarrow \qquad E_\phi = \frac{K}{r}$$

where K_1 and K are constants. The functional form of E_ϕ is thus determined. Note that we could alternatively have found $\Phi(r)$ using Laplace's equation, namely $\nabla^2 \Phi = 0$, and subsequently determined **E** from $\mathbf{E} = -\nabla \Phi$. The voltage between the end faces is thus

$$\Phi_{12} = -\int_1^2 \mathbf{E} \cdot d\mathbf{l} = -\int_{\phi_0}^{0} \frac{K}{r} \hat{\boldsymbol{\phi}} \cdot \hat{\boldsymbol{\phi}} r \, d\phi = \int_0^{\phi_0} \frac{K}{r} r \, d\phi = K\phi_0 = \frac{Kl}{a}$$

where we have implicitly assumed the end face at the lower potential to be at $\phi = \phi_0$, so that $\mathbf{E} = \hat{\boldsymbol{\phi}}(K/r)$. The total current between the end faces is given by

$$I = \int_S \mathbf{J} \cdot d\mathbf{s} = \int_S \sigma \mathbf{E} \cdot d\mathbf{s} = \sigma w K \ln \frac{b}{a}$$

Hence the resistance is given by

$$R = \frac{\Phi_{12}}{I} = \frac{l}{\sigma w a \ln(b/a)}$$

Note that the preceding expression for R applies for any arc length l of the curved bar. For example, consider the flat washer shown in Figure 5.6c, with resistance between the edges of a saw-cut through the radius of the washer (i.e., the shaded surface and the one facing it). In this case, and assuming the arc length of the cut region is very small ($\ll 2\pi a$), we simply replace the length l in the previous expression with $l \simeq 2\pi a$. Note from Figure 5.6d that the resistance of the flat washer between the indicated faces can also be thought of as being the parallel combination of elemental flat washer resistances, each of radial thickness Δr.

(B

⑦ 20211004/ Nonlinear Time Series...|
Nonlinear Time Series Analysis with R....pdf
Pág. 111 (122) 4.5
Poincaré Map.

Komorač
funcho = Koromač

20211005

AD-03-20210827/
3-Aberto/
videos-Geral/
videos-por-data/
2020/
08/

bluntly put = colorado sem modos

20200811/
20200813-20/ ⇒ 24 Gb
22-31
25
52
45
97 Gb

09/
04-06 ⇒ 30 Gb
07-09 ⇒ 29 Gb
10-15 ⇒ 30 Gb
16-20 → 21 Gb
22-27 ⇒ 29 Gb
28-30 ⇒

Komorač
(Foeniculum vulgare)

Kush
(erva medicinal)

! -256- - -* ! - ' 8 ! 9 2 ! 3
5 1"
(' $! M !1

Provenzale (1989)). On the other hand, there are deterministic dynamical systems for which time series appear stochastic, for instance when the noise level is high. A further consideration with regard to reliable estimation of the correlation dimension is the length of the embedded time series. Ding *et al.* (1993) showed that the estimated correlation dimension for rather short series falls systematically below that for longer series. The estimation of invariants is a somewhat sensitive matter – indeed, it has been said that *'[it] is often as much an art as it is a science'* (Diks, 1999).

To conclude this discussion of the estimation of invariants, we wish to draw attention to the problem of 'error bars'. For instance, suppose that we have measured a time series and, in the reconstructed space, estimated the MCLE as $\hat{\lambda} = 0.072$. May we believe that we have a chaotic system? This depends on how precise our estimate is. If the standard error were, say, 0.8, it would actually be difficult to be confident about the chaotic nature of the system – hence the need to assign measures of precision to estimates. In an imaginary world of Oz, if we had a measured series with one million observations, then we could divide this series into, say, 100 subseries, each 10 000 observations long, and estimate the MCLE and the correlation dimension in each of them. From this sample, it should be straightforward to derive the estimates of the whole original series and the confidence intervals. In our actual world, something similar can be done, even with only a single available series of reasonable length. Suppose that we have measured a series of length, say, 10 000 observations. Then, by means of so-called ***bootstrap methods*** (see Appendix C), we can generate 100 new series, each 10 000 observations long, since these are in all respects like 100 new measurements of the same process. We have shown that on the basis of such techniques, we can assign the standard deviation and confidence intervals to estimation of the MCLE and of the correlation dimension; the reader is referred to Giannerini *et al.* (2007*b*) and Borovkova *et al.* (2011), respectively. Incidentally, the estimate $\hat{\lambda} = 0.072$ refers to the Rössler attractor, introduced by Rössler (1976) in his studies on chemical kinetics, and its standard error is calculated in Giannerini and Rosa (2002) to be 0.006.

4.5 Poincaré Map

Imagine that the phase space is something real and so is the representative point. Suppose that the representative point follows the trajectory sketched in Figure 4.15a, namely a periodic orbit. Imagine now putting a piece of paper P parallel to the plane (x, z). The representative point intersecting the paper cuts a 'hole' in it. If the orbit has period 1, there is only one 'hole', A (Figure 4.15a). At every period, the representative point passes through the hole A. If the orbit is doubly periodic, there are two 'holes', A and B (Figure 4.15b), so the representative point passes through A and B alternately. The 'piece of paper' is the ***Poincaré section***, and the 'holes' form the ***Poincaré map***, also known as *first-return map*. In Figure 4.15, the 'holes' in the Poincaré section are mathematically points, but are enlarged for the sake of clarity, as will done also in the following. The process at the basis of the Poincaré section construction consists in converting the trajectory of a continuous dynamical system in an N-dimensional phase space into a discrete trajectory (i.e. a map) in an $(N-1)$-dimensional phase space. We can look at this geometrical construction as a 'stroboscopic' representation

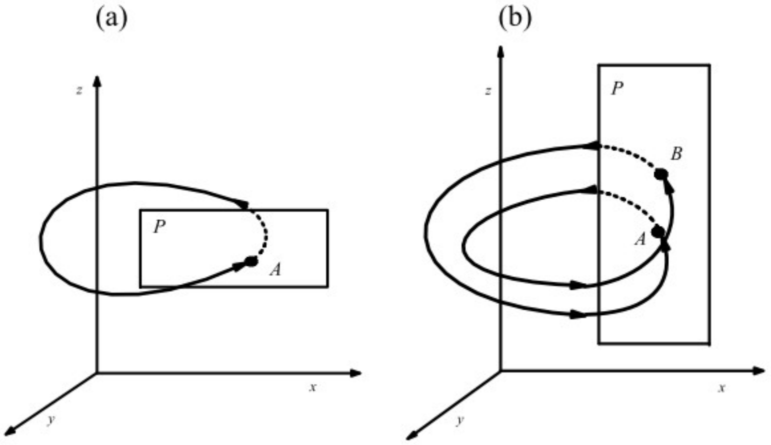

Fig. 4.15 Poincaré map for (a) a period-1 cycle and (b) a period-2 cycle.

of the dynamics of a system. We have seen that for a period-1 cycle, the Poincaré map consists of only one point in the Poincaré section; that is, at every period, the points are plotted upon each other. For a period-2 cycle, the Poincaré map consists of two points; for a period-n cycle, there are n points in the Poincaré section. Note that the intersection points that are considered are only those produced by trajectories in a specified direction, for instance those that cross the plane from the same side.

In the following, we will exploit the Poincaré map to show the dynamics of the *damped driven pendulum*. This pendulum derives from the ideal pendulum already discussed, but now it is more 'realistic', since, in addition to the force of gravity, two further forces acting on it are considered: the damping force and the driving force. The damping force takes into account possible friction with the suspension point and air resistance. The amplitude of oscillation will diminish gradually with time until eventually the system stops. However, an external force acts on the bob, supplying energy to the system, which can continue to oscillate. The interaction between the damping and driving forces makes the system exhibit an extraordinary variety of behaviour, including chaotic behaviour, as shown in Baker and Gollub (1996).

The damping force and the driving force are chosen with very simple forms: the former as being linearly dependent on the velocity of the bobs and the latter as periodic. The the equation of motion of the pendulum, in dimensionless form, is then

$$\frac{d^2x}{dt^2} + \gamma\frac{dx}{dt} + \sin x = F\cos(wt) \tag{4.9}$$

where γ is the *damping constant* (or damping coefficient). In the literature, the quantity $Q = 1/\gamma$, called the *quality factor*, or Q-factor, is also introduced Note that the small-angle approximation, $\sin x \approx x$, is *not* assumed. The parameters F and w are the amplitude and angular frequency of the periodic driving force. Do not confuse

w with ω in eqn (3.2) – the latter is the 'natural' angular frequency, that is, the pendulum frequency without the damping and driving forces. Note that ω does not appear explicitly in eqn (4.9), since it is taken equal to 1, so the driving angular frequency w is a multiple of ω.

Equation (4.9) can be rewritten as a system of three first-order differential equations:

$$\begin{cases} \dfrac{dx}{dt} = y \\ \dfrac{dy}{dt} = -\gamma y - \sin x + F \cos z \\ \dfrac{dz}{dt} = w \end{cases} \tag{4.10}$$

where $z = wt$ is the phase of the driving force. The phase space has three coordinate axes (x, y, z); recall that physically x is the angle $\theta(t)$ and y is the angular velocity $d\theta(t)/dt$ (see Figure 3.1). This system is clearly nonlinear owing to the presence of the terms $\sin x$ and $\cos z$.

For the system (4.10), the solutions depends on γ and ω (in our particular case, ω is fixed to be 1). Let $\gamma > 0$ and $\gamma^2 < 4\omega^2$. The solution is then a sine curve modulated by an exponential; in phase space, the orbits are spirals converging to the asymptotically stable fixed point, the origin $(x(0), y(0)) = (0, 0)$, which is an attractor, called the *focus* (see Figure 6.3b). The basin of attraction is the entire (x, y) plane. Recall that the term *basin of attraction* describes the set of initial points from which the trajectories start and will converge on the attractor at long times. The basin of attraction can be a region of phase space or the whole phase space. There are also trajectories that are not attracted by attractors and are called *separatrices* (singular *separatrix*). They separate regions of phase space inside which the trajectories have different qualitative and quantitative behaviours. In the chaotic regime, the basins of attraction have fractal boundaries.

Mathematically, we can consider also $\gamma < 0$. In this case, the origin $(0, 0)$ becomes an unstable focus, that is, a repeller. If $\gamma^2 < 4\omega^2$, with $\gamma > 0$, the attractor is a *node* and the trajectories approach it monotonically, without spiralling. We are in the presence of *overdamped* motion. See the discussion in Section 5.2. We can solve the system (4.10) by means of Code 4.4 `p_damp_driv_1.R`, which follows the same logic as `pend_id_s_o.R`, but now with three difference equations. As in Baker and Gollub (1996), the parameters γ (`gamma`) and w (`w`) are fixed ($\gamma = \frac{1}{2}$ and $w = \frac{2}{3}$), while F (`F`) assumes different values.

```
#Code 4.4     Damped driven pendulum (p_damp_driv_1.R)
# F=0.9      period-1 cycle
# F=1.07     period-2 cycle
# F=1.15     chaotic regime

parms<- c(1/2,0.9,2/3)   # F = 0.9 or 1.07 or 1.15

tinit<-0
tfin<-100
```

```
step<- 3*pi/100
times<-seq(tinit,tfin,by=step)
funct<- function(t,init,parms){
x<-init[1]
y<-init[2]
z<-init[3]

gamma<-parms[1]
F<-parms[2]
w<-parms[3]
dx<- y
dy<- -gamma*y - sin(x) + F*cos(z)
dz<- w
list(c(dx,dy,dz))
                             }     # end of funct
require(deSolve)
require(scatterplot3d)
cinit<-c(5,4,0)
xyz<-lsoda(cinit,times,funct,parms)
scatterplot3d(xyz[,2],xyz[,3],xyz[,4],type="l",cex.lab=1.5,cex.axis=1.2,lwd=3,
xlab = "x(t)",ylab = "y(t)",zlab = "z(t)",xlim=c(4,16),ylim=c(-2,4),zlim=c(0,50))
windows()
par(mai=c(1.02,1.,0.82,0.42)+0.1)
plot(xyz[,2],xyz[,3],type="p",pch=20,cex=0.8,xlab="x(t)",ylab="y(t)",
cex.lab=1.5,cex.axis=1.2,lwd=3,lty=1,xlim=c(5,15),ylim=c(-2,4))

## second initial point
cinit<-c(13,0,0)
xyz<-lsoda(cinit,times,funct,parms)
points(xyz[,2],xyz[,3],col="black",pch=1,cex=1)
```

The period T_D of one driving cycle is $T_D = 2\pi/w$, and the driving angular frequency w is taken equal to $\frac{2}{3}$. The time step (`step`) is $3\pi/100 \approx 0.09425$, so in 100 iterations the system makes one complete phase orbit.

With $F = 0.9$, the trajectory in the three-dimensional phase space starts from the point $(5, 4, 0)$ and, after a brief transient, the motion becomes periodic and we see the trajectory spiral along the z axis, which represents the flow of time, as shown in Figure 4.16a. For clarity, the three-dimensional trajectory is projected onto the (x, y) plane (as we will also do in the following) in Figure 4.16b (full circles), with a further trajectory (open circles) starting from the point $(13, 0, 0)$ also being shown. After some steps the two trajectories, independently of the initial conditions, converge on a unique closed curve, which is an attractor called the *limit cycle*, more specifically, the period-1 limit cycle. Physically, the pendulum, after the initial transient is extinguished, oscillates at a single frequency, exactly the driving angular frequency w.

Two improvements are introduced in the code. The angle x (i.e. θ) is restricted to the interval $[-\pi, \pi]$, so that if the trajectory exits at the right edge ($> \pi$), it is reported at the left edge, and vice versa. The second improvement concerns the Poincaré map. The `R` package `nonlinearTseries` contains the function **poincareMap** to compute the Poincaré map. So Code 4.4 becomes Code 4.5 `p_damp_driv_2.R`, which is run for $F = 0.9$ and $F = 1.07$.

```
#Code 4.5     Damped driven pendulum (p_damp_driv_2.R)
# Angle x restricted in [-pi, pi], nonperiodic transient discarded.
#Poincare map computed
```

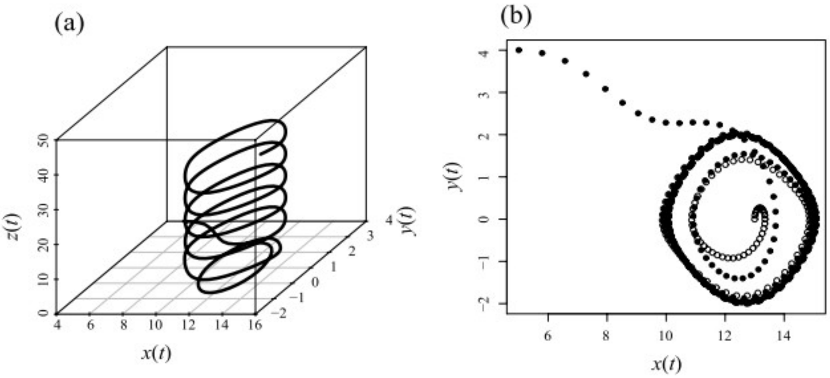

Fig. 4.16 (a) Trajectory in the three-dimensional phase space of the damped driven pendulum with parameters $\gamma = \frac{1}{2}$ and $w = \frac{2}{3}$ and $F = 0.9$ and with initial conditions $(5, 4, 0)$. (b) The same trajectory, but projected onto the (x, y) plane (full circles); a further trajectory starting from $(13, 0, 0)$ is also shown (open circles).

```
# F=0.9      period-1 cycle
# F=1.07     period-2 cycle
# F=1.15     chaotic regime

parms<- c(1/2,0.9,2/3)   # F = 0.9 or 1.07 or 1.15

tinit<-0
tfin<-100
step<- 3*pi/100
nstep<-tfin/step
itrans<-nstep*0.3    # the first 30% steps are discarted
times<-seq(tinit,tfin,by=step)
funct<- function(t,init,parms){
x<-init[1]
y<-init[2]
z<-init[3]

gamma<-parms[1]
F<-parms[2]
w<-parms[3]
dx<- y
dy<- -gamma*y - sin(x) + F*cos(z)
dz<- w
list(c(dx,dy,dz))
                             }     # end of funct
require(deSolve)
cinit<-c(5,4,0)
xyz<-lsoda(cinit,times,funct,parms)
# up to here it is the same as p\_damp\_driv\_2.R

axis.x<- numeric()
axis.y<- numeric()
axis.x<-xyz[itrans:nstep,2]
axis.y<-xyz[itrans:nstep,3]

for(i in 1:length(axis.x)){
```

```
if(axis.x[i]< -pi)    {             #  the trajectory is lesser than -pi ntimes
ntimes<-trunc(axis.x[i]/pi)
axis.x[i]<- axis.x[i]%%pi
if(ntimes%%2!=0)axis.x[i]<- axis.x[i]+pi
                      }
}

for(i in 1:length(axis.x)){
if(axis.x[i]> +pi)    {             #  the trajectory is greater than +pi ntimes
ntimes<-trunc(axis.x[i]/pi)
axis.x[i]<- axis.x[i]%%pi
if(ntimes%%2!=0)axis.x[i]<- axis.x[i]-pi
                      }
}
par(mai=c(1.02,1.,0.82,0.42)+0.1)
plot(axis.x,axis.y,type="p",pch=20,cex=0.3,xlab="x(t)",ylab="y(t)",
cex.lab=1.5,cex.axis=1.2,lwd=3,lty=1,xlim=c(-pi,pi),ylim=c(-pi,pi))

require(nonlinearTseries)
# compute poincare sections
takens=cbind(axis.x,axis.y)
pm=poincareMap(takens=takens,normal.hiperplane.vector=c(0,1),hiperplane.point=c(0,0))
points(pm$pm.pos,col="black",pch=19,cex=2.8,lwd=2.5)
```

To consider only the trajectory on the limit cycle, the nonperiodic transient is discarded. The line `takens=cbind(axis.x,axis.y)` means that the Takens vectors are known – that is, the real phase space, instead of the reconstructed phase space, is known to the user. If necessary, the embedding procedure can be performed with, for example, the instruction

```
pm=poincareMap(time.series=my.series, embedding.dim=3, time.lag=1,
takens=NULL, normal.hiperplane.vector=c(0,0,1), hiperplane.point=c(0,0,0))
```

The parameters `normal.hiperplane.vector=c(0,1)` and `hiperplane.point=c(0,0)` are used to select the Poincaré section. With `pm$pm.pos`, only intersection points produced by trajectories in the positive direction are considered. The result is shown in Figure 4.17 for $F = 0.9$ (a) and 1.07 (b).

Figure 4.17a shows the limit cycle in Figure 4.16b (initial condition $(5, 4, 0)$), but translated in $[-\pi, \pi]$. The result of the computer code is that the limit cycle in Figure 4.16b is inside $[10.06367, 15.07358]$. The right-hand limit is greater than $4 \times \pi$, so $x(t) = 15.07358$ has to be translated as $(15.07358 - 4 \times \pi = 2.507209)$. The Poincaré section is a plane perpendicular to the (x, y) plane and passing through the $y = 0$ axis. The Poincaré map appears as a single point – actually, single points are plotted in the same location on top of each other, at each integer multiple of T_D. When $F = 1.07$ (Figure 4.17b), the period doubles, $T_D = 6\pi$, so the Poincaré map consists of two points plotted within 3π of each other, signifying that the motion of the pendulum is characterized by two different periodic oscillations. Note that the small patch of trajectory on the right is due to the translation in $[-\pi, \pi]$ from left to right of the patch of trajectory less than $-\pi$.

Lastly, let us see what happens when $F = 1.15$. Figure 4.18a shows two trajectories starting from two very close points $A = (5, 5, 0)$ (continuous line) and $B = (5.001, 5.001, 0)$ (dashed line). Here the motion of the pendulum is chaotic: we see that, after a few steps, the trajectories separate themselves at an exponential rate and

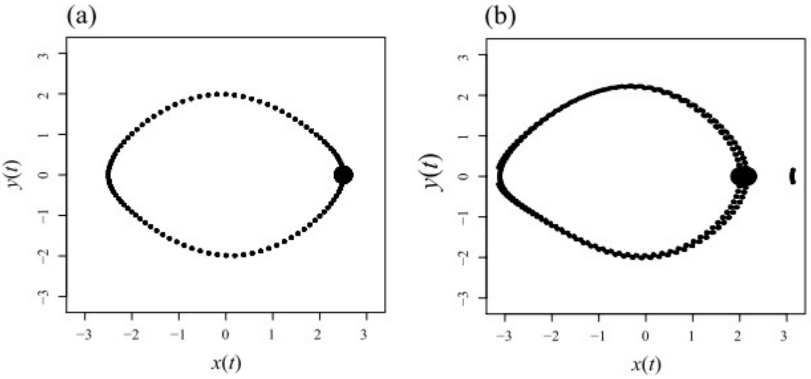

Fig. 4.17 Trajectory in the (x, y) plane of the damped driven pendulum with parameters $\gamma = \frac{1}{2}$ and $w = \frac{2}{3}$ and with initial conditions $(5, 4, 0)$. The trajectory is restricted to the interval $[-\pi, \pi]$ and the transient is discarded. The Poincaré map is also shown as full circles. (a) $F = 0.9$, (b) $F = 1.07$.

we again find a sensitivity to initial conditions, owing to the stretching and folding mechanism. Note that sometimes the single trajectory seems to intersect, although this is not possible for a deterministic system – actually, this impression is due to the projection of the trajectory onto the (x, y) plane. In Figure 4.18b, the trajectory starting from $(5, 5, 0)$ is 10 times longer and is translated in $[-\pi, \pi]$. The figure shows manifestly the 'strange' structure of the attractor, which can be seen in greater detail in the Poincaré map in Figure 4.19. There is no evidence of periodic behaviour, every cycle is different to every other and every point of phase space is visited by the system only once.

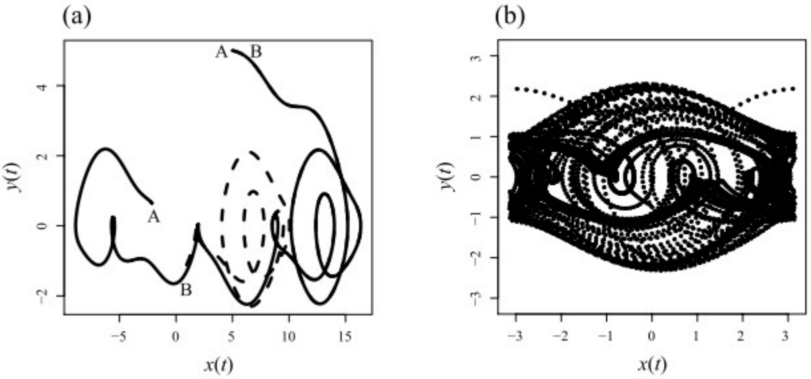

Fig. 4.18 (a) Trajectories in the (x, y) plane of the damped driven pendulum with parameters $\gamma = \frac{1}{2}$, $w = \frac{2}{3}$ and $F = 1.15$ and with initial conditions $A = (5, 5, 0)$ (continuous line) and $B = (5.001, 5.001, 0)$ (dashed line). (b) The trajectory starting from $A = (5, 5, 0)$ is restricted to the interval $[-\pi, \pi]$ and the transient is discarded. Compared with the trajectories in (a), the length of the trajectory here is 10 times greater.

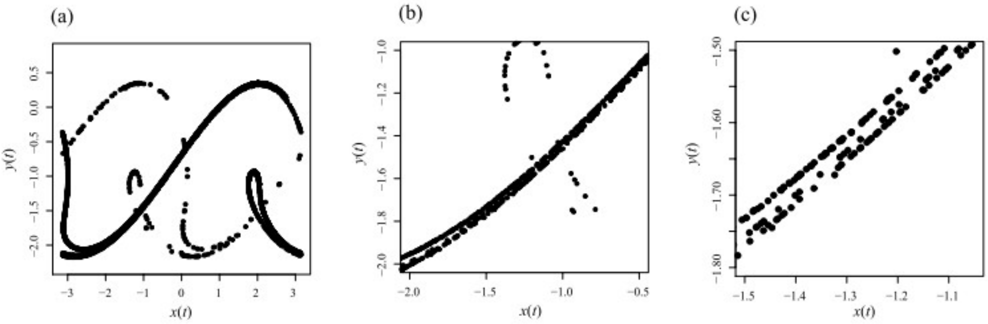

Fig. 4.19 (a) Poincaré map of the damped driven pendulum with parameters $\gamma = \frac{1}{2}$, $w = \frac{2}{3}$ and $F = 1.15$, and with initial conditions $(5, 5, 0)$, as in Figure 4.18b. (b, c) Zooming in on the Poincaré map highlights its fractal structure.

To compute the Poincaré map in this case, a 'stroboscopic picture' of the trajectory in Figure 4.18b is taken at every period of the driving force T_D. Figure 4.19a shows that the Poincaré map is no longer composed either of single points or of random points, but of a set of points with involute structure. If a 'line' on which the points are located is magnified, a group of similar lines emerges (Figure 4.19b, c; note the scales on the axes). Thus, the Poincaré map can be seen as an attractor having a fractal dimension. Looking at Figure 4.17, the cycle in (a), with $F = 0.9$, is symmetric with respect to the x and y axes, while cycle in (b), with $F = 1.07$, is asymmetric with respect to the x axis. If F is increased further, a period doubling occurs up to $F = 1.15$, when the system becomes chaotic. We will see such a route to chaos in Section 4.5.1.

4.5.1 Symmetry breaking

As we did for the logistic map, we construct a bifurcation diagram by computing the angular value $x(t)$ as a function of F, after the transient has been discarded, as shown in Figure 4.20. We have a cascade of bifurcations up to $F \approx 1.10$, even though only the first of these is clearly distinguishable. As F increases, a segment of period 3 becomes visible, around $F \approx 1.12$. For $F \approx 1.15$, a chaotic region begins, and, for $F \approx 1.30$, a period-1 window is born, which doubles for $F \approx 1.45$ and doubles again with a small increase in F. Around $F \approx 1.50$, the chaotic region reappears.

Let us examine in greater detail the first bifurcation occurring when F is very close to 1.066, as shown in Figure 4.21a and already seen in Figure 4.17b. It appears also that the period doubles again for $F \approx 1.08$. For Figure 4.21a, the initial conditions are $(5, 5, 0)$, but if the trajectory starts from $(5, 7, 0)$, we obtain Figure 4.21b. We can see that the arm between $F \approx 1.025$ and $F \approx 1.060$ is the same as that in Figure 4.21a, but as if it has been specularly 'reflected'.

To better understand what has happened, let us examine Figure 4.22a. It is immediately clear that the two innermost cycles are symmetric with respect to the x and y axes, while for the outermost cycles the symmetry with respect to the y axis is no longer apparent, as we have already seen in Figure 4.17. The orbit depicted by circles,

*D 3 3D !.

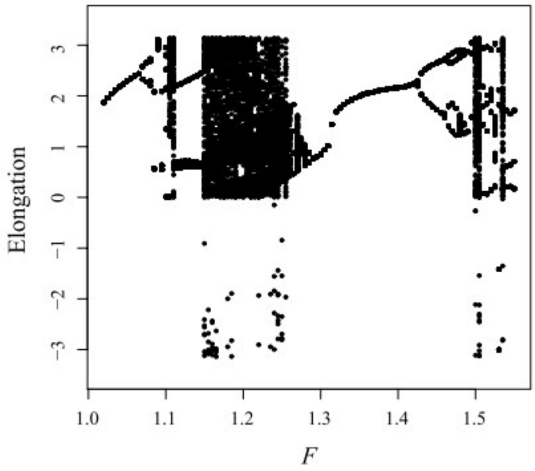

Fig. 4.20 Bifurcation diagram of the damped driven pendulum with parameters $\gamma = \frac{1}{2}$, $w = \frac{2}{3}$ and $F \in [1.020, 1.550]$.

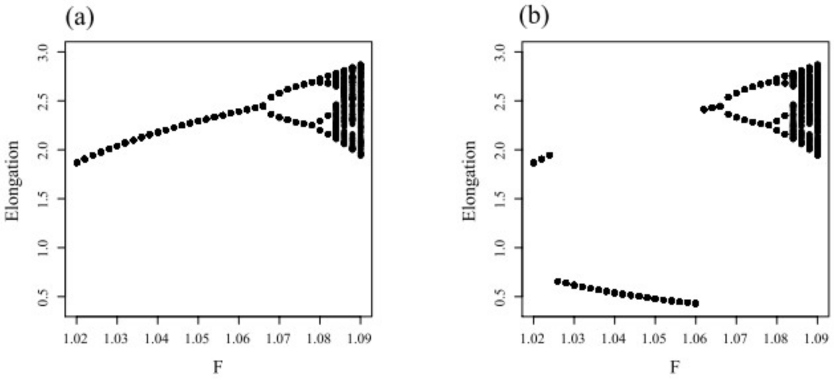

Fig. 4.21 Bifurcation diagram of the damped driven pendulum with parameters $\gamma = \frac{1}{2}$, $w = \frac{2}{3}$ and $F \in [1.020, 1.090]$, with initial conditions $(5, 5, 0)$ (a) and $(5, 7, 0)$ (b).

with initial conditions $(5, 7, 0)$, spends more time in the region with $x(t) < 0$; that is, the region to the left of the vertical line is visited more frequently than the region to the right. Specular behaviour occurs for the orbit depicted by open squares, with initial conditions $(5, 5, 0)$. Up to a value of $F \approx 1.024$, the trajectories converge on symmetric limit cycles, regardless of the initial conditions, but when F goes beyond this critical value, convergence occurs on two possible asymmetric limit cycles, each being the reflected image of the other. In other words, the trajectory may converge on a limit cycle or on its specular reflection, with both solutions being correct from a physical point of view and depending only on the initial conditions but not on the physical parameters. The existence of specular limit cycles explains the appearance of

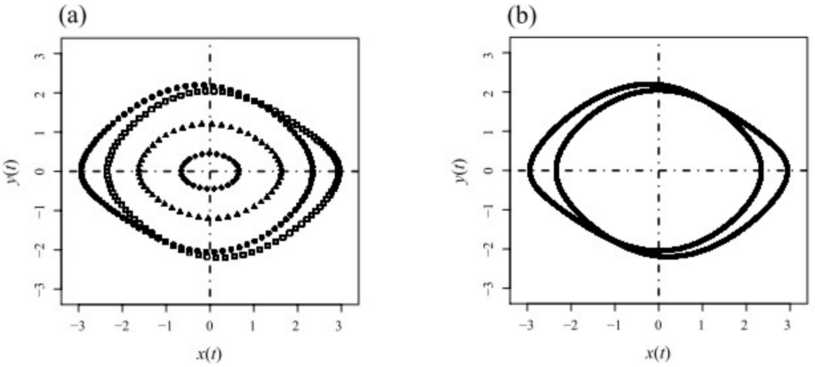

Fig. 4.22 Evolution in phase space of the damped driven pendulum with parameters $\gamma = \frac{1}{2}$ and $w = \frac{2}{3}$. (a) For the two innermost cycles, $F = 0.4$ (diamonds) and $F = 0.7$ (triangles), with initial conditions $(5, 7, 0)$. For the two outermost cycles, $F = 1.025$, with initial conditions $(5, 7, 0)$ (circles) and $(5, 5, 0)$ (open squares). (b) Only the two outermost cycles are plotted, with lines connecting the points.

the bifurcation plots in Figure 4.21, since the points on the specular arms are the coordinates x of the possible specular attractors. For instance, when $F = 1.026$, we find $x(t) = 0.657$ with initial conditions $(5, 7, 0)$ and $x(t) = 1.977$ with initial conditions $(5, 5, 0)$.

Nothing in the equations of the damped driven pendulum can reveal that the region to the right of the vertical line is more preferred than that on the left, or vice versa. This type of phenomenon is called *symmetry breaking*. Such symmetry breaking does not occur if the system evolution is described by linear equations, but is intimately connected with the nonlinear behaviour of the system. Note, however, that symmetry breaking of a single limit cycle leads to the emergence of a new symmetry if we imagine joining together the two specular attractors. We thereby obtain two limit cycles that are symmetric with respect to the vertical line. Figure 4.22b shows the two outermost cycles in Figure 4.22a, with the points now connected with a line, and the impression is of two symmetric limit cycles one inside the other.

The interconnected concepts of symmetry, symmetry breaking, conservation laws and degeneracy play a central role in the physical sciences (see e.g. Brading and Castellani (2013) and references therein). It is common knowledge that symmetry principles (or invariance principles) can be invoked to explain conservation laws. Examples from classical mechanics are energy conservation (related to invariance of the Hamiltonian under time translation) and momentum and angular momentum conservation (related to invariance of the Hamiltonian under spatial translation and rotation, respectively). In quantum mechanics, symmetry principles are even more powerful – recall, for instance, that so-called quantum statistics are related to a new kind of symmetry, the interchange of identical particles. A further consequence of symmetry is degeneracy of energy levels: greater symmetry in the Hamiltonian leads to more degeneracy in the energy spectrum. A further concept connected to symmetry is that of *symmetry breaking*. As David Gross (1996) writes: ‘The secret of nature is symmetry, but much of the texture of the world is due to mechanisms of symmetry breaking.' As we have seen, the laws of mechanics describing the dynamics of the pendulum are symmetric, but when F is greater than a critical value, the states visited by the system are no longer symmetric in phase space.

4.6 Summary

The strangeness of strange attractors is exhibited in the dynamics of trajectories and in the fractal warp woven by the system in (reconstructed) phase space. One of the main features of deterministic chaotic systems is their well-known sensitive dependence on initial conditions. Recall here that historically von Mises, as far back as 1939, recognized sensitivity to initial conditions as the primary source of randomness even in stochastic systems.

The maximal characteristic Lyapunov exponent (MCLE) is an important measure of stability and one of the indicators of the presence of chaos. There are two classes of estimators for the MCLE. The first measures the exponential divergence of trajectories that are initially close through estimation of the Jacobian matrix along the trajectories. The second directly measures the evolution in time of the distance between initially close points. We have focused in this chapter on the performance of the estimator of the MCLE implemented in `tseriesChaos`, which belongs to the second class. For this type of estimator, computation is easy compared with that of Jacobian estimators, and they can also detect whether a system is non-chaotic. The presence of a sharp linear region in the plot of the evolution of the logarithm of the mean distance between points is a strong signal of chaos. For stochastic systems, in contrast, the linear region tends to become flatter with increasing embedding dimension. However, there are situations (the presence of a dynamic noise component and/or series of moderate size) in which these estimators can produce a linear scaling region giving a positive estimate of the MCLE even for non-chaotic systems.

The 'strangeness' of an attractor's geometry is measured by its fractal dimension. There are a number of definitions of fractal dimension, one of which is the correlation dimension D_2. This is estimated via the so-called correlation integral, which is the proportion of all pairs of reconstruction vectors closer than a given distance ϵ in some norm. In the Grassberger–Procaccia correlation integral, if a point is less than a threshold ϵ, it counts as 1; otherwise it counts as 0. In the Gaussian correlation integral, two points that are a small distances apart are weighted (using a Gaussian weighting) more than those separated by large distances. An estimate of the correlation dimension is given by the slope of the log–log plot of the correlation integral versus ϵ. If this estimate is repeated for increasing embedding dimensions m, for chaotic systems, it initially increases with m, but eventually – at least theoretically – it asymptotically reaches a plateau. The value m^* at which this plateau begins is the next integer above D_2. For instance, if the estimated correlation dimension is 2.056, as for the Lorenz attractor, then the minimum number of variables required to describe the system is 3. Other techniques exist to estimate D_2. In Chapter 7, the entire procedure to estimate D_2 for the Lorenz attractor will be repeated, except that, finally, it will be derived through a maximum likelihood procedure (the Takens estimator).

Jules Henri Poincaré makes two appearances this chapter: in *Poincaré's recurrence theorem* and in the *Poincaré section*. He introduced the concept of recurrence, namely that systems, under certain conditions and after a sufficiently long but finite time, will return to a state as close as you like to the initial state. A graphical analysis of the recurrence states of a dynamical system can be executed through recurrence plots (RPs), which can be quantified using recurrence quantification analysis (RQA). In this chapter, we have applied RPs to the logistic map and the Lorenz attractor. It is important to note that methods based on RPs allow trajectories in the m-dimensional phase space of a system to be analysed in terms of a two-dimensional representation of the system's recurrences.

For the study of the *damped driven pendulum*, we have exploited the Poincaré section. This is constructed by converting the trajectory of a continuous dynamical system in an N-dimensional phase space into a discrete trajectory (i.e. a map) in an $(N-1)$-dimensional phase space. We can view this geometric construction as a 'stroboscopic' representation of the dynamics of the system. The damped driven pendulum is the ideal pendulum, introduced in Chapter 3, made more realistic by the inclusion, in addition to the force of gravity, of two further forces acting on it: a damping force and a driving force. The damping force takes into account possible friction with the suspension point and air resistance. The driving force is an external force acting on the bob and supplying energy to the system. A wide variety of behaviours, including chaotic ones, are shown by the damped driven pendulum. Here, we have studied the bifurcation diagram, as for the logistic map, and the particular phenomenon called *symmetry breaking*.

(B'

	(71,43)					
^{1}160	ARQ_TMP_	120	^{1}165	230	^{1}180	^{1}210
180	20200207	230	115	165	140	160
100	e	110	110	395	170	80
100	ARQ-TMP	225	115	Math1	100	450
540	FULL	685	120	math2	590	
105	(49,05)	ARQ BASICS	625		105	
645		02-a-04	ARQ-BAS	750	695	
100		20210603	02-04	ARQ BAS	003-004	
745		(94,43)	20210716	02-04	20210227	
			(84,68)	20220110 (84,63)	(19,63)	

2021-New-20210601/1-Copy-Geral/
Books p/
HD01/1-Copy-Geral-2/Books

---

começar pelo
smprof002@gmail.com (amanhã de manhã)
Já usei 450 Gb.

Books:
Books-020-2021/
20211006/
Equations of Mathematical Physics/
— 10.1007_978-3-030-63253-3.pdf
obs: Differential Geometry.

D -256- - -* ! - ' l/ % % B% $% % 1"
E 4 ! 7 3J 4 ! => ! J ! " ! ! ! 3

-256- - -* ! - ' / B$" / 1"
E => ! ! " 2 ! 3 ! 9 ! ! ! 3
($ %

Up to this stage, the solution to Eq. (6.76) is expected to be represented in a series form of

$$V=\sum_{n=0}^{\infty}\mathcal{P}_n(r)\Theta_n(\theta)=\frac{1}{\sqrt{r}}\sum_{n=0}^{\infty}C_n\mathrm{J}_{n+\frac{1}{2}}(kr)\mathrm{P}_n(\cos\theta) \tag{6.83}$$

with the coefficients C_n to be determined.

Inserting the boundary condition of $V|_{r=1}=1-\cos^2\theta$ into Eq. (6.83) suggests

$$1-\cos^2\theta=\sum_{n=0}^{\infty}C_n\mathrm{J}_{n+\frac{1}{2}}(k)\mathrm{P}_n(\cos\theta). \tag{6.84}$$

For simplicity, we introduce $x=\cos\theta$, and Eq. (6.84) then becomes

$$1-x^2=\sum_{n=0}^{\infty}C_n\mathrm{J}_{n+\frac{1}{2}}(k)\mathrm{P}_n(x). \tag{6.85}$$

For determining C_n in Eq. (6.85), we note that the left side of Eq. (6.85) is a degree 2 polynomial. Thus $C_n=0$ for $n>2$. Besides, since $1-x^2$, the left side of Eq. (6.85), is an even function, $C_1=0$ as $\mathrm{P}_1(x)$ is an odd function. Equation (6.85) is further simplified to be

$$1-x^2=C_0\mathrm{J}_{\frac{1}{2}}(k)\mathrm{P}_0(x)+C_2\mathrm{J}_{\frac{5}{2}}(k)\mathrm{P}_2(x). \tag{6.86}$$

Incorporating it with Eq. (6.46) sets up a linear equation system for C_0 and C_2, and we obtain $C_0=2/(3J_{1/2}(k))$ and $C_2=-2/(3J_{5/2}(k))$. Inserting them back into Eq. (6.83), we finally express the solution to problem (6.78) by

$$\begin{aligned}V(r,\theta)&=\frac{1}{\sqrt{r}}\left(\frac{2}{3J_{\frac{1}{2}}(k)}\cdot\mathrm{J}_{\frac{1}{2}}(kr)\mathrm{P}_0(\cos\theta)-\frac{2}{3J_{\frac{5}{2}}(k)}\cdot\mathrm{J}_{\frac{5}{2}}(kr)\mathrm{P}_2(\cos\theta)\right)\\&=\frac{1}{\sqrt{r}}\left(\frac{2}{3J_{\frac{1}{2}}(k)}\cdot\mathrm{J}_{\frac{1}{2}}(kr)-\frac{1}{3J_{\frac{5}{2}}(k)}\cdot\mathrm{J}_{\frac{5}{2}}(kr)\left(3\cos^2\theta-1\right)\right).\end{aligned} \tag{6.87}$$

6.4.2 *Legendre–Gauss Quadrature*

The concept of the Legendre–Gauss quadrature has been intensively used for evaluating definite integrals. First, we derive the rule of the Legendre–Gauss quadrature. Then the idea is extended to general orthogonal polynomials, and the rule of general Gauss quadrature is presented.

To this end, we start with evaluating $\int_{-1}^{1} f(x)\,\mathrm{d}x$, where $f(x)$ is a polynomial of degree no greater than $2n-1$. Suppose that the roots of $\mathrm{P}_n(x)$ are denoted by $-1 < \zeta_1^{(n)} < \cdots < \zeta_n^{(n)} < 1$, i.e.

$$\mathrm{P}_n(x) = (x - \zeta_1^{(n)}) \cdots (x - \zeta_n^{(n)}). \tag{6.88}$$

Now we construct a function $F(x)$ through the Lagrange interpolation as

$$F(x) = \sum_{k=1}^{n} f(\zeta_k^{(n)}) \cdot \frac{\mathrm{P}_n(x)}{\mathrm{P}_n'(\zeta_k^{(n)})(x - \zeta_k^{(n)})}, \tag{6.89}$$

where P_n' denotes the first derivative of P_n. It can be checked that $F(x)$ is a polynomial of degree $n-1$ and

$$F(\zeta_k^{(n)}) = \lim_{x \to \zeta_k^{(n)}} F(x) = f(\zeta_k^{(n)}) \tag{6.90}$$

for $k = 1, \ldots, n$.

This implies that the function $F - f$ is essentially a polynomial of degree no greater than $2n-1$, and $\zeta_1^{(n)}, \ldots, \zeta_n^{(n)}$ are all its roots. In other words, the polynomial $F - f$ can be factorised by

$$F(x) - f(x) = (x - \zeta_1^{(n)})(x - \zeta_2^{(n)}) \cdots (x - \zeta_n^{(n)}) r(x) = \mathrm{P}_n(x) r(x), \tag{6.91}$$

where $r(x)$ is another polynomial of degree $\leq n-1$, and Eq. (6.88) is employed to derive the second identity.

Replacing $F(x)$ in Eq. (6.91) by Eq. (6.89) and rearranging the result, we get

$$f(x) = \sum_{k=1}^{n} f(\zeta_k^{(n)}) \cdot \frac{\mathrm{P}_n(x)}{\mathrm{P}_n'(\zeta_k^{(n)})(x - \zeta_k^{(n)})} - r(x)\mathrm{P}_n(x). \tag{6.92}$$

Integrating both sides on $(-1, 1)$ gives

$$\int_{-1}^{1} f(x)\,\mathrm{d}x = \sum_{k=1}^{n} \omega_k^{(n)} f(\zeta_k^{(n)}) - \int_{-1}^{1} r(x)\mathrm{P}_n(x)\,\mathrm{d}x, \tag{6.93}$$

where $w_k^{(n)}$ are sometimes termed as the *Christoffel numbers* defined by

$$w_k^{(n)} = \frac{1}{\mathrm{P}_n'(\zeta_k^{(n)})} \cdot \int_{-1}^{1} \frac{\mathrm{P}_n(x)\,\mathrm{d}x}{x - \zeta_k^{(n)}}. \tag{6.94}$$

Table 6.2 Roots of the Legendre polynomials and the Christoffel numbers defined by Eq. (6.94) up to degree 5

P_n	$\zeta_k^{(n)}$	$w_k^{(n)}$
$n=2$	$\pm\frac{1}{\sqrt{3}}$	1
$n=3$	0 $\pm\sqrt{\frac{3}{5}}$	$\frac{8}{9}$ $\frac{5}{9}$
$n=4$	$\pm\sqrt{\frac{15-2\sqrt{30}}{35}}$ $\pm\sqrt{\frac{15+2\sqrt{30}}{35}}$	$\frac{1}{36}\left(18+\sqrt{30}\right)$ $\frac{1}{36}\left(18-\sqrt{30}\right)$
$n=5$	0 $\pm\frac{1}{21}\sqrt{245-14\sqrt{70}}$ $\pm\frac{1}{21}\sqrt{245+14\sqrt{70}}$	$\frac{128}{225}$ $\frac{1}{900}\left(322+13\sqrt{70}\right)$ $\frac{1}{900}\left(322-13\sqrt{70}\right)$

With reference to Eq. (6.66), the integral on the right side of Eq. (6.93) should vanish. Hence, the *rule of the Legendre–Gauss quadrature* is derived to be

$$\int_{-1}^{1} f(x)\,\mathrm{d}x = \sum_{k=1}^{n} w_k^{(n)} f(\zeta_k^{(n)}). \tag{6.95}$$

Several properties shown by the rule of the Legendre–Gauss quadrature make it particularly attractive. Firstly, the integral about a polynomial of degree $2n-1$ is simply calculated by a summation of n terms. Secondly, the Christoffel numbers $w_k^{(n)}$ do not depend on the actual form of $f(x)$, but only on the roots of P_n. Thus, they can be computed beforehand during an "offline" stage as these roots are applied for all cases of $f(x)$. During the "online" stage when the target function $f(x)$ is specified, one can directly add together the n terms on the right side of Eq. (6.95), so as to calculate the integral $\int_{-1}^{1} f(x)\,\mathrm{d}x$.

Table 6.2 lists the roots of the Legendre polynomials and the corresponding Christoffel numbers up to $n=5$. Note that the roots that are symmetric to each other about $x=0$ share a same Christoffel number, and $\sum_{k=1}^{n} w_k^{(n)} = 2$ which is just the length of the domain of integration. Thus, the Christoffel numbers actually correspond to a partition of the domain $(-1, 1)$, and $f(\zeta_k^n)$ can be treated as a representative function value in the kth subinterval.

Equation (6.95) is exact for $f(x)$ being a polynomial of degree $2n-1$. When $f(x)$ is an arbitrary bounded function on $(-1, 1)$, it becomes an approximation. Rigorous analysis can be made to give an upper bound for the approximating error. However, we have seen from Eq. (6.64) that $f(x)$ can be approximated by a linear combination of Legendre polynomials, as shown in demonstrative Fig. 6.4. Therefore, the approximating error is expected to diminish quickly as n is relatively large.

Table 6.3 Calculation of the integral of $\int_{-1}^{1} \mathrm{e}^{2x}\,\mathrm{d}x$ using the rule of the Legendre–Gauss quadrature by Eq. (6.95) and the corresponding relative errors

	Exact	$n = 2$	$n = 3$	$n = 4$	$n = 5$
Value	3.6269	3.4882	3.6223	3.6268	3.6269
Rel. err.	–	0.0382	0.0013	2.26×10^{-5}	1.87×10^{-7}

In Table 6.3, the computed values for $\int_{-1}^{1} \mathrm{e}^{2x}\,\mathrm{d}x$ using the rule of the Legendre–Gauss quadrature are given for various n. It is observed that the approximating accuracy on using Eq. (6.95) is up to six decimal digits as n is as small as 5. From this example, we see that the rule of the Legendre–Gauss quadrature offers an effective algorithm for evaluating definite integrals.

Finally, the derivation for Eq. (6.95) is generalisable to cope with other types of classical orthogonal polynomials. The key difference then is one needs to multiply the corresponding weighted function $\rho(x)$ when integrating Eq. (6.92). And this gives rise to the *general rule of Gauss quadrature*:

$$\int_I \rho(x) f(x)\,\mathrm{d}x = \sum_{k=1}^{n} w_k^{(n)} f(\eta_k^{(n)}), \tag{6.96}$$

where I is the domain of the orthogonal polynomial $p_n(x)$ in consideration and $\eta_k^{(n)}$ denotes the kth root of $p_n(x)$. And the Christoffel numbers then become

$$w_k^{(n)} = \frac{1}{\mathrm{P}_n'(\eta_k^{(n)})} \cdot \int_{-1}^{1} \frac{\rho(x) p_n(x)\,\mathrm{d}x}{x - \eta_k^{(n)}}. \tag{6.97}$$

6.5 The Associated Legendre Functions

In this section, we further consider the solutions to the associated Legendre equation (6.19), that is,

$$(1 - x^2)\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} - 2x\frac{\mathrm{d}y}{\mathrm{d}x} + \left(n(n+1) - \frac{m^2}{1 - x^2}\right) y = 0. \tag{6.98}$$

As discussed in Sect. 6.1, the cases when n and m are both non-negative integers are of particular importance for implementing the method of separation of variables. So the discussion in this section is confined to such situations. Since the present section is aimed to disseminate basic concepts, most of the related proofs are skipped in the main texts.

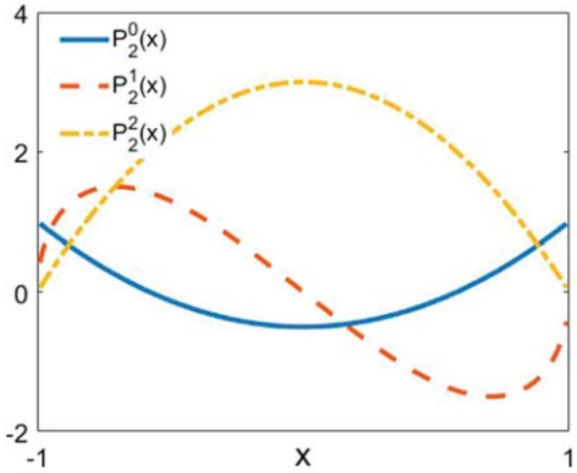

Fig. 6.5 Profiles of the associated Legendre function of the first kind $P_2^m(x)$, with $m = 0$, 1 and 2

We have demonstrated in Sect. 6.2.1 that a fundamental system of solutions to the Legendre equation (6.21) can be formed by the Legendre polynomial $P_n(x)$ given by Eq. (6.40) and the Legendre function of the second kind $Q_n(x)$ given by Eq. (6.44). Actually, the solutions to Eq. (6.98) can be induced from $P_n(x)$ and $Q_n(x)$.

For this purpose, we introduce

$$P_n^m(x) = (-1)^m \left(1 - x^2\right)^{\frac{m}{2}} \frac{d^m}{dx^m} P_n(x), \quad x \in (-1, 1) \tag{6.99}$$

and

$$Q_n^m(x) = (-1)^m \left(1 - x^2\right)^{\frac{m}{2}} \frac{d^m}{dx^m} Q_n(x), \quad x \in (-1, 1). \tag{6.100}$$

It can be checked that Eqs. (6.99) and (6.100) are both solutions to the associated Legendre equation (6.98). Here $P_n^m(x)$ is called the *associated Legendre function of the first kind*, and $Q_n^m(x)$ is called the *associated Legendre function of the second kind*. Here the integer coefficient n is termed as its *degree* and m as its *order*. Note that the term "degree" is inherited from that of Legendre polynomials, although P_n^m is not necessarily a polynomial.

From its definition (6.99), we see that $P_n^m(x)$ is uniformly zero if $n < m$. Thus, we require $n \geq m$ when defining $P_n^m(x)$. Besides, $P_n^0(x)$ is simply the Legendre polynomial of degree n. This coincides with the fact that Eq. (6.98) becomes the Legendre equation when $m = 0$.

In Fig. 6.5, the profiles of $P_2^m(x)$ up to order 2 are shown.

Some of the properties of the associated Legendre functions can be deduced straight forward from that of Legendre functions. For instance, the symmetry of $P_n^m(x)$ about $x = 0$ depends on the value of $n - m$. When $n - m$ is an odd number, $P_n^m(x)$ is an odd function. Moreover, $P_n^m(x)$ are bounded on $x \in (-1, 1)$, but $Q_n^m(x)$ are unbounded on $x \in (-1, 1)$.

Zeros

As for the zeros of $\mathrm{P}_n^m(x)$, we conclude

- $\mathrm{P}_n^m(x)$ has $(n-m)$ simple zeros within the interval $(-1, 1)$.
- The zeros of $\mathrm{P}_n^m(x)$ and $\mathrm{P}_{n+1}^m(x)$ alternate on $(-1, 1)$.
- $\mathrm{P}_n^m(x)$ and $\mathrm{P}_{n+1}^m(x)$ do not have common zeros.
- The zeros of $\mathrm{P}_n^m(x)$ are symmetric about $x = 0$.

Orthogonal systems

If replacing $n(n+1)$ in Eq. (6.98) by λ and rearranging it, we obtain an eigenvalue problem induced by a second-order linear ODE formulated by

$$\frac{\mathrm{d}}{\mathrm{d}x}\left((1-x^2)\frac{\mathrm{d}y}{\mathrm{d}x}\right) - \frac{m^2}{1-x^2}\cdot y + \lambda y = 0, \quad x \in (-1, 1). \tag{6.101}$$

A comparison of Eq. (6.101) with Eq. (2.76) $[k(x) = 1 - x^2 > 0, q = \frac{m^2}{1-x^2} \geq 0$ and $\rho(x) = 1]$ suggests that Eq. (6.101) is a Sturm–Liouville equation. Besides, since $k(\pm 1) = 0$, only natural boundary conditions are needed.

In comparison with Eq. (6.98), the eigenvalues for Eq. (6.101) should be

$$\lambda_n = n(n+1), \quad n \geq m. \tag{6.102}$$

Since only bounded solutions are considered, the corresponding eigenfunctions are

$$y_n(x) = \mathrm{P}_n^m(x), \quad n \geq m. \tag{6.103}$$

Hence, $\{\mathrm{P}_n^m(x)\}_{n=m}^{\infty}$ should form a complete set of orthogonal functions, and the following conclusions are naturally drawn.

For a bounded function $f(x)$ defined on $x \in (-1, 1)$, it can be expressed as a series expansion of the associated Legendre polynomials of order m

$$f(x) = \sum_{n=m}^{\infty} a_n \mathrm{P}_n^m(x), \tag{6.104}$$

where the coefficient a_n is given by

$$a_n = \frac{1}{M_n^m}\int_{-1}^{1} f(x)\mathrm{P}_n^m(x)\,\mathrm{d}x \tag{6.105}$$

with the modulus M_n^m given by

$$M_n^m = \int_{-1}^{1} (\mathrm{P}_n^m(x))^2\,\mathrm{d}x = \frac{2}{2n+1}\frac{(n+m)!}{(n-m)!}. \tag{6.106}$$

It is noted that when an orthogonal system is formed, we have already fixed the order m, and the summation is made over n, with n starting from m.

The aforementioned knowledge should enable us to access the solutions to the Helmholz equation defined within a sphere for more generally imposed boundary conditions.

Exercise

1. Verify that $\mathcal{P}_\alpha(r)$ given by Eq. (6.15) solves Eq. (6.11a).
2. Take the following steps to show that $\mathcal{P}(r) = \frac{1}{\sqrt{r}} J_{\alpha+\frac{1}{2}}(kr)$ is a solution to the Cauchy–Euler equation of Eq. (6.11a), when $k \neq 0$.
   (a) Suppose $\mathcal{P}(r) = F(kr)$, and derive the governing differential equation for $y = F(x)$.
   (b) Assume a series form of $F(x) = \sum_{m=0} a_m x^{m+c}$ and show that $c = \alpha$ if the series solves Eq. (6.11a).
   (c) Carry on to determine all the coefficients a_m in the series, and finalise the proof.
3. Verify that $\Theta(\theta) = F(\cos\theta)$ solves Eq. (6.17b), where $F(\cdot)$ is a solution to the associated Legendre equation (6.19).
4. Solve the Legendre equation [Eq. (6.21)] using the following series form:

$$y = \sum_{k=1}^{\infty} a_k x^k,$$

   and show that the resulting solution is effectively y_2 given by Eq. (6.31b).
5. Verify that Q_0 and Q_n defined by Eqs. (6.43) and (6.44), respectively, solve correspondingly the Legendre equation of degree 0 and of degree n.
6. Use the Rodrigues formula (6.45) to show that $P_n(1) = 1$, $P_n(-1) = (-1)^n$.
7. Based on Eq. (6.46) and recurrence relation (6.47) to express the Legendre polynomials of degrees 6, 7 and 8.
8. Verify the orthogonality condition of Eq. (6.55) for Legendre polynomials with the use of Rodrigues formula (6.45).
9. Write down the orthogonality condition, in comparison with Eq. (6.55), for other types of orthogonal polynomials listed in Table 6.1.
10. Consider the Helmholtz equation $\Delta u + k^2 u = 0$ in $\mathbb{R}^3$.
    (a) Derive the general form of its radial solution, that is, $u = u(\sqrt{x^2+y^2+z^2})$;
    (b) Find one of its fundamental solution following the strategy detailed in Sect. 3.4.
11. Consider the following problem governed by Laplace's equation:

$$\begin{cases} \Delta u = 0, \text{ in } \Omega = \left\{(x, y, z) | \sqrt{x^2+y^2+z^2} < R\right\}; \\ u|_{\partial\Omega} = 1 - z^2. \end{cases}$$

(a) Solve the problem using the method of separation of variables;
(b) Solve the problem using Poisson's formula (3.74) based on the image method for electric potentials in Sect. 3.5.

12. Solve the following problem imposed with Robin boundary condition

$$\begin{cases} \Delta u = 0, \text{ in } \Omega = \left\{(x, y, z) | \sqrt{x^2 + y^2 + z^2} < 1\right\}; \\ \left(u + \dfrac{\partial u}{\partial n}\right)\Big|_{\sqrt{x^2+y^2+z^2}=1} = z. \end{cases}$$

13. Given a spherical shell of

$$\Omega = \left\{\frac{R}{2} < \sqrt{x^2 + y^2 + z^2} < R\right\},$$

use the method of separation of variables to solve

$$\begin{cases} \Delta u + k^2 u = 0, \text{ in } \Omega; \\ u|_{r=\frac{R}{2}} = 0, \ \ u|_{r=R} = z. \end{cases}$$

14. Solve the problem modelling wave propagation in a unit sphere formulated as follows:

$$\begin{cases} u_{tt} - \Delta u = 0, \text{ in } \Omega = \left\{(x, y, z) | r = \sqrt{x^2 + y^2 + z^2} < 1\right\}; \\ u|_{t=0} = r^{\frac{1}{2}}(1 - r^2), \ \ u_t|_{t=0} = 0; \\ u|_{r=1} = 0. \end{cases}$$

Differential Equations, and Mathematical Physics.

- 9811654409.pdf

  Equations and Analytical Tools in Mathematical Physics

  Legendre-Gauss quadrature

  Pág. 225(231).

  Pág. 231(236)

  → Gauss hypergeometric functions

- Equations of Mathematical Physics.pdf

  Pág. 114(125)

  Gauss's theorem in an infinitesimal volume centered at R=0 and enclosed by the surface S,

  Pág. 216(227) 7.7

  Maxwell Equations

  → Gauss's law

  The first two equations are the Gauss's laws, the third equation is the Faraday's law, the

Gauss's theorem in an infinitesimal volume centered at $\mathbf{R}=0$ and enclosed by the surface S,

$$\begin{aligned}&\int d\mathbf{R}\Delta\tfrac{1}{R}=\int d\mathbf{R}div\cdot grad\tfrac{1}{R}=\\&=\int d\mathbf{S}grad\tfrac{1}{R}=-\int d\mathbf{S}\tfrac{\mathbf{R}}{R^3}\rightarrow-4\pi\ \ ,\end{aligned}\tag{5.6}$$

which shows that $\Delta(1/R)=-4\pi\delta(\mathbf{R})$.

Therefore, $1/R$ (or $-1/4\pi R$) is the Green function for the laplacian in $3D$, and we can write the solution of the equation

$$\Delta u=f\tag{5.7}$$

as

$$u(\mathbf{R})=-\frac{1}{4\pi}\int d\mathbf{R}^{'}\frac{1}{|\mathbf{R}-\mathbf{R}^{'}|}f(\mathbf{R}^{'})\ .\tag{5.8}$$

The above equation (which is an inhomogeneous Laplace equation) is the Poisson equation. Its solution vanishes at infinity (as the Green function $1/R$ does). With this boundary condition the solution is unique.

It is worth noting that

$$\Delta\frac{1}{|\mathbf{R}-\mathbf{R}_0|}=-4\pi\delta(\mathbf{R}-\mathbf{R}_0)\tag{5.9}$$

for any $\mathbf{R}_0$; also, we note that derivatives of the form $\partial_x^m\partial_y^n\partial_z^p(1/R)$ are harmonic functions for $\mathbf{R}\neq 0$.

In two dimensions we can check easily that

$$\Delta\ln R=2\pi\delta(\mathbf{R})\ ;\tag{5.10}$$

in one dimension

$$\Delta\,|x|=\frac{d^2}{dx^2}\,|x|=\frac{d}{dx}sgn(x)=2\delta(x)\ ;\tag{5.11}$$

$\ln R$ and $|x|$ are Green functions for the laplacian in $2D$ and $1D$, respectively. It is worth noting that they do not vanish at infinity, as $1/R$ does in $3D$. The Poisson equation can be solved in $2D$ or $1D$ by using these Green functions.

In $2D$ the harmonic functions are related to the analytical functions of complex variable $z = x + iy$ (the complex conjugate is $\overline{z} = x - iy$). An analytic function f is the function which has the derivative df/dz, which means that there exists the derivative along any path approaching the point where the derivative is calculated, and all these derivatives are equal; therefore, $df/dx = df/dy$; with $f = u + iv$ we get

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \; , \; \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x} \; ; \tag{5.12}$$

these are the Cauchy-Riemann relations for analytic functions. We can see immediately that $\Delta u = \Delta v = 0$.

5.2 Green theorem (formulae)

Since

$$v\Delta u = div(vgradu) - gradv \cdot gradu \tag{5.13}$$

we have

$$\begin{gathered} \int_V d\mathbf{R} gradv \cdot gradu + \int_V d\mathbf{R} v\Delta u = \\ = \int_S d\mathbf{S} vgradu = \int_S dSv\frac{\partial u}{\partial n} \; , \end{gathered} \tag{5.14}$$

where the surface S encloses the volume V and n is the coordinate along the normal $\mathbf{n}$ to the surface S (taken along the outward direction). This is the first Green theorem (formula); it applies to spaces with any dimensions.

There exist a few interesting consequences of this theorem. For instance, if $v = 1$ and u is harmonic ($\Delta u = 0$), then

$$\int_S dS\frac{\partial u}{\partial n} = 0 \; . \tag{5.15}$$

Next, if $v = u$ and they are harmonic, then

$$\int_V d\mathbf{R}(gradu)^2 = \int_S dSu\frac{\partial u}{\partial n} \; ; \tag{5.16}$$

(this integral is called the Dirichlet integral); we can see that $u = 0$ in V if $u = 0$ on S and $u = const$ in V if $\partial u/\partial n = 0$ on S.

for the external pressure, the boundary condition $p_0 = \rho\omega^2\Phi(r = R)$ leads to

$$\frac{2}{\pi}\int dkk^2\frac{\varphi_{lm}(k)}{c^2k^2-\omega^2}j_l(kR) + C_{lm}j_l(\omega R/c) = \frac{1}{\rho\omega^2}p_{lm} \ , \qquad (7.92)$$

an equation which determines the coefficients C_{lm}.

7.7 Maxwell equations

E and **H** denote the electric and magnetic field, respectively, ρ and $\mathbf{j}$ are the densities of charge and current, respectively, all real functions of time t and position $\mathbf{r}$. Maxwell equations read

$$\begin{gathered} div\mathbf{E} = 4\pi\rho \ , \ div\mathbf{H} = 0 \ , \\ curl\mathbf{E} = -\tfrac{1}{c}\tfrac{\partial\mathbf{H}}{\partial t} \ , \ curl\mathbf{H} = \tfrac{1}{c}\tfrac{\partial\mathbf{E}}{\partial t} + \tfrac{4\pi}{c}\mathbf{j} \ , \end{gathered} \qquad (7.93)$$

where c is the speed of light. The first two equations are the Gauss's laws, the third equation is the Faraday's law, the fourth equation is the Maxwell-Ampere equation (Michael Faraday (1791-1867), James Clerk Maxwell (1831-1979), Andre-Marie Ampere (1775-1836)). The charges and currents are subjected to the continuity equation

$$\frac{\partial\rho}{\partial t} + div\mathbf{j} = 0 \ . \qquad (7.94)$$

Taking the *div* of the *curl* and using the continuity equation, we can see that only the last two are independent equations.

The fields can be obtained from the protentials **A** (vector potential) and Φ (scalar potential):

$$\mathbf{E} = -\frac{1}{c}\frac{\partial\mathbf{A}}{\partial t} - grad\Phi \ , \ \mathbf{H} = curl\mathbf{A} \ , \qquad (7.95)$$

defined up to the gauge transformation

$$\mathbf{A} = \mathbf{A}^{'} + grad\chi \ , \ \Phi = \Phi^{'} - \frac{1}{c}\frac{\partial\chi}{\partial t} \ , \qquad (7.96)$$

which leaves the fields unchanged; χ is an arbitrary function. We can see that the representation given by equation (7.95) satisfies automatically the homogeneous Maxwell equations (charge- and current-free equations). The same representation used in the inhomogeneous equations leads to

$$\begin{aligned} &div\mathbf{E} = -\tfrac{1}{c}div\tfrac{\partial \mathbf{A}}{\partial t} - \Delta\Phi = 4\pi\rho \ , \\ &curl\mathbf{H} = -\Delta\mathbf{A} + grad \cdot div\mathbf{A} = \\ &= -\tfrac{1}{c^2}\tfrac{\partial^2 \mathbf{A}}{\partial t^2} - \tfrac{1}{c}grad\tfrac{\partial\Phi}{\partial t} + \tfrac{4\pi}{c}\mathbf{j} \ . \end{aligned} \tag{7.97}$$

Imposing the Lorenz ("Lorentz") gauge

$$div\mathbf{A} + \frac{1}{c}\frac{\partial\Phi}{\partial t} = 0 \ , \tag{7.98}$$

these equations become the wave equations with sources

$$\frac{1}{c^2}\frac{\partial^2\mathbf{A}}{\partial t^2} - \Delta\mathbf{A} = \frac{4\pi}{c}\mathbf{j} \ , \ \frac{1}{c^2}\frac{\partial^2\Phi}{\partial t^2} - \Delta\Phi = 4\pi\rho \tag{7.99}$$

for electromagnetic waves propagating with the (light) velocity c (Ludvig Lorenz (1829-1891), Hendrik Lorentz (1853-1928)). Subjected to the Lorenz gauge, the potentials are still defined up to a χ-gauge which satisfies the free wave equation $\partial^2\chi/\partial t^2 - \Delta\chi = 0$. The two wave equations (7.99) are not independent; they are coupled by the Lorenz gauge. It is sufficient to solve one equation, for instance the equation for the vector potential **A** and derive the scalar potential by using the Lorenz gauge. Apart from free-waves solutions of the homogeneous equations, the wave equations (7.99) are solved by the Kirchhoff potentials

$$\begin{aligned} \mathbf{A}(t,\mathbf{r}) &= \tfrac{1}{c}\int d\mathbf{r}'\tfrac{\mathbf{j}(\mathbf{r}', t-\left|\mathbf{r}-\mathbf{r}'\right|/c)}{\left|\mathbf{r}-\mathbf{r}'\right|} \ , \\ \Phi(t,\mathbf{r}) &= \int d\mathbf{r}'\tfrac{\rho(\mathbf{r}', t-\left|\mathbf{r}-\mathbf{r}'\right|/c)}{\left|\mathbf{r}-\mathbf{r}'\right|} \end{aligned} \tag{7.100}$$

(particular solutions), which exhibit retardation (according to the causality). Similar wave equations and solutions hold for the fields **E** and **H**.

The propagation of the electromagnetic waves with the speed of light $c = 3 \times 10^8 m/s$ in vacuum is a fact experimentally established. For small distances, we can limit ourselves to quasi-static (quasi-stationary, non-retarded, sub-wavelength regime), or even static solutions. This amounts to neglecting $|\mathbf{r} - \mathbf{r}'| /c$ in equations (7.100). We get then the well-known Coulomb law and Biot-Savart (Ampere) law (Charles-Augustin de Coulomb (1736-1806), Jean-Baptiste Biot (1774-1862), Felix Savart (1791-1841)). In the absence of charges and currents, we may take $\Phi = 0$ (*e.g.*, by a gauge transformation); then, we have $div\mathbf{A} = 0$ and a pure radiation field. We may preserve Φ and, instead of the Lorenz gauge (7.98), impose the Coulomb gauge $div\mathbf{A} = 0$; then, we have the quasi-static (quasi-stationary solution, non-retarded, non-propagating, Coulomb) solution to the first equation (7.97) (Poisson's equation) and a corresponding wave equation for **A**.

We note that the free-wave equation has plane waves $\mathbf{A} \sim e^{-i\omega t + i\mathbf{kr}}$ as elementary solutions (with vanishing boundary conditions at infinity); $\omega = ck$ is the frequency, **k** is the wavevector, $2\pi/k$ is the wavelength, $2\pi/\omega$ is the period. The condition $div\mathbf{A} = 0$ pertains to transverse waves (**A** perpendicular to **k**); $curl\mathbf{A} = 0$ refers to "longitudinal" waves (**A** parallel with **k**). The Lorenz gauge relates the logitudinal waves to the scalar ones (the latter pertaining to Φ). For finite regions of space, the elementary solutions are given by the eigenfunctions of the laplacian (Helmholtz equation) with specific boundary conditions; then, in general, a separation between transverse and longitudinal waves is not possible.

As it is well known, the Maxwell equations are invariant under Lorentz transformations of coordinates, with constant speed of light. They can be cast into the covariant form, the Lorenz gauge is also covariant. We note also that the Maxwell's equations are invariant under spatial inversion ($\mathbf{E} \rightarrow -\mathbf{E}$, $\mathbf{j} \rightarrow -\mathbf{j}$ are polar vectors), providing **H** is not changed (it is an axial vector); the temporal inversion entails the change $\mathbf{H} \rightarrow -\mathbf{H}$.

The above equations are valid in vacuum. In matter there exists a polarization charge density $\rho_p = -div\mathbf{P}$, where **P** is called polarization; it gives a current density $\mathbf{j_p} = \frac{\partial \mathbf{P}}{\partial t}$; in addition, in matter there exists also a magnetization **M** which gives a current density $\mathbf{j}_m = c \cdot curl\mathbf{M}$;

by convention, $\mathbf{H}$ is replaced in matter by $\mathbf{B}$, which is called magnetic induction. The Maxwell equations in matter read

$$\begin{gathered} div\mathbf{D} = 4\pi\rho \ , \ div\mathbf{B} = 0 \ , \\ curl\mathbf{E} = -\tfrac{1}{c}\tfrac{\partial \mathbf{B}}{\partial t} \ , \ curl\mathbf{H} = \tfrac{1}{c}\tfrac{\partial \mathbf{D}}{\partial t} + \tfrac{4\pi}{c}\mathbf{j} \ , \end{gathered} \tag{7.101}$$

where $\mathbf{D} = \mathbf{E} + 4\pi\mathbf{P}$ and $\mathbf{B} = \mathbf{H} + 4\pi\mathbf{M}$, where $\mathbf{D}$ is called electric displacement, or electric induction; additional relations are necessary between the four unknowns $\mathbf{D}$, $\mathbf{E}$, $\mathbf{B}$, $\mathbf{H}$. At the interface of two bodies the tangential components of the electric field $\mathbf{E}$ and the normal components of the magnetic induction $\mathbf{B}$ must be continuous, while the normal components of the electric induction $\mathbf{D}$ and the tangential components of the magnetic field $\mathbf{H}$ may have a jump $4\pi\rho_s$ and $4\pi\mathbf{j}_s/c$, respectively, where ρ_s and $\mathbf{j}_s$ are the surface charge and current densities, respectively.

We note that $\mathbf{E} = -(1/c)\dot{\mathbf{A}} - grad\Phi$, $\mathbf{B} = curl\mathbf{A}$ and the gauge conditions are still useful in matter; in addition, the representation $\mathbf{A} = (1/c)\mathbf{P}_e + curl\mathbf{P}_m$, $\Phi = -div\mathbf{P}_e$ makes the gauge condition an identity; the potentials $\mathbf{P}_{e,m}$ satisfy the wave equation with sources $\mathbf{P}$ and $\mathbf{M}$, respectively; they are known as Hertz potentials.

7.8 Hansen vectors

For vector equations, like the vector Helmholtz equation

$$\Delta\mathbf{u} + K^2\mathbf{u} = 0 \ , \tag{7.102}$$

it is useful to have three complete sets of independent, orthonormal solutions in separable coordinates. The vector

$$\mathbf{L} = \frac{1}{K} grad\Phi \tag{7.103}$$

satisfies the equation, if Φ satisfies it too. The vector

$$\mathbf{M} = curl(\mathbf{a}\psi) \ , \tag{7.104}$$

where $\mathbf{a} = w\mathbf{e}_1$ is perpendicular to $\mathbf{a}$ and tangential to the surface $q_1 = const$, where (q_1, q_2, q_3) are curvilinear coordinates, $\mathbf{e}_1$ is the coordinates q^m and x^m respectively), the expansions

$$\begin{aligned} \eta^m &= (x^n - q^n)\nabla_n \eta^m\,|_q + \mathcal{O}(2), \\ X \equiv \eta^m \eta_m &= (x^m - q^m)(x^n - q^n)\mathcal{H}_{mn}\,|_q + \mathcal{O}(3) \end{aligned} \tag{19.2}$$

hold and therefore the *regularity condition*

$$X_{,a}X^{,a}/4X \to 1 \tag{19.3}$$

in the limit at the rotation axis is satisfied. Otherwise, if (19.3) is violated, there are conical singularities (rods or struts) on the axis.

The behaviour of the metric near the axis was carefully studied by Carlson and Safko (1980) using pseudo-Cartesian coordinates and their transformation to polar coordinates perpendicular to the symmetry axis. Carot (2000) generalized these results and proved some other implications of axial symmetry. Remarkably, a group of motions G_2 containing an axial symmetry must be Abelian. The 'extended' regularity conditions (Van den Bergh and Wils 1985b) do not necessarily hold even when (19.3) is satisfied. A *cyclic* Killing vector has closed orbits but – contrary to the axisymmetric case – fixed points and a regular axis are not assumed to exist, cp. Szabados (1987). The possible Bianchi types of a G_3 containing a cyclic symmetry are determined in Barnes (2000).

In this chapter we demand that there are two commuting Killing vectors $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$,

$$\xi^a{}_{;b}\eta^b - \eta^a{}_{;b}\xi^b = 0, \quad \xi^a\xi_a < 0, \quad \eta^a\eta_a > 0, \tag{19.4}$$

i.e. they generate an Abelian group G_2I, see §17.1. Carter (1970) has shown that the cases of the greatest physical importance, *asymptotically flat* stationary axisymmetric gravitational fields, necessarily admit an Abelian group G_2, so that for them (19.4) does not impose an additional restriction. If the cyclic Killing vector $\boldsymbol{\eta}$ has a non-null orbit, than any G_2 containing $\boldsymbol{\eta}$ is Abelian (Ernotte 1980).

The Killing vectors $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are *uniquely* determined if we demand that (i) $\boldsymbol{\eta}$ has compact trajectories (in general other subgroups G_1 will have non-compact trajectories), (ii) $\boldsymbol{\eta}$ is normalized by (19.3) and (iii) space-time is asymptotically flat and $\boldsymbol{\xi}$ is normalized so that $F = \xi^a\xi_a \to -1$ holds in the asymptotic region.

19.2 Orthogonal surfaces

The Killing trajectories form two-dimensional orbits T_2; the simple bivector

$$v_{ab} = 2\xi_{[a}\eta_{b]}, \quad v_{ab}v^{ab} < 0, \tag{19.5}$$

is surface-forming (see (19.4) and (6.12)). The surface element orthogonal to the group orbits is spanned by the dual bivector

$$\tilde{v}_{ab} = \tfrac{1}{2}\varepsilon_{abcd}v^{cd} = 2m_{[a}\overline{m}_{b]}, \quad m_a\xi^a = 0 = m_a\eta^a. \tag{19.6}$$

The simple bivector $\tilde{v}_{ab}$ is surface-forming and the metric is said to be *orthogonally transitive* if and only if the condition

$$\begin{aligned} v_{ab}\tilde{v}^{bc}{}_{;c} = 2\xi_{[a}\eta_{b]}(\overline{m}^c m^b{}_{;c} - m^c\overline{m}^b{}_{;c}) = 0 \\ \Leftrightarrow \quad \varepsilon_{abcd}\eta^a\xi^b\xi^{c;d} = 0 = \varepsilon_{abcd}\xi^a\eta^b\eta^{c;d} \end{aligned} \tag{19.7}$$

is satisfied (cp. (6.12)). In general, the Killing vectors of an arbitrary stationary axisymmetric space-time do not obey (19.7), but only solutions obeying (19.7) are explicitly known.

The condition (19.7) can be formulated as a restriction on the algebraic form of the Ricci tensor:

Theorem 19.1 *Stationary axisymmetric fields admit 2-spaces orthogonal to the group orbits if and only if the conditions*

$$\xi^d R_{d[a}\xi_b\eta_{c]} = 0 = \eta^d R_{d[a}\xi_b\eta_{c]} \tag{19.8}$$

are satisfied (Kundt and Trümper 1966).

Proof (see e.g. also Carter 1972): For a Killing vector $\boldsymbol{\xi}$ the identity (8.22),

$$\xi_{a;bc} = R_{abcd}\xi^d, \tag{19.9}$$

holds. From (19.9) (and an analogous relation for $\boldsymbol{\eta}$) we get

$$(\xi_{[a;b}\xi_{c]})^{;c} = \tfrac{2}{3}\xi^d R_{d[a}\xi_{b]}, \quad (\eta_{[a;b}\eta_{c]})^{;c} = \tfrac{2}{3}\eta^d R_{d[a}\eta_{b]}. \tag{19.10}$$

With the aid of these equations, together with the cyclic symmetry relation of the curvature tensor and the commutativity of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$, we obtain

$$(\xi_{[a;b}\xi_c\eta_{d]})^{;d} = -\tfrac{1}{2}\xi^d R_{d[a}\xi_b\eta_{c]}, \quad (\eta_{[a;b}\eta_c\xi_{d]})^{;d} = -\tfrac{1}{2}\eta^d R_{d[a}\eta_b\xi_{c]}. \tag{19.11}$$

Using the fact that a completely antisymmetric tensor is proportional to the ε-tensor, we convert (19.11) into the equivalent dual equations

$$\begin{aligned} (\varepsilon^{abcd}\xi_{a;b}\xi_c\eta_d)^{;e} &= -2\varepsilon^{abce}\xi^d R_{da}\xi_b\eta_c, \\ (\varepsilon^{abcd}\eta_{a;b}\eta_c\xi_d)^{;e} &= -2\varepsilon^{abce}\eta^d R_{da}\eta_b\xi_c. \end{aligned} \tag{19.12}$$

From these equations we conclude that the conditions (19.8) are a necessary consequence of the original criterion (19.7). Conversely, if the conditions (19.8) are satisfied, the twist scalars $\varepsilon_{abcd}\eta^a\xi^b\xi^{c;d}$ and $\varepsilon_{abcd}\xi^a\eta^b\eta^{c;d}$

are constants. Both twist scalars vanish on the rotation axis ($\boldsymbol{\eta} = 0$). Hence they vanish in a connected domain which intersects the rotation axis if the conditions (19.8) hold in that domain. Then the geometry admits 2-spaces orthogonal to the orbits of the Abelian group of motions G_2.

For vacuum fields ($R_{ab} = 0$), the existence of orthogonal 2-spaces was first demonstrated by Papapetrou (1966) in his pioneering work on this subject.

Conditions (19.8) hold for a wide class of energy-momentum tensors. For example, decomposition of the metric into orthogonal 2-spaces is possible for perfect fluid solutions, and for Einstein–Maxwell fields, provided that the four-velocity of the fluid, and the electromagnetic 4-current density vector, respectively, satisfy the so-called circularity condition

$$u_{[a}\xi_b\eta_{c]} = 0 = j_{[a}\xi_b\eta_{c]}, \tag{19.13}$$

i.e. the trajectories of $\boldsymbol{u}$ and $\boldsymbol{j}$ lie in the surfaces of transitivity of the group G_2. In this terminology, conditions (19.8) mean that $R^a_b\xi^b$ and $R^a_b\eta^b$ are circular vector fields. The circularity condition (19.13) is a natural generalization of the condition (18.40) (four-velocity and current parallel to the Killing vector) to the case of stationary axisymmetric fields. Obviously, for perfect fluids the condition $u_{[a}\xi_b\eta_{c]} = 0$ implies (19.8).

Now we show that $j_{[a}\xi_b\eta_{c]} = 0$ for Einstein–Maxwell fields also implies (19.8). The Maxwell equations

$$F^{*ab}{}_{;b} = j^a, \tag{19.14}$$

and the assumption that the Lie derivatives of F^*_{ab} with respect to both the commuting Killing vectors $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ vanish, lead to

$$(F^*_{ab}\xi^a\eta^b)_{;d} = \mathrm{i}\varepsilon_{abcd}\xi^a\eta^b j^c = 0. \tag{19.15}$$

$F^*_{ab}\xi^a\eta^b$ is constant, and it is zero on the axis of rotation. Hence,

$$F^*_{ab}\xi^a\eta^b = 0 = \xi_{[a}\eta_b F^*_{cd]}. \tag{19.16}$$

(In flat space-time electrodynamics this relation says that the azimuthal components of the electric and magnetic field vectors are zero.) Relation (19.16) in turn implies (19.8).

If in an Einstein–Maxwell geometry, outside the sources, two Killing vectors $\boldsymbol{\xi}$, $\boldsymbol{\eta}$ satisfy the conditions (19.4), (19.5), (19.7), then the Maxwell field shares the space-time symmetry: $\mathcal{L}_{\boldsymbol{\xi}}F_{ab} = 0 = \mathcal{L}_{\boldsymbol{\eta}}F_{ab}$ (Michalski and Wainwright 1975).

Gourgoulhon and Bonazzola (1993) considered the more general case when the conditions (19.8) do not hold, e.g. for toroidal magnetic fields. The field equations can be written in an entirely two-dimensional covariant form.

19.3 The metric and the projection formalism

Theorem 19.1 is of fundamental importance in the construction of exact solutions of the Einstein equations. If we adapt the coordinate system to the Killing vectors, the line element of a stationary axisymmetric field admitting 2-spaces orthogonal to the Killing vectors $\boldsymbol{\xi} = \partial_t$ and $\boldsymbol{\eta} = \partial_\varphi$ can be written in the form (Lewis 1932, Papapetrou 1966)

$$\mathrm{d}s^2 = \mathrm{e}^{-2U}(\gamma_{MN}\mathrm{d}x^M\mathrm{d}x^N + W^2\mathrm{d}\varphi^2) - \mathrm{e}^{2U}(\mathrm{d}t + A\mathrm{d}\varphi)^2, \qquad (19.17)$$

where the metric functions U, γ_{MN} , W, A depend only on the coordinates $x^M = (x^1, x^2)$ which label the points on the 2-surfaces S_2 orthogonal to the orbits. The form (19.17) is preserved under the transformations

$$x^{M'} = x^{M'}(x^N), \quad t' = at + b\varphi, \quad \varphi' = c\varphi + dt, \qquad (19.18)$$

($a, \dots, d$ constants). The functions W, U, A behave like scalars under $x^{M'} = x^{M'}(x^N)$. The metric (19.17) has block-diagonal form and exhibits the reflection symmetry $(t, \varphi) \to (-t, -\varphi)$.

Equation (19.17) is the specialization of (18.8) (including the conformal transformation (18.23)) to the stationary axisymmetric fields obeying the conditions (19.7). In the coordinate system (19.17), the scalar products of the Killing vectors are

$$\begin{aligned} \xi^a\xi_a &= -\mathrm{e}^{2U}, & \eta^a\eta_a &= \mathrm{e}^{-2U}W^2 - \mathrm{e}^{2U}A^2, \\ \xi^a\eta_a &= -\mathrm{e}^{2U}A, & 2\xi_{[a}\eta_{b]}\xi^a\eta^b &= W^2. \end{aligned} \qquad (19.19)$$

If a linear combination of the two Killing vectors leads to the relation $A = \pm W\mathrm{e}^{-2U}$, the space-time admits a null Killing vector. This special case will be considered in Chapter 24. Vacuum solutions with this property belong to the class $S = S(A)$ (§20.4), provided that W is not constant.

In general, $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are not hypersurface-orthogonal. However, it can easily be verified (Bardeen 1970) that the timelike vector $\boldsymbol{\zeta} = \boldsymbol{\xi} - \xi^a\eta_a(\eta^b\eta_b)^{-1}\boldsymbol{\eta}$ is orthogonal to the hypersurfaces $t = \text{const}$. In general, it is not a Killing vector.

Without loss of generality we can introduce *isotropic coordinates* in V_2,

$$\gamma_{MN} = \mathrm{e}^{2k}\delta_{MN}. \qquad (19.20)$$

Moreover, if W is a non-constant potential function, it can be transformed to $W = x^1$ with the aid of a coordinate transformation $\zeta' = f(\zeta)$, $\sqrt{2}\zeta = x^1 + \mathrm{i}x^2$, preserving the isotropic coordinate condition. The case $W = \text{const}$ is not of interest here because it is not compatible with the regularity condition (19.3); vacuum metrics of this type are plane waves (24.35) with $F = 0$, $f = f(\zeta)$.

The isotropic coordinates with $W = x^1 = \rho$ are called *Weyl's canonical coordinates* ($x^1 = \rho$, $x^2 = z$; $\sqrt{2}\zeta = \rho + \mathrm{i}z$). In these coordinates, the space-time metric (19.17) is given by

$$\mathrm{d}s^2 = \mathrm{e}^{-2U}[\mathrm{e}^{2k}(\mathrm{d}\rho^2 + \mathrm{d}z^2) + \rho^2\mathrm{d}\varphi^2] - \mathrm{e}^{2U}(\mathrm{d}t + A\mathrm{d}\varphi)^2. \tag{19.21}$$

For stationary axisymmetric fields, one can develop a projection formalism (Geroch 1972) similar to that for stationary fields given in §18.1. We introduce the notation

$$\lambda_{AB} \equiv \xi_A{}^a\xi_{Ba}, \quad A, B = 1, 2, \tag{19.22}$$

for the scalar products of the Killing vectors $\boldsymbol{\xi}_1 = \boldsymbol{\xi}$, $\boldsymbol{\xi}_2 = \boldsymbol{\eta}$, and identify the projection tensor

$$H_{ab} \equiv g_{ab} + W^{-2}\lambda^{AB}\xi_{Aa}\xi_{Bb}, \quad W^2 \equiv -\tfrac{1}{2}\lambda^{AB}\lambda_{AB}, \tag{19.23}$$

with the metric tensor on a differentiable manifold Σ_2. Capital Latin indices are raised and lowered with the aid of the alternating symbols (3.66), as in (3.67). Covariant derivatives (denoted by D_a) and the Riemann curvature tensor on Σ_2 are defined in analogy to (18.3) and (18.5) ($h_{ab} \to H_{ab}$).

If 2-spaces V_2 orthogonal to the group orbits exist, Σ_2 can be identified with these 2-spaces and the two scalars $C_A = \varepsilon^{MN}\varepsilon^{abcd}\xi_{Ma}\xi_{Nb}\xi_{Ac;d}$ vanish. Then the Ricci tensor R_{ab} in V_4 can be written in terms of tensors and metric operations on V_2 as follows,

$$\overset{2}{R}_{ab} = (2W)^{-2}\lambda^{AB}{}_{,a}\lambda_{AB,b} + W^{-1}D_aW_{,b} + H_a^cH_b^dR_{cd}, \tag{19.24}$$

$$D^a(W^{-1}\lambda_{AB,a}) = \tfrac{1}{2}W^{-3}\lambda_{AB}H^{ab}\lambda^{CD}{}_{,a}\lambda_{CD,b} - 2W^{-1}R_{ab}\xi_A^a\xi_B^b, \tag{19.25}$$

$$H_a^c\xi_A^dR_{cd} = 0. \tag{19.26}$$

The contracted Bianchi identities (2.82) in the projection formalism are given by Whelan and Romano (1999) and read (for $C_A = 0$)

$$D_a(W\xi_A^bH^{ac}G_{bc}) = 0, \quad D_b(WH_a^cH^{bd}G_{cd}) = \tfrac{1}{2}W\xi_A^c\xi_B^dG_{cd}(W^{-2}\lambda^{AB})_{,a}, \tag{19.27}$$

where G_{ab} is the Einstein tensor.

Kitamura (1978) gave an invariant characterization of the stationary axisymmetric metric (19.17) in terms of a tetrad and its first covariant derivatives. This characterization is independent of the field equations.

(/ %% / ' 9 " > ! ! ! 3 !L 3 = !L "
(' ' '$B 7! ! A!
(' $ '$ 7! ! A! +
(' $ ' 7! ! A!
($ % $$$ 7! ! A! 1 " 1!
($ $ $$ %' % 1 L A! ! ! !
($ B ' 9 7! ! " "!,, > !
($ % B $ % % 7! "!,, 5 > !

-256- - -* ! - ' ! ! 4 ! => ! J
1 ! 9 3J ! " ! ! ! 3 9 1"
(/' 7! ! !

in $L^2(\mathbb{R}^d)$,

$$\mathcal{U}_h^{\varepsilon}(t,t_0)g_{z_0}^{\Gamma_0,\varepsilon} = \mathrm{e}^{\frac{i}{\varepsilon}S(t,t_0,z_0)}\left(g_{\Phi^{t,t_0}(z_0)}^{\Gamma(t,t_0,z_0),\varepsilon} + \sum_{j=1}^{N_0}\varepsilon^{\frac{j}{2}}\mathrm{WP}^{\varepsilon}_{\Phi^{t,t_0}(z_0)}(\varphi_j(t))\right) + O(\varepsilon^{\frac{N_0+1}{2}}) \tag{2.7}$$

with

$$\Gamma(t,t_0,z_0) = (C(t,t_0,z_0) + D(t,t_0,z_0)\Gamma_0)(A(t,t_0,z_0) + B(t,t_0,z_0)\Gamma_0)^{-1} \tag{2.8}$$

and

$$c_{\Gamma(t,t_0,z_0)} = c_{\Gamma_0}\det{}^{-\frac{1}{2}}(A(t,t_0,z_0) + B(t,t_0,z_0)\Gamma_0),$$

where the complex square root is continuous in time.

Note that we have $\Gamma(t_0,t_0,z_0) = \Gamma_0$ in the statement above. Actually, explicit information is obtained on the profiles $(\varphi_j(t))_{j\in\mathbb{N}}$, in particular about $\varphi_1(t)$ (see [5, Section 4.1.2] or [8, Proposition 2.3]). As we shall see in Section 5, this result can be a starting point for proving the Herman–Kluk approximation of Theorem 2.1.

The result above also holds for general wave packets as in (2.6) with profiles that are not necessarily Gaussians (see [5, Section 4.1.2]). Moreover, also the norm can be generalized to

$$\|f\|_{\Sigma_k^{\varepsilon}} = \sup_{|\alpha|+|\beta|\leqslant k}\|x^{\alpha}(\varepsilon\partial_x)^{\beta}f\|_{L^2}, \quad k\in\mathbb{N}.$$

Example 2.3. Both Theorem 2.1 and 2.2 are exact when the Hamiltonian h is quadratic. For the Hamiltonians $h_{\pm}(z) = p \pm kq$ associated with the toy model (see (1.3)), the classical flow is linear

$$\Phi_{\pm}^{t,t_0}(z_0) =: (q_{\pm}(t), p_{\pm}(t)) = z_0 + (t-t_0)(1,\mp k), \quad z_0 = (q_0,p_0),$$

the width of the Gaussian is constant, $\Gamma_{\pm}(t,t_0,z_0) = \Gamma_0$, and the actions only depend on q_0 and are given by

$$S_{\pm}(t,t_0,q_0) = \mp kq_0(t-t_0) \mp \frac{k}{2}(t-t_0)^2.$$

3 Herman–Kluk formula in the adiabatic setting

We need adaptations for generalizing the scalar ideas to systems. The first one replaces the scalar Herman–Kluk prefactor $u_0(t,t_0,z_0)$ by a vector-valued one $\vec{U}(t,t_0,z_0)$ that is expanded in a basis of eigenvectors of $H(t,z)$, taken along the classical trajectory $\Phi_{h_j}^{t,t_0}(z)$ of the corresponding eigenvalues (denoted here by $h_j(t,z)$). This is done by parallel transport, and it is sufficient for an order ε approximation as long as the system is gapped.

3.1 Parallel transport

The following construction generalizes [4, Proposition 1.9], which was inspired by the work of G. Hagedorn, see [13, Proposition 3.1]. The details are given in [8]. In the sequel, we denote the complementary orthogonal projector by

$$\Pi^{\perp}(t,z) = \mathbb{I}_{\mathbb{C}^N} - \Pi(t,z)$$

and assume that

$$H(t,z) = h(t,z)\Pi(t,z) + h^{\perp}(t,z)\Pi^{\perp}(t,z) \tag{3.1}$$

with the second eigenvalue given by $h^{\perp}(t,z) = \mathrm{tr}(H(t,z)) - h(t,z)$. We introduce the auxiliary matrices

$$\begin{aligned}
\Omega(t,z) &= -\frac{1}{2}\big(h(t,z) - h^{\perp}(t,z)\big)\Pi(t,z)\{\Pi,\Pi\}(t,z)\Pi(t,z),\\
K(t,z) &= \Pi^{\perp}(t,z)\big(\partial_t \Pi(t,z) + \{h,\Pi\}(t,z)\big)\Pi(t,z),\\
\Theta(t,z) &= i\Omega(t,z) + i(K - K^*)(t,z),
\end{aligned}$$

that are smooth and satisfy some algebraic properties. In particular, Ω is skew-symmetric and Θ is self-adjoint: $\Omega = -\Omega^*$ and $\Theta = \Theta^*$.

Proposition 3.1 ([8]). *Let $H(t,z)$ be a smooth Hamiltonian that satisfies (1.6) and has a smooth spectral decomposition (3.1). We assume that both eigenvalues are of subquadratic growth as well. We consider $\vec{V}_0 \in \mathcal{C}_0^{\infty}(\mathbb{R}^{2d},\mathbb{C}^N)$ and $z_0 \in \mathbb{R}^{2d}$ such that there exists a neighborhood U of z_0 where for all $z \in U$*

$$\vec{V}_0(z) = \Pi(t_0,z)\vec{V}_0(z) \quad \textit{and} \quad \|\vec{V}_0(z)\|_{\mathbb{C}^N} = 1.$$

Then there exists a smooth normalized vector-valued function $\vec{V}(t,t_0)$ satisfying

$$\vec{V}(t,t_0,z) = \Pi(t,z)\vec{V}(t,t_0,z) \quad \textit{for all } z \in \Phi_h^{t,t_0}(U),$$

and such that for all $t \in \mathbb{R}$ and $z \in \Phi_h^{t,t_0}(U)$,

$$\partial_t \vec{V}(t,t_0,z) + \{h,\vec{V}\}(t,t_0,z) = -i\Theta(t,z)\vec{V}(t,t_0,z), \quad \vec{V}(t_0,t_0,z) = \vec{V}_0(z). \tag{3.2}$$

We note that the results of this Proposition are valid as long as smooth eigenprojectors and eigenvalues do exist. It does not require an explicit adiabatic situation and we will use that observation in Section 4. In the case of Schrödinger systems, one refers to (3.2) as parallel transport because the vectors $\partial_t\vec{V}(t) + \{h,\vec{V}\}(t)$ belong to the range of $\Pi^{\perp}(t)$ at any time of the evolution.

Example 3.2. For the toy model $H_{k,\theta}(x,\xi)$, the auxiliary matrices are

$$\Omega_{\pm} = 0, \quad K_{\pm} = \frac{i\theta}{4}\begin{pmatrix} 1 & \pm\,\mathrm{e}^{i\theta x} \\ \mp\,\mathrm{e}^{-i\theta x} & -1 \end{pmatrix}, \quad \Theta_{\pm} = \frac{\theta}{2}\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Initiating the parallel transport equation by the eigenvectors given in (1.3), we obtain

$$\vec{V}_{\pm}(t,t_0,q) = \frac{1}{\sqrt{2}}\begin{pmatrix} \mathrm{e}^{i\frac{\theta}{2}(t-t_0)}\,\mathrm{e}^{i\theta(q-t+t_0)} \\ \pm\,\mathrm{e}^{-i\frac{\theta}{2}(t-t_0)} \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} \mathrm{e}^{-\frac{i}{2}\theta(t-t_0)}\,\mathrm{e}^{i\theta q} \\ \pm\,\mathrm{e}^{-\frac{i}{2}\theta(t-t_0)} \end{pmatrix}.$$

Note that they do not depend on p.

3.2 Herman–Kluk approximation in an adiabatic setting

In the context of the preceding section, and in presence of an eigenvalue gap,

$$\exists\delta > 0: \quad \mathrm{dist}(h(t,z), h^{\perp}(t,z)\}) \geqslant \delta \quad \text{for all } (t,z),$$

it is well known that one has *adiabatic decoupling*: If the initial data are scalar multiples of an eigenvector associated with the eigenvalue h, then the solution keeps this property to leading order in ε. For an approximation to higher order in ε, one has to consider perturbations of the eigenprojector, the so-called *superadiabatic projectors* (see [2, 27] and the new edition of [5] that should appear soon). Adiabatic theory implies a Herman–Kluk approximation of the propagator that we state next.

Theorem 3.3 ([8]). *In the situation of Proposition 3.1, we assume the existence of an eigenvalue gap for the eigenvalue h of H and consider initial data of the form*

$$\psi_0^{\varepsilon} = \widehat{\vec{V}}_0 v_0^{\varepsilon} + O(\varepsilon) \quad \text{in } L^2(\mathbb{R}^d),$$

where $\vec{V}_0$ is a smooth eigenvector and $(v_0^{\varepsilon})_{\varepsilon>0}$ a family of functions uniformly bounded in $L^2(\mathbb{R}^d,\mathbb{C})$. Then, in $L^2(\mathbb{R}^d)$,

$$\mathcal{U}_H^{\varepsilon}(t,t_0)\psi_0^{\varepsilon} = (2\pi\varepsilon)^{-d}\int_{\mathbb{R}^{2d}} \langle g_z^{\varepsilon}, v_0^{\varepsilon}\rangle \vec{U}(t,t_0,z)\,\mathrm{e}^{\frac{i}{\varepsilon}S(t,t_0,z)}\, g^{\varepsilon}_{\Phi_h^{t,t_0}(z)}\, dz + O(\varepsilon)$$

with

$$\vec{U}(t,t_0,z) = u_0(t,t_0,z)\vec{V}(t,t_0,\Phi_h^{t,t_0}(z)),$$

where $u_0(t,t_0,z)$ is given by (2.5) and the eigenvector $\vec{V}(t,t_0,z)$ by (3.2).

We describe in Section 5 a proof of this result that crucially uses the approximate evolution of a Gaussian state, that is [24, Section 3],

$$\begin{aligned}&\mathcal{U}_H^{\varepsilon}(t,t_0)(\widehat{\vec{V}}_0 g_z^{\Gamma,\varepsilon})\\ &\quad = \mathrm{e}^{\frac{i}{\varepsilon}S(t,t_0,z)}\,\widehat{\vec{V}(t,t_0)} g^{\Gamma(t,t_0,z),\varepsilon}_{\Phi_h^{t,t_0}(z)}\left(1 + \sqrt{\varepsilon}\vec{a}(t)\cdot\frac{(x-q(t))}{\sqrt{\varepsilon}}\right) + O(\varepsilon)\end{aligned} \tag{3.3}$$

with $\vec{a}(t) = \vec{a}(t,t_0,z)$ a smooth and bounded map. Note that using superadiabatic projectors [24], one can push the asymptotics further and exhibit $O(\varepsilon)$ contributions that will have components on the other mode. Here again, a formula similar to (3.3) can be proved for general wave packets [5].

3.2 The semigroup $e^{-t\sqrt{H_\alpha}}$

By the spectral theorem,

$$e^{-t\sqrt{H_\alpha}} = \sum_{m\in\mathbb{Z}} \oplus\big(e^{-t\sqrt{h_m}} \otimes \mathrm{id}\big)\Pi_m. \tag{3.3}$$

Denote by $p_m(r, r', t)$ the integral kernel of $e^{-t\sqrt{h_m}}$ in $L^2(\mathbb{R}_+, r\,dr)$. From (3.1) it follows that

$$p_m(r, r', t) = \frac{1}{\sqrt{rr'}} e^{-t\sqrt{L_m}}(r, r').$$

On the other hand, in view of (3.2) we get

$$\begin{aligned}\big(e^{-t\sqrt{L_m}} g\big)(r) &= \big(W_m^{-1} e^{-t\sqrt{p}} W_m g\big)(r)\\ &= \frac{1}{2}\int_0^\infty \sqrt{rr'} \int_0^\infty e^{-t\sqrt{p}} J_{|m+\alpha|}(r\sqrt{p}) J_{|m+\alpha|}(r'\sqrt{p})\, dp\, g(r')\, dr'.\end{aligned}$$

Hence

$$\begin{aligned}p_m(r, r', t) &= \frac{1}{2}\int_0^\infty e^{-t\sqrt{p}} J_{|m+\alpha|}(r\sqrt{p}) J_{|m+\alpha|}(r'\sqrt{p})\, dp\\ &= \int_0^\infty e^{-tp} J_{|m+\alpha|}(rp) J_{|m+\alpha|}(r'p)\, p\, dp.\end{aligned}$$

In order to simplify the notation we will use in the sequel the shorthand

$$z := \frac{r^2}{t^2}. \tag{3.4}$$

From [4, equation 4.14.(16)] we then get the explicit expression for p_m on the diagonal

$$\begin{aligned}p_m(r, r, t) &= \int_0^\infty e^{-tp} J^2_{|m+\alpha|}(rp)\, p\, dp \qquad (3.5)\\ &= \frac{4^\nu}{\pi}(2\nu+1)\, t^{-2}\Big(\frac{r}{t}\Big)^{2\nu} \frac{\Gamma^2(\nu+\frac{1}{2})}{\Gamma(2\nu+1)} F\Big(\nu+\frac{1}{2}, \nu+\frac{3}{2}, 2\nu+1; -4z\Big),\end{aligned}$$

where

$$\nu = |m+\alpha| \tag{3.6}$$

and $F(a, b, c; w)$ denotes the Gauss hypergeometric series, see for example [1, equation 15.1.1]. Using its integral representation

$$F(a, b, c; w) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 s^{b-1}(1-s)^{c-b-1}(1-sw)^{-a}\, ds,$$

where $\mathrm{Re}\, c > \mathrm{Re}\, b > 0$, see [1, equation 15.3.1], in combination with the transformation formula [1, equation 15.3.5]

$$F(a, b, c; w) = (1-w)^{-b} F\Big(b, c-a, c; \frac{w}{w-1}\Big),$$

we find that

$$
\begin{aligned}
&p_m(r,r,t)\\
&\quad= \frac{2\nu+1}{\pi t^2}(4z)^\nu \frac{\Gamma^2(\nu+\frac{1}{2})}{\Gamma(2\nu+1)}(1+4z)^{-\nu-\frac{3}{2}} F\left(\nu+\frac{3}{2}, \nu+\frac{1}{2}, 2\nu+1; \frac{4z}{4z+1}\right)\\
&\quad= \frac{2\nu+1}{\pi t^2}(4z)^\nu (1+4z)^{-\nu-\frac{3}{2}} \int_0^1 s^{\nu-\frac{1}{2}}(1-s)^{\nu-\frac{1}{2}}\left(1-\frac{4zs}{4z+1}\right)^{-\nu-\frac{3}{2}} ds\\
&\quad= \frac{2\nu+1}{\pi t^2}(4z)^\nu \int_0^1 s^{\nu-\frac{1}{2}}(1-s)^{\nu-\frac{1}{2}}(1+4z(1-s))^{-\nu-\frac{3}{2}}\, ds\\
&\quad= \frac{2\nu+1}{\pi t^2}(4z)^\nu \int_0^1 s^{\nu-\frac{1}{2}}(1-s)^{\nu-\frac{1}{2}}(1+4zs)^{-\nu-\frac{3}{2}}\, ds. \qquad (3.7)
\end{aligned}
$$

We have:

Lemma 3.1. *There exists* $\varepsilon_0 > 0$ *such that for every* $\varepsilon \in (0, \varepsilon_0)$ *there is* $C_\varepsilon > 0$ *for which the upper bound*

$$
(1+r)^{-\frac{3}{2}-\varepsilon}(1+r')^{-\frac{3}{2}-\varepsilon} p_m(r,r',t) \le C_\varepsilon \frac{t^{-2-2\kappa}}{(|m+\alpha|+1)^{1+\varepsilon}}
$$

holds for all $m \in \mathbb{Z}$ *and all* $t \ge 1$.

Proof. Put

$$
\varepsilon_0 := \min\left\{ |m+\alpha| - \frac{3}{2} : m \in \mathbb{Z} \wedge |m+\alpha| > \frac{3}{2} \right\}. \qquad (3.8)
$$

Clearly we have $\varepsilon_0 > 0$. Let $\varepsilon < \varepsilon_0$ and put $\rho = \frac{3}{2} + \varepsilon$. Keeping in mind the notation (3.6), we will distinguish two cases depending on the value of ν.

Assume first that $\nu > \frac{3}{2}$. In view of (3.4) and (3.7),

$$
\begin{aligned}
\frac{p_m(r,r,t)}{(1+r^2)^\rho} &= \frac{4\nu+2}{\pi t^3}\frac{r(4z)^{\nu-\frac{1}{2}}}{(1+r^2)^\rho} \int_0^1 s^{\nu-\frac{1}{2}}(1-s)^{\nu-\frac{1}{2}}(1+4zs)^{-\nu-\frac{3}{2}}\, ds\\
&\lesssim t^{-3}\nu (4z)^{\nu-\rho} \int_0^1 s^{\nu-\frac{1}{2}}(1-s)^{\nu-\frac{1}{2}}(1+4zs)^{-\nu-\frac{3}{2}}\, ds\\
&= t^{-3}\nu (4z)^{\nu-\rho} \int_0^1 s^{\nu-\frac{1}{2}}(1+4zs)^{\rho-\nu}(1-s)^{\nu-\frac{1}{2}}(1+4zs)^{-\rho-\frac{3}{2}}\, ds,
\end{aligned}
$$

where we have used the fact that $z \le r^2$ by assumption. From (3.8) it follows that $\rho < \nu$. Hence

$$
\begin{aligned}
(1+r^2)^{-\rho} p_m(r,r,t) &\lesssim t^{-3}\nu (4z)^{\nu-\rho} \int_0^1 s^{\nu-\frac{1}{2}}(4zs)^{\rho-\nu}(1-s)^{\nu-\frac{1}{2}}\, ds\\
&= t^{-3}\nu \int_0^1 s^{\rho-\frac{1}{2}}(1-s)^{\nu-\frac{1}{2}}\, ds = t^{-3}\nu B\left(\rho+\frac{1}{2}, \nu+\frac{1}{2}\right)\\
&= t^{-3}\nu \frac{\Gamma(\rho+\frac{1}{2})\Gamma(\nu+\frac{1}{2})}{\Gamma(\nu+\rho+1)},
\end{aligned}
$$

Blow-ups for the Horn–Kapranov parametrization of the classical discriminant

Evgeny Mikhalkin, Vitaly Stepanenko and Avgust Tsikh

Dedicated to the 70th birthday of our colleague and friend Ari Laptev

The Horn–Kapranov parametrizations describe the singular sets of hypergeometric functions in several variables. These parametrizations are inverses of logarithmic Gauss maps for A-discriminants. In this paper we demonstrate that, despite the multivalued nature of the indicated parametrizations, their blow-ups properties are the same as for single-valued meromorphic mappings. As an application, a new proof of factorization identities for the classical discriminant is given.

1 Introduction

Given a meromorphic mapping $f : X \to Y$ of complex analytic sets (spaces), it is reasonable to treat it as an analytic subset of $X \times Y$ that is the closure $\overline{G}_f$ of the graph of f. Fibers in $\overline{G}_f$ over uncertainty points can be regarded as result of “blow-up” or “shrinkage” procedures [13]. The most transparent scheme for such procedures can be seen for mappings that are inverses of the logarithmic Gauss map [9]. In the modern theory of hypergeometric functions [8], these inverses are known as Horn–Kapranov parametrizations [2, 14, 15].

Our goal in this paper is to study the Horn–Kapranov parametrization and to give a new proof of factorization identities for truncations of the classical discriminant. Such identities were obtained in [7, Chapter 10], using a complicated technique of the theory of A-determinants. In our approach, instead of meromorphic mappings we consider multivalued mappings with abelian monodromy groups.

The classical discriminant relates to a polynomial in one variable

$$f(y) = a_0 + a_1 y + \cdots + a_n y^n. \tag{1.1}$$

Keywords: Discriminant, Horn–Kapranov parametrization, Newton polytope, logarithmic Gauss map, blow-up

2020 Mathematics Subject Classification: Primary 32H04; secondary 32S25

It is an irreducible polynomial

$$\Delta_n = \Delta_n(a_0, a_1, \ldots, a_n)$$

with integer coefficients that vanishes if and only if f has multiple roots. Discriminants play an important role in mathematics, which is demonstrated in the fundamental monographs [7, 16]. The interest in the topic of discriminant identities considered in this article is inspired by studying the universal algebraic function [1, 6, 11, 12, 16] and by research in singularity theory [17], algebraic (tropical) geometry and mathematical physics [3, 5].

It is well known [4, 7] that the Newton polytope $\mathcal{N}(\Delta_n) \subset \mathbb{R}^{n+1}$ of the indicated discriminant is combinatorially equivalent to an $(n-1)$-dimensional cube. Since such a cube has 2^{n-1} vertices, it is natural to encode these vertices by all possible subsets of the set $\{1, \ldots, n-1\}$. The polytope $\mathcal{N}(\Delta_n)$ has $n-1$ hyperfaces $\{h_k^0\}$ lying in the coordinate hyperplanes $\{t_k = 0\}$, $k = 1, \ldots, n-1$ (it is assumed that the coordinates $t_0, t_1, \ldots, t_n$ are chosen in the ambient space $\mathbb{R}^{n+1}$). Each hyperface h_k^0 has 2^{n-2} vertices, which are encoded by subsets $I \subset \{1, \ldots, n-1\}$ that do not contain k. Let h_k denote the face opposite to h_k^0, whose vertices are encoded by subsets I containing k (an explicit equation for a hyperplane containing k is given by formula (2.7) in Section 2).

Our main objects of interest in this paper are truncations of the discriminant Δ_n to the faces of its Newton polytope. Recall that the *truncation* of a polynomial Δ to a face h of its Newton polytope $\mathcal{N}(\Delta)$ is the sum of all monomials from Δ whose exponents belong to h. This truncation is denoted by $\Delta|_h$.

The truncations $\Delta_n|_{h_k^0}$ to the coordinate hyperfaces h_k^0 are irreducible polynomials. In the recent paper [10] we wrote a sketch of proof of the following formula:

$$\Delta_n\Big|_{h_k} = a_k^2\, \Delta_k(a_0, a_1, \ldots, a_k)\Delta_{n-k}(a_k, a_{k+1}, \ldots, a_n), \tag{1.2}$$

where Δ_k, Δ_{n-k} *are the discriminants of polynomials*

$$a_0 + a_1 y + \cdots + a_k y^k, \quad a_k + a_{k+1} y + \cdots + a_n y^{n-k}$$

of degrees k and $n-k$, respectively.

Note that for $k = 1$ the first discriminant on the right-hand side of (1.2) is identically equal to 1: $\Delta_1(a_0, a_1) \equiv 1$. The same is true for the second factor $\Delta_1(a_{n-1}, a_n)$ when $k = n-1$. Thus, the truncations of Δ_n to the hyperfaces h_1 and h_{n-1} coincide, up to the monomials a_1^2 and a_{n-1}^2, respectively, with the discriminants Δ_{n-1} of polynomials of degree $n-1$.

As an example consider the case $n = 3$ when f is a cubic polynomial

$$f = a_0 + a_1 y + a_2 y^2 + a_3 y^3$$

and the discriminant Δ_3 is equal to

$$\Delta_3(a_0, a_1, a_2, a_3) = -27a_0^2 a_3^2 - 4a_1^3 a_3 - 4a_0 a_2^3 + 18a_0 a_1 a_2 a_3 + a_1^2 a_2^2.$$

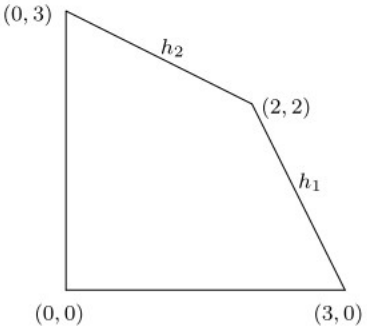

Figure 1.1. The Newton polytope for $\mathcal{N}(\Delta_3^{\mathrm{red}})$.

For a better understanding the geometry of the Newton polytope $\mathcal{N}(\Delta_3) \subset \mathbb{R}^4$ it is convenient to consider its projection onto the subspace $\mathbb{R}^2$ of variables t_1, t_2. For this it is enough to pass from the polynomial f to its reduction f^{red} setting $a_0 = a_3 = 1$. The corresponding reduced discriminant is the following:

$$\Delta_3^{\mathrm{red}}(a_1, a_2) = -27 - 4a_1^3 - 4a_2^3 + 18a_1a_2 + a_1^2a_2^2.$$

For its Newton polytope see Figure 1.1.

The truncations to h_1 and h_2 of the reduced discriminant are

$$\begin{aligned} a_1^2a_2^2 - 4a_1^3 &= a_1^2 \cdot 1 \cdot (a_2^2 - 4a_1) = a_1^2 \cdot \Delta_1(1, a_1) \cdot \Delta_2(a_1, a_2, 1), \\ a_1^2a_2^2 - 4a_2^3 &= a_2^2 \cdot (a_1^2 - 4a_2) \cdot 1 = a_2^2 \cdot \Delta_2(1, a_1, a_2)\Delta_1(a_2, 1), \end{aligned}$$

respectively. For the original discriminant these expressions become as

$$\begin{aligned} \Delta_3\big|_{h_1} &= a_1^2a_2^2 - 4a_1^3a_3 = a_1^2 \cdot \Delta_1(a_0, a_1) \cdot \Delta_2(a_1, a_2, a_3), \\ \Delta_3\big|_{h_2} &= a_1^2a_2^2 - 4a_0a_2^3 = a_2^2 \cdot \Delta_2(a_0, a_1, a_2)\Delta_1(a_2, a_3), \end{aligned}$$

which corresponds to (1.2).

In the present paper we prove more general formulas concerning the truncations of Δ_n to faces

$$h_K := h_{k_1} \cap \cdots \cap h_{k_p}$$

of codimension p, obtained by intersection of p non-coordinate hyperfaces. The multi-index $K = \{k_1, \dots, k_p\}$ determines a partition of the tuple $\{0, 1, \dots, n\}$ into $p + 1$ subtuples (segments)

$$K_i = \{k_i, k_i + 1, \dots, k_{i+1}\}, \quad i = 0, 1, \dots, p,$$

assuming that $k_0 = 0$ and $k_{p+1} = n$. Let $l_i := k_{i+1} - k_i$ denote the length of K_i and

$$f_{K_i} := a_{k_i} + a_{k_i+1}y + \cdots + a_{k_{i+1}}y^{l_i}.$$

Theorem 1.1. *In previous notations, the truncation of* Δ_n *to* h_K *admits the following factorization:*

$$\Delta_n\big|_{h_K} = a_K^2 \prod_{i=0}^{p} \Delta_{l_i}(f_{K_i}), \tag{1.3}$$

where $a_K^2 = a_{k_1}^2 \dots a_{k_p}^2$, *and* Δ_{l_i} *are the discriminants of polynomials of degrees* l_i.

In order to prove this theorem, we start with a similar assertion for the extremal part of Δ_n (Lemma 2.1). Thus we conclude that the Newton polytopes of the polynomials in both sides of (1.3) coincide. Then we prove that the zero set of the truncation $a_K^{-2}\Delta_n|_{h_K}$ contains the union of zero sets of discriminants $\Delta_{l_i}(f_{K_i})$ (Lemma 4.2). To this end, we use the Horn–Kapranov parametrizations of the zero sets of these discriminants. The final conclusion is based on general facts of the intersection theory.

2 Extremal discriminant and factorization of its truncations

Given a discriminant Δ_n, the sum of its monomials with exponents ranging over the vertex set of the polytope $\mathcal{N}(\Delta_n)$ we denote by $\widehat{\Delta}_n$ and call it the *extremal discriminant*.

Lemma 2.1. *The truncation of* $\widehat{\Delta}_n$ *to the face* h_K *admits the factorization*

$$\widehat{\Delta}_n\big|_{h_K} = a_K^2 \prod_{i=0}^{p} \widehat{\Delta}_{l_i}(f_{K_i}), \tag{2.1}$$

where $a_K^2 = a_{k_1}^2 \dots a_{k_p}^2$, *and* Δ_{l_i} *are the discriminants of polynomials of degrees* l_i.

For the proof of this statement we use the formulas for monomials of $\widehat{\Delta}_n$, i.e., for extremal monomials of Δ_n.These formulas were derived by Gelfand, Kapranov and Zelevinsky (see [7, Chapter 12, Theorems 2.2 and 2.3]) with the help of complicated algebraic technique of the theory of A-determinants. Note that another proof of this theorem was given by Batyrev [4] using only the classical Newton method for finding branches of an algebraic function. This assertion about monomials is as follows.

Claim. *The Newton polytope of the discriminant of polynomial* (1.1) *is combinatorially equivalent to an* $(n-1)$*-dimensional cube; it contains* 2^{n-1} *vertices, which are in bijective correspondence with all possible subsets* $I \subset \{1, 2, \dots, n-1\}$. *The vertex* $\mathrm{v}(I)$ *corresponding to a subset* $I = \{i_1 < i_2 < \dots < i_s\}$ *has the coordinates*

$$\begin{aligned} &\mathrm{v}_0 = i_1 - 1, \quad \mathrm{v}_n = n - i_s - 1, \\ &\mathrm{v}_{i_\nu} = i_{\nu+1} - i_{\nu-1} && \text{for } i_\nu \in I, \\ &\mathrm{v}_i = 0 && \text{for } i \notin I \cup \{0, n\}. \end{aligned} \tag{2.2}$$

(BB

fourth equation is the
Maxwell-Ampere equation.

- Exact Solutions of Einstein's
Field Equations - Cambridge
Univers. pdf
Pág. 293 (325)
19.2 Orthogonal Surfaces
Pág. 298 (330)
19.4 The field equations for
stationary axisymmetric
Einstein-Maxwell fields
→ Gaussian curvature of the
2-space $\hat{V}_2$ with metric γ_{MN}.
Pág. 424 (456)
Gaussian curvature K of the
2-surface $2\,d\zeta\,d\bar{\zeta}/P^2$ could be
normalized to $K=0, \pm 1$.

(B

Pág. 425 (457)
Gaussian curvature K
Pág. 472 (504)
Gaussian curvature
Obs: Look: 31.3
the Ricci tensor components.
Pág. 523 (555)
Gaussian curvatures of the corresponding 2-spaces.
Pág. 525 (557)
34.13 Complex invariance transformations
→ Gaussian curvature and are related by the complex transformation
Pág. 572 (604) -
the Gauss and Codazzi equations in the form...
Pág. 583 (615) →

(B

37.3 Some theorems on local isometric embedding

Gauss-Codazzi-Ricci equations.
etc

obs: Gaussian Curvature Pág 101

— Partial Differential Equations...pdf
Pág. 94 (107)
Gaussian is constant.
Pág. 285 (298)
Gauss Hypergeometric series
Pág. 315 (328)
Logarithmic Gauss map.
Pág. 323 (336) Figure 1.2.
Truncated factorization for
Δ_4^{red}.

— Symmetries and Conservation
Laws for Differential Equations
of Mathematical Ph..pdf

-256- - -* ! - ' 3 ! " A! ?!
4 ! => ! ! ! ! 1"
($ /$!1 $ A! ?! 7! 9

-256- - -* ! - '
9 7 ! 9 3 ! ! I2I 1 ! I 1"

-256- - -* ! - ' 9 ! ! ! ! "
< ! ! 4 ! => ! 1"
(B % :31 31

$$\lim_{x\to 0} k_\nu(x) = \frac{\Gamma(-\nu)}{2^{2\nu+1}\Gamma(1+\nu)}, \qquad \nu < 0, \qquad -\nu \notin \mathbb{N}, \tag{1.27}$$

$$\lim_{x\to 0} x^\alpha k_0(x) = 0, \qquad \alpha > 0, \qquad \lim_{x\to 0} \frac{1}{\ln x} k_0(x) = -1, \tag{1.28}$$

$$\lim_{x\to 0} x^{2\nu+1} \frac{dk_\nu(x)}{dx} = -1, \qquad \nu > -1. \tag{1.29}$$

We will use notations

$$\mathbf{j}_\gamma(x,\xi) = \prod_{i=1}^{n} j_{\frac{\gamma_i-1}{2}}(x_i\xi_i) \tag{1.30}$$

and

$$\mathbf{i}_\gamma(x,\xi) = \prod_{i=1}^{n} i_{\frac{\gamma_i-1}{2}}(x_i\xi_i), \tag{1.31}$$

where $\gamma = (\gamma_1, ..., \gamma_n)$, $\gamma_1 > 0, ..., \gamma_n > 0$.

Information about the Bessel functions is taken from [591].

The Struve function is a solution $y(x)$ of the nonhomogeneous Bessel differential equation:

$$x^2 \frac{d^2y}{dx^2} + x\frac{dy}{dx} + \left(x^2 - \alpha^2\right) y = \frac{4\left(\frac{x}{2}\right)^{\alpha+1}}{\sqrt{\pi}\,\Gamma\left(\alpha + \frac{1}{2}\right)}.$$

Struve functions, denoted as $\mathbf{H}_\alpha(x)$, have the power series form

$$\mathbf{H}_\alpha(x) = \sum_{m=0}^{\infty} \frac{(-1)^m}{\Gamma\left(m+\frac{3}{2}\right)\Gamma\left(m+\alpha+\frac{3}{2}\right)} \left(\frac{x}{2}\right)^{2m+\alpha+1}. \tag{1.32}$$

Another definition of the Struve function, for values of α satisfying $\operatorname{Re}\alpha > -\frac{1}{2}$, is possible using an integral representation:

$$\mathbf{H}_\alpha(x) = \frac{2\left(\frac{x}{2}\right)^\alpha}{\sqrt{\pi}\,\Gamma\left(\alpha+\frac{1}{2}\right)} \int_0^{\frac{\pi}{2}} \sin(x\cos\tau)\sin^{2\alpha}(\tau)\,d\tau.$$

1.1.3 Hypergeometric type functions

The hypergeometric Gauss function inside the circle $|z|<1$ is determined as the sum of the hypergeometric series (see [2], p. 373, formula (15.3.1))

$${}_2F_1(a,b;c;z) = F(a,b,c;z) = \sum_{k=0}^{\infty} \frac{(a)_k(b)_k}{(c)_k} \frac{z^k}{k!}, \tag{1.33}$$

and for $|z| \geq 1$ it is obtained by analytic continuation of this series. In (1.33) parameters a, b, c and variable z can be complex, and $c \neq 0, -1, -2, \ldots$. The multiplier $(a)_k$ is the Pochhammer symbol (1.10).

Since the hypergeometric series (1.33) converges only in the unit circle of the complex plane, it is necessary to construct an analytic continuation of the hypergeometric function beyond the boundary of this circle, to the entire complex plane. One of the ways to continue analytically is to use the Euler integral representation of the form

$$
\begin{aligned}
&{}_2F_1(a, b; c; z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(b-c)} \int\limits_0^1 t^{b-1}(1-t)^{c-b-1}(1-tz)^{-a}\,dt, \qquad (1.34)\\
&0 < \operatorname{Re} b < \operatorname{Re} c, \qquad |\arg(1-z)| < \pi,
\end{aligned}
$$

in which the right side is defined under the specified conditions, ensuring the convergence of the integral.

An important property of the hypergeometric function is that many special and elementary functions can be obtained from it with certain values of the parameters and the transformation of an independent argument.

Examples for elementary functions are

$$
\begin{aligned}
&(1+x)^n = {}_2F_1(-n, \beta, \beta; -x), \qquad \frac{1}{x}\ln(1+x) = {}_2F_1(1, 1, 2; -x),\\
&e^x = \lim_{n\to\infty} {}_2F_1\left(1, n, 1; \frac{x}{n}\right),\\
&\cos x = \lim_{\alpha,\,\beta\to\infty} {}_2F_1\left(\alpha, \beta, \frac{1}{2}; -\frac{x^2}{4\alpha\beta}\right),\\
&\cosh x = \lim_{\alpha,\,\beta\to\infty} {}_2F_1\left(\alpha, \beta, \frac{1}{2}; \frac{x^2}{4\alpha\beta}\right).
\end{aligned}
$$

The first-kind Bessel function and the Gauss hypergeometric function are related by the formula

$$
J_\nu(z) = \lim_{\alpha,\,\beta\to\infty} \left[\frac{\left(\frac{z}{2}\right)^\nu}{\Gamma(\nu+1)} {}_2F_1\left(\alpha, \beta, \nu+1; -\frac{z^2}{4\alpha\beta}\right) \right].
$$

The degenerate Kummer hypergeometric function ${}_1F_1(a; b; z)$ is

$$
{}_1F_1(a; b; z) = \sum_{n=0}^{\infty} \frac{a^{(n)} z^n}{b^{(n)} n!},
$$

associated with the Gauss hypergeometric function by the limiting relation

$$
{}_1F_1(a; c; z) = \lim_{b\to\infty} {}_2F_1\left(a, b; c; \frac{z}{b}\right).
$$

The degenerate Tricomi hypergeometric function $\Psi(a;b;z)$ is determined by the equality

$$\Psi(a;b;z) = \frac{\Gamma(1-b)}{\Gamma(a+1-b)}\,{}_1F_1(a;b;z) + \frac{\Gamma(b-1)}{\Gamma(a)} z^{1-b}\,{}_1F_1(a+1-b;2-b;z).$$

Next we present Whittaker functions which appear in kernels of integral transforms connected with the fractional Bessel integral.

Whittaker functions $M_{\kappa,\mu}(z)$ and $W_{\kappa,\mu}(z)$ are special solutions of Whittaker's equation

$$\frac{d^2w}{dz^2} + \left(-\frac{1}{4} + \frac{\kappa}{z} + \frac{1/4-\mu^2}{z^2}\right) w = 0.$$

They are modified forms of Kummer's confluent hypergeometric functions and were introduced by Edmund Taylor Whittaker by

$$\begin{aligned} M_{\kappa,\mu}(z) &= \exp(-z/2)\, z^{\mu+\frac{1}{2}} M\left(\mu-\kappa+\frac{1}{2}, 1+2\mu; z\right), \\ W_{\kappa,\mu}(z) &= \exp(-z/2)\, z^{\mu+\frac{1}{2}} U\left(\mu-\kappa+\frac{1}{2}, 1+2\mu; z\right), \end{aligned} \tag{1.35}$$

where

$$M(a,b,z) = \sum_{n=0}^{\infty} \frac{a^{(n)} z^n}{b^{(n)} n!} = {}_1F_1(a;b;z)$$

and

$$U(a,b,z) = \frac{\Gamma(1-b)}{\Gamma(a+1-b)} M(a,b,z) + \frac{\Gamma(b-1)}{\Gamma(a)} z^{1-b} M(a+1-b,2-b,z)$$

are Kummer's functions.

The Whittaker functions $M_{\kappa,\mu}(z)$ and $W_{\kappa,\mu}(z)$ are the same as those with opposite values of μ; in other words, considered as a function of μ at fixed κ and z, they are even functions. When κ and z are real, the functions give real values for real and imaginary values of μ.

A *generalized hypergeometric function* is defined as a power series

$${}_pF_q(a_1,\ldots,a_p;b_1,\ldots,b_q;z) = \sum_{n=0}^{\infty} \frac{(a_1)_n \cdots (a_p)_n}{(b_1)_n \cdots (b_q)_n} \frac{z^n}{n!}.$$

The functions of the form ${}_0F_1(;a;z)$ are called confluent hypergeometric limit functions and are closely related to the Bessel functions J_α and I_α. The relationships are

$$J_\alpha(x) = \frac{(\frac{x}{2})^\alpha}{\Gamma(\alpha+1)} {}_0F_1\left(;\alpha+1; -\tfrac{x^2}{4}\right),$$
$$I_\alpha(x) = \frac{(\frac{x}{2})^\alpha}{\Gamma(\alpha+1)} {}_0F_1\left(;\alpha+1; \tfrac{x^2}{4}\right)$$

or

$${}_0F_1\left(;\alpha+1; -\tfrac{x^2}{4}\right) = j_\alpha(x), \qquad {}_0F_1\left(;\alpha+1; \tfrac{x^2}{4}\right) = i_\alpha(x).$$

We also need the function ${}_1F_2(;a;z)$. It is known (see [456]) that for $\alpha > 0$, $\xi \geq 0$, $t > 0$,

$$\int_0^t \left(t^2-u^2\right)^{\alpha-1} u^{1-\gamma} J_\gamma(u\xi)dt = \frac{\xi^\gamma t^{2\alpha}}{2^{\gamma+1}\alpha\Gamma(\gamma+1)} \times {}_1F_2\left(1;\alpha+1,\gamma+1; -\frac{t^2\xi^2}{4}\right)$$

and for $\gamma < 2$, $\alpha > 0$, $\xi \geq 0$, $t > 0$,

$$\int_0^t \left(t^2-u^2\right)^{\alpha-1} u^{1-\gamma} I_\gamma(u\xi)dt = \frac{\xi^\gamma t^{2\alpha}}{2^{\gamma+1}\alpha\Gamma(\gamma+1)} \times {}_1F_2\left(1;\alpha+1,\gamma+1; \frac{t^2\xi^2}{4}\right).$$

The *Appell hypergeometric function* $F_4(a,b,c_1,c_2;x,y)$ (see [456], p. 658) for $|x|^{1/2}+|y|^{1/2}<1$ has the form

$$F_4(a,b,c_1,c_2;x,y) = \sum_{m,n=0}^{\infty} \frac{(a)_{m+n}(b)_{m+n}}{(c_1)_m(c_2)_n\, m!\, n!} x^m y^n. \tag{1.36}$$

For $|x|^{1/2}+|y|^{1/2} \geq 1$, function $F_4(a,b;c_1,c_2;x,y)$ is understood as an analytical continuation, which is determined by the formulas from [130].

Struve functions (of any order) can be expressed in terms of the generalized hypergeometric function ${}_1F_2$:

$$\mathbf{H}_\alpha(z) = \frac{z^{\alpha+1}}{2^\alpha\sqrt{\pi}\,\Gamma\left(\alpha+\frac{3}{2}\right)} {}_1F_2\left(1, \tfrac{3}{2}, \alpha+\tfrac{3}{2}, -\tfrac{z^2}{4}\right).$$

The Lauricella function ([500], p. 33) is

$$\begin{aligned}&\mathbf{F}_{\gamma}^{(n)}(a_1, ..., a_n, b_1, ..., b_n; c; z_1, ..., z_n)\\&= \sum_{m_1,...,m_n}^{\infty} \frac{(a_1)_{m_1}...(a_n)_{m_n}(b_1)_{m_1}...(b_n)_{m_n}}{(c)_{m_1+...+m_n}} \frac{z_1^{m_1}}{m_1!}...\frac{z_n^{m_n}}{m_n!},\\&\max\{|z_1|, ..., |z_n|\} < 1.\end{aligned} \tag{1.37}$$

The *Fox–Wright function* ${}_p\Psi_q(z)$ is defined for $z \in \mathbb{C}$, $a_l, b_j \in \mathbb{C}$, $\alpha_l, \beta_j \in \mathbb{R}$, $l = 1, ..., p$, $j = 1, ..., q$ by the series

$${}_p\Psi_q(z) = {}_p\Psi_q\left[\begin{matrix}(a_l, \alpha_l)_{1,p}\\(b_j, \beta_j)_{1,q}\end{matrix}\middle| z\right] = \sum_{k=0}^{\infty} \frac{\prod\limits_{l=1}^{p} \Gamma(a_l + \alpha_l k)}{\prod\limits_{j=1}^{q} \Gamma(b_j + \beta_j k)} \frac{z^k}{k!}. \tag{1.38}$$

If the condition

$$\sum_{j=1}^{q} \beta_j - \sum_{l=1}^{p} \alpha_l > -1$$

is satisfied, the series in (1.38) is convergent for any $z \in \mathbb{C}$. Let

$$\begin{aligned}\delta &= \prod_{l=1}^{p} |\alpha_l|^{-\alpha_l} \prod_{j=1}^{q} |\beta_j|^{\beta_j},\\ \mu &= \sum_{j=1}^{q} b_j - \sum_{l=1}^{p} a_l + \frac{p-q}{2}.\end{aligned}$$

If

$$\sum_{j=1}^{q} \beta_j - \sum_{l=1}^{p} \alpha_l = -1,$$

then the series in (1.38) is absolutely convergent for $|z| < \delta$ and for $|z| = \delta$ and $\operatorname{Re}\mu > \frac{1}{2}$.

The *Mittag-Leffler function* $E_{\alpha,\beta}(z)$ is the entire function of order $1/\alpha$ defined by the following series when the real part of α is strictly positive:

$$E_{\alpha,\beta}(z) = \sum_{n=0}^{\infty} \frac{z^n}{\Gamma(\alpha n + \beta)}, \ z \in \mathbb{C}, \ \alpha, \beta \in \mathbb{C}, \ \operatorname{Re}\alpha > 0, \ \operatorname{Re}\beta > 0. \tag{1.39}$$

Function (1.39) was introduced by Gesta Mittag-Leffler in 1903 for $\alpha = 1$ and by A. Wiman in 1905 in the general case. The first applications of these functions by

Mittag-Leffler and Wiman were applications in complex analysis (nontrivial examples of entire functions with noninteger orders of growth and generalized summation methods). In the USSR, these functions became mainly known after the publication of the famous monograph by M. M. Dzhrbashyan [98] (see also his later monograph [106]). The most famous application of the Mittag-Leffler functions in the theory of integro-differential equations and fractional calculus is the fact that through them the resolvent of the Riemann–Liouville fractional integral is explicitly expressed in accordance with the famous Hille–Tamarkin–Dzhrbashyan formula [494]. In view of the numerous applications to the solution of fractional differential equations, this function was deservedly named in [202] "*Royal function of fractional calculus*."

The derivative of the Mittag-Leffler function is calculated by the formula

$$E'_{\alpha,\beta}(z) = \frac{E_{\alpha,\beta}(z)}{dz} = \sum_{k=0}^{\infty} \frac{(1+k)z^k}{\Gamma(\beta+\alpha(1+k))}.$$

Note that

$$\begin{aligned} &E_{\alpha,\beta}(0) = 1, \\ &E_{0,1}(z) = \sum_{k=0}^{\infty} z^k = \frac{1}{1-z}, \\ &E_{1,1}(z) = e^z, \qquad E_{1,2}(z) = \frac{e^z-1}{z}, \qquad E_{2,2}(z) = \frac{\sinh(\sqrt{z})}{\sqrt{z}}. \end{aligned}$$

Using the Fox–Wright function (1.38) we can write

$$E_{\alpha,\beta}(z) = {}_1\Psi_1\left[\left.\begin{matrix}(1,1)\\(\beta,\alpha)\end{matrix}\right| z\right]. \tag{1.40}$$

A general definition of the Meijer G-function is given by the following line integral in the complex plane (see [20], p. 206):

$$\begin{aligned} &G^{m,n}_{p,q}\left(\left.\begin{matrix}a_1,\ldots,a_p\\b_1,\ldots,b_q\end{matrix}\right| z\right) \\ &\quad = \frac{1}{2\pi i}\int_L \frac{\prod\limits_{j=1}^{m}\Gamma(b_j-s)\prod\limits_{j=1}^{n}\Gamma(1-a_j+s)}{\prod\limits_{j=m+1}^{q}\Gamma(1-b_j+s)\prod\limits_{j=n+1}^{p}\Gamma(a_j-s)} z^s\,ds. \end{aligned} \tag{1.41}$$

The general Legendre equation reads

$$(1-x^2)\,y'' - 2xy' + \left[\lambda(\lambda+1) - \frac{\mu^2}{1-x^2}\right] y = 0,$$

where the numbers λ and μ may be complex. This differential equation has two linearly independent solutions, which can both be expressed in terms of the hypergeometric function ${}_2F_1$:

$$P_\lambda^\mu(z) = \frac{1}{\Gamma(1-\mu)}\left[\frac{1+z}{1-z}\right]^{\mu/2} {}_2F_1\left(-\lambda, \lambda+1; 1-\mu; \frac{1-z}{2}\right), \tag{1.42}$$

for $|1-z| < 2$,

$$Q_\lambda^\mu(z) = \frac{\sqrt{\pi}\,\Gamma(\lambda+\mu+1)}{2^{\lambda+1}\Gamma(\lambda+3/2)} \frac{e^{i\mu\pi}(z^2-1)^{\mu/2}}{z^{\lambda+\mu+1}} {}_2F_1\left(\frac{\lambda+\mu+1}{2}, \frac{\lambda+\mu+2}{2}; \lambda+\frac{3}{2}; \frac{1}{z^2}\right), \tag{1.43}$$

for $|z| > 1$.

Functions P_λ^μ and Q_λ^μ are generally known as Legendre functions of the first and second kind of noninteger degree, with the additional qualifier "associated" if μ is nonzero.

We will use also Legendre functions with $z = x$, where $-1 < x < 1$ (see [19]). The interval $(-1, 1)$ is "the cut." If μ is even integer, then the values of $P_\lambda^\mu(z)$ on both sides of the cut are equal and in this case it is sufficient to take the branch cut along the real axis from -1 to $-\infty$. In all other cases $P_\lambda^\mu(x - i0)$ and $P_\lambda^\mu(x + i0)$ are different (here $f(x \pm i0)$ means $\lim\limits_{\varepsilon\to 0} f(x \pm i\varepsilon)$, $\varepsilon > 0$). In order to avoid ambiguity it is usual to introduce slightly modified Legendre functions. These will be denoted by $\mathbb{P}_\nu^\mu(z)$ and $\mathbb{Q}_\nu^\mu(z)$:

$$\mathbb{P}_\nu^\mu(x) = \frac{1}{2}\left[e^{i\frac{\mu\pi}{2}} P_\nu^\mu(x+i0) + e^{-i\frac{\mu\pi}{2}} P_\nu^\mu(x-i0)\right], \tag{1.44}$$

$$\mathbb{Q}_\nu^\mu(x) = \frac{1}{2}e^{-i\mu\pi}\left[e^{-i\frac{\mu\pi}{2}} Q_\nu^\mu(x+i0) + e^{i\frac{\mu\pi}{2}} Q_\nu^\mu(x-i0)\right]. \tag{1.45}$$

1.1.4 Polynomials

Gegenbauer polynomials or ultraspherical polynomials $C_n^{(\alpha)}(x)$ are orthogonal polynomials on the interval $[-1, 1]$ with respect to the weight function $(1-x^2)^{\alpha-\frac{1}{2}}$ that can be defined by the recurrence relation

$$C_0^\alpha(x) = 1, \qquad C_1^\alpha(x) = 2\alpha x,$$
$$C_n^\alpha(x) = \frac{1}{n}[2x(n+\alpha-1)C_{n-1}^\alpha(x) - (n+2\alpha-2)C_{n-2}^\alpha(x)].$$

The next decomposition is valid:

$$\frac{1}{(1-2xt+t^2)^\alpha} = \sum_{n=0}^\infty C_n^{(\alpha)}(x)t^n.$$

for any constant C, which are real for $x < 0$. The characteristics comprise two families of semicubical parabolas lying in the half-plane $x < 0$, with cusps on the line $x = 0$. This is of hyperbolic degeneracy, for which the two characteristic families coincide, perpendicularly to the line $x = 0$.

For $\pm x > 0$, set $\tau = \frac{2}{3}(\pm x)^{\frac{3}{2}}$. Then the Tricomi equation becomes the classical elliptic or hyperbolic Euler–Poisson–Darboux equation:

$$u_{\tau\tau} \pm u_{yy} + \frac{\beta}{\tau} u_\tau = 0.$$

The index $\beta = \frac{1}{3}$ determines the singularity of solutions near $\tau = 0$, equivalently, $x = 0$.

Many important problems in fluid mechanics and differential geometry can be reduced to corresponding problems for the Tricomi equation, particularly transonic flow problems and isometric embedding problems. The Tricomi equation is a prototype of the generalized Tricomi equation:

$$u_{xx} + K(x) u_{yy} = 0.$$

For a steady-state transonic flow in $\mathbb{R}^2$, $u(x; y)$ is the stream function of the flow, $K(x)$ and x are functions of the velocity, which are positive at subsonic and negative at supersonic speeds, and y is the angle of inclination of the velocity. The solutions $u(x; y)$ also serve as entropy generators for entropy pairs of the potential flow system for the velocity. For the isometric embedding problem of two-dimensional Riemannian manifolds into $\mathbb{R}^3$, the function $K(x)$ has the same sign as the Gaussian curvature.

A closely related partial differential equation is the Keldysh equation

$$x u_{xx} + u_{yy} = 0.$$

It is hyperbolic when $x < 0$, is elliptic when $x > 0$, and degenerates on the line $x = 0$. Its characteristics are

$$y \pm \frac{1}{2}(-x)^{\frac{1}{2}} = C$$

for any constant C, which are real for $x < 0$. The two characteristic families are (quadratic) parabolas lying in the half-plane $x < 0$ and coincide tangentially to the degenerate line $x = 0$, which is of parabolic degeneracy. For $\pm x > 0$, the Keldysh equation becomes the elliptic or hyperbolic Euler–Poisson–Darboux equation with index $\beta = -\frac{1}{4}$ by setting $\tau = \frac{1}{2}(\pm x)\frac{1}{2}$. Many important problems in continuum mechanics can also be reduced to corresponding problems for the Keldysh equation, particularly shock reflection–diffraction problems in gas dynamics.

1.4.4 Abstract Euler–Poisson–Darboux equation

The abstract Euler–Poisson–Darboux equation has the form

$$Au = (B_\gamma)_t \, u, \qquad u = u(x, t; \gamma), \tag{1.108}$$

where A is a linear operator acting only by variable $x = (x_1, ..., x_n)$.

Lemma 7. *Let* $u^k = u(x, t; k)$ *be a solution to Eq.* (1.108)*. The two recurrent formulas*

$$u^k = t^{1-k} u^{2-k}, \tag{1.109}$$

$$u^k_t = t u^{k+2} \tag{1.110}$$

are valid.

Proof. Let us show (1.109). Putting $w = t^{k-1} v$, $v = u^k$, we obtain

$$\begin{aligned}
w_t &= (k-1)t^{k-2} v + t^{k-1} v_t = \frac{k-1}{t}\, w + t^{k-1} v_t,\\
w_{tt} &= (k-1)(k-2)t^{k-3} v + (k-1)t^{k-2} v_t + (k-1)t^{k-2} v_t + t^{k-1} v_{tt}\\
&= \frac{(k-1)(k-2)}{t^2}\, w + 2(k-1)t^{k-2} v_t + t^{k-1} v_{tt},\\
\frac{2-k}{t}\, w_t &= -\frac{(k-1)(k-2)}{t^2}\, w + (2-k)t^{k-2} v_t,\\
w_{tt} + \frac{2-k}{t}\, w_t &= 2(k-1)t^{k-2} v_t + t^{k-1} v_{tt} + (2-k)t^{k-2} v_t\\
&= t^{k-1}\left(v_{tt} + \frac{k}{t}\, v_t\right)
\end{aligned}$$

or

$$w_{tt} + \frac{2-k}{t}\, w_t = t^{k-1}\left(v_{tt} + \frac{k}{t}\, v_t\right). \tag{1.111}$$

If $w = t^{k-1} v$ satisfies the equation

$$A w = w_{tt} + \frac{2-k}{t}\, w_t,$$

then using (1.111) we get

$$t^{k-1} A v = t^{k-1}\left(v_{tt} + \frac{k}{t}\, v_t\right),$$

which means that v satisfies the equation

$$A v = v_{tt} + \frac{k}{t}\, v_t.$$

Denoting $w = u^{2-k}$ we obtain (1.109).

Let us prove now (1.110). Putting $t w = v_t$, $v = u^k$, we can write

$$w_t = -\frac{1}{t^2}\, v_t + \frac{1}{t}\, v_{tt},$$

$$w_{tt} = \frac{2}{t^3}\,v_t - \frac{2}{t^2}\,v_{tt} + \frac{1}{t}\,v_{ttt},$$

and

$$\frac{k+2}{t}\,w_t = -\frac{k+2}{t^3}\,v_t + \frac{k+2}{t^2}\,v_{tt}.$$

We have

$$\begin{aligned} w_{tt} + \frac{k+2}{t}\,w_t &= \frac{2}{t^3}\,v_t - \frac{2}{t^2}\,v_{tt} + \frac{1}{t}\,v_{ttt} - \frac{k+2}{t^3}\,v_t + \frac{k+2}{t^2}\,v_{tt} \\ &= \frac{1}{t}\,v_{ttt} - \frac{k}{t^3}\,v_t + \frac{k}{t^2}\,v_{tt} = \frac{1}{t}\left(v_{ttt} - \frac{k}{t^2}\,v_t + \frac{k}{t}\,v_{tt}\right) \\ &= \frac{1}{t}\frac{\partial}{\partial t}\left(v_{tt} + \frac{k}{t}\,v_t\right) \end{aligned}$$

or

$$w_{tt} + \frac{k+2}{t}\,w_t = \frac{1}{t}\frac{\partial}{\partial t}\left(v_{tt} + \frac{k}{t}\,v_t\right). \tag{1.112}$$

If $w = \frac{1}{t}\,v_t$ satisfies the equation

$$Aw = w_{tt} + \frac{k+2}{t}\,w_t,$$

then using (1.112) we obtain

$$\frac{1}{t}\frac{\partial}{\partial t}Av = \frac{1}{t}\frac{\partial}{\partial t}\left(v_{tt} + \frac{k}{t}\,v_t\right),$$

which means that v satisfies the equation

$$Av = v_{tt} + \frac{k}{t}v_t.$$

Denoting $w = u^{k+2}$, $v = u^k$, we get (1.110). □

For $A = \Delta_\gamma$ formula (1.109) is proved in [147] and formula (1.110) is proved in [16]. Both formulas are present in Weinsteins' article [595], but in the case when in Eq. (1.108) for each of variables x_i, $i = 1, ..., n$, the second derivative acts. These recurring formulas will allow using the solution u^k to (1.108) to get a solution to the same equation, but with the parameters $k+2$ and $2-k$, respectively.

So, if the function $u(x,t;k)$ is a solution of the abstract Euler–Poisson–Darboux equation $Au = (B_k)_t\, u$, where A is a linear operator acting only by $x = (x_1, ..., x_n)$, then the function $t^{1-k}u(x,t;2-k)$ also is a solution to this equation.

Now let us construct integral operators of Buschman–Erdélyi with more general functions as kernels. Consider an operator ${}_1S^{\nu}_{0+}$. It has the form

$$ {}_1S^{\nu}_{0+} = \frac{d}{dx}\int\limits_0^x K\left(\frac{x}{y}\right) f(y)\,dy, \tag{5.61} $$

with kernel K expressed by $K(z) = P_\nu(z)$. Simple properties of special functions lead to the fact that ${}_1S^{\nu}_{0+}$ is a special case of (5.61) with the Gegenbauer function kernel

$$ K(z) = \frac{\Gamma(\alpha+1)\,\Gamma(2\beta)}{2^{p-\frac{1}{2}}\Gamma(\alpha+2\beta)\,\Gamma(\beta+\frac{1}{2})}(z^\alpha-1)^{\beta-\frac{1}{2}}C^{\beta}_{\alpha}(z) \tag{5.62} $$

with $\alpha = \nu$, $\beta = \frac{1}{2}$, or with the Jacobi function kernel

$$ K(z) = \frac{\Gamma(\alpha+1)}{2^\rho\Gamma(\alpha+\rho+1)}(z-1)^\rho(z+1)^\sigma P^{(\rho,\sigma)}_{\alpha}(z) \tag{5.63} $$

with $\alpha = \nu$, $\rho = \sigma = 0$. More general are operators with the Gauss hypergeometric function kernel ${}_2F_1$ or Meijer G-function $G^{m,n}_{p,q}$ (1.41) or Fox–Wright $E_{\alpha,\beta}$ function (1.40) kernels (cf. [494]). For studying such operators inequalities for kernel functions are very useful (e.g., [228,528]).

Define the first class of generalized operators.

Definition 40. *Define Gauss–Buschman–Erdélyi operators by formulas*

$$ {}_1F_{0+}(a,b,c)[f] = \frac{1}{2^{c-1}\Gamma(c)}, \tag{5.64} $$

$$ \int\limits_0^x \left(\frac{x}{y}-1\right)^{c-1}\left(\frac{x}{y}+1\right)^{a+b-c} {}_2F_1\left({}^{a,b}_{c}|\frac{1}{2}-\frac{1}{2}\frac{x}{y}\right) f(y)\,dy, $$

$$ {}_2F_{0+}(a,b,c)[f] = \frac{1}{2^{c-1}\Gamma(c)}, \tag{5.65} $$

$$ \int\limits_0^x \left(\frac{y}{x}-1\right)^{c-1}\left(\frac{y}{x}+1\right)^{a+b-c} {}_2F_1\left({}^{a,b}_{c}|\frac{1}{2}-\frac{1}{2}\frac{y}{x}\right) f(y)\,dy, $$

$$ {}_1F_{-}(a,b,c)[f] = \frac{1}{2^{c-1}\Gamma(c)}, \tag{5.66} $$

$$ \int\limits_0^x \left(\frac{y}{x}-1\right)^{c-1}\left(\frac{y}{x}+1\right)^{a+b-c} {}_2F_1\left({}^{a,b}_{c}|\frac{1}{2}-\frac{1}{2}\frac{y}{x}\right) f(y)\,dy, $$

$$
{}_2F_{-}(a,b,c)[f] = \frac{1}{2^{c-1}\Gamma(c)}, \tag{5.67}
$$

$$
\int\limits_0^x \left(\frac{x}{y}-1\right)^{c-1}\left(\frac{x}{y}+1\right)^{a+b-c} {}_2F_1\left({}^{a,b}_{c}|\frac{1}{2}-\frac{1}{2}\frac{x}{y}\right) f(y)\,dy,
$$

$$
{}_3F_{0+}[f] = \frac{d}{dx}\,{}_1F_{0+}[f], \quad {}_4F_{0+}[f] = {}_2F_{0+}\frac{d}{dx}[f], \tag{5.68}
$$

$$
{}_3F_{-}[f] = {}_1F_{-}\left(-\frac{d}{dx}\right)[f], \quad {}_4F_{-}[f] = \left(-\frac{d}{dx}\right){}_2F_{-}[f]. \tag{5.69}
$$

The symbol ${}_2F_1$ in definitions (5.65) and (5.67) means the Gauss hypergeometric function on the natural domain and in (5.64) and (5.66) the main branch of its analytical continuation.

Operators (5.64)–(5.67) generalize Buschman–Erdélyi ones (5.5)–(5.8), respectively. They reduce to Buschman–Erdélyi for the choice of parameters $a=-(\nu+\mu)$, $b=1+\nu-\mu$, $c=1-\mu$. For operators (5.64)–(5.67) the above results are generalized with necessary changes. For example they are factorized via more simple operators (5.68) and (5.69) with a special choice of parameters.

Operators (5.68) and (5.69) are generalizations of (5.14) and (5.15). For them the following result is true.

Theorem 64. *Operators* (5.68) *and* (5.69) *may be extended to isometric in* $L_2(0,\infty)$ *if and only if they coincide with Buschman–Erdélyi operators of zero order smoothness* (5.14) *and* (5.15) *for integer values of* $\nu = \frac{1}{2}(b-a-1)$.

This theorem singles out Buschman–Erdélyi operators of zero order smoothness at least in the class (5.64)–(5.69). Operators (5.64)–(5.67) are generalizations of fractional integrals. Analogously may be studied generalizations to (5.33)–(5.34), (5.43), and (5.46)–(5.47).

More general are operators with G-function kernel,

$$
\begin{aligned}
{}_1G_{0+}(\alpha,\beta,\delta,\gamma)[f] &= \frac{2^{\delta}}{\Gamma(1-\alpha)\Gamma(1-\beta)} \\
&\times \int\limits_0^x \left(\frac{x}{y}-1\right)^{-\delta}\left(\frac{x}{y}+1\right)^{1+\delta-\alpha-\beta} G^{1\,2}_{2\,2}\left(\frac{x}{2y}-\frac{1}{2}\Big|^{\alpha,\,\beta}_{\gamma,\,\delta}\right) f(y)\,dy.
\end{aligned} \tag{5.70}
$$

Other operators are with different intervals of integration and parameters of the G-function. For $\alpha=1-a$, $\beta=1-b$, $\delta=1-c$, $\gamma=0$ (5.70) reduces to (5.64), for $\alpha=1+\nu$, $\beta=-\nu$, $\delta=\gamma=0$ (5.70) reduces to Buschman–Erdélyi operators of zero order smoothness ${}_1S^{\nu}_{0+}$.

Further generalizations are in terms of Wright or Fox functions. They lead to Wright–Buschman–Erdélyi and Fox–Buschman–Erdélyi operators. These classes are connected with Sonine–Dimovski and Poisson–Dimovski transmutations [89,93], and also with generalized fractional integrals introduced by Kiryakova [252].

V. Katrakhov found a new approach for boundary value problems for elliptic equations with strong singularities of infinite order. For example for the Poisson equation he studied problems with solutions of arbitrary growth. At singular points he proposed a new kind of boundary condition: the K-trace. His results are based on constant usage of Buschman–Erdélyi transmutations of the first kind for definition of norms, solution estimates, and correctness proofs [225,227].

Moreover in joint papers with I. Kipriyanov he introduced and studied new classes of pseudodifferential operators based on transmutational techniques [248–250]. These results were paraphrased in a reorganized manner in [52].

5.3 Multi-dimensional integral transforms of Buschman–Erdélyi type with Legendre functions in kernels

In this section we consider generalizations of Buschman–Erdélyi operators for the multi-dimensional case. This case was studied by O. V. Skoromnik and S. M. Sitnik (see [538]).

5.3.1 *Basic definitions*

Let $x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n_+$, $t = (t_1, t_2, ..., t_n) \in \mathbb{R}^n_+$, $x \cdot t = \sum\limits_{n=1}^{n} x_n t_n$ denotes their scalar product, and in particular, $x \cdot 1 = \sum\limits_{n=1}^{n} x_n$. The expression $x > t$ means that $x_1 > t_1, ..., x_n > t_n$, the nonstrict inequality $\geq$ has a similar meaning, by $\mathbb{N} = \{1, 2, ...\}$ we denote the set of positive integers, $\mathbb{N}_0 = \mathbb{N} \bigcup \{0\}$, and $\mathbb{N}^n_0 = \mathbb{N}_0 \times \mathbb{N}_0 \times ... \times \mathbb{N}_0$. We denote

$\mathbf{m} = (m_1, m_2, ..., m_n) \in \mathbb{N}^n_0$, $m_1 = m_2 = ... = m_n$;
$\mathbf{n} = (\overline{n}_1, \overline{n}_2, ..., \overline{n}_n) \in \mathbb{N}^n_0$, $\overline{n}_1 = \overline{n}_2 = ... = \overline{n}_n$;
$\mathbf{p} = (p_1, p_2, ..., p_n) \in \mathbb{N}^n_0$, $p_1 = p_2 = ... = p_n$;
$\mathbf{q} = (q_1, q_2, ..., q_n) \in \mathbb{N}^n_0$, $q_1 = q_2 = ... = q_n$, $(0 \leq \mathbf{m} \leq \mathbf{q}, 0 \leq \mathbf{m} \leq \mathbf{p})$.

We put

$\sigma = (\sigma_1, \sigma_2, ..., \sigma_n) \in \mathbb{C}^n$;
$\kappa = (\kappa_1, \kappa_2, ..., \kappa_n) \in \mathbb{C}^n$;
$\delta = (\delta_1, \delta_2, ..., \delta_n) \in \mathbb{R}^n$;
$\gamma = (\gamma_1, \gamma_2, ..., \gamma_n) \in \mathbb{R}^n$, $0 < \gamma < 1$;
$a_i = (a_{i_1}, a_{i_2}, ..., a_{i_n})$, $1 \leq i \leq p$, $a_{i_1}, a_{i_2}, ..., a_{i_n} \in \mathbb{C}$, $1 \leq i_1 \leq p_1, ..., 1 \leq i_n \leq p_n$;
$b_j = (b_{j_1}, b_{j_2}, ..., b_{j_n})$, $1 \leq j \leq q$, $b_{j_1}, b_{j_2}, ..., b_{j_n} \in \mathbb{C}$, $1 \leq j_1 \leq q_1, ..., 1 \leq j_n \leq q_n$;
$\alpha_i = (\alpha_{i_1}, \alpha_{i_2}, ..., \alpha_{i_n})$, $1 \leq i \leq p$, $\alpha_{i_1}, \alpha_{i_2}, ..., \alpha_{i_n} \in \mathbb{R}^1_+$, $1 \leq i_1 \leq p_1, ..., 1 \leq i_n \leq p_n$;

-256- - -* ! - ' 2 D ! 3 2 1 D
A !11 ! 2 ! ! ! ! 3 % 1"
(%B/ % !1 B !A ! " 3
K -256- - -* ! - ' 9 ! 2 7! 3 2 ! G
" 1"

(B/

Pag. 185(195) Chapter 5
Conservation Laws
Gauss theorem.
- The General Theory of Particle
Mechanics...pdf
- Transmutations-Singular-and-
Fractional-Differential-Equations.pdf
Pag. 06(32)-
1.1.3 Hypergeometric type
functions
The hypergeometric Gauss function
inside the circle |z|<1
Pag. 50(76)-
Gaussian curvature
Pag. 227(252)- Definition 40
Gauss-Buschman-Erdélyi
operators by formulas...

→ Riemann

- 20211006/ Electricity Maxwell/ Algebraic geometry...pdf
- 20211006/ Elect.. Maxwell/ Thomas.. All the Math you missed(2022).pdf
Riemann-Roch Theorem for smooth curves - Pag. 369(380)
(veja os 2 arquivos) Chapter 6 - Sheaves and Cohomology.

110	50	014 2021	100	200	210	100	91-2020-(7,65)	
105	50	20210714 (24,69)	105	110	220	115	~~100~~	
100	100	130	200	210	150	105	91-100	
100	200	120	220	110	580	120	320	590
105	110	130	625	630	110	120	110	
100	120	380	Books	007-2021	690	560	420	126-2020
800	105	240	2021 10	01-20210726	104-2020	110	100	(7,39)
162	635	620	20210812	(17,34)	(25,25)	670	50	
	85	018-2021	(44,89)			05-2021 20210731	570	
	720	20210913 (50,49)				(39,60)	20 590	

100 100 > 400 + 100 > 600 + 90 = 690 010-e-011-2021
100 100 100 50 740 → (56,32)

210 Data-NovoVivo-102-2020 (28,83) DATA (continuar

105

20211007 (20211007)

US$ 5 / 3,78 (gal) = 1,322751

x 5,5 (1US$ = R$ 5,50) = 7,275132

Books

20211007/ Geometric Approaches to ~~of~~ Quantum Field Theory/

3030852687.pdf

Pág. 15(31) Chapter 2

Field Space Covariance

"God always geometrises" Plato.

2.1 Riemannian Geometry.

Pág. 34(50) - Gaussian integral

2021 10 07 / Précis de Géométrie.../

Précis de géométrie .. Papelier...pdf

Pág. 292(301) Fig. 120.

Obs. Nº 380 - Tangentes à
l'ellipse.

$x = a \cos \varphi$; $y = b \operatorname{sen} \varphi$;

Obs: Ellipse use estas equações para simular a deformação do espaço-tempo com os devidas adequações.

Na realidade a ellipse é uma deformação do círculo, ou seja, com os devidos ajustes das equações matemáticas poderemos simular a curvatura do espaço-tempo do Universo.

Obs: $x = a \cos \varphi$,

$$y_1 = \frac{a}{b} y = a \operatorname{sen} \varphi.$$

! -256- - -* ! - '/ % % $ B 1"
($ % !1 < " 1! A! ! N7 " ! !3 O !
(%' $ 7! ! !

D -256- - -* ! - '/
M I"0! PD I"0! ! 3 I I" IQ R !1 I7 ID1 B. ' $%% 1"

D -256- - -* ! - '/
M I" I M M I! ! 3 > ISIQ R !1 I7 ID1 B./%$ 1"
(/ % <

D% -256- - -* ! - '/
M I" I M M I! ! 3 > ISIQ R !1 I7 ID1 B./' % > 1"

D' -256- - -* ! - '/
M I" I M ! > ISI 0 ! IQ R !1 I7 ID1 B. ' / 1"

*D
D -256- - -* ! - '/
M I" I M M I! ! 3 > ISIQ R !1 I7 ID1 B./%$ 1"
(/ % <

On obtiendrait la même condition en supposant $v \neq 0$.

Si l'on a $a^2u^2 + b^2v^2 - w^2 < 0$, la droite rencontre l'ellipse en des points imaginaires; et enfin si $a^2u^2 + b^2v^2 - w^2 = 0$, la droite rencontre l'ellipse en deux points confondus, elle est tangente à l'ellipse.

Il résulte de là que l'équation tangentielle de cette courbe est

$$a^2u^2 + b^2v^2 - w^2 = 0.$$

Tangentes à l'ellipse.

380. Tangente en un point de la courbe. — La tangente en un point (x, y) de la courbe a pour équation

$$\frac{Xx}{a^2} + \frac{Yy}{b^2} - 1 = 0,$$

x et y vérifiant la relation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - 1 = 0.$$

L'équation de la tangente renferme donc deux paramètres liés par une relation.

On peut exprimer les coordonnées d'un point de l'ellipse en fonction d'un seul paramètre. En effet, la relation $\frac{x^2}{a^2} + \frac{y^2}{b^2} - 1 = 0$ nous conduit à poser $\frac{x}{a} = \cos\varphi$, $\frac{y}{b} = \sin\varphi$, ou

$$x = a\cos\varphi, \qquad y = b\sin\varphi;$$

en faisant varier φ de 0 à 2π, ces formules donnent les coordonnées de tous les points de l'ellipse.

L'équation de la tangente au point (x, y) peut alors s'écrire

$$\frac{X\cos\varphi}{a} + \frac{Y\sin\varphi}{b} - 1 = 0;$$

elle ne contient plus qu'un seul paramètre.

Posons $\operatorname{tg}\frac{\varphi}{2} = t$, nous avons

$$x = \frac{a(1 - t^2)}{1 + t^2}, \qquad y = \frac{2bt}{1 + t^2};$$

x et y sont alors exprimés rationnellement en fonction d'un paramètre, ce qui montre que *l'ellipse est une courbe unicursale* (229).

L'équation de la tangente est alors

$$\frac{X(1-t^2)}{a}+\frac{2tY}{b}-(1+t^2)=0. \quad (*)$$

L'angle φ a une signification géométrique remarquable. Soit M le point de l'ellipse ayant pour coordonnées

$$x=a\cos\varphi, \qquad y=b\sin\varphi;$$

le point correspondant M_1 du cercle décrit sur AA' comme diamètre a pour coordonnées

$$x=a\cos\varphi, \qquad y_1=\frac{a}{b}y=a\sin\varphi;$$

on en conclut que l'angle φ est l'angle de la demi-droite OM_1 avec la demi-droite OA (*fig.* 120).

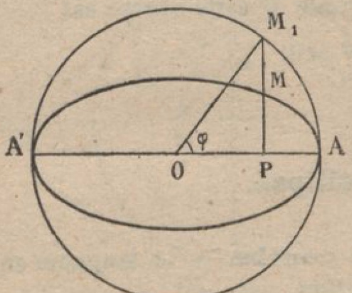

Fig. 120.

L'angle φ est appelé *l'anomalie excentrique* (**) du point M.

381. Condition pour qu'une droite soit tangente à l'ellipse. — Soit la droite $uX+vY+w=0$; pour qu'elle soit tangente à l'ellipse, il faut et il suffit qu'il existe un système de valeurs de x et de y vérifiant la relation $\frac{x^2}{a^2}+\frac{y^2}{b^2}-1=0$ et telles que les deux équations

$$\frac{Xx}{a^2}+\frac{Yy}{b^2}-1=0, \qquad uX+vY+w=0$$

représentent la même droite.

On doit donc avoir

$$\frac{x}{a^2u}=\frac{y}{b^2v}=\frac{-1}{w},$$

d'où $x=-\frac{a^2u}{w}$, $y=-\frac{b^2v}{w}$ et, en portant ces valeurs dans la relation $\frac{x^2}{a^2}+\frac{y^2}{b^2}-1=0$, on a

$$a^2u^2+b^2v^2-w^2=0.$$

C'est l'équation tangentielle de l'ellipse, que nous avons déjà obtenue au nº 379.

(*) Cette équation étant du deuxième degré par rapport à t, la droite qu'elle représente enveloppe une courbe de deuxième classe (289), c'est-à-dire une conique.

(**) Cette appellation est empruntée à l'astronomie.

382. Tangentes par un point non situé sur la courbe. — 1° Les points de contact sont à l'intersection de l'ellipse et de la polaire du point donné P(α, β),

$$\frac{\alpha x}{a^2}+\frac{\beta y}{b^2}-1=0.$$

Cette droite rencontre l'ellipse en des points réels si l'on a (379)

$$\frac{\alpha^2}{a^2}+\frac{\beta^2}{b^2}-1>0,$$

c'est-à-dire si le point P est dans la région positive de l'ellipse (269).

Or le centre de la courbe est dans la région négative; nous appellerons cette région la *région intérieure*, l'autre région qui ne contient pas le centre est appelée la *région extérieure*.

Par conséquent, pour que les tangentes issues d'un point P à une ellipse soient réelles, il faut et il suffit que le point P soit situé dans la région extérieure.

2° On a l'équation aux coefficients angulaires de ces tangentes en écrivant que la droite $y-\beta=t(x-\alpha)$ est tangente, ce qui donne

$$a^2t^2+b^2-(\beta-t\alpha)^2=0,$$

ou

$$t^2(a^2-\alpha^2)+2t\alpha\beta+b^2-\beta^2=0,$$

et cette équation a ses racines réelles si l'on a $\frac{\alpha^2}{a^2}+\frac{\beta^2}{b^2}-1>0.$

3° Enfin on obtient l'équation de l'ensemble des tangentes issues du point P en remplaçant dans cette dernière équation t par $\frac{y-\beta}{x-\alpha}$.

L'équation obtenue peut aussi se mettre sous la forme (292)

$$\left(\frac{\alpha x}{a^2}+\frac{\beta y}{b^2}-1\right)^2-\left(\frac{x^2}{a^2}+\frac{y^2}{b^2}-1\right)\left(\frac{\alpha^2}{a^2}+\frac{\beta^2}{b^2}-1\right)=0.$$

383. Lieu géométrique des points d'où l'on peut mener deux tangentes rectangulaires à l'ellipse. — Soient α, β les coordonnées d'un point du lieu. Les coefficients angulaires des tangentes passant par ce point sont racines de l'équation

$$t^2(a^2-\alpha^2)+2t\alpha\beta+b^2-\beta^2=0;$$

pour que ces tangentes soient perpendiculaires, il faut que le produit des racines de cette équation soit égal à -1, ce qui donne

$$\alpha^2+\beta^2-a^2-b^2=0;$$

en y remplaçant α, β par x, y, on a l'équation du lieu

$$x^2+y^2-a^2-b^2=0.$$

Le lieu cherché est le cercle circonscrit au rectangle construit sur les axes de l'ellipse; ce cercle est appelé le *cercle orthoptique* de l'ellipse.

384. Tangentes parallèles à une direction donnée. — Les points de contact de ces tangentes sont à l'intersection de l'ellipse et du diamètre conjugué de la direction donnée; ces points sont toujours réels, car le diamètre passe par le centre.

Cherchons maintenant les équations de ces tangentes. Soit m le coefficient angulaire de la direction donnée; écrivons que la droite $y = mx + \lambda$ est tangente, nous avons

$$a^2m^2 + b^2 - \lambda^2 = 0;$$

nous en tirons $\lambda = \pm\sqrt{a^2m^2 + b^2}$, et par suite les tangentes à l'ellipse parallèles à la direction donnée ont pour équations

$$y = mx \pm \sqrt{a^2m^2 + b^2}.$$

Normales à l'ellipse.

385. Normale en un point de la courbe. — La normale au point (x, y) a pour équation

$$\frac{X - x}{\frac{x}{a^2}} = \frac{Y - y}{\frac{y}{b^2}}, \qquad \text{ou} \qquad \frac{a^2X}{x} - a^2 = \frac{b^2Y}{y} - b^2.$$

En supposant toujours $a > b$, $a^2 - b^2$ est une quantité positive que l'on désigne par c^2; l'équation de la normale est alors

$$\frac{a^2X}{x} - \frac{b^2Y}{y} - c^2 = 0,$$

avec la condition

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - 1 = 0.$$

En remplaçant x par $a\cos\varphi$ et y par $b\sin\varphi$, on obtient l'équation de la normale en fonction d'un seul paramètre φ,

$$\frac{aX}{\cos\varphi} - \frac{bY}{\sin\varphi} - c^2 = 0;$$

on peut aussi l'écrire, en posant $\operatorname{tg}\frac{\varphi}{2} = t$,

$$\frac{aX(1 + t^2)}{1 - t^2} - \frac{bY(1 + t^2)}{2t} - c^2 = 0,$$

ou

$$2at(1 + t^2)X - b(1 - t^4)Y - 2c^2t(1 - t^2) = 0.$$

Cette équation étant du quatrième degré par rapport à t, on reconnaît (289) que par un point quelconque du plan on peut mener quatre normales à une ellipse. Par suite, l'enveloppe des normales, ou la développée de l'ellipse, est une courbe algébrique, unicursale et de quatrième classe.

386. Condition pour qu'une droite soit normale à l'ellipse. — Pour que la droite $uX+vY+w=0$ soit normale à l'ellipse, il faut et il suffit qu'il existe un système de valeurs de x et de y vérifiant l'équation $\frac{x^2}{a^2}+\frac{y^2}{b^2}-1=0$, et telles que les deux équations

$$\frac{a^2X}{x}-\frac{b^2Y}{y}-c^2=0, \qquad uX+vY+w=0$$

représentent la même droite.

On doit donc avoir

$$\frac{a^2}{ux}=-\frac{b^2}{vy}=-\frac{c^2}{w},$$

d'où

$$x=-\frac{a^2w}{c^2u}, \qquad y=\frac{b^2w}{c^2v},$$

et, en portant ces valeurs dans la relation $\frac{x^2}{a^2}+\frac{y^2}{b^2}-1=0$, on a

$$\frac{a^2w^2}{c^4u^2}+\frac{b^2w^2}{c^4v^2}-1=0,$$

ou

$$w^2(a^2v^2+b^2u^2)-c^4u^2v^2=0.$$

On retient aisément cette équation en la mettant sous la forme

$$\frac{a^2}{u^2}+\frac{b^2}{v^2}-\frac{c^4}{w^2}=0.$$

C'est l'équation tangentielle de la développée de l'ellipse; et nous voyons de nouveau que cette développée est de quatrième classe.

387. Développée de l'ellipse. — Cherchons maintenant l'équation ponctuelle de la développée.

La normale à l'ellipse au point qui a pour coordonnées λ et μ a pour équation

$$(1) \qquad \frac{a^2x}{\lambda}-\frac{b^2y}{\mu}-c^2=0,$$

λ et μ vérifiant la relation

$$(2) \qquad \frac{\lambda^2}{a^2}+\frac{\mu^2}{b^2}-1=0.$$

Pour avoir l'enveloppe de cette normale (274), il faut éliminer λ et μ entre les équations (1), (2) et la suivante

$$\frac{-\frac{a^2x}{\lambda^2}}{\frac{2\lambda}{a^2}}=\frac{\frac{b^2y}{\mu^2}}{\frac{2\mu}{b^2}},$$

ou

$$\frac{a^4x}{\lambda^3}=-\frac{b^4y}{\mu^3}.$$

On en tire

$$\frac{\lambda}{(a^4x)^{\frac{1}{3}}} = -\frac{\mu}{(b^4y)^{\frac{1}{3}}},$$

et, en égalant ces deux rapports égaux à une inconnue auxiliaire ρ, nous avons

$$\lambda = \rho(a^4x)^{\frac{1}{3}}, \qquad \mu = -\rho(b^4y)^{\frac{1}{3}}.$$

Remplaçons λ et μ par ces valeurs dans les équations (1) et (2), nous obtenons

$$\frac{(ax)^{\frac{2}{3}} + (by)^{\frac{2}{3}}}{\rho} - c^2 = 0, \qquad \rho^2[(ax)^{\frac{2}{3}} + (by)^{\frac{2}{3}}] - 1 = 0.$$

Il ne nous reste plus qu'à éliminer ρ entre ces deux équations, ce qui nous donne

$$[(ax)^{\frac{2}{3}} + (by)^{\frac{2}{3}}]^3 = c^4,$$

ou, en extrayant la racine cubique,

$$a^{\frac{2}{3}}x^{\frac{2}{3}} + b^{\frac{2}{3}}y^{\frac{2}{3}} - c^{\frac{4}{3}} = 0. \tag{3}$$

Telle est l'équation ponctuelle de la développée.

388. Elle se présente sous forme irrationnelle; on peut la rendre rationnelle de la manière suivante.

On vérifie aisément l'identité

$$u^3 + v^3 + w^3 - 3uvw \equiv (u + v + w)\mathrm{P},$$

en posant

$$\mathrm{P} \equiv u^2 + v^2 + w^2 - vw - wu - uv \equiv \frac{1}{2}[(v-w)^2 + (w-u)^2 + (u-v)^2],$$

ce qui montre que si u, v, w sont réels, P est toujours positif, excepté dans le cas où $u = v = w$.

D'autre part, nous avons

$$\mathrm{A}^3 - \mathrm{B}^3 \equiv (\mathrm{A} - \mathrm{B})(\mathrm{A}^2 + \mathrm{AB} + \mathrm{B}^2),$$

et si A et B sont réels, $\mathrm{A}^2 + \mathrm{AB} + \mathrm{B}^2$ est positif.

Remplaçons dans cette identité A par $u^3 + v^3 + w^3$ et B par $3uvw$; nous obtenons

$$(u^3 + v^3 + w^3)^3 - 27u^3v^3w^3 \equiv (u^3 + v^3 + w^3 - 3uvw)\mathrm{Q},$$

Q désignant ce que devient $\mathrm{A}^2 + \mathrm{AB} + \mathrm{B}^2$ par la substitution, ou en remplaçant $u^3 + v^3 + w^3 - 3uvw$ par sa valeur précédente,

$$(u^3 + v^3 + w^3)^3 - 27u^3v^3w^3 \equiv (u + v + w)\mathrm{PQ}.$$

Faisons maintenant

$$u = a^{\frac{2}{3}}x^{\frac{2}{3}}, \qquad v = b^{\frac{2}{3}}y^{\frac{2}{3}}, \qquad w = -c^{\frac{4}{3}},$$

nous avons

$$(4)\qquad (a^2x^2+b^2y^2-c^4)^3+27a^2b^2c^4x^2y^2\equiv\left(a^{\frac{2}{3}}x^{\frac{2}{3}}+b^{\frac{2}{3}}y^{\frac{2}{3}}-c^{\frac{4}{3}}\right)\mathrm{R},$$

et R est toujours positif, car on ne peut avoir $a^{\frac{2}{3}}x^{\frac{2}{3}}=b^{\frac{2}{3}}y^{\frac{2}{3}}=-c^{\frac{4}{3}}$ pour des valeurs réelles de x et de y.

Il en résulte que les équations

$$(3)\qquad a^{\frac{2}{3}}x^{\frac{2}{3}}+b^{\frac{2}{3}}y^{\frac{2}{3}}-c^{\frac{4}{3}}=0,$$

$$(5)\qquad (a^2x^2+b^2y^2-c^4)^3+27a^2b^2c^4x^2y^2=0$$

ont mêmes ensembles de solutions réelles; par suite l'équation (5) est l'équation rationnelle de la développée.

Cette courbe est donc du sixième degré.

389. Mais pour la construire il sera plus facile d'utiliser l'équation (3).

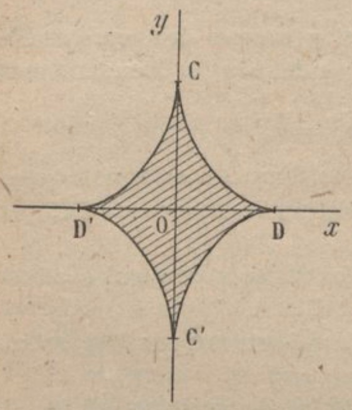

Fig. 121.

Cette équation ne change pas si on change x en $-x$ ou y en $-y$; par suite, la courbe est symétrique par rapport aux deux axes. Nous construirons seulement la portion de cette courbe qui correspond aux valeurs positives de x et de y.

Résolvons l'équation par rapport à y; nous avons

$$y=\frac{a}{b}\left[\left(\frac{c^2}{a}\right)^{\frac{2}{3}}-x^{\frac{2}{3}}\right]^{\frac{3}{2}}.$$

Pour que y soit réel, il faut que x soit plus petit que $\frac{c^2}{a}$, et quand x croît de 0 à $\frac{c^2}{a}$, y décroît de $\frac{c^2}{b}$ à 0; nous obtenons ainsi la branche de courbe CD (*fig.* 121).

D'autre part, la dérivée de y par rapport à x est

$$y'=-\frac{a}{bx^{\frac{1}{3}}}\left[\left(\frac{c^2}{a}\right)^{\frac{2}{3}}-x^{\frac{2}{3}}\right]^{\frac{1}{2}};$$

par suite, au point $\mathrm{C}\left(0,\frac{c^2}{b}\right)$ la tangente est Oy, au point $\mathrm{D}\left(\frac{c^2}{a},0\right)$, la tangente est O$x$.

On achève par symétrie, et on a la courbe représentée par la figure 121. Elle admet quatre points de rebroussement.

390. Normales par un point non situé sur l'ellipse. Détermination des pieds. — Soient α, β les coordonnées du point donné P. Désignons par x, y les coordonnées du pied d'une normale issue du point P; nous avons d'abord

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}-1=0 \tag{1}$$

puisque ce point est sur l'ellipse; en outre, en écrivant que la normale en ce point, $\dfrac{X-x}{\dfrac{x}{a^2}}=\dfrac{Y-y}{\dfrac{y}{b^2}}$, passe par le point P, nous avons

$$\frac{\alpha-x}{\dfrac{x}{a^2}}=\frac{\beta-y}{\dfrac{y}{b^2}}. \tag{2}$$

Les équations (1) et (2) déterminent les coordonnées des pieds des normales issues du point P.

Considérons dans ces équations x et y comme des coordonnées courantes, la première représente l'ellipse donnée; la deuxième peut s'écrire

$$c^2xy+b^2\beta x-a^2\alpha y=0,$$

elle représente une hyperbole équilatère dont les asymptotes sont parallèles aux axes de l'ellipse et qui passe par le centre O de l'ellipse et par le point P.

Cette hyperbole rencontre l'ellipse en quatre points qui sont les pieds des normales issues du point P à l'ellipse.

On l'appelle *l'hyperbole aux pieds des normales* ou *l'hyperbole d'Apollonius* relative au point P.

La branche d'hyperbole qui passe au point O rencontre l'ellipse en deux points réels; il en résulte que sur les quatre normales passant par le point P deux sont toujours réelles.

391. Nous allons chercher dans quel cas les deux autres sont aussi réelles.

Pour cela, nous exprimerons les coordonnées d'un point quelconque de l'hyperbole d'Apollonius en fonction d'un paramètre. Revenons à l'équation (2) et désignons par t la valeur commune des deux membres,

$$\frac{\alpha-x}{\dfrac{x}{a^2}}=\frac{\beta-y}{\dfrac{y}{b^2}}=t;$$

nous en tirons

$$x=\frac{a^2\alpha}{t+a^2},\qquad y=\frac{b^2\beta}{t+b^2}. \tag{3}$$

Ce sont les coordonnées d'un point de l'hyperbole en fonction de t.

*D ? . ! % / < (% B H! "
1 ## # # B# > ! !1 ! .. B '
E :(T ! A !

(%

Pág. 201 (210) – Fig. 89

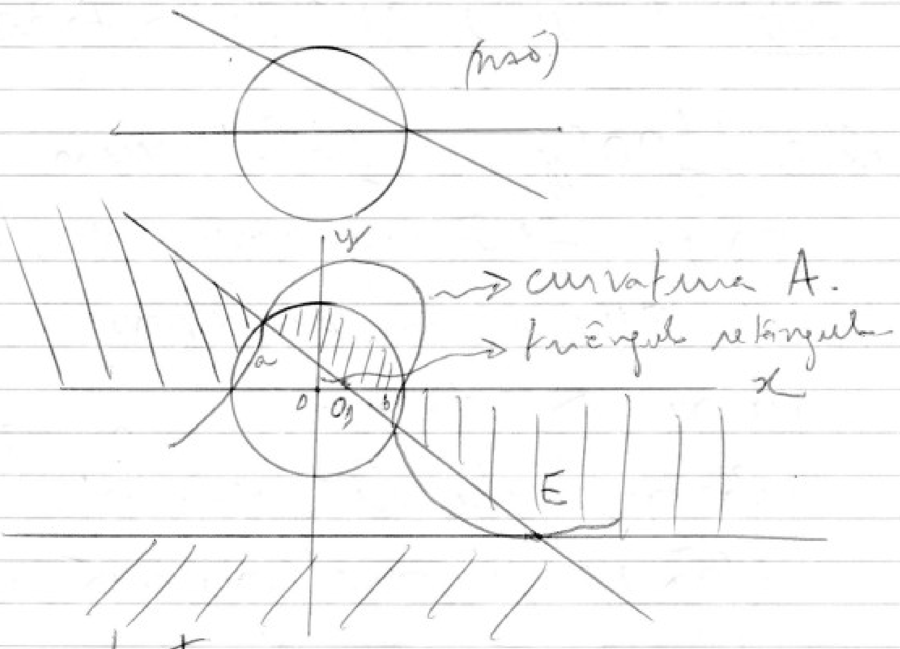
(não)
y
curvatura A.
triângulo retângulo
x
a
0
O_1
b
E

Important:

Estudar a razão entre a e b em relação ao ângulo no ponto O_1, observando se quando os dois são ângulos opostos, ou seja, tem o mesmo ângulo.

Veja como este comportamento matemático interfere na curvatura de A, simule até obter a igualdade com a curvatura tempo-espaço do universo.

Observe que O e O_1 formam um triângulo retângulo (Pitágoras).

Use tudo isto com a ellipse do caso anterior para calcular

data
S T Q Q S S D

a curvatura do Universo e que repercutiu no primeiro Mus (TT).

Obs: Estudar este livro todo!!!

20211008 10h15min

porque - Não fui ao cinema porque perdi a hora.
↳ substituir por pois, visto que...

por que ⇒ pelo qual
Por que ela não veio hoje?
Não sei por que você não consegue...
Subst: por que motivo

por quê ⇒
Você não foi à festa por quê?

porquê ⇒
As crianças não entenderam
o porquê do castigo.

In September of 476 A.D., the barbarian commander Odoacer forced the teenaged Western Roman emperor Romulus Augustus

($

to resign his office.

"Odoacer, king of the Goths, took control of Rome".

↳ the Constantinopolitan chronicler Marcellinus Comes in the 510s.

The capture of Rome by the East came 60 years, and three months after Odoacer's coup of 476.

Western Roman Empire had clearly fallen by the 560s.

smp (02)	(03)	(04)	(05)	(06)	(07)	spf						
100	160	100	120	90	110	110	160	225	180	150	170	170
105	150	130	170	200	110	150	220	170	210	190	170	
100	150	120	140	120	140	110	150	195	450	170	130	
110	120	120	120	110	220	120	180	200	130	510	120	
100	580	130	100	100	110	150	710	600	160	190	590	
100	110	600	650	620	690	640	60 02 001	80	690	700		
	690	006-2023		01.002	03 03.002			680	(P01 01 001)	(P02 07)		
159	016-2021	(26,24)	136-2020	(47.9)	(43,42)			(2019 01)				
(3,84)	20210813		(12,55)									
	(141,03)											

[illegible]	no	MVSISS		Prof4590	240	200	200	90	45
29	105	244	479	320	250	200	300	60	100
69	82	133	151	220	490	200	200	150	145
98	187	104	630	540	40	90	√		190
7	57	479		120	530	690			335
105	244		85	660	PHOTOS 5482 001 (50)				SV-64 (9,78)

data

Books:

- 20211008/ College Algebra from Basic/
College_Algebra_from_Basic....pdf
Pág. 681(696) Chapter 8
Analytic Geometry.
- 20211008/ Krushna's-Dynamics.../
Krishna's-Dynamics....pdf
Pág. D-159(164) Chapter 3
Constrained Motion on Smooth and
Rough Plane Curves
(Vertical Circle and
Cycloid)
- Algebra-Linear.pdf
Pág. 7 Chapter 1 Vector Spaces.
- Algebra Trigonometry.pdf
Pág. 187 Chapter 3
Logarithms of Complex Numbers
- Mathematics.pdf
Pág. 119 Chapter 7 the Riemann Integral.

$
! -256- - -* ! - $
I2 D !I< I)! I I2"A! "I?!A 1"
(B B/B !1 2 ! 3 7 3
D -256- - -* ! - $ + !0 43 ! J =" 1"
(4 $/ B' !1 %
! " ! " 5 ! A
! ! " 3 "
-256- - -* ! - $ 2 D ! ? ! 1"
(!1 1!
" -256- - -* ! - $ 2 D ! ; 9 3 1"
(!1 % ? ! 1 L 8 D
(
-256- - -* ! - $! ! 1"
(!1 % ! !11
-256- - -* ! - $ + !0 < A !
! !D ! " ! ! 4 ! => ! 1"
-256- - -* ! - $ + !0 A " !1
! ! G ! C== ! 1"
-256- - -* ! - $ + !0 2 ! 3 J =" B 1"
(!1 ' 9 5 ! & !
-256- - -* ! - $ + !0 1 L 2 ! 3 J
=" % 1"
(!1 1 L 8 D ! " 7 ! 5 1 !
K -256- - -* ! - $ + !0 2 D ! ; 9 3J
" =" 1"
. -256- - -* ! - $ + !0 ! "J
=") 1"
(
-256- - -* ! - $ + !0 ! J " =" 1"
(' !1 A! ! A!
*D ? A " UUU
-256- - -* ! - $ + !0 ! ;
4 ! => ! J =" B 1"
(B !1 $ 4 ! ! " & !
*D ? A " UUU
(BB !1 7 0 7! 0 ! " . 0 9
-256- - -* ! - $ + !0 = ! J
=" '= 1"
(%/ ' ! 0 9 ?!
-256- - -* ! - $ + !0 = 6 !
! J 2 ; ! 1 ! % 1"
($ H " => !
(/
1 -256- - -* ! - $ + !0 & ! => ! ;
) "! 3 ! D J =") 1"
> -256- - -* ! - $ + !0 ! ; @!A
J =") 1"
-256- - -* ! - $ 2 D ! 1"
-256- - -* ! - $! 1"

20211008(cont.)///

Pa _20211008/Krishna's.../
Mathematics.pdf (cont.)
Pág. 212 Chapter 3
Conformal Mappings

- Krishna's - Functions of Several Variables...pdf
- Krishna's - Solved Papers Mathematics..pdf
- Krishna's - Analysis, Edition-6.pdf

Pág. 178 Chapter 4
The Riemann Integral

- Krishna's - Complex Analysis, Edition-3.pdf

Pág. 8 - Chapter 1
Complex Numbers and their Geometrical Representation.

- Krishna's - Algebra Trigonometry 2nd Edition.pdf
- Krishna's - Mechanics of Solid.pdf

1B

- Krishna's - Calculus, 2nd Edition.pdf
  Pág. 114 - Chapter 7
  Curvature - Meaning of Curvature.
  Obs: Livro todo!!!
- Krishna's - Matrices Vector....pdf
  Pág. 106 - Chapter 5
  Differentiation and Integration of Vectors
  Obs: Livro todo!!!
  cont. Pág. 166 - Chapter 7
  Green's, Gauss's and Stoke's theorems.
- Krishna's - Engineering Mechanics ...4E.pdf
  Pág. 39 2.4
  Varignon's theorem (Law of Moment).
- Krishna's - Elements of Quantum Mechanics, Atomic... pdf

$ (/

Pág. 85 - Unit 2

Schrödinger Equation and its application

2.1 → One Dimensional Time Independent Schrödinger's

- Krishna's Integral Equations 27B.pdf
- Krishna's - Mechanics Wave Motion.pdf

Books:

Books-020-2021/Important/
IMPORTANT-20211008/

Obs: Diretório todo!!!

- An Expedition to Geometry...pdf

Pág. 53 - Projective Geometry

Obs: Livro todo!!!

- Calculus without Limit Almost.pdf
- Linear Algebra Illustrated.pdf

Obs: Livro todo!!!

$

-V& %-4292 8 A A $,-4292 8 A A -) .
- -& *59289 -
*D 4 W " UUU

! -256- - -* ! - $ 2 =L1 " 7 3
: " !) . 2 3 $ 1"
($% K A 7 3

D -256- - -* ! - $! @ ? 2 1"
(' : 11 ! 0 ? ! ")

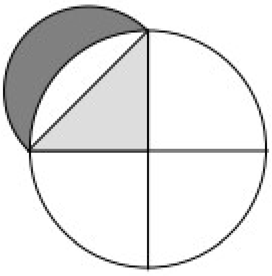

Hippocrates' Lune: Circa 440BC

This is the earliest known geometric figure having two curvilinear boundaries for which a planar area could be exactly determined.

(
-256- - -* ! - $? ! 2 D ! & ! " 1"
*D ? A " UUU
(% % 9 ! !L ! 1 !

" -256- - -* ! - $ + ! ! 2
5 ! ! 7 3 9! ! & // 1"
-256- - -* ! - $ + ! ! 2 5 ! !
7 3 9! ! & // "KA
(' $$ ' 7! 9 3 ! 5X%
($$ B% $ A! ! " ! ! 9 ! 1
($ B ' 7! 9 ! =

-256- - -* ! - $ + ! ! 9 1 3
1! 2 1 ! & ! ! J ? " $ 1"
(%$!1 B 1 1!

-256- - -* ! - $ + ! ! 2
4 ! 7 3 ! " ? 7 1 : " !) . 2 3 1"
*D ? A " UUU

Pág. 177 3.3
Transforming the axes transforms the plane.

— S. Kumaresan – A Course in Riemannian Geometry.pdf
Pág. 46 2.4
Gauss Theory of Surfaces in $\mathbb{R}^3$.
Pág. 54 2.5
Curvature and Parallel Transport

— S. Kumaresan – Topology of Metric Space.pdf
Pág. 122 (135) Chapter 6
Complete Metric Spaces

— S. Kumaresan – A Course in Differential Geometry...pdf
Obs. Livro todo!!!

$ (
4292 8 A A $,-4292 8 A A -) .
- -2 !-
4292 8 A A $,-4292 8 A A -) .
- /-2 !-

! -256- - -* ! - $ 9 5! ! ! < L D
= !) " B%) . Y)! /B 1"
(B 25&*99= 284 ?=&)8&V *8 =?2 9&)=2 B '

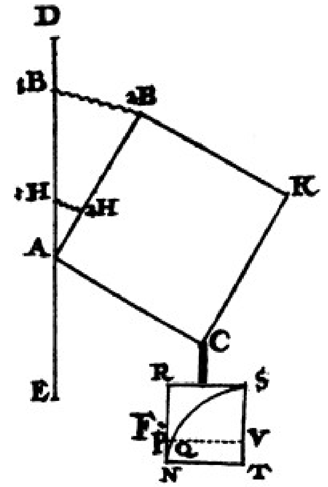
D
1B
2B
K
1H
2H
A
C
R
S
E
F
P
Q
V
N
T

Figure 19. Leibniz's figure for calculating the bending moment acting on the cross-section of a terminally loaded beam (1684)

D -256- - -* ! - $ 2 < 5! !
! 7 ! 1 2 !" & // 1"
*D ? A " UUU

20211009 ///

Books:

— 20211009/ Astronomics/
The Rational Mechanics of Flexible or Elastic Bodies ...(1960).pdf

Pag. 61 Mariotte and Leibniz on Elastic Beams (1684)

Figure 19.

Leibniz's figure for calculating the bending moment acting on the cross-section of a terminally load beam (1684)

— A first course in rational continuum mechanics – Volume 1 ... (1977).pdf

Obs: Livro todo!!!

04	05	06	07	08	zip.1	
130	100	210	100	250	190	Class 15 Test 2..mp4
210	100	210	120	60	210	(86, 9)
100	190	110	140	310	400	Pro:
100	105	530	110	150	210	Class 16
110	100	90	130	220	610	—
40	595	620	600	680	70	Video-Math
690	5V78	60	100	(201812)	680	705 files
74-10	20201220	680	700			
(17,31)	(3,28)	5V78 20210102 (12,05)	86-70 20210802 (50)			

data

Can a Circle be a Straight Line
Space Time PBS

—

69,8 (20211009)
8,6
78,4 TB
31,24
109,64 TB

— Terra - $1,26^{21}$ l H_2O

—

Books:

- 20211009/Numerical Geometry.../
3030767973.pdf
Pág. 37(53) -
1 The Gauss Integral for the Linking Number of Curves
- 20211009/Comparison Finsler Geometry/Ohta2021_Book_ComparisonFinslerGeometry.pdf
Pág. 29(50) Lemma 3.18
(Gauss's Lemma)

$%
! -256- - -* ! - $% % % B / % 1"
(% $% 9 7! & ! ? . 8 D A

D -256- - -* ! - $%
* ! I) .I 1! < 7 3 1"
(/ $? ! % 7! 0 ? !

We may write $B^+(x,r)$ instead of $B^+_M(x,r)$ for simplicity (when the space in question is transparent), or $B^+_F(x,r)$ to stress that the distance structure is induced from F. Open and closed balls in T_xM are defined in a similar way by using the (asymmetric) distance function $d_{T_xM}(v,w) := F(w-v)$.

Proposition 3.17 (Properties of the Exponential Map)

(i) *The domain* $\mathcal{D}$ *of the exponential map* exp *is an open set including the zero section* $\mathbf{0}$.
(ii) *For any* $x \in M$*, there exist* $\varepsilon > 0$ *and an open neighborhood* U *of* x *such that any two points in* U *are joined by a unique solution to the geodesic equation with length less than* ε.
(iii) *The exponential map* exp *is* C^1 *on* $\mathcal{D}$*, and its restriction to* $\mathcal{D} \setminus \mathbf{0}$ *is* C^∞.
(iv) *Given* $x \in M$*, for sufficiently small* $\varepsilon > 0$*, the map* $\exp_x : B^+_{T_xM}(0,\varepsilon) \longrightarrow B^+_M(x,\varepsilon)$ *is a* C^1*-diffeomorphism and its restriction* $\exp_x : B^+_{T_xM}(0,\varepsilon) \setminus \{0\} \longrightarrow B^+_M(x,\varepsilon) \setminus \{x\}$ *is a* C^∞*-diffeomorphism.*

These assertions essentially follow from the ODE theory. In (iii) and (iv), by construction, we find that the derivative $d(\exp_x)_0 : T_0(T_xM) \longrightarrow T_xM$ of $\exp_x$ at $0 \in T_xM$ is the identity map on T_xM (by identifying its domain $T_0(T_xM)$ with T_xM); we refer to [25, Sect. 5.3] for details. The exponential map becomes C^2 at the zero section $\mathbf{0}$ if and only if F belongs to the special class of Berwald metrics (see Proposition 6.11).

To prove that solutions of the geodesic equation are locally minimizing, we can follow the standard line as follows.

Lemma 3.18 (Gauss' Lemma) *Let* $v \in U_xM$ *and* $\eta : [0,l] \longrightarrow M$ *be the solution to the geodesic equation* (3.15) *with* $\dot\eta(0) = v$*. For* $t \in (0,l)$ *and* $w \in T_v(T_xM)$ *tangent to* U_xM*, we have*

$$g_{\dot\eta(t)}\big(\dot\eta(t), d(\exp_x)_{tv}(w)\big) = 0,$$

where we regard as $w \in T_{tv}(T_xM)$ *by identifying both* $T_v(T_xM)$ *and* $T_{tv}(T_xM)$ *with* T_xM *in the canonical way.*

Proof Fix $T \in (0,l)$ and, for sufficiently small $\varepsilon > 0$, define a variation $\sigma : [0,T] \times (-\varepsilon,\varepsilon) \longrightarrow M$ by $\sigma(t,s) := \exp_x(t(v+sw))$. Then we have on one hand

$$\mathcal{E}\big(\sigma(\cdot,s)\big) = \frac{T}{2} F^2(v+sw),$$

which takes its minimum at $s = 0$ since w is tangent to $U_x M$. On the other hand, observe from the calculation in (3.12) and the geodesic equation (3.15) that

$$\frac{d}{ds}\Big[\mathcal{E}\big(\sigma(\cdot, s)\big)\Big]_{s=0} = \sum_{i,j=1}^{n} g_{ij}\big(\dot\eta(T)\big)\partial_s \sigma^i(T,0) \cdot \dot\eta^j(T)$$
$$= g_{\dot\eta(T)}\big(d(\exp_x)_{Tv}(Tw), \dot\eta(T)\big).$$

This shows the claim at $t = T$, and completes the proof. □

Note that $w \in T_v(T_x M)$ being tangent to $U_x M$ is equivalent to $g_v(v, w) = 0$.

Exercise 3.19 Let $\varepsilon > 0$ and $U \subset M$ be given as in Proposition 3.17(ii), $y, z \in U$, and $\eta : [0, 1] \longrightarrow M$ be the unique solution to the geodesic equation satisfying $\eta(0) = y$, $\eta(1) = z$ and $L(\eta) < \varepsilon$. Then, for any C^1-curve $\xi : [0, 1] \longrightarrow M$ from y to z, show that we have $L(\eta) \le L(\xi)$ and equality holds only if ξ is a reparametrization of η.

We may consider the spherical coordinates about y and make use of Lemma 3.18 (see [66, Sect. I.6] for the Riemannian case). By a similar argument to the characterization of equality in the above exercise, one can also find that a geodesic in the sense of Definition 2.17 is necessarily a solution to the geodesic equation.

3.5 Completenesses and the Hopf–Rinow Theorem

Here we introduce completeness conditions and prove the Hopf–Rinow theorem.

Definition 3.20 (Completenesses) We say that (M, F) is *forward complete* if $\mathcal{D} = TM$, i.e., if the exponential map is defined on the whole tangent bundle TM. If the reverse Finsler structure $(M, \overleftarrow{F})$ (recall Definition 2.19) is forward complete, then (M, F) is said to be *backward complete*.

We remark that the forward completeness is not necessarily equivalent to the backward completeness, when the *reversibility constant*

$$\Lambda_F := \sup_{v \in TM \setminus \mathbf{0}} \frac{F(-v)}{F(v)} \in [1, \infty] \tag{3.17}$$

is infinite. For instance, the Funk metric appearing in Sect. 6.5 is forward complete but not backward complete (see Exercise 6.22).

In order to state the Finsler analogue of the *Hopf–Rinow theorem*, we define two notions for asymmetric metric spaces. We say that a set $A \subset M$ is *forward bounded* if $A \subset B^+_M(x, r)$ for some $x \in M$ and $r > 0$. A sequence $(x_i)_{i \in \mathbb{N}}$ in M is called a *forward Cauchy sequence* if, for any $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such

(/

Chapter 8
Some Comparison Theorems

Comparison theorems are the main subjects of this book. They are concerned with quantitative or qualitative properties of a space with a certain condition on its curvature. In this book, we will consider Finsler manifolds whose flag or (weighted) Ricci curvature is bounded from below or above by a constant. This chapter is devoted to some fundamental examples of geometric comparison theorems. The first two of them (the *Bonnet–Myers* and *Cartan–Hadamard theorems*) are verbatim analogues of the Riemannian counterparts. Then we study the convexity and concavity of the distance function using some non-Riemannian quantities besides the flag curvature. We refer to [66] for the Riemannian case.

8.1 The Bonnet–Myers Theorem

The Riemannian characterization of the Finsler curvature as in Theorem 5.12 has some straightforward applications in comparison geometry. We discuss two such applications in this and the next sections. Although they can be essentially reduced to the Riemannian case, we give detailed proofs for the sake of self-containedness.

We will say that $\mathrm{Ric} \ge K$ holds for some constant $K \in \mathbb{R}$ if $\mathrm{Ric}(v) \ge KF^2(v)$ holds for all $v \in TM$ (or, equivalently, $\mathrm{Ric}(v) \ge K$ for all unit vectors $v \in UM$). The *diameter* of M is defined by

$$\mathrm{diam}(M) := \sup\{d(x, y) \mid x, y \in M\}.$$

Theorem 8.1 (Bonnet–Myers Theorem (Unweighted)) *Assume that* (M, F) *is forward or backward complete and satisfies* $\mathrm{Ric} \ge K$ *for some* $K > 0$*. Then we have*

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2021

S.-i. Ohta, *Comparison Finsler Geometry*, Springer Monographs in Mathematics,
https://doi.org/10.1007/978-3-030-80650-7_8

$$\mathrm{diam}(M) \le \pi\sqrt{\frac{n-1}{K}}.$$

In particular, M is compact and has finite fundamental group.

Proof We shall prove by contradiction. Assume that there are $x, y \in M$ with $l := d(x, y) > \pi\sqrt{(n-1)/K}$. By the Hopf–Rinow theorem (Theorem 3.21), there exists a unit vector $v \in U_x M$ such that $\eta(t) := \exp_x(tv)$ is a minimal geodesic from x to $y = \eta(l)$. We follow the same lines as the Riemannian case to see that there necessarily exists a conjugate point $\eta(t_0)$ of $\eta(0)$ for some $t_0 \in (0, \pi\sqrt{(n-1)/K}]$.

Choose an orthonormal basis $\{e_i\}_{i=1}^{n-1} \cup \{v\}$ of $(T_x M, g_v)$, and consider the vector fields $E_i(t) := d(\exp_x)_{tv}(te_i) \in T_{\eta(t)}M$ along η for $i = 1, 2, \dots, n-1$. Note that E_i is the variational vector field of the geodesic variation

$$\sigma_i(t, s) := \exp_x\bigl(t(v + se_i)\bigr), \qquad (t, s) \in [0, l) \times (-\varepsilon, \varepsilon).$$

Hence E_i is a Jacobi field (for both F and $g_{\dot\eta}$; recall Proposition 4.6), and we deduce from Gauss' lemma (Lemma 3.18) that $g_{\dot\eta}(\dot\eta, E_i) = 0$ for all i. Then, since $l \le \rho(v)$, we find that $\{E_i(t)\}_{i=1}^{n-1}$ is a basis of the orthogonal complement of $\dot\eta(t)$ in $(T_{\eta(t)}M, g_{\dot\eta})$ for every $t \in (0, l)$ (similarly to the proof of Lemma 7.13).

For brevity, we will denote by E_i' the covariant derivative $D_{\dot\eta}^{\dot\eta} E_i$ along η. Observe from $g_{\dot\eta}(\dot\eta, E_i) = 0$ and Lemma 4.9 that $g_{\dot\eta}(\dot\eta, E_i') = 0$. We define $(n-1)\times(n-1)$-matrices $A(t) = (a_{ij}(t))$ and $B(t) = (b_{ij}(t))$ for $t \in (0, l)$ by

$$a_{ij}(t) = g_{\dot\eta}\bigl(E_i(t), E_j(t)\bigr), \qquad E_i'(t) = \sum_{j=1}^{n-1} b_{ij}(t) E_j(t).$$

We summarize some fundamental properties of A and B in the next claim.

Claim 8.2

(i) We have $BA = AB^{\mathsf{T}}$ and $A' = 2BA$, where B^{T} denotes the transpose matrix of B.

(ii) The matrix $A^{-1/2}BA^{1/2}$ is symmetric. Here we set

$$A^{1/2} := P^{-1}\begin{pmatrix} \sqrt{a_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{a_{n-1}} \end{pmatrix} P$$

and $A^{-1/2} := (A^{1/2})^{-1}$, where $a_1, \dots, a_{n-1} > 0$ are the eigenvalues of A and P is an orthogonal matrix which diagonalizes A as

$$A = P^{-1} \begin{pmatrix} a_1 & & 0 \\ & \ddots & \\ 0 & & a_{n-1} \end{pmatrix} P.$$

(iii) We have the *Riccati equation*

$$B' + B^2 + RA^{-1} = 0, \tag{8.1}$$

where the matrix $R(t) = (R_{ij}(t))$ is defined by $R_{ij}(t) := g_{\dot\eta}(R_{\dot\eta}(E_i), E_j)(t) = g_{\dot\eta}(E_i, R_{\dot\eta}(E_j))(t)$.

Proof (i) Observe from Lemma 4.9 that $A' = BA + AB^{\mathsf{T}}$. Moreover, it follows from the Jacobi equation (5.3) and Lemma 5.8 that

$$\begin{aligned} \frac{d}{dt}\Big[g_{\dot\eta}(E_i', E_j) - g_{\dot\eta}(E_i, E_j')\Big] &= g_{\dot\eta}(E_i'', E_j) - g_{\dot\eta}(E_i, E_j'') \\ &= -g_{\dot\eta}\big(R_{\dot\eta}(E_i), E_j\big) + g_{\dot\eta}\big(E_i, R_{\dot\eta}(E_j)\big) = 0. \end{aligned}$$

Combining this with $E_i(0) = E_j(0) = 0$, we find that $g_{\dot\eta}(E_i', E_j) = g_{\dot\eta}(E_i, E_j')$ (similarly to the proof of Lemma 7.13), which yields $BA = AB^{\mathsf{T}}$. Thus we have $A' = 2BA$.

(ii) It follows from $BA = AB^{\mathsf{T}}$ that BA is symmetric. Since A is symmetric, we find that $A^{-1/2}(BA)A^{-1/2} = A^{-1/2}BA^{1/2}$ is also symmetric.

(iii) On one hand, we deduce from

$$a_{ij}''(t) = -g_{\dot\eta}\big(R_{\dot\eta}(E_i), E_j\big) - g_{\dot\eta}\big(E_i, R_{\dot\eta}(E_j)\big) + 2g_{\dot\eta}(E_i', E_j'),$$

Lemma 5.8 and (i) that

$$A'' = -2R + 2BAB^{\mathsf{T}} = -2R + 2B^2A.$$

On the other hand, it follows from (i) that

$$A'' = 2B'A + 2BA' = 2B'A + 4B^2A.$$

Comparing these expressions of A'', we obtain the claim. □

Now, we consider the function $h(t) := (\det[A(t)])^{1/(2(n-1))}$ (which can be regarded as the *area form* of the sphere $\partial B^+(x,t)$ at $\eta(t)$ with respect to $g_{\dot\eta}$). Observe from $A' = 2BA$ that

$$(n-1)h' = \frac{h}{2}(\det A)^{-1}(\det A)' = \frac{h}{2}\operatorname{trace}(A'A^{-1}) = h\operatorname{trace}(B).$$

Moreover, by (8.1), we have

$$(n-1)h'' = h' \operatorname{trace}(B) + h \operatorname{trace}(B')$$
$$= \frac{h}{n-1} \operatorname{trace}(B)^2 - h \operatorname{trace}(B^2) - h \operatorname{trace}(RA^{-1}).$$

Since $A^{-1/2}BA^{1/2}$ is symmetric, we deduce from the Cauchy–Schwarz inequality that

$$\operatorname{trace}(B)^2 = \operatorname{trace}(A^{-1/2}BA^{1/2})^2 \le (n-1)\operatorname{trace}(A^{-1/2}B^2A^{1/2})$$
$$= (n-1)\operatorname{trace}(B^2).$$

Furthermore, it follows from the definitions of A and R that

$$\operatorname{trace}(RA^{-1}) = \operatorname{trace}(A^{-1/2}RA^{-1/2}) = \operatorname{Ric}(\dot\eta).$$

Thus we obtain a Finsler analogue of the *Bishop inequality* asserting

$$h''(t) \le -\frac{\operatorname{Ric}(\dot\eta(t))}{n-1}h(t). \tag{8.2}$$

If $\operatorname{Ric} \ge K > 0$, then (8.2) yields

$$h''(t) \le -\frac{K}{n-1}h(t) \tag{8.3}$$

for $t \in (0, l)$. We compare this differential inequality with

$$\mathbf{s}''(t) = -\frac{K}{n-1}\mathbf{s}(t), \qquad \text{where} \quad \mathbf{s}(t) := \sin\left(\sqrt{\frac{K}{n-1}}t\right).$$

Put $f := h/\mathbf{s}$ on $(0, \pi\sqrt{(n-1)/K})$. Observe that $f' = (h'\mathbf{s} - h\mathbf{s}')/\mathbf{s}^2$ and

$$(h'\mathbf{s} - h\mathbf{s}')' = h''\mathbf{s} - h\mathbf{s}'' \le 0$$

by (8.3). Moreover, we have $h(t) = O(t)$ as $t \to 0$ and $\lim_{t\to 0} h'(t) = h'(0) = 1$ (we leave $h'(0) = 1$ as an exercise below). Combining these, we find that f is non-increasing. Hence $h(t) \le (h(\varepsilon)/\mathbf{s}(\varepsilon)) \cdot \mathbf{s}(t)$ holds for $t > \varepsilon > 0$. However, since $\mathbf{s}(\pi\sqrt{(n-1)/K}) = 0$, this implies that $h(t_0) = 0$ necessarily holds at some $t_0 \in (0, \pi\sqrt{(n-1)/K}]$. Then $\eta(t_0)$ is a conjugate point of $\eta(0)$, and η is no longer minimizing beyond t_0 (recall Proposition 7.16). This contradicts the choice of v, therefore we conclude that $\operatorname{diam}(M) \le \pi\sqrt{(n-1)/K}$.

The compactness of M follows from the Hopf–Rinow theorem (Theorem 3.21). To show the assertion on the fundamental group, let us consider the universal cover $\widetilde{M}$ of M equipped with the induced Finsler structure $\widetilde{F}$. Then $\widetilde{F}$ again satisfies $\operatorname{Ric} \ge K$, and $(\widetilde{M}, \widetilde{F})$ is forward (and backward) complete since geodesics in M

can be lifted to $\widetilde{M}$. Therefore, $\widetilde{M}$ is also compact, and hence the fundamental group of M is finite. □

Exercise 8.3 Show that $h'(0) = 1$ holds in the above proof.

The Bonnet–Myers theorem for the weighted Ricci curvature will be discussed in Sect. 9.4. This is the reason why we called Theorem 8.1 the *unweighted* one.

Remark 8.4 (Bishop Inequality) What is commonly called the Bishop inequality (or *Bishop's comparison theorem*) in Riemannian geometry is an upper bound of the volume of a ball by that in a model space of constant curvature (see [36, Sect. 11.10] and, e.g., [66, Theorem III.4.4]). A generalization to a Berwald space equipped with the Busemann–Hausdorff measure will be given in Corollary 10.3. Since the inequality (8.2) is an essential step from which the volume comparison can be deduced rather straightforwardly, we call the inequality (8.2) (and its generalizations) a Bishop inequality.

Remark 8.5 (Reversing Stability) In the proof of Theorem 8.1, we used the Hopf–Rinow theorem for finding a minimal geodesic from x to y and for showing that the finiteness of the diameter implies the compactness. For these assertions, it is sufficient to assume either the forward completeness or the backward completeness. Moreover, if we consider the reverse Finsler structure $\overleftarrow{F}$, then we have $\overleftarrow{\mathrm{Ric}}(v) = \mathrm{Ric}(-v)$ (see Sect. 2.5) and hence the curvature bound $\mathrm{Ric}(v) \ge KF^2(v)$ is equivalent to $\overleftarrow{\mathrm{Ric}}(v) \ge K\overleftarrow{F}^2(v)$. Therefore, Theorem 8.1 is stable when we take the reverse structure.

Finally we explain how to reduce Theorem 8.1 to the Riemannian case. Thanks to Remark 8.5 above, we may assume the forward completeness. Fix arbitrary $x \in M$ and define a vector field V on $M \setminus (\{x\} \cup \mathrm{Cut}(x))$ by $V(\eta_v(t)) := \dot{\eta}_v(t)$ for $t \in (0, \rho(v))$, along each unit speed geodesic $\eta_v : [0, \rho(v)) \longrightarrow M$ with $\dot{\eta}_v(0) = v \in U_xM$. This vector field V is indeed well-defined, and all of its integral curves are geodesic. Therefore, g_V enjoys $\mathrm{Ric}^{g_V}(V) \ge K$ by Theorem 5.12, and the Riemannian Bonnet–Myers theorem asserts that $\rho(v) \le \pi\sqrt{(n-1)/K}$ for all $v \in U_xM$. Hence we obtain $\mathrm{diam}(M) \le \pi\sqrt{(n-1)/K}$.

8.2 The Cartan–Hadamard Theorem

By analyzing Jacobi fields again, we can show another fundamental comparison theorem, under the nonpositive flag curvature this time.

Theorem 8.6 (Cartan–Hadamard Theorem) *Assume that (M, F) is forward complete and satisfies $\mathbf{K} \le 0$. Then we have the following:*

(i) *Any geodesic does not have conjugate points.*
(ii) *For any $x \in M$, the exponential map $\exp_x : T_xM \longrightarrow M$ is a C^1-covering.*

(/ ' !1

Chapter 10
Examples of Measured Finsler Manifolds

In this chapter, we analyze the weighted Ricci curvature for some examples appearing in Chap. 6. We remark that a suitable choice of a measure is unclear in some cases. In fact, the theory of weighted Ricci curvature has so far been focused on general theory and there are not many investigations on concrete examples.

One of the most important observations in this chapter is the absence of "canonical measures" in some Randers spaces. Precisely, we will see that some Randers spaces do not admit any measure whose **S**-curvature vanishes identically (see Remark 10.11).

10.1 Minkowski Normed Spaces

Let $(\mathbb{R}^n, \|\cdot\|)$ be a Minkowski normed space (recall Example 2.12(a), Sect. 6.1) and consider a measure $\mathfrak{m} = \mathrm{e}^{-\psi}\,\mathfrak{L}_n$ with $\psi \in C^\infty(\mathbb{R}^n)$. Then the weighted Ricci curvature of $(\mathbb{R}^n, \|\cdot\|, \mathfrak{m})$ is given by

$$\mathrm{Ric}_N(v) = (\psi \circ \eta_v)''(0) - \frac{(\psi \circ \eta_v)'(0)^2}{N-n}$$

for $v \in T_x\mathbb{R}^n$, where $\eta_v(t) := x + tv$ for $t \in \mathbb{R}$ (by identifying $T_x\mathbb{R}^n$ and $\mathbb{R}^n$). We also find that $\mathbf{S}(v) = (\psi \circ \eta_v)'(0) = d\psi(v)$ (recall Remark 9.14), and hence the **S**-curvature is identically 0 if and only if ψ is a constant function (see Exercise 10.7 below for a related fact on general measured Finsler manifolds).

Exercise 10.1 Prove the above expression of the weighted Ricci curvature $\mathrm{Ric}_N(v)$.

In the case of $N = \infty$, the weighted Ricci curvature Ric_∞ is directly related to the convexity (or concavity) of ψ. If ψ is K-convex in the sense that $(\psi \circ \eta_v)''(0) \geq K\|v\|^2$ for all $v \in T\mathbb{R}^n$, then we have $\mathrm{Ric}_\infty \geq K$. When $K = 0$ (i.e., ψ

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2021

S.-i. Ohta, *Comparison Finsler Geometry*, Springer Monographs in Mathematics,
https://doi.org/10.1007/978-3-030-80650-7_10

is convex), we encounter a log-concave measure as in Remark 9.9(c). Another important example is a Gaussian-type measure

$$\mathfrak{m}(dx) = \mathrm{e}^{-K\|x\|^2/2}\, \mathfrak{L}_n(dx)$$

with $K > 0$, which satisfies $\mathrm{Ric}_\infty \ge K\mathsf{C}^{-1}$ for the uniform convexity constant C of $(\mathbb{R}^n, \|\cdot\|)$ as in (1.2). Here, precisely, $\mathrm{Ric}_\infty \ge K\mathsf{C}^{-1}$ is understood in the weak sense (around the origin) since the weight function $\psi(x) = K\|x\|^2/2$ is only C^1 at the origin (recall Subsect. 1.2.2).

10.2 Berwald Spaces

We next consider Berwald spaces (recall Sect. 6.3). In this case, the Busemann–Hausdorff measure can be regarded as a canonical measure since its **S**-curvature vanishes identically. Recall Definition 9.1 for the definition of the Busemann–Hausdorff measure.

Proposition 10.2 (S-Curvature Vanishes for $\mathfrak{m}_{\mathrm{BH}}$) *Let $(M, F, \mathfrak{m}_{\mathrm{BH}})$ be a Berwald space equipped with the Busemann–Hausdorff measure. Then the corresponding* **S***-curvature is identically* 0.

Proof Given $v \in U_xM$, let $\eta : (-\varepsilon, \varepsilon) \longrightarrow M$ be the geodesic with $\dot\eta(0) = v$. Fix an orthonormal basis $\{e_i\}_{i=1}^{n-1} \cup \{v\} \subset T_xM$ with respect to g_v, and let P_i be the parallel vector field along η with $P_i(0) = e_i$ for $i = 1, 2, \dots, n-1$. (Recall that the covariant derivative is independent of the choice of a reference vector in Berwald spaces.) We consider a local coordinate system around x given by

$$\exp_{\eta(t)}\left(\sum_{i=1}^{n-1} a_i P_i(t)\right) \longmapsto (a_1, a_2, \dots, a_{n-1}, t) \in \mathbb{R}^n.$$

We remark that the exponential map $\exp_{\eta(t)}$ is C^∞ on a neighborhood of 0 by Proposition 6.11. We shall compare $\mathrm{vol}_{g_{\dot\eta}}$ and $\mathfrak{m}_{\mathrm{BH}}$ along η by using this coordinate system.

On one hand, observe that

$$g_{ij}\big(\dot\eta(t)\big) = g_{\dot\eta}\left(\frac{\partial}{\partial x^i}\bigg|_{\eta(t)}, \frac{\partial}{\partial x^j}\bigg|_{\eta(t)}\right) = g_{\dot\eta}\big(P_i(t), P_j(t)\big),$$

where we set $P_n(t) := \dot\eta(t)$. Since P_i is a parallel vector field for all $1 \le i \le n$, we have $g_{ij}(\dot\eta(t)) = \delta_{ij}$ for all t. Therefore we obtain $\mathrm{vol}_{\dot\eta}(dx) = dx^1 dx^2 \cdots dx^n$ along η. On the other hand, since the parallel transport along η preserves F (by Proposition 6.5; here we essentially need the Berwald condition), we find that

$$= \int_{\Omega} d\phi(\nabla u)\, d\mathfrak{m}. \tag{11.12}$$

If u is a minimizer of $\mathcal{E}_{\Omega}$ in the sense above, then it follows from (11.12) that $\int_{\Omega} d\phi(\nabla u)\, d\mathfrak{m} = 0$ necessarily holds for all $\phi \in H_0^1(\Omega)$. Hence, u is harmonic.

Conversely, let u be harmonic and suppose that there is some $\phi \in H_0^1(\Omega)$ such that $\mathcal{E}_{\Omega}(u+\phi) < \mathcal{E}_{\Omega}(u)$. Then the convexity of $\mathcal{E}_{\Omega}$ as in (11.3) implies

$$\mathcal{E}_{\Omega}(u+t\phi) \le (1-t)\mathcal{E}_{\Omega}(u) + t\mathcal{E}_{\Omega}(u+\phi)$$

for any $t \in (0,1)$, and hence

$$\frac{d}{dt}\big[\mathcal{E}_{\Omega}(u+t\phi)\big]_{t=0} \le \mathcal{E}_{\Omega}(u+\phi) - \mathcal{E}_{\Omega}(u) < 0.$$

This contradicts the harmonicity of u due to (11.12). Therefore u is a minimizer of $\mathcal{E}_{\Omega}$ as desired. □

11.3 Laplacian Comparison Theorem

In this section, we show the *Laplacian comparison theorem* as an analytic counterpart to the (directional) area comparison (9.8) (see Remark 11.21 below). This was shown in [206, Theorem 5.2] via a reduction to the Riemannian case, and here we give a direct proof by a calculation similar to Theorem 9.15. We refer to, e.g., [155, Theorem 4.1] for the Riemannian case.

Theorem 11.20 (Laplacian Comparison) *Let $(M, F, \mathfrak{m})$ be forward or backward complete and assume $\mathrm{Ric}_N \ge K$ for some $K \in \mathbb{R}$ and $N \in [n, \infty)$. Then, for any $z \in M$, the distance function $u(x) := d(z,x)$ from z satisfies*

$$\boldsymbol{\Delta} u(x) \le (N-1)\frac{\mathbf{s}'_{K/(N-1)}(d(z,x))}{\mathbf{s}_{K/(N-1)}(d(z,x))} \tag{11.13}$$

pointwise on $M \setminus (\{z\} \cup \mathrm{Cut}(z))$, and in the weak sense on M.

Recall (9.4) for the definition of $\mathbf{s}_{K/(N-1)}$, and Definition 7.10 for the definition of the cut locus $\mathrm{Cut}(z)$. For example, when $K = 0$, the inequality (11.13) means

$$\boldsymbol{\Delta} u(x) \le \frac{N-1}{d(z,x)}.$$

We remark that, in the case of $K > 0$, $\mathrm{diam}(M) \le \pi\sqrt{(N-1)/K}$ holds by the Bonnet–Myers theorem (Corollary 9.17), and hence $d(z,x) = \pi\sqrt{(N-1)/K}$ can be achieved only when x is a cut point of z.

Proof As in the proof of Theorem 9.15, we fix a unit vector $v \in U_zM$ and put $\eta(t) := \exp_z(tv)$. We also take an orthonormal basis $\{e_i\}_{i=1}^{n-1} \cup \{v\}$ of (T_zM, g_v) and define $E_i(t) := d(\exp_z)_{tv}(te_i)$ for $t \in [0, \rho(v))$ and $i = 1, 2, \ldots, n-1$.

Now, let us consider polar coordinates $(x^i)_{i=1}^n$ on a neighborhood of $\eta((0, \rho(v)))$ such that $x^n = u$ and $(\partial/\partial x^i)|_{\eta(t)} = E_i(t)$ for $t \in (0, \rho(v))$ (one may recall Gauss' lemma (Lemma 3.18)). Then we have $\boldsymbol{\nabla} u = \partial/\partial x^n$ and, by (11.9),

$$\boldsymbol{\Delta} u\big(\eta(t)\big) = -\psi_\eta'(t) + \frac{d}{dt}\left[\log\left(\sqrt{\det\big[g_{ij}\big(\dot{\eta}(t)\big)\big]}\right)\right].$$

Notice that

$$g_{ij}\big(\dot{\eta}(t)\big) = g_{\dot{\eta}(t)}\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right) = g_{\dot{\eta}(t)}\big(E_i(t), E_j(t)\big),$$

and consider the function $h := (\det[g_{ij}(\dot{\eta})])^{1/(2(n-1))}$ as in the proofs of Theorems 8.1 and 9.15. Then we find

$$\boldsymbol{\Delta} u\big(\eta(t)\big) = -\psi_\eta'(t) + \frac{(h^{n-1})'}{h^{n-1}}(t) = \frac{(\mathrm{e}^{-\psi_\eta} h^{n-1})'}{\mathrm{e}^{-\psi_\eta} h^{n-1}}(t) = \frac{f'}{f}(t), \tag{11.14}$$

where $f(t) := \mathrm{e}^{-\psi_\eta(t)} h(t)^{n-1}$. In the proof of Theorem 9.15, we showed

$$\left(\frac{f(t)}{\mathbf{s}_{K/(N-1)}(t)^{N-1}}\right)' \le 0$$

as a consequence of the weighted Bishop inequality (9.6) under $\mathrm{Ric}_N \ge K$. This is equivalent to

$$\frac{f'}{f}(t) \le (N-1)\frac{\mathbf{s}_{K/(N-1)}'}{\mathbf{s}_{K/(N-1)}}(t), \tag{11.15}$$

which yields the claim on $M \setminus (\{z\} \cup \mathrm{Cut}(z))$.

In order to extend (11.13) to M in the weak sense, let $\phi \in C_c^\infty(M)$ be an arbitrary nonnegative function with compact support. Then the claim is

$$-\int_M d\phi(\boldsymbol{\nabla} u)\, d\mathfrak{m} \le \int_M \phi(x)\Theta\big(d(z,x)\big)\, \mathfrak{m}(dx),$$

where we denote by $\Theta(d(z,x))$ the right-hand side of (11.13). We remark that $\Theta(d(z,\cdot))$ is integrable around z since $\Theta(t) = O(t^{-1})$ as $t \to 0$ and $n \ge 2$. Given a unit vector $v \in U_zM$, let $f_v : (0, \rho(v)) \longrightarrow (0, \infty)$ be the function f as in the first part of this proof associated with v. Recall that

$$d\mathfrak{m} = f_v(t)\, dt\, \mathbf{A}_z(dv)$$

on $M \setminus (\{z\} \cup \mathrm{Cut}(z)) = \{\exp_z(tv) \mid v \in U_zM,\ t \in (0, \rho(v))\}$, where $\mathbf{A}_z$ is the measure on U_zM induced from g_v as in the proof of Theorem 9.15. Thanks to this decomposition of $\mathfrak{m}$ and $f_v'(t) \le \Theta(t) f_v(t)$ from (11.15), we have

$$\begin{aligned}\int_M \phi(x)\Theta\bigl(d(z,x)\bigr)\,\mathfrak{m}(dx) &= \int_{U_zM}\int_0^{\rho(v)} \phi\bigl(\exp_z(tv)\bigr)\Theta(t) f_v(t)\,dt\,\mathbf{A}_z(dv)\\ &\ge \int_{U_zM}\int_0^{\rho(v)} \phi\bigl(\exp_z(tv)\bigr) f_v'(t)\,dt\,\mathbf{A}_z(dv).\end{aligned}$$

Moreover, for each $v \in U_zM$,

$$\begin{aligned}&\int_0^{\rho(v)} \phi\bigl(\exp_z(tv)\bigr) f_v'(t)\,dt\\ &= \Bigl[\phi\bigl(\exp_z(tv)\bigr) f_v(t)\Bigr]_0^{\rho(v)} - \int_0^{\rho(v)} \frac{\partial[\phi(\exp_z(tv))]}{\partial t} f_v(t)\,dt\\ &\ge -\int_0^{\rho(v)} d\phi\Bigl(\nabla u\bigl(\exp_z(tv)\bigr)\Bigr) f_v(t)\,dt\end{aligned}$$

since $\phi \ge 0$ and $f_v(0) = 0$. Therefore we obtain

$$\begin{aligned}\int_M \phi(x)\Theta\bigl(d(z,x)\bigr)\,\mathfrak{m}(dx) &\ge -\int_{U_zM}\int_0^{\rho(v)} d\phi\Bigl(\nabla u\bigl(\exp_z(tv)\bigr)\Bigr) f_v(t)\,dt\,\mathbf{A}_z(dv)\\ &= -\int_M d\phi(\nabla u)\,d\mathfrak{m},\end{aligned}$$

which completes the proof. □

Remark 11.21 (Laplacian and Area Comparisons) As seen in the above proof, the Laplacian comparison is equivalent to the inequality (11.15), which is essentially the integration of the Bishop inequality (9.6). More precisely, (11.15) can be regarded as a directional and infinitesimal version of the area comparison (9.8). This observation, in particular, shows that the Laplacian comparison is closely related to the measure contraction property (see Remark 9.16 and Subsect. 18.5.3).

For the distance function $\bar{u}(x) := d(x,z)$ to z, we have

$$\overleftarrow{\mathbf{\Delta}}\bar{u}(x) = -\mathbf{\Delta}(-\bar{u})(x) \le (N-1)\frac{\mathbf{s}_{K/(N-1)}'(d(x,z))}{\mathbf{s}_{K/(N-1)}(d(x,z))}$$

by Theorem 11.20 for the reverse Finsler structure $\overleftarrow{F}$, since the Ricci curvature bound $\mathrm{Ric}_N \ge K$ is common between $(M, F, \mathfrak{m})$ and $(M, \overleftarrow{F}, \mathfrak{m})$ (recall Sect. 2.5).

(9 $

Clearly we have $\mathcal{I}_{(M,F,\mathfrak{m})}(0) = \mathcal{I}_{(M,F,\mathfrak{m})}(1) = 0$. An *isoperimetric inequality* means a lower bound of $\mathcal{I}_{(M,F,\mathfrak{m})}(\theta)$, i.e., a lower bound of the measure (area) of the boundary of any set of prescribed volume. The main result of this chapter is the following (established in [204]).

Theorem 15.1 (Bakry–Ledoux Isoperimetric Inequality) *Assume that $(M, F, \mathfrak{m})$ is compact and satisfies $\mathfrak{m}(M) = 1$ and $\mathrm{Ric}_\infty \geq K$ for some $K > 0$. Then we have*

$$\mathcal{I}_{(M,F,\mathfrak{m})}(\theta) \geq \mathcal{I}_K(\theta) \tag{15.3}$$

for all $\theta \in (0, 1)$, where

$$\mathcal{I}_K(\theta) := \sqrt{\frac{K}{2\pi}}\mathrm{e}^{-Kc^2(\theta)/2} \quad \textit{with } \theta = \int_{-\infty}^{c(\theta)} \sqrt{\frac{K}{2\pi}}\mathrm{e}^{-Ka^2/2}\,da.$$

In the weighted Riemannian case, Theorem 15.1 (under completeness instead of compactness) is due to Bakry and Ledoux [23]. The inequality (15.3) can be regarded as a dimension-free version of *Lévy–Gromov's isoperimetric inequality*. Lévy–Gromov's classical isoperimetric inequality states that the isoperimetric profile of an n-dimensional (unweighted) Riemannian manifold (M, g) with $\mathrm{Ric}_g \geq n-1$ is bounded from below by the profile of the unit sphere $\mathbb{S}^n$ (both spaces are equipped with the normalized volume measures; see [112, 152, 153]). In $\mathbb{S}^n$, isoperimetric minimizers (i.e., sets attaining the infimum in (15.2)) are given by balls.

In (15.3), the role of the unit sphere is played by the real line $\mathbb{R}$ equipped with the Euclidean distance and the Gaussian measure $\mathfrak{m}_K(dx) := \sqrt{K/(2\pi)}\,\mathrm{e}^{-Kx^2/2}\,dx$, thus (15.3) is also called the *Gaussian isoperimetric inequality*. Recall that $(\mathbb{R}, |\cdot|, \mathfrak{m}_K)$ satisfies $\mathrm{Ric}_\infty = K$. Isoperimetric minimizers in $(\mathbb{R}, |\cdot|, \mathfrak{m}_K)$ are the half-lines $(-\infty, c(\theta)]$ and $[-c(\theta), \infty)$, thus $\mathcal{I}_K$ is indeed the isoperimetric profile of $(\mathbb{R}, |\cdot|, \mathfrak{m}_K)$. Higher dimensional Gaussian spaces $(\mathbb{R}^n, |\cdot|, (K/(2\pi))^{n/2}\,\mathrm{e}^{-K|x|^2/2}\,dx)$ have the same isoperimetric profile $\mathcal{I}_K$, and then isoperimetric minimizers are again half-spaces of the form $\{x \in \mathbb{R}^n \mid \langle x, w\rangle \leq c(\theta)\}$ with $w \in \mathbb{S}^{n-1}$. We refer to [40, 236] for the Gaussian isoperimetric inequality in the classical Euclidean or Hilbert setting, and to [13] for the case of metric measure spaces satisfying the Riemannian curvature-dimension condition $\mathrm{RCD}(K, \infty)$ (by means of the Γ-calculus).

Exercise 15.2 Prove that, in $(\mathbb{R}, |\cdot|, \mathfrak{m}_K)$, the half-lines $(-\infty, c(\theta)]$ and $[-c(\theta), \infty)$ are isoperimetric minimizers.

We will prove Theorem 15.1 by generalizing the argument in [23]. This approach based on the Γ-calculus is known to give a sharp estimate only in the case of $\mathrm{Ric}_\infty \geq K$. As for general Ric_N, we will discuss Lévy–Gromov-type isoperimetric inequalities in Sect. 19.4 as an application of another powerful technique called the *needle decomposition* (or the *localization*). This more geometric method is also available for singular spaces, precisely, for essentially non-branching metric measure spaces satisfying the curvature-dimension condition (see [59]). This wide class of spaces includes, in particular, reversible Finsler manifolds of weighted Ricci curvature bounded below. However, in the non-reversible case, the needle decomposition gives only weaker estimates, e.g.,

$$\mathcal{I}_{(M,F,\mathfrak{m})}(\theta) \ge \Lambda_F^{-1} \cdot \mathcal{I}_K(\theta)$$

instead of (15.3), where $\Lambda_F \ge 1$ is the reversibility constant as in (3.17). See Chap. 19 (and [203]) for more details.

15.2 Poincaré–Lichnerowicz Inequality and Variance Decay

Throughout this and the next chapters, the behavior of heat semigroups as $t \to \infty$ will play an essential role. In this section, as an auxiliary step towards the Bakry–Ledoux isoperimetric inequality, we consider the outcome of the *ergodicity*

$$u_t \to \int_M u_0 \, d\mathfrak{m} \quad \text{in } L^2(M) \tag{15.4}$$

for all global solutions $(u_t)_{t\ge 0}$ to the heat equation (Proposition 13.16(iii)). Another ingredient is the *mass conservation*

$$\int_M P_{s,t}^{\nabla u}(f)\, d\mathfrak{m} = \int_M f \, d\mathfrak{m} \tag{15.5}$$

for any global solution $(u_t)_{t\ge 0}$ to the heat equation, $f \in L^2(M)$ and for any $0 \le s < t < \infty$ (this is a consequence of Lemma 11.5, (13.19) and Proposition 13.20(iii)).

We shall show a Poincaré-type inequality as a crucial step. Since the proof is common, here we consider the more general setting of $\mathrm{Ric}_N \ge K > 0$ with $N \in (-\infty, 0) \cup [n, \infty]$, yielding a Finsler analogue to the *Lichnerowicz inequality* (compare Theorem 15.4 below with Exercise 12.14). We begin with a simple application of the Bochner inequality.

Lemma 15.3 *Assume that* $(M, F, \mathfrak{m})$ *is compact and satisfies* $\mathrm{Ric}_N \ge K$ *for some* $K \in \mathbb{R}$ *and* $N \in (-\infty, 0) \cup [n, \infty]$. *Then we have, for any* $u \in H^2(M) \cap C^1(M)$ *such that* $\boldsymbol{\Delta} u \in H^1(M)$,

$$\int_M \left\{ d(\boldsymbol{\Delta} u)(\nabla u) + K F^2(\nabla u) + \frac{(\boldsymbol{\Delta} u)^2}{N} \right\} d\mathfrak{m} \le 0.$$

In particular, if $K > 0$, *then it holds that*

(%

Pag. 92(112)

$$\mathrm{diam}(M) \leq \pi \sqrt{\frac{n-1}{k}}$$

In particular,

M is compact and has finite fundamental group.

Obs: The Bonnet-Myers theorem (Pag. 91(111))

The Riemannian characterization of the Finsler curvature as in theorem 5.12 (Pag. 54(75) Riemannian Characterizations of Curvatures).

has some straightforward applications in comparison geometry.

Pag. 129(148) - Chapter 10

Examples of Measured Finsler Manifolds.

→ Ricci curvature

10.1 Minkowski Normed Spaces

Pag. 130(149) — Gaussian-type measure

$$m(dx) = e^{-k\|x\|^2/2} \mathcal{L}_n(dx) \ldots$$

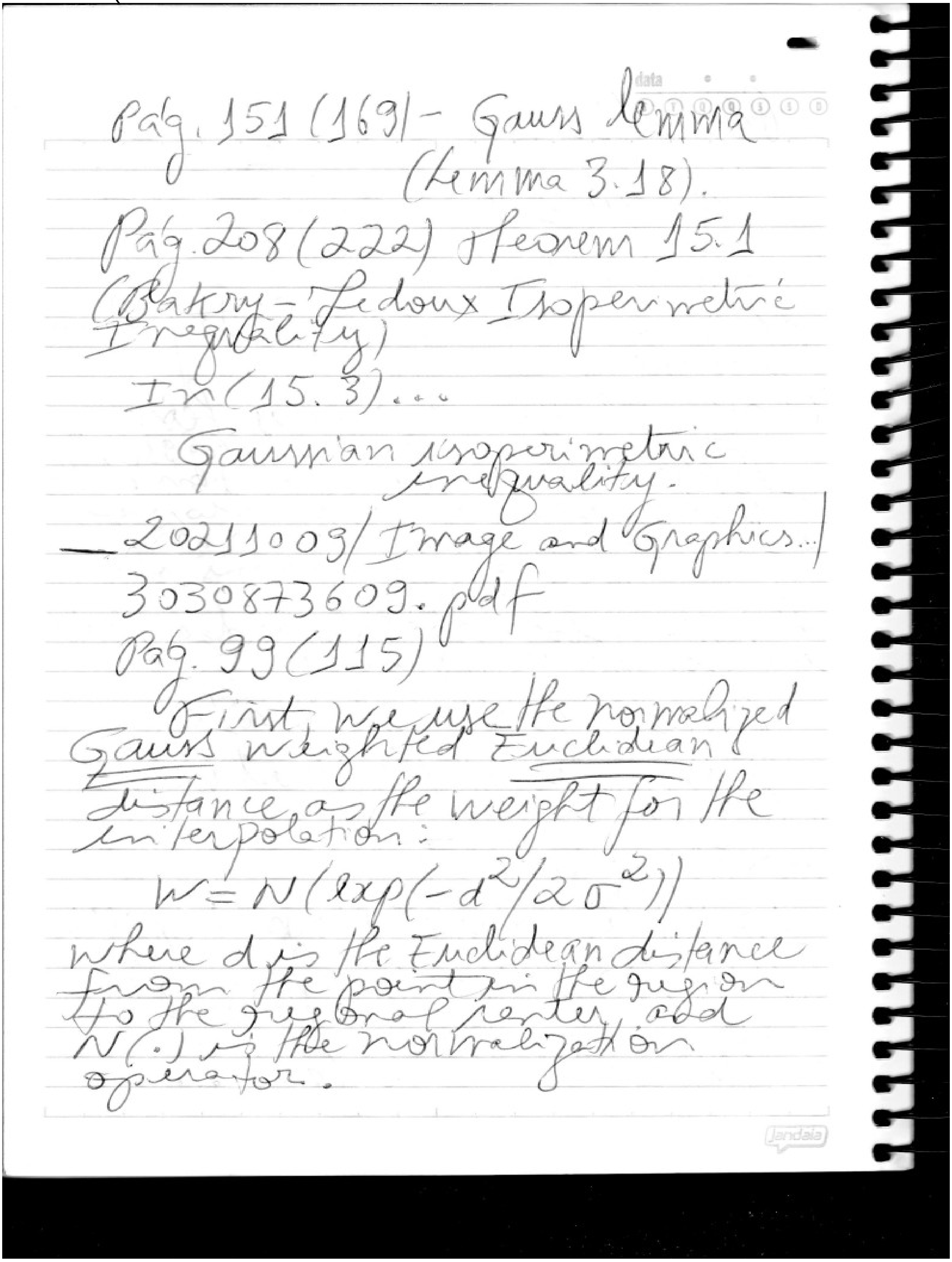

Pág. 151 (169) - Gauss lemma
(Lemma 3.18).

Pág. 208 (222) Theorem 15.1
(Bakry-Ledoux Isoperimetric Inequality)

In (15.3)...

Gaussian isoperimetric inequality.

— 20211009/Image and Graphics...|
3030873609.pdf

Pág. 99 (115)

First, we use the normalized Gauss weighted Euclidean distance as the weight for the interpolation:

$$W = N(\exp(-d^2/2\sigma^2))$$

where d is the Euclidean distance from the point in the region to the regional center, and N(.) is the normalization operator.

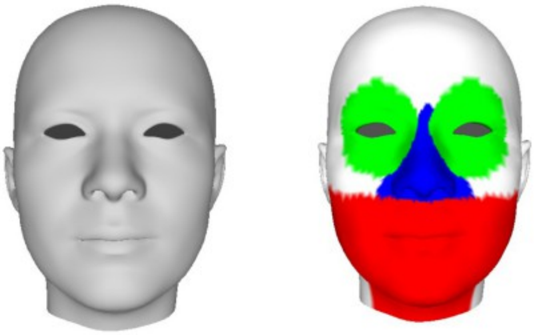

Fig. 1. Visualization of average face and AU-related mask

3D Face Data. In our work, we use the FaceWarehouse dataset [4], which contains 150 individuals aged 7–80. For each person, the dataset includes its neutral expression and 19 other expressional 3D mesh models, all of which have been densely corresponded to meet the conditions for constructing a 3DMM. Each mesh model includes 11,510 points and 11,400 faces. As for the AU labels, we use an open-source AU detection tool, OpenFace [1], to perform AU detection on the RGB images corresponding to all 3D faces in the data. In order to capture a more physical variance of AUs, we select AUs with a high occurrence rate from the data. In this work, we finally selected 8 primary AUs, namely AU10, AU14, AU17, AU20, AU23, AU25, AU28, AU43.

Parameterization of AUs. First, we get the average face $\bar{S}$ and the PCA bases of the identity attribute from 150 faces. Next, we need to obtain the PCA bases of each AU. Considering that the location of muscle movement represented by AU is relatively fixed, so we utilize a region mask to obtain the deformation of the face shape represented by the area where the AU is located and perform PCA on the deformation matrix of each AU to obtain the PCA basis B_i and eigenvalues σ_i for a single AU. Finally, our 3D face model can be expressed as:

$$\hat{S} = \bar{S} + B_{id} \cdot diag(\sigma_{id}) \cdot \alpha_{id} + \sum_{i \in U} f \cdot M_i \cdot B_i \cdot diag(\sigma_i) \cdot \alpha_i, \tag{4}$$

where U is a set of AUs, i is the index of AU, f is a parameter that controls the intensity of AU, and M_i is the mask of the i-th AU. We divide the face into several regions according to the AU positions and get the corresponding region mask in this work. As shown in Fig. 1, we show the average face $\bar{S}$ of the model and the mask area on the average face.

Smooth Processing. At this point, we can randomly generate a set of AU parameters to generate the corresponding 3D face according to Eq. 4. However, the face we generated in this way will appear a little unnatural at the mask's boundary, so we added a smoothing operation to make a face more realistic. First, we use the normalized Gaussian weighted Euclidean distance as the weight for the interpolation:

$$W = N(exp(-d^2/2\sigma^2)) \tag{5}$$

where d is the Euclidean distance from the point in the region to the regional center, and $N(\cdot)$ is the normalization operation. Then, a more natural result is obtained by linear interpolation.

$$\hat{P} = W \cdot P + (1 - W) \cdot P', \tag{6}$$

where $\hat{P}$ is the smoothed final position, P is the origin position, P' is the position before smoothing.

4 Experiments

In this section, we evaluate our proposed model from multiple aspects. Our model can edit the state of 8 AUs, namely AU10, AU14, AU17, AU20, AU23, AU25, AU28, AU43. We analyze our proposed model qualitatively and quantitatively and show our experimental results as below.

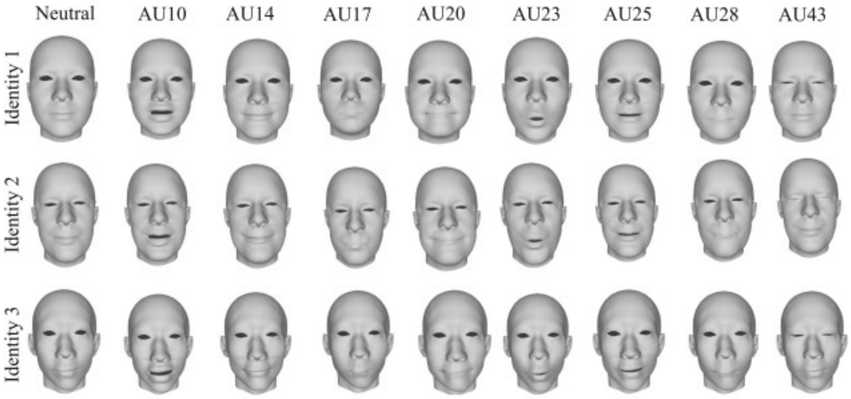

Fig. 2. Synthetic results of different AUs under different identities

Qualitative Results. Our proposed improved 3DMM model can change one of the identity attribute and the AU attribute of the 3D face without changing the other attribute. As shown in Fig. 2, we ensure that one attribute remains unchanged and displays the 3D face after another attribute is changed. Each row represents the result of changing different AU with a fixed identity parameter. In addition, we can observe that the same AU parameters show similar shape deformations on different 3D faces, which also verifies our model's control ability over a single AU.

*D 7! ! "= "! " ! ! 1 !

(% 'B 7! 8 !" ! "

Geometric Context Sensitive Loss and Its Application for Nonrigid Structure from Motion

Fudong Nian[1,2], Shimeng Yang[3], Xia Chen[2,4(✉)], Jun Wang[1], and Gang Lv[1]

[1] School of Advanced Manufacturing Engineering, Hefei University, Hefei 230601, China
[2] Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University, Hefei 230601, China
[3] School of Computer Science and Technology, Anhui University, Hefei 230601, China
[4] School of Information and Computer, Anhui Agricultural University, Hefei 230036, China
`xiachen@ahau.edu.cn`

Abstract. Coordinate prediction is an important signal processing task, which aims to predict 2D or 3D coordinates from a single image or a series of images. Previous techniques utilize Mean Square Error (MSE) loss function to measure the difference between the prediction coordinates and its corresponding ground-truth in training step, which usually assumes that coordinates are independent to each other without considering their correlations, neglecting the geometric context of the object. To address the issue, this paper presents a novel loss function named Geometric Context Sensitive (GCS) loss to model the geometric shape context of general objects by measuring the difference between any pair of prediction coordinates and their corresponding ground-truth coordinates. Our proposed method has several advantages: (1) The proposed GCS loss is trainable and can be optimized by Gauss-Newton in the traditional models. (2) GCS loss can be formulated in both 2D and 3D forms. Thus, the proposed GCS loss is easy to be implemented and can be integrated into current popular 2D/3D coordinates prediction models naturally and effectively, e.g., nonrigid structure from motion. (3) No additional learnable parameters. Though the proposed GCS loss is intuitive in theory, extensive experimental results on several public NRSFM datasets show that GCS loss can significantly boost the performance. We share the implementation code and models of our proposed methods at https://github.com/nianfudong/GCS/tree/master/NRSFM.

Keywords: Loss function design · Geometric context constraint · NRSFM

1 Introduction

Coordinate prediction (or detection) plays a major role in computer vision/multimedia area, e.g., as illustrated in Fig. 1(a), facial landmark detection aims

© Springer Nature Switzerland AG 2021
Y. Peng et al. (Eds.): ICIG 2021, LNCS 12890, pp. 130–142, 2021.
https://doi.org/10.1007/978-3-030-87361-5_11

(B 7! ! 1! ! " A! A ! , ! ! " A ! "

problems. On this basis, Broad Learning System (BLS) [22] was proposed as shown in Fig. 2. Later, authors extended BLS to the Fuzzy BLS [23] which can be applied to regression and classification. Firstly, BLS used the characteristics of the input data mapping as the “feature nodes” of the network. Secondly, the features of the mapping are enhanced to “enhanced nodes” that generate weight randomly. Finally, all mapped features and enhancement nodes are directly connected to the output, and the corresponding output coefficients can be derived from the pseudo-reverse of the courier.

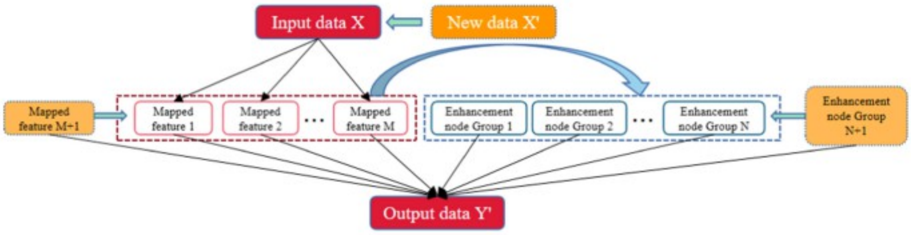

Fig. 2. Broad learning system [23]

The complex structure and numerous hyper parameters lead that DNN has long training time. To obtain higher testing accuracy, we have to increase the number of layers or add the number of parameters. Beyond that, BLS is an alternative method of DL network. The main contributions of this manuscript are as follows:

- A novel NR measure based on a flat neural network;
- A novel NR technique which used statistical distribution of gradient magnitude and the information of luminance;
- An effective fusion of the introduced NR measures;
- Avoiding excessive hyper parameters and deep layers.

3 NR IQA Based BLS

NR IQA is a critical measure in image processing. How to select various features that can describe the image logically is an important factor in IQA. The structure and contour of the image will be changed by extracting gradient features on the other hand the locally normalized luminance coefficients will be used to describe the statistical characteristics of the image.

3.1 Structure Features

Gradients can reflect the structural information of the image, such as Fig. 3. Considering the importance of the Gradient amplitude and LOG response of an image, we combine two common local contrast features and joint statistical normalized [24] as the first set of features. Given an image I,

$$G_I = \sqrt{[I \otimes h_x]^2 + [I \otimes h_y]^2} \tag{1}$$

$$L_I = I \otimes h_{LOG} \tag{2}$$

$$\begin{aligned} h_{LOG}(x, y|\sigma) &= \frac{\partial^2}{\partial x^2} g(x, y|\sigma) + \frac{\partial^2}{\partial y^2} g(x, y|\sigma) \\ &= \frac{1}{2\pi\sigma^2} \frac{x^2 + y^2 - 2\sigma^2}{\sigma^4} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \end{aligned} \tag{3}$$

where "⊗" is a linear convolution operator, h_d, $d = \{x, y\}$ is Gaussian partial derivative filter along the horizontal and vertical direction

$$\begin{aligned} h_d(x, y|\sigma) &= \frac{\partial}{\partial d} g(x, y|\sigma) \\ &= -\frac{1}{2\pi\sigma^2} \frac{d}{\sigma^2} \exp(-\frac{x^2 + y^2}{2\sigma^2}), d \in \{x, y\} \end{aligned} \tag{4}$$

In Eq. (4). $g(x, y|\sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right)$ is an isotropic Gaussian function with scale parameter σ. We use the Weibull distribution [25] and the generalized Gaussian distribution [26] to fit the gradient feature LOG response of image. Firstly, we set

$$F_I(i, j) = \sqrt{G_I^2(i, j) + L_I^2(i, j)} \tag{5}$$

Subsequently, we an add adaptive normalization factor at a local location (i, j)

$$N_I(i, j) = \sqrt{\sum \sum\nolimits_{l,k \in \Omega_{i,j}} \omega(l, k) F_I(l, k)} \tag{6}$$

where $\Omega_{i,j}$ is a window, $\omega(l, k)$ is a weight which obeys Gaussian distribution, and satisfies $\sum\limits_{l,k} \omega(l, k) = 1$. Normalizing the Eq. (1) and Eq. (2) as

$$\overline{G_I} = G_I/(N_I + \varepsilon) \tag{7}$$

$$\overline{L_I} = L_I/(N_I + \varepsilon) \tag{8}$$

where ε is a normal small number to prevent N_I being too small and resulting unstable values.

Then quantified $\overline{G_I}(i, j)$ and $\overline{L_I}(i, j)$ as m levels $\{g_1, g_2, g_3, ...g_m\}$ and n levels $\{l_1, l_2, l_3, ...l_n\}$, expressed as G and L respectively. The joint probability distribution, the marginal distribution and the independence of G and L are defined as follows

$$K_{m,n} = P(G = g_m, L = l_n), m = 1, ...M; n = 1, ...N \tag{9}$$

$$\begin{cases} P_G(G = g_m) = \sum_{n=1}^{N} K_{m,n} \\ P_L(L = l_n) = \sum_{m=1}^{M} K_{m,n} \end{cases} \tag{10}$$

$$D_{m,n} = \frac{K_{m,n}}{P(G = g_m) \times P(L = l_n)} \tag{11}$$

$D_{m,n} = 1$ for all m and n if G and L are independent. Hence

$$\begin{aligned} Q_G(G = g_m) &= \frac{1}{N}\sum\nolimits_{n=1}^{N} \frac{P(G = g_m, L = l_n)}{P(L = l_n)} \\ &= \frac{1}{N}\sum\nolimits_{n=1}^{N} P(G = g_m | L = l_n) \end{aligned} \tag{12}$$

$$\begin{aligned} Q_L(L = l_n) &= \frac{1}{M}\sum\nolimits_{m=1}^{M} \frac{P(G = g_m, L = l_n)}{P(G = g_m)} \\ &= \frac{1}{M}\sum\nolimits_{m=1}^{M} P(L = l_n | G = g_m) \end{aligned} \tag{13}$$

Equation (12) and Eq. (13) describe the independence between G and L. We set the ratio parameter of the filter as 0.5 which can capture more details in the feature maps, we extract 40 dimensional features.

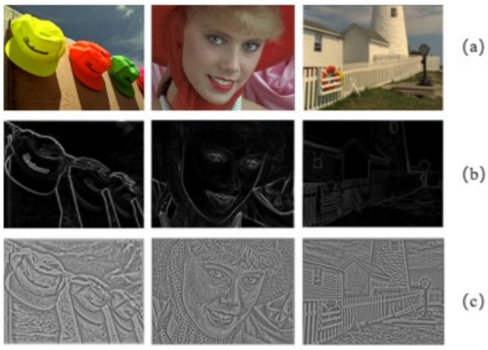

Fig. 3. (a) three example images; (b) the gradient magnitude of (a); (c) locally normalized luminance of (a)

3.2 Statistical Features

Using the scene statistics of the locally normalized luminance coefficients [31] to quantify the loss caused by the distortion. Which as follows

$$\hat{I}(i,j) = \frac{I(i,j) - \mu(i,j)}{\delta(i,j) + C} \tag{14}$$

where

$$\mu(i,j) = \sum_{s=-S}^{S}\sum_{t=-T}^{T} \varpi_{s,t} I_{s,t}(i,j) \tag{15}$$

Table 1. PSNR and SSIM of the denoising methods evaluated on DnD [26] dataset.

Method	Blind/non-blind	PSNR	SSIM
CDnCNN-B [32]	Blind	32.43	0.7900
TNRD [12]	Non-blind	33.65	0.8306
LP [11]	Non-blind	34.23	0.8331
FFDNet [34]	Non-blind	34.40	0.8474
BM3D [13]	Non-blind	34.51	0.8507
WNNM [16]	Non-blind	34.67	0.8646
KSVD [3]	Non-blind	36.49	0.8978
MCWNNM [30]	Non-blind	37.38	0.9294
FFDNet+ [34]	Non-blind	37.61	0.9415
TWSC [29]	Non-blind	37.96	0.9416
CBDNet [17]	Blind	38.06	0.9421
RIDNet [4]	Blind	39.25	0.9528
IERD [6]	Blind	39.30	0.9531
Ours	Blind	**39.48**	**0.9580**

Table 2. The quantitative results (PSNR) for the SIDD dataset [1].

Method					
BM3D	FFDNet	CBDNet	RIDNet	IERD	Ours
30.88	29.20	30.78	38.71	38.82	**39.55**

4.3 Ablation Study

The major contribution of this work is proposing a novel implementation of the SPA attention module based on weighting wavelet layers. Although previous SPA-based attention mechanisms have already experienced remarkable performance, our approach offers a new perspective to this problem and answers new questions of "how does the pyramid level influence the denoising performance of the SPA blocks" and "if the SPA blocks can boost the performance of other networks with arbitrary architectures". Our results prove the superiority of the proposed SPA block compared to the naive channel attention block. At the same time, it also justifies the effect of different levels of SPA blocks. We also conducted an ablation study on the proposed network architecture. We test PSNR on the SIDD validation set for levels-0 to level-4. Notably, we define the naive channel attention as a level-0 pyramid (which means no pyramid). The second row of Table 3 shows that the higher pyramid level of the SPA blocks leads to a high PSNR and a better denoising performance. For evaluating the generalization of the SPA blocks, we conducted another ablation study by deploying SPA blocks at different levels in the RIDNet network as an alternative for its channel attention blocks. The third row of Table 3 shows the results of RIDNet when equipped with SPA blocks of different pyramid levels. The results show that the

SPA blocks generalize well in RIDNet while expecting to confirm that pyramids with more layers lead to better results.

Table 3. Investigation of effects of the pyramid level of SPA block and generalization on RIDNet.

Networks	Level	0	1	2	3	4
Ours	PSNR	39.24	39.33	39.47	39.55	39.57
RIDNet	PSNR	38.71	38.87	38.99	39.04	39.05

The proposed SPA attention is also generalized on the most commonly used AWGN model instead of only specific for real-world noise models. To investigate the AWGN noise, we use images with synthetic noise. We trained the blind denoising RIDNet and our network equipped with levels-0 to level-4 SPA attention on BSD500 [23] dataset with random AWGN (where the deviation σ varies between 5 to 55). The trained models were evaluated on BSD68 [27] dataset with noise level 15, 25, and 50. Table 4 compared the performance of each model. It consistently proves the generalizability.

Table 4. Comparisons on AWGN image denoising.

Noise levels	Method (RIDNet/Ours)			
	Level 1	Level 2	Level 3	Level 4
15	34.01/34.12	34.11/34.25	34.22/34.37	34.22/34.38
25	31.37/31.44	31.49/31.55	31.55/31.65	31.56/31.66
50	28.14/28.23	28.22/28.30	28.31/28.42	28.32/28.42

To further evaluate the performance of the proposed method, denoising is performed for Poisson-Gaussian noise model, which is defined as $y = P(x) + N(0, \sigma^2)$. This model includes Poisson-distribution $P(x)$ related to the clean image x, and followed by a Gaussian-distribution $N(0, \sigma^2)$. Here, we set the standard deviation σ to 15. We trained our network equipped with levels-0 to level-4 SPA attention on DIV2K [2] training set, and test them on DIV2K validation set. Table 5 presents the results of each level. The results show that the proposed SPA works better than naive channel attention (Level 0), and the performance gain still followed the obey the rules we observed in the previous experiments, that higher level pyramid brings more improvement.

Table 5. Comparisons on Poisson-Gaussian noise denoising.

Pyramid level	w/o CA	0	1	2	3	4
PSNR	31.11	31.52	32.21	32.57	33.04	33.06

Table 6. Comparisons on GPU runtime (seconds) and Model size (MB).

Level		w/o CA	0	1	2	3	4
Model size/GPU runtime	RIDNet	5.71/0.16	5.72/0.19	5.73/0.22	5.74/0.23	5.74/0.25	5.75/0.27
	Ours	17.11/0.22	17.13/0.24	17.16/0.29	17.18/0.32	17.21/0.35	17.23/0.41

4.4 Computational Overhead

As mentioned before, the SPA blocks improves the denoising performance while not adding considerable computational overhead. In this section, we compared the model sizes[1] when deployed different-level SPA blocks on both RIDNet and our network. It can be seen in Table 6 that the SPA blocks would not significantly increase the model size. Moreover, comparisons on average GPU runtime were conducted to show the proposed SPA module achieve a good trade-off between the calculation and performance. We fed images with a size of 512 × 512 for runtime evaluation. We tested each model 100 times and calculated the average runtime.

5 Conclusion

A novel channel attention module called Sub-band Pyramid Attention (SPA) is proposed in this work. The SPA blocks are built upon wavelet decomposition to realize a joint sub-band channel attention. The SPA is implemented as a plug-and-play module, hence can replace the naive channel attention in arbitrary deep learning networks. The proposed SPA block performs a more precise feature re-calibration that re-scales both the feature channels and the multi-level frequency components. The achieved gain for the proposed method over the conventional filtering methods and DL methods is considerable, ranging from 0.18 dB to 3.97 dB on the benchmark dataset of DnD. The generalization of the SPA blocks is verified by the ablation study, which suggests that the SPA module is compatible with other network architectures and can be widely used in other networks to boost image restoration performance.

References

1. Abdelhamed, A., Lin, S., Brown, M.S.: A high-quality denoising dataset for smartphone cameras. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1692–1700 (2018)

---

[1] Model size was calculated by torchsummary package (https://github.com/sksq96/pytorch-summary).

Pág. 146 - Gauss-Newton in the traditional models.

Pág. 171 (186) -
Gaussian partial derivative filter along the horizontal and vertical direction

Pág. 311 (326) -
Poisson - Gaussian noise model,

—

20211010,,,

Michelson and the speed of light
Bernard Jaffe 1886-1986.

The speed of light
Ron Carlson

Smarandache Geometries and Curves

https://archive.org/details/smarandache-geometries-and-curves

https://archive.org/details/mgstheory

—

$$

! 1 ##! A #" ! # ! ! "! ! " A

! 1 ##! A #" ! # !19 3

-256- - -* ! - $$!19 3 1"

! 1 ##! A #" ! # 1!

-256- - -* ! - $$ 1! 1"

data

Books:

- 20211010/ Géométrie Contemporaine - 3 Volumes//
- Star and Relativity... Vol.1.pdf (somente imagem)
- Star and Relativity... djvu (texto)

Source: 20211010/Géométrie Contemporaine - 3 Volumes/Star-and-Relativity-Zel'dovich/Star and Relativity... (2011).djvu

Pág. 16(33)

The curvature at a given point of a three-dimensional space is determined in the following manner.

A geodesic surface (analogue of a plane in Euclidean space) is drawn through the point, and its curvature is calculated. This curvature is called the Riemann curvature of space in the two-dimensional direction of the chosen surface. The curvature may be

$B
! -256- - -* ! - $B ! ! " 5 ! A 3 5 ! A
! 1 3 A " ! " ! " ! 5 ! " 1"
E !
! -256- - -* ! - $B ! ! " ! A 3 4 A
D ! "KA
E L

-256- - -* ! - $B ! ! " ! A 3 4 A D !
1"
(B %% 9 A! ! ! A 1 ! " ! 1! " "
!
*D ! ! J A! ! ! " D 1 A J
9 L
9 A! ! ! A 1 ! " ! 1! " "
! 2 " ! ! ! ! 1 ! = " !
1! " ! 1 J ! " A! ! ! " 9
A! ! " 5 ! A! 1! " !
" ! 9 A! !3 D " "
" 9 A! !A ! " A ! " ! " 7! !
A! 1! @ A ! 1
A! ! ! < 1 ! 1 L! 1 J
! ! " ! " Z
9 J ! ! ! L1! " !
!D A ! ! A A!
& " " ! " ! ! !
= " ! 3 ! ! ! !
" ! 1! 9 " ! 1! ! ! !
! [! [" ! 31 ! [[1!
" ! 1! 9 1! ! " ! ! !
! A " " ! 1! & ! " 3
1 ! 3 A " = " ! 9
! 1 3 ! ! 3 ! A " " !
! ! A ! ! " ! 1! & 1 !
! 1! ! J 3 A " ! = " !

Let us sum up: Even in conventional flat spacetime it is impossible to establish on an accelerated body a rigid coordinate system with a flat (Euclidean) space metric and with synchronized time. Such a coordinate system can be tied only to an inertially moving body. With the exception of special cases (for example, a rigidly rotating disk), any given frame of reference will become deformed in the course of time; its geometrical properties will change, and the rates of clocks tied to it will change.

In Newtonian physics, a rigid Cartesian frame of reference is determined by specifying at each moment of time the location of its origin and the orientation of its axes. In special relativity, to establish an accelerated reference frame one must give not only the motion of one of its points (the origin) but also the motion of all other points. We will mean by reference frame a collection of test particles (each carrying a clock) which fill all of spacetime, and which move along arbitrary paths.[7]

A similar situation occurs in GTR. The key difference is that in STR, where gravitational fields are absent, one can always pass from any noninertial frame of reference to an inertial frame of reference, and use that inertial frame throughout all of spacetime. In GTR, this cannot be done because of the curvature of spacetime.

We turn now to the mathematical formulation of these ideas. All formulae that we shall obtain will be valid not only in flat spacetime but also in the curved spacetime of general relativity, because they are local formulae and by virtue of the equivalence principle a gravitational field is locally indistinguishable from an accelerated frame.

1.4 THE MEASUREMENT OF SPACETIME INTERVALS

In an inertial frame of reference in STR, it is not absolutely necessary to use Cartesian space coordinates. Any curvilinear space coordinates (for example, spherical coordinates) can be used. The transformation from one system of space coordinates to another has the form

$$x^{\alpha} = x^{\alpha}(x^{1'}, x^{2'}, x^{3'}) \, . \qquad (1.4.1)$$

Here the index α takes on the values 1, 2, 3; and x^1, x^2, x^3 denote the three space coordinates.

The expression for ds^2 now takes on a form different from equation (1.2.1). Instead of the three squares of the differentials of the Cartesian coordinates, we have the expression for the square of the space interval in curvilinear coordinates (taken with a minus sign). In spherical coordinates this is $-dl^2 = -(dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta\, d\phi^2)$, and the squared spacetime interval is written

---

[7] *Editors' note.*—More precisely the "reference frame" is identified with the "collection of test particles" moving in a specified manner. A reference frame is not a coordinate system, which includes in addition a parameterization of the particles (by coordinates x^1, x^2, x^3) and a time variable x^0 (not necessarily physical time!) for each particle.

This terminology is used widely in the Soviet Union but is not common in the West.

$$ds^2 = (dct)^2 - dr^2 - r^2 d\theta^2 - r^2 \sin^2\theta\, d\phi^2 \,. \tag{1.4.2}$$

In cylindrical coordinates

$$ds^2 = (dct)^2 - dr^2 - r^2 d\phi^2 - dz^2 \,. \tag{1.4.3}$$

The transformation (1.4.1) corresponds to a passage from one three-dimensional space grid of coordinates to another. However, the frame of reference (cf. § 1.3) remains unchanged. In passing to a different frame of reference, which moves arbitrarily relative to the inertial frame, one must use a time-dependent coordinate transformation

$$x^a = x^a(x^{1'}, x^{2'}, x^{3'}, t) \,. \tag{1.4.4}$$

The spatial part ($a = 1, 2, 3$) of this transformation defines completely the motion of the new reference system. The time part ($a = 0$) of the transformation (which involves coordinate time, not physical time; see below) can be arbitrary.

The transformation to a uniformly rotating reference frame has the form

$$\begin{aligned} x^1 &= x^{1'} \cos \Omega t - x^{2'} \sin \Omega t \,, \\ x^2 &= x^{1'} \sin \Omega t + x^{2'} \cos \Omega t \,, \\ x^3 &= x^{3'} \,, \end{aligned}$$

where Ω is the angular velocity; or, in cylindrical coordinates,

$$r = r' \,, \quad z = z' \,, \quad \phi = \phi' + \Omega t \,. \tag{1.4.5}$$

(The symbol t will be reserved for the time coordinate [not physical time!] in the new reference frame.) In this case, substituting equation (1.4.4) or (1.4.5) into equation (1.4.3), we find that the entire expression for ds^2 changes:

$$ds^2 = (c^2 - \Omega^2 r'^2) dt^2 - 2\Omega r'^2\, d\phi' dt - dz'^2 - r'^2 d\phi'^2 - dr'^2 \tag{1.4.6}$$

for the rotating disk; and

$$ds^2 = g_{ik} dx^i dx^k \tag{1.4.7}$$

in general. Here and throughout this book any index which appears twice must be summed; latin indices take on the values 0, 1, 2, 3; x^1, x^2, x^3 denote space coordinates (generally curvilinear); $x^0 = ct$ is the time coordinate; and g_{ik} are functions of the spacetime coordinates. Each term with $i \neq k$ occurs twice in the sum (1.4.7); for example, $g_{01} dx^0 dx^1$ and $g_{10} dx^1 dx^0$. By definition, g_{ik} always equals g_{ki}.

The quantities g_{ik} are called the components of the *metric* tensor; they define the so-called spacetime metric. The specific forms of g_{ik}, as functions of the coordinates, are determined not only by the properties of spacetime but also by the particular coordinate system used, that is, by the choice of reference frame, by the choice of space coordinates in that frame, and by the choice of the method of measuring time (i.e., the time coordinate).

The expression (1.4.7) contains complete information about the geometric properties of the reference frame and about the geometric properties of spacetime itself. How is this expression to be used? First, to examine the passage of time at a given point it is obviously necessary to regard x^1, x^2, x^3 as constant: $dx^1 = dx^2 = dx^3 = 0$. Then ds will be the time separation between two nearby events; i.e., it will be the interval of time multiplied by the speed of light, $ds = cd\tau$. Consequently,

$$d\tau = [(g_{00})^{1/2}/c]dx^0 \,. \qquad (1.4.8)$$

In any reference frame formed by real bodies, it is always true that $g_{00} > 0$.

Let us now calculate the spatial distance dl between two particles at rest in our reference frame. This cannot be done by simply setting $dx^0 = 0$, $x^0 = \text{constant}$ in expression (1.4.7). The problem with such a procedure is that the same moment of proper time does not always correspond to identical clock readings at different points in space. Therefore, before performing the computation we must determine which value of x^0 at a neighboring point is "simultaneous" with a given value of x^0 at a given point. Clock synchronization between two nearby points is performed by means of light signals.[8] We will not go into the computation here; instead, we refer anyone interested to the textbook of Landau and Lifshitz (1962). The final formula for the square of the spatial distance dl^2 is

$$dl^2 = \left(-g_{\alpha\beta} + \frac{g_{0\alpha}g_{0\beta}}{g_{00}}\right) dx^\alpha dx^\beta, \quad \alpha, \beta = 1, 2, 3 \,. \qquad (1.4.9)$$

The terms in parentheses will be denoted $\gamma_{\alpha\beta}$. They determine the metric of the three-dimensional space in our frame of reference.[9] The volume element of this three-dimensional space is given by the expression $dV = \sqrt{\gamma}\, dx^1 dx^2 dx^3$, where $\gamma = |\gamma_{\alpha\beta}|$ is the determinant of the matrix composed of elements $\gamma_{\alpha\beta}$.

---

[8] See n. 6.

[9] *Editors' note.*—This three-dimensional space is defined only locally, in the following sense: At each event in spacetime it is an infinitesimal three-dimensional surface orthogonal to the world line $(x^1, x^2, x^3) = \textit{constant}$. These infinitesimal "three-surfaces" usually cannot be meshed together to form a family of three-surfaces extending throughout spacetime. According to a theorem of differential geometry, the infinitesimal three-surfaces mesh to form finite three-surfaces if and only if the world lines $(x^1, x^2, x^3) = \textit{constant}$ have zero rotation—which they usually do not—or, equivalently, if and only if the angular-velocity tensor $A_{\alpha\beta}$ of equation (1.6.3) vanishes. Consequently, while the space metric $\gamma_{\alpha\beta}$ is defined at each event in spacetime, it is not usually the metric of a large-scale three-surface embedded in spacetime. Nevertheless, it *is* the metric of a (fictitious) three-surface whose geometric properties (Riemannian and Gaussian curvatures—cf. end of this section and end of § 1.8—deformations [eq. (1.6.4)], etc.) describe the geometric and kinematic properties of the reference frame.

Consider again the rotating disk as an example. Using formula (1.4.8), we obtain from equation (1.4.6) for the time interval (in the future we omit the prime on the coordinates)

$$d\tau = (1 - \Omega^2 r^2/c^2)^{1/2} dt\ .$$

On the axis of rotation, $r = 0$ and $d\tau = dt$. Hence the preceding formula can be rewritten in the form

$$d\tau = (1 - \Omega^2 r^2/c^2)^{1/2} d\tau_{\text{axis}}\ .$$

Time flows more slowly, the further removed one is from the axis of rotation.

For the element of spatial distance, we obtain from equations (1.4.6) and (1.4.9).

$$dl = \left[dr^2 + \frac{r^2 d\phi^2}{1 - \Omega^2 r^2/c^2} + dz^2\right]^{1/2}. \tag{1.4.10}$$

Using equation (1.4.10), we find that, when r and z are constant, then $dl = (1 - \Omega^2 r^2/c^2)^{-1/2} r\, d\phi$, and the ratio of the circumference of the circle to its diameter is

$$\frac{l_{\text{circle}}}{D} = \frac{\pi}{(1 - \Omega^2 r^2/c^2)^{1/2}}\ ,$$

in accordance with the analysis of the preceding section.

For comparison, we present an example of a nonstatic reference frame (i.e., a reference frame deforming with time) in which, as for our rotating disk, the ratio of a circle's circumference to its diameter is less than π.

In Minkowski spacetime consider a collection of particles ejected in all directions with all possible speeds, from a given point in space at some moment of time. The world lines of these particles fill the interior of the light cone. The reference frame tied to the particles is called the Milne frame (after Milne, who introduced such a cosmological model).

The transformation from the spherical coordinates (r', θ', ϕ') of a nondeforming reference frame and from laboratory time, t', to Milne's space coordinates (ψ, θ, ϕ) and the proper time of the particles, t, is given by

$$t^2 = t'^2 - r'^2\ , \quad \tanh\psi = r'/t'\ ,$$

$$\theta = \theta'\ , \quad \phi = \phi'\ .$$

Substituting these transformations into $ds^2 = c^2dt^2 - dr'^2 - r'^2(d\theta'^2 + \sin^2\theta'\, d\phi'^2)$, we obtain the expression for ds^2 in the Milne frame

$$ds^2 = d(ct)^2 - (ct)^2\,[d\psi^2 + \sinh^2\psi(d\theta^2 + \sin^2\theta\, d\phi^2)]\ .$$

At a fixed moment in time, $t = const.$, we examine the curvature of an equatorial "plane," $\theta = \frac{1}{2}\pi$. The ratio of the circumference of a circle in this equatorial plane, $l = 2\pi ct \sinh\psi$, to its diameter, $D = 2ct\psi$, is greater than π: $l/D = \pi(\sinh\psi)/\psi > \pi$.

If at a given point of a two-dimensional surface the ratio of the circumference of a small circle to its diameter, l/D, is smaller than π, then the curvature of the surface is said to be positive, and its geometry is similar to the geometry on a sphere. If $l/D > \pi$, then the curvature is negative, and the geometry is similar to the geometry on a pseudosphere (saddle-shaped surface).

Numerically, the curvature is defined by the square of the radius of curvature, a^2, which is determined in the following manner. A small triangle is drawn on the surface; the sides of the triangle are the shortest lines (geodesics) between its vertices. Let Σ denote the sum of the angles of the triangle. It can be demonstrated that the difference $\Sigma - \pi$ is proportional to the area of the triangle, S:

$$\Sigma - \pi = CS\,.$$

The coefficient of proportionality C is called the curvature, and the quantity $a = 1/C^{1/2}$ is the radius of curvature. If $\Sigma > \pi$, then $C = 1/a^2 > 0$, and the curvature is positive. If $\Sigma < \pi$, then $C = 1/a^2 < 0$, the curvature is negative, and a is imaginary. The smaller $|a^2|$, the larger the curvature and the more the geometry differs from Euclidean geometry.

The curvature at a given point of a three-dimensional space is determined in the following manner. A geodesic surface (analogue of a plane in Euclidean space) is drawn through the point, and its curvature is calculated. This curvature is called the Riemann curvature of space in the two-dimensional direction of the chosen surface. The curvature may be different in different directions. The curvature averaged over all directions is called the Gaussian curvature of the three-space. We will not give here formulae for computing the curvature in the general case. For important specific examples, the formulae are listed at the end of § 1.8.

Thus, a surface of constant time in the expanding Milne reference frame (see above) has negative curvature.

It is not difficult to understand the mathematical source of the non-Euclidean three-geometry of a noninertial reference frame in flat four-dimensional spacetime. The three-dimensional space of an *inertial* frame is a "flat" three-dimensional hypersurface or "slice" passing through the four-dimensional spacetime. The space of a three-dimensional *noninertial* frame results from a curved slice through four-dimensional spacetime. It is hardly surprising that the geometry of this curved slice is non-Euclidean. The situation is completely analogous to the geometry of a curved two-dimensional surface in a conventional (flat) three-dimensional space. In spite of the fact that the space is flat, the geometry on the curved surface is non-Euclidean.

1.5 SOME FORMULAE FOR CURVILINEAR COORDINATES

In STR one introduces the concept of a four-dimensional vector (four-vector) B_i as an aggregate of four quantities (functions of coordinates and time) which, under Lorentz transformations, transform as the coordinates x^i do. Examples of four-vectors are the four-velocity, $u^i = dx^i/ds$, the four-momentum, $p^i = mu^i$, and the four-potential of the electromagnetic field, A_i. A four-tensor of second rank, B_{ik}, is defined as an aggregate of quantities which transform like the product of the coordinates $x^i x^k$. An example is the electromagnetic-field tensor

$$F_{ik} = \partial A_k/\partial x^i - \partial A_i/\partial x^k \, .$$

Tensors of third rank and higher are defined in an analogous manner.

In inertial frames of reference in STR one uses Galilean coordinates, in which the interval takes the form (1.2.1). When passing to curvilinear coordinates in four-dimensional spacetime one must describe vectors and tensors in terms of this new type of coordinates. Most important of all is the introduction of the *covariant* and *contravariant* components of a vector.

A contravariant four-vector is any aggregate of quantities B^i (with index up) which, under a change of coordinates, $x^i = x^i(x^{0'}, x^{1'}, x^{2'}, x^{3'})$, transforms according to the rule

$$B^i = (\partial x^i/\partial x^{k'})B^{k'} \, . \tag{1.5.1}$$

An example of a contravariant vector is the set of coordinate differentials dx^i, since $dx^i = (\partial x^i/\partial x^{k'})dx^{k'}$.

The covariant components of the same vector B_i (with index down) are defined by

$$B_i = g_{ik}B^k \, . \tag{1.5.2}$$

From the determination of g_{ik} as the coefficients in equation (1.4.7), their law of transformation, $g_{ik} = (\partial x^{l'}/\partial x^i)(\partial x^{m'}/\partial x^k)g_{l'm'}$, follows. Using this law, and referring to equation (1.5.2), we obtain the law of transformation for the covariant components of a vector

$$\begin{aligned} B_i = g_{ik}B^k &= (\partial x^{l'}/\partial x^i)(\partial x^{m'}/\partial x^k)g_{l'm'}(\partial x^k/\partial x^{n'})B^{n'} \\ &= (\partial x^{l'}/\partial x^i)B_{l'} \, . \end{aligned} \tag{1.5.3}$$

The concept of a tensor is generalized in the same way: The transformation law for a contravariant tensor, B^{ik}, is

$$B^{ik} = (\partial x^i/\partial x^{l'})(\partial x^k/\partial x^{m'})B^{l'm'} \, , \tag{1.5.4}$$

and for its covariant components the transformation law is

$$B_{ik} = g_{li}g_{mk}B^{lm} = (\partial x^{l'}/\partial x^i)(\partial x^{m'}/\partial x^k)B_{l'm'} \, . \tag{1.5.5}$$

One also uses mixed components, which transform as

$$B^i{}_k = B^{il}g_{lk} = (\partial x^{l'}/\partial x^k)(\partial x^i/\partial x^{m'})B^{m'}{}_{l'} \, . \tag{1.5.6}$$

The concept of a tensor of a higher rank is generalized similarly.

The components g_{ik}, as indicated by their law of transformation, comprise a tensor. This tensor, which plays a fundamental role in the theory, is called the *fundamental metric tensor*.

The determinant

$$g = |g_{ik}| \tag{1.5.7}$$

is called the *fundamental determinant.*

The quantities g^{ik}, defined by

$$g_{ik}g^{im} = \delta^m{}_k \, , \tag{1.5.8}$$

where $\delta^m{}_k$ is the Kronecker delta, are called the *contravariant components* of the metric tensor. From equation (1.5.8), it follows that

$$g^{ik} = (A^{ik}/g) \, , \tag{1.5.9}$$

where A^{ik} is the minor of the matrix g_{ik}.

From equation (1.5.8), by using equation (1.5.5), we find that

$$B^{ik} = g^{il}g^{mk}B_{lm} \, . \tag{1.5.10}$$

Thus, just as the lowering of indices is performed with the covariant components g_{ik}, so the raising of indices is performed with the contravariant components g^{ik}. The mixed tensor $g^i{}_k$ is equal to the Kronecker symbol, $g^i{}_k = \delta^i{}_k$.

Consider the quantity A^iB_i. It is the scalar product of the vectors A^i and B^j, and it is unaffected by a transformation of coordinates. Specifically, the square of the length of a vector is

$$A^2 = A^iA_i \, . \tag{1.5.11}$$

Similarly, we can construct a scalar from two tensors

$$A^{ik}B_{ik} = A_i{}^kB^i{}_k = A_{ik}B^{ik} \, .$$

All three notations are equivalent. If the second tensor is the fundamental tensor, then the scalar $A^{ik}g_{ik} = A^i$, is called the *trace* of the tensor.

In a similar manner, lower-rank tensors can be formed from higher-rank tensors. For example,

$$A_{iklm}g^{mi} = A^i{}_{kli} = A_{kl} \, .$$

This operation is called *tensor contraction*.

In curvilinear coordinates, the concept of differentiation of vectors and tensors must also be generalized. The covariant derivatives (denoted by a semicolon) of a contravariant vector and of a covariant vector are the quantities

$$B^i{}_{;k} = (\partial B^i/\partial x^k) + \Gamma^i{}_{lk}B^l \, , \tag{1.5.12}$$

and

$$B_{i;k} = (\partial B_i/\partial x^k) - \Gamma^l{}_{ik}B_l \, , \tag{1.5.13}$$

respectively. Here $\Gamma^l{}_{mn}$ are Christoffel symbols (not tensors!) which are defined by the expressions

$$\Gamma^l{}_{mn} = \tfrac{1}{2}g^{lk}\left[(\partial g_{km}/\partial x^n) + (\partial g_{kn}/\partial x^m) - (\partial g_{mn}/\partial x^k)\right] . \tag{1.5.14}$$

Obviously, in Cartesian coordinates all the $\Gamma^l{}_{mn}$ vanish, and covariant differentiation reduces to conventional differentiation.

Tensors of higher rank are differentiated in a similar manner:

$$B^{ik}{}_{;l} = (\partial B^{ik}/\partial x^l) + \Gamma^i{}_{ml}B^{mk} + \Gamma^k{}_{ml}B^{im}\,, \tag{1.5.15}$$

$$B^i{}_{k;l} = (\partial B^i{}_k/\partial x^l) - \Gamma^m{}_{kl}B^i{}_m + \Gamma^i{}_{ml}B^m{}_k\,, \tag{1.5.16}$$

$$B_{ik;l} = (\partial B_{ik}/\partial x^l) - \Gamma^m{}_{il}B_{mk} - \Gamma^m{}_{kl}B_{im}\,. \tag{1.5.17}$$

It is important to note that from equations (1.5.12) and (1.5.14) and the expression for ds^2 one can obtain the following expression for the covariant divergence of a vector

$$B^i{}_{;i} = [1/(-g)^{1/2}]\partial[(-g)^{1/2}B^i]/\partial x^i\,. \tag{1.5.18}$$

Finally, we give the equation in curvilinear coordinates which defines the geodesic line connecting two points in four-dimensional space (in flat space, in which we are working in this section, it is a straight line):

$$(d^2x^i/ds^2) + \Gamma^i{}_{kl}\,(dx^k/ds)(dx^l/ds) = 0\,. \tag{1.5.19}$$

In Minkowski space, as we know from STR, a body of finite rest mass moves along a straight (timelike) line; i.e., along a geodesic. Hence, equations (1.5.19) are the equations of motion of a body of finite rest mass, written in the curvilinear coordinates of a noninertial reference frame. The differential equation for a geodesic in curved spacetime has precisely the same form as equation (1.5.19) for a straight line in flat spacetime.

1.6 DYNAMIC AND KINEMATIC QUANTITIES

The quantities g_{ik} of equation (1.4.7) are constructed from derivatives of the transformation (1.4.4), which determines the motion of the reference frame used with respect to the original inertial frame. In particular, g_{ik} includes $\partial x^{i'}/\partial x^0$, the relative velocity of the two frames. Hence, it is natural that g_{ik} contains information not only about the flow of time and the geometry of the reference frame but also about its accelerations and deformation. We shall cite here only the final formulae for the computation of dynamic and kinematic quantities; for more detail see the contributions of Zel'manov (1944, 1959*b*). The three-dimensional vector F^α, describing the acceleration relative to the reference frame of a freely moving body momentarily at rest in it, is determined by the expression

$$F^\alpha = -c^2\Gamma^\alpha{}_{00}/g_{00} \qquad (\alpha = 1, 2, 3)\,, \tag{1.6.1}$$

as we shall demonstrate below (eq. [1.6.1a]). The quantities $\Gamma^\alpha{}_{00}$ were defined in the preceding section (eq. [1.5.14]). The vector F^α forms a field of inertial forces in the reference frame. F^α is a three-dimensional vector; in working with it one must use the tensor $\gamma_{\alpha\beta}$ (see § 4 of this chapter). Rein STR, then, as in GTR, the formula for the number of particles would take the form (1.8.12). Thus, in GTR the law of conservation of particles is written in exactly the same form as when gravity is absent. This is hardly surprising, since gravitation obviously does not change the number of particles.

Finally, let us write an expression for the Gaussian curvature (cf. § 1.4) of the three-dimensional space of the comoving reference frame in the static case (that is, when all time derivatives vanish):

$$C_G = \tfrac{8}{3}\pi(G/c^2)\rho\,. \tag{1.8.13}$$

In the case of an isotropic deformation of matter and vanishing rotation, the Gaussian curvature is

$$C_G = \tfrac{8}{3}\pi(G/c^2)\rho - \tfrac{1}{9}D^2\,. \tag{1.8.14}$$

These formulae are derived, for example, by Zel'manov (1959*b*).

1.9 THE COSMOLOGICAL CONSTANT

The general requirements usually placed on the equations of the theory of gravitation permit one to write a variational principle with the action in the form (where V is the four-dimensional volume)

$$S = -mc\int ds - \frac{c^3}{16\pi G}\left(\int R\,dV + \int 2\Lambda\,dV\right). \tag{1.9.1}$$

The corresponding field equations have the form

$$R_{ik} - \tfrac{1}{2}g_{ik}R - \Lambda g_{ik} = (\kappa/c^2)T_{ik}\,. \tag{1.9.2}$$

Here Λ is the so-called cosmological constant, and quantities that are proportional to it (ΛdV, Λg_{ik}) are called cosmological terms. These field equations obviously satisfy the condition of local Lorentz-invariance and include the equations of motion in the same sense as the equations without Λ do; $T^k{}_{i;k} = 0$, as before.

Einstein initially selected Λ in such a manner as to obtain a stationary cosmological solution with a mean density $T_0{}^0 = \rho c^2 = \text{const.}$ that is different from zero; for this, it is necessary that $\Lambda = 8\pi G\rho/3c^2$. After the discovery of the cosmological redshift, Einstein preferred the equations with $\Lambda \equiv 0$.

Both stationary and nonstationary cosmological solutions with $\Lambda \neq 0$ were investigated in detail before 1930; however, until 1967 there were no observational indications as to the necessity or even desirability of introducing Λ. Since 1967, observational data on quasars have suggested that Λ might not actually be zero, but might instead have a value of the order of $\Lambda \approx 10^{-55}\ \text{cm}^{-2}$. For details, see § 14.1. At present this hypothesis is not at all proved; in fact, it encounters difficulties in explaining the quasar observations. However, in the course of the discussions it has become apparent that the simplest assumption of $\Lambda \equiv 0$, while not refuted, has not been distinctly and uniquely proved. Predictably, difficult and extensive work will be required in years to come, to determine the quantity Λ, or at least to determine limits on it. After a genie is let out of a bottle (i.e., now that the possibility is admitted that $\Lambda \neq 0$), legend has it that the genie can be chased back in only with great difficulty. Even if the specific hypothesis that $\Lambda \approx 10^{-55}\,\mathrm{cm}^{-2}$ is invalidated, we will return not to the simplistic assumption that $\Lambda \equiv 0$, but to a careful $-a < \Lambda < b$, and to a gradual and painful decrease of the limits a and b.

How can the physical meaning of the cosmological constant be understood? Why, in fact, is it interesting for physics as a whole?

One approach was prompted by the dimensions of Λ (cm^{-2}). In this approach one views Λ as the curvature of empty space. But the theory of gravitation links the curvature to the energy, momentum, and pressure of matter. Putting the terms with Λ onto the right side of the field equation, we obtain

$$R_{ik} - \tfrac{1}{2}g_{ik}R = (8\pi G/c^4)T_{ik} + g_{ik}\Lambda\,. \tag{1.9.3}$$

The assumption that $\Lambda \neq 0$ means that empty space creates a gravitational field identical with that in the theory with $\Lambda = 0$, but with matter filling all of spacetime with a mass density $\rho_\Lambda = (c^2\Lambda/8\pi G)$, an energy density $\epsilon_\Lambda = (c^4\Lambda/8\pi G)$, and a pressure $P_\Lambda = -\epsilon_\Lambda$. In this sense one can speak about the energy density and the pressure (stress tensor) of the vacuum.

Notice that our assumptions about ϵ_Λ and P_Λ were formulated in such a manner that the relativistic invariance of the theory was not broken; ϵ_Λ and P_Λ are the same in all coordinate systems moving relative to each other (Lorentz-transformed).

These quantities do not make themselves felt either in elementary-particle experiments or in atomic and molecular physics: the vacuum energy of the vessel in which an experiment takes place plays the role of a constant term which can be canceled in the law of energy conservation.

The only sort of phenomenon in which ϵ_Λ and P_Λ manifest themselves is a gravitational one. In this case ϵ_Λ and P_Λ "work" not only in empty space; they are, as is clear from the above formulae, full and equal members of the field equations even when normal matter is present. Thus, a Cavendish experiment in principle can serve for the discovery and measurement of ϵ_Λ and P_Λ. The attraction of a sphere of lead depends on the sum of the lead density ($11\ \mathrm{g\ cm}^{-3}$), and the vacuum density ($|\rho_\Lambda|$ is less than 10^{-28} g cm^{-3}), integrated over the volume of the sphere.

A practical measurement of the influence of P_Λ and ϵ_Λ is not possible, either in laboratory experiments or in observations of planetary motions in the solar system or stellar motions in the Galaxy. Indeed, the mean density of solar-system matter in a sphere with the radius of the Earth's orbit is $\langle\rho\rangle \approx 10^{-7}\ \mathrm{g\ cm}^{-3}$. The mean density of matter in the Galaxy is of the order of 10^{-24} g cm^{-3}. The influence of ρ_Λ will be significant only on the largest of scales—the scale of the whole Universe—i.e., in cosmology.

Let us dwell on the nature of Λ. One can take the viewpoint that a certain Λ and corresponding ρ_Λ, ϵ_Λ, and P_Λ are universal constants that do not require any further explanation. A different viewpoint is also conceivable: assume that in some zeroth-order approximation $\Lambda = \rho_\Lambda = \epsilon_\Lambda = P_\Lambda \equiv 0$. Higher-order values that are different from zero and characterize the vacuum might be derivable from considerations of the theory of elementary particles. There is no such derivation at the present time, any more than there is a theory proving that $\Lambda = 0$.

We consider the second point of view in more detail. The very first attempts to quantize the electromagnetic field led to the paradoxical conclusion of an infinite energy density for the vacuum. The vacuum is defined as the lowest energy state of the system considered, and can be characterized by the Maxwell equations. The particles—photons in this case—represent the elementary excitations of the system. In the analogous quantum-mechanical problem of the motion of the atomic nuclei in a crystal lattice, the situation is similar: the elementary excitations are phonons (sound quanta), and the ground state of the crystal is that which contains no phonons, i.e., which has a temperature of absolute zero. This state is similar to the vacuum.

The ground-state energy in the case of a crystal has a quite certain value which can be measured. The difference in ground-state energies for different isotopes of the same element leads to the dependence of the heat of evaporation on the atomic weight of the isotope. In the simplest variant of field theory the ground-state energy of the vacuum is infinite. However, the theory can be reformulated so that the ground-state energy of the free field becomes identically zero.

In classical Maxwell theory the energy density is equal to $\epsilon = (E^2 + H^2)/8\pi$, where E and H are the electric and magnetic field strengths. As has been emphasized by Berestetsky, Lifshitz, and Pitaevsky (1969), there is no formulation of quantum electrodynamics in which the mean value of E^2 or H^2 is zero in the vacuum (i.e., far from all charges and in the absence of real photons). Consequently, in order to formulate these theories with the help of the usual products of operators, in such a manner that in vacuum $\epsilon = 0$, one must abandon the classical relation between ϵ and the fields.

A second source of the energy of the vacuum arises from the Dirac theory of the electron: the concept of filled levels with negative energy leads literally to an infinite negative value for the energy density. In this case also the theory was soon reformulated to make ϵ identically zero for the vacuum of noninteracting particles. However, it is not at all guaranteed that, when the interaction of the particles is taken into account, the vacuum energy will remain zero. The peculiarity of present-day theory is that the interaction between particles is effective not only when real particles are present to interact, but also when the particles are virtual.

In the above discussion we must remind ourselves that the term "interaction" is not used in the sense of classical physics. There one speaks about the interaction of two colliding bodies, and about the (Coulomb) interaction of a proton and electron. In quantum field theory one speaks about a four-fermion interaction when a neutron breaks up into an electron, a proton, and a neutrino; and one speaks about a photon-electron interaction when an electron produces a photon.

A freely moving electron cannot emit a real photon which is capable of being detected far from the electron. However, one can say that the free electron emits and then reabsorbs photons, and that this causes a change in the properties of the electron (for example, its mass, its magnetic moment, etc.), as has been proved by the Lamb-Rutherford experiment.

Experimental measurement of the change in the electron mass is impossible because there are no experiments which could, even in principle, measure the mass of an electron stripped of its photons. However, the change in the electron's magnetic moment has been corroborated by experiment, with all the precision which is now available.

There are many other processes similar to the above which take place in the vacuum—for example, the creation and annihilation of electron-positron pairs with the absorption and emission of γ-radiation.

The theory of the vacuum state and its properties is not so simple and obvious today as it was about fifty years ago!

Several possible viewpoints may be delineated.

The first is the supposition that the energy of the vacuum is identically zero when fields and interactions are ignored. When they are taken into account, the vacuum energy becomes nonzero; it acquires an additive constant when processes involving real particles are investigated. Particle theory based on this viewpoint faces the problem of calculating all observable processes in a manner which yields answers that do not depend on the unknown—or indeterminate, or even infinite—energy of the vacuum.

It is in this manner that Feynman formulated the problem and devised a successful solution. In his formulation the transition amplitude A_{12} (amplitude for particles plus vacuum in the initial state to change into particles plus vacuum in the final state) is divided by the transition amplitude A_v (vacuum → vacuum); and only the ratio A_{12}/A_v represents the real value corresponding to the real particle interaction. Such a method of avoiding the question of the energy of the vacuum is good everywhere except in gravitation theory! The energy density of the vacuum in gravitational problems is, as we mentioned above, a real, measurable quantity!

There is a second, so-called axiomatic, viewpoint on the theory of particles: it is taken as an axiom that the energy density of the vacuum and the corresponding pressure are identically zero.

We have no objection if such a statement is postulated openly as one possibility among others. However, one often meets in the literature statements that such an axiom is necessary; that it and no other is in agreement with relativistic invariance. Such a supposition is quite wrong. It was already noted above that the relation between the pressure and energy density of the vacuum which is provided by the cosmological constant, $P_\Lambda = -\epsilon_\Lambda$, is relativistically invariant.

We shall show below how particle theory can give an order-of-magnitude estimate of ϵ_Λ which is nonzero, while maintaining relativistic invariance.

An error frequently committed arises from introducing a particular normalizing volume V, and examining the energy $E = V\epsilon$. The (three-dimensional) momentum $\boldsymbol{P}$ of the vacuum is obviously zero because there is no preferred direction for it. Energy and momentum comprise the four-dimensional vector $(E, \boldsymbol{P})$, which in our case, for a given normalizing volume, is $(E, \boldsymbol{0})$. Such a combination is clearly noninvariant and will give $\boldsymbol{P} \neq 0$ in other coordinate systems, unless one chooses $E = 0$ (and, therefore $\epsilon = 0$).

The error lies in choosing a particular volume, since this itself violates relativistic invariance. An unbounded medium, and the vacuum in particular, must be characterized with an energy density which is the T_{00} component of the second-order stress-energy tensor. The tensor as a whole includes components $T_{0\alpha} = T_{\alpha 0}$ (where $\alpha = 1, 2, 3$ is a spatial index), which describe the energy flow and simultaneously the momentum density in space, $P_\alpha = T_{0\alpha}V$. The components $T_{\alpha\beta}$ of the stress-energy tensor correspond to the stress which enters into the theory of elasticity. In the case of a gas or fluid (without shear stresses) $T_{\alpha\beta} = \delta_{\alpha\beta}P$.

These well-known truths are repeated here in order to emphasize that the problem lies not in the question of whether the vacuum has an energy-momentum vector, but whether it has a stress-energy tensor. There is no relativistically invariant vector (it is identically zero), but there can perfectly well be a relativistically invariant tensor. It must have the form

$$\text{const.} \times \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{vmatrix},$$

and it is about precisely such a tensor that one speaks in the case $\Lambda \neq 0$. We cannot rule out such a tensor associated with the vacuum a priori.

The following questions remain:

1. Is there any principle which requires one to set $\Lambda \neq 0$?

2. Must we regard $\Lambda \neq 0$ as a new, independent, universal constant?

3. Is it possible to calculate Λ (even in order of magnitude) from the other universal constants?

An attempt to answer the third question, without touching on the first two, will be made below. This answer (based only on dimensional analysis and comparisons of orders of magnitude) will probably be useful in the construction of a future theory which is more nearly correct and logically consistent.

The theory of elementary particles yields a quantity with the dimensions of ϵ_Λ: from the fundamental constants of the theory one can construct the energy mc^2, the length $\hbar/mc$, the number density $n = (mc/\hbar)^3$, and the energy density $\epsilon_\Lambda = mc^2(mc/\hbar)^3$. This value of ϵ_Λ certainly is not applicable to gravitation theory; for the electron mass we get ϵ_Λ of the order of 10^{25} ergs cm^{-3}, and for the proton mass we get ϵ_Λ of the order of 10^{38}. Apparently the subconscious aversion of physicists to $\Lambda \neq 0$ has been based on the comparison of such a "theoretical" ϵ_Λ with what is admissible in cosmology: If we can't take a large Λ, then let there be none at all!

Under the influence of suggestions from astronomers, it was pointed out recently (Zel'dovich 1967*b*) that a sensible ϵ_Λ can be derived by multiplying $mc^2(mc/\hbar)^3$ by the dimensionless quantity $Gm^2/\hbar c$ which is characteristic of gravitation. We can interpret this expression in the following manner: In the vacuum, virtual particles with mass m and average separation $\lambda = \hbar/mc$ are created. Their total proper energy is identically zero, so the gravitational interaction of the neighboring particles determines the energy density of the vacuum:

$$\epsilon_\Lambda = (Gm^2/\lambda)(1/\lambda^3) = Gm^6c^4/\hbar^4\,,$$

corresponding to the relations mentioned above.

The connection between the theory of particles and the cosmological constant is treated in greater detail by Zel'dovich (1968).

There is another side to the problem: Often it is said that $\Lambda \neq 0$ means that gravitons have a nonzero rest mass. But $\Lambda \neq 0$ leads also to the result that, even in the absence of matter, spacetime cannot be flat everywhere. In curved spacetime the very definition of the mass of a graviton is no longer clear. For a thorough discussion of this point, see Treder (1968).

1.10 NEWTON'S LAW AND WEAK GRAVITATIONAL FIELDS

Consider a weak gravitational field. In this case one obviously can construct a reference frame in which the components of the metric tensor differ very little from their Galilean values [denoted by the superscript (0)]:

$$\left.\begin{aligned} g_{ik} &= g^{(0)}{}_{ik} + h_{ik}\,, \\ g^{(0)}{}_{\alpha\beta} &= -\delta_{\alpha\beta}\,, \quad g^{(0)}{}_{0\alpha} = 0\,, \quad g^{(0)}{}_{00} = 1\,. \end{aligned}\right\} \tag{1.10.1}$$

Actually, the Gaussian curvature (see § 1.4) of space, C_G, at the center of the star is larger than it is at other places. The quantity C_G is given by the formula (see eq. [1.8.13]):

$$C_G = (8\pi/3)(G\rho/c^2) . \qquad (3.3.6)$$

Inasmuch as the density is a maximum at the center of the star, C_G is also a maximum there. Of course, we should not assume on the basis of equation (3.3.6) that space is Euclidean outside the star (where $\rho = 0$), when the gravitational field near the star is strong. Formula (3.3.6) gives only the mean curvature of space averaged over all two-dimensional directions; and this mean curvature actually is zero. However, beyond the star, the Riemannian space curvature (see § 1.4) is nonzero; and, depending on the two-dimensional direction, it may be positive or negative. At the center of the star, all directions are equivalent; there the curvature for any orientation is given by formula (3.3.6), and it is always positive.

The gravitational field obtained by joining the star's interior and exterior solutions has no physical singularities of the type exhibited on the Schwarzschild sphere; everywhere

$$1 \le e^{\lambda} < \infty , \quad \text{and} \quad 0 < e^{\nu} < 1 .$$

3.4 THE RADIAL MOTION OF LIGHT RAYS AND OF ULTRA-RELATIVISTIC PARTICLES

Consider now the radial motion, in a spherical gravitational field, of particles with the fundamental speed c; for example, photons and neutrinos. Since an observer can locally introduce coordinates for which $ds^2 = c^2dt^2 - dx^2 - dy^2 - dz^2$, and since the speed of light c is independent of reference frame, an equation of motion for a particle with zero rest mass is $ds = 0$ in any coordinate system. Because of spherical symmetry, a particle that starts out moving radially will always move radially. Along such a particle's trajectory θ and ϕ are constant, so the condition $ds = 0$ implies

$$dr/dt = ce^{(\nu-\lambda)/2} . \qquad (3.4.1)$$

Everywhere inside the star $e^{(\nu-\lambda)/2} < 1$. Outside the star, in vacuum, $e^{(\nu-\lambda)/2} = 1 - 2GM/c^2r < 1$; and this quantity approaches unity as $r \to \infty$. It follows, for example, that for a neutrino emitted from the star's center, the radial coordinate r, as a function of the time t of a distant observer, must have the form shown in Figure 7. The dotted line in this diagram indicates the motion of the neutrino in the absence of a gravitational field.

In the future we will frequently consider the motion of bodies in the immediate vicinity of the gravitational radius of the central body (where the field is especially strong), and even in the limit $r \to r_g$. As noted in § 3.2, a static body cannot have dimensions $r \le r_g$; and Bondi (1964) has demonstrated that the dimensions of a body in equilibrium cannot be smaller than

$ (

different in different directions.
The curvature averaged over all directions is called the Gaussian curvature of the three-space.

Pág. 28(45) – Gaussian curvature (cf. § 1.4)

$$C_G = \frac{8}{3}\pi(G/c^2)\rho.$$

Pág. 89(106) Gaussian curvature of space, C_G, at the center of the star is larger than it is at other places.

$$C_G = (8\pi/3)(G\rho/c^2).$$

– String Gravity and Physics at the Planck Energy Scale (1996).pdf

Pág. 45(60) –
Quadratic curvature corrections (in terms of the Gauss – Bonnet term) to the Einstein's action are included [36].

Pág. 122(134) – 1.1
Point particles and open strings in magnetic field (exactly solvable)
becomes gaussian if the field

strength is constant,

Pag. 179 (189) 4.

Covariant Double-null Dynamics

The geometrical framework is the embedding formalism of Gauss and Codazzi for the foliation of spacetime by spacelike hypersurfaces.

Pag. 259 (266)

Euclidean Hamiltonian constraint [13]

$$R + K^{ab}K_{ab} - K^2 = 0,$$

rather than the Lorentzian Hamiltonian constraint

$$R - K^{ab}K_{ab} + K^2 = 0,$$

...

Pag. 266 (273)

if we had only the Gauss and the diffeomorphism constraints, the total volume of Σ would be a Dirac observable.

---

https://archive.org/details/@librosmir

$
-256- - -* ! - $ 7 !A 3 ! " 3 ! ! .
= 3 ! 1 8 ! " //B 1"
('$ B 6 !" ! A!

$$
\begin{aligned}
R_{\mp} = &\quad C_1' \, \eta^{\pm \frac{1}{\sqrt{D-1} \mp 1}} \to 0 \\
e^{\Psi_{\mp}} = &\, C_2' \, \eta^{-\frac{\sqrt{D-1}}{\sqrt{D-1} \mp 1}} \to \begin{cases} \infty \\ 0 \end{cases}
\end{aligned}
\tag{6.16}
$$

Where C_1' and C_2' are constants. The branch $(-)$ would describe an expanding non-inflationary behaviour near the initial singularity $T = 0$, while the branch $(+)$ describes a 'big crunch' situation and is rather unphysical.

Similarly, for $R \to \infty$ and $e^{\Phi} \to \infty$, we find

$$
\begin{aligned}
R_{\mp} = &\, D_1 \, T^{\mp 1/\sqrt{D-1}} \to \infty \quad , \\
e^{\Psi_{\mp}} = &\quad D_2 \, T^{-1} \to \begin{cases} \infty \\ 0 \end{cases}
\end{aligned}
\tag{6.17}
$$

Where D_1 and D_2 are constants. Here again, the branches $(-)$ and $(+)$ correspond to $T \to 0$ and to $T \to \infty$ respectively, but now in both regimes $R_{\mp} \to \infty$ and $e^{\Phi_{\mp}} \to \infty$. (In this limit, one is not guaranteed that U can be consistently neglected). In terms of the conformal time, eqs.(6.17) read

$$
\begin{aligned}
R_{\mp} = &\quad D_1' \, \eta^{\mp \frac{1}{\sqrt{D-1} \pm 1}} \to \infty \\
e^{\Psi_{\mp}} = &\, D_2' \, \eta^{-\frac{\sqrt{D-1}}{\sqrt{D-1} \pm 1}} \to \begin{cases} \infty \\ 0 \end{cases}
\end{aligned}
\tag{6.18}
$$

The branch $(+)$ describes a noninflationary expanding behaviour for $T \to \infty$ faster than the standard matter dominated expansion, while the branch $(-)$ describes a super-inflationary behaviour $\eta^{-\alpha}$, since $0 < \alpha < 1$, for all D.

The behaviours (6.15) for $R_{\mp} \to 0$ and (6.17) for $R_{\mp} \to \infty$ are related by duality $R \leftrightarrow 1/R$.

B. String driven inflation?

Let us consider now the question of whether de Sitter spacetime may be a self-consistent solution of the effective string equations (6.7) with the string sources included. The strings in cosmological universes like de Sitter spacetime have the equation of state (5.4)-(5.6). Since $e^{\Psi} = e^{\Phi} \, R^{1-D}$:

$$
\bar{\rho} = e^{\Psi} \left(u \, R + \frac{d}{R} + s \right) \tag{6.19}
$$

$$
\bar{p} = \frac{e^{\Psi}}{D-1} \left(\frac{d}{R} - u \, R \right) \tag{6.20}
$$

In the absence of dilaton potential and cosmological constant term, the string sources do not generate de Sitter spacetime as discussed in sec. V.A. We see that for $U = c = 0$, and $R = e^{HT}$, eqs.(6.11) yields to a contradiction (unless $D = 0$) for the value of Ψ , required to be $-HT +$ constant.

A self-consistent solution describing asymptotically de Sitter spacetime self-sustained by the string equation of state (6.19)-(6.20) is given by

$$\begin{aligned} R &= e^{HT} \quad , \quad H = \text{constant} > 0 \quad , \\ 2U - c &= D\,H^2 = \text{constant} \\ \Psi_{\pm} &= \mp HT \pm i\pi + \log \frac{(D-1)\,H^2}{\rho_{\pm}} \\ \rho_+ &\equiv u \quad , \quad \rho_- \equiv d \end{aligned} \tag{6.21}$$

The branch Ψ_+ describes the solution for $R \to \infty$ ($T \to +\infty$), while the branch Ψ_- corresponds to $R \to 0$ ($T \to -\infty$). De Sitter spacetime with lorentzian signature self-sustained by the strings necessarily requires a constant imaginary piece $\pm i\pi$ in the dilaton field. This makes $e^{\Psi} < 0$ telling us that the gravitational constant $G \sim e^{\Psi} < 0$ here describes antigravity.

Is interesting to notice that in the euclidean signature case, i. e. $(+++\ldots++)$, the Ansatz $\dot{H} = 0$, $2U - c$ =constant, yields a constant curvature geometry with a real dilaton, but which is of Anti-de Sitter type. This solution is obtained from eqs.(6.20)-(6.21) through the transformation

$$\hat{X}^0 = iT \quad , \quad \hat{H} = -iH \quad , \quad X^i = X^i \quad , \quad \Psi = \Psi \tag{6.22}$$

which maps the Lorentzian de Sitter metric into the positive definite one

$$d\hat{s}^2 = (d\hat{X}^0)^2 + e^{\hat{H}\hat{X}^0}\,(d\vec{X})^2. \tag{6.23}$$

The equations of motion (6.11) within the constant curvature Ansatz ($\dot{\hat{H}} = \ddot{\Psi} = 0$) are mapped onto the equations

$$\begin{aligned} (D-1)\,\hat{H}^2 - \left.\frac{\partial U}{\partial \Psi}\right|_R &= \bar{\rho} \\ \hat{H}\,\frac{d\Psi}{d\hat{X}^0} + \frac{R}{D-1}\left.\frac{\partial U}{\partial R}\right|_{\Psi} &= \bar{p} \\ -(\frac{d\Psi}{d\hat{X}^0})^2 + (D-1)\,\hat{H}^2 - 2\,\bar{\rho} - 2U + c = 0 \, , \end{aligned} \tag{6.24}$$

with the solution

$$\begin{aligned} R &= e^{\hat{H}\hat{X}^0} \quad , \quad \hat{H} = \text{constant} > 0 \quad , \\ c - 2U &= D\,\hat{H}^2 = \text{constant} \\ \Psi_{\pm} &= \mp \hat{H}\hat{X}^0 + \log \frac{(D-1)\,\hat{H}^2}{\rho_{\pm}} \\ \rho_+ &\equiv u \quad , \quad \rho_- \equiv d \end{aligned} \tag{6.25}$$

Both solutions (6.25) and (6.21) are mapped one into another through the transformation (6.22).

It could be recalled that in the context of (point particle) field theory, de Sitter spacetime (as well as anti-de Sitter) emerges as an exact selfconsistent solution of the semiclassical Einstein equations with the back reaction included [34] - [35]. (Semiclassical in this context, means that matter fields including the graviton are quantized to the one-loop level and coupled to the (c-number) gravity background through the expectation value of the energy-momentum tensor T_A^B . This expectation value is given by the trace anomaly: $< T_A^A >= \bar{\gamma}\ R^2$). On the other hand, the α' expansion of the effective string action admits anti-de Sitter spacetime (but not de Sitter) as a solution when the quadratic curvature corrections (in terms of the Gauss-Bonnet term) to the Einstein action are included [36]. It appears that the corrections to the anti-de Sitter constant curvature are qualitatively similar in the both cases, with α' playing the rôle of the trace anomaly parameter $\bar{\gamma}$ [35].

The fact that de Sitter inflation with true gravity $G \sim e^{\Psi} > 0$ does not emerge as a solution of the effective string equations does not mean that string theory excludes inflation. What means is that the effective string equations are not enough to get inflation. The effective string action is a low energy field theory approximation to string theory containing only the *massless* string modes (*massless* background fields).

The vacuum energy scales to start inflation (physical or true vacuum) are typically of the order of the Planck mass [26] - [27] where the effective string action approximation breaks down. One must consider the massive string modes (which are absent from the effective string action) in order to properly get the cosmological condensate yielding de Sitter inflation. We do not have at present the solution of such problem.

TABLE 4. Asymptotic solution of the string effective equations (including the dilaton).

EFFECTIVE STRING EQUATIONS SOLUTIONS IN COSMOLOGY

Effective String equations	$R(T) \to 0$ behaviour	$R(T) \to \infty$ behaviour
$T \to 0$	$\simeq T^{+1/\sqrt{D-1}}$	$\simeq T^{-1/\sqrt{D-1}}$
$T \to \infty$	$\simeq T^{-1/\sqrt{D-1}}$	$\simeq T^{+1/\sqrt{D-1}}$

VII. MULTI-STRINGS AND SOLITON METHODS IN DE SITTER UNIVERSE

Among the cosmological backgrounds, de Sitter spacetime occupies a special place. This is, in one hand relevant for inflation and on the other hand string propagation turns to be specially interesting there [2] - [8]. String unstability, in the sense that the string proper length grows indefinitely is particularly present in de Sitter. The string dynamics in de Sitter universe is described by a generalized sinh-Gordon model with a potential unbounded from below [4]. The sinh-Gordon function $\alpha(\sigma, \tau)$ having a clear physical meaning : $H^{-1}e^{\alpha(\sigma,\tau)/2}$ determines the string proper length. Moreover the classical string equations of motion (plus the string constraints) turn to be integrable in de Sitter universe [4,5]. More precisely, they are equivalent

(%' 1! ! " 1 ! "
L! 3 A!D

magnetic background is accompanied by a curvature of space-time. In spite of that, certain closed string magnetic models can be solved exactly. In particular, these are static flux tube type configurations with approximately uniform magnetic field which generalize the Melvin solution of the Einstein-Maxwell theory [1]. Such backgrounds are exact solutions of (super)string theory [2,3] and, moreover, the spectrum of the corresponding unitary conformal string models can be explicitly determined [3,4].

Below we shall first review the solution [4] of this class of magnetic models. In the last section we shall discuss related compact models which may be used for string compactifications and the issue of supersymmetry breaking induced by 'magnetic' twist parameters.

Before turning to closed string theory let us first recall the solution of similar uniform magnetic field problems in particle (field) theory and open string theory.

1.1. Point particles and open strings in magnetic field

The reason why the quantum-mechanical or field-theoretical problem of a particle in a uniform abelian (electro)magnetic field is exactly solvable is that the action $I = \int d\tau[\dot{x}^\mu \dot{x}^\mu + ie\dot{x}^\mu A_\mu(x)]$ (which determines the Hamiltonian in quantum mechanics and the heat kernel in field theory) becomes gaussian if the field strength is constant, $A_\mu = -\frac{1}{2}F_{\mu\nu}x^\nu$, $F_{\mu\nu} = const.$ Assuming the magnetic field is directed along x^3-axis (so that $F_{ij} = H\epsilon_{ij}$, $i,j = 1,2$) and introducing $x = x_1 + ix_2$, $a = -i(\partial_x^* + \frac{1}{4}eHx)$, $a^\dagger = -i(\partial_x - \frac{1}{4}eHx^*)$, $[a, a^\dagger] = \frac{1}{2}eH$, one can reduce the corresponding quantum-mechanical problem to a free oscillator one. The resulting energy spectrum is the special ($S = 0$) case of the Landau spectrum for a particle of charge e, mass M_0 and third component of the spin S (we assume $eH > 0$)

$$E^2 = M_0^2 + p_3^2 - 2eHJ \ , \quad J \equiv -l - \tfrac{1}{2} + \tfrac{1}{2}gS \ , \quad l = 0, 1, 2, \ldots \ . \qquad (1.1)$$

Here l is the Landau level number (which replaces the continuous momenta p_1, p_2) and g is the gyromagnetic ratio which is $1/S$ for minimally coupled particles but can be equal to 2 for non-minimally coupled ones. Thus E^2 can become negative for large enough values of H, e.g., $H > H_{crit} = M_0^2/e$ for spin 1 charged states. That applies, for example, to W-bosons in the context of electroweak theory [5] suggesting the presence of a transition to a phase with a W-condensate. In the case of unbroken gauge theory with massless charged vector particles the instability is present for any (e.g., infinitesimal) value of the magnetic field [5]. This infra-red instability of a magnetic background is not cured by supersymmetry, i.e. it remains also in supersymmetric gauge theories (e.g., in ultra-violet finite $N = 4$ supersymmetric Yang-Mills theory) since the small fluctuation operator for the gauge field $-\delta_{\mu\nu}D^2 - 2F_{\mu\nu}$ still has negative modes due to the 'anomalous magnetic moment' term. This is not surprising given that the magnetic spin-dependent coupling breaks supersymmetry.

The meaning of this instability is that the expansion is carried out near a classical solution of the Yang-Mills equations (abelian $A_i = -\frac{1}{2}H\epsilon_{ij}x^j$ belonging to the Cartan subalgebra) which is *not* a vacuum one: the energy is proportional to H^2 and thus is minimal for $H = 0$. As a result, a non-zero magnetic field will tend to dissipate. The presence of tachyonic modes in the magnetic background with infinitesimal (but non-vanishing H) does not indicate an instability of the trivial vacuum but only that of a configuration with $H \neq 0$ (this corresponds to a resummation of the expansion in small perturbations near the trivial vacuum).

The solution of the uniform magnetic field problem in the open string theory is similar to the one in the particle case. Indeed, the open string is coupled to the (abelian) vector field only at its ends, $I = \frac{1}{4\pi\alpha'}\int d^2\sigma\ \partial_a x^\mu \partial^a x^\mu + ie\int d\tau\ A_\mu(x)\dot{x}^\mu$, and thus for the abelian background $A_\mu = -\frac{1}{2}F_{\mu\nu}x^\nu$ the resulting gaussian path integral can be computed exactly [6]. This is a consistent 'on-shell' problem since $F_{\mu\nu} = \text{const}$ is an exact solution of the effective field equations of the open string theory. The corresponding 2d world-sheet theory is a conformal field theory [7] which can be solved explicitly in terms of free oscillators (thus representing a string generalization of the Landau problem in quantum particle mechanics). As a result, one is able to determine the spectrum of an open string moving in a constant magnetic field [7,8,9]. The Hamiltonian (L_0) of the open (super)string in a constant magnetic field is given by [7,9]

$$\hat{\mathcal{H}} = \tfrac{1}{2}\alpha'\big(-E^2 + p_\alpha^2\big) + \hat{N} - \gamma\hat{J}\ , \tag{1.2}$$

$$\hat{\gamma} \equiv \frac{2}{\pi}\Big|\arctan\big(\tfrac{1}{2}\alpha'\pi e_1 H\big) + \arctan\big(\tfrac{1}{2}\alpha'\pi e_2 H\big)\Big|\ , \quad 0 \leq \hat{\gamma} < 1\ .$$

Here e_1, e_2 are charges at the two ends of the string and H is the magnetic field, $F_{ij} = H\epsilon_{ij}$. $\hat{N} = 0, 1, 2...$ is the number of states operator and $\hat{J} = -l - \frac{1}{2} + S$ is the angular momentum operator of the open string in the $(1,2)$ plane. The energy spectrum is found from the constraint $\hat{\mathcal{H}} = 0$ and for small $e_i H$ is in agreement with (1.1) with $g = 2$. A novel feature of this spectrum as compared to the free (super)string one is the presence of tachyonic states above certain critical values of the magnetic field ($H_{crit} \sim \alpha'^{-1}\hat{N}/S \sim \alpha'^{-1}$). Thus the constant magnetic field background is unstable in the open string theory (as it is in the non-abelian gauge theory). A qualitative reason for this instability is that the free open string spectrum contains electrically charged higher spin massive particle states and the latter have (approximately) the Landau spectrum (1.1) ($\hat{\gamma} \approx \alpha'(e_1 + e_2)H$ for a weak field). The tachyonic states appear only in the bosonic (Neveu-Schwarz) sector and only on the leading Regge trajectory ($l = 0$, maximal S for a given mass level $\hat{N}$, $S = \hat{N} + 1$).

1.2. Closed strings and Melvin-type magnetic flux tube backgrounds

Similar background magnetic field problem is also exactly solvable in *closed* (super)string theories [10,3,4]. This may be unexpected since the abelian vector field is now coupled to the internal points of the string, and such interaction terms, e.g., $\Delta L = \partial y \bar{\partial} y + A_\mu(x) \partial x^\mu \bar{\partial} y + \tilde{A}_\mu(x) \bar{\partial} x^\mu \partial y + \ldots$, do not become gaussian for $A_\mu = -\frac{1}{2} F_{\mu\nu} x^\nu$. Here $y \in (0, 2\pi R)$ is a compact internal Kaluza-Klein coordinate that 'charges' the string. The spectrum of the closed string compactified on a circle contains states with arbitrarily large masses, spins *and* charges $Q_{L,R} = mR^{-1} \pm w\alpha'^{-1} R$, $(m, w = 0, \pm 1, \ldots)$. One difference compared to the open string case is that there are infinitely many possible values of charges. Another is that both Q_L and Q_R are, in general, non-vanishing and coupled to (a combination of) two abelian vector fields ($G_{5\mu}$ and $B_{5\mu}$).[1]

The important observation is that in contrast to the tree level abelian open string case, the $F_{\mu\nu} =$ const background in flat space does not represent a solution of a *closed* string theory, i.e. the above interaction terms added to the free string Lagrangians do not lead to conformally invariant 2d σ-models. Since the closed string theory contains gravity, a uniform magnetic field, which has a non-vanishing energy, must curve the space (as well as possibly induce other 'massless' background fields). One should thus first find a consistent conformal model which is a closed string analogue of the uniform magnetic field background in the flat space field (or open string) theory and then address the question of its solvability. Remarkably, it turns out that extra terms which should be added to the closed string action in order to satisfy the conformal invariance condition (i.e. to satisfy the closed string effective field equations) produce solvable 2d models. Like the particle and open string models, these models are exactly solvable in terms of free oscillators (one is able to find explicitly the spectrum, partition function, etc).

A simple example of an approximately uniform magnetic field background in the Einstein-Maxwell theory is the static cylindrically symmetric Melvin 'magnetic universe' or 'magnetic flux tube' solution [1]. It has R^4 topology and can be considered [12] as a gravitational analogue of the abelian Higgs model vortex [13] with the magnetic pressure (due to repulsion of Faraday's flux lines) being balanced not by the Higgs field but by the gravitational attraction. The magnetic field is approximately constant inside the tube and decays to zero at infinity in the direction orthogonal to x_3-axis. The space is not, however, asymptotically euclidean: the (ρ, φ) 2-plane orthogonal to x_3 asymptotically closes at large ρ (so the solution should be interpreted not as a flux tube embedded in an approximately flat space, but rather as a 'magnetic universe'). Several interesting features of the Melvin solution in the context of Kaluza-Klein (super)gravity (e.g., instability against monopole or magnetic black hole pair creation) were discussed in [12,14,15].

---

[1] As a result, the gyromagnetic ratio of closed string states is $g \leq 2$, e.g., $g = 1$ for standard non-winding Kaluza-Klein states [11,3].

(/ / ' A! ! 4 D

4. Covariant Double-null Dynamics

The classic analysis of Arnowitt, Deser and Misner [15] formulates gravitational dynamics in terms of the evolution of a spatial 3-geometry. The geometrical framework is the imbedding formalism of Gauss and Codazzi for the foliation of spacetime by spacelike hypersurfaces.

Quite often, however, —as in the analysis of Cauchy horizon singularities, but also in other contexts—one encounters circumstances where a lightlike foliation is especially suitable. Because of various degeneracies that arise in the lightlike case, the imbedding relations are quite different and the situation not quite as familiar and under control. To bypass the degeneracies, one is forced to fall back to a foliation of co-dimension 2, by spacelike 2-surfaces.

Several such (2+2)-formalisms are extant, notably Geroch, Held and Penrose [16], d'Inverno and Smallwood [17] and Hayward [18]. Of course, all have basically the same content, but they take very different forms.

The distinguishing feature of the approach I shall present here is that it maintains manifest two-dimensional covariance and operates directly with geometrically significant objects. Two-dimensional covariance permits reduction of the Einstein field equations to an especially concise and geometrically transparent form: the ten Ricci components are embraced in a set of just three compact, two-dimensionally covariant expressions.

$(2 + 2)$-formalisms are a versatile tool having a range of applications which go far beyond the specific problem of Sec. 3. They are useful in the analysis and numerical integration of the characteristic initial-value problem of general relativity, the dynamics of strings, of real and apparent horizons, the study of Planck-energy collisions and light-cone quantization. The covariant formalism has been applied by our group ([20]) to elucidate the singular geometry of the black hole Cauchy horizon.

5. Basic Metric Notions

I shall briefly sketch the elements of the formalism.

We consider a double foliation of spacetime by a net of two intersecting families of lightlike hypersurfaces Σ^0 (with equations $u^0 =$ const.) and Σ^1 (given by $u^1 =$ const.), where $u^A(x^\alpha)$ $(A, B, \ldots, = 0, 1;\ \alpha, \beta, \ldots, = 1, \ldots, 4)$ are a given pair of scalar fields over spacetime, with lightlike gradients:

$$\nabla u^A \cdot \nabla u^B = e^{-\lambda} \eta^{AB},$$

and where the matrix

$$\eta^{AB} = \eta_{AB} = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$$

will be used to raise and lower upper-case Latin indices, and $\lambda(x^\alpha)$ is a scalar function. The generators $\ell^{(A)}_\alpha$ of Σ^A are conveniently defined as

$$\ell^{(A)} = e^\lambda \nabla u^A.$$

Two hypersurfaces Σ^0 and Σ^1 intersect in a 2-surface S, with parametric equations

$$x^\alpha = x^\alpha(u^A, \theta^a) \qquad (a, b, \cdots = 2, 3)$$

where (θ^2, θ^3) are intrinsic co-ordinates of S. Both generators $\ell^{(A)}$ are orthogonal to S.

Holonomic basis vectors $e_{(a)}$ and the intrinsic metric of S may now be defined:

$$e^\alpha_{(a)} = \frac{\partial x^\alpha}{\partial \theta^a}, \qquad g_{ab} = e_{(a)} \cdot e_{(b)}.$$

The matrix g_{ab} and its inverse g^{ab} are used to lower and raise lower-case Latin indices, so that $e^{(a)} = g^{ab} e_{(b)}$ are the dual basis vectors tangent to S.

Two-dimensional shift vectors s^a_A are defined by

$$s^a_A = \frac{\partial x^a}{\partial u^A}\, e^{(a)}_\alpha = -\ell^\alpha_{(A)}\, \frac{\partial \theta^a}{\partial x^\alpha}.$$

As in the Arnowitt-Deser-Misner formalism, the shift vector s^a_A measures how much one has to deviate from the normal direction $\ell_{(A)}$ to connect points on different 2-surfaces having the same intrinsic co-ordinates θ^a. An infinitesimal four-dimensional displacement dx^α can be decomposed as

$$dx^\alpha = \ell^\alpha_{(A)} du^A + e^\alpha_{(a)}(d\theta^a + s^a_A du^A).$$

Together with the completeness relation

$$g_{\alpha\beta} = e^{-\lambda}\eta_{AB}\ell^{(A)}_\alpha \ell^{(B)}_\beta + g_{ab} e^{(a)}_\alpha e^{(b)}_\beta$$

for the basis $(\ell^{(A)}, e_{(a)})$, this implies that the spacetime metric is decomposable as

$$g_{\alpha\beta}dx^\alpha\, dx^\beta = e^\lambda \eta_{AB} du^A\, du^B + g_{ab}(d\theta^a + s^a_A du^A)(d\theta^b + s^b_B\, du^B).$$

6. Covariant Geometrical Objects Embodying First Derivatives of Metric

Associated with its two normals $\ell_{(A)}$, a 2-surface S has two extrinsic curvatures, defined by

$$K_{Aab} = (\nabla_\beta \ell_{(A)\alpha}) e^\alpha_{(a)}\, e^\beta_{(b)}$$

and easily shown to be symmetric in a, b. (Since we are free to rescale the null vectors $\ell_{(A)}$, a certain scale-arbitrariness is inherent in this definition.)

A further basic geometrical property of the double foliation is given by the Lie bracket of $\ell_{(0)}$ and $\ell_{(1)}$. One finds

$$[\ell_{(B)}, \ell_{(A)}] = \epsilon_{AB}\omega^a e_{(a)} \tag{4}$$

where

$$\omega^a = \epsilon^{AB}(\partial_B s^a_A - s^b_B s^a_{A;b}),$$

the semicolon indicates two-dimensional covariant differentiation associated with metric g_{ab}, and ϵ_{AB} is the two-dimensional permutation symbol.

These are the "simplest" equations in the following sense: Among non-trivial gauge covariant expressions, the left side of the first equation is the only one which is at most linear in E and A; that of the second is the only one that is at most linear in E and quadratic in A; and, that of the third is the only one that is at most quadratic in each of E and A. (There is no gauge covariant expression which is linear in A and at most quadratic in E.) Somewhat surprisingly, the first two equations provide us precisely with the 6 of 7 constraints of general relativity: the Gauss constraint $\mathcal{G}_i$ which generates internal frame rotations and the vector constraint $\mathcal{V}_b$ that generates spatial diffeomorphisms [12]. The third equation is almost –but not quite– the last (scalar or Hamiltonian) constraint we seek. To see this, let us translate it in to geometrodynamical variables. It then reduces to the familiar *Euclidean* Hamiltonian constraint [13]

$$R + K^{ab}K_{ab} - K^2 = 0, \tag{3a}$$

rather than the Lorentzian Hamiltonian constraint

$$R - K^{ab}K_{ab} + K^2 = 0, \tag{3b}$$

which, as one might expect, carries opposite signs in front of terms which are quadratic in momenta. (Intuitively, the "Wick" rotation should map the momentum to i times the momentum.) The Gauss and the vector constraints are insensitive to the signature.

To get the Lorentzian theory, the most straightforward avenue is to consider *complex* connections ${}^{\mathbb{C}}A_a{}^i := \Gamma_a^i - iK_a^i$. This avenue has been pursued vigorously in the literature [8-11]. In this article, however, I will present another possibility which has arisen from a recent idea of Thomas Thiemann's and which may allow one to pass from the solutions to the Euclidean quantum constraints to the Lorentzian ones via a generalized Wick transformation. This strategy will be discussed in Section 5. For the moment, let us note only that $SU(2)$ connections can be regarded as the configuration variables also in general relativity. Thus, the classical configuration space is $\mathcal{A}/\mathcal{G}$, the space of suitably smooth connections modulo gauge transformations. Our task is to find the corresponding quantum configuration space and to develop functional calculus on it.

I will conclude this discussion with a curiosity: The Euclidean signature seems to be easier to handle in other approaches as well, in particular, Connes' non-commutative geometry and string theory. In the Connes' "geometrization" of the standard model, one is naturally led to work with the Euclidean signature and the extension of the framework to the Lorentzian regime is still to be worked out. In string theory, while one can allow the target space geometry to be Lorentzian, in all the discussions that I am aware of, the world sheet is kept Euclidean. In perturbation theory off Minkowskian backgrounds, as far as the S-matrix theory is concerned, this seems like an obvious extension of the strategy one adopts in field theory. However, in non-perturbative contexts, such as the discussion of mirror symmetry, restriction to Riemannian world-sheets appears artificial from a physical the expression of the operator is a sum over intersections. At each intersection, the graph has a number of edges. So, all that the operator can do is to act on the copies of groups associated with these edges. That is, thanks to diffeomorphism invariance, the action of operators is reduced to simple algebraic operations in the representation theory of $SU(2)$!

2. Since the two operators are self-adjoint, one can compute their spectra. *They are purely discrete!* Each of the Penrose spin network state is an eigenstate of *both* the operators. The area operators reduce to a sum of $SU(2)$-Laplacians, each of which has the familiar eigenvalues $\sqrt{j(j+1)}$. The eigenvalue of the volume operator on these states is, however, zero [30, 28]. Thus, higher valent spin networks are essential. Both operators again leave the (finite dimensional) Hilbert space associated with any of these networks invariant. They can therefore be diagonalized separately on each network. (Already in the 4-valent case, there exist states with non-zero volume.) This overall picture shows that *the microscopic geometry at Planck scale is very different from what the continuum picture suggests.* As we argued in the Introduction, this basic fact may well be the essential reason why the perturbation theory fails.

3. The geometrical observables can be defined on the full phase space of the classical theory; one need not restrict oneself to the constraint surface. Similarly, in the quantum theory, the analogous operators have been defined on the kinematic Hilbert space H_o, before imposing quantum constraints. The results are thus robust; they are not sensitive to the details of quantum dynamics. For example, they will continue to hold also in supergravity, and more generally, in any theory in which the triads and connections feature as a basic canonical pair.

4. There is, however, a subtlety. To see this, recall a basic difference between quantum mechanics and quantum field theory. In quantum mechanics, kinematics and dynamics are largely decoupled in the sense the same representation of the canonical commutation relations supports all interesting Hamiltonians. In quantum field theory, on the other hand, kinematics and dynamics are always weakly coupled: there are infinitely many *inequivalent* representations of the canonical commutation relations and, in general, different Hamiltonians are meaningful on different representations. In the present case, then, the key question is whether the quantum constraints of a given theory can be meaningfully imposed on H_o. Results to date indicate (but do not yet prove) that this is likely to be the case for constraints of general relativity. My general expectation is that this will also be the case for a class of theories such as supergravity which are "near general relativity" as far as geometry is concerned. What we have here is a glimpse into the nature of quantum geometry underlying *this* class of quantum gravity theories.

5. Note that $\hat{A}(S)$ and $\hat{V}(R)$ have been defined for *any* surface S and *any* region R in Σ and will therefore not be Dirac observables; they are not expected to commute with the quantum constraints. To obtain Dirac observables, one would have to specify S and R *intrinsically.* (For example, if we had only the Gauss and the diffeomorphism constraints, the total volume of Σ would be a Dirac observable.) A natural strategy is to specify S and R using matter fields (see, e.g.,[31]). In view of the Hamiltonian constraint, the problem of providing an explicit specification of

Relaciones y geometría analítica
López Quiles, Antonio
https://archive.org/details/
isbn_978684441514/

cont-

20211010/ Géométrie Contemporaine.../
_Modern Geometry - Methods and
Applications - Part I....pdf
Pág. 77(92) 8.1.1. Definition
Gaussian curvature K,

Pág. 81(96)
Recall that the ratio of these quadratic forms in the components of the tangent vector to the section of the surface by a plane orthogonal at the surface at a point (i.e. a "normal section"), gives the curvature of the section (i.e. "normal curvature") at that point (Corollary 8.1.3)
↳ Pág. 79(94)

8.3.1. Definition →
the product of the principal curvatures

(/ 4 7! ! A! +J

(iii) By (19), the induced metric g_{ij} on a surface $r = r(u, v)$ is given by

$$g_{12} = \langle r_u, r_v \rangle, \qquad g_{11} = \langle r_u, r_u \rangle, \qquad g_{22} = \langle r_v, r_v \rangle.$$

It is then easy to verify that $g = g_{11}g_{22} - g_{12}^2 = |[r_u, r_v]|^2$, whence the result. ☐

Using Definition 7.4.1 to give us a definition of area in an arbitrary 2-dimensional space with a Riemannian metric, we see that the concept of area, like arc length and angle, is defined in terms of the scalar product of vectors attached to each point, or, what amounts to the same thing, the metric g_{ij}.

7.5. Exercises

1. The torus T^2 in Euclidean 3-space can be realized as the surface of revolution obtained by revolving a circle about a straight line which it does not intersect, and which lies in the same plane. Find parametric equations for this torus, and calculate the induced metric on it.
2. Find the first fundamental form of the ellipsoid of revolution
$$\frac{x^2}{a^2} + \frac{y^2 + z^2}{b^2} = 1.$$
3. Find the metric induced on the surface of revolution
$$r(u, \varphi) = (\rho(u) \cos \varphi, \rho(u) \sin \varphi, z(u)).$$
Verify that its meridians ($\varphi = \text{const.}$) and circles of latitude ($u = \text{const.}$) form an orthogonal net. Find those curves which bisect the angles between the meridians and circles of latitude.
4. Find those curves on a sphere which intersect the meridians at a fixed angle α ("loxodromes"). Find the length of a loxodrome.
5. Let $F(x, y, z)$ be a smooth homogeneous function of degree n (i.e. $F(cx, cy, cz) = c^nF(x, y, z)$). Prove that away from the origin the induced metric on the conical surface $F(x, y, z) = 0$ is Euclidean.

§8. The Second Fundamental Form

8.1. Curvature of Curves on a Surface in Euclidean Space

Suppose we are given a surface in Euclidean 3-space and a non-singular point (x_0, y_0, z_0) on it. We shall assume initially that the z-axis is perpendicular to the tangent plane to the surface at the point (x_0, y_0, z_0), in which case the x-axis and y-axis will be parallel to it. The surface may then be given locally about (i.e. in a neighbourhood of) the point (x_0, y_0, z_0) by an equation of the form $z = f(x, y)$, where $z_0 = f(x_0, y_0)$ and

$$\left.\frac{\partial f}{\partial x}\right|_{\substack{x=x_0\\y=y_0}} = \left.\frac{\partial f}{\partial y}\right|_{\substack{x=x_0\\y=y_0}} = 0, \quad \text{i.e. grad } f\,|_{\substack{x=x_0\\y=y_0}} = 0. \tag{1}$$

We now consider the second differential of the function $z = f(x, y)$ (which we now assume has continuous second derivatives), i.e. $d^2l = f_{xx}\,dx^2 + 2f_{xy}\,dx\,dy + f_{yy}\,dy^2$. The matrix of this quadratic form, i.e. the matrix (a_{ij}) where $a_{ij} = f_{x^ix^j}$, $x^1 = x$, $x^2 = y$, is known as the *Hessian* of the function f.

8.1.1. Definition. Given a surface $z = f(x, y)$, and a point (x_0, y_0, z_0) on it at which grad $f = 0$, we define the *principal curvatures* of the surface at the point to be the eigenvalues of the matrix (a_{ij}) evaluated at the point. (These eigenvalues are real since (a_{ij}) is symmetric.) In addition we call $\det(a_{ij})$ at the point the *Gaussian curvature* K, and the trace of (a_{ij}), $\operatorname{tr}(a_{ij})$, at the point, the *mean curvature*, of the surface at (x_0, y_0, z_0).

Thus the eigenvalues k_1, k_2 say, of (a_{ij}) evaluated at (x_0, y_0, z_0) are the principal curvatures, while the Gaussian curvature is $K = k_1k_2 = a_{11}a_{22} - a_{12}a_{21}$, and the mean curvature is $a_{11} + a_{22} = k_1 + k_2$. As we shall see later (in Chapter 4), the Gaussian curvature of the surface is an "intrinsic" invariant of the surface, i.e. depends only on the internal metrical properties of the surface; this is Gauss' famous "Theorema Egregium."

The disadvantage of Definition 8.1.1 is that it tells us what the curvatures at a point of the surface are in terms of a rather special co-ordinate system, influenced by the point under scrutiny in that the z-axis was chosen perpendicular to the surface at the point. To define these quantities in terms of arbitrary co-ordinates, we turn to the theory of the curvature of curves lying in the surface.

Thus suppose the surface given in parametric form

$$r = r(u, v). \tag{2}$$

We saw in §7.3 that the vector $[r_u, r_v]$ is normal to the surface at each non-singular point; hence if m denotes a unit vector in the same direction, then $[r_u, r_v] = |[r_u, r_v]|m$. If now $r = r(u(t), v(t))$ is a curve on the surface, then we have $\dot{r} = r_u\dot{u} + r_v\dot{v}$, and

$$\ddot{r} = (r_{uu}\dot{u}^2 + 2r_{uv}\dot{u}\dot{v} + r_{vv}\dot{v}^2) + (r_u\ddot{u} + r_v\ddot{v}).$$

Using $r_u \perp m$ and $r_v \perp m$, we deduce that

$$\langle \ddot{r}, m\rangle = \langle r_{uu}, m\rangle\dot{u}^2 + 2\langle r_{uv}, m\rangle\dot{u}\dot{v} + \langle r_{vv}, m\rangle\dot{v}^2 = b_{11}\dot{u}^2 + 2b_{12}\dot{u}\dot{v} + b_{22}\dot{v}^2. \tag{3}$$

From (3) we conclude that: *In terms of local co-ordinates $u = x^1$, $v = x^2$ the projection $\langle \ddot{r}, m\rangle$ of the acceleration vector $\ddot{r}$ on the normal to the surface, is a quadratic form in the components $\dot{u}$, $\dot{v}$ of the tangent vector.*

If we put $b_{11} = L$, $b_{12} = M$, $b_{22} = N$, then in terms of differentials (3) can be rewritten as

$$\langle \ddot{r}, m\rangle\, dt^2 = b_{ij}\, dx^i\, dx^j = L(du)^2 + 2M\, du\, dv + N(dv)^2. \tag{4}$$

We call this expression for $\langle \ddot{r}, m\rangle\, dt^2$, the *second fundamental form of the surface* (2). (Note that its coefficients L, M, N depend only on the point of the surface under scrutiny, and not on the curve $r = r(u(t), v(t))$.)

Suppose now that the curve $r = r(u(t), v(t))$ is parametrized by the natural parameter, i.e. $t = l$, where l is arc length. (Thus with this parametrization the curve is a "unit-speed curve.") By the first of the Serret–Frenet formulae (see 5.2.3), we have

$$\ddot{r} = \frac{d^2 r}{dl^2} = kn,$$

where n is the principal normal to the curve, and k is its curvature. Hence $\langle \ddot{r}, m\rangle = k\langle n, m\rangle = k \cos \theta$, where θ is the angle between the unit vectors m and n. From this and (4) we obtain

$$k \cos \theta (dl)^2 = \langle \ddot{r}, m) \, dl^2 = L(du)^2 + 2M\, du\, dv + N(dv)^2 = b_{ij}\, dx^i\, dx^j.$$

Comparing this with the formula for the induced metric on the surface:

$$dl^2 = g_{ij}\, dx^i\, dx^j = E(du)^2 + 2F\, du\, dv + G(dv)^2,$$

where $u = x^1$, $v = x^2$, we conclude that

$$k \cos \theta = \frac{b_{ij}\dot{x}^i \dot{x}^j}{g_{ij}\dot{x}^i \dot{x}^j},$$

where here the dots denote differentiation with respect to any parameter t. In the convenient differential notation we may rewrite this as

$$k \cos \theta = \frac{b_{ij}\, dx^i\, dx^j}{g_{ij}\, dx^i\, dx^j}, \qquad x^1 = u, \quad x^2 = v, \tag{5}$$

(although, as usual with this notation, we need to bear in mind that strictly speaking here dx^i is not an independent differential, but just $\dot{x}^i\, dt$).

We restate (5) in words, as a

8.1.2. Theorem. *The curvature at a point of a curve on a surface in Euclidean 3-space, when multiplied by the cosine of the angle between the normal to the surface and the principal normal of the curve at the point, is the same as the ratio of the second and first fundamental forms in the components of the tangent vector to the curve at the point.*

8.1.3. Corollary. *If the curve is obtained by sectioning the surface with a plane normal to the surface at the point, then (since* $\cos\theta = \pm 1$*) the corresponding "normal curvature" is given by*

$$k = \frac{b_{ij}\dot{x}^i\dot{x}^j}{g_{ij}\dot{x}^i\dot{x}^j} \tag{6}$$

(where here we allow k to be negative if m and n have opposite directions).

8.2. Invariants of a Pair of Quadratic Forms

Our main achievement so far in this chapter, is the association with each point of a surface in Euclidean 3-space, of the pair of quadratic ("fundamental") forms

$$dl^2 = g_{ij}\,dx^i\,dx^j, \tag{7}$$

$$\langle \ddot{r}, m\rangle\,dt^2 = b_{ij}\,dx^i\,dx^j, \tag{8}$$

the first of which is positive definite. Our aim is to deduce from these quadratic forms invariants of the surface. What are the known algebraic invariants of a pair of quadratic forms?

To answer this, consider any pair of quadratic forms in the plane with co-ordinates x, y, having matrices relative to these co-ordinates

$$G = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}, \qquad Q = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}, \tag{9}$$

(using the same symbols as in (7) and (8)), where G is positive definite and $g_{12} = g_{21}$, $b_{12} = b_{21}$. (Note that in contrast with the general context of (7) and (8), the g_{ij} and b_{ij} are assumed here to be fixed numbers.) Next consider the equation

$$\det(Q - \lambda G) = 0, \tag{10}$$

which, when written out in detail, becomes

$$(b_{11} - \lambda g_{11})(b_{22} - \lambda g_{22}) - (b_{12} - \lambda g_{12})^2 = 0. \tag{11}$$

We call the roots λ_1, λ_2 of this equation *the eigenvalues of the pair of quadratic forms.*

For each $i = 1, 2$ we look for solutions of the system of equations

$$\left.\begin{aligned}(b_{11} - \lambda_i g_{11})\xi_i^1 + (b_{12} - \lambda_i g_{12})\xi_i^2 = 0,\\ (b_{12} - \lambda_i g_{12})\xi_i^1 + (b_{22} - \lambda_i g_{22})\xi_i^2 = 0,\end{aligned}\right\} \tag{12}$$

in the unknowns ξ_i^1, ξ_i^2. Since λ_1, λ_2 satisfy (11), the coefficient matrix of each of the systems (12) is singular, so that there exist non-trivial solutions, say

$$f_1 = \xi_1^1 e_1 + \xi_1^2 e_2, \qquad f_2 = \xi_2^1 e_1 + \xi_2^2 e_2,$$

(/B 5 ! ! ! > !" !

and all solutions will be scalar multiples of the appropriate one of f_1, f_2 (i.e. the solution space of (12) has dimension 1). The directions of the vectors f_1 and f_2 are called the *principal directions* in the plane of the pair of quadratic forms, that of f_1 corresponding to λ_1, and that of f_2 to λ_2.

We now equip our plane with a scalar product defined in terms of the positive definite quadratic form with matrix G, by setting $\langle e_i, e_j \rangle = g_{ij}$, where $e_1 = (1, 0)$, $e_2 = (0, 1)$.

8.2.1. Lemma. *If the eigenvalues of the pair of quadratic forms* (9) *are distinct, then the principal directions are orthogonal.*

Proof. We wish to show that the two vectors $f_1 = \xi_1^1 e_1 + \xi_1^2 e_2$ and $f_2 = \xi_2^1 e_1 + \xi_2^2 e_2$ are orthogonal, i.e. that

$$\langle f_1, f_2 \rangle = g_{ij}\xi_1^i \xi_2^j = 0.$$

Multiplying the first of the equations (12), with $i = 1$, by ξ_2^1, and the second by ξ_2^2, we obtain

$$(b_{11} - \lambda_1 g_{11})\xi_1^1\xi_2^1 + (b_{12} - \lambda_1 g_{12})\xi_1^2\xi_2^1 = 0,$$

$$(b_{11} - \lambda_1 g_{12})\xi_1^1\xi_2^2 + (b_{22} - \lambda_1 g_{22})\xi_1^2\xi_2^2 = 0.$$

On adding these equations we get

$$b_{ij}\xi_1^i\xi_2^j - \lambda_1 g_{ij}\xi_1^i\xi_2^j = 0. \tag{13}$$

If now we take $i = 2$ in the equations (12), and multiply the first of them by ξ_1^1 and the second by ξ_1^2, and again add the resulting equations, we obtain

$$b_{ij}\xi_1^i\xi_2^j - \lambda_2 g_{ij}\xi_1^i\xi_2^j = 0. \tag{14}$$

Finally, subtracting (13) from (14), and using the hypothesis that $\lambda_1 \neq \lambda_2$, we obtain the desired conclusion. ☐

This lemma may be regarded as a variant of the theorem from linear algebra which states that a real quadratic form assumes diagonal form relative to a suitable orthogonal basis; hence a real quadratic form can be brought to diagonal form by a rotation of the given orthogonal co-ordinate system.

8.3. Properties of the Second Fundamental Form

We now return to the first and second fundamental forms of a surface $r = r(u, v)$ in Euclidean 3-space:

$$dl^2 = g_{ij}dx^i\,dx^j, \tag{15}$$

$$\langle \ddot{r}, m \rangle\, dt^2 = b_{ij}dx^i\,dx^j, \qquad x^1 = u, \quad x^2 = v. \tag{16}$$

Recall that the ratio of these quadratic forms in the components of the tangent vector to the section of the surface by a plane orthogonal to the surface at a point (i.e. a "normal section"), gives the curvature of the section (i.e. "normal curvature") at that point (Corollary 8.1.3).

Having at our disposal the concepts introduced in the preceding subsection, we can now formulate the general definition of the principal, Gaussian, and mean curvatures of a surface at a point.

8.3.1. Definition. The eigenvalues of the pair of quadratic forms (15) and (16) (see §8.2) are called the *principal curvatures* of the surface at the point under investigation. The product of the principal curvatures is called the *Gaussian curvature* K of the surface at the point, and their sum the *mean curvature*.†

8.3.2. Example. As at the beginning of §8.1 (see especially 8.1.1), we suppose we are given our surface in the form $z = f(x, y)$, and further that at the point (x_0, y_0, z_0) of it under study, $f_x = f_y = 0$ (i.e. that the z-axis is perpendicular to the tangent plane to the surface at (x_0, y_0, z_0)). Putting $x = u$, $y = v$, the first and second fundamental forms at the point (x_0, y_0, z_0) are given by

$$g_{11} = 1, \qquad g_{22} = 1, \qquad g_{12} = g_{21} = 0 \qquad (g_{ij} = \delta_{ij}), \tag{17}$$

$$\begin{aligned} L = b_{11} &= \langle r_{uu}, m\rangle = f_{xx}(x_0, y_0), \\ M = b_{12} &= \langle r_{uv}, m\rangle = f_{xy}(x_0, y_0), \\ N = b_{22} &= \langle r_{vv}, m\rangle = f_{yy}(x_0, y_0). \end{aligned} \tag{18}$$

(Here of course the unit normal m to the surface at (x_0, y_0, z_0) is in the direction of the positive z-axis.) Thus at the point under scrutiny, the second fundamental form of the surface is

$$b_{11}\,dx^2 + 2b_{12}\,dx\,dy + b_{22}\,dy^2 = f_{xx}\,dx^2 + 2f_{xy}\,dx\,dy + f_{yy}\,dy^2 = d^2 f,$$

so that since $g_{ij} = \delta_{ij}$, the principal curvatures are the eigenvalues λ_1, λ_2 of the matrix

$$\begin{pmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{pmatrix}, \tag{19}$$

i.e. the roots of the equation $(f_{xx} - \lambda)(f_{yy} - \lambda) - (f_{xy})^2 = 0$, and the Gaussian curvature is $f_{xx}f_{yy} - (f_{xy})^2$, the determinant of the Hessian (19) of the function f. The mean curvature is $f_{xx} + f_{yy} = \lambda_1 + \lambda_2$. Thus our general Definition 8.3.1 does agree with the earlier one (8.1.1) where the various curvatures were defined in terms of a co-ordinate system dependent (as in the present example) very much on the particular point under consideration.

---

† It is perhaps more usual to define the mean curvature as the arithmetic mean of the principal curvatures.

Continuing with our example, let us now consider the principal directions at our point, of the pair of quadratic forms given by (17) and (18). As defined in the preceding subsection (see (12)), these are (the directions of) the non-zero vectors in the (x, y)-plane $f_1 = (\xi_1^1, \xi_1^2)$ and $f_2 = (\xi_2^1, \xi_2^2)$, where ξ_i^1, ξ_i^2 are non-trivial solutions of the following system of linear equations ($i = 1, 2$):

$$(f_{xx} - \lambda_i \delta_{11})\xi_i^1 + (f_{xy} - \lambda_i \delta_{12})\xi_i^2 = 0,$$
$$(f_{xy} - \lambda_i \delta_{12})\xi_i^1 + (f_{yy} - \lambda_i \delta_{22})\xi_i^2 = 0.$$

If we assume that the eigenvalues λ_1, λ_2 are distinct, then by Lemma 8.2.1, we have $f_1 \perp f_2$ (in the usual Euclidean sense since $g_{ij} = \delta_{ij}$). Hence we can replace the co-ordinate system x, y by a new system x', y', with axes in the principal directions, obtained from the old system by means of a rotation of the (x, y)-plane through an angle φ. Then in terms of the new co-ordinates x', y', z, we have

$$z = f(x(x', y'), y(x', y')),$$

where $x = x' \cos\varphi + y' \sin\varphi$, $y = -x' \sin\varphi + y' \cos\varphi$. Relative to these new co-ordinates, the second fundamental form becomes (at the point (x_0, y_0, z_0))

$$\lambda_1 (dx')^2 + \lambda_2 (dy')^2.$$

By Corollary 8.1.3, the curvature at our point (x_0, y_0, z_0) of a normal section of the surface through that point (i.e. the "normal curvature" at the point) is therefore

$$k = \frac{\lambda_1 (\dot{x}')^2 + \lambda_2 (\dot{y}')^2}{(\dot{x}')^2 + (\dot{y}')^2}, \tag{20}$$

where we are taking the normal section to have parametric equations $x' = x'(t)$, $y' = y'(t)$, $z = z(t)$.

Clearly the tangent vector to this normal section at the point (x_0, y_0, z_0) is $e = (\dot{x}', \dot{y}', 0)$. Hence the angle α between this tangent vector and the x'-axis satisfies

$$\cos^2\alpha = \frac{(\dot{x}')^2}{(\dot{x}')^2 + (\dot{y}')^2}, \qquad \sin^2\alpha = \frac{(\dot{y}')^2}{(\dot{x}')^2 + (\dot{y}')^2}. \tag{21}$$

From (20) and (21) we deduce that at the point (x_0, y_0, z_0), at least for our special choice of co-ordinates,

$$k = \lambda_1 \cos^2\alpha + \lambda_2 \sin^2\alpha.$$

This formula, called "Euler's formula," holds in any system of co-ordinates. We shall now prove this. (Thus our example ends here, although the argument we have used actually comprises the bulk of the proof.)

8.3.3. Theorem. *The curvature of a normal section* (*i.e. the* "*normal curvature*") *through a point on a surface, is given by the formula*

$$k = \lambda_1 \cos^2\alpha + \lambda_2 \sin^2\alpha, \tag{22}$$

where λ_1, λ_2 *are the principal curvatures of the surface at the point, and* α *is the angle between the tangent vector to the normal section at the point, and the principal direction corresponding to* λ_1.

PROOF. In Example 8.3.2 above we deduced Euler's formula in the case where the surface is given in the form $z = f(x, y)$, and the z-axis is perpendicular to the surface at the point (x_0, y_0, z_0) being studied. However since nothing of this special co-ordinate system is involved in the formula, for any particular (non-singular) point of the surface we simply choose a co-ordinate system with these properties (as we can, locally at least, by §7.1), i.e. so that the z-axis is normal to the surface at the point, and the x- and y-axes mutually orthogonal, and parallel to the tangent plane through the point (we may even choose them parallel to the principal directions). Then in a neighbourhood of the point the surface can be given by a function $z = f(x, y)$, with, of course, $f_x = f_y = 0$ at the point. Moreover if the x-axis and y-axis are chosen in the principal directions then $f_{xy} = f_{yx} = 0$, and $\lambda_1 = f_{xx}, \lambda_2 = f_{yy}$ at the point. We may then argue as in the example, provided $\lambda_1 \neq \lambda_2$ (which is assumed implicitly in the theorem since otherwise the principal directions remain undefined). This completes the proof. □

Note that if we reverse the direction of the z-axis in the above, then the signs of the principal curvatures are also reversed.

8.3.4. Corollary. *If* λ_1 *and* λ_2 *are the principal curvatures of a surface at a point and* $\lambda_1 > \lambda_2$, *then* λ_1 *and* λ_2 *are respectively the greatest and smallest values of the normal curvature (i.e. curvature of a normal section) at the point. (Here the signs are taken into account—see Corollary* 8.1.3.)

We next derive useful formulae for the second fundamental form and the Gaussian curvature in the case that the surface is given in the form $z = f(x, y)$ (but where we are no longer assuming a particular direction for the z-axis at any point). In that case, with $x = u$, $y = v$, and $r(u, v) = (u, v, f(u, v))$ we have

$$r_u = (1, 0, f_x), \qquad r_v = (0, 1, f_y),$$

$$[r_u, r_v] = (-f_x, -f_y, 1),$$

$$r_{uu} = (0, 0, f_{xx}), \qquad r_{uv} = (0, 0, f_{xy}), \qquad r_{vv} = (0, 0, f_{yy}),$$

$$m = \frac{[r_u, r_v]}{|[r_u, r_v]|} = \frac{(-f_x, -f_y, 1)}{\sqrt{1 + f_x^2 + f_y^2}}, \tag{23}$$

$$L = b_{11} = \frac{f_{xx}}{\sqrt{1 + f_x^2 + f_y^2}}, \qquad M = b_{12} = b_{21} = \frac{f_{xy}}{\sqrt{1 + f_x^2 + f_y^2}},$$

$$N = b_{22} = \frac{f_{yy}}{\sqrt{1 + f_x^2 + f_y^2}}.$$

Hence

$$b_{ij}\,dx^i\,dx^j = \frac{1}{\sqrt{1+f_{x^1}^2+f_{x^2}^2}}\,f_{x^ix^j}\,dx^i\,dx^j), \tag{24}$$
$$x^1 = u = x, \qquad x^2 = v = y.$$

Recall also the formulae (20) of §7 for the g_{ij}:

$$\begin{aligned} g_{11} &= 1 + f_x^2, \qquad g_{12} = g_{21} = f_x f_y, \qquad g_{22} = 1 + f_y^2, \\ g &= g_{11}g_{22} - g_{12}^2 = 1 + f_x^2 + f_y^2. \end{aligned} \tag{25}$$

We shall use formulae (24) and (25) in the proof of the following

8.3.5. Theorem. *The Gaussian curvature of a surface is equal to the ratio of the determinants of the second and the first fundamental forms:*

$$K = \frac{b_{11}b_{22} - b_{12}^2}{g_{11}g_{22} - g_{12}^2}. \tag{26}$$

In particular, if the surface is given in the form $z = f(x, y)$, then we have

$$K = \frac{f_{xx}f_{yy} - f_{xy}^2}{(1 + f_x^2 + f_y^2)^2}. \tag{27}$$

Proof. By definition, the principal curvatures λ_1, λ_2 are the solutions of equation (10) above:

$$\det(Q - \lambda G) = 0,$$

where $Q = (b_{ij})$ is the matrix of the second fundamental form, and $G = (g_{ij})$. Since the first fundamental form is positive definite, its matrix G is nonsingular. Hence

$$\det(Q - \lambda G) = \det G \det(G^{-1}Q - \lambda \cdot 1),$$

from which it follows that λ_1 and λ_2 are just the eigenvalues of the matrix $G^{-1}Q$. Since the product of the eigenvalues of a matrix is the same as the determinant (a fact we have used before!), we deduce that

$$K = \lambda_1\lambda_2 = \det(G^{-1}Q) = \frac{\det Q}{\det G},$$

which proves (26).

To get formula (27), simply substitute in (26) the expressions for the b_{ij} and g_{ij} given in (24) and (25). □

8.3.6. Corollary. *If a surface is given in the form $z = f(x, y)$, then the sign of the Gaussian curvature K is the same as the sign of the determinant $f_{xx}f_{yy} - f_{xy}^2$ of the Hessian of $f(x, y)$.*

8.3.7. Example. Suppose that we are given a surface in the form $z = f(x, y)$, where the function $f(x, y)$ is a solution of Laplace's equation $f_{xx} + f_{yy} = 0$. Then for all points of the surface we shall have $f_{xx}f_{yy} - f_{xy}^2 \leqq 0$, since $f_{xx} = -f_{yy}$. Hence by Corollary 8.3.6, at all points of the surface where at least one of the partial derivatives f_{xx}, f_{xy} is non-zero, we shall have that the Gaussian curvature $K < 0$.

Finally we explain the geometric significance of the sign of the Gaussian curvature at a point. As usual we choose a co-ordinate system x, y, z, with mutually orthogonal co-ordinate axes, such that the z-axis is perpendicular to the surface at the point (x_0, y_0, z_0). Then locally (about the point) the surface is given by a function $z = f(x, y)$, where $f_x = f_y = 0$. We consider three cases:

(i) $K > 0$; $\lambda_1 > 0, \lambda_2 > 0$. Since the smallest value of the normal curvature is positive, all values of it are positive, i.e. at the point the surface bends upwards in all directions, so the function $f(x, y)$ has a local minimum at the point (x_0, y_0) (see Figure 10(a)).

(ii) $K > 0$; $\lambda_1 < 0, \lambda_2 < 0$. The function $f(x, y)$ has a local maximum at (x_0, y_0) (see Figure 10(b)).

(iii) $K < 0$. In this case we must have $\lambda_1 > 0, \lambda_2 < 0$, or $\lambda_1 < 0, \lambda_2 > 0$. The function has a "saddle" or "pass" at the point (Figure 10(c)).

We conclude that: *If $K > 0$ there is a neighbourhood of the point throughout which the surface lies on one side of the tangent plane at the point. If $K < 0$ then the surface intersects the tangent plane at the point arbitrarily close to the point.*

If the Gaussian curvature is positive at every point of a surface, then we say that the surface is *strictly convex*.

Note that in the above, by using Corollary 8.3.6, and the fact that λ_1, λ_2 are the eigenvalues of the Hessian of $f(x, y)$ (since $g_{ij} = \delta_{ij}$ at the point), it is easy to see that cases (i) and (ii) are respectively equivalent to: $K > 0$; $f_{xx}, f_{yy} > 0$, and $K > 0$; $f_{xx}, f_{yy} < 0$. In this form, (i), (ii) and (iii) comprise the familiar "second-partials test" for determining the nature of a surface in a neighbourhood of a point at which the tangent plane is "horizontal."

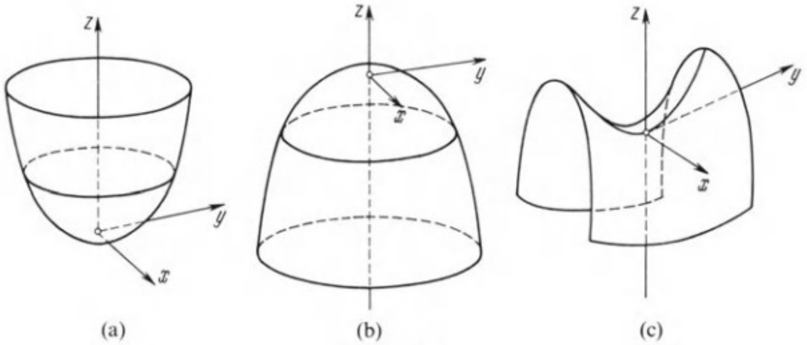

(a) (b) (c)

Figure 10

(/

is called the Gaussian curvature K of the surface at the point, and their sum the mean curvature.

Pag. 84 (99)-

8.3.5. Theorem

The Gaussian curvature of a surface is equal to the ratio of the determinants of the second and the first fundamental forms:

$$K = \frac{b_{11}\, b_{22} - b_{12}^2}{g_{11}\, g_{22} - g_{12}^2} . \;\ldots$$

Pág. 85 (100) 8.3.7. Example

Gaussian curvature $K < 0$

Finally we explain the geometric significance of the sign of the Gaussian curvature at a point.

Pág. 95 (110) the Language of Complex Numbers in Geometry

Then by partial analogy with Theorem 8.3.5, we define Gaussian curvature K of a space-like surface in R_1^3 by

('// %$ 9 9 7! ! A! ! ! >! !
" ! "!" "! ! (/

Hence

$$b_{ij}\,dx^i\,dx^j = \frac{1}{\sqrt{1 + f_{x^1}^2 + f_{x^2}^2}}\, f_{x^ix^j}\,dx^i\,dx^j), \tag{24}$$

$$x^1 = u = x, \qquad x^2 = v = y.$$

Recall also the formulae (20) of §7 for the g_{ij}:

$$\begin{aligned} g_{11} &= 1 + f_x^2, \qquad g_{12} = g_{21} = f_x f_y, \qquad g_{22} = 1 + f_y^2, \\ g &= g_{11}g_{22} - g_{12}^2 = 1 + f_x^2 + f_y^2. \end{aligned} \tag{25}$$

We shall use formulae (24) and (25) in the proof of the following

8.3.5. Theorem. *The Gaussian curvature of a surface is equal to the ratio of the determinants of the second and the first fundamental forms*:

$$K = \frac{b_{11}b_{22} - b_{12}^2}{g_{11}g_{22} - g_{12}^2}. \tag{26}$$

In particular, if the surface is given in the form $z = f(x, y)$, *then we have*

$$K = \frac{f_{xx}f_{yy} - f_{xy}^2}{(1 + f_x^2 + f_y^2)^2}. \tag{27}$$

PROOF. By definition, the principal curvatures λ_1, λ_2 are the solutions of equation (10) above:

$$\det(Q - \lambda G) = 0,$$

where $Q = (b_{ij})$ is the matrix of the second fundamental form, and $G = (g_{ij})$. Since the first fundamental form is positive definite, its matrix G is nonsingular. Hence

$$\det(Q - \lambda G) = \det G \det(G^{-1}Q - \lambda \cdot 1),$$

from which it follows that λ_1 and λ_2 are just the eigenvalues of the matrix $G^{-1}Q$. Since the product of the eigenvalues of a matrix is the same as the determinant (a fact we have used before!), we deduce that

$$K = \lambda_1\lambda_2 = \det(G^{-1}Q) = \frac{\det Q}{\det G},$$

which proves (26).

To get formula (27), simply substitute in (26) the expressions for the b_{ij} and g_{ij} given in (24) and (25). ☐

8.3.6. Corollary. *If a surface is given in the form* $z = f(x, y)$, *then the sign of the Gaussian curvature* K *is the same as the sign of the determinant* $f_{xx}f_{yy} - f_{xy}^2$ *of the Hessian of* $f(x, y)$.

8.3.7. Example. Suppose that we are given a surface in the form $z = f(x, y)$, where the function $f(x, y)$ is a solution of Laplace's equation $f_{xx} + f_{yy} = 0$. Then for all points of the surface we shall have $f_{xx}f_{yy} - f_{xy}^2 \leqq 0$, since $f_{xx} = -f_{yy}$. Hence by Corollary 8.3.6, at all points of the surface where at least one of the partial derivatives f_{xx}, f_{xy} is non-zero, we shall have that the Gaussian curvature $K < 0$.

Finally we explain the geometric significance of the sign of the Gaussian curvature at a point. As usual we choose a co-ordinate system x, y, z, with mutually orthogonal co-ordinate axes, such that the z-axis is perpendicular to the surface at the point (x_0, y_0, z_0). Then locally (about the point) the surface is given by a function $z = f(x, y)$, where $f_x = f_y = 0$. We consider three cases:

(i) $K > 0$; $\lambda_1 > 0$, $\lambda_2 > 0$. Since the smallest value of the normal curvature is positive, all values of it are positive, i.e. at the point the surface bends upwards in all directions, so the function $f(x, y)$ has a local minimum at the point (x_0, y_0) (see Figure 10(a)).

(ii) $K > 0$; $\lambda_1 < 0$, $\lambda_2 < 0$. The function $f(x, y)$ has a local maximum at (x_0, y_0) (see Figure 10(b)).

(iii) $K < 0$. In this case we must have $\lambda_1 > 0$, $\lambda_2 < 0$, or $\lambda_1 < 0$, $\lambda_2 > 0$. The function has a "saddle" or "pass" at the point (Figure 10(c)).

We conclude that: *If $K > 0$ there is a neighbourhood of the point throughout which the surface lies on one side of the tangent plane at the point. If $K < 0$ then the surface intersects the tangent plane at the point arbitrarily close to the point.*

If the Gaussian curvature is positive at every point of a surface, then we say that the surface is *strictly convex*.

Note that in the above, by using Corollary 8.3.6, and the fact that λ_1, λ_2 are the eigenvalues of the Hessian of $f(x, y)$ (since $g_{ij} = \delta_{ij}$ at the point), it is easy to see that cases (i) and (ii) are respectively equivalent to: $K > 0$; $f_{xx}, f_{yy} > 0$, and $K > 0$; $f_{xx}, f_{yy} < 0$. In this form, (i), (ii) and (iii) comprise the familiar "second-partials test" for determining the nature of a surface in a neighbourhood of a point at which the tangent plane is "horizontal."

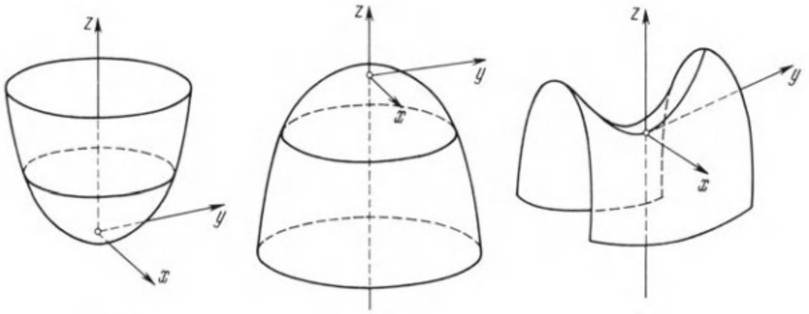

(a) (b) (c)

Figure 10

where m is the unit normal to the surface (at the point under study). Then by partial analogy with Theorem 8.3.5, we define the *Gaussian curvature K of a space-like surface* in $\mathbb{R}_1^3$ by

$$K = -\frac{\det(b_{ij})}{\det(g_{ij})}. \tag{14}$$

(The minus sign is explained below.) It follows that if the surface is given in the form $t = f(x, y)$, with $1 - f_x^2 - f_y^2 > 0$ (see above), then (cf. §8.3 (27))

$$K = \frac{f_{xy}^2 - f_{xx}f_{yy}}{(1 - f_x^2 - f_y^2)^2}. \tag{15}$$

In particular for the hyperboloid $t^2 - x^2 - y^2 = 1$, we obtain $K \equiv -1$.

Remarks. 1. In Chapter 4 we shall compute the curvature of the Lobachevskian plane in terms of its intrinsic geometry. We shall again arrive at $K \equiv -1$, vindicating the choice of sign in the definition (14).

2. By analogy with the definition given in §8.3, we shall say that a space-like surface in $\mathbb{R}_1^3$ is *convex* if $K < 0$ at all points of the surface.

§11. The Language of Complex Numbers in Geometry

11.1. Complex and Real Co-ordinates

Many geometrical problems are most conveniently formulated and solved in terms of the language of complex numbers. We therefore expound here the simple facts about complex numbers that we shall be needing.

We denote by $\mathbb{C}^n$ the vector space of dimension n over the field of complex numbers $\mathbb{C}$. (A vector space is determined up to isomorphism by its dimension.) If $e_1, \ldots, e_n$ form a basis for this vector space, then every vector $\xi \in \mathbb{C}^n$ has the unique form

$$\xi = z^k e_k, \qquad z^k = x^k + iy^k, \tag{1}$$

where the z^k are the (complex) components.

We can regard $\mathbb{C}^n$ also as a $2n$-dimensional vector space $\mathbb{R}^{2n}$ over the reals; as a basis for this space we might then take

$$e_1, \ldots, e_n, ie_1, \ldots, ie_n. \tag{2}$$

Since

$$\xi = z^k e_k = x^k e_k + y^k(ie_k), \tag{3}$$

the real components of the vector ξ relative to the basis (2), are the numbers $x^1, \ldots, x^n, y^1, \ldots, y^n$. This process of going from a complex vector space to a real one is called *realization.*

The non-singular (i.e. invertible) linear transformations of the vector space $\mathbb{C}^n$ form a group, called the *general linear group of degree n over* $\mathbb{C}$, and denoted by $GL(n, \mathbb{C})$. This group is of course isomorphic to the group of $n \times n$ non-singular matrices with complex entries (so that, as usual, we shall often think of $GL(n, \mathbb{C})$ as consisting of matrices). Upon realization of $\mathbb{C}^n$, each of these transformations becomes a linear transformation of the real vector space $\mathbb{R}^{2n}$. This defines a monomorphism (i.e. one-to-one homomorphism)

$$r\colon GL(n, \mathbb{C}) \to GL(2n, \mathbb{R}), \tag{4}$$

called the *realization map.*

11.1.1. Example. In the case $n = 1$, our space is just the one-dimensional complex vector space $\mathbb{C}$ co-ordinatized by $z = x + iy$. The invertible linear transformations are just multiplications by a complex number $\lambda \neq 0$ (one such λ for each linear transformation):

$$z \mapsto \lambda z. \tag{5}$$

Writing $\lambda = a + ib$, $a^2 + b^2 \neq 0$, this takes the form

$$z = x + iy \mapsto (a + ib)(x + iy) = (ax - by) + i(bx + ay).$$

Hence the corresponding linear transformation of $\mathbb{R}^2$ has matrix

$$\begin{pmatrix} a & b \\ -b & a \end{pmatrix} = r(\lambda), \qquad a^2 + b^2 \neq 0. \tag{6}$$

Thus we see that certainly not all (invertible) linear transformation of $\mathbb{R}^2$ are obtained in this way from the complex linear transformations of $\mathbb{C}$; in other words, r is not onto.

This example generalizes to the n-dimensional case: It is not difficult to show that if Λ is a matrix in $GL(n, \mathbb{C})$, and $\Lambda = A + iB$ where A and B are real matrices, then

$$r(\Lambda) = \begin{pmatrix} A & B \\ -B & A \end{pmatrix}. \tag{7}$$

11.1.2 Exercise. Prove that $\det(r(\Lambda)) = |\det \Lambda|^2$.

Here is a quick proof of (7) which incidentally yields an alternative characterization of the matrices of the form $r(\Lambda)$, i.e. of the matrices occurring as images under the map $r\colon GL(n, \mathbb{C}) \to GL(2n, \mathbb{R})$. If Λ is any element of $GL(n, \mathbb{C})$, and ξ is any vector in $\mathbb{C}^n$, then

$$\Lambda(i\xi) = i\Lambda(\xi). \tag{8}$$

Now the linear transformation of $\mathbb{C}^n$ which simply multiplies each vector by the scalar i, goes under r to a linear transformation I of $\mathbb{R}^{2n}$, given by

$$r(i) = I = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad \text{where } 1 = n \times n \text{ identity matrix.} \tag{9}$$

(To see this consider the action of i on the basis vectors e_k, ie_k.) It follows from (8) that the matrix I commutes with the matrix $r(\Lambda)$ (since ξ was arbitrary). But then it is straightforward to check that matrices commuting with I must have the form $\begin{pmatrix} A & B \\ -B & A \end{pmatrix}$. We leave to the reader the verification that a matrix of this form does indeed arise from a "complex operator" on $\mathbb{C}^n$.

We conclude by mentioning an important subgroup of $GL(n, \mathbb{C})$, namely that consisting of all matrices of determinant 1, called the *special linear group of degree n over* $\mathbb{C}$, and denoted by $SL(n, \mathbb{C})$.

11.2. The Hermitian Scalar Product

We define a scalar product on the space $\mathbb{C}^n$ as follows:

$$\langle \xi_1, \xi_2 \rangle_{\mathbb{C}} = \sum_{k=1}^{n} z_1^k \bar{z}_2^k, \tag{10}$$

where $\xi_1 = z_1^i e_i$, $\xi_2 = z_2^i e_i$, and the bar indicates the complex conjugate. It follows that if $\xi = z^i e_i$, then

$$\langle \xi, \xi \rangle_{\mathbb{C}} = \sum_{k=1}^{n} |z^k|^2. \tag{11}$$

This scalar product has the following main properties:

$$\begin{gathered} \langle \lambda\xi, \eta \rangle_{\mathbb{C}} = \lambda \langle \xi, \eta \rangle_{\mathbb{C}}, \\ \langle \xi, \lambda\eta \rangle_{\mathbb{C}} = \bar{\lambda} \langle \xi, \eta \rangle_{\mathbb{C}}, \\ \langle \xi, \eta \rangle_{\mathbb{C}} = \overline{\langle \eta, \xi \rangle}_{\mathbb{C}}, \\ \langle \xi_1 + \xi_2, \eta \rangle_{\mathbb{C}} = \langle \xi_1, \eta \rangle_{\mathbb{C}} + \langle \xi_2, \eta \rangle_{\mathbb{C}}, \\ \langle \xi, \xi \rangle_{\mathbb{C}} > 0 \quad \text{if } \xi \neq 0. \end{gathered} \tag{12}$$

Any scalar product on $\mathbb{C}^n$ with the properties (12) is called an *Hermitian scalar product*, or *Hermitian form.*

If on the realized space $\mathbb{R}^{2n}$ we introduce the usual Euclidean scalar product, defined for a pair of vectors

$$\xi_1 = (x_1^1, \ldots, x_1^n, y_1^1, \ldots, y_1^n), \ \xi_2 = (x_2^1, \ldots, x_2^n, y_2^1, \ldots, y_2^n)$$

by

$$\langle \xi_1, \xi_2 \rangle_{\mathbb{R}} = \sum_{k=1}^{n} (x_1^k x_2^k + y_1^k y_2^k), \tag{13}$$

Let $z = z(w)$ define a complex analytic co-ordinate change, so that $\partial z/\partial \bar{w} = 0$. Then as in the derivation of (33) we have

$$dz = \left(\frac{dz}{dw}\right) dw, \qquad d\bar{z} = \overline{\left(\frac{dz}{dw}\right)}\, d\bar{w},$$

whence

$$dl^2 = h(z, \bar{z})\, dz\, d\bar{z} = h(z(w, \bar{w}), \overline{z(w, \bar{w})}) \left|\frac{dz}{dw}\right|^2 dw\, d\bar{w},$$

which still has conformal form. That complex conjugation preserves the conformal form, is easily verified.

For the converse, suppose that we have a co-ordinate change $z = z(w) = z(w, \bar{w})$, which is neither analytic (so that $\partial z/\partial \bar{w} \not\equiv 0$), nor the composite of an analytic function and complex conjugation (so that $\partial z/\partial w \not\equiv 0$). It follows from our underlying assumption of continuity of the partial derivatives of $z(u, v)$ (where $w = u + iv$), that there exist points at which $\partial z/\partial \bar{w} \neq 0 \neq \partial z/\partial w$ (since otherwise from (6) we should have, at every point, $\partial z/\partial u = \pm i(\partial z/\partial v)$ with neither equality holding everywhere). Now by Lemma 12.1.1, $dz = (\partial z/\partial w)\, dw + (\partial z/\partial \bar{w})\, d\bar{w}$; hence

$$dl^2 = h\left|\frac{\partial z}{\partial w} dw + \frac{\partial z}{\partial \bar{w}} d\bar{w}\right|^2 = h(a_{11}\, du^2 + 2a_{12}\, du\, dv + a_{22}\, dv^2),$$

where $w = u + iv$. Calculation of the a_{ij} shows that the equations $a_{11} = a_{22}$, $a_{12} = 0$ imply that either $\partial z/\partial \bar{w} = 0$ or $\partial z/\partial w = 0$. Since there are points at which neither of these hold, we deduce that our transformation does not preserve the conformal form, completing the proof. ☐

§13. The Conformal Form of the Metric on a Surface

13.1. Isothermal Co-ordinates. Gaussian Curvature in Terms of Conformal Co-ordinates

Suppose that we have in Euclidean space $\mathbb{R}^3$, a 2-dimensional surface with parametric equations

$$x = x(p, q), \qquad y = y(p, q), \qquad z = z(p, q), \tag{1}$$

where (p, q) ranges over some region of $\mathbb{R}^2$. Then, as we saw in §7.3, the Euclidean metric in $\mathbb{R}^3$ induces on this surface a Riemannian metric

$$dl^2 = E(dp)^2 + 2F\, dp\, dq + G(dq)^2, \tag{2}$$

where $g = EG - F^2 > 0$ (since the form given by (2) is positive definite). It turns out that by means of a change of the local co-ordinates p, q, this metric can be brought into conformal form. (We have already seen this in the case of the sphere (§9).) We state this result precisely as

13.1.1. Theorem. *With the above notation, suppose that E, F, G are real-valued analytic functions of the real variables p, q (i.e. are representable as power series in p, q). Then there exist new (real) local co-ordinates u, v for the surface in terms of which the induced metric* (2) *takes the form*

$$dl^2 = f(u, v)(du^2 + dv^2). \tag{3}$$

Co-ordinates with this property are called *isothermal* or *conformal* (cf. Definition 12.3.3). Thus for local co-ordinates on a surface to be isothermal the induced metric must have conformal form (i.e. the form (3)) in those co-ordinates.

Proof of Theorem 13.1.1. Factorizing the right-hand side of (2), we obtain

$$dl^2 = \left(\sqrt{E}\,dp + \frac{F + i\sqrt{g}}{\sqrt{E}}\,dq\right)\left(\sqrt{E}\,dp + \frac{F - i\sqrt{g}}{\sqrt{E}}\,dq\right).$$

We are seeking new co-ordinates u, v as functions $u = u(p, q)$, $v = v(p, q)$ of p and q, with the property that in terms of u, v the metric (2) takes the form (3). We shall certainly have achieved this if we can find an "integrating factor," i.e. a complex-valued function $\lambda = \lambda(p, q)$ such that

$$\begin{aligned} \lambda\left(\sqrt{E}\,dp + \frac{F + i\sqrt{g}}{\sqrt{E}}\,dq\right) &= du + i\,dv, \\ \bar{\lambda}\left(\sqrt{E}\,dp + \frac{F - i\sqrt{g}}{\sqrt{E}}\,dq\right) &= du - i\,dv, \end{aligned} \tag{4}$$

since on multiplying these two equations together, we obtain

$$|\lambda|^2\,dl^2 = du^2 + dv^2, \qquad dl^2 = |\lambda|^{-2}(du^2 + dv^2), \tag{5}$$

so that as $f(u, v)$ the function $|\lambda|^{-2}$ will serve. Since the second of the equations (4) is just the complex conjugate of the first, our problem has now become that of finding functions $u(p, q)$, $v(p, q)$, $\lambda(p, q)$ satisfying the equation

$$\lambda\left(\sqrt{E}\,dp + \frac{F + i\sqrt{g}}{\sqrt{E}}\,dq\right) = du + i\,dv = \left(\frac{\partial u}{\partial p} + i\frac{\partial v}{\partial p}\right)dp + \left(\frac{\partial u}{\partial q} + i\frac{\partial v}{\partial q}\right)dq, \tag{6}$$

or, equivalently,

$$\lambda\sqrt{E} = \frac{\partial u}{\partial p} + i\frac{\partial v}{\partial p}, \qquad \lambda\frac{F + i\sqrt{g}}{\sqrt{E}} = \frac{\partial u}{\partial q} + i\frac{\partial v}{\partial q}. \tag{7}$$

Eliminating λ, we get

$$(F + i\sqrt{g})\left(\frac{\partial u}{\partial p} + i\frac{\partial v}{\partial p}\right) = E\left(\frac{\partial u}{\partial q} + i\frac{\partial v}{\partial q}\right), \tag{8}$$

which is equivalent to

$$F\frac{\partial u}{\partial p} - \sqrt{g}\frac{\partial v}{\partial p} = E\frac{\partial u}{\partial q}, \qquad \sqrt{g}\frac{\partial u}{\partial p} + E\frac{\partial v}{\partial p} = E\frac{\partial v}{\partial q}.$$

Hence

$$\begin{aligned} \frac{\partial v}{\partial p} &= \frac{F\dfrac{\partial u}{\partial p} - E\dfrac{\partial u}{\partial q}}{\sqrt{g}}, \qquad & \frac{\partial v}{\partial q} &= \frac{G\dfrac{\partial u}{\partial p} - F\dfrac{\partial u}{\partial q}}{\sqrt{g}}, \\ \frac{\partial u}{\partial p} &= \frac{E\dfrac{\partial v}{\partial q} - F\dfrac{\partial v}{\partial p}}{\sqrt{g}}, \qquad & \frac{\partial u}{\partial q} &= \frac{F\dfrac{\partial v}{\partial q} - G\dfrac{\partial v}{\partial p}}{\sqrt{g}}, \end{aligned} \tag{9}$$

where the two equations on the right-hand side of (9) follow on multiplying equation (8) by $(F - i\sqrt{g})/E$. If we now also demand of the functions u, v that their second partial derivatives be continuous, we shall have $\partial^2 u/\partial p\,\partial q = \partial^2 u/\partial q \partial p$ and $\partial^2 v/\partial p \partial q = \partial^2 v/\partial q \partial p$, and then from (9) we get $Lu = 0$, $Lv = 0$ where the differential operator L is given by

$$L = \frac{\partial}{\partial q}\left[\frac{F\dfrac{\partial}{\partial p} - E\dfrac{\partial}{\partial q}}{\sqrt{EG - F^2}}\right] + \frac{\partial}{\partial p}\left[\frac{F\dfrac{\partial}{\partial q} - G\dfrac{\partial}{\partial p}}{\sqrt{EG - F^2}}\right]. \tag{10}$$

The partial differential equation $Lf = 0$ is known as "Beltrami's equation," and the operator Las the "Beltrami operator." Thus the upshot is that u, v are solutions of Beltrami's equation. Now it is known from the theory of partial differential equations (but we shall not prove it here) that if the functions E, F, G are real analytic, then Beltrami's equation always has solutions of the type we seek, namely functions u, v such that in the region over which (p, q) ranges, the map $(p, q) \mapsto (u(p, q), v(p, q))$ is one-to-one. (Note that once u, say is chosen (as a solution of Betrami's equation) then the equations in the first row of (9) determine grad u, and therefore (essentially) u. The function λ is then determined using (7).) With this we conclude the proof. ☐

Hence (assuming E, F, G real analytic) isothermal co-ordinates always exist on a surface in Euclidean 3-space. Lemma 12.3.4 then tells us exactly which local co-ordinate changes lead to other isothermal co-ordinates p, q on the surface. For if we write $z = p + iq$, $w = u + iv$, then regarding w as usual as a function of $z, \bar{z}$, i.e. $w = w(z, \bar{z})$, we know from 12.3.4 that p, q are isothermal co-ordinates on the surface if and only if $w(z, \bar{z})$ is analytic in either z or $\bar{z}$.

13.1.2. Example. Consider the complex plane (or line) $\mathbb{C}$ with complex co-ordinate $z = x + iy$. The Euclidean metric $dl^2 = dx^2 + dy^2 = dz\, d\bar{z}$ certainly has conformal form. Since linear-fractional transformations are analytic, we should expect (in view of 12.3.4) that they will preserve the conformal form of the Euclidean metric; and indeed if

$$w = \frac{az + b}{cz + d}, \qquad ad - bc = 1,$$

is any linear-fractional transformation, then $dw = [-1/(cz + d)^2]\, dz$, so that

$$dl^2 = dz\, d\bar{z} = |cz + d|^4\, dw\, d\bar{w},$$

as expected.

We next compute the Gaussian curvature of a surface in Euclidean 3-space, in terms of conformal co-ordinates on the surface.

13.1.3. Theorem. *If u, v are conformal co-ordinates on a surface in Euclidean 3-space, in terms of which the induced metric has the form*

$$dl^2 = g(u, v)(du^2 + dv^2),$$

then the Gaussian curvature of the surface is given by

$$K = -\frac{1}{2g(u, v)} \Delta \ln g(u, v), \tag{11}$$

where $\Delta = \partial^2/\partial u^2 + \partial^2/\partial v^2$ is the Laplace operator.

PROOF. Suppose that in terms of the conformal co-ordinates u, v, the surface is given (locally) by $r = r(u, v)$; $r = (x, y, z)$ (where x, y, z are Euclidean co-ordinates for $\mathbb{R}^3$). Since the metric on the surface in terms of these co-ordinates is given by $dl^2 = g(u, v)(du^2 + dv^2)$, we have (by §7.3 (19)) that

$$\langle r_u, r_u \rangle = \langle r_v, r_v \rangle = g(u, v), \qquad \langle r_u, r_v \rangle = 0. \tag{12}$$

On differentiating these equations with respect to u and v (and, as usual, assuming enough continuity and differentiability), we obtain

$$\begin{aligned} \frac{1}{2}\frac{\partial g(u, v)}{\partial u} &= \langle r_{uu}, r_u \rangle = \langle r_{uv}, r_v \rangle, \\ \frac{1}{2}\frac{\partial g(u, v)}{\partial v} &= \langle r_{vv}, r_v \rangle = \langle r_{uv}, r_u \rangle, \\ \langle r_{uu}, r_v \rangle + \langle r_u, r_{uv} \rangle &= 0 = \langle r_{uv}, r_v \rangle + \langle r_u, r_{vv} \rangle. \end{aligned} \tag{13}$$

Define unit vectors e_1, e_2, n by

$$e_1 = \frac{r_u}{\sqrt{g(u, v)}}, \qquad e_2 = \frac{r_v}{\sqrt{g(u, v)}}, \qquad n = [e_1, e_2]. \tag{14}$$

By (12) and the properties of the vector product the frame e_1, e_2, n is orthonormal at each point of the surface; in addition the vectors e_1, e_2 are tangent to the surface, and the vector n normal to it. By definition (§8.1(3)), the coefficients of the second fundamental form of the surface are

$$\begin{aligned} b_{11} &= L = \langle r_{uu}, n\rangle, \qquad b_{12} = M = \langle r_{uv}, n\rangle, \\ b_{22} &= \langle r_{vv}, n\rangle = N. \end{aligned} \tag{15}$$

It follows from (13) and (15) that relative to the basis e_1, e_2, n, the components of the vectors r_{uu}, r_{uv}, r_{vv} are as follows:

$$\begin{aligned} r_{uu} &= \left(\frac{1}{2\sqrt{g}}\frac{\partial g}{\partial u}, -\frac{1}{2\sqrt{g}}\frac{\partial g}{\partial v}, L\right), \\ r_{uv} &= \left(\frac{1}{2\sqrt{g}}\frac{\partial g}{\partial v}, \frac{1}{2\sqrt{g}}\frac{\partial g}{\partial u}, M\right), \\ r_{vv} &= \left(-\frac{1}{2\sqrt{g}}\frac{\partial g}{\partial u}, \frac{1}{2\sqrt{g}}\frac{\partial g}{\partial v}, N\right), \end{aligned} \tag{16}$$

where of course $g = g(u, v)$. Hence

$$\langle r_{uu}, r_{vv}\rangle - \langle r_{uv}, r_{uv}\rangle = LN - M^2 - \frac{1}{2g}\left[\left(\frac{\partial g}{\partial v}\right)^2 + \left(\frac{\partial g}{\partial u}\right)^2\right].$$

From this and (13) we obtain

$$\begin{aligned} \frac{1}{2}\frac{\partial^2 g}{\partial u^2} &= \langle r_{uuv}, r_v\rangle + \langle r_{uv}, r_{uv}\rangle \\ &= \frac{\partial}{\partial v}\langle r_{uu}, r_v\rangle - \langle r_{uu}, r_{vv}\rangle + \langle r_{uv}, r_{uv}\rangle \\ &= -\frac{1}{2}\frac{\partial^2 g}{\partial v^2} - (LN - M^2) + \frac{1}{2g}\left[\left(\frac{\partial g}{\partial u}\right)^2 + \left(\frac{\partial g}{\partial v}\right)^2\right], \end{aligned}$$

whence it follows (using §8(26) and a little manipulation) that the Gaussian curvature is given by

$$K = \frac{\det(b_{ij})}{\det(g_{ij})} = \frac{LN - M^2}{(g(u, v))^2} = -\frac{1}{2g(u, v)}\Delta \ln g(u, v),$$

concluding the proof of the theorem. □

Note finally that in complex notation the metric $dl^2 = g(u, v)(du^2 + dv^2)$ becomes (setting $z = u + iv$)

$$dl^2 = g(z, \bar{z})\, dz\, d\bar{z},$$

and the formula (11) for the Gaussian curvature of the surface with this metric becomes (again after a little calculation)

$$K = -\frac{2}{g}\frac{\partial^2}{\partial z\,\partial\bar{z}}(\ln g). \tag{17}$$

13.2. Conformal Form of the Metrics on the Sphere and the Lobachevskian Plane

We have already found (in §9) conformal co-ordinates x, y for the sphere: they were obtained by means of the stereographic projection of the sphere onto the plane with Euclidean co-ordinates x, y. Thus for the sphere of radius $R = 1$, the induced metric was found to have the form (see §9(4))

$$dl^2 = 4\frac{dx^2 + dy^2}{(1 + x^2 + y^2)^2}. \tag{18}$$

Setting $z = x + iy$, this takes the form

$$dl^2 = \frac{4}{(1 + |z|^2)^2}\,dz\,d\bar{z}, \tag{19}$$

where z co-ordinatizes the complex plane $\mathbb{C}$.

In §10 we derived (again by means of a stereographic projection) the following formula for the metric of the Lobachevskian plane (with $R = 1$; see §10(10)):

$$dl^2 = \frac{4}{(1 - (x^2 + y^2))^2}(dx^2 + dy^2), \qquad x^2 + y^2 < 1. \tag{20}$$

In complex notation this formula becomes

$$dl^2 = \frac{4}{(1 - |z|^2)^2}\,dz\,d\bar{z}, \qquad |z| < 1, \tag{21}$$

where the complex co-ordinate z ranges over the open unit disc. Recall that the open unit disc with metric (21) (or (20)) is known as the "Poincaré model" of the hyperbolic (i.e. Lobachevskian) plane. If we map the unit disc onto the the upper half-plane Im $w > 0$ by means of the map $z = (1 + iw)/(1 - iw)$, then we obtain the metric in "Klein's model" of the hyperbolic plane (see §10(13)):

$$dl^2 = -\frac{4}{(w - \bar{w})^2}\,dw\,d\bar{w}, \qquad \operatorname{Im} w > 0. \tag{22}$$

We shall now determine in terms of complex co-ordinates, the isometry groups of the sphere and the Lobachevskian plane. As we saw in the preceding section (Lemma 12.3.4), the conformal form of a metric on a surface is complex number k say. Then from $ad - bc = 1$ it follows that we may take $k = 1$ or i. However if $k = i$, then it is easy to see that, for example, $(ai + b)/(ci + d)$ is in the lower half-plane. The upshot is that the transformation (23) maps the upper half-plane onto itself precisely if a, b, c, d are all real.

It turns out that all transformations (23) with a, b, c, d real are in fact isometries of the Klein model of the Lobachevskian plane, i.e. they all preserve the metric

$$dl^2 = \frac{4\,dz\,d\bar{z}}{-(z - \bar{z})^2}.$$

We shall omit the verification of this, which is similar to that of the preceding two cases. It is noteworthy that in both models of the Lobachevskian plane, all linear-fractional transformations preserving the region of the complex plane on which the metric is defined, turn out to be isometries.

We gather the above results together in the following

13.2.3. Theorem. *The group of direct isometries of the Lobachevskian plane is isomorphic to:*

(i) $SU(1, 1)/\{\pm 1\}$ *(using the Poincaré model)*;
(ii) $SL(2, \mathbb{R})/\{\pm 1\}$ *(using Klein's model)*;
(iii) *the connected component of the identity of* $SO(1, 2)$ *(see* §6.2).

Again we are assuming the result (to be established in Chapter 4 of Part II) that these groups do indeed account for all direct isometries of the Lobachevskian plane. The statement (iii) follows in part from the fact that the connected component of the identity of $SO(1, 2)$, consists of just those isometries of $\mathbb{R}^3_1$ preserving the upper sheet of the pseudo-sphere (see §10.1).

Note also that, analogously to the case of the sphere, the full isometry group of the Lobachevskian plane is generated by all the direct isometries of the Poincaré model (i.e. by $SU(1, 1)/\{\pm 1\}$) together with the map $z \mapsto \bar{z}$, which sends the unit disc onto itself, and is obviously an isometry of that model. In terms of Klein's model, the full isometry group is obtained by including (as a generator) the map $z \mapsto -\bar{z}$, along with the transformations (23) with a, b, c, d real (forming a group isomorphic to $SL(2, \mathbb{R})/\{\pm 1\}$).

13.2.4. Corollary. *The groups* $SU(1, 1)/\{\pm 1\}$, $SL(2, \mathbb{R})/\{\pm 1\}$, *and the connected component of the identity of* $SO(1, 2)$, *are isomorphic.*

As an exercise, prove this directly by using appropriate co-ordinate changes!

13.3. Surfaces of Constant Curvature

Before broaching the theme of this section, we need to answer the following natural question: Are there surfaces in Euclidean 3-space whose induced metrics are hyperbolic (i.e. Lobachevskian)? It turns out that the answer is affirmative: there are surfaces in Euclidean 3-space which are, as they say "locally isometric" to the hyperbolic plane. (However there is no such surface that is globally isometric to the hyperbolic plane; i.e. from which there is a one-to-one metric-preserving map onto, say, the Poincaré model of the hyperbolic plane.) An example is provided by the following hornshaped surface of revolution (to which the name "pseudo-sphere" was first applied):

$$x = \operatorname{sech} u \cos v, \qquad y = \operatorname{sech} u \sin v, \qquad z = u - \tanh u, \qquad u \geqq 0,$$

where x, y, z are Euclidean co-ordinates for $\mathbb{R}^3$. The induced metric is easily calculated to be given by

$$dl^2 = (\tanh^2 u)\, du^2 + (\operatorname{sech}^2 u)\, dv^2,$$

which by a change of local co-ordinates can be brought into one of the usual forms for the Lobachevskian metric (we omit the verification of this). It is also easy to calculate that the Gaussian curvature $K \equiv -1$ (cf. the first of the remarks at the end of §10).

To return to our theme, suppose we are given a surface in Euclidean 3-space whose metric (in terms of certain local co-ordinates) has complex conformal form $dl^2 = g(z, \bar{z})\, dz\, d\bar{z}$. (By 13.1.1 the metrics of many surfaces in Euclidean 3-space can (locally) be brought into this form. By the preceding remarks there are such surfaces locally isometric to the Lobachevskian plane.) Since $g(z, \bar{z}) > 0$, we can define a new function $\varphi(z, \bar{z})$ unambiguously by $g(z, \bar{z}) = e^{\varphi}$. In terms of φ, the formulae (11) and (17) for the Gaussian curvature become

$$K = -\tfrac{1}{2} e^{-\varphi} \Delta\varphi, \qquad K = -2e^{-\varphi} \frac{\partial^2 \varphi}{\partial z\, \partial \bar{z}}. \tag{29}$$

If K is constant, then in the following form these equations are known as "Liouville's equations":

$$\Delta\varphi = -2Ke^{\varphi}, \qquad \frac{\partial^2 \varphi}{\partial z\, \partial \bar{z}} = -\frac{K}{2} e^{\varphi}. \tag{30}$$

13.3.1. Theorem. *A surface in Euclidean 3-space with metric $dl^2 = g(z, \bar{z})\, dz\, d\bar{z}$, and with constant Gaussian curvature K is locally isometric to:*

(i) *the sphere if $K > 0$;*
(ii) *the Euclidean plane if $K = 0$;*
(iii) *the Lobachevskian plane if $K < 0$.*

PROOF. From (29), with K assumed constant, we have

$$0 = \frac{\partial}{\partial z}\left(-\frac{K}{2}\right) = \frac{\partial}{\partial z}\left(e^{-\varphi} \frac{\partial^2 \varphi}{\partial z\, \partial \bar{z}}\right) = e^{-\varphi}\left(\frac{\partial^3 \varphi}{\partial z^2\, \partial \bar{z}} - \frac{\partial \varphi}{\partial z} \frac{\partial^2 \varphi}{\partial z\, \partial \bar{z}}\right)$$
$$= e^{-\varphi} \frac{\partial}{\partial \bar{z}}\left(\frac{\partial^2 \varphi}{\partial z^2} - \frac{1}{2}\left(\frac{\partial \varphi}{\partial z}\right)^2\right),$$

so that the function $\partial^2\varphi/\partial z^2 - \frac{1}{2}(\partial\varphi/\partial z)^2 = \psi(z)$ say, is analytic. Under a complex analytic co-ordinate change $z = f(w)$ the metric takes the form (see Lemma 12.1.1)

$$dl^2 = g(z, \bar{z}) \left| \frac{df}{dw} \right|^2 dw\, d\bar{w} = \tilde{g}(w, \bar{w})\, dw\, d\bar{w}.$$

If we then define $\tilde{\varphi}(w, \bar{w})$ by $\tilde{g}(w, \bar{w}) = e^{\tilde{\varphi}}$, it follows that

$$\tilde{\varphi}(w, \bar{w}) = \varphi(z, \bar{z}) + \ln \frac{df}{dw} + \ln \frac{d\bar{f}}{d\bar{w}}. \tag{31}$$

Of course since $\tilde{\varphi}$ was defined completely analogously to φ, the function $\partial^2\tilde{\varphi}/\partial w^2 - \frac{1}{2}(\partial\tilde{\varphi}/\partial w)^2 = \tilde{\psi}(w)$ say, will also be analytic (as a function of w). From this and (31) (and a little computation), we obtain

$$\tilde{\psi}(w) = \psi(z)(f')^2 + \frac{f'''}{f'} - \frac{3}{2}\left(\frac{f''}{f'}\right)^2,$$

where $f' = df/dw$. We wish to choose the function f so that $\tilde{\psi}(w)$ is identically zero, i.e. we seek an analytic function f satisfying the differential equation

$$\frac{f'''}{f'} - \frac{3}{2}\left(\frac{f''}{f'}\right)^2 = -\psi(f(w))(f')^2. \tag{32}$$

It can be shown using the theory of differential equations that (32) has an analytic solution; we shall not give the proof. (In the theory of functions of a complex variable, the left-hand side of (32) is known as the "Schwarz derivative.")

With f chosen to satisfy (32), we have $\tilde{\psi}(w) \equiv 0$, whence

$$\frac{\partial^2 e^{-\tilde{\varphi}/2}}{\partial w^2} = -\tfrac{1}{2} e^{-\tilde{\varphi}/2}\left(\frac{\partial^2\tilde{\varphi}}{\partial w^2} - \frac{1}{2}\left(\frac{\partial\tilde{\varphi}}{\partial w}\right)^2\right) = 0. \tag{33}$$

If we put $w = u + iv$, and analyse the meaning of (33) in terms of the original definition of the operator $\partial^2/\partial w^2$ (see §12(6)), we find that (33) is equivalent to

$$\frac{\partial^2 e^{-\tilde{\varphi}/2}}{\partial u\, \partial v} = 0, \qquad \frac{\partial^2 e^{-\tilde{\varphi}/2}}{\partial u^2} = \frac{\partial^2 e^{-\tilde{\varphi}/2}}{\partial v^2}.$$

It follows easily from these equations that

$$e^{-\tilde{\varphi}/2} = a(u^2 + v^2) + \text{a linear function of } u \text{ and } v, \tag{34}$$

which in complex notation takes the (anticipated) form

$$e^{-\tilde{\varphi}/2} = aw\bar{w} + bw + \bar{b}\bar{w} + c,$$

where a, c are real. Hence in terms of the new co-ordinates w, $\bar{w}$, our metric is given by

$$dl^2 = \frac{dw\, d\bar{w}}{(aw\bar{w} + bw + \bar{b}\bar{w} + c)^2}. \tag{35}$$

From the first of the formulae (29) (with $\tilde{\varphi}$ replacing φ), and (34), it follows that the Gaussian curvature is $K = 4(ac - b\bar{b})$. By means of linear-fractional transformations (which the reader may like to construct), the formula (35) can be brought into the forms:

(i) $\dfrac{4R^4\,dz\,d\bar{z}}{(1 + |z|^2)^2}$ if $K = 4(ac - b\bar{b}) > 0$;

(ii) $dz\,d\bar{z}$ if $K = 4(ac - b\bar{b}) = 0$;

(iii) $\dfrac{4R^4\,dz\,d\bar{z}}{(1 - |z|^2)^2}$ if $K = 4(ac - b\bar{b}) < 0$,

which are the familiar forms of the metrics of the sphere, Euclidean plane, and hyperbolic plane respectively. □

13.4. Exercises

1. Suppose the metric on a surface has the form

$$dl^2 = dx^2 + f(x)\,dy^2,$$

where $f(x)$ is a positive (real-valued) function. Prove that this metric can be brought into conformal form $dl^2 = g(u, v)(du^2 + dv^2)$.

2. Prove that a 2-dimensional pseudo-Riemannian metric (of type (1, 1)) with real analytic coefficients, takes the form

$$dl^2 = \lambda(t, x)(dt^2 - dx^2)$$

after a suitable co-ordinate change.

§14. Transformation Groups as Surfaces in N-Dimensional Space

14.1. Co-ordinates in a Neighbourhood of the Identity

Consider the group $GL(n, \mathbb{R})$ of invertible matrices A, i.e.

$$A = (a^i_j), \qquad \det(a^i_j) \neq 0. \tag{1}$$

Condition (1) defines a region in the space $M(n, \mathbb{R})$ of all $n \times n$ real matrices. If we regard $M(n, \mathbb{R})$ as a vector space (under matrix addition, and multiplication of matrices by scalars), then from this point of view the general linear group is a region of the vector space $\mathbb{R}^{n^2} = M(n, \mathbb{R})$. There is a natural system of co-ordinates for this space, namely the entries a^i_j of the general

$$K = \frac{\det(b_{ij})}{\det(g_{ij})} \qquad (14)$$

Pág. 109 (125)

§13. The Conformal Form of the Metric on a Surface

13.1 Isothermal Co-ordinates. Gaussian Curvature in terms of Conformal Co-ordinattes.

... the Euclidean metric in $\mathbb{R}^3$ induces on this surface a Riemannian metric

$$dl^2 = E(dp)^2 + 2F\,dp\,dq + G(dq)^2, \ldots$$

Important:

Pág. 117 (132)

13.2 Surfaces of Constant Curvature

Are there surfaces in Euclidean 3-space whose induced metric are hyperbolic (i.e. Lobachevskian)?

... there is a one-to-one metric-preserving map onto, say, the Poincaré ~~model~~ of the hyperbolic plane.

(%% % % 9 2 ! = " ! % 1!
-256- - -* ! - $ " 7 3 \ " ! "
211 ! ! & 9 7 3 ! J 9 ! ! 7 1 J ! " < " 1"

affirmative: there are surfaces in Euclidean 3-space which are, as they say "locally isometric" to the hyperbolic plane. (However there is no such surface that is globally isometric to the hyperbolic plane; i.e. from which there is a one-to-one metric-preserving map onto, say, the Poincaré model of the hyperbolic plane.) An example is provided by the following hornshaped surface of revolution (to which the name "pseudo-sphere" was first applied):

$$x = \operatorname{sech} u \cos v, \qquad y = \operatorname{sech} u \sin v, \qquad z = u - \tanh u, \qquad u \geqq 0,$$

where x, y, z are Euclidean co-ordinates for $\mathbb{R}^3$. The induced metric is easily calculated to be given by

$$dl^2 = (\tanh^2 u)\, du^2 + (\operatorname{sech}^2 u)\, dv^2,$$

which by a change of local co-ordinates can be brought into one of the usual forms for the Lobachevskian metric (we omit the verification of this). It is also easy to calculate that the Gaussian curvature $K \equiv -1$ (cf. the first of the remarks at the end of §10).

To return to our theme, suppose we are given a surface in Euclidean 3-space whose metric (in terms of certain local co-ordinates) has complex conformal form $dl^2 = g(z, \bar{z})\, dz\, d\bar{z}$. (By 13.1.1 the metrics of many surfaces in Euclidean 3-space can (locally) be brought into this form. By the preceding remarks there are such surfaces locally isometric to the Lobachevskian plane.) Since $g(z, \bar{z}) > 0$, we can define a new function $\varphi(z, \bar{z})$ unambiguously by $g(z, \bar{z}) = e^{\varphi}$. In terms of φ, the formulae (11) and (17) for the Gaussian curvature become

$$K = -\tfrac{1}{2} e^{-\varphi} \Delta\varphi, \qquad K = -2e^{-\varphi} \frac{\partial^2 \varphi}{\partial z\, \partial \bar{z}}. \tag{29}$$

If K is constant, then in the following form these equations are known as "Liouville's equations":

$$\Delta\varphi = -2Ke^{\varphi}, \qquad \frac{\partial^2 \varphi}{\partial z\, \partial \bar{z}} = -\frac{K}{2} e^{\varphi}. \tag{30}$$

13.3.1. Theorem. *A surface in Euclidean 3-space with metric $dl^2 = g(z, \bar{z})\, dz\, d\bar{z}$, and with constant Gaussian curvature K is locally isometric to*:

(i) *the sphere if $K > 0$;*
(ii) *the Euclidean plane if $K = 0$;*
(iii) *the Lobachevskian plane if $K < 0$.*

PROOF. From (29), with K assumed constant, we have

$$0 = \frac{\partial}{\partial z}\left(-\frac{K}{2}\right) = \frac{\partial}{\partial z}\left(e^{-\varphi} \frac{\partial^2 \varphi}{\partial z\, \partial \bar{z}}\right) = e^{-\varphi}\left(\frac{\partial^3 \varphi}{\partial z^2\, \partial \bar{z}} - \frac{\partial \varphi}{\partial z} \frac{\partial^2 \varphi}{\partial z\, \partial \bar{z}}\right)$$

$$= e^{-\varphi} \frac{\partial}{\partial \bar{z}}\left(\frac{\partial^2 \varphi}{\partial z^2} - \frac{1}{2}\left(\frac{\partial \varphi}{\partial z}\right)^2\right),$$

so that the function $\partial^2\varphi/\partial z^2 - \frac{1}{2}(\partial\varphi/\partial z)^2 = \psi(z)$ say, is analytic. Under a complex analytic co-ordinate change $z = f(w)$ the metric takes the form (see Lemma 12.1.1)

$$dl^2 = g(z, \bar{z})\left|\frac{df}{dw}\right|^2 dw\,d\bar{w} = \tilde{g}(w, \bar{w})\,dw\,d\bar{w}.$$

If we then define $\tilde{\varphi}(w, \bar{w})$ by $\tilde{g}(w, \bar{w}) = e^{\tilde{\varphi}}$, it follows that

$$\tilde{\varphi}(w, \bar{w}) = \varphi(z, \bar{z}) + \ln\frac{df}{dw} + \ln\frac{d\bar{f}}{d\bar{w}}. \tag{31}$$

Of course since $\tilde{\varphi}$ was defined completely analogously to φ, the function $\partial^2\tilde{\varphi}/\partial w^2 - \frac{1}{2}(\partial\tilde{\varphi}/\partial w)^2 = \tilde{\psi}(w)$ say, will also be analytic (as a function of w). From this and (31) (and a little computation), we obtain

$$\tilde{\psi}(w) = \psi(z)(f')^2 + \frac{f'''}{f'} - \frac{3}{2}\left(\frac{f''}{f'}\right)^2,$$

where $f' = df/dw$. We wish to choose the function f so that $\tilde{\psi}(w)$ is identically zero, i.e. we seek an analytic function f satisfying the differential equation

$$\frac{f'''}{f'} - \frac{3}{2}\left(\frac{f''}{f'}\right)^2 = -\psi(f(w))(f')^2. \tag{32}$$

It can be shown using the theory of differential equations that (32) has an analytic solution; we shall not give the proof. (In the theory of functions of a complex variable, the left-hand side of (32) is known as the "Schwarz derivative.")

With f chosen to satisfy (32), we have $\tilde{\psi}(w) \equiv 0$, whence

$$\frac{\partial^2 e^{-\tilde{\varphi}/2}}{\partial w^2} = -\tfrac{1}{2}e^{-\tilde{\varphi}/2}\left(\frac{\partial^2\tilde{\varphi}}{\partial w^2} - \frac{1}{2}\left(\frac{\partial\tilde{\varphi}}{\partial w}\right)^2\right) = 0. \tag{33}$$

If we put $w = u + iv$, and analyse the meaning of (33) in terms of the original definition of the operator $\partial^2/\partial w^2$ (see §12(6)), we find that (33) is equivalent to

$$\frac{\partial^2 e^{-\tilde{\varphi}/2}}{\partial u\,\partial v} = 0, \qquad \frac{\partial^2 e^{-\tilde{\varphi}/2}}{\partial u^2} = \frac{\partial^2 e^{-\tilde{\varphi}/2}}{\partial v^2}.$$

It follows easily from these equations that

$$e^{-\tilde{\varphi}/2} = a(u^2 + v^2) + \text{a linear function of } u \text{ and } v, \tag{34}$$

which in complex notation takes the (anticipated) form

$$e^{-\tilde{\varphi}/2} = aw\bar{w} + bw + \bar{b}\bar{w} + c,$$

where a, c are real. Hence in terms of the new co-ordinates w, $\bar{w}$, our metric is given by

$$dl^2 = \frac{dw\,d\bar{w}}{(aw\bar{w} + bw + \bar{b}\bar{w} + c)^2}. \tag{35}$$

From the first of the formulae (29) (with $\tilde{\varphi}$ replacing φ), and (34), it follows that the Gaussian curvature is $K = 4(ac - b\bar{b})$. By means of linear-fractional transformations (which the reader may like to construct), the formula (35) can be brought into the forms:

$$\text{(i)}\ \frac{4R^4\,dz\,d\bar{z}}{(1+|z|^2)^2} \quad \text{if } K = 4(ac - b\bar{b}) > 0;$$

$$\text{(ii)}\ dz\,d\bar{z} \qquad \text{if } K = 4(ac - b\bar{b}) = 0;$$

$$\text{(iii)}\ \frac{4R^4\,dz\,d\bar{z}}{(1-|z|^2)^2} \quad \text{if } K = 4(ac - b\bar{b}) < 0,$$

which are the familiar forms of the metrics of the sphere, Euclidean plane, and hyperbolic plane respectively. □

13.4. Exercises

1. Suppose the metric on a surface has the form

$$dl^2 = dx^2 + f(x)\,dy^2,$$

where $f(x)$ is a positive (real-valued) function. Prove that this metric can be brought into conformal form $dl^2 = g(u, v)(du^2 + dv^2)$.

2. Prove that a 2-dimensional pseudo-Riemannian metric (of type (1, 1)) with real analytic coefficients, takes the form

$$dl^2 = \lambda(t, x)(dt^2 - dx^2)$$

after a suitable co-ordinate change.

§14. Transformation Groups as Surfaces in N-Dimensional Space

14.1. Co-ordinates in a Neighbourhood of the Identity

Consider the group $GL(n, \mathbb{R})$ of invertible matrices A, i.e.

$$A = (a^i_j), \qquad \det(a^i_j) \neq 0. \tag{1}$$

Condition (1) defines a region in the space $M(n, \mathbb{R})$ of all $n \times n$ real matrices. If we regard $M(n, \mathbb{R})$ as a vector space (under matrix addition, and multiplication of matrices by scalars), then from this point of view the general linear group is a region of the vector space $\mathbb{R}^{n^2} = M(n, \mathbb{R})$. There is a natural system of co-ordinates for this space, namely the entries a^i_j of the general so that the eigenvector (ξ^i) is a (non-zero) solution of the linear system of equations

$$T_{kj}\xi^j = \lambda g_{ki}\xi^i, \qquad k = 1, \dots, n. \tag{11}$$

It is clear that the trace tr $T^i_j = T^i_i = g^{ik}T_{ik}$, and determinant det T^i_j of the linear operator $T^i_j = g^{ik}T_{kj}$, are metrical invariants of the form T_{ij}, i.e. they are invariant under co-ordinate changes, but of course depend on the metric g_{ij}. As the following example will recall to the reader, we have met these invariants before, albeit in a slightly different guise.

19.2.4. Example. In Chapter 2 we associated with (each point of) a surface $r = r(u, v)$; $r = (x, y, z)$, $x^1 = u$, $x^2 = v$, in Euclidean 3-space, two quadratic forms:

(i) the induced metric $g_{ij}\,dx^i\,dx^j$, or tensor g_{ij};
(ii) the second fundamental form $b_{ij}\,dx^i\,dx^j$, or tensor b_{ij}.

(Here the summation is from 1 to 2, since the surface is 2-dimensional.)

In §8 we obtained the following formulae expressing the Gaussian and mean curvatures of the surface in terms of these two tensors:

$$\text{Gaussian curvature } K = \frac{\det(b_{ij})}{\det(g_{ij})};$$

$$\text{mean curvature } H = b^i_i = g^{ij}b_{ij}.$$

Thus we see that, in our present terminology, the mean curvature is just the trace of the tensor b_{ij} relative to the metric g_{ij}, while since

$$K = (\det(g_{ij}))^{-1}\det(b_{ij}) = \det(g^{ik}b_{kj}) = \det(b^i_j),$$

the Gaussian curvature is the second of the above-mentioned metrical invariants of the second fundamental form b_{ij}.

19.3. The Operator $*$

The presence of a metric g_{ij} allows us to derive from any skew-symmetric tensor T of type $(0, k)$ a skew-symmetric tensor $*T$ of type $(0, n-k)$, defined as follows.

19.3.1. Definition. Given a skew-symmetric tensor T of type $(0, k)$, with components $T_{i_1 \dots i_k}$, we define the skew-symmetric tensor $*T$ of type $(0, n-k)$ by the formula

$$(*T)_{i_{k+1}\dots i_n} = \frac{1}{k!}\sqrt{|g|}\;\varepsilon_{i_1\dots i_n}T^{i_1\dots i_k}, \tag{12}$$

(%/B ' % ' 9

coincides (to within a constant factor) with the expression (24) for the energy-momentum tensor of an electromagnetic field.

In the case of so-called "empty" space, i.e. in the absence of matter, we have $T_{ik} \equiv 0$, so that the gravitational field equations take the form $R_{ik} - \frac{1}{2}Rg_{ik} = 0$, or more simply $R_{ik} = 0$ (this was noted in §30.3(iii)). To see how the latter equations are arrived at, note first that (36) may be rewritten as

$$R_k^\alpha - \tfrac{1}{2}R\delta_k^\alpha = \frac{8\pi G}{c^4} T_k^\alpha.$$

On contracting the indices α and k this yields $R - 2R = (8\pi G/c^4)T$, where $T = T_\alpha^\alpha$, whence

$$R = -\frac{8\pi G}{c^4} T.$$

Substituting for R from this into the original equations (36) for the general gravitational field we obtain

$$R_{ik} = \frac{8\pi G}{c^4}(T_{ik} - \tfrac{1}{2}g_{ik}T).$$

Thus if $T_{ik} = 0$, then $R_{ik} = 0$, as claimed.

It does not follow from the equations $R_{ik} \equiv 0$ that empty space-time is flat: the vanishing of the Ricci tensor does not in general entail the vanishing of the Riemann curvature tensor. (Recall that on the other hand if our space were 3-dimensional then $R_{jkl}^i \equiv 0$ would follow from $R_{ik} \equiv 0$, since in this case the Riemann curvature tensor is expressible in terms of the Ricci tensor (see §30.3(ii)).

From Theorem 37.4.1 (or, more precisely, its proof) we obtain the following consequence for 2-dimensional spaces.

37.4.2. Theorem. *Let g_{ij}, $i, j = 1, 2$, be a Riemannian metric on a 2-dimensional space, and let K denote as usual the Gaussian curvature of the space. Then the integral $S[g] = \int K\,dS$, where $dS = \sqrt{g}\,dx^1 \wedge dx^2$, is invariant under smooth local perturbations of the metric.*

PROOF. From Theorem 30.3.3 and its proof we obtain: $K = R/2$, $R_{ij} = \frac{1}{2}Rg_{ij}$, whence $R_{ij} - \frac{1}{2}Rg_{ij} \equiv 0$. Hence from Theorem 37.4.1 (whose proof was, in essence, independent of the number of dimensions and the fact that the metric was pseudo-Riemannian), we have

$$\frac{1}{\sqrt{g}}\frac{\delta S}{\delta g^{ij}} = R_{ij} - \tfrac{1}{2}Rg_{ij} \equiv 0,$$

which yields the desired conclusion. ☐

Since a closed surface in Euclidean $\mathbb{R}^3$ is by definition compact (and without boundary), we deduce immediately the following celebrated result of Gauss and Bonnet.

37.4.3. Corollary (Gauss–Bonnet). *The integral over a closed surface in 3-dimensional Euclidean space of the Gaussian curvature of the surface, is not changed by smooth deformations of the surface*:

$$\int K\,dS = \text{const.}$$

The value of this constant, and its significance, will be discussed in Part II.

37.5. Soap Films

Consider a smooth hypersurface V^{n-1} in Euclidean $\mathbb{R}^n$ with Euclidean coordinates $x^1, \ldots, x^n$, which we shall assume to be given in the graphical form $x^n = f(x^1, \ldots, x^{n-1})$, where the domain of definition of the function f is a bounded region D of $\mathbb{R}^{n-1}$. In this subsection our concern will be with the area functional $S[f]$ defined on the space of all functions f with the fixed domain D. Thus

$$S[f] = \int_D \sqrt{\det A}\, d^{n-1}x,$$

where $A = (g_{ij}(x))$, $x \in D$, is the induced Riemannian metric on the surface V^{n-1} defined by f, and $d^{n-1}x = dx^1 \wedge \cdots \wedge dx^{n-1}$.

We shall now express $\sqrt{\det A}$ explicitly in terms of the function f. We might for this purpose appeal directly to formula (31) of §7.4 (or rather to its generalization to surfaces of $(n-1)$ dimensions). Alternatively we may proceed as follows. Denote by $d\tau^{n-1}$ the form representing the $(n-1)$-dimensional volume element of V^{n-1}; thus $S[f] = \int_D d\tau^{n-1}$. For each point P of V^{n-1}, let $\alpha(P)$ denote the angle between $n(P)$, the unit normal to V^{n-1} at P, and $e_n = (0, \ldots, 0, 1)$. Then as Figure 39 suggests (at least in the case $n = 3$), we have

$$S[f] = \int_D d\tau^{n-1} = \int_D \frac{d^{n-1}x}{\cos\alpha(P)}.$$

Now since $n(P)$ is given by (cf. the formula for the unit normal in the case $n = 3$ given in §8.3(23)):

$$n(P) = \frac{1}{\sqrt{1 + \sum_{i=1}^{n-1} (f_{x^i})^2}}(-f_{x^1}, \ldots, -f_{x^{n-1}}, 1),$$

Pág. 118(133) -

13.3.1 Theorem

A surface in Euclidean 3-space with metric $dl^2 = g(z,\bar{z})\,dz\,d\bar{z}$, and with constant Gaussian curvature K is locally isometric to:

(i) the sphere if $K>0$;

(ii) the Euclidian plane if $K=0$;

(iii) the Lobachevskian plane if $K<0$.

Pág. 171(186)

Gaussian curvature $K = \dfrac{\det(b_{ij})}{\det(g_{ij})}$;

mean curvature $H = b^i_i = g^{ij} b_{ij}$.

Pág. 396(411)

37.4.2 Theorem

Let g_{ij}, $i, j = 1, 2$, be a Riemannian metric on a 2-dimensional space, and

$/ (/%

-

let K denote as usual the Gaussian curvature of the space.

Important:

Pág. 397(412) –

37.4.3. Corollary (Gauss-Bonnet)

The integral over a closed surface in 3-dimensional Euclidean space of the Gaussian curvature of the surface is not changed by smooth deformations of the surface:

$$\int K\,dS = \text{const}.$$

– Modern Geometry – Methods and Applications – Part II ... pdf

Pág. 94(110) –

where K is the Gaussian curvature of M

Pág. 95(111) – Figure 27

Gauss map r to S^{q-1}.

Pág. 112(128) 14.2

$/
-256- - -* ! - $/ " 7 3\ " ! "
211 ! ! && 9 7 3 ! " 9 1 3 ! " 1 ! 8 1"
(/'

$M \to \mathbb{R}^{2n}$, are no longer everywhere dense in the space of all smooth maps $M \to \mathbb{R}^{2n}$. In general no further reduction of the dimension is possible: we invite the reader to prove that a non-orientable, closed 2-manifold is not embeddable in $\mathbb{R}^3$ (see §2.1). However in particular cases, i.e. for more restricted classes of manifolds, the dimension can be further reduced; for instance, it can be shown on the basis of the known classification of compact, orientable, 2-dimensional manifolds, that any such manifold can be embedded in $\mathbb{R}^3$.

11.2. The Construction of Morse Functions as Height Functions

With the aid of an embedding $M \to \mathbb{R}^N$, we shall now give a different proof of the existence of Morse functions on a compact smooth manifold M. We shall show that Morse functions can be found in a rather narrow class of functions on manifolds, called "height functions", which we shall now define.

Let M be smoothly embedded in $\mathbb{R}^N$, and let (in distinction to the previous subsection) $\xi_l(t)$ denote a (parametrized) straight line through the origin of $\mathbb{R}^N$ with direction vector l. Define a function h_l on M by setting $h_l(x)$ equal to t_x where t_x is that value of the parameter t at the point of orthogonal projection $\xi_l(t_x)$ of $x \in M$ onto the line $\xi_l(t)$. (Thus if t is distance (positive or negative) on $\xi_l(t)$ from the origin, then $h_l(x)$ is the distance of $x \in M$ from the origin (or "height above the origin") in the direction l. Such functions h_l are called *height functions* on $M \subset \mathbb{R}^N$.

The reader will easily verify the following two properties of height functions:

(i) The set of height functions on M is (neglecting the particular parametrizations of the $\xi_l(t)$) in one-to-one correspondence with the set of pairs of diametrically opposite points of the (unit) sphere S^{N-1}, or, equivalently, the points of $\mathbb{R}P^{N-1}$;
(ii) A point x_0 of M is a critical point of a height function h_l if and only if the vector l is orthogonal to the manifold M at the point x_0 (i.e. $l \perp T_{x_0}M$).

Our next task is to discover conditions under which a critical point $x_0 \in M$ of a height function h_l is non-degenerate. We shall at this point restrict our considerations to manifolds M embedded in $\mathbb{R}^{m+1}$, where $m = \dim M$, i.e. to hypersurfaces in $\mathbb{R}^{m+1}$ (but see the "Important Remark" below). Recall (from §26.2 of Part I) that for hypersurfaces $M \subset \mathbb{R}^{m+1}$ the *Gauss map* $r\colon M \to S^m \to \mathbb{R}P^m$ is defined by $r(x) = n(x)$, where $n(x)$ is the unit normal vector to M at the point x (translated back to the origin, i.e. with its tail at the origin).

11.2.1. Lemma. *A point x_0 of the hypersurface $M \subset \mathbb{R}^{m+1}$ is a non-degenerate critical point of a height function h_l if and only if it is a regular* (*i.e. non-critical*) *point of the Gauss map $r\colon M \to \mathbb{R}P^m$* (*and of course also $l \perp T_{x_0}M$*).

Proof. By means of a suitable orthogonal change of co-ordinates on $\mathbb{R}^{m+1}$ we can arrange that the x^{m+1}-axis is parallel to l (so that the other m-axes are perpendicular to l). The hyperplane $\mathbb{R}^m(x^1, \ldots, x^m)$ can then be regarded as the tangent space to M at x_0, and in some sufficiently small neighbourhood of x_0 the manifold M will be given by an equation of the form $x^{m+1} = \varphi(x^1, \ldots, x^m)$, with $d\varphi|_{x_0} = 0$, and $x^1, \ldots, x^m$ will serve as local co-ordinates. Then on that neighbourhood the "height" $h_l(x)$ is just the function $\varphi(x^1, \ldots, x^m) = x^{m+1}$. By defining in terms of $x^1, \ldots, x^{m+1}$, as in the proof of Theorem 26.2.2 of Part I, suitable local co-ordinates $y^1, \ldots, y^m$ on the sphere S^m in a neighbourhood of the image point $r(x_0)$, and repeating the argument given there (with $m+1$ in place of 3) leading to the formula $K\, d\sigma = r^*(\Omega)$, where K is the Gaussian curvature of M (which may be taken to be defined by the equation below), and $d\sigma$, Ω are the (induced) volume elements on M and S^m, respectively, we obtain the following equations:

$$\left(\frac{\partial^2 h_l(x_0)}{\partial x^\alpha\, \partial x^\beta}\right) = \left(\frac{\partial^2 \varphi(x_0)}{\partial x^\alpha\, \partial x^\beta}\right) = \left(\left.\frac{\partial y^\alpha}{\partial x^\beta}\right|_{x_0}\right);$$

$$\det\left(\frac{\partial^2 h_l(x_0)}{\partial x^\alpha\, \partial x^\beta}\right) = K,$$

where here K is the Gaussian curvature of M at x_0, and $y^\alpha = r^\alpha(x^1, \ldots, x^m)$, $1 \le \alpha \le m$, is the Gauss map expressed in terms of the local co-ordinates $x^1, \ldots, x^m$; $y^1, \ldots, y^m$. It is immediate from the first of the above equations that the condition that the Gauss map r be regular at x_0, namely $\det(\partial y^\alpha/\partial x^\beta|_{x_0}) \ne 0$, is equivalent to the defining condition for non-degeneracy of the critical point x_0 of h_l, namely $\det(\partial^2 h_l(x_0)/\partial x^\alpha\, \partial x^\beta) \ne 0$. This completes the proof of the lemma. □

11.2.2. Theorem. *A height function h_l on a hypersurface $M \subset \mathbb{R}^{m+1}$ ($m = \dim M$) is a Morse function precisely if the point $\{\pm l\}$ of $\mathbb{R}P^m$ is a regular value of the Gauss map r: $M \to \mathbb{R}P^m$. It follows that the Morse functions are everywhere dense in the set of height functions on M.*

Proof. The first statement is immediate from the lemma. The second follows from Sard's theorem, which tells us that the regular values of the Gauss map are everywhere dense in $\mathbb{R}P^m$. □

Important Remark. This theorem is valid without the restriction that M be a hypersurface in $\mathbb{R}^m$, i.e. it holds quite generally for any smooth embedding $M \subset \mathbb{R}^q$ ($q \ge m+1$). We now sketch an argument allowing us to deduce this general result from Lemma 11.2.1. Thus suppose $M \subset \mathbb{R}^q$ with $q > m+1$, where $m = \dim M$. We shall now define a $(q-1)$-dimensional manifold N, as the boundary of a so-called "tubular neighbourhood" of the manifold M. To this end consider the set of $(q-m)$-dimensional discs D_x^{q-m} in $\mathbb{R}^q$ (i.e. balls) orthogonal to the submanifold M, of radius $\varepsilon > 0$, and with centres at the points x of M. For sufficiently small $\varepsilon > 0$, the union of these discs is a q-

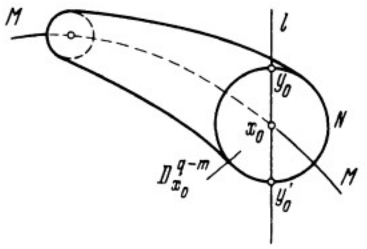

Figure 27

dimensional submanifold of $\mathbb{R}^q$ (a "tubular neighbourhood" of M), whose boundary we denote by N (see Corollary 11.3.3 below). Since N has dimension $q-1$, it can be mapped via the Gauss map r to S^{q-1}.

Now let $h_l(x)$ be any height function defined for points x in M or N (i.e. on $M \cup N$). It is intuitively clear that with each critical point $x_0 \in \mathrm{M}$ of the function $h_l(x)$ there are associated exactly two critical points y_0 and y_0' of h_l restricted to N, namely the two points of intersection with N of the straight line parallel to l passing through x_0 and orthogonal to M at x_0 (see Figure 27).

It may be shown that a critical point $x_0 \in M$, of the function h_l restricted to M, is non-degenerate precisely if the points y_0, y_0' are non-degenerate for h_l restricted to N. (They are always either both degenerate or both non-degenerate.) Hence if h_l is a Morse function on N, then it is also a Morse function when restricted to M. The existence and everywhere denseness (in the set of all height functions on N) of Morse functions on N, guaranteed by Theorem 11.2.2, therefore implies the analogous conclusion for M. We have thus shown that the restriction that M be a hypersurface in $\mathbb{R}^{m+1}$ was inessential.

11.3. Focal Points

There exist yet other simple methods of obtaining a plentiful supply of Morse functions on a smooth manifold. Without going into great detail, we describe one such method.

Consider a smooth m-manifold M smoothly embedded as a submanifold of $\mathbb{R}^q$. With each point $p \in \mathbb{R}^q$, we associate a smooth function L_p on the submanifold M, defined by $L_p(x) = |p-x|^2$, where $x \in M$ and $|p-x|$ is the Euclidean distance between the points $p, x \in \mathbb{R}^q$. We shall show that for almost all points $p \in \mathbb{R}^q$ (i.e. for all points except those in a set of measure zero) the function L_p is a Morse function on M. (Clearly the set of functions of the form $L_p(x)$ is different from the set of height functions $h_l(x)$.)

For the proof we first define the "focal points" for $M \subset \mathbb{R}^q$; our ultimate aim will then be to show that the set of focal points, which has measure zero,

(' 9 4 ! < " ! :31 !

over the respective regions are equal in absolute value (which is not surprising if f is one-to-one, i.e. if f is a co-ordinate change).

14.2. The Degree of a Vector Field on a Hypersurface

Let $\xi=(\xi^\alpha(x))$, $\alpha=1,\dots,n$, denote a smooth, non-vanishing vector field defined on a region U of n-dimensional Euclidean space $\mathbb{R}^n$ with Euclidean co-ordinates $x^1,\dots,x^n$. Since $\xi(x)\neq 0$ at all points of U, the vector field ξ gives rise to a unit vector field $n(x)=\xi(x)/|\xi(x)|$ on U, which leads in turn to the *spherical* (*Gauss*) *map* $f\colon U\to S^{n-1}$, where for each point $x\in U$, $f(x)$ is the tip of the unit vector $n(x)$ when its tail is at the origin. (Thus $f(x)$ is a point on the unit $(n-1)$-sphere, as required.) If now Q is any hypersurface entirely within U, then the degree of the restriction map $f|_Q\colon Q\to S^{n-1}$, is called the *degree of the vector field* ξ *on the hypersurface* Q.

Consider the following closed $(n-1)$-form Ω on S^{n-1} (or, if you like, on $\mathbb{R}^n\backslash\{0\}$):

$$\Omega=\frac{1}{\gamma_n}\,\frac{\sum_{i=1}^{n}(-1)^{i+1}x^i\,dx^1\wedge\cdots\wedge\widehat{dx^i}\wedge\cdots\wedge dx^n}{((x^1)^2+\cdots+(x^n)^2)^{n/2}}. \tag{3}$$

Here the hatted symbols are to be omitted, and γ_n is the normalizing constant chosen so that $\int_{S^{n-1}}\Omega=1$. (It is not difficult to verify that the form $\gamma_n\Omega$ is just the volume element on the unit sphere S^{n-1}, whence γ_n is the volume of S^{n-1}.) For instance in the plane $\mathbb{R}^2\backslash\{0\}$ with Euclidean co-ordinates x, y, we have

$$\Omega=\frac{1}{2\pi}\left(\frac{x\,dy-y\,dx}{x^2+y^2}\right),$$

or, in polar co-ordinates, $\Omega=d\varphi/2\pi$. In $\mathbb{R}^3\backslash\{0\}$ with Euclidean co-ordinates x, y, z, we have

$$\Omega=\frac{1}{4\pi}\left(\frac{x\,dy\wedge dz-y\,dx\wedge dz+z\,dy\wedge dx}{(x^2+y^2+z^2)^{3/2}}\right),$$

or, in spherical co-ordinates, $\Omega=(1/4\pi)|\sin\theta|\,d\theta\,d\varphi$.

From Theorem 14.1.1 we deduce the following

14.2.1. Corollary. *The degree of a non-vanishing vector field* $\xi(x)$ *on a closed hypersurface* Q *in Euclidean n-space, is equal to the integral* $\int_Q f^*\Omega$, *where* $f\colon Q\to S^{n-1}$ *is the Gauss map determined by the vector field, and* Ω *is the* $(n-1)$*-form on* S^{n-1} *defined in* (3). *It follows that if* Q *is given locally by a system of equations*

$$x^\alpha=x^\alpha(u^1,\dots,u^{n-1}),\qquad \alpha=1,\dots,n,$$

(*where* $x^1,\dots,x^n$ *are Euclidean co-ordinates in* $\mathbb{R}^n$), *then*

$$\deg f = \deg_Q \xi = \frac{1}{\gamma_n}\int_Q \frac{1}{|\xi|^n} \det \begin{pmatrix} \xi^1 & \cdots & \xi^n \\ \dfrac{\partial \xi^1}{\partial u^1} & \cdots & \dfrac{\partial \xi^n}{\partial u^1} \\ \cdots & \cdots & \cdots \\ \dfrac{\partial \xi^1}{\partial u^{n-1}} & \cdots & \dfrac{\partial \xi^n}{\partial u^{n-1}} \end{pmatrix} du^1 \wedge \cdots \wedge du^{n-1}. \tag{4}$$

(The first statement is immediate from Theorem 14.1.1. The second follows easily from the definitions of Ω, of f (via $f(x)=(1/|\xi(x)|)\,(\xi^1(x), \ldots, \xi^n(x))$), and of $f^*\Omega$ (see §22.1 of Part I).)

In the case $n=2$, the formula (4) becomes:

$$\deg f = \frac{1}{2\pi}\oint \frac{dt}{|\xi|^2}\left(\xi^1 \frac{d\xi^2}{dt} - \xi^2 \frac{d\xi^1}{dt}\right),$$

where t is a parameter on the closed curve over which the integral is taken, and $\xi(t)=(\xi^1(t), \xi^2(t))$ is a non-vanishing vector field on the curve. In the case $n=3$, we have $\gamma_3=4\pi$ whence (4) becomes:

$$\deg f = \frac{1}{4\pi}\iint_Q \frac{du\,dv}{|\xi|^3} \det \begin{pmatrix} \xi^1 & \xi^2 & \xi^3 \\ \dfrac{\partial \xi^1}{\partial u} & \dfrac{\partial \xi^2}{\partial u} & \dfrac{\partial \xi^3}{\partial u} \\ \dfrac{\partial \xi^1}{\partial v} & \dfrac{\partial \xi^2}{\partial v} & \dfrac{\partial \xi^3}{\partial v} \end{pmatrix} \tag{5}$$

$$= \frac{1}{4\pi}\iint_Q \frac{du\,dv}{|\xi|^3}\left\langle \xi, \left[\frac{\partial \xi}{\partial u}, \frac{\partial \xi}{\partial v}\right]\right\rangle,$$

where $[\ ,\ ]$ denotes the vector (or cross) product.

A particular situation of interest is that where the vector field $\xi(x)$ is a unit field (i.e. $|\xi|=1$), and at every point of the closed manifold Q has the direction of the exterior normal to Q. We know (from §26.2 of Part I) that in this case

$$\gamma_n(f^*\Omega) = K\,d\sigma = K\sqrt{g}\,du^1 \wedge \cdots \wedge du^{n-1}, \tag{6}$$

where K is the Gaussian curvature of the hypersurface (defined as the product of the principal curvatures), and $d\sigma=\sqrt{g}\,du^1 \wedge \cdots \wedge du^{n-1}$ is the standard volume element on the hypersurface Q determined by the metric induced on Q by the Euclidean metric on the ambient space $\mathbb{R}^n$. (Recall that when $n=2$ we have $d\sigma=dl$, the element of length on the curve Q, and $K=k$, the curvature of the curve, while for $n=3$, K is the usual Gaussian curvature of a 2-dimensional surface, and $d\sigma=\sqrt{EG-F^2}\,du \wedge dv$ is the usual element of area (see §7 of Part I).

Combining (6) with the first statement of Corollary 14.2.1, we obtain immediately the following

14.2.2. Theorem. *The integral of the Gaussian curvature over a closed hypersurface in Euclidean n-space is equal to the degree of the Gauss map of the surface, multiplied by* γ_n *(the Euclidean volume of the unit* $(n-1)$*-sphere).*

14.3. The Whitney Number. The Gauss–Bonnet Formula

Our aim in this subsection is to discover the relationship between the degree of the Gauss map of a hypersurface in Euclidean n-space and certain striking geometrical features of the hypersurface (or at least of the way it is embedded in $\mathbb{R}^n$). We shall restrict attention to the most important cases $n=2, 3$.

The case $n=2$: *curves in* $\mathbb{R}^2$. Let $\gamma=(x(t), y(t))$ be a closed curve "in general position" in Euclidean $\mathbb{R}^2$; by this we mean that: for all t, $x(t+2\pi)=x(t)$, $y(t+2\pi)=y(t)$; the tangent vector $(\dot{x}, \dot{y})$ is non-zero for all t; and finally every self-intersection (if there are any) is double and has the property that the two tangent vectors at the point of intersection are linearly independent (as for instance in Figure 34).

To each point of self-intersection of the closed curve γ we attach a sign (± 1) in the following way. Fix on any point t_0 of the curve which is not a point of self-intersection. Denote by $[1, 2]$ an orienting frame on $\mathbb{R}^2$ (e.g. that determined by the given Euclidean co-ordinates x, y, in that order). We now trace out the curve γ once starting from the point t_0 in the direction of increase of the parameter; when any particular point of self-intersection is first encountered we label with the numeral 1 the tangent vector at that point to the branch we are on, and when we meet that point of self-intersection again we label the tangent vector to the second branch (on which we now find ourselves) with the numeral 2 (see Figure 34). Once our tour is completed, each point of self-intersection of γ has associated with it a frame whose orientation class of course coincides either with that of the prescribed frame $[1, 2]$, or its opposite; in the former case we attach the sign $+1$ to the point, and in the latter case the sign -1. We then define the *Whitney number* $W(\gamma)$ of the plane closed curve γ (in general position) to be the sum of the signs of all its points of self-intersection.

The following result links the degree of the Gauss map to the Whitney number.

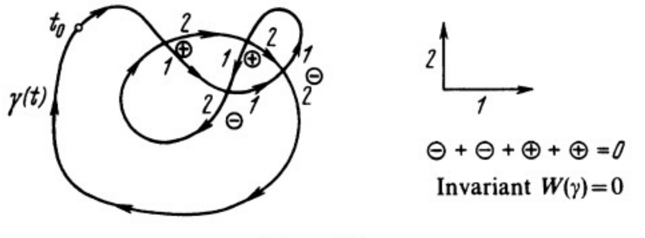

Figure 34

(' % '%9 @ 3 8 D 9 7!) < !

14.2.2. Theorem. *The integral of the Gaussian curvature over a closed hypersurface in Euclidean n-space is equal to the degree of the Gauss map of the surface, multiplied by γ_n (the Euclidean volume of the unit $(n-1)$-sphere).*

14.3. The Whitney Number. The Gauss–Bonnet Formula

Our aim in this subsection is to discover the relationship between the degree of the Gauss map of a hypersurface in Euclidean n-space and certain striking geometrical features of the hypersurface (or at least of the way it is embedded in $\mathbb{R}^n$). We shall restrict attention to the most important cases $n=2, 3$.

The case $n=2$: curves in $\mathbb{R}^2$. Let $\gamma=(x(t), y(t))$ be a closed curve "in general position" in Euclidean $\mathbb{R}^2$; by this we mean that: for all t, $x(t+2\pi)=x(t)$, $y(t+2\pi)=y(t)$; the tangent vector $(\dot{x}, \dot{y})$ is non-zero for all t; and finally every self-intersection (if there are any) is double and has the property that the two tangent vectors at the point of intersection are linearly independent (as for instance in Figure 34).

To each point of self-intersection of the closed curve γ we attach a sign (± 1) in the following way. Fix on any point t_0 of the curve which is not a point of self-intersection. Denote by $[1, 2]$ an orienting frame on $\mathbb{R}^2$ (e.g. that determined by the given Euclidean co-ordinates x, y, in that order). We now trace out the curve γ once starting from the point t_0 in the direction of increase of the parameter; when any particular point of self-intersection is first encountered we label with the numeral 1 the tangent vector at that point to the branch we are on, and when we meet that point of self-intersection again we label the tangent vector to the second branch (on which we now find ourselves) with the numeral 2 (see Figure 34). Once our tour is completed, each point of self-intersection of γ has associated with it a frame whose orientation class of course coincides either with that of the prescribed frame $[1, 2]$, or its opposite; in the former case we attach the sign $+1$ to the point, and in the latter case the sign -1. We then define the *Whitney number* $W(\gamma)$ of the plane closed curve γ (in general position) to be the sum of the signs of all its points of self-intersection.

The following result links the degree of the Gauss map to the Whitney number.

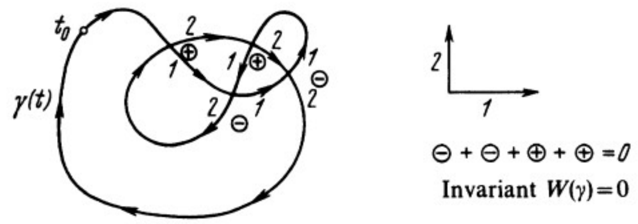

Figure 34

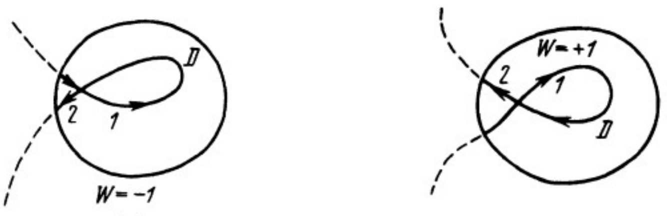

(a) (b)

Figure 35

14.3.1. Theorem. *Let γ be a non-singular, plane closed curve in general position. Then the parity of the number of self-intersections of γ (which is the same as the parity of the Whitney number of γ) is opposite to that of the degree of the Gauss map f of γ.*

PROOF. For a simple (i.e. non-self-intersecting) plane closed curve γ it is easy to see that $\deg f = 1$ (assuming the curve γ and S^1 traced in the same sense), while clearly $W(\gamma) = 0$; hence the theorem is true in this case. With this as our point of departure we shall prove the theorem by induction on the number of self-intersections. Suppose first that our curve γ has a loop with the properties that the "starting point" t_0 is not on the loop, and no points of γ are in its interior (see Figure 35). Enclose this loop in a region D which excludes all points of γ not on the loop (as in Figure 35); subsequent alterations of γ will be carried out entirely in the region D. The (direction-preserving) alterations in question are depicted in Figure 36(a) or (b) (depending on the sense in which the loop is traced). The operation of replacing γ by $\gamma_1 \cup \gamma_2$ and then removing γ_2, decreases the degree of f by 1 in case (a), and increases the degree by 1 in case (b); clearly the Whitney number is affected in the same way. Thus the inductive step is completed in the situation where γ has such a "minimal" loop.

Suppose now that γ has no "minimal" loops. It can be shown that there will still be a loop γ_2 say (see Figure 37) which does not have the point t_0 on it, but which intersects the remainder of the curve (denoted by γ_1 in Figure 37). If we now alter the curve γ appropriately within some sufficiently small neighbourhood of the point of self-intersection of our loop we shall arrive

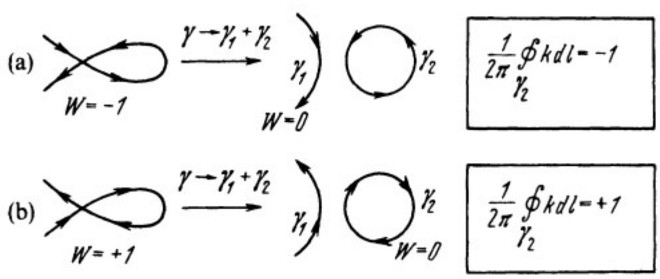

Figure 36

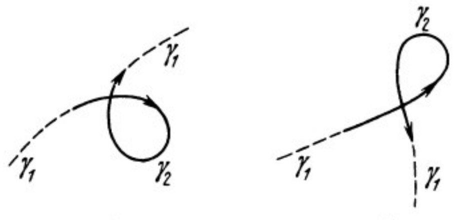

(a) (b)

Figure 37

again at the situation depicted in Figure 36(a) or (b), except that now the simple closed curve γ_2 of that figure will intersect γ_1. However since the number of such points of intersection will of course be even, the argument concerning parities goes through as before. This completes the proof of the theorem. □

Remark. It can be shown that under the above-described alteration of γ and subsequent removal of the simple closed component γ_2, the Whitney number and the degree of f change by the same amount (and not just modulo 2). This leads to the more precise result that *the Whitney number* $W(\gamma)$ *is equal to either* $\deg f-1$ *or* $\deg f+1$. Whether 1 is to be added or subtracted depends on the initial choice of the starting point t_0 (as illustrated in Figure 38).

The case $n=3$: *surfaces in* $\mathbb{R}^3$. Our aim is to express the degree of the Gauss map $f: Q\rightarrow S^2$, of a smooth, closed, oriented surface Q in $\mathbb{R}^3$, more directly in terms of the geometry of Q. Let $y_0\in S^2$ be a regular value of the map f. By applying a suitable orthogonal transformation we may clearly arrange for the point y_0 to be at the north pole of S^2, i.e. to have co-ordinates $(0, 0, 1)$, and, by means of a suitable, small deformation (i.e. homotopy), we may further arrange for the south pole $y_0^*=(0, 0, -1)$ to be also a regular value of f. (The simultaneous regularity of two diametrically opposite points of S^2 is equivalent to the regularity of their common image point in the real projective plane $\mathbb{R}P^2$ under the composite map

$$Q\xrightarrow{f}S^2\rightarrow\mathbb{R}P^2,$$

from Q to $\mathbb{R}P^2$.)

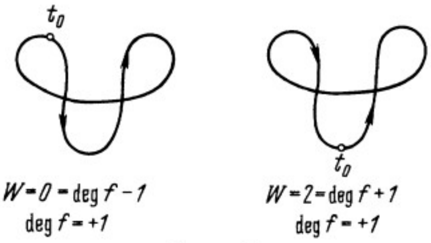

Figure 38

14.3.2. Lemma. *Let Q be a hypersurface in Euclidean 3-space, with f (the Gauss map of Q), y_0, y_0^* as above, and let φ be the height function on Q (i.e. for each $P \in Q$, $\varphi(P)$ is the z-co-ordinate of P). Then the set of critical points of the function φ coincides with the union $f^{-1}(y_0) \cup f^{-1}(y_0^*)$, and the critical points are all non-degenerate.*

Proof. In some neighbourhood of each critical point of $\varphi: Q \to \mathbb{R}$, the surface Q is defined by an equation of the form $z = \varphi(x, y)$, and the critical points in that neighbourhood are just those satisfying grad $\varphi = 0$. Now the gradient of $\varphi(x, y)$ is zero at precisely those points of Q where the direction of the z-axis is perpendicular to Q; but by definition of the Gauss map f of Q, the latter are exactly the points sent by f to the north or south poles of S^2. Hence the set of critical points of φ is $f^{-1}(y_0) \cup f^{-1}(y_0^*)$.

It is immediate from the proof of (6) above given in §26.2 of Part I, that at each point of $f^{-1}(y_0)$ the Jacobian of the Gauss map $f: Q \to S^2$ (in terms of suitable local co-ordinates on S^2) coincides with the determinant of the Hessian of the function $z = \varphi(x, y)$ (and hence with the Gaussian curvature). Since y_0 is a regular value of the map f it follows (from the definition (10.4.1) of non-degeneracy of critical points of a function) that the points of $f^{-1}(y_0)$ are indeed non-degenerate (critical) points of φ. (Alternatively this could have been inferred immediately from Lemma 11.2.1.) By applying the same argument with the z-axis reversed, we draw the same conclusion for the points in $f^{-1}(y_0^*)$, completing the proof of the lemma. □

In connexion with this proof note the obvious equality

$$\det\left(\frac{\partial^2 \varphi}{\partial u^i\, \partial u^j}\right) = (-1)^{n-1} \det\left(\frac{\partial^2(-\varphi)}{\partial u^i\, \partial u^j}\right), \tag{7}$$

where $u^1, \ldots, u^{n-1}$ are local co-ordinates in some neighbourhood of any point of the surface Q; this equality is crucial in the distinction arising between the cases of odd and even n. Since in our case $n-1=2$, it follows from (7) and the above proof that at every point of the set $f^{-1}(y_0) \cup f^{-1}(y_0^*)$ the Jacobian of the map f agrees in sign with the Gaussian curvature K (independently of the choice of the direction of the z-axis, i.e. of whether we work with φ or $-\varphi$). This, the above lemma, and the definition of the degree of a map, together yield immediately the following

14.3.3. Lemma. *With $f: Q \to S^2$ as above, we have*

$$2 \deg f = \sum_{P_j} (-1)^{\alpha(P_j)}, \tag{8}$$

where the summation is over the set of critical points P_j of the height function $z = \varphi$, and where $\alpha(P_j) = 0$ for those critical points where φ has a local maximum or minimum (where sgn $K = +1$*), and $\alpha(P_j) = 1$ for saddle points (where* sgn $K = -1$*).*

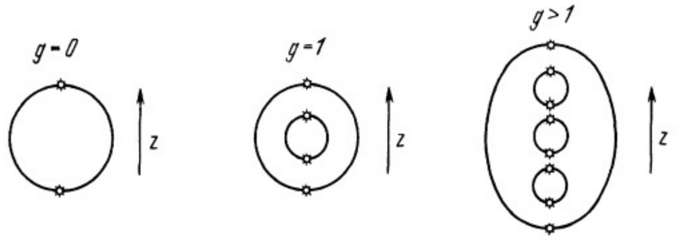

Figure 39

We shall now show that when Q is the sphere-with-g-handles, the number on the right-hand side of (8) is $2-2g$. It is easy to imagine this surface embedded in $\mathbb{R}^3$ in such a way that the height function φ has exactly one minimum point, one maximum point, and $2g$ saddle points (see Figure 39, where the critical points are indicated). For such an embedding we have by (8) together with Theorem 14.2.2 that

$$2 \deg f = \frac{1}{2\pi} \int_Q K \, d\sigma = 2-2g.$$

Of course other embeddings of the surface Q in $\mathbb{R}^3$ are possible. However at the conclusion of §37.4 of Part I we showed that the quantity $\int_Q K \, d\sigma$ is not altered by smooth variations in the metric on Q. Let $g_{ij}^{(0)}$ and $g_{ij}^{(1)}$ be any two Riemannian metrics on the surface Q, and consider the family of metrics

$$g_{ij}(t) = t g_{ij}^{(1)} + (1-t) g_{ij}^{(0)}, \qquad 0 \le t \le 1.$$

Clearly $g_{ij}(0) = g_{ij}^{(0)}$, $g_{ij}(1) = g_{ij}^{(1)}$, and $g_{ij}(t)$ is positive definite for all $t \in [0, 1]$. Hence the integral $\int_Q K \, d\sigma$ is the same for both metrics. We have thus proved (in more precise form than Corollary 37.4.3 of Part I) the following celebrated result.

14.3.4. Theorem (Gauss–Bonnet). *For the sphere-with-g-handles Q endowed with any Riemannian metric we have*

$$\frac{1}{2\pi} \int_Q K \, d\sigma = 2-2g.$$

14.4. The Index of a Singular Point of a Vector Field

In this subsection we shall study the Gauss map of a vector field in a neighbourhood of an "isolated singular point" of the vector field. We first need to define these terms. Thus let $\xi = \xi(x)$ be a vector field defined on some neighbourhood of a particular point x_0 of Euclidean space $\mathbb{R}^n$. In the conventional terminology, the point x_0 is called a *singular point of the field* ξ if $\xi(x_0) = 0$, and an *isolated* singular point of the field if it is a singular point but at all other points of some (small) neighbourhood of it $\xi(x)$ does not vanish. A singular point x_0 is said to be *non-degenerate* if

$$\det\left(\left.\frac{\partial\xi^\alpha}{\partial x^\beta}\right|_{x=x_0}\right)\neq 0.$$

14.4.1. Lemma. *A non-degenerate singular point of a vector-field is necessarily isolated.*

PROOF. Consider $\xi(x)$ as a map from $\mathbb{R}^n$ to $\mathbb{R}^n$. Since the derivative $\det(\partial\xi^\alpha/\partial x^\beta)\neq 0$ at the non-degenerate singular point x_0, the Inverse Function Theorem tells us that ξ is one-to-one on some neighbourhood of x_0. Hence x_0 is an isolated singular point, as claimed. □

By the *roots* of a non-degenerate singular point x_0 of a vector field ξ we shall mean the eigenvalues $\lambda_1, \ldots, \lambda_n$ of the matrix $(\partial\xi^\alpha/\partial x^\beta)_{x=x_0}$, and by the *index* of such a point the sign

$$\operatorname{sgn}\det\left(\left.\frac{\partial\xi^\alpha}{\partial x^\beta}\right|_{x=x_0}\right)=\operatorname{sgn}(\lambda_1\ldots\lambda_n).$$

Note that if the vector field ξ happens to be the gradient field of some function f, i.e. $\xi^\alpha=\partial f/\partial x^\alpha$, then the index of a non-degenerate singular point x_0 coincides with the sign of the determinant of the Hessian:

$$\operatorname{sgn}\det\left(\left.\frac{\partial\xi^\alpha}{\partial x^\beta}\right|_{x=x_0}\right)=\operatorname{sgn}\det\left(\left.\frac{\partial^2 f}{\partial x^\alpha\,\partial x^\beta}\right|_{x=x_0}\right)=(-1)^{i(x_0)},$$

where $i(x_0)$ denotes the number of negative squares occurring in the quadratic form $d^2f|_{x=x_0}$ when brought into canonical form.

Let $Q_\varepsilon\cong S^{n-1}$ be a sphere with centre an isolated singular point x_0 of a vector field ξ, and with radius $\varepsilon>0$ sufficiently small for the field ξ to be non-vanishing on Q_ε and its interior. We can then define, as in §14.2 above, the spherical Gauss map on Q_ε (relative to the field ξ):

$$f_{x_0}\colon Q_\varepsilon\to S^{n-1}.$$

14.4.2. Definition. The *index of an isolated singular point x_0 of a vector field ξ* is the degree of the Gauss map f_{x_0}:

$$\operatorname{ind}_{x_0}(\xi)=\deg f_{x_0}.$$

If the point x_0 is non-degenerate, then this definition of index is equivalent to the previous one; this is the import of the following result.

14.4.3. Theorem. *For a non-degenerate singular point x_0 of a vector field $\xi(x)$, we have*

$$\deg f_{x_0}=\operatorname{sgn}\det\left(\left.\frac{\partial\xi^\alpha}{\partial x^\beta}\right|_{x=x_0}\right),\tag{9}$$

where f_{x_0} is the Gauss map defined above.

Proof. In some neighbourhood of x_0 we have by Taylor's theorem that

$$\xi(x)=\xi^{(1)}(x)+\xi^{(2)}(x),$$

where $\xi^{(1)\beta}(x)=\partial\xi^\beta/\partial x^\gamma|_{x=x_0}(x^\gamma-x_0^\gamma)$, and $|\xi^{(2)}(x)|=o(|\xi^{(1)}(x)|)$. The homotopy $\xi(x,t)$ defined by

$$\xi(x,t)=\xi^{(1)}(x)+(1-t)\xi^{(2)}(x), \qquad 0\le t\le 1,$$

clearly has the properties that $\xi(x,0)=\xi(x)$, $\xi(x,1)=\xi^{(1)}(x)$, and in some sufficiently small neighbourhood of x_0, for all t, $0\le t\le 1$, $\xi(x,t)$ vanishes only at x_0. We have thus shown that in some neighbourhood of x_0 of the field ξ is smoothly homotopic to the linear vector field $\xi^{(1)}$ via fields for which x_0 is the only singular point (in the neighbourhood). Let ε be small enough for the sphere Q_ε with centre x_0 to be entirely contained in this neighbourhood. In the course of the above-defined deformation of ξ into $\xi^{(1)}$, the linear part of $\xi(x,t)$ remains fixed, and the map $f_{x_0}\colon Q_\varepsilon\to S^{n-1}$ undergoes a smooth homotopy; hence both sides of equation (9) remain constant throughout the homotopy, and it therefore suffices to verify (9) for the linear field $\xi^{(1)}$, and the corresponding map

$$f_{x_0}^{(1)}(x)=\frac{\xi^{(1)}(x)}{|\xi^{(1)}(x)|}.$$

Since $\xi^{(1)}$ is a non-singular linear transformation (from a neighbourhood of x_0 to a neighbourhood of the origin of $\mathbb{R}^n$), it follows that it is one-to-one, and that the map $f_{x_0}^{(1)}\colon Q_\varepsilon\to S^{n-1}$, is also one-to-one and without critical points. Hence each point y_0 on the sphere S^{n-1} is a regular value with exactly one preimage, and it is clear that the sign of the determinant of the linear transformation $\xi^{(1)}$ determines whether the orientation is preserved or reversed by $f_{x_0}^{(1)}$. This completes the proof. □

We now list the possible types of non-degenerate singular points in the simplest cases $n=2$ and 3.

(i) ($n=2$). The non-degenerate singular points of vector fields on the plane will be of one of the following types (note that in this case the index of a singular point is independent of the direction of the field):

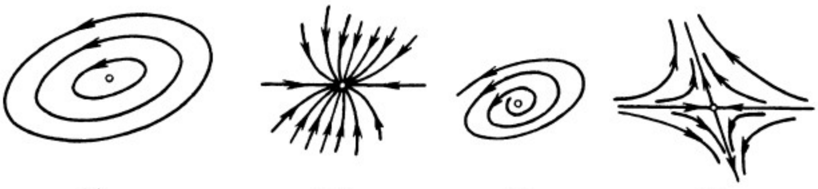

Figure 40

		Index
a centre (when both eigenvalues are purely imaginary)	Figure 40(a)	$+1$
a node (eigenvalues real and of the same sign)	Figure 40(b)	$+1$
a focal point (complex conjugate eigenvalues, not purely imaginary)	Figure 40(c)	$+1$
a saddle point (eigenvalues real and of opposite signs)	Figure 40(d)	-1

If the field happens to be the gradient of some two-variable function f then the possible singularities are restricted to the following types:

	Index
a local minimum point of f (Figure 40(b) with the arrows reversed—a "source")	$+1$
a saddle point of f (Figure 40(d))	-1
a local maximum point of f (Figure 40(b)—a "sink")	$+1$

(ii) ($n=3$). If the field is a gradient ($\xi^\alpha = \partial f/\partial x^\alpha$), a non-degenerate singular point is of one of the following types:

	Index
a local minimum point of f ("source")	$+1$
a saddle point of type 1 (when the form d^2f has only one negative square)	-1
a saddle point of type 2 (when the form d^2f has two negative squares)	$+1$
a local maximum of f ("sink")	-1

The classification of non-degenerate singular points of a general vector field in a region of $\mathbb{R}^3$ is as follows (note that either all three eigenvalues λ_i are real, or else one is real and the other two are complex conjugates):

	Index
a source (Re $\lambda_i \geq 0$, $i=1, 2, 3$)	$+1$
a saddle of type 1 (Re $\lambda_1 \geq 0$, Re $\lambda_2 \geq 0$, λ_3 real and negative)	-1
a saddle of type 2 (λ_1 real and positive, Re $\lambda_2 \leq 0$, Re $\lambda_3 \leq 0$)	$+1$
a sink (Re $\lambda_i \leq 0$, $i=1, 2, 3$)	-1

14.4.4. Theorem. *Let $\xi = \xi(x)$ be a vector field in $\mathbb{R}^n$ having only isolated singular points $x_1, \ldots, x_m$, and let Q be a closed oriented hypersurface in $\mathbb{R}^n$ avoiding the singular points of the vector field, and bounding a region D of $\mathbb{R}^n$. Then the degree of the vector field ξ on the surface Q, i.e. the degree of the Gauss map $Q \to S^{n-1}$ determined by the field, is equal to the sum of the indices of the singular points $x_{i_1}, \ldots, x_{i_k}$ lying in the region D.*

Proof. For each $j=1,\ldots,m$ let $Q_{j\varepsilon}$ be the sphere of radius $\varepsilon>0$ about the point x_j, choosing ε sufficiently small for the closed balls bounded by the $Q_{j\varepsilon}$ to miss the hypersurface Q and each other. The $\tilde{D}$ obtained from D by removing from it all open balls bounded by those spheres $Q_{j\varepsilon}$ in D, clearly has boundary

$$\partial\tilde{D}=Q\cup Q_{i_1\varepsilon}\cup\cdots\cup Q_{i_k\varepsilon}.$$

Consider now the form $f^*\Omega$ on $\tilde{D}$, where $f\colon\tilde{D}\to S^{n-1}$ is the Gauss map relative to ξ, and Ω is the form defined by the formula (3). Dimensional considerations yield immediately that $d\Omega=0$ on the sphere S^{n-1}, whence also $df^*\Omega=f^*(d\Omega)=0$ on S^{n-1}. Hence by the general Stokes formula (see §8.2) (and the additivity property of integrals) we have

$$0=\int_{\tilde{D}}df^*\Omega=\int_{\partial\tilde{D}}f^*\Omega=-\int_Q f^*\Omega+\sum_{q=1}^{k}\int_{Q_{i_q\varepsilon}}f^*\Omega,$$

where the difference in sign in the final expression is accounted for by the fact that in applying the general Stokes formula, the "external" component Q of the boundary $\partial\tilde{D}$ of $\tilde{D}$ is taken with orientation opposite to that on the boundaries of the $Q_{i_q\varepsilon}$. The theorem now follows from Definition 14.4.2 and Corollary 14.2.1. □

14.5. Transverse Surfaces of a Vector Field. The Poincaré–Bendixson Theorem

A situation of particular interest is that where the hypersurface Q is itself a sphere (of large radius) with the property that at no point of it is the vector field $\xi(x)$ tangential to it (or zero on it). We call a closed surface with the latter property a *transverse surface* to the vector field ξ. In this situation the following simple result holds.

14.5.1. Lemma. *The degree of a vector field ξ on a transverse hypersurface in $\mathbb{R}^n$ is (up to sign) equal to the (normalized) integral of the Gaussian curvature over the surface. (If the transverse surface is a sphere then this (normalized) integral is (in absolute value) unity.)*

Proof. A vector field which is transverse to a closed hypersurface $Q\subset\mathbb{R}^n$ is homotopic (in the class of fields on Q nowhere tangent to Q) to one or the other of the unit normal vector fields $\pm n(x)$ to Q, so that, degree being a homotopy invariant, we may assume our vector field on Q to be $\pm n(x)$. We know from Theorem 14.2.2 that the degree of the Gauss map (determined by $\pm n(x)$) is $\pm(1/\gamma_n)\int_Q K\,d\sigma$, where K is the Gaussian curvature on Q, $d\sigma$ is the volume element on Q, and γ_n is the volume of the unit $(n-1)$-sphere. If Q is a sphere then it is easy to calculate that $(1/\gamma_n)\int_Q K\,d\sigma=1$. This completes the proof of the lemma. □

they each have three components, say $A_a = A_a^0 e_0 + A_a^1 e_1 + A_a^2 e_2$, where e_0, e_1, e_2 form a canonical basis for L:

$$[e_0, e_1] = e_2, \qquad [e_1, e_2] = e_0, \qquad [e_2, e_0] = e_1, \tag{47}$$

chosen so that e_0 generates L_0, e_1, e_2 span L_1, and e_1, e_2 are of length 1 with respect to the scalar product $\langle\ ,\ \rangle_{L_1}$; thus

$$\langle e_0, e_0\rangle_{L_1} = 0; \qquad \langle e_1, e_1\rangle_{L_1} = \langle e_2, e_2\rangle_{L_1} = 1; \qquad \langle e_i, e_j\rangle_{L_1} = 0, \qquad i \neq j.$$

Regarding A as defining a trivial connexion on the principal G-bundle over $\mathbb{R}^2$, we deduce (as earlier) that $F_{ab} = 0$ (see e.g. Part I, §41.2), i.e.

$$\frac{\partial A_b^\beta}{\partial x^a} - \frac{\partial A_a^\beta}{\partial x^b} + [A_a, A_b]^\beta = 0, \qquad a, b = 1, 2; \quad \beta = 0, 1, 2. \tag{48}$$

Setting

$$\begin{aligned} &B_a = A_a^1 + iA_a^2, && \bar{B}_a = A_a^1 - iA_a^2, \\ &A_a^0 = if_a, && f_{ab} = \frac{\partial A_a^0}{\partial x^b} - \frac{\partial A_b^0}{\partial x^a}, \qquad a, b = 1, 2, \end{aligned} \tag{49}$$

we obtain from (48) by direct calculation (using (47)) that

$$f_{12} = \frac{i}{2}(B_1 \bar{B}_2 - B_2 \bar{B}_1), \tag{50}$$

and also

$$\frac{\partial B_1}{\partial x^2} - \frac{\partial B_2}{\partial x^1} = B_2 f_1 - B_1 f_2. \tag{51}$$

Defining a "covariant" differential operator $D = (D_1, D_2)$ in the usual way by (cf. §25.3 and Part I, §41.1)

$$D_a = \frac{\partial}{\partial x^a} + f_a,$$

(51) becomes

$$D_1 B_2 - D_2 B_1 = 0.$$

The Euler–Lagrange equations for the functional (45) turn out to have the form

$$\frac{\partial B_a}{\partial x^a} = 2f_a B_a, \qquad a = 1, 2. \tag{52}$$

Since the field $f_a = -iA_0^a$ is a gauge field (which, using (46), can be shown to have its transformation rule (under gauge transformations $g(x) \to e^{i\varphi(x)}g(x)$) of the purely gradient form $f_a \to f_a + \partial\varphi/\partial x^a$), and since $B = (B_a)$ can be shown (again via (46)) to transform according to the rule

$$B_a \to e^{i\varphi} B_a,$$

it follows that we have reduced our variational problem for n-fields to that of solving the equations (52) for the complex vector field (B_a) and the gauge field f_a, on whose stress tensor f_{ab} there is imposed the constraint

$$f_{12} = \frac{\partial f_1}{\partial x^2} - \frac{\partial f_2}{\partial x^1} = \frac{i}{2}(B_1 \bar{B}_2 - B_2 \bar{B}_1). \tag{53}$$

It can be shown that for any n-field n we have

$$n^*(\Omega) = \text{const.} \times f_{12}\, dx^1 \wedge dx^2, \tag{54}$$

where as before Ω is the $SO(3)$-invariant 2-form on S^2 given by the element of area on S^2, normalized so that $\int_{S^2} \Omega = 1$. (Thus

$$\Omega = \text{const.} \times \frac{dz \wedge d\bar{z}}{(1 + |z|^2)^2} \tag{55}$$

in terms of a suitable complex co-ordinate z *on* $S^2 \backslash \{\infty\}$ (cf. (41).) Hence the homotopy invariant d, the degree of the n-field, is also given (to within a constant factor) by the integral of the stress tensor:

$$\begin{aligned} d &= \int_{\mathbb{R}^2} n^*(\Omega) = \text{const.} \times \int_{\mathbb{R}^2} f_{12}\, dx^1 \wedge dx^2 \\ &= \text{const.} \times \oint_{\Gamma \cup \Gamma_\infty} (f_1\, dx^1 + f_2\, dx^2) = \lambda \int_{\mathbb{R}^2} (B_1 \bar{B}_2 - B_2 \bar{B}_1)\, dx^1 \wedge dx^2, \end{aligned} \tag{56}$$

where λ is a constant depending only on the degree d, and where to obtain the second equality we have used Stokes' formula (see Part I, §26.3), Γ_∞ denoting a circle in $\mathbb{R}^2$ of sufficiently large radius R and $\Gamma = \cup_i \Gamma_i$ consisting of appropriately oriented circles of small radius ε around the singular points x_i of the fields B or f. (Fields B, f with finitely many singular points in the "finite" part of S^2, i.e. in $\mathbb{R}^2$, must be admitted for consideration in order for the present formulation of the theory (of n-fields) to be valid, since in the present formalism the points x_i such that $n(x_i) = \infty$, will be singular for (B_a, f_a). A given map $n: S^2 \to S^2$ need not be globally covered by a map $g: S^2 \to S^3$, so that at one point (∞) of S^2 the corresponding cross-section (defining g) of the bundle $S^3 \to S^2$ may be many-valued, i.e. $g(x)$ may be many-valued at the points x_i in $n^{-1}(\infty)$, and consequently B or f will fail to be defined at these x_i.) (Note also that the form $f_1\, dx^1 + f_2\, dx^2$ appearing in (56) can be regarded as a connexion on the standard Hopf bundle $S^3 \to S^2$ with group $G \cong S^1$, and projection given by the natural map $SU(2) \to SU(2)/U(1)$; see §24.3, Example (a).)

We conclude the present discussion by finding an explicit expression for the B_a in terms of a holomorphic field n. Let w be the complex co-ordinate on $\mathbb{R}^2$ given by $w = x^1 + ix^2$; if the function $n = z(w)$ defined by a given n-field is holomorphic, then by (55) and Part I, Lemma 12.2.2,

$$n^*(\Omega) = \text{const.} \times \left|\frac{dz}{dw}\right|^2 \frac{dw \wedge d\bar{w}}{(1+|z(w)|^2)^2}. \tag{57}$$

Since

$$S = \int_{\mathbb{R}^2} \langle A, A\rangle_{L_1} d^2x = \int_{\mathbb{R}^2} (B_1\bar{B}_1 + B_2\bar{B}_2)\, dx^1 \wedge dx^2, \tag{58}$$

it follows (via 56) that

$$\begin{aligned} S + \frac{i}{\lambda}\int_{\mathbb{R}^2} n^*(\Omega) &= \int_{\mathbb{R}^2} (B_1\bar{B}_1 + B_2\bar{B}_2 + i(B_1\bar{B}_2 - B_2\bar{B}_1))\, dx^1 \wedge dx^2 \\ &= \int_{\mathbb{R}^2} |B_1 + iB_2|^2\, dx^1 \wedge dx^2 \geq 0. \end{aligned} \tag{59}$$

Since we are assuming $n = z(w)$ to be holomorphic, it follows (essentially from the result established earlier to the effect that the functional (58) attains the lower bound given by (59) precisely for the holomorphic n-fields of degree d) that

$$B_1 + iB_2 = 0. \tag{60}$$

Equations (53), (54), (57) and (60) together give

$$\begin{aligned} \left|\frac{dz}{dw}\right|^2 \frac{dw \wedge d\bar{w}}{(1+|z|^2)^2} &= \text{const.} \times (B_1\bar{B}_2 - B_2\bar{B}_1)\, dw \wedge d\bar{w} \\ &= \text{const.} \times B_1^2\, dw \wedge d\bar{w}, \end{aligned}$$

whence finally

$$B_1 = \frac{\text{const.}}{1+|z|^2}\frac{dz}{dw}. \tag{61}$$

We end this section with a proof of the following recent result: *The Lagrange–Euler equation $\delta S = 0$ for the extremals of the functional S given by* (40), *with the metric g_{ab} taken to be the Minkowski metric (so that the n-fields are defined on $\mathbb{R}_1^2$) rather than the Euclidean one, is equivalent to the "sine–Gordon equation"*. (The latter equation was derived in Part I, §30.4, as a necessary and sufficient condition for a surface of constant negative Gaussian curvature to be embeddable in Euclidean space $\mathbb{R}^3$.)

To prove this we first change from the usual Minkowski co-ordinates x, y of $\mathbb{R}_1^2$, in terms of which the metric has the form $ds^2 = dx^2 - dy^2$, to the co-ordinates $\eta = x + y$, $\xi = x - y$, in terms of which the metric is given by $ds^2 = d\eta\, d\xi$. If we write as before n^1, n^2, n^3 for the components of $n(x) \in \mathbb{R}^3 \supset S^2$, where $(n^1)^2 + (n^2)^2 + (n^3)^2 = 1$, the functional (40) (with g_{ab} now given by the Minkowski metric) becomes

$$S\{n\} = \int_{\mathbb{R}^2} \left(\sum_{\alpha=1}^{3} n_\xi^\alpha n_\eta^\alpha\right) d\xi\, d\eta. \tag{62}$$

The Degree of a Vector Field
on a Hypersurface.

Pág. 114 (130) –
14.3 The Whitney Number.
The Gauss-Bonnet Formula.
$\Rightarrow$ Gauss map

Pág. 118 (134)
14.3.4 Theorem (Gauss-Bonnet)

$$\frac{1}{2\pi}\int_Q K\, d\sigma = 2-2g.$$

...

Pág. 412 (428)
Lagrange-Euler equation
$\delta S = 0$

... surface of constant negative Gaussian curvature to be imbeddable in Euclidean space R^3.

Geometria Moderna;

- Geometría moderna ...(1987).pdf
  Dubrovin
- 0755. Geometría analítica.pdf
  Obs: Livro todo!!!

Pág. 127(137) —

Ángulo entre dos rectas

Los principios básicos de la geometría establecen que um ángulo exterior de um triángulo es igual a la suma de los ángulos interiores opuestos

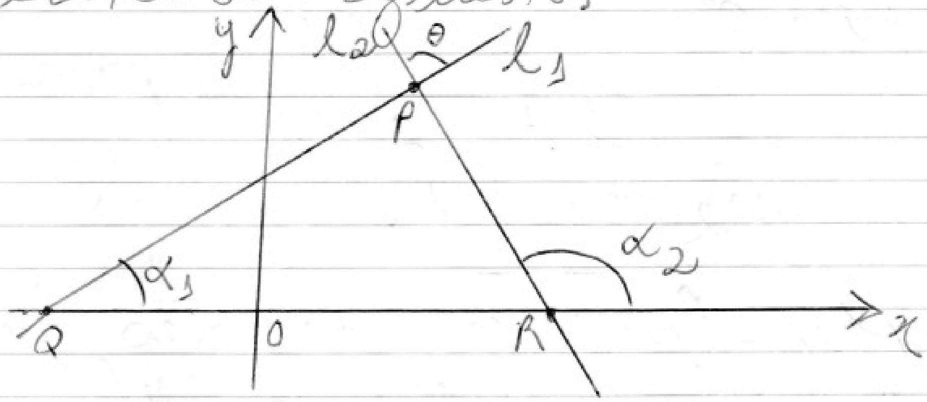

$$\alpha_2 = \alpha_1 + \theta$$

Ángulo entre las rectas es:

$$\theta = \alpha_2 - \alpha_1$$

Al aplicar la fórmula trigonométrica

B
! -256- - -* ! - B
^_`abcdecf`gha^dic`jhaklamno`dp_jqarlastp_udpavawlaxdecjudlyiz

D -256- - -* ! - B $$ 7 {! ! ! { ! 1"
(% | " !

Escribe los números correspondientes

Competencias genéricas ☐

Competencias disciplinares ☐

2. Un hexágono regular tiene su centro en el polo y dos lados paralelos al eje polar; si la longitud de un lado es de 3 unidades, determina el par principal de coordenadas polares de cada uno de sus seis vértices.
3. Dos de los vértices de un triángulo equilátero son $A(0,0°)$ y $B(4,0°)$, encuentra el par principal de coordenadas polares del tercer vértice.
4. Determina la ecuación polar de la recta paralela al eje polar que está situada a 5 unidades por debajo del eje.
5. Demuestra por coordenadas polares que los siguientes puntos $A(-1,1)$, $B(3,1)$ y $C(1,4.46)$ son los vértices de un triángulo equilátero y determina su área.
6. Demuestra por coordenadas polares que los siguientes puntos $A(-5,0)$, $B(0,2)$ y $C(0,-2)$ son los vértices de un triángulo isósceles y determina su área.
7. Demuestra por medio de distancias que los puntos dados $A(3,30°)$, $B\left(\frac{3\sqrt{3}}{2},60\right)$ y $C(3,90°)$ son colineales.
8. Encuentra el área, perímetro y semiperímetro para un triángulo cuyos vértices son el polo y los puntos $Q(2,60°)$ y $R(1,135°)$.
9. Demuestra que las ecuaciones polares de las rectas que son perpendiculares y paralelas al eje polar se expresan en la forma $r = \pm p \sec \theta$ y $r = \pm p \csc \theta$, respectivamente, donde p es la distancia del polo a la recta.
10. Determina la ecuación polar de la recta que pasa por el punto $A(5,60°)$ y forma un ángulo de 120° con el eje polar.
11. Encuentra la ecuación polar de la recta que pasa por el punto $Q(4,120°)$ y por el polo.
12. Determina la ecuación polar de la recta que pasa por el punto $M\left(2,\frac{2\pi}{3}\right)$ y es perpendicular a la recta que une el polo con dicho punto.
13 Encuentra la ecuación de la recta que pasa por el punto $K(7,20°)$ y forma un ángulo de 140° con el eje polar.
14. Encuentra la ecuación de la recta que pasa por el punto $R(4,120°)$ y es perpendicular a la que une al polo con dicho punto.
15. Dos de los vértices de un paralelogramo son los puntos $A(3,-80°)$ y $B\left(5,\frac{3\pi}{14}\right)$; si el punto de intersección de sus diagonales es el polo, encuentra las coordenadas de los otros dos vértices del paralelogramo.

Verifica tus resultados en la sección de respuestas correspondiente. . . . .

Ángulo entre dos rectas

Sean ℓ_1 y ℓ_2 dos rectas que se cortan en el punto P y que cortan al eje x en los puntos Q y R; las cuales forman con el eje x ángulos α_1 y α_2, de tal forma que sus pendientes son $m_1 = \tan \alpha_1$ y $m_2 = \tan \alpha_2$, donde α_1 y α_2 son diferentes de $\frac{\pi}{2}$.

Al cortarse, las rectas forman dos ángulos suplementarios, uno de ellos es el ángulo θ, cuya relación con las rectas se muestra en los siguientes casos.

Caso 1

Gráficamente, se tiene:

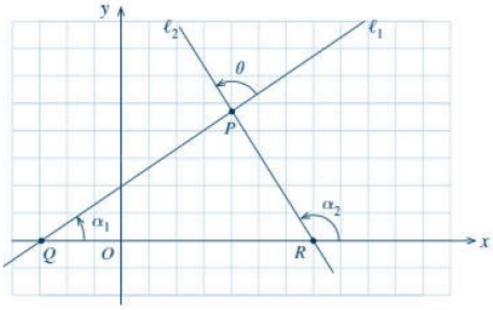
y
x
ℓ_1
ℓ_2
θ
P
α_1
α_2
Q
O
R

Los principios básicos de la geometría establecen que un ángulo exterior de un triángulo es igual a la suma de los dos ángulos interiores opuestos, es decir:

$$\alpha_2 = \alpha_1 + \theta$$

De tal forma, que el ángulo entre las rectas es:

$$\theta = \alpha_2 - \alpha_1$$

Al aplicar la fórmula trigonométrica: $\tan(\alpha_2 - \alpha_1) = \dfrac{\tan\alpha_2 - \tan\alpha_1}{1 + \tan\alpha_2 \tan\alpha_1}$, se tiene:

$$\tan\theta = \frac{\tan\alpha_2 - \tan\alpha_1}{1 \pm \tan\alpha_2 \tan\alpha_1}$$

Al sustituir $m_1 = \tan\alpha_1$ y $m_2 = \tan\alpha_2$ resulta:

$$\tan\theta = \frac{m_2 - m_1}{1 + m_2 m_1}, \text{ para } m_2 m_1 \neq -1.$$

Caso 2

Gráficamente, se tiene:

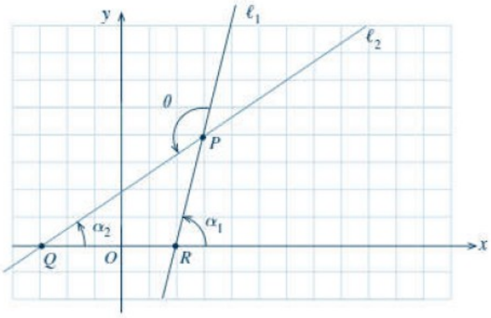
y
x
ℓ_1
ℓ_2
θ
P
α_2
α_1
Q
O
R

Por geometría elemental, se tiene que:

$$\begin{aligned} \theta &= \alpha_2 + (180^\circ - \alpha_1) \\ -180^\circ + \theta &= \alpha_2 - \alpha_1 \\ 180^\circ - \theta &= \alpha_1 - \alpha_2 \end{aligned}$$

Al aplicar la misma fórmula trigonométrica del caso anterior para las tangentes, resulta:

$$\tan(180^\circ - \theta) = \frac{\tan\alpha_1 - \tan\alpha_2}{1 + \tan\alpha_1 \tan\alpha_2}$$

Por la fórmula trigonométrica: $\tan(180^\circ - \theta) = -\tan\theta$, se tiene:

$$-\tan\theta = \frac{\tan\alpha_1 - \tan\alpha_2}{1 + \tan\alpha_1 \tan\alpha_2} \qquad \tan\theta = \frac{\tan\alpha_2 - \tan\alpha_1}{1 + \tan\alpha_1 \tan\alpha_2}$$

Al sustituir $m_1 = \tan\alpha_1$ y $m_2 = \tan\alpha_2$, resulta:

$$\tan\theta = \frac{m_2 - m_1}{1 + m_1 m_2}, \text{ para } m_1 m_2 \neq -1.$$

Teorema

Sean ℓ_1 y ℓ_2 dos rectas que tienen pendientes m_1 y m_2, respectivamente; si θ es el ángulo dirigido de ℓ_1 a ℓ_2, se establece que:

$$\tan\theta = \frac{m_2 - m_1}{1 + m_1 m_2}, \text{ para } m_1 m_2 \neq -1.$$

Para saber cuál de los dos ángulos se va determinar (Caso I o Caso II) se debe considerar que el ángulo se mide girando en el sentido contrario a las manecillas del reloj, por lo que la recta ℓ_1 será la que corresponde al **lado inicial** del ángulo y la recta ℓ_2 la que corresponde al **lado final**.

Deducción de las condiciones de paralelismo y perpendicularidad de dos rectas

Cuando dos rectas son paralelas, el ángulo que se forma entre ellas es de 0° o de 180°. Al aplicar la fórmula para determinar el ángulo de dos rectas, resulta:

Para $\theta = 0^\circ$

$$\begin{aligned} \tan\theta &= \frac{m_2 - m_1}{1 + m_1 m_2} \\ \tan 0^\circ &= \frac{m_2 - m_1}{1 + m_1 m_2} \\ 0 &= \frac{m_2 - m_1}{1 + m_1 m_2} \end{aligned}$$

$$m_2 - m_1 = 0$$

$$m_2 = m_1$$

Para $\theta = 180^\circ$

$$\begin{aligned} \tan\theta &= \frac{m_2 - m_1}{1 + m_1 m_2} \\ \tan 180^\circ &= \frac{m_2 - m_1}{1 + m_1 m_2} \\ 0 &= \frac{m_2 - m_1}{1 + m_1 m_2} \end{aligned}$$

$$m_2 - m_1 = 0$$

$$m_2 = m_1$$

Por el carácter simétrico o recíproco de la igualdad, se tiene que $m_1 = m_2$. Por consecuencia, se establece: **si dos rectas son paralelas, sus pendientes son iguales**.

Cuando dos rectas son perpendiculares, el ángulo que se forma entre ellas es de 90°, entonces, al aplicar la fórmula para determinar el ángulo de dos rectas, resulta:

$$\tan\theta = \frac{m_2 - m_1}{1 + m_1 m_2}$$

$$\tan 90^\circ = \frac{m_2 - m_1}{1 + m_1 m_2}$$

Como $\tan 90^\circ = \infty$, la ecuación anterior no apoya el principio que se desea demostrar, por lo que es necesario emplear la siguiente identidad:

$$\cot\theta = \frac{1}{\tan\theta}$$

Al sustituir en la expresión anterior, resulta:

$$\cot\theta = \frac{1}{\dfrac{m_2 - m_1}{1 + m_1 m_2}}$$

$$\cot 90^\circ = \frac{1 + m_1 m_2}{m_2 - \mathbf{m}_1}$$

$$0 = \frac{1 + m_1 m_2}{m_2 - m_1}$$

$$1 + m_1 m_2 = 0$$

$$m_1 m_2 = -1$$

Por consecuencia, se establece: **dos rectas son perpendiculares entre sí, cuando el producto de sus pendientes es igual a -1.**

También de la ecuación $m_1 m_2 = -1$, se tiene:

$$m_1 = -\frac{1}{m_2} \qquad m_2 = -\frac{1}{m_1}$$

Lo que da lugar al siguiente criterio equivalente: *dos rectas son perpendiculares si sus pendientes son recíprocas y de signos contrarios.*

EJEMPLOS

Ejemplos

1 Determina los ángulos interiores del triángulo cuyos vértices son los puntos $A(-2,1)$, $B(3,4)$ y $C(5,-2)$ y comprueba los resultados.

Solución

Al elaborar la gráfica de los puntos dados, se tiene:

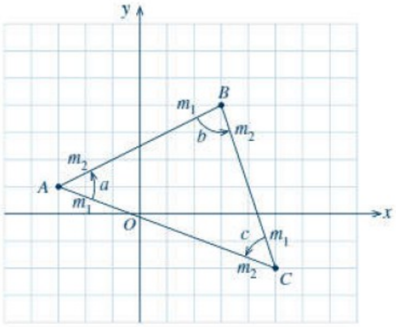

Se determinan las pendientes de los lados del triángulo, es decir:

$$m\,\overline{AB}=\frac{y_A-y_B}{x_A-x_B}=\frac{1-4}{-2-3}=\frac{-3}{-5}=\frac{3}{5}$$

$$m\,\overline{BC}=\frac{y_B-y_C}{x_B-x_C}=\frac{4+2}{3-5}=\frac{6}{-2}=-3$$

$$m\,\overline{AC}=\frac{y_A-y_C}{x_A-x_C}=\frac{1+2}{-2-5}=\frac{3}{-7}=-\frac{3}{7}$$

Al aplicar la fórmula $\tan\theta=\dfrac{m_2-m_1}{1+m_1m_2}$, resulta:

Para el $\sphericalangle a$

$$\tan a=\frac{m_{\overline{AB}}-m_{\overline{AC}}}{1+m_{\overline{AC}}m_{\overline{AB}}}$$

$$\tan a=\frac{\frac{3}{5}-\left(-\frac{3}{7}\right)}{1+\left(-\frac{3}{7}\right)\left(\frac{3}{5}\right)}$$

$$\tan a=\frac{\frac{21+15}{\cancel{35}}}{\frac{35-9}{\cancel{35}}}$$

$$\tan a=\frac{36}{26}=1.3846$$

$$\sphericalangle a=\arctan(1.3846)$$

$$\sphericalangle a=54°09'44''$$

Para el $\sphericalangle b$

$$\tan b=\frac{m_{\overline{BC}}-m_{\overline{AB}}}{1+m_{\overline{AB}}m_{\overline{BC}}}$$

$$\tan b=\frac{-3-\frac{3}{5}}{1+\left(\frac{3}{5}\right)(-3)}$$

$$\tan b=\frac{\frac{-18}{\cancel{5}}}{\frac{5-9}{\cancel{5}}}$$

$$\tan b=\frac{-18}{-4}=4.5$$

$$\sphericalangle \boldsymbol{b}=\arctan(4.5)$$

$$\sphericalangle b=77°28'16''$$

Para el $\sphericalangle c$

$$\tan c=\frac{m_{\overline{AC}}-m_{\overline{BC}}}{1+m_{\overline{BC}}m_{\overline{AC}}}$$

$$\tan c=\frac{-\frac{3}{7}-(-3)}{1+(-3)\left(-\frac{3}{7}\right)}$$

$$\tan c=\frac{\frac{-3+21}{\cancel{7}}}{\frac{7+9}{\cancel{7}}}$$

$$\tan c=\frac{18}{16}=1.125$$

$$\sphericalangle c=\arctan(1.125)$$

$$\sphericalangle c=48°21'59''$$

Como los ángulos interiores de todo triángulo suman 180°, se tiene una aproximación por redondeo:

$$\begin{aligned} \sphericalangle a + \sphericalangle b + \sphericalangle c &= 180^\circ \\ 54^\circ 09' 44'' + 77^\circ 28' 16'' + 48^\circ 21' 59'' &= 180^\circ \\ 179^\circ 59' 59'' &\approx 180^\circ \end{aligned}$$

2 ●●· Dos rectas se cortan formando un ángulo de 45°, la recta inicial pasa por los puntos $A(-2,1)$ y $B(9,7)$ y la recta final pasa por el punto $P(3,9)$ y por el punto Q, cuya abscisa es -2; determina la ordenada de Q.

Solución

De acuerdo con los datos dados, es necesario determinar las pendientes de las rectas inicial y final, dado que se conoce el ángulo entre las dos rectas, se tiene:

$$m_i = \frac{y_A - y_B}{x_A - x_B} = \frac{1-7}{-2-9} = \frac{-6}{-11} = \frac{6}{11}$$

$$m_f = \frac{y_P - y_Q}{x_P - x_Q} = \frac{9-y}{3+2} = \frac{9-y}{5}$$

Al sustituir en la fórmula del ángulo de dos rectas, resulta:

$$\tan\theta = \frac{m_f - m_i}{1 + m_i m_f}$$

$$\tan 45^\circ = \frac{\frac{9-y}{5} - \frac{6}{11}}{1 + \left(\frac{6}{11}\right)\left(\frac{9-y}{5}\right)}$$

$$\tan 45^\circ = \frac{\frac{11(9-y)-(5)(6)}{55}}{1 + \frac{54-6y}{55}}$$

$$\tan 45^\circ = \frac{\frac{99-11y-30}{\cancel{55}}}{\frac{55+54-6y}{\cancel{55}}}$$

$$1 = \frac{69-11y}{109-6y}$$

$$\begin{aligned} 1(109-6y) &= 69-11y \\ 109-6y+11y &= 69 \\ 5y &= 69-109 \\ y &= \frac{-40}{5} \\ y &= -8 \end{aligned}$$

La ordenada del punto Q es -8.

Gráficamente, se tiene:

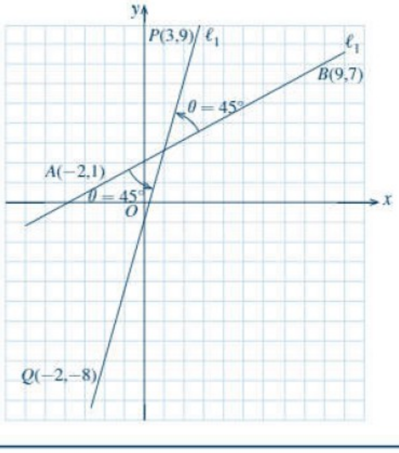

Posiciones relativas de dos rectas

A continuación discutiremos las condiciones analíticas de las posiciones relativas de dos rectas, cuyas ecuaciones se expresan en las formas generales $Ax + By + C = 0$ y $A'x + B'y + C' = 0$, bajo las cuales estas dos rectas son:

1. **Paralelas**. De la recta $Ax + By + C = 0$, su pendiente es $m = -\frac{A}{B}$, para $B \neq 0$; para la recta $A'x + B'y + C' = 0$, su pendiente es $m' = -\frac{A}{B}$, para $B' \neq 0$.

   Con base en el principio que establece que **si dos rectas son paralelas, sus pendientes son iguales**, se tiene que: $-\frac{A}{B} = -\frac{A'}{B'}$, o bien, $\frac{A}{B} = \frac{A'}{B'}$. Esta expresión puede escribirse como:

   $$AB' = A'B$$

   Al dividir la igualdad entre $A'B'$, resulta:

   $$\frac{A\cancel{B'}}{A'\cancel{B'}} = \frac{\cancel{A'}B}{\cancel{A'}B'}$$

   $$\frac{A}{A'} = \frac{B}{B'} \quad \Big\} \text{ Esto indica que los coeficientes de } x \text{ y } y \text{ deben ser proporcionales.}$$

   La expresión anterior sólo se aplica cuando $A' \neq 0$ y $B' \neq 0$. Si $A' = 0$, la recta es paralela al eje x, de esta forma, ambas rectas serán horizontales; cuando $B' = 0$, la recta es parelala al eje y, entonces, ambas rectas serán verticales.

   Entonces, la relación equivalente $AB' = A'B$ es verdadera para cualquier caso.

2. **Perpendiculares**. Con base en el principio que establece que **si dos rectas son perpendiculares entre sí, el producto de sus pendientes es igual a -1**, se tiene que:

   $$m = -\frac{A}{B} \qquad\qquad m' = -\frac{A'}{B'}$$

   De tal forma que:

   $$\left(-\frac{A}{B}\right)\left(-\frac{A'}{B'}\right) = -1$$

   Lo anterior da lugar a la relación $AA' + BB' = 0$, sea verdadera para cualquier caso, siempre que las rectas sean perpendiculares.

3. **Coinciden**. Cuando dos rectas tienen todos los puntos comunes y la misma pendiente, se dice que son coincidentes.

   Dadas las rectas $Ax + By + C = 0$ y $A'x + B'y + C = 0$, su intersección con el eje x, será el punto de abscisa, es decir:

   $$x = -\frac{C}{A} \quad x = -\frac{C'}{A'}$$

   Por lo anterior, se establece que $-\frac{C}{A} = -\frac{C'}{A'}$, lo que da lugar a la siguiente relación:

   $$\frac{A}{A'} = \frac{C}{C'}$$

Como las pendientes son iguales, se tiene que $-\frac{A}{B}=-\frac{A'}{B'}$, o la relación: $\frac{A}{A'}=\frac{B}{B'}$. Al relacionar ambos criterios, queda: $\frac{A}{A'}=\frac{B}{B'}=\frac{C}{C'}$, es decir, dos rectas son coincidentes, si sus coeficientes correspondientes son proporcionales.

Lo anterior es verdadero sólo si todos los coeficientes son diferentes de cero. Las relaciones equivalentes $A = kA'$, $B = kB'$ y $C = kC'$, para alguna constante k, caracterizan a dos rectas coincidentes cualquier caso.

4. **Se cortan en uno y solamente un punto**. Dos rectas no paralelas, por geometría, se cortan en uno y solamente en un punto; analíticamente, se deduce que sus pendientes no son iguales, por lo que sus coeficientes no son proporcionales, es decir:

$$\frac{A}{A'}\neq\frac{B}{B'},\ \text{o bien, } AB'-A'B\neq 0.$$

EJEMPLOS

Ejemplos

1 La ecuación de una recta es $3x - 2y - 6 = 0$; determina la ecuación que representa a todas las rectas paralelas a la recta dada. A partir de esta última ecuación encuentra la ecuación de la recta paralela a la recta dada y que pasa por el punto $A(-3,5)$.

Solución

La ecuación de la recta que representa a todas las rectas paralelas a la ecuación $3x - 2y - 6 = 0$, es:

$$Ax + By + C = 0$$

La condición de paralelismo establece que:

$$\frac{A}{A'}=\frac{B}{B'}\ \text{es decir, } \frac{A}{3}=\frac{B}{-2}$$

$$B=-\frac{2A}{3}$$

Al sustituir en la ecuación $Ax + By + C = 0$, se tiene:

$$Ax+\left(-\frac{2A}{3}\right)y+C=0$$

$$3Ax - 2Ay + 3C = 0$$

Al dividir entre A, resulta:

$$\frac{3\cancel{A}x}{\cancel{A}}-\frac{2\cancel{A}y}{\cancel{A}}+\frac{3C}{A}=0$$

Al hacer $k=\frac{3C}{A}$, se tiene:

$$3x-2y+\frac{3C}{A}=0$$
$$3x-2y+k=0$$

Ecuación que representa a todas las rectas paralelas a $3x - 2y - 6 = 0$

Específicamente, la recta paralela a la recta dada debe pasar por el punto $A(-3{,}5)$, es decir:

$$\begin{aligned} 3x-2y+k&=0\\ 3(-3)-2(5)+k&=0\\ -9-10+k&=0\\ k&=19 \end{aligned}$$

La ecuación de la recta que pasa por el punto $A(-3{,}5)$ y que es paralela a la recta dada es $3x-2y+19=0$.

Gráficamente, se tiene:

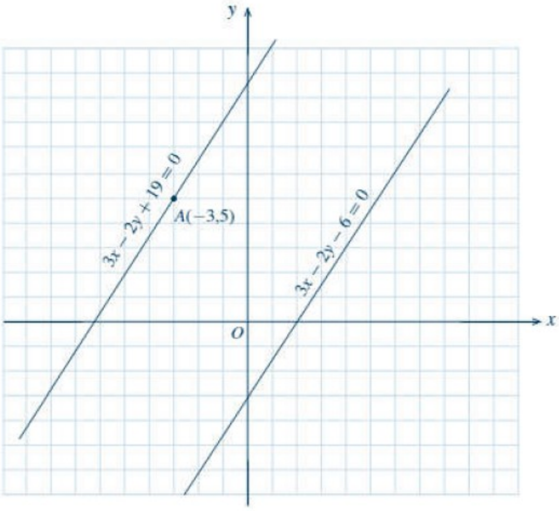

2 ●●• Demuestra que las rectas $3x-15y-21=0$ y $x-5y-7=0$, son coincidentes.

Solución

La condición que establece la coincidencia, indica que dos rectas tienen todos los puntos comunes y la misma pendiente, es decir:

$$m_1=-\frac{A}{B}=-\frac{3}{-15}=\frac{1}{5} \qquad m_2=-\frac{A}{B}=-\frac{1}{-5}=\frac{1}{5}$$

Como las pendientes son iguales, se debe cumplir la siguiente relación:

$$\frac{A}{A'}=\frac{B}{B'}$$

Al sustituir los coeficientes de las ecuaciones en la relación, se tiene:

$$\frac{3}{1}=\frac{-15}{-5}$$
$$3=3$$

La intersección de las rectas dadas con el eje x son:

$$x_1 = -\frac{C}{A} = -\frac{-21}{3} = 7 \qquad x_2 = -\frac{-7}{1} = 7$$

Como ambas rectas tienen el mismo punto en común, se debe cumplir la siguiente relación:

$$\frac{A}{A'} = \frac{C}{C'}$$

Al sustituir los coeficientes de las ecuaciones en la relación, se tiene

$$\frac{3}{1} = \frac{-21}{-7}$$
$$3 = 3$$

El principio de rectas coincidentes establece que:

$$\frac{A}{A'} = \frac{B}{B'} = \frac{C}{C'}$$

Al sustituir los coeficientes de las ecuaciones en la relación, se tiene:

$$\frac{3}{1} = \frac{-15}{-5} = \frac{-21}{-7}$$
$$3 = 3 = 3$$

Por lo tanto, se demuestra que las dos rectas dadas son coincidentes.

Gráficamente, se tiene:

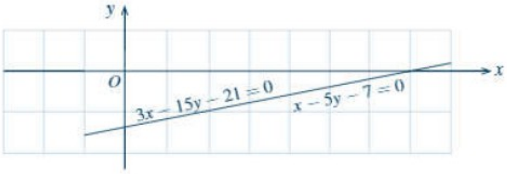

3 ●•·Demuestra que las rectas $3x + 4y - 12 = 0$ y $5x + 12y - 60 = 0$, se cortan en uno y solamente en un punto; determina las coordenadas del punto de intersección.

Solución

La condición que establece la intersección en uno y solamente un punto, indica que las dos rectas no tienen pendientes iguales y sus coeficientes no son proporcionales, es decir:

$$m_1 = -\frac{A}{B} = -\frac{3}{4} \quad m_2 = -\frac{A}{B} = -\frac{5}{12} \quad \therefore \quad \begin{matrix} m_1 \neq m_2 \\ -\frac{3}{4} \neq -\frac{5}{12} \end{matrix}$$

Como las pendientes no son iguales, se debe cumplir la siguiente relación:

$$\frac{A}{A'} \neq \frac{B}{B'} \qquad \text{o} \qquad AB' - A'B \neq 0$$

Al sustituir los coeficientes de las ecuaciones en la relación, se tiene:

$$\frac{3}{5} \neq \frac{4}{12} \qquad \text{o} \qquad (3)(12) - (5)(4) \neq 0$$

$$0.666\ldots \neq 0.333\ldots \qquad 16 \neq 0$$

Por lo anterior, se demuestra que las dos rectas dadas
se cortan en uno y solamente en un punto.

Al resolver el sistema de ecuaciones dadas, se tiene:

$$3x + 4y - 12 = 0 \quad \textbf{(1)}$$
$$5x + 12y - 60 = 0 \quad \textbf{(2)}$$

Al multiplicar la ecuación (**1**) por -3, resulta:

$$-3(3x + 4y - 12) = 0 \longrightarrow \begin{array}{r} -9x - \cancel{12y} + 36 = 0 \\ 5x + \cancel{12y} - 60 = 0 \\ \hline -4x - 24 = 0 \\ -4x = 24 \\ x = \dfrac{24}{-4} \\ x = -6 \end{array}$$

Al sustituir el valor de x en la ecuación (**1**):

$$\begin{array}{r} 5(-6) + 12y - 60 = 0 \\ -30 + 12y - 60 = 0 \\ 12y - 90 = 0 \\ 12y = 90 \\ y = \dfrac{90}{12} \\ y = 7.5 \end{array}$$

Las cordenadas del punto de intersección de las rectas dadas son $(-6{,}7.5)$.

Al elaborar la gráfica correspondiente, se tiene:

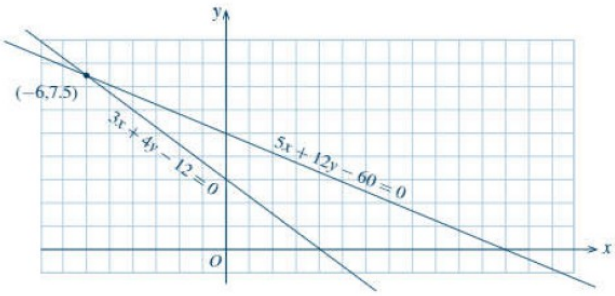

$\tan(\alpha_2 - \alpha_1) = \dfrac{\tan\alpha_2 - \tan\alpha_1}{1 + \tan\alpha_2 \tan\alpha_1}$, se tiene:

$$\tan\theta = \frac{\tan\alpha_2 - \tan\alpha_1}{1 \pm \tan\alpha_2 \tan\alpha_1}$$

Al substituir $m_1 = \tan\alpha_1$ y
$m_2 = \tan\alpha_2$
resulta:

$$\tan\theta = \frac{m_2 - m_1}{1 + m_2 m_1},$$ para

$m_2 m_1 \neq 1$.

— Mir_Problemas_de_Geometria_Analitica_Kletenik.pdf

— Mir_Problemas_de_Geometría_Diferencial_Fedenko.pdf

Obs: Livro todo!!!

Curvas and Superficies.

- 20211010/ Géométrie Contemporaine.../
Map-Theory/
- 9781599731346.pdf ⌉ same
- maptheory.pdf ⌋ book.

Pág. 09(24) - Fig 1.3
Sphere and projective plane
Obs. livro todo!!!

- /Map-theory/
1931233144.pdf
Smarandache Multi-Space Theory
Pág. 173(183) Chapter 4
Planar Map Geometries

Important!
Utilizar este livro com estes dois livros posteriores:

Source: 20211010/ Géométrie Contemporaine.../

B
! -256- - -* ! - B / $// % %'B 1"
! -256- - -* ! - B !19 3 1"
*D !1 9 3
(/ ' < % 1 ! " 1 K A 1 !
*D ? A " UUU

Because the system of equations obtained by deleting all the equations for all the edges on the circuit from (1.6) is equivalent to the original system of equations (1.6), in virtue of the finite recursion principle a solution of (1.6) can always be extracted . $\square$

When $n = 2$, this theorem has a variety of applications. In [Liu5], some applications can be seen. Further, its extension on a nonAbelian group can also be done while the system of equations are not yet linear but quadratic.

I.2 Surfaces

In topology, a surface is a compact 2-dimensional manifold without boundary. In fact, it can be seen as what is obtained by identifying each pair of edges on a polygon of even edges pairwise.

For example, in Fig.1.3, two ends of each dot line are the same point. The left is a sphere, or the plane when the infinity is seen as a point. The right is the projective plane. From the symmetry of the sphere, a surface can also seen as a sphere cutting off a polygon with pairwise edges identified.

The two surfaces in Fig.1.3 are formed by a polygon of two edges pairwise as a.

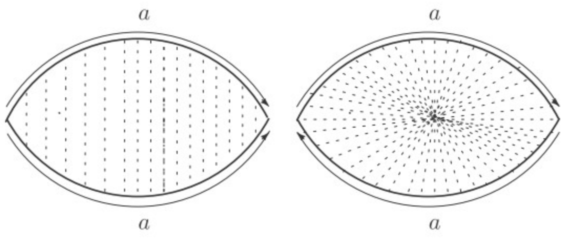

Sphere(Plane) Projective plane

Fig.1.3 Sphere and projective plane

Surface closed curve axiom A closed curve on a surface has one of the two possibilities: one side and two sides.

A curve with two sides is called a *double side curve* ; otherwise, a *single side curve* . As shown in Fig.1.3, any closed curve on a sphere is a double side curve(In fact, this is the Jordan curve axiom). However, it is different from the sphere for the projective plane. there are both a single(shown by a dot line) and a double side curve.

How do we justify whether a closed curve on a surface is of single side, or not?

In order to answer this question, the concept of contractibility of a curve has to be clarified. If a closed curve on a surface can be continuously contracted along one side into a point, then it is said to be *contractible*, or *homotopic* to 0.

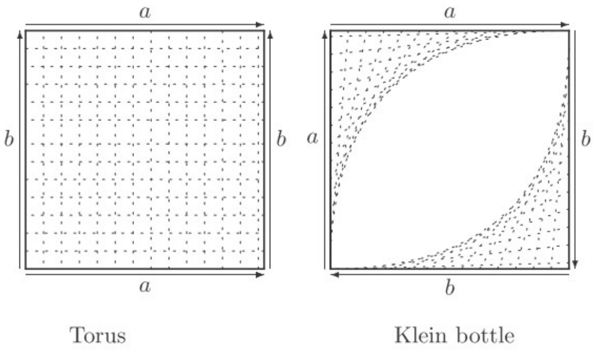

Fig.1.4 Torus and Klein bottle

It is seen that a single side curve is never homotopic to 0 and a double side curve is not always homotopic to 0. For example, in Fig.1.4, the left, *i.e.*, the torus, each of the dot lines is of double side but not contractible. The right, *i.e.*, the Klein bottle, all the dot lines are of single side , and hence, none of them is contractible.

A surface with all closed curves of double side is called *orientable*; otherwise, *nonorientable* .

For example, in Fig.1.3, the sphere is orientable and the projective plane is nonorientable. In Fig.1.4, the torus is orientable and the Klein bottle is nonorientable.

The maximum number of closed curves cutting along without destroying the continuity on a surface is called the *pregenus* of the surface.

In view of Jordan curve axiom, there is no such closed curve on the sphere. Thus, the pregenus of sphere is 0. On the projective plane, only one such curve is available (each of dot lines is such a closed curve in Fig.1.3) and hence the pregenus of projective plane is 1.

Similarly, the pregenera of torus and Klein bottle are both 2 as shown in Fig.1.4.

Theorem 1.5 The pregenus of an orientable surface is a nonnegative even number.

A formal proof can not be done until Chapter 5. Based on this theorem, the genus of an orientable surface can be defined to be half its pregenus, called the *orientable genus*. The genus of a nonorientable surface, called *nonorientable genus*, is its pregenus itself.

The sphere is written as aa^{-1} where a^{-1} is with the opposite direction of a on the boundary of the polygon. Thus, the projective plane, torus and Klein bottle are, respectively, aa, $aba^{-1}b^{-1}$ and $aabb$. In general,

$$\begin{aligned} O_p &= \prod_{i=1}^{p} a_i b_i a_i^{-1} b_i^{-1} \\ &= a_1 b_1 a_1^{-1} b_1^{-1} a_2 b_2 a_2^{-1} b_2^{-1} \cdots a_p b_p a_p^{-1} b_p^{-1} \end{aligned} \tag{1.8}$$

and

$$Q_q = \prod_{i=1}^{q} a_i a_i = a_1 a_1 a_2 a_2 \cdots a_q a_q \tag{1.9}$$

denote, respectively, a surface of orientable genus p and a surface of nonorientable genus q. Of course, O_0, Q_1, O_1 and Q_2 are, respectively, the sphere, projective plane, torus and Klein bottle.

It is easily checked that whenever an even polygon is with a pair of its edges in the same direction, the polygon represents a nonorientable surface. Thus, all O_p, $p \geq 0$, orientable and all Q_q, $q \geq 1$, are nonorientable.

Forms (1.8) and (1.9) are said to be *standard*.

If the *form* of a surface is defined by its orientability and its genus, then the operations 1–3 and their inverses shown as in Fig.1.5–7, do not change the surface form.

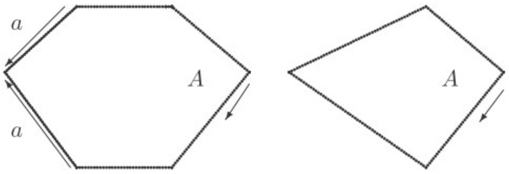

Fig.1.5 Operation 1: $Aaa^{-1} \Longleftrightarrow A$

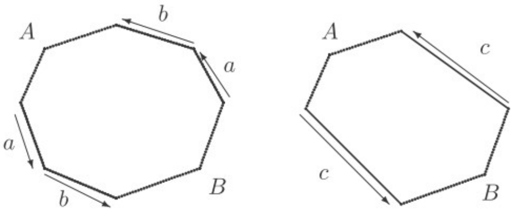

Fig.1.6 Operation 2: $AabBab \Longleftrightarrow AcBc$

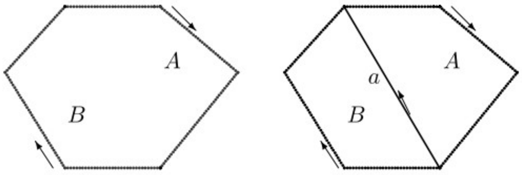

Fig.1.7 Operation 3: $AB \Longleftrightarrow (Aa)(a^{-1}B)$

In fact, what is determined under these operations is just a topological equivalence, denoted by $\sim_{\text{top}}$.

Notice that A and B are all linear order of letters and permitted to be empty as degenerate case in these operations.

The parentheses stand for cyclic order when more than one cyclic orders occur for distinguishing from one to another.

Relation 0 On a surface (A, B), if A is a surface itself then $(A, B) = ((A)x(B)x^{-1}) = ((A)(B))$.

Relation 1 $(AxByCx^{-1}Dy^{-1}) \sim_{\text{top}} ((ADCB)(xyx^{-1}y^{-1}))$.

Relation 2 $(AxBx) \sim_{\text{top}} ((AB^{-1})(xx))$.

Relation 3 $(Axxyzy^{-1}z^{-1}) \sim_{\text{top}} ((A)(xx)(yy)(zz))$.

In the four relations, A, B, C, and D are permitted to be empty. $B^{-1} = b_s^{-1} \cdots b_3^{-1}b_2^{-1}b_1^{-1}$ is also called the inverse of $B = b_1b_2b_3 \cdots b_s$, $s \geq 1$. Parentheses are always omitted when unnecessary to distinguish cyclic or linear order.

On a surface S, the operation of cutting off a quadrangle $aba^{-1}b^{-1}$ and then identifying each pair of edges with the same letter is called a *handle* as shown in the left of Fig.1.8.

If the quadrangle $aba^{-1}b^{-1}$ is replaced by aa, then such an operation is called a *crosscap* as shown in the right of Fig.1.8.

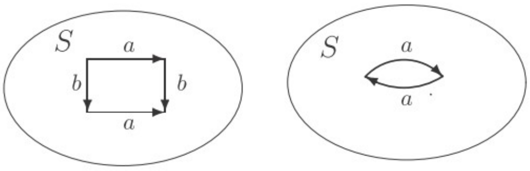

Handle Crosscap

Fig.1.8 Handle and crosscap

The following theorem shows the result of doing a handle on an orientable surface.

Theorem 1.6 What is obtained by doing a handle on an orientable surface is still orientable with its genus 1 added.

Proof Suppose S is the surface obtained, then

$$\begin{aligned}S &\sim_{\text{top}} \prod_{i=1}^{p} a_i b_i a_i^{-1} b_i^{-1} x a_{p+1} b_{p+1} a_{p+1}^{-1} b_{p+1}^{-1} x^{-1} \text{ (Relation 0)}\\ &\sim_{\text{top}} \prod_{i=1}^{p} a_i b_i a_i^{-1} b_i^{-1} x x^{-1} a_{p+1} b_{p+1} a_{p+1}^{-1} b_{p+1}^{-1} \text{ (Relation 1)}\\ &\sim_{\text{top}} \prod_{i=1}^{p} a_i b_i a_i^{-1} b_i^{-1} a_{p+1} b_{p+1} a_{p+1}^{-1} b_{p+1}^{-1} \text{ (Operation 1)}\\ &= \prod_{i=1}^{p+1} a_i b_i a_i^{-1} b_i^{-1}.\end{aligned}$$

This is the theorem. □

In the above proof, x and x^{-1} are a line connecting the two boundaries to represent the surface as a polygon shown in Fig.1.8. This procedure can be seen as the degenerate case of operation 3.

In what follows, observe the result by doing a crosscap on an orientable surface.

Theorem 1.7 On an orientable surface of genus p, $p \geq 0$, what is obtained by doing a crosscap is nonorientable with its genus $2p+1$.

Proof Suppose N is the surface obtained, then

$$\begin{aligned}N &\sim_{\text{top}} \prod_{i=1}^{p} a_i b_i a_i^{-1} b_i^{-1} x a a x^{-1} \text{ (Relation 0)}\\ &\sim_{\text{top}} \prod_{i=1}^{p} a_i b_i a_i^{-1} b_i^{-1} x x^{-1} c_1 c_1 \text{ (Relation 2)}\\ &\sim_{\text{top}} \prod_{i=2}^{p} a_i b_i a_i^{-1} b_i^{-1} x x^{-1} c_1 c_1 c_2 c_2 c_3 c_3 \text{ (Relation 3)}\\ &\sim_{\text{top}} \prod_{i=1}^{2p+1} c_i c_i. \text{ (Relation 3 by } p-1 \text{ times).}\end{aligned}$$

This is the theorem. □

By doing a handle on a nonorientable surface, 2 more genus should be added with the same nonorientability.

Theorem 1.8 On a nonorientable surface, what is obtained by doing a handle is nonorientable with its genus 2 added.

Proof Suppose N is the obtained surface, then

$$\begin{aligned} N &\sim_{\text{top}} \prod_{i=1}^{q} a_i a_i x a b a^{-1} b^{-1} x^{-1} \text{ (Relation 0)} \\ &\sim_{\text{top}} \prod_{i=1}^{q} a_i a_i x x^{-1} a b a^{-1} b^{-1} \text{ (Relation 1)} \\ &\sim_{\text{top}} \prod_{i=1}^{q} a_i a_i a b a^{-1} b^{-1} \text{ (Operation 1)} \\ &\sim_{\text{top}} \prod_{i=1}^{q+2} c_i c_i. \text{ (Relation 3)}. \end{aligned}$$

This is the theorem. □

By doing a crosscap on a nonorientable surface, 1 more genus produced with the same nonorientability

Theorem 1.9 On a nonorientable surface, what is obtained by doing a crosscap is nonorientable with its genus 1 added.

Proof Suppose N is the obtained surface, then

$$\begin{aligned} N &\sim_{\text{top}} \prod_{i=1}^{q} a_i a_i x a a x^{-1} \text{ (Relation 0)} \\ &\sim_{\text{top}} \prod_{i=1}^{q} a_i a_i x x^{-1} a a \text{ (Relation 2)} \\ &\sim_{\text{top}} \prod_{i=1}^{q} a_i a_i a a \text{ (Operation 1)} \\ &\sim_{\text{top}} \prod_{i=1}^{q+1} a_i a_i. \text{ (Relation 3).} \end{aligned}$$

This is the theorem. □

I.3 Embedding

Let $G = (V, E)$, $V = \{v_1, v_2, \cdots, v_n\}$, be a graph. A point in the 3-dimensional space is represented by a real number t as the parameter, *e.g.*, $(x, y, z) = (t, t^2, t^3)$. Write the vertices as

$$v_i = (x_i, y_i, z_i) = (t_i, t_i^2, t_i^3)$$

such that $t_i \neq t_j$, $i \neq j$, $1 \leq i, j \leq n$, and an edge as

$$(u, v) = u + \lambda v, \ 1 \leq \lambda \leq 1,$$

i.e., the straight line segment between u and v. Because for any four vertices v_i, v_j, v_l and v_k,

$$\det \begin{pmatrix} x_i - x_j & x_i - x_l & x_i - x_k \\ y_i - y_j & y_i - y_l & y_i - y_k \\ z_i - z_j & z_i - z_l & z_i - z_k \end{pmatrix}$$

$$= \det \begin{pmatrix} t_i - t_j & t_i - t_l & t_i - t_k \\ t_i^2 - t_j^2 & t_i^2 - t_l^2 & t_i^2 - t_k^2 \\ t_i^3 - t_j^3 & t_i^3 - t_l^3 & t_i^3 - t_k^3 \end{pmatrix}$$

D -256- - -* ! - B /% %% " 1"
*D 252842 := H?9& 2 = 9:=*5G
(% % !1 ' ! ! !1 7

B (/

— Géométrie Contemporaine
1re partie ...(1982). djvu
(Doubrovine)

Pág. 86(84) — §8.
Deuxième forme fondamentale
1. Courbure des courbes sur la surface dans l'espace euclidien.

obs: Pág. 87(85) —
On appelle courbure totale ou courbure de Gauss le déterminant de la matrice (a_{ij}).

Pág. 253(251)
C'est la formule de Gauss–Ostrogradski dans l'espace euclidien à trois dimensions.

— Géométrie Contemporaine
2e partie ...djvu

B
! -256- - -* ! - B 7M M 1 !
1! M M " ! J " 1 " ! ! " ! 1 M"
/ 1"

! -256- - -* ! - B 7M M 1 !
1! M M " ! J " 1 " ! ! " ! 1 M"
/ "KA

! -256- - -* ! - B 7M M 1 !
1! M M 1 " A! M M M" 1"

! -256- - -* ! - B 7M M 1 !
1! M M 1 " A! M M M" "KA

!% -256- - -* ! - B 7M M 1 ! %
1! M " " ! M " } M" 1"

!% -256- - -* ! - B 7M M 1 ! %
1! M " " ! M " } M" "KA

*D

-256- - -* ! - B 7M M 1 ! 1!
M M " ! J " 1 " ! ! " ! 1 M" / 1"
(B ' Z

Nous venons de voir donc que l'aire se définit, de même que la longueur, par la donnée du produit scalaire (g_{ij}) de deux vecteurs tangents en chaque point.

Exercices. 1. Le tore T^2 dans l'espace euclidien à trois dimensions est une surface de révolution engendrée par un cercle qui tourne autour d'une droite contenue dans le plan du cercle. On demande d'écrire les équations paramétriques du tore et la métrique induite sur le tore.

2. Chercher la première forme fondamentale pour l'ellipsoïde de révolution

$$\frac{x^2}{a^2} + \frac{y^2 + z^2}{b^2} = 1.$$

3. Chercher la métrique induite sur la surface de révolution

$$r(u, \varphi) = \{\rho(u) \cos \varphi, \rho(u) \sin \varphi, z(u)\}.$$

Vérifier que ses méridiens $\{\varphi = \text{const}\}$ et parallèles $\{u = \text{const}\}$ forment un réseau orthogonal. Définir les bissectrices des angles formés par les méridiens et les parallèles.

4. Définir sur la sphère les loxodromies, c'est-à-dire les courbes qui coupent les méridiens sous un angle donné α. Chercher la longueur d'une loxodromie.

5. Soit $F(x, y, z)$ une fonction homogène différentiable (c'est-à-dire telle que $F(cx, cy, cz) = c^n F(x, y, z)$). Montrer que la métrique sur une surface conique $F(x, y, z) = 0$ est euclidienne partout sauf à l'origine des coordonnées.

§ 8. Deuxième forme fondamentale

1. Courbure des courbes sur la surface dans l'espace euclidien. Soit une surface dans l'espace euclidien à trois dimensions. Soit un point non singulier (x_0, y_0, z_0) de cette surface. Supposons d'abord que l'axe des z soit normal au plan tangent en (x_0, y_0, z_0) à la surface: dans ce cas les axes des x et des y sont parallèles à ce plan. Alors la surface se définit localement, dans le voisinage du point (x_0, y_0, z_0), par l'équation $z = f(x, y)$, $z_0 = f(x_0, y_0)$, où

$$\left.\frac{\partial f}{\partial x}\right|_{\substack{x=x_0\\ y=y_0}} = \left.\frac{\partial f}{\partial y}\right|_{\substack{x=x_0\\ y=y_0}} = 0, \text{ donc } \operatorname{grad} f|_{\substack{x=x_0\\ y=y_0}} = 0. \tag{1}$$

Considérons la différentielle seconde de la fonction $z = f(x, y)$, c'est-à-dire $d^2 f = f_{xx} dx^2 + 2 f_{xy}\, dx\, dy + f_{yy}\, dy^2$, et écrivons la matrice (appelée *hessienne)* $a_{ij} = f_{x^i x^j}$, où $x^1 = x$, $x^2 = y$. Etudions cette matrice de la forme quadratique dans le point (x_0, y_0, z_0) où $\operatorname{grad} f = 0$.

DÉFINITION 1. Par *courbures principales* de la surface $z = f(x, y)$ dans le point (x_0, y_0, z_0) où grad $f = 0$, on entend les valeurs propres de la matrice (a_{ij}). On appelle *courbure totale* ou *courbure de Gauss* le déterminant de la matrice (a_{ij}) dans ce point, et *courbure moyenne*, la trace de la matrice dans ce point: $\operatorname{Tr} a_{11} + a_{22} = k_1 + k_2$; k_1 et k_2 sont les valeurs propres; la courbure totale est $K = k_1 k_2 = \det(a_{ij}) = a_{11}a_{22} - a_{12}a_{21}$.

Il sera montré dans le chapitre 4 que la courbure totale de la surface ne dépend que des propriétés métriques intrinsèques de cette surface.

Nous avons défini la notion de courbure en coordonnées spéciales attachées au point en question, où l'axe des z est normal à la surface et les axes des x et y sont tangents à celle-ci en ce point, ou, localement, $z = f(x, y)$ et grad $f = 0$ en (x_0, y_0). Afin de définir ces grandeurs en coordonnées quelconques, rappelons les faits de la théorie de la courbure des courbes sur la surface.

Supposons que la surface soit définie sous forme paramétrique:

$$r = r(u, v). \tag{2}$$

On a alors $[r_u, r_v] = |[r_u, r_v]| m$, où m est le vecteur unité de la normale à la surface, $|m| = 1$. Considérons une courbe $r = r(u(t), v(t))$ sur la surface. On a $\dot{r} = r_u \dot{u} + r_v \dot{v}$, $\ddot{r} = (r_{uu}\dot{u}^2 + 2r_{uv}\dot{u}\dot{v} + r_{vv}\dot{v}^2) + (r_u \ddot{u} + r_v \ddot{v})$. Puisque $r_u \perp m$ et $r_v \perp m$, il vient

$$\langle \ddot{r}, m \rangle = \langle r_{uu}, m \rangle \dot{u}^2 + 2\langle r_{uv}, m \rangle \dot{u}\dot{v} + \langle r_{vv}, m \rangle \dot{v}^2 = \\ = b_{11}\dot{u}^2 + 2b_{12}\dot{u}\dot{v} + b_{22}\dot{v}^2. \tag{3}$$

CONCLUSION. *La projection normale du vecteur accélération* $\langle \ddot{r}, m \rangle$ *est la forme quadratique du vecteur vitesse* $(\dot{u}, \dot{v})$ *en coordonnées locales* $u = x^1$, $v = x^2$.

Posons $b_{11} = L$, $b_{12} = M$, $b_{22} = N$. Il vient

$$\langle \ddot{r}, m \rangle dt^2 = b_{ij}\, dx^i\, dx^j = L(du)^2 + 2M\, du\, dv + N(dv)^2.$$

DÉFINITION 2. La forme quadratique $\langle \ddot{r}, m \rangle dt^2$ est appelée *deuxième forme fondamentale* de la surface (2).

Supposons que la courbe $u(t)$, $v(t)$ soit rapportée au paramètre naturel $t = l$. En vertu des formules de Frenet on a pour la courbe $r = r(t) = r(u(t), v(t))$

$$\ddot{r} = \frac{d^2 r}{dl^2} = kn,$$

où n est la normale principale à la courbe, et k la courbure de cette courbe. On a donc $\langle \ddot{r}, m \rangle = k \langle n, m \rangle = k \cos\theta$, où θ est l'angle entre m et n, d'où l'on tire

$$k \cos\theta (dl)^2 = \langle \ddot{r}, m \rangle dl^2 = L(du)^2 + 2M\, du\, dv + N(dv)^2 = \\ = b_{ij}\, dx^i\, dx^j, \quad x^1 = u, \quad x^2 = v,$$

où $dl^2 = g_{ij}\, dx^i\, dx^j = E(du)^2 + 2F\, du\, dv + G(dv)^2$.

CONCLUSION. On a

$$k\cos\theta = \frac{b_{ij}dx^i\,dx^j}{g_{ij}dx^i\,dx^j}, \quad x^1 = u, \quad x^2 = v, \tag{4}$$

ce qui achève la démonstration du

THÉORÈME 1. *Le produit de la courbure, dans un point donné, d'une courbe sur une surface dans l'espace à trois dimensions par le cosinus de l'angle entre la normale à la surface et la normale principale à la courbe est égal au rapport de la deuxième et de la première formes fondamentales sur le vecteur tangent à la courbe dans ce point.*

COROLLAIRE. *Si la courbe est la section d'une surface par un plan normal, on a*

$$\cos\theta = 1$$

et

$$k = \frac{b_{ij}dx^i\,dx^j}{g_{ij}dx^i\,dx^j}. \tag{5}$$

2. Invariants d'un couple de formes quadratiques. On a donc, en tout point de la surface, un couple de formes quadratiques:

1) $$dl^2 = g_{ij}\,dx^i\,dx^j, \tag{6}$$

2) $$\langle \ddot{r}, m\rangle\, dt^2 = b_{ij}\,dx^i\,dx^j, \tag{7}$$

la forme dl^2 étant toujours positive.

Quels sont les invariants d'un couple de formes quadratiques connus de l'Algèbre?

Soit, sur le plan, un couple de formes quadratiques (dont une positive) de matrices

$$G = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}, \qquad Q = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \tag{8}$$

($g_{21} = g_{12}$, $b_{21} = b_{12}$). Faisons l'équation

$$\det(Q - \lambda G) = 0, \tag{9}$$

ou, sous forme développée,

$$(b_{11} - \lambda g_{11})(b_{22} - \lambda g_{22}) - (b_{12} - \lambda g_{12})^2 = 0.$$

Les racines λ_1, λ_2 de cette équation sont appelées *valeurs propres du couple de formes quadratiques.*

Proposons-nous de résoudre les équations linéaires

$$\left.\begin{aligned} (b_{11} - \lambda_i g_{11})\xi_i^1 + (b_{12} - \lambda_i g_{12})\xi_i^2 = 0, \\ (b_{12} - \lambda_i g_{12})\xi_i^1 + (b_{22} - \lambda_i g_{22})\xi_i^2 = 0, \end{aligned}\right\} \tag{10}$$

où ξ_i^1, ξ_i^2 sont des inconnues.

Si λ_1, λ_2 sont les valeurs propres, le système (10) admet des solutions non triviales

$$f_1 = (\xi_1^1, \xi_1^2) \quad \text{et} \quad f_2 = (\xi_2^1, \xi_2^2).$$

Les directions des vecteurs f_1, f_2 sont dites *directions principales* du couple de formes quadratiques; f_1 correspond à λ_1 et f_2 à λ_2.

De même qu'auparavant, les produits scalaires $\langle e_i, e_j \rangle$ des vecteurs de base du plan sont désignés par g_{ij}, $i, j = 1, 2$ (la métrique riemannienne étant définie par la forme g_{ij}).

LEMME 1. *Si les valeurs propres du couple de formes quadratiques sont différentes, les directions principales sont orthogonales.*

Nous avons deux directions principales f_1, f_2,

$$f_1 = \xi_1^1 e_1 + \xi_1^2 e_2, \qquad f_2 = \xi_2^1 e_1 + \xi_2^2 e_2.$$

Par définition, leur orthogonalité signifie que

$$\langle f_1, f_2 \rangle = g_{ij} \xi_1^i \xi_2^j = 0.$$

Démonstration. Choisissons un couple de vecteurs plans d_1, d_2 tels que $\langle d_i, d_j \rangle = \delta_{ij}$. Cela est possible, parce que la forme quadratique de matrice g_{ij} est définie positive et qu'on peut l'exprimer comme une somme de carrés par une transformation linéaire. Etudions maintenant la deuxième forme quadratique par rapport à la nouvelle base d_1, d_2 en posant

$$e_i = a_i^j d_j, \qquad A = (a_i^j). \tag{11}$$

Dans cette nouvelle base d_1, d_2 les matrices $\overline{G}$ et $\overline{Q}$ de la première et de la deuxième forme quadratique vérifient les relations

1) $\overline{G} = (\langle d_i, d_j \rangle) = (\delta_{ij})$ ou $\overline{G} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, donc $G = AA^{\mathrm{T}}$;

2) $Q = A\overline{Q}A^{\mathrm{T}}$.

Puisque $G = AA^{\mathrm{T}}$ et $Q = A\overline{Q}A^{\mathrm{T}}$, on a

$$\begin{gathered} Q - \lambda G = A(\overline{Q} - \lambda \cdot 1)A^{\mathrm{T}}, \\ \det(Q - \lambda G) = (\det A)^2 \det(\overline{Q} - \lambda \cdot 1). \end{gathered} \tag{12}$$

Remarquons que $\det A = \det A^{\mathrm{T}} = \sqrt{g} = \sqrt{\det G} \neq 0$. Aussi l'équation (9) est-elle équivalente à

$$\det(\overline{Q} - \lambda \cdot 1) = 0. \tag{13}$$

Les solutions de n'importe laquelle de ces deux équations aux valeurs propres λ_1 et λ_2 donnent les directions principales f_1 et f_2.

Si la base est d_1, d_2, le produit scalaire est euclidien. Il est défini par la matrice unité $\overline{G} = 1 = (\delta_{ij})$. On apprend dans le cours d'Algèbre que toute forme quadratique $\overline{Q}$ peut être ramenée, au moyen d'une rotation, à une forme définie par une matrice diagonale dont les vecteurs propres f_1, f_2 sont orthogonaux au sens euclidien usuel. Le lemme est démontré.

Ce lemme n'est qu'une variante du théorème sur la diagonalisation par rotation d'une forme quadratique dans le plan euclidien.

3. Propriétés de la deuxième forme fondamentale. Revenons maintenant aux deux formes fondamentales de la surface dans l'espace euclidien à trois dimensions:

$$g_{ij}\,dx^i\,dx^j = dl^2, \tag{14}$$

$$b_{ij}dx^i\,dx^j, \qquad x^1 = u, \qquad x^2 = v. \tag{15}$$

Le rapport de ces formes quadratiques est la courbure de la section normale.

DÉFINITION 3. Les valeurs propres de ce couple de formes quadratiques s'appellent *courbures principales de la surface* dans le point donné. Le produit des courbures principales est la *courbure de Gauss*, ou *courbure totale de la surface*, et leur somme, la *courbure moyenne* [*] *de la surface*.

Exemple. Supposons que la surface soit définie sous la forme $z = f(x, y)$ avec $f_x = f_y = 0$ dans un point donné (x_0, y_0) (l'axe des z étant normal au plan tangent à la surface en ce point). Soit $x = u$, $y = v$, $z = f(u, v)$. Dans le point x_0, y_0 on a pour la première et la deuxième forme fondamentale

$$\begin{aligned}
&\text{(I)} && g_{11} = 1,\ g_{22} = 1,\ g_{12} = g_{21} = 0 \quad (g_{ij} = \delta_{ij}),\\
&\text{(II)} && L\ = b_{11} = \langle r_{uu}, m\rangle = f_{xx}(x_0, y_0),\\
& && M = b_{12} = \langle r_{uv}, m\rangle = f_{xy}(x_0, y_0),\\
& && N\ = b_{22} = \langle r_{vv}, m\rangle = f_{yy}(x_0, y_0).
\end{aligned} \tag{16}$$

Le vecteur m se confond ici avec le vecteur unité porté par l'axe des z.

Ainsi donc, la deuxième forme fondamentale se présente dans le point considéré comme suit:

$$b_{ij}\,dx^i\,dx^j = f_{x^ix^j}\,dx^i\,dx^j = d^2f. \tag{17}$$

La courbure totale se confond dans ce cas avec le déterminant $\det\begin{pmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy}\end{pmatrix}$. Les valeurs propres se laissent déduire de l'équation $(f_{xx} - \lambda)(f_{yy} - \lambda) - (f_{xy})^2 = 0$, puisque $g_{ij} = \delta_{ij}$.

Le plan tangent à la surface dans ce point est parallèle au plan (x, y). On obtient les directions principales dans ce point par la résolution des équations

$$(f_{x^ix^j} - \lambda_1\delta_{ij})\xi_1^j = 0, \qquad i, j = 1, 2 \text{ pour } f_1, \tag{18}$$

$$(f_{x^ix^j} - \lambda_2\delta_{ij})\xi_2^j = 0, \qquad i, j = 1, 2 \text{ pour } f_2. \tag{19}$$

---

[*] Généralement, on entend par courbure moyenne la demi-somme des courbures principales.

Puisque $f_1 \perp f_2$, nous pouvons retenir les vecteurs unités des directions principales comme les nouveaux axes de coordonnées x', y', c'est-à-dire effectuer une rotation dans le plan (x, y), avec la seule condition que $\lambda_1 \neq \lambda_2$.

En ces nouvelles coordonnées (z, x', y') on a

$$z = f(x(x', y'),\ y(x', y')),$$

où $x = x' \cos\varphi + y' \sin\varphi$, $y = -x' \sin\varphi + y' \cos\varphi$, en désignant par φ l'angle de rotation.

En coordonnées nouvelles la deuxième forme fondamentale se présente (dans le point considéré seulement!) comme suit:

$$\lambda_1(dx')^2 + \lambda_2(dy')^2. \tag{20}$$

La courbure de la section normale dans le point donné sera alors

$$k = \frac{\lambda_1(dx')^2 + \lambda_2(dy')^2}{(dx')^2 + (dy')^2}. \tag{21}$$

Le vecteur tangent e à la section normale de la surface se présente dans le point donné comme $e = (\dot{x}', \dot{y}')$, avec

$$dx' = \dot{x}'\, dt, \qquad dy' = \dot{y}'\, dt,$$

aussi

$$\cos^2\alpha = \frac{(dx')^2}{(dx')^2 + (dy')^2}, \qquad \sin^2\alpha = \frac{(dy')^2}{(dx')^2 + (dy')^2}, \tag{22}$$

où α est l'angle que fait l'axe des x avec le vecteur tangent e à la section normale.

Nous constatons donc que dans notre système de coordonnées privilégiées $k = \lambda_1 \cos^2\alpha + \lambda_2 \sin^2\alpha$. Cette formule porte le nom de *formule d'Euler*. Remarquons qu'elle reste valable dans n'importe quel système de coordonnées.

THÉORÈME 2. *La courbure de la section normale dans un point donné est égale à*

$$k = \lambda_1 \cos^2\alpha + \lambda_2 \sin^2\alpha,$$

où λ_1, λ_2 sont les courbures principales dans le point considéré, et α l'angle sur la surface entre le vecteur tangent à la section normale et la direction principale correspondante.

Démonstration. La formule d'Euler vient d'être établie pour le cas où la surface est définie sous la forme $z = f(x, y)$ et $f_x = f_y = 0$ dans le point donné (x_0, y_0). Or, puisque le résultat obtenu ne dépend pas du système de coordonnées choisi, nous pouvons toujours choisir des coordonnées au point considéré de telle façon que l'axe des z soit normal à la surface dans ce point et que les axes des x et y soient tangents à la surface et orthogonaux entre eux (on peut même faire confondre ces axes et les directions principales). La surface se définit alors dans le voisinage du point considéré par l'expression

$$z = f(x, y), \qquad f_x = f_y = 0,$$

et même $f_{xy} = f_{yx} = 0$ (toujours dans le point considéré) si les axes se confondent avec les directions principales. On a alors en ce point $\lambda_1 = f_{xx}$, $\lambda_2 = f_{yy}$. Puisque la formule d'Euler a été établie justement dans ces coordonnées, le théorème est démontré.

Remarque. Quand on inverse le sens de l'axe des z, le signe des courbures principales se trouve inversé lui aussi.

La formule d'Euler donne lieu à un corollaire: si $\lambda_1 > \lambda_2$, alors λ_1 et λ_2 sont les courbures maximale et minimale des sections normales dans le point donné (en tenant compte du signe).

Citons quelques expressions utiles de la deuxième forme fondamentale.

Si la surface est définie par l'équation $z = f(x, y)$, les coefficients de la deuxième forme fondamentale s'écriront (avec $x = u$, $y = v$)

$$r_u = (1, 0, f_x), \qquad r_v = (0, 1, f_y),$$
$$[r_u, r_v] = (-f_x, -f_y, 1), \tag{23}$$

$$r_{uu} = (0, 0, f_{xx}), \quad r_{uv} = (0, 0, f_{xy}), \quad r_{vv} = (0, 0, f_{yy}),$$

$$m = \frac{[r_u, r_v]}{\|[r_u, r_v]\|} = \frac{(-f_x, -f_y, 1)}{\sqrt{1 + f_x^2 + f_y^2}},$$

$$L = b_{11} = \frac{f_{xx}}{\sqrt{1 + f_x^2 + f_y^2}}, \qquad M = b_{12} = b_{21} = \frac{f_{xy}}{\sqrt{1 + f_x^2 + f_y^2}},$$

$$N = b_{22} = \frac{f_{yy}}{\sqrt{1 + f_x^2 + f_y^2}}.$$

D'où nous déduisons

$$b_{ij}\, dx^i\, dx^j = \frac{1}{\sqrt{1 + f_{x^1}^2 + f_{x^2}^2}} (f_{x^i x^j}\, dx^i\, dx^j), \tag{24}$$
$$x^1 = u = x, \qquad x^2 = v = y.$$

Rappelons que nous avons établi les formules (7.15) pour les coefficients g_{ij}:

$$g_{11} = 1 + f_x^2, \qquad g_{12} = g_{21} = f_x f_y, \qquad g_{22} = 1 + f_y^2,$$
$$g = g_{11} g_{22} - g_{12}^2 = 1 + f_x^2 + f_y^2. \tag{25}$$

On a le

THÉORÈME 3. *La courbure totale de la surface est égale au rapport des déterminants de ses deuxième et première formes fondamentales:*

$$K = \frac{b_{11} b_{22} - b_{12}^2}{g_{11} g_{22} - g_{12}^2}. \tag{26}$$

En particulier, si la surface est définie sous la forme d'un graphique $z = f(x, y)$, *on a*

$$K = \frac{f_{xx}f_{yy} - f_{xy}^2}{(1 + f_x^2 + f_y^2)^2} \cdot$$

Démonstration. Nous avons déduit les valeurs propres λ_1 et λ_2 de l'équation (9)

$$\det(Q - \lambda G) = 0$$

dans laquelle $Q = (b_{ij}) = \begin{pmatrix} L & M \\ M & N \end{pmatrix}$ est la matrice de la deuxième forme fondamentale, $G = (g_{ij})$. Remarquons que $Q - \lambda G = (b_{ij} - \lambda g_{ij})$ et $\det(Q - \lambda G) = (b_{11} - \lambda g_{11}) = (b_{22} - \lambda g_{22}) - (b_{12} - \lambda g_{12})^2$. La matrice $G = (g_{ij})$ est définie positive, donc non dégénérée. Il vient donc

$$\det(Q - \lambda G) = \det G \det(G^{-1}Q - \lambda \cdot 1),$$

où $\det G \neq 0$. Les valeurs propres λ_1 et λ_2 sont solutions de l'équation $\det(G^{-1}Q - \lambda \cdot 1) = 0$. Rappelons un fait connu de l'Algèbre: le produit de toutes les valeurs propres d'une matrice est égal à son déterminant; plus exactement, on a $\det A = \lambda_1\lambda_2$ pour les matrices du second ordre. En posant $A = G^{-1}Q$, nous voyons que

$$\lambda_1\lambda_2 = \det(G^{-1}Q) = \frac{\det Q}{\det G} = K. \tag{27}$$

Il s'ensuit que la courbure totale est égale au rapport des déterminants des matrices de la deuxième et de la première formes fondamentales.

Ensuite, si la surface est définie par $z = f(x, y)$, les coefficients b_{ij} et g_{ij} sont donnés par les formules (23) et (7.15). En calculant les déterminants et en faisant leurs rapports, nous aboutissons à la formule de la courbure totale.

COROLLAIRE. *Si la surface est définie par le graphique* $z = f(x, y)$, *le signe de la courbure totale* K *est le même que celui du déterminant* $f_{xx}f_{yy} - f_{xy}^2$, *parce que*

$$K = \frac{f_{xx}f_{yy} - f_{xy}^2}{(1 + f_x^2 + f_y^2)^2} \cdot$$

Exemple. Soit une surface définie par $z = f(x, y)$, où la fonction $f(x, y)$ vérifie l'équation de Laplace $f_{xx} + f_{yy} = 0$. On a alors $f_{xx}f_{yy} - f_{xy}^2 \leqslant 0$, car $f_{xx} = -f_{yy}$. Aussi la courbure totale K est-elle inférieure à zéro chaque fois que l'une au moins des deux dérivées f_{xx}, f_{xy} est non nulle.

Précisons la signification géométrique de la courbure totale. Plaçons dans le point considéré de la surface un repère orthonormé (x, y, z) tel que l'axe des z soit normal à la surface. La surface se définira alors localement par $z = f(x, y)$, où $f_x = f_y = 0$ dans le point en question.

On a dans le point donné $g_{ij} = \delta_{ij}$, parce que $g_{11} = 1 + f_x^2$, $g_{12} = f_x f_y$, $g_{22} = 1 + f_y^2$. Il vient ensuite $L = b_{11} = f_{xx}$, $M = b_{12} = f_{xy}$, $N = b_{22} = f_{yy}$. Considérons les trois cas suivants.

1) $K > 0$, $\lambda_1 > 0$, $\lambda_2 > 0$ (la fonction $f(x, y)$ admet un minimum pour $x = x_0$, $y = y_0$, fig. 10, *a*).

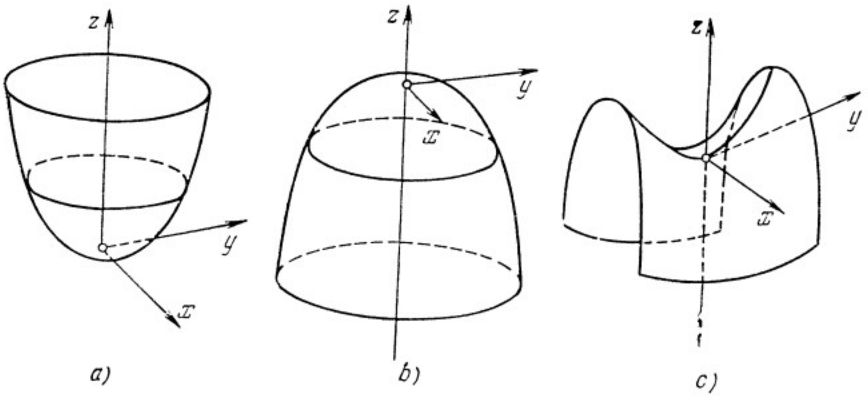

Fig. 10.

2) $K > 0$, $\lambda_1 < 0$, $\lambda_2 < 0$ (la fonction admet un maximum pour $x = x_0$, $y = y_0$, fig. 10, *b*).

3) $K < 0$, donc $\lambda_1 > 0$, $\lambda_2 < 0$ ou inversement: la fonction passe par un « point selle » ou « col » (fig. 10, *c*).

CONCLUSION. *Pour* $K > 0$ *la surface est localement située d'un même côté du plan tangent dans le voisinage du point donné. Pour* $K < 0$ *la surface coupe toujours le plan tangent aussi près que l'on veut du point de contact.*

Si la courbure totale est partout positive, il s'agit d'une surface *strictement convexe*.

Exercices. 1. Chercher la surface dont toutes les normales viennent concourir en un point.

2. Calculer la deuxième forme fondamentale sur la surface de révolution

$$r(u, \varphi) = \{x(u),\ \rho(u)\cos\varphi,\ \rho(u)\sin\varphi\}, \qquad \rho(u) > 0.$$

3. Calculer les courbures totale et moyenne de la surface d'équation

$$z = f(x) + g(y).$$

4. Montrer que si la surface plongée dans l'espace euclidien à trois dimensions a ses courbures totale et moyenne identiquement nulles, il s'agit d'un plan.

5. Montrer que la courbure moyenne H de la surface $z = f(x, y)$ est

$$H = \operatorname{div}\left(\frac{\operatorname{grad} f}{\sqrt{1 + |\operatorname{grad} f|^2}}\right).$$

6. Soit S une surface formée par des droites tangentes à la courbe donnée de courbure $k(l)$. Montrer que si la courbure $k(l)$ se conserve le long de la courbe, la surface S conserve sa métrique.

7. Si la métrique d'une surface est

$$dl^2 = A^2 du^2 + B^2 dv^2, \qquad A = A(u, v), \qquad B = B(u, v),$$

la courbure totale est

$$K = -\frac{1}{AB}\left[\left(\frac{A_v}{B}\right)_v + \left(\frac{B_u}{A}\right)_u\right].$$

8. Montrer que les seules surfaces de révolution admettant une courbure moyenne nulle sont le plan et la caténoïde engendrée par la rotation de la courbe $\left(\frac{\operatorname{ch}(at + b)}{a}, t\right)$.

§ 9. Métrique de la sphère

La sphère $S^2 \subset \mathbf{R}^3$ de rayon R centrée à l'origine des coordonnées a pour équation

$$x^2 + y^2 + z^2 = R^2. \tag{1}$$

En coordonnées sphériques r, θ, φ cette sphère a pour équation $r = R$, avec θ et φ quelconques. Aussi les paramètres (θ, φ) peuvent-ils servir de coordonnées locales sur la sphère, sauf en ses pôles Nord et Sud (en lesquels on a $\theta = 0$, $\theta = \pi$; ce sont les points singuliers du système de coordonnées sphériques; voir § 1). On sait (voir § 3, n° 1) que la métrique euclidienne $dx^2 + dy^2 + dz^2$ s'écrit en coordonnées sphériques comme suit:

$$dx^2 + dy^2 + dz^2 = dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\varphi^2). \tag{2}$$

Sur la surface $r = R$ la différentielle dr s'annule, si bien qu'on a pour la métrique de la sphère dans les coordonnées (θ, φ)

$$dl^2 = R^2(d\theta^2 + \sin^2\theta\, d\varphi^2). \tag{3}$$

Ici $0 \leqslant \varphi \leqslant 2\pi$, $0 \leqslant \theta \leqslant \pi$ (fig. 11). Dans un petit voisinage du point 0 on a $\sin\theta \sim \theta$, ce qui fait que $\frac{1}{R^2} dl^2$ est approximativement égal à la métrique euclidienne $d\theta^2 + \theta^2(d\varphi)^2$ (en coordonnées polaires θ, φ). Considérons la projection stéréographique de la sphère sur le plan (voir fig. 12 où l'on montre la section plane de la sphère). Ici (θ, φ) sont les coordonnées sur la sphère et (r, φ) les coordonnées polaires sur le plan. On remarque sur la figure 12 que $\varphi = \varphi$, $r = R \operatorname{cotg} \frac{\theta}{2}$. En utilisant ces formules de changement de coordonnées, nous pouvons transcrire la métrique (3) dans les coordonnées (r, φ) (ou dans les coordonnées x, y):

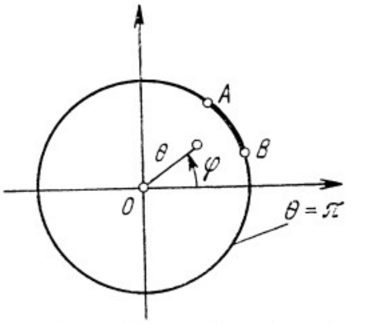

Fig. 11. La distance des points A et B mesurée le long du cercle de rayon $\theta = \pi$ est égale à zéro, c'est-à-dire que la frontière du cercle se réduit à un point; on obtient donc une sphère à deux dimensions.

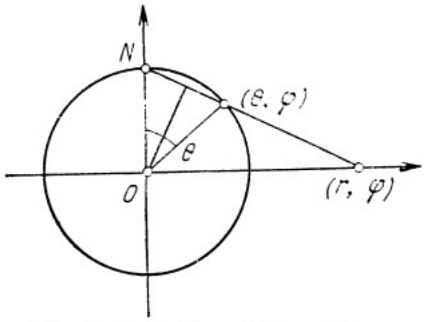

Fig. 12. Projection stéréographique $(\theta, \varphi) \to (r, \varphi)$.

$$dl^2 = \frac{4R^4}{(R^2 + r^2)^2}(dr^2 + r^2 d\varphi^2) = \frac{4R^4}{(R^2 + x^2 + y^2)^2}(dx^2 + dy^2). \quad (4)$$

Il est clair qu'on obtient la métrique de la sphère en multipliant la métrique du plan euclidien par la fonction $\frac{4R^4}{(R^2 + x^2 + y^2)^2}$, soit

$$dl^2_{\text{sphère}} = \frac{4R^4}{(R^2+x^2+y^2)^2} dl^2_{\text{plan}}. \quad (5)$$

Exemple. Proposons-nous de chercher la longueur du cercle et l'aire du disque de rayon ρ sur une sphère. Supposons que le disque soit centré au pôle Nord (noté N) de la sphère (point $\theta = 0$, fig. 13). Le rayon ρ du cercle est alors $\rho = R\theta_0$. Le cercle

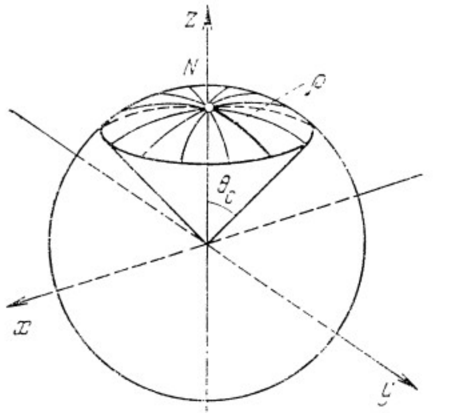

Fig. 13. Le disque de rayon ρ sur la sphère.

a alors pour équation $\theta = \rho/R$, $0 \leqslant \varphi \leqslant 2\pi$. Le disque de rayon ρ est le domaine $\theta \leqslant \rho/R$, $0 \leqslant \varphi \leqslant 2\pi$. Sur la courbe $\theta = \rho/R = \text{const}$ nous avons

$$dl^2 = R^2(d\theta^2 + \sin^2\theta\, d\varphi^2) = R^2 \sin^2 \frac{\rho}{R}\,(d\varphi)^2,$$

aussi la longueur l_ρ est-elle

$$l_\rho = \int_0^{2\pi} R \sin\left(\frac{\rho}{R}\right) d\varphi = 2\pi R \sin\left(\frac{\rho}{R}\right)\cdot \tag{6}$$

Nous voyons que pour $\rho = R\pi/2$ la longueur est maximale (c'est l'équateur), et pour $\rho = \pi$ elle est minimale (le pôle Sud). De la formule (6) il ressort que le rapport de la longueur du cercle l_ρ au rayon ρ sur la sphère est toujours plus petit que 2π:

$$\frac{l_\rho}{\rho} = 2\pi \frac{\sin(\rho/R)}{\rho/R} < 2\pi \text{ pour } \rho > 0.$$

Cherchons maintenant l'aire σ_ρ du disque de rayon ρ. Il ressort de la représentation (3) de dl^2 que $\sqrt{g} = R^2|\sin\theta|$. L'aire cherchée est donc

$$\sigma_\rho = \iint\limits_{0 \leqslant \theta \leqslant \rho/R} R^2 |\sin\theta|\, d\theta\, d\varphi = R^2 \int_0^{\rho/R} \sin\theta\, d\theta \int_0^{2\pi} d\varphi =$$

$$= 2\pi R^2 \left(1 - \cos\frac{\rho}{R}\right), \qquad \rho \leqslant \frac{\pi}{R}\cdot \tag{7}$$

Pour $\rho = \pi R$ notre disque va envahir la sphère tout entière, si bien que l'aire de la sphère est égale à $4\pi R^2$.

Si le rayon ρ est faible, on a $\sin\frac{\rho}{R} \approx \frac{\rho}{R}$ et $1 - \cos\frac{\rho}{R} \approx \frac{\rho^2}{2R^2}$. Il vient donc

$$l_\rho = 2\pi R \sin\frac{\rho}{R} \approx 2\pi\rho,$$

$$\sigma_\rho = 2\pi R^2\left(1 - \cos\frac{\rho}{R}\right) \approx \pi\rho^2,$$

ce qui signifie que, lorsque les rayons sont faibles, les longueurs et les aires s'expriment sensiblement par les mêmes formules qu'en géométrie euclidienne.

Cherchons maintenant les courbures totale et moyenne de la sphère de rayon R. Notons que les sections normales de la sphère sont des cercles de rayon R (appelés grands cercles de la sphère). Aussi toute section normale a-t-elle une courbure constante égale à R^{-1} (voir § 5, n° 1). Ainsi donc, les deux valeurs propres de la deuxième forme fondamentale sont égales entre elles et à R^{-1}. Il s'ensuit que la courbure totale de la sphère est égale à $1/R^2$, et sa courbure moyenne, à $2/R$.

Passons au groupe des déplacements conservant la métrique de la sphère. Toute rotation de l'espace effectuée autour de l'origine des coordonnées applique la sphère de rayon R dans elle-même. Cette rotation définie par une matrice orthogonale conserve la métrique riemannienne sur la sphère. Ainsi donc, les déplacements de la sphère S^2 sont définis par des matrices orthogonales. Rappelons (§ 4, n° 3) que le groupe des matrices orthogonales se note $O(3)$ et s'appelle groupe orthogonal. Toute rotation se laisse définir à l'aide de trois paramètres. Donc, la métrique de la sphère admet un ensemble de déplacements au moins tridimensionnel (triparamétrique).

Remarque importante. Nous avons vu que toutes les transformations de la sphère définies par des matrices orthogonales sont des déplacements. Aussi le groupe $O(3)$ est-il au moins un *sous-groupe* du groupe des déplacements de la sphère S^2. Nous n'avons pas encore démontré que le groupe $O(3)$ est le groupe des déplacements. C'est pourtant ce qui a lieu. Toute transformation conservant la métrique sur S^2 est une transformation linéaire et orthogonale dans $\mathbf{R}^3$. Or, la démonstration rigoureuse de cette assertion est fondée sur la notion de géodésique; nous reviendrons sur cette question dans le chapitre 4.

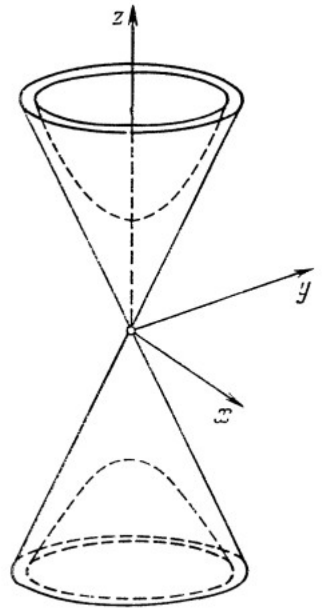

Fig. 14.

§ 10. Surfaces spatiales dans l'espace pseudo-euclidien

1. Pseudosphère. Considérons un espace pseudo-euclidien à trois dimensions muni de coordonnées (t, x, y) dans lesquelles la métrique pseudo-euclidienne s'écrit

$$(dl^2) = (dt)^2 - (dx)^2 - (dy)^2. \qquad (1)$$

La pseudosphère de rayon R se définit dans l'espace $\mathbf{R}^3_1$ par l'équation

$$t^2 - x^2 - y^2 = R^2. \qquad (2)$$

C'est un hyperboloïde à deux nappes dans l'espace à trois dimensions (fig. 14). Contenue tout entière à l'intérieur du cône de lumière $t^2 - x^2 -$

$- y^2 = 0$, la pseudosphère se définit en coordonnées pseudosphériques ρ, χ, φ (voir (3.11), (3.12)) par les équations

$$\rho = R \text{ (nappe supérieure)},$$
$$\rho = - R \text{ (nappe inférieure)}.$$

Nous ne considérerons que la nappe supérieure de l'hyperboloïde. Rappelons que la métrique $dt^2 - dx^2 - dy^2$ exprimée en coordonnées pseudosphériques ρ, χ, φ a pour équation (3.12):

$$dl^2 = - \rho^2(d\chi^2 + \operatorname{sh}^2\chi\, d\varphi^2) + d\rho^2. \tag{3}$$

Aussi a-t-on pour l'hyperboloïde, où $d\rho = 0$,

$$-dl^2 = R^2(d\chi^2 + \operatorname{sh}^2\chi\, d\varphi^2). \tag{4}$$

Ainsi donc, la métrique induite par la métrique pseudo-euclidienne (donc indéfinie) sur la surface de la pseudosphère est définie négative.

Cela revient à dire que la pseudosphère $t^2 - x^2 - y^2 = R^2$ dans $\mathbf{R}_1^3$ est une hypersurface spatiale: en effet, les vecteurs tangents à cette surface sont toujours spatiaux (avec $dl^2 < 0$).

DÉFINITION 1. La métrique (4) porte le nom de *métrique lobatchevskienne*.

Par analogie avec le paragraphe précédent, on peut construire la projection stéréographique de la pseudosphère sur le plan (x, y). La pseudosphère a pour centre l'origine des coordonnées O, son pôle Nord est le point de coordonnées $(R, 0, 0)$, et son pôle Sud, le point de coordonnées $(- R, 0, 0)$. La nappe supérieure de la pseudosphère aura comme image le disque ouvert $x^2 + y^2 < R^2$ (fig. 15).

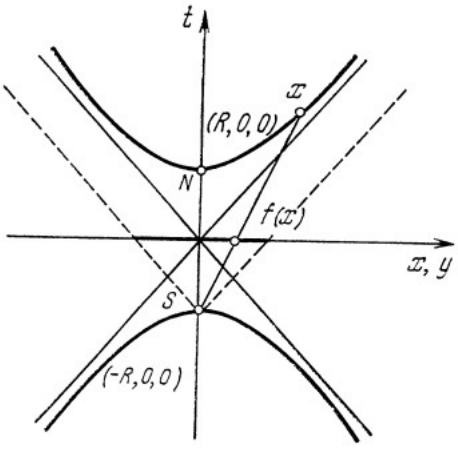

Fig. 15.

Soient t, x, y les coordonnées d'un point P sur la pseudosphère (avec $t > 0$), et u, v les coordonnées du point $f(P)$, où f est la projection stéréographique. Cherchons à expliciter la relation entre ces coordonnées. En regardant la figure 15, on s'assure que

$$\frac{x}{u} = \frac{t+R}{R}, \qquad \frac{y}{v} = \frac{t+R}{R},$$

d'où $x = u\left(1 + \frac{t}{R}\right)$, $y = v\left(1 + \frac{t}{R}\right)$. En portant ces expressions de x et y dans l'équation de la surface $t^2 - x^2 - y^2 = 0$, on obtient

$$t = -R\left(1 + \frac{2R^2}{u^2 + v^2 - R^2}\right), \tag{5}$$

d'où

$$x = \frac{2R^2u}{R^2 - u^2 - v^2}, \qquad y = \frac{2R^2v}{R^2 - u^2 - v^2}. \tag{6}$$

Les formules (5) et (6) sont les relations cherchées pour la projection stéréographique. Leur connaissance permet d'exprimer la métrique (4) sur la pseudosphère en coordonnées (u, v). Par calcul direct on obtient

$$-dl^2 = -(dt^2 - dx^2 - dy^2) = \frac{4R^2}{(u^2 + v^2 - R^2)^2}(du^2 + dv^2). \tag{7}$$

On voit qu'ici encore, la métrique de la pseudosphère est proportionnelle en coordonnées u, v à la métrique du plan euclidien (comme dans le cas de la sphère S^2). La métrique (7) sur le disque $u^2 + v^2 < R^2$ affectée du signe négatif s'appelle *métrique de Poincaré* de la géométrie de Lobatchevski. En introduisant des coordonnées polaires (r, φ) sur le disque $u^2 + v^2 < R^2$, la métrique lobatchevskienne s'écrira

$$dl^2 = \frac{4R^2}{(R^2 - r^2)^2}(dr^2 + r^2 d\varphi^2). \tag{8}$$

L'analogie avec la métrique de la sphère S^2 est évidente (voir (9.4)).

Résumons les différents types de métriques de la sphère S^2 et du plan lobatchevskien L^2 dans un tableau (en posant $R = 1$):

S^2	L^2
$d\theta^2 + \sin^2\theta\,(d\varphi)^2$	$d\chi^2 + \text{sh}^2\chi\,(d\varphi)^2$
$4\dfrac{dx^2 + dy^2}{(1 + x^2 + y^2)^2}$	$4\dfrac{dx^2 + dy^2}{(1 - x^2 - y^2)^2}$, $x^2 + y^2 < 1$

Nous donnerons maintenant une autre notation de la métrique lobatchevskienne. Elle sera basée sur un fait bien connu de la Théorie des fonctions d'une variable complexe: il existe une transformation homographique du plan complexe par laquelle le demi-plan supérieur se transforme en un disque unité. Indiquons à titre d'exemple la transformation $z = (1 + iw)/(1 - iw)$ (fig. 16). En posant $z = u + iv$, $w = x + iy$, nous intro-

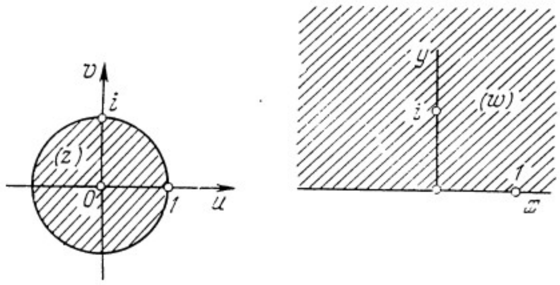

Fig. 16.

duisons sur le disque unité un nouveau système de coordonnées (x, y) avec $y > 0$. Le calcul direct montre que la métrique lobatchevskienne se présente dans les coordonnées (x, y) comme suit:

$$dl^2 = \frac{dx^2 + dy^2}{y^2}, \qquad y > 0. \tag{9}$$

La métrique (9) est dite *métrique de Klein* de la géométrie de Lobatchevski.

Cherchons le groupe des déplacements du plan lobatchevskien. Puisque toute transformation pseudo-orthogonale de l'espace $\mathbf{R}_1^3$ (voir § 6, n° 2) conserve la forme $t^2 - x^2 - y^2$, la pseudosphère $t^2 - x^2 - y^2 = R^2$ revient à elle-même. Or, si la transformation appartient au groupe $O(1, 2)$, il se peut que les nappes supérieure et inférieure de la pseudosphère soient permutées: le groupe des déplacements du plan lobatchevskien contient donc le groupe des transformations orthochrones appartenant à $O(1, 2)$. Nous verrons dans le chapitre 4 que ces groupes se confondent. Ainsi donc, le plan lobatchevskien admet, au même titre que la sphère et le plan euclidien, un ensemble de déplacements triparamétrique.

2. Courbure des surfaces spatiales dans $\mathbf{R}_1^3$. Nous dirons qu'une surface dans $\mathbf{R}_1^3$ est spatiale si tel est tout vecteur tangent à cette surface. En d'autres termes, la métrique minkowskienne $dl^2 = dt^2 - dx^2 - dy^2$ induit sur la surface une métrique définie négative.

Dans le cas où la surface est définie sous forme d'un graphique de fonction $t = f(x, y)$, la métrique de la surface se présente comme suit:

$$\begin{gathered} -dl^2 = -(dt^2 - dx^2 - dy^2) = \\ = (1 - f_x^2)\,dx^2 - 2f_x f_y\,dx\,dy + (1 - f_y^2)\,dy^2 = g_{ij}\,dx^i\,dx^j, \\ x^1 = x, \ x^2 = y. \end{gathered}$$

On a $\det g_{ij} = 1 - f_x^2 - f_y^2$; pour que la surface soit spatiale, il faut qu'il y ait $\det g_{ij} = 1 - f_x^2 - f_y^2 > 0$. Le vecteur unité de la normale à la surface est

$$m = \frac{(1, f_x, f_y)}{\sqrt{1 - f_x^2 - f_y^2}}$$

(vecteur spatial). La deuxième forme fondamentale de la surface a pour équation

$$b_{ij} = \left\langle \frac{\partial^2 r}{\partial x^i \, \partial x^j}, m \right\rangle, \qquad i, j = 1, 2.$$

Introduisons la courbure totale K de la surface spatiale en posant

$$K = -\frac{\det b_{ij}}{\det g_{ij}}. \tag{10}$$

Pour les surfaces d'équation $t = f(x, y)$, $1 - f_x^2 - f_y^2 > 0$, il vient par analogie avec le § 8

$$K = \frac{f_{xy}^2 - f_{xx} f_{yy}}{(1 - f_x^2 - f_y^2)^2}. \tag{11}$$

En particulier, pour l'hyperboloïde $t^2 - x^2 - y^2 = 1$ (géométrie de Lobatchevski) on a $K \equiv -1$.

Remarques. 1. Dans le chapitre 4 nous calculerons la courbure du plan lobatchevskien en ne tenant compte que de sa géométrie intrinsèque. Ce calcul nous donnera encore $K \equiv -1$, ce qui explique le choix du signe dans la définition (10).

2. De même que dans le § 8, nous constatons que les surfaces spatiales de courbure totale **négative** $K < 0$ dans l'espace $\mathbf{R}_1^3$ sont convexes.

§ 11. Le langage des nombres complexes en géométrie

1. Coordonnées complexes et réelles. Dans bien des cas, le langage des nombres complexes offre des avantages certains en géométrie. L'utilisation de ce langage repose sur quelques faits élémentaires que nous exposerons dans le texte qui suit.

Soit un espace linéaire $\mathbf{C}^n$ de dimension n sur le corps des nombres complexes $\mathbf{C}$ avec la base $e_1, \ldots, e_n$. Chaque vecteur $\xi \in \mathbf{C}^n$ se présente comme suit:

$$\xi = z^k e_k, \qquad z^k = x^k + i y^k, \tag{1}$$

où z^k sont les coordonnées complexes.

L'espace $\mathbf{C}^n$ peut être regardé comme un espace linéaire $\mathbf{R}^{2n}$ de dimension $2n$ sur le corps des nombres réels. La base dans $\mathbf{R}^{2n}$ se définit comme suit:

$$e_1, \ldots, e_n, \qquad i e_1, \ldots, i e_n. \tag{2}$$

C'est la formule de Gauss-Ostrogradski dans l'espace euclidien à trois dimensions.

b) Soient $k = 2$, U un domaine sur la surface $x^i = x^i(z^1, z^2)$, $i = 1, 2, 3$, et Γ le bord de ce domaine (une courbe). On a

$$\oint_\Gamma T_\alpha dx^\alpha = \iiint_U \left[\left(\frac{\partial T_1}{\partial x^2} - \frac{\partial T_2}{\partial x^1} \right) dx^1 \wedge dx^2 + \right.$$

$$\left. + \left(\frac{\partial T_1}{\partial x^3} - \frac{\partial T_3}{\partial x^1} \right) dx^1 \wedge dx^3 + \left(\frac{\partial T_2}{\partial x^3} - \frac{\partial T_3}{\partial x^2} \right) dx^2 \wedge dx^3 \right]. \quad (37)$$

Dans le cas euclidien, où l'on n'a pas à distinguer les vecteurs et les covecteurs et lorsqu'on peut associer à un tenseur alterné (T_{ij}) un vecteur $T = (T^i)$, on retrouve la formule de Stokes:

$$\oint_\Gamma T_\alpha dx^\alpha = \iint_U \langle \text{rot } T, n \rangle \sqrt{|g|} dz^1 \wedge dz^2, \quad (38)$$

où rot T est le vecteur associé au tenseur alterné $(\text{rot } T)_{\alpha\beta} = \frac{\partial T_\alpha}{\partial x^\beta} - \frac{\partial T_\beta}{\partial x^\alpha}$.

On constate donc que dans tous les cas envisagés la formule générale de Stokesse réduit, au moyen du théorème 1, à des formules intégrales différentes connues du cours d'Analyse, ce qui achève la démonstration pour l'espace à trois dimensions.

Notons en conclusion que l'énoncé de la formule générale de Stokes n'implique pas nécessairement que le bord Γ soit constitué d'un seul morceau. Si Γ se compose de plusieurs morceaux (fig. 31), les intégrales prises sur les morceaux différents interviendront avec leur signe + ou — ; le seul problème consiste à choisir correctement le signe. Nous l'avons déjà vu dans le cas trivial de $k = 1$, lorsque le bord Γ du segment de courbe est formé de deux points: un point (extrémité) de signe + et un point (origine) de signe — . Il y a lieu de remarquer ici que le choix du signe dans l'expression $\iint_\Gamma \langle T, n \rangle \sqrt{|g|} dz^1 \wedge dz^2$ (ou, comme l'on dit, de l'« orientation » du bord Γ) est défini par le sens de la normale unité n.

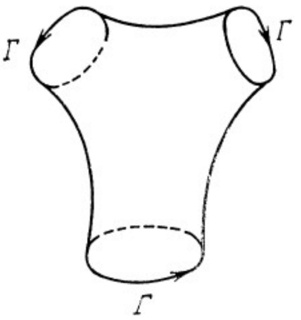

Fig. 31.

Citons une application importante de la formule générale de Stokes.

Considérons l'espace à quatre dimensions $x^0 = ct,\ x^1, x^2, x^3$ muni de la métrique $g_{ij} = \begin{pmatrix} +1 & & & 0 \\ & -1 & & \\ & & -1 & \\ 0 & & & -1 \end{pmatrix}$, où c est la vitesse de la lumière et t le temps.

Soit $F_{ik} = - F_{ki}$ le tenseur du champ électromagnétique, $i, k = 0, 1, 2, 3$. Les propriétés de ce tenseur seront étudiées en laissant le temps $x^0 = ct$ inchangé et en faisant varier seules les coordonnées spatiales x^i, $i = 1, 2, 3$:

$$x'^0 = x^0, \qquad x'^i = x'^i(x^1, x^2, x^3).$$

Dans ce cas le tenseur (F_{ik}) se définit dans l'espace à quatre dimensions par le covecteur du champ électrique $E_\alpha = F_{0\alpha}$, $\alpha = 1, 2, 3$, et le tenseur du champ magnétique $H_{\alpha\beta} = - H_{\beta\alpha} = F_{\alpha\beta}$, $\alpha, \beta = 1, 2, 3$. Si les coordonnées x^1, x^2, x^3 sont euclidiennes, le champ magnétique se définit par le vecteur du champ magnétique (§ 21)

$$H^1 = H_{23}, \qquad H^2 = - H_{13}, \qquad H^3 = H_{12}.$$

Dans le langage des formes différentielles, on a (voir § 25)

$d(F_{ij} dx^i \wedge dx^j) = 0$ (premier couple d'équations de Maxwell),

ou, sous forme euclidienne à trois dimensions,

$$\text{a)} \qquad \operatorname{div} \mathbf{H} = \frac{\partial H^i}{\partial x^i} = 0;$$

$$\text{b)} \qquad \operatorname{rot} \mathbf{E} + \frac{1}{c} \frac{\partial \mathbf{H}}{\partial t} = 0.$$

De l'équation a) et de la formule de Gauss-Ostrogradski on déduit (Γ étant le bord d'un domaine U)

$$\iiint_U \operatorname{div} \mathbf{H}\, dx^1 \wedge dx^2 \wedge dx^3 = \iint_\Gamma \langle \mathbf{H} \cdot n \rangle d\sigma = 0$$

(« le flux du champ magnétique à travers une surface fermée est toujours égal à zéro »).

De l'équation b) et de la formule de Stokes on déduit

$$\iint_U \langle \operatorname{rot} \mathbf{E}, n \rangle d\sigma = \oint_\Gamma E_\alpha\, dx^\alpha$$

(Γ étant le bord du domaine U sur la surface),

$$-\iint_U \left\langle \frac{1}{c}\frac{\partial \mathbf{H}}{\partial t}, n \right\rangle d\sigma = \oint_\Gamma E_\alpha \, dx^\alpha$$

(« la dérivée par rapport au temps du flux du champ magnétique à travers une surface est égale à la circulation du champ électrique le long du bord de la surface »).

Le deuxième couple d'équations de Maxwell (voir § 25) se présente comme suit :

$$\sum_{k=1}^{3} \frac{\partial F_{ik}}{\partial x^k} - \frac{1}{c}\frac{\partial F_{0i}}{\partial t} = \frac{4\pi}{c} j_{(4)i},$$

où $j_{(4)} = (c\rho, j_1, j_2, j_3)$, ρ est la densité de la charge, $\mathbf{j} = (j_1, j_2, j_3)$ le vecteur courant tridimensionnel (la divergence de dimension quatre du tenseur F_{ik} pour la métrique $\begin{pmatrix} +1 & & & 0 \\ & -1 & & \\ & & -1 & \\ 0 & & & -1 \end{pmatrix}$ est égale au produit du vecteur courant de dimension quatre par $4\pi/c$).

Sous forme tridimensionnelle, on a

a) $\operatorname{div} \mathbf{E} = 4\pi\rho$;

b) $\operatorname{rot} \mathbf{H} + \dfrac{1}{c}\dfrac{\partial \mathbf{E}}{\partial t} = \dfrac{4\pi}{c}\mathbf{j}$.

De a) et de la formule de Gauss-Ostrogradski on déduit

$$\iiint_U 4\pi\rho \, dx^1 \wedge dx^2 \wedge dx^3 = \iint_\Gamma \langle \mathbf{E}, n \rangle d\sigma$$

(« le flux du champ électrique à travers le bord d'un domaine est égal — à 4π près — à la charge totale dans ce domaine »).

De b) et de la formule de Stokes on déduit

$$-\frac{1}{c}\iint_U \frac{\partial \mathrm{E}}{\partial t} n \, d\sigma + \iint_U \frac{4\pi}{c} \mathbf{j} n \, d\sigma = \oint_\Gamma H_\alpha \, dx^\alpha,$$

où Γ est le bord du domaine U sur la surface (« le courant total à travers une surface diminué de la dérivée du flux du champ électrique à travers cette surface est égal à la circulation du champ magnétique le long du bord de la surface »).

Nous voyons que le premier et le deuxième couple d'équations de Maxwell revêtent ici des significations géométriques différentes. Le premier couple d'équations de Maxwell n'implique aucune métrique particulière dans l'espace, tandis que le deuxième couple n'existe pas en dehors de la métrique. La forme usuelle de ces équations est étroitement liée aux coordonnées euclidiennes.

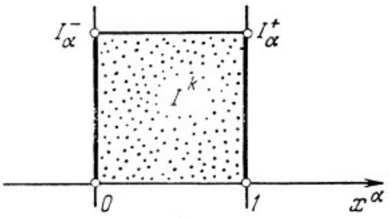

Fig. 32.

4. Démonstration de la formule générale de Stokes pour le cube. Nous venons de montrer que les formules intégrales classiques de Green et de Gauss-Ostrogradski sont des cas particuliers de la formule générale de Stokes. Dans ce n°, nous nous appliquerons à démontrer la formule générale de Stokes pour le cas où l'intégration s'étend à un cube de dimension k.

Le *cube singulier* σ dans l'espace $\mathbf{R}^n$ est une application différentiable

$$\sigma : I^k \to \mathbf{R}^n, \qquad n \geqslant k,$$

où I^k est un cube cartésien de dimension k:

$$I^k = \{(x^1, \ldots, x^k) | \, 0 \leqslant x^\alpha \leqslant 1\}.$$

Les équations $x^\alpha = 0$ et $x^\alpha = 1$ définissent deux faces à $(k-1)$ dimensions I_α^- et I_α^+ respectivement (fig. 32). Désignons par ∂I^k le bord du cube I^k, en sorte que $\partial I^k = \bigcup\limits_\alpha (I_\alpha^+ \cup I_\alpha^-)$. Soit φ^{k-1} une $(k-1)$-forme dans $\mathbf{R}^n$, et soit $d\varphi^{k-1}$ sa différentielle extérieure; soit enfin $\sigma : I^k \to \mathbf{R}^n$ le cube singulier.

THÉORÈME 4. *On a l'égalité*

$$\int\limits_{\sigma(\partial I^k)} \varphi^{k-1} = \int\limits_{\sigma(I^k)} d\varphi^{k-1}$$

(l'orientation du bord ∂I^k étant induite par l'orientation naturelle du cube I^k, compte tenu du sens de la normale extérieure).

Remarque. Dans cette formule, on entend par l'intégrale $\int\limits_{\sigma(\partial I^k)} \varphi$ la somme prise pour toutes les faces du cube.

Démonstration. Considérons la forme $\omega = \sigma^*(\varphi)$; les opérations σ^* et d étant permutables, on a $d\omega = d\sigma^*(\varphi) = \sigma^*(d\varphi)$, ce qui veut dire qu'il suffit de démontrer la proposition avancée sous la forme

$$\int\limits_{\partial I^k} \omega = \int\limits_{I^k} d\omega.$$

Nous nous sommes servi du fait qu'on a par définition

$$\int\limits_{\sigma(I^k)} d\varphi = \int\limits_{I^k} \sigma^*(d\varphi) \qquad \text{et} \qquad \int\limits_{\sigma(\partial I^k)} \varphi = \int\limits_{\partial I^k} \sigma^*(\varphi).$$

Soit $\omega = a_\alpha(x^1, \ldots, x^k)\, dx^1 \wedge \ldots \wedge \widehat{dx^\alpha} \wedge \ldots \wedge dx^k$, où $a_\alpha(x^1, \ldots, x^k)$ sont des fonctions différentiables; la différentielle dx^α est omise. Il vient alors

$$d\omega = \sum_\alpha \frac{\partial a_\alpha}{\partial x^\alpha}\, dx^\alpha \wedge dx^1 \wedge \ \ldots \wedge \widehat{dx^\alpha} \wedge \ldots \wedge dx^k = \sum_\alpha (-\, 1)^{\alpha-1} \frac{\partial a_\alpha}{\partial x^\alpha}\, d^k x,$$

où $d^k x = dx^1 \wedge \ldots \wedge dx^k$. Pour simplifier le raisonnement qui va suivre, mettons les fonctions $a_\alpha(x^1, \ldots, x^k)$ sous la forme de produits:

$$a_\alpha(x^1, \ldots, x^k) = \prod_{q=1}^{k} b_\alpha^q(x^q),$$

ou les fonctions b_α^q sont des fonctions différentiables d'une variable unique x^q. (Rappelons un théorème démontré dans le cours d'Analyse: toute fonction différentiable se laisse uniformément approcher par des combinaisons linéaires de produits de fonctions différentiables d'une même variable; il va sans dire que la démonstration de ce théorème n'a pas sa place ici.)

Explicitons $\int\limits_{I^k} d\omega$:

$$\int\limits_{I^k} d\omega = \int\limits_{I^k} \Bigl(\sum_\alpha (-\, 1)^{\alpha-1} \frac{\partial a_\alpha}{\partial x^\alpha} \Bigr)\, d^k x =$$

$$= \int\limits_{I^k} \Bigl(\sum_\alpha (-\, 1)^{\alpha-1} \frac{\partial}{\partial x^\alpha} \Bigl(\prod_{q=1}^{k} b_\alpha^q(x^q) \Bigr) \Bigr)\, d^k x =$$

$$= \int\limits_{I^k} \Bigl(\sum_\alpha (-\, 1)^{\alpha-1}\, b_\alpha^1(x^1) \ldots \frac{\partial b_\alpha^\alpha(x^\alpha)}{\partial x^\alpha} \ldots\ b_\alpha^k(x^k) \Bigr)\, d^k x =$$

$$= \sum_\alpha \int\limits_{(x^1)} \cdots \int\limits_{(\hat{x}^\alpha)} \cdots \int\limits_{(x^k)} ((-\, 1)^{\alpha-1} b_\alpha(x^1) \ldots b_\alpha^{\alpha-1}(x^{\alpha-1}) b_\alpha^{\alpha+1}(x^{\alpha+1}) \ldots$$

$$\ldots b_\alpha^k(x^k) \Bigl[(-\, 1)^{\alpha-1} \int\limits_{(x^\alpha)} \frac{\partial b_\alpha^\alpha(x^\alpha)}{\partial x^\alpha}\, dx^\alpha \Bigr]\, dx^1 \wedge \ldots \wedge \widehat{dx^\alpha} \wedge \ldots \wedge dx^k =$$

$$= \sum_\alpha \int\limits_{(x^1)} \cdots \int\limits_{(\hat{x}^\alpha)} \cdots \int\limits_{(x^k)} (b_\alpha^1(x^1) \ldots \hat{b}_\alpha^\alpha(x^\alpha) \ldots$$

$$\ldots b_\alpha^k(x^k)[b_\alpha^\alpha(1) - b_\alpha^\alpha(0)])\, dx^1 \wedge \ldots \wedge \widehat{dx^\alpha} \wedge \ldots \wedge dx^k =$$

$$= \sum_\alpha \int_{(x^1)} \cdots \int_{(\hat{x}^\alpha)} \cdots \int_{(x^k)} (b_\alpha^1(x^1) \ldots b_\alpha^\alpha(1) \ldots b_\alpha^k(x^k) -$$

$$- b_\alpha^1(x^1) \ldots b_\alpha^\alpha(0) \ldots b_\alpha^k(x^k))\, dx^1 \wedge \ldots \wedge \widehat{dx^\alpha} \wedge \ldots \wedge dx^k =$$

$$= \sum_\alpha \int_{(x^1)} \cdots \int_{(\hat{x}^\alpha)} \cdots \int_{(x^k)} (a_\alpha(x^1, \ldots, x^k)|_{x^\alpha = 1} -$$

$$- a_\alpha(x^1, \ldots, x^k)|_{x^\alpha = 0}\, dx^1 \wedge \ldots \wedge \widehat{dx^\alpha} \wedge \ldots \wedge dx^k = \int_{\partial I^k} \omega.$$

Le théorème est démontré.

Exercices. 1. Calculer le volume des groupes $SO(3, \mathbf{R})$ et $SU(2)$ pour la métrique de Killing.

2. Soit X^i un champ de vecteurs dans l'espace de Minkowski $\mathbf{R}_1^4$. Définissons l'intégrale de ce champ de vecteurs sur l'hypersurface à trois dimensions (le «flux» du champ de vecteurs à travers l'hypersurface) comme l'intégrale d'une 3-forme $X^i dS_i$ avec

$$dS_i = \frac{1}{6} \sqrt{|g|}\, \varepsilon_{jkli}\, dx^j \wedge dx^k \wedge dx^l$$

et ε_{jkli}, tenseur antisymétrique de rang 4. Déduire de la formule générale de Stokes l'égalité suivante:

$$\int_{\partial V} X^i\, dS_i = \int_V \frac{\partial X^i}{\partial x^i}\, dV.$$

3. Soient U un domaine borné à bord différentiable dans l'espace muni de la métrique riemannienne g_{ij} et Ω_U^p l'espace des formes différentiables qui s'annulent en dehors de U. Définissons le produit scalaire dans Ω_U^p en posant

$$\langle \omega_1, \omega_2 \rangle = \int_U \omega_1 \wedge * \omega_2$$

pour les p-formes ω_1, ω_2 de Ω_U^p.

2) Montrer que Ω_U^p muni du produit scalaire est un espace euclidien (voir § 19, exercice 2).

b) L'opérateur $*$ est orthogonal:

$$\langle * \omega_1, * \omega_2 \rangle = \langle \omega_1, \omega_2 \rangle.$$

c) Les opérateurs $\delta = (-1)^{np+n+1} * d *$ et d sont adjoints entre eux, en sorte que

$$\langle d\omega_1, \omega_2 \rangle = \langle \omega_1, \delta\omega_2 \rangle.$$

d) Le carré de l'opérateur δ est nul, $\delta\delta = 0$.

e) Soit $\Delta = d\delta + \delta d$. Montrer que l'opérateur Δ est autoadjoint dans $\Omega^p_U : \langle \Delta\omega_1, \omega_2 \rangle = \langle \omega_1, \Delta\omega_2 \rangle$. Vérifier que les relations de commutativité suivantes ont lieu:

$$\Delta d = d\Delta, \qquad \Delta\delta = \delta\Delta, \qquad \Delta * = * \Delta.$$

§ 27. Formes différentielles dans les espaces complexes

1. Opérateurs d′ et d″. Soit D un domaine dans l'espace complexe $\mathbf{C}^n$ muni de coordonnées complexes $z^1, \ldots, z^n$. On peut y distinguer un domaine «décomplexifié» $D^{\mathbf{R}}$ muni de coordonnées réelles $x^1, \ldots, x^n, \; y^1, \ldots, y^n$ dans lequel

$$z^k = x^k + iy^k, \qquad k = 1, \ldots, n. \tag{1}$$

La base dans l'espace tangent à $D^{\mathbf{R}}$ (en un point quelconque) est constituée par les vecteurs $\frac{\partial}{\partial x^1}, \ldots, \frac{\partial}{\partial x^n}, \; \frac{\partial}{\partial y^1}, \ldots, \frac{\partial}{\partial y^n}$. Considérons les vecteurs tangents complexes, c'est-à-dire les combinaisons linéaires

$$\sum_{k=1}^{n} a^k \frac{\partial}{\partial x^k} + \sum_{k=1}^{n} b^k \frac{\partial}{\partial y^k} \tag{2}$$

à coefficients complexes. Il est commode d'introduire dans l'espace des vecteurs de cette espèce une base complexe $\frac{\partial}{\partial z^k}, \frac{\partial}{\partial \bar{z}^k}$ en posant

$$\frac{\partial}{\partial z^k} = \frac{1}{2}\left(\frac{\partial}{\partial x^k} - i\,\frac{\partial}{\partial y^k}\right), \quad \frac{\partial}{\partial \bar{z}^k} = \frac{1}{2}\left(\frac{\partial}{\partial x^k} + i\,\frac{\partial}{\partial y^k}\right), \quad k = 1, \ldots, n. \tag{3}$$

Afin de faciliter la compréhension des formules qui seront établies par la suite, nous adopterons une convention d'écriture qui consiste à surligner toujours les indices tensoriels correspondant aux coordonnées surlignées, exception faite pour les lettres qui étaient surlignées auparavant. C'est ainsi que la notation $a_{k\bar{l}} z^k \bar{z}^l$ implique la sommation pour les deux indices (voir aussi les formules (4), (7), (9), etc. ci-après).

Tout vecteur complexe ξ peut donc se présenter comme suit:

$$\xi = \xi^k \frac{\partial}{\partial z^k} + \xi^{\bar{k}} \frac{\partial}{\partial \bar{z}^k}, \qquad \xi = (\xi^k, \; \xi^{\bar{k}})\,; \tag{4}$$

CHAPITRE 3

DEGRÉ DE L'APPLICATION. INDICE D'INTERSECTION. APPLICATIONS DE CES NOTIONS

§ 12. Notion d'homotopie

1. Définition de l'homotopie. Possibilité de l'approximation différentiable. Soient deux variétés différentiables N, M (par exemple de classe C^∞) et une application différentiable $f: N \to M$.

DÉFINITION 1. On entend par *homotopie* (ou *déformation*) *différentiable* (resp. *différentiable par morceaux*, *continue*) d'une application f, une application différentiable (resp. différentiable par morceaux, continue) du cylindre

$$F: N \times I \to M \qquad (I = [0, 1])$$

telle que, pour tout $x \in N$, $F(x, 0) = f(x)$. On dit que les applications $f_t: N \to M$, $f_t(x) = F(x, t)$ sont *homotopes* à l'application initiale $f = f_0$; quant à l'application du cylindre F dans son ensemble, elle est appelée *homotopie* ou « opération d'homotopie ».

L'ensemble des applications homotopes à f forme une classe d'applications homotopes deux à deux, ou classe d'homotopie. Il va sans dire que les classes d'homotopie différentiable existent pour tout degré de différentiabilité l. En particulier, pour $l = 0$ on a une classe d'applications continues et continûment homotopes. D'autre part, le théorème selon lequel toute application continue peut être approchée par une application différentiable donne lieu à quelques propriétés élémentaires de l'homotopie:

1) toute application continue se laisse approcher par une application différentiable de classe quelconque (C^∞ entre autres) qui est voisine de l'application donnée et homotope à celle-ci;

2) deux applications continues sont différentiablement homotopes chaque fois qu'elles sont continûment homotopes.

Examinons ces propriétés plus en détail.

THÉORÈME 1. *Soient deux variétés différentiables compactes M, N et deux applications continues f, g: $M \to N$. Alors, étant donnée une métrique riemannienne quelconque sur N, il existe un $\varepsilon > 0$ tel que la condition $\rho(f, g) < \varepsilon$ implique l'homotopie de f et g (ρ est la distance des deux applications, voir plus haut).*

Démonstration. Munissons M d'une métrique riemannienne et désignons par ρ_0 la distance définie par cette métrique. Puisque N est une variété compacte, il existe un $\varepsilon > 0$ tel que pour $p, q \in N$ et $\rho_0(p, q) < \varepsilon$ les points p et q peuvent être reliés par une seule géodésique réalisant le plus court chemin de p à q (voir Première partie, § 29). Soient maintenant

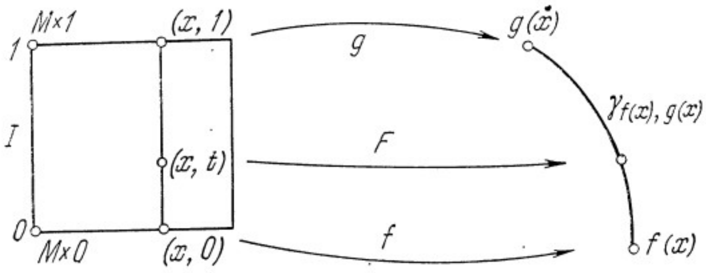

Fig. 28.

deux applications continues $f, g: M \to N$ telles que $\rho\,(f, g) < \varepsilon$. Faisons l'homotopie de $F: M \times I \to N$ avec $F(x, 0) = f(x)$, $F(x, 1) = g(x)$. A cet effet, définissons la valeur que prend $F(x, t)$ en un point quelconque $(x, t) \in M \times I$. Considérons deux points $f(x)$, $g(x) \in N$. Puisque $\rho_0(f(x), g(x)) < \varepsilon$, le plus court chemin entre ces points sera réalisé par une géodésique unique $\gamma_{f(x), g(x)}(\tau)$. Puisque nous connaissons non seulement x mais aussi t, nous pouvons diviser le segment de la géodésique $\gamma_{f(x), g(x)}$ entre $f(x)$ et $g(x)$ dans le rapport $t: (1 - t)$ et identifier le point de division au point $(x, t) \in M \times I$ (fig. 28).

Le fait que l'application $F(x, t)$ ainsi construite est continue résulte du théorème selon lequel les solutions dépendent continûment des conditions initiales; ce théorème s'applique à tout système d'équations différentielles ordinaires, ce qui est précisément le cas pour les équations des géodésiques. Le théorème est démontré.

THÉORÈME 2. *Soit f une application continue quelconque d'une variété différentiable M dans une variété différentiable N. Cette application est homotope alors à une application différentiable $g: M \to N$.*

Démonstration. Nous avons montré plus haut que toute application continue $f: M \to N$ admet toujours une application différentiable aussi voisine que l'on veut. L'assertion du théorème résulte du théorème précédent.

Exercice. Montrer que si l'application f est déjà différentiable sur une certaine sous-variété $X \subset M$, on peut toujours l'approcher par une application différentiable g de telle sorte que les restrictions de g et de f à X soient identiques.

THÉORÈME 3. *Deux applications différentiables* f, $g: M \to N$ *continûment homotopes sont différentiablement homotopes.*

Ce théorème résulte de l'existence d'une approximation différentiable pour toute application continue (en l'occurrence, l'homotopie initiale $F: M \times I \to N$), ainsi que des théorèmes 1 et 2.

On démontre par analogie l'assertion suivante: *toute application continue d'une variété différentiable M dans une variété différentiable de dimension suffisamment grande est homotope à un plongement différentiable. Si l'application donnée est différentiable, il peut en être de même pour l'homotopie en question.*

DÉFINITION 2. Nous disons que deux plongements différentiables f, $g: M \to N$ sont (différentiablement) *isotopes* s'ils sont liés par une homotopie différentiable $F: M \times I \to N$ telle que l'application $f_t: M \to N$ définie par l'égalité $f_t(x) = F(x, t)$ reste un plongement différentiable pour tout $t \in [0, 1]$.

THÉORÈME 4. *Deux plongements différentiables quelconques f, g d'une variété différentiable M de dimension n dans un espace euclidien* $\mathbf{R}^q$ *de dimension q suffisamment élevée (plus exactement* $q \geqslant 2n + 2$*) sont isotopes.*

Démonstration. Si la dimension q est suffisamment grande, deux plongements f, g admettent toujours une homotopie continue (donc aussi différentiable) $F: M \times I \to \mathbf{R}^q$ (par exemple ils peuvent être tous les deux homotopes à une application ponctuelle). Ceci étant, on peut transformer l'homotopie différentiable $F: M \times I \to N$ en un plongement par une déformation différentiable (la déformation étant fixe sur les « bases » $M \times 0$ et $M \times 1$ du cylindre), ce qui nous conduit justement à l'isotopie de f et g.

Grâce aux théorèmes démontrés, nous cesserons désormais de faire la distinction entre homotopie continue et homotopie différentiable. Dans le texte qui suit, nous considérerons des classes d'homotopie continue, différentiable par morceaux ou continûment différentiable de classe C^∞, suivant le cas.

La classe d'homotopie d'une application f se note $[f]$; l'ensemble des classes d'homotopie des applications $N \to M$ se note $[N; M]$.

2. Homotopies relatives. Dans le texte qui suit, nous rencontrerons des homotopies et des classes d'homotopie auxquelles on a imposé certaines contraintes. Citons-en quelques-unes:

a) Soient deux variétés N et M dont chacune présente un point de base $x_0 \in N$ et $y_0 \in M$. On demande que x_0 se transforme en y_0 par toute application $f: N \to M$ (de même que par toute homotopie). On dit alors que la classe d'applications (classe d'homotopie) est *liée en un point*.

b) Soient deux variétés à bord N et M de bords $\Gamma = \partial N$ et $\Delta = \partial M$. On demande que le bord Γ se transforme en le bord Δ par toute application $f: N \to M$ et par toute homotopie. Une telle classe d'applications est dite *relative* (relativement aux bords).

c) Soient deux variétés N, M non compactes (ouvertes). On demande que l'image réciproque d'un point quelconque par l'application $f: N \to M$ soit compacte. La même condition est demandée pour les homotopies $F: N \times I \to M$. Une telle classe d'applications (d'homotopie) est dite *propre.*

d) La classe la plus générale d'applications relatives est la suivante. Soit, dans chacune des variétés N et M, un ensemble de base $A \subset N$ et $B \subset M$. On demande que B soit l'image de A par toute application f et par toute homotopie F. L'ensemble des classes d'homotopie correspondantes se note $[(N, A); (M, B)]$.

§ 13. Degré de l'application

1. Définition du degré. Dans ce paragraphe, nous examinerons surtout les classes d'homotopie des applications de deux fermés orientés N et M de même dimension n, particulièrement lorsque M est une sphère. Soit une application différentiable $f: N \to M$. Choisissons un point $y_0 \in M$ et supposons que f soit une application régulière pour $y_0 \in M$. Cela revient à dire que l'image réciproque de y_0 est constituée par un nombre fini de points x_i $(i = 1, \ldots, m)$ de la variété N aux coordonnées locales (x_i^β) dans le voisinage des x_i et (y_0^α) dans le voisinage de y_0, les jacobiens det $\left(\dfrac{\partial y_0^\alpha}{\partial x_i^\beta}\right)$ étant non nuls, $i = 1, \ldots, m$. Rappelons que les variétés N et M sont munies d'une orientation, c'est-à-dire que tous les jacobiens de passage des coordonnées locales d'un voisinage à celles d'un autre sont positifs.

DÉFINITION 1. Le *degré* d'une application différentiable $f: N \to M$ de deux variétés fermées orientées connexes en un point régulier par rapport à une valeur régulière $y_0 \in M$ est la quantité

$$\deg f = \sum_{f(x_i) = y_0} \operatorname{sgn} \det\left(\frac{\partial y_0^\alpha}{\partial x_i^\beta}\right)$$

(les signes des jacobiens des applications étant bien déterminés dans tous les x_i).

On a le théorème important que voici:

THÉORÈME 1. *Le degré d'une application est indépendant de la valeur régulière choisie y_0 et est invariant par toute homotopie.*

Démonstration. Montrons d'abord que le degré ne dépend pas du choix de la valeur régulière. Ce résultat est immédiat pour toute valeur régulière voisine de y_0, car les images réciproques sont au même nombre et de même signe.

Soient y_0 et y_1 deux valeurs régulières de l'application f. Relions-les par un chemin différentiable non croisé sur la variété M à tangente non dégénérée: ce chemin représente une variété à une dimension γ. En vertu du

(/ /$ < 211 ! ! 1 P

cations $M \to \mathbf{R}^{2n}$. Dans le cas général, on en reste là. C'est ainsi qu'une variété fermée non orientable à deux dimensions ne peut pas être plongée dans $\mathbf{R}^3$ (vérifier!).

Il y a cependant des cas où l'on peut faire mieux: en particulier, toute variété orientable à deux dimensions peut se réaliser dans $\mathbf{R}^3$, ce qui ressort de la classification des variétés pareilles.

2. Fonctions de Morse comme fonctions de la hauteur. Le plongement $M \to \mathbf{R}^N$ peut fournir une nouvelle démonstration de l'existence des fonctions de Morse sur des variétés compactes différentiables. Nous montrerons en effet que les fonctions de Morse peuvent être recherchées dans une classe assez restreinte de fonctions sur M, dites «fonctions de la hauteur».

Etant donné un plongement différentiable de M dans $\mathbf{R}^N$ et une droite $\xi_l(t)$ parallèle à un vecteur l et passant par l'origine des coordonnées dans $\mathbf{R}^N$, considérons une fonction $h_l(x)$ (*fonction de la hauteur*) dont la valeur en $x \in M$ est égale à la projection orthogonale de x sur $\xi_l(t)$ (on suppose que $\xi(t)$ est la droite numérique réelle).

La fonction de la hauteur $h_l(x)$ possède quelques propriétés évidentes (d'ailleurs faciles à vérifier):

1) L'ensemble des fonctions de la hauteur correspond biunivoquement aux couples de points diamétralement opposés de la sphère S^{N-1} ou aux points de l'espace projectif $\mathbf{R}P^{N-1}$.

2) Le point $x_0 \in M$ est un point critique de la fonction de la hauteur $h_l(x)$ si et seulement si le vecteur l est orthogonal à M en x_0 (c'est-à-dire si $l \perp T_{x_0}M$).

Voyons dans quelles conditions le point critique $x_0 \in M$ est un point non dégénéré pour $h_l(x)$. Plaçons-nous dans le cas particulier où M est plongé dans $\mathbf{R}^{m+1}$ avec $m = \dim M$ (à la façon d'une hypersurface).

Rappelons que pour toute hypersurface $M \subset \mathbf{R}^{m+1}$ il existe une application gaussienne $r\colon M \to S^m \to \mathbf{R}P^m$ telle que $r(x) = n(x)$, où $n(x) \perp T_xM$ est le vecteur unité de la normale à M en x (le vecteur $n(x)$ étant transporté par parallélisme à l'origine des coordonnées $0 \in \mathbf{R}^{m+1}$).

LEMME 1. *Le point* $x_0 \in M \subset \mathbf{R}^{m+1}$ *est un point critique non dégénéré pour la fonction de la hauteur* $h_l(x)$ *si et seulement si* x_0 *est en même temps un point régulier de l'application gaussienne* $r\colon M \to \mathbf{R}P^m$, *avec* $l \perp T_{x_0}M$.

Démonstration. Orientons l'axe de coordonnée x^{m+1} parallèlement au vecteur l, et les axes de coordonnées $x^1, \ldots, x^m$ perpendiculairement à l; le plan $\mathbf{R}^m(x^1, \ldots, x^m)$ sera alors tangent à M en x_0. Dans le voisinage du point $x_0 \in M$, la variété M admet pour équation $x^{m+1} = \varphi(x^1, \ldots$ $\ldots, x^m)$, avec $d\varphi\,|_{x_0} = 0$. Dans un voisinage suffisamment petit U de x_0, on peut admettre que $(x^1, \ldots, x^m)$ soient des coordonnées locales introduites sur l'hypersurface M, tandis que la «hauteur» $h_l(x)$ soit précisément $x^{m+1} =$ $= \varphi(x^1, \ldots, x^m)$. Des coordonnées analogues $y^1, \ldots, y^m$ seront introduites sur la sphère dans le voisinage du point l. Un calcul analogue à celui qui a été entrepris pour déduire la formule $K d\sigma = f^*\Omega$ montre qu'on a dans le point x_0, en coordonnées locales $(x^1, \ldots, x^m)$ et $(y^1, \ldots, y^m)$, l'égalité

$$\left(\frac{\partial^2 h_l(x_0)}{\partial x^\alpha \partial x^\beta}\right) = \left(\frac{\partial^2 \varphi(x_0)}{\partial x^\alpha \partial x^\beta}\right) = \left(\frac{\partial y^\alpha}{\partial x^\beta}\bigg|_{x_0}\right);$$

$$\det\left(\frac{\partial^2 h_l(x_0)}{\partial x^\alpha \partial x^\beta}\right) = K.$$

Ici K est la courbure totale de l'hypersurface au point x_0, et $y^\alpha = r^\alpha(x^1, \ldots$ $\ldots, x^m)$, $1 \leqslant \alpha \leqslant m$, est l'application gaussienne transcrite au moyen des coordonnées (x) et (y). Ainsi donc, si l'on veut que l'application gaussienne r admette x_0 comme point régulier, $\det\left(\frac{\partial y^\alpha}{\partial x^\beta}\bigg|_{x_0}\right) \neq 0$, il faut et il suffit que la hessienne de la fonction de la hauteur $h_l(x_0)$ soit non dégénérée, $\det\left(\frac{\partial^2 h_l(x_0)}{\partial x^\alpha \partial x^\beta}\right) \neq 0$. Le lemme est démontré.

Remarque importante. Etant donné un plongement $M \subset \mathbf{R}^q$, où $q > m + 1$ ($m = \dim M$), il existe une variété différentiable N à $q - 1$ dimensions qui constitue la frontière du «voisinage tubulaire» de M. Considérons en effet l'ensemble des disques D^{q-m}_x de dimension $q - m$ orthogonaux à M, de centre en des points $x \in M$ et de rayon $\varepsilon > 0$ suffisamment petit. La réunion de ces boules (avec $\varepsilon > 0$ petit) forme une variété q-dimensionnelle appelée quelquefois voisinage tubulaire de M; c'est sa frontière (bord) qu'on désigne par N. L'image de cette sous-variété (plongée dans $\mathbf{R}^q$ pour ε petit) par l'application gaussienne r est une sphère S^{q-1}. Soit $h_l(x)$ une fonction de la hauteur sur M et sur N. Il est facile de voir (vérifier!) que, pour tout point critique $x_0 \in M$ de la fonction $h_l(x)$, on a exactement deux points critiques y_0, y_0' sur N: ce sont les points où N vient couper la droite l qui passe par x_0 et qui est orthogonale à M (fig. 27).

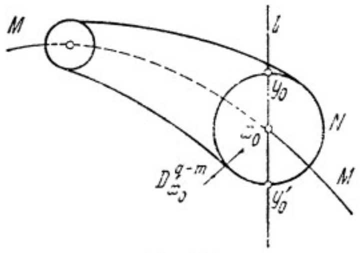

Fig. 27.

On vérifie sans peine ensuite que le point critique $x_0 \in M$ de la fonction $h_l(x)$ est non dégénéré si et seulement si les points y_0, y_0' sont non dégénérés pour $h_l(x)$ (les points y_0 et y_0' sont tous les deux dégénérés ou non). Si donc $h_l(x)$ est une fonction de Morse sur N, elle l'est aussi sur M. Il s'ensuit que toute proposition selon laquelle les fonctions de Morse existent et sont partout denses dans l'ensemble des fonctions de la hauteur sur N est toujours vraie pour M. Cela prouve qu'on ne diminue pas la généralité en se bornant à considérer, dans le lemme ci-dessus, une hypersurface $M \subset \mathbf{R}^{m+1}$.

THÉORÈME 2. *La fonction de la hauteur* $h_l(x)$ *sur l'hypersurface* $M \subset \mathbf{R}^{m+1}$, $m = \dim M$, *est une fonction de Morse si et seulement si le point* $(\pm l) \in \mathbf{R}P^m$ *est une valeur régulière de l'application gaussienne* $r: M \to \mathbf{R}P^m$. *En particulier, presque toutes les fonctions de la hauteur* $h_l(x)$ *sont des fonctions de Morse.*

Démonstration. La première assertion du théorème découle du lemme 1. La seconde résulte du lemme de Sard: en effet, les valeurs régulières de l'application gaussienne $r: M \to \mathbf{R}P^m$ sont partout denses dans $\mathbf{R}P^m$. Le théorème est démontré.

Remarque. En vertu des observations supplémentaires faites ci-dessus au sujet de $M \subset \mathbf{R}^q$ pour $q > m + 1$, le théorème reste valable pour toute sous-variété différentiable dans $\mathbf{R}^q$ de codimension quelconque (donc pour tout $q - m > 0$).

3. Points focaux. Il y a beaucoup d'autres procédés qui permettent de construire sans peine de nombreuses fonctions de Morse sur des variétés différentiables. Sans entrer dans le détail, donnons la description d'un procédé.

Considérons une variété différentiable M et son plongement différentiable quelconque dans $\mathbf{R}^q$. Choisissons une fois pour toutes un point $p \in \mathbf{R}^q$ et associons à ce point une fonction différentiable L_p construite sur le plongement M^m en posant $L_p(x) = |p - x|^2$, où $x \in M$ et $|p - x|$ est la longueur du vecteur $p - x$. On montre que $L_p(x)$ est une fonction de Morse sur M pour presque tous les points $p \in \mathbf{R}^q$. L'ensemble des fonctions du type $L_p(x)$ ne se confond pas avec celui des fonctions de la hauteur $h_l(x)$.

On veut savoir pour quels points p la fonction $L_p(x)$ est une fonction de Morse. Désignons par N l'ensemble des couples (x, v) où $x \in M$ et $v \in \mathbf{R}^q$ est un vecteur orthogonal à M en x. Il est évident que N est une variété différentiable à q dimensions (vérifier!). Considérons une application différentiable $f: N \to \mathbf{R}^q$ qui à tout couple $(x, v) \in N$ fait correspondre l'extrémité d'un vecteur ayant son origine en x.

DÉFINITION 1. On dit que le point $p \in \mathbf{R}^q$ est un *point focal de multiplicité* $\mu > 0$ pour M si $p = f(x, v)$ et le jacobien de f en (x, v) est de rang $q - \mu$.

En vertu du lemme de Sard, on ne trouve presque pas de point focal pour le plongement $M \subset \mathbf{R}^q$ parmi les points $p \in \mathbf{R}^q$; en particulier, les points focaux $p \in \mathbf{R}^q$ forment un ensemble de mesure nulle.

Considérons à présent la «deuxième forme fondamentale du plongement $M \subset \mathbf{R}^q$». Supposons que la variété M soit (localement) définie dans $\mathbf{R}^q$ sous forme paramétrique $x = x(u^1, \ldots, u^m)$, où u^α, $1 \leqslant \alpha \leqslant m$, sont des coordonnées locales sur M; $x = (x^1, \ldots, x^q)$. Considérons un vecteur $\dfrac{\partial^2 x}{\partial u^i \partial u^j} = x_{ij}$; soit v la normale à M en $x \in M$. Il existe alors le produit sca-

(//

Pág. 101(99) Chapitre 3
Degré de l'application.
Indice d'intersection.
Applications de ces notions.

§12 Notion d'homotopie

Obs: Teoria da homotopia

é um estudo sistemático de situações em que os mapas vêm com homotopias entre eles. Sabe-se que, se existe uma função contínua entre os espaços topológicos, significa que são homotópicos.

Important: Use este conceito para formular uma equação na qual poderemos representar uma superfície curva (Gauss) em uma superfície plana, ou seja, esta equação irá representar uma superfície curva projetada ortogonalmente sobre uma superfície.

Usaremos a homotopia para representar a projeção do espaço curvo (espaço-tempo) sobre um plano.

Pág. 97(95) – Fig. 27

application gaussienne si est une sphère S^{n-1}.

- Modern Geometry ... Part I ... djvu
  (Novikov)
- Modern Geometry .. Part II ... djvu
- Modern Geometry ... Part III . djvu

Obs: Nestes 3 volumes há inúmeras referências ao Gauss
(estudá-las !!!)

cont.

20211010/ Géométrie Contemporaine .../
Stars-and-Relativity-Zel'dovich/
Stars and Relativity (2011).djvu

Pág. 89 (106) –
Gaussian curvature (see §1.4) of space, C_G at the center of star is larger than it is at other places.

$$C_G = (8\pi/3)(G\rho/c^2).$$

Pág. 28 (45) – Gaussian curvature for the three-dimensional space of the comoving reference frame in the static

$$C_G = \frac{8}{3}\pi(G/c^2)\rho$$

B%
! -256- - -* ! - B% ! ! " ! A 3 4 A
D ! 1"
! -256- - -* ! - B% ! ! " ! A 3 4 A
D ! "KA
(/ B 7! ! A! 1! J Z '

Actually, the Gaussian curvature (see § 1.4) of space, C_G, at the center of the star is larger than it is at other places. The quantity C_G is given by the formula (see eq. [1.8.13]):

$$C_G = (8\pi/3)(G\rho/c^2)\ . \tag{3.3.6}$$

Inasmuch as the density is a maximum at the center of the star, C_G is also a maximum there. Of course, we should not assume on the basis of equation (3.3.6) that space is Euclidean outside the star (where $\rho = 0$), when the gravitational field near the star is strong. Formula (3.3.6) gives only the mean curvature of space averaged over all two-dimensional directions; and this mean curvature actually is zero. However, beyond the star, the Riemannian space curvature (see § 1.4) is nonzero; and, depending on the two-dimensional direction, it may be positive or negative. At the center of the star, all directions are equivalent; there the curvature for any orientation is given by formula (3.3.6), and it is always positive.

The gravitational field obtained by joining the star's interior and exterior solutions has no physical singularities of the type exhibited on the Schwarzschild sphere; everywhere

$$1 \leq e^{\lambda} < \infty\ , \quad \text{and} \quad 0 < e^{\nu} < 1\ .$$

3.4 THE RADIAL MOTION OF LIGHT RAYS AND OF ULTRA-RELATIVISTIC PARTICLES

Consider now the radial motion, in a spherical gravitational field, of particles with the fundamental speed c; for example, photons and neutrinos. Since an observer can locally introduce coordinates for which $ds^2 = c^2dt^2 - dx^2 - dy^2 - dz^2$, and since the speed of light c is independent of reference frame, an equation of motion for a particle with zero rest mass is $ds = 0$ in any coordinate system. Because of spherical symmetry, a particle that starts out moving radially will always move radially. Along such a particle's trajectory θ and ϕ are constant, so the condition $ds = 0$ implies

$$dr/dt = ce^{(\nu-\lambda)/2}\ . \tag{3.4.1}$$

Everywhere inside the star $e^{(\nu-\lambda)/2} < 1$. Outside the star, in vacuum, $e^{(\nu-\lambda)/2} = 1 - 2GM/c^2r < 1$; and this quantity approaches unity as $r \to \infty$. It follows, for example, that for a neutrino emitted from the star's center, the radial coordinate r, as a function of the time t of a distant observer, must have the form shown in Figure 7. The dotted line in this diagram indicates the motion of the neutrino in the absence of a gravitational field.

In the future we will frequently consider the motion of bodies in the immediate vicinity of the gravitational radius of the central body (where the field is especially strong), and even in the limit $r \to r_g$. As noted in § 3.2, a static body cannot have dimensions $r \leq r_g$; and Bondi (1964) has demonstrated that the dimensions of a body in equilibrium cannot be smaller than (up to terms of first order in a). Expression (4.3.4) is identical with the one given by Landau and Lifshitz (1962) for a weak field. In reality, it is valid also in a strong field when $a \ll r_g$ (to first order in the dimensionless parameter a/r_g). The expression (4.3.4) for a stationary sphere was independently obtained by Gurovich (1965). For beautiful discussions of its relationship to Mach's Principle see Brill and Cohen (1966), and Cohen and Brill (1968).

In this solution for perturbations of the Schwarzschild field, only terms linear in a were retained in $h_{\mu\nu}$; terms proportional to a^2 and higher powers were neglected. Notice that the surfaces S_m and S_{hor} coincide in this case; they differ only by $O(a^2)$. Those aspects of the Kerr solution at $g_{00} = 0$ which depend on corrections to $g_{\mu\nu}$ which are linear in a are retained by this solution. In particular, here $C_{g_{00}} = 0 = 12/r_g{}^4 < \infty$.

Hartle and Thorne (1968) have derived the general solution for the exterior gravitational field of a slowly rotating, centrifugally deformed star with a precision of second order in the angular momentum $(a/r_g)^2$ and of first order in the quadrupole moment. Their solution has this curious property: it has no singularity at S_m or S_{hor} if and only if the quadrupole moment Q, the angular momentum K, and the star's mass M satisfy the algebraic identity

$$Q = K^2/M\,.$$

It is widely speculated that in order for the exact solution to have no real singularities at S_{hor}, every higher multipole moment must be linked to the angular momentum and mass by a relation similar to the above; and it is thought that the only exact solution with such a property (an absence of singularities at the event horizon) is probably the Kerr solution (cf. Hernandez 1967).

Can a solution with a "rotational" deviation from spherical symmetry be realized in nature for the entire domain down to $g_{00} = 0$?

Such a solution cannot be created by a stationary body for the same reason that it cannot in the case of a body with a quadrupole moment. However, for a nonstationary source we cannot apply the same reasoning as we did in § 4.2, because here in the "rotating" Kerr solution the invariant C is finite at $g_{00} = 0$ and at the event horizon, S_{hor}. It will be demonstrated later that a stationary solution with "rotational" perturbations and without a singularity at the event horizon can actually be realized, throughout the domain down to the horizon, as the asymptotic solution for the collapse of a rotating body.

4.4 THE SCHWARZSCHILD SPHERE IN AN EXTERNAL QUADRUPOLE FIELD

We conclude our discussion of nonspherical static fields with the following comments.

There exist solutions of the Einstein equations containing Schwarzschild surfaces S_m which do not differ qualitatively from the Schwarzschild surface for the spherical case. However, in these cases the deviations from spherical symmetry must be caused by external fields. For example, if we consider a spherical mass in an external quadrupole field (which increases with distance from the mass m), then an exact solution of the Einstein equations in vacuum has the form (the notation is the same as used in § 2 of this chapter)

$$\psi = \tfrac{1}{2} \ln [(\lambda - 1)/(\lambda + 1)] + \tfrac{1}{4} q(3\lambda^2 - 1)(3\mu^2 - 1) ,$$

$$\gamma = \tfrac{1}{2} \ln [(\lambda^2 - 1)/(\lambda^2 - \mu^2)] - 3q\lambda(1 - \mu^2)$$

$$- \tfrac{9}{16} q^2(\lambda^2 - 1)(1 - \mu^2)[9\mu^2\lambda^2 - \lambda^2 - \mu^2 + 1] .$$

The surface $g_{00} = 0$ is determined by the condition $\lambda = 1$. This surface is a Schwarzschild sphere deformed by an external field. The Gaussian curvature of the two-dimensional surface S_m (*not* the Gaussian curvature of the three-dimensional surface created by S_m as it moves through spacetime),

$$C_G{}^* = \frac{1}{4m^2} e^q[1 + 3q - 12q\mu^2 - 9q^2\mu^2 + 9q^2\mu^4] ,$$

is different for different μ and is always finite. The physical properties of this S_m are the same as those of the Schwarzschild sphere.

A constant external quadrupole field could be created by distant masses fastened on supports, which prevent them from falling inward. For a limited interval of time the same field could be created approximately also by freely moving distant masses, whose velocities under the influence of reciprocal gravitation are small initially. In this case the field would be almost static.

4.5 THE GRAVITATIONAL CONTRACTION OF A SLOWLY ROTATING BODY WITH SMALL DEVIATIONS FROM SPHERICAL SYMMETRY[6]

Let us briefly review the contraction of a uniform spherical dust cloud of radius r (see § 3.12). For a distant observer, the picture tends to "freeze" as $r \longrightarrow r_g$ because of time retardation. For this observer, the contracting ball will never have dimensions smaller than r_g. By contrast, an observer on the surface of the contracting ball reaches $R_0 = r_g$ in finite proper time. For him the contraction does not "freeze" at all, but continues through the Schwarzschild sphere into the T-region. The density of the matter in the ball, when $R_0 = r_g$ and the mass is large, has no singularity whatsoever. It can be estimated easily that

$$\rho_g \approx \frac{M}{\frac{4}{3}\pi r_g{}^3} \approx 2 \times 10^{16}(M_\odot/M)^2 \text{ g cm}^{-3} ,$$

where $M_\odot \approx 2 \times 11^{33}$ g is the mass of the Sun.

---

[6] We will not consider here large-scale magnetic and electric fields surrounding the contracting body. Such fields will be discussed later in § 13.6.

For $M = 10^8 M_\odot$, for example, $\rho_g \approx 2 \text{ g cm}^{-3}$. After the contracting ball passes its gravitational radius, light beams from its surface (see Fig. 25) cannot escape back out through the Schwarzschild sphere and *cannot* reach the distant observer.

If there are small initial perturbations in the density and velocity of the matter inside the body, then they will increase as the contraction proceeds; this was first investigated in detail by Lifshitz (1946) (see Vol. II). However, the surface of the contracting body will at some point cross a surface (event horizon) through which light rays cannot escape to infinity. Although this horizon will differ from the Schwarzschild surface due to the small perturbations, the difference will be so small that we need not make a distinction between the actual horizon and the Schwarzschild surface $r = r_g$. The moment when the star's surface crosses the Schwarzschild surface, $R_0 = r_g$, is in no way special so far as the dynamics of the body's matter are concerned; the density of the matter then is still far from infinite.

Thus, if the initial perturbations are sufficiently small, then at the moment when $R_0 = r_g$ they have not had enough time as yet to become large. The surface of the body in the frame of the comoving observer intersects the Schwarzschild sphere when the perturbations in its matter and in the field around it are still small.

Subsequently, the perturbations in the body increase. But, as demonstrated in § 3.12, an external observer does not receive any information whatsoever from under the Schwarzschild sphere. Consequently, the increasing perturbations inside S_m have no effect whatsoever on the spacetime domain near the Schwarzschild surface, nor on the external domain of the distant observer (R-region). The reader who is inclined to accept this without further explanation may skip the following paragraph. A more formal mathematical proof is included in the appendix to this section.

In fact, disturbances in the gravitational field propagate away from the body at the speed of light. It follows however, from Figure 25 that the trajectories of the rays which escape from the body in the T-region do not approach the Schwarzschild surface. Even trajectories that are fairly large perturbations of null rays do not reach it. This means that the disturbances in the vacuum near the Schwarzschild surface are always small, and that its properties remain unchanged. In particular, no information and no radiation pass through the Schwarzschild sphere to an external observer.[7] Consequently, even when there are disturbances in the collapsing body, the external observer has only a finite interval of the body's evolution accessible to him.

---

[7] Let us emphasize once more that, because of the small perturbations, the event horizon actually lies deeper than $g_{00} = 0$ (see § 4.3), so that the properties of the Schwarzschild surface are now smeared out over a narrow region (between the horizon and $g_{00} = 0$). However, that region is so very narrow that we shall simply call it the "Schwarzschild surface."

He can follow the development of the disturbances in the body and in its surrounding field, but only up to the moment when $R_0 = r_g$.

It should be clear by now that the external field of the collapsing body for a distant observer must become stationary in the limit $t \rightarrow \infty$; all $\partial/\partial t \rightarrow 0$. As a matter of fact, in the reference frame of an external observer, time-dependent perturbations that occur prior to the moment when the surface reaches the sphere r_g must spread out as gravitational waves, and

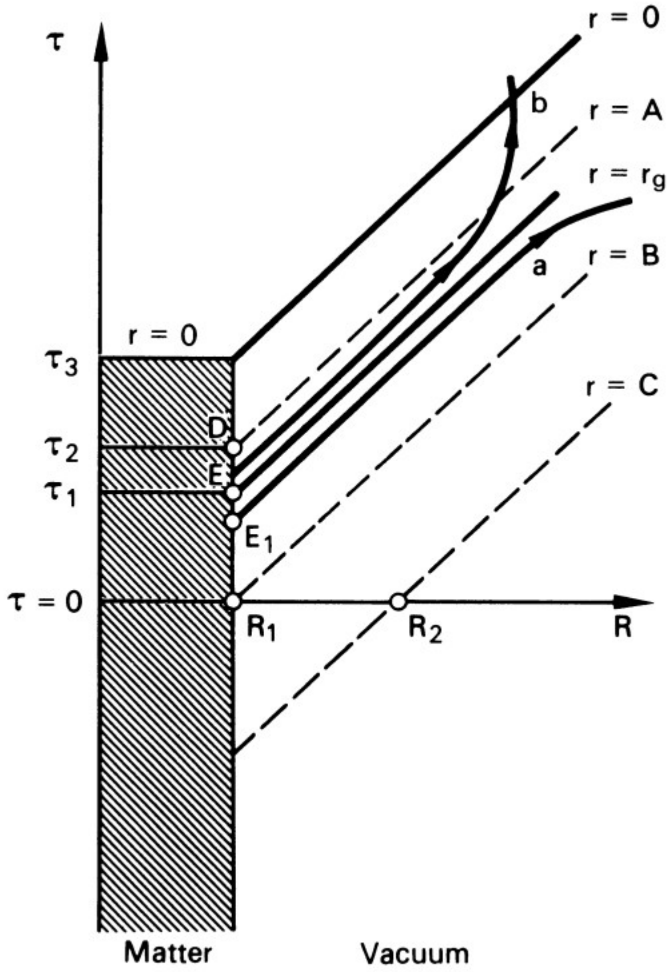

FIG. 25—.Collapse of a spherical dust cloud in a freely falling reference frame (for discussion see appendix to § 4.5). Here *a* and *b* are the world lines of light rays. Ray *a*, emitted from E_1 near time τ_1, propagates near $r = r_g$ for a long time (according to the time of *any* reference frame).

no new disturbances can be received from under the Schwarzschild sphere. Naturally, all these conclusions are valid not only for a collapsing body made of dust but also for matter with nonzero pressure.

Thus, the asymptotic field of a contracting body is stationary, with small deviations from sphericity, when $t \to \infty$.

The rotation of the body brings about an oblateness, i.e., a deviation from sphericity. The perturbations in the external field due to such deviations from sphericity are quantities of second order in the body's angular velocity, as compared with the first-order perturbations of the field due to the rotation itself. The effect of rotation upon the external field was considered in detail in § 4.3.

During the contraction of a rotating ball the external gravitational field does not change with time (up to terms linear in the angular momentum K), and the asymptotic field will be a Schwarzschild field plus a correction caused by the rotation

$$g_{03} = \sin^2 \theta \, (2K/r) . \tag{4.5.1}$$

Such a conclusion is not surprising. During the contraction of the ball, the mass M and the angular momentum K are conserved. Hence, the asymptotic field depends on both of these quantities.

What is the nature of the asymptotic field of a contracting body whose nonsphericity is caused not by rotation but, for example, by a nonsymmetrical distribution of mass? The asymptotic field must be stationary. It was demonstrated in § 4.2 that, if corrections to the Schwarzschild field for quadrupole and higher multipole moments (produced by a change in the field's source) are time-independent, then, however small they might be at finite distances from S_m, they become arbitrarily large near S_m and create a true singularity, $C = \infty$, there. On the other hand, as we mentioned at the beginning of this section, in the comoving reference frame of a contracting body with small initial, nonspherical, density perturbations the moment when the body's surface crosses the Schwarzschild surface is not special; it is not accompanied by a true singularity, either in the metric or in the density. Comparison of these results suggests that the quadrupole and higher multipole moments of the external gravitational field must attenuate in the relativistic stages of contraction of a nonsymmetrical body.

Let us attempt to establish the law which governs such attenuation. Consider the approach to the asymptotic field (in the first approximation, and near the Schwarzschild surface) as a passage through a sequence of static states. (About the validity of this assumption, see below.) We shall call this sequence of states "quasistatic" and we shall denote its quadrupole moment by $q(t)$. Examination of the equations for static, axisymmetric, quadrupole perturbations of the Schwarzschild metric indicates that when $g_{00} \to 0$, the components of the metric perturbations that are different from zero take the following form:

$$\begin{aligned} h_{00} &\sim q(1 - r_g/r) \ln (1 - r_g/r) , \\ h_{11} &\sim q(1 - r_g/r) \ln (1 - r_g/r) , \\ h_{22} &\sim h_{33} \sim q \ln (1 - r_g/r) , \end{aligned} \tag{4.5.2}$$

where q is the quadrupole parameter of the perturbation. In these equations multiplicative factors that do not depend on $(1 - r_g/r)$ are omitted.

When the body contracts with $q \neq 0$, all quantities $h_{\mu\nu}$ must remain finite in the comoving frame. Since h_{22} and h_{33} are not affected by a transformation from a comoving coordinate system to the Schwarzschild system, it is obvious that, as $r \rightarrow r_g$,

$$q \lesssim 1/\ln (1 - r_g/r) \sim 1/t . \tag{4.5.3}$$

The last relation in equation (4.5.3) was obtained from the equation (3.5.1) for the law of free fall in a Schwarzschild field. Near the Schwarzschild surface all processes occur at the rate given by equation (4.5.3) (including, e.g., the propagation of gravitational waves). Thus, the quadrupole moment q decays at least as rapidly as $q \sim t^{-1}$. The same applies to higher multipole moments.

More careful analyses of collapse with small nonspherical perturbations (Price 1970; Israel 1970) reveal that the decay of the perturbations is even more rapid than $q \sim t^{-1}$. The body always remains deformed, as seen by a comoving observer. However, as it approaches its Schwarzschild sphere, the body radiates away all nonspherical deformations of its external gravitational field, except the rotational perturbations discussed above. If the gravitational waves did not backscatter off the Schwarzschild curvature of spacetime, the metric perturbations would die exponentially at late times as

$$h_{ij} \sim \exp \left\{ -\frac{1}{\tau} [t - r - r_g \ln (r/r_g - 1)] \right\} , \quad \tau \sim r_g/c .$$

However, Price (1970) has shown that, because of backscatter, the outgoing gravitational waves leave behind a "tail." The strength of this tail dies out, at fixed radius, as

$$h_{ij} \sim t^{-(2l+2)} \ln t .$$

Here l is the order of the multipole considered ($l = 2$ for quadrupole perturbations, etc.). Interestingly, this decay law for the wave tail is valid for long-wavelength electromagnetic waves, as well as for the gravitational waves which carry off the deformations of the metric.

An important consequence of the above is this: At very late times the asymptotic, external field of a contracting, deformed, slowly rotating body is (to first order) a Schwarzschild field plus stationary "rotational perturbations" in the off-diagonal terms.

An analysis of the motions of test particles and light rays in such a field suggests that the properties of the motion are qualitatively the same as they are in the Schwarzschild field (§§ 3.9 and 3.10). For an external observer, a particle with an impact parameter that is smaller than the critical one is gravitationally captured; upon completing a finite number of spiraling revolutions, it approaches the Schwarzschild surface $g_{00} = 0$ asymptotically as $t \rightarrow \infty$.

The same is true for light rays. No radiation or information can reach an external observer from under the sphere S_m; a gravitational self-closure takes place.

These conclusions are especially important for analyses of the catastrophic contraction of stars, which will be discussed in the next part of this book.

4.6 WHAT HAPPENS TO MATTER AFTER IT FALLS THROUGH THE EVENT HORIZON?

The problem posed in the heading of this section inevitably must be faced. Actually, in the case of strict spherical symmetry, as was demonstrated in § 3.12, the contracting body is necessarily compressed to a point soon after it crosses $r_g = 2Gm/c^2$, so far as the comoving observer is concerned; all of the matter in the sphere is compressed to infinite density. What happens after that? True, the external observer cannot learn anything about it; so far as he is concerned, as $t \rightarrow \infty$ the body's surface approaches r_g, and everything that happens after that is for him in the absolute future (see § 3.12). But what is the final fate of the contracting body, not for an outside observer, but for a comoving observer on the surface of the body?

The relativistic contraction of a spherical body, as noted previously, is unstable. In the course of contraction, perturbations increase without bound as $\rho \rightarrow \infty$ (see Vol. II). Can the evolution of the asymmetries prevent the creation of a singularity in the solution and transform the contraction into expansion after a certain maximum density has been reached? A partial answer to this question has been given by Penrose (1965, 1967) and by Hawking and Penrose (1969). They have shown that if a body contracts to dimensions smaller than r_g, and if other, reasonable conditions are satisfied, then the evolution of a true singularity in the solution is unavoidable. Whether or not *all* of the matter in the body reaches infinite density cannot be established by using the methods of Penrose and Hawking. However, other methods developed by Lifshitz and Khalatnikov (1960*a*, *b*; 1963), Lifshitz, Sudakov, and Khalatnikov (1961), and Belinsky and Khalatnikov (1969) have yielded considerable information about the nature of the singularities for a wide class of solutions to the Einstein field equations. These results and the "singularity theorems" of Penrose, Hawking, and Geroch will be discussed in connection with cosmology in Vol. II of this work.

(%

In the case of an isotropic deformation of matter and vanishing rotation, the Gaussian curvature is

$$C_G = \frac{8}{3}\pi(G/c^2)\rho - \frac{1}{9}D^2$$

1.9 The Cosmological Constant

V is the four-dimensional volume

$$S = -mc\int ds - \frac{c^{\textcircled{3}}}{16\pi G}\left(\int R\,dV + \int 2\Lambda\,dV\right).$$

Pág. 138 (155) - 4.4

The Schwarzschild Sphere in an External Quadrupole Field.

...

estudar!!!

Pág. 139 (156)

This surface is a Schwarzschild sphere deformed by an external field. The Gaussian curvature of the two-dimensional surface S_m (not the Gaussian curvature of the three-dimensional surface created by S_m as it moves through spacetime),

$$C_G^{*} = \frac{1}{4m^2}e^{3}[1+\textcircled{3}q - 12q\mu^2 - 9q^2\mu^2 + 9q^2\mu^4],$$

is different for different μ and is

(%

always finite.

The physical properties of this S_M are the same of those of the Schwarzschild sphere!

Pag. 441 (458) –

We can rewrite the condition for linkage of the protons to the field lines as

$$H > 10^{-12}(r_K/r)^{3/2} \text{ gauss.} \quad (13.4.12)$$

Pag. 456 (473) –

455 (472) –

What could cause the necessary quadrupole moment?

The magnetic field, inclined at an angle relative to the axis of rotation, will produce some deformation.

However, Ostriker and Gunn have shown that the dipole magnetic field needed for the magnetic deceleration is too weak to produce the necessary deformation.

⇒ "One has to assume an internal field of the order of 10^{15} gauss with magnetic

ϕ is measured in units of c^2 (0.1 c^2 is the typical magnitude of the gravitational potential near the surface of a neutron star, cf. the next section); and $u \equiv -dr/dt \approx 1.2 \times 10^{13}\, r^{-1/2}\, (M/M_\odot)^{1/2}$.

In our search for the link between γ and the temperature change during free fall, let us consider first the case in which one can neglect the role of the last term of equation (13.4.9). In this case the falling gas will cool due to radiation reaction if the temperature "at infinity" is small enough (i.e., if the second term on right-hand side of eq. [13.4.9] dominates). In this case, the cooling law will be $T \approx (A \ln (r/r_0) + T_0{}^{1/2})^2$ initially; however, since $T_0 \sim 10000°$ K, recombination will quickly come into play, and the temperature will level off at a constant value, $T =$ const., corresponding to $\gamma = 1$. On the other hand, if the temperature "at infinity" is large, then the role of radiation reaction will diminish continuously during the fall, the first term of equation (13.4.9) will dominate, and the temperature change will approach $T \sim 1/r$, corresponding to $\gamma \rightarrow \frac{5}{3}$. The critical temperature separating these two extreme cases is

$$T_k = 2 \times 10^6 (L/L_0)^2 (0.1/\phi)^2 (M_\odot/M)^2 \beta^2 \text{ °K} . \qquad (13.4.10)$$

The existence of two limiting cases, together with the great difference between the values of γ for those cases, shows the importance of taking into account the last term in equation (13.4.9). The general situation, unfortunately, has not yet been investigated. However, by limiting ourselves to a consideration of Compton scattering between the plasma's electrons and the outward-flying photons, we can write the last term of equation (13.4.9) in the form (Zel'dovich and Shakura 1969)

$$-\frac{dQ_i/dt}{\frac{3}{2}R^* u} = 2 \times 10^6 \left(\frac{L}{L_0}\right)(T - T_p)\left(\frac{1}{r^{3/2}}\right), \qquad (13.4.11)$$

where T_p is the photon temperature. By inserting this term into equation (13.4.9) one readily sees that the temperature is practically constant for $r < 3 \times 10^{12}\, (L/L_\odot)^2$, and is equal to T_p. Other processes will obviously have only weak influence on this conclusion, but they will be important in determining what happens at larger radii.

We should note, however, that there is probably a justification, independent of the above considerations, for idealizing the accreting matter as a fluid. This justification is the existence of a magnetic field, frozen into the plasma. More particularly, the Larmor radius for a proton moving with thermal velocity turns out to be smaller than the characteristic length scale of the infalling matter whenever $H > 1.3\, T^{1/2}\, r^{-1}$ gauss. Taking account of the fact that $T \leq T_\infty(r_c/r)$, and assuming $T_\infty = 10000°$ K, $r_c = 10^{14}$ cm, we can rewrite the condition for linkage of the protons to the field lines as

$$H > 10^{-12}(r_c/r)^{3/2} \text{ gauss} . \qquad (13.4.12)$$

Tucker 1968). Thus, the problem of the energy balance of the Crab Nebula has found a solution.

The next question is how the transfer of energy and angular momentum occurs. The most plausible idea relies on magnetic dipole radiation (Pacini 1968; Gunn and Ostriker 1969). As shown by Gunn and Ostriker (1969) a magnetic field of the order of 3×10^{12} gauss at the surface gives the needed braking torque. For magnetic dipole radiation

$$dE/dt = I\omega\ (d\omega/dt) = -\text{const.}\ (d^2\mathfrak{M}/dt^2)^2 = -\text{const.}\ \mathfrak{M}^2\omega^4 .$$

With a constant magnetic moment $\mathfrak{M}$ this gives

$$d\omega/dt = -a\omega^3 , \quad (2/\omega^2) - (2/\omega_0^2) = at , \quad \omega = (2at + 1/\omega_0^2)^{-1/2} ;$$

and with $\omega_0^2 \gg \omega^2$,

$$\omega = 1/(2at)^{1/2} , \qquad d\omega/dt = -\omega/2t .$$

This law is fairly well confirmed by observations; the initial rotational velocity, which could be different for different pulsars, cancels out. The constant a varies by not more than a factor of about 10 over the life of a pulsar, $dP/dt = 3 \times 10^{-15}/P$, which means $d\omega/dt = 10^{-13}\omega^3$.

For the Crab pulsar, the exact time of birth is known so that the deceleration law can be checked. One finds $d\omega/dt = -\omega/2370$ years $= -\omega/2t$. In actuality, the explosion was in the year 1053 A.D., $t = 916$ years ago. The agreement is not too bad.

There is another source of deceleration—gravitational radiation emitted by the rotating star. Of course, if the star is symmetric around its axis of rotation, it will not radiate. Assume that the star possesses a quadrupole moment which is constant in the corotating frame. Then

$$dE_r/dt = I\omega\dot{\omega} = -\text{const.}\ (d^3Q/dt^3)^2 = -\text{const.}\ Q^2\omega^6 .$$

The corresponding law of deceleration is

$$\omega = \left(at + \frac{1}{\omega_0^4}\right)^{-1/4} \rightarrow bt^{-1/4} ,$$

i.e.,

$$d\omega/dt = -\omega/4t ,$$

which is in worse accord with the observations, but with a discrepancy of the opposite sign.

A combination of gravitational and magnetic dipole radiation can be found which matches exactly the age and current deceleration rate of the Crab Nebula's pulsar. By assuming an initial angular velocity of $\Omega_0 = 10^4$, the authors obtain $(L_0)_{\text{grav}} = 3 \times 10^{48}$ ergs sec^{-1}, $(L_0)_{\text{mag}} = 5 \times 10^{45}$ ergs sec^{-1}; and for the current epoch, $L_{\text{grav}} = 2 \times 10^{38}$ ergs sec^{-1}, $L_{\text{mag}} = 10^{39}$ ergs sec^{-1}.

What could cause the necessary quadrupole moment? The magnetic field, inclined at an angle relative to the axis of rotation, will produce some deformation. However, Ostriker and Gunn have shown that the dipole magnetic field needed for the magnetic deceleration is too weak to produce the necessary deformation. One has to assume an internal field of the order of 10^{15} gauss with magnetic field lines hidden inside the star in order to sustain the necessary asymmetry against the force of gravity.[14]

Although the numbers given are plausible, it is obvious that caution is needed; the results are too new to be definitive.

The gravitational radiation escapes completely from the pulsar and its surroundings. In connection with Weber's experiment the important point is the existence of a rigorous upper limit on the action of a pulsar on a resonant detector of gravitational waves; this limit does not depend on any assumption about the star's asymmetry. In the absence of other mechanisms of deceleration, the total energy emitted in gravitational waves per unit angular frequency is obviously

$$\Delta E_\omega/\Delta\omega = \Delta\left(\tfrac{1}{2}\, I\, \omega_r^2\right)/\Delta\omega = I\, \omega_r \Delta\omega_r/\Delta\omega = \tfrac{1}{4} I\omega\ .$$

Here it is taken into account that the frequency of the emitted waves is twice the angular velocity of rotation ω_r. For the maximum possible $I = 3.5 \times 10^{45}$ g cm^2 of a neutron star, and for the $\omega = 10^4$ of Weber's apparatus, we find

$$\Delta E_\omega/\Delta\omega = \tfrac{1}{4} \times 3.5 \times 10^{45} \times 10^4 = 10^{49} \text{ ergs}\ .$$

The corresponding averaged flux at a distance of 10 kpc. (galactic nucleus) is

$$F_\omega = \frac{\Delta E_\omega}{\Delta\omega 4\pi r^2} = 10^5 \text{ ergs cm}^{-2}\text{ Hz}^{-1}\ .$$

This is definitely below the limit of Weber's sensitivity.

Slowly rotating pulsars, such as those we observe now including the Crab pulsar, lose their energy largely by electromagnetic processes. We shall not attempt to give a detailed picture of these processes, nor shall we present a detailed theory for the directed radiation. As mentioned above, Ostriker and Gunn calculate the energy emitted by a rotating magnetic dipole. Other authors are considering charged particles and plasma which rotate with the magnetic field out to the radius where the linear velocity of rotation attains the speed of light (Gold 1968, 1969; Goldreich 1969). These two different points of view give the same order of magnitude for the rates of loss of energy and angular momentum. How can one understand this coincidence?

Consider the electromagnetic field in vacuum. It is a static dipole field $|H| \sim |M|/r^3$ in the near zone, i.e., at distances less than the reduced wavelength of the radiation emitted, $r < \lambdabar = c/\omega$, where $\lambdabar = \lambda/(2\pi)$. At greater distances it is an outgoing-wave field, with an energy flux of the order of $q = H^2/4\pi c$. The formula for the radiation is obtained by joining the fields at $r = ƛ$: we find $q = (M/ƛ^3)^2(c/4\pi)$:

---

[14] The idea of a strong internal magnetic field, much greater than the external dipole field, is familiar for ordinary stars.

$$Q = 4\pi r^2 q = 4\pi ƛ^2(M/\lambda^3)^2(c/4\pi) \sim M^2\omega^4/c^3 \sim c^{-3}(d^2M/dt^2)^2 .$$

(Note that for a rotating dipole $r = ƛ = c/\omega$ coincides with the condition $v = \omega r = c$; we have calculated the Poynting vector at the radius where $v = c$.) If the field carries charged particles, their density is of the order of the energy density of the field itself. Therefore, the energy flux carried by a plasma injected with the help of the field should be equal to the Poynting flux. This explains the coincidence.

If a rotating body loses energy, it always loses angular momentum as well at a corresponding rate: d(angular momentum)$/dt$ = (angular velocity)^{-1}d(energy)$/dt$. A rigorous Newtonian proof is given by Ostriker and Gunn (1969); and the analogous proof in general relativity is given by Hartle (1970).

By ignoring the superconductivity of the neutron star, one estimates that the time required for the magnetic field to decay is $\sim 10^7$ years. This is just the time of the active life of typical pulsars. Schwartzman (1970*d*) has pointed out that as the magnetic field decreases, particle ejection can be changed into accretion of surrounding matter onto the neutron star. Gas accretion onto a pulsar with a magnetic field is accompanied by Langmuir plasma vibrations, and by a transformation of the energy of such vibrations into electromagnetic radiation. Thus the pulsar radiation, which decays at first, can grow again because of accretion. However, the radiation must finally terminate when the pulsar is $\sim 10^7$ years old.

13.11 THE SUPERFLUIDITY AND SUPERCONDUCTIVITY OF HIGHLY COMPRESSED MATTER, AND THEIR INFLUENCE ON THE BEHAVIOR OF NEUTRON STARS

The interaction of neutrons in the highly compressed matter inside neutron stars is likely to lead to superfluidity. Experimental studies of atomic nuclei reveal that nuclear forces between like particles are attractive: two neutrons in vacuum cannot be bound; but nevertheless, when they are in the fields of other particles, for example in a nucleus, they are paired. The same effect probably occurs in a homogeneous, degenerate nuclear fluid.

Pairs of neutrons are Bose particles and therefore their behavior must be analogous to that of ^{4}He atoms in liquid helium. But as Kapitsa found experimentally, liquid ^{4}He at temperatures below 2° K is a superfluid. In neutron matter, superfluidity may exist at temperatures below ~ 1 MeV $= 10^{10}$ ° K. The same reasoning leads to the prediction that the protons in nuclear matter (whose number is of the order of a few percent of all baryons) will be paired, and will give rise to the phenomenon of superconductivity. A discussion of these properties and of the order of magnitude of the critical

(% %

field lines hidden inside the star in order to sustain the necessary asymmetry against the force of gravity.[14]

Pág. 466(477) Fig. 72

Topology of the magnetic field of stars (a), and (b) a metagalactic field.

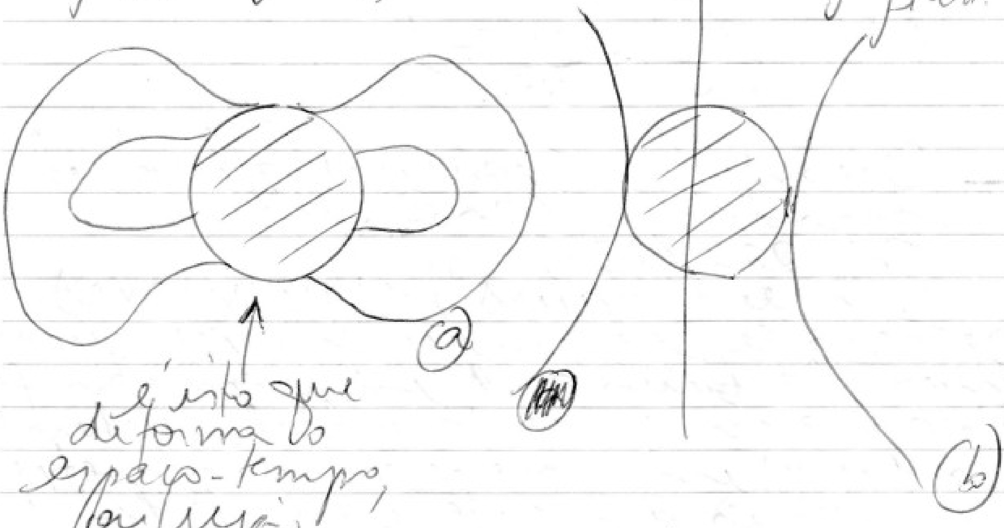

é isto que deforma o espaço-tempo, ou seja, o field é que gera a gravidade. A gravidade é decorrência das forças fundamentais da matéria.

Fundamental, melhor!!!

Pág. 463(480) ⇒ $M = 10^8 M_\odot$, the external regular field will have a strength $H_0 \approx H_i\, n^{-1/2} \approx 10^{-7}$ gauss.

(with $\nabla \times v = 0$; i.e., without vorticity) only on the microscopic scale. It is known that discrete, quantized vortex lines distributed in the liquid make the macroscopic behavior similar to that of a normal, viscous liquid. The same can be said of convection and convective heat transfer in the star. Still, there is a difference: the transfer of velocity from the superfluid neutron liquid to the proton component (which is tied to the magnetic field) is slow. Consequently, the relaxation of the stars fluid motions to an equilibrium, rigid-body rotation after a change in the rotation velocity can take a long time.

This discussion of hydrodynamic effects may sound pessimistic; but in the case of pulsars, even minute changes in period can be measured. There are very interesting hypotheses which explain the detailed variations in the periods of some pulsars by the superfluidity of the nuclear matter and the rigidity of the outer layers of the neutron star. Even if the effects are small, they are of great intrinsic interest, because they provide a means for studying the properties of matter under extreme conditions which are far outside the realm of possible laboratory physics.

Perhaps the details of the period variations (with jumps, etc., as observed in the Vela and Crab pulsars) will give, with the help of a complete theory, additional information on the mass, density, etc., of the pulsar.

13.12 Magnetic and Magnetohydrodynamic Phenomena in Collapsing Bodies

Ginzburg (1964) was the first to point out that the collapse of a star must be accompanied by a great increase of its magnetic field. Since then several contributions have concentrated on this problem (see below). In analyzing the collapse of stars and of very large gas masses, one must take account of the qualitatively different topologies of the external magnetic fields of these objects (Kardashev 1964). According to observational data and contemporary theories of stellar origin, the magnetic field of a normal star (e.g., the Sun) has a quasi-dipole nature. The topology of such a field is shown in Figure 72. Most of the magnetic field lines close not too far from the star. The topology of the external field of a galaxy or of a metagalactic object is different. According to contemporary ideas (Pikel'ner 1965), the field lines are not closed; they extend practically to infinity, thus connecting the body with the surrounding medium and with other objects.

Magnetic processes in collapse, including phenomena in the surrounding envelope, are rather complex; they have probably been investigated even less than the effects of rotation. Thus, we will limit ourselves to a very general analysis of such processes. Let us begin with the collapse of normal stars. It is known from observations that the surface magnetic fields of normal stars can be as large as $\sim 10^4$ gauss. (For comparison, the Sun's magnetic field is ~ 1 gauss.) In all cases, the magnetic energy is considerably smaller than the gravitational energy.

The conductivity of stellar matter is rather high; for stars similar to the Sun, $\sigma_{\odot} = 10^{16}\ \mathrm{sec}^{-1}$. Hence, the field attenuation time is (Pikel'ner 1961)

$$t_{\odot} \approx 4\pi\sigma_{\odot}R^{2}_{\odot}/c^{2} \approx 7 \times 10^{17}\ \mathrm{sec} \approx 2 \times 10^{10}\ \mathrm{years}\,,$$

which is considerably greater than the age of the Sun ($\sim 5 \times 10^{9}$ years). It is clear that the field attenuation time is considerably greater than the characteristic contraction time for any star (Ginzburg 1964). When a normal star is transformed to a neutron star, its conductivity increases enormously.

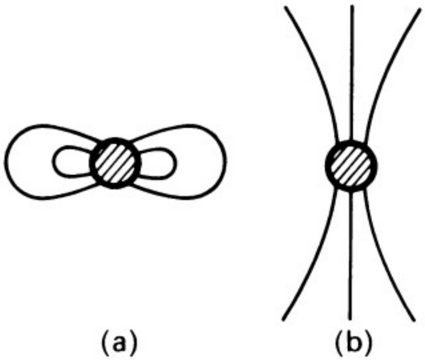

FIG. 72.—Topology of the magnetic field of (a) a star, and (b) a metagalactic object

Thus, we can always regard the field as "frozen into" the stellar matter. Under such conditions, $H \sim R^{-2}$ during contraction, and the magnetic energy increases as $E_{\mathrm{magnetic}} \sim H^{2}R^{3} \sim R^{-1}$; i.e., it increases by the same factor as the gravitational energy. In normal stars, as mentioned above, $E_{\mathrm{magnetic}} \ll E_{\mathrm{grav}}$; thus the magnetic field has no influence whatsoever upon the dynamics of the stellar collapse, for the star as a whole.

How will the magnetic field change during relativistic collapse, when a normal star is transformed to a frozen star? This question has been investigated by Ginzburg (1964) and by Ginzburg and Ozernoy (1964).

As we have emphasized, the field is "frozen" into the stellar matter; hence, in contraction to dimensions $\sim r_g$ it must reach the colossal magnitude of $\sim 10^{10}$–10^{14} gauss as seen by an observer comoving with the matter. To a stationary observer, however, the field will appear to behave very differently as $R \rightarrow r_g$. To find this behavior, Ginzburg and Ozernoy first considered the static problem.

Visualize a sequence of static, gravitating, magnetic spheres with successively smaller dimensions, and with fixed magnetic flux. Examine their external (dipole) magnetic fields. In the classical theory, the dipole moment d of each sphere is proportional to R: $d = d_0(R/R_0)$, where d_0 is the dipole moment for radius R_0; d approaches zero as $R \to 0$. Ginzburg (1964) has demonstrated that in the relativistic theory, in the limit as $R \to r_g$, the dipole moment varies according to the modified law

$$d = \frac{d_0 r_g}{R_0 \times 3 \ln [r_g/(R - r_g)]} .$$

Thus, $d \to 0$ when $R \to r_g$.

Turn now from a sequence of static stars to the variation with time of the magnetic moment of a collapsing star. At the surface, the magnetic field is computed exactly as in the stationary case, so the dipole field in the near zone decreases exactly as before. Knowing the dependence of the star's radius R on time (cf. § 11.6), Ginzburg and Ozernoy find finally that, as $t \to \infty$,

$$d = d_0 r_g^2/(3R_0 ct) .$$

[*Note added by K. S. Thorne:* Actually, this result is not correct because in the late stages of collapse, as R nears r_g, every radius r enters the wave zone within a time $\Delta t \sim r_g/c$ after the star's surface passes $R = r$. Thus, a correct analysis cannot avoid dealing with wave-zone phenomena. An analysis taking them into account (Price 1970, 1971*a*, *b*) reveals the same decay law for the magnetic field at late times,

$$(\text{field strength at fixed } r) \propto t^{-(2l+3)} \ln t = t^{-5} \ln t ,$$

as for nonspherical perturbations in the gravitational field (§ 5.4).]

What is the physical explanation for the attenuation of the magnetic field? The magnetic field is produced by circular electric currents in the star. Since all processes attenuate as seen by the external observer, the rotation speed of the charges, v, approaches 0; since $v \to 0$, the current attenuates, $I \to 0$. This attenuation of the current results in an attenuation of the external magnetic field of a collapsing star, as measured by an observer at fixed Schwarzschild radius r.

The change with time of the magnetic field during contraction produces an electrical eddy field. In the near (non-wave) zone, this eddy field may generate a curent-carrying layer in the plasma that surrounds the star, and/or may generate magnetohydrodynamic waves (Ginzburg and Ozernoy 1964). These processes have never been investigated, so we will limit ourselves to some comments about the distant (wave) zone (Novikov 1964*a*). Let r_g/c and r_g be the characteristic time and scale of the near-zone phenomena, respectively. Then the wave zone begins at a radius $l > r_g$. Let the external magnetic field of the star include a dipole term with a magnetic moment $d = \Phi R$, where $\Phi = \text{const}$. Estimate the total energy radiated by the external magnetic field during its attenuation. It is clear a priori that the energy radiated will be of the order of the total energy of the magnetic field at $R \sim r_g$, because when $R \sim r_g$, the contraction velocity is of the order c, and there exists no dimensionless parameter in which the radiated energy could be expanded. We will make a more precise calculation under the assumption that, in the process of collapse, the matter falls almost freely. Then

$$R_{,tt} = -GM/R^2\,, \quad R_{,t} = (2GM/R)^{1/2}\,. \tag{13.12.1}$$

The total power radiated is

$$I = (2/3c^3)(d_{,tt})^2\,. \tag{13.12.2}$$

Substituting equation (13.12.1) into equation (13.12.2), we get for the instantaneous power when the star is at radius R,

$$I = (\Phi^2 c/6r_g{}^2)(r_g/R)^4\,, \tag{13.12.3}$$

and, for the total energy radiated during the contraction to radius R,

$$E = (\Phi^2/15r_g)(r_g/R)^{2.5}\,. \tag{13.12.4}$$

If $\Phi \approx 3 \times 10^{21}$ (erg cm)$^{1/2}$, as it is for the Sun, and $r_g \approx 3 \times 10^5$ cm, then

$$I \approx 5 \times 10^{41}(r_g/R)^4 \text{ ergs sec}^{-1}\,, \tag{13.12.5}$$

$$E \approx 3 \times 10^{36}(r_g/R)^{2.5} \text{ ergs}\,. \tag{13.12.6}$$

Most of the radiation emerges in a single burst which lasts for a time $\sim r_g/c$.[15] This calculation was performed in the nonrelativistic theory; and the formulae used lose their validity near the Schwarzschild sphere. However, they should yield a correct estimate of the order of magnitude of the power and energy radiated. The total energy radiated in this case is not large. We will see below that the energy radiated is much greater when the contracted body is $\sim 10^5$–$10^8 M_\odot$ rather than $\sim 1\ M_\odot$ as assumed above.

Up to now we have tacitly assumed that the plasma which surrounds the star does not interfere with the outflow of the radiation. The radiation frequencies are low, $\omega \sim c/r_g$, especially for large masses. Even when the plasma has very low density, the natural frequency of plasma vibrations,

$$\omega_0 = (4\pi e^2 n_e/m_e)^{1/2} \approx 2 \times 10^4\, n_e{}^{1/2} \text{ sec}^{-1}\,,$$

is considerably larger than ω. This suggests that radiation might not occur at all. However, if the occurrence of radiation is to be prevented, then, in addition to the condition $\omega > \omega_0$, it is necessary that the maximum possible current which the plasma can support, $I_{\max} = n_e ec$ (where n_e is the electron number density and e is the electron charge), be capable of compensating for the variation of the magnetic field in the near zone. Let us calculate the critical value of the density of the surrounding plasma for which radiation no longer occurs (Novikov 1964*a*). From Maxwell's equations, we obtain

$$|\nabla \times H| = 4\pi\, I_{\max}/c = 4\pi n_e e\,. \tag{13.12.7}$$

---

[15] The decay of the pulse is caused by the relavitistic freezing of all processes when $R \to r_g$; see above.

We restrict ourselves to order-of-magnitude estimates. Consider the characteristic stage of contraction when $(R - r_g) \sim r_g$. Since $|\nabla \times H| \sim H/r_g$ and $H \sim \Phi/r_g$, we obtain from equation (13.12.7)

$$\Phi_{\rm crit} \approx 4\pi\, n_e e r_g^3 \, .$$

Thus, if the inequality

$$n_e < \Phi/(4\pi r_g^3) \, , \tag{13.12.8}$$

is satisfied, electromagnetic waves are generated in the wave zone (beginning at distances $L > r_g$), even when $c/r_g \ll \omega_0$. Substituting $\Phi \approx 3 \times 10^{21}$ and $(r_g)_\odot \approx 3 \times 10^5$ cm into equation (13.12.8), we find

$$n_e < \sim 10^{15}\ \mathrm{cm}^{-3} \, . \tag{13.12.9}$$

For comparison, the electron concentration in the solar corona near the surface of the Sun is $\sim 10^8$ cm^{-3}.

Thus, when condition (13.12.9) is satisfied, radiation is emitted by the star in the form of a single pulse, though it may be absorbed by the plasma in the wave zone far from the star. Let us recall again that the total energy radiated by normal stars is relatively small.

Let us now turn our attention from normal stars to the contraction of very massive gas clouds. The gravitational contraction of huge clouds with magnetic fields was presumably responsible for the formation of various types of galaxies and of radio galaxies. This has been particularly emphasized recently in contributions by Piddington (1964) and Pikel'ner (1965). This raises the question whether, under certain conditions, a contracting gas cloud with a magnetic field can become a supermassive star (quasar).

Here we deal with a different magnetic-field topology than in the case of a normal star (see Fig. 72). The lines of force link the contracting cloud to the surrounding medium. At the onset of contraction, the magnetic energy of the cloud is probably already of the order of its rotational energy, and the contraction is accompanied by a continual loss of angular momentum. Contraction causes "constrictions" in the lines of the magnetic field.[16] The processes which occur during this contraction and constriction are only poorly understood, so we will confine ourselves to very rough estimates. In interstellar clouds there exist magnetic fields with an intensity of $\sim 10^{-5}$ gauss. Consider an initial cloud with $M = 10^8\, M_\odot$. The external regular field will have a strength $H_0 \approx H_1 n^{-1/2} \approx 10^{-7}$ gauss. After equilibrium evolution and collapse down to r_g, the density of the cloud will have changed by approximately twenty-four orders of magnitude (from the initial density of $\sim 10^{-24}$ g cm^{-3} to a final density of ~ 1 g cm^{-3}), and the field will have grown by sixteen orders of magnitude to an intensity $H \approx 10^9$ gauss. The increase of the field energy is accompanied by a growth of the energy of the particles which are frozen onto the field.

Will the magnetic field emit radiation during its contraction? Criterion

---

[16] On this subject, see contributions by Mestel (1959).

(13.12.8), for the emission of radiation in the above case when the initial field at the surface of the cloud is $H_0 \approx 10^{-7}$ gauss, reads

$$n_e < 10^4(10^8 M_\odot/M)^{7/3}\ \mathrm{cm}^{-3}\,. \qquad (13.12.10)$$

Since it is likely that $H_0 \sim M^{1/2}$, the exponent on the expression in the parentheses might be 3 rather than $\frac{7}{3}$. Thus, the plasma density required to prevent the emission of radiation ranges from $\sim 10^4\ \mathrm{cm}^{-3}$ for $M \approx 10^8\ M_\odot$ to 10^{11}–$10^{13}\ \mathrm{cm}^{-3}$ for $M \approx 10^5\ M_\odot$.

If the plasma density satisfies criterion (13.12.8), then radiation is emitted. Let us estimate the total power and energy radiated for $M \approx 10^8\ M_\odot$, using formulae (13.12.5) and (13.12.6):

$$I = 10^{54}(r_g/R)^4\ \mathrm{ergs\ sec^{-1}}\,, \qquad E = 5 \times 10^{56}(r_g/R)^{2.5}\ \mathrm{ergs}\,.$$

If H_0 scales with mass as $H_0 \sim M^{-1/2}$, then I and E scale as $I \sim M^{-5/3}$, $E \sim M^{-2/3}$. If H_0 is independent of mass, then $I \sim M^{-2/3}$, $E \sim M^{1/3}$. Thus, the electromagnetic radiation from the external magnetic field is quite impressive for the collapse of superlarge masses, in contrast to the collapse of normal stars; possibly such radiation plays a significant role in the energy balance of quasars during their collapse stage.

13.13 THE STATISTICS OF STARS AT THE ENDPOINT OF STELLAR EVOLUTION

Do astronomical observations yield any indications about the final fates of massive stars, and about the conditions for the creation of neutron stars and frozen stars?

If the collapse of a star produces an explosion which destroys the entire star or almost the entire star, then the explosion, of course, must be visible from a tremendous distance. It is natural to identify such catastrophes with supernova outbursts, which are observed by astronomers. In such an outburst, an energy of the order of 10^{52} ergs or even more is released.

Unfortunately, it is entirely unclear at present just what portion of the mass of the star is emitted in a supernova outburst of Type I or Type II. For details see the review by the Burbidges (1958), and the monograph by Shklovsky (1966).

Assume that every star with a mass that is greater than $\sim 1.5\ M_\odot$ terminates its evolution by a supernova outburst; then the number of supernovae in a time interval Δt must equal the number of massive stars which terminate their evolution during that time interval. Let us estimate this number (Hoyle and Fowler 1963; Zwicky 1958; Schwarzschild 1958; Novikov and Ozernoy 1964; Hoyle, Fowler, Burbidge, and Burbidge 1964) and compare it with the observational data.

Assume that each massive star in the process of its evolution does not lose a substantial amount of mass, or at least that the mass losses are not so great

Books: (Important)

20211010/ Gravity/

- College-Algebra ... to-Advanced-Level.pdf

Pág. 681(696) - 8 Analytic Geometry

Pág. 684(699) Equation (proof)

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$$

- Quantum Gravity. Mathematical models ... (2007).pdf

Pág. 222(232)

Indeed, Gauss' "Disquisitiones generale circa superficies curvas" was the basis and inspiration for Riemann's work which ushered (conduziu) in a new era in geometry.

It is an extension of this geometry that is the cornerstone (pilar) of relativity theory.

Pág. 246(255)

When adds a Gauss-Bonnet term to the Einstein Hilbert action one obtains the same entropy formula (10) as when using the full Weyl multiplet [38,37].

Pág. 298(306) -

B'
! -256- - -* ! - B'
I2 D !I< I)! I I2"A! "I?!A 1"
(B B/B 2 ! 3 7 3

D -256- - -* ! - B' 6 ! !A 3 ! ! !
" ! " L1 ! D ") . ! 1"
(% 7! N4 > ! ! 1 A! O

Quantum Gravity
B. Fauser, J. Tolksdorf and E. Zeidler, Eds., 221–235
© 2006 Birkhäuser Verlag Basel/Switzerland

Topological Quantum Field Theory as Topological Quantum Gravity

Kishore Marathe

Abstract. We discuss a special case of the gauge theory to string theory correspondence in the Euclidean version of the theories, where exact results are available. We show how the Witten-Reshetikhin-Turaev invariant in $SU(n)$ Chern-Simons theory on S^3 is related via conifold transition to the all-genus generating function of the topological string amplitudes on a Calabi-Yau manifold. This result can be thought of as an interpretation of TQFT (Topological Quantum Field Theory) as TQG (Topological Quantum Gravity).

Mathematics Subject Classification (2000). Primary 81T45; Secondary 58D29.

Keywords. Chern-Simons theory, string amplitudes, conifold transition, Witten-Reshetikhin-Turaev invariants, knot invariants, topological field theory, topological quantum gravity.

1. Introduction

This talk[1] and in fact, our conference premise ignores the strong admonition by Galileo against[2] "disputar lungamente delle massime questioni senza conseguir verità nissuna". However, I do follow his most important advice: "To read the book of Universe which is wide open in front of our eyes, you must know the language in which it is written. This is the language of Geometry." I will discuss the geometric setting which has had great success in the study of fundamental particles and forces. I will also follow Gallileo's preference and discuss "a small truth" which shows a beautiful and quite unexpected relationship between topological quantum field theory and string theory amplitudes, both calculated in the Euclidean version of the theories. The result is thus primarily of interest in geometric topology. Quantum group computations initiated by Reshetikhin and Turaev and Kohno's special functions corresponding to representations of mapping class groups in the space of conformal blocks also lead to the same result.

---

[1] Given at the workshop on Mathematical and Physical Aspects of Quantum Gravity

[2] lengthy discussions about the greatest questions that fail to lead to any truth whatever.

Topological quantum field theory was ushered in by Witten in his 1989 paper [4] "QFT and the Jones' polynomial". WRT (Witten–Reshetikhin–Turaev) invariants arose as a byproduct of the quantization of Chern-Simons theory used to characterize the Jones' polynomial. At this time, it is the only known geometric characterization of the Jones' polynomial, although the Feynman integrals used by Witten do not yet have a mathematically acceptable definition. Space-time manifolds in such theories are compact Riemannian manifolds. They are referred to as Euclidean theories in the physics literature. Their role in physically interesting theories is not clear at this time and they should be regarded as toy models.

The unification of electric and magnetic fields by Maxwell is one of the most important chapters in mathematical physics. It is the only field theory which has had great predictive success in classical physics and an extension to the quantum domain. Predictions of Quantum Electrodynamics are in agreement with experimental observations to a very high degree of accuracy. Yang-Mills theory predicted massless particles and was unused for over two decades until the mechanism of symmetry breaking led to the electro-weak theory and the standard model. String theory and certain other supersymmetric theories seem to be the most promising candidates to lead to the so called grand unification of all four fundamental forces. Unifying different string theories into a single theory (such as M-theory) would seem to be the natural first step. Here even the physical foundations are not yet clear. From a mathematical point of view we would be lucky if in a few years we know what are the right questions to ask.

This year we are celebrating a number of special years. The Gauss' year and the 100th anniversary of Einstein's "Annus Mirabilis" (the miraculous year) are the most important among these. Indeed, Gauss' "Disquisitiones generale circa superficies curvas" was the basis and inspiration for Riemann's work which ushered in a new era in geometry. It is an extension of this geometry that is the cornerstone of relativity theory. More recently, we have witnessed the marriage between Gauge Theory and the Geometry of Fiber Bundles from the sometime warring tribes of Physics and Mathematics. Marriage brokers were none other than Chern and Simons. The 1975 paper by Wu and Yang [5] can be regarded as the announcement of this union. It has led to many wonderful offspring. The theories of Donaldson, Chern-Simons, Floer-Fukaya, Seiberg-Witten, and TQFT are just some of the more famous members of their extended family. Quantum Groups, CFT, Supersymmetry, String Theory, Gromov-Witten theory and Gravity also have close ties with this family. In this talk we will discuss one particular relationship between gauge theory and string theory, that has recently come to light. The qualitative aspects of Chern-Simons theory as string theory were investigated by Witten [17] almost ten years ago. Before recounting the main idea of this work we review the Feynman path integral method of quantization which is particularly suited for studying topological quantum field theories. For general background on gauge theory and geometric topology see, for example, [12, 13, 14].

In section 2 we introduce the Feynman path integral approach to QFT. The Euclidean version of this theory is applied in section 3 to the Chern-Simons Lagrangian to obtain the skein relations for the Jones-Witten polynomial of a link in S^3. A by product of this is the family of WRT invariants of 3-manifolds. They are discussed in section 4. Sections 5 to 7 are devoted to studying the relation between WRT invariants of S^3 with gauge group $SU(n)$ and the open and closed string amplitudes in generalized Calabi-Yau manifolds. This result is a special case of the general program introduced by Witten in [17]. A realization of this program even within the Euclidean field theory promises to be a rich and rewarding area of research. We have given some indication of this at the end of section 7. Various formulations of Einstein's equations for gravitational field are discussed in section 8. They also make a surprising appearance in Perelman's proof of Thurston's Geometrization conjecture. No physical application of Euclidean gravitational instantons is known at this time. So their appearance in supersymmetric string theory does not give the usual field equations of gravitation. However, topological amplitudes calculated in this theory can be thought of as Euclidean TQG. We give some further arguments in the concluding section 9 in support of supersymmetric string theory as a candidate for the unification of all four fundamental forces.

2. Quantum Observables

A **quantum field theory** may be considered as an assignment of the **quantum expectation** $< \Phi >_\mu$ to each gauge invariant function $\Phi : \mathcal{A}(M) \to \mathbb{C}$, where $\mathcal{A}(M)$ is the space of gauge potentials for a given gauge group G and the base manifold (space-time) M. Φ is called a **quantum observable** or simply an **observable** in quantum field theory. Note that the invariance of Φ under the group of gauge transformations $\mathcal{G}$ implies that Φ descends to a function on the moduli space $\mathcal{B} = \mathcal{A}/\mathcal{G}$ of gauge equivalence classes of gauge potentials. In the Feynman path integral approach to quantization the quantum or vacuum expectation $< \Phi >_\mu$ of an observable is given by the following expression.

$$< \Phi >_\mu = \frac{\int_{\mathcal{B}(M)} e^{-S_\mu(\omega)} \Phi(\omega) \mathcal{D}\mathcal{B}}{\int_{\mathcal{B}(M)} e^{-S_\mu(\omega)} \mathcal{D}\mathcal{B}}, \tag{1}$$

where $e^{-S_\mu}\mathcal{D}\mathcal{B}$ is a suitably defined measure on $\mathcal{B}(M)$. It is customary to express the quantum expectation $< \Phi >_\mu$ in terms of the **partition function** Z_μ defined by

$$Z_\mu(\Phi) := \int_{\mathcal{B}(M)} e^{-S_\mu(\omega)} \Phi(\omega) \mathcal{D}\mathcal{B}. \tag{2}$$

Thus we can write

$$< \Phi >_\mu = \frac{Z_\mu(\Phi)}{Z_\mu(1)}. \tag{3}$$

In the above equations we have written the quantum expectation as $< \Phi >_\mu$ to indicate explicitly that, in fact, we have a one-parameter family of quantum where $F_\Upsilon = \frac{\partial F}{\partial \Upsilon}$. Thus F_Υ, which depends on the higher derivative couplings $F^{(g)}$, $g \geq 1$, encodes the corrections to the area law.

If the prepotential is sufficiently simple, one can find explicit solutions of the attractor equations [15, 35, 47].[17] In particular, we can now compare $\mathcal{S}_{\text{macro}}$ to the $\mathcal{S}_{\text{micro}}$ computed from state counting (10) [15]:

$$\begin{aligned}
&\mathcal{S}_{\text{macro}}(q,p) \\
&= \frac{A}{4} + \text{Correction term} \\
&= 2\pi \frac{\frac{1}{6}|q_0|(C_{ABC}p^Ap^Bp^C + \frac{1}{2}c_{2A}p^A)}{\sqrt{\frac{1}{6}|q_0|(C_{ABC}p^Ap^Bp^C + c_{2A}p^A)}} + 2\pi \frac{\frac{1}{12}|q_0|c_{2A}p^A}{\frac{1}{6}|q_0|(C_{ABC}p^Ap^Bp^C + c_{2A}p^A)} \\
&= 2\pi \sqrt{\tfrac{1}{6}|q_0|(C_{ABC}p^Ap^Bp^C + c_{2A}p^A)} \\
&= \mathcal{S}_{\text{micro}} \ . \qquad (22)
\end{aligned}$$

In the second line we can see explicitly how the higher derivative terms modify the area. But, when sticking to the naive area law, one finds that $\frac{A}{4}$ differs from $\mathcal{S}_{\text{micro}}$ already in the first subleading term in an expansion in large charges. In contrast, when taking into account the modification of the area law, $\mathcal{S}_{\text{macro}}$ and $\mathcal{S}_{\text{micro}}$ agree completely. In other words 'string theory state counting knows about the modification of the area law.' This provides strong evidence that string theory captures the microscopic degrees of freedom of black holes, at least of supersymmetric ones.

At this point one might wonder about the role of other types of higher derivatives terms. So far, we have only included a very particular class, namely those which can be described using the Weyl multiplet. The full string effective action also contains other higher derivative terms, including terms which are higher powers in the curvature. Naively, one would expect that these also contribute to the black hole entropy. However, as we will see in the next sections, one can obtain an even more impressive agreement between microscopic and macroscopic entropy by just using the terms encoded in the Weyl multiplet. One reason might be the close relationship between the terms described by the Weyl multiplet and the topological string, which we are going to review in the next section. There are two other observation which indicate that the Weyl multiplet encodes all contributions relevant for the entropy.[18] The first observation is that when one just adds a Gauss-Bonnet term to the Einstein Hilbert action, one obtains the same entropy formula (10) as when using the full Weyl multiplet [38, 37]. The second is that (10) can also be derived using gravitational anomalies [41, 40]. Both suggest that the black hole entropy is a robust object, in the sense that it does not seem to depend sensitively on details of the Lagrangian.

---

[17] One can also construct the solution away from the horizon, at least iteratively [13, 64].

[18] For toroidal compactifications of type-II string theory there are no R^2-corrections, but the entropy of string states is non-vanishing. This case seems to require the presence of higher derivative terms which are not captured by the Weyl multiplet. See [4] for further discussion.

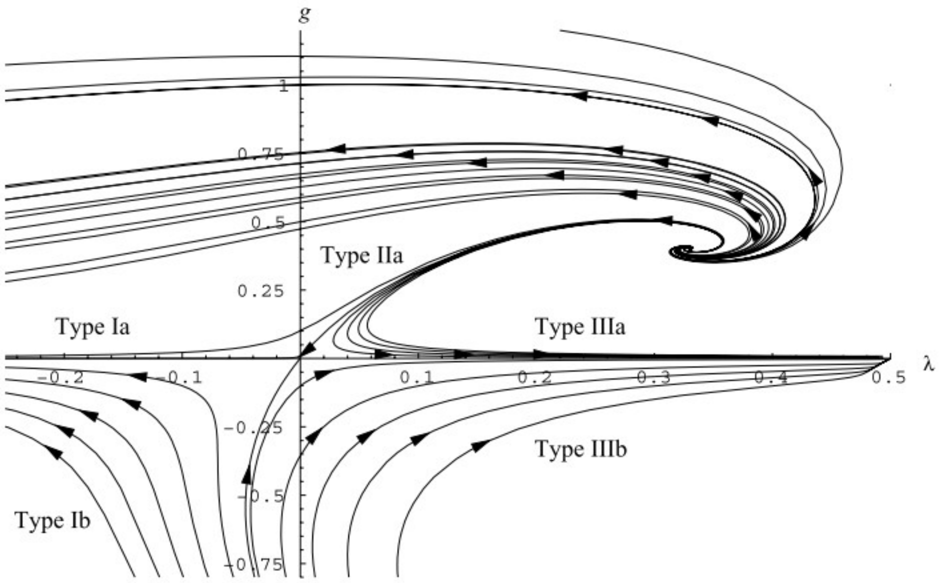

Figure 1. Part of theory space of the Einstein-Hilbert truncation with its RG flow. The arrows point in the direction of decreasing values of k. The flow is dominated by a non-Gaussian fixed point in the first quadrant and a trivial one at the origin. (From [9].)

exact RG equations no practical tool was known which would have allowed a nonperturbative calculation of the $\boldsymbol{\beta}$-functions in the physically interesting case of $d = 4$ spacetime dimensions. However, as we saw above, the effective average action in the Einstein-Hilbert approximation does indeed predict the existence of a NGFP in a nonperturbative setting. It was first analyzed in [13, 8, 9], and also first investigations of its possible role in black hole physics [25] and cosmology [26, 27] were performed already.

The detailed analyses of refs. [8, 9] demonstrated that the NGFP found has all the properties necessary for asymptotic safety. In particular one has a pair of complex conjugate critical exponents $\theta' \pm \mathrm{i}\,\theta''$ with $\theta' > 0$, implying that the NGFP, for $k \to \infty$, attracts all trajectories in the half-plane $g > 0$. (The lower half-plane $g < 0$ is unphysical probably since it corresponds to a negative Newton constant.) Because of the nonvanishing imaginary part $\theta'' \neq 0$, all trajectories spiral around the NGFP before hitting it.

The question of crucial importance is whether the fixed point predicted by the Einstein-Hilbert truncation actually approximates a fixed point in the exact theory, or whether it is an artifact of the truncation. In refs. [8, 10, 9] evidence was found which, in our opinion, strongly supports the hypothesis that there does indeed exist a non-Gaussian fixed point in the exact 4-dimensional theory, with exactly

Figure 1. Part of theory space of the Einstein-Hilbert truncation with its RG flow.

→ flow is dominated by a non-Gaussian fixed point in the first quadrant and a trivial one at the origin.

— Gravity and Strings...(2004).pdf Pág. 24(45) —

Gauss-Ostrogradski theorem

Pág 218(239) Rodapé

Coulomb force between two charges is

5. $\frac{1}{4\pi} \frac{q_1 q_2}{r_{12}^2}$ (8.14) ↙

In the Gaussian system we should replace 1/(4κ) by 1/(16πκ) and the factor of 4π disappears from the Coulomb's force.

Pág. 243(264) Euler characteristic of the Euclidean ERN solution using the Gauss-Bonnet theorem adapted to manifolds with boundaries. The

B$
-256- - -* ! - B$ 7 !A 3 ! " H ' 1"
(' '$ 7! * !" .

Since

$$\left({}^\star d\,{}^\star\omega\right)_{\rho_1\cdots\rho_{k-1}} = (-1)^{k(d-k+1)-1}\,\text{sign}\, g\,\nabla_\mu \omega^\mu_{\rho_1\cdots\rho_{k-1}}, \tag{1.133}$$

we find that the relation between δ and the divergence is

$$(\delta\omega)_{\rho_1\cdots\rho_{k-1}} = (-1)^d \nabla_\mu \omega^\mu{}_{\rho_1\cdots\rho_{k-1}}. \tag{1.134}$$

Only k-forms can be integrated on k-dimensional manifolds. If ω is a $(d-1)$-form defined on a d-dimensional manifold M with boundary $\partial\mathrm{M}$, then *Stokes' theorem* states that

$$\int_{\mathrm{M}} d\omega = \int_{\partial\mathrm{M}} \omega. \tag{1.135}$$

It is convenient to define *volume forms* for a manifold and its lower-dimensional submanifolds. Their contraction with other tensors results in differential forms that can be integrated. Thus, we define in a d-dimensional manifold, for $(d-n)$-dimensional submanifolds M^{d-n}, $0 \leq n \leq d$, the volume forms

$$d^{d-n}\Sigma_{\mu_1\cdots\mu_n} \equiv dx^{\nu_1}\cdots dx^{\nu_{d-n}}\frac{1}{(d-n)!\sqrt{|g|}}\epsilon_{\nu_1\cdots\nu_{d-n}\mu_1\cdots\mu_n}. \tag{1.136}$$

Observe that the standard invariant-volume form for the total manifold M^d is just $d^d\Sigma$ up to a sign (we now use the signature $(+-\cdots-)$):

$$d^d\Sigma = (-1)^{d-1}dx^1 \wedge \cdots \wedge\ dx^d\sqrt{|g|} \equiv (-1)^{d-1}d^dx\sqrt{|g|}. \tag{1.137}$$

Now, if we have a rank-n completely antisymmetric contravariant tensor $T^{\mu_1\cdots\mu_n}$ and contract it with the volume element $d^{d-n}\Sigma_{\mu_1\cdots\mu_n}$, we have constructed a $(d-n)$-form that can be integrated over a $(d-n)$-dimensional submanifold. Up to numerical factors, that form is the Hodge dual of the n-form that one gets by lowering the indices of $T_{\mu_1\cdots\mu_n}$:

$$\frac{1}{n!}d^{d-n}\Sigma_{\mu_1\cdots\mu_n}T^{\mu_1\cdots\mu_n} = {}^\star T. \tag{1.138}$$

We can also take the divergence of the tensor and contract it with the volume element $d^{d-n-1}\Sigma_{\mu_1\cdots\mu_{n-1}}$. The result is

$$\frac{(-1)^{d-n}}{(n-1)!}d^{d-n+1}\Sigma_{\mu_1\cdots\mu_{n-1}}\nabla_\rho T^{\rho\mu_1\cdots\mu_{n-1}} = d\,{}^\star T. \tag{1.139}$$

Stokes' theorem for the exterior derivative of form ${}^\star T$ integrated over a $(d-n+1)$-dimensional submanifold M^{d-n+1} with $(d-n)$-dimensional boundary $\partial\mathrm{M}^{d-n+1}$ is now

$$\int_{\mathrm{M}^{d-n+1}} d^{d-n+1}\Sigma_{\mu_1\cdots\mu_{n-1}}\nabla_\rho T^{\rho\mu_1\cdots\mu_{n-1}} = \frac{(-1)^{d-n}}{n}\int_{\partial\mathrm{M}^{d-n+1}} d^{d-n}\Sigma_{\mu_1\cdots\mu_n}T^{\mu_1\cdots\mu_n}. \tag{1.140}$$

The $n=1$ case is the *Gauss–Ostrogradski theorem*,

$$\int_{\mathrm{M}^d} d^dx\sqrt{|g|}\,\nabla_\mu v^\mu = (-1)^{d-1}\int_{\partial\mathrm{M}} d^{d-1}\Sigma_\mu v^\mu. \tag{1.141}$$

The Vielbein and spin-connection 1-forms and the torsion 2-form are

$$e^a = e_\mu{}^a dx^\mu, \qquad \omega^{ab} = \omega_\mu{}^{ab} dx^\mu, \qquad T^a = \tfrac{1}{2} T_{\mu\nu}{}^a dx^\mu \wedge dx^\nu, \tag{1.142}$$

These 1-forms are related by the *structure equation*

$$de^a + \omega_b{}^a \wedge e^b + T^a = 0, \tag{1.143}$$

which (in the absence of torsion) gives a convenient way of finding ω. The curvature 2-form and the Ricci-tensor 1-form are given by

$$\begin{aligned} R^{ab} &= \tfrac{1}{2} R^{ab}{}_{\mu\nu} dx^\mu \wedge dx^\nu = d\omega^{ab} + \omega_c{}^a \wedge \omega^{cb}, \\ R^a &= R_\mu{}^a dx^\mu = R_{\mu\lambda}{}^{ab} e^\lambda{}_b dx^\mu. \end{aligned} \tag{1.144}$$

1.8 Extrinsic geometry

Let us consider a hypersurface Σ embedded in a d-dimensional spacetime with metric $g_{\mu\nu}$ and with normal unit vector n^μ:

$$n^\mu n_\mu = \varepsilon, \qquad \begin{cases} \varepsilon = +1, & \Sigma \text{ spacelike}, \\ \varepsilon = -1, & \Sigma \text{ timelike}. \end{cases} \tag{1.145}$$

The metric induced on Σ by $g_{\mu\nu}$ is defined by

$$h_{\mu\nu} = g_{\mu\nu} - \varepsilon n_\mu n_\nu. \tag{1.146}$$

$h_{\mu\nu}$ has $(d-1)$-dimensional character but it is written in d-dimensional form and it is evidently singular and cannot be inverted. Its indices are raised and lowered with g. Observe that $h_{\mu\nu} n^\nu = 0$ and thus h can be used to project tensors onto the hypersurface Σ.

A way to measure how Σ is curved inside the spacetime would be to measure the variation of the normal unit vector along it. Mathematically this would be expressed by

$$\mathcal{K}_{\mu\nu} \equiv h_\mu{}^\alpha h_\nu{}^\beta \nabla_{(\alpha} n_{\beta)}, \tag{1.147}$$

where $\mathcal{K}_{\mu\nu}$ is the *extrinsic curvature* or *second fundamental form*.

We can consider a field of unit vectors n^μ defined in the whole spacetime determining a family of hypersurfaces. Then we can calculate the Lie derivative of the induced metrics in the direction of the normal unit vectors. We find that this is twice the extrinsic curvature,

$$\mathcal{K}_{\mu\nu} = \tfrac{1}{2} \mathcal{L}_n h_{\mu\nu}. \tag{1.148}$$

The trace of the extrinsic curvature is denoted by $\mathcal{K}$, and given by

$$\mathcal{K} = h^{\mu\nu} \mathcal{K}_{\mu\nu} = h^{\mu\nu} \nabla_\mu n_\nu. \tag{1.149}$$

From this point of view, if BHs can be understood as particle-like objects, looking for BHs with a well-defined scalar charge is utterly hopeless. One should look instead for "black strings" and instantons and in due time we will do so and find them.[4]

There is another point of view concerning scalar fields: in some cases they should be interpreted not as matter fields but as "local coupling constants" (as in the case of the string-theory dilaton) or, more generally, as *moduli fields*, which we will define in Chapter 11, in which case they should be treated as backgrounds and there would be no room for the notion of scalar charge.

In conclusion, if we want to find new BH solutions, we need to couple the Einstein–Hilbert action to matter fields that have associated conserved charges. The charges must be those of point-particles or we will naturally obtain solutions describing extended objects instead of black holes. Thus, we have to consider vector fields and the simplest one is an Abelian vector field A_μ. We are going to study in some detail the resulting system because later we will find generalizations of all the concepts and formulae developed here.

8.2 The Einstein–Maxwell system

The action for gravity coupled to an Abelian vector field A_μ is the so-called Einstein–Maxwell action[5] obtained by adding the Einstein–Hilbert and the Maxwell action with $\eta_{\mu\nu}$, ∂_μ, and d^4x replaced by $g_{\mu\nu}$, ∇_μ, and d^4x:

$$\boxed{S_{\rm EM}[g_{\mu\nu}, A_\mu] = S_{\rm EH}[g] + \frac{1}{c}\int d^4x\sqrt{|g|}\,\left[-\tfrac{1}{4}F^2\right].} \tag{8.15}$$

$F_{\mu\nu}$ is the field strength of the electromagnetic vector field A_μ and is again given by

$$F_{\mu\nu} = 2\partial_{[\mu}A_{\nu]}, \qquad F^2 = F_{\mu\nu}F^{\mu\nu}, \tag{8.16}$$

since, in the absence of torsion, $\nabla_{[\mu}A_{\nu]} = \partial_{[\mu}A_{\nu]}$. The components of A_μ and $F_{\mu\nu}$ in a given coordinate system are customarily split in this way,

$$(A_\mu) = (\phi, -\vec{A}), \qquad (F_{\mu\nu}) = \begin{pmatrix} 0 & E_1 & E_2 & E_3 \\ -E_1 & 0 & -B_3 & B_2 \\ -E_2 & B_3 & 0 & -B_1 \\ -E_3 & -B_2 & B_1 & 0 \end{pmatrix}, \tag{8.17}$$

---

[4] This argument really applies to pseudoscalar fields.

[5] In this section we work in the Heaviside system of units, so the Coulomb force between two charges is

$$\frac{1}{4\pi}\frac{q_1 q_2}{r_{12}^2}. \tag{8.14}$$

In the Gaussian system we should replace $1/(4c)$ by $1/(16\pi c)$ and the factor of 4π disappears from the Coulomb force. The dimensions of the vector field A_μ are $M^{1/2}L^{1/2}T^{-1}$ (that is, L^{-1} in natural units $\hbar = c = 1$) and the electric charge's units are $M^{1/2}L^{3/2}T^{-1}$, so it is dimensionless in natural units. At the end we will introduce another system of units, which will be the one we more often will work with, taking $c = 1$ and replacing the factor of $1/(4c)$ in front of F^2 by $1/\left(64G_{\rm N}^{(4)}\right)$.

where $\vec{E} = (E_1, E_2, E_3)$ and $\vec{B} = (B_1, B_2, B_3)$ are the electric and magnetic 3-vector fields in that coordinate system, and, thus, with $\vec{\nabla} = (\partial_{\underline{1}}, \partial_{\underline{2}}, \partial_{\underline{3}})$

$$\begin{cases} E_i = F_{\underline{0i}}, \\ B_i = -\frac{1}{2}\epsilon_{ijk} F_{\underline{kl}}, \end{cases} \qquad \Leftrightarrow \qquad \begin{cases} \vec{E} = -\vec{\nabla}\phi - \dfrac{1}{c}\dfrac{\partial}{\partial t}\vec{A}, \\ \vec{B} = \vec{\nabla} \times \vec{A}. \end{cases} \tag{8.18}$$

The field strength (and the action) is invariant under the Abelian gauge transformations

$$A'_\mu = A_\mu + \partial_\mu \Lambda \tag{8.19}$$

with smooth, gauge parameter Λ. Depending on which gauge group we consider ($\mathbb{R}$ or U(1)), Λ must be a single-valued or multivalued function.[6] In differential-forms language

$$A = A_\mu dx^\mu, \qquad A' = A + d\Lambda, \qquad F = \tfrac{1}{2} F_{\mu\nu} dx^\mu \wedge dx^\nu = dA, \tag{8.20}$$

and the gauge invariance of F is a consequence of $d^2 = 0$. Using these differential forms, the Maxwell action can be rewritten as follows:

$$S_{\mathrm{M}}[A] = \frac{1}{8c} \int F \wedge {}^\star F. \tag{8.21}$$

Observe that there is no matter charged with respect to A_μ in this system. This is analogous to the presence of no matter fields in the Einstein–Hilbert action. However, the Einstein–Hilbert action contains the self-coupling of gravity and therefore the presence of a coupling constant in it makes sense, whereas in the Maxwell theory there are no *direct* interactions between photons and, in principle, there is neither an electromagnetic coupling constant nor a unit of electric charge. We will see that things are a bit more complicated in the presence of gravity, through which photons do interact.

The equations of motion of $g_{\mu\nu}$ and A_μ are

$$G_{\mu\nu} - \frac{8\pi G_{\mathrm{N}}^{(4)}}{c^4} T_{\mu\nu} = 0, \tag{8.22}$$

$$\nabla_\mu F^{\mu\nu} = 0 \quad \text{(Maxwell's equation)}, \tag{8.23}$$

where

$$T_{\mu\nu} = \frac{-2c}{\sqrt{|g|}} \frac{\delta S_{\mathrm{M}}[A]}{\delta g^{\mu\nu}} = F_{\mu\rho} F_\nu{}^\rho - \tfrac{1}{4} g_{\mu\nu} F^2 \tag{8.24}$$

is the energy–momentum tensor of the vector field, which is traceless[7] in $d = 4$. The tracelessness of the electromagnetic energy–momentum tensor implies that $R = 0$ and the

---

[6] If the gauge group is $\mathbb{R}$, the elements of the group will be $e^{\Lambda/L}$, whereas, if it is U(1), they will be $e^{i\Lambda/L}$, where L is a constant introduced to make the exponent dimensionless because Λ is dimensionful. In the second case Λ will have to be identified with $\Lambda + 2\pi L$. When there is a unit of charge, L is related to it.

[7] This property is associated with the invariance of the Maxwell Lagrangian in curved spacetime under Weyl rescalings of the metric,

$$g'_{\mu\nu} = \Omega^2(x) g_{\mu\nu}. \tag{8.25}$$

In fact, if $\Omega = e^\sigma$, then for infinitesimal transformations $\delta_\sigma g_{\mu\nu} = 2\sigma(x) g_{\mu\nu}$ we have

$$\delta_\sigma S_{\mathrm{M}} = \frac{\delta S_{\mathrm{M}}}{\delta g_{\mu\nu}} \delta_\sigma g_{\mu\nu} \sim \sigma T^{\mu\nu} g_{\mu\nu} = 0. \tag{8.26}$$

Einstein equation takes the simpler form

$$R_{\mu\nu} = \frac{8\pi G_{\rm N}^{(4)}}{c^4} T_{\mu\nu}. \tag{8.27}$$

On taking the divergence of the Einstein equation and using the contracted Bianchi identity for the Einstein tensor $\nabla_\mu G^{\mu\nu} = 0$, we find

$$F_{\nu\rho}\nabla_\mu F^{\mu\rho} - \tfrac{3}{2} F^{\mu\rho}\nabla_{[\mu}F_{\rho\nu]} = 0. \tag{8.28}$$

Since the Levi-Cività connection is symmetric,

$$\nabla_{[\mu}F_{\rho\nu]} = \partial_{[\mu}F_{\rho\nu]} = 0 \qquad \text{(the Bianchi identity)} \tag{8.29}$$

identically, using the definition of $F_{\mu\nu}$, and then we see that the Einstein equation implies generically the Maxwell equation. Using Eq. (1.62), the Maxwell equation can also be written in a simpler, equivalent, form:

$$\partial_\mu\left(\sqrt{|g|}F^{\mu\nu}\right) = 0. \tag{8.30}$$

The equations are written in terms of the field strength F and usually they are solved in terms of it. However, we are ultimately interested in the vector field A itself and we have to make sure that the F we obtain is such that it is related to some vector field by Eq. (8.16) or Eq. (8.20). It turns out that, locally, A exists if the electromagnetic Bianchi identity Eq. (8.29) is satisfied.[8]

The Bianchi identity can also be written in this form (by contracting Eq. (8.29) with $\epsilon^{\mu\nu\rho\sigma}$, introducing it into the partial derivative (because it is constant), and using the

---

[8] In fact, if we are given F and the Bianchi identity is satisfied, we can always find the corresponding vector potential by using, e.g., the formula

$$\boxed{A_\mu(x) = -\int_0^1 d\lambda\lambda x^\nu F_{\mu\nu}(\lambda x).} \tag{8.31}$$

To check this formula it is necessary to use the Bianchi identity: taking the curl of the l.h.s.,

$$\partial_{[\rho}A_{\mu]}(x) = -\int_0^1 d\lambda\lambda\partial_{[\rho}[x^\nu F_{\mu]\nu}(\lambda x)], \tag{8.32}$$

and operating,

$$\partial_{[\rho}\big(F_{\mu]\nu}(\lambda x)\big) = \lambda\partial_{[\rho}F_{\mu]\nu}(\lambda x) - \tfrac{1}{2}\lambda\big(\partial_\nu F_{\rho\mu}\big)(\lambda x), \tag{8.33}$$

where the Bianchi identity Eq. (8.29) has been used in the last identity, one obtains

$$\begin{aligned}\partial_{[\rho}A_{\mu]}(x) &= \tfrac{1}{2}\int_0^1 d\lambda\left\{\lambda^2 x^\nu\partial_\nu F_{\rho\mu}(\lambda x) - \lambda F_{\mu\rho}(\lambda x)\right\} = \tfrac{1}{2}\int_0^1 d\lambda\frac{d}{d\lambda}\left[\lambda^2 F_{\rho\mu}(\lambda x)\right]\\ &= \frac{1}{2}F_{\rho\mu}(x).\end{aligned} \tag{8.34}$$

*D)! " 3

definition of the Hodge dual and Eq. (8.30) for the divergence):

$$\nabla_\mu {}^\star F^{\mu\sigma} = 0. \tag{8.35}$$

In the language of differential forms, the Maxwell equation and Bianchi identity are

$$d^\star F = 0, \tag{8.36}$$

$$dF = 0, \tag{8.37}$$

and the Bianchi identity is just a consequence of the definition Eq. (8.20) and $d^2 = 0$.

Then, if we work with the field strength, we find that there are two pairs of equations, (8.23) and (8.35) and (8.36) and (8.23), which are (as pairs) invariant if one replaces F by ${}^\star F$ (by virtue of ${}^{\star\star}F = -F$). This is an *electric–magnetic-duality* transformation. The name is due to the fact that this transformation interchanges the electric and magnetic fields in any given coordinate system according to

$$\vec{E}\,' = \vec{B}, \qquad \vec{B}\,' = -\vec{E}. \tag{8.38}$$

Actually, this pair of homogeneous equations (the Maxwell equation and the Bianchi identity) would be invariant under the (invertible) substitution for F of any linear combination of F and ${}^\star F$. We would have a symmetry of all the equations of motion if the Einstein equation were also invariant under this replacement. We will later see in Section 8.7 that this is the case and that the Einstein–Maxwell theory is invariant under electric–magnetic duality.

The four Maxwell equations in Minkowski spacetime can be deduced from the Maxwell equation and the Bianchi identity (two of them imply the existence of the potential A_μ and are equivalent to the latter). We have

$$\begin{aligned} \partial_\mu F^{\mu\nu} = 0 \quad &\Leftrightarrow \quad \begin{cases} \vec{\nabla}\cdot\vec{E} = 0, \\ \vec{\nabla}\times\vec{B} - \dfrac{1}{c}\dfrac{\partial}{\partial t}\vec{E} = 0, \end{cases} \\ \partial_\mu {}^\star F^{\mu\nu} = 0 \quad &\Leftrightarrow \quad \begin{cases} \vec{\nabla}\cdot\vec{B} = 0, \\ \vec{\nabla}\times\vec{E} + \dfrac{1}{c}\dfrac{\partial}{\partial t}\vec{B} = 0. \end{cases} \end{aligned} \tag{8.39}$$

8.2.1 Electric charge

The electric charge can be defined in terms of a source coupled to the electromagnetic field (this is analogous to the energy–momentum-pseudotensor approach for the gravitational field) or in terms of the Noether current associated with the gauge invariance (the approach that leads to Komar's formula and its generalizations for the gravitational field). The two definitions are equivalent and are very closely related to each other because the gauge invariance of the free theory imposes strong constraints on the possible couplings.

(% B

Euler characteristic χ is a topological invariant whose (de quem) value is an integer and the Gauss-Bonnet theorem states that the integral of the 4-form (naquele)

$$\frac{1}{32\pi^2}\epsilon_{abcd} R^{ab}\wedge R^{cd} \quad (8.115).$$

over a four-dimensional compact manifold M is precisely χ.

Pag. 355(376) →

which is in turn (que é por sua vez) related to the modification of the Gauss law introduced by the dilaton.

→ partícula hipotética

Trata-se de uma partícula de campo escalar Φ, um campo escalar que sempre surge com a gravidade.

Pag. 507(528) →

Using the generalization of the Gauss law Eq. (18.37) and Stokes' theorem.

— Approaches to quantum gravity (2009).pdf

-256- - -* ! - B$ 7 !A 3 ! " H
' 1"
(%$$ % B " ! 7! ! " !

The next values of interest from the string-theory-supergravity point of view are $a = 1$;

$$\boxed{\begin{aligned} ds^2 &= H^{-1} W dt^2 - H\left[W^{-1} dr^2 + r^2 d\Omega^2_{(2)}\right], \\ e^{-2\varphi} &= e^{-2\varphi_0} H, \qquad \omega = h\left[1 - (\alpha/\sqrt{2})^2\right], \end{aligned}} \tag{12.23}$$

and $a = 1/\sqrt{3}$;

$$\boxed{\begin{aligned} ds^2 &= H^{-\frac{3}{2}} W dt^2 - H^{\frac{3}{2}}\left[W^{-1} dr^2 + r^2 d\Omega^2_{(2)}\right], \\ e^{-2\varphi} &= e^{-2\varphi_0} H^{\frac{\sqrt{3}}{2}}, \qquad \omega = h\left[1 - \alpha^2/3\right]. \end{aligned}} \tag{12.24}$$

$d = 4$ stringy solutions with these metrics will appear in the compactification of solutions that describe the intersection of two and three extended objects, respectively, instead of just one as in the previous case. We are going to see how this comes about in Section 20.1.

Finally, we have $a = 0$, the RN BH:

$$\boxed{\begin{aligned} ds^2 &= H^{-2} W dt^2 - H^{2}\left[W^{-1} dr^2 + r^2 d\Omega^2_{(2)}\right], \\ e^{-2\varphi} &= e^{-2\varphi_0}, \qquad \omega = h\left[1 - (\alpha/2)^2\right]. \end{aligned}} \tag{12.25}$$

This case will be seen to arise from the intersection of four extended objects in higher dimensions.

In four dimensions, we can define the mass M, electric charge q, and "scalar charge" Σ more precisely by the asymptotic expansions

$$g_{tt} \sim 1 - \frac{2G_{\rm N}^{(4)} M}{r}, \qquad A_t \sim \frac{4G_{\rm N}^{(4)} e^{2a\varphi_0} q}{r}, \qquad \varphi \sim \varphi_0 + \frac{G_{\rm N}^{(4)} \Sigma}{r}. \tag{12.26}$$

The dilaton-dependent factor $e^{2a\varphi_0}$ in the definition of the electric charge is related to the integral definition

$$q = \frac{1}{16\pi G_{\rm N}^{(4)}} \int_{S^2_\infty} e^{-2a\varphi \star} F, \tag{12.27}$$

which is in turn related to the modification of the Gauss law introduced by the dilaton.

For $a \neq 1$ the integration constants in the solutions are given by

$$\begin{aligned} h &= \frac{1+a^2}{1-a^2} G_{\rm N}^{(4)} \left(M \pm \sqrt{M^2 - 4(1-a^2) e^{2a\varphi_0} q^2}\right), \\ \alpha &= \frac{1-a^2}{1+a^2} \frac{4e^{a\varphi_0} q}{M \pm \sqrt{M^2 - 4(1-a^2) e^{2a\varphi_0} q^2}}, \\ \omega &= \frac{2a^2}{1-a^2} G_{\rm N}^{(4)} M \pm \frac{2}{1-a^2} G_{\rm N}^{(4)} \sqrt{M^2 - 4(1-a^2) e^{2a\varphi_0} q^2}, \end{aligned} \tag{12.28}$$

and Σ is related to the conserved charges by

$$\Sigma = \frac{4a^2 e^{2a\varphi_0} q^2}{M \pm \sqrt{M^2 - 4(1-a^2) e^{2a\varphi_0} q^2}}. \tag{12.29}$$

For $a = 1$

$$h = \frac{4G_N^{(4)} e^{2\varphi_0} q^2}{M}, \qquad \alpha = -\frac{M}{e^{\varphi_0} q}, \qquad \omega = -2G_N^{(4)} \frac{M^2 - 2e^{2\varphi_0} q^2}{M}, \tag{12.30}$$

and Σ is related to the conserved charges by

$$\Sigma = -\frac{2G_N^{(4)} e^{2\varphi_0} q^2}{M}. \tag{12.31}$$

The extremality condition always takes the form

$$\left(\frac{\omega}{2G_N^{(4)}}\right)^2 = M^2 + \frac{1}{a^2}\Sigma^2 - 4e^{2\varphi_0} q^2 = 0. \tag{12.32}$$

Thermodynamics of four-dimensional dilaton BHs. The coupling to scalar fields requires a modification of the first law of BH thermodynamics, which has to include a new term [259, 260, 446] in order to take into account possible variations of the energy due to variations of the scalar fields. This term is proportional to the "scalar charges" and to the variations of the values of the scalar fields at infinity (*moduli*) that characterize the vacuum of the theory,

$$\Sigma_a d\varphi^a. \tag{12.33}$$

Apart from this new term, the temperature and the entropy of dilaton BHs are related to the area and surface gravity of the event horizon by the standard formulae. When $\omega \leq 0$ (as we will assume) the event horizon is placed at $r = -\omega$. Its area is given by

$$A = 4\pi H^{\frac{2}{1+a^2}} r^2 \Big|_{r=-\omega}, \tag{12.34}$$

and so the entropy is given by

$$S = \pi (h + |\omega|)^{\frac{2}{1+a^2}} |\omega|^{\frac{2a^2}{1+a^2}}. \tag{12.35}$$

The temperature is given by

$$T = \frac{1}{4\pi} (h + |\omega|)^{-\frac{2}{1+a^2}} |\omega|^{\frac{1-a^2}{1+a^2}}. \tag{12.36}$$

These expressions can be compared with those in [539], where the thermodynamics of four-dimensional dilaton BHs was studied, with $|\omega| = r_+ - r_-$ and $h + |\omega| = r_+$.

The behavior of T and S in the extreme limit depends on the value of the parameter a:

$$\lim_{\omega \to 0} \begin{cases} T \to 0, & S \to \pi h^{\frac{2}{1+a^2}}, \quad 0 \leq a < 1, \\ T \to h, & S \to 0, \quad a = 1, \\ T \to \infty, & S \to 0, \quad a > 1, \end{cases} \tag{12.37}$$

Below $a = 1$ the behavior is similar to that of the RN BH, which also means that near the extreme limit the specific heat is positive. Above $a = 1$ the behavior is similar to that of the Schwarzschild BH in the zero-mass limit and the specific heat is negative.

Electric–magnetic duality in the four-dimensional a-model. In four dimensions the a-model has electric–magnetic duality: the equations of motion are invariant under the discrete transformation

$$F' = \tilde{F} \equiv e^{-2a\varphi}\,{}^{\star}F, \qquad \varphi' = \tilde{\varphi} \equiv -\varphi. \tag{12.38}$$

Tilded fields are, by definition, the S-dual fields. This symmetry allows us to transform the above electrically charged solutions into magnetic solutions that have the same (Einstein-frame) metric. All the properties that depend on the Einstein metric (for instance, thermodynamical properties) are not affected by this transformation. However, in some cases we are interested in properties that depend on the metric given in a different frame (such as the string frame, that we will study, or the KK frame that we studied in Chapter 11) that is related to Einstein's by a conformal rescaling by a function of the dilaton. Since the dilaton changes in this electric–magnetic transformation, so does the (KK or stringy) metric. A good example is provided by the electric–magnetic-duality rotation of the electrically charged KK BH studied on page 328.

For special values of the parameter a there are also dyonic dilaton BH solutions, carrying both electric and magnetic charges.[4] This trivially happens for $a = 0$, the Einstein–Maxwell plus uncoupled scalar case, because in this case (as we have already seen) the electric–magnetic-duality symmetry is a continuous symmetry and one can continuously rotate the purely electric solution into the purely magnetic one. In the case $a = 1$ there is no obvious reason for this to happen. However, the $a = 1$ model is a truncation of the $N = 4, d = 4$ SUEGRA action that we are going to see next, which does have a continuous electric–magnetic-duality symmetry.

The dyonic solutions take the form[5] [432, 447, 612]

$$\boxed{\begin{aligned}
ds^2 &= (H_1 H_2)^{-1} W dt^2 - H_1 H_2 \big[W^{-1} dr^2 + r^2 d\Omega^2_{(2)}\big],\\
A_t &= \frac{-4G_{\rm N}^{(4)} e^{\varphi_0} q}{r_- - G_{\rm N}^{(4)}\Sigma}(H_1^{-1} - 1), \qquad \tilde{A}_t = \frac{1}{4\pi}\frac{e^{-\varphi_0} p}{r_- + G_{\rm N}^{(4)}\Sigma}(H_2^{-1} - 1),\\
e^{-2\varphi} &= e^{-2\varphi_0} H_1/H_2,\\
H_1 &= 1 + \frac{r_- - G_{\rm N}^{(4)}\Sigma}{r}, \qquad H_2 = 1 + \frac{r_- + G_{\rm N}^{(4)}\Sigma}{r},\\
W &= 1 - \frac{2r_0}{r}, \qquad r_\pm = M \pm r_0,\\
r_0^2 &= M^2 + \Sigma^2 - 4\left[e^{2\varphi_0} q^2 + e^{-2\varphi_0}\left(\frac{p}{16\pi G_{\rm N}^{(4)}}\right)^2\right],\\
\Sigma &= \frac{2}{M}\left[e^{2\varphi_0} q^2 - e^{-2\varphi_0}\left(\frac{p}{16\pi G_{\rm N}^{(4)}}\right)^2\right].
\end{aligned}} \tag{12.39}$$

---

[4] There is no theorem ensuring this, but all attempts to build dyonic solutions for other values of a have been unsuccessful.

[5] Solutions with additional scalar hair are also possible, but we will not deal with them any further.

Here we have used the S-dual potential $\tilde{A}_\mu$ which is the potential related to the S-dual field strength,

$$\tilde{F}_{\mu\nu} = 2\partial_{[\mu}\tilde{A}_{\nu]}, \tag{12.40}$$

whose existence is ensured by the equation of motion of A_μ, which is just the Bianchi identity for $\tilde{F}_{\mu\nu}$. Knowledge of the electric components F_{tr} and $\tilde{F}_{tr}$ and of the metric and dilaton is enough to find all the components $F_{\mu\nu}$, but this form of presenting the result is more elegant and convenient since it exhibits the symmetries of the theory acting on the solution. In particular, we see that S duality interchanges A_μ and $\tilde{A}_\mu$ and q and $p/(16\pi G_N{}^{(4)})$, and takes φ_0 to $-\varphi_0$, which also takes Σ to $-\Sigma$.

The purely electric dilaton BH solutions with $a = 1$ are recovered when $H_2 = 1$ and the purely magnetic ones when $H_1 = 1$. When $H_1 = H_2 = H$ the scalar becomes trivial and we recover the RN solutions. Thus, these solutions are the most general from the point of view of electric–magnetic duality.

As usual, when $W = 1$, H_1 and H_2 can be arbitrary harmonic functions in three-dimensional Euclidean space. They may but need not have coincident poles and, thus, the solutions describe electric and magnetic monopoles and dyons in static equilibrium.

Solutions of the four-dimensional ($a = 1$)-model with primary scalar hair and electric charge have been presented in [19] and probably can be generalized to all values of a and to higher dimensions. We will not pursue this issue any further.

12.2 Dilaton/axion black holes

The a-model is a good starting point from which to study BH solutions of supergravity/superstring theories, but it is clearly too simple. It is natural to introduce successive generalizations to this model that make it closer to the real thing. In higher dimensions we can introduce differential-form potentials of higher rank, but these are associated with extended objects. In four dimensions we can introduce, as a first step, additional vector fields, all of them coupled in the same way to the scalar field. Then, we can introduce new scalars or different couplings of the scalar(s) to the vector fields. We would have an action of the form

$$S = \frac{1}{16\pi G_{\rm N}^{(4)}} \int d^4x \sqrt{|g|} \left[R + \tfrac{1}{2} g_{ij} \partial_\mu \varphi^i \partial^\mu \varphi^j - \tfrac{1}{4} M_{ij} F^i{}_{\mu\nu} F^{j\,\mu\nu}\right], \tag{12.41}$$

where $g^{ij}(\varphi)$ and $M_{ij}(\varphi)$ are some square matrices depending on the scalars. g^{ij} can be interpreted as the inverse metric of some space of which the scalars φ_i are the coordinates. The scalar kinetic term is a *σ-model*.

A good example of an action of this kind is provided by the four-dimensional KK action that one obtains from $\hat{d} = 4 + N$ dimensions by compactification on T^N, Eq. (11.195). The scalars parametrize an $\mathbb{R}^+ \times \mathrm{SL}(N, \mathbb{R})/\mathrm{SO}(2)$ coset space.

There is another kind of couplings of scalars to vectors that we can introduce in four dimensions: couplings of the form

$$-\tfrac{1}{4} N_{ij}(\varphi) F^i{}_{\mu\nu}{}^\star F^{j\,\mu\nu}. \tag{12.42}$$

($ $! ,! 7! ! => % ! " .

The conservation law for the current $d^{\star} j = 0$ suggests the following definition for the charge q_p associated with $A_{(p+1)}$:

$$q_p \equiv \int_{\mathrm{B}^{(d-p-1)}} {}^{\star} j, \tag{18.38}$$

where, by definition, $\mathrm{B}^{(d-p-1)}$ is a *capping surface* whose boundary is (topologically) $\partial \mathrm{B}^{(d-p-1)} = \mathrm{S}^{(d-p-2)}$. More precisely, this is the charge contained in the capping surface. The total charge would be calculated by integrating over a capping surface whose boundary is the $(d-p-2)$-sphere at infinity $\mathrm{S}^{(d-p-2)}_{\infty}$. As usual this definition is invariant under smooth deformations of the capping surface in source-free ($j = 0$) regions.

Using the generalization of the Gauss law Eq. (18.37) and Stokes' theorem, we obtain the usual definition

$$\boxed{q_p = (-1)^{d+p} \int_{\mathrm{S}^{(d-p-2)}} {}^{\star} F_{(p+2)},} \tag{18.39}$$

which is also invariant under smooth deformations of the $(d-p-2)$-dimensional surface in source-free regions.

Given a p-brane sweeping out a $(p+1)$-dimensional worldvolume $\mathrm{W}^{(p+1)}$, we can immediately construct a conserved current for it that generalizes Eq. (8.49) for the current of a point-like charged object:

$$\begin{aligned} j^{\mu_1 \cdots \mu_{p+1}}(x) &= q_p \int_{\mathrm{W}^{(p+1)}} dX^{\mu_1} \wedge \cdots \wedge dX^{\mu_{p+1}} \frac{\delta^{(d)}(x - X(\xi))}{\sqrt{|g(X)|}} \\ &= q_p \int_{\mathrm{W}^{(p+1)}} d^{p+1}\xi \frac{\partial (X^{\mu_1} \cdots X^{\mu_{p+1}})}{\partial (\xi^{\mu_1} \cdots \xi^{\mu_{p+1}})} \frac{\delta^{(d)}(x - X(\xi))}{\sqrt{|g(X)|}}. \end{aligned} \tag{18.40}$$

The charge associated with this current is q_p, except when the p-brane worldvolume has boundaries, since, in that case, we can continuously deform the surface of integration and contract it to a point without meeting the p-brane source, obtaining zero charge. This is easy to visualize for a string of finite length in a four-dimensional target space: the string charge with respect to a 2-form potential is calculated through a closed line integral around the string that can be slid off the string and contracted to a point as shown in Figure 18.1. In a sense this explains why the KR 2-form does not appear in the open-string spectrum. It is clear now that $q_p \neq 0$ only when the p-brane has compact topology or extends to infinity. A brane with boundaries can also carry charge if the boundaries are attached to another object just as open strings are attached to D-branes. This case, which is more complicated (and interesting), will be studied in Section 19.6 since it depends strongly on the theory we are considering.

On substituting the above p-brane current into the interaction term Eq. (18.34), we obtain the standard form of the WZ term for p-branes which appears in the p-brane worldvolume action,

$$(-1)^{(p+1)} q_p \int_{\mathrm{W}^{(p+1)}} A_{(p+1)}. \tag{18.41}$$

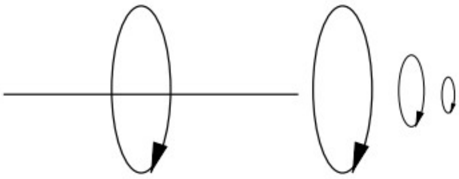

Fig. 18.1. Open strings cannot carry charge.

Electric–magnetic duality for extended objects. The electric–magnetic dual of a charged p-brane is, by definition, an object that couples to the electric-magnetic dual of the $(p+1)$-form potential $A_{(p+1)}$. We have seen several examples, for instance, in Section 16.5.5, in which we used Poincaré duality to replace the KR 2-form completely by its dual in various dimensions and also in Sections 16.1.3 and 17.1.1 in which we defined the on-shell duals of the RR potentials that appear in the $N = 2A, B, d = 10$ supergravity actions, not being able to replace the original potentials completely by their duals.

In all (massless) cases, the dual is a $(\tilde{p}+1)$-form $\tilde{A}_{(\tilde{p}+1)}$ with $\tilde{p} = d - p - 4$, whose field strength is, by definition, the Hodge dual of $\tilde{F}_{\tilde{p}+2} = {}^\star F_{(p+2)}$. The electric-magnetic dual of a p-brane is therefore a $\tilde{p}$-brane with $\tilde{p} \neq p$ in general, which is electrically charged with respect to $\tilde{A}_{(\tilde{p}+1)}$. Only in even dimensions do objects with $p = (d-4)/2$ have duals of the same dimension, and then we can have p-brane dyons.

A new feature with respect to point-particles is that there can be *self-dual p-branes*, charged with respect to a self-dual potential: point-particles couple only to vectors, which are dual to vectors in $d = 4$, but, in $d = 4$, self- or anti-self-duality is consistent only with a Euclidean signature. However, self-dual 2-forms in $d = 6$ and 4-forms in $d = 10$ can be consistently defined. They occur in $N = (1, 0), d = 6$ SUGRA (Section 13.4.1) and in $N = 2B, d = 10$ SUEGRA (Chapter 17), and are associated with chiral theories.

Dirac charge quantization. The charge of a p-brane moving in the background $A_{(p+1)}$ field sourced by a dual $\tilde{p}$-brane is quantized as in the case of point-particles [718, 881, 882].[5]

Let us consider the quantum propagation of a charged p-brane moving along a closed path[6] so its worldvolume is topologically a $(p+1)$-sphere $\mathrm{S}^{(p+1)}$. The interesting term in the path integral is the WZ term which, using Stokes' theorem, can be written in the form

$$(-1)^{(p+1)} q_p \int_{\mathrm{B}^{(p+2)}} dA_{(p+1)} = (-1)^{(p+1)} q_p \int_{\mathrm{B}^{(p+2)}} F_{(p+2)}, \tag{18.42}$$

where $\mathrm{B}^{(p+2)}$ is one of the many possible $(p+2)$-dimensional capping surfaces with $\partial \mathrm{B}^{(p+2)} = \mathrm{S}^{(p+1)}$. To avoid having any ambiguities in the path integral, the difference between choosing two different capping surfaces $\mathrm{B}_1^{(p+2)}$ and $\mathrm{B}_2^{(p+2)}$ must be an integer multiple of 2π:

---

[5] The derivation of the Dirac quantization condition we are about to explain is different from those used in Section 8.7.2 for point-particles, which can also be generalized to charged p-branes.

[6] This will be the analog of a point-particle moving along a closed path encircling a Dirac string.

$$(-1)^{(p+1)} q \int_{\mathrm{B}_1^{(p+2)}} F_{(p+2)} - (-1)^{(p+1)} q \int_{\mathrm{B}_2^{(p+2)}} F_{(p+2)} = 2\pi n. \tag{18.43}$$

Now, the first capping surface plus the second (with reversed orientation) form (topologically) a $(p+2)$-sphere $\mathrm{S}^{(p+2)}$, and we arrive at the condition

$$q_p \int_{\mathrm{S}^{(p+2)}} F_{(p+2)} = (-1)^{(p+1)} 2\pi n. \tag{18.44}$$

Defining now the "magnetic" charge, or, better, the electric charge of the dual $\tilde{p}$-brane, by

$$\boxed{p_p = q_{\tilde{p}} = (-1)^{d+\tilde{p}} \int_{\mathrm{S}^{p+2}} F_{(p+2)},} \tag{18.45}$$

the above result takes the form of the generalization of the Dirac quantization condition we were after:

$$\boxed{q_p q_{\tilde{p}} = 2\pi n, \qquad n \in \mathbb{Z}.} \tag{18.46}$$

The charge quantization condition for p-brane dyons was found in [305].

Observe that, in the string σ-model action Eq. (15.31), the constant that plays the role of the charge with respect to the KR 2-form is the string tension T. This identity, which has the form of a BPS bound, indicates that the object described by the action is a BPS object. All the extended objects we are going to deal with have the same property.

Observe also that, even though we have defined a scalar charge q_p because this is a useful quantity, this number does not give all the information. Actually, the charge should be the tensor

$$\boxed{Z^{\mu_1\cdots\mu_p} \sim \int_{\Sigma} d^{d-1}\Sigma_{\mu} j^{\mu\mu_1\cdots\mu_p}} \tag{18.47}$$

(where Σ is a spacelike hypersurface) and includes information about spatial orientation, etc. This explains why objects with the same tensions and q_ps can be in equilibrium if they are parallel (so their $Z^{\mu_1\cdots\mu_p}$s are identical) but not when one of them is tilted. These tensorial charges appear naturally in supersymmetry algebras, as we are going to see in Section 19.5.

18.1.3 The coupling of p-branes to scalar fields

A spacetime scalar $K(X)$ can be introduced in only one place in the NG action if we want to preserve reparametrization invariance and the gauge invariance of the WZ term,

$$S_{\mathrm{NG}}^{(p)}[X^{\mu}(\xi)] = -T_{(p)} \int d^{p+1}\xi (K/K_0) \sqrt{|g_{ij}|}. \tag{18.48}$$

Pág. 80(102), 5.4
A spacetime approach to
Quantum Gravity Theory
Lorentzian metric is quantized...

Pág. 122(144) Fig. 8.2.
The flow in the Einstein-Hilbert
truncation
(Gaussian FP).

Pág. 244(266) –
Newton's constant and $\iota > 0$
is any positive real number
Define the Gauss constraint

$$G_j := \partial_a E_j^a + \epsilon_{jkl} A_a^k E_l^a .$$

Pág. 245(267)
From the Gauss constraint and
spatial diffeomorphism constraint
respectively.

Pág. 276(298) –
– Poisson bracket
– Gauss law and the curvature
constraint respectively.

BB
-256- - -* ! - BB 211 ! > ! !A 3 ! "
! " ! " 1! J J ! " ! ! D " H A 3 / 1"
($ ' 2 1! !11 ! 6 ! 7 !A 3 9 3
? , ! > ! , "

5.4 A spacetime approach to Quantum Gravity theory

5.4.1 Motivation

The histories approach to General Relativity suggests a new, spacetime-focussed, approach to Quantum Gravity, characterized by two features that are not implemented in the existing Quantum Gravity schemes.

First, the Lorentzian metric is quantized, analogous to the 'external' quantum field in the history approach to scalar quantum field theory [22; 14]. This contrasts with conventional canonical Quantum Gravity where only a spatial three-metric is quantised. Second, the history scheme incorporates general covariance via a manifest representation of the spacetime diffeomorphism group.

The canonical quantisation scheme was originally developed with the hope of providing a background independent formulation of Quantum Gravity. The general procedure involves (i) a $3+1$ splitting of spacetime; (ii) the construction of a suitable Hilbert space to accommodate the basic kinematical quantities of the theory; and (iii) the definition of self-adjoint operators that represent the Hamiltonian constraints. The imposition of the constraints on the state vectors then projects out the physical degrees of freedom.

The canonical treatment of Quantum Gravity introduces a spacelike foliation that enters the quantum description. However, the physical predictions should be independent of the choice of this foliation. This is part of the famous 'problem of time', as are attempts to understand the spacetime diffeomorphism group in this context. These issues are significantly addressed by the histories formulation with its genuine spacetime description of physical quantities.

The definition of the history group provides the HPO formalism with a quantisation scheme that follows the general lore of canonical quantisation, providing however a fully covariant description – see for example the quantum treatment of minisuperspace models in [3].

The obvious technical problem in a histories-based quantisation is the rigorous implementation of the dynamics by a history analogue of the Hamiltonian constraint operator. As in standard canonical theory, the classical expression is non-quadratic – indeed non-polynomial – in the field variables, and so the construction of an operator for the Hamiltonian constraint seems a hopeless task using conventional methods. For this reason, we intend to exploit the basic ideas of loop quantum gravity, which has been a promising approach for the construction this operator.

5.4.2 Towards a histories analogue of loop quantum gravity

Loop quantum gravity is a successful canonical theory in many respects. The basic algebra is defined with reference to objects that have support on loops in the three-dimensional surface Σ.

The first step in the application of the histories formalism to loop quantum gravity is to develop the histories analogue of the connection formalism of General Relativity. The original formulation of the programme involved the consideration of self-dual $SL(2, \mathbf{C})$ connections on spacetime, together with a field of tetrads for a Lorentzian metric [26; 4]. However, the mainline approach in loop quantum gravity finds it more convenient to employ in quantisation a real $SU(2)$ connection, proposed by Barbero [5]. The Barbero connection may be obtained as a variable in a state space, extending that of canonical General Relativity; or it may be obtained from a Lagrangian action (the Holst Lagrangian by [11]). However, the latter procedure involves gauge fixing, and it is not clear whether the connection may be defined in its absence – see [19; 20] for related discussions.

The histories description for classical gravity in term of the Holst Lagrangian has been developed in [25]. The basic variables at the covariant level is an $SL(2, \mathbf{C})$ connection and a field of tetrads on spacetime M, together with their conjugate variables. The corresponding history space carries a symplectic action of the group $Diff(M)$ of spacetime diffeomorphisms. The introduction of an equivariant foliation functional allows the translation of the spacetime description into that of an one-parameter family of canonical structures. The results of the metric-based theory can be fully reproduced in this construction: the set of constraints corresponding to the Holst Lagrangian is invariant under the action of the spacetime $Diff(M)$ group. Hence the generators of the spacetime diffeomorphisms group can also be projected onto the reduced state space.

The next step would involve choosing the basic variables for quantisation. Following the spirit of loop quantum gravity, we may try to identify a loop algebra, and then construct a histories Hilbert space by studying its representation theory. The obvious place to start would be the loop algebra corresponding to the spacetime $SL(2, \mathbf{C})$ connection of the covariant description. This, however, would involve a representation theory for loop variables with a non-compact gauge group, which to the best of our knowledge has not yet been fully developed. Moreover, we would have to identify a new role for the tetrad fields, because at this level they commute with the connection variables.

It may be more profitable to work with ‘internal’ fields, namely the ones that correspond to one-parameter families of the standard canonical variables. This would allow the consideration of connections with compact gauge group. However, a complication arises, because of the gauge-dependence of the definition of the Barbero connection. A gauge-fixing condition, at this level, breaks the background independence of the theory. In [25] we show that a connection sharing all properties of the Barbero connection can be defined in a gauge invariant way, albeit in a larger space than the one usually employed.

A history quantisation may be therefore envisioned that will employ variables defined with support on a two-dimensional cylinder – giving a history analogue of the T_0 variables – and a three-dimensional space $S \times R$ (where S is a spatial two-surface), as a history analogue of the T_1 variables. The price of gauge-invariance is that additional canonical variables have to be quantised: they correspond to a unit timelike vector field that determines the possible ways that the group $SU(2)$ can be embedded into $SL(2, C)$ – see [14] for the quantisation of a related structure in the context of quantum field theories.

The above is a potential point of divergence between histories quantisation and the canonical loop approach, which is necessary in light of the strong restrictions placed by our requirement of full spacetime general covariance. At present the research is focussed on finding a proper algebra for quantisation. An interesting possibility is that the histories formalism may provide spacetime geometric operators: for example, an operator for spacetime volume; or 'length' operators that distinguish between spacelike and timelike curves.

Acknowledgement

The preparation of this publication was supported by the EP/C517687 EPSRC grant.

References

[1] J. Ambjorn, J. Jurkiewicz, & R. Loll, A non-perturbative Lorentzian path integral for gravity. *phys. Rev. Lett.* **85** (2000) 924.

[2] C. Anastopoulos, Continuous-time histories: observables, probabilities, phase space structure and the classical limit. *J. Math. Phys.* **42** (2001) 3225.

[3] C. Anastopoulos & K. N. Savvidou, Minisuperspace models in history theory. *Class. Quant. Grav.* **22** (2005) 1841.

[4] A. Ashtekar, New variables for classical and quantum gravity. *Phys. Rev. Lett.*, **57**(18) (1986) 2244–2247.

[5] J. F. Barbero, Reality conditions and Ashtekar variables: a different perspective. *Phys.Rev.* **D51** (1995) 5498.

[6] L. Bombelli, J. H. Lee, D. Meyer, & R. Sorkin, Spacetime as a causal set. *Phys. Rev. Lett.* **59** (1989) 521.

[7] A. Burch, Histories electromagnetism. *J. Math. Phys.* **45**(6) (2004) 2153.

[8] M. Gell-Mann & J. B. Hartle, Quantum mechanics in the light of quantum cosmology. In *Complexity, Entropy and the Physics of Information*, ed. W. Zurek. (Addison Wesley, Reading, 1990).

[9] R. B. Griffiths, Consistent histories and the interpretation of quantum mechanics. *J. Stat. Phys.* **36** (1984) 219.

[10] J. Hartle, Spacetime quantum mechanics and the quantum mechanics of spacetime. In *Proceedings on the 1992 Les Houches School,Gravitation and Quantisation* (1993).

[11] S. Holst, Barbero's Hamiltonian derived from a generalized Hilbert–Palatini action. *Phys. Rev.* **D53** (1996) 5966.

[12] C. Isham, Structural issues in quantum gravity. Plenary Talk at GR14 conference, (1995) gr-qc/9510063.

[13] C. J. Isham, & N. Linden, Continuous histories and the history group in generalised quantum theory. *J. Math. Phys.* **36** (1995) 5392.

[14] C. J. Isham, & K. N. Savvidou, Quantising the foliation in history quantum field theory, *J. Math. Phys* **43** (2002).

[15] C. Isham, N. Linden, K. Savvidou, & S. Schreckenberg, Continuous time and consistent histories. *J. Math. Phys.* **37** (1998) 2261.

[16] K. Kuchar, The problem of time in canonical quantization. In *Conceptual Problems of Quantum Gravity*, eds. A. Ashtekar & J. Stachel (Birkhäuser, Boston, 1992) pp. 141–171.

[17] R. Omnès, Logical reformulation of quantum mechanics: I Foundations. *J. Stat. Phys.* **53** (1998) 893.

[18] R. Omnès, *The Interpretation of Quantum Mechanics* (Princeton University Press, Princeton, 1994).

[19] J. Samuel, Is Barbero's Hamiltonian formulation a gauge theory of Lorentzian gravity? *Class. Quant. Grav.* **17** (2000) L141.

[20] J. Samuel, Canonical gravity, diffeomorphisms and objective histories. *Class. Quant. Grav.* **17** (2000) 4645.

[21] K. N. Savvidou, The action operator in continuous time histories. *J. Math. Phys.* **40** (1999) 5657.

[22] K. N. Savvidou, Poincaré invariance for continuous-time histories. *J. Math. Phys.* **43** (2002) 3053.

[23] K. N. Savvidou, General relativity histories theory II: Invariance groups. *Class. Quant. Grav.* **21** (2004) 631.

[24] K. N. Savvidou, General relativity histories theory I: The spacetime character of the canonical description. *Class. Quant. Grav.* **21** (2004) 615.

[25] K. N. Savvidou, Histories analogue of the general relativity connection formalism (in the press).

[26] A. Sen, Gravity as a spin system. *Phys. Lett.* **B119** (1982) 89–91.

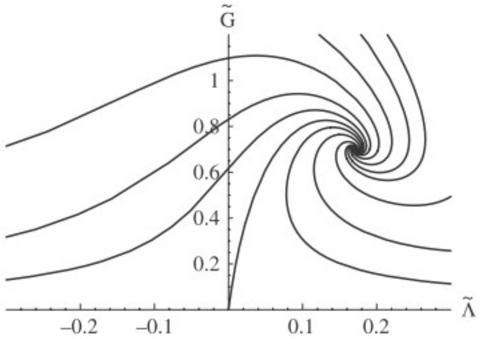

Fig. 8.2. The flow in the Einstein–Hilbert truncation, see eqs. (8.23) and (8.24). The nontrivial FP at $\tilde{\Lambda} = 0.171$, $\tilde{G} = 0.701$ is UV-attractive with eigenvalues $-1.69 \pm 2.49i$. The Gaussian FP is attractive along the $\tilde{\Lambda}$-axis with eigenvalue -2 and repulsive in the direction (0.04, 1.00) with eigenvalue 2.

Lauscher & Reuter [22] and Reuter & Saueressig [37] have studied the gauge- and cutoff-dependence of the FP in the Einstein–Hilbert truncation. The dimensionless quantity $\Lambda' = \Lambda G$ (the cosmological constant in Planck units) and the critical exponents have a reassuringly weak dependence on these parameters. This has been taken as a sign that the FP is not an artifact of the truncation. Lauscher & Reuter [23] have also studied the ERGE including a term R^2 in the truncation. They find that in the subspace of $\tilde{\mathcal{Q}}$ spanned by $\tilde{\Lambda}$, $\tilde{G}$, $1/\xi$, the non-Gaussian FP is very close to the one of the Einstein–Hilbert truncation, and is UV-attractive in all three directions. More recently, the FP has been shown to exist if the Lagrangian density is a polynomial in R of order up to six (Codello, Percacci and Rahmede, in preparation). In this truncation the UV critical surface is three dimensional.

There have been also other generalizations. Niedermaier [28] considered the RG flow for dimensionally reduced $d = 4$ gravity, under the hypothesis of the existence of two Killing vectors. This subsector of the theory is parametrized by infinitely many couplings, and has been proved to be asymptotically safe.

Matter couplings have been considered by Percacci & Perini [31; 32]. Consider the general action

$$\Gamma_k(g_{\mu\nu}, \phi) = \int d^4x \sqrt{g} \left(-\frac{1}{2} g^{\mu\nu} \partial_\mu \phi \partial_\nu \phi - V(\phi^2) + F(\phi^2) R \right), \qquad (8.25)$$

where V and F are arbitrary functions of ϕ^2, analytic at $\phi^2 = 0$. This action has a so-called Gaussian-Matter FP, meaning that only the coefficients of the ϕ-independent terms in (8.25) (namely $g^{(0)}$ and $g^{(1)}$) are nonzero. The critical surface has dimension four and there are no marginal operators. In the presence of other, minimally coupled matter fields, the dimension of the critical surface can be

We consider spacetime manifolds M which are diffeomorphic to $\mathbb{R} \times \Sigma$ where Σ is a 3D manifold of arbitray topology. Consider any principal $SU(2)$ bundle[7] $\mathcal{P}$ over Σ and denote by A the pull back by local sections of a connection on P. Likewise, consider an associated, under the adjoint representation of $SU(2)$, vector bundle whose local sections are $su(2)$ valued vector densities E. We now consider a new phase space with the following symplectic structure

$$\{E^a_j(x), A^k_b(y)\} = \kappa \iota \delta^a_b \delta^k_j \delta(x, y) \tag{13.11}$$

where $\kappa = 8\pi G_N,\ G_N$ is Newton's constant and $\iota > 0$ is any positive real number [19; 20]. All other brackets vanish.

We now want to establish the connection with the ADM formulation. Consider a Dreibein whose local sections are $su(2)$ valued one forms e. Then $q_{ab} = \delta_{ij} e^i_a e^j_b$ and $E^a_j = |\det(e)| e^a_j$ where $e^a_j e^j_b = \delta^a_b,\ e^a_j e^k_a = \delta^k_j$ and q is the pull back of the spacetime metric to Σ. Next, denote the pull back by local sections of the spin connection[8] associated with e by Γ. Then $A^j_a = \Gamma^j_a + \iota K_{ab} e^b_j$ where K_{ab} is the pull back to Σ of the extrinsic curvature of the foliation $M \cong \mathbb{R} \times \Sigma$.

The non-trivial result is now the following. We may use the relations just displayed to invert $q_{ab} = H_{ab}(A, E),\quad P^{ab} = \Pi^{ab}(A, E)$ where H, Π are functions which can be easily derived by the reader from the definition $P^{ab} = \sqrt{\det(q)}[q^{ac}q^{bd} - q^{ab}q^{cd}]K_{cd}$ of the momentum conjugate[9] to q. Define the Gauss constraint

$$C_j := \partial_a E^a_j + \epsilon_{jkl} A^k_a E^a_l. \tag{13.12}$$

Then one can show that the Poisson brackets of H, Π, using the symplectic structure (13.11) are precisely the Poisson brackets of q, P up to terms which vanish when $C_j = 0$. In other words, the phase space spanned by A, E with symplectic structure (13.11) and constraint (13.12) imposed is precisely the ADM phase space of geometrodynamics. Notice that H, Π are Dirac observables with respect to the Gauss constraint.

We can now define the algebra $\mathfrak{P}$ of elementary observables based on A, E. The analogy with an $SU(2)$ gauge theory familiar from the electroweak theory naturally suggests the use of techniques which are standard in (lattice) gauge theory. To that end, let s, S be a one- and two-dimensional submanifold of Σ respectively, which

---

use real valued ι and multiply the Hamiltonian constraint by a sufficiently high power of $\sqrt{|\det(E)|}$ in order to make it polynomial. The work of [25; 26; 27; 28; 29; 30; 31; 32] showed that the representation for $\mathfrak{A}$ supports the Hamiltonian constraint only if one uses the non-polynomial form. The reason for this is again background independence.

[7] One can show that principal $SU(2)$ bundles over 3D Σ are necessarily trivial but we will not need this in what follows.

[8] It is defined by the covariant constance of e, i.e. $D_a e^j_b = 0$.

[9] I.e. $\{P^{ab}(x), q_{cd}(y)\} = 16\pi G_N \delta^a_{(c} \delta^b_{d)} \delta(x, y)$.

will play the role of labelling smearing functions. Furthermore, let k be an $su(2)$ valued, smooth function on Σ. We define holonomy and flux functions

$$A(s) := \mathcal{P}\exp(\int_p A), \quad E_k(S) := \int_S \mathrm{Tr}(k \ *E) \tag{13.13}$$

where $*E$ is the metric independent, pseudo-two form dual to E. Now consider an arbitrary, finite collection of paths s. Their union forms a finite[10] graph γ and we may compose the paths s from the edges of the resulting graph. We now call a function cylindrical over a graph provided it is a complex valued function of the $A(s)$ where s runs through the edges of the graph. The cylindrical functions form an Abelean *-algebra which we denote by Cyl. Next, denote by $Y_{k,S}$ the Hamiltonian vector field of $E_k(S)$. Then $\mathfrak{P}$ is defined as the Lie algebra of cylindrical functions f and vector fields v equipped with the following Lie bracket $[(f, v), (f', v')] = (v[f'] - v'[f], [v, v'])$. The most important building block in that algebra is

$$Y_{k,S}[A(s)] = \{E_{k,S}, A(s)\} = \iota\kappa A(s_1)k(s \cap S)A(s_2) \tag{13.14}$$

where $s = s_1 \circ s_2$ and we have assumed that $s \cap S$ is precisely one point, the general case being similar.

13.3.1.1 The quantum algebra $\mathfrak{A}$ and its representations

The corresponding $\mathfrak{A}$ is defined by formally following the procedure of section 13.2. We now consider its representation theory. Since we are dealing with a field theory the representation theory of $\mathfrak{A}$ will be very rich so we have to downsize it by imposing additional physical requirements. The natural requirement is that the representation derives from a state invariant under the automorphisms of the bundle P. Locally these automorphisms can be identified with the semidirect product $\mathfrak{G} := \mathcal{G} \rtimes \mathrm{Diff}(\Sigma)$ of local $SU(2)$ gauge transformations and spatial diffeomorphisms. The requirement of $\mathfrak{G}$-invariance is natural because both groups are generated canonically, that is by the exponential of the respective Hamiltonian vector fields, from the Gauss constraint and spatial diffeomorphism constraint respectively. For instance we have with $C(\Lambda) = \int_\Sigma \ d^3x \ \Lambda^j C_j$ and $C(\vec{N}) = \int_\Sigma \ d^3x \ N^a C_a$ that

$$\begin{aligned} \alpha_{g_\Lambda}(A(s)) \ &:= \ \exp(X_{C(\Lambda)}) \cdot A(s) = g_\Lambda(b(s)A(s)g_\Lambda(f(s)) \\ \alpha_{\varphi_{\vec{N}}}(A(s)) \ &:= \ \exp(X_{C(\vec{N})}) \cdot A(s) = A(\varphi_{\vec{N}}(s)), \end{aligned} \tag{13.15}$$

where $g_\Lambda = \exp(\Lambda)$ and $\varphi_{\vec{N}}$ is the diffeomorphism defined by the integral curves of the vector field $\vec{N}$. Here, $b(s)$, $f(s)$ respectively denote beginning and final point of the path s and X_F denotes the Hamiltonian vector field of F.

---

[10] Technically paths and surfaces must be semi-analytic and compactly supported in order for that to be true [1; 2; 21] but we will not go into these details here.

where $M = \Sigma \times \mathbb{R}$ (for Σ an arbitrary Riemann surface), ω is an $SU(2)$-connection and the triad e is an $su(2)$-valued 1-form. The gauge symmetries of the action are the local $SU(2)$ gauge transformations

$$\delta e = [e, \alpha]\,, \qquad \delta\omega = d_{\omega}\alpha, \tag{15.3}$$

where α is a $su(2)$-valued 0-form, and the "topological" gauge transformation

$$\delta e = d_{\omega}\eta, \qquad \delta\omega = 0, \tag{15.4}$$

where d_{ω} denotes the covariant exterior derivative and η is a $\mathtt{su(2)}$-valued 0-form. The first invariance is manifest from the form of the action, while the second is a consequence of the Bianchi identity, $d_{\omega}F(\omega) = 0$. The gauge symmetries are so large that all the solutions to the equations of motion are locally pure gauge. The theory has only global or topological degrees of freedom.

Upon the standard $2+1$ decomposition, the phase space in these variables is parametrized by the pull back to Σ of ω and e. In local coordinates one can express them in terms of the two-dimensional connection A_a^i and the triad field $E_j^b = \epsilon^{bc} e_c^k \delta_{jk}$ where $a = 1, 2$ are space coordinate indices and $i, j = 1, 2, 3$ are $su(2)$ indices. The Poisson bracket is given by

$$\{A_a^i(x), E_j^b(y)\} = \delta_a^{\ b}\ \delta^i_{\ j}\ \delta^{(2)}(x, y). \tag{15.5}$$

Local symmetries of the theory are generated by the first class constraints

$$D_b E_j^b = 0, \quad F_{ab}^i(A) = 0, \tag{15.6}$$

which are referred to as the Gauss law and the curvature constraint respectively. This simple theory has been quantized in various ways in the literature [5], here we will use it to introduce the spin foam representation.

15.3.2 Spin foams from the Hamiltonian formulation

The physical Hilbert space, $\mathscr{H}_{\text{phys}}$, is defined by those "states in $\mathscr{H}_{\text{kin}}$" that are annihilated by the constraints. As discussed in the chapter by Thiemann (see also [2; 4]), spin network states solve the Gauss constraint – $\widehat{D_a E_i^a}|s\rangle = 0$ – as they are manifestly $SU(2)$ gauge invariant. To complete the quantization one needs to characterize the space of solutions of the quantum curvature constraints $\widehat{F}_{ab}^i$, and to provide it with the physical inner product. As discussed in Section 15.2 we can achieve this if we can make sense of the following formal expression for the generalized projection operator P:

$$P = \int D[N] \exp(i \int_{\Sigma} \mathrm{Tr}[N\widehat{F}(A)]) = \prod_{x\subset\Sigma} \delta[\widehat{F(A)}], \tag{15.7}$$

(%

Pag. 389(411)

For instance the Gauss law of Yang-Mills theory of gravity written in the new variables formulation.

Pag 512(534)

— Poincaré algebra

— GR is an example of a constrained theory: there is a set of first class constraints that encode the diffeomorphism symmetry.¹

Gauss Law

Pag. 558(580)

Gauss's law constraint which requires that

$$D_a \tilde{E}^a_i = D_a \frac{\delta S(A^i)}{\delta A^i_a} = 0 \quad (28.15)$$

— Tomás Ortín — Gravity and Strings .. (2015).pdf

— Modifications of Einstein's Theory of Gravity .. (2015).pdf

-256- - -* ! - BB 211 ! > ! !A 3
! " ! " ! " 1! J J ! " ! ! D " H A 3 / 1"
(% / ' 7! ! G! 3 !A 3

The transformation has to be such that it implements the evolution equations as operatorial relations acting on the space of wavefunctions in the Heisenberg representation, where

$$U(q|q') = \langle n+1, q'|n, q\rangle, \tag{20.11}$$

and where $|n+1, q'\rangle$ and $|n, q\rangle$ are the eigenvectors of the configuration operators $\hat{q}$ in the Heisenberg representation at levels $n+1$ and n respectively. The evolution equations take the form,

$$\begin{aligned} \langle n+1, q|\hat{q}_{n+1} - f(\hat{q}_n, \hat{p}_n)|n, q'\rangle &= 0, & (20.12)\\ \langle n+1, q|\hat{p}_{n+1} - g(\hat{q}_n, \hat{p}_n)|n, q'\rangle &= 0, & (20.13) \end{aligned}$$

with f, g the quantum evolution equations, which are chosen to be self-adjoint in order for the transformation to be unitary. Explicit examples of this construction for cosmological models can be seen in [17]. If at the end of this process one has constructed a transformation that is truly unitary the quantization is complete in the discrete space and one has a well defined framework to rigorously compute the conditional probabilities that arise when one uses a relational time to describe the physical system. This is a major advantage over attempts to construct the relational picture with systems where one has constraints. There are some caveats to this construction that are worth pointing out. As we mentioned, our construction generically yields discrete theories that are constraint-free. To be more precise, the theories do not have the constraints associated with space-time diffeomorphisms. If the theory under consideration has other symmetries (for instance the Gauss law of Yang–Mills theory or gravity written in the new variable formulation), such symmetries may be preserved upon discretization (we worked this out explicitly for Yang–Mills and BF theory in [4]). The resulting discrete theory therefore will have some constraints. If this is the case, the above construction starts by considering as wavefunctions states that are gauge invariant and endowed with a Hilbert space structure given by a gauge invariant inner product. The resulting theory has true (free) Lagrange multipliers associated with the remaining constraints. The unitary transformation will depend on such parameters. An alternative is to work in a representation where the constraints are solved automatically (like the loop representation for the Gauss law). There one has no constraints left and the inner product is the kinematical one in the loop representation and the unitary transformation does not depend on free parameters. Other issues that may arise have to do with the fact that in many situations canonical transformation do not correspond quantum mechanically to unitary transformations. This problem has been discussed, for instance, by Anderson [1]. He noted that the only canonical transformations that can be implemented as unitary transformations are those that correspond to an isomorphism of a phase space into itself. This is important for the discrete theories in the following way. If one has a continuum constrained theory, its physical phase space is on the constraint surface. The discrete theories have a phase space that includes the constraint surface of the continuum theory. However, the discrete phase space variables cover only a subspace of the kinematical phase space of the continuum theory. There are inaccessible sectors that correspond to complex values of the Lagrange multipliers in the discrete theory. Therefore, in order that the canonical transformation of the discrete theory be an isomorphism, one may have to choose a physical Hilbert space for the discrete theory that is a subspace of the kinematical space instead of just taking it to be coincident. This has to be done carefully, since restricting the Hilbert space may imply that some physical quantities fail to be well defined in the physical Hilbert space. We have explored some of these issues in some quantum mechanical models that have a relational description. We have shown that one can successfully recover the traditional quantum mechanical results in a suitable continuum limit by carefully imposing a restriction on the kinematical Hilbert space, and that one can define variables that approximate any dynamical variable of the continuum theory in the continuum limit in the restricted Hilbert space (see [6]).

20.5 The quantum continuum limit

As we argued in the discussion of the model analyzed by Rovelli, a good measure of how close one is to the continuum theory in a given solution of the discrete theory is to evaluate the constraint of the continuum theory. Such constraint is only exactly satisfied in the continuum limit. An alternative way of presenting this is to consider the construction of a "Hamiltonian" such that exponentiated would yield the unitary evolution between n and $n+1$, $\hat{U} = \exp(i\hat{H})$ where $\hbar = 1$ and $\hat{H}$ has units of action. Such a Hamiltonian can only be constructed locally since in some points of the evolution the logarithm of the unitary transformation is not well defined. Such a Hamiltonian can be written as a formal expansion in terms of the constraint of the continuum theory (a way of seeing this is to notice that in the continuum limit this Hamiltonian has to vanish since it incorporates the timestep). If one chooses an initial state such that $\langle\hat{H}\rangle \ll 1$ the evolution will preserve this ($\hat{H}$ is an exact constant of the motion). This will continue until one reaches a point where $\hat{H}$ is not well defined. The evolution will continue, but it will not necessarily remain close to the continuum limit. In certain cosmological examples this point coincides with the point where the continuum theory has the singularity, for example [17]. Therefore a first condition on the quantum states in the continuum limit $\langle\hat{H}\rangle \ll 1$. A second condition is that the expectation values of the physical variables should not take values in the points where $\hat{H}$ is not well defined. A third condition is not to

The discussion can be made more precise when addressed in the Hamiltonian formalism. GR is an example of a constrained theory: there is a set of first class constraints that encode the diffeomorphism symmetry.[1] Observable quantities are functions on phase space that commute with the constraints. It is pretty hard to construct a general complete set of observables. However, taking advantage of the fact that physics should be relational allowed us to construct a large set of such observables [13].

To simplify the analysis let us consider the relativistic free particle: in this case, we have time reparametrization invariance, encoded in the mass shell constraint $H = p^2 - m^2 = 0$. It is easy to construct the algebra of observables: it is given by the Poincaré algebra $\{J_{\mu\nu},\ p_\mu\}$. This set of observables does not include the important notion of position. To define this concept, we need to introduce the following Rovelli terminology of a *partial observable*[2] $b(\tau)$ as a clock. A natural observable[3] is then the value of another partial observable $a(\tau)$ when $b(\tau)$ is T. If separately a and b are not observable, since not commuting with the mass shell constraint, the quantity

$$a(b^{-1}(T)) = \int d\tau a(\tau)\dot{b}(\tau)\delta\left(b(\tau) - T\right) \tag{26.1}$$

is clearly time reparametrization invariant, and therefore observable. For example, if we take b to be x_0, and a to be x_i, we obtain the trajectories of the relativistic particle in terms of the time x_0:

$$x_\mu(T) = x_\mu + \frac{p_\mu}{p_0}(x_0 - T).$$

Notice that of course this observable can be constructed from the Poincaré algebra [14].

In the language of constrained mechanics, $b(\tau) = T$ is a gauge fixing, or a second class constraint. From the physics point of view, the degree of freedom b is the reference frame. Obviously there is the issue of the invertibility of b. In general the choice of clock might not lead to a function which is invertible everywhere. This means that the clock ceased to be a good clock. This issue has to be studied in a case per case analysis.

Introducing second class constraints means that we can reduce the phase space to obtain the physical phase space. The reduced symplectic form is called the Dirac bracket, and is not in general identical to the canonical Poisson bracket. This leads to complications when one wants to quantize such system.

---

[1] There might be more constraints according to the choice of variables. For example, if using the pair (tetrad, connection), there is also the Gauss law.

[2] Let be f a function on phase space, not commuting with the first class constraint H, then we define $f(\tau) = e^{\tau\{H,.\}}f$.

[3] That is the outcome of a measurement or, according to Rovelli, a complete observable [12].

The general theory of quantization of a constrained system has been set up by Dirac [15]. Let us deal only with a set of first class constraints $\{C_i\}$. We first quantize the algebra of partial observables $\mathcal{A}_k$ and then construct the kinematical Hilbert space $\mathcal{H}_k$ carrying the representation of $\mathcal{A}_k$. We quantize then the constraints C_i:

$$\mathcal{A}_k \to \hat{\mathcal{A}}_k,\ C_i \to \hat{C}_i.$$

In general it might be difficult to construct $\mathcal{H}_k$ but also to quantize the constraints C_i, since they might be non-polynomial functions of partial observables. This is precisely what happened in the first try to quantize GR [16].

The great accomplishment of LQG was to describe GR in terms of variables that allowed one to construct the Hilbert space and quantize the constraints.

From the kinematical Hilbert space, we construct physical Hilbert $\mathcal{H}_{\text{phys}}$ space, which is in the kernel of the constraints $\hat{C}_i$. For this, it is useful to introduce the projector[4]

$$P : \mathcal{H}_k \to \mathcal{H}_{\text{phys}} \equiv \{P|\psi\rangle_k\} \text{ with } P \sim \int d\lambda e^{i\lambda\hat{C}} \sim \delta(\hat{C}).$$

The physical quantum observables are also obtained upon projection

$$\hat{A} \in \hat{\mathcal{A}}_k \to P\hat{A}P \in \hat{\mathcal{A}}_{\text{phys}}.$$

In order to have some physical quantities, we can construct the relational observables analog to (26.1):

$$P\delta(\hat{B}-b)\hat{A}\delta(\hat{B}-b)P, \text{ with } \hat{A}, \hat{B} \in \hat{\mathcal{A}}_k,$$

where $\delta(\hat{B}-b)$ denotes the projection of $\hat{B}$ on the eigenspace with eigenvalue b. The degrees of freedom $\hat{B}$ can be identified as a quantum reference frame (QRF): the physics of QG should be understood in terms of QRF. This new type of physics is extremely rich and interesting to explore. Since in the context of QG it is a bit hard to explore this, we can look for some toy models to mimic this structure.

In fact interestingly the notion of QRF has already been introduced in quantum information theory (QIT), in concrete models that can also be experimentally tested! Let us choose, for example,[5] a quantum universe made of N qubits $\vec{\sigma}^i$,[6] which is globally invariant under $SO(3)$ rotations [17].

---

[4] It will be a projector if the constraints have zero in their discrete spectrum. If this is not the case, it is not a projector; we need to use distributions. This a mathematical subtlety important in QG, but not relevant to the current discussion.

[5] There are many other little toy models where one can play around to mimic gravitational effects. For example, one can look at *analog gravity* models [20], or constrained harmonic oscillators [21]. Of course all these models are finite dimensional, which (over?)simplifies drastically the analysis.

[6] I use the notational shortcut $\vec{\sigma}^i \equiv \mathbf{1} \otimes \ldots \otimes \vec{\sigma} \otimes \mathbf{1} \otimes \ldots$, the Pauli matrices being at the ith position.

the Ashtekar formulation, which are the $SU(2)_L$ connection A^i_a and its conjugate momenta $\tilde{E}^a_i \approx e \wedge e$ this becomes

$$F^i_{ab} + \frac{\Lambda}{3}\epsilon_{abc}\tilde{E}^c_i = 0. \tag{28.13}$$

One can solve this with a Hamilton–Jacobi function on configuration space [51; 47], which is a function $S(A^i)$ such that $\tilde{E}^a_i = \frac{\delta S(A^i)}{\delta A^i_a}$. This leads to the equation

$$F^i_{ab} + \frac{\Lambda}{3}\epsilon_{abc}\frac{\delta S(A^i)}{\delta A^i_a} = 0. \tag{28.14}$$

There is also the Gauss's law constraint which requires that

$$\mathcal{D}_a\tilde{E}^a_i = \mathcal{D}_a\frac{\delta S(A^i)}{\delta A^i_a} = 0. \tag{28.15}$$

These have a unique solution

$$S(A^i) = -\frac{k}{4\pi}\int Y_{\mathrm{CS}}(A^i) \tag{28.16}$$

where $Y(A^i)$ is the Chern–Simons invariant. Thus we can consider the Chern–Simons invariant to be a time functional on the Euclidean configuration space.

If we choose $\Sigma = S^3$ we find that there is a periodicity due to the property that under large gauge transformations with winding number n

$$\int Y_{\mathrm{CS}}(A^i) \rightarrow \int Y_{\mathrm{CS}}(A^i) + 8\pi^2 n \tag{28.17}$$

This means that the Euclidean configuration space is a cylinder which further implies that *all correlation functions, for any fields in the theory* are periodic in an imaginary time variable given by the Chern–Simons functional $S(A^i)$. But by the KMS theorem, this means that the theory is at a finite temperature. If one works out the periodicity one finds precisely the temperature (28.12).

This applies to the full Quantum Gravity theory because it means that any quantum state on the full configuration space of the theory will be periodic in imaginary time. Thus, with very little effort we greatly extend the significance of the de Sitter temperature. This is an example of the power of seeing General Relativity in terms of connection variables and it is also an example of the importance of topological field theory to the physics of Quantum Gravity.

28.4 The problem of the emergence of classical spacetime

We have just seen that LQG gets several things about gravitational physics right, including the entropy of horizons and the temperature of de Sitter spacetime. There

D -256- - -* ! - BB 9 (* { 7 !A 3 ! "
! D " H A 3 $ 1"

-256- - -* ! - BB " ! = 0 9 3
7 !A 3 ! ?! 4 ! 1 & ! ! D $ 1"
(/ '$ =7) = 7!) 9 3

2.2 The Lovelock and Horndeski Uniqueness Theorems

2.2.1 Lovelock Theory

Our purpose in this lecture is to present Lovelock theory in relation to Horndeski theory. To this end it is mostly sufficient to truncate Lovelock theory to what is usually called Einstein-Gauss-Bonnet (EGB) theory in the literature. Unlike the name suggests, this is the five or six dimensional version of Lovelock theory originally discussed by Lanczos [3]. Let us start with the uniqueness theorem defining Lovelock theory and stick to six dimensions in order to fix notation. The five dimensional theory is identical. Consider $\mathscr{L} = \mathscr{L}(\mathscr{M}, g, \nabla, \nabla^2)$ where $(\mathscr{M}, g)$ is a six dimensional locally differentiable Lorentzian manifold without boundary[3] and ∇ is the Levi-Civita metric connexion over $\mathscr{M}$. The field equations obtained upon metric variation of the action,

$$S^{(6)} = \frac{{M_{(6)}}^4}{2} \int \sqrt{-g^{(6)}} \left[R - 2\Lambda + \alpha \hat{G} \right], \tag{2.2}$$

are unique and admit the following properties:

- they depend on a symmetric two-tensor $\mathscr{E}_{AB}$
- the equations of motion are second-order PDE's with respect to the metric field variables
- satisfying Bianchi identities.

Here, $M_{(6)}$ is the fundamental mass scale in six-dimensional spacetime, $\hat{G}$ is the Gauss-Bonnet density reading,

$$\hat{G} = R_{ABCD} R^{ADCB} - 4 R_{AB} R^{AB} + R^2 , \tag{2.3}$$

and Λ is the cosmological constant. The field equations in vacuum are

$$\mathscr{E}_{AB} = G_{AB} + \Lambda g_{AB} + \alpha H_{AB} = 0 , \tag{2.4}$$

where G_{AB} stands for the standard Einstein tensor. Uppercase Latin indices will refer to six-dimensional coordinates whereas Greek indices will always refer to four dimensional space-time. We have also introduced the Lanczos or second order Lovelock tensor,

$$H_{AB} = \frac{g_{AB}}{2} \hat{G} - 2 R R_{AB} + 4 R_{AC} R^C{}_B + 4 R_{CD} R^C{}_A{}^D{}_B - 2 R_{ACDE} R_B{}^{CDE} . \tag{2.5}$$

---

[3] The result depicted here is easily extended to manifolds with boundaries [12].

Naturally, the Lanczos tensor is also divergence free, $\nabla^A H_{AB} = 0$. It is important to note that, just like GR in four dimensions, i.e. under the same set of hypotheses, EGB theory is the unique and most general metric theory with second order PDE's in five or six space-time dimensions. This is a non-trivial statement since the terms appearing in the action already contain second order derivatives. In a moment we will see that in higher than six dimensions this property is generalized by adding the relevant higher order Lovelock terms. Furthermore in four dimensions the tensor (2.5) is identically zero. Therefore we can note as a prelude that Lovelock theory is the unique massless metric theory in arbitrary dimensions identical to GR with a cosmological constant in four dimensional spacetime.

Before moving on it is useful to discuss some tensorial properties. The Lanczos tensor (2.5) can be elegantly written (in arbitrary dimension) using the following rank four tensor that will be useful to us later on,

$$\begin{aligned} P_{ABCD} = R_{ABCD} + R_{BC}\, g_{AD} - R_{BD}\, g_{AC} - R_{AC}\, g_{BD} + R_{AD}\, g_{BC} \\ +\frac{1}{2} R\, g_{AC}\, g_{BD} - \frac{1}{2} R\, g_{BC}\, g_{AD}, \end{aligned} \tag{2.6}$$

as

$$H_{AB} = -2 P_{ACDE} R_B{}^{CDE} + \frac{g_{AB}}{2}\hat{G}\,. \tag{2.7}$$

The 4 index tensor P_{ABCD} has several interesting tensorial properties. For a start it is divergence free (in all indices) since Bianchi identities of the curvature tensor are simply written as $\nabla^D P_{ABCD} = 0$. It has the same index symmetries as the Riemann curvature tensor. Its bi-tensor obtained by tracing two of its non-consecutive indices yields

$$P^B{}_{ACB} = (D-3) G_{AC}, \tag{2.8}$$

the Einstein tensor. In fact divergence freedom of the Einstein tensor can be seen to originate from this relation. In a nutshell, one can say that P_{ABCD} is the curvature tensor whose bi-tensor is the Einstein tensor, just as the Ricci tensor is the bi-tensor of the Riemann tensor. A last interesting property is that a metric is an Einstein space, $R_{AB} = \frac{g_{AB}}{D} R$, if and only if $P_{ABCD} = R_{ABCD}$.

In four dimensions, the $P_{\mu\nu\rho\sigma}$ tensor is even more very special. Indeed it has all the above properties but, additionally it can be pictured in a very similar way to the Faraday tensor in electromagnetism,

$$\star\, F_{\mu\nu} = \tfrac{1}{2}\epsilon_{\mu\nu\kappa\rho}\, F^{\kappa\rho} \tag{2.9}$$

In analogy here, $P_{\mu\nu\rho\sigma}$ is a 4 tensor, and coincides with the double dual (i.e. for each pair of indices) of the Riemann tensor defined as,

$$P^{\mu\nu}{}_{\rho\sigma} = ({}^\star R^\star)^{\mu\nu}{}_{\rho\sigma} \doteq -\frac{1}{2}\epsilon^{\rho\sigma\lambda\kappa}\, R_{\lambda\kappa}{}^{\xi\tau}\, \frac{1}{2}\epsilon_{\xi\tau\mu\nu}, \tag{2.10}$$

where $\epsilon_{\mu\nu\rho\sigma}$ is the rank 4 Levi-Civita tensor. Finally since in four dimensions we have that $H_{\mu\nu} = 0$ we obtain the Lovelock identity,

$$P_{\alpha\nu\rho\sigma} R_{\beta}{}^{\nu\rho\sigma} = \frac{g_{\alpha\beta}}{4}\hat{G} \tag{2.11}$$

which will be useful to us later on (see [19] for extensions).

In order to define the generic Lovelock densities one can use the elegant language of differential forms [12]. Alternatively we take the route taken by Lovelock using the generalized Kronecker delta symbols; the same route taken later on by Horndeski,

$$\delta^{A_1\dots A_h}_{B_1\dots B_h} = \begin{vmatrix} \delta^{A_1}_{B_1} & \dots & \delta^{A_1}_{B_h} \\ \vdots & & \vdots \\ \delta^{A_h}_{B_1} & \dots & \delta^{A_h}_{B_h} \end{vmatrix} \tag{2.12}$$

$$= h!\delta^{A_1}_{[B_1} \dots \delta^{A_h}_{B_h]} \tag{2.13}$$

which is antisymmetric in any pair of upper or lower indices. In fact we have $\delta^{A_1\dots A_h}_{B_1\dots B_h} = \epsilon_{B_1\dots B_h}\epsilon^{A_1\dots A_h}$ with respect to the Levi-Civita symbols. Once this has been digested the Lovelock densities are the complete contraction of the above with the Riemann curvature tensor,

$$L_{(h/2)} = \frac{1}{2^h}\delta^{A_1 A_2 \dots A_h}_{B_1 B_2 \dots B_h}\; R_{A_1 A_2}{}^{B_1 B_2} \;\dots\; R_{A_{(h-1)} A_h}{}^{B_{(h-1)} B_h} \tag{2.14}$$

As such we can check that $L_{(1)}$ is the Einstein-Hilbert term whereas $L_{(2)}$ is the Gauss-Bonnet combination. This immediately means that for $h > D$ all Lovelock densities vanish. Therefore the Lovelock Langrangian is given by,

$$L = \sum_{h=0}^{k} c_h L_h \tag{2.15}$$

where $k = [(D-1)/2]$. The case $h = D$ is quite special because then the Lovelock density is a topological one. Indeed we can query what is special about Lovelock densities. The answer lies in differential geometry (see for example [20]). One can trace the origin of such terms in the early works of Gauss who measuring geodesic distances noted that scalar (Gauss) curvature of two dimensional surfaces depended only on the first fundamental form, in other words the intrinsic metric of the surface and its derivatives. This was the basis of what he called the Egregium theorem; scalar curvature (unlike other extrinsic curvature components) does not depend on the variation of the normal vector field on the surface i.e. on how the surface is embedded in three dimensional space. Then later on Euler in his work on surface triangulations noted that two dimensional surfaces can be topologically classified by their "Euler" number, χ: $\chi[\mathscr{M}] = 2 - 2h$ where h is the number of topological handles. So one can take an arbitrary surface with no boundary and continuously deform it to a sphere, a torus a double torus and so on.[4] This completely classifies topologically two dimensional surfaces. In other words all topological properties of two dimensional surfaces can be understood or characterized by their Euler number. Gauss and Bonnet essentially related this topological number to a differentiable geometric quantity, the scalar curvature, resulting in the celebrated relation,

$$\chi[\mathscr{M}_2] = \frac{1}{4\pi} \int_{\mathscr{M}} R. \tag{2.16}$$

The Gauss-Bonnet theorem on surfaces has nothing to do with the Gauss-Bonnet term given above (2.3). For our purposes the above Gauss-Bonnet relation means that the Einstein-Hilbert term is in two dimensions is a topological invariant i.e. the Einstein tensor in two dimensional space-time is identically zero. This analogy goes through for all Lovelock terms as a corollary to the works of Chern [21] who generalized the theorem of Gauss and Bonnet to higher dimensions finding the relevant higher order curvature scalars. For example we have,

$$\chi[\mathscr{M}_4] = \frac{1}{32\pi^2} \int_{\mathscr{M}} \hat{G} \tag{2.17}$$

and thus the Lanczos or Gauss-Bonnet density is a topological invariant in four dimensions whose integral is the generalised Euler or Chern topological number. Beware this does not mean that the Gauss-Bonnet scalar is zero or constant in four dimensions. It means that the Lanczos density is identically zero $H_{\mu\nu} = 0$ as we admitted earlier.

Dimensionally extending the Chern scalar densities we obtain the Lovelock densities (2.14) i.e. just those densities whose variation leads to second order field equations. Any higher order derivatives present in the variation of Lovelock densities conveniently end up as total divergent terms and thus do not contribute to the field equations. In a similar way for example, in seven or eight dimensions, the six-dimensional Euler density will be promoted to a Lovelock density of third power in the curvature tensor and so forth. This explains the nice and unique properties of the Lovelock densities and Lovelock theory in general. For more details the reader can consult [12].

2.2.2 *Horndeski Theory*

So much for the moment concerning higher dimensional metric theories. In four space-time dimensions we know that the unique classical metric theory is GR with

---

[4] When a surface has a boundary an analogous result holds.

(%/ $$ 9 7!) ! ! ! 1!

discussion on stability of EGB vacua see [37]). The lower branch is ghost free [38] and is the branch that we will consider from now on omitting the '+' branch. The horizon sections, parametrized by the four dimensional metric $h^{(4)}_{\mu\nu}$ are constrained by the field equations to be Einstein spaces,

$$R^{(4)}_{\mu\nu} = \frac{R^{(4)}}{D-2} h^{(4)}_{\mu\nu} \tag{2.38}$$

But as we can see from the lapse function which has to be a function of the radial variable r, a new geometric condition now appears whereupon the horizon quantity ${R^{(4)}}^2 - 6\hat{G}^{(4)}$ has to be constant. Combining the two conditions gives,

$$C^{\alpha\beta\gamma\mu} C_{\alpha\beta\gamma\nu} = \Theta \delta^{\mu}_{\nu} \tag{2.39}$$

where Θ is a given constant and $C_{\alpha\beta\gamma\nu}$ is the four dimensional Weyl tensor associated to $h^{(4)}_{\mu\nu}$. This is a supplementary condition for EGB theories, the Dotti-Gleiser condition, (2.39) imposed in addition to the usual Einstein space condition (2.38) for higher dimensional general relativity. Clearly then, for EGB theory, the lapse function for the black hole carries a supplementary information particular to the type of horizon section for the black hole solution. For example, the Euclidean Schwarszchild metric is not a legitimate internal metric anymore since it does not verify (2.39). Both of the conditions (2.38) and (2.39) present a geometric similarity in that we ask for (part of) *the curvature tensor to be analogous to the spacetime metric*. The main difference being that the curvature tensor in (2.39) is the Weyl tensor and, given its symmetries, it is actually its square which has to be analogous to the spacetime metric. Clearly horizons with $\Theta \neq 0$ will not be homogeneous spaces and not even asymptotically so. Another interesting point is that the Gauss-Bonnet scalar, whose spacetime integral is the Euler characteristic of the horizon, has to be constant for these solutions to be valid. The Gauss-Bonnet scalar of the internal space then reads $\hat{G}^{(4)} = 4\Theta + 24\kappa^2$ and the potential [33, 34],

$$V(r) = \kappa + \frac{r^2}{12\alpha}\left(1 - \sqrt{1 + \frac{12}{5}\alpha\Lambda - 24\frac{\alpha^2\Theta}{r^4} + 24\frac{\alpha M}{r^5}}\right) \tag{2.40}$$

since $R^{(4)} = 12\kappa$. For $\Theta = 0$, we obtain the black holes first discussed by Boulware and Deser (see [39]). In this case since the Weyl curvature is zero the horizon sections are geometries of constant curvature. Taking $\Lambda = 0$ we note that these black holes are asymptotically flat and are an extension of the higher dimensional version of the Schwarzschild solution. In fact taking the limit of α small and large r one obtains precisely the latter GR solution. Recently these black holes have been reported to have a spin 2 instability for small enough mass parameter [40]. This result has been extended to Lovelock black holes [41]. It is not yet understood what is the physical nature of this "short distance scale instability" and if it is somehow related to thermodynamic instability and quantum Hawking radiation.

Putting this aside we now want to examine cases of Einstein metrics whose squared Weyl curvature is not zero but constant. This is a special case of Einstein metric and the Dotti-Gleiser condition is much like a supplementary requirement. What is already clear is that any such solution will not be asymptotically "usual" as the fall off the relevant term is 5 rather than six dimensional. Indeed notice that the Θ-term in (2.40) has a fall off rate of a five-dimensional Boulware-Deser black hole [39] and is therefore dominant over the "usual" mass term contribution, [42]. We now investigate a simple example which will have interesting four dimensional consequences.

Consider a four-dimensional space which is a product of two 2-spheres,

$$ds^2 = \rho_1^2 \left(d\theta_1^2 + \sin^2\theta_1 d\phi_1^2\right) + \rho_2^2 \left(d\theta_2^2 + \sin^2\theta_2 d\phi_2^2\right) , \tag{2.41}$$

where the (dimensionless) radii ρ_1 and ρ_2 of the spheres are constant. The entire six-dimensional metric reads,

$$ds_{(4)}^2 = -V(r)\, dt^2 + \frac{dr^2}{V(r)} + r^2\rho_1^2 \left(d\theta_1^2 + \sin^2\theta_1 d\phi_1^2\right) + r^2\rho_2^2 \left(d\theta_2^2 + \sin^2\theta_2 d\phi_2^2\right) , \tag{2.42}$$

with lapse function

$$V(r) = \frac{R^{(4)}}{12} + \frac{r^2}{12\alpha}\left(1 - \sqrt{1 - 24k^2\alpha - 24\Theta\frac{\alpha^2}{r^4} + 24\alpha\frac{M}{r^5}}\right) . \tag{2.43}$$

In order for (2.42) to be a solution to the equations of motion the spheres have to be of equal radius, $\rho_1 = \rho_2$. This ensures that (2.42) is an Einstein space. The second condition is then immediately verified for a product of 2-spheres. We have $\kappa = \frac{1}{3\rho_1^2} > 0$ and $\Theta = \frac{4}{3\rho_1^4}$. Note that even when the 2-sphere curvature is normalized to $\rho_1 = 1$ then $\kappa \neq 1$. A linear redefinition of the r coordinate then shows that the area of the four dimensional space is reduced compared to the homogeneous 4-sphere. In other words space-time is asymptotically altered by an overall solid angle deficit. This results in a genuine curvature singularity at $r = 0$. Of course when we have $M \neq 0$ there is central curvature singularity at $r = 0$ anyway. But, for (2.42) the $r = 0$ singularity is present *even for zero mass whenever* $\Theta \neq 0$! This is not an artefact of EGB theory. In fact it is easy to see, taking the combined limit of $\alpha \to 0$ and large r, that the resulting GR black hole with $V(r) = \frac{R^{(4)}}{12} + r^2k^2 - \frac{M}{r^3}$ has exactly the same problem at the origin independently of the value of M. For $M = 0$ the GR solution has a naked singularity at the origin. Note again that the lapse function for higher dimensional GR is the same with the higher dimensional Schwarzschild black hole modulo the horizon curvature term. The zero mass solution is singular at the origin wether we are in GR or Lovelock theory. But for Lovelock theory an interesting effect occurs due to the presence of the Θ term in the lapse function.

To see this consider for the moment $M = 0$ in the lapse function (2.42). Then the Θ term in (2.42) is identical to the mass term in the Boulware Deser black hole in five dimensions. Therefore, as we know from the Boulware-Deser solution [39] this extra Θ term generically generates an event horizon cloaking the central $r = 0$ singularity as long as $\alpha \neq 0$! In fact the length scale of this event horizon is given by the coupling constant $\alpha \sim length^2$ which we know from string theory effective actions [43] is related to the fundamental string tension α'. One then can interpret this horizon as a higher order 'quantum' cloak of an otherwise naked singularity present in GR. Details for the horizon structure can be found in [34]. We will come back to this solution in order to construct a Galileon black hole.

Let us, before moving on, make some final remarks regarding these solutions. First we should note that most probably these multiple sphere solutions can be unstable to linear perturbations. It has been shown in GR [31] that there is a "balloon instability" whereupon one of the spheres wants to deflate with respect to the other. This geometric effect may remain true in the above EGB version [33] although the perturbation equations for EGB in lesser symmetry change completely compared to GR. It is also probable that this instability may be stabilized by the inclusion of a magnetic field in the relevant solution [44]. Secondly we should note that the above construction involving multiples of equal radius spheres, can be undertaken in arbitrary even dimensional spacetime as long as we truncate Lovelock theory to EGB. If one considers higher order Lovelock terms it is not known under what geometric conditions the horizon sections will be admissible. One may expect a higher order condition of the type (2.39) in third and higher curvature order... We expect the horizon sections to be more and more constrained as higher order Lovelock terms come into play. At the same time since horizon sections will be of higher dimension this will allow for a richer geometry. This is also an open question. Finally, putting it all together we have arrived to the following result concerning EGB theory: *given the anzatz (2.30), the only asymptotically flat solution of six dimensional EGB theory with zero cosmological constant, is the Boulware Deser solution [39]*. This is because whenever $\Theta \neq 0$ the solution is not asymptotically flat for six dimensional space-time. Therefore we can deduce that EGB theory is very similar in this aspect to four dimensional GR lifting the degeneracy present in higher dimensional GR due to the additional elegant geometric condition (2.39).

2.4 From Lovelock to Horndeski Theory: Kaluza-Klein Reduction

In order to apply higher dimensional Lovelock theory to cosmology or gravity in four dimensional space-time one needs some means of approach to four dimensional gravity. There are at least two routes, braneworlds and Kaluza Klein reduction. In the recent past Lovelock theory had an important implication in the braneworld paradigm [45]. Braneworlds consist of higher dimensional spacetimes endowed with a distributional brane where standard matter is localized. The idea "inspired" in rather loose terms from string theory, is that gravity perceives all the space-time dimensions while matter is localized on a four dimensional braneworld. Since the set-up involves junction or matching conditions an essential feature is the number of extra dimensions, namely codimension, yielding for example a wall or string type of defect. There is a long literature of articles on the subject treating codimension one [12] and codimension two braneworlds (see [46] and references within) involving respectively five and six dimensional EGB theory. In particular Lovelock theory permits, due to the generalized junction conditions [47], well defined codimension two braneworld cosmology [48]. This leads to important consequences since in GR one cannot consider distributional sources for cosmological symmetry and codimension 2. Again, the richer structure of Lovelock theory permits solutions with distributional sources not available in higher dimensional GR. For more details on these aspects see [12] and [46] and references within. Here, we will focus on the more classical Kaluza-Klein compactification since it will give us a direct connection to higher order scalar tensor terms, found in Galileon/Horndeski theory. It will also provide a way to obtain exact black hole solutions [49].

It has been known since a long time [50] that a consistent Kaluza Klein reduction of Lovelock theory gives a scalar-tensor theory with higher order derivatives, but crucially, with second order equations of motion. In this sense many of the Galileon terms discussed later on were known from previous work on Kaluza Klein compactifications and braneworlds [51]. This is the direction we will take here. The most generic of Kaluza-Klein reduction to four dimensions has recently been given in the nice paper of [52]. There it has been shown that only up to the third order Lovelock terms contribute to the Kaluza-Klein compactification in four dimensions. Here we will concentrate on EGB theory i.e. up to second order Lovelock theory. We will consider the simplest consistent toroidal compactification giving rise to one extra scalar degree of freedom.

Start by taking D-dimensional Einstein Gauss-Bonnet theory which is the five or six-dimensional Lovelock theory truncated to arbitrary dimension. The arbitrary dimension D will be important when we end up promoting dimension from a positive integer to a real parameter once we have undertaken a consistent Kaluza-Klein reduction. We have the EGB action with a cosmological constant,

$$S = \frac{1}{16\pi G_N} \int d^D x \sqrt{-g} \left[-2\Lambda + R + \hat{G} \right] \tag{2.44}$$

Consider now the simplest but consistent diagonal reduction along some arbitrary n-dimensional internal curved space $\tilde{\mathbf{K}}$. We aim to reduce this theory down to four space-time dimensions with $D = 4 + n$:

$$ds^2_{(4+n)} = d\bar{s}^2_{(4)} + e^{\phi} d\tilde{K}^2_{(n)} \,. \tag{2.45}$$

This particular frame is chosen in such a way as so there is no conformal factor of ϕ in front of the four-dimensional metric. As such the asymptotic character (i.e. radial fall off) of a Lovelock D dimensional solution will be similar to the four dimensional one. All terms with a tilde refer to the curved n-dimensional internal space, while terms with a bar refer to the (4)-dimensional space-time. One can show for the given metric Anzatz that the KK reduction for n arbitrary is consistent, i.e. that the reduced equations of motion are derived from the reduced action [53]. This reduction is therefore generalised in the manner defined in [54, 55]. The important result of this is that the integer n corresponding to the compact Kaluza-Klein space can be analytically continued to a real parameter of the reduced action. Naturally n corresponds to a dimension only for n integer. The solutions from the four dimensional point of view are still solutions of the resulting effective action for arbitrary n. The four dimensional effective action reads after integrating out the internal space,

$$\begin{aligned}\bar{S}_{(4)} = \int d^4x\, \sqrt{-\bar{g}}\, e^{\frac{n}{2}\phi} \Big\{ & \bar{R} - 2\Lambda + \alpha\bar{G} + \frac{n}{4}(n-1)\partial\phi^2 - \alpha n(n-1)\bar{G}^{\mu\nu}\partial_\mu\phi\partial_\nu\phi \\ & -\frac{\alpha}{4}n(n-1)(n-2)\partial\phi^2\nabla^2\phi + \frac{\alpha}{16}n(n-1)^2(n-2)\left(\partial\phi^2\right)^2 \\ & + e^{-\phi}\tilde{R}\left[1 + \alpha\bar{R} + \alpha 4(n-2)(n-3)\partial\phi^2\right] + \alpha\tilde{G}e^{-2\phi}\Big\}\ , \end{aligned} \tag{2.46}$$

For $\alpha = 0$ this effective action is just the usual toroidal KK effective action. The higher order Gauss-Bonnet term gives rise to several higher order scalar-tensor Galileon (or equivalently Horndeski) terms, [14, 15, 56, 57] with very particular potentials. The Galileon field ϕ can then simply be understood to be the scalar field parametrising the volume of the internal space.

Indeed, apart from the usual lower order terms appearing in standard Kaluza Klein compactification of Einstein dilaton theories, we see the emergence of several higher order terms. For a start we have the four dimensional Gauss Bonnet term $\bar{G}$ which will contribute to the scalar field variation although it is a topological term for four dimensional GR. Secondly we have $\bar{G}^{\mu\nu}\partial_\mu\phi\partial_\nu\phi$ involving the coupling of the Einstein tensor with the kinetic term. Here, rather than metric-scalar interaction, as for the standard kinetic term of ϕ we have a curvature-scalar interaction which we will see has very interesting consequences in the forthcoming section. This term has equations of motion of second order essentially due to the divergence free property of the Einstein tensor $G_{\mu\nu}$. For example if one considers $\bar{R}^{\mu\nu}\partial_\mu\phi\partial_\nu\phi$ this is not true. It has also shift symmetry in the scalar field typical of certain Galileon terms. Furthermore we have, what is often called the DGP term, $\partial\phi^2\nabla^2\phi$ appearing in the decoupling limit of the DGP model [25] and then the standard Galileon term $\left(\partial\phi^2\right)^2$ which are also shift symmetric in ϕ. The last line in the effective action takes part only for a curved internal space in the face of Ricci and Gauss Bonnet curvature. Reducing from the EGB action yields terms up to quartic order in derivatives (either of the metric or the scalar, or a mixed combination of the two). Reducing higher order Lovelock densities yields terms with a higher number of derivatives. A typical example is the higher order permissible curvature-scalar interaction,

$$P^{\mu\nu\alpha\beta}\nabla_\mu\phi\nabla_\alpha\phi\nabla_\nu\nabla_\beta\phi \tag{2.47}$$

which involves six derivatives and one can show [9] originates from the Kaluza-Klein reduction of the third order Lovelock density [52]. Again the reader will recollect the divergence freedom of the double dual tensor giving second order field equations. Taking for example $R^{\mu\nu\alpha\beta}\nabla_\mu\phi\nabla_\alpha\phi\nabla_\nu\nabla_\beta\phi$ would fail this Galileon property.

Although this effective action is very complex, and its field equations even more so, it is "simple" to generate solutions for the above (2.46) in four dimensions [53]. One starts from a convenient Lovelock solution in D dimensions. Since we want the four dimensional solution to have, at least locally, spherical horizon sections we have to consider a solution where the $(D-2)$ dimensional horizon sections are $(D-2)/2$-products of two spheres. This is precisely the extension of the six dimensional solution we discussed in the previous section (2.42) generalized to arbitrary dimensions [53]. The solution reads,

$$d\bar{s}^2_{(4)} = -V(R)dt^2 + \frac{dR^2}{V(R)} + \frac{R^2}{n+1} d\bar{K}^2_{(2)}\,, \tag{2.48}$$

$$V(R) = \kappa + \frac{R^2}{\tilde{\alpha}_r}\left[1 \mp \sqrt{1 - \frac{2\tilde{\alpha}_r}{l^2} - \frac{2\tilde{\alpha}_r^2\kappa^2}{(n-1)R^4} + \frac{4\tilde{\alpha}_r m}{R^{3+n}}}\right], \tag{2.49}$$

$$\tilde{\alpha}_r = 2\alpha n(n+1), \qquad \frac{1}{\ell^2} = \frac{-2\Lambda}{(n+2)(n+3)} \tag{2.50}$$

$$e^\phi = \frac{R^2}{n+1}\,, \tag{2.51}$$

Here, n is the dimension of the internal space minus one 2-sphere in other words, $n = D-4$. This is the higher dimensional interpretation of n but once the solution is written out we simply take n an arbitrary real number and (2.48) is still an exact solution and n parametrizes the theory. In our notation here $\kappa = 0, 1, -1$ is the normalised horizon curvature and we have redefined for this section the constants $\tilde{\alpha}_r$ and ℓ. Taking carefully the $\tilde{\alpha}_r \to 0$ limit, gives a standard Einstein dilaton solution with a Liouville potential [49]. Set $\Lambda = 0$, $\kappa = 1$ and let us start by making some qualitative remarks describing properties of the solution without entering into technical details. Note that, taking carefully the $n = 0$ limit switches off the scalar field and the higher-derivative corrections, and we obtain pure GR in (2.46) and a Schwarzschild black hole (2.48). This is particularly interesting since the scalar-tensor solution given above for arbitrary n is a continuous deformation of the Schwarzschild solution. When n is in the neighborhood of zero we are closest to the GR black hole. As we hinted in the previous section the topology of the solution is not that of GR. Indeed the warp factor of the 2-sphere, in (2.48), is recovered only at $n = 0$, i.e. the GR limit. Otherwise the area of the reduced spherical horizon is given by $\frac{4\pi R^2}{n+1}$ rather than the 2-sphere area, $4\pi R^2$. This is again a solid deficit angle (and not a conical deficit angle) the same one we encountered for the Lovelock solution in the previous section. As stressed in the previous sections this will give, at $R = 0$, a true curvature singularity even if $m = 0$. For large R, we have a spacetime metric very similar to that of a gravitational monopole [58]. Expanding 2.49 for small $\tilde{\alpha}_r$ and large R gives,

$$V(R) = 1 + \frac{\tilde{\alpha}_r}{(n-1)R^2} - \frac{2m}{R^{n+1}} + \dots \tag{2.52}$$

This solution is reminiscent of a RN black hole solution where the role of the electric charge is undertaken by the leading horizon curvature correction in $\tilde{\alpha}_r$. This is the particular Θ term we discussed in the previous section. This term dominates the mass term close to the horizon and for $n < 1$. Note that it can be of negative sign depending on the value of n and $\tilde{\alpha}_r$. The further we are from $n = 0$, the GR limit, the further we deviate from a standard four-dimensional radial fall-off. The first important question we want to deal with is the central curvature singularity at $R = 0$, which is due to the solid deficit angle and is present even if $m = 0$. Also note that whenever the square root in the lapse function (2.49) is zero we also have a branch singularity which is also a dangerous curvature singularity. Setting $m = 0$, we find that for $-1 < n < 1$ and $\tilde{\alpha}_r > 0$ the singularity at $R = 0$ is covered by an event horizon created by the higher-order curvature correction. In its absence ($\tilde{\alpha}_r = 0$), this solution would have been singular...The UV (small R) behaviour of the solution is therefore regularised by the presence of the higher-order terms. If $n > 1$ or $n < -1$, then $\tilde{\alpha}_r < 0$ is needed in order to preserve the event horizon. The remaining cases are singular.

Now let us switch on the mass, $m \neq 0$. Whenever $\tilde{\alpha}_r > 0$, we have a single event horizon. When $-1 < n < 1$, there is no branch singularity however small m is. On the contrary, when $n > 1$, the mass is bounded from below in order to avoid a branch singularity:

$$m > \left(\frac{2}{n+3}\right)^{\frac{n+3}{4}} \frac{\tilde{\alpha}_r^{\frac{n+1}{2}}}{n-1}\,. \tag{2.53}$$

When $n < -1$, the solution is also a black hole but the mass term is not falling off at infinity. The region of most immediate interest is whenever n is small but not zero.

The black hole properties are rather different for $\tilde{\alpha}_r < 0$. When $-1 < n < 1$, there is an inner and an outer event horizon as long as the following condition is fulfilled:

$$\left(\frac{1}{2}\right)^{\frac{n+1}{2}} < \frac{m(1-n)}{|\tilde{\alpha}_r|^{\frac{n+1}{2}}} < \left(\frac{2}{n+3}\right)^{\frac{n+3}{4}}\,. \tag{2.54}$$

When $n > 1$, a single event horizon exists, covering a single branch singularity with $R_s < R_h$.

Overall we can say that the KK solution given here has an interesting horizon structure and presents again a quantum cloaking of an otherwise Einstein-Dilaton singular solution. It is however not of ordinary asymptotics bifurcating in this way the no hair paradigm for Galileons [59]. In the last section we will see a way to construct asymptotically ordinary solutions with fake black hole hair.

2.5 Self-tuning and the Fab 4

As we saw in the previous sections Horndeski or Galileon theory encompasses all the possible (single) scalar tensor terms one can consider in order for the equations of motion to be of second order. This is an essential requirement for a well-defined classical modification of gravity [5]. In this section we will question which of the terms in scalar-tensor theory have "self-tuning" properties. Self-tuning is a rather simple and quite old idea with application to the cosmological constant problem. The basic principle consists of finding solutions for flat (or possibly maximally symmetric vacua) of some gravity action endowed with a bulk cosmological constant, independently of the value of the cosmological constant in the action. In order for self-tuning (and not fine tuning) to be effective the cosmological constant should not be fixed with respect to any of the coupling constants in the gravitational action. The idea then is that the cosmological constant is absorbed by a dynamical solution involving the non-trivial scalar field without affecting the gravitational background. This can only be a "partial" solution to the cosmological constant problem since radiative corrections will destabilize this vacuum solution beyond a certain energy scale, the cutoff of the effective gravity theory. It is however an interesting first step especially since no theories were known, before [9], to have such a property without some hidden effective fine tuning of the action coupling constants as for example in codimension one braneworld models (see for example [60]). We should note that recently there has been considerable progress on protecting the cosmological constant from standard model radiative corrections [61] and we refer the interested reader to this article for the model in question which interestingly is a rather minimal extension of GR. Rather, for our purposes, having at hand the general scalar tensor theory we will formulate the following question: is there a subset of Horndeski theory with self-tuning properties? The answer is affirmative as shown in [9], yielding a rather simple and neat geometrical theory which was dubbed by the authors as Fab 4 theory. We start by presenting the theory and then give a specific self tuning solution which elegantly and non-technically gives the general idea. We close the section by showing a simple method to obtain regular black hole solutions in fab 4 and Horndeski theory, independently of self-tuning.

The Fab 4 potentials make up the most general scalar-tensor theory capable of self-tuning. They are given by the following geometric terms,

$$\mathscr{L}_{john} = \sqrt{-g}V_{john}(\phi)G^{\mu\nu}\nabla_\mu\phi\nabla_\nu\phi \tag{2.55}$$

$$\mathscr{L}_{paul} = \sqrt{-g}V_{paul}(\phi)P^{\mu\nu\alpha\beta}\nabla_\mu\phi\nabla_\alpha\phi\nabla_\nu\nabla_\beta\phi \tag{2.56}$$

& B (% /#%
-256- - -* ! - B 2 V = !A 3 !
H A 3 % 1"
(B$: " . ! 1! A " •
*D ~ ? A " UUU

This procedure gives us an easy way to determine the metric on a sphere. Simply take the metric (2) $ds^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\varphi^2$ and set $r = a$. Then $dr = 0$, so that ds^2 collapses to $a^2(d\theta^2 + \sin^2\theta d\varphi^2)$. From 3-dimensional flat space we have "lost" the coordinate r and gone to a 2-dimensional curved space with coordinates $x^\mu = (x^1, x^2) = (\theta, \varphi)$. Without loss of generality, we can take a as the unit of distance and set $a = 1$. So, on the unit 2-dimensional sphere S^2

$$ds^2 = d\theta^2 + \sin^2\theta d\varphi^2 \tag{3}$$

with a metric given by $g_{11} = g_{\theta\theta} = 1$, $g_{22} = g_{\varphi\varphi} = \sin^2\theta$, and $g_{12} = g_{\theta\varphi} = g_{21} = g_{\varphi\theta} = 0$.

The take-home message here is that curved space is just a skip and a hop away from the familiar change of coordinates. This is fortunate for students of physics: when you learned to change coordinates, you were actually also learning about curved spaces. We are now going to develop a general formalism for changing coordinates. Even though you already know how to change coordinates, it pays to learn this formalism, because we can also use it to study curved space and curved spacetime (which, as you have surely heard, plays a central role in Einstein gravity).

Change of coordinates, curved space, and curved spacetime: basically the same deal, as you will see.

How do we know whether a space is curved or not?

This raises an exceedingly interesting and crucial question: given a space with the metric $g_{\mu\nu}(x)$, how do we know whether it is curved or flat?

A complicated looking metric does not necessarily mean that the space is curved, since somebody could have simply chosen an especially gunky coordinate system. It could be flat space in disguise. To forcefully bring home this point, I invite you to consider $ds^2 = (1 + u^2)du^2 + (1 + 4v^2)dv^2 + 2(2v - u)dudv$ and $ds^2 = (1 + u^2)du^2 + (1 + 2v^2)dv^2 + 2(2v - u)dudv$. One describes flat space, the other a space that at some points is violently curved. Which is which?

Puzzled, you reply: "How could I possibly tell?"

That's in fact the correct answer at this stage of this discussion. The two metrics I just gave you look almost identical except for one single 2 → 4. In one of the most famous episodes in mathematics, Carl Friedrich Gauss (1777–1855) solved this problem for 2-dimensional spaces. His work was then generalized by his student Bernhard Riemann (1826–1866). Later, in chapter VI.1, given any metric in any number of dimensions, you will be able to calculate, and even better, to train the computer to calculate, something called the Riemann curvature tensor, which will tell you once and for all if the space is flat or curved. No more thinking involved! Gauss and Riemann did it for you.

But for now, let me ask you to think about two simple examples in good old 2-dimensional space, for which our intuition is allegedly pretty good. We know that $ds^2 = dr^2 + r^2 d\theta^2$ describes flat space. Consider

$$ds^2 = d\rho^2 + \sin^2\rho \, d\theta^2 \tag{4}$$

Is the space being described flat or curved? Or consider the space described by

$$ds^2 = \cos^2 \rho \, d\rho^2 + \sin^2 \rho \, d\theta^2 \tag{5}$$

Is it flat or curved? You should think about this before reading on. The answers are given in appendix 1.

Remember the civilization of mites in the prologue? You are in the same position as the mite professors of geometry: they can measure the distance between infinitesimally separated points, and from that they have to figure out whether their world is curved. We will face the same problem as the mites when we get to cosmology in parts V and VI.

The logic of differential geometry

Differential geometry, as developed by Gauss and Riemann, tells us that given the metric, we can calculate the curvature. The logic goes as follows. The metric tells you the distance between two nearby points. Integrating, you can obtain the distance along any curve joining two points, not necessarily nearby. Find the curve with the shortest distance. By definition, this curve is the "straight line" between these two points. Once you know how to find the "straight line" between any two points, you can test all of Euclid's theorems to see whether our space is flat. For example, as described in the prologue, the mite geometers could now draw a small circle around any point, measure its circumference, and see if it is equal to 2π times the radius. (See appendix 1.) Thus, the metric can tell us about curvature.

Take an everyday example: given an airline table of distances, you can deduce that the world is curved without ever going outside. If I tell you the three distances between Paris, Berlin, and Barcelona, you can draw a triangle on a flat piece of paper with the three cities at the vertices. But now if I also give you the distances between Rome and each of these three cities, you would find that you can't extend the triangle to a planar quadrangle (figure 1). So the distances between four points suffice to prove that the world is not flat. But the metric tells you the distances between an infinite number of points.

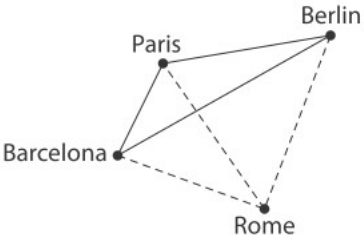

Figure 1 The distances between four cities suffice to prove that the world is not flat.

(%/

✡

Pág. 29(45) - EGB

Einstein-Gauss-Bonnet theory.

Pág. 31(47)

One can trace the origin of such terms in the early works of GAUSS who measuring geodesic distances noted that scalar (Gauss) curvature of two dimensional surfaces depended only on the first fundamental form, in other words the intrinsic metric of the surface and its derivatives.

This was the basis of what he called the Egregium theorem; scalar curvature (unlike (diferente) other extrinsic curvature (components) does not depend on the variation of the normal vector field on the surface i.e. on how the surface is embedded in three dimensional space.

Then later, for Euler in his work on surface triangulations noted that two dimensional surfaces can be topologically classified by their "Euler" number,

$X: \chi[M] = 2-2h$ where h is

(%

the number of topological handles.

Pág. 32(48) → página toda

_ Gauss-Bonnet theorem

– Einstein-Hilbert term is in two dimensions is a topological invariant, i.e. the Einstein tensor in two dimensional space-time is identically zero.

Pág. 39(55)

the Gauss-Bonnet scalar of the internal space...

Pág. 43(59)

the last line in the effective action takes part only for a curved internal space in the face of Ricci and Gauss-Bonnet curvature./

— A. Zee - Einstein gravity (2013).pdf

Pág. 65(88) -

How do we know whether a space is curved or not?

Obs: Livro todo!!!

B
-256- - -* ! - B 2 V = !A 3 !
H A 3 % 1"
(% $ B A " 1! 7! ! " 5 !

1.6

Curved Spaces: Gauss and Riemann

Fear of curves

Surely you have heard that Einstein gravity involves curved spacetime. In my experience, the very mention of the Riemann curvature tensor, which we will come to in due time, inspires fear and trembling in many beginning students of general relativity. “I was doing well in the course,” students would moan, “until we started doing Riemannian geometry!” I actually have a great deal of sympathy for these students: very little in their background prepares them for the Riemann curvature tensor. Unless they have taken an exceptionally good mechanics course, with studies of particle motion on curved surfaces, they typically have no prior exposure to the massive amount of rather technical material needed to calculate the Riemann curvature tensor. Dear reader, take heart: even Einstein had to struggle to master differential geometry.

Given this frightening reputation of Riemannian geometry, we will take baby steps toward curvature. My pedagogical strategy is to first deal with surfaces that you can practically imagine holding in your hands.

Introducing Professor Flat

As I was encouraging a fearful student, Professor Flat came ambling along. A mild mannered man with a somewhat disheveled look, he inquired gently, “Why all the trembling?” We told him, and he nodded. Then he asked the fearful student, “Do you know why it took humans so long to realize that the world was round?”

FS: “Because the world is locally flat, and humans can’t walk very fast.”

PF: “Exactly! In everyday life, we have no need to know that the world is actually round, as long as the distance scale of interest is small compared to the earth’s radius. If you focus on a small enough region on any surface—or any space in any dimension for that matter—it's going to look flat, and you could simply study the small deviation from flat space. Nothing terribly frightening at all!"

Tangent plane

So indeed, let's follow Professor Flat's advice. To get some feel for curvature, let us keep things simple and be content to take some nice curved space, roundish like the sphere. Imagine approaching it with a flat plane until it touches the surface at some point P. The plane is then known as the tangent plane. To focus our minds, consider a sphere of radius L. Since all points on it are equivalent, we might as well have the plane touch the south pole and be perpendicular to the z-axis joining the north and south poles. The southern hemisphere is defined by $z = -\sqrt{L^2 - x^2 - y^2} + L$, where we added the constant L to the usual definition of z so that for convenience, $z = 0$ at the point P we are interested in (the south pole in this case). Near the south pole, $z \simeq \frac{1}{2L}(x^2 + y^2)$: the sphere is locally a parabolic bowl and is well approximated by the tangent plane. It is in this sense that Professor Flat says that the surface is locally flat.

So, consider the tangent plane touching the roundish surface we are studying at some point P. See figure 1. Let z denote the coordinate perpendicular to the plane, and (x, y) the coordinates in the plane, with the point P coordinatized by $(x, y) = (0, 0)$. Locally, we have a quadratic expansion in the small quantities x and y: $z = \frac{1}{2}ax^2 + cxy + \frac{1}{2}by^2$. Here a, b, c have dimension of inverse length, and thus the local region is defined by x, y small compared to a^{-1}, b^{-1}, c^{-1}. (For the sphere, $c = 0$ and $a = b = 1/L$.) Applying Pythagoras, we obtain the distance squared between two nearby points with coordinates (x, y) and $(x + dx, y + dy)$:

$$\begin{aligned} ds^2 &= dx^2 + dy^2 + dz^2 = dx^2 + dy^2 + [(ax + cy)dx + (by + cx)dy]^2 \\ &\equiv g_{xx}dx^2 + g_{yy}dy^2 + 2g_{xy}dxdy \end{aligned} \tag{1}$$

with the metric

$$g_{xx} = 1 + (ax + cy)^2,\ g_{yy} = 1 + (by + cx)^2,\ g_{xy} = (ax + cy)(by + cx) \tag{2}$$

Note that even for the sphere, the metric regarded as a 2-by-2 matrix contains an off-diagonal term in this locally flat coordinate system. Of course, for $x, y \to 0$, the off-diagonal term vanishes and the metric approaches the identity matrix.

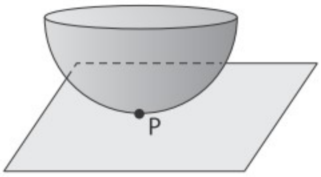

Figure 1 The tangent plane to a curved surface.

We could also write $z \simeq \frac{1}{2}\vec{x}^T M\vec{x}$, with $M = \begin{pmatrix} a & c \\ c & b \end{pmatrix}$ and $\vec{x} = (x, y)$ a column vector which we are writing as a row vector for typographical reasons mentioned in chapter I.3. But we can always rotate the coordinates $\vec{x} = (x, y)$ to some other coordinates $\vec{x}' = (u, v)$ given by $\vec{x}' = \mathcal{R}\vec{x}$. We want the curvature to be an invariant geometric concept not dependent on our coordinate choices. In the new coordinates, M is replaced by $M' = R^{-1}MR$. We know that M possesses two invariant quantities, namely its two eigenvalues μ and ν. But this dovetails perfectly with our discussion in the prologue.

Remember the ant going for her honey? We learned from that story that at a given point, a curved space has an intrinsic and an extrinsic curvature. It all makes sense, then: these two curvatures correspond to the two invariant quantities contained in M. As you will see, we could exploit similar reasoning, asking how things transform under a change of coordinates, to determine the curvature of Riemannian spaces of any dimension.

Let us diagonalize M, so that in the new coordinate basis* $z = \frac{1}{2}\mu u^2 + \frac{1}{2}\nu v^2$. Thus, we have the intuitive result that any surface is locally the sum of two parabolas, $z = \frac{1}{2}\mu u^2 + \frac{1}{2}\nu v^2$, with μ^{-1} and ν^{-1} the radius of curvature of the parabola in the u and in the v direction, respectively. (Expand a circle of radius L around some point, in the same way that we expanded a sphere earlier, $y = -\sqrt{L^2 - x^2} + L \simeq \frac{1}{2L}x^2$, and we see that it is locally a parabola with radius of curvature equal to L.)

Intrinsic versus extrinsic

For a 2-by-2 matrix, its determinant and trace, or equivalently, its two eigenvalues, constitute its two basis-independent attributes. Our insistence that the measure of surface curvature does not depend on whether we use the x, y or the u, v coordinates means that we can have two measures of surface curvature:

$$\text{Intrinsic curvature} = \det M = \mu\nu = ab - c^2 \tag{3}$$

$$\text{Extrinsic curvature} = \left(\tfrac{1}{2}\text{tr}M\right)^2 = \tfrac{1}{4}(\mu + \nu)^2 = \tfrac{1}{4}(a + b)^2 \tag{4}$$

(Note that I have defined the extrinsic curvature to have the same dimension as the extrinsic curvature. The normalization factor $\frac{1}{4}$ is such that the intrinsic and extrinsic curvatures of the sphere, for which $a = b$, $c = 0$, are equal.)

How do I know which is which, intrinsic versus extrinsic curvature? We appeal to the cylinder in the prologue as an example, knowing that it is intrinsically flat. Indeed, for a cylinder† of radius L, $z = -\sqrt{L^2 - x^2} + L$ independent of y and hence $b = c = 0$. We have intrinsic curvature $= 0$ and extrinsic curvature $= a^2 = 1/L^2$, as we would expect. In

---

[*] Note that while M is diagonal, the metric is not.

[†] The cylinder example also underlines the fact that we are always talking about local curvature, as explained in the preceding chapter. Globally, the mites could of course go all the way round and come back to the same place.

Einstein gravity, we are normally only interested in the intrinsic curvature of spacetime, since we can't get out* of spacetime and look at its extrinsic curvature.

Negative curvature

A sphere of radius L has intrinsic and extrinsic curvatures both equal to $1/L^2$. Since every point on the sphere is the same as the south pole, the sphere has constant intrinsic and extrinsic curvatures.

A beginning student named Confusio[1] looked, well, confused. He said: "You mean the curvature at the bottom of a bowl and at the top of a dome are not opposite in sign? Everybody knows that the bowl is curved upward, the dome downward."

Confusio's everyday intuition caused him to think that he was somehow inside the bowl and outside the dome. In fact, the sphere is an infinitely thin surface. There is no conceptual distinction between being on the "inside" and the "outside" of the surface. Of course, we could also turn the bowl upside down. Let's calculate the curvature at "the top of the dome," as Confusio calls it, otherwise known as the north pole. Near the north pole, $z = \sqrt{L^2 - x^2 - y^2} - L \simeq -\frac{1}{2L}(x^2 + y^2)$, so that $a = b = -\frac{1}{L}$, $c = 0$. Thus, the intrinsic curvature is $ab - c^2 = \frac{1}{L^2}$, as expected. Compared to the neighborhood around the south pole, z flips sign, but dz^2 does not, so that the metric, and hence the curvature, have the same value at the north pole and at the south pole, as they should.

Thus, in contrast to everyday perception, a negative curvature surface[2] is one in which the two parabolas bend in opposite directions, $\mu = -\nu$, such as the proverbial saddle. A more contemporary example is the kind of potato chips that come in a cylindrical container. I am surmising that the typical reader of this text is more likely to eat potato chips than to gallop across the steppes with the Golden Horde. An example is a surface that goes like $z = xy$ for $x \simeq 0$, $y \simeq 0$. Then $a = b = 0$, $c = 1$, and $\mu = 1$, $\nu = -1$, and it has negative curvature $R = -1$ at the origin.

Professor Flat: "Confusio, you still look confused. Think of a donut or a bagel, otherwise known as the torus. Do you see that along its hole, the curvature is negative? Along the outer edge, in contrast, the curvature is positive."

Incidentally, it then follows that somewhere on the torus the curvature (we are always talking about intrinsic curvature) vanishes. Do you see where? You could always use the tangent space method here to calculate the curvature and thus check your intuition.

Embedding of curved spaces in higher dimensional flat spaces

In general, we could embed[3] a D-dimensional space in N-dimensional Euclidean space E^N. Write $X^A(x^1, \cdots, x^D)$ for $A = 1, \cdots, N$. For everyday surfaces, $D = 2$ and $N = 3$, but the formalism to be given presently works for arbitrary values of $N > D$. (For the iconic sphere, $(x^1, x^2) = (\theta, \varphi)$ and $(X^1, X^2, X^3) = (X, Y, Z) = (\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta)$.) We see that x represents coordinates on the surface: as x^μ varies, the points $X^A(x)$ sweep out a D-dimensional space in E^N. In this age of spectacular computer graphics, you can easily generate all kinds of interesting surfaces by playing around with three functions $X^A(x)$ of two variables.

---

* Except in some speculative and unproven theories!

Consider two neighboring points: P described by x^μ and Q described by $x^\mu + dx^\mu$. If the Euclidean coordinates of P are given by X^A, then the Euclidean coordinates of Q are given by $X^A + dX^A = X^A + \frac{\partial X^A}{\partial x^\mu} dx^\mu$. Pythagoras gives us the distance squared between P and Q

$$ds^2 \equiv \sum_A (dX^A)^2 = \sum_A \frac{\partial X^A}{\partial x^\mu} dx^\mu \frac{\partial X^A}{\partial x^\nu} dx^\nu \equiv g_{\mu\nu} dx^\mu dx^\nu \tag{5}$$

with the metric

$$g_{\mu\nu} = \sum_A \frac{\partial X^A}{\partial x^\mu} \frac{\partial X^A}{\partial x^\nu} \tag{6}$$

Here I choose to display the summation over A explicitly. The attentive reader wlll realize that (1) and (2) are examples of these two equations. The metric $g_{\mu\nu}$ is said to be induced by the ambient Euclidean metric.

Another common embedding method is to restrict X^A to satisfy certain conditions. Again, we have the familiar example of the unit sphere defined by $X^2 + Y^2 + Z^2 = 1$. Indeed, writing (X, Y, Z) in terms of (θ, φ) amounts to solving this equation. Instead, we could choose to eliminate* Z. In the notation used above, $(x^1, x^2) = (X, Y)$. Then, $ds^2 = dX^2 + dY^2 + dZ^2 = dX^2 + dY^2 + \frac{(XdX + YdY)^2}{1 - X^2 - Y^2}$. We can clearly exploit the rotational invariance in the $(X$-$Y)$ plane and write[4] $X = r\cos\varphi$, $Y = r\sin\varphi$, $X^2 + Y^2 = r^2$, and $XdX + YdY = rdr$, thus obtaining

$$ds^2 = dr^2 + r^2 d\varphi^2 + \frac{r^2 dr^2}{1 - r^2} = \frac{dr^2}{1 - r^2} + r^2 d\varphi^2 \tag{7}$$

Since the metric in (7) and the usual $ds^2 = d\theta^2 + \sin^2\theta d\varphi^2$ both describe the sphere, they must be related by a coordinate transformation. Do you see how? Another question for you: Why does (7) become singular as $r \to 1$? See appendix 1. Further examples of embedding are given in appendix 2.

Locally flat

The fearful student looks much more relaxed. He asks, "All this is clear enough for actual 2-dimensional surfaces I can visualize, but is it obvious that we can always choose our neighborhood to be locally flat for any space of any dimension D?"

---

* The two solutions of the quadratic equation for Z define two coordinate patches covering the northern and southern hemispheres (minus the equator, strictly speaking), as mentioned in the preceding chapter.

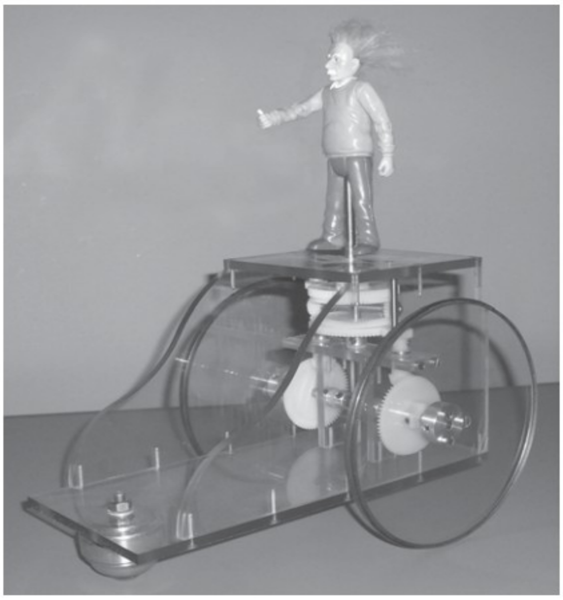

Figure 5 A modern version of the south-pointing carriage described in ancient Chinese chronicles.

Gauss's strategy

To determine the curvature of the surface at a given point P, Gauss's strategy was to study curves on the surface passing through P and calculate their curvatures. This sounds puzzling at first. How did Gauss expect to determine[6] the curvature of a surface by studying the curvature of curves lying on the surface?

To motivate his reasoning, consider the curvature at the saddle point P with coordinates $(x = 0,\ y = 0)$ on the surface $z = \frac{1}{2}px^2 - \frac{1}{2}qy^2$, taking $p > 0,\ q > 0$ for definiteness (a special case of the surface considered in chapter I.5). Imagine yourself walking along the ridge defined by $y = 0$: on both sides of you, the land falls away (see figure 6). When you reach the lowest point $x = 0$ on the ridge, the land in front of you and behind you rises up. Consider the family of curves going through P. For a curve pointing in the y direction, the curvature is positive, while for a curve pointing in the x direction, the curvature is negative. Evidently, for a curve pointing in some other direction, the curvature is intermediate between two extremal values. Gauss proposed to find these two extremal values.

So for a given point P on a curved space, let's look at some curve going through that point and study its tangent vector $\vec{t} = t^\mu \vec{e}_\mu$ with $t^\mu = \frac{dX^\mu}{dl}$. Then

$$\frac{d\vec{t}}{dl} = \frac{dt^\mu}{dl}\vec{e}_\mu + t^\mu \frac{d\vec{e}_\mu}{dl} \tag{13}$$

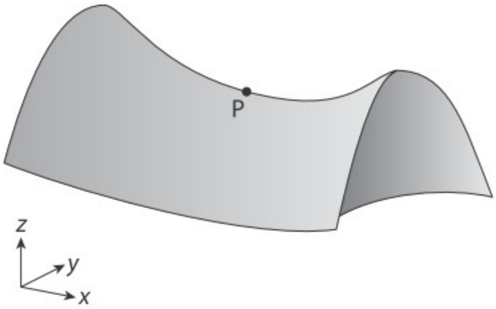

Figure 6 Determining the curvature of a surface with a saddle point P.

Write the right hand side as some linear combination of $\vec{e}_1$, $\vec{e}_2$, and $\vec{n}$, so that

$$\frac{d\vec{t}}{dl} = \kappa_g \vec{e} + \kappa_n \vec{n} \tag{14}$$

with $\vec{e}$ some unit vector given by some linear combination of $\vec{e}_1$ and $\vec{e}_2$. Note in passing that since $\vec{t} \cdot \vec{t} = 1$, we have $\vec{t} \cdot \frac{d\vec{t}}{dl} = 0$ and so $\vec{t} \cdot \vec{e} = 0$. The three unit vectors $\vec{n}$, $\vec{t}$, and $\vec{e}$ form an orthonormal triad. Note also that $\vec{e}_\mu$ and $\vec{n}$ pertain to the surface, while $\vec{t}$ and $\vec{e}$ pertain to the curve.

To get a feel for the two quantities κ_g and κ_n in (14), consider a couple of surfaces. Draw a couple of pictures as you read the next two paragraphs. That will make it easy to follow what I am talking about, which is not much more than common spatial sense.

First, a plane flat as Kansas. Through P draw a curve as curvy as you like, and you can make κ_g as large as you like. But try as you may, $\frac{d\vec{t}}{dl}$ will stay in the plane by definition, and κ_n will remain stubbornly equal to 0. Evidently, κ_g does not tell us about the curvature of the surface, merely the curvature of the curve.

Next, picture a sphere, and let P be Copenhagen, say. Following Gauss, consider a variety of curves going through Copenhagen. For example, think of the circle of constant latitude. Then $\frac{d\vec{t}}{dl}$ points toward the axis of the earth joining the north and south poles and can be written as in (14) with $\vec{e}$ pointing north along some street in Copenhagen and $\vec{n}$ pointing upward* at the sky (as always, independent of where we are). As we try different curves, κ_g and κ_n vary. If we choose the curve to be the circle of constant longitude (rather than latitude) going through Copenhagen, κ_g vanishes, while $|\kappa_n|$ is maximized, given by the inverse of the earth's radius. Indeed, $|\kappa_n|$ attains its maximum value for the two great circles going through Copenhagen. If you have trouble with this, try drawing a picture.

This example shows clearly that it is κ_n, not κ_g, that tells us about the curvature of the surface. You can further convince yourself by picturing other examples, such as the ridge you were walking on earlier. While there is an infinite number of surfaces with different global properties, if we focus on a local patch of the surface around a point P, there are only a couple prototypical surfaces, sphere-like or ridge-like, so to speak.

---

* With this convention (rather than pointing downward toward the center of the earth, as would be mathematically more natural), $\kappa_g > 0$ while $\kappa_n < 0$.

From curvature of curves to curvature of surface

Using (13) and (7), we have (suppressing terms not of interest to us)

$$\frac{d\vec{t}}{dl} = \cdots + t^\mu \frac{d\vec{e}_\mu}{dl} = \cdots + t^\mu \frac{\partial \vec{e}_\mu}{\partial x^\nu} t^\nu = \cdots + t^\mu (K_{\mu\nu}\vec{n}) t^\nu \tag{15}$$

Comparing with (14), we can extract

$$\kappa_n = K_{\mu\nu} t^\mu t^\nu \tag{16}$$

A digressive word about what appears to physicists as rather quaint terminology. Strip the dl off $t^\mu = \frac{dx^\mu}{dl}$ in κ_n and we encounter the combination $K_{\mu\nu} dx^\mu dx^\nu$, known to classical differential geometers as the second fundamental form.[7] Sounds pretty important. Then what do these guys call the first fundamental form? None other than our beloved infinitesimal distance squared: $g_{\mu\nu} dx^\mu dx^\nu$.

Next, differentiate $\vec{t} \cdot \vec{n} = 0$ to obtain (using (14))

$$\vec{t} \cdot \frac{d\vec{n}}{dl} = -\vec{n} \cdot \frac{d\vec{t}}{dl} = -K_{\mu\nu} t^\mu t^\nu = -\kappa_n \tag{17}$$

known as Weingarten's equation after Julius Weingarten (1836–1910).

Gauss's idea was to look at all curves passing through the point P. For each curve, calculate κ_n. Following the great man, we are supposed to find the extremal values of κ_n. So, let us extremize $K_{\mu\nu} t^\mu t^\nu$. But t^μ is not entirely free to vary: it has to satisfy $g_{\mu\nu} t^\mu t^\nu = 1$. Thus, we have a constrained extremization problem.

But we know how to deal with this: introduce a Lagrange multiplier[8] k. In other words, vary $K_{\mu\nu} t^\mu t^\nu - k(g_{\mu\nu} t^\mu t^\nu - 1)$ with respect to t^μ and set the result to 0, thus obtaining $(K_{\mu\nu} - k g_{\mu\nu}) t^\nu = 0$. Multiplying by $g^{\lambda\mu}$, we obtain an equation

$$(g^{\lambda\mu} K_{\mu\nu} - k\delta^\lambda_\nu) t^\nu = 0 \tag{18}$$

for the eigenvalues of the 2-by-2 matrix $M^\lambda_{\ \mu} \equiv g^{\lambda\mu} K_{\mu\nu}$. Contracting (18) with $t_\lambda \equiv g_{\lambda\rho} t^\rho$, we find $k = K_{\mu\nu} t^\mu t^\nu = \kappa_n$. In other words, the two eigenvalues k_1 and k_2 give the extremal values of κ_n.

As usual, the eigenvalues are determined by

$$\det(g^{-1}K - kI) = 0 \tag{19}$$

which (since $g^{-1}K$ is a 2-by-2 matrix) amounts to the quadratic equation $k^2 - \mathrm{tr}(g^{-1}K)k + \det(g^{-1}K) = 0$. Thus, the product of the eigenvalues is given by

$$\mathcal{G} \equiv \det(g^{-1}K) = \frac{\det K}{\det g} \tag{20}$$

and the sum by

$$\mathcal{S} \equiv \mathrm{tr}(g^{-1}K) \tag{21}$$

($ B ' = : D ! A " 1!

Alas for Nordström, he chose the wrong road.* Nature does not† make use of this possibility. Incidentally, he did this work[5] before Einstein formulated his theory of general relativity.

After special relativity was established, which after all was to make mechanics compatible with electromagnetism and its Lorentz invariance, Einstein was not the only one to realize that Newtonian gravity (1) also had to be made compatible with Lorentz invariance. Others in the race included Max Abraham (1875–1922) and Gustav Mie (1869–1957).

A less traveled road to Einstein gravity

> Compared with understanding gravity, the special theory of relativity was mere child's play.
>
> —A. Einstein writing to Arnold Sommerfeld, 1912

While ρ becomes a component of the tensor $T^{\mu\nu}$, the left hand side of $\partial^2\Phi = 4\pi G\rho$, for the equation to make sense, is also compelled to be a component of a tensor. We are thus forced to promote $\Phi(x)$ to a symmetric tensor field $h^{\mu\nu}(x)$ and write

$$\partial^2 h_{\mu\nu} = 8\pi G T_{\mu\nu} \tag{3}$$

where we have defined $h_{00} = -2\Phi$ to agree with our earlier discussion.

After the preceding chapter, it does not take much to guess that the field $h_{\mu\nu}$ will rather naturally turn out to be the deviation of the metric $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$ from the Minkowski metric. Thus, once we decide not to wander off with Nordström, we are practically committed to curved spacetime. Indeed, as soon as we write down (3), we are led inexorably to Einstein gravity, as was shown by Stanley Deser, collaborating with David Boulware and echoing various earlier and later works of Suraj Gupta, Robert Kraichnan,[6] Richard Feynman,[7] Steve Weinberg,[8] and others.

I merely sketch how we will end up with Einstein gravity, suppressing indices for clarity. The action that would lead to (3) has the form $S_1 \sim \int d^4x(\frac{1}{G}\partial h\partial h + hT)$, with the (invisible) indices contracted by $\eta_{\mu\nu}$. Indeed, $\delta S_1 \sim \int d^4x(\frac{1}{G}\partial h\partial\delta h + \delta hT) \sim \int d^4x(-\frac{1}{G}\partial^2 h + T)\delta h = 0$, giving us $\frac{1}{G}\partial^2 h \sim T$. I must confess that living the unindexed life has its charms.

Keep on iterating

But this action can't be the end of the story. The very fact that the field $h_{\mu\nu}(x)$ is endowed with dynamics—that it can wriggle in spacetime—means that it carries energy and mo-

---

* Had Nature chosen this road, you wouldn't have to learn Riemannian geometry to master gravity.

† This is one of my favorite examples of the need to read with sophistication Einstein's dictum about making physics as simple as possible. Simple does not necessarily mean less math. Nature couldn't care less about how much or how little math you learned in school.

mentum. In S_1, we include in $T^{\mu\nu}$ only the contribution of the matter fields. Here the word "matter" is used in the same sense as in chapter VI.4 and includes anything but the gravitational field $h_{\mu\nu}$. We are now forced to include the contribution of $h_{\mu\nu}$ to $T^{\mu\nu}$ as well.

In chapter VI.4, we learned, given an action, how to determine its contribution to $T^{\mu\nu}$, even if the action is given merely in flat spacetime. We pretend momentarily that $\eta_{\mu\nu}$ is actually $g_{\mu\nu}$, we vary $g_{\mu\nu}$ to obtain $T^{\mu\nu}$, and then quickly set $g_{\mu\nu}$ back to $\eta_{\mu\nu}$, doing all this in our heads.

Apply this procedure to the term $\frac{1}{G}\partial h \partial h$ in S_1. Since this term carries 6 indices (which we have suppressed), there are actually 3 "invisible" ηs lurking in this term, in the schematic form $\sim \frac{1}{G}\eta\eta\eta\partial h\partial h$. Promoting $\eta_{\mu\nu}$ to $g_{\mu\nu}$, varying $g_{\mu\nu}$, and then setting $g_{\mu\nu}$ back to $\eta_{\mu\nu}$, we obtain a contribution to $T^{\mu\nu}$ of the schematic form $\frac{1}{G}\eta\eta\partial h\partial h$, where two indices are left dangling to match the $\mu\nu$ indices on $T^{\mu\nu}$. In other words, we have to shift $hT \to h(T + \frac{1}{G}\eta\eta\partial h\partial h)$. Including this contribution, we are forced to the action $S_2 \sim \int d^4x(\frac{1}{G}\partial h\partial h + hT + \frac{1}{G}h\partial h\partial h)$. (Our notation is evidently such that after every step, the letter T in the schematic form of the action once again includes only the contribution of the matter fields. We also suppress the many ηs lurking in the action.)

You see how the game goes: we keep on iterating. The term $\frac{1}{G}h\partial h\partial h$ in S_2 will now force us to include a term of the form $\frac{1}{G}h\partial h\partial h$ in T. Hence we are led by the nose to the action $S_3 \sim \int d^4x(\frac{1}{G}\partial h\partial h + hT + \frac{1}{G}h\partial h\partial h + \frac{1}{G}h^2\partial h\partial h)$.

Before you can shout "Here comes S_4!", you see that the action will iterate to an infinite series $S \sim \int d^4x\{\frac{1}{G}(\partial h\partial h + h\partial h\partial h + h^2\partial h\partial h + \cdots) + hT\}$. This much is easy to understand and makes sense physically, as I will explain presently. The hard part is to show, as Deser and company did, that the series $\frac{1}{G}(\partial h\partial h + h\partial h\partial h + h^2\partial h\partial h + \cdots)$ sums to the Lagrangian $\frac{1}{G}\sqrt{-g}R$ in Einstein gravity.*

In other words, the claim is that, upon substituting $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$ into the Einstein-Hilbert Lagrangian and expanding in h, we will obtain the series we found by iterating. But in a sense this is hardly surprising, since general covariance forces us to the Einstein-Hilbert Lagrangian uniquely (as we saw in chapter VI.1).

In chapter V.2, we extolled the far-reaching power of the equivalence principle. If we know the action governing any interaction in flat Minkowskian spacetime, we immediately know the action governing that interaction in curved spacetime, in the presence of gravity: all we have to do, we learned, is to promote $\eta_{\mu\nu}$ to $g_{\mu\nu}$. But there is one interaction we cannot apply this wondrous stupendous trick to, so take back the word "any" in the preceding sentence. That very special interaction is the gravitational interaction itself! We knew gravity only in Newtonian, not Minkowskian, spacetime. What we have learned in this chapter is that if we try to construct the gravitational action in Minkowskian spacetime iteratively, we end up with the Einstein-Hilbert action in curved spacetime.

Instead of working with the action, we could also have worked with the equation of motion (3). Our task would then be to find a combination, involving two derivatives and the metric, with which to replace $\partial^2 h_{\mu\nu}$ on the left hand side, as we had already anticipated way way back in part II. Gauss, Riemann, and Ricci solved this highly nontrivial problem for us.

---

* To keep things simple, we did not mention that starting with hT, we generate an infinite series that will also sum up appropriately.

Geometry emerges

This discussion underlines the importance of the geometrical view of special relativity that we have emphasized in part III, in contrast to the view of special relativity as a series of apparent paradoxes. We could certainly work through the theory of special relativity applying the Lorentz transformation to various apparent paradoxes, without once mentioning the word "geometry." But that would be impoverishing physics.

Imagine yourself in a galaxy far far away, where people have never heard of Einstein gravity. But people know about the Newtonian equation (1) $\nabla^2\Phi(\vec{x}, t) = 4\pi G\rho(\vec{x}, t)$, and Lorentz invariance has just been discovered. Suppose you then try to make this equation Lorentz invariant, following the road less traveled outlined here, and thus promoting Φ to $h_{\mu\nu}$. If you had understood special relativity in terms of the geometry of spacetime (that is, understood $\eta_{\mu\nu}dx^\mu dx^\nu$ as the generalized distance between two nearby points), then you would naturally interpret $(\eta_{\mu\nu} + h_{\mu\nu})dx^\mu dx^\nu$ as an even more generalized distance between two nearby points. Geometry of curved spacetime naturally emerges. But the poor sap who knows special relativity as a series of paradoxes may have a hard time seeing curved spacetime.

The graviton interacts with itself

The physics behind the infinite series $\frac{1}{G}(\partial h\partial h + h\partial h\partial h + h^2\partial h\partial h + \cdots)$ is easy to understand with a minimal knowledge[9] of quantum field theory. You have no doubt heard, and I have already mentioned, that when we quantize the electromagnetic field, we obtain the photon, and when we quantize the gravitational field, we obtain the graviton. A huge difference between the photon and the graviton is that the photon couples to charged fields, such as the electron field, but is not charged itself. In contrast, the graviton couples to anything carrying energy and momentum, and since it certainly carries energy and momentum, it couples to itself. In the electromagnetic action given in chapter IV.2, the photon couples to the charged fields that comprise the electromagnetic current J^μ via the term $A_\mu J^\mu$ in the action. Similarly, in the Einstein-Hilbert action, the graviton couples to matter fields that comprise the energy momentum tensor $T^{\mu\nu}$ via the term $h_{\mu\nu}T^{\mu\nu}$. But in addition, there are an infinite[10] number of terms of the form $h \cdots h\partial h\partial h$ describing the complicated interactions of many gravitons with one another. This partly accounts for the intractability of quantum gravity at present.

To say all this in a slightly different way, we recall from chapter VI.4 that $T^{\mu\nu}$ is not locally conserved, in contrast to the electromagnetic current J^μ. The physics behind this fact is the ability of the gravitational field to exchange energy momentum with $T^{\mu\nu}$.

Appendix 1: From electrostatics to Maxwell

It is instructive to repeat for electromagnetism what we did in this chapter. Suppose we start with electrostatics, with Poisson's equation $\nabla^2\Phi(\vec{x}, t) = \rho(\vec{x}, t)$ determining the electrostatic potential Φ in terms of the charge density ρ. Indeed, it is essentially the same equation as (1) that we started this chapter with.

($ ' 211 " L %

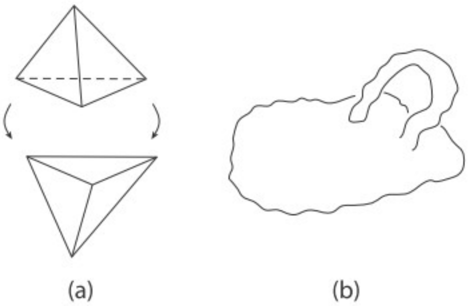

Figure 1 (a) Gluing 2 tetrahedra together. (b) A spherical blob grows a trunk.

What about the Einstein-Hilbert action? Let's try $S = \int R_{\alpha\beta} * (e^\alpha e^\beta)$. Again, the answer is easy to guess. We need a d-form to integrate over, involving 1 power of the curvature 2-form and no power of derivative. To obtain a d-form, we multiply the curvature 2-form by a $(d-2)$-form: the only possibility in this context is the star of the 2-form $e^\alpha e^\beta$. Again, between us friends, do you doubt this? Okay, let's check (but, to save writing and to spare you the stream of indices, only for 4-dimensional spacetime):

$$\begin{aligned} R_{\alpha\beta} * \left(e^\alpha e^\beta\right) &= \left(R_{\alpha\beta\omega\psi} dx^\omega dx^\psi\right)\left(\epsilon_{\rho\sigma}^{\ \ \mu\nu}\sqrt{-g}\, e^\alpha_\mu e^\beta_\nu dx^\rho dx^\sigma\right) \\ &= R_{\mu\nu\omega\psi}\epsilon_{\rho\sigma}^{\ \ \mu\nu}\sqrt{-g}\, dx^\omega dx^\psi dx^\rho dx^\sigma = R_{\mu\nu\omega\psi}\epsilon_{\rho\sigma}^{\ \ \mu\nu}\sqrt{-g}\,\epsilon^{\omega\psi\rho\sigma} d^4x \\ &= R_{\mu\nu\omega\psi} g^{\mu\omega} g^{\nu\psi}\sqrt{-g}\, d^4x = \sqrt{-g}\, d^4x\, R \end{aligned} \tag{16}$$

(using the generalization of (6) and again ignoring overall numerical factors), as expected.

Appendix 3: Topological invariants: Euler characteristic, Gauss-Bonnet theorem, and all that

We are now ready to look at some celebrated invariants in topology. Our discussion will be heuristic rather than rigorous, hitting some highlights rather than being exhaustive. I keep the discussion as elementary as possible.

I suspect that many readers probably first encountered the Euler characteristic, like me, in a popular book of mathematics. For me, it was a real eye opener. First, let us look at the empirical data. The cube has 8 vertices or corners, 12 edges (4 on the top, 4 on the bottom, and 4 on the side), and 6 faces. Hence, $V = 8$, $E = 12$, $F = 6$. Next, the tetrahedron has 4 vertices, 6 edges (3 on the bottom and 3 on the side), and 4 faces, that is, $V = 4$, $E = 6$, $F = 4$. We see that the combination, $\chi = V - E + F$, known as the Euler characteristic, is equal to 2 in both cases. (Note that in the sum, we add the number of geometrical entities, vertices, edges, and faces, with an alternating sign according to whether the dimension of the entity is even or odd.) The joke is that theoretical physicists would proclaim this to be a theorem at this point, but in fact it is easy to prove. Let me give a proof that would satisfy most physicists (but not mathematicians) and is suitable for elementary school children.

Glue another tetrahedron to the tetrahedron we have, producing a 6 faced "diamond-shaped" object. See figure 1a. We gain 4 faces (the four of the second tetrahedron) but lose 2 faces (the two that are glued together). Thus, $\Delta F = +4 - 2 = 2$. Similarly, we can see that $\Delta V = +4 - 3 = 1$ and $\Delta E = +6 - 3 = 3$. Hence, we have $\Delta\chi = \Delta V - \Delta E + \Delta F = 1 - 3 + 2 = 0$. Now we can glue zillions of these tetrahedra together to approximate any object we like. At every gluing, the Euler characteristic does not change. Hence $\chi = 2$ for any spherical-looking blob. The discussion makes clear that we are counting the V, E, and F on the surface. (The interior of the object, consisting of the faces we have glued together and their edges and vertices, has been "lost forever," so to speak.) Throughout, we will be studying a surface or a 2-manifold, not the 3-dimensional blob enclosed by the surface.

The Euler characteristic is evidently a topological quantity. We can lengthen and shorten the edges of the tetrahedra we are gluing together without changing V, E, and F.

Suppose we now go beyond deforming the spherical blob to changing its topology. We show presently that χ "measures" the topology of the object.

By gluing more tetradehedra on, we can make the spherical blob grow a trunk like that of an elephant. Let us slowly extend the trunk back toward another part of the blob, sort of like an elephant about to scratch its back with its trunk. The Euler characteristic χ remains equal to 2, until we glue the tetrahedron at the tip of the trunk to a tetrahedron on the back of the elephant. See figure 1b. Now we lose 2 faces, 3 vertices, and 3 edges, and thus $\Delta\chi = \Delta V - \Delta E + \Delta F = -3 + 3 - 2 = -2$. The resulting surface has the topology of a torus and Euler characteristic $\chi = 2 - 2 = 0$.

Growing another trunk and attaching it somewhere else causes the Euler characteristic χ to decrease further, with $\Delta\chi = -2$ every time we do it. We have thus derived the general result $\chi = V - E + F = 2(1 - g)$, where g denotes the genus ($g = 0$ for the sphere, $g = 1$ for the torus, and so on). Some people call the genus the number of "holes." (The torus is said to have 1 hole, but as we will soon see, in this context, the word "genus" or "handle" is preferable to the word "hole.") The Euler characteristic χ is manifestly a topological invariant, independent of the size or "shape" of the surface, and only dependent on its genus.

A trivial generalization is to punctured surfaces. (In everyday parlance, a punctured sphere is a sphere with a hole in its surface, like a rapidly shrinking balloon with a hole in it. This is one reason why the use of the word "hole" for genus or handle, as indulged in by some, is ill advised.) Since we can puncture a surface by removing a triangular face from the surface (so that $\Delta V = 0$, $\Delta E = 0$, and $\Delta F = -1$ and hence $\Delta\chi = -1$), we have, more generally,

$$\chi = 2 - 2g - h \tag{17}$$

with h the number of holes or punctures in the surface.[7]

Another way of proving (17) is to start with $\chi = 2$ for a spherical surface, and to punch any number of holes in it, thus obtaining $\chi = 2 - h$. Deforming the surface to bring 2 holes near each other and then to glue them together, we decrease h by 2 and increase the genus by 1. Hence we have (17). This derivation also makes clear the relative factor of 2 in the coefficients of g and h in χ.

We obtain what is known as a triangulation of the surface. (Indeed, that is what surveyors do: they triangulate the surface of the earth.) At the level of rigor of physics, any surface can be approximated to arbitrary accuracy by making the triangles small enough. In other words, we physicists would take the continuum limit without further ado.

At the same level of rigor, we can also approximate spacetime by a large number of discrete elements. This represents the first step in a program to discretize Einstein gravity and to put it on the computer for numerical analysis.*

That we used tetrahedra is not essential. Take the tetrahedron we started with. Call the vertices on the triangle on its "base" A, B, and C. Pick a point X on the edge joining A and B, and draw a line connecting X to the other vertex C. Then $\Delta V = 1$, $\Delta E = 2$, and $\Delta F = 1$, and so $\Delta\chi = 0$. By "pulling" on the point X, we can deform the tetrahedron, if we feel like it, to a pyramid with a square base. As another example, we can pick a triangle on the surface we are studying and draw a line from one side of the triangle to another side, so that we divide the triangle into a smaller triangle and a quadrilateral. In the process, $\Delta V = 2$, $\Delta E = 3$, and $\Delta F = 1$, and so again $\Delta\chi = 0$. You can make up your own "moves" and show that instead of triangles, the surface could be composed of polygons with any number of sides you like.

Indeed, Descartes had already published, in his progymnasmata to the study of solids, a theorem on angular deficits that foreshadowed the Euler characteristic. Here is what Descartes said. At each vertex of a cube, 3 squares meet. The 3 angles at the vertex add up to $3(\frac{1}{2}\pi) = \frac{3}{2}\pi$. The amount by which this is less than 2π is known as the angular deficit, in this case equal to $(2 - \frac{3}{2})\pi = \frac{1}{2}\pi$. The total angular deficit, namely the sum of the angular deficits at each of the 8 vertices, is then equal to $8(\frac{1}{2}\pi) = 4\pi$. Descartes stated that the total angular deficit of any polyhedron topologically equivalent to the sphere is equal to 4π.

Let's try it for a tetrahedron. At each vertex, 3 equilateral triangles meet, with angles adding up to $3(\frac{\pi}{3}) = \pi$, so that the deficit equals $(2 - 1)\pi = \pi$. There are 4 vertices, and so the total deficit is indeed 4π. We have verified 2 cases, so it is surely a theorem. We are proud physicists, but still, perhaps a third example would be good.

So consider the dodecahedron with 12 pentagonal faces. Since there are $E = (12 \times 5)/2 = 30$ edges, according to Euler's theorem, the number of vertices equals $V = 2 + E - F = 2 + 30 - 12 = 20$. The 12 pentagons have $12 \times 5 = 60$ vertices, and hence $60/20 = 3$ pentagons meet at each vertex. For a regular polygon with n sides, the angle α at each vertex is given by $n\alpha + 2\pi = n\pi$ (to see this, divide the polygon into n triangles), that is, $\alpha = (n-2)\pi/n$. In particular, for a pentagon, we have $\alpha = \frac{3}{5}\pi$. The angular deficit at each vertex is then $\left(2 - 3\left(\frac{3}{5}\right)\right)\pi = \frac{1}{5}\pi$. Thus, the total angular deficit equals to $20 \times \frac{1}{5}\pi = 4\pi$. Descartes was right!

---

* Lattice gravity is a thriving subject of research. Historically, this first step is known as the Regge calculus.

Let's now show that the total angular deficit is topologically invariant. Consider some triangulated surface as described earlier, and visualize it as a wire framework. (For ease of presentation, let the faces be triangles. Thus, for the cube mentioned above, simply divide the squares by their diagonals.) Pick an edge AB, shared by two triangles ABC_1 and ABC_2. Call the 6 angles contained in these 2 triangles α_j, β_j, γ_j, with $j = 1, 2$, using an almost self-evident notation (for example, α_1 is the angle of ABC_1 at the vertex A, γ_1 is the angle of ABC_1 at the vertex C_1, and so on). Now imagine lengthening the edge AB slightly, thus increasing γ_j and decreasing α_j and β_j. Thus, the change in the angular deficit at the vertex A is $-(\delta\alpha_1 + \delta\alpha_2)$. But since the angles in a triangle have to add up to π, $\delta(\alpha_j + \beta_j + \gamma_j) = 0$. The angular deficit at the vertices A, B, C_1, and C_2 all vary, but the total angular deficit stays the same: it is a topological invariant, as Descartes taught us.

Here is the previous proof dressed up to make it look more sophisticated. Label the vertices by p. At vertex p, a set $s(p)$ of triangles meet. The sum of the angles meeting at vertex p is then $\sum_{i \in s(p)} \alpha_i$, with α_i the angle extended by the ith triangle at that vertex. The total angular deficit is then $\sum_p \left(2\pi - \sum_{i \in s(p)} \alpha_i\right)$. Let us now deform the surface infinitesimally by lengthening or shortening each of the zillions of edges. The variation in the total angular deficit is equal to $-\sum_p \sum_{i \in s(p)} \delta\alpha_i$. Rearranging to sum over 1 triangle at a time, we see that this vanishes, since the sum of the angles in each triangle is constrained to add up to π.

With this background explanation of what the Euler characteristic χ is, we now return to the subject of this appendix and of this chapter. Can we write χ as an integral? In other words, how do we calculate χ, originally defined to be $V - E + F$, in the continuum limit where V, E, and F are not defined?

We are given a closed surface, that is, a 2-manifold M without boundary (in other words, we are setting $h = 0$ for simplicity). What is a 2-form that we could integrate over M? The curvature 2-form $R^{\alpha\beta}$ comes to mind; so let's try $\int_M \varepsilon_{\alpha\beta} R^{\alpha\beta}$. (It is also instructive to try the other possibility: $e^\alpha e^\beta$. It turns out that $\int_M \varepsilon_{\alpha\beta} e^\alpha e^\beta$ is the area of the surface, as you might have guessed from the fact that it does not contain any derivative.) We first work out the integrand in terms of a more elementary notation:

$$\begin{aligned} \varepsilon_{\alpha\beta} R^{\alpha\beta} &= \varepsilon_{\alpha\beta} R^{\alpha\beta}_{\mu\nu} dx^\mu dx^\nu = \varepsilon_{\alpha\beta} e^\alpha_\rho e^\beta_\sigma R^{\rho\sigma}_{\mu\nu} dx^\mu dx^\nu \\ &= (\det e)\varepsilon_{\rho\sigma} R^{\rho\sigma}_{\mu\nu} \varepsilon^{\mu\nu} d^2x = 2d^2x\sqrt{g}\,\delta^\mu_\rho \delta^\nu_\sigma R^{\rho\sigma}_{\mu\nu} = d^2x\sqrt{g}R \end{aligned} \tag{18}$$

(In the next to last step, we used (6).)

We have known, for quite a while now, that for a sphere of radius a, the scalar curvature $R = 2/a^2$. The area is $\int_M d^2x\sqrt{g} = 4\pi a^2$. Thus, the radius cancels out in the integral $\int_M \varepsilon_{\alpha\beta} R^{\alpha\beta} = 2\int_M d^2x\sqrt{g}R$, and this integral is indeed, as we might suspect, topological in character, equal to some constant like 16π.

The scalar curvature of a torus was calculated back in exercise I.6.2 and again in exercise I.5.16; it assumes both positive and negative values. I will let you check that the integral gives 0 for a torus. We claim that, up to some overall factor, this integral gives the Euler characteristic χ.

In fact, Descartes' theorem is just the statement, up to some overall constant, that the Euler characteristic χ equals 2 for a surface with the topology of the sphere. Go back to the spherical-looking blob. Under the microscope, we see that the surface is formed out of zillions of triangles. Inside any triangle, we have a flat surface; the surface curvature is concentrated on the edges and at the vertices. Indeed, the angular deficit measures our intuitive understanding of curvature: that which we cannot iron flat is curvature. The smaller the angular deficit is at a vertex, the less that vertex sticks out. When the angular deficit vanishes, the surface around that vertex is flat. The angular deficit "measures" the curvature.

The generalization to a 4-manifold M almost immediately suggests itself. Let's parallel the discussion embodied in (18) and integrate the 4-form $\varepsilon_{\alpha\beta\gamma\delta} R^{\alpha\beta} R^{\gamma\delta}$ over M. As in the discussion above, the characteristic length a of M will cancel out in the integral $\int_M \varepsilon_{\alpha\beta\gamma\delta} R^{\alpha\beta} R^{\gamma\delta}$. I will let you have the fun of working this out in an exercise. That this integral is a topological invariant[8] is known as the Gauss-Bonnet theorem in the physics literature.

In the interest of keeping this appendix to a manageable size, I have not proven why the 2 integrals mentioned here are topological invariant,[9] but instead, have "merely" shown you why they must be so (in the spirit of what the American Mathematical Society said about my field theory textbook, as quoted in a footnote in chapter V.6).

Exercises

1 Check Descartes' theorem for the icosahedron, constructed out of 20 equilateral triangular faces. Hint: Use Euler's theorem.

(/B / B A " 1! 7! ! " 5 !

18 Change coordinates by $u = \log r,\ v = \theta$, and we obtain $ds^2 = (dr^2 + r^2 d\theta^2)$. The space is just the plane in disguise.

I.6 Curved Spaces: Gauss and Riemann

4 With no loss of generality, we can pick the point P to be $(x, y) = (0, y_*)$. A particular set of locally flat coordinates (u, v) is given by $x = y_*(u + uv + \cdots)$, $y = y_*(v + \frac{1}{2}(v^2 - u^2) + \cdots)$, where the dots represent terms cubic and higher in u, v. We could of course trivially rotate (u, v) (and also translate) to obtain another set of locally flat coordinates.

5 From $g_{11} = (1 + \mu^2 u^2)$, we have $B_{11,22} = 0$. Similarly, $B_{22,11} = 0$. From $B_{12,12} = \frac{1}{2}\mu\nu$. So $2B_{12,12} - B_{11,22} - B_{22,11} = \mu\nu$, which is indeed the intrinsic curvature.

8 We have $ds^2 = dx^2 + dz^2/\left(pz^{\frac{p-1}{p}}\right)^2$. For example, for $z = y^2$, $g_{zz} = 1/(4z)$ blows up at $z = 0$.

10 Let one line segment go from the point x to $x + (\Delta x)_1$, and the other to $x + (\Delta x)_2$. The angle between the two line segments is given by

$$\cos\theta = (\Delta x)_1 \cdot (\Delta x)_2 / \sqrt{((\Delta x)_1 \cdot (\Delta x)_1)((\Delta x)_2 \cdot (\Delta x)_2)}$$

(defining $(\Delta x)_1 \cdot (\Delta x)_2 \equiv g_{\mu\nu}(\Delta x)_1^\mu (\Delta x)_2^\nu$ and generalizing the standard high school formula for the scalar dot product). Suppose we now calculate the angle with the metric $\tilde{g}$: the factors of Ω evidently cancel between the numerator and the denominator.

11 Solution by counting. We can choose D functions, but we have to satisfy $\frac{1}{2}D(D+1) - 1$ conditions. We have enough freedom for $D = 2$, but not for $D > 2$.

13 $R = \frac{x+y}{2x^2y^2}$

I.7 Differential Geometry Made Easy, But Not Any Easier!

2 We calculate the components of $\vec{V}$ along the two basis vectors, namely $\vec{V} \cdot \vec{e}_\mu(y)$, and then use these two components to form a linear combination of the two basis vectors. These words translate into an expression for $\vec{V}$ projected into the tangent plane: $\vec{V}_{\rm P}(y) \equiv (\vec{V} \cdot \vec{e}_\mu(y)) g^{\mu\nu}(y)\vec{e}_\nu(y)$, where $g^{\mu\nu}(y)$ is the inverse of the 2-by-2 matrix $g_{\mu\nu}(y)$ defined by $g^{\mu\nu}(y)g_{\nu\lambda}(y) = \delta^\mu_\lambda$. (For the sphere, $g_{\mu\nu} = \begin{pmatrix} 1 & 0 \\ 0 & \sin^2\theta \end{pmatrix}$ and $g^{\mu\nu} = \begin{pmatrix} 1 & 0 \\ 0 & 1/\sin^2\theta \end{pmatrix}$.) To see that the inverse is needed here, take the dot product of $\vec{V}_{\rm P}(y)$ and a basic vector $\vec{e}_\lambda(y)$:

$$\vec{V}_{\rm P}(y) \cdot \vec{e}_\lambda(y) = (\vec{V} \cdot \vec{e}_\mu(y)) g^{\mu\nu}(y) \vec{e}_\nu(y) \cdot \vec{e}_\lambda(y) = (\vec{V} \cdot \vec{e}_\mu(y)) g^{\mu\nu}(y) g_{\nu\lambda}(y) = (\vec{V} \cdot \vec{e}_\mu(y))\delta^\mu_\lambda = (\vec{V} \cdot \vec{e}_\lambda(y))$$

In other words, $(\vec{V}_{\rm P}(y) - \vec{V}) \cdot \vec{e}_\lambda(y) = 0$, which is just another way of saying that $\vec{V}_{\rm P}(y)$ and $\vec{V}$ differ by a vector normal to the surface. In other words, we subtracted out the component of $\vec{V}$ normal to the surface from $\vec{V}$ to obtain $\vec{V}_{\rm P}(y)$.

3 Multiply the two eigenvalue equations (for $i = 1, 2$)

$$(K_{\mu\nu} - k_i g_{\mu\nu}) t_i^\nu = 0$$

by t_j^μ (with $j \neq i$) and subtract one from the other. We obtain $(k_1 - k_2) g_{\mu\nu} t_1^\mu t_2^\nu = 0$. For those readers who have studied quantum mechanics, does this remind you of the proof of wave function orthogonality?

6 Take the dot product: $\vec{e}_\rho \cdot \vec{e}_{\mu,\nu} = g_{\rho\lambda}\Gamma^\lambda_{\mu\nu} = \Gamma_{\rho\cdot\mu\nu}$. Interchange $\rho \leftrightarrow \mu$: $\vec{e}_\mu \cdot \vec{e}_{\rho,\nu} = \Gamma_{\mu\cdot\rho\nu}$. Add to obtain $\partial_\nu(\vec{e}_\rho \cdot \vec{e}_\mu) = \partial_\nu g_{\rho\mu} = \Gamma_{\rho\cdot\mu\nu} + \Gamma_{\mu\cdot\rho\nu}$, which we can solve for Γ using I.4.15.

II.1 The Hanging String and Variational Calculus

2 Once again, we can solve $\nabla^2\Phi = GM\delta^{(3)}(\vec{x})$ by dimensional analysis. It is also easy to verify the solution explicitly. By rotational invariance, Φ can only depend on $r^2 \equiv \sum_{i=1}^D x^i x^i$. Differentiate this to obtain $r\,dr = \sum_{i=1}^D x^i dx^i$, so that $\frac{\partial r}{\partial x^i} = \frac{x^i}{r}$. Then we have

$$\sum_{i=1}^{D} \frac{\partial}{\partial x^i}\frac{\partial}{\partial x^i}\frac{1}{r^p} = \sum_{i=1}^{D}\frac{\partial}{\partial x^i}\left(-\frac{p x^i}{r^{p+2}}\right) = \frac{p(p+2-D)}{r^{p+2}}$$

We see that $\Phi \propto 1/r^p$ with $p = D-2$ solves the equation for $\vec{x}$ away from the origin. The potential goes like $1/r^{D-2}$, and so the force law is an inverse $(D-1)$ law.

3 Varying S with respect to b gives $a' = 0$, and varying it with respect to a gives $(r(1-b))' = 0$. Fitting to the boundary conditions at spatial infinity gives $a = 1$ and $b = 1 - \frac{2M}{r}$.

5 Describe the desired curve by $y(x)$, with y the vertical axis. Released at rest at $y = 0$, the bead attains a speed of $v(y) = \sqrt{2gy}$ after falling a distance of y (with the coordinate y chosen to point downward). We see that the transit time (in suitable units) is given by

$$T = \int \sqrt{\frac{dx^2 + dy^2}{y}} = \int dx \left[\frac{1}{y}\left(1 + \left(\frac{dy}{dx}\right)^2\right)\right]^{\frac{1}{2}}$$

You could proceed from here, derive the Euler-Lagrange equation, solve for $y(x)$, and “rival” Newton more than 300 years later. Note that in spite of my remark in the text, in this context, the notation $y(x)$ seems quite natural.

Here is the instructive part of the problem. We could just as well have chosen y as the variable and solved for $x(y)$. Then

$$T = \int dy \left[\frac{1}{y}\left(1 + \left(\frac{dx}{dy}\right)^2\right)\right]^{\frac{1}{2}}$$

You can verify that, in contrast to the case with the previous choice, the second order differential equation can now be integrated trivially to yield the first order differential equation

$$\left(\frac{dx}{dy}\right)^2 = \frac{y}{y^* - y}$$

with y^* an integration constant. You could solve this easily with the change of variable $y = y^* \sin^2\theta$.

Moral of the story: in solving variational problems, it pays to choose the independent variable wisely.

II.2 The Shortest Distance between Two Points

6 Varying the stated quantity, we obtain $\frac{d}{d\lambda}\left(2g_{\sigma\nu}\frac{dX^\nu}{d\lambda}\right) = (\partial_\sigma g_{\mu\nu})\frac{dX^\mu}{d\lambda}\frac{dX^\nu}{d\lambda}$ almost instantly, but this is just (16).

8 Simply plug $g_{\mu\sigma} = e^{2\Omega}\delta_{\mu\sigma}$ into (24) to obtain $\frac{d}{dl}\left(e^{2\Omega}\delta_{\mu\sigma}V^\mu\right) = \partial_\sigma\Omega$ (since $g_{\mu\nu}V^\mu V^\nu = 1$). Noting that $\frac{d}{dl}e^{2\Omega} = 2V^\nu\partial_\nu\Omega e^{2\Omega}$, we find, after cleaning up a bit,

B (%

Pág. 82(105) 1.6.
Curved Spaces: Gauss and Riemann
Pág. 104(127).
Gauss's strategy
Pág. 581(604)
– Einstein-Hilbert action in curved spacetime
– Gauss, Riemann, and Ricci solved this highly nontrivial problem for us.
Pág. 725(748) Appendix 3:
Topological invariants: Euler characteristic, Gauss-Bonnet theorem, and all that.
Pág. 796(819) – Solution 1.6.
– Pietro.. Gravity, a Geometrical Course - Volume 1 ..(2012).pdf
Pág. 87(100) Chapter 3
Connections and Metrics
3.2.1 Gauss Introduces Intrinsic

B
! -256- - -* ! - B 7 11 < P 7 !A 3J !
7 ! 1"
! -256- - -* ! - B 7 11 < P 7 !A 3J !
7 ! 1"
*D

(!1 % ! "
% 7! & " &

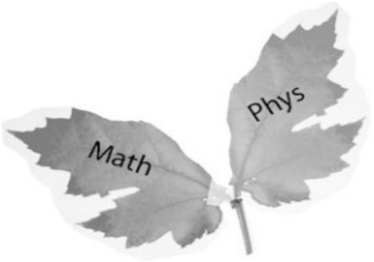

Fig. 3.1 A symbolic assessment by C.N. Yang that Mathematics and Physics have a common root

3.2 A Historical Outline

One spring in the late seventies, C.N. Yang, the father of gauge-theories and one among the most extraordinary scientific minds of the XXth century, was invited by Scuola Normale di Pisa to give the traditional yearly series of Fermian Lectures. It was just the eve of the 1983 experimental discovery of the W and Z bosons, which finally confirmed that gauge theories are indeed the language adopted by Nature to express the fundamental forces binding matter together and driving the evolution of the physical world. Yang chose to start his recollection of gauge-theories and his personal assessment of the entire subject from a symbolic picture of the kind shown in Fig. 3.1.

Two leaves depart from the same stem: one is Mathematics, the other is Physics. The two leaves live parallel lives but have a common root and, as it happens in non-Euclidian geometry, parallels can and indeed do intersect. They frequently intersect and, equally frequently, interchange the role of guiding pivot. Examples are numberless both in recent and less recent history of science. This is a relevant but somehow trivial observation.

The most profound allusion in the typical Chinese symbolism of Yang's picture is the common stem of the two leaves. Mathematics and Physics were not that much socially and academically separated in the XIXth as they became in the XXth century and were further unified by a common denominator: the shared philosophical attitude of both the mathematician and the physicist, the latter assessing himself a natural philosopher.

The common root of Modern Mathematics and Modern Theoretical Physics is most prominently evident in the case of the concepts of *connections* and *metrics* which constitute the main topics of the present chapter. From the physical side these mathematical notions encode all the *fundamental interactions* among *elementary matter constituents*. From the mathematical side they are the corner stones of Differential Geometry, Algebraic Topology and allied subjects.

The historical developments of both the notion of a *connection* and the notion of a *metric* are quite long, spreading over more than a century. They are strongly intertwined and involve both Physics and Mathematics in an alternate and entangled fashion.

Two fundamental geometrical problems are at the basis of both notions: the problem of *length* and the problem of *parallel transport*, namely how can we measure distances in a general continuous space and how do we assess the parallelism of two lines going through different points of that space. It turns out that these two apparently different problems are intimately related. Such a relation is the reason why connections and metrics, although genuinely different and independent mathematical notions can be, under certain conditions, related in a one-to-one fashion. The case where connection and metric are one-to-one related corresponds to Gravity and is the play-ground of General Relativity, while the case where the connection stands on his own feet is the case of non-gravitational interactions and encodes all the rest of Physics.

Historically the first notion of connection that was discovered is that of the *metric connection* named after its systematizer, the Italian mathematician Levi Civita. The *Levi Civita connection* is that appropriate to a Riemannian manifold, namely to a manifold equipped with a metric and it is analytically described by the 3-index symbols $\{{}^{\mu}_{\nu\rho}\}$, introduced in the XIXth century by their German inventor, Elwin Bruno Christoffel. The development of such a notion is embedded in the development of Riemannian geometry, embracing the fall of the XIXth century and the dawn of the XXth, which is a mathematical tale strongly intertwined with the physical tale of Einstein's quest for General Covariance and the Theory of Gravity.

It took several decades in the first half of the XXth century and the work of several mathematicians to single out the notion of a *connection on a principal fibre bundle*, purified from association with a metric. In a completely independent way in 1954 Yang and Mills introduced the physical notion of non-Abelian gauge fields which extends to all Lie groups G the structure of the electromagnetic theory based on the simplest of all such groups, namely $\mathrm{U}(1)$. In the course of due time the mathematical notion of a connection and the physical one of a Yang-Mills field were recognized to be identical.

Let us sketch the main outline of this crucial, century long intellectual development.

3.2.1 Gauss Introduces Intrinsic Geometry and Curvilinear Coordinates

The first appearance of a metric is in the 1828 essay of Gauss (see Fig. 3.2) on curved surfaces. Written in Latin, the *Disquisitiones Generales circa superficies curvas* [1] contains the major revolutionary step forward that was necessary to overcome the precincts of Euclidian geometry and found a new *differential science* of spaces able to treat both flat and curved ones. Up to Gauss' paper, Geometry was either formulated abstractly in terms of Euclidian axioms or analytically in terms of *Cartesian coordinates*. By Geometry it was meant the study of global properties of plane figures like triangles, squares and other polygons, or solids like the regular polyhedra. All such objects were conceived as immersed in an external space where it was implicitly assumed that one could always define the absolute distance $d(A,B)$ between any two given points A and B. Distance is the basic brick of the whole Euclidian building and it is calculated as the length of the segment with end-

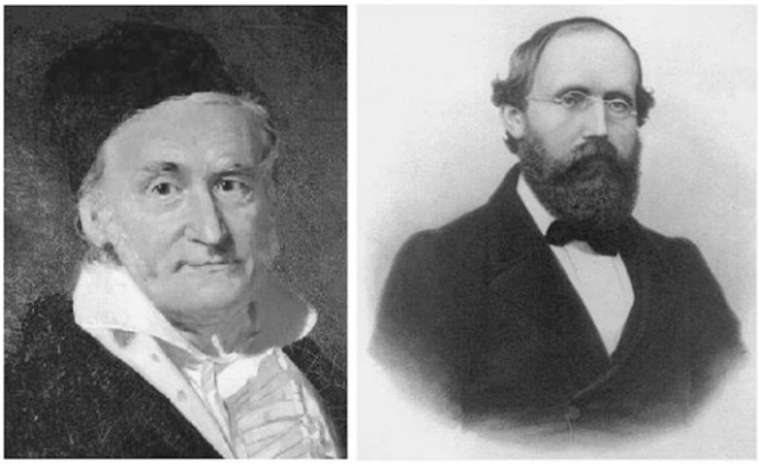

Fig. 3.2 Carl Friedrich Gauss (1777–1855) on the *left* and Georg Friedrich Bernhard Riemann (1826–1866) on the *right*. Gauss, *the King of Mathematicians*, was Professor at the University of Göttingen for many decades up to the very end of his long life. His contributions to all fields of Mathematics were enormous and most profound. Bernhard Riemann, another genious and a giant of human thought, had a very short and not too happy life. He was Gauss' student both at the level of Diploma and of *Habilitationsschrift*. The whole of his work is contained in no more than 11 papers for a total of little more than 200 pages. Yet each of his contributions was a milestone in Mathematics and set the path for century long future developments. The foundations of Riemannian geometry were laid in the 16 page long dissertation of his Habilitation. Riemann had important ties with the Scuola Normale di Pisa and died in Italy at the age of 39

points in A and B, lying on the unique straight line which goes through any such pair of distinct points.

Curved surfaces were obviously known before Gauss, yet their shape and properties were conceived only through their immersion in three-dimensional space, considered unique and absolute, as pretended by Immanuel Kant who promoted Euclidian geometry to an a priori truth lying at the basis of any sensorial experience. Gauss revolutionary starting point was that of reformulating the geometrical study of surfaces from an intrinsic rather than extrinsic viewpoint. He wondered how a little being, confined to live on the surface, might have perceived the geometry of his world. Rather than viewing the global shape of the surface M, unaccessible to his observations, the little creature would have explored its local properties in the vicinity of a point $p \in M$.

In order to study curved surfaces in these terms, Gauss understood that it was necessary to abandon Cartesian coordinates as a system of point identification. In analytic geometry every point $p \in \mathbb{E} \equiv \mathbb{R}^3$ of Euclidian space is singled out by a triplet of real numbers X, Y, Z which determine the distance of p from the three coordinate axes. As long as the points of the surface M are particular points of $\mathbb{E}$, they admit the labeling in terms of three Cartesian coordinates, yet in such a description there is an excess of superfluous information. Why three coordinates when we are talking about a two-dimensional surface? Two should suffice. Gauss

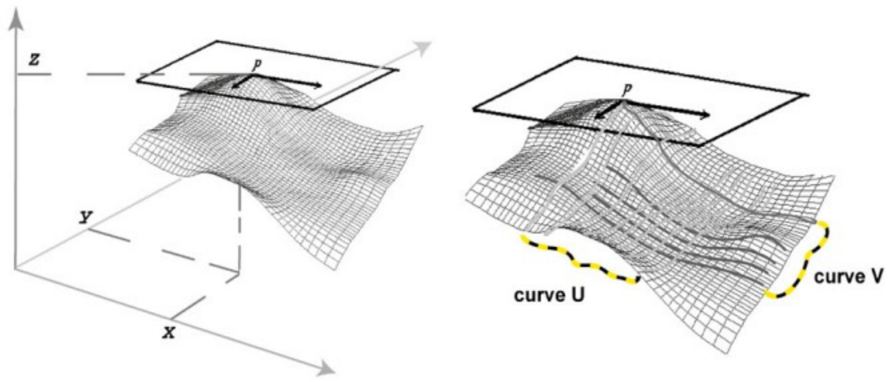

Fig. 3.3 The points p of a curved surface M can be labeled with the three Cartesian coordinates X, Y, Z of the Euclidian space in which M is immersed. Yet this is a redundant information. The two Gaussian coordinates u and v are given through the construction of two systems of curves U and V on the surface

was the first to grasp the notion of curvilinear coordinates and invented *Gaussian coordinates*. A very simple but revolutionary idea.

On the surface M let us consider a family of curves U such that each element of the family never intersects any other element of the same family, at least in the neighborhood $\mathscr{U}_p$ of the point p (for instance the lighter curves in Fig. 3.3) and such that the family covers the entire considered neighborhood. Let us next introduce a second family of curves V, with the same properties among themselves, yet such that each element of the family V intersects all elements of the U family at least in the neighborhood $\mathscr{U}_p$ (for instance the darker curves Fig. 3.3). Once such systems of curves have been constructed, any point $q \in \mathscr{U}_p$ in the neighborhood of p can be localized by stating on which U-curve and on which V-curve it lies. Assuming that u and v are the real parameters respectively enumerating the U and V curves, the pair of real numbers (u, v) provides the new (*Gaussian*) system of coordinates to label surface points. Using these coordinates we no longer need to make any reference to the exterior space in which M is immersed that might also be non-existing!

By introducing *curvilinear Gaussian coordinates* the King of Mathematicians freed the study of surfaces from their immersion in the external Euclidian space $\mathbb{E}$ but he immediately had to cope with a new fundamental problem. Having abolished from the list of one's mathematical instruments the straight line segments that join any two points A and B of the surface M, how can we calculate their distance? The great intuitions of Gauss were the *tangent plane* $\mathrm{T_p}M$ and the linear element ds^2, namely the metric.

Defining the absolute distance between two points A and B was no longer possible but also not interesting. In the external Euclidian space, the distance between A and B is the length of the segment which joins them, but which relevance has this datum for the little creature confined to live on the surface M, if such a segment does not lie on it? For the little two-dimensional being the only interesting datum is

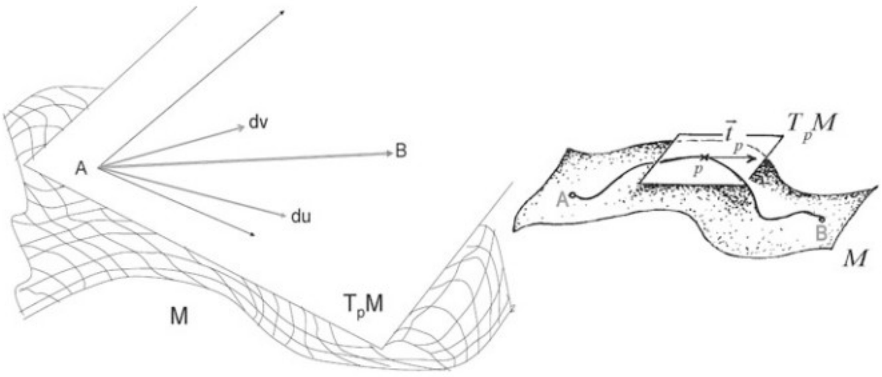

Fig. 3.4 The tangent plane to a point p of a surface M is a two-dimensional Euclidian space $\mathbb{E}^2$ where the notion of distance is defined. In the tangent plane which approximates infinitesimal regions of the surface surrounding p, we can define the line element as the Euclidian distance between two infinitesimally close points of the surface

the length of a road going from A to B: the length of any possible road with such a property! The small ant has exactly the same needs as any contemporary car-driver who wants to start on a journey. Both need to know the length of all possible paths from their origin to their destination.

Hence the problem addressed by Gauss was to give an answer to the following question: *Can we define the length of any curve departing from* $p \in M$ *and arriving at* $q \in M$ *in terms of data completely intrinsic to the surface* M*?*

Gauss' answer was positive and based on the change of perspective at the basis of the new *differential geometry*.

Let us reformulate the initial question whether we might define the absolute distance between two arbitrary points $A, B \in M$ of the surface, adding the extra condition that A and B should be only infinitesimally apart from each other. Analytically this means that if the Gaussian coordinates of A are (u, v), then those of B should be $(u + du, v + dv)$ where du and dv are infinitesimal. Gauss crucial observation is that a very small portion of the surface M around any point $p \in M$ can be approximated by a portion of the tangent plane to the surface at the point p, namely T_pM. Smaller the considered region of M better the approximation (see Fig. 3.4). This being the case we observe that Euclidian geometry makes sense in the tangent plane. Gauss remarked that the distance of two infinitesimally close points can be defined as the length of the infinitesimally short segment which joins them and which lies in the tangent plane.

Recalling Pithagora's theorem one might be tempted to say that the square length of the segments joining A and B, named ds^2, is the sum of the squared differences of the Gaussian coordinates, namely: $ds^2 = du^2 + dv^2$. Yet this is not necessarily true. In order to apply Pithagora's theorem it is required that the axes **u** and **v** be orthogonal, which is generically false. Indeed a priori it is by no means guaranteed that the curve u and the curve v, meeting at point A, should intersect there at right angle. In order to calculate ds^2 one has therefore to find the components dx and

dy of the infinitesimal segment AB in an orthogonal system of axes $\mathbf{x}$ and $\mathbf{y}$. Once these components are known one can apply Pithagora's theorem and write $ds^2 = dx^2 + dy^2$. The components dx and dy depend on the Gaussian shifts du and dv linearly:

$$\begin{pmatrix} dx \\ dy \end{pmatrix} = \begin{pmatrix} a(u,v) & b(u,v) \\ c(u,v) & d(u,v) \end{pmatrix} \begin{pmatrix} du \\ dv \end{pmatrix} \tag{3.2.1}$$

with matrix coefficients that vary from place to place on the surface, namely are functions of the Gaussian coordinate u, v. Taking this into account Gauss wrote the line element in the following way:

$$ds^2 = F(u,v)\,du^2 + G(u,v)\,dv^2 + H(u,v)\,du\,dv \tag{3.2.2}$$

where $F = a^2 + c^2$, $G = b^2 + d^2$, $H = 2ab + 2cd$.

Formula (3.2.2), written in 1828 provided the first example (a two-dimensional one) of a Riemannian metric, although Riemann was at that time only a two-year old child.

3.2.2 Bernhard Riemann Introduces n-Dimensional Metric Manifolds

The name of Riemann is associated in Mathematics with so many different and fundamental objects that the contemporary student is instinctively led to think about the scientific production of this giant of human thought as composed by a countless number of papers, books and contributions. Actually the entire corpus of Riemann's works is constituted only by 225 pages distributed over 11 articles published during the life-time of their author to which one has to add the 102 pages of the 4 posthumous publications. Among the latter there are the 16 pages of the *Ueber die Hypothesen, welche der Geometrie zu Grunde liegen* [2] which, in 1854, was debated by the candidate in front the Göttingen Faculty of Philosophy as *Habilitationsschrift*. The habilitation to teach courses was the traditional first step in the academic career foreseen by most European universities all over their very long history. In XIXth century Germany the procedure to access habilitation consisted of the writing of a dissertation on a topic chosen by the Faculty from a list of three proposed by the candidate. Typical time allowed for the preparation of such a dissertation was a couple of months and in the case of Riemann it amounted to exactly seven weeks.

Obsessed the whole of his short life by extreme poverty and by a very poor health that eventually led him to death from pulmonary consumption at the quite young age of thirty-nine, the shy and meek Bernhard Riemann, who was nonetheless quite conscious of his own talents, had already profoundly impressed Gauss with his diploma thesis. Written in 1851 and entitled *Grundlagen für eine allgemeine Theorie der Functionen einer veränderlichen complexen Grösse* which can be translated as *Principles of a General Theory of the Functions of one complex*

-256- - -* ! - B 7 11 < P 7 !A 3J !
7 ! 1"
(/ / < ' B 7! ! L

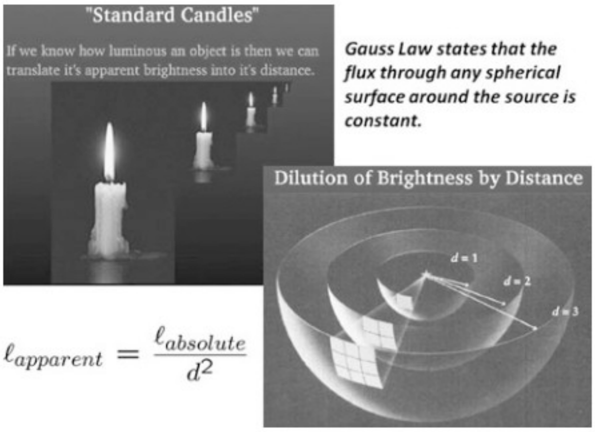
"Standard Candles"
If we know how luminous an object is then we can translate it's apparent brightness into it's distance.
Gauss Law states that the flux through any spherical surface around the source is constant.
Dilution of Brightness by Distance
d = 1
d = 2
d = 3
$\ell_{apparent} = \frac{\ell_{absolute}}{d^2}$

Fig. 4.6 The concept of standard candle is based on Gauss Law of fluxes which is satisfied by any type of massless radiation like the electromagnetic one. Since the flux through any spherical surface that surrounds the source is constant, it follows that the apparent luminosity decreases, with respect to the absolute one, with the square of the distance from the surface. If we happen to know the absolute luminosity then, by measuring the apparent one, we can easily evaluate the distance which separates our terrestrial environment from the observed object

Fig. 4.7 Henrietta Leavitt (1868–1921) was completely deaf. This notwithstanding she was one of the first women who made a scientific career in Astronomy. After graduation from the Radcliffe College, an allied institution of Harvard University, she obtained a job from the Harvard Observatory as a human-computer, namely as a human resource for those calculations that at that time, deprived of electronic computers, had to be done by hand. In 1912 she made her great discovery about variable Cepheides. She died from cancer in 1921. Proposed for the Nobel Prize by Mittag-Leffler she could not get it since she had already died at the time of assignment

The union of (6.1.2) with (6.1.3) constitutes the Neveu Schwarz algebra [7] which is a first example of a *super Lie Algebra*. The novelty that entitles it to the new qualifier "*super*" is the presence of a grading that splits the set of all generators in two classes, the *even* ones (in our case the L_n) and the *odd* ones (in our case the $G_{n+\frac{1}{2}}$). The algebra is specified by providing the *commutators* of the *even* operators with the *even* ones that necessarily produces another *even* operator, of the *even*s with the *odd*s that produces an *odd* and finally the *anti-commutator* of two *odd*s that necessarily produces an *even*.

Chronologically the Neveu Schwarz superalgebra was not really the very first, since a few months before, also in 1971, Pierre Ramond had found another very similar extension of the Virasoro algebra adding to it a set of *odd generators* G_n that are in correspondence with the integer numbers. Ramond superalgebra [4] is obtained by adjoining to (6.1.2) the following commutation, anti-commutation relations:

$$\begin{aligned} [L_m, G_n] &= \left(\frac{1}{2}m - n\right) G_{m+n} \\ \{G_m, G_n\} &= 2L_{n+m} + \frac{c}{3}\left(m^2 - \frac{1}{4}\right)\delta_{m+n,0} \\ [c, G_n] &= 0 \end{aligned} \tag{6.1.4}$$

Which algebra was the right one for superstrings? Both were right since they have a common origin in an extension of Nambu string theory by means of a new fermionic field that lives on the world-sheet traced by the string while propagating through space-time. The evolution history of the string, for instance that drawn in Fig. 6.3 or in Fig. 6.4 is described by giving the coordinate X^μ of the ambient space as functions $X^\mu(\sigma, \tau)$ of the two Gaussian coordinates $\{\sigma, \tau\}$ that label the world-sheet points. So doing we can regard the world-sheet itself as a two-dimensional space-time and the functions $X^\mu(\sigma, \tau)$ as a set of scalar fields living on it. Adopting this point of view why not consider, besides spin 0 fields also new fermionic fields of spin $\frac{1}{2}$ living in the same space-time? Let us do it and let us introduce as many new fields of this type as there were scalar fields: let us name such newcomers $\Psi^\mu(\sigma, \tau)$. What has it happened at the end of such a procedure? A surprising miracle! The little field theory that we have constructed on the world-sheet has the following marvelous properties:

- It is *supersymmetric* since it is invariant under a set of appropriate transformations that exchange the bosonic fields $X^\mu(\sigma, \tau)$ with the fermionic ones $\Psi^\mu(\sigma, \tau)$.
- It is anomaly free and quantum consistent no longer in a $D = 26$ space-time, rather in a significantly smaller one $D = 10$.
- The states encompassed in the spectrum of this quantum theory divide in two sectors, one named **NS** realizes the Neveu Schwarz algebra (6.1.2)+(6.1.3), the second named **R** realizes the Ramond algebra (6.1.2)+(6.1.4). Both sectors are necessary to construct Veneziano amplitudes for bosonic and fermionic particles.

6.1.2 Supersymmetry

While trying to insert fermions into the structure of Veneziano amplitudes, a new algebraic structure had been discovered, *supersymmetry*, which was destined to mark heavily the development of Theoretical Physics in the subsequent years. Actually in the same 1971 year, the supersymmetry algebra had been constructed in a completely different context by two Russian scientists, Golfand and Likhtman [1, 2], whose almost tragic personal story is a sort of emblem of the incredible contradictions of Soviet times, also in relation with pure science.[3] Results similar to those of Golfand and Likhtman were obtained also in Kharkov, by other two Soviet scientists, Volkov and Akulov [5] who constructed a non-linear field theoretical realization of the same super Poincaré Lie algebra discovered in Moscow.

If we analyse the meaning of the operators in the Virasoro algebra and in its extensions we understand the following: the operators L_n correspond to a mode-expansion of the stress-energy tensor, namely of the Noether current of translations P_μ, while the fermionic operators G_n correspond to the mode-expansion of a spinor-vector current J^α_μ. Which symmetry is such a Noether current associated with? The answer is unique: to some new symmetry generator Q^α which transforms as a spinor under the Lorentz group and whose anti-commutator with itself must be proportional to the translation generator P_μ.

In $D = 4$ the super Lie Algebra corresponding to such a symmetry, contains $\mathscr{N}$ of such spinor generators, named supercharges, and has the following general structure:

$$\begin{aligned}
&[J_{\mu\nu}, J_{\rho\sigma}] = -\eta_{\mu\rho} J_{\nu\sigma} + \eta_{\nu\rho} J_{\mu\sigma} - \eta_{\nu\sigma} J_{\mu\rho} + \eta_{\mu\sigma} J_{\nu\rho} \\
&[J_{\mu\nu}, P_\rho] = -\eta_{\mu\rho} P_\nu + \eta_{\nu\rho} P_\mu \\
&[J_{\mu\nu}, \overline{Q}_{A\beta}] = -\frac{1}{4}\overline{Q}_{A\alpha}(\gamma_{\mu\nu})^\alpha{}_\beta \\
&\{\overline{Q}_{A\alpha}, \overline{Q}_{B\beta}\} = \mathrm{i}\big(\mathscr{C}\gamma^\mu\big)_{\alpha\beta} P_\mu \delta_{AB} - \mathscr{C}_{\alpha\beta} Z_{AB}; \quad (A, B = 1, \ldots, \mathscr{N}) \\
&[Z_{AB}, \text{anything}] = 0
\end{aligned} \tag{6.1.5}$$

---

[3] Born in Kharkov in 1922, Yuri Abramovich Golfand got his mathematical-physical education in that Ukrainian city. Later, since 1951, he joined the Tamm group at the Lebedev Physical Institute of the Soviet Academy of Sciences in Moscow (FIAN), an institution that collected seven Nobel Prizes in Physics in the course of sixty years. Golfand and his student Likhtman conducted there, at the end of the 1960s, the studies that led them to discover the super Poincaré Lie algebra and to construct its first field theoretical realizations, published in 1971 after a long procedure of checks and inspections by the Soviet censorship authorities (see [10] for a detailed account of these facts). The next year, in the course of a routine campaign of personnel cuts, Yuri Golfand was fired from FIAN and decided to apply for an exit visa to Israel. This put him in a very bad light in front of Soviet authorities who refused the visa and treated him as a renegate. For 7 years he lived unemployed and was readmitted to FIAN only in 1980. Golfand obtained permission to emigrate to Israel only in 1990 and there he lived his last four years in Haifa, where he died in 1994. Because of his association with the renegate Golfand, also Likhtman had very difficult times with Soviet authorities and could never get a proper academic position.

*D 1 3 3

the frontiers of current research in gravitational theories, little by little absorbing a remarkable wealth of geometrical lore and of consciousness about its physical meaning.

Hopefully our reader should by now be convinced that the geometrical seeds first implanted in the XIXth century by Gauss and Riemann, not only inspired Einstein and, through his mind, produced a beautiful and so far fully verified theory, but have still a lot to say about Gravity. Whether supersymmetric or not, Gravitation is certainly the most fundamental interaction among the fundamental ones, it governs the structure of the Universe and it is a manifestation of Geometry. Which geometry the humans will still debate for a long time.

References

1. Ferrara, S., Kallosh, R.: Supersymmetry and attractors. Phys. Rev. D **54**, 1514–1524 (1996). hep-th/9602136
2. Ferrara, S., Kallosh, R., Strominger, A.: $N=2$ extremal black holes. Phys. Rev. D **52**, 5412–5416 (1995). hep-th/9508072
3. Breitenlohner, P., Maison, D., Gibbons, G.W.: Four-dimensional black holes from Kaluza-Klein theories. Commun. Math. Phys. **120**, 295 (1988)
4. Ferrara, S., Sabharwal, S.: Quaternionic manifolds for type II superstring vacua of Calabi-Yau spaces. Nucl. Phys. B **332**, 317 (1990)
5. Bergshoeff, E., Chemissany, W., Ploegh, A., Trigiante, M., Van Riet, T.: Generating geodesic flows and supergravity solutions. Nucl. Phys. B **812**, 343 (2009). arXiv:0806.2310
6. Gibbons, G.W., Kallosh, R., Kol, B.: Moduli, scalar charges, and the first law of black hole thermodynamics. Phys. Rev. Lett. **77**, 4992–4995 (1996). hep-th/9607108
7. Goldstein, K., Iizuka, N., Jena, R.P., Trivedi, S.P.: Non-supersymmetric attractors. Phys. Rev. D **72**, 124021 (2005). arXiv:hep-th/0507096
8. Tripathy, P.K., Trivedi, S.P.: Non-supersymmetric attractors in string theory. J. High Energy Phys. **0603**, 022 (2006). arXiv:hep-th/0511117
9. Kallosh, R.: New attractors. J. High Energy Phys. **0512**, 022 (2005). arXiv:hep-th/0510024
10. Giryavets, A.: New attractors and area codes. J. High Energy Phys. **0603**, 020 (2006). arXiv:hep-th/0511215
11. Kallosh, R., Sivanandam, N., Soroush, M.: The non-BPS black hole attractor equation. J. High Energy Phys. **0603**, 060 (2006). arXiv:hep-th/0602005
12. Bellucci, S., Ferrara, S., Marrani, A.: On some properties of the attractor equations. Phys. Lett. B **635**, 172 (2006). arXiv:hep-th/0602161
13. Bellucci, S., Ferrara, S., Gunaydin, M., Marrani, A.: Charge orbits of symmetric special geometries and attractors. Int. J. Mod. Phys. A **21**, 5043–5098 (2006). hep-th/0606209
14. Andrianopoli, L., D'Auria, R., Ferrara, S., Trigiante, M.: Extremal black holes in supergravity. Lect. Notes Phys. **737**, 661–727 (2008). hep-th/0611345
15. Frè, P., Gargiulo, F., Rulik, K., Trigiante, M.: The general pattern of Kac Moody extensions in supergravity and the issue of cosmic billiards. Nucl. Phys. B **741**, 42 (2006). arXiv:hep-th/0507249
16. Frè, P., Sorin, A.S.: The arrow of time and the Weyl group: All supergravity billiards are integrable. Nucl. Phys. B **815**, 430 (2009). arXiv:0710.1059
17. Frè, P., Sorin, A.S.: The integration algorithm for nilpotent orbits of $G/H^\star$ lax systems: For extremal black holes. arXiv:0903.3771
18. Frè, P., Sorin, A.S., Trigiante, M.: Integrability of supergravity black holes and new tensor classifiers of regular and nilpotent orbits. arXiv:1103.0848 [hep-th]

&　B/　(　%
!　-256-　-　-*　! - B/　7 !A　3J　1　!　5　!　A　3J ! "
<　2)　=　")　"　<　!　:!"　1
1"

D　-256-　-　-*　! - B/　8　0　7 !A　3　1
1"
(　$　9　7!　!　!　!　!

The Gaussian Constant in Celestial Mechanics

The asteroid Ceres had been discovered in 1803, but was soon lost as its path took it into the Sun's glare. The tools then available to astronomers were not so well developed as to determine the whole orbit of a thing from just a few observations. Then a young man named Karl Friedrich Gauss developed a method to determine such orbits. His method was used by astronomers to recover Ceres after its passage around the Sun. It was a stunning success and Gauss became instantly famous. Gauss evolved into one of the greatest mathematicians of all time, and contributed brilliantly to a broad range of problems. Of interest to us here is what became known as the *Gaussian constant*.

Gauss sought a constant for use with Kepler's Third Law that would be valid for all objects in any orbit around the Sun, and would make orbital calculations simpler. He derived the constant k that bears his name in his famous 1809 work, *Theory of Motion of the Heavenly Bodies Moving About the Sun in Conic Sections,* sometimes referred to by its shortened Latin title, *Theoria Motus*. Gauss used the Earth's orbit around the Sun as his reference. He took the distance between the Earth and Sun as the unit distance, and the Sun's mass as one unit.

Earlier we saw how when we changed units the equations became much simpler. The mass unit was one solar mass; the distance unit was one astronomical unit (AU). The secondary mass was assumed to be a particle of near-zero mass is orbiting the Sun. When the unit of time was in seconds, the inertial constant became this:

$$\underbrace{G = \frac{4\pi^2}{P^2}}_{G \text{ where units are solar masses, AU and seconds}}$$

When the year is the time unit (such that $P = 1$) the constant further simplified to this:

$$\underbrace{G = 4\pi^2}_{G \text{ where units are solar masses, AU and years}}$$

If, on the other hand, we use *days* as our unit of time, as Gauss did, then we have a constant that will tell us orbital periods of planets in days, which can also be very useful. In Gauss' day the year's length was not as precisely known as it is today, and with a slight adjustment for the mass of the Moon, he arrived at the value for $\sqrt{G}$[4]:

$$\sqrt{G} = .01720209895$$

---

[4] In Gauss' day the period of revolution of the earth was taken as 365.2563835 days. The combined mass of the Earth-Moon system was assumed by Gauss to be 1/354710 of the Sun's mass.

Gauss called this value k so,

$$k = .01720209895$$

This is the *Gaussian constant* or the *Gaussian gravitational constant*. Thus, with these units (solar masses, AU and days),

$$\underbrace{G = k^2}_{G \text{ where units are solar masses, AU and days}}$$

This value of k is used today in celestial mechanics for the solar system, where the mass and distance units are solar and AU, respectively, and where the time units desired are days.[5] It has fundamental importance as an astronomical standard constant.

Use of the Gaussian constant can simplify calculations in celestial mechanics. And even though the values for mass of the Earth-Moon system and Earth's orbital period have been refined over the years since the early 1800s, the Gaussian constant has been adopted as a fundamental constant in astronomy.

Using the Gaussian Constant to Find Heliocentric Periods

For bodies of negligible secondary mass orbiting the Sun, where the units are AU and days, Kepler's equation in terms of the Gaussian constant can be written as:

$$P = \frac{2\pi}{k} a^{3/2}$$

Using the Gaussian constant, we can as an example find the orbital period of Mars (neglecting its own mass), given that its semi-major axis is about 1.52 AU (or, more precisely, 1.52371034 AU)[6] from the Sun:

$$P = \frac{2\pi}{.01720209895} (1.52371034)^{3/2}$$

$$P = 686.99 \text{ days}$$

which is 1.881 years. This is an easy method to get quite accurate results in celestial mechanics calculations.

For bodies of more considerable mass, the equation reverts to this general form (where the 1 is the Sun's mass in solar mass units):

---

[5] The modern astronomical unit has been redefined (a little) to, in effect, take out the mass adjustment Gauss made in determining his constant.

[6] See http://ssd.jpl.nasa.gov

$$P = \frac{2\pi}{k\sqrt{1+m}} a^{3/2}$$

where m is the mass of the body in heliocentric orbit, in solar masses. The Gaussian constant was derived from bodies orbiting the Sun, and is commonly used for such bodies. It applies to *any* body: planet, comet, asteroid, or spacecraft in solar orbit. It was derived from the particular facts of the Earth-Sun system, because they were known, but the same constant could just as easily have been derived from another planet's data in the solar system.

Applying the Gaussian Constant to Find Heliocentric Orbital Velocities

The Gaussian constant is expressed in terms of solar masses, astronomical units and days. For finding the period of heliocentric orbits, this is very useful. Later we will be calculating velocities in heliocentric orbits. The usual result in such calculations is kilometers per second. So it would seem the AU/days units for the Gaussian constant would not be much of a time-saver. However, we can derive a conversion factor for such purposes. For the conversion we divide 1 AU in meters by the length of the day, which is 86,400 s. Hence the Gaussian AU/day units can become meters per second. To find its value, we divide

$$\underbrace{\frac{\text{AU}}{\text{Day}}}_{\text{Units for Gaussian constant k}} \quad \rightarrow \quad \underbrace{\frac{\text{AU in kilometers}}{\text{Day in seconds}}}_{\text{Conversion factor when seeking result in km/s}}$$

Inserting the correct values for the numerator and denominator,

$$\frac{149,597,871 \times 10^{8}}{86,400}$$

$$\underbrace{1731.46}_{\text{Conversion factor from AU/days to km/ sec}}$$

---

[7] The precise result is 1731.45684 which can be used where more precision is desired.

This number is good to remember.[7] To see how this is useful, recall the circular velocity in a heliocentric orbit is given by

$$v = \sqrt{\frac{Gm_\odot}{a}}$$

Using the units for the Gaussian constant k (where mass units are solar, time units are days and distances are in AU) the equation becomes,

$$v = \frac{k}{\sqrt{a}}$$

To take the example of Mars again, its *mean* velocity should be,

$$v = \frac{.01720209895}{\sqrt{1.5237}}$$

$$v = .0139358.\mathrm{AU}/\mathrm{day}$$

These are awkward units! Converting to kilometers per second, multiply the result by the conversion factor derived above and we obtain, for a mean orbital velocity of Mars, the value,

$$1731.46 \times .0139358 = 24.1\,\mathrm{km}/\,\mathrm{sec}$$

We used two constants, the Gaussian constant and a conversion factor. Combining these into one "modified" Gaussian constant for a heliocentric orbit, and calling the result the Greek letter kappa, we get

$$\kappa = 1731.46k$$
$$\kappa = 29.785$$

We will use this number frequently in heliocentric velocity calculations.[8] The equation for mean velocity in kilometers per second thus becomes,

$$v = \frac{\kappa}{\sqrt{a}}$$

---

[8] The result for κ to more decimal places is 29.78469189. We will use the rounded value for the instructional purposes of this book.

*D
($/ % 7! ! A! + W E (% !" '

(%

Geometry and Curvilinear Coordinates.

— Pietro .. Gravity, a Geometrical Course - Volume 2.. (2012).pdf

Pág. 79(97) Fig. 4.6

Gauss law of fluxes.

Pág. 217(233)

Gaussian coordinates $\{\sigma, \tau\}$

Pág. 404(418) -

Gauss and Riemann → Einstein

Obs: Usar estes 2 volumes !!!

↓ — Gravity, Special Relativity, and the Strong Force ..(2012).pdf

↓ — Newton's Gravity ...(2012).pdf

Pág. 258(272)

the Gaussian Constant in Celestial Mechanics

(% %

$\sqrt{G} = .01720209895$

In Gauss' day the periods of revolution of the earth was taken as

365.2563835 days.

the combined mass of the Earth-Moon was assumed by Gauss to be

1/354710 of the Sun's mass

Pag. 259(273)

Gauss called this value k so,

K = .01720209895

this is the Gaussian constant or the Gaussian gravitational constant.

— Metric theories of Gravity (2017).pdf

Pag. 371(396) 8

Conserved quantities in the Einstein-Gauss-Bonnet gravity (EGB)

-256- - -* ! - 9 7 !A 3
D! ! " A! ?! @! " 7 3 1"
(% %/B A " > ! = 7!) !A 3 =7)

8 Conserved quantities in the Einstein-Gauss-Bonnet gravity

8.1 Superpotentials and currents in the EGB gravity

8.1.1 Action and field equations in the EGB gravity

In the present chapter we develop and apply the results of the previous chapter to Einstein-Gauss-Bonnet (EGB) gravity. This modification of general relativity is the one of the most well-known of all modified gravitational theories. From the one side, it is the second order of the Lovelock type theories, Lagrangians of which are polynomials of the curvature tensors leaving the field equations second order [299]. From the other side, the EGB theory is the low energy limit of some string theories.

The action of the n-dimensional EGB theory with a bare cosmological term Λ_0 is

$$\begin{aligned} S &= -\frac{1}{2\kappa}\int d^n x \mathscr{L}_{EGB} + \int d^n x \mathscr{L}^M \\ &= -\frac{1}{2\kappa}\int d^n x \sqrt{-g}\,[R - 2\Lambda_0 + \alpha L_{GB}] + \int d^n x \mathscr{L}^M\,, \end{aligned} \tag{8.1.1}$$

where $\alpha > 0$, we restrict ourselves to $\Lambda_0 \le 0$, and the Gauss-Bonnet combination is

$$L_{GB} = R^2_{\mu\nu\rho\sigma} - 4R^2_{\mu\nu} + R^2\,, \tag{8.1.2}$$

see, for example, [126]. Thus $\mathscr{L}_{EGB}$ plays the role of $\mathscr{L}^G$ in (7.2.1), or in (7.2.60). Below, the subscripts "$_E$" is related to the pure Einstein part of the action (8.1.1), and the subscript "$_{GB}$" is related to the Gauss-Bonnet part connected with the α-correction.

The equations of motion that follow from (8.1.1) are

$$\mathscr{E}_{\mu\nu} = \kappa \mathscr{T}_{\mu\nu}\,, \tag{8.1.3}$$

where

$$\mathscr{E}_{\mu\nu} \equiv \frac{\delta}{\delta g^{\mu\nu}}\mathscr{L}_{EGB} = \sqrt{-g}\left[\left(R_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}R + g_{\mu\nu}\Lambda_0\right) + \alpha H_{\mu\nu}\right]\,; \tag{8.1.4}$$

$$\begin{aligned} H_{\mu\nu} \equiv 2\Big[&RR_{\mu\nu} - 2R_{\mu\sigma\nu\rho}R^{\sigma\rho} - R_{\mu\sigma\rho\tau}R_\nu{}^{\sigma\rho\tau} - 2R_{\mu\sigma}R_\nu{}^\sigma \\ &- \tfrac{1}{4}g_{\mu\nu}\left(R^2_{\tau\lambda\rho\sigma} - 4R^2_{\rho\sigma} + R^2\right)\Big]\,, \end{aligned} \tag{8.1.5}$$

and $\mathscr{T}_{\mu\nu}$ is the standard matter energy-momentum defined as in (1.3.21). The expression $H_{\mu\nu}$ is a result of varying L_{GB}, and $H_{\mu\nu} \equiv 0$ if $n \le 4$.

DOI 10.1515/9783110351781-008

The equations (8.1.3) are the variant of the equations (7.2.3) for the case of the EGB gravity. The equations (8.1.3) are also

$$\mathscr{E}_{\mu\nu} = 0 \tag{8.1.6}$$

in the vacuum case.

8.1.2 Three types of superpotentials

In deriving concrete expressions for conserved quantities in the EGB gravity, we change the order of the presentation used in Sections 7.1 and 7.2, where currents have been described firstly. Indeed, considering an abstract theory, it is more natural to begin the construction from currents, then to continue the construction of superpotentials. Here, inversely, we begin from superpotentials, and finish with the currents. The reasons are as follow. First, anyway we keep in mind the principal scheme of construction in Chapter 7 and already have at hand the basic general formulae for such a construction. Second, historically superpotentials have been more popular and more desirable in applications because they give a direct way for constructing conserved charges. Third, superpotential expressions are significantly simpler than the expressions for currents.

The canonical prescription

At first, one has to define the KBL type Lagrangian for the EGB gravity. Following to (7.2.15), we derive it as

$$\mathscr{L}_{KBL} = -\frac{1}{2\kappa}\left(\mathscr{L}^{c}_{EGB} - \bar{\mathscr{L}}^{c}_{EGB} + \partial_\alpha \mathscr{D}^\alpha\right) . \tag{8.1.7}$$

Following the recommendation (7.2.18), we construct the canonical superpotential in the EGB gravity,

$$\boldsymbol{j}^{*\alpha\beta}_{C}(\xi) = \boldsymbol{i}^{*\alpha\beta}(\xi) - \bar{\boldsymbol{i}}^{*\alpha\beta}(\xi) + \boldsymbol{i}'^{\alpha\beta}(\xi) . \tag{8.1.8}$$

From the beginning we construct the *starred* expressions, coefficients for which are derived in Appendix C. Possibilities of the developed approaches in Section 7.2 permit us to construct families of superpotentials easily. Here we use these possibilities by applying concrete expressions in the EGB case. Thus, the pure Noether's canonical starred superpotential of the general form (7.1.46) in EGB gravity acquires the concrete form:

$$\begin{aligned}
\boldsymbol{i}^{*\alpha\beta} = {} & \frac{1}{\kappa}\left(\mathfrak{g}^{\rho[\alpha}\bar{\nabla}_\rho \xi^{\beta]} + \mathfrak{g}^{\rho[\alpha}\Delta^{\beta]}_{\rho\sigma}\xi^\sigma\right)\\
& - 2\alpha\frac{1}{\kappa}\left\{\Delta^{\rho}_{\lambda\sigma}\mathscr{R}_{\rho}{}^{\lambda\alpha\beta} + 4\Delta^{\rho}_{\lambda\sigma}g^{\lambda[\alpha}\mathscr{R}^{\beta]}_{\rho} + \Delta^{[\alpha}_{\rho\sigma}g^{\beta]\rho}\mathscr{R}\right\}\xi^\sigma\\
& - 2\alpha\frac{1}{\kappa}\left\{\mathscr{R}_{\sigma}{}^{\lambda\alpha\beta} + 4g^{\lambda[\alpha}\mathscr{R}^{\beta]}_{\sigma} + \delta^{[\alpha}_{\sigma}g^{\beta]\lambda}\mathscr{R}\right\}\bar{\nabla}_\lambda\xi^\sigma ,
\end{aligned} \tag{8.1.9}$$

(' $ '$ = 7!) 9 3

a testament. All the terms, parameters, contribute, in a highly non-trivial way, to the masses which are only defined at a given constant curvature background. Let us consider the three and the four dimensional cases as two explicit examples as they offer interesting limits and some other special theories.

The case of $n = 3$ dimensions

The masses of the spin-2 and spin-0 modes respectively read

$$m_g^2 = -\frac{1}{\kappa\beta} - 4\left(1 + 3\frac{\alpha}{\beta}\right)\Lambda, \qquad m_s^2 = \frac{1}{(8\alpha + 3\beta)\kappa} - \frac{4(3\alpha + \beta)}{(8\alpha + 3\beta)}\Lambda, \tag{9.2.76}$$

which were also found with the usual canonical analysis in [210] for the de Sitter case. It is also important to note that, since there is no massless graviton these are the 3 degrees of freedom in 3 dimensions. This fact has remarkable consequences, for the choice $8\alpha + 3\beta = 0$, the scalar mode decouples and one can make the massive spin-2 mode unitary. This theory is called the "New Massive Gravity" (NMG) theory noted above [39] which is a theory of massive spin-2 graviton at the non-linear level. [2]

The case of $n = 4$ dimensions

The masses of the spin-2 and spin-0 modes respectively read

$$m_g^2 = -\frac{1}{\kappa\beta} - 2\left(1 + 4\frac{\alpha}{\beta}\right)\Lambda, \qquad m_s^2 = \frac{1}{2(3\alpha + \beta)\kappa}. \tag{9.2.77}$$

It is interesting to note that only in four dimensions, for (A)dS backgrounds, the massless mode does not receive a correction from the background. In addition to these two modes, there is the Einsteinian massless spin-2 graviton which altogether make up the 8 degrees of freedom whose flat space versions were given by Stelle [423]. An often studied sub-case of quadratic gravity is the Weyl-square corrected Einstein's gravity for which $3\alpha + \beta = 0$ in four dimensions. In this case the scalar mode decouples but the theory has a massive spin-2 ghost. In addition, if one drops out the Einstein term and keeps only the quadratic part, one has the so called conformal gravity [309, 389] since the combination $R_{\mu\nu}^2 - \frac{1}{3}R^2$ is equal to the Weyl-square up to a boundary term.

Einstein-Gauss-Bonnet theory

For the Einstein-Gauss-Bonnet (Einstein-GB) theory in n dimensions, since $\alpha = \beta = 0$, both massive modes decouple and only the massless mode survives as expected since the field equations are second order. But there is a constraint on the theory for unitarity since the effective Newton's constant now reads

---

[2] As a side remark, let us note that NMG suffers from bulk-boundary unitarity clash, namely the theory does not have a unitary boundary conformal field theory in its 2 dimensional boundary of AdS_3 as one of its central charges of the double copy of the Virasoro algebra turns out to be negative. This unitarity clash seems to persist in all extensions of NMG with massive spin-2 gravitons [212].

$$\frac{1}{\kappa_{\text{eff}}} \equiv \frac{1}{\kappa} + \frac{4\Lambda\,(n-3)\,(n-4)}{(n-1)\,(n-2)}\gamma = \frac{1}{\kappa} + 4\Lambda k, \tag{9.2.78}$$

which must be positive to get a positive kinetic energy for the massless graviton or equivalently, to get attractive gravity and this condition must be compatible with the existence of a maximally symmetric solution that is $1 + 8k\kappa\Lambda_0 \geq 0$. This theory arises in low energy string theory [71, 477].

Critical gravity

Critical gravity corresponds to the decoupling of the massive spin-0 mode with the choice $4\alpha(n-1)+\beta n = 0$ while tuning the cosmological constant in such a way that the massive spin-2 mode also decouples. The resulting theory is a higher derivative theory with an action of the form [123, 300]

$$S = \int d^n x\, \sqrt{-g}\left(\frac{1}{\kappa}\,(R - 2\Lambda_0) + \alpha R^2 - \frac{4\,(n-1)}{n}\alpha R^{\mu\nu}R_{\mu\nu} + \gamma\chi_{\text{GB}}\right), \tag{9.2.79}$$

where the bare cosmological constant must be tuned as

$$\Lambda_0 = -\frac{n^2\,(n-1)\,(n-2)\,\left[(n-1)\,(n-2)\,\alpha + (n-3)\,(n-4)\,\gamma\right]}{8\kappa\left[(n-1)\,(n-2)^2\,\alpha + n\,(n-3)\,(n-4)\,\gamma\right]^2}. \tag{9.2.80}$$

Even though this theory seems promising as a possible ghost-free, perhaps renormalizable gravity theory, it turns out this is not the case. The theory has asymptotically non-AdS logarithmic modes [3, 215] which are of the wave-type and arise as both exact and as perturbative solutions. It was shown in [381] that they are ghosts and truncation of these modes renders the Hilbert space trivial, devoid of anything but the vacuum. Hence the above discussion teaches us something important: for $n \geq 4$ one must have $\beta = 0$ to avoid the massive spin-2 ghost, $n = 3$ is the exceptional case.

9.3 Particle spectrum of $f(R_{\mu\nu\sigma\rho})$ gravity in (A)dS

Now we turn our attention to the following more general theory with f an arbitrary analytic function of the Riemann tensor

$$S = \int d^n x\, \sqrt{-g} f\left(R^{\mu\nu}_{\alpha\beta}\right). \tag{9.3.1}$$

In this formulation, with two up and two down indices on the Riemann tensor, the (inverse) metric is not needed for contractions and this simplifies the ensuing discussion a little bit. For example, in this notation, the Einstein-Hilbert action (up to a multiplicative constant) simply reads

data

Pag. 425 (450)
Einstein-Gauss-Bonnet Theory
— Equations of Motion in Relativistic Gravity (2015).pdf
Pag. 12 (25)

The difference between their values for two different choices of Σ can be converted by Gauss's theorem to volume integrals. The integrands then vanish by the conservation equation for $T^{\alpha\beta}$ and because it is symmetric.

Pag. 253 (263)
The Dependence of the Particle's Momenta on Σ
Gauss theorem (see Fig. 7)
Pag. 254 (264) shadowed regions
⇒ hypersurfaces. IMPORTANT!

! -256- - -* ! - => ! 5 ! A
7 !A 3 1 $ 1"
($

way, progression from the dipole to the full quadrupole case would have added a third evolution equation and all three would have involved time derivatives of quadrupole components. Each successive approximation would add a further equation in this way, and a complete solution of the variational equation would have an infinite set of equations, each involving time derivatives of an infinite set of moments.

In Newtonian mechanics there would only be two evolution equations, those for momentum and angular momentum. Mathisson's new mechanics should aim at maintaining this simplicity. This would require eliminations between the full set of equations to leave only two, those governing momentum and angular momentum. The issue is whether or not this can be achieved, and if so how it can be done.

The final issue concerns the supplementary condition (18). Equation (17) is of second order in $S^{\alpha\beta}$. By use of (18) it can be put in a form that is of first order in $S^{\alpha\beta}$ but then it becomes second order in v^α. The set of equations (15), (17) and (18) then has solutions corresponding to arbitrary initial values of position, velocity, spin *and acceleration*. This is greater freedom than classical mechanics allows.

If the equations of motion are considered to represent a point particle then it is a matter of fact whether (18) is or is not satisfied. In a second paper [12] in 1937, Mathisson showed that this additional freedom in the solutions provides a classical analogue of the *zitterbewegung* of the quantum theory of the electron. But if the equations of motion are to represent an extended body then the supplementary condition selects the world line used to represent its position. It is no longer a matter of fact, but of choice, whether (18) holds. The freedom to specify an initial acceleration is now a defect. It does not correspond to a physical freedom but merely to the failure of (18) to determine a unique world line. It points to the need for a different supplementary condition that does not admit such solutions.

5 The Unphysical Solutions

The unphysical solutions allowed by the supplementary condition (18) are present also in special relativity. In this context they were studied by a number of authors in the period 1939–1950. It was noted above that they must be avoided when modelling an extended body. To achieve this, their origin needs to be properly understood. The following account is based on the work of Pryce [13] and Møller [14] and uses a Minkowskian coordinate system in the flat spacetime of special relativity.

Consider an extended body with symmetric energy-momentum tensor $T^{\alpha\beta}$ that satisfies the conservation equation (3). Given a spacelike hypersurface Σ that spans the body and a point z not necessarily on Σ, let

$$p^\alpha \equiv \int_\Sigma T^{\alpha\beta} \mathrm{d}\Sigma_\beta \tag{26}$$

and

$$S^{\alpha\beta} \equiv 2\int_{\Sigma}(x^{[\alpha} - z^{[\alpha}(s))T^{\beta]\gamma}\mathrm{d}\Sigma_{\gamma}. \tag{27}$$

The difference between their values for two different choices of Σ can be converted by Gauss's theorem to volume integrals. The integrands then vanish by the conservation equation for $T^{\alpha\beta}$ and because it is symmetric. These quantities are therefore independent of the choice of Σ. As the definition of p^{α} does not involve z, it follows that p^{α} is constant while $S^{\alpha\beta}$ becomes a function only of the point z. These properties have led to them becoming the generally accepted definitions of momentum and angular momentum (spin) for an isolated body in special relativity. Generalisations to more complex situations such as general relativity or the presence of external forces need to reduce to these in this special case.

Let z move along an arbitrary timelike world line L parameterised as $z(s)$ by proper time s and let $v^{\alpha} = \mathrm{d}z^{\alpha}/\mathrm{d}s$, which satisfies $v^{\alpha}v_{\alpha} = 1$. Differentiate (27) with respect to s, holding Σ fixed as the values of the integrals do not depend on it. This gives

$$\frac{\mathrm{d}S^{\alpha\beta}}{\mathrm{d}s} = 2p^{[\alpha}v^{\beta]} \tag{28}$$

from which

$$p^{\alpha} = Mv^{\alpha} + v_{\beta}\frac{\mathrm{d}S^{\alpha\beta}}{\mathrm{d}s} \tag{29}$$

where $M \equiv p^{\alpha}v_{\alpha}$. Since p^{α} is constant, (29) gives

$$\frac{\mathrm{d}}{\mathrm{d}s}\left(Mv^{\alpha} + v_{\beta}\frac{\mathrm{d}S^{\alpha\beta}}{\mathrm{d}s}\right) = 0 \tag{30}$$

while (28) and (29) together give

$$\frac{\mathrm{d}S^{\alpha\beta}}{\mathrm{d}s} + v^{\alpha}v_{\gamma}\frac{\mathrm{d}S^{\beta\gamma}}{\mathrm{d}s} - v^{\beta}v_{\gamma}\frac{\mathrm{d}S^{\alpha\gamma}}{\mathrm{d}s} = 0. \tag{31}$$

These are precisely the forms taken in flat spacetime by Mathisson's equations (15) and (17) in his dipole approximation, but now they are exact and the supplementary condition (18) has not yet been imposed. The variables also agree, to the accuracy of his approximation. To see this, note first that $m^{\alpha\beta\gamma} = M^{\alpha\beta\gamma}$ in the notation of (24) when quadrupole and higher contributions are neglected, and that the definition (27) of $S^{\alpha\beta}$ corresponds to $S^{\alpha\beta} = 2M^{[\alpha\beta]\gamma}v_{\gamma}$ to this accuracy. But (14) gives $2m^{[\alpha\beta]\gamma}v_{\gamma} = S^{\alpha\beta}$, so it follows that the two tensors $S^{\alpha\beta}$ are in agreement. The vector v^{α} is the unit tangent to L in both cases and in order for the right hand side of (29) to be constant with both definitions of M, these must also agree.

This agreement enables the origin of the unphysical solutions arising from the supplementary condition (18) to be studied in a context where the equations of motion are exact. First suppose that z is any point on some L that satisfies (18). Choose Σ in (27) to be the hyperplane $v_\alpha(x^\alpha - z^\alpha) = 0$ through z. For this choice, $\mathrm{d}\Sigma_\alpha$ is parallel to v_α so that $\mathrm{d}\Sigma_\alpha = v_\alpha v^\gamma \mathrm{d}\Sigma_\gamma$. Hence (18) with (27) gives

$$z^\alpha(s) = \int x^\alpha T^{\beta\gamma} v_\beta v_\gamma v^\delta \mathrm{d}\Sigma_\delta \Big/ \int T^{\beta\gamma} v_\beta v_\gamma v^\delta \mathrm{d}\Sigma_\delta, \tag{32}$$

showing that z^α is a centre of mass for the mass (or energy) density $T^{\alpha\beta} v_\alpha v_\beta$ in the instantaneous rest frame. It follows that for a body with positive energy density throughout, i.e., $T^{\alpha\beta} v_\alpha v_\beta > 0$ for all choices of timelike v^α, L lies within the body (suitably interpreted if spacelike cross-sections of the body are not convex).

This is a restriction on the solutions for L but it does not make L unique. To study the non-uniqueness in more detail, let

$$K^{\alpha\beta} \equiv 2 \int x^{[\alpha} T^{\beta]\gamma} \mathrm{d}\Sigma_\gamma \tag{33}$$

so that (27) gives

$$S^{\alpha\beta}(s) = K^{\alpha\beta} - 2z^{[\alpha}(s) p^{\beta]}. \tag{34}$$

This shows that there is a unique world line parameterised by a scalar λ, given by

$$z^\alpha = \lambda p^\alpha + \frac{p_\beta K^{\alpha\beta}}{p^\gamma p_\gamma}, \tag{35}$$

on which

$$p_\beta S^{\alpha\beta} = 0. \tag{36}$$

If the Minkowskian coordinate system is chosen so that this line is the time axis then it is also given by $z^\alpha = \lambda p^\alpha$, so (34) gives $S^{\alpha\beta}(s) = K^{\alpha\beta}$ and (36) becomes

$$p_\beta K^{\alpha\beta} = 0. \tag{37}$$

Let Latin indices take values 1, 2, 3 and let m be the magnitude of p^α, so $m^2 = p^\alpha p_\alpha$. Then with these coordinates $p^i = 0$, $p^0 = m$ and hence from (37) also $K^{i0} = 0$. From (34) the supplementary condition (18) can be put in the form

$$v_\beta \left(K^{\alpha\beta} - 2z^{[\alpha}(s) p^{\beta]} \right) = 0. \tag{38}$$

This has only three linearly independent components, which can be taken as those with $\alpha = 1, 2, 3$, since multiplication by v_α gives an identity. It is therefore equivalent to

$$v_j K^{ij} - m z^i v_0 = 0 \tag{39}$$

in the chosen Minkowskian coordinates. Since K^{ij} is antisymmetric it can be represented by a three-dimensional vector K_i such that $K^{ij} = \epsilon^{ikj} K_k$ where ϵ^{ijk} is the three-dimensional Levi-Civita symbol. If $t \equiv x^0$ is coordinate time then $v^i/v^0 = \mathrm{d}z^i/\mathrm{d}t$. These results enable (39) to be expressed in three-dimensional vector notation as

$$m\mathbf{z} - \frac{\mathrm{d}\mathbf{z}}{\mathrm{d}t} \times \mathbf{K} = 0. \tag{40}$$

Since $\mathbf{K}$ is constant, the time derivative of (40) shows that $\mathrm{d}\mathbf{z}/\mathrm{d}\mathbf{t}$ and $\mathbf{K}$ are orthogonal. The vector product of (40) with $\mathbf{K}$ therefore gives

$$\frac{\mathrm{d}\mathbf{z}}{\mathrm{d}t} = \frac{m}{\mathbf{K}^2}\mathbf{K} \times \mathbf{z}. \tag{41}$$

This corresponds to motion about the origin in a circle with angular velocity m/K, where $K = |\mathbf{K}|$, in a plane perpendicular to $\mathbf{K}$. From (40) the radius of the circle is vK/m where $v = |\mathrm{d}\mathbf{z}/\mathrm{d}t|$. Units here have the velocity of light $c = 1$ so since L is timelike, $v < 1$. The possible circular motions therefore all lie in a disc of radius K/m in this plane and centred on the origin, but all such motions give an L that satisfies the supplementary condition, so confirming the non-uniqueness claimed above. For a body with positive energy density throughout, it was seen above that all choices of L that satisfy the supplementary condition lie within the body, suitably interpreted. It follows that this disc lies within the body, which gives a minimum size for such a body when its spin is nonzero.

A different perspective arises if the Eqs. (30) and (31) are taken in isolation, so that the quantities M and $S^{\alpha\beta}$ are considered to be properties of the body rather than constructs from the energy-momentum tensor. This is the situation when the equations of motion are used to model a point particle whose motion satisfies (18) as a matter of physical fact. The same results hold but they need to be cast in terms of M and $S^{\alpha\beta}$ rather than m and K. With the same choice of Minkowskian coordinate system used above, (34) gives $S^{ij} = K^{ij}$ but whereas $K^{i0} = 0$, S^{i0} is not zero. From (18),

$$S^{i0} = -S^{ij} v_j/v_0 = -K^{ij} v_j/v_0 = -\epsilon^{ikj}\frac{\mathrm{d}z_j}{\mathrm{d}t}K_k. \tag{42}$$

But $\mathrm{d}\mathbf{z}/\mathrm{d}t$ and $\mathbf{K}$ are orthogonal three-dimensional vectors and $S_{i0} = -S^{i0}$ due to the form of the metric tensor, so

$$S^{i0}S_{i0} = -(vK)^2. \tag{43}$$

With the magnitude S of the spin tensor $S^{\alpha\beta}$ defined by $S^2 = S^{\alpha\beta}S_{\alpha\beta}/2$, (43) gives

$$S^2 = K^2(1 - v^2). \tag{44}$$

Also

$$M \equiv p^{\alpha} v_{\alpha} = m v_0 = m/\sqrt{1 - v^2} \tag{45}$$

so

$$\frac{vK}{m} = \frac{vS}{M(1 - v^2)} \tag{46}$$

which gives the radius of the circular orbit in terms of S and M as required. But whereas $v < 1$, which gave a limiting size to the radius, $v/(1 - v^2)$ is unbounded as $v \to 1$. With this perspective there is no limit on the size of the circular orbits. As described above, the initial acceleration of the particle can be arbitrary and it is this choice that determines which orbit is followed.

This form of solution to the Eqs. (30) and (31) in flat spacetime was first given by Weyssenhoff and Raabe [15]. Mathisson [12] had earlier given a solution in the non-relativistic approximation that he had used to provide a classical analogue of the *zitterbewegung* of the electron. But as discussed in the preceding section, when the equations are used to model an extended body then these unphysical solutions must be avoided. The above analysis provides the clue to their avoidance, for it was seen that (36) *does* determine a unique world line. So in special relativity, if (36) is adopted as the supplementary condition in place of (18) then unphysical solutions are avoided. The issue in a curved spacetime is to find appropriate meanings for p^{α} and $S^{\alpha\beta}$. This will be addressed in Sect. 21.

6 The Period 1950–1970

At his death in 1940, Mathisson [9] had been studying his variational equation in special relativity for a body under the influence of external forces, maybe looking for clues as to how to progress further with it in general relativity. This work led to a paper by Shanmugadhasan [16] in 1946 that continued this development in special relativity. The multipole approach to equations of motion in general relativity went dormant, however, until the appearance of a paper by Papapetrou [17] in 1951 entitled "Spinning test-particles in general relativity. I", a title that does not show anything like the same vision as did Mathisson's.

Papapetrou defined a set of non-covariant moment integrals identical to those of Bielecki et al. [8] given as (24) above. He continued by developing equations of motion in terms of these quantities and other similarly defined ones, also non-covariant, constructing combinations of them along the way that he subsequently proved did transform as tensors. This was done first for a particle in which only the monopole moment is retained, and then when the dipole moment is also retained, which he calls a pole-dipole particle. In this way he finally arrives at the Eqs. (15) and (17) first obtained by Mathisson. Mathisson's paper [1] of 1937 is given as a reference, but with the statement that it gives a discussion of the pole-dipole particle in special relativity. Since Papapetrou's paper of 1951 in Proceedings of the Royal and a is the largest dimension of the body. Thus, when $\lambda \ll 1$, i.e. when the curvature is not too strong compared to the scale of the size of the body,[13] $P^{\hat{\alpha}}_{\text{Dix}} \simeq P^{\hat{\alpha}}$. The two schemes are actually indistinguishable in a pole-dipole approximation, where only terms to linear order in x are kept in the integrals defining the moments; the resulting equations of motion are the same (compare Eqs. (43), (49) of [25] with Eqs. (6.31) and (6.32) of [11], or Eqs. (7.1) and (7.2) of [26]), both schemes leading to the well known Mathisson-Papapetrou equations (the latter derived using less sophisticated formalisms). These conclusions are natural, for the metric in Riemann normal coordinates is (e.g. [27]) of the form $g_{\hat{\alpha}\hat{\beta}} = \eta_{\hat{\alpha}\hat{\beta}} + \mathcal{O}(\|\mathbf{R}\|x^2)$; hence the assumption $\lambda \ll 1$ amounts to say that, *for the computation* of P^α and $S^{\alpha\beta}$, one may, to a good approximation, take the spacetime as nearly flat throughout the body.

The Dependence of the Particle's Momenta on Σ

The momenta (1) and (2) depend, in general, on the spacelike hypersurface $\Sigma(z,u) \equiv \Sigma(z(\tau),u)$ on which the integration is performed, see e.g. [11, 25, 26]. This is so even in flat spacetime; when forces and torques act on the body, it is clear that $P^\alpha(z,u)$, $S^{\alpha\beta}(z,u)$ depend on $z^\alpha(\tau)$, and also on the argument u^α of Σ. Curvature brings additional complications, as u^α is no longer a "free vector", and Σ itself is in principle point dependent. Herein we shall show that, in the absence of electromagnetic field ($F^{\alpha\beta} = 0$), and under the assumption $\lambda \ll 1$ made above, for hypersurfaces $\Sigma(z,u)$ *through a point* z^α *within the body*'s convex hull, the u^α dependence of the momentum and angular momentum is negligible.

Denote by $\boldsymbol{\xi} = dx^{\hat{\alpha}}$ a particular basis 1-form of the Riemann normal coordinate system $\{x^{\hat{\alpha}}\}$; $P^{\boldsymbol{\xi}} \equiv P^\alpha \xi_\alpha$ is thus the $\boldsymbol{\xi}$ component of P^α. From definition (1), and since ξ_α has constant components, we may write the $\boldsymbol{\xi}$ component of the momentum as the integral of a vector $A^\alpha \equiv T^{\alpha\beta}\xi_\beta$ on a 3-surface,

$$P^{\boldsymbol{\xi}}(z,u) = \xi_{\hat{\alpha}} \int_{\Sigma(z,u)} T^{\hat{\alpha}\hat{\beta}} d\Sigma_{\hat{\beta}} = \int_{\Sigma(z,u)} T^{\hat{\alpha}\hat{\beta}} \xi_{\hat{\alpha}} d\Sigma_{\hat{\beta}} = \int_{\Sigma(z,u)} A^\beta d\Sigma_\beta \ .$$

Take $u^\alpha = P^\alpha/M$, and consider another vector u'^α at the same point z^α; the $\boldsymbol{\xi}$ component of the difference between the momenta computed in the hypersurfaces $\Sigma(z,u')$ and $\Sigma(z,u)$, $\Delta P^{\boldsymbol{\xi}} \equiv P^{\boldsymbol{\xi}}(z,u') - P^{\boldsymbol{\xi}}(z,u)$ is, from an application of the Gauss theorem (see Fig. 7),

$$\Delta P^{\boldsymbol{\xi}} = \int_{\Sigma(z,u)} A^\beta d\Sigma_\beta - \int_{\Sigma(z,u')} A^\beta d\Sigma_\beta = \int_{V_{Left}} A^\beta{}_{;\beta} dV - \int_{V_{Right}} A^\beta{}_{;\beta} dV \ .$$

---

[13] For example, in the case of the Schwarzschild spacetime, $\|\mathbf{R}\| \sim m_S/r^3$, $\lambda = (m_S/r)(a^2/r^2)$; since $m_S/r < 1$ for any point outside the horizon, $\lambda \ll 1$ is guaranteed just by taking the size of the body much smaller than its the distance to the source, $r^2 \gg a^2$.

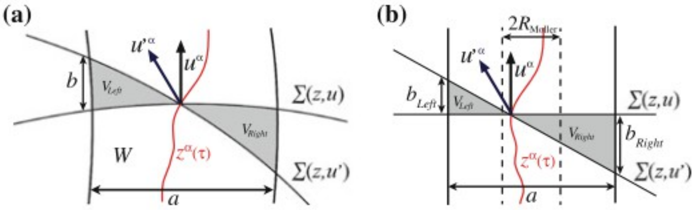

Fig. 7 *Shadowed regions* V_{Left} and V_{Right} are the 4-volumes delimited by the hypersurfaces $\Sigma(z,u')$, $\Sigma(z,u)$, and the boundary of the body's worldtube, of convex hull W. u^α is chosen parallel to P^α. **a** Curved spacetime; **b** flat spacetime

Here V_{Left} and V_{Right} denote the shadowed regions of Fig. 7 (where $A^\alpha \neq 0$), i.e. the "left" and "right" 4-volumes delimited by $\Sigma(z,u')$, $\Sigma(z,u)$ and the boundary of the body's worldtube. Now, using the conservation law $T^{\alpha\beta}_{\ \ \ ;\beta} = 0$, one notes that

$$A^\beta_{\ ;\beta} = T^{\alpha\beta}_{\ \ \ ;\beta}\xi_\alpha + T^{\alpha\beta}\xi_{\alpha;\beta} = T^{\alpha\beta}\xi_{\alpha;\beta};$$

thus

$$\Delta P^{\boldsymbol{\xi}} = \int_{V_{Left}} T^{\alpha\beta}\xi_{\alpha;\beta}dV - \int_{V_{Right}} T^{\alpha\beta}\xi_{\alpha;\beta}dV.$$

Since $\boldsymbol{\xi}$ is a basis 1-form, $\xi_{\hat{\alpha},\hat{\beta}} = 0$, and

$$\xi_{\hat{\alpha};\hat{\beta}} = -\Gamma^{\hat{\gamma}}_{\hat{\alpha}\hat{\beta}}\xi_{\hat{\gamma}} = \mathcal{O}(\|\mathbf{R}\|x);$$

therefore

$$|\Delta P^{\boldsymbol{\xi}}| \lesssim \|\mathbf{R}\| \int_V T^{\hat{0}\hat{0}}|x|dV = \|\mathbf{R}\| V \left\langle T^{\hat{0}\hat{0}}|x| \right\rangle,$$

where $V \equiv V_{Left} + V_{Right}$, $\langle\cdot\rangle$ denotes the average on the shadowed region of Fig. 7a, and we noted that $T^{\hat{0}\hat{0}}$ is the largest component of $T^{\alpha\beta}$ and always positive. Since $b < av(u',u)$ (see Fig. 7), with $v^\alpha(u',u)$ defined by Eq. (9), and $v(u',u) < 1$, then $\langle|x|\rangle < a$; moreover (assuming $\partial_{\hat{0}} = \mathbf{u}$ at z), $V\left\langle T^{\hat{0}\hat{0}}\right\rangle < Mav(u',u)$; hence we get

$$|\Delta P^{\boldsymbol{\xi}}| \lesssim M\lambda v(u',u) = \|\mathbf{P}\|\lambda v(u',u)\ , \tag{70}$$

showing that ΔP^α is negligible[14] compared to P^α under the restriction above on the strength of the gravitational field, $\lambda \ll 1$ (the same under which the different

---

[14] The inequality (69) means not only that the components of $\Delta\mathbf{P}$ in the system $\{x^{\hat{\alpha}}\}$ (where $P^{\hat{i}} = 0$) are much smaller than M, but also much smaller than the typical spatial momentum in other frames. For instance, in normal coordinates $\{x^{\alpha'}\}$ comoving with u'^α, one has $|P^{i'}| \sim \gamma(u',u)v(u',u)M$ and $|\Delta P^{i'}| \lesssim |\Delta P^{\boldsymbol{\xi}}|\gamma(u',u)$, thus $|\Delta P^{i'}| \ll |P^{i'}|$ when $\lambda \ll 1$.

*D & 1 ! = "! ! A! "! < !

multipole schemes become equivalent, and one can take local Lorentz coordinates as nearly rectangular throughout the extension of the body; see also footnote 13). In the application in Sect. 3.4—Schwarzschild spacetime, far field limit—we can write

$$|\Delta P^{\boldsymbol{\xi}}| \lesssim \frac{M m_S}{r^3} a^2 v(u', u) \simeq \| P_{\text{hidI}} \| \frac{a}{R_{\text{Moller}}} \frac{a}{r} v(u', u)$$

where P^{α}_{hidI} is the inertial hidden momentum *of the CP condition*, Eq. (57) (P^{α}_{hidI} is zero or negligible for the other solutions). Thus ΔP^{α} is negligible compared to P^{α}_{hidI} under the condition $\frac{R_{\text{Moller}}}{a} \gg \frac{a}{r} v(u', u)$, which is reasonable in a problem where the particle's spin is worth taking into account (e.g. in the problem of nearly circular motion in Sect. 3.4 this amounts to taking $\omega_{\text{body}} \gg \omega_{\text{orbit}}$, where ω_{body} and ω_{orbit} are the body's rotation and orbital angular velocities).

Through an analogous procedure, one can show that the dependence of $S^{\alpha\beta}$ on u^{α} is negligible in this regime. Let $\mathscr{J}^{\hat{\alpha}\hat{\beta}\hat{\gamma}} \equiv 2x^{[\hat{\alpha}}T^{\hat{\beta}]\hat{\gamma}}$, so that $S^{\hat{\alpha}\hat{\beta}} = \int_{\Sigma(z,u)} \mathscr{J}^{\hat{\alpha}\hat{\beta}\hat{\gamma}} d\Sigma_{\hat{\gamma}}$; and consider the two basis *spatial* 1-forms $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$. Constructing the vector $\mathscr{J}^{\gamma} \equiv \mathscr{J}^{\alpha\beta\gamma}\xi_{\alpha}\eta_{\beta}$, we can write the $\boldsymbol{\xi} \otimes \boldsymbol{\eta}$ component of $S^{\alpha\beta}$ as $S^{\boldsymbol{\xi\eta}}(z, u) = \int_{\Sigma(z,u)} \mathscr{J}^{\beta} d\Sigma_{\beta}$. By the Gauss theorem,

$$\begin{aligned} \Delta S^{\boldsymbol{\xi\eta}} = \int_{V_{Left}} \mathscr{J}^{\beta}_{\ ;\beta} dV - \int_{V_{Right}} \mathscr{J}^{\beta}_{\ ;\beta} dV \ \sim \ \|\mathbf{R}\| \int_V x^2 |\vec{J}| dV = \|\mathbf{R}\| V \langle |\vec{J}| x^2 \rangle \\ < \|\mathbf{R}\| a^2 V \langle |\vec{J}| \rangle = \lambda V \langle |\vec{J}| \rangle \lesssim \lambda a^4 v(u', u) \langle |\vec{J}| \rangle \ , \end{aligned}$$

where $J^{\hat{i}} = T^{\hat{0}\hat{i}}$. In the second relation again we used $\Gamma^{\hat{\gamma}}_{\hat{\alpha}\hat{\beta}} = \mathcal{O}(\|\mathbf{R}\| x)$. Since $S = \mathcal{O}(a^4 \langle |\vec{J}| \rangle)$, cf. Eq. (13), we see that indeed when $\lambda \ll 1$, $\|\Delta S^{\boldsymbol{\xi\eta}}\| \ll S$.

The Case with Electromagnetic Field

When $F^{\alpha\beta} \neq 0$, the conservation law is $T^{\alpha\beta}_{\ \ ;\beta} = F^{\alpha\beta} j_{\beta}$ (denoting by $T^{\alpha\beta}$ *the particle's* energy momentum tensor). Consider for simplicity flat spacetime, and let $\boldsymbol{\xi}$ be a basis 1-form of a *global* Lorentz system; then $T^{\alpha\beta}_{\ \ ;\beta}\xi_{\alpha} = (T^{\alpha\beta}\xi_{\alpha})_{;\beta} \equiv A^{\beta}_{\ ;\beta} = F^{\alpha\beta} j_{\beta}\xi_{\alpha}$. Note that $f^{\alpha} \equiv F^{\alpha\beta} j_{\beta}$ is the Lorentz force density. It follows (see Fig. 7b)

$$\Delta P^{\boldsymbol{\xi}} = \int_{V_{Left}} A^{\beta}_{\ ;\beta} dV - \int_{V_{Right}} A^{\beta}_{\ ;\beta} dV = \xi_{\alpha} \left(V_{Left} \langle f^{\alpha} \rangle_{Left} - V_{Right} \langle f^{\alpha} \rangle_{Right} \right) .$$

We have $V_{Left} = V^{(3)}_{Left} b_{Left}/2$, $V_{Right} = V^{(3)}_{Right} b_{Right}/2$ (where $V^{(3)}$ denote 3-volumes orthogonal to u^{α}). Herein we allow z^{α} to be any point within the worldtube of centroids; it follows that

Electromagnetic forces and torques defined in this way first arise at dipole order. The (monopole) Lorentz force depends—unlike any other electromagnetic terms—on $\dot{z}_s^a$. It has therefore been separated out explicitly in (192). There is no similarly velocity-dependent electromagnetic torque.

Defining the force to exclude the Lorentz force has the advantage that if ψ^a is Killing and preserves $\hat{F}_{ab}$, the associated conservation law implies that

$$\hat{F}_a \psi^a + \frac{1}{2} \hat{N}_{ab} \nabla^a \psi^b = 0. \tag{195}$$

This does not involve q, and is directly analogous to the scalar result (106).

4.3.5 Center of Mass

The electromagnetic laws of motion (192)–(194) depend on both the foliation $\{\mathfrak{B}_s\}$ and the worldline $\mathcal{Z}$. Center of mass conditions may be used to fix these structures as described in Sect. 3.8. An evolution equation for the center of mass position γ_s can then be obtained by differentiating $\hat{p}_a \hat{S}^{ab} = 0$. The result differs slightly from (111) due to the additional velocity-dependence associated with the Lorentz force. Adapting the methods of [31], the electromagnetic momentum-velocity relation may be shown to be

$$\hat{m}\dot{\gamma}_s^a = \hat{p}^a - \hat{N}^a{}_b u^b - \frac{\hat{S}^{ab}[\hat{m}\hat{F}_b + (\hat{p}^c - \hat{N}^c{}_h u^h)(q\hat{F}_{bc} - \frac{1}{2}R_{bcdf}\hat{S}^{df})]}{\hat{m}^2 - \frac{1}{2}\hat{S}^{pq}(q\hat{F}_{pq} - \frac{1}{2}R_{pqrs}\hat{S}^{rs})} \tag{196}$$

when s has been chosen such that $\dot{\gamma}_s^a \hat{p}_a = -\hat{m}$ and u^a is the unit timelike vector satisfying $\hat{p}^a = \hat{m} u^a$.

Although a precise set of assumptions which imply this are not known, it is assumed here that the center of mass condition (111) admits a unique timelike solution in a sufficiently broad class of physical systems. Inspection of (196) shows that a breakdown of the center of mass condition must occur whenever

$$\hat{m}^2 - \frac{1}{2}\hat{S}^{bc}\left(q\hat{F}_{bc} - \frac{1}{2}R_{bcdf}\hat{S}^{df}\right) = 0. \tag{197}$$

This is sufficient but not necessary. While the $q = 0$ case was dismissed in Sect. 3.8 as unlikely to be physically relevant, the charged case is potentially more interesting. Consider an electron[24] in a magnetic field of order B. Setting $q = e$, $\hat{m} = m_e$, $R_{abc}{}^d = 0$, and $\hat{S} = \hbar/2$, the denominator in (196) can diverge when $B \sim 2m_e^2/e\hbar \sim 10^{14}$ Gauss. This may be viewed as the field strength at which an electron's dipole energy $\hbar B/2$ becomes comparable to its rest mass. Quantum mechanically, it is also the field strength at which the separation between (Landau) energy levels becomes comparable to the rest mass energy. Indeed, this is the scale at which quantum electrodynamics is expected to become dominant. Although systems with 10^{14} Gauss magnetic fields are far from direct laboratory experience, such fields are believed to exist around some neutron stars [45]. Even in somewhat smaller magnetic fields, hidden momentum effects predicted by the classical theory can become very large. Whether or not this has qualitative consequences for neutron star astrophysics is an open question.

---

[24] It is unclear that there is any sense in which an electron's behavior can be modeled using equations derived for classical extended charges. Nevertheless, the example appears to be suggestive.

If the center of mass can be defined and the classical laws of physics remain valid, (196) may be combined with (192)–(194) to yield very general laws of motion. As in the scalar case, $\hat{S}^{ab}$ can also be reduced to a single spin vector $\hat{S}^a$ satisfying (113). Similarly, the mass varies according to (114). Unlike in the scalar case, matter coupled to electromagnetic fields can change mass only at dipole and higher orders. Such effects are related to changes in a body's internal energy due to work performed by (or against) the ambient fields [24].

4.3.6 Monopole Approximation

In simple cases, the laws of motion governing an extended charge distribution may be truncated at monopole order. Within this approximation, $\hat{F}_a = \hat{N}_{ab} = 0$ and (192) reduces to

$$\frac{\mathrm{D}\hat{p}^a}{\mathrm{d}s} = q\hat{F}^a{}_b\dot{\gamma}_s^b + \frac{1}{2}R_{bcd}{}^a\hat{S}^{bc}\dot{\gamma}_s^d, \qquad \frac{\mathrm{D}\hat{S}_{ab}}{\mathrm{d}s} = 2\hat{p}^{[a}\dot{\gamma}_s^{b]}. \tag{198}$$

Contracting the second of these equations with $\hat{S}_{ab}$ while using (110) again shows that $\hat{S}^2 := \hat{S}^a\hat{S}_a =$ (constant). An object which is initially not spinning therefore remains non-spinning in this approximation. Consider these cases for simplicity. The momentum-velocity relation then reduces to $\hat{p}^a = \hat{m}\dot{\gamma}_s^a$, the mass remains constant, and the body accelerates via the Lorentz force law

$$\hat{m}\ddot{\gamma}_s^a = q\hat{F}^a{}_b\dot{\gamma}_s^b \tag{199}$$

in the effective electromagnetic field $\hat{F}_{ab}$. The effective field always satisfies the vacuum Maxwell equations. It is generically distinct from the physical field F_{ab} which governs the acceleration of nearby test charges.

In many cases of interest, it is useful to model F_{ab} as the sum of some external field F_{ab}^{ext} and the retarded field associated with a body's charge distribution:

$$F_{ab} = F_{ab}^{\mathrm{ext}} + 2\int_{\mathfrak{W}} \nabla_{[a}G^{-}_{b]b'}J^{b'}\mathrm{d}V'. \tag{200}$$

(% $

a. Curved spacetime

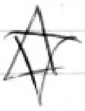

b. flat spacetime

Pág. 382 (390) 4.3.5

Center of Mass

Obs: Usar isto para calcular o centro de massa que gera a curvatura do espaço-tempo.

Obs: Lorentz force.

— Geometrical theory of Satellite (2018).pdf

Pág. 264 (275) 18.5

Environmental Effects on the Galileo Clock Parameters

⇒ (gauss equal 10^{-4} tesla)

— Quantum Gravity - World Scientific Publishing ..(1997).pdf

⇒ Gauss map

— ~~Overcoming Gravity...(2016).pdf~~
(error)

-256- - -* ! - 7 ! 9 3 ! * D
! " 7 !A 3 < " 1 & ! ! D 1"
(B' $ $ = A ! = 7! . 1! ! 7! > ! X '
!

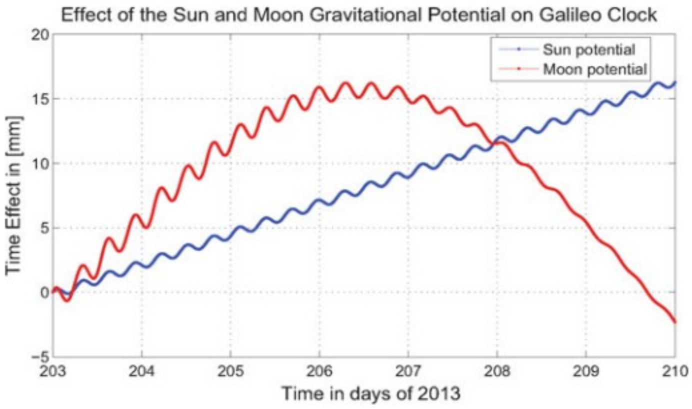

Fig. 18.9 Accumulated time along the orbit of the Galileo E11 satellite due to the gravitational potential of Sun and Moon for a selected period of 6 days with low Sun elevation $4^\circ > \beta > -2^\circ$ above the orbital plane. When the Sun is in the Galileo orbital plane, the satellite orbit spans about 2×29600 km per orbit revolution in the Sun’s gravitational field (max.-min. distance to the Sun). The maximum is reached for a Moon elevation of 28.5°

18.5 Environmental Effects on the Galileo Clock Parameters

In this section, we discuss the impact of the in-orbit environment (magnetic field and temperature variations) on the Galileo clock performance. Magnetic field perturbations can be estimated by using the magnetic sensitivity coefficient of $<3 \times 10^{-13}$/G (one gauss equals 10^{-4} tesla) in fractional frequency, as measured during ground tests (Boving et al. 2009) and (Rochat et al. 2012). Magnetic filed variations along the Galileo orbit were calculated by using the International Geomagnetic Reference Field (IGRF) model (International Association of Geomagnetism and Aeronomy et al. 2010) in the direction of the satellite X, Y and Z axes. The time accumulated along the orbit was obtained by integrating the fractional frequency variations due to the ambient magnetic field (see Fig. 18.11). Considering that the magnetic field is in the order of 300–550 nT along the Galileo orbit (days 100–116 in 2013), the contribution of magnetic perturbations to the estimated residual clock parameters is in the order of several millimeters. However, assuming the orientation of the Galileo maser cavity along the satellite X-axis (that never faces the Sun), the maximum effect of the magnetic field is below 0.8 mm, see Fig. 18.11. When applied as a correction, the standard deviation of the residual clock parameters in Fig. 18.7 was improved by only 0.1 mm. We can therefore conclude that the impact of magnetic field variations on the Galileo clock parameters is very small and negligible. In addition, shielding of the satellite further reduces their effect. However, this would not be the case if the same clock were placed in a LEO orbit, where the magnetic field strength is higher by two orders of magnitude.

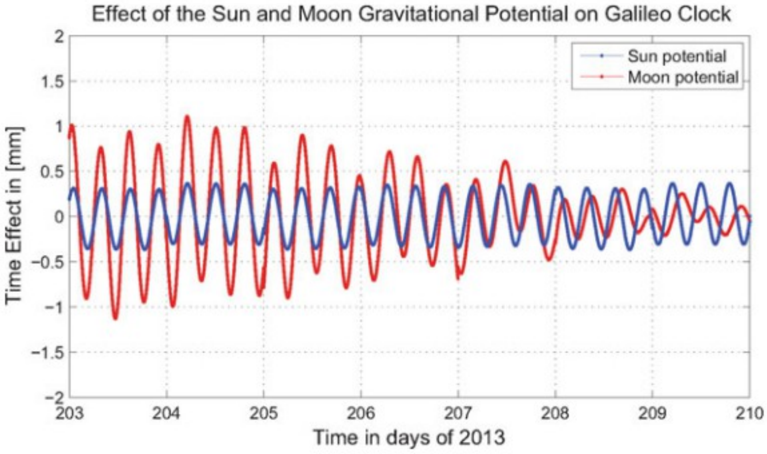

Fig. 18.10 Accumulated time along the orbit of the Galileo E11 satellite due to the gravitational potential of Sun and Moon after removing daily time offset and drift. One can clearly see a distinct twice per revolution effect for the Sun potential. After removing daily time offset and time drift, the remaining effect on the residual clock parameters is below 0.4 mm for the Sun and up to 1 mm for the Moon potential

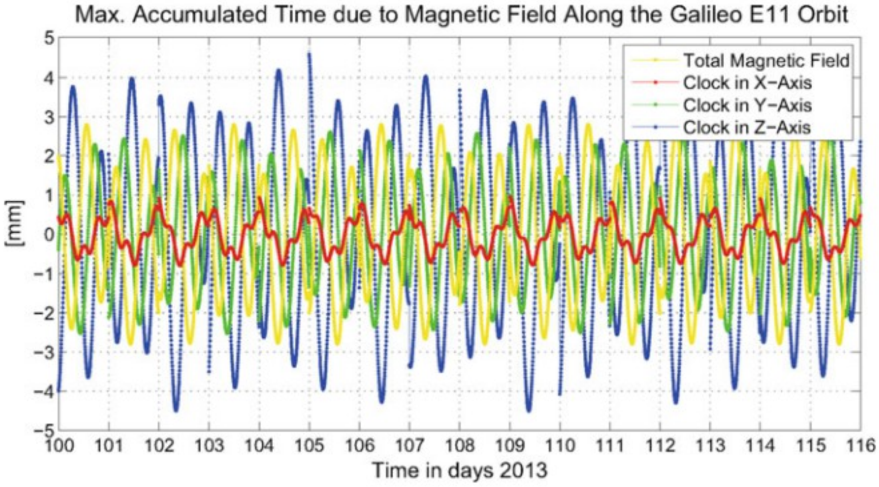

Fig. 18.11 Maximum accumulated time in (mm) along the orbit of Galileo E11 satellite assuming a maximum magnetic sensitivity coefficient of $3 \times 10^{-13}/\text{Gauss}$. The magnetic field along the X, Y and Z satellite axes was calculated using the IGRF model giving a magnetic field variation of 300–550 nT along the orbit. Assuming the orientation of the H-maser cavity along the X satellite axis (never faces Sun), the maximum effect of the magnetic field on residual clock parameters is below 0.8 mm. Along the Z axis, the effect is about three times higher. Daily bias and drift were removed

Unfortunately, no public data is available on the in-orbit temperature at the clock reference point, therefore not much can be said about thermal perturbations. Boving et al. (2009) reported a thermal sensitivity coefficient of the Galileo H-maser as measured on the ground of $\leq 2 \times 10^{-14}/\,^{\circ}\mathrm{C}$. The cavity temperature of the Galileo H-maser is stabilized by a two-stage thermal control and an additional electronic Automatic Cavity Tuning (ACT) system is used to optimize the cavity frequency pulling effect caused by the residual thermal drift (Mattioni et al. 2002). From Mattioni et al. (2002), one can see that, for platform temperature variations of 5 °C, the cavity thermal control stabilizes the temperature within 3 m°C.

Temperature variations at the PHM reference point on-board the Galileo satellite are expected to have two different periods, the orbital period and the period of the stabilization loop. An analysis of the Allan deviation and PSD curves in Figs. 18.7 and 18.8, reveals no perturbation at the orbital period nor for periods shorter than 1000 s that could be attributed to temperature effects. This allows us to conclude that the temperature stability at the PHM reference point is at the level of a few tenths of a Kelvin.

References

Beutler G, Brockmann E, Gurtner W et al (1994) Extended orbit modeling techniques at the CODE processing center of the international GPS service for geodynamics (IGS): theory and initial results. Manuscr Geod 19:367–386

Boving D, Droz F, Mosset P et al (2009) Space passive hydrogen maser, performances, lifetime data and GIOVE-B related telemetries. In: IGNSS2009 conference proceedings

ESA Portal (2014) Galileo facts and figures. http://www.esa.int/Our_Activities/Navigation

International Association of Geomagnetism and Aeronomy WGV-MP members, Finlay CC, Maus S, et al (2010) International geomagnetic reference field: the eleventh generation. Geophys J Int 183:1216–1230. https://doi.org/10.1111/j.1365-246x.2010.04804.x

Kouba J (2004) Improved relativistic transformations in GPS. GPS Solutions 8:170–180. https://doi.org/10.1007/s10291-004-0102-x

Mattioni L, Belloni M, Berthoud P et al (2002) The development of a passive hydrogen maser cclock for the Galileo navigation system. In proceeding of: 34th annual precise time and time interval (PTTI) meeting, pp 161–170

Montenbruck O (2013) MultiGNSS working group. In: International GNSS service technical report 2012. IGS Central Bureau, JPL, Pasadena

Montenbruck O, Steigenberger P, Hugentobler U (2014) Enhanced solar radiation pressure modeling for Galileo satellites. J Geod 1–15. https://doi.org/10.1007/s00190-014-0774-0

Petit G, Luzum B (2010) IERS Conventions. Verlag des Bundesamts für Kartographie und Geodäsie

Riley W (2014) The Stable32 program for frequency stability analysis. Hamilton Technical Services, Beaufort, SC, p 29907

Rochat P, Droz F, Wang Q, Froidevaux S (2012) Atomic clocks and timing systems in global navigation satellite systems

Steigenberger P, Hugentobler U, Loyer S et al (2014) Galileo orbit and clock quality of the IGS Multi-GNSS experiment. Advances in space research. https://doi.org/10.1016/j.asr.2014.06.030

Svehla D (2010a) Complete relativistic modelling of the GIOVE-B clock parameters and its impact on POD, track-to-track ambiguity resolution and precise timing applications. In: IGS Workshop 2010. Newcastle, UK

Svehla D, Navarro-Reyes D (2011) First laser measurements of Europe's Galileo satellites made from Chile. ESA press release (Navigation): http://www.esa.int/Our_Activities/Operations/First_laser_measurements_of_Europe_s_Galileo_satellites_made_from_Chile

Svehla D, Cacciapuoti L, Rothacher M (2015) Noise model of the Galileo "mm-Clock"—geometrical mapping of the orbit perturbations using a clock on board Galileo satellites. IUGG General Assembly 2015

Svehla D, Rothacher M, Cacciapuoti L (2016) Thermal re-radiation acceleration in the GNSS orbit modelling based on galileo clock parameters. In: IGS workshop 2016, Sydney, Australia

Svehla D, Cacciapuoti L, Petit G, Rothacher M (2017) How stable is the primary clock onboard galileo satellites? First geometrical mapping of the orbit perturbations using a clock in space. Submitted to journal of geodesy (reviewed)

Waller P, Gonzalez F, Binda S et al (2009) Update on the In-Orbit performances of GIOVE clocks. In: IEEE international frequency control symposium, 2009 Joint with the 22nd EFTF, pp 388–392

Wang Q, Mosset P, Droz F, Rochat P (2013) Lifetime of space passive hydrogen maser. In: (EFTF/IFC), 2013, pp 973–976

Wolf P, Petit G (1995) Relativistic theory for clock syntonization and the realization of geocentric coordinate times. Astron Astrophys 304:653–661

D -256- - -* ! - 6 ! 7 !A 3 @ "
D ? " // 1"
E 7! !1

data

- Principles of Astrophysics (2014).pdf
- Beyond Spacetime - the Foundations of Quantum Gravity (2020)).pdf

Pag. 90(99) Gaussianity of the spectrum

- Baggott - Quantum Space (2018).pdf

Pag. 36(57)

Einstein's friend Marcel Grossman directed him to the work of the great German mathematician Carl Friedrich Gauss and his student, Bernhard Riemann.

Einstein was introduced to the Riemann curvature tensor, the definitive mathematical expression for the curvature of a generalized geometry in any set of coordinates.

- Einstein's general theory of relativity (2010).pdf

Pag. 8(26) 1.5

Local form of Newton's Gravitational Law

%
! -256- - -* ! - % 1 2 1 3 H
7 !A 3 ! " ! 3 =L1 1 ' 1"

D -256- - -* ! - %)! J C 6 ! 1! ? 1
> ! !A 3 ! " ! 1! J J ! " A *L " H A 3
1"
(/ // 7! ! 3 1

corresponds to an angular scale of about 1° and we need to explain the absence of order one anisotropies on the angular scales of the full sky). Thus the first requirement on a successful early universe scenario is that the horizon at late times be much larger than the Hubble radius.

In order to be able to have a causal mechanism of structure formation, scales that are currently observed in cosmology must originate in the early universe with a wavelength smaller than the Hubble radius. This is the second requirement.

Third, the scales that we observe today must propagate for a long time at super-Hubble lengths. This is in order that the fluctuations are squeezed and enter the Hubble radius at late times as standing waves. This is required in order to provide an explanation for the observed acoustic oscillations in the angular power spectrum of the CMB.

Finally, the generation mechanism that acts on sub-Hubble scales must yield a nearly scale-invariant spectrum of cosmological perturbations.

The inflationary universe scenario provides a simple realization of these four criteria. However, there are alternative realizations as explained in the following section.

4.3 Paradigms of Early Universe Cosmology

4.3.1 Cosmological Inflation

Cosmological inflation (Brout, Englert, and Gunzig, 1978; Fang, 1980; Guth, 1981; Sato, 1981) assumes that space underwent a period of almost exponential expansion during a finite time interval in the very early universe, starting at some time t_i and ending at a time t_R.[4] The first model of inflation was based on a modified gravitational action (Starobinsky, 1980) where higher curvature terms dominate at early times and lead to accelerated expansion. Later models assumed that the accelerated expansion of space is realized in the context of Einstein gravity, but assuming the presence of a slowly rolling (Albrecht and Steinhardt, 1982; Linde, 1982, 1983) scalar field whose energy-momentum tensor is dominated by the scalar field potential energy. After the end of inflation at time t_R, the evolution of the universe is like in SBB cosmology.

Figure 4.5 represents a spacetime sketch of inflationary cosmology. The vertical axis is time, the horizontal axis denotes physical spatial distance. The solid curve labeled H^{-1} represents the Hubble radius, the dashed line is the horizon, and the curve labeled k represents the physical wavelength of a fluctuation mode.

During the phase of accelerated expansion for $[t_i < t < t_R]$, the horizon is growing very fast while the Hubble radius increases only slowly. For almost exponential

---

[4] The subscript R stands for *reheating* since at this time the conditions of a hot early universe must be created.

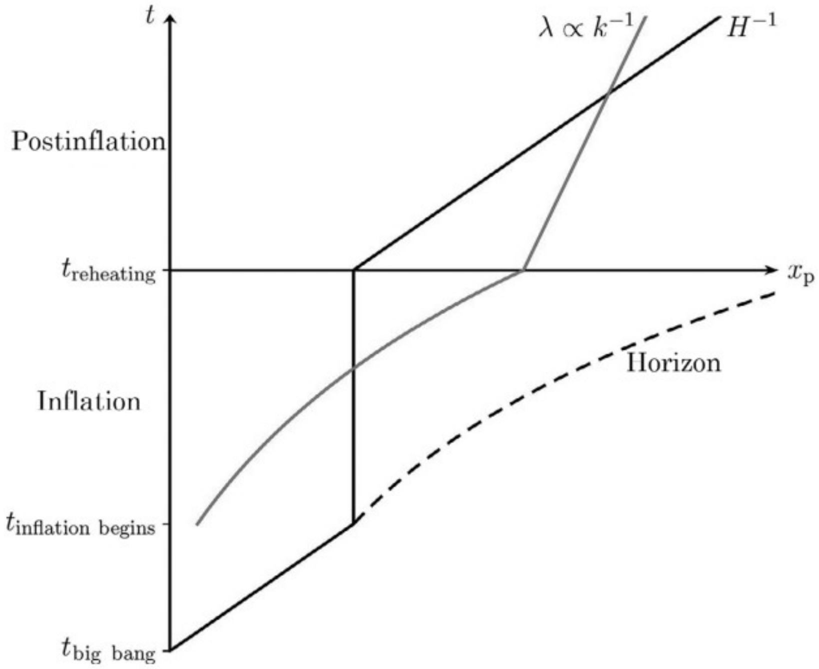

Figure 4.5 Space-time sketch of inflationary cosmology. The vertical axis is time, the horizontal axis is physical distance. The curve labeled H^{-1} is the Hubble radius, and the curve labeled k is the physical wavelength of the fluctuation on a particular (comoving) scale. Note that during the period inflation both the wavelength of a fluctuation and the horizon grow exponentially (for exponential inflation), which is indicated by the curvature of the lines. After the end of inflation, the wavelength grows as $a(t) \sim t^{1/2}$ which is indicated by the near linearity of the curve.

expansion the horizon increases nearly exponentially while the Hubble radius is almost constant. Only a very short interval of accelerated expansion is needed to solve the horizon problem (the horizon needs to increase by a factor of greater than 100 relative to the Hubble radius).

If the energy scale of inflation corresponds to the scale of particle physics Grand Unification (an energy scale of about 10^{16} GeV), then about 60 e-foldings of exponential expansion are required in order that scales that we observe now in cosmology had a wavelength smaller than the Hubble radius at the beginning of the inflationary period. In this case, the second of the general criteria for a successful early universe cosmology is satisfied.

As is evident from Figure 4.5, the wavelength of fluctuations is larger than the Hubble radius for a long time. Hence the squeezing of the fluctuations that is required in order to obtain the acoustic oscillations in the CMB angular power spectrum is realized.

Finally, as initially argued in Starobinsky (1979) for gravitational waves and in Mukhanov and Chibisov (1981) for the curvature perturbations, the power spectrum of fluctuations will be approximately invariant of scale. This is a consequence of the time-translation invariance of the phase of exponential expansion (Press, 1980; Sato, 1981). It is usually assumed that fluctuations originate as quantum vacuum perturbations at early times during the inflationary phase. This assumption can be justified since, in the absence of interactions with the agent driving the accelerated expansion, any initial fluctuations are redshifted away, leaving the matter in a vacuum state. If the agent driving inflation interacts with matter, it is possible to realize inflation while maintaining the dominance of thermal fluctuations. This is the *warm inflation* scenario (Berera, 1995).

The scale invariance of the spectrum of curvature fluctuations can also be seen using the general formalism described in the previous section. The curvature fluctuation variable, starting out with a deep blue vacuum spectrum on sub-Hubble scales, gets squeezed on super-Hubble scales. Large wavelength fluctuations exit the Hubble radius earlier and are thus squeezed more. For almost exponential inflation the squeezing is just right to turn the initial vacuum spectrum into a scale-invariant one. To see this, consider first the Hubble crossing condition:

$$a^{-1}(t_H t(k))k = H. \tag{4.11}$$

Before $t_H(k)$, the amplitude of v_k is constant, afterwards it increases as z. Thus the power spectrum of ζ at some late time t is:

$$\begin{aligned} P_\zeta(k,\eta) &= k^3 z^{-2}(\eta)\left(\frac{z(\eta)}{z((\eta_H(k))}\right)^2 |v_k(t_i)|^2 \\ &\simeq \frac{1}{2}\left(\frac{a(t_H(k))}{z(t_H(k))}\right)^2 H^2. \end{aligned} \tag{4.12}$$

The k-dependence has canceled out, and we thus obtain a scale-invariant spectrum.

Gravitational waves evolve as curvature fluctuations, with the function $z(\eta)$ being replaced by $a(\eta)$. Hence inflationary cosmology predicts a scale invariant spectrum of gravitational waves whose amplitude is suppressed compared to that of curvature fluctuations by the slow-roll parameter ϵ, which gives the ratio between z and a during the inflationary phase.

Of the early universe scenarios discussed here, inflationary cosmology is the only one that predicted (as opposed to postdicted) various observational results, e.g., the spatial flatness of the universe, the almost scale invariance of the spectrum of cosmological perturbations, and the near Gaussianity of this spectrum. In fact, inflation predicted (Mukhanov and Chibisov, 1981) a slight red tilt of the spectrum, a tilt that has now been established by the recent CMB observations (see, e.g., Ade et al. [2014]). Inflation predicts a nearly scale-invariant spectrum of gravitational waves (Starobinsky, 1979) with a slight red tilt. Note that, whereas the tilt of the spectrum of cosmological perturbations can be made blue by complicating the scalar field model of inflation, the spectrum of gravitational waves has a red tilt unless the matter that is responsible for inflation violates standard energy conditions.[5]

A drawback of the inflationary scenario is the *trans-Planckian problem* for cosmological perturbations (Brandenberger and Martin, 2001; Martin and Brandenberger, 2001). (See e.g., Brandenberger and Martin [2013] for a review with an extended list of references.) If the period of inflation lasts only slightly longer than the minimal amount of time it needs to in order to solve the flatness and structure formation problems of SBB cosmology, then the wavelengths of the fluctuation modes that we observe today are smaller than the Planck length at the beginning of inflation. We do not understand the physics on these wavelengths, and hence it is unclear if the initial conditions for the fluctuations that are used are well justified. Some other problems of inflationary cosmology are discussed in later sections.

4.3.2 Matter Bounce

Another paradigm to obtain successful structure formation is the *matter bounce* scenario (Wands, 1999; Finelli and Brandenberger, 2002). It is assumed that there is new physics[6] that resolves the cosmological singularity. Time runs from $-\infty$ to ∞. The universe begins in a contracting phase that is the mirror inverse of the expanding standard big bang cosmology evolution, which is to say that at very early times during contraction, a dark energy component may dominate, followed by a period of matter-dominated contraction, followed by a radiation-dominated phase, and then a nonsingular bounce. The bounce point (minimal value of the scale factor) is $t = 0$.

Figure 4.6 presents a sketch of the resulting spacetime diagram. The vertical axis is time, the horizontal axis represents comoving spatial distance. The Hubble radius is symmetric about $t = 0$. The perpendicular curve indicates the wavelength of a fluctuation. Since there is no origin of time, the horizon is infinite and there is no horizon problem. As obvious from the figure, all scales originate inside the Hubble

---

[5] At this point, we have not yet observed a stochastic background of gravitational waves, and observing the small tilts predicted by various early universe models will be very challenging.

[6] The new physics could be string theory, loop quantum gravity, or some effective field theory that corresponds to Einstein gravity coupled to some matter that violated the usual Hawking–Penrose energy conditions. For a review of bouncing cosmologies see, e.g., Brandenberger and Peter (2017). I am emphasizing here ways to obtain a spectrum of cosmological fluctuations in agreement with observations, and the specific dynamics of the bounce phase has little to say about this issue.

-256- - -* ! - %)! J C 6 ! 1! ? 1
> ! !A 3 ! " ! 1! J J ! " A *L " H A 3
1"
(%B $ = 0 " ! 7 !

It's a common misconception that Einstein the genius physicist must also have been a great mathematician. He was not. Responding to his cry for help, his friend Marcel Grossman directed him to the work of the great German mathematician Carl Friedrich Gauss and his student, Bernhard Riemann. Einstein was introduced to the Riemann curvature tensor, the definitive mathematical expression for the curvature of a generalized geometry, in any set of coordinates. This was the connection that would later bring the sixteen-year-old Smolin from architecture to general relativity.

But the Riemann tensor didn't quite fit the bill, and as Einstein struggled to find a way to generalize it, he got 'lost' in the algebra, for two years making mistakes and following alleys that each in turn proved to be blind.

He eventually found the generalization required to make the field equations conform satisfactorily to the demands of the physics. The result is a set of equations which connect the curvature of spacetime—on the left-hand side—with the distribution and flow of mass-energy on the right-hand side. He presented his equations to the Prussian Academy of Sciences in Berlin in a series of four lectures which concluded on 25 November 1915.

From the beginning of his journey to the gravitational field equations of general relativity, Einstein was aware of four potential empirical tests.

Actually, the first is not so much a test, more the resolution of a mystery. We've known since the work of Johannes Kepler in the seventeenth century that a planet moves in an elliptical orbit around the Sun. But the planetary orbits are not exact ellipses. If they were, each planet's point of closest approach to the Sun (which is called the *perihelion*) would be fixed, the planet always passing through the same point with each and every orbit. But astronomical observations show that with each orbit the perihelion shifts slightly, or *precesses*.

Much of the observed precession is caused by the cumulative gravitational pull of all the other planets in the Solar System, effects which can be predicted using Newton's law. For the planet Mercury, lying closest to the Sun, this 'Newtonian precession' is predicted to be 532 arc-seconds per century.* However, the observed precession is rather more, about 574 arc-seconds per century, a difference of 42 arc-seconds. Though small, this difference accumulates and is equivalent to one 'extra' orbit every three million years or so.

Newton's gravity can't account for this small difference and other explanations—such as the existence of another planet, closer to the Sun than Mercury (named Vulcan)—were suggested. Astronomers searched for it in vain. Einstein was delighted to discover that the field equations of general relativity predict a further 'relativistic' contribution of 43 arc-seconds per century, due to the curvature of spacetime around the Sun in the vicinity of Mercury.† This discovery gave Einstein the strongest emotional experience of his life in science: 'I was beside myself with joy and excitement for days.'[10]

Perhaps the most famous prediction of general relativity concerns the bending of starlight passing close to the Sun. Once again, it's not the 'bending' per se that represents a novel prediction, but rather its extent. Newton's gravity suggests that light grazing the surface of the Sun should bend through 0.85 arc-seconds.‡ But the curvature of spacetime predicted by general

---

* A full circle is 360°, and an arc-minute is one-sixtieth of 1°. An arc-second is then one-sixtieth of an arc-minute. So 532 arc-seconds represents about 0.15 of a degree.

† The perihelia of other planets are also susceptible to precession caused by the curvature of spacetime, but as these planets are further away from the Sun the contributions are much less pronounced.

‡ Newton favoured a 'corpuscular' theory in which light is imagined to consist of individual particles. It's therefore not difficult to imagine why such 'bullets' of light might be affected by the gravitational pull of a large object such as a star.

Einstein was well versed in the geometries of things that curve. We know well enough that the Earth is finite—it has a mass that can be estimated (it's about six million billion billion kilograms). And, although from any human vantage point on its surface the Earth looks flat, we know that it is spherical, and so has no edge. The Earth is finite but unbounded.

So, in Einstein's universe spacetime curves back on itself like the surface of a sphere, and contains a finite amount of mass-energy. Note that the 'sphere' in question is a sphere in four spacetime dimensions (what mathematicians refer to as a 3-sphere or, more generally, a hypersphere—see Figure 11). The surface of

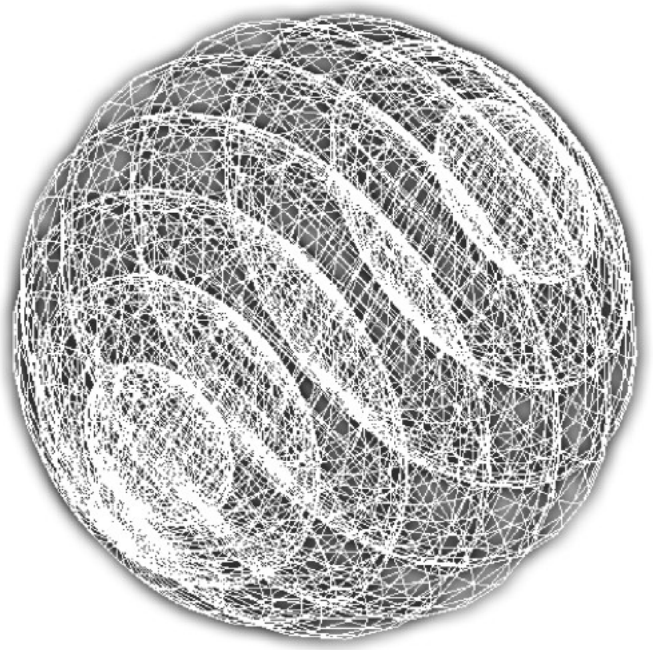

Figure 11. This representation of a 3-sphere (or hypersphere) is a projection into three-dimensional space, showing the structure as a stack of regular spheres (2-spheres), which is then projected onto two dimensions.

the Earth, in contrast, is a 2-sphere, a sphere in three-dimensional space. If it were possible for us to set off together across Einstein's universe on a journey along a straight path, we would eventually find ourselves returning to our point of departure.

Curiously, this was not the first example of a finite, but unbounded universe. Carlo Rovelli was delighted to discover that Einstein had been anticipated by the great Italian poet Dante Alighieri, who had studied at the University of Bologna in the fourteenth century. In *Paradiso*, the third and final part of the *Divine Comedy*, Dante describes his ascent to the Empyrean, the abode of God and the angels. Confronted with the challenge of describing the universe at its very edge, he fixes on the same solution that Einstein would choose, five hundred years later: 'Dante embeds the model in four dimensions, which does, as we know, solve the problem.'[2]

Einstein now had to deal with a more stubborn second problem. Gravity pulls objects together but it doesn't push them apart. Unlike electricity or magnetism, gravity doesn't have opposite charges or opposite poles, which attract *and* repel. Put another way, mass-energy curves spacetime only in ways which cause objects to move towards each other, never away.

Newton well understood where this leads. The mutual gravitational attraction between all the objects in an infinite universe would cause the universe to collapse in on itself. He had no choice but to invoke the ultimate solution, arguing in the *Mathematical Principles* that God had placed the stars sufficiently far apart to prevent such a catastrophic collapse from happening.[3]

Several centuries of astronomy had certainly produced no evidence that all the stars in the universe are rushing towards each other and in 1917 plain common sense suggested that the universe is surely *static*, not dynamic—we don't call them 'fixed stars' for nothing. But, just as in Newton's gravity, there was nothing in Einstein's field equations that would help to hold the universe steady. Einstein needed a physical mechanism that would do much the same job that Newton had asked of God.

His solution was to introduce a rather unsatisfactory 'fudge'. The left-hand side of the gravitational field equations describes the extent of spacetime curvature that will determine the motions of all the mass-energy, which is summarized on the right-hand side. Despite their extraordinary success, Einstein concluded that his equations were still 'unbalanced'. To engineer a static universe, he attempted to balance the equations by introducing a new term on the left-hand side, imbuing spacetime with a curious kind of antigravitational force, a negative gravitational pressure which builds in strength over long distances, counteracting the effects of the curvature caused by all the mass-energy on the right-hand side.

The new term is now characterized by a parameter Λ (Greek lambda), which is referred to as the 'cosmological constant'. By carefully selecting the value of Λ, Einstein found that he could achieve perfect balance: a static universe. He admitted that he 'had to introduce an extension to the field equations that is not justified by our actual knowledge of gravitation'.[4] But at the time it must have seemed like a neat solution. The cosmological constant didn't alter the way general relativity works over short distances, so all the theory's successful predictions were preserved.

However, Einstein had also made some significant errors. Not only was there no evidence or even a real theoretical justification for the cosmological constant, it actually doesn't do what Einstein claimed. The universe that he described in his 1917 paper is actually not static and is instead rather unstable. The modified field equations still didn't completely exclude the possibility that the universe might contract or expand.

Fourteen years later, Einstein's universe was overtaken by events.

" -256- - -* ! - % = 0 ! 3 ! A 3
1 1"
(B $? ! 8 0 7 !A ! ! ?!

A point mass M at a position $\mathbf{r}' = \mathbf{r}_0$ represents a mass density

$$\rho(\mathbf{r}') = M\delta(\mathbf{r}' - \mathbf{r}_0). \tag{1.20}$$

Substitution into Eq. (1.17) gives the potential of the point mass

$$\boxed{\phi(\mathbf{r}) = -\frac{GM}{|\mathbf{r} - \mathbf{r}_0|}.} \tag{1.21}$$

1.5 Local form of Newton's Gravitational law

Newton's law of gravitation cannot be a relativistically correct law, because it permits action at a distance. A point mass at one place may then act instantaneously on a point mass at another remote position. According to the special theory of relativity, instantaneous action at a distance is impossible. An action which is instantaneous in one reference frame, is not instantaneous in another frame, moving with respect to the first. This is due to the relativity of simultaneity (see Ch. 2). *Instantaneous action at a distance can only exist in a theory with absolute simultaneity.* As a first step towards a relativistically valid theory of gravitation, we shall give a local form of Newton's law of gravitation.

We shall now show how Newton's law of gravitation leads to a field equation for gravity. Consider a continuous mass-distribution $\rho(\mathbf{r}')$. Equations (1.16) and (1.17) lead to

$$\boldsymbol{\nabla}\phi(\mathbf{r}) = G\int \rho(\mathbf{r}')\frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3}d^3r', \tag{1.22}$$

which gives

$$\nabla^2\phi(\mathbf{r}) = G\int \rho(\mathbf{r}')\boldsymbol{\nabla}\cdot\frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3}d^3r'. \tag{1.23}$$

Furthermore

$$\begin{aligned}\boldsymbol{\nabla}\cdot\frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} &= \frac{\boldsymbol{\nabla}\cdot\mathbf{r}}{|\mathbf{r} - \mathbf{r}'|^3} + (\mathbf{r} - \mathbf{r}')\cdot\boldsymbol{\nabla}\frac{1}{|\mathbf{r} - \mathbf{r}'|^3}\\ &= \frac{3}{|\mathbf{r} - \mathbf{r}'|^3} + (\mathbf{r} - \mathbf{r}')\cdot\left(-3\frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^5}\right) = 0, \quad \mathbf{r} \neq \mathbf{r}'.\end{aligned} \tag{1.24}$$

In general the volume of integration encompasses the point $\mathbf{r}' = \mathbf{r}$ where the field is measured. Thus, we have to find an expression for $\boldsymbol{\nabla}\cdot(\mathbf{r} - \mathbf{r}')/|\mathbf{r} - \mathbf{r}'|^3$ which is also valid at this point. Equation (1.24) indicates that $\boldsymbol{\nabla}\cdot(\mathbf{r} - \mathbf{r}')/|\mathbf{r} - \mathbf{r}'|^3$ is proportional to Dirac's δ-function. According to Eq. (1.19) the proportionality factor can be found by calculating the integral $\int \boldsymbol{\nabla}\cdot(\mathbf{r} - \mathbf{r}')/|\mathbf{r} - \mathbf{r}'|^3 d^3r'$. Using Gauss' integral theorem

$$\int_V \boldsymbol{\nabla}\cdot\mathbf{A}d^3r' = \oint_S \mathbf{A}\cdot d\mathbf{S}' \tag{1.25}$$

where S is the surface enclosing V, we get

$$\int_V \boldsymbol{\nabla}\cdot\frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3}d^3r' = \oint_S \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3}\cdot d\mathbf{S}'. \tag{1.26}$$

Note that the gradient $\nabla\phi$ in Eq. (1.22) is directed away from the source. Thus the divergence of this vector in Eq. (1.23) must be positive. The direction of the surface element $d\mathbf{S}'$ in Fig. 1.3 is chosen to satisfy this criterion.

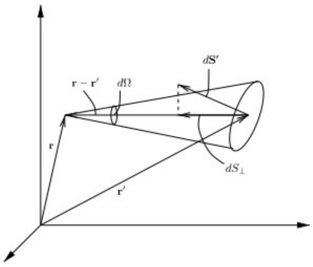

Figure 1.3: Definition of surface elements

With reference to Fig. 1.3 the solid angle element $d\Omega$ is defined by

$$d\Omega = -\frac{dS_\perp}{|\mathbf{r}-\mathbf{r}'|^2} \tag{1.27}$$

where

$$dS_\perp = \frac{\mathbf{r}-\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|}\cdot d\mathbf{S}'. \tag{1.28}$$

It follows that

$$d\Omega = -\frac{\mathbf{r}-\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|^3}\cdot d\mathbf{S}' \tag{1.29}$$

and

$$\int\limits_V \boldsymbol{\nabla}\cdot\frac{\mathbf{r}-\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|^3}d^3r' = \int d\Omega = \begin{cases} 4\pi, & P \text{ inside } V \\ 0, & P \text{ outside } V \end{cases} \tag{1.30}$$

Thus we get

$$\boldsymbol{\nabla}\cdot\frac{\mathbf{r}-\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|^3} = 4\pi\delta(\mathbf{r}-\mathbf{r}'). \tag{1.31}$$

Substituting this into Eq. (1.23) and using Eq. (1.19) with $f(\mathbf{r}) = 1$ we have

$$\boxed{\nabla^2\phi(\mathbf{r}) = 4\pi G\rho(\mathbf{r}).} \tag{1.32}$$

This Poisson equation is the local form of Newton's law of gravitation. Newton's 2nd law applied to a particle falling freely in a gravitational field gives the acceleration of gravity

$$\boxed{\mathbf{g} = -\boldsymbol{\nabla}\phi.} \tag{1.33}$$

Newton's theory of gravitation can now be summarized in the following way: *Mass generates a gravitational field according to Poisson's equation, and the gravitational field generates acceleration according to Newton's second law.*

Stoke's Theorem: *Let M be a smooth n-dimensional (oriented) compact manifold with intrinsic boundary ∂M, and let $\boldsymbol{\alpha}$ be an $(n-1)$-form. Then*

$$\boxed{\int\limits_M \mathbf{d}\boldsymbol{\alpha} = \int\limits_{\partial M} \boldsymbol{\alpha}.} \tag{6.71}$$

This is the generalisation of Gauss' law for vector calculus.

In particular we note that if $\partial M = 0$ then the integral vanishes. This is why any closed loop integral over an exact one-form yields zero, a loop has no boundary.

We will not get into details in this book, but we will just see how this law can be applied to the electromagnetic case.

Consider a bounded 3-dimensional region M which is purely spatial. From Stoke's theorem and Maxwell's equations we have

$$\int\limits_M \mathbf{d}\star\mathbf{F} = \int\limits_M \star\mathbf{J} = \int\limits_{\partial M} \star\mathbf{F}. \tag{6.72}$$

Since the region is purely spatial, so will its intrinsic boundary be. Thus

$$\int\limits_M \star\mathbf{J} = \int\limits_M \rho\mathbf{dx}^1\wedge\mathbf{dx}^2\wedge\mathbf{dx}^3 \equiv Q; \tag{6.73}$$

i.e. Q is the total charge inside the spatial region. On the other hand

$$\begin{aligned}\int\limits_{\partial M} \star\mathbf{F} &= \int\limits_{\partial M}\left(E_1\mathbf{dx}^2\wedge\mathbf{dx}^3 + E_2\mathbf{dx}^3\wedge\mathbf{dx}^1 + E_3\mathbf{dx}^1\wedge\mathbf{dx}^2\right)\\ &= \int\limits_{\partial M} \mathbf{E}\cdot\mathbf{dS},\end{aligned} \tag{6.74}$$

where $\mathbf{E}\cdot\mathbf{dS}$ can be interpreted as the electric flux out of the surface element $\mathbf{dS}$. Thus in all,

$$\int\limits_{\partial M} \mathbf{E}\cdot\mathbf{dS} = Q. \tag{6.75}$$

This is the famous Gauss' law in electromagnetism.

The corresponding law for the magnetic field is

$$0 = \int\limits_M \mathbf{dF} = \int\limits_{\partial M} \mathbf{F} = \int\limits_{\partial M} \mathbf{B}\cdot\mathbf{dS} \tag{6.76}$$

and can be interpreted as the lack of magnetic monopoles in electromagnetism.

Examples

Example 6.4 (The Electromagnetic Field outside a static point charge)
Let us investigate the electromagnetic field outside a point charge.

Assume that the field is a function of the radial coordinate, r, only. A pure static electric field can be generated by a field potential $A_0 = \varphi(r)$. Using

$$F_{\mu\nu} = \frac{\partial A_\nu}{\partial x^\mu} - \frac{\partial A_\mu}{\partial x^\nu}, \tag{6.77}$$

the only non-zero components of the field tensor are

$$F_{i0} = \frac{\partial \varphi}{\partial x^i} = \frac{\partial \varphi}{\partial r}\frac{x^i}{r} \tag{6.78}$$

using the chain rule and $r = \sqrt{(x^1)^2 + (x^2)^2 + (x^3)^2}$. The electric field is now just $E_i = F_{i0}$. If M is the interior of a spherical region so that ∂M is a sphere of radius r, the area surface element $\mathbf{dS}$ has components

$$(\mathbf{dS})_i = \frac{x^i}{r} \cdot r^2 d\Omega \tag{6.79}$$

Thus Gauss's law, eq. (6.75), gives

$$Q = \int\limits_{\partial M} \mathbf{E} \cdot \mathbf{dS} = \frac{\partial \varphi}{\partial r} r^2 \int d\Omega = \frac{\partial \varphi}{\partial r} 4\pi r^2 \tag{6.80}$$

which can be integrated to yield

$$\varphi(r) = -\frac{Q}{4\pi r} \tag{6.81}$$

We should note that even though this result was derived in Minkowski coordinates, the result is highly general. Around any point we can introduce geodesic normal coordinates (this will be shown later in this chapter). Hence, this result is valid for all spherically symmetric spacetimes. The only requirement is that the coordinate system is such that the area of a spherical surface with radius r is $4\pi r^2$.

Example 6.5 (Gauss' integral theorem)

Let us apply Stoke's theorem to the one-form $\boldsymbol{\alpha}$, and let $\mathbf{A}$ be the vector associated with this form. Define $\boldsymbol{\beta} = \star\boldsymbol{\alpha}$ so that according to equations (6.24) and (6.25) we have

$$\mathbf{d}\boldsymbol{\beta} = (\nabla \cdot \mathbf{A})\boldsymbol{\epsilon}. \tag{6.82}$$

According to Stoke's theorem

$$\int\limits_{M} \mathbf{d}\boldsymbol{\beta} = \int\limits_{M} (\nabla \cdot \mathbf{A})\boldsymbol{\epsilon} = \int\limits_{\partial M} \boldsymbol{\beta}. \tag{6.83}$$

Also

$$\boldsymbol{\beta} \;\; = \;\; \star\boldsymbol{\alpha} = \sqrt{|g|}\,\varepsilon_{\mu|\mu_1 \dots \mu_{n-1}|}\alpha^{\mu}\mathbf{dx}^{\mu_1} \wedge \dots \wedge \mathbf{dx}^{\mu_{n-1}}. \tag{6.84}$$

If we now assume that $x^1, \dots, x^{n-1}$ are coordinates on the surface ∂M, and that x^n is the orthogonal direction, then this can be written

$$\boldsymbol{\beta} \;\; = \;\; \sqrt{\frac{|g|}{g_{nn}}}\,\varepsilon_{1\dots(n-1)\hat{n}}\alpha^{\hat{n}}\mathbf{dx}^1 \wedge \dots \wedge \mathbf{dx}^{n-1}. \tag{6.85}$$

Thus

$$\int\limits_{\partial M} \boldsymbol{\beta} = \int\limits_{\partial M} \mathbf{A} \cdot \mathbf{dS} \tag{6.86}$$

where $\mathbf{dS} = \hat{\mathbf{n}}\sqrt{\frac{|g|}{g_{nn}}}dx^1 \dots dx^{n-1}$ is the invariant surface element. In this case Stoke's theorem takes the form

$$\boxed{\int\limits_{M} (\nabla \cdot \mathbf{A})\boldsymbol{\epsilon} = \int\limits_{\partial M} \mathbf{A} \cdot \mathbf{dS},} \tag{6.87}$$

which is *Gauss' integral theorem.*

The variation of $\mathbf{n}$ along the curve is now given by

$$\frac{d\mathbf{n}}{ds} = \frac{d\mathbf{b}}{ds} \times \mathbf{t} + \mathbf{b} \times \frac{d\mathbf{n}}{ds} = -\tau \mathbf{n} \times \mathbf{t} + \kappa \mathbf{b} \times \mathbf{n} \tag{7.11}$$

so

$$\boxed{\frac{d\mathbf{n}}{ds} = \tau \mathbf{b} - \kappa \mathbf{t}.} \tag{7.12}$$

Equations (7.6), (7.9) and (7.12) are called the *Serret-Frenet equations*.

7.2 Surfaces

Consider a two-dimensional surface embedded in three-dimensional Euclidean space. Let u and v be coordinates (or parameters) on the surface. Then, at every point on the surface the basis vectors

$$\mathbf{e}_u = \frac{\partial}{\partial u}, \qquad \mathbf{e}_v = \frac{\partial}{\partial v} \tag{7.13}$$

define a tangent plane of the surface.

The line element on the surface is

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu \tag{7.14}$$

where $x^1 = u$ and $x^2 = v$. This is often called *the first fundamental form* of the surface.

The directional derivative of $\mathbf{e}_\mu$ along $\mathbf{e}_\nu$ has generally one component in the tangent plane of the surface and one component orthogonal to the surface,

$$\boxed{\mathbf{e}_{\mu,\nu} = \Gamma^\alpha_{\ \mu\nu} \mathbf{e}_\alpha + K_{\mu\nu} \mathbf{n}} \tag{7.15}$$

where $\Gamma^\alpha_{\ \mu\nu}$ is the connection coefficients of the u, v system and $\mathbf{n}$ is the unit normal vector field on the surface. The coefficients $K_{\mu\nu}$ are defined by this equation. Since $\mathbf{e}_{\mu,\nu} = \frac{\partial^2}{\partial x^\mu \partial x^\nu} = \frac{\partial^2}{\partial x^\nu \partial x^\mu} = \mathbf{e}_{\nu,\mu}$ these coefficients are symmetric. Note also that the eq. (7.15) provides us with an interpretation of the covariant derivative and the connection coefficients $\Gamma^\alpha_{\ \mu\nu}$. The covariant derivative of a surface embedded in an Euclidean space is the ordinary derivative in Euclidean space, projected onto the tangent space of the surface.

The equation (7.15) is usually called *Gauss' equations*. Using

$$K_{\mu\nu} = \mathbf{e}_{\mu,\nu} \cdot \mathbf{n} \tag{7.16}$$

we can calculate the coefficients $K_{\mu\nu}$. Let now $\mathbf{u}$ be the tangent vector of a curve in the surface, parametrized by λ. Using equation (7.15) we have

$$\frac{d\mathbf{u}}{d\lambda} = u^\mu_{\ ;\nu} u^\nu \mathbf{e}_\mu + K_{\mu\nu} u^\mu u^\nu \mathbf{n}. \tag{7.17}$$

The λ-independent coefficient in front of $\mathbf{n}$, $K_{\mu\nu} dx^\mu dx^\nu$ is called *the second fundamental form* of the surface. Whereas the first fundamental form determines the intrinsic geometry of the surface, the second fundamental form reflects the extrinsic geometry, i.e. how the surface curves in the ambient three-dimensional Euclidean space in which it is embedded.

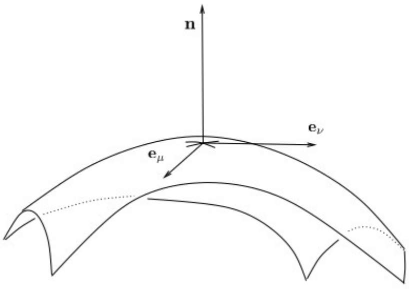

Figure 7.2: A two-dimensional surface embedded in three-dimensional Euclidean space

If λ is the arc-length, we can write equation (7.17) as

$$\frac{d\mathbf{u}}{d\lambda} = \kappa_g \mathbf{e} + \kappa_n \mathbf{n}, \quad \kappa_g \equiv |u^\mu{}_{;\nu} u^\nu \mathbf{e}_\mu|, \quad \kappa_n \equiv K_{\mu\nu} u^\mu u^\nu \tag{7.18}$$

where $\mathbf{e}$ is a unit vector. The quantities κ_g and κ_n are called *the geodesic curvature* and *the normal curvature* of the surface, respectively.

Since $\frac{d\mathbf{u}}{d\lambda}$ is orthogonal to $\mathbf{u}$ and $\mathbf{n}$, the unit vector $\mathbf{e}$ in the surface is given by

$$\mathbf{e} = \pm \mathbf{n} \times \mathbf{u}. \tag{7.19}$$

We see that the normal curvature can also be written

$$\kappa_n = \frac{d\mathbf{u}}{d\lambda} \cdot \mathbf{n}. \tag{7.20}$$

By differentiating $\mathbf{u} \cdot \mathbf{n}$ we get

$$\mathbf{u} \cdot \frac{d\mathbf{n}}{d\lambda} = -\kappa_n = -K_{\mu\nu} u^\mu u^\nu. \tag{7.21}$$

This is *Weingarden's equation*.

The geodesic and normal curvatures, taken separately, characterize the extrinsic geometry of the surface. We shall now find a quantity, called the Gaussian curvature, which is a measure of the intrinsic geometry of the surface.

At an arbitrary point on the surface we consider geodesic curves through the point with tangent vectors $\mathbf{u} = u^\mu \mathbf{e}_\mu$. In order to compare the normal curvature of the geodesics having different directions, we choose tangent vectors of unit length,

$$\mathbf{u} \cdot \mathbf{u} = g_{\mu\nu} u^\mu u^\nu = 1. \tag{7.22}$$

The directions with maximal and minimal values of the normal curvature are found by extremizing κ_n as given by eq. (7.18) with the constraint eq. (7.22). Thus we have to solve the variational problem δF for arbitrary u^μ, where

$$F = K_{\mu\nu} u^\mu u^\nu - k(g_{\mu\nu} u^\mu u^\nu - 1). \tag{7.23}$$

Here k is a Lagrange multiplier. Variation with respect to u^μ gives

$$\delta F = 2(K_{\mu\nu} - k g_{\mu\nu}) u^\nu \delta u^\mu. \tag{7.24}$$

Thus we must have

$$(K_{\mu\nu} - kg_{\mu\nu})u^\nu = 0. \tag{7.25}$$

This set of equations has non-trivial solutions whenever

$$\det\left(K_{\mu\nu} - kg_{\mu\nu}\right) = 0, \tag{7.26}$$

which yields the following quadratic equation for k

$$k^2 \det(g_{\mu\nu}) - k(g_{11}K_{22} - 2g_{12}K_{12} + g_{22}K_{11}) + \det(K_{\mu\nu}) = 0 \tag{7.27}$$

with solutions k_1 and k_2. These are the extremal values of k.

To see the meaning of k we multiply eq. (7.25) by u^μ and use eq. (7.18). This gives

$$k = \kappa_n. \tag{7.28}$$

The extremal values of k are called the *principal curvatures* of the surface. The *Gaussian curvature* of the surface is defined as

$$K = k_1 k_2. \tag{7.29}$$

From eq. (7.27) it follows that

$$\boxed{K = \frac{\det(K_{\mu\nu})}{\det(g_{\mu\nu})}.} \tag{7.30}$$

Let the directions corresponding to the principal curvatures be characterized by the vectors $\mathbf{u}$ and $\mathbf{v}$. From equation (7.25) and the symmetry of $K_{\mu\nu}$ follows

$$(k_1 - k_2)(\mathbf{u} \cdot \mathbf{v}) = 0. \tag{7.31}$$

For $k_1 \neq k_2$ this implies that $\mathbf{u}$ and $\mathbf{v}$ are orthogonal. The principal curvatures are found in orthogonal directions.

A positive Gaussian curvature means that the principal curvatures are of the same sign. The surface is locally similar to a distorted sphere, and the geometry is locally elliptic. If the Gaussian curvature vanishes, one of the principal curvatures have to vanish and the geometry is locally planar and the geometry is locally Euclidean. Finally, if the Gaussian curvature is negative the surface is locally saddle-shaped, i.e. like a hyperbolic surface. In this case the geometry is said to be locally hyperbolic.

In the following sections we will show that the Gaussian curvature represents a measure of the intrinsic geometry of the surface. This will be achieved by expressing it in terms of the components of the Riemann tensor, which we will introduce in the next section. Note however that curves have no intrinsic curvature.

7.3 The Riemann Curvature Tensor

In this section we will not restrict ourselves to two-dimensional surfaces, but describe spaces of arbitrary number of dimensions. We will introduce the important concept of curvature in an invariant way, described by the Riemann curvature tensor.

Consider two nearby points Q and P connected by a vector $\delta\lambda\mathbf{v}$ of infinitesimal length. Let $\mathbf{A}$ be a vector field. Then the difference between the vector field at Q, denoted $\mathbf{A}_Q$, and the vector $\mathbf{A}_P$ parallel-transported from P to Q (which is denoted $\mathbf{A}_{PQ}$) is to first order in $\delta\lambda$ given by the directional derivative at Q of $\mathbf{A}$ in the $\mathbf{v}$-direction

$$\delta\lambda\nabla_{\mathbf{v}}\mathbf{A} \approx \mathbf{A}_Q - \mathbf{A}_{PQ} \tag{7.32}$$

Thus

$$\mathbf{A}_{PQ} \approx (1 - \delta\lambda\nabla_{\mathbf{v}})\,\mathbf{A}_Q \tag{7.33}$$

To second order $\mathbf{A}_{PQ}$ is given by the first terms of the Taylor expansion

$$\mathbf{A}_{PQ} \approx \left(1 - \delta\lambda\nabla_{\mathbf{v}} + \frac{1}{2}\delta\lambda^2\nabla_{\mathbf{v}}\nabla_{\mathbf{v}}\right)\mathbf{A}_Q \tag{7.34}$$

We are now going to parallel-transport the vector $\mathbf{A}_P$ around the polygon shown in Fig.7.3. Parallel-transporting $\mathbf{A}_{PQ}$ from Q to R gives

$$\mathbf{A}_{PQR} \approx \left(1 - \delta\lambda\nabla_{\mathbf{u}} + \frac{1}{2}\delta\lambda^2\nabla_{\mathbf{u}}\nabla_{\mathbf{u}}\right)\left(1 - \delta\lambda\nabla_{\mathbf{v}} + \frac{1}{2}\delta\lambda^2\nabla_{\mathbf{v}}\nabla_{\mathbf{v}}\right)\mathbf{A}_R. \tag{7.35}$$

Proceeding around the polygon gives

$$\begin{aligned}\mathbf{A}_{PQRSTP} \;\approx\;& \left(1 + \delta\lambda\nabla_{\mathbf{u}} + \frac{1}{2}\delta\lambda^2\nabla_{\mathbf{u}}\nabla_{\mathbf{u}}\right)\left(1 + \delta\lambda\nabla_{\mathbf{v}} + \frac{1}{2}\delta\lambda^2\nabla_{\mathbf{v}}\nabla_{\mathbf{v}}\right)\\ \times\;& \left(1 - \delta\lambda^2\nabla_{[\mathbf{u},\mathbf{v}]}\right)\\ \times\;& \left(1 - \delta\lambda\nabla_{\mathbf{u}} + \frac{1}{2}\delta\lambda^2\nabla_{\mathbf{u}}\nabla_{\mathbf{u}}\right)\left(1 - \delta\lambda\nabla_{\mathbf{v}} + \frac{1}{2}\delta\lambda^2\nabla_{\mathbf{v}}\nabla_{\mathbf{v}}\right)\mathbf{A}_R.\end{aligned} \tag{7.36}$$

Thus, to second order, the change of the vector after parallel-transport around the polygon is

$$\delta\mathbf{A} = \mathbf{A}_{PQRSTP} - \mathbf{A}_P = \left([\nabla_{\mathbf{u}}, \nabla_{\mathbf{v}}] - \nabla_{[\mathbf{u},\mathbf{v}]}\right)\delta\lambda^2\mathbf{A}_P \tag{7.37}$$

In flat space there would be no such change of $\mathbf{A}$. This change is due to the curvature of the space. It can be shown that $\delta\mathbf{A}$ as given in eq. (7.37) is a vector which is linear in $\mathbf{A}$, $\mathbf{u}$ and $\mathbf{v}$ (see e.g. [vW81]). Thus $\delta\mathbf{A}$ may be expressed by a tensor of rank $\{^1_3\}$. This tensor $\mathbf{R}$ is called the *Riemann curvature tensor* and is defined by

$$\boxed{\mathbf{R}(\mathbf{u},\mathbf{v})\mathbf{A} \equiv \left([\nabla_{\mathbf{u}}, \nabla_{\mathbf{v}}] - \nabla_{[\mathbf{u},\mathbf{v}]}\right)\mathbf{A}.} \tag{7.38}$$

The components of this tensor are given by

$$\mathbf{e}_\mu R^\mu{}_{\nu\alpha\beta} = \left([\nabla_\alpha, \nabla_\beta] - \nabla_{[\mathbf{e}_\alpha,\mathbf{e}_\beta]}\right)\mathbf{e}_\nu \tag{7.39}$$

It follows that the Riemann curvature tensor is antisymmetric in $\mathbf{u}$ and $\mathbf{v}$ i.e. in α and β. We can therefore define a matrix of two-forms

$$\boxed{\mathbf{R}^\mu_\nu = \frac{1}{2}R^\mu{}_{\nu\alpha\beta}\boldsymbol{\omega}^\alpha \wedge \boldsymbol{\omega}^\beta} \tag{7.40}$$

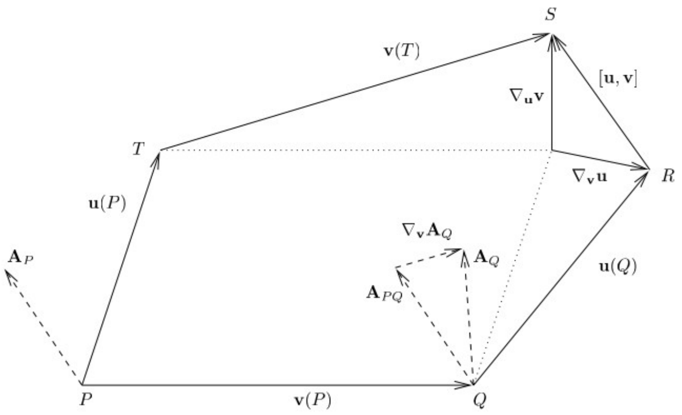
S
$\mathbf{v}(T)$
$[\mathbf{u},\mathbf{v}]$
$\nabla_{\mathbf{u}}\mathbf{v}$
T
$\nabla_{\mathbf{v}}\mathbf{u}$
R
$\mathbf{u}(P)$
$\nabla_{\mathbf{v}}\mathbf{A}_Q$
$\mathbf{A}_P$
$\mathbf{A}_Q$
$\mathbf{u}(Q)$
$\mathbf{A}_{PQ}$
P
$\mathbf{v}(P)$
Q

Figure 7.3: The vector $\mathbf{A}_P$ parallel transported around the polygon $PQRSTP$

which are called the *curvature forms*.

Equation (7.37) may now be written

$$\delta\mathbf{A} = \mathbf{e}_\mu R^\mu{}_{\nu\alpha\beta} A^\nu u^\alpha v^\beta. \tag{7.41}$$

The infinitesimal area of the polygon is to lowest order in $\mathbf{u}$ and $\mathbf{v}$

$$\Delta S^{\alpha\beta} = u^\alpha v^\beta - u^\beta v^\alpha. \tag{7.42}$$

Using the antisymmetry of $R^\mu{}_{\nu\alpha\beta}$ we may then write

$$\delta\mathbf{A} = \frac{1}{2} A^\nu R^\mu{}_{\nu\alpha\beta} \Delta S^{\alpha\beta} \mathbf{e}_\mu. \tag{7.43}$$

This shows that the change of a vector by parallel transport around a closed curve is proportional to the curvature of the space and to the area enclosed by the curve.

We shall now show how $R^\mu{}_{\nu\alpha\beta}$ may be expressed by the connection- and structure coefficients of a given basis. We find

$$\begin{aligned}
\mathbf{e}_\mu R^\mu{}_{\nu\alpha\beta} &= \left([\nabla_\alpha, \nabla_\beta] - \nabla_{[\mathbf{e}_\alpha,\mathbf{e}_\beta]}\right)\mathbf{e}_\nu \\
&= \left(\nabla_\alpha\nabla_\beta - \nabla_\beta\nabla_\alpha - c^\rho{}_{\alpha\beta}\nabla_\rho\right)\mathbf{e}_\nu \\
&= \left(\Gamma^\mu{}_{\nu\beta,\alpha} + \Gamma^\rho{}_{\nu\beta}\Gamma^\mu{}_{\rho\alpha} - \Gamma^\mu{}_{\nu\alpha,\beta} - \Gamma^\rho{}_{\nu\alpha}\Gamma^\mu{}_{\rho\beta} - c^\rho{}_{\alpha\beta}\Gamma^\mu{}_{\nu\rho}\right)\mathbf{e}_\mu
\end{aligned} \tag{7.44}$$

which implies that

$$\boxed{R^\mu{}_{\nu\alpha\beta} = \Gamma^\mu{}_{\nu\beta,\alpha} - \Gamma^\mu{}_{\nu\alpha,\beta} + \Gamma^\rho{}_{\nu\beta}\Gamma^\mu{}_{\rho\alpha} - \Gamma^\rho{}_{\nu\alpha}\Gamma^\mu{}_{\rho\beta} - c^\rho{}_{\alpha\beta}\Gamma^\mu{}_{\nu\rho}.} \tag{7.45}$$

The connection coefficients and structure coefficients are calculated from a basis vector field in the space considered, without reference to any higher dimensional space. Thus eq. (7.45) shows that the Riemann tensor describes the intrinsic geometry of space.

The curvature forms may now be written in component form as

$$\mathbf{R}^{\mu}{}_{\nu} = \left(\Gamma^{\mu}{}_{\nu\beta,\alpha} - \frac{1}{2} c^{\rho}{}_{\alpha\beta} \Gamma^{\mu}{}_{\nu\rho} + \Gamma^{\rho}{}_{\nu\beta} \Gamma^{\mu}{}_{\rho\alpha} \right) \boldsymbol{\omega}^{\alpha} \wedge \boldsymbol{\omega}^{\beta}. \tag{7.46}$$

Using equations (6.136) and (6.165) we find

$$\boxed{\mathbf{R}^{\mu}{}_{\nu} = \mathbf{d}\boldsymbol{\Omega}^{\mu}{}_{\nu} + \boldsymbol{\Omega}^{\mu}{}_{\lambda} \wedge \boldsymbol{\Omega}^{\lambda}{}_{\nu}.} \tag{7.47}$$

This is *Cartan's 2nd structural equation.*

We shall now deduce some identities fulfilled by the Riemann tensor. Using eqs. (6.118) and (6.144) we get

$$\begin{aligned} A^{\mu}{}_{;\beta\alpha} &= A^{\lambda} \Gamma^{\mu}{}_{\lambda\beta,\alpha} - A^{\lambda} \Gamma^{\tau}{}_{\lambda\alpha} \Gamma^{\mu}{}_{\tau\beta} - A^{\mu}{}_{;\lambda} \Gamma^{\lambda}{}_{\beta\alpha} + A^{\mu}{}_{,\beta\alpha} \\ &\quad + \left(A^{\lambda}{}_{;\alpha} \Gamma^{\mu}{}_{\lambda\beta} + A^{\lambda}{}_{;\beta} \Gamma^{\mu}{}_{\lambda\alpha} \right). \end{aligned} \tag{7.48}$$

The two latter terms are symmetric in α and β and do not contribute to the antisymmetric combination $A^{\mu}{}_{;\beta\alpha} - A^{\mu}{}_{;\alpha\beta}$. Hence, by using eqs. (7.45) and (6.134), we get

$$A^{\mu}{}_{;\beta\alpha} - A^{\mu}{}_{;\alpha\beta} = A^{\lambda} R^{\mu}{}_{\lambda\alpha\beta} - A^{\mu}{}_{;\lambda} T^{\lambda}_{\alpha\beta}. \tag{7.49}$$

Since this is a tensor equation it is valid in an arbitrary basis. Eq. (7.49) is called the *Ricci identity*. In Riemannian geometry it reduces to

$$\boxed{A^{\mu}{}_{;\beta\alpha} - A^{\mu}{}_{;\alpha\beta} = R^{\mu}{}_{\lambda\alpha\beta} A^{\lambda}.} \tag{7.50}$$

Combining this with eq. (6.163) leads to

$$\mathbf{d}^2 \mathbf{A} = \frac{1}{2} R^{\mu}{}_{\nu\alpha\beta} A^{\nu} \mathbf{e}_{\mu} \otimes \boldsymbol{\omega}^{\alpha} \wedge \boldsymbol{\omega}^{\beta} \tag{7.51}$$

By exterior differentiation of Cartan's first structural equation, eq. (6.180) combined with Cartan's second structural equation and Poincaré's Lemma, we find

$$\mathbf{R}^{\mu}{}_{\nu} \wedge \boldsymbol{\omega}^{\nu} = \mathbf{d}\mathbf{T}^{\mu} + \boldsymbol{\Omega}^{\mu}{}_{\nu} \wedge \mathbf{T}^{\nu}. \tag{7.52}$$

By means of eq. (6.216), this equation takes the form

$$\mathbf{R}^{\mu}{}_{\nu} \wedge \boldsymbol{\omega}^{\nu} = \mathbf{D}\mathbf{T}^{\mu}. \tag{7.53}$$

This is *Bianchi's first identity*. In Riemannian geometry it reduces to

$$\boxed{\mathbf{R}^{\mu}{}_{\nu} \wedge \boldsymbol{\omega}^{\nu} = 0.} \tag{7.54}$$

On component form this equation is

$$\boxed{R^{\mu}{}_{[\nu\alpha\beta]} = 0.} \tag{7.55}$$

By exterior differentiation of Cartan's second structural equation, and using Poincaré's Lemma, we get

$$\mathbf{d}\mathbf{R}^{\mu}{}_{\nu} = \mathbf{d}\boldsymbol{\Omega}^{\mu}{}_{\lambda} \wedge \boldsymbol{\Omega}^{\lambda}{}_{\nu} - \boldsymbol{\Omega}^{\mu}{}_{\lambda} \wedge \mathbf{d}\boldsymbol{\Omega}^{\lambda}{}_{\nu} = \mathbf{R}^{\mu}{}_{\lambda} \wedge \boldsymbol{\Omega}^{\lambda}{}_{\nu} - \boldsymbol{\Omega}^{\mu}{}_{\lambda} \wedge \mathbf{R}^{\lambda}{}_{\nu}, \tag{7.56}$$

which is more usually written

$$\mathbf{dR}^{\mu}{}_{\nu} + \mathbf{\Omega}^{\mu}{}_{\lambda} \wedge \mathbf{R}^{\lambda}{}_{\nu} - \mathbf{R}^{\mu}{}_{\lambda} \wedge \mathbf{\Omega}^{\lambda}{}_{\nu} = 0. \tag{7.57}$$

Applying eq. (6.214) to a matrix of two-forms this equation may be written

$$\boxed{\mathbf{DR}^{\mu}{}_{\nu} = 0.} \tag{7.58}$$

This is *Bianchi's second identity*. On component form this identity becomes

$$\boxed{R^{\mu}{}_{\nu[\alpha\beta;\gamma]} = 0.} \tag{7.59}$$

An additional symmetry of the Riemann tensor is most easily found by decomposing it in an orthonormal basis field. Applying eq. (6.169) and Cartan's second structural equation (7.47), we get

$$\boxed{R_{\mu\nu\alpha\beta} = -R_{\nu\mu\alpha\beta}} \tag{7.60}$$

The fourth and last symmetry of the Riemann tensor is found by applying geodesic normal coordinates. From eq. (7.45), (6.110) and (6.232) we then get

$$R_{\mu\nu\alpha\beta} = \frac{1}{2}(g_{\mu\beta,\nu\alpha} - g_{\mu\alpha,\nu\beta} + g_{\nu\alpha,\mu\beta} - g_{\nu\beta,\mu\alpha}). \tag{7.61}$$

It follows that

$$\boxed{R_{\mu\nu\alpha\beta} = R_{\alpha\beta\mu\nu}.} \tag{7.62}$$

The four symmetries of the Riemann tensor reduce its number of independent components in an n-dimensional space from n^4 to $\frac{1}{12}n^2(n^2-1)$, in four-dimensional space-time from 256 to 20.

We shall now construct a curvature tensor of rank $\{{0 \atop 2}\}$ by contraction of the Riemann tensor. Note first that

$$R^{\alpha}{}_{\alpha\mu\nu} = 0 \tag{7.63}$$

because of the antisymmetry in the first two indices. Furthermore

$$R^{\alpha}{}_{\mu\nu\alpha} = -R^{\alpha}{}_{\mu\alpha\nu}. \tag{7.64}$$

Thus there exists only one independent non-vanishing contraction of the Riemann tensor. This is called the *Ricci tensor* and is usually written

$$\boxed{R_{\mu\nu} = R^{\alpha}{}_{\mu\alpha\nu}.} \tag{7.65}$$

From eq. (7.62) follows that it is symmetrical. It has $\frac{1}{2}n(n+1)$ independent components in an n-dimensional space; 10 components in four-dimensional space-time.

Contraction of the Ricci tensor gives the *Ricci scalar*

$$\boxed{R = R^{\mu}{}_{\mu}.} \tag{7.66}$$

Let us calculate the divergence of the Ricci tensor. Contraction of Bianchi's second identity eq.(7.58) (μ with α), yields

$$\begin{aligned} R^{\mu}{}_{\nu\mu\beta;\gamma} + R^{\mu}{}_{\nu\gamma\mu;\beta} + R^{\mu}{}_{\nu\beta\gamma;\mu} &= \\ R_{\nu\beta;\gamma} - R_{\nu\gamma;\beta} + R^{\mu}{}_{\nu\beta\gamma;\mu} &= 0. \end{aligned} \tag{7.67}$$

Raising the index ν and contracting with γ leads to

$$R^{\nu}{}_{\beta;\nu} - R_{;\beta} + R^{\mu}{}_{\beta;\mu} = 0. \tag{7.68}$$

Hence the divergence of the Ricci tensor is

$$R^{\nu}{}_{\mu;\nu} = \frac{1}{2}\left(\delta^{\nu}{}_{\mu}R\right)_{;\nu}. \tag{7.69}$$

Thus the tensor

$$E^{\nu}{}_{\mu} = R^{\nu}{}_{\mu} - \frac{1}{2}\delta^{\nu}{}_{\mu}R \tag{7.70}$$

is divergence free. This is *Einstein's curvature tensor*. Its covariant components are

$$\boxed{E_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu}.} \tag{7.71}$$

It follows immediately that this tensor is symmetric.

In an n-dimensional space the vanishing of the divergence of the Einstein tensor represents n equations. Thus the Einstein tensor has $\frac{1}{2}n(n-1)$ independent components in general, and 6 independent components in four-dimensional space-time.

The Weyl Curvature Tensor

Let us now focus on the case where the dimension of the manifold is 4, like our four-dimensional spacetime. The symmetries of the Riemann tensor imply that the Riemann tensor has 20 independent components in four dimensions. The Ricci tensor, on the other hand, has only 10 independent components. The components of the Riemann tensor which is not captured in the Ricci part form what is called the *Weyl curvature tensor*.

In four dimensions the Weyl curvature tensor is defined by

$$C_{\alpha\beta\gamma\delta} = R_{\alpha\beta\gamma\delta} - g_{\alpha[\gamma}R_{\delta]\beta} + g_{\beta[\gamma}R_{\delta]\alpha} + \frac{1}{3}Rg_{\alpha[\gamma}g_{\delta]\beta}. \tag{7.72}$$

It possesses the same symmetries as the Riemann tensor,

$$C_{\alpha\beta\gamma\delta} = C_{\gamma\delta\alpha\beta}, \quad C_{\alpha\beta\gamma\delta} = -C_{\beta\alpha\gamma\delta}, \quad C_{\alpha[\beta\gamma\delta]} = 0,$$

and in addition, *contraction over any pair of indices yields zero*:

$$C^{\alpha}{}_{\beta\alpha\delta} = 0. \tag{7.73}$$

Hence, when contracting the Riemann tensor over two indices only the Ricci part of it will survive. This gives, as we will see later, a physical interpretation of the Weyl tensor. Einstein's equations, which will be introduced in the next chapter, will only involve the Ricci tensor and hence the Weyl tensor represents the *free gravitational field*. Thus even if the Ricci tensor is zero, there can be a free gravitational field encoded in the Weyl tensor. This property gives rise to many interesting phenomena; many of which will be discussed in this book. Two important examples are gravitational waves, and the Schwarzschild vacuum solution.

The definition of the Weyl tensor in any dimension, see section 18.4. This section also discusses some further properties of the Weyl tensor.

7.4 Extrinsic and Intrinsic Curvature

From eq. (7.45) it is seen that the components of the Riemann tensor may be calculated from the components of the metric tensor. They are defined in terms of basis-vectors in a space which has a curvature specified by the Riemann tensor. Thus the specification of the metric does not presuppose any embedding of the curved space in a higher-dimensional flat space. The Riemann curvature tensor represents intrinsic geometric properties of space, which may be measured by inhabitants of that space. (Such inhabitants are always assumed to be creatures with the same number of dimensions as that of the space they inhabit.) Therefore one says that the Riemann tensor is a measure of the *intrinsic* curvature of space.

In the case of a two-dimensional surface, say a balloon, the intrinsic geometry is measured by two-dimensional creatures on the surface, "flatlanders". In general, if the Riemann tensor in a space vanishes, the space is flat. For a two-dimensional surface this means that the surface may be rolled onto an Euclidean plane without any local changes of the geometry. If the surface is an elastic membrane, no stresses or strains are introduced by this "rolling out" of it. The intrinsic geometry of a cylindrical surface for example, is Euclidean, and has no intrinsic curvature.

However, as seen from an external Euclidean three-dimensional space, the cylindrical surface looks curved. The surface has external curvature. We shall now introduce a tensor measuring the external curvature of a space, which is embedded in a space of one more dimension.

In the following we shall consider a curved n-dimensional space $\mathcal{M}^n$ embedded in an $(n+1)$-dimensional space $\mathcal{M}^{n+1}$ which also may be curved. Such a space embedded in an space with one dimensional higher, is called a *hypersurface*. Furthermore, Greek indices are associated with $\mathcal{M}^{n+1}$ and Latin indices with $\mathcal{M}^n$.

A measure of the extrinsic curvature of a space is obtained by considering how the direction of the unit normal vector $\mathbf{N}$ to the hypersurface changes with position on the hypersurface. *The extrinsic curvature tensor* $\mathbf{K}$ is a tensor on $\mathcal{M}^n$ of rank $\{^0_2\}$ defined up to a sign ambiguity by

$$\boxed{K_{ab} = -\mathbf{e}_b \cdot \nabla_a \mathbf{n}} \tag{7.74}$$

where the covariant derivative is taken in the ambient space $\mathcal{M}^{n+1}$. Since $\mathbf{N}$ is orthogonal to the basis vectors $\mathbf{e}_b$ on $\mathcal{M}^n$, so that $\nabla_a(\mathbf{e}_b \cdot \mathbf{n}) = 0$ we get

$$K_{ab} = \mathbf{n} \cdot \nabla_a \mathbf{e}_b = \mathbf{n} \cdot \mathbf{e}_\alpha \Gamma^\alpha{}_{ba} = n^\alpha g_{\alpha\beta} \Gamma^\beta{}_{ba} = n_\alpha \Gamma^\alpha{}_{ba}, \tag{7.75}$$

where $\mathbf{n} = n^\alpha \mathbf{e}_\alpha$. We introduce an orthonormal basis $\{\mathbf{e}_{\hat{a}}\}$ on $\mathcal{M}^n$ and a normal unit vector $\mathbf{n} = \mathbf{e}_{\hat{n}}$. By defining $\mathbf{n} \cdot \mathbf{n} = \epsilon = \pm 1$ we get

$$K_{\hat{a}\hat{b}} = \Gamma^{\hat{n}}{}_{\hat{a}\hat{b}}\,, \qquad g_{\hat{n}\hat{n}} = \epsilon. \tag{7.76}$$

Equation (7.75) shows that the extrinsic curvature is symmetric.

Let us now deduce the relation between the Riemann tensors of $\mathcal{M}^{n+1}$ and $\mathcal{M}^n$ and the extrinsic curvature of $\mathcal{M}^n$.

We will calculate the Riemann tensors using orthonormal frames. The Riemann tensor in $\mathcal{M}^{n+1}$ projected onto $\mathcal{M}^n$ is given by

$$\left(\mathbf{d}^2 \mathbf{e}_{\hat{b}}\right)_\perp = \left(\frac{1}{2} {}^{(n+1)}R^{\hat{\lambda}}{}_{\hat{b}\hat{\alpha}\hat{\beta}} \mathbf{e}_{\hat{\lambda}} \otimes \boldsymbol{\omega}^{\hat{\alpha}} \wedge \boldsymbol{\omega}^{\hat{\beta}}\right)_\perp = \frac{1}{2} {}^{(n+1)}R^{\hat{a}}{}_{\hat{b}\hat{c}\hat{d}} \mathbf{e}_{\hat{a}} \otimes \boldsymbol{\omega}^{\hat{c}} \wedge \boldsymbol{\omega}^{\hat{d}}. \tag{7.77}$$

We can also write this in another way, using instead the Riemann tensor in $\mathcal{M}^n$

$$\begin{aligned} \left(\mathbf{d}^2\mathbf{e}_{\hat b}\right)_\perp &= \left(\mathbf{d}\left[\mathbf{e}_{\hat\alpha}\otimes\mathbf{\Omega}^{\hat\alpha}{}_{\hat b}\right]\right)_\perp = \left(\mathbf{d}\mathbf{e}_{\hat\alpha}\otimes\mathbf{\Omega}^{\hat\alpha}{}_{\hat b}\right)_\perp + \left(\mathbf{e}_{\hat\alpha}\otimes\mathbf{d}\mathbf{\Omega}^{\hat\alpha}{}_{\hat b}\right)_\perp \\ &= \mathbf{e}_{\hat a}\otimes\left(\mathbf{d}\mathbf{\Omega}^{\hat a}{}_{\hat b} + \mathbf{\Omega}^{\hat a}{}_{\hat\alpha}\wedge\mathbf{\Omega}^{\hat\alpha}{}_{\hat b}\right)_\perp . \end{aligned} \tag{7.78}$$

We can now use Cartan's second structural equation in $\mathcal{M}^n$ by decomposing the wedge product

$$\begin{aligned} \left(\mathbf{d}^2\mathbf{e}_{\hat b}\right)_\perp &= \mathbf{e}_{\hat a}\otimes\left(\mathbf{d}\mathbf{\Omega}^{\hat a}{}_{\hat b} + \mathbf{\Omega}^{\hat a}{}_{\hat\lambda}\wedge\mathbf{\Omega}^{\hat\lambda}{}_{\hat b} + \mathbf{\Omega}^{\hat a}{}_{\hat n}\wedge\mathbf{\Omega}^{\hat n}{}_{\hat b}\right)_\perp \\ &= \mathbf{e}_{\hat a}\otimes\left({}^{(n)}\mathbf{R}^{\hat a}{}_{b} + \left[\mathbf{\Omega}^{\hat a}{}_{\hat n}\wedge\mathbf{\Omega}^{\hat n}{}_{\hat b}\right]_\perp\right) \\ &= \frac{1}{2}\left({}^{(n)}R^{\hat a}{}_{\hat b\hat c\hat d} + \Gamma^{\hat a}{}_{\hat n\hat c}\Gamma^{\hat n}{}_{\hat b\hat d}\right)\mathbf{e}_{\hat a}\otimes\boldsymbol{\omega}^{\hat c}\wedge\boldsymbol{\omega}^{\hat d} \\ &= \frac{1}{2}\left({}^{(n)}R^{\hat a}{}_{\hat b\hat c\hat d} \pm K^{\hat a}{}_{\hat c}K_{\hat b d}\right)\mathbf{e}_{\hat a}\otimes\boldsymbol{\omega}^{\hat c}\wedge\boldsymbol{\omega}^{\hat d}. \end{aligned} \tag{7.79}$$

From equations (7.77) and (7.79) it follows that

$${}^{(n)}R^{\hat a}{}_{\hat b\hat c\hat d} = {}^{(n+1)}R^{\hat a}{}_{\hat b\hat c\hat d} \pm 2K^{\hat a}{}_{[\hat c}K_{\hat d]\hat b}. \tag{7.80}$$

This is an expression of *Gauss' Theorema Egregium*, which says that an inhabitant of, say, a three-dimensional space, may perform measurements, within the three-dimensional space, which reveals to him the curvature of that space. It is not necessary to embed the space in a higher-dimensional one.

Let us write equation (7.80) in a covariant manner. The Riemann tensor on the right side in eq. (7.80) is the projected Riemann tensor of the ambient space $\mathcal{M}^{n+1}$. If we split the metric tensor $g_{\alpha\beta}$ into

$$g_{\alpha\beta} = h_{\alpha\beta} + \epsilon n_\alpha n_\beta, \tag{7.81}$$

where n_α are the components of the unit normal vector $\mathbf{n}$, then the tensor $h_{\alpha\beta}$ will act as a projection tensor onto the space $\mathcal{M}^n$. In addition, the tensor $h_{\alpha\beta}$ will be the metric tensor of the space $\mathcal{M}^n$ for vectors on $\mathcal{M}^n$. Since the exterior curvature and the Riemann tensor in $\mathcal{M}^n$ already are projected we have

$$h^\lambda{}_\alpha K_{\lambda\beta} = K_{\alpha\beta} \tag{7.82}$$

i.e. they are eigentensors to the projection map $h^\lambda{}_\alpha$. The equation (7.80) can now be written covariantly as

$$\boxed{{}^{(n)}R^\alpha{}_{\beta\mu\nu} = {}^{(n+1)}R^\lambda{}_{\gamma\sigma\rho}h^\alpha{}_\lambda h^\gamma{}_\beta h^\sigma{}_\mu h^\rho{}_\nu + \epsilon\left(K^\alpha{}_\mu K_{\beta\nu} - K^\alpha{}_\nu K_{\beta\mu}\right).} \tag{7.83}$$

Similarly, we can show that

$$\boxed{{}^{(n)}\nabla_\alpha K_{\beta\mu} - {}^{(n)}\nabla_\mu K_{\beta\alpha} = {}^{(n+1)}R^\lambda{}_{\sigma\rho\delta}n_\lambda h^\sigma{}_\beta h^\rho{}_\alpha h^\delta{}_\mu} \tag{7.84}$$

where ${}^{(n)}\nabla_\alpha = h^\beta{}_\alpha\nabla_\beta$ is the n-dimensional connection. The derivation of this is left as a problem (see problem 7.4). This equation is called the *Codazzi equation*.

In the special case where $\mathcal{M}^{n+1}$ is flat, eq. (7.80) reduce to

$${}^{(n)}R_{\hat a\hat b\hat c\hat d} = \pm\left(K_{\hat a\hat c}K_{\hat d\hat b} - K_{\hat a\hat d}K_{\hat c\hat b}\right). \tag{7.85}$$

For a two dimensional surface $\mathcal{M}^2$ these equations reduce to the single equation

$$R_{\hat{1}\hat{2}\hat{1}\hat{2}} = K_{\hat{1}\hat{1}} K_{\hat{1}\hat{2}} - (K_{\hat{1}\hat{2}})^2 = \det(K_{\hat{a}\hat{b}}). \tag{7.86}$$

Comparing with eq. (7.30) of a surface (with $\det(g_{\hat{a}\hat{b}}) = 1$) we obtain

$$K = R_{\hat{1}\hat{2}\hat{1}\hat{2}}. \tag{7.87}$$

Thus, the Gaussian curvature represents the *intrinsic* geometry of the surface.

Example 7.2 (The curvature of a straight circular cone) **Example**

From the figure 7.4 it is seen that the line element on the conical surface is

$$ds^2 = dl^2 + r^2 d\theta^2 = dl^2 + \left(\frac{R}{H}\right)^2 l^2 d\theta^2. \tag{7.88}$$

Here are l and θ coordinates on the surface, and r a coordinate normal to the axis of the cone.

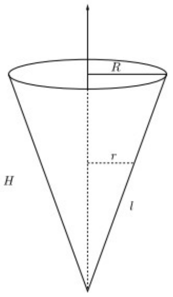

Figure 7.4: A cone is intrinsically flat

We introduce an orthonormal basis on the surface

$$\boldsymbol{\omega}^{\hat{l}} = \mathbf{dl}, \quad \boldsymbol{\omega}^{\hat{\theta}} = \left(\frac{R}{H}\right) \mathbf{d}\boldsymbol{\theta}. \tag{7.89}$$

Exterior differentiation yields

$$\mathbf{d}\boldsymbol{\omega}^{\hat{l}} = 0, \quad \mathbf{d}\boldsymbol{\omega}^{\hat{\theta}} = \frac{1}{l} \boldsymbol{\omega}^{\hat{l}} \wedge \boldsymbol{\omega}^{\hat{\theta}}. \tag{7.90}$$

Cartan's first structural equation now give the non-vanishing connection forms

$$\boldsymbol{\Omega}^{\hat{\theta}}{}_{\hat{l}} = -\boldsymbol{\Omega}^{\hat{l}}{}_{\hat{\theta}} = \frac{1}{l} \boldsymbol{\omega}^{\hat{\theta}}. \tag{7.91}$$

Thus

$$\mathbf{d}\boldsymbol{\Omega}^{\hat{\theta}}{}_{\hat{l}} = \left(\mathbf{d}\frac{1}{l}\right) \boldsymbol{\omega}^{\hat{\theta}} + \frac{1}{l} \mathbf{d}\boldsymbol{\omega}^{\hat{\theta}} = 0. \tag{7.92}$$

and Cartan's second structural equation gives

$$\mathbf{R}^{\hat{\theta}}{}_{\hat{l}} = \mathbf{R}^{\hat{l}}{}_{\hat{\theta}} = 0 \tag{7.93}$$

which shows that the intrinsic geometry of the conical surface is Euclidean.

The extrinsic curvature is given by eq. (7.75). In the present case the single non-vanishing component of this tensor is

$$K_{\hat{\theta}\hat{\theta}} = \Gamma^{\hat{n}}{}_{\hat{\theta}\hat{\theta}}. \tag{7.94}$$

In example 6.9 we found that $\Gamma^{\hat{r}}{}_{\hat{\theta}\hat{\theta}} = -\frac{1}{r}$, so

$$\mathbf{d}\mathbf{e}_{\hat{\theta}} = -\frac{1}{r}\mathbf{e}_{\hat{r}} \otimes \boldsymbol{\omega}^{\hat{\theta}} \tag{7.95}$$

From Fig.7.4 we see that

$$\mathbf{e}_{\hat{r}} = \frac{H}{L}\mathbf{e}_{\hat{n}} + \frac{R}{L}\mathbf{e}_{\hat{l}}, \quad L = (H^2 + L^2)^{\frac{1}{2}} \tag{7.96}$$

This yields

$$\mathbf{d}\mathbf{e}_{\hat{\theta}} = -\frac{H}{Lr}\mathbf{e}_{\hat{n}} \otimes \boldsymbol{\omega}^{\hat{\theta}} - \frac{R}{Lr}\mathbf{e}_{\hat{l}} \otimes \boldsymbol{\omega}^{\hat{\theta}} \tag{7.97}$$

which shows that

$$\Gamma^{\hat{n}}{}_{\hat{\theta}\hat{\theta}} = -\frac{H}{Lr} \tag{7.98}$$

Hence, the conical surface has a non-vanishing extrinsic curvature given by

$$K_{\hat{\theta}\hat{\theta}} = -\left(1 + \frac{R^2}{H^2}\right)^{-\frac{1}{2}} \cdot \frac{1}{r} \tag{7.99}$$

The limit $H \longrightarrow \infty, r \longrightarrow R$ represents a straight cylinder with $K_{\hat{\theta}\hat{\theta}} = -\frac{1}{R}$. The limit $H \longrightarrow 0$ represents a plane with $K_{\hat{\theta}\hat{\theta}} = 0$.

7.5 The equation of geodesic deviation

Consider two nearby geodesic curves, both parametrised by a parameter λ. Let s be a vector connecting points on the two curves with the same value of λ. The connecting vector s is said to measure the *geodesic deviation* of the curves.

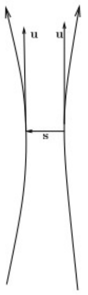

Figure 7.5: The two solid lines are neighbouring geodesics. They are connected by an infinitesimal vector s that obeys the equation of geodesic deviation

In order to deduce an equation describing how the geodesic deviation varies along the curves, we consider the covariant directional derivative of s along the curve $\nabla_{\mathbf{u}}\mathbf{s}$ where $\mathbf{u}$ is the tangent vector to the curve.

Let $\mathbf{u}$ and s be coordinate basis vectors of a coordinate system. Then $[\mathbf{s}, \mathbf{u}] = 0$, so that

$$\nabla_{\mathbf{u}}\mathbf{s} = \nabla_{\mathbf{s}}\mathbf{u} \tag{7.100}$$

giving

$$\nabla_{\mathbf{u}}\nabla_{\mathbf{u}}\mathbf{s} = \nabla_{\mathbf{u}}\nabla_{\mathbf{s}}\mathbf{u}. \tag{7.101}$$

Furthermore

$$\begin{aligned}\mathbf{R}(\mathbf{u},\mathbf{s})\mathbf{u} &= \left([\nabla_{\mathbf{u}}, \nabla_{\mathbf{s}}] - \nabla_{[\mathbf{u},\mathbf{s}]}\right)\mathbf{u} \\ &= [\nabla_{\mathbf{u}}, \nabla_{\mathbf{s}}]\,\mathbf{u}.\end{aligned} \tag{7.102}$$

Thus

$$\nabla_{\mathbf{u}}\nabla_{\mathbf{u}}\mathbf{s} = \nabla_{\mathbf{s}}\nabla_{\mathbf{u}}\mathbf{u} + \mathbf{R}(\mathbf{u},\mathbf{s})\mathbf{u}. \tag{7.103}$$

Since the curves are geodesics $\nabla_{\mathbf{u}}\mathbf{u} = 0$, and the equation reduces to

$$\boxed{\nabla_{\mathbf{u}}\nabla_{\mathbf{u}}\mathbf{s} + \mathbf{R}(\mathbf{s},\mathbf{u})\mathbf{u} = 0} \tag{7.104}$$

where we have used the antisymmetry of the Riemann tensor. Equation (7.104) is called the *equation of geodesic deviation*. The component form of the equation is

$$\left(\frac{d^2\mathbf{s}}{d\lambda^2}\right)^{\mu} + R^{\mu}{}_{\alpha\nu\beta}u^{\alpha}s^{\nu}u^{\beta} = 0. \tag{7.105}$$

This equation shows that the Riemann tensor can be determined entirely from measurements of geodesic deviation.

In comoving geodesic normal coordinates with $\mathbf{u} = (1, 0, 0, 0)$ the equation reduces to

$$\left(\frac{d^2\mathbf{s}}{d\lambda^2}\right)^{i} + R^{i}{}_{0j0}s^{j} = 0. \tag{7.106}$$

7.6 Spaces of constant curvature

We have in section 6.11 seen how important Killing vectors may be for the solvability of the equations that govern particle motion. Highly symmetric spaces are therefore important both in mathematics and in physics. We will study maximally symmetric spaces with a Riemannian metric, S^n, $\mathbb{P}^n$, $\mathbb{E}^n$ and $\mathbb{H}^n$. The first 2 of these differ only at a global scale, that is they are locally the same but their topology is different. Maybe the mathematically most interesting one is the hyperbolic space $\mathbb{H}^n$. We will start out with the most familiar one, the Euclidean space $\mathbb{E}^n$.

The Euclidean space, $\mathbb{E}^n$

We are all familiar with this space. The metric can be written

$$ds^2 = g_{ab}\mathbf{dx}^a \otimes \mathbf{dx}^b \tag{7.107}$$

where $g_{ab} = \delta_{ab}$. We can easily generalize this to the Minkowskian case by letting the components of the metric tensor be of either sign, $g_{ab} = \pm\delta_{ab}$. As claimed, this is a maximally symmetric space, thus having $\frac{1}{2}n(n+1)$ linearly independent Killing vectors. We should stress that when we say linearly independent in this context, we mean linearly independent *solutions* of Killing's equations. The Killing vectors are not linearly independent in the space itself. Killing's equations do in this case reduce to

$$\xi_{i,j} + \xi_{j,i} = 0 \tag{7.108}$$

because the connection coefficients all vanish identically. We can easily find n linearly independent solutions to this equation,

$$\boldsymbol{\xi}_{(a)} = \frac{\partial}{\partial x^b}. \tag{7.109}$$

Another set of solutions are[1]

$$\boldsymbol{\xi}_{(ab)} = x^b \frac{\partial}{\partial x^a} - x^a \frac{\partial}{\partial x^b} \tag{7.110}$$

The solutions $\boldsymbol{\xi}_{(ab)}$ are antisymmetric in the indices so they represent $\frac{1}{2}n(n-1)$ linearly independent solutions. Thus all in all we have found $n + \frac{1}{2}n(n-1) = \frac{1}{2}n(n+1)$ linearly independent solutions of Killing's equation. Hence, as claimed, the Euclidean spaces admit $\frac{1}{2}n(n+1)$ linearly independent Killing vectors and is therefore maximally symmetric.

The one-parameter group of diffeomorphisms associated to these Killing vectors can be found by solving the equations

$$\frac{d}{dt}\phi_t = \boldsymbol{\xi}. \tag{7.111}$$

Representing ϕ_t by a vector $\mathbf{V}$, the equations for $\boldsymbol{\xi}_{(a)}$ can be written

$$\frac{d}{dt}\mathbf{V}_{(a)} = \mathbf{e}_a. \tag{7.112}$$

It has solutions

$$\mathbf{V}_{(a)} = \mathbf{V}_0 + t\mathbf{e}_a \tag{7.113}$$

where $\mathbf{V}_0$ is a constant vector corresponding to the initial condition. These mappings are mere translations a distance t in the a-direction. The Euclidean spaces have translation invariance.

Defining $\mathsf{J}_{(ab)}$ as the antisymmetric matrix with indices

$$\left(\mathsf{J}_{(ab)}\right)_{ij} = \delta_{ai}\delta_{bj} - \delta_{bi}\delta_{aj} \tag{7.114}$$

---

[1] Since we use Cartesian coordinates we have $x_i \equiv g_{ij}x^j = x^i$. Hence, the position of the index does not matter.

we can write for the Killing vectors $\boldsymbol{\xi}_{(ab)}$

$$\frac{d}{dt}\mathbf{V}_{(ab)} = \mathsf{J}_{(ab)}\mathbf{V}_{(ab)}. \tag{7.115}$$

This equation have solutions

$$\mathbf{V}_{(ab)} = e^{t\mathsf{J}_{(ab)}} \cdot \mathbf{V}_0 \tag{7.116}$$

where

$$e^{t\mathsf{J}_{(ab)}} = 1 + t\mathsf{J}_{(ab)} + \frac{1}{2}t^2\mathsf{J}^2_{(ab)} + ... = \mathsf{R}_{(ab)}(t) \tag{7.117}$$

is the rotation matrix through an angle t with respect to the (ab)-plane. Showing this is left as an exercise, see the problems in the end of this chapter. Thus, the Euclidean spaces are also rotationally symmetric. The three-dimensional Euclidean space has 6 linear independent Killing vectors, representing translation and rotational invariance. There is one significant difference between these operations. While the translations move "the whole space", the rotations leave an axis fixed. For a point p, the subgroup of the isometry group that leaves p fixed, is called *the isotropy subgroup*. In the case of $\mathbb{E}^3$, the isotropy subgroup of any point is the group of rotations with respect to this point. We will come back to these concepts in a later chapter, where we will define these groups more rigorously.

The elliptic spaces, S^n and $\mathbb{P}^n$

Let us first start out by defining the spheres S^n. Consider the $(n+1)$-dimensional Euclidean space and look at the hypersurface

$$(x^1)^2 + (x^2)^2 + ... + (x^{n+1})^2 = 1. \tag{7.118}$$

This hypersurface with the induced metric is the sphere S^n. We claim now that the Killing vectors of the sphere is the Killing vectors from the Euclidean space that generate maps that map the sphere onto itself. These are the rotations with respect to the origin, with corresponding Killing vectors

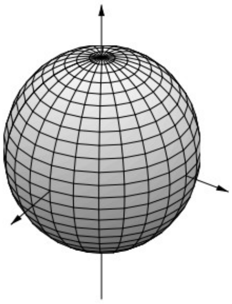

Figure 7.6: The two-dimensional sphere embedded in a Euclidean space

$$\boldsymbol{\xi}_{(\alpha\beta)} = x^\beta\frac{\partial}{\partial x^\alpha} - x^\alpha\frac{\partial}{\partial x^\beta} \tag{7.119}$$

where Latin indices run from 1 to $(n+1)$. Here we have $\frac{1}{2}n(n+1)$ of them, and when constrained to the sphere, they form the Killing vectors of the sphere. Thus this space is maximally symmetric. Let us use Gauss' Theorema Egregium to calculate its curvature properties. Given a point on the sphere, then for one of the coordinates $x^a \neq 0$. Assume that $x^{n+1} \neq 0$. Then we can define the following basis vectors

$$\mathbf{e}_a = \boldsymbol{\xi}_{(a(n+1))} = x^{n+1}\frac{\partial}{\partial x^a} - x^a \frac{\partial}{\partial x^{n+1}} \tag{7.120}$$

where Greek indices run from 1 to n. The metric in this basis is

$$\begin{aligned} g_{ab} &= \mathbf{e}_a \cdot \mathbf{e}_b = \left(x^{n+1}\frac{\partial}{\partial x^a} - x^a \frac{\partial}{\partial x^{n+1}}\right) \cdot \left(x^{n+1}\frac{\partial}{\partial x^b} - x^b \frac{\partial}{\partial x^{n+1}}\right) \\ &= x^a x^b + \left(x^{n+1}\right)^2 \delta_{ab}. \end{aligned} \tag{7.121}$$

The radial vector

$$\mathbf{e}_r = x^\alpha \frac{\partial}{\partial x^\alpha} \tag{7.122}$$

has unit length on the sphere. Using this we can calculate the components of the extrinsic curvature

$$\begin{aligned} K_{ab} &= \mathbf{e}_r \cdot (\mathbf{d}\mathbf{e}_a(\mathbf{e}_b)) = \mathbf{e}_r \cdot \left[\frac{\partial}{\partial x^a}\left(\mathbf{d}\mathbf{x}^{n+1}(\mathbf{e}_b)\right) - \frac{\partial}{\partial x^{n+1}}\left(\mathbf{d}\mathbf{x}^a(\mathbf{e}_b)\right)\right] \\ &= \mathbf{e}_r \cdot \left[x^b \frac{\partial}{\partial x^a} + \delta_{ab}(x^{n+1})\frac{\partial}{\partial x^{n+1}}\right] = x^a x^b + \left(x^{n+1}\right)^2 \delta_{ab}. \end{aligned} \tag{7.123}$$

Thus we have

$$K_{ab} = g_{ab}. \tag{7.124}$$

The Gauss' Theorema Egregium now gives us the Riemann tensor

$$R_{abcd} = g_{ac}g_{bd} - g_{ad}g_{bc}. \tag{7.125}$$

Contracting once,

$$R_{bd} = (n-1)g_{bd} \tag{7.126}$$

we see that the Ricci tensor is proportional to the metric[2]. Thus in an orthonormal frame, the Ricci tensor will be everywhere positive and constant. Contracting once more to obtain the Ricci scalar

$$R = n(n-1). \tag{7.127}$$

The projective space $\mathbb{P}^n$ can now be obtained by identifying opposite points on the sphere S^n. The projective space is therefore basically half of the sphere. There is a pathology however, for n even, the projective space is *non-orientable*. That means there is no globally defined unit normal vector field on the space. However, for n odd (including therefore dimension 3), the space is orientable, and no problem of such kind exists.

---

[2] More generally, spaces for which the Ricci tensor can be written $R_{ab} = \lambda g_{ab}$ where λ is a constant, are called *Einstein spaces*. All the constant curvature spaces are Einstein spaces.

Let us derive a useful form of the metric of the sphere. We introduce a radial coordinate r by

$$r^2 = \sum_a (x^a)^2 \tag{7.128}$$

so that

$$x^{n+1} = \pm\sqrt{1-r^2}. \tag{7.129}$$

Hence,

$$\left(dx^{n+1}\right)^2 = \frac{r^2 dr^2}{1-r^2}. \tag{7.130}$$

Assuming now that $d\Omega_{n-1}$ is the metric on the $(n-1)$-dimensional sphere, S^{n-1}, then

$$\sum_a (dx^a)^2 = dr^2 + r^2 d\Omega_{n-1}^2 \tag{7.131}$$

which is just the expression for the Euclidean metric in spherical coordinates. The metric on the sphere, S^n, is now

$$ds^2 = \left(dx^{n+1}\right)^2 + \sum_a (dx^a)^2 = \frac{dr^2}{1-r^2} + r^2 d\Omega_{n-1}^2. \tag{7.132}$$

Note that this metric only covers half of the sphere. For the case $\mathbb{P}^n$ it covers the whole space, except for the points on the equator which forms a set of measure zero.

The Hyperbolic spaces, $\mathbb{H}^n$

As for the sphere, the hyperbolic space can be viewed upon as a hypersurface in a flat $(n+1)$-dimensional space. But now we have to use the flat $(n+1)$-dimensional Minkowski space with metric

$$ds^2 = \eta_{\alpha\beta}\mathbf{dx}^\alpha \otimes \mathbf{dx}^\beta = -\mathbf{dx}^{n+1} \otimes \mathbf{dx}^{n+1} + \sum_a \mathbf{dx}^a \otimes \mathbf{dx}^a. \tag{7.133}$$

The hyperbolic space is defined as the hyperboloid

$$-(x^{n+1})^2 + (x^1)^2 + ... + (x^n)^2 = -1, \quad x^{n+1} > 0. \tag{7.134}$$

This surface is space-like and has a Riemannian metric. Its symmetries can be found by using the same argument as for the sphere. The symmetries for the Minkowski space that leave the hyperboloid invariant, are the Lorentz transformations in $(n+1)$ dimensions. The Killing vectors for the Lorentz transformations come in two classes, boosts

$$\boldsymbol{\xi}_{(a)} = x^{n+1}\frac{\partial}{\partial x^a} + x^a \frac{\partial}{\partial x^{n+1}} \tag{7.135}$$

and rotations

$$\boldsymbol{\xi}_{(ab)} = x^b\frac{\partial}{\partial x^a} - x^a \frac{\partial}{\partial x^b}. \tag{7.136}$$

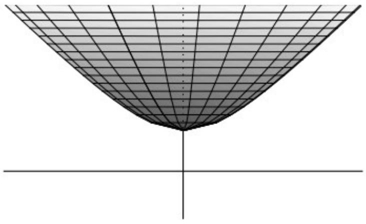

Figure 7.7: The Hyperbolic plane embedded in three-dimensional Minkowski space

As for the sphere we can now choose a set of basis vectors by

$$\mathbf{e}_a = \boldsymbol{\xi}_{(a)} \tag{7.137}$$

but these will now be globally defined since $x^{n+1} \geq 1$. The metric in this basis is

$$\begin{aligned} g_{ab} &= \mathbf{e}_a \cdot \mathbf{e}_b = \left(x^{n+1} \frac{\partial}{\partial x^a} + x^a \frac{\partial}{\partial x^{n+1}} \right) \cdot \left(x^{n+1} \frac{\partial}{\partial x^b} + x^b \frac{\partial}{\partial x^{n+1}} \right) \\ &= -x^a x^b + \left(x^{n+1} \right)^2 \delta_{ab} \end{aligned} \tag{7.138}$$

since $\frac{\partial}{\partial x^{n+1}} \cdot \frac{\partial}{\partial x^{n+1}} = -1$. The radial vector

$$\mathbf{e}_r = x^\alpha \frac{\partial}{\partial x^\alpha} \tag{7.139}$$

has on the hyperboloid a length

$$\mathbf{e}_r \cdot \mathbf{e}_r = -(x^{n+1})^2 + \sum_a (x^a)^2 = -1. \tag{7.140}$$

Following an almost identical procedure to the case of the sphere, we can calculate the extrinsic curvature

$$\begin{aligned} K_{ab} &= \mathbf{e}_r \cdot (\mathbf{d}\mathbf{e}_a(\mathbf{e}_b)) \\ &= g_{ab}. \end{aligned} \tag{7.141}$$

The Gauss's Theorema Egregium now turns into (we have to choose the negative sign due to the negative length of the normal unit vector)

$$R_{abcd} = -(g_{ac}g_{bd} - g_{ad}g_{bc}). \tag{7.142}$$

Contracting once,

$$R_{bd} = -(n-1)g_{bd}. \tag{7.143}$$

This space has negative curvature, and basically just the same curvature properties as for the sphere, just with an opposite sign of the curvature. A metric on the hyperbolic space can written by the following. Introduce spherical coordinates,

$$r^2 = \sum_a (x^a)^2 \tag{7.144}$$

and following the procedure as for the sphere, the metric can be written

$$ds^2 = \frac{dr^2}{1+r^2} + r^2 d\Omega_{n-1}^2 \tag{7.145}$$

where $d\Omega_{n-1}^2$ is the metric on the $(n-1)$-dimensional sphere. In this case the metric covers the whole of $\mathbb{H}^n$.

As we see, this space has infinite volume in contrast to the sphere. Note that there exist (and a lot of them) compact hyperbolic spaces as well, but these breaks the isometries for the $\mathbb{H}^n$ at a global scale. Locally, however, they have a hyperbolic metric and are locally isometric to $\mathbb{H}^n$.

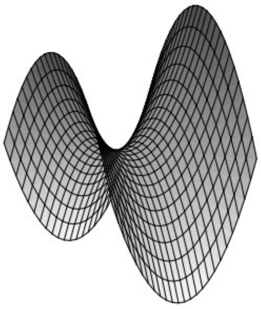

Figure 7.8: The hyperbolic plane can be seen upon as a saddle-surface

A more intuitive view of the hyperbolic plane is by considering it as the surface of a saddle, Fig. 7.8. A saddle surface in Euclidean three-space has a negative curvature and has therefore much of the same properties of the constantly curved hyperbolic plane. We should emphasize that the whole hyperbolic plane given by eq. (7.134) for $n = 2$ cannot be embedded in the three-dimensional Euclidean space. So the surface in Fig. 7.8 has not a constant curvature.

We can find a constant negatively curved space in Euclidean space. Consider the surface generated by the rotation around the z-axis of the *tractrix*

$$z = \ln\left|\frac{1 \pm \sqrt{1-r^2}}{r}\right| \mp \sqrt{1-r^2} \tag{7.146}$$

where $r^2 = x^2 + y^2$. This space will have constant negative curvature and are sometimes called a *pseudo-sphere*. The tractrix and the pseudo-sphere are depicted in Fig.7.9.

The name tractrix is due to the fact that this is the curve which is traced out by an object on the end of a rope of unit length held by a running child along the z-axis. If the object and the child start in the xz-plane at $(1,0)$ and $(0,0)$, respectively, then the curve is precisely the curve eq. (7.146) with $y = 0$. The points along the circle $r = 1$ correspond to a singular line, and hence, this cannot be the whole hyperbolic space.

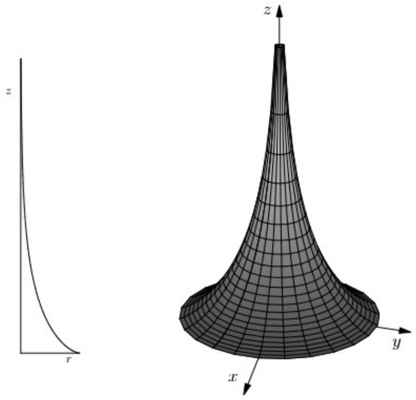

(a) The tractrix (b) The pseudo-sphere

Figure 7.9: The tractrix and the pseudo-sphere. The pseudo-sphere is obtained by rotating the tractrix around the z-axis.

Problems

7.1. *Rotation matrices*
Show that

$$e^{t\mathsf{J}_{(ab)}} = 1 + t\mathsf{J}_{(ab)} + \frac{1}{2}t^2\mathsf{J}^2_{(ab)} + ... = \mathsf{R}_{(ab)}(t), \tag{7.147}$$

i.e. verify eq. (7.117).

7.2. *Inverse metric on* S^n
In this problem we will use the basis vectors $\mathbf{e}_a$ defined in eq. (7.120):

$$\mathbf{e}_a = \boldsymbol{\xi}_{(a(n+1))} = x^{n+1}\frac{\partial}{\partial x^a} - x^a\frac{\partial}{\partial x^{n+1}}. \tag{7.148}$$

(a) Assume that the basis forms on the sphere S^n has the form

$$\boldsymbol{\omega}^a = C^a_{\ \mu}\mathbf{dx}^\mu \tag{7.149}$$

Find the coefficients $C^a_{\ \mu}$.

(b) Find g^{ab}, and verify that $g^{ab}g_{bc} = \delta^a_{\ c}$.

7.3. *The curvature of a curve*
Consider a curve $y = y(x)$ in a two-dimensional plane. Utilize that the differential of the arc length is $ds = \left(1 + y'^2\right)^{1/2} dx$ and show that the tangent vector, the curvature vector and the curvature of the curve are given by

$$\begin{aligned} \mathbf{t} &= \left(1 + y'^2\right)^{-1/2}(\mathbf{e}_x + y'\mathbf{e}_y), \\ \mathbf{k} &= \left(1 + y'^2\right)^{-2} y''(y'\mathbf{e}_x + \mathbf{e}_y), \\ \kappa &= \left(1 + y'^2\right)^{-3/2} y'', \end{aligned} \tag{7.150}$$

respectively.

7.4. *The Gauss-Codazzi equations*
The Gauss-Codazzi equations are of great importance and tells us much about the geometry of hypersurfaces. We will need these equations for a later chapter, and in this problem we will show an important relation between the Ricci curvatures in the ambient space and on the hypersurface.

(a) We will first show that the projection operator $h^{\alpha}_{\ \beta}$, defined by eq. (7.81) is what it is claimed to be. We define $h_{\alpha\beta}$ with eq. (7.81):

$$g_{\alpha\beta} = h_{\alpha\beta} + \epsilon n_{\alpha} n_{\beta} \tag{7.151}$$

where $g_{\alpha\beta}$ is the metric in the ambient space, and n_{α} is the components of the unit normal vector. The operator $h^{\alpha}_{\ \beta}$ is now defined by $h^{\alpha}_{\ \beta} = g^{\alpha\lambda} h_{\lambda\beta}$. Show the following:

$$\begin{aligned} h_{\alpha\beta} &= h_{\beta\alpha} \\ h^{\alpha}_{\ \lambda} h^{\lambda}_{\ \beta} &= h^{\alpha}_{\ \beta}. \end{aligned}$$

These properties defines a projection map. We now have to show that it projects onto the space that is orthogonal to n^{α}. Hence, if v^{β} is a vector, we have to show that

$$\begin{aligned} h^{\alpha}_{\ \beta} v^{\beta} &= 0, && \text{iff } v^{\beta} \text{ is parallel to } n^{\beta}. \\ h^{\alpha}_{\ \beta} v^{\beta} &= v^{\alpha}, && \text{iff } v^{\beta} \text{ is orthogonal to } n^{\beta}. \end{aligned}$$

Prove this. Explain now we can define the projection of an arbitrary tensor by projection the tensor, index by index.

(b) Verify the Codazzi equation, eq. (7.84). (Hint: Calculate $\left(\mathbf{d}^2 \mathbf{e}_{\hat{b}}\right) \cdot \mathbf{n}$.)

(c) Show that

$${}^{(n+1)}R = {}^{(n)}R - \epsilon\left(K^2 - K^{\alpha\beta}K_{\alpha\beta}\right) + 2\epsilon^{(n+1)}R_{\alpha\beta}n^{\alpha}n^{\beta}, \tag{7.152}$$

which can be written

$$-2\epsilon^{(n+1)}E_{\alpha\beta}n^{\alpha}n^{\beta} = {}^{(n)}R - \epsilon\left(K^2 - K^{\alpha\beta}K_{\alpha\beta}\right). \tag{7.153}$$

7.5. *The Poincaré half-space*
Consider half of $\mathbb{R}^3$, $z > 0$ with metric

$$ds^2 = \frac{1}{z^2}\left(dx^2 + dy^2 + dz^2\right) \tag{7.154}$$

(a) Calculate the connection forms and the curvature forms using the structural equations of Cartan.

(b) Calculate the Riemann tensor, the Ricci tensor and the Ricci scalar.

(c) Show that

$$R_{abcd} = -(g_{ac}g_{bd} - g_{ad}g_{bc}) \tag{7.155}$$

Compare this with the three-dimensional hyperbolic space. Are there any way we can differentiate between these two cases? Are they different manifestations of the same space?

7.6. *The pseudo-sphere*
Show that the tractrix, eq. (7.146), obeys the differential equation

$$\left(\frac{dr}{dz}\right)^2 = \frac{r^2}{1-r^2}. \tag{7.156}$$

Substitute this into the line-element for flat space, do the substitution $r = \sin\theta$, and show that the metric on the pseudo-sphere can be written

$$ds^2 = d\theta^2 + \sinh^2\theta d\phi^2. \tag{7.157}$$

Show that this metric has constant negative curvature.

7.7. *A non-Cartesian coordinate system in two dimensions*
Consider the following metric on a two-dimensional surface:

$$ds^2 = v^2du^2 + u^2dv^2. \tag{7.158}$$

We will show, in two different ways, that this is only the flat Euclidean plane in disguise.

(a) Use the orthonormal frame approach and find the connection one-forms $\mathbf{\Omega}^{\hat{a}}{}_{\hat{b}}$. Find also the curvature two-forms $\mathbf{R}^{\hat{a}}{}_{\hat{b}}$ and show that they are identically zero.

(b) Show that the metric can be put onto the form

$$ds^2 = dx^2 + dy^2$$

by finding a transformation matrix $\mathsf{M} = \left(M^i{}_a\right)$ connecting the basis vectors $\mathbf{e}_u$ and $\mathbf{e}_v$, and $\mathbf{e}_x$ and $\mathbf{e}_y$. This can be done using the following relations

$$\begin{aligned} g_{ab} &= g_{ij}M^i{}_aM^j{}_b \\ \frac{\partial M^i{}_a}{\partial x^b} &= \frac{\partial M^i{}_b}{\partial x^a}. \end{aligned} \tag{7.159}$$

Where do these relations come from?

7.8. *The curvature tensor of a sphere*
Introduce an orthonormal basis on the sphere, S^2, and use Cartan's structural equations to find the physical components of the Riemann curvature tensor.

7.9. *The curvature scalar of a surface of simultaneity*
The spatial line-element of a rotating disc is

$$d\ell^2 = dr^2 + \frac{r^2}{1-\frac{\omega^2r^2}{c^2}}d\phi^2. \tag{7.160}$$

Introduce an orthonormal basis on this surface and use Cartan's structural equations to find the Ricci scalar.

7.10. *The tidal force pendulum and the curvature of space*
We will again consider the tidal force pendulum from Example 1.3. Here we shall use the equation for geodesic deviation, eq. (7.104), to find the period of the pendulum.

(a) Why can the equation for geodesic equation be used to find the period of the pendulum in spite of the fact that the particles do not move along geodesics? Explain also why the equation can be used even though the centre of the pendulum does not follow a geodesic.

(b) Assume that the centre of the pendulum is fixed at a distance R from the centre of mass of the Earth. Introduce an orthonormal basis $\{\mathbf{e}_{\hat{a}}\}$ with the origin at the centre of the pendulum (see Fig.7.10).

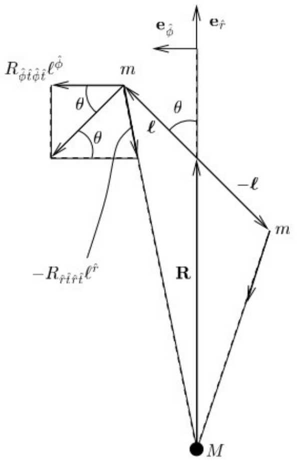

Figure 7.10: The tidal force pendulum

Show that, to first order in v/c and ϕ/c^2, where v is the three-velocity of the masses and ϕ the gravitational potential at the position of the pendulum, that the equation of geodesic equation takes the form

$$\frac{d^2\ell_{\hat i}}{dt^2} + R_{\hat i\hat 0\hat j\hat 0}\ell^{\hat j} = 0. \tag{7.161}$$

(c) Find the period of the pendulum expressed in terms of the components of Riemann's curvature tensor.

7.11. *The Weyl tensor vanishes for spaces of constant curvature*
Use the definition eq. (7.72) to show that the constant curvature spaces S^4, $\mathbb{E}^4$, and $\mathbb{H}^4$ all have zero Weyl tensor.

7.12. *Frobenius' Theorem*
In this problem we will consider integrability conditions for vector fields. In particular, we shall obtain a necessary and sufficent criterion for when a vectorfield is hypersurface orthogonal. Assume that we have a m-dimensional manifold M and a smooth collection of r-dimensional subspaces $\mathcal{D}_p \in T_pM$, one for each $p \in M$. A submanifold N of M is called an *integral submanifold* if $\mathcal{D}_p = T_pN$ for each $p \in N$. Frobenius' theorem can now be stated:

Frobenius' Theorem: Given a smooth collection of spaces $\mathcal{D}_p$ as above. Then there exists an integral submanifold N at every p if and only if $\{\mathcal{D}_p\}$ is involute; i.e., for all $p \in M$,

$$[\mathbf{X}_i, \mathbf{X}_j] \in \mathcal{D}_p, \quad \forall X_k \in \mathcal{D}_p.$$

(a) Assume that $\boldsymbol{\omega}^a$ is a set of one-forms that span the orthogonal complement of $\mathcal{D}_p$; i.e. $\boldsymbol{\omega}^a(\mathbf{X}_i) = 0$ for all $\mathbf{X}_i \in \mathcal{D}_p$. Show that $\mathbf{d}\boldsymbol{\omega}^a$ must be on the form

$$\mathbf{d}\boldsymbol{\omega}^a = \sum_b \boldsymbol{\omega}^b \wedge \boldsymbol{\rho}^b,$$

where each $\boldsymbol{\rho}^b$ is in T_pM. (Hint: Show first that $X^\mu Y^\nu \nabla_{[\mu}\omega^a_{\nu]} = 0$ for all $X^\mu, Y^\nu \in T_pN$.)

(b) In particular, this means if M is a m-dimensional manifold and ξ_μ is a covariant vector field, then ξ_μ is hypersurface orthogonal to some $(m-1)$-dimensional submanifold if and only if $\nabla_{[\mu}\xi_{\nu]} = \xi_{[\mu}\rho_{\nu]}$. Show that a sufficent condition for ξ_μ to be hypersurface orthogonal is

$$\xi_{[\mu}\nabla_\nu\xi_{\sigma]} = 0. \tag{7.162}$$

(a) Show that in this case the radial geodesic equation reduces to

$$\Gamma^r{}_{tt}dt^2 + 2\Gamma^r{}_{\phi t}d\phi dt + \Gamma^r_{\phi\phi}d\phi^2 = 0.$$

(b) Calculate the Christoffel symbols and show that the equation takes the form

$$\left(\frac{dt}{d\phi}\right)^2 - 2a\frac{dt}{d\phi} + a^2 - \frac{r^3}{M} = 0,$$

where M is the mass of the rotating body and a its angular momentum per unit mass, $a = J/M$.

(c) Use the solution of the geodesic equation and the four-velocity identity to show that the proper time interval $d\tau$ shown on a clock moving an angle $d\phi$ is

$$d\tau = \pm\sqrt{1 - \frac{3M}{r} \pm 2a\omega_0}d\phi,$$

where $\omega_0 = \left(M/r^3\right)^{1/2}$ is the angular velocity of a clock moving in the Schwarzschild spacetime in accordance with Kepler's 3rd law. The plus and minus sign apply to direct and retrograde motion, respectively.

(d) Show that to first order in a the proper time difference for one closed orbit ($\phi \to \phi + 2\pi$) in the direct and the retrograde direction is $\tau_+ - \tau_- \approx 4\pi a = 4\pi J/M$, or in S.I. units, $\tau_+ - \tau_- \approx 4\pi a = 4\pi J/mc^2$.

Estimate this time difference for clocks in satellites moving in the equatorial plane of the Earth. (The mass of the Earth is $m = 6 \cdot 10^{26}$kg and its angular momentum $J = 10^{34}$kg m^{2}s^{-1}.)

10.9. *The photon sphere radius of a Reissner-Nordström black hole*
Show that there exists a sphere of radius

$$r_{PS} = \frac{3M}{2}\left(1 + \sqrt{1 - \frac{8Q^2}{9M^2}}\right) \tag{10.276}$$

in the Reissner-Nordström black hole spacetime where photons will have circular orbits around the black hole.

10.10. *Curvature of 3-space and 2-surfaces of the internal and the external Schwarzschild spacetimes*

(a) The 3-space of the internal Schwarzschild solution has a geometry given by the line-element

$$d\ell_I^2 = \frac{dr^2}{1 - \frac{R_S}{R^2}r^2} + r^2(d\theta^2 + \sin^2\theta d\phi^2)$$

where $R_S = 2M$ is the Schwarzschild radius of the mass distribution and R its radius. The corresponding line-element for the external Schwarzschild solution is

$$d\ell_E^2 = \frac{dr^2}{1 - \frac{R_S}{r}} + r^2(d\theta^2 + \sin^2\theta d\phi^2)$$

Find the spatial curvature $k = k(r) = \frac{1}{6}R$ of the 3-spaces, where R is the Ricci scalar.

(b) We shall now consider the equatorial surfaces $\theta = \pi/2$. The line-elements of these surfaces are, for the internal solution

$$d\sigma_I^2 = \frac{dr^2}{1 - \frac{R_S}{R^2} r^2} + r^2 d\phi^2,$$

and for the external solution

$$d\sigma_E^2 = \frac{dr^2}{1 - \frac{R_S}{r}} + r^2 d\phi^2.$$

For these line-elements the Gaussian curvatures of the surfaces they describe are given by

$$K = -\frac{1}{2g} g_{\phi\phi}'' + \frac{g_{\phi\phi}}{4g^2} g_{rr}' g_{\phi\phi}' + \frac{g_{rr}}{4g^2} \left(g_{\phi\phi}' \right)^2,$$

where $g = g_{rr} g_{\phi\phi}$ and differentiation is with respect to r. Show that the Gaussian curvature of the equatorial surfaces are for
The internal solution: $K = R_S/R^3$. What sort of surface is this?
The external solution: $K = -(1/2)(R_S/r^3)$.

(c) The equatorial surfaces shall now be compared to the embedding surfaces. The Gaussian curvature of a surface of revolution given by $z = z(r)$, is

$$K = \frac{z' z''}{r(1 + z'^2)^2}.$$

Calculate the Gaussian curvatures of the embedding surfaces of the internal Schwarzschild solution, as given in problem 10.2, and of the external solution, as given in eq.(10.98). Compare the results with those of the previous point.

10.11. *Proper radial distance in the external Schwarzschild space*
Show that the proper radial distance from a coordinate position r to the horizon R_S in the external Schwarzschild space is

$$\ell_r = \sqrt{r}\sqrt{r - R_S} + R_S \ln \left(\sqrt{\frac{r}{R_S}} - \sqrt{\frac{r}{R_S} - 1} \right).$$

Find the limit of this expression for $R_S \ll r$.

10.12. *Gravitational redshift in the Schwarzschild spacetime*
Define z, describing the redshift of light, by

$$z = \frac{\Delta\lambda}{\lambda_e}, \tag{10.277}$$

where $\Delta\lambda$ is the change in the photons wavelength and λ_e the wavelength of the photon when emitted.

Show that the gravitational redshift of light emitted at r_E and received at r_R in the Schwarzschild spacetime outside a star of mass M is

$$z = \left(\frac{r_R - R_S}{r_E - R_S} \right)^{\frac{1}{2}} - 1$$

where $R_S = 2M$ is the Schwarzschild radius of the star. What is the gravitational redshift of light emitted from the surface of a neutron star as observed by a faraway observer? A neutron star has typically a mass of 1.2 solar masses and a radius of about 20km.

10.13. *The Reissner-Nordström repulsion*
Consider a radially infalling neutral particle in the Reissner-Nordström spacetime with $M > |Q|$. Show that when the particle comes inside the radius $r = Q^2/M$ it will feel a repulsion away from $r = 0$ (i.e. that $d^2r/d\tau^2 < 0$ for τ the proper time of the particle). Is this inside or outside the outer horizon r_+? Show further that the particle can never reach the singularity at $r = 0$.

10.14. *Light-like geodesics in the Reissner-Nordström spacetime*
We will in this problem consider radial photon paths in the Reissner-Nordström spacetime. The horizons of this spacetime are at $r_\pm = M \pm \sqrt{M^2 - Q^2}$, and we will assume that $M > |Q|$.

(a) Show that the radial light rays obey the differential equation

$$\frac{dr}{dt} = \pm\left(1 - \frac{2M}{r} + \frac{Q^2}{r^2}\right).$$

(b) It is convenient to introduce two *null coordinates* u and v by

$$\begin{aligned} u &= t - r^* \\ v &= t + r^*, \end{aligned} \qquad (10.278)$$

where

$$r^* = \int \frac{dr}{1 - \frac{2M}{r} + \frac{Q^2}{r^2}}.$$

Show that u is a constant of motion for outgoing photons, while v is a constant of motion for ingoing photons. Show further that

$$r^* = r + \frac{r_+^2}{r_+ - r_-}\ln\left[\frac{1}{2M}|r - r_+|\right] + \frac{r_-^2}{r_+ - r_-}\ln\left[\frac{1}{2M}|r - r_-|\right] \qquad (10.279)$$

(c) Draw the light-cones in the tr-plane for the three regions $r < r_-$, $r_- < r < r_+$ and $r_+ < r$.

10.15. *Birkhoff's theorem*
We will in this problem consider a spherically symmetric metric describing the spacetime external to some region. We will first assume that the metric is time dependent, but will show that, under some assumptions, that this cannot be possible.

A spherically symmetric metric outside a source can always be put onto the canonical form

$$ds^2 = -e^{2\alpha(r,t)}dt^2 + e^{2\beta(r,t)}dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2). \qquad (10.280)$$

Assume also that the spacetime is asymptotically flat; i.e.

$$\lim_{r\to\infty}\alpha(r,t) = \lim_{r\to\infty}\beta(r,t) = 0.$$

(a) Outside some r_0 we have $T_{\mu\nu} = 0$. Denote the derivative $\frac{\partial}{\partial r}$ with a prime and $\frac{\partial}{\partial t}$ with a dot. Show that Einstein's field equations in vacuum (for

$r > r_0$) can be written as

$$
\begin{aligned}
e^{-2\beta}\left(\frac{2\alpha'}{r}+\frac{1}{r^2}\right)-\frac{1}{r^2} &= 0 && (10.281)\\
e^{-2\beta}\left(\frac{1}{r^2}-\frac{2\beta'}{r}\right)-\frac{1}{r^2} &= 0 && (10.282)\\
2e^{-2\beta}\frac{\dot{\beta}}{r} &= 0 && (10.283)\\
e^{-2\beta}\left(\alpha''+\alpha'^2+\frac{\alpha'-\beta'}{r}-\alpha'\beta'\right) & && \\
-e^{-2\alpha}\left(\ddot{\beta}+\dot{\alpha}^2-\dot{\alpha}\dot{\beta}\right) &= 0. && (10.284)
\end{aligned}
$$

(b) Show that for $r > r_0$ we have $\beta(r,t) = \beta(r)$. Show also that $\alpha' = -\beta'$, and by integrating, $\alpha(r,t) = -\beta(r)$ for $r > r_0$. Explain that the metric has to have the form of metric (10.4).

This is what is called Birkhoff's theorem: *If a spacetime contains a region which is spherically symmetric, asymptotically flat, and empty ($T_{\mu\nu} = 0$) for $r > r_0$, then the metric in this region is time independent and hence independent of the dynamical properties of its source.*

10.16. *Gravitational mass*

(a) Use the line-element (10.4) and show that the surface gravity of a Schwarzschild black hole can be written

$$\kappa = -e^{\alpha-\beta}\alpha'. \tag{10.285}$$

(b) Show, using Einstein's field equations, that

$$4\pi r^2 e^{\alpha+\beta}\left(T^0_{\ 0}-T^1_{\ 1}-T^2_{\ 2}-T^3_{\ 3}\right) = \left(r^2 e^{\alpha-\beta}\alpha'\right)'. \tag{10.286}$$

Hence, deduce that the surface gravity can be written

$$\kappa = -\frac{4\pi}{r^2}\int\limits_0^r \left(T^0_{\ 0}-T^1_{\ 1}-T^2_{\ 2}-T^3_{\ 3}\right)e^{\alpha+\beta}r^2 dr. \tag{10.287}$$

(c) Define the gravitational mass M_G inside a radius r of a spherical mass distribution by

$$\kappa = -\frac{M_G}{r^2}, \tag{10.288}$$

and deduce that

$$M_G = 4\pi\int\limits_0^r \left(T^0_{\ 0}-T^1_{\ 1}-T^2_{\ 2}-T^3_{\ 3}\right)e^{\alpha+\beta}r^2 dr. \tag{10.289}$$

This is the *Tolman-Whittaker* expression for the gravitational mass of a system.

What is the condition for repulsive gravitation?

('/" 7! 9 !=

and hence, there must be a coordinate transformation on Σ which relates $h^+_{\mu\nu}$ with $h^-_{\mu\nu}$. Thus we can set $h^\pm_{\mu\nu} = h_{\mu\nu}$.

From the Gauss' Theorema Egregium, eq. (7.83), and the Codazzi equation (7.84) we have

$$\begin{aligned} E_{\mu\nu}n^\mu n^\nu|^\pm &= -\frac{1}{2}\epsilon^{(3)}R + \frac{1}{2}\left(K^2 - K_{\alpha\beta}K^{\alpha\beta}\right)^\pm && (16.9) \\ E_{\mu\nu}h^\mu{}_\alpha n^\nu|^\pm &= -\left({}^{(3)}\nabla_\mu K^\mu_\alpha - {}^{(3)}\nabla{}_\alpha K\right)^\pm && (16.10) \\ E_{\mu\nu}h^\mu{}_\alpha h^\nu{}_\beta|^\pm &= {}^{(3)}E_{\alpha\beta} + \epsilon n^\mu \nabla_\mu \left(K_{\alpha\beta} - h_{\alpha\beta}K\right)^\pm - 3\epsilon K_{\alpha\beta}K|^\pm \\ &\quad + 2\epsilon K^\mu_\alpha K_{\mu\beta}|^\pm + \frac{1}{2}\epsilon h_{\alpha\beta}(K^2 + K^{\mu\nu}K_{\mu\nu})^\pm. && (16.11) \end{aligned}$$

16.2 Einstein's field equations

We will now relate the curvature tensors to the energy-momentum tensor according to Einstein's field equations. We will allow for the energy-momentum tensor to be discontinuous at Σ, but continuous elsewhere. The metric tensor is required to be continuous across the whole spacetime.

The Einstein tensor contains second derivatives of the metric tensor, but we will allow the derivative $g_{\mu\nu;\alpha}n^\alpha$ to be discontinuous. The second derivative can therefore be a delta-function since $\theta'(x) = \delta(x)$ where $\theta(x)$ is the step-function

$$\theta(x) = \begin{cases} 0, & x < 0 \\ 1, & x > 1. \end{cases} \tag{16.12}$$

Let y be an orthogonal coordinate so that $\frac{\partial}{\partial y} = \mathbf{n}$ and $y = 0$ at the surface Σ. The most general energy-momentum tensor we can have across the boundary is therefore

$$T_{\alpha\beta} = S_{\alpha\beta}\delta(y) + T^+_{\alpha\beta}\theta(y) + T^-_{\alpha\beta}\theta(-y). \tag{16.13}$$

The energy-momentum tensor of the surface $S_{\alpha\beta}$ can be defined as the integral over the thickness of the surface Σ as the thickness goes to zero

$$S_{\alpha\beta} = \lim_{\tau\to 0} \int_{-\tau/2}^{\tau/2} T_{\alpha\beta}dy. \tag{16.14}$$

For this to be well defined the tensor $S_{\alpha\beta}$ has to live on the hypersurface so that

$$h^\alpha{}_\mu h^\beta{}_\nu S_{\alpha\beta} = S_{\mu\nu}. \tag{16.15}$$

This way of defining the tensor $S_{\alpha\beta}$ is called the *thin shell approximation*.

Introduce a set of coordinates so that x^i are coordinates on the hypersurface and y is the coordinate in the orthogonal direction (as before). Using the thin shell approximation we can find an expression for S_{ij}.

From Gauss' equation (16.11), we have

$$\lim_{\tau\to 0} \int_{-\tau/2}^{\tau/2} E_{ij}dy = \lim_{\tau\to 0} \int_{-\tau/2}^{\tau/2} \left[\epsilon n^\mu \nabla_\mu \left(K_{ij} - h_{ij}K\right) + U_{ij}\right] dy \tag{16.16}$$

(%

Pag. 9(27)

$$\nabla^2 \phi(r) = 4\pi G\rho(r) \qquad (1.32)$$

this Poisson equation is the local form of Newton's law of gravitation.

→ Mass generates a gravitational field according to Poisson's equation, and the gravitational field generates acceleration according to Newton's second law.

Pag. 118(136)

$$\int_M d\alpha = \int_{\partial M} \alpha . \qquad (6.71)$$

this is the generalization of Gauss's law for Vector calculus.

$$\int_{\partial M} E.dS = Q \qquad (6.75)$$

This is the famous Gauss's law in electromagnetism.

Pag. 151(169) - 7.2 Surfaces

⇒

(%

$$e_{\mu,\nu} = T^{\alpha}_{\mu\nu} e_{\alpha} + K_{\mu\nu} n \quad (7.15)$$

The Equation (7.15) is usually called Gauss' equation.

Important:

Pág. 152 (170) - Figure 7.2

A two-dimensional surface embedded in three-dimensional Euclidean space

The geodesic and normal curvatures, taken separately, characterize the extrinsic geometry of the surface.

We shall now find a quantity, called the Gaussian curvature, which is a measure of the intrinsic geometry of the surface.

Pág. 153 (171)

The Gaussian curvature of the surface is defined as

$$K = k_1 k_2 \quad (7.29)$$

(% /

Obs: Usar livro todo!!!

Pág. 153(171)

Gaussian curvature represents a measure of the intrinsic geometry of surface.

This will be achieved by expressing it in terms of the components of the Riemann tensor.

Pág. 160(178) -

Gauss' Theorem Egregium

Pág. 161(179) Fig 7.4

A cone is intrinsically flat.

Pág. 166(184)

Let us use Gauss' theorem Egregium to calculate its curvature properties.

The Gauss' theorema Egregium now gives us the Riemann tensor (7.125)

Pág. 171(189) 7.4

The Gauss-Codazzi equations

Pág. 255(273) →

' (%

Gaussian curvatures of the surfaces they describe are given by..

$$K = -\frac{1}{2g} g'_{\phi\phi} + \frac{g_{\phi\phi}}{4g^2} g'_{rr} g'_{\phi\phi} + \frac{g_{rr}}{4g^2} (g'_{\phi\phi})^2, \ldots$$

Pág. 429 (447)

Gauss' Theorema Egregium and the Codazzi equation, we have..

(16.9) (16.10) e (16.11).

- Einstein's theory - a rigorous introduction to general relativity (2002).pdf

Pág. 349

The intrinsic curvature K_T is called the Gaussian curvature.

- Conversations on Quantum Gravity .. (2021).pdf

Pág. 16 (28)

What is asymptotic safety?

Gaussian fixed point

Pág. 445 (457) non-Gaussian fixed point contrary to a Gaussian fixed point?

'
-256- - -* ! - ' = 0 3 !
" ! ! A 3 ! ! ! 3 ! " 2" & 1"
(%'/ 9 A! + ! " 7! ! A!

is $K = 2/(1+4x^2)^{3/2}$. This expression shows that the curvature of the parabola is largest, equal to 2, at $x = 0$, and decreases towards zero for large values of x. Far away from the y-axis the parabola approaches a straight line. Looking at Fig. 2.4 you see that this is in accordance with the shape of the curve. Note also that if we had tried to define the curvature as the rate of change of slope per unit distance along the x-axis, i.e. as y'', we would have found a constant curvature, equal to 2 for the parabola. A constant curvature corresponds to a circular arc, which clearly does not correspond to the shape of the parabola.

9.2 The curvature of surfaces

Let us inspect a simple surface; the *surface* of a cylinder. Suppose the cylinder you look at is of paper and that you roll it out like you would do with a rolled-up newspaper fetched by your dog. If there were drawings of triangles and other figures on the cylinder they would all retain their shapes and lengths. There would be no trace of distortions or deformations. The geometry of a cylindrical surface is the same as that of a flat paper. The right triangle drawn on a paper with lengths 3, 4, and 5 cm, retain these quantities when we roll the paper into a cylinder. The theorem of Pythagoras holds good also on the curved cylindrical surface. The sum of the angles remains 180 degrees. Two-dimensional creatures on a cylinder develop the plane Euclidean geometry. They have no concept of a cylinder as we see it. But of course a sufficiently long *straight* journey on the surface may lead back to where they started. This is an astonishing fact for the cylindrians, showing that the topology of their world is different from that of an Euclidean plane.

As three-dimensional creatures we may curl the paper cylinder more and more. The diameter of the cylinder gets smaller and smaller, the paper curves 'more and more' per cm. The 5 cm straight line of our drawn triangle is eventually changed into a spiral. But the cylindrians, as *mathematicians*, may not notice anything whatsoever. The geometrical measurements of the cylindrians show no difference. Technically we say that the *intrinsic curvature* of the cylindrical surface is zero. Embedded in a three-dimensional space, it has curvature which is greater or smaller. The curvature of a two-dimensional surface as measured when embedded in a three-dimensional space is called the *extrinsic curvature*.

In general relativity we are concerned with curved four-dimensional spacetime, without reference to any higher-dimensional flat space in which spacetime could be imagined to be embedded. Thus by 'curvature of spacetime' is always meant 'intrinsic curvature'. We might speak of the extrinsic curvature of spacetime, embedded in for example a ten-dimensional Euclidean space, but there is no need for such an exploit.

We shall deduce an expression showing how the intrinsic curvature of a two-dimensional surface can be found by measurements on the surface itself. The circumference and radius of a circle can be measured by means of a ruler. The quotient between these quantities is equal to 2π on a plane. We shall find how the intrinsic curvature of a surface is given by the deviation of this quotient from 2π.

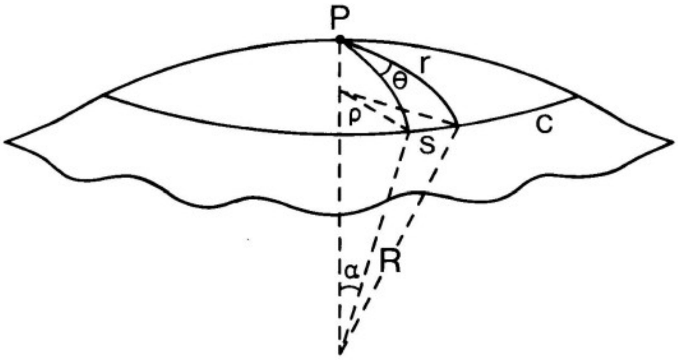

Figure 9.2: Surface of a dome

Consider a small part of any surface, for instance a part of the surface of a dome which surrounds a point P. By making the neighborhood around P small enough, the form of the considered mathematical surface will depart as little as one pleases from a spherical form. This is illustrated in Fig. 9.2.

From this figure and the definition of an angle as measured in radians (see Sect. 4.1) we see that

$$\alpha = \frac{r}{R}.$$

The flatter the dome, the larger is R, and the smaller is the angle α. Furthermore

$$\sin\alpha = \frac{\rho}{R}.$$

Thus

$$\rho = R\,\sin\alpha = R\,\sin(r/R).$$

Also, as shown more clearly in Fig. 9.3

$$\theta = \frac{s}{\rho},$$

where s is the distance along the circle C extended by the angle θ. We get

$$s = \theta\rho = \theta\,R\,\sin(r/R). \tag{9.5}$$

This expression shall be used to find the curvature at a point P of the surface depicted on Fig. 9.2. In general the curvature of a surface is a function of the position on the surface. The curvature at P is a local quantity, generally different from the curvature at other points on the surface. Therefore, we shall need to develop an expression for s, valid for very small values of r. This is performed by means of a series expansion of the expression (9.5).

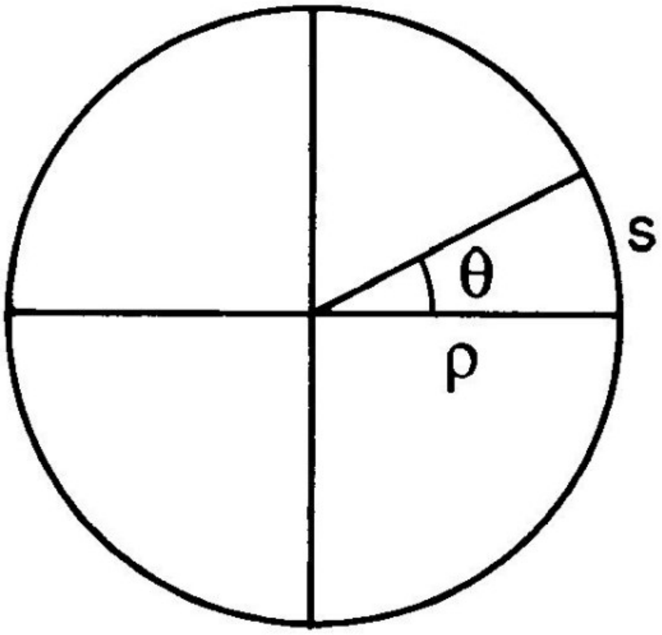

Figure 9.3: The dome seen from above

The first two terms of the MacLaurin series for $\sin x$ is (see Sect. 4.1.1)

$$\sin x = x - \frac{1}{6}x^3 + \cdots . \tag{9.6}$$

Here $x = r/R$, and we get from Eq. (9.5)

$$s = \theta\, R \left(\frac{r}{R} - \frac{1}{6}\frac{r^3}{R^3} + \cdots \right) = \theta \left(r - \frac{1}{6}\frac{1}{R^2} r^3 + \cdots \right). \tag{9.7}$$

We have obtained our desired expression for s. This will be used to find the curvature of our surface, Fig. 9.2, at P.

Here R is the radius of the sphere that the surface around P is part of. The *extrinsic curvature*—the curvature measured from the outside of the surface—is called 'K_E', and is defined as

$$K_E \equiv \frac{1}{R}.$$

It is called extrinsic because it refers to the radius of curvature R of the two-dimensional surface as embedded in the three-dimensional space external to the surface. Equation (9.7) may now be written as

$$s = \theta \left(r - \frac{1}{6}{K_E}^2\, r^3 + \cdots \right).$$

The length ℓ of the circle C around P with radius r on the surface, is

$$\ell = 2\pi\left(r - \frac{1}{6}{K_E}^2\, r^3 + \cdots\right), \tag{9.8}$$

since the angle θ around the circumference is 2π. Dividing each side of Eq. (9.8) by 2π we get

$$\frac{l}{2\pi} \approx r - \frac{1}{6}\,{K_E}^2\, r^3.$$

Thus

$$\frac{1}{6}\,{K_E}^2\, r^3 \approx r - \frac{\ell}{2\pi} = \frac{2\pi r\; - \ell}{2\pi}.$$

Multiplying by 6 and dividing by r^3 we get

$${K_E}^2 \approx \frac{3}{\pi}\,\frac{2\pi r - \ell}{r^3}.$$

Taking the limit $r \to 0$ we get

$${K_E}^2 = \frac{3}{\pi}\lim_{r\to 0}\frac{2\pi r - \ell}{r^3}. \tag{9.9}$$

We have now succeeded in expressing the extrinsic curvature of the considered surface in terms of quantities that can be measured by a two-dimensional creature on the surface.

The *intrinsic curvature* K_I of a surface at a point P is given by the expression at the right-hand side of Eq. (9.9). Thus

$$K_I = \frac{3}{\pi}\lim_{r\to 0}\frac{2\pi r - \ell}{r^3}. \tag{9.10}$$

The intrinsic curvature K_I is called the *Gaussian curvature*. From Eqs. (9.9) and (9.10) it follows that for the considered surface $K_I = {K_E}^2$. This extremely simple relation is not easily seen intuitively, and is *not* valid in general. It is due to the special property of the considered surface; that this surface was assumed to be indistinguishable from a spherical cap at a sufficiently small region around a point P. This is not a general property of surfaces. The *cylindrical surface* we mentioned above, for example, is locally similar to a plane. It has $K_E \neq 0$ and $K_I = 0$.

In general the extrinsic curvature of a surface may be characterized as follows. At every point on the surface we consider the straightest possible curves with different directions. We then find the curvature of these curves. They curve in the direction normal to the surface. Generally the curvature varies with the direction of the curve (except in the isotropic case that we considered above). The maximal and minimal curvatures are called the 'principal curvatures' of the surface, and are denoted by κ_1 and κ_2. The *extrinsic curvature* is defined by

$$K_E \equiv \frac{1}{2}\left(\kappa_1 + \kappa_2\right). \tag{9.11}$$

The *intrinsic* (Gaussian) curvature is given by

$$K_I \equiv \kappa_1\, \kappa_2. \tag{9.12}$$

For a spherical surface $\kappa_1 = \kappa_2$, which gives $K_I = {K_E}^2 =$

$1/R^2$. In the case of a cylindrical surface $\kappa_1 \neq 0$ and $\kappa_2 = 0$, which gives $K_E = \kappa_1/2 \neq 0$ and $K_I = 0$.

If $K_I \neq 0$, K_I may be either positive or negative. The case we have considered in Fig. 9.2 has $\ell < 2\pi r$, giving $K_I > 0$.

The geometry of such surfaces is called *parabolic*. In the case that $l > 2\pi r$, the spherical surface is replaced by a surface similar to a saddle. In this case $K_I < 0$. The geometry of such surfaces is called *hyperbolic*.

In elementary school the famous π is a symbol for the ratio between the circumference and diameter of a circle, and the value of π is given as $3.14159\cdots$. True or false? On a flat surface: true. But what about circles drawn on curved surfaces? Think of a circle on a spherical surface. Let the centre of the circle be at the North Pole. From the point of view of a Euclidean three-dimensional space in which the spherical surface is embedded, the radius of the circle is a curved path on the surface, from the North Pole to the circle. Assume that the radius of the circle is made larger, so that the circle approaches the Equator, and even passes the Equator. Then the circumference of the circle *decreases* while the radius increases. Obviously, the ratio of the length of the circumference and the length of the radius of the circle is not constant when the radius of the circle increases. It is due to the curvature of the surface.

What we have seen is that if one maintains the school-definition of π, then the magnitude of π will depend upon the diameter of the circle that it refers to. This may be maintained, but not the implicit assumption that the circle must be drawn on a flat surface. But both mathematical and physical flatness is an extremely special condition.

Defining the concept of π as quotient between circumference and diameter, its value may vary from zero to infinity, dependent upon the curvature of a surface! We shall in what follows mercifully retain the value of π you always find, by modifying the school definition of π, since it is admittedly not very practical that π might sometimes be equal to 2.15 and sometimes for example equal to 8.50. In order not to violate the near-sacredness of $3.14159\cdots$ we need only to add four words to the definition of π: "on a flat surface". This more precise definition has a consequence with respect to the number of radians we get for the angle around a circle (see Sect. 4.1.1). In order that a radian shall be a geometry-independent measure of an angle, we shall demand that the number of radians around a circle is the same whether it refers to circles on flat or curved surfaces. Then the definition of a radian must also refer to a flat surface. We shall illustrate all this by example 9.1.

Example 9.1 *Consider a circle on a spherical surface, as shown in Fig. 9.4.*

Here ℓ is the circumference of a circle with radius r on we must perform a difficult operation. We must 'parallel transport' to the one we want to compare it to (see Fig. 1.3). In daily life this can be done for example by means of straight threads along which we slide the ends of the arrows. Such procedures work well on a flat surface. However, difficulties will appear when we try to perform such parallel transport over great distances on the Earth, as the curvature of the Earth's spherical surface will disturb the result (see Fig. 9.2).

1.3 Vector fields

From now on vectors are no longer material arrows, but arrows in a purely abstract sense. In what follows a 'vector' is a geometric quantity with geometrically defined 'magnitude' and 'direction'. However, just as there is a correspondence between ordinary numbers and certain physical quantities, such as for example temperature (number of degrees) and length (number of centimetres), there is a correspondence between vectors and a special class of physical quantities; namely, those having a direction.

When we are outdoors in the wind, the moving air fills the region around us. There is a measurable velocity of the air everywhere in the region. In Fig. 1.4 we have shown, as an illustration, a weather map with the wind velocity field over Europe on February 3, 1988. The velocity has a magnitude and a direction. It is a vector. Thus a velocity field is linked conceptually with every point of the region. These abstract vectors are everywhere. If one can think of God as omnipresent, then one might also be able to think of vectors as omnipresent. In such a region there is said to be a vector field.

In order to be able to calculate with vectors, we must have a practical means of specifying their directions and magnitudes. Let us consider vectors on a flat surface, say on the sheet of paper you are writing on. Then we place unit vectors along the horizontal edge at the bottom and along the vertical left-hand edge of the sheet. These are reference vectors.

Drawing equidistant parallels with unit distance between each, we obtain a grid (see Fig. 1.5). Each corner is given a number containing figures; the first figure gives the distance of the point from the left edge of the sheet, and the second figure its distance from the downward edge. Imagine that each square is divided into four smaller squares. Again each corner is given a two-figured number. (The figures are not generally integer numbers.) If this division is proceeded indefinitely, we can associate a number with every point of our sheet. Note also that each point have different numbers. Such a set of numbers, by which we may specify the positions of arbitrary points in a region, is called a coordinate system. There is no reference to the grid we introduced originally, in the

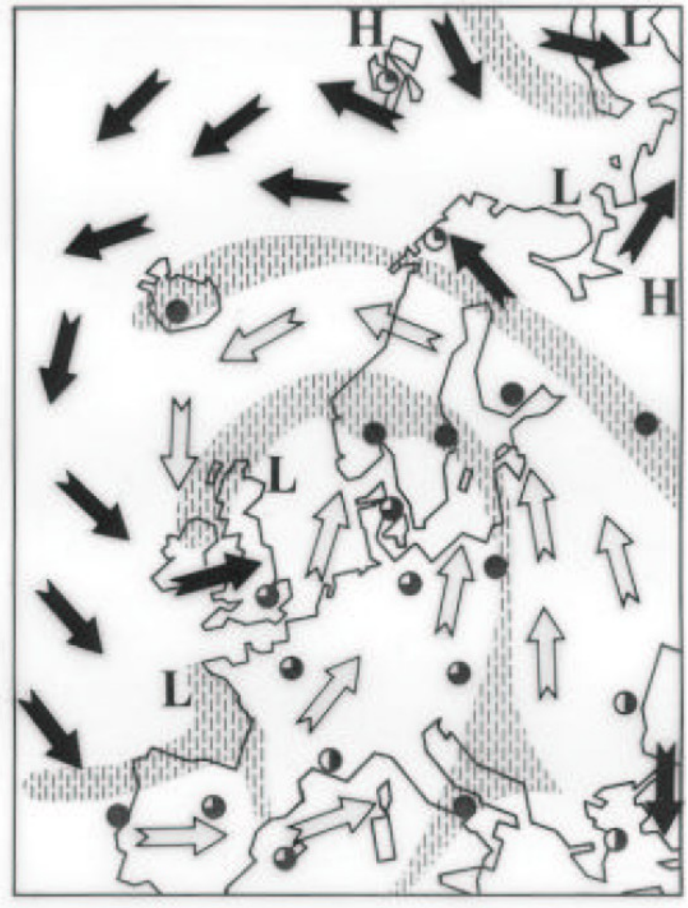

Figure 1.4: Wind velocity field over Europe on February 3, 1988

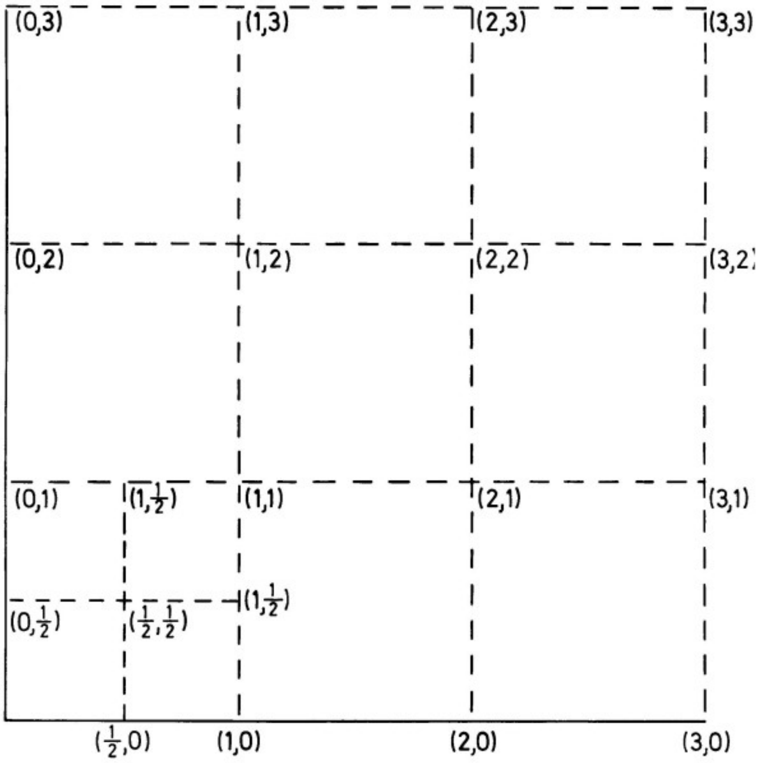
(0,3)
(1,3)
(2,3)
(3,3)
(0,2)
(1,2)
(2,2)
(3,2
(0,1)
(1,½)
(1,1)
(2,1)
(3,1)
(0,½)
(½,½)
(1,½)
(½,0)
(1,0)
(2,0)
(3,0)

Figure 1.5: Grid

definition of a coordinate system. The grid which is usually drawn on maps, or a globe, for example, is only of heuristic help, making it easier to read off the position of a point in the coordinate system.

As is well known by users of atlases, one can map a region by means of different sorts of coordinate systems. The particular coordinate system that we have described above, with vertical and horizontal straight axes, is called a Cartesian coordinate system, in honour of the great mathematician René Descartes (1596–1650). The horizontal axis is usually called the x-axis, and the vertical axis the y-axis. The position of a point is given by a number (x, y) relative to a chosen origin, which has coordinates $(0, 0)$.

By means of our coordinate-system we can specify the magnitudes of the vectors in our vector-field. In order to be able to write a vector as the sum of 'vector components' (see below) along the coordinate lines we introduce reference vectors (see above) along these lines. They are called basis vectors. The basis vectors are introduced at every point in the region of interest. Generally the basis vectors may have unequal directions and magnitudes at different positions (see for example section 4.1.1, where a system of plane polar coordinates is considered). Only in the special case of a Cartesian coordinate system (in a flat space) do they have the same magnitude and direction all over space.

At every point of an n-dimensional space there are n basis vectors. There are, for example, two basis vectors at every point of a surface. Having a coordinate system and the associated basis vector field at our disposal, we can specify the components of, for example, a velocity vector field at every point in a region. By this means we shall be able to develop the calculus of vectors.

1.4 Calculus of vectors. Two dimensions

As to notation, vectors shall be denoted by placing an arrow above the letter. Thus, 'A-vector' is denoted by $\vec{A}$. Basis vectors are denoted by a the letter 'e' with an arrow and a subscript, which tells which basis vector we are talking about, for example $\vec{e}_x$.

Consider a vector $\vec{A}$. In Fig. 1.6 this vector is placed in a two-dimensional coordinate-system. The coordinate axes are straight lines. The difference between their directions amounts to 90°. They are said to be *orthogonal* to each other. (This term will be used frequently in the present text.) We have a Cartesian coordinate system. The basis vectors along the x-axis and y-axis, are written $\vec{e}_x$ and $\vec{e}_y$, respectively. The magnitude of a vector $\vec{A}$ is designated by $|\vec{A}|$. The basis vectors $\vec{e}_x$ and $\vec{e}_y$ have per definition magnitude equal to 1, $|\vec{e}_x| = |\vec{e}_y| = 1$.

D -256- - -* ! - ' A ! 6 ! 7 !A 3
! D " H A 3 1"
(B @ ! ! 3 1 ! 3• 7! ! L " 1

So if you cannot really test such theories, what are they useful for?

Well . . . it's not really ruled out that they can't be tested but I think that any test has to come from cosmological data in one way or another. It is hard to test it at any other scale, in particular by scattering particles. Such cumulative processes could in the end add up to something significant but it would require some very detailed calculations to show that it would be indeed the case. Such scenarios have been supported by some people in a not so far past but I think that it is still preliminary. At the moment, it seems that the best hope of constructing and testing a consistent theory is to apply the theory to our current universe. As such, I do not see such theory as providing us with multiple predictions for the future in such a way that we can live long enough to verify them (laughs).

Since it is hard to test these theories, would it be reasonable for someone to stand up and say, "This is not really science but religion" because it's a matter of belief in a specific theory of quantum gravity?

Clearly, science might come to a limit in which this will be the case because in order to construct an accurate theory of the universe you need to conduct experiments. However, you don't really have the possibility of making many experiments since you just have this one universe. From this point of view, you might have to establish different kinds of standards for what should be called *science* and *scientific predictions*. At the moment, I don't consider this to be a problem. On the other hand, I must say that you will probably always be able to wave your hands in the air and use something like the anthropic principle to explain the workings of the universe. I tend not to refer to that approach as *science*.

Can you tell me what is the story behind the approach to quantum gravity that you pursue? How did it begin?

The approach that I pursue was developed by myself and collaborators, as well as other groups, in the mid-1980s in order to provide a non-perturbative regularisation of Polyakov's string theory [4–11] and it has been extremely successful in describing two-dimensional quantum gravity and non-critical string theory. In the beginning of the 1990s, it became clear that string theory was not going anywhere and it also became unclear whether it would be able to make any predictions. At that point, one had to think whether or not there could be other descriptions of quantum gravity. The generalisation of our approach to higher dimensions, which is related to what is referred to as *asymptotic safety* [12], was possible and could potentially lead to an alternative theory. Thus, we began pursuing it.

What is asymptotic safety?

Asymptotic safety is a kind of generalisation of Wilsonian standard renormalisation group philosophy according to which a quantum theory needs not to be renormalisable in the ordinary sense [12]. By a theory being *renormalisable* one usually means that it has a Gaussian

Why is this fixed point referred to as a non-Gaussian fixed point contrary to a Gaussian fixed point?

Yes, it's a little bit of a misnomer which stems from scalar theories. According to the definition we adopted with Max Niedermaier in an early review [6], what it means to be a non-Gaussian fixed point is that the critical exponents are not the canonical ones. Basically, it is synonymous to saying that, in the language of lattice field theory, you take the continuum limit at a fixed point different from the one of perturbation theory [11].

Now going back to these flows from the fixed point, I guess you are looking at flows that go from pure classical Einstein gravity to something else in the ultraviolet?

We don't start in the infrared. I mean, logically, the first step is to find the fixed point, then to find its finite-dimensional UV-critical hypersurface, and only later look for trajectories going downward to classical Einstein gravity.

But how many fixed points are out there and how many different trajectories exist that flow to the same infrared physics?

The problem is that in the approximate calculations that we can do, we cannot yet fully explore the full size of this hypersurface, so . . .

Even in truncated theory spaces?

If I generalise the truncation then I make the full space bigger. But a priori, I don't know how many dimensions this adds to the embedded critical surface. For instance, in $f(R)$ truncations, what happens is that up to R^2 you get relevant directions [3–5, 9]. So the cosmological constant, R and R^2 give rise to parameters in the surface and so it's three-dimensional. Adding R^3 and so on, you only get irrelevant directions. In this class of truncations, you can add more and more powers of the Ricci scalar and you will not increase the number of parameters you have to determine experimentally. So, fixing three parameters experimentally, you can predict all the other coefficients of all R powers. This has been discovered by Roberto Percacci's and Frank Saueressig's groups.

What is the reason for that?

This is just the intrinsic three-dimensionality of this hypersurface. If you look at what is the unstable manifold of the fixed point, then you see that in three directions you're driven away when you lower the cutoff, while in all others you're attracted to it. But now you could ask whether, if I add something like the Riemann tensor squared, will it add a relevant direction or an irrelevant one? A priori, I do not know and I have to do the calculation. Whenever you generalise a truncation, you may add a relevant or an irrelevant direction. We have some intuition from Gaussian fixed points for which canonical dimensions are still a guide. So you can partly understand whether specific terms give rise to relevant or irrelevant dimensions. However, strictly speaking, you cannot determine the exact number of dimensions from any given truncation as it is a property of the full theory.

& $ (%

20211011|||

BOOKS:

Gravity - Hartle

Gravity - Misner

- 20211010/ Gravity-Hartle/ Hartle-Gravity ...(2014).pdf

Pag. 16(19) - Triangle of Gauss.

- 20211011/ An Introduction to Calculus with Hyperbolic Functions.../ An_Introduction_to_Calculus...pdf

- 20211011/ Gravity-02/

Gravity Newtonian, Post-Newtonian, Relativistic.pdf

Pag. 09(25) - Proof that

$$\nabla^2 |x-x'|^{-1} = 4\pi\,\delta(x-x')$$

→ Gauss's theorem.

Pag. 309(325) - Box 6.5

Green's function for the wave equation

→ We can use Gauss's theorem to get

- 20211011/ Gravity-Hartle/ Misner-Gravitation...(1973).pdf

& $
4292 8 A A $,-4292 8 A A -) .
- -
! -256- - -* ! - $:! 7 !A 3 2 & "
= 0 7 ! 5 ! A 3 ! =" ! ? " ' 1"
-256- - -* ! - $:! 7 !A 3 2
& " = } 7 ! 5 ! A 3 D 1"
(B / 9 ! 7!

BOX 2.1 (*continued*)

unmanned spacecraft in 1973. Since 1969 a systematic program to determine the Moon's orbit using these devices has been carried out mainly at the McDonald Observatory at Mt. Locke, in Texas, and the Observatoire de Côte d'Azur station in Grasse, France. The lasers currently in use send pulses lasting 200 picoseconds, each containing about 10^{18} photons, about 10 times per second. Diffraction, refraction in the atmosphere, and other effects spread the beam over a 7-km radius on the Moon so that only 10^{-9} of the photons that are sent impinge on the retroreflector. On return, the reflected spot is spread over 20 km so that a 1-m telescope would detect only 10^{-9} of the returning photons. In the end, only one reflected photon is detected every few seconds. Returning photons have been detected for more than thirty years since 1970 at the time of writing.

some initial velocity, decelerates, reaches a maximum height, accelerates downward, and returns to the surface of the Earth. Any other body thrown upward from the same initial position with the same initial velocity would follow *exactly the same curve*.

This uniqueness of trajectory in space and time is a special property of gravity. The motion of a body in a magnetic field depends on what kind of charge it has. Bodies with one sign of charge will be deflected one way, bodies with the opposite charge will be deflected the other, and bodies with no charge will not be deflected at all. Only in a gravitational field do all bodies with the same initial conditions follow the same curve in space and time.

Einstein's idea was that this uniqueness of path in spacetime could be explained in terms of the geometry of the four-dimensional union of space and time called *spacetime*. Specifically, he proposed that the presence of a mass such as Earth curves the geometry of spacetime nearby, and that, in the absence of any other forces, all bodies move on the straight paths in this curved spacetime. Bodies free from forces move on straight lines of three-dimensional Euclidean space in Newtonian mechanics—that's part of Newton's first law. Einstein's idea is that the Earth moves in its orbit around the Sun, not because a force of gravity acts on it, but because it is following the straightest possible path in the slightly non-Euclidean geometry of spacetime produced by the Sun.

2.2 Experiments in Geometry

There is a story that in the late 1820s the great mathematician C. F. Gauss carried out an experiment to verify one of the standard theorems of the Euclidean geometry of space—that the interior angles of a triangle add up to 180°. Using the mountaintops of Hohenhagen, Brocken, and Inselsberg as vertices and assuming that light rays move on straight lines, he is said to have measured the angles, found the sum, and determined 180° to the accuracy with which the angles could be measured. (See Figure 2.2.)

The historical evidence is not conclusive as to whether Gauss actually carried out this experiment. However, he might have done so, and that emphasizes an important point. The sum of the angles was not guaranteed to be 180° from logic alone. Many geometries of physical space are possible that are different from

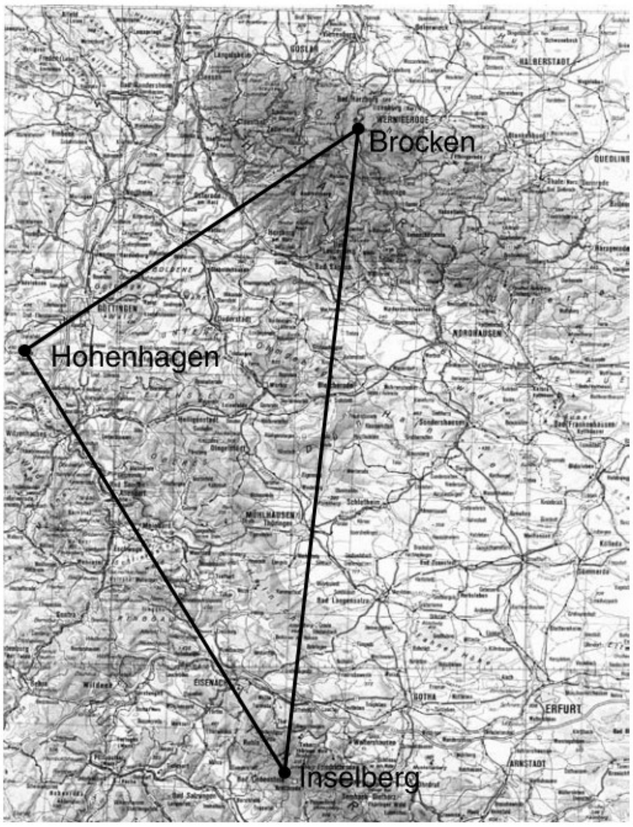

FIGURE 2.2 A modern map of Germany showing the locations of the peaks Hohenhagen, Brocken, and Inselberg, which form the vertices of a triangle that Gauss could have surveyed to check whether the interior angles add up to 180°, as predicted by Euclidean geometry.

Euclid's. These predict different results for the sum of the interior angles of a triangle. The geometry of space is an empirical question. It is a question in physics, subject to measurement, hypothesis, and test. By the end of this book you will know that if Gauss had been able to carry out his experiment with sufficient accuracy, he would have found a small difference in the sum of the angles just due to the mass of the Earth, $M_\oplus$, of order

$$\left|\begin{pmatrix}\text{sum of interior angles}\\ \text{of a triangle in radians}\end{pmatrix}\right| - \pi \sim \frac{(\text{area of triangle})}{R_\oplus^2}\left(\frac{GM_\oplus}{R_\oplus c^2}\right) \tag{2.1}$$

(where $R_\oplus$ is the Earth's radius) together with contributions from the Sun and the other planets. Note the appearance of the ratio GM/Rc^2, which is characteristic of weak relativistic effects, as discussed in Chapter 1. The distances between the mountains are 69 km, 85 km, and 107 km. Using these, this works out to be a difference of order 10^{-15} radians (!). So small a discrepancy could not be detected even with present technology, but modern experiments can detect the deviations from Euclidean geometry produced by the Sun and measure the geometry of space on the very large scales of cosmology. (See Box 2.2.)

BOX 2.2 Determining the Spatial Geometry of the Universe

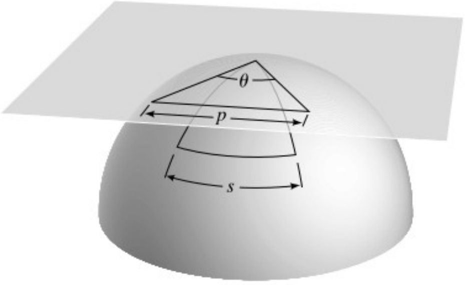

Modern measurements—not so different in character from that attributed to Gauss—determine the curvature of space on the distance scale of the visible universe. General relativity plus observations of the distribution of galaxies and radiation in the universe suggest only a few possibilities for the large-scale geometry of three-dimensional space at a moment of time, as we will see in detail in Chapter 18. The flat geometry of the plane, the positively curved geometry of the surface of a sphere, or the negatively curved geometry of a surface locally like some potato chips[a] are two-dimensional analogs of the possible flat, positively curved, and negatively curved large-scale geometries for three-dimensional space. How can the geometry of space in our universe be measured?

To understand a little of one method, imagine the geometry of space to be fixed in time (in contrast to the geometry of the actual universe, which is expanding). Imagine further that objects of known size p could be identified a known distance, d, away. If the geometry were flat like a plane, the angle θ subtended by these objects would be p/d. But, as illustrated in the accompanying figure, if the geometry were positively curved like the surface of a sphere, an object of smaller size s would subtend the same angle.[b] Alternatively, an object of a given size and distance away will subtend a larger angle on a positively curved surface like a sphere than it will in a flat plane (Problem 6). In a similar way, the angle subtended in a negatively curved space will be smaller. (For details, see Problem 18.12.) This discussion will be corrected to include the expansion of the universe in Chapter 19, but the qualitative result is the same: measuring the angular size of features of known size and distance is one way of determining whether the geometry of intervening space is flat, positively curved, or negatively curved. The cosmic background radiation provides such features.

The cosmic microwave background radiation (CMB) is light from the hot big bang that began the universe. The radiation started from the moment the universe had

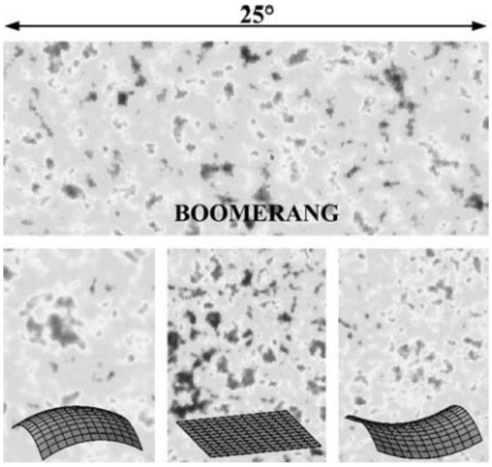

---

[a] "Crisps" in the UK and elsewhere.

[b] If that is not clear from thinking about the figure, imagine the sphere is made of rubber and could be flattened out on a plane tangent to the North Pole. The angles between lines of longitude at the North Pole do not change, but the equator and other lines of latitude have to be stretched. Thus the angle subtended by an object spanning a range of longitude on a sphere is the same as that subtended by a larger object in a plane.

BOX 2.2 (*continued*)

expanded and cooled enough for the matter to be transparent to radiation. It has been propagating freely to us over the intervening approximately 14 billion years. Were the universe unchanging in time, it would have traveled approximately 14 billion light-years. The temperature of the radiation has cooled to 2.73°K and is very nearly the same in all directions but not quite. Tiny temperature fluctuations of only a few tens of millionths of a degree are observed. The theory of the origin of these fluctuations predicts a spectrum of sizes that is characterized by a known length scale. The fluctuations are, therefore, features with a known spectrum of sizes a known distance away. Observations of their angular size can thus measure the spatial geometry of the universe. The right hand figure on the previous page shows a map of the temperature fluctuations in a 25°-wide region of the sky that were observed by the Boomerang experiment (deBernardis et al. 2000). The three figures on the bottom show simulations of what the map would look like based on the theoretical spectrum of original sizes if the geometry were positively curved (left), flat (middle), or negatively curved (right). Quantitative comparisons of the spectrum of angular sizes show that the geometry is very close to flat. (In the near future there will be a more accurate result, but the idea will be the same.) The geometry of space is a measurable, physical question.

2.3 Different Geometries

The idea of different geometries is easily illustrated in two dimensions. In your studies of the Euclidean geometry of the plane, you met the notions of point, straight line, distance, angle, parallel, triangle, circle, chord, etc. Familiar theorems include the one just discussed for a triangle:

$$\sum_{\text{vertices}} \begin{pmatrix}\text{interior}\\ \text{angle}\end{pmatrix} = \pi. \tag{2.2}$$

Another relates the ratio of the circumference to the radius of a circle:

$$\frac{(\text{circumference})}{(\text{radius})} = \frac{C}{r} = 2\pi. \tag{2.3}$$

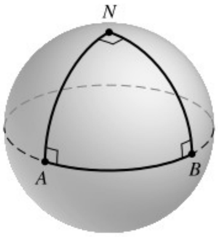

FIGURE 2.3 A spherical triangle NAB where the sum of the interior angles is 270°. The triangle consists of the parts of two lines of longitude 90° apart from the North Pole to the equator and the part of the equator between them. These are all segments of great circles and, therefore, straight lines in the geometry of the sphere.

The surface of a sphere provides an example of a *different* two-dimensional geometry in which such results of plane geometry are replaced by different theorems. Straight lines can be defined on a sphere as the shortest distance between two points, that is, as segments of great circles. Triangles are made up of three intersecting great circles. A circle is the locus of points equidistant (as measured on the surface) from a point which is its center, etc. For a spherical triangle of area A,

$$\sum_{\text{vertices}} \begin{pmatrix}\text{interior}\\ \text{angle}\end{pmatrix} = \pi + \frac{A}{a^2} \tag{2.4}$$

where a is the radius of the sphere.

Equation 2.4 shows that the sum of the interior angles of a spherical triangle is always *greater* than π. An example is shown in Figure 2.3. As the size of the triangle becomes smaller and smaller compared with the radius of curvature a, it

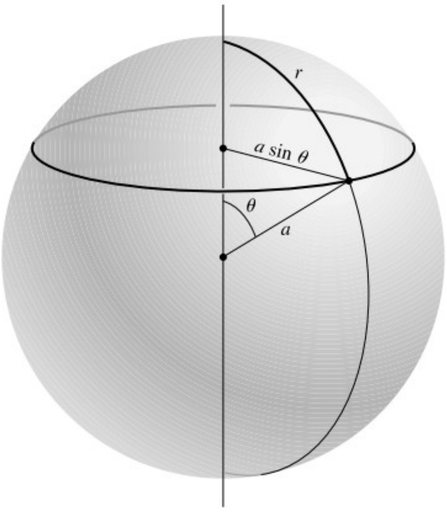

FIGURE 2.4 The relation of the circumference to the radius of a circle in the geometry of the surface of a sphere. A circle is the locus of points on the surface that are equidistant (as measured on the surface) from its center point. In this figure the North Pole has been chosen to coincide with the center of a circle that is a line of constant latitude labeled by θ. The radius r is the distance from the North Pole to this latitude measured along any line of constant longitude.

becomes increasingly difficult to tell the difference between a flat plane and the curved surface of the sphere. For triangles with very small areas ($A/a^2 \ll 1$), the result (2.4) is well approximated by the flat space result (2.2).

With the bit of geometry shown in Figure 2.4, the ratio of the circumference to the radius of a circle on a sphere can be calculated to be

$$\frac{\text{(circumference)}}{\text{(radius)}} = \frac{C}{r} = 2\pi \frac{\sin(r/a)}{(r/a)}. \tag{2.5}$$

Again, if $r \ll a$, the right-hand side reduces to the flat-space result (2.3).

It is not necessary to leave the surface of the earth to determine its geometry. Surveyors (such as Gauss) working on the surface of the earth can measure such things as the interior angles of a triangle and the circumference and radius of circles. By fitting to formulas such as (2.4) and (2.5), they could, in principle, tell if the geometry of the surface was spherical and determine the radius of curvature a. Similarly, by surveying in three dimensions we can, in principle, determine the geometry of space without needing any extra dimensions.

Visualization of three-dimensional curved geometries is not as easy as for two-dimensional curved geometries, which can often be represented as surfaces in Euclidean three-dimensional space. However some simple three-dimensional geometries can be thought of as curved surfaces in a hypothetical four-dimensional Euclidean space. For example, the three-dimensional geometry analogous to the two-dimensional sphere discussed before is the three-dimensional surface of a sphere in four dimensions—a *three-sphere*. If space had a fixed three-sphere geometry, a journey in a straight line in any direction would eventually bring one back to the starting point. However, more detailed information about the geometry of space can be determined locally. For example, it turns out[1] that the volume inside a two-dimensional sphere of radius r in such a spatial geometry is given by

$$V = 4\pi a^3 \left\{ \frac{1}{2} \sin^{-1}\left(\frac{r}{a}\right) - \frac{r}{2a}\left[1 - \left(\frac{r}{a}\right)^2\right]^{1/2} \right\} \tag{2.6a}$$

$$\approx \frac{4\pi r^3}{3}\left[1 + \begin{pmatrix} \text{corrections} \\ \text{of order } (r/a)^2 \end{pmatrix}\right], \qquad \text{for small } \frac{r}{a}, \tag{2.6b}$$

where a is the characteristic radius of curvature of the three-sphere geometry. For a two-sphere whose radius is much smaller than a, the volume-radius relation approaches the Euclidean flat-space result, as (2.6b) shows. If three-dimensional space had such a three-sphere geometry, the characteristic radius of curvature a could be determined by careful measurements of the radii and volume of two-spheres. As we will discover in Chapter 18, Einstein's theory predicts this three-sphere geometry as one possibility for the spatial geometry of a uniform universe on very large distance scales. Box 2.2 on p. 17 describes one effort to survey space on these scales.

2.4 Specifying Geometry

In addition to the geometry of the plane and the geometry of the sphere, there are an infinite number of other two-dimensional geometries. For example, there is the geometry on the surface of an egg or the geometry of the surface of a plane with a few hills on it. In three dimensions there are a similarly infinite number of geometries. How are these different geometries described and compared mathematically?

One way to describe a geometry is to embed it as a surface in a higher-dimensional space whose geometry is Euclidean. We have made use of this method in describing two-dimensional geometries as surfaces embedded in three-dimensional Euclidean geometry—planes, spheres, eggs, etc. However, it becomes almost impossible to think of any but the simplest three- and four-dimensional geometries as surfaces in four and five dimensions. Further, the extra dimension is physically superfluous. An *intrinsic* description of geometry that makes use of just the physical dimensions that can be measured is what is needed.

Another idea is to specify a geometry by giving a small number of *axioms*, or postulates, from which the other results of geometry can be derived as theorems. For the geometry of the flat plane, for example, there are Euclid's five axioms: Two points determine a unique line, parallel lines never intersect, etc. Some other simple geometries can be characterized in this way with different axioms. For example, the geometry on the surface of a sphere can be summarized by a set of axioms like Euclid's in which the parallel postulate is replaced by the axiom that two parallel lines *always* meet in two points. However, this method also is limited. What axioms describe the geometry on the surface of a potato? We need a more local and detailed description.

---

[1] This result will be derived explicitly in Example 7.6.

The key to a general description of geometry is to use differential and integral calculus to reduce all geometry to a specification of the distance between each pair of nearby points. From the distance between nearby points, the distances along curves can be built up by integration. Straight lines are the curves of the shortest distance between two points. Angles are ratios of the lengths of arcs to their radii when those radii are small. Areas, volumes, etc., can be constructed by multiple integrals over area and volume elements, themselves specified by the distances between nearby points. By specifying the distances between nearby points and using differential and integral calculus, the most general geometry may be specified. This area of mathematics is called *differential geometry*. We will explore just a few ideas of this subject in the next section.

2.5 Coordinates and Line Element

The Euclidean Geometry of a Plane

A systematic way of labeling points is a prerequisite to a specification of the distance between nearby ones. A system of coordinates assigns unique labels to each point, and there are many systems that do so. In two dimensions, for instance, there are Cartesian coordinates (x, y), polar coordinates (r, ϕ) about some origin, etc. (Figure 2.5).

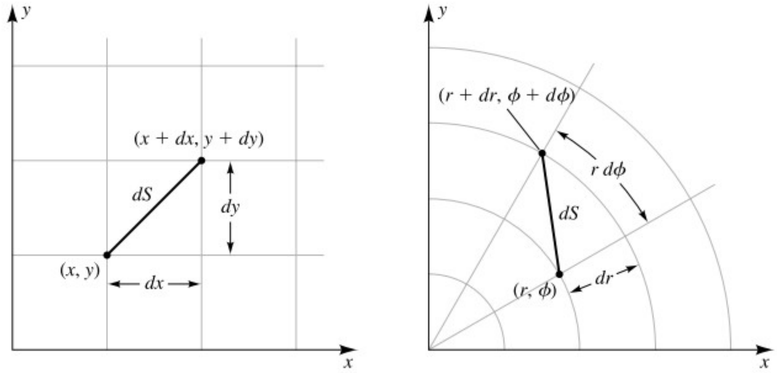

FIGURE 2.5 Cartesian and polar coordinates. Cartesian and polar coordinates are both systematic ways of labeling points in the plane, and the distance between nearby points can be expressed in terms of either.

Nearby points have nearby values of their coordinates. For example, the points (x, y) and $(x+dx, y+dy)$ are nearby when dx and dy are infinitesimal. Similarly, (r, ϕ) and $(r + dr, \phi + d\phi)$ are nearby.

In Cartesian coordinates (x, y), the distance dS between the points (x, y) and $(x + dx, y + dy)$ is (see Figure 2.5)

$$dS = \left[(dx)^2 + (dy)^2\right]^{1/2}. \tag{2.7}$$

The same rule can be expressed in polar coordinates where the distance between the nearby points (r, ϕ) and $(r + dr, \phi + d\phi)$ is (see Figure 2.5)

$$dS = \left[(dr)^2 + (r\, d\phi)^2\right]^{1/2}. \tag{2.8}$$

Expression (2.8) and others like it are valid only if dr and $d\phi$ are small. However, large distances can be built up from these infinitesimal relations by integration. Let's, for example, calculate the ratio of the circumference to the diameter of a circle of radius R. Choosing the origin at the center, the equation for such a circle in Cartesian coordinates is

$$x^2 + y^2 = R^2. \tag{2.9}$$

The circumference C is the integral of dS around the circle. Using (2.7) this is

$$C = \oint dS = \oint \left[(dx)^2 + (dy)^2\right]^{1/2} \tag{2.10a}$$

$$= 2\int_{-R}^{+R} dx \left[1 + \left(\frac{dy}{dx}\right)^2\right]^{1/2}_{x^2+y^2=R^2} \tag{2.10b}$$

$$= 2\int_{-R}^{+R} dx \sqrt{\frac{R^2}{R^2 - x^2}}. \tag{2.10c}$$

Changing variables by writing $x = R\xi$, we have

$$C = 2R\int_{-1}^{1} \frac{d\xi}{\sqrt{1-\xi^2}} = 2\pi R. \tag{2.11}$$

This is the correct answer. The integral could even be taken to define π; by doing it numerically, one could discover that $\pi = 3.1415926535\ldots$.

Deriving the relation between radius and circumference is even easier in polar coordinates, where the equation of the circle is just $r = R$. Evaluating (2.8) on the circle and integrating the resulting dS over it gives

$$C = \oint dS = \int_0^{2\pi} R\, d\phi = 2\pi R. \tag{2.12}$$

The ease of using polar coordinates to arrive at (2.12) shows that, for a given problem, some coordinates are better than others.

By proceeding in this way we could derive all the theorems of Euclidean plane geometry. The angle between two intersecting lines, for example, can be defined as the ratio of the length ΔC of the part of a circle centered on their intersection that lies between the lines to the circle's radius R.

$$\theta \equiv \frac{\Delta C}{R} \qquad \text{(radians)}. \tag{2.13}$$

With this definition we could prove that the sum of the interior angles of a triangle is π. Indeed, we could verify the *axioms* of Euclidean plane geometry from (2.7) or (2.8). All geometry can be reduced to relations between distances; all distances can be reduced to integrals of distances between nearby points; all Euclidean plane geometry is contained in (2.7) or (2.8).

Line Element

To summarize, a geometry is specified by the *line element*, such as (2.7) or (2.8), which gives the distance between nearby points in terms of the coordinate intervals between them in some coordinate system. Conventionally, a line element is written as a quadratic relation for dS^2, e.g.,

$$dS^2 = dx^2 + dy^2 \tag{2.14}$$

with no brackets around the differentials. The form of the line element for a geometry varies from coordinate system to coordinate system [e.g., (2.7) and (2.8)], but the geometry remains the same.

The Non-Euclidean Geometry of a Sphere

An example of a non-Euclidean geometry is provided by the surface of a two-dimensional sphere of radius a. We can use the angles (θ, ϕ) of three-dimensional polar coordinates to label points on the sphere. The distance between points (θ, ϕ) and $(\theta + d\theta, \phi + d\phi)$ can be seen after a little work (Figure 2.6) to be

Line Element for a Sphere

$$dS^2 = a^2\left(d\theta^2 + \sin^2\theta\, d\phi^2\right). \tag{2.15}$$

This is the line element of the surface of a sphere.

Let's use the line element (2.15) to calculate the ratio of the circumference to the radius of a circle on the sphere. By circle we mean the locus of points *on the surface* that are a constant distance (the radius) *along the surface* from a fixed point (the center) *in the surface*. Since no one point is distinguished geometrically from any other on the sphere, we may conveniently orient our polar coordinate system so that the polar axis is at the center of the circle. A circle is then a curve of constant θ. Consider the circle defined by the equation

$$\theta = \Theta \tag{2.16}$$

for constant Θ. The circumference is the distance around this curve. Nearby points along the curve are separated by $d\phi$ but have $d\theta = 0$. Thus, (2.15) gives $dS =$

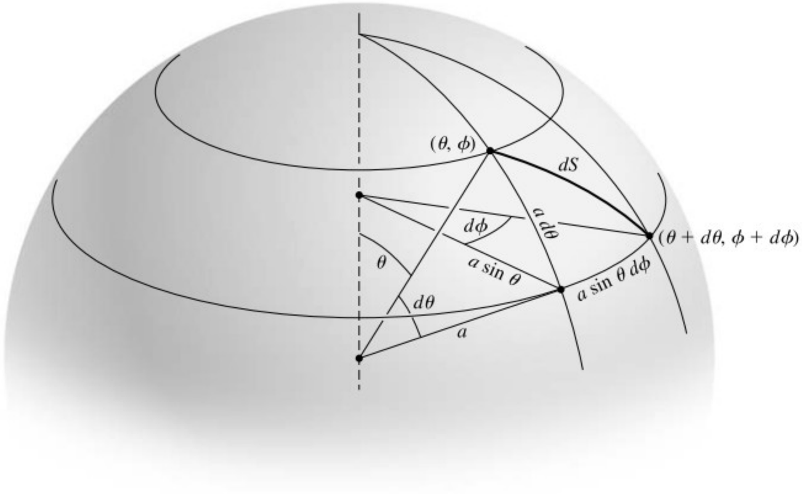

FIGURE 2.6 Deriving the line element on the sphere. The derivation makes use of the fact that the two-dimensional sphere is a surface in three-dimensional Euclidean space. Two infinitesimally separated points at locations (θ, ϕ) and $(\theta+d\theta, \phi+d\phi)$ are indicated. The construction shows that the distance between ϕ and $\phi + d\phi$ along a line of constant latitude θ is $a \sin\theta\, d\phi$. The distance between θ and $\theta+d\theta$ along a line of constant longitude is $a\, d\theta$. Because the θ and ϕ coordinate lines are orthogonal, the sum of the squares of these two differentials gives the square of the distance dS between the two points when $d\theta$ and $d\phi$ are infinitesimally small. This gives (2.15).

$a \sin\Theta\, d\phi$ along the circle, and the circumference is

$$C = \oint dS = \int_0^{2\pi} a \sin\Theta\, d\phi = 2\pi a \sin\Theta. \tag{2.17}$$

The radius is the distance from the center to the circle along a curve for which θ varies but $d\phi = 0$. Along this curve, (2.15) gives $dS = a\, d\theta$, and the radius is

$$r = \int_{\text{center}}^{\text{circle}} dS = \int_0^{\Theta} a\, d\theta = a\Theta. \tag{2.18}$$

Using (2.18) to eliminate Θ in (2.17), the relation between the circumference and radius of a circle in the non-Euclidean geometry of a sphere becomes

$$C = 2\pi a \sin\left(\frac{r}{a}\right). \tag{2.19}$$

In this expression a is a fixed number characterizing the geometry. It measures the scale on which the geometry is curved. When the radius of the circle is much smaller than the radius of the sphere, $r \ll a$, then we have approximately

$$C \approx 2\pi r, \tag{2.20}$$

which is the familiar result in Euclidean geometry. The geometry of the surface of the Earth is the same as a sphere to a good approximation.

The many different projections used to make maps of its surface are just different coordinate systems for expressing the geometry of a sphere as described in Box 2.3.

BOX 2.3 Map Projections

The various projections used to make planar maps of the Earth's surface are examples of a familiar geometry expressed in different systems of coordinates. The geometry is that of the two-dimensional surface of the sphere to an excellent approximation. In usual polar coordinates the line element is given by (2.15), with a being the radius of the Earth. On the surface of the Earth, the angle ϕ is the longitude (measured in radians rather than degrees). The latitude λ is $\pi/2 - \theta$. Expressed in terms of latitude and longitude, the line element is

$$dS^2 = a^2(d\lambda^2 + \cos^2\lambda\, d\phi^2). \tag{a}$$

To make a map we introduce new coordinates x and y on the sphere, defined by relations of the form

$$x = x(\lambda, \phi), \qquad y = y(\lambda, \phi), \tag{b}$$

and use these as rectangular coordinates in the plane to plot the outlines of the continents, locations of cities, etc. Different projections correspond to different choices for the functions $x(\lambda, \phi)$ and $y(\lambda, \phi)$. One can think of these functions as providing a map in the mathematical sense from the sphere into the plane.

There are as many projections as there are different functions. The simplest example is

$$x = (L\phi)/2\pi, \qquad y = (L\lambda)/\pi, \tag{c}$$

where L is the width of the map. This just plots ϕ and λ as x and y on rectangular axes. The result, shown in the accompanying figure, is called an *equirectangular* projection. However, there are more useful projections which preserve some properties of the geometry of the sphere on a plane. Not all properties can be preserved because the geometry of a sphere is different from that of the plane!

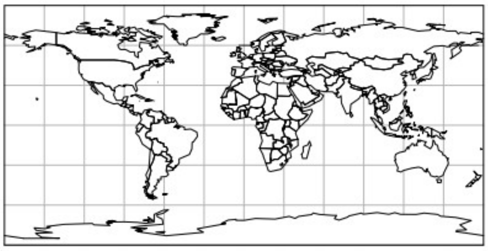

Equirectangular projection.

A wide class of useful projections send longitude linearly into x:

$$x = \frac{L\phi}{2\pi}, \qquad y = y(\lambda). \tag{d}$$

For projections of this kind, the true distances are given by the line element

$$dS^2 = a^2\left[\left(\frac{2\pi}{L}\cos[\lambda(y)]\right)^2 dx^2 + \left(\frac{d\lambda}{dy}\right)^2 dy^2\right]. \tag{e}$$

A simple example of a projection of this kind is the Mercator projection, invented by G. Kramer in 1569 and illustrated below. Kramer's idea was that angles on the map should equal compass bearings on the sphere. That is, the map from the sphere to the plane should preserve angles between different directions from a point. A mariner wishing to sail between Caracas and Lisbon would draw the straight line on the map connecting these two ports. The angle between that line and the y-axis would be the bearing from north that when held constant during the voyage would bring the ship from Caracas

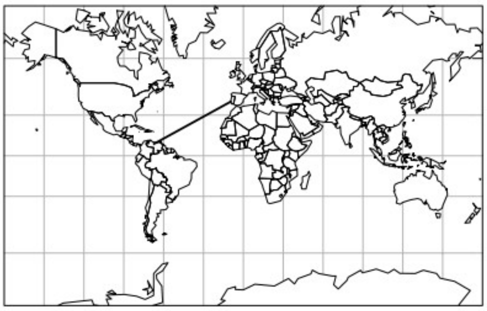

Mercator projection.

to Lisbon. What choice of function $y(\lambda)$ or $\lambda(y)$ would make a map like this?

Angles are *ratios* of distances, as we saw in (2.13). The angle between two directions on a sphere will equal the angle between the corresponding directions on the plane if the line element on the sphere is *proportional* to the line element on the flat plane, $dS^2_{\text{flat}} = dx^2 + dy^2$. Thus, to implement Kramer's idea we seek a function $\lambda(y)$ such that (e) can be written

$$dS^2 = \Omega^2(x, y)(dx^2 + dy^2) \tag{f}$$

for some function $\Omega(x, y)$. Clearly we need

$$\frac{d\lambda}{dy} = \frac{2\pi}{L}\cos\lambda. \tag{g}$$

Choosing $y = 0$ to coincide with $\lambda = 0$ gives

$$y(\lambda) = \frac{L}{2\pi}\int_0^\lambda \frac{d\lambda'}{\cos\lambda'} = \frac{L}{2\pi}\log\left[\tan\left(\frac{\pi}{4} + \frac{\lambda}{2}\right)\right]. \tag{h}$$

Equations (h) and (d) define the Mercator projection. The equator $\lambda = 0$ is mapped to the line $y = 0$. The poles $\lambda = \pm\pi/2$ are mapped to $y = \pm\infty$, respectively.

The proportionality factor between the spherical metric and the flat metric on the plane $\Omega(x, y)$ that was defined in (f) is

$$\Omega(y) = \frac{2\pi a}{L}\cos\lambda(y) = \frac{4\pi a}{L}\frac{e^{2\pi y/L}}{1 + e^{4\pi y/L}}. \tag{i}$$

Most of the familiar properties of the Mercator projection follow from this factor. For example, consider two points at the same latitude separated by a difference in longitude, Δx. The physical distance between these points, ΔS, is

$$\Delta S = \Omega(y)\Delta x \tag{j}$$

and depends on latitude. As $y \to \infty$, the North Pole, this distance shrinks to zero, as it should. True distances in x at higher latitudes are smaller than coordinate distances because of the factor $\Omega(y)$.

The same holds true for areas. A small rectangle on the map of coordinate dimensions Δx and Δy has area

$$\Delta A = [\Omega(y)\Delta x][\Omega(y)\Delta y] = \Omega^2(y)\Delta x\Delta y. \tag{k}$$

Thus, although Greenland looms large on the Mercator projection in coordinate area when compared with South Americal, for example, its actual area is much smaller.

The Geometry of Some More General Surfaces

The line element of the plane and the sphere were worked out before, starting from a clear picture of these geometries as surfaces in Euclidean space. However, in general relativity it is more usual to be confronted with a line element and have to figure out the properties of the geometry it represents.

Consider as an example the line element

$$dS^2 = a^2(d\theta^2 + f^2(\theta)\,d\phi^2) \tag{2.21}$$

for various possible choices of the function $f(\theta)$. The choice $f(\theta) = \sin\theta$ gives the geometry of the surface of a sphere (2.15). But what surfaces in three-

D -256- - -* ! - $
2 I& " I I ! I@ I:31 D I< I? I4 A! A I! "I 1"

-256- - -* ! - $ 7 !A 3 8 ! J 8 ! J
5 ! A 1"
(/ $! E 7! 0

where

$$\nabla^2 := \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{1.9}$$

is the familiar Laplacian operator; Poisson's equation (known after its originator Siméon Denis Poisson, who unfortunately is not related to either author of this book), is the generalization of Eq. (1.4b) to continuous matter.

As was stated previously, these equations will be properly introduced in the following two sections. To complete the formulation of the theory we must impose a relationship between the pressure and the density of the fluid. This relationship, known as the *equation of state*, takes the general form of

$$p = p(\rho, T, \cdots), \tag{1.10}$$

in which the pressure is expressed as a function of the density, temperature, and possibly other relevant variables such as chemical composition. The equation of state encodes information about the microphysics that governs the fluid, and this information must be provided as an input in most applications of the theory.

A complete description of a physical situation involving gravity and a distribution of matter can be obtained by integrating Eqs. (1.5), (1.6), and (1.8) simultaneously and self-consistently. The solutions must be subjected to suitable boundary conditions, which will be part of the specification of the problem. All of Newtonian gravity is contained in these equations, and all associated phenomena follow as consequences of these equations.

1.3 Newtonian field equation

In this section we examine the equations that govern the behavior of the gravitational field, and show how Eq. (1.8) is an appropriate generalization of the more primitive form of Eq. (1.4b).

We recall that the relation $U = GM/r$ applies to a point body of active gravitational mass M situated at the origin of the coordinate system. Suppose that we are given an arbitrary number N of point bodies, and that we assign to each one a label $A = 1, 2, \cdots, N$. The mass and position of each body are then denoted M_A and $\boldsymbol{r}_A(t)$, respectively. If we *assume* that the total Newtonian potential U is a linear superposition of the individual potentials U_A created by each body, we have that the potential at position $\boldsymbol{x}$ is given by

$$U = \sum_A U_A = G \sum_A \frac{M_A}{|\boldsymbol{x} - \boldsymbol{r}_A|}. \tag{1.11}$$

The generalization of this relation to a continuous distribution of matter is straightforward. We convert the discrete sum $\sum_A M_A$ to a continuous integral $\int d^3x' \rho(t, \boldsymbol{x}')$, and we replace the discrete positions $\boldsymbol{r}_A$ with the continuous integration variable $\boldsymbol{x}'$. The result is

$$U(t, \boldsymbol{x}) = G \int \frac{\rho(t, \boldsymbol{x}')}{|\boldsymbol{x} - \boldsymbol{x}'|} d^3x', \tag{1.12}$$

one of the key defining equations for the Newtonian potential. The integral can be evaluated as soon as the density field $\rho(t, \boldsymbol{x}')$ is specified, regardless of whether ρ is a proper solution to the remaining fluid equations. As such, Eq. (1.12) gives U as a *functional* of an arbitrary function ρ. The potential, however, will be physically meaningful only when ρ itself is physically meaningful, which means that it must be a proper solution to the continuity and Euler equations.

The integral equation (1.12) can easily be transformed into a differential equation for the Newtonian potential U. The transformation relies on the identity

$$\nabla^2 \frac{1}{|\boldsymbol{x} - \boldsymbol{x}'|} = -4\pi \delta(\boldsymbol{x} - \boldsymbol{x}') , \tag{1.13}$$

in which $\delta(\boldsymbol{x} - \boldsymbol{x}') := \delta(x - x')\delta(y - y')\delta(z - z')$ is a three-dimensional delta function defined by the properties

$$\delta(\boldsymbol{x} - \boldsymbol{x}') = 0 \qquad \text{when } \boldsymbol{x} \neq \boldsymbol{x}', \tag{1.14a}$$

$$f(\boldsymbol{x})\delta(\boldsymbol{x} - \boldsymbol{x}') = f(\boldsymbol{x}')\delta(\boldsymbol{x} - \boldsymbol{x}') \qquad \text{for any smooth function } f(\boldsymbol{x}), \tag{1.14b}$$

$$\int \delta(\boldsymbol{x} - \boldsymbol{x}')\, d^3x' = 1 \qquad \text{for any domain of integration that encloses } \boldsymbol{x}. \tag{1.14c}$$

These properties further imply that $\delta(\boldsymbol{x}' - \boldsymbol{x}) = \delta(\boldsymbol{x} - \boldsymbol{x}')$. The identity of Eq. (1.13) is derived in Box 1.2. If we apply the Laplacian operator on both sides of Eq. (1.12) and exchange the operations of integration and differentiation on the right-hand side, we obtain

$$\begin{aligned} \nabla^2 U &= G \int \rho(t, \boldsymbol{x}') \nabla^2 \frac{1}{|\boldsymbol{x} - \boldsymbol{x}'|}\, d^3x' \\ &= -4\pi G \int \rho(t, \boldsymbol{x}')\delta(\boldsymbol{x} - \boldsymbol{x}')\, d^3x' \\ &= -4\pi G\rho(t, \boldsymbol{x}) ; \end{aligned}$$

the identity was used in the second step, and the properties of the delta function displayed in Eq. (1.14) allowed us to evaluate the integral. The end result is Poisson's equation,

$$\nabla^2 U = -4\pi G\rho , \tag{1.15}$$

whose formulation was anticipated in Eq. (1.8).

It is possible to proceed in the opposite direction, and show that Eq. (1.12) provides a solution to Poisson's equation (1.15). A powerful tool in the integration of differential equations is the *Green's function* $G(\boldsymbol{x}, \boldsymbol{x}')$, a function of a field point $\boldsymbol{x}$ and a source point $\boldsymbol{x}'$. In the specific context of Poisson's equation, the Green's function is required to be a solution to

$$\nabla^2 G(\boldsymbol{x}, \boldsymbol{x}') = -4\pi \delta(\boldsymbol{x} - \boldsymbol{x}') , \tag{1.16}$$

which is recognized as a specific case of the general differential equation, corresponding to a point mass situated at $\boldsymbol{x}'$. Armed with such an object, a formal solution to Eq. (1.15) can be expressed as

$$U(t, \boldsymbol{x}) = G \int G(\boldsymbol{x}, \boldsymbol{x}')\rho(t, \boldsymbol{x}')\, d^3x' ; \tag{1.17}$$

Box 1.2

Proof that $\nabla^2|\boldsymbol{x}-\boldsymbol{x}'|^{-1} = -4\pi\,\delta(\boldsymbol{x}-\boldsymbol{x}')$

To simplify the proof of Eq. (1.13) we set $\boldsymbol{x}' = \boldsymbol{0}$ without loss of generality; this can always be achieved by a translation of the coordinate system. This gives rise to the simpler equation

$$\nabla^2 r^{-1} = -4\pi\,\delta(\boldsymbol{x})\,, \tag{1}$$

with $r := |\boldsymbol{x}|$.

We first show that $\nabla^2 r^{-1} = 0$ whenever $\boldsymbol{x} \neq \boldsymbol{0}$. Derivatives of r^{-1} can be evaluated with the help of the identities

$$\frac{\partial r}{\partial x^j} = n_j\,, \qquad \frac{\partial n_j}{\partial x^k} = \frac{\partial n_k}{\partial x^j} = \frac{1}{r}\big(\delta_{jk} - n_j n_k\big)\,,$$

where $x^j := (x, y, z)$ is a component notation for the vector $\boldsymbol{x}$, $n^j := x^j/r$, and δ_{jk} is the Kronecker delta, equal to one when $j = k$ and zero otherwise. These equations hold provided that $r \neq 0$. According to this we have that

$$\frac{\partial}{\partial x^j} r^{-1} = -\frac{1}{r^2} n_j$$

and

$$\frac{\partial^2}{\partial x^j \partial x^k} r^{-1} = \frac{1}{r^3}\big(3n_j n_k - \delta_{jk}\big)\,.$$

Because $\boldsymbol{n}$ is a unit vector, it follows that $\nabla^2 r^{-1} = 0$ whenever $r \neq 0$.

To handle the special case $r = 0$ we introduce the vector $\boldsymbol{j} := \boldsymbol{\nabla} r^{-1}$ and write the left-hand side of Eq. (1) as $\boldsymbol{\nabla} \cdot \boldsymbol{j}$. Integrating this over a volume V bounded by a spherical surface S of radius η, we obtain

$$\int_V \boldsymbol{\nabla} \cdot \boldsymbol{j}\, d^3x = \oint_S \boldsymbol{j} \cdot d\boldsymbol{S}$$

by virtue of Gauss's theorem. Here $d\boldsymbol{S}$ is an outward-directed surface element on S, which can be expressed as $d\boldsymbol{S} = \boldsymbol{n}\eta^2\, d\Omega$, with $d\Omega$ denoting an element of solid angle centered at $\boldsymbol{n}$. The vector $\boldsymbol{j}$ is equal to $-\eta^{-2}\boldsymbol{n}$ on S, and evaluating the surface integral returns -4π.

Because $\nabla^2 r^{-1}$ vanishes when $\boldsymbol{x} \neq 0$ and integrates to -4π whenever the integration domain encloses $\boldsymbol{x} = \boldsymbol{0}$, we conclude that it is distributionally equal to $-4\pi\,\delta(\boldsymbol{x})$. The proof is complete.

the steps involved in establishing that this U is indeed a solution to Poisson's equation are identical to those that previously led us to Eq. (1.15) from Eq. (1.12). The difference is that in the earlier derivation the identity of the Green's function was already known. In the approach described here, the result follows simply by virtue of Eq. (1.16). It is not difficult, of course, to identify the Green's function: comparison with Eq. (1.13) allows us to write

$$G(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{|\boldsymbol{x} - \boldsymbol{x}'|}\,. \tag{1.18}$$

Not surprisingly, the Green's function represents the potential of a point mass situated at $\boldsymbol{x}'$.

" -256- - -* ! - $ 7 !A ! @ : < !
/ % 1"
(% $/ A!

Figure 1.10.
Curvature as manifested in the "acceleration of the separation" of two nearby geodesics. Two geodesics, originally parallel, and separated by the distance ("geodesic deviation") ξ_0, are no longer parallel when followed a distance s. The separation is $\xi = \xi_0 \cos\phi = \xi_0 \cos(s/a)$, where a is the radius of the sphere. The separation follows the equation of simple harmonic motion, $d^2\xi/ds^2 + (1/a^2)\,\xi = 0$ ("equation of geodesic deviation").

The direction of the separation vector, $\boldsymbol{\xi}$, is fixed fully by its orthogonality to the fiducial geodesic. Hence, no reference to the direction of $\boldsymbol{\xi}$ is needed or used in the equation of geodesic deviation: only the magnitude ξ of $\boldsymbol{\xi}$ appears there, and only the magnitude, not direction, of the relative acceleration appears.

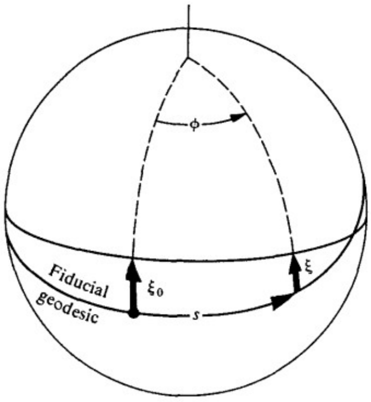

In a space of more than two dimensions, an equation of the same general form applies, with several differences. In two dimensions the *direction* of acceleration of one geodesic relative to a nearby, fiducial geodesic is fixed uniquely by the demand that their separation vector, $\boldsymbol{\xi}$, be perpendicular to the fiducial geodesic (see Figure 1.10). Not so in three dimensions or higher. There $\boldsymbol{\xi}$ can remain perpendicular to the fiducial geodesic but rotate about it (Figure 1.11). Thus, to specify the relative acceleration uniquely, one must give not only its magnitude, but also its direction.

The relative acceleration in three dimensions and higher, then, is a vector. Call it "$D^2\boldsymbol{\xi}/ds^2$," and call its four components "$D^2\xi^\alpha/ds^2$." Why the capital D? Why not "$d^2\xi^\alpha/ds^2$"? Because our coordinate system is completely arbitrary (cf. §1.2). The twisting and turning of the coordinate lines can induce changes from point to point in the components ξ^α of $\boldsymbol{\xi}$, even if the vector $\boldsymbol{\xi}$ is not changing at all. Consequently, the accelerations of the components $d^2\xi^\alpha/ds^2$ are generally not equal to the components $D^2\xi^\alpha/ds^2$ of the acceleration!

How, then, in curved spacetime can one determine the components $D^2\xi^\alpha/ds^2$ of the relative acceleration? By a more complicated version of the equation of geodesic deviation (1.6). Differential geometry (Part III of this book) provides us with a geometrical object called the *Riemann curvature tensor*, "***Riemann***." ***Riemann*** is

Curvature is characterized by Riemann tensor

(*continued on page 34*)

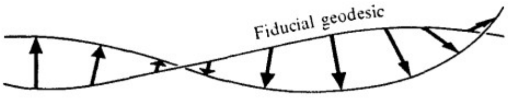

Figure 1.11.
The separation vector $\boldsymbol{\xi}$ between two geodesics in a curved three-dimensional manifold. Here $\boldsymbol{\xi}$ can not only change its length from point to point, but also rotate at a varying rate about the fiducial geodesic. Consequently, the relative acceleration of the geodesics must be characterized by a direction as well as a magnitude: it must be a vector, $D^2\boldsymbol{\xi}/ds^2$.

Box 1.6 CURVATURE OF WHAT?

Nothing seems more attractive at first glance than the idea that gravitation is a manifestation of the curvature of space (A), and nothing more ridiculous at a second glance (B). How can the tracks of a ball and of a bullet be curved so differently if that curvature arises from the geometry of space? No wonder that great Riemann did not give the world a geometric theory of gravity. Yes, at the age of 28 (June 10, 1854) he gave the world the mathematical machinery to define and calculate curvature (metric and Riemannian geometry). Yes, he spent his dying days at 40 working to find a unified account of electricity and gravitation. But if there was one reason more than any other why he failed to make the decisive connection between gravitation and curvature, it was this, that he thought of space and the curvature of space, not of spacetime and the curvature of spacetime. To make that forward step took the forty years to special relativity (1905: time on the same footing as space) and then another ten years (1915: general relativity). Depicted in spacetime (C), the tracks of ball and bullet appear to have comparable curvature. In fact, however, neither track has any curvature at all. They both look curved in (C) only because one has forgotten that the spacetime they reside in is itself curved—curved precisely enough to make these tracks the straightest lines in existence ("geodesics").

If it is at first satisfying to see curvature, and curvature of spacetime at that, coming to the fore in so direct a way, then a little more reflection produces a renewed sense of concern. Curvature with respect to what? Not with respect to the labo-

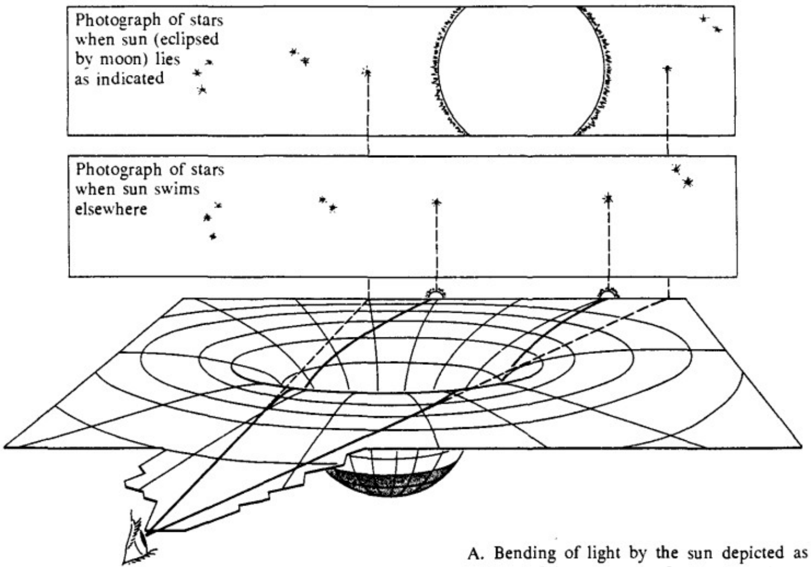
Photograph of stars when sun (eclipsed by moon) lies as indicated
Photograph of stars when sun swims elsewhere

A. Bending of light by the sun depicted as a consequence of the curvature of space near the sun. Ray of light pursues geodesic, but geometry in which it travels is curved (actual travel takes place in spacetime rather than space; correct deflection is twice that given by above elementary picture). Deflection inversely proportional to angular separation between star and center of sun. See Box 40.1 for actual deflections observed at time of an eclipse.

ratory. The earth-bound laboratory has no simple status whatsoever in a proper discussion. First, it is no Lorentz frame. Second, even to mention the earth makes one think of an action-at-a-distance version of gravity (distance from center of earth to ball or bullet). In contrast, it was the whole point of Einstein that physics looks simple only when analyzed locally. To look at local physics, however, means to compare one geodesic of one test particle with geodesics of other test particles traveling (1) nearby with (2) nearly the same directions and (3) nearly the same speeds. Then one can "look at the separations between these nearby test particles and from the second time-rate of change of these separations and the 'equation of geodesic deviation' (equation 1.8) read out the curvature of spacetime."

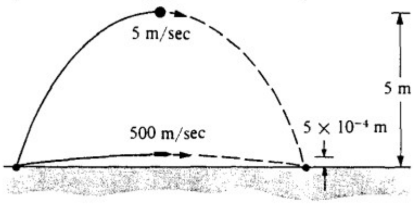
10 m
5 m/sec
500 m/sec
5×10^{-4} m
5 m

B. Tracks of ball and bullet through space as seen in laboratory have very different curvatures.

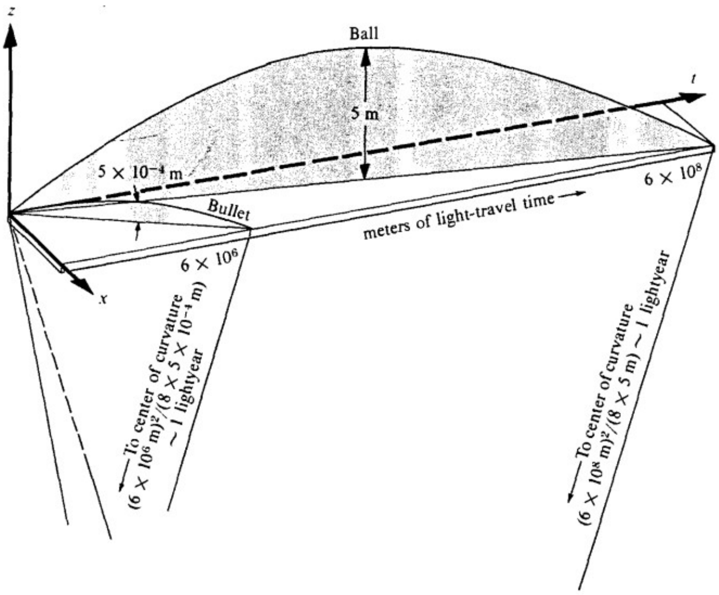
z
Ball
t
5 m
5×10^{-4} m
Bullet
6×10^8
meters of light-travel time →
6×10^6
x
← To center of curvature $(6 \times 10^6\ \text{m})^2/(8 \times 5 \times 10^{-4}\ \text{m}) \sim 1$ lightyear
← To center of curvature $(6 \times 10^8\ \text{m})^2/(8 \times 5\ \text{m}) \sim 1$ lightyear

C. Tracks of ball and bullet through spacetime, as recorded in laboratory, have comparable curvatures. Track compared to arc of circle: (radius) = (horizontal distance)^2/8 (rise).

the higher-dimensional analog of the Gaussian curvature R of our apple's surface. ***Riemann*** is the mathematical embodiment of the bends and warps in spacetime. And ***Riemann*** is the agent by which those bends and warps (curvature of spacetime) produce the relative acceleration of geodesics.

Riemann, like the metric tensor $\boldsymbol{g}$ of Box 1.3, can be thought of as a family of machines, one machine residing at each event in spacetime. Each machine has three slots for the insertion of three vectors:

$$\boldsymbol{Riemann}\;(\overset{\text{slot 1}}{\downarrow}\quad,\quad\overset{\text{slot 2}}{\downarrow}\quad,\quad\overset{\text{slot 3}}{\downarrow}).$$

Choose a fiducial geodesic (free-particle world line) passing through an event $\mathscr{Q}$, and denote its unit tangent vector (particle 4-velocity) there by

$$\boldsymbol{u} = \boldsymbol{dx}/d\tau;\ \text{components},\ u^\alpha = dx^\alpha/d\tau. \tag{1.7}$$

Choose another, neighboring geodesic, and denote by $\boldsymbol{\xi}$ its perpendicular separation from the fiducial geodesic. Then insert $\boldsymbol{u}$ into the first slot of ***Riemann*** at $\mathscr{Q}$, $\boldsymbol{\xi}$ into the second slot, and $\boldsymbol{u}$ into the third. ***Riemann*** will grind for awhile; then out will pop a new vector,

$$\boldsymbol{Riemann}\;(\boldsymbol{u}, \boldsymbol{\xi}, \boldsymbol{u}).$$

Riemann tensor, through equation of geodesic deviation, produces relative accelerations

The equation of geodesic deviation states that this new vector is the negative of the relative acceleration of the two geodesics:

$$D^2\boldsymbol{\xi}/d\tau^2 + \boldsymbol{Riemann}\;(\boldsymbol{u}, \boldsymbol{\xi}, \boldsymbol{u}) = 0. \tag{1.8}$$

The Riemann tensor, like the metric tensor (Box 1.3), and like all other tensors, is a linear machine. The vector it puts out is a linear function of each vector inserted into a slot:

$$\begin{aligned}&\boldsymbol{Riemann}\;(2\boldsymbol{u}, a\boldsymbol{w} + b\boldsymbol{v}, 3\boldsymbol{r})\\ &= 2 \times a \times 3\;\boldsymbol{Riemann}\;(\boldsymbol{u}, \boldsymbol{w}, \boldsymbol{r}) + 2 \times b \times 3\;\boldsymbol{Riemann}\;(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{r}).\end{aligned} \tag{1.9}$$

Consequently, in any coordinate system the components of the vector put out can be written as a "trilinear function" of the components of the vectors put in:

$$\boldsymbol{r} = \boldsymbol{Riemann}\;(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \Longleftrightarrow r^\alpha = R^\alpha{}_{\beta\gamma\delta}\, u^\beta\, v^\gamma\, w^\delta. \tag{1.10}$$

(Here there is an implied summation on the indices β, γ, δ; cf. Box 1.1.) The $4 \times 4 \times 4 \times 4 = 256$ numbers $R^\alpha{}_{\beta\gamma\delta}$ are called the "components of the Riemann tensor in the given coordinate system." In terms of components, the equation of geodesic deviation states

$$\frac{D^2\xi^\alpha}{d\tau^2} + R^\alpha{}_{\beta\gamma\delta}\frac{dx^\beta}{d\tau}\xi^\gamma\frac{dx^\delta}{d\tau} = 0. \tag{1.8$'$}$$

In Einstein's geometric theory of gravity, this equation of geodesic deviation summarizes the entire effect of geometry on matter. It does for gravitation physics what the Lorentz force equation,

Equation of geodesic deviation is analog of Lorentz force law

$$\frac{D^2x^\alpha}{d\tau^2} - \frac{e}{m}F^\alpha{}_\beta\frac{dx^\alpha}{d\tau} = 0, \tag{1.11}$$

does for electromagnetism. See Box 1.7.

The units of measurement of the curvature are cm^{-2} just as well in spacetime as on the surface of the apple. Nothing does so much to make these units stand out clearly as to express mass in "geometrized units":

Geometrized units

$$\begin{aligned} m(\mathrm{cm}) &= (G/c^2)m_{\mathrm{conv}}(\mathrm{g}) \\ &= (0.742 \times 10^{-28}\ \mathrm{cm/g})m_{\mathrm{conv}}(\mathrm{g}). \end{aligned} \tag{1.12}$$

Box 1.7 EQUATION OF MOTION UNDER THE INFLUENCE OF A GRAVITATIONAL FIELD AND AN ELECTROMAGNETIC FIELD, COMPARED AND CONTRASTED

	Electromagnetism [Lorentz force, equation (1.11)]	*Gravitation [Equation of geodesic deviation (1.8′)]*
Acceleration is defined for one particle?	Yes	No
Acceleration defined how?	Actual world line compared to world line of uncharged "fiducial" test particle passing through same point with same 4-velocity.	Already an uncharged test particle, which can't accelerate relative to itself! Acceleration measured relative to a nearby test particle as fiduciary standard.
Acceleration depends on all four components of the 4-velocity of the particle?	Yes	Yes
Universal acceleration for all test particles in same locations with same 4-velocity?	No; is proportional to e/m	Yes
Driving field	Electromagnetic field	Riemann curvature tensor
Ostensible number of distinct components of driving field	$4 \times 4 = 16$	$4^4 = 256$
Actual number when allowance is made for symmetries of tensor	6	20
Names for more familiar of these components	3 electric 3 magnetic	6 components of local Newtonian tide-producing acceleration

This conversion from grams to centimeters by means of the ratio

$$G/c^2 = 0.742 \times 10^{-28}\ \mathrm{cm/g}$$

is completely analogous to converting from seconds to centimeters by means of the ratio

$$c = \frac{9.4605460000 \times 10^{17}\ \mathrm{cm}}{31{,}556{,}925.974\ \mathrm{sec}}$$

(see end of §1.5). The sun, which in conventional units has $m_{\mathrm{conv}} = 1.989 \times 10^{33}$ g, has in geometrized units a mass $m = 1.477$ km. Box 1.8 gives further discussion.

Components of Riemann tensor evaluated from relative accelerations of slowly moving particles

Using geometrized units, and using the Newtonian theory of gravity, one can readily evaluate nine of the most interesting components of the Riemann curvature tensor near the Earth or the sun. The method is the gravitational analog of determining the electric field strength by measuring the acceleration of a slowly moving test particle. Consider the separation between the geodesics of two nearby and slowly moving ($v \ll c$) particles at a distance r from the Earth or sun. In the standard, nearly inertial coordinates of celestial mechanics, all components of the 4-velocity of the

Box 1.8 GEOMETRIZED UNITS

Throughout this book, we use "geometrized units," in which the speed of light c, Newton's gravitational constant G, and Boltzman's constant k are all equal to unity. The following alternative ways to express the number 1.0 are of great value:

$$1.0 = c = 2.997930 \cdots \times 10^{10}\ \mathrm{cm/sec}$$

$$1.0 = G/c^2 = 0.7425 \times 10^{-28}\ \mathrm{cm/g};$$

$$1.0 = G/c^4 = 0.826 \times 10^{-49}\ \mathrm{cm/erg};$$

$$1.0 = Gk/c^4 = 1.140 \times 10^{-65}\ \mathrm{cm/K};$$

$$1.0 = c^2/G^{1/2} = 3.48 \times 10^{24}\ \mathrm{cm/gauss^{-1}}.$$

One can multiply a factor of unity, expressed in any one of these ways, into any term in any equation without affecting the validity of the equation. Thereby one can convert one's units of measure from grams to centimeters to seconds to ergs to . . . . For example:

$$\begin{aligned}\text{Mass of sun} = M_\odot &= 1.989 \times 10^{33}\ \mathrm{g}\\ &= (1.989 \times 10^{33}\, g) \times (G/c^2)\\ &= 1.477 \times 10^{5}\ \mathrm{cm}\\ &= (1.989 \times 10^{33}\, g) \times (c^2)\\ &= 1.788 \times 10^{54}\ \mathrm{ergs}.\end{aligned}$$

The standard unit, in terms of which everything is measured in this book, is centimeters. However, occasionally conventional units are used; in such cases a subscript "conv" is sometimes, but not always, appended to the quantity measured:

$$M_{\odot\mathrm{conv}} = 1.989 \times 10^{33}\ \mathrm{g}.$$

fiducial test particle can be neglected except $dx^0/d\tau = 1$. The space components of the equation of geodesic deviation read

$$d^2\xi^k/d\tau^2 + R^k{}_{0j0}\xi^j = 0. \tag{1.13}$$

Comparing with the conclusions of Newtonian theory, equations (1.5), we arrive at the following information about the curvature of spacetime near a center of mass:

$$\begin{Vmatrix} R^{\hat{x}}{}_{\hat{0}\hat{x}\hat{0}} & R^{\hat{y}}{}_{\hat{0}\hat{x}\hat{0}} & R^{\hat{z}}{}_{\hat{0}\hat{x}\hat{0}} \\ R^{\hat{x}}{}_{\hat{0}\hat{y}\hat{0}} & R^{\hat{y}}{}_{\hat{0}\hat{y}\hat{0}} & R^{\hat{z}}{}_{\hat{0}\hat{y}\hat{0}} \\ R^{\hat{x}}{}_{\hat{0}\hat{z}\hat{0}} & R^{\hat{y}}{}_{\hat{0}\hat{z}\hat{0}} & R^{\hat{z}}{}_{\hat{0}\hat{z}\hat{0}} \end{Vmatrix} \begin{matrix} = \\ = \\ = \end{matrix} \begin{Vmatrix} m/r^3 & 0 & 0 \\ 0 & m/r^3 & 0 \\ 0 & 0 & -2m/r^3 \end{Vmatrix} \tag{1.14}$$

(units cm^{-2}). Here and henceforth the caret or "hat" is used to indicate the components of a vector or tensor in a local Lorentz frame of reference ("physical components," as distinguished from components in a general coordinate system). Einstein's theory will determine the values of the other components of curvature (e.g., $R^{\hat{x}}{}_{\hat{z}\hat{x}\hat{z}} = -m/r^3$); but these nine terms are the ones of principal relevance for many applications of gravitation theory. They are analogous to the components of the electric field in the Lorentz equation of motion. Many of the terms not evaluated are analogous to magnetic field components—ordinarily weak unless the source is in rapid motion.

This ends the survey of the effect of geometry on matter ("effect of curvature of apple in causing geodesics to cross"—especially great near the dimple at the top, just as the curvature of spacetime is especially large near a center of gravitational attraction). Now for the effect of matter on geometry ("effect of stem of apple in causing dimple")!

§1.7. EFFECT OF MATTER ON GEOMETRY

> *The weight of any heavy body of known weight at a particular distance from the center of the world varies according to the variation of its distance therefrom; so that as often as it is removed from the center, it becomes heavier, and when brought near to it, is lighter. On this account, the relation of gravity to gravity is as the relation of distance to distance from the center.*
>
> AL KHĀZINĪ (*Merv, A.D. 1115*), *Book of the Balance of Wisdom*

Figure 1.12 shows a sphere of the same density, $\rho = 5.52$ g/cm^3, as the average density of the Earth. A hole is bored through this sphere. Two test particles, A and B, execute simple harmonic motion in this hole, with an 84-minute period. Therefore their geodesic separation $\boldsymbol{\xi}$, however it may be oriented, undergoes a simple periodic motion with the same 84-minute period:

$$d^2\xi^j/d\tau^2 = -\left(\frac{4\pi}{3}\rho\right)\xi^j, \qquad j = x \text{ or } y \text{ or } z. \tag{1.15}$$

Thereby conclude that the general relativistic line element, accurate to linearized order, is

$$ds^2 = -(1 - 2M/r)\,dt^2 + (1 + 2M/r)(dr^2 + r^2\,d\theta^2 + r^2\sin^2\theta\,d\phi^2).$$

(c) Derive this general, static, spherically symmetric, Lorentz-gauge, vacuum solution to the linearized field equations from scratch, working entirely in spherical coordinates. [*Hint:* As discussed at the end of Box 18.2, $\eta_{\mu\nu}$ in equation (18.8c) must be replaced by $g_{\mu\nu\,\text{flat}}$; and in the field equations and gauge conditions (18.8a, b), all commas (partial derivatives) must be replaced by covariant derivatives, whose connection coefficients come from $g_{\mu\nu\,\text{flat}}$.]
(d) Calculate the Riemann curvature tensor for this gravitational field. The answer should agree with equation (1.14).

---

§18.2. GRAVITATIONAL WAVES

Linearized theory and electromagnetic theory compared

The gauge conditions and field equations (18.8a, b) of linearized theory bear a close resemblance to the equations of electromagnetic theory in Lorentz gauge and flat spacetime,

$$A^{\alpha}{}_{,\alpha} = 0, \tag{18.10a}$$

$$-A^{\mu}{}_{,\alpha}{}^{\alpha} = 4\pi J^{\mu}. \tag{18.10b}$$

They differ only in the added index ($h^{\mu\nu}$ versus A^{μ}, $T^{\mu\nu}$ versus J^{μ}). Consequently, from past experience with electromagnetic theory, one can infer much about linearized gravitation theory.

For example, the field equations (18.8b) must have gravitational-wave solutions. The analog of the electromagnetic plane wave

$$A^x = A^x(t-z), \qquad A^y = A^y(t-z), \qquad A^z = 0, \qquad A^0 = 0,$$

Plane gravitational waves

will be the gravitational plane wave

$$\begin{gathered} \bar{h}^{xx} = \bar{h}^{xx}(t-z), \qquad \bar{h}^{xy} = \bar{h}^{xy}(t-z), \qquad \bar{h}^{yy} = \bar{h}^{yy}(t-z), \\ \bar{h}^{\mu 0} = \bar{h}^{\mu z} = 0 \text{ for all } \mu. \end{gathered} \tag{18.11}$$

Although a detailed study of such waves will be delayed until Chapters 35–37, some properties of these waves are explored in the exercises at the end of the next section.

§18.3. EFFECT OF GRAVITY ON MATTER

How to analyze effects of weak gravity on matter

The effects of weak gravitational fields on matter can be computed by using the linearized metric (18.1) and Christoffel symbols (18.2) in the appropriate equations of motion—i.e., in the geodesic equation (for the motion of particles or light rays), in the hydrodynamic equations (for fluid matter), in Maxwell's equations (for electromagnetic waves), or in the equation $\boldsymbol{\nabla} \cdot \boldsymbol{T} = 0$ for the total stress-energy tensor of whatever fields and matter may be present. Exercises 18.5, 18.6 and 18.7 provide examples, as do the Newtonian-limit calculations in exercises 16.1 and 16.4, and in §17.4. If, however, the lowest-order (linearized) gravitational "forces" (Christoffel-symbol terms) have a significant influence on the motion of the sources of the gravitational field, one finds that the linearized field equation (18.7) is inadequate, and better approximations to Einstein's equations must be considered. [Thus emission of gravitational waves by a mechanically or electrically driven oscillator falls within the scope of linearized theory, but emission by a double-star system, or by stellar oscillations that gravitational forces maintain, will require discussion of nonlinear terms (gravitational "stress-energy") in the Einstein equations; see §§36.9 to 36.11.]

The above conclusions follow from a consideration of conservation laws associated with the linearized field equation. Just as the electromagnetic equations (18.10a, b) guarantee charge conservation

$$J^{\mu}{}_{,\mu} = 0, \qquad \int_{\text{all space}} J^0(t, \boldsymbol{x})\, dx\, dy\, dz \equiv Q = \text{const},$$

Conservation of 4-momentum and angular momentum in linearized theory

so the gravitational equations (18.8a, b) guarantee conservation of the total 4-momentum and angular momentum of any body bounded by vacuum:

$$T^{\mu\nu}{}_{,\nu} = 0, \tag{18.12a}$$

$$\int_{\text{body}} T^{\mu 0}(t, \boldsymbol{x})\, dx\, dy\, dz \equiv P^{\mu} = \text{const}; \tag{18.12b}$$

$$(x^{\alpha} T^{\beta\mu} - x^{\beta} T^{\alpha\mu})_{,\mu} = 0, \tag{18.13a}$$

$$\int_{\text{body}} (x^{\alpha} T^{\beta 0} - x^{\beta} T^{\alpha 0})\, dx\, dy\, dz \equiv J^{\alpha\beta} = \text{const}. \tag{18.13b}$$

(See §5.11 for the basic properties of angular momentum in special relativity. The angular momentum here is calculated relative to the origin of the coordinate system.) Now it is important that the stress-energy components $T^{\mu\nu}$, which appear in the linearized field equations (18.7) and in these conservation laws, are precisely the components one would calculate using special relativity (with $g_{\mu\nu} = \eta_{\mu\nu}$). As a result, the energy-momentum conservation formulated here contains no contributions or effects of gravity! From this one sees that linearized theory assumes that gravitational forces do no significant work. For example, energy losses due to gravitational radiation-damping forces are neglected by linearized theory. Similarly, conservation of 4-momentum P^{μ} for each of the bodies acting as sources of $h_{\mu\nu}$ means that each body moves along a geodesic of $\eta_{\mu\nu}$ (straight lines in the nearly Lorentz coordinate system) rather than along a geodesic of $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$. Thus, linearized theory can be used to calculate the motion of test particles and fields, using $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$; but to include gravitational corrections to the motion of the sources themselves—to allow them to satisfy $T^{\mu\nu}{}_{;\nu} = 0$ rather than $T^{\mu\nu}{}_{,\nu} = 0$—one must reinsert into the field equations the nonlinear terms that linearized theory discards. (See, e.g., Chapter 20 on conservation laws; §§36.9–36.11 on the generation of gravitational waves and radiation reaction; and Chapter 39 on the post-Newtonian approximation.)

Limit on validity of linearized theory: gravity must not affect motions of sources significantly

The energy, momentum, and angular momentum radiated by gravitational waves in linearized theory can be calculated by special-relativistic methods analogous to those used in electromagnetic theory for electromagnetic waves [Fiertz and Pauli (1939)], but it will be more informative and powerful to use a fully gravitational approach (Chapters 35 and 36).

EXERCISES

Exercise 18.4. SPACETIME CURVATURE FOR A PLANE GRAVITATIONAL WAVE

Calculate the components of the Riemann curvature tensor [equations (18.9)] for the gravitational plane wave (18.11). [*Answer:*

$$R_{x0x0} = -R_{y0y0} = -R_{x0xz} = +R_{y0yz} = +R_{xzxz} = -R_{yzyz} = -\frac{1}{4}(\bar{h}_{xx} - \bar{h}_{yy})_{,tt};$$

$$R_{x0y0} = -R_{x0yz} = +R_{xzyz} = -R_{xzy0} = -\frac{1}{2}\bar{h}_{xy,tt};$$

all other components vanish except those obtainable from the above by the symmetries $R_{\alpha\beta\gamma\delta} = R_{[\alpha\beta][\gamma\delta]} = R_{\gamma\delta\alpha\beta}$.

Exercise 18.5. A PRIMITIVE GRAVITATIONAL-WAVE DETECTOR (see Figure 18.1)

Two beads slide almost freely on a smooth stick; only slight friction impedes their sliding. The stick falls freely through spacetime, with its center moving along a geodesic and its ends attached to gyroscopes, so they do not rotate. The beads are positioned equidistant (distance $\frac{1}{2}\ell$) from the stick's center. Plane gravitational waves [equation (18.11) and exercise 18.4], impinging on the stick, push the beads back and forth ("geodesic deviation"; "tidal gravitational forces"). The resultant friction of beads on stick heats the stick; and the passage of the waves is detected by measuring the rise in stick temperature.* (Of course, this is not the best of all conceivable designs!) Neglecting the effect of friction on the beads' motion, calculate the proper distance separating them as a function of time. [*Hints:* Let $\boldsymbol{\xi}$ be the separation between the beads; and let $\boldsymbol{n} = \boldsymbol{\xi}/|\boldsymbol{\xi}|$ be a unit vector that points along the stick in the stick's own rest frame. Then their separation has magnitude $\ell = \boldsymbol{\xi} \cdot \boldsymbol{n}$. The fact that the stick is nonrotating is embodied in a parallel-transport law for $\boldsymbol{n}$, $\nabla_{\boldsymbol{u}}\boldsymbol{n} = 0$. ("Fermi-Walker transport" of §§6.5, 6.6, and 13.6 reduces to parallel transport, because the stick moves along a geodesic with $\boldsymbol{a} = \nabla_{\boldsymbol{u}}\boldsymbol{u} = 0$.) Thus,

$$d\ell/d\tau = \nabla_{\boldsymbol{u}}(\boldsymbol{\xi} \cdot \boldsymbol{n}) = (\nabla_{\boldsymbol{u}}\boldsymbol{\xi}) \cdot \boldsymbol{n},$$

$$d^2\ell/d\tau^2 = \nabla_{\boldsymbol{u}}\nabla_{\boldsymbol{u}}(\boldsymbol{\xi} \cdot \boldsymbol{n}) = (\nabla_{\boldsymbol{u}}\nabla_{\boldsymbol{u}}\boldsymbol{\xi}) \cdot \boldsymbol{n},$$

where τ is the stick's proper time. But $\nabla_{\boldsymbol{u}}\nabla_{\boldsymbol{u}}\boldsymbol{\xi}$ is produced by the Riemann curvature of the wave (geodesic deviation):

$$\nabla_{\boldsymbol{u}}\nabla_{\boldsymbol{u}}\boldsymbol{\xi} = \text{projection along } \boldsymbol{n} \text{ of } [-\textbf{Riemann}(\ldots, \boldsymbol{u}, \boldsymbol{\xi}, \boldsymbol{u})].$$

(The geodesic-deviation forces perpendicular to the stick, i.e., perpendicular to $\boldsymbol{n}$, are coun-

---

*This thought experiment was devised by Bondi [1957, 1965; Bondi and McCrea (1960)] as a means for convincing skeptics of the reality of gravitational waves.

EXERCISE

Exercise 19.3. GRAVITATIONAL FIELD FAR FROM A STATIONARY, FULLY RELATIVISTIC SOURCE

Derive the line element (19.13) for the special case of a source that is time-independent ($g_{\mu\nu,t} = 0$). This can be a difficult problem, if one does not proceed carefully along the following outlined route. (1) Initially ignore all nonlinearities in the Einstein field equations. The field is weak far from the source. These nonlinearities will be absent from the dominant terms. (2) Calculate the dominant terms using linearized theory in the Lorentz gauge [equations (18.8)]. (3) In particular, write the general solution to the vacuum, time-independent wave equation (18.8b) in the following form involving $n^j \equiv x^j/r \equiv$ (unit vector in radial direction):

$$\begin{aligned}
\bar{h}_{00} &= \frac{A^0}{r} + \frac{B^j n^j}{r^2} + 0\left(\frac{1}{r^3}\right),\\
\bar{h}_{0j} &= \frac{A^j}{r} + \frac{B^{jk} n^k}{r^2} + 0\left(\frac{1}{r^3}\right),\\
\bar{h}_{jk} &= \frac{A^{jk}}{r} + \frac{B^{jkl} n^l}{r^2} + 0\left(\frac{1}{r^3}\right),\\
A^{jk} &= A^{(jk)}, \qquad B^{jkl} = B^{(jk)l}.
\end{aligned} \tag{19.17}$$

(Round brackets denote symmetrization.) (4) Then impose the Lorentz gauge conditions $\bar{h}_{\alpha}{}^{\beta}{}_{,\beta} = 0$ on this general solution, thereby learning

$$\begin{aligned}
&A^j = 0, \qquad A^{jk} = 0,\\
&B^{jk}(\delta^{jk} - 3n^j n^k) = 0,\\
&B^{jkl}(\delta^{kl} - 3n^k n^l) = 0.
\end{aligned} \tag{19.18}$$

(5) Write B^{jk} as the sum of its trace $3B$, its traceless symmetric part S^{jk}, and its traceless antisymmetric part (these are its "irreducible parts"):

$$B^{jk} = B\delta^{jk} + S^{jk} + \epsilon^{jkl}F^l, \qquad S^{jj} = 0. \tag{19.19}$$

Show that any tensor B^{jk} can be put into such a form. Then show that the gauge conditions (19.18) imply $S^{jk} = 0$. (6) Similarly show that any tensor B^{jkl} that is symmetric on its first two indices can be put into the form

$$\begin{aligned}
&B^{jkl} = \delta^{jk}A^l + C^{(j}\delta^{k)l} + \epsilon^{ml(j}E^{k)m} + S^{jkl},\\
&E^{km} \text{ symmetric and traceless, i.e.,} \quad E^{km} = E^{(km)}, \qquad E^{kk} = 0,\\
&S^{jkl} \text{ symmetric and traceless, i.e.,} \quad S^{jkl} = S^{(jkl)},\\
&\qquad S^{jjl} = S^{jkk} = S^{jkj} = 0.
\end{aligned} \tag{19.20}$$

Then show that the gauge conditions (19.18) imply $C^j = -2A^j$ and $E^{km} = S^{jkl} = 0$. (7) Combining all these results, conclude that

$$\begin{aligned}
\bar{h}_{00} &= \frac{A^0}{r} + \frac{B^j n^j}{r^2} + 0\left(\frac{1}{r^3}\right),\\
\bar{h}_{0j} &= \frac{\epsilon^{jkl} n^k F^l}{r^2} + \frac{B n^j}{r^2} + 0\left(\frac{1}{r^3}\right),\\
\bar{h}_{jk} &= \frac{\delta^{jk} A^l n^l - A^j n^k - A^k n^j}{r^2} + 0\left(\frac{1}{r^3}\right).
\end{aligned} \tag{19.21}$$

Then use gauge transformations, which stay within Lorentz gauge, to eliminate B and A^j from $\bar{h}_{0j}$ and $\bar{h}_{jk}$; so

$$\begin{aligned}\bar{h}_{00} &= \frac{A^0}{r} + \frac{(B^j + A^j)n^j}{r^2} + 0\left(\frac{1}{r^3}\right),\\ \bar{h}_{0j} &= \frac{\epsilon^{jkl}n^k F^l}{r^2} + 0\left(\frac{1}{r^3}\right),\\ \bar{h}_{jk} &= 0\left(\frac{1}{r^3}\right).\end{aligned} \tag{19.22}$$

(8) Translate the origin of coordinates so ${x^j}_{\text{new}} = {x^j}_{\text{old}} - (B^j + A^j)/A^0$; in the new coordinate system $\bar{h}_{\alpha\beta}$ has the same form as (19.22), but with $B^j + A^j$ removed. From the resultant $\bar{h}_{\alpha\beta}$, construct the metric and redefine the constants A^0 and F^l to agree with equation (19.13). (9) All linear terms in the metric are now accounted for. The dominant nonlinear terms must be proportional to the square, $(M/r)^2$, of the dominant linear term. The easiest way to get the proportionality constant is to take the Schwarzschild geometry for a fully relativistic, static, spherical source [equation (31.1)], by a change of coordinates put it in the form

$$ds^2 = -\left(\frac{1 - M/2r}{1 + M/2r}\right)^2 dt^2 + \left(1 + \frac{M}{2r}\right)^4 (dx^2 + dy^2 + dz^2) \tag{19.23}$$

(exercise 25.8), and expand it in powers of M/r.

§19.4. MASS AND ANGULAR MOMENTUM OF A CLOSED UNIVERSE

> *"There are no snakes in Ireland."*
>
> Statement of St. Patrick after driving the snakes out of Ireland (legend*)

For a closed universe the total mass-energy M and angular momentum S are undefined and undefinable

There is no such thing as "the energy (or angular momentum, or charge) of a closed universe," according to general relativity, and this for a simple reason. To weigh something one needs a platform on which to stand to do the weighing.

To weigh the sun, one measures the periods and semimajor axes of planetary orbits, and applies Kepler's "1-2-3" law, $M = \omega^2 a^3$. To measure the angular momentum, S, of the sun (a task for space technology in the 1970's or 1980's!), one measures the precession of a gyroscope in a near orbit about the sun, or one examines some other aspect of the "dragging of inertial frames." To determine the electric charge

---

[*] Stokes (1887) and other standard references deny this legend. In part I of Stokes the basic manuscript references are listed, including especially codex manuscript Rawlinson B.512 in 154 folios, in double columns, written by various hands in the fourteenth and fifteenth centuries (*cf. Catalogi codicum manuscriptorum Bibliothecae Bodleianae Partis Quintae Fasciculus Primus*, Oxford, 1862, col. 728–732). In this manuscript, folio 97b.1, line 14, reads in the translation of Stokes, Part I, p. xxx: "as Paradise is without beasts, without a snake, without a lion, without a dragon, without a scorpion, without a mouse, without a frog, so is Ireland in the same manner without any harmful animal, save only the wolf. . ."

B (%

Pag. 32(59) _ Curvature.

Others:

- 20211010/ Gravity/
  Quantum theory of Gravity...(1984).djvu
  Pag. 456(470)
  the covariance of the relevant Gaussian is a Green's function of the Jacobi operator with boundary conditions determined by the domain of integration.

- 20211010/ Gravity/
  Quantum gravity - Cambridge (2004).djvu
  Pag. 33(51)
  Riemann geometry.
  The tensor g equips the spacetime manifold M with a metric structure: it defines a distance between any two points, and this distance is a smooth function on M.
  (More precisely, it defines a pseudo-metric structure, as distance can be imaginary).
  Riemann studied the structure defined by (M, g), called today a riemannian manifold,

B
! -256- - -* ! - B 6 ! 9 3 7 !A 3 = !3
B D "!3) 3 4 @ 2 : / ' "KA

-256- - -* ! - B 6 ! 9 3 7 !A 3 = !3
B D "!3) 3 4 @ 2 : / ' 1"
('$B ' A! ! A! 7! ! ! 7 0 C! D 1 !

with

$$S[q,p]=\int p\,\mathrm{d}q-H(q,p)\,\mathrm{d}t \tag{1.4}$$

the covariance of the relevant Gaussian is a Green's function of the Jacobi operator with boundary conditions determined by the domain of integration. We recall that the Jacobi operator (alias the small disturbance operator, alias the operator of geodetic deviation) is obtained from the second variation of the action. For instance, on the space of paths with values in the configuration space, $q: T\to\mathbb{R}^n$, $T=[t_a,t_b]$, the expansion of the action around the classical path Q is

$$S[Q+q]=S[Q]+(\text{boundary terms})+\tfrac{1}{2}S''[Q]qq+\ldots \tag{1.5}$$

and the Jacobi operator $\mathscr{J}[Q(t)]\equiv\mathscr{J}_t$ is the differential operator (of second-order for Lagrangians that depend quadratically on velocities) defined by

$$S''[Q]qq=\int_{t_a}^{t_b} q(t)(\mathscr{J}_t q(t))\,\mathrm{d}t\equiv\langle q,\mathscr{J}q\rangle. \tag{1.6}$$

On the space of paths (q,p) with values in phase space, $\mathbb{R}^{2n}$, the expansion of the action around the classical path (Q,P) is

$$S[Q+q,P+p]=S[Q,P]+(\text{boundary terms})+\tfrac{1}{2}(q,p)S''[Q,P]\begin{pmatrix}q\\p\end{pmatrix}+\ldots \tag{1.7}$$

and the phase space Jacobi operator is defined by the equation

$$(q,p)S''[Q,P]\begin{pmatrix}q\\p\end{pmatrix}=\int_{t_a}^{t_b}(q(t),p(t))\mathscr{J}_t^{(\text{p.s.})}\begin{pmatrix}q(t)\\p(t)\end{pmatrix}\mathrm{d}t \tag{1.8}$$

where $\mathscr{J}_t^{(\text{p.s.})}$ is a $2n\times 2n$ matrix built up with first-order differential operators.

Although one might expect the configuration space version and the phase space version of the path integrals for simple systems, say systems without constraints, to be totally equivalent, we shall find it not to be so. Even for a free particle in $\mathbb{R}^n$ there is an important difference: given equivalent boundary conditions, the configuration space Jacobi operator may have only positive eigenvalues whereas the corresponding phase space Jacobi operator has both positive and negative eigenvalues. It follows that the Green's function of the configuration space Jacobi operator can be used as a covariance, whereas the Green's function of the corresponding phase space Jacobi operator cannot.†

In the following, we consider the Lagrangian‡

$$L=\tfrac{1}{2}\dot{q}^2-V(q) \tag{1.9}$$

---

† Path integration using Green's functions of the phase space Jacobi operator has been developed in DeWitt-Morette *et al* (1979) and in DeWitt-Morette *et al* (1982). The path integrals in these papers are defined in terms of prodistributions and the non-positivity of the Green's function is not an issue.

‡ See for instance DeWitt-Morette *et al* (1979) for the configuration and phase space Jacobi operators for arbitrary systems.

D -256- - -* ! - B 6 ! !A 3 ! D "
H A 3 ' "KA
-256- - -* ! - B 6 ! !A 3 ! D " H A 3
' 1"
(%% $ 5 ! 3

The Riemann tensor can be defined via

$$R^{\mu}{}_{\nu\rho\sigma}\; e^I_\mu = R^I{}_{J\;\rho\sigma}\; e^J_\nu, \tag{2.90}$$

The Ricci tensor is

$$R_{\mu\nu} \quad = \quad R^I_\mu\; e_{I\nu} \tag{2.91}$$

(R^I_μ is defined in (2.11)), and the energy momentum tensor (see footnote after (2.37))

$$T_{\mu\nu} \quad = \quad T^I_\mu\; e_{I\nu}. \tag{2.92}$$

Using this, the Einstein equations (2.37) read

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu}(R+\lambda) = 8\pi G T_{\mu\nu}. \tag{2.93}$$

The Minkowski solution is

$$g_{\mu\nu}(x) = \eta_{\mu\nu}, \tag{2.94}$$

where we see clearly that the spacetime Minkowski metric is nothing but a particular value of the gravitational field. With a straightforward calculation, the action (2.10) reads

$$S[g] = \frac{1}{16\pi G}\int (R+\lambda)\; \sqrt{-\det g}\; d^4x. \tag{2.95}$$

The matter action cannot be written in metric variables.

Riemann geometry. The tensor g equips the spacetime manifold M with a metric structure: it defines a distance between any two points, and this distance is a smooth function on M. (More precisely, it defines a pseudo-metric structure, as distance can be imaginary.) Riemann studied the structure defined by (M, g), called today a riemannian manifold, and defined the Riemann curvature tensor, as a generalization of Gauss theory of curved surfaces to an arbitrary number of dimensions. Riemann presented this mathematical theory as a general theory of "geometry" that generalizes Euclidean geometry. Einstein utilized this mathematical theory for describing the physical dynamics of the gravitational field. In retrospect, the reason this was possible is because –as understood by Einstein– the Euclidean structure of the physical space in which we live is determined by the local gravitational field. Therefore elementary physical geometry is simply a description of the local properties of the gravitational field, as revealed by matter (rigid bodies) interacting with it. This point is discussed in more detail below in section 2.2.3.

The basic equations of GR presented in this section do not look too different from the equations of a prerelativistic[9] field theory, such as QED or the standard model. But the similarity can be very misleading. The physical interpretation of a general relativistic theory is very different from the interpretation of a prerelativistic one. In particular, the meaning of the coordinates x^μ is different than in prerelativistic physics, and the gauge invariant observables are not related to the fields as they are in prerelativistic physics.

Understanding the physical meaning of the GR formalism has been a long process. This process has taken decades and perhaps it is not entirely concluded yet. For several decades after Einstein's discovery of the theory, for instance, it was not clear whether or not the theory predicted gravitational waves. The prevailing opinion was that wave solutions were only a coordinate artifact and did not represent physical waves capable of carrying energy and momentum, or, as Bondi put it, capable of "boiling a glass of water". This opinion was wrong of course. Einstein himself got badly wrong on the meaning of the Schwarzshild singularity. Wrong high precision measurements of the Earth-Moon distance have been in the literature for a while, because of a mistake due to a conceptual confusion between physical and coordinate distance. . .

I do not want to give the impression that GR is foggy. On the very contrary, the fact that in all these and similar instances consensus has eventually emerged indicates that the conceptual

---

[9] Recall that in this book "relativistic" means *general* relativistic.

The Riemann tensor can be defined via

$$R^{\mu}{}_{\nu\rho\sigma}\, e^I_\mu = R^I{}_{J\ \rho\sigma}\, e^J_\nu, \tag{2.90}$$

The Ricci tensor is

$$R_{\mu\nu} \quad = \quad R^I_\mu\, e_{I\nu} \tag{2.91}$$

(R^I_μ is defined in (2.11)), and the energy momentum tensor (see footnote after (2.37))

$$T_{\mu\nu} \quad = \quad T^I_\mu\, e_{I\nu}. \tag{2.92}$$

Using this, the Einstein equations (2.37) read

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu}(R+\lambda) = 8\pi G T_{\mu\nu}. \tag{2.93}$$

The Minkowski solution is

$$g_{\mu\nu}(x) = \eta_{\mu\nu}, \tag{2.94}$$

where we see clearly that the spacetime Minkowski metric is nothing but a particular value of the gravitational field. With a straightforward calculation, the action (2.10) reads

$$S[g] = \frac{1}{16\pi\, G}\int (R+\lambda)\ \sqrt{-\det g}\ d^4x. \tag{2.95}$$

The matter action cannot be written in metric variables.

Riemann geometry. The tensor g equips the spacetime manifold M with a metric structure: it defines a distance between any two points, and this distance is a smooth function on M. (More precisely, it defines a pseudo-metric structure, as distance can be imaginary.) Riemann studied the structure defined by (M, g), called today a riemannian manifold, and defined the Riemann curvature tensor, as a generalization of Gauss theory of curved surfaces to an arbitrary number of dimensions. Riemann presented this mathematical theory as a general theory of "geometry" that generalizes Euclidean geometry. Einstein utilized this mathematical theory for describing the physical dynamics of the gravitational field. In retrospect, the reason this was possible is because –as understood by Einstein– the Euclidean structure of the physical space in which we live is determined by the local gravitational field. Therefore elementary physical geometry is simply a description of the local properties of the gravitational field, as revealed by matter (rigid bodies) interacting with it. This point is discussed in more detail below in section 2.2.3.

The basic equations of GR presented in this section do not look too different from the equations of a prerelativistic[9] field theory, such as QED or the standard model. But the similarity can be very misleading. The physical interpretation of a general relativistic theory is very different from the interpretation of a prerelativistic one. In particular, the meaning of the coordinates x^μ is different than in prerelativistic physics, and the gauge invariant observables are not related to the fields as they are in prerelativistic physics.

Understanding the physical meaning of the GR formalism has been a long process. This process has taken decades and perhaps it is not entirely concluded yet. For several decades after Einstein's discovery of the theory, for instance, it was not clear whether or not the theory predicted gravitational waves. The prevailing opinion was that wave solutions were only a coordinate artifact and did not represent physical waves capable of carrying energy and momentum, or, as Bondi put it, capable of "boiling a glass of water". This opinion was wrong of course. Einstein himself got badly wrong on the meaning of the Schwarzshild singularity. Wrong high precision measurements of the Earth-Moon distance have been in the literature for a while, because of a mistake due to a conceptual confusion between physical and coordinate distance...

I do not want to give the impression that GR is foggy. On the very contrary, the fact that in all these and similar instances consensus has eventually emerged indicates that the conceptual structure of GR is secure. But to understand this conceptual structure, to understand how to use the equations of GR correctly and how to relate the quantities appearing in these equations to the numbers measured in the laboratory or observed by the astronomers, is definitely a non-trivial problem. More generally, the problem is to understand what precisely GR says about the world. On the other hand, clarity is this respect is essential if we want to understand the quantum physics of the theory.

---

[9] Recall that in this book "relativistic" means *general* relativistic.

In order to shed light on this problem, it is illuminating to retrace the conceptual path and the problems that led to the discovery of the theory. This is done in the following section 2.2. The impatient reader may skip section 2.2 and jump to section 2.3, where the interpretation of GR is compactly presented (but impatience slows understanding. . .)

and defined the Riemann curvature tensor as a generalization of Gauss' theory of curved surfaces to an arbitrary number of dimensions.

Riemann presented this mathematical theory as a general theory of "geometry" that generalized Euclidean geometry.

Einstein utilized this mathematical theory for describing the physical dynamics of the gravitational field.

_20211011/ Gravity-Hartle/

Gravitation-Misner-1973.pdf

Pág. 32(65) - Curvature of what?

Pág. 146(179) - §5.9.

Conservation of 4-Momentum: Differential formulation

→ Gauss's theorem in four dimensions...

_20211011/ A Book of Curves Lockwood/

EH Lockwood - A book of Curves (1961).pdf

Obs. Livro todo!!!

_20211011/ Gauss/PPN578628864.pdf

Die Theorie der Parallellinien _

-256- - -* ! - 7 !A ! @ : < !
/ % 1"
(% B A! @ ! •
('B % $ /

Case (d)

Here the contribution to the integral comes entirely from two arbitrary spacelike hypersurfaces, $\mathcal{S}_A$ and $\mathcal{S}_B$, cutting all the way across spacetime. As in cases (a) and (b), the integral form of the conservation law says

Total 4-momentum independent of hypersurface where measured

$$\boldsymbol{p}_{\text{on } \mathcal{S}_A} = \boldsymbol{p}_{\text{on } \mathcal{S}_B}; \tag{5.33}$$

i.e., *the total 4-momentum on a spacelike slice through spacetime is independent of the specific slice chosen*—so long as the energy-momentum flux across the "hypersurface at infinity" connecting $\mathcal{S}_A$ and $\mathcal{S}_B$ is zero.

Case (e)

This case concerns a box whose walls oscillate and accelerate as time passes. The three-dimensional boundary $\partial\mathcal{V}$ is made up of (1) the interior $\mathcal{S}$ of the box, at an initial moment of time $t =$ constant in the box's initial Lorentz frame, taken with nonstandard orientation; (2) the interior $\bar{\mathcal{S}}$ of the box, at $\bar{t} =$ constant in its final Lorentz frame, with standard orientation; (3) the 3-volume $\mathcal{T}$ swept out by the box's two-dimensional faces between the initial and final states, with positive sense oriented outward. The integral conservation law $\int_{\partial\mathcal{V}} \boldsymbol{T} \cdot d^3\boldsymbol{\Sigma} = 0$ says

Change with time of 4-momentum in a box equals flux of 4-momentum across its faces

$$\begin{aligned}\begin{pmatrix}\text{total 4-momentum}\\ \text{in box at } \bar{\mathcal{S}}\end{pmatrix} - \begin{pmatrix}\text{total 4-momentum}\\ \text{in box at } \mathcal{S}\end{pmatrix}&\\ = \begin{pmatrix}\text{total 4-momentum that enters box through}\\ \text{its faces between states } \mathcal{S} \text{ and } \bar{\mathcal{S}}\end{pmatrix}&.\end{aligned} \tag{5.34}$$

§5.9. CONSERVATION OF 4-MOMENTUM: DIFFERENTIAL FORMULATION

Complementary to any "integral conservation law in flat spacetime" is a "differential conservation law" with identical information content. To pass back and forth between them, one can use Gauss's theorem.

Gauss's theorem in four dimensions, applied to the law of 4-momentum conservation, converts the surface integral of $T^{\mu\alpha}$ into a volume integral of $T^{\mu\alpha}{}_{,\alpha}$:

$$0 = \oint_{\partial\mathcal{V}} T^{\mu\alpha}\, d^3\Sigma_\alpha = \int_{\mathcal{V}} T^{\mu\alpha}{}_{,\alpha}\, dt\, dx\, dy\, dz. \tag{5.35}$$

(See Box 5.3 for elementary discussion; Box 5.4 for sophisticated discussion.) If the integral of $T^{\mu\alpha}{}_{,\alpha}$ is to vanish, as demanded, for any and every 4-volume $\mathcal{V}$, then $T^{\mu\alpha}{}_{,\alpha}$ must itself vanish everywhere in spacetime:

Differential conservation law for 4-momentum: $\boldsymbol{\nabla} \cdot \boldsymbol{T} = 0$

$$T^{\mu\alpha}{}_{,\alpha} = 0;\ \text{i.e., } \boldsymbol{\nabla} \cdot \boldsymbol{T} = 0 \text{ everywhere.} \tag{5.36}$$

(*continued on page 152*)

Box 5.3 VOLUME INTEGRALS, SURFACE INTEGRALS, AND GAUSS'S THEOREM IN COMPONENT NOTATION

A. Volume Integrals in Spacetime

1. By analogy with three-dimensional space, the volume of a "hyperparallelepiped" with vector edges $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$ is

$$\text{4-volume} \equiv \Omega \equiv \epsilon_{\alpha\beta\gamma\delta} A^\alpha B^\beta C^\gamma D^\delta = \det \begin{Vmatrix} A^0 & A^1 & A^2 & A^3 \\ B^0 & B^1 & B^2 & B^3 \\ C^0 & C^1 & C^2 & C^3 \\ D^0 & D^1 & D^2 & D^3 \end{Vmatrix}$$

$$= {}^*(\boldsymbol{A} \wedge \boldsymbol{B} \wedge \boldsymbol{C} \wedge \boldsymbol{D}).$$

Here, as for 3-volumes, orientation matters; interchange of any two edges reverses the sign of Ω. The *standard orientation* for any 4-volume is the one which makes Ω positive; thus, $\boldsymbol{e}_0 \wedge \boldsymbol{e}_1 \wedge \boldsymbol{e}_2 \wedge \boldsymbol{e}_3$ has standard orientation if $\boldsymbol{e}_0$ points toward the future and $\boldsymbol{e}_1$, $\boldsymbol{e}_2$, $\boldsymbol{e}_3$ are a righthanded triad.

2. The "volume element" whose edges in a specific, standard-oriented Lorentz frame are

$$A^\alpha = (\Delta t, 0, 0, 0),\ B^\alpha = (0, \Delta x, 0, 0),\ C^\alpha = (0, 0, \Delta y, 0),\ D^\alpha = (0, 0, 0, \Delta z)$$

has a 4-volume, according to the above definition, given by

$$\Delta^4\Omega = \epsilon_{0123}\, \Delta t\, \Delta x\, \Delta y\, \Delta z = \Delta t\, \Delta x\, \Delta y\, \Delta z.$$

3. Thus, the volume integral of a tensor $\boldsymbol{S}$ over a four-dimensional region $\mathcal{V}$ of spacetime, defined as

$$\boldsymbol{M} \equiv \lim_{\left(\substack{\text{number of} \\ \text{elementary} \\ \text{volumes} \\ \to\infty}\right)} \sum_{\left(\substack{\text{elementary} \\ \text{volumes } \mathcal{A} \\ \text{in } \mathcal{V}}\right)} \boldsymbol{S}_{\text{at center of } \mathcal{A}}\, (\text{volume of } \mathcal{A}),$$

can be calculated in a Lorentz frame by

$$M^\alpha{}_{\beta\gamma} = \int_{\mathcal{V}} S^\alpha{}_{\beta\gamma}\, d^4\Omega = \int_{\mathcal{V}} S^\alpha{}_{\beta\gamma}\, dt\, dx\, dy\, dz.$$

Box 5.3 *(continued)*

B. Integrals over 3-Surfaces in Spacetime

1. Introduce arbitrary coordinates a, b, c on the three-dimensional surface. The elementary volume bounded by coordinate surfaces

   $$a_0 < a < a_0 + \Delta a, \qquad b_0 < b < b_0 + \Delta b,$$
   $$c_0 < c < c_0 + \Delta c$$

   has edges

   $$A^\alpha = \frac{\partial x^\alpha}{\partial a}\Delta a,\ B^\beta = \frac{\partial x^\beta}{\partial b}\Delta b,\ C^\gamma = \frac{\partial x^\gamma}{\partial c}\Delta c;$$

   so its volume 1-form is

   $$\Delta^3\Sigma_\mu = \epsilon_{\mu\alpha\beta\gamma}\frac{\partial x^\alpha}{\partial a}\frac{\partial x^\beta}{\partial b}\frac{\partial x^\gamma}{\partial c}\Delta a\,\Delta b\,\Delta c.$$

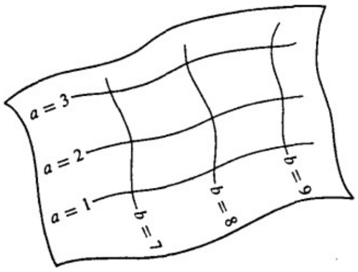

2. The integral of a tensor $\boldsymbol{S}$ over the 3-surface $\mathcal{S}$ thus has components

   $$N^\alpha{}_\beta = \int_{\mathcal{S}} S^\alpha{}_\beta{}^\gamma\, d^3\Sigma_\gamma = \int_{\mathcal{S}} S^\alpha{}_\beta{}^\gamma \epsilon_{\gamma\mu\nu\lambda}\frac{\partial x^\mu}{\partial a}\frac{\partial x^\nu}{\partial b}\frac{\partial x^\lambda}{\partial c}\, da\, db\, dc.$$

   An equivalent formula involving a Jacobian is often used (see exercise 5.5):

   $$N^\alpha{}_\beta = \int_{\mathcal{S}} S^\alpha{}_\beta{}^\gamma \frac{1}{3!}\epsilon_{\gamma\mu\nu\lambda}\frac{\partial(x^\mu, x^\nu, x^\lambda)}{\partial(a, b, c)}\, da\, db\, dc.$$

C. Gauss's Theorem Stated

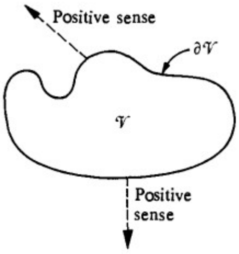

1. Consider a bounded four-dimensional region of spacetime $\mathcal{V}$ with closed boundary $\partial\mathcal{V}$. Orient the volume 1-forms on $\partial\mathcal{V}$ so that the "positive sense" is away from $\mathcal{V}$.
2. Choose a tensor field $\boldsymbol{S}$. Integrate its divergence over $\mathcal{V}$, and integrate it itself over $\partial\mathcal{V}$. The results must be the same (*Gauss's theorem*):

   $$\int_{\mathcal{V}} S^\alpha{}_\beta{}^\gamma{}_{,\gamma}\, d^4\Omega = \oint_{\partial\mathcal{V}} S^\alpha{}_\beta{}^\gamma\, d^3\Sigma_\gamma.$$

D. Proof of Gauss's Theorem

1. The indices α and β of $S^{\alpha}{}_{\beta}{}^{\gamma}$ "go along for a free ride," so one can suppress them from the proof. Then the equation to be derived is

$$\int_{\mathcal{V}} S^{\gamma}{}_{,\gamma}\, dt\, dx\, dy\, dz = \oint_{\partial\mathcal{V}} S^{\gamma}\, d^3\Sigma_{\gamma}.$$

2. Since the integral of a derivative is just the original function, the volume integral of $S^0{}_{,0}$ is

$$\int_{\mathcal{V}} S^0{}_{,0}\, dt\, dx\, dy\, dz$$

$$= \int_{\text{“up”}} S^0\, dx\, dy\, dz - \int_{\text{“down”}} S^0\, dx\, dy\, dz.$$

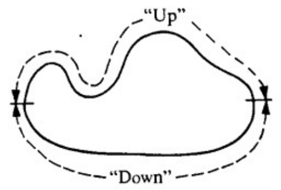

3. The surface integral $\int_{\partial\mathcal{V}} S^0\, d^3\Sigma_0$ can be reduced to the same set of terms:
   a. Use x, y, z as coordinates on $\partial\mathcal{V}$. On the "up" side, $d^3\Sigma_0$ must be positive to achieve a "positive" sense pointing away from $\mathcal{V}$, so (see part B above)

$$d^3\Sigma_0 = \epsilon_{0\alpha\beta\gamma} \frac{\partial x^{\alpha}}{\partial x} \frac{\partial x^{\beta}}{\partial y} \frac{\partial x^{\gamma}}{\partial z}\, dx\, dy\, dz = \epsilon_{0123}\, dx\, dy\, dz = dx\, dy\, dz.$$

   b. On the "down" side, $d^3\Sigma_0$ must be negative, so

$$d^3\Sigma_0 = -dx\, dy\, dz.$$

   c. Hence,

$$\int_{\partial\mathcal{V}} S^0\, d^3\Sigma_0 = \int_{\text{“up”}} S^0\, dx\, dy\, dz - \int_{\text{“down”}} S^0\, dx\, dy\, dz.$$

4. Equality is proved for the other components in the same manner. Adding components produces the result desired:

$$\int_{\mathcal{V}} S^{\gamma}{}_{,\gamma}\, d^4\Omega = \oint_{\partial\mathcal{V}} S^{\gamma}\, d^3\Sigma_{\gamma}.$$

FOR THE READER WHO HAS STUDIED CHAPTER 4

Box 5.4 I. EVERY INTEGRAL IS THE INTEGRAL OF A FORM.
II. THE THEOREM OF GAUSS IN THE LANGUAGE OF FORMS.

I. *Every integral encountered* in Chapter 5 can be interpreted as *the integral of an exterior differential form.* This circumstance shows up in fourfold and threefold integrals, for example, in the fact that

$$d^4\Omega = \boldsymbol{\varepsilon} = {}^*1 = \varepsilon_{0123}\, \boldsymbol{d}t \wedge \boldsymbol{d}x \wedge \boldsymbol{d}y \wedge \boldsymbol{d}z$$

and

$$d^3\Sigma_\mu = \varepsilon_{\mu|\alpha\beta\gamma|}\, \boldsymbol{d}x^\alpha \wedge \boldsymbol{d}x^\beta \wedge \boldsymbol{d}x^\gamma$$

are basis 4- and 3-forms. (Recall: the indices $\alpha\beta\gamma$ between vertical bars are to be summed only over $0 \leq \alpha < \beta < \gamma \leq 3$.) A more extensive glossary of notations is found in C below.

II. *Gauss's Theorem for a tensor integral* in flat space reads

$$\int_{\mathscr{V}} (\boldsymbol{\nabla} \cdot \boldsymbol{S})\, d^4\Omega = \oint_{\partial\mathscr{V}} \boldsymbol{S} \cdot d\boldsymbol{\Sigma}$$

for any tensor, such as $\boldsymbol{S} = S^\alpha{}_\beta{}^\gamma \boldsymbol{e}_\alpha \otimes \boldsymbol{\omega}^\beta \otimes \boldsymbol{e}_\gamma$ (see Box 5.3 for component form). It is an application of the generalized Stokes Theorem (Box 4.1), and *depends on the fact that the basis vectors* $\boldsymbol{e}_\alpha$ *and* $\boldsymbol{\omega}^\beta$ *of a global Lorentz frame are constants,* i.e., are independent of x. The definitions follow in A; the proof is in B.

A. *Tensor-valued integrals* can be defined in flat spaces because one uses constant basis vectors. Thus one defines

$$\int \boldsymbol{S} \cdot d^3\boldsymbol{\Sigma} = \boldsymbol{e}_\alpha \otimes \boldsymbol{\omega}^\beta \int S^\alpha{}_\beta{}^\gamma\, d^3\Sigma_\gamma$$

for a tensor of the indicated rank. One justifies pulling basis vectors and forms outside the integral sign because they are constants, independent of location in spacetime. Each of the numbers $\int S^\alpha{}_\beta{}^\gamma\, d^3\Sigma_\gamma$ (for $\alpha, \beta = 0, 1, 2, 3$) is then evaluated by substituting *any* properly oriented parametrization of the hypersurface into the 3-form $S^\alpha{}_\beta{}^\gamma\, d^3\Sigma_\gamma$ as described in Box 4.1 (arbitrary curvilinear parametrization in the part of the calculation not involving the "free indices" α and β). In other words, $\boldsymbol{S} \cdot d^3\boldsymbol{\Sigma} = \boldsymbol{e}_\alpha \otimes \boldsymbol{\omega}^\beta \otimes S^\alpha{}_\beta{}^\gamma\, d^3\Sigma_\gamma$ is considered a "tensor-valued 3-form." Under an integral sign, it is contracted with the hyperplane element tangent to the 3-surface $\mathscr{P}(\lambda^1, \lambda^2, \lambda^3)$ of integration to form the integral

$$\begin{aligned}\int \boldsymbol{S} \cdot d^3\boldsymbol{\Sigma} &= \int \left\langle \boldsymbol{S} \cdot d^3\boldsymbol{\Sigma}, \frac{\partial\mathscr{P}}{\partial\lambda^1} \wedge \frac{\partial\mathscr{P}}{\partial\lambda^2} \wedge \frac{\partial\mathscr{P}}{\partial\lambda^3} \right\rangle d\lambda^1\, d\lambda^2\, d\lambda^3 \\ &= \boldsymbol{e}_\alpha \otimes \boldsymbol{\omega}^\beta \int S^\alpha{}_\beta{}^\gamma \varepsilon_{\gamma|\lambda\mu\nu|} \underbrace{\frac{\partial(x^\lambda, x^\mu, x^\nu)}{\partial(\lambda^1, \lambda^2, \lambda^3)}}_{\substack{\text{Jacobian}\\ \text{determinant}}} d\lambda^1\, d\lambda^2\, d\lambda^3.\end{aligned}$$

Although constant basis vectors $\boldsymbol{e}_\alpha$, $\boldsymbol{\omega}^\beta$ derived from rectangular coordinates are essential here, *a completely general parametrization of the hypersurface may be used.*

B. *The proof of Gauss's Theorem* is a computation:

$$\oint_{\partial\mathcal{V}} \boldsymbol{S}\cdot d^3\boldsymbol{\Sigma} = \boldsymbol{e}_\alpha \otimes \boldsymbol{\omega}^\beta \oint_{\partial\mathcal{V}} S^\alpha{}_\beta{}^\gamma\, d^3\Sigma_\gamma \qquad (\boldsymbol{e}_\alpha, \boldsymbol{\omega}^\beta \text{ are constant})$$

$$= \boldsymbol{e}_\alpha \otimes \boldsymbol{\omega}^\beta \int_{\mathcal{V}} \boldsymbol{d}(S^\alpha{}_\beta{}^\gamma\, d^3\Sigma_\gamma) \qquad \text{(Stokes Theorem)}$$

$$= \boldsymbol{e}_\alpha \otimes \boldsymbol{\omega}^\beta \int_{\mathcal{V}} S^\alpha{}_\beta{}^\gamma{}_{,\gamma}\, {}^*1 \qquad \text{(see below)}$$

$$= \int_{\mathcal{V}} (\boldsymbol{\nabla}\cdot\boldsymbol{S})\, d^4\Omega. \qquad \text{(merely notation)}$$

The missing computational step above is

$$\begin{aligned} \boldsymbol{d}(S^\alpha{}_\beta{}^\gamma\, d^3\Sigma_\gamma) &= (\partial S^\alpha{}_\beta{}^\gamma/\partial x^\rho)\, \boldsymbol{d}x^\rho \wedge d^3\Sigma_\gamma \\ &= (\partial S^\alpha{}_\beta{}^\gamma/\partial x^\gamma)\, {}^*1. \end{aligned}$$

Here the first step uses $\boldsymbol{d}(d^3\Sigma_\gamma) = 0$ (which follows from $\varepsilon_{\mu\alpha\beta\gamma} = \text{const}$ in flat spacetime). The second step uses

$$\boldsymbol{d}x^\rho \wedge d^3\Sigma_\gamma = \delta^\rho_\gamma\, {}^*1.$$

[Write the lefthand side of this identity as $\varepsilon_{\gamma|\mu\nu\lambda|}\, \boldsymbol{d}x^\rho \wedge \boldsymbol{d}x^\mu \wedge \boldsymbol{d}x^\nu \wedge \boldsymbol{d}x^\lambda$. The only possible non-zero term in the sum over $\mu\nu\lambda$ is the one with $\mu < \nu < \lambda$ all different from ρ. The righthand side is the value of this term.]

C. *Glossary of notations.*

Charge density 3-form:

$$\begin{aligned} {}^*\boldsymbol{J} &= J^\mu\, d^3\Sigma_\mu = \boldsymbol{J}\cdot d^3\boldsymbol{\Sigma} \\ &= \rlap{$\underbrace{\phantom{J^\mu\varepsilon_{\mu\alpha\beta\gamma}}}_{({}^*J)_{\alpha\beta\gamma}}$}J^\mu \underbrace{\varepsilon_{\mu\alpha\beta\gamma}\, \boldsymbol{d}x^\alpha \wedge \boldsymbol{d}x^\beta \wedge \boldsymbol{d}x^\gamma/3!}_{d^3\Sigma_\mu}. \end{aligned}$$

Maxwell and Faraday 2-forms:

$${}^*\boldsymbol{F} = \frac{1}{2} F^{\mu\nu}\, d^2S_{\mu\nu};$$

$$\boldsymbol{F} = \frac{1}{2} F_{\mu\nu}\, \boldsymbol{d}x^\mu \wedge \boldsymbol{d}x^\nu.$$

Basis 2-forms:

$$\boldsymbol{d}x^\alpha \wedge \boldsymbol{d}x^\beta; \qquad \text{(one way to label)}$$

$$d^2S_{\mu\nu} = \varepsilon_{\mu\nu|\alpha\beta|}\, \boldsymbol{d}x^\alpha \wedge \boldsymbol{d}x^\beta. \qquad \text{(dual way to label)}$$

Energy-momentum density 3-form:

$$\boldsymbol{T}\cdot d^3\boldsymbol{\Sigma} \equiv \boldsymbol{e}_\mu T^{\mu\nu}\, d^3\Sigma_\nu \equiv \overbrace{{}^*\boldsymbol{T};}^{}$$

$$\text{dual on last index, } ({}^*T)^\mu{}_{\alpha\beta\gamma} = T^{\mu\nu}\varepsilon_{\nu\alpha\beta\gamma}.$$

Angular momentum density 3-form:

$$\boldsymbol{\mathcal{J}}\cdot d^3\boldsymbol{\Sigma} \equiv \frac{1}{2}\boldsymbol{e}_\mu \wedge \boldsymbol{e}_\nu\, \mathcal{J}^{\mu\nu\alpha}\, d^3\Sigma_\alpha \equiv {}^*\boldsymbol{\mathcal{J}};$$

$$({}^*\mathcal{J})^{\mu\nu}{}_{\alpha\beta\gamma} = \mathcal{J}^{\mu\nu\lambda}\varepsilon_{\lambda\alpha\beta\gamma}.$$

(In the frame-independent equation $\boldsymbol{\nabla} \cdot \boldsymbol{T} = 0$, one need not worry about which slot of $\boldsymbol{T}$ to take the divergence on; the slots are symmetric, so either can be used.)

The equation $\boldsymbol{\nabla} \cdot \boldsymbol{T} = 0$ is *the differential formulation of the law of 4-momentum conservation*. It is also called the *equation of motion for stress-energy*, because it places constraints on the dynamic evolution of the stress-energy tensor. To examine these constraints for simple systems is to realize the beauty and power of the equation $\boldsymbol{\nabla} \cdot \boldsymbol{T} = 0$.

§5.10. SAMPLE APPLICATIONS OF $\boldsymbol{\nabla} \cdot \boldsymbol{T} = 0$

Newtonian fluid characterized by $|v^j| \ll 1$, $p \ll \rho$

The equation of motion $\boldsymbol{\nabla} \cdot \boldsymbol{T} = 0$ makes contact with the classical (Newtonian) equations of hydrodynamics, when applied to a nearly Newtonian fluid. Such a fluid has low velocities relative to the Lorentz frame used, $|v^j| \ll 1$; and in its rest frame its pressure is small compared to its density of mass-energy, $p/\rho = p/\rho c^2 \ll 1$. For example, the air in a hurricane has

$$|v^j| \sim 100 \text{ km/hour} \sim 3{,}000 \text{ cm/sec} \sim 10^{-7}\, c = 10^{-7} \ll 1,$$

$$\frac{p}{\rho} \sim \frac{1 \text{ atmosphere}}{10^{-3} \text{ g/cm}^3} \sim \frac{10^6 \text{ dynes/cm}^2}{10^{-3} \text{ g/cm}^3} = 10^9 \,\frac{\text{cm}^2}{\text{sec}^2} \sim 10^{-12}\, c^2 = 10^{-12} \ll 1.$$

Stress-energy tensor and equation of motion for a Newtonian fluid

The stress-energy tensor for such a fluid has components

$$T^{00} = (\rho + p)u^0u^0 - p \approx \rho, \tag{5.37a}$$

$$T^{0j} = T^{j0} = (\rho + p)u^0u^j \approx \rho v^j, \tag{5.37b}$$

$$T^{jk} = (\rho + p)u^ju^k + p\,\delta^{jk} \approx \rho v^jv^k + p\,\delta^{jk}; \tag{5.37c}$$

and the equation of motion $\boldsymbol{\nabla} \cdot \boldsymbol{T} = 0$ has components

$$T^{00}{}_{,0} + T^{0j}{}_{,j} = \partial\rho/\partial t + \boldsymbol{\nabla} \cdot (\rho \boldsymbol{v}) = 0 \quad \text{("equation of continuity");} \tag{5.38a}$$

and

$$T^{j0}{}_{,0} + T^{jk}{}_{,k} = \partial(\rho v^j)/\partial t + \partial(\rho v^j v^k)/\partial x^k + \partial p/\partial x^j = 0,$$

or, equivalently (by combining with the equation of continuity),

$$\frac{\partial \boldsymbol{v}}{\partial t} + (\boldsymbol{v} \cdot \boldsymbol{\nabla})\boldsymbol{v} = -\frac{1}{\rho}\boldsymbol{\nabla} p \qquad \text{("Euler's equation").} \tag{5.38b}$$

Box 5.5 derives and discusses these results from the Newtonian viewpoint.

Application of $\boldsymbol{\nabla} \cdot \boldsymbol{T} = 0$ to an electrically charged, vibrating rubber block

As a second application of $\boldsymbol{\nabla} \cdot \boldsymbol{T} = 0$, consider a composite system: a block of rubber with electrically charged beads imbedded in it, interacting with an electromagnetic field. The block of rubber vibrates, and its accelerating beads radiate electromagnetic waves; at the same time, incoming electromagnetic waves push on the beads, altering the pattern of vibration of the block of rubber. The interactions shove 4-momentum back and forth between beaded block and electromagnetic field.

D -256- - -* ! - = : ? . " 2 D . A
H A 3 /B 1"
D -256- - -* ! - = : ? . " 2) .
A ! D " H A 3 "KA
D -256- - -* ! - = : ? . " 2) .
A ! D " H A 3 1"
*D ? A " UUU
-256- - -* ! - 8$ B B' 1"
E 4 9 " ! ! A = . " D ! 7!

data

20211012///
https://archive.org/details/
euclidoreinstein0000unse
→ Euclid or Einstein: a proof
of the parallel theory and a
critique of metageometry.
1931 Callahan, Jeremiah
Joseph.

ISBN 9781258776879

=

- 20211011/Gauss/ED058058.pdf
Pág. 16 The Trisection Problem

—

Important: Leibniz inspired
↓
"Numero impari gaudet" Virgil
DEUS
Book

- 20211012/Classical Analysis.../
Classical Analysis in the
Complex Plane.pdf
Obs: Livro todo!!!

! 1 ##! A #" ! # "

! 1 ##! A #" ! # " " LI /

! -256- - -* ! - " !' 1"
! -256- - -* ! - = "0 = 5 " L 4!
! ! ! 1"

!% 1 ##! A #" ! # " 1! ! 1

!% -256- - -* ! - " 1! ! 1 1"
!% -256- - -* ! - " 1! ! 1 ID 1"
!% % -256- - -* ! - " 1! ! 1 "KA

D -256- - -* ! -) !" = "J 8 J ! "
= 1"

-256- - -* ! - ! ! ! 3 1!
1" 1"

" -256- - -* ! - = 7= 1"

-256- - -* ! - @) 1"

-256- - -* ! - =4 $ $ 1"
*D 9 9 D & 1 ! UUU
(B 9 9 D

*D N8 4 1! ! " O ? D , 1 "

-256- - -* ! - ! ! 2 ! 3 1 L
! 1"
*D ? A " UUU

-256- - -* ! - 9 1 L & ! 3 ? D ,}
! ! 3 9 3 3 ! A J !1 3 ! 1 ! " + 1 ! & 1 !
) . Y 1"

& 1 ! UUU
-256- - -* ! - ! ! 2 ! 3 1 L
! 1"
($%

9 % 7 3 ? D ,

Remarks: (i) There is an alternative proof of the limit in 3.21 which is short and transparent but a bit less elementary, and affords no estimate of the speed of convergence. Let L be the function L_1 of 3.19. For integers $n > N > 0$, $z \in \overline{D}(0, N) \setminus \{0\}$ implies $1 + z/n \in D(1, 1)$. Consequently, by 3.19

$$nL(1 + z/n) = z \cdot \frac{L(1 + z/n) - L(1)}{z/n} \to zL'(1) = z \quad \text{as} \quad n \to \infty,$$

uniformly in such z, whence $(1+z/n)^n = [\exp L(1+z/n)]^n = \exp(nL(1+z/n)) \to \exp(z)$. [Similarly, $(1 + z_n/n)^n \to e^z$ whenever $z_n \to z$.]

(ii) According to 3.21 the numbers $a_n := n!\, n^{-n}$ satisfy $a_{n+1}/a_n = (1 + 1/n)^{-n} \to 1/e$, in view of which 3.9 tells us that $a_n^{1/n} = n^{-1}(n!)^{1/n} \to 1/e$. After a little algebra, involving $n^{-1} \sum_{j=1}^{n} j = \frac{1}{2}(n + 1)$, another interesting feature of e comes to light:

$$\frac{1}{2}e = \lim_{n\to\infty} \frac{\text{Arithmetic Mean of } \{1, 2, \ldots, n\}}{\text{Geometric Mean of } \{1, 2, \ldots, n\}}.$$

(A generalization to means of $\{1^s, 2^s, \ldots, n^s\}$ $(s > -1)$ was proved by different [elementary] means by Xu [2010].)

Although the constant π is ubiquitous in this subject (see Castellanos [1988]), one seldom needs estimates of it finer than (8) and (8)′ in the proof of 3.14. Nevertheless, we will look at a few of the myriad series and integrals by which it can be explicitly calculated.

Theorem 3.22 (Gregory-Leibniz series) $\dfrac{\pi}{4} = \displaystyle\int_0^1 \frac{dx}{1 + x^2} = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n + 1}$.

Proof: The quadratic polynomial $X^2 - i$ has exactly two zeros in $\mathbb{C}$, $\pm(1 + i)/\sqrt{2}$. Since, in the notation of the proof of 3.14,

$$\left[C\left(\frac{\pi}{4}\right) + iS\left(\frac{\pi}{4}\right)\right]^2 = E\left(i\frac{\pi}{4}\right)^2 = E\left(i\frac{\pi}{2}\right) = i$$

and C, S are non-negative in $\left[0, \frac{\pi}{2}\right]$, it must be that

$$C\left(\frac{\pi}{4}\right) = S\left(\frac{\pi}{4}\right) = \frac{1}{\sqrt{2}}. \tag{1}$$

Since C is an even function and is never zero in $[0, \pi/2)$, by definition of π, we may form the function $T := S/C$ on $(-\pi/2, \pi/2)$ and have from equations (5) of the proof of 3.14 that

$$T' = \frac{S'C - SC'}{C^2} = \frac{C^2 + S^2}{C^2} = 1 + (S/C)^2 = 1 + T^2 > 0. \tag{2}$$

Therefore T is strictly increasing. Since $T(0) = 0$ and $T(t) \to +\infty$ as $t \uparrow \frac{\pi}{2}$, the range of the odd function T is $\mathbb{R}$. Now $T^{-1} \colon \mathbb{R} \mapsto (-\pi/2, \pi/2)$ exists and by (2)

$$(3) \qquad \int_0^x \frac{1}{1+t^2}\,dt = \int_{T^{-1}(0)}^{T^{-1}(x)} \frac{T'}{1+T^2} = \int_{T^{-1}(0)}^{T^{-1}(x)} 1 = T^{-1}(x) \qquad \forall x \in \mathbb{R}$$

and in particular, because of (1),

$$(3)' \qquad \int_0^1 \frac{1}{1+t^2}\,dt = T^{-1}(1) = \frac{\pi}{4}.$$

For every $N \in \mathbb{N}$, $t \in \mathbb{R}$

$$\sum_{n=0}^{N-1} \left(-t^2\right)^n = \frac{1-\left(-t^2\right)^N}{1-\left(-t^2\right)} = \frac{1}{1+t^2} - \frac{\left(-1\right)^N t^{2N}}{1+t^2}$$

and therefore

$$\begin{aligned} \left|\sum_{n=0}^{N-1} \frac{\left(-1\right)^n x^{2n+1}}{2n+1} - \int_0^x \frac{1}{1+t^2} dt\right| &= \left|\int_0^x \left[\sum_{n=0}^{N-1} \left(-t^2\right)^n - \frac{1}{1+t^2}\right] dt\right| \\ (4) \qquad &= \left|\int_0^x \frac{\left(-1\right)^{N-1} t^{2N}}{1+t^2} dt\right| \qquad \forall x \in \mathbb{R}. \end{aligned}$$

The trivial inequalities $\frac{1}{2}t^{2N} \leq t^{2N}(1+t^2)^{-1} \leq t^{2N}$ for $t \in [-1,1]$ let us estimate the last integral in (4), below and above, yielding for $x \in [-1,1]\,, N \in \mathbb{N}$.

$$(5) \qquad \frac{|x|^{2N+1}}{4N+2} \leq \left|\sum_{n=0}^{N-1} \frac{\left(-1\right)^n x^{2n+1}}{2n+1} - \int_0^x \frac{1}{1+t^2} dt\right| \leq \frac{|x|^{2N+1}}{2N+1}.$$

In particular, using (3)

$$(6) \qquad T^{-1}(x) = \sum_{n=0}^{\infty} \frac{\left(-1\right)^n x^{2n+1}}{2n+1} \qquad \forall x \in [-1,1]\,.$$

Now equations $(3)'$ and (6) prove the theorem, but $(3)'$ and (5) also furnish the estimate

$$(7) \qquad \frac{1}{4N+2} \leq \left|\sum_{n=0}^{N-1} \frac{\left(-1\right)^n}{2n+1} - \frac{\pi}{4}\right| \leq \frac{1}{2N+1} \qquad \forall N \in \mathbb{N}.$$

Remarks: (i) See 3.37(vii), 3.38(viii) and 11.37(viii) [with $z := 1/2$] for different proofs. This series converges excruciatingly slowly, as (7) shows; for one converging to π with geometric rapidity, see 5.82(ii). After checking (see 3.30(ii) Hints) that $T(\pi/6) = 1/\sqrt{3}$, one can also take $x := 1/\sqrt{3}$ in (6) and get the even more rapidly converging representation

$$\frac{\pi}{6} = \frac{1}{\sqrt{3}} \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} \cdot \frac{1}{3^n}. \tag{8}$$

In 3.45(v) we will see that for each $x \in \overline{\mathbb{D}} \setminus \{-i, i\}$ the series (6) converges to a solution w of the equation $\tan w = x$. (See def. 3.27.)

(ii) Besides using the partial sums of the series or Riemann-sum approximations to the integral, there is a third (more efficient?) way to exploit 3.22 for rational approximations to π. Each $n \in \mathbb{N}$ determines a polynomial p_n with integer coefficients such that $x^{4n}(1-x)^{4n} = p_n(x)(1+x^2) + (-4)^n$. (One checks that $p_1(x) := x^6 - 4x^5 + 5x^4 - 4x^2 + 4$ works and the rest is induction.) Therefore $\int_0^1 x^{4n}(1-x)^{4n}(1+x^2)^{-1}dx$ is a rational number minus $(-4)^{n-1}\pi$. This integral is positive and less than $[\max\{x(1-x) : 0 \le x \le 1\}]^{4n} = 4^{-4n}$. A rational approximation to π closer than $4^{-4n-(n-1)} = 4^{-5n+1}$ emerges. In case $n = 1$, $\int_0^1 x^4(1-x)^4(1+x^2)^{-1}dx = 22/7 - \pi$. Since

$$\frac{1}{630} = \int_0^1 x^4(1-x)^4 dx > \int_0^1 x^4(1-x)^4(1+x^2)^{-1}dx > \int_0^1 x^4(1-x)^4 2^{-1} dx = \frac{1}{1260},$$

we get $1/630 > 22/7 - \pi > 1/1260$, and so

$$3 + \frac{10}{71} < \frac{22}{7} - \frac{1}{630} < \pi < \frac{22}{7} - \frac{1}{1260} < 3 + \frac{1}{7}.$$

This is from Dalzell [1944], with more in Lucas [2009]. The extreme inequalities are Archimedes' famous estimate; his geometric construction can be found on pp. 31–34 of C. H. Edwards [1979].

(iii) Yet a fourth way to utilize 3.22 is to express $\pi/4 = \arctan(1)$ as a rational linear combination of $\arctan(\varepsilon_j)$ for small ε_j (for which the series is more rapidly convergent). One of these goes back to Euler: $\pi/4 = 4\arctan(1/5) - \arctan(1/70) + \arctan(1/99)$. There are numerous such relations; see pp. 274–275 of Berggren, Borwein and Borwein [2004], pp. 59–65 of Eymard and Lafon [2004] and pp. 90–98 of Castellanos [1988].

(iv) In 1755, Euler found a series which, in contrast to (6), converges to $T^{-1}(x)$ throughout $\mathbb{R}$:

$$T^{-1}(x) = \frac{x}{1+x^2} \sum_{k=0}^{\infty} \frac{2^{2k}(k!)^2}{(2k)!} \, \frac{1}{2k+1} \left(\frac{x^2}{1+x^2}\right)^k \qquad \forall x \in \mathbb{R}. \tag{9}$$

To confirm it, change the variable of integration in (3) via $t =: xC(\theta)$, $x \neq 0$, $\theta \in [0, \pi/2]$, to get

$$(1+x^2)/(1+t^2) = 1/[1 - \frac{x^2}{1+x^2} S^2(\theta)] = \sum_{k=0}^{\infty} \left(\frac{x^2}{1+x^2}\right)^k S^{2k}(\theta).$$

After commuting the uniformly converging series in $S(\theta)$ with the integration, the k^{th} coefficient is $\int_0^{\pi/2} S^{2k+1}(\theta) d\theta$; that it equals $2^{2k}(k!)^2/(2k+1)!$ is confirmed by induction, reduction of k being accomplished by an integration by parts. [The Yin & Yang relation between (9) and 5.82(iv) is noteworthy.]

(v) Since $(T^{-1})'(t) = 1/(1+t^2)$, we have

$$\frac{1}{2}(T^{-1}(x))^2 = \int_0^x T^{-1}(t)/(1+t^2)\ dt.$$

Use (9) on the integrand, permute integration and summation, and make the change of integration variable $u := t^2/(1+t^2)$. You get a combinatorially remarkable companion to (9):

$$(9)^2 \qquad (T^{-1}(x))^2 = \sum_{k=0}^{\infty} \frac{2^{2k+1}(k!)^2}{(2k+2)!} \left(\frac{x^2}{1+x^2}\right)^{k+1} \qquad \forall x \in \mathbb{R}.$$

Exercise 3.23 *With the convention* $0^0 := 1$, *the function* x^{-x} *is continuous on* $[0, 1]$. *(Cf.* 3.18(4) *and recall the customary definition* $x^{-x} := e^{-x \log x}$ *for* $x > 0$.*) Prove the satisfyingly symmetric formula*

$$\int_0^1 x^{-x}\, dx = \sum_{n=1}^{\infty} n^{-n}.$$

Hints: $x^{-x} = \sum_0^\infty (-1)^n (x \log x)^n / n!$. Interchange summation and integration and evaluate $\int_0^1 x^n \log^n(x)\, dx$ by repeated integration-by-parts.

Exercise 3.24 *Consider* $f(x) := \int_0^1 (1+t^2)^{-1} e^{-x^2(1+t^2)} dt$ *and* $g(x) := \int_0^x e^{-t^2}\, dt$, $x \in [0, \infty)$.

(i) *Show that* $0 < f(x) < e^{-x^2} f(0)$ *for all* $x > 0$.

(ii) *Use* 2.19(i) *to show that* $f'(x) = -2e^{-x^2} g(x)$ *for all* $x \geq 0$*; that is,* $f' = \left(-g^2\right)'$.

(iii) *Prove that* $\int_0^\infty e^{-t^2}\, dt = \sqrt{\pi}/2$.

Hints: Integrate in (ii) from 0 to x, let $x \to \infty$ and use (i); evaluate $f(0)$ via 3.22.

(iv) *Show that* $\int_0^\infty t^k e^{-\pi t^2} \cos(2\pi xt)\,dt$ *and* $\int_0^\infty t^k e^{-\pi t^2} \sin(2\pi xt)\,dt$ *exist for each* $x \in \mathbb{R}$, $k \in \mathbb{N} \cup \{0\}$, *and they define functions* $C_k(x)$, $S_k(x)$ *which satisfy* $C_k' = -2\pi S_{k+1}$. (Cf. 2.19(i) and glance ahead to definition 3.27)

Hints: The power series shows that $e^x \geq x^{k+1}/(k+1)!$ for all $x \geq 0$, $k \in \mathbb{N} \cup \{0\}$. Hence $t^k e^{-\pi t^2} \leq (k+1)!/\pi^{k+1} t^{k+2}$, an inequality whose validity for all $t \geq 1$ clearly insures the existence of the relevant improper integrals. For real $\alpha < \beta$

$$\left|\frac{\cos\beta - \cos\alpha}{\beta - \alpha} + \sin\alpha\right| = \frac{1}{\beta - \alpha}\left|\int_\alpha^\beta (\sin\alpha - \sin t)dt\right| \leq \sup_{t \in [\alpha,\beta]} |\sin\alpha - \sin t|$$
$$= \sup_{t \in [\alpha,\beta]} \left|\int_\alpha^t \cos\right| \leq \beta - \alpha.$$

If we apply this inequality in the relevant integrand, we see that for all real $x \neq x_0$

$$\left|\frac{C_k(x) - C_k(x_0)}{x - x_0} + 2\pi S_{k+1}(x_0)\right| \leq \int_0^\infty 2\pi t^{k+1} e^{-\pi t^2} |2\pi xt - 2\pi x_0 t|\,dt$$
$$= 4\pi^2 C_{k+2}(0)|x - x_0|.$$

(v) *Identify the function* C_0 *of* (iv) *as being* $C_0(x) = \frac{1}{2} e^{-\pi x^2}$ *for all* $x \in \mathbb{R}$.

Hints: Using (iv) and integration by parts deduce that $e^{\pi x^2} C_0(x)$ has derivative $\mathbf{0}$. The constant value $C_0(0)$ of the function $e^{\pi x^2} C_0(x)$ is found via (iii).

(vi) *Show that* $\int_{-\infty}^\infty e^{-\pi t^2} e^{-2\pi ixt}\,dt = e^{-\pi x^2}$ *for all* $x \in \mathbb{R}$.

(vii) *In the integral defining* f *replace* $-x^2$ *by* ix^2 *and in the integral defining* g *replace* $-t^2$ *by* it^2. *Call the new functions* f_i, g_i, *respectively, and carry through for them the program in* (ii) *to get*

$$g_i^2(x) = \frac{\pi i}{4} - ie^{ix^2} \int_0^1 \frac{e^{ix^2 t^2}}{1 + t^2}\,dt \qquad \forall x \in \mathbb{R}. \tag{1}$$

Infer from this and 3.22 *that*

$$|g_i^2(x)| \leq \pi/2 \qquad \forall x \in \mathbb{R}. \tag{2}$$

(viii) *Use integration by parts to show that*

$$\int_0^1 \frac{e^{ix^2 t^2}}{1 + t^2}\,dt = \frac{1}{2x}\,g_i(x) + \frac{1}{x}\int_0^1 \frac{2t}{(1 + t^2)^2}\,g_i(xt)dt \qquad \forall x > 0.$$

Infer from this and (2) *that* $\int_0^1 (1+t^2)^{-1} e^{ix^2t^2}\,dt \to 0$ *as* $x \to \infty$; *then from the latter and* (1) *infer that*

$$\lim_{x\to\infty} g_i^2(x) = \frac{\pi i}{4}. \tag{3}$$

(ix) *From* (3) *and continuity of* g_i *infer that*

$$\int_0^\infty e^{it^2}\,dt := \lim_{x\to\infty} g_i(x) \text{ exists and equals } \sqrt{\frac{\pi}{2}}\left(\frac{1+i}{2}\right).$$

Hints: From (3) we know that $\operatorname{Im} g_i^2(x) > \pi/8$ for all large x. Consequently $\operatorname{Im} g_i(x)$ is of constant sign for large x. Since $\sin \ge 0$ in $[2k\pi, (2k+1)\pi]$ for all $k \in \mathbb{Z}$ (see definition 3.27 and the proof of 3.14),

$$\begin{aligned}
0 &\le \sum_{k=0}^{n-1} \int_{2k\pi}^{(2k+1)\pi} \sin t\left(\frac{1}{\sqrt{t}} - \frac{1}{\sqrt{t+\pi}}\right)dt \\
&= \sum_{k=0}^{n-1} \left(\int_{2k\pi}^{(2k+1)\pi} \frac{\sin t}{\sqrt{t}}\,dt + \int_{(2k+1)\pi}^{2(k+1)\pi} \frac{\sin t}{\sqrt{t}}\,dt\right) \\
&= \int_0^{2n\pi} \frac{\sin t}{\sqrt{t}}\,dt = \int_0^{\sqrt{2n\pi}} 2\sin(t^2)\,dt = 2\operatorname{Im} g_i(\sqrt{2n\pi}) \qquad \forall n \in \mathbb{N}.
\end{aligned}$$

Therefore the eventual constant sign of $\operatorname{Im} g_i(x)$ is $+$. Setting $A := \sqrt{\frac{\pi}{2}}(\frac{1+i}{2})$, we have $A^2 = \frac{\pi i}{4}$ and (3) reads

$$\lim_{x\to\infty} [g_i(x) - A][g_i(x) + A] = 0.$$

As $\operatorname{Im}[g_i(x)+A] \ge \operatorname{Im} A = \sqrt{\pi/8}$ for all large x, this product can only converge to 0 if its first factor does.

Remarks: The results (iii) and (ix) will be derived by complex-analysis and Fourier analysis methods in 11.32 and its Remark (iii). The real-variables techniques here are adapted from Anonyme [1889], Bloch [1922], R. M. Young [1991], and especially Výborný [1991], where $\int_0^\infty e^{-ct^2}\,dt$ is treated for arbitrary non-zero $c \in \overline{\mathbb{H}}$, thereby getting (iii) and (ix) in one stroke. For history see footnotes 1433–1436 in Burkhardt [1914].

Exercise 3.25 (i) *Show that if* f *is differentiable in an open, connected subset* Ω *of* $\mathbb{C}$ *or* $\mathbb{R}$ *and* $f' = cf$ *for some constant* c, *then*

$$f(z) = f(a)e^{-ca}e^{cz} \qquad \forall a, z \in \Omega.$$

Hint: Let $g(z) := e^{-cz}$ and note that $(f \cdot g)' = 0$ throughout Ω.

(ii) *Find all continuous homomorphisms f of the additive group $\mathbb{R}$ into the multiplicative groups $\mathbb{T}$, $\mathbb{C} \setminus \{0\}$ and $\mathbb{R} \setminus \{0\}$.* Cf. 4.23(iii).

Hints: Since f is a homomorphism, $f(0) = 1$. By continuity then there exists $a > 0$ such that $\int_0^a f = b \neq 0$. The homomorphic property $f(x + t) = f(x)f(t)$ (all $x, t \in \mathbb{R}$) then implies that

$$bf(x) = f(x) \int_0^a f(t)\,dt = \int_0^a f(x+t)\,dt = \int_x^{x+a} f = \int_0^{x+a} f - \int_0^x f.$$

From this and the fundamental theorem of calculus we learn that the function f is not only continuous but is differentiable, and

$$bf'(x) = f(x+a) - f(x) = f(x)[f(a) - 1] \qquad \forall x \in \mathbb{R},$$

so $f' = cf$, with $c := [f(a) - 1]/b$. This leads via (i) to $f(x) = f(0)e^{cx} = e^{cx}$. It is thus obvious that f maps into $\mathbb{T}$ if and only if $\operatorname{Re} c = 0$, and maps into $\mathbb{R}$ if and only if c is real.

(iii) *Show that every continuous homomorphism h of the multiplicative group $\mathbb{R}\setminus\{0\}$ into the additive group $\mathbb{R}$ has the form $c \log \circ |\ \ |$ for some $c \in \mathbb{R}$.*

Hints: According to (ii) there is a $c \in \mathbb{R}$ such that $e^{h(e^x)} = e^{cx}$ for all $x \in \mathbb{R}$. For $x = \log y$ this gives $h(y) = cx = c \log y$, for all $y > 0$. But $0 = h(1) = h((-1)(-1)) = 2h(-1)$, so $h(-y) = h(-1) + h(y) = h(y)$ for all $y \in \mathbb{R} \setminus \{0\}$.

(iv) *Show that* (iii) *remains valid if at every occurrence $\mathbb{R}$ is replaced by $\mathbb{C}$.*

Hint: Apply (iii) to $\operatorname{Re} h$ and $\operatorname{Im} h$ restricted to $\mathbb{R} \setminus \{0\}$ to come up with $c \in \mathbb{C}$ such that $h(x) = c \log x$ for all $x > 0$. Since continuous homomorphism h maps $\mathbb{T}$ onto a compact additive subgroup of $\mathbb{C}$, $h(\mathbb{T}) = \{0\}$. Thus $h(z) = h(|z|) + h(z/|z|) = h(|z|)$, for all $z \in \mathbb{C} \setminus \{0\}$.

Exercise 3.26 *Let $\phi\colon [0,1] \mapsto \mathbb{C}$ be continuous and satisfy $\phi(1) = \phi(0)$. Show that there exists a continuous $\Phi\colon \mathbb{T} \mapsto \mathbb{C}$ satisfying*

$$(*) \qquad \Phi\big(e^{2\pi i t}\big) = \phi(t) \qquad \forall t \in [0,1].$$

Hints: By 3.14(v) and the $2\pi i$-periodicity of the exponential function, the map $t \mapsto e^{2\pi i t}$ sends $[0, 1)$ one-to-one onto the circle $\mathbb{T}$. Hence we can define Φ on $\mathbb{T}$ by the decree

$$(1) \qquad \Phi\big(e^{2\pi i t}\big) = \phi(t) \qquad \forall t \in [0,1)\,.$$

The condition $\phi(1) = \phi(0)$ ensures that in fact $(*)$ holds. Suppose now that Φ is not continuous at some $z_0 \in \mathbb{T}$. This means that there exists an $\varepsilon > 0$ and a sequence $\{z_n\} \subset \mathbb{T}$ such that $z_n \to z_0$ and

$$(2) \qquad |\Phi(z_n) - \Phi(z_0)| \geq \varepsilon, \quad n \in \mathbb{N}.$$

For each $n \in \mathbb{N}$ there is a unique $t_n \in [0, 1)$ such that $z_n = e^{2\pi i t_n}$ and there is a subsequence of $\{t_n\}$ which converges to some $t_0 \in [0, 1]$. Passing to that subsequence, we can simply suppose that $\{t_n\}$ itself converges to t_0. Then $z_n = e^{2\pi i t_n} \to e^{2\pi i t_0}$, and so $z_0 = e^{2\pi i t_0}$. But then

$$\Phi(z_n) = \phi(t_n) \to \phi(t_0) \overset{(*)}{=} \Phi\big(e^{2\pi i t_0}\big) = \Phi(z_0),$$

in violation of (2).

It is customary to introduce functions cos, sin, tan and cot by

Definition 3.27 The functions cos and sin are defined for all $z \in \mathbb{C}$ by

$$\cos(z) := \frac{e^{iz} + e^{-iz}}{2} = \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n}}{(2n)!}$$
$$\sin(z) := \frac{e^{iz} - e^{-iz}}{2i} = \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n+1}}{(2n+1)!}.$$

The function tan is defined as sin/cos. Since the zeros of cos are evidently the odd integer multiples of $\pi/2$, the domain of tan is $\mathbb{C} \setminus \pi(\mathbb{Z} + 1/2)$. The function cot is defined as cos/sin, with the appropriate domain, viz., $\mathbb{C} \setminus \pi\mathbb{Z}$ (cf. 3.30(iii)). The closely related entire functions $z \mapsto \cos(iz) = \frac{1}{2}(e^z + e^{-z})$ and $z \mapsto -i\sin(iz) = \frac{1}{2}(e^z - e^{-z})$, the even and odd parts of the exponential function, are designated $\cosh(z)$ and $\sinh(z)$ [in French literature, $ch(z)$ and $sh(z)$].

Remarks: The functions C and S in the proof of 3.14 were just cos and sin restricted to $\mathbb{R}$ and are known collectively as the *circular* or *cyclometric* or *trigonometric* functions. The functions cos and sin are infinitely differentiable throughout $\mathbb{C}$ and $\cos' = -\sin$, $\sin' = \cos$. For each $x \in \mathbb{R} \setminus \mathbb{Z}$, $y \in \mathbb{R} \setminus \{0\}$ they map the lines $\frac{1}{2}\pi x + i\mathbb{R}$, $\mathbb{R} + iy$ onto semi-hyperbolas and ellipses with foci ± 1, as the reader can easily confirm. The (only occasionally encountered) name cis for the function $z \mapsto e^{iz}$ is suggested by the identity $e^{iz} = \cos(z) + i\sin(z)$. The functions cos, sin, and tan map $[0, \pi]$, $[-\pi/2, \pi/2]$, and $(-\pi/2, \pi/2)$ bijectively onto $[-1, 1]$, $[-1, 1]$, and $\mathbb{R}$, respectively. The inverses of these restrictions are sometimes denoted by $\cos^{-1}$, $\sin^{-1}$, and $\tan^{-1}$ (a slight abuse of language), but preferably by arccos, arcsin, and arctan. The rationale for the latter notation is the fact (see the Remarks following 3.16) that every $x+iy \in \mathbb{T}_+$ is $\cos\theta + i\sin\theta$ for a unique $\theta \in (0, \pi)$, and this θ is the length of the counterclockwise *arc* of $\mathbb{T}$ with initial point 1 and terminal point $x+iy$; so $\theta = \cos^{-1}(x)$ is an arclength.

Cosh and sinh are known as *hyperbolic trigonometric* functions. Notice that cosh is an even function which effects a strictly increasing mapping of $[0, \infty)$ onto $[1, \infty)$, and sinh is an odd function which effects a strictly increasing mapping of $\mathbb{R}$ onto $\mathbb{R}$. The inverses of these restrictions are denoted arcosh and arsinh, respectively. Just as $t \mapsto (\cos t, \sin t)$, $t \in [-\pi/2, \pi/2]$, parameterizes the right half of the circle $x^2 + y^2 = 1$, the function $t \mapsto (\cosh t, \sinh t)$, $t \in \mathbb{R}$, parameterizes the right half of the "equilateral" hyperbola $x^2 - y^2 = 1$. But the parameter t appears to lose its arclength significance in the latter case. Rather it is now an area: The sector bounded by the circle and its two radii $[0, x \pm iy]$ $(x > 0, y > 0)$ has area $\arccos(x) = \arcsin(y)$, while the analogous region bounded by the hyperbola and its "radii" $[0, x \pm iy]$ has area $\operatorname{arcosh}(x) = \operatorname{arsinh}(y)$, as the reader can routinely compute. These facts elucidate the use of the adjective "hyperbolic" and the fact that the prefix "ar" is intended to suggest "area".

In the spirit of *ars gratia artis* I want to briefly divert from business to offer the reader a glimpse into "the BOOK," Paul Erdös' famous fantasy repository of proofs that the Almighty itself judges to be *ne plus ultra*. The idea of this proof goes back to Charles Hermite, in 1873, but the exposition here follows AIGNER and ZIEGLER [2010]. T. W. JONES [2010], ZHOU and MARKOV [2010] and ZHOU [2011] analyze the mechanism of the π-proof and its evolution. They also show how these ideas can prove the irrationality of $\cos(r)$ and $\tan(r)$ for every non-zero rational r. Other elementary proofs will be found in HANČL [1986], DESBROW[1990] (via complex analysis), LACZKOVICH [1997], and HUYLEBROUCK [2001].

Theorem 3.28 *π^2 is irrational, and e^m is irrational for every $m \in \mathbb{N}$.*

Proof: A central role in both proofs is played by the $2n$-degree polynomial

$$p(x) := x^n(1-x)^n/n!, \tag{1}$$

$n \in \mathbb{N}$ to be specified later. The properties needed are

$$0 < p(x) < 1/n! \qquad \forall x \in (0,1), \tag{2}$$

$$p^{(k)}(0), p^{(k)}(1) \in \mathbb{Z} \qquad \forall k \in \mathbb{Z}_+. \tag{3}$$

Since $0 < x(1-x) < 1$ for $x \in (0,1)$, (2) is clear. Evidently the polynomial $q(x) := x^n(1-x)^n$ has the form $\sum_{j=n}^{2n} c_j x^j$ with coefficients $c_j \in \mathbb{Z}$ (use, say, induction on n), and so $q^{(k)}(0) = 0$ if $k < n$ or $k > 2n$. If, on the other

*D ! !

hand, $n \le k \le 2n$, then $q^{(k)}(x) = \sum_{j=k}^{2n} j(j-1)\dots(j-k+1)c_j x^{j-k}$, so $q^{(k)}(0) = k!c_k$, an integer multiple of $n!c_k$. Hence $p^{(k)}(0) = q^{(k)}(0)/n! \in \mathbb{Z}$. And since $p(x) = p(1-x)$, $p^{(k)}(1) = (-1)^k p^{(k)}(0) \in \mathbb{Z}$.

Now suppose first that for some $m, a, b \in \mathbb{N}$

$$e^m = a/b. \tag{4}$$

Since the power series for e^{m^2} converges, we may pick $n \in \mathbb{N}$ so that

$$(m^2)^n/n! < 1/am. \tag{5}$$

With the p in (1) form

$$P := \sum_{k=0}^{2n} (-1)^k m^{2n-k} p^{(k)}.$$

Since $p^{(2n+1)} = \mathbf{0}$, we see that

$$\begin{aligned} P' = \sum_{k=0}^{2n-1} (-1)^k m^{2n-k} p^{(k+1)} &= \sum_{j=1}^{2n} (-1)^{j-1} m^{2n-(j-1)} p^{(j)} \\ &= -m \sum_{j=1}^{2n} (-1)^j m^{2n-j} p^{(j)} = -m[P - m^{2n}p]. \end{aligned}$$

Hence

$$\frac{d}{dx}[e^{mx}P(x)] = me^{mx}P(x) + e^{mx}P'(x) = m^{2n+1}e^{mx}p(x).$$

Consequently, the mean-value theorem supplies $x \in (0,1)$ such that $e^m P(1) - e^0 P(0) = m^{2n+1} e^{mx} p(x)$. That is, in view of (4),

$$N := aP(1) - bP(0) = bm^{2n+1}e^{mx}p(x), \tag{6}$$

which is an integer, due to (3) and the definition of P. From (2) and (6)

$$0 < N \le bm^{2n+1}e^m/n! \overset{(4)}{=} m^{2n+1}a/n! \overset{(5)}{<} 1,$$

inequalities which contradict the fact $N \in \mathbb{Z}$ and thus prove that (4) is untenable.

Turning to the π^2-claim, its proof being similar, a few computations can be left to the reader. Suppose then that for some $a, b \in \mathbb{N}$

$$\pi^2 = a/b, \tag{7}$$

and select $n \in \mathbb{N}$ such that

$$a^n/n! < 1/\pi. \tag{8}$$

With the p in (1) form

$$P := b^n \sum_{k=0}^{n} (-1)^k \pi^{2(n-k)} p^{(2k)}$$

and calculate that

$$P'' = -\pi^2 P + b^n \pi^{2n+2} p.$$

Use this to further calculate

$$\begin{aligned}\frac{d}{dx}\Big[\pi^{-1}P'(x)\sin\pi x - P(x)\cos\pi x\Big] &= \Big[\pi^{-1}P''(x) + \pi P(x)\Big]\sin\pi x\\ &= b^n\pi^{2n+1}p(x)\sin\pi x\end{aligned}$$

and from this and the mean-value theorem get an $x \in (0,1)$ such that

$$N := P(1) + P(0) = b^n \pi^{2n+1} p(x) \sin \pi x. \tag{9}$$

Because of (7), $b^n\pi^{2(n-k)} \in \mathbb{N}$ for each $k = 0, 1, \ldots, n$ and therefore (3) and the definition of P tell us that $P(0)$, $P(1) \in \mathbb{Z}$. Hence $N \in \mathbb{Z}$. But from (9) and (2)

$$0 < N \le b^n\pi^{2n+1}/n! \overset{(7)}{=} a^n\pi/n! \overset{(8)}{<} 1,$$

inequalities which contradict $N \in \mathbb{Z}$ and thus prove that (7) is untenable.

I think readers will agree: a proof “from the BOOK.”

Exercise 3.29 (i) $|\sin x| \le |x|$ *for all real* x.

Hint: For each $x \neq 0$ the Mean-Value Theorem provides an x' such that

$$\left|\frac{\sin x}{x}\right| = \left|\frac{\sin(x) - \sin(0)}{x - 0}\right| = \left|\sin'(x')\right| = |\cos(x')| \le 1.$$

(ii) $\cos x \ge 1 - 2x/\pi$ *for all* x *in* $[0, \pi/2]$.

Hints: The function $g(x) := \cos x + 2x/\pi$ satisfies $g(0) = g(\pi/2) = 1$. Also $g' = 2/\pi - \sin$. From equation (10) in the proof of 3.14, sin increases on $[0, \pi/2]$ from $\sin(0) = 0$ to $\sin(\pi/2) = 1 > 2/\pi$. Therefore g' has a unique zero $x_0 \in [0, \pi/2]$, $g' > 0$ on $[0, x_0)$ and $g' < 0$ on $(x_0, \pi/2]$. It follows that g is increasing on $[0, x_0]$, decreasing on $[x_0, \pi/2]$ and so $\min g[0, \pi/2] = g(0) = g(\pi/2) = 1$.

Exercise 3.30 (i) *$\sin x \geq 2x/\pi$ for all $x \in [0, \pi/2]$.*

Hints: The function $h(x) := \sin x - 2x/\pi$ satisfies $h(0) = h(\pi/2) = 0$ and $h' = -2/\pi + \cos$. Now $\cos' = -\sin$ is negative on $(0, \pi/2]$, so h' decreases from $h'(0) = -2/\pi + 1 > 0$ to $h'(\pi/2) = -2/\pi < 0$. As in the hints of 3.29(ii), it follows that $\min h[0, \pi/2] = h(0) = h(\pi/2) = 0$. (For refinements of (i), see **MR** 2006m:26027.)

(ii) *$|\sin(\pi z)| > 1/2$ for all $z \in \mathbb{C} \setminus \bigcup_{n\in\mathbb{Z}} D(n, 1/4)$.*

Hints: Consider the hyperbolic sine function

$$h(t) := \frac{1}{2}\left(e^t - e^{-t}\right) = \sum_{k=0}^{\infty} \frac{t^{2k+1}}{(2k+1)!}, \quad t \in \mathbb{R}.$$

Evidently $h(|t|) \geq |t|$. By elementary algebra, if $x, y \in \mathbb{R}$ and $z := x + iy$, then $|\sin(\pi z)|^2 = \sin^2(\pi x) + h^2(\pi y)$. Suppose $|\sin(\pi z)|^2 \leq 1/4$. Then

$$\sin^2(\pi x) \leq 1/4 \quad \text{and} \quad h^2(\pi y) \leq 1/4. \tag{1}$$

Since $(\sqrt{3} + i)/2$ and $e^{\pi i/6}$ each lie in $\mathbb{U} \cap \mathbb{H}$ and each is a cube root of i, they must be equal. So $\sin(\pi/6) = 1/2$. Since sin is an odd function, strictly increasing on $[-\pi/2, \pi/2]$, it follows that $t \in [-\pi/2, \pi/2]$ & $|\sin t| \leq 1/2 \Rightarrow |t| \leq \pi/6$. For the x in (1) choose $n \in \mathbb{Z}$ such that $|x - n| \leq 1/2$. Then $\pi x - n\pi \in [-\pi/2, \pi/2]$ and $|\sin(\pi x - n\pi)| = |\sin(\pi x)| \leq 1/2$, so $|\pi x - n\pi| \leq \pi/6$. Furthermore, since

$$(\pi y)^2 = |\pi y|^2 \leq h^2(|\pi y|) = h^2(\pi y),$$

inequality (1) yields $(\pi y)^2 \leq 1/4$ and so

$$|z - n|^2 = (x - n)^2 + y^2 \leq \left(\frac{1}{6}\right)^2 + \frac{1}{4\pi^2}. \tag{2}$$

But by $(8)'$ in the proof of 3.14, $\pi > 3$, so (2) gives

$$|z - n|^2 < \left(\frac{1}{6}\right)^2 + \left(\frac{1}{6}\right)^2 < \left(\frac{1}{4}\right)^2.$$

(iii) *Show that the only zeros in* $\mathbb{C}$ *of* sin *are those in* $\pi\mathbb{Z}$ *and the only periods in* $\mathbb{C}$ *of* sin *are those in* $2\pi\mathbb{Z}$.

Hints: $\sin z = 0 \Leftrightarrow e^{2iz} - 1 = 0 \Leftrightarrow 2iz \in 2\pi i\mathbb{Z}$. From the (routinely confirmed) identity

$$\sin(z+2p) - \sin(z) = 2\cos(z+p)\sin(p) \qquad \forall z, p \in \mathbb{C}$$

it follows that $2p$ is a period of sin if and only if $\sin(p) = 0$.

Exercise 3.31 $\int_0^{\pi/4} e^{-R\cos 2x}\,dx \leq \pi/4R$ *for all* $R > 0$.

Hint: By 3.29(ii), $e^{-R\cos 2x} \leq e^{-R(1-4x/\pi)} = e^{-R}e^{(4R/\pi)x}$, so

$$\int_0^{\pi/4} e^{-R\cos 2x}\,dx \leq \left[e^{-R}\frac{\pi}{4R}e^{(4R/\pi)x}\right]_{x=0}^{x=\pi/4} < e^{-R}\cdot\frac{\pi}{4R}\cdot e^{4(R/\pi)(\pi/4)}.$$

Exercise 3.32 *Suppose* $\lim_{N\to\infty}\sum_{n=-N}^{N} c_n z^n$ *exists uniformly on* $C(0,r)$, *for some* $r > 0$. *Denote this limit by* $f(z)$.

(i) *Prove* Parseval's Formula

$$\frac{1}{2\pi}\int_0^{2\pi} \left|f\left(re^{it}\right)\right|^2 dt = \sum_{n=-\infty}^{\infty} |c_n|^2 r^{2n} := \lim_{N\to\infty}\sum_{n=-N}^{N} |c_n|^2 r^{2n}.$$

Hints: The functions $S_N(z) := \sum_{n=-N}^{N} c_n z^n$ satisfy

$$\begin{aligned}\left|S_N\left(re^{it}\right)\right|^2 &= \left[\sum_{n=-N}^{N} c_n r^n e^{int}\right]\left[\sum_{k=-N}^{N} \bar{c}_k r^k e^{-ikt}\right]\\ &= \sum_{n=-N}^{N}\sum_{k=-N}^{N} c_n\bar{c}_k r^{n+k} e^{i(n-k)t},\end{aligned}$$

and so

$$\int_0^{2\pi} \left|S_N\left(re^{it}\right)\right|^2 dt = \sum_{n=-N}^{N}\sum_{k=-N}^{N} c_n\bar{c}_k r^{n+k}\int_0^{2\pi} e^{i(n-k)t}\,dt = \sum_{n=-N}^{N} 2\pi|c_n|^2 r^{2n}.$$

By hypothesis, $S_N\left(re^{it}\right) \to f(re^{it})$ uniformly in $t \in \mathbb{R}$ for $N \to \infty$.

(ii) *Show that if* $c_n = 0$ *for all* $n < 0$, *then*

$$\frac{1}{2\pi}\int_0^{2\pi} \left[\operatorname{Re} f\left(re^{it}\right)\right]^2 dt = (\operatorname{Re} c_0)^2 + \frac{1}{2}\sum_{n=1}^{\infty} |c_n|^2 r^{2n}$$

and infer that if additionally $|\operatorname{Re} f| \leq A$, *then*

$$2(\operatorname{Re} c_0)^2 + \sum_{n=1}^{\infty} |c_n|^2 r^{2n} \leq 2A^2.$$

Hints: Define

$$c_n^* := \begin{cases} c_n & \text{if } n > 0 \\ c_0 + \bar{c}_0 & \text{if } n = 0 \\ \bar{c}_{-n} & \text{if } n < 0 \end{cases}$$

and note that $c_k^* c_{-k}^* = |c_k^*|^2$ for all k. Hence

$$\begin{aligned} 2 \operatorname{Re} S_N\bigl(re^{it}\bigr) &= \sum_{n=0}^{N} c_n \bigl(re^{it}\bigr)^n + \sum_{n=0}^{N} \bar{c}_n \bigl(\overline{re^{it}}\bigr)^n = \sum_{k=-N}^{N} c_k^* r^{|k|} e^{ikt} \\ \bigl[2 \operatorname{Re} S_N\bigl(re^{it}\bigr)\bigr]^2 &= \sum_{j=-N}^{N} \sum_{k=-N}^{N} c_k^* c_j^* r^{|k|+|j|} e^{i(k+j)t}, \quad \text{whence} \\ \int_0^{2\pi} \bigl[2 \operatorname{Re} S_N\bigl(re^{it}\bigr)\bigr]^2 \, dt &= \sum_{k=-N}^{N} c_k^* c_{-k}^* r^{2|k|} \cdot 2\pi = 2\pi \sum_{k=-N}^{N} |c_k^*|^2 r^{2|k|} \\ &= 2\pi \left[(2 \operatorname{Re} c_0)^2 + 2 \sum_{n=1}^{N} |c_n|^2 r^{2n} \right]. \end{aligned}$$

Now let $N \to \infty$

(iii) *For every* $j \in \mathbb{N}$ *the coefficient* c_j *in* (ii) *can be expressed as an integral of* $\operatorname{Re} f$ *over* $C(0, r)$. *Find explicit formulae.*

Hint: Multiply the above sum for $2 \operatorname{Re} S_N(re^{it})$ by e^{-ijt}, integrate and let $N \to \infty$ to learn that $\pi r^j c_j = \int_0^{2\pi} e^{-ijt} \operatorname{Re} f(re^{it}) \, dt$.

(iv) *Consider* $R(z) := \sum_{j=-n}^{n} c_j z^j$ $(n \in \mathbb{N}, c_j \in \mathbb{C})$, $M(r) := \max\{|R(z)| : |z| = r\}$ *and* $m(r) := \min\{|R(z)| : |z| = r\}$ *for* $r > 0$. *The useful inequalities*

$$m^2(r) \leq \sum_{j=-n}^{n} |c_j|^2 r^{2j} \leq M^2(r) \qquad \forall r > 0$$

are a simple consequence of (i). *But they can be proved without recourse to either integrals or the exponential function, by a clever averaging technique that has broader applicability.*

Hints: (suggested by Oswaldo Rio Branco de Oliveira) Fix $u \in \mathbb{T}$ such that $1 \notin \{u^m : 1 \leq m \leq 2n\}$. Then define $K := \max\{|1 - u^m|^{-1} : 1 \leq m \leq 2n\}$ zand note that

$$(*) \qquad \left|\sum_{k=0}^{N-1} u^{km}\right| = \left|\frac{1 - u^{Nm}}{1 - u^m}\right| \leq 2K \quad \text{for} \quad 1 \leq m \leq 2n, N \in \mathbb{N}.$$

Direct calculation delivers, for $z \neq 0$ and $k \in \mathbb{Z}$

$$|R(u^k z)|^2 = \sum_{-n \leq j \leq n} |c_j|^2 |z|^{2j} + 2 \sum_{-n \leq j < l \leq n} \operatorname{Re}[\bar{c}_j \bar{z}^j c_l z^l u^{k(l-j)}].$$

Summing over k gives

$$\begin{aligned} N \sum_{-n \leq j \leq n} |c_j|^2 |z|^{2j} &\leq \sum_{k=0}^{N-1} |R(u^k z)|^2 + 2 \sum_{-n \leq j < l \leq n} |c_j c_l| |z|^{j+l} \left|\sum_{k=0}^{N-1} u^{k(l-j)}\right| \\ &\overset{(*)}{\leq} \sum_{k=0}^{N-1} |R(u^k z)|^2 + 4K \sum_{-n \leq j < l \leq n} |c_j c_l| |z|^{j+l}. \end{aligned}$$

Taking $z \in C(0, r)$, this yields

$$N \sum_{-n \leq j \leq n} |c_j|^2 r^{2j} \leq N M^2(r) + 4K \sum_{-n \leq j < l \leq n} |c_j c_l| r^{j+l} \qquad \forall r > 0, N \in \mathbb{N}.$$

Divide by N and let $N \to \infty$. ("O frabjous day! Callooh! Callay!") Obvious sign changes in this argument lead to the $m^2(r)$ inequality.

(v) *Prove* Gutzmer's Inequalities:

$$m^2(r) \leq \sum_{n=-\infty}^{\infty} |c_n|^2 r^{2n} \leq M^2(r).$$

Hints: This too follows trivially from (i), but in Weierstrass' approach to function theory a proof without integration theory was desirable, and follows routinely from (iv): Given $\varepsilon > 0$, there is an $N(\varepsilon) \in \mathbb{N}$ such that the partial sums $S_N(z) := \sum_{n=-N}^{N} c_n z^n$ satisfy

$$|S_N(z) - f(z)| \leq \varepsilon \qquad \forall z \in C(0, r), N \geq N(\varepsilon).$$

The numbers $M_N := \max\{|S_N(z)| : |z| = r\}$ consequently satisfy

$$M_N \leq \varepsilon + \sup\{|f(z)| : |z| = r\} = \varepsilon + M(r),$$

and (iv) delivers

$$\sum_{n=-N}^{N} |c_j|^2 r^{2n} \le M_N^2 \le (\varepsilon + M(r))^2 \qquad \forall N \ge N(\varepsilon).$$

Let $N \to \infty$ and then let $\varepsilon \downarrow 0$.

Exercise 3.33 *The series $\sum_{n=0}^{\infty} c_n z^n$ converges to $f(z)$ for all $z \in D(0,R)$ and some $0 < R \le \infty$.*

(i) (Liouville) *Deduce from either* 3.32(i) *or* (v) *that if f is bounded and $R = \infty$, then f is constant.*

(ii) *Deduce from either* 3.32(i) *or* (v) *that if $|f(z)| \le |f(0)|$ for all $z \in C(0,r)$ and some $0 < r < R$, then f is constant.*

Hint: Parseval gives

$$\sum_{n=0}^{\infty} |c_n|^2 r^{2n} \le \frac{1}{2\pi} \int_0^{2\pi} |f(0)|^2 \, dt = |c_0|^2,$$

whence $\sum_{n=1}^{\infty} |c_n|^2 r^{2n} \le 0$.

(iii) (Maximum-Modulus Principle for functions of power series type). $F : \Omega \mapsto \mathbb{C}$ *is bounded in the region Ω, each point of which is the center of a disk in which F is represented by a power series. Show that if there is some $a \in \Omega$ such that $|F(a)| = M := \sup |F(\Omega)|$, then F is constant.*

Hints: Rotating F lets us assume $F(a) = M$. Consider the set $S := F^{-1}(M)$. By continuity of F it is relatively closed in Ω. If $c \in S$, we can apply (ii) to infer that $F = M$ throughout some whole disk centered at c. Thus S is open. Finally, $S \ne \varnothing$, since it contains a.

Exercise 3.34 (i) *Deduce from* 3.32(i) *or* (v) *the so-called* Cauchy estimates: *If $c_{-n} = 0$ for all $n \in \mathbb{N}$, and if we set*

$$M(r) := \sup_{|z|=r} |f(z)|,$$

then for every non-negative integer k

$$\frac{\left|f^{(k)}(0)\right|}{k!} = |c_k| \le \frac{M(r)}{r^k},$$

and equality holds for some k if and only if f is a monomial of degree k.

Edwards, C. H., Jr.

The Historical Development of the Calculus, Springer-Verlag (1979), New York. **Zbl** 452.01001. **MR** 81c:01003.

Egan, M. F.

"The integrals $\int_0^{\frac{\pi}{2}} \frac{\sin^2 mx}{\sin^2 x}\,dx$, $\int_0^{\infty} \frac{\sin^2 x}{x^2}\,dx$ and $\int_0^{\infty} \frac{\sin x}{x}\,dx$," *Math. Gazette* 8 (1915–16), 299–300.

Eggleston, H. G.

"A Tauberian lemma," *Proc. Lon. Math. Soc.* (3) 1 (1951), 28–45. **Zbl** 45, p. 335. **MR** 13, p. 22.

Eilenberg, S.

"Sur les transformations d'espaces métriques en circonférence," *Fund. Math.* 24 (1935), 160–176. **FM** 61, p. 621. **Zbl** 10, p. 277.

"Transformations continues en circonférence et la topologie du plan," *Fund. Math.* 26 (1936), 61–112. **FM** 62, p. 680. **Zbl** 13, p. 420.

"An invariance theorem for subsets of S^n," *Bull. Amer. Math. Soc.* 47 (1941), 73–75. **FM** 67, p. 738. **Zbl** 25, p. 234. **MR** 2, p. 179.

Lectures on Topology, mimeographed lecture notes, New York University (1948).

Eisenack, G. and Fenske, C.

Fixpunkttheorie, Bibliographisches Institut (1978), Mannheim. **Zbl** 369.47001. **MR** 80g:58001.

Eisenberg, M. and Guy, R.

"A proof of the hairy ball theorem," *Amer. Math. Monthly* 86 (1979), 572–574. **Zbl** 433.57011. **MR** 80i:57018.

Eisermann, M.

"The fundamental theorem of algebra made effective: An elementary real-algebraic proof via Sturm chains," *Amer. Math. Monthly* 119 (2012), 715–752. **Zbl** 1267.26016. **MR** 2990932.

Elekes, M.

"Level sets of differentiable functions of two variables with non-vanishing gradient," *Jour. Math. Anal. and Applic.* 270 (2002), 369–382. **Zbl** 1006.26006. **MR** 2003d:26015.

Eljoseph, N.

"On the iteration of linear fractional transformations," *Amer. Math. Monthly* 75 (1968), 362–366. **Zbl** 159, p. 367.

Elkins, J. M.

"A Borel-Carathéodory inequality and approximation of entire functions by polynomials with restricted zeros," *Jour. Approx. Theory* 4 (1971), 274–278. **Zbl** 234.30030. **MR** 44 #4191.

Elstrodt, J.

"Iterierte Potenzen," *Math. Semesterber.* 41 (1994), 167–178. **Zbl** 817.40001. **MR** 96e:11032.

Author: Robert B. Burckel

burckel@math.ksu.edu

https//people.math.ksu.edu/
personnel_details.php?
person_id=1324

Procurando por GAUSS neste livro eu encontrei na Pág. XIX (highlights) (19) no item 6, mas olhando o anterior:

5. The Remarks following 3.22 give an interesting analytic derivation of Archimedes' estimate of π and 3.28 outlines short elementary proofs of the 'irrationality' of e and π.

Procurando por 3.22, encontrei:

Pág. 127(153) §3 The Complex Exponential Function.

Theorem 3.22 (Gregory-Leibniz series)

$$\frac{\pi}{4} = \int_0^1 \frac{dx}{1+x^2} = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1}$$

Proof: ...

$$C\left(\frac{\pi}{4}\right) = S\left(\frac{\pi}{4}\right) = \frac{1}{\sqrt{2}} \longleftarrow$$

número errado!!!

Pág. 129(155) → (8)

/ (% B

$$\frac{\pi}{6} = \frac{1}{\sqrt{3}} \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} \cdot \frac{1}{3^n} \cdot \qquad (8)$$

Important: Pag. 129 (155).

(ii) Besides using the partial sums of the series or Riemann-sum approximations to the integral there is a third (more efficient?) way to exploit 3.22 for rational approximations to π.

We get 1/630 > 22/7 − π > 1/1260, and so

$$3+\frac{10}{71} < \frac{22}{7}-\frac{1}{630} < \pi < \frac{22}{7}-\frac{1}{1260} < 3+\frac{1}{7} \;///$$

his geometric construction can be found on pp. 31-34 of C.H. Edwards [1979].

↳ Bibliography

Pag. 945 (967) —

Edwards, C.H., Jr.
The Historical Development of the Calculus - Verlag (1979).

/
-V& %-4292 8 A A $,-4292 8 A A -) .
- - ! ! 2 ! 3 1 L ! -9 : ! 4 A 1
! -

! -256- - -* ! - / 9 : ! 4 A 1
! 1 ! 8 G . / / 1"
(% ' 9 ! !

The Measurement of a Circle

As we saw in Chapter 1, the awareness (on some level) that the area of a circle is proportional to the square of its radius, $A = \pi_1 r^2$ for some constant π_1, dates back to earliest times. Similarly, the proportionality between a circle's circumference and diameter, $C = \pi_2 d$ for some constant π_2, is an ancient one. However, it is not clear when it was first realized that the two proportionality constants are the same, $\pi_1 = \pi_2 = \pi$. In the *Measurement of a Circle*, Archimedes provided the first rigorous proof of this fact by showing that the area of a circle is equal to that of a triangle with base equal to its circumference and height equal to its radius,

$$A = \tfrac{1}{2} rC. \tag{1}$$

To see that (1) implies that $\pi_1 = \pi_2$, simply substitute $A = \pi_1 r^2$ and $C = 2\pi_2 r$. He then showed that

$$3\tfrac{10}{71} < \pi < 3\tfrac{1}{7} \tag{2}$$

by explicitly establishing this inequality for the ratio of the circumference of a circle to its diameter.

Formula (1) was certainly known before Archimedes, and it was probably deduced by regarding the circle as the union of indefinitely many isosceles triangles with the center as their common vertex, and with their bases forming an inscribed regular polygon with each of its indefinitely many sides almost coinciding with a small arc of the circle. Since the height of each triangle will virtually equal the radius of the circle, and the sum of their bases will virtually equal its circumference, this picture makes the truth of (1) seem evident.

This heuristic derivation supplies the motivation for Archimedes' rigorous proof. In it he extends the method of exhaustion to what has been termed the "method of compression." Instead of dealing only with inscribed polygons, he employs both inscribed and circumscribed polygons. The area of the circle is then "compressed" between the areas of inscribed and circumscribed polygons that closely approximate the circle (Fig. 1). The following two exercises are preliminaries to the proof.

EXERCISE 1. Consider a circle with circumference C, and let P and Q be inscribed and circumscribed polygons, respectively. Show that the perimeter of P is less than C, while the perimeter of Q is greater than C. Use the facts that $\sin\theta < \theta < \tan\theta$ if θ is an angle less than $\pi/2$ radians, and that a central angle of θ radians subtends an arc of length $r\theta$ in a circle of radius r.

EXERCISE 2. In Chapter 1 we saw that, given a circle with area A and $\epsilon > 0$, there is an inscribed regular polygon P with $a(P) > A - \epsilon$. Show similarly that there is a circumscribed regular polygon Q with $a(Q) < A + \epsilon$.

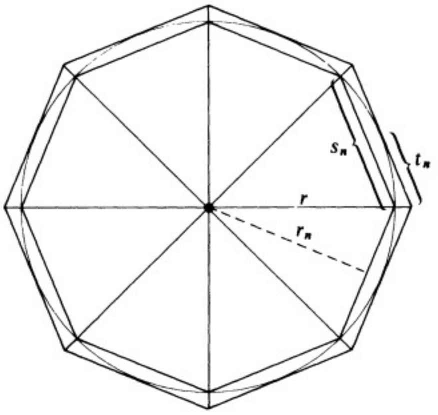

Figure 1

The proof of (1) is a typical *reductio ad absurdum* argument. Assuming that $A > \frac{1}{2}rC$, let $\epsilon = A - \frac{1}{2}rC$, and choose a regular n-sided polygon P inscribed in the circle such that

$$a(P) > A - \epsilon = \tfrac{1}{2}rC.$$

If s_n is the length of a side of P, and r_n is the length of a perpendicular from the center to a side of P, then

$$r_n < r \quad \text{and} \quad ns_n < C$$

by Exercise 1. Because P is the union of n isosceles triangles with base s_n and height r_n, it follows that

$$a(P) = n \cdot \tfrac{1}{2}r_n s_n = \tfrac{1}{2}r_n(ns_n) < \tfrac{1}{2}rC.$$

But this is a contradiction, so A is not greater than $\frac{1}{2}rC$.

Assuming that $A < \frac{1}{2}rC$, let $\epsilon = \frac{1}{2}rC - A$, and choose a regular n-sided circumscribed polygon Q such that

$$a(Q) < A + \epsilon = \tfrac{1}{2}rC.$$

For the purpose of subsequent computations, let t_n denote *half* the length of a side of Q. Then

$$a(Q) = n \cdot \tfrac{1}{2}r(2t_n) = \tfrac{1}{2}r(2nt_n) > \tfrac{1}{2}rC,$$

because the perimeter $2nt_n$ of Q is greater than C (Exercise 1). This contradiction completes the proof of (1). □

EXERCISE 3. Let A_n and C_n denote the area and perimeter, respectively, of a regular polygon with n sides inscribed in a circle of radius r. Show that

$$A_n = nr^2 \sin\frac{\pi}{n}\cos\frac{\pi}{n} \quad \text{and} \quad C_n = 2nr\sin\frac{\pi}{n}.$$

Deduce that $A = \frac{1}{2}rC$ by taking the limit of A_n/C_n as $n \to \infty$.

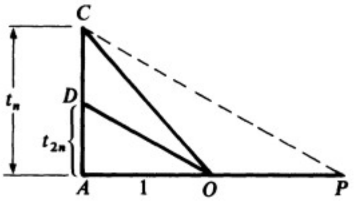

Figure 2

EXERCISE 4. With A_n and C_n as in Exercise 3, and r_n, s_n as in Figure 1, write $A_n = \frac{1}{2}nr_ns_n$ and $C_n = ns_n$. Deduce that $A = \frac{1}{2}rC$ without using trigonometric functions, by taking the limit of A_n/C_n as $n \to \infty$. What must you assume to be "obvious"?

In order to obtain the approximation (2) for π, Archimedes began with regular hexagons inscribed in and circumscribed about a circle of radius one. By successively doubling the number of sides he obtained pairs of inscribed and circumscribed regular polygons with 12, 24, 48, and 96 sides, and calculated their perimeters to find upper and lower bounds for π.

Consider first the circumscribed polygons. If t_n denotes *half* the length of a side of a regular circumscribed polygon with n sides, then the relationship between t_n and t_{2n} is indicated in Figure 2, where O is the center of the circle, and OD bisects angle AOC. If CP is parallel to OD, it is easily seen that $OP = CO$. Since triangles ADO and ACP are similar, it follows that

$$\frac{AD}{AO} = \frac{AC}{AO+OP} = \frac{AC}{AO+OC},$$

or

$$t_{2n} = \frac{t_n}{1+\sqrt{1+t_n^2}}. \tag{3}$$

Now consider the inscribed polygons. If s_n denotes the side of the regular inscribed polygon with n sides, the relationship between s_n and s_{2n} is indicated by Figure 3, where $s_n = BC$, $s_{2n} = BD$, and AD bisects angle BAC (Why?). It is easily checked that the triangles ABD, BPD, and APC

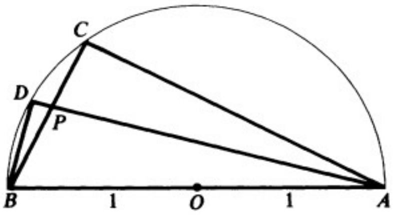

Figure 3

are similar. Hence

$$\frac{AB}{AD} = \frac{BP}{BD} \quad \text{and} \quad \frac{AC}{AD} = \frac{PC}{BD},$$

so

$$\frac{AB + AC}{AD} = \frac{BP + PC}{BD} = \frac{BC}{BD},$$

or

$$\frac{2 + \sqrt{4 - s_n^2}}{\sqrt{4 - s_{2n}^2}} = \frac{s_n}{s_{2n}}.$$

After cross-multiplying and squaring, the resulting equation yields

$$s_{2n}^2 = \frac{s_n^2}{2 + \sqrt{4 - s_n^2}}. \tag{4}$$

EXERCISE 5. Observe from the familiar geometry of the regular hexagon that $s_6 = 1$, $t_6 = 1/\sqrt{3}$. Apply formulas (3) and (4) to compute $s_{12}, s_{24}, s_{48}, s_{96}$ and $t_{12}, t_{24}, t_{48}, t_{96}$ recursively (using a hand calculator) so as to obtain

$$s_{96} = 0.065438 \quad \text{and} \quad t_{96} = 0.032737.$$

Since $48 s_{96} < \pi < 96 t_{96}$, conclude that

$$3\tfrac{10}{71} < \pi < 3\tfrac{1}{7}.$$

Hence $\pi = 3.14$, rounded off to two decimal places.

Of course, Archimedes did not have a hand calculator available. He started with the approximation

$$\tfrac{265}{153} < \sqrt{3} < \tfrac{1351}{780}$$

and proceeded manually, carefully rounding down in calculating the s_n's and rounding up in calculating the t_n's, finally obtaining

$$3\frac{10}{71} < \frac{6336}{2017\frac{1}{4}} < \pi < \frac{14688}{4673\frac{1}{2}} < 3\frac{1}{7}.$$

It is generally believed that the extant *Measurement of a Circle* is only a fragment of Archimedes' original and more comprehensive treatment of the circle. In a recent article W. R. Knorr [10] argues persuasively that, to obtain a more accurate approximation to π, Archimedes started with inscribed and circumscribed decagons (regular 10-sided polygons) and successively doubled sides six times to obtain inscribed and circumscribed regular polygons with 640 sides.

Exercise 6. Starting with the fact that the side of a regular decagon inscribed in the unit circle is $s_{10} = 2\sin 18° = (\sqrt{5} - 1)/2$, apply formulas (3) and (4) to recursively calculate s_{640} and t_{640}, carrying 8 decimal places on a hand calculator. Thence verify that $\pi = 3.1416$, rounded off to 4 decimal places.

Exercise 7. Let p_n and P_n denote the perimeters of the inscribed and circumscribed regular n-sided polygons for the unit circle. Noting that

$$s_n = 2\sin\frac{\pi}{n} \quad \text{and} \quad t_n = \tan\frac{\pi}{n},$$

show that

$$p_{2n} = \sqrt{p_n P_{2n}} \quad \text{and} \quad P_{2n} = \frac{2p_n P_n}{p_n + P_n}.$$

Starting with $p_4 = 4\sqrt{2}$ and $P_4 = 8$ for inscribed and circumscribed squares, use these recursive formulas to calculate p_{64} and P_{64}. What bounds on π does this computation give?

Exercise 8. If a_n and A_n are the areas of the inscribed and circumscribed polygons in the previous exercise, show that

$$a_{2n} = \sqrt{a_n A_n} \quad \text{and} \quad A_{2n} = \frac{2a_{2n} A_n}{a_{2n} + A_n}.$$

The Quadrature of the Parabola

A *segment* of a convex curve is a region bounded by a straight line and a portion of the given curve (Fig. 4). In the preface to the *Quadrature of the Parabola*, Archimedes remarks that earlier mathematicians had successfully attempted to find the area of a segment of a circle or hyperbola, but that apparently no one had previously attempted the quadrature of a segment of a parabola—precisely the one that can be carried out by the method of exhaustion.

The parabola was originally defined by the Greeks as a conic section. That is, given a circular (double) cone with vertical axis, a parabola is the

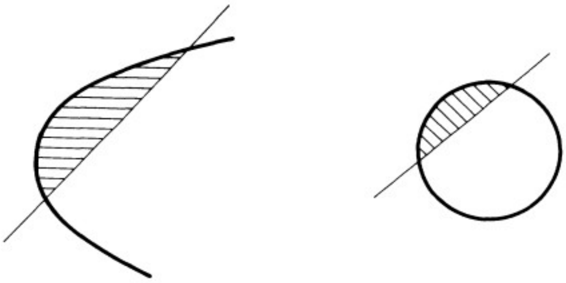

Figure 4

D -256- - -* ! - / ! ! 2 ! 3 1 L
! 1"
(/ $$

Remarks: (i) There is an alternative proof of the limit in 3.21 which is short and transparent but a bit less elementary, and affords no estimate of the speed of convergence. Let L be the function L_1 of 3.19. For integers $n > N > 0$, $z \in \overline{D}(0, N) \setminus \{0\}$ implies $1 + z/n \in D(1, 1)$. Consequently, by 3.19

$$nL(1 + z/n) = z \cdot \frac{L(1 + z/n) - L(1)}{z/n} \to zL'(1) = z \quad \text{as} \quad n \to \infty,$$

uniformly in such z, whence $(1+z/n)^n = [\exp L(1+z/n)]^n = \exp(nL(1+z/n)) \to \exp(z)$. [Similarly, $(1 + z_n/n)^n \to e^z$ whenever $z_n \to z$.]

(ii) According to 3.21 the numbers $a_n := n!\,n^{-n}$ satisfy $a_{n+1}/a_n = (1 + 1/n)^{-n} \to 1/e$, in view of which 3.9 tells us that $a_n^{1/n} = n^{-1}(n!)^{1/n} \to 1/e$. After a little algebra, involving $n^{-1}\sum_{j=1}^{n} j = \frac{1}{2}(n+1)$, another interesting feature of e comes to light:

$$\frac{1}{2}e = \lim_{n\to\infty} \frac{\text{Arithmetic Mean of } \{1, 2, \ldots, n\}}{\text{Geometric Mean of } \{1, 2, \ldots, n\}}.$$

(A generalization to means of $\{1^s, 2^s, \ldots, n^s\}$ $(s > -1)$ was proved by different [elementary] means by Xu [2010].)

Although the constant π is ubiquitous in this subject (see Castellanos [1988]), one seldom needs estimates of it finer than (8) and (8)$'$ in the proof of 3.14. Nevertheless, we will look at a few of the myriad series and integrals by which it can be explicitly calculated.

Theorem 3.22 (Gregory-Leibniz series) $\dfrac{\pi}{4} = \displaystyle\int_0^1 \frac{dx}{1+x^2} = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1}.$

Proof: The quadratic polynomial $X^2 - i$ has exactly two zeros in $\mathbb{C}$, $\pm(1 + i)/\sqrt{2}$. Since, in the notation of the proof of 3.14,

$$\left[C\left(\frac{\pi}{4}\right) + iS\left(\frac{\pi}{4}\right)\right]^2 = E\left(i\frac{\pi}{4}\right)^2 = E\left(i\frac{\pi}{2}\right) = i$$

and C, S are non-negative in $\left[0, \frac{\pi}{2}\right]$, it must be that

$$C\left(\frac{\pi}{4}\right) = S\left(\frac{\pi}{4}\right) = \frac{1}{\sqrt{2}}. \tag{1}$$

Since C is an even function and is never zero in $[0, \pi/2)$, by definition of π, we may form the function $T := S/C$ on $(-\pi/2, \pi/2)$ and have from equations (5) of the proof of 3.14 that

$$T' = \frac{S'C - SC'}{C^2} = \frac{C^2 + S^2}{C^2} = 1 + (S/C)^2 = 1 + T^2 > 0. \tag{2}$$

Therefore T is strictly increasing. Since $T(0) = 0$ and $T(t) \to +\infty$ as $t \uparrow \frac{\pi}{2}$, the range of the odd function T is $\mathbb{R}$. Now $T^{-1}\colon \mathbb{R} \mapsto (-\pi/2, \pi/2)$ exists and by (2)

$$\int_0^x \frac{1}{1+t^2}\,dt = \int_{T^{-1}(0)}^{T^{-1}(x)} \frac{T'}{1+T^2} = \int_{T^{-1}(0)}^{T^{-1}(x)} 1 = T^{-1}(x) \qquad \forall x \in \mathbb{R} \tag{3}$$

and in particular, because of (1),

$$\int_0^1 \frac{1}{1+t^2}\,dt = T^{-1}(1) = \frac{\pi}{4}. \tag{3$'$}$$

For every $N \in \mathbb{N}$, $t \in \mathbb{R}$

$$\sum_{n=0}^{N-1} \left(-t^2\right)^n = \frac{1-\left(-t^2\right)^N}{1-(-t^2)} = \frac{1}{1+t^2} - \frac{(-1)^N t^{2N}}{1+t^2}$$

and therefore

$$\begin{aligned} \left|\sum_{n=0}^{N-1} \frac{(-1)^n x^{2n+1}}{2n+1} - \int_0^x \frac{1}{1+t^2}dt\right| &= \left|\int_0^x \left[\sum_{n=0}^{N-1} \left(-t^2\right)^n - \frac{1}{1+t^2}\right]dt\right| \\ &= \left|\int_0^x \frac{(-1)^{N-1} t^{2N}}{1+t^2}dt\right| \qquad \forall x \in \mathbb{R}. \end{aligned} \tag{4}$$

The trivial inequalities $\frac{1}{2}t^{2N} \le t^{2N}(1+t^2)^{-1} \le t^{2N}$ for $t \in [-1,1]$ let us estimate the last integral in (4), below and above, yielding for $x \in [-1,1]\,, N \in \mathbb{N}$.

$$\frac{|x|^{2N+1}}{4N+2} \le \left|\sum_{n=0}^{N-1} \frac{(-1)^n x^{2n+1}}{2n+1} - \int_0^x \frac{1}{1+t^2}dt\right| \le \frac{|x|^{2N+1}}{2N+1}. \tag{5}$$

In particular, using (3)

$$T^{-1}(x) = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{2n+1} \qquad \forall x \in [-1,1]\,. \tag{6}$$

Now equations (3)$'$ and (6) prove the theorem, but (3)$'$ and (5) also furnish the estimate

$$\frac{1}{4N+2} \le \left|\sum_{n=0}^{N-1} \frac{(-1)^n}{2n+1} - \frac{\pi}{4}\right| \le \frac{1}{2N+1} \qquad \forall N \in \mathbb{N}. \tag{7}$$

Remarks: (i) See 3.37(vii), 3.38(viii) and 11.37(viii) [with $z := 1/2$] for different proofs. This series converges excruciatingly slowly, as (7) shows; for one converging to π with geometric rapidity, see 5.82(ii). After checking (see 3.30(ii) Hints) that $T(\pi/6) = 1/\sqrt{3}$, one can also take $x := 1/\sqrt{3}$ in (6) and get the even more rapidly converging representation

$$\frac{\pi}{6} = \frac{1}{\sqrt{3}} \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} \cdot \frac{1}{3^n}. \tag{8}$$

In 3.45(v) we will see that for each $x \in \overline{\mathbb{D}} \setminus \{-i, i\}$ the series (6) converges to a solution w of the equation $\tan w = x$. (See def. 3.27.)

(ii) Besides using the partial sums of the series or Riemann-sum approximations to the integral, there is a third (more efficient?) way to exploit 3.22 for rational approximations to π. Each $n \in \mathbb{N}$ determines a polynomial p_n with integer coefficients such that $x^{4n}(1-x)^{4n} = p_n(x)(1+x^2) + (-4)^n$. (One checks that $p_1(x) := x^6 - 4x^5 + 5x^4 - 4x^2 + 4$ works and the rest is induction.) Therefore $\int_0^1 x^{4n}(1-x)^{4n}(1+x^2)^{-1}dx$ is a rational number minus $(-4)^{n-1}\pi$. This integral is positive and less than $[\max\{x(1-x) : 0 \le x \le 1\}]^{4n} = 4^{-4n}$. A rational approximation to π closer than $4^{-4n-(n-1)} = 4^{-5n+1}$ emerges. In case $n = 1$, $\int_0^1 x^4(1-x)^4(1+x^2)^{-1}dx = 22/7 - \pi$. Since

$$\frac{1}{630} = \int_0^1 x^4(1-x)^4 dx > \int_0^1 x^4(1-x)^4(1+x^2)^{-1}dx > \int_0^1 x^4(1-x)^4 2^{-1}dx = \frac{1}{1260},$$

we get $1/630 > 22/7 - \pi > 1/1260$, and so

$$3 + \frac{10}{71} < \frac{22}{7} - \frac{1}{630} < \pi < \frac{22}{7} - \frac{1}{1260} < 3 + \frac{1}{7}.$$

This is from DALZELL [1944], with more in LUCAS [2009]. The extreme inequalities are Archimedes' famous estimate; his geometric construction can be found on pp. 31–34 of C. H. EDWARDS [1979].

(iii) Yet a fourth way to utilize 3.22 is to express $\pi/4 = \arctan(1)$ as a rational linear combination of $\arctan(\varepsilon_j)$ for small ε_j (for which the series is more rapidly convergent). One of these goes back to Euler: $\pi/4 = 4\arctan(1/5) - \arctan(1/70) + \arctan(1/99)$. There are numerous such relations; see pp. 274–275 of BERGGREN, BORWEIN and BORWEIN [2004], pp. 59–65 of EYMARD and LAFON [2004] and pp. 90–98 of CASTELLANOS [1988].

(iv) In 1755, Euler found a series which, in contrast to (6), converges to $T^{-1}(x)$ throughout $\mathbb{R}$:

$$T^{-1}(x) = \frac{x}{1+x^2} \sum_{k=0}^{\infty} \frac{2^{2k}(k!)^2}{(2k)!} \; \frac{1}{2k+1} \left(\frac{x^2}{1+x^2}\right)^k \qquad \forall x \in \mathbb{R}. \tag{9}$$

To confirm it, change the variable of integration in (3) via $t =: xC(\theta)$, $x \neq 0$, $\theta \in [0, \pi/2]$, to get

$$(1+x^2)/(1+t^2) = 1/[1 - \frac{x^2}{1+x^2} S^2(\theta)] = \sum_{k=0}^{\infty} \left(\frac{x^2}{1+x^2}\right)^k S^{2k}(\theta).$$

After commuting the uniformly converging series in $S(\theta)$ with the integration, the k^{th} coefficient is $\int_0^{\pi/2} S^{2k+1}(\theta) d\theta$; that it equals $2^{2k}(k!)^2/(2k+1)!$ is confirmed by induction, reduction of k being accomplished by an integration by parts. [The Yin & Yang relation between (9) and 5.82(iv) is noteworthy.]

(v) Since $(T^{-1})'(t) = 1/(1+t^2)$, we have

$$\frac{1}{2}(T^{-1}(x))^2 = \int_0^x T^{-1}(t)/(1+t^2)\ dt.$$

Use (9) on the integrand, permute integration and summation, and make the change of integration variable $u := t^2/(1+t^2)$. You get a combinatorially remarkable companion to (9):

$$(9)^2 \qquad (T^{-1}(x))^2 = \sum_{k=0}^{\infty} \frac{2^{2k+1}(k!)^2}{(2k+2)!} \left(\frac{x^2}{1+x^2}\right)^{k+1} \qquad \forall x \in \mathbb{R}.$$

Exercise 3.23 *With the convention $0^0 := 1$, the function x^{-x} is continuous on $[0,1]$. (Cf. 3.18(4) and recall the customary definition $x^{-x} := e^{-x \log x}$ for $x > 0$.) Prove the satisfyingly symmetric formula*

$$\int_0^1 x^{-x}\, dx = \sum_{n=1}^{\infty} n^{-n}.$$

Hints: $x^{-x} = \sum_0^{\infty} (-1)^n (x \log x)^n / n!$. Interchange summation and integration and evaluate $\int_0^1 x^n \log^n(x)\, dx$ by repeated integration-by-parts.

Exercise 3.24 *Consider $f(x) := \int_0^1 (1+t^2)^{-1} e^{-x^2(1+t^2)} dt$ and $g(x) := \int_0^x e^{-t^2}\, dt$, $x \in [0, \infty)$.*

(i) *Show that $0 < f(x) < e^{-x^2} f(0)$ for all $x > 0$.*

(ii) *Use 2.19(i) to show that $f'(x) = -2e^{-x^2} g(x)$ for all $x \geq 0$; that is, $f' = \left(-g^2\right)'$.*

(iii) *Prove that $\int_0^{\infty} e^{-t^2}\, dt = \sqrt{\pi}/2$.*

here, the definition of cos and sin!] can consult the two-volume work of Naux [1966], [1971] (being wary of its deficiencies, as pointed out in the *Zentralblatt* reviews) and the numerous papers of J. W. L. Glaisher, Florian Cajori, and Joseph E. Hofmann.

3.22 inspired Leibniz to quote Virgil (on gambling!): "Numero Deus impari gaudet." Its interesting history and independent discovery by an Indian mathematician are detailed in the prize-winning paper Roy [1990]; also Parameśwaran [1992]. A beautiful elementary derivation is given by Kazarinoff [1964], Matsumoto [1995] and Zheng [2002]. Essentially the first equality in 3.22 was adopted by Weierstrass (in 1841) as his *definition* of π.) (This was also the definition adopted in the alternative treatment of the exponential function given in 2.22.) The present definition of π as "twice the least positive zero of the cosine function" dates at latest from an 1875 textbook of Richard Baltzer, and appears in Pringsheim [1925] and E. Landau [1950]. When Landau used it in his Göttingen lectures in 1934, it occasioned a vicious attack by the Nazis—see chapter 6 of Ebbinghaus, *et al.* [1990].

3.23 is from the Euler-Johann Bernoulli correspondence (see **MR** 88f:01011, **Zbl** 631.01011 for its history); different proofs may be found in Hoppe [1877] and Selberg [1933], and the 3.23 method is pushed a little farther in Glaister [2001]. Lord [2003a] shows that many other continuous functions g satisfy $\int_0^1 g(x)\,dx = \sum_{n=1}^{\infty} g(n)$, including $g(x) := c^x$ for a unique $c > 0$.

A very nice alternative proof of 3.25 can be found on page 35 of Guichardet [1968] or page 235 of Stromberg [1981]. Yet another proof runs thus: use an analog of 4.1 to find a continuous function $\phi\colon \mathbb{R} \mapsto \mathbb{C}$ such that $f = e^{\phi}$ and $\phi(0) = 0$. Deduce that ϕ is additive, whence $\phi(r) = r\phi(1)$ for all rational r, whence by continuity $\phi(x) = cx$ for all $x \in \mathbb{R}$ $(c := \phi(1))$. The integration trick in the text originates with Schimmack [1908], rediscovered by Kac [1936/37]. It is further capable of showing that the only endomorphisms of $(\mathbb{R}, +)$ which are integrable over $[0, 1]$ are the constant multiples of the identity. Since the rate of change of many quantities in nature is (approximately) a constant multiple of their momentary magnitude, 3.25(i) is the reason for the ubiquity of the exponential function in science.

Though proofs of 3.32(i) appear in Gutzmer [1887], [1888], [1889], the result pre-dates this; he acknowledges the priority of A. Harnack, but in fact the formula goes back to an 1806 memoir of M. A. Parseval (see §23 of Burkhardt [1914]) and is customarily referred to in the literature as Parseval's formula. See remark (iii) following 5.20 and Chapter XI. The Cauchy estimates date from 1831 and were discovered independently in 1841 by Weierstrass (*Werke* I, pp. 67–68 and II, pp. 224–226 and Ullrich [1990]), who gave an integral-free treatment. On the latter see also Méray [1891b], pp. 278, 280, 606 of Pringsheim [1925], A. Kneser [1920], Montel [1936], and pp.

Observe that for $0 < x \leq 2$ and $n \geq 1$

$$\frac{x^2}{(4n)!} \leq \frac{4}{(4n)(4n-1)(4n-2)!} < \frac{1}{(4n-2)!}$$

and so

$$\begin{aligned} C(x) &= 1 + \left[\frac{x^2}{4!} - \frac{1}{2!}\right] x^{4-2} + \sum_{n=2}^{\infty} \left[\frac{x^2}{(4n)!} - \frac{1}{(4n-2)!}\right] x^{4n-2} \\ &< 1 + \left[\frac{x^2}{4!} - \frac{1}{2!}\right] x^{4-2}. \end{aligned}$$

Therefore $C(\sqrt{3}) < -\frac{1}{8}$. Since $C(0) = \operatorname{Re} E(0) = 1$, we have that $C\big[0, \sqrt{3}\,\big]$ is a connected subset of $\mathbb{R}$ which contains positives and negatives and hence zero; that is, the compact set $\big[0, \sqrt{3}\,\big] \cap C^{-1}(0)$ is not void. Let $\pi/2$ denote its least element:

$$\pi/2 := \min\Big(\big[0, \sqrt{3}\,\big] \cap C^{-1}(0)\Big). \tag{7}$$

Thus

$$0 < \pi/2 < \sqrt{3} \quad \text{and} \quad C(\pi/2) = 0. \tag{8}$$

[The notation π, meant to suggest "perimeter" (of a circle having unit diameter), anticipates the remarks following 3.16.] An insertion-of-parentheses argument similar to the above shows that

$$C(x) > 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} =: P(x), \quad x \in \big(0, 3\sqrt{10}\big].$$

This polynomial is decreasing on $\big[0, \sqrt{6}\big]$ (look at its derivative), so

$$C(x) > P(x) \geq P(3/2) > 0 \qquad \forall x \in (0, 3/2]\,.$$

We infer that $\pi/2 > 3/2$, i.e.,

$$\pi > 3. \tag{8$'$}$$

From (8) and (6) we get

$$S(\pi/2) = \pm 1. \tag{9}$$

However by (5) and the definition of π

$$S' = C > 0 \quad \text{in } [0, \pi/2)\,, \tag{10}$$

the fact (see the Remarks following 3.16) that every $x+iy \in \mathbb{T}_+$ is $\cos\theta + i\sin\theta$ for a unique $\theta \in (0,\pi)$, and this θ is the length of the counterclockwise *arc* of $\mathbb{T}$ with initial point 1 and terminal point $x+iy$; so $\theta = \cos^{-1}(x)$ is an arclength.

Cosh and sinh are known as *hyperbolic trigonometric* functions. Notice that cosh is an even function which effects a strictly increasing mapping of $[0,\infty)$ onto $[1,\infty)$, and sinh is an odd function which effects a strictly increasing mapping of $\mathbb{R}$ onto $\mathbb{R}$. The inverses of these restrictions are denoted arcosh and arsinh, respectively. Just as $t \mapsto (\cos t, \sin t)$, $t \in [-\pi/2, \pi/2]$, parameterizes the right half of the circle $x^2 + y^2 = 1$, the function $t \mapsto (\cosh t, \sinh t)$, $t \in \mathbb{R}$, parameterizes the right half of the "equilateral" hyperbola $x^2 - y^2 = 1$. But the parameter t appears to lose its arclength significance in the latter case. Rather it is now an area: The sector bounded by the circle and its two radii $[0, x \pm iy]$ $(x > 0, y > 0)$ has area $\arccos(x) = \arcsin(y)$, while the analogous region bounded by the hyperbola and its "radii" $[0, x \pm iy]$ has area $\operatorname{arcosh}(x) = \operatorname{arsinh}(y)$, as the reader can routinely compute. These facts elucidate the use of the adjective "hyperbolic" and the fact that the prefix "ar" is intended to suggest "area".

In the spirit of *ars gratia artis* I want to briefly divert from business to offer the reader a glimpse into "the BOOK," Paul Erdös' famous fantasy repository of proofs that the Almighty itself judges to be *ne plus ultra.* The idea of this proof goes back to Charles Hermite, in 1873, but the exposition here follows AIGNER and ZIEGLER [2010]. T. W. JONES [2010], ZHOU and MARKOV [2010] and ZHOU [2011] analyze the mechanism of the π-proof and its evolution. They also show how these ideas can prove the irrationality of $\cos(r)$ and $\tan(r)$ for every non-zero rational r. Other elementary proofs will be found in HANČL [1986], DESBROW[1990] (via complex analysis), LACZKOVICH [1997], and HUYLEBROUCK [2001].

Theorem 3.28 *π^2 is irrational, and e^m is irrational for every $m \in \mathbb{N}$.*

Proof: A central role in both proofs is played by the $2n$-degree polynomial

$$p(x) := x^n(1-x)^n/n!, \tag{1}$$

$n \in \mathbb{N}$ to be specified later. The properties needed are

$$0 < p(x) < 1/n! \qquad \forall x \in (0,1), \tag{2}$$

$$p^{(k)}(0), p^{(k)}(1) \in \mathbb{Z} \qquad \forall k \in \mathbb{Z}_+. \tag{3}$$

Since $0 < x(1-x) < 1$ for $x \in (0,1)$, (2) is clear. Evidently the polynomial $q(x) := x^n(1-x)^n$ has the form $\sum_{j=n}^{2n} c_j x^j$ with coefficients $c_j \in \mathbb{Z}$ (use, say, induction on n), and so $q^{(k)}(0) = 0$ if $k < n$ or $k > 2n$. If, on the other hand, $n \le k \le 2n$, then $q^{(k)}(x) = \sum_{j=k}^{2n} j(j-1)\dots(j-k+1)c_j x^{j-k}$, so $q^{(k)}(0) = k!c_k$, an integer multiple of $n!c_k$. Hence $p^{(k)}(0) = q^{(k)}(0)/n! \in \mathbb{Z}$. And since $p(x) = p(1-x)$, $p^{(k)}(1) = (-1)^k p^{(k)}(0) \in \mathbb{Z}$.

Now suppose first that for some $m, a, b \in \mathbb{N}$

$$e^m = a/b. \tag{4}$$

Since the power series for e^{m^2} converges, we may pick $n \in \mathbb{N}$ so that

$$(m^2)^n/n! < 1/am. \tag{5}$$

With the p in (1) form

$$P := \sum_{k=0}^{2n} (-1)^k m^{2n-k} p^{(k)}.$$

Since $p^{(2n+1)} = \mathbf{0}$, we see that

$$\begin{aligned} P' = \sum_{k=0}^{2n-1} (-1)^k m^{2n-k} p^{(k+1)} &= \sum_{j=1}^{2n} (-1)^{j-1} m^{2n-(j-1)} p^{(j)} \\ &= -m \sum_{j=1}^{2n} (-1)^j m^{2n-j} p^{(j)} = -m[P - m^{2n}p]. \end{aligned}$$

Hence

$$\frac{d}{dx}[e^{mx}P(x)] = me^{mx}P(x) + e^{mx}P'(x) = m^{2n+1}e^{mx}p(x).$$

Consequently, the mean-value theorem supplies $x \in (0,1)$ such that $e^m P(1) - e^0 P(0) = m^{2n+1}e^{mx}p(x)$. That is, in view of (4),

$$N := aP(1) - bP(0) = bm^{2n+1}e^{mx}p(x), \tag{6}$$

which is an integer, due to (3) and the definition of P. From (2) and (6)

$$0 < N \le bm^{2n+1}e^m/n! \overset{(4)}{=} m^{2n+1}a/n! \overset{(5)}{<} 1,$$

inequalities which contradict the fact $N \in \mathbb{Z}$ and thus prove that (4) is untenable.

Turning to the π^2-claim, its proof being similar, a few computations can be left to the reader. Suppose then that for some $a, b \in \mathbb{N}$

$$\pi^2 = a/b, \tag{7}$$

and select $n \in \mathbb{N}$ such that

$$a^n/n! < 1/\pi. \tag{8}$$

With the p in (1) form

$$P := b^n \sum_{k=0}^{n} (-1)^k \pi^{2(n-k)} p^{(2k)}$$

and calculate that

$$P^{''} = -\pi^2 P + b^n \pi^{2n+2} p.$$

Use this to further calculate

$$\begin{aligned} \frac{d}{dx}\Big[\pi^{-1}P'(x)\sin\pi x - P(x)\cos\pi x\Big] &= \Big[\pi^{-1}P^{''}(x) + \pi P(x)\Big]\sin\pi x \\ &= b^n\pi^{2n+1}p(x)\sin\pi x \end{aligned}$$

and from this and the mean-value theorem get an $x \in (0,1)$ such that

$$N := P(1) + P(0) = b^n \pi^{2n+1} p(x) \sin \pi x. \tag{9}$$

Because of (7), $b^n \pi^{2(n-k)} \in \mathbb{N}$ for each $k = 0, 1, \ldots, n$ and therefore (3) and the definition of P tell us that $P(0)$, $P(1) \in \mathbb{Z}$. Hence $N \in \mathbb{Z}$. But from (9) and (2)

$$0 < N \le b^n \pi^{2n+1}/n! \overset{(7)}{=} a^n \pi/n! \overset{(8)}{<} 1,$$

inequalities which contradict $N \in \mathbb{Z}$ and thus prove that (7) is untenable.

I think readers will agree: a proof “from the BOOK.”

Exercise 3.29 (i) $|\sin x| \le |x|$ *for all real* x.

Hint: For each $x \neq 0$ the Mean-Value Theorem provides an x' such that

$$\left|\frac{\sin x}{x}\right| = \left|\frac{\sin(x) - \sin(0)}{x - 0}\right| = \left|\sin'(x')\right| = |\cos(x')| \le 1.$$

(ii) $\cos x \ge 1 - 2x/\pi$ *for all* x *in* $[0, \pi/2]$.

Hints: The function $g(x) := \cos x + 2x/\pi$ satisfies $g(0) = g(\pi/2) = 1$. Also $g' = 2/\pi - \sin$. From equation (10) in the proof of 3.14, sin increases on $[0, \pi/2]$ from $\sin(0) = 0$ to $\sin(\pi/2) = 1 > 2/\pi$. Therefore g' has a unique zero $x_0 \in [0, \pi/2]$, $g' > 0$ on $[0, x_0)$ and $g' < 0$ on $(x_0, \pi/2]$. It follows that g is increasing on $[0, x_0]$, decreasing on $[x_0, \pi/2]$ and so $\min g[0, \pi/2] = g(0) = g(\pi/2) = 1$.

Exercise 3.30 (i) *$\sin x \geq 2x/\pi$ for all $x \in [0, \pi/2]$.*

Hints: The function $h(x) := \sin x - 2x/\pi$ satisfies $h(0) = h(\pi/2) = 0$ and $h' = -2/\pi + \cos$. Now $\cos' = -\sin$ is negative on $(0, \pi/2]$, so h' decreases from $h'(0) = -2/\pi + 1 > 0$ to $h'(\pi/2) = -2/\pi < 0$. As in the hints of 3.29(ii), it follows that $\min h[0, \pi/2] = h(0) = h(\pi/2) = 0$. (For refinements of (i), see **MR** 2006m:26027.)

(ii) *$|\sin(\pi z)| > 1/2$ for all $z \in \mathbb{C} \setminus \bigcup_{n\in\mathbb{Z}} D(n, 1/4)$.*

Hints: Consider the hyperbolic sine function

$$h(t) := \frac{1}{2}\left(e^t - e^{-t}\right) = \sum_{k=0}^{\infty} \frac{t^{2k+1}}{(2k+1)!}, \quad t \in \mathbb{R}.$$

Evidently $h(|t|) \geq |t|$. By elementary algebra, if $x, y \in \mathbb{R}$ and $z := x + iy$, then $|\sin(\pi z)|^2 = \sin^2(\pi x) + h^2(\pi y)$. Suppose $|\sin(\pi z)|^2 \leq 1/4$. Then

$$\text{(1)} \qquad \sin^2(\pi x) \leq 1/4 \quad \text{and} \quad h^2(\pi y) \leq 1/4.$$

Since $(\sqrt{3} + i)/2$ and $e^{\pi i/6}$ each lie in $\mathbb{U} \cap \mathbb{H}$ and each is a cube root of i, they must be equal. So $\sin(\pi/6) = 1/2$. Since sin is an odd function, strictly increasing on $[-\pi/2, \pi/2]$, it follows that $t \in [-\pi/2, \pi/2]$ & $|\sin t| \leq 1/2 \Rightarrow |t| \leq \pi/6$. For the x in (1) choose $n \in \mathbb{Z}$ such that $|x - n| \leq 1/2$. Then $\pi x - n\pi \in [-\pi/2, \pi/2]$ and $|\sin(\pi x - n\pi)| = |\sin(\pi x)| \leq 1/2$, so $|\pi x - n\pi| \leq \pi/6$. Furthermore, since

$$(\pi y)^2 = |\pi y|^2 \leq h^2(|\pi y|) = h^2(\pi y),$$

inequality (1) yields $(\pi y)^2 \leq 1/4$ and so

$$\text{(2)} \qquad |z - n|^2 = (x - n)^2 + y^2 \leq \left(\frac{1}{6}\right)^2 + \frac{1}{4\pi^2}.$$

But by $(8)'$ in the proof of 3.14, $\pi > 3$, so (2) gives

$$|z - n|^2 < \left(\frac{1}{6}\right)^2 + \left(\frac{1}{6}\right)^2 < \left(\frac{1}{4}\right)^2.$$

(%

Livro disponível: Source

20211012/ The Historical Development of the Calculus/

digo

20211012/ Classical Analysis in the Complex Plane/
The Historical Development of the Calculus/
The Historical Development of the Calculus... (1979).pdf

Pág. 31(42) The Measurement of a Circle

$$3\frac{10}{71} < \pi < 3\frac{1}{7}$$

other:

$$\frac{265}{153} < \sqrt{3} < \frac{1351}{780}$$

$$3\frac{10}{71} < \frac{6336}{2017\frac{1}{4}} < \pi < \frac{14688}{4673\frac{1}{2}} < 3\frac{1}{7}$$

return:

20211012/ Classical Analysis in the Complex Plane/
Classical Analysis in the Complex Plane.pdf

Pág. 129(155) (iii) →

(%

Yet a fourth way to utilize 3.22 is to express

$\pi/4 = \arctan(1)$ as a rational linear combination of

$\arctan(\varepsilon_j)$ for small ε_j

(for which the series is more rapidly convergent).

One of these goes back to Euler:

$$\pi/4 = 4\arctan(1/5) - \arctan(1/70) + \arctan(1/99).$$

Pag. 180 (206) –

3.22 inspired Leibniz to quote Virgil (on gambling!):

"Numero Deus impari gaudet".

Essentially the first equality in 3.22 was adopted by Weierstrass (in 1841) as his definition of π.

(This was also the definition adopted in the alternative treatment of the exponential function give in 2.22).

The present definition of π as "twice the least positive zero of the cosine function"

↗

other: (cont.)

Pág. 114 (140) (8)

thus $0 < \pi/2 < \boxed{\sqrt{3}}$ and

estudar!!!

$C(\pi/2) = 0.$

Pág. 135 (161) -

Theorem 3.28

π is irrational, and e^m is irrational for every $m \in \mathbb{N}$.

— 20211012/ Classical Analysis in the Complex Plane/ the Complex Itinerary of Leibniz's Planetary theory...pdf

Pág. 96 (105) - Gravity

letter to Huygens in which Leibniz explicitly claimed:

I distinguish the aether which causes the gravity (pesanteur) (and perhaps also the direction of parallelism of the axes) from that which carries the planets, which is rather coarser.[43]

! -256- - -* ! - 9 1 L & ! 3 ? D ,}
! ! 3 9 3 3 ! A J !1 3 ! 1 ! " + 1 ! & 1 !
) . Y 1"
(/B $ 7 !A 3

c) The items 6)–8) introduce a further hypothesis on the origin of the *conatus explosivus*: it can be originated by an impulse, through which the gravific fluid is pushed out from earth's centre in straight lines towards any direction. Item 8) specifies how such a fluid could act to produce gravity. This can happen by transport of matter or by a wave.
d) Item 9) concerns the specific weight.
e) Item 10) is strictly connected to the hypothesis that gravity acts by straight lines as light. For, if one wishes to reduce the fact that gravity acts with an intensity which decreases as the square of the distance to the fact that light acts in this manner, he has to show that gravity acts with a mechanism similar to light's. This is the reason why Leibniz embraced the hypothesis of the *conatus explosivus*.

Some commentaries:

Aiton stresses that the fluid responsible for gravity is different from that causing the harmonic vortex.[42] He quotes a passage of the already mentioned letter to Huygens in which Leibniz explicitly claimed:

> I distinguish the aether which causes the gravity (*pesanteur*) (and perhaps also the direction of parallelism of the axes) from that which carries the planets, which is rather coarser.[43]

The quotation mentioned by Aiton is perfectly coherent with what Leibniz asserted in the *Zweite Bearbeitung*. For, if the earth is born by means of the condensation of a certain matter, it is quite plausible that the least dense part of the original matter—namely a liquid part—has been pushed outside. It met the original fluid responsible for the harmonic circulation and—being less dense than this fluid—begun to orbit around the earth. The dense harmonic vortex, which surrounded the terrestrial vortex, prevented it to escape along the tangent. Given this picture, it is then necessary to add, that the fluid necessary for Leibniz's theory were, in fact, three and not only two:

a) fluid responsible for the harmonic vortex;
b) fluid responsible for gravity;
c) fluid responsible for the different specific weights of the bodies.

For, in the previous quotation the word *pesanteur* seems clearly to refer to gravity as a general phenomenon, not to the different specific weights of the bodies.

What Leibniz explained in the *Zweite Bearbeitung* is interesting from many aspects and unsatisfactory for other aspects. First of all the argument is complex as far as the line of reasoning is concerned because there is a triple causal level: the

---

[42] Aiton (1972, p. 131).

[43] Leibniz 1690, in Huygens, 1901, *Oeuvres Complétes*, IX p. 526. See also Leibniz (1690a, 1860, 1962, VI, p. 192). Aiton quotes this passage, translated into English, in Aiton (1972, p. 134). I have quoted Aiton's translation. Original French text: "Cependant je distingue l'aether qui fait la pesanteur (et puetestre aussi la direction ou le parallelisme des axes) de celuy qui defere les planetes, qui est bien plus grossier".

*D : 3 ? D ,

(%%

Leibniz 1690, in Huygens, 1901, Oeuvres Completes, IX, p. 526.

- 20211012/ Cosmology in Gauge Field theory ~~of~~ and String Theory/

Cosmology in Gauge Field...pdf

Pág. 292 (305)

- 20211012/ Modular Forms...|

Modular_Forms.pdf

Search: Riemann

Pág. 16(22) theorem 4.1.

$$G_K(z)=2\zeta(K)+\frac{2(2\overset{\downarrow}{\pi} i)^K}{(K-1)!}\sum_{n=1}^{\infty}\sigma_{K-1}(n)\,q^n,$$

Pág. 19(25)

Riemann surfaces (like $SL_2(\mathbb{Z}\backslash \mathbb{H} \cup \{\infty\})$, we have to admit poles of finite order in ∞.

Pág. 31 (37) 5.2

L-functions of Modular Forms

→ Riemann zeta function.

! -256- - -* ! - 3 7! < "9 3
! " 9 3 1"
(/ % $

Nevertheless, using a similar treatment, we shall see later that string theory *can* account for the entropy but the associated microstates are *non*-perturbative. Even then, the accounting has so far only been successful for certain special types of black hole called 'extreme' black holes. We therefore first describe the features of extreme black holes that are important for understanding the microscopic origin of their entropy.

10.5 Extreme black holes

It might, in any case, be objected that associating perturbative string states with black holes is absurd. The former are obtained by quantizing the string in a flat background spacetime. How could they be equivalent to a black hole? Nevertheless, certain perturbative states *can* be associated with 'extreme' black holes. We have already noted the constraint $M \geq |Q|$ for the Reissner–Nordström (RN) metrics (10.33). Metrics saturating this constraint are called 'extreme' (RN) black holes. The usual (supersymmetric) perturbative states in string theory satisfy $M \geq |Q|$ and states saturating this inequality are called 'BPS states' after Bogomolnyi [14] and Prasad and Sommerfield [15]. They have the crucial property that their mass cannot receive quantum corrections and it is this that allows their association with extreme black holes. As the string coupling strength g_s increases, the mass M of the perturbative state is unaltered, since it is independent of g_s classically, and because of supersymmetry, there are no quantum corrections. However, the gravitational field of the state is determined by $G_N M$ (in four dimensions) and G_N is proportional to g_s^2. Thus, as g_s increases, the gravitational field increases, there is a back reaction on the perturbative state and, eventually, it may be described by a curved spacetime with large curvature. Thus, in principle at least, it might be possible to associate extreme black-hole spacetimes with perturbative states.

The RN black hole looks like a natural starting place in the search for black holes in string theory. We can think of the Einstein–Maxwell action (10.31) as the bosonic part of the $\mathcal{N} = 2$ supergravity theory in four dimensions. The (massless) gravity supermultiplet contains the graviton, two (fermionic) gravitinos and a vector boson called the 'graviphoton'. The supersymmetry algebra is [13]

$$\begin{aligned} \{Q^A_\alpha, \overline{Q}_{B\dot{\beta}}\} &= 2\delta^A_B \sigma^\mu_{\alpha\dot{\beta}} P_\mu \\ \{Q^A_\alpha, Q^B_\beta\} &= 2\epsilon_{\alpha\beta} Z^{AB} \end{aligned} \tag{10.91}$$

where Q^A_α, with $\alpha = 1, 2$ a (Weyl) spinor index and $A = 1, 2$, are the two supersymmetry generators; the $\overline{Q}$s are defined by $\overline{Q}_{A\dot{\alpha}} \equiv (Q^A_\alpha)^\dagger$; the 2×2 matrices σ^μ with $\mu = 0, 1, 2, 3$ are defined by $\sigma^\mu = (I_2, \sigma^i)$, with σ^i ($i = 1, 2, 3$) the standard Pauli matrices; $\epsilon_{\alpha\beta}$ and the 'central charge' Z^{AB} are antisymmetric with $\epsilon_{12} = +1$ and $Z^{12} \equiv Z$ and, without loss of generality, we may choose $Z \geq 0$. (The graviphoton is a $U(1)$ gauge boson coupled to a charge which is, in fact, the central charge.) For massive representations, we may work in the rest frame where $P_\mu = (M, \mathbf{0})$. Then it is easy to see (exercise 9) that we may form two linear combinations a_α and b_α of the generators that satisfy

$$\begin{aligned}\{a_\alpha, a_\beta^\dagger\} &= 4(M + Z)\delta_{\alpha\beta}\\ \{b_\alpha, b_\beta^\dagger\} &= 4(M - Z)\delta_{\alpha\beta}\\ \{a_\alpha, b_\beta^\dagger\} &= 0 = \{a_\alpha, b_\beta\}\end{aligned} \tag{10.92}$$

Up to a normalization factor, these are just the anticommutation relations obeyed by two independent sets of fermion annihilation and creation operators. In general, starting with a state $|\psi\rangle$ that is annihilated by a_α and b_α, we can construct a total of 16 states using the creation operators $a_\alpha^\dagger$ and $b_\alpha^\dagger$. Since $\langle\phi|\{a_\alpha, a_\alpha^\dagger\}|\phi\rangle \geq 0$ for any state $|\phi\rangle$ and, similarly, for b_α, the 'BPS bound'

$$M \geq |Z| \tag{10.93}$$

follows. The inequality is saturated by representations (BPS states) for which $|\psi\rangle$ is also annihilated by one set of creation operators ($b_\alpha^\dagger$ if $Z > 0$). Thus, such states are 'short' massive representations of the supersymmetry algebra, since they are constructed using only the $a_\beta^\dagger$ creation operators. They are invariant under half of the supersymmetry algebra.

The extreme RN black hole, obtained from (10.33) and (10.34) by setting $M = Q$, *is* part of such a short hypermultiplet [16]. In this case, the two horizons are both at $r = M = Q$ and, defining $\rho \equiv r - Q$, we may write the solution in the 'isotropic" form in which the spatial part of the metric is conformal to flat space:

$$ds^2 = H^{-2}\,dt^2 - H^2(d\rho^2 + \rho^2\,d\Omega_2^2) \tag{10.94}$$

$$A = (1 - H^{-1})\,dt \tag{10.95}$$

where

$$H \equiv 1 + \frac{Q}{\rho}. \tag{10.96}$$

Both the temporal and spatial 'warp' factors (the factors multiplying the two parts of the metric), as well as the electromagnetic vector-potential 1-form, are determined by a single (harmonic) function H. Extreme solutions have the important property that they are easily generalized to a case representing N extreme black holes with charges $q_i = m_i$ (with $i = 1, 2, \ldots, N$) by replacing the function H given in (10.96) by

$$H = 1 + \sum_{i=1}^{N} \frac{q_i}{|\mathbf{r} - \mathbf{r}_i|}. \tag{10.97}$$

By Gauss' law, the total charge is $Q = \sum_{i=1}^{N} q_i$ which, by the BPS bound, is also the total mass

$$M = \sum_{i=1}^{N} m_i = \sum_{i=1}^{N} q_i = Q. \tag{10.98}$$

There is, thus, no binding energy between the individual black holes: the gravitational binding is precisely cancelled by the electrostatic repulsion. Near the horizon $|\rho| \ll Q$, the metric is approximated by $\mathrm{Ad}S_2 \times S^2$ (exercise 11). Thus, the extreme black hole may also be regarded as a soliton that interpolates between the Minkowski vacuum when $\rho \gg Q$ and $AdS_2 \times S^2$ when $|\rho| \ll Q$. Generalizations of this special form, involving extra dimensions, extended objects and other charges, are important in what follows.

In particular, the simplest example of a string-theory black hole for which a microscopic description can be found is provided by a five-dimensional analogue of the RN solution (10.33) and (10.34). The static, solution of the five-dimensional Einstein–Maxwell equations outside of a (S^3)-spherically symmetric body of mass M and charge Q is

$$\mathrm{d}s^2 = \left(1 - \frac{2M}{r^2} + \frac{Q^2}{r^4}\right) \mathrm{d}t^2 - \left(1 - \frac{2M}{r^2} + \frac{Q^2}{r^4}\right)^{-1} \mathrm{d}r^2 - r^2\,\mathrm{d}\Omega_3^2 \quad (10.99)$$

and the gauge potential 1-form

$$A = \frac{Q}{r^2}\,\mathrm{d}t. \quad (10.100)$$

(The r^{-2} dependence of the potential arises, of course, because there are now four spatial dimensions.) As before, the extreme case where $M = Q$, with the horizon at $r = \sqrt{Q}$, may be cast in an isotropic form by using the coordinate $\rho \equiv \sqrt{r^2 - Q}$. Then the horizon is at $\rho = 0$ and

$$\mathrm{d}s^2 = H^{-2}\,\mathrm{d}t^2 - H(\mathrm{d}\rho^2 + \rho^2\,\mathrm{d}\Omega_3^2) \quad (10.101)$$

$$A = (1 - H^{-1})\,\mathrm{d}t \quad (10.102)$$

where now the harmonic function is

$$H = 1 + \frac{Q}{\rho^2}. \quad (10.103)$$

To connect this black hole with string theory, we need to consider the effective field theory describing string theory in the low-energy limit. This is a generalization of Einstein–Maxwell theory both with respect to the number of spacetime dimensions (ten) and the fields involved. The fields involved are just the massless modes that arise in (perturbative) string theory and the field theory that desribes them is type II supergravity.

10.6 Type II supergravity

The massless states that arise in superstring theory include both bosons and fermions. Black holes are, of course, solutions of the classical bosonic field

D -256- - -* ! - " ! I< 1"
! 5 !
(B 9 '

$$\begin{aligned} G_k(Sz) &= G_k\left(-\frac{1}{z}\right) = \sum_{\substack{c,\, d \in \mathbb{Z} \\ (c,d) \neq (0,0)}} \frac{1}{\left(c\left(-\frac{1}{z}\right) + d\right)^k} \\ &= z^k \sum_{\substack{c,\, d \in \mathbb{Z} \\ (c,d) \neq (0,0)}} \frac{1}{(dz - c)^k} = (1 \cdot z + 0)^k G_k(z). \end{aligned}$$

We have thus shown the necessary transformation property for the matrices S and T and hence all of $\mathrm{SL}_2(\mathbb{Z})$.

The holomorphicity in ∞ follows from the following statement about the Fourier expansion of the Eisenstein series.

Theorem 4.1 *Let $k > 2$ be even. It holds*

$$G_k(z) = 2\zeta(k) + \frac{2(2\pi i)^k}{(k-1)!} \sum_{n=1}^{\infty} \sigma_{k-1}(n) q^n,$$

where $q = e^{2\pi i n z}$, $\sigma_\ell(n) := \sum_{0<d|n} d^\ell$ denotes the divisor sum function, and $\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$ denotes the Riemann zeta function.

Remark 4.1 In the following, we often consider the *normalized Eisenstein series* $E_k(z) = \frac{1}{2\zeta(k)} G_k(z)$ with Fourier expansion

$$E_k(z) = 1 - \frac{2k}{B_k} \sum_{n=1}^{\infty} \sigma_{k-1}(n) q^n.$$

Here, B_k denotes the kth Bernoulli number defined by $\frac{t}{e^t - 1} = \sum_{n=0}^{\infty} \frac{B_n}{n!} t^n$ and Euler's formula applies

$$2\zeta(k) = -\frac{(2\pi i)^k}{k!} \cdot B_k, \; k \geq 2 \text{ even}.$$

Example 4.1 We have

$$E_4(z) = 1 + 240\sum_{n=1}^{\infty}\sigma_3(n)q^n = 1 + 240q + 2160q^2 + 6720q^3 + \dots .$$

A simple conclusion from Corollary 3.1 is now that

$$M_k = \mathbb{C}E_k$$

for $k = 4, 6, 8, 10$.

Example 4.2 The following applies

$$E_4(z)^2 = E_8(z),$$

because the dimension of M_8 is 1 and both $E_4(z)^2$ and $E_8(z)$ have leading coefficient 1, so they agree.

From this follows directly the following relationship between divisor sum functions

$$\sum_{m=1}^{n-1}\sigma_3(m)\sigma_3(n-m) = \frac{\sigma_7(n) - \sigma_3(n)}{120}.$$

4.2 The Discriminant

We now come to the definition of the *discriminant*, a cusp form of weight 12.

Definition 4.2 The *discriminant* or Δ-*function* is defined by

$$\Delta(z) := \frac{E_4^3(z) - E_6^2(z)}{1728} = q - 24q^2 + 252q^3 + \dots .$$

Based on the definition and because of $\Delta(z) = q + \dots$, one immediately sees that $\Delta \in S_{12}$ holds.

Remark 4.2 One also denotes the Fourier coefficients of the discriminant by $\tau(n)$.

5.2 *L*-functions of Modular Forms

We first recall important results about the Riemann zeta function

$$\zeta(s) := \sum_{n=1}^{\infty} n^{-s},\ s \in \mathbb{C}, \Re(s) > 1.$$

This series converges normally (locally uniform and absolute) and the function defined by it is holomorphic. The function $\zeta(s) - \frac{1}{s-1}$ has a holomorphic continuation to $\mathbb{C}$.

It is easy to show that

$$\zeta(s) = \prod_{p \text{ prime}} \frac{1}{1 - p^{-s}},\ \Re(s) > 1,$$

and via this product expansion one obtains another proof for the fact that there are infinitely many primes. For if there were only finitely many primes, the above product would be finite and thus $\zeta(1)$ would be rational and in particular finite. But we have just noted that $\zeta(s)$ has a pole at $s = 1$.

Riemann realized that the connection between the arithmetic of primes and the analytic properties of the Riemann zeta function is much deeper, and that $\zeta(s)$ should provide the key to studying the distribution of primes. One compares with the corresponding versions of the Riemann conjecture.

Theorem 5.4 (Riemann) *The Riemann* ζ *-function has a meromorphic continuation to* $\mathbb{C}$. *The completed Riemann zeta function* $\zeta^*(s) := \pi^{-\frac{s}{2}}\Gamma\left(\frac{s}{2}\right)\zeta(s)$ *satisfies the functional equation*

$$\zeta^*(s) = \zeta^*(1-s).$$

Here $\Gamma(s) = \int_0^{\infty} e^{-t} t^{s-1} dt$ *is for* $\Re(s) > 0$ *the gamma function.*

Analogous to the Riemann zeta function, one now considers *L*-functions of cusp forms.

Definition 5.5 Let $f(z) = \sum_{n=1}^{\infty} a(n)q^n$, $q = e^{2\pi i z}$, be a cusp form of weight $k > 0$ for $\mathrm{SL}_2(\mathbb{Z})$. We call

$$L(f, s) := \sum_{n=1}^{\infty} a(n)n^{-s}, s \in \mathbb{C},$$

the *L-function* (or *L-series*) of f.

One can show that $L(f, s)$ converges normal (locally uniform and absolute) for $s \in \mathbb{C}$ with $\Re(s) > \frac{k}{2} + 1$. This follows directly from the estimate for the Fourier coefficients that we derived in Theorem 3.4.

Theorem 5.5 *Let $f \in S_k$. Then the associated L-function has a holomorphic continuation to all $\mathbb{C}$. The function*

$$\Lambda(f, s) := (2\pi)^{-s}\Gamma(s)L(f, s)$$

is entire and satisfies the functional equation

$$\Lambda(f, s) = (-1)^{\frac{k}{2}}\Lambda(f, k - s).$$

The function $\Lambda(f, s)$ is bounded on every vertical strip, that is, for every $T > 0$ there exists a $C_T > 0$ such that $|\Lambda(f, s)| \leq C_T$ for all $s \in \mathbb{C}$ with $\Re(s) \leq T$.

Similar to the Riemann zeta function, the L-functions of cusp forms also have a representation as an *Euler product.*

Theorem 5.6 *Let $f \in S_k$, $k > 14$, $a_f(1) = 1$. Then the following are equivalent:*

1. *f is a simultaneous eigenform with respect to all Hecke operators.*
2. *$L(f, s)$ has a product expansion of the form*

$$L(f, s) = \prod_{p \text{ prime}} \frac{1}{1 - a_f(p)p^{-s} + p^{k-1-2s}}$$

for $\Re(s) > 1 + \frac{k}{2}$.

The *modularity theorem* proved by Wiles et al. [Wil95, BCDT01] in the course of proving Fermat's Last Theorem establishes a far-reaching connection between the L-functions of newforms and those of elliptic curves.

Let E be an elliptic curve over $\mathbb{Q}$ given by the solutions $(x, y) \in \mathbb{Q}^2$ of the equation $y^2 = x^3 + ax + b$, with $a, b \in \mathbb{Q}$ such that $-4a^3 - 27b^2 \neq 0$, together with a point O at infinity.

Given E (of conductor N_E), we now define an L-function as follows. We set

$$L(E, s) = \prod_{p \text{ prime}} L_p(E, s),$$

where

$$L_p(E, s) = \begin{cases} \left(1 - a(p)p^{-s} + p^{1-2s}\right)^{-1}, & \text{if } p \nmid N_E, \\ \left(1 - a(p)p^{-s}\right)^{-1}, & \text{if } p \mid N_E,\ p^2 \nmid N_E, \\ 1, & \text{if } p^2 \mid N_E. \end{cases}$$

Here one defines $a(p) = p + 1 -$ (#points on E modulo p).

In the 1950s, Taniyama conjectured that such an L-function coincides with that of a weight 2 newform. In the 1960s, Eichler and Shimura showed that for any weight 2 newform with integer Fourier coefficients, one can find an elliptic curve E over $\mathbb{Q}$ such that $L(E, s) = L(G, s)$. The other direction was then proved in the 1990s by Wiles et al. [Wil95, BCDT01] and has far-reaching consequences for various areas of mathematics.

Theorem 5.7 (Modularity Theorem) *Let E be an elliptic curve over $\mathbb{Q}$ of conductor N_E. There exists a newform $G_E(z) = \sum_{n=1}^{\infty} a_E(n)q^n \in S_2(\Gamma_0(N_E))$ of weight 2 such that*

$$L(G_E, s) = L(E, s), s \in \mathbb{C}.$$

(%%

20211012/ Cosmology...|

Supersymmetric gauge field theory (1994).djvu

Pag. 316(324) Appendix D.

Closed-String Integrals

This integral separates into standard (Gaussian) integrals when Cartesian coordinates are used for $z = x + iy$. Then (D7)

$$J(\lambda, \mu) = \frac{1}{s+t} \exp\left(\frac{\lambda t + \mu s - st}{s+t} \right)$$

20211012/ Image and Graphics...|

3030873544.pdf

Pag. 669(694) Spatial Structure Feature

The gradient amplitude map of an I-frame can be defined as:

$$GM = \sqrt{[I \otimes h_x]^2 + [I \otimes h_y]^2}, \quad (4)$$

where I is a luminance channel of the I-frame, and $\otimes$ represents the convolution operation.

→

! -256- - -* ! - 1 3 ! " 3
! " 3 & 3 D //' "KA
-256- - -* ! - 1 3 ! " 3 ! "
3 & 3 D //' 1"
(% B % ' 211 " L 4 " & !

APPENDIX D

CLOSED-STRING INTEGRALS

The integral (12.109), which is required in order to evaluate the four-tachyon amplitude in the closed-bosonic-string theory, is a special case of the more general integral

$$I(\alpha, n; \beta, m) = \int \frac{\mathrm{d}^2 z}{\pi} |z|^{\alpha} |1 - z|^{\beta} z^{n} (1 - z)^{m} \tag{D1}$$

which is convergent for

$$\begin{aligned} &\mathrm{Re}(\alpha + \beta + m + n + 2) < 0 \\ &\mathrm{Re}(\alpha + n + 2) > 0 \\ &\mathrm{Re}(\beta + m + 2) > 0. \end{aligned} \tag{D2}$$

We substitute an integral representation of $|z|^{\alpha}$

$$|z|^{\alpha} = \frac{1}{\Gamma(-\alpha/2)} \int_0^{\infty} \mathrm{d}s\, s^{-\alpha/2 - 1}\, \mathrm{e}^{-s|z|^2} \tag{D3}$$

which follows from the standard definition of the Γ function

$$\Gamma(p) = \int_0^{\infty} \mathrm{d}u\, u^{p-1}\, \mathrm{e}^{-u} \tag{D4}$$

by changing the integration variable to $u = s|z|^2$. Similarly for $|1 - z|^{\beta}$. Then

$$I = \frac{1}{\Gamma(-\alpha/2)\Gamma(-\beta/2)} \int_0^{\infty} \mathrm{d}s\, s^{-\alpha/2 - 1} \int_0^{\infty} \mathrm{d}t\, t^{-\beta/2 - 1} \times \int \frac{\mathrm{d}^2 z}{\pi} z^{n} (1 - z)^{m}\, \mathrm{e}^{-s|z|^2 - t|1 - z|^2}. \tag{D5}$$

Next we do the z-integral. This is easily done using the generating integral

$$J(\lambda, \mu) \equiv \int \frac{\mathrm{d}^2 z}{\pi} \exp[-s|z|^2 - t|1 - z|^2 + \lambda z + \mu(1 - z)]. \tag{D6}$$

This integral separates into standard (Gaussian) integrals when Cartesian coordinates are used for $z = x + \mathrm{i}\, y$. Then

$$J(\lambda, \mu) = \frac{1}{s + t} \exp\left(\frac{\lambda t + \mu s - st}{s + t}\right) \tag{D7}$$

so the z-integral in (D5) is obtained by differentiating with respect to λ, μ:

$$\int \frac{\mathrm{d}^2 z}{\pi} z^n (1-z)^m \,\mathrm{e}^{-s|z|^2 - t|1-z|^2} = \frac{t^n s^m}{(s+t)^{n+m+1}} \mathrm{e}^{-st/(s+t)}. \tag{D8}$$

Of course this result follows only when n, m are integral, but we can continue to the case when they are non-integral. Next we change integration variables to u and x where

$$u = \frac{st}{s+t} \tag{D9a}$$

$$x = \frac{t}{s+t}. \tag{D9b}$$

The domain of integration is

$$0 \leqslant u < \infty \qquad 0 \leqslant x \leqslant 1. \tag{D10}$$

Then the required integral (D5) is$^{(1)}$

$$\begin{aligned} I &= \frac{1}{\Gamma(-\alpha/2)\Gamma(-\beta/2)} \int_0^1 \mathrm{d}x\, x^{n+\alpha/2} (1-x)^{m+\beta/2} \int_0^\infty \mathrm{d}u\, u^{-\alpha/2-\beta/2-2}\, \mathrm{e}^{-u} \\ &= \frac{\Gamma(1+n+\alpha/2)\Gamma(1+m+\beta/2)\Gamma(-\alpha/2-\beta/2-1)}{\Gamma(-\alpha/2)\Gamma(-\beta/2)\Gamma(2+m+n+\alpha/2+\beta/2)}. \end{aligned} \tag{D11}$$

Reference

1 Gross D J, Harvey J A, Martinec E and Rohm R 1986 *Nucl. Phys.* B **267** 75

D -256- - -* ! - % % %$" 1"
(BB/ B/' 1! ! < !

and the coding features generated during the video coding process. The coding features mainly use the target bit-rate and the SATD value.

Target Bit-Rate Feature: It refers to the target bit-rate allocated to the current I-frame, denoted as R_t.

Sum of Absolute Transformed Difference Feature: SATD refers to the sum of the absolute values of transform coefficients, which is generated by the *Hadamard* transform. It is suggested in [3] as the content complexity evaluation value for the I-frame, which is adopted in HEVC and also utilized in VTM [13].

Here, the content complexity of an I-frame is defined as the sum of the SATD values of all 8 × 8 blocks in the frame. We use SATD as one of the input features, and its calculation is:

$$SATD = \sum_{i=1}^{8} \sum_{j=1}^{8} |c_{i,j}|, \tag{3}$$

where $c_{i,j}$ denotes the *Hadamard* transform coefficient.

Spatial Structure Feature: Gradient amplitude response can effectively detect most of the spatial structure information of the video frame. Some previous studies have proposed using gradient information to reflect the frame content complexity [7,8]. Therefore, the gradient feature is also considered as one of the input features. The gradient amplitude map of an I-frame can be defined as:

$$GM = \sqrt{[I \otimes h_x]^2 + [I \otimes h_y]^2}, \tag{4}$$

where I is a luminance channel of the I-frame, and $\otimes$ represents the convolution operation.

$h_d, d \in \{x, y\}$, represents the *Gaussian* partial derivative filter along the horizontal or vertical direction respectively. It is formulated as:

$$\begin{aligned} h_d(x, y|\sigma) &= \frac{\partial}{\partial d} g(x, y|\sigma) \\ &= -\frac{1}{2\pi\sigma^2} \frac{d}{\sigma^2} exp^{\{-\frac{x^2+y^2}{2\sigma^2}\}} \end{aligned} \tag{5}$$

where $g(x, y|\sigma)$ denotes the isotropic *Gaussian* function with a scale parameter of σ.

Then, the two-dimensional feature matrix GM is converted into the form of a histogram, and the feature data are divided into 10 groups to better retain the information of overall structure and local shape. The obtained feature vector is named HG.

Frequency-Domain Energy Feature: Luminance change is an important feature in expressing the frame content, and this change can be reflected more significantly in the frequency domain. We hence design a feature extraction method in the DCT transform [10]. First, the input I-frame is divided and transformed in the DCT domain with N × N blocks, and the obtained DC coefficients reflect the

h_d, $d \in \{x, y\}$, represents the Gaussian partial derivative filter along the horizontal or vertical direction respectively.

- 2021 10 12 / Mastering Uncertainty in Mechanical Engineering / 3030783537.pdf

Pág. 37(50) → Gauss

A scaled Gaussian Process is used ~~in~~ for ~~modell~~ (model) calibration and prediction.

Pág. 49(62) –

Gaussian uncertainty propagation is not adequate.

Pág. 50(63) – Fig. 3.4. ↙

(a) Gaussian uncertainty propagation shown to be associated to a coordinate transformation or mapping $\theta_1 \rightarrow \theta = h(\theta_1)$;

(b) approximation $\bar{\theta}_g \approx \bar{\theta}$ due to the nonlinearity in $h(\theta_1)$.

(estudar os gráficos).

%
! -256- - -* ! - % % % %$% 1"
(% $ 7!

In accordance to this book's notation, deterministic approaches do not master stochastic uncertainty in a mathematical model. However, they may be suited to quantify and master incertitude, e.g. when only extreme values like minima and maxima are known or assumed, or when parts of a model are missing to detect ignorance, cf. Fig. 2.2. Probabilistic approaches, however, describe and master all three types of uncertainty. In the following, we give an overview of the two approaches.

Deterministic framework

On the one hand, a *deterministic* analysis can be used, for example, in fault diagnosis [56] to quantify uncertainty in the model equations. To do so, the model is adjusted by a model discrepancy function $\delta\eta$ which is assumed to stay within a bounded uncertainty set $\mathcal{U}_\eta$:

$$\eta(u, m) + \delta\eta(u) = z, \quad \delta\eta \in \mathcal{U}_\eta. \tag{2.6}$$

A standard residual analysis then enables the engineer to distinguish component failures from the effects of model uncertainty. However, this method strongly relies on the assumption that the model discrepancy function $\delta\eta$ stays within a bounded uncertainty set $\mathcal{U}_\eta$.

On the other hand, a deterministic analysis may be used in model verification and validation processes. In [9, 57], the possibility to approximate the residual between the model output and the observed quantity of interest via a polynomial p is mentioned. In this case, the polynomial p takes the role of the discrepancy function:

$$\eta(u, m) + p(u, m, \theta) = z. \tag{2.7}$$

This necessarily leads to an augmented parameter set (m, θ) consisting of the original physical axiomatic and empiric parameters m as well as the non-physical polynomial parameters θ, which do not give an enhanced physics-based understanding of the model's uncertainty or shortcoming in predicting reality [9]. The augmented parameter set needs to be calibrated, which is usually performed by an optimisation scheme.

Probabilistic framework

Within the probabilistic framework, *Bayesian inference-based* approaches are frequently used to assess the prediction quality of a mathematical model under given experimental data [9, 16, 52, 61]. In [34] or [37], Bayesian calibration techniques and a plausibility prior argument are used for model selection. Another important approach introduces a stochastic process δ for the discrepancy function [22]:

$$\eta(u, m) + \delta(u, m, \phi) + \varepsilon = z\,, \tag{2.8}$$

where ε is stochastic noise, usually produced by the measurement error, and ϕ are hyperparameters, which need to be tuned by an optimisation scheme. Based on this setting, [64] introduce a technique for model validation. A method designed for time-dependent systems is proposed in [3]. Combinations of a high- and a low-fidelity model and experimental data are used to construct a multi-fidelity model, which is proposed in [13] as an improved predictor. A scaled Gaussian Process is used in [17] for model calibration and prediction. It is claimed that the method bridges the gap between the least-squares calibration and the Gaussian Process calibration and, thus, is an improvement of the method introduced in [22]. Furthermore, a Bayesian interval hypotheses-based approach [36] and a Bayesian inference-based approach [37] are used to compare different models based on their internal functional relations from axiomatic or empiric assumptions, see also Sect. 4.3.3 for assessing model uncertainty in the Modular Active Spring-Damper System, see Sect. 3.6.1.

In order to apply the Bayesian methods, it is necessary to select prior distributions. Often, this is subjective and it is unclear how to choose them, which may lead to unrealistic assumptions. Furthermore, it is pointed out in [60] that the approach in [22] may lead to inadequate results in case of an imperfect computer model. As a consequence, further verification of the computer codes needs to be conducted.

From the *frequentist's* perspective, a simple approach to quantify model uncertainty is to use validation metrics, such as the area validation metric [35] or the Mahalanobis distance [68]. They give a quantitative measure of disagreement between the model output and the observed quantity of interest based upon observed measurements. In order to select the most adequate model, an arbitrary threshold on the metric is imposed, or classical hypothesis testing is performed. Surrogate models are often used to quantify the uncertainty in a technical system. Usually, these methods use computer simulations and a small sample of experimental data to estimate properties of probability distributions, such as quantiles [25–27] or densities [14, 23, 24]. A detailed description for a method based on an imperfect computer model is given in Sect. 4.3.8. A case where computer models are assumed to fit the reality is shown in Sect. 5.2.6. In [12], another method to detect model uncertainty is proposed which is based on optimum experimental design and hypothesis testing, see also Sect. 4.3.1. In [65], another approach to quantify the model error is developed. Here, the model discrepancy function is estimated on bootstrap samples via a regression estimation, e.g. smoothing splines or artificial neural networks.

However, the application of frequentist methods often demands a larger sample size, which can become a problem, since generating data is an expensive and time-consuming process.

(molecules, ...) in an arbitrarily short time at a time $t = 0$ into a constant flow Q. If the process has the volume V, then we measure the temporal change of the concentration given by $N/(V\bar{t})\ p(t)$ at the process output. Here, $\bar{t}$ is the mean residence time given by $\bar{t} = V/Q$. As always, things become clearer in a dimensionless representation: if we measure time in multiples of the mean residence time, i.e. $\theta := \frac{t}{\bar{t}}$, and concentration c in multiples of the typical concentration N/V, we measure $\delta(\theta) \rightarrow p(\theta)$, i.e. a Gaussian normal distribution for a purely diffusive process centred at $\theta = 1$. In a real experiment it is not possible to realise a Dirac delta function $\delta(\theta)$ but it is easy to realise a Heaviside function $\mathcal{H}(\theta)$. This is achieved by changing the concentration at the inlet at a time $\theta = 0$ suddenly from zero to one. In practice this means switching a valve. At the process outlet we then measure the cumulative density function $\mathcal{H}(\theta) \rightarrow P(\theta) = \int_0^\theta p(\tau)\mathrm{d}\tau$.

Secondly (ii), in the general case, the dependent variable θ is a nonlinear function of one, two or more independent variables $\theta = h(\theta_1, \theta_2)$. A perturbation of h at the operation point $(\theta_{10}, \theta_{20})$ marked by the index 0 is given by

$$h = h_0 + \left.\frac{\partial h}{\partial \theta_1}\right|_0 \delta\theta_1 + \left.\frac{\partial h}{\partial \theta_2}\right|_0 \delta\theta_2 + \mathrm{O}(\delta\theta^2).$$

Inserting this perturbation into the general convolution integral and neglecting terms of the order of magnitude $\mathcal{O}(\delta\theta^2)$ yields the Gaussian uncertainty propagation [29]

$$\sigma_{\mathrm{g}}^2 \approx \left[\frac{\partial h}{\partial \theta_1}\right]_0^2 \sigma_1^2 + \left[\frac{\partial h}{\partial \theta_1}\frac{\partial h}{\partial \theta_2}\right]_0 \sigma_{12}^2 + \left[\frac{\partial h}{\partial \theta_2}\right]_0^2 \sigma_2^2.$$

The index g indicates the Gaussian approximation of the variance. Here,

$$\sigma_{12}^2 := \int_0^\infty \int_0^\infty (\theta_1 - \bar{\theta}_1)(\theta_2 - \bar{\theta}_2) h(\theta_1, \theta_2)\, \mathrm{d}\theta_1 \mathrm{d}\theta_2 = \overline{\theta_1\theta_2} - \bar{\theta}_1\bar{\theta}_2$$

denotes the co-variance of the two independent variables.

The Gaussian uncertainty propagation is the simplest and most commonly used method of uncertainty propagation, but it is rarely critically questioned. From the derivation it is clear that the trust into the method should decrease with increasing nonlinearity in $h(\theta_1, \theta_2)$. This becomes clear for the simpler case $\theta = h(\theta_1)$. This coordinate transformation shown in Fig. 3.4b leads to a distortion of a symmetric input density function $p_1(\theta_1) \rightarrow p(\theta)$. As a result, the mean $\bar{\theta} = \int_0^\infty h(\theta_1) p_1(\theta_1)\, \mathrm{d}\theta_1$ differs from the approximation $\bar{\theta}_{\mathrm{g}} = h(\theta_{10})$ and further σ_{g}^2 is only an approximation of σ^2, Fig. 3.4b. In summary we recognise that the usage of the Gaussian approximation $(\delta\theta)^2 \approx (\delta\theta)_{\mathrm{g}}^2 = \sum (\frac{\partial h}{\partial}\theta_k)^2 (\delta\theta_k)^2$ for uncorrelated input data θ_k needs some justification.

For the strongly nonlinear process chain being treated in the following Sect. 3.3, the moment method, i.e. the Gaussian uncertainty propagation is not adequate. In Sect. 3.3 we use the bootstrapping method [10]. Other methods and examples for uncertainty propagation are found in Sect. 6.1.6.

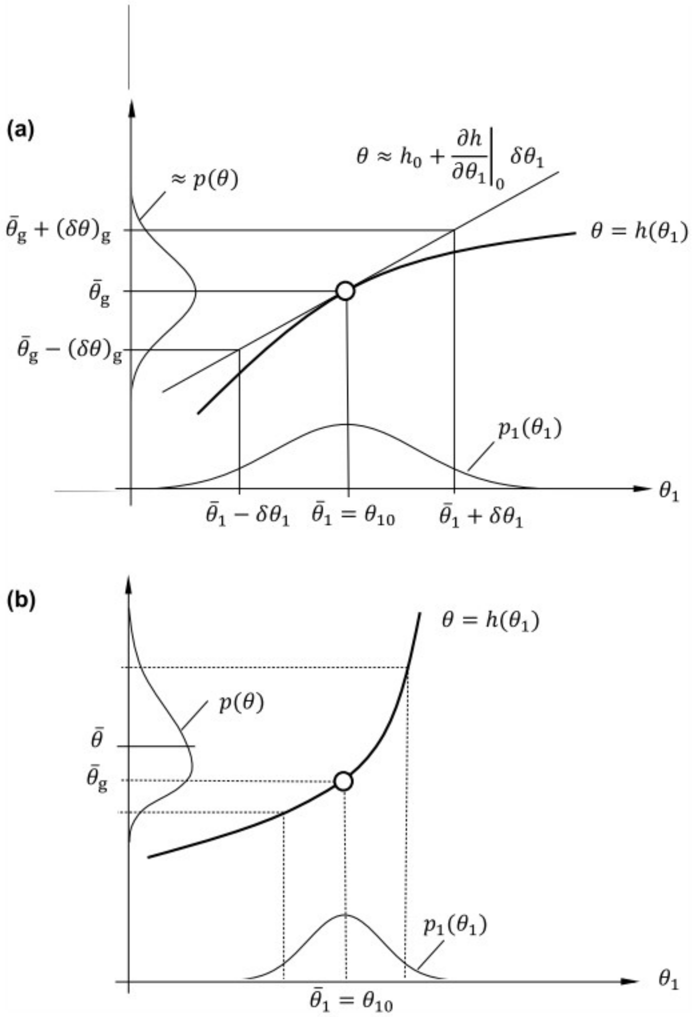
(a)
$\theta \approx h_0 + \left.\frac{\partial h}{\partial \theta_1}\right|_0 \delta\theta_1$
$\approx p(\theta)$
$\bar{\theta}_g + (\delta\theta)_g$
$\theta = h(\theta_1)$
$\bar{\theta}_g$
$\bar{\theta}_g - (\delta\theta)_g$
$p_1(\theta_1)$
θ_1
$\bar{\theta}_1 - \delta\theta_1$
$\bar{\theta}_1 = \theta_{10}$
$\bar{\theta}_1 + \delta\theta_1$
(b)
$\theta = h(\theta_1)$
$p(\theta)$
$\bar{\theta}$
$\bar{\theta}_g$
$p_1(\theta_1)$
θ_1
$\bar{\theta}_1 = \theta_{10}$

Fig. 3.4 **a** Gaussian uncertainty propagation shown to be associated to a coordinate transformation or mapping $\theta_1 \rightarrow \theta = h(\theta_1)$; **b** approximation $\overline{\theta}_g \approx \overline{\theta}$ due to the nonlinearity in $h(\theta_1)$

3.3 Five Complementary Methods for Mastering Uncertainty in Process Chains

Peter F. Pelz

In Sect. 3.2 we applied moment methods of uncertainty propagation to process chains. In several parts of the book we resume these and use more in-depth methods of uncertainty propagation, cf. Sect. 6.1.6. Uncertainty quantification includes uncertainty propagation [30]. Uncertainty propagation is often recognised as uncertainty

$p^\star$ and covariance matrix C. Then the set K is an n_p-dimensional ellipsoid determined by the covariance matrix C:

$$K(p^\star, C, \alpha) := \left\{ p \in \mathbb{R}^{n_\mathrm{p}} : \left(p - p^\star\right)^\top C^{-1}\left(p - p^\star\right) \leq \chi^2_{n_\mathrm{p}}(1-\alpha) \right\},$$

where $\chi^2_{n_\mathrm{p}}$ is the quantile function of the χ^2 probability distribution with n_p degrees of freedom. We consider the following approximations for the covariance matrix C coming from a Gauss–Newton approach [19, 38] and a sensitivity analysis [10, 38], respectively:

$$C_\mathrm{GN} = \left(J^\top \Omega \Sigma^{-1} J\right)^{-1}, \qquad C_\mathrm{S} = H^{-1} J^\top \Omega \Sigma^{-1} J H^{-\top},$$

where J is the total derivative of the residual vector r with respect to the model parameters p and

$$H := J^\top \Omega \Sigma^{-1} J + \sum_{i=1}^{n_z} r_i \Omega_{ii} \Sigma_{ii}^{-1} \frac{\mathrm{d}^2 r_i}{\mathrm{d}p^2}. \tag{4.17}$$

Our choice is determined depending on the application and the computational effort for the Hessian H, which requires the calculation of second order derivatives of the solution operator, compare Eq. (4.17) with Eq. (4.15).

Optimal design of experiments

In general, the goal in optimum experimental design is to minimise the confidence region of the parameter estimates by changing the experimental setup, namely, sensor locations and types represented by the variable ω, boundary and initial conditions described by the inputs u_j, etc. Since we employ a linear approximation of the confidence region the aim is to minimise the "size" of the covariance matrix C. There exists extensive research on different design criteria Ψ that measure the "size" of a matrix in the context of optimum experimental design [49, 143, 158]. We list a few prominent options:

$$\Psi_A(C) = \mathrm{trace}\,(C), \qquad \Psi_D(C) = \det\,(C), \qquad \Psi_E(C) = \lambda_\mathrm{max}\,(C).$$

Depending on the application, the computational effort, and the adaptability of the experimental setup, we formulate slightly different optimisation problems. If the calculation of the Hessian H is fast, the number of sensors is small and the experimental setup is limited to adapting sensor positions only, we consider the matrix C_S from above in the optimisation model:

$$\begin{aligned} \min_{\omega} \quad & \Psi\left[C_\mathrm{S}(\omega, p(z,\omega), \mathcal{S}(u, p(z,\omega)))\right] \\ \text{s.t.} \quad & \omega \in \{0,1\}^{n_\mathrm{s}}, \quad G(\omega) \leq 0. \end{aligned} \tag{4.18}$$

where $\bar{\mathcal{F}}$ is a set of functions.

The improved surrogate model is constructed as in Sect. 4.3.7, i.e. it is defined by

$$\hat{m}_n(x) = \hat{m}_{L_n}(x) + \hat{m}_n^{\epsilon}(x) \quad (x \in \mathbb{R}). \tag{4.69}$$

In order to estimate the probability density function g of Y, the kernel density estimator of [123, 135] is applied on the sample $\hat{m}_n(\bar{X}_{N_{n,1}+1}), \ldots, \hat{m}_n(\bar{X}_{N_{n,1}+N_{n,2}})$, i.e. it is defined by

$$\hat{g}_{N_{n,2}}(y) = \frac{1}{N_{n,2} \cdot h_{N_{n,2}}} \sum_{i=N_{n,1}+1}^{N_{n,1}+N_{n,2}} K\left(\frac{y - \hat{m}_n(\bar{X}_i)}{h_{N_{n,2}}}\right), \tag{4.70}$$

for some bandwidth $h_{N_{n,2}} > 0$ and some kernel $K\colon \mathbb{R} \to \mathbb{R}$, which is usually chosen as a symmetric and bounded density, e.g. the Gaussian kernel $K(t) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}t^2)$.

Application

As an example we consider a lateral vibration attenuation system with piezo-elastic supports. A visualisation of the technical system can be found in [103, Fig. 1].

This system consists of a beam with circular cross-section embedded in two piezo-elastic supports A and B. Support A is used for lateral beam vibration excitation and support B for lateral beam vibration attenuation, as proposed in [61]. The two piezo-elastic supports A and B are located at the beam's end; each consists of one elastic membrane-like spring element made of spring steel, two piezoelectric stack transducers arranged orthogonally to each other and mechanically prestressed with disc springs, as well as the relatively stiff axial extension made of hardened steel that connects the piezoelectric transducers with the beam. For vibration attenuation in support B, optimally tuned electrical shunt circuits are connected to the piezoelectric transducers [63].

Our aim is to quantify uncertainty, i.e. to estimate the probability density function of the maximal amplitude of the vibration occurring in an experiment with this attenuation system. Five parameters vary during the construction of the attenuation system and influence the maximal vibration amplitude: the lateral stiffness $k_{lat,y}$ and $k_{lat,z}$ in direction of y and z, the rotatory stiffness $k_{rot,y}$ and $k_{rot,z}$ in direction of y and z, and the height of the membrane h_x. In our setting these five values are the input X of the experiment with the technical system. A computer model (above denoted by m) is available with which we can compute an approximation $m(X)$ of the maximal vibration amplitude Y to a corresponding input value X. To apply the density estimator (4.70) we measured the corresponding parameters for ten real built systems. As a result we got the data in Table 4.4.

Since the parameters vary in scale, it does not make sense to estimate the surrogate model $\hat{m}_{L_n}$ on $U_{i,n} \sim U([-c \cdot \log(L_n), c \cdot \log(L_n)]^d)$. Instead we rescale the

' (%%%

a) $\theta = h(\theta_1)$

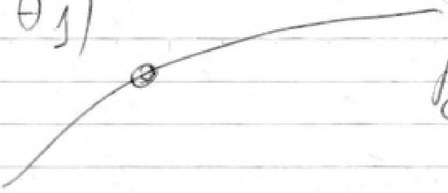

faz uma curvatura.

Pag. 147 (159) - Gauss-Newton approach

Pag. 181 (193) - Gaussian kernel

$$K(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} t^2\right)$$ ⟸

2021 1012 / Nonlinear Differential Equations.. /

Nonlinear Differential Equations ... pdf

Pag. 10 (11) - Gaussian distribution

Pag. 6 (307) - 2.1 Numerical approximation

Pag 7 (308) - Gauss-Legendre

Galerkin method using the quadrature rule of Gauss-Legendre (as present in [9]).

Pag. 8 (309) - , using Gauss-Legendre quadrature twice yields. (Bordas).

'

-256- - -* ! - '
8 ! 4 ! => ! &9 L? 1"
(7! ! " D

Johnson [15] solved Eqs. (31), (32), (36), and (37), **Figures 4** and **5** compare the numerical and analytical solutions for the SLS and Maxwell models, respectively. The solid lines are the numerical solutions obtained by the author, and the dashed lines are the analytical solutions solved by Johnson [15]. In **Figures 4** and **5**, r_h is the contacting region at time zero, p_h is the maximum contacting pressure at time zero, and T is the relaxation time.

Figures 4 and **5** indicate the deviation between the numerical and analytical results is less than 10%. Therefore, the numerical code can be used to simulate the visco-elastic deformations of rough fractures. For the two validation cases, the numerical simulation accuracy is not strongly dependent on the total number of elements, but on the time interval Δt. The deviation between numerical and analytical solutions will be smaller if the time interval Δt is reduced.

3. One example: visco-elastic deformations of rough rock fractures

3.1 Brief introduction of Brown's (1995) model

In this chapter, synthetic fracture surface pairs are generated based on Brown's model [10]. Brown's probabilistic model assumes that the surface is self-affine, and the surface height distribution follows Gaussian distribution [10]. The surface geometry can be completely described by three parameters: the Hurst exponent H, the mismatch length λ_σ and the root mean square roughness RMS.

Mathematically, a self-affine surface is defined as:

$$z(x) \sim \varepsilon^{-H} z(\varepsilon x) \tag{38}$$

where H is the Hurst exponent, z is the surface height, and ε is a constant for scaling at the x-direction. The H value is between 0 and 1, and it describes local roughness. A smaller H value corresponds to a rougher local surface profile.

The H value can be obtained from the power spectrum of surface height. The power spectrum of a surface can be obtained by decomposing the surface profile into a series of sinusoidal waves via Fourier transform, and each sinusoidal wave has its own amplitude A, wavelength λ, and phase. **Figure 6** shows the schematic of the decomposition process. The power (A^2) is defined as the square of the amplitude A; and the plot of power versus the wavelength number (the inverse of wavelength, which is 2π/λ) is defined as the power spectrum. **Figure 7** shows the schematic of power spectrum.

For a self-affine fracture surface, the power C ($=A^2$) can be related to the wavelength number q (=2π/λ) as:

$$\mathrm{C(q)} \sim q^{-2(1+H)} \tag{39}$$

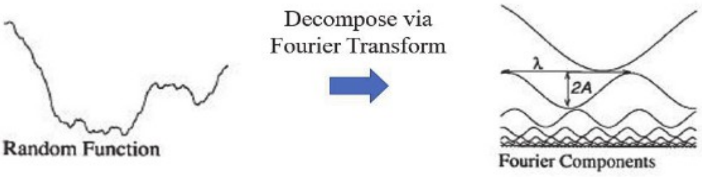

Figure 6.
Schematic of wave decomposition via Fourier transform (Kang et al. [13]).

In **Figure 7**, the q has an upper bound and a lower bound. For the lower bound, $q_{min} = 2\pi / \lambda_L$, where λ_L is the surface dimension; for the upper bound, $q_{max} = 2\pi/\lambda_1$, where λ_1 is the surface measurement resolution.

The second parameter is the mismatch length, λ_c. As illustrated in **Figure 6**, each wave component has its own wavelength λ. Glover et al. [16] and Brown [10, 17, 18] stated that for most natural rock joints, the two surfaces have relative shear displacements. At long wavelengths, the wave components match well; at short wavelengths, the wave components are not identical. Based on the above observation, Brown [10] proposed a parameter: critical wavelength λ_c, which is also called the mismatch length scale. Brown [10] assumed that above the mismatch wavelength, the decomposed wave components of two surfaces match perfectly; they have the same amplitudes, wavelengths, and phases. On the contrary, below the mismatch wavelength, the decomposed wave components of two surfaces do not match; they have the same amplitudes and wavelengths, but the phases are independent. **Figure 8** illustrates the concept of the mismatch wavelength.

The third parameter is the root mean square roughness, RMS. It represents the absolute scale of surface asperity elevation. Mathematically, the RMS is defined as:

$$\sigma^2 = \int_{q_{min}}^{q_{max}} C(q)dq \tag{40}$$

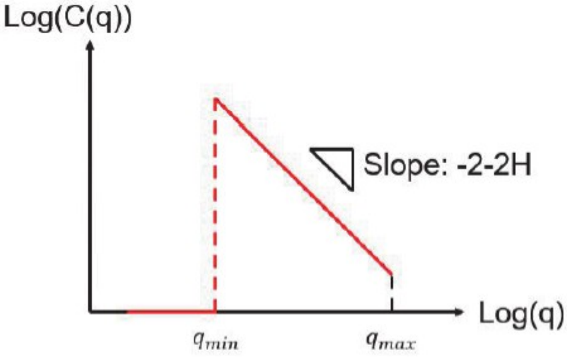

Figure 7.
Schematic of a power spectrum (Kang et al. [13]).

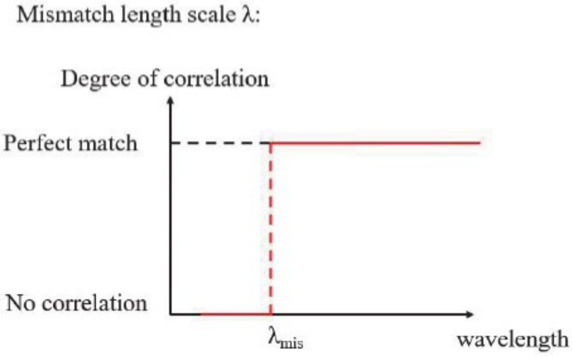

Figure 8.
Illustration of the mismatch wavelength (Kang et al. [13]).

where C is the power, q is the wavelength number, and σ is the RMS value. When generating the synthetic surface, the surface heights are normalized by its own RMS value, σ_{ini}, and then multiplied by the designated RMS value, σ_{des}:

$$z_{des} = z_{ini}\,\frac{\sigma_{des}}{\sigma_{ini}} \tag{41}$$

where z_{ini} is the initial surface height and z_{des} is the surface height after linear scaling. In this chapter, only the key mathematical concepts of Brown's [10] model is introduced; other details can be found in [10].

3.2 Generated synthetic surface pairs

Brown [10] measured the Hurst exponent H, mismatch length λ_σ and RMS for 23 natural rock joints. His measurement results imply that the H value is normally between 0.5 and 1.0; the normalized λ_c value (λ_c /fracture profile length) is normally between 0.02 and 0.2, and the normalized RMS value (RMS/fracture profile length) is normally between 0.005 and 0.015. Based on the above conclusion, seven synthetic fracture surface pairs are generated, with different H, λ_c, and RMS values. **Table 1** summarizes the parameters of the seven synthetic surface pairs. It is worth noting that surface pair No. 2 is the reference surface pair.

Table 1 shows that between surface pairs 1, 2, and 3, the H value is varied; between surface pairs 2, 4, and 5, the λ_c value is varied; between surface pairs 2, 6, and 7, the RMS value is varied. For each surface pair, the aperture distribution field can be plotted. **Figure 9** plots the aperture fields for surface pairs 1, 2, and 3; **Figure 10** plots the aperture fields for surface pairs 2, 4, and 5, and **Figure 11** plots the aperture fields for surface pairs 2, 6, and 7.

Based on **Figures 9–11**, we have the following observations:

1. According to **Figure 9**, when H increases, the average and standard deviation of the aperture decreases;

2. According to **Figure 10**, when λ_c deceases, the average and standard deviation of aperture decreases;

3. According to **Figure 11**, the average and standard deviation of aperture scales linearly with the RMS value.

Surface Pair No.	Profile length L (mm)	H	λ_c		RMS	
			λ_c/L	Absolute value (μm)	RMS/L	Absolute value (μm)
1	10	0.6	0.1	1000	0.005	50
2	10	0.8	0.1	1000	0.005	50
3	10	1.0	0.1	1000	0.005	50
4	10	0.8	0.2	2000	0.005	50
5	10	0.8	0.3	3000	0.005	50
6	10	0.8	0.1	1000	0.010	100
7	10	0.8	0.1	1000	0.015	150

Table 1.
The parameters of the seven synthetic surface pairs.

2.1 Numerical approximation

This subsection is devoted to the numerical approximation of the obtained IBVP (17) for $\alpha = 0$ and B given in (20). Our strategy is to seek numerical simulations that permits to avoid any boundary reflections and in some way renders the fully discrete scheme unconditionally stable.

Let N, M be integers, we define time step $\Delta t = \frac{T}{M}$ and spatial step $h = \frac{b-a}{N}$. The grids $t_n = n\Delta t, 0 \leq n \leq M$ and $x_i = a + ih, 0 \leq i \leq N$ are used to discretize Ω_T. Throughout this paper, we denote u_i^n the considered approximation of $u(t_n, x_i)$ and the set $\mathbb{N}_k^l = \{m \in \mathbb{N},\ k \leq m \leq l\}$.

2.1.1 Approximation of the governing equation

We describe a discretization for the BBM equation by the Crank-Nicholson time scheme as follows

$$\frac{u(t_{n+1},x) - u(t_n,x)}{\Delta t} - \beta \frac{\partial_{xx}u(t_{n+1},x) - \partial_{xx}u(t_n,x)}{\Delta t} + \gamma \frac{\partial_x u(t_{n+1},x) + \partial_x u(t_n,x)}{2} = 0, \tag{21}$$

$$u(t_0,x) = u_0(x), \qquad n \in \mathbb{N}_0^{M-1}.$$

For the space finite difference scheme, we use the approximations

$$\partial_x u(t,x) \approx \frac{1}{2h}(u(t,x+h) - u(t,x-h)),$$
$$\partial_{xx} u(t,x) \approx \frac{1}{h^2}(u(t,x+h) - 2u(t,x) + u(t,x-h)).$$

The fully discretization then writes,

$$\begin{aligned} &\frac{u_i^{n+1} - u_i^n}{\Delta t} - \beta \frac{\left(u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}\right) - \left(u_{i+1}^n - 2u_i^n + u_{i-1}^n\right)}{h^2 \Delta t} \\ &\qquad + \gamma \frac{\left(u_{i+1}^{n+1} - u_{i-1}^{n+1}\right) + \left(u_{i+1}^n - u_{i-1}^n\right)}{4h} = 0, \quad (i,n) \in \mathbb{N}_1^{N-1} \times \mathbb{N}_0^{M-1} \\ &u_i^0 = u_0(x_i), \qquad\qquad i \in \mathbb{N}_0^N. \end{aligned} \tag{22}$$

2.1.2 Approximation of the boundary condition

The constructed boundary conditions (BC) contains time convolutions that are non-local and introduces many difficulties, for example, using a direct implementation leads to long and low accuracy. Several techniques have been used to overcome these problems by trying to localize the BC, see [5–8] for more details. The resulting localized BC are easy to implement and more efficient but tends to depend sensitively on the initial data. In our case, we utilize the Discrete Galerkin Method. The BC are formulated as Fredholm integral equations of second kind. The basic idea is to write the boundary condition on (20) in the form

$$\partial_n u(t,a) - \int_0^t \overline{K}_1(t,s)u(s,a)ds = 0, \tag{23}$$

$$\partial_n u(t,b) - \int_0^t \overline{K}_2(t,s)u(s,b)ds = 0. \tag{24}$$

where, the introduced Kernels $\overline{K}_1, \overline{K}_2$ represent a linear combination of the two Bessel functions of order 0 and 1 at the time $(t - s)$. After a space discretization we obtain

$$u(t, x_0) - \int_0^t K_1(t, s) u(s, x_0) ds = u(t, x_1), \tag{25}$$

$$u(t, x_N) - \int_0^t K_2(t, s) u(s, x_N) ds = u(t, x_{N-1}), \tag{26}$$

where $K_1 = -h\overline{K}_1$ and $K_2 = h\overline{K}_2$. Both resulting Eqs. (25) and (26) can be identified to the linear integral equation

$$y(t) - \int_D K(t, s) y(s) d\sigma(s) = z(t), \quad t \in D. \tag{27}$$

The Eq. (27) is a Fredholm integral equation of second kind, where D is a closed bounded set in $\mathbb{R}^m$, with $m \geq 1$. The approximation of such integral equation could be made by a discrete Galerkin method using the quadrature rule of Gauss-Legendre as presented in [9]. Based on this, the BC can be similarly discretized while considering the domain D as the time interval $[0, t]$. Precisely, we use the Gauss Legendre Quadrature of order q, labeled GLQ_q with zeros ξ_j and weights w_j being in the interval $[-1, 1]$ for $j \in \mathbb{N}_0^q$. Let $i \in \mathbb{N}_0^{N-1}$, we introduce the following transformation

$$\begin{array}{llll} F_i: & [-1, 1] & \rightarrow & [t_i, t_{i+1}] \\ & \xi & \mapsto & t_i \dfrac{1-\xi}{2} + t_{i+1} \dfrac{1+\xi}{2} \end{array} \tag{28}$$

The approximation of the BC is now given by

$$u(t_{n+1}, x_0) - \int_0^{t_{n+1}} K_1(t_{n+1}, s) u(s, x_0) ds = u(t_{n+1}, x_1) \tag{29}$$

$$u(t_{n+1}, x_N) - \int_0^{t_{n+1}} K_2(t_{n+1}, s) u(s, x_N) ds = u(t_{n+1}, x_{N-1}) \tag{30}$$

that is

$$u_0^{n+1} - \int_0^{t_{n+1}} K_1(t_{n+1}, s) u(s, x_0) ds = u_1^{n+1} \tag{31}$$

$$u_N^{n+1} - \int_0^{t_{n+1}} K_2(t_{n+1}, s) u(s, x_N) ds = u_{N-1}^{n+1} \tag{32}$$

For the seek of simplicity, we rewrite the integral terms of (31) and (32) in the form

$$\begin{aligned} \int_0^{t_{n+1}} K_i(t_{n+1}, s) u(s, x_k) ds = A_i \int_0^{t_{n+1}} u(s, x_k) ds + B_i \int_0^{t_{n+1}} J_0^c(t_{n+1} - s) \left(\int_0^s u(r, x_k) dr \right) ds \\ + D_i \int_0^{t_{n+1}} J_0^c(t_{n+1} - s) u(s, x_k) ds \\ := A_i I_1(t_{n+1}, x_k) + B_i I_2(t_{n+1}, x_k) + D_i I_3(t_{n+1}, x_k), \end{aligned}$$

such that $(i, k) \in \{(1, 0), (2, N)\}$, $A_1 = A_2 = -hC_1, B_2 = -B_1 = hC_2$ and $D_1 = -D_2 = hC_3$, all constants C_i are defined in (??). Thus, basing on the approach presented in [9], the GLQ_q applied to the integrals previously defined is described by the following, for $n \in \mathbb{N}_0^{M-1}$ and $k \in \{0, N\}$,

$$I_1(t_{n+1}, x_k) = \frac{\Delta t}{2} \sum_{i=0}^{n} \sum_{j=0}^{q} w_j u\Big(F_i\big(\xi_j\big), x_k\Big). \tag{33}$$

The second integral is more complicated since it involves two composing integrals, using Gauss-Legendre quadrature twice yields

$$I_2(t_{n+1}, x_k) = \frac{\Delta t^2}{4} \sum_{i=0}^{n} \sum_{j=0}^{q} w_j J_0\Big(t_{n+1} - F_i\big(\xi_j\big)\Big) \sum_{l=0}^{j} \sum_{m=0}^{q} w_m u(F_l(\xi_m), x_k) \tag{34}$$

and the remained integral is approximated by

$$I_3(t_{n+1}, x_k) = \frac{\Delta t}{2} \sum_{i=0}^{n} \sum_{j=0}^{q} w_j J_1\Big(t_{n+1} - F_i\big(\xi_j\big)\Big) u\Big(F_i\big(\xi_j\big), x_k\Big). \tag{35}$$

From approximations (33)–(35), the numerical solution on the interface of (17) can be given by

$$u_0^{n+1} = u_1^{n+1} + f_1^{n+1}, \tag{36}$$

$$u_N^{n+1} = u_{N-1}^{n+1} + f_2^{n+1}. \tag{37}$$

We accomplish this by simply adding (36) and (37) to the discretization of the interior governing Eq. (22). We obtain an implicit scheme that we illustrate by the following system in matrix form

$$\begin{cases} \big(I + \tilde{A} + \tilde{B}\big) U^{n+1} = (I + A - B) U^n + F^n, \quad n \in \mathbb{N}_0^{M-1}, \\ u_0^{n+1} = u_1^{n+1} + f_1^{n+1}, \\ u_N^{n+1} = u_{N-1}^{n+1} + f_2^{n+1}. \end{cases} \tag{38}$$

with

$$U^n = \begin{pmatrix} u_1^n \\ \vdots \\ u_{N-1}^n \end{pmatrix} \quad I = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}, \quad F^n = \begin{pmatrix} \left(C_{d_1} + C_{d_2}\right)\left(f_1^n - u_1^n\right) \\ 0 \\ \vdots \\ 0 \\ \left(C_{d_1} - C_{d_2}\right)\left(f_2^n - u_{N-1}^n\right) \end{pmatrix},$$

$$A = C_{d_1} \begin{pmatrix} 2 & -1 & 0 \\ -1 & \ddots & -1 \\ 0 & -1 & 2 \end{pmatrix}, \quad B = C_{d_2} \begin{pmatrix} 0 & 1 & 0 \\ -1 & \ddots & 1 \\ 0 & -1 & 0 \end{pmatrix}$$

and

$$\tilde{A} = C_{d_1}\begin{pmatrix} 2 - C_{d_1} & -1 & 0 \\ -1 & \ddots & -1 \\ 0 & -1 & 2 - C_{d_1} \end{pmatrix}, \quad \tilde{B} = C_{d_2}\begin{pmatrix} -C_{d_2} & 1 & 0 \\ -1 & \ddots & 1 \\ 0 & -1 & C_{d_2} \end{pmatrix}.$$

where the discretization constants are $C_{d_1} = \frac{\beta}{h^2}, C_{d_2} = \frac{\gamma \Delta t}{4h}$.

3. Non standard boundary conditions for the BBM equation

In [4], we have presented a new method to derive transparent boundary conditions for the BBM equation. These boundary conditions have the advantage of being local in time but needs an additional function construct the BC which means bigger system to be solved. We recall that the problem designed to be the restriction in Ω_T of the BBM initial Eq. (1) with $\alpha = 0$ is given by

$$\begin{cases} \partial_t u - \beta \partial_{txx} u + \gamma \partial_x u = f & \text{in}]0, T] \times]a, b[\\ \partial_t v - \beta \partial_{txx} v + \gamma \partial_x v = g & \text{in}]0, T] \times]a, b[\\ \partial_x u = v & \text{on}]0, T] \times \{a, b\} \\ \beta \partial_{tx} v - \gamma v = \partial_t u - f & \text{on }]0, T] \times \{a, b\} \\ u(0, x) = u_0(x) & \text{on} [a, b] \\ v(0, x) = v_0(x) & \text{on } [a, b] \end{cases} \tag{39}$$

4. Numerical examples

We take an initial condition as solitary wave like function locally supported in Ω. The evolution of the solutions are plotted in different time steps before, under and after traveling the right boundary of the considered bounded domain. We save a reference solution that is numerically calculated in a broaden domain of Ω with Dirichlet boundary condition. We compute infinite error between numerical solutions using both formulations presented in this paper and the reference solution. We denote GLQ_i for approached solution with BEM and Gauss Legendre Quadrature in (2) for $i \in \{0, 1, 2\}$, while NSBC refers to numerical solution with non standard boundary conditions given in (3). We define the following errors

$$\|(u - u_{ref})(t, a)\|_\infty = \sup_{t \in [0, T]} |u(t, a) - u_{ref}(t, a)|, \tag{40}$$

$$\|(u - u_{ref})(t, b)\|_\infty = \sup_{t \in [0, T]} |u(t, b) - u_{ref}(t, b)|, \tag{41}$$

$$\|(u - u_{ref})(t, x)\|_\infty = \sup_{(t, x) \in [0, T] \times]a, b\left[|u(t, x) - u_{ref}(t, x)|.\right.} \tag{42}$$

Let $\beta = \gamma = 1$, $T = 10$ and $u_0(x) = sech^2(x)$. The considered initial data is locally supported in the interval $\Omega =]-10, 10[$, since $u_0(10) = u_0(-10) \approx 8,25.10^{-9}$. We fix $h = 10^{-2}$ and we vary the time step Δt. For a better comparison of these methods, we compute CPU time, in seconds, needed for each one to obtain numerical solution.

_ 20211012/ Nuclear ...|
Nuclear and Radiochemistry ... pdf
Pág. 471 (488) Figure 12.93

The Gaussian curves were obtained by fits taking into account the yields for a total of about sixty detected nuclides, ...

—

Gaussian distribution - pages
209, 277, 278, 396, 409, 423, 424, 439, 449, 454,

_ 20211012/ Vertical External Cavity Surface Emitting Lasers/
Vertical... pdf
Pág. 111 (124) 4.2
The "Ideal" Laser (Space)
→ Laguerre-Gauss or
Hermite-Gauss basis
Pág. 131 (147) 4.10 Spatial Coherence

We can consider that a real-world Gaussian TEM_{00} mode is an ideal

$
! -256- - -* ! - $ 8 ! ! " 5!" 3
< "! ! ! " 211 ! 1"
(' ' < /% 7! ! A

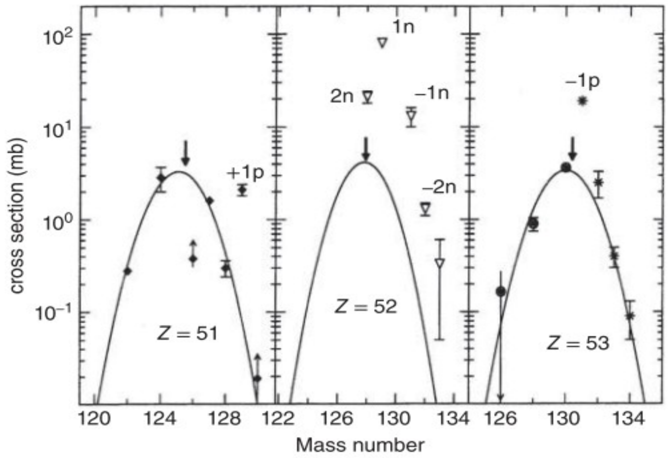

Figure 12.93 Yields for Sb, Te/Kr, and I ejectiles in the reaction ^{86}Kr + ^{130}Te at E_{cm} = 203.0 MeV. The Gaussian curves were obtained by fits taking into account the yields for a total of about sixty detected nuclides, but leaving out the predominantly quasi-elastic channels that are indicated in the figure. The arrows close to the top of the curves indicate the most favored mass at fixed charge from potential energy considerations. Source: Reisdorf et al. (1992)/Springer Nature.

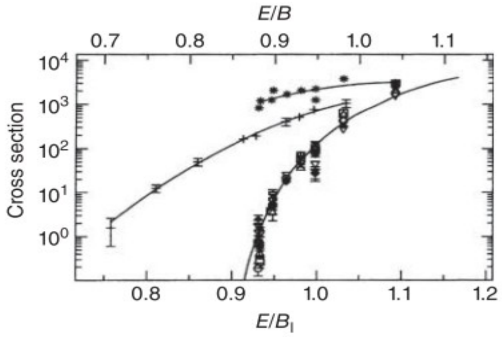

Figure 12.94 Comparison of excitation functions for one-neutron transfer, massive transfer, and fusion. Crosses, one-neutron transfer (in mb) for ^{238}U + ^{238}U; asterisks: one-neutron transfer (to ^{237}U in mb) for ^{238}U + ^{197}Au. All other symbols stand for complex transfer channels (in μb) ($^{223-225}$Ra, $^{225-226}$Ac, ^{227}Th) in the system ^{238}U + ^{197}Au. The cross sections for each isotope have been renormalized with an energy-independent factor to agree best with the values for ^{224}Ra, in order to stress the common shape of the excitation functions. While the smooth curves going through the neutron-transfer data are just to guide the eye, the curve going through the complex transfer data represents the measured fusion cross section for ^{40}Ar + ^{154}Sm scaled down by a factor of 100 and shifted towards 5% higher E/B_I (see the upper abscissa). The transfer data are from Funke et al. (1991). B_I is the interaction barrier, B is the adjusted neck formation barrier. Source: Reisdorf (1994/IOP Publishing.

7! ! " D 1! /J J J%/BJ' /J' %J' 'J'%/J''/J'$'

D -256- - -* ! - $! =L ! !A 3 !
= ?! I " 1"
(' 9 N&" ! N ?! : 7! D!

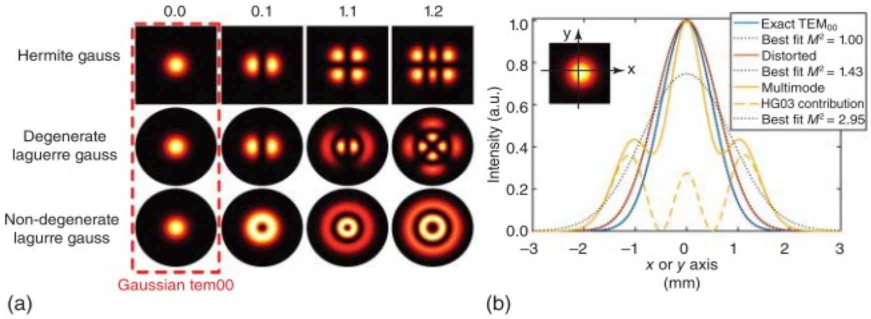

Figure 4.2 (a) Transverse intensity distribution of some possible transverse modes of Fabry–Perot cavities. (b) Exact TEM_{00} compared with distorted and multimode beam, with associated M^2 values.

spatial profile of the light. The VeCSEL is composed of a semiconductor chip, an output coupler, and in most cases an optical pump (electrical pumping is often possible) [9]. The semiconductor chip embeds both the gain medium and a multilayer High-Reflectivity (HR) backside mirror. The mm–cm long cavity can be adjusted for the continuous-wavelength tunability purposes, over wide spectral ranges (close to THz demonstrated [10]). In the following, we focus on the design parameters of a VeCSEL that demonstrates tunability, compactness, and high coherence.

4.2 The “Ideal” Laser

The purpose of what we call here “the ideal laser” in continuous-wave regime is to generate a high-power light (from some mW to multiwatt operation), which is strongly confined on three physical observable quantities:

- *Space*: the generated light has to be focusable in the smallest possible spot, as predicted.[2] by the diffraction. This property is called “spatial coherence” and leads to “brightness” in the case of high-power lasers. The spatial coherence also induces the possibility to collimate the beam over a very long distance. Most spatially coherent lasers emit solely the *transverse* TEM_{00} mode from the Laguerre–Gauss or Hermite–Gauss basis [11], (see Figure 4.2a). An exact TEM_{00} emission is said to be “diffraction limited.” Any mismatching observed between the emitted light profile and the exact TEM_{00} mode is usually estimated by means of the propagation factor $M^2 \geq 1$. $\sqrt{M^2}$ can be understood like the unexpected radius increase of the beam [12], compared with the diffraction-limited (theoretical) one, defined to be w_0 in near field and θ_0 in far field.[3] Both are related

---

[2] In extreme focusing conditions, its radius tends to be $2\lambda/\pi^2 \approx 0.2\lambda$ where λ is the wavelength.
[3] Both w_0 and θ_0 are evaluated at $1/\exp(2)$ from the maximum on the intensity profile. However, choosing an appropriate indicator to define the size of a beam is not something easy. We encourage the reader to examine the very interesting work of Siegman [12] for free-space beams, and the works of Petermann [13, 14] for confined/guided light.

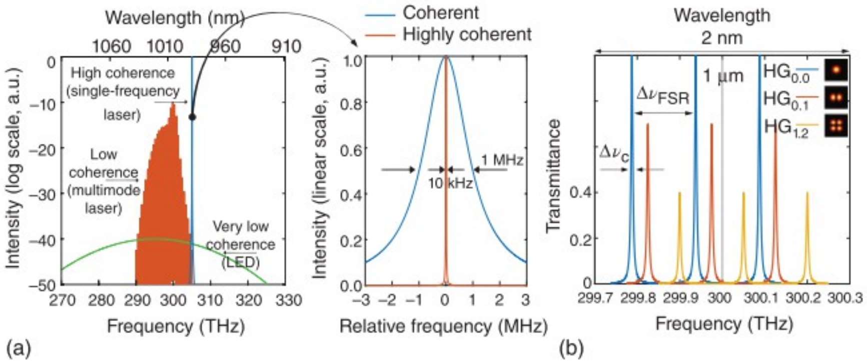

Figure 4.3 (a) Various degrees of time coherence, depending on multimode operation and also on the ability of the cavity to filter the spontaneous emission, thus on its Finesse $\mathcal{F}$. Please note the range ratio ($\approx 10^7$) between the two frequency ranges. (b) Longitudinal resonances of HG_{00} (TEM_{00}) for a passive Fabry–Perot cavity, as well as the resonances of two higher-order transverse modes. We also depict two fundamental parameters of the cavity transmittance spectrum: its periodicity $\Delta\nu_{FSR}$ and its FWHM linewidth $\Delta\nu_c$.

by the invariant parameter from the Gaussian beam theory $w_0\theta_0 = \lambda/\pi M^2$ [11]. However, it makes more sense, for better quantification or physical analysis, to perform a more advanced investigation, by observing the unexpected phase variations in the transverse plane, as well as performing a Zernike decomposition of the wavefront [15].

- *Time/spectrum*: the spectrum of the light source must be as narrow as possible. This property is called "monochromatic wave" or "long coherence time" and means that the electric field follows a perfect, pure, and unmodulated sine wave oscillation at the optical frequency ν_0,[4] that was born a long time ago and that will die in a very far future. Whatever the physical origin of the spectral width (multimode operation and/or temporal phase distortion, see Figure 4.3a and b), a classical way to quantify this time coherence is the Full-Width-Half-Maximum (FWHM) of the laser spectrum, which is, from high to low coherence, in the kHz–GHz range for most single-frequency free-running lasers.[5] However, because the possible physical reasons for broadening are many, studying the coherence of a laser by means of the spectrum FWHM is not always relevant. For better analysis and more accuracy, one may look at the time-domain fluctuations of the laser field instant frequency, often quantified its the Allan Variance, but more usually through its spectral density [17]. That's what we do in the following.
- *Polarization*: the polarization state of the electrical field must be unique in space and time. In most cases, it will be linear, following a specific direction, but it can also be circular, left or right-handed, or elliptical. This property is called "polarization coherence."

---

[4] This frequency is in the 100–1000 THz range for usual ultraviolet to infrared wavelengths.

[5] It can go close to Hz level for lasers actively locked on a frequency reference [16].

Aiming at this kind of coherence, the laser design will have to *induce the selection of a single transverse, longitudinal, and polarization mode* of the cavity (see Figure 4.3b). Of course, this ability for the laser strongly depends on its technology and design; however, *reaching a single-mode operation is only the first step*: it is necessary but not sufficient for high coherence, and it needs to be discussed in the following. These two topics are the purpose of the two next sections.

4.3 Toward Single-Mode Operation

In most cases, each mode of a passive cavity is identified by a specific optical resonance frequency (see Figure 4.3b). These resonances can however be displayed in a more convenient way (see Figure 4.4a-c), by using the dispersion relation that exists between the wave vector k and the optical frequency ν in free space:

$$\nu = c/\lambda = c/2\pi \times k \tag{4.1}$$

On the longitudinal axis, these resonance frequencies are well-known (-a) as each longitudinal mode resonates at a frequency located by one of the tooth of a frequency comb, periodicity of which is given by the Free Spectral Range, defined as follows:

$$\Delta\nu_{\text{FSR}} = c/(2nD) \tag{4.2}$$

where c is the speed of light in vacuum, n the optical index of the medium in the cavity, and D the cavity length.

Each mode of the transverse plane supports a frequency comb with exactly the same periodicity $\Delta\nu_{\text{FSR}}$ on the longitudinal axis. However, the frequency comb associated to a given transverse mode is shifted by an offset that depends on the base of modes supported by the cavity. For a VeCSEL, it is usually the Hermite–Gauss or

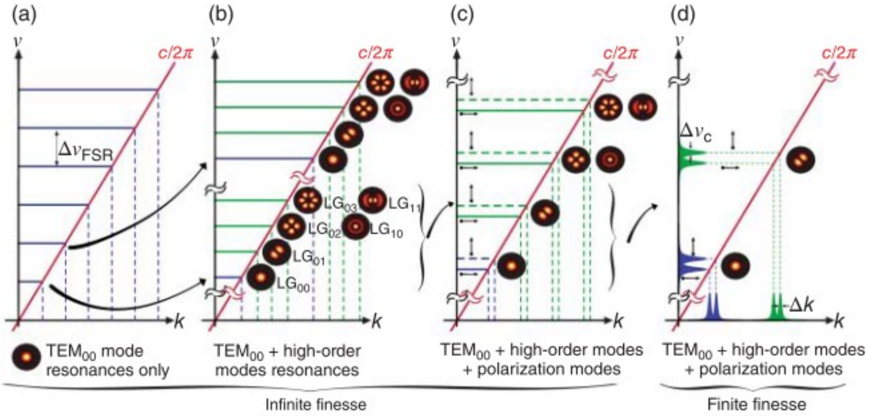
(a)
(b)
(c)
(d)
ν
c/2π
Δν_FSR
k
LG_03
LG_11
LG_02
LG_10
LG_01
LG_00
Δν_c
Δk
TEM_00 mode resonances only
TEM_00 + high-order modes resonances
TEM_00 + high-order modes + polarization modes
TEM_00 + high-order modes + polarization modes
Infinite finesse
Finite finesse

Figure 4.4 Relation of dispersion (Eq. 4.1) for optical modes. (a) Longitudinal modes for *TEM_{00}*. (b) *Zoom of (a)*: Transverse modes added. (c) *Zoom of (b)*: Polarization modes added. (d) *Zoom of (c)*: Impact of a finite Finesse on the wave vector k.

(% ' ' 1! ! 7! ! 9=

So, we made the theoretical proof as well as the experimental proof of the exceptional purity of the single-mode operation of the VeCSEL. We revealed its excessively pure photon state on all the three coherence axis thanks to the radio-frequency metrology, which offers much stronger measurement dynamics (by decades compared with usual methods!).

4.10 Spatial Coherence

As demonstrated above, VeCSELs can select a single transverse mode, which is a first important step to reach spatial coherence. However, this pure mode can be polluted by some spatial noise, (see Figure 4.7a). In this section we get a closer look at this *transverse spatial noise*, i.e. the static (time-domain independent) fluctuations superimposed on an ideal wavefront beam: the *time domain phase fluctuations* will be investigated in the next section.

We can consider that a real-world Gaussian TEM_{00} mode is an ideal spherical Gaussian wave, wavefront of which is polluted by some "spatial phase noise" in the transverse dimension of the wave, as illustrated in Figure 4.14. This leads to M^2 propagation factors higher than 1. In the case of VeCSELs, we measured that the wavefront deformation, compared with an ideal spherical Gaussian wave, is less than 1%, leading to excellent M^2 values (typically lower than 1.2). This kind of wavefront deformation typically originates from some residual, nonspheric, thermal lensing, or from the slight irregularities of various surfaces that the beam meets inside the cavity.

4.11 Time Domain Coherence and Noise

An accurate mean for the quantification of the time domain coherence is the study of noise. When talking about laser coherence, two kinds of fluctuations need to be discussed: *magnitude fluctuations* $A(t)$ (or *intensity fluctuation*) and *phase fluctuations* $\varphi(t)$ of the total electrical field $E(t) = (A_0 + A(t))\exp(j(\omega_0 t + \varphi(t)))$ as a function of the time.

The main characteristic of $A(t)$ and $\varphi(t)$ is that they are random functions of the time, i.e. not predictable, so they are called "noise." The word "noise" is often thought in terms of parasitics, which are external to the device. This is a narrow way of thinking: noise has fundamental reasons to be ingrained by the very physics of the device. Lasers are no exception.

4.11.1 Noise in Photonics: Basics

For example, *about* $A(t)$, the mere fact that a photonic source emits photons at random instants[19] leads to *intensity noise*. Actually, the optical power can be described in

---

[19] This is the case of most of them, even most lasers. A few counter-example exist in the literature, see a pedagogical example with high-efficiency Light-Emitting Diode [41] or in the case of lasers [42, 43].

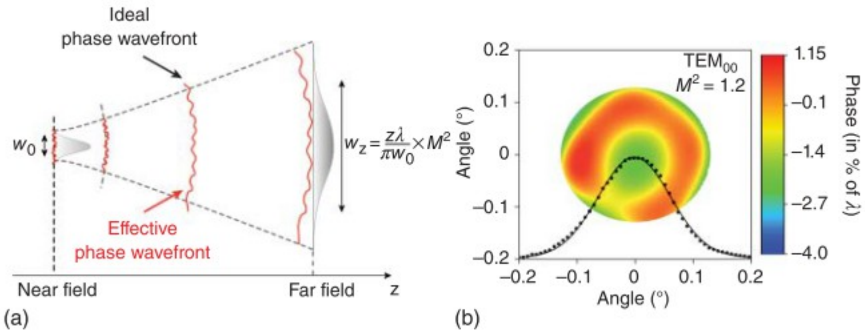
Ideal
phase wavefront
w_0
$w_z = \frac{z\lambda}{\pi w_0} \times M^2$
Effective
phase wavefront
Near field
Far field
z
(a)
0.2
0.1
0
−0.1
−0.2
Angle (°)
−0.2
−0.1
0
0.1
0.2
Angle (°)
TEM$_{00}$
$M^2 = 1.2$
1.15
−0.1
−1.4
−2.7
−4.0
Phase (in % of λ)
(b)

Figure 4.14 Spatial coherence of the VeCSEL beam. (a) Propagation of a nonideal Gaussian beam. (b) Typical transverse phase fluctuation in a VeCSEL.

a pure corpuscular vision like a flux of photons going through a surface, for example, the laser output coupler itself, or the surface of a detector. It is then defined like:

$$P_{\text{out}} = h\nu \frac{N}{\tau} \tag{4.15}$$

where N is the number of photon going through the considered surface during the time τ of the experiment. This is a very high rate signal: for a monochromatic emission of 1 mW at $\lambda = 1\ \mu\text{m}$, this leads to $N \approx 10^{16}$ for an observation of one second, with an intensity modulation of 100%, because at each instant, a photon is emitted or not, randomly. As we use bandwidth-limited detectors, we filter this high rate signal and observe an average value with moderated fluctuations (see Figure 4.15a). Due to its random nature, it is complex to speak about the time function of a noise because the magnitude in time is totally unpredictable, and we thus prefer to work with a statistically stable indicator such as its variance, or the square root of the variance, called effective value or RMS (Root Mean Squared) value.

Whatever the physical origin of the noise for a given photonic source, it should always be compared with a fundamental, nearly unavoidable, noise reference: *the shot noise*. The shot noise is nearly unavoidable, because optical attenuation,

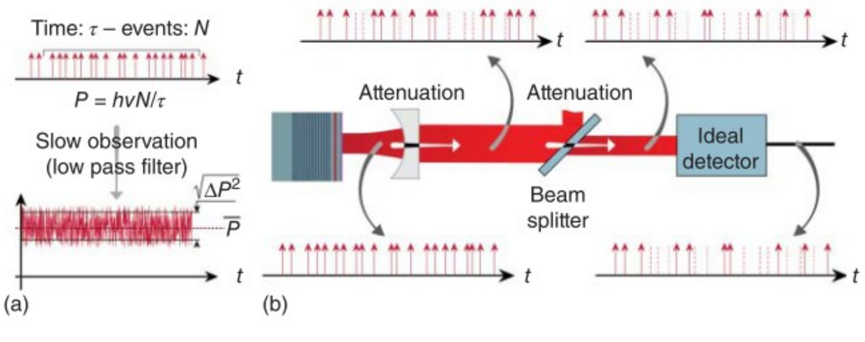
Time: τ – events: N
t
$P = h\nu N/\tau$
Slow observation
(low pass filter)
$\sqrt{\Delta P^2}$
$\overline{P}$
t
(a)
t
t
Attenuation
Attenuation
Ideal
detector
Beam
splitter
t
t
(b)

Figure 4.15 (a) Macroscopic intensity noise related to microscopic events. (b) effect of the attenuation on the photon statistics, while extracting light from a cavity or passing through a lossy optical system.

whatever its origin (attenuator, absorption, beam-splitter, diffraction, …), is nearly inescapable in practical photonic systems. Because attenuation selects randomly the photons that can pass, it alters the statistic of the attenuated beam, compared with the incident one. In the case of lasers, and especially for high finesse cavity lasers, the shot noise is usually unavoidable because the intracavity power experiences a huge attenuation when being coupled to the outside of the cavity [44] (see Figure 4.15b). We must point out that, in most usual cases, the magnitude of the shot noise is low compared with other kinds of noise, and for that reason the existence of the shot noise is rarely a limit[20]: it is more to be understood like a practical reference level of noise. This shot noise is quite easy to quantify. We simply have to evaluate the average value and the variance of N_1, the amount of photons passing through an attenuator of transmittance T [44]:

$$\overline{N_1} = T \times \overline{N} \quad \text{and} \quad \overline{\Delta N_1^2} = \overline{\left(N_1 - \overline{N_1}\right)^2} = \overline{N_1^2} - \overline{N_1}^2 \tag{4.16}$$

Assuming all the attenuation events are independent, we can calculate $\overline{\Delta N_1^2}$ for a single photon and apply a scaling factor by $\overline{N}$ to take into account all the photon selection events: $\overline{\Delta N_1^2} = \overline{N}T(1 - T)$. Then, Eq. (4.15) leads to the fluctuation that occurs on an optical power $P_{\text{att}} = P_{\text{out}} \times T$ that experienced some attenuation [44]:

$$\overline{P_{\text{att}}} = h\nu\frac{\overline{N_1}}{\tau} \quad \text{and} \quad \overline{\Delta P_{\text{att}}^2} = h^2\nu^2\frac{\overline{\Delta N_1^2}}{\tau^2} = \frac{h\nu}{\tau}\overline{P_{\text{out}}}T(1 - T) \tag{4.17}$$

This relation is dependent on the observation time τ, which is not practical. Usually, people prefer to work with the power spectral density $S_\text{P}(f)$ of the signal $P(t)$, which is simply related to the variance or its square root P_{RMS} as follows:

$$P_{\text{RMS}} = \sqrt{\overline{\Delta P^2}} = \sqrt{\int_{f_{\text{low}}}^{f_{\text{high}}} S_P(f)\text{d}f} \tag{4.18}$$

$S_\text{P}(f)$ can be understood like the spectral distribution of the noise.[21]

Because observing a signal over a time τ is equivalent to observing it with a spectral bandwidth[22] $\Delta f = (2\tau)^{-1}$, we can deduce from Eqs. (4.17) and (4.18) that the spectral density due to the shot noise is given by $S_{\text{shot}} = 2h\nu\overline{P_{\text{out}}}T(1 - T)$. In the case of laser experiments, strong attenuations are usual ($T \rightarrow 0$), and thus $T(1 - T) \rightarrow T$, which leads to:

$$S_{\text{shot}}(f) = 2h\nu\overline{P_{\text{att}}} \tag{4.19}$$

---

[20] Except in extreme setup like spin noise spectroscopy [45], or in very long-range doppler detection [46].

[21] As depicted above, due to its random nature, the time function of a noise is complex to describe with mathematics. It is exactly the same for the Fourier transform of the noise because it exhibits the same random feature as the time function. We thus use the power spectral density $S_\text{P}(f)$, which is a stable indicator of the spectral repartition of the noise.

[22] This equivalence is due to the fact that the Fourier transform of a time window τ is a cardinal sine which width is $\Delta f = (2\tau)^{-1}$.

This expression does not depend on the frequency, and this means that the shot noise spectrum is flat, with uniform value for all frequencies: it is a *white noise*. Intuitively, we feel that there is no reason for an attenuation process to have a specific time constant associated. If we take a source with a self noise, spectral density of which is $S_{\text{laser}}(f)$, because the attenuation process is not dependent on the source statistics, the observation of this source through the attenuator should lead to:

$$S_{P_{\text{att}}}(f) \approx S_{\text{laser}}(f) \times T^2 + 2h\nu\overline{P_{\text{out}}} \times T \tag{4.20}$$

Finally, increasing the attenuation ($T \to 0$) leads to observe only shot noise and hides the contribution of $S_{\text{laser}}(f)$.

When detected by a photodiode, a quantum efficiency $\eta_q < 1$ can be understood as if all the photon events were not transfered into electron events. This is similar to the detection with an ideal photodiode (that means $\eta_q = 1$, thus a sensitivity $= hc/q\lambda$ in A/W) placed after an optical attenuator, transmittance of which is equal to the quantum efficiency η_q. And so the simple detection process with a real-life detector can introduce a shot noise contribution too. Because of all the possible attenuation processes described above, it is usually assumed that the detected current contains a part of shot noise, value of which is $S_{\text{shot}}(f) = 2qi_{\text{ph}}$, where i_{ph} is the average detected photocurrent. For that reason, intensity noise is usually evaluated by the "Relative Intensity Noise" (RIN), which does take into account the fact that the shot noise may come from the measurement setup itself:

$$\text{RIN} = \frac{S_{\text{iph}} - 2qi_{\text{ph}}}{i_{\text{ph}}^2} \tag{4.21}$$

where S_{iph} is the spectral density of the photocurrent measured in an experiment similar to the one depicted in Figure 4.15b.

The measurement is usually performed thanks a low-noise-amplified photodiode, output of which will be processed to perform power spectral density spectra averaging. Averaging is necessary because the power spectral density is evaluated during a finite measurement time (the one of the experiment); so, its experimental estimation undergoes a statistical error, which can be reduced M times by accumulating M averages (due to the law of the large numbers). The power spectral density may be estimated by a RF sweeping spectrum analyzer or by measuring the time signal with an acquisition card and some Fast-Fourier-Transform (FFT) processing [47, 48]. The latter is more interesting, when possible, because in most practical cases, the computational efficiency of the FFT algorithm leads to petty computation times.[23] For that reason, the time necessary for the acquisition of a single spectrum is the one necessary for the acquisition of the time domain signal, defined as RBW^{-1} where RBW is the Resolution BandWidth of the spectrum. It is far quicker than with a RF sweeping analyzer, which requires at least $K \times \text{RBW}^{-1}$ with K the amount of points in the spectrum. This necessary time for the acquisition is an important topic because, as explained, accurate noise measurements rely on averaging spectra. An example of result of this process is given for S_{iph}, as well as for the corresponding RIN spectrum, in Figure 4.16a.

---

[23] This is true in lots of practical cases, when the wanted spectral resolution is narrow. In very specific cases, dedicated electronics can be involved to speed up the FFT computation [45].

To sum up, noise is usually quantified through the *power spectral density*, which indicates the spectral repartition of its fluctuations. The attenuation impedes an accurate laser power spectral density measurement because the *shot-noise* is superimposed. For that reason, the intensity noise is usually given by the RIN that takes into account the impact of the shot-noise on the measurement.

4.11.2 Intensity Noise of a VeCSEL

Like in most lasers, the VeCSEL noise has a fundamental cause: the spontaneous emission. However, this contribution is weak due to the high Finesse cavity design, and other technical noise, mainly coming from the pump, usually has a much stronger contribution (see Figure 4.16b and c). The pump noise itself depends on the kind of pump: usually single-mode laser diodes experience mainly $1/f$ noise at low frequency [49] due to the impact of the semiconductor cristal defects on the electronic transport [50]. Multimode pumps can deliver much higher power but may produce extra contributions to noise for optical reasons, coming from the multimode operation itself [51].

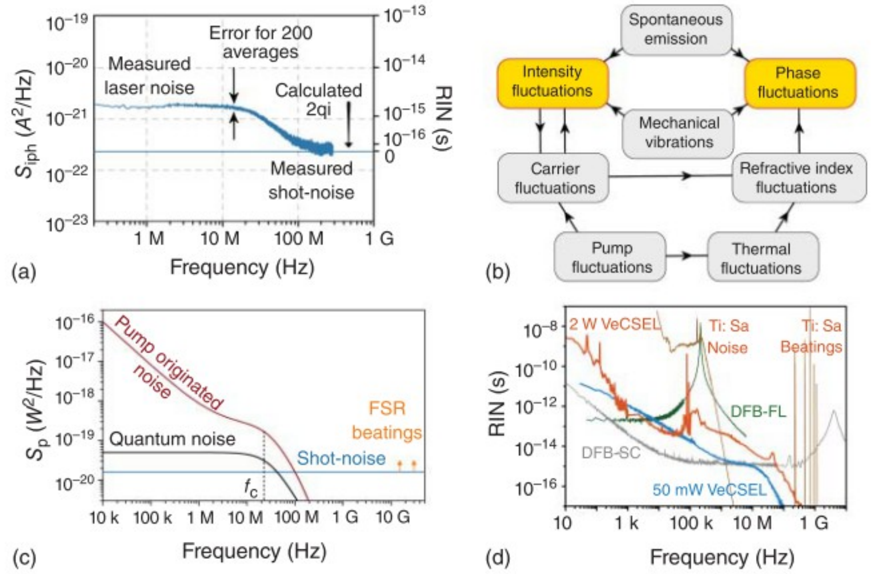

Figure 4.16 (a) Example of noise spectrum, exhibiting both RIN and spectral density. (b) Model of the VeCSEL Noise [18]. (c) Result from the model described in *(b)* in the case of intensity noise. (d) VeCSEL intensity noise compared with other technologies. For the VeCSEL, the cutoff is always below 100 MHz, and the intensity noise inside the VeCSEL bandwidth depends on the kind of pump: single-mode for the 50 mW VeCSEL, multimode in the case of the 2 W VeCSEL. The VeCSEL beatings were outside of the available spectral range (beyond 10 GHz) at the moment of the experiment.

-

spherical Gaussian wave, wavefront of which is polluted by some "spatial phase noise" in the transverse dimension of the wave, as illustrated in Figure 4.14.

- 20211012/ Semiconductors for Optoelectronics...../

Semiconductors for optoelectronics...pdf

Pág. 225(233) – Gaussian distribution

(7.19

- 20211012/ Mathematics and the physical world/

Morris Kline - Mathematics in western culture (1964).pdf

Pág. 396(425) –

Gauss expressed a horror of infinite quantities:

"I protest against the use of infinite magnitude... which is never permissible in mathematics".

B
! -256- - -* ! - B "
*1 1"
($ %% 7! ! " D /

normalised ***Doppler broadened lineshape*** and is, in fact, a Gaussian distribution:

$$g(\nu) = \frac{c}{\nu_0}\left(\frac{M}{2\pi k_B T}\right)^{1/2} \exp\left[\frac{-Mc^2}{2k_B T \nu_0^2}(\nu - \nu_0)^2\right] \tag{7.19}$$

where M is atomic mass. We can now obtain the transition rate for monochromatic light at a single frequency $W_i(\nu)$ by replacing the spectral energy density $\rho(\nu)$ in Eq. 7.15 with that for monochromatic light density ρ_ν and also by including line shape broadening:

$$W_i(\nu) = \frac{c^3 \rho_\nu}{8\pi n^3 h \nu^3 \tau_{21}} g(\nu) \tag{7.20}$$

Equation 7.20 can be written in terms of the optical intensity I_ν, in (W/m^2):

$$\begin{aligned} I_\nu &= c\rho_\nu / n \\ W_i(\nu) &= \frac{c^2 I_\nu}{8\pi n^2 h \nu^3 \tau_{21}} g(\nu) = \frac{\lambda^2 I_\nu}{8\pi n^2 h \nu \tau_{21}} g(\nu) \end{aligned} \tag{7.21}$$

where c is the speed of light, n is the refractive index and λ is the wavelength in vacuum.

Example A light bulb is operated at a temperature of $T = 1000$ K and has an average emission wavelength of 0.5 µm. Calculate the ratio of the stimulated emission to spontaneous emission rates.

Average emission frequency:

$$\nu = \frac{c}{\lambda} = \frac{3 \times 10^8 m/s}{0.5 \times 10^{-6} m} = 6 \times 10^{14}\,Hz$$

From Eq. 7.14 the ratio of stimulated and spontaneous emission rates are:

$$\frac{W_{21}^i}{W_{21}} = \frac{1}{\exp\left(\frac{6.626 \times 10^{-34} Js \times 6 \times 10^{14} s^{-1}}{1.38 \times 10^{-23} J/K \times 1000} K\right) - 1} = 3.1 \times 10^{-13}$$

This simple exercise proves that, for a system in thermal equilibrium, spontaneous emission will always dominate the stimulated emission. Ordinary light sources emit light in a wide spectrum and their emission is incoherent.

In order to produce coherent radiation, the stimulated emission rate needs to be increased. This implies that for a two level system, in Eq. 7.7, the population of the higher energy level (N_2) has to be increased over that of the lower energy level (N_1).

D -256- - -* ! - B + ! !
@ 7! !L3) . /B' 1"
(%/B ' ' 7! L1 " > ! N& 1 ! !
! " A 1 D ! ! O

XXV

The Paradoxes of the Infinite

> *We admit, in geometry, not only infinite magnitudes, that is to say, magnitudes greater than any assignable magnitude, but infinite magnitudes infinitely greater, the one than the other. This astonishes our dimension of brains, which is only about six inches long, five broad, and six in depth, in the largest heads.*
>
> VOLTAIRE

Tristram Shandy was hopelessly perplexed. He had begun to write his autobiography and found that he could record only half a day's experiences in one day of writing. Consequently, even if he were to start writing at birth and even if he were to live forever, he could not record his whole life, for at any time only half of his life would be recorded. And yet if he did live on indefinitely he ought to be able to record his whole life, for the experiences of his first ten years would be recorded by the end of his twentieth year; the experiences of his first twenty years by the end of his fortieth year; and so on. Thereby every year of his life would be reached at some time. Hence, depending on which way he reasoned, he could or could not complete his autobiography. The longer Tristram puzzled over this paradox the more confused he became and the farther he appeared to be from a decision.

Tristram's inability to resolve the paradox was really to be expected, for his problem involved an infinitude of time. The greatest mathematicians and philosophers from Greek times on had been plagued with problems involving infinite quantities and had not fared any better. For example, Galileo recognized that the number of whole numbers is infinite; that is, the number of these whole numbers is greater than any finite number that can be named. He recognized, also, that the number of even whole numbers is also infinite. Which of these two infinite sets, he asked, is the larger one? On the one hand, it seemed as though the first should be, for it contains all the numbers of the second set and more besides. On the other hand, to each number in the first set there corresponds exactly one number in the second, as 5 corresponds to 10. Also, to each number in the second set there corresponds exactly one in the first, as 10 corresponds to 5. In view of this *one-to-one correspondence* between the two sets there should be as many in the first as in the second. Galileo concluded that it was impossible to compare infinite quantities and abandoned further thought on the subject. He says 'infinity and indivisibility are in their very nature incomprehensible to us.' Leibniz, too, considered the very same question and concluded that the notion of the number of whole numbers is self-contradictory and should be rejected.

Not many years before a successful attack was finally made on the problems of the infinite, the supreme nineteenth-century mathematician, Karl Friedrich Gauss, expressed a horror of infinite quantities: 'I protest against the use of infinite magnitude . . . which is never permissible in mathematics.'

However much mathematicians recoiled from or repudiated the thought of infinite quantities, by the middle of the nineteenth century mathematics could no longer dispense with the concept. During the period from 1600 to 1850 mathematics had made gigantic strides. In this heroic age great intellectual adventurers had dared to leap over chasms of difficulties to goals envisaged by their genius and farsightedness. These trail blazers expected that others would build the bridges to support the circumscribed steps of the more cautious thinkers who were to follow.

But the bridges were not easily built. The attempts to fill the gaps left during the heroic age were frustrated by paradoxes, contradictions, and more paradoxes. There developed an imperative need for critical thinkers with imagination and daring of another kind, the kind that would be able to dispense with and even override intuition and 'common sense.' This need was finally met. Neither the more circumspect workers, however, nor the trail blazers could have anticipated the astonishing and profound disclosures which the critical efforts brought forth.

The first successful attack on the problems of the infinite was made by Georg Cantor. His father had urged him to study engineering, a more profitable pursuit than teaching, and Cantor had started to follow this down-to-earth career; he ended by contributing to the the most abstract regions of mathematics. His work got the reception that innovation and originality usually encounter—neglect, ridicule, and even abuse. One fellow-mathematician, Leopold Kronecker, attacked it viciously. Somewhat milder and more typical of the response to it was the remark made in 1908 by Henri Poincaré, the most famous of the late nineteenth-century mathematicians: 'Later generations will regard [Cantor's] *Mengenlehre* as a disease from which one has recovered.' Mathematicians, let it be known, are often no less illogical, no less closed-minded, and no less predatory than most men. Like other closed minds they shield their obtuseness behind the curtain of established ways of thinking while they hurl charges of madness against the men who would tear apart the fabric. So severe were the attacks on his work that Cantor began to doubt himself, became depressed, and suffered mental breakdowns.

Toward the end of his life (he died in 1918), the uncommon sense of his logic finally gained some recognition from a few colleagues. It is comforting to contrast Poincaré's statement above with one made just a little later by David Hilbert, the greatest mathematician of this century: 'No one shall expel us from the paradise which Cantor has created for us.' Today Cantor's work is so widely and completely accepted that many profound mathematicians are quite willing to devote themselves to the solution of further problems which the acceptance of Cantor's work brought in its train.

And now let us see how Cantor attacked the problem of infinite quantities. The most familiar examples of infinite collections are the collection of whole numbers, the collection of fractions, and the collection of all real numbers, that is, whole numbers, fractions, and irrational numbers such as $\sqrt{2}$, $\sqrt{3}$, and π. To obtain the number of objects in such collections by counting is impossible because the process is endless. On the other hand, describing them as infinite sheds very little light on them, for all this word says is that they are not finite. Such a description is about as informative as the statement that *Pithecanthropus erectus* is not a cow. We must substitute, if possible, a positive answer to the question of how many objects there are in infinite collections.

Cantor recognized, of course, that the number of objects in an infinite collection or class cannot be obtained by counting. He also recognized the deeper significance of another seemingly superficial observation. Suppose we have two classes of objects such that to each object in the first class there corresponds one and only one object in the second and conversely. For example, if a squad of soldiers, each carrying a gun, were to pass before us there would be just such a correspondence between soldiers and guns. Technically, the relation between the two classes, soldiers and guns, is described by the phrase *one-to-one correspondence.* Obviously, two classes that are in one-to-one correspondence must contain the same number of objects. Moreover, it is not necessary to count the classes to reach this conclusion.

Cantor's greatness lies in his perception of the importance of the one-to-one correspondence principle and in his courage to pursue its consequences. *If two infinite classes can be put into one-to-one correspondence then,* according to Cantor, *they have the same number of objects in them.* For example, the class of positive whole numbers

$$1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad \ldots$$

and the class of reciprocals of these numbers

$$1 \quad \tfrac{1}{2} \quad \tfrac{1}{3} \quad \tfrac{1}{4} \quad \tfrac{1}{5} \quad \tfrac{1}{6} \quad \ldots$$

are in the one-to-one correspondence whereby each number in the first class corresponds to one and only one number in the second, namely, its reciprocal. In like manner, to each number in the second class there corresponds one and only one in the first. Hence, these two classes have the same number of objects in them. The number that represents the quantity of objects in these particular classes Cantor designated by $\aleph_0$ (aleph-null). It is called a transfinite number.

To say that the number of positive integers, as well as the number of any set of objects in one-to-one correspondence with the positive integers, is $\aleph_0$ does not seem to answer the fundamental question of how many there are in each set. The reader may say that $\aleph_0$ is a stranger to him and gives him no information on the number of positive integers. The objection is not a valid one. This number is as informative as the number one billion billion. The latter number is no more than a symbol representing the quantity of objects in a particular collection just as $\aleph_0$ represents the number of positive integers.

XXVI

New Geometries, New Worlds

> *I have made such wonderful discoveries that I am myself lost in astonishment: Out of nothing I have created a new and another world.*
>
> JOHN BOLYAI

The first man to challenge Euclid was Euclid himself. The creator of the most widely and most completely accepted system of thought—the abode of truth, and the birthplace of philosophies and sciences—doubted his results even before he issued them to the world. Euclid's questioning of himself marked the beginning of a two-thousand-year 'behind the scenes' attack on the obvious.

It is well known that Euclidean geometry is founded on ten axioms whose truth appears so self-evident that no 'sane' man would dare question them. From this sound basis impeccable logic produced more 'truths' just as appealing and as immediately acceptable as the axioms. Two millenniums of application climaxed by the successes of the Newtonian era added practically incontrovertible evidence of the soundness and reliability of these truths. Century after century buttressed logic with experience and common sense with tradition, until Euclid's system acquired inviolate sanctity. By 1800 educated people were far more likely to swear by the theorems of Euclid than by any statement in the Bible.

Whether a person appealed to experience, accepted the Kantian philosophy, or inclined to the obvious, the inescapable conclusion appeared to be that Euclid was truth and truth, Euclid. Despite this enviable position, which Euclidean geometry possessed from the outset and which time enhanced, a few thinkers, including Euclid, were not at ease. They were disturbed by two apparently innocent axioms.

The first of these says that a line segment can be extended *as far as one pleases* in either direction. The second one is the axiom on parallels, which says that through a point P not on a line L there passes one and only one line M (in the plane of P and L) that does not meet L *no matter how far M and L are extended* (fig. 77). If the axioms of Euclidean geometry are accepted because experience with physical space warrants our doing so, then these axioms are open to some doubt. No man has had direct experience with what happens in space more than a few miles beyond the Earth. All we can really say is that these axioms appear to be true in the limited regions in which we actually move. And even here we cannot be too sure of our assertions, for it was pointed out in the chapter on projective geometry that we never *see* parallel lines even in the portion of space immediately around us. As we look into the distance along lines Euclid refers to as parallel, we find that they seem to meet.

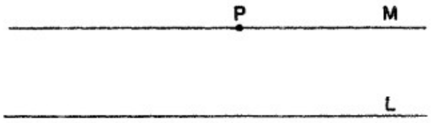

Figure 77. Euclid's parallel axiom

Euclid reveals his concern about these axioms by the manner in which he uses them. He does not use the parallel axiom, which is the more questionable of the two, until he has proved as many theorems as he can without it. He is equally cautious about the unlimited extensibility of the straight line. An examination of the theorems of his geometry shows that he uses line segments (portions of lines between two points) but never supposes that he has an infinite straight line to start with. When necessary he extends a segment in either direction only as far as the theorem requires. It should not be inferred that Euclid doubted the truth of these axioms; rather, because of their seemingly weighty implications, he would have preferred, no doubt, to derive their contents as consequences of simpler axioms.

A few hypercritical thinkers in every age following that of Euclid also hesitated to use as axioms statements on which any hard-headed businessman would bet his bottom dollar. To eliminate any lingering qualms these men from Missouri all tried the same thing. They concentrated on the parallel axiom and sought either to deduce it from the other axioms or to find a more acceptable substitute. Several hundred such worthy attempts on the part of the best mathematicians ended in failure, however. By 1800 the parallel axiom had come to be labeled the scandal of geometry.

It is hardly desirable, nor would it be too profitable, to review most of the efforts. The work of one man, however, the Jesuit priest Girolamo Saccheri, a professor of mathematics at the University of Pavia and a keen student of logic, merits our attention. Saccheri had a brand new idea. His novel attack on the problem of the parallel axiom was to argue in effect: Given a line L and a point P, then either (a) there is exactly one parallel to L through P, or (b) there are no parallels to L through P, or (c) there are at least two parallels to L through P. Alternative (a) was Euclid's parallel axiom. Suppose it were replaced by alternative (b), and the latter together with the other nine axioms of Euclid were shown to lead to contradictory theorems. Then surely alternative (b) could not be correct. Similarly, if the use of alternative (c) and the other nine Euclidean axioms led to contradictory theorems, then alternative (c) could not be correct. Then it would follow that Euclid's parallel axiom is the only possible one.

By using alternative (b) together with the other nine Euclidean axioms Saccheri did deduce theorems that contradicted each other. But he failed to deduce contradictions from the nine Euclidean axioms and the alternative axiom postulating the existence of at least two parallels. Though his efforts were determined and extensive, and though some of his deductions were indeed strange when compared with analogous results in Euclidean geometry, contradiction there was not.

Saccheri was on the threshold of an epoch-making discovery but he refused to step over it. For the moment we shall leave it to the reader to determine the conclusion Saccheri should have drawn from his failure to deduce inconsistencies. As for Saccheri himself, he was so unprepared for the strange theorems he established from his set of axioms that he decided Euclid's parallel axiom must be right. Accordingly, in 1733 he published his results in a book titled *Euclid Vindicated from All Defects*. Apparently when one man sets out to vindicate another he will most likely do so regardless of the facts.

One explanation for Saccheri's failure as well as the many others is that great as the mathematicians were who approached the problem presented by the parallel axiom, none was discerning enough to recognize and reject a two-thousand-year-old habit of thought. But within the mathematical world of the early nineteenth century a change in the intellectual milieu took place, bringing with it a sweeping, critical re-examination of fundamental beliefs. Undoubtedly this change accounts for the fact that three men, Gauss, Lobatchevsky, and Bolyai, with no knowledge of each other's thoughts, discovered the correct interpretation of Saccheri's work at about the same time; Lobatchevsky and Bolyai published their results within a few years of each other.

Of these three the greatest, and one who ranks with Newton and Archimedes, was Karl Friedrich Gauss. Karl showed unbelievable precocity in many fields and a particular predilection for mathematics. When as a young man he proved that the regular polygon of 17 sides could be constructed with straightedge and compass, he was so delighted that he abandoned his intention of becoming a philologist in order to study mathematics. He soon contributed masterful work to many branches of the subject and also achieved note as an inventor and experimentalist. Though his contributions were no less numerous and no less profound than those of other mathematicians, Gauss was extremely modest. He said, 'If others would but reflect on mathematical truths as deeply and as continuously as I have, they would make my discoveries.' Those who believe that genius is 99 per cent perspiration as well as those who despair of their mathematical abilities may find comfort in Gauss's statement.

Gauss was still a youth when the problem of the parallel axiom first came to his attention. At first he worked hard to replace the parallel axiom by a simpler one, and he failed. He then followed Saccheri's line of thought by adopting a parallel axiom contradicting Euclid's--essentially Saccheri's third alternative--and by deducing consequences from this new axiom and the other nine of Euclid's. Like Saccheri he arrived at strange theorems. Instead of allowing the strangeness to frighten him Gauss fought fire with fire. He drew the brand new, astounding conclusion which the great and near-great had failed to consider. He decided that *there can be other geometries as valid as Euclid's.*

Gauss had the intellectual courage to create non-Euclidean geometry but not the moral courage to face the mobs who would have called the creator mad, for the scientists of the early nineteenth century lived in the shadow of Kant whose pronouncement that there could be no geometry other than Euclidean geometry ruled the intellectual world. Gauss's work on non-Euclidean geometry was found among his papers after his death.

Of the two other men who deserve honor for the creation of non-Euclidean geometry the first was the gifted Nicholas Lobatchevsky. Born in 1793 to a poor Russian family, he studied at the University of Kasan and at the age of twenty-three became a full professor there. Lobatchevsky, too, was attracted to the problem of the parallel axiom. He says that he was struck by the fact that two thousand years of effort by the greatest mathematicians had failed to produce a better axiom. And so, like Saccheri and Gauss, he built a new geometry on the basis of a parallel axiom contradicting Euclid's. The almost unbelievable theorems to which he was led did not discourage him any more than they had Gauss. Sound reasoning had led to them and sound reasoning was the unquestionable guide. And so Lobatchevsky, too, affirmed the radical but inescapable conclusion: There are geometries different from Euclid's and just as valid.

The man who shares honors with Lobatchevsky for the discovery and the courage to publish his work on non-Euclidean geometry is the Hungarian, John Bolyai. Like the other two he was blessed by the gods and, in addition, was encouraged and developed by his father Wolfgang, also a mathematician. Wolfgang had himself been bitten by the bug of the parallel axiom problem and had spent many vain years in work upon it. He bequeathed it to his son who, in 1825, at the age of twenty-three, suddenly saw the light. There are axioms contradicting Euclid's, he contended, that can nevertheless serve as the basis for new geometries. John proceeded to build one. On the urging of his father, he published his work in 1833 as an appendix to his father's text.

How were the epoch-making documents of Lobatchevsky and Bolyai received? How did scientists react to the startling news that Euclidean geometry now had rivals? How did the very rational philosophers greet the most thoroughgoing refutation of the leading philosophy of the times? The works of Lobatchevsky and Bolyai were completely neglected. Moreover, in 1847 Lobatchevsky was dismissed by the University, despite brilliant contributions and unselfish devotion to his work. If Bolyai had been a professor rather than an Austrian army officer, he might have suffered the same fate.

About thirty years after Lobatchevsky and Bolyai published their monumental work, Gauss's correspondence on non-Euclidean geometry was published posthumously along with other papers. His name attracted attention to the subject and shortly thereafter the mathematical world began to read Lobatchevsky and Bolyai.

To appreciate their work on the problem of the parallel axiom we must go back for a moment. Consider any straight line L (fig. 78) and any point P not on L. Euclid's parallel axiom asserts that there is one and only one line K through P which does not meet L. Now let Q be any point on L. As Q moves to the right, the line PQ revolves counterclockwise about P and seems to approach the line K. In like manner, as Q moves to the left along L, the line PQ rotates clockwise about P and again approaches K. In each case, then, PQ approaches one and the same limiting line K.

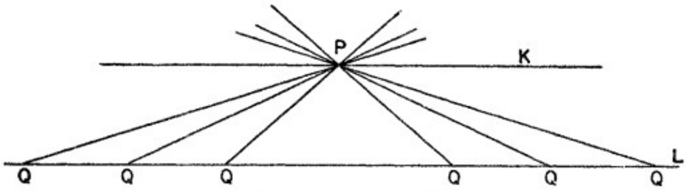

Figure 78. Euclid's parallel as a unique limiting line

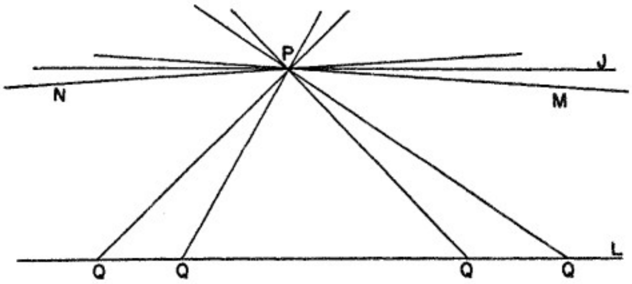

Figure 79. The parallel axiom of Lobatchevsky and Bolyai

Bolyai and Lobatchevsky, however, assumed that the two limiting positions of PQ are not the same line K but two different lines through P, and that these limiting lines, M and N (fig. 79), do not meet L. Moreover, they assumed that *every* line through P and between M and N, such as J, does not meet L. Hence Bolyai's and Lobatchevsky's parallel axiom affirms the existence of an infinite set of parallels to L through P. (These men reserved the word parallel for just the limiting lines M and N, but we shall use it to denote any line through P that does not meet L.)

The reader will perhaps feel, along with the mathematicians of Bolyai's and Lobatchevsky's time, that this is a ridiculous assumption to make. The diagram suggests that M and N will meet L if all three lines are extended far enough. Let us remember, however, that Bolyai and Lobatchevsky were interested in picking an axiom which, regardless of whether or not it describes the space we believe we live in, would be a logical alternative to Euclid's. And since the theorems to be derived from this axiom and the remaining Euclidean axioms would depend only on reasoning and not at all on accord with diagrams, the failure of the axiom to correspond with visual sensations is irrelevant.

What theorems were Bolyai and Lobatchevsky able to prove with their axioms? Of course, all the theorems of Euclid's geometry proved without the use of his parallel axiom are, automatically, theorems in Bolyai's and Lobatchevsky's geometry, since these men retained the other axioms of Euclid. As examples of such theorems we might mention these: Vertical angles are equal; from a point P at most one perpendicular can be drawn to a straight line; and, in a triangle with equal sides, the angles opposite these sides are equal.

Astonishing is the word for the theorems in the geometry of Bolyai and Lobatchevsky that do depend on their parallel axiom and therefore are not found in Euclid. These theorems, like all mathematical theorems, are proved by deductive methods of reasoning familiar to the reader; unlike the case in Euclidean geometry, however, figures are not nearly so useful in suggesting the steps of the proofs or in illustrating the meanings of the theorems.

Most unexpected is the theorem that *the sum of the angles of any triangle is always less than 180°*. Moreover, *of two triangles, the one with a larger area has a smaller angle sum*. Even more surprising is the fact that the new geometry wipes out a vital concept of Euclidean geometry, that is, that two geometric figures may have the same shape but different sizes. We say, in such a case, that the figures are similar but not congruent. In the new geometry *two similar triangles must also be congruent*. As a final example of a new theorem, we mention this one: *The distance between two parallel lines approaches zero in one direction along the lines and becomes infinite in the other direction.*

Bolyai and Lobatchevsky had succeeded in erecting a new geometry with many surprising theorems. But was their work any more than an exercise in logic? Let us realize, first, that hundreds of deductions in the new geometry had produced no theorems contradicting each other. This meant that the old parallel axiom could not be deduced from the other Euclidean axioms; otherwise the assumption of the new one would surely have led to contradictions within the system. It was not exactly news that the Euclidean parallel axiom could not be deduced from other Euclidean axioms. This fact had been suspected before.

The second implication in the work of Bolyai and Lobatchevsky was not anticipated. It is that we could not hope to establish the incontrovertible truth of the Euclidean parallel axiom by showing that any alternative produced contradictions. It was clear, therefore, that both of the schemes used by earlier mathematicians to vindicate the parallel axiom would never have been successful.

But the greatest significance of the new geometry was completely unexpected. Though the logical exercise was over the conclusion lingered on in people's minds: *There are geometries different from Euclid's.* A mathematician who possessed this knowledge was like a boy with an air rifle in his hands. The temptation to use it was too strong to resist. Euclidean geometry was known to be an accurate description of physical space. The non-Euclidean geometry of Bolyai and Lobatchevsky, on the other hand, did not seem and was not intended to apply to the physical world—but could it?

First reactions to this question are generally negative. If Euclidean geometry is correct, how can this new, conflicting geometry also be correct? Moreover, how can such absurd theorems apply to our familiar world? A little thought indicates that first reactions may be too hasty. What guarantee do we have that Euclidean geometry is correct? True it has been used for thousands of years. It also has the favor of long-established habits of thought. But let us recall Euclid's own reason for being concerned about the parallel axiom. Was it not that it makes an affirmation about regions of space far removed from man's daily experience, a space so vast that the accessible regions are in comparison only a dot on the surface of the Earth?

Who of us knows the geometry of the universe in the vicinity of Mars, or for that matter even ten miles above the surface of the Earth? By what right do we assume that it must be the same as what seems to apply on Earth? Euclidean geometry may be no better than the hundreds of scientific laws that served well enough in their day but ultimately had to be discarded.

After carefully considering just this problem Gauss suggested a criterion for determining the truth of Euclidean geometry. The sum of the angles of a triangle equals 180° in this geometry but is less than 180° in the new geometry. Hence measuring the angles of a triangle should decide which geometry fits the physical world. For two reasons a very large triangle had to be chosen. In the first place, the error in sighting is greater in a smaller triangle. In the second place, it is a theorem of Lobatchevsky's and Bolyai's geometry that the angle sum of a triangle approaches 180° as the triangle shrinks in size. For a small triangle the sum might be so close to 180° that measuring instruments might not be sensitive to the difference.

Gauss, himself, performed the experiment. He stationed an observer on each of three mountain peaks. Each observer measured the angle formed by his lines of sight to the other two observers. The sum of the angles of the triangle turned out to be within 2″ of 180°, so close that the difference could be charged to errors of measurement. Hence the experiment was not decisive.

The provoking aspect of Gauss's triangle test is that even under the best of experimental conditions it could never prove that space is Euclidean, for even if the measured angle sum should be 180° allowance would always have to be made for errors of measurement and hence for the *possibility* that the true sum is less than 180°. Actually the test involved two unwarranted assumptions, either one of which could invalidate a conclusion drawn from it. The first of these was that the triangle formed by three mountain tops is large enough to be decisive. The second assumption was that the light rays which formed the sides of the triangle follow straight lines. The rays may actually curve slightly but imperceptibly.

While Gauss's test can be dismissed as an interesting but inconclusive experiment, the larger question of the applicability of non-Euclidean geometry still deserves attention. The surprising fact that emerges from all attempts to decide which of the two geometries fits physical space is that both fit *equally well.* We have already indicated that in a small triangle the new geometry calls for an angle sum close to 180° and the smaller the triangle the closer must the sum be to 180°. If we applied non-Euclidean geometry and accordingly used angle sums that are slightly less than 180° no harm would be done from the practical standpoint. Similarly, no harm would be done if we assumed that given a point *P* and a line *L* an infinite number of parallels to *L* exist through *P* and in the plane of *P* and *L*.

We might think that the new geometry cannot be applied to the physical world because it asserts that similar triangles must be congruent. It certainly seems possible to construct two physical triangles that are similar but not congruent. In fact, one triangle could be made very large and the other very small. No matter how carefully the two triangles are constructed, however, we cannot be sure they really are similar—that is, that corresponding angles are exactly equal. The smaller triangle could have a larger angle sum, in accordance with the new, non-Euclidean geometry, but the difference might not be measurable. Hence for all practical purposes it would not matter whether we accepted the assertion of the new geometry. In other words, there is no way of deciding which geometry applies to physical space; either one could be used. Our prejudices and habits favor Euclidean geometry, which may also be somewhat simpler than the non-Euclidean geometry. But these reasons for preferring it do not discredit the applicability of the new geometry.

No doubt the reader is not satisfied. Perhaps he can be diverted with other, more pleasant arguments which show that non-Euclidean geometry can be applied to the physical world.

Let us return for a moment to Euclidean geometry. Imagine an enormous sheet of paper extending indefinitely in all directions. This sheet of paper is a physical illustration of the mathematical plane, the plane in which the theorems of Euclid's geometry hold. Now suppose we alter the shape of this vast sheet of paper by bending the left and right sides upward somewhat (fig. 80) so that it forms a curved surface, which, however, continues to extend indefinitely in all directions as did the plane. Such a surface is known as a cylindrical surface. As a result of the change in shape, most straight lines of the former plane become curves which, like the straight lines in the plane, are the shortest paths between the points they join on the surface. We shall call such curves *geodesics*. Two straight lines that were parallel in the plane become parallel geodesics, that is geodesics which do not meet on the surface. Triangles of the original plane become figures formed by arcs of geodesics on the surface. We shall call the new figures 'triangles' also. Circles on the plane give rise to figures we shall call 'circles.'

We come now to a very startling fact. Every axiom of Euclidean geometry holds for the figures on the cylindrical surface with the one provision—that we interpret the words line, triangle, and circle as

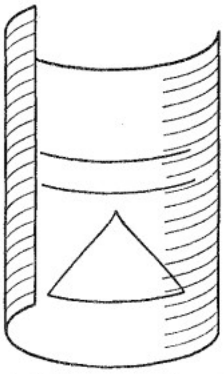

Figure 80. A new pictorial interpretation of Euclidean geometry

just suggested above. Hence *the theorems of Euclidean geometry, which follow from the axioms by deductive reasoning, a process completely independent of the pictures we may draw, also hold for figures on the curved surface.* To take one instance, the sum of the angles of a 'triangle' on the surface is 180°.

The reader may object to the argument above on the ground that straight lines and figures defined in terms of straight lines no longer have the proper meaning; they have lost their straightness. Now, however, we take advantage of a fact first pointed out in Chapter IV, namely, that the basic concepts of geometry such as point and line are undefined. We use only the properties of these concepts that are explicitly stated in the axioms. Hence if some new physical picture of line, say, has the properties required by the axioms, it is possible to adopt this new picture. Therefore it is logically justifiable to associate a completely new physical picture with all of Euclidean geometry.

The argument just made for the new physical interpretation of Euclidean geometry applies as well to non-Euclidean geometry. And if we do take advantage of our liberty to choose the physical interpretation of the line and other figures we obtain an intuitively acceptable interpretation of the new geometry.

Figure 81 illustrates a surface known as a *pseudosphere*. The curves on the pseudosphere which are the shortest paths between points on the surface—these special curves are also called geodesics—have the properties that straight lines possess in the axioms of Lobatchevsky and Bolyai. For example, the axiom that two points determine one and only one straight line applies to these geodesics. Two points on the pseudosphere (C and D in fig. 81) determine one and only one geodesic, or shortest path, between them. In like manner, the parallel axiom of Lobatchevsky and Bolyai, which says that through a point P not on a line L there is an infinite number of lines that do not meet L, applies to the geodesics on the pseudosphere.

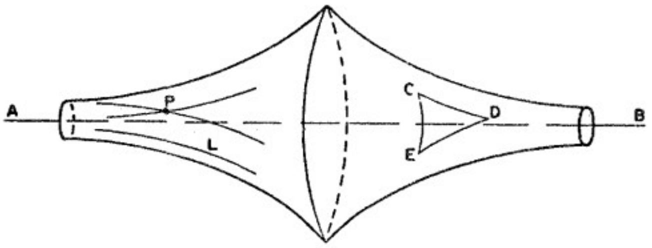

Figure 81. Pictorial interpretation of Lobatchevsky's and Bolyai's non-Euclidean geometry

Because the axioms of Lobatchevsky and Bolyai fit the geodesics on the pseudosphere, the theorems, as logical consequences of the axioms, must also apply. Thus the theorem that the sum of the angles of a triangle is less than 180° holds for triangle CDE formed by the arcs of geodesics. We have therefore obtained a visualization of the non-Euclidean geometry at the cost merely of a slight and justifiable change in the picture of the straight line.

Having made 'sense' of the new geometry let us return to our original question. Can the new geometry be descriptive of the physical world in which we live? The answer, as the reader may have anticipated, is that the geometry of physical space depends on the physical meaning we attach to the concept of straight lines. Experience tells us that if straight line is taken to be a stretched string, Euclidean geometry applies very well. It is neither necessary nor desirable, however, to allow straight line to mean a stretched string in all physical applications. Let us consider for a moment people who live in mountainous country and who are interested in the geometry of the surface of their country. The most useful physical interpretation of the straight line for them is the geodesic, that is, the curve of shortest distance between two points. The first surprising fact about these 'straight lines' is that they change shape from one part of the country to another, depending on the shapes of the hills and valleys. What axioms do these 'straight lines' obey? Almost surely not the Euclidean ones. For example, the topography of the area may be such that there are several shortest paths between some pairs of points. There may be many geodesics through a given point that do not meet a particular geodesic; and so forth.

In astronomical measurements, too, the stretched string is not the practical interpretation of the straight line. Here the light ray must serve instead. And what geometry fits best when light rays are used as straight lines? We shall leave this question with the reader until the next chapter. Meanwhile, we had better return to the mathematical account of non-Euclidean geometry. There are more mathematical worlds to be examined.

Lobatchevsky and Bolyai concentrated on Euclid's parallel axiom but accepted the other axiom in Euclid that is almost as questionable, namely, the axiom that says a line segment can be extended indefinitely in either direction. Here, too, is an axiom purporting to describe what happens in space trillions of miles from our Earth. How can we be sure of its truth, that is, its applicability to the physical world?

Not long after the work of Bolyai and Lobatchevsky on the concept of parallelism, the piercing glances of mathematicians lighted on the infiniteness of the straight line and sought to determine the wisdom of this axiom. The sickly, precocious Bernhard Riemann (1826-66), who had to beg his father, a German Lutheran pastor, for permission to abandon his training for the ministry so that he might study mathematics, undertook to pursue possible alternatives to this axiom.

One of his novel thoughts was that we must distinguish between endlessness and infiniteness. For example, the equator on the Earth is endless but finite. In view of this distinction Riemann proposed an alternative to Euclid's axiom on the infiniteness of the straight line, namely, the axiom that all lines are finite in length but endless.

This thought was then followed by reflections on the parallel axiom similar to those of Lobatchevsky and Bolyai but leading in this case to a different conclusion. As *R* moves to the left along *L* (fig. 82) and as *Q* moves to the right, both points must ultimately meet, for Riemann supposed the line *L* to be finite. The line *PR* will, as a result, rotate around *P* into *PQ* without ever losing contact with *L*. That is, there should be *no* line through *P* parallel to *L*. Figure 82 does not tell us how this complete rotation of *PR* about *P* is possible with our usual conception of the straight line, but the drawing is not intended to do more than suggest Riemann's thought. These reflections suggested to Riemann that he adopt along with the finiteness of the straight line an axiom to the effect that there are no parallel lines.

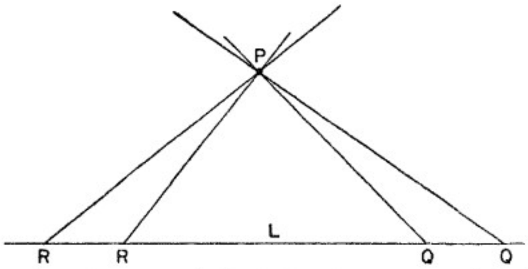

Figure 82. The geometrical basis for Riemann's parallel axiom

As if two radical departures from Euclid were not enough, Riemann proposed a third one: instead of requiring that two points determine one and only one line, Riemann adopted the axiom that two points may determine more than one line.

Before proceeding we remind the reader that these axioms are to be accepted for the moment purely as the basis for a logical development of a new geometry. The relation between this rather arbitrary system and the real world will be considered later.

The geometry of Riemann, like that of Lobatchevsky and Bolyai, has some theorems in common with those of Euclid. The theorem that vertical angles are equal and the theorem that angles opposite equal sides of a triangle are also equal hold in all three geometries, because these theorems depend only on axioms that are common to the three geometries.

Some of the theorems in Riemann's geometry that differ from Euclid's are striking. For example: *All the perpendiculars to a straight line meet in a point* (fig. 83). Another fact in this strange new world is that *two straight lines enclose an area* (fig. 84). As in the geometry of Lobatchevsky and Bolyai we find that *triangles which are similar are also congruent.* Two other theorems might almost be anticipated. The first states that *the sum of the angles of any triangle is greater than 180°.* The second states that *of two triangles the one with larger area has the greater angle sum.*

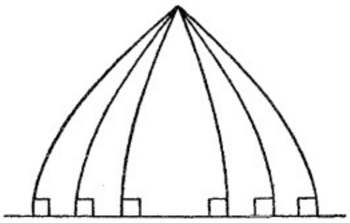

Figure 83. All the perpendiculars to a line meet in one point

Figure 84. Two straight lines enclose an area

We may now raise the same question that we raised with regard to the geometry of Lobatchevsky and Bolyai. Has Riemannian geometry any possible significance beyond that of an intellectual exercise for the mathematicians? Here again the answer is yes. It is possible to apply Riemann's geometry to the physical world with the usual understanding of straight line and never detect any differences between the assertions of the geometry and the physical situation. The argument here is precisely the one that was made in connection with the geometry of Lobatchevsky and Bolyai.

Moreover, by changing the picture of the straight line we may find other, intuitively satisfying interpretations of Riemann's geometry. Just as we were able to picture Euclidean geometry on a cylindrical surface and Lobatchevsky's and Bolyai's geometry on a pseudosphere, so we may picture Riemann's geometry on the familiar sphere. The curve that connects two points on a sphere by the shortest route—that is, the curve that will be our picture of the straight line—is the arc of the great circle through the two points. By a great circle we mean one whose center is also the center of the sphere. Thus of the two circles through *A* and *B* (fig. 85), the circle *ABCDE* is the great circle whereas the circle *ABFGH* is not.

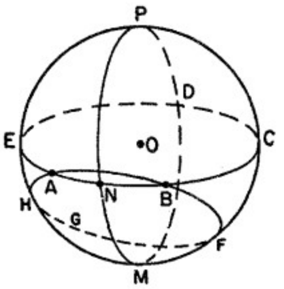

Figure 85. A pictorial interpretation of Riemann's geometry

Let us see whether the axioms of Riemann's geometry apply to the sphere, provided we interpret straight line of the axioms to mean great circle on the sphere. In the first place, great circles are endless and finite in length. Second, there are no parallel lines on the sphere, for any two great circles meet. In fact, they meet not once but twice. For example, the great circles *ABCDE* and *MNPD* meet at *N* and *D*. The axiom that two points may determine more than one line is also fulfilled on the sphere. Two points such as *N* and *D* in figure 85 have more than one great circle through them, while through two points such as *A* and *B* only one great circle passes.

Since the axioms of Riemann's geometry correctly describe facts about the sphere, the theorems, which are derived by valid deductive reasoning, must also be true on the sphere. Let us check a few. One theorem states that all the perpendiculars to a straight line meet in a point. Taking the great circle *L* of figure 86 as our straight line, we find that all the perpendiculars to *L* meet at *P*. If, for example, *L* were the equator on the Earth, *P* would be the North or South Pole.

Another theorem states that the sum of the angles of a triangle is greater than 180°. Since the straight lines of our axioms are great circles, a triangle is the figure formed by arcs of great circles. Such a triangle is illustrated by *ABP* in figure 86. Since two of the angles of this triangle are right angles, the sum of the three angles is necessarily greater than 180°. This fact is true of every 'triangle' on a sphere.

It is not necessary to labor an already obvious point. *Every theorem of Riemann's geometry can be interpreted on the sphere merely by thinking of the straight lines in the theorems as great circles on the sphere.* Hence we can give geometrical and intuitively satisfying meaning to Riemann's geometry. More than that, this geometry supplies exact answers to practical and scientific problems involving geometrical relationships *on the surface* of the sphere. Hence it is, certainly to that extent, a geometry of the physical world. In fact, every argument for the theory that our physical world could be non-Euclidean in the sense of Lobatchevsky's and Bolyai's geometry applies equally well to Riemann's geometry. The applicability of non-Euclidean geometries to the world in which we live will be discussed further under the subject of relativity.

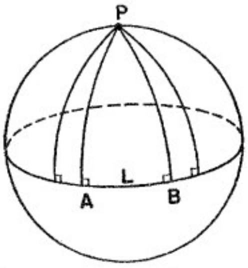

Figure 86. All the perpendiculars to a great circle on a sphere meet in one point

In retrospect, the history of the creation of non-Euclidean geometries is the history of the blindness of human beings, great and small. Man lives on the surface of the Earth. Suppose he were to set about constructing a geometry to fit *this surface directly* instead of regarding it as a special surface in the three-dimensional Euclidean world. What kind of geometry would he develop? The 'line' in this geometry for the surface of a sphere should obviously be the curve that joins any two points by the shortest route, for this curve would be the most useful. This curve, as we have seen, is the great circle joining the points. On the other hand, the straight line in the familiar sense of Euclidean geometry would certainly not be chosen as the basic curve, for it does not even exist *on* the surface of a sphere.

What axioms would a geometer choose for his great circles? Why, none other than those Riemann selected, a system of axioms in which no parallel lines exist and in which the line is finite in length. In other words, the natural geometry, the practical geometry, the common-sense geometry for us earth-bound mortals is Riemannian geometry.

For thousands of years this geometry has been right under the feet of man. Yet during all those years the greatest mathematicians never once sought to test their attack on the parallel axiom by checking with the geometry of the sphere. And as a climax to these thousands of years the great Kant built his profound philosophy on the incontrovertible truth of Euclidean geometry, indeed on the impossibility of conceiving of any other geometry. Yet all this time he was living *on*, if not in, a non-Euclidean world.

How is it, then, that though geometry arose from measurements made on the Earth, Euclidean geometry was developed first? The answer is that to human beings living in a very limited region, the Earth does indeed appear flat, and the shortest distance on a flat surface is indeed the straight line in the commonly accepted sense of the term. With this stretched-string picture of the line the axioms and theorems of Euclidean geometry followed naturally. Once the geometry of flat surfaces was developed, the sphere had to be introduced within the framework of Euclidean geometry. No one, not even the Greeks, who were especially fond of the sphere, thought to approach the geometry of the sphere through a set of axioms designed to fit such a surface directly. This history shows that men are ruled as much by habits of thought as by physical habits, social customs, and conventions. Surely the unsuccessful precursors of Lobatchevsky and Bolyai did not lack technical skill or the ability to master difficult mathematics. They failed to solve the problem of the parallel axiom only because they were unable to break a habit of thought—Euclidean geometry. The history of this mental inertia is an excellent example of what Lecky, in his *History of Rationalism in Europe*, has described as that spirit, that *Zeitgeist* of an age, which predisposes people to points of view or beliefs independent of arguments for or against. So it was with Kant and all the mathematicians up to 1800. The belief in the truth, the unassailableness, the uniqueness of Euclidean geometry precluded anyone from even considering the possibility of another geometry, even though a non-Euclidean geometry lay right before them.

The importance of non-Euclidean geometry in the general history of thought cannot be exaggerated. Like Copernicus' heliocentric theory, Newton's law of gravitation, and Darwin's theory of evolution, non-Euclidean geometry has radically affected science, philosophy, and religion. It is fair to say that no more cataclysmic event has ever taken place in the history of all thought.

First, the creation of non-Euclidean geometry brought into clear light a distinction that had always been implicit but never recognized—the distinction between a mathematical space and physical space. The original identification of the two was due to a misunderstanding. Fleeting visitors to our minds, sensations of sight and touch, suggested that the axioms of Euclidean geometry were true of physical space. The theorems deduced from these axioms were checked with further sensations of sight and touch and, behold, they checked perfectly—at least as far as these sensations could reveal. Euclidean geometry was therefore held to be an exact description of physical space. This habit of thought became so well established over hundreds of years that the very notion of a new geometry failed to make sense. Geometry meant the geometry of physical space and that geometry was Euclid's. With the creation of non-Euclidean geometry, however, mathematicians, scientists, and laymen were ultimately compelled to appreciate the fact that *systems of thought based on statements about physical space are different from that physical space.*

This distinction is vital to an understanding of the developments in mathematics and science since 1880. We must say now that a mathematical space takes on the nature of a scientific theory. It is applied to the study of physical space as long as it fits the facts of experience and serves the needs of science. However, if one mathematical space can be replaced by another in closer agreement with the expanding results of scientific work, then it will be replaced just as the Ptolemaic theory of the motion of the heavenly bodies was replaced by the Copernican theory. Nor should the reader be surprised if he discovers that this possibility has materialized by the time he reaches the next chapter.

We should regard any theory about physical space, then, as a purely subjective construction and not impute to it objective reality. Man constructs a geometry, Euclidean or non-Euclidean, and decides to view space in those terms. The advantages in doing so, even though he cannot be sure space possesses any of the characteristics of the structure he has built up in his own mind, are that he can then think about space and use his theory in scientific work. This view of space and of nature generally does not deny that there is such a thing as an objective physical world. It merely recognizes the fact that man's judgments and conclusions about space are purely of his own making.

The creation of non-Euclidean geometry cut a devastating swath through the realm of truth. Like religion in ancient societies mathematics occupied a revered and unchallenged position in Western thought. In the temple of mathematics reposed all truth, and Euclid was its high priest. But the cult, its high priest, and all its attendants were stripped of divine sanction by the work of the unholy three: Bolyai, Lobatchevsky, and Riemann. It is true that in undertaking their research these audacious intellects had in mind only the logical problem of investigating the consequences of a new parallel axiom. It certainly did not occur to them at the outset that they were challenging Truth itself. And as long as their work was regarded merely as an ingenious bit of mathematical hocus-pocus, no serious questions were raised. The moment men realized, however, that the non-Euclidean geometries could be valid descriptions of physical space, an inescapable problem presented itself. How could mathematics, which had always claimed to present the truth about quantity and space, now offer several contradictory geometries? No more than one of these could be the truth. Indeed, and even more disturbing, perhaps the truth was different from all these geometries. The creation of the new geometries, therefore, forced recognition of the fact that there could be an 'if' about all mathematical axioms. *If* the axioms of Euclidean geometry are truths about the physical world then the theorems are. But, unfortunately, we cannot decide on *a priori* grounds that the axioms of Euclid, or of any other geometry, are truths.

In depriving mathematics of its status as a collection of truths, the creation of non-Euclidean geometries robbed man of his most re-calculation suggests a general statement, namely, that the sum of the first n odd numbers, where n is any positive integer, is the square of n. Of course this possible theorem is not proved by the calculations above. Nor could it ever be proved by such calculations, for no mortal man could make the infinite set of calculations that would be required to establish the conclusion for *every* n. The calculations do, however, give the mathematician something to work on.

Consider another instance of generalization as a source of suggestions for theorems. A triangle is a polygon of three sides. Now in Euclidean geometry the sum of the angles of a triangle is 180°. Is it not natural to ask whether any general theorem could be found about the sum of the angles of any polygon? This question is answered by a very old theorem. The sum of the angles of any polygon is found by subtracting two from the number of sides and then multiplying the result by 180°.

We have already seen how the purely logical problem of deducing the assertion contained in Euclid's parallel axiom from more acceptable axioms led to non-Euclidean geometry. Once the idea of such geometries was grasped numerous suggestions for theorems were obtained by seeking the analogues of theorems that held in Euclidean geometry. For example, what is the analogue of the theorem that the sum of the angles of a quadrilateral is 360°?

These few indications of how the mathematician secures suggestions for theorems do not tell the whole story. Even if we add the more fortuitous sources such as pure chance, guesswork, and blundering about until a theorem is found, we still have left out the most valuable source of possible theorems—the imagination, intuition, and insight of creative genius. Most people could look at a quadrilateral indefinitely without becoming aware that if the midpoints of the four sides are joined (fig. 88), the figure formed is a parallelogram. Such knowledge is not the product of logic but of a sudden flash of insight.

In the domains of algebra, calculus, and advanced analysis especially, the first-rate mathematician depends on the kind of inspiration that we usually associate with the composer of music. The composer feels that he has a theme, a phrase which, when properly developed and embroidered, will produce beautiful music. Experience and a knowledge of music aid him in developing it. Similarly the mathematician divines that he has a conclusion which will follow from the

(%%B

Pag. 410(439) XXVI

New Geometries, New Worlds.

Pag. 412 (441) - Saccheri

Euclid's parallel axiom must be right.

Book: Euclid Vindicated from All Defects. (1733)

"Apparently one man set out to vindicate another he will most likely do so regardless of the facts".

Gauss, Lobatchevsky and Bolyai (Pag. 413 (442)).

Undoubtedly this (this) change accounts for the fact that three men, Gauss, Lobatchevsky and Bolyai] with no knowledge of each other's thoughts, discovered the correct interpretation of Saccheri's work at about the same time

Pag. 413(442)

Gauss was extremely modest. He said,

"If others would but reflect on mathematical truths as deeply and as continuously as I have, they would make my discoveries".

Those who believe that genius is 99 per cent perspiration as well as those who despair of their mathematical abilities may find comfort in Gauss's statement.

Gauss was still a youth when the problem of the parallel axiom first came to his attention.

Pag. 414(443)

Gauss's work on non-Euclidean geometry was found among his papers after his death.

Pag. 418(447) Gauss suggested a criterion for determining the truth of Euclidean Geometry.

The sum of the angles of a triangle equals 180° in this geometry but is less than 180° in the new geometry.

_ Morris Kline.. Nonmathematician (1985).pdf

Pág. 456(473) Gauss

He pursued the logical implications of a system of axioms which included the assumption that more than one parallel to a given line passed through a given point, and thus created non-Euclidean geometry.

Pág. 542(559)

Some of the axioms of Euclidean geometry are that two points determine a line, all right angles are equal, and the axiom on parallel lines.

The non-Euclidean geometry of Gauss, Lobatchevsky, and Bolyai contains the same axioms as Euclidean geometry does, except for the parallel axiom.

Look: T2 /// (estudar!!!).

_ 20211012/Mathematics.../ Geometry of Vector Sheaves associative, commutative, and distributive properties and the axioms about equalities. Some of the axioms of Euclidean geometry are that two points determine a line, all right angles are equal, and the axiom on parallel lines. The non-Euclidean geometry of Gauss, Lobatchevsky, and Bolyai contains the same axioms as Euclidean geometry does, except for the parallel axiom. From the concepts and axioms, theorems are deduced. Hence, from the standpoint of structure, the concepts, axioms, and theorems are the essential components. We shall discuss these in turn.

The basic concepts of the elementary branches of mathematics are abstractions from experience. Whole numbers and fractions were certainly suggested by obvious physical counterparts. But it is noteworthy that many more concepts are introduced which are, in essence, creations of the human mind with or without partial help from experience. Irrational numbers, such as $\sqrt{2}$, were forced upon mathematicians to represent all the lengths occurring in Euclidean geometry, for example, the length of the hypotenuse of a right triangle whose arms are both one unit long. The notion of a negative number, though perhaps suggested by the need to distinguish debits from credits, is nevertheless not wholly derived from experience, for the mind had to conceive of the notion of an entirely new type of number to which operations such as addition, multiplication, and the like can be applied. The notion of a variable to represent the quantitative values of some changing physical phenomenon, such as temperature or time, is also at least one mental step beyond the mere observation of change. The concept of a function, or a relationship between variables, is almost entirely a mental creation. The farther one proceeds with the mathematics of number, the more remote from experience are the concepts introduced and the larger is the creative role played by the mind. The derivative of a function, for example, the instantaneous rate of change of distance compared with time, is a wholly man-made and, one might also say, an ingenious construction. We do not experience instants or instantaneous velocities but, rather, small intervals of time and speeds over small intervals of time.

The gradual introduction of new concepts which more and more depart from forms of experience finds its parallel in geometry. Though point, line, triangle, circle, and a few other elementary concepts are no more than abstractions from experience, this is not true of most of the curves which geometry considers. The approach to the conic sections as sections of a cone made by a plane or by the locus definitions (Chapter 6) subsequently used to define parabola, ellipse, and hyperbola was conceived by the mind. The notions of projection and section of projective geometry and many of the specific concepts of projective geometry such as cross ratio are entirely mental creations.

This brief review of the origin of mathematical concepts may serve to emphasize several major facts. The first of these is growth. As mathematicians continue to work in any given branch, they discover new concepts which are worth introducing and developing. Secondly, as their work advances, the new concepts are less and less drawn from experience and more and more from the recesses of human minds. Moreover, the development of concepts is progressive, later concepts being built on earlier ones. These facts have unfortunate consequences. Because the more advanced ideas are purely mental creations rather than abstractions from immediate experience and because they are defined in terms of prior concepts, it is more difficult to comprehend them. At least, one cannot usually find simple and familiar physical pictures or experiences to illustrate their meanings.

Axioms constitute the second major component of any branch of mathematics. The axioms of the mathematics of number and of Euclidean geometry were suggested by experience. Up to the introduction of non-Euclidean geometry, the idea that man chose axioms suggested by experience would have been regarded as absurd. Axioms were understood to be basic, self-evident truths about the concepts involved. While minor variations in the choice and wording were tolerated, the universal belief was that man had to accept what was clearly true. These truths were supposed to be written into the universe and were considered to be inescapable. Some philosophers, such as Plato and Descartes, believed that these truths were already planted in our minds by God. We know now that this view must be discarded. The seemingly self-evident nature of the axioms of the mathematics of number and of Euclidean geometry is really the consequence of limited experience and relatively superficial observation. In the simpler uses of numbers and in the limited regions of the universe accessible to man until recent times, the well-known axioms of number and Euclidean geometry seemed inescapably true. But we can choose other axioms and produce other geometries and even other number systems. Though we can make arbitrary choices, we do not do so because we wish our new axioms to yield systems of mathematics as significant and as useful as the older systems. Were we able to continue our study of mathematics proper, we would find that there are many other systems of axioms which lead to important branches of mathematics.

The fruit of mathematical activity consists of the theorems deduced from a set of axioms. The theorems offer new knowledge by no means immediately discernible in the axioms. Whether or not the reader may prize the knowledge, he learns by pursuing the implications of the Euclidean axioms that the sum of the angles of a triangle is 180° and that the area of a circle is π times the square of the radius. The amount of information that can be deduced from some sets of axioms is almost incredible. Euclid deduced about 500 theorems from his set. The axioms of number give rise to the results of algebra, properties of functions, the theorems of the calculus, the solutions of various types of differential equations, and many other results we have not surveyed.

Mathematical theorems, as we know, must be deductively established. Since observation, measurement, induction, and a variety of other methods of obtaining knowledge are available and are used in all other scientific pursuits, trying to replace Euclid's parallel axiom by a more acceptable one, that is, one which did not involve any assertion about what must happen far off in space. But he did not succeed. He also appreciated at the age of 15, as he told his friend Schumacher, that one could not hope to prove an assertion about parallel lines on the basis of the other nine axioms. That is, he realized that these nine axioms did not in themselves dictate the form of the parallel axiom. Gauss was too brilliant a man to overlook the implication of this fact. If there was some freedom in the choice of a parallel axiom, then one might choose an axiom different from Euclid's and build a new kind of geometry. Gauss did just this. He pursued the logical implications of a system of axioms which included the assumption that *more than one* parallel to a given line passed through a given point, and thus created non-Euclidean geometry. (Gauss himself finally adopted the term non-Euclidean after having called his system anti-Euclidean geometry, and later astral geometry.)

Though Gauss realized that such a geometry could conceivably apply to physical space and hence was indeed significant, he did not publish his results. Gauss was far ahead of his times in concluding that Euclidean geometry was not necessarily the correct description of physical space and that some non-Euclidean geometry might prove as accurate. Hence he feared that he would be laughed at. In a letter written in 1829 to his friend, the mathematician Friedrich Wilhelm Bessel, Gauss confessed that he feared the clamor of the Boeotians, a figurative reference to the dull-witted, for the Boeotians had been one of the more simple-minded Greek tribes. Gauss's work on non-Euclidean geometry was found among his papers after his death in 1855.

The men who usually receive credit for creating non-Euclidean geometry because they published their results are Nicholas I. Lobachevsky (1793–1856) and John Bolyai (1802–1860). Lobachevsky, born to a poor family and displaying brightness even as a youngster, attended the University of Kazan in Russia. One of his teachers was the German mathematician J. M. C. Bartels, a friend of Gauss, and it is very likely that the problem of the parallel axiom was called to Lobachevsky's attention by Bartels. Lobachevsky became a professor at Kazan at the age of 23 and continued to work on the problem. By 1823 he realized that there can be other geometries and that Euclidean geometry need not be the correct description of physical space. He said later that two thousand years of fruitless attempts to put the parallel axiom on an unquestionable basis had led him to suspect that it could not be done. From 1829 on Lobachevsky published books and papers in which he expounded the theorems that hold in his non-Euclidean geometry. Despite valuable services to mathematics, his university, and the Russian government, he was dismissed in 1846, but continued to work in his field until his death.

Bolyai, a Hungarian, was an Austrian army officer. He learned mathematics from his father, Wolfgang Bolyai, who set his son thinking about the parallel axiom. In 1823 John arrived at the same conclusion that had been

D -256- - -* ! - 7 3 !A 2
2L ! 211 ! 4 ! 7 3 & !A 7 ! 9 3
1 // 1"
D -256- - -* ! - 7 3 !A 2
2L ! 211 ! 4 ! 7 3 && 7 3 =L! 1 ! " 211 !
1 8 1"
($ 5 ! ! 2 5 ! ! A ! ••• 1 !

for instance, the *Fundamental Lemma of Riemannian Geometry,* pertaining here to the *existence of* an *$\mathcal{A}$-connection compatible with a* given *Riemannian $\mathcal{A}$-metric,* for every *Riemannian vector sheaf,* when the corresponding *differential triad,* as in (i) above, is *suitably adapted* to the present (riemannian) framework.

On the other hand, by contrast with the previous situation, concerning the existence of an $\mathcal{A}$-connection, one realizes that, *given an $\mathcal{A}$-connection, we can always associate with it* what we might consider here as the analogue of the classical notion of *curvature.* Thus, the study of this concept, together with the deduction of several standard results, within the present axiomatic framework, connected with the latter notion, constitute the subject matter of Chapter VIII. In this concern, we further note that, in order to formulate the same notion, one has to supplement the initially given differential triad, as in (i) above, with what we call a *curvature datum*

$$(viii) \qquad (\mathcal{A}, \partial, \Omega^1, d, \Omega^2).$$

Thus, in effect, by another *"differential operator"*

$$(ix) \qquad d \equiv d^1 : \Omega \equiv \Omega^1 \longrightarrow \Omega^2 \equiv \wedge^2\Omega^1,$$

satisfying the appropriate *compatibility conditions,* with respect to ∂. We also speak then of a *curvature space* X. So the curvature of a given $\mathcal{A}$-connection D is now defined by the formula

$$(x) \qquad R(D) \equiv D^1 \circ D,$$

where D^1 is the *1st prolongation of* D, by analogy with the classical case, the previous operator $d^1 \equiv d$ being thus the *1st prolongation of* ∂. So one of the compatibility conditions between d and ∂, referred to above, is the assumption that

$$(xi) \qquad R(\partial) \equiv d^1 \circ \partial \equiv d \circ \partial = 0,$$

that is, the *flatness of the* given *$\mathcal{A}$-connection* ∂ of $\mathcal{A}$ (in the strict sense of the word, or else that the *curvature of* ∂ is *zero*). We also examine *flat $\mathcal{A}$-connections,* in general, in the sense that the corresponding curvature, as in (x) above, vanishes. The notion of *flatness* is further related with the analogues here of other classical concepts, as, for example, with those of *complete parallelism, integrability* and *triviality* of the corresponding $\mathcal{A}$-connections; yet, with the *flatness of* a given *vector sheaf* E, relative to the so-called *co-*

(%%/

data

..1(1998).djvu

38 occurrences - Riemann

- Geometry of Vector Sheaves .. II.pdf

65 occurrences - Riemann.

Pág. 15 - Riemannian A-metric, for every Riemannian vector sheaf, when the corresponding differential triad.

- Morris Kline - Storia del pensiero matematico - I (1999).pdf

Pág. 333(361) - 14

La origini della geometria proiettiva

- Morris Kline ... II (1999).pdf

Pág. 1004 (291) XXXVI

~~La~~ La geometria non euclidea

Pág. 1008(295) Figura 36.2

Imagem:

! -256- - -* ! - + ! " 1
! ! 4! 02 S ! & = ! " /// 1"
(%%% %B ' ?! " ! ! 1 A!

Capitolo quattordicesimo

Le origini della geometria proiettiva

> Devo confessare apertamente che non ho mai avuto gusto per lo studio e per la ricerca né in geometria né in fisica se non in quanto esse potevano servirmi come mezzi per arrivare a una qualche sorta di conoscenza delle cause prossime... per il bene e la comodità della vita, nel mantenere la salute, nella pratica di qualche arte... avendo osservato che buona parte delle arti si basa sulla geometria, come fra le altre il taglio delle pietre in architettura, l'arte delle meridiane e in particolare l'arte della prospettiva.
>
> GIRARD DESARGUES

1. *La rinascita della geometria.*

In geometria il risorgere di un'attività creativa significativa venne in ritardo rispetto all'algebra. A parte la creazione del sistema matematico della prospettiva e le ricerche geometriche incidentali degli artisti rinascimentali (cfr. cap. XII, § 2), pochissimi risultati degni di nota vennero ottenuti in geometria dai tempi di Pappo fino al 1600 circa. Un certo interesse fu suscitato dalla pubblicazione di numerose edizioni a stampa delle *Sezioni coniche* di Apollonio, in particolare della notevole traduzione latina dei libri I-IV di Federigo Commandino (1509-75) apparsa nel 1566. Altri tradussero i libri V-VII e un certo numero di studiosi, compresi Vieta, Willebrord Snell (1580-1626) e Ghetaldi, intrapresero il lavoro di ricostruzione dell'ottavo libro che era andato perduto.

Quelli che erano necessari, e che in effetti si presentarono, per dirigere le menti dei matematici entro nuovi canali erano dei problemi nuovi. Uno di questi era già stato sollevato da Alberti: quali proprietà geometriche hanno in comune due sezioni della stessa proiezione di una figura reale? Un gran numero di problemi venne dalla scienza e dalle necessità pratiche. L'uso fatto da Kepler delle sezioni coniche nella sua opera del 1609 diede un enorme impulso al riesame di queste curve e alla ricerca di loro proprietà utili per l'astronomia. L'ottica, che aveva destato l'interesse dei matematici fin dai tempi dei Greci, ricevette un'attenzione molto piú viva dopo l'invenzione del telescopio e del microscopio avvenuta all'inizio del Seicento. La progettazione delle lenti di questi strumenti divenne un problema fondamentale; esso implicava un aumento d'interesse per le forme delle superfici o, poiché si trattava di superfici di rotazione, per le curve che le generano. Le esplorazioni geografiche avevano creato un gran bisogno di carte e un interesse per lo studio delle rotte delle navi quali sono rappresentate sulla sfera e sulla carta. L'introduzione del concetto del moto della Terra richiedeva nuovi principî di meccanica che dessero conto delle traiettorie degli oggetti mobili e anche questo implicava lo studio di curve. Fra gli oggetti mobili i proiettili diventarono sempre piú importanti perché i cannoni potevano ora lanciare le loro palle a centinaia di metri di distanza ed era vitale essere in grado di predirne la traiettoria e la gittata. Il problema pratico del calcolo delle aree e dei volumi incominciò ad attrarre un'attenzione sempre maggiore. La *Nova stereometria doliorum vinariorum* (1615) di Kepler diede inizio a una nuova esplosione di attività in questo campo.

Un altro tipo di problemi si presentò come conseguenza dell'assimilazione delle opere greche. I matematici cominciarono a rendersi conto che i metodi di dimostrazione greci mancavano di generalità, in quanto bisognava escogitare un metodo speciale quasi per ogni problema. Questa osservazione era stata fatta da Agrippa di Nettesheim (1486-1532) fin dal 1527, e da Maurolico, che aveva tradotto delle opere greche e aveva scritto libri sulle sezioni coniche e su altri argomenti matematici.

Molte delle risposte date ai nuovi problemi si ridussero a variazioni minime su vecchi temi. Venne cambiato l'approccio alle sezioni coniche. Le curve erano definite fin dall'inizio come luoghi piani invece che come sezioni di un cono, come avveniva in Apollonio. Guidobaldo Del Monte, ad esempio, nel 1579 definí l'ellisse come il luogo dei punti tali che la somma delle loro distanze dai fuochi sia costante. Vennero ristudiate non soltanto le coniche ma anche altre antiche curve greche quali la concoide di Nicomede, la cissoide di Diocle, la spirale di Archimede e la quadratrice di Ippia. Vennero anche introdotte nuove curve, in particolare la cicloide (cfr. cap. XVII, § 2). Mentre tutte queste ricerche risultarono utili per diffondere i contributi greci, nessuna di esse forní nuovi teoremi o nuovi metodi di dimostrazione. La prima innovazione portatrice di importanti conseguenze venne in risposta ai problemi sollevati dai pittori.

2. *I problemi sollevati dalle ricerche sulla prospettiva.*

L'idea fondamentale nel sistema di prospettiva focale creato dai pittori è il principio di proiezione e sezione (cfr. cap. XII, § 1). Una scena

D -256- - -* ! - + ! " 1
! ! 4! ! && = ! " /// 1"
(' / FFF &

Capitolo trentaseiesimo

La geometria non euclidea

> … poiché sembra essere vero che molte cose hanno, per cosí dire, un'epoca in cui vengono scoperte simultaneamente in molti luoghi, proprio come le viole sbocciano da tutte le parti in primavera.
>
> WOLFGANG BOLYAI

> Il carattere di necessità ascritto alle verità matematiche e anche la peculiare certezza ad esse attribuita sono un'illusione.
>
> JOHN STUART MILL

1. *Introduzione.*

Fra tutte le creazioni tecnicamente complesse della matematica dell'Ottocento, la piú profonda, la geometria non euclidea, fu anche la piú semplice dal punto di vista tecnico. Questa creazione diede origine a nuove importanti branche della matematica, ma la sua conseguenza piú significativa sta nel fatto che obbligò i matematici a rivedere radicalmente le loro idee sulla natura della matematica e sui suoi rapporti con il mondo fisico. Essa diede anche origine a problemi nei fondamenti della matematica con i quali sta ancora lottando il xx secolo. Come vedremo, la geometria non euclidea fu il culmine di una lunga serie di sforzi nel campo della geometria euclidea. I frutti di questo lavoro furono colti all'inizio del xix secolo, nel corso degli stessi decenni in cui rinasceva e si estendeva la geometria proiettiva. Tuttavia, le due discipline non erano in quel momento collegate.

2. *Lo stato della geometria euclidea intorno al 1800.*

Nonostante i Greci avessero riconosciuto che lo spazio matematico astratto è distinto dalle percezioni sensoriali dello spazio e Newton avesse sottolineato questo punto[1], tutti i matematici fino al 1800 circa erano convinti che la geometria euclidea fosse l'idealizzazione corretta delle proprietà dello spazio fisico e delle figure di questo spazio. In effetti, come abbiamo già notato, vi furono molti tentativi per costruire l'aritmetica, l'algebra e l'analisi, la cui fondazione logica era oscura, sulla geometria euclidea, garantendo cosí anche la verità di queste discipline.

---

[1] I. NEWTON, *Principia*, libro I, def. 8, scolio.

Molti matematici proclamarono ad alta voce la loro fiducia nella verità della geometria euclidea. Isaac Barrow, ad esempio, che costruí la sua matematica, compreso il calcolo infinitesimale, sulla geometria, elenca otto ragioni in favore della certezza della geometria: la chiarezza dei concetti, le definizioni non ambigue, la sicurezza intuitiva e la verità universale degli assiomi, la chiara possibilità e la facile immaginabilità dei suoi postulati, il piccolo numero dei suoi assiomi, la chiara concepibilità del modo in cui sono generate le grandezze, il facile ordine delle dimostrazioni e l'assenza di cose non note.

Barrow si chiese: come possiamo essere sicuri che i principî geometrici si applicano alla natura? La sua risposta fu che essi sono derivati dalla ragione innata e gli oggetti sensibili sono dei semplici agenti che li risvegliano in noi. Inoltre, i principî della geometria erano stati confermati costantemente dall'esperienza e avrebbero continuato ad esserlo perché il mondo, progettato da Dio, è immutabile. La geometria è quindi la scienza perfetta e certa.

È degno di nota il fatto che i filosofi della fine del Seicento e del Settecento sollevarono pure il problema di come si potesse essere certi che il corpo di conoscenze piú ampio prodotto dalla scienza newtoniana fosse vero. Quasi tutti, e in particolare Hobbes, Locke e Leibniz, risposero che le leggi matematiche, come la geometria euclidea, erano inerenti al disegno dell'universo. Leibniz lasciò qualche spazio al dubbio quando tracciò la distinzione fra mondi attuali e mondi possibili, ma l'unica eccezione significativa fu David Hume, che nel *Treatise of Human Nature* (1739) negò l'esistenza di leggi o di successioni necessarie di eventi nell'universo, sostenendo che dall'osservazione di queste successioni gli esseri umani erano indotti a concludere che esse si sarebbero sempre ripetute nello stesso modo. La scienza è quindi puramente empirica. In particolare, le leggi della geometria euclidea non sono verità fisiche necessarie.

L'influenza di Hume fu controbattuta e in effetti soppiantata da Immanuel Kant. La risposta data da Kant, nella *Critica della ragion pura* (1781), al problema di come si può essere sicuri che la geometria euclidea si applica al mondo fisico è molto originale. Egli sosteneva che le nostre menti forniscono certi modi di organizzazione (che chiamava intuizioni) dello spazio e del tempo e che l'esperienza viene assorbita e organizzata dalle nostre menti in accordo con queste intuizioni. Le nostre menti sono costruite in modo tale da obbligarci a vedere il mondo esterno in un unico modo. Di conseguenza, certi principî concernenti lo spazio sono anteriori all'esperienza; questi principî e le loro conseguenze logiche, che Kant chiamava giudizi sintetici a priori, sono quelli della geometria euclidea. La natura del mondo esterno ci è nota soltanto nella maniera in cui le nostre menti ci obbligano a interpretarla. Sulle basi ora descritte, Kant affermava, e i suoi contemporanei lo accettarono, che il mondo fisico deve essere euclideo. In ogni caso, sia che si facesse appello all'esperienza basata sulle verità innate, sia che si accettasse l'opinione di Kant, la conclusione comune era l'unicità e la necessità della geometria euclidea.

3. *Le ricerche sull'assioma delle parallele.*

Sebbene la fiducia nella geometria euclidea come idealizzazione corretta dello spazio fisico sia rimasta inalterata dal 300 a.C. fino al 1800 circa, una preoccupazione tenne occupati i matematici durante quasi tutto questo lungo periodo. Gli assiomi adottati da Euclide erano considerati verità evidenti sullo spazio fisico e sulle figure di questo spazio. Tuttavia, l'assioma delle parallele, nella forma enunciata da Euclide (cfr. cap. IV, § 3), era considerato un po' troppo complicato. Nessuno ne metteva realmente in dubbio la verità, ma tuttavia ad esso mancava la irresistibile evidenza degli altri assiomi. Apparentemente anche lo stesso Euclide non era soddisfatto della sua versione dell'assioma delle parallele, perché non vi fece ricorso se non dopo aver dimostrato tutti i teoremi che si potevano provare senza di esso.

Un problema connesso, che non preoccupò altrettante persone ma che finí per rivelarsi ugualmente fondamentale, è quello di stabilire se si possa assumere l'esistenza nello spazio fisico di rette infinite. Euclide aveva fatto attenzione a postulare soltanto che è possibile prolungare una retta (finita) fintanto che è necessario, cosicché anche la retta estesa continuava a essere finita. Anche la particolare formulazione da parte di Euclide dell'assioma delle parallele, e cioè che due rette si incontreranno dalla parte della trasversale in cui la somma degli angoli interni è minore di due angoli retti, era un modo per evitare l'affermazione esplicita che esistono coppie di rette che non s'incontrano mai, comunque vengano prolungate. Ciò nondimeno, quanto Euclide aveva postulato implicava l'esistenza di rette infinite perché, se fossero state finite, non avrebbero potuto essere prolungate tanto quanto fosse stato necessario in ogni contesto dato ed egli *dimostrò* l'esistenza di rette parallele.

*D 2L ! ! 1 ! { UUU

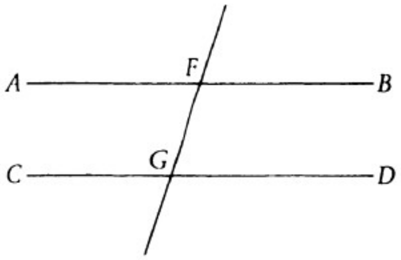

Figura 36.1.

La storia della geometria non euclidea comincia con gli sforzi per eliminare i dubbi intorno all'assioma euclideo delle parallele. A partire dall'epoca dei Greci fino all'inizio dell'Ottocento furono tentati due approcci. Il primo consisteva nel sostituire l'assioma delle parallele con un enunciato piú evidente. Il secondo consisteva nel tentare di dedurlo dagli altri nove assiomi di Euclide; se ciò fosse stato possibile, esso sarebbe diventato un teorema e tutti i dubbi sarebbero stati eliminati. Non faremo un resoconto dettagliato di tutte queste ricerche perché la loro storia è facilmente reperibile[2].

Il primo tentativo importante fu fatto da Tolomeo nel suo trattato sul postulato delle parallele. Egli tentò di dimostrarlo deducendolo dagli altri nove assiomi e dai teoremi 1-28 di Euclide che non dipendono dal postulato delle parallele. Tolomeo fece però l'assunzione inconscia che due rette non racchiudano una porzione di spazio e che, se *AB* e *CD* sono parallele (fig. 36.1), allora tutto ciò che vale per gli angoli interni da una delle parti di *FG* deve valere anche dall'altra.

Proclo, il commentatore del v secolo, fu molto esplicito nelle sue critiche al postulato delle parallele. Egli dice: «Questo postulato dovrebbe essere eliminato del tutto dai postulati perché esso è un teorema che implica molte difficoltà, che Tolomeo in un certo libro si propose di risolvere, e richiede per la sua dimostrazione un certo numero di definizioni e di teoremi; il suo inverso, inoltre, fu dimostrato come teorema da Euclide stesso». Proclo osserva che, mentre è necessario credere che due linee tendano l'una verso l'altra dalla parte della trasversale in cui la somma degli angoli interni è minore di due angoli retti, non è cosí chiaro che queste due rette si intersecheranno effettivamente in un punto finito. Questa conclusione è soltanto probabile. Infatti, continua, ci sono certe curve che si avvicinano l'una all'altra sempre di piú senza mai incontrarsi. Cosí, un'iperbole si avvicina al suo asintoto ma non l'incontra mai. Di conseguenza, non potrebbe ciò essere vero per le due rette del postulato di Euclide? Aggiunge poi che fino a un certo valore

---

[2] Si veda, ad esempio, il libro di Bonola citato nella bibliografia.

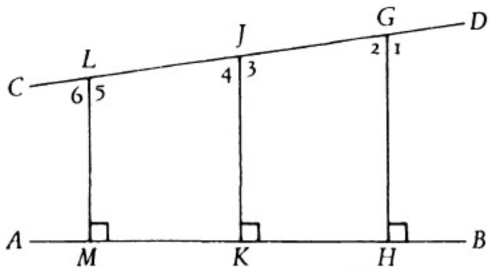

Figura 36.2.

della somma degli angoli interni da una delle parti della trasversale le due rette potrebbero effettivamente incontrarsi, mentre per un valore leggermente piú grande, ma ancora inferiore a due angoli retti, le rette potrebbero essere asintotiche.

Proclo basava la sua dimostrazione del postulato delle parallele su un assioma che Aristotele aveva usato per dimostrare che l'universo è finito. L'assioma dice: «Se da un punto si prolungano indefinitamente due rette formanti un certo angolo, le successive distanze fra le due rette finiranno per superare ogni grandezza finita». La dimostrazione di Proclo era sostanzialmente corretta, ma si limitava a sostituire un assioma discutibile con un altro altrettanto discutibile.

Nasîr-Eddin (1201-74), l'editore persiano di Euclide, diede anche lui una «dimostrazione» del postulato euclideo delle parallele assumendo che due rette non parallele si avvicinino l'una all'altra in una direzione e si allontanino nell'altra. Specificamente, se *AB* e *CD* (fig. 36.2) sono due rette tagliate da *GH*, *JK*, *LM*, ..., se queste ultime sono perpendicolari ad *AB* e se gli angoli 1, 3, 5, ... sono ottusi, mentre gli angoli 2, 4, 6, ... sono acuti, allora $GH > JK > LM$ ... Questo fatto, dice Nasîr-Eddin, lo si vede chiaramente.

Wallis fece nel 1663 alcune ricerche sull'assioma delle parallele, che pubblicò nel 1693[3]. Anzitutto riprodusse le ricerche di Nasîr-Eddin sull'assioma delle parallele che si era fatto tradurre da un professore di arabo di Oxford. Incidentalmente, fu cosí che il lavoro di Nasîr-Eddin divenne noto in Europa. Wallis criticò poi la dimostrazione di Nasîr-Eddin ed espose la sua dimostrazione dell'enunciato di Euclide. Essa si basa sull'assunzione esplicita che, per ogni triangolo, ne esiste uno simile i cui lati hanno un qualsiasi rapporto dato con i lati del triangolo di partenza. Wallis riteneva che questo assioma fosse piú evidente di quelli che si basano sulla possibilità di suddivisioni arbitrariamente piccole e di estensioni arbitrariamente grandi. In effetti, diceva, l'assioma di Euclide che afferma che si può costruire un cerchio di centro e raggio dati presuppone di avere a disposizione un raggio arbitrariamente grande. Si può perciò ben assumere che valga l'analogo per le figure rettilinee quali il triangolo.

---

[3] J. WALLIS, *Opera*, II, 669-78.

Il piú semplice degli assiomi sostitutivi fu proposto da Joseph Fenn nel 1769. Esso afferma che due rette che si intersecano non possono essere entrambe parallele a una terza. Questo assioma compare anche nel commento di Proclo alla proposizione 31 del libro I degli *Elementi* di Euclide. L'enunciato di Fenn è completamente equivalente all'assioma dato nel 1795 da John Playfair (1748-1819) e afferma che, per un dato punto *P* che non stia su una retta *l*, passa un'unica retta che giace nel piano di *P* ed *l* e che non incontra *l*. È questo l'assioma usato nei testi moderni (in cui per semplicità si dice di solito che passa «una e una sola retta...»)

Legendre lavorò sul problema del postulato delle parallele per un periodo di circa vent'anni. I suoi risultati furono pubblicati in libri e articoli, comprese le molte edizioni dei suoi *Eléments de géométrie*[4]. In uno dei suoi attacchi al problema dimostrò il postulato delle parallele assumendo l'esistenza di triangoli simili di dimensioni diverse; la sua dimostrazione era analitica, ma egli presuppose che l'unità di lunghezza non avesse alcuna influenza sulla dimostrazione. Diede poi una dimostrazione basata sull'assunzione che, dati tre punti non allineati, esiste un cerchio che passa per essi. In un altro approccio ancora si serví di tutti i postulati di Euclide ad eccezione di quello delle parallele e dimostrò che la somma degli angoli di un triangolo non può essere maggiore di due angoli retti. Osservò poi che, con queste stesse assunzioni, l'area è proporzionale al difetto, cioè alla differenza fra due angoli retti e la somma degli angoli. Cercò perciò di costruire un triangolo che avesse dimensione doppia di quella di un triangolo dato e quindi anche difetto doppio. Procedendo in questo modo sperava di ottenere dei triangoli con difetti sempre maggiori le cui somme degli angoli avrebbero perciò teso a zero. Questo risultato, pensava, sarebbe stato assurdo e perciò la somma degli angoli avrebbe dovuto essere uguale a 180°. Questo fatto avrebbe a sua volta implicato l'assioma euclideo delle parallele. Legendre trovò però che la costruzione si riduceva a dimostrare che, per un dato punto interno a un angolo minore di 60°, si può sempre tracciare una retta che incontri entrambi i lati dell'angolo e questo non era in grado di farlo senza il postulato delle parallele. In ciascuna delle dodici edizioni (l'ultima nel 1813) della sua versione degli *Elementi* di Euclide, Legendre aggiunse delle appendici che apparentemente davano delle dimostrazioni del postulato delle parallele, ma ciascuna era manchevole perché assumeva qualche cosa che non poteva esserlo o perché assumeva un assioma altrettanto discutibile di quello di Euclide.

---

[4] 1ª ed., 1794.

Nel corso delle sue ricerche Legendre[5], servendosi degli assiomi euclidei diversi da quello delle parallele, dimostrò i seguenti teoremi significativi: se la somma degli angoli di un triangolo è uguale a due angoli retti, lo stesso avviene per ogni triangolo. Analogamente, se la somma è minore di due angoli retti in un triangolo, lo stesso avviene per tutti. Egli dà poi la dimostrazione del fatto che, se la somma degli angoli di un triangolo è uguale a due angoli retti, allora vale il postulato euclideo delle parallele. Queste ricerche sulla somma degli angoli di un triangolo furono infruttuose anche perché Legendre non riuscí a dimostrare (senza l'aiuto dell'assioma delle parallele o di un assioma equivalente) che la somma degli angoli di un triangolo non può essere minore di due angoli retti.

Gli sforzi descritti finora erano soprattutto dei tentativi di trovare un sostituto dell'assioma euclideo delle parallele che fosse piú evidente, e molti degli assiomi proposti sembravano in effetti intuitivamente piú evidenti. Perciò i loro creatori erano soddisfatti e ritenevano di aver raggiunto il loro obiettivo. Tuttavia, un esame piú attento rivelava che questi assiomi sostitutivi non erano in realtà piú soddisfacenti. Alcuni di essi facevano delle affermazioni su ciò che accade infinitamente lontano nello spazio. Cosí, richiedere l'esistenza di un cerchio passante per tre qualsiasi punti non allineati implica l'esistenza di cerchi sempre piú grandi quando i tre punti tendono a essere allineati. D'altro lato, gli assiomi sostitutivi che non coinvolgevano direttamente l'«infinito», come ad esempio l'assioma che afferma che esistono due triangoli simili ma diversi, si rivelavano essere assunzioni abbastanza complesse e in nessun modo preferibili all'assioma euclideo delle parallele.

Il secondo gruppo di sforzi per risolvere il problema dell'assioma delle parallele fu rivolto a dedurre l'asserzione di Euclide dagli altri nove assiomi. La deduzione poteva essere diretta o indiretta. Tolomeo aveva tentato una dimostrazione diretta. Il metodo indiretto consiste nell'assumere una qualche asserzione che contraddice l'assioma euclideo e nel tentare di dedurre una contraddizione nel nuovo corpo di teoremi che ne conseguono. Ad esempio, poiché l'assioma euclideo delle parallele è equivalente all'assioma che dice che per un punto P che non sta su una retta l passa una e una sola retta parallela a l, esistono due alternative a questo assioma. Una consiste nell'affermare che non vi sono

---

[5] A.-M. LEGENDRE, in «Mém. de l'Acad. des Sci., Paris», XII (1833), 367-410.

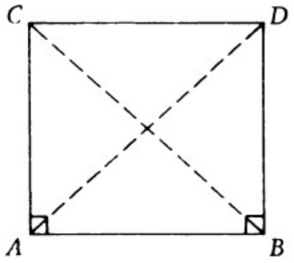

Figura 36.3.

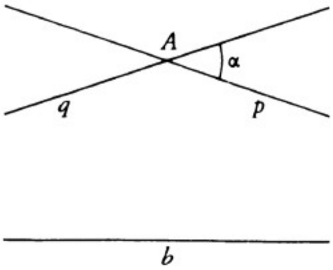

Figura 36.4.

parallele a l passanti per P, l'altra che ve ne sono piú di una. Se adottando ciascuna di queste alternative in luogo dell'assioma sull'«unica parallela» si potesse dimostrare che i nuovi insiemi di assiomi conducono a una contraddizione, allora entrambe dovrebbero essere respinte e l'asserzione sull'«unica parallela» risulterebbe *dimostrata*.

Lo sforzo piú significativo di questo tipo fu compiuto da Gerolamo Saccheri (1667-1733), un gesuita che fu professore all'università di Pavia. Egli studiò attentamente l'opera di Nasîr-Eddin e di Wallis e adottò poi il suo approccio personale. Saccheri partí con il quadrilatero $ABCD$ (fig. 36.3) in cui gli angoli in A e in B sono retti e $AC = BD$. È poi facile provare che $\sphericalangle C = \sphericalangle D$. Ora, l'assioma euclideo delle parallele è equivalente all'asserzione che gli angoli in C e in D sono retti. Saccheri prese perciò in considerazione le due alternative possibili: 1) ipotesi dell'angolo ottuso: $\sphericalangle C$ e $\sphericalangle D$ sono ottusi; 2) ipotesi dell'angolo acuto: $\sphericalangle C$ e $\sphericalangle D$ sono acuti. Sulla base della prima ipotesi (e degli altri nove assiomi di Euclide) Saccheri dimostrò che gli angoli in C e in D devono essere retti. L'ipotesi dell'angolo ottuso conduce perciò a una contraddizione.

Saccheri passò poi a considerare l'ipotesi dell'angolo acuto e dimostrò molti teoremi interessanti. Continuò fino a che giunse al seguente teorema: dati un punto A e una retta b (fig. 36.4), l'ipotesi dell'angolo acuto implica che nel fascio di rette per A esistano due rette p e q che lo dividono in due parti. La prima di esse è costituita dalle rette che intersecano b, la seconda è costituita dalle rette (comprese nell'angolo α) che posseggono una perpendicolare comune con b. Le rette p e q stesse sono asintotiche a b. Da questo risultato e da un lungo ragionamento successivo Saccheri dedusse che p e b dovrebbero avere una perpendicolare comune nel loro punto comune, che è all'infinito. Anche se non era giunto ad alcuna contraddizione, Saccheri trovò questa e altre conclusioni cosí ripugnanti che decise che l'ipotesi dell'angolo acuto doveva essere falsa.

Rimaneva soltanto l'ipotesi che gli angoli in C e in D della figura

36.3 fossero retti. Saccheri aveva dimostrato in precedenza che, quando questi angoli sono retti, la somma degli angoli di un qualsiasi triangolo è uguale a 180° e che questo fatto implica il postulato euclideo delle parallele. Si considerò perciò giustificato nel concludere che Euclide aveva ragione e pubblicò perciò l'*Euclides ab Omni Naevo Vindicatus* (1733). Tuttavia, poiché Saccheri non aveva ricavato una contraddizione dall'ipotesi dell'angolo acuto, il problema dell'assioma delle parallele rimaneva aperto.

Gli sforzi per trovare un sostituto accettabile per l'assioma euclideo delle parallele o per dimostrare che esso deve essere un teorema furono cosí numerosi e futili che nel 1759 d'Alembert chiamò il problema dell'assioma delle parallele «lo scandalo degli elementi di geometria».

4. *I primi presagi della geometria non euclidea.*

Nella sua dissertazione del 1763 Georg S. Klügel (1739-1812), professore di matematica all'università di Helmstädt, che conosceva il libro di Saccheri, fece la notevole osservazione che la certezza con cui veniva accettata la verità dell'assioma euclideo delle parallele era basata sull'esperienza. Questa osservazione introdusse per la prima volta l'idea che gli assiomi fossero convalidati dall'esperienza piuttosto che dall'evidenza. Klügel espresse dei dubbi circa il fatto che l'asserzione euclidea potesse essere provata. Egli si rendeva conto che Saccheri non era pervenuto a una contraddizione, ma soltanto a risultati che sembravano essere in disaccordo con l'esperienza.

Il lavoro di Klügel suggerí a Lambert delle ricerche sull'assioma delle parallele. Nella sua *Theorie der Parallellinien*, scritta nel 1766 e pubblicata nel 1786[6], Lambert, un po' come Saccheri, considerò un quadrilatero con tre angoli retti e prese in considerazione la possibilità che il quarto angolo fosse retto, ottuso o acuto. Tuttavia, a differenza di Saccheri, Lambert non concluse di aver ottenuto una contraddizione dall'ipotesi dell'angolo acuto.

Le conseguenze che Lambert dedusse rispettivamente dall'ipotesi dell'angolo ottuso e dell'angolo acuto, anche se la prima conduceva a una contraddizione, sono degne di nota. Il suo risultato piú notevole è che, in entrambe le ipotesi, l'area di un poligono con n lati è proporzionale alla differenza fra la somma degli angoli e $2n-4$ angoli retti (Saccheri aveva ottenuto questo risultato per il triangolo). Notò anche che l'ipotesi dell'angolo ottuso dava origine a teoremi simili a quelli che valgono per le figure tracciate sulla superficie di una sfera e fece la congettura che i teoremi che seguivano dall'ipotesi dell'angolo acuto si applicassero alle figure tracciate su una sfera di raggio immaginario. Questo lo indusse a scrivere un lavoro[7] sulle funzioni trigonometriche degli angoli immaginari, cioè della forma iA, dove A è un angolo reale e $i = \sqrt{-1}$, in cui in effetti introdusse le funzioni iperboliche (cfr. cap. XIX, § 2). Vedremo un po' piú chiaramente in seguito che cosa significano le affermazioni di Lambert.

---

[6] J. H. LAMBERT, in «Magazin für reine und angewandte Mathematik», 1786, 137-64, 325-58.

Le opinioni di Lambert sulla geometria erano molto avanzate. Egli si rendeva conto che ogni corpo di ipotesi che non conducessero a una contraddizione forniva una possibile geometria. Una tale geometria sarebbe stata una struttura logica valida anche se avesse avuto poco a che fare con le figure reali. Queste ultime potevano suggerire una particolare geometria, ma non controllavano la varietà di geometrie logicamente sviluppabili. Lambert non giunse alla conclusione piú radicale, che fu tratta un po' di tempo dopo da Gauss.

Un ulteriore passo avanti fu compiuto da Ferdinand Karl Schweikart (1780-1859), un professore di giurisprudenza che dedicava il tempo libero alla matematica. Egli lavorò sulla geometria non euclidea nel periodo in cui Gauss dedicò qualche pensiero all'argomento, ma Schweikart giunse alle sue conclusioni indipendentemente. Fu tuttavia influenzato dall'opera di Saccheri e di Lambert. In un promemoria del 1816 inviato a Gauss nel 1818 per l'approvazione, Schweikart distingueva due geometrie. C'è la geometria di Euclide e c'è una geometria basata sull'assunzione che la somma degli angoli di un triangolo non è uguale a due angoli retti. Egli chiamava quest'ultima geometria astrale perché potrebbe valere nello spazio stellare e i suoi teoremi erano quelli che Saccheri e Lambert avevano stabilito sulla base dell'ipotesi dell'angolo acuto.

Franz Adolph Taurinus (1794-1874), nipote di Schweikart, seguí il suggerimento dello zio di studiare la geometria astrale. Anche se stabilí nei suoi *Geometriae Prima Elementa* (1826) alcuni risultati nuovi, in particolare di geometria analitica, ne concluse che soltanto la geometria di Euclide poteva essere vera per lo spazio fisico, ma che la geometria astrale era *logicamente coerente*. Taurinus dimostrò anche che le formule che varrebbero su una sfera di raggio immaginario sono esattamente quelle che valgono nella sua geometria astrale.

Le ricerche di Lambert, Schweikart e Taurinus costituiscono dei progressi che meritano di essere ricapitolati. Tutti e tre, e insieme con loro altri matematici come Klügel e Abraham G. Kästner (1719-1800), un professore di Göttingen, erano convinti che l'assioma delle parallele non potesse essere dimostrato, cioè che fosse indipendente dagli altri assiomi di Euclide. Inoltre, Lambert, Schweikart e Taurinus erano convinti che fosse possibile adottare un assioma alternativo che contraddicesse quello di Euclide e costruire con esso una geometria logicamente coerente. Lambert non si pronunciò circa l'applicabilità di una tale geometria, Taurinus pensava che non fosse possibile applicarla allo spazio fisico, mentre Schweikart riteneva che essa potesse essere applicata allo spazio interstellare. Questi tre matematici notarono anche che la geometria su una sfera reale ha le stesse proprietà della geometria basata sull'ipotesi dell'angolo ottuso (lasciando da parte le contraddizioni che conseguono da quest'ultima) e che la geometria su una sfera con raggio immaginario ha le stesse proprietà della geometria basata sull'ipotesi dell'angolo acuto. Cosí, tutti e tre riconobbero l'esistenza di una geometria non euclidea, ma mancarono di cogliere un punto essenziale, e cioè che la geometria euclidea non è l'unica geometria che descrive le proprietà dello spazio fisico nei limiti della precisione che l'esperienza può garantire.

---

[7] ID., in «Hist. de l'Acad. de Berlin», XXIV (1768), 327-54, pubbl. 1770 = *Opera Mathematica*, II, 245-69.

5. *La creazione della geometria non euclidea.*

Nessuna branca fondamentale della matematica né alcun importante risultato specifico sono opera di un solo uomo; al massimo, è possibile attribuire a un singolo un passo o una dimostrazione decisivi. Questo sviluppo collettivo della matematica si applica in particolare alla geometria non euclidea. Se si intende per creazione della geometria non euclidea il riconoscimento che esistono delle geometrie alternative a quelle di Euclide, allora il merito deve essere attribuito a Klügel e a Lambert. Se geometria non euclidea significa lo sviluppo tecnico delle conseguenze di un sistema di assiomi contenente un'alternativa all'assioma euclideo delle parallele, allora la maggior parte del merito deve essere attribuita a Saccheri, anche se beneficiò delle ricerche di molti matematici che avevano tentato di trovare un assioma piú accettabile che sostituisse quello di Euclide. Tuttavia, il fatto piú significativo concernente la geometria non euclidea è che essa può essere usata per descrivere le proprietà dello spazio fisico con altrettanta precisione della geometria euclidea. Quest'ultima non è la geometria necessaria dello spazio fisico e la sua verità fisica non può essere garantita su basi aprioristiche. Questa presa di coscienza, che non richiedeva nessuno sviluppo tecnico perché questi erano già stati ottenuti, fu conseguita per la prima volta da Gauss.

Carl Friedrich Gauss (1777-1855) era figlio di un muratore della città tedesca di Braunschweig e sembrava destinato al lavoro manuale. Il direttore della scuola in cui ricevette l'educazione elementare fu però colpito dalla sua intelligenza e richiamò su di lui l'attenzione del duca Carlo Guglielmo. Il duca mandò Gauss a una scuola superiore e poi, nel 1795, all'università di Göttingen. Qui Gauss incominciò a lavorare duramente sulle sue idee. A diciotto anni inventò il metodo dei minimi quadrati e a diciannove dimostrò che il poligono regolare di 17 lati è costruibile. Questi successi lo convinsero a passare dalla filologia alla matematica. Nel 1798 si trasferí all'università di Helmstädt dove fu notato da Johann Friedrich Pfaff, che diventò suo maestro e amico. Dopo aver ottenuto il dottorato, Gauss ritornò a Braunschweig, dove scrisse alcuni dei suoi lavori piú famosi, che gli fruttarono nel 1807 la nomina a professore di astronomia e a direttore dell'osservatorio di Göttingen. Se si eccettua un viaggio a Berlino per partecipare a un congresso scientifico, Gauss rimase a Göttingen per il resto della sua vita. Non amava insegnare e lo diceva. Gli piaceva tuttavia la vita di società, si sposò due volte ed ebbe dei bambini.

La prima opera importante di Gauss fu la sua tesi di dottorato, in cui dimostrò il teorema fondamentale dell'algebra. Nel 1801 pubblicò le classiche *Disquisitiones arithmeticae*. Il suo lavoro di geometria differenziale, le *Disquisitiones generales circa superficies curvas* (1827), che, incidentalmente, fu il risultato dei suoi interessi per la geodesia, la topografia e la cartografia, è una pietra miliare della matematica (cfr. cap. XXXVII, § 2). Portò molti altri contributi all'algebra, alla teoria delle funzioni complesse e alla teoria del potenziale. In lavori non pubblicati sono riportate le sue ricerche innovatrici sulla teoria delle funzioni ellittiche e sulla geometria non euclidea.

I suoi interessi per la fisica furono ugualmente ampi e ad essi dedicò la maggior parte delle sue energie. Quando Giuseppe Piazzi (1746-1826) scoprí nel 1801 il pianeta Cerere, Gauss si propose di determinarne l'orbita. Questo fu l'inizio delle sue ricerche astronomiche, l'attività che l'assorbí maggiormente e a cui dedicò circa vent'anni. Una delle sue grandi pubblicazioni in questo settore è la *Theoria Motus Corporum Coelestium* (1809). Gauss ottenne anche grande fama per le sue ricerche sul magnetismo teorico e sperimentale. Maxwell dice nell'*Electricity and Magnetism* che gli studi di Gauss sul magnetismo ricostruirono l'intera scienza, gli strumenti usati, i metodi di osservazione e il calcolo dei risultati. I lavori di Gauss sul magnetismo terrestre sono dei modelli di ricerche fisiche e fornirono il metodo migliore per misurare il campo magnetico terrestre. Le sue ricerche sull'astronomia e sul magnetismo aprirono un nuovo e brillante periodo di cooperazione fra la matematica e la fisica.

Anche se Gauss e Whilhelm Weber (1804-91) non inventarono il telegrafo, nel 1833 migliorarono le tecniche precedenti con un dispositivo pratico che faceva ruotare un ago a destra e a sinistra a seconda della direzione della corrente fatta passare in un filo. Gauss si occupò anche di ottica, che era stata trascurata dai tempi di Euclide, e le sue ricerche del 1838-41 diedero una base completamente nuova alla trattazione dei problemi ottici.

L'universalità delle attività di Gauss è tanto piú degna di nota in quanto i suoi contemporanei avevano incominciato a confinarsi in ricerche specialistiche. Sebbene sia considerato il piú grande matematico almeno dopo Newton, Gauss non fu tanto un innovatore quanto una figura di transizione fra il Settecento e l'Ottocento. Anche se elaborò delle idee nuove che occuparono in seguito molti altri matematici, egli fu orientato piú verso il passato che verso il futuro. Felix Klein descrive la posizione di Gauss con queste parole: «Potremmo avere un quadro dello sviluppo della matematica se immaginassimo una catena di alte montagne, rappresentanti i matematici del Settecento, che culmini con una cima imponente – Gauss – e poi un'ampia e ricca regione riempita da nuovi elementi di vita. I contemporanei di Gauss ne apprezzarono il genio e all'epoca della sua morte, avvenuta nel 1855, era universalmente venerato e chiamato il "principe dei matematici"».

Gauss pubblicò relativamente poco delle sue ricerche perché amava rifinire tutto ciò che faceva, in parte per conferirgli eleganza, in parte per dare alle sue dimostrazioni il massimo di concisione senza sacrificare il rigore (quello del suo tempo, almeno). Nel caso della geometria non euclidea non pubblicò nessun lavoro definitivo. In una lettera a Bessel del 27 gennaio 1829 dice che non avrebbe probabilmente mai pubblicato le sue scoperte in questo campo perché temeva «le strida dei Beoti». Può darsi che Gauss sia stato troppo cauto, ma bisogna ricordare che, anche se alcuni matematici avevano gradualmente raggiunto il vertice delle ricerche sulla geometria non euclidea, il mondo intellettuale in senso lato era ancora dominato dalle idee kantiane. Ciò che sappiamo delle ricerche gaussiane sulla geometria non euclidea è tratto dalle lettere agli amici, da due brevi recensioni apparse sulle «Göttingische gelehrte Anzeigen» del 1816 e del 1822 e da alcune note del 1831 trovate fra le sue carte dopo la morte[8].

---

[8] C. F. GAUSS, *Werke*, VIII, 157-268; contiene tutto ciò e anche la lettera discussa sotto.

Gauss era pienamente consapevole della vanità degli sforzi per stabilire il postulato euclideo delle parallele, perché questo era un fatto comunemente noto a Göttingen e tutta la storia di questi sforzi era pienamente nota al suo maestro Kästner. Gauss disse all'amico Schumacher che fin dal 1792 (quando aveva quindici anni) si era fatto l'idea che fosse possibile l'esistenza di una geometria logica in cui non fosse valido il postulato euclideo delle parallele. Con il 1794 Gauss aveva trovato che nel suo concetto di geometria non euclidea l'area di un quadrangolo deve essere proporzionale alla differenza fra 360° e la somma dei suoi angoli interni. In questo periodo, però, e fino al 1799, Gauss stava ancora tentando di dedurre il postulato euclideo delle parallele da altre assunzioni piú plausibili e riteneva ancora che la geometria euclidea fosse la geometria dello spazio fisico anche se poteva immaginare altre geometrie non euclidee logicamente corrette. Tuttavia, il 17 dicembre 1799 Gauss scrisse al matematico ungherese suo amico Wolfgang Farkas Bolyai (1775-1856):

> Quanto a me ho già fatto alcuni progressi nel mio lavoro. Tuttavia, il cammino che ho scelto non conduce affatto alla meta che cerchiamo [la deduzione dell'assioma delle parallele] e che tu mi assicuri di aver raggiunto. Essa sembra piuttosto obbligarmi a dubitare della verità della stessa geometria. È vero che sono giunto a qualche cosa che la maggior parte delle persone riterrebbe che costituisca una dimostrazione, ma ai miei occhi non prova un bel nulla. Ad esempio, se potessimo dimostrare la possibilità dell'esistenza di un triangolo rettilineo la cui area sia maggiore di ogni area data, allora sarei pronto a provare tutta la geometria [euclidea] in maniera assolutamente rigorosa.
>
> La maggior parte delle persone accetterebbero certamente questo fatto come assioma, ma io no! Potrebbe in effetti essere possibile che l'area rimanga sempre al di sotto di un certo limite, per quanto distanti siano presi i tre vertici del triangolo.

Questo passo dimostra che nel 1799 Gauss era abbastanza convinto che l'assioma delle parallele non potesse essere dedotto dagli altri assiomi euclidei e incominciava a prendere piú sul serio lo sviluppo di una nuova geometria suscettibile di applicazioni.

A partire dal 1813 circa Gauss sviluppò la sua nuova geometria che all'inizio chiamava geometria antieuclidea, poi geometria astrale e infine geometria non euclidea. Egli si convinse che essa era logicamente coerente ed era piuttosto sicuro che potesse essere applicata. Nelle recensioni del 1816 e del 1822 e nella lettera a Bessel del 1829 Gauss riaffermò che il postulato delle parallele non poteva essere dimostrato sulla base degli altri assiomi di Euclide. La lettera a Olbers scritta nel 1817[9] costituisce una pietra miliare. In essa Gauss dice: «Mi sto convincendo sempre di piú che la necessità fisica della nostra geometria [euclidea] non può essere dimostrata, almeno non dalla ragione umana né per ragioni umane. Forse in un'altra vita penetreremo la natura dello spazio, che è per ora irraggiungibile. Fino ad allora dovremo porre la geometria non nella stessa classe dell'aritmetica, che è puramente a priori, ma in quella della meccanica».

Per verificare l'applicabilità della geometria euclidea e della sua geometria non euclidea Gauss misurò effettivamente la somma degli angoli del triangolo formato dalle cime delle tre montagne Brocken, Hohehagen e Inselberg. I lati di questo triangolo erano rispettivamente lunghi 69, 85 e 197 chilometri. Gauss trovò[10] che la somma degli angoli superava i 180° di 14".85. L'esperimento non dimostrò niente perché l'errore sperimentale era molto piú grande dell'eccedenza trovata, cosicché la somma corretta avrebbe potuto benissimo essere uguale a 180° o anche meno. Come Gauss si rese conto, il triangolo era troppo piccolo e poiché nella geometria non euclidea il difetto è proporzionale all'area, soltanto un triangolo molto grande avrebbe potuto eventualmente rivelare una differenza significativa rispetto ai 180° nella somma degli angoli.

Non discuteremo i teoremi specifici di geometria non euclidea dovuti a Gauss. Egli non scrisse un'esposizione completamente deduttiva delle sue ricerche e i teoremi che provò sono in gran parte simili a quelli che incontreremo nell'opera di Lobatchevsky e di Bolyai, i due matematici a cui viene generalmente attribuita la creazione della geometria non euclidea. Che cosa debba loro essere attribuito lo discuteremo in seguito, ma essi pubblicarono delle presentazioni organiche di una geometria non euclidea su basi deduttive sintetiche rendendosi pienamente conto che questa nuova geometria era dal punto di vista logico altrettanto legittima di quella di Euclide.

Nikolai Ivanovich Lobatchevsky (1793-1856), russo, studiò all'università di Kazan di cui fu professore e rettore dal 1827 al 1846. Espose le sue vedute sui fondamenti della geometria in un lavoro letto di fronte al dipartimento di matematica e fisica dell'università nel 1826. Tuttavia, il lavoro non fu mai stampato e andò perduto. In seguito espose il suo approccio alla geometria non euclidea in una serie di lavori, i primi due dei quali furono pubblicati in riviste di Kazan e il terzo nel «Journal für Mathematik»[11]. Il primo era intitolato *Sui fondamenti della geometria* e apparve nel 1829-30. Il secondo, intitolato *Nuovi fondamenti della geometria con una teoria completa delle parallele* (1835-37), è una presentazione migliore delle idee di Lobatchevsky. Egli chiamava la sua nuova geometria «geometria immaginaria» per motivi che sono forse già chiari e che lo diventeranno di piú in seguito. Nel 1840 pubblicò un libro in tedesco, intitolato *Geometrische Untersuchungen zur Theorie der Parallellinien* (Ricerche geometriche sulla teoria delle parallele)[12]. In esso si lamenta per lo scarso interesse nei confronti dei suoi scritti. Pur essendo diventato cieco, dettò un'esposizione completamente nuova della sua geometria e la pubblicò nel 1855 con il titolo di *Pangéométrie*.

---

[9] *Ibid.*, 177.
[10] *Ibid.*, IV, 258.

János Bolyai (1802-60), figlio di Wolfgang Bolyai, era un ufficiale ungherese. Sulla geometria non euclidea, che chiamava geometria assoluta, scrisse un lavoro di ventisei pagine intitolato *La scienza dello spazio assoluto*[13], che fu pubblicato in appendice al libro del padre intitolato *Tentamen Juventutem Studiosam in Elementa Matheseos.* Anche se quest'opera in due volumi apparve nel 1832-33, e quindi dopo il primo lavoro di Lobatchevsky, sembra che Bolyai abbia elaborato le sue idee sulla geometria non euclidea prima del 1825 e che entro quel periodo si fosse convinto che la nuova geometria non era contraddittoria. In una lettera al padre datata 23 novembre 1823 János scrive: «Ho fatto delle scoperte cosí meravigliose che sono io stesso sconvolto per lo stupore». Le ricerche di Bolyai erano cosí simili a quelle di Lobatchevsky che quando Bolyai vide per la prima volta nel 1835 i lavori del russo pensò che li avesse copiati dalla sua pubblicazione del 1832-33. D'altro lato, Gauss lesse l'articolo di János Bolyai nel 1832 e scrisse a Wolfgang[14] che non poteva lodarlo perché, cosí facendo, avrebbe lodato le proprie ricerche.

6. *Il contenuto tecnico della geometria non euclidea.*

Gauss, Lobatchevsky e Bolyai si erano resi conto che l'assioma euclideo delle parallele non poteva essere dimostrato sulla base degli altri nove assiomi e che esso era necessario per fondare la geometria euclidea. Poiché l'assioma delle parallele era indipendente, doveva quindi

---

[11] N. I. LOBATCHEVSKY, in «Jour. für Math.», XVII (1837), 295-320.

[12] Se ne trova una traduzione in inglese nel libro di Bonola citato nella bibliografia.

[13] Se ne trova una traduzione in inglese nel libro di Bonola citato nella bibliografia.

[14] C. F. GAUSS, *Werke*, VIII, 220-21.

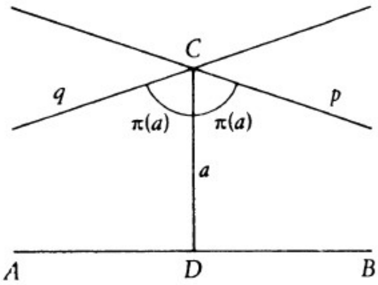

Figura 36.5.

essere possibile, almeno dal punto di vista logico, adottare un enunciato che lo contraddicesse e sviluppare le conseguenze del nuovo insieme di assiomi.

Per studiare il contenuto tecnico delle loro scoperte tanto vale prendere in considerazione l'opera di Lobatchevsky perché tutti e tre fecero praticamente le stesse cose. Lobatchevsky diede, come sappiamo, numerose versioni che differiscono soltanto nei dettagli. Come base per la nostra esposizione ci serviremo del lavoro del 1835-37.

Poiché, come negli *Elementi* di Euclide, si possono dimostrare molti teoremi che non dipendono affatto dall'assioma delle parallele, tali teoremi sono validi anche nella nuova geometria. Lobatchevsky dedica i primi sei capitoli del suo lavoro alla dimostrazione di questi teoremi fondamentali. Egli assume all'inizio che lo spazio sia infinito e può poi dimostrare che due rette non possono intersecarsi in piú di un punto e che due perpendicolari alla stessa retta non possono intersecarsi.

Nel settimo capitolo Lobatchevsky respinge bravamente l'assioma euclideo delle parallele e fa la seguente assunzione: dati una retta AB e un punto C (fig. 36.5), le rette per C si dividono in due classi rispetto ad AB, e cioè, la classe delle rette che incontrano AB e la classe delle rette che non la incontrano. A quest'ultima appartengono due rette p e q che costituiscono il confine fra le due classi. Queste due rette sono dette parallele. Piú precisamente, se C è un punto a distanza a dalla retta AB, allora esiste un angolo[15] $\pi(a)$ tale che tutte le rette per C che formano con la perpendicolare CD un angolo minore di $\pi(a)$ intersecheranno AB, mentre tutte le altre rette per C non intersecheranno AB[16]. Le due rette che formano l'angolo $\pi(a)$ con AB sono le parallele e $\pi(a)$ è detto angolo di parallelismo. Le rette per C diverse dalle parallele e che non incontrano AB sono dette linee non intersecanti, anche se nel senso di Euclide esse sono parallele ad AB, e perciò in questo senso la geometria di Lobatchevsky contiene un numero infinito di parallele passanti per C.

---

[15] Il simbolo $\pi(a)$ è standard e perciò viene usato qui. Va da sé che il π che compare in $\pi(a)$ non ha nulla a che fare con il numero π.

[16] L'idea che si possa associare un angolo specifico a ciascuna lunghezza è dovuta a Lambert.

Se $\pi(a) = \pi/2$ si ha l'assioma euclideo. In caso contrario, ne segue che $\pi(a)$ cresce e tende a $\pi/2$ quando a tende a 0, mentre $\pi(a)$ decresce e tende a 0 quando a diventa infinito. La somma degli angoli di un triangolo è sempre minore di π, decresce quando l'area del triangolo cresce e tende a π quando l'area tende a zero. Se due triangoli sono simili, allora sono congruenti.

Lobatchevsky affronta poi l'aspetto trigonometrico della sua geometria. Il primo passo è la determinazione di $\pi(a)$. Il risultato, se un angolo giro è uguale a 2π, è[17]:

$$\operatorname{tg} \frac{\pi(x)}{2} = e^{-x}, \tag{1}$$

da cui segue che $\pi(0) = \pi/2$ e $\pi(+\infty) = 0$. La relazione (1) è importante, in quanto a ogni lunghezza x associa un angolo definito $\pi(x)$. Se $x = 1$ si ha $\operatorname{tg} [\pi(1)/2] = e^{-1}$ e quindi $\pi(1) = 40° 24'$. L'unità di lunghezza è perciò quella lunghezza il cui angolo di parallelismo è uguale a 40° 24'. Questa unità di lunghezza non ha un significato fisico diretto. Dal punto di vista fisico può essere uguale a un centimetro o a un chilometro. Bisognerà scegliere l'interpretazione fisica che renda applicabile la geometria[18].

Lobatchevsky deduce poi le formule che legano fra loro i lati e gli angoli dei triangoli piani della sua geometria. In un lavoro del 1834 aveva definito cos x e sin x, per x reale, come le parti reale e immaginaria di e^{ix}. L'obiettivo di Lobatchevsky era quello di dare una fondazione puramente analitica alla trigonometria, in modo da renderla indipendente dalla geometria euclidea. Le principali formule trigonometriche della sua geometria sono le seguenti (fig. 36.6):

$$\begin{aligned} \operatorname{cotg} \pi(a) &= \operatorname{cotg} \pi(c) \sin A \\ \sin A &= \cos B \sin \pi(b) \\ \sin \pi(c) &= \sin \pi(a) \sin \pi(b). \end{aligned}$$

---

[17] Questa è una formulazione particolare. Nel lavoro del 1840 Lobatchevsky dà la formula che viene usata solitamente nei testi moderni e che è contenuta anche in Gauss, e cioè:

$$\operatorname{tg} \frac{1}{2} \pi(x) = e^{x/k}, \tag{a}$$

dove k è una costante che viene detta costante spaziale. Dal punto di vista teorico il valore di k è irrilevante. Anche Bolyai dà la formula (*a*).

[18] Nel caso della relazione $\operatorname{tg} [\pi(x)/2] = e^{-x/k}$, la scelta del valore di x che corrisponde, ad esempio, a 40° 24' determina k.

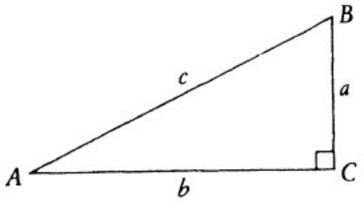

Figura 36.6.

Queste formule sono valide nella trigonometria sferica ordinaria purché i lati abbiano lunghezze immaginarie. In altri termini, se si sostituiscono a, b e c nelle formule usuali della trigonometria sferica con ia, ib e ic si ottengono le formule di Lobatchevsky. Poiché le funzioni trigonometriche degli angoli immaginari possono essere sostituite dalle funzioni iperboliche, ci si potrebbe aspettare di veder comparire queste ultime funzioni nelle formule di Lobatchevsky. Esse possono essere introdotte servendosi della relazione $\operatorname{tg} [\pi(x)/2] = e^{-x/k}$. In questo modo la prima delle formule precedenti diventa:

$$\sinh \frac{a}{k} = \sinh \frac{c}{k} \sin A.$$

Inoltre, mentre nella geometria sferica usuale l'area di un triangolo di angoli A, B e C è uguale a $r^2(A + B + C - \pi)$, nella geometria non euclidea essa è uguale a $r^2[\pi - (A + B + C)]$, il che equivale a sostituire il raggio r della sfera con ir nella formula usuale.

Servendosi di un triangolo infinitesimo, Lobatchevsky aveva derivato nel suo primo lavoro (1829-30) la formula

$$ds = \sqrt{(dy)^2 + \frac{(dx)^2}{\sin^2 \pi(x)}}$$

per l'elemento d'arco di una curva di equazione $y = f(x)$ nel punto (x, y). Servendosi di essa si può calcolare la lunghezza della circonferenza di un cerchio di raggio r, che risulta essere uguale a

$$C = \pi(e^r - e^{-r}).$$

Analogamente, il valore dell'area del cerchio risulta essere uguale a

$$A = \pi(e^{r/2} - e^{-r/2})^2.$$

Egli dà anche dei teoremi sulle aree di regioni piane e curve e sui volumi dei solidi.

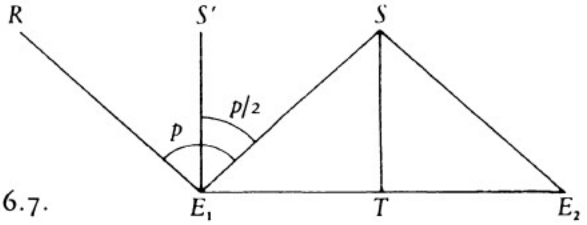

Figura 36.7.

Le formule della geometria euclidea seguono da quelle non euclidee quando le grandezze in gioco sono piccole. Cosí, se si usa il fatto che

$$e^r = 1 + r + \frac{r^2}{2!} + \dots$$

e si trascurano, per r piccolo, tutti i termini tranne i primi due, si ottiene ad esempio

$$C = \pi(e^r - e^{-r}) \sim \pi\{1 + r - (1 - r)\} = 2\pi r.$$

Nel primo lavoro (1829-30) Lobatchevsky prese anche in esame le possibilità di applicazione della sua geometria allo spazio fisico. L'essenza del suo ragionamento si basa sulla parallasse delle stelle. Supponiamo che E_1 ed E_2 (fig. 36.7) siano le posizioni della Terra a distanza di sei mesi e che S sia una stella. La parallasse p di S è la differenza fra le direzioni di E_1S e di E_2S misurata, ad esempio, a partire dalla perpendicolare E_1S'. Se E_1R è la parallela euclidea a E_2S, allora, poiché E_1SE_2 è un triangolo isoscele, $\pi/2 - \sphericalangle SE_1E_2$ è la metà del cambiamento di direzione della stella, cioè è uguale a $p/2$. Questo angolo è uguale a $1''.24$ per la stella Sirio (valore dato da Lobatchevsky). Fintanto che questo angolo non è uguale a zero, la retta da E_1 alla stella non può essere parallela a TS perché essa taglia TS. Se invece ci fosse un estremo inferiore per le varie parallassi di tutte le stelle, allora ogni retta uscente da E_1 e formante con E_1S' un angolo minore di questo estremo inferiore potrebbe essere considerata una parallela a TS passante per E_1 e questa geometria sarebbe ugualmente utile fintanto che si abbia a che fare con le misure stellari. Tuttavia, Lobatchevsky dimostra che l'unità di lunghezza di questa geometria dovrebbe essere uguale a piú di mezzo milione di volte il raggio dell'orbita terrestre. In altre parole, le geometria di Lobatchevsky potrebbe essere applicata soltanto in un triangolo enormemente grande.

7. *Le rivendicazioni di priorità di Lobatchevsky e di Bolyai.*

La geometria non euclidea viene spesso usata come esempio del come la stessa idea possa venire nello stesso momento, ma indipendentemente, a piú persone. Talvolta ciò viene considerato come una pura coincidenza e talvolta come una prova del fatto che lo spirito dei tempi esercita la sua influenza in direzioni molto distanti fra loro. La creazione della geometria non euclidea da parte di Gauss, Lobatchevsky e Bolyai non è un esempio di creazione simultanea né è giustificato il grande merito attribuito a Lobatchevsky e a Bolyai. È vero, come abbiamo già notato, che essi furono i primi a pubblicare una geometria dichiaratamente non euclidea e con questo atto diedero prova di un coraggio maggiore di quello di Gauss. Tuttavia, la creazione della geometria non euclidea può difficilmente esser loro attribuita. Abbiamo già osservato che anche Gauss era stato preceduto da Lambert, che Schweikart e Taurinus furono dei creatori indipendenti, che Lambert e Taurinus pubblicarono le loro ricerche e, inoltre, che fu Gauss a rendersi conto per primo che la nuova geometria può essere applicata allo spazio fisico.

Sia Lobatchevsky che Bolyai devono molto a Gauss. Il maestro di Lobatchevsky a Kazan fu Johann Martin Bartels (1769-1836), un buon amico di Gauss. In effetti, essi passarono insieme a Braunschweig gli anni che vanno dal 1805 al 1807. Successivamente Gauss e Bartels rimasero in contatto fra di loro. È estremamente improbabile che Bartels non abbia comunicato a Lobatchevsky, che rimase a Kazan come suo collega, i progressi di Gauss nella geometria non euclidea. In particolare, Bartels deve essere stato al corrente dei dubbi di Gauss circa la verità della geometria euclidea.

Quanto a János Bolyai, il padre Wolfgang era anch'esso un intimo amico di Gauss, di cui era stato compagno a Göttingen dal 1796 al 1798. Wolfgang Bolyai e Gauss non soltanto continuarono a rimanere in comunicazione, ma discussero anche l'argomento dell'assioma delle parallele, come indica una delle citazioni precedenti. Wolfgang Bolyai continuò a lavorare intensamente sul problema dell'assioma delle parallele e mandò a Gauss nel 1804 quella che riteneva essere una dimostrazione, ma Gauss gli dimostrò che era sbagliata. Entro il 1817 Gauss era sicuro non soltanto che l'assioma non poteva essere dimostrato, ma anche che era possibile costruire una geometria non euclidea logicamente coerente e applicabile fisicamente. Oltre alla sua comunicazione in tal senso del 1799, Gauss trasmise liberamente le sue idee successive all'amico, il quale continuò a occuparsi del problema fino a quando pubblicò il

(%'

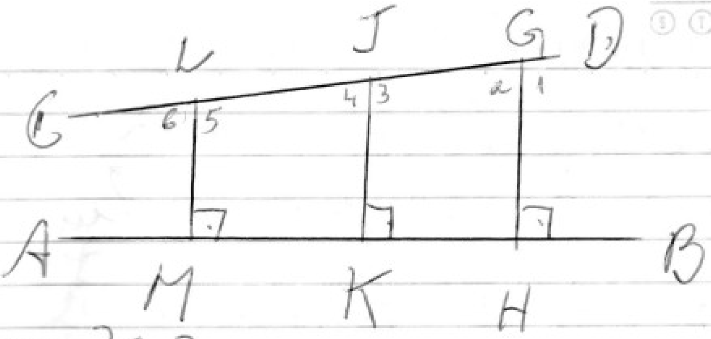

Figura 36.2

⇒ essere asintotiche.

Nasîr-Eddin (1201-74),
l'editore di Euclide, diede anche
lui una « dimostrazione »
del postulato euclideo delle
parallele

assumendo che due rette non
parallele si avvicinino
l'una all'altra in una
(l'una) direzione e si
allontanino
nell'altra.

Pág. 1021 (308) -
Pag. 1020 (307) ⇒

Nel settimo capitolo
Lobatchevsky respinge bravamente
l'assioma euclideo delle parallele

/ (%'

e fa la seguente assunzione:
→ aborda a questão do π.
Pág. 1021(308)

$$\operatorname{tg} \frac{\pi(x)}{2} = e^{-x}$$

- Morris Kline - Calculus (1998).pdf
- Anastasios Mallios - Differential Sheaves and Connections (2015).pdf

---

- Morris Kline - Mathematics and the Physical World. (1981).djvu

83 occurrences — Gauss.

- Morris Kline - Mathematical Thought from Ancient to Modern Times (1972).djvu

Pág. 261(277) — When, in 1801, Gauss adopted x^2 for xx, the former became standard.

/
! -256- - -* ! - / + ! 2 & A
! " 3 ! 211 ! 4 A D ! // 1"

D -256- - -* ! - / 2 ! ! ! 4 !
!A ! " 2 8! ! 211 ! 3 ! 7 3 @ " $ 1"

-256- - -* ! - / + ! ! ! "
3 ! @ " 4 A D ! / "KA

-256- - -* ! - / + ! ! ! "
3 ! @ " 4 A D ! / 1"
*D % 7!

" -256- - -* ! - / + ! ! !
9 2 " 9 *L " H A 3 / "KA

" -256- - -* ! - / + ! ! !
9 2 " 9 *L " H A 3 / 1"

" -256- - -* ! - / + ! ! !
9 2 " 9 *L " H A 3 J H 2 / "KA

" -256- - -* ! - / + ! ! !
9 2 " 9 *L " H A 3 J H 2 / 1"

"% -256- - -* ! - / + ! ! !
9 2 " 9 % *L " H A 3 J H 2 / "KA

"% -256- - -* ! - / + ! ! !
9 2 " 9 % *L " H A 3 J H 2 / 1"

*D
-256- - -* ! - / + ! ! ! 9
2 " 9 *L " H A 3 / 1"
(
B J 7! !" 1 " LX LLJ D ! ! "! "

Claude Bachet de Méziriac (1581–1638) preferred to write $x^3 + 13x^2 + 5x + 2$ as $1C + 13Q + 5N + 2$. Vieta used the same notation for equations with numerical coefficients.

Descartes made rather systematic use of positive integral exponents. He expressed

$$1 + 3x + 6x^2 + x^3 \quad \text{as} \quad 1 + 3x + 6xx + x^3.$$

Occasionally he and others used x^2 also. For higher powers he used x^4, $x^5, \ldots$ but not x^n. Newton used positive, negative, integral, and fractional exponents, as in $x^{5/3}$ and x^{-3}. When, in 1801, Gauss adopted x^2 for xx, the former became standard.

The most significant change in the character of algebra was introduced in connection with symbolism by François Vieta. Trained as a lawyer, he served in that capacity for the parliament in Brittany; he was later privy councillor to Henry of Navarre. When he was out of office from 1584 to 1589, as a result of political opposition, he devoted himself entirely to mathematics. Generally, he pursued mathematics as a hobby and printed and circulated his work at his own expense—a guarantee, as one writer put it, of oblivion.

Vieta was a humanist in spirit and intention; he wished to be the preserver, rediscoverer, and continuator of ancient mathematics. Innovation was for him renovation. He describes his *In Artem Analyticam Isagoge*[6] as the "work of the restored mathematical analysis." For this book he drew on the seventh book of Pappus' *Mathematical Collection* and on Diophantus' *Arithmetica*. He believed that the ancients had used a general algebraic type of calculation, which he reintroduced in his algebra, thus merely reactivating an art known and approved of in antiquity.

During the hiatus in his political career Vieta studied the works of Cardan, Tartaglia, Bombelli, Stevin, and Diophantus. From them and particularly from Diophantus he got the idea of using letters. Though a number of men, including Euclid and Aristotle, had used letters in place of specific numbers, these uses were infrequent, sporadic, and incidental. Vieta was the first to use letters purposefully and systematically, not just to represent an unknown or powers of the unknown but as general coefficients. Usually he used consonants for the known quantities and vowels for the unknown quantities. He called his symbolic algebra *logistica speciosa*, as opposed to *logistica numerosa*. Vieta was fully aware that when he studied the general quadratic equation $ax^2 + bx + c = 0$ (in our notation), he was studying an entire class of expressions. In making the distinction between *logistica numerosa* and *logistica speciosa* in his *Isagoge*, Vieta drew the line between arithmetic and algebra. Algebra, the *logistica speciosa*, he said, was a method of operating on species or forms of things. Arithmetic, the *numerosa*,

---

[6] Introduction to the Analytic Art, 1591 = *Opera*, 1–12.

-256- - -* ! - / + ! ! ! 9
2 " 9 *L " H A 3 J H 2 / 1"
(' ' $ 7! ,

$\phi = \pi$. The notation $\Delta(k, \phi)$ for $\sqrt{1 - k^2 \sin^2 \phi}$ was also introduced by Legendre.

These forms can be converted, by the change of variable $x = \sin \phi$, to the Jacobi forms

$$F(k, x) = \int_0^x \frac{dx}{\sqrt{1 - x^2}\sqrt{1 - k^2x^2}} \tag{42}$$

$$E(k, x) = \int_0^x \sqrt{\frac{1 - k^2x^2}{1 - x^2}}\, dx \tag{43}$$

$$\pi(n, k, x) = \int_0^x \frac{dx}{(1 + nx^2)\sqrt{(1 - x^2)(1 - k^2x^2)}}. \tag{44}$$

The quantity k is called the modulus of each elliptic integral. If the limits of integration are $\phi = \pi/2$ or $x = 1$, then the integrals are called complete; otherwise, incomplete.

There was much merit in Legendre's work on elliptic integrals. He drew many inferences, previously unstated, from the work of his predecessors and organized the mathematical subject; but he did not add any basic ideas, nor did he attain the new insight of Abel and Jacobi (Chap. 27, sec. 6) who inverted these integrals and so conceived of the elliptic *functions*. Legendre did get to know the work of Abel and Jacobi and praised them, with much humility and, probably, some bitterness. In devoting the supplements to his 1825 work to their new ideas, he understood very well that this material threw into the shade all he had done on the subject. He had overlooked one of the great discoveries of his epoch.

5. *Further Special Functions*

The elliptic indefinite integrals are new transcendental functions. As the analytical work of the eighteenth century developed, more transcendental functions were obtained. Of these the most important is the gamma function, which arose from work on two problems, interpolation theory and antidifferentiation. The problem of interpolation had been considered by James Stirling (1692–1770), Daniel Bernoulli, and Christian Goldbach (1690–1764). It was posed to Euler and he announced his solution in a letter of October 13, 1729, to Goldbach.[33] A second letter, of January 8, 1730, brought in the integration problem.[34] In 1731 Euler published results on both in a paper, "De Progressionibus. . . ."[35]

[33] Fuss, *Correspondance*, 1, 3–7.

[34] Fuss, *Correspondance*, 1, 11–18.

[35] *Comm. Acad. Sci. Petrop.*, 5, 1730/1, 36–57, pub. 1738 = *Opera*, (1), 14, 1–24.

The interpolation problem was to give meaning to $n!$ for nonintegral values of n. Euler noted that

$$(45)\qquad \begin{aligned} n! &= \left[\left(\frac{2}{1}\right)^n \frac{1}{n+1}\right]\left[\left(\frac{3}{2}\right)^n \frac{2}{n+2}\right]\left[\left(\frac{4}{3}\right)^n \frac{3}{n+3}\right]\cdots \\ &= \prod_{k=1}^{\infty} \left(\frac{k+1}{k}\right)^n \frac{k}{k+n}. \end{aligned}$$

The equation is seen to be formally correct if one cancels common factors in the infinite product. However, this analytic expression for $n!$, unlike the basic definition $n(n-1)\cdots 2\cdot 1$, has a meaning for all n except negative integral values. Euler noticed that for $n = 1/2$ the right side yields, after a bit of manipulation, the infinite product of Wallis

$$(46)\qquad \frac{\pi}{2} = \left(\frac{2\cdot 2}{1\cdot 3}\right)\left(\frac{4\cdot 4}{3\cdot 5}\right)\left(\frac{6\cdot 6}{5\cdot 7}\right)\left(\frac{8\cdot 8}{7\cdot 9}\right)\cdots.$$

In the notation $\Gamma(n+1) = n!$, introduced later by Legendre, Euler also showed that $\Gamma(n+1) = n\Gamma(n)$ and so obtained $\Gamma(3/2)$, $\Gamma(5/2)$, and so forth.

Euler could have used (45) as his generalization of the factorial concept. In fact, it is often introduced today in the equivalent form, which Euler also gave, namely,

$$(47)\qquad \lim_{m\to\infty} \frac{m!\,(m+1)^n}{(n+1)(n+2)\cdots(n+m)}.$$

But the connection with Wallis's result led Euler to take up an integral already considered by Wallis, namely,

$$(48)\qquad \int_0^1 x^e(1-x)^n\,dx$$

wherein for Euler e and n are now arbitrary. Euler evaluated this integral by expanding $(1-x)^n$ by the binomial theorem and obtained

$$(49)\qquad \begin{aligned} &\int_0^1 x^e(1-x)^n\,dx \\ &\quad = \frac{1}{e+1} - \frac{n}{1(e+2)} + \frac{n(n-1)}{1\cdot 2(e+3)} - \frac{n(n-1)(n-2)}{1\cdot 2\cdot 3(e+4)} + \cdots. \end{aligned}$$

For $n = 0, 1, 2, 3, \ldots$ the sums on the right side are respectively

$$(50)\qquad \frac{1}{e+1}, \frac{1}{(e+1)(e+2)}, \frac{1\cdot 2}{(e+1)(e+2)(e+3)}, \frac{1\cdot 2\cdot 3}{(e+1)(e+2)(e+3)(e+4)}, \ldots.$$

Hence for positive integral n Euler found that

$$\int_0^1 x^e(1-x)^n\,dx = \frac{n!}{(e+1)(e+2)\cdots(e+n+1)}. \tag{51}$$

Euler now sought an expression for $n!$ with n arbitrary. By a series of transformations that would not be wholly acceptable to us today, Euler arrived at

$$n! = \int_0^1 (-\log x)^n\,dx. \tag{52}$$

This integral has meaning for almost all arbitrary n. It is called the second Eulerian integral, or, as Legendre later called it, the gamma function, and is denoted by $\Gamma(n+1)$. [Gauss let $\pi(n) = \Gamma(n+1)$]. Later, in 1781 (pub. 1794), Euler gave the modern form, which is obtained from (52) by letting $t = -\log x$,

$$\Gamma(n+1) = \int_0^\infty x^n e^{-x}\,dx. \tag{53}$$

The integral (48), which Legendre called the first Eulerian integral, became standardized as the beta function,

$$B(m,n) = \int_0^1 x^{m-1}(1-x)^{n-1}\,dx. \tag{54}$$

Euler discovered the relation between the two integrals,[36] namely,

$$B(m,n) = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}.$$

In his *Exercices de calcul intégral*, Legendre made a profound study of the Eulerian integrals and arrived at the duplication formula

$$\Gamma(2x) = (2\pi)^{-1/2}\,2^{2x-(1/2)}\Gamma(x)\Gamma\left(x+\frac{1}{2}\right). \tag{55}$$

Gauss studied the gamma function in his work on the hypergeometric function[37] and extended Legendre's result to what is called the multiplication formula:

$$\Gamma(nx) = (2\pi)^{(1-n)/2}\,n^{nx-(1/2)}\Gamma(x)\Gamma\left(x+\frac{1}{n}\right)\Gamma\left(x+\frac{2}{n}\right)\cdots\Gamma\left(x+\frac{n-1}{n}\right). \tag{56}$$

---

[36] *Novi Comm. Acad. Sci. Petrop.*, 16, 1771, 91–139, pub. 1772 = *Opera*, (1), 17, 316–57.

[37] *Comm. Soc. Gott.*, II, 1813 = *Werke*, 3, 123–162, p. 149 in part.

($/ B 1!1 7! 1 A" ! " 1 3 ! ! D L1 "! !1 "
! ! " > !" ! ! !

out that complex roots occur in conjugate pairs, so that the product of $x - (a + b\sqrt{-1})$ and $x - (a - b\sqrt{-1})$, wherein $a + b\sqrt{-1}$ and $a - b\sqrt{-1}$ are a conjugate pair, gives a quadratic expression with real coefficients. Euler then showed that this was true for Bernoulli's example. But Goldbach, too, rejected the idea that every polynomial with real coefficients can be factored into real factors and gave the example $x^4 + 72x - 20$. Euler then showed Goldbach that the latter had made a mistake and that he (Euler) had proved his theorem for polynomials up to the sixth degree. However, Goldbach was not convinced, because Euler did not succeed in giving a general proof of his assertion.

The kernel of the problem of factoring a real polynomial into linear and quadratic factors with real coefficients was to prove that every such polynomial had at least one real or complex root. The proof of this fact, called the fundamental theorem of algebra, became a major goal.

Proofs offered by d'Alembert and Euler were incomplete. In 1772[1] Lagrange, in a long and detailed argument, "completed" Euler's proof. But Lagrange, like Euler and his contemporaries, applied freely the ordinary properties of numbers to what were supposedly the roots without establishing that the roots must at worst be complex numbers. Since the nature of the roots was unknown, the proof was actually incomplete.

The first substantial proof of the fundamental theorem, though not rigorous by modern standards, was given by Gauss in his doctoral thesis of 1799 at Helmstädt.[2] He criticized the work of d'Alembert, Euler, and Lagrange and then gave his own proof. Gauss's method was not to calculate a root but to demonstrate its existence. He points out that the complex roots $a + ib$ of $P(x + iy) = 0$ correspond to points (a, b) of the plane, and if $P(x + iy) = u(x, y) + iv(x, y)$ then (a, b) must be an intersection of the curves $u = 0$ and $v = 0$. By a qualitative study of the curves he shows that a continuous arc of one joins the points of two distinct regions separated by the other. Then the curve $u = 0$ must cut the curve $v = 0$. The argument was highly original. However, he depended on the graphs of these curves, which were somewhat complicated, to show that they must cross In this same paper Gauss proved that the nth degree polynomial can be expressed as a product of linear and quadratic factors with real coefficients.

Gauss gave three more proofs of the theorem. In the second proof[3] he dispensed with geometrical arguments. In this proof he also showed that the product of the differences of each two roots (which we, following Sylvester, call the discriminant) can be expressed as a linear combination of the polynomial and its derivative, so that a necessary and sufficient condition that the polynomial and its derivative have a common root is that the discriminant vanish. However, this second proof assumed that a polynomial cannot change signs at two different values of x without vanishing in between. The proof of this fact lay beyond the rigor of the time.

---

[1] *Nouv Mém de l'Acad de Berlin*, 1772, 222 ff = *Œuvres*, 3, 479–516.

[2] *Werke*, 3, 1–30; also reproduced in Euler, *Opera*, (1), 6, 151–69.

[3] *Comm. Soc. Gott.*, 3, 1814/15, 107–42 = *Werke*, 3, 33–56.

The third proof[4] used in effect what we now call the Cauchy integral theorem (Chap. 27, sec. 4).[5] The fourth proof[6] is a variation of the first as far as method is concerned. However, in this proof Gauss uses complex numbers more freely because, he says, they are now common knowledge. It is worth noting that the theorem was, in many proofs, not proved in all generality. Gauss's first three proofs and later proofs by Cauchy, Jacobi, and Abel assumed that the (literal) coefficients represented real numbers, whereas the full theorem includes the case of complex coefficients. Gauss's fourth proof did allow the coefficients of the polynomial to be complex numbers.

Gauss's approach to the fundamental theorem of algebra inaugurated a new approach to the entire question of mathematical existence. The Greeks had wisely recognized that the existence of mathematical entities must be established before theorems about them can be entertained. Their criterion of existence was constructibility. In the more explicit formal work of the succeeding centuries, existence was established by actually obtaining or exhibiting the quantity in question. For example, the existence of the solutions of a quadratic equation is established by exhibiting quantities that satisfy the equation. But in the case of equations of degree higher than four, this method is not available. Of course a proof of existence such as Gauss's may be of no help at all in computing the object whose existence is being established.

While the work that ultimately showed that every polynomial equation with real coefficients has at least one root was under way, the mathematicians also pushed hard to solve equations of degree higher than four by algebraic processes. Leibniz and his friend Tschirnhausen were the first to make serious efforts. Leibniz[7] reconsidered the irreducible case of the third degree equation and convinced himself that one could not avoid the use of complex numbers to solve this type. He then tackled the solution of the fifth degree equation but without success. Tschirnhausen[8] thought he had solved this problem by transforming the given equation into a new one by means of a transform $y = P(x)$, where $P(x)$ is a suitable fourth degree polynomial.

---

[4] *Comm. Soc. Gott*, 3, 1816 = *Werke*, 3, 59–64.

[5] For a discussion of Gauss's third proof see M. Bocher, "Gauss's Third Proof of the Fundamental Theorem of Algebra," *Amer. Math. Soc., Bull.*, 1, 1895, 205–9. A translation of the third proof can be found in H. Meschkowski, *Ways of Thought of Great Mathematicians*, Holden-Day, 1964.

[6] *Abhand. der Ges. der Wiss. zu Gött.*, 4, 1848/50, 3–34 = *Werke*, 3, 73–102.

[7] *Der Briefwechsel von Gottfried Wilhelm Leibniz mit Mathematikern*, Georg Olms (reprint), 1961, Vol. 1, 547–64.

[8] *Acta Erud.*, 2, 1683, 204–07.

This transformation eliminated all but the x^5 and constant terms of the equation. But Leibniz showed that to determine the coefficients of $P(x)$ one had to solve equations of degree higher than five, and so the method was useless.

For a while the problem of solving the nth degree equation centered on the special case $x^n - 1 = 0$, called the binomial equation. Cotes and De Moivre showed, through the use of complex numbers, that the solution of this problem amounts to the division of the circumference of a circle into n equal parts. To obtain the roots by radicals (trigonometric solutions are not necessarily algebraic) it is sufficient to consider the case of n an odd prime, because if $n = pm$, where p is a prime, then one can consider $(x^m)^p - 1 = 0$. If this equation can be solved for x^m, then $x^m - A$ can be considered where A is any of the roots of the preceding equation. Alexandre-Théophile Vandermonde (1735–96) affirmed in a paper of 1771[9] that every equation of the form $x^n - 1 = 0$, where n is a prime, is solvable by radicals. However, Vandermonde merely verified that this is so for values of n up to 11. The decisive work on binomial equations was done by Gauss (Chap. 31, sec. 2).

The major effort toward solving equations of degree higher than four concentrated on the general equation, and toward this end some subsidiary work on symmetric functions proved important. The expression $x_1x_2 + x_2x_3 + x_3x_1$ is a symmetric function of x_1, x_2, and x_3 because replacement throughout of any x_i by an x_j and x_j by x_i leaves the entire expression unaltered. The interest in symmetric functions arose when the seventeenth-century algebraists noted and Newton proved that the various sums of the products of the roots of a polynomial equation can be expressed in terms of the coefficients. For example, for $n = 3$, the sum of the products taken two at a time,

$$a_1a_2 + a_2a_3 + a_3a_1$$

is an elementary symmetric function; and if the equation is written as

$$x^3 - c_1x^2 + c_2x - c_3 = 0,$$

the sum equals c_2. The progress made by Vandermonde in the 1771 paper was to show that any symmetric function of the roots can be expressed in terms of the coefficients of the equation.

The outstanding work of the eighteenth century on the problem of the solution of equations by radicals, after efforts by many men including Euler,[10] was by Vandermonde in the 1771 paper and Lagrange, in his massive paper "Réflexions sur la résolution algébrique des équations."[11] Vandermonde's

---

[9] *Hist. de l'Acad. des Sci., Paris*, 1771, 365–416, pub. 1774.

[10] *Comm. Acad. Sci. Petrop.*, 6, 1732/3, 216–31, pub. 1738 = *Opera*, (1), 6, 1–19 and *Novi Comm. Acad. Sci. Petrop.*, 9, 1762/63, 70–98, pub. 1764 = *Opera*, (1), 6, 170–96.

[11] *Nouv. Mém. de l'Acad. de Berlin*, 1770, 134–215, pub. 1772 and 1771, 138–254, pub. 1773 = *Œuvres*, 3, 205–421.

Elimination theory was also applied to two equations, $f(x, y) = 0$ and $g(x, y) = 0$, of degree higher than 1. The motivation was to establish the number of common solutions of the two equations or, geometrically, to find the number of intersections of the curves corresponding to the equations. The outstanding method for eliminating one unknown from $f(x, y) = 0$ and $g(x, y) = 0$, sketched first by Bezout in a paper of 1764, was presented by him in his *Théorie générale des équations algébriques* (1779). Bezout's idea was that by multiplying $f(x, y)$ and $g(x, y)$ by suitable polynomials, $F(x)$ and $G(x)$ respectively, he could form

$$R(y) = F(x)f(x, y) + G(x)g(x, y). \tag{11}$$

Moreover he sought an F and G such that the degree of $R(y)$ would be as small as possible.

The question of the degree of the eliminant was also answered by Bezout in his *Théorie* (and independently by Euler in the 1764 paper). Both gave the answer as mn, the product of the two degrees of f and g, and both proved this theorem by reducing the problem to one of elimination from an auxiliary set of linear equations. This product is the number of points of intersection of the two algebraic curves. Jacobi[22] and Minding[23] also gave Bezout's method of elimination for the two equations, but neither mentions Bezout. It could be that Bezout's work was not known to them.

4. *The Theory of Numbers*

The theory of numbers in the eighteenth century remained a series of disconnected results. The most important works in the subject were Euler's *Anleitung zur Algebra* (German edition, 1770) and Legendre's *Essai sur la théorie des nombres* (1798). The second edition appeared in 1808 under the title *Théorie des nombres*, and an expanded third edition in two volumes appeared in 1830. The problems and results to be described are a small sample of what was done.

In 1736 Euler established[24] the minor theorem of Fermat, namely, if p is a prime and a any positive integer, $a^p - a$ is divisible by p. Many proofs of this same theorem were given by other men in the eighteenth and nineteenth centuries. In 1760 Euler generalized this theorem[25] by introducing the ϕ function or totient of n; $\phi(n)$ is the number of integers $< n$ and prime to n so that if n is a prime, $\phi(n)$ is $n - 1$. [The notation $\phi(n)$ was introduced by Gauss.] Then Euler proved that if a is relatively prime to n,

$$a^{\phi(n)} - 1$$

is divisible by n.

---

[22] *Jour. für Math*, 15, 1836, 101–24 = *Gesam. Werke*, 3, 297–320.

[23] *Jour. fur Math.*, 22, 1841, 178–83.

[24] *Comm. Acad. Sci. Petrop.*, 8, 1736, 141–46, pub. 1741 = *Opera*, (1), 2, 33–37.

[25] *Novi Comm. Acad. Sci. Petrop.*, 8, 1760/1, 74–104, pub. 1763 = *Opera*, (1), 3, 531–55.

As for the famous Fermat conjecture on $x^n + y^n = z^n$, Euler proved[26] this was correct for $n = 3$ and $n = 4$; the case $n = 4$ had been proven by Frénicle de Bessy. This work of Euler had to be completed by Lagrange, Legendre, and Gauss. Legendre then proved[27] the conjecture for $n = 5$. As we shall see, the history of the efforts to prove Fermat's conjecture is extensive.

Fermat had also conjectured (Chap. 13, sec. 7) that the formula

$$2^{2^n} + 1$$

yielded primes for an unspecified set of values of n. This is true for $n = 0$, 1, 2, 3, and 4. Fermat did not claim it was always true, and Euler showed in 1732[28] that 641 was a factor when $n = 5$. In fact the formula is now known not to yield primes for many other values of n, whereas no other value beyond 4 has been found for which the formula does yield primes. However, the formula is of interest in that it reappears in the work of Gauss on the constructibility of regular polygons (Chap. 31, sec. 2).

A topic with many subdivisions concerns the decomposition of integers of various types into other classes of integers. Fermat had affirmed that each positive integer is the sum of at most four squares (repetition of a square as in $8 = 4 + 4$ is permitted if it is counted the number of times it appears). Over a period of forty years, Euler kept trying to prove this theorem and gave partial results.[29] Using some of Euler's work, Lagrange[30] proved the theorem. Neither Euler nor Lagrange obtained the number of representations.

Euler, in the 1754/5 paper just referred to and in another paper in the same journal[31] proved Fermat's assertion that every prime of the form $4n + 1$ is decomposable uniquely into a sum of two squares. However, Euler's proof did not follow the method of descent that Fermat had sketched for this theorem. In another paper[32] Euler also proved that every divisor of the sum of two relatively prime squares is the sum of two squares.

Edward Waring (1734–98) stated, in his *Meditationes Algebraicae* (1770), the theorem known now as "Waring's theorem," that every integer is either a cube or the sum of at most nine cubes; also every integer is either a fourth power or the sum of at most 19 fourth powers. He conjectured also that every positive integer can be expressed as the sum of at most r kth powers, the r depending upon k. These theorems were not proven by him.[33]

---

[26] *Algebra*, Part II, Second Section, 509–16 = *Opera*, (1), 1, 484–89 (for $n = 3$); and *Comm. Acad. Sci. Petrop.*, 10, 1738, 125–46, pub. 1747 = *Opera*, (1), 2, 38–59 (for $n = 4$).

[27] *Mém. de l'Acad. des Sci., Paris*, 6, 1823, 1–60, pub. 1827.

[28] *Comm. Acad Sci Petrop.*, 6, 1732/3, 103–07 = *Opera*, (1), 2, 1–5.

[29] *Novi Comm Acad. Sci. Petrop.*, 5, 1754/5, 13–58, pub. 1760 = *Opera*, (1), 2, 338–72.

[30] *Nouv. Mém. de l'Acad. de Berlin*, 1, 1770, 123–33, pub. 1772 = *Œuvres*, 3, 189–201.

[31] 5, 1754/5, 3–13, pub. 1760 = *Opera*, (1), 2, 328–37.

[32] *Novi Comm. Acad. Sci. Petrop.*, 4, 1752/53, 3–40, pub. 1758 = *Opera*, (1) 2, 295–327.

[33] The general theorem was proven by David Hilbert (*Math. Ann.*, 67, 1909, 281–300).

In a letter to Euler of June 7, 1742, Christian Goldbach, a Prussian envoy to Russia, stated without proof that every even integer is the sum of two primes and every odd integer is either a prime or a sum of three primes. The first part of the assertion is now known as Goldbach's conjecture and is still an unsolved problem. The second assertion actually follows from the first, because if n is odd substract any prime p from it. Then $n - p$ is even.

Among somewhat more specialized results dealing with the decomposition of numbers is Euler's proof that $x^4 - y^4$ and $x^4 + y^4$ cannot be squares.[34] Euler and Lagrange proved many of Fermat's assertions to the effect that certain primes can be expressed in particular ways. For example Euler proved[35] that a prime of the form $3n + 1$ can be expressed uniquely in the form $x^2 + 3y^2$.

Amicable and perfect numbers continued to interest the mathematicians. Euler[36] gave 62 pairs of amicable numbers, including the three pairs already known. Two of his pairs were incorrect. He also proved in a posthumously published paper[37] the converse of Euclid's theorem: Every even perfect number is of the form $2^{p-1}(2^p - 1)$ where the second factor is prime.

John Wilson (1741–93), who had been a prize student of mathematics at Cambridge but became a lawyer and judge, stated a theorem still named after him: For every prime p, the quantity $(p - 1)! + 1$ is divisible by p; in addition, if the quantity is divisible by p, then p is a prime. Waring published the statement in his *Meditationes Algebraicae* and Lagrange proved it in 1773.[38]

The problem of solving the equation $x^2 - Ay^2 = 1$ in integers has already been discussed (Chap. 13, sec. 7). Euler, in a paper of 1732/33, erroneously called it Pell's equation, and the name has stuck. He became interested in this equation because he needed its solutions to solve $ax^2 + bx + c = y^2$ in integers; he wrote several papers on this last theme. In 1759 Euler gave a method of solving the Pellian equation by expressing $\sqrt{A}$ as a continued fraction.[39] Euler's idea was that the values x and y which satisfy the equation are such that the ratios x/y are convergents (in the sense of continued fractions) to $\sqrt{A}$. He failed to prove that his method always gives solutions and that all its solutions are given by the continued fraction development of $\sqrt{A}$. The existence of solutions of Pell's equation was shown in 1766 by Lagrange[40] and then more simply in later papers.[41]

---

[34] *Comm. Acad. Sci. Petrop.*, 10, 1738, 125–46, pub. 1747 = *Opera*, (1), 2, 38–59; also in *Algebra* (1770), Part II, Ch. 13, arts. 202–8 = *Opera*, (1), 1, 436–43.

[35] *Novi Comm. Acad. Sci. Petrop.*, 8, 1760/1, 105–28, pub. 1763 = *Opera*, (1), 2, 556–75.

[36] "De numeris amicabilibus," *Opuscula varii argumenti*, 2, 1750, 23–107 = *Opera*, (1), 2, 86–162.

[37] "De numeris amicabilibus," *Comm. Arith.*, 2, 1849, 627–36 = *Opera postuma*, 1, 1862, 85–100 = *Opera*, (1), 5, 353–65.

[38] *Nouv. Mém. de l'Acad. de Berlin*, 2, 1771, 125 ff., pub. 1773 = *Œuvres*, 3, 425–38.

[39] *Novi Comm. Acad. Sci. Petrop.*, 11, 1765, 28–66, pub. 1767 = *Opera*, (1), 3, 73–111.

Fermat had asserted that he could determine when the more general equation $x^2 - Ay^2 = B$ was solvable in integers and that he could solve it when solvable. The equation was solved by Lagrange in the two papers just mentioned.

The problem of giving all integral solutions of the general equation

$$ax^2 + 2bxy + cy^2 + 2dx + 2ey + f = 0,$$

where the coefficients are integers, was also tackled. Euler gave incomplete classes of solutions; then Lagrange[42] gave the complete solution. In the next volume of the *Mémoires*[43] he gave a simpler proof.

The most original and perhaps the most consequential discovery of the eighteenth century in the theory of numbers is the law of quadratic reciprocity. It uses the notion of quadratic remainders or residues. In the language introduced by Euler in the 1754/55 paper and adopted by Gauss, if there exists an x such that $x^2 - p$ is divisible by q, then p is said to be a quadratic residue of q; if there is no such x, p is said to be a quadratic nonresidue of q. Legendre (1808) invented a symbol that is now used to represent either state of affairs. The symbol is (p/q) and means the following: For any number p and any prime q,

$$(p/q) = \begin{cases} 1 & \text{if } p \text{ is a quadratic residue of } q \\ -1 & \text{if } p \text{ is a quadratic nonresidue of } q. \end{cases}$$

It is also understood that $(p/q) = 0$ if p divides evenly into q.

The law of quadratic reciprocity, in symbolic form, states that if p and q are distinct odd primes, then

$$(p/q)(q/p) = (-1)^{(p-1)(q-1)/4}.$$

This means that if the exponent of (-1) is even, p is a quadratic residue of q and q is a quadratic residue of p or neither is a quadratic residue of the other. When the exponent is odd, which occurs when p and q are of the form $4k + 3$, one prime will be a quadratic residue of the other but not the second of the first.

The history of this law is detailed. Euler in a paper of 1783[44] gave four theorems and a fifth summarizing theorem that states the law of quadratic reciprocity very clearly. However, he did not prove these theorems. The work on this paper dates from 1772 and incorporates even earlier work. Kronecker observed in 1875[45] that the statement of the law is actually contained in a much earlier paper by Euler.[46] However, Euler's "proof" was based on calculations. In 1785 Legendre announced the law independently, though he cites another paper by Euler, in the same volume of the *Opuscula*, in his own paper on the subject. His proof[47] was incomplete. In his *Théorie des nombres*[48] he again stated the law and gave another proof. However, this one, too, was incomplete, because he assumed that there are an infinite number of primes in certain arithmetic progressions. The desire to find what is behind the law and to derive its many implications has been a key theme in investigations since 1800 and has led to important discoveries, some of which we shall consider in a later chapter.

---

[40] *Misc. Taur*, 4, 1766/69, 19 ff. = *Œuvres*, 1, 671–731.

[41] *Mém de l'Acad. de Berlin*, 23, 1767, 165–310, pub. 1769, and 24, 1768, 181–256, pub. 1770 = *Œuvres*, 2, 377–535 and 655–726; also in Lagrange's additions to his translation of Euler's *Algebra*; see the bibliography

[42] *Mém de l'Acad. de Berlin*, 23, 1767, 165–310, pub 1769 = *Œuvres*, 2, 377–535.

[43] 24, 1768, 181–256, pub 1770 = *Œuvres*, 2, 655–726.

[44] *Opuscula Analytica*, 1, 1783, 64–84 = *Opera*, (1), 3, 497–512

The eighteenth-century work on the theory of numbers closes with Legendre's classic *Théorie* of 1798. Though it contains a number of interesting results, in this domain as in others (such as the subject of elliptic integrals), Legendre made no great innovations. One could reproach him for presenting a collection of propositions from which general conceptions could have been extracted but were not. This was done by his successors.

Bibliography

Cajori, Florian: "Historical note on the Graphical Representation of Imaginaries Before the Time of Wessel," *Amer. Math. Monthly*, 19, 1912, 167–71.

Cantor, Moritz: *Vorlesungen über Geschichte der Mathematik*, B. G. Teubner, 1898 and 1924; Johnson Reprint Corp., 1965, Vol. 3, Chap. 107; Vol. 4, pp. 153–98.

Dickson, Leonard E.: *History of the Theory of Numbers*, 3 vols., Chelsea (reprint), 1951.

— — -: "Fermat's Last Theorem," *Annals of Math.*, 18, 1917, 161-87.

Euler, Leonhard: *Opera Omnia*, (1), Vols. 1 5, Orell Fussli, 1911-44.

— - : *Vollstandige Anleitung zur Algebra* (1770) = *Opera Omnia*, (1), 1.

Fuss, Paul H. von, ed.: *Correspondance mathématique et physique de quelques célèbres géomètres du XVIIIème siècle*, 2 vols. (1843), Johnson Reprint Corp., 1967.

Gauss, Carl Friedrich: *Werke*, Königliche Gesellschaft der Wissenschaften zu Göttingen, 1876, Vol. 3, pp. 3–121.

Gerhardt, C. I.: *Der Briefwechsel von Gottfried Wilhelm Leibniz mit Mathematikern*, Mayer und Muller, 1899; Georg Olms (reprint), 1962.

Heath, Thomas L.: *Diophantus of Alexandria*, 1910, Dover (reprint), 1964, pp. 267–380.

Jones, P. S.: "Complex Numbers: An Example of Recurring Themes in the

---

[45] *Werke*, 2, 3–10

[46] *Comm Acad. Sci Petrop.*, 14, 1744/46, 151–81, pub 1751 = *Opera*, (1), 2, 194–222.

[47] *Hist. de l'Acad. des Sci*, *Paris*, 1785, 465–559, pub 1788.

[48] 1798, 214–26; 2nd ed., 1808, 198–207.

Development of Mathematics," *The Mathematics Teacher*, 47, 1954, 106–14, 257–63, 340–45.

Lagrange, Joseph-Louis: *Œuvres*, Gauthier-Villars, 1867–69, Vols. 1–3, relevant papers.

———: "Additions aux éléments d'algèbre d'Euler," *Œuvres*, Gauthier-Villars, 1877, Vol. 7, pp. 5–179.

Legendre, Adrien-Marie, *Théorie des nombres*, 4th ed., 2 vols., A. Blanchard (reprint), 1955.

Muir, Thomas: *The Theory of Determinants in the Historical Order of Development*, 1906, Dover (reprint), 1960, Vol. 1, pp. 1–52.

Ore, Oystein: *Number Theory and its History*, McGraw-Hill, 1948.

Pierpont, James: "Lagrange's Place in the Theory of Substitutions," *Amer. Math. Soc. Bulletin*, 1, 1894/5, 196–204.

———: "Zur Geschichte der Gleichung des V. Grades (bis 1858)," *Monatshefte fur Mathematik und Physik*, 6, 1895, 15–68.

Smith, H. J. S.: *Report on the Theory of Numbers*, 1867, Chelsea (reprint), 1965; also in Vol. 2 of the *Collected Mathematical Papers of H. J. S. Smith*, 1894, Chelsea (reprint), 1965.

Smith, David Eugene: *A Source Book in Mathematics*, 1929, Dover (reprint), 1959, Vol. 1, relevant selections. One of the selections is an English translation of Gauss's second proof of the fundamental theorem of algebra.

Struik, D. J.: *A Source Book in Mathematics, 1200–1800*, Harvard University Press, 1969, pp. 26–54, 99–122.

Vandiver, H. S.: "Fermat's Last Theorem," *Amer. Math. Monthly*, 53, 1946, 555–78.

Whiteside, Derek T.: *The Mathematical Works of Isaac Newton*, Johnson Reprint Corp., 1967, Vol. 2, pp. 3–134. This section contains Newton's *Universal Arithmetic* in English.

Wussing, H. L.: *Die Genesis der abstrakten Gruppenbegriffes*, VEB Deutscher Verlag der Wissenschaften, 1969.

31

Galois Theory

> Unfortunately what is little recognized is that the most worthwhile scientific books are those in which the author clearly indicates what he does not know; for an author most hurts his readers by concealing difficulties. EVARISTE GALOIS

1. *Introduction*

The basic problem of algebra, the solution of polynomial equations, continued to occupy the center of the stage in the algebra of the early nineteenth century. In this period, the broad question of which equations are solvable by algebraic operations was definitely and comprehensively answered by Galois. Moreover, he not only created the first significant coherent body of algebraic theory but he introduced new notions which were to be developed into still other broadly applicable theories of algebra. In particular the concepts of a group and a field emerged from his work and Abel's.

2. *Binomial Equations*

We have already discussed (Chap. 25, sec. 2) the fruitless efforts of Euler, Vandermonde, Lagrange, and Ruffini to solve algebraically equations of degree greater than 4 and the binomial equation $x^n - 1 = 0$. A major success was achieved by Gauss. In the last section of his *Disquisitiones Arithmeticae*[1] Gauss considered the equation

$$(1) \qquad x^p - 1 = 0,$$

where p is a prime.[2] This equation is often called the cyclotomic equation or the equation for the division of a circle. The latter term refers to the fact that the roots of this equation are, by de Moivre's theorem,

$$(2) \qquad x_j = \cos\frac{k2\pi\theta}{p} + i\sin\frac{k2\pi\theta}{p}, \qquad k = 1, 2, \ldots, p,$$

---

[1] 1801, *Werke*, I.

[2] The case of p prime takes care of $x^n - 1 = 0$ for if $n = pq$, let $y = x^q$. But $y^p - 1 = 0$ is solvable. Hence x^q = const. can be solved if q is a prime and if not q can be decomposed in the same manner that n is.

and the complex numbers x_j when plotted geometrically are the vertices of a p-sided regular polygon that lie on the unit circle.

Gauss showed that the roots of this equation may be rationally expressed in terms of the roots of a sequence of equations

$$Z_1 = 0, Z_2 = 0, \cdots, \tag{3}$$

whose coefficients are rational in the roots of the preceding equations of the sequence. The degrees of the equations (3) are precisely the prime factors of $p - 1$. There is a Z_i for each factor even if repeated. Moreover, each of the $Z_i = 0$ can be solved by radicals and so equation (1) can also be so solved.

This result is of course of great significance for the problem of solving the general nth-degree equation algebraically. It shows that some equations of high degree can be solved by radicals, for example a fifth degree equation if 5 is a factor of $p - 1$ or a seventh degree equation if 7 is a factor.

The result is of major importance also for the geometric problem of constructing regular polygons of p sides. If $p - 1$ contains no factors other than 2, then the polygon is constructible with straightedge and compass because the degrees of the equations (3) are each 2 and each of its roots is constructible in terms of its coefficients. Thus we are able to construct all polygons of a prime number of sides p if $p - 1$ is a power of 2. Such primes are 3, 5, 17, 257, 65537, ... Alternatively, a regular polygon can be constructed if p is a prime of the form $2^{2^h} + 1$.[3] Gauss remarks (Art. 365) that although the geometric construction of regular polygons of 3, 5, and 15 sides and those immediately derivable from them—for example, 2^n, $2^n \cdot 3$, $2^n \cdot 5$, $2^n \cdot 15$, wherein n is a positive integer—was known in Euclid's time, in an interval of 2000 years no new constructible polygons had been discovered and geometers had been unanimous in declaring that no others could be constructed.

Gauss thought that his result might lead to all sorts of attempts to find new constructible polygons of a prime number of sides. He warns then: "As often as $p - 1$ contains other prime factors besides 2, we arrive at higher equations, namely, to one or more cubic equations if 3 enters once or oftener as a factor of $p - 1$, to equations of the fifth degree if $p - 1$ is divisible by 5, etc. And we can prove with all rigor that these higher equations cannot be avoided or made to depend upon equations of lower degree; and although the limits of this work do not permit us to give the demonstration here, we still thought it necessary to note this fact in order that one should not seek to construct other polygons [of prime number of sides] than those given by our theory, as for example, polygons of 7, 11, 13, 19 sides, and so employ his time in vain."

Gauss then considers (Art. 366) polygons of any number of sides n and asserts that a regular polygon of n sides is constructible if and only if

---

[3] In a prime of the form $2^{\mu} + 1$, μ is necessarily of the form 2^h, but $2^{2^h} + 1$ is not necessarily prime.

$n = 2^l p_1 \cdot p_2 \cdots p_n$, where $p_1, p_2, \ldots, p_n$ are distinct primes of the form $2^{2^h} + 1$ and where l is any positive integer or 0. The sufficiency of this condition does follow readily from Gauss's work on polygons of a prime number of sides but the necessity is not at all obvious and was not proven by Gauss.[4]

The construction of regular polygons had interested Gauss since 1796 when he conceived the first proof that the 17-sided polygon is constructible. There is a story about this discovery which may bear repeating. This construction problem was already famous. One day Gauss approached his professor A. G. Kästner at the University of Göttingen with the proof that this polygon is constructible. Kästner was incredulous and sought to dismiss Gauss, much as university teachers today dismiss angle-trisectors. Rather than take the time to examine Gauss's proof and find the supposed error in it, Kästner told Gauss the construction was unimportant because practical constructions were known. Of course Kästner knew that the existence of practical or approximate constructions was irrelevant for the theoretical problem. To interest Kästner in his proof Gauss pointed out that he had solved a seventeenth degree algebraic equation. Kästner replied that the solution was impossible. But Gauss rejoined that he had reduced the problem to solving an equation of lower degree. "Oh well," scoffed Kästner, "I have already done this." Later Gauss repaid Kästner, who also prided himself on his poetry, by lauding Kästner as the best poet among mathematicians and the best mathematician among poets.

3. *Abel's Work on the Solution of Equations by Radicals*

Abel read Lagrange's and Gauss's work on the theory of equations and while still a student in high school tackled the problem of the solvability of higher degree equations by following Gauss's treatment of the binomial equation. At first Abel thought he had solved the general fifth degree equation by radicals. But soon convinced of his error he tried to prove that such a solution was not possible (1824–26). First he succeeded in proving the theorem: The roots of an equation solvable by radicals can be given such a form that each of the radicals occurring in the expressions for the roots is expressible as a rational function of the roots of the equation and certain roots of unity. Abel then used this theorem to prove[5] the impossibility of solving by radicals the general equation of degree greater than four.

Abel's proof, done in ignorance of Ruffini's work (Chap. 25, sec. 2), is

---

[4] See James Pierpont, "On an Undemonstrated Theorem of the *Disquisitiones Arithmeticae*," *Amer. Math. Soc. Bull.*, 2, 1895–96, 77–83. This article gives the proof. The fact that the Gauss condition is necessary was first proved by Pierre L. Wantzel (1814–48), *Jour. de Math.*, 2, 1837, 366–72.

[5] *Jour. für Math.*, 1, 1826, 65–84 = *Œuvres*, 1, 66–94.

-256- - -* ! - / + ! ! ! 9
2 " 9 % *L " H A 3 J H 2 / 1"
7! %
(% $ 9 9 3

34

The Theory of Numbers in the Nineteenth Century

> It is true that Fourier had the opinion that the principal object of mathematics was public use and the explanation of natural phenomena; but a philosopher like him ought to know that the sole object of the science is the honor of the human spirit and that under this view a problem of [the theory of] numbers is worth as much as a problem on the system of the world.
>
> C. G. J. JACOBI

1. *Introduction*

Up to the nineteenth century the theory of numbers was a series of isolated though often brilliant results. A new era began with Gauss's *Disquisitiones Arithmeticae*[1] which he composed at the age of twenty. This great work had been sent to the French Academy in 1800 and was rejected but Gauss published it on his own. In this book he standardized the notation; he systematized the existing theory and extended it; and he classified the problems to be studied and the known methods of attack and introduced new methods. In Gauss's work on the theory of numbers there are three main ideas: the theory of congruences, the introduction of algebraic numbers, and the theory of forms as the leading idea in Diophantine analysis. This work not only began the modern theory of numbers but determined the directions of work in the subject up to the present time. The *Disquisitiones* is difficult to read but Dirichlet expounded it.

Another major nineteenth-century development is analytic number theory, which uses analysis in addition to algebra to treat problems involving the integers. The leaders in this innovation were Dirichlet and Riemann.

2. *The Theory of Congruences*

Though the notion of congruence did not originate with Gauss—it appears in the work of Euler, Lagrange, and Legendre—Gauss introduced the notation in the first section of *Disquisitiones* and used it systematically thereafter. The basic idea is simple. The number 27 is congruent to 3 modulo 4,

---

[1] Published 1801 = *Werke*, 1.

$$27 \equiv 3 \text{ modulo } 4,$$

because $27 - 3$ is exactly divisible by 4. (The word modulo is often abbreviated to mod.) In general, if a, b, and m are integers

$$a \equiv b \text{ modulo } m$$

if $a - b$ is (exactly) divisible by m or if a and b have the same remainders on division by m. Then b is said to be a residue of a modulo m and a is a residue of b modulo m. As Gauss shows, all the residues of a modulo m, for fixed a and m, are given by $a + km$ where $k = 0, \pm 1, \pm 2, \ldots$.

Congruences with respect to the same modulus can be treated to some extent like equations. Such congruences can be added, subtracted, and multiplied. One can also ask for the solution of congruences involving unknowns. Thus, what values of x satisfy

$$2x \equiv 25 \text{ modulo } 12?$$

This equation has no solutions because $2x$ is even and $2x - 25$ is odd. Hence $2x - 25$ cannot be a multiple of 12. The basic theorem on polynomial congruences, which Gauss re-proves in the second section, had already been established by Lagrange.[2] A congruence of the nth degree

$$Ax^n + Bx^{n-1} + \cdots + Mx + N \equiv 0 \text{ modulo } p$$

whose modulus is a prime number p which does not divide A cannot have more than n noncongruent roots.

In the third section Gauss takes up residues of powers. Here he gives a proof in terms of congruences of Fermat's minor theorem, which, stated in terms of congruences, reads: If p is a prime and a is not a multiple of p then

$$a^{p-1} \equiv 1 \text{ modulo } p.$$

The theorem follows from his study of congruences of higher degree, namely,

$$x^n \equiv a \text{ modulo } m$$

where a and m are relatively prime. This subject was continued by many men after Gauss.

The fourth section of *Disquisitiones* treats quadratic residues. If p is a prime and a is not a multiple of p and if there exists an x such that $x^2 \equiv a$ mod p, then a is a quadratic residue of p; otherwise a is a quadratic nonresidue of p. After proving some subordinate theorems on quadratic residues Gauss gave the first rigorous proof of the law of quadratic reciprocity (Chap.

---

[2] *Hist. de l'Acad. de Berlin*, 24, 1768, 192 ff., pub. 1770 = *Œuvres*, 2, 655–726.

25, sec. 4). Euler had given a complete statement much like Gauss's in one paper of his *Opuscula Analytica* of 1783 (Chap. 25, sec. 4). Nevertheless in article 151 of his *Disquisitiones* Gauss says that no one had presented the theorem in as simple a form as he had. He refers to other work of Euler including another paper in the *Opuscula* and to Legendre's work of 1785. Of these papers Gauss says rightly that the proofs were incomplete.

Gauss is supposed to have discovered a proof of the law in 1796 when he was nineteen. He gave another proof in the *Disquisitiones* and later published four others. Among his unpublished papers two more were found. Gauss says that he sought many proofs because he wished to find one that could be used to establish the biquadratic reciprocity theorem (see below). The law of quadratic reciprocity, which Gauss called the gem of arithmetic, is a basic result on congruences. After Gauss gave his proofs, more than fifty others were given by later mathematicians.

Gauss also treated congruences of polynomials. If A and B are two polynomials in x with, say, real coefficients then one knows that one can find unique polynomials Q and R such that

$$A = B \cdot Q + R,$$

where the degree of R is less than the degree of B. One can then say that two polynomials A_1 and A_2 are congruent modulo a third polynomial P if they have the same remainder R on division by P.

Cauchy used this idea[3] to define complex numbers by polynomial congruences. Thus if $f(x)$ is a polynomial with real coefficients then under division by $x^2 + 1$

$$f(x) \equiv a + bx \bmod x^2 + 1$$

because the remainder is of lower degree than the divisor. Here a and b are necessarily real by virtue of the division process. If $g(x)$ is another such polynomial then

$$g(x) \equiv c + dx \bmod x^2 + 1.$$

Cauchy now points out that if A_1, A_2, and B are any polynomials and if

$$A_1 = BQ_1 + R_1 \quad \text{and} \quad A_2 = BQ_2 + R_2,$$

then

$$A_1 + A_2 \equiv R_1 + R_2 \bmod B, \quad \text{and} \quad A_1A_2 \equiv R_1R_2 \bmod B.$$

We can now see readily that

$$f(x) + g(x) \equiv (a + c) + (b + d)x \bmod x^2 + 1$$

---

[3] *Exercices d'analyse et de physique mathématique*, 4, 1847, 84 ff. = *Œuvres*, (1), 10, 312–23 and (2), 14, 93–120.

and since $x^2 \equiv -1 \bmod x^2 + 1$ that

$$f(x)g(x) \equiv (ac - bd) + (ad + bc)x \bmod x^2 + 1.$$

Thus the numbers $a + bx$ and $c + dx$ combine like complex numbers; that is, they have the formal properties of complex numbers, x taking the place of i. Cauchy also proved that every polynomial $g(x)$ not congruent to 0 modulo $x^2 + 1$ has an inverse, that is, a polynomial $h(x)$ such that $h(x)g(x)$ is congruent to 1 modulo $x^2 + 1$.

Cauchy did introduce i for x, i being for him a real indeterminate quantity. He then showed that for any

$$f(i) = a_0 + a_1 i + a_2 i^2 + \cdots$$

that

$$f(i) \equiv a_0 - a_2 + a_4 - \cdots + (a_1 - a_3 + a_5 - \cdots)i \text{ modulo } i^2 + 1.$$

Hence any expression involving complex numbers behaves as one of the form $c + di$ and one has all the apparatus needed to work with complex expressions. For Cauchy, then, the polynomials in i, with his understanding about i, take the place of complex numbers and one can put into one class all those polynomials having the same residue modulo $i^2 + 1$. These classes are the complex numbers.

It is interesting that in 1847 Cauchy still had misgivings about $\sqrt{-1}$. He says, "In the theory of algebraic equivalences substituted for the theory of imaginary numbers the letter i ceases to represent the symbolic sign $\sqrt{-1}$, which we repudiate completely and which we can abandon without regret since one does not know what this supposed sign signifies nor what sense to attribute to it. On the contrary we represent by the letter i a real quantity but indeterminate and in substituting the sign $\equiv$ for $=$ we transform what has been called an imaginary equation into an algebraic equivalence relative to the variable i and to the divisor $i^2 + 1$. Since this divisor remains the same in all the formulas one can dispense with writing it."

In the second decade of the century Gauss proceeded to search for reciprocity laws applicable to congruences of higher degree. These laws again involve residues of congruences. Thus for the congruence

$$x^4 \equiv q \bmod p$$

one can define q as a biquadratic residue of p if there is an integral value of x satisfying the equation. He did arrive at a law of biquadratic reciprocity (see below) and a law of cubic reciprocity. Much of this work appeared in papers from 1808 to 1817 and the theorem proper on biquadratic residues was given in papers of 1828 and 1832.[4]

---

[4] *Comm. Soc. Gott.*, 6, 1828, and 7, 1832 = *Werke*, 2, 65–92 and 93–148; also pp. 165–78.

To attain elegance and simplicity in his theory of cubic and biquadratic residues Gauss made use of complex integers, that is, numbers of the form $a + bi$ with a and b integral or 0. In Gauss's work on biquadratic residues it was necessary to consider the case where the modulus p is a prime of the form $4n + 1$ and Gauss needed the complex factors into which prime numbers of the form $4n + 1$ can be decomposed. To obtain these Gauss realized that one must go beyond the domain of the ordinary integers to introduce the complex integers. Though Euler and Lagrange had introduced such integers into the theory of numbers it was Gauss who established their importance.

Whereas in the ordinary theory of integers the units are $+1$ and -1 in Gauss's theory of complex integers the units are ± 1 and $\pm i$. A complex integer is called composite if it is the product of two such integers neither of which is a unit. If such a decomposition is not possible the integer is called a prime. Thus $5 = (1 + 2i)(1 - 2i)$ and so is composite, whereas 3 is a complex prime.

Gauss showed that complex integers have essentially the same properties as ordinary integers. Euclid had proved (Chap. 4, sec. 7) that every integer is uniquely decomposable into a product of primes. Gauss proved that this unique decomposition, which is often referred to as the fundamental theorem of arithmetic, holds also for complex integers provided we do not regard the four unit numbers as different factors. That is, if $a = bc = (ib)(-ic)$, the two decompositions are the same. Gauss also showed that Euclid's process for finding the greatest common divisor of two integers is applicable to the complex integers.

Many theorems for ordinary primes carry over to the complex primes. Thus Fermat's theorem carries over in the form: If p be a complex prime $a + bi$ and k any complex integer not divisible by p then

$$k^{Np-1} \equiv 1 \text{ modulo } p$$

where Np is the norm $(a^2 + b^2)$ of p. There is also a law of quadratic reciprocity for complex integers, which was stated by Gauss in his 1828 paper.

In terms of complex integers Gauss was able to state the law of biquadratic reciprocity rather simply. One defines an uneven integer as one not divisible by $1 + i$. A primary uneven integer is an uneven integer $a + bi$ such that b is even and $a + b - 1$ is even. Thus -7 and $-5 + 2i$ are primary uneven numbers. The law of reciprocity for biquadratic residues states that if α and β are two primary uneven primes and A and B are their norms, then

$$\left(\frac{\alpha}{\beta}\right)_4 = (-1)^{(1/4)(A-1)(1/4)(B-1)}\left(\frac{\beta}{\alpha}\right)_4.$$

The symbol $(\alpha/\beta)_4$ has the following meaning: If p is any complex prime and k is any biquadratic residue not divisible by p, then $(k/p)_4$ is the power i^e of i which satisfies the congruence

$$k^{(Np-1)/4} \equiv 1 \text{ modulo } p$$

wherein Np stands for the norm of p. This law is equivalent to the statement: The biquadratic characters of two primary uneven prime numbers with respect to one another are identical, that is, $(\alpha/\beta)_4 = (\beta/\alpha)_4$, if either of the primes is congruent to 1 modulo 4; but if neither of the primes satisfies the congruence, then the two biquadratic characters are opposite, that is, $(\alpha/\beta)_4 = -(\beta/\alpha)_4$.

Gauss stated this reciprocity theorem but did not publish his proof. This was given by Jacobi in his lectures at Königsberg in 1836–37. Ferdinand Gotthold Eisenstein (1823–52), a pupil of Gauss, published five proofs of the law, of which the first two appeared in 1844.[5]

For cubic reciprocity Gauss found that he could obtain a law by using the "integers" $a + b\rho$ where ρ is a root of $x^2 + x + 1 = 0$ and a and b are ordinary (rational) integers but Gauss did not publish this result. It was found in his papers after his death. The law of cubic reciprocity was first stated by Jacobi[6] and proved by him in his lectures at Königsberg. The first published proof is due to Eisenstein.[7] Upon noting this proof Jacobi claimed[8] that it was precisely the one given in his lectures but Eisenstein indignantly denied any plagiarism.[9] There are also reciprocity laws for congruences of degree greater than four.

3. *Algebraic Numbers*

The theory of complex integers is a step in the direction of a vast subject, the theory of algebraic numbers. Neither Euler nor Lagrange envisioned the rich possibilities which their work on complex integers opened up. Neither did Gauss.

The theory grew out of the attempts to prove Fermat's assertion about $x^n + y^n = z^n$. The cases $n = 3$, 4, and 5 have already been discussed (Chap. 25, sec. 4). Gauss tried to prove the assertion for $n = 7$ but failed. Perhaps because he was disgusted with his failure, he said in a letter of 1816 to Heinrich W. M. Olbers (1758–1840), "I confess indeed that Fermat's theorem as an isolated proposition has little interest for me, since a multitude of such propositions, which one can neither prove nor refute, can easily be formulated." This particular case of $n = 7$ was disposed of by Lamé in 1839,[10] and Dirichlet established the assertion for $n = 14$.[11] However, the general proposition was unproven.

---

[5] *Jour. für Math.*, 28, 1844, 53–67 and 223–45.

[6] *Jour. für Math.*, 2, 1827, 66–69 = *Werke*, 6, 233–37.

[7] *Jour. für Math.*, 27, 1844, 289–310.

[8] *Jour. für Math.*, 30, 1846, 166–82, p. 172 = *Werke*, 6, 254–74.

[9] *Jour. für Math.*, 35, 1847, 135–274 (p. 273).

It was taken up by Ernst Eduard Kummer (1810–93), who turned from theology to mathematics, became a pupil of Gauss and Dirichlet, and later served as a professor at Breslau and Berlin. Though Kummer's major work was in the theory of numbers, he made beautiful discoveries in geometry which had their origin in optical problems; he also made important contributions to the study of refraction of light by the atmosphere.

Kummer took $x^p + y^p$ where p is prime, and factored it into

$$(x + y)(x + \alpha y)\cdots(x + \alpha^{p-1}y),$$

where α is an imaginary pth root of unity. That is, α is a root of

$$(1) \qquad \alpha^{p-1} + \alpha^{p-2} + \cdots + \alpha + 1 = 0.$$

This led him to extend Gauss's theory of complex integers to algebraic numbers insofar as they are introduced by equations such as (1), that is, numbers of the form

$$f(\alpha) = a_0 + a_1\alpha + \cdots + a_{p-2}\alpha^{p-2},$$

where each a_i is an ordinary (rational) integer. (Since α satisfies (1), terms in α^{p-1} can be replaced by terms of lower power.) Kummer called the numbers $f(\alpha)$ complex integers.

By 1843 Kummer made appropriate definitions of integer, prime integer, divisibility, and the like (we shall give the standard definitions in a moment) and then made the mistake of assuming that unique factorization holds in the class of algebraic numbers that he had introduced. He pointed out while transmitting his manuscript to Dirichlet in 1843 that this assumption was necessary to prove Fermat's theorem. Dirichlet informed him that unique factorization holds only for certain primes p. Incidentally, Cauchy and Lamé made the same mistake of assuming unique factorization for algebraic numbers. In 1844 Kummer[12] recognized the correctness of Dirichlet's criticism.

To restore unique factorization Kummer created a theory of ideal numbers in a series of papers starting in 1844.[13] To understand his idea let us consider the domain of $a + b\sqrt{-5}$, where a and b are integers. In this domain

$$6 = 2\cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5})$$

---

[10] *Jour. de Math.*, 5, 1840, 195–211.

[11] *Jour. für Math.*, 9, 1832, 390–93 = *Werke*, 1, 189–94.

[12] *Jour. de Math.*, 12, 1847, 185–212.

[13] *Jour. für Math.*, 35, 1847, 319–26, 327–67.

and all four factors can readily be shown to be prime integers. Then unique decomposition does not hold. Let us introduce, for this domain, the ideal numbers $\alpha = \sqrt{2}$, $\beta_1 = (1 + \sqrt{-5})/\sqrt{2}$, $\beta_2 = (1 - \sqrt{-5})/\sqrt{2}$. We see that $6 = \alpha^2\beta_1\beta_2$. Thus 6 is now uniquely expressed as the product of four factors, all ideal numbers as far as the domain $a + b\sqrt{-5}$ is concerned.[14] In terms of these ideals and other primes factorization in the domain is unique (apart from factors consisting of units). With ideal numbers one can prove some of the results of ordinary number theory in all domains that previously lacked unique factorization.

Kummer's ideal numbers, though ordinary numbers, do not belong to the class of algebraic numbers he had introduced. Moreover, the ideal numbers were not defined in any general way. As far as Fermat's theorem is concerned, with his ideal numbers Kummer did succeed in showing that it was correct for a number of prime numbers. In the first hundred integers only 37, 59, and 67 were not covered by Kummer's demonstration. Then Kummer in a paper of 1857[15] extended his results to these exceptional primes. These results were further extended by Dimitry Mirimanoff (1861–1945), a professor at the University of Geneva, by perfecting Kummer's method.[16] Mirimanoff proved that Fermat's theorem is correct for each n up to 256 if x, y, and z are prime to that exponent n.

Whereas Kummer worked with algebraic numbers formed out of the roots of unity, Richard Dedekind (1831–1916), a pupil of Gauss, who spent fifty years of his life as a teacher at a technical high school in Germany, approached the problem of unique factorization in an entirely new and fresh manner. Dedekind published his results in supplement 10 to the second edition of Dirichlet's *Zahlentheorie* (1871) which Dedekind edited. He extended these results in the supplements to the third and fourth editions of the same book.[17] Therein he created the modern theory of algebraic numbers.

Dedekind's theory of algebraic numbers is a generalization of Gauss's complex integers and Kummer's algebraic numbers but the generalization is somewhat at variance with Gauss's complex integers. A number r that is a root of

$$(2) \qquad a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n = 0,$$

where the a_i's are ordinary integers (positive or negative), and that is not a root of such an equation of degree less than n is called an algebraic number of degree n. If the coefficient of the highest power of x in (2) is 1, the solutions are called algebraic integers of degree n. The sum, difference, and product of algebraic integers are algebraic integers, and if an algebraic integer is a rational number it is an ordinary integer.

---

[14] With the introduction of the ideal numbers, 2 and 3 are no longer indecomposable, for $2 = \alpha^2$ and $3 = \beta_1\beta_2$.

[15] *Abh. König. Akad. der Wiss. Berlin*, 1858, 41–74.

[16] *Jour. für Math.*, 128, 1905, 45–68.

[17] 4th ed., 1894 = *Werke*, 3, 2–222.

We should note that under the new definitions an algebraic integer can contain ordinary fractions. Thus $(-13 + \sqrt{-115})/2$ is an algebraic integer of the second degree because it is a root of $x^2 + 13x + 71 = 0$. On the other hand $(1 - \sqrt{-5})/2$ is an algebraic number of degree 2 but not an algebraic integer because it is a root of $2x^2 - 2x + 3 = 0$.

Dedekind introduced next the concept of a number field. This is a collection F of real or complex numbers such that if α and β belong to F then so do $\alpha + \beta$, $\alpha - \beta$, $\alpha\beta$ and, if $\beta \neq 0$, α/β. Every number field contains the rational numbers because if α belongs then so does α/α or 1 and consequently $1 + 1$, $1 + 2$, and so forth. It is not difficult to show that the set of all algebraic numbers forms a field.

If one starts with the rational number field and θ is an algebraic number of degree n then the set formed by combining θ with itself and the rational numbers under the four operations is also a field of degree n. This field may be described alternatively as the smallest field containing the rational numbers and θ. It is also called an extension field of the rational numbers. Such a field does not consist of all algebraic numbers and is a specific algebraic number field. The notation $R(\theta)$ is now common. Though one might expect that the members of $R(\theta)$ are the quotients $f(\theta)/g(\theta)$ where $f(x)$ and $g(x)$ are any polynomials with rational coefficients, one can prove that if θ is of degree n, then any member α of $R(\theta)$ can be expressed in the form

$$\alpha = a_0\theta^{n-1} + a_1\theta^{n-2} + \cdots + a_{n-1},$$

where the a_i are ordinary rational numbers. Moreover, there exist algebraic integers $\theta_1, \theta_2, \ldots, \theta_n$ of this field such that all the algebraic integers of the field are of the form

$$A_1\theta_1 + A_2\theta_2 + \cdots + A_n\theta_n,$$

where the A_i are ordinary positive and negative integers.

A ring, a concept introduced by Dedekind, is essentially any collection of numbers such that if α and β belong, so do $\alpha + \beta$, $\alpha - \beta$, and $\alpha\beta$. The set of all algebraic integers forms a ring as does the set of all algebraic integers of any specific algebraic number field.

The algebraic integer α is said to be divisible by the algebraic integer β if there is an algebraic integer γ such that $\alpha = \beta\gamma$. If j is an algebraic integer which divides every other integer of a field of algebraic numbers then j is called a unit in that field. These units, which include $+1$ and -1, are a generalization of the units $+1$ and -1 of ordinary number theory. The algebraic integer α is a prime if it is not zero or a unit and if any factorization of α into $\beta\gamma$, where β and γ belong to the same algebraic number field, implies that β or γ is a unit in that field.

Now let us see to what extent the fundamental theorem of arithmetic holds. In the ring of *all* algebraic integers there are no primes. Let us consider the ring of integers in a specific algebraic number field $R(\theta)$, say the field $a + b\sqrt{-5}$, where a and b are ordinary rational numbers. In this field unique factorization does not hold. For example,

$$21 = 3 \cdot 7 = (4 + \sqrt{-5})(4 - \sqrt{-5}) = (1 + 2\sqrt{-5})(1 - 2\sqrt{-5}).$$

Each of these last four factors is prime in the sense that it cannot be expressed as a product of the form $(c + d\sqrt{-5})(e + f\sqrt{-5})$ with c, d, e, and f integral.

On the other hand let us consider the field $a + b\sqrt{6}$ where a and b are ordinary rational numbers. If one applies the four algebraic operations to these numbers one gets such numbers. If a and b are restricted to integers one gets the algebraic integers (of degree 2) of this domain. In this domain we can take as an equivalent definition of unit that the algebraic integer M is a unit if $1/M$ is also an algebraic integer. Thus 1, -1, $5 - 2\sqrt{6}$, and $5 + 2\sqrt{6}$ are units. Every integer is divisible by any one of the units. Further, an algebraic integer of the domain is prime if it is divisible only by itself and the units. Now

$$6 = 2 \cdot 3 = \sqrt{6} \cdot \sqrt{6}.$$

It would seem as though there is no unique decomposition into primes. But the factors shown are not primes. In fact

$$6 = 2 \cdot 3 = \sqrt{6} \cdot \sqrt{6} = (2 + \sqrt{6})(-2 + \sqrt{6})(3 + \sqrt{6})(3 - \sqrt{6}).$$

Each of the last four factors is a prime in the domain and unique decomposition does hold in this domain.

In the ring of integers of a specific algebraic number field factorization of the algebraic integers into primes is always possible but *unique* factorization does not generally hold. In fact for domains of the form $a + b\sqrt{-D}$, where D may have any positive integral value not divisible by a square, the unique factorization theorem is valid only when $D = 1, 2, 3, 7, 11, 19, 43, 67$, and 163, at least for D's up to 10^9.[18] Thus the algebraic numbers themselves do not possess the property of unique factorization.

4. *The Ideals of Dedekind*

Having generalized the notion of algebraic number, Dedekind now undertook to restore unique factorization in algebraic number fields by a scheme quite different from Kummer's. In place of ideal numbers he introduced

---

[18] H. M. Stark has shown that the above values of D are the only ones possible. See his "On the Problem of Unique Factorization in Complex Quadratic Fields," *Proceedings of Symposia in Pure Mathematics*, XII, 41–56, Amer. Math. Soc., 1969.

(!L !3 = 3 ! " ! ! 7! 0 " !
" J "J " D A! J ! "
! !

does. The latter is not the necessary geometry of physical space; its physical truth cannot be guaranteed on any a priori grounds. This realization, which did not call for any technical mathematical development because this had already been done, was first achieved by Gauss.

Carl Friedrich Gauss (1777–1855) was the son of a mason in the German city of Brunswick and seemed destined for manual work. But the director of the school at which he received his elementary education was struck by Gauss's intelligence and called him to the attention of Duke Karl Wilhelm. The Duke sent Gauss to a higher school and then in 1795 to the University of Göttingen. Gauss now began to work hard on his ideas. At eighteen he invented the method of least squares and at nineteen he showed that the 17-sided regular polygon is constructible. These successes convinced him that he should turn from philology to mathematics. In 1798 he transferred to the University of Helmstädt and there he was noticed by Johann Friedrich Pfaff who became his teacher and friend. After finishing his doctor's degree Gauss returned to Brunswick where he wrote some of his most famous papers. This work earned for him in 1807 the appointment as professor of astronomy and director of the observatory at Göttingen. Except for one visit to Berlin to attend a scientific meeting he remained at Gottingen for the rest of his life. He did not like to teach and said so. However, he did enjoy social life, was married twice, and raised a family.

Gauss's first major work was his doctoral thesis in which he proved the fundamental theorem of algebra. In 1801 he published the classic *Disquisitiones Arithmeticae*. His mathematical work in differential geometry, the "Disquisitiones Generales circa Superficies Curvas" (General Investigations of Curved Surfaces, 1827), which, incidentally, was the outcome of his interest in surveying, geodesy, and map-making, is a mathematical landmark (Chap. 37, sec. 2). He made many other contributions to algebra, complex functions, and potential theory. In unpublished papers he recorded his innovative work in two major fields: the elliptic functions and non-Euclidean geometry.

His interests in physics were equally broad and he devoted most of his energy to them. When Giuseppe Piazzi (1746–1826) discovered the planet Ceres in 1801 Gauss undertook to determine its path. This was the beginning of his work in astronomy, the activity that absorbed him most and to which he devoted about twenty years. One of his great publications in this area is his *Theoria Motus Corporum Coelestium* (Theory of Motion of the Heavenly Bodies, 1809). Gauss also earned great distinction in his physical research on theoretical and experimental magnetism. Maxwell says in his *Electricity and Magnetism* that Gauss's studies of magnetism reconstructed the whole science, the instruments used, the methods of observation, and the calculation of results. Gauss's papers on terrestrial magnetism are models of physical research and supplied the best method of measuring the earth's magnetic

(%'

_ Morris Kline - Mathematical thought from Ancient - Volume 2 (1972).djvu

Pag. 424(52) -

[Gauss let $\Pi(n) = \Gamma(n+1)$].

Pág. 598(226) -

In the same paper Gauss proved that nth degree polynomial can be expressed as a product of linear and quadratic factors with real coefficients.

Pág. 609(237)

Fermat conjecture on

$$x^n + y^n = z^n$$

Euler proved this was correct for $n=3$ and $n=4$. The case $n=4$ had been proven by Frénicle de Bessy.

This work of Euler had to be completed by Lagrange, Legendre and Gauss.

Legendre then proved the conjecture for $n=5$.

Fermat had also conjectured that the formula

$$2^{2^n}+1$$

(Chapter 13, sec. 7).

yielded primes for an unspecified set of values of n.

Obs: (complitar com esta obs.)

Pág. 753(381)

Alternatively, a regular polygon can be constructed iff p is ~~prime~~ of the

$$2^{2^h}+1.$$

Obs: 2^u+1

$$u=2^h \Rightarrow 2^{2^h}=1$$

→ Ao todo, 105 occurrences (Gauss)

— Morris Kline - Mathematical Through. Volume 3 (1972). djvu

Gauss - 273 occurrences.

Pág. 813(16) 2. The Theory of Congruences.

/ (%"

Pag. 870 (72) –

Maxwell says in his Electricity and Magnetism that Gauss's studies of magnetism reconstructed the whole science, the instruments used, the methods of observation, and the calculation of results.

Obs:

Gauss – 273 occurrences

---

www.wilbourhall.org

GAUSS SURFACES

https://math.dartmouth.edu/~doyle/docs/102.2020/sources/gauss_surfaces.pdf

→ English.

Gauss and the Definition of the Plane concept in Euclidean Elementary Geometry.

20211012/GAUSS/82058304.pdf

Gauss Werke
Göttinger Digitalisierungszentrum
https://gdz.sub.uni-goettingen.de/volumes/id/PPN235957348

– 20211012/ GAUSS-2/ ⟵
Obs: Todo o diretório!!!
– Curves and surfaces.pdf
407 ocorrências (Gauss)
Pág. VIII (8) Preface
Finally, in chapter 7, we discuss a few important results about the global theory of surfaces (one of them, the Hartman-Nirenberg theorem, proved only in 1959), mainly (principalmente) focusing on the connection between the sign of Gaussian curvature and the global (topological and differential) structures of surfaces.
Obs: Livro todo!!!

/
1 ## ", D " #A # "# 8 %$/$ %'
5 4292 8 A A $,-4292 8 A A -) .
- -72H -

7! #

4292 8 A A $,-4292 8 A A -) .
- -72H -

! -256- - -* ! - / A ! " ! 1"
*D 7! '
(&&& !
< ! 3J !1 J " ! 1 ! !D D! 3 !

of surfaces, from the (modern) definition of a surface to Gauss's *Theorema Egregium*. Chapter 5 studies geodesics on a surface, up to the proof of their existence and local uniqueness.

The remaining chapters (and some of the chapters' supplementary material; see below) are devoted instead to global properties. Chapter 2 discusses some fundamental results in the global theory of plane curves, including the *Jordan curve theorem*, which is a good example of the relatively belated development of the interest in global properties of curves: the statement of the Jordan curve theorem seems obvious until you try to actually prove it — and then it turns out to be surprisingly difficult, and far deeper than expected.

Chapter 6 is devoted to the Gauss-Bonnet theorem, with a complete proof is given, and to some of its innumerable applications. Finally, in Chapter 7, we discuss a few important results about the global theory of surfaces (one of them, the Hartman-Nirenberg theorem, proved only in 1959), mainly focusing on the connections between the sign of Gaussian curvature and the global (topological and differential) structure of surfaces. We shall confine ourselves to curves and surfaces in the plane and in space; but the language and the methods we shall introduce are compatible with those used in differential geometry of manifolds of any dimension, and so this book can be useful as a training field before attacking n-dimensional manifolds.

We have attempted to provide several possible paths for using this book as a textbook. A minimal path consists in just dealing with local theory: Chapters 1, 3, and 4 can be read independently of the remaining ones, and can be used to offer, in a two-month course, a complete route from initial definitions to Gauss's Theorema Egregium; if time permits, the first two sections of Chapter 5 can be added, describing the main properties of geodesics. Such a course is suitable and useful both for Mathematics (of course) and Physics students, from the second year on, and for Engineering and Computer Science students (possibly for a Master degree) who need (or are interested in) more advanced mathematical techniques than those learnt in a first-level degree.

In a one-semester course, on the other hand, it is possible to adequately cover the global theory too, introducing Chapter 2 about curves, the final section of Chapter 5 about vector fields, and, at the professor's discretion, Chapter 6 about the Gauss-Bonnet theorem or Chapter 7 about the classification of closed surfaces with constant Gaussian curvature. In our experience, in a one-semester course given to beginning or intermediate undergraduate Mathematics or Physics students, it is difficult to find time to cover both topics, so the two chapters are completely independent; but in the case of an one-year course, or in an one-semester course for advanced students, it might be possible to do so, possibly even touching upon some of the supplementary material (see below).

Each chapter comes with several guided problems, that is, solved exercises, both computational (one of the nice features of local differential geometry of curves and surfaces is that almost everything can be explicitly computed — well, with the exception of geodesics) and of a more theoretical character,

D -256- - -* ! - / ! I ! 1"

- 20211012/GAUSS-2/
Differential Geometry of Curves and Surfaces.pdf
Pág. 160(178) The Gauss Map
Pág. 189(207) Gaussian and Mean Curvature.

Pág. 153(171) Figure 6.3
≅ Gravity.

- 20211012/GAUSS-2/Remmert-Theory.pdf
⇒ Theory of Complex Functions.

- 20211012/Measure and Integration theory/Measure and Integration (2002).pdf

- Outra cópia:

- 20211012/GAUSS-2/
Burckel-Classical Analysis in the Complex Plane...
(2021).pdf.

Obs: Livro todo!!!

Os diretórios GAUSS and GAUSS-2 deverão ser explorados em detalhes.

Source: Data-Nova/Vivo-a20-2021/Books-a20-2021/20211012/

/
! -256- - -* ! - / 4 ! 7 3 A ! "
! 1"
(B 9 7! !1

6.2 The Gauss Map

In Section 2.2, given any closed, simple, regular curve $\gamma : I \to \mathbb{R}^2$ in the plane, we defined the tangential indicatrix to be the curve given by $\vec{T} : I \to \mathbb{R}^2$, the unit tangent vector of $\gamma(t)$. The image of $\vec{T}$ lies on the unit circle, and though $\vec{T}$ is a closed curve, it need not be regular as its locus may stop and double back. Regardless of the parametrization, the tangential indicatrix is well defined up to a change in sign.

On any regular surface S in $\mathbb{R}^3$, the tangent plane T_pS at a point $p \in S$ is a two-dimensional subspace of $\mathbb{R}^2$, and hence the vectors that are normal to S at p form a one-dimensional subspace. Thus, there exist exactly two possible choices for a unit normal vector. Proposition 5.5.4 implies that if the surface is orientable, we can specify its orientation by a continuous function $n : S \to \mathbb{S}^2$, where $\mathbb{S}^2$ means the unit sphere in $\mathbb{R}^3$.

Definition 6.2.1. Let S be an oriented regular surface in $\mathbb{R}^3$ with an orientation $n : S \to \mathbb{S}^2$. In the classical theory of surfaces, the function n is also called the *Gauss map*.

It is important to remain aware of the distinction between the functions n and $\vec{N}$. Let S be an oriented surface with orientation $n : S \to \mathbb{S}^2$. Suppose that $\vec{X} : U \to \mathbb{R}^3$, where U is an open subset of $\mathbb{R}^2$, parametrizes a neighborhood of S, then the function $\vec{N} : U \to \mathbb{S}^2$ is defined in reference to $\vec{X}$ by Equation (5.9). The functions n and $\vec{N}$ are related by the fact that if $\vec{X}$ is a positively oriented parametrization, then

$$\vec{N} = n \circ \vec{X}.$$

Example 6.2.2. Consider as an example a sphere S in $\mathbb{R}^3$ equipped with the orientation n, with the unit vectors normal to S pointing away from the center $\vec{c}$ of the sphere. The sphere can be given as the solution to the equation

$$\|\vec{x} - \vec{c}\|^2 = R^2. \tag{6.10}$$

Recall that $\|\vec{v}\|^2 = \vec{v} \cdot \vec{v}$. If $\vec{x}(t)$ is any curve on the sphere, then by differentiating the relationship in Equation (6.10), one obtains

$$\vec{x}\,'(t) \cdot (\vec{x}(t) - \vec{c}) = 0.$$

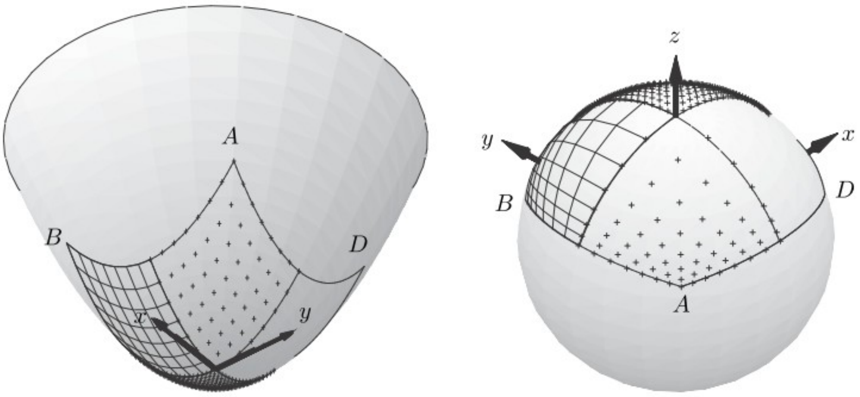

(a) Elliptic paraboloid (b) Unit sphere

Figure 6.4. Gauss map on the elliptic paraboloid.

The tangent plane to the sphere S at a point $\vec{p}$ consists of all possible vectors $\vec{x}'(t_0)$, where $\vec{x}(t)$ is a curve on S, with $\vec{x}(t_0) = \vec{p}$. Therefore, at any point $\vec{p}$ on the sphere, there are two options for unit normal vectors:

$$\frac{\vec{p}-\vec{c}}{\|\vec{p}-\vec{c}\|} \qquad \text{or} \qquad -\frac{\vec{p}-\vec{c}}{\|\vec{p}-\vec{c}\|}$$

However, it is the former that provides the outward-pointing orientation n. Thus the Gauss map for the sphere of radius R and center $\vec{c}$ is explicitly

$$n(\vec{p}) = \frac{\vec{p}-\vec{c}}{R}.$$

Furthermore, if S is itself the unit sphere centered at the origin, then the Gauss map is the identity function.

Example 6.2.3. Consider the elliptic paraboloid S defined by the equation $z = x^2 + y^2$. This is an orientable surface, and suppose it is oriented with the unit normal always pointing in a positive z-direction. (Figure 6.4 shows this is possible.) Using the function $F(x, y, z) = z - x^2 - y^2$, the elliptic paraboloid is given by the equation $F(x, y, z) = 0$. By Proposition 5.2.13, for all $(x, y, z) \in S$, the gradient $\vec{\nabla} F(x,y,z)$ is a normal vector, so (since the z-direction is positive) the Gauss map is

$$n(x,y,z) = \frac{(-2x,-2y,1)}{\sqrt{4x^2+4y^2+1}}.$$

Figure 6.4 shows how the Gauss map acts on the patch

$$\{(x,y,z) \in S \mid -1 \leq x \leq 1 \text{ and } -1 \leq y \leq 1\}.$$

It is not hard to show that the image of the Gauss map on the unit sphere is the upper hemisphere. Indeed, using cylindrical coordinates, the Gauss map is

$$n(r,\theta,z) = \frac{(-2r\cos\theta, -2r\sin\theta, 1)}{\sqrt{4r^2+1}}.$$

The function $f(r) = 1/\sqrt{4r^2+1}$ is a decreasing bijection from $[0,+\infty)$ to $(0,1]$. Thus, for any $z_0 \in (0,1]$, there exists a unique $r_0 \in [0,+\infty)$ with $f(r_0) = z_0$. Then, with $\theta \in [0,2\pi]$, the image of $n(r_0,\theta,z_0)$ is a circle on the unit sphere at height z_0. Thus, the image of n is

$$n(S) = \{(x,y,z) \in \mathbb{S}^2 \mid z > 0\}.$$

Example 6.2.4. In contrast to Example 6.2.3, consider the hyperbolic paraboloid given by the parametrization $\vec{X} : \mathbb{R}^2 \to \mathbb{R}^3$, with

$$\vec{X}(u,v) = (u,v,u^2-v^2).$$

We calculate

$$\vec{X}_u = (1,0,2u) \qquad \text{and} \qquad \vec{X}_v = (0,1,-2v),$$

and thus,

$$\vec{N} = \frac{\vec{X}_u \times \vec{X}_v}{\|\vec{X}_u \times \vec{X}_v\|} = \frac{(-2u,2v,1)}{\sqrt{4u^2+4v^2+1}}.$$

Like Figure 6.4, Figure 6.5 shows the Gauss map for a “square” on the hyperbolic paraboloid. These two examples manifest a central behavior of the Gauss map. On the elliptic paraboloid, the Gauss

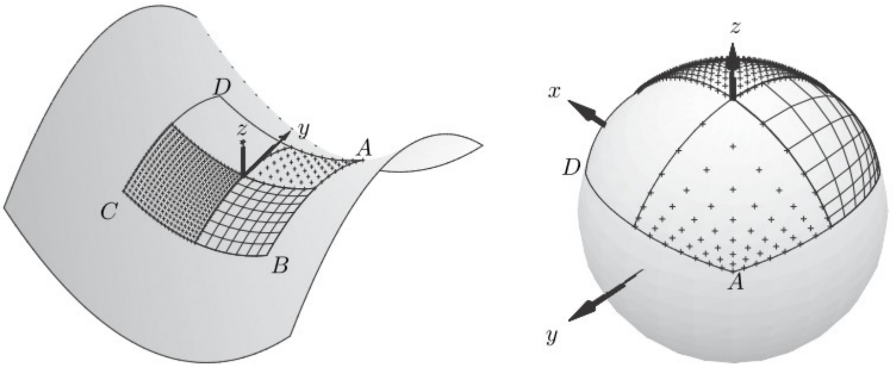

(a) Hyperbolic paraboloid (b) Unit sphere

Figure 6.5. Gauss map on the hyperbolic paraboloid.

map preserves the orientation of the square in the sense that if one travels along the boundary in a clockwise sense on the surface, one also travels along the boundary of the image of the square mapped by n in a clockwise sense. On the other hand, with the hyperbolic paraboloid, the Gauss map reverses the orientation of the square.

Proposition 6.2.5. *Let S be a regular surface in $\mathbb{R}^3$, and let $\vec{X} : U \to \mathbb{R}^3$ be a parametrization of a coordinate patch $V = \vec{X}(U)$, where U is an open set in $\mathbb{R}^2$. Define the vector function $\vec{N} : U \to \mathbb{R}^3$ by*

$$\vec{N} = \frac{\vec{X}_u \times \vec{X}_v}{\|\vec{X}_u \times \vec{X}_v\|}.$$

If $\vec{X}$ is of class C^r, then $\vec{N}$ is of class C^{r-1}.

Proof: By Problem 3.1.10, we deduce that if $\vec{F} : U \to \mathbb{R}^3$, $\vec{G} : U \to \mathbb{R}^3$, and $h : U \to \mathbb{R}$ are of class C^1, then the three functions $\vec{F} \cdot \vec{G} : U \to \mathbb{R}$, $\vec{F} \times \vec{G} : U \to \mathbb{R}^3$, and $h\vec{F} : U \to \mathbb{R}^3$ are also of class C^1, and their partial derivatives follow appropriate product rules. By definition,

$$\vec{N} = \frac{\vec{X}_u \times \vec{X}_v}{\sqrt{(\vec{X}_u \times \vec{X}_v) \cdot (\vec{X}_u \times \vec{X}_v)}}.$$

One can see that any partial derivative of any order of $\vec{N}$ will involve the partial derivative of a combination of the above three types of products and the derivative of an expression of the form

$$f(u,v) = \Big((\vec{X}_u \times \vec{X}_v) \cdot (\vec{X}_u \times \vec{X}_v) \Big)^{-k/2},$$

where k is an odd positive integer. Since S is regular, $\|\vec{X}_u \times \vec{X}_v\|$ is never 0, so a particular higher derivative of $\vec{N}$ will exist if that higher derivative exists for $\vec{X}_u$ and for $\vec{X}_v$. The proposition follows. □

One would be correct to think that the different behaviors in Figures 6.4 and 6.5 can be illustrated and quantified by a function between tangent spaces to S and $\mathbb{S}^2$, respectively. This perspective is encapsulated in the notion of the differential of functions between regular surfaces. We develop the more general theory of differentials of functions between manifolds in Chapter 3 of [22]. However, the differential of the Gauss map, though only a particular case of the differentials of functions between surfaces, serves a central role in the rest of this chapter.

Let S be a regular oriented surface with orientation n, and let $\vec{X} : U \to \mathbb{R}^3$ be a regular positively oriented parametrization of a coordinate patch $\vec{X}(U)$ of S. Suppose also that S is of class C^2, which, according to Proposition 6.2.5, implies that $\vec{N}$ is of class C^1. Since $\vec{N} \cdot \vec{N} = 1$ for all $(u,v) \in U$, we have

$$\vec{N} \cdot \vec{N}_u = 0 \qquad \text{and} \qquad \vec{N} \cdot \vec{N}_v = 0 \tag{6.11}$$

for all $(u,v) \in U$. Therefore, both derivatives $\vec{N}_u$ and $\vec{N}_v$ are vector functions such that $\vec{N}_u(q)$ and $\vec{N}_v(q)$ are in T_pS when $\vec{X}(q) = p$.

The vector function $\vec{N} : U \to \mathbb{R}^3$ is itself a parametrized surface, with its image lying in the unit sphere, though not necessarily giving a regular parametrization of $\mathbb{S}^2$. The simple fact in Equation (6.11) indicates that if $\vec{N}$ admits a tangent plane at $q \in U$, then

$$T_pS = T_{\vec{N}(q)}(\mathbb{S}^2) \tag{6.12}$$

as subspaces of $\mathbb{R}^3$. In other words, the sets of vectors $\{\vec{X}_u, \vec{X}_v\}$ and $\{\vec{N}_u, \vec{N}_v\}$ span the same subspace T_pS.

The differential of the Gauss map at a point $p \in S$ is a linear transformation $dn_p : T_pS \to T_{n(p)}(\mathbb{S}^2)$, but by virtue of Equation (6.12), one identifies it as a linear transformation $dn_p : T_pS \to T_pS$. For all $\vec{v} \in T_pS$, one defines $dn_p(\vec{v}) = \vec{w}$ if there exists a curve $\vec{\alpha} : I \to U$ such that

$$\begin{cases} \vec{\alpha}(0) = q, \text{ with } \vec{X}(q) = p, \\ \dfrac{d}{dt}\left(\vec{X} \circ \vec{\alpha}\right)(t)\big|_{t=0} = \vec{v}, \text{ and} \\ \dfrac{d}{dt}\left(\vec{N} \circ \vec{\alpha}\right)(t)\big|_{t=0} = \vec{w}. \end{cases} \tag{6.13}$$

In other words, by writing $\vec{X}(t) = \vec{X} \circ \vec{\alpha}(t)$ and $\vec{N}(t) = \vec{N} \circ \vec{\alpha}(t)$, the differential of the Gauss map satisfies $dn_p(\vec{X}'(0)) = \vec{N}'(0)$.

Using coordinate lines $\vec{X}(t, v_0)$ or $\vec{X}(u_0, t)$ for the curve $\vec{\alpha}(t)$ in the above definition, it is easy to see that

$$\begin{aligned} dn_p(\vec{X}_u) &= \vec{N}_u, \\ dn_p(\vec{X}_v) &= \vec{N}_v. \end{aligned} \tag{6.14}$$

Though the Gauss map $n : S \to \mathbb{S}^2$ is independent of any coordinate systems on S, the differential dn_p requires reference to a regular positively oriented parametrization of a neighborhood of p. However, though different parametrizations of a neighborhood of p induce different coordinate bases on T_pS, the definition in Equation 6.13 remains unchanged and consequently, as a linear transformation of T_pS to itself, is independent of the parametrization.

Problems

6.2.1. Describe the region of the unit sphere covered by the image of the Gauss map for the following surfaces:

(a) The hyperbolic paraboloid given by $z = x^2 - y^2$.
(b) The hyperboloid of one sheet given by $x^2 + y^2 - z^2 = 1$.
(c) The cone with opening angle α given by $z^2 \tan^2 \alpha = x^2 + y^2$.
(d) The right circular cylinder $x^2 + y^2 = R^2$, with R a constant.

6.2.2. What regular surface S has a single point on $\mathbb{S}^2$ as the image of the Gauss map?

6.5 Gaussian and Mean Curvature

We now arrive at two fundamental geometric invariants that encapsulate a considerable amount of useful information about the local shape of a surface. Again, in this section, we must assume that S is a regular surface of class C^2.

Definition 6.5.1. Let κ_1 and κ_2 be the principal curvatures of a regular surface S at a point p. Define

1. the *Gaussian curvature* of S at p as the product $K = \kappa_1 \kappa_2$;
2. the *mean curvature* of S at p as the average $H = \frac{\kappa_1+\kappa_2}{2}$ of the principal curvatures.

By Problem 6.3.6, the matrix for the Gauss map in terms of any particular coordinate system on a neighborhood of p on S is conjugate to the corresponding matrix for a different coordinate system via the change of basis matrix $\left(\frac{\partial x_i}{\partial \bar{x}_j}\right)$. Therefore, since $\det(BAB^{-1}) = \det(A)$ and $\operatorname{Tr}(BAB^{-1}) = \operatorname{Tr}(A)$, which is a standard result in linear algebra, the eigenvalues of the Gauss map, the principal curvatures, the Gaussian curvature, and the mean curvature are invariant under coordinate changes. Furthermore, since $\det(-A) = \det(A)$ when A is a square matrix with an even number of rows, then

$$\begin{aligned} K &= \kappa_1 \kappa_2 = \det(a_i^j) = \det(dn_p), \\ H &= \frac{\kappa_1 + \kappa_2}{2} = -\frac{1}{2}\operatorname{Tr}(a_i^j) = -\frac{1}{2}\operatorname{Tr}(dn_p). \end{aligned}$$

Consequently, as claimed above, the Gaussian curvature and the mean curvature of S at p are geometric invariants. This means that they do not depend on the orientation or position of S in space, and they do not depend on any particular coordinate system on S in a neighborhood of the point p.

In order to calculate the mean curvature $H(u,v)$, one has no choice but to calculate the matrix of the differential of the Gauss map $(a_j^i) = [dn_p]$. However, from a computational perspective, Equation (6.26) leads to a much simpler formula for the Gaussian curvature function $K(u,v)$. Recall that $\det(AB) = \det(A)\det(B)$ for all square matrices of the same size and also that $\det(A^{-1}) = 1/\det(A)$. Then Equation (6.26) implies that

$$K = \frac{\det(L_{ij})}{\det(g_{ij})} = \frac{L_{11}L_{22} - L_{12}^2}{g_{11}g_{22} - g_{12}^2}. \tag{6.37}$$

This equation lends itself readily to calculations since one does not need to fully compute the matrix of the Gauss map, let alone find its eigenvalues. However, we can also obtain an alternate characterization of the Gaussian curvature.

Proposition 6.5.2. *Let S be a regular surface of class C^2, and let V be a neighborhood parametrized by $\vec{X} : U \subset \mathbb{R}^2 \to \mathbb{R}^3$. Define the unit normal vector $\vec{N}$ as*

$$\vec{N} = \frac{\vec{X}_u \times \vec{X}_v}{\|\vec{X}_u \times \vec{X}_v\|}.$$

Then over the domain U, the Gaussian curvature is the unique function $K(u, v)$ satisfying

$$\vec{N}_u \times \vec{N}_v = K(u, v)\vec{X}_u \times \vec{X}_v$$

Proof: From the definition of the (a_j^i) functions in Equation (6.22), we have

$$\vec{N}_u \times \vec{N}_v = (a_1^1\vec{X}_u + a_1^2\vec{X}_v) \times (a_2^1\vec{X}_u + a_2^2\vec{X}_v).$$

It is then easy to see that

$$\vec{N}_u \times \vec{N}_v = \det(a_j^i)\vec{X}_u \times \vec{X}_v.$$

The result follows since $K = \det(a_j^i)$. □

Example 6.5.3 (Spheres). Consider the sphere parametrized by the vector function

$$\vec{X}(u, v) = (R\cos u \sin v, R\sin u \sin v, R\cos v),$$

with $(u, v) \in (0, 2\pi) \times (0, \pi)$. Example 6.1.5 and Example 6.3.3 gave us

$$(g_{ij}) = \begin{pmatrix} R^2\sin^2 v & 0 \\ 0 & R^2 \end{pmatrix} \quad \text{and} \quad (L_{ij}) = \begin{pmatrix} -R\sin^2 v & 0 \\ 0 & -R \end{pmatrix}.$$

Using Equation (6.37), it is then easy to see that the Gaussian curvature function on the sphere is

$$K(u,v)=\frac{R^2\sin^2 v}{R^4\sin^2 v}=\frac{1}{R^2},$$

which is a constant function. (Note that we assumed that $v\neq 0,\ \pi$, which correspond to the north and south poles of the parametrization. However, another parametrization that includes the north and south poles would show that at these points as well we have $K=1/R^2$.)

The matrix of the Gauss map is then

$$[dn_p]=-Lg^{-1}=\begin{pmatrix}-\frac{1}{R} & 0\\ 0 & -\frac{1}{R}\end{pmatrix},$$

from which one immediately deduces that the principal curvatures are $\kappa_1(u,v)=\kappa_2(u,v)=\frac{1}{R}$. This shows that all points on the sphere are umbilical points. In addition, the mean curvature is also a constant function $H(u,v)=\frac{1}{R}$.

Example 6.5.4 (Function Graphs). Consider the graph of a function $f:U\subset\mathbb{R}^2\to\mathbb{R}$. The graph can be parametrized by $\vec{X}:U\to\mathbb{R}^3$, with

$$\vec{X}(u,v)=(u,v,f(u,v)).$$

Problem 6.3.7 asks the reader to calculate the matrix (L_{ij}). It is not hard to show that

$$(g_{ij})=\begin{pmatrix}1+(f_u)^2 & f_uf_v\\ f_vf_u & 1+(f_v)^2\end{pmatrix}\qquad\text{and}\qquad (L_{ij})=\frac{1}{\sqrt{1+f_u^2+f_v^2}}\begin{pmatrix}f_{uu} & f_{uv}\\ f_{vu} & f_{vv}\end{pmatrix},$$

where f_u is the typical shorthand to mean $\frac{\partial f}{\partial u}$. But then $\det(g)=1+f_u^2+f_v^2$, and we find that the Gaussian curvature function on a function graph is

$$K(u,v)=\frac{1}{(1+f_u^2+f_v^2)^2}\begin{vmatrix}f_{uu} & f_{uv}\\ f_{vu} & f_{vv}\end{vmatrix}=\frac{f_{uu}f_{vv}-f_{uv}^2}{(1+f_u^2+f_v^2)^2}.$$

This result allows us to rephrase the second derivative test in the calculus of a function f from $\mathbb{R}^2$ to $\mathbb{R}$ as follows: if f has continuous second partial derivatives and (u_0,v_0) is a critical point (i.e., $f_u(u_0,v_0)=f_v(u_0,v_0)=0$), then

1. (u_0, v_0) is a local maximum if $K(u_0, v_0) > 0$ and $f_{uu}(u_0, v_0) < 0$;
2. (u_0, v_0) is a local minimum if $K(u_0, v_0) > 0$ and $f_{uu}(u_0, v_0) > 0$;
3. (u_0, v_0) is a saddle point if $K(u_0, v_0) < 0$;
4. the test is inconclusive if $K(u_0, v_0) = 0$.

In the language we have introduced in this chapter, local minima and local maxima of the function $z = f(u, v)$ are elliptic points, saddle points are hyperbolic points, and points where the second derivative test is inconclusive are either parabolic or planar points.

Example 6.5.5 (Pseudosphere). A *tractrix* is a curve in the plane with parametric equations

$$\vec{x}(t) = (t - \tanh t, \operatorname{sech} t),$$

and it has the x-axis as an asymptote. The *pseudosphere* is defined as half of the surface of revolution of a tractrix about its asymptote. More precisely, we can parametrize the pseudosphere by

$$\vec{X}(u, v) = (\operatorname{sech} v \cos u, \operatorname{sech} v \sin u, v - \tanh v),$$

with $u \in [0, 2\pi)$ and $v \in [0, \infty)$. (See Figure 6.10 for a picture of the tractrix and the pseudosphere.) We leave it as an exercise for the reader to prove that

$$(g_{ij}) = \begin{pmatrix} \operatorname{sech}^2 v & 0 \\ 0 & \tanh^2 v \end{pmatrix} \quad \text{and} \quad (L_{ij}) = \begin{pmatrix} \operatorname{sech} v \tanh v & 0 \\ 0 & -\operatorname{sech} v \tanh v \end{pmatrix}.$$

We find that the Gaussian curvature for the pseudosphere is $K = -1$, which motivates the name “pseudosphere” since it is analogous to the unit sphere but with a constant Gaussian curvature of -1 instead of 1.

Proposition 6.5.6. *Let S be a regular oriented surface of class C^2 with Gauss map $n : S \to \mathbb{S}^2$, and let p be a point on S. Call $K(p)$ the Gaussian curvature of S at p, and suppose that $K(p) \neq 0$. Let B_ε be the ball of radius ε around p and define $V_\varepsilon = B_\varepsilon \cap S$. Calling $K(p)$ the Gaussian curvature of S at p, we have*

$$|K(p)| = \lim_{\varepsilon \to 0} \frac{\operatorname{Area}(n(V_\varepsilon))}{\operatorname{Area}(V_\varepsilon)}.$$

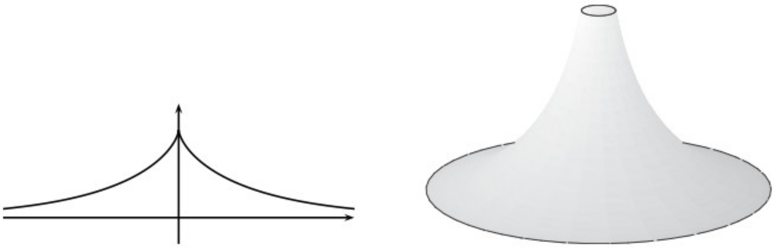

Figure 6.10. Tractrix and pseudosphere.

In other words, the absolute value of the Gaussian curvature at a point p *is the limit around* p *of the ratio of the surface area on* $\mathbb{S}^2$ *mapped under the Gauss map to the corresponding surface area on* S*.*

Proof: Let $\vec{X} : U \to \mathbb{R}^3$ be a regular parametrization of a neighborhood of $p = \vec{X}(u_0, v_0)$, where U is an open subset of $\mathbb{R}^2$. Since $\vec{X}(U)$ is an open set, $V_\varepsilon \subset \vec{X}(U)$ for all ε small enough. Furthermore, since we assumed that $K(p) \neq 0$, by the continuity of the Gaussian curvature function $K : U \to \mathbb{R}$, we deduce that K does not change sign for ε chosen small enough. Therefore, from now on, we assume that K does not change sign.

Define $U_\varepsilon = \vec{X}^{-1}(V_\varepsilon)$, the preimage of the neighborhood V_ε under $\vec{X}$. Since $\vec{X}$ is bijective as a regular parametrization, $\{(u_0, v_0)\}$ is the unique point in all U_ε for all $\varepsilon > 0$. Then by the formula for surface area,

$$\text{Area}(V_\varepsilon) = \iint_{V_\varepsilon} dA = \iint_{U_\varepsilon} \|\vec{X}_u \times \vec{X}_v\| \, du \, dv.$$

Similarly, on the unit sphere,

$$\text{Area}(n(V_\varepsilon)) = \iint_{U_\varepsilon} \|\vec{N}_u \times \vec{N}_v\| \, du \, dv.$$

However, by Proposition 6.5.2, we also have

$$\text{Area}(n(V_\varepsilon)) = \iint_{U_\varepsilon} |K(u, v)| \|\vec{X}_u \times \vec{X}_v\| \, du \, dv.$$

By the Mean Value Theorem for double integrals, for every $\varepsilon > 0$, there exist points $(u_\varepsilon, v_\varepsilon)$ and $(u'_\varepsilon, v'_\varepsilon)$ in the open set U_ε such that

$$\iint_{U_\varepsilon} \|\vec{X}_u \times \vec{X}_v\| \, du \, dv = \|\vec{X}_u(u_\varepsilon, v_\varepsilon) \times \vec{X}_v(u_\varepsilon, v_\varepsilon)\|$$

and

$$\iint_{U_\varepsilon} \|\vec{N}_u \times \vec{N}_v\| \, du \, dv = |K(u'_\varepsilon, v'_\varepsilon)| \, \|\vec{X}_u(u'_\varepsilon, v'_\varepsilon) \times \vec{X}_v(u'_\varepsilon, v'_\varepsilon)\|.$$

Thus, since $\lim_{\varepsilon\to 0}(u_\varepsilon, v_\varepsilon) = \lim_{\varepsilon\to 0}(u'_\varepsilon, v'_\varepsilon) = (u_0, v_0)$, we have

$$\begin{aligned}\lim_{\varepsilon\to 0} \frac{\operatorname{Area}(n(V_\varepsilon))}{\operatorname{Area}(V_\varepsilon)} &= \lim_{\varepsilon\to 0} \frac{|K(u'_\varepsilon, v'_\varepsilon)| \, \|\vec{X}_u(u'_\varepsilon, v'_\varepsilon) \times \vec{X}_v(u'_\varepsilon, v'_\varepsilon)\|}{\|\vec{X}_u(u_\varepsilon, v_\varepsilon) \times \vec{X}_v(u_\varepsilon, v_\varepsilon)\|} \\ &= \frac{|K(u_0, v_0)| \|\vec{X}_u(u_0, v_0) \times \vec{X}_v(u_0, v_0)\|}{\|\vec{X}_u(u_0, v_0) \times \vec{X}_v(u_0, v_0)\|} \\ &= |K(u_0, v_0)| = |K(p)|. \qquad \square\end{aligned}$$

Problems

6.5.1. Calculate the Gaussian curvature of the torus parametrized by

$$\vec{X}(u, v) = ((a + b\cos v)\cos u, (a + b\cos v)\sin u, b\sin v)$$

where $a > b$ are constants and $(u, v) \in (0, 2\pi) \times (0, 2\pi)$.

6.5.2. Consider a regular space curve $\vec{\gamma} : I \to \mathbb{R}^3$ and the tangential surface defined by

$$\vec{X}(t, u) = \vec{\gamma}(t) + u\vec{\gamma}'(t)$$

for $(t, u) \in I \times \mathbb{R}$. Calculate the mean and Gaussian curvature functions.

6.5.3. Let $\vec{\alpha}(t)$ and $\vec{\beta}(t)$ be differentiable vector functions with common domain I. Define the *secant surface* between the two resulting curves by

$$\vec{X}(t, u) = (1 - u)\vec{\alpha}(t) + u\vec{\beta}(t)$$

for $(t, u) \in I \times \mathbb{R}$. Assume that the corresponding surface is regular.

Proof: Recall the formula for surface area introduced in multivariable calculus that gives the surface area of the region R on S as

$$A(R) = \iint\limits_R dA = \iint\limits_Q \|\vec{X}_u \times \vec{X}_v\| \, du \, dv. \tag{6.5}$$

(See [23, Section 7.4] or [28, Section 17.7] for an explanation of Equation (6.5).)

By Problem 3.1.3, for any two vectors $\vec{v}$ and $\vec{w}$ in $\mathbb{R}^3$, the following identity holds:

$$(\vec{v} \times \vec{w}) \cdot (\vec{v} \times \vec{w}) = (\vec{v} \cdot \vec{v})(\vec{w} \cdot \vec{w}) - (\vec{v} \cdot \vec{w})^2.$$

Therefore,

$$\|\vec{X}_u \times \vec{X}_v\|^2 = (\vec{X}_u \cdot \vec{X}_u)(\vec{X}_v \cdot \vec{X}_v) - (\vec{X}_u \cdot \vec{X}_v)^2 = g_{11}g_{22} - g_{12}^2, \tag{6.6}$$

and the proposition follows because $g_{12} = g_{21}$. ☐

Proposition 6.1.7 is interesting in itself as it leads to another property of the first fundamental form. It is obvious from the definition of the first fundamental form that $g_{11}(u,v) \geq 0$ and $g_{22}(u,v) \geq 0$ for all $(u,v) \in U$. However, since $\|\vec{X}_u \times \vec{X}_v\|^2 = \det(g)$ and since $\vec{X}_u \times \vec{X}_v$ is never $\vec{0}$ on a regular surface, we deduce that $\det(g) = g_{11}g_{22} - g_{12} > 0$ for all $(u,v) \in U$. In other words, the matrix of functions g is always a *positive definite* matrix.

Example 6.1.8. Consider the regular surface S parametrized by $\vec{X}(u,v) = (v\cos u, v \sin u, \ln v)$, with $(u,v) \in [0, 2\pi] \times (0, \infty) = U$. (From a practical standpoint, it is not so important that U is not an open subset of $\mathbb{R}^2$, but it should be understood that S requires two coordinate neighborhoods to satisfy the criteria for being a regular surface. See Figure 6.3.) Let $Q = [0, 2\pi] \times [1, 2]$, and let's calculate the area of $R = \vec{X}(Q)$.

It's not hard to show that

$$(g_{ij}) = \begin{pmatrix} v^2 & 0 \\ 0 & 1 + \frac{1}{v^2} \end{pmatrix}.$$

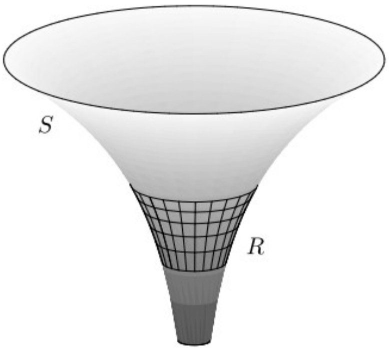

Figure 6.3. Surface area patch for Example 6.1.8.

Then Proposition 6.1.7 gives

$$\begin{aligned} A(R) &= \int_0^{2\pi}\int_1^2 \sqrt{v^2+1}\,dv\,du = 2\pi\Big[\frac{1}{2}v\sqrt{v^2+1}+\frac{1}{2}\sinh^{-1}v\Big]_1^2 \\ &= \pi\left(2\sqrt{5}-\sqrt{2}+\ln\left(\frac{2+\sqrt{5}}{1+\sqrt{2}}\right)\right). \end{aligned}$$

As we have done for many quantities on curves and surfaces, we wish to see how the functions in the first fundamental form change under a coordinate transformation. In order to avoid confusion with various coordinate systems, we introduce some notations that will become more common in future topics. Suppose that an open subset V of a regular surface S can be parametrized by two different sets of coordinates (x_1, x_2) and $(\bar{x}_1, \bar{x}_2)$. If we write $\vec{X}(x_1, x_2)$ and $\vec{Y}(\bar{x}_1, \bar{x}_2)$ for the specific parametrizations, then the function $F = \vec{Y}^{-1} \circ \vec{X}$ is a diffeomorphism between two open subsets of $\mathbb{R}^2$ and

$$(\bar{x}_1, \bar{x}_2) = F(x_1, x_2).$$

For convenience, we will often simply write

$$\begin{cases} \bar{x}_1 = \bar{x}_1(x_1, x_2), \\ \bar{x}_2 = \bar{x}_2(x_1, x_2) \end{cases}$$

to indicate the dependency of one set of variables on the other. Vice versa, we can write

$$(x_1, x_2) = F^{-1}(\bar{x}_1, \bar{x}_2) \quad \text{or simply} \qquad \begin{cases} x_1 = x_1(\bar{x}_1, \bar{x}_2), \\ x_2 = x_2(\bar{x}_1, \bar{x}_2). \end{cases}$$

Now $\vec{Y}(\bar{x}_1, \bar{x}_2) = \vec{X}(x_1(\bar{x}_1, \bar{x}_2), x_2(\bar{x}_1, \bar{x}_2))$ for the respective parametrizations. Let g_{ij} be the coefficient functions in the first fundamental form for S parametrized by x_1, x_2, and let $\bar{g}_{ij}$ be the coefficient functions for S parametrized by $\bar{x}_1, \bar{x}_2$. Then by the chain rule,

$$\bar{g}_{ij} = \vec{Y}_{\bar{x}_i} \cdot \vec{Y}_{\bar{x}_j} = \left(\vec{X}_{x_1} \frac{\partial x_1}{\partial \bar{x}_i} + \vec{X}_{x_2} \frac{\partial x_2}{\partial \bar{x}_i} \right) \cdot \left(\vec{X}_{x_1} \frac{\partial x_1}{\partial \bar{x}_j} + \vec{X}_{x_2} \frac{\partial x_2}{\partial \bar{x}_j} \right).$$

After some reorganization, we find

$$\bar{g}_{ij} = \sum_{k=1}^{2} \sum_{l=1}^{2} \frac{\partial x_k}{\partial \bar{x}_i} \frac{\partial x_l}{\partial \bar{x}_j} g_{kl}. \tag{6.7}$$

This type of transformation of coordinates is our first (nontrivial) encounter with tensors. In modern terminology, the collection of the quantities g_{ij} defined at each point of $p \in S$ is referred to as the components of the *metric tensor* of S. Though the first fundamental form, as a bilinear form on T_pS, is independent of any particular parametrization, the specific functions g_{ij} depend on the parametrization, but one knows that they behave according to Equation (6.7) under coordinate changes. Because the g_{ij} functions satisfy this particular identity, we call the matrix of functions $g = (g_{ij})$ a tensor of type $(0, 2)$. Later on, as further study will lead to more complicated geometric objects, we will develop a comprehensive theory of tensors in differential geometry.

The value of the first fundamental form is that, given the metric tensor, one can use the appropriate formulas for arc length, for angles between curves and for areas of regions without knowing the specific parametrization of the surface. An additional benefit of the first fundamental form and the metric tensor is that one can still use them to study the geometry of surfaces in higher dimensions.

Suppose that S is a regular surface in $\mathbb{R}^n$ and consider a parametrization of a coordinate patch $\vec{X} : U \to \mathbb{R}^n$, where U is an open subset of $\mathbb{R}^2$. Define the functions $g_{ij} = \vec{X}_i \cdot \vec{X}_j$ as before. Then Equations (6.1) and (6.2) still provide the appropriate formulas for arc length and the angle between vectors in T_pS. In addition, Proposition 6.1.7 generalizes to the following.

Proposition 6.1.9. *Let S be a regular surface in $\mathbb{R}^n$, with $n \geq 3$, and let $\vec{X} : U \to \mathbb{R}^n$ be a parametrization of a coordinate patch of S, where U is an open subset of $\mathbb{R}^2$. Let Q be a compact subset of U and call $R = \vec{X}(Q)$ the compact subset of S. The area of R is given by*

$$A(R) = \iint_Q \sqrt{\det(g)}\, du\, dv. \tag{6.8}$$

Proof: The proof of Equation (6.4) relied on the fact that S is a regular surface in $\mathbb{R}^3$ and made reference to the cross product of $\vec{X}_u$ and $\vec{X}_v$, namely, using the formula

$$\iint_Q \|\vec{X}_u \times \vec{X}_v\|\, du\, dv$$

from multivariable calculus. The Riemann sum behind this integral involves approximating the surface area traced out by $\vec{X}$ over $[u_i, u_{i+1}] \times [v_i, v_{i+1}]$ by a parallelogram spanned on two sides by $\vec{X}_u(u^*, v^*)$ and $\vec{X}_v(u^*, v^*)$, where (u^*, v^*) is a selection point in $[u_i, u_{i+1}] \times [v_i, v_{i+1}]$.

In order to find a formula for the surface area of a regular surface in $\mathbb{R}^n$, one cannot refer to any formula using the cross product since the cross product is only defined in $\mathbb{R}^3$. However, the approach used by the Riemann sum is still the right one. Therefore, one must find a convenient formula for the area of the parallelogram spanned by $\vec{X}_u$ and $\vec{X}_v$ in $\mathbb{R}^n$.

Regardless of the dimension of the ambient Euclidean space, we can calculate the area of these parallelograms as $\|\vec{X}_u\|\,\|\vec{X}_v\| \sin\theta$, where θ is the angle between the two vectors. One cannot get $\sin\theta$ directly but one can obtain $\cos\theta$ from $\vec{X}_u \cdot \vec{X}_v = \|\vec{X}_u\|\,\|\vec{X}_v\| \cos\theta$.

Thus,

$$\begin{aligned}
& & (\vec{X}_u \cdot \vec{X}_v)^2 &= (\vec{X}_u \cdot \vec{X}_u)(\vec{X}_v \cdot \vec{X}_v)\cos^2\theta \\
\Longrightarrow & & (\vec{X}_u \cdot \vec{X}_v)^2 &= (\vec{X}_u \cdot \vec{X}_u)(\vec{X}_v \cdot \vec{X}_v)(1-\sin^2\theta) \\
\Longrightarrow & & (\vec{X}_u \cdot \vec{X}_u)(\vec{X}_v \cdot \vec{X}_v) - (\vec{X}_u \cdot \vec{X}_v)^2 &= (\vec{X}_u \cdot \vec{X}_u)(\vec{X}_v \cdot \vec{X}_v)\sin^2\theta \\
\Longleftrightarrow & & \sqrt{(\vec{X}_u \cdot \vec{X}_u)(\vec{X}_v \cdot \vec{X}_v) - (\vec{X}_u \cdot \vec{X}_v)^2} &= \|\vec{X}_u\|\,\|\vec{X}_v\|\sin\theta.
\end{aligned}$$

Thus, at each point $(u,v) \in U$, we have $\sqrt{\det(g)} = \|\vec{X}_u\|\,\|\vec{X}_v\|\sin\theta$. Taking the limit of the Riemann sum as the norm of a mesh that goes to 0 gives Equation (6.8). □

Because of Equation (6.1), a few mathematicians and most physicists use an alternative notation for the metric tensor. These authors describe the metric tensor by saying that the "line element" in a coordinate patch is given by

$$ds^2 = g_{11}(u,v)\,du^2 + 2g_{12}(u,v)\,du\,dv + g_{22}(u,v)\,dv^2. \tag{6.9}$$

This notation has the advantage that one can write it concisely on a single line instead of in matrix format. However, one should not forget that it not only leads to Equation (6.1) for arc length of a curve on a surface but that it also leads to the Equation (6.2) for angles between curves on the surface and the formula in Proposition 6.1.9 that gives the area of regions on the surface.

Problems

6.1.1. Calculate the metric tensor for the following regular surfaces:

- (a) The plane parametrized by $\vec{X}(s,t) = \vec{u} + s\vec{v} + t\vec{w}$ with $(s,t) \in \mathbb{R}^3$.
- (b) The surface with parametrization $\vec{X}(u,v) = (u^2 - v, uv, u)$.
- (c) The helicoid parametrized by $\vec{X}(u,v) = (v\cos u, v\sin u, au)$ for some $a \in \mathbb{R}$.

6.1.2. Calculate the metric tensor for the following families of conic surfaces:

- (a) The ellipsoid $\vec{X}(u,v) = (a\cos u\sin v, b\sin u\sin v, c\cos v)$.
- (b) The elliptic paraboloid $\vec{X}(u,v) = (av\cos u, bv\sin u, v^2)$.

! -256- - -* ! - / 4 ! 7 3 A ! "
! 4<4 A 1"
(/ 2 " 0 9

between the horizontal and vertical directions, it increases the purple–green angle and decreases the green–red angle. Thus, not all angles are preserved.

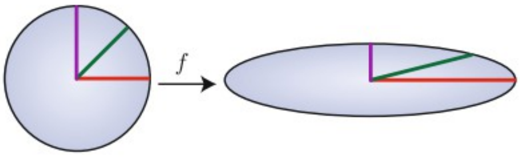

FIGURE 3.30. A linear equiareal map with distortion factor $\lambda = 2$

The above example is very special, because f is linear, but it illustrates the infinitesimal picture for a general equiareal map. That is, if $f : S_1 \to S_2$ is any equiareal map, then for every $p \in S_1$, orthonormal bases of T_pS_1 and $T_{f(p)}S_2$ can be chosen such that the matrix representing the linear transformation $df_p : T_pS_1 \to T_{f(p)}S_2$ has the form $A = \begin{pmatrix} \lambda & 0 \\ 0 & 1/\lambda \end{pmatrix}$ for some $\lambda > 0$ (Exercise 3.88). Thus, an equiareal map avoids distorting infinitesimal area through the balancing act of expanding one direction while compressing the other to compensate. The expansion–contraction factor λ is a function of p.

The most famous example of an equiareal map comes from Archimedes. The domain is the sphere with the north and south poles removed: $S^2 - \{(0,0,\pm 1)\}$. The range is the following cylinder:

$$C = \{(x,y,z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1 \text{ and } -1 < z < 1\}.$$

The function $f : S^2 - \{(0,0,\pm 1)\} \to C$ is pictured in Fig. 3.31. It sends each $p \in S^2 - \{(0,0,\pm 1)\}$ to the point $f(p) \in C$ that is nearest to p. This nearest point will have the same z-value as p. If you imagine that the z-axis emits light, and the sphere is made of glass with a speck of paint at p, then $f(p)$ is the speck's shadow on the cylinder. A formula for this shadowing of the sphere onto the cylinder is

$$f(x,y,z) = \left(\frac{x}{\sqrt{x^2+y^2}}, \frac{y}{\sqrt{x^2+y^2}}, z \right). \tag{3.18}$$

THEOREM 3.69 (**Archimedes's Theorem**).
The function f in Eq. 3.18 is equiareal.

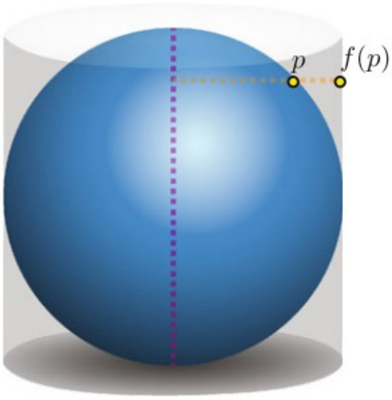

Archimedes used this theorem to discover the area of a sphere (it equals the area of the cylinder, which was previously known). According to legend, Archimedes asked that an image of his proof be inscribed on his tombstone. Since he predated the invention of calculus, his proof was much more involved than this one:

FIGURE 3.31. An equiareal map from the sphere to the cylinder

PROOF. The spherical coordinate chart $\sigma : \underbrace{(0, 2\pi) \times (0, \pi)}_{U} \to S^2$ was defined in Example 3.24 (on page 129) as

$$\sigma(\theta, \phi) = (\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi).$$

Let $q = (\theta, \phi) \in U$ and let $p = \sigma(q)$. We computed in that example that $|\sigma_\theta \times \sigma_\phi| = \sin\phi$. Since $\{\sigma_\theta, \sigma_\phi\}$ is a basis of T_pS^2, the definition of area distortion gives

$$\|df_p\| = \frac{|df_p(\sigma_\theta) \times df_p(\sigma_\phi)|}{|\sigma_\theta \times \sigma_\phi|} = \frac{|df_p(\sigma_\theta) \times df_p(\sigma_\phi)|}{\sin\phi}.$$

It remains to demonstrate that $|df_p(\sigma_\theta) \times df_p(\sigma_\phi)| = \sin\phi$, so that $\|df_p\| = 1$, as desired.

The composition $\tilde{\sigma} = f \circ \sigma : U \to C$ is a surface patch for C, whose formula simplifies to

$$\tilde{\sigma}(\theta, \phi) = (\cos(\theta), \sin(\theta), \cos\phi).$$

Since $df_p(\sigma_\theta) = \tilde{\sigma}_\theta$ and $df_p(\sigma_\phi) = \tilde{\sigma}_\phi$, we learn that

$$|df_p(\sigma_\theta) \times df_p(\sigma_\phi)| = |\tilde{\sigma}_\theta \times \tilde{\sigma}_\phi| = |(-\sin\theta, \cos\theta, 0) \times (0, 0, -\sin\phi)| = \sin\phi,$$

which completes the proof. □

From the formulas in the previous proof, you can verify what is apparent in Fig. 3.32; namely, $\|df_p\| = 1$, because an expansion of length in the latitudinal (σ_θ) direction is balanced by a contraction of length in the longitudinal (σ_ϕ) direction. In fact, with respect to the orthonormal bases $\left\{\frac{\sigma_\theta}{|\sigma_\theta|}, \frac{\sigma_\phi}{|\sigma_\phi|}\right\}$ and $\left\{\frac{\tilde{\sigma}_\theta}{|\tilde{\sigma}_\theta|}, \frac{\tilde{\sigma}_\phi}{|\tilde{\sigma}_\phi|}\right\}$, the linear transformation df_p is represented by the matrix $\begin{pmatrix} 1/\sin\phi & 0 \\ 0 & \sin\phi \end{pmatrix}$; compare to Example 3.68. The expansion–contraction factor, $\lambda = 1/\sin\phi$, equals 1 along the equator and approaches infinity as p approaches the poles.

Archimedes's theorem can be used to create a flat rectangular map of the spherical Earth with desirable properties. Specifically,the **Lambert projection** is obtained by projecting a globe onto the cylinder (via the function f from Eq. 3.18) and then unwrapping the cylinder into a flat rectangle (via the inverse of the cylinder's surface patch in Example 3.23 on page 128). This process and the resulting map are pictured in Fig. 3.33. The orange ellipses help you visualize the separate vertical and horizontal distortion factors of the projection at several points. Technically, they are the images of equal-sized round circles under the derivative of the projection (from globe to map) at a grid of equally spaced points along the map.

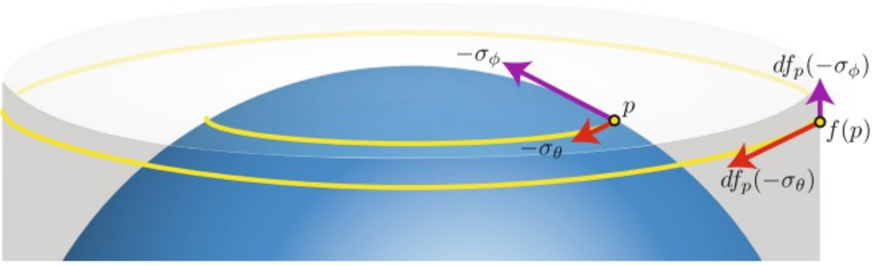

FIGURE 3.32. $\|df_p\| = 1$ because the vectors at p span the same area as the vectors at $f(p)$

According to Archimedes's theorem, the Lambert projection creates an "equal-area" map. The area of every continent or country on the spherical globe equals the area of its projection onto this rectangular map. The map therefore correctly displays relative areas, such as the fact that Africa is about 3.7 times the size of Australia. A common trivial modification is to compress the Lambert map horizontally. For example, the **Gall–Peters map** is nothing more than the Lambert map horizontally compressed until the expansion–contraction factor equals 1 along the 45° parallels north and south ($\phi = \pi/4$ and $\phi = 3\pi/4$), rather than along the equator. Shapes appear undistorted along these two special parallels, which are shown as thick orange lines in Fig. 3.33.

Although these equal-area maps represent area accurately, we saw in Fig. 3.30 that they distort angles and shapes. The distortion is extreme near the poles. Could a map of the Earth preserve angles? This question leads to our second generalization of isometries:

DEFINITION 3.70.
Let S_1, S_2 be a pair of regular surfaces. A diffeomorphism $f : S_1 \to S_2$ is called ***conformal*** *if df preserves angles, that is, if*

$$\angle(x, y) = \angle(df_p(x), df_p(y)) \text{ for all } p \in S_1 \text{ and all } x, y \in T_pS_1.$$

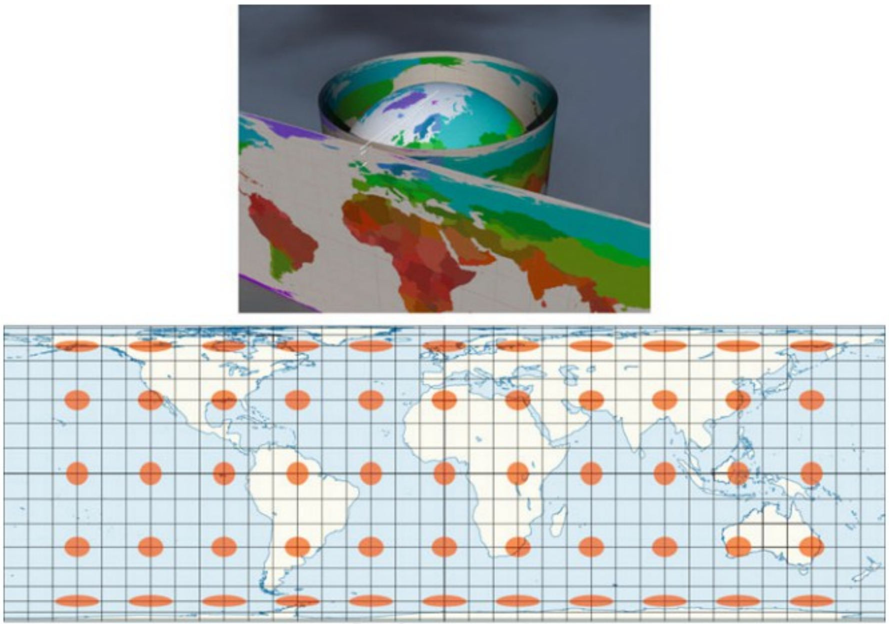

Lambert projection

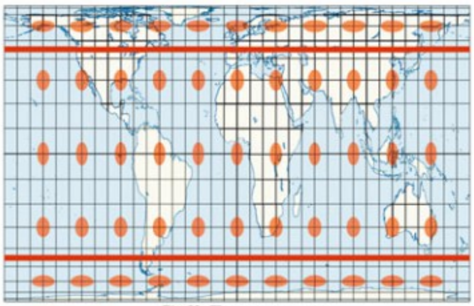

Gall-Peters map

FIGURE 3.33. Equal-area maps of the Earth

EXAMPLE 3.71. *Let* $\lambda > 0$. *Let* $f = L_A : \mathbb{R}^2 \to \mathbb{R}^2$ *denote the linear transformation represented by the matrix* $A = \left(\begin{smallmatrix} \lambda & 0 \\ 0 & \lambda \end{smallmatrix}\right)$, *that is,* $f(p) = \lambda p$. *Let* $p \in \mathbb{R}^2$ *and* $x, y \in T_p\mathbb{R}^2 = \mathbb{R}^2$. *Notice that* $df_p(x) = f(x) = \lambda x$. *Thus*

$$\begin{aligned} \angle(df_p(x), df_p(y)) = \cos^{-1}\left(\frac{\langle df_p(x), df_p(y)\rangle}{|df_p(x)||df_p(y)|}\right) &= \cos^{-1}\left(\frac{\langle \lambda x, \lambda y\rangle}{|\lambda x||\lambda y|}\right) \\ &= \cos^{-1}\left(\frac{\langle x, y\rangle}{|x||y|}\right) = \angle(x, y), \end{aligned}$$

so f *is conformal. Figure 3.34 shows the effect of* f *(with* $\lambda = 2$*) on a circle and some angles.*

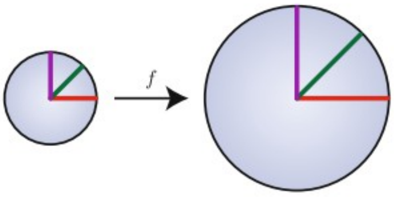

FIGURE 3.34. A linear conformal map with distortion factor $\lambda = 2$

The above example is very special, because f is linear, but it illustrates the infinitesimal picture for a general conformal map. That is, if $f : S_1 \to S_2$ is any conformal map, then for every $p \in S_1$, there exist orthonormal bases of T_pS_1 and $T_{f(p)}S_2$ with respect to which the linear transformation df_p is represented by a matrix of the form $A = \left(\begin{smallmatrix} \lambda & 0 \\ 0 & \lambda \end{smallmatrix}\right)$. The distortion factor λ typically depends on p, and is the factor by which df_p uniformly scales the norms of all vectors. In summary:

PROPOSITION 3.72.
A diffeomorphism $f : S_1 \to S_2$ is conformal if and only if there exists a smooth positive-valued function $\lambda : S_1 \to \mathbb{R}$ such that
$\langle df_p(x), df_p(y)\rangle_{f(p)} = \lambda(p)^2 \cdot \langle x, y\rangle_p$ *for all $p \in S_1$ and all $x, y \in T_pS_1$.*

PROOF. Follows from Exercise 3.89. ☐

The best-known example of a conformal map is the **stereographic projection** function $f : S^2 - \{(0, 0, 1)\} \to \mathbb{R}^2$. By definition, $f(p)$ gives the xy-coordinates of the intersection of the plane $z = -1$ with the line containing $(0, 0, 1)$ and p; see Fig. 3.35 (some authors instead project onto the plane $z = 0$; see Exercise 3.93).

PROPOSITION 3.73.
The stereographic projection function f is conformal.

PROOF. We first must describe a formula for f. An arbitrary point $p \in S^2$ can be expressed as $p = (\underbrace{r\cos\theta}_{x}, \underbrace{r\sin\theta}_{y}, \underbrace{\pm\sqrt{1-r^2}}_{z})$ for some angle θ and some $r > 0$. Its image $f(p)$ will have the form $(R\cos\theta, R\sin\theta)$ for some $R > 0$. To complete our description of f (and to understand f^{-1}), we must relate R and r. By similar triangles,

$$R = \frac{2r}{1-z} = \frac{2r}{1 \pm \sqrt{1-r^2}} \overset{\text{solve for } r}{\Longrightarrow} r = \frac{4R}{R^2+4}. \tag{3.19}$$

In summary, a formula for f^{-1} is

$$f^{-1}(R\cos\theta, R\sin\theta) = \left(r\cos\theta, r\sin\theta, \pm\sqrt{1-r^2}\right), \quad \text{where} \quad r = \frac{4R}{R^2+4}.$$

The symbol "±" equals 1 if $R > 2$ and equals -1 if $R < 2$.

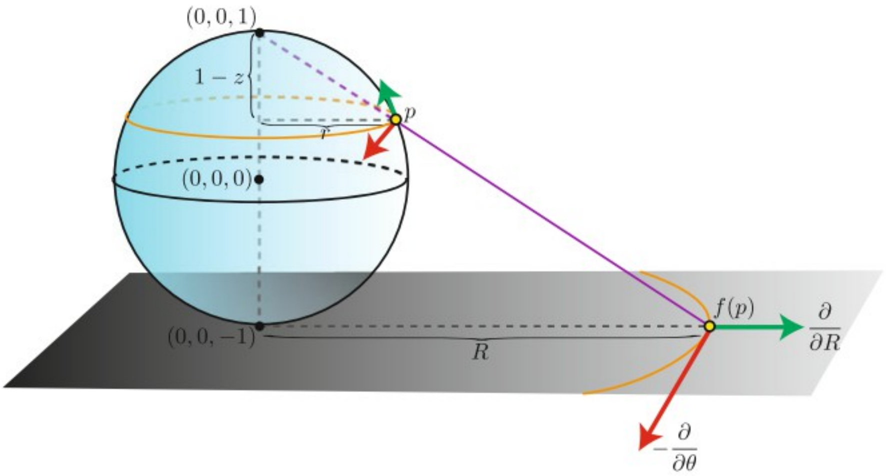
$(0,0,1)$
$1-z$
p
r
$(0,0,0)$
$(0,0,-1)$
R
$f(p)$
$\frac{\partial}{\partial R}$
$-\frac{\partial}{\partial \theta}$

FIGURE 3.35. Stereographic projection is conformal because its derivative scales the green and red vectors by the same factor

Now we prove that f is conformal. It is equivalent to prove that its inverse $f^{-1} : \mathbb{R}^2 \to S^2 - \{(0,0,1)\}$ is conformal. Let $q = (R\cos\theta, R\sin\theta) \in \mathbb{R}^2$. A basis for $T_q\mathbb{R}^2 = \mathbb{R}^2$ is

$$\frac{\partial}{\partial\theta} = (-R\sin\theta, R\cos\theta), \quad \frac{\partial}{\partial R} = \frac{q}{|q|} = (\cos\theta, \sin\theta).$$

These two vectors are orthogonal. Their images under $d(f^{-1})_q$ are also orthogonal:

$$d(f^{-1})_q\left(\frac{\partial}{\partial\theta}\right) = (-r\sin\theta, r\cos\theta, 0), \quad d(f^{-1})_q\left(\frac{\partial}{\partial R}\right) = \frac{dr}{dR}\left(\cos\theta, \sin\theta, \frac{\pm r}{\sqrt{1-r^2}}\right).$$

We have identified a pair of orthogonal vectors sent by $d(f^{-1})_q$ to a pair of orthogonal vectors, so it will suffice to verify that these vectors are scaled by the same amount; think about why this will suffice, or see Exercise 3.89 for more details. For this, it is straightforward to compute

$$\lambda = \frac{|d(f^{-1})_q\left(\frac{\partial}{\partial\theta}\right)|}{|\frac{\partial}{\partial\theta}|} = \frac{|d(f^{-1})_q\left(\frac{\partial}{\partial R}\right)|}{|\frac{\partial}{\partial R}|} = \frac{4}{4+R^2}. \tag{3.20}$$

□

Figure 3.36 shows a map made from a stereographic projection of the globe (turned upside down in order to depict the northern hemisphere). This map even includes part of the southern hemisphere (overlaid white), although an unbounded map would be required to include all of the southern hemisphere. This figure also depicts an artistic representation of stereographic projection as the shadowing from a light source positioned at the pole. In addition to these still images, we recommend the YouTube video *Stereographic projection of the Riemann sphere* for visualizing the conformal property of stereographic projection.

Stereographic projection is important within mathematics, but not within map-making. Cartographers care much more about a different conformal map—the **Mercator projection**, illustrated in Fig. 3.37. This well-recognized projection has greatly influenced our collective mental image of the Earth.

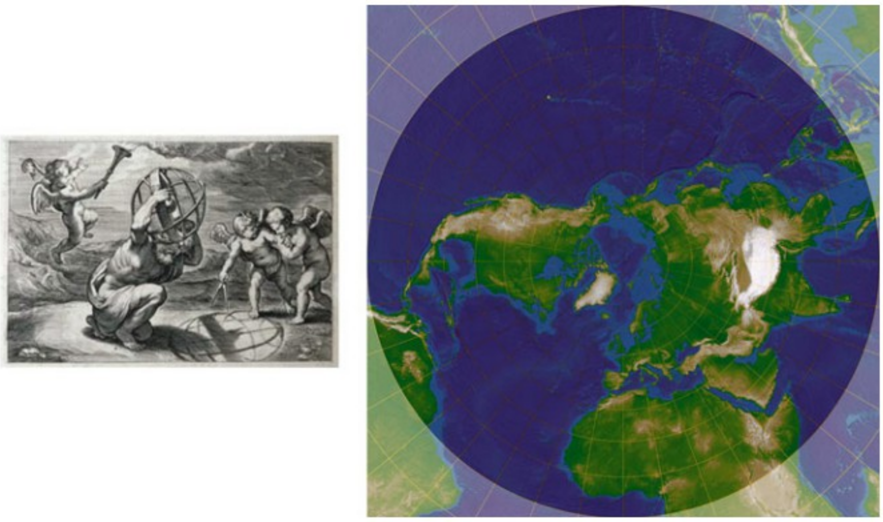

FIGURE 3.36. Stereographic projection map of northern hemisphere

EXAMPLE 3.74 (The Mercator Projection). *The idea is to modify the spherical coordinate chart so it becomes conformal. Recall from Example 3.24 on page 129 the spherical coordinate chart* $\sigma : \underbrace{(0,2\pi)\times(0,\pi)}_{U} \to S^2$ *defined as*

$$\sigma(\theta,\phi) = (\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi).$$

We computed in that example that

$$\sigma_\theta = (-\sin\phi\sin\theta, \sin\phi\cos\theta, 0), \qquad \sigma_\phi = (\cos\phi\cos\theta, \cos\phi\sin\theta, -\sin\phi).$$

The vectors σ_θ *and* σ_ϕ *are orthogonal, but* σ *is not conformal, because*

$$\sin\phi = |\sigma_\theta| \neq |\sigma_\phi| = 1.$$

We can fix this problem by precomposing σ *with a function that appropriately warps the vertical axis of the domain. That is, we define the function*

$$\Psi : \underbrace{(0,2\pi)\times(a,b)}_{\tilde{U}} \to \underbrace{(0,2\pi)\times(0,\pi)}_{U}$$

!% -256- - -* ! - / 4 ! 7 3 A ! "
! 5 A " ! " H1"! " " =" 1"
($$ 9 7! !1 ? ! " !

***18.** Let $\lambda_1, \dots, \lambda_m$ be the normal curvatures at $p \in S$ along directions making angles $0, 2\pi/m, \dots, (m-1)2\pi/m$ with a principal direction, $m > 2$. Prove that

$$\lambda_1 + \cdots + \lambda_m = mH,$$

where H is the mean curvature at p.

***19.** Let $C \subset S$ be a regular curve in S. Let $p \in C$ and $\alpha(s)$ be a parametrization of C in p by arc length so that $\alpha(0) = p$. Choose in $T_p(S)$ an orthonormal positive basis $\{t, h\}$, where $t = \alpha'(0)$. The *geodesic torsion* τ_g of $C \subset S$ at p is defined by

$$\tau_g = \left\langle \frac{dN}{ds}(0), h \right\rangle.$$

Prove that

a. $\tau_g = (k_1 - k_2)\cos\varphi \sin\varphi$, where φ is the angle from e_1 to t and t is the unit tangent vector corresponding to the principal curvature k_1.

b. If τ is the torsion of C, n is the (principal) normal vector of C and $\cos\theta = \langle N, n \rangle$, then

$$\frac{d\theta}{ds} = \tau - \tau_g.$$

c. The lines of curvature of S are characterized by having geodesic torsion identically zero.

***20.** (*Dupin's Theorem.*) Three families of surfaces are said to form a *triply orthogonal system* in an open set $U \subset R^3$ if a unique surface of each family passes through each point $p \in U$ and if the three surfaces that pass through p are pairwise orthogonal. Use part c of Exercise 19 to prove Dupin's theorem: *The surfaces of a triply orthogonal system intersect each other in lines of curvature.*

3-3. The Gauss Map in Local Coordinates

In the preceding section, we introduced some concepts related to the local behavior of the Gauss map. To emphasize the geometry of the situation, the definitions were given without the use of a coordinate system. Some simple examples were then computed directly from the definitions; this procedure, however, is inefficient in handling general situations. In this section, we shall obtain the expressions of the second fundamental form and of the differential of the Gauss map in a coordinate system. This will give us a systematic method for computing specific examples. Moreover, the general expressions thus obtained are essential for a more detailed investigation of the concepts introduced above.

All parametrizations $\mathbf{x}\colon U \subset R^2 \to S$ considered in this section are assumed to be compatible with the orientation N of S; that is, in $\mathbf{x}(U)$,

$$N = \frac{\mathbf{x}_u \wedge \mathbf{x}_v}{|\mathbf{x}_u \wedge \mathbf{x}_v|}.$$

Let $\mathbf{x}(u, v)$ be a parametrization at a point $p \in S$ of a surface S, and let $\alpha(t) = \mathbf{x}(u(t), v(t))$ be a parametrized curve on S, with $\alpha(0) = p$. To simplify the notation, we shall make the convention that all functions to appear below denote their values at the point p.

The tangent vector to $\alpha(t)$ at p is $\alpha' = \mathbf{x}_u u' + \mathbf{x}_v v'$ and

$$dN(\alpha') = N'(u(t), v(t)) = N_u u' + N_v v'.$$

Since N_u and N_v belong to $T_p(S)$, we may write

$$\begin{aligned} N_u &= a_{11}\mathbf{x}_u + a_{21}\mathbf{x}_v, \\ N_v &= a_{12}\mathbf{x}_u + a_{22}\mathbf{x}_v, \end{aligned} \tag{1}$$

and therefore,

$$dN(\alpha') = (a_{11}u' + a_{12}v')\mathbf{x}_u + (a_{21}u' + a_{22}v')\mathbf{x}_v;$$

hence,

$$dN\begin{pmatrix} u' \\ v' \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}\begin{pmatrix} u' \\ v' \end{pmatrix}.$$

This shows that in the basis $\{\mathbf{x}_u, \mathbf{x}_v\}$, dN is given by the matrix (a_{ij}), $i, j = 1, 2$. Notice that this matrix is not necessarily symmetric, unless $\{\mathbf{x}_u, \mathbf{x}_v\}$ is an orthonormal basis.

On the other hand, the expression of the second fundamental form in the basis $\{\mathbf{x}_u, \mathbf{x}_v\}$ is given by

$$\begin{aligned} II_p(\alpha') &= -\langle dN(\alpha'), \alpha'\rangle = -\langle N_u u' + N_v v', \mathbf{x}_u u' + \mathbf{x}_v v'\rangle \\ &= e(u')^2 + 2fu'v' + g(v')^2, \end{aligned}$$

where, since $\langle N, \mathbf{x}_u\rangle = \langle N, \mathbf{x}_v\rangle = 0$,

$$\begin{aligned} e &= -\langle N_u, \mathbf{x}_u\rangle = \langle N, \mathbf{x}_{uu}\rangle, \\ f &= -\langle N_v, \mathbf{x}_u\rangle = \langle N, \mathbf{x}_{uv}\rangle = \langle N, \mathbf{x}_{vu}\rangle = -\langle N_u, \mathbf{x}_v\rangle, \\ g &= -\langle N_v, \mathbf{x}_v\rangle = \langle N, \mathbf{x}_{vv}\rangle. \end{aligned}$$

We shall now obtain the values of a_{ij} in terms of the coefficients e, f, g. From Eq. (1), we have

$$\begin{aligned} -f &= \langle N_u, \mathbf{x}_v \rangle = a_{11}F + a_{21}G, \\ -f &= \langle N_v, \mathbf{x}_u \rangle = a_{12}E + a_{22}F, \\ -e &= \langle N_u, \mathbf{x}_u \rangle = a_{11}E + a_{21}F, \\ -g &= \langle N_v, \mathbf{x}_v \rangle = a_{12}F + a_{22}G, \end{aligned} \tag{2}$$

where E, F, and G are the coefficients of the first fundamental form in the basis $\{\mathbf{x}_u, \mathbf{x}_v\}$ (cf. Sec. 2-5). Relations (2) may be expressed in matrix form by

$$-\begin{pmatrix} e & f \\ f & g \end{pmatrix} = \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{pmatrix} \begin{pmatrix} E & F \\ F & G \end{pmatrix}; \tag{3}$$

hence,

$$\begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{pmatrix} = -\begin{pmatrix} e & f \\ f & g \end{pmatrix} \begin{pmatrix} E & F \\ F & G \end{pmatrix}^{-1},$$

where $(\ )^{-1}$ means the inverse matrix of $(\ )$. It is easily checked that

$$\begin{pmatrix} E & F \\ F & G \end{pmatrix}^{-1} = \frac{1}{EG - F^2} \begin{pmatrix} G & -F \\ -F & E \end{pmatrix},$$

whence the following expressions for the coefficients (a_{ij}) of the matrix of dN in the basis $\{\mathbf{x}_u, \mathbf{x}_v\}$:

$$\begin{aligned} a_{11} &= \frac{fF - eG}{EG - F^2}, \\ a_{12} &= \frac{gF - fG}{EG - F^2}, \\ a_{21} &= \frac{eF - fE}{EG - F^2}, \\ a_{22} &= \frac{fF - gE}{EG - F^2}. \end{aligned}$$

For completeness, it should be mentioned that relations (1), with the above values, are known as the *equations of Weingarten*.

From Eq. (3) we immediately obtain

$$K = \det(a_{ij}) = \frac{eg - f^2}{EG - F^2}. \tag{4}$$

To compute the mean curvature, we recall that $-k_1$, $-k_2$ are the eigenvalues of dN. Therefore, k_1 and k_2 satisfy the equation

$$dN(v) = -kv = -kIv \quad \text{for some } v \in T_p(S),\, v \neq 0,$$

where I is the identity map. It follows that the linear map $dN + kI$ is not invertible; hence, it has zero determinant. Thus,

$$\det\begin{pmatrix} a_{11}+k & a_{12} \\ a_{21} & a_{22}+k \end{pmatrix} = 0$$

or

$$k^2 + k(a_{11} + a_{22}) + a_{11}a_{22} - a_{21}a_{12} = 0.$$

Since k_1 and k_2 are the roots of the above quadratic equation, we conclude that

$$H = \frac{1}{2}(k_1 + k_2) = -\frac{1}{2}(a_{11} + a_{22}) = \frac{1}{2}\frac{eG - 2fF + gE}{EG - F^2}; \tag{5}$$

hence,

$$k^2 - 2Hk + K = 0,$$

and therefore,

$$k = H \pm \sqrt{H^2 - K}. \tag{6}$$

From this relation, it follows that if we choose $k_1(q) \geq k_2(q)$, $q \in S$, then the functions k_1 and k_2 are continuous in S. Moreover, k_1 and k_2 are differentiable in S, except perhaps at the umbilical points ($H^2 = K$) of S.

In the computations of this chapter, it will be convenient to write for short

$$\langle u \wedge v, w\rangle = (u, v, w) \quad \text{for any } u, v, w \in R^3.$$

We recall that this is merely the determinant of the 3×3 matrix whose columns (or lines) are the components of the vectors u, v, w in the canonical basis of R^3.

Example 1. We shall compute the Gaussian curvature of the points of the torus covered by the parametrization (cf. Example 6 of Sec. 2-2)

$$\mathbf{x}(u, v) = ((a + r\cos u)\cos v, (a + r\cos u)\sin v, r\sin u),$$
$$0 < u < 2\pi, \quad 0 < v < 2\pi.$$

For the computation of the coefficients e, f, g, we need to know N (and thus $\mathbf{x}_u$ and $\mathbf{x}_v$), $\mathbf{x}_{uu}$, $\mathbf{x}_{uv}$, and $\mathbf{x}_{vv}$:

$$\begin{aligned}
\mathbf{x}_u &= (-r\sin u\cos v, -r\sin u\sin v, r\cos u),\\
\mathbf{x}_v &= (-(a + r\cos u)\sin v, (a + r\cos u)\cos v, 0),\\
\mathbf{x}_{uu} &= (-r\cos u\cos v, -r\cos u\sin v, -r\sin u),\\
\mathbf{x}_{uv} &= (r\sin u\sin v, -r\sin u\cos v, 0),\\
\mathbf{x}_{vv} &= (-(a + r\cos u)\cos v, -(a + r\cos u)\sin v, 0).
\end{aligned}$$

From these, we obtain

$$\begin{aligned}
E &= \langle \mathbf{x}_u, \mathbf{x}_u\rangle = r^2, \quad F = \langle \mathbf{x}_u, \mathbf{x}_v\rangle = 0,\\
G &= \langle \mathbf{x}_v, \mathbf{x}_v\rangle = (a + r\cos u)^2.
\end{aligned}$$

Introducing the values just obtained in $e = \langle N, \mathbf{x}_{uu}\rangle$, we have, since $|\mathbf{x}_u \wedge \mathbf{x}_v| = \sqrt{EG - F^2}$,

$$e = \left\langle \frac{\mathbf{x}_u \wedge \mathbf{x}_v}{|\mathbf{x}_u \wedge \mathbf{x}_v|}, \mathbf{x}_{uu} \right\rangle = \frac{(\mathbf{x}_u, \mathbf{x}_v, \mathbf{x}_{uu})}{\sqrt{EG - F^2}} = \frac{r^2(a + r\cos u)}{r(a + r\cos u)} = r.$$

Similarly, we obtain

$$\begin{aligned}
f &= \frac{(\mathbf{x}_u, \mathbf{x}_v, \mathbf{x}_{uv})}{r(a + r\cos u)} = 0,\\
g &= \frac{(\mathbf{x}_u, \mathbf{x}_v, \mathbf{x}_{vv})}{r(a + r\cos u)} = \cos u(a + r\cos u).
\end{aligned}$$

Finally, since $K = (eg - f^2)/(EG - F^2)$, we have that

$$K = \frac{\cos u}{r(a + r\cos u)}.$$

From this expression, it follows that $K = 0$ along the parallels $u = \pi/2$ and $u = 3\pi/2$; the points of such parallels are therefore parabolic points. In the region of the torus given by $\pi/2 < u < 3\pi/2$, K is negative (notice that $r > 0$ and $a > r$); the points in this region are therefore hyperbolic points. In the region given by $0 < u < \pi/2$ or $3\pi/2 < u < 2\pi$, the curvature is positive and the points are elliptic points (Fig. 3-15).

As an application of the expression for the second fundamental form in coordinates, we shall prove a proposition which gives information about the

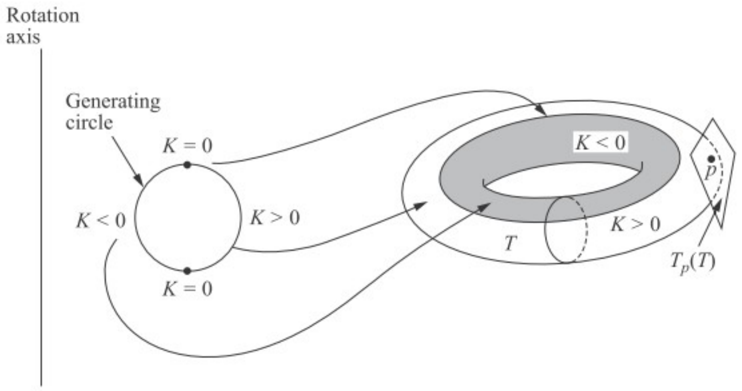

Figure 3-15

position of a surface in the neighborhood of an elliptic or a hyperbolic point, relative to the tangent plane at this point. For instance, if we look at an elliptic point of the torus of Example 1, we find that the surface lies on one side of the tangent plane at such a point (see Fig. 3-15). On the other hand, if p is a hyperbolic point of the torus T and $V \subset T$ is any neighborhood of p, we can find points of V on both sides of $T_p(S)$, however small V may be. This example reflects a general local fact that is described in the following proposition.

PROPOSITION 1. *Let* $\mathrm{p} \in \mathrm{S}$ *be an elliptic point of a surface* S. *Then there exists a neighborhood* V *of* p *in* S *such that all points in* V *belong to the same side of the tangent plane* $\mathrm{T_p(S)}$. *Let* $\mathrm{p} \in \mathrm{S}$ *be a hyperbolic point. Then in each neighborhood of* p *there exist points of* S *in both sides of* $\mathrm{T_p(S)}$.

Proof. Let $\mathbf{x}(u, v)$ be a parametrization in p, with $\mathbf{x}(0, 0) = p$. The distance d from a point $q = \mathbf{x}(u, v)$ to the tangent plane $T_p(S)$ is given by (Fig. 3-16)

$$d = \langle \mathbf{x}(u, v) - \mathbf{x}(0, 0), N(p) \rangle.$$

Since $\mathbf{x}(u, v)$ is differentiable, we have Taylor's formula:

$$\mathbf{x}(u, v) = \mathbf{x}(0, 0) + \mathbf{x}_u u + \mathbf{x}_v v + \tfrac{1}{2}(\mathbf{x}_{uu} u^2 + 2\mathbf{x}_{uv} uv + \mathbf{x}_{vv} v^2) + \bar{R},$$

where the derivatives are taken at $(0, 0)$ and the remainder $\bar{R}$ satisfies the condition

$$\lim_{(u,v)\to(0,0)} \frac{\bar{R}}{u^2 + v^2} = 0.$$

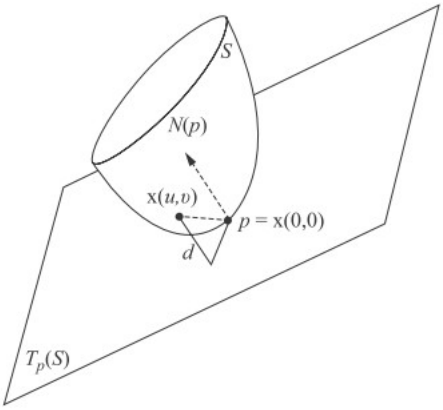

Figure 3-16

It follows that

$$\begin{aligned} d &= \langle \mathbf{x}(u, v) - \mathbf{x}(0, 0), N(p)\rangle \\ &= \tfrac{1}{2}\{\mathbf{x}_{uu}, N(p)\rangle u^2 + 2\langle \mathbf{x}_{uv}, N(p)\rangle uv + \langle \mathbf{x}_{vv}, N(p)\rangle v^2\} + R \\ &= \tfrac{1}{2}(eu^2 + 2fuv + gv^2) + R = \tfrac{1}{2}II_p(w) + R, \end{aligned}$$

where $w = \mathbf{x}_u u + \mathbf{x}_v v$, $R = \langle \bar{R}, N(p)\rangle$, and $\lim_{w\to 0}(R/|w|^2) = 0$.

For an elliptic point p, $II_p(w)$ has a fixed sign. Therefore, for all (u, v) sufficiently near p, d has the same sign as $II_p(w)$; that is, all such (u, v) belong to the same side of $T_p(S)$.

For a hyperbolic point p, in each neighborhood of p there exist points (u, v) and $(\bar{u}, \bar{v})$ such that $II_p(w/|w|)$ and $II_p(\bar{w}/|\bar{w}|)$ have opposite signs (here $\bar{w} = \mathbf{x}_u\bar{u} + \mathbf{x}_v\bar{v}$); such points belong therefore to distinct sides of $T_p(S)$.

Q.E.D.

No such statement as Prop. 1 can be made in a neighborhood of a parabolic or a planar point. In the above examples of parabolic and planar points (cf. Examples 3 and 6 of Sec. 3-2) the surface lies on one side of the tangent plane and may have a line in common with this plane. In the following examples we shall show that an entirely different situation may occur.

Example 2. The "monkey saddle" (see Fig. 3-17) is given by

$$x = u, \quad y = v, \quad z = u^3 - 3v^2u.$$

A direct computation shows that at (0, 0) the coefficients of the second fundamental form are $e = f = g = 0$; the point (0, 0) is therefore a planar point. In any neighborhood of this point, however, there are points in both sides of its tangent plane.

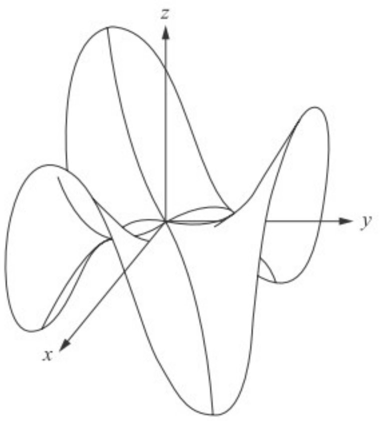

Figure 3-17

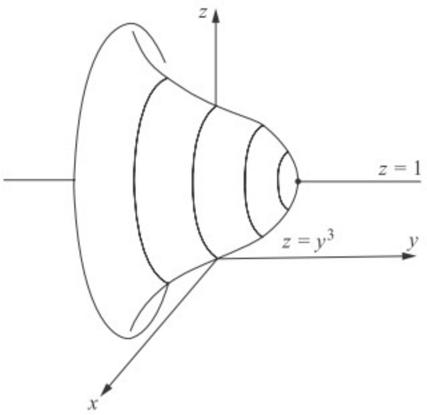

Figure 3-18

Example 3. Consider the surface obtained by rotating the curve $z = y^3$, $-1 < z < 1$, about the line $z = 1$ (see Fig. 3-18). A simple computation shows that the points generated by the rotation of the origin O are parabolic points. We shall omit this computation, because we shall prove shortly (Example 4) that the parallels and the meridians of a surface of revolution are lines of curvature; this, together with the fact that, for the points in question, the meridians (curves of the form $y = x^3$) have zero curvature and the parallel is a normal section, will imply the above statement.

Notice that in any neighborhood of such a parabolic point there exist points in both sides of the tangent plane.

The expression of the second fundamental form in local coordinates is particularly useful for the study of the asymptotic and principal directions. We first look at the asymptotic directions.

Let $\mathbf{x}(u, v)$ be a parametrization at $p \in S$, with $\mathbf{x}(0, 0) = p$, and let $e(u, v) = e$, $f(u, v) = f$, and $g(u, v) = g$ be the coefficients of the second fundamental form in this parametrization.

We recall that (see Def. 9 of Sec. 3-2) a connected regular curve C in the coordinate neighborhood of $\mathbf{x}$ is an asymptotic curve if and only if for any parametrization $\alpha(t) = \mathbf{x}(u(t), v(t))$, $t \in I$, of C we have $II(\alpha'(t)) = 0$, for all $t \in I$, that is, if and only if

$$e(u')^2 + 2fu'v' + g(v')^2 = 0, \quad t \in I. \tag{7}$$

Because of that, Eq. (7) is called the *differential equation of the asymptotic curves.* In the next section we shall give a more precise meaning to this expression. For the time being, we want to draw from Eq. (7) only the following useful conclusion: *A necessary and sufficient condition for a parametrization in a neighborhood of a hyperbolic point* ($\mathrm{eg} - \mathrm{f}^2 < 0$) *to be such that* *the coordinate curves of the parametrization are asymptotic curves is that* $\mathrm{e} = \mathrm{g} = 0$.

In fact, if both curves $u = \text{const.}$, $v = v(t)$ and $u = u(t)$, $v = \text{const.}$ satisfy Eq. (7), we obtain $e = g = 0$. Conversely, if this last condition holds and $f \neq 0$, Eq. (7) becomes $fu'v' = 0$, which is clearly satisfied by the coordinate lines.

We shall now consider the principal directions, maintaining the notations already established.

A connected regular curve C in the coordinate neighborhood of $\mathbf{x}$ is a line of curvature if and only if for any parametrization $\alpha(t) = \mathbf{x}(u(t), v(t))$ of C, $t \in I$, we have (cf. Prop. 3 of Sec. 3-2)

$$dN(\alpha'(t)) = \lambda(t)\alpha'(t).$$

It follows that the functions $u'(t)$, $v'(t)$ satisfy the system of equations

$$\begin{aligned} \frac{fF - eG}{EG - F^2}u' + \frac{gF - fG}{EG - F^2}v' &= \lambda u', \\ \frac{eF - fE}{EG - F^2}u' + \frac{fF - gE}{EG - F^2}v' &= \lambda v'. \end{aligned}$$

By eliminating λ in the above system, we obtain the *differential equation of the lines of curvature*,

$$(fE - eF)(u')^2 + (gE - eG)u'v' + (gF - fG)(v')^2 = 0,$$

which may be written, in a more symmetric way, as

$$\begin{vmatrix} (v')^2 & -u'v' & (u')^2 \\ E & F & G \\ e & f & g \end{vmatrix} = 0. \tag{8}$$

Using the fact that the principal directions are orthogonal to each other, it follows easily from Eq. (8) that *a necessary and sufficient condition for the coordinate curves of a parametrization to be lines of curvature in a neighborhood of a nonumbilical point is that* $F = f = 0$.

Example 4 (*Surfaces of Revolution*)**.** Consider a surface of revolution parametrized by (cf. Example 4 of Sec. 2-3; we have replaced f and g by φ and ψ, respectively)

$$\begin{aligned} \mathbf{x}(u, v) = (\varphi(v)\cos u, \varphi(v)\sin u, \psi(v)), \\ 0 < u < 2\pi, \quad a < v < b, \quad \varphi(v) \neq 0. \end{aligned}$$

The coefficients of the first fundamental form are given by

$$E = \varphi^2, \quad F = 0, \quad G = (\varphi')^2 + (\psi')^2.$$

It is convenient to assume that the rotating curve is parametrized by arc length, that is, that

$$(\varphi')^2 + (\psi')^2 = G = 1.$$

The computation of the coefficients of the second fundamental form is straightforward and yields

$$\begin{aligned} e &= \frac{(\mathbf{x}_u, \mathbf{x}_v, \mathbf{x}_{uu})}{\sqrt{EG - F^2}} = \frac{1}{\sqrt{EG - F^2}} \begin{vmatrix} -\varphi \sin u & \varphi' \cos u & -\varphi \cos u \\ \varphi \cos u & \varphi' \sin u & -\varphi \sin u \\ 0 & \psi' & 0 \end{vmatrix} \\ &= -\varphi\psi' \\ f &= 0, \\ g &= \psi'\varphi'' - \psi''\varphi'. \end{aligned}$$

Since $F = f = 0$, we conclude that the parallels ($v =$ const.) and the meridians ($u =$ const.) of a surface of revolution are lines of curvature of such a surface (this fact was used in Example 3).

Because

$$K = \frac{eg - f^2}{EG - F^2} = -\frac{\psi'(\psi'\varphi'' - \psi''\varphi')}{\varphi}$$

and φ is always positive, it follows that the parabolic points are given by either $\psi' = 0$ (the tangent line to the generator curve is perpendicular to the axis of rotation) or $\varphi'\psi'' - \psi'\varphi'' = 0$ (the curvature of the generator curve is zero). A point which satisfies both conditions is a planar point, since these conditions imply that $e = f = g = 0$.

It is convenient to put the Gaussian curvature in still another form. By differentiating $(\varphi')^2 + (\psi')^2 = 1$ we obtain $\varphi'\varphi'' = -\psi'\psi''$. Thus,

$$K = -\frac{\psi'(\psi'\varphi'' - \psi''\varphi')}{\varphi} = -\frac{(\psi')^2\varphi'' + (\varphi')^2\varphi''}{\varphi} = -\frac{\varphi''}{\varphi}. \qquad \textbf{(9)}$$

Equation (9) is a convenient expression for the Gaussian curvature of a surface of revolution. It can be used, for instance, to determine the surfaces of revolution of constant Gaussian curvature (cf. Exercise 7).

To compute the principal curvatures, we first make the following general observation: *If a parametrization of a regular surface is such that* $F = f = 0$, *then the principal curvatures are given by* e/E and g/G. In fact, in this case, the Gaussian and the mean curvatures are given by (cf. Eqs. (4) and (5))

$$K = \frac{eg}{EG}, \quad H = \frac{1}{2}\frac{eG + gE}{EG}.$$

Since K is the product and $2H$ is the sum of the principal curvatures, our assertion follows at once.

Thus, the principal curvatures of a surface of revolution are given by

$$\frac{e}{E} = -\frac{\psi'\varphi}{\varphi^2} = -\frac{\psi'}{\varphi}, \quad \frac{g}{G} = \psi'\varphi'' - \psi''\varphi'; \tag{10}$$

hence, the mean curvature of such a surface is

$$H = \frac{1}{2}\frac{-\psi' + \varphi(\psi'\varphi'' - \psi''\varphi')}{\varphi}. \tag{11}$$

Example 5. Very often a surface is given as the graph of a differentiable function (cf. Prop. 1, Sec. 2-2) $z = h(x, y)$, where (x, y) belong to an open set $U \subset R^2$. It is, therefore, convenient to have at hand formulas for the relevant concepts in this case. To obtain such formulas let us parametrize the surface by

$$\mathbf{x}(u, v) = (u, v, h(u, v)), \quad (u, v) \in U,$$

where $u = x$, $v = y$. A simple computation shows that

$$\mathbf{x}_u = (1, 0, h_u), \quad \mathbf{x}_v = (0, 1, h_v), \quad \mathbf{x}_{uu} = (0, 0, h_{uu}),$$
$$\mathbf{x}_{uv} = (0, 0, h_{uu}), \quad \mathbf{x}_{vv} = (0, 0, h_{vv}).$$

Thus

$$N(x, y) = \frac{(-h_x, -h_y, 1)}{(1 + h_x^2 + h_y^2)^{1/2}}$$

is a unit normal field on the surface, and the coefficients of the second fundamental form in this orientation are given by

$$e = \frac{h_{xx}}{(1 + h_x^2 + h_y^2)^{1/2}},$$
$$f = \frac{h_{xy}}{(1 + h_x^2 + h_y^2)^{1/2}},$$
$$g = \frac{h_{yy}}{(1 + h_x^2 + h_y^2)^{1/2}}.$$

From the above expressions, any needed formula can be easily computed. For instance, from Eqs. (4) and (5) we obtain the Gaussian and mean curvatures:

$$K = \frac{h_{xx}h_{yy} - h_{xy}^2}{(1 + h_x^2 + h_y^2)^2},$$

$$2H = \frac{(1 + h_x^2)h_{yy} - 2h_x h_y h_{xy} + (1 + h_y^2)h_{xx}}{(1 + h_x^2 + h_y^2)^{3/2}}.$$

There is still another, perhaps more important, reason to study surfaces given by $z = h(x, y)$. It comes from the fact that locally any surface is the graph of a differentiable function (cf. Prop. 3, Sec. 2-2). Given a point p of a surface S, we can choose the coordinate axis of R^3 so that the origin O of the coordinates is at p and the z axis is directed along the positive normal of S at p (thus, the xy plane agrees with $T_p(S)$). It follows that a neighborhood of p in S can be represented in the form $z = h(x, y)$, $(x, y) \in U \subset R^2$, where U is an open set and h is a differentiable function (cf. Prop. 3, Sec. 2-2), with $h(0, 0) = 0$, $h_x(0, 0) = 0$, $h_y(0, 0) = 0$ (Fig. 3-19).

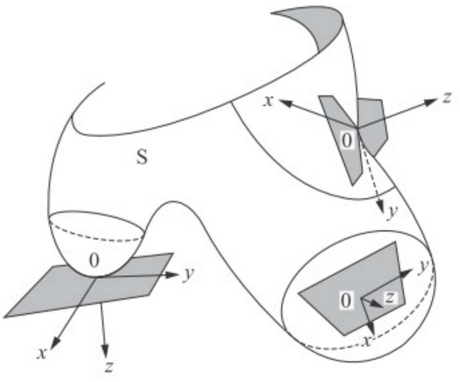

Figure 3-19. Each point of S has a neighborhood that can be written as $z = h(x, y)$.

The second fundamental form of S at p applied to the vector $(x, y) \in R^2$ becomes, in this case,

$$h_{xx}(0, 0)x^2 + 2h_{xy}(0, 0)xy + h_{yy}(0, 0)y^2.$$

In elementary calculus of two variables, the above quadratic form is known as the *Hessian* of h at $(0, 0)$. Thus, the Hessian of h at $(0, 0)$ is the second fundamental form of S at p.

Let us apply the above considerations to give a geometric interpretation of the Dupin indicatrix. With the notation as above, let $\epsilon > 0$ be a small number such that

$$C = \{(x, y) \in T_p(S);\, h(x, y) = \epsilon\}$$

is a regular curve (we may have to change the orientation of the surface to achieve $\epsilon > 0$). We want to show that if p is not a planar point, the curve C is "approximately" similar to the Dupin indicatrix of S at p (Fig. 3-20).

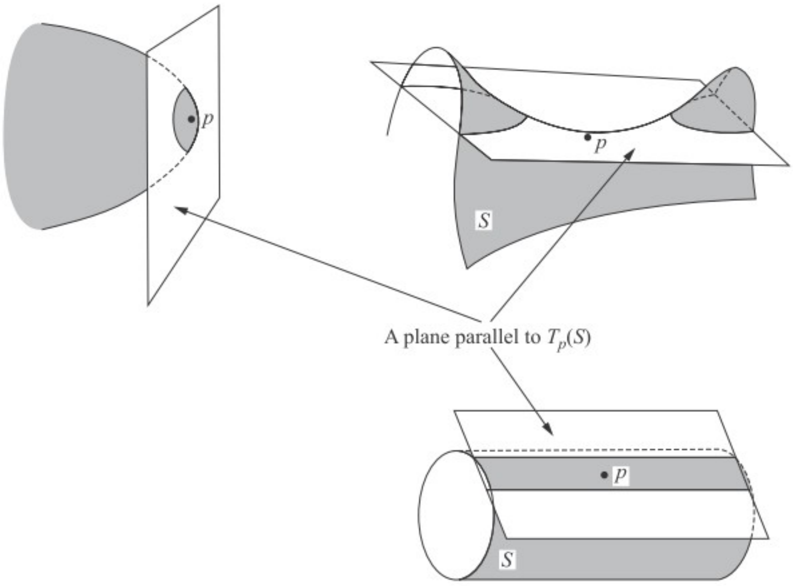

Figure 3-20

To see this, let us assume further that the x and y axes are directed along the principal directions, with the x axis along the direction of maximum principal curvature. Thus, $f = h_{xy}(0, 0) = 0$ and

$$k_1(p) = \frac{e}{E} = h_{xx}(0, 0), \quad k_2(p) = \frac{g}{G} = h_{yy}(0, 0).$$

By developing $h(x, y)$ into a Taylor's expansion about $(0, 0)$, and taking into account that $h_x(0, 0) = 0 = h_y(0, 0)$, we obtain

$$\begin{aligned} h(x, y) &= \tfrac{1}{2}(h_{xx}(0, 0)x^2 + 2h_{xy}(0, 0)xy + h_{yy}(0, 0)y^2) + R \\ &= \tfrac{1}{2}(k_1x^2 + k_2y^2) + R, \end{aligned}$$

where

$$\lim_{(x,y)\to(0,0)} \frac{R}{x^2 + y^2} = 0.$$

Thus, the curve C is given by

$$k_1x^2 + k_2y^2 + 2R = 2\epsilon.$$

Now, if p is not a planar point, we can consider $k_1x^2 + k_2y^2 = 2\epsilon$ as a first-order approximation of C in a neighborhood of p. By using the similarity transformation,

$$x = \bar{x}\sqrt{2\epsilon}, \quad y = \bar{y}\sqrt{2\epsilon},$$

we have that $k_1x^2 + k_2y^2 = 2\epsilon$ is transformed into the curve

$$k_1\bar{x}^2 + k_2\bar{y}^2 = 1,$$

which is the Dupin indicatrix at p. This means that *if* p *is a nonplanar point, the intersection with* S *of a plane parallel to* $\mathrm{T_p(S)}$ *and close to* p *is, in a first-order approximation, a curve similar to the Dupin indicatrix at* p.

If p is a planar point, this interpretation is no longer valid (cf. Exercise 11).

To conclude this section we shall give a geometrical interpretation of the Gaussian curvature in terms of the Gauss map $N: S \to S^2$. Actually, this was how Gauss himself introduced this curvature.

To do this, we first need a definition.

Let S and $\bar{S}$ be two oriented regular surfaces. Let $\varphi: S \to \bar{S}$ be a differentiable map and assume that for some $p \in S$, $d\varphi_p$ is nonsingular. We say that φ is *orientation-preserving* at p if given a positive basis $\{w_1, w_2\}$ in $T_p(S)$, then $\{d\varphi_p(w_1), d\varphi_p(w_2)\}$ is a positive basis in $T_{\varphi(p)}(\bar{S})$. If $\{d\varphi_p(w_1), d\varphi_p(w_2)\}$ is not a positive basis, we say that φ is *orientation-reversing* at p.

We now observe that both S and the unit sphere S^2 are embedded in R^3. Thus, an orientation N on S induces an orientation N in S^2. Let $p \in S$ be such that dN_p is nonsingular. Since for a basis $\{w_1, w_2\}$ in $T_p(S)$

$$dN_p(w_1) \wedge dN_p(w_2) = \det(dN_p)(w_1 \wedge w_2) = K\, w_1 \wedge w_2,$$

the Gauss map N will be orientation-preserving at $p \in S$ if $K(p) > 0$ and orientation-reversing at $p \in S$ if $K(p) < 0$. Intuitively, this means the following (Fig. 3-21): An orientation of $T_p(S)$ induces an orientation of small closed curves in S around p; the image by N of these curves will have the same or the opposite orientation to the initial one, depending on whether p is an elliptic or hyperbolic point, respectively.

To take this fact into account we shall make the convention that the area of the image by N of a region contained in a connected neighborhood $V \subset S$ where $K \neq 0$ is positive if $K > 0$ and negative if $K < 0$ (since V is connected, K does not change sign in V).

Now we can state the promised geometric interpretation of the Gaussian curvature K, for $K \neq 0$.

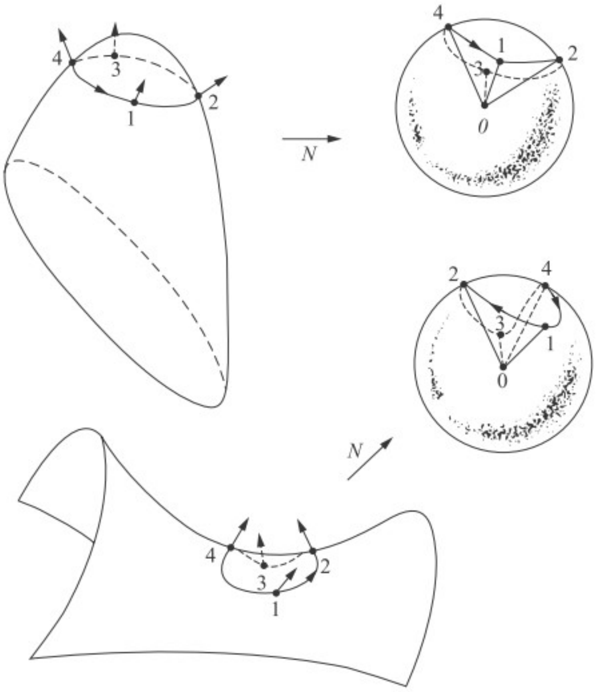

Figure 3-21. The Gauss map preserves orientation at an elliptic point and reverses it at a hyperbolic point.

PROPOSITION 2. *Let* p *be a point of a surface* S *such that the Gaussian curvature* $\mathrm{K(p)} \neq 0$, *and let* V *be a connected neighborhood of* p *where* K *does not change sign. Then*

$$\mathrm{K(p)} = \lim_{\mathrm{A}\to 0} \frac{\mathrm{A}'}{\mathrm{A}},$$

where A *is the area of a region* $\mathrm{B} \subset \mathrm{V}$ *containing* p*,* A′ *is the area of the image of* B *by the Gauss map* $\mathrm{N}\colon \mathrm{S} \to \mathrm{S}^2$*, and the limit is taken through a sequence of regions* $\mathrm{B_n}$ *that converges to* p*, in the sense that any sphere around* p *contains all* $\mathrm{B_n}$*, for* n *sufficiently large.*

Proof. The area A of B is given by (cf. Sec. 2-5)

$$A = \iint_R |\mathbf{x}_u \wedge \mathbf{x}_v|\, du\, dv,$$

where $\mathbf{x}(u, v)$ is a parametrization in p, whose coordinate neighborhood contains V (V can be assumed to be sufficiently small) and R is the region in the uv plane corresponding to B. The area A' of $N(B)$ is

$$A' = \iint_R |N_u \wedge N_v|\, du\, dv.$$

Using Eq. (1), the definition of K, and the above convention, we can write

$$A' = \iint_R K|\mathbf{x}_u \wedge \mathbf{x}_v|\, du\, dv. \qquad \textbf{(12)}$$

Going to the limit and denoting also by R the area of the region R, we obtain

$$\lim_{A\to 0} \frac{A'}{A} = \lim_{R\to 0} \frac{A'/R}{A/R} = \frac{\lim\limits_{R\to 0}(1/R) \iint_R K|\mathbf{x}_u \wedge \mathbf{x}_v|\, du\, dv}{\lim\limits_{R\to 0}(1/R) \iint_R |\mathbf{x}_u \wedge \mathbf{x}_v|\, du\, dv}$$
$$= \frac{K|\mathbf{x}_u \wedge \mathbf{x}_v|}{|\mathbf{x}_u \wedge \mathbf{x}_v|} = K$$

(notice that we have used the mean value theorem for double integrals), and this proves the proposition. **Q.E.D.**

Remark. Comparing the proposition with the expression of the curvature

$$k = \lim_{s\to 0} \frac{\sigma}{s}$$

of a plane curve C at p (here s is the arc length of a small segment of C containing p, and σ is the arc length of its image in the indicatrix of tangents; cf. Exercise 3 of Sec. 1-5), we see that the Gaussian curvature K is the analogue, for surfaces, of the curvature k of plane curves.

EXERCISES

1. Show that at the origin $(0, 0, 0)$ of the hyperboloid $z = axy$ we have $K = -a^2$ and $H = 0$.

***2.** Determine the asymptotic curves and the lines of curvature of the helicoid $x = v\cos u$, $y = v\sin u$, $z = cu$, and show that its mean curvature is zero.

***3.** Determine the asymptotic curves of the catenoid

$$\mathbf{x}(u, v) = (\cosh v \cos u, \cosh v \sin u, v).$$

4. Determine the asymptotic curves and the lines of curvature of $z = xy$.

5. Consider the parametrized surface (Enneper's surface)

$$\mathbf{x}(u, v) = \left(u - \frac{u^3}{3} + uv^2,\ v - \frac{v^3}{3} + vu^2,\ u^2 - v^2\right)$$

D -256- - -* ! - / 5 9 3 1L< 1"
-256- - -* ! - / ! ! " & ! 9 3 4
7 3 1"
" -256- - -* ! - /) . ! ! 2 ! 3
1 L !) . Y 1"

20211013//

orthogonal Procrustes problem

Source: 20211013/Linear Algebra
with Applications.../
Analysis and Linear Algebra
(2021).pdf.

"Obstáculos são aqueles perigos que você vê quando tira os olhos do seu objetivo"
Orlando Drummond (Seu Peru).

"Aquilo que se faz por amor está ~~al~~ sempre além do bem e do mal"
Friedrich Nietzsche

O privilégio de um é uma desvantagem para o outro.

Books:

20211013/Lectures on Differential Topology/
Lectures on Differential Topology.pdf

Pág. 248(279) – 14.7

E-P characteristic of tubular neighbourhoods and the Gauss Map. (Map).

Pág. 249(280) – Corollary 14.10.

/%
! -256- - -* ! - /% 2 ! 3 ! " ? ! 2 D ! 9
! ! 4 1 ! " 211 ! 2 ! ! ! ! 3 1"

D -256- - -* ! - /% ? 4 ! 9 1 3 1"
(' / ' 7! !1

The lemma follows by using two adapted gradient fields to compute the relative characteristics.

Note that $\chi(D^n) = 1$: use a Morse function on $(D^n, \emptyset, S^{n-1})$ with just one minimum.

If X is boundaryless and Y is with boundary, then the very same argument used when Y is boundaryless allows us to extend the multiplicative property.

Lemma 14.8.

$$\chi(X \times Y) = \chi(X)\chi(Y) \ .$$

In particular,

$$\chi(X \times D^n) = \chi(X) \ .$$

14.6.1. Additive property of χ. If (W, V_0, V_1), (W', V_0', V_1') are triads and $\phi : V_1 \to V_0'$ is a diffeomorphism, we get a new composite triad (W'', V_0, V_1'), where $W'' = W \amalg_\phi W'$. Any couple of relative fields v and v' with isolated zeros on the given two triads, respectively, can be glued together to produce a relative field v'' having as zeros the union of the zeros of v and v', each one keeping its index. Then we have

$$\chi(W'', V_0, V_1') = \chi(W, V_0, V_1) + \chi(W', V_0', V_1') \ .$$

This additive property of χ has remarkable consequences.

14.6.2. A baby TQFT. In Section 10.8, we have roughly outlined the axioms of a so-called TQFT and posed the question about the existence of any "nontrivial" one. We use χ to provide a first nontrivial example. Consider $\mathbf{CAT}_\eta(n+1)$. Associate to every object M the vector space $Z(M) = \mathbb{C}$. To every arrow f, carried by a triad (W, M_0, M_1), associate the unitary $\mathbb{C}$-linear map

$$Z(f) : Z(M_0) \to Z(M_1), \ \ z \to e^{i\chi(W, M_0)} z \ .$$

By the additive property of χ, it is easy to check that all axioms are satisfied. This shows, at least, that there are not logical contradictions within the given pattern of axioms.

14.7. E-P characteristic of tubular neighbourhoods and the Gauss map

The identity $\chi(X \times D^n) = \chi(X)$ is a special case of the following.

Proposition 14.9. *Let $p : U \to X$ be a closed tubular neighbourhood of a submanifold X of some Y. Then $\chi(U) = \chi(X)$.*

Proof. It is enough to show the equality for an ϵ-neighbourhood $N_\epsilon(X)$ of the zero section X of a vector bundle $\pi : E \to X$ endowed with a field of positive definite scalar products on every fibre. Let v be a tangent field on X with nondegenerate zeros. Define the field on $N_\epsilon(X)$,

$$w(z) = (z - p(z)) + v(p(z)) \ .$$

We check that w is a field on the triad $(N_\epsilon(X), \emptyset, \partial N_\epsilon(X))$ and that the zeros of w coincide with the zeros of v, are nondegenerate, and keep the sign. The proposition follows. ■

In the special case $X \subset \mathbb{R}^k$, assume that U has been constructed using the standard metric on $\mathbb{R}^k$. By removing from the interior of U a system of pairwise disjoint small open disks $\mathcal{D}_p$ around each zero of w, we get a manifold W with boundary $(\amalg_p S^{k-1}_p) \amalg \partial U$ on which the normalized field $\mathfrak{w} := w/||w||$ is well-defined, as well as the map $\mathfrak{w} : W \to S^{k-1}$. The restriction of $\mathfrak{w}$ to ∂U is a field of unitary vectors pointing out from U (in fact normal to the boundary). This restriction, say $g_{\partial U}$, is called the *Gauss map of the hypersurface* ∂U.

Corollary 14.10. *Let $p : U \to X$ be a tubular neighbourhood of X in $\mathbb{R}^k$. Then*

$$\chi(X) = \deg(g_{\partial U})$$

where $g_{\partial U}$ is the Gauss map of the hypersurface ∂U.

The corollary follows by computing the characteristic as the sum of zero indices and using the bordism invariance of the degree.

If X itself is an oriented hypersurface in $\mathbb{R}^k$, $k \geq 2$, we can define its Gauss map $g_X : X \to S^{k-1}$, where this field of oriented unitary normal vectors along X is determined by the fact that, for every $x \in X$, $g_X(x) \oplus T_xX = T_x\mathbb{R}^k$ as oriented vector spaces, where $\mathbb{R}^k$ is endowed with the standard orientation. In this case, ∂U consists of two parallel copies of X with opposite orientations. The last corollary specializes to the following:

Corollary 14.11. *Let X be an oriented hypersurface of $\mathbb{R}^k$.*

(1) *If k is even, then $\chi(X) = 0$.*

(2) *If k is odd, then*

$$\chi(X) = 2\deg(g_X)$$

where g_X is the Gauss map of X.

The corollary follows by using that the degree of the antipodal map $a_{k-1} : S^{k-1} \to S^{k-1}$ is equal to $(-1)^k$.

Remark 14.12. We can compute inductively the characteristic of real and complex projective spaces by decomposing $\mathbf{P}^n(K)$ as the union of a tubular neighbourhood of $\mathbf{P}^{n-1}(K) \subset \mathbf{P}^n(K)$ and its complement and applying Proposition 14.9 together with the additivity of χ.

14.8. Nontriviality of $\eta_\bullet$ and $\Omega_\bullet$

The integer E-P characteristic is *not* invariant up to bordism. For example, $[S^2] = [S^1 \times S^1] = 0 \in \eta_2$, but $\chi(S^2) = 2 \neq 0 = \chi(S^1 \times S^1)$. On the other hand, the E-P characteristic mod(2) is bordism invariant.

Proposition 14.13. *Let* $[X] = 0 \in \eta_n$. *Then* $\chi_{(2)}(X) = 0 \in \mathbb{Z}/2\mathbb{Z}$.

Proof. If n is odd, we know in general that $\chi(X) = 0$. Assume that n is even. Let $X = \partial W$, W being a compact manifold with boundary. Take the double $D(W)$. The double $D(W)$ can be presented as the composition of the triad

$$(X \times D^1, \emptyset, (X \times \{-1\} \amalg (X \times \{1\})))$$

followed by two copies of the triad

$$(W, X, \emptyset)$$

glued to $X \times D^1$ along $X \times \{\pm 1\}$, respectively. By the additive property

$$\chi(D(W)) = \chi(X \times D^1) + 2\chi(W, X, \emptyset) \ .$$

By Lemmas 14.8 and 14.7 and the fact that $n + 1$ is odd and the double is boundaryless, we have

$$\chi(X) = \chi(D(W)) - 2\chi(W, X, \emptyset) = 2\chi(W) \in \mathbb{Z}$$

so that

$$\chi_{(2)}(X) = 0 \in \mathbb{Z}/2\mathbb{Z} \ . \qquad \blacksquare$$

As an immediate corollary, we have the *nontriviality of* η_{2n} *and* Ω_{4n}.

Corollary 14.14. *For every* $n \geq 1$, $\eta_{2n} \neq 0$ *and* $\Omega_{4n} \neq 0$.

Proof. We know that $\chi(\mathbf{P}^{2n}(\mathbb{R})) = 1$; hence $[\mathbf{P}^{2n}(\mathbb{R})] \neq 0 \in \eta_{2n}$. Similarly $\chi_{(2)}(\mathbf{P}^{2n}(\mathbb{C})) = 1$; hence $[\mathbf{P}^{2n}(\mathbb{C})] \neq 0 \in \Omega_{4n}$. ■

By using the multiplicative property of χ and the obvious fact that it is additive under disjoint union, we also have the following.

Gauss map of the hypersurface $\mathcal{V}$.

Obs: Riemann

Pág. 13 - 1.7

Tangent vector fields,
Riemannian metrics,
gradient fields.

$$\widetilde{\pi}|U \circ X = id_U .$$

_ 20211013/ Satellitengeodäsie../
Satellitengeodäsie...pdf
Search: geometrischen
Obs: Livro todo!!! German language.

Pág. 124 (142) 7
Transformationen

Abb 7.1. Imagem.
Pág. 126 (144) Abb 7.2
Pág. 127 (145) Abb 7.3

_ 20211013/ Satellitengeodäsie/
Satellite Geodesy - Foundations..pdf

/'
-256- - -* ! - /' ! "Y 7 " ! "
3 .! " ! ! 7 "Y 1"
(' ' 9 ! !

von Bedeutung sind, findet man im Band 4 *Bezugssysteme* der *Grundlagen der Physikalischen und Mathematischen Geodäsie.*

Die Koordinaten der Beobachtungsstation $\mathbf{R}_S$ beziehen sich auf ein erdfestes Bezugssystem und die des topozentrischen Ortsvektors $\mathbf{R}$ entweder auf ein raumfestes oder ein erdfestes Bezugssystem. Zur Modellbildung muss deshalb ein einheitlicher Bezugsrahmen vereinbart werden, auf den sich die Koordinaten der Vektoren der geometrischen Grundgleichung beziehen. Im Prinzip kann auch ein beliebig bewegtes Bezugssystem gewählt werden, also beispielsweise auch das erdfeste Bezugssystem. In diesem Fall muss die Bewegungsgleichung (7.2) um die entsprechenden Trägheitskräfte erweitert werden.

Im Folgenden wird der Fall angenommen, dass die in einem erdfesten Bezugsrahmen gewonnenen Beobachtungen des topozentrischen Ortsvektors $\mathbf{R}(t)$ wie auch die geozentrischen Ortsvektoren der Beobachtungsstationen $\mathbf{R}_S(t)$ in das Newton-System transformiert werden, in dem die Bewegungsgleichung für die geozentrische Satellitenposition $\mathbf{r}(t)$ gilt. Häufig wird als Bezugssystem, in dem die Beobachtungsgleichungen formuliert werden, das wahre raumfeste Äquatorsystem zum momentanen Zeitpunkt t (CRFt, auch mit $\tilde{\mathcal{A}}(t)$ bezeichnet) gewählt. Man beachte, dass es sich dabei ebenfalls um ein bewegtes Bezugssystem handelt, das sich bzgl. eines raumfesten Bezugssystems um die Effekte der Nutation und Präzession bewegt. Deshalb sind in der Bewegungsgleichung entsprechende Trägheitskräfte zu berücksichtigen. Im Prinzip sind aber auch andere Festlegungen möglich (siehe auch Schneider 1990, 1988, 1999; Newcomb 1892, 1960).

7.1.1 Transformation der Ortsvektoren der Beobachtungsstationen

Als „erdraumbezogenes Bezugssystem“, in dem beispielsweise die Koordinaten der Beobachtungsstationen definiert sind, wird ein vereinbarter erdfester Bezugsrahmen einer bestimmten Epoche t_S gewählt. Verwendet man die Bezugsrahmen des IERS (International Earth Rotation and Reference System Service – McCarthy, Petit (Hrsg.), 2004; IERS Conventions, 2010), so handelt es sich hierbei um den ITRF t_S (ITRF International conventional Terrestrial Reference Frame zur Epoche t_S), auch mit $\mathcal{G}(t_S)$ bezeichnet. Als Bezugssystem, in dem die Beobachtungsgleichungen formuliert werden, wird das wahre raumfeste Äquatorsystem zum momentanen Zeitpunkt t (CRFt, auch mit $\tilde{\mathcal{A}}(t)$ bezeichnet) gewählt.

Die Transformation der Stationskoordinaten $\mathbf{R}_S$, definiert im vereinbarten erdfesten Bezugsrahmen der Epoche t_s, $\mathcal{G}(t_S)$, in den wahren raumfesten Bezugsrahmen (true Celestial Reference Frame CRFt), entspricht dem wahren raumfesten Äquatorsystem $\tilde{\mathcal{A}}(t)$) zum Zeitpunkt t, beinhaltet die Drehungen zufolge der Polbewegung und die tägliche Rotation der Erde, bestehend aus der:

- Transformation vom vereinbarten erdfesten Bezugsrahmen der Epoche t_s (ITRF t_s auch mit $\mathcal{G}(\boldsymbol{t_s})$ bezeichnet) in den wahren erdfesten Bezugsrahmen derselben Epoche t_s (Terrestrial Reference Frame TRF t_s, auch mit $\tilde{\mathcal{G}}(\boldsymbol{t_s})$ bezeichnet): Polbewegung sowie der

- Transformation vom wahren erdfesten Bezugsrahmen t_{s} (true Terrestrial Reference Frame – TRF t_{s}) in den wahren raumfesten Bezugsrahmen der Epoche t (true celestial reference frame – CRFt, auch mit $\tilde{\mathcal{A}}(t)$ bezeichnet): tägliche Drehung.

Die Transformation der rechtwinklig-kartesischen Koordinaten des vereinbarten erdfesten Bezugsrahmens der Epoche t_{S} (ITRFt_{S}) in Koordinaten des wahren raumfesten Bezugsrahmens der Epoche t (CRFt) gelingt durch die Folge von drei Drehungen

$$\begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \end{pmatrix}_{\mathrm{CRF}t} = \mathbf{S}(x_P, y_P, -\,\mathrm{GAST}) \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \end{pmatrix}_{\mathrm{ITRF}t_S} \tag{7.3}$$

mit der Drehmatrix $\mathbf{S}$, die zwei Drehungen zufolge der Polbewegung, ausgedrückt durch die Polkoordinaten (x_P, y_P), und die tägliche Drehung, ausgedrückt durch die scheinbare Sternzeit des Meridians von Greenwich, beschreibt

$$\mathbf{S}(x_P, y_P, -\mathrm{GAST}) := \mathbf{D}_3(-\mathrm{GAST})\ \mathbf{D}_1(y_P)\ \mathbf{D}_2(x_P) \tag{7.4}$$

Die beteiligten Koordinatensysteme sind in der Abb. 7.1 dargestellt. Das wahre raumfeste Äquatorsystem $\tilde{\mathcal{A}}(t)$ ist durch den wahren raumfesten Äquator (true celestial equator) und durch den wahren raumfesten Ephemeridenpol (true Celestial Ephemeris Pole – CEP) in Richtung $(\mathbf{e}_3)_{\tilde{\mathcal{C}}(t_{\mathsf{S}})} \equiv (\mathbf{e}_3)_{\tilde{\mathcal{A}}(t)}$ definiert. Das mittlere (vereinbarte) globale geozentrische Bezugssystem $\mathcal{C}(t_{\mathsf{S}})$ ist durch den vereinbarten mittleren erdfesten Äquator (conventional mean terrestrial equator) und den vereinbarten erdfesten Pol (Conventional Terrestrial Pole – CTP) in Richtung $(\mathbf{e}_3)_{\mathcal{C}(t_{\mathsf{S}})}$ zur Epoche t_{S} definiert. Die Lage des CEP bzgl. dem CTP ist durch die Parameter der Polbewegung (Polkoordinaten (x_P, y_P)) beschrieben.

Der Winkel zwischen der x-Achse $(\mathbf{e}_1)_{\tilde{\mathcal{A}}(t)}$ des wahren raumfesten Äquatorsystems $\tilde{\mathcal{A}}(t)$ (Richtung des wahren Frühlingspunktes zur Epoche t) und der x-Achse $(\mathbf{e}_1)_{\tilde{\mathcal{C}}(t_{\mathsf{S}})}$ des wahren globalen geozentrischen Koordinatensystems $\tilde{\mathcal{C}}(t_{\mathsf{S}})$ (die in der mittleren astronomischen Meridianebene von Greenwich liegt) ist durch den Stundenwinkel $t_{\tilde{\mathcal{C}}(t_{\mathsf{S}})}$ des wahren Frühlingspunktes von Greenwich, der scheinbaren Sternzeit des Meridians von Greenwich (GAST), definiert. Die scheinbare Sternzeit Greenwich GAST unterliegt einer ständigen i. Allg. ungleichförmigen Veränderung wegen der Rotation der Erde. Sie kann aus der mittleren Sternzeit Greenwich (Greenwich Mean Siderial Time – GMST) berechnet werden. Die mittlere Sternzeit Greenwich GMST selbst erhält man aus einem zeitlichen Polynomansatz, der aus der Differenz der aus Beobachtungen erhaltenen Weltzeit UT1 (Universal Time 1 – UT1) und der Atomzeit UTC (Universal Time Coordinated – UTC) abgeleitet wird:

$$\Delta t = UTC - UT1. \tag{7.5}$$

Die Integration von Δt über einen Tag ergibt die Veränderungen der Tageslänge (Length Of the Day – LOD). Die Zeitdifferenz Δt kann nicht in Form

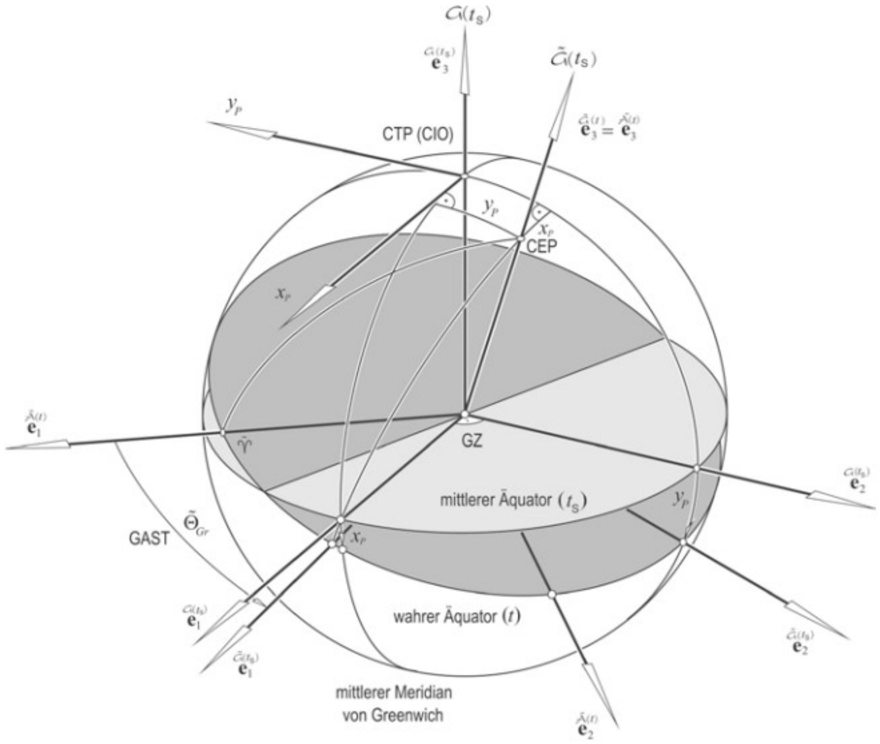
$\mathcal{G}(t_{\mathrm{S}})$
$\mathbf{e}_3^{\mathcal{G}(t_{\mathrm{S}})}$
$\tilde{\mathcal{G}}(t_{\mathrm{S}})$
y_P
CTP (CIO)
$\mathbf{e}_3^{\tilde{\mathcal{G}}(t)} = \mathbf{e}_3^{\tilde{\mathcal{A}}(t)}$
y_P
x_P
CEP
x_P
$\mathbf{e}_1^{\tilde{\mathcal{A}}(t)}$
$\tilde{\Upsilon}$
GZ
$\mathbf{e}_2^{\mathcal{G}(t_{\mathrm{S}})}$
mittlerer Äquator (t_{S})
y_P
GAST
$\tilde{\Theta}_{Gr}$
x_P
$\mathbf{e}_1^{\mathcal{G}(t_{\mathrm{S}})}$
wahrer Äquator (t)
$\mathbf{e}_2^{\tilde{\mathcal{G}}(t_{\mathrm{S}})}$
$\mathbf{e}_1^{\tilde{\mathcal{G}}(t_{\mathrm{S}})}$
mittlerer Meridian
von Greenwich
$\mathbf{e}_2^{\tilde{\mathcal{A}}(t)}$

Abb. 7.1 Transformation vom vereinbarten erdfesten Bezugsrahmen der Epoche t_{S} (ITRF t_{S}, auch mit $\mathcal{G}(t_{\mathrm{S}})$ bezeichnet) in den wahren raumfesten Bezugsrahmen der Epoche t (CRFt, auch mit $\tilde{\mathcal{A}}(t)$ bezeichnet)

einer analytischen Funktion angegeben werden. Sie muss zusammen mit den Polkoordinaten (x_P, y_P) durch Beobachtung erfasst werden und kann erst im Nachhinein bekannt gegeben werden.

7.1.2 Transformation der topozentrischen Ortsvektoren zum Satelliten

Für die Transformation des topozentrischen Satellitenortes $\mathbf{R}(t)$ in das wahre raumfeste Äquatorsystem zum momentanen Zeitpunkt t, $\tilde{\mathcal{A}}(t)$, ist entscheidend, in welchem System die topozentrischen Richtungen (Rektaszensionen und Deklinationen) zum Satelliten erhalten worden waren. I. Allg. wird es sich um einen mittleren raumfesten Bezugsrahmen zur Epoche $t_0 = \mathrm{J}2000{,}0$ (ICRFt_0) handeln. Die Transformation der rechtwinklig-kartesischen Koordinaten bzgl. des mittleren raumfesten Bezugsrahmens der Epoche $t_0 = \mathrm{J}2000{,}0$ (ICRFt_0, auch mit $\mathcal{A}(t_0)$ bezeichnet) in Koordinaten des wahren raumfesten Äquatorsystems zum momentanen Zeitpunkt t (CRFt, auch mit $\tilde{\mathcal{A}}(t)$ bezeichnet) gelingt durch die Folge von sechs Drehungen $\mathbf{D}$, die in zwei Drehmatrizen, die Präzession $\mathbf{P}$ und die Nutation $\mathbf{N}$ zusammengefasst sind (Mathews et al. 2002):

- Transformation vom mittleren raumfesten Bezugsrahmen der Epoche $t_0 = \text{J2000,0}$ ($\mathcal{A}(t_0)$), in den mittleren raumfesten Bezugsrahmen t ($\mathcal{A}(t)$): ***Präzession,***
- Transformation vom mittleren raumfesten Bezugsrahmen der Epoche t (ICRFt) in den wahren raumfesten Bezugsrahmen zur selben Epoche t ($\tilde{\mathcal{A}}(t)$): ***Nutation.***

Die gesamte Transformation lautet:

$$\begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \end{pmatrix}_{\text{CRF}t} = \mathbf{D}(-\zeta, \vartheta, -z, \varepsilon_s, -\Delta\psi, -\varepsilon_s - \Delta\varepsilon) \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \end{pmatrix}_{\text{ICRF}t_0} \tag{7.6}$$

mit der Drehmatrix

$$\mathbf{D}(-\zeta, \vartheta, -z, \varepsilon_s, -\Delta\psi, -\varepsilon_s - \Delta\varepsilon) = \mathbf{N}(\varepsilon_s, -\Delta\psi, -\varepsilon_s - \Delta\varepsilon)\ \mathbf{P}(-\zeta, \vartheta, -z). \tag{7.7}$$

Die ersten drei hintereinandergeschalteten Drehungen beschreiben nach Newcomb die allgemeine (lunisolare und planetare) Präzession $\mathbf{P}$ für den Zeitraum zwischen den Epochen $t_0 = \text{J2000,0}$ und t, (Abb. 7.2)

$$\mathbf{P}(-\zeta, \vartheta, -z) := \mathbf{D}_3(-z)\ \mathbf{D}_2(\vartheta)\ \mathbf{D}_3(-\zeta) \tag{7.8}$$

mit den äquatorialen Präzessionsparametern

$$\begin{aligned} \zeta &= 2306{,}''2181\,T + 0{,}''30188\,T^2 + 0{,}''017998\,T^3, \\ \vartheta &= 2004{,}''3109\,T - 0{,}''42665\,T^2 - 0{,}''041833\,T^3, \\ z &= 2306{,}''2181\,T + 1{,}''09468\,T^2 + 0{,}''018203\,T^3, \end{aligned} \tag{7.9}$$

wobei die Zeitdifferenz T in Julianischen Jahrhunderten der baryzentrischen dynamischen Zeit TDB ***(Barycentric Dynamical Time – TDB)*** einzusetzen ist

$$T = (t - J2000{,}0)\big/36525{,}0. \tag{7.10}$$

Die Größe t ist das Julianische Datum der Berechnungsepoche. Diese Parameter entstammen einem numerischen Modell, das von der IAU im Jahre 1976 verbindlich angenommen worden ist. Obwohl in der Zwischenzeit genauere Werte bekannt sind, werden diese vereinbarten Parameter noch beibehalten, um die Gültigkeit und Konsistenz der Systeme über einen angemessenen Zeitraum zu gewährleisten.

Die Transformation vom mittleren raumfesten Äquatorsystem $\mathcal{A}(t)$ ins wahre raumfeste Äquatorsystem $\tilde{\mathcal{A}}(t)$ wird durch die astronomische Nutation $\mathbf{N}$ zur Epoche t beschrieben:

$$\begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \end{pmatrix}_{\tilde{\mathcal{A}}(t)} = \mathbf{N}(\varepsilon_s, -\Delta\psi, -\varepsilon_s - \Delta\varepsilon) \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \end{pmatrix}_{\mathcal{A}(t)}. \tag{7.11}$$

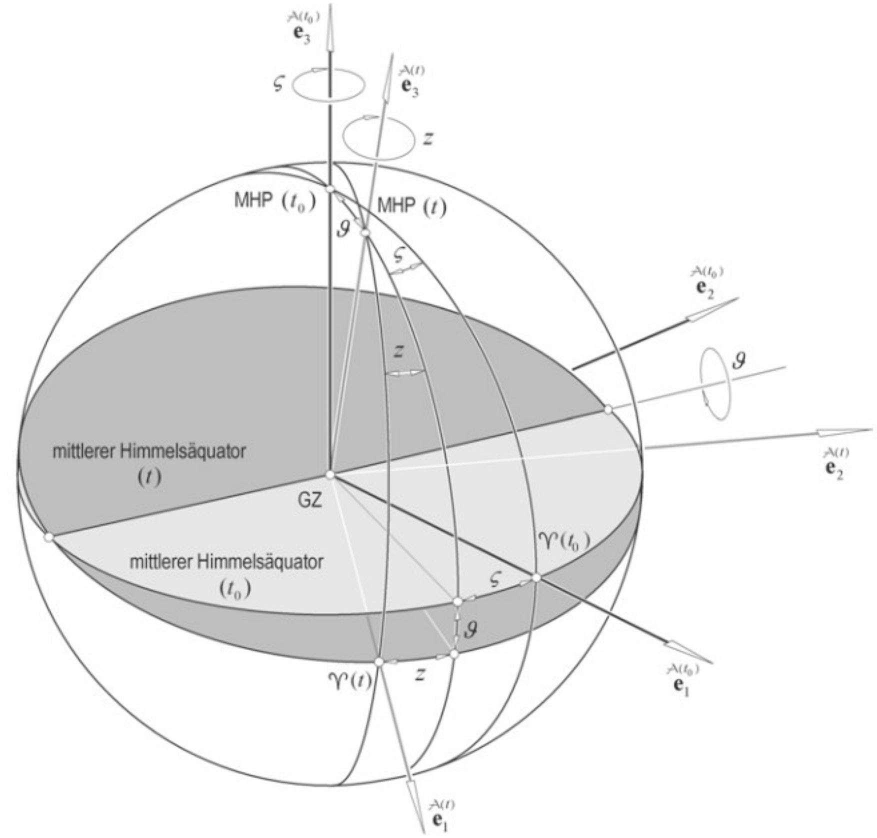
$\overset{\mathcal{A}(t_0)}{\mathbf{e}_3}$
ς
$\overset{\mathcal{A}(t)}{\mathbf{e}_3}$
z
MHP (t_0)
MHP (t)
ϑ
ς
$\overset{\mathcal{A}(t_0)}{\mathbf{e}_2}$
z
ϑ
mittlerer Himmelsäquator
(t)
$\overset{\mathcal{A}(t)}{\mathbf{e}_2}$
GZ
$\Upsilon(t_0)$
mittlerer Himmelsäquator
(t_0)
ς
ϑ
$\Upsilon(t)$
z
$\overset{\mathcal{A}(t_0)}{\mathbf{e}_1}$
$\overset{\mathcal{A}(t)}{\mathbf{e}_1}$

Abb. 7.2 Definition der Newcombschen Präzessionswinkel

Sie setzt sich ebenfalls aus drei Einzeldrehungen zusammen (Abb. 7.3):

$$\mathbf{N}(\varepsilon_s, -\Delta\psi, -\varepsilon_s - \Delta\varepsilon) := \mathbf{D}_1(-\varepsilon_s - \Delta\varepsilon)\mathbf{D}_3(-\Delta\psi)\mathbf{D}_1(\varepsilon_s). \qquad (7.12)$$

Die Drehungen enthalten die mittlere Schiefe der Ekliptik ε_s und die astronomischen Nutationskomponenten in Länge $\Delta\psi$ sowie in Schiefe $\Delta\varepsilon$. Diese Parameter basieren auf einem numerischen Modell von Wahr (Wahr 1981), das 1980 von der IAU als verbindlich angenommen worden ist. Die mittlere Schiefe der Ekliptik zur Epoche t, ε_s, die den Winkel zwischen dem mittleren Ekliptiksystem und dem mittleren Äquatorsystem repräsentiert, kann aus der folgenden Formel berechnet werden:

$$\varepsilon_s = 23^\circ 26' 21'', 448 - 46'', 8150\, T - 0,'' 00059\, T^2 + 0,'' 001813\, T^3. \qquad (7.13)$$

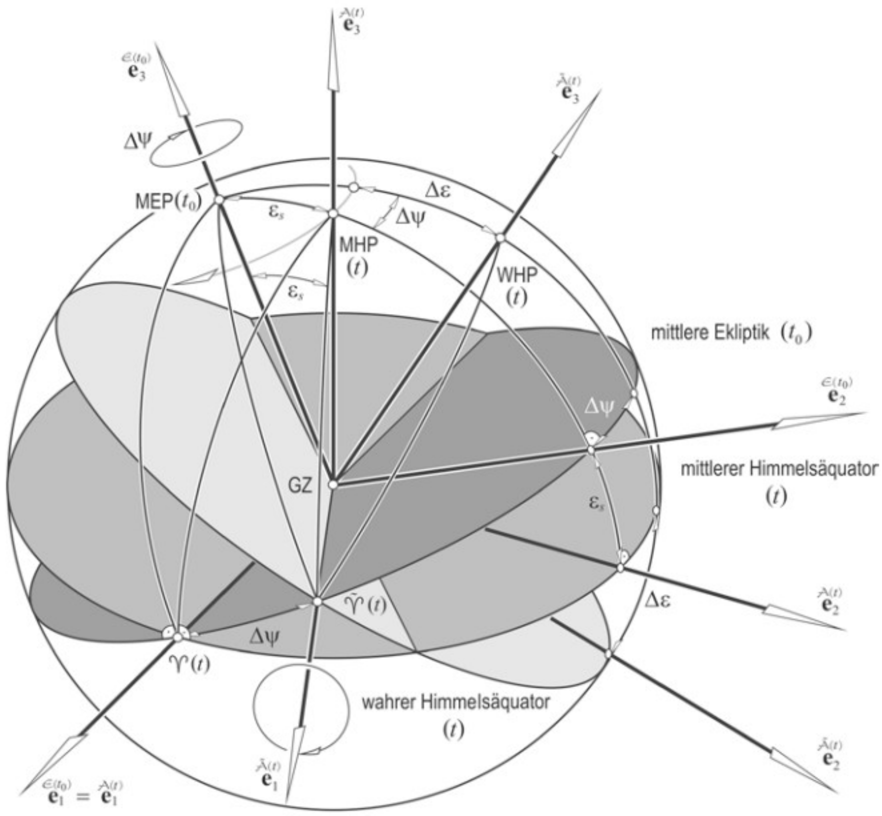
$\overset{\mathcal{A}(t)}{\mathbf{e}}_3$
$\overset{\mathcal{E}(t_0)}{\mathbf{e}}_3$
$\Delta\psi$
$\overset{\tilde{\mathcal{A}}(t)}{\mathbf{e}}_3$
$\Delta\varepsilon$
MEP(t_0)
ε_s
$\Delta\psi$
MHP (t)
ε_s
WHP (t)
mittlere Ekliptik (t_0)
$\overset{\mathcal{E}(t_0)}{\mathbf{e}}_2$
$\Delta\psi$
mittlerer Himmelsäquator (t)
GZ
ε_s
$\tilde{\Upsilon}(t)$
$\Delta\varepsilon$
$\overset{\mathcal{A}(t)}{\mathbf{e}}_2$
$\Delta\psi$
$\Upsilon(t)$
wahrer Himmelsäquator (t)
$\overset{\tilde{\mathcal{A}}(t)}{\mathbf{e}}_2$
$\overset{\tilde{\mathcal{A}}(t)}{\mathbf{e}}_1$
$\overset{\mathcal{E}(t_0)}{\mathbf{e}}_1 = \overset{\mathcal{A}(t)}{\mathbf{e}}_1$

Abb. 7.3 Transformation vom mittleren raumfesten Äquatorsystem $\mathcal{A}(t)$ ins wahre raumfeste Äquatorsystem $\widetilde{\mathcal{A}}(t)$

Die wahre Ekliptschiefe zur Epoche t ergibt sich aus der Summe $\varepsilon_s + \Delta\varepsilon$. Die astronomischen Nutationskomponenten in Länge $\Delta\psi$ und in Schiefe $\Delta\varepsilon$ können aus Reihenentwicklungen entnommen werden. Für die ersten Terme der Nutation in Länge ergibt sich

$$\begin{aligned}\Delta\psi = (-17'',1996 - 0'',01742\,T)\sin(\Omega)+\\ +(0'',2062 + 0'',00002\,T)\sin(2\Omega)+\\ +(0'',0046 + 0'',00000\,T)\sin(-2l + 2F + \Omega) + \ldots\end{aligned} \tag{7.14}$$

und entsprechend für die Nutation in Schiefe:

$$\begin{aligned}\Delta\varepsilon = (9'',2025 + 0'',00089\,T)\cos(\Omega)+\\ +(-0'',0895 + 0'',00005\,T)\cos(2\Omega)+\\ +(-0'',0024 + 0'',00000\,T)\cos(-2l + 2F + \Omega) + \ldots\end{aligned} \tag{7.15}$$

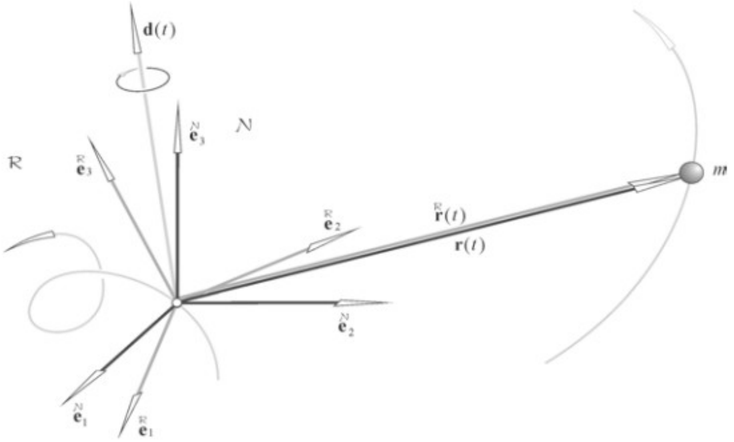
$\mathbf{d}(t)$
$\overset{\mathcal{N}}{\mathbf{e}}_3$
$\mathcal{N}$
$\mathcal{R}$
$\overset{\mathcal{R}}{\mathbf{e}}_3$
m
$\overset{\mathcal{R}}{\mathbf{e}}_2$
$\overset{\mathcal{R}}{\mathbf{r}}(t)$
$\mathbf{r}(t)$
$\overset{\mathcal{N}}{\mathbf{e}}_2$
$\overset{\mathcal{N}}{\mathbf{e}}_1$
$\overset{\mathcal{R}}{\mathbf{e}}_1$

Abb. 7.4 Ortsvektoren $\mathbf{r}(t)$ und $\overset{\mathcal{R}}{\mathbf{r}}(t)$ eines Teilchens, bezogen auf die Bezugssysteme $\mathcal{N}$ und $\mathcal{R}$

mit der mittleren Knotenlänge des Mondes Ω, der mittleren Anomalie des Mondes, bezogen auf das Perigäum l und dem mittleren Knotenabstand des Mondes F. Die vollständigen Reihenentwicklungen sind beispielsweise in den IERS-Konventionen gegeben, wie auch verbesserte Theorien für Präzession und Nutation (IERS Conventions 2010).

7.2 Bewegungsgleichung des Satelliten im wahren Äquatorsystem

Für die Formulierung der Bewegungsgleichung des Satelliten $\mathbf{r}(t)$ im wahren raumfesten Äquatorsystem $\widetilde{\mathcal{A}}(t)$ müssen die Trägheitskräfte berücksichtigt werden. Der geozentrische Ortsvektor $\mathbf{r}(t)$ hängt über die Bewegungsgleichung

$$\ddot{\mathbf{r}}(t) = \frac{d^2\mathbf{r}}{dt^2} = \mathbf{a}(t;\ \boldsymbol{\alpha}, \mathbf{x}_\mathrm{I},\ \mathbf{x}_\mathrm{R}) \tag{7.16}$$

bezogen auf ein im Geozentrum definiertes Newton-System (im vorliegenden Fall das mittlere raumfeste Äquatorsystem) $\mathcal{N} \equiv \mathcal{A}(t)$, von den auf den Satelliten wirkenden Feldstärken des Gravitationsfeldes und der Störbeschleunigungen ab

$$\mathbf{a}(t_k;\ \mathbf{x}_i,\ \mathbf{x}_\mathrm{R}) = \mathbf{a}_\mathrm{K}(t_k) + \mathbf{a}_\mathrm{R}(t_k) + \mathbf{a}_\mathrm{St}(t_k) \tag{7.17}$$

Die Zeitableitung der Bahnfunktion, $\ddot{\mathbf{r}}(t) = d^2\mathbf{r}/dt^2$, bezieht sich dabei auf das Newton-System $\mathcal{A}(t)$. Die Drehmatrix

$$\mathbf{D}(-\zeta, \vartheta, -z, \varepsilon_s, -\Delta\psi, -\varepsilon_s - \Delta\varepsilon) = \mathbf{N}(\varepsilon_s, -\Delta\psi, -\varepsilon_s - \Delta\varepsilon)\ \mathbf{P}(-\zeta, \vartheta, -z) \tag{7.18}$$

D -256- - -* ! - /' ! 7 " 3 < "! J
" J 211 ! % " $$ @! " 7 3 % 1"

(%'/ !" '

-256- - -* ! - /' 7 D ! 7 " 3
< "! J " J 211 ! % 1"

" -256- - -* ! - /' 3 ! 7 " 3 1
$ 1"

-256- - -* ! - /' :! "D . ! ! ! " 3
znj„f_djh…†hjh…‡f_„†hji†ydfcjf_h…†fˆcd`cf_„†ecfˆdi‰Šk_`uˆ‹n‰c`†Œ•Ž••‘lyiz

-256- - -* ! - /' 7 " 3 ! =! 1
1"

-256- - -* ! - /' ! !
7 "Y l ! ! ! 7 " 3 :! "D " 7 "Y J ! D A @ < " "
5 5 1) : " D 1"

(%'/

obs: Livro todo!!!

cont.

- 20211013/Satellitengeodäsie/

obs: Diretório todo!!!

- 20211013/ Geodesy/
  - Gunter Seeber - Satellite ... (2003).pdf
  - Physical Geodesy ... (2005).pdf
  - Handbook of Mathematical Geodesy ... (2018).pdf

    938 pages.
  - Geodesy for Planet Earth ... (2012).pdf

    991 pages.
  - Mathematische Geodäsie - Mathematical Geodesy.pdf

    1459 pages.

obs: Livro todo!!!

obs: Diretório todo também!!!

data

20211013/ Geodesy/...

→ Tom Herring - Geodesy..(2009).pdf

Pág. 14(20) Figure 2

Ellipsoidal coordinates

Pág. 15 (21) Figure 3

Relative geometry of geoid and ellipsoid

→ orthometric height

Pág. 14(20)

Their relation to Cartesian coordinates is given by

$$x = \sqrt{u^2 + E^2} \sin \delta \cos \lambda$$

$$y = \sqrt{u^2 + E^2} \sin \delta \cos \lambda \qquad (2)$$

$$z = u \cos \delta$$

Pág. 19(25)

Gauss' divergence theorem (analogous to Stoke's theorem, eqn [19]),

$$\iiint_V \nabla \cdot f \, dv = \iint_S f \cdot n \, ds \qquad [26]$$

/$
! -256- - -* ! - /$ 9 : 7 " 3 9 !
7 1 3 = A / 1"
(' < = 1 "! " !

general relativistic effects) and will not be expounded here. We are primarily concerned with gravitation on and external to the solid and liquid Earth since this is the domain of most applications. The internal field can also be modeled for specialized purposes (such as submarine navigation), but internal geophysical modeling, for example, is done usually in terms of the sources (mass density distribution), rather than the resulting field.

3.02.1.2 Coordinate Systems

Modeling the Earth's gravitational field depends on the choice of coordinate system. Customarily, owing to the Earth's general shape, a spherical polar coordinate system serves for most applications, and virtually all global models use these coordinates. However, the Earth is slightly flattened at the poles, and an ellipsoidal coordinate system has also been advocated for some near-Earth applications. We note that the geodetic coordinates associated with a geodetic datum (based on an ellipsoid) are never used in a foundational sense to model the field since they do not admit to a separation-of-variables solution of Laplace' differential equation (Section 3.02.4.1).

Spherical polar coordinates, described with the aid of **Figure 1**, comprise the spherical colatitude, θ, the longitude, λ, and the radial distance, r. Their relation to Cartesian coordinates is

$$\begin{aligned} x &= r\sin\theta\cos\lambda \\ y &= r\sin\theta\sin\lambda \\ z &= r\cos\theta \end{aligned} \qquad [1]$$

Considering Earth's polar flattening, a better approximation, than a sphere, of its (ocean) surface is an ellipsoid of revolution. Such a surface is generated by rotating an ellipse about its minor axis (polar axis). The two focal points of the best-fitting, Earth-centered ellipsoid (ellipse) are located in the equator about $E = 522$ km from the center of the Earth. A given ellipsoid, with specified semiminor axis, b, and linear eccentricity, E, defines the set of ellipsoidal coordinates, as described in **Figure 2**. The longitude is the same as in the spherical case. The colatitude, δ, is the complement of the so-called reduced latitude; and the distance coordinate, u, is the semiminor axis of the confocal ellipsoid through the point in question. We call (δ, λ, u) ellipsoidal coordinates; they are also known as spheroidal coordinates, or Jacobi ellipsoidal coordinates. Their relation to Cartesian coordinates is given by

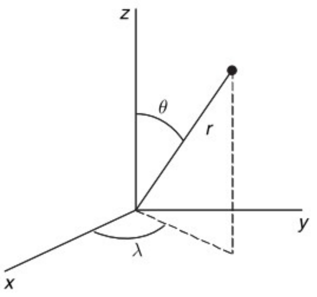

Figure 1 Spherical polar coordinates.

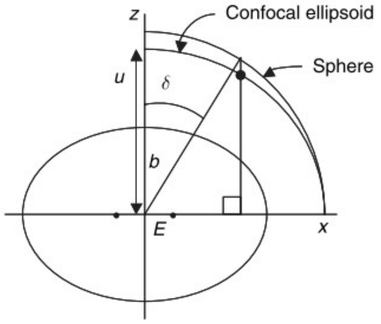

Figure 2 Ellipsoidal coordinates.

$$\begin{aligned} x &= \sqrt{u^2+E^2}\sin\delta\cos\lambda \\ y &= \sqrt{u^2+E^2}\sin\delta\sin\lambda \\ z &= u\cos\delta \end{aligned} \qquad [2]$$

Points on the given ellipsoid all have $u = b$; and, all surfaces, $u =$ constant, are confocal ellipsoids (the analogy to the spherical case, when $E = 0$, should be evident).

3.02.1.3 Preliminary Definitions and Concepts

The gravitational potential, V, of the Earth is generated by its total mass density distribution. For applications on the Earth's surface it is useful to include the potential, ϕ, associated with the centrifugal acceleration due to Earth's rotation. The sum, $W = V + \phi$, is then known, in geodetic terminology, as the gravity potential, distinct from gravitational potential. It is further advantageous to define a relatively simple reference potential, or normal potential, that accounts for the bulk of the gravity potential

(Section 3.02.5.2). The normal gravity potential, U, is defined as a gravity potential associated with a best-fitting ellipsoid, the normal ellipsoid, which rotates with the Earth and is also a surface of constant potential in this field. The difference between the actual and the normal gravity potentials is known as the disturbing potential: $T = W - U$; it thus excludes the centrifugal potential. The normal gravity potential accounts for approximately 99.999 5% of the total potential.

The gradient of the potential is an acceleration, gravity or gravitational acceleration, depending on whether or not it includes the centrifugal acceleration. Normal gravity, γ, comprises 99.995% of the total gravity, g, although the difference in magnitudes, the gravity disturbance, δg, can be as large as several parts in 10^4. A special kind of difference, called the gravity anomaly, Δg, is defined as the difference between gravity at a point, P, and normal gravity at a corresponding point, Q, where $W_P = U_Q$, and P and Q are on the same perpendicular to the normal ellipsoid.

The surface of constant gravity potential, W_0, that closely approximates mean sea level is known as the geoid. If the constant normal gravity potential, U_0, on the normal ellipsoid is equal to the constant gravity potential of the geoid, then the gravity anomaly on the geoid is the difference between gravity on the geoid and normal gravity on the ellipsoid at respective points, P_0, Q_0, sharing the same perpendicular to the ellipsoid. The separation between the geoid and the ellipsoid is known as the geoid undulation, N, or also the geoid height (**Figure 3**). A simple Taylor expansion of the normal gravity potential along the ellipsoid perpendicular yields the following important formula:

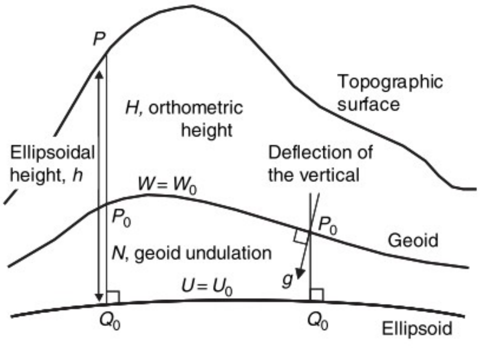

Figure 3 Relative geometry of geoid and ellipsoid.

$$N = \frac{T}{\gamma} \qquad [3]$$

This is Bruns' equation, which is accurate to a few millimeters in N, and which can be extended to $N = T/\gamma - (W_0 - U_0)/\gamma$ for the general case, $W_0 \neq U_0$. The gravity anomaly (on the geoid) is the gravity disturbance corrected for the evaluation of normal gravity on the ellipsoid instead of the geoid. This correction is $N\,\partial\gamma/\partial h = (\partial\gamma/\partial h)(T/\gamma)$, where h is height along the ellipsoid perpendicular. We have $\delta g = -\partial T/\partial h$, and hence

$$\Delta g = -\frac{\partial T}{\partial h} + \frac{1}{\gamma}\frac{\partial \gamma}{\partial h} T \qquad [4]$$

The slope of the geoid with respect to the ellipsoid is also the angle between the corresponding perpendiculars to these surfaces. This angle is known as the deflection of the vertical, that is, the deflection of the plumb line (perpendicular to the geoid) relative to the perpendicular to the normal ellipsoid. The deflection angle has components, ξ, η, respectively, in the north and east directions. The spherical approximations to the gravity disturbance, anomaly, and deflection of the vertical are given by

$$\begin{aligned} \delta g &= -\frac{\partial T}{\partial r}, & \Delta g &= -\frac{\partial T}{\partial r} - \frac{2}{r} T \\ \xi &= \frac{1}{\gamma}\frac{\partial T}{r \partial \theta}, & \eta &= -\frac{1}{\gamma}\frac{\partial T}{r \sin\theta\, \partial \lambda} \end{aligned} \qquad [5]$$

where the signs on the derivatives are a matter of convention.

3.02.2 Newton's Law of Gravitation

In its original form, Newton's law of gravitation applies only to idealized point masses. It describes the force of attraction, **F**, experienced by two such solitary masses as being proportional to the product of the masses, m_1 and m_2; inversely proportional to the distance, ℓ, between them; and directed along the line joining them:

$$\mathbf{F} = G\frac{m_1 m_2}{\ell^2}\mathbf{n} \qquad [6]$$

G is a constant, known as Newton's gravitational constant, that takes care of the units between the left- and right-hand sides of the equation; it can be

(Section 3.02.5.2). The normal gravity potential, U, is defined as a gravity potential associated with a best-fitting ellipsoid, the normal ellipsoid, which rotates with the Earth and is also a surface of constant potential in this field. The difference between the actual and the normal gravity potentials is known as the disturbing potential: $T = W - U$; it thus excludes the centrifugal potential. The normal gravity potential accounts for approximately 99.999 5% of the total potential.

The gradient of the potential is an acceleration, gravity or gravitational acceleration, depending on whether or not it includes the centrifugal acceleration. Normal gravity, γ, comprises 99.995% of the total gravity, g, although the difference in magnitudes, the gravity disturbance, δg, can be as large as several parts in 10^4. A special kind of difference, called the gravity anomaly, Δg, is defined as the difference between gravity at a point, P, and normal gravity at a corresponding point, Q, where $W_P = U_Q$, and P and Q are on the same perpendicular to the normal ellipsoid.

The surface of constant gravity potential, W_0, that closely approximates mean sea level is known as the geoid. If the constant normal gravity potential, U_0, on the normal ellipsoid is equal to the constant gravity potential of the geoid, then the gravity anomaly on the geoid is the difference between gravity on the geoid and normal gravity on the ellipsoid at respective points, P_0, Q_0, sharing the same perpendicular to the ellipsoid. The separation between the geoid and the ellipsoid is known as the geoid undulation, N, or also the geoid height (**Figure 3**). A simple Taylor expansion of the normal gravity potential along the ellipsoid perpendicular yields the following important formula:

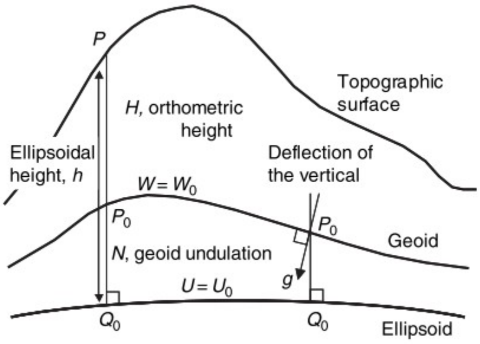

Figure 3 Relative geometry of geoid and ellipsoid.

$$N = \frac{T}{\gamma} \qquad [3]$$

This is Bruns' equation, which is accurate to a few millimeters in N, and which can be extended to $N = T/\gamma - (W_0 - U_0)/\gamma$ for the general case, $W_0 \neq U_0$. The gravity anomaly (on the geoid) is the gravity disturbance corrected for the evaluation of normal gravity on the ellipsoid instead of the geoid. This correction is $N\,\partial\gamma/\partial h = (\partial\gamma/\partial h)(T/\gamma)$, where h is height along the ellipsoid perpendicular. We have $\delta g = -\partial T/\partial h$, and hence

$$\Delta g = -\frac{\partial T}{\partial h} + \frac{1}{\gamma}\frac{\partial \gamma}{\partial h} T \qquad [4]$$

The slope of the geoid with respect to the ellipsoid is also the angle between the corresponding perpendiculars to these surfaces. This angle is known as the deflection of the vertical, that is, the deflection of the plumb line (perpendicular to the geoid) relative to the perpendicular to the normal ellipsoid. The deflection angle has components, ξ, η, respectively, in the north and east directions. The spherical approximations to the gravity disturbance, anomaly, and deflection of the vertical are given by

$$\begin{aligned} \delta g &= -\frac{\partial T}{\partial r}, \qquad & \Delta g &= -\frac{\partial T}{\partial r} - \frac{2}{r} T \\ \xi &= \frac{1}{\gamma}\frac{\partial T}{r \partial \theta}, \qquad & \eta &= -\frac{1}{\gamma}\frac{\partial T}{r \sin\theta\, \partial \lambda} \end{aligned} \qquad [5]$$

where the signs on the derivatives are a matter of convention.

3.02.2 Newton's Law of Gravitation

In its original form, Newton's law of gravitation applies only to idealized point masses. It describes the force of attraction, $\mathbf{F}$, experienced by two such solitary masses as being proportional to the product of the masses, m_1 and m_2; inversely proportional to the distance, ℓ, between them; and directed along the line joining them:

$$\mathbf{F} = G\frac{m_1 m_2}{\ell^2}\mathbf{n} \qquad [6]$$

G is a constant, known as Newton's gravitational constant, that takes care of the units between the left- and right-hand sides of the equation; it can be determined by experiment and the current best value is (Groten, 2004):

$$G = (6.67259 \pm 0.00030) \times 10^{-11}\ \mathrm{m}^3\,\mathrm{kg}^{-1}\,\mathrm{s}^{-2} \quad [7]$$

The unit vector $\mathbf{n}$ in eqn [6] is directed from either point mass to the other, and thus the gravitational force is attractive and applies equally to one mass as the other. Newton's law of gravitation is universal as far as we know, requiring reformulation only in Einstein's more comprehensive theory of general relativity which describes gravitation as a characteristic curvature of the space–time continuum (Newton's formulation assumes instantaneous action and differs significantly from the general relativistic concept only when very large velocities or masses are involved).

We can ascribe a gravitational acceleration to the gravitational force, which represents the acceleration that one mass undergoes due to the gravitational attraction of the other. Specifically, from the law of gravitation, we have (for point masses) the gravitational acceleration of m_1 due to the gravitational attraction of m_2:

$$\mathbf{g} = G\frac{m_2}{\ell^2}\mathbf{n} \quad [8]$$

The vector $\mathbf{g}$ is independent of the mass, m_1, of the body being accelerated (which Galileo found by experiment).

By the law of superposition, the gravitational force, or the gravitational acceleration, due to many point masses is the vector sum of the forces or accelerations generated by the individual point masses. Manipulating vectors in this way is certainly feasible, but fortunately a more appropriate concept of gravitation as a scalar field simplifies the treatment of arbitrary mass distributions.

This more modern view of gravitation (already adopted by Gauss (1777–1855) and Green (1793–1841)) holds that it is a field having a gravitational potential. Lagrange (1736–1813) fully developed the concept of a field, and the potential, V, of the gravitational field is defined in terms of the gravitational acceleration, $\mathbf{g}$, that a test particle would undergo in the field according to the equation

$$\nabla V = \mathbf{g} \quad [9]$$

where ∇ is the gradient operator (a vector). Further elucidation of gravitation as a field grew from Einstein's attempt to incorporate gravitation into his special theory of relativity where no reference frame has special significance above all others. It was necessary to consider that gravitational force is not a real force (i.e., it is not an applied force, like friction or propulsion) – rather, it is known as a kinematic force, that is, one whose action is proportional to the mass on which it acts (like the centrifugal force; see Martin, 1988). Under this precept, the geometry of space is defined intrinsically by the gravitational fields contained therein.

We continue with the classical Newtonian potential, but interpret gravitation as an acceleration different from the acceleration induced by real, applied forces. This becomes especially important when considering the measurement of gravitation (Section 3.02.6.1). The gravitational potential, V, is a 'scalar' function, and, as defined here, V is derived directly on the basis of Newton's law of gravitation. To make it completely consistent with this law and thus declare it a Newtonian potential, we must impose the following conditions:

$$\lim_{\ell\to\infty} \ell V = Gm \quad \text{and} \quad \lim_{\ell\to\infty} V = 0 \quad [10]$$

Here, m is the attracting mass, and we say that the potential is regular at infinity. It is easy to show that the gravitational potential at any point in space due to a point mass, in order to satisfy eqns [8]–[10], must be

$$V = \frac{Gm}{\ell} \quad [11]$$

where, again, ℓ is the distance between the mass and the point at which the potential is expressed. Note that V for $\ell = 0$ does not exist in this case; that is, the field of a point mass has a singularity. We use here the convention that the potential is always positive (in contrast to physics, where it is usually defined to be negative, conceptually closer to potential energy).

Applying the law of superposition, the gravitational potential of many point masses is the sum of the potentials of the individual points (see **Figure 4**):

$$V_P = G\sum_j \frac{m_j}{\ell_j} \quad [12]$$

And, for infinitely many points in a closed, bounded region with infinitesimally small masses, $\mathrm{d}m$, the summation in eqn [12] changes to an integration,

$$V_P = G\int_{\text{mass}} \frac{\mathrm{d}m}{\ell} \quad [13]$$

or, changing variables (i.e., units), $\mathrm{d}m = \rho\,\mathrm{d}v$, where ρ represents density (mass per volume) and $\mathrm{d}v$ is a volume element, we have (**Figure 5**)

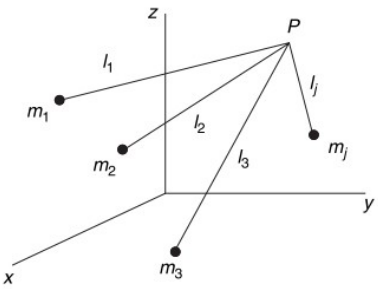

Figure 4 Discrete set of mass points (superposition principle).

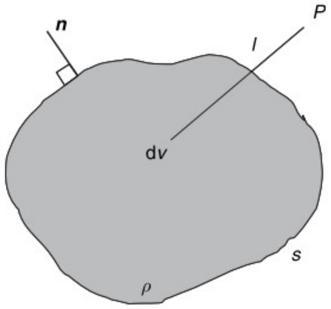

Figure 5 Continuous density distribution.

$$V_P = G\iiint_{\text{volume}} \frac{\rho}{\ell}\,\mathrm{d}v \qquad [14]$$

In eqn [14], ℓ is the distance between the evaluation point, P, and the point of integration. In spherical polar coordinates (Section 3.02.1.2), these points are (θ, λ, r) and (θ', λ', r'), respectively. The volume element in this case is given by $\mathrm{d}v = r'^2 \sin\theta'\,\mathrm{d}\lambda'\,\mathrm{d}\theta'\,\mathrm{d}r'$. V and its first derivatives are continuous everywhere – even in the case that P is on the bounding surface or inside the mass distribution, where there is the apparent singularity at $\ell = 0$. In this case, by changing to a coordinate system whose origin is at P, the volume element becomes $\mathrm{d}v = \ell^2 \sin\psi\,\mathrm{d}\alpha\,\mathrm{d}\psi\,\mathrm{d}\ell$ (for some different colatitude and longitude ψ and α); and, clearly, the singularity disappears – the integral is said to be weakly singular.

Suppose the density distribution over the volume depends only on radial distance (from the center of mass): $\rho = \rho(r')$, and that P is an exterior evaluation point. The surface bounding the masses necessarily is a sphere (say, of radius, R) and because of the spherically symmetric density we may choose the integration coordinate system so that the polar axis passes through P. Then

$$\ell = \sqrt{r'^2 + r^2 - 2r'r\cos\theta'}, \qquad \mathrm{d}\ell = \frac{1}{\ell} r' r \sin\theta'\,\mathrm{d}\theta' \qquad [15]$$

It is easy to show that with this change of variables (from θ to ℓ) the integral [14] becomes simply

$$V(\theta, \lambda, r) = \frac{GM}{r}, \qquad r \geq R \qquad [16]$$

where M is the total mass bounded by the sphere. This shows that to a very good approximation the external gravitational potential of a planet such as the Earth (with concentrically layered density) is the same as that of a point mass.

Besides volumetric mass (density) distributions, it is of interest to consider surface distributions. Imagine an infinitesimally thin layer of mass on a surface, s, where the units of density in this case are those of mass per area. Then, analogous to eqn [14], the potential is

$$V_P = G\iint_s \frac{\rho}{\ell}\,\mathrm{d}s \qquad [17]$$

In this case, V is a continuous function everywhere, but its first derivatives are discontinuous at the surface. Or, one can imagine two infinitesimally close density layers (double layer, or layer of mass dipoles), where the units of density are now those of mass per area times length. It turns out that the potential in this case is given by (see Heiskanen and Moritz, 1967, p. 8)

$$V_P = G\iint_s \rho \frac{\partial}{\partial n}\left(\frac{1}{\ell}\right)\mathrm{d}s \qquad [18]$$

where $\partial/\partial n$ is the directional derivative along the perpendicular to the surface (**Figure 5**). Now, V itself is discontinuous at the surface, as are all its derivatives. In all cases, V is a Newtonian potential, being derived from the basic formula [11] for a point mass that follows from Newton's law of gravitation (eqn [6]).

The following properties of the gravitational potential are useful for subsequent expositions. First, consider Stokes's theorem, for a vector function, **f**, defined on a surface, **s**:

$$\iint_s (\nabla \times \mathbf{f}) \cdot \mathbf{n}\ \mathrm{d}s = \oint_p \mathbf{f} \cdot \mathrm{d}\mathbf{r} \qquad [19]$$

where p is any closed path in the surface, $\mathbf{n}$ is the unit vector perpendicular to the surface, and $\mathrm{d}\mathbf{r}$ is a differential displacement along the path. From eqn [9], we find

$$\nabla \times \mathbf{g} = \mathbf{0} \quad [20]$$

since $\nabla \times \nabla = 0$; hence, applying Stokes's theorem, we find with $F = mg$ that

$$w = \oint \mathbf{F} \cdot \mathbf{ds} = 0 \quad [21]$$

That is, the gravitational field is conservative: the work, w, expended in moving a mass around a closed path in this field vanishes. In contrast, dissipating forces (real forces!), like friction, expend work or energy, which shows again the special nature of the gravitational force.

It can be shown (Kellogg, 1953, p. 156) that the second partial derivatives of a Newtonian potential, V, satisfy the following differential equation, known as Poisson's equation:

$$\nabla^2 V = -4\pi G\rho \quad [22]$$

where $\nabla^2 = \boldsymbol{\nabla} \cdot \boldsymbol{\nabla}$ formally is the scalar product of two gradient operators and is called the Laplacian operator. In Cartesian coordinates, it is given by

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \quad [23]$$

Note that the Laplacian is a scalar operator. Eqn [22] is a local characterization of the potential field, as opposed to the global characterization given by eqn [14]. Poisson's equation holds wherever the mass density, ρ, satisfies certain conditions similar to continuity (Hölder conditions; see Kellogg, 1953, pp. 152–153). A special case of eqn [22] applies for those points where the density vanishes (i.e., in free space); then Poisson's equation turns into Laplace' equation,

$$\nabla^2 V = 0 \quad [24]$$

It is easily verified that the point mass potential satisfies eqn [24], that is,

$$\nabla^2 \left(\frac{1}{\ell}\right) = 0 \quad [25]$$

where $\ell = \sqrt{(x-x')^2+(y-y')^2+(z-z')^2}$ and the mass point is at (x', y', z').

The solutions to Laplace' equation [24] (that is, functions that satisfy Laplace' equations) are known as harmonic functions (here, we also impose the conditions [10] on the solution, if it is a Newtonian potential and if the mass-free region includes infinity). Hence, every Newtonian potential is a harmonic function in free space. The converse is also true: every harmonic function can be represented as a Newtonian potential of a mass distribution (Section 3.02.3.1).

Whether as a volume or a layer density distribution, the corresponding potential is the sum or integral of the source value multiplied by the inverse distance function (or its normal derivative for the dipole layer). This function depends on both the source points and the computation point and is known as a Green's function. It is also known as the 'kernel' function of the integral representation of the potential. Functions of this type also play a dominant role in representing the potential as solutions to boundary-value problems (BVPs), as shown in subsequent sections.

3.02.3 Boundary-Value Problems

If the density distribution of the Earth's interior and the boundary of the volume were known, then the problem of determining the Earth's gravitational potential field is solved by the volume integral of eqn [14]. In reality, of course, we do not have access to this information, at least not the density, with sufficient detail. (The Preliminary Reference Earth Model (PREM), of Dziewonsky and Anderson (1981), still in use today by geophysicists, represents a good profile of Earth's radial density, but does not attempt to model in detail the lateral density heterogeneities.) In this section, we see how the problem of determining the exterior gravitational potential can be solved in terms of surface integrals, thus making exclusive use of accessible measurements on the surface.

3.02.3.1 Green's Identities

Formally, eqn [24] represents a partial differential equation for V. Solving this equation is the essence of the determination of the Earth's external gravitational potential through potential theory. Like any differential equation, a complete solution is obtained only with the application of boundary conditions, that is, imposing values on the solution that it must assume at a boundary of the region in which it is valid. In our case, the boundary is the Earth's surface and the exterior space is where eqn [24] holds (the atmosphere and other celestial bodies are neglected for the moment). In order to study the solutions to these BVPs (to show that solutions exist and are unique), we take advantage of some very important theorems and identities. It is noted that only a rather elementary introduction to BVPs is offered here with no attempt to address the much larger field of solutions to partial differential equations.

The first, seminal result is Gauss' divergence theorem (analogous to Stokes' theorem, eqn [19]),

$$\iiint_v \nabla\cdot\mathbf{f}\ \mathrm{d}v = \iint_s f_n\,\mathrm{d}s \qquad [26]$$

where **f** is an arbitrary (differentiable) vector function and $f_n = \mathbf{n}\cdot\mathbf{f}$ is the component of **f** along the outward unit normal vector, **n** (see **Figure 5**). The surface, s, encloses the volume, v. Equation [26] says that the sum of how much **f** changes throughout the volume, that is, the net effect, ultimately, is equivalent to the sum of its values projected orthogonally with respect to the surface. Conceptually, a volume integral thus can be replaced by a surface integral, which is important since the gravitational potential is due to a volume density distribution that we do not know, but we do have access to gravitational quantities on a surface by way of measurements.

Equation [26] applies to general vector functions that have continuous first derivatives. In particular, let U and V be two scalar functions having continuous second derivatives, and consider the vector function $\mathbf{f} = U\nabla V$. Then, since $\mathrm{n}\cdot\nabla = \partial/\partial n$, and

$$\nabla\cdot(U\nabla V) = \nabla U\cdot\nabla V + U\nabla^2 V \qquad [27]$$

we can apply Gauss' divergence theorem to get Green's first identity,

$$\iiint_v \left(\nabla U\cdot\nabla V + U\nabla^2 V\right)\mathrm{d}v = \iint_s U\frac{\partial V}{\partial n}\,\mathrm{d}s \qquad [28]$$

Interchanging the roles of U and V in eqn [28], one obtains a similar formula, which, when subtracted from eqn [28], yields Green's second identity,

$$\iiint_v \left(U\nabla^2 V - V\nabla^2 U\right)\mathrm{d}v = \iint_s \left(U\frac{\partial V}{\partial n} - V\frac{\partial U}{\partial n}\right)\mathrm{d}s \qquad [29]$$

This is valid for any U and V with continuous derivatives up to second order.

Now let $U = 1/\ell$, where ℓ is the usual distance between an integration point and an evaluation point. And, suppose that the volume, v, is the space exterior to the Earth (i.e., Gauss' divergence theorem applies to any volume, not just volumes containing a mass distribution). With reference to **Figure 6**, consider the evaluation point, P, to be inside the volume (free space) that is bounded by the surface, s; P is thus outside the Earth's surface. Let V be a solution to eqn [24], that is, it is the gravitational potential of the Earth. From the volume, v, exclude the volume bounded by a small sphere, σ, centered at P. This sphere becomes part of the surface that bounds the volume, v. Then, since U, by our definition, is a point mass potential, $\nabla^2 U = 0$ everywhere in v (which excludes the interior of the small sphere around P); and, the second identity [29] gives

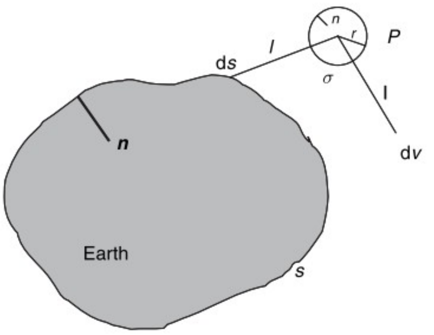

Figure 6 Geometry for special case of Green's third identity.

$$\iint_s \left(\frac{1}{\ell}\frac{\partial V}{\partial n} - V\frac{\partial}{\partial n}\left(\frac{1}{\ell}\right)\right)\mathrm{d}s + \iint_\sigma \left(\frac{1}{\ell}\frac{\partial V}{\partial n} - V\frac{\partial}{\partial n}\left(\frac{1}{\ell}\right)\right)\mathrm{d}\sigma = 0 \qquad [30]$$

The unit vector, **n**, represents the perpendicular pointing away from v. On the small sphere, **n** is opposite in direction to $\ell = r$, and the second integral becomes

$$\begin{aligned}\iint_\sigma \left(-\frac{1}{r}\frac{\partial V}{\partial r} + V\frac{\partial}{\partial r}\left(\frac{1}{r}\right)\right)\mathrm{d}\sigma &= -\iint_\Omega \frac{1}{r}\frac{\partial V}{\partial r}r^2\,\mathrm{d}\Omega \\ &\quad -\iint_\Omega V\,\mathrm{d}\Omega \\ &= -\iint_\Omega \frac{\partial V}{\partial r}r\,\mathrm{d}\Omega - 4\pi\bar{V}\end{aligned} \qquad [31]$$

where $\mathrm{d}\sigma = r^2\mathrm{d}\Omega$, Ω is the solid angle, 4π, and $\bar{V}$ is an average value of V on σ. Now, in the limit as the radius of the small sphere shrinks to zero, the right-hand side of eqn [31] approaches $0 - 4\pi V_P$. Hence, eqn [30] becomes (Kellogg, 1953, p. 219)

$$V_P = \frac{1}{4\pi}\iint_s \left(\frac{1}{\ell}\frac{\partial V}{\partial n} - V\frac{\partial}{\partial n}\left(\frac{1}{\ell}\right)\right)\mathrm{d}s \qquad [32]$$

with **n** pointing down (away from the space outside the Earth). This is a special case of Green's third identity. A change in sign of the right-hand side transforms **n** to a normal unit vector pointing into v, away from the masses, which conforms more to an Earth-centered coordinate system.

The right-hand side of eqn [32] is the sum of single- and double-layer potentials and thus shows that every harmonic function (i.e., a function that satisfies Laplace' equation) can be written as a Newtonian potential. Equation [32] is also a solution to a BVP; in this case, the boundary values are independent values of V and of its normal derivative, both on s (Cauchy problem). Below and in Section 3.02.4.1, we encounter another BVP in which the potential and its normal derivative are given in a specified linear combination on s. Using a similar procedure and with some extra care, it can be shown (see also Courant and Hilbert, 1962, vol. II, p. 256 (footnote)) that if P is on the surface, then

$$V_P = \frac{1}{2\pi}\iint_s \left(\frac{1}{\ell}\frac{\partial V}{\partial n} - V\frac{\partial}{\partial n}\left(\frac{1}{\ell}\right)\right) \mathrm{d}s \qquad [33]$$

where n points into the masses. Comparing this to eqn [32], we see that V is discontinuous as one approaches the surface; this is due to the double-layer part (see eqn [18]).

Equation [32] demonstrates that a solution to a particular BVP exists. Specifically, we are able to measure the potential (up to a constant) and its derivatives (the gravitational acceleration) on the surface and thus have a formula to compute the potential anywhere in exterior space, provided we also know the surface, s. Other BVPs also have solutions under appropriate conditions; a discussion of existence theorems is beyond the present scope and may be found in (Kellogg, 1953). Equation [33] has deep geodetic significance. One objective of geodesy is to determine the shape of the Earth's surface. If we have measurements of gravitational quantities on the Earth's surface, then conceptually we are able to determine its shape from eqn [33], where it would be the only unknown quantity. This is the basis behind the work of Molodensky *et al.* (1962), to which we return briefly at the end of this section.

3.02.3.2 Uniqueness Theorems

Often the existence of a solution is proved simply by finding one (as illustrated above). Whether such as solution is the only one depends on a corresponding uniqueness theorem. That is, we wish to know if a certain set of boundary values will yield just one potential in space. Before considering such theorems, we classify the BVPs that are typically encountered when determining the exterior potential from measurements on a boundary. In all cases, it is an exterior BVP; that is, the gravitational potential, V, is harmonic ($\nabla^2 V = 0$) in the space exterior to a closed surface that contains all the masses. The exterior space thus contains infinity. Interior BVPs can be constructed, as well, but are not applicable to our objectives.

- *Dirichlet problem* (*or, BVP of the first kind*). Solve for V in the exterior space, given its values everywhere on the boundary.
- *Neumann problem* (*or, BVP of the second kind*). Solve for V in the exterior space, given values of its normal derivative everywhere on the boundary.
- *Robin problem* (*mixed BVP, or BVP of the third kind*). Solve for V in the exterior space, given a linear combination of it and its normal derivative on the boundary.

Using Green's identities, we prove the following theorems for these exterior problems; similar results hold for the interior problems.

Theorem 1. If V is harmonic (hence continuously differentiable) in a closed region, v, and if V vanishes everywhere on the boundary, s, then V also vanishes everywhere in the region, v.

Proof. Since $V = 0$ on s, Green's first identity (eqn [28]) with $U = V$ gives

$$\iiint_v (\nabla V)^2 \,\mathrm{d}v = \iint_s V\frac{\partial V}{\partial n}\,\mathrm{d}s = 0 \qquad [34]$$

The integral on the left side is therefore always zero, and the integrand is always non-negative. Hence, $\boldsymbol{\nabla} V = \mathbf{0}$ everywhere in v, which implies that $V = \text{constant}$ in v. Since V is continuous in v and $V = 0$ on s, that constant must be zero; and so $V = 0$ in v.

This theorem solves the Dirichlet problem for the trivial case of zero boundary values and it enables the following uniqueness theorem for the general Dirichlet problem.

Theorem 2 (*Stokes' theorem*). If V is harmonic (hence continuously differentiable) in a closed region, v, then V is uniquely determined in v by its values on the boundary, s.

Proof. Suppose the determination is not unique: that is, suppose there are V_1 and V_2, both harmonic

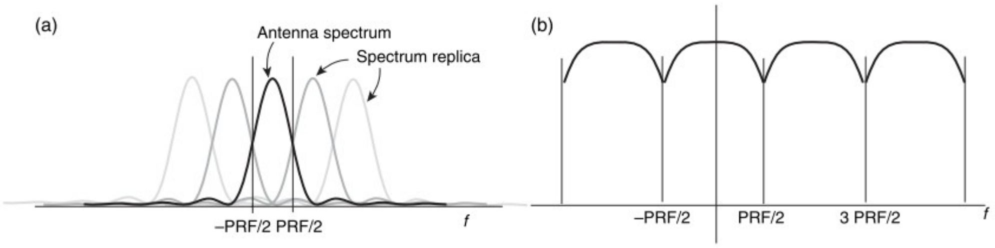

Figure 38 The azimuth signal is sampled at the PRF. The resulting discrete time signal in azimuth has a circular spectrum with all energy at frequency higher than the PRF aliased into the circular bandwidth of the PRF. (a) Illustrates the replicas of the continuous spectrum that are aliased, and (b) the resulting circular spectrum.

azimuth signals are inherently derived from pulses acquired along track, the azimuth signal is a discrete time signal sampled at the PRF. The actual discrete time spectrum is then a circular spectrum, with all energy at frequencies higher than the PRF aliased into the range of one PRF (**Figure 38**).

Thus without knowledge of the geometry of the spacecraft and the observation, it is often difficult to know the actual Doppler centroid, since one usually estimates it from the spectrum only in the range of the PRF. One can process the data with an incorrect Doppler centroid, off by some number of PRFs (also called 'ambiguities') for instance, and the imagery will look almost correct. Since the signal is sampled, the matched filter spectrum for the actual centroid will be the same as for the incorrect centroid as long as the centroid is wrong by an integer number of ambiguities. However, the range migration correction is in terms of absolute Doppler, so if the wrong ambiguity is used, the images will be poorly focused in range (because energy is spread in range for each scatterer) and in azimuth (because the compression uses incomplete information for each scatterer).

If the attitude of the platform is known, then the Doppler can be computed from

$$f_{\mathrm{D}} = -\frac{2\vec{v}\cdot\hat{l}}{\lambda} = -\frac{2v}{\lambda}\sin\theta\sin\phi_{\mathrm{sq}} \qquad [116]$$

where $\vec{v}$ is the velocity vector, $\hat{l}$ is the look vector from antenna to target, θ is the look angle, and ϕ_{sq} is the squint angle. If the attitude parameters are not known, then one has to estimate the ambiguity. This can be done by trial and error – examine the focus for a variety of integer ambiguity choices – or by one of several automated techniques. A popular method of ambiguity estimation is to split the azimuth spectrum into two pieces and process each with a particular Doppler ambiguity assumption. If the correction is the right one, both side bands will be adjusted to the correct range and the two images produced will be properly registered in range. If the correction is not right, there will be a residual shift between the two images in range. From the magnitude of this shift it is often possible to estimate the correct ambiguity. Unfortunately the imagery are lower resolution in azimuth and defocused in range, so estimating the shift is sometimes difficult.

Speckle

Radar images are naturally noisy because of the coherent interaction of the electro-magnetic wave with scatterers within a resolution cell, known as speckle. Speckle is an important part of the radar literature, as it is an important limitation of the imagery that many people have worked hard to minimize. It is also an important concept for interferometry, which exploits the spatially fixed properties of speckle to extract phase differences between images. The normalized backscatter cross-section σ_0 is an average quantity for a naturally varying random surface. For homogenous surfaces made up of a continuum of scatterers, the backscatter cross-section is a measure of the roughness of the surface and its natural reflectivity, as shown in a previous section. Viewing the SAR imaging process as a lens (**Figure 39**), after imaging, the random (complex) surface reflectivity is convolved with the sinc-function impulse response of the imaging system, accounting for the random differential path length introduced by surface roughness.

In a given resolution element, we consider the sum of the collection of small independent scatterers, where each scatterer has its own amplitude and phase

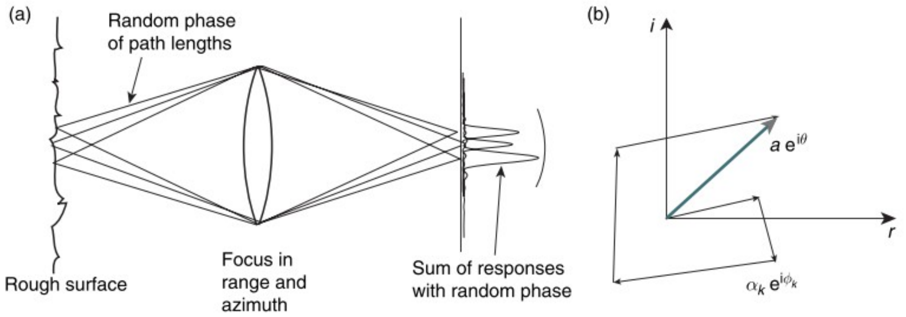

Figure 39 Illustration of the concept of radar imaging of a random surface (a) and the coherent sum of independent scatterers as a phasor (b).

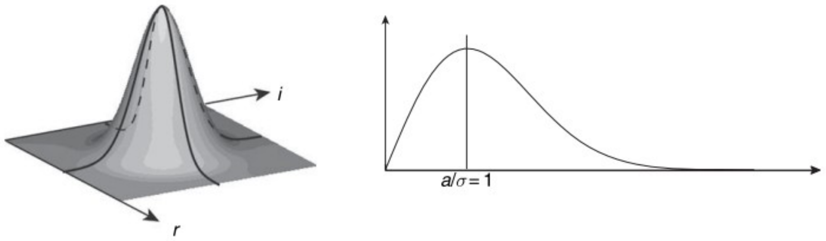

Figure 40 The real and imaginary parts of a SAR image are circularly Gaussian (*left*). The amplitude of the SAR pixel is Rayleigh distributed, with peak at a = σ.

$$a e^{j\theta} = \sum_{k=1}^{N} \alpha_k e^{j\phi_k} \qquad [117]$$

If α_k and ϕ_k are statistically independent of each other and α_k are identically distributed with mean $\bar{\alpha}$ and second moment $\overline{\alpha^2}$, and ϕ_k are uniformly distributed in the interval $[-\pi, \pi]$, then one can show that for large N

$$r = a\cos\theta \qquad [118]$$

$$i = a\sin\theta \qquad [119]$$

are Gaussian-distributed zero-mean random variables with joint probability density function

$$p_{\mathrm{RI}}(r, i) = \frac{1}{2\pi\sigma^2} e^{-(r^2+i^2)/2\sigma^2} \qquad [120]$$

where $\sigma^2 = \overline{\alpha^2}/2$ (**Figure 40**). The variance of the real and imaginary parts of the SAR image are related to the mean power of the intrinsic scatterers but are zero mean random variables themselves. As such for homogeneous areas, averaging many complex pixels together in a region will eventually reduce the signal to zero on average and not improve the quality of the image. This is a common mistake among beginners. It is the amplitude of the image, a, or its intensity (power), $I = a^2$, that interests us.

For the amplitude $a = \sqrt{r^2 + i^2}$, we can use the rules of transformation of random variables (multiply the distribution by the Jacobian of the transformation and substitute in the new variables) to get the distribution

$$p_A(a) = \frac{a}{\sigma^2} e^{-a^2/2\sigma^2}, \quad a \geq 0 \qquad [121]$$

which is known as the Rayleigh distribution (**Figure 40**). The phase is uniformly distributed over the interval $(-\pi, \pi)$

$$p_\Theta(\theta) = \frac{1}{2\pi}, \quad -\pi < \theta < \pi \qquad [122]$$

The mean and variance of the amplitude are

$$\bar{a} = \sqrt{\frac{\pi}{2}}\sigma \qquad [123]$$

$$\sigma_a^2 = \left(2 - \frac{\pi}{2}\right)\sigma^2 \qquad [124]$$

Clearly the mean amplitude of the image scales with the variability in the image, which is related to the mean backscattered power $\overline{\alpha^2}$ of the scatterers comprising the resolution element. The brighter the image, the noisier the image, and vice-versa. Thus, at some signal level such that the signal level returned is sufficiently larger than the thermal noise in an image, having a more powerful radar does not necessarily improve image quality.

The intensity, or power, of the image $I = a^2$ has an exponential distribution

$$p_I(I) = \frac{1}{\bar{I}} e^{-I/\bar{I}}, \quad I \geq 0 \qquad [125]$$

with mean and variance

$$\bar{I} = 2\sigma^2, \quad \sigma_I^2 = 4\sigma^4 \qquad [126]$$

The mean power is equal to its standard deviation, as expected from the mean amplitude scaling with σ. In the display and analysis of an image with only 'one look,' meaning processed at full resolution using the full synthetic aperture, the SNR of the intensity image is just I since $\bar{I}^2/\sigma_j^2 = 1_j$ the brighter the image, the noisier. In the design of a radar system, one must have sufficient power to receive a return from the surface above the thermal noise (satisfying the radar equation, where $\overline{\alpha^2}$ is related to the backscattering cross-section), but not so high that we are wasting power. For interferometry, we will soon see however that more power in general is better!

The radiometric resolution is defined as the ability to discriminate surfaces of different brightnesses. It has been defined to be

$$Q = 10 \log\left(\frac{\bar{I} + \sigma_I}{\bar{I} - \sigma_I}\right) \qquad [127]$$

which is something like the distance apart in amplitude of discriminable amplitudes. For a single-look image, that distance is infinite. In order to improve this situation, it is often necessary to average many pixels together. The act of averaging N independent (Gaussian, but also approximately true for non-Gaussian) random variables with identical distributions reduces the variance by a factor of N. Thus for

$$I' = \frac{1}{N} \sum_{k=1}^{N} I_k \qquad [128]$$

we have $\bar{I}' = \bar{I}$, $\sigma_{I'}^2 = \sigma_I^2/N$, and the radiometric resolution becomes

$$Q = 10 \log\left(\frac{\bar{I}' + \sigma_{I'}}{\bar{I}' - \sigma_{I'}}\right) = 10 \log\left(\frac{\sqrt{N} + 1}{\sqrt{N} - 1}\right) \qquad [129]$$

The resolution distance approaches 0 for arbitrarily large N. This averaging technique is often called 'taking looks' because it can be shown to be equivalent to segmenting the azimuth spectrum into separate frequency bands, forming a power image from each band, then stacking the images in an average. Since azimuth band corresponds to a particular look direction relative to a target on the ground, this technique is called taking looks.

Of course averaging together a large number of pixels will lower the intrinsic resolution of the image, so designing the system to have a larger bandwidth is desirable if radiometric resolution is important. Note also that there is a limit to how far one can go to improve the statistics. The improvement above assumes that all the pixels come from the same distribution. If the scene is variable in intrinsic brightness, not all the benefits will be realized.

If there is a bright target within a resolution cell, then there is a mean value associated with the Gaussian random variables (**Figure 41**). This complex random variable has a Rician distribution given by

$$p_A(a) = \frac{a}{\sigma^2} e^{-(a^2+s^2)/2\sigma^2} I_0\left(\frac{as}{\sigma^2}\right), \quad a \geq 0 \qquad [130]$$

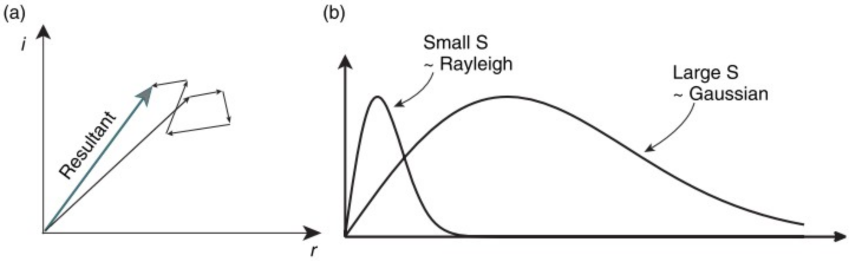

Figure 41 (a) A phasor diagram with a bright specular target (within a scattering cell. (b) The Rician distribution of the amplitude.

where I_0 is the modified Bessel function of the first kind and s is the specular target strength. This distribution approaches a Gaussian distribution for large s relative to σ. The distribution of the phase, which is uniform for $s=0$, becomes more peaked as s grows, eventually approximating a Gaussian also:

$$p_{\Theta}(\theta) \rightarrow \frac{s/\sigma}{\sqrt{2\pi}} \mathrm{e}^{-s^2\theta^2/2\sigma^2} \qquad [131]$$

Often specular targets are used for calibration of the radar. If the specular target is not sufficiently bright, then the amplitude and phase will be corrupted by the background random scatterers.

Another way to model the surface is in terms of a general reflectivity function that is statistical in nature. Since

$$\Gamma_{\mathrm{zcc}}(x', r') = \iint \mathrm{d}r\, \mathrm{d}x \Gamma(x, r) \times \operatorname{sinc}\Big(\frac{\pi}{\Delta X}(x'-x)\Big) \operatorname{sinc}\Big(\frac{\pi}{\Delta R}(r'-r)\Big) \qquad [132]$$

if Γ is a statistical function, the power that is meaningful is the expected power scattered back from the resolution cell, which is related by scale factors and geometric normalizations to $E\{\Gamma\Gamma^*\}$. Writing this out explicitly, we have

$$E\{\Gamma_{\mathrm{zcc}}\Gamma^*_{\mathrm{zcc}}\}(x', r') = \int \mathrm{d}x_1 \int \mathrm{d}x_2 \int \mathrm{d}r_1 \int \mathrm{d}r_2 \times E\{\Gamma(x_1, r_1)\Gamma^*(x_2, r_2)\} \qquad [133]$$

$$= \operatorname{sinc}\Big(\frac{\pi}{\Delta X}(x'-x_1)\Big) \operatorname{sinc}\Big(\frac{\pi}{\Delta X}(x'-x_2)\Big) \qquad [134]$$

$$= \operatorname{sinc}\Big(\frac{\pi}{\Delta R}(r'-r_1)\Big) \operatorname{sinc}\Big(\frac{\pi}{\Delta R}(r'-r_2)\Big) \qquad [135]$$

If the scatterers that comprise the reflectivity are uncorrelated, then the expected value of the product of the surface with itself when not coaligned would be zero:

$$E\{\Gamma(x_1, r_1)\Gamma^*(x_2, r_2)\} = E\{\Gamma(x_1, r_1)\}E\{\Gamma^*(x_2, r_2)\} = 0 \qquad [136]$$

However, when the surface is self-aligned

$$E\{\Gamma(x_1, r_1)\Gamma^*(x_2, r_2)\} = |\Gamma(x_1, r_1)|^2 \qquad [137]$$

More succinctly,

$$E\{\Gamma(x_1, r_1)\Gamma^*(x_2, r_2)\} = E\{|\Gamma(x_1, r_1)|^2\} \times \delta(x_1-x_2)\delta(r_1-r_2) \qquad [138]$$

Using this expression in eqn [135] simplifies the integrals

$$E\{\Gamma_{\mathrm{zcc}}\Gamma^*_{\mathrm{zcc}}\}(x', r') = \int \mathrm{d}x \int \mathrm{d}r\, E\{|\Gamma(x, r)|^2\} \qquad [139]$$

$$\operatorname{sinc}^2\Big(\frac{\pi}{\Delta X}(x'-x)\Big) \operatorname{sinc}^2\Big(\frac{\pi}{\Delta R}(r'-r)\Big) \qquad [140]$$

$$= E\{|\Gamma|^2\}\Delta X \Delta R \qquad [141]$$

For homogeneous surfaces, we can expect the mean power to be essentially constant over the resolution cell, so we can take the expectation outside the integration. Also the variables x' and r' can profitably be set to zero since there is no dependence on space. Then the integration over the sinc functions produces a constant, which is proportional to the area under them. This turns out to be ΔX and ΔR. So we see that $E\{\Gamma_{\mathrm{zcc}}\Gamma^*_{\mathrm{zcc}}\}$ is related to the normalized backscattering cross-section σ_0 through the function Γ.

Doppler and the Interferometric Baseline

The precise definition of interferometric baseline and phase, and consequently the topographic mapping process, depends on how the SAR data comprising the interferometer are processed. Consequently, a brief overview of the salient aspects of SAR processing is in order.

Processed data from SAR systems are sampled images. Each sample, or pixel, represents some aspect of the physical process of radar backscatter. A resolution element of the imagery is defined by the spectral content of the SAR system. Fine resolution in the range direction is achieved typically by transmitting pulses of either short time duration with high peak power, or of a longer time duration with a wide, coded signal bandwidth at lower peak transmit power. Resolution in range is inversely proportional to this bandwidth. In both cases, the received echo for each pulse is sampled at the required radar signal bandwidth.

For ultra-narrow pulsing schemes, the pulse width is chosen at the desired range resolution, and no further data manipulation is required. For coded pulses, the received echoes are typically processed with a matched filter technique to achieve the desired range resolution. Most spaceborne platforms use chirp-encoding to attain the desired bandwidth and consequent range resolution, where the frequency is linearly changed across the pulse.

($!" '
D -256- - -* ! - /$ " ! !
1 $ "KA

D -256- - -* ! - /$ " ! !
1 $ 1"

D -256- - -* ! - /$ " ! !
1 $ "KA

D -256- - -* ! - /$ " ! !
1 $ 1"

-256- - -* ! - /$: J 9 ! 9 !
7 1 3 % 7 " 3 % = A $ 1"
*D ? A " UUU
(% % < %B

$$2\theta_{\mathrm{b.s.}} = X_{\mathrm{ill}}/R = \lambda R/(LR) = \lambda/L \qquad [77]$$

as expected. Thus, the Doppler bandwidth is $f_{\mathrm{D,ill}}$ and by this derivation is also $2v(2\sin\theta_{\mathrm{b.s.}})/\lambda = 2v/L$.

At any given instant, the received signal is composed of the full complement of Doppler frequencies because targets are distributed throughout the beam, each with their own Doppler frequency. In forming the synthetic aperture at some reference time, we are rearranging each of the targets in the range-Doppler image to lie at the Doppler frequency appropriate to that geometry at that reference time. This is not immediately obvious from the signal-theoretical approach to azimuth filtering that was just derived. To be more explicit, we pick a collection of pulses that will be used to form our synthetic aperture, noting that this establishes a position in space at a time instant as a reference for defining Doppler frequencies. We can also model the ground as a collection of point targets arranged in azimuth, such that

$$r_{\mathrm{zc\delta,tot}} = \sum_{k=1}^{N} A_k \mathrm{e}^{-\mathrm{j}\pi(x-x_k)^2/F^2} \operatorname{sinc}\frac{\pi}{\Delta R}\left(R - R_0 - (x-x_k)^2/2R_0\right) \qquad [78]$$

For the sake of this discussion, we allow the sinc function to become a delta function in range and ignore its variability with x, which is small and correctable. Thus,

$$r_{\mathrm{zc\delta}} = \sum_{k=1}^{N} A_k \mathrm{e}^{-\mathrm{j}\pi(x-x_k)^2/F^2} \qquad [79]$$

The azimuth compression process can be written then as

$$\begin{aligned} r_{\mathrm{zcc\delta}}(x) &= \int_{x-X_{\mathrm{ill}}/2}^{x+X_{\mathrm{ill}}/2} r_{\mathrm{zc\delta}}(x')\mathrm{e}^{\mathrm{j}\pi(x'-x)^2/F^2}\,\mathrm{d}x' \\ &= \int_{x-X_{\mathrm{ill}}/2}^{x+X_{\mathrm{ill}}/2} \sum_{k=1}^{N} A_k \mathrm{e}^{-\mathrm{j}\pi(x'-x_k)^2/F^2} \mathrm{e}^{\mathrm{j}\pi(x'-x)^2/F^2}\,\mathrm{d}x' \\ &= \mathrm{e}^{\mathrm{j}\pi x^2/F^2} \int_{x-X_{\mathrm{ill}}/2}^{x+X_{\mathrm{ill}}/2} \sum_{k=1}^{N} A_k \mathrm{e}^{-\mathrm{j}\pi x_k^2/F^2} \mathrm{e}^{\mathrm{j}2\pi x' x_k/F^2} \mathrm{e}^{-\mathrm{j}2\pi x' x/F^2}\,\mathrm{d}x' \\ &= \sum_{k=1}^{N} \mathrm{e}^{\mathrm{j}\pi x^2/F^2} \int_{x-X_{\mathrm{ill}}/2}^{x+X_{\mathrm{ill}}/2} A'_k \mathrm{e}^{-\mathrm{j}2\pi x'(f_{\mathrm{Dx}} - f_{k,\mathrm{Dx}})}\,\mathrm{d}x' \end{aligned} \qquad [80]$$

where $f_{\mathrm{Dx}}(x) = x/F^2$ is the Doppler spatial frequency. From this arrangement of terms, we can recognize that the compressed azimuth signal is just the sum of Fourier transforms of the individual complex exponentials of complex amplitude A'_k. Because the argument of the exponential is the Doppler frequency, we see the transform domain is the Doppler domain, and each exponential (being representative of a scatterer at a particular azimuth location x_k relative to the reference location) transforms to an individual Doppler frequency $f_{k,\mathrm{Dx}}$. Thus, for a general continuous surface, the summation becomes an integration, and we see the compression process consists of focusing the array by applying a quadratic phase correction, followed by a transform that sorts the scatterers into correct Doppler bins for that time reference.

The reason we represent (and in fact execute) the imaging as a convolution in spatial variables is because the processed data the sample spacing independent of range on input and output, which is often desirable. In the focusing Fourier transform approach, the Doppler sample spacing is controlled by the implementation of Fourier transform, which is usually accomplished by fast Fourier transform (FFT). In this case, if the input data are sampled at a spacing $L/2$, the Doppler spacing will be $2/L/N$, where N is the number of points in the FFT.

A.1.5 Other Doppler considerations

If the antenna pattern was isotropic and we have a continuous time system, the phase and Doppler of a single point target would as depicted in Figure 35. Imposing an antenna pattern where the antenna pattern is pointing 'broadside,' that is, orthogonal to the velocity vector, the phase and Doppler history would be truncated according to the extent of the beam (see same figure). It is most commonly the case that the finite extent beam is 'squinted' forward or back with respect to the velocity vector of the sensor. In an aircraft, this is most commonly due to cross winds causing the fuselage to crab. In spacecraft, there is a natural latitude dependent angular offset between the velocity vector of the inertial orbit and the rotating Earth below causing a natural squint (Figure 36). The effect of this squint on the phase and Doppler history of a target is shown in Figure 37. Note the shape of the phase or Doppler does not change, just the range of phases or frequencies considered. Thus, there is a unique relationship between the pointing and the location of the scatterers. In processing data, the extent of the antenna beam (Doppler bandwidth) and the *centroid* of the Doppler, or the degree of squint as represented by the center of the Doppler bandwidth, are specified in one way or another. Figure 38 shows that the spectra are shifted to be centered on the Doppler spread associated with a particular

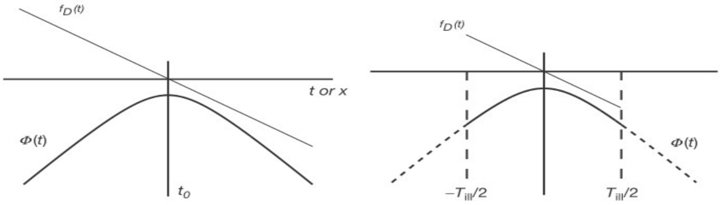

Figure 35 Phase and Doppler history over time for a point target. At left, the trace for a system with an isotropic antenna: the trace keeps going. At right, the finite illumination time of a finite dimension antenna truncates the phase and Doppler history.

squint. It is important to know the squint to ensure that the processing is centered on the proper part of the azimuth spectrum. The centroid specified in processing defines direction from which the ground is imaged after processing.

The squint can be derived from information provided by the spacecraft and radar manufacturer. Often, spacecrafts have sensitive position and attitude sensors that record the information needed to calculate the direction of the antenna boresight. When this information is not available, the Doppler centroid itself can be estimated from the data. This estimated centroid can be used directly in the processing. The centroid generally varies across range, so often the centroid is specified as a function of range.

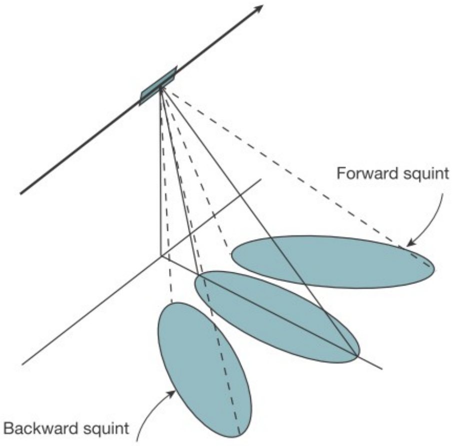

Figure 36 Illustration of forward and backward squint.

To understand the arrangement of spectral energy for a collection of scatterers, consider a collection of point target scatterers all at the same range, but at different azimuths:

$$\begin{aligned} r_{\mathrm{zc\delta,tot}} = {} & A_0 \mathrm{e}^{-\mathrm{j}\pi(x-x_0)^2/F^2} \operatorname{sinc}\frac{\pi}{\Delta R}\left(R-R_0-(x-x_0)^2/2R_0\right) \\ & + A_1 \mathrm{e}^{-\mathrm{j}\pi(x-x_1)^2/F^2} \operatorname{sinc}\frac{\pi}{\Delta R}\left(R-R_0-(x-x_1)^2/2R_0\right) \\ & + \cdots \\ & + A_n \mathrm{e}^{-\mathrm{j}\pi(x-x_n)^2/F^2} \operatorname{sinc}\frac{\pi}{\Delta R}\left(R-R_0-(x-x_n)^2/2R_0\right) \end{aligned} \quad [81]$$

If we ignore the antenna pattern modulation of target energy and denote the azimuth spectrum of an individual scatter response as

$$F(\omega) = \mathrm{F.T.}\left\{\mathrm{e}^{-\mathrm{j}\pi x^2/F^2} \operatorname{sinc}\frac{\pi}{\Delta R}\left(R-R_0-x^2/2R_0\right)\right\} \quad [82]$$

where F.T. denotes the Fourier transform, the spectrum of $r_{\mathrm{zc\delta,tot}}$ would be

$$R_{\mathrm{zc\delta,tot}}(\omega) = F(\omega)\left(A_0\mathrm{e}^{-\mathrm{j}\omega x_0} + A_1\mathrm{e}^{-\mathrm{j}\omega x_1} + \cdots + A_n\mathrm{e}^{-\mathrm{j}\omega x_n}\right) \quad [83]$$

Thus, except for a phase ramp in azimuth frequency space, all scatterers have the same spectrum. Clearly, the process of applying the matched filter neutralizes this common component, and the linear phase terms in frequency distinguish the

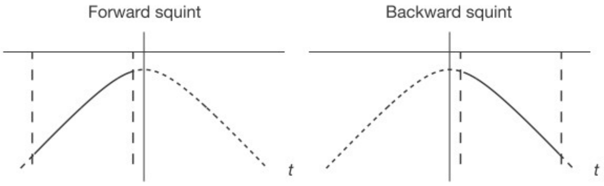

Figure 37 Phase history over time for a point target. At left, the trace for a forward squinted system. At right, the trace for a backward squinted system. Only phase is shown, but with reference to the previous figures, the Doppler history is similarly constrained. Forward (backward) squinted systems see more positive (negative) Doppler frequencies.

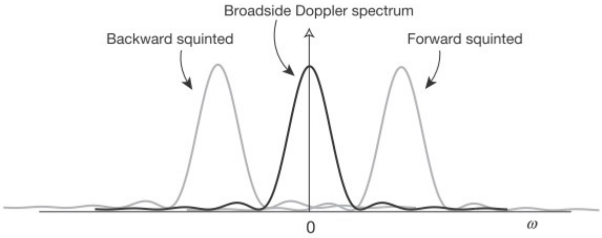

Figure 38 Hypothetical Doppler spectra of broadside, forward and backward squinted signals. For a collection of scatterers over the beam pointed with a particular squint, all scatterers experience the same Doppler frequency spread, so the spectra would be as shown, controlled in extent and magnitude by the antenna pattern in azimuth.

targets in position upon inverse transformation. This again shows the spatial-Doppler duality. The x dependence of the energy's localization in range $x^2/2R_0$ is called the range migration and is seen to be part of the function $F(\omega)$ in the frequency domain. Hence, any effects this may introduce in can be dealt with for all targets simultaneously by manipulating the spectrum.

We ignored the *range migration* term $(x-x_0)2/2R_0$ in the argument to the sinc function in range, but frequently this is not advisable. The processing of data on an orthogonal grid is enforced by the need for high-speed computations, and the most efficient processing methods are those that are separable in range and azimuth and for which FFT-based convolutions can be performed. For broadside imaging, the term can be ignored when $X_{\text{ill}}^2/2R_0$ is a fraction of the range resolution ΔR. In words, when the migration over the full extent of the azimuth matched filter (equivalent to the beam extent on the ground) is smaller than a range resolution element, the azimuth matched filter will catch all the energy for a given scatterer in a single range bin.

A.1.6 The sampled azimuth spectrum and range migration

Since the range migration correction terms depend on the Doppler frequency, one must know the correct Doppler to properly compress the imagery. This is also true for computing the correct azimuth matched filter; however, the sampled nature of the azimuth signal adds an interesting wrinkle to the processing that is often confusing. Since the azimuth signals in inherently derived from pulses acquired along-track, the azimuth signal is a discrete time signal sampled at the PRF. The actual discrete time spectrum is then a circular spectrum, with all energy at frequencies higher than the PRF aliased into the range of one PRF (Figure 39). Thus, without knowledge of the geometry of the spacecraft and the observation, it is often difficult to know the actual Doppler centroid, since one usually estimates it from the spectrum only in the range of the PRF. One can process the data with an incorrect Doppler centroid, off by some number of PRFs (also called *ambiguities*), for instance, and the imagery will look almost correct. Since the signal is sampled, the matched-filter spectrum for the actual centroid will be the same as for the incorrect centroid as long as the centroid is wrong by an integer number of ambiguities. However, the range migration correction is in terms of absolute Doppler, so if the wrong ambiguity is used, the images will be poorly focused in range (because energy is spread in range still for each scattered) and in azimuth (because the compression uses incomplete information for each scatterer).

If the attitude of the platform is known, then the Doppler can be computed from

$$f_{\text{D}}=-\frac{2\,\vec{v}\cdot\hat{l}}{\lambda}=-\frac{2v}{\lambda}\sin\theta\sin\phi_{\text{sq}} \qquad [84]$$

where $\vec{v}$ is the velocity vector, $\hat{l}$ is the look vector from antenna to target, θ is the look angle, and ϕ_{sq} is the squint angle. If the attitude parameters are not known, then one has to estimate the ambiguity. This can be done by trial and error – examine the focus for a variety of integer ambiguity choices – or by one of several automated techniques. A popular method of ambiguity estimation is to split the azimuth spectrum into two pieces and process each with a particular Doppler ambiguity assumption. If the correction is the right one, both side bands will be adjusted to the correct range and the two images produced will be properly registered in range. If the correction is not right, then there will be a residual shift between the two images in range. From the magnitude of this shift, it is often possible to estimate the correct ambiguity. Unfortunately, the imagery is lower resolution in azimuth and defocused in range, so estimating the shift is sometimes difficult.

A.2 The Derivation of the Interferogram from Two Images

As shown leading up to eqn [72], for a general complex scene $\Gamma(x, r)$, the SAR image after compression is given by convolution of r with the point target response

$$\Gamma_{\text{zcc}}(x', r') = \iint \Gamma(x, r)\, r_{\text{zcc}\delta}(x', r'; x, r)\,\mathrm{d}x\,\mathrm{d}r$$
$$= \int \mathrm{d}x \int \mathrm{d}r\,\Gamma(x, r)\mathrm{e}^{-\mathrm{j}4\pi r/\lambda}\cdot\operatorname{sinc}\left(\frac{\pi}{\Delta R}(r'-r)\right)\operatorname{sinc}\left(\frac{\pi}{\Delta X}(x'-x)\right) \qquad [85]$$

which can be verified by substituting $\Gamma(x, r)=\delta(x-x_0, r-R_0)$ to recover the impulse response $r_{\text{zcc}\delta}$.

The SAR image Γ_{zcc} is the convolution of the actual reflectivity with a two-dimensional function similar to a delta function but with finite width (the sinc function). This convolution smears out the intrinsic reflectivity as a point spread function in optics. If the system bandwidths were to become infinite, the sinc functions would become delta functions

$$\operatorname{sinc}\left(\frac{\pi}{\Delta R}(r'-r)\right) \rightarrow \delta(r'-r) \quad \text{as } \Delta R \rightarrow 0 \qquad [86]$$

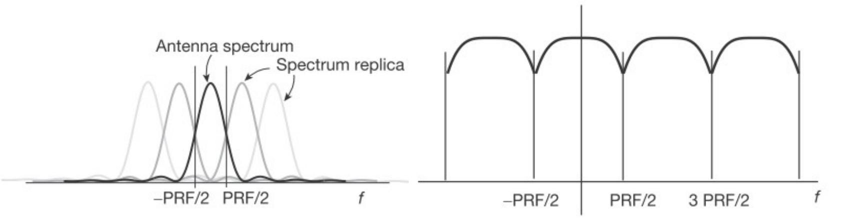

Figure 39 The azimuth signal is sampled at the PRF. The resulting discrete time signal in azimuth has a circular spectrum with all energy at frequency higher than the PRF aliased into the circular bandwidth of the PRF. Left illustrates the replicas of the continuous spectrum that are aliased, right the resulting circular spectrum.

(%$

Figure 6 - Geometry for special case of Green's third identity.

We can apply Gauss' divergence theorem to get Green's first identity, [28]

Pág. 440 (441) -

Gaussian-distributed zero-mean random variables with joint probability density function

$$p_{RI}(r,i) = \frac{1}{2\pi\sigma^2} e^{-(r^2+i^2)/2\sigma^2}$$ [120]

- Methods of Celestial Mechanics - Volume I. .(2004).pdf
- Methods of Celestial Mechanics - Volume II .. pdf

obs: Estes dois volumes existem em formato .djvu.

- 20211013 / Geodesy /

Herring, Thomas - Treatise on Geophysics - Volume 3 ..(2015).pdf

obs: Livro todo!!!

Pag. 377 (370) – Figure 36

Illustration of forward and backward squint. (estrabismo).

(important !!!).

Obs: Look [81] sinc $\frac{\pi}{\Delta R}$

– 20211013 [Lectures on Differential Topology / Lectures on Differential Topology.pdf

Pag. 249 (280)

Corollary 14.11

Let X be an oriented hypersurface of $\mathbb{R}^K$.

(1) If K is even, then $\chi(X) = 0$.

(2) If K is odd, then

$\chi(X) = 2\deg(g_X)$ where

g_X is the Gauss map of X.

/B
! -256- - -* ! - /B ? 4 ! 9 1 3 1"
('/ ! 3 '

The lemma follows by using two adapted gradient fields to compute the relative characteristics.

Note that $\chi(D^n) = 1$: use a Morse function on $(D^n, \emptyset, S^{n-1})$ with just one minimum.

If X is boundaryless and Y is with boundary, then the very same argument used when Y is boundaryless allows us to extend the multiplicative property.

Lemma 14.8.

$$\chi(X \times Y) = \chi(X)\chi(Y)\ .$$

In particular,

$$\chi(X \times D^n) = \chi(X)\ .$$

14.6.1. Additive property of χ. If (W, V_0, V_1), (W', V_0', V_1') are triads and $\phi : V_1 \to V_0'$ is a diffeomorphism, we get a new composite triad (W'', V_0, V_1'), where $W'' = W \amalg_\phi W'$. Any couple of relative fields v and v' with isolated zeros on the given two triads, respectively, can be glued together to produce a relative field v'' having as zeros the union of the zeros of v and v', each one keeping its index. Then we have

$$\chi(W'', V_0, V_1') = \chi(W, V_0, V_1) + \chi(W', V_0', V_1')\ .$$

This additive property of χ has remarkable consequences.

14.6.2. A baby TQFT. In Section 10.8, we have roughly outlined the axioms of a so-called TQFT and posed the question about the existence of any "nontrivial" one. We use χ to provide a first nontrivial example. Consider $\mathbf{CAT}_\eta(n+1)$. Associate to every object M the vector space $Z(M) = \mathbb{C}$. To every arrow f, carried by a triad (W, M_0, M_1), associate the unitary $\mathbb{C}$-linear map

$$Z(f) : Z(M_0) \to Z(M_1),\ \ z \to e^{i\chi(W, M_0)} z\ .$$

By the additive property of χ, it is easy to check that all axioms are satisfied. This shows, at least, that there are not logical contradictions within the given pattern of axioms.

14.7. E-P characteristic of tubular neighbourhoods and the Gauss map

The identity $\chi(X \times D^n) = \chi(X)$ is a special case of the following.

Proposition 14.9. *Let $p : U \to X$ be a closed tubular neighbourhood of a submanifold X of some Y. Then $\chi(U) = \chi(X)$.*

Proof. It is enough to show the equality for an ϵ-neighbourhood $N_\epsilon(X)$ of the zero section X of a vector bundle $\pi : E \to X$ endowed with a field of positive definite scalar products on every fibre. Let v be a tangent field on X with nondegenerate zeros. Define the field on $N_\epsilon(X)$,

$$w(z) = (z - p(z)) + v(p(z)) \ .$$

We check that w is a field on the triad $(N_\epsilon(X), \emptyset, \partial N_\epsilon(X))$ and that the zeros of w coincide with the zeros of v, are nondegenerate, and keep the sign. The proposition follows. ■

In the special case $X \subset \mathbb{R}^k$, assume that U has been constructed using the standard metric on $\mathbb{R}^k$. By removing from the interior of U a system of pairwise disjoint small open disks $\mathcal{D}_p$ around each zero of w, we get a manifold W with boundary $(\amalg_p S_p^{k-1}) \amalg \partial U$ on which the normalized field $\mathfrak{w} := w/||w||$ is well-defined, as well as the map $\mathfrak{w} : W \to S^{k-1}$. The restriction of $\mathfrak{w}$ to ∂U is a field of unitary vectors pointing out from U (in fact normal to the boundary). This restriction, say $g_{\partial U}$, is called the *Gauss map of the hypersurface* ∂U.

Corollary 14.10. *Let* $p : U \to X$ *be a tubular neighbourhood of* X *in* $\mathbb{R}^k$. *Then*

$$\chi(X) = \deg(g_{\partial U})$$

where $g_{\partial U}$ *is the Gauss map of the hypersurface* ∂U.

The corollary follows by computing the characteristic as the sum of zero indices and using the bordism invariance of the degree.

If X itself is an oriented hypersurface in $\mathbb{R}^k$, $k \geq 2$, we can define its Gauss map $g_X : X \to S^{k-1}$, where this field of oriented unitary normal vectors along X is determined by the fact that, for every $x \in X$, $g_X(x) \oplus T_xX = T_x\mathbb{R}^k$ as oriented vector spaces, where $\mathbb{R}^k$ is endowed with the standard orientation. In this case, ∂U consists of two parallel copies of X with opposite orientations. The last corollary specializes to the following:

Corollary 14.11. *Let* X *be an oriented hypersurface of* $\mathbb{R}^k$.

(1) *If* k *is even, then* $\chi(X) = 0$.

(2) *If* k *is odd, then*

$$\chi(X) = 2\deg(g_X)$$

where g_X *is the Gauss map of* X.

The corollary follows by using that the degree of the antipodal map $a_{k-1} : S^{k-1} \to S^{k-1}$ is equal to $(-1)^k$.

Remark 14.12. We can compute inductively the characteristic of real and complex projective spaces by decomposing $\mathbf{P}^n(K)$ as the union of a tubular neighbourhood of $\mathbf{P}^{n-1}(K) \subset \mathbf{P}^n(K)$ and its complement and applying Proposition 14.9 together with the additivity of χ.

14.8. Nontriviality of $\eta_\bullet$ and $\Omega_\bullet$

The integer E-P characteristic is *not* invariant up to bordism. For example, $[S^2] = [S^1 \times S^1] = 0 \in \eta_2$, but $\chi(S^2) = 2 \neq 0 = \chi(S^1 \times S^1)$. On the other hand, the E-P characteristic mod(2) is bordism invariant.

Proposition 14.13. *Let $[X] = 0 \in \eta_n$. Then $\chi_{(2)}(X) = 0 \in \mathbb{Z}/2\mathbb{Z}$.*

Proof. If n is odd, we know in general that $\chi(X) = 0$. Assume that n is even. Let $X = \partial W$, W being a compact manifold with boundary. Take the double $D(W)$. The double $D(W)$ can be presented as the composition of the triad

$$(X \times D^1, \emptyset, (X \times \{-1\} \amalg (X \times \{1\})))$$

followed by two copies of the triad

$$(W, X, \emptyset)$$

glued to $X \times D^1$ along $X \times \{\pm 1\}$, respectively. By the additive property

$$\chi(D(W)) = \chi(X \times D^1) + 2\chi(W, X, \emptyset)\ .$$

By Lemmas 14.8 and 14.7 and the fact that $n+1$ is odd and the double is boundaryless, we have

$$\chi(X) = \chi(D(W)) - 2\chi(W, X, \emptyset) = 2\chi(W) \in \mathbb{Z}$$

so that

$$\chi_{(2)}(X) = 0 \in \mathbb{Z}/2\mathbb{Z}\ .$$ ■

As an immediate corollary, we have the *nontriviality of η_{2n} and Ω_{4n}*.

Corollary 14.14. *For every $n \geq 1$, $\eta_{2n} \neq 0$ and $\Omega_{4n} \neq 0$.*

Proof. We know that $\chi(\mathbf{P}^{2n}(\mathbb{R})) = 1$; hence $[\mathbf{P}^{2n}(\mathbb{R})] \neq 0 \in \eta_{2n}$. Similarly $\chi_{(2)}(\mathbf{P}^{2n}(\mathbb{C})) = 1$; hence $[\mathbf{P}^{2n}(\mathbb{C})] \neq 0 \in \Omega_{4n}$. ■

By using the multiplicative property of χ and the obvious fact that it is additive under disjoint union, we also have the following.

Chapter 1

The smooth category of open subsets of Euclidean spaces

In this book, we will deal with manifolds. Roughly, a manifold is a topological space locally modeled on a Euclidean space $\mathbb{R}^n$, $n \in \mathbb{N}$. Let's recall a few facts about these local models. Since many of them should be familiar to the reader, we will at times omit the proofs or just sketch them out.

1.1. Basic structures on $\mathbb{R}^n$

Every space $\mathbb{R}^n$, $n \in \mathbb{N}$, is endowed with a variety of structures that will be employed case by case.

$\mathbb{R}^n$ is the *vector space* of column vectors (with n rows). We stipulate that if $v \in \mathbb{R}^n$ occurs as a vector in any linear algebra formula, then it is considered as a column.

The space $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ of linear maps $L : \mathbb{R}^n \to \mathbb{R}^m$ coincides with the space of $m \times n$ matrices, $M(m, n, \mathbb{R})$, so that for every $v \in \mathbb{R}^n$, $v \to Lv$ via the usual "rows by column" product. Using the lexicographic order on the entries of any matrix $L = (l_{i,j})_{i=1,\dots,m;j=1,\dots,n}$, we also fix the identification of $M(m, n, \mathbb{R})$ with $\mathbb{R}^{mn}$.

As with every vector space, $\mathbb{R}^n$ has a canonical structure of *affine space* over itself determined by the map that associates to every pair of *points* $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$ the *vector* $\overrightarrow{xy} := y - x$. Every element of $\mathbb{R}^n$ can be considered either as a *point* x of the space or as a *vector* v. For every point x and every vector v, there is a unique point $y := x + v$ such that $v = \overrightarrow{xy}$; for every point x, the map $v \to x+v$ is a bijection. Consider $T(\mathbb{R}^n) := \mathbb{R}^n \times \mathbb{R}^n$ where the first (second) factor consists of points (vectors) and the natural projection onto the space factor $\pi_n : T(\mathbb{R}^n) \to \mathbb{R}^n$, $\pi_n(x,v) = x$. For every point x, the fibre $T_x\mathbb{R}^n = \pi^{-1}(x)$ is a copy of the vector space $\mathbb{R}^n$, called the *tangent vector space* to the affine space at the point x. We call $(T(\mathbb{R}^n), \pi_n)$ the *affine tangent vector bundle* of the affine space $\mathbb{R}^n$. The space of points $\mathbb{R}^n$ can be included in $T(\mathbb{R}^n)$ via the *zero section* $s_0(x) = (x, 0)$, so that for every point x, $\pi \circ s_0(x) = x$. For every *affine map* $f : \mathbb{R}^n \to \mathbb{R}^m$, there is a unique linear map $L_f : \mathbb{R}^n \to \mathbb{R}^m$, called the *linear part* of the affine map, such that for every point x and every vector v, $f(x+v) = f(x) + L_f v$. The map f is an affine isomorphism if and only if L_f is a linear isomorphism (in such a case, $n = m$). In other words, the *tangent map* of f,

$$Tf := (f, L_f) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m \times \mathbb{R}^m,\ (x,v) \to (f(x), L_f v)\ ,$$

verifies $\pi_m \circ Tf = f \circ \pi_n$, so that for every point x, the tangent space $T_x\mathbb{R}^n$ goes to the tangent space $T_{f(x)}\mathbb{R}^m$ via the linear map L_f.

$\mathbb{R}^n$ is a *complete metric space* endowed with the *Euclidean distance* $d = d_n$ defined by

$$d(x,y) = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2}\ .$$

The *standard positive definite scalar product* $(*,*) = (*,*)_n$ on $\mathbb{R}^n$ is defined by

$$(x,y) := \sum_{j=1}^{n} x_j y_j = x^t I y$$

with the associated *norm* $||x|| = \sqrt{(x,x)}$. We note that

$$d^2(x,y) = (x-y, x-y)$$

and that the familiar formula

$$(x,y) = ||x|| \cdot ||y|| \cos\theta$$

allows us to recover the measure of the angle formed by the ordered and oriented lines spanned by two nonzero vectors x, y; in particular, they are orthogonal iff $(x,y) = 0$. We see that many basic objects of elementary geometry can be expressed analytically using the standard scalar product.

$\mathbb{R}^n$ is a *topological space* endowed with the topology $\tau = \tau_n$ induced by the distance d_n. As for any metrizable topological space, a subset U of $\mathbb{R}^n$ is *open* (i.e., it belongs to τ) if and only if for every $x \in U$, there is $r > 0$ such that the *"open" n-ball of centre x and radius r*

$$B^n(x,r) := \{y \in \mathbb{R}^n;\ d(x,y) < r\}$$

is contained in U. We will denote by

$$D^n = \overline{B}^n(0,1)$$

the *closed unitary n-ball* also called the unitary *n-disk*, and by

$$S^{n-1} = \partial D^n = \{x \in \mathbb{R}^n;\ d(0,x) = ||x|| = 1\}$$

the *unitary sphere*. It is not hard to verify that "open" balls are indeed open sets and that the open balls with the centre in $\mathbb{Q}^n \subset \mathbb{R}^n$ and rational radius form a *countable basis of open sets of τ* (every open set is a union of such balls). Any other scalar product

$$(x,y)_A := x^t A y$$

defined by a positive definite symmetric matrix $A = A^t$ determines (by the same formulas as above) a norm $||.||_A$, a distance d_A, and an associated topology τ_A. All these distances are *topologically equivalent*; that is, every $\tau_A = \tau$. This can be proved using the version of the elementary *spectral theorem* stating that there exists a basis of $\mathbb{R}^n$ which is simultaneously orthonormal for $(*,*)$ and orthogonal for $(*,*)_A$. Another topologically equivalent distance on $\mathbb{R}^n$ is defined by $\delta(x,y) := \max\{|x_j - y_j|;\ j = 1,\dots,n\}$.

According to general topological definitions, for every $X \subset \mathbb{R}^n$, $\tau \cap X = \{U \cap X;\ U \in \tau\}$ is the topology on X that makes it a *topological subspace* of $(\mathbb{R}^n, \tau)$; given subspaces $X \subset \mathbb{R}^n$, $Y \subset \mathbb{R}^m$, a map $f : X \to Y$ is *continuous* if for every open set $U \subset Y$, the *inverse image* $f^{-1}(U) = \{x \in X;\ f(x) \in U\}$ is an open set of X. A continuous map $f : X \to Y$ is a *homeomorphism* if it is bijective and also the *inverse map* $f^{-1} : Y \to X$ is continuous.

Every subspace $X \subset \mathbb{R}^n$ is metrizable (hence, in particular, *Hausdorff*) by the restriction to X of the distance d (or of any distance topologically equivalent to d); the restriction of any (countable) basis of open sets of τ is a (countable) basis of $\tau \cap X$.

As for every Hausdorff space with a countable basis, a subspace X of $\mathbb{R}^n$ is *compact* (i.e., every open covering of X admits a *finite* subcovering) if and only if it is *sequentially compact* (i.e., every sequence a_n of points of X admits a subsequence a_{j_n} converging to some point x of X). A subspace is compact if and only if it is *closed* (i.e., the complementary is open) and *bounded* (i.e., it is contained in some ball $B^n(0,r)$). $\mathbb{R}^n$ is *locally compact* (for every $x \in \mathbb{R}^n$ the family of *closed balls* $\overline{B}^n(x,r) = \{y \in \mathbb{R}^n;\ d(x,y) \leq r\}$, when $r > 0$ varies, is a basis of compact neighbourhoods of x). The same holds for every subspace X which is a closed subset of $\mathbb{R}^n$.

We recall that a nonempty open subset $U \subset \mathbb{R}^n$ is *connected* if U is the only open-and-closed nonempty subset of U; U is *path-connected* if for every two points x_0, x_1 of U, there is a continuous path $\alpha : [0,1] \to U$ such that

$\alpha(0) = x_0$, $\alpha(1) = x_1$. We have:

Proposition 1.1. *A nonempty open set U of $\mathbb{R}^n$ is connected if and only if it is path-connected.*

Proof. The "if" implication holds in general for arbitrary topological spaces and it is due to the basic fact that intervals in the real line are connected; for "only if", note that "being connected by a continuous path" defines an equivalence relation on U. The equivalence classes are called the *path-connected components* of U. As every open ball $B^n(x, r) \subset U$ is contained in the path-connected component of U which contains $x \in U$, then every path-connected component of U is open; hence there is only one if U is connected. ■

1.2. Differential calculus

Another fundamental structure carried by the spaces $\mathbb{R}^n$ is the *differential calculus*. Let $U \subset \mathbb{R}^n$, $W \subset \mathbb{R}^m$ be open sets. A map

$$f = (f_1, \dots, f_m) : U \to W$$

is said to be $\mathcal{C}^0$ if it is continuous. The map is *differentiable* at $x \in U$ if there is a (necessarily unique) linear map $d_x f \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ that "well" approximates $g(h) = f(x+h) - f(x)$ in a neighbourhood of $h = 0$. Precisely, for every $\epsilon > 0$, there is $\delta > 0$ such that for every h such that $||h|| < \delta$, $x + h \in U$ and

$$||g(h) - d_x f(h)|| \leq \epsilon ||h|| \ .$$

The linear map $d_x f$ is called the *differential of f at x*. The map f is (globally) *differentiable* if it is differentiable at every point $x \in U$. In such a case it is defined as the *differential map*

$$df : U \to M(m, n, \mathbb{R}), \ df(x) := d_x f \ .$$

We say f is $\mathcal{C}^1$ if it is differentiable and df is continuous (being $M(m, n, \mathbb{R})$, that is, $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$, identified with $\mathbb{R}^{mn}$ as we said above). Every $\mathcal{C}^1$-map is $\mathcal{C}^0$. By induction, for every $r \geq 1$, we say that f is $\mathcal{C}^r$ if df is $\mathcal{C}^{r-1}$. In practice, f is $\mathcal{C}^r$, $r \geq 1$, if and only if it is $\mathcal{C}^0$ and for every multi-index $J = j_1 \dots j_n$ of *order* $|J| := j_1 + \cdots + j_m \leq r$, for every $i = 1, \dots, m$, the *partial derivative function*

$$\frac{\partial^J f_i}{\partial^{j_1} x_1 \dots \partial^{j_n} x_n} : U \to \mathbb{R}$$

is defined and continuous. For every $x \in U$, the partial derivatives of the first order can be organized in an $m \times n$ matrix so that

$$d_x f := \left(\frac{\partial f_i}{\partial x_j}(x) \right)_{i=1,\dots,m; \ j=1,\dots,n} \in M(m, n, \mathbb{R}) \ .$$

$f(x_n) = 0$ for every n. Hence $||\frac{f(x_n)-x_n}{x_n}|| = 1$ against the fact that $d_0 f = I_n$. As for the third point, consider the function $g(x) = f(x) - x$, so that

$$||x - z|| - ||f(x) - f(z)|| \leq ||g(x) - g(z)|| \; .$$

As

$$\frac{\partial g_i}{\partial x_j}(x) = \frac{\partial f_i}{\partial x_j}(x) - \frac{\partial f_i}{\partial x_j}(0)$$

we can take ϵ to make $|\frac{\partial g_i}{\partial x_j}(x)| < \frac{1}{2n^2}$ uniformly on $\overline{B}$. Then the conclusive inequality

$$||g(x) - g(z)|| \leq \frac{1}{2}||x - z||$$

is obtained by applying several times the main value theorem for functions of one variable.

Proof of Claim 2. It is enough to prove that for every $y \in B'$ there is a unique $x \in U$ such that $f(x) = y$. The smooth function $h(x) = ||y - f(x)||^2$ has a minimum point p on the compact set $\overline{B}$ and by construction p belongs necessarily to the open ball B. A simple computation shows that $d_p f(y - f(p)) = 0$; hence $y - f(p) = 0$ because $d_p f$ is invertible. As for the uniqueness, this follows by the inequality $||p_1 - p_2|| \leq 2||f(p_1) - f(p_2)||$, so that $p_1 = p_2$ if $f(p_1) = f(p_2) = y$.

Proof of Claim 3. The same inequality implies that $||\phi^{-1}(y_1) - \phi^{-1}(y_2)|| \leq 2||y_1 - y_2||$ and the continuity of ϕ^{-1} follows.

Proof of Claim 4. As we know, it is enough to show that ϕ^{-1} is differentiable. By using directly the definition of the differential, you can prove that $d_y\phi^{-1} = (d_x\phi)^{-1}$, where $y = \phi(x)$.

■

1.9. Topologies on spaces of smooth maps

Let $U \subset \mathbb{R}^n$, $W \subset \mathbb{R}^m$ be open sets. For every $k \geq 0$, we define a topology δ_k on $\mathcal{C}^k(U, W)$; we will denote by $\mathcal{E}^k(U, W)$ the set $\mathcal{C}^\infty(U, W) \subset \mathcal{C}^k(U, W)$ endowed with the subspace topology. We determine δ_k by giving, for every $f = (f_1, \dots, f_m) \in \mathcal{C}^k(U, W)$, a basis of open neighbourhoods $\mathcal{U}_k(f, K, \epsilon)$ where the varying arguments are a compact subset $K \subset U$ and a real $\epsilon > 0$. Then, by definition, $g \in \mathcal{C}^k(U, W)$ belongs to $\mathcal{U}_k(f, K, \epsilon)$ if and only if:

(1) For every $x \in K$, $||g(x) - f(x)|| < \epsilon$.

(2) For every multi-index J such that $|J| = r \leq k$, for every $i = 1, \dots, m$, for every $x \in K$, we have

$$\left\| \frac{\partial^J (g_i - f_i)}{\partial x_1^{j_1} \dots \partial x_n^{j_n}}(x) \right\| < \epsilon \ .$$

We omit the verification that these define bases of neighbourhoods of a topology.

We denote by $\mathcal{E}(U, W)$ the set $\mathcal{C}^\infty(U, W)$ endowed with the *union topology* $\delta = \bigcup_k \delta_k$.

All of these are called *weak topologies*. There are other so-called *strong* topologies, σ_k, on the same sets of maps. By considering, for example, $\mathcal{E}(\mathbb{R}^n, \mathbb{R})$, we can control the difference of two functions, up to an arbitrarily prescribed order, on an arbitrarily given compact set K, but we have not any control "at infinity". The strong topologies σ_k, which contain δ_k being *heavily finer*, allow instead such control at infinity. On another hand, the weak topologies δ_k have nice properties; for example, we can prove that they are metrizable; hence every f has a countable basis of open neighbourhoods. On the contrary, this is not the case for the strong topologies; for example, if we have a converging sequence $g_n \to f$ in $\mathcal{C}^k(\mathbb{R}^n, \mathbb{R})$ for the strong topology, then there exists a compact set K in $\mathbb{R}^n$ such that g_n definitely equals f on the complement of K. *We do not define the strong topologies.* For our purposes, the control at compact sets will suffice.

Let us recall (a particular case of) the classical *Stone-Weierstrass theorem* (see [**Stone**]).

Theorem 1.9. *For every $f \in \mathcal{C}^k(U, \mathbb{R}^m)$, for every $k \geq 0$, for every neighbourhood $\mathcal{U} = \mathcal{U}_k(f, K, \epsilon)$, there exists a polynomial map $p : \mathbb{R}^n \to \mathbb{R}^m$ such that the restriction of p to U belongs to $\mathcal{U}$. In other words, the polynomial maps are dense in $\mathcal{C}^k(U, \mathbb{R}^m)$ for every $k \geq 0$ and in $\mathcal{E}(U, W)$.*

1.10. Stability of submersions and immersions at a compact set

Let $f \in \mathcal{C}^k(U, W)$ be as above, $k \geq 1$, $K \subset U$ a compact set. We say that f is *a submersion (resp. an immersion) at* K if, for every $x \in K$, $d_x f$ is surjective (resp. injective). This is equivalent to the fact that there exists an open neighbourhood $U' \subset U$ of K such that the restriction of f to U' is a submersion (immersion). Here are the stability results.

Proposition 1.10. *If f is either* (1) *a submersion,* (2) *an immersion, or* (3) *an injective immersion at K, then there is a neighbourhood $\mathcal{U} = \mathcal{U}_1(f, K, \epsilon)$ such that every $g \in \mathcal{U}$ shares the same properties of f, respectively.*

Proof. If $n \geq m$ (resp. $n < m$), then every $m \times n$ matrix A has $\binom{n}{m}$ $m \times m$ (resp. $\binom{m}{n}$) submatrices A_j; in any case define

$$\delta(A) = \sum_j (\det A_j)^2 \ .$$

In both cases (1) and (2) the hypothesis is equivalent to $d(x) := \delta(d_x f) > 0$ for every $x \in K$. As d is continuous and K is compact, then

$$\sup_{x \in K} \{d(x)\} = \max_{x \in K} \{d(x)\} = d_0 > 0 \ .$$

It is clear that if g is close enough to f at K in $\mathcal{C}^1(U, W)$, then $\delta(d_x g) > 0$ for every $x \in K$. As for (3), assume that the thesis fails. Then there exist a sequence $g_n \in \mathcal{C}^1(U, W)$ and sequences of points x_n, y_n in the compact set K such that the follwoing hold:

(1) Every g_n is an immersion at K (by (2) as above).

(2) $g_n \to f$ and $dg_n \to df$ uniformly on K.

(3) $x_n \to x$, $y_n \to y$ in K, $x_n \neq y_n$ and $g_n(x_n) = g_n(y_n)$ for every n.

(4) $v_n := \frac{x_n - y_n}{||x_n - y_n||} \to v \in S^{n-1}$ (by the compactness of the unitary sphere S^{n-1}).

Then $g_n(x_n) \to f(x)$, $g_n(y_n) \to f(y)$; hence $x = y$ because f is injective. Hence

$$||g_n(x_n) - g_n(y_n) - d_{y_n} g_n(x_n - y_n)|| / ||x_n - y_n|| \to 0$$

uniformly, so that

$$||d_{y_n} g_n(v_n)|| \to ||d_x f(v)|| = 0 \ .$$

This is absurd because $d_x f$ is injective. ■

1.11. Morse lemma

Let $f \in \mathcal{C}^\infty(U, \mathbb{R})$, U an open set of $\mathbb{R}^n$. A point $p \in U$ is *regular* for f if $d_p f \neq 0$ (that is, f is a submersion near p); we say that p is *critical* (or also *singular*) otherwise. We are interested in the local behaviour of f at p (actually to the germ $[f]_p$). Up to pre- or post-composition with a translation, we can normalize the situation so that $p = 0$ and $f(0) = 0$. Moreover, we can assume that $U = B^n(0, \epsilon_0)$ for some $\epsilon_0 > 0$ and, case by case, we can restrict f to any $U_\epsilon = B^n(0, \epsilon)$, $0 < \epsilon \leq \epsilon_0$. For every ϵ, the commutative ring $\mathcal{C}^\infty(U_\epsilon, \mathbb{R})$ has a canonical ideal

$$\mathfrak{m}_\epsilon = \{g \in \mathcal{C}^\infty(U_\epsilon, \mathbb{R}); \ g(0) = 0\}$$

a bit of direct computation, we see that it is indeed an immersion. Finally, if Ω is a sufficiently small neighbourhood of H in $S(n,\mathbb{R})$ and $W=\Omega\cap\mathfrak{G}_{n,k}$, then for every $A\in W$, V_A belongs $\tilde{U}$; the restriction to W of $A\to L_V$ is a chart of $\mathfrak{G}_{n,k}$ with values in an open neighbourhood U of $0\in M(n-k,k,\mathbb{R})$. We have now proved again that $\mathfrak{G}_{n,k}$ is an embedded smooth manifold of dimension $k(n-k)$ in $S(n,\mathbb{R})$.

3.4. Stiefel manifolds as fibre bundles over Grassmann manifolds

There are natural surjective maps

$$\mathfrak{l}_{n,k}: L_{n,k}\to\mathfrak{G}_{n,k}\ ,$$

$$s_{n,k}: S_{n,k}\to\mathfrak{G}_{n,k}$$

defined in both cases by $B\to A_{[B]}$ where $[B]$ denotes the linear k-subspace of $\mathbb{R}^n$ generated by the columns of B.

Let us concentrate on the map $s_{n,k}$. Note that $[B]=[C]$ if and only if there exists $Q\in O(k)$ such that $C=BQ$, and note that $A_{[B]}=H$ if and only if it is of the form

$$B=\begin{pmatrix}Q\\0\end{pmatrix},\ Q\in O(k)\ .$$

It follows that every fibre of $s_{n,k}$ is diffeomorphic to $O(k)$ and there is a transitive action (on the *right*) of $O(k)$ itself on each fibre.

The map $s_{n,k}$ is *equivariant* for the actions of $O(n)$: $(P,B)\to PB$ on $S_{n,k}$, $(P,A)\to P^tAP$ on $\mathfrak{G}_{n,k}$, respectively. Recall that "equivariant" means that for every (P,B), $A_{[PB]}=P^tA_{[B]}P$. Then it is enough to analyze the behaviour of the restriction of the map to the inverse image $\tilde{\Omega}:=s^{-1}_{n,k}(\Omega)$ (which is an open neighbourhood of J in $S_{n,h}$) of some open neighbourhood Ω of H in $\mathfrak{G}_{n,h}$. For every $B\in S_{n,k}$, if P is the top $k\times k$ submatrix of B, let us express this by writing $B=(P|D)$. Let $\tilde{\Omega}$ be the open neighbourhood of J in $S_{n,k}$ formed by the matrices $B=(P|D)$ such that P is nonsingular. If $B\in\tilde{\Omega}$, then $[B]\cap\mathbb{R}^{n-k}=\{0\}$; hence its image Ω in $\mathfrak{G}_{n,k}$ is an open set. Moreover, if $[(P|D)]=[(R|S)]$, then there is $Q\in O(k)$ such that $(P|D)=(RQ|SQ)$. If P is nonsingular, then R is also necessarily nonsingular. This means that $\tilde{\Omega}=s^{-1}_{n,k}(\Omega)$ is a *saturated* open set of $S_{n,k}$ for the surjective map $s_{n,k}$. We can make explicit $s_{n,k}(B)$ on $\tilde{\Omega}$ by applying to every $[B]$ and its orthonormal basis given by B itself the construction already used above to construct a local parametrization of $\mathfrak{G}_{n,k}$ at H. This shows that $s_{n,k}$ is smooth. Moreover, define $\phi:\tilde{\Omega}\to M(k,n-k,\mathbb{R})$ by $\phi((P|D))=DP^{-1}$. If

$(P|D) = (RQ|SQ)$ as above, then $SQQ^{-1}R^{-1} = SR^{-1}$. Then there is an induced smooth map $\Omega \to M(k, n-k, \mathbb{R})$ whose inverse map is

$$\psi : M(k, n-k, \mathbb{R}) \to \Omega,\ \psi(Z) = A_{[\mathfrak{r}_{n,k}(I_k|Z)]}$$

providing once again a local parametrization of $\mathfrak{G}_{n,k}$ at H. We can summarize this discussion by saying that there is a locally trivializing commutative diagram at H

$$\begin{tikzcd}
\Omega \times O(k) \arrow[r, "\Psi"] \arrow[d, "\pi_\Omega"] & \tilde{\Omega} \arrow[d, "s_{n,k}"] \\
\Omega \arrow[r, "\mathrm{id}_\Omega"] & \Omega
\end{tikzcd}$$

where $\Psi(A, Q) = \psi(Z)Q$, $A = \psi(Z)$. Its orbit by the action of $O(n)$ provides a fibred atlas for the submersion $s_{n,k}$. So we have proved

Proposition 3.2. *The map* $s_{n,k} : S_{n,k} \to \mathfrak{G}_{n,k}$ *is a fibre bundle with fibre* $O(k)$. *Every change of trivialization*

$$\Phi' \circ \Psi(\Omega \cap \Omega') \times O(k) \to (\Omega \cap \Omega') \times O(k)$$

is of the form

$$(p, P) \to (p, PQ(p))$$

where $p \to Q(p)$ *defines a smooth map* $\Omega \cap \Omega' \to O(k)$.

We also have the following topological consequences.

Corollary 3.3. *Every* $\mathfrak{G}_{n,k}$ *is a compact and connected embedded smooth manifold. As a topological space it has the quotient space topology* $S_{n,k}/s_{n,k}$.

- *Real projective spaces.* A particular case of the above discussion is when $k = 1$. In such a case $\mathfrak{G}_{n,1}$ is also denoted by $\mathbf{P}^{n-1}(\mathbb{R})$ and is called the (real) $(n-1)$-*projective space.* In this case, $S_{n,1} = S^{n-1}$ and the map $s = s_{n,1} : S^{n-1} \to \mathbf{P}^{n-1}(\mathbb{R})$ is a *smooth covering map of degree* 2.

- *Complex Stiefel and Grassmann manifolds.* As a smooth manifold $\mathbb{C}^n = \mathbb{R}^{2n}$; hence $M(n, \mathbb{C})$ is a submanifold of $M(2n, \mathbb{R})$, etc. Throughout the above discussion, let us replace:
  - $\mathbb{R}^n$ with $\mathbb{C}^n$. The real linear subspaces of $\mathbb{R}^n$ with the *complex* linear subspaces of $\mathbb{C}^n$.
  - The standard positive definite scalar product on $\mathbb{R}^n$ with the standard positive definite *Hermitian product* on $\mathbb{C}^n$, $\langle v, w \rangle = v^t \bar{w}$.
  - The (real) orthogonal groups $O(n)$ with the *unitary groups*

    $$U(n) := \{A \in \mathrm{GL}(n, \mathbb{C});\ A^{-1} = A^* := \bar{A}^t\}\ .$$

  - The spaces of real symmetric matrices $S(n, \mathbb{R})$ with the spaces of *Hermitian matrices*

    $$H(n, \mathbb{C}) = \{A \in M(n, \mathbb{C});\ A = A^*\}\ .$$

- The spectral theorem for real symmetric matrices with the *spectral theorem for complex Hermitian matrices.*

Then, by repeating *verbatim* the above constructions, for every (n,k) as above, we realize the (unitary) *complex Stiefel manifold* $S_{n,k}(\mathbb{C})$ as a compact embedded smooth manifold in $M(n,k,\mathbb{C})$, the complex Grassmann manifold $\mathfrak{G}_{n,k}(\mathbb{C})$ as a compact embedded smooth manifold in $H(n,\mathbb{C})$ (defined by the usual equations $A^2 - A = 0$, $\mathrm{trace}(A) = k$), the complex projective spaces $\mathbf{P}^{n-1}(\mathbb{C}) = \mathfrak{G}_{n,1}(\mathbb{C})$, and so on. Although we are dealing with spaces based on the *complex numbers*, we stress that in this way we have realized them as *real* embedded smooth manifolds.

We understand that also the next considerations about Stiefel and Grassmann manifolds would have a counterpart for the complex version.

3.5. A cellular decomposition of the Grassmann manifolds

We describe a natural partition of $\mathfrak{G}_{n,k}$ by a finite number of subsets, each one diffeomorphic to some $\mathbb{R}^h$, $0 \leq h \leq \dim \mathfrak{G}_{n,k}$ (i.e., an *open h-cell*) and such that its closure in $\mathfrak{G}_{n,k}$ is a union of cells of lower dimension. Let $L \in \mathfrak{G}_{n,k}$; that is, L is a k-dimensional linear subspace of $\mathbb{R}^n$ (here we make the abuse of identifying $G_{n,k}$ and $\mathfrak{G}_{n,k}$). For every $i = 0, \dots, n$, denote by

$$p_i : \mathbb{R}^n \to \mathbb{R}^{n-i}$$

the projection to the first $n-i$ coordinates, $p_i((x_1, \dots, x_n)^t) = (x_1, \dots, x_{n-i})^t$. The dimensions of $p_i(L) \subset \mathbb{R}^{n-i}$ decrease from k to 0 in exactly k steps; that is, there are integers

$$1 \leq \sigma_1 < \sigma_2 < \cdots < \sigma_k \leq n$$

such that for j that decreases from k to 1,

$$\dim p_{\sigma_j + 1}(L) - \dim p_{\sigma_j}(L) = 1 \ .$$

We call

$$\sigma(L) := (\sigma_1, \dots, \sigma_k)$$

the *Schubert symbol* of L. There is a concrete, elementary way to determine $\sigma(L)$:

- Fix any rank k, $n \times k$ matrix $A \in L_{n,k}$ which projects to $L \in \mathfrak{G}_{n,k}$.

- Apply to A the Gauss algorithm via elementary operations on the columns and get a matrix

$$\hat{A} \in L_{n,k}$$

in *column echelon form* which also projects to L.

(%$%

Pág. 2(33) →

$\mathbb{R}^n$ is a topological space endowed with the topological $T = T_n$ induced by the distance d_n.

As for any metrizable topological space, a subset U of $\mathbb{R}^n$ is open (i.e., it belongs to T) if and only if for every $x \in U$ there is $r > 0$ such that (de tal modo que) the "open" n-ball of centre x and radius r

$$B^n(x, r) := \{y \in \mathbb{R}^n ; d(x, y) < r\}$$

Pág. 21 (52) 1.9

Topologies on spaces of smooth maps

Pág. 22 (53)

classical Stone-Weierstrass theorem (see Stone).

Pág. 51 (82) - Corollary 3.3

- Real projective spaces
- Complex Stiefel and Grassmann manifolds.

$\mathbb{R}^n$ with $\mathbb{C}^n$.
Complex linear subspaces of $\mathbb{C}^n$.
Hermitian product on $\mathbb{C}^n$
orthogonal groups $O_{(n)}$
Hermitian ~~spaces~~ matrices
spectral theorem for complex Hermitian matrices

—

- 20211013/ Geometry the Line and the Circle/ Geometry - The Line and the Circle (2018).pdf
- 20211013/ Squaring the Circle.../ Ernest Hobson - Squaring the circle the Circle ... (1988).pdf

Editorial:

- Ernest Hobson - Squaring ... djvu
Pag. 12(18) - Pi, being transcendental, ...
- Ernest .. the theory of functions .. (2008).djvu

/
! -256- - -* ! - / 7 3 9 ? ! "
22 1"

D -256- - -* ! - / = : D > !
; * !1 / 1"

-256- - -* ! - / = : D > !
! " * !1 1"

-256- - -* ! - / = : D > ! ! "
* !1 "KA

" -256- - -* ! - / = @ ! : D 9 3
! ! A! !D ! " 3 < 0 + D J
?? 1"
-256- - -* ! - / = @ ! : D 9 3
! ! A! !D ! " 3 < 0 + D J ??
"KA

D

-256- - -* ! - / = : D > ! ! "
* !1 1"

(J D ! " ! J €

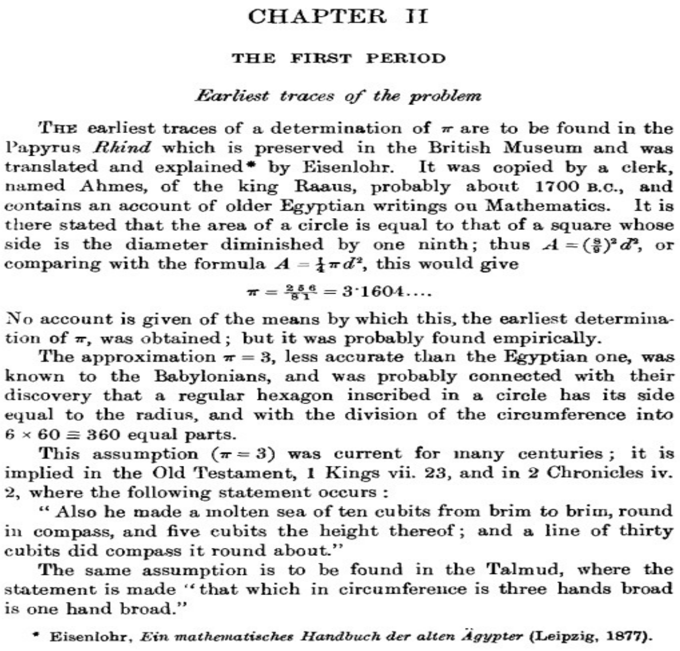

CHAPTER II

THE FIRST PERIOD

Earliest traces of the problem

THE earliest traces of a determination of π are to be found in the Papyrus *Rhind* which is preserved in the British Museum and was translated and explained* by Eisenlohr. It was copied by a clerk, named Ahmes, of the king Raaus, probably about 1700 B.C., and contains an account of older Egyptian writings on Mathematics. It is there stated that the area of a circle is equal to that of a square whose side is the diameter diminished by one ninth; thus $A = (\frac{8}{9})^2 d^2$, or comparing with the formula $A = \frac{1}{4}\pi d^2$, this would give

$$\pi = \frac{256}{81} = 3 \cdot 1604 \ldots .$$

No account is given of the means by which this, the earliest determination of π, was obtained; but it was probably found empirically.

The approximation $\pi = 3$, less accurate than the Egyptian one, was known to the Babylonians, and was probably connected with their discovery that a regular hexagon inscribed in a circle has its side equal to the radius, and with the division of the circumference into $6 \times 60 \equiv 360$ equal parts.

This assumption ($\pi = 3$) was current for many centuries; it is implied in the Old Testament, 1 Kings vii. 23, and in 2 Chronicles iv. 2, where the following statement occurs:

"Also he made a molten sea of ten cubits from brim to brim, round in compass, and five cubits the height thereof; and a line of thirty cubits did compass it round about."

The same assumption is to be found in the Talmud, where the statement is made "that which in circumference is three hands broad is one hand broad."

---

* Eisenlohr, *Ein mathematisches Handbuch der alten Ägypter* (Leipzig, 1877).

The earlier Greek Mathematicians

It is to the Greek Mathematicians, the originators of Geometry as an abstract Science, that we owe the first systematic treatment of the problems of the quadrature and rectification of the circle. The oldest of the Greek Mathematicians, Thales of Miletus (640—548 B.C.) and Pythagoras of Samos (580—500 B.C.), probably introduced the Egyptian Geometry to the Greeks, but it is not known whether they dealt with the quadrature of the circle. According to Plutarch (in *De exilio*), Anaxagoras of Clazomene (500—428 B.C.) employed his time during an incarceration in prison on Mathematical speculations, and constructed the quadrature of the circle. He probably made an approximate construction of an equal square, and was of opinion that he had obtained an exact solution. At all events, from this time the problem received continuous consideration.

About the year 420 B.C. Hippias of Elis invented a curve known as the τετραγωνίζουσα or Quadratrix, which is usually connected with the name of Dinostratus (second half of the fourth century) who studied the curve carefully, and who shewed that the use of the curve gives a construction for π.

This curve may be described as follows, using modern notation.

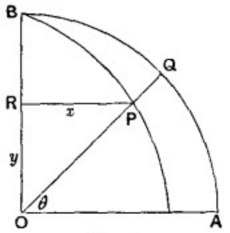

Fig. 1.

Let a point Q starting at A describe the circular quadrant AB with uniform velocity, and let a point R starting at O describe the radius OB with uniform velocity, and so that if Q and R start simultaneously they will reach the point B simultaneously. Let the point P be the intersection of OQ with a line perpendicular to OB drawn from R. The locus of P is the quadratrix. Letting $\angle QOA = \theta$, and $OR = y$, the ratio y/θ is constant, and equal to $2a/\pi$, where a denotes the radius of the circle. We have

$$x = y \cot\theta, \text{ or } x = y \cot \frac{\pi y}{2a},$$

the equation of the curve in rectangular coordinates. The curve will intersect the x axis at the point

$$x = \lim_{y=0} \left(y \cot \frac{\pi y}{2a} \right) = 2a/\pi.$$

If the curve could be constructed, we should have a construction for the length $2a/\pi$, and thence one for π. It was at once seen that the construction of the curve itself involves the same difficulty as that of π.

The problem was considered by some of the Sophists, who made futile attempts to connect it with the discovery of "cyclical square numbers," *i.e.* such square numbers as end with the same cipher as the number itself, as for example $25 = 5^2$, $36 = 6^2$; but the right path to a real treatment of the problem was discovered by Antiphon and further developed by Bryson, both of them contemporaries of Socrates (469—399 B.C.). Antiphon inscribed a square in the circle and passed on to an octagon, 16agon, &c., and thought that by proceeding far enough a polygon would be obtained of which the sides would be so small that they would coincide with the circle. Since a square can always be described so as to be equal to a rectilineal polygon, and a circle can be replaced by a polygon of equal area, the quadrature of the circle would be performed. That this procedure would give only an approximate solution he overlooked. The important improvement was introduced by Bryson of considering circumscribed as well as inscribed polygons; in this procedure he foreshadowed the notion of upper and lower limits in a limiting process. He thought that the area of the circle could be found by taking the mean of the areas of corresponding in- and circum-scribed polygons.

Hippocrates of Chios who lived in Athens in the second half of the fifth century B.C., and wrote the first text book on Geometry, was the first to give examples of curvilinear areas which admit of exact quadrature. These figures are the menisci or lunulae of Hippocrates.

If on the sides of a right-angled triangle ACB semi-circles are described on the same side, the sum of the areas of the two lunes AEC, BDC is equal to that of the triangle ACB. If the right-angled triangle is isosceles, the two lunes are equal, and each of them is half the area of the triangle. Thus the area of a lunula is found.

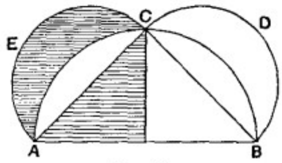

FIG. 2.

If $AC = CD = DB =$ radius OA (see Fig. 3), the semi-circle ACE is $\frac{1}{4}$ of the semi-circle $ACDB$. We have now

$$\mathrm{D}AB - 3\mathrm{D}AC = ACDB - 3 \,.\, \text{meniscus } ACE,$$

and each of these expressions is $\frac{1}{4}\mathrm{D}AB$ or half the circle on $\frac{1}{2}AB$ as diameter. If then the meniscus AEC were quadrable so also would be the circle on $\frac{1}{2}AB$ as diameter. Hippocrates recognized the fact that the meniscus is not quadrable, and he made attempts to find other quadrable lunulae in order to make the quadrature of the circle depend on that of such quadrable lunulae. The question of the existence of various kinds of quadrable lunulae was taken up by Th. Clausen* in 1840, who discovered four other quadrable lunulae in addition to the one mentioned above. The question was considered in a general manner by Professor Landau † of Göttingen in 1890, who pointed out that two of the four lunulae which Clausen supposed to be new were already known to Hippocrates.

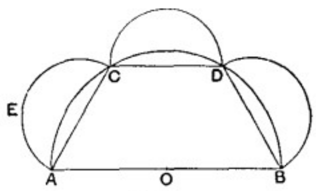

FIG. 3.

From the time of Plato (429—348 B.C.), who emphasized the distinction between Geometry which deals with incorporeal things or images of pure thought and Mechanics which is concerned with things in the external world, the idea became prevalent that problems such as that with which we are concerned should be solved by Euclidean determination only, equivalent on the practical side to the use of two instruments only, the ruler and the compass.

The work of Archimedes

The first really scientific treatment of the problem was undertaken by the greatest of all the Mathematicians of antiquity, Archimedes (287—212 B.C.). In order to understand the mode in which he actually established his very important approximation to the value of π it is necessary for us to consider in some detail the Greek method of dealing with problems of limits, which in the hands of Archimedes provided a method of performing genuine integrations, such as his determination of the area of a segment of a parabola, and of a considerable number of areas and volumes.

This method is that known as the method of exhaustions, and rests on a principle stated in the enunciation of Euclid X. 1, as follows:

"Two unequal magnitudes being set out, if from the greater there be subtracted a magnitude greater than its half, and from that which

---

* *Journal für Mathematik*, vol. 21, p. 375.

† *Archiv Math. Physik* (3) 4 (1903).

20211014///

superlativo
prisioneiros

Advogado constitucionalista
Freiburg. Dietrich Murswiek

Books:

- ~~R~~ 20211014/ Essential University Physics - 2 Volumes/
  - Essential_University_Physics_Volume1.pdf
  - Essential_University_Physics_Volume_2.pdf

Pág. 405(40) Chapter 21 Gauss' Law
Pág. 413(48) Example 21.1
Gauss's Law: A Hollow Spherical Shell
Look: Pág. 414(49) - pictures Sphere

Charges at B contribute $\vec{E}_B$ to the field at P...
and charges at A contribute $\vec{E}_A$..
the two fields cancel.

/
! -256- - -* ! - /
= ! IH A 3I 3 I I I7 D! I=" 1"

D -256- - -* ! - /
= ! IH A 3I 3 I I I7 D! I=" 1"
('$' !1 7! 0 ?!

19 The Second Law of Thermodynamics
20 Electric Charge, Force, and Field
21
22 Electric Potential
23 Electrostatic Energy and Capacitors

Gauss's Law

Skills & Knowledge You'll Need

- Coulomb's law (Section 20.2)
- The dot product of two vectors (Section 6.2)

Learning Outcomes

After finishing this chapter you should be able to:

LO 21.1 Describe electric fields using field lines.

LO 21.2 Explain the concept of electric flux in relation to field lines.

LO 21.3 Describe Gauss's law and explain how it relates to Coulomb's law.

LO 21.4 Use Gauss's law to determine electric fields for charge distributions with sufficient symmetry.

LO 21.5 Approximate electric fields of general charge distributions using what you know about simpler distributions.

LO 21.6 Describe electrostatic equilibrium in conductors and use Gauss's law to determine electric fields associated with charged conductors.

Huge sparks jump to the operator's cage in the Hall of Electricity at the Boston Museum of Science, but the operator is unharmed. Why?

In this chapter we introduce an elegant way of describing electric fields that makes it much easier to calculate the fields of certain charge distributions. In the process we'll formulate one of the four fundamental laws of electromagnetism—a statement that embodies Coulomb's law but that gives deeper insights into the electric field.

21.1 Electric Field Lines

LO 21.1 *Describe electric fields using field lines.*

We can visualize electric fields by drawing **electric field lines**, continuous lines whose direction is everywhere the same as that of the electric field. To draw a field line, determine the field direction at some point. Move a small distance in the direction of the field, and evaluate the field direction at the new point. Extending the process in both directions from your starting point traces out an electric field line. You'll find that field lines begin on positive charges and either end on negative charges or extend to infinity. Drawing many field lines gives a picture of the overall field.

To explore the field of a positive point charge, as shown in Fig. 21.1*a*, start near the charge. The field points radially outward, so move a little way outward. The field is still radial. Repeat the process, and you'll trace a straight line that extends indefinitely. So the field lines of a positive point charge are straight lines that begin on the charge and extend radially to infinity (Fig. 21.1*b*).

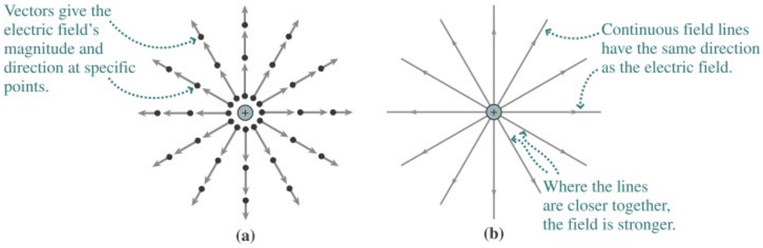
Vectors give the electric field's magnitude and direction at specific points.
Continuous field lines have the same direction as the electric field.
Where the lines are closer together, the field is stronger.
(a)
(b)

FIGURE 21.1 Vectors (a) and field lines (b) provide two ways to visualize the electric field.

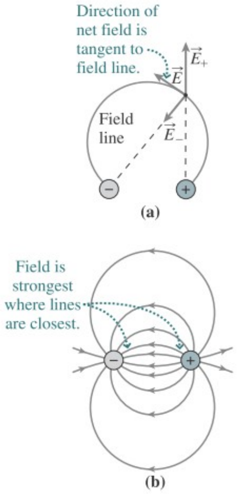
Direction of net field is tangent to field line.
$\vec{E}_+$
$\vec{E}$
Field line
$\vec{E}_-$
(a)
Field is strongest where lines are closest.
(b)

FIGURE 21.2 Field of an electric dipole. (a) At each point, the field-line direction is that of the *net* electric field, $\vec{E} = \vec{E}_+ + \vec{E}_-$ (b) Tracing many field lines shows the overall dipole field.

In Fig. 21.1*b* the field lines spread apart as they extend farther from the point charge. Coulomb's law shows that the field weakens farther from the charge, so in Fig. 21.1*b* the field is stronger where the lines are closer and weaker where they're farther apart. This is generally true, and lets us infer the field's relative magnitude as well as direction from field-line pictures.

To trace the field lines of a charge distribution, follow the net field—the vector sum of the field contributions from all charges in the distribution. Usually the field direction varies, so the line is curved. Figure 21.2*a* shows the details for one field line of a dipole. Figure 21.2*b* shows a number of dipole field lines; here you can see that the field is strongest near the individual charges and in the region between them. The electric field exists everywhere, so there are really infinitely many field lines. We can't draw them all, but in order to make field-line pictures somewhat precise, we associate a fixed number of field lines with a charge of a given magnitude. In Fig. 21.3, for example, eight field lines correspond to a charge of magnitude q. Study the figure, and you'll see how all the fields shown are consistent with this convention.

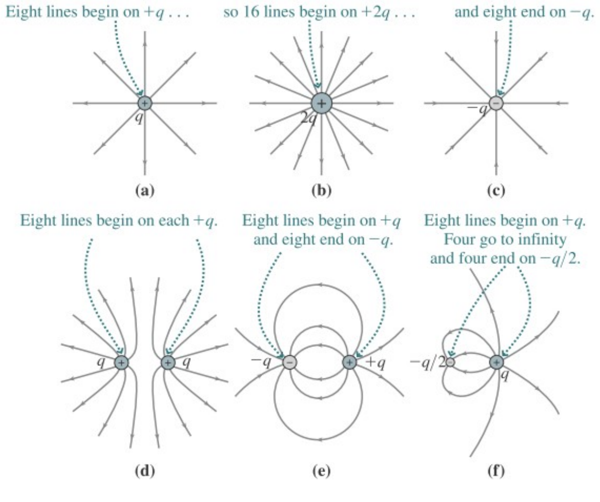
Eight lines begin on $+q$ . . .
so 16 lines begin on $+2q$ . . .
and eight end on $-q$.
q
$2q$
$-q$
(a)
(b)
(c)
Eight lines begin on each $+q$.
Eight lines begin on $+q$ and eight end on $-q$.
Eight lines begin on $+q$. Four go to infinity and four end on $-q/2$.
q
q
$-q$
$+q$
$-q/2$
q
(d)
(e)
(f)

FIGURE 21.3 Field lines for six charge distributions, using the convention that eight lines correspond to a charge of magnitude q.

GOT IT?

21.1 Which figure represents the electric field of a charge distribution consisting of $+2q$ and $-q/2$, using the convention that eight lines correspond to a charge of magnitude q?

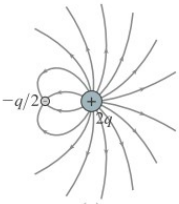

(a)

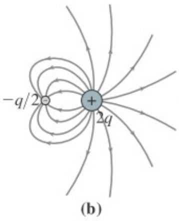

(b)

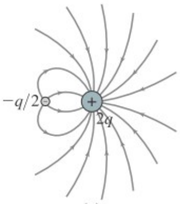

(c)

21.2 Electric Field and Electric Flux

LO 21.2 *Explain the concept of electric flux in relation to field lines.*

In Fig. 21.4 we've surrounded each charge distribution from Fig. 21.3 with several surfaces. Each surface is closed, so it's impossible to get from inside to outside without crossing the surface. (Figure 21.4 shows only the two-dimensional cross section of each surface.) How many field lines emerge from within each surface?

In Fig. 21.4*a* the answer for surfaces 1 and 2 is eight. For surface 3 one field line crosses twice going out and once going in; if we count the inward-going line as negative, then there's still a net of eight lines going out. *Any* closed surface surrounding $+q$ will have eight lines emerging from it, for the simple reason that eight lines begin on the charge and extend to infinity; to get there they all have to cross the closed surface.

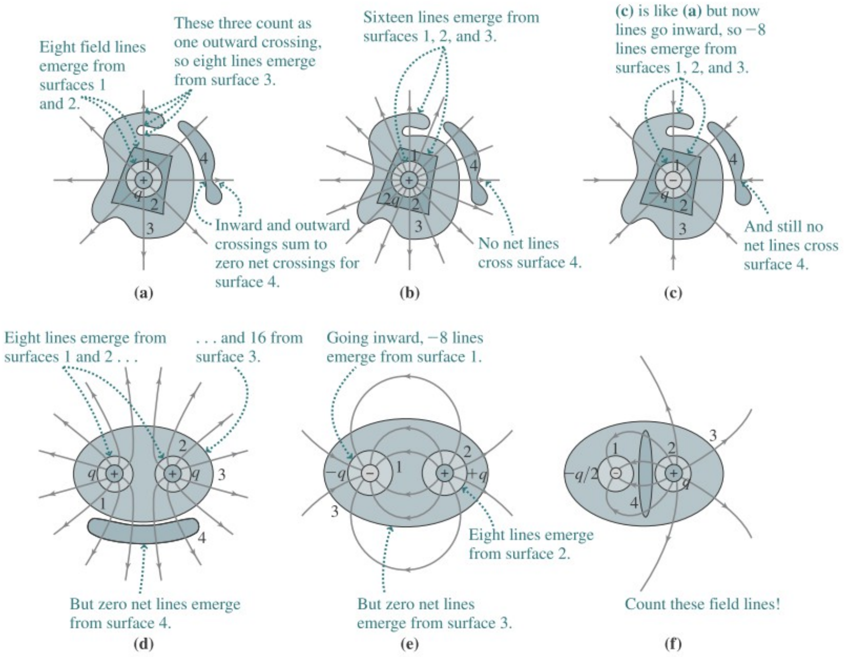
Eight field lines emerge from surfaces 1 and 2.
These three count as one outward crossing, so eight lines emerge from surface 3.
Inward and outward crossings sum to zero net crossings for surface 4.
q
1
2
3
4
(a)
Sixteen lines emerge from surfaces 1, 2, and 3.
No net lines cross surface 4.
$2q$
1
2
3
4
(b)
(c) is like (a) but now lines go inward, so −8 lines emerge from surfaces 1, 2, and 3.
And still no net lines cross surface 4.
$-q$
1
2
3
4
(c)
Eight lines emerge from surfaces 1 and 2 . . .
. . . and 16 from surface 3.
But zero net lines emerge from surface 4.
q
q
1
2
3
4
(d)
Going inward, −8 lines emerge from surface 1.
Eight lines emerge from surface 2.
But zero net lines emerge from surface 3.
$-q$
$+q$
1
2
3
(e)
Count these field lines!
$-q/2$
q
1
2
3
4
(f)

FIGURE 21.4 The number of field lines emerging from a closed surface depends only on the net charge enclosed.

If $\vec{E}$ is parallel to some parts of the area—as happens in line and plane symmetry— then $\vec{E} \perp d\vec{A}$, so $\vec{E} \cdot d\vec{A} = 0$ and there's no contribution to the flux from those areas.

- Evaluate the *enclosed* charge. This may or may not be the same as the total charge, depending on whether the position at which you're evaluating the field lies outside or inside the charge distribution.
- Evaluate the field E by invoking Gauss's law, equating the flux to $q_{\text{enclosed}}/\epsilon_0$, and solving for E. This is the field magnitude; the direction should be evident from symmetry.

ASSESS Does your answer make sense? Does the field behave as you would expect given what you know of simpler charge distributions—point charges, line charges, or charged sheets, depending on the symmetry?

You'll quickly get the hang of this strategy because Gauss's law is useful for finding the field only in the three common symmetries. That means you need to evaluate the integral for the flux just once for a given symmetry.

Spherical Symmetry

A charge distribution has spherical symmetry when the charge density depends only on the radial distance r from the center of the distribution—also called the *point of symmetry*. A point charge is one example, as are the three charge distributions we'll now treat in Examples 21.1 through 21.3. In fact, *any* spherical charge distribution has spherical symmetry, provided the charge density varies solely with distance from the center. The only electric field consistent with spherical symmetry is a field that points in the radial direction, either away from or toward the point of symmetry. We'll start with the simplest case, a charged spherical shell.

EXAMPLE 21.1 Gauss's Law: A Hollow Spherical Shell

A thin, hollow spherical shell of radius R has total charge Q distributed uniformly over its surface. Find the electric field (a) outside and (b) inside the shell.

INTERPRET This is a spherically symmetric charge distribution, so we can use Gauss's law to find the field.

DEVELOP With spherical symmetry, the field lines have to point radially outward for a positive charge and inward for a negative one. We've sketched the charged shell and some field lines outside it in Fig. 21.11. With spherical symmetry, the appropriate Gaussian surfaces are also spheres centered on the center of the shell. Since we're asked for the field both outside and inside the shell, we've sketched one such surface outside the shell and one inside.

EVALUATE

- We first evaluate the flux integral $\Phi = \oint \vec{E} \cdot d\vec{A}$ that appears on the left-hand side of Gauss's law. With our choice of a spherical Gaussian surface, the field is everywhere perpendicular to the surface, or parallel to the normal vector $d\vec{A}$ (recall Fig. 21.9). So $\cos\theta = 1$, and the dot product becomes just $E\,dA$. With our Gaussian surface being a sphere centered on the point of symmetry, the field magnitude E is the same all over the Gaussian surface. So E comes outside the integral, leaving $\Phi = E\oint dA$. The remaining integral is the sum of all the little areas dA over the spherical Gaussian surface—and that's the surface area of a sphere, $A = 4\pi r^2$. The flux then becomes $\Phi = 4\pi r^2 E$. Note that this result doesn't depend on whether the surface is outside or inside the shell or on the particular value of its radius r; it follows from the symmetry and our choice of a spherical Gaussian surface alone. So it holds for both Gaussian surfaces shown in Fig. 21.11, outside and inside the shell.

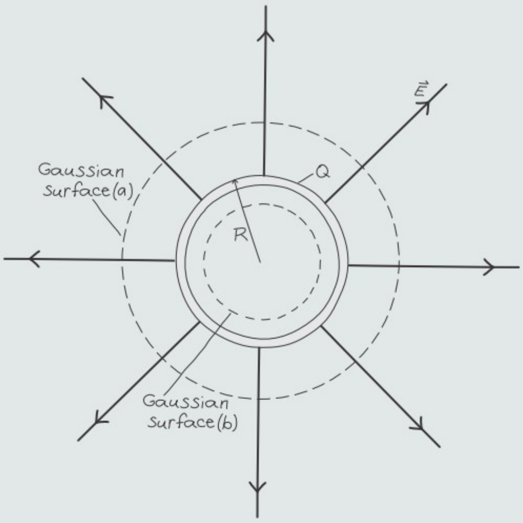

FIGURE 21.11 Finding the field of a charged shell.

(continued)

- Now we evaluate the *enclosed* charge. That *is* different for the two surfaces, so we'll deal first with (a), the field outside the shell. For this we consider the Gaussian surface marked (a); it's outside the shell, so it encloses the entire charge Q. Substituting our expression for the flux, $\Phi = 4\pi r^2 E$, into Gauss's law, $\Phi = \oint \vec{E} \cdot d\vec{A} = q_{\text{enclosed}}/\epsilon_0$, gives $4\pi r^2 E = Q/\epsilon_0$. We can then solve for E to get

$$E = \frac{Q}{4\pi\epsilon_0 r^2} \quad \text{(field outside any spherical charge distribution)} \qquad (21.4)$$

- Now let's consider the interior of the shell. With all the charge on the shell's surface, there's no charge inside. So Gaussian surface (b) doesn't enclose any charge. But, as we argued previously, the flux is still $4\pi r^2 E$, so Gauss's law becomes $4\pi r^2 E = 0$. Solving for E gives $E = 0$. We could have drawn Gaussian surface (b) anywhere inside the shell, including right up to its surface, so there's no electric field anywhere inside the shell.

ASSESS

- Consider first the field outside the shell (Equation 21.4). Since $1/4\pi\epsilon_0$ is the Coulomb constant k, this field can be written $E = kQ/r^2$—precisely the field of a point charge! In fact, our argument leading to this result didn't depend on the charge being on a spherical shell; as long as a charge distribution is spherically symmetric and we're evaluating the field *outside* the distribution, we would find the same result. That's why we've made the result a numbered equation and labeled it "field outside any spherical charge distribution." Outside *any* spherical charge distribution, then, its field is *exactly* the same as that of a point charge with the same total charge, located at the center of symmetry. Incidentally, this result also holds for gravity because it, too, obeys an inverse-square law. That's why we can treat planets as though they were point masses located at their centers.
- Now the interior field: How can it be exactly zero everywhere inside the shell? Figure 21.12 shows that this remarkable result is a consequence of the inverse-square law.

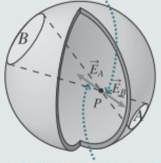

Charges at B contribute $\vec{E}_B$ to the field at P . . .

. . . and charges at A contribute $\vec{E}_A$. The two fields cancel.

FIGURE 21.12 At any point *P* inside a charged shell, the field from the relatively few nearby charges at *A* is exactly canceled by the field from the more numerous but more distant charges at *B*. The result is zero field everywhere inside the shell.

TIP **SYMMETRY MATTERS!** We used the fact that there's no charge inside an empty spherical shell to conclude that the field inside the shell is zero. But be careful: That conclusion follows only when there's enough symmetry, as Fig. 21.13 shows. Here, the Gaussian surface encloses zero net charge, so the flux through the surface is zero. But the electric field both on and inside the surface isn't zero.

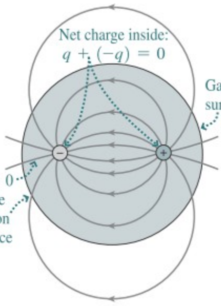

Net charge inside: $q + (-q) = 0$

Gaussian surface

$E \neq 0$ inside and on surface

FIGURE 21.13 A spherical Gaussian surface surrounds a dipole. We have $q_{\text{enclosed}} = 0$, but $E \neq 0$ inside the surface because the dipole isn't spherically symmetric.

GOT IT? **21.4** A spherical shell carries charge Q distributed uniformly over its surface. If the charge on the shell doubles, what happens to the electric field (1) inside and (2) outside the shell?

Now let's consider a slightly more complicated charge distribution: a spherical shell with a point charge at the center.

EXAMPLE 21.2 Gauss's Law: A Point Charge within a Shell

Worked Example with Variation Problems

A positive point charge $+q$ is at the center of a spherical shell of radius R carrying charge $-2q$, distributed uniformly over its surface. Find expressions for the field strength inside and outside the shell.

INTERPRET This problem is about a charge distribution with spherical symmetry.

DEVELOP We want to find the fields both inside and outside the distribution, so we show two spherical Gaussian surfaces in our sketch, Fig. 21.14. We've also drawn some field lines, which will help in assessing our answer.

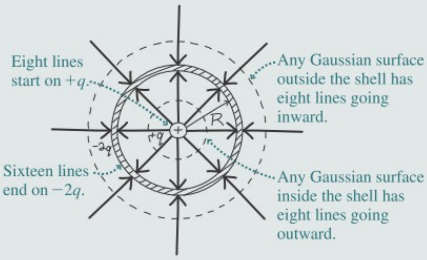

Eight lines start on $+q$.

Sixteen lines end on $-2q$.

Any Gaussian surface outside the shell has eight lines going inward.

Any Gaussian surface inside the shell has eight lines going outward.

FIGURE 21.14 A shell carrying charge $-2q$ surrounds a point charge $+q$. Note that the number of field lines beginning on the point charge and the number ending on the shell correctly reflect their respective charges.

EVALUATE

- We already know that the flux through a spherical Gaussian surface is $\Phi = 4\pi r^2 E$ when we have spherical symmetry, so that's the flux for both surfaces.
- The outer surface encloses the charge $+q$ at the center and $-2q$ on the shell, so the net charge enclosed is $q_{\text{enclosed}} = -q$ for $r > R$. The inner surface encloses only the point charge $+q$, so $q_{\text{enclosed}} = +q$ for $r < R$.
- Now we find the field by equating the flux $\Phi = 4\pi r^2 E$ to $q_{\text{enclosed}}/\epsilon_0$. The result for $r > R$ is $E = -q/4\pi\epsilon_0 r^2$, where the minus sign appears because the enclosed charge is $-q$. For $r < R$ with enclosed charge $+q$, the result is $E = q/4\pi\epsilon_0 r^2$.

ASSESS Make sense? We've seen that the field outside a spherically symmetric charge distribution is that of an equivalent point charge at the center. Here the net charge is $-q$, and our result for $r > R$ is indeed the field of a point charge $-q$. Another way to see this is to apply superposition: The field outside due to the shell alone is that of a point charge $-2q$; that adds to the field of the central point charge $+q$ to give, again, the field of a point charge $-q$ in the region outside the shell. In Example 21.1 we found that the shell produces no field in its interior, so here superposition leaves us with the field of the central point charge alone, just as our result shows. Sketching the field lines, as we've done in Fig. 21.14, also shows how our results make sense.

What if we have charge distributed throughout a volume? As long as the distribution is spherically symmetric, we can find the electric field using Gauss's law just as we did in Examples 21.1 and 21.2. The only difference comes when we evaluate the enclosed charge. Example 21.3 considers this situation.

EXAMPLE 21.3 Gauss's Law: A Uniformly Charged Sphere

A charge Q is distributed uniformly throughout a sphere of radius R. Find the electric field at all points, (a) outside and (b) inside the sphere.

INTERPRET The charge distribution has spherical symmetry, so we can use Gauss's law to find the field.

DEVELOP Figure 21.15 is a sketch of the spherical charge distribution. As you've already seen, the field in any case of spherical symmetry must be radial, so we've sketched some field lines on the figure. As in Examples 21.1 and 21.2, Gaussian surfaces appropriate to spherical symmetry are themselves spheres. Since we're asked for the field both outside and inside the charge distribution, we've drawn two such Gaussian surfaces, one outside and one inside the distribution.

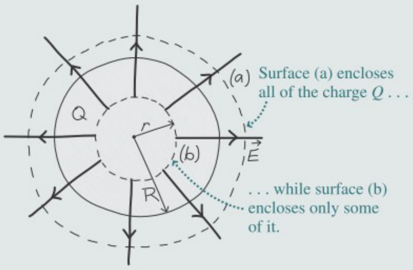

Surface (a) encloses all of the charge Q . . .

. . . while surface (b) encloses only some of it.

FIGURE 21.15 Finding the field of a uniformly charged sphere. The field also exists within the inner Gaussian surface, but we haven't shown it.

EVALUATE

- We already know the answer to (a), the field outside the charge distribution. That's because, as we found in Example 21.1, the field outside *any* spherically symmetric charge distribution is exactly that of a point charge located at the center. So the field outside the charged sphere has magnitude $E_{\text{out}} = Q/4\pi\epsilon_0 r^2$ and points radially

(*continued*)

outward for positive Q. You could also recover this result by considering the flux through surface (a) in Fig. 21.15 and equating it to $q_{\text{enclosed}}/\epsilon_0$—which would essentially repeat the calculation for the field outside the spherical shell in Example 21.1.

- For (b), the field inside the charged sphere, we need to apply Gauss's law to surface (b) in Fig. 21.15. In Example 21.1 we found that the left-hand side of Gauss's law—the flux—is given by $4\pi r^2 E$ whenever we have spherical symmetry and are using a spherical Gaussian surface. However, we still need to find the charge enclosed by surface (b). Since surface (b) lies *within* the sphere of charge, it doesn't enclose the entire charge Q. We can find just how much charge it encloses by comparing its volume with that of the entire sphere. Surface (b) has radius r, so its volume is $\frac{4}{3}\pi r^3$. The charged sphere has radius R, so its volume is $\frac{4}{3}\pi R^3$. So our Gaussian surface, (b), encloses a fraction r^3/R^3 of the total volume of the charged sphere. Because charge is distributed *uniformly* throughout the sphere, this is also the fraction of the total charge enclosed by our Gaussian surface. So q_{enclosed} in Gauss's law becomes Qr^3/R^3 for a Gaussian surface of radius r that's inside the charged sphere. We can now write Gauss's law for this situation, equating the flux, $4\pi r^2 E$, to $q_{\text{enclosed}}/\epsilon_0$: $4\pi r^2 E = Qr^3/\epsilon_0 R^3$. Solving for the field magnitude E gives

$$E = \frac{Qr}{4\pi\epsilon_0 R^3} \quad \text{(field inside uniformly charged sphere)} \qquad (21.5)$$

ASSESS: Make sense? The field inside the sphere increases linearly with distance r from the center. This reflects two opposite effects: First, as we move outward from the center, the enclosed charge grows in proportion to the volume enclosed—that is, as r^3. But the distance from the center grows as well, and that causes a $1/r^2$ decrease in the field. The combined effect is a field that grows linearly from the center to the surface of the sphere, as shown in Equation 21.5. But once we're outside the sphere, the enclosed charge no longer grows, so we see only the $1/r^2$ decrease of the point-charge field. Figure 21.16 plots the resulting field magnitude E both inside and outside the sphere. As the figure shows, and as you can convince yourself by comparing Equations 21.4 and 21.5, our results for the exterior and interior fields agree at the sphere's surface, where $r = R$.

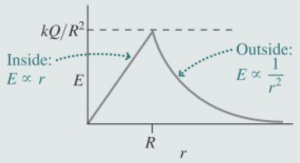

FIGURE 21.16 Field strength versus radial distance for a uniformly charged sphere of radius R.

Before moving on, consider the three examples of Gauss's law in spherical symmetry that you've just seen. In each example we equated the left-hand side of Gauss's law to the right-hand side and solved for the electric field magnitude E. We only had to find the left-hand side once: In every case of spherical symmetry with a spherical Gaussian surface, the left-hand side of Gauss's law is $\Phi = \oint \vec{E} \cdot d\vec{A} = 4\pi r^2 E$. We discovered this in Example 21.1, and you never need to do that calculation again as long as you've got spherical symmetry. So we used $4\pi r^2 E$ for the left-hand side of Gauss's law in both Examples 21.2 and 21.3. Example 21.1 showed us that the field *outside* any spherically symmetric charge distribution is *exactly* that of a point charge located at the center. Again, that's an exact result—not just an approximation that holds at large distances. It's true right up to the surface of the charge distribution, as Figure 21.16 shows graphically.

Example 21.1 also showed us that the field everywhere *inside* a spherical shell of charge is zero. You could think of any spherically symmetric charge distribution as made up of concentric shells—and that means there's never a contribution to the electric field from charge that's further from the center than the position where you're evaluating the field. In that light, here's another take on the most difficult of our three examples, Example 21.3 for the field within a spherical volume of charge: At any position inside that volume, there's no contribution from the charge that's further out. And, as Example 21.1 also showed, the charge interior to that position acts like a point charge. Work out that charge—essentially the calculation leading to q_{enclosed} in Example 21.3—and you'll get Equation 21.5 for the field inside a uniformly charged sphere. So all spherically symmetric charge distributions are similar, in that we apply Gauss's law in exactly the same way. The only difference is in finding the enclosed charge. Problems 69 and 71 provide additional examples involving charge distributions that are spherically symmetric but not uniform.

We'll now consider the two other symmetries where Gauss's law provides a quick route to finding the electric field. As you work through the examples for these symmetries, consider how they're similar to the spherical examples you've just seen. The application of Gauss's law works just the same, but now there's a different expression for the flux on the left-hand side, and calculations of enclosed charge are different.

Line Symmetry

A charge distribution has line symmetry when its charge density depends only on the perpendicular distance r from a line, called the *symmetry axis*. Symmetry then requires that the field point radially and that the field magnitude depend only on distance from the axis. It also requires the charge distribution to be infinitely long, so there's no variation parallel to the line. That's impossible, but nevertheless the infinite line is a reasonable approximation to elongated structures like wires. The next two examples explore the application of Gauss's law to line symmetry.

EXAMPLE 21.4 Gauss's Law: An Infinite Line of Charge

Worked Example with Variation Problems

Use Gauss's law to find the electric field of an infinite line charge carrying charge density λ in coulombs per meter.

INTERPRET An infinite line has line symmetry, so we can apply Gauss's law to find the electric field.

DEVELOP Symmetry requires that the field point radially from the line of charge and that its magnitude be the same at a given distance r from the line. Our job is to find a Gaussian surface that exploits this symmetry—a surface on which the surface integral will turn into a simple product of the field magnitude E and the surface area A. We've sketched an appropriate surface in Fig. 21.17: It's a cylinder of radius r and length L, concentric with the line charge. All points on the cylinder's curved surface are the same distance r from the line charge, so the field magnitude E is the same over the curved part of the surface. And because the field must be radial, the angle between $\vec{E}$ and the normal vectors $d\vec{A}$ is zero, making $\cos\theta$ equal to 1 in the dot product $\vec{E}\cdot d\vec{A}$.

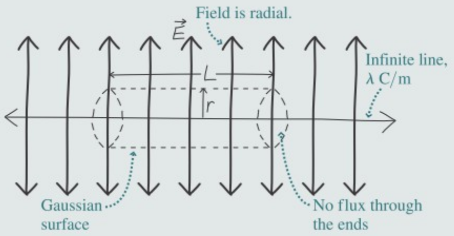

FIGURE 21.17 A cylindrical Gaussian surface surrounds a line charge. The sketch shows a slice in the plane of the page; the field extends radially outward from the line into and out of the page as well.

EVALUATE

- We begin with the left-hand side of Gauss's law, the flux through our Gaussian surface. Unlike the spherical case, this Gaussian surface has two different types of regions—the curved part of the cylinder and its circular ends. For the curved part, we've just seen that $\cos\theta = 1$ in the dot product, so the flux through the curved part becomes $\int \vec{E}\cdot d\vec{A} = \int E\,dA = E\int dA$, where the last equality follows because the magnitude E doesn't change over the curved part of the cylinder. Note that we took the circle off the integral sign because we're just considering the curved part, which alone doesn't constitute a *closed* surface. The remaining integral is just the area of the curved part of the cylinder. Imagine unrolling this curved surface and flattening it out: You'd get a rectangle of length L and width equal to the circumference $2\pi r$, so the flux through the curved part becomes $\Phi = 2\pi rLE$. What about the ends? No field lines emerge from them, so here the flux is zero; mathematically, $\vec{E}\cdot d\vec{A} = 0$ on the ends because the field $\vec{E}$ and the vectors $d\vec{A}$ normal to the ends are perpendicular. The left-hand side of Gauss's law, then, is $2\pi rLE$. This result depends only on the symmetry, so it applies in all situations involving line symmetry.
- Next we need the enclosed charge. The line carries λ C/m and our Gaussian cylinder encloses L meters of the line, so $q_{\text{enclosed}} = \lambda L$.
- Finally, we invoke Gauss's law by equating the flux Φ to $q_{\text{enclosed}}/\epsilon_0$. The result, $2\pi rLE = \lambda L/\epsilon_0$, solves to give

$$E = \frac{\lambda}{2\pi\epsilon_0 r} \quad \text{(field of a line charge)} \tag{21.6}$$

ASSESS Make sense? We worked this same problem in Example 20.7 in a much more difficult calculation involving a complicated integral. Since $1/2\pi\epsilon_0 = 2k$, our result here is the same. But Gauss's law makes the problem much easier!

Although Example 21.4 involved an infinitesimally thin line of charge, you can see from Fig. 21.18 that our result must hold *outside* any charge with line symmetry. And, as we argued in Example 20.7, it's a good approximation for the field of any long, cylindrical structure as long as we're not too near its ends.

Our two examples of line symmetry correspond, respectively, to the case of a point charge and a hollow spherical shell in spherical geometry. In Example Variation 42 you can explore the cylindrical analog of Example 21.3's solid sphere of charge.

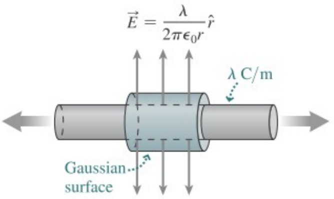

FIGURE 21.18 The arguments of Example 21.4 apply outside *any* cylindrical charge distribution.

EXAMPLE 21.5 Gauss's Law: A Hollow Pipe

A thin-walled pipe 3.0 m long and 2.0 cm in radius carries a net charge $q = 5.7\ \mu\text{C}$ distributed uniformly over its surface. Find the electric field both 1.0 cm and 3.0 cm from the pipe axis, far from either end.

INTERPRET Although the pipe has finite length, both distances are small compared with that length, so we can approximate the pipe as an infinitely long structure with line symmetry.

DEVELOP With line symmetry the appropriate Gaussian surface is a cylinder coaxial with the pipe. We've drawn two such cylinders in Fig. 21.19, one for each radius where we're asked to evaluate the field.

EVALUATE

- We showed in Example 21.4 that the flux integral in line symmetry gives $\Phi = 2\pi rLE$.

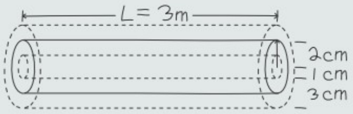

FIGURE 21.19 Gaussian surfaces for Example 21.5.

- Next we need the enclosed charge. At 3 cm we're outside the pipe, so the Gaussian surface with this radius encloses all the charge: $q_{\text{enclosed}} = 5.7\ \mu\text{C}$. The pipe is hollow, so at 1 cm the enclosed charge is zero.
- Equating the flux to $q_{\text{enclosed}}/\epsilon_0$ and solving for E give

$$E = \frac{q_{\text{enclosed}}}{2\pi\epsilon_0 rL} = \frac{5.7\ \mu\text{C}}{(2\pi\epsilon_0)(3.0\times10^{-2}\ \text{m})(3.0\ \text{m})} = 1.1\ \text{MN/C}$$

for the field at 3 cm and $E = 0$ for the field inside the pipe.

ASSESS Make sense? *Inside* the pipe, there's no field, and for the same reason as inside a uniformly charged hollow sphere—namely, that fields from near and far parts of the pipe cancel due, ultimately, to the inverse-square law. Again, be careful: That result follows because of the symmetry, although we'll soon see that with *conducting* pipes and shells there's no interior field even without that symmetry. We argued earlier that the field *outside* any line-symmetric distribution should be given by Equation 21.6, $E = \lambda/2\pi\epsilon_0 r$. In our result, the quantity q_{enclosed}/L is the line charge density λ, so our result is indeed consistent with that equation.

Plane Symmetry

A charge distribution has plane symmetry when its charge density depends only on the perpendicular distance from a plane. The only electric-field direction consistent with this symmetry is perpendicular to the plane. As with line symmetry, true plane symmetry implies a charge distribution that's infinite in extent. That's impossible—but plane symmetry remains a good approximation when charge is spread uniformly over large, flat surfaces or slabs. The next example applies Gauss's law to plane symmetry.

EXAMPLE 21.6 Gauss's Law: A Sheet of Charge

An infinite sheet of charge carries uniform surface charge density σ in coulombs per square meter. Find the resulting electric field.

INTERPRET Since the sheet is infinite, we have plane symmetry with the sheet itself as the symmetry plane.

DEVELOP We need a Gaussian surface chosen so the flux integral $\oint \vec{E}\cdot d\vec{A}$ reduces to a simple multiplication, at least over some parts of the area; over any remaining parts it should be zero. Because the electric field must be perpendicular to our charged sheet, *any* surface with ends parallel to the sheet will have $\cos\theta$ equal to 1 in the dot product. Also, the symmetry dictates that E can't change as we move in a direction parallel to the sheet, since we'd remain the same distance from it. Furthermore, any surface whose sides are perpendicular to the charged sheet will have zero flux through the sides. So we're free to choose for our Gaussian surface any surface with ends parallel to the sheet and sides perpendicular. To ensure that E has the same value over both ends of our surface, we'll have it extend equal distances on either side of the charged sheet. A simple surface that meets these criteria is a cylinder that straddles the charged sheet, as shown in Fig. 21.20.

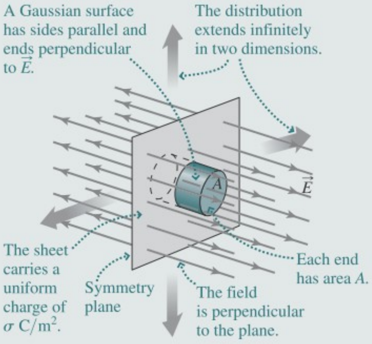
A Gaussian surface has sides parallel and ends perpendicular to $\vec{E}$.
The distribution extends infinitely in two dimensions.
A
$\vec{E}$
The sheet carries a uniform charge of σ C/m^2.
Symmetry plane
Each end has area A.
The field is perpendicular to the plane.

FIGURE 21.20 An infinite sheet of charge, with a Gaussian surface straddling the sheet.

EVALUATE

- First we evaluate the flux through our Gaussian cylinder. The field is parallel to the sides, so there's no flux contribution here. Then the total flux is through the ends, each of which has area A. Since the field is uniform over each end and perpendicular to the ends, the flux through each end is EA. So the total flux through both ends is $\Phi = 2EA$. This result is independent of the details of the charge distribution, so it holds in all cases of plane symmetry.
- Next we need the enclosed charge. The Gaussian surface encloses an area of the sheet equal to the area A of its ends. With surface charge density σ on the sheet, the enclosed charge is $q_{\text{enclosed}} = \sigma A$.
- Now we apply Gauss's law, equating the flux to $q_{\text{enclosed}}/\epsilon_0$. Thus $2EA = \sigma A/\epsilon_0$, so

$$E = \frac{\sigma}{2\epsilon_0} \quad \text{(field of a charged sheet)} \tag{21.7}$$

The direction of this field on either side of the sheet is outward from the sheet if it's positively charged and inward if negative.

ASSESS Make sense? With an infinite plane, symmetry requires that the field lines be perpendicular to the plane. So they don't spread out, and that means the field doesn't vary with distance—as Equation 21.7 shows because it doesn't involve the distance from the sheet. Although our result is exact only for a truly infinite sheet, it's a good approximation near any large, flat, uniformly charged surface as long as we're not close to an edge.

21.5 Fields of Arbitrary Charge Distributions

LO 21.5 *Approximate electric fields of general charge distributions using what you know about simpler distributions.*

Although Gauss's law is always true, most charge distributions lack the symmetry needed to apply Gauss's law to find the field. The alternative, Coulomb's law, is hard to use in all but the simplest cases. But we can often learn a lot by considering the distributions whose fields we calculated here and in Chapter 20. Figure 21.21 summarizes

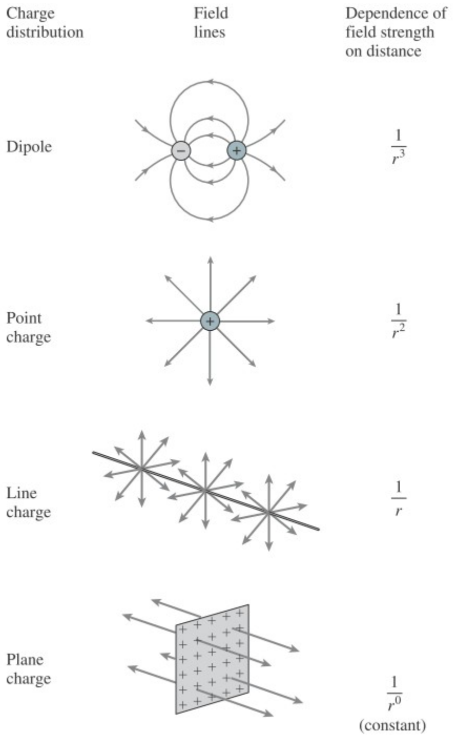

FIGURE 21.21 Electric fields of a dipole, a point charge, a charged line, and a charged plane.

CONCEPTUAL EXAMPLE 21.1 A Charged Disk

Sketch some electric field lines for a uniformly charged disk, starting at the disk and extending out to several disk diameters.

EVALUATE When we're near the disk and not close to its edge, the disk looks like a large, flat, charged plane. Its field is essentially the uniform field of an infinite plane charge, so we draw straight field lines emanating perpendicular to the disk. Far from *any* finite distribution carrying a nonzero net charge, the field approximates that of a point charge, so farther away we draw field lines going radially outward. Field lines begin only on charges—in this case on the charged disk—so we have to connect close-in and far-out lines. We don't know exactly how the field looks in the intermediate region either close to or far from the disk, so we connect them as best we can. Figure 21.22 is the result.

ASSESS Our sketch is a good approximation to the field of a charged disk. And it obeys Gauss's law because the same number of field lines—namely, all 16 lines we chose to draw—cross any closed surface surrounding the entire disk.

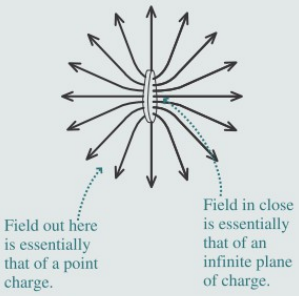
Field out here is essentially that of a point charge.
Field in close is essentially that of an infinite plane of charge.

FIGURE 21.22 The electric field of a charged disk.

MAKING THE CONNECTION Suppose the disk is 1.0 cm in diameter and carries charge 20 nC spread uniformly over its surface. Find the electric field strength (a) 1.0 mm from the disk surface and (b) 1.0 m from the disk.

EVALUATE (a) Close to the disk, assuming we're not near the edge, Equation 21.7 applies: $E = \sigma/2\epsilon_0 = 14\ \text{MN/C}$, where we used the total charge and the disk area to get the surface charge density σ. (b) At 1 meter, the disk is so small it looks essentially like a point charge, so Equation 20.3 applies: $E = kq/r^2 = 180\ \text{N/C}$.

four of these fields. For the last three, note the simple relation between the number of dimensions and the behavior of the field. The plane has two dimensions, and its field doesn't decrease with distance. The line has one dimension, and its field decreases as $1/r$. The point has no dimensions, and its field falls as $1/r^2$. In a sense, the dipole continues this progression. It consists of opposite point charges whose effects nearly cancel; no wonder its field decreases faster still, as $1/r^3$. In fact, there's a hierarchy of charge distributions whose fields decrease ever faster, as dipoles nearly cancel dipoles, and so on. Scientists and engineers use this hierarchy in modeling charged structures ranging from molecules to radio antennas.

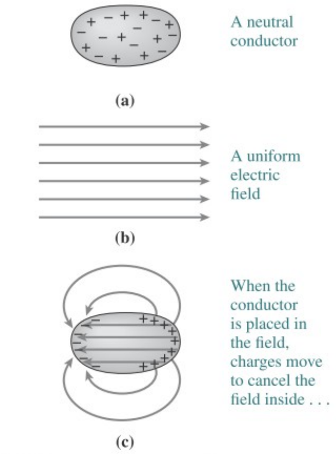
A neutral conductor
(a)
A uniform electric field
(b)
When the conductor is placed in the field, charges move to cancel the field inside . . .
(c)
. . . resulting in this net field.
(d)

FIGURE 21.23 A conductor in a uniform electric field.

GOT IT? **21.5** (1) If you're close to a *finite* line of charge (and not near its ends), does its field vary as (a) $1/r^3$, (b) $1/r^2$, or (c) $1/r$? (2) Repeat for the case when you're far from the line (i.e., much farther than its length).

21.6 Gauss's Law and Conductors

LO 21.6 *Describe electrostatic equilibrium in conductors and use Gauss's law to determine electric fields associated with charged conductors.*

Electrostatic Equilibrium

We've defined conductors as materials that contain free charges, like the free electrons in metals. Figure 21.23 shows what happens when an electric field is applied to a conductor. Free charges respond to the electric force $q\vec{E}$ by moving—in the direction of the field if they're positive, opposite the field if negative. The resulting charge separation gives rise to an electric field within the material that's opposite to the applied field. As more charge

Pág. 416 (51)

I've found that the left-hand side of Gauss's law - the flux - is given $4\pi r^2 E$ whenever we have spherical symmetry and are using a spherical Gaussian surface.

Obs: Livro todo!!!

_20211014/AMS/

Bonahon - Low-dimensional geometry ... (2009).djvu

Pág. 130(141) Exercise 5.16 (The Gauss-Bonnet formula).

Obs: Search: Riemann
24 occurrences.

_20211014/AMS/

Convexity...(1963).djvu

Pág. 239(251)

(iii) In E^3 the point R(K) is the centroid of the Gauss Curvature of the boundary of K (Gericke [2]).

//
! -256- - -* ! - //) ! A 7! ! !
2 // "KA
! -256- - -* ! - //) ! A 7! ! !
2 // 1"

D -256- - -* ! - //) ! ? " !
3 < = " ! ! 31 D . 2 / "KA
D -256- - -* ! - //) ! ? " !
3 < = " ! ! 31 D . 2 / 1"

-256- - -* ! - // A L 3 2 J /B% "KA

-256- - -* ! - // A L 3 2 J /B% 1"
(%/ $ " 7! A!

(iii). In E^3 the point $k(K)$ is the centroid of the Gauss curvature of the boundary of K (Gericke [2]). The geometric properties of $k(K)$ in higher dimensions do not seem to have been investigated.

Another problem which seems to be open is whether (in the plane, or in higher dimensions) $k(K)$ is the only s.i. point satisfying (iii).

The additivity of $k(K)$ is of interest in connection with certain properties of some similarity invariant measures of symmetry (see §§ 4, 5, 6). Considering the related properties of affine invariant measures of symmetry, one is naturally led to the question whether there exist a.i. points $p(K)$ with the additivity property (iii). The negative answer follows at once from the following remarks.

(a) The only a.i. point of a triangle, or a square, is its centroid $g(K)$.

(b) Let $T_{\alpha,\beta}$, $T_{\beta,\alpha}$ and $K_{\alpha+\beta}$ denote the sets represented in Figure 2. Obviously $T_{\alpha,\beta} + T_{\beta,\alpha} = K_{\alpha+\beta}$ and therefore, if $p(K)$ had property (iii), we would have $p(K_{\alpha+\beta}) = ((\alpha + \beta)/3; (\alpha + \beta)/3)$. When $\alpha \to 0$ then $p(K_{\alpha+\beta}) \to (\beta/3, \beta/3)$ while $K_{\alpha+\beta}$ tends to a square S for which $p(S) = (\beta/2, \beta/2)$.

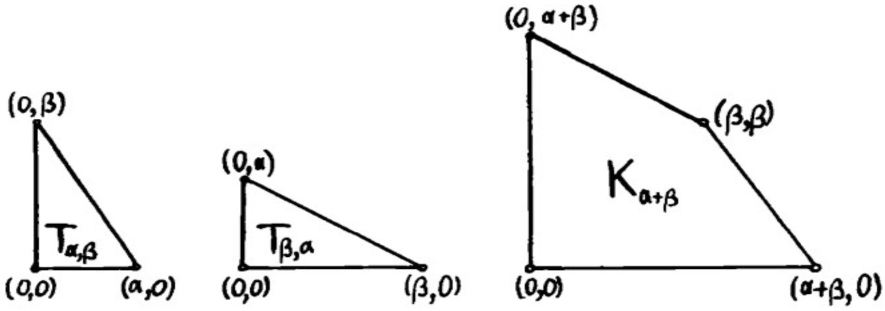

FIGURE 2

Open, and more difficult, is the question whether there exists an a.i. *set* $p(K)$ (with $p(K)$ not identical with K) which satisfies (iii), or the weaker property

$$p(K + K') \supset p(K) + p(K') .$$

If K is a convex body such that there exists a group of affine maps of K onto itself which leave one and only one point of K fixed, then obviously all the a.i. points of K coincide (with that fixed point). Similarly for s.i. points. This implies, in particular, that centrally symmetric sets, and simplices, have a unique a.i. point, an observation which may be used to compute the value of some measures of symmetry for simplices (see, e.g., Fáry-Rédei [1]). The question whether the converse of the above statement is true seems to be open. The same holds for the following problems.

If K is a convex body let K^0 denote the set of all points of E^n (individually) fixed under every affine map of K onto itself. Is the family of all a.i. points

$p(K)$ big enough to ensure $K^0 = \{p(K)\}$? In particular, if $K^0 = E^n$, is $\{p(K)\} = E^n$?

Given a.i. points $p_i(K)$ or K, $1 \leqq i \leqq m$, and reals α_i with $\sum_{i=1}^{m} \alpha_i = 1$, then the combination $\sum_{i=1}^{m} \alpha_i p_i(K)$ is another a.i. point of K. Does there exist, among all the a.i. points of K, a finite "basis" such that each a.i. point is a combination of them?

The same types of problems arise with respect to s.i. points of K.

An additional set of problems is exemplified by the following. For every convex body $K \in \mathfrak{K}^2$, with centroid g and circumellipse-center c, it is easy to show that $g + 2(c - g) \in K$; on the other hand, it is easy to verify that for the set K on Figure 3, $g + 6(g - c) \in K$. What is the greatest interval $[\alpha, \beta]$ such that for every $\lambda \in [\alpha, \beta]$ and every $K \in \mathfrak{K}^2$ the relation $g + \lambda(g - c) \in K$ holds?

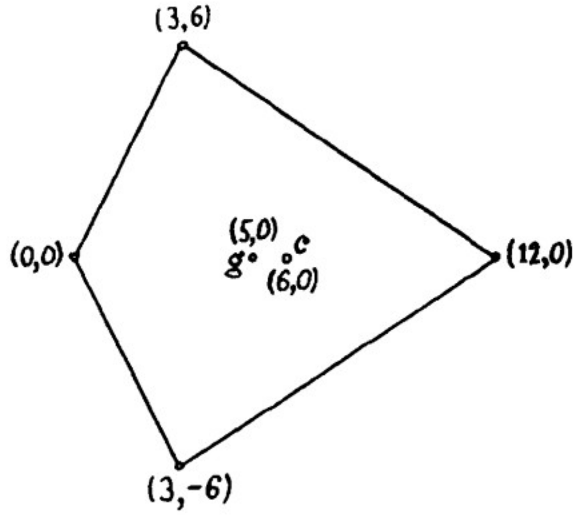

FIGURE 3

A similar problem was suggested by Fenchel (see Pál [1]): In a convex set $K \in \mathfrak{K}^2$ of diameter 1, how far can the centroid of K be from the center of the circumcircle? A partial answer was given by Pál [1]; for a related problem, and an improvement of Pál's result, see Grünbaum [7].

3.3. The most important a.i. and s.i. points considered in the literature are the centroids (centers of gravity) of different mass-distributions related to geometric properties of convex bodies.

(i) The *centroid* $g = g(K)$ of a convex body $K \in \mathfrak{K}^n$ is the centroid of a homogeneous distribution of mass throughout the convex body K. A summary of older results on the properties of the centroid $g(K)$ is given by Bonnesen-Fenchel [1, p. 53]. It would lead us too far to attempt to summarize here the more recent results on this subject (see, however, pp. 249 and 255 for some open problems).

(ii) The *surface-area centroid* $a(K)$ of a convex body $K \in \mathfrak{K}^n$ is the centroid of a homogeneous distribution of mass on the boundary of K. The surface-area centroid received much less attention than $g(K)$. Of the relevant results, not contained in Bonnesen-Fenchel [1], one should mention those by Bose and Roy (published in a series of papers in Bull. Calcutta Math. Soc., starting with Bose-Roy [1]) which generalize and unify many previously known theorems. It seems quite certain that the results of Bose and Roy are susceptible to far-reaching generalizations.[3]

A problem, representative of a whole class of questions about whose solutions nothing seems to be known, is the following. Does there exist, for each convex body $K \subset E^n$, a nonsingular affine transformation T of E^n onto itself such that $a(T(K)) = g(T(K))$?

(iii) The curvature centroid $k(K)$ of a plane convex body K is the centroid of a mass-distribution on the boundary of K, with density proportional to the curvature. (See p. 238 for further properties.) For convex bodies in E^n there are $n-1$ analogs of $k(K)$; very little seems to be known about their properties.

3.4. The circumellipsoid-center $c(K)$ and inellipsoid-center $i(K)$. It is well known that for each convex body $K \in \mathfrak{K}^n$ there exists a unique ellipsoid of minimal volume containing K,—the Löwner-ellipsoid of K. Its center $c(K)$ is an a.i. point of K. The existence of the Löwner-ellipsoids was established by different authors (Behrend [1; 2] for $n = 2$; in the general case, John [1], Danzer-Laugwitz-Lenz [1], Zaguskin [1]).

In a sense dual to Löwner's ellipsoid is the unique ellipsoid of maximal volume contained in the convex body $K \in \mathfrak{K}^n$. This ellipsoid was considered by Behrend [1; 2] for $n = 2$, and by Danzer-Laugwitz-Lenz [1] and Zaguskin [1] for all n. Its center $i(K)$ is an a.i. point of K.

Of considerable interest is the following theorem, first established by John [1] (see also Leichtweiss [2]): For each [centrally symmetric] convex body $K \in \mathfrak{K}^n$ there exists an ellipsoid $E \subset K$ such that the concentric ellipsoid E^*, homothetic to E in the ratio n [resp. $\sqrt{n}$] contains K. If K is not a simplex, the minimal ratio is less than n. (In other words, the d_1 distance (see § 2) of the ball B^n to any element of $\mathfrak{K}^n$ is $\leq \ln n$, while the distance of B^n to any centrally symmetric element of $\mathfrak{K}^n$ is $\leq \ln \sqrt{n}$. See also § 7.)

It is well known (see John [1], Leichtweiss [2]) that the Löwner-ellipsoid of K may be taken for E^* in John's theorem. On the other hand, it is an easy exercise in analytic geometry to prove that the ellipsoid of maximal volume contained in K verifies John's theorem if taken as E. This remark obviously implies the statements about F_1^c and F_1^i (on p. 247).

3.5. The following two a.i. sets should be considered in connection with measures of symmetry of convex bodies in the plane.

(i) The set of *six-partite points* $s(K)$. R. C. Buck and E. F. Buck [1] called *six-partite point* of K each point s with the property that there exist three straight lines through s such that each of the six wedges determined by them contains 1/6 of the area of K. The existence of at least one six-partite point follows from simple continuity arguments applicable not only to the partition of the area of a convex body, but also to the partition of an arbitrary, continuous, distribution of non-negative masses; with a slight change in the definition ("each of the six closed wedges contains at least 1/6 of the total mass") it applies to arbitrary distributions of non-negative masses. Six-partite points, and some of their properties, have been considered by Yaglom-Boltyanskii [1], Eggleston [2; 3; 5], Newman [1].

---

[3] See Note 3 on p. 264.

A convex set may have more than one six-partite point (Eggleston [3, p. 126]); no characterization of convex bodies with a unique six-partite point, or of sets which may serve as the set of all six-partite points of a convex body, seem to be known.

No generalizations of the notion of six-partite points to higher dimensions have been established; for some problems in this direction see Grünbaum [3].

One analogous notion seems to deserve mention. In the plane, for each continuous mass-distribution there exists at least one pair of mutually perpendicular straight lines such that each of the four wedges contains 1/4 of the total mass. Does this generalize in any way to mass-distributions in E^n and to the 2^n wedges formed by n concurrent hyperplanes? The problem seems to be open even for E^3. The only known result is that of Zindler [1]: For every convex body $K \in \mathfrak{K}^3$ and for every plane P_1 bisecting the surface-area of K, there exist planes P_2 and P_3 such that each of the eight wedges determined by P_1, P_2, P_3 contains 1/8 of the surface area of K.

(ii) The set of *hexagonal points* $h(K)$. An affine-regular hexagon is the image of a regular hexagon under a nonsingular affine transformation. Simple continuity arguments prove that each convex body $K \in \mathfrak{K}^2$ admits at least one inscribed affine-regular hexagon (i.e., such that its vertices belong to the boundary of K). This result was established by very many authors, usually as a lemma needed for other results in the geometry of numbers or Minkowski spaces; see Besicovitch [1], Fáry [1], Laugwitz [1], Rešetnyak [1] for typical applications. The center of any affine-regular hexagon inscribed into K is called a hexagonal point $h(K)$ of K. The set of all hexagonal points of K is an a.i. set of K. For a centrally symmetric convex body $K \in \mathfrak{K}^2$, each boundary point of K is the vertex of at least one inscribed affine-regular hexagon. (For further results on inscribed affine-regular hexagons and other regular or affine-regular sets see, e.g., Asplund-Grünbaum [1], Böhme [1], Eggleston [4], Grünbaum [1], Kelly [1], Pucci [1].)

No characterization of convex sets with exactly one inscribed affine-regular hexagon, or of those with only one hexagonal point, seems to be known; the set of all hexagonal points may be a 1-dimensional continuum, but the general nature of such sets is still undetermined.

Another unsolved problem, whose affirmative solution would probably be of interest in various branches of combinatorial geometry, is as follows. The affine-regular hexagon is the vector sum of a triangle and its negative. In analogy with the above, does there exist for each convex body $K \in \mathfrak{K}^n$ at least one inscribed "hyper-hexagon" (i.e., vector sum of a simplex and its negative)? The question is open even in E^3, where the "hyper-hexagon" is a nonsingular affine transform of the Archimedean "cubooctahedron."

It should be mentioned that for circumscribed affine-regular hexagons results completely analogous to those mentioned above hold (L. Danzer, oral communication; it seems that these results are not to be found in the literature). In the same way, the results (Böhme [1], Grünbaum [1]) on inscribed affine-regular pentagons and octagons have their analogues for circumscribed ones.

Another direction which seems to be completely unexplored is that of existence of inscribed affine-regular hexagons (and similar sets) into arbitrary closed Jordan curves.

4. A property of some measures of symmetry. In trying to find additional conditions which are to be satisfied by a function in order to qualify for the name "measure of symmetry," the following property was accepted as "reasonable" by quite a few of the people we consulted.

"The addition of a 'more' symmetric set to a 'less' symmetric one should not decrease the measure of symmetry of the latter set."

More precisely, this "superminimality" condition is: For all $K_1, K_2 \in \mathfrak{K}^n$,

$$f(K_1 + K_2) \geqq \min\{f(K_1), f(K_2)\} . \tag{*}$$

Though an "objective" justification for imposing, or trying to impose, this condition is missing (and seems rather likely to remain so), the intuitive appeal of (*) seems to be sufficient to explain our interest in it. One's enthusiasm is likely to cool off somewhat after noticing that such intuitively "right" measures of symmetry as that of Kovner-Besicovitch (see § 6) fail to have the superminimality property.

We feel that, despite this, the superminimality is of more than marginal interest. Our enlarging on it is due mostly to the fact that (*) is stringent enough to insure that every measure of symmetry (as defined by conditions (1) to (4) in § 1) satisfying it has at least a certain resemblance to the intuitive notion of a measure of symmetry. More precisely, the following theorem holds:

(A) *For every affine [similarity] invariant measure of symmetry for* $\mathfrak{K}^2$ *which satisfies* (*) *we have* $\min\{f(K)\colon K \in \mathfrak{K}^2\} = f(\textit{simplex})$ [*resp.* $\inf\{f(K)\colon K \in \mathfrak{K}^2\} = \inf\{f(T)\colon T \textit{ a simplex } \subset E^2\}$].

Theorem (A) is an immediate consequence of

(B) *Every convex body in* E^2 *is the limit* (*in the Hausdorff metric*) *of finite*

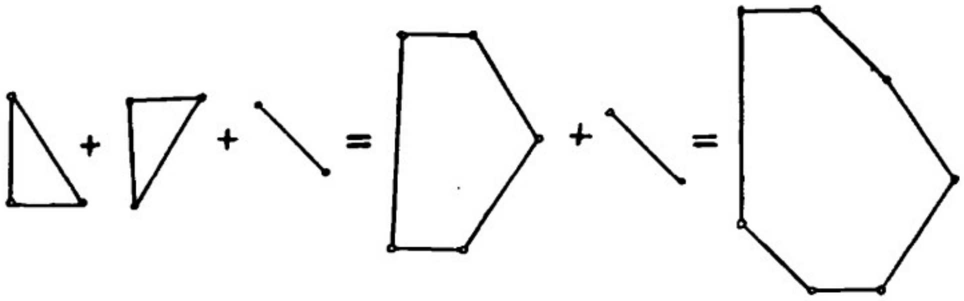

FIGURE 4

*D ! 3 3

(B % // AL 3

9.8. Order-convexity. A set K in a partially ordered space is *order-convex* provided $y \in K$ whenever $x, z \in K$ with $x < y < z$. For results on extreme points and convex hulls with respect to order-convexity, see Franklin [1]. Apparently the literature contains no results of Helly-type about order-convexity in general partially ordered linear spaces, though in any *linearly* ordered space the family of all order-convex sets is intersectional and has Helly-number 2.

9.9. Metric convexity. The usual definition of metric convexity is due to Menger [1]: A set K in a metric space (M, ρ) is *metrically convex* if for each pair of distinct points $x, z \in K$ there exists a point $y \in K$, different from x and z, such that $\rho(x, z) = \rho(x, y) + \rho(y, z)$. When K is not only convex but also metrically complete, then the points x and z of K must lie in a subset of K which is isometric with a linear interval of length $\rho(x, z)$. A closed subset of E^n (or of any strictly convex Banach space) is metrically convex if and only if it is convex in the usual sense. For geodesic distance on the sphere S^n, a closed set is metrically convex if and only if it is weakly convex in the sense of 9.1.

For various geometric developments involving Menger's and other closely related notions of metric convexity, see Blumenthal [3], Busemann [2], Pauc [1], Aronszajn-Panitchpakdi [1], and references listed by these authors. A theorem of Whitehead [1] asserts that for any point p of a Riemannian manifold with positive definite metric, all sufficiently small "spherical" neighborhoods of p are metrically convex in a rather strong sense. This result was sharpened by Nijenhuis [1].

Several authors have studied the relationship between the topological structure of a metrizable space and the existence of a convex metric compatible with the given topology. Menger [1] asked whether every continuous curve admits a convex metric, and after several partial results the problem was finally settled affirmatively by Bing [1]. See Bing [1; 2; 3], Plunkett [1], and Lelek-Nitka [1] for this and related results and references.

In order to extend some of his selection theorems, Michael [1] formulated the notion of a *convex structure* on a metric space, describing a situation in which, as in a Riemannian manifold, it is meaningful to form "convex combinations" of certain (but not necessarily all) k-tuples of points in the space. For another metric notion related to convexity, see Michael [2].

Danzer [2; 3; 4] formulates and applies yet another concept of convexity in metric spaces. For each pair x and z of distinct points of M, let $H(x, z) = \{y \in M : \rho(x, y) < \rho(y, z)\}$. Such a set is called a *halfspace*, and a set is called convex provided it is an intersection of halfspaces. This is closely related to Leibniz's idea (see Busemann [1]) of defining a plane as the set of all points equidistant from two given points. From results of Busemann [1] it follows that when (M, ρ) is a normed linear space with metric generated by the norm, then the closed Danzer-convex sets in M are all metrically convex (in the sense of Menger) only when (M, ρ) is an inner-product space. Sets of the form $\{y \in M : \rho(x, y) = \rho(y, z)\}$ are studied also by Kalisch and Straus [1].

Defining the convex hull of a set as the intersection of all halfspaces containing it, Danzer [5] proves that in a metric space, *any family of spherical cells having nonempty intersection covers the convex hull of the set of its centers.* This was the principal tool in his proof of 7.12 above.

9.10. Convexity in topological groups. Though another approach to convexity in topological groups has been formulated by Ghika [4; 5], we consider here only the approach of Rådström [1; 2]. Let G be a topological group and 2^G the class of all nonempty subsets of G. A *one-parameter semigroup of subsets of G* is defined by Rådström [2] to be a biunique mapping A of $]0, \infty[$ into 2^G such that $A(\delta_1 + \delta_2) = A(\delta_1) + A(\delta_2)$ for all $\delta_i \in]0, \infty[$, where the right-hand "+" indicates the usual addition of sets. Rådström [2] proves the following, suggested by the observation that in E^n, "high powers of small sets are in some sense almost convex": *In a locally convex Hausdorff linear space, the one-parameter semigroups of compact sets are exactly those of the form $A(\delta) = f(\delta) + \delta K$, where f is a one-parameter semigroup of points and K is a compact convex set.*

In Rådström's earlier paper [1], a slightly different definition of one-parameter semigroup leads to the conclusion that in R^n, the one-parameter semigroups of sets are exactly those of the form $A(\delta) = \delta K$ for compact convex K. This suggests that in an arbitrary topological group G, one might define convex sets in terms of the one-parameter semigroups, or at least that these semigroups may play a role similar to that played by convex sets in R^n. Rådström [1] shows that if G is a Lie group, then every one-parameter semigroup of subsets of G has, in a sense, an infinitesimal generator which is a convex set.

9.11. Complex convexity. We turn finally to the generalized convexity which plays such an important role in the theory of functions of several complex variables. This generalizes simultaneously the notions of convex set and convex function, with striking parallelism between the real and complex theories. (Of course they differ in details of proof, the complex theory being much more difficult in some respects.) See Bremermann [1; 2; 3] for an excellent description of this parallelism and for references to other work in the field. In particular, Bremermann [1] supplies a "real-complex dictionary" with the following translation of terms:

Real

linear function of one real variable
convex function of one real variable
convex function of n real variables
convex region in R^n

Complex

harmonic function of one complex variable
subharmonic function of one complex variable
plurisubharmonic function of n complex variables
region of holomorphy in C^n (the space of n complex variables).

20211014(cont.)|||

The geometric properties of k(K) in higher dimensions do not seem to have been investigated.

Pág.161(173) – 9.9 Metric convexity

A theorem of Whitehead [1] asserts that for any point p of a Riemannian manifold with positive definite metric, all sufficiently small "spherical" neighborhoods of p are metrically convex in a rather strong sense.

This result was sharpened by Nijenhuis [1]. (Nijenhuis)

–20211014/AMS/

Mathematics Frontiers and perspectives/

Mathematics-Frontiers...(2000).djvu

Pág. 354(364) – A.

Isometric embedding of surfaces.

An important well-known question is to characterize those ~~metric~~

intrinsic metric on a surface which can be realized as embedding into three space.

-256- - -* ! - ! ! < ! "
1 1 A 2 "KA

-256- - -* ! - ! ! < ! "
1 1 A 2 1"
(%$' %B' 2 & D "" !

A. Isometric embedding of surfaces. An important well-known question is to characterize those intrinsic metrics on a surface which can be realized as embeddings into three space. Minkowski made the first major progress on this problem by proving that any convex polyhedron can be so realized [**148**]. For smooth surfaces with positive curvature, this problem is called the Weyl problem, since Weyl found the first significant estimate for the problem [**216**]. H. Lewy solved the problem in the real analytic category [**120**], and Pogorelov [**165**] and Nirenberg [**158**] solved it in the smooth category. There is also recent work by P. Gaun and Y.Y. Li [**83**].

An important reason that the Weyl problem can be solved is that its solutions must be unique a theorem due to Cohn-Vossen [**46**] and Pogorelov [**164**]. Uniqueness here means that any isometric embedding of the surface is related to any other by a rigid motion in three space. Such global rigidity is difficult to prove for nonconvex surfaces. A.D. Alexandrov studied the next easiest case, namely, those surfaces for which the integral of the positive part of the curvature is 4π [**1**]. He could only prove the rigidity for real analytic metrics. Nirenberg made a beautiful attempt to settle the smooth case [**159**]. Nirenberg's argument would succeed if one can prove that there is at most one closed asymptotic curve in the surface. It is still an outstanding question to understand the behavior of the asymptotic curves on a surface. They are of fundamental importance for the global isometric embedding problem for surfaces with negative curvature, as they are the characteristic curves for the isometric embedding problem. The isometric problem for surfaces of negative curvature is a very interesting nonlinear hyperbolic problem. As such, it is very difficult to prove global existence theorems for such surfaces. In fact, the celebrated Hilbert-Efimov theorem says that a complete surface with strongly negative curvature cannot be isometrically embedded into three-dimensional Euclidean space [**54**].

A significant existence theorem was proved by Hong [**96**], who assumes the curvature decays in the right way. It is a challenging problem to give a transparent and quantitative proof of Efimov's theorem. In other words, take a geodesic disk of radius r whose curvature is not greater than -1 and equal to -1 at the center. What is the largest r so that it can be embedded into $\mathbb{R}^3$ and the second fundamental form is bounded by a given constant?

Global rigidity is false for a general compact surface even when the metric is real analytic. However, an important outstanding problem in classical geometry is whether there exists a continuous family of isometric embeddings of compact surfaces in $\mathbb{R}^3$ which does not arise from rigid motions.

R. Connelly [**49**] and D. Bleecker [**15**] gave beautiful counterexamples to this rigidity conjecture among polyhedra. It is unlikely that those methods can be improved to give smooth examples. There should be (extrinsic) invariants for isometric motions of closed surfaces. Hopefully there are only a finite number of such invariants, so that the space of isometric motions is finite-dimensional. Bleecker's examples show that the volume enclosed by such surfaces is not an invariant. One should be able to generalize these polyhedral examples to piecewise-smooth examples and understand how the motions relate to the motions of the edges.

Also, Cohn-Vossen developed a theory of infinitesimal rigidity for closed surfaces [**47**]. The equation involved is linear and thus is easier to handle. But it is only elliptic if the curvature is positive. Hence for surfaces whose curvature changes sign, it is a difficult problem to understand the equation.

We know that there do exist nontrivial first order isometric deformations for some closed surfaces. The question remains how to characterize such surfaces. These are natural uniqueness questions for mixed-type equations on global surfaces.

Of course, we can also ask all the above questions for open surfaces. Cartan proved that every real analytic surface can be locally isometrically embedded into $\mathbb{R}^3$ [**35**]. The question for a smooth metric is much more difficult. It is an important theorem of C.S. Lin that smooth surfaces with nonnegative curvature can be locally isometrically embedded into $\mathbb{R}^3$ [**136**]. He also settled the problem when the gradient of the curvature is not zero [**137**]. It remains an open problem whether such an assumption can be dropped.

One can formulate a boundary value problem for isometric embeddings of compact surfaces with boundary in two different ways. One is the Neumann problem, which is to require the mean curvature of the boundary to be a given function H. (Clearly we require $H^2 \geq K$.) Of course, one may also require the image of the curve to be a subset of some surface. The other is the Dirichlet problem, which is to require the image of the boundary to be a given Jordan curve. In the first problem, if the curvature is positive, the mean curvature can be bounded and regularity can be guaranteed. In the latter problem, an important necessary condition on the given curve is that the length of its second fundamental form must dominate the geodesic curvature of the original boundary. Hong has made important progress on this problem [**95**].

For closed surfaces with no boundary in $\mathbb{R}^3$, there are many constraints on their intrinsic metrics. The first important constraint is that the curvature must be positive at one point and that $\int_\Sigma K^+ \geq 4\pi$, where $K^+ = \max(K, 0)$.

Nirenberg [**159**] proved that if $\int_\Sigma K^+ = 4\pi$, each component of the set $\{K < 0\}$ is bounded and tangential to parallel planes along two closed convex planar curves. The total geodesic curvature of each closed curve must be equal to 2π. This fact gives a constraint on metrics that can be isometrically immersed into $\mathbb{R}^3$. Is there any constraint on the topology of the sets $\{K > 0\}$ and $\{K < 0\}$ in general?

If Σ is embedded into $\mathbb{R}^3$, it bounds a domain Ω. It is interesting to relate the spectrum of Σ to the spectrum of Ω. A simple argument shows that the volume of Ω is bounded by $\frac{\sqrt{2}}{3}\text{Area}(\Sigma)\lambda_1^{-\frac{1}{2}}$, where λ_1 is the first eigenvalue of Σ. It follows from this statement and the work of Korevaar [**112**] that

$\frac{\mu_i(\Omega)^{3/2}}{\lambda_i(\Sigma)\sqrt{\lambda_1(\Sigma)}} \geq C_g > 0$, where $\mu_i(\Omega)$ is the i-th Dirichlet eigenvalue of Ω and $\lambda_i(\Sigma)$ is the i-th eigenvalue of Σ. Here C_g is a constant depending only on the genus of Σ. What is the optimal C_g and is there a surface that realizes such an optimal constant? Does the set $\{\mu_i(\Omega)\}$ determine $\{\lambda_i(\Sigma)\}$ and vice versa? If a surface Σ can be isometrically embedded into $\mathbb{R}^3$, the spectrum $\{\lambda_i(\Sigma)\}$ is probably constrained. What is the constraint? It is known that asymptotic behavior of $\{\lambda_i(\Sigma)\}$ is closely related to the dynamics of the geodesic flow of Σ. Is the dynamics of geodesic flow constrained? Since Σ cannot have curvature everywhere nonpositive, should the geodesic flow of a closed embedded surface Σ be nonergodic? If it is not ergodic, can one describe the invariant regions of the geodesic flow in terms of extrinsic geometry? What is the entropy of the flow?

B. Different geometries. One can study properties of surfaces that are invariant under a bigger group than the orthogonal group. There are the special linear group, the conformal group and the projective group each acting on $\mathbb{R}^3$. We can ask for properties of a surface Σ that are invariant under these groups. They are called affine, conformal and projective geometries respectively. Many interesting questions remain unanswered for these geometries.

For affine geometry, important invariants are the concepts of the affine metric (which is the second fundamental form divided by the fourth root of the Gaussian curvature) and the affine normal. When the affine normals of the surface focus to a point, the surface is called an affine sphere. Such surfaces were studied extensively by Calabi [**23, 24**], Calabi-Nirenberg [**26**] and Cheng-Yau [**39, 40**]. From these works, we know that for any convex cone, there is a complete affine sphere asymptotic to the cone. It would be very interesting to construct such an affine sphere efficiently. Can one find a representation in terms of pseudoholomorphic functions (see [**213**])? Recently, in his Harvard thesis, John Loftin discussed how to construct an affine sphere by a holomorphic section of the tricanonical line bundle of a Riemann surface. Are there many cases that such spheres can be written in closed form? (The surface $xyz = 1$ is asymptotic to the coordinate cone.) This is especially interesting when the cone is a polyhedral cone. Can we compute the affine metric in these cases? Of course, one asks the same question for higher dimensions, and it is related to the linear optimization problem [**106**].

Up to now, most of the progress assumes convexity of the surfaces. What happens if the surface has negative curvature? What are those affine spheres whose induced metrics are complete and have negative curvature?

There is a concept of an affine maximal surface, which is defined by a fourth-order elliptic equation (see e.g. [**25**]). One can therefore fix the boundary and the normal direction along the boundary for the surface. If we fix the boundary to be a triangle, is there an efficient way to solve the boundary value problem for the affine maximal surface? If we fix a polyhedron in $\mathbb{R}^3$, we can attempt to form a C^1-surface that passes through all the edges of the polyhedron so that in each face, the surface is affine maximal. Among all such C^1-surfaces, we can try to find one with maximum total affine area. Recently, Trudinger and Wang [**209**] solved the Bernstein problem for affine maximal surfaces. Hopefully, the future will bring more activities for affine maximal surfaces.

As we discuss in Cheng-Yau [**40**], the affine metric of the affine sphere in $\mathbb{R}^{n+1}$, after a Legendre transform, becomes a projective invariant metric on a convex domain in $\mathbb{R}^n$. It is therefore interesting to relate projective-invariant properties of surfaces in $\mathbb{R}^n$ to affine properties of hypersurfaces of the affine sphere in $\mathbb{R}^{n+1}$.

It is known that if a closed surface in $\mathbb{R}^3$ is infinitesimally (metrically) rigid, its image under a projective transformation is also infinitesimally rigid. Is the statement true for the global rigidity of a closed surface? Can one formulate the rigidity problem in terms of projective or affine geometry?

Algebraic surfaces in $\mathbb{RP}^3$ are clearly important surfaces. However, we know virtually nothing about them. There is some information about the number of their components. But their geometry is clearly special. The number of umbilical points, the number of closed geodesics, the number of closed asymptotic lines, the topology of the regions of positive curvature, and the eigenfunctions of the Laplacian are all interesting invariants related to the polynomials that define the surface. On the one hand, we have techniques and information that we learn from algebraic geometry. We can complexify the surfaces and obtain information related to the topology and the incidence relations among them. On the other hand, we want to understand the classical geometry of these surfaces using classical methods–partial differential equations or connections from real projective geometry. It will be important to bridge these two approaches. When we complexify the surfaces, in most cases, we can find a complete Kähler-Einstein metric with negative scalar curvature [**202**]. Since the metric is invariant under an anti-holomorphic involution, we have a canonical metric on the surface. (Canonical in the sense that if there is a nonsingular bijective projective linear map from one algebraic surface to another one, it is an isometry of the canonical metrics.) Some theorems on real algebraic varieties can be proved using the canonical metrics. How do we relate this canonical metric to the Euclidean metric induced on the surface?

In complex algebraic geometry, it is very important to understand algebraic curves in an algebraic manifold. What is the role of real algebraic curves in a real algebraic surface? What are the topology and geometry of these curves? How do we bound the number of components and the geodesic curvatures of the curves in terms of the degree of polynomials defining them?

Potentially, every closed surface can be approximated by a real algebraic surface of large degree. What is the minimal degree needed to approximate a given surface? In other words, given $\varepsilon > 0$, what is the minimal degree of an algebraic surface whose Hausdorff distance is less than ε from a given surface? Can one estimate it in terms of a suitable calculable geometric quantity of the surface (e.g., the L^2 norm of the mean curvature of the surface)? If we minimize the L^2 norm of the mean curvature of surfaces with degree less than a given integer, what kind of algebraic surfaces do we get?

Real algebraic surfaces provide a rich class of natural singularities. It is very interesting to study the behavior of the principal curvatures near singularities. The behavior of the eigenfunctions of the Laplacian near such points should also be very interesting.

When a surface is defined by a real algebraic set, there is a (singular) foliation attached to it. The study of such a continuous family should be very helpful in understanding the algebraic surface.

Conversely, if we have an algebraic one-form which defines a (possibly singular) foliation, it will be interesting to find the closed leaves and bound the number of such leaves. When are the leaves algebraic?

C. Minimal surfaces in $\mathbb{R}^3$. There are several important surfaces studied in the literature. Most of these are obtained by variational methods.

The first major class is minimal surfaces. These are surfaces with mean curvature equal to zero. When we are given a closed curve in $\mathbb{R}^3$, we look for surfaces bounded by this curve which minimize area. We can restrict the topology of these surfaces to be within certain classes. For example, we can assume all surfaces have genus g. Then the minimum of the area is a number A_g which may depend on the genus. In general, $A_g \geq A_{g+1} \geq \cdots$ [**145**]. When the boundary curve is smooth, it is a celebrated theorem of Hardt-Simon that for some g_0, $A_{g_0} = A_{g_0+1} = \cdots$ [**88**]. There is no effective estimate of g_0 and it will be challenging to find such an estimate.

It is not known whether a smooth simple closed curve in $\mathbb{R}^3$ can bound an infinite number of minimal surfaces. (If the curve is real analytic, this cannot happen (Tomi [**207**]).) Is there a general algorithm to find all minimal surfaces bounded by a given closed curve (or a set of closed curves)? It is still a difficult problem to calculate unstable minimal surfaces bounded by given curves. (There is also no efficient way to calculate area-minimizing surfaces.) It is not clear how big the genus of a stable or unstable minimal surface must be if it is bounded by a given set of smooth curves. Can it be arbitrarily large?

Many fundamental questions remain unsolved regarding the quantitative behavior of minimal surfaces. One well-known question is that of the best constant for the isoperimetric inequality, i.e., what is the best upper bound of $\frac{4\pi A}{L^2}$, where A is the area and L is the length of the boundary. Naturally one conjectures it to be 1. This is known if the boundary is a Jordan curve (see Li-Schoen-Yau [**128**]). This question is related to finding the best constant

D -256- - -* ! - ! %4 7 3J J
! " A! ! ! 22 9 L D . 1"
(% %% "!1M 7! 0 7! 0 * !" .3 9

the region. We shall always assume that our region is regular, so that the boundary is piecewise regular; for this theorem we orient the boundary *outward*.

Theorem 5.8.5 (Divergence Theorem). [17] *Suppose*

$$\vec{F}(x,y,z) = P(x,y,z)\vec{\imath} + Q(x,y,z)\vec{\jmath} + R(x,y,z)\vec{k}$$

is a $\mathcal{C}^1$ *vector field on the regular region* $\mathfrak{D} \subset \mathbb{R}^3$.

Then the flux integral of $\vec{F}$ *over the boundary* $\partial\mathfrak{D}$*, oriented outward, equals the (triple) integral of its divergence over the interior of* $\mathfrak{D}$:

$$\iint_{\partial\mathfrak{D}} \vec{F} \cdot d\vec{S} = \iiint_{\mathfrak{D}} \operatorname{div}\vec{F}\, dV. \tag{5.33}$$

The proof of this theorem is given in Appendix A.12

We note that, as in the case of Green's Theorem, we can extend the Divergence Theorem to any region which can be partitioned into regular regions.

It can also be extended to regular regions with "holes" that are themselves regular regions. For example, suppose $\mathfrak{D}$ is a regular region and $B_\varepsilon\left(\overrightarrow{x_0}\right)$ is a ball of radius $\varepsilon > 0$ centered at $\overrightarrow{x_0}$ and contained in the interior of $\mathfrak{D}$ (that is, it is inside $\mathfrak{D}$ and disjoint from its boundary). Then the region $\mathfrak{D} \setminus B_\varepsilon\left(\overrightarrow{x_0}\right)$ consisting of points in $\mathfrak{D}$ but at distance at least ε from $\overrightarrow{x_0}$ is "$\mathfrak{D}$ with a hole at $\overrightarrow{x_0}$"; it has two boundary components: one is $\partial\mathfrak{D}$, oriented outward, and the other is the sphere of radius ε centered at $\overrightarrow{x_0}$, and oriented *into* the ball. Suppose for a moment that F is defined inside the ball, as well. Then the flux integral over the boundary of $\mathfrak{D} \setminus B_\varepsilon\left(\overrightarrow{x_0}\right)$ is the flux integral over the boundary of $\mathfrak{D}$, oriented outward, *minus* the flux integral over the boundary of the ball (also oriented outward). The latter is the integral of $\operatorname{div}\vec{F}$ over the ball, so it follows that the flux integral over the boundary of $\mathfrak{D} \setminus B_\varepsilon\left(\overrightarrow{x_0}\right)$ is the integral of $\operatorname{div}\vec{F}$ over its interior. Now, this last integral is independent of what F does inside the hole, provided it is $\mathcal{C}^1$ and agrees with the given value along the boundary. Any $\mathcal{C}^1$ vector field F defined on and outside the sphere can be extended to its interior (Exercise 14), so we have

Corollary 5.8.6. *If the ball* $B_\varepsilon\left(\overrightarrow{x_0}\right)$ *is interior to the regular region* $\mathfrak{D}$*, then the flux integral of a* $\mathcal{C}^1$ *vector field* $\vec{F}$ *over the* boundary *of* $\mathfrak{D}$ *with a hole* $\iint_{\partial(\mathfrak{D}\setminus B_\varepsilon(\overrightarrow{x_0}))} \vec{F} \cdot d\vec{S}$ *equals the integral of* $\operatorname{div}\vec{F}$ *over the interior of* $\mathfrak{D} \setminus B_\varepsilon\left(\overrightarrow{x_0}\right)$:

$$\iint_{\partial(\mathfrak{D}\setminus B_\varepsilon(\overrightarrow{x_0}))} \vec{F} \cdot d\vec{S} = \iiint_{\mathfrak{D}\setminus B_\varepsilon(\overrightarrow{x_0})} \operatorname{div}\vec{F}\, dV. \tag{5.34}$$

In particular, if $\vec{F}$ *s divergence-free in* $\mathfrak{D} \setminus B_\varepsilon\left(\overrightarrow{x_0}\right)$ *then the outward flux of* $\vec{F}$ *over* $\partial(\mathfrak{D} \setminus B_\varepsilon\left(\overrightarrow{x_0}\right))$ *equals the outward flux of* $\vec{F}$ *over the sphere of radius* ε *centered at* $\overrightarrow{x_0}$.

Like Stokes' Theorem, the Divergence Theorem allows us to compute the same integral two different ways. We consider a few examples.

First, let us calculate directly the flux of the vector field

$$\vec{F}(x,y,z) = x\vec{\imath} + y\vec{\jmath} + z\vec{k}$$

out of the sphere $\mathfrak{S}$ of radius R about the origin.

---

[17] This theorem was published by Carl Friedrich Gauss (1777-1855) in 1838 [17] and independently by Mikhail Vasilevich Ostrogradski (1801-1862) in 1831. (see [1]). It is often called *Gauss's Theorem* or the *Gauss-Ostrogradsky Theorem*.

The natural parametrization of this sphere uses spherical coordinates: $x = R\sin\phi\cos\theta$, $y = R\sin\phi\sin\theta$, $z = R\cos\phi$, for $0 \le \phi \le \pi$ and $0 \le \theta \le 2\pi$. The partials are $\frac{\partial\vec{p}}{\partial\phi} = (R\cos\phi\cos\theta)\vec{\imath} + (R\cos\phi\sin\theta)\vec{\jmath} - (R\sin\phi)\vec{k}$ and $\frac{\partial\vec{p}}{\partial\theta} = (-R\sin\phi\sin\theta)\vec{\imath} + (R\sin\phi\cos\theta)\vec{\jmath}$, with cross product

$$\begin{aligned}\frac{\partial\vec{p}}{\partial\phi} \times \frac{\partial\vec{p}}{\partial\theta} &= \begin{vmatrix} \vec{\imath} & \vec{\jmath} & \vec{k} \\ R\cos\phi\cos\theta & R\cos\phi\sin\theta & -R\sin\phi \\ -R\sin\phi\sin\theta & R\sin\phi\cos\theta & 0 \end{vmatrix} \\ &= (R^2\sin^2\phi\cos\theta)\vec{\imath} + (R^2\sin^2\phi\sin\theta)\vec{\jmath} + (R^2\sin\phi\cos\phi)\vec{k}.\end{aligned}$$

To check whether this gives the outward orientation, we compute the direction of this vector at a point where it is easy to find, for example at $(1,0,0) = \vec{p}\left(\frac{\pi}{2}, 0\right)$:

$$\left(\frac{\partial\vec{p}}{\partial\phi} \times \frac{\partial\vec{p}}{\partial\theta}\right)\left(\frac{\pi}{2}, 0\right) = R^2\vec{\imath}$$

which points out of the sphere at $(1,0,0)$. Thus, the element of outward oriented surface area is

$$d\vec{S} = \Big((R^2\sin^2\phi\cos\theta)\vec{\imath} + (R^2\sin^2\phi\sin\theta)\vec{\jmath} + (R^2\sin\phi\cos\phi)\vec{k}\Big)\,d\phi\,d\theta.$$

On the surface, the vector field is $\vec{F}(\phi,\theta) = (R\sin\phi\cos\theta)\vec{\imath} + (R\sin\phi\sin\theta)\vec{\jmath} + (R\cos\phi)\vec{k}$, so

$$\begin{aligned}\vec{F}\cdot d\vec{S} &= \Big((R\sin\phi\cos\theta)(R^2\sin^2\phi\cos\theta) + (R\sin\phi\sin\theta)(R^2\sin^2\phi\sin\theta) \\ &\qquad + (R\cos\phi)(R^2\sin\phi\cos\phi)\Big)\,d\phi\,d\theta \\ &= R^3(\sin^3\phi\cos^2\theta + \sin^3\phi\sin^2\theta + \sin\phi\cos^2\phi)\,d\phi\,d\theta \\ &= R^3\sin\phi(\sin^2\phi + \cos^2\phi)\,d\phi\,d\theta = R^3\sin\phi\,d\phi\,d\theta.\end{aligned}$$

The flux integral is therefore

$$\begin{aligned}\iint_{\mathfrak{S}} \vec{F}\cdot d\vec{S} = \int_0^{2\pi}\int_0^{\pi} R^3\sin\phi\,d\phi\,d\theta &= \int_0^{2\pi} \left(-R^3\cos\phi\right)_{\phi=0}^{\pi}\,d\theta \\ &= 2R^3\int_0^{2\pi} d\theta = 2R^3(2\pi) = 4\pi R^3.\end{aligned}$$

Now let us see how the same calculation looks using the Divergence Theorem. The divergence of our vector field is $\operatorname{div}\vec{F} = 1 + 1 + 1 = 3$, so the Divergence Theorem tells us that

$$\iint_{\mathfrak{S}} \vec{F}\cdot d\vec{S} = \iiint_{\mathfrak{D}} \operatorname{div}\vec{F}\,dV = \iiint_{\mathfrak{D}} 3\,dV = 3\mathcal{V}(\mathfrak{D}),$$

where $\mathfrak{D}$ is the sphere of radius R, with volume $\frac{4\pi R^3}{3}$, and our integral is

$$\iint_{\mathfrak{S}} \vec{F}\cdot d\vec{S} = 4\pi R^3.$$

As another example, let us calculate the flux of the vector field $\vec{F}(x,y,z) = x^2\vec{\imath} + y^2\vec{\jmath} + z^2\vec{k}$ over the same surface. We have already calculated that the element of outward oriented surface area is

$$d\vec{\mathcal{S}} = \left((R^2\sin^2\phi\cos\theta)\vec{\imath} + (R^2\sin^2\sin\theta)\vec{\jmath} + (R^2\sin\phi\cos\phi)\vec{k}\right)d\phi\,d\theta.$$

This time, our vector field on the surface is

$$\vec{F}(\phi,\theta) = (R^2\sin^2\phi\cos^2\theta)\vec{\imath} + (R^2\sin^2\phi\sin^2\theta)\vec{\jmath} + (R^2\cos^2\phi)\vec{k}$$

and its dot product with $d\vec{\mathcal{S}}$ is

$$\begin{aligned}\vec{F}\cdot d\vec{\mathcal{S}} &= \Big((R^2\sin^2\phi\cos^2\theta)(R^2\sin^2\phi\cos\theta) + (R^2\sin^2\phi\sin^2\theta)(R^2\sin^2\sin\theta)\\ &\qquad + (R^2\cos^2\phi)(R^2\sin\phi\cos\phi)\big)\,d\phi\,d\theta\\ &= R^4\left(\sin^4\phi\cos^3\theta + \sin^4\phi\sin^3\theta + \sin\phi\cos^3\phi\right)d\phi\,d\theta\\ &= R^4\big(\frac{1}{4}\left(1 - 2\cos 2\phi + \cos^2\phi\right)\left(\cos^3\theta + \sin^3\theta\right) + \sin\phi\cos^3\phi\big)\,d\phi\,d\theta\\ &= R^4\big(\Big(\frac{3}{8} - \frac{1}{2}\cos 2\phi + \frac{1}{8}\cos 4\phi\Big)\big)(\cos^3\theta + \sin^3\theta) + \sin\phi\cos^3\phi\,d\phi\,d\theta\end{aligned}$$

and our flux integral is

$$\begin{aligned}&\iint_{\mathfrak{S}}\vec{F}\cdot d\vec{\mathcal{S}}\\ &= \int_0^{2\pi}\int_0^{\pi} R^4\big(\Big(\frac{3}{8} - \frac{1}{2}\cos 2\phi + \frac{1}{8}\cos 4\phi\Big)\big)(\cos^3\theta + \sin^3\theta) + \sin\phi\cos^{3\phi}\ d\phi\,d\theta\\ &= R^4\int_0^{2\pi}\big(\Big(\frac{3}{8}\theta - \frac{1}{4}\sin 2\phi + \frac{1}{32}\sin 4\phi\Big)(\cos^3\theta + \sin^3\theta) - \frac{1}{4}\cos^4\phi\big)_{\phi=0}^{\pi}\,d\phi\,d\theta\\ &\qquad = R^4\int_0^{2\pi}\Big(\frac{3\pi}{8}\Big)(\cos^3\theta + \sin^3\theta)\cos^3 + \sin^3)\,d\theta\\ &\qquad = \frac{3\pi}{8}R^4\int_0^{2\pi}\left((1 - \sin^2\theta)\cos\theta + (1 - \sin^2\theta)\cos\theta\right)d\theta\\ &\qquad\qquad = \frac{3\pi}{8}R^4\left(\sin\theta - \frac{1}{3}\sin^3\theta - \cos\theta + \frac{1}{3}\cos^3\theta\right)_0^{2\pi} = 0.\end{aligned}$$

Now if we use the Divergence Theorem instead, we see that the divergence of our vector field is $\operatorname{div}\vec{F} = 2x + 2y + 2z$ so by the Divergence Theorem $\iint_{\mathfrak{S}}\vec{F}\cdot d\vec{\mathcal{S}} = \iiint_{\mathfrak{D}}\operatorname{div}\vec{F}\ dV = \iiint_{\mathfrak{D}} 2(x + y + z)\ dV$, which is easier to do in spherical coordinates:

$$\begin{aligned}
&\int_0^{2\pi}\int_0^{\pi}\int_0^{R} 2(\rho\sin\phi\cos\theta+\rho\sin\phi\sin\theta+\rho\cos\phi)\rho^2\sin\phi\,d\rho\,d\phi\,d\theta\\
&\quad=\int_0^{2\pi}\int_0^{\pi}\int_0^{R} 2\rho^3\sin\phi(\sin\phi\cos\theta+\sin\phi\sin\theta+\cos\phi)\,d\rho\,d\phi\,d\theta\\
&\quad=\int_0^{2\pi}\int_0^{\pi}\left(\frac{\rho^4}{4}\right)_0^R\sin\phi(\sin\phi\cos\theta+\sin\phi\sin\theta+\cos\phi)\,d\rho\,d\phi\,d\theta\\
&\quad=\frac{R^4}{4}\int_0^{2\pi}\int_0^{\pi}\big(\sin^2\phi(\cos\theta+\sin\theta)+\sin\phi\cos\phi\big)\,d\phi\,d\theta\\
&\quad=\frac{R^4}{4}\int_0^{2\pi}\int_0^{\pi}\big(\frac{1}{2}(1-\cos 2\phi)(\cos\theta+\sin\theta)+\sin\phi\cos\phi\big)\,d\theta\\
&\quad=\frac{R^4}{4}\int_0^{2\pi}\Big(\Big(\frac{\phi}{2}-\frac{1}{2}\sin 2\phi\Big)(\cos\theta+\sin\theta)+\frac{1}{2}\sin^2\phi\Big)_{\phi=0}^{\pi}\,d\theta\\
&\quad=\frac{R^4}{4}\int_0^{2\pi}\big(\frac{\pi}{2}(\cos\theta+\sin\theta)+0\big)\,d\theta\\
&\quad=\frac{\pi R^4}{8}(\sin\theta-\cos\theta)_0^{2\pi}=0.
\end{aligned}$$

We note in passing that this triple integral could have been predicted to equal zero on the basis of symmetry considerations. Recall that the integral of an odd function of one real variable $f(t)$ (*i.e.*, if $f(-t)=-f(t)$) over a symmetric interval $[-a,a]$ is zero. We call a region $\mathfrak{D}\subset\mathbb{R}^3$ **symmetric in** z if it is unchanged by reflection across the xy-plane, that is, if whenever the point (x,y,z) belongs to $\mathfrak{D}$, so does $(x,y,-z)$. (The adaptation of this definition to symmetry in x or in y is left to you in Exercise 10.) We say that a function $f(x,y,z)$ is **odd in** z (*resp.* **even in** z) if reversing the sign of z but leaving x and y unchanged reverses the sign of f (*resp.* does not change f): for odd, this means $f(x,y,-z)=-f(x,y,z)$ while for even it means $f(x,y,-z)=f(x,y,z)$.

Remark 5.8.7. *If $f(x,y,z)$ is odd in z and $\mathfrak{D}$ is z-regular and symmetric in z, then*

$$\iiint_{\mathfrak{D}} f(x,y,z)\;dV=0.$$

(To see this, just set up the triple integral and look at the innermost integral.)

Recall that one of the useful consequences of the Fundamental Theorem for Line Integrals was that the line integral of a conservative vector field depends only on the endpoints of the curve, not on the curve itself; more generally, if the curl of a vector field is zero in a region, then the line integral of the field over a curve is not changed if we deform it within that region, holding the endpoints fixed. A similar use can be made of the Divergence Test. We illustrate with an example.

Let us find the flux integral over $\mathfrak{S}$ the upper hemisphere $z=\sqrt{1-x^2-y^2}$, oriented up, of the vector field $\vec{F}(x,y,z)=(1+z)e^{y^2}\vec{\imath}-(z+1)e^{x^2}\vec{\jmath}+(x^2+y^2)\vec{k}$. Whether we think of the hemisphere as the graph of a function or parametrize it using spherical coordinates, the terms involving exponentials of squares are serious trouble. However, note that this vector field is divergence-free:

$$\operatorname{div}\vec{F} = \frac{\partial}{\partial x}\left[(1+z)e^{y^2}\right] + \frac{\partial}{\partial y}\left[(z+1)e^{x^2}\right] + \frac{\partial}{\partial z}\left[(x^2+y^2)\right] = 0.$$

Thus, if we consider the half-ball $\mathfrak{D}$ bounded above by the hemisphere and below by the unit disc in the xy-plane, the Divergence Theorem tells us that

$$\iint_{\partial\mathfrak{D}} \vec{F} \cdot d\vec{\mathcal{S}} = \iiint_{\mathfrak{D}} \operatorname{div}\vec{F}\, dV = 0.$$

Now, the boundary of $\mathfrak{D}$ consists of two parts: the hemisphere, $\mathfrak{S}$, and the disc, $\mathcal{D}$. The outward orientation on $\partial\mathfrak{D}$ means an upward orientation on the hemisphere $\mathfrak{S}$, but a *downward* orientation on the disc $\mathcal{D}$. Thus, the flux integral over the whole boundary equals the flux integral over the upward-oriented hemisphere, plus the flux integral over the downward-oriented disc—which is to say, *minus* the flux over the *upward*-oriented disc. Since the difference of the two upward-oriented discs equals zero, they are equal. Thus

$$\iint_{\mathfrak{S}} \vec{F} \cdot d\vec{\mathcal{S}} = \iint_{\mathcal{D}} \vec{F} \cdot d\vec{\mathcal{S}}.$$

But on $\mathcal{D}$, $d\vec{\mathcal{S}} = \vec{k}\, dA$, so substituting $z = 0$ we see that the vector field on the disc is $\vec{F}(x,y) = e^{y^2}\vec{\imath} - e^{x^2}\vec{\jmath} + (x^2+y^2)\vec{k}$ and $\vec{F} \cdot d\vec{\mathcal{S}} = (x^2+y^2)\, dA$. This is easy to integrate, especially when we use polar coordinates:

$$\iint_{\mathcal{D}} \vec{F} \cdot d\vec{\mathcal{S}} = \int_0^{2\pi} \int_0^1 (r^2)(r\, dr\, d\theta) = \int_0^{2\pi} (\frac{r^4}{4})\, d\theta = \int_0^{2\pi} \frac{1}{4}\, d\theta = \frac{\pi}{2}.$$

Exercises for § 5.8

Answers to Exercises 1a, 2a, 3a, and 5a are given in Appendix A.13.

Practice problems:

(1) Use Green's Theorem to calculate the integral $\int_{\mathcal{C}} \vec{F} \cdot \vec{N}\, d\mathfrak{s}$, where $\vec{N}$ is the outward unit normal and $\mathcal{C}$ is the ellipse $x^2 + 4y^2 = 4$, traversed counterclockwise:
  - (a) $\vec{F}(x,y) = x\vec{\imath} + y\vec{\jmath}$
  - (b) $\vec{F}(x,y) = y\vec{\imath} + x\vec{\jmath}$
  - (c) $\vec{F}(x,y) = x^2\vec{\imath} + y^2\vec{\jmath}$
  - (d) $\vec{F}(x,y) = x^3\vec{\imath} + y^3\vec{\jmath}$

(2) Find the flux integral $\iint_{\mathfrak{S}} \vec{F} \cdot d\vec{\mathcal{S}}$, where $\mathfrak{S}$ is the unit sphere oriented outward, for each vector field below:
  - (a) $\vec{F}(x,y,z) = (x+y^2)\vec{\imath} + (y-z^2)\vec{\jmath} + (x+z)\vec{k}$
  - (b) $\vec{F}(x,y,z) = (x^3+y^3)\vec{\imath} + (y^3+z^3)\vec{\jmath} + (z^3-x^3)\vec{k}$
  - (c) $\vec{F}(x,y,z) = 2xz\vec{\imath} + y^2\vec{\jmath} + xz\vec{k}$
  - (d) $\vec{F}(x,y,z) = 3xz^2\vec{\imath} + y^3\vec{\jmath} + 3x^2z\vec{k}$

(3) For each vector field $\vec{F}$ below, find the flux integral $\iint_{\mathfrak{S}} \vec{F} \cdot d\vec{\mathcal{S}}$, where $\mathfrak{S}$ is the boundary of the unit cube $[0,1] \times [0,1] \times [0,1]$, oriented outward, in two different ways: (*i*) directly (you will need to integrate over each face separately and then add up the results) and (*ii*) using the Divergence Theorem.
  - (a) $\vec{F}(x,y,z) = x\vec{\imath} + y\vec{\jmath} + z\vec{k}$
  - (b) $\vec{F}(x,y,z) = \vec{\imath} + \vec{\jmath} + \vec{k}$
  - (c) $\vec{F}(x,y,z) = x^2\vec{\imath} + y^2\vec{\jmath} + z^2\vec{k}$

-256- - -* ! - 7 3 9 ? ! "
22 1"
(' B/ < %' 1 & B

Exercises 11.4

1. Use geometric software to produce Dürer's construction of the regular pentagon.

2. The following set of instructions is for constructing a pentagon given a side AB of length d.

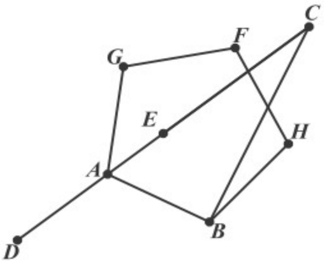

Figure 11.32. Exercise 11.4.2: Constructing a regular pentagon given side AB

- Construct CB perpendicular to AB such that $CB = 2d$.
- Construct segment CA and extend ray CA to a point D such that $AD = d$, where A lies between C and D.
- Bisect CD at point E.
- Let F be the intersection of the two circles centered at A and B, each with radius CE.
- Let G be the intersection of circles with centers A and F and radius AB, and let H be the intersection of circles with centers B and F and radius AB.
- Form pentagon ⬠$ABHFG$ by connecting vertices A, B, H, F and G.

(a) Use geometric software to produce this construction of the regular pentagon.

(b) Prove that if $AB = 1$, then $CE = \dfrac{1+\sqrt{5}}{2}$, and hence, $\triangle FAB$ is a golden triangle.

(c) Use part (b) to prove that ⬠$ABHFG$ is a regular pentagon.

3. Prove that the construction given in Proposition IV.13 works for any regular n-gon.

4. Prove that the construction given in Proposition IV.14 works for any regular n-gon.

5. Prove that the construction given above using central angles of 72° produces a regular pentagon.

6. Prove Lemma 11.13.

11.5 Constructing regular polygons

Euclid constructed both the equilateral triangle and the square in Book I. We had to wait until the eleventh proposition of the fourth book for the construction of the next regular polygon, the more complex regular pentagon. Its construction and proof relies on results from all three of the previous books. In the two remaining propositions, Euclid constructs a regular hexagon and then finishes Book IV with a regular 15-gon. In addition to these two constructions, we consider the more general question of which regular n-gons *can* be constructed using only an unmarked straightedge and compass.

In Proposition IV.15, Euclid inscribes a regular hexagon in a circle. We give a modified version of his construction and leave the verification step to the reader.

Proposition IV.15. *In a given circle to inscribe an equilateral and equiangular hexagon.*

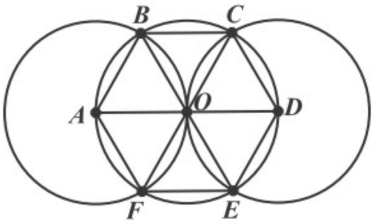

Figure 11.33. Regular hexagon

Construction: Given a circle with center O and point A, construct the diameter AD. Construct two circles, one with center A, the other with center D, and both with radius AO. Label the intersections of these circles with our original circle B, C, E and F so that the letters are in order from A either clockwise or counter-clockwise.

Claim: The hexagon $\hexagon ABCDEF$ is a regular hexagon. The justification is left to the reader. □

Another way to construct a regular hexagon is to notice that it has twice as many sides as an equilateral triangle, which is itself constructible. So, if we circumscribe a circle about the equilateral triangle, bisect each of the central angles and mark the intersection of the angle bisectors with the circle, then we'll have the six vertices of a regular hexagon marked on the circle. We will exploit this method to construct a regular $2n$-gon from any constructible n-gon: start with a constructible regular n-gon, circumscribe a circle around it, bisect each of the central angles and add the points where the angle bisectors meet the circle to the original set of vertices. Alternatively, we could bisect each of the n arcs into which the regular n-gon divides the circle. Either way, we now have $2n$ vertices which, when connected, produce a regular $2n$-gon. While Euclid did not include the following proposition in Book IV, his far more complicated constructions leads us to believe that he chose not to include it. We leave the proof to the reader.

Theorem 11.14. *If a regular n-gon is constructible, then so is a regular $2n$-gon.*

Using this doubling method on an inscribed equilateral triangle, square or regular pentagon, we are thus able to construct regular n-gons for $n = 6, 8, 10, 12, 16, 20, 32, 40, \ldots$.

For the last proposition in Book IV, Euclid constructs a regular 15-gon, also called a regular **pentadecagon**. We give a slightly revised version of his construction.

Proposition IV.16. *In a given circle to inscribe a fifteen-angled figure which shall be both equilateral and equiangular.*

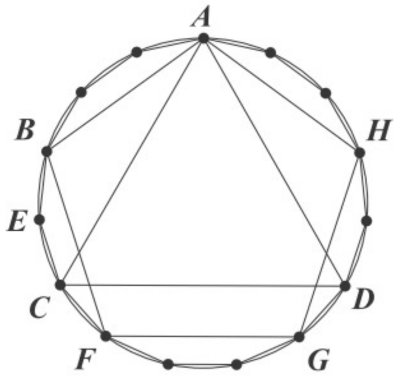

Figure 11.34. Proposition IV.16

Construction: Given a circle with a point A on it, inscribe an equilateral triangle $\triangle ACD$ with vertex A. Also inscribe a regular pentagon ⬠ $ABFGH$. The triangle divides the circumference into three arcs, each one-third of the whole, while the pentagon divides the circumference into five arcs, each one-fifth of the whole. The arc that lies between B and C must be two-fifteenths of the whole since $\frac{1}{3} - \frac{1}{5} = \frac{2}{15}$. If we bisect this arc at E, then $\widehat{BE}$ will be exactly one-fifteenth of the whole. Copying this arc around the circle, we construct a regular 15-gon, as desired. □

At first glance, the construction technique employed in the proof appears fruitful, as if it would yield other new regular n-gons from those that we already know to be constructible. Alas, it does not. (Why not? What makes the regular 3- and 5-gons special?) Euclid ends his regular polygon constructions with the 15-gon, but the hunt for other constructible regular polygons (besides those obtained by doubling) continued for nearly 2000 years until a 19-year-old finally found one. In 1796, a young Gauss constructed the **heptadecagon**, or 17-gon. Better still, he gave the formula to completely classify which regular n-gons are constructible. The reason the resolution to this problem took so long is that it required tools outside of geometry, specifically, tools from the new mathematical areas of abstract algebra and complex numbers. With those tools, Gauss completely classified the constructibility of regular n-gons as given in Theorem 11.15.

Theorem 11.15. [Gauss's Theorem]. *A regular n-gon is constructible if and only if $n = 2^k p_1 p_2 p_3 \ldots p_m$ where $n \geq 3$, $k \geq 0$, and p_i are distinct primes of the form $2^{2^j} + 1$.*

Primes of the form $2^{2^j} + 1$ are called **Fermat primes**, named for Pierre de Fermat (1601–1665). The known Fermat primes are:

$$\begin{gathered} 2^{2^0} + 1 = 3 \\ 2^{2^1} + 1 = 5 \\ 2^{2^2} + 1 = 17 \\ 2^{2^3} + 1 = 257 \\ 2^{2^4} + 1 = 65537. \end{gathered}$$

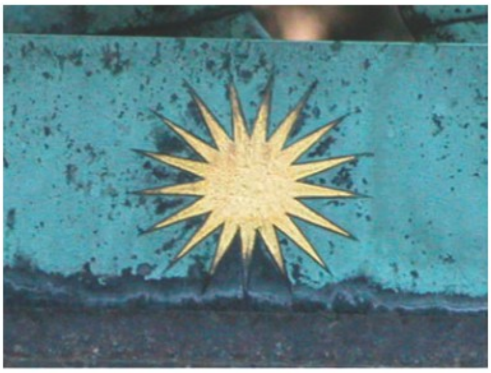

Figure 11.35. Close-up of 17-point star on monument to Gauss, Brunswick, Germany

Euler famously showed that $2^{2^5} + 1 = 4,294,967,297$ is not prime. It is known that $2^{2^j} + 1$ is not prime for $5 \leq j \leq 32$ and a smattering of other values for larger j. However, it is **unknown** whether there are only five primes of this form.[1] Therefore, as of this edition, we have the following:

A regular n-gon is constructible if $n = 2^k \cdot 3^{b_1} \cdot 5^{b_2} \cdot 17^{b_3} \cdot 257^{b_4} \cdot 65537^{b_5}$ *where* $n \geq 3$*,* $k \geq 0$*, and* $b_i \in \{0, 1\}$ *for all i.*

If one day it is shown that there are no other Fermat primes, then this implication becomes an "if and only if."

As for the constructions of the regular 257-gon and the regular 65537-gon, Gauss left them to two other German mathematicians. Friedrich Julius Richelot (1808–1875) constructed the regular 257-gon in 1831, and Johann Gustav Hermes (1846–1912) spent ten years on the regular 65537-gon, finishing in 1894. The manuscript detailing this impossible-to-imagine construction sits in a large box at the University of Göttingen. For now, we are left to wonder why it is impossible to construct a regular n-gon when n does not take this form. For example, why can't we construct a 7-gon or a 9-gon? We will discuss this, as well as the three famous construction problems of antiquity, in Chapter 16.

Surely owing to its personal significance, Gauss remarked that the heptadecagon was a result worthy of being on his tombstone. His desire to memorialize his great accomplishment puts him in fine company. The greatest mathematician of antiquity and the subject of the following section, Archimedes of Syracuse (287–212 BCE), requested a representation of his most significant accomplishment, a cylinder circumscribing a sphere, on his tomb. The Roman philosopher Cicero (106–43 BCE) writes of finding a figure of a sphere in a cylinder when in search of Archimedes' grave [**24**]. On Gauss's simple grave in Göttingen there is no such polygon, but a 17-point star does grace the bottom of the bronze statue that sits in his honor in his native town of Brunswick. In Figure 11.36 we include a similar homage to Gauss in the form of a 17-point star made by Swiss artist Eugen Jost and appearing in a lovely book on art and mathematics, *Beautiful Geometry*.

---

[1] Current information on Fermat numbers can be found at *www.fermatsearch.org/news.html*.

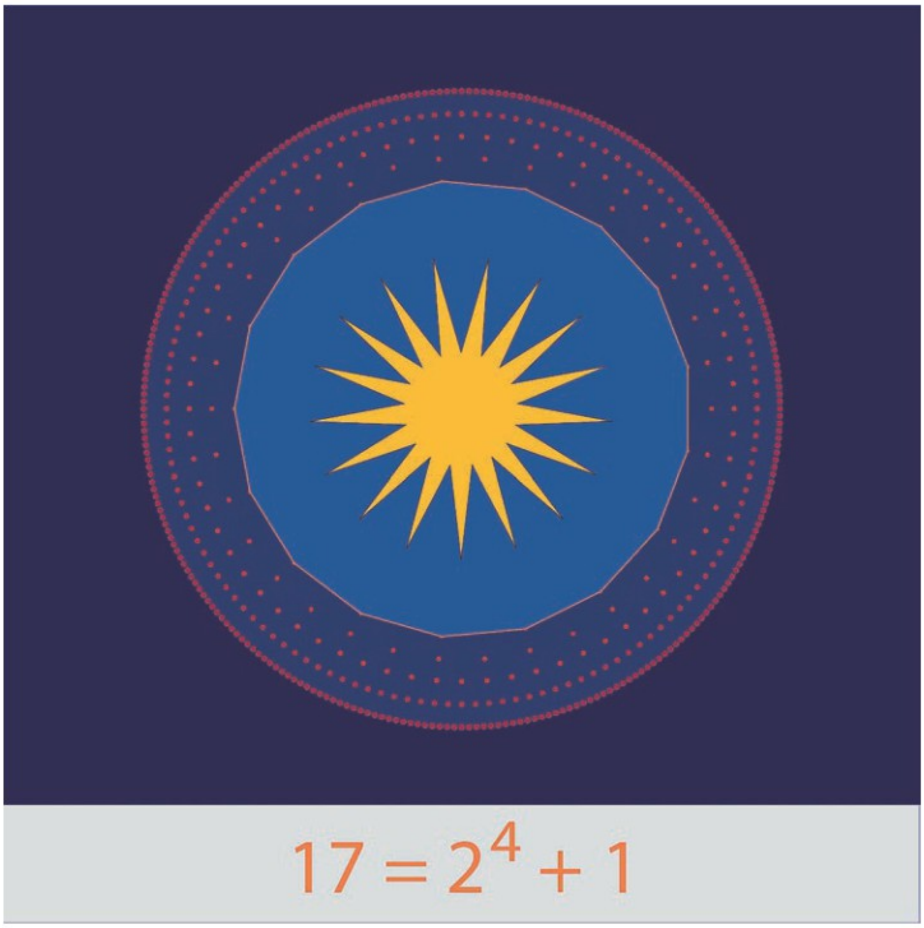

Figure 11.36. Homage to Carl Friedrich Gauss by Eugen Jost (2014) [86]

Exercises 11.5

1. Prove that the construction given for Proposition IV.15 is valid.

2. Construct a regular $2n$-gon given a regular n-gon. Prove that this construction is valid.

3. Given a regular m-gon that is constructible and a regular n-gon that is constructible, prove that if m and n are relatively prime ($\gcd(m, n) = 1$), then a regular mn-gon is also constructible.

4. Find all n such that $3 \leq n \leq 100$ and the regular n-gon is constructible.

5. Find all n such that $101 \leq n \leq 200$ and the regular n-gon is constructible.

6. Use geometric software to produce the following "quadruple quadrisection" construction of a 17-gon due to Herbert W. Richmond (1863–1948), King's College,

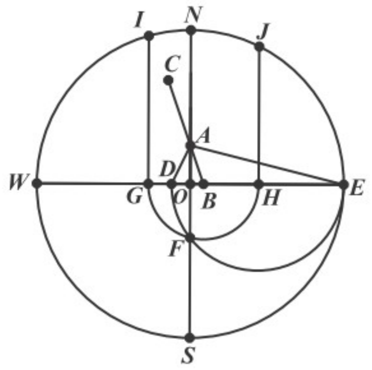

Figure 11.37. Exercise 11.5.6: Richmond's construction of a 17-gon

Cambridge, who gave it in the *Quarterly Journal of Pure and Applied Mathematics* in 1893. To **quadrisect** is to divide something into four equal parts. (Hence, to bisect into halves, and then, bisect each half.)

- Quadrisect the perimeter of the circle of center O by points N, E, S, W.
- Quadrisect the radius ON by the point A where A is closest to O.
- Quadrisect the angle $\angle OAE$ by the line AB (where B lies on OE and is closest to O). See Figure 11.37.
- Form line BAC.
- Quadrisect the straight angle BAC in the direction of W by the line AD, where D is on radius OW, and angle $\angle BAD$ is half of a right angle.
- Construct semicircle DFE, cutting OS at F.
- Construct semicircle GFH, centered at B, cutting OW at G and OE at H.
- Construct $GI \perp WE$ where I is on circle WNE in the 2nd quadrant.
- Construct $HJ \perp WE$ where J is on circle WNE in the 1st quadrant.
- Since arc $\widehat{IJ}$ is two seventeenths of the circumference of the circle, bisecting arc $\widehat{IJ}$ at K produces vertices I, K and J of the regular heptadecagon. Copy arc $\widehat{IK}$ around circle $ENWS$ to construct the other vertices of the heptadecagon.

11.6 The area of a circle & π

At the end of Chapter 10, we gave a decidedly modern proof using limits to show that the ratio of any circle's circumference to its diameter is a constant, and noted that this constant was named π in the eighteenth century. Euclid does not include a proof that this ratio is constant in the *Elements*, which is curious since there is evidence that it was known long before Euclid took brush to papyrus. There are estimates for π in the Bible, in ancient Egyptian papyri, and in ancient Chinese books, all of which predate Euclid. Certainly, a calculation of an estimate for π would have been out of place in the *Elements* for it is not a book that shows any concern for absolute measurement. Most of its propositions are relationships between geometric objects. So, while the lack of an estimate is unsurprising, we are left to wonder why he does not show that all circles satisfy this ratio. On the other hand, Euclid *does* present a different ratio that is satisfied by all circles. With the second proposition of his twelfth book, he proves the following: *Circles are to one another as the squares on the diameters.* Here, given any two circles with areas, $Area(circle_1)$ and $Area(circle_2)$, and diameters, $diameter_1$ and $diameter_2$, respectively, we have

$$XII.2 \qquad \frac{Area(circle_1)}{Area(circle_2)} = \frac{diameter_1^2}{diameter_2^2}.$$

His proof relies on inscribing a sequence of regular polygons of an increasing number of sides. We will not prove this particular result, but we will revisit it after looking at a related proposition from Archimedes, a mathematician who followed closely on the heels of Euclid and is widely regarded as the greatest of antiquity. We present two results from Archimedes' treatise, *Measurement of the Circle.* The first is equivalent to our modern formula for the area of the circle, and the second is an estimate for π. Both of Archimedes' results rely on the same method Euclid uses in XII.2, inscribing and circumscribing regular polygons in and around a circle.

The most important document containing the works of Archimedes was only rediscovered in Constantinople in 1906. Called the *Archimedes Palimpsest*, it contains seven treatises, including *Measurement of the Circle*, which contains the two propositions given in this section, and, notably, the only surviving copy *On Floating Bodies.* A palimpsest is a manuscript that has been erased so that the precious writing material, in this case goatskin parchment, can be reused. Disguised as a book of prayers since the thirteenth century, it laid undiscovered for hundreds of years in a monastery library. It was sold to a private collector in the 1920s, and resold at auction in 1998 at Christie's Auction House in New York City, where it sold to an anonymous buyer for over two million dollars. It has been an object of study at the Walters Art Museum in Baltimore, Maryland, since 1999.

Archimedes' Proposition 1. *The area of a circle is equal to the area of a right triangle in which one of the legs is equal to the radius and the other to the circumference.*

Restating this as a formula, we have $Area = \frac{1}{2}r \cdot Circumference$, where r is the radius of the circle. Since $Circumference = 2\pi r$, we have the familiar formula for area of a circle, $Area = \pi r^2$. Archimedes' proof (as does Euclid's XII.2) employs the **method of exhaustion**, a technique attributed to Eudoxus (408–355 BCE), a Greek mathematician who predates Euclid. Anyone familiar with integral calculus will recognize this method, which entails exhausting the difference between a known area and an unknown area. Here, the area of the circle is approximated by systematically inscribing regular polygons of an increasing number of sides in the circle. As the number of sides increases, the easily computed area of the polygon approaches the area of the circle from below, so that the difference between the unknown area of the circle and the known area of the polygon is less than any given quantity, no matter how small. At heart, this is the same limit method we saw at play in the proof that the ratio of any circle's circumference to its diameter is constant in Theorem 10.3, but here we are working with the two-dimensional area of the inscribed polygon instead of the one-dimensional perimeter. It is the idea that this process can continue forever, theoretically, and thus, that the difference between the unknown and the known can be exhausted, that assuredly places Greek mathematics as a forerunner to calculus.

As a visual aid, Figure 11.38 shows a systematic progression of regular n-gons inscribed in a circle. At this size, it is difficult to distinguish the circle from the regular 24-gon. Clearly, the area of the regular n-gon is a good approximation for the area of the circle when n is large.

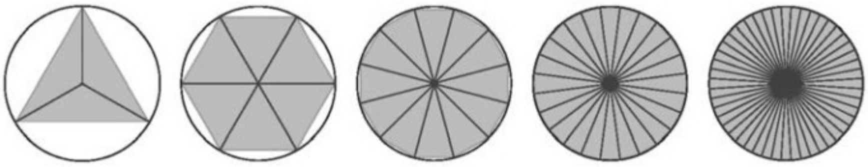

Figure 11.38. Inscribed regular 3-gon, 6-gon, 12-gon, 24-gon and 48-gon

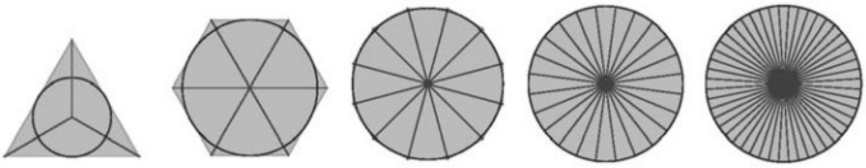

Figure 11.39. Circumscribed regular 3-gon, 6-gon, 12-gon, 24-gon and 48-gon

Archimedes takes Eudoxus' method of exhaustion one step further by approaching the area of the circle from above with circumscribed regular polygons, knowing that the actual area of the circle lies between these bounds. In fact, he uses a double reductio ad absurdum argument to show that the circle and triangle have the equal area, arguing that the other possibilities, when either is bigger than the other, lead to a contradiction. In approaching the area from above and below, those familiar with the work of Isaac Newton (1642–1727) and Gottfried Leibniz (1646–1716) will no doubt see the seeds of integral calculus in this method and want to turn this into a limit problem. We do just that, giving an updated proof of Archimedes' first proposition which relies on the limit concept. First, we need a new definition and a formula for the area of a regular n-gon.

Definition 11.16. The **apothem** of a regular polygon is a line segment from the center to the midpoint of one of its sides.

Lemma 11.17. *A regular polygon with perimeter P and apothem h has an area of* $\frac{1}{2}h \cdot P$.

Proof of Lemma. Suppose the regular polygon has n sides, each of length s. Join each vertex to the center, O. The polygon is composed of n congruent triangles, each with area $\frac{1}{2}hs$. Hence, the area of the polygon is $n \cdot (\frac{1}{2}hs) = \frac{1}{2}h \cdot (ns) = \frac{1}{2}h \cdot P$, as desired. □

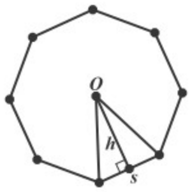

Figure 11.40. Area of a regular polygon

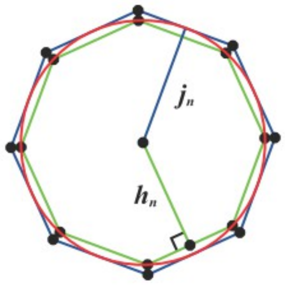

Figure 11.41. Archimedes' Proposition 1: Area of a circle

Proof of Archimedes' Proposition 1. Let *Area(circle)* represent the area of the given circle, B_n be the area of an inscribed regular polygon with n sides, and D_n be the area of a circumscribed regular polygon with n sides. Clearly, $B_n \leq \mathit{Area(circle)} \leq D_n$ for all $n \geq 3$. Let h_n be the apothem of B_n, and P_n be the perimeter of B_n. Likewise, let j_n be the apothem and Q_n the perimeter of D_n. By Lemma 11.17, we have

$$\frac{1}{2}h_n \cdot P_n \;\leq\; \mathit{Area(circle)} \;\leq\; \frac{1}{2}j_n \cdot Q_n.$$

Taking the limit as n approaches infinity, we have

$$\lim_{n\to\infty} \frac{1}{2}h_n \cdot P_n \;\leq\; \mathit{Area(circle)} \;\leq\; \lim_{n\to\infty} \frac{1}{2}j_n \cdot Q_n.$$

But clearly, $\lim\limits_{n\to\infty} h_n = \lim\limits_{n\to\infty} j_n = r$, and $\lim\limits_{n\to\infty} P_n = \lim\limits_{n\to\infty} Q_n = \mathit{Circumference}$. Thus,

$$\frac{1}{2}r \cdot \mathit{Circumference} \;\leq\; \mathit{Area(circle)} \;\leq\; \frac{1}{2}r \cdot \mathit{Circumference},$$

and the area of the circle is $\mathit{Area(circle)} = \frac{1}{2}r \cdot \mathit{Circumference}$. ☐

In light of this proposition, let's revisit Euclid's Proposition XII.2. A notable difference between these results is that Euclid relates geometric figures of the same dimension, whereas Archimedes relates figures of differing dimensions. Consequently, in combination, these two results tell us something very interesting. Neither result relies on the constancy of the ratio of any circle's circumference to its diameter, and together, they provide a proof of this result.

$$\begin{aligned}\frac{\mathit{Circumference}_1}{\mathit{Circumference}_2} &= \frac{\frac{2}{r_1}\cdot \mathit{Area}_1}{\frac{2}{r_2}\cdot \mathit{Area}_2} \quad \text{by Archimedes Prop. 1}\\ &= \frac{2r_2}{2r_1}\cdot\frac{\mathit{Area}_1}{\mathit{Area}_2} \quad \text{by rearranging}\\ &= \frac{d_2}{d_1}\cdot\frac{d_1^2}{d_2^2} \quad \text{by Euclid XII.2}\\ &= \frac{d_1}{d_2}.\end{aligned}$$

Therefore, we have another proof that the ratio of any circle's circumference to its diameter is constant, that is

$$\frac{\mathit{Circumference}_1}{\mathit{diameter}_1} = \frac{\mathit{Circumference}_2}{\mathit{diameter}_2}.$$

For our second result, we consider Archimedes' third proposition in which he provides an estimate for this constant ratio, π.

Archimedes' Proposition 3. *The ratio of the circumference of any circle to its diameter is less than* $3\frac{1}{7}$ *but greater than* $3\frac{10}{71}$.

Archimedes method entails bounding this quantity between two fractions with the use of an inscribed and circumscribed regular 96-gon. He begins by inscribing a regular hexagon in the circle. He then proceeds to double the number of sides, successively producing an inscribed regular 12, 24, 48 and 96-gon. He relates the circumference of the inscribed $2n$-gon to that of the n-gon, and thus, generates a recursive formula to estimate a lower bound for π. Of course, he does all of this without the benefit of algebra or the ease of calculation afforded by our decimal number system. He then circumscribes a regular hexagon around the circle and, once again, doubles the number of sides, successively producing a circumscribed regular 12, 24, 48 and 96-gon. This gives him an upper bound for π. We present a revised argument, and where Archimedes used fractions to estimate square roots, we take full advantage of modern technology when giving our lower and upper bounds.

We begin by inscribing a regular hexagon in a circle of radius 1. Notice that the perimeter of the regular 6-sided polygon, $P_6 = 6$, is an underestimate for the circumference, 2π. More generally, if a_n denotes the length of the side of a regular n-sided polygon inscribed in a circle of radius 1, then its perimeter, P_n, is given by $P_n = n \cdot a_n$. Since the perimeter of any polygon inscribed in the circle is less than the circumference of the circle, that is $P_n < 2\pi$, we have

$$\pi \geq \frac{n \cdot a_n}{2}.$$

We now give the following lemma.

Lemma 11.18. *If a_n denotes the length of the side of a regular n-sided polygon inscribed in a circle of radius* 1*, then*

$$a_{2n} = \sqrt{2 - \sqrt{4 - a_n^2}}.$$

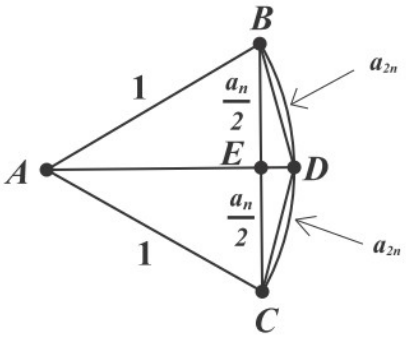

Figure 11.42. Sides of n-gon and $2n$-gon

Proof of Lemma. Let A be the center of the given circle of radius 1, $BC = a_n$ and $CD = a_{2n}$, as shown in Figure 11.42. Note that AD bisects arc $\widehat{BC}$, meeting BC at E. Since angle $\angle AEB$ is a right angle (by proof of Theorem 11.14), then by the Pythagorean Theorem we have $1 = \left(\frac{a_n}{2}\right)^2 + AE^2$, and $a_{2n}^2 = \left(\frac{a_n}{2}\right)^2 + DE^2$. Since $DE = 1 - AE$, this gives

$$\begin{aligned} a_{2n}^2 &= \left(\tfrac{a_n}{2}\right)^2 + (1-AE)^2 \\ &= \left(\tfrac{a_n}{2}\right)^2 + \left(1 - \sqrt{1-\left(\tfrac{a_n}{2}\right)^2}\right)^2 \\ &= \left(\tfrac{a_n}{2}\right)^2 + 1 - 2\sqrt{1-\left(\tfrac{a_n}{2}\right)^2} + 1 - \left(\tfrac{a_n}{2}\right)^2 \\ &= 2 - \sqrt{4-a_n^2}. \end{aligned}$$

□

Substituting the appropriate values for n gives the following:

$$\begin{aligned} a_{12} &= \sqrt{2-\sqrt{3}} \approx 0.5176381 \\ a_{24} &= \sqrt{2-\sqrt{2+\sqrt{3}}} \approx 0.2610524 \\ a_{48} &= \sqrt{2-\sqrt{2+\sqrt{2+\sqrt{3}}}} \approx 0.1308063 \\ a_{96} &= \sqrt{2-\sqrt{2+\sqrt{2+\sqrt{2+\sqrt{3}}}}} \approx 0.0654382. \end{aligned}$$

Thus,

$$\pi \geq 48a_{96} \approx 3.14103 \geq 3\frac{10}{71}.$$

Let's consider the circumscribed polygons that produce the upper bound for π. Suppose b_n denotes the length of the side of a regular n-sided polygon circumscribed about a circle of radius 1. We have two possible plans of attack for the circumscribed polygons. The first is to use the fact that $b_6 = \frac{2}{\sqrt{3}}$, and then prove the following recurrence relationship.

Lemma 11.19. *If b_n denotes the length of the side of a regular n-gon circumscribed about a circle of radius* 1*, then*

$$b_{2n} = \frac{2\left(\sqrt{4+b_n^2}-2\right)}{b_n}.$$

An alternative choice is to establish the following relationship between a_n and b_n.

Lemma 11.20. *If a_n and b_n denote the lengths of the sides of regular n-gons inscribed in and circumscribed about a circle of radius* 1*, respectively, then*

$$b_n = \frac{2a_n}{\sqrt{4-a_n^2}}.$$

A representative diagram for regular pentagons is shown in Figure 11.43. We leave both proofs to the reader, but note that $b_{96} \approx 0.0654732$. Hence,

$$\pi \;\le\; 48b_{96} \;\approx\; 3.142715 \;\le\; 3\frac{1}{7}.$$

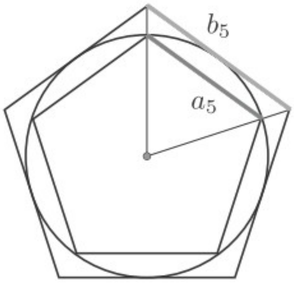

Figure 11.43. Sides of inscribed and circumscribed regular pentagon

Exercises 11.6

1. Prove Lemma 11.19.

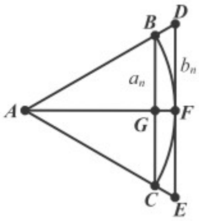

Figure 11.44. Exercise 11.6.2: Sides of inscribed and circumscribed regular n-gons

2. Using Figure 11.44 as a guide, prove Lemma 11.20. Note that the circle of radius 1 is centered at A, $AB = AC = 1$, $BC = a_n$, $DE = b_n$, and DE is tangent to the circle at F. Be sure to justify why $AG \perp BC$, $BG = \frac{a_n}{2}$ and $DF = \frac{b_n}{2}$.

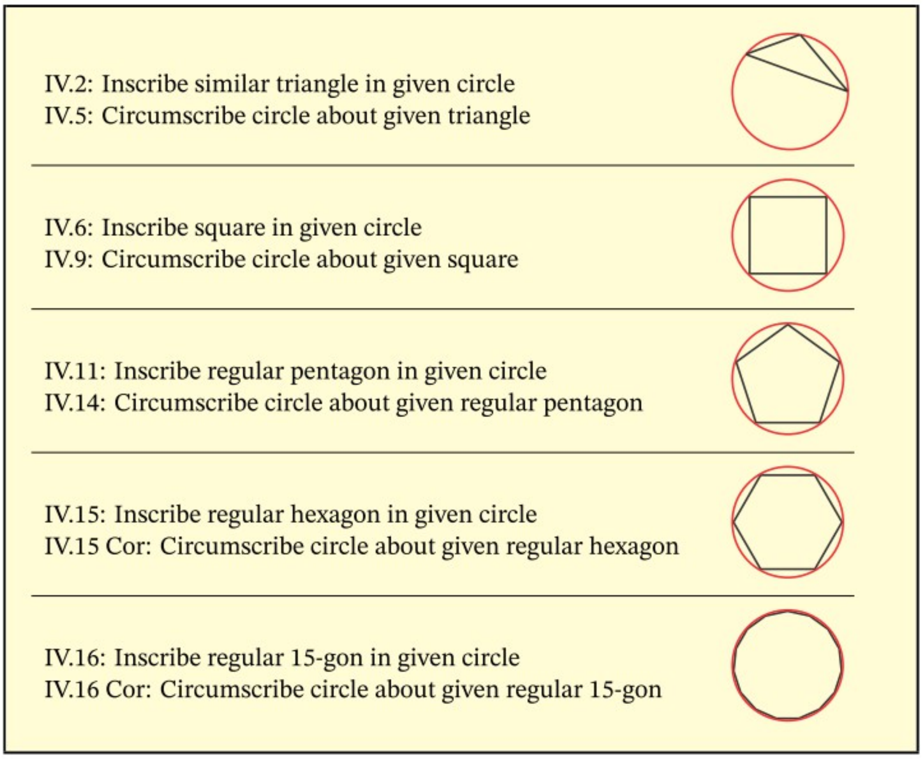

IV.2: Inscribe similar triangle in given circle
IV.5: Circumscribe circle about given triangle

IV.6: Inscribe square in given circle
IV.9: Circumscribe circle about given square

IV.11: Inscribe regular pentagon in given circle
IV.14: Circumscribe circle about given regular pentagon

IV.15: Inscribe regular hexagon in given circle
IV.15 Cor: Circumscribe circle about given regular hexagon

IV.16: Inscribe regular 15-gon in given circle
IV.16 Cor: Circumscribe circle about given regular 15-gon

Figure 11.45. Inscribed polygon or circumscribed circle propositions in Book IV

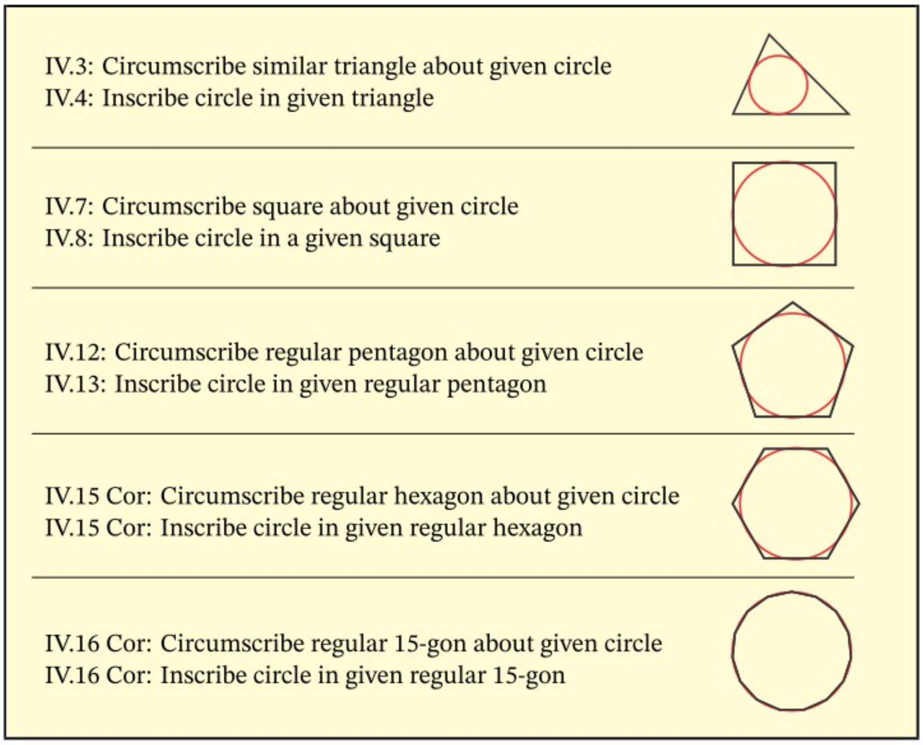
IV.3: Circumscribe similar triangle about given circle
IV.4: Inscribe circle in given triangle
IV.7: Circumscribe square about given circle
IV.8: Inscribe circle in a given square
IV.12: Circumscribe regular pentagon about given circle
IV.13: Inscribe circle in given regular pentagon
IV.15 Cor: Circumscribe regular hexagon about given circle
IV.15 Cor: Inscribe circle in given regular hexagon
IV.16 Cor: Circumscribe regular 15-gon about given circle
IV.16 Cor: Inscribe circle in given regular 15-gon

Figure 11.46. Circumscribed polygon or inscribed circle propositions in Book IV

Minkowski made the first major progress on this problem by proving that any convex polyhedron can be so realized [148].

Pag. 356(366) B. Different geometries

One can study properties of surfaces that are invariant under a bigger group than the orthogonal group.

the There are the special linear group, the conformal group and the projective group each acting on $\mathbb{R}^3$!

For affine geometry important invariants are the concepts of the affine metric (which is the second fundamental form divided by the fourth root of the GAUSSIAN curvature) and the affine normal.

John Loftin discussed how to construct an affine sphere by a holomorphic section of the tricanonical line bundle of a Riemann surface.

- 20211014/ Calculus in 3D Geometry ...!

(%$/

- Calculus in 3D Geometry, Vectors, and Multivariate... pdf

Pág. 321(332) - rodapé 17

Gauss's theorem or the Gauss's-Ostrogradsky Theorem

- Geometry - the Line and the Circle (2018).pdf

Pág. 248(269) Figure 11.34

Proposition IV. 16

In a given circle to inscribe a fifteen-angled figure which shall be both equilateral and equiangular

Theorem 11.15 [Gauss's Theorem]

38 ocurrences

- 20211014/ Krishna's - Differential Equations... pdf

Pág. 183 - Unit-3

Chapter 5

Gauss Hypergeometric Equation.

-256- - -* ! - + !0 4 ! => ! < 3
9 " =" 1"
(% H % !1 $ 7! :31 => !

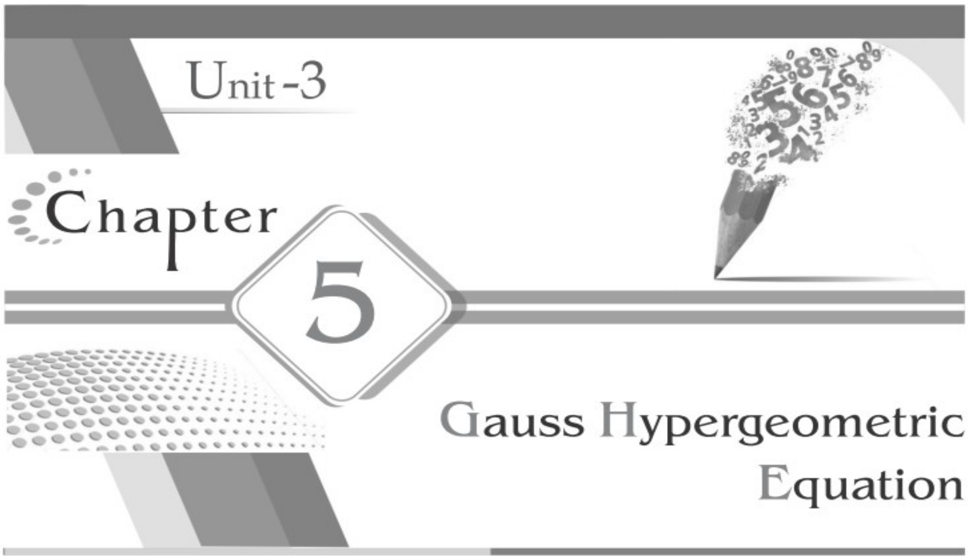

Unit-3

Chapter 5

Gauss Hypergeometric Equation

5.1 The Pochhammer Symbol (Def.)

If n is a positive integer, then the abbreviation

$$(\alpha)_n = \alpha\ (\alpha + 1)\ (\alpha + 2) \ldots (\alpha + n - 1)$$

with $(\alpha)_0 = 1$

is called the **Pochhammer symbol.**

Another Def. We have

$$(\alpha)_n = \alpha\ (\alpha + 1)\ (\alpha + 2) \ldots. (\alpha + n - 1)$$

$$= \frac{1.2.3 \ldots\ldots (\alpha - 1).\alpha\ (\alpha + 1)\ (\alpha + 2) \ldots. (\alpha + n - 1)}{1.2.3. \ldots. (\alpha - 1)}$$

$$= \frac{\Gamma(\alpha + n)}{\Gamma(\alpha)}$$

Thus, **Pochhammer symbol** in the form of gamma function can also be defined as

$$(\alpha)_n = \frac{\Gamma(\alpha + n)}{\Gamma(\alpha)}, n = 0, 1, 2, 3, \ldots\ldots\ldots$$

Note : The Pochhammer Symbol $(\alpha)_n$ is sometimes replaced by Appel's symbol (α, n). But throughout this chapter we shall use Pochhammer symbol.

5.2 Identities Satisfied by Pochhammer Symbol $(\alpha)_n$

1. $(1)_n = n!$
2. $(\alpha + n)(\alpha)_n = \alpha(\alpha+1)_n = (\alpha)_{n+1}$
3. $\binom{-\alpha}{n} = \frac{(-1)^n}{n!}(\alpha)_n$
4. $(\alpha)_{n+1} = \alpha(\alpha+1)_n$
5. $(\alpha)_{n+k} = (\alpha)_k(\alpha+k)_n = (\alpha)_n(\alpha+n)_k$ (Addition formula)
6. $(\alpha)_{k-n} = (-1)^n(\alpha)_k/(1-\alpha-k)_n$
7. $(\alpha)_{-n} = (-1)^n/(1-\alpha)_n,\ n = 1, 2, 3, \ldots.$
8. $(-k)_n = \begin{cases} \frac{(-1)^n k!}{(k-n)!}, & 0 \le n \le k \\ 0, & n > k \end{cases}$

Proof: 1. $(1)_n = 1(1+1)(1+2)\ldots(1+n-1) = 1.2.3\ldots.n = n!$

2. $(\alpha+n)(\alpha)_n = (\alpha)_n.(\alpha+n) = \alpha(\alpha+1)(\alpha+2)\ldots.(\alpha+n-1).(\alpha+n)$

$$= \alpha.\{(\alpha+1)(\alpha+2)\ldots.(\alpha+1+n-1)\} = \alpha(\alpha+1)_n$$

Also $(\alpha+n)(\alpha)_n = \alpha(\alpha+1)(\alpha+2)\ldots(\alpha+n-1).(\alpha+n+1-1) = (\alpha)_{n+1}$

Hence $(\alpha+n)(\alpha)_n = \alpha(\alpha+1)_n = (\alpha)_{n+1}$

3. $$\binom{-\alpha}{n} = \frac{(-\alpha)(-\alpha-1)(-\alpha-2)\ldots(-\alpha-n+1)}{n!}$$

$$= \frac{(-1)^n \alpha(\alpha+1)(\alpha+2)\ldots.(\alpha+n-1)}{n!} = \frac{(-1)^n}{n!}(\alpha)_n$$

4. $(\alpha)_{n+1} = \alpha(\alpha+1)(\alpha+2)\ldots.(\alpha+n+1-1)$

$$= \alpha.(\alpha+1)(\alpha+1+1)\ldots(\alpha+1+n-1) = \alpha.(\alpha+1)_n$$

5. $(\alpha)_{n+k} = \alpha(\alpha+1)(\alpha+2)\ldots(\alpha+k-1)(\alpha+k)\ldots(\alpha+n+k-1)$

$$= \{\alpha(\alpha+1)(\alpha+2)\ldots(\alpha+k-1)\}\{(\alpha+k)(\alpha+k+1)\ldots\ldots(\alpha+k+n-1)\}$$

$$= (\alpha)_k.(\alpha+k)_n \quad \text{if } k < n.$$

Also,

$$(\alpha)_{n+k} = \{\alpha(\alpha+1)(\alpha+2)\ldots(\alpha+n-1)\}\{(\alpha+n)(\alpha+n+1)\ldots.(\alpha+n+k-1)\}$$

$$= (\alpha)_n.(\alpha+n)_k \quad \text{if } k > n$$

Hence $(\alpha)_{n+k} = (\alpha)_k(\alpha+k)_n = (\alpha)_n.(\alpha+n)_k$

6. $(\alpha)_{k-n} = \alpha(\alpha+1)(\alpha+2)\dots(\alpha+k-n-1)$

$$= \frac{\alpha(\alpha+1)(\alpha+2)\dots(\alpha+k-n-1)(\alpha+k-n)(\alpha+k-n+1)\dots(\alpha+k-1)}{(\alpha+k-n)(\alpha+k-n+1)\dots(\alpha+k-1)}$$

$$= \frac{(\alpha)_k}{(-1)^n(1-\alpha-k)(1-\alpha-k+1)\dots(1-\alpha-k+n-1)}$$

$$= \frac{(-1)^n(\alpha)_k}{(1-\alpha-k)_n}$$

7. In (6) we have proved that

$$(\alpha)_{k-n} = \frac{(-1)^n(\alpha)_k}{(1-\alpha-k)_n}$$

Putting $k = 0$, we get $(\alpha)_{-n} = \dfrac{(-1)^n}{(1-\alpha)_n}, \quad \because (\alpha)_0 = 1$

8. $(-k)_n = (-k)(-k+1)(-k+2)\dots\dots(-k+n-1)$

$$= \frac{(-1)^n . k(k-1)(k-2)\dots(k-(n-1)).(k-n)\dots 2.1}{(k-n)(k-n-1)\dots 2.1} \text{ if } 0 \le n \le k$$

$$= \frac{(-1)^n k!}{(k-n)!} \text{ if } 0 \le n \le k$$

$$= 0 \text{ if } n > k, \quad \because \quad (-p)! = \infty, p > 0$$

5.3 Hypergeometric Series

The series defined by

$$1 + \frac{\alpha\beta}{\gamma}x + \frac{\alpha(\alpha+1)\beta(\beta+1)}{\gamma(\gamma+1)}\frac{x^2}{2!} + \dots\dots = \sum_{r=0}^{\infty}\frac{(\alpha)_r(\beta)_r}{(\gamma)_r}\frac{x^r}{r!} \qquad \dots(1)$$

is called the **hypergeometric series.**

Note : 1. Throughout this chapter α, β, γ are assumed to be real.

2. Above series is given the name hypergeometric series from the fact that, for $\alpha = 1, \gamma = \beta$, this series reduces to the geometric series

$$1 + x + x^2 + \dots = \sum_{n=0}^{\infty} x^n$$

The above series (1) converges absolutely if $|x| < 1$ and it is divergent for all $|x| > 1$. if $|x| = 1$ then the series converges absolutely if $\gamma - \alpha - \beta > 0$ *i.e.*, $\gamma > \alpha + \beta$

5.4 Hypergeometric Function

The function defined by the hypergeometric series $\sum\limits_{r=0}^{\infty} \frac{(\alpha)_r (\beta)_r}{(\gamma)_r} \frac{x^r}{r!}, |x| < 1$ is called the **hypergeometric function.** It is commonly denoted by the symbol ${}_2F_1(\alpha, \beta; \gamma; x)$ or simply $F(\alpha, \beta, \gamma, x)$

Thus $${}_2F_1(\alpha, \beta; \gamma; x) = \sum_{r=0}^{\infty} \frac{(\alpha)_r (\beta)_r}{(\gamma)_r} \frac{x^r}{r!} \qquad \ldots(1)$$

where 2 and 1 refer to the number of numerator and denominator parameters respectively, in its series representation. The semicolons separate the numerator parameter α and β (which are themselves separated by a comma), the denominator parameter γ and the argument x.

Note : 1. If $\gamma = 0$ or a negative integer, then the series (1) does not exist and so the function ${}_2F_1(\alpha, \beta; \gamma; x)$ is not defined.

2. If either α or β (or both) is zero or a negative integer, then the series is finite and thus converges for all x.

$\because$ If $\alpha = -m$ then

$(\alpha)_n = (-m)_n = 0$ when $n > m$

5.5 Different Forms of Hypergeometric Function

The hypergeometric function $F(\alpha, \beta, \gamma, x)$ can be put in the following forms as well

$$\begin{aligned} F(\alpha, \beta, \gamma, x) &= (1-x)^{\gamma-\alpha-\beta} F(\gamma-\alpha, \gamma-\beta; \gamma; x) \\ &= (1-x)^{-\alpha} F\left(\alpha, \gamma-\beta; \gamma; \frac{x}{x-1}\right) \\ &= (1-x)^{-\beta} F\left(\beta, \gamma-\alpha; \gamma; \frac{x}{x-1}\right) \end{aligned}$$

5.6 Confluent Hypergeometric Function (Kummer Function)

The function denoted by ${}_1F_1(\alpha; \beta; x)$ and defined by

$${}_1F_1(\alpha; \beta; x) = \sum_{r=0}^{\infty} \frac{(\alpha)_r}{(\beta)_r} \frac{x^r}{r!}$$

is called the **Confluent Hypergeometric Function** or **Kummer Function**

The Confluent Hypergeometric Function ${}_1F_1(\alpha\,;\beta\,;x)$ is also denoted by $M(\alpha,\beta,x)$

5.7 General Hypergeometric Function

The general hypergeometric function denoted by ${}_mF_n(\alpha_1,\alpha_2,\dots\alpha_m;\beta_1,\beta_2,\dots\beta_n;x)$ is defined by

$$ {}_mF_n(\alpha_1,\alpha_2,\dots\alpha_m;\beta_1,\beta_2,\dots\beta_n;x)=\sum_{r=0}^{\infty}\frac{(\alpha_1)_r(\alpha_2)_r\dots(\alpha_m)_r}{(\beta_1)_r(\beta_2)_r\dots(\beta_n)_r}\frac{x^r}{r!} $$

Another notation often used for general hypergeometric function is

$$ {}_mF_n\left[\begin{matrix}\alpha_1,\alpha_2,\dots\dots,\alpha_m\,; \\ \beta_1,\beta_2,\dots\dots,\beta_n\,;\end{matrix}\;x\right] $$

Note : 1. Confluent hypergeometric function (5.6) is special case of general hypergeometric function when $m=n=1$.

2. Hypergeometric function (5.4) is special case of general hypergeometric function when $m=2$ and $n=1$.

5.8 Symmetric Property of Hypergeometric Function

[Kanpur 2002, 03]

Hypergeometric function does not change if the parameters α and β are interchanged, keeping γ fixed *i.e.*, $F(\alpha,\beta;\gamma;x)=F(\beta,\alpha\,;\gamma;x)$

Proof: By the definition of hypergeometric function, we have

$$ F(\alpha,\beta;\gamma;x)=\sum_{r=0}^{\infty}\frac{(\alpha)_r(\beta)_r}{(\gamma)_r}\frac{x^r}{r!} $$

$$ =\sum_{r=0}^{\infty}\frac{(\beta)_r(\alpha)_r}{(\gamma)_r}\frac{x^r}{r!}=F(\beta,\alpha\,;\gamma;x) $$

5.9 Particular Cases of Hypergeometric Series

Here we give some well known series which can be denoted by hypergeometric function.

1. $e^x={}_1F_1(\alpha\,;\alpha\,;x)$

2. $(1-x)^{-1}={}_2F_1(\alpha,1;\alpha\,;x)={}_2F_1(1,\alpha\,;\alpha\,;x)\,,|x|<1$ **[Kanpur 2004]**

3. $(1-x)^{-\alpha}={}_2F_1(\alpha,\beta;\beta;x)={}_2F_1(\beta,\alpha\,;\beta;x)$

4. $(1+x)^n = {}_2F_1\,(-n, 1; 1; -x) = {}_2F_1\,(1, -n; 1; -x)$ **[Meerut 2011 (Sem. I)]**

5. $\log_e\,(1+x) = x\,{}_2F_1\,(1, 1; 2; -x)$

6. $\log_e\,(1-x) = -x\,{}_2F_1\,(1, 1; 2; x)$ **[Kanpur 2000]**

7. $\log\left(\frac{1+x}{1-x}\right) = 2x\,{}_2F_1\,(1/2, 1; 3/2; x^2) = 2x\,{}_2F_1\,(1, 1/2; 3/2; x^2)$

8. $\sin^{-1} x = x\,{}_2F_1\,(1/2, 1/2; 3/2; x^2)$

9. $\tan^{-1} x = x\,{}_2F_1\,(1/2, 1; 3/2; -x^2) = x\,{}_2F_1\,(1, 1/2; 3/2; -x^2)$

Proof: 1. $e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots$

$$= 1 + \frac{\alpha}{\alpha}\cdot\frac{x}{1!} + \frac{\alpha\,(\alpha+1)}{\alpha\,(\alpha+1)}\cdot\frac{x^2}{2!} + \ldots. = {}_1F_1,(\alpha; \alpha; x)$$

2. $(1-x)^{-1} = 1 + x + x^2 + x^3 + \ldots.$

$$= 1 + \frac{\alpha\,.1}{\alpha}\,\frac{x}{1!} + \frac{\alpha\,(\alpha+1)\,.1\,(1+1)}{\alpha\,(\alpha+1)}\,\frac{x^2}{2!} + \frac{\alpha\,(\alpha+1)\,(\alpha+2)\,.1\,(1+1)\,(1+2)}{\alpha\,(\alpha+1)\,(\alpha+2)}\,\frac{x^3}{3!} + \ldots.$$

$$= {}_2F_1\,(\alpha, 1; \alpha; x) = {}_2F_1\,(1, \alpha; \alpha; x)$$

(By symmetric property)

3. $(1-x)^{-\alpha} = 1 + (-\alpha)\,(-x) + \frac{(-\alpha)\,(-\alpha-1)}{2!}\,(-x)^2 + \frac{(-\alpha)\,(-\alpha-1)\,(-\alpha-2)}{3!}\,(-x)^3 + \ldots$

$$= 1 + \alpha\,x + \frac{\alpha\,(\alpha+1)}{2!}\,x^2 + \frac{\alpha\,(\alpha+1)\,(\alpha+2)}{3!}\,x^3 + \ldots.$$

$$= 1 + \frac{\alpha\,.\beta}{\beta}\,x + \frac{\alpha\,(\alpha+1)\,.\beta\,(\beta+1)}{\beta\,(\beta+1)}\,\frac{x^2}{2!} + \frac{\alpha\,(\alpha+1)\,(\alpha+2)\,.\,\beta\,(\beta+1)\,(\beta+2)}{\beta\,(\beta+1)\,(\beta+2)}\,\frac{x^3}{3!} + \ldots..$$

$$= {}_2F_1\,(\alpha, \beta; \beta; x) = {}_2F_1\,(\beta, \alpha; \beta; x)$$

(By symmetric property)

4. $(1+x)^n = 1 + n\,x + \frac{n\,(n-1)}{2!}\,x^2 + \frac{n\,(n-1)\,(n-2)}{3!}\,x^3 + \ldots.$

$$= 1 + \frac{(-n)\,.1}{1}\cdot\frac{(-x)}{1!} + \frac{(-n)\,(-n+1)\,.1\,(1+1)}{1\,(1+1)}\cdot\frac{(-x)^2}{2!}$$

$$+ \frac{(-n)\,(-n+1)\,(-n+2)\,.1\,(1+1)\,(1+2)}{1\,(1+1)\,(1+2)}\cdot\frac{(-x)^3}{3!} + \ldots.$$

$$= {}_2F_1\,(-n, 1; 1; -x) = {}_2F_1\,(1, -n; 1; -x)$$

(By symmetric property)

5. $$\log_e (1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots$$

$$= x\left[1 - \frac{x}{2} + \frac{x^2}{3} - \frac{x^3}{4} + \dots\right]$$

$$= x\left[1 + \frac{1.1}{2}\cdot\frac{(-x)}{1!} + \frac{1\,(1+1).1\,(1+1)}{2\,(2+1)}\cdot\frac{(-x)^2}{2\,!} + \frac{1\,(1+1)\,(1+2)\,1\,(1+1)\,(1+2)}{2\,(2+1)\,(2+2)}\cdot\frac{(-x)^3}{3\,!} + \dots\right]$$

$$= x\;{}_2F_1\,(1, 1; 2\,; -x)$$

6. $$\log_e (1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \dots = -x\left[1 + \frac{x}{2} + \frac{x^2}{3} + \frac{x^3}{4} + \dots\right]$$

$$= -x\left[1 + \frac{1.1}{2}\cdot\frac{x}{1!} + \frac{1\,(1+1)\,.\,1\,(1+1)}{2\,(2+1)}\cdot\frac{x^2}{2\,!} + \frac{1\,(1+1)\,(1+2)\,.\,1(1+1)\,(1+2)}{2\,(2+1)\,(2+2)}\cdot\frac{x^3}{3\,!} + \dots\right]$$

$$= -x\;{}_2F_1\,(1, 1; 2\,; x)$$

7. $$\log\left(\frac{1+x}{1-x}\right) = 2\left(x + \frac{x^3}{3} + \frac{x^5}{5} + \dots\dots\right)$$

$$= 2x\left(1 + \frac{x^2}{3} + \frac{x^4}{5} + \dots\right)$$

$$= 2x\left[1 + \frac{(1/2).1}{3/2}\cdot\frac{x^2}{1!} + \frac{(1/2)\,(1/2+1)\,.\,1\,(1+1)}{(3/2)\,(3/2+1)}\cdot\frac{(x^2)^2}{2\,!} + \dots\right]$$

$$= 2x\;{}_2F_1\,(1/2, 1; 3/2\,; x^2) = 2x\;{}_2F_1\,(1, 1/2\,; 3/2\,; x^2) \qquad \text{(By symmetric property)}$$

8. $$\sin^{-1} x = x + 1^2\cdot\frac{x^3}{3\,!} + 1^2\cdot 3^2\cdot\frac{x^5}{5\,!} + 1^2\cdot 3^2\cdot 5^2\cdot\frac{x^7}{7!} + \dots$$

$$= x\left[1 + 1^2\cdot\frac{x^2}{3\,!} + 1^2\cdot 3^2\cdot\frac{x^4}{5\,!} + 1^2\cdot 3^2\cdot 5^2\cdot\frac{x^6}{7!} + \dots\right]$$

$$= x\left[1 + \frac{1/2\,.\,1/2}{3/2}\cdot\frac{x^2}{1!} + \frac{(1/2)\,(1/2+1)\,(1/2)\,(1/2+1)}{(3/2)\,(3/2+1)}\cdot\frac{(x^2)^2}{2\,!} + \dots\right]$$

$$= x\;{}_2F_1\,(1/2, 1/2\,; 3/2\,; x^2)$$

Aliter We have ${}_2F_1\,(\alpha, \beta; \gamma; x) = 1 + \dfrac{\alpha\,.\,\beta}{\gamma}\,\dfrac{x}{1!} + \dfrac{\alpha\,(\alpha+1)\,\beta\,(\beta+1)}{\gamma\,(\gamma+1)}\,\dfrac{x^2}{2\,!} + \dots$

Replacing α, β, γ, x by 1/2, 1/2, 3/2, x^2 respectively, we get

$$x \,. {}_2F_1\,(1/2, 1/2; 3/2; x^2)$$

$$= x\left[1 + \frac{1/2 \,.\, 1/2}{3/2}\,\frac{x^2}{1!} + \frac{(1/2)\,.\,(1/2+1)\,.\,1/2\,(1/2+1)}{(3/2)\,(3/2+1)}\,.\,\frac{(x^2)^2}{2\,!} + \ldots .\right]$$

$$= x + 1^2 \,.\, \frac{x^2}{3\,!} + 1^2 \,.\, 3^2 \,.\, \frac{x^5}{5\,!} + 1^2 \,.\, 3^2 \,.\, 5^2\, \frac{x^7}{7\,!} + \ldots .$$

9. $\tan^{-1} x = x - \dfrac{x^3}{3} + \dfrac{x^5}{5} - \dfrac{x^7}{7} + \ldots .$

$$= x\left[1 - \frac{x^2}{3} + \frac{x^4}{5} - \frac{x^6}{7} + \ldots\right]$$

$$= x\left[1 + \frac{1/2 \,.\, 1}{3/2}\,\frac{(-x^2)}{1!} + \frac{(1/2)\,(1/2+1)\,.\,1\,(1+1)}{(3/2)\,(3/2+1)}\,\frac{(-x)^2}{2\,!} + \ldots .\right]$$

$$= x\;{}_2F_1\,(1/2, 1; 3/2; -x^2) = x\;{}_2F_1\,(1, 1/2; 3/2; -x^2)$$

(By symmetric property)

Aliter : We have, ${}_2F_1\,(\alpha, \beta; \gamma; x) = 1 + \dfrac{\alpha \,.\, \beta}{\gamma}\,\dfrac{x}{1!} + \dfrac{\alpha\,(\alpha+1)\,.\,\beta\,(\beta+1)}{\gamma\,(\gamma+1)}\,.\,\dfrac{x^2}{2\,!} + \ldots .$

Replacing α, β, γ, x by 1/2, 1, 3/2, $-x^2$, we get

$$x \,. {}_2F_1\,(1/2, 1; 3/2; -x^2)$$

$$= x\left[1 + \frac{1/2 \,.\, 1}{3/2}\,.\,\frac{(-x^2)}{1!} + \frac{(1/2)\,(1/2+1)\,.\,1\,(1+1)}{(3/2)\,(3/2+1)}\cdot\frac{(-x^2)^2}{2\,!} + \ldots .\right]$$

$$= x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \ldots .$$

5.10 Gauss's Hypergeometric Equation or Gauss's Equation or Hypergeometric Equation

The differential equation

$$x\,(1-x)\,\frac{d^2 y}{dx^2} + \{\gamma - (\alpha + \beta + 1)\,x\}\,\frac{dy}{dx} - \alpha\beta\; y = 0$$

is called **Gauss's hypergeometric equation** or **Gauss's equation** or **hypergeometric equation.**

5.11 Solution of the Hypergeometric Equation

[Meerut 2010 (Sem. I); Kanpur 2006]

Hypergeometric equation is

$$x(1-x)\frac{d^2y}{dx^2}+\{\gamma-(\alpha+\beta+1)x\}\frac{dy}{dx}-\alpha\beta y=0 \qquad \ldots(1)$$

Dividing by $x(1-x)$, we get $\dfrac{d^2y}{dx^2}+\dfrac{\gamma-(\alpha+\beta+1)x}{x(1-x)}\dfrac{dy}{dx}-\dfrac{\alpha\beta}{x(1-x)}y=0$

Comparing with $\dfrac{d^2y}{dx^2}+P(x)\dfrac{dy}{dx}+Q(x)\,y=0$, we get

$$P(x)=\frac{\gamma-(\alpha+\beta+1)x}{x(1-x)} \text{ and } Q(x)=-\frac{\alpha\beta}{x(1-x)}$$

Since $P(x)$ and $Q(x)$ are not finite at $x=0$ and $x=1$, and $x\,P(x)$ and $x^2Q(x)$ are both finite and so analytic at $x=0$ also $(x-1)\,P(x)$ and $(x-1)^2\,Q(x)$ are also both finite and so analytic at $x=1$.

Thus the points $x=0$ and $x=1$ are both regular singular points of (1).

To find series solution of (1) near $x=0$ **[Kanpur 2006]**

Let $$y=\sum_{r=0}^{\infty}C_r x^{m+r},\quad C_0\neq 0 \qquad \ldots(2)$$

be series solution of (1) near $x=0$

Diff. (2), we get

$$\frac{dy}{dx}=\sum_{r=0}^{\infty}C_r(m+r)\,x^{m+r-1} \text{ and } \frac{d^2y}{dx^2}=\sum_{r=0}^{\infty}C_r\,(m+r)(m+r-1)\,x^{m+r-2}$$

Substituting in (1), we get

$$\sum_{r=0}^{\infty}C_r\,[(x-x^2)(m+r)(m+r-1)\,x^{m+r-2}$$
$$+\{\gamma-(\alpha+\beta+1)x\}(m+r)\,x^{m+r-1}-\alpha\beta\,x^{m+r}]=0$$

or $$\sum_{r=0}^{\infty}C_r\,[\{(m+r)(m+r-1)+\gamma(m+r)\}\,x^{m+r-1}$$
$$-\{(m+r)(m+r-1)+(\alpha+\beta+1)(m+r)+\alpha\beta\}\,x^{m+r}]=0$$

or $$\sum_{r=0}^{\infty} C_r \left[(m+r)(m+r+\gamma-1)\, x^{m+r-1} - \{(m+r)^2 + (\alpha+\beta)(m+r) + \alpha\beta\}\, x^{m+r}\right] = 0$$

or $$\sum_{r=0}^{\infty} C_r \left[(m+r)(m+r+\gamma-1)\, x^{m+r-1} - (m+r+\alpha)(m+r+\beta)\, x^{m+r}\right] = 0 \quad \ldots(3)$$

which is an identity, so we can equate to zero the coefficients of various powers of x,

Equating to zero the coefficient of lowest power of x *i.e.*, of x^{m-1}, the *indicial equation* is :

$$C_0 m (m+\gamma-1) = 0 \Rightarrow m = 0 \text{ or } m = 1-\gamma$$

which are unequal and not differing by an integer as γ is not an integer.

Also equating to zero the coefficient of x^{m+r-1}, we get

$$C_r (m+r)(m+r+\gamma-1) - (m+r-1+\alpha)(m+r-1+\beta)\, C_{r-1} = 0$$

$$\Rightarrow \quad C_r = \frac{(m+r-1+\alpha)(m+r-1+\beta)}{(m+r)(m+r+\gamma-1)}\, C_{r-1}, \quad \text{for } r \geq 1 \quad \ldots(4)$$

Case I. When $m = 0$ Putting $m = 0$, in (4), we get

$$C_r = \frac{(\alpha+r-1)(\beta+r-1)}{r(\gamma+r-1)}\, C_{r-1}, \quad \text{for } r \geq 1$$

Putting $r = 1, 2, 3$ .... ,

$$C_1 = \frac{\alpha . \beta}{1 . \gamma} C_0,\ C_2 = \frac{(\alpha+1)(\beta+1)}{2(\gamma+1)} C_1 = \frac{\alpha(\alpha+1)\beta(\beta+1)}{2!\,\gamma(\gamma+1)} C_0$$

$$C_3 = \frac{(\alpha+2)(\beta+2)}{3(\gamma+2)} C_2 = \frac{\alpha(\alpha+1)(\alpha+2)\beta(\beta+1)(\beta+2)}{3!\,\gamma(\gamma+1)(\gamma+2)} C_0 \text{ and so on.}$$

Putting in (2), we get

$$y = C_0 \left[1 + \frac{\alpha . \beta}{\gamma} \frac{x}{1!} + \frac{\alpha(\alpha+1)\beta(\beta+1)}{\gamma(\gamma+1)} \frac{x^2}{2!} + \frac{\alpha(\alpha+1)(\alpha+2)\beta(\beta+1)(\beta+2)}{\gamma(\gamma+1)(\gamma+2)} \frac{x^3}{3!} + \ldots\right]$$

or $y = C_0\, {}_2F_1(\alpha, \beta; \gamma; x)$

Taking $C_0 = 1$, we see that $\quad y = {}_2F_1(\alpha, \beta; \gamma; x) \quad \ldots(5)$

called hypergeometric function is one solution of (1).

Case II. When $m = 1 - \gamma$

Putting $m = 1 - \gamma$ in (4), we get

$$C_r = \frac{(1-\gamma+r-1+\alpha)\,(1-\gamma+r-1+\beta)}{(1-\gamma+r)\,(1-\gamma+r+\gamma-1)}\,C_{r-1}$$

$$\Rightarrow \qquad C_r = \frac{(\alpha'+r-1)\,(\beta'+r-1)}{r\,(\gamma'+r-1)}\,C_{r-1},\ r \geq 1 \qquad \text{...(6)}$$

where $\alpha' = 1-\gamma+\alpha$, $\beta' = 1-\gamma+\beta$ and $\gamma' = 2-\gamma$

Putting $r = 1, 2, 3, \text{.........}$ in (6), we get

$$C_1 = \frac{\alpha', \beta'}{1.\gamma'}\,C_0,\ C_2 = \frac{(\alpha'+1)\,(\beta'+1)}{2\,(\gamma'+1)}\,C_1 = \frac{\alpha'\,(\alpha'+1)\,.\,\beta'\,(\beta'+1)}{2\,!\,\gamma'\,(\gamma'+1)}\,C_0$$

$$C_3 = \frac{(\alpha'+2)\,(\beta'+2)}{3\,(\gamma'+2)}\,C_2 = \frac{\alpha'\,(\alpha'+1)\,(\alpha'+2)\,.\,\beta'\,(\beta'+1)(\beta'+2)}{3\,!\,\gamma'\,(\gamma'+1)\,(\gamma'+2)}\,C_0, \text{ and so on.}$$

Putting in (2), we get

$$y = C_0 x^{1-\gamma}\left[1 + \frac{\alpha'\,.\,\beta'}{\gamma'}\cdot\frac{x}{1!} + \frac{\alpha'\,(\alpha'+1)\,\beta'\,(\beta'+1)}{\gamma'\,(\gamma'+1)}\cdot\frac{x^2}{2\,!} + \frac{\alpha'\,(\alpha'+1)\,(\alpha'+2)\,.\,\beta'\,(\beta'+1)(\beta'+2)}{\gamma'\,(\gamma'+1)\,(\gamma'+2)}\,\frac{x^3}{3\,!} + \ldots\right]$$

or $\qquad y = C_0 x^{1-\gamma}\ {}_2F_1\,(\alpha', \beta'; \gamma'; x)$

or $\qquad y = C_0 x^{1-\gamma}\ {}_2F_1\,(1-\gamma+\alpha, 1-\gamma+\beta; 2-\gamma; x)$

Taking $C_0 = 1,\ y = x^{1-\gamma}\ {}_2F_1\,(1-\gamma+\alpha, 1-\gamma+\beta; 2-\gamma; x)$ ...(7)

Which is another independent solution of (1).

Hence the general solution of (1) near $x = 0$ is

$$y = a.{}_2F_1\,(\alpha, \beta; \gamma; x) + b.x^{1-\gamma}\ {}_2F_1\,(1-\gamma+\alpha, 1-\gamma+\beta, 2-\gamma, x) \qquad \text{...(8)}$$

where a and b are arbitrary constants.

To find series solution of (1) near $x = 1$.

To find series solution of diff. equation (1) near $x = 1$ *i.e.*, in powers of $x - 1$, we change the independent variable from x to z, by the substitution

$$1 - x = z, \text{ so that } \frac{dz}{dx} = -1, \frac{dy}{dx} = \frac{dy}{dz}\frac{dz}{dx} = -\frac{dy}{dz}$$

and $$\frac{d^2y}{dx^2} = \frac{d}{dx}\left(\frac{dy}{dx}\right) = \frac{d}{dx}\left(-\frac{dy}{dz}\right) = -\frac{d^2y}{dz^2}\cdot\frac{dz}{dx} = \frac{d^2y}{dz^2}$$

∴ Diff., equation (1), reduces to

$$(1-z)\,z\,.\frac{d^2y}{dz^2} + \{\gamma - (\alpha+\beta+1)(1-z)\}\left(-\frac{dy}{dz}\right) - \alpha\beta y = 0$$

or $$z(1-z)\frac{d^2y}{dz^2} + \{(\alpha+\beta-\gamma+1) - (\alpha+\beta+1)\,z\}\frac{dy}{dz} - \alpha\beta y = 0 \qquad \ldots(9)$$

which is identical with (1) with γ replaced by $\alpha+\beta-\gamma+1$ and

Hence from (8), solution of (9) is

$$y = a.{}_2F_1(\alpha,\beta;\alpha+\beta-\gamma+1;z) + b\,.\,z^{1-(\alpha+\beta-\gamma+1)}.$$
$$_2F_1(\gamma-\beta,\gamma-\alpha,\gamma-\alpha-\beta+1;z)$$

⇒ Solution of (1) near $x = 1$ is

$$y = a.{}_2F_1(\alpha,\beta;\alpha+\beta-\gamma+1;1-x) + b(1-x)^{\gamma-\alpha-\beta}.$$
$$_2F_1(\gamma-\alpha,\gamma-\beta;\gamma-\alpha-\beta+1;1-x) \quad \ldots(10)$$

5.12 Derivatives of Hypergeometric Functions

We have, $$_2F_1(\alpha,\beta;\gamma;x) = \sum_{n=0}^{\infty}\frac{(\alpha)_n(\beta)_n}{(\gamma)_n}\cdot\frac{x^n}{n!}$$

Differentiating both sides, w.r.t. x

$$\frac{d}{dx}\,{}_2F_1(\alpha,\beta;\gamma;x) = \sum_{n=0}^{\infty}\frac{(\alpha)_n(\beta)_n}{(\gamma)_n}\cdot\frac{n\,x^{n-1}}{n!}$$
$$= \sum_{n=1}^{\infty}\frac{(\alpha)_n(\beta)_n}{(\gamma)_n}\cdot\frac{x^{n-1}}{(n-1)!}$$
$$= \sum_{m=0}^{\infty}\frac{(\alpha)_{m+1}(\beta)_{m+1}}{(\gamma)_{m+1}}\cdot\frac{x^m}{m!}, \quad \text{Taking } n-1 = m$$

Since $$(\alpha)_{m+1} = \alpha(\alpha+1)(\alpha+2)\ldots.(\alpha+m+1-1)$$
$$= \alpha\,.(\alpha+1)(\alpha+2)\ldots.(\alpha+1+m-1)$$
$$= \alpha(\alpha+1)_m \qquad \ldots(1)$$

$$\therefore \qquad \frac{d}{dx}\,{}_2F_1(\alpha,\beta;\gamma;x)=\frac{\alpha\beta}{\gamma}\sum_{m=0}^{\infty}\frac{(\alpha+1)_m(\beta+1)_m}{(\gamma+1)_m}\cdot\frac{x^m}{m!} \qquad \ldots(2)$$

$$\Rightarrow \qquad \frac{d}{dx}\,{}_2F_1(\alpha,\beta;\gamma;x)=\frac{(\alpha)_1\cdot(\beta)_1}{(\gamma)_1},\ {}_2F_1(\alpha+1,\beta+1;\gamma+1;x)$$

$\because \qquad (\alpha)_1=\alpha$ **etc.**

[Kanpur 2005; Kurukshetra 2006; Garhwal 2005]

Differentiating again

$$\frac{d^2}{dx^2}\,{}_2F_1(\alpha,\beta;\gamma;x)=\frac{\alpha\beta}{\gamma}\sum_{m=1}^{\infty}\frac{(\alpha+1)_m\cdot(\beta+1)_m}{(\gamma+1)_m}\cdot\frac{mx^{m-1}}{m!}$$

$$=\frac{\alpha\beta}{\gamma}\sum_{m=1}^{\infty}\frac{(\alpha+1)_m\cdot(\beta+1)_m}{(\gamma+1)_m}\cdot\frac{x^{m-1}}{(m-1)!}$$

$$=\frac{\alpha\beta}{\gamma}\sum_{n=0}^{\infty}\frac{(\alpha+1)_{n+1}\cdot(\beta+1)_{n+1}}{(\gamma+1)_{n+1}}\cdot\frac{x^n}{n!} \qquad \text{Taking } m-1=n$$

$$=\frac{\alpha\beta}{\gamma}\sum_{n=0}^{\infty}\frac{(\alpha+1)(\alpha+2)_n(\beta+1)(\beta+2)_n}{(\gamma+1)(\gamma+2)_n}\cdot\frac{x^n}{n!}$$

$[\because$ From (1), $(\alpha+1)_{n+1}=(\alpha+1)(\alpha+2)_n)]$

$$=\frac{\alpha(\alpha+1)\beta(\beta+1)}{\gamma(\gamma+1)}\cdot\sum_{n=0}^{\infty}\frac{(\alpha+2)_n\cdot(\beta+2)_n}{(\gamma+2)_n}\cdot\frac{x^n}{n!} \qquad \ldots(3)$$

$$\Rightarrow \qquad \frac{d^2}{dx^2}\,{}_2F_1(\alpha,\beta;\gamma;x)=\frac{(\alpha)_2\cdot(\beta)_2}{(\gamma)_2}\,{}_2F_1(\alpha+2,\beta+2;\gamma+2;x)$$

$\because \qquad (\alpha)_2=\alpha(\alpha+2-1)=\alpha(\alpha+1)$ **etc.**

5.13 n-th Derivative of Hypergeometric Function

To Prove $\dfrac{d^n}{dx^n}\,{}_2F_1(\alpha,\beta;\gamma;x)=\dfrac{(\alpha)_n(\beta)_n}{(\gamma)_n}\,{}_2F_1(\alpha+n,\beta+n;\gamma+n;x) \qquad \ldots(1)$

where *n* is a positive integer. **[Meerut 2010 (Sem. I)]**

Proof: In 5.12 we have proved that

$$\frac{d}{dx}\,{}_2F_1(\alpha,\beta;\gamma;x)=\frac{(\alpha)_1(\beta)_1}{(\gamma)_1}\,{}_2F_1(\alpha+1,\beta+1;\gamma+1;x) \qquad \ldots(2)$$

i.e., (1) is true for $n=1$

Now we assume that (1) is true for a positive integer $m<n$, then

$$\frac{d^m}{dx^m}\,{}_2F_1(\alpha,\beta;\gamma;x)=\frac{(\alpha)_m(\beta)_m}{(\gamma)_m}\,{}_2F_1(\alpha+m,\beta+m;\gamma+m;x)$$

Differentiating both sides, w.r.t. x, we get

$$\frac{d^{m+1}}{dx^{m+1}}\,{}_2F_1(\alpha,\beta;\gamma;x)=\frac{(\alpha)_m(\beta)_m}{(\gamma)_m}\cdot\frac{d}{dx}\,{}_2F_1(\alpha+m,\beta+m;\gamma+m;x)$$

$$=\frac{(\alpha)_m(\beta)_m}{(\gamma)_m}\cdot\frac{(\alpha+m)(\beta+m)}{(\gamma+m)}\;{}_2F_1(\alpha+m+1,\beta+m+1;\gamma+m+1;x)$$

(using (1) for $\alpha+m,\beta+m,\gamma+m$ in place of α,β,γ respectively and with $(\alpha)_1=\alpha$)

$$=\frac{(\alpha)_{m+1}(\beta)_{m+1}}{(\gamma)_{m+1}}\cdot\;{}_2F_1(\alpha+m+1,\beta+m+1;\gamma+m+1;x) \qquad ...(3)$$

$$\because \quad (\alpha)_m\,.\,(\alpha+m)=(\alpha)_{m+1}$$

from (3) it follows that (1) is true for $m+1$.

i.e., assuming that (1) is true for positive integer $m<n$, (1) is also true for $m+1$.

But from (2), it is obvious that (1) is true for $n=1$, hence by mathematical induction (1) is true for all positive integral value of n.

5.14 Derivatives of Hypergeometric Function at x = 0

1. $\frac{d}{dx}\,{}_2F_1(\alpha\;\;\beta;\gamma;0)=1$
2. $\frac{d^2}{dx^2}\,{}_2F_1(\alpha,\beta;\gamma;0)=\frac{\alpha.\beta}{\gamma}=\frac{(\alpha)_1(\beta)_1}{(\gamma)_1}$
3. $\frac{d^n}{dx^n}\,{}_2F_1(\alpha,\beta;\gamma;0)=\frac{(\alpha)_n(\beta)_n}{(\gamma)_n}$

Proof: We have,

$${}_2F_1(\alpha,\beta;\gamma;x)=\sum_{n=0}^{\infty}\frac{(\alpha)_n\cdot(\beta)_n}{(\gamma)_n}\cdot\frac{x^n}{n!}$$

$$=\frac{(\alpha)_0\cdot(\beta)_0}{(\gamma)_0}\cdot\frac{1}{0!}+\frac{(\alpha)_1\cdot(\beta)_1}{(\gamma)_1}\cdot\frac{x}{1!}+\frac{(\alpha)_2\cdot(\beta)_2}{(\gamma)_2}\cdot\frac{x^2}{2!}+\ldots+\frac{(\alpha)_n\cdot(\beta)_n}{(\gamma)_n}\cdot\frac{x^n}{n!}+\ldots$$

$$\therefore \quad \frac{d}{dx}\,{}_2F_1(\alpha,\beta;\gamma;x) = 0 + \frac{(\alpha)_1(\beta)_1}{(\gamma)_1}\cdot 1 + \frac{(\alpha)_2(\beta)_2}{(\gamma)_2}\cdot\frac{2x}{2!} + \frac{(\alpha)_3(\beta)_3}{(\gamma)_3}\cdot\frac{3x^2}{3!} + \ldots$$

$$+ \frac{(\alpha)_n(\beta)_n}{(\gamma)_n}\cdot\frac{n\,x^{n-1}}{n!} + \ldots.$$

$$\frac{d^2}{dx^2}\,{}_2F_1(\alpha,\beta;\gamma;x) = 0 + \frac{(\alpha)_2(\beta)_2}{(\gamma)_2}\cdot 1 + \frac{(\alpha)_3(\beta)_3}{(\gamma)_3}\cdot\frac{3.2x}{3!} + \ldots.$$

$$+ \frac{(\alpha)_n(\beta)_n}{(\gamma)_n}\cdot\frac{n(n-1)\,x^{n-2}}{n!} + \ldots$$

Proceeding similarly

$$\frac{d^n}{dx^n}\,{}_2F_1(\alpha,\beta;\gamma;x) = 0 + \frac{(\alpha)_n(\beta)_n}{(\gamma)_n}\cdot\frac{n!}{n!} + \frac{(\alpha)_{n+1}(\beta)_{n+1}}{(\gamma)_{n+1}}\cdot\frac{(n+1)\,n\,(n-1)\ldots 2}{n!}\,x + \ldots$$

Putting $x = 0$ we get

$${}_2F_1(\alpha,\beta;\gamma;x) = \frac{(\alpha)_0(\beta)_0}{(\gamma)_0} = 1$$

$$\frac{d}{dx}\,{}_2F_1(\alpha,\beta;\gamma;0) = \frac{(\alpha)_1(\beta)_1}{(\gamma)_1} \quad \text{or} \quad \frac{\alpha\beta}{\gamma}$$

$$\frac{d^2}{dx^2}\,{}_2F_1(\alpha,\beta;\gamma;0) = \frac{(\alpha)_2(\beta)_2}{(\gamma)_2} \quad \text{or} \quad \frac{\alpha(\alpha+1)\beta(\beta+1)}{\gamma(\gamma+1)}$$

$$\frac{d^n}{dx^n}\,{}_2F_1(\alpha,\beta;\gamma;0) = \frac{(\alpha)_n(\beta)_n}{(\gamma)_n}$$

5.15 Integral Formula for Hypergeometric Function

If $|x| < 1$ ***and*** $\gamma > \beta > 0$***, then***

$${}_2F_1(\alpha,\beta;\gamma;x) = \frac{\Gamma(\gamma)}{\Gamma(\beta)\,\Gamma(\gamma-\beta)}\int_0^1 t^{\beta-1}(1-t)^{\gamma-\beta-1}(1-xt)^{-\alpha}\,dt$$

or
$${}_2F_1(\alpha,\beta;\gamma;x) = \frac{1}{B(\beta,\gamma-\beta)}\int_0^1 t^{\beta-1}(1-t)^{\gamma-\beta-1}(1-xt)^{-\alpha}\,dt$$

[Rajasthan 2000, 09; Garhwal 2004; Kanpur 2004, 07; Purvanchal 2007]

Proof: We have the hypergeometric function

$${}_2F_1(\alpha,\beta;\gamma;x) = \sum_{n=0}^{\infty}\frac{(\alpha)_n(\beta)_n}{(\gamma)_n}\cdot\frac{x^n}{n!} \qquad \ldots(1)$$

where $(\alpha)_n = \alpha(\alpha+1)(\alpha+2)\ldots.(\alpha+n-1)$

$$= \frac{1.2 \ldots. (\alpha - 1)\, \alpha\, (\alpha + 1) \ldots. (\alpha + n - 1)}{1.2 \ldots\ldots (\alpha - 1)} = \frac{\Gamma(\alpha + n)}{\Gamma(\alpha)}$$

$$\therefore \qquad \frac{(\beta)_n}{(\gamma)_n} = \frac{\Gamma(\beta + n)}{\Gamma(\beta)} \cdot \frac{\Gamma(\gamma)}{\Gamma(\gamma + n)} = \frac{\Gamma(\gamma)\, \Gamma(\beta + n)}{\Gamma(\beta)\, \Gamma(\gamma + n)} \cdot \frac{\Gamma(\gamma - \beta)}{\Gamma(\gamma - \beta)}$$

$$= \frac{\Gamma(\gamma)}{\Gamma(\beta)} \cdot \frac{\Gamma(\beta + n)\, \Gamma(\gamma - \beta)}{\Gamma(\beta + n + \gamma - \beta)} \cdot \frac{1}{\Gamma(\gamma - \beta)}$$

$$= \frac{\Gamma(\gamma)}{\Gamma(\beta)\, \Gamma(\gamma - \beta)} \cdot \mathrm{B}(\beta + n, \gamma - \beta)$$

$$= \frac{\Gamma(\gamma)}{\Gamma(\beta)\, \Gamma(\gamma - \beta)} \cdot \int_0^1 t^{\beta + n - 1} \cdot (1 - t)^{\gamma - \beta - 1}\, dt$$

$\therefore$ From (1) ${}_2F_1 (\alpha, \beta; \gamma; x)$

$$= \sum_{n=0}^{\infty} \frac{\Gamma(\gamma)}{\Gamma(\beta)\, \Gamma(\gamma - \beta)} \int_0^1 t^{\beta + n - 1} (1 - t)^{\gamma - \beta - 1} \cdot (\alpha)_n \cdot \frac{x^n}{n!}\, dt$$

$$= \sum_{n=0}^{\infty} \frac{\Gamma(\gamma)}{\Gamma(\beta)\, \Gamma(\gamma - \beta)} \int_0^1 t^{\beta - 1} (1 - t)^{\gamma - \beta - 1} (\alpha)_n \cdot \frac{(tx)^n}{n!}\, dt$$

$$= \frac{\Gamma(\gamma)}{\Gamma(\beta)\, \Gamma(\gamma - \beta)} \int_0^1 t^{\beta - 1} (1 - t)^{\gamma - \beta - 1} \left\{ \sum_{n=0}^{\infty} (\alpha)_n \cdot \frac{(xt)^n}{n!} \right\} dt$$

$$= \frac{\Gamma(\gamma)}{\Gamma(\beta)\, \Gamma(\gamma - \beta)} \int_0^1 t^{\beta - 1} (1 - t)^{\gamma - \beta - 1} \cdot \left\{ (\alpha)_0 \,.\, 1 + (\alpha)_1 \cdot \frac{xt}{1!} + (\alpha)_2 \cdot \frac{(xt)^2}{2!} + \ldots. \right\} dt$$

$$= \frac{\Gamma(\gamma)}{\Gamma(\beta)\, \Gamma(\gamma - \beta)} \int_0^1 t^{\beta - 1} (1 - t)^{\gamma - \beta - 1} \cdot \left\{ 1 + \alpha \,.\, xt + \alpha\, (\alpha + 1) \,.\, \frac{(xt)^2}{2!} + \ldots \right\} dt$$

or
$$ {}_2F_1 (\alpha, \beta; \gamma; x) = \frac{\Gamma(\gamma)}{\Gamma(\beta)\, \Gamma(\gamma - \beta)} \cdot \int_0^1 t^{\beta - 1} (1 - t)^{\gamma - \beta - 1} (1 - xt)^{-\alpha}\, dt$$

which is the required integral formula.

Which can also be written as

$${}_2F_1 (\alpha, \beta; \gamma; x) = \frac{1}{B\, (\beta, \gamma - \beta)} \int_0^1 t^{\beta - 1} (1 - t)^{\gamma - \beta - 1} (1 - xt)^{-\alpha}\, dt$$

Integral formula for hypergeometric function is valid for $|x| < 1, \gamma > \beta > 0$.

5.16 Kummer's Theorem (For Hypergeometric Function)

$$ {}_2F_1(\alpha,\beta;\beta-\alpha+1;-1)=\frac{\Gamma(\beta-\alpha+1)\,\Gamma(\beta/2+1)}{\Gamma(\beta+1)\,\Gamma(\beta/2-\alpha+1)} $$ **[Purvanchal 2005]**

Proof: The integral formula for hypergeometric function is

$$ {}_2F_1(\alpha,\beta;\gamma;x)=\frac{\Gamma(\gamma)}{\Gamma(\beta)\,\Gamma(\gamma-\beta)}\cdot\int_0^1 t^{\beta-1}(1-t)^{\gamma-\beta-1}(1-xt)^{-\alpha}\,dt $$

Putting $x=-1$ and $\gamma=\beta-\alpha+1$, we get

$$ {}_2F_1(\alpha,\beta;\beta-\alpha+1;-1)=\frac{\Gamma(\beta-\alpha+1)}{\Gamma(\beta)\,\Gamma(1-\alpha)}\int_0^1 t^{\beta-1}(1-t)^{-\alpha}(1+t)^{-\alpha}\,dt $$

$$ =\frac{\Gamma(\beta-\alpha+1)}{\Gamma(\beta)\,\Gamma(1-\alpha)}\int_0^1 t^{\beta-1}(1-t^2)^{-\alpha}\,dt $$

$$ =\frac{\Gamma(\beta-\alpha+1)}{\Gamma(\beta)\,\Gamma(1-\alpha)}\int_0^1 z^{(\beta-1)/2}(1-z)^{-\alpha}\cdot\frac{1}{2}z^{-1/2}\,dz $$

Putting $t^2=z\,,\ t=z^{1/2}$ and $dt=\frac{1}{2}z^{-1/2}\,dz$

$$ =\frac{1}{2}\frac{\Gamma(\beta-\alpha+1)}{\Gamma(\beta)\,\Gamma(1-\alpha)}\int_0^1 z^{\beta/2-1}(1-z)^{(1-\alpha)-1}\,dz $$

$$ =\frac{1}{2}\frac{\Gamma(\beta-\alpha+1)}{\Gamma(\beta)\,\Gamma(1-\alpha)}.\mathrm{B}(\beta/2,1-\alpha) $$

$$ =\frac{1}{2}\frac{\Gamma(\beta-\alpha+1)}{\Gamma(\beta)\,\Gamma(1-\alpha)}\cdot\frac{\Gamma(\beta/2).\Gamma(1-\alpha)}{\Gamma(\beta/2+1-\alpha)} $$

$$ =\frac{\Gamma(\beta-\alpha+1).(\beta/2)\,\Gamma(\beta/2)}{\beta\Gamma(\beta)\,\Gamma(\beta/2-\alpha+1)} $$

or $$ {}_2F_1(\alpha,\beta;\beta-\alpha+1;-1)=\frac{\Gamma(\beta-\alpha+1).\Gamma(\beta/2+1)}{\Gamma(\beta+1)\,\Gamma(\beta/2-\alpha+1)} $$

which is the required Kummer's theorem (or formula).

5.17 Gauss's Theorem

$$ {}_2F_1(\alpha,\beta;\gamma;1)=\frac{\Gamma(\gamma).\Gamma(\gamma-\beta-\alpha)}{\Gamma(\gamma-\alpha).\Gamma(\gamma-\beta)} $$

[Kanpur 2005, 06, 07; Meerut 2001, 03, 12 (Sem. I); Rajasthan 2000, 07, 08]

Proof: The integral formula for hypergeometric function is

$${}_2F_1(\alpha, \beta; \gamma; x) = \frac{\Gamma(\gamma)}{\Gamma(\beta)\,\Gamma(\gamma-\beta)} \int_0^1 t^{\beta-1}(1-t)^{\gamma-\beta-1}(1-xt)^{-\alpha}\,dt$$

Putting $x = 1$, we get

$${}_2F_1(\alpha, \beta; \gamma; 1) = \frac{\Gamma(\gamma)}{\Gamma(\beta)\,\Gamma(\gamma-\beta)} \int_0^1 t^{\beta-1}(1-t)^{\gamma-\beta-1}(1-t)^{-\alpha}\,dt$$

$$= \frac{\Gamma(\gamma)}{\Gamma(\beta)\,\Gamma(\gamma-\beta)} \int_0^1 t^{\beta-1}(1-t)^{(\gamma-\beta-\alpha)-1}\,dt$$

$$= \frac{\Gamma(\gamma)}{\Gamma(\beta)\,\Gamma(\gamma-\beta)}\, B(\beta, \gamma-\beta-\alpha)$$

$$= \frac{\Gamma(\gamma)}{\Gamma(\beta)\,\Gamma(\gamma-\beta)} \cdot \frac{\Gamma(\beta).\Gamma(\gamma-\beta-\alpha)}{\Gamma(\beta+\gamma-\beta-\alpha)}$$

or $${}_2F_1(\alpha, \beta; \gamma; x) = \frac{\Gamma(\gamma).\Gamma(\gamma-\beta-\alpha)}{\Gamma(\gamma-\alpha)\,\Gamma(\gamma-\beta)}$$

which is the required Gauss's theorem.

5.18 Vandermonde's Theorem

$${}_2F_1(-n, \beta; \gamma; 1) = \frac{(\gamma-\beta)_n}{(\gamma)_n}$$ **[Kanpur 2000, 01; Meerut 2001, 03]**

Proof: The integral formula for hypergeometric function is

$${}_2F_1(\alpha, \beta; \gamma; x) = \frac{\Gamma(\gamma)}{\Gamma(\beta)\,\Gamma(\gamma-\beta)} \int_0^1 t^{\beta-1}(1-t)^{\gamma-\beta-1}(1-xt)^{-\alpha}\,dt$$

Putting $x = 1$ and $\alpha = -n$, we have

$${}_2F_1(-n, \beta; \gamma; 1) = \frac{\Gamma(\gamma)}{\Gamma(\beta)\,\Gamma(\gamma-\beta)} \int_0^1 t^{\beta-1}(1-t)^{\gamma-\beta-1}(1-t)^{n}\,dt$$

$$= \frac{\Gamma(\gamma)}{\Gamma(\beta)\,\Gamma(\gamma-\beta)} \int_0^1 t^{\beta-1}(1-t)^{(\gamma+n-\beta)-1}\,dt$$

$$= \frac{\Gamma(\gamma)}{\Gamma(\beta)\,\Gamma(\gamma-\beta)}\, B(\beta, \gamma+n-\beta)$$

$$= \frac{\Gamma(\gamma)}{\Gamma(\beta)\,\Gamma(\gamma-\beta)} \cdot \frac{\Gamma(\beta).\Gamma(\gamma+n-\beta)}{\Gamma(\beta+\gamma+n-\beta)} = \frac{\Gamma(\gamma)\,\Gamma(\gamma-\beta+n)}{\Gamma(\gamma-\beta)\,\Gamma(\gamma+n)}$$

$$= \frac{\Gamma(\gamma)\,(\gamma-\beta+n-1)\,(\gamma-\beta+n-2)\ldots(\gamma-\beta)\,\Gamma(\gamma-\beta)}{\Gamma(\gamma-\beta)\,.\,(\gamma+n-1)\,(\gamma+n-2)\,\ldots.\,\gamma\,.\,\Gamma(\gamma)}$$

$$= \frac{(\gamma-\beta)\,(\gamma-\beta+1)\ldots(\gamma-\beta+n-2)\,(\gamma-\beta+n-1)}{\gamma\,(\gamma+1)\,\ldots.\,(\gamma+n-2)\,(\gamma+n-1)}$$

or $\;{}_2F_1\,(-n,\beta;\gamma;1) = \dfrac{(\gamma-\beta)_n}{(\gamma)_n}$

which is the required Vandermonde's theorem.

5.19 Confluent Hypergeometric Equation (Or Kummer's Equation) and its Solution

The hypergeometric differential equation is

$$x\,(1-x)\,\frac{d^2y}{dx^2}\,\{\gamma-(\alpha+\beta+1)\,x\}\,\frac{dy}{dx}-\alpha\beta y=0 \qquad \ldots(1)$$

Replacing x by x/β in (1), we get

$$\frac{x}{\beta}\left(1-\frac{x}{\beta}\right)\beta^2\,\frac{d^2y}{dx^2}+\left\{\gamma-(\alpha+\beta+1)\,\frac{x}{\beta}\right\}\beta\,\frac{dy}{dx}-\alpha\beta y=0$$

$$\Rightarrow \quad x\left(1-\frac{x}{\beta}\right)\frac{d^2y}{dx^2}+\left\{\gamma-\left(1+\frac{1+\alpha}{\beta}\right)x\right\}\frac{dy}{dx}-\alpha y=0 \qquad \ldots(2)$$

Since ${}_2F_1\,(\alpha,\beta;\gamma;x)$ is the solution of (1), $\therefore\ {}_2F_1\,(\alpha,\beta;\gamma;x/\beta)$ is the solution of (2)

When $\beta \to \infty$, the equation (2) reduces to

$$x\,\frac{d^2y}{dx^2}+(\gamma-x)\,\frac{dy}{dx}-\alpha y=0 \qquad \ldots(3)$$

whose solution is given by

$$\lim_{\beta\to\infty}\,{}_2F_1\,(\alpha,\beta;\gamma;x/\beta)=\lim_{\beta\to\infty}\sum_{r=0}^{\infty}\frac{(\alpha)_r\,(\beta)_r}{(\gamma)_r}\cdot\frac{1}{r\,!}\left(\frac{x}{\beta}\right)^r$$

$$=\sum_{r=0}^{\infty}\frac{(\alpha)_r}{(\gamma)_r}\cdot\frac{x^r}{r\,!}={}_1F_1\,(\alpha;\gamma;x)$$

Since $\quad \displaystyle\lim_{\beta\to\infty}\frac{(\beta)_r}{\beta^r}=\lim_{\beta\to\infty}\frac{\beta\,(\beta+1)\,(\beta+2)\,\ldots.\,(\beta+r-1)}{\beta^r}$

$$= \lim_{\beta \to \infty} 1.\left(1+\frac{1}{\beta}\right)\left(1+\frac{2}{\beta}\right)\ldots\left(1+\frac{r-1}{\beta}\right) = 1$$

i.e., solution of equation (3) is ${}_1F_1\,(\alpha\,;\gamma;x)$

This function ${}_1F_1\,(\alpha\,;\gamma;x)$ is called **confluent hypergeometric function** which is the solution of the equation (3) called **Confluent hypergeometric equation or Kummer's equation.**

5.20 Series Solution of Confluent Hypergeometric Differential Equation near x = 0 when γ is not an Integer

The confluent hypergeometric differential equation is

$$x\frac{d^2y}{dx^2} + (\gamma - x)\frac{dy}{dx} - \alpha y = 0 \qquad \ldots(1)$$

Comparing with $\frac{d^2y}{dx^2} + P(x)\frac{dy}{dx} + Q(x)\,y = 0$, we get

$$P(x) = \frac{\gamma - x}{x} \text{ and } Q(x) = -\frac{\alpha}{x}$$

None of which is finite *i.e.*, analytic at $x = 0$ and $xP(x) = \gamma - x$ and $x^2Q(x) = -\alpha x$, which are both finite *i.e.*, analytic at $x = 0$.

So $x = 0$ is a regular singular point of (1).

Let the solution of (1), near the point $x = 0$, be given by the infinite series

$$y = \sum_{r=0}^{\infty} C_r\, x^{m+r}, \quad C_0 \neq 0 \qquad \ldots(2)$$

Diff. $$\frac{dy}{dx} = \sum_{r=0}^{\infty} C_r\,(m+r)\,x^{m+r-1}$$

and $$\frac{d^2y}{dx^2} = \sum_{r=0}^{\infty} C_r\,(m+r)\,(m+r-1)\,x^{m+r-2}$$

Substituting in (1), we get

$$\sum_{r=0}^{\infty} C_r\,[(m+r)\,(m+r-1)\,x^{m+r-1} + (\gamma - x)\,(m+r)\,x^{m+r-1} - \alpha\,x^{m+r}] = 0$$

or $$\sum_{r=0}^{\infty} C_r\,[(m+r)\,(m+r-1+\gamma)\,x^{m+r-1} - (m+r+\alpha)\,x^{m+r}] = 0 \qquad \ldots(3)$$

Which is an identity, so we can equate to zero the coefficients of various powers of x.

Equating to zero the coefficient of lowest power of x *i.e.*, of x^{m-1}, the *indicial equation* is

$$C_0 m(m-1+\gamma)=0 \Rightarrow m=0 \text{ or } 1-\gamma \qquad \because \quad C_0 = 0$$

which are unequal and not differing by an integer.

Now equating to zero the coefficient of x^{m+r-1}, we get

$$C_r (m+r)(m+r-1+\gamma)-(m+r-1+\alpha)\, C_{r-1}=0$$

$$\Rightarrow \qquad C_r = \frac{m+\alpha+r-1}{(m+r)(m+r-1+\gamma)}\, C_{r-1}, \qquad \text{for } r \geq 1 \qquad \ldots(4)$$

Case I : When $m = 0$:

In this case $C_r = \dfrac{\alpha+r-1}{r(r+\gamma-1)}\, C_{r-1}, \qquad r \geq 1$

Putting $\quad r = 1, 2, 3, \ldots\ldots\ldots,$

$$C_1 = \frac{\alpha}{\gamma} C_0,\ C_2 = \frac{\alpha+1}{2.(\gamma+1)} C_1 = \frac{\alpha(\alpha+1)}{1.2\,\gamma(\gamma+1)} C_0$$

$$C_3 = \frac{\alpha+2}{3(\gamma+2)} C_2 = \frac{\alpha(\alpha+1)(\alpha+2)}{1.2.3\,\gamma(\gamma+1)(\gamma+2)} C_0 \text{ and so on.}$$

Putting in (2), we get

$$y = \sum_{r=0}^{\infty} C_r x^r, \text{ when } m=0$$

$$= C_0 + C_1 x + C_2 x^2 + C_3 x^3 + \ldots.$$

or
$$y = C_0 \left[1 + \frac{\alpha}{\gamma}\cdot x + \frac{\alpha(\alpha+1)}{\gamma(\gamma+1)}\cdot\frac{x^2}{2!} + \frac{\alpha(\alpha+1)(\alpha+2)}{\gamma(\gamma+1)(\gamma+2)}\cdot\frac{x^3}{3!} + \ldots.\right]$$

Taking $\quad C_0 = 1$

$$y = \sum_{r=0}^{\infty} \frac{(\alpha)_r}{(\gamma)_r}\cdot\frac{x^r}{r!} = {}_1F_1(\alpha;\gamma;x)$$

which is a solution of (1) and is called the **confluent hypergeometric function of first kind.**

Case II. When $m = 1-\gamma$: In this case, from (4)

$$C_r = \frac{1-\gamma+\alpha+r-1}{(1-\gamma+r)(1-\gamma+r-1+\gamma)} C_0 = \frac{(\alpha-\gamma+1)+r-1}{r\{(2-\gamma)+r-1\}} C_0$$

$$\Rightarrow \qquad C_r = \frac{\alpha' + r - 1}{r\,(\gamma' + r - 1)}\,C_0 \qquad \text{for } r \geq 1$$

where $\alpha' = \alpha - \gamma + 1$ and $\gamma' = 2 - \gamma$

Putting $r = 1, 2, 3, \ldots\ldots\ldots$, we get

$$C_1 = \frac{\alpha'}{1.\gamma'}\,C_0,\ C_2 = \frac{\alpha' + 1}{2\,(\gamma' + 1)}\,C_1 = \frac{\alpha'\,(\alpha' + 1)}{1.2\,\gamma'\,(\gamma' + 1)}\,C_0$$

$$C_3 = \frac{\alpha' + 2}{3\,.(\gamma' + 2)}\,C_2 = \frac{\alpha'\,(\alpha' + 1)\,(\alpha' + 2)}{1.2\,.3\,\gamma'\,(\gamma' + 1)\,(\gamma' + 2)}\,C_0 \text{ and so on.}$$

$\therefore$ from (2), we have

$$y = \sum_{r=0}^{\infty} C_r x^{1-\gamma+r} = C_0 x^{1-\gamma} + C_1 x^{2-\gamma} + C_2 x^{3-\gamma} + \ldots..$$

$$= C_0 x^{1-\gamma}\left[1 + \frac{\alpha'}{\gamma'}\,\frac{x}{1!} + \frac{\alpha'\,(\alpha' + 1)}{\gamma'\,(\gamma' + 1)}\cdot\frac{x^2}{2\,!} + \frac{\alpha'\,(\alpha' + 1)\,(\alpha' + 2)}{\gamma'\,(\gamma' + 1)\,(\gamma' + 2)}\,\frac{x^3}{3\,!} + \ldots\right]$$

Taking $C_0 = 1$

$$y = x^{1-\gamma} \sum_{r=0}^{\infty} \frac{(\alpha')_r}{(\gamma')_r}\cdot\frac{x^r}{r\,!} = x^{1-\gamma}\,{}_1F_1\,(\alpha'\,;\gamma'\,;x)$$

i.e., $y = x^{1-\gamma}\,{}_1F_1\,(\alpha - \gamma + 1; 2 - \gamma; x)$

which is other independent solution of (1) and is called the **confluent hypergeometric function of second kind.**

Hence the general solution of confluent hypergeometric equation (1) is

$$y = A\,{}_1F_1(\alpha\,;\gamma;x) + \mathrm{B}\,x^{1-\gamma}\,{}_1F_1\,(\alpha - \gamma + 1; 2 - \gamma; x)$$

where $\gamma > 0$ is not integral and A, B are arbitrary constants.

5.21 Derivatives of Confluent Hypergeometric Function

The confluent hypergeometric function is given by

$${}_1F_1\,(\alpha\,;\gamma;x) = \sum_{n=0}^{\infty} \frac{(\alpha)_n}{(\gamma)_n}\cdot\frac{x^n}{n!}$$

Differentiating both sides w.r.t. x, we get

$$\frac{d}{dx}\,{}_1F_1\,(\alpha\,;\gamma;\,x) = \sum_{n=0}^{\infty} \frac{(\alpha)_n}{(\gamma)_n}\cdot\frac{n\,x^{n-1}}{n!} = \sum_{n=1}^{\infty} \frac{(\alpha)_n}{(\gamma)_n}\cdot\frac{x^{n-1}}{(n-1)!}$$

$$= \sum_{m=0}^{\infty} \frac{(\alpha)_{m+1}}{(\gamma)_{m+1}}\cdot\frac{x^m}{m!} \qquad \text{Putting } n-1=m$$

$$= \frac{\alpha}{\gamma}\sum_{m=0}^{\infty} \frac{(\alpha+1)_m}{(\gamma+1)_m}\,\frac{x^m}{m!} \qquad [\because (\alpha)_{m+1} = \alpha\,(\alpha+1)_m]$$

$$\therefore \qquad = \frac{\alpha}{\gamma}\,{}_1F_1\,(\alpha+1;\gamma+1;\,x)$$

i.e., $$\boldsymbol{\frac{d}{dx}\,{}_1F_1\,(\alpha\,;\,\gamma;\,x) = \frac{\alpha}{\gamma}\,{}_1F_1\,(\alpha+1;\,\gamma+1;\,x)} \qquad \ldots(1)$$

Differentiating (1) again, w.r.t x we get

$$\frac{d^2}{dx^2}\,{}_1F_1\,(\alpha\,;\gamma;\,x) = \frac{\alpha}{\gamma}\sum_{m=0}^{\infty} \frac{(\alpha+1)_m}{(\gamma+1)_m}\,\frac{m\,x^{m-1}}{m!}$$

$$= \frac{\alpha}{\gamma}\sum_{m=1}^{\infty} \frac{(\alpha+1)_m}{(\gamma+1)_m}\cdot\frac{x^{m-1}}{(m-1)!}$$

$$= \frac{\alpha}{\gamma}\sum_{n=0}^{\infty} \frac{(\alpha+1)_{n+1}}{(\gamma+1)_{n+1}}\,\frac{x^n}{n!} \qquad \text{Taking } m-1=n$$

$$= \frac{\alpha}{\gamma}\sum_{n=0}^{\infty} \frac{(\alpha+1)\,(\alpha+2)_n}{(\gamma+1)\,(\gamma+2)_n}\cdot\frac{x^n}{n!} \quad [\because (\alpha+1)_{n+1} = (\alpha+1)\,(\alpha+2)_n]$$

$$= \frac{\alpha\,(\alpha+1)}{\gamma\,(\gamma+1)}\sum_{n=0}^{\infty} \frac{(\alpha+2)_n}{(\gamma+2)_n}\,\frac{x^n}{n!} = \frac{\alpha\,(\alpha+1)}{\gamma\,(\gamma+1)}\,{}_1F_1\,(\alpha+2\,,\gamma+2\,;x)$$

i.e., $$\frac{d^2}{dx^2}\,{}_1F_1\,(\alpha\,;\gamma;\,x) = \frac{\alpha\,(\alpha+1)}{\gamma\,(\gamma+1)}\,{}_1F_1\,(\alpha+2\,;\gamma+2\,;x) = \frac{(\alpha)_2}{(\gamma)_2}\,{}_1F_1\,(\alpha+2\,;\gamma+2\,;x) \qquad \ldots(2)$$

Proceeding similarly, we can assume that

$$\frac{d^p}{dx^p}\,{}_1F_1\,(\alpha\,;\gamma;\,x) = \frac{(\alpha)_p}{(\gamma)_p}\,{}_1F_1\,(\alpha+p\,;\gamma+p\,;x) \qquad \ldots(3)$$

where p is a positive integer.

Differentiating again, w.r.t. 'x', we get

$$\frac{d^{p+1}}{dx^{p+1}}\,{}_1F_1(\alpha;\gamma;x)=\frac{(\alpha)_p}{(\gamma)_p}\cdot\frac{d}{dx}\,{}_1F_1(\alpha+p;\gamma+p;x)$$

$$=\frac{(\alpha)_p}{(\gamma)_p}\cdot\frac{\alpha+p}{\gamma+p}\,{}_1F_1(\alpha+p+1;\gamma+p+1;x)$$

$$=\frac{(\alpha)_{p+1}}{(\gamma)_{p+1}}\,{}_1F_1(\alpha+p+1;\gamma+p+1;x)$$

$$\because\quad (\alpha)_p.(\alpha+p)=(\alpha)_{p+1}$$

i.e., if (3) is true for positive integer p, then it is also true for positive integer $p+1$. But from (1) it is true for $p=1$. Hence by mathematical induction, (3) is the for every positive integer n.

$$\frac{d^n}{dx^n}\,{}_1F_1(\alpha;\gamma;x)=\frac{(\alpha)_n}{(\gamma)_n}\cdot{}_1F_1(\alpha+n;\gamma+n;x)$$

5.22 Integral Formula for Confluent Hypergeometric Function ${}_1F_1(\alpha;\gamma;x)$

$${}_1F_1(\alpha;\gamma;x)=\frac{\Gamma(\gamma)}{\Gamma(\alpha)\,\Gamma(\gamma-\alpha)}\int_0^1(1-t)^{\gamma-\alpha-1}.t^{\alpha-1}e^{xt}dt$$

or $${}_1F_1(\alpha;\gamma;x)=\frac{1}{B(\alpha,\gamma-\alpha)}\int_0^1(1-t)^{\gamma-\alpha-1}.t^{\alpha-1}e^{xt}dt$$

[Rajasthan 2003; Purvanchal 2006]

Proof: We have, the confluent hypergeometric function

$${}_1F_1(\alpha;\gamma;x)=\sum_{n=0}^{\infty}\frac{(\alpha)_n}{(\gamma)_n}\cdot\frac{x^n}{n!}\qquad\ldots(1)$$

Since $(\alpha)_n=\dfrac{\Gamma(\alpha+n)}{\Gamma(\alpha)}$ (See 5.15)

$$\therefore\quad\frac{(\alpha)_n}{(\gamma)_n}=\frac{\Gamma(\alpha+n)}{\Gamma(\alpha)}\cdot\frac{\Gamma(\gamma)}{\Gamma(\gamma+n)}=\frac{\Gamma(\gamma)}{\Gamma(\alpha)\,\Gamma(\gamma-\alpha)}\cdot\frac{\Gamma(\gamma-\alpha)\,\Gamma(\alpha+n)}{\Gamma(\gamma+n)}$$

$$=\frac{\Gamma(\gamma)}{\Gamma(\alpha)\,\Gamma(\gamma-\alpha)}\cdot\frac{\Gamma(\gamma-\alpha)\,\Gamma(\alpha+n)}{\Gamma(\gamma-\alpha+\alpha+n)}=\frac{\Gamma(\gamma)}{\Gamma(\alpha)\,\Gamma(\gamma-\alpha)}\cdot B(\alpha+n,\gamma-\alpha)$$

$$=\frac{\Gamma(\gamma)}{\Gamma(\alpha)\,\Gamma(\gamma-\alpha)}\int_0^1(1-t)^{\gamma-\alpha-1}\,t^{\alpha+n-1}\,dt$$

$\therefore$ from (1), we have

$$ {}_1F_1\,(\alpha\,;\gamma;x) = \sum_{n=0}^{\infty} \frac{\Gamma(\gamma)}{\Gamma(\alpha)\,\Gamma(\gamma-\alpha)} \int_0^1 (1-t)^{\gamma-\alpha-1}\, t^{\alpha+n-1} \cdot \frac{x^n}{n!}\,dt $$

$$ = \frac{\Gamma(\gamma)}{\Gamma(\alpha)\,\Gamma(\gamma-\alpha)} \int_0^1 (1-t)^{\gamma-\alpha-1}\, t^{\alpha-1} \left(\sum_{n=0}^{\infty} \frac{(x\,t)^n}{n!} \right) dt $$

$$ = \frac{\Gamma(\gamma)}{\Gamma(\alpha)\,\Gamma(\gamma-\alpha)} \int_0^1 (1-t)^{\gamma-\alpha-1}\, t^{\alpha-1} \,.\, e^{xt}\,dt $$

or $$ {}_1F_1\,(\alpha\,;\gamma;x) = \frac{1}{\mathrm{B}\,(\alpha\,,\gamma-\alpha)} \cdot \int_0^1 (1-t)^{\gamma-\alpha-1} . t^{\alpha-1} e^{xt}\,dt $$

5.23 Kummer's Theorem (for Confluent Hypergeometric Function)

$$ {}_1\boldsymbol{F}_1\,(\alpha\,;\,\beta;\,\boldsymbol{x}) = \boldsymbol{e}^{\boldsymbol{x}}\ {}_1\boldsymbol{F}_1\,(\beta-\alpha\,;\,\beta;\,-\boldsymbol{x}) $$

[Kanpur 2000, 01; Rajasthan 2002, 03]

Proof: Integral formula for confluent hypergeometric function is :

$$ {}_1F_1\,(\alpha\,;\beta;x) = \frac{\Gamma(\beta)}{\Gamma(\alpha)\,\Gamma(\beta-\alpha)} \int_0^1 (1-t)^{\beta-\alpha-1}\, t^{\alpha-1} \,.\, e^{xt}\,dt \qquad \ldots(1) $$

Replacing α by $\beta-\alpha$ and x by $-x$, we get

$$ {}_1F_1\,(\beta-\alpha\,;\beta;x) = \frac{\Gamma(\beta)}{\Gamma(\beta-\alpha)\,\Gamma[\beta-(\beta-\alpha)]} \int_0^1 (1-t)^{\beta-(\beta-\alpha)-1} t^{\beta-\alpha-1} e^{-xt}\,dt $$

$$ = \frac{\Gamma(\beta)}{\Gamma(\beta-\alpha)\,\Gamma(\alpha)} \int_0^1 (1-t)^{\alpha-1}\, t^{\beta-\alpha-1}\, e^{-xt}\,dt $$

$$ = \frac{\Gamma(\beta)}{\Gamma(\beta-\alpha)\,\Gamma(\alpha)} \int_1^0 u^{\alpha-1}\,(1-u)^{\beta-\alpha-1} e^{-x\,(1-u)}\,(-1)\ du $$

Putting $1-t=u,\ dt=-du$

$$ = e^{-x}\, \frac{\Gamma(\beta)}{\Gamma(\alpha)\,\Gamma(\beta-\alpha)} \int_0^1 (1-u)^{\beta-\alpha-1}\, u^{\alpha-1}\, e^{xu}\,du $$

$$ = e^{-x}\, \frac{\Gamma(\beta)}{\Gamma(\alpha)\,\Gamma(\beta-\alpha)} \int_0^1 (1-t)^{\beta-\alpha-1}\, t^{\alpha-1}\, e^{xt}\,dt $$

By Prop of D.I.

$$ = e^{-x}\ F(\alpha\,;\beta;x) \text{ from } (1) $$

$$ \Rightarrow \quad {}_1F_1\,(\alpha\,;\beta;x) = e^{x}\ {}_1F_1\,(\beta-\alpha\,;\beta;\ x) $$

5.24 Whittaker's Confluent Hypergeometric Function

The Confluent Hypergeometric equation is

$$x\frac{d^2y}{dx^2}+(\gamma-x)\frac{dy}{dx}-\alpha y=0 \qquad \ldots(1)$$

whose general solution is

$$y=Ay_1(x)+By_2(x) \qquad \ldots(2)$$

where A and B are arbitrary constants and

$$y_1(x)={}_1F_1\,(\alpha\,;\gamma;x) \qquad \ldots(3)$$

and $$y_2(x)=x^{1-\gamma}\;{}_1F_1\,(\alpha-\gamma+1;2-\gamma;x) \qquad \ldots(4)$$

Putting $$y(x)=x^{-\gamma/2}e^{x/2}W(x) \qquad \ldots(5)$$

so that, $$\frac{dy}{dx}=\left\{-\left(\frac{\gamma}{2}\right)x^{-\gamma/2-1}e^{x/2}+\frac{1}{2}x^{-\gamma/2}e^{x/2}\right\}W(x)+x^{-\gamma/2}e^{x/2}\frac{dW}{dx}$$

$$\frac{d^2y}{dx^2}=\left\{-(\gamma/2).x^{-\gamma/2-1}e^{x/2}+\frac{1}{2}x^{-\gamma/2}e^{x/2}\right\}\frac{dW}{dx}$$

$$+\left\{\left(\frac{\gamma}{2}\right)\left(\frac{\gamma}{2}+1\right)e^{-(\gamma/2)-2}.e^{x/2}-\left(\frac{\gamma}{2}\right)x^{-(\gamma/2)-1}.\frac{1}{2}e^{x/2}\right.$$

$$\left.-\frac{\gamma}{4}x^{-(\gamma/2)-1}e^{x/2}+\frac{1}{4}x^{-\gamma/2}e^{x/2}\right\}W(x)$$

$$+x^{-\gamma/2}e^{x/2}\frac{d^2W}{dx^2}+\left\{\left(-\frac{\gamma}{2}\right)x^{-(\gamma/2)-1}e^{x/2}+\frac{1}{2}x^{-\gamma/2}e^{x/2}\right\}\frac{dW}{dx}$$

$$=\{-\gamma\,x^{-\gamma/2-1}e^{x/2}+x^{-\gamma/2}e^{x/2}\}\frac{dW}{dx}$$

$$+\left\{\left(\gamma\frac{(\gamma+2)}{4}\right)x^{-(\gamma/2)-2}e^{x/2}-\left(\frac{\gamma}{2}\right)x^{-(\gamma/2)-1}e^{x/2}+\left(\frac{1}{4}\right)e^{-\gamma/2}e^{x/2}\right\}.$$

$$W(x)+x^{-\gamma/2}e^{x/2}\frac{d^2W}{dx^2}$$

in (1), it reduces to

$$x\,.\,x^{-\gamma/2}e^{x/2}\left[\frac{d^2W}{dx^2}+(-\gamma\,x^{-1}+1)\frac{dW}{dx}+\left\{\left(\gamma\frac{(\gamma+2)}{4}\right)x^{-2}-\left(\frac{\gamma}{2}\right)x^{-1}+\frac{1}{4}\right\}W(x)\right]$$

$$+(\gamma-x)\,.\,x^{-\gamma/2}e^{x/2}\left[\frac{dw}{dx}+\left\{\left(-\frac{\gamma}{2}\right)x^{-1}+\frac{1}{2}\right\}W(x)\right]-\alpha\,x^{-\gamma/2}e^{x/2}\,W(x)=0$$

or $$x\frac{d^2W}{dx^2}+\left[\left(\frac{\gamma(\gamma+2)}{4}\right)x^{-1}-\frac{\gamma}{2}+\frac{x}{4}+(\gamma-x)\left\{\left(-\frac{\gamma}{2}\right)x^{-1}+\frac{1}{2}\right\}-\alpha\right]W(x)=0$$

or $$\frac{d^2W}{dx^2}+\left[-\frac{1}{4}+\frac{\gamma/2-\alpha}{x}-\frac{(-\gamma^2/4+\gamma/2)}{x^2}\right]W(x)=0$$

or $$\frac{d^2W}{dx^2}+\left[-\frac{1}{4}+\frac{\gamma/2-\alpha}{x}+\frac{1/4-(\gamma/2-1/2)^2}{x^2}\right]W(x)=0$$

or $$\frac{d^2W}{dx^2}+\left[-\frac{1}{4}+\frac{k}{x}+\frac{1/4-m^2}{x^2}\right]W(x)=0 \qquad \text{....(6)}$$

where $k=\gamma/2-\alpha$ and $m=\gamma/2-1/2$

The solutions $W(x)$ of (6) are called the Whittaker's Confluent Hypergeometric Functions

If $2m$ is neither 1 nor an integer then the solution of equation (6) is given by (2) where $\gamma=2m+1$ and $\alpha=\gamma/2-k=1/2+m-k$

Thus the solutions of (6) *i.e.*, **Whittaker's Confluent hypergeometric functions** are given by

$$M_{k,\,m}(x)=x^{\gamma/2}\,e^{-x/2}\,y_1(x)$$

$$=x^{m+1/2}\,.e^{-x/2}\,{}_1F_1\,(1/2+m-k;2m+1;x)$$

and $$M_{k,\,-m}(x)=x^{\gamma/2}e^{-x/2}\,y_2(x)$$

$$=x^{1/2-m}\,.e^{-x/2}\,{}_1F_1\,(1/2-m-k;1-2m;x)$$

5.25 Contiguous Hypergeometric Functions

According to Gauss, the function $F(\alpha';\beta';\gamma';x)$ is said to be contiguous to function $F(\alpha,\beta;\gamma;x)$ if it is obtained by increasing or decreasing one and only one of the parameters α, β, γ by unity.

Thus there exist six hypergeometric functions contiguous to $F(\alpha,\beta;\gamma;x)$ and are denoted as given below

$$F(\alpha+1,\beta;\gamma;x)=F_{\alpha+}, \qquad F(\alpha-1,\beta;\gamma;x)=F_{\alpha-}$$
$$F(\alpha,\beta+1;\gamma;x)=F_{\beta+}, \qquad F(\alpha,\beta-1;\gamma;x)=F_{\beta-}$$
$$F(\alpha,\beta;\gamma+1;x)=F_{\gamma+}, \qquad F(\alpha,\beta;\gamma-1;x)=F_{\gamma-}$$

5.26 Theorem : Contiguity Relationship

There exists a linear relation with polynomial coefficients between a hypergeometric function $F(\alpha, \beta; \gamma; x)$ and its any two contiguous hypergeometric functions

i.e., $(\alpha - \beta)\, F(\alpha, \beta; \gamma; x) = \alpha . F_{\alpha +} - \beta . F_{\beta +}$ ***(etc.)***

Which holds for $|x| < 1$.

Proof: We have,

$$\alpha . F_{\alpha+} - \beta . F_{\beta+} = \alpha . F(\alpha + 1, \beta; \gamma; x) - \beta . F(\alpha, \beta + 1; \gamma; x)$$

$$= \alpha \cdot \frac{(\alpha + 1)_r (\beta)_r}{(\gamma)_r} \cdot \frac{x^r}{r\,!} - \beta \cdot \frac{(\alpha)_r (\beta + 1)_r}{(\gamma)_r} \cdot \frac{x^r}{r\,!}$$

$$= \Big[\, [\alpha\, (\alpha + 1)]_r\, (\beta)_r - (\alpha)_r\, \beta\, (\beta + 1)_r \Big] \cdot \frac{x^r}{(\gamma)_r\, r\,!}$$

$$= \Big[[(\alpha)_r\, .\, (\alpha + r)\, .\, (\beta)_r - (\alpha)_r\, (\beta)_r\, (\beta + r)] \Big] \frac{x^r}{(\gamma)_r\; r\,!}$$

$\because$ $$\alpha\, (\alpha + 1)_r = \alpha\, (\alpha + 1)\, (\alpha + 2) \ldots .\, (\alpha + 1 + r - 1)$$

$$= \{\alpha\, (\alpha + 1)\, (\alpha + 2) \ldots .\, (\alpha + r - 1)\}\, (\alpha + r)$$

$$= (\alpha)_r . (\alpha + r)$$

Similarly, $\beta\, (\beta + 1)_r = (\beta)_r (\beta + r)$

$\therefore$ $$\alpha . F_{d+} - \beta . F_{\beta+} = [(\alpha + r) - (\beta + r)]\, \frac{(\alpha)_r (\beta)_r}{(\gamma)_r} \cdot \frac{x^r}{r\,!}$$

$$= (\alpha - \beta)\, \frac{(\alpha)_r\, .\, (\beta)_r}{(\gamma)_r} \cdot \frac{x^r}{r\,!} = (\alpha - \beta)\; F\, (\alpha, \beta; \gamma; x)$$

Solved Examples

Example 1: ***Show that*** $(1 + n)_n = 4^n\; (1/2)_n$

Solution: $$(1 + n)_n = (1 + n)\, (1 + n + 1)\, (1 + n + 2) \ldots .\, (1 + n + n - 1)$$

$$= (n + 1)\, (n + 2)\, (n + 3) \ldots .\, (2n)$$

$$= \frac{n!}{n!} \cdot (n + 1)\, (n + 2) \ldots (2n) = \frac{2n!}{n!}$$

$$= \frac{2n\, (2n - 1)\, (2n - 2)\, (2n - 3) \ldots 4 . 3 . 2 . 1}{n!}$$

& (%B

data

20211015/// (20211014)/// Cont.

→ 20211014/Circle/

Circle in a box ... (2009).pdf

→ Gauss – 45 occurrences

– Renormalization and 3-Manifolds ... (1996).pdf

Pag. 46(52) –

The injectivity radius of X is bounded above in terms of S (because its area is determined by Gauss-Bonnet), so we obtain an upper bound on the injectivity radius in the convex core of M.

– New Approaches to Circle Packing in a square ... pdf

– Barry Simon – The $P(\phi)_2$ Euclidean .. (1974).pdf

Gauss – 123 occurrences

Pag. 15(37) Definition

Let H be a real Hilbert space

$\sqrt{2}$ Gaussian random process
index

! -256- - -* ! - ! D L ! ! !
5 ! & 2 ! ! ! ! 3 / 1"

D -256- - -* ! - 5 ! ,! ! " % ! "
@ < D A H A 3 //B 1"
('B $ K A 3 !"

through every point of X there is a geodesic segment which is mapped isometrically to a geodesic segment in M. The injectivity radius of X is bounded above in terms of S (because its area is determined by Gauss-Bonnet), so we obtain an upper bound on the injectivity radius in the convex core of M.

To find sufficiently many pleated surfaces, one interpolates between pairs of such surfaces by interpolating between their pleating laminations. This idea completes the proof when M is quasifuchsian, since the boundary faces of the convex core are pleated surfaces. In general one must establish that the ends of the thick part of the convex core of M are *geometrically tame*, meaning any neighborhood of the end contains a pleated surface. This is established for $M \in \overline{QF(S)}$ in [Th1] and for all $M \in AH(S)$ in [Bon]. (It is conjectured that these two sets of manifolds are actually the same.) A detailed proof of the Theorem above appears in [Can, Thm. 6.2].

The same discussion leads to a proof of:

Theorem 3.5 (Tameness) *Any M in $AH(S)$ is homeomorphic to* $\mathrm{int}(S) \times \mathbb{R}$.

3.3 The mapping class group

The *mapping class group* $\mathrm{Mod}(S)$ is the group of isotopy classes of orientation-preserving homeomorphisms from S to itself. It is canonically isomorphic to the group of outer automorphisms of $\pi_1(S)$ that preserve orientation and stabilize the set of conjugacy classes of boundary curves.

A mapping class $\psi : S \to S$ acts on $\mathrm{Teich}(S)$ by sending (f, X) to $(f \circ \psi^{-1}, X)$. There are similar actions on $\mathrm{Teich}(\overline{S})$ and on $AH(S)$. These actions are compatible with the mapping

$$Q : \mathrm{Teich}(S) \times \mathrm{Teich}(\overline{S}) \to AH(S);$$

that is, $\psi(Q(X, Y)) = Q(\psi(X), \psi(Y))$.

A mapping class also acts on a Bers slice B_Y by $Q(X, Y) \mapsto Q(\psi(X), Y)$. The action of $\mathrm{Mod}(S)$ on B_Y is *not* the restriction of its action on $AH(S)$; the latter sends B_Y to $B_{\psi(Y)}$.

A mapping class ψ is *reducible* if there is a finite system of disjoint essential simple closed curves on S, none parallel to ∂S, which are

-256- - -* ! - 8 211 ! ! .
! > ! @ ! " 1 *1 ,! ! " & 211 ! 1"
($ %B

$$\frac{\left\lfloor \frac{2\sigma}{\sqrt{3}} \right\rfloor}{2}\left(\left\lfloor \sigma - \frac{1}{2}\right\rfloor + \lfloor\sigma\rfloor + 2\right) + \lfloor\sigma\rfloor + 1, \quad \text{if } \left\lfloor\frac{2\sigma}{\sqrt{3}}\right\rfloor + 1 \text{ is odd.}$$

In both cases, it is easy to see that we have placed at least $\frac{2}{\sqrt{3}}\sigma^2$ points in total, as because of $\lfloor \sigma - \frac{1}{2}\rfloor + \lfloor\sigma\rfloor + 2 > 2\sigma$, $\left\lfloor \frac{2\sigma}{\sqrt{3}}\right\rfloor > \frac{2}{\sqrt{3}}\sigma - 1$, and $\lfloor\sigma\rfloor > \sigma - 1$, we get

$$\frac{\left\lfloor \frac{2\sigma}{\sqrt{3}} \right\rfloor + 1}{2}\left(\lfloor\sigma\rfloor + \left\lfloor \sigma - \frac{1}{2}\right\rfloor + 2\right) > \frac{\frac{2\sigma}{\sqrt{3}} - 1 + 1}{2}2\sigma = \frac{2}{\sqrt{3}}\sigma^2$$

in the even case, and

$$\frac{\left\lfloor \frac{2\sigma}{\sqrt{3}} \right\rfloor}{2}\left(\left\lfloor \sigma - \frac{1}{2}\right\rfloor + \lfloor\sigma\rfloor + 2\right) + \lfloor\sigma\rfloor + 1 > \frac{\frac{2\sigma}{\sqrt{3}} - 1}{2}2\sigma + \sigma - 1 + 1 = \frac{2}{\sqrt{3}}\sigma^2$$

in the odd case, respectively. □

Theorem 3.1 supplies lower bounds of general validity; that is, the inequality (3.1) is applicable for each $n \geq 2$ integer. For individual n values, computer methods, say, can be used to deliver good lower bounds for the optimum values—provided that either a geometric construction with exact values is derived from the numerical solutions, or the mathematical correctness of the numerical bounds is *validated* (verified) by some guaranteed reliability numerical method. A stochastic algorithm for finding approximate numerical solutions of the point arrangement problems is introduced in Chapter 4.

Besides this, it is known that for several values of n, the optimal solutions are regular arrangements. Keeping this observation in mind, one can speculate that other types of regular arrangements may also provide good lower bounds for the optimum values of Problem 2.2 (see also as (2.1)). Indeed, a detailed investigation of these special patterns resulted in further improvements on the theoretical lower bounds for several n values. These results will be discussed in Chapter 10.

3.2 Upper bounds for the optimal radius $\overline{r}_n$

As one might expect, it is somewhat harder to determine non-trivial upper bounds for $\overline{m}_n$ and $\overline{r}_n$ than those for lower bounds. The main difficulty for this is that lower bounds can be obtained in a constructive way by trying out proper approximate (feasible) solutions, but for the upper bounds some knowledge about the properties of the optimal solutions is required beforehand.

As we already mentioned in Subsection 1.2.5, in [21] it was proved that the densest packing of circles in the plane is a regular (hexagonal) arrangement, and the density of this packing is $\pi/\sqrt{12} \approx 0.9068996821$. Moreover, it can be proved that this value is an upper bound for the density of any circle packings in the unit square (see Chapter 11). However, when Problem 2.2 itself is investigated, this general upper bound can be substantially improved:

Theorem 3.2. *For each $n \geq 2$ integer, an upper bound of the $\overline{r}_n$ optimal radius is*

$$\overline{r}_n \leq \min\left(U_1(n), U_2(n)\right),$$

with

$$U_1(n) := \frac{1}{\sqrt{2\sqrt{3}n + (4\lfloor\sqrt{n}\rfloor - 2)(2-\sqrt{3})}},$$

and

$$U_2(n) := \frac{1 + \sqrt{1 + \frac{2}{\sqrt{3}}(n-1)}}{2n + 2\sqrt{1 + \frac{2}{\sqrt{3}}(n-1)}},$$

where $\lfloor\cdot\rfloor$ denotes the greatest integer not greater than the argument.

PROOF. The proof is in two parts. First we show that $U_1(n)$ is the upper bound of $\overline{r}_n$, then we do likewise with $U_2(n)$.

- $U_1(n)$: Suppose that for a fixed $n \geq 2$ an optimal circle packing in the unit square $[0,1]^2$ exists, and that the common radius of the circles is $\overline{r}_n$. Partition the square into *Dirichlet cells* (*Voronoi polygons*) according to the centres of the circles: the interior of each cell consists of exactly those points of the square which are closer to a circle centre than to any other. Then, each cell belonging to a circle centre will contain the respective circle itself as well. Obviously the total area of the n cells will be equal to 1.

  We will bound the area of the individual cells from below. To achieve this, we divide the cells into two groups, depending on whether or not they share a point with one of the sides of the square.

  Case 1: The cell does not share a point with a side of the square. Then the area of a regular hexagon drawn around the circle in the cell provides a lower bound for the area of the cell. This follows from the following facts. First, among the polygons with k sides drawn around a circle, the regular polygon has the smallest area. Secondly, out of the equilateral triangle, the square, the regular pentagon and the regular hexagon drawn around a circle, the hexagon has the smallest area. In a special case, the area of the cell equals the area of the respective regular hexagon—namely when a circle of the packing is surrounded by six other circles. From this it follows that for Case 1 a lower bound for the area of the cells is the area of a regular hexagon of side $\overline{r}_n$, that is $2\sqrt{3}\overline{r}_n^2$.

  Case 2: The cell shares a point with one or more sides of the square. Although the lower bound obtained in case 1 is also valid for these types of cells, this bound can be improved because the cell makes contact with at least one side of the square: First, notice that the convex hull of the circle centres of the packing is located in the square $[\overline{r}_n, 1-\overline{r}_n]^2$. Let us denote that region of the unit square by R, which is outside of this inner square, i.e. $R := [0,1]^2 \setminus [\overline{r}_n, 1-\overline{r}_n]^2$. It is easy to see

in the odd case, respectively. □

Theorem 3.1 supplies lower bounds of general validity; that is, the inequality (3.1) is applicable for each $n \geq 2$ integer. For individual n values, computer methods, say, can be used to deliver good lower bounds for the optimum values—provided that either a geometric construction with exact values is derived from the numerical solutions, or the mathematical correctness of the numerical bounds is *validated* (verified) by some guaranteed reliability numerical method. A stochastic algorithm for finding approximate numerical solutions of the point arrangement problems is introduced in Chapter 4.

Besides this, it is known that for several values of n, the optimal solutions are regular arrangements. Keeping this observation in mind, one can speculate that other types of regular arrangements may also provide good lower bounds for the optimum values of Problem 2.2 (see also as (2.1)). Indeed, a detailed investigation of these special patterns resulted in further improvements on the theoretical lower bounds for several n values. These results will be discussed in Chapter 10.

3.2 Upper bounds for the optimal radius $\overline{r}_n$

As one might expect, it is somewhat harder to determine non-trivial upper bounds for $\overline{m}_n$ and $\overline{r}_n$ than those for lower bounds. The main difficulty for this is that lower bounds can be obtained in a constructive way by trying out proper approximate (feasible) solutions, but for the upper bounds some knowledge about the properties of the optimal solutions is required beforehand.

As we already mentioned in Subsection 1.2.5, in [21] it was proved that the densest packing of circles in the plane is a regular (hexagonal) arrangement, and the density of this packing is $\pi/\sqrt{12} \approx 0.9068996821$. Moreover, it can be proved that this value is an upper bound for the density of any circle packings in the unit square (see Chapter 11). However, when Problem 2.2 itself is investigated, this general upper bound can be substantially improved:

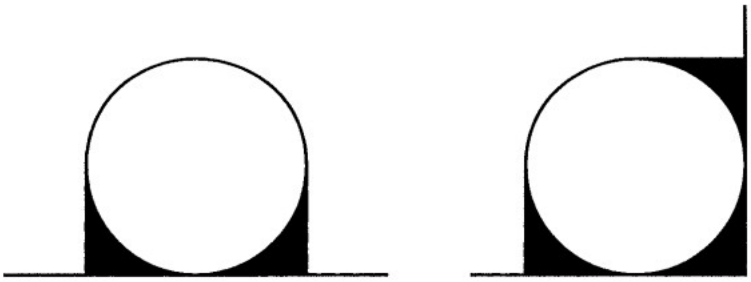

Fig. 3.2. Gaps occurring for circles touching the sides of the square.

that the densest possible way of packing circles into R is to place $4\sqrt{n}-4$ circles on the sides of the unit square.

In this packing, the non-covered regions of R have two different types of 'gaps': the regions defined by two circles and one side of the square, and the regions defined by two sides and a circle located in the corner of $[0,1]^2$ (see Figure 3.2). The sum of the non-covered areas is a lower bound for the area of the non-covered regions of R *for every optimal circle packing* in $[0,1]^2$. A gap increases the estimated lower bound of a regular hexagon by $(2-\sqrt{3})\overline{r}_n^2$. The overall total gap is $(4\lfloor\sqrt{n}\rfloor-2)(2-\sqrt{3})\overline{r}_n^2$. Then the sum of the lower bounds of Dirichlet cells is less than 1 (the area of the unit square), so

$$\overline{r}_n \le \frac{1}{\sqrt{2\sqrt{3}n+(4\lfloor\sqrt{n}\rfloor-2)(2-\sqrt{3})}}.$$

- $U_2(n)$: Using the Oler Theorem [90], J. H. Folkman and R. L. Graham proved the following statement [29]: if X is a compact convex subset of the plane, then the number of points in X with a mutual distance of at least 1 cannot be more than

$$\frac{2}{\sqrt{3}}A(X)+\frac{1}{2}P(X)+1, \tag{3.2}$$

where $A(X)$ is the area, and $P(X)$ is the perimeter of X, respectively.

Applying this result to an $A'(1,\overline{\sigma}_n)$ associated point arrangement (that is, for $X=[0,\overline{\sigma}_n]^2$, $A(X)=\overline{\sigma}_n^2$, and $P(X)=4\overline{\sigma}_n$), we obtain

$$n \le \frac{2}{\sqrt{3}}\overline{\sigma}_n^2+2\overline{\sigma}_n+1.$$

From the relation $\overline{\sigma}_n=(1-2\overline{r}_n)/2\overline{r}_n$ of Corollary 2.15, the lower bound

$$\overline{r}_n \leq \frac{1+\sqrt{1+\frac{2}{\sqrt{3}}(n-1)}}{2n+2\sqrt{1+\frac{2}{\sqrt{3}}(n-1)}}$$

follows, i.e. $U_2(n)$ is also a valid upper bound for $\overline{r}_n$. This completes the proof.

□

T. Tarnai and Zs. Gáspár in some special cases improved the previous upper bound in [122].

3.3 Asymptotic estimation and its error

Based on Corollary 2.15 $\left(\overline{m}_n = \frac{2\overline{r}_n}{1-2\overline{r}_n}\right)$, Theorems 3.1 and 3.2 suggest the following inequality bounds:

$$\sqrt{\frac{2}{\sqrt{3}n}} \leq \overline{m}_n \leq \frac{1}{n-1} + \sqrt{\frac{1}{(n-1)^2} + \frac{2}{\sqrt{3}(n-1)}}.$$

If we multiply these inequalities by $\sqrt{n}$, and let n tend to infinity, then

$$\lim_{n\to\infty} \sqrt{n}\overline{m}_n = \sqrt{\frac{2}{\sqrt{3}}}$$

holds, and this means $\overline{m}_n$ is asymptotically equal to $\sqrt{\frac{2}{\sqrt{3}n}}$ [14].

Theorem 3.3. *An upper bound of the absolute error for the* $\overline{m}_n$ $(\approx \sqrt{\frac{2}{\sqrt{3}n}})$ *asymptotic estimation is*

$$\frac{2}{n-1}.$$

PROOF. The above inequality

$$\frac{1+\sqrt{1+\frac{2}{\sqrt{3}}(n-1)}}{n-1} - \sqrt{\frac{2}{\sqrt{3}n}} < \frac{2}{n-1}$$

is equivalent to

$$\sqrt{1+\frac{2}{\sqrt{3}}(n-1)} - \sqrt{\frac{2(n-1)^2}{\sqrt{3}n}} < 1.$$

After a short calculation we obtain the following relation:

" -256- - -* ! -)! 3 9 * = "!
H A / ' 1"
($% 4 ? :D ! ! : D 1! >! 7! ! !" 1

LEMMA I.5. *Let* $\{f_\alpha\}_{\alpha\epsilon I}$ *be a family of random variables on* (M, Σ, μ). *Let* $\mathfrak{A} = \{F(f_{\alpha_1}, \ldots, f_{\alpha_n}) \mid F \epsilon \mathcal{S}(R^n), \alpha_i \epsilon I\}$. *The following are equivalent*:

(1) $\{f_\alpha\}_{\alpha\epsilon I}$ *is full*

(2) $\mathfrak{A}$, *as a subset of* $L^2(M, d\mu)$, *is dense*

(3) *The von Neumann algebra* $\mathfrak{A}''$, *generated by* $\mathfrak{A}$ *viewed as a family of all multiplication operators, is* $L^\infty(M, d\mu)$.

Proof. We concentrate on (1) $\Longrightarrow$ (2) which is what we will use below. The rest is fairly standard (but somewhat tedious). We will show (1) $\Longrightarrow$ (3) and (3) $\Longrightarrow$ (2). For suppose $\{f_\alpha\}_{\alpha\epsilon I}$ is full and let S be the family of projections in $\mathfrak{A}''$. Each of these projections is also a multiplication operator so S is a subset of $\Sigma/\mathfrak{I}_\mu$. It is easy to see it is a submeasure algebra and that each f_α is measurable with respect to S. Thus S is all of $\Sigma/\mathfrak{I}_\mu$ so $\mathfrak{A}''$ is all multiplication operators. To prove that (3) $\Longrightarrow$ (2) we note that $\overline{\mathfrak{A}1}$ contains $\mathfrak{A}''1$ since $\mathfrak{A}''$ is the strong closure of $\mathfrak{A}$. Thus $\overline{\mathfrak{A}1}$ contains L^∞ which is dense in L^2. ■

Remark:

Here and throughout our L^2 *spaces are always spaces of complex-valued functions*. Similarly $\mathcal{S}$ denotes complex-valued functions. We use $L^2_{\mathbf{R}}$ and $\mathcal{S}_{\mathbf{R}}$ to denote the spaces of real-valued functions.

A preliminary version of the second main object of this section is:

DEFINITION. A random process indexed by V is called a Gaussian random process indexed by V if and only if:

(a) $\{\phi(v) \mid v \epsilon V\}$ is full

(b) Each $\phi(v)$ is a g.r.v.

Remark:

Since V is a vector space, (b) implies that any $\phi(v_1), \ldots, \phi(v_n)$ are jointly Gaussian.

Given a Gaussian random process, the map $v, w \mapsto \langle\phi(v)\phi(w)\rangle$ defines a (semi-definite) inner product on V. In the usual way, we can quotient out by $\{v \mid <\phi(v)^2> = 0\}$ and complete to a Hilbert space, $\mathcal{H}$. It is not hard to see (after developing a part of the theory below!) that nothing is lost by supposing V to be a Hilbert space to begin with:

DEFINITION. Let $\mathcal{H}$ be a real Hilbert space. *The Gaussian random process indexed by* $\mathcal{H}$ is a random process indexed by $\mathcal{H}$ so that

(a) $\{\phi(v) \mid v \in \mathcal{H}\}$ is full.
(b) Each $\phi(v)$ is a g.r.v.
(c) $<\phi(v)\phi(w)> = \frac{1}{2} <v, w>$, the inner product on $\mathcal{H}$.

Remarks:

1. The $\frac{1}{2}$ in $<\phi(v)\phi(w)> = \frac{1}{2}<v, w>$ is somewhat unnatural. We add the $\frac{1}{2}$ so that when we consider the connection with Fock space in Section I.3, we can write $\phi(v) = \frac{1}{\sqrt{2}}(A^*(v) + A(v))$. If one suppressed the $\sqrt{2}$ in this last formula, one could eliminate the $\frac{1}{2}$ above, but we prefer to conform to the usual physicist's convention. Many authors (e.g., [90, 135]) do not include the $\frac{1}{2}$.
2. Notice that the phrase "Gaussian random process indexed by $\mathcal{H}$" means more if $\mathcal{H}$ is a Hilbert space than if $\mathcal{H}$ is merely a vector space. In Section I.6, we will have occasion to deal with two Gaussian processes indexed by the vector space, $\mathcal{H}$. We will then use the phrase "Gaussian random process with general covariance."

Of course, we will have to justify the use of the article "the" by proving a uniqueness theorem. In fact:

THEOREM I.6. *Let* $\mathcal{H}$ *be a real Hilbert space. Let* ϕ *and* ϕ' *be two Gaussian random processes indexed by* $\mathcal{H}$ *on probability measure spaces* (M, Σ, μ) *and* (M', Σ', μ') *respectively. Then there exists an isomorphism between the two probability measure spaces so that for every* $v \in \mathcal{H}$, $\phi(v)$ *corresponds to* $\phi'(v)$ *under the isomorphism.*

—

— Orthogonal Polynomials on the Unit Circle - Part 1 ...(2009).pdf
Pág. 17(44) — (1.2.43)

One place that the Christoffel function enters is in the Gauss-Jacobi or Gauss-Christoffel quadrature formula that

→ π a polynomial of degree 2n

Obs: Gauss - 32 occurences

— Orthogonal Polynomials on the Unit Circle - Part 2 ...(2004).djvu
Livro ref.
Pág. 991 (548)
[202] E. B. Christoffel,
Über die Gaussische Quadratur und eine Verallgemeinerung derselben,
J. Reine Angew. Math. 55 (1858), 61–82 (Cited on 24, 684).

— The Infinite in Mathematics
...(1978).pdf

—

%

! -256- - -* ! - % * ! 3 ! H

! ! ! 9 3 2 ! ! ! ! 3 ' "KA

-256- - -* ! - % * ! 3 ! H

! ! ! 9 3 2 ! ! ! ! 3 ' 1"

D -256- - -* ! - % * ! 3 ! H

! 1 ! 9 3 2 ! ! ! ! 3 ' "KA

-256- - -* ! - % * ! 3 ! H

! 1 ! 9 3 2 ! ! ! ! 3 ' 1"

*D

(" 7! > !" ! !

PROPOSITION 1.2.1. *Let $\varphi \in \mathcal{H}$, a Hilbert space with $\varphi \neq 0$. Then*

$$\min\{\|\psi\|^2 \mid \langle \psi, \varphi \rangle = 1\} = \frac{1}{\|\varphi\|^2}$$

The minimizing ψ is given by $\psi = \varphi/\|\varphi\|^2$. If ψ is restricted to lie in a closed subspace, $\mathcal{L}$, and P is the projection onto $\mathcal{L}$, then the minimum is $1/\|P\varphi\|^2$.

PROOF. Let $\psi_0 = \varphi/\|\varphi\|^2$. If $\langle \psi, \varphi \rangle = 1$, then $\langle \psi - \psi_0, \varphi \rangle = 0$ so $\|\psi\|^2 = \|\psi_0\|^2 + \|\psi - \psi_0\|^2$, so the minimum occurs at $\psi = \psi_0$. Since $\|\psi_0\|^2 = 1/\|\varphi\|^2$, the result is proven.

For the subspace problem, make $\mathcal{L}$ the Hilbert space and note that for $\psi \in \mathcal{L}$, $\langle \psi, \varphi \rangle = \langle \psi, P\varphi \rangle$ and $P\varphi \in \mathcal{L}$. □

For a moment, it is worthwhile relaxing our requirement that $\mu(\mathbb{R}) = 1$. For general μ, if $\tilde{\mu} = \mu/\mu(\mathbb{R})$, then $P_n(x)$ are unchanged, that is, $P_n(x; d\mu) = P_n(x; d\tilde{\mu})$, but $p_n(x; d\mu) = p_n(x; d\tilde{\mu})\mu(\mathbb{R})^{-1/2}$, and consistent with (1.2.39), $\lambda_n(\zeta, d\mu) = \mu(\mathbb{R})\lambda_n(\zeta, d\tilde{\mu})$. The left, and so, the right side of (1.2.39) is clearly monotone in μ, that is, $\mu < \nu \Rightarrow \lambda_n(\zeta, d\mu) \leq \lambda_n(\zeta, d\mu)$, and we thus have

$$\sum_{j=0}^{n} |p_j(\zeta; d\mu)|^2 \geq \sum_{j=0}^{n} |p_j(\zeta; d\nu)|^2 \tag{1.2.40}$$

a fact that is not obvious without (1.2.39).

Remarks. 1. It is known (see [**17, 974**]) that the moment problem is determinate if and only if for some $\zeta \in \mathbb{C}_+$, $\sum_{j=0}^{\infty} |p_j(\zeta; d\mu)|^2 = \infty$. (1.2.40) implies that if c_n and d_n are two sets of moments and $c_n - d_n$ is also a set of moments, then if the moment problem for c_n is determinate, the same is true for d_n. Alternately, we note if d_n is indeterminate, we can get multiple measures for c_n by adding solutions for d_n to a solution for $c_n - d_n$.

2. If the moment problem is determinate, it is known (see [**974**]) for all $\zeta \in \mathbb{C}$:

$$\sum_{j=0}^{\infty} |p_j(\zeta, d\mu)|^2 = \begin{cases} \mu(\{\zeta\})^{-1} & \text{if } \zeta \in \mathbb{R} \\ \infty & \text{if } \zeta \notin \mathbb{R} \end{cases} \tag{1.2.41}$$

and thus, if the moment problem is determinate and μ has no pure points, then

$$\lim_{n \to \infty} \lambda_n(\zeta) = 0 \tag{1.2.42}$$

for all ζ. We will see the situation for OPUC is very different.

One place that the Christoffel function enters is in the *Gauss-Jacobi* or *Gauss-Christoffel quadrature formula* that

$$\int \pi(x)\, d\mu(x) = \sum_{j=1}^{n} \lambda_n(x_j^{(n)}(d\mu), d\mu)\pi(x_j^{(n)}(d\mu)) \tag{1.2.43}$$

for π a polynomial of degree $2n-1$ and where $\{x_j^{(n)}(d\mu)\}_{j=1}^{n}$ are the zeros of $P_n(x)$. The point of (1.2.43) is that generally if one fixed n points, one can only hope to fit polynomials up to degree $n-1$ by adjusting the constants. But here we get a space of almost twice the dimension. The weights $\lambda_n(x_j^{(n)}, d\mu)$ are called *Cotes numbers.*

Closely related to (1.2.43) is the fact that in the case where the moment problem is determinate,

$$\sum_{j=1}^{n} \lambda_n(x_j^{(n)}(d\mu), d\mu)\delta_{x_j^{(n)}(d\mu)} \to d\mu \tag{1.2.44}$$

weakly, that is, when integrated over with functions in $C_\infty(\mathbb{R})$, the functions vanishing at infinity. Given (1.2.43), we can integrate over any polynomially bounded function.

We will give (nonstandard) proofs of (1.2.43) and (1.2.44) in Subsection 12, but note the following which is part of a proof of (1.2.43): By Lagrange interpolation, there are unique polynomials $L_j^{(n)}(x, d\mu)$ of degree $n-1$ with $L_j^{(n)}(x_k^{(n)}(d\mu), d\mu) = \delta_{jk}$. Then if π is any polynomial, $\pi(x) - \sum_{j=1}^{n} \pi(x_j^{(n)})L_j^{(n)}(x)$ vanishes at the zeros of P_n and so divisible by P_n, that is,

$$\pi(x) - \sum_{j=1}^{n} \pi(x_j^{(n)})L_j^{(n)}(x) = P_n(x)R(x)$$

Since $\deg R \leq \max(\deg(\pi), n-1) - n$, we see that if $\deg(\pi) \leq 2n-1$, $\int P_n(x)R(x)\, d\mu(x) = 0$, that is, (1.2.43) holds with weights $\int L_j^{(n)}(x)\, d\mu$.

Christoffel functions have been turned by Freud and Nevai (see [**363, 809**] and [**697, 717, 721**] for more recent related developments) into powerful tools for analyzing asymptotics and spectral theory.

9. Christoffel-Darboux Formula

If we write (1.2.16) at two real points x and y, multiply the x equation by $p_n(y)$ and the y equation by $p_n(x)$, and subtract, we get

$$(x-y)p_n(x)p_n(y) = Q_{n+1}(x,y) - Q_n(x,y) \tag{1.2.45}$$

where

$$Q_j(x,y) = a_j(p_j(x)p_{j-1}(y) - p_j(y)p_{j-1}(x))$$

Since $Q_0(x,y) = 0$ (since $p_{-1}(x) = 0$), if we write (1.2.45) for $j = 0, 1, \dots$, and sum, we get the Christoffel-Darboux formula,

$$K_n(x,y) = a_{n+1}\, \frac{p_{n+1}(x)p_n(y) - p_n(x)p_{n+1}(y)}{x-y} \tag{1.2.46}$$

where K_n is the reproducing kernel given by (1.2.36). This also has a limit

$$K_n(x,x) = a_{n+1}(p'_{n+1}(x)p_n(x) - p'_n(x)p_{n+1}(x)) \tag{1.2.47}$$

We will not apply the formula here except to note it is often used to prove many of the facts we have obtained by other means, for example, interlacing of zeros and the Gauss-Jacobi quadrature formula. We will use its analog in the OPUC case; see especially Section 2.2.

10. Second Kind Polynomials, Transfer Matrices, and the m-function

Define $a_0 \equiv 1$. The numbers $u_n \equiv p_n(x)$ are the unique solution of the difference equation

$$a_{n+1}u_{n+1} = (x - b_{n+1})u_n - a_n u_{n-1} \qquad n = 0, 1, 2, \dots \tag{1.2.48}$$

with the boundary conditions $u_{-1} = 0$, $u_0 = 1$. The *second kind polynomials*, $q_n(x)$, are the unique solution of (1.2.48) with boundary conditions $u_{-1} = -1$, $u_0 = 0$

-256- - -* ! - % * ! 3 !
H ! 1 ! 9 3 2 ! ! ! ! 3 ' 1"
(// $' Q R =) 6 !" !

[184] C. Carathéodory, *Über den Variabilitätsbereich der Fourier'schen Konstanten von positiven harmonischen Funktionen*, Rend. Circ. Mat. Palermo **32** (1911), 193–217. (Cited on 37, 38.)

[185] C. Carathéodory and L. Fejér, *Über den Zusammenhang der Extremen von harmonischen Funktionen mit ihren Koeffizienten und über den Picard-Landau'schen Satz*, Rend. Circ. Mat. Palermo **32** (1911), 218–239. (Cited on 37.)

[186] R. Carey and J. Pincus, *Perturbation vectors*, Integral Equations Operator Theory **35** (1999), 271–365. (Cited on 333.)

[187] T. Carleman, *Zur theorie der linearen integralgleichungen*, Math. Zeit. **9** (1921), 196–217. (Cited on 55.)

[188] R. Carmona, *One-dimensional Schrödinger operators with random or deterministic potentials: New spectral types*, J. Funct. Anal. **51** (1983), 229–258. (Cited on 630, 706, 972.)

[189] R. Carmona and J. Lacroix, *Spectral Theory of Random Schrödinger Operators. Probability and Its Applications*, Birkhäuser, Boston, 1990. (Cited on 605, 616, 617, 847.)

[190] K.M. Case, *Orthogonal polynomials from the viewpoint of scattering theory*, J. Math. Phys. **15** (1974), 2166–2174. (Cited on xiii, 937.)

[191] K.M. Case, *Orthogonal polynomials, II* J. Math. Phys. **16** (1975), 1435–1440. (Cited on 937.)

[192] L. Cesari, *Un nuovo criterio di stabilità per le soluzioni delle equazioni differenziali lineari*, Ann. Scuola Norm. Super. Pisa (2) **9** (1940), 163–186. (Cited on 824.)

[193] K. Chadan and P.C. Sabatier, *Inverse Problems in Quantum Scattering Theory*, second edition. Texts and Monographs in Physics, Springer, New York, 1989. (Cited on 630.)

[194] J. Charris and M.E.H. Ismail, *On sieved orthogonal polynomials, V. Sieved Pollaczek polynomials*, SIAM J. Math. Anal. **18** (1987), 1177–1218. (Cited on 178.)

[195] P.L. Chebyshev, *On boundary values of integrals* in "Collected Works," Moscow, 1918. (Cited on 24.)

[196] T.S. Chihara, *An Introduction to Orthogonal Polynomials*, Mathematics and Its Applications, 13, Gordon and Breach, New York-London-Paris, 1978. (Cited on 24.)

[197] G. Choquet, *Lectures on Analysis, Volume 1: Integration and Topological Vector Spaces; Volume II: Representation Theory; Volume III: Infinite Dimensional Measures and Problem Solutions*, W.A. Benjamin, New York-Amsterdam, 1969. (Cited on 605.)

[198] Y.S. Chow and H. Teicher, *Probability Theory. Independence, Interchangeability, Martingales*, Springer, New York-Heidelberg, 1978. (Cited on 410.)

[199] D. Chowdhury and D. Stauffer, *Principles of Equilibrium Statistical Mechanics*, Wiley-VCH, Weinheim, 2000. (Cited on 352.)

[200] M. Christ and A. Kiselev, *WKB and spectral analysis of one-dimensional Schrödinger operators with slowly varying potentials*, Comm. Math. Phys. **218** (2001), 245–262. (Cited on 673.)

[201] M. Christ, A. Kiselev, and C. Remling, *The absolutely continuous spectrum of one-dimensional Schrödinger operators with decaying potentials*, Math. Res. Lett. **4** (1997), 719–723. (Cited on xii.)

[202] E.B. Christoffel, *Über die Gaussische Quadratur und eine Verallgemeinerung derselben*, J. Reine Angew. Math. **55** (1858), 61–82. (Cited on 24, 684.)

[203] V.A. Chulaevsky, *An inverse spectral problem for limit-periodic Schrödinger operators*, Functional Anal. Appl. **18** (1984), 220–233; Russian original in Funktsional. Anal. i Prilozhen. **18** (1984), 63–66. (Cited on 707.)

[204] V.A. Chulaevsky and Ya. G. Sinai, *Anderson localization and KAM-theory*, in "Analysis, et cetera," pp. 237–249, Academic Press, Boston, 1990. (Cited on 707.)

[205] S. Clark and F. Gesztesy, *Weyl-Titchmarsh M-Function asymptotics for matrix-valued Schrödinger operators*, Proc. London Math. Soc. **82** (2001), 701–724. (Cited on 216.)

[206] S. Clark and F. Gesztesy, *Weyl-Titchmarsh M-function asymptotics, local uniqueness results, trace formulas, and Borg-type theorems for Dirac operators*, Trans. Amer. Math. Soc. **354** (2002), 3475–3534. (Cited on 216.)

[207] S. Clark and F. Gesztesy, *On Povzner-Wienholtz-type self-adjointness results for matrix-valued Sturm-Liouville operators*, Proc. Math. Soc. Edinburgh **133A** (2003), 747–758. (Cited on 216.)

-256- - -* ! - % 9 & ! ! ?
! ! ! 1 8 ! " / 1"

(%B

data

20211015m

500
40
145
95
780
75
855
120
98
1.073m
(Livros)
1073
130
1203m

Brestova.

Edentulismo: perda dos dentes.

Applications of Harmonic Analysis
Vilenkin

Integral Geometry and Representation Theory - Vilenkin

Books:

- 20211015/ A History of Geometrical Methods/ Julian ... A Treatise on algebraic plane curves ... (1959).pdf
- 20211015/ Integral Measure .../ Integral, Measure ... pdf
- 20211015/ Calculus of Variations/ Robert... Calculus of Variations...pdf
- 20211016/ Cálculo/
  - Ecuaciones Diferenciales Elementales Rainville..pdf

Pag. 95 Capitulo 5

Jandaia

'

! -256- - -* ! - ' C ! ? " 2 !
! D ! 1 ! A 8 G .J 4 A D ! /$/ 1"

D -256- - -* ! - ' & ! J ! ! " 4 A! A 2
H " 211 ! 4 A D ! / I 1"

-256- - -* ! - ' 5 D @ . ! *
! ! J @ 211 ! 9 3 2 " = @ . / ' 1"

" -256- - -* ! - ' = ! 4 !
= ! 5! A 4<4 A 1"
(/$!1{ $ 9 K ! !

-256- - -* ! - ' = ! 4 !
4 A!"! ! ! 1"

-256- - -* ! - ' = ! " ! 1"

-256- - -* ! - ' = ! " ! 3 1 D !
A! ! !I D 4 ! => ! ! " ! D 1"
*D ? A " UUU

-256- - -* ! - '
4 ! I=> ! I! "I? ! I2 D !I7 D! I=" 1"
(B % ! ! " 7! ! = !
(7! C "! !
(/ 7! ! ! ! 0
(%% D 1! 5%
(B%B 9 ? " => !
(B$' ' 7! ! 6 !" !

-256- - -* ! - ' BB/B$I 1"
E ' M " 7M M 4 1 A C!A! 3

K -256- - -* ! - ' ! " D 1"
K -256- - -* ! - ' ! " D 1"
K% -256- - -* ! - ' ! " %D 1"

. -256- - -* ! - ' ! " 1"
*D 7M M 5 M
. -256- - -* ! - ' ! " 1"

-256- - -* ! - ' ! " ! ! ! 1"
-256- - -* ! - ' ! " ! ! ! 1"

-256- - -* ! - ' 9 ! ! & 3 2 7 "
7 ! &" ! 1"
(/ / 7! " A " " 3
*D : D 0 & :

intersects finitely many entries in the array and each entry is met by some arrow.

We take it as clear that f is one-to-one. Let $A_{nm} \in \mathbf{L}_n$. By the above scheme we see that the subscripts lying along the first arrow add up to $1+1=2$, the subscripts lying along the second arrow add up to 3, the subscripts lying along arrow #3 add up to 4, and in general the subscripts lying along arrow $\#k$ add up to $k+1$. Thus A_{nm} lies along arrow $\#n+m-1$. Now consider the number that we will assign A_{nm} with f.

On arrow #1 there is 1 entry. On arrow #2 there are 2 entries, and in general on arrow $\#n$ there are n entries. Thus, when we have finished counting the entries from A_{11} to A_{1n}, we will have counted a total of

$$\begin{array}{llllllll} \text{from arrow} & \#1 & \#2 & \ldots & \#\ell & & \\ & 1 \;+ & 2 \;+ & \ldots \;+ & \ell & = & \dfrac{\ell(\ell+1)}{2} \end{array}$$

entries. We will see why this sum is $\frac{\ell(\ell+1)}{2}$ in a little bit. For now just accept it knowing that there is the promise of a proof. Meanwhile back at the argument, we see that A_{nm} is on arrow $\#n+m-1$ so that

$$\text{there is some } k \leq \frac{(n+m-1)(n+m)}{2} \text{ such that } f(k) = A_{nm}.$$

This proves that $f(k) = A_{mn}$ for some number k, and this in turn proves that f is onto. Since we have already stated that f is one-to-one, $f : \mathbb{N} \longrightarrow A$ is a bijection. We would then send Rationalist A_{nm} to room k. The assignment of rooms is complete.

Let us take a break at this time and examine the sum $1+2+\ldots+\ell$. A German school teacher was feeling a bit hungry so he prepared to go out for coffee and a danish. To keep his pupils busy he instructed them to add up the numbers from 1 to 100. We will do this problem for an arbitray positive integer ℓ. A child, Carl F. Gauss, decided that there must be a faster way. In a flash of genius the 11 year old Gauss discovered the following identity. Let us call the sum S.

$$S = 1+2+\ldots+\ell.$$

Find the sum in two different ways. Start at 1 and end at ℓ, and in the other start at ℓ and end at 1.

$$\begin{array}{ccccccccccc} S & = & 1 & + & 2 & + & \ldots & + & \ell-1 & + & \ell \\ S & = & \ell & + & \ell-1 & + & \ldots & + & 2 & + & 1 \end{array}$$

Notice that in each column we have a sum

$$\ell+1.$$

Add these two sums together to find that

$$\begin{array}{ccccccccccc} S & = & 1 & + & 2 & + & \ldots & + & \ell-1 & + & \ell \\ +S & = & \ell & + & \ell-1 & + & \ldots & + & 2 & + & 1 \\ \hline 2S & = & (\ell+1) & + & (\ell+1) & + & \ldots & + & (\ell+1) & + & (\ell+1) \end{array}$$

There are ℓ terms $\ell+1$ so we have $2S = \ell(\ell+1)$ or equivalently

$$S = 1+2+\ldots+\ell = \frac{\ell(\ell+1)}{2}.$$

Quite a pretty little argument for a child, eh? Before the school teacher could button his coat, Gauss turned his slate over and folded his hands. The teacher, thinking something was amiss, examined the slate and saw the correct sum written there. History does not record whether or not the school teacher got his coffee that day or if he suffered for a lack of it.

The purpose behind the above stories is to develop an intuition for the infinite. The story about Hilbert's Infinite Hotel shows us that in an infinite set there is always *plenty of room.* It would seem that we can make room for any number of guests. Or can we?

The conventions and conferences are over so that the Hotel is empty. Hilbert breathes a sigh of relief and counts his revenue. Accommodating all of those guests has taxed him. There is a knock at the door and Hilbert sees that an infinite collection of Realists

$$\{R_x \mid x \in \mathbb{R} \text{ and } 0 \leq x \leq 1\}$$

is in town for a conference on complex issues. Hilbert sets out to assign each of them a private room, but his first attempt at assigning rooms leaves infinitely many Realists without a room. He tries again and still he cannot assign every Realist a single room. For some reason he does not yet fathom, no matter how he assigns rooms to the Realists, some Realists do not get a single room. The mathematics behind this conundrum is a jolt to our *common sense* and will have to wait until Theorem 4.3.1. The problem of assigning rooms to Realists shows us that our *intuition* about finite sets is useless when we consider matters about infinite sets. We will have to develop entirely new insights if we are to understand the infinite. Hilbert does not yet have the mathematical experience to address this problem. He sends most of these potential guests packing.

Let us discover what Hilbert did not fathom. Let us see why there is no way to accommodate these realists with single rooms in the Infinite Hotel.

Suppose, for the sake of contradiction, that we can make an implied list of these guests and their single rooms. That is, we assign Realists to single rooms and then we make an implied list of the rooms n and the associated occupants R_{x_n}.

room	occupant	decimal					
1	R_{x_1}	$x_1 =$	$.\boxed{d_{11}}$	d_{12}	d_{13}	d_{14}	$\dots$
2	R_{x_2}	$x_2 =$	$.d_{21}$	$\boxed{d_{22}}$	d_{23}	d_{24}	$\dots$
3	R_{x_3}	$x_3 =$	$.d_{31}$	d_{32}	$\boxed{d_{33}}$	d_{34}	$\dots$
4	R_{x_4}	$x_4 =$	$.d_{41}$	d_{42}	d_{43}	$\boxed{d_{44}}$	$\dots$
$\vdots$	$\vdots$						

We have also given the decimal expansion of x_n. In this implied list, each d_{ij} is a digit, a number in the set $\{0, 1, \dots, 9\}$, and we choose the shortest decimal expansion for x_n. That is, if the decimal expansion for x_n ends with $99\bar{9}$ then we replace it with the number that ends in 0's. For instance, the following numbers are equal.

$$\begin{aligned} 1.00\bar{0} &= .99\bar{9} \\ .100\bar{0} &= .099\bar{9} \\ .123\bar{0} &= .12299\bar{9}. \end{aligned}$$

In these cases we choose the shorter finite decimal expansion, *except for* 1. The number 1 is written as a decimal $.99\bar{9}$.

Furthermore, d_{11} represents the first digit in the decimal form of x_1, d_{23} represents digit number 3 in the decimal form of x_2, and in general d_{nm} represents digit number m in the decimal form of x_n. We can now produce a Realist R_x who does not get a room.

Let us define a digit d_1 as follows.

$$d_1 = \begin{cases} d_{11}+1 & \text{if } d_{11} \neq 9 \\ 0 & \text{if } d_{11} = 9 \end{cases} .$$

Notice that we have defined d_1 so that $d_1 \neq d_{11}$. For example, if $d_{11} = 5$ then $d_1 = 6$ and if $d_{11} = 9$ then $d_1 = 0$.

We do the same thing for d_{22}. Let

$$d_2 = \begin{cases} d_{22}+1 & \text{if } d_{22} \neq 9 \\ 0 & \text{if } d_{22} = 9 \end{cases} .$$

Notice that we have defined d_2 so that $d_2 \neq d_{22}$. For example, if $d_{22} = 0$ then $d_2 = 1$ and if $d_{22} = 7$ then $d_2 = 8$.

In general, let us define

$$d_n = \begin{cases} d_{nn}+1 & \text{if } d_{nn} \neq 9 \\ 0 & \text{if } d_{nn} = 9 \end{cases} .$$

Then $d_n \neq d_{nn}$.

The roomless Realist R_x is given by writing down the decimal expansion

$$\boxed{x = .d_1 d_2 d_3 d_4 \ldots .}$$

If x ends in $99\bar{9}$ then we replace it with the shorter decimal representation that ends in $00\bar{0}$. Then $x \neq x_1$ because by our choice of d_1, x and x_1 differ in the first decimal place, $d_1 \neq d_{11}$. Also, $x \neq x_2$ because by our choice of d_2, x and x_2 differ in decimal place number 2, $d_2 \neq d_{22}$. In general, $x \neq x_n$ because by our choice of d_n, x and x_n differ in decimal place number n, $d_n \neq d_{nn}$. Thus R_x is not on the implied list of Realists who got a room. This is the person we have sought. This is why Hilbert could not give each Realist a single room. There are simply too many Realists! No matter what rooms he assigned there would always be a Realist R_x who does not get a room. There are more Realists than there are rooms.

How can that be, reader? There are infinitely many rooms and infinitely many Realists. How can there be more Realists than rooms? Read on.

Let's examine the above choices for d_n and the construction of x on a concrete example. Suppose that we are given specific real numbers x_1, x_2, x_3, ... that form the following implied list.

$$\begin{array}{rclllllll}
x_1 &=& .\boxed{1} & 2 & 3 & 4 & 5 & \ldots \\
x_2 &=& .0 & \boxed{9} & 0 & 1 & 0 & \ldots \\
x_3 &=& .1 & 4 & \boxed{9} & 5 & 7 & \ldots \\
x_4 &=& .4 & 1 & 4 & \boxed{2} & 4 & \ldots \\
x_5 &=& .0 & 1 & 0 & 1 & \boxed{3} & \ldots \\
 & \vdots & & & & & &
\end{array}$$

Then

$$x = .2\ 0\ 0\ 3\ 4\ \cdots.$$

The first decimal of x is 2 because the first decimal of x_1 is 1.
The second decimal of x is 0 because the second decimal of x_2 is 9.
The second decimal of x is 0 because the third decimal of x_3 is 9.
The fourth decimal of x is 3 because the fourth decimal of x_4 is 2.
The fifth decimal of x is 4 because the fifth decimal of x_5 is 3.

The comparisons of x to the numbers on our implied list are obvious.
$x \neq x_1$ because they differ in decimal place 1,
$x \neq x_2$ because they differ in decimal place 2,
$x \neq x_3$ because they differ in decimal place 3.
These comparisons continue indefinitely. In this way we show that x is not on the implied list $x_1, x_2, x_3, \ldots$.

There is a person at the Hotel who haunts the Lobby. Each day he tries to walk a 2 meter long straight line across the Lobby floor. He starts at one end of the line and wants to get to the other end, but he has a peculiar way of walking. His next step is always half the length of his previous step. He starts with a step of one meter. One meter remains to be crossed.

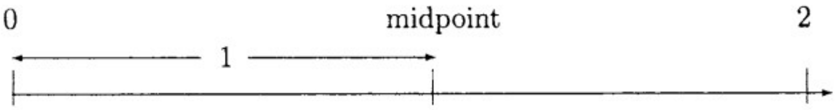
0
midpoint
2
1

He follows this with a $\frac{1}{2}$ meter step. Half a meter remains.

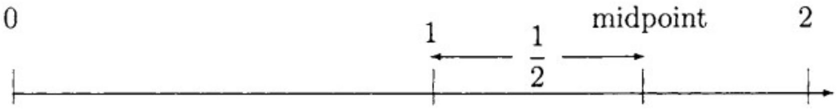
0
1
$\frac{1}{2}$
midpoint
2

Then he steps off a $\frac{1}{4} = \frac{1}{2^2}$ meter step. A $\frac{1}{4}$ meter remains.

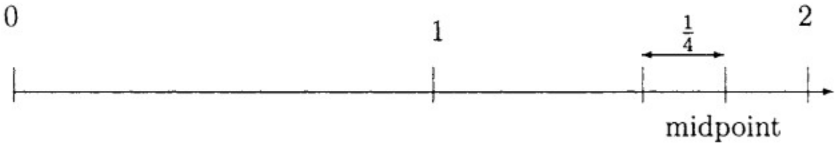
0
1
$\frac{1}{4}$
2
midpoint

Continuing on, on his nth step he moves $\frac{1}{2^n}$ meters, and that leaves $\frac{1}{2^n}$ meters left to be covered. The scale is magnified.

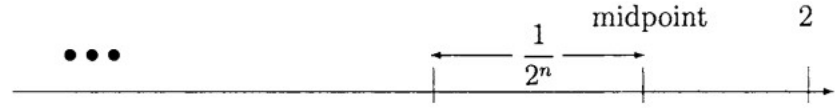
• • •
$\frac{1}{2^n}$
midpoint
2

He continues on indefinitely. Upon completion there are 0 meters left to cover and so he has crossed

$$2 = 1 + \frac{1}{2} + \frac{1}{4} + \ldots + \frac{1}{2^n} + \ldots$$

meters. He has walked the entire 2 meter length.

Each of the rooms at Hilbert's Infinite Hotel has to be cleaned and there is one and only one cleaning lady to do the job. Her name is Mary. (Hilbert's mother needed a job.) Her routine is to clean the rooms in order. She begins in the morning with room 1, followed by room 2, and then room 3, and on inductively. She is an industrious woman who picks up momentum as her day progresses so she cleans the first room in 1 hour, the second room in $\frac{1}{2}$ hour, the next in $\frac{1}{2^2}$ hour, and so on, cleaning room n in $\frac{1}{2^n}$ hours. When she has cleaned each room she is done. Thus she works

$$1 + \frac{1}{2} + \frac{1}{4} + \ldots + \frac{1}{2^n} + \ldots = 2$$

hours a day. She certainly has a good job. She only has a 2 hour work day. But something else is wrong here. Where is she at the end of her day? There is no last room for her to clean because the Hotel has *infinitely many rooms*. However, we can be assured that each of the rooms is spotless. We will learn the fate of the cleaning lady when we finish our discussion of ordinals on page 184.

3.3 Equivalent Sets and Cardinality

In this section we will define the equivalence of cardinalities.

Definition 3.3.1 *Sets A and B are said to be* equivalent *if there is a bijection $f : A \longrightarrow B$.*

Theorem 3.3.2 *A and B are equivalent exactly when* $\mathrm{card}(A) = \mathrm{card}(B)$.

If we look at the different steps in Hilbert's Infinite Hotel we see that in making space for the lone stranger Hilbert used the fact that the function

$$f : \{1, 2, 3, \ldots\} \longrightarrow \{2, 3, 4, \ldots\}$$

defined by

$$f(n) = n + 1$$

(%B%

Trajectorias ortogonales

Así una curva que corta otras formando angulos rectos, se llama trayectoria ortogonal de aquellas otras.

Pág. 98 24. Trayectorias ortogonales, coordenadas polares

$$\tan \psi = r \frac{d\theta}{dr}$$

Veja: Figura 11.

_20211016/ Cálculo/

Ecuaciones en derivadas parciales con series de Fourier y problemas de contorno...pdf

Obs: olhar livro todo!!.

Interessante.

_20211016/ Cálculo/

Ecuaciones diferenciales...pdf

Gauss - 20 occurrences

Obs: Livro todo!!.

_20211016/ Cálculo/ Ecuaciones

(%B'

diferenciales y problemas con valores
en la frontera ...pdf

Obs: Estudar este livro!!!

20211016 ///

Books:

- 20211016 / Differential Equations... /

Differential Equations and Linear
Algebra.().pdf

Gauss - 42 ocurrencias

Pág. 160 3.2

Matrices and Gaussian Elimination

Pág. 171 Gauss-Jordan elimination

→ Algorithm Gauss-Jordan
Elimination

Pág. 229 (using Gaussian elimination
or Cramer's rule)

Pág. 233 Subspaces of $\mathbb{R}^3$

Pág. 636 the Legendre Equation

→ Gaussian quadrature

Pág. 654 42.

This problem is a brief introduction to Gauss's hypergeometric equation.

_20211016/ Traite de Geometrie Descriptive Javary/

obs: Há os 2 volumes.

000066965_02.pdf

781 pages.

obs: /Books-others/

_traitdegom01breinoft.pdf
02
03

3 volumes - Traité de Géométrie Descriptive - Breithof

_traitdegom01rouchoft.pdf
02

2 volumes - Traité de Géométrie Rouché

_20211016/ Grigori-Perelman-Poincaré/

obs: Diretório todo!!!

_20211016/ Grigori.../ →

The Mathematics of Infinity
A guide to great Ideas ... .pdf

Pág. 92(109) – Gauss
Gauss discovered the following identity

obs: Hilbert's Infinite Hotel

Pág. 108(125)
sum as the identity

Pág. 233(250)
little Gauss on page 92

– 20211016| Gugou ... |
Visions of Infinity the great mathematical problems ... .pdf

– Pág. 40(53) Chapter 3
The puzzle of pi
Squaring the circle.

obs: Gauss – 58 occurrences

Pág. 22(34) – Gauss introduced modular arithmetic in the
Disquisitiones Arithmeticae

3

The puzzle of pi

Squaring the Circle

PRIMES ARE AN OLD IDEA, but circles are even older. Circles led to a great problem that took more than 2000 years to solve. It is one of several related geometric problems that have come down to us from antiquity. The central character in the story is the number π (Greek 'pi') which we meet at school in connection with circles and spheres. Numerically it is 3.14159 and a bit; often the approximation 22/7 is used. The digits of π never stop, and they never repeat the same sequence over and over again. The current record for calculating digits of π is 10 trillion digits, by Alexander Yee and Chigeru Kondo in October 2011.[14] Computations like this are significant as ways to test fast computers, or to inspire and test new methods to calculate π, but very little hinges on the numerical results. The reason for being interested in π is not to calculate the circumference of a circle. The same strange number appears all over mathematics, not just in formulas related to circles and spheres, and it leads into very deep waters indeed. The school formulas are important, even so, and they reflect π's origins in Greek geometry.

There, one of the great problems was the unsolved task of squaring the circle. This phrase is often employed colloquially to indicate a wrong-headed approach to something, rather like trying to fit a square peg into a round hole. Like many common phrases extracted from science, this one's meaning has changed over the centuries.[15] In Greek times, trying to square the circle was a perfectly reasonable idea. The difference in the two shapes – straight or curved – is totally irrelevant:

similar problems have valid solutions.[16] However, it eventually turned out that this particular problem cannot be solved using the specified methods. The proof is ingenious and technical, but its general nature is comprehensible.

In mathematics, squaring the circle means constructing a square whose *area* is the same as that of a given circle, using the traditional methods of Euclid. Greek geometry actually permitted other methods, so one aspect of the problem is to pin down which methods are to be used. The impossibility of solving the problem is then a statement about the limitations of those methods; it doesn't imply that we can't work out the area of a circle. We just have to find another approach. The impossibility proof explains why the Greek geometers and their successors failed to find a construction of the required kind: there isn't one. In retrospect, that explains why they had to introduce more esoteric methods. So the solution, despite being negative, clears up what would otherwise be a big historical puzzle. It also stops people wasting time in a continuing search for a construction that doesn't exist – except for a few hardy souls who regrettably seem unable to get the message, no matter how carefully it is explained.[17]

In Euclid's *Elements* the traditional methods for constructing geometric figures are idealised versions of two mathematical instruments: the ruler and the compass. To be pedantic, compass*es*, for the same reason that you cut paper with scissor*s*, not with a scissor – but I will follow common parlance and avoid the plural. These instruments are used to 'draw' diagrams on a notional sheet of paper, the Euclidean plane.

Their form determines what they can draw. A compass comprises two rigid rods, hinged together. One has a sharp point, the other holds a sharp pencil. The instrument is used to draw a circle, or part of one, with a specific centre and a specific radius. A ruler is simpler: it has a straight edge, and is used to draw a straight line. Unlike the rulers you buy in stationery shops, Euclid's rulers have no marks on them, and this is an important restriction for the mathematical analysis of what they can create.

The sense in which the geometer's ruler and compass are idealisations is straightforward: they are assumed to draw infinitely thin lines. Moreover, the straight lines are exactly straight and the circles are perfectly round. The paper is perfectly flat and even. The other key ingredient of Euclid's geometry is the notion of a point, another ideal. A point is a dot on the paper, but it is a physical impossibility: it has no size. 'A point', said Euclid, in the first sentence of the *Elements*, 'is that which has no part.' This sounds a bit like an atom, or if you're clued into modern physics, a subatomic particle, but compared to a geometric point, those are gigantic. From an everyday human perspective, however, Euclid's ideal point, an atom, and a pencil dot on a sheet of paper, are similar enough for the purposes of geometry.

These ideals are not attainable in the real world, however carefully you make the instruments and sharpen the pencil, and however smooth you make the paper. But idealism can be a virtue, because these requirements make the mathematics much simpler. For instance, two pencil lines cross in a small fuzzy region shaped like a parallelogram, but mathematical lines meet at a single point. Insights gained from ideal circles and lines can often be transferred to real, imperfect ones. This is how mathematics works its magic.

Two points determine a (straight) line, the unique line that passes through them. To construct the line, place your ideal ruler so that it passes through the two points, and run your ideal pencil along it. Two points also determine a circle: choose one as the centre, and place the compass point there; then adjust it so that the tip of the pencil lies on the other point. Now swing the pencil round in an arc, keeping the central point fixed. Two lines determine a unique point, where they cross, unless they are parallel, in which case they don't cross, but a Pandora's box of logical issues yawns wide. A line and a circle determine two points, if they cross; one point, if the line cuts the circle at a tangent; nothing at all if the circle is too small to meet the line. Similarly two circles either meet in two points, one, or none.

Distance is a fundamental concept in the modern treatment of Euclidean geometry. The distance between any two points is measured along the line that joins them. Euclid managed to get his geometry working without an explicit concept of distance, by finding a way to say that two line segments have the *same* length without defining length itself. In fact, this is easy: just stretch a compass between the ends of one segment, transfer it to the second, and see if the ends fit. If they do, the lengths are equal; if they don't, they're not. At no stage do you measure an actual length.

From these basic ingredients, geometers can build up more interesting shapes and configurations. Three points determine a triangle unless they all lie on the same line. When two lines cross, they form an angle. A right angle is especially significant; a straight line corresponds to two right angles joined together. And so on, and so on, and so on. Euclid's *Elements* consists of 13 books, delving ever deeper into the consequences of these simple beginnings.

The bulk of the *Elements* consists of theorems – valid features of geometry. But Euclid also explains how to solve geometric problems, using 'constructions' based on ruler and compass. Given two points joined by a segment of a line, construct their midpoint. Or trisect the segment: construct a point exactly one third of the way along it. Given an angle, construct one that bisects it – is half the size. But some simple constructions proved elusive. Given an angle, construct one that trisects it – is one third the size. You can do that for line segments, but no one could find a method for angles. Approximations, as close as you wish, yes. Exact constructions using only an unmarked ruler and a compass: no. However, no one really needs to trisect angles exactly anyway, so this particular issue didn't cause much trouble.

More embarrassing was a construction that could not be ignored: given a circle, construct a square that has the same area. This is the problem of squaring the circle. From the Greek point of view, if you couldn't solve that, you weren't entitled to claim that a circle *had* an area. Even though it visibly encloses a well-defined space, and intuitively the area is *how much* space. Euclid and his successors, notably Archimedes, settled for a pragmatic solution: assume circles have areas, but don't expect to be able to construct squares with the same area. You can still say a lot; for instance, you can prove, in full logical rigour, that the area of a circle is proportional to the square of its diameter. What you can't do, without squaring the circle, is to construct a line whose length is the constant of proportionality.

The Greeks couldn't square the circle using ruler and compass, so they settled for other methods. One used a curve called a quadratrix.[18] The importance they attached to using only ruler and compass was exaggerated by some later commentators, and it's not even clear that the Greeks considered squaring the circle to be a vital issue. By the nineteenth century, however, the problem was becoming a major nuisance. Mathematics that was unable to answer such a straightforward question was like a cordon bleu cook who didn't know how to boil an egg.

Squaring the circle sounds like a problem in geometry. That's because it is a problem in geometry. But its solution turned out to lie not in geometry at all, but in algebra. Making unexpected connections between apparently unrelated areas of mathematics often lies at the heart of solving a great problem. Here, the connection was not entirely unprecedented, but its link to squaring the circle was not at first appreciated. Even when it was, there was a technical difficulty, and dealing with that required yet another area of mathematics: analysis, the rigorous version of calculus. Ironically, the first breakthrough came from a fourth area: number theory. And it solved a geometric problem that the Greeks would never in their wildest dreams have believed to possess a solution, and as far as we can tell never thought about: how to construct, with ruler and compass, a regular polygon with 17 sides.

It sounds mad, especially if I add that no such construction exists for regular polygons with 7, 9, 11, 13, or 14 sides, but one does for 3, 4, 5, 6, 8, 10, and 12. However, there is method behind the madness, and it is the method that enriched mathematics.

First: what is a regular polygon? A polygon is a shape bounded by straight lines. It is regular if those lines have equal length and meet at equal angles. The most familiar example is the square: all four sides have the same length and all four angles are right angles. There are other shapes with four equal sides or four equal angles: the rhombus and rectangle respectively. Only a square has both features. A regular 3-sided polygon is an equilateral triangle, a regular 5-sided polygon is a regular pentagon, and so on, Figure 4. Euclid provides ruler-and-compass constructions for regular polygons with 3, 4, and 5 sides. The Greeks also knew how to repeatedly double the number of sides, giving 6, 8, 10, 12, 16, 20, and so on. By combining the constructions for 3- and 5-sided regular polygons they could obtain a 15-sided one. But there, their knowledge stopped. And for about 2000 years that's how it remained. No one imagined that any other numbers were feasible.

They didn't even ask, it just seemed obvious that nothing more could be done.

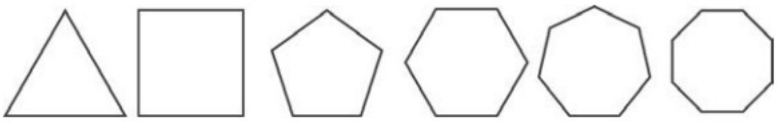

Fig 4 The first few regular polygons. *From left to right*: equilateral triangle, square, pentagon, hexagon, heptagon, octagon.

It took one of the greatest mathematicians who have ever lived to think the unthinkable, ask the unaskable, and discover a truly astonishing answer. Namely, Gauss Carl Friedrich. Gauss was born into a poor, working-class family in the city of Braunschweig (Brunswick) in Germany. His mother Dorothea could not read or write, and failed to write down the date of his birth, but she did remember that it was on a Wednesday, eight days before the feast of the ascension, in 1777. Gauss later worked out the exact date from a mathematical formula he devised for the date of Easter. His father Gebhard came from a farming family, but made a living in a series of low-level jobs: gardener, canal labourer, street butcher, funeral parlour accountant. Their son was a child prodigy who is reputed to have corrected his father's arithmetic at the age of three, and his abilities, which extended to languages as well as mathematics, led the Duke of Braunschweig to fund his university studies at the Collegium Carolinum. While an undergraduate Gauss independently rediscovered several important mathematical theorems that had been proved by illustrious people such as Euler. But his theorem about the regular 17-sided polygon came as a bolt from the blue.

By then, the close link between geometry and algebra had been understood for 140 years. In an appendix to *Discours de la Méthode* ('Discourse on the method') René Descartes formalised an idea that had been floating around in rudimentary form for some time: the notion of a coordinate system. In effect, this takes Euclid's barren plane, a blank sheet of paper, and turns it into paper ruled into squares, which engineers and scientists call graph paper. Draw two straight lines on the paper, one horizontal, the other vertical: these are

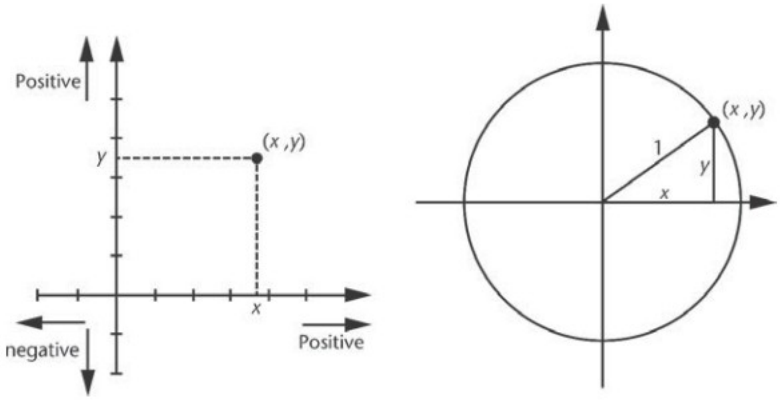

Fig 5 *Left*: Coordinates in the plane. *Right*: How to derive the equation for the unit circle.

called axes. Now you can pin down the location of any point of the plane by asking how far it lies in the direction along the horizontal axis, and how far up the vertical axis, Figure 5 (left). These two numbers, which may be positive or negative, provide a complete description of the point, and they are called its coordinates.

All geometric properties of points, lines, circles, and so on can be translated into algebraic statements about the corresponding coordinates. It's very difficult to talk meaningfully about these connections without using some actual algebra – just as it's hard to talk sensibly about football without mentioning the word 'goal'. So the next few pages will include some formulas. They are there to ensure that the main players in the drama have names and the relationship between them is clear. 'Romeo' is so much simpler to follow than 'the son of an Italian patriarch who falls in love with his father's sworn enemy's beautiful daughter'. Our Romeo will bear the prosaic name x, and his Juliet will be y.

As an example of how geometry converts into algebra, Figure 5 (right) shows how to find the equation for a circle of unit radius centred at the origin, where the two axes cross. The marked point has coordinates (x, y), so the right-angled triangle in the figure has horizontal side of length x and vertical side of length y. The longest side of the triangle is the radius of the circle, which is 1. Pythagoras's theorem now tells us that the sum of the squares of the two coordinates is 1. In symbols, a point with coordinates x and y lies on the circle if (and only if) it satisfies the condition $x^2 + y^2 = 1$. This symbolic characterisation of the circle is brief and precise, and it shows that we really are talking algebra. Conversely, any algebraic property of pairs of numbers, any equation involving x and y, can be reinterpreted as a geometric statement about points, lines, circles, or more elaborate curves.[19]

The basic equations of algebra involve polynomials, combinations of powers of an unknown quantity x, where each power is multiplied by some number, called a coefficient. The largest power of x that occurs is the degree of the polynomial. For example, the equation

$$x^4 - 3x^3 - 3x^2 + 15x - 10 = 0$$

involves a polynomial starting with x^4, so its degree is 4. The coefficients are 1, -3, -3, 15, and -10. There are four distinct solutions: $x = 1, 2, \sqrt{5}$, and $\sqrt{5}$. For these numbers the left-hand side of the equation is equal to zero – the right-hand side. Polynomials of degree 1, like $7x + 2$, are said to be linear, and they involve only the first power of the unknown. Equations of degree 2, like $x^2 - 3x + 2$, are said to be quadratic, and they involve the second power – the square. The equation for a circle involves a second variable, y. However, if we know a second equation relating x and y, for example the equation defining some straight line, then we can solve for y in terms of x and reduce the equation for a circle to one that involves only x. This new equation tells us where the line meets the circle. In this case the new equation is quadratic, with two solutions; this is how the algebra reflects the geometry, in which a line meets the circle at two distinct points.

This feature of the algebra has an important implication for ruler-and-compass constructions. Such a construction, however complicated, breaks up into a sequence of simple steps. Each step produces new points at places where two lines, two circles, or a line and a circle, meet. Those lines and circles are determined by previously constructed points. By translating geometry into algebra, it can be proved that the algebraic equation that corresponds to the intersection of two lines is always linear, while that for a line and a circle, or two circles, is quadratic. Ultimately this happens because the equation for a circle involves x^2 but no higher power of x. So every individual step in a construction corresponds to solving an equation of degree 1 or 2 only.

More complex constructions are sequences of these basic operations, and a certain amount of algebraic technique lets us deduce that each coordinate of any point that can be constructed by ruler and compass is a solution of a polynomial equation, with integer coefficients, whose degree is a power of 2. That is, the degree has to be one of the numbers 1, 2, 4, 8, 16, and so on.[20] This condition is necessary for a construction to exist, but it can be beefed up into a precise characterisation of which regular polygons are constructible. Suddenly a tidy algebraic condition emerges from a complicated geometric muddle – and it applies to *any construction whatsoever.* You don't even need to know what the construction is: just that it uses only ruler and compass.

Gauss was aware of this elegant idea. He also knew (indeed, any competent mathematician would quickly realise) that the question of which regular polygons can be constructed by ruler and compass boils down to a special case, when the polygon has a prime number of sides. To see why, think of a composite number like 15, which is 3×5. Any hypothetical construction of a 15-sided regular polygon automatically yields a 3-sided one (consider every fifth vertex) and a 5-sided one (consider every third vertex), Figure 6. With a bit more effort you can combine constructions for a 3-gon and a 5-gon to get a 15-gon.[21] The numbers 3 and 5 are prime, and the same idea applies in general. So Gauss focused on polygons with a prime number of sides, and asked what the relevant equation looked like. The answer was surprisingly neat. Constructing a regular 5-sided polygon, for example, is equivalent to solving the equation $x^5 - 1 = 0$. Replace 5 by any other prime, and the corresponding statement is true.

The degree of this polynomial is 5, which is *not* one of the powers of 2 that I listed; even so, a construction exists. Gauss quickly figured out why: the equation splits into two pieces, one of degree 1 and the other of degree 4. Both 1 and 4 are powers of 2, and it turns out that the degree-4 equation is the crucial one. To see why, we need to connect the equation to the geometry. That involves a new kind of number, one that is largely ignored in school mathematics but is indispensable for

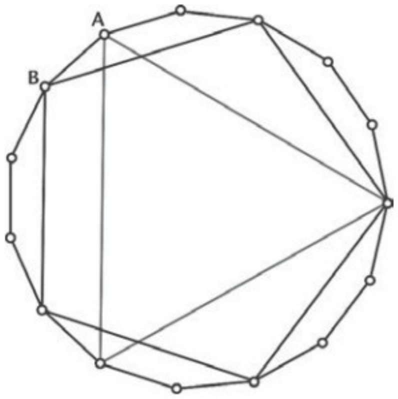

Fig 6 Constructing an equilateral triangle and a regular pentagon from a regular 15-gon. For the reverse, observe that A and B are consecutive points on the regular 15-gon.

anything beyond that. They are called complex numbers, and their defining feature is that in the complex number system −1 has a square root.[22]

An ordinary 'real' number is either positive or negative, and either way, its square is positive, so −1 can't be the square of any real number. This is such a nuisance that mathematicians invented a new kind of 'imaginary' number whose square is −1. They needed a new symbol for it, so they called it i (for 'imaginary'). The usual operations of algebra – adding, subtracting, multiplying, dividing – lead to combinations of real and imaginary numbers such as $3 + 2i$. These are said to be complex, which doesn't mean 'complicated', but indicates that they come in two parts: 3 and 2i. Real numbers lie on the famous number line, like the numbers on a ruler. Complex numbers lie in a number plane, in which an imaginary ruler is placed at right angles to a real one, and the two together form a system of coordinates, Figure 7 (left).

For the last 200 years, mathematicians have considered complex numbers to be fundamental to their subject. We now recognise that logically they are on the same footing as the more familiar 'real' numbers – which, like all mathematical structures, are abstract concepts, not real physical things. Complex numbers were in widespread use before the time of Gauss, but their status was still

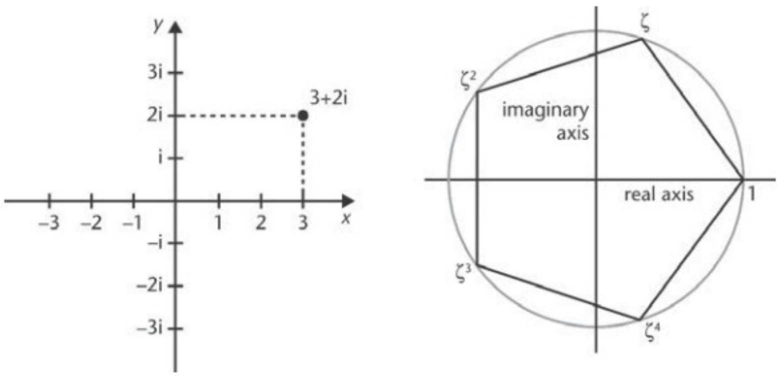

Fig 7 *Left*: The complex plane. *Right:* The complex fifth roots of unity.

mysterious until Gauss and several others demystified them. The source of their attraction was paradoxical: despite the mystery surrounding their meaning, complex numbers were much better behaved than real numbers. They supplied a missing ingredient that the real numbers lacked. They provided a complete set of solutions for an algebraic equation.

Quadratic equations are the simplest example. Some quadratics have two real solutions, while others have none. For example $x^2 - 1 = 0$ has the solutions 1 and -1, but $x^2 + 1 = 0$ has no solutions. In between is $x^2 = 0$, whose sole solution is 0, but there is a sense in which this is the same solution 'repeated twice'.[23] If we allow complex solutions, however, then $x^2 + 1 = 0$ also has two solutions: i and –i. Gauss had no qualms about using complex numbers; in fact, his doctoral thesis provided the first logically sound proof of the fundamental theorem of algebra: the number of complex solutions to any polynomial equation (with multiplicities counted correctly) is equal to the degree of the equation. So quadratics (degree 2) always have two complex solutions, cubics (degree 3) always have three complex solutions, and so on.

The equation $x^5 - 1 = 0$, which I claimed defines a regular pentagon, has degree 5. Therefore it has five complex solutions. There is just one real solution: $x = 1$. What about the other four? They provide four vertexes of a perfect regular pentagon in the complex plane, with $x = 1$ being the fifth, Figure 7 (right). This correspondence is an example of mathematical beauty: an elegant geometric shape becomes an elegant equation.

Now, the equation whose solutions are these five points has degree 5, which is not a power of 2. But, as mentioned earlier, the degree-5 equation splits into two pieces with degrees 1 and 4, called its irreducible factors:

$$x^5 - 1 = (x-1)\left(x^4 + x^3 + x^2 + x + 1\right)$$

('Irreducible' means that no further factors exist, just like prime numbers.) The first factor yields the real solution $x = 1$. The other factor yields the four complex solutions and the other four vertexes of the pentagon. So everything makes much more sense, and is far more elegant, when we use complex numbers.

It is often difficult to reconstruct how the mathematicians of the past arrived at new discoveries, because they had a habit of presenting only the final outcome of their deliberations, not the many false steps they took along the way. This problem is often compounded, because the natural thought patterns in past ages were different from today's. Gauss in particular was notorious for covering his tracks and publishing only his final, highly polished, analysis. But when it comes to Gauss's research on the 17-sided polygon, we are on fairly safe ground; the analysis that he eventually published provides several useful clues.

His starting-point was not new. Several earlier mathematicians were well aware that the above analysis of regular pentagons works in complete generality. Constructing a polygon with any number n of sides is equivalent to solving the equation $x^n - 1 = 0$ in complex numbers. Moreover, this polynomial factorises as

$$(x-1)\left(x^{n-1} + x^{n-2} + \cdots + x^2 + x + 1\right)$$

Again the first factor gives the real solution $x = 1$ and the remaining $n - 1$ solutions come from the second factor. When n is odd, these are all complex; when n is even, one of them is a second real solution $x = -1$.

What Gauss noticed, and everyone else had missed, is that sometimes the second factor can be expressed using a series of quadratic equations. Not by representing it as a product of simpler factors, because that's not possible, but by using equations whose coefficients solve other equations. The key fact here – the weak point in the problem – is an elegant property of algebraic equations, which arises when we solve several of them in turn in this manner. The calculation is always equivalent to solving a single equation, but the degree generally gets bigger. So the price we pay for having fewer equations is an increase in the degree. It can be messy, but there is one feature that we can predict: how big the degree becomes. Just multiply together the degrees of the successive polynomials.

If they are all quadratics, the result is $2 \times 2 \times \cdots \times 2$, a power of 2. So $n - 1$ must be a power of 2 if a construction exists. However, this condition is not always sufficient. When $n = 9$, $n - 1 = 8$, which is a power of 2. But Gauss discovered that no construction exists for the regular 9-gon. The reason is that 9 is not prime.[24] What about the next case, in which we solve a series of four quadratic equations? Now the degree $n - 1$ of the corresponding single equation is $2 \times 2 \times 2 \times 2 = 16$. So $n = 17$, and this is prime.

By this point Gauss must have known he was on to something, but there is a further technical point, possibly fatal. Gauss had convinced himself that in order for construction for a regular pentagon with a prime number of sides to exist, that prime must be a power of 2, plus 1. So this condition is necessary for a construction to exist: if it fails, there is no such construction. However, the condition might not be sufficient: in fact there are plenty of equations of degree 16 that do not reduce to a series of four quadratics.

There was a reason to be optimistic, however: the Greek constructions. Which primes occurred there? Only three: 2, 3, and 5. These are all 1 more than a power of 2, namely $2^0 + 1$, $2^1 + 1$, and $2^2 + 1$. The algebra associated with the pentagon provides further clues. Thinking it all through, Gauss proved that the degree-16 polynomial associated with the 17-sided polygon can indeed be reduced to a series of quadratics. Therefore a ruler-and-compass construction must exist. A similar method proved that the same is true whenever the number of sides is a prime that is 1 greater than some power of 2. The ideas are a tribute to Gauss's ability to understand mathematical patterns. At their heart are some general theorems in number theory, which I won't go into here. The point is, none of this was accidental. There were solid structural reasons for it to work. You just had to be a Gauss to notice them.

Gauss didn't provide an explicit construction, but he did give a formula for the solutions of the degree-16 equation that can be turned into such a construction if you really want one.[25] When he wrote down his ideas in the *Disquisitiones Arithmeticae*, he omitted quite a few of the details, but he did assert that he possessed complete proofs. His epic discovery convinced him that he should devote his life to mathematics rather than languages. The Duke continued to support Gauss financially, but Gauss wanted something more permanent and reliable. When the astronomer Giuseppe Piazzi discovered the first asteroid, Ceres, only a few observations could be made before this new world became invisible against the glare of the Sun. Astronomers were worried that they wouldn't be able to find it again. In a *tour de force* that involved new techniques for calculating orbits, Gauss predicted where it would reappear – and he was right. This led to him being appointed Professor of Astronomy and Director of the Göttingen Observatory. He continued to hold the post for the rest of his life.

It turns out that 17 is not the only new number of this kind. Two more are known: $2^8 + 1 = 257$ and $2^{16} + 1 = 65,537$. (A bit of algebra shows that the power of 2 that occurs must itself be a power of 2; if not, then the number cannot be prime.) However, the pattern stops at that point, because $2^{32} + 1 = 4,294,967,297$ is equal to $641 \times 6,700,417$, hence is not prime. The so-called Fermat numbers $2^{2^n} + 1$ are known not to be prime for $n = 5, 6, 7, \ldots$ up to 32. Many larger Fermat numbers are also known not to be prime. No further prime Fermat numbers have been found, but their existence is by no means impossible.[26] A construction for the 257-sided polygon is known. One mathematician devoted many years to the 65,537-sided polygon, a somewhat pointless task, and his results contain errors anyway.[27]

The upshot of Gauss's analysis is that a regular polygon can be constructed with ruler and compass if and only if the number of sides is a product of a power of 2 and *distinct* odd prime Fermat numbers. In particular, a regular 9-sided polygon cannot be constructed in this manner. This immediately implies that at least one angle cannot be trisected, because the angle in an equilateral triangle is 60 degrees, and one third of that is 20 degrees. Given this angle, it is easy to construct a regular 9-sided polygon. Since that is impossible, there is no general ruler-and-compass construction to trisect an angle.

Gauss omitted many details of the proofs when he wrote up his results, and mathematicians couldn't simply take his word for it. In 1837 the French mathematician Pierre Wantzel published a complete proof of Gauss's characterisation of constructible regular polygons, and deduced the impossibility of trisecting a general angle by ruler-and-compass construction. He also proved that it is impossible to construct a cube whose volume is twice that of a given cube, another ancient Greek problem known as 'duplicating the cube'.

Both angle-trisection and cube-duplication turn out to be impossible because the lengths involved satisfy irreducible *cubic* equations – degree 3. Since 3 is not a power of 2, this knocks the question on the head. However, this method didn't seem to work for the problem of squaring the circle, for interesting reasons. A circle of unit radius has area π, and square of that area has side $\sqrt{\pi}$. Geometric constructions for square roots exist, and so do constructions for squares, so squaring the circle boils down to starting with a line of length 1 and constructing one of length π. If π happened to satisfy an irreducible cubic equation – or any irreducible equation whose degree is not a power of 2 – then Wantzel's methods would prove that it's impossible to square the circle.

However, no one knew any algebraic equation that was satisfied exactly by π, let alone one whose degree is not a power of 2. The school value 22/7 satisfies $7x - 22 = 0$, but that's just an approximation to π, a tiny bit too large, so it doesn't help. If it could be proved that no such equation exists – and many suspected this on the grounds that it would have been found if it did – the impossibility of squaring the circle would follow. Unfortunately, no one could prove that there is no such equation. The algebraic status of π was in limbo. The eventual solution employed methods that did not just go beyond geometry: they went beyond algebra as well.

To appreciate the main issue here, we need to start with a simpler idea. There is an important distinction in mathematics between numbers that can be expressed as exact fractions p/q, where p and q

are whole numbers, and those that can't be so expressed. The former are said to be rational (they are ratios of whole numbers), and the latter are irrational. The approximation 22/7 to π is rational, for example. There are better approximations; a famous one is 355/113, correct to six decimal places. However, it is known that no fraction can represent π exactly: it is irrational. This long-suspected property was first proved by the Swiss mathematician Johann Heinrich Lambert in 1768. His proof is based on a clever formula for the tangent function in trigonometry, which he expressed as a continued fraction: an infinite stack of ordinary fractions.[28] In 1873 Charles Hermite found a simpler proof, based on formulas in calculus, which went further: it proved that π^2 is irrational. Therefore π is not the square root of a rational number either.

Lambert suspected something much stronger. In the article that proved π to be irrational, he conjectured that π is transcendental; that is, π does not satisfy any polynomial equation with integer coefficients. It transcends algebraic expression. Subsequent discoveries proved him right. The breakthrough came in two stages. Hermite's new method for proving irrationality set the scene by hinting that calculus – more precisely, its rigorous version, analysis – might be a useful strategy. By pushing that idea further, Hermite found a wonderful proof that the other famous curious number in mathematics, the base e of natural logarithms, is transcendental. Numerically, e is roughly 2.71828, and if anything it is even more important than π. Hermite's transcendence proof is magical, a rabbit extracted with a flourish from the top hat of analysis. The rabbit is a complicated formula related to a hypothetical algebraic equation that e is assumed to satisfy. Using algebra, Hermite proves that this formula is equal to some nonzero integer. Using analysis, he proves that it must lie between $-\frac{1}{2}$ and $\frac{1}{2}$. Since the only integer in this range is zero, these results are contradictory. Therefore the assumption that e satisfies an algebraic equation must be false, so e is transcendental.

In 1882 Ferdinand Lindemann added some bells and whistles to Hermite's method, and proved that if a nonzero number satisfies an algebraic equation, then e raised to the power of that number does not satisfy an algebraic equation. He then took advantage of a relationship that was known to Euler involving π, e, and the imaginary number i: the famous formula $e^{i\pi} = -1$. Suppose that π satisfies some algebraic equation. Then so does $i\pi$, and Lindemann's theorem implies that -1 does *not* satisfy an algebraic equation. However, it visibly does: it is the solution of $x + 1 = 0$. The only way out of this logical contradiction is that π does not satisfy an algebraic equation; that is, it is transcendental. And that means you can't square the circle.

It was a long and indirect journey from Euclid's geometry to Lindemann's proof, and it took more than 2000 years, but mathematicians finally got there. The story doesn't just tell us that the circle can't be squared. It is an object lesson in how great mathematical problems get solved. It required mathematicians to formulate carefully what they meant by 'geometric construction'. They had to pin down general features of such constructions that might place limits on what they could achieve. Finding those features required making connections with another area of mathematics: algebra. Solving the algebraic problem, even in simpler cases such as the construction of regular polygons, also involved number theory. Dealing with the difficult case of π required further innovations, and the problem had to be transported into yet another area of mathematics: analysis.

None of these steps was simple or obvious. It took about a century to complete the proof, even when the main ideas were in place. The mathematicians involved were among the best of their age, and at least one was among the best of *any* age. Solving great problems requires a deep understanding of mathematics, plus persistence and ingenuity. It can involve years of concentrated effort, most of it apparently fruitless. But imagine how it must feel when your persistence pays off, and you crack wide open something that has baffled the rest of humanity for centuries. As President John F. Kennedy said in 1962 when announcing the Moon landing project: 'We choose to ... do [these] ... things, not because they are easy, but because they are hard.'

Few stories in mathematics end, and π is no exception. Every so often, amazing new discoveries about π appear. In 1997 Fabrice Bellard announced that the trillionth digit of π, in binary notation, is 1.[29] What made the statement remarkable was not the answer. The amazing feature was that he didn't calculate any of the earlier digits. He just plucked one particular digit from the air.

The calculation was made possible by a curious formula for π discovered by David Bailey, Peter Borwein, and Simon Plouffe in 1996. It may seem a bit complicated, but let's take a look anyway:

$$\pi = \sum_{n=0}^{\infty} \frac{1}{2^{4n}} \left(\frac{4}{8n+1} - \frac{2}{8n+4} - \frac{1}{8n+5} - \frac{1}{8n+6} \right)$$

The big Σ means 'add up', over the range specified. Here n runs from 0 to infinity (∞). Bellard actually used a formula that he had derived using similar methods, which is marginally faster for computations:

$$\pi = \frac{1}{64} \sum_{n=0}^{\infty} \frac{(-1)^n}{2^{10n}} \left(-\frac{32}{4n+1} - \frac{1}{4n+3} + \frac{256}{10n+1} - \frac{64}{10n+3} \right.$$
$$\left. - \frac{4}{10n+5} - \frac{4}{10n+7} + \frac{1}{10n+9} \right)$$

The key point is that many of the numbers that occur here – 1, 4, 32, 64, 256, and also 2^{4n} and 2^{10n} – are powers of 2, which of course are very simple in the binary system used for the internal workings of computers. This discovery stimulated a flood of new formulas for π, and for several other interesting numbers. The record for finding a single binary digit of π is broken regularly: in 2010 Yahoo's Nicholas Sze computed the two-quadrillionth binary digit of π, which turns out to be 0.

The same formulas can be used to find isolated digits of π in arithmetic to the bases 4, 8, and 16. Nothing of the kind is known for any other base; in particular, we can't compute decimal digits in isolation. Do such formulas exist? Until the Bailey-Borwein-Plouffe formula was found, no one imagined it could be done in binary.

whole numbers, you now allow yourself to replace any multiple of 12 by zero. For example, $5 \times 5 = 25$, but 24 is twice 12, so subtracting 24 we obtain $5 \times 5 = 1$ to the modulus 12. Modular arithmetic is very pretty, because nearly all of the usual rules of arithmetic still work. The main difference is that you can't always divide one number by another, even when it's not zero. Modular arithmetic is also useful, because it provides a tidy way to deal with questions about divisibility: which numbers are divisible by the chosen modulus, and what is the remainder when they're not? Gauss introduced modular arithmetic in the *Disquisitiones Arithmeticae*, and today it is widely used in computer science, physics, and engineering, as well as mathematics.

Fermat's theorem states that if we choose a prime modulus p, and take any number a that is not a multiple of p, then the $(p-1)$ th power of a is equal to 1 in arithmetic to the modulus p. Suppose, for example, that $p = 17$ and $a = 3$. Then the theorem predicts that when we divide 3^{16} by 17, the remainder is 1. As a check,

$$3^{16} = 43,046,721 = 2,532,160 \times 17 + 1$$

No one in their right mind would want to do the sums that way for, say, 100-digit primes. Fortunately, there is a clever, quick way to carry out this kind of calculation. The point is that if the answer is not equal to 1 then the modulus we started with is composite. So Fermat's theorem forms the basis of an efficient test that provides a necessary condition for a number to be prime.

Unfortunately, the test is not sufficient. Many composite numbers, known as Carmichael numbers, pass the test. The smallest is 561, and in 2003 Red Alford, Andrew Granville, and Carl Pomerance proved, to general amazement, that there are infinitely many. The amazement was because they found a proof; the actual result was less of a surprise. In fact, they showed that there are at least $x^{2/7}$ Carmichael numbers less than or equal to x if x is large enough.

However, more sophisticated variants of Fermat's theorem can be turned into genuine tests for primality, such as one published in 1976 by Gary Miller. Unfortunately, the proof of the validity of Miller's test depends on an unsolved great problem, the generalised Riemann hypothesis, chapter 9. In 1980 Michael Rabin turned Miller's test into a probabilistic one, a test that might occasionally give the wrong answer. The exceptions, if they exist, are very rare, but they can't be ruled out altogether. The most efficient deterministic (that is, guaranteed correct) test to date is the Adleman-Pomerance-Rumely test, named for Leonard Adleman, Pomerance, and Robert Rumely. It uses ideas from number theory that are more sophisticated than Fermat's theorem, but in a similar spirit.

I still vividly recall a letter from one hopeful amateur, who proposed a variant of trial division. Try all possible divisors, but start at the square root and work *downwards*. This method sometimes gets the answer more quickly than doing things in the usual order, but as the numbers get bigger it runs into the same kind of trouble as the usual method. If you try it on my example above, the 22-digit number 1,080,813,321,843,836,712,253, then the square root is about 32,875,725,419. You have to try 794,582,971 prime divisors before you find one that works. This is *worse* than searching in the usual direction.

In 1956 The famous logician Kurt Gödel, writing to John von Neumann, echoed Gauss's plea. He asked whether trial division could be improved, and if so, by how much. Von Neumann didn't pursue the question, but over the years others answered Gödel by discovering practical methods for finding primes with up to 100 digits, sometimes more. These methods, of which the best known is called the quadratic sieve, have been known since about 1980. However, nearly all of them are either probabilistic, or they are inefficient in the following sense.

How does the running time of a computer algorithm grow as the input size increases? For primality testing, the input size is not the number concerned, but how many digits it has. The core distinction in such questions is between two classes of algorithms called P and not-P. If the running time grows like some fixed power of the input size, then the algorithm is class P; otherwise, it's not-P. Roughly speaking, class P algorithms are useful, whereas not-P algorithms are impractical, but there's a stretch of no-man's-land in between where other considerations come into play. Here P stands for 'polynomial time', a fancy way to talk about powers, and we return to the topic of efficient algorithms in chapter 11.

By the class P standard, trial division performs very badly. It's all right in the classroom, where the numbers that occur have two or three digits, but it's completely hopeless for 100-digit numbers. Trial division is firmly in the not-P class. In fact, the running time is roughly $10^{n/2}$ for an n-digit number, which grows faster than any fixed power of n. This type of growth, called exponential, is *really* bad, computational cloud-cuckoo-land.

Until the 1980s all known algorithms for primality testing, excluding probabilistic ones or those whose validity was unproved, had exponential growth rate. However, in 1983 an algorithm was found that lies tantalisingly in the no-man's-land adjacent to P territory: the aforementioned Adleman-Pomerance-Rumely test. An improved version by Henri Cohen and Hendrik Lenstra has running time n raised to the power log log n, where log denotes the logarithm. Technically, log log n can be as large as we wish, so this algorithm is not in class P. But that doesn't prevent it being practical: if n is a googolplex, 1 followed by 10^{100} zeros, then log log n is about 230. An old joke goes: 'It has been proved that log log n tends to infinity, but it has never been observed doing it.'

The first primality test in class P was discovered in 2002 by Manindra Agrawal and his students Neeraj Kayal and Nitin Saxena, who were undergraduates at the time. I've put some details in the Notes.[10] They proved that their algorithm has running time proportional to at most n^{12}; this was quickly improved to $n^{7.5}$. However, even though their algorithm is class P, hence classed as 'efficient', its advantages don't show up until the number n becomes very large indeed. It should beat the Adleman-Pomerance-Rumely test when the number of *digits* in n is about 10^{1000}. There isn't room to fit a number that big into a computer's memory, or, indeed, into the known universe. However, now that we *know* that a class P algorithm for primality testing exists, it becomes worthwhile to look for better ones. Lenstra and Pomerance reduced the power from 7.5 to 6. If various other conjectures about primes are true, then the power can be reduced to 3, which starts to look practical.

The most exciting aspect of the Agrawal-Kayal-Saxena algorithm, however, is not the result, but the method. It is simple – to mathematicians, anyway – and novel. The underlying idea is a variant of Fermat's theorem, but instead of working with numbers, Agrawal's team used a polynomial. This is a combination of powers of a variable x, such as $5x^3 + 4x - 1$. You can add, subtract, and multiply polynomials, and the usual algebraic laws remain valid. Chapter 3 explains polynomials in more detail.

This is a truly lovely idea: expand the domain of discourse and transport the problem into a new realm of thought. It is one of those ideas that are so simple you have to be a genius to spot them. It developed from a 1999 paper by Agrawal and his PhD supervisor Somenath Biswas, giving a probabilistic primality test based on an analogue of Fermat's theorem in the world of polynomials. Agrawal was convinced that the probabilistic element could be removed. In 2001 his students came up with a crucial, rather technical, observation. Pursuing that led the team into deep number-theoretic waters, but eventually everything was reduced to a single obstacle, the existence of a prime p such that $p - 1$ has a sufficiently large prime divisor. A bit of asking around and searching the Internet led to a theorem proved by Etienne Fouvry in 1985 using deep and technical methods. This was exactly what they needed to prove that their algorithm worked, and the final piece of the jigsaw slotted neatly into place.

In the days when number theory was safely tucked away inside its own little ivory tower, none of this would have mattered to the rest of the world. But over the last 20 years, prime numbers have become important in cryptography, the science of secret codes. Codes aren't just important for military use; commercial companies have secrets too. In this Internet age, we all do: we don't want criminals to gain access to our bank accounts, credit card numbers, or, with the growth of identity theft, the name of our cat. But the Internet is such a convenient way to pay bills, insure cars, and book holidays, that we have to accept some risk that our sensitive, private information might fall into the wrong hands.

Computer manufacturers and Internet service providers try to reduce that risk by making various encryption systems available. The involvement of computers has changed both cryptography and cryptanalysis, the dark art of code-breaking. Many novel codes have been devised, and one of the most famous, invented by Ron Rivest, Adi Shamir, and Leonard Adleman in 1978, uses prime numbers. Big ones, about a hundred digits long. The Rivest-Shamir-Adleman system is employed in many computer operating systems, is built into the main protocols for secure Internet communication, and is widely used by governments, corporations, and universities. That doesn't mean that every new result about primes is significant for the security of your Internet bank account, but it adds a definite frisson of excitement to any discovery that relates primes to computation. The Agrawal-Kayal-Saxena test is a case in point. Mathematically, it is elegant and important, but it has no direct practical significance.

It does, however, cast the general issue of Rivest-Shamir-Adleman cryptography in a new and slightly disturbing light. There is still no class P algorithm to solve Gauss's second problem, factorisation. Most experts think nothing of the kind exists, but they're not quite as sure as they used to be. Since new discoveries like the Agrawal-Kayal-Saxena test can lurk unsuspected in the wings, based on such simple ideas as polynomial versions of Fermat's theorem, cryptosystems based on prime factorisation might not be quite as secure as we fondly imagine. Don't reveal your cat's name on the Internet just yet.

Even the basic mathematics of primes quickly leads to more advanced concepts. The mystery becomes even deeper when we ask subtler questions. Euclid proved that the primes go on for ever, so we can't just list them all and stop. Neither can we give a simple, useful algebraic formula for successive primes, in the way that x^2 specifies squares. (There do exist simple formulas, but they 'cheat' by building the primes into the formula in disguise, and don't tell us anything new.[11]) To grasp the nature of these elusive, erratic numbers, we can carry out experiments, look for hints of structure, and try to prove that these apparent patterns persist no matter how large the primes become. For instance, we can ask how the primes are distributed among all whole numbers. Tables of primes strongly suggest that they tend to thin out as they get bigger. Table 1 shows how many primes there are in various ranges of 1000 consecutive numbers.

The numbers in the second column mostly decrease as we move down the rows, though sometimes there are brief periods when they go the other way: 114 is followed by 117, for instance. This is a symptom of the irregularity of the primes, but despite that, there is a clear general tendency for primes to become rarer as their size increases. The

range	number of primes
1–1000	168
1001–2000	135
2001–3000	127
3001–4000	119
4001–5000	118
5001–6000	114
6001–7000	117
7001–8000	106
8001–9000	110
9001–10,000	111

Table 1 The number of primes in successive intervals of 1000 numbers.

reason is not far to seek: the bigger a number becomes, the more potential factors there are. Primes have to avoid all of these factors. It's like fishing for non-primes with a net: the finer the net becomes, the fewer primes slip through.

The 'net' even has a name: the sieve of Eratosthenes. Eratosthenes of Cyrene was an ancient Greek mathematician who lived around 250 BC. He was also an athlete with interests in poetry, geography, astronomy, and music. He made the first reasonable estimate of the size of the Earth by observing the position of the Sun at noon in two different locations, Alexandria and Syene – present-day Aswan. At noon, the Sun was directly overhead at Syene, but about 7 degrees from the vertical at Alexandria. Since this angle is one fiftieth of a circle, the Earth's circumference must be 50 times the distance from Alexandria to Syene. Eratosthenes couldn't measure that distance directly, so he asked traders how long it took to make the journey by camel, and estimated how far a camel typically went in a day. He gave an explicit figure in a unit known as a *stadium*, but we don't know how long that unit was. Historians generally think that Eratosthenes's estimate was reasonably accurate.

His sieve is an algorithm to find all primes by successively eliminating all multiples of numbers already known to be prime. Figure 2 illustrates the method on the numbers up to 102, arranged to

Pág. 23(36) -

In 1956, the famous logician Kurt Gödel, writing to John von Neumann, echoed) Gauss's plea - (primer)

_20211016/ Cálculo/ matesimp.pdf
Obs: Matemáticas Simplificadas
1641 pages.
⇒ Livro todo!!!

_20211016/ Einstein/
Besse - Einstein Manifolds ... (1987).pdf
Obs Gauss - 34 occurrences
Pág. 02(13) 0.5 Introduction ↙
Important:

For surfaces ($n=2$) the answer to the above question is known.

The best Riemannian structures on a given compact surface M are those of constant curvature.

In dimension 2 there is only one notion of curvature, namely the Gauss curvature, which is a function $K: M \to \mathbb{R}$. If M is embedded in $\mathbb{R}^3$

B
! -256- - -* ! - B ! 1 1"
*D & 1 ! { UUU

D -256- - -* ! - B) = ! " 1
) : " D / "KA

-256- - -* ! - B) = ! " 1
) : " D / 1"
(% $ & "

to the *compact* case. If one looks on Einstein manifolds as critical points of the U functional (see 0.12 or Chapter 4) then it is reasonable to hope for an existence result in the compact case (even if this approach is disappointing in the end, see 4.1). The non-compact case remains interesting nevertheless (see Chapter 15).

0.5. For surfaces ($n = 2$) the answer to the above question is known. The best Riemannian structures on a given compact surface M are those of *constant curvature*. In dimension 2, there is only one notion of curvature, namely the Gauss curvature, which is a function $K\colon M \to \mathbb{R}$. If M is embedded in $\mathbb{R}^3$ with the induced natural Riemannian metric, the Gauss curvature is the product of the principal curvatures, or the inverse of the product of the principal radii of curvature. There exists at least one Riemannian metric of constant curvature on any compact surface and the constant can be normalized by scaling to $+1, 0, -1$. Moreover, on a given M, the Riemannian structures of constant given curvature 1, 0 or -1 form a nice finite dimensional submanifold—with singularities—of $\mathscr{M}/\mathfrak{D}$, the so-called *Moduli Space*. For example, on S^2 (the 2-dimensional sphere) and on $\mathbb{R}P^2$ (the real projective plane) there exists exactly one Riemannian structure of constant curvature equal to 1. On a compact surface M, orientable or not, with Euler characteristic $\chi(M) < 0$, the Riemannian structures with constant curvature -1 depend on $-3\chi(M)$ real parameters. For example, $6\gamma - 6$ parameters when M is orientable and of genus γ. For more detail, see 12.B.

0.6. What would a natural generalization of the concept of constant curvature be, for a Riemannian manifold, when the dimension is greater than 2? We claim that a good generalization would be the notion of *constant Ricci curvature*.

0.7. For Ricci curvature the reader can refer to Chapter 1. Here we shall simply say this. The Ricci curvature of a Riemannian manifold (M, g) is a quadratic differential form (alias a bilinear symmetric form), *denoted by* r (or r_g if the metric needs to be specified). By diagonal restriction r defines a function (still denoted by r) on the unit tangent bundle UM of M. This function r is the trace of the curvature endomorphism defined as follows by the curvature 4-tensor R: $r(x) = \operatorname{trace}(z \mapsto R(x, z)x)$. Equivalently, for x in UM, the value $r(x)$ is the sum $\sum_{i=2}^{n} K(x, x_i)$ of the sectional curvatures $K(x, x_i)$ for any completion of x to an orthonormal basis $(x, x_2, \ldots, x_n)$.

0.8. A Riemannian manifold has three main notions of curvature. i) The Riemann *curvature tensor* R (equivalent to the *sectional curvature* function, K, defined on tangent planes) which is a biquadratic form giving complete information, at the curvature level, on g. ii) The Ricci curvature above, r, the trace of R with respect to g. iii) Finally, the scalar curvature, s, a scalar function on M, which is the trace $s = \operatorname{trace}_g r$ with respect to g of the quadractic form r. Note first that when $n = 2$ the three curvatures are equivalent. When $n = 3$ the Ricci curvature contains as much information as the Riemann curvature tensor. This is the main reason why our book is devoted to manifolds of dimension greater than or equal to 4.

1.61 Definitions. Let (M, g) be a Riemannian manifold.

(a) Given a piecewise smooth curve $c\colon [a, b] \to M$, the *length* of c is

$$L(c) = \int_a^b \sqrt{g(\dot{c}, \dot{c})}\, dt.$$

(b) For each pair of points x and y in M, we *denote* by $d(x, y)$ the infimum of the lengths of all piecewise smooth curves starting from x and ending at y.

Note that in (a), if c is a geodesic, then $g(\dot{c}, \dot{c})$ is constant and $L(c) = (b - a)\sqrt{g(\dot{c}, \dot{c})}$; in (b), the infimum $d(x, y)$ may or may not be realized by a curve.

1.62 Theorem. *Given a Riemannian manifold (M, g), the function d is a distance on M, and the topology of the metric space (M, d) is the same as the manifold topology of M.*

A corollary of the Gauss Lemma 1.46 is that the distance is realized (at least locally) by geodesics. More precisely

1.63 Theorem. *For each x in M, there is a neighborhood U_x in M such that, for any y in U_x, the distance $d(x, y)$ is the length of the unique geodesic from x to y in U_x.*

1.64 Corollary. *Any geodesic minimizes the length between any pair of sufficiently near points on it; conversely, any curve having this property is (up to reparameterization) a geodesic.*

Also there is a notion of completeness of a metric space. Fortunately, these notions are as compatible as they can be. This is the content of the following theorem.

1.65 Theorem (H. Hopf-W. Rinow). *For a Riemannian manifold (M, g), the following conditions are equivalent.*

a) *(M, g) is complete for the Levi-Civita connection;*

b) *(M, d) is a complete metric space;*

c) *the bounded subsets of M are relatively compact.*

And these properties imply that

d) *for any two points x, y in M, there exists at least a geodesic starting at x and ending at y.*

Note however that there may be many more than one geodesic connecting x and y and that property d) does not imply the completeness of (M, g).

1.66 Corollary. *A compact Riemannian manifold is complete.*

(This need not be true for pseudo-Riemannian manifolds, even in the homogeneous case, see Chapter 7).

Note that a geodesic connecting two points does not necessarily realize the component of the fixed point set of the isometry group generated by reflection of $\mathbb{R}$ about t and reflections of S^2 about great circles through p. The vector field $\frac{\partial}{\partial t}$ is a Jacobi field along $\gamma_{(t,p)}$ and hence

$$K\left(\frac{\partial}{\partial t} \wedge \frac{\partial}{\partial \rho}\right) = \left\langle D^2_{\partial/\partial\rho}\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right\rangle \Big/ \left\|\frac{\partial}{\partial t} \wedge \frac{\partial}{\partial \rho}\right\|^2 = F''/F. \tag{3.35a}$$

Similarly, for each unit vector x tangent to S^2, $\partial/\partial\theta$ is a Jacobi field along γ, where θ is the angular coordinate on the great circle S^1 in S^2 generated by x, and hence

$$K\left(\frac{\partial}{\partial \rho} \wedge x\right) = \left\langle D^2_{\partial/\partial\rho}\frac{\partial}{\partial \theta}, \frac{\partial}{\partial \theta}\right\rangle \Big/ \left\|\frac{\partial}{\partial \theta} \wedge \frac{\partial}{\partial \rho}\right\|^2 = G''/G. \tag{3.35b}$$

In other words, we have computed the curvature of 2-dimensional totally geodesic subspaces in Gaussian coordinates.

We can obtain the rest of the relevant sectional curvatures by observing that on each $\rho =$ constant hypersurface $\mathbb{R} \times \{\rho\} \times S^2$ the induced metric is the product metric

$$g_\rho = F^2(\rho)\,dt^2 - G^2(\rho)\,d\sigma^2$$

and the second fundamental form of this hypersurface with respect to the unit normal vector field $\frac{\partial}{\partial \rho}$ is given by $\left(\frac{D}{\partial \rho} g_\rho\right)(x, y) = -2g_\rho(S \cdot x, y)$, hence

$$S = \begin{pmatrix} -FF' & 0 \\ 0 & -GG'h \end{pmatrix},$$

where $h = d\sigma^2$. Hence, by the Gauss equation,

$$K\left(\frac{\partial}{\partial t} \wedge x\right) = \frac{F'}{F}\cdot\frac{G'}{G}, \tag{3.35c}$$

whenever x is tangent to S^2, and

$$K(x \wedge y) = -\frac{1}{G^2} + \left(\frac{G'}{G}\right)^2, \tag{3.35d}$$

whenever both x and y are tangent to S^2.

3.36. It follows that the eigenvalues of the Ricci tensor are

$$r_t = \frac{F''}{F} + 2\frac{F'}{F}\cdot\frac{G'}{G}, \tag{3.36a}$$

$$r_\rho = \frac{F''}{F} + 2\frac{G''}{G}, \tag{3.36b}$$

and

$$r_x = \frac{G''}{G} + \frac{F'}{F}\cdot\frac{G'}{G} - \frac{1}{G^2} + \left(\frac{G'}{G}\right)^2 \qquad \text{(repeated)}. \tag{3.36c}$$

(B % '' 5 ! .

See [Koo] or [Laf 3] for further study of the equation $s_g'^* \cdot f = 0$. It is not difficult to see that the preceding proposition does not give the best possible result. Indeed, for the standard product metric on $S^2 \times S^3$, $s/4$ is an eigenvalue of the Laplacian, but $\operatorname{Ker} s_g'^* = 0$. Anyway, the condition that $s_g/(n-1)$ does not belong to the spectrum will appear again as an important property in the next paragraph.

4.38 Lemma (see [Ka-Wa 2] for the proof). *If* $\dim M \geqslant 2$ *and if* $f \in C^0(M)$, *then an* L^p*-function belongs to the* L^p*-closure of the set* $\{f \circ \varphi, \varphi \in \mathfrak{D}\}$ *if and only if* $\inf f \leqslant f_1(x) \leqslant \sup f$ *almost everywhere.*

4.39 Lemma. *Let* $f_0 \in \mathscr{C}M$ *be the scalar curvature of some metric. If* $f_1 \in \mathscr{C}M$ *is such that, for a positive real number* λ, *the inclusion* $\lambda(\operatorname{Im}(f_0)) \subset \operatorname{Im}(f_1)$ *holds, then* f_1 *is also the scalar curvature of some metric.*

Proof of the lemma. After a homothetic change of metrics, we may suppose that $\lambda = 1$. If $f_0 = s_{g_0}$ and $\operatorname{Ker} s_{g_0}'^* = 0$, apply the inverse function theorem 4.36 to g_0, and take φ in $\mathfrak{D}$ such that $\|f - f_1 \circ \varphi\|_p < \varepsilon$. If $\operatorname{Ker} s_{g_0}'^* \neq 0$ (which by 4.37 implies that s_{g_0} is constant), we can find a metric g_2 such that $\operatorname{Im}(s_{g_2}) \subset \operatorname{Im}(f_1)$ still holds (take g_2 close enough to g_0), but s_{g_2} is no longer a constant. Then apply the preceding argument to g_2.

4.40. *End of the proof of Theorem* 4.35. If M carries a metric whose scalar curvature is non-negative everywhere, but not identically zero, from 4.32 ii), there are also on M metrics with constant positive, constant negative and zero scalar curvatures. Using Lemma 4.39 with those constant functions as functions f_0, we get that any function is a scalar curvature. The other cases follow just in the same way. That $s_g = 0$ implies $r_g = 0$ in case (B) will be proved in 4.49. □

4.41 Remarks. i) If $\dim M = 2$, using the same technique and only replacing Theorem 4.32 by the conformal representation theorem, one gets the following converse to the Gauss-Bonnet theorem ([Ka-Wa 1]): if $\chi(M) > 0$ (resp. $\chi(M) = 0$, resp. $\chi(M) < 0$) a function f on M is the gaussian curvature of some metric if and only if f is positive somewhere (resp. is identically zero or takes both signs, resp. is negative somewhere).

ii) The use of Sobolev spaces L_k^p (and not only L_k^2 as in most discussions of linear elliptic theory) was crucial: some L^p-norm is needed for the approximation lemma 4.38, which is trivially false for the uniform norm. Then the use of the inverse function theorem in 4.36 makes it necessary to take $p > \dim M$.

F. The Manifold of Metrics with Constant Scalar Curvature[1])

4.42. Since the manifold $\mathscr{M}$ of Riemannian metrics and its submanifolds are Fréchet manifolds modelled on $\mathscr{S}^2M$, the usual inverse function theorem cannot be applied to them. This difficulty is taken care of by the technique of ILH-manifolds, developed

---

[1]) This section may be skipped on first reading.

If $R_{XYZU} \neq 0$, the coefficient matrix satisfies the cofactor relations $(\lambda - \xi)(\lambda + \xi - \mu - \nu) = (\nu - \lambda)(\nu + \lambda - \mu - \xi) = (\lambda - \mu)(\lambda + \mu - \nu - \xi) = 0$, which easily imply that λ is equal to one of μ, ν, ξ. Therefore, evaluating R_x on four eigenvectors of B_x yields zero if more than two eigenspaces are involved. Hence $R_{XY}Z = 0$ if λ, μ, ν are mutually distinct. On the other hand, if $\lambda = \mu \neq \nu$, (16.16) gives $0 = B(R^G_{XY}Z + R^G_{YZ}X + R^G_{ZX}Y) = R_{XY}(BZ) + R_{YZ}(BX) + R_{ZX}(BY) = (\nu - \lambda)R_{XY}Z$, which completes the proof. ☐

16.17 Corollary (J.P. Bourguignon, [Bou 9]). *Let b be a Codazzi tensor on a Riemannian manifold (M, g). Then*

(i) *b commutes with r, $\mathring{R}(b)$ and with $Dd(\mathrm{tr}_g b) + \delta Db$.*

(ii) *The endomorphisms R, $g \owedge r$* and *W* of *$\bigwedge^2 TM$* commute with *$g \owedge b$.*

Proof. By 16.14, $r(X, Y) = \mathring{R}(b)(X, Y) = 0$ for eigenvectors X, Y of b corresponding to distinct eigenvalues, so that r and $\mathring{R}(b)$ commute with b. Hence (i) follows from Formula (16.9) (which also directly implies that r and b commute). Since $\bigwedge^2 T_xM$ is spanned by the subspaces $V_\lambda(x) \wedge V_\mu(x)$ and $(g \owedge b)_x$ restricted to such a subspace is $\lambda + \mu$ times the identity, (ii) is immediate from 16.14. ☐

16.18. An orthonormal basis $(X_1, \ldots, X_n)$ of a Euclidean space E is said to *diagonalize* an algebraic curvature tensor $R \in \mathscr{C}E$ if all exterior products $X_i \wedge X_j$, $i < j$, are eigenvectors of R (viewed as an endomorphism of $\bigwedge^2 E$), i.e., if $R(X_i, X_j)X_k = 0$ whenever i, j, k are mutually distinct. Following H. Maillot ([Mai 1], [Mai 2]) we shall call $R \in \mathscr{C}E$ *pure* if it is diagonalized by some orthonormal basis of E. A Riemannian manifold (M, g) will be said to have *pure curvature operator* (resp., *pure Weyl tensor*) if for any $x \in M$, R_x (resp., W_x) is pure. If $n = \dim E = 2$, every orthonormal basis of E diagonalizes any $R \in \mathscr{C}E$, while, for $n \geqslant 3$, such a basis diagonalizes R if and only if it diagonalizes $W(R)$ and $c(R)$. Thus, each of the following conditions implies that the Riemannian manifold (M, g) has pure curvature operator:

(i) $\dim M = 2$, or $W = 0$ (e.g., $\dim M = 3$, cf. 1.119);

(ii) (M, g) is a hypersurface in a space of constant curvature (cf. the Gauss equation (1.72c));

(iii) (M, g) is a Riemannian product of manifolds with pure curvature operators.

Moreover, it is obvious that

(iv) the property of having pure Weyl tensor is conformally invariant.

16.19. If $R \in \mathscr{C}E$ has pure Weyl component $W = W(R)$ (e.g., if R is pure), then all the Pontryagin forms $p_i(R) \in \bigwedge^{4i} E^*$, $i \geqslant 1$, are zero (see [Mai 1]). In fact, it is easy to see (cf. [Gre]) that $p_i(R) = p_i(W)$ and the subalgebra $p(R) = p(W)$ of $\bigwedge E^*$ generated by the $p_i(W)$ has another system of generators $\Omega_i \in \bigwedge^{4i} E^*$, Ω_i being obtained by alternating the map $(Y_1, \ldots, Y_{4i}) \to \mathrm{tr}[W(Y_1 \wedge Y_2) \circ \cdots \circ W(Y_{4i-1} \wedge Y_{4i})]$. However, if an orthonormal basis $(X_1, \ldots, X_n)$ diagonalizes W, then $W(X_i \wedge X_j) \circ W(X_k \wedge X_l) = 0$ for mutually distinct i, j, k, l, which implies $p(W) = \bigwedge^0 E^* = \mathbb{R}$.

16.20 Lemma. *Let E be an oriented 4-dimensional Euclidean space and let $W \in \mathscr{W}(E)$*

it may be useful to point out that they are both optimal and easy to discover. To see this, for example with $l = 0$, let $\varphi \in C_0^\infty(|x| < 1)$, $\varphi \not\equiv 0$, and let $f_\lambda(x) = \varphi(\lambda x) \in C_0^\infty(|x| < 1)$ for $\lambda \geqslant 1$. Applying (8) to the f_λ and doing a brief computation, one obtains

$$\|\varphi\|_{L^q} \leqslant c\lambda^{k+n(1/q-1/p)}\|\varphi\|_{p,k}.$$

Letting $\lambda \to \infty$ there is a contradiction unless (7) holds. This example uses the conformal map $x \to \lambda x$; it leads one to suspect that conformal defomations of metrics lead one to the limiting case of the Sobolev inequality. This suspicion is verified in Chapter 4.

D. Differential Operators

Here and below we will use the notation $F(x, \partial^k u)$, $a(x, \partial^k u)$, etc. to represent any (possibly nonlinear) differential operator of order k. We use local coordinates on $\mathbb{R}^n$ and consider a differential operator on vector valued functions u:

$$F(x, \partial^k u) = \sum_{|\alpha|=k} a_\alpha(x, \partial^l u)\partial^\alpha u + f(x, \partial^l u),$$

where α is a multi-index and $l \leqslant k$. The equation $F(x, \partial^k u) = 0$ is *linear* if the coefficients a_α—which may be complex matrices—and f do not depend on u or its derivatives, while it is *quasilinear* if a_α and f depend only on the derivatives $\partial^\beta u$ of at most order $k - 1$, so $l \leqslant k - 1$. In this case, F is linear in its highest order derivatives. The usual formulas for the various curvatures of a Riemannian manifold are examples of second order quasilinear differential operators in the metric g. Other examples are the minimal surface equation—as well as all Euler-Lagrange equations for variational problems. On the other hand, the Gaussian curvature, K, of a surface $u\colon \mathbb{R}^2 \to \mathbb{R}$ in $\mathbb{R}^3$ (in the notation $u_x = \partial u/\partial x$ etc) satisfies

$$u_{xx}u_{yy} - u_{xy}^2 - K(x, y)(1 + u_x^2 + u_y^2)^2 = 0, \tag{12}$$

which is not quasilinear, it is fully nonlinear. Another fully nonlinear equation is the Monge-Ampère equation for Kähler-Einstein metrics [it is customary to refer to nonlinear equations involving the determinant of the hessian as "Monge-Ampère equations"].

One often meets second order quasilinear equations in *divergence form*

$$\operatorname{div} A(x, u, Du) = f(x, u, Du),$$

where A is a vector field. For example, on $\mathbb{R}^n$ this equation is

$$-\sum_i \frac{\partial}{\partial x^i} A^i(x, u, Du) = f(x, u, Du).$$

Clearly, if $u \in C^2$ is a solution of this in an open set $\Omega \subset \mathbb{R}^n$, then for any $\varphi \in L_1^2$ with support in Ω we have

$$\int \sum_i A^i(x, u, Du) \frac{\partial \varphi}{\partial x^i} \mu_g = \int f(x, u, Du) \varphi \mu_g. \tag{13}$$

In many circumstances, this last formula makes sense even for $u \in L_1^2(\Omega)$, rather than $C^2(\Omega)$ (the model is where $A(x, u, Du) = \operatorname{grad} u$). Then we say that $u \in L_1^2(\Omega)$ is a *weak solution* of the original equation if (13) holds for all $\varphi \in L_1^2$ with support in Ω. Caution: there are several other useful, but not equivalent, definitions of "weak solution".

The main application of these notions is to calculus of variations problems, where one seeks a critical point of

$$J(u) = \int F(x, u, Du) \mu_g.$$

Then the Euler-Lagrange equation is automatically in divergence form, so one may think of (13) as a weak form of an Euler-Lagrange equation.

The virtue of enlarging the class of admissible solutions to allow weak solutions is that this may make it much easier to prove the existence of a solution of the equation. On the other hand, one is then faced with the often difficult regularity problem of determining to what extent this weak solution is actually smooth.

E. Adjoint

If E and F are smooth hermitian vector bundles over M and if $P: C^\infty(E) \to C^\infty(F)$ is a *linear* differential operator, then one can use the L^2 inner product to define the *formal adjoint*, P^*, by the usual rule

$$\langle Pu, v \rangle_F = \langle u, P^* v \rangle_E$$

for all smooth sections $u \in C^\infty(E)$, $v \in C^\infty(F)$. Since the supports of u and v can be assumed to be in a coordinate patch, one can compute P^* locally using integration by parts. For example if $u: \mathbb{R} \to \mathbb{R}^N$ and if $Pu = Du$ (where here $D = d/dx$), then $P^* v = -Dv$, while if

$$Pu = A(x) D^2 u, \quad \text{then} \quad P^* v = D^2(A^*(x) v),$$

where A^* is the adjoint of the matrix A. We use the term "formal adjoint" to avoid confusion with the strict Hilbert space adjoint of an unbounded operator.

F. Principal Symbol

To a linear differential operator $P: C^\infty(E) \to C^\infty(F)$ of order k, at every point $x \in M$ and for every $\xi \in T_x^* M$ one can associate an algebraic object, the *principal symbol* $\sigma_\xi(P; x)$, often written simply $\sigma_\xi(P)$. If, in local coordinates, With the induced natural Riemannian metric the Gauss curvature is the product of the principal curvatures, or the inverse of the product of the principal radii (radii) of curvature. ...

Pag. 36(47) 1.62 Theorem

Given a Riemannian manifold (M, g), the function d is a distance on M, and the topology of the metric space (M, d) is the same as the manifold topology of M.

→ A corollary of the Gauss Lemma 1.46 is that the distance is realized (at least locally) by geodesics. More precisely.

Pag. 103(114)

In other words, we have computed the curvature of 2-dimensional totally geodesic subspaces in Gaussian coordinates.

We can obtain the rest of the relevant sectional curvatures by observing that on each

$\rho =$ constant hypersurface

(%B/

$\mathbb{R} \times \{p\} \times S^2$ the induced metric is the product metric

$$g_p = f^2(p)\,dt^2 - G^2(p)\,d\sigma^2$$

. . .

Pag. 126(137) – 4.41 Remarks

If dim $M = 2$

a function f on M is the gaussian curvature of some metric if and only if f is positive somewhere (resp. is identically zero or takes both signs, resp. is negative somewhere).

F. The Manifold of Metric with Constant Scalar Curvature:

Pag. 439(450) – 16.18

Thus, each of the following conditions implies that the Riemannian manifold (M, g) has pure curvature operator:

(i) dim $M = 2$, or $W = 0$ (e.g., dim $M = 3$, cf. 1.119);

(ii) (M, g) is a hypersurface in a space of constant curvature (cf. the Gauss equation (1.72c));

(%

(iii) (M, g) is a Riemannian
product of manifolds with
pure curvature operator.

Moreover, it is obvious that

(iv) the property of having pure
Weyl tensor is conformally
invariant.

Pág. 459(470)

On the other hand, the Gaussian
curvature, K, of a surface
$u: \mathbb{R}^2 \Rightarrow \mathbb{R}$ in $\mathbb{R}^3$
(in the notation $u_x = \partial u / \partial x$ etc)
satisfies (12)

$$u_{xx} u_{yy} - u^2_{yy} - K(x,y)(1 + u_x^2 + u_y^2) = 0,$$

which is not quasilinear, it is
fully nonlinear.

=20211016/Einstein/

Quantum Relativity – A Synthesis
of the Ideas of Einstein and
Heisenberg

Pág. 241(256) – $\Delta P \gg \frac{\hbar}{a}$, $\Delta Q \gg a$,
the coherent states are practically orthogonal.

-256- - -* ! - 6 ! 5 ! A 3 2 3
&" ! = ! " : D 1 !) : " D //B 1"
(' $B ! 4 !

where the functions $P(t)$ and $Q(t)$ are just the time-dependent coordinates of an classical free particle with initial position and momentum

$$Q(0) = Q, \quad P(0) = P\,. \tag{5}$$

Coherent states are not orthogonal, and so there is no variable having the coherent states $|Q, P\rangle$ as its eigenvectors and the parameters P (or Q) as corresponding eigenvalues. The parameters P and Q are not eigenvalues of any quantum variables. They cannot be identified with the results of exact measurements.

On the other hand, coherent states can be nearly orthogonal. The inner product between distinct coherent states $|P + \Delta P, Q + \Delta Q\rangle$ and $|Q, P\rangle$ is called the *form factor* for these coherent states and is given by

$$\langle P + \Delta P, Q + \Delta Q|P, Q\rangle = \frac{1}{\sqrt{2\pi}} e^{-(\Delta P^2 + \Delta Q^2)/2} = f(\Delta P, \Delta Q) \tag{6}$$

in units with $a = 1$. This falls off rapidly with the differences $\Delta P, \Delta Q$, so that for macroscopic position and momentum differences,

$$\Delta P \gg \frac{\hbar}{a}, \quad \Delta Q \gg a\,, \tag{7}$$

the coherent states are practically orthogonal.

The wave functions of coherent states $|Q, P\rangle$ in the q frame of basis vectors $|q\rangle$ are harmonically modulated Gaussian wave packets

$$\langle q|P, Q\rangle \sim \exp\left[-\tfrac{1}{2}(q - Q)^2 + iPq\right] \tag{8}$$

P is the modulating wavenumber and Q is the center. In the p frame the wave function has center P and modulating wavenumber $-Q$.

The main point is that the P, Q parameters of the coherent state change according to the classical equations of motion when the mode develops according to Schrödinger's equation. For the harmonic oscilator this is exact. In the general case, one settles for equality up to higher order terms in $\hbar$, which are regarded as negligible for high quantum numbers and large action.

The term "coherent state" is apropriate even though quantum systems do not have states, because the quantum modes called coherent provide a quantum origin for classical statelike behavior. To be sure, an initial vector for one quantum is not a state in a classical sense; but in practice we apply coherent states to such large populations of quanta that we can determine a one-quantum ket from a small part which we then discard, hardly changing the whole.

We now examine how coherent states aid us to understand the classical limit.

path tensor by taking the diagonal elements in one frame, specifically a spacetime frame.

A determination of the quantum path is complementary to a determination of the Hamiltonian of the quantum, and so to its mass and other couplings. We may ask the path that a quantum takes, but then we cannot know what kind of quantum takes that path.

12.5.3 Hamiltonian and Lagrangian Theories

Quantum dynamics was historically discovered in canonical coordinates and expressed as Hamiltonian equations of motion. The Hamiltonian quantum theory looks for the propagator G. It expresses the propagator for infinitesimal time interval δt in units with $\hbar = 1$ in the form

$$G = e^{-iH\delta t} \tag{1}$$

with Hamiltonian $H = H(q, p)$. The propagator for finite intervals x'', x' is found by decomposing the interval into N subintervals, using (1) for each, multiplying in chronological order, and taking the limit $N \to \infty$. This finite propagator is also called the time-ordered product integral of the infinitesimal propagator (1).

This is not a relativistic theory. It inherits this flaw from classical Hamiltonian mechanics. To be sure, a Lorentz transformation is a canonical transformation, so that the form of the classical equations of motion was relativistically invariant. But the main concepts entering into the Hamiltonian equations, the Hamiltonian H and the time derivative d/dt, are not relativistic objects, in that they do not have LORENTZ transformation laws.

Since the differential action $dS = L\,dt$ is a LORENTZ scalar and the Hamiltonian is not, it was natural to found a relativistic quantum theory on an action principle, expressed like the classical action principle in terms of relativistic entities, instead of a Hamiltonian. Dirac posed and partly solved the problem of formulating a quantum action principle. He regarded the Lagrangian $L(x, \dot{x})$ as a matrix element $L = (x'|x)$ relating two positions x, x' at times t, $t' = t + dt$ separated by a differential time dt, and showed, schematically speaking, that the propagator may be expressed equivalently as an exponential in the action and as the matrix element of an exponential in the Hamiltonian. More fully:

Suppose that L and H are of the form

$$L = \tfrac{1}{2}m\dot{x}^2 - V(x) =: K - V\,, \tag{2}$$

$$p = \frac{\partial L}{\partial \dot{x}}\,,$$

$$H(p, x) = p\dot{x} - L = \frac{1}{2m}p^2 + V(x) = K + V\,.$$

Then for infinitesimal $\delta t = t'' - t'$, the propagator (1) may be expressed in terms of the Lagrangian by

$$\langle x''|G|x'\rangle := [x'']e^{-iH\delta t}[x'] = \text{const}\, e^{iL(x,\dot{x})\delta t} + \mathrm{O}(\delta t^{3/2})\,. \tag{3}$$

Here const is a normalization constant whose magnitude is fixed so that (3) defines a unitary G; and the arguments of the Lagrangian function L in (3) are defined by

$$x = \frac{x' + x''}{2} \tag{4}$$

$$\dot{x} = \frac{x'' - x'}{\delta t}\,. \tag{5}$$

To prove this, one completes the square. Consider first a free particle ($V \equiv 0$). Without loss of generality set $m = 1$. The free propagator G_0 is given by 12.5 (1) with $H = K$. We express its matrix element as a Fourier integral by inserting the spectral projection $[p]$ for momentum, using the transformation function $\langle p|x\rangle$ from position eigenvectors $|x\rangle$ to momentum ones $|p\rangle$,

$$\langle p|x\rangle = \frac{1}{\sqrt{2\pi}}e^{ipx} \tag{6}$$

and completing the square:

$$\begin{aligned}\langle x''|G_0|x'\rangle &= \langle x''|e^{-iK\delta t}|x'\rangle \\ &= \langle x''|e^{-ip^2\delta t/2}|x'\rangle \\ &= \int dp[x''|e^{-ip^2\delta t/2}|p|p|x'] \\ &= \frac{1}{2}\int dp \exp\Big[-\frac{i}{2}p^2\delta t - ip\Delta x\,\delta t\Big]\end{aligned} \tag{7}$$

where $\Delta x := x'' - x'$. To complete the square we introduce a translated momentum variable $P = p + a$ and determine a so that

$$p^2 + 2p\Delta x + a^2 = P^2\,. \tag{8}$$

It follows that

$$a = \Delta x\,, \tag{9}$$

$$\langle x''|G_0|x'\rangle = \frac{1}{2\pi}\Big[\int dP \exp\Big[-\frac{i}{2}P^2\delta t\Big]\Big] \exp\Big[\frac{i}{2}\Delta x^2\delta t\Big]\,.$$

This integral is the standard Gaussian, integrated along a line in the complex plane rotated from the real axis by $\pi/4$:

$$\int dP\, e^{-iP^2\delta t/2} = \sqrt{\frac{2p}{i\delta t}}\,. \tag{10}$$

Therefore

$$\langle x''|G_0|x'\rangle = \frac{1}{\sqrt{2\pi i\delta t}} \exp\Big[\frac{i}{2}\Delta x^2\delta t\Big] = \frac{1}{\sqrt{2\pi i\delta t}}e^{iL_0\delta t}\,. \tag{11}$$

where $L_0 = K$ is the free Lagrangian. This establishes the desired result for the free particle and gives the complex normalization constant const $= (2\pi i\delta t)^{-1/2}$.

Next we turn on the potential. The infinitesimal propagator from t' to $t'' = t'+\delta t$ may be approximately factored as

$$G = e^{-iH\delta t} = e^{-i(K+V)\delta t} = e^{-iK\delta t}e^{-iV\delta t} + \mathrm{O}(\delta t^2) = G_0 e^{-iV\delta t} + \mathrm{O}(\delta t^2)\,. \tag{12}$$

The kinetic and potential energy now act alternately in succession instead of simultaneously. The error arises from neglecting the commutator $[K\delta t, V\delta t] = \mathrm{O}(\delta t^2)$. Let t be the interpolated time dividing the interval from t' to t'' into two equal halves, and let x be the variable position at time t. We propagate with potential energy for the first half of the time interval and with kinetic for the second. The matrix element of the potential propagator is

$$\langle x|e^{-iV\delta t}|x'\rangle = \delta(x - x')e^{-iV(x)\delta t}\,. \tag{13}$$

The matrix element of G is therefore

$$\begin{aligned} &\langle x''|G|x'\rangle \\ &= \frac{1}{\sqrt{2\pi i\delta t}}\int dx \exp\left[\frac{i}{2}\Delta x^2\delta t\right]\delta(x - x')\exp[-iV(x')\delta t] + \mathrm{O}(\delta t^{3/2}) \\ &= \frac{1}{\sqrt{2\pi i dt}}e^{iL\delta t} + \mathrm{O}(\delta t^{3/2})\,, \\ &L := \langle x''|L|x'\rangle = L(x, \dot{x})\,. \end{aligned} \tag{14}$$

This proves the infinitesimal result. It then follows for finite intervals by subdivision and composition. The number of intervals is $\mathrm{O}(\delta t^{-1})$, and so the accumulated error is $\mathrm{O}(\delta t^{1/2})$ and vanishes in the limit $\delta t \to 0$. Evidently the normalization constant const diverges as $\delta t \to 0$.

Note the sign difference between (1) and (2).

The propagator for finite intervals x'', x' is found by decomposing the interval into N subinterval with $N + 1$ points $x' = x_0, x_1, x_2, \cdots, x_N = x''$, using (2) for each, multiplying in chronological order, integrating over the intermediate variables $x_1, \cdots, x_{N-1}$, and taking the limit $N \to \infty$.

This shows the equivalence of the two forms of theory.

12.5.4 Schwinger Variational Principle

Neither of these equivalent forms of 12.5.3 (3) is quite satisfactory for a quantum dynamical principle. The Hamiltonian form 12.5.3 (1) is a quantum tensor but not a spacetime one. Since the exponential of the tensor $iL\,dt$ in 12.5.3 (3) is not a tensor, the Lagrangian form is a spacetime tensor (namely a scalar) but not a quantum one. This problem was resolved by Richard P. Feynman and Julian Schwinger by integral and differential methods respectively. We have already seen an example of Feynman's path integral theory of quantum dynamics. Now we illustrate the Schwinger variational theory.

All three of these modifications require us to give new meaning to the path integral that appears in our action principle. If there are basic odd quanta we must learn to integrate over first-grade Grassmann variables. If there are gauge potentials we must do integrals over non-dynamical degrees of freedom as well as dynamical. Since there are interactions we must integrate functions other than Gaussian exponentials, leading to divergences.

13.2 Gauge Physics

An adequate treatment of gauge physics takes a sizable volume. Here we merely summarize some of the features of gauge physics to which we will refer in later chapters.

13.2.1 Gauge History

The word "gauge" comes from the machinist's precision standard of length, the gauge or gauge block. After Einstein showed that gravity could be understood as a disturbance in the remote comparison of direction, Hermann Weyl (1922) proposed that electromagnetism was a disturbance in the remote comparison of length. He suggested a spacetime physics in which each experimenter not only provided a spacetime chart x^μ but also a field of "gauges" $g(x)$, one gauge (block) at each spacetime point, with which to measure local lengths. He postulated that physical laws and concepts must not only be invariant under chart transformations, as Einstein assumed, but also under arbitrary gauge transformations, $\Lambda(x) : g(x) \to g'(x) = \Lambda(x)g(x)$, by any non-zero real function Λ: $M^4 \to \Gamma$. Here Γ is the original gauge group, the multiplicative group $\mathbb{R}^+$ of positive real numbers. In gauge geometry a local physical quantity $\phi = \phi(x)$ not only defines a representation of the local group GL_4 by its tensor index structure but also defines a representation of the local gauge group Γ by its gauge weight n, the power of Λ in its gauge transformation law $\phi \to \Lambda^n\phi$ when $g \to \Lambda g$.

Under such a gauge transformation, the Lie derivative $\partial_\mu\phi$ and Einstein's covariant derivative $D_\mu\phi$ of any physical quantity ϕ both change by an additive term $\sim \partial_\mu\Lambda(x) \cdot \phi(x)$, which brings in the independent object ϕ. Therefore they do not have a gauge transformation by themselves, and are not gauge geometrical objects. Following the analogy of the covariant derivative of general relativity, Weyl postulated a new gauge-covariant derivative, necessarily of the form

$$D_\mu := \partial_\mu + \Gamma_\mu + A_\mu \tag{1}$$

where Γ_μ is the usual gravitational connection and A_μ is a new field, the gauge connection, with a change that compensates that of ∂_μ. The gauge field is the commutator $F_{\mu\nu} := [D_\mu, D_\nu]$. The action of the theory is the sum of the usual action for the sources, with the substitution $\partial \to D$, and the gauge field action $\alpha F_{\mu\nu}F^{\mu\nu}/4$. The analogue of the Bianchi identity for A_μ is one of Maxwell's equations, and the other Maxwell equation looks like a mere definition of the current. Therefore Weyl proposed to identify A_μ with the electromagnetic vector potential up to a factor.

To accord with the usual electromagnetic theory, Weyl identified α with the fine structure constant $\sim 1/137.04$. The quantum of electrical charge e now appears only in front of the electromagnetic action, as a kind of stiffness of the electromagnetic field, and nowhere else in physics, because Weyl absorbed it into the connection.

Perhaps we should do this today with $e/3$; the corresponding fine structure constant is then $\alpha = 9/137.04 = 1/15.23$.

Instead of accounting for the coupling constant, which is now formally 1, we must understand the stiffness α of the field. The electromagnetic coupling between unit charges is relatively weak because the field is stiff, as with masses supported against gravity on a taut elastic sheet. The fundamental variable of gauge physics is the connection A_μ, or better the gauge (covariant) derivation $D_\mu(x) = \partial_\mu + A_\mu$, a dual vector.

Weyl's theory was unviable because it implied that an atom of the solar atmosphere convecting around the sun (say) should grow in size by a fixed factor on each revolution, due to the sun's magnetic field, and therefore that we should see a mix of many atomic sizes and spectral wavelengths. We do not.

Dirac salvaged Weyl's theory by redefining its gauge group. Instead of stretching literal gauges, Dirac's gauge transformation shifts the phase of the wavefunction of all charged particles by $\epsilon^{i\theta}$. Dirac's gauge group is U_1, not $\mathbb{R}_+$, and acts on different variables. Nevertheless Dirac's gauge (Lie) algebra is the same as Weyl's: the trivial algebra on the vector space $\mathbb{R}$, with $[\alpha, \beta] \equiv 0$. Dirac's theory inherited the main formal properties of Weyl and agreed with Maxwell's theory and experiment, but no longer explained what electrical charge was or how charged particles differ from neutral ones in spacetime terms.

As a result of this success Weyl's theory was generalized still further. A gauge theory today has a local gauge variable γ in the episystem, generalizing Weyl's gauge block g, subject to a local gauge group Γ, not necessarily abelian, generalizing the gauge group $\mathbb{R}_+$. For convenience we coordinate the variable γ with a variable group element $\gamma \in \Gamma$. The theory has global gauge transformations, each described locally by a field $\gamma(x)$: $M^4 \to \Gamma$ of local gauge transformations, multiplied point by point. It has a gauge derivative $D(x)$, which changes locally under an infinitesimal gauge transformation $\Lambda = (\Lambda(x)) = (1+\lambda(x))$, where $\lambda \to 0$, according to

$$\Lambda : D_\mu \to D^\Lambda_\mu(x) = \Lambda(x)D(x)\Lambda(x)^{-1} = \partial_\mu + A^\Lambda_\mu\,. \tag{2}$$

Its gauge connection A_μ represents the infinitesimal change that the gauge variable γ undergoes during a congruent transport along the segment dx^μ and so takes on values in the Lie algebra of Γ:

$$d\gamma = A_\mu dx^\mu \tag{3}$$

A_μ transforms under (2) according to

$$A_\mu \to A_\mu^\Lambda = A_\mu + [\lambda, A_\mu] - \partial_\mu \lambda \,. \tag{4}$$

The gauge flux through a loop $x^\mu(\tau)$ is the gauge group element $\delta\gamma$ that acts on a gauge when it is carried congruently around the loop. In Weyl's theory, this is a magnetic flux for an infinitesimal loop $\delta_1 x \vee \delta_2 x$ if both δx_1 and δx_2 are along space axes, and an electric flux if one δx is along the time axis. In general, the transport around a loop $\delta_1 x \vee \delta_2 x$ results in a net second-order gauge transformation

$$\begin{aligned} \delta^{(2)}\gamma &= F_{\mu\nu}\delta_1 x^\mu \delta_2 x^\nu \,, \\ F_{\mu\nu} &:= [D_\mu, D_\nu] \end{aligned} \tag{5}$$

and the commutator F defining the loop transport is called the gauge field.

Gravity may be regarded as an exceptional kind of gauge theory where the gauge group acts on spacetime vectors dx^μ themselves. In all other gauge theories today, spacetime vectors are assumed to be invariant under the gauge group. For gravity, the gauge conection is the Christoffel connection $\Gamma_\kappa = (\Gamma^\mu_{\lambda\kappa})$ and the gauge field is the curvature tensor $F_{\lambda\kappa} = (R^\nu_{\mu\lambda\kappa})$. In this theory we may use the metric tensor $g_{\mu\nu}(x)$ to form the Hilbert action

$$L = \kappa R + \lambda \,, \tag{6}$$

with only two terms; L also has a common factor of the relative scalar $\sqrt{\det(-g_{\mu\nu})}$, which for simplicity we have absorbed into the volume element d^4x. The first term κR in this Lagrangian is linear in the gauge field and its coefficient κ is equivalent to Newton's gravitational constant G; the second term λ is the cosmological constant.

Gauge theory inherits one great virtue from Einstein's theory of gravity: Its action principle $L(x)$ is narrowly constrained by two natural postulates, gauge invariance and maximal locality. The gauge invariance postulate is inherent in the gauge theory itself. By maximal locality we mean that the action $L(x)$ formed from $A_\mu(x)$ has the lowest possible differential order. Since higher order terms vary as higher powers of energy-momentum, this postulate may be justified physically as a low-energy approximation to a more general action. For gravity it gives us the Hilbert action, and for electromagnetism the Maxwell action. Roughly speaking, one need merely specify a gauge group and a full-fledged dynamical theory follows.

Now to see the effect of the sun's magnetic field on an electron (say) we would have to do an interference experiment, splitting and recombining an electron beam and noticing the shift in the interference fringes as we slowly move the sun in and out of the beam loop. As Aharanov and Bohm emphasized, one should see a phase shift even if there is no magnetic field at the electron beam, provided only that the beam loop encloses flux. The Aharanov-Bohm experiment has been done (with magnetized iron whiskers instead of the sun) and is consistent with gauge theory.

The moral is that to describe locally the effect of electromagnetism (or other gauge field) on a quantum we must know the potential A_μ, not just the field $F_{\mu\nu}$.

Nevertheless all we can measure locally is a flux or a field $F_{\mu\nu}$, never a potential A_μ.

For ordinary gauge groups, which leave spacetime vectors fixed, the maximally local action $S = \int L\, dx$ that resembles the Maxwell action in being quadratic in the gauge field has the density

$$L = -\tfrac{1}{4}\alpha F^{\mu\nu} F_{\mu\nu}\,. \tag{7}$$

The quanta of the gauge connection are called gauge vector bosons. According to the Lagrangian (7) the physical quanta have spin 1 and (rest) mass 0. They are accompanied by unphysical longitudinal and timelike fields of spin 0, transmitting static forces that fall off as the inverse square of distance, with no length parameter to set their scale, and are therefore said to have infinite range.

In the early decades of gauge theory it was accepted that gauge forces inevitably have infinite range and massless quanta, like the prototype gauge fields, gravity and electromagnetism. Certainly a mass term like $\mu^2 A^2$ spoiled gauge invariance at the same time that it gave the force a range μ^{-1}. This made it hard to understand the short-range weak and strong forces as gauge forces.

E.C.G. Stückelberg then showed that a gauge vector boson could acquire mass and short range by interacting with a scalar field $f = f(x)$ that transformed additively under gauge transformations, according to

$$\Lambda : f \mapsto f + \ln \Lambda\,. \tag{8}$$

The standard gauge covariant Lagrangian for the field f,

$$\begin{aligned} L_f &= -\tfrac{1}{2} D_\mu f D^\mu f\,, \\ D_\mu f &:= \partial_\mu f + \partial_\mu \ln \Lambda \cdot f\,, \end{aligned} \tag{9}$$

then provides a mass for the gauge field.

Instead of the Stückelberg field f one could equivalently introduce its exponential, the field $\eta = e^f = \eta(x)$, which transforms multiplicatively with gauge weight 1, but then η must have a non-zero vacuum value $\eta = e^f = e^0 = 1$, corresponding to $f = 0$.

Quaternion quantum mechanics required just such a field, also called η, on fundamental grounds [Finkelstein, Jauch, Schiminovich and Speiser (1961a, b)]. The quaternionic field η replaces the constant complex factor $i\hbar$ that is used in complex quantization, just as Einstein's variable spacetime metric $g_{\mu\nu}$ replaces the constant Minkowski metric of special relativity. The η field is a pure imaginary quaternion with vacuum value subject to the constraint $|\eta| = \hbar = 1$ and defines the electromagnetic axis in the three-dimensional space of imaginary quaternions. It gives mass to a charged vector gauge boson, forming an SU_2 triplet with a massless photon-like boson . This seems to be the only field theory that provides charged gauge bosons with mass and is not ad hoc, though it leaves other problems of electroweak physics unresolved.

Higgs (1964) and others proposed to account for massive gauge vector bosons by couplings to such fields, called Higgs fields, which are essentially exponentials of Stückelberg fields. Instead of the nonlinear constraint $|\eta| = 1$, which is computationally intractable, Higgs fields have non-zero vaccuum expectation values because their non-quadratic action has a minimum at $\eta \neq 0$. Then t'Hooft (1971) showed that gauge theories of that kind retained the renormalizability of the massless gauge theories. This opened a new era of gauge theory.

13.2.2 Standard Model

We now sketch the standard model for the non-gravitational part of physics today, which has strong experimental support for energies up to 100 GeV and cries out for explanation. It sums up work of Sheldon Glashow, Abdul Salam, Steven Weinberg and others.

The standard model rests on the standard gauge group

$$G_{\text{standard}} = U_1 \otimes SU_2 \otimes SU_3 = \text{hypercharge}\otimes\text{isospin}\otimes\text{color}\,. \tag{1}$$

The U_1, SU_2 and SU_3 groups have dimensions 1, 3 and 8 and rank (maximal number of commuting generators) 1, 1 and 2. When used to represent particle symmetries, their respective charges or infinitesimal generators are called hypercharge Y, (weak) isospin T^i $(i = 1, 2, 3)$ and color C^c $(\gamma = 1, 2, \cdots, 8)$.

These are the matrix generators of the identity representations of these unitary groups. These are therefore one 1×1 fixed complex hypercharge matrix Y; three 2×2 isospin charges T^i; and eight 3×3 color charges C^c, respectively.

There is a gauge vector field for each of these $1+3+8$ generators. A 2-vector $|\tau\rangle$ supporting isospin is called an isospinor or SU_2 spinor and its two components are labeled *up* and *down*. A 3-vector $|\gamma\rangle$ supporting color SU_3 is called a color or SU_3 spinor and the three components are labeled *blue*, *green*, and *red*.

For orientation we mention that electric charge is $Q = Y + T_3$, where 3 labels a special direction in the three-dimensional isospin space called the electromagnetic axis. This direction is singled out by the Higgs field.

The fundamental fields of the standard model fall into two broad species by statistics, odd and even.

The odd fields are all left-handed spin 1/2 fields and their right-handed antiparticles. These divide again, but now in two independent ways:

- by symmetries into leptons, which are colorless and so have color operator

  $$C^{\gamma}_{\text{lepton}} = 0$$

  and quarks, which have the 3×3 color operators of the identity representation

  $$C^{\gamma}_{\text{quark}} = C^c\,;$$

- and by masses into three families also called flavors, labelled with a flavor index $F = 1, 2, 3$.

D -256- - -* ! - 9 ! 3 <! !L J 4 !
! " = => ! 2 ") " 211 ! 1 !) 1"
(/ ' % 4 ! 7 3 : ") "

4.3 Differential Geometry in the Hodge Bundle

4.3.1 Riemannian and Lorentzian Structures on M

Next we introduce on M a smooth metric field $\mathbf{g} \in \sec T_2^0 M$ and gives the

Definition 229 *A pair* $(M, \mathbf{g})$*,* $\dim M = n$ *is a* n*-dimensional Riemann structure (or Riemann manifold) if* $\mathbf{g} \in \sec T_2^0 M$ *is a smooth* metric *of signature* $(n, 0)$*. If* $\mathbf{g}$ *has signature* (p, q) *with* $p + q = n$*,* $p \neq n$ *or* $q \neq n$ *then the pair* $(M, \mathbf{g})$ *is called a pseudo Riemannian manifold. When* $\mathbf{g}$ *has signature* $(1, n-1)$ *the pair* $(M, \mathbf{g})$ *is called an hyperbolic manifold. When* $\dim M = 4$ *and* $\mathbf{g}$ *has signature* $(1, 3)$ *the pair* $(M, \mathbf{g})$ *is called a Lorentzian manifold.*[8]

We already defined the concept of oriented manifold. Thus, we say that a Riemannian (or pseudo Riemannian or Lorentzian) manifold is orientable if and only if it admits a continuous metrical volume element field $\tau_{\mathbf{g}} \in \sec \bigwedge^n T^*M$ given in local coordinate functions $\{\mathbf{x}^i\}$ covering $U \subset M$ by

$$\tau_{\mathbf{g}} = \sqrt{|\det \mathbf{g}|} dx^1 \wedge ... \wedge dx^n \ , \tag{4.91}$$

where

$$\det \mathbf{g} = \det \left[\mathbf{g}\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right) \right] \ . \tag{4.92}$$

Proposition 230 *Any* C^r *manifold* M*,* $\dim M = n$ *admits a* C^{r-1} *Riemannian metric* $\mathbf{g}$ *(signature* $(n, 0)$*) if, and only if, it is paracompact.*

Proof. For a proof see, e.g. [3]. ☐

Let us consider now a smooth oriented metric manifold $\mathsf{M} = (M, \mathbf{g}, \tau_{\mathbf{g}})$, $\dim M = n$, where $\mathbf{g}$ is a smooth metric field of signature (p, q) and $\tau_{\mathbf{g}} \in \sec \bigwedge^n T^*M$. We denote by $g \in \sec T_0^2 M$ the *metric* tensor of the cotangent bundle. Also we denote the scalar product induced on $\bigwedge T^*M$ by the metric tensor $g \in \sec T_0^2 M$ by[9] $\underset{\mathbf{g}}{\cdot} : \sec \bigwedge T^*M \times \sec \bigwedge T^*M \to \sec \bigwedge^0 T^*M$. If $A, B \in \sec \bigwedge^p T^*M$ we have (recall (2.121))

$$(A \underset{\mathbf{g}}{\cdot} B)\tau_{\mathbf{g}} = A \wedge \underset{\mathbf{g}}{\star} B. \tag{4.93}$$

---

[8] When Lorentzian manifolds serve as models of spacetimes it is also imposed that M is noncompact. See Sect. 4.4.1.

[9] When there is no chance of confusion we eventually used the symbol $\cdot$ instead of the symbol $\underset{\mathbf{g}}{\cdot}$ in order to simplify the notation.

4.3.2 Hodge Bundle

Definition 231 *The* Hodge bundle *of the structure* M*, is the triple*

$$\bigwedge(\mathsf{M}) = (\bigwedge T^*M, \underset{\mathbf{g}}{\cdot}, \tau_{\mathbf{g}}) \,. \tag{4.94}$$

The importance of the Hodge bundle is that besides the exterior derivative operator, we can now introduce a new differential operator called the Hodge codifferential. Equipped with these two operators we can write, e.g. Maxwell equations (with currents) in a *diffeomorphism* invariant way[10] (see Sect. 4.7.1). This is a very important fact, which is often not well known as it should be.

Definition 232 *The* Hodge codifferential *operator in the Hodge bundle of* $\bigwedge(\mathsf{M})$ *is the mapping* $\underset{\mathbf{g}}{\delta} : \sec \bigwedge T^*M \to \sec \bigwedge T^*M$*, given, for homogeneous multiforms, by:*

$$\underset{\mathbf{g}}{\delta} = (-1)^r \underset{\mathbf{g}}{\star}^{-1} d \underset{\mathbf{g}}{\star} \,, \tag{4.95}$$

where $\underset{\mathbf{g}}{\star}$ *is the Hodge star operator associated with the scalar product* $\underset{\mathbf{g}}{\cdot}$ *.*

Definition 233 *The* Hodge Laplacian *operator is the mapping*

$$\underset{\mathbf{g}}{\lozenge} : \sec \bigwedge T^*M \to \sec \bigwedge T^*M,$$

given by:

$$\underset{\mathbf{g}}{\lozenge} = -(d\underset{\mathbf{g}}{\delta} + \underset{\mathbf{g}}{\delta} d) \,. \tag{4.96}$$

The exterior derivative, the Hodge codifferential and the Hodge Laplacian satisfy the relations:

$$\begin{aligned}
& dd = \underset{\mathbf{g}}{\delta}\underset{\mathbf{g}}{\delta} = 0; \quad \underset{\mathbf{g}}{\lozenge} = (d - \underset{\mathbf{g}}{\delta})^2 \\
& d\underset{\mathbf{g}}{\lozenge} = \underset{\mathbf{g}}{\lozenge} d; \quad \underset{\mathbf{g}}{\delta}\underset{\mathbf{g}}{\lozenge} = \underset{\mathbf{g}}{\lozenge}\underset{\mathbf{g}}{\delta} \\
& \underset{\mathbf{g}}{\delta}\underset{\mathbf{g}}{\star} = (-1)^{r+1}\underset{\mathbf{g}}{\star} d; \quad \underset{\mathbf{g}}{\star}\underset{\mathbf{g}}{\delta} = (-1)^r d\underset{\mathbf{g}}{\star} \\
& d\underset{\mathbf{g}}{\delta}\underset{\mathbf{g}}{\star} = \underset{\mathbf{g}}{\star}\underset{\mathbf{g}}{\delta} d; \quad \underset{\mathbf{g}}{\star} d \underset{\mathbf{g}}{\delta} = \underset{\mathbf{g}}{\delta} d \underset{\mathbf{g}}{\star}; \quad \underset{\mathbf{g}}{\star}\underset{\mathbf{g}}{\lozenge} = \underset{\mathbf{g}}{\lozenge}\underset{\mathbf{g}}{\star} \,.
\end{aligned} \tag{4.97}$$

Remark 234 *When it is clear from the context which metric field is involved we use the symbols* $\star$*,* δ *and* $\lozenge$ *in place of the symbols* $\underset{\mathbf{g}}{\star}$*,* $\underset{\mathbf{g}}{\delta}$ *and* $\underset{\mathbf{g}}{\lozenge}$ *in order to simplify the writing of equations.*

---

[10] For the exact meaning of the concept of diffeomorphism invariance of a spacetime physical theory (as used in this text) see Sect. 5.5.3.

4.4 Classification of Geometries on M and Spacetimes

Definition 251 *Given a triple* $(M, \mathbf{g}, \nabla)$:

(a) *it is called a Riemann*-Cartan space if and only if

$$\nabla \mathbf{g} = 0 \qquad \text{and} \qquad \Theta[\nabla] \neq 0 \,. \tag{4.131}$$

(b) *it is called* Weyl space if and only if

$$\nabla \mathbf{g} \neq 0 \qquad \text{and} \qquad \Theta[\nabla] = 0 \,. \tag{4.132}$$

(c) *it is called a* Riemann space *if and only if*

$$\nabla \mathbf{g} = 0 \qquad \text{and} \qquad \Theta[\nabla] = 0 \,, \tag{4.133}$$

and in that case the pair $(\nabla, \mathbf{g})$ *is called* Riemannian structure *for* M*.*

(d) *it is called* Riemann-Cartan-Weyl space *if and only if*

$$\nabla \mathbf{g} \neq 0 \qquad \text{and} \qquad \Theta[\nabla] \neq 0 \,. \tag{4.134}$$

(e) *it is called (Riemann) flat if and only if*

$$\nabla \mathbf{g} = 0 \qquad \text{and} \qquad \mathbf{R}[\nabla] = 0 \,,$$

(f) *it is called teleparallel if and only if*

$$\nabla \mathbf{g} = 0, \;\; \Theta[\nabla] \neq 0 \text{ and } \mathbf{R}[\nabla] = 0. \tag{4.135}$$

For each metric tensor defined on the manifold M there exists one and only one connection in the conditions of (4.133). It is called *Levi-Civita connection* of the metric considered, and is denoted by D. If in a given context it is necessary to distinguish between the Levi-Civita connections of two different metric tensors $\mathring{\mathbf{g}}$ and $\mathbf{g}$ on the same manifold, we write $\mathring{D}$, D.

Remark 252 *When* $\dim M = 4$ *and the metric* $\mathbf{g}$ *has signature* $(1,3)$ *we sometimes substitute Riemann by Lorentz in the previous definitions* (c),(e) *and* (f)*.*

4.4.1 Spacetimes

From nowhere besides the constraints already imposed (Hausdorff and paracompact) on M, we suppose also that it is connected and noncompact [10, 29]. We now introduce the concept of *time orientability* on an oriented Lorentzian manifold structures $(M, \mathbf{g}, \tau_{\mathbf{g}})$, which plays a key role in physical theories.

Definition 253 *Let* $(M, \mathbf{g})$ *be a Lorentzian manifold,* $TM = \bigcup_{\mathfrak{e} \in M} T_{\mathfrak{e}} M$ *its tangent bundle and* $\pi : TM \to M$ *the canonical projection (see* Appendix*). The causal character of* $(\mathfrak{e}, \mathbf{v}) \in TM$ *is the causal character of* $\mathbf{v}$ (Definition 59)*.*

Definition 254 *A line element at* $x \in M$ *is a one-dimensional subspace of* $T_x M$.

Proposition 255 *Let* M *be a* C^1 *paracompact and Hausdorff manifold, with* $\dim M = 4$. *Then the existence of a continuous line element field on* M *is equivalent to the existence of a Lorentzian structure on* M.

Proof. For a proof see [3]. $\square$

Proposition 256 *The set* $\mathfrak{T} \subset TM$ *of timelike points is an open manifold and it has either one (connected) component or two.*

Proof. A proof of this important result can be found in [29]. $\square$

Definition 257 *A connected Lorentzian manifold* $(M, \mathbf{g})$ *is said to be time orientable if and only if* $\mathfrak{T}$ *has two components and one of the components is labeled the future* $\mathfrak{T}^+$ *and the other component* $\mathfrak{T}^-$ *is labeled the past. We denote by* ↑ *the time orientability of a Lorentzian manifold.*

Definition 258 *A spacetime is a pentuple* $\mathfrak{M} = (M, \mathbf{g}, \nabla, \tau_{\mathbf{g}}, \uparrow)$ *where* $(M, \mathbf{g})$ *is a Lorentzian oriented and time oriented manifold and* ∇ *is an arbitrary covariant derivative operator on* M.

Definition 259 *When* $(M, \mathbf{g}, \nabla, \tau_{\mathbf{g}}, \uparrow)$ *is a spacetime and* $\nabla = D$ *is the Levi-Civita connection of* $\mathbf{g}$ *the spacetime is said to be Lorentzian. When* $\nabla \mathbf{g} = 0$ *and* $\Theta[\nabla] \neq 0$ *we call the structure* $(M, \mathbf{g}, \nabla, \tau_{\mathbf{g}}, \uparrow)$ *a Riemann-Cartan spacetime. The particular Riemann-Cartan spacetime for which* $\mathbf{R}(D) = 0$ *is called teleparallel (or Weitzenböck) spacetime.*

Definition 260 *A Lorentzian spacetime structure* $\mathcal{M} = (M, \boldsymbol{\eta}, D, \tau_{\boldsymbol{\eta}}, \uparrow)$ *is said to be Minkowski spacetime if and only if* $M \simeq \mathbb{R}^4$ *and* $\mathbf{R}(D) = 0$.

Exercise 261 *Show that we can give to* $\mathcal{M}$ *the structure of a metric* affine space, i.e. a structure $(M, \mathbf{V}, -, \eta)$ *where* $\eta : \mathbf{V} \times \mathbf{V} \to \mathbb{R}$ *is a Minkowski metric and such that there exists a mapping* $- : M \times M \to \mathbf{V}, (x - y) \in V$ *satisfying the triangle equality* $(x - y) + (y - z) = (x - z)$. *Also, if* $\boldsymbol{\eta} \in \sec T_2^0 M$ *is the metric of* $\mathcal{M}$ *and if* $\{e_\mu\}$ *is an orthonormal basis of* TM *then there is an orthonormal basis* $\{\mathbf{e}_\mu\}$ *of* $\mathbf{V}$ *such that* $\boldsymbol{\eta}_x(e_\mu, e_\nu) = \eta(\mathbf{e}_\mu, \mathbf{e}_\nu) = \eta_{\mu\nu}$, $\forall x \in M$.

Remark 262 *We just establish that any Lorentzian manifold admits a continuous element field. If it is also time orientable, we can choose a direction for the continuous element field, and say that it is a* timelike *vector field pointing to the future. This is a nontrivial result and very important for our discussion of the Principle of Relativity* (Chap. 5).

4.5 Differential Geometry in the Clifford Bundle

It is well known [19] that the natural operations on metric vector spaces, such as, e.g. direct sum, tensor product, exterior power, etc. carry over canonically to vector bundles with metric tensors. Then we give the following definition:

Definition 263 *The* Clifford bundle *of differential forms of the pair* $(M, \mathbf{g})$ *is:*

$$\mathcal{C}\ell(M,\mathbf{g}) = \frac{\mathcal{T}M}{J_{\mathbf{g}}} = \bigcup_{x \in M} \mathcal{C}\ell(T_x^*M, \mathbf{g}_x) \ , \tag{4.136}$$

where $\mathcal{T}M$ *denotes the (covariant) tensor bundle of* M, $J_{\mathbf{g}} \subset \mathcal{T}M$ *is the bilateral ideal of* $\mathcal{T}M$ *generated by the elements of the form* $\alpha \otimes \beta + \beta \otimes \alpha - 2\mathbf{g}(\alpha, \beta)$, *with* $\alpha, \beta \in \sec T^*M \subset \mathcal{T}M$ *and* $\mathcal{C}\ell(T_x^*M, \mathbf{g}_x)$ *is the Clifford algebra of the metric vector space structure* $(T_x^*M, \mathbf{g}_x)$.

The even subbundle of $\mathcal{C}\ell(M,\mathbf{g})$ is denoted by $\mathcal{C}\ell^{(0)}(M,\mathbf{g})$.

It can be shown that the Clifford bundle $\mathcal{C}\ell(M,\mathbf{g})$ (as defined by (4.136)) is a vector bundle associated with the principal bundle of orthonormal frames $\mathbf{P}_{\mathrm{SO}^e(p,q)}$, i.e.

$$\mathcal{C}\ell(M,\mathbf{g}) = \mathbf{P}_{\mathrm{SO}^e_{p,q}} \times_{Ad} \mathbb{R}_{p,q} \ . \tag{4.137}$$

In (4.137) Ad is the *adjoint representation* of $\mathrm{Spin}^e(p,q)$, i.e. $Ad : \mathrm{SO}^e_{p,q} \to \mathrm{Aut}(\mathbb{R}_{p,q})$, $u \mapsto Ad_u$, with $Ad_u A = u^{-1}$, $\forall u \in \mathrm{SO}^e_{p,q}$, $\forall A \in \mathbb{R}_{p,q} \simeq \mathcal{C}\ell(T_x^*M, \mathbf{g}_x)$. Details on these groups have been given in Chap. 3. In Chap. 6 we scrutinize the vector bundle structure of the Clifford bundle of differential forms over a general Riemann-Cartan manifold modeling spacetime, showing in particular (4.137).

4.5.1 Representation of Clifford Fields as Sums of Non-homogeneous Differential Forms

Definition 264 *Sections of* $\mathcal{C}\ell(M,\mathbf{g})$ *are called Clifford fields.*

We recall some notations and conventions. By $F(U)$ we denote the frame bundle (see Appendix A.3) of $U \subset M$. A section of $F(U)$ will be denoted by $\{\mathbf{e}_\alpha\} \in \sec F(U)$. The dual frame of a frame $\{\mathbf{e}_\alpha\}$ will be denoted by $\{\theta^\alpha\}$, where $\theta^\alpha \in \sec T^*U \subset T^*M$. When $\{\mathbf{e}_\alpha\}$ is a coordinate frame associated with the coordinate functions $\{\mathbf{x}^\mu\}$ of a local chart covering U we use instead of $\mathbf{e}_\alpha$ the notation $e_\alpha = \partial_\alpha$ and in this case $\theta^\alpha = dx^\alpha$. When $\{\mathbf{e}_\alpha\}$ refers to an orthonormal frame we use instead of $\mathbf{e}_\alpha$ the notation $\mathbf{e}_{\mathbf{a}}$ and instead of θ^α the notation $\theta^{\mathbf{a}}$.

Recall that as a vector space over $\mathbb{R}$, $\mathcal{C}\ell(T_x^*M, \mathtt{g}_x)$ is isomorphic to the exterior algebra $\bigwedge T_x^*M$ of the cotangent space and

$$\bigwedge T_x^*M = \bigoplus\nolimits_{k=0}^{n} \bigwedge{}^k T_x^*M \ , \tag{4.138}$$

where $\bigwedge^k T_x^*M$ is the $\binom{n}{k}$-dimensional space of k-forms. Then, there is a natural embedding $\bigwedge T^*M \hookrightarrow \mathcal{C}\ell(M, \mathtt{g})$ [16] and sections of $\mathcal{C}\ell(M, \mathtt{g})$ –Clifford fields (Definition 264)– can be represented as a sum of non homogeneous differential forms.[17] Let $\{\mathbf{e_a}\}$ be an orthonormal basis for $TU \subset TM$, i.e. $\mathbf{g}(\mathbf{e_a}, \mathbf{e_b}) = \eta_{\mathbf{ab}} = \mathrm{diag}(1, 1, ..., -1, ..., -1)$ and $(\mathbf{a}, \mathbf{b}, \mathbf{i}, \mathbf{j}, ... = 1, 2, ..., n)$. Moreover, let $\{\theta^{\mathbf{a}}\} \in \sec \bigwedge^1 T^*M \hookrightarrow \sec \mathcal{C}\ell(M, \mathtt{g})$ such that the set $\{\theta^{\mathbf{a}}\}$ is the dual basis of $\{\mathbf{e_i}\}$. We denote by $\{\theta_{\mathbf{i}}\}$ the *reciprocal basis* of $\{\theta^{\mathbf{i}}\}$, i.e. $\theta_{\mathbf{i}} \underset{\mathtt{g}}{\cdot} \theta^{\mathbf{j}} = \delta_{\mathbf{i}}^{\mathbf{j}}$.

For the particular case of a 4−dimensional spacetime, of course, the range of the bold labels are $\mathbf{a}, \mathbf{b}, \mathbf{i}, ... = 0, 1, 2, 3$. Recall that the fundamental *Clifford product* is generated by

$$\theta^{\mathbf{i}}\theta^{\mathbf{j}} + \theta^{\mathbf{j}}\theta^{\mathbf{i}} = 2\eta^{\mathbf{ij}}. \tag{4.139}$$

If $\mathcal{C} \in \sec \mathcal{C}\ell(M, \mathtt{g})$ is a Clifford field, we have:

$$\mathcal{C} = s + v_{\mathbf{i}}\theta^{\mathbf{i}} + \frac{1}{2!}b_{\mathbf{ij}}\theta^{\mathbf{i}}\theta^{\mathbf{j}} + \frac{1}{3!}t_{\mathbf{ijk}}\theta^{\mathbf{i}}\theta^{\mathbf{j}}\theta^{\mathbf{k}} + p\theta^{\mathbf{5}} \ , \tag{4.140}$$

where $\theta^{\mathbf{5}} = \theta^{\mathbf{0}}\theta^{\mathbf{1}}\theta^{\mathbf{2}}\theta^{3}$ is the volume element and

$$s, v_{\mathbf{i}}, b_{\mathbf{ij}}, t_{\mathbf{ijk}}, p \in \sec \bigwedge{}^0 T^*M \hookrightarrow \sec \mathcal{C}\ell(M, \mathtt{g}) \ . \tag{4.141}$$

4.5.2 Pullbacks and Relation Between Hodge Star Operators

Let M be a n-dimensional manifold and $\mathbf{g}$,$\mathring{\mathbf{g}} \in \sec T_2^0M$ two metrics of the same signature with corresponding metrics $\mathtt{g}$,$\mathring{\mathtt{g}} \in \sec T_0^2M$. Let g and $\mathring{g}$ be the extensor fields associated with $\mathtt{g}$ and $\mathring{\mathtt{g}}$. Let $\mathtt{h}$: $M \to M$ be a diffeomorphism and suppose that

$$\mathbf{g} \ = \mathtt{h}^*\mathring{\mathbf{g}} \ . \tag{4.142}$$

From the algebraic results of Sect. 2.7 we easily infer that there exists a metric gauge extensor field $h = h^\dagger$ such that

$$g(a) \underset{\mathring{\mathtt{g}}}{\cdot} b = h(a) \underset{\mathring{\mathtt{g}}}{\cdot} h(b) \ , \tag{4.143}$$

for any $a, b \in \sec \bigwedge^1 T^*M$ and we write $g = h^2$. Then, as in the purely algebraic case discussed in Sect. 2.7 we can also show that we have the following relation between the Hodge star operators associated with $\mathbf{g}$ and $\mathring{\mathbf{g}}$,

$$\underset{\mathbf{g}}{\star} = \underline{h}^{-1} \underset{\mathring{\mathbf{g}}}{\star} \underline{h} \ . \tag{4.144}$$

---

[17] For this reason Clifford fields are also called (non homogeneous) multiform fields.

In this case we say that the metric gauge extensor h is related to the pullback mapping $\mathtt{h}^*$.

We now show the

Proposition 265 *Let* $\mathtt{h} : M \longrightarrow M$ *a diffeomorphism. Let* $\mathbf{g}$,$\mathring{\mathbf{g}} \in \sec T_2^0 M$ *two metrics of the same signature. Then for any* $\omega \in \sec \bigwedge^p T^*M$ *we have*

$$\underset{\mathbf{g}}{\star}\,\mathtt{h}^*\omega = \mathtt{h}^*\underset{\mathring{\mathbf{g}}}{\star}\,\omega\ . \tag{4.145}$$

Proof. As in Remark 180 take two charts (U,φ) and (V,χ), U, $\mathtt{h}(U), \subset V$ with coordinate functions $\boldsymbol{x}^i$ and $\boldsymbol{y}^i$ such that and $\mathbf{x}^i(\mathfrak{e}) = \boldsymbol{y}^i(\mathtt{h}(\mathfrak{e}))$, i.e. calling $\mathbf{x}^i(\mathfrak{e}) = x^i$, $\boldsymbol{y}^i(\mathtt{h}(\mathfrak{e})) = y^i$ we have $\partial y^i/\partial x^j = \delta^i_j$, $dx^i = dy^i$. Let also $\mathring{\mathbf{g}}(\partial/\partial y^k, \partial/\partial y^l) = \mathring{\mathbf{g}}_{kl}(y^i)$. Then it follows that $\mathbf{g}_{kl}(x^i) = \mathring{\mathbf{g}}_{kl}(y^j(x^i)) = \mathring{\mathbf{g}}_{kl}(x^i)$ and $\det \mathbf{g} = \det \mathring{\mathbf{g}}$. Now, if $\omega = \frac{1}{p!}\omega_{i_1 \ldots i_p}\left(x^i\right) dx^{i_1} \wedge \ldots \wedge dx^{i_p}$, we can write (taking into account that $\bigwedge^p T^*M \hookrightarrow \mathcal{C}\ell(M,\mathbf{g})$ and also $\bigwedge^p T^*M \hookrightarrow \mathcal{C}\ell(M,\mathring{\mathbf{g}})$)

$$\begin{aligned}
\underset{\mathbf{g}}{\star}\mathtt{h}^*\omega &= \widetilde{\mathtt{h}^*\omega}\underset{\mathbf{g}}{\lrcorner}\tau_{\mathbf{g}} = \widetilde{\mathtt{h}^*\omega}\tau_{\mathbf{g}} \\
&= \frac{1}{p!}\omega_{i_1 \ldots i_p}\left(y^i(x^j)\right)dx^{i_1}\,\widetilde{\wedge \ldots \wedge}\,dx^{i_p}\sqrt{|\det \mathbf{g}|}dx^1 \wedge \ldots \wedge dx^n \\
&= \frac{1}{p!}\omega_{i_1 \ldots i_p}(y^i(x^j))dy^{i_1}\,\widetilde{\wedge \ldots \wedge}\,dy^{i_p}\sqrt{|\det \mathring{\mathbf{g}}|}dy^1 \wedge \ldots \wedge dy^n \\
&= \frac{1}{p!}\omega_{i_1 \ldots i_p}\left(y^i\right)dy^{i_1}\,\widetilde{\wedge \ldots \wedge}\,dy^{i_p}\underset{\mathring{\mathbf{g}}}{\lrcorner}\sqrt{|\det \mathring{\mathbf{g}}|}dy^1 \wedge \ldots \wedge dy^n \\
&= \mathtt{h}^*\underset{\mathring{\mathbf{g}}}{\star}\,\omega\ ,
\end{aligned}$$

and the proposition is proved. □

Remark 266 *When* $\mathbf{g} = \mathtt{h}^*\mathring{\mathbf{g}}$, *there exists an associated metric gauge extensor field* h *satisfying* (4.143), *i.e.* $g = h^2$. *The relation* $\underset{\mathbf{g}}{\star}\mathtt{h}^*\omega = \mathtt{h}^*\underset{\mathring{\mathbf{g}}}{\star}\,\omega$ *and* $\underset{\mathbf{g}}{\star} = \underline{h}^{-1}\underset{\mathring{\mathbf{g}}}{\star}\,\underline{h}$ *permit us to write the suggestive* operator *identity*

$$\underset{\mathring{\mathbf{g}}}{\star}\,\underline{h}\mathtt{h}^*\omega = \underline{h}\mathtt{h}^*\underset{\mathring{\mathbf{g}}}{\star}\,\omega\ . \tag{4.146}$$

Exercise 267 *Consider any diffeomorphism* $\mathtt{h} : M \to M$, *and two metrics* $\mathbf{g}$ *and* $\mathring{\mathbf{g}}$ *such that* $\mathbf{g} = \mathtt{h}^*\mathring{\mathbf{g}}$. *Show that*

$$\underset{\mathbf{g}}{\star}\,d\underset{\mathbf{g}}{\star}\,\mathtt{h}^*\omega = \mathtt{h}^*\underset{\mathring{\mathbf{g}}}{\star}\,d\underset{\mathring{\mathbf{g}}}{\star}\,\omega\ , \tag{4.147}$$

for any $\omega \in \bigwedge T^*M$.

Solution: The first member of (4.147) can be writing successively using (4.145) as

$$
\begin{aligned}
\underset{\mathbf{g}}{\star} d \underset{\mathbf{g}}{\star} \mathtt{h}^{*} \omega &= \underset{\mathbf{g}}{\star} d \mathtt{h}^{*} \underset{\mathring{\mathbf{g}}}{\star} \omega \\
&= \underset{\mathbf{g}}{\star} \mathtt{h}^{*} d \underset{\mathring{\mathbf{g}}}{\star} \omega \\
&= \mathtt{h}^{*} \underset{\mathring{\mathbf{g}}}{\star} d \underset{\mathring{\mathbf{g}}}{\star} \omega \,.
\end{aligned}
$$

4.5.3 Dirac Operators

We now equip the Riemannian (pseudo-Riemannian, or Lorentzian) manifold $(M, \mathring{\mathbf{g}})$ with a *standard* (or fiducial) structure $(M, \mathring{\mathbf{g}}, \mathring{D})$, where $\mathring{D}$ is the Levi-Civita connection of $\mathring{\mathbf{g}}$.

We are going to introduce in the Clifford bundle of differential forms $\mathcal{C}\ell(M, \mathring{\mathtt{g}})$ a differential operator $\mathring{\boldsymbol{\partial}}$, called the standard Dirac operator[18], which is associated with the Levi-Civita connection of the structure $(M, \mathring{\mathbf{g}}, \mathring{D})$ and we study the properties of that operator. Next we define new Dirac-like operators associated with a connection different from the Levi-Civita one, i.e. to connections ∇ defining a general Riemann-Cartan-Weyl geometry $(M, \mathring{\mathbf{g}}, \nabla)$. Moreover, making use of the results developed in Sect. 2.7, we show that it is possible to introduce infinitely many others Dirac-like operators, one for each bilinear form field defined on the manifold M of the structure $(M, \mathring{\mathbf{g}}, \mathring{D})$. These constructions enable us to formulate the geometry of a Riemann-Cartan-Weyl space in the Clifford bundle $\mathcal{C}\ell(M, \mathring{\mathtt{g}})$. Some interesting geometrical concepts, like the Dirac *commutator* and *anticommutator*, are introduced. Moreover, we show a new decomposition of a general linear connection, identifying some relevant tensors which are important for a clear understanding of any formulation of the gravitational theory in flat Minkowski spacetime (Chap. 9) and other related subjects appearing in the literature.

The Standard Dirac Operator

Given $\mathbf{u} \in \sec TM$ and $A \in \sec \bigwedge^{r} T^{*}M \hookrightarrow \sec \mathcal{C}\ell(M, \mathring{\mathtt{g}})$ consider the tensorial mapping $A \mapsto \mathring{D}_{\mathbf{u}} A \in \sec \bigwedge^{r} T^{*}M \hookrightarrow \sec \mathcal{C}\ell(M, \mathring{\mathtt{g}})$. Since $\mathring{D}_{\mathbf{u}} J_{\mathring{\mathtt{g}}} \subseteq J_{\mathring{\mathtt{g}}}$, where $J_{\mathring{\mathtt{g}}}$ is the ideal used in the definition of $\mathcal{C}\ell(M, \mathring{\mathtt{g}})$, we see immediately that the notion of covariant derivative (related to the Levi-Civita connection[19]) pass to the quotient bundle $\mathcal{C}\ell(M, \mathring{\mathtt{g}})$, i.e. given $A, B \in \sec \bigwedge^{r} T^{*}M \hookrightarrow \sec \mathcal{C}\ell(M, \mathring{\mathtt{g}})$

---

[18] It is crucial to distinguish the Dirac operators introduced in this chapter and which act on sections of Clifford bundles with the spin Dirac operator introduced in Chap. 6 and which act on sections of spin-Clifford bundles.

[19] And more generally, to any metric compatible connection.

(%

Pág. 392 (403) -

→ transformation function $\langle p|x\rangle$

We express its matrix element as a Fourier integral by inserting the spectral projection $[p]$ for momentum, keying the transformation function $\langle p|x\rangle$ from position eigenvectors $|x\rangle$ to momentum ones $|p\rangle$.

$$\langle p|x\rangle = \frac{1}{\sqrt{2\pi}} e^{ipx} \qquad (6).$$

∴ this integral is the standard Gaussian, integrated along a line in the complex plane from the real axis by $\pi/4$:

$$\int dp\, e^{-ip^2 \delta t/2} = \sqrt{\frac{2p}{i\delta t}}. \qquad (10)$$

Pág. 407 (418) 13.2 Gauge Physics

—

_ 20211016 / Einstein /

The Many Faces of Maxwell, Dirac and Einstein Equations..pdf

Obs: Riemann → 120 occurrences ↙

Gauss → 0 (???) estudar!!!

(%

Pág. 117 (129)

4.3° Differential Geometry in the Hodge Bundle

4.3.1 Riemannian and Lorentzian Structures on M.

Pág. 124 (136)

4.4 Classification of Geometries on M and Spacetimes

... "For each metric tensor defined on the manifold M there exists one and only one connection in the conditions of (4.133). It is called Levi-Civita connection of the metric considered, and is denoted by D.

Definition 251 Given a triple (M, g, ∇):

→ Riemann-Cartan space (4.131).

4.4.1 Spacetimes

oriented Lorentzian manifold structures (M, g, τ_g), which play a key role in physical theories.

(% %

Pág. 129(141)-

4.5.3 Dirac operators

∇ defining a general

Riemann-Cartan-Weyl geometry

$\underline{(M, \mathring{g}, \nabla)}$-

20211017_{III}

_20211016/New Perpectives in Algebra, Topology and Categories|

3030843181.pdf

Pág. 190(199)

3.9 Classical and Generalized (Point-Free) spaces

"sobrification" of a space →

"spatialization" of a locale

Pág. 183(192) – 3 Frames and Spaces

3.1 The Category of Frames

Similarly we have coframe ↙

HOMOMORPHISMS

! -256- - -* ! - % % '% 1"
(/ // % % ! ! ! " 7 ! , " < 1!

3.9 Classical and Generalized (Point-Free) Spaces

Now we are ready for a rough outline of the relation of the point-free and classical spaces.

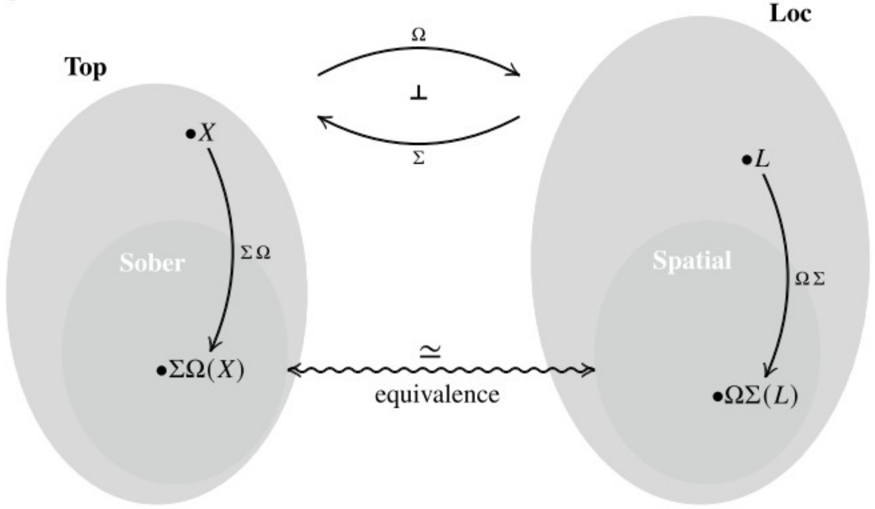

"sobrification" of a space *"spatialization" of a locale*

On the one hand the point-free theory extends the classical one (or, at least a substantial part of it; precisely, the subcategory of sober spaces). On the other hand, the scope of the generalized is much larger; we will present two simple examples shortly. The reader may of course ask whether such an extension is useful. It turns out that it is, as will be hopefully apparent from the following sections.

3.9.1 Two Easy Examples First, consider a complete Boolean algebra B without atoms (e.g., the lattice of all *regular open sets*[2] of any Euclidean space). There are no primes at all: indeed, let p be one. Since it is not an atom, there is an x with $p < x < 1$. We have $x^* \wedge x = 0 \leq p$, hence $x^* = 0$ and $x = 1$, a contradiction.

Next, take such a B again and consider $L = \{(x, y) \in B \times B \mid x \leq y\}$. Since all the $(0, x)$ are in L we obtain by the same reasoning as above that a prime in L has to be of the form $(q, 1)$ and hence there are no primes to separate distinct $(0, a)$ and $(0, b)$.

The latter example seems to be very similar to the former, but it is in fact much more interesting. While Boolean frames are something like a generalization of discrete spaces (albeit constituting a much more colorful class), the frames L of this example are geometrically rather peculiar (do not forget that a subframe is, due to the contravariance, geometrically more like a quotient space, not like a subspace):

---

[2] That is, the opens $U = \operatorname{int}\overline{U}$, "open sets without lesions", the open sets one thinks about first.

interpreted as spaces they are behaving like Hausdorff ones, while on the other hand they are not even subfit (see 7.6.2 below) which is a property weaker than T_1!

4 Categorical Remarks

4.1 *Semilattices and a Free Functor*

Under a *semilattice* we here understand a *meet semilattice with 0 and 1*, and *semilattice homomorphisms* preserve $\wedge$, 0 and 1.

4.1.1 Biproduct Note that in the category of semilattices (similarly like in abelian groups), the cartesian product with the injections and projections as in the following diagram

$$\begin{tikzcd}
L_1 \arrow[dr, bend right=10, "\iota_1=(a\mapsto(a,1))"] & & L_1 \\
& L_1\times L_2 \arrow[ur, two heads, bend left=10, "p_1=((a_1,a_2)\mapsto a_1)"] \arrow[dr, two heads, bend right=10, "p_2=((a_1,a_2)\mapsto a_2)"'] & \\
L_2 \arrow[ur, bend left=10, "\iota_2=(a\mapsto(1,a))"'] & & L_2
\end{tikzcd}$$

constitutes a biproduct (note that $p_1\iota_1 \wedge p_2\iota_2 = \mathrm{id}$ and check the coproduct and product properties).

4.1.2 A Free Construction For a semilattice L consider the down-set lattice

$$\mathfrak{D}(L) = \{U \subseteq L \mid \emptyset \neq U = {\downarrow}U\}$$

ordered by inclusion. Further, define $\lambda = \lambda_L : L \to \mathfrak{D}(L)$ by setting $\lambda(a) = {\downarrow}a$. Obviously,

$\mathfrak{D}(L)$ is a frame and λ is a semilattice homomorphism

(since we take only the non-empty down-sets, the zero of $\mathfrak{D}(L)$ is $\{0\}$; all the other joins are the unions).

Proposition *Let M be a frame and let $h: L \to M$ be a semilattice homomorphism. Then there is precisely one frame homomorphism $\widetilde{h}: \mathfrak{D}(L) \to M$ such that the diagram*

2.6.4 Observation *Let* b *have a complement* b^{c} *in a distributive lattice* L. *Then*

$$a \wedge b \leq c \text{ iff } a \leq b^{\mathsf{c}} \vee c, \quad \text{and} \quad a \vee b \geq c \text{ iff } a \geq c \wedge b^{\mathsf{c}}.$$

Thus, in a Heyting algebra we have for any complemented element $b, b \to c = b^* \vee c$ *and in a coHeyting algebra we have for any complemented element* $b, c \smallsetminus b = c \wedge b^{\#}$.

Note All the assumptions are essential, though. In particular the formulas $b \to c = b^* \vee c$ resp. $c \smallsetminus b = c \wedge b^{\#}$ hold for complemented elements only; the Heyting resp. coHeyting operation cannot be thus reduced to pseudocomplementing resp. supplementing in no other case.

2.7 *Boolean Algebras*

A *Boolean algebra* is a distributive lattice in which every element is complemented. From 2.6.4 we immediately obtain

Corollary *A Boolean algebra is both a Heyting and a coHeyting algebra.*

3 Frames and Spaces

3.1 *The Category of Frames*

A *frame* (resp. *coframe*) is a complete lattice L satisfying the distributivity law

$$(\bigvee A) \wedge b = \bigvee \{a \wedge b \mid a \in A\} \tag{frm}$$

$$\text{(resp. } (\bigwedge A) \vee b = \bigwedge \{a \vee b \mid a \in A\}\text{)} \tag{cofrm}$$

for all $A \subseteq L$ and $b \in L$. A *frame homomorphism* $h\colon L \to M$ between two frames is a mapping preserving all joins and all finite meets. The resulting category will be denoted by

$$\mathsf{Frm}.$$

Similarly we have *coframe homomorphisms* between coframes preserving all meets and finite joins.

3.2 Spaces and Frames. The Functor Ω

A typical frame is the lattice $\Omega(X)$ of all open subsets of a topological space. Furthermore, if $f\colon X \to Y$ is a continuous mapping we have a frame homomorphism $\Omega(f)\colon \Omega(Y) \to \Omega(X)$ defined by $\Omega(f)(U) = f^{-1}[U]$. Thus we obtain a *contravariant* functor

$$\Omega\colon \mathsf{Top} \to \mathsf{Frm},$$

a basic link between classical spaces and what will turn out to be the generalized ones (see already in 3.5.2 below, then in 3.9, and again and again).

3.2.1 A Warning Let us agree that our spaces will be, from now on, always T_0: the frames $\Omega(X)$ will be central in our approach to spaces and it will make no sense to discuss classical points that cannot be distinguished by open sets. In particular we will, without further particular mentioning, use the following

Observation *Let $f, g\colon X \to Y$ be distinct continuous maps, and let Y be T_0. Then $\Omega(f) \neq \Omega(g)$.*

(Indeed, if $f(x) \neq g(x)$ consider a U such that, say, $g(x) \notin U \ni f(x)$. Then $x \in \Omega(f)(U) \smallsetminus \Omega(g)(U)$.)

3.2.2 Notes (a) Unlike plain (finite) distributivity, the frame distributivity typically does not carry over to the dual, that is, a frame is seldom simultaneously a coframe. Take e.g. any T_1-space X with a non-isolated point x and an open set $W \ni x$. Set $V = W \smallsetminus \{x\}$ and $\mathcal{U} = \{U \in \Omega(X) \mid x \in U\}$. Then $\bigwedge \mathcal{U} = \operatorname{int} \bigcap \mathcal{U} = \emptyset$ and hence $(\bigwedge \mathcal{U}) \cup V = V$ while $\bigwedge\{U \cup V \mid U \in \mathcal{U}\} = W \neq V$.
(b) As the example shows, coframes will seldom come as models of (generalized) spaces. They will play, however, a fundamental role in the study of the structure of generalized subspaces.

3.3 The Heyting Structure

Recall 2.6 (and 2.2). The distributivity rule (frm) makes a frame a Heyting algebra and computing with the Heyting operation will be extensively used; similarly we will use computing with the difference in coframes. But we have to keep in mind that the category Frm is not that of (complete) Heyting algebras: frame homomorphisms generally do not respect the Heyting operation.

data

between coframes preserving all
meets and finite joins.

Quantum theory of Solids

Marilena Barnabei

www.unibo.it/sitoweb/marilena.
barnabei

marilena.barnabei@unibo.it

Igreja de Nossa Senhora de Sião
Axum, Tigré, Etiópia (Arca da
Aliança)
14°07'N, 38°43'L

⇒ Google Drive Share:
https://drive.google.com/
drive/u/1/folders/0ACQD97Je9R76Uk9PVA

STENIOMUSICH.com

https://drive.google.com/drive/u/1/folders/
0AAV8wobLXAu8Uk9PVA

/ (% $

Books:

1) _20211017/Dictionary of Distances/
Dictionary of Distances...pdf
Pág. 73(90) - 6.4
Non-Euclidean Geometry
obs.: Li mo todo!!!

Gauss - 24 occurrences.

Pág. 112(129)

For surfaces, two kinds of curvature are considered: Gaussian curvature, and mean curvature.

The Gaussian curvature is $K = k_1 k_2$ (it can be given entirely in terms of the first fundamental form).
The mean curvature is

$$H = \frac{1}{2}(k_1 + k_2).$$

Pág. 113(130) - Metric of positive curvature

Gaussian curvature - Metric of negative curvature

/
-256- - -* ! - / 4 ! 3 4 ! D3
! 4 ,! 1"
(% / B ' 8 = " ! 7 3
*D ? A " UUU

6.4. NON-EUCLIDEAN GEOMETRY

The term **non-Euclidean Geometry** describes both *Hyperbolic Geometry* (or *Lobachevsky Geometry*, *Lobachevsky–Bolyai–Gauss Geometry*) and *Elliptic Geometry* (sometimes called also *Riemannian Geometry*) which are contrasted with *Euclidean Geometry* (or *Parabolic Geometry*). The essential difference between Euclidean and non-Euclidean Geometry is the nature of parallel lines. In Euclidean Geometry, if we start with a line l and a point a, which is not on l, then we can only draw one line through a that is parallel to l. In Hyperbolic Geometry there are infinitely many lines through a parallel to l. In Elliptic Geometry, parallel lines do not exist.

The *Spherical Geometry* is also “non-Euclidean”, but it fails the axiom that any two points determine exactly one line.

- **Spherical metric**

  Let $S^n(0,r) = \{x \in \mathbb{R}^{n+1} : \sum_{i=1}^{n+1} x_i^2 = r^2\}$ be a sphere in $\mathbb{R}^{n+1}$ with the center 0 and the radius $r > 0$.

  The **spherical metric** (or **great circle metric**) d_{sph} is a metric on $S^n(0,r)$, defined by

$$r \arccos\left(\frac{|\sum_{i=1}^{n+1} x_i y_i|}{r^2}\right),$$

  where arccos is the inverse cosine in $[0, \pi]$. It is the length of the *great circle* arc, passing through x and y. In terms of the standard *inner product* $\langle x, y\rangle = \sum_{i=1}^{n+1} x_i y_i$ on $\mathbb{R}^{n+1}$, the spherical metric can be written as $r \arccos \frac{|\langle x,y\rangle|}{\sqrt{\langle x,x\rangle\langle y,y\rangle}}$.

  The metric space $(S^n(0,r), d_{sph})$ is called *n-dimensional spherical space*. It is a space of curvature $1/r^2$, and r is the radius of curvature. It is a model of n-dimensional *Spherical Geometry*. The great circles of the sphere are its geodesics, all geodesics are closed and of the same length. (See, for example, [Blum70].)

- **Elliptic metric**

  Let $\mathbb{R}P^n$ be the real n-dimensional projective space. The **elliptic metric** d_{ell} is a metric on $\mathbb{R}P^n$, defined by

$$r \arccos \frac{|\langle x, y\rangle|}{\sqrt{\langle x, x\rangle\langle y, y\rangle}},$$

  where, for any $x = (x_1 : \ldots : x_{n+1})$, $y = (y_1 : \ldots : y_{n+1}) \in \mathbb{R}P^n$, one has $\langle x, y\rangle = \sum_{i=1}^{n+1} x_i y_i$, r is a fixed positive constant, and arccos is the inverse cosine in $[0, \pi]$.

  The metric space $(\mathbb{R}P^n, d_{ell})$ is called *n-dimensional elliptic space*. It is a model of n-dimensional *Elliptic Geometry*. It is the space of curvature $1/r^2$, and r is the radius of curvature. As $r \to \infty$, the metric formulas of Elliptic Geometry yield formulas of Euclidean Geometry (or become meaningless).

  If $\mathbb{R}P^n$ is viewed as the set $E^n(0,r)$, obtained from the sphere $S^n(0,r) = \{x \in \mathbb{R}^{n+1} : \sum_{i=1}^{n+1} x_i^2 = r^2\}$ in $\mathbb{R}^{n+1}$ with the center 0 and the radius r by identifying diametrically-opposite points, then the elliptic metric on $E^n(0,r)$ can be written as $d_{sph}(x, y)$ if

$d_{sph}(x, y) \leqslant \frac{\pi}{2}r$, and as $\pi r - d_{sph}(x, y)$ if $d_{sph}(x, y) > \frac{\pi}{2}r$, where d_{sph} is the **spherical metric** on $S^n(0, r)$. Thus, no two points of $E^n(0, r)$ have distance exceeding $\frac{\pi}{2}r$. The elliptic space $(E^2(0, r), d_{ell})$ is called *Poincaré sphere*.

If $\mathbb{R}P^n$ is viewed as the set E^n of lines through the zero element 0 in $\mathbb{R}^{n+1}$, then the elliptic metric on E^n is defined as the angle between the corresponding subspaces.

An n-dimensional elliptic space is a *Riemannian space* of constant positive curvature. It is the only such space which is topologically equivalent to a projective space. (See, for example, [Blum70], [Buse55].)

- **Hermitian elliptic metric**

Let $\mathbb{C}P^n$ be the n-dimensional complex projective space. The **Hermitian elliptic metric** d^H_{ell} (see, for example, [Buse55]) is a metric on $\mathbb{C}P^n$, defined by

$$r \arccos \frac{|\langle x, y\rangle|}{\sqrt{\langle x, x\rangle\langle y, y\rangle}},$$

where, for any $x = (x_1 : \dots : x_{n+1})$, $y = (y_1 : \dots : y_{n+1}) \in \mathbb{C}P^n$, one has $\langle x, y\rangle = \sum_{i=1}^{n+1} \overline{x}_i y_i$, r is a fixed positive constant, and arccos is the inverse cosine in $[0, \pi]$.

The metric space $(\mathbb{C}P^n, d^H_{ell})$ is called n-dimensional *Hermitian elliptic space* (cf. **Fubini–Study metric**).

- **Elliptic plane metric**

The **elliptic plane metric** is the **elliptic metric** on the *elliptic plane* $\mathbb{R}P^2$. If $\mathbb{R}P^2$ is viewed as the *Poincaré sphere* (i.e., a sphere in $\mathbb{R}^3$ with identified diametrically-opposite points) of diameter 1 tangent to the extended complex plane $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ at the point $z = 0$, then, under the stereographic projection from the "north pole" $(0, 0, 1)$, $\overline{\mathbb{C}}$ with identified points z and $-\frac{1}{\overline{z}}$ is a model of the elliptic plane, and the elliptic plane metric d_{ell} on it is defined by its *line element* $ds^2 = \frac{|dz|^2}{(1+|z|^2)^2}$.

- **Pseudo-elliptic distance**

The **pseudo-elliptic distance** (or *elliptic pseudo-distance*) dp_{ell} is a distance on the extended complex plane $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ with identified points z and $-\frac{1}{\overline{z}}$, defined by

$$\left|\frac{z - u}{1 + \overline{z}u}\right|.$$

In fact, $dp_{ell}(z, u) = \tan d_{ell}(z, u)$, where d_{ell} is the **elliptic plane metric**.

- **Hyperbolic metric**

Let $\mathbb{R}P^n$ be the n-dimensional real projective space. Let, for any $x = (x_1 : \dots : x_{n+1})$, $y = (y_1 : \dots : y_{n+1}) \in \mathbb{R}P^n$, the *scalar product* $\langle x, y\rangle = -x_1 y_1 + \sum_{i=2}^{n+1} x_i y_i$ be considered.

Chapter 8

Distances on Surfaces and Knots

8.1. GENERAL SURFACE METRICS

A *surface* is a real two-dimensional *manifold* M^2, i.e., a **Hausdorff space**, each point of which has a neighborhood which is homeomorphic to a plane $\mathbb{E}^2$, or a closed half-plane (cf. Chapter 7).

A compact orientable surface is called *closed* if it has no boundary, and it is called *surface with boundary*, otherwise. There are compact non-orientable surfaces (closed or with boundary); the simplest such surface is the *Möbius strip*. Non-compact surfaces without boundary are called *open*.

Any closed connected surface is homeomorphic to either a sphere with, say, g (cylindric) handles, or a sphere with, say, g *cross-caps* (i.e., caps with a twist like Möbius strip in them). In both cases the number g is called *genus* of the surface. In the case of handles, the surface is orientable; it is called *torus* (doughnut), *double torus*, and *triple torus* for $g = 1, 2$ and 3, respectively. In the case of cross-caps, the surface is non-orientable; it is called *real projective plane*, *Klein bottle*, and *Dyck's surface* for $g = 1, 2$ and 3, respectively. The genus is the maximal number of disjoint simple closed curves which can be cut from a surface without disconnecting it (the *Jordan curve theorem* for surfaces).

The *Euler–Poincaré characteristic* of a surface is (the same for all polyhedral decompositions of a given surface) the number $\chi = v - e + f$, where v, e and f are, respectively, the number of vertices, edges and faces of the decomposition. It holds $\chi = 2 - 2g$ if the surface is orientable, and $\chi = 2 - g$ if not. Every surface with boundary is homeomorphic to a sphere with appropriated number of (disjoint) *holes* (i.e., what remains if an open disk is removed) and handles or cross-caps. If h is the number of holes, then $\chi = 2 - 2g - h$ holds if the surface is orientable, and $\chi = 2 - g - h$ if not.

The *connectivity number* of a surface is the largest number of closed cuts that can be made on the surface without separating it into two or more parts. This number is equal to $3 - \chi$ for closed surfaces, and $2 - \chi$ for surfaces with boundaries. A surface with connectivity number 1, 2 and 3 is called, respectively, *simply, doubly* and *triply connected*. A sphere is simply connected, while a torus is triply connected.

A surface can be considered as a metric space with its own **intrinsic metric**, or as a figure in space. A surface in $\mathbb{E}^3$ is called *complete* if it is a **complete** metric space with respect to its intrinsic metric.

A surface is called *differentiable*, *regular*, or *analytic*, respectively, if in a neighborhood of each of its points it can be given by an expression

$$r = r(u, v) = r(x_1(u, v), x_2(u, v), x_3(u, v)),$$

where the *position vector* $r = r(u, v)$ is a differentiable, *regular* (i.e., a sufficient number of times differentiable), or *real analytic*, respectively, vector function satisfying the condition $r_u \times r_v \neq 0$.

Any regular surface has the intrinsic metric with the *line element* (or *first fundamental form*)

$$ds^2 = dr^2 = E(u, v)\,du^2 + 2F(u, v)\,du\,dv + G(u, v)\,dv^2,$$

where $E(u, v) = \langle r_u, r_u \rangle$, $F(u, v) = \langle r_u, r_v \rangle$, $G(u, v) = \langle r_v, r_v \rangle$. The length of a curve, defined on the surface by the equations $u = u(t)$, $v = v(t)$, $t \in [0, 1]$, is computed by

$$\int_0^1 \sqrt{Eu'^2 + 2Fu'v' + Gv'^2}\,dt,$$

and the distance between any points $p, q \in M^2$ is defined as the infimum of the lengths of all curves on M^2, connecting p and q. A **Riemannian metric** is a generalization of the first fundamental form of a surface.

For surfaces, two kinds of *curvature* are considered: *Gaussian curvature*, and *mean curvature*. To compute those curvatures at a given point of the surface, consider the intersection of the surface with a plane, containing a fixed *normal vector*, i.e., a vector which is perpendicular to the surface at this point. This intersection is a plane curve. The *curvature* k of this plane curve is called *normal curvature* of the surface at the given point. If we vary the plane, the normal curvature k will change, and there are two extremal values – the *maximal curvature* k_1, and the *minimal curvature* k_2, called *principal curvatures* of the surface. A curvature is taken to be *positive* if the curve turns in the same direction as the surface's chosen normal, otherwise it is taken to be *negative*. The *Gaussian curvature* is $K = k_1 k_2$ (it can be given entirely in terms of the first fundamental form). The *mean curvature* is $H = \frac{1}{2}(k_1 + k_2)$.

A *minimal surface* is a surface with mean curvature zero, or, equivalently, a surface of minimum area subject to constraints on the location of its boundary.

A *Riemann surface* is an one-dimensional *complex manifold*, or a two-dimensional real manifold with a complex structure, i.e., in which the local coordinates in neighborhoods of points are related by complex analytic functions. It can be thought as a deformed version of the complex plane. All Riemann surfaces are orientable. Closed Riemann surfaces are geometrical models of *complex algebraic curves*. Every connected Riemann surface can be turned into a *complete* two-dimensional *Riemannian manifold* with constant curvature -1, 0, or 1. The Riemann surfaces with the curvature -1 are called *hyperbolic*, the *unit disk* $\Delta = \{z \in \mathbb{C}: \ |z| < 1\}$ is the canonical example. The Riemann surfaces with the curvature 0 are called *parabolic*, $\mathbb{C}$ is a typical example. The Riemann surfaces with the curvature 1 are called *elliptic*, the *Riemann sphere* $\mathbb{C} \cup \{\infty\}$ is a typical example.

- **Regular metric**

The intrinsic metric of a surface is called **regular** if it can be specified using the *line element*

$$ds^2 = E\,du^2 + 2F\,du\,dv + G\,dv^2,$$

where the coefficients of the form ds^2 are regular functions.

Any regular surface, given by an expression $r = r(u, v)$, has a regular metric with the *line element* ds^2, where $E(u, v) = \langle r_u, r_u \rangle$, $F(u, v) = \langle r_u, r_v \rangle$, $G(u, v) = \langle r_v, r_v \rangle$.

- **Analytic metric**

  The intrinsic metric on a surface is called **analytic** if it can be specified using the *line element*

  $$ds^2 = E\,du^2 + 2F\,du\,dv + G\,dv^2,$$

  where the coefficients of the form ds^2 are real analytic functions.

  Any analytic surface, given by an expression $r = r(u, v)$, has an analytic metric with the *line element* ds^2, where $E(u, v) = \langle r_u, r_u \rangle$, $F(u, v) = \langle r_u, r_v \rangle$, $G(u, v) = \langle r_v, r_v \rangle$.

- **Metric of positive curvature**

  A **metric of positive curvature** is the intrinsic metric on a *surface of positive curvature*.

  A *surface of positive curvature* is a surface in $\mathbb{E}^3$ that has positive Gaussian curvature at every point.

- **Metric of negative curvature**

  A **metric of negative curvature** is the intrinsic metric on a *surface of negative curvature*.

  A *surface of negative curvature* is a surface in $\mathbb{E}^3$ that has negative Gaussian curvature at every point. A surface of negative curvature locally have a saddle-like structure. The intrinsic geometry of a surface of constant negative curvature (in particular, of a *pseudo-sphere*) locally coincides with the geometry of *Lobachevsky plane*. There exists no surface in $\mathbb{E}^3$, whose intrinsic geometry coincides completely with the geometry of Lobachevsky plane (i.e., a complete regular surface of constant negative curvature).

- **Metric of non-positive curvature**

  A **metric of non-positive curvature** is the intrinsic metric on a *saddle-like surface*.

  A *saddle-like surface* is a generalization of a surface of negative curvature: a twice continuously-differentiable surface is a saddle-like surface if and only if at each point of the surface its Gaussian curvature is non-positive. These surfaces can be seen as antipodes of *convex surfaces*, but they do not form such a natural class of surfaces as do convex surfaces.

- **Metric of non-negative curvature**

  A **metric of non-negative curvature** is the intrinsic metric on a *convex surface*.

  A *convex surface* is a *domain* (i.e., a connected open set) on the boundary of a *convex body* in $\mathbb{E}^3$ (in some sense, it is an antipode of saddle-like surface). The entire boundary of a convex body is called *complete convex surface*. If the body is finite (bounded), the complete convex surface is called *closed*. Otherwise, it is called *infinite* (an infinite convex surface is homeomorphic to a plane or to a circular cylinder).

Any convex surface M^2 in $\mathbb{E}^3$ is a *surface of bounded curvature*. The *total Gaussian curvature* $w(A) = \iint_A K(x)\,d\sigma(x)$ of a set $A \subset M^2$ is always non-negative (here $\sigma(.)$ is the *area*, and $K(x)$ is the *Gaussian curvature* of M^2 at a point x), i.e., a convex surface can be seen as a *surface of non-negative curvature*.

The intrinsic metric of a convex surface is a **convex metric** in the sense of surface theory, i.e., it displays the *convexity condition*: the sum of the angles of any triangle whose sides are shortest curves is not less that π.

- **Metric with alternating curvature**

A **metric with alternating curvature** is the intrinsic metric on a surface with alternating (positive or negative) Gaussian curvature.

- **Flat metric**

A **flat metric** is the intrinsic metric on a *developable surface*, i.e., a surface, on which the Gaussian curvature is everywhere zero.

- **Metric of bounded curvature**

A **metric of bounded curvature** is the intrinsic metric ρ on a *surface of bounded curvature*.

A surface M^2 with an intrinsic metric ρ is called *surface of bounded curvature* if there exists a sequence of **Riemannian metrics** ρ_n, defined on M^2, such that for any compact set $A \subset M^2$ one has $\rho_n \to \rho$ uniformly, and the sequence $|w_n|(A)$ is bounded, where $|w|_n(A) = \iint_A |K(x)|\,d\sigma(x)$ is *total absolute curvature* of the metric ρ_n (here $K(x)$ is the Gaussian curvature of M^2 at a point x, and $\sigma(.)$ is the *area*).

- ***Λ*-metric**

A ***Λ*-metric** (or *metric of type Λ*) is a **complete** metric on a surface with curvature bounded from above by a negative constant.

An Λ-metric does not have embeddings into $\mathbb{E}^3$. It is a generalization of the classical result of Hilbert (1901): no complete regular surface of constant negative curvature (i.e., a surface whose intrinsic geometry coincides completely with the geometry of Lobachevsky plane) exists in $\mathbb{E}^3$.

- **(*h*, *Δ*)-metric**

An (h, Δ)**-metric** is a metric on a surface with a slowly-changing negative curvature.

A **complete** (h, Δ)-metric does not permit a regular *isometric embedding* in three-dimensional Euclidean space (cf. ***Λ*-metric**).

- ***G*-distance**

A connected set G of points on a surface M^2 is called *geodesic region* if, for each point $x \in G$, there exists a *disk* $B(x, r)$ with center at x, such that $B_G = G \cap B(x, r)$ has one of the following forms: $B_G = B(x, r)$ (x is a *regular interior point* of G); B_G is a semi-disk of $B(x, r)$ (x is a *regular boundary point* of G); B_G is a sector of $B(x, r)$

other than a semi-disk (x is an *angular point* of G); B_G consists of a finite number of sectors of $B(x, r)$ with no common points except x (x is a *nodal point* of G).

The **G-distance** between any x and $y \in G$ is defined as the greatest lower bound of the lengths of all rectifiable curves connecting x and $y \in G$, and completely contained in G.

- **Conformally invariant metric**

Let R be a Riemann surface. A *local parameter* (or *local uniformizing parameter*, *local uniformizer*) is a complex variable z considered as a continuous function $z_{p_0} = \phi_{p_0}(p)$ of a point $p \in R$ which is defined everywhere in some neighborhood (*parametric neighborhood*) $V(p_0)$ of a point $p_0 \in R$ and which realizes a homeomorphic mapping (*parametric mapping*) of $V(p_0)$ onto the disk (*parametric disk*) $\Delta(p_0) = \{z \in \mathbb{C}\colon\ |z| < r(p_0)\}$, where $\phi_{p_0}(p_0) = 0$. Under a parametric mapping, any point function $g(p)$, defined in the parametric neighborhood $V(p_0)$, goes into a function of the local parameter z: $g(p) = g(\phi_{p_0}^{-1}(z)) = G(z)$.

A **conformally invariant metric** is a differential $\rho(z)|dz|$ on the Riemann surface R which is invariant with respect to the choice of the local parameter z. Thus, to each local parameter z ($z : U \to \overline{\mathbb{C}}$) a function $\rho_z : z(U) \to [0, \infty]$ is associated such that, for any local parameters z_1 and z_2, we have:

$$\frac{\rho_{z_2}(z_2(p))}{\rho_{z_1}(z_1(p))} = \left|\frac{dz_1(p)}{dz_2(p)}\right| \quad \text{for any } p \in U_1 \cap U_2.$$

Every linear differential $\lambda(z)dz$ and every *quadratic differential* $Q(z)dz^2$ induce conformally invariant metrics $|\lambda(z)\|dz|$ and $|Q(z)|^{1/2}\|dz|$, respectively (cf. **Q-metric**).

- **Q-metric**

An **Q-metric** is a **conformally invariant metric** $\rho(z)|dz| = |Q(z)|^{1/2}|dz|$ on a Riemann surface R, defined by a *quadratic differential* $Q(z)dz^2$.

A *quadratic differential* $Q(z)dz^2$ is a non-linear differential on a Riemann surface R which is invariant with respect to the choice of the local parameter z. Thus, to each local parameter z ($z : U \to \overline{\mathbb{C}}$) a function $Q_z : z(U) \to \overline{\mathbb{C}}$ is associated such that, for any local parameters z_1 and z_2, we have:

$$\frac{Q_{z_2}(z_2(p))}{Q_{z_1}(z_1(p))} = \left(\frac{dz_1(p)}{dz_2(p)}\right)^2 \quad \text{for any } p \in U_1 \cap U_2.$$

- **Extremal metric**

An **extremal metric** is a **conformally invariant metric** in the *modulus problem* for a family Γ of locally rectifiable curves on a Riemann surface R which realizes the infimum in the definition of the *modulus* $M(\Gamma)$.

Formally, let Γ be a family of locally rectifiable curves on a Riemann surface R, let P be a non-empty class of conformally invariant metrics $\rho(z)|dz|$ on R such that $\rho(z)$ is square-integrable in the z-plane for every local parameter z, and the integrals

$$A_\rho(R) = \iint_R \rho^2(z)\,dx\,dy \quad \text{and} \quad L_\rho(\Gamma) = \inf_{\gamma\in\Gamma}\int_y \rho(z)|dz|$$

are not simultaneously equal to 0 or ∞ (each of the above integrals is understood as a Lebesgue integral). The *modulus of the family of curves* Γ is defined by

$$M(\Gamma) = \inf_{\rho\in P}\frac{A_\rho(R)}{(L_\rho(\Gamma))^2}.$$

The *extremal length of the family of curves* Γ is equal to $\sup_{\rho\in P}\frac{(L_\rho(\Gamma))^2}{A_\rho(R)}$, i.e., is the reciprocal of $M(\Gamma)$.

The modulus problem for Γ is defined as follows: let P_L be the subclass of P such that, for any $\rho(z)|dz| \in P_L$ and any $\gamma \in \Gamma$, one has $\int_\gamma \rho(z)|dz| \geqslant 1$. If $P_L \neq \emptyset$, then the modulus $M(\Gamma)$ of the family Γ can be written as $M(\Gamma) = \inf_{\rho\in P_L} A_\rho(R)$. Every metric from P_L is called **admissible metric** for the modulus problem on Γ. If there exists ρ^* for which

$$M(\Gamma) = \inf_{\rho\in P_L} A_\rho(R) = A_{\rho^*}(R),$$

the metric $\rho^*|dz|$ is called **extremal metric** for the modulus problem on Γ.

- **Fréchet surface metric**

Let (X, d) be a metric space, M^2 be a compact two-dimensional manifold, f be a continuous mapping $f : M^2 \to X$, called *parameterized surface*, and $\sigma : M^2 \to M^2$ be a homeomorphism of M^2 onto itself. Two parameterized surfaces f_1 and f_2 are called *equivalent* if $\inf_\sigma \max_{p\in M^2} d(f_1(p), f_2(\sigma(p))) = 0$, where the infimum is taken over all possible homeomorphisms σ. A class f^* of parameterized surfaces, equivalent to f, is called *Fréchet surface*. It is a generalization of the notion of a surface in an Euclidean space to the case of an arbitrary metric space (X, d).

The **Fréchet surface metric** is a metric on the set of all Fréchet surfaces, defined by

$$\inf_\sigma \max_{p\in M^2} d\big(f_1(p), f_2\big(\sigma(p)\big)\big)$$

for any Fréchet surfaces f_1^* and f_2^*, where the infimum is taken over all possible homeomorphisms σ (cf. **Fréchet metric**).

8.2. INTRINSIC METRICS ON SURFACES

In this section we list intrinsic metrics, given by their *line elements* (which, in fact, are two-dimensional **Riemannian metrics**), for some selected surfaces.

- **Quadric metric**

A *quadric* (or *quadratic surface*, *surface of second order*) is a set of points in $\mathbb{E}^3$, whose coordinates in a Cartesian coordinate system satisfy an algebraic equation of degree two. There are 17 classes of such surfaces, among them are: *ellipsoids*, *one-sheet* and *two-sheet hyperboloids*, *elliptic paraboloids*, *hyperbolic paraboloids*, *elliptic*, *hyperbolic* and *parabolic cylinders*, and *conical surfaces*.

For example, a *cylinder* can be given by the following parametric equations:

$$x_1(u,v)=a\cos v,\quad x_2(u,v)=a\sin v,\quad x_3(u,v)=u.$$

The intrinsic metric on it is given by the *line element*

$$ds^2=du^2+a^2\,dv^2.$$

An *elliptic cone* (i.e., a cone with elliptical cross-section) has the following parametric equations:

$$x_1(u,v)=a\frac{h-u}{h}\cos v,\quad x_2(u,v)=b\frac{h-u}{h}\sin v,\quad x_3(u,v)=u,$$

where h is the *height*, a is the *semi-major axis*, and b is the *semi-minor axis* of the cone. The intrinsic metric on it is given by the *line element*

$$ds^2=\frac{h^2+a^2\cos^2 v+b^2\sin^2 v}{h^2}\,du^2+2\frac{(a^2-b^2)(h-u)\cos v\sin v}{h^2}\,du\,dv + \frac{(h-u)^2(a^2\sin^2 v+b^2\cos^2 v)}{h^2}\,dv^2.$$

- **Sphere metric**

A *sphere* is a *quadric*, given by the Cartesian equation $(x_1-a)^2+(x_2-b)^2+(x_3-c)^2=r^2$, where the point (a,b,c) is the *center* of the sphere, and $r>0$ is the *radius* of the sphere. The sphere of radius r, centered at the origin, can be given by the following parametric equations:

$$x_1(\theta,\phi)=r\sin\theta\cos\phi,\quad x_2(\theta,\phi)=r\sin\theta\sin\phi,\quad x_3(\theta,\phi)=r\cos\theta,$$

where the *azimuthal angle* $\phi\in[0,2\pi)$, and the *polar angle* $\theta\in[0,\pi]$. The intrinsic metric on it (in fact, the two-dimensional **spherical metric**) is given by the *line element*

$$ds^2=r^2\,d\theta^2+r^2\sin^2\theta\,d\phi^2.$$

A sphere of radius r has constant positive Gaussian curvature equal to r.

- **Ellipsoid metric**

  An *ellipsoid* is a *quadric* given by the Cartesian equation $\frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} + \frac{x_3^2}{c^2} = 1$, or by the following parametric equations:

  $$x_1(\theta, \phi) = a\cos\phi\sin\theta, \quad x_2(\theta, \phi) = b\sin\phi\sin\theta, \quad x_3(\theta, \phi) = c\cos\theta,$$

  where the *azimuthal angle* $\phi \in [0, 2\pi)$, and the *polar angle* $\theta \in [0, \pi]$. The intrinsic metric on it is given by the *line element*

  $$\begin{aligned} ds^2 = {} & \big(b^2\cos^2\phi + a^2\sin^2\phi\big)\sin^2\theta\, d\phi^2 + \big(b^2 - a^2\big)\cos\phi\sin\phi\cos\theta\sin\theta\, d\theta\, d\phi \\ & + \big(\big(a^2\cos^2\phi + b^2\sin^2\phi\big)\cos^2\theta + c^2\sin^2\theta\big)\, d\theta^2. \end{aligned}$$

- **Spheroid metric**

  A *spheroid* is an *ellipsoid* having two axes of equal length. It is also a *rotation surface*, given by the following parametric equations:

  $$x_1(u, v) = a\sin v\cos u, \quad x_2(u, v) = a\sin v\sin u, \quad x_3(u, v) = c\cos v,$$

  where $0 \leqslant u < 2\pi$, and $0 \leqslant v \leqslant \pi$. The intrinsic metric on it is given by the *line element*

  $$ds^2 = a^2\sin^2 v\, du^2 + \frac{1}{2}\big(a^2 + c^2 + \big(a^2 - c^2\big)\cos(2v)\big)\, dv^2.$$

- **Hyperboloid metric**

  A *hyperboloid* is a *quadric* which may be one- or two-sheeted. The one-sheeted hyperboloid is a *surface of revolution* obtained by rotating a hyperbola about the perpendicular bisector to the line between the foci, while the two-sheeted hyperboloid is a surface of revolution obtained by rotating a hyperbola about the line joining the foci. The one-sheeted circular hyperboloid, oriented along the x_3-axis, is given by the Cartesian equation $\frac{x_1^2}{a^2} + \frac{x_2^2}{a^2} - \frac{x_3^2}{c^2} = 1$, or by the following parametric equations:

  $$x_1(u, v) = a\sqrt{1 + u^2}\cos v, \quad x_2(u, v) = a\sqrt{1 + u^2}\sin v, \quad x_3(u, v) = cu,$$

  where $v \in [0, 2\pi)$. The intrinsic metric on it is given by the *line element*

  $$ds^2 = \left(c^2 + \frac{a^2u^2}{u^2 + 1}\right) du^2 + a^2\big(u^2 + 1\big)\, dv^2.$$

- **Rotation surface metric**

  A *rotation surface* (or *surface of revolution*) is a surface generated by rotating a two-dimensional curve about an axis. It is given by the following parametric equations:

  $$x_1(u, v) = \phi(v)\cos u, \quad x_2(u, v) = \phi(v)\sin u, \quad x_3(u, v) = \psi(v).$$

The intrinsic metric on it is given by the *line element*

$$ds^2 = \phi^2\, du^2 + \left(\phi'^2 + \psi'^2\right) dv^2.$$

- **Pseudo-sphere metric**

  A *pseudo-sphere* is a half of the *rotation surface*, generated by rotating a *tractrix* about its asymptote. It is given by the following parametric equations:

  $$x_1(u, v) = \operatorname{sech} u \cos v, \quad x_2(u, v) = \operatorname{sech} u \sin v, \quad x_3(u, v) = u - \tanh u,$$

  where $u \geqslant 0$, and $0 \leqslant v < 2\pi$. The intrinsic metric on it is given by the *line element*

  $$ds^2 = \tanh^2 u\, du^2 + \operatorname{sech}^2 u\, dv^2.$$

  The pseudo-sphere has constant negative Gaussian curvature equal to -1, and in this sense is an analog of the sphere which has constant positive Gaussian curvature.

- **Torus metric**

  A *torus* is a surface having genus one. A torus azimuthally symmetric about the x_3-axis is given by the Cartesian equation $(c - \sqrt{x_1^2 + x_2^2})^2 + x_3^2 = a^2$, or by the following parametric equations:

  $$x_1(u, v) = (c + a\cos v)\cos u, \quad x_2(u, v) = (c + a\cos v)\sin u, \quad x_3(u, v) = a\sin v,$$

  where $c > a$, and $u, v \in [0, 2\pi)$. The intrinsic metric on it is given by the *line element*

  $$ds^2 = (c + a\cos v)^2\, du^2 + a^2\, dv^2.$$

- **Helical surface metric**

  A *helical surface* (or *surface of screw motion*) is a surface described by a plane curve γ which, while rotating around an axis at a uniform rate, also advances along that axis at a uniform rate. If γ is located in the plane of the axis of rotation x_3 and is defined by the equation $x_3 = f(u)$, the position vector of the helical surface is

  $$r = (u\cos v,\ u\sin v,\ f(u) = hv), \quad h = \text{const},$$

  and the intrinsic metric on it is given by the *line element*

  $$ds^2 = \left(1 + f'^2\right) du^2 + 2hf'\, du\, dv + \left(u^2 + h^2\right) dv^2.$$

  If $f = \text{const}$, one has a *helicoid*; if $h = 0$, one has a *rotation surface*.

- **Catalan surface metric**

  The *Catalan surface* is a *minimal surface*, given by the following parametric equations:

$$x_1(u,v) = u - \sin u \cosh v, \quad x_2(u,v) = 1 - \cos u \cosh v,$$
$$x_3(u,v) = 4\sin\left(\frac{u}{2}\right)\sinh\left(\frac{v}{2}\right).$$

  The intrinsic metric on it is given by the *line element*

$$ds^2 = 2\cosh^2\left(\frac{v}{2}\right)(\cosh v - \cos u)\,du^2 + 2\cosh^2\left(\frac{v}{2}\right)\left(\cosh v - \cos u\right)dv^2.$$

- **Monkey saddle metric**

  The *monkey saddle* is a surface, given by the Cartesian equation $x_3 = x_1(x_1^2 - 3x_2^2)$, or by the following parametric equations:

$$x_1(u,v) = u, \quad x_2(u,v) = v, \quad x_3(u,v) = u^3 - 3uv^2.$$

  This is a surface which a monkey can straddle with both legs and his tail. The intrinsic metric on it is given by the *line element*

$$ds^2 = \left(1 + \left(su^2 - 3v^2\right)^2\right)du^2 - 2\left(18uv\left(u^2 - v^2\right)\right)du\,dv + \left(1 + 36u^2v^2\right)dv^2.$$

8.3. DISTANCES ON KNOTS

A *knot* is a closed, non-self-intersecting curve that is embedded in S^3. The *trivial knot* (or *unknot*) O is a closed loop that is not knotted. A knot can be generalized to a link which is a set of disjoint knots. Every link has its *Seifert surface*, i.e., a compact oriented surface with given link as boundary. Two knots (links) are called *equivalent* if one can be smoothly deformed into another. Formally, a link is defined as a smooth one-dimensional *submanifold* of the 3-sphere S^3; a knot is a link consisting of one component; links L_1 and L_2 are called *equivalent* if there exists an orientation-preserving homeomorphism $f : S^3 \to S^3$ such that $f(L_1) = L_2$.

All the information about a knot can be described using a *knot diagram*. It is a projection of a knot onto a plane such that no more than two points of the knot are projected to the same point on the plane, and at each such point it is indicated which strand is closest to the plane, usually by erasing part of the lower strand. Two different knot diagrams may both represent the same knot. Much of knot theory is devoted to telling when two knot diagrams represent the same knot.

An *unknotting operation* is an operation which changes the overcrossing and the undercrossing at a double point of a given knot diagram. The *unknotting number* of a knot K is the minimum number of unknotting operations needed to deform a diagram of K into

D -256- - -* ! - / :! J < 2 "
6 ! 9 3 " @ " D 1! 3 & 1 ! " 1"

-256- - -* ! - / + !0 = ! !
&&&J =") 1"
(B !1 $

Unit-5

Chapter 15

Solution of System of Linear Equation

Introduction

There are many direct methods to solve simultaneous linear equations. In this chapter, we shall discuss Crout's method to solve a system of Linear Equations. Crout's method provides the solution after a computation that is known in advance. This chapter also describes an iterative or indirect method: Gauss-Seidel Method to solve a system of Linear Equations. In Gauss-Seidel Method, the solution starts from an approximation to the true solution and if convergent, we get a sequence of closer solution. The computations are repeated till the required accuracy is obtained. In direct method: Crout's method, amount of computation is fixed, while in iteration method: Gauss-Seidel Method, the amount of computation depends upon the required accuracy of the solution.

15.1 Crout's Method

This method is based upon the fact that every square matrix A can be expressed as the product of a lower triangular matrix (L), an upper triangular matrix (U). Let AX = B is a system of linear equation. So, if a matrix decomposition of a matrix A is such that A = LDU being L is a unit lower triangular matrix, D is a diagonal matrix and U is a unit upper triangular matrix, then Crout's method produces A = (LD)U. After the LU decomposition of A , the solution to the system of linear equations AX = L U X = B is solved by solving the system of linear equations LY = B by forward substitution for Y, and then solving the system of linear equations UX = Y by backward substitution for X. Now consider we want solve the system of linear equation $AX = B$, which can be written as,

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_2 \end{bmatrix} \qquad \text{...(15.1)}$$

Let the cofficient matrix can be written as

$$A = LU$$

where

$$L = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \text{and} \, U = \begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 1 & 1 \end{bmatrix}$$

Then Eqn (15.1) becomes

Let $LUX = B$ ...(15.2)

$$UX = Y \qquad \text{....(15.3)}$$

so that Eqn (15.2) becomes

$$LY = B \qquad \text{....(15.4)}$$

i.e. $$\begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

$$\therefore l_{11}y_1 = b_1;\ \ l_{21}y_1 + l_{22}y_2 = b_2;\ \ l_{31}y_1 + l_{32}y_2 + l_{33}y_3 = b_3 \qquad \text{...(15.5)}$$

By forward substitution, y_1, y_2 and y_3 can be found out if L is known. Now L and U can be found from $LU = A$.

$$\begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

or $$\begin{bmatrix} l_{11} & l_{11}u_{12} & l_{11}u_{13} \\ l_{21} & l_{21}u_{12} + l_{22} & l_{21}u_{13} + l_{22}u_{23} \\ l_{31} & l_{31}u_{12} + l_{32} & l_{31}u_{13} + l_{32}u_{23} + l_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

Equating the corresponding elements on both sides, we get

$$l_{11} = a_{11}, l_{21} = a_{21}, l_{31} = a_{31}$$
$$l_{11}u_{12} = a_{12} \ \therefore u_{12} = a_{12}/a_{11}$$
$$l_{11}u_{13} = a_{13} \ \therefore u_{13} = a_{13}/a_{11}$$
$$l_{21}u_{12} + l_{22} = a_{22} \ \therefore l_{22} = a_{22} - l_{21}u_{12}$$
$$l_{31}u_{12} + l_{32} = a_{32} \ \therefore u_{32} = \{a_{12} - l_{21}u_{13}\}/u_{23}$$
$$l_{31}u_{13} + l_{32}u_{23} + l_{33} = a_{33} \ \therefore l_{33} = a_{33} - l_{31}u_{13} - l_{31}u_{23}$$

Thus all the 12 unknowns are determined and the solution, as show already, is obtained from Eqn (15.5)The whole procedure is illustrated by some suitable examples which are given below.

Illustrative Examples

Ex.1: Solve the system

$$2x + y + 4z = 12, \quad 8x + 3y + 2z = 20, \quad 4x + 11y + z = 33$$

by Crout's method. (*M.U, B.E., 1991*)

Sol : Argumented matrix,

$$[A/B] = \left[\begin{array}{ccc|c} 2 & 1 & 4 & 12 \\ 8 & -3 & 2 & 20 \\ 4 & 11 & -1 & 23 \end{array}\right]$$

Let the derived matrix (DM)

$$= \begin{bmatrix} l_{11} & u_{12} & u_{13} & y_1 \\ l_{21} & l_{22} & u_{23} & y_2 \\ l_{31} & l_{32} & l_{33} & y_3 \end{bmatrix}$$

$$\begin{bmatrix} 1 & -1 & 2 \\ 0 & 1 & -\frac{2}{5} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & y_{12} & y_{13} \\ 0 & 1 & y_{13} \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

We get $y_{12} = 1,\ y_{13} = -8/5;\ y_{23} = 2/5$

$$\therefore \quad L^{-1} = \begin{bmatrix} \frac{1}{2} & 0 & 2 \\ -\frac{1}{5} & \frac{1}{5} & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{and}\, U^{-1} = \begin{bmatrix} 1 & 1 & -\frac{8}{5} \\ 0 & 1 & \frac{2}{5} \\ 0 & 0 & 1 \end{bmatrix}$$

$$A^{-1} = U^{-1}L^{-1} = \begin{bmatrix} -\frac{1}{2} & \frac{1}{5} & \frac{-8}{5} \\ 0 & \frac{1}{5} & \frac{2}{5} \\ \frac{1}{2} & 0 & 1 \end{bmatrix}$$

15.3 Gauss-Seidel Method

In certain cases, when a system of equations is large, iterative methods of solving equations are more advantageous. Elimination methods, such as Gaussian elimination, are prone to large round-off errors for a large set of equations. Iterative methods, such as the Gauss-Seidel method, give the user control of the round-off error. After reading Gauss-Seidel Method, you should be able to solve a set of equations using Gauss-Seidel method, recognize the advantages and pitfalls of Gauss-Seidel method, determine under what conditions Gauss Seidel method always converges. Given a general set of n equations and n unknowns, we have

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots\ldots + a_{1n}x_n &= c_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots\ldots + a_{2n}x_n &= c_2 \\ \cdots \qquad \cdots \qquad &\cdots \\ \cdots \qquad \cdots \qquad &\cdots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \ldots\ldots + a_{nn}x_n &= c_n \end{aligned}$$

If the diagonal elements are non-zero, each equation is rewritten for the corresponding unknown, that is, the first equation is rewritten with x_2 on the left hand side, second equation is rewritten with on the left hand side and so on as follows

$$x_2 = \frac{c_2 - a_{21}x_1 - a_{23}x_3 \ldots\ldots\ldots\ldots - a_{2n}x_n}{a_{22}}$$

$$\vdots$$

$$\vdots$$

$$x_{n-1} = \frac{c_{n-1} - a_{n-1,1}x_1 - a_{n-1,2}x_2 \ldots\ldots\ldots - a_{n-1,n-2}x_{n-2} - a_{n-1,n}x_n}{a_{n-1,n-1}}$$

$$x_n = \frac{c_n - a_{n1}x_1 - a_{n2}x_2 - \ldots\ldots - a_{n,n-1}x_{n-1}}{a_{nn}}$$

These equations can be rewritten in a summation form as

$$x_1 = \frac{c_1 - \sum_{\substack{j=1 \\ j \neq 1}}^{n} a_{1j}x_j}{a_{11}}, \; x_2 = \frac{c_2 - \sum_{\substack{j=1 \\ j \neq 2}}^{n} a_{2j}x_j}{a_{22}}$$

$$x_{n-1} = \frac{c_{n-1} - \sum_{\substack{j=1 \\ j \neq n-1}}^{n} a_{n-1,j}x_j}{a_{n-1,n-1}}, \; x_n = \frac{c_n - \sum_{\substack{j=1 \\ j \neq 2}}^{n} a_{nj}x_j}{a_{nn}}$$

Hence for any row i,

$$x_i = \frac{c_i - \sum_{\substack{j=1 \\ j \neq 1}}^{n} a_{ij}x_j}{a_{ii}}, i = 1, 2, \ldots, n.$$

Now to find x_i's, one assumes an initial guess for the x_i 's and then uses the rewritten equations to calculate the new estimates. Remember, one always uses the most recent estimates to calculate the next estimates x_i. At the end of each iteration, one calculates the absolute relative approximate error for each x_i as

$$|\varepsilon_a|_i = \left| \frac{x_i^{new} - x_i^{old}}{x_i^{new}} \right| \times 100$$

where x_i^{new} is the recently obtained value of x_i, and x_i^{old}s the previous value of x_i. When the absolute relative approximate error for each x_i is less than the pre-specified tolerance, the iterations are stopped.

for the better understanding of Gauss-Seidel's method some examples are given below.

Ex.5: The upward velocity of a rocket is given at three different times in the following table

Time, t is second	Velocity, v m/s
5	106.8
8	177.2
12	279.2

The velocity data is approximated by a polynomial as $v(t) = a_1 t^2 + a_2 t + a_3$, $5 \le t \le 12$.

Find the values of a_1, a_2 and a_3 using Gauss-Seidel Method. Assume an initial guess of the solution as $a_1 = 1, a_2 = 2, a_3 = 5$.

Sol: The polynomial is going through three data points $(t_1, v_1), (t_2, v_2)$, and (t_3, v_3) where from the above table

$$t_1 = 5, v_1 = 106.8$$
$$t_2 = 8, v_2 = 177.2$$
$$t_3 = 12, v_3 = 279.2$$

Requiring that $v(t) = a_1 t^2 + a_2 t + a_3$,passes through the three data points gives

$$v(t_1) = v_1 = a_1 t_1^2 + a_2 t_1 + a_3$$
$$v(t_2) = v_2 = a_1 t_2^2 + a_2 t_2 + a_3$$
$$v(t_3) = v_3 = a_1 t_3^2 + a_2 t_3 + a_3$$

Substituting the data $(t_1, v_1), (t_2, v_2), (t_3, v_3)$ gives

$$a_1(5^2) + a_2(5) + a_3 = 106.8$$
$$a_1(8^2) + a_2(8) + a_3 = 177.2$$
$$a_1(12^2) + a_2(12) + a_3 = 279.2$$

or

$$25a_1 + 5a_2 + a_3 = 106.8$$
$$64a_1 + 8a_2 + a_3 = 177.2$$
$$144a_1 + 12a_2 + a_3 = 279.2$$

The coefficients for the above expression are given by

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

Rewriting the equations gives

$$a_1 = \frac{106.8 - 5a_2 - a_3}{25}$$
$$a_2 = \frac{177.2 - 64a_1 - a_3}{8}$$
$$a_3 = \frac{279.2 - 144a_1 - 12a_2}{1}$$

Iteration #1

Given the initial guess of the solution vector as $a_1 = 1, a_2 = 2, a_3 = 5$

we get

$$a_1 = \frac{106.8 - 5(2) - (5)}{25} = 3.6720$$
$$a_2 = \frac{177.2 - 64(3.6720) - (5)}{8} = -7.8150,$$
$$a_3 = \frac{279.2 - 144(3.6720) - 12(-7.8510)}{1} = -155.36$$

The absolute relative approximate error for each x_i then is

$$|\varepsilon_a|_3 = \left|\frac{3.6720 - 1}{3.6720}\right| \times 100 = 72.76\%$$

$$|\varepsilon_a|_2 = \left|\frac{-7.8510 - 2}{-7.8510}\right| \times 100 = 125.47\%$$

$$|\varepsilon_a|_3 = \left|\frac{155.36 - 5}{-155.36}\right| \times 100 = 103.22\%$$

At the end of the first iteration, the estimate of the solution vector is

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 3.6720 \\ -7.8510 \\ -155.36 \end{bmatrix}$$

and the maximum absolute relative approximate error is 125.47%.

Iteration #2

The estimate of the solution vector at the end of Iteration #1 is

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 3.6720 \\ -7.8510 \\ -155.36 \end{bmatrix}$$

Now we get

$$a_1 = \frac{106.8 - 5(-7.8510) - (-155.36)}{25} = 12.056$$

$$a_2 = \frac{177.2 - 64(12.056) - (-155.36)}{8} = -54.882$$

$$a_3 = \frac{279.2 - 144(12.056) - 12(-54.882)}{1} = -798.34$$

The absolute relative approximate error for each then is

$$|\in_a|_1 = \left|\frac{12.056 - 3.6720}{12.056}\right| \times 100 = 69.543\%$$

$$|\in_a|_2 = \left|\frac{-54.882 - (-7.8510)}{-54.882}\right| \times 100 = 85.695\%$$

$$|\in_a|_3 = \left|\frac{-798.34 - (-155.36)}{-798.34}\right| \times 100 = 80.540\%$$

At the end of the second iteration the estimate of the solution vector is

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 12.056 \\ -54.882 \\ -798.54 \end{bmatrix}$$

and the maximum absolute relative approximate error is 85.695%.

Conducting more iteration gives the following values for the solution vector and the corresponding absolute relative approximate errors.

Iteration	a_1	$\lvert\in_a\rvert_1\%$	a_2	$\lvert\in_a\rvert_2\%$	a_3	$\lvert\in_a\rvert_3\%$
1	3.6720	72.767	-7.8510	125.47	-155.36	103.22
2	12.056	69.543	-54.882	85.695	-798.34	80.540
3	47.182	74.447	-255.51	78.521	-3448.9	76.852
4	193.33	75.595	-1093.4	76.632	-14440	76.116
5	800.53	75.850	-4577.2	76.112	-60072	75.963
6	3322.6	75.906	-19049	75.972	-249580	75.931

As seen in the above table, the solution estimates are not converging to the true solution of $a_1 = 0.29048$, $a_2 = 19.690$, $a_3 = 1.0857$

The above system of equations does not seem to converge? Why?

Well, a pitfall of most iterative methods is that they may or may not converge. However, the solution to a certain classes of systems of simultaneous equations does always converge using the Gauss-Seidel method. This class of system of equations is where the coefficient matrix $[A]$ in $[A][X] = [C]$ is diagonally dominant, that is $|a_{ii}| \geq \sum_{\substack{j=1 \\ j\neq 1}}^{n} |a_{ij}|$ for all, and $|a_{ii}| \geq \sum_{\substack{j=1 \\ j\neq 1}}^{n} |a_{ij}|$ for at least one i. If a system of equations has a coefficient matrix that is not diagonally dominant, it may or may not converge. Fortunately, many physical systems that result in simultaneous linear equations have a diagonally dominant coefficient matrix, which then assures convergence for iterative methods such as the Gauss-Seidel method of solving simultaneous linear equations.

Ex. 6: Find the solution to the following system of equations using the Gauss-Seidel method

$$12x_1 + 3x_2 - 5x_3 = 1$$
$$x_1 + 5x_2 + 3x_3 = 28$$
$$3x_1 + 7x_2 + 13x_3 = 76$$

Use $x_1 = 1, x_2 = 0, x_3 = 1$, as the initial guess and conduct two iterations.

Sol: The coefficient matrix

$$[A] = \begin{bmatrix} 12 & 3 & -5 \\ 1 & 5 & 3 \\ 3 & 7 & 13 \end{bmatrix}$$

is diagonally dominant as

$$|a_{11}| = |12| = 12 \geq |a_{12}| + |a_{13}| = |3| + |-5| = 8$$
$$|a_{22}| = |5| = 5 \geq |a_{21}| + |a_{23}| = |1| + |3| = 4$$
$$|a_{33}| = |13| = 13 \geq |a_{31}| + |a_{32}| = |3| + |7| = 10$$

and the inequality is strictly greater than for at least one row. Hence, the solution should converge using the Gauss-Seidel method.

Rewriting the equations, we get

$$x_1 = \frac{1-3x_2+5x_3}{12},\ x_2 = \frac{28-x_1-3x_3}{5},\ x_3 = \frac{76-3x_1-7x_2}{13}$$

Assuming an initial guess of . $x_1 = 1, x_2 = 0, x_3 = 1.$

Iteration#1:

$$x_1 = \frac{1-3(0)+5(1)}{12} = 0.50000,$$

$$x_2 = \frac{28-(0.50000)-3(1)}{5} = 4.9000$$

$$x_3 = \frac{76-3(0.50000)-7(4.9000)}{13} = 3.0923$$

The absolute relative approximate error at the end of first iteration is

$$|\in_a|_1 = \left|\frac{0.50000-1}{0.50000}\right| \times 100 = 100.000\%$$

$$|\in_a|_2 = \left|\frac{4.9000-0}{4.9000}\right| \times 100 = 100.000\%$$

$$|\in_a|_3 = \left|\frac{3.0923-1}{3.0923}\right| \times 100 = 67.662\%$$

The maximum absolute relative approximate error is 100.000%

Iteration 2:

$$x_1 = \frac{1-3(4.9000)+5(3.0923)}{12} = 0.14679$$

$$x_2 = \frac{28-(0.14679)-3(3.0923)}{12} = 3.7153$$

$$x_3 = \frac{76-3(0.14679)-7(3.7153)}{13} = 3.8118$$

At the end of second iteration, the absolute relative approximate error is

$$|\in_a|_1 = \left|\frac{0.14679-0.50000}{0.14679}\right| \times 100 = 240.61\%$$

$$|\in_a|_2 = \left|\frac{3.7153-4.9000}{3.7153}\right| \times 100 = 31.889\%$$

$$|\in_a|_3 = \left|\frac{3.8118-3.0923}{3.8118}\right| \times 100 = 18.874\%$$

The maximum absolute relative approximate error is 240.61%. This is greater than the value of 100.00% we obtained in the first iteration. Is the solution diverging? No, as you conduct more iteration, the solution converges as follows.

Iteration	a_1	$\|\in_a\|_1\%$	a_2	$\|\in_a\|_2\%$	a_3	$\|\in_a\|_3\%$
1	0.50000	100.00	4.9000	100.00	3.0923	67.662
2	0.14679	240.61	3.7153	31.889	3.8118	18.874
3	0.74275	80.236	3.1644	17.408	3.9708	4.0064
4	0.94675	21.546	3.0281	4.4996	3.9971	0.65772
5	0.99177	4.5391	3.0034	0.82499	4.0001	0.074383
6	0.99919	0.74307	3.0001	0.10856	4.0001	0.00101

This is close to the exact solution $x_1 = 1, x_2 = 3, x_3 = 4$

Ex.7: For a system of equations

$$\begin{aligned} 3x_1 + 7x_2 + 13x_3 &= 76 \\ x_1 + 5x_2 + 3x_3 &= 28 \\ 12x_1 + 3x_2 - 5x_3 &= 1 \end{aligned}$$

find the solution using Gauss-Seidel method. Use $[x_1, x_2, x_3] = [1 \ \ 0 \ 1]$ as the initial guess.

Sol: Rewriting the equations, we get

$$x_1 = \frac{76 - 7x_2 - 13x_3}{3}, x_2 = \frac{28 - x_1 - 3x_3}{5}, \ x_3 = \frac{1 - 12x_1 - 3x_2}{-5}$$

Assuming an initial guess of $x_1 = 1, x_2 = 0, x_3 = 1$,the next six iterative values are given in the table below

Iteration	a_1	$\|\in_a\|_1\%$	a_2	$\|\in_a\|_2\%$	a_3	$\|\in_a\|_3\%$
1	21.000	95.238	0.80000	100.00	50.680	98.027
2	-196.15	110.71	14.421	94.453	-462.30	110.96
3	1995.0	109.83	-116.02	112.43	4718.1	109.80
4	-20149	109.90	1204.6	109.63	-47636	109.90
5	2.0364×10^5	109.89	-12140	109.92	4.8144105	109.89
6	-2.0579×10^6	109.89	1.2272×10^5	109.89	-4.8653×10^6	109.89

You can see that this solution is not converging and the coefficient matrix is not diagonally dominant. The coefficient matrix

$$[A] = \begin{bmatrix} 3 & 7 & 13 \\ 1 & 5 & 3 \\ 12 & 3 & -5 \end{bmatrix}$$

is not diagonally dominant as $|a_{11}| = |3| = 3 \le |a_{12}| + |a_{13}| = |7| + |13| = 20$ Hence, the Gauss-Seidel method may or may not converge. However, it is the same set of equations as the previous example and that converged. The only difference is that we exchanged first and the third equation with each other and that made the coefficient matrix not diagonally dominant. Therefore, it is possible that a system of equations can be made diagonally dominant if one exchanges the equations with each other. However, it is not possible for all cases. For example, the following set of equations.

$$x_1 + x_2 + x_3 = 3$$
$$2x_1 + 3x_2 + 4x_3 = 9$$
$$x_1 + 7x_2 + x_3 = 9$$

cannot be rewritten to make the coefficient matrix diagonally dominant.

Ex. 8 : Solve the system of Equation by Gauss - Seidel's Method.

$$8x - 3y + 2z = 20$$
$$6x + 3y + 12z = 35$$
$$4x + 11y - z = 33$$

Sol. From the given equations, we have

$$x = \frac{1}{8}(20 + 3y - 2z) \qquad \text{....(i)}$$

$$y = \frac{1}{11}(33 - 4x + z) \qquad \text{....(ii)}$$

$$z = \frac{1}{12}(35 - 6x - 3y) \qquad \text{....(iii)}$$

Putting $y = 0,\ z = 0$ in RHS of (i), we get $x = \frac{20}{8} = 2.5$

Putting $x = 2.5$, $z = 0$ in RHS of (ii), we get

$$y = \frac{1}{11}[33 - 4(2.5)] = 2.0909091$$

Putting $x = 2.5$, $y = 2.0909091$in RHS of (iii), we get

$$z_1 = \frac{1}{12}[35 - 6(2.5) - 3(2.0909091)] = 1.1439394$$

For the *second approximation*,

$$x_2 = \frac{1}{8}[20 + 3y_1 - 2z_1]$$
$$= \frac{1}{8}[20 + 3(2.0909091) - 2(1.1439394)] = 2.9981061$$

$$y_2 = \frac{1}{11}[33 - 4x_2 + z_1]$$
$$= \frac{1}{11}[33 - 4(2.9981061) + 1.1439394] = 2.0137741$$

$$z_2 = \frac{1}{12}[35 - 6x_2 - 3y_2]$$
$$= \frac{1}{12}[35 - 6(2.9981061) - 3(2.0137741)] = 0.9141701$$

Third approximation :

$$x_3 = \frac{1}{8}[20 + 3(2.0137741) - 2(0.9141701)] = 3.0266228$$

$$y_3 = \frac{1}{11}[33-4(3.0266228+0.9141701)] = 1.9825163$$

$$z_3 = \frac{1}{12}[35-6(3.0266228)-3(1.9825163)] = 0.9077262$$

Fourth approximation :

$$x_4 = \frac{1}{8}[20-3(1.9825163)-2(0.9077262)] = 3.0165121$$

$$y_4 = \frac{1}{11}[33-4(3.0165121)+(0.9077262)] = 1.9856071$$

$$z_4 = \frac{1}{12}[35-6(3.0165121)-3(1.9856071)] = 0.9120088$$

Fifth approximation :

$$x_5 = \frac{1}{8}[20+3(1.9856071)-2(0.9120088)] = 3.0166005$$

$$y_5 = \frac{1}{11}[33-4(3.0166005)+0.9120088] = 1.9859643$$

$$z_5 = \frac{1}{12}[35-6(3.0166005)-3(1.9859643)] = 0.9118753$$

Sixth approximation :

$$x_6 = \frac{1}{8}[20+3(1.9859643)-2(0.9118753)] = 3.0167678$$

$$y_6 = \frac{1}{11}[33-4(3.0167678)+0.9118753] = 1.9858913$$

$$z_6 = \frac{1}{12}[35-6(3.0167678)-3(1.9858913)] = 0.9118099$$

Seventh approximation :

$$x_7 = \frac{1}{8}[20+3(1.9858913)-2(0.9118099)] = 3.0167568$$

$$y_7 = \frac{1}{11}[33-4(3.0167568)+0.9118099] = 1.9858894$$

$$z_7 = \frac{1}{12}[35-6(3.0167568)-3(1.9858894)] = 1.9118159$$

Since at the sixth and seventh approximations, the values of x,y,z are the same, correct four decimal places, we can stop the iteration process.

$\therefore\ x = 3.0167, \quad y = 1.9858, \ z = 0.9118$

We find that 12 iterations are needed in Gauss–Jacobi method to get the same accuracy as achieved by 7 iterations in Gauss–Seidel method.

Ex. 9: Solve by Gauss–Seidel method, the following system of equations

$$28x + 4y - z = 32, x + 3y + 10z = 24,$$

and $2x + 17y + 4z = 35$ (*B.U,1997, M.U, 1991*)

Sol : The coefficient matrix of the given system is not diagonally dominant. Hence, we rearrange the equations as follows, such that the elements in the coefficient matrix are diagonally dominant.

$$28x + 4y - z = 32$$

$$2x + 17y + 4z = 35$$

$$x + 3y + 10z = 24$$

Hence we can apply Gauss–Seidel iteration method.

From the above equations

$$x = \frac{1}{28}[32 - 4y + z] \qquad \text{....(i)}$$

$$y = \frac{1}{17}[35 - 2x - 4z] \qquad \text{....(ii)}$$

$$z = \frac{1}{10}[24 - x - 3y] \qquad \text{....(iii)}$$

First approximation

Putting $y = z = 0$ in (i), we get

$$x_1 = \frac{1}{28}(32) = 1.1428571$$

Putting $x = 1.1428571,\ z = 0$ in (ii), we get

$$y_1 = \frac{1}{17}[35 - 2(1.1428571)] = 1.9243697$$

Putting $x = 1.1428571,\ y = 1.9243697$ in (iii), we get

$$z_1 = \frac{1}{10}[24 - 1.1428571 - 3(1.9243697)] = 1.7084034$$

Second approximation

$$x_2 = \frac{1}{28}[32 - 4(1.9243697) + 1.7084034] = 0.9289615$$

$$y_2 = \frac{1}{17}[35 - 2(0.9289615) - 4(1.7084034)] = 1.5475567$$

$$z_2 = \frac{1}{10}[24 - 0.9289615 - 3(1.5475567)] = 1.8428368$$

Third approximation

$$x_3 = \frac{1}{28}[32 - 4(1.5475667) + 1.8428368] = 0.9875932$$

$$y_3 = \frac{1}{17}[35 - 2(0.9875932) - 4(1.8428368)] = 1.5090274$$

$$z_3 = \frac{1}{10}[24 - 0.9875932 - 3(1.5090274)] = 1.8485325$$

Fourth approximation

$$x_4 = \frac{1}{28}[32 - 4(1.5090274) + 1.8485325] = 0.9933008$$

$$y_4 = \frac{1}{17}[35 - 2(0.9933008)] = 4(1.8485325)] = 1.5070158$$

$$z_4 = \frac{1}{10}[24 - 0.9933008 - 3(1.5070158)] = 1.8485625$$

Fifth approximation

$$x_5 = \frac{1}{28}[32 - 4(1.5070158) + 1.8485625] = 0.9935893$$

$$y_5 = \frac{1}{17}[35 - 2(0.9935893) - 4(1.8485652)] = 1.5069741$$

$$z_5 = \frac{1}{10}[24 - 0.9935893 - 3(1.5069741)] = 1.8485488$$

Sixth approximation

$$x_6 = \frac{1}{28}[32 - 4(1.5069741) + 1.8485488] = 0.9935947$$

$$y_6 = \frac{1}{17}[35 - 2(0.9935947) - 4(1.8485488)] = 1.5069774$$

$$z_6 = \frac{1}{10}[24 - 0.9935947 - 3(1.5069774)] = 1.8485473$$

$\because$ The values of x, y, z in the fourth and fifth iteration are same upto four decimals, we stop the process here.

Hence $\quad x = 0.9936,\ y = 1.5069,\ z = 1.8485$

Ex.10: Using Gauss–Seidel iteration method, solve the system of equations.

$$10x - 2y - z - w = 3; -2x + 10y - z - w = 15;$$

$$-x - y + 10z - 2w = 27; -x - y - 2z + 10w = -9$$

(M.U, B.E., 1987)

Sol : The coefficient matrix of the given system is diagonally dominant. Hence we can apply Gauss–Seidel iteration method.

From the given equations, we can write

$$x = \frac{1}{10}[3 + 2y + z + w] \quad \text{....(i)}$$

$$y = \frac{1}{10}[15 + 2x + z + w] \quad \text{....(ii)}$$

$$z = \frac{1}{10}[27 + x + y + 2w] \quad \text{....(iii)}$$

$$w = \frac{1}{10}[-9 + x + y + 2z] \quad \text{....(iv)}$$

First approximation :

Putting $y = z = w = 0$ on RHS of (i), we get $x_1 = \frac{3}{10} = 0.3$

Putting $x = 0.3,\ z = w = 0$ on RHS of (ii),

we get $y_1 = \frac{1}{10}[15 + 2(0.3)] = 1.56$

Putting $x = 0.3, y = 1.56, w = 0$ on RHS of (iii),

we get $z_1 = \frac{1}{10}[27 + 0.3 + 1.56,] = 2.886$

Putting $x = 0.3,\ y = 1.56,\ z = 2.886$ on RHS of (iv),

we get $w_1 = \frac{1}{10}[-9 + 0.3 + 1.56 + 2(2.886)] = -0.1368$

Second approximation :

$$x_2 = \frac{1}{10}[3 + 2(1.56)] + 2.886 - 0.1368] = 0.88692$$

$$y_2 = \frac{1}{10}[15 + 2(0.88692) + 2.886 - 0.1368] = 1.952304$$

$$z_2 = \frac{1}{10}[27 + 0.88692 + 1.952304 + 2(-0.1368)] = 2.9565624$$

$$w_2 = \frac{1}{10}[-9 + 0.88692 + 1.952304 + 2(2.9565624)] = -0.0247651$$

Third approximation :

$$x_3 = \frac{1}{10}[3 + 2(1.952304) + 2.9565624 - 0.0247651] = 0.9836405$$

$$y_3 = \frac{1}{10}[15 + 2(0.9836405) + 2.9565624 - 0.0247651] = 1.9899087$$

$$z_3 = \frac{1}{10}[27 + 0.9836405 + 1.9899087 + 2(-0.0247651)] = 2.9924019$$

$$w_3 = \frac{1}{10}[-9 + 0.983405 + 1.9899087 + 2(2.9924019)] = -0.0041647$$

Fourth approximation :

$$x_4 = \frac{1}{10}[3 + 2(1.9899087) + 2.9924019 - 0.0041647] = 0.9968054$$

$$y_4 = \frac{1}{10}[15 + 2(0.9968054) + 2.9924019 - 0.0041647] = 1.9981848$$

$$z_4 = \frac{1}{10}[27 + 0.9968054 + 1.9981848 + 2(-0.0041647)] = 2.9986661$$

$$w_4 = \frac{1}{10}[-9 + 0.9968054 + 1.9981848 + 2(2.9986661)] = -0.0007677$$

Fifth approximation :

$$x_5 = \frac{1}{10}[3 + 2(1.9981848) + 2.9986661 - 0.0007677] = 0.9994268$$

$$y_5 = \frac{1}{10}[15 + 2(0.9994268) + 2.998666 - 0.0007677] = 1.9996752$$

$$z_5 = \frac{1}{10}[27 + 0.9994268 + 1.9996752 + 2(-0.0007677)] = 2.9997567$$

$$w_5 = \frac{1}{10}[-9 + 0.9994268 + 1.9996752 + 2(2.9997567)] = -0.0001384$$

Sixth approximation :

$$x_6 = \frac{1}{10}[3 + 2(1.9996752) + 2.9997567 - 0.0001384] = 0.9998968$$

$$y_6 = \frac{1}{10}[15 + 2(0.9998968) + 2.9997567 - 0.0001384] = 1.9999412$$

$$z_6 = \frac{1}{10}[27 + 0.9998968 + 1.9999412 + 2(-0.0001384)] = 2.9999561$$

$$w_6 = \frac{1}{10}[-9 + 0.9998968 + 1.9999412 + 2(2.9999561)] = -0.0002498$$

Seventh approximation :

$$x_7 = \frac{1}{10}[3 + 2(1.9999412) + 2.9999561 - 0.0002498] = 0.9999588$$

$$y_7 = \frac{1}{10}[15 + 2(0.9999588) + 2.9999561 - 0.0002498] = 1.9999624$$

$$z_7 = \frac{1}{10}[27 + 0.9999588 + 1.9999264 + 2(-0.0002498)] = 2.9999422$$

$$w_7 = \frac{1}{10}[-9 + 0.9999588 + 1.9999624 + 2(2.9999422)] = -0.0001945$$

Now, from sixth and seventh approximations the values of x, y, z and w correct to four decimal places are

$$x = 0.9999 \quad y = 1.9999 \quad z = 2.9999 \quad w = 0.0002$$

Exercises

Objective Type Questions

Multiple Choice Questions

Tick the correct answer :

1. Crout's method is also known as :

(a) direct method (b) iterative method

(c) Indirect method (d) rotational method.

2. Crout's method is suited for :

(a) computer operation (b) manipulating the data

(c) everywhere (d) none of these.

3. Crout's method can be applied to get :

(a) inverse of a Matrix (b) transpose of a Matrix

(c) the value of a determinant (d) no where in numerical problems.

4. Gauss-Seidel's method is an :

(a) direct method (b) iterative method

(c) indirect method (d) none of these.

5. Gauss-Seidel's method is used to solve :

(a) a system of linear equation (b) a system of non linear equation

(c) a system of algebsic equation (d) none of these.

Fill in the Blanks

1. In Gauss-Seidel method the $(K+1)^{th}$ iterations are given as below,

$x_{k+1} = \dots\dots\dots$

$y_{k+1} = \dots\dots\dots$

and $z_{k+1} = \dots\dots\dots$

where $a_1, b_1, c_1, a_2, b_2, c_2, a_3, b_3, c_3, d_1, b_2$ and d_3 are taken from

$$\begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix}\begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix}$$

2. If x_0, y_0, z_0 are approximation, then the next approximations x_1, y_1, z_1 are given as below,

"   -256-   -   -*   ! - /
!   !   ! I "   I   I   A!K .3 1"
(%' $/   ' !L   > !   7!   ?!

4 Maxwell equations

Previously, electrical vibrations in a wire were considered. In a general case, ***electromagnetic waves*** are considered[25]. The electromagnetic field is characterized by four vector functions that are ***electric field strength*** **E**, ***magnetic field strength*** **H**, ***electric induction*** **D** and ***magnetic induction*** **B**. Each of these values has three components.

The mathematical models of electrodynamics are based on ***Maxwell equations*** connecting the above values. These include the ***Gauss law***

$$\mathrm{div}\mathbf{D} = 4\pi\rho,$$

according to which an electric charge is a source of electric induction. Here ρ is the charge density, and the divergence div of the vector $\mathbf{D}= \ (D_1, D_2, D_3)$ is characterized by the equality[26]

$$\mathrm{div}\mathbf{D} = \frac{\partial D_1}{\partial x} + \frac{\partial D_2}{\partial y} + \frac{\partial D_3}{\partial z}.$$

Then we have the ***Gauss law of magnetic field***

$$\mathrm{div}\mathbf{B} = 0,$$

postulating the absence of magnetic charges. By the ***Faraday law of induction***

$$\mathrm{rot}\mathbf{E} = -\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t}.$$

the change in magnetic induction generates a vortex electric field, where c is the ***velocity of light*** in vacuum, and the ***rotor*** rot of the vector $\mathbf{E} = (E_1, E_2, E_3)$ is characterized by the equality

$$\mathrm{rot}\mathbf{E} = \Big(\frac{\partial E_3}{\partial y} - \frac{\partial E_2}{\partial z}, \frac{\partial E_1}{\partial z} - \frac{\partial E_3}{\partial x}, \frac{\partial E_2}{\partial x} - \frac{\partial E_1}{\partial y}\Big).$$

Finally, the equality

$$\mathrm{rot}\mathbf{H} = \frac{1}{c}\frac{\partial \mathbf{D}}{\partial t} + \frac{4\pi}{c}\mathbf{j},$$

means that electric current and a change in electric induction generate a vortex magnetic field, where **j** is the ***density of the electric current***. The above relations are supplemented by ***material field equations***

$$\mathbf{D} = \varepsilon\mathbf{E}, \ \ \mathbf{B} = \mu\mathbf{H}, \ \ \mathbf{j} = \sigma\mathbf{E}, \tag{12.32}$$

where ε is the ***dielectric constant***, μ is the ***magnetic permeability***, and σ is the ***conductivity of the medium***.

Under certain conditions, the Maxwell equations can be simplified[27]. In particular, in a vacuum in the absence of charges and currents for $\varepsilon = 1$, $\mu = 1$, after excluding the induction of electric and magnetic fields using equalities (12.32), we obtain the following relations

$$\mathrm{div}\mathbf{E} = 0, \ \ \mathrm{div}\mathbf{H} = 0, \ \ \mathrm{rot}\mathbf{E} = -\frac{1}{c}\frac{\partial \mathbf{H}}{\partial t}, \ \ \mathrm{rot}\mathbf{H} = \frac{1}{c}\frac{\partial \mathbf{E}}{\partial t}. \tag{12.33}$$

Acting on the third equality (12.33) by the operator rot and taking into account the possibility of its permutation with the time derivative, we have

$$\mathrm{rot}\,\mathrm{rot}\mathbf{E} = -\frac{1}{c}\frac{\partial}{\partial t}\mathrm{rot}\mathbf{H} = -\frac{1}{c^2}\frac{\partial^2 \mathbf{E}}{\partial t^2} \tag{12.34}$$

by fourth equality (12.33).

It can be shown that for any vector function **E**, the following equality holds

$$\operatorname{rot}\operatorname{rot}\mathbf{E} = \nabla \operatorname{div}\mathbf{E} - \Delta\mathbf{E},$$

where ∇ is the ***gradient***, i.e., the vector whose components are partial derivatives with respect to spatial variables, and Δ is the Laplace operator, i.e., sum of second order derivatives with respect to spatial variables. Then, taking into account the first equality (12.33), we transform the relation (12.34) to the following form

$$\mathbf{E}_{tt} = c^2\Delta\mathbf{E}. \tag{12.35}$$

Analogically, from the equalities (12.33) it follows

$$\mathbf{H}_{tt} = c^2\Delta\mathbf{H}. \tag{12.36}$$

Note that the strength of electric and magnetic fields are vector quantities. Thus, under the assumptions made, it turns out that all its components satisfy the wave equation (12.31), and the velocity of light acts as the parameter a here. In Section 3, it was noted that this coefficient is associated with the wave velocity. Thus, electromagnetic waves propagate at the light velocity.

5 Finite difference method for the vibrating string equation

As already noted, an analytical solution of differential equations can be found only in exceptional cases. In the previous chapter, the finite difference method for the approximate solving of the heat equation was described. It can also be used to solve boundary value problems for the vibrating string equation[28].

Considered the non-homogeneous vibrating string equation

$$u_{tt}(x,t) = a^2u_{xx}(x,t) + f(x,t),\ \ 0 < x < L,\ 0 < t < T \tag{12.37}$$

with initial conditions

$$u(x,0) = \varphi(x),\ u_t(x,0) = \psi(x),\ 0 < x < L \tag{12.38}$$

and boundary conditions

$$u(0,t) = \alpha(t),\ u(L,t) = \beta(t),\ t > 0, \tag{12.39}$$

where the numbers a and L and the functions f, φ, ψ, α, and β are known.

The interval $[0,T]$ is divided into parts with a step $\tau = T/N$ by the points $t_j = \tau j$, $j = 0,...,N$, and the interval $[0,L]$ is divided with a step $h = L/M$ by the points $x_i = hi$, $i = 0,...,M$. A solution to the problem is sought at the grid nodes (x_i,t_j). The second derivatives in equation (12.37) are approximated as follows

$$\frac{\partial^2u(x_i,t_j)}{\partial t^2} \approx \frac{u(x_i,t_{j-1}) - 2u(x_i,t_j) + u(x_i,t_{j+1})}{\tau^2},$$

$$\frac{\partial^2u(x_i,t_j)}{\partial x^2} \approx \frac{u(x_{i-1},t_j) - 2u(x_i,t_j) + u(x_{i+1},t_j)}{h^2}.$$

Then we obtain the equality

$$\frac{u_i^{j-1} - 2u_i^j + u_i^{j+1}}{\tau^2} = a^2\frac{u_{i-1}^j - 2u_i^j + u_{i+1}^j}{h^2} + f_i^j,$$

-256- - -* ! - / 7 3 ? J 2 ! J ! "
9 4 ! 1! 1"
(' !1 %

CHAPTER 3

Areas by Slicing and Scaling

The most fruitful and powerful ideas in the study of areas and volumes are *slicing* and *scaling*. The Greeks used these ideas in finding the fundamental shape that must be taken by formulas for lengths, areas and volumes.

Archimedes showed that the formulas for the circumference and area of a circle and the surface area and volume of a sphere are related by constants involving π. His ideas eventually led to the notion of integration in calculus.

It is easy to see that the results Archimedes obtained are, in many cases, easy consequences of the first results on integration learned in first-semester calculus. The point is, however, that calculus as a body of knowledge did not exist at the time of Archimedes and that the extremely clever arguments used by Archimedes were the arguments that eventually led to calculus.

Wallis used basic calculations of calculus to discover a beautiful product formula for π. Fourier used Wallis's formula for π in solving a special case of the heat equation which he had developed. And this solution led Fourier to what are now called Fourier series. This chapter will relate those discoveries.

3.1. Slicing and Scaling

The principles of slicing and scaling are most easily described in a coordinate plane or coordinate 3-space, though versions were developed before Euclidean coordinates became common and standard. We shall take the principles of slicing and scaling as given. Here is a description.

Slicing principle: Two (reasonably nice) objects A and B in the plane have the same area if each of the horizontal slices in A and B have the same length. See Figure 1. In 3-space, objects A and B have the same volume if each of the horizontal slices in A and B have the same area.

We saw in the chapter on the Pythagorean Theorem how the slicing principle was proved by Euclid in the case of two parallelograms. A fairly rigorous proof for planar sets can be adapted from this Euclidean theorem. This slicing principle is also at the basis of the area and volume formulas in first-year calculus. An advanced version for quite general sets requires Lebesgue measure and is called Fubini's Theorem.

Scaling principle: Let A be a reasonably nice object in 2- or 3-space and k a positive number. Then $k \cdot A$ is said to be a *scaled version* of A with *scale factor* k. If B is congruent to a scaled version kA of A, then A and B are said to be *similar*. The *scaling principle* says that lengths ℓ in A correspond to length $k\ell$ in B. Areas a in A correspond to areas k^2a in B. Volumes v in A correspond to volumes k^3v in B.

The scaling principle is obvious for the interval, the square, and the cube. See Figure 2. For a nice object A, lengths, areas, and volumes are nicely approximated

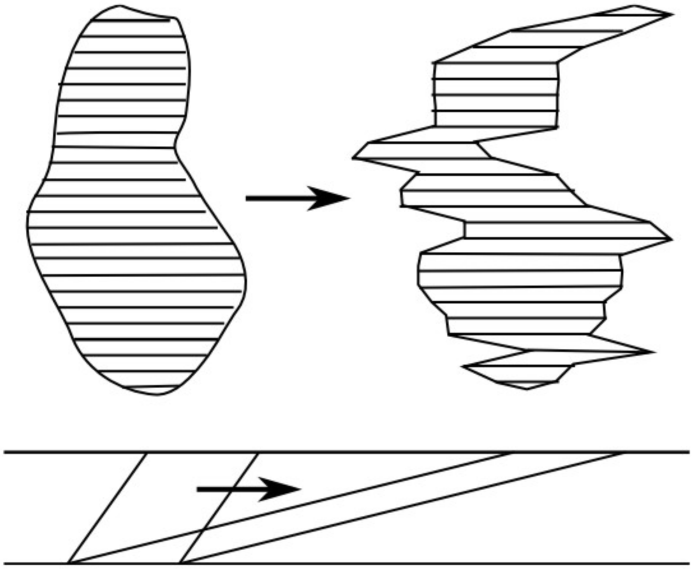

FIGURE 1. The slicing principle

by intervals, squares, and cubes, so that the truth of the scaling principle is apparent. Figure 2 illustrates the principle for the interval, square, and cube with scale factor $k = 2$.

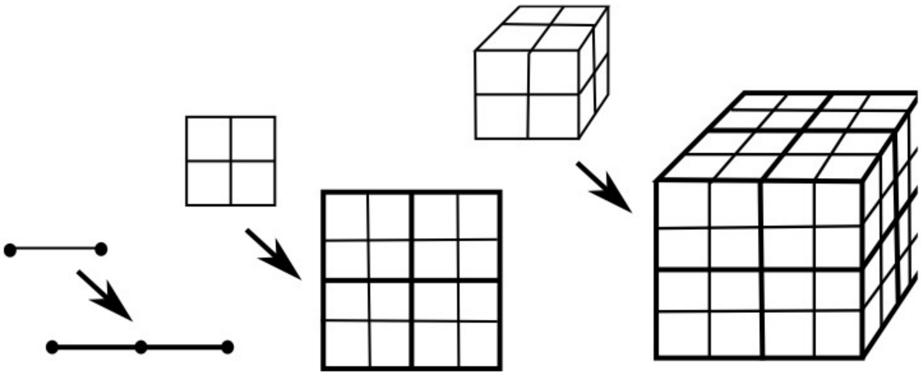

FIGURE 2. The scaling principle again

We need to make a careful distinction between the circle, which is 1-dimensional (a curve) and is called a simple closed curve, and the union of a planar circle with its interior, which is 2-dimensional and which we call a circular 2-dimensional disk or ball. Likewise, we distinguish between the 1-dimensional boundary of a triangle and the 2-dimensional set that is the union of that 1-dimensional set and its interior, which is called a disk or triangular disk. Likewise, we distinguish between the 2-dimensional sphere that is the boundary of a 3-dimensional (solid) ball. The former is called a 2-sphere, the latter a 3-ball or 3-disk.

For our purposes, we used the slicing principle in the Euclidean proof of the Pythagorean Theorem and the scaling principle in the Einstein proof. We now observe that the scaling principle yields the basic form of the standard classical formulas for the circumference of a circle and area of a circular disk, and for the surface area of a 2-sphere and volume of the 3-ball. (Compare Dunham's treatment of Archimedes in [23].)

Theorem 3.1. *Let $C(r)$ denote the circumference of a circle of radius r, and let $A(r)$ denote the area of the circular disk of radius r. Let $S(r)$ denote the surface area of the sphere of radius r, and let $V(r)$ denote the volume of the round 3-ball of radius r. Then*

$$C(r) = r \cdot C(1); \quad A(r) = r^2 \cdot A(1); \quad S(r) = r^2 \cdot S(1); \quad \textit{and } V(r) = r^3 \cdot V(1).$$

Proof. The circle of radius r is a scaled version of the unit circle with scale factor r. The 2-disk of radius r is a scaled version of the unit 2-disk with scale factor r. The 2-sphere is a scaled version of the unit sphere with scale factor r. The 3-ball of radius r is a scaled version of the unit ball with scale factor r. Hence the theorem follows directly from the scaling principle. ☐

3.2. What Is π?

Of course, we all know that $C(1)$, $A(1)$, $S(1)$, and $V(1)$ are closely related: $C(1) = 2\pi$, $A(1) = \pi$, $S(1) = 4\pi$, and $V(1) = (4/3)\pi$. But what is π? and how do we find these basic formulas?

The beginning is easy:

Definition 3.2. π is *by definition* the area $\pi = A(1)$ of the unit 2-disk.

We shall see next, from the slicing principle, that $C(1) = 2 \cdot A(1)$ and that $S(1) = 3 \cdot V(1)$. It remains then only to find the volume $V(1) = (4/3)\pi$ of the unit ball. Given first-year calculus, that is an easy calculation. But we need to remember that Archimedes, who discovered the formula, did not have calculus. His wonderful derivation—extremely clever—suggested ideas that eventually *led* to calculus. The story of his discovery, which I learned from Georg Pólya's *Mathematics and Plausible Reasoning*, is one of my favorite stories. Remember that integration had not as yet been invented.

But first we apply the slicing principle:

Theorem 3.3. *The area of a triangular disk is $(1/2)ab$, where a is the altitude and b is the base.*

Proof. This fact is essentially a consequence of the slicing principle for parallelograms. If the parallelogram is right angled, then the area of the parallelogram is obviously ab, where a is the altitude and b is the base. If the parallelogram is not right angled, then the same result is a consequence of the slicing principle. The triangle is half of the appropriate parallelogram. ☐

Theorem 3.4. *The volume of a 3-dimensional triangular cone is $(1/3)ab$, where a is the altitude and b is the area of the base.*

*D

Proof. This fact is essentially a consequence of the slicing principle for triangular pyramids:

If the pyramid is right angled, then the volume is obviously ab, where a is the altitude and b is the base. If the pyramid is not right angled, then the same result is a consequence of the slicing principle.

It therefore suffices to show that the triangular cone is $1/3$ the volume of the triangular pyramid. This fact requires two further applications of the slicing principle. See Figure 3.

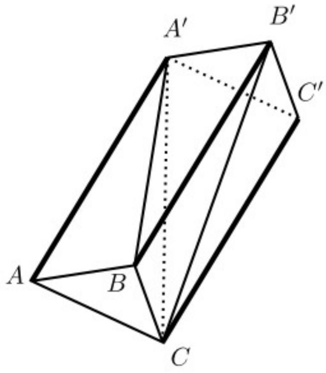

Figure 3. The Triangular Pyramid Cut in Thirds

We may assume that the triangular pyramid has parallel ends ABC and $A'B'C'$. Then the pyramid is the union of three triangular cones, namely $ABCA'$, $BCA'B'$, and $CA'B'C'$. The first two share a triangular base BCA' with altitudes determined by A and B', respectively, at the same distance from the common base. Hence their volumes are equal by the slicing principle. Likewise, $BCA'B'$ and $CA'B'C'$ have the same volume. This completes the proof that the triangular cone is $1/3$ the volume of the triangular pyramid. ☐

Theorem 3.5. *$C(1) = 2A(1) = 2\pi$, so that $C(r) = 2\pi r$ and $A(r) = \pi r^2$.*

Proof. See Figure 4. Approximate the circle of radius 1 closely by a polygonal path made up of segments $S_1, S_2, \ldots, S_n$ of lengths $\ell_1, \ell_2, \ldots, \ell_n$. Let x denote the center of the circle. Then the circular disk is closely approximated by the triangles $xS_1, xS_2, \ldots, xS_n$. The altitude of each triangle is approximately 1, so that the area of the triangle xS_i is approximately $(1/2) \cdot 1 \cdot \ell_i$. Then

$$A(1) \approx (1/2)(\ell_1 + \ell_2 + \cdots + \ell_n) \approx (1/2)C(1).$$

☐

Theorem 3.6. *$S(1) = 3V(1)$.*

Proof. See Figure 5. Approximate the surface of the ball of radius 1 closely by a polyhedral surface made up of triangles $T_1, T_2, \ldots, T_n$ of areas $a_1, a_2, \ldots, a_n$. Let x denote the center of the ball. Then the ball is closely approximated by the triangular cones $xT_1, xT_2, \ldots, xT_n$. The altitude of each cone is approximately 1, so that the volume of the triangular cone xT_i is approximately $(1/3) \cdot 1 \cdot a_i$. Then

$$V(1) \approx (1/3)(a_1 + a_2 + \cdots + a_n) \approx (1/3)S(1).$$

☐

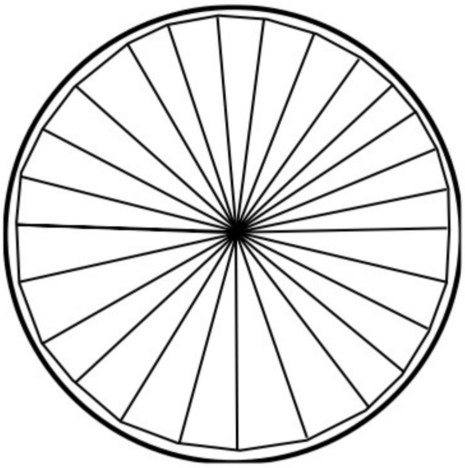

FIGURE 4. The disk as a cone over its center

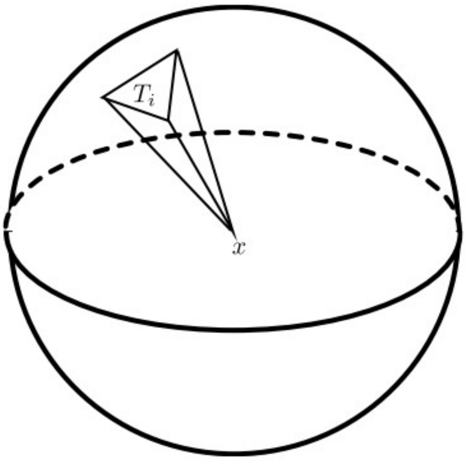

FIGURE 5. The ball as a cone over its center

3.3. Archimedes Discovers the Volume of a Sphere

See [10].

Archimedes knew:

(1) The Pythagorean Theorem, which gives the following equation for the circle in the plane centered at $(1, 0)$ of radius 1: $(x - 1)^2 + (y - 0)^2 = 1^2$, or

$$x^2 + y^2 = 2x.$$

(2) The area of a circle: $A(r) = \pi r^2$.

(2) The volume of a cylinder: $V_{cyl} = a \cdot b$, where a is the altitude and b is the area of the base.

(3) The volume of a cone: $V_{cone} = (1/3)a \cdot b$, where a is the altitude and b is the area of the base. Archimedes also realized that the formula was derived by

examining cross sections in the cone. (Was this fact originally discovered by a form of the slicing principle?).

(4) The principle of the balance: See Figure 6. Suppose weight w_1 is suspended at distance d_1 from the center of the balance and that weight w_2 is suspended at distance d_2 from the center on the opposite side of the balance. Then the condition for balance is that $d_1 \cdot w_1 = d_2 \cdot w_2$.

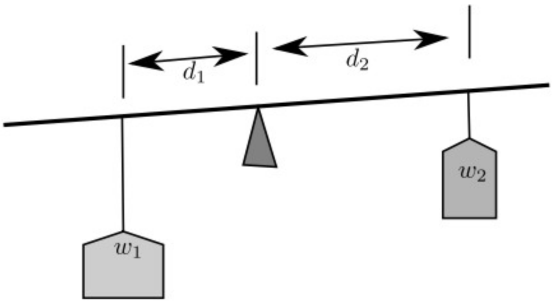

FIGURE 6. The principle of the balance

Archimedes derived the principle of the balance from the principle of the center of mass, which he had found experimentally:

(5) The principle of the center of mass: See Figure 7. Starting in a position of balance, any individual portion of the weights may be rearranged symmetrically about its own individual center of mass. [Thus, if we have dominos stacked on the balance arm in a position of balance, then we may move any domino to the left n units if we also move another domino to the right n units. For example, in the figure, we have 9 dominos stacked at the center, obviously balancing by symmetry. It is only the total number of units moved in each direction that matters. Thus 6 left, one unit each, balances 3 right, two units each. It is by such arguments that Archimedes developed the principle of the balance.]

It is great fun to make a simple balance at home and experiment with stacks of dominos on the balance. We don't need to make the balance perfectly symmetric. After all, even commercial balances recognize the necessity of adjustments. Just move the original stacks of dominos on our imperfect balance to a position of balance and then start moving dominos, a pair at a time, one left a unit, one right a unit. The resulting positions that balance often look totally unlikely.

Now what did Archimedes set out to do? He identifies volume with weight. He considers the solid ball B of radius 1, and the right circular cone C and cylinder D with base radius 2 and height (altitude) 2. By an unbelievably clever argument he shows that B and C suspended at distance 2 from the center of the balance on one arm of the balance should balance the cylinder D suspended at distance 1 from the center on the other arm of the balance. See Figure 8.

In other words, by the principle of the balance,

$$2 \cdot (\mathrm{Vol}_{sph} + \mathrm{Vol}_{cone}) = 1 \cdot \mathrm{Vol}_{cyl}\,.$$

Since he knows that $\mathrm{Vol}_{cone} = (1/3)\,\mathrm{Vol}_{cyl} = (1/3)(\pi \cdot 2^2 \cdot 2)$, he can solve for the one unknown Vol_{sph}:

$$V(1) = \mathrm{Vol}_{sph} = (1/2)\,\mathrm{Vol}_{cyl} - (1/3)\,\mathrm{Vol}_{cyl} = (1/6)\,\mathrm{Vol}_{cyl} = (8/6)\pi = (4/3)\pi.$$

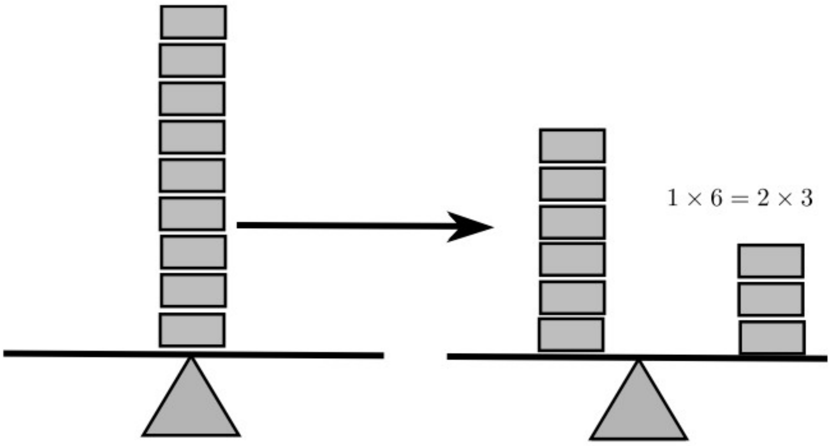

FIGURE 7. Balancing dominos — the center of mass

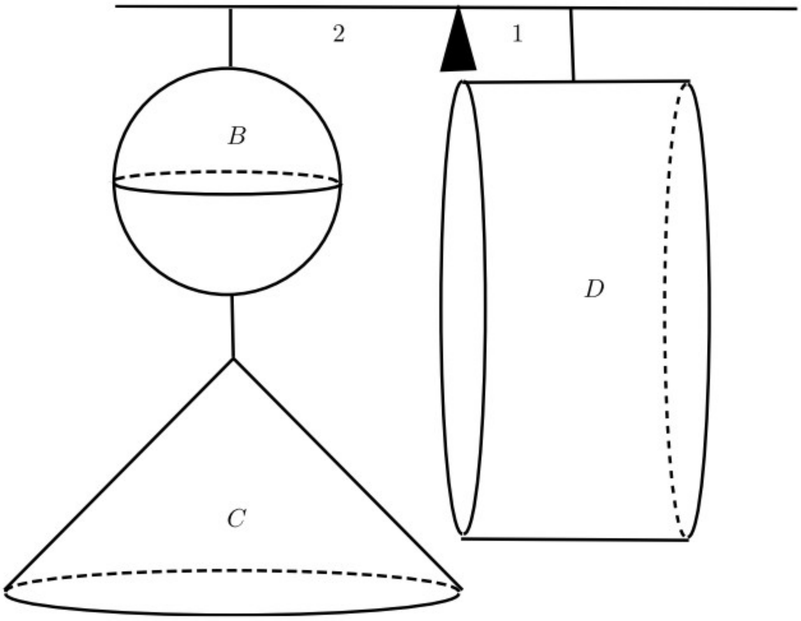

FIGURE 8. The final balance

What is Archimedes's argument?

Archimedes begins with the circle $x^2 + y^2 = 2x$ of radius 1 centered at $(1, 0)$ and given by the Pythagorean Theorem. He rotates this circle about the x-axis to obtain the desired boundary sphere of the ball being considered. Thinking of the fact that the volume of the cone was discovered by considering cross sections in the cone, he considers cross sections in the ball defined by setting x equal to a constant. The area of that cross section is πy^2. Since y^2 appears in the circle equation, Archimedes multiplies the equation by π and seeks to interpret the terms obtained:

$$\pi x^2 + \pi y^2 = \pi 2x.$$

His interpretation is obtained by rotating Figure 9 about the x-axis. The result is pictured in Figure 10.

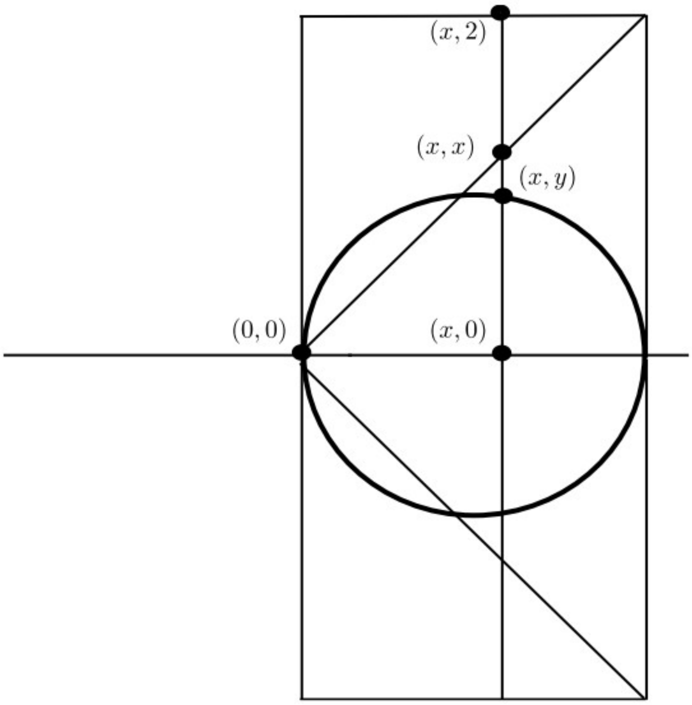

FIGURE 9. The sphere, cone, and cylinder of Archimedes

He observes that πx^2 is also the area of a circle. What circle? As x varies, the circles of radius x concentric with the spherical cross sections of radius y form a cone, namely a cone of base radius 2 and altitude 2. This cone is obtained by rotating the large triangle of the diagram about the x-axis.

Since Archimedes knows the volume of the cone, this seems like progress. Is it possible to interpret the final term $\pi 2x$ in terms of a circular cross-section? It is not at all obvious what to do. What does Archimedes do? He multiplies by 2. He obtains the equation

$$2 \cdot (\pi x^2 + \pi y^2) = x \cdot (\pi 2^2).$$

The factor $\pi 2^2$ is the area of a circle of radius 2. As x varies, the circles become cross sections of constant radius 2 that sweep out a circular cylinder of base radius 2 and altitude 2. This large cylinder is created by rotating the large rectangle of the diagram about the x-axis.

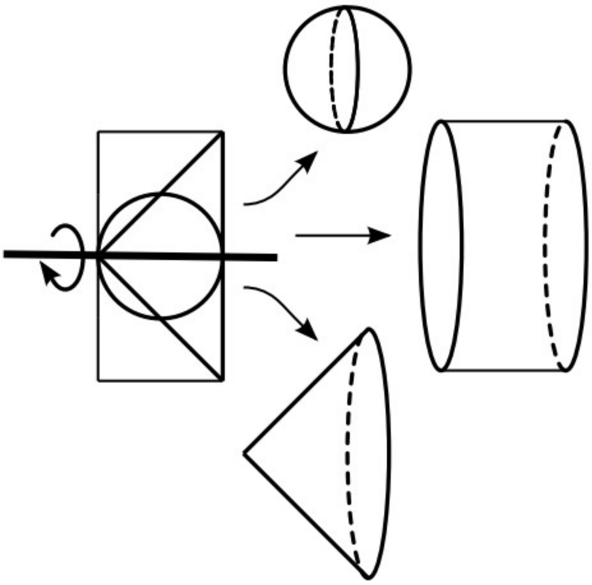

FIGURE 10. The rotation

Archimedes interprets each of the circular cross sections as a weight. He interprets the factor 2 on the left-hand side of the equation as the length of a lever arm and the factor x on the right-hand side of the equation as the length of another lever arm. Then the equation becomes the condition for a balance: Hang the sphere and cone cross sections at distance 2 from the balance center on the left; hang the cylinder cross-section at distance x from the balance center on the right. That configuration balances.

If this configuration balances for one set of cross-sections, why not for all cross sections taken together (addition of infinitely many infinitesimal weights!). That is, hang the entire ball and cone on the left-hand side at distance 2 from the center. Leave the cylinder sections where they are, so that the cylinder remains unmoved. But now apply the principle of center of mass: the cylinder has an obvious center of mass at $(1, 0)$. Hence we may hang the entire cylinder at distance 1 from the center. This completes Archimedes's construction.

The idea of adding infinitely many infinitesimal weights became one of the central ideas leading to calculus. For many years, integration was interpreted precisely as such an operation.

3.4. Wallis Discovers a Product Formula for π

See [11].

In the last section, we spent a good deal of time in finding how Archimedes originally discovered formulas for the area and circumference of a disk and the surface area and volume of a ball. We used slicing and scaling principles, as well as properties of balance and center of mass.

In this section we will admit that we know the basics of calculus and explain two nice results that are usually passed over in our undergraduate classes.

(1) Wallis's formula for π:

$$\frac{\pi}{4} = \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdot \frac{6}{7} \cdot \frac{8}{7} \cdot \frac{8}{9} \cdot \frac{10}{9} \cdots .$$

(2) The volume of the higher dimensional ball $\mathbb{B}^n$:

$$\mathrm{Vol}(\mathbb{B}^n) = \mathrm{Vol}\{(x_1, \cdots, x_n) \mid x_1^2 + \cdots + x_n^2 \leq 1\}.$$

The calculations used for (2) give precisely the results needed for (1). The calculation is by an induction on dimension which uses both the slicing principle and the scaling principle, but in all finite dimensions. A ball of dimension $n+1$ has sections or slices that are balls of dimension n. Assuming that the n-dimensional volume of an n-dimensional ball of radius 1 is known, the n-volume of each slice or section is given by the scaling principle. Thus we obtain a specifically known integral whose value will be the required $n+1$-dimensional volume. That integral can be simplified by one integration by parts (a process well-known from first-year calculus, but which we shall review).

The formula for volume that we obtain is slightly strange. Does the numerical value of the n-volume of the n-ball get larger or smaller as n increases? After all, the n-ball is contained in the $n+1$-ball. We shall see.

Wallis's formula is not particularly useful for approximating π because it converges to $\pi/4$ so slowly. Nevertheless, it is a beautiful formula, and it played a very interesting role in Fourier's development of the heat equation. We will spend a little time explaining Fourier's account without proofs.

3.4.1. Volume of the n-dimensional ball. Our goal is to calculate the volume $V^{(n)}(r)$ of the n-ball $\mathbb{B}^n = \{(x_1, \ldots, x_n) : x_1^2 + \cdots + x_n^2 \leq r^2\}$ of radius r. It suffices to calculate the volume $V^{(n)} = V^{(n)}(1)$, since, by the similarity principle, $V^{(n)}(r) = V^{(n)} \cdot r^n$.

It is obvious that $V^{(1)} = 2$ since $\mathbb{B}^1$ is an interval of length 2. By the very definition of π, $V^{(2)} = \pi$. We proceed by induction.

For $(n+1) \geq 2$, we slice $\mathbb{B}^{n+1}$ by the plane $x_{n+1} = t \in (-1, 1)$ and obtain as cross-section in $\mathbb{B}^{n+1}$ an n-ball of radius $r(t) = (1-t^2)^{\frac{1}{2}}$ and of volume $V^{(n)} \cdot r(t)^n$. Hence

$$V^{(n+1)} = \int_{-1}^{1} V^{(n)} \cdot r(t)^n \, dt = V^{(n)} \cdot \int_{-1}^{1} (1-t^2)^{\frac{n}{2}} \, dt.$$

Setting $\square[n] = \int_{-1}^{1} (1-t^2)^{\frac{n}{2}} \, dt$, we obtain the basic formula

$$V^{(n+1)} = V^{(n)} \cdot \square[n].$$

Since $V^{(1)} = \square[0] = 2$, it follows immediately from equation (1) that

$$V^{(n+1)} = \square[0] \cdot \square[1] \cdot \square[2] \cdots \square[n]. \tag{1}$$

Thus it suffices to calculate $\square[n]$. The integrals $\square[n]$ are precisely the integrals that John Wallis calculated in developing his infinite product representation of $\pi/4$.

We will need to integrate once by parts. The formula for integration by parts is simply the application of the product rule for derivatives to integration:

$$\frac{d}{dx}(u(x) \cdot v(x)) = u(x) \cdot v'(x) + u'(x) \cdot v(x).$$

Hence

$$\int_a^b u(x)dv(x) = u(x)v(x)\Big|_a^b - \int_a^b v(x)du(x),$$

where $dv(x) = v'(x)dx$ and $du(x) = u'(x)dx$. In shorthand,

$$\int udv = uv - \int vdu.$$

The easy cases are

$$\Box[0] = 2 = V^{(1)} \text{ and } \Box[1] = V^{(2)}/V^{(1)} = \pi/2. \tag{2}$$

For $n \geq 2$, we peel off one copy of $(1-t^2)$ in order to split the integral $\Box[n]$ in two:

$$\Box[n] = \int_{-1}^{1} (1-t^2)^{\frac{n-2}{2}}\, dt + \int_{-1}^{1} t \cdot [-t(1-t^2)^{\frac{n-2}{2}}]\, dt.$$

The first of the two resulting integrals is obviously $\Box[n-2]$. Since the derivative of $v(t) = (1/n)\cdot(1-t^2)^{\frac{n}{2}}$ is $[-t(1-t^2)^{\frac{n-2}{2}}]$, the second integral can be integrated by parts, with $u(t) = t$:

$$\int_{-1}^{1} u(t)dv(t) = \Big[u(t)\cdot v(t)\Big]_{-1}^{1} - (1/n)\int_{-1}^{1} (1-t^2)^{\frac{2}{2}}\, dt = 0 - \frac{1}{n}\Box[n].$$

Hence we obtain the reduction formula

$$\Box[n] = \frac{n}{n+1}\Box[n-2]. \tag{3}$$

Equations (1), (2), and (3) yield recursive definitions of all integrals $\Box[n]$ and all volumes $V^{(n)}$. We easily obtain the results

$$\Box[2n] = 2\cdot\frac{2}{3}\cdot\frac{4}{5}\cdots\frac{2n}{2n+1} = n!\cdot\frac{2}{1}\cdot\frac{2}{3}\cdot\frac{2}{5}\cdots\frac{2}{2n+1} \quad \text{and} \tag{4}$$

$$\Box[2n-1] = \frac{\pi}{2}\cdot\frac{3}{4}\cdot\frac{5}{6}\cdots\frac{2n-1}{2n} = \frac{\pi}{n!}\cdot\frac{1}{2}\cdot\frac{3}{2}\cdot\frac{5}{2}\cdots\frac{2n-1}{2}, \tag{5}$$

from which it is easy to deduce that

$$V^{(2n)} = \frac{\pi^n}{n!},$$

$$V^{(2n+1)} = \pi^n\cdot\frac{2}{1}\cdot\frac{2}{3}\cdot\frac{2}{5}\cdots\frac{2}{2n+1}.$$

EXERCISE 3.7. Calculate $\lim_{n\to\infty} V^{(n)}$.

Those familiar with the Γ function will recognize the latter two formulas as expressing

$$V^{(m)} = \frac{\pi^{\frac{m}{2}}}{\Gamma(\frac{m}{2}+1)}.$$

3.4.2. Volume of the n-dimensional ball. Since equations (4) and (5) so easily imply Wallis's product formula, we include the derivation here:

The integral $\square[n]$ is a decreasing function of n. Hence

$$\square[2n] < \square[2n-1] < \square[2n-2].$$

That is,

$$2 \cdot \frac{2}{3} \cdot \frac{4}{5} \cdots \frac{2n}{2n+1} < \frac{\pi}{2} \cdot \frac{3}{4} \cdot \frac{5}{6} \cdots \frac{2n-1}{2n} < 2 \cdot \frac{2}{3} \cdot \frac{4}{5} \cdots \frac{2n-2}{2n-1}.$$

Solving these inequalities for $\pi/4$, we find

$$\frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdots \frac{2n-2}{2n-1} \cdot \frac{2n}{2n-1} \cdot \frac{2n}{2n+1} < \frac{\pi}{4} < \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdots \frac{2n-2}{2n-1} \cdot \frac{2n}{2n-1}.$$

The difference between the larger right-hand side and the smaller left-hand side is $1/(2n)$ times the left-hand side. Hence both the left-hand expression and the right-hand expression converge to $\pi/4$ as $n \to \infty$. Thus we obtain Wallis's expression:

$$\pi/4 = \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdot \frac{6}{7} \cdot \frac{8}{7} \cdot \frac{8}{9} \cdot \frac{10}{9} \cdots.$$

PROBLEM 3.8. Can you devise a way to use Wallis's formula so as to obtain more rapid approximations to π?

3.5. Fourier Discovers Fourier Series

See [13].

Contrary to almost everything else in this book, the following will be only an outline of some mathematics, rather than a complete proof.

Fourier saw, in the problem of heat, his chance to make an historic contribution to physics. He developed a differential equation called the heat equation, and sought to understand its behavior in basic situations. In particular, he considered an infinite, solid, homogeneous slab whose planar slice looked like Figure 11.

The vertical sides were to be cooled to freezing. Fourier chose units so that freezing was a temperature of 0. The bottom was to be heated to boiling. He chose units so that boiling was a temperature of 1. He wrote:

> The present question consists in determining the permanent temperature of a rectangular infinite solid, when the sides are maintained at freezing temperature and the bottom at boiling temperature.
>
> The consideration of such simple and basic questions is one of the most certain means of discovering the laws of natural phenomena, and we see by the history of science that all theories are formed following this method.

Fourier applies the general *heat equation*

$$\frac{d\nu}{dt} = \frac{K}{CD} \times \left(\frac{\partial^2 \nu}{\partial x^2} + \frac{\partial^2 \nu}{\partial y^2} + \frac{\partial^2 \nu}{\partial z^2} \right).$$

The variable ν represents temperature. The parameters K, C, and D are constants determined experimentally: K refers to how rapidly a quantity of heat can move through the solid, C to the capacity of the substance to store heat, D the density of the substance.

-256- - -* ! - / 1 1"
E 2 7 3
('% $ B

On the other hand, the formalism of modular symbols allows one to write for this number an expression having the same structure as (2.4):

$$1 - \lambda_p + p = \sum_{p=\Delta\Delta'+\delta\delta'} y_{11}(\Delta, \delta). \tag{2.6}$$

This time, however, $y_{11}(\Delta, \delta)$ is not a polynomial: it depends only on $(\Delta : \delta) \bmod 11$: for the values of the latter $0, \infty, \pm 1, \pm 2, \pm 3, \pm 4, \pm 5 \bmod 11$, the values of y_{11} are respectively $2, -2, 0, 10, 5, -5, -10$.

Thus, we have connected solutions modulo p of the equation (2.5) "depending on 11" as its conductor with solutions modulo 11 of the equation $p = \Delta\Delta' + \delta\delta'$ depending on p. This justifies the name "non–commutative reciprocity law" suggested for (2.6) and its generalizations in [**Man72**].

Such formulas can be used to make more explicit the exact arithmetic content of special cases of the very general and therefore somewhat abstruse Langlands formalism.

Proofs of formulas for coefficients such as (2.4), (2.6) consist of two steps. For simplicity, we will illustrate this for the case of weight two cusp form $f(z)$ which is an eigenform with respect to a Hecke operator T_n so that $T_n f = a_n f$. We integrate this identity, say, from 0 to $i\infty$ and get

$$\int_0^{\infty} T_n f dz = a_n \int_0^{i\infty} f dz.$$

Now, use the explicit definition of the Hecke operator T_n on the left hand side and make a change of variables. We will get a sum of modular symbols. Using a continued fraction trick and a lemma initially proved by Heilbronn, we finally reduce the left hand side to a sum over solutions of $n = \Delta\Delta' + \delta\delta'$.

2.6. Relations with noncommutative geometry and a real analog of p–adic integration. The role of the upper half-plane in our constructions is of course explained by the fact that it parametrizes elliptic curves: complex tori $\mathbf{C}/\langle 1, \tau \rangle$, $\tau \in H$. The action of the modular group extends to this family, and the respective quotient is a non–complete algebraic variety. The cusps $\tau \in \mathbf{P}^1(\mathbf{Q})$ can be added to compactify this quotient by degenerate elliptic curves. However, for irrational values $\theta \in \mathbf{R} \setminus \mathbf{Q}$, the quotient $\mathbf{C}/\langle 1, \theta \rangle = \mathbf{C}^*/(e^{2\pi i \theta})$ is a "bad" topological group, and the common wisdom is that it is best represented by a non–commutative space, (a version of) the quantum torus T_θ.

Tori T_θ are parametrized by $\theta \in \mathbf{R}$. However, if one considers only tori modulo Morita equivalence, then they are parametrized by $PGL\,(2, \mathbf{Z}) \setminus \mathbf{P}^1(\mathbf{R})$. Set–theoretically, $PGL\,(2, \mathbf{Z}) \setminus \mathbf{P}^1(\mathbf{R})$ = the set of equivalence classes of $\alpha \in [0, 1)$ modulo the relation

$$\alpha \equiv \beta \ \Leftrightarrow\ \exists\, n_0, n_1\ \forall n > 0, \quad k_{n+n_0}(\alpha) = k_{n+n_1}(\beta).$$

Here $k_n(\alpha)$ are successive components of the continued fraction of α.

Thus, we can imagine an "invisible boundary" of the modular tower supporting a family of non–commutative spaces, the phantom of the classical modular family.

This viewpoint was discussed in [**MM02**], see also [**Mar05**], and in particular the Gauss problem on the distribution of continued fractions and its generalizations were treated as a measure theory on the "non–commutative modular curves".

We will describe here one result of this study, which produces an "∞–adic analogue" of the theory of p–adic integration used to construct p–adic Mellin transforms of cusp forms in [**Man73**].

Fix a prime number $N > 0$ and put $G_0 = \Gamma_0(N)$. We will assume that the genus of $X_{G_0} = X_0(N)$ is ≥ 1. Consider a $\Gamma_0(N)$–invariant differential $\omega = f(z)dz$ on H such that $f(z)$ is a cusp eigenform of weight two for all Hecke operators and denote by $L_f^{(N)}(s)$ (resp. $\zeta^{(N)}(s)$) its Mellin transform (resp. Riemann's zeta) with omitted Euler N–factor. More precisely, the coefficients of $L_f^{(N)}(s)$ are Hecke eigenvalues of f.

For $\alpha \in (0,1)$, denote by $p_n(\alpha)/q_n(\alpha)$ the n–th convergent of α.

Theorem. *We have for* $\mathrm{Re}\, t > 0$*:*

$$\int_0^1 d\alpha \sum_{n=0}^{\infty} \frac{q_{n+1}(\alpha) + q_n(\alpha)}{q_{n+1}(\alpha)^{1+t}} \int_0^{q_n(\alpha)/q_{n+1}(\alpha)} f(z)\, dz =$$

$$\left[\frac{\zeta(1+t)}{\zeta(2+t)} - \frac{L_f^{(N)}(2+t)}{\zeta^{(N)}(2+t)^2}\right] \int_0^{i\infty} f(z)\, dz. \tag{2.7}$$

If $\int_0^{i\infty} f(z)\, dz \neq 0$, we can read (2.7) as an expression for $L_f^{(N)}(s)$ which has striking structural similarities to the p–adic Mellin integral. In particular, both formulas involve a construction of a measure out of modular symbols, on $(0,1)$ and on $\mathbf{Z}_p^*$ respectively.

The proof of (2.7) given in [**MM02**] combines an old lemma by P. Lévy with the continued fractions trick alluded to above.

The Theorem above does not involve directly the non–commutative geometry of the invisible boundary. However, it was shown in [**MM02**], Sec. 4, and [**Mar05**], Sec. 6 of Ch. 4, that modular symbols themselves can be identified with specific elements in the K–theory of this space, giving additional weight to the geometric intuition behind this picture.

3. Iterated modular symbols

3.1. Multiple zeta values and iterated integrals. The theory of iterated modular symbols (cf. [**Man06**], [**Man05**]) is a simultaneous generalization of two constructions–of classical modular symbols and of multiple zeta values–and is an elaboration of a special case of Chen's iterated integrals theory ([**Che77**]) in a holomorphic setting.

Multiple zeta values are the numbers given by the k–multiple Dirichlet series

$$\zeta(m_1, \ldots, m_k) = \sum_{0<n_1<\cdots<n_k} \frac{1}{n_1^{m_1} \cdots n_k^{m_k}} \tag{3.1}$$

which converge for all integer $m_i \geq 1$ and $m_k > 1$, or equivalently by the m–multiple iterated integrals, $m = m_1 + \cdots + m_k$,

$$\zeta(m_1, \dots, m_k) = \int_0^1 \frac{dz_1}{z_1} \int_0^{z_1} \frac{dz_2}{z_2} \int_0^{z_2} \cdots \int_0^{z_{m_{k-1}}} \frac{dz_{m_k}}{1 - z_{m_k}} \cdots \tag{3.2}$$

where the sequence of differential forms in the iterated integral consists of consecutive subsequences of the form $\dfrac{dz}{z}, \dots, \dfrac{dz}{z}, \dfrac{dz}{1-z}$ of lengths $m_k, m_{k-1}, \dots, m_1$.

Easy combinatorial considerations allow one to express in two different ways products $\zeta(l_1, \dots, l_j) \cdot \zeta(m_1, \dots, m_k)$ as linear combinations of multiple zeta values.

If one uses for this the integral representation (3.2), one gets a sum over shuffles which enumerate the simplices of highest dimension occurring in the natural simplicial decomposition of the product of two integration simplices.

If one uses instead (3.1), one gets sums over shuffles with repetitions which enumerate some simplices of lower dimension as well.

These relations and their consequences are called double shuffle relations. Both types of relations can be succinctly written down in terms of formal series on free noncommuting generators. One can include in these relations regularized multiple zeta values for arguments where the convergence of (3.1), (3.2) fails. A clear and systematic exposition of these results can be found in [**Del01**] and [**Rac00**], [**Rac02**].

In fact, the formal generating series for (regularized) iterated integrals (3.2) appeared in the famous Drinfeld paper [**Dri90**], essentially as *the Drinfeld associator*, and more relations for multiple zeta values were implicitly deduced there. The question about interdependence of (double) shuffle and associator relations does not seem to be settled at the moment of writing this: cf. [**Rac04**]. The problem of completeness of these systems of relations is equivalent to some difficult transcendence questions.

Multiple zeta values are interesting, because they and their generalizations appear in many different contexts involving mixed Tate motives ([**DG05**], [**Ter02**]), deformation quantization ([**Kon99**]), knot invariants, etc.

In order to make contact with modular symbols, notice first that the differentials $\dfrac{dz}{z}, \dfrac{dz}{1-z}$ span the space of meromorphic differential forms with no more than logarithmic singularities at points $\{0, 1, \infty\}$ of $\mathbf{P}^1(\mathbf{C})$. We can identify

$$(\mathbf{P}^1(\mathbf{C}), \{0, 1, \infty\}) \cong \Gamma_0(4) \setminus (\overline{H}, cusps).$$

Then $\dfrac{dz}{z}, \dfrac{dz}{1-z}$ lift to Eisenstein series of weight two for $\Gamma_0(4) \subset SL(2, \mathbf{Z})$.

In the general theory sketched below, $\Gamma_0(4)$ is replaced by an arbitrary (congruence) subgroup Γ of $SL(2, \mathbf{Z})$, Eisenstein series of weight two are replaced by (cusp form + Eisenstein series) with respect to Γ, multiplied by $z^{s-1}dz$ for appropriate s. (We mostly focus on cusp forms; in the presence of logarithmic singularities, the necessary regularization procedure is described for weight two in Sec. 3.6.)

Finally, ordinary integrals along geodesics connecting two cusps are replaced by iterated integrals.

characteristic 0 enjoys similar properties, through the Gauss-Manin connection and the Kodaira-Spencer map.

Theorem 2.4. $\boxed{\textbf{BB}}$ (Grothendieck-Messing) *Let* $X_0 \to \mathrm{Spec}(R_0)$ *be an* p*-divisible group over a commutative ring* R_0.

(i) *To every locally nilpotent* DP *extension* $(R, I, (\gamma_i)_{i\in\mathbb{N}})$ *of* R_0 *there is a functorially attached locally free* R*-module* $\mathbb{D}(X_0)_R = \mathbb{D}(X0)_{(R,I,(\gamma_i))}$ *of rank* $\mathrm{ht}(X_0)$. *The functor* $\mathbb{D}(X_0)$ *is called the covariant Dieudonné crystal attached to* X_0.

(ii) *Let* $(R, I, (\gamma_i)_{i\in\mathbb{N}})$ *be a locally nilpotent* DP *extension of* R_0. *Suppose that* $X \to \mathrm{Spec}(R)$ *is a* p*-divisible group extending* $X_0 \to \mathrm{Spec}(R_0)$. *Then there is a functorial short exact sequence*

$$0 \to \mathrm{Lie}(X^t/R)^\vee \to \mathbb{D}(X_0)_R \to \mathrm{Lie}(X/R) \to 0\,.$$

Here $\mathrm{Lie}(X/R)$ *is the tangent space of the* p*-divisible group* $X \to \mathrm{Spec}(R)$, *which is a projective* R*-module of rank* $\dim(X/R)$, *and* $\mathrm{Lie}(X^t/R)^\vee$ *is the* R*-dual of the tangent space of the Serre dual* $X^t \to \mathrm{Spec}(R)$ *of* $X \to \mathrm{Spec}(R)$.

(iii) *Let* $(R, I, (\gamma_i)_{i\in\mathbb{N}})$ *be a locally nilpotent* DP *extension of* R_0. *Suppose that* $A \to \mathrm{Spec}(R)$ *is an abelian scheme such that there exists an isomorphism*

$$\beta : A[p^\infty] \times_{\mathrm{Spec}(R)} \mathrm{Spec}(R_0) \xrightarrow{\sim} X_0$$

of p*-divisible groups over* R_0. *Then there exists a natural isomorphism*

$$\mathbb{D}(X_0)_R \to \mathrm{H}_1^{\mathrm{DR}}(A/R)\,,$$

where $\mathrm{H}_1^{\mathrm{DR}}(A/R)$ *is the first de Rham homology of* $A \to \mathrm{Spec}(R)$. *Moreover the above isomorphism identifies the short exact sequence*

$$0 \to \mathrm{Lie}(A[p^\infty]^t/R)^\vee \to \mathbb{D}(X_0)_R \to \mathrm{Lie}(A[p^\infty]/R) \to 0$$

described in (ii) *with the Hodge filtration*

$$0 \to \mathrm{Lie}(A^t/R)^\vee \to \mathrm{H}_1^{\mathrm{DR}}(A/R) \to \mathrm{Lie}(A/R) \to 0$$

on $\mathrm{H}_1^{\mathrm{DR}}(A/R)$.

(iv) *Let* $(R, I, (\gamma_i)_{i\in\mathbb{N}})$ *be a locally nilpotent* DP *extension of* R_0. *Denote by* $\mathfrak{C}$ *the category whose objects are short exact sequences*

$$0 \to F \to \mathbb{D}(X_0)_R \to Q \to 0$$

such that F *and* Q *are projective* R*-modules, plus an isomorphism from the short exact sequence*

$$(0 \to F \to \mathbb{D}(X_0)_R \to Q \to 0) \otimes_R R_0$$

of projective R_0*-modules to the short exact sequence*

$$0 \to \mathrm{Lie}(X_0^t)^\vee \to \mathbb{D}(X_0)_{R_0} \to \mathrm{Lie}(X) \to 0$$

attached to the p*-divisible group* $X_0 \to \mathrm{Spec}(R_0)$ *as a special case of* (ii) *above. The morphisms in* $\mathfrak{C}$ *are maps between diagrams. Then the functor from the category of* p*-divisible groups over* R *lifting* X_0 *to the category* $\mathfrak{C}$ *described in* (ii) *is an equivalence of categories.*

Theorem 2.4 is a summary of the main results in Chapter IV of [**49**].

Corollary 2.5. *Let* X_0 *be a* p*-divisible group over a perfect field* $K \supset \mathbb{F}_p$*. Let* $d = \dim(X_0)$*,* $c = \dim(X_0^t)$*. The deformation functor* $\mathrm{Def}(X_0/W(K))$ *of* X_0 *is representable by a smooth formal scheme over* $W(K)$ *of dimension* cd*. In other words,* $\mathrm{Def}(X_0/W(K))$ *is non-canonically isomorphic to the functor represented by the formal spectrum* $\mathrm{Spf}\,(W(K)[[x_1, \ldots, x_{cd}]])$*.*

Proof. Recall that formal smoothness of $\mathrm{Def}(X_0/W(K))$ means that the natural map

$$\mathrm{Def}(X_0/W(K))(R) \to \mathrm{Def}(X_0/W(K))(R')$$

attached to any surjective ring homomorphism $R \to R'$ is surjective, for any Artinian local rings $R, R' \in \mathrm{Art}_{W(K)}$. To prove this, we may and do assume that the kernel I of the surjective homomorphism $R \to R'$ satisfies $I^2 = (0)$. Apply Theorem 2.4 to the trivial DP structure on pairs (R, I) with $I^2 = (0)$, we see that $\mathrm{Def}(X_0/W(K))$ is formally smooth over $W(K)$. Applying Theorem 2.4 again to the pair $K[t]/(t^2), tK[t]/(t^2)$, we see that the dimension of the tangent space of $\mathrm{Def}(X_0/K)$ is equal to cd. □

2.6. We set up notation for the Serre-Tate Theorem 2.7, which says that deforming an abelian variety A_0 over a field of characteristic p is the same as deforming the p-divisible group $A_0[p^\infty]$ attached to A_0. Recall that $A_0[p^\infty]$ is the inductive system formed by the p^n-torsion subgroup schemes $A_0[p^n]$ of A_0, where n runs through positive integers. In view of Theorem 2.4 one can regard the p-divisible group $A_0[p^\infty]$ as a refinement of the first homology group of A_0.

Let p be a prime number. Let S be a scheme such that p is locally nilpotent in $\mathcal{O}_S$. Let $I \subset \mathcal{O}_S$ be a coherent sheaf of ideals such that I is locally nilpotent. Let $S_0 = \mathrm{Spec}(\mathcal{O}_S/I)$. Denote by AV_S the category of abelian schemes over S. Denote by $\mathrm{AVBT}_{S_0,S}$ the category whose objects are triples $(A_0 \to S_0, X \to S, \epsilon)$, where $A_0 \to S_0$ is an abelian scheme over S_0, $X \to S$ is a p-divisible group over S, and $\epsilon : X \times_S S_0 \to A_0[p^\infty]$ is an isomorphism of p-divisible groups. A morphism from $(A_0 \to S_0, X \to S, \epsilon)$ to $(A_0' \to S_0, X' \to S, \epsilon')$ is a pair (h, f), where $h_0 : A_0 \to A_0'$ is a homomorphism of abelian schemes over S_0, and $f : X \to X'$ is a homomorphism of p-divisible groups over S, such that $h[p^\infty] \circ \epsilon = \epsilon' \circ (f \times_S S_0)$. Let

$$\mathbb{G}_{S_0,S} : \mathrm{AV}_S \to \mathrm{AVBT}_{S_0,S}$$

be the functor which sends an abelian scheme $A \to S$ to the triple

$$((A \times_S S_0, A[p^\infty], \mathrm{can}),$$

where can is the canonical isomorphism $A[p^\infty] \times_S S_0 \xrightarrow{\sim} (A \times_S S_0)[p^\infty]$.

Theorem 2.7. $\boxed{\textbf{BB}}$ (Serre-Tate) *Notation and assumptions as in the above paragraph. The functor* $\mathbb{G}_{S_0,S}$ *is an equivalence of categories.*

Remark. See [**46**]. A proof of Theorem 2.7 first appeared in print in [**49**]. See also [**41**].

Corollary 2.8. *Let* A_0 *be an abelian variety over a perfect field* K*. Let*

$$\mathbb{G} : \mathrm{Def}(A_0/W(K)) \to \mathrm{Def}(A_0[p^\infty]/W(K))$$

be the functor which sends any object

$$\Big(A \to \mathrm{Spec}(R), \epsilon : A \times_{\mathrm{Spec}(R)} \mathrm{Spec}(R/\mathfrak{m}_R) \xrightarrow{\sim} A_0 \times_{\mathrm{Spec}(k)} \mathrm{Spec}(R/\mathfrak{m}_R)\Big)$$

(% B

Pag. 117(134) - Sphere Metric

$ds^2 = r^2 d\theta^2 + r^2 \sin^2\theta \, d\phi^2$.

A sphere of radius r has constant positive Gaussian curvature equal to r.

2) _20211017/ Quantum theory of Solids/ Han.. A Modern Course in the Quantum theory of Solid...pdf

3) _20211017/ Krishna's - Engineering Mathematics III ..pdf

Obs: Gauss - 68 occurences.

Pag. 677 Chapter 15

Solution of System of Linear Equation

Gauss - Seidel Method to solve a system of Linear Equations.

Pag. 684 - 15.3

Gauss - Seidel Method

4) _20211017/ Mathematical Modelling.pdf

(%

Pág. 234(259) - 4 Maxwell equations
Gauss law $\operatorname{div} D = 4\pi\rho$,

_20211017/ Geometry/
Geometry of Lengths, Areas...pdf
Pág. 27(40) - Chapter 3
Areas by Slicing and Scaling

Wallis used basic calculations of calculus to discover a beautiful product formula for π.

Fourier used Wallis's formula for π in solving a special ~~product~~ for case of the heat equation which he had developed.

_20211017/ Geometry/
cmip08_c.pdf (Arithmetic Geometry)
Pág. 143(150) 2.6

Relations with noncommutative geometry and a real analog of p-adic integration.

→ Pág. 144(151) This viewpoint was discussed in [MM02], see also [Mar05], and in particular the GAUSS problem on the distribution of continued fractions and its generalizations were ~~the~~ treated as a measure theory on the "non-commutative modular curves".

Pág. 454(461) -

through the Gauss-Manin connection and the Kodaria-Spencer map.

_ 20211017/ Geometry/

Projective Geometry (Oxford).pdf

Pág. 15(28) 2 Introduction

the Desargues' and Pappus' configurations

_ collinear
_ concurrent

Obs. Livro todo!!

_ 20211017/ Geometry/

Classical Geometry Euclidean, transformational, Inversive, and Projective. pdf

Obs. Gauss - 07 occurrences.

!   -256-   -   -*   ! -   K   A 7   3 2 &   "
*L   " :! "D .   1"
($ &   "

2 Introduction

The Desargues' and Pappus' configurations

Definition 2.1 *In the Euclidean plane,* $\mathbb{R}^2 = \{(x, y) \mid x, y \in \mathbb{R}\}$

1. *Points which lie on the same line are said to be* **collinear**.
2. *Lines which pass through the same point are said to be* **concurrent**.

Convince yourself (by drawing with the help of a straight edge and a pencil) that the following two theorems *may well in general be true* in the Euclidean plane $\mathbb{R}^2 = \{(x, y) : x, y \in \mathbb{R}\}$.

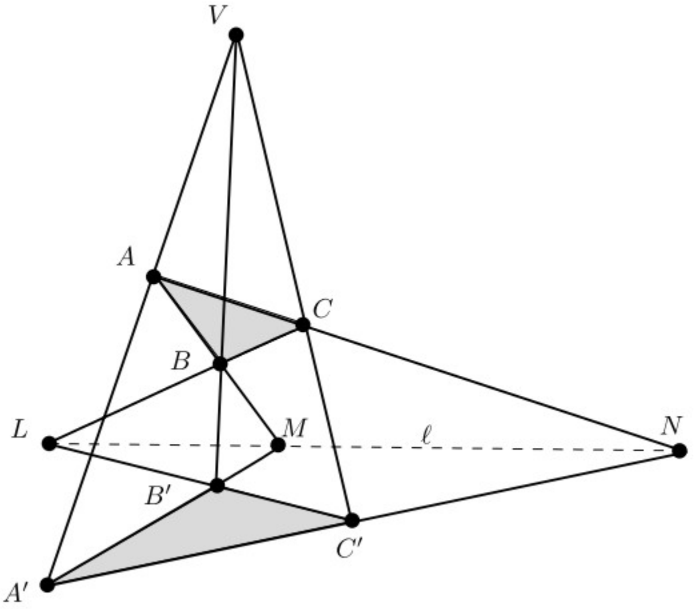

The Desargues' configuration

Theorem 2.2 *(***Desargues' Theorem***) Let* $ABC, A'B'C'$ *be two triangles, with distinct vertices, such that the lines* AA', BB', CC' *are concurrent in the point* V*. Then the points*

$$L = BC \cap B'C',$$
$$M = AB \cap A'B',$$
$$N = AC \cap A'C'$$

are collinear (on a line ℓ*).*

Definition 2.3 *With respect to the Desargues' configuration above,*

1. *The two triangles* $ABC, A'B'C'$ *are said to be* **in perspective from** *the point* V*, and from the line* ℓ.
2. *The point* V *is called the* **vertex**, *and the line* ℓ *is called the* **axis** *of the* **perspectivity**.

Theorem 2.4 *(***Pappus' Theorem***) Let* ℓ, ℓ' *be two lines of the plane.*

Let

1. A, B, C *be points of* ℓ*;*
2. A', B', C' *be points of* ℓ'*;*
3. *all these points be distinct from* $\ell \cap \ell'$.

Then, $L = AB' \cap A'B$, $M = AC' \cap A'C$, $N = BC' \cap B'C$ *are collinear (i.e. L, M, and N lie on a line).*

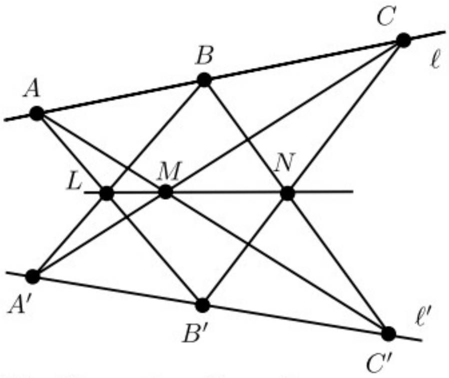

The Pappus' configuration

The extended Euclidean plane

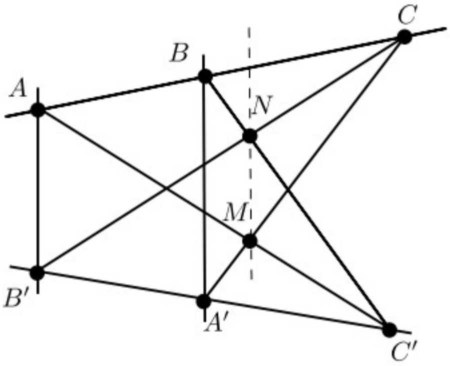

A few drawings do not constitute a proof. But your drawings may have convinced you that in the Euclidean plane, the theorems of Pappus and Desargues (Theorems 2.4 and 2.2) may well be true in general; these theorems are definitely not true in the cases where some of the relevant lines are parallel. For example, (see the adjoining diagram), if AB' and BA' are parallel, then $L = AB' \cap A'B$ does not exist, and therefore in this case Pappus' Theorem does not hold. It is to be noted that, in this case,

1. If M and N exist, then the line MN is parallel to AB' and BA'.
2. If M does not exist, which is the case when AC' is parallel to $A'C$, then N also does not exist.

Parallelism is a feature of the Euclidean plane. We need to consider whether there is anything to be gained if the Euclidean plane is *modified* to allow parallel lines to intersect.

In order to develop the theory, it is convenient to use the following terminology. We define a **triple $(\mathcal{P}, \mathcal{L}, \mathcal{I})$** to be

- A set $\mathcal{P}$ whose elements are called *points*, together with
- A set $\mathcal{L}$ of subsets of $\mathcal{P}$ whose elements are called *lines* (or *blocks*),
- An *incidence relation* $\mathcal{I}$ which describes precisely which points belong to which lines (blocks).

For example, the Euclidean plane is a triple $(\mathcal{P}, \mathcal{L}, \mathcal{I})$, where

$\mathcal{P} = \{(x, y) \mid x, y \in \mathbb{R}\}$,

$\mathcal{L} = \{[a, b, c] \mid a, b, c \in \mathbb{R}, a, b$ not both zero$\}$;

$\mathcal{I}$: point (x, y) lies on line $[a, b, c]$ if and only if $ax + by + c = 0$.

Note 3 *It is sometimes convenient, particularly when $\mathcal{P}$ and $\mathcal{L}$ are finite sets, to express $\mathcal{I}$ as a subset of $\mathcal{P} \times \mathcal{L}$. For example, consider the triple $(\mathcal{P}, \mathcal{L}, \mathcal{I})$ given by*

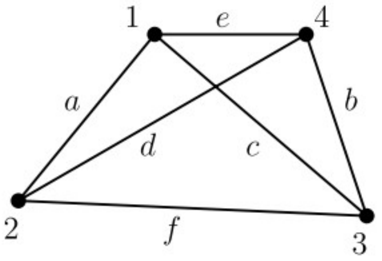

$\mathcal{P} = \{1, 2, 3, 4\}$,

$\mathcal{L} = \{a, b, c, d, e, f\}$,

$\mathcal{I} = \{(1, a), (2, a), (3, b), (4, b), (1, c), (3, c), (2, d), (4, d), (1, e), (4, e), (2, f), (3, f)\}$.

This means that points 1 and 2 lie on line a, points 3 and 4 lie on line b, and so on.

Definition 2.5 (The extended Euclidean plane)

1. *Consider the Euclidean plane $\mathbb{R}^2$.*
2. *Given a line ℓ of $\mathbb{R}^2$, the set consisting of ℓ, and all lines parallel to ℓ is called a* **pencil** *of parallel lines. To each pencil of parallel lines, add an abstract entity P_∞ called the* **point at infinity** *of the pencil. The lines of the pencil are said to intersect at P_∞. A line ℓ together with its point at infinity is called an* **extended line**, *and is denoted by ℓ^*.*
3. *Stipulate that distinct pencils of lines have distinct points at infinity.*
4. *The set of all points at infinity is called the* **line at infinity**, *and is denoted by ℓ_∞.*

Then the **extended Euclidean Plane** *(EEP) is a triple* $(\mathcal{P}, \mathcal{L}, \mathcal{I})$ *with:*

- *Points* $\mathcal{P}$ *are the points of* $\mathbb{R}^2$ *and the points* P_∞ *at infinity,*
- *Lines* $\mathcal{L}$ *are the extended lines and the line* ℓ_∞.
- *Incidence relation* $\mathcal{I}$ *is the inherited one, namely*
  - (a) *the point* P *not at infinity lies on* ℓ^* *if and only if* P *lies on* ℓ.
  - (b) P_∞ *lies on* ℓ^* *if and only if* P_∞ *is the point at infinity of the pencil of parallel lines determined by* ℓ.
  - (c) *All the points at infinity lie on* ℓ_∞.

Note 4 *We will prove that the theorems of Pappus and Desargues, as stated in Theorems 2.4 and 2.2, are valid in the EEP.*

Theorem 2.6 *In the EEP, we have*

1. *Two distinct points* lie *on a* unique *line.*
2. *Two distinct lines* intersect *in a* unique *point.*

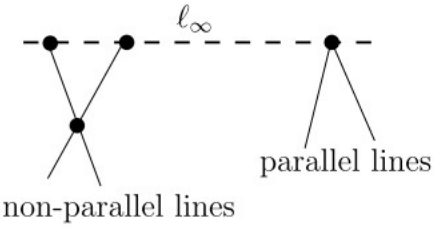

Proof. As in Definition 2.5, a line $\ell \in \mathbb{R}^2$ when extended is denoted by ℓ^*.

1. Let A and B be two distinct points of the EEP.
   - (a) If A and B are two distinct points of $\mathbb{R}^2$, then $(AB)^*$ is the unique line of the EEP which contains them.
   - (b) If A and B are two distinct points at ∞, then ℓ_∞ is the unique line of the EEP which contains them.
   - (c) Suppose $A \in \mathbb{R}^2$, and $B \in \ell_\infty$. Then B is the point at infinity of a unique pencil of parallel lines. There is a unique line ℓ of that pencil passing through A. Then ℓ^* is the unique line of the EEP which contains A and B.
2. Three cases to be considered.
   - (a) Let ℓ and m be two distinct non-parallel lines of $\mathbb{R}^2$. Therefore they intersect in a unique point P of $\mathbb{R}^2$. Then P is the unique point of intersection of ℓ^* and m^*.
   - (b) Let ℓ and m be two distinct parallel lines of $\mathbb{R}^2$. Then the pencil of parallel lines containing ℓ and m has a unique point P_∞ at infinity: P_∞ is the unique point of intersection of ℓ^* and m^*.
   - (c) The unique point at infinity of a line ℓ^* is the unique point of intersection of ℓ^* and ℓ_∞.

□

Definition 2.7 *In the EEP, a set of four points $A, B, C, D,$* no three collinear *is called a* **quadrangle**. *A set of four lines $a, b, c, d,$* no three concurrent *is called a* **quadrilateral**. *(The diagram here represents a parallelogram $ABCD$ in $\mathbb{R}^2$, that is a quadrilateral with opposite sides a, c parallel, since a and c intersect at a point at infinity, and opposite sides b, d parallel since b and d intersect at a point at infinity.)*

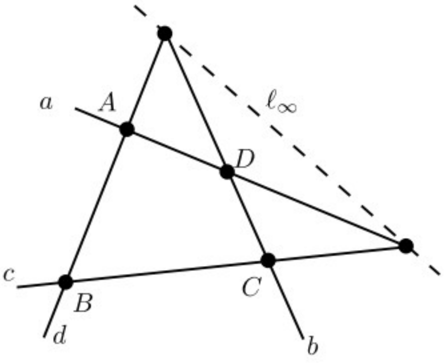

Definition 2.8 *If K is a configuration, then its* **dual configuration** *is the configuration obtained by making the following changes:*

$$\begin{aligned} point &\longleftrightarrow line, \\ collinear &\longleftrightarrow concurrent, \\ join &\longleftrightarrow intersection. \end{aligned}$$

Example 2.9 A quadrangle and a quadrilateral are dual configurations, since
a quadrangle is a set of four *points*, no three *collinear*,
a quadrilateral is a set of four *lines*, no three *concurrent*.

Similarly, a triangle and a **trilateral** are dual configurations, since
a triangle is a set of three *non-collinear points*,
a trilateral is a set of three *non-concurrent lines*.

Thus, a triangle (considered as a set of three points and three sides) is a self-dual configuration.

Example 2.10 Consider the following configurations: (1) two lines l, m intersecting in the point P, (2) three collinear points A, B, C, and (3) a line intersecting three concurrent lines l, m, n in the points B, C, D.

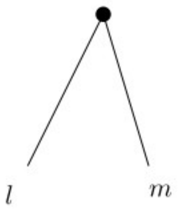

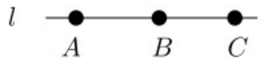

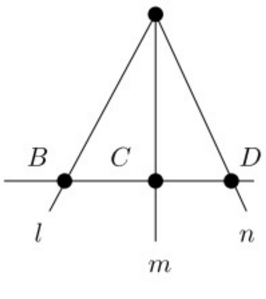

The dual configurations are respectively: (1) two points L, M on line p, (2) three concurrent lines a, b, c through the point L, and (3) three concurrent lines b, c, d intersected by line a in the points L, M, N.

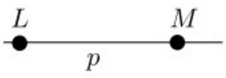

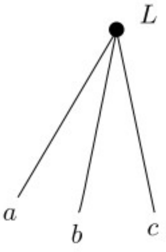

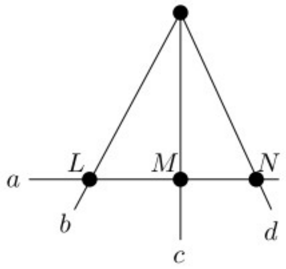

Consider the two statements in Theorem 2.6, namely that in the EEP, two distinct points lie on a *unique* line and two distinct lines intersect in a *unique* point. Either statement can be obtained from the other by the simple interchange of the words 'point' and 'line' and of the words 'lie' and 'intersect'. It follows that any theorem deducible from these two statements remains valid when these words are interchanged. We formally present this important fact as a theorem.

Theorem 2.11 *(***The Principle of Duality***) If $\mathcal{T}$ is a theorem valid in the EEP, and $\mathcal{T}'$ is the statement obtained from $\mathcal{T}$ by making the following changes:*

$$\begin{aligned} \textit{point} &\longleftrightarrow \textit{line}, \\ \textit{collinear} &\longleftrightarrow \textit{concurrent}, \\ \textit{join} &\longleftrightarrow \textit{intersection}, \end{aligned}$$

and whatever grammatical adjustments that are necessary, then $\mathcal{T}'$ (called the **Dual Theorem***) is a valid theorem in the EEP.*

Example 2.12 The following example shows how to dualise a theorem:

Desargues' Theorem	**Dual**
If two triangles $ABC, A'B'C'$	If two trilaterals $abc, a'b'c'$
with distinct vertices	with distinct sides
are such that the lines	are such that the points
AA', BB', CC'	$a \cap a', b \cap b', c \cap c'$
are concurrent in the point V	are collinear on the line v
then, the points	then, the lines
$L = BC \cap B'C'$,	$\ell = \langle b \cap c, b' \cap c' \rangle$,
$M = AC \cap A'C'$,	$m = \langle a \cap c, a' \cap c' \rangle$,
$N = AB \cap A'B'$	$n = \langle a \cap b, a' \cap b' \rangle$
are collinear	are concurrent
on the line v	in the point V

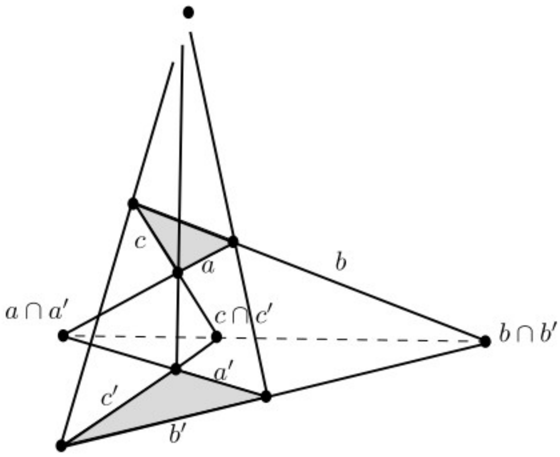

By the Principle of Duality, if Desargues′ Theorem is valid in the EEP, so is its dual.

Exercise 2.1

1. Recall that the affine plane $\mathbb{C}^2$ contains the Euclidean plane $\mathbb{R}^2$ as a *subplane*: the points of the Euclidean plane $\mathbb{R}^2$ are points of the affine plane $\mathbb{C}^2$, and lines of the Euclidean plane $\mathbb{R}^2$ are sublines of lines of the affine plane $\mathbb{C}^2$. Using Definition 2.5 as a model, define the **Extended Plane** $\mathbb{C}^2$ by adding *points at infinity*, and *a line* ℓ_∞^* *at infinity* to the affine plane $\mathbb{C}^2$. Explain carefully the following concepts:
   (a) The lines of EEP are *sublines* of the Extended Plane $\mathbb{C}^2$.
   (b) The EEP is a *subplane* of the Extended Plane $\mathbb{C}^2$.

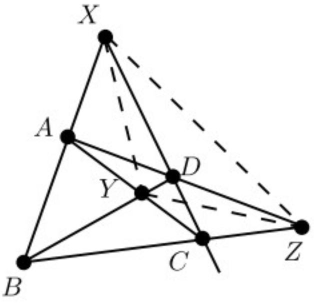

2. Let $ABCD$ be a quadrangle in the EEP.

   Let $X = AB \cap CD, Y = BD \cap CA, Z = AD \cap BC$. The triangle XYZ is called the **diagonal triangle**. Draw the dual configuration (the quadrilateral and the *diagonal trilateral*).

3. Assume that Desargues' Theorem is valid in the EEP.
   (a) State the dual of Desargues' Theorem.
   (b) Draw the dual Desargues' configuration.
   (c) State the converse of Desargues' Theorem.
   (d) Compare your statements in (a) and (c).
   (e) Deduce that the converse of Desargues' Theorem is also valid in the EEP.

4. (a) Dualise Pappus' Theorem.
   (b) Draw the dual Pappus' configuration.

D -256- - -* ! - K A 7 3J " =" 1"
E L

-256- - -* ! - ! ! 7 3 = " ! J
9 ! ! ! J & A A J ! " K A 1"
(% % '

12. Construct a triangle given the foot F of an altitude, the orthocenter H, and the center N of the 9-point circle.

13. Given a diameter $\overline{AB}$ of a circle and a point P as shown, construct a perpendicular from P to $\overleftrightarrow{AB}$, *with a straightedge alone*.

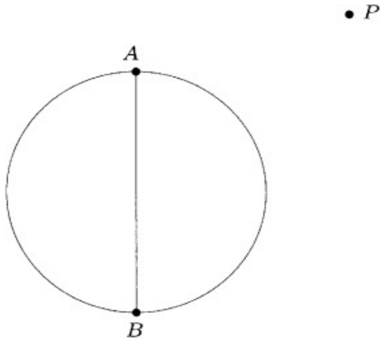

14. (a) List all the regular n-gons with $n \leq 100$ sides that are constructible with a straightedge and compass.
    (b) Use Gauss' Theorem to prove that an angle of $20°$ is not constructible.
    (c) Use Gauss' Theorem to decide whether or not an angle of $6°$ is constructible.

15. Construct $\triangle ABC$ given a line segment $\overline{XY}$ and two adjacent angles, as in the figure, where the length of the perimeter is XY, $\alpha = \angle B$, and $\beta = \angle C$.

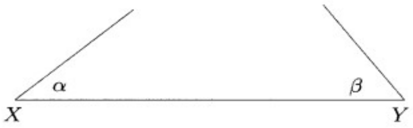

16. Construct triangle ABC given the location of its circumcenter S, the location of its orthocenter H, and the foot of an altitude F.

17. Construct a triangle given the foot F of an altitude, the circumcenter S, and the center N of the 9-point circle.

18. Construct triangle ABC given the location of vertex A, the location of its circumcenter S, and the center N of the 9-point circle.

19. Show that the incenter of a triangle is the Nagel point of its medial triangle.

"   -256-   -   -*   ! -   !   ! 7   "
L   7   3   5 ! &   "   1!   9   " ="   1"

-256-   -   -*   ! -   9   .   7   ! 3 2   A 3
7   1"
('$ 7!   !   ! 3   " !   !   /B

constructible. The only known primes of this form, called Fermat primes, are 3, 5, 17, 257, and 65,536. Gauss actually constructed a regular 17-gon in 1796. He conjectured and Pierre Wantzel proved in 1837 that no other regular polygons are constructible by straightedge and compass alone.

1.2.14. Assume that congruent figures have the same area.

(a) Given $\triangle ABC$ construct, as in Figure 1.23, midpoints D and E of sides $\overline{AB}$ and $\overline{AC}$, respectively, and perpendiculars $\overline{AF}$, $\overline{BG}$, and $\overline{CH}$ to $\overleftrightarrow{DE}$. Show that $\triangle ABC$ has the same area as the quadrilateral $BCHG$. (You may assume that the order of the points on $\overleftrightarrow{DE}$ is as shown in Figure 1.23.)

(b) What orderings of the points D, E, F, G, and H are possible, other than the order in Figure 1.23? Draw some of them. It is a fact that in all cases the areas of $\triangle ABC$ and of $BCHG$ are equal.

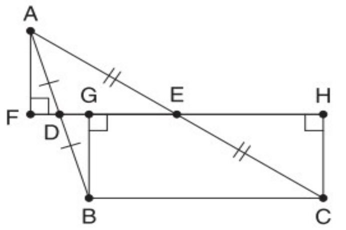

Figure 1.23

*1.2.15. Let $\triangle ABC$ be a triangle and G be the intersection of the perpendicular bisector $\overline{GD}$ of $\overline{AB}$ and the perpendicular bisector $\overline{GE}$ of $\overline{BC}$. (See Figure 1.24.)

(a) Prove that $\overline{AG}$, $\overline{BG}$, and $\overline{CG}$ are congruent. (Thus G is called the *circumcenter*, the center of the *circumscribed* circle of the triangle.)

(b) Let F be the midpoint of $\overline{AC}$. Prove that $\overline{GF}$ is the perpendicular bisector of $\overline{AC}$. (Thus the perpendicular bisectors of a triangle are *concurrent*, that is, they meet in a common point.)

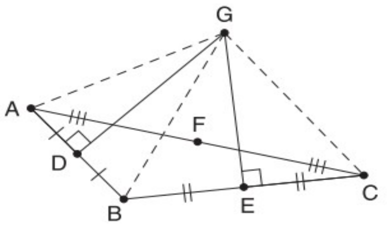

Figure 1.24 G is the circumcenter of $\triangle ABC$.

needed an essentially inborn understanding of geometry and space before we could experience anything in space. (Studies of blind people and people with recovered sight indicate that Kant's argument is incorrect; sight is essential to develop the usual conception of space. See Sacks [13, 124].) Kant also thought that geometry had to be Euclidean.

In the eighteenth century, the Age of Enlightenment, mathematicians and philosophers built on the perceived absolute truth of mathematics in a second way. The astounding success of Newton's calculus and physics convinced his successors that mathematics wasn't just true in some metaphysical sense, but also in a tangible sense. The ideal world of mathematics, it seemed, was the real world. The physical meaning of much of the mathematics developed in the eighteenth century was so convincing that the rigorous deductive methods of Greek geometry seemed superfluous. The shock of the radically different mathematical results of the nineteenth century, starting with non-Euclidean geometries, forced mathematicians to reintroduce careful proofs.

The first person to break from the world view of Euclidean geometry, its unquestionable truth, and its applicability to the physical world was Carl Friedrich Gauss (1777–1855). Despite his fame, Gauss never published anything on non-Euclidean geometries because he feared ridicule. Nicolai Lobachevsky (1793–1856) and János Bolyai (1802–1860), the two young mathematicians who did publish on non-Euclidean geometry, were greeted with silence for years after their publications in 1829 and 1832, respectively. Indeed, only with the publication of Gauss's notes after his death did the wider community of mathematicians start investigating non-Euclidean geometries.

4.1.1 The Advent of Hyperbolic Geometry

Hyperbolic geometry, the non-Euclidean geometry that Gauss, Lobachevsky, and Bolyai developed, retains Euclid's first four postulates and changes the fifth postulate to the following axiom. In Sections 4.2, 4.3, and 4.4, we investigate hyperbolic geometry in much the same way that they did. However, to make the process clearer, we make explicit certain logically necessary assumptions, overlooked until the end of the nineteenth century. (See Sections 2.1 and 2.2.) Hyperbolic geometry is sometimes called Lobachevskian geometry to honor Lobachevsky's priority in publishing. Felix Klein called the geometry *hyperbolic* in his classification of geometries, which we discuss in Section 7.5.

Characteristic Axiom of Hyperbolic Geometry Given a line k and a point P not on k, there are at least two lines on P that do not intersect k (See Figure 4.1).

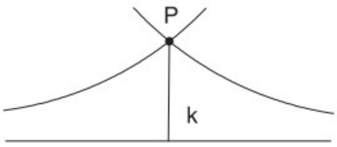

Figure 4.1 The characteristic axiom of hyperbolic geometry.

Various consequences follow from this change, including the many that Saccheri found. The most startling is the theorem that the measures of the angles of a triangle do not add up to π (or 180°), as they do in Euclidean geometry. The goal in our study of hyperbolic geometry, Theorem 4.1.1, goes even further, relating the angle sum to the area of the triangle. Figure 4.2 illustrates that the greater the area of the triangle is, the smaller its angle sum is. (We state the theorem using radians to emphasize the analogy with Theorem 1.5.3. Otherwise we will measure angles in degrees.) Consequently, there is a constant k so that each triangle's area is less than k. Since the sides of triangles can become indefinitely long, this consequence seems paradoxical.

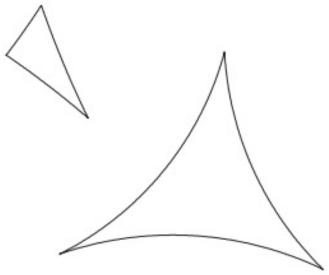

Figure 4.2 The angle sum of a hyperbolic triangle depends on its area.

Theorem 4.1.1. *In hyperbolic geometry the difference,* $\pi - (m\angle A + m\angle B + m\angle C)$, *between* π *and the angle sum of a triangle is proportional to the area of the triangle.*

Exercise 4.1.1.* Compare Theorem 4.1.1 with Theorem 1.5.3. Restate the theorem using degrees.

Theorem 4.1.1 suggests that we could decide which geometry is true by measuring real triangles. Since at least 1890, mathematicians and physicists have realized the futility of empirically deciding which geometry is correct because physical assumptions must be made to test mathematical relations. For example, we would have to assume that the path of a light ray is a straight line to measure even a moderately large triangle.

The first notable response to the advent of hyperbolic geometry came in 1854 when Georg Riemann (1826–1866) gave his introductory lecture to the faculty at Göttingen University. Only Riemann's mentor, the aging Gauss, apparently caught its point, entitled "On the Hypotheses which underlie Geometry." However, the talk, published after Riemann's death, focused geometric thought on a new field, differential geometry, and spurred an active debate on non-Euclidean geometries. Riemann had realized that the work of Gauss, Lobachevsky, and Bolyai was more than playing abstractly with a postulate. He recognized that the revised postulate implied that space had to be shaped differently than what Euclid's fifth postulate implied. He then articulated how infinitely many different geometries could be created, each with its own shape.

Differential geometry, the field pioneered by Riemann, Gauss, and others, investigates geometries by looking at how they behave in small regions and, in particular, how they curve. Gauss had shown how to measure the curvature of surfaces. Where the flat Euclidean geometry had zero curvature and a sphere had positive curvature, hyperbolic geometry has a uniform negative curvature, similar to the saddle shape of Figure 4.3. Riemann defined curvature in higher dimensions and envisioned geometries in any number of dimensions with changing curvatures throughout. The general theory of relativity builds on Riemann's work on the curvature of space. Einstein used a four-dimensional geometry (three spatial dimensions interacting with time) curved by the gravitational forces. Light waves travel along *geodesics*, paths of shortest length following the curvature of the surrounding space. Einstein's theory settled the nineteenth century question of the empirical truth of Euclidean and hyperbolic geometries in a surprising way: both are false! (See Chapter 9 as well as do Carmo [3], Kline [8], McCleary [9], Rucker [12], and Weeks [16] for further information.)

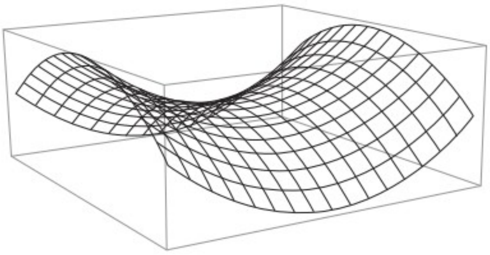

Figure 4.3 A saddle shape with negative curvature.

The abstraction throughout mathematics and the strangeness of the new geometries led mathematicians to search again for certainty in their mathematical arguments. No longer could they rely on an intuitive model to reveal the essential idea behind an argument. They examined the axioms for geometry, culminating in the modern axioms for Euclidean geometry, including Hilbert's (see Section 2.2). Hilbert also obtained an axiomatic system for hyperbolic geometry by replacing Playfair's axiom with the characteristic axiom above. (See Bonola [1] and Kline [8, Chapter 36] for further historical information.)

4.1.2 Models of Hyperbolic Geometry

In the latter part of the nineteenth century, various mathematicians developed models of hyperbolic geometry. We partially treat some of them here. Each has disadvantages, but they all help give a feel for the geometry. They are based in part on a Euclidean plane or space, with suitable interpretations of the undefined terms. For the Poincaré model, we first need a definition: Two (Euclidean) circles are *orthogonal* if and only if their radii at the circles' points of intersection are perpendicular, illustrated in Figure 4.4. In all the models the terms *on* and *between* have their usual meaning. In the language of Section 2.3, the formal models show the relative consistency of hyperbolic geometry based on Euclidean geometry.

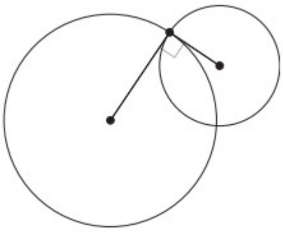

Figure 4.4 Orthogonal circles.

Informal Models. A saddle shape, as in Figure 4.3, provides a direct feel for a small part of hyperbolic geometry, although the negative curvature of actual saddles varies considerably from point to point. A physical version of this model has the advantage of facilitating actual measurements on its surface. Crocheted surfaces, as in Figure 4.5, developed by Daina Taimina, provide a more extended realization of approximately constant negatively curved surfaces. (See Henderson and Taimina [6, 49 ff].) The *pseudosphere*, represented in Figure 4.6 and discussed below, has constant negative curvature except on its "equator." In all the models, *lines* are interpreted as geodesics—intuitively shortest paths along the surface. Distances are measured along them and angles are measured using tangents to them where they meet. The surfaces give direct insight into hyperbolic geometry, at least in the region where the curvature is a constant negative.

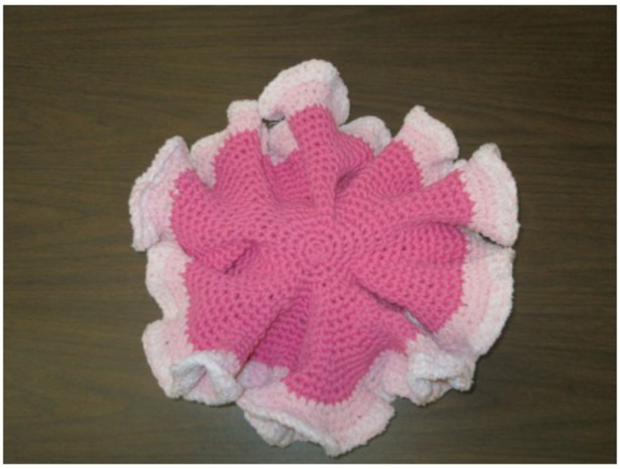

Figure 4.5 An extended crocheted surface with negative curvature.

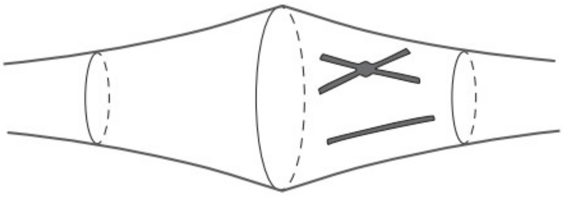

Figure 4.6 "pseudosphere" illustrating the characteristic axiom.

The Pseudosphere (1868). Ferdinand Minding (1806–1885) investigated this surface of constant negative curvature, shown in Figure 4.6, as early as 1840. Eugenio Beltrami (1835–1899) was the first to realize that it and all other surfaces of constant negative curvature provide (local) models of hyperbolic geometry. Beltrami's work suggested the relative consistency of hyperbolic geometry for the first time. The models now named for Felix Klein (1849–1925) and Henri Poincaré (1854–1912) resolved logical concerns about this geometry.

The Klein Model (1868). *Point* means a point in the interior of a Euclidean circle. *Line* means the portion of a Euclidean line in its interior. Figure 4.7 illustrates that the characteristic axiom holds in the model. Its biggest advantage is the matrix representation of its transformations, considered in Chapter 7. However, distances and angle measures are distorted and complicated in it.

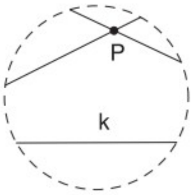

Figure 4.7 The characteristic axiom in the Klein model.

The Poincaré Model (1882). *Point* means a point in the interior of a Euclidean circle. *Line* means the portion interior to the circle of one of its diameters or of a Euclidean circle orthogonal to it. The chief advantage of this model is that angles are measured as they would be for curves in Euclidean geometry. In Figure 4.8 the curved sides of the triangle in bold show visually that the angle sum is less than 180°. The picture at the beginning of this chapter uses the Poincaré model. Because the angle measures are Euclidean, the repeated objects in this design are recognizably the same, although somewhat distorted to our eyes. Their hyperbolic size is the same, which gives a sense of how distances are measured using the same complicated formula as the Klein model. In Section 5.6 we consider transformations for this model and their connections with complex numbers. There are software programs, such as NonEuclid, that facilitate constructions in this model.

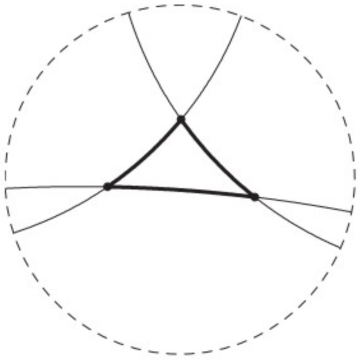

Figure 4.8 A triangle in the Poincaré model.

The Half-Plane Model (1882). *Point* means a point in the upper half of the Euclidean plane; that is, points (x, y) with $y > 0$. *Line* means the portion above the x-axis of either a vertical line or a circle with its center on the x-axis. In Figure 4.9, $\triangle ABC$ has an angle sum less than 180°, although that fact may not be as clear as in Figure 4.8. Henri Poincaré developed the model and the closely related Poincaré model at the same time. It measures angles as Euclidean angles and, in addition, its lines have equations that are easy to find. However, distances appear even more distorted than in the previous two models.

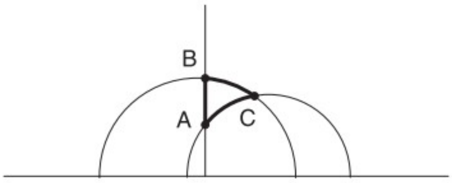

Figure 4.9 A triangle in the half plane model.

Unless explicitly stated, the figures in the rest of this chapter are not based on any of the models. This will encourage thinking of the geometry as an entire system, not an isolated exercise applying only to an artificial model. Many of the lines drawn in the figures will be curved so that they will not appear to intersect. This corresponds to one of the findings of Saccheri, mentioned earlier, that lines can approach one another in this geometry without intersecting.

4.1.3 Exercises for Section 4.1

Throughout the problems "line" refers to a hyperbolic line in the relevant model. Euclidean objects are identified as Euclidean.

4.1.2. Use a physical saddle-shaped surface, as in Figure 4.3, to explore basic properties of hyperbolic geometry. (The curves shown in that figure are not geodesics; they are the intersections of the surface with vertical planes to convey the three-dimensional curving.) An L-shaped piece of flexible plastic with straight sides and a right angle will lie on the surface closely enough to draw approximate perpendicular hyperbolic lines (geodesics). Use the protractor of Figure 2.16 to find the approximate measure of angles.

(a) Construct different sized triangles on the surface and measure their angle sums. Verify Theorem 4.1.1.

(b) Construct a quadrilateral $ABCD$ with right angles at A and B and with sides $\overline{AD}$ and $\overline{BC}$ congruent. Measure the angles at B and C. What can you say about them? Compare the lengths of $\overline{AB}$ and $\overline{CD}$.

(c) Make a large right triangle. Measure the lengths of its sides. Compare the squares of their lengths with the equation of the Pythagorean theorem (Theorem 1.1.2).

(d) Draw lines on the surface to illustrate the characteristic axiom of hyperbolic geometry.

*4.1.3 Use the Klein model of hyperbolic geometry to investigate how many lines through a point P do not intersect a line k that is not on P. Explain your answer.

4.1.4. Use the points inside the Euclidean unit circle $x^2 + y^2 = 1$ for the Poincaré model.

*(a) Show that the circle $C : (x - 1.25)^2 + y^2 = (0.75)^2$ is orthogonal to the unit circle. Graph the unit circle and the part of C in the unit circle on the same axes.

(b) We form a hyperbolic triangle with vertices $P = (0, 0)$, Q the intersection of the circle C with $y = (9/13)x$ (inside the unit circle), and R the intersection of the circle C with $y = -(9/13)x$ (inside the unit circle). Assume that $x > 0$. Verify that Q and R have x-coordinate 0.65. Find their y-coordinates. Include $\triangle PQR$ in the graph for part (a).

(c) Verify visually that the angle sum of the triangle in part (b) is less than 180°.

*(d) Find the measures of the angles of the triangle in part (b) and verify that their sum is less than 180°. Hints: The angle made by a Euclidean line and a Euclidean circle is the angle made by the Euclidean line and the tangent to the circle at their intersection. The slope of a Euclidean line equals $\tan\alpha$, where α is the angle that line makes with the x-axis. The arctangent function converts slopes to angles.

*4.1.5 Although three-dimensional Euclidean space satisfies the characteristic axiom, it doesn't satisfy all the axioms of plane geometry. Find which of Hilbert's axioms in Appendix C fail when we interpret a *plane* as three-dimensional Euclidean space and a *line* as a Euclidean line.

4.1.6. Use the half-plane model of hyperbolic geometry.

*(a) Find the equation of the line (Euclidean half circle) that is on the points $(1, 1)$ and $(4, 2)$.

(b) Explain why the equations of lines in the model are either $x = c$ or $y = \sqrt{r^2 - (x - c)^2}$, for appropriate constants c and r.

(c) Use a Euclidean argument about Euclidean circles and lines to explain why every two points in this model have one (hyperbolic) line that is on both of them.

*(d) Verify that the point $(-4, 10)$ is not on the line $y = \sqrt{25 - (x + 4)^2}$. Find the equations of two lines on $(-4, 10)$ that do not intersect $y = \sqrt{25 - (x + 4)^2}$. Graph the lines.

(e) Use a Euclidean argument to explain why the characteristic axiom always holds in this model. Hint: Consider the two kinds of lines separately.

*(f) Verify that the lines $y = \sqrt{25 - x^2}$ and $y = \sqrt{16 - (x - 1)^2}$ do not have a point of the model in common, although they intersect in a Euclidean point as curves in the Euclidean plane. Find the Euclidean point. We will call these lines *sensed parallels*.

*(g) Describe the type of intersection of the lines in part (f).

(h) Verify that every point in the model has at least one line that is a sensed parallel to the line $y = \sqrt{25 - x^2}$ in the sense that the only Euclidean intersection of the line with $y = \sqrt{25 - x^2}$ is the Euclidean point you found in part (f). Explain why, for a hyperbolic line k represented by a semicircle, that there are exactly two sensed parallels to k through a point not on k. How can we interpret sensed parallels so this property remains correct if a vertical line represents k?

4.1.7. (Calculus) In the half-plane model of hyperbolic geometry, consider the triangle formed by the lines $y = \sqrt{5 - (x - 3)^2}$, $y = \sqrt{10 - (x - 4)^2}$, and $x = 4$. Find the vertices of the triangle and graph it. Use the method of Exercise 4.1.4(d) to find the measures of the angles of this triangle. Hint: Use calculus and trigonometry. The angle sum is approximately 161.6°.

4.1.8. *(a) In the half plane model, verify that the line $y = \sqrt{16 - x^2}$ is perpendicular to the lines $y = \sqrt{9 - (x - 5)^2}$ and $y = \sqrt{9 - (x + 5)^2}$ at their points of intersection.

*(b) Find the intersections the line $y = \sqrt{34 - x^2}$makes with the lines $y = \sqrt{9 - (x \pm 5)^2}$ and the angles it makes with them. (See Exercise 4.1.4(d) and 4.1.7 for finding the angles.)

(c) Compare the angles in part (b) with Exercise 4.1.2(b).

4.1.9. A curious property of hyperbolic geometry is that two lines that do not intersect and are not sensed parallels (described informally in Exercise 4.1.6) have a common perpendicular. Use the half-plane model of hyperbolic geometry to illustrate this property with the lines $y = \sqrt{1 - x^2}$ and $x = 2$. Graph the lines first. Then use relevant Euclidean concepts, including orthogonal circles, to find the hyperbolic line perpendicular to both of them.

4.1.4 Carl Friedrich Gauss

Carl Friedrich Gauss (1777–1855) dominated mathematics during the first half of the nineteenth century, making fundamental contributions to virtually every area of mathematics. The prince of his small German state sponsored his education after learning of his prodigious abilities as a child. By age five he had found an arithmetic error in his father's accounts. In another story his grade school teacher made the class add the numbers from one to one hundred. After some thought Gauss found a formula and simply wrote down the correct answer, whereas his classmates toiled and made mistakes.

Gauss soon won acclaim as a mathematician. At 18 he developed the method of least squares, which is used extensively in statistics. A year later he constructed with straightedge and compass a regular 17-sided polygon and later characterized all constructible regular polygons, the first such advances since the ancient Greeks. He earned his Ph.D. at 22 by giving the first proof of the fundamental theorem of algebra, which says that every nth degree polynomial with real or complex coefficients has n roots, counting repetitions and these roots are all complex numbers. Two years later, he published his first major treatise on number theory, rejuvenating that ancient area. That same year, 1801, Gauss astonished astronomers by determining the orbit of the first asteroid, Ceres, based on only a few observations and after it had been lost from view owing to weather conditions. To do so he had to generalize his method of least squares. Gauss subsequently became a professor of astronomy and director of the observatory in Göttingen.

By 1800 Gauss had become convinced that Euclid's parallel postulate could not be proven, and in installments he developed what he called non-Euclidean geometry. Although none of this work was published until after his death, he did correspond with a number of mathematicians about it. Gauss also made seminal contributions to differential geometry, including the curvature of a surface and geodesics. He showed that Gaussian curvature determined all the properties of a surface not related to how it is placed in a surrounding space.

Gauss's contributions on complex numbers persuaded mathematicians that complex numbers were essential in mathematics. He represented complex numbers geometrically, as some others before him had. Additionally, he extended number theory to complex integers, and contributed fundamentally to what is now called complex analysis. He advanced knowledge in astronomy, magnetism, optics, and other applied fields.

Gauss's contemporaries revered him as the "Prince of Mathematicians," and he is widely considered the greatest mathematician since Newton. He was also one of the last who could claim to know all the mathematics in existence at his time. Gauss's work can be seen both as crowning the great expansion of mathematics in the sixteenth and seventeenth centuries and as igniting the explosion of specialized, abstract mathematics since then.

4.2 Properties of Lines and Omega Triangles

We develop hyperbolic geometry following the path of Saccheri, Gauss, Lobachevsky, and Bolyai, but add a modern foundation, due to Hilbert and others. We can thus accentuate the geometric intuition of earlier work while having the logical basis required to avoid the flaws discovered later.

Our undefined terms are *point*, *line, on, between,* and *congruent*. Our axioms are Hilbert's axioms for plane geometry (see Appendix C), with the exception of axiom IV-1 (Playfair's axiom), which we replace with the characteristic axiom of hyperbolic geometry. (See Section 4.1.) We also avail ourselves of Euclid's first 28 propositions, the ones he proved without using the parallel postulate. (See Appendix A.) They historically provided the key properties about lines and triangles needed to development of hyperbolic geometry. They are provable in Hilbert's axiomatic system without either Axiom IV-1 or the characteristic axiom. We will also make use of some of the SMSG postulates that hold in hyperbolic geometry. We assume the SMSG postulates in Appendix B, except 16, 20, 21, and 22. (They are false in hyperbolic geometry.)

Like the developers of hyperbolic geometry, we use proofs by contradiction for many theorems in this chapter, including Theorem 4.2.1. Until 1868, when Beltrami built the first model of hyperbolic geometry, mathematicians wondered whether the characteristic axiom itself was potentially a contradiction. The heavy use of proofs by contradiction increased mathematicians' unease before Beltrami's model. A rigorous proof of even Theorem 4.2.1 requires considerable development from Hilbert's axioms. For example, in Figure 4.10 we would need to show that B can always be chosen so that the line m enters $\triangle ABC$ at P. (By "enter" we mean that the line through P also has a point inside of the triangle on it, following Pasch's axiom.) The use of real numbers from the SMSG postulates does simplify the proof. Moise [10, Chapter 24] gives a careful development of the basics from Hilbert's axioms, including the justification of right- and left-sensed parallels in the definition below.

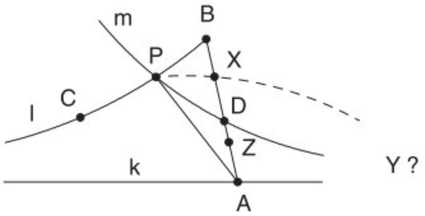

Figure 4.10 Constructing non-intersecting lines.

Theorem 4.2.1. *Given a point P not on a line k, there are infinitely many lines on P that have no points on k.*

Proof. Let the two lines indicated by the characteristic axiom be l and m. Pick points A on k and B and C on l such that line m enters $\triangle ABC$ at P. By Pasch's axiom, m intersects the triangle on another side, without loss of generality at D on $\overline{AB}$. SMSG postulate 3 permits us to match the points on $\overleftrightarrow{AB}$ with the real numbers, with 0 at B and 1 at D. Now, for every point X on the segment $\overline{BD}$, we can draw $\overleftrightarrow{PX}$. (See Figure 4.10.)

We claim that $\overleftrightarrow{PX}$ does not intersect k. Suppose, for a contradiction, that $\overleftrightarrow{PX}$ and k had the point Y in common. Then the line m would enter triangle $\triangle XYA$ at D, and by Pasch's axiom m would have to intersect the triangle again on $\overline{XY}$ or on $\overline{YA}$. However, m already intersects $\overleftrightarrow{PX}$ at P, eliminating $\overline{XY}$. Furthermore, $\overline{YA}$ is part of k, which has no point in common with m, which is a contradiction. The points X between B and D match with numbers between 0 and 1. Since there are infinitely many such numbers, there are infinitely many lines $\overleftrightarrow{PX}$ that have no point in common with k. ■

For the same k and P as in the preceding proof, the lines through P split into two categories, those that intersect k and those that do not. In Figure 4.10, for Z on the segment $\overline{AD}$, some of the lines $\overleftrightarrow{PZ}$, such as $\overleftrightarrow{PA}$, intersect k, but others don't, such as $\overleftrightarrow{PD}$. The points Z match with numbers greater than 1, but less than the number matched with A. By the continuity of the real numbers, some number ω must separate the numbers of points with lines intersecting k from those not intersecting k. Let W be the point matching ω, so W must separate these lines. Does $\overleftrightarrow{PW}$ intersect k? Draw a figure illustrating the following argument by contradiction showing that $\overleftrightarrow{PW}$ cannot intersect k. If T is the supposed intersection of $\overleftrightarrow{PW}$ and k, consider a point S on k with T between A and S. Now draw $\overleftrightarrow{PS}$, which clearly intersects k at S. Therefore $\overleftrightarrow{PW}$ would not be the last line, contradicting the assumption about W. A similar situation occurs on the left side of Figure 4.10.

Definition. Given a point P not on line k, the first line on P in each direction that does not intersect k is the (*right-* or *left-*) *sensed parallel* to k at P. Other lines on P that do not intersect k are called *ultraparallel* to k. Let A be on k with $\overline{AP} \perp k$. Call the smaller of the angles a sensed parallel makes with $\overline{AP}$ the *angle of parallelism at P*. (If the angles are equal, either is the angle of parallelism.)

The goal of this section has two parts. First, we start building inequalities leading to the big inequality in Theorem 4.1.1 about the angle sum of a triangle. Corollary 4.2.3, on the angle of parallelism, provides the starting point. Theorems 4.2.8 and 4.2.9 accomplish the other goal, showing that the angle of parallelism depends only on the length of the segment $\overline{AP}$.

Theorem 4.2.2. *If l and m are the two sensed parallels to k at P, they have the same angle of parallelism.*

Proof. As shown in Figure 4.11, let $\overline{AP} \perp k$ and $\overleftrightarrow{PB}$ and $\overleftrightarrow{PC}$ be the sensed parallels to k at P. For a contradiction, let the angles of parallelism differ, say, $m\angle APB < m\angle APC$. Construct inside $\angle APC$ a new angle $\angle APD \cong \angle APB$. Because $\overleftrightarrow{PC}$ is a sensed parallel, $\overleftrightarrow{PD}$ intersects k, say at E, giving a triangle $\triangle APE$. Let F be on k with $\overline{AF} \cong \overline{AE}$. Then $\triangle APF \cong \triangle APE$ by SAS (Euclid I-4). But $\angle APF \cong \angle APD \cong \angle APB$, which means that $\overleftrightarrow{PF}$ and $\overleftrightarrow{PB}$ are the same line by Hilbert's axiom III-4. However, $\overleftrightarrow{PB}$ is a sensed parallel to k, so it cannot intersect k at F, which is a contradiction. Hence the angles of parallelism must be the same. ■

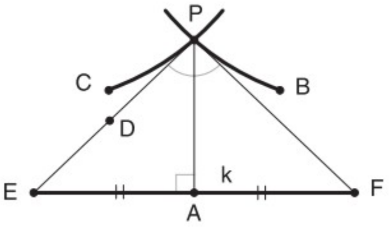

Figure 4.11 Angles of parallelism.

The proof of Theorem 4.2.2 may seem on first reading overly complicated. However, in order to use the Euclidean proposition SAS, we need "ordinary" triangles.

Corollary 4.2.3. *All angles of parallelism are less than a right angle. Two lines with a common perpendicular are ultraparallel.*

Proof. See Exercise 4.2.5. ■

Theorem 4.2.4. *Let P be a point not on k and let m be a sensed parallel to k at P. If S is any other point on m, then m also is a sensed parallel to k at S.*

Proof. Without loss of generality, let m be a right-sensed parallel to k at P.

Case 1. Suppose that S is on m to the left of P and U is on m with S between P and U. Because m and k do not intersect, m is either the right-sensed parallel to k at S or m is ultraparallel to k at S. For a contradiction, suppose that l, not m, is the right-sensed parallel to k at S, as in Figure 4.12. Pick T on l and on the opposite side of m from k. Then $\overleftrightarrow{TP}$ crosses m at

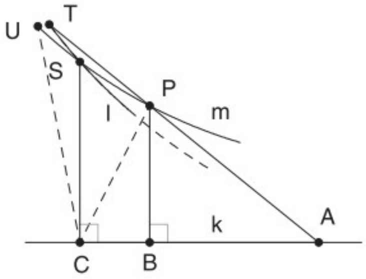

Figure 4.12 The sensed parallel to k at P is also sensed parallel at S.

P. As m is the right-sensed parallel to k at P, $\overleftrightarrow{TP}$ must intersect k, say at A. Let U be on m with S between P and U and C on k with $\overline{SC} \perp k$. Now l enters $\triangle UPC$ at S. Since l is right-sensed parallel, it must exit on $\overline{PC}$, rather than $\overline{UC}$. In turn l enters $\triangle CPB$ on $\overline{CP}$ and, by Pasch's axiom, l would need to intersect either $\overline{CB}$ or $\overline{PB}$. But l is a right-sensed parallel of $k = \overleftrightarrow{CB}$, so l intersects $\overline{PB}$. Finally l enters $\triangle PBA$. Again by Pasch's axiom, either l intersects $\overline{BA}$, which is k, or it intersects $\overline{PA}$. As before, k can't intersect l. However, l already intersected $\overline{PA}$ at T, so that also is impossible. We thus have a contradiction from assuming l was the right-sensed parallel. Hence m is the right-sensed parallel to k at S.

Case 2. Let S be on m to the right of P. (See Exercise 4.2.6.) ■

4.2.1 Omega Triangles

In both the Klein model (Figure 4.13) and the Poincaré model (Figure 4.14) of hyperbolic geometry m and n, the sensed parallels to k at P, "meet" k on the circular boundary. Only points interior to the boundary qualify as points in these models. Following the historical development, we say that two sensed parallels meet at an imaginary point called an *omega point*. Although the originators of hyperbolic geometry didn't have models in the modern sense, they benefited greatly from thinking of sensed parallels "meeting at infinity." They used *omega triangles*, which have one omega point, to prove theorems about ordinary triangles. They first showed, as do we here, that omega triangles share some key properties with regular triangles. Figures 4.15 and 4.16 illustrate an omega triangle in the Klein and Poincaré models with one vertex (labelled Π)

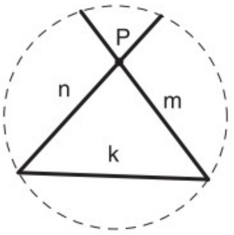

Figure 4.13 Sensed parallels in the Klein model.

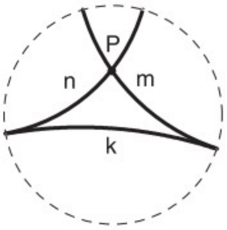

Figure 4.14 Sensed parallels in the Poincaré model.

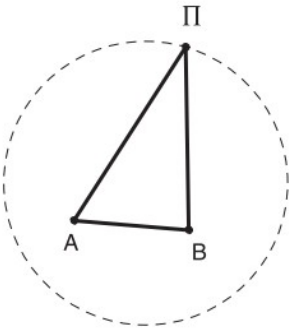

Figure 4.15 An omega triangle in the Klein model.

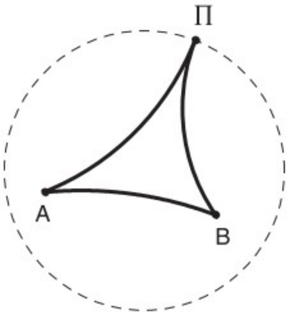

Figure 4.16 An omega triangle in the Poincaré model.

9

Differential Geometry

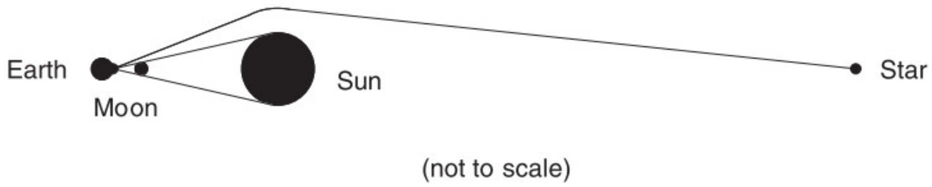

Figure 9.0 In 1919 Sir Arthur Eddington (1882–1944) spectacularly confirmed a key prediction of Einstein's general theory of relativity during a solar eclipse. Astronomers had determined that the sun should have blocked a starlight's path if it followed a Euclidean straight line from the star to earth. Instead because the sun's gravity distorted the shape of space, Eddington saw the star when the moon blocked the sun's light during the eclipse. The figure greatly distorts sizes to illustrate the setting of this famous experiment. In the language of differential geometry, the light from the star followed a geodesic—a locally shortest path. The gravitational field of the sun induced enough curvature to space to make an observable difference between the geodesic and a Euclidean line. While relativity theory is beyond the level of this chapter, we will investigate curvature and geodesics.

9.1 Overview and History

A plane flying from Tokyo to Los Angeles flies far north of either city, depicted in Figure 9.1. However, the two cities lie at virtually the same latitude. Why don't planes fly straight east from Tokyo to Los Angeles? Actual airline routes minimize the distance on the curved surface of the earth. A straight eastward route follows a circle of latitude, which doesn't minimize distances (except the equator). Differential geometry provides the language and theory to understand the geometry of curved surfaces. We will investigate how to find the curvature of a surface and geodesics, which are the locally straight paths on them.

Intuitively, Euclidean straight lines do not bend at all, while circles bend at constant rates depending on their radii. (See Figure 9.2.) The ancient Greeks realized that other curves, such as the parabola in Figure 9.3, bend more at some points (for example, A) than others (such as B or C). However, real advances in defining and measuring curvature, the amount of bending, needed the language of calculus. By 1671 Sir Isaac Newton (1642–1727), one of the originators of calculus, had already figured out the correct formulation for the curvature of a plane curve. (See Theorem 9.2.1.) However, Newton delayed publishing that work and much else for decades. By that time Gottlieb Wilhelm Leibniz (1646–1716), the other founder of calculus, and his followers

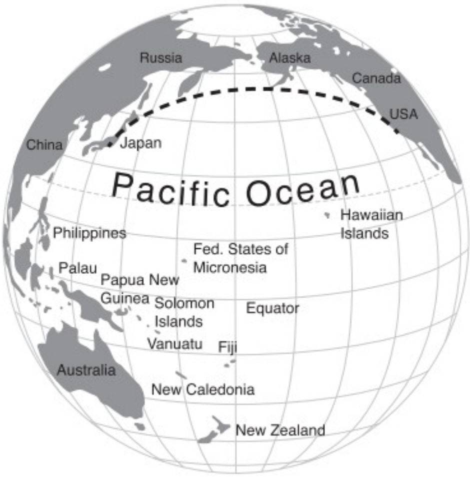
Russia
Alaska
Canada
USA
China
Japan
Pacific Ocean
Hawaiian Islands
Philippines
Fed. States of Micronesia
Palau
Papua New Guinea
Solomon Islands
Equator
Vanuatu
Fiji
Australia
New Caledonia
New Zealand

Figure 9.1 A polar air route.

had published on many of the same topics, including curvature. Jakob Bernoulli (1654–1705), a follower of Leibniz, determined geodesics on surfaces of revolution in 1698.

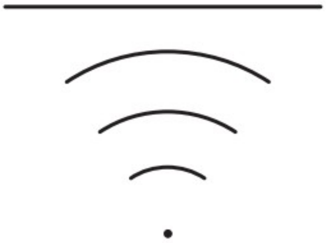

Figure 9.2

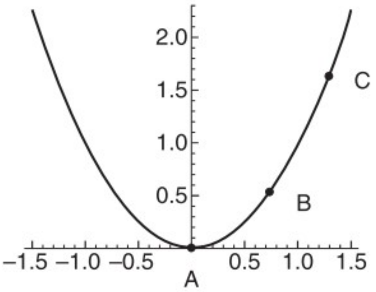

Figure 9.3 Curvature on a parabola.

In the 1730s Alexis-Claude Clairaut (1713–1765) and Leonhard Euler (1707–1783) considered space curves, such as the helix of Figure 9.4. They analyzed how much such curves twisted out of a plane, now called the *torsion* of a curve. (See Project 1.) More central to our focus, they and others investigated general surfaces and geodesics.

The inherent distortions of a flat map of portions of a curved earth motivated many mathematicians from Euler to Carl Friedrich Gauss (1777–1855). In fact, as part of his employment, Gauss spent some years surveying the state of Hanover, which spurred his profound investigations in differential geometry. His major paper in 1827 shaped the direction of this subject. Gauss studied a space on its own terms, not dependent on how it is embedded in Euclidean space. For example, measurements of a portion of the surface of the earth could determine its curvature there. One of his profound theorems (Theorem 9.4.3) generalizes Theorem 1.5.3 and Theorem 4.1.1. They link the area of a triangle in spherical and hyperbolic geometry to their angle sums. In the generalization Gauss used geodesics for the sides of the triangles. Then, as

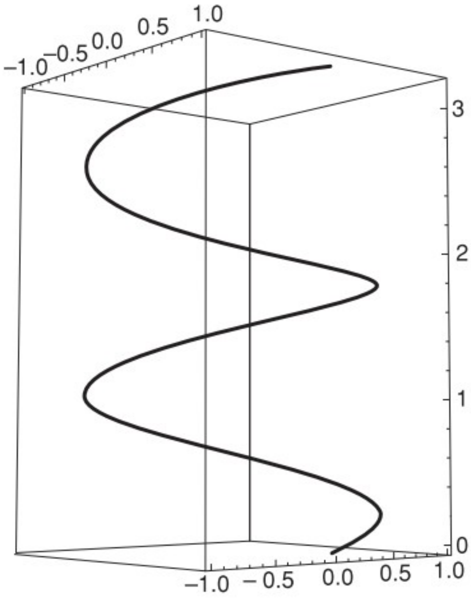

Figure 9.4 A helix.

he proved, the difference between the angle sum of the triangle and π was determined by the area of the triangle and the curvature within it. The sphere has constant positive curvature and the hyperbolic plane has constant negative curvature, accounting for the difference in Theorems 1.5.3 and 4.1.1.

In 1854 Georg Riemann (1826–1866) generalized Gauss's results to any number of dimensions and so transformed differential geometry again. Many mathematicians developed Riemann's insights into rigorous mathematics with computational methods. Albert Einstein (1879–1955) built on Riemann's profound insights to revolutionize physics with the general theory of relativity. Einstein needed a four-dimensional space built from three spatial dimensions and time, all with varying curvature corresponding to the strength of gravitational fields at different points.

9.1.1 Exercises for Section 9.1

*9.1.1. Use a computer graphing capability to find the circle best approximating the bend of $y = x^2$ at $x = 0$ as follows.

(a) Explain why the circle should have a common tangent with $y = x^2$ at $x = 0$. On what line does the center of the circle lie? Use your answers to simplify the general equation of a circle: $(x - a)^2 + (y - b)^2 = r^2$.

(b) Solve the equation in part (a) for y and graph the bottom half along with $y = x^2$ for various radii. Which center and radius appear to give the best fit?

9.1.2. *(a) Repeat Exercise 9.1.1 for the function $y = \cos(x)$ at $x = 0$.

(b) Repeat Exercise 9.1.1 for the function $y = x^4$ at $x = 0$.

9.3 Surfaces and Curvature

The bulging of a sphere makes the angle sum of spherical triangles greater than 180°. Indeed, the excess of the angle sum over 180° depends on the area of the triangle, as Theorem 1.5.3 makes precise. The angle sum of a hyperbolic triangle also depends on the triangle's area. But, as proven in Theorem 4.1.1, the angle sum is less than 180°. What kind of anti-bulging would a surface need to have to match hyperbolic geometry? Carl Friedrich Gauss, one of the founders of hyperbolic geometry, answered that question with differential geometry, although it is unclear whether he saw the connection. From Gauss's definition of curvature, a sphere has constant positive curvature and a Euclidean plane has 0 curvature. Eugenio Beltrami (1835–1900) realized that a surface with constant negative Gaussian curvature gave a model of (part of) hyperbolic geometry. In this section we investigate surfaces and their curvatures. To simplify the theory and calculations we will use parametric equations to embed the surfaces in $\mathbb{R}^3$, Euclidean 3-dimensional space. (Gauss showed how to study the geometry of a surface independently of embedding it in $\mathbb{R}^3$.) We focus on surfaces of revolution, further easing our work. We often use the formula $\cos^2(\theta) + \sin^2(\theta) = 1$ to simplify formulas for surfaces of revolution.

9.3.1 Surfaces

Example 1. We can represent the surface of a sphere of radius R in more than one way. The pair of functions $z = \pm\sqrt{R^2 - x^2 - y^2}$ determine the upper and lower hemispheres. The computer drawing in Figure 9.15 uses the curves determined from them by holding one of x or y fixed and varying the other. The familiar and more useful longitude and latitude coordinates, depicted in Figure 9.16, fit with the parametric equations $x(u, v) = R\cos(u)\cos(v)$, $y(u, v) = R\sin(u)\cos(v)$, and $z(u, v) = R\sin(v)$, where $-\pi < u \leq \pi$ and $-\frac{\pi}{2} \leq v \leq \frac{\pi}{2}$. Here u gives the longitude and v gives the latitude. The equations come from rotating the circle $\overrightarrow{c(v)} = (R\cos(v), R\sin(v))$ around the z-axis. If we restrict v to $-\frac{\pi}{2} < v < \frac{\pi}{2}$, they map the points (u, v) in a rectangle to the points $(x(u, v), y(u, v), z(u, v))$ on the sphere, minus the north and south poles. The latitudes and longitudes in Figure 9.16 come from holding v or u fixed and varying the other. This parameterization doesn't work well at the north and south poles, which are the images of infinitely many points $(u, \frac{\pi}{2})$ and $(u, -\frac{\pi}{2})$, respectively. Both representations of the sphere present difficulties at the edges of the domain—where $z = 0$ for the first one and where $u = \pm\pi$ or $v = \pm\frac{\pi}{2}$ for the second one. A more rigorous development of differential geometry covers a surface with overlapping patches, each with its own parameterization. (See Do Carmo [1, 55–57].) ◊

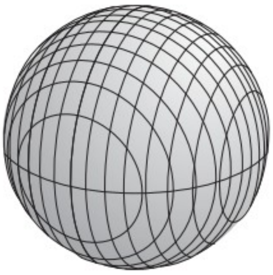

Figure 9.15 A sphere represented using cross sections.

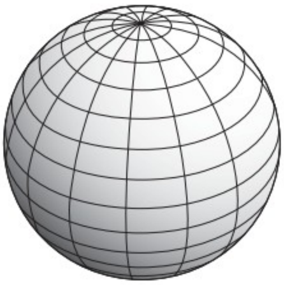

Figure 9.16 Longitude and latitude lines on a sphere.

Exercise 9.3.1. Replace x and y in the first form for the sphere with their parametric formulas and reduce the expression to $z(u, v)$ to verify that the representations of a sphere match.

Exercise 9.3.2. Most texts parameterize the sphere as $x(u, v) = R\cos(u)\sin(v)$, $y(u, v) = R\sin(u)\sin(v)$, and $z(u, v) = R\cos(v)$, where $0 < u \leq 2\pi$ and $0 \leq v \leq \pi$. Convert it to the second one in Example 1.

Example 2. We generate a *torus*, or more informally an inner tube, by rotating the circle $(y - R)^2 + z^2 = r^2$ in the yz-plane around the z-axis. (See Figures 9.17 and 9.18.) The rotation replaces the y-value in the equation with corresponding points (x, y). The distance of (x, y) from the origin, $\sqrt{x^2 + y^2}$, takes the place of y, giving the formula $z = \pm\sqrt{r^2 - (\sqrt{x^2 + y^2} - R)^2}$.

Figure 9.19 depicts the torus as a surface of revolution using cross sections from the parametric form $((r\cos(v) + R)\cos(u), (r\cos(v) + R)\sin(u), r\sin(v))$, for $0 \leq u < 2\pi$ and $-\frac{\pi}{2} \leq v < \frac{3\pi}{2}$.

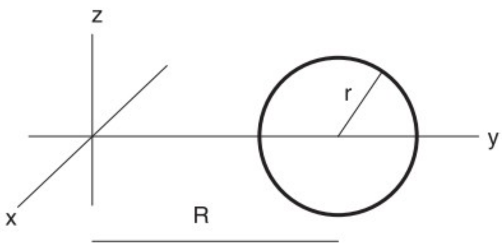

Figure 9.17

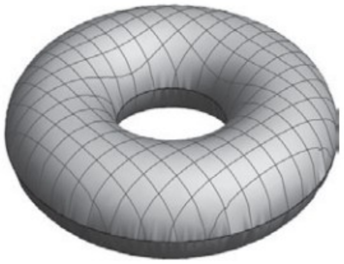

Figure 9.18

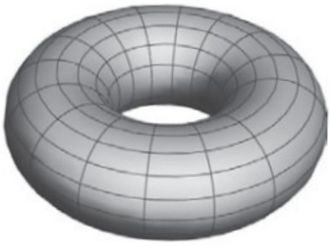

Figure 9.19

Points on the bulging outside surface of the torus, where $-\frac{\pi}{2} < v < \frac{\pi}{2}$, have positive curvature. Points on the inside surface, where $\frac{\pi}{2} < v < \frac{3\pi}{2}$, have negative curvature. For them the surface has a saddle-like shape, curving inward for the horizontal cross section and outward for the vertical cross section. Consider the two circles of points when $v = \pm\frac{\pi}{2}$, between the bulging points and the saddle points. Although the torus may not appear flat at them, its curvature can't be positive or negative, and so must be 0. The situation is analogous to Example 3 of Section 9.2, where $y = x^4$ had a curvature of 0 at $x = 0$. ◊

Exercise 9.3.3. Replace x and y in the first form for the torus with the parametric formulas and reduce the expression to $z(u, v)$ to verify that the two representations of a torus match.

In general, a surface is a function $\mathbf{s}$ from a region of $\mathbb{R}^2$ to $\mathbb{R}^3$ that we represent using parametric equations: $\mathbf{s}(u, v) = (x(u, t),\ y(u, v),\ z(u, v))$. There are many conditions on them to make the patches smooth enough and able to fit together. We'll avoid the technical details and say $\mathbf{s}$ is a *smooth surface* provided it satisfies the conditions found in Oprea [7, 71–73].

9.3.2 Curvature

The curvature of a surface at a point is more subtle than the curvature of curves. We use curves on the surface and through the point, in particular the cross sections, as discussed in Examples 1 and 2, to lead us to a definition of surface curvature. When we fix v, the values of $\mathbf{s}(u, v)$ vary with u and so the cross section curve has tangent vectors $\mathbf{s}_u = (\frac{\partial x}{\partial u}, \frac{\partial y}{\partial u}, \frac{\partial z}{\partial u})$. Similarly, when we fix u, the tangent vectors are $\mathbf{s}_v = (\frac{\partial x}{\partial v}, \frac{\partial y}{\partial v}, \frac{\partial z}{\partial v})$. From multivariable calculus other tangent vectors to the surface at this point are linear combinations of them and so form a plane. However, it is easier to keep track of the *unit normal*, a unit vector orthogonal to the tangent plane. The cross product $\mathbf{s}_u \times \mathbf{s}_v$ gives us a vector orthogonal to both tangent vectors and so it is orthogonal to the entire tangent plane. The definition of $N(u, v)$ divides the cross product by its length to get a unit vector. We'll use the change of the unit normal to determine how the surface bends. This is analogous to (9.4) which gave the curvature of a curve in terms of the change of the unit tangent vector.

Definition. For a smooth surface $\mathbf{s}(u, v)$, the *unit normal* is $N(u, v) = \frac{\mathbf{s}_u \times \mathbf{s}_v}{\|\mathbf{s}_u \times \mathbf{s}_v\|}$.

Example 3. For the sphere of Example 1, $\mathbf{s}(u, v) = (R\cos(u)\cos(v),\ R\sin(u)\cos(v),\ R\sin(v))$. Then

$$\begin{aligned} \mathbf{s}_u &= (-R\sin(u)\cos(v),\ R\cos(u)\cos(v), 0) \\ \text{and } \mathbf{s}_v &= (-R\cos(u)\sin(v),\ -R\sin(u)\sin(v),\ R\cos(v)). \end{aligned}$$

We find

$$\mathbf{s}_u \times \mathbf{s}_v = R^2(\cos(u)\cos^2(v),\ \sin(u)\cos^2(v),\ \cos(v)\sin(v)).$$

From the fact that $\cos(v) > 0$ for $-\frac{\pi}{2} < v < \frac{\pi}{2}$ we find

$$\begin{aligned} \|\mathbf{s}_u \times \mathbf{s}_v\| &= R^2\sqrt{\cos^2(u)\cos^4(v) + \sin^2(u)\cos^4(v) + \cos^2(v)\sin^2(v)} \\ &= R^2\sqrt{\cos^4(v) + \cos^2(v)\sin^2(v)} = R^2\sqrt{\cos^2(v)} = R^2\cos(v). \end{aligned}$$

Thus $N(u, v) = (\cos(u)\cos(v), \sin(u)\cos(v), \sin(v))$. The vectors point directly out from the sphere at each point $\mathbf{s}(u, v)$. ◊

Exercise 9.3.4.* Find $N_u = \frac{\partial N}{\partial u}$ and $N_v = \frac{\partial N}{\partial v}$ and verify that $N_u = \frac{1}{R}\mathbf{s}_u$ and $N_v = \frac{1}{R}\mathbf{s}_v$.

Example 4. Exercise 9.3.16 shows that the normal to the torus in Example 2 is $N(u, v) = (\cos(u)\cos(v), \sin(u)\cos(v), \sin(v))$. To determine the curvature of the surface at a point (a, b) we need to know more than the normal. We need to know how it is changing as (u, v) approaches (a, b) along the cross sections. Exercise 9.3.16 gives formulas for N_u and N_v and shows that $N_u = \frac{\cos(v)}{R + r\cos(v)}\mathbf{s}_u$ and $N_v = \frac{1}{r}\mathbf{s}_v$. The ratios indicate how quickly the normal is changing in the u and v directions. We can find the curvature of the curves of the cross sections from them. ◊

In a more complicated but analogous manner to Theorem 9.2.2 the partial derivatives N_u and N_v are linear combinations of $\mathbf{s}_u$ and $\mathbf{s}_v$. That is, there is a 2×2 matrix M with entries possibly dependent on u and v so that $M\begin{bmatrix} \mathbf{s}_u \\ \mathbf{s}_v \end{bmatrix} = \begin{bmatrix} N_u \\ N_v \end{bmatrix}$. In Example 3, $M = \begin{bmatrix} 1/R & 0 \\ 0 & 1/R \end{bmatrix}$ and in Example 4, $M = \begin{bmatrix} \frac{\cos(v)}{R+r\cos(v)} & 0 \\ 0 & \frac{1}{r} \end{bmatrix}$. For a surface $\mathbf{s}(u, v)$, M is called the *shape operator*. While the terms of M depend on the parameterization of the surface, the Gaussian curvature defined below is independent of it. (See Oprea [7, 83–105] for a more extensive exposition of shape operators and their connection with curvature.)

Definition. For a point (p, q) on the surface $\mathbf{s}(u, v)$, the *Gaussian curvature* is $\kappa(p, q) = \det M$, where M is the shape operator of $\mathbf{s}(u, v)$ at (p, q).

Exercise 9.3.5. Use Exercise 9.3.4 to verify that at a point on a sphere of radius R the curvature is $1/R^2$.

Example 5. The curvature of the torus of Examples 2 and 4 is $\kappa(u, v) = \det M = \frac{\cos(v)}{r(R+r\cos(v))}$. For $-\frac{\pi}{2} < v < \frac{\pi}{2}$ the torus bulges out and so should have a positive curvature, matching $\cos(v) > 0$ and $\kappa(u, v) > 0$ for these values. For $v = 0$, $\kappa(u, 0) = \frac{1}{r(R+r)}$, the product of the curvatures of the circles that are the cross sections for fixed u and v. Similarly, for $\frac{\pi}{2} < v < \frac{3\pi}{2}$, the torus has a saddle shape and negative curvature, matching the negative value of $\cos(v)$ and so $\kappa(u, v)$. Also $\kappa(u, \pi) = \frac{-1}{r(R-r)}$, which is the negative of the product of the curvatures of the cross sections for fixed u and v. Finally, $\cos(\pm\frac{\pi}{2}) = 0$, giving a curvature $\kappa(u, \pm\frac{\pi}{2}) = 0$, as discussed in Example 2. ♢

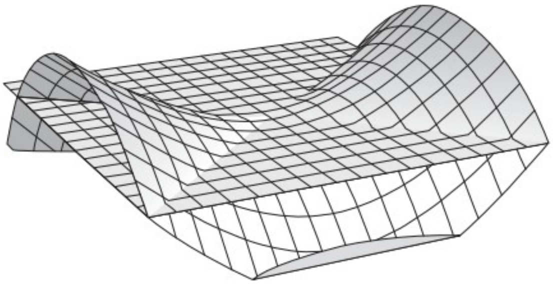

Figure 9.20 A tangent plane at a saddle point.

We can interpret the curvature of a surface in terms of "signed curvatures" for certain cross sections, generalizing Example 5. As noted in that example, the curvature of the torus at (u, π) is negative. As Figure 9.20 indicates, at a saddle point, the tangent plane cuts through the surface. Thus the osculating circles for the cross sections fixing u and v have centers on opposite sides of the tangent plane at (u, π). We pick one side of the tangent plane to carry a positive sign for curvature of curves and the other side to carry a negative sign. For points in the inside, saddle shaped part, the product of the signed curvatures will be negative. For points where the torus bulges outward, the signed curvatures have the same sign and so the product is positive. Euler noticed this relationship using the cross sections to a smooth surface through the normal at a point. Gauss proved that his definition of curvature at a point of a surface was the product of the maximum and minimum (signed) curvatures for curves formed from the cross sections.

9.3.3 Surfaces of Revolution

We derive the general equation of a surface of revolution and verify its curvature formula. Given a parametric curve $(f(v), g(v))$, as in Figure 9.21, we rotate it around the axis of the second coordinate (z) to get a surface of revolution, depicted in Figure 9.22. The radius $f(v)$ is split into its x and y components: $x = f(v)\cos(u)$ and $y = f(v)\sin(u)$. (Since $f(v)$ is a radius, we assume that $f(v) \geq 0$.) Thus the surface is $\mathbf{s}(u, v) = (f(v)\cos(u), f(v)\sin(u), g(v))$.

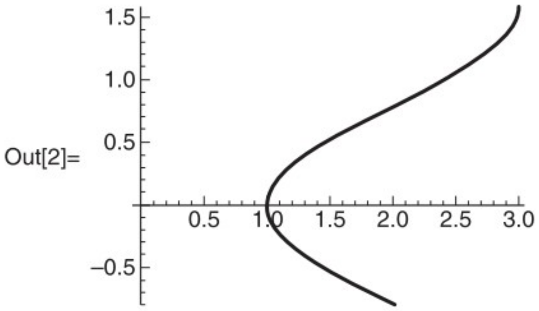

Figure 9.21 A parametric curve.

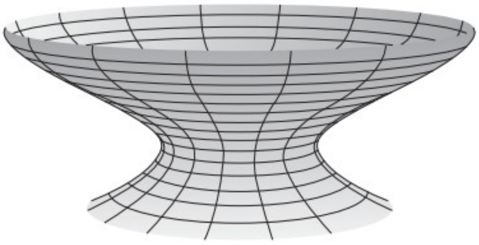

Figure 9.22 The corresponding surface of revolution.

Exercise 9.3.6. Verify the following calculations.

$$\mathbf{s}_u = (-f(v)\sin(u), f(v)\cos(u), 0)$$

$$\mathbf{s}_v = (f'(v)\cos(u), f'(v)\sin(u), g'(v))$$

$$\mathbf{s}_u \times \mathbf{s}_v = (f(v)g'(v)\cos(u),\ f(v)g'(v)\sin(u),\ -f(v)f'(v))$$

$$\|\mathbf{s}_u \times \mathbf{s}_v\| = f(v)\sqrt{f'(v)^2 + g'(v)^2}$$

$$N(u, v) = \left(\frac{g'(v)\cos(u)}{\sqrt{f'(v)^2 + g'(v)^2}}, \frac{g'(v)\sin(u)}{\sqrt{f'(v)^2 + g'(v)^2}}, \frac{-f'(v)}{\sqrt{f'(v)^2 + g'(v)^2}}\right)$$

$$N_u = \left(\frac{-g'(v)\sin(u)}{\sqrt{f'(v)^2 + g'(v)^2}}, \frac{g'(v)\cos(u)}{\sqrt{f'(v)^2 + g'(v)^2}}, 0\right) \quad N_u = \left(\frac{g'(v)}{f(v)\sqrt{f'(v)^2 + g'(v)^2}}\right)\mathbf{s}_u.$$

In addition to the computations of Exercise 9.3.6, we also need N_v to determine the curvature of a surface of revolution. We can simplify the complicated expression for N_v using the substitution $K = \frac{g''(v)f'(v)-g'(v)f''(v)}{(f'(v)^2+g'(v)^2)^{3/2}}$. In (9.3) K is the signed curvature of the parametric curve $(f(v), g(v))$, the curve we are revolving to make the surface.

Exercise 9.3.7. Verify $\left(\frac{g'(v)}{\sqrt{f'(v)^2+g'(v)^2}}\right)' = f'(v)K$ and $\left(\frac{-f'(v)}{\sqrt{f'(v)^2+g'(v)^2}}\right)' = g'(v)K$.

From Exercise 9.3.7, $N_v = K(f'(v)\cos(u), f'(v)\sin(u), g'(v)) = K\mathbf{s}_v$. The preceding work shows that the shape operator for a surface of revolution is $M = \left[\begin{smallmatrix} J & 0 \\ 0 & K \end{smallmatrix}\right]$, where $J = \frac{g'(v)}{f(v)\sqrt{f'(v)^2+g'(v)^2}}$ and K is given above. Now the proof of Theorem 9.3.1 only requires multiplying J and K.

Theorem 9.3.1. *Suppose the surface of revolution* $\mathbf{s}(u, v) = (f(v)\cos(u), f(v)\sin(u), g(v))$ *is a smooth surface. Then the curvature of* $\mathbf{s}(u, v)$ *is* $\kappa(u, v) = \frac{g'(v)(g''(v)f'(v)-g'(v)f''(v))}{f(v)(f'(v)^2+g'(v)^2)^2}$.

Exercise 9.3.8. Verify that $\mathbf{s}_u$ and $\mathbf{s}_v$ are orthogonal for surfaces of revolution.

9.3.4 Exercises for Section 9.3

Use computer software to make the requested graphs.

9.3.9. *(a) Explain why $\mathbf{s}(u, v) = (R\cos(u), R\sin(u), v)$ is a cylinder. Explain what curve is rotated around the z-axis to generate it.

(b) Find its curvature. Explain why the formula is reasonable.

(c) Verify that the helices of Example 6 of Section 9.2 are curves on the surface of the cylinder.

9.3.10. *(a) From Example 1 we can represent the sphere as $\mathbf{s}(x, y) = (x, y, \pm\sqrt{R^2 - x^2 - y^2})$. Find $\mathbf{s}_x$, $\mathbf{s}_y$, and $N(x, y)$. Are $\mathbf{s}_x$ and $\mathbf{s}_y$ orthogonal? Are N_x and N_y linear combinations of $\mathbf{s}_x$ and $\mathbf{s}_y$? This part illustrates some advantages of surfaces of revolution.

(b) Graph the surface $\mathbf{s}(u, v) = (u, v, u^2 - v^2)$ for $-2 \le u \le 2$ and $-2 \le v \le 2$. Describe what you think the curvature of this surface is at various points.

(c) At $(0, 0)$ the curvature of the surface of part (b) is the negative of the product of the curvatures of the cross sections when $u = 0$ and when $v = 0$. Find the curvatures of the cross sections and so the curvature of the surface at $(0, 0)$. What does the value tell you about the curvature of the surface at other points?

9.3.11. *(a) Graph the surface $(u, v, \cos(u)\cos(v))$ for $-\pi \le u \le \pi$ and $-\pi \le v \le \pi$. Describe what you think the curvature of this surface is at various points.

(b) At points $(p\pi, q\pi)$ the curvature of the surface of part (a) is the product of the curvatures of the cross sections when $u = p\pi$ and when $v = q\pi$, for integers p and q. Find the curvatures of the cross sections and so the curvature of the surface at $(p\pi, q\pi)$.

(c) Verify that the curvatures of the cross sections $u = \frac{\pi}{2}$ and $v = \frac{\pi}{2}$ are 0. Explain from the graph of part (a) why the curvature at $(\frac{\pi}{2}, \frac{\pi}{2})$ is not 0.

(d) The biggest (positive and negative signed) curvatures at $(\frac{\pi}{2}, \frac{\pi}{2})$ occur in cross sections bisecting the angles between $u = \frac{\pi}{2}$ and $v = \frac{\pi}{2}$. Explain why the cross sections have equations $z = \cos^2(u)$ and $z = -\cos^2(u)$. Find their curvatures and so the curvature of the surface in part (c) at $(\frac{\pi}{2}, \frac{\pi}{2})$.

9.3.12. The surface of a paraboloid results from rotating a parabola about its axis. The parametric form is $\mathbf{s}(u, v) = (v\cos(u), v\sin(u), bv^2)$.

(a) Explain how to derive the parametric form of the paraboloid from the equation of a parabola.

(b) Graph the paraboloid with $b = 1$. How does increasing the value of b affect its shape? Decreasing b?

(c) Verify that the curvature of a paraboloid is $\kappa(u, v) = \frac{4b}{4b^2v^4+1}$. For $b = 1$ graph the osculating sphere at $v = 0$ with the graph in part (b).

(d) Another way to represent a paraboloid is $\mathbf{s}(x, y) = (x, y, bx^2 + by^2)$. Find $\mathbf{s}_x$, $\mathbf{s}_y$, and $N(x, y)$. Are $\mathbf{s}_x$ and $\mathbf{s}_y$ orthogonal? Are N_x and N_y linear combinations of $\mathbf{s}_x$ and $\mathbf{s}_y$? This part illustrates some advantages of surfaces of revolution.

9.3.13. (a) Graph $\mathbf{s}(u, v) = (v\cos(u), v\sin(v), v^4 - 2v^2)$ for $0 \le u \le 2\pi$ and $0 \le v \le \sqrt{2}$.

(b) Describe the values of v for which you think the curvature of the surface in part (a) is positive. Repeat for negative curvature.

(c) Graph the cone with formula $\mathbf{c}(u, v) = (v\cos(u), v\sin(u), mv)$ for $0 \le u \le 2\pi$ and $0 \le v \le \sqrt{2}$. Explain why the surface is not smooth when $v = 0$.

(d) Assume $v > 0$ and find the curvature of the surface in part (c). Explain why this curvature is reasonable. What happens at $v = 0$?

9.3.14. (a) Write the equations for the surface of revolution for the curve $\overrightarrow{c(t)} = (t, \cos(t))$.

(b) Graph the surface in part (a) with $0 \le u \le 2\pi$ and $0 \le v \le 3\pi$. For which values of v does the curvature appear positive? negative?

(c) Assume that the curvature of the surface in part (a) is $\kappa(u, v) = \frac{\sin(v)\cos(v)}{v(1+\sin^2(v))^2}$ for $v > 0$. Use a limit to determine $\kappa(u, 0)$. For which values of v is $\kappa(u, v) > 0$? Repeat for $\kappa(u, v) < 0$. Compare the values with your answers in part (b).

(d) At $t = 0$, $cos(t)$ has a derivative of 0, so the surface of revolution is smooth when $v = 0$. Investigate as in parts (a) and (b) the surface $\mathbf{g}(u, v)$ made by rotating $(t, \sin(t))$. What happens when $v = 0$? Graph the surface $\mathbf{g}(u, v)$ for $0 \le u \le 2\pi$ and $0 \le v \le 2\pi$.

9.3.15. Figure 9.23 depicts an ellipse, which has parametric form $\overrightarrow{c(t)} = (\cos(t), a\sin(t))$. When we rotate the ellipse around the vertical axis, we get a *spheroid*. When $a > 1$, we call it a *prolate* spheroid, like a rounded football. When $a < 1$, as the ellipse in Figure 9.23, the spheroid is *oblate*. The earth is, as Newton predicted, slightly oblate, with the diameter at the equator less than 1% bigger than the distance between the poles.

*(a) Find the parametric form for the surface $\mathbf{s}(u, v)$ of a spheroid. Graph the spheroid with $a = 2$. Repeat with $a = \frac{1}{2}$.

*(b) Find $\mathbf{s}_u$, $\mathbf{s}_v$, and $\mathbf{s}_u \times \mathbf{s}_v$.

(c) Verify that $\|\mathbf{s}_u \times \mathbf{s}_v\| = \cos(v)\sqrt{a^2\cos^2(v)+\sin^2(v)}$ and $N(u,v) = \left(\frac{a\cos(u)\cos(v)}{\sqrt{a^2\cos^2(v)+\sin^2(v)}}, \frac{a\sin(u)\cos(v)}{\sqrt{a^2\cos^2(v)+\sin^2(v)}}, \frac{\sin(v)}{\sqrt{a^2\cos^2(v)+\sin^2(v)}}\right)$.

*(d) Find N_u and write N_u as a multiple of $\mathbf{s}_u$.

(e) Use a computer algebra system to find and simplify N_v. Verify that $N_v = \frac{a}{(\sin^2(v)+a^2\cos^2(v))^{3/2}}\mathbf{s}_v$ and so the curvature of a spheroid is $\kappa(u,v) = \frac{a^2}{(\sin^2(v)+a^2\cos^2(v))^2}$.

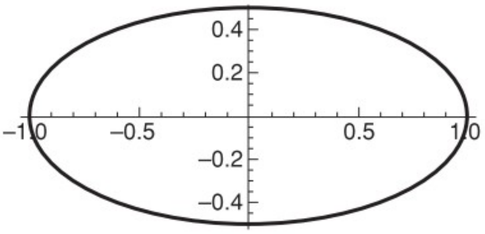

Figure 9.23 An ellipse.

9.3.16. Let $\mathbf{s}(u,v) = ((r\cos(v)+R)\cos(u), (r\cos(v)+R)\sin(u), r\sin(v))$, a parametric representation of the torus.

*(a) Find $\mathbf{s}_u$ and $\mathbf{s}_v$.

(b) Verify that

$$\mathbf{s}_u \times \mathbf{s}_v = (r\cos(u)\cos(v)(R+r\cos(v))$$

$$r\sin(u)\cos(v)(R+r\cos(v)),\ \ r\sin(v)(R+r\cos(v)).$$

(c) Verify that $\|\mathbf{s}_u \times \mathbf{s}_v\| = r(R+r\cos(v))$.

(d) Find $N(u,v)$, N_u, and N_v.

(e) Verify that $N_u = \frac{\cos(v)}{R+r\cos(v)}\mathbf{s}_u$ and $N_v = \frac{1}{r}\mathbf{s}_v$.

9.3.17. Define the hyperbolic trigonometry functions $\tanh(x) = \frac{e^x-e^{-x}}{e^x+e^{-x}}$ and $\operatorname{sech}(x) = \frac{2}{e^e+e^{-x}}$.

(a) Verify that $\tanh^2(x)+\operatorname{sech}^2(x) = 1$, $\tanh(x)' = \operatorname{sech}^2(x)$, $\operatorname{sech}(x)' = -\tanh(x)\operatorname{sech}(x)$, and $\operatorname{sech}(x)'' = -\operatorname{sech}^3(x)+\tanh^2(x)\operatorname{sech}(x)$.

(b) Verify that the pseudosphere $\mathbf{s}(u,v) = (\cos(u)\operatorname{sech}(v), \sin(u)\operatorname{sech}(v), v-\tanh(v))$ is a surface of constant negative curvature for $v \neq 0$.

(c) Graph the plane curve $\overrightarrow{c(t)} = (\operatorname{sech}(t), t-\tanh(t))$. Explain why we need to require $v \neq 0$ in part (b) both visually and from the definition of a curve in Section 9.2.

(Sir Isaac Newton, among others in the early days of calculus, studied the curve $\overrightarrow{c(t)} = (\operatorname{sech}(t), t-\tanh(t))$, called the *tractrix*. Revolved it generates the pseudosphere. Beltrami realized in 1868 that the pseudosphere gave a model of hyperbolic geometry.)

9.3.18. Figure 9.24 depicts one branch of the hyperbola $x^2 - v^2 = 1$, for $x > 0$. When we rotate it around the vertical axis, as in Example 2, we get a hyperboloid of one sheet.

*(a) Find the parametric form for it and graph it.

(b) Use a computer algebra system to verify that the curvature of the surface in part (b) is $\kappa(u,v) = \frac{-1}{(1+2v^2)^2}$.

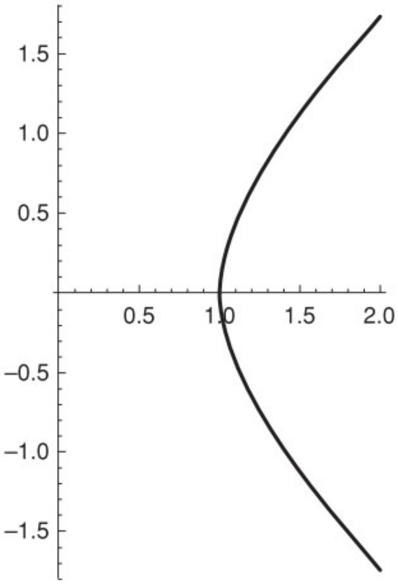

Figure 9.24 A branch of a hyperbola.

*(c) Repeat part (a) for the hyperboloid of two sheets obtained from $v^2 - x^2 = 1$, for $x > 0$ and $v \geq 1$. From the graph what can you say about its curvature?

9.4 Geodesics and the Geometry of Surfaces

9.4.1 Geodesics

Imagine driving a car in hilly terrain. You can easily distinguish when the car goes straight versus turning to the right or left, independently of the curvature of the hills. Intuitively, a straight path following the terrain is a *geodesic*. It is a locally shortest path on the surface. Geodesics are the natural generalization to surfaces of Euclidean lines. Great circles are geodesics on a sphere, discussed in Section 1.5. And on a sphere, the shortest path between any two points follows a great circle. On a surface of constant negative curvature the geodesics become the lines of hyperbolic geometry, discussed in Chapter 4. On a torus the geodesics are more complicated. Figure 9.25 depicts some geodesics on a torus that are indeed shortest paths between two points on them. As the figure illustrates, the curvature can be positive or negative and still the geodesic can be the shortest path. The solid curve $\overrightarrow{g}$ between points A and B in Figure 9.26 gives a geodesic, even though the curve $\overrightarrow{c}$ seems to provide a shorter path. Curve $\overrightarrow{g}$ would, however, feel straight to someone traveling along it, unlike $\overrightarrow{c}$. For nearby points on $\overrightarrow{g}$, it is locally a shortest path. In this section we investigate geodesics on surfaces and their relationship to curvature. In addition we explore more deeply the geometry of a surface, leading to some profound insights of Gauss. The ideas extend to higher dimensions, which we consider briefly.

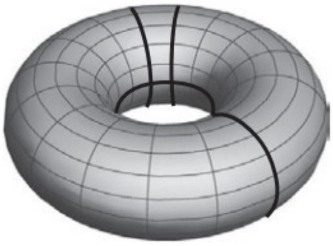

Figure 9.25 Geodesics on a torus that are shortest paths.

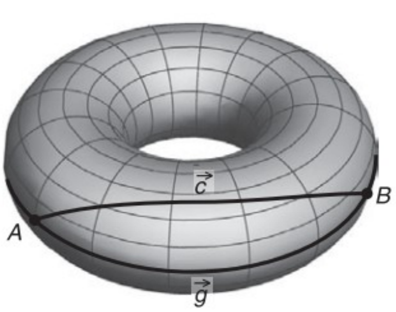

Figure 9.26 A geodesic that is not a shortest path globally.

Imagine attaching a unit tangent vector $T(t)$ to a point (or the car of the previous paragraph) moving along a curve on a surface. In Figure 9.27, the point is "going straight up" the surface. Here the acceleration vector $T'(t)$ is perpendicular to the tangent. In this case the curve is a geodesic, following the surface. In contrast, the plane curve in Figure 9.28 is not a geodesic and here $T'(t)$ is in the plane, which is the tangent plane.

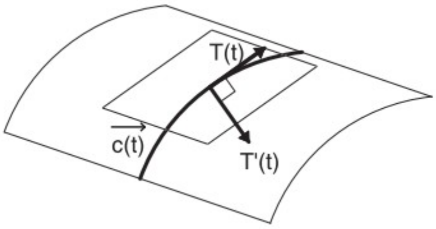

Figure 9.27 A geodesic.

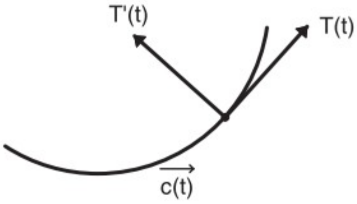

Figure 9.28 A non-geodesic.

Suppose $\overrightarrow{c(t)}$ is a curve in the surface $\mathbf{s}(u, v)$. At a point $\overrightarrow{c(t_0)} = \mathbf{s}(u_0, v_0)$ the unit tangent vector $T(t_0)$ to the curve and the unit normal $N(u_0, v_0) = N(t_0)$ are perpendicular to each other. The tangent plane at this point is determined by $T(t_0)$ and the cross product $T(t_0) \times N(t_0) = B(t_0)$, which is called the *binormal* to the curve. The vectors $T(t_0)$, $N(t_0)$, and $B(t_0)$, form a basis of $\mathbb{R}^3$. So every vector from this point on the curve can be written uniquely as a linear combination of them, in particular, the acceleration vector $T'(t_0)$. By Theorem 9.2.2, T' is orthogonal to T, so T' is a linear combination of the normal and binormal vectors. The binormal component indicates the amount of turning of the curve in the tangent plane and so the deviation from the geodesic. In Figure 9.27, $T'(t_0)$ is a scalar multiple of the binormal, whereas in Figure 9.28 it is a scalar multiple of the normal. The difference motivates our definition of a geodesic. For simplicity, whenever possible we make the curve move at unit velocity so that $\overrightarrow{c'(t_0)} = T(t_0)$ and $\overrightarrow{c''(t_0)} = T'(t_0)$.

Definition. A curve $\overrightarrow{c(t)}$ on a surface $\mathbf{s}(u, v)$ is a *geodesic* if and only if for every value t, there is a scalar $\alpha(t)$ so that $T'(t) = \alpha(t)N(t)$.

Example 1. The lines of longitude on a sphere run through the north and south pole and are, as proved here, geodesics. (See Figure 9.29.) A line of longitude has the form

$$\overrightarrow{c(t)} = \left(R\cos(u_0)\cos\left(\frac{t}{R}\right), R\sin(u_0)\cos\left(\frac{t}{R}\right), R\sin\left(\frac{t}{R}\right)\right) \quad \text{for a fixed } u_0$$

and the sphere is

$$\mathbf{s}(u, v) = (R\cos(u)\cos(v), R\sin(u)\cos(v), R\sin(v)).$$

By Exercise 9.2.13 and Example 3 of Section 9.3 we have

$$\overrightarrow{c'(t)} = T(t) = \left(-\cos(u_0)\sin\left(\frac{t}{R}\right), -\sin(u_0)\sin\left(\frac{t}{R}\right), \cos\left(\frac{t}{R}\right)\right),$$
$$\overrightarrow{c''(t)} = T'(t) = \left(-\frac{1}{R}\cos(u_0)\cos\left(\frac{t}{R}\right), -\frac{1}{R}\sin(u_0)\cos\left(\frac{t}{R}\right), -\frac{1}{R}\sin\left(\frac{t}{R}\right)\right),$$

and

$$N\left(u_0, \frac{t}{R}\right) = \left(\cos(u_0)\cos\left(\frac{t}{R}\right), \sin(u_0)\cos\left(\frac{t}{R}\right), \sin\left(\frac{t}{R}\right)\right).$$

Then for every value of t we have $T'(t) = -\frac{1}{R}N(u_0, \frac{t}{R})$. Hence these curves are geodesics.

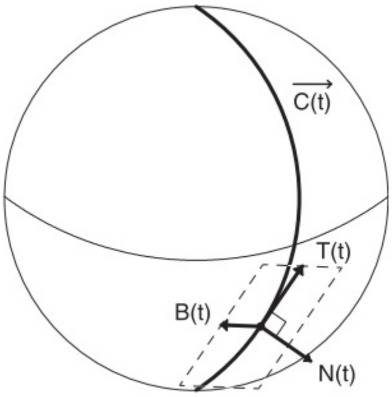

Figure 9.29 Longitude lines are geodesics.

In contrast we show that the circles of latitude, depicted in Figure 9.30, are not geodesics, except the equator (when $v_0 = 0$). Thus we consider values v_0 strictly between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$, other than 0. (The north and south poles, where $v_0 = \pm\frac{\pi}{2}$, are points, not curves.) Here

$$\overrightarrow{c(t)} = (R\cos(t)\cos(v_0), R\sin(t)\cos(v_0), R\sin(v_0))$$

for a fixed v_0. Then

$$\overrightarrow{c'(t)} = (-R\sin(t)\cos(v_0), R\cos(t)\cos(v_0), 0),$$

which is generally not a unit vector. For v_0 other than $\pm\frac{\pi}{2}$, we can divide $\overrightarrow{c'(t)}$ by $R\cos(v_0)$ to find

$$T(t) = (-\sin(t), \cos(t), 0) \qquad \text{and}$$
$$T'(t) = (-\cos(t), -\sin(t), 0).$$

However,

$$N(t) = N(t, v_0) = (\cos(t)\cos(v_0), \sin(t)\cos(v_0), \sin(v_0))$$

and $T'(t)$ is not a scalar multiple of just $N(t)$. That is, circles of latitude are generally not geodesics. Exercise 9.4.5 asks you to show that the equator is a geodesic. ◊

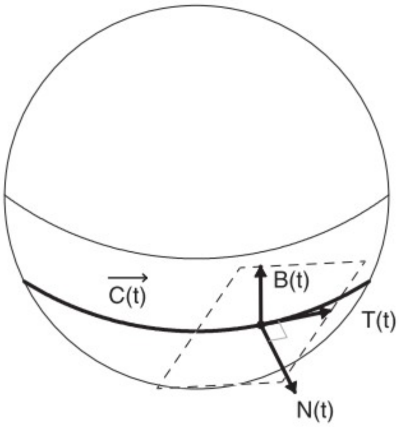

Figure 9.30 Latitude lines generally are not geodesics.

Exercise 9.4.8 asks you to verify that great circles besides longitudes are geodesics.

9.4.2 Geodesics on Surfaces of Revolution

Example 1 generalizes to a surface of revolution $\mathbf{s}(u, v) = (f(v)\cos(u), f(v)\sin(u), g(v))$. Curves with u_0 fixed are geodesics, as Theorem 9.4.1 shows, but curves with v_0 fixed generally are not. Surfaces of revolution have geodesics in other directions, but they can be quite difficult to determine. We add the assumption $f'(v)^2 + g'(v)^2 = 1$ in Theorem 9.4.1 only to simplify the proof. It ensures that $\overrightarrow{c'(t)}$ will have unit length. However, the result remains correct with different parameterizations of the curve. (See Do Carmo [1, 255–258 and 356].)

Exercise 9.4.1. Verify that $f'(v)^2 + g'(v)^2 = 1$ holds in Example 1.

Theorem 9.4.1. *Let* $\mathbf{s}(u, v) = (f(v)\cos(u), f(v)\sin(u), g(v))$ *be a smooth surface of revolution with* $f'(v)^2 + g'(v)^2 = 1$ *for all* v. *Then for all fixed* u_0, *the curves* $\overrightarrow{c(t)} = (f(t)\cos(u_0), f(t)\sin(u_0), g(t))$ *are geodesics.*

Proof. From the assumptions,

$$\overrightarrow{c'(t)} = T(t) = (f'(t)\cos(u_0),\, f'(t)\sin(u_0), g'(t))$$
$$\text{and } T'(t) = (f''(t)\cos(u_0),\, f''(t)\sin(u_0), g''(t)).$$

By Theorem 9.2.2, $T(t) \cdot T'(t) = 0$. That is,

$$f'(t)f''(t)\cos^2(u_0) + f'(t)f''(t)\sin^2(u_0) + g'(t)g''(t) = 0 \qquad \text{or}$$
$$f'(t)f''(t) + g'(t)g''(t) = 0.$$

For now assume that $f'(t) \neq 0$ and rewrite the last equality as $f''(t) = \frac{-g'(t)g''(t)}{f'(t)}$. From subsection 9.3.3

$$N(t) = (g'(t)\cos(u_0), g'(t)\sin(u_0), -f'(t)).$$

For $\overrightarrow{c(t)}$ to be a geodesic, $T'(t)$ must be a scalar multiple of $N(t)$ and so the third coordinates must match. That is, we need

$$g''(t) = \alpha(t)(-f'(t)) \quad \text{or} \quad \alpha(t) = \frac{-g''(t)}{f'(t)}.$$

Now

$$\begin{aligned}\alpha(t)N(t) &= \frac{-g''(t)}{f'(t)}(g'(t)\cos(u_0), g'(t)\sin(u_0), -f'(t)) \\ &= \left(\frac{-g'(t)g''(t)}{f'(t)}\cos(u_0), \frac{-g'(t)g''(t)}{f'(t)}\sin(u_0), g''(t)\right) \\ &= (f''(t)\cos(u_0),\, f''(t)\sin(u_0), g''(t)) = T'(t).\end{aligned}$$

Thus $\overrightarrow{c(t)}$ is a geodesic. Exercise 9.4.9 considers the case $f'(t) = 0$. ■

9.4.3 Arc Length on Surfaces

Figure 9.31 depicts a straight line and a right triangle on a flat piece of paper. When we roll the paper into a cylinder, as in Figure 9.32, the line becomes a helix, which by Exercise 9.4.10 is a geodesic of the cylinder. Distances on the surface of the cylinder work just like Euclidean distance, as long as we don't go too far around it. That is, in Figure 9.32 we can use the Pythagorean theorem and the lengths a and b to find c, provided the triangle is made from geodesic segments on the cylinder. In this sense, the cylinder remains flat, in spite of the bending of the paper. And, as found in Exercise 9.3.9, the cylinder has 0 curvature everywhere. The length of a curve on a surface with nonzero curvature requires a more sophisticated approach.

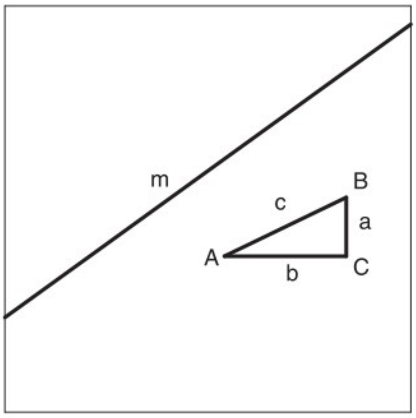

Figure 9.31 Shapes in the plane.

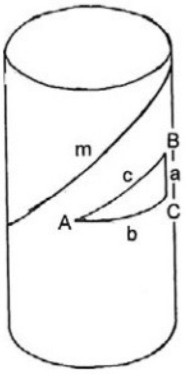

Figure 9.32 Corresponding shapes on a cylinder.

Gauss generalized the Pythagorean theorem to curved surfaces. Since general surfaces vary from point to point, his approach had to be local. He used the language of differentials, which we still use. For a surface $\mathbf{s}(u, v)$, Gauss expressed the distance along a geodesic in terms of the distance along the cross section curves. In Figure 9.33 we want to determine the differential (infinitesimal distance) ds along the geodesic from (u_0, v_0) to (u_1, v_1) in terms of the differentials du and dv. Gauss found the following amazing formula, now called the first fundamental form. (See Do Carmo [1, 92].)

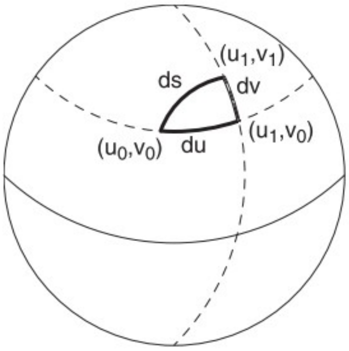

Figure 9.33 A differential on a surface.

$$ds^2 = Edu^2 + 2Fdudv + Gdv^2, \tag{9.5}$$

where $E = \mathbf{s}_u \cdot \mathbf{s}_u$, $F = \mathbf{s}_u \cdot \mathbf{s}_v$, and $G = \mathbf{s}_v \cdot \mathbf{s}_v$.

Exercise 9.4.2.* For the cylinder $\mathbf{s}(u, v) = (R\cos(\frac{u}{R}), R\sin(\frac{u}{R}), v)$ find E, F, and G from equation 9.5. Explain why the values validate the text's assertion about distances on the surface of the cylinder.

(9.5) generalizes the distance formulas of exercises 3.1.9 and 3.1.10 when the scales of the axes differ and are at an angle to one another. In linear algebra terms, for each point of the surface (9.5) determines an inner product for the tangent plane. From Exercise 9.3.8, for surfaces of revolution $\mathbf{s}_u$ and $\mathbf{s}_v$ are orthogonal, and so F is always 0.

Exercise 9.4.3.* For a surface of revolution as in Theorem 9.4.1 find $E(u,v)$ and $G(u,v)$ to verify that $ds^2 = f(v)^2du^2 + dv^2$.

Theorem 9.4.2 uses (9.5) to find the length along a curve on a surface. Calculus texts derive the arc length of a plane curve through approximation with the lengths Δs_i of line segments, as in Figure 9.34. A larger number of segments can improve the approximation. The plane curve $\overrightarrow{c(t)} = (x(t), y(t))$ has arc length $s = \int_a^b \sqrt{x'(t)^2 + y'(t)^2}dt$ for $a \leq t \leq b$. On a surface we similarly approximate the length of a curve $\overrightarrow{c(t)}$ using geodesic segments, as in Figure 9.35, giving $s \approx \sum \Delta s_i$. In the limit we get an integral: $s = \int_a^b ds = \int_a^b \sqrt{Edu^2 + 2Fdudv + Gdv^2}$. From calculus $du = u'(t)dt$ and $dv = v'(t)dt$. The reasoning of this paragraph gives us Theorem 9.4.2.

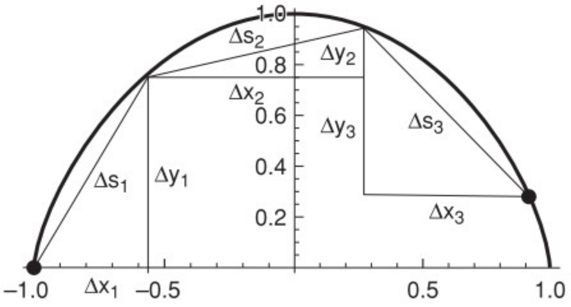

Figure 9.34 Arc length of a plane curve.

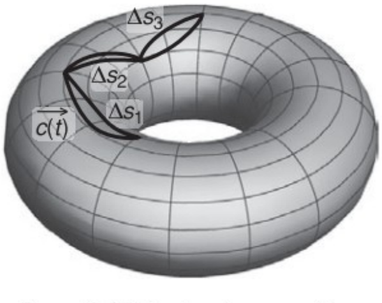

Figure 9.35 Arc length on a surface.

Theorem 9.4.2. *On a smooth surface* $\mathbf{s}(u,v)$, *the length of a curve* $\overrightarrow{c(t)}$ *from* $t = a$ *to* $t = b$ *is*

$$s = \int_a^b \sqrt{E(t)u'(t)^2 + 2F(t)u'(t)v'(t) + G(t)v'(t)^2}dt. \tag{9.6}$$

Example 2. For the sphere of radius R, $\mathbf{s}(u, v) = (R\cos(u)\cos(\frac{v}{R}), R\sin(u)\cos(\frac{v}{R}), R\sin(\frac{v}{R}))$. From Exercise 9.4.3, $E(u, v) = R^2\cos^2(\frac{v}{R})$, $F = 0$, and $G = 1$. A circle of latitude with $v = v_0$ has $\overrightarrow{c(t)} = (R\cos(t)\cos(\frac{v_0}{R}), R\sin(t)\cos(\frac{v_0}{R}), R\sin(\frac{v_0}{R}))$ for $0 \leq t \leq 2\pi$. Here $u(t) = t$ and $v(t) = v_0$. We can substitute these values into (9.6) to obtain

$$s = \int_0^{2\pi} \sqrt{R^2\cos^2\left(\frac{v_0}{R}\right) \cdot 1 + 0 + 0}\,dt = \int_0^{2\pi} R\left|\cos\left(\frac{v_0}{R}\right)\right| dt = 2\pi R\left|\cos\left(\frac{v_0}{R}\right)\right|,$$

the circumference of a circle with radius $R\cos(\frac{v_0}{R})$. ♢

9.4.4 The Gauss-Bonnet Theorem

Gauss coupled his insights on curvature and geodesics to relate areas of triangles and curvature. The proof of Theorem 9.4.3 goes beyond the level of this text. However, its statement unites Theorem 1.5.3 and Theorem 4.1.1 and so shows the unity of spherical and hyperbolic geometries. The notation $d\sigma$ in the theorem is a shorthand for the differential of area taking into account the values of E, F, and G in (9.5).

Theorem 9.4.3 (Gauss part of Gauss-Bonnet Theorem, 1827). *Suppose a triangular region T of a smooth surface is bounded by geodesics and the measures of the angles made by the geodesics are α, β, and γ. Assume that T can be shrunk continuously to a point on the surface. Then $\alpha + \beta + \gamma - \pi = \int\int_T \kappa(u, v)d\sigma$, where $\kappa(u, v)$ is the curvature at (u, v).*

Proof. See Do Carmo [1, 264–270]. ■

Example 3. A sphere of radius R has constant curvature $\frac{1}{R^2}$. Then $\int\int_T \kappa(u, v)d\sigma = \frac{1}{R^2}$ area(T). Thus we can rearrange the equation of Theorem 9.4.3 to give Theorem 1.5.3 for spherical triangles: $R^2(\alpha + \beta + \gamma - \pi) = \text{area}(T)$. In Chapter 1 we called $\alpha + \beta + \gamma - \pi$ the spherical excess. Analogously, Theorem 4.1.1 tells us about the area of triangles in hyperbolic geometry. The hyperbolic plane has constant negative curvature, say $-\frac{1}{k}$. (See Exercise 9.3.17.) The minus sign changes Gauss's formula to $-k(\alpha + \beta + \gamma - \pi) = k(\pi - (\alpha + \beta + \gamma)) = \text{area}(T)$. We called $\pi - (\alpha + \beta + \gamma)$ the defect of a triangle. In between spherical and hyperbolic geometries is Euclidean geometry, where the curvature is 0. In the Euclidean case, Theorem 9.4.3 reduces to Theorem 1.1.1, and the angle sum of a Euclidean triangle is π or 180°. ♢

Pierre Ossian Bonnet (1819–1892) generalized Gauss's result in 1848 to regions bounded by curves, whether or not they are geodesics. (See Do Carmo [1, 264–270].) He also considered regions that can't be shrunk to a point. For example, the shaded band on the torus in Figure 9.36 can't be shrunk continuously on the surface to a point. The general Gauss-Bonnet theorem needs to take into account the topological type of the surface, using the Euler characteristic. (See Project 9. Also, Project 12 of Chapter 1 considers the generalized Euler's formula, which relates to the Euler characteristic.)

9.4.5 Higher Dimensions

Bernhard Riemann (1826–1866) boldly extended Gauss's vision from two-dimensional surfaces to what are called n-dimensional *manifolds*. All our work has embedded curves and surfaces

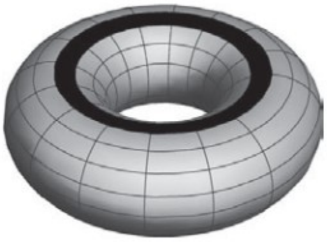

Figure 9.36 A band on a torus.

in 2- and 3-dimensional Euclidean space. While the embedding simplified our presentation, it would have made Riemann's advances impossible. For example, we heavily used the cross product, which is defined only in $\mathbb{R}^3$. Riemann needed the results Gauss proved about surfaces independent of whether they were embedded in a space. Riemann generalized the concepts of curvature, geodesics, and more. For example, a 3-dimensional manifold $\mathbf{m}(u, v, w)$ has lengths generalizing the 3-dimensional Pythagorean theorem. Riemann's version of (9.5) for the 3-manifold $\mathbf{m}(u, v, w)$ is

$$ds^2 = Edu^2 + 2Fdudv + Gdv^2 + 2Hdudw + 2Idvdw + Jdw^2,$$

where

$$E = \mathbf{m}_u \cdot \mathbf{m}_u,\ F = \mathbf{m}_u \cdot \mathbf{m}_v,\ G = \mathbf{m}_v \cdot \mathbf{m}_v,\ H = \mathbf{m}_u \cdot \mathbf{m}_w,$$
$$I = \mathbf{m}_v \cdot \mathbf{m}_w, \quad \text{and} \quad J = \mathbf{m}_w \cdot \mathbf{m}_w \tag{9.7}$$

Exercise 9.4.4. Verify that the three-dimensional Pythagorean theorem (Exercise 1.5.11) corresponds to a 3-dimensional manifold with $E = G = J = 1$ and $F = H = I = 0$.

Einstein's general theory of relativity is related to 4-dimensional manifolds since physics uses three spatial dimensions and one time dimension. The values of the parameters in the formula analogous to (9.5) and (9.6) depend on the gravitational field. (However, the value corresponding to ds^2 can come out negative, so the space isn't Riemannian. See the discussions in Section 5.5 and 6.6.) Alternatively, we could take the curvature of the manifold as determining what gravity is at each point. In any case, light travels on geodesics of it and so bends from a Euclidean straight line. In 1919 Sir Arthur Eddington provided the most well known confirmation of this bending of light, mentioned in the introduction to Chapter 9. The general theory of relativity has everyday effects as well. The calculations for global positioning systems need to compensate for relativistic effects. The effects include the difference of the earth's gravitational field around the satellites compared with the field on the earth's surface where we use devices to find position. The gravitational differences modify how fast the highly accurate atomic clocks on the satellites run compared to earth clocks. Hence GPS systems must compensate for relativistic effects to avoid significant errors. (See Love [5].)

Differential geometry has provided profound understanding in mathematics and physics. At the heart of its insights are the geometric concepts of curvature and geodesics.

(%/

Pág. 203(218) 13 e 14

14 → (a) List all the regular n-gons with $n \leq 100$ sides that are constructible with straightedge and compass.

(b) Use Gauss' theorem to prove that an angle of 20° is not constructible.

(c) Use Gauss' theorem to decide whether or not an angle of 6° is constructible

Obs: Livro todo!!!

Veja Transformation.

—

— 20211017/ Geometry/

Thinking Geometrically ... pdf

Obs: Gauss - 59 occurences

Livro todo!!

Pág. 20(45) Gauss actually constructed a regular 17-gon in 1796.

Pág. 153(178) - 4.1.1 ⇙

The Advent of Hyperbolic Geometry

(%

Characteristic Axiom of Hyperbolic Geometry

Given a line K and a point P not on K, there are at least two lines on P that do not intersect K (Fig 4.1).

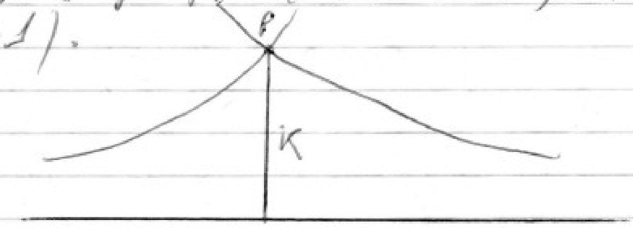

Various consequences from this change, including the many that Saccheri found.

The most startling is the theorem that the measures of the angles of a triangle do not add up to

π (or 180°), as they do in Euclidean geometry.

Pag 154 (179) theorem 4.1.1

In hyperbolic geometry the difference,

$\pi - (m\angle A + m\angle B + m\angle C)$

between π and the angle sum of a triangle is proportional to the area of triangle.

(%

Pág. 160 (185) 4.1.4

Carl Friedrich Gauss

Gauss → regular 17-sided polygon

Obs: Exercises 4.1.9.

A curious property of hyperbolic geometry is that two lines do not intersect and are not sensed parallel (described informally in Exercise 4.1.6) have a common perpendicular.

2º By 1800 Gauss had become convinced that Euclid's parallel postulate could not be proven, and in installments he developed what he called non-Euclidean geometry.

IMPORTANT:

Pág. 410 (435)

Gauss → they link the area of a triangle in spherical and hyperbolic geometry to their angle sums.

In the generalization, Gauss used geodesics for the sides of the triangles. Then, as he proved, the difference

(%

Pag. 411 (436) cont.

between the angle sum of the triangle and π was determined by the area of the triangle and the curvature within it.

The sphere has constant positive curvature and the hyperbolic plane has constant negative curvature, accounting for the difference in theorems 1.5.3 and 4.1.1.

Obs: Página importante.

Einstein
Riemann
Gauss

In 1854, Georg Riemann (1826–1866) generalized Gauss's results to any number of dimensions and so transformed differential geometry again.

Many mathematicians developed Riemann's insights into rigorous mathematics with computational methods.

Albert Einstein (1879–1955) built on Riemann's profound insights to revolutionize physics with the general theory of relativity.

Einstein needed a four-dimensional space built from three dimensions and time, all with varying CURVATURE corresponding to the strength of gravitational fields.

(% %

—

at different points.

Pag. 424 (449) 9.3

Surfaces and Curvature

From Gauss's definition of curvature, a sphere has constant positive curvature and a Euclidean plane has 0 curvature.

We often use the formula

$$\cos^2(\theta) + \sin^2(\theta) = 1$$

to simplify formulas for surfaces of revolution.

9.3.1 Surfaces

$$z = \pm\sqrt{R^2 - x^2 - y^2}$$

determine the upper and lower hemispheres.

Pag. 427 (452) → Página importante

M is called the shape operator.

Pag. 436 (461) - 9.4.3

Arc Length and Surfaces

(%'

Pág. 437 (461) –

Figure 9.31 Shapes in the plane

Figure 9.32 Corresponding shapes on a cylinder.

Important:

→ Figure 9.33 A differential on a surface.

Obs: Estudar livro todo!!!

– 20211017/ Geometry/
Perspectives on Projective Geometry.pdf

Obs, Livro todo!!!
Estudar com o livro anterior e
Pappus' theorem.

→ Gauss – 31 ocurrences.

– 20211017/ Geometry/
Vector analysis versus vector calculus.pdf

! -256- - -* ! - 1 A K A 7 3
2 7 " " 9 9 5 ! ! " 1 L 7 3 1"
D -256- - -* ! - ! ! 3 A A
! 1"
(% ' % / 9 / B 7! 0 !

Lemma 9.6.1. *Let B be the closed ball centered at a point $\boldsymbol{P}$ and with radius $R > 0$. The flux of the vector field $\boldsymbol{F}$ across ∂B is*

$$-4\pi GMm.$$

Proof. The unit normal vector to ∂B at a point $(x,y,z) \in \partial B$, pointing to the *exterior* of B, is given by

$$\boldsymbol{N} = \frac{\boldsymbol{r}}{\| \boldsymbol{r} \|}.$$

Hence, for

$$\boldsymbol{F} = -\frac{GMm}{\| \boldsymbol{r} \|^3}\,\boldsymbol{r},$$

it turns out that

$$\langle \boldsymbol{F},\boldsymbol{N} \rangle = -GMm\frac{1}{\| \boldsymbol{r} \|^2} = -GMm\frac{1}{R^2}.$$

That is, $\langle \boldsymbol{F},\boldsymbol{N} \rangle$ is constant on the sphere. Consequently, the flux of the vector field $\boldsymbol{F}$ across ∂B coincides with the product of such a constant with the area of ∂B. That is,

$$\text{Flux} \;=\; -4\pi GMm.$$

□

Theorem 9.6.1 (Gauss's law). *Let U be a regular 3-surface with boundary of class C^2 in $\mathbb{R}^3$ and assume that*

$$S := \partial U$$

is a compact regular surface that is oriented according to the normal vector pointing to the exterior of U.

(1) If $\boldsymbol{P} \notin U$, then the flux of $\boldsymbol{F}$ across S is zero.
(2) If $\boldsymbol{P} \in U \setminus S$, then the flux of $\boldsymbol{F}$ across S is $-4\pi GMm$.

Proof. We first suppose that $\boldsymbol{P} \notin U$. In this case, $\boldsymbol{F}$ is a vector field of class C^1 on a neighborhood of U, and by the divergence theorem, the flux of $\boldsymbol{F}$ across S coincides with

$$\iiint_U \operatorname{Div}\boldsymbol{F}(x,y,z)\;\mathrm{d}(x,y,z) = 0.$$

Next, suppose that $\boldsymbol{P} \in U \setminus S$. This time, it is not so easy to apply the divergence theorem, because $\boldsymbol{F}$ is not defined at the point $\boldsymbol{P} \in U$. To avoid this complication, take $R > 0$ small enough that the closed ball B centered at $\boldsymbol{P}$ and with radius R is contained in the interior of U (Fig. 9.6). Now consider

$$\Omega = \{(x,y,z) \in U \;:\, \| (x-x_0, y-y_0, z-z_0) \| \geq R\},$$

which is a regular 3-surface with boundary

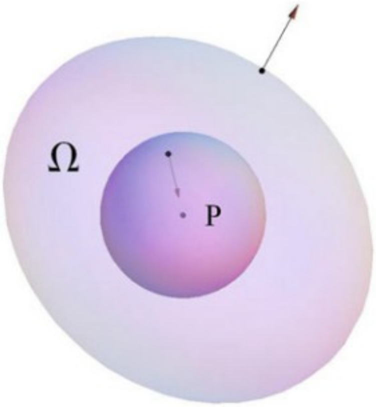

Fig. 9.6 Illustration of the proof of Theorem 9.6.1

$$\partial\Omega = S \cup \partial B.$$

As usual, the boundary of Ω is oriented according the normal vector exterior to Ω. This means that the orientation of $\partial\Omega$ at the points of S is given by the normal vector to S that points to the exterior of U, and the orientation of $\partial\Omega$ at points of ∂B is given by the normal vectors to the sphere that point to the *interior* of the ball.

So, if we consider in ∂B the orientation according to the normal vector *exterior* to the ball, then

$$\int_{\partial\Omega} \langle \boldsymbol{F}, \boldsymbol{N} \rangle \ \mathrm{d}S = \int_{S} \langle \boldsymbol{F}, \boldsymbol{N} \rangle \ \mathrm{d}S - \int_{\partial B} \langle \boldsymbol{F}, \boldsymbol{N} \rangle \ \mathrm{d}S.$$

On the other hand, since $\boldsymbol{F}$ is a vector field of class C^1 on Ω, we can apply the divergence theorem to conclude that

$$\int_{\partial\Omega} \langle \boldsymbol{F}, \boldsymbol{N} \rangle \ \mathrm{d}S = \iiint_{\Omega} \mathrm{Div}\boldsymbol{F}(x,y,z) \ \mathrm{d}(x,y,z) = 0.$$

Consequently, from the previous lemma,

$$\int_{S} \langle \boldsymbol{F}, \boldsymbol{N} \rangle \ \mathrm{d}S = \int_{\partial B} \langle \boldsymbol{F}, \boldsymbol{N} \rangle \ \mathrm{d}S = -4\pi GMm.$$

□

It should be noted that Gauss's law is usually formulated in terms of the electrical field generated by a point charge.

In what follows, we assume for simplicity that $m = 1$. Note that the flux of the gravitational field generated by a point mass M located at $\boldsymbol{P}$ in the interior of U depends only on M and not on the position $\boldsymbol{P}$. If we consider a series of point masses

$M_1, M_2, \ldots$ in the interior of U, the resulting vector field $\boldsymbol{F}$ is the sum of the vector fields generated by each one of the point masses, and the flux of $\boldsymbol{F}$ across S is given by

$$\sum_j -4\pi G M_j = -4\pi G M,$$

where

$$M = \sum_j M_j$$

is the total mass. Therefore, it is no surprise that *if $\boldsymbol{F}$ is the gravitational field generated by an object of total mass M located in the interior of U, then the flux of $\boldsymbol{F}$ across $S = \partial U$ equals $-4\pi GM$*. (see [11]).

-256- - -* ! - 2 D ! & ! ! ! "
1"

" -256- - -* ! - ! ! 3 ! "

-256- - -* ! - ! ! . “ 3 . "
&)! “ "! 7 " " "KA

-256- - -* ! - ! ! . “ 3 . " &
)! “ "! 7 " " 1"

-256- - -* ! - ! ! . “ 3 . "
& "KA

-256- - -* ! - G! ! ! ! !
1 D ! 3 4 A / 1"

-256- - -* ! - G! ! ! ! !
1 D ! 3 4 A 1"

-256- - -* ! - ! J ? ! 2 D !J ! "
4 ! < 2 H " 211 ! ! L =" $ 1"

—

Gauss - 9 occurrences.

Pág. 314 (329) - ⇙

Theorem 9.6.1 (Gauss's law)

Let U be a regular 3-surface with boundary of class C^2 in $\mathbb{R}^3$ and assume that

$$S := \partial U$$

is a compact regular surface that is oriented to the normal vector pointing to the exterior of U.

(1) If $P \in U$, then the flux of F across S is zero.

(2) If $P \in U \setminus S$, then the flux of F across S is $-4\pi G M m$.

- 20211017T Geometry/ Vector Algebra I Scalar and Vector...pdf (nome errado = ao livro seguinte)

Obs: Gauss - 45 occurrences.

Do mesmo autor: Klaus Weltner

(% B

- Mathematics for Physicists and Engineers.pdf

Obs: Gauss - 49 ocurrências -
Mais 2 volumes:

= Mathematik für Physiker 1.djvu
= Mathematik für Physiker 2.djvu

- 20211017/ Geometry/
Yaglom - Challeging mathematical problems with elementary solutions (1987).pdf

Yaglom - Challeging mathematical ... 2.pdf

Obs:

Volume 1 → integers

Volume 2 → Four formulas for π
(Pág 22)

- 20211017/ Geometry/
Important-Vector/ Vector Calculus (2015).pdf

Obs: Gauss - 125 ocurrências.
Livro todo!!!

Veja: Todo diretório: Important-Vector

(%

20211017 ||| (20211017) || cont.

- 20211017/ Geometry/Important Vector/ Groups, Matrices, and Vector Spaces..(2017).pdf
  Pág. 126 (136) - 5.2.7
  A beautiful formula: Lewis Carroll's identity.
  → theory of determinantes.
- Linear Algebra and Vector Calculus (21100115)...(2017).pdf
  Obs.: Gauss - 57 occurences.
  Pág. 570 (570) - Example 11 (chapter 7)
  $\bar{F} = x\hat{i} - z\hat{j} + 2y\hat{k}$
  displacing the particle along the triangle OAB
  Fig 7.8 (Pág. 571)
  Obs: Livro todo!!!
- 20211017/ Geometry/Important-Vector/ a) Calculus...(2017).pdf
  Obs. Pág. 207 Chapter 3

! -256- - -* ! - 7 1 J ! J ! " 1!
2 7 1 9 211 ! ? ! 2 D ! 1 1"
(B %B $

Clearly, $\det(A) = 0$ unless P is a full permutation matrix. Furthermore, we have the following proposition.

Proposition 5.10. *If $A = LPDU$ is nonsingular, then*

$$\det(A) = \det(P)\det(D) = \pm\det(D).$$

Thus, up to sign, the determinant of an invertible matrix is the product of its pivots. If $A = LDU$, then $\det(A) = \det(D)$.

Proof. Just use the product rule and the fact that $\det(L) = \det(U) = 1$, since L and U are triangular and have ones on their diagonals. ☐

Recall from Proposition 4.17 that an invertible $A \in \mathbb{F}^{n\times n}$ has an LDU decomposition if and only if each A_k is also invertible, where A_k is the $k \times k$ submatrix in the upper left-hand corner of A, and that the pivots of A are the diagonal entries of D. Let $d_1, \dots, d_n$ be these pivots. Then $A_k = L_k D_k U_k$, so

$$\det(A_k) = \det(D_k) = d_1 \cdots d_k.$$

This gives us the following result.

Proposition 5.11. *If $A \in \mathbb{F}^{n\times n}$ satisfies the condition that each A_k, $1 \le k \le n$, is invertible, then A has an LDU decomposition, and the kth pivot of A in its LDU decomposition is*

$$d_k = \frac{\det(A_k)}{\det(A_{k-1})}. \tag{5.10}$$

5.2.7 A beautiful formula: Lewis Carroll's identity

The theory of determinants is a remarkably rich topic, and we have barely scratched its surface. Many of the foremost mathematicians of the nineteenth century, among them Gauss, Laplace, Lagrange, Cayley, Sylvester, and Jacobi, discovered some of its important properties and applications. An example is the Jacobian of a mapping and the change of variables formula. In this section, we will consider a determinantal curiosity from the nineteenth century that has recently been found to have connections to contemporary mathematics. The formula $\det\begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$ turns out to have a generalization to $n \times n$ matrices that was discovered by Charles Dodgson, a professor of mathematics at Oxford who, as is well known, wrote *Alice's Adventures in Wonderland* and *Through the Looking Glass* under the pseudonym Lewis Carroll. Dodgson found an identity involving the determinant of an $n \times n$ matrix A that is analogous to the formula in the 2×2 case. Let A_C be the

$(n-2) \times (n-2)$ submatrix of A obtained by deleting the first and last rows and the first and last columns. If $n = 2$, put $\det(A_C) = 1$. Also, let A_{NW} denote the $(n-1) \times (n-1)$ submatrix in the upper left-hand corner of A, and define A_{NE}, A_{SW}, and A_{SE} to be the $(n-1) \times (n-1)$ submatrices in the other three corners of A. Dodgson's formula asserts that

$$\det(A_C)\det(A) = \det(A_{NW})\det(A_{SE}) - \det(A_{NE})\det(A_{SW}) \qquad (5.11)$$

(see C.L. Dodgson, *Proc. Royal Soc. London* **17**, 555–560 (1860)). If $\det(A_C) \neq 0$, this substantially cuts down on the difficulty of finding $\det(A)$. Of course, if $\det(A_C) = 0$, then the identity says nothing about $\det(A)$, although it does say something about the four determinants at the corners. This formula is referred to as Dodgson condensation. Dodgson himself supplied the term condensation. In the 1980s, it was noticed that Dodgson condensation is related to the problem of counting alternating sign matrices (ASMs). Eventually, the problem of enumerating the ASMs was given an elegant solution using statistical mechanics.

D -256- - -* ! - ? ! 2 D ! ! " !
$ 7 K! ! 9 ! H A 3 7 ! : =" ! 1"
($ =L! 1 !1

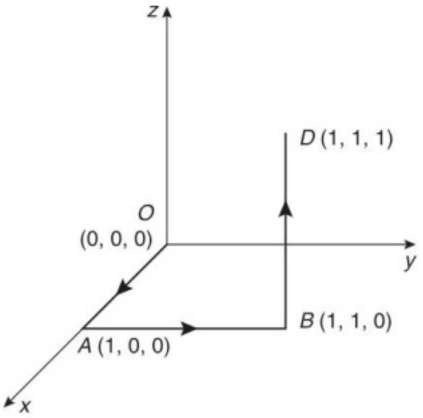

Fig. 7.7

(a) Along $OA : y = 0, \quad z = 0$

$dy = 0, \quad dz = 0$

x varies from 0 to 1.

$$\int_{OA} \overline{F} \cdot d\overline{r} = \int_0^1 3x^2 dx = \left|x^3\right|_0^1 = 1$$

(b) Along $AB : x = 1, \quad z = 0$

$dx = 0, \quad dz = 0$

y varies from 0 to 1.

$$\int_{AB} \overline{F} \cdot d\overline{r} = \int_0^1 0\, dy = 0$$

(c) Along $BD : x = 1, \quad y = 1$

$dx = 0, \quad dy = 0$

z varies from 0 to 1.

$$\int_{BD} \overline{F} \cdot d\overline{r} = \int_0^1 20z^2 dz = 20\left|\frac{z^3}{3}\right|_0^1 = \frac{20}{3}$$

Substituting in Eq. (1),

$$\text{Work done } = 1 + 0 + \frac{20}{3}$$

$$= \frac{23}{3}$$

Example 11: Find the work done by the force $\overline{F} = x\hat{i} - z\hat{j} + 2y\hat{k}$ in displacing the particle along the triangle OAB, where

$$OA : 0 \le x \le 1, \quad y = x, \quad z = 0$$

$$AB : 0 \le z \le 1, \quad x = 1, \quad y = 1$$

$$BO : 0 \le x \le 1, \quad y = z = x$$

Solution: Work done $= \int_C \bar{F}\cdot \mathrm{d}\bar{r}$

(i) Let $\bar{r} = x\hat{i} + y\hat{j} + z\hat{k}$

$$\mathrm{d}\bar{r} = \hat{i}\,\mathrm{d}x + \hat{j}\mathrm{d}y + \hat{k}\mathrm{d}z$$

(ii) $\bar{F}\cdot \mathrm{d}\bar{r} = (x\hat{i} - z\hat{j} + 2y\hat{k})\cdot(\hat{i}\,\mathrm{d}x + \hat{j}\mathrm{d}y + \hat{k}\mathrm{d}z)$

$$= x\mathrm{d}x - z\mathrm{d}y + 2y\mathrm{d}z$$

(iii) Path of integration is the triangle *OAB*.

$$\text{Work done } = \int_C \bar{F}\cdot \mathrm{d}\bar{r}$$
$$= \int_{OA} \bar{F}\cdot \mathrm{d}\bar{r} + \int_{AB} \bar{F}\cdot \mathrm{d}\bar{r} + \int_{BO} \bar{F}\cdot \mathrm{d}\bar{r} \qquad \dots(1)$$

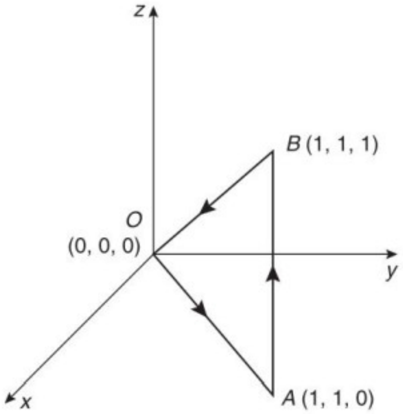

Fig. 7.8

(a) Along OA : $y = x,\quad z = 0$
$\mathrm{d}y = \mathrm{d}x,\quad \mathrm{d}z = 0$
x varies from 0 to 1.

$$\int_{OA} \bar{F}\cdot \mathrm{d}\bar{r} = \int_0^1 x\,\mathrm{d}x = \left|\frac{x^2}{2}\right|_0^1 = \frac{1}{2}$$

(b) Along AB: $x = 1,\quad y = 1$
$\mathrm{d}x = 0,\quad \mathrm{d}y = 0$
z varies from 0 to 1.

$$\int_{AB} \bar{F}\cdot \mathrm{d}\bar{r} = \int_0^1 2\,\mathrm{d}z = |2z|_0^1 = 2$$

(c) Along BO: $x = y = z$
$\mathrm{d}x = \mathrm{d}y = \mathrm{d}z$
x varies from 1 to 0.

$$\int_{BO} \bar{F}\cdot \mathrm{d}\bar{r} = \int_1^0 (x\,\mathrm{d}x - x\,\mathrm{d}x + 2x\,\mathrm{d}x) = |x^2|_1^0 = -1$$

Substituting in Eq. (1),

$$\int_C \bar{F}\cdot \mathrm{d}\bar{r} = \frac{1}{2} + 2 - 1 = \frac{3}{2}$$

Example 12: Find the work done by the force $\bar{F} = 16\,y\hat{i} + (3x^2 + 2)\,\hat{j}$ in moving a particle once round the right half of the ellipse $x^2 + a^2y^2 = a^2$ from (0, 1) to (0, −1).

Solution: Work done $= \int_C \bar{F}\cdot \mathrm{d}\bar{r}$

(i) Let $\bar{r} = x\hat{i} + y\hat{j}$

$$\mathrm{d}\bar{r} = \hat{i}\mathrm{d}x + \hat{j}\mathrm{d}y$$

-256- - -* ! - ! 9 " =" 7 K! !
9 ! H A 3 7 ! : =" ! 1"
(!1 %

CHAPTER 3

Curve Sketching

3.1 INTRODUCTION

Curve sketching is a procedure to obtain an approximate shape of the curve without plotting a large number of points on it. The graph of a given function is helpful in giving a visual presentation of the behaviour of the function involving a study of symmetries, asymptotes, the interval of rising up or falling down and of the concavity upwards and downwards, etc. Before drawing curves, we will study monotonic functions, concavity and convexity of a curve and points of inflection.

3.2 MONOTONIC FUNCTIONS

A function $f(x)$ is called *monotonic increasing* if

(i) $f(x)$ is continuous in the closed interval $[a, b]$,
(ii) $f(x)$ is differentiable in the open interval (a, b) and
(iii) $f'(x) \geq 0$ throughout the interval (a, b).
If $f'(x) > 0$, the function is called *strictly increasing*.

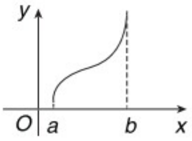

Fig. 3.1

A function $f(x)$ is called *monotonic decreasing* if

(i) $f(x)$ is continuous in the closed interval $[a, b]$,
(ii) $f(x)$ is differentiable in the open interval (a, b) and
(iii) $f'(x) \leq 0$ throughout the interval (a, b).
If $f'(x) < 0$, the function is called *strictly decreasing*.

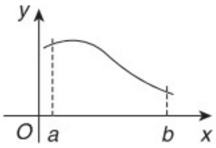

Fig. 3.2

A function is called monotonic in the interval $[a, b]$, if it is increasing or decreasing in the interval $[a, b]$.

3.3 CONCAVITY, CONVEXITY AND POINTS OF INFLECTION OF A CURVE

A function $f(x)$ is concave upwards (convex downwards) if

(i) $f(x)$ is continuous in the closed interval $[a, b]$,
(ii) $f(x)$ is differentiable in the open interval (a, b), and
(iii) $f''(x) > 0$ throughout the interval (a, b).

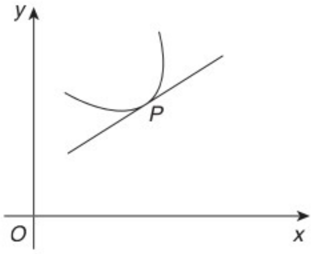

Fig. 3.3

A function $f(x)$ is concave downwards (convex upwards) if

(i) $f(x)$ is continuous in the closed interval $[a, b]$,
(ii) $f(x)$ is differentiable in the open interval (a, b), and
(iii) $f''(x) < 0$ throughout the interval (a, b).

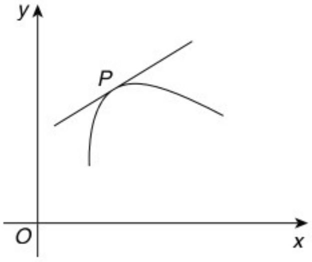

Fig. 3.4

A point at which $f''(x) = 0$ is known as point of inflection where the function changes the direction of concavity from downwards to upwards or vice versa.

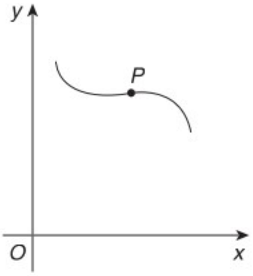

Fig. 3.5

3.4 MAXIMA AND MINIMA

Maxima and minima are largest and smallest value respectively that a function takes either within a given neighbourhood or on the entire function domain.

First Derivative Test for Local Extremum

(a) $f(x)$ has a local maximum at $x = a$, if $f'(a) = 0$ and $f'(x)$ changes sign from positive to negative at $x = a$.

(b) $f(x)$ has a local minimum at $x = a$, if $f'(a) = 0$ and $f'(x)$ changes sign from negative to positive at $x = a$.

Second Derivative Test for Local Extremum

(a) $f(x)$ has a local maximum at $x = a$, if $f'(a) = 0$ and $f''(a) < 0$.
(b) $f(x)$ has a local minimum at $x = a$, if $f'(a) = 0$ and $f''(a) > 0$.
(c) The test fails if $f'(a) = 0$ and $f''(a) = 0$. $f(x)$ may have a local maximum, a local minimum or neither.

Example 1

Discuss the function $f(x) = x^4 - 4x^3$ with respect to increasing or decreasing nature, concavity, points of inflection and local maxima and minima.

Solution

$$f(x) = x^4 - 4x^3$$
$$f'(x) = 4x^3 - 12x^2 = 4x^2 (x - 3)$$
$$f''(x) = 12x^2 - 24x = 12x (x - 2)$$

(*i*) *Increasing or decreasing function*

(a) $f(x)$ is an increasing function if $f'(x) > 0$ i.e. $4x^2(x - 3) > 0$, i.e. $x^2(x - 3) > 0$, i.e. $x - 3 > 0$ since $x^2 > 0$ i.e. $x > 3$.
Hence, $f(x)$ is increasing in the interval $(3, \infty)$.

(b) $f(x)$ is a decreasing function if $f'(x) < 0$, i.e. $4x^2(x - 3) < 0$, i.e. $x^2(x - 3) < 0$, i.e. $x - 3 < 0$ since $x^2 > 0$, i.e. $x < 3$.

Hence, $f(x)$ is decreasing in the interval $(-\infty, 3)$.

(*ii*) *Concavity*

(a) $f(x)$ is concave upwards if $f''(x) > 0$, i.e. $12x(x - 2) > 0$, i.e. $x(x - 2) > 0$.
Now, $x(x - 2) > 0$ if

Case (I) $x > 0$ and $x - 2 > 0$, i.e. $x > 0$ and $x > 2$, i.e. $x > 2$.
Case (II) $x < 0$ and $x - 2 < 0$, i.e. $x < 0$ and $x < 2$, i.e. $x < 0$

Hence, $f''(x) > 0$ in the intervals $(-\infty, 0)$ and $(2, \infty)$ and the function $f(x)$ is concave upwards in these intervals.

(b) $f(x)$ is concave downwards if $f''(x) < 0$, i.e. $12x(x - 2) < 0$, i.e. $x(x - 2) < 0$ i.e. $x(x - 2) < 0$
Now, $x(x - 2) < 0$ if

Case (I) $x > 0$ and $x-2 < 0$, i.e. $x > 0$ and $x < 2$, i.e. $0 < x < 2$.
Case (II) $x < 0$ and $x-2 > 0$, i.e. $x < 0$ and $x > 2$ but this is not possible.

Hence, $f''(x) < 0$ in the interval $(0, 2)$ and the function $f(x)$ is concave downwards in this interval.

(*iii*) *Points of inflection*
At the points of inflection, $f''(x) = 0$, i.e. $12x(x - 2) = 0$, i.e. $x = 0, x = 2$.
At $x = 0$, $f(x) = 0$

"   -256-   -   -*   ! -   =   !   !   !   !
!11 !   9! !   7 ! :   =" !   A!   ? "   1"

-256-   -   -*   ! -   !   !   J 79H   7 !
:   =" !   1"
($ '

CHAPTER 4

Applications of Definite Integrals

Chapter Outline

4.1 Introduction
4.2 Volume using Cross-sections
4.3 Length of Plane Curves
4.4 Area of Surface of Solid of Revolution

4.1 INTRODUCTION

In this chapter, we will explore few applications of definite integrals. Applications of definite integrals in finding volume by cross-sections, length of plane curves and areas of surfaces of revolution are discussed in detail.

4.2 VOLUME USING CROSS-SECTIONS

The volume of a solid of known integrable cross-section area $A(x)$ formed by a plane perpendicular to the x-axis at any point between $x = a$ to $x = b$ is

$$V = \int_a^b A(x)\mathrm{d}x$$

Note: If the cross-section $A(y)$ is perpendicular to the y-axis at any point between $y = c$ to $y = d$, then the volume of the solid is

$$V = \int_c^d A(y)\mathrm{d}y$$

Example 1

Find the volume of the solid generated by rotating the plane region bounded by $y=\dfrac{1}{x}$, $x=1$ *and* $x = 3$ *about the x-axis.*

Solution

The volume is generated by rotating the dotted region about the x-axis.

The cross-section PQ of the generated volume is a circle of radius y perpendicular to x-axis.

Area of cross-section, $A = \pi y^2$

$$= \pi\cdot\frac{1}{x^2}$$

For the region shown, x varies from 1 to 3.

Volume,
$$V = \int_1^3 \frac{\pi}{x^2}\,\mathrm{d}x$$
$$= \pi\left|-\frac{1}{x}\right|_1^3$$
$$= \pi\left(-\frac{1}{3}+1\right)$$
$$= \frac{2}{3}\pi$$

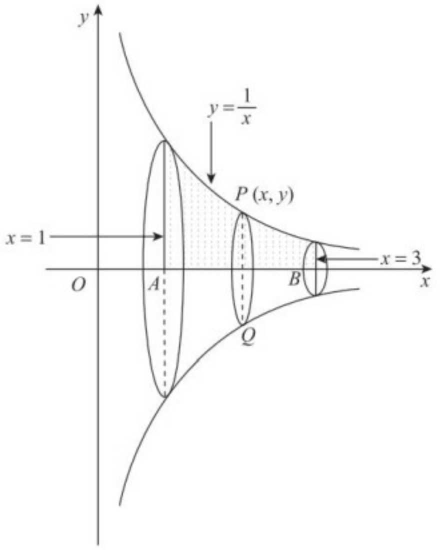

Fig. 4.1

Example 2

Find the volume of the solid that lies between planes perpendicular to the x-axis at $x = 0$ *and* $x = 4$. *The cross-sections perpendicular to the x-axis between these planes run from one side of the parabola* $x = y^2$ *to the other. The cross-sections are squares with bases in the xy-plane.*

Solution

The cross-section of the solid is a square $PQRS$ with side runs from $y=-\sqrt{x}$ to $y=\sqrt{x}$

Length $PQ = 2y = 2\sqrt{x}$

Area of cross-section, $A = (2\sqrt{x})^2 = 4x$

For the region shown, x varies from 0 to 4.

Volume, $$V = \int_0^4 4x\,dx$$

$$= 4\left|\frac{x^2}{2}\right|_0^4$$

$$= 2(16-0)$$

$$= 32$$

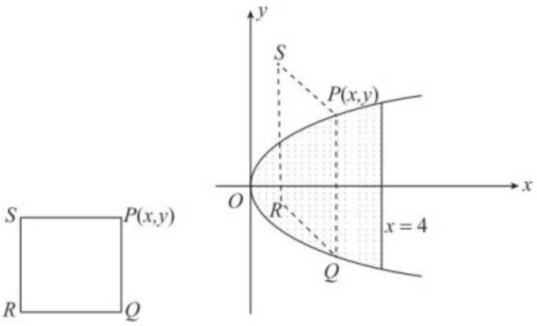

Fig. 4.2

Example 3

Find the volume of a right circular cone of base radius r and height h.

Solution

The cone is generated by rotating the line AB about the y-axis. Equation of the line AB passing through the points $A(r, 0)$ and $B\ (0, h)$ is

$$y - 0 = \frac{h-0}{0-r}(x-r)$$

$$x = r\left(1-\frac{y}{h}\right)$$

The cross-section PQ of the cone is a circle of radius x perpendicular to the y-axis.

Area of cross-section, $A = \pi x^2$

$$= \pi r^2\left(1-\frac{y}{h}\right)^2$$

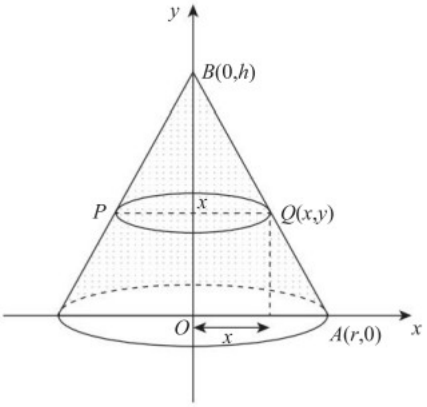

Fig. 4.3

For the region shown, y varies from 0 to h.

Volume, $$V = \int_o^h \pi r^2\left(1-\frac{y}{h}\right)^2 dy$$

$$= \pi r^2\left|\frac{\left(1-\frac{y}{h}\right)^3}{3\left(-\frac{1}{h}\right)}\right|_o^h$$

$$= -\frac{\pi r^2 h}{3}(0-1)$$

$$= \frac{1}{3}\pi r^2 h$$

Example 4

Find the volume of the solid with a circular base of radius 5 and whose cross-sections perpendicular to the base and parallel to the x-axis are equilateral triangles.

Solution

Equation of the circular base of radius 5 is,

$$x^2 + y^2 = 25 \qquad \ldots(1)$$

The cross-section of the solid is an equilateral triangle PQR with base $2x$.

Area of cross-section,
$$\begin{aligned} A &= \frac{1}{2} \times (\text{base}) \times (\text{height}) \\ &= \frac{1}{2}(2x)(h) \\ &= \frac{1}{2}(2x)(x \tan 60^\circ) \\ &= x^2\sqrt{3} \\ &= (25 - y^2)\sqrt{3} \qquad \text{[from Eq. (1)]} \end{aligned}$$

For the region shown, y varies from −5 to 5.

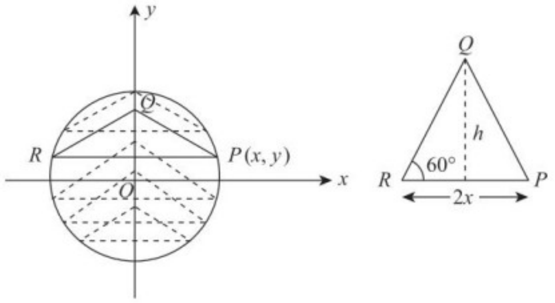

Fig. 4.4

Volume,
$$\begin{aligned} V &= \int_{-5}^{5} \sqrt{3}(25 - y^2)\,dy \\ &= 2\int_{0}^{5} \sqrt{3}(25 - y^2)\,dy \\ &= 2\sqrt{3}\left|25y - \frac{y^3}{3}\right|_0^5 \\ &= 2\sqrt{3}\left(125 - \frac{125}{3}\right) \\ &= \frac{500\sqrt{3}}{3} \end{aligned}$$

Example 5

Use the method of slicing to find the volume of solid with semicircular base defined by $y = 5\sqrt{\cos x}$ *on the interval* $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. *The cross-sections of the solid are squares perpendicular to the x-axis with base running from x-axis to the curve.* **[Winter 2015]**

Solution

The cross-section of the solid is a square with side runs from $y = 0$ (x-axis) to $y = 5\sqrt{\cos x}$.

Length of one side of the square = $5\sqrt{\cos x}$

Area of cross-section, $A = \left(5\sqrt{\cos x}\right)^2$

$= 25\cos^2 x$

In the given interval $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, x varies from $-\frac{\pi}{2}$ to $\frac{\pi}{2}$.

Volume,

$$V = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} 25\cos^2 x\,dx$$

$$= 25\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \left(\frac{1+\cos 2x}{2}\right)dx$$

$$= \frac{25}{2}\left|x + \frac{\sin 2x}{2}\right|_{-\frac{\pi}{2}}^{\frac{\pi}{2}}$$

$$= \frac{25}{2}\left[\left(\frac{\pi}{2} + \frac{\pi}{2}\right) + 0\right] \quad [\because \sin\pi = 0]$$

$$= \frac{25}{2}\pi$$

EXERCISE 4.1

---

1. Find the volume of the solid that lies between planes perpendicular to the y-axis at $y = 0$ and $y = 2$. The cross-sections perpendicular to the

y-axis are circular disks with diameters running from the y-axis to the parabola $x = \sqrt{5}y^2$.

[Ans.: 8π]

2. Find the volume of the solid that lies between the planes perpendicular to the x-axis at $x = -1$ and $x = 1$. The cross-sections perpendicular to the x-axis are circular disks whose diameters run from the parabola $y = x^2$ to the parabola $y = 2 - x^2$.

$\left[\textbf{Ans.: } \frac{16}{15}\pi\right]$

3. Show that the volume of a sphere of radius r is $\frac{4}{3}\pi r^3$.

4. Find the volume of the solid whose base is the region bounded between the curves $y = x$ and $y = x^2$, and whose cross-sections perpendicular to the x-axis are squares.

$\left[\textbf{Ans.: } \frac{1}{30}\right]$

5. Find the volume of the solid whose base is a triangle with vertices (0, 0), (2, 0) and (0, 2) and whose cross-sections perpendicular to the base and parallel to the y-axis are semicircles.

$\left[\textbf{Ans.: } \frac{\pi}{3}\right]$

4.3 LENGTH OF PLANE CURVES

The process of determining the length of an arc of a plane curve is known as rectification of curves.

4.3.1 Length of Arc in Cartesian Form

We know from differential calculus that for the curve $y = f(x)$,

$$\frac{ds}{dx} = \sqrt{1 + \left(\frac{dy}{dx}\right)^2}$$

The length of the arc of the curve $y = f(x)$ between $x = a$ and $x = b$ is given by,

$$s = \int_a^b \frac{ds}{dx}\,dx$$

$$= \int_a^b \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\,dx$$

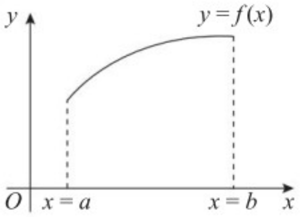

Fig. 4.5

Similarly, the length of the arc of the curve $x = f(y)$ between $y = c$ and $y = d$ is given by,

$$s = \int_c^d \frac{\mathrm{d}s}{\mathrm{d}y}\mathrm{d}y$$
$$= \int_c^d \sqrt{1+\left(\frac{\mathrm{d}x}{\mathrm{d}y}\right)^2}\,\mathrm{d}y$$

Example 1

Show that the length of the arc of the curve $4ax = y^2 - 2a^2\log\frac{y}{a} - a^2$ *from* $(0, a)$ *to any point* (x, y) *is given by* $\frac{y^2}{2a} - \frac{a}{2} - x$.

Solution

$$4ax = y^2 - 2a^2\log\frac{y}{a} - a^2 \qquad \ldots (1)$$
$$4a\frac{\mathrm{d}x}{\mathrm{d}y} = 2y - 2a^2\cdot\frac{a}{y}\cdot\frac{1}{a}$$
$$\frac{\mathrm{d}x}{\mathrm{d}y} = \frac{y}{2a} - \frac{a}{2y} = \frac{y^2-a^2}{2ay}$$

For the required arc, y varies from a to y.

Length of the arc, $$s = \int_a^y \sqrt{1+\left(\frac{\mathrm{d}x}{\mathrm{d}y}\right)^2}\,\mathrm{d}y$$
$$= \int_a^y \sqrt{1+\left(\frac{y^2-a^2}{2ay}\right)^2}\,\mathrm{d}y$$
$$= \int_a^y \sqrt{\frac{(y^2+a^2)^2}{(2ay)^2}}\,\mathrm{d}y$$
$$= \frac{1}{2a}\int_a^y \frac{y^2+a^2}{y}\mathrm{d}y$$
$$= \frac{1}{2a}\left(\int_a^y y\,\mathrm{d}y + a^2\int_a^y \frac{1}{y}\mathrm{d}y\right)$$
$$= \frac{1}{2a}\left|\frac{y^2}{2} + a^2\log y\right|_a^y$$

$$= \frac{1}{2a}\left(\frac{y^2}{2} + a^2 \log\frac{y}{a} - \frac{a^2}{2}\right)$$

$$= \frac{1}{2a}\left[\frac{y^2}{2} + \left(\frac{y^2}{2} - 2ax - \frac{a^2}{2}\right) - \frac{a^2}{2}\right] \qquad \text{… [From Eq. (1)]}$$

$$= \frac{1}{2a}(y^2 - 2ax - a^2)$$

$$= \frac{y^2}{2a} - x - \frac{a}{2}$$

$$= \frac{y^2}{2a} - \frac{a}{2} - x$$

Example 2

Find the length of the arc of the curve y = e^x from the point (0, 1) *to* (1, *e*).

Solution

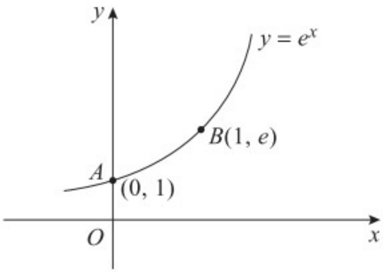

Fig. 4.6

$$y = e^x$$

$$\frac{dy}{dx} = e^x$$

For the required arc, x varies from 0 to 1.

Length of the arc AB, $$s = \int_0^1 \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx$$

$$= \int_0^1 \sqrt{1 + e^{2x}}\, dx$$

Putting $1 + e^{2x} = t^2$,

$$2e^{2x} dx = 2t\, dt$$

$$dx = \frac{t}{t^2 - 1} dt$$

When $x = 0, \quad t = \sqrt{2}$

When $x = 1, \quad t = \sqrt{1 + e^2}$

Length of the arc, $$s = \int_{\sqrt{2}}^{\sqrt{1+e^2}} t \cdot \frac{t}{t^2 - 1} dt$$

$$= \int_{\sqrt{2}}^{\sqrt{1+e^2}} \frac{t^2 - 1 + 1}{t^2 - 1} dt$$

$$= \int_{\sqrt{2}}^{\sqrt{1+e^2}} \left(1 + \frac{1}{t^2 - 1}\right) dt$$

$$= \left| t + \frac{1}{2}\log\frac{t-1}{t+1} \right|_{\sqrt{2}}^{\sqrt{1+e^2}}$$

$$= \sqrt{1+e^2} - \sqrt{2} + \frac{1}{2}\left(\log\frac{\sqrt{1+e^2}-1}{\sqrt{1+e^2}+1} - \log\frac{\sqrt{2}-1}{\sqrt{2}+1} \right)$$

$$= \sqrt{1+e^2} - \sqrt{2} + \frac{1}{2}\left[\log\left\{ \frac{\left(\sqrt{1+e^2}-1\right)^2}{1+e^2-1} \right\} - \log\left\{ \frac{\left(\sqrt{2}-1\right)^2}{2-1} \right\} \right]$$

$$= \sqrt{1+e^2} - \sqrt{2} + \log\left(\sqrt{1+e^2}-1\right) - \frac{1}{2}\log e^2 - \log\left(\sqrt{2}-1\right)$$

$$= \sqrt{1+e^2} - \sqrt{2} + \log\left(\sqrt{1+e^2}-1\right) - 1 - \log\left(\sqrt{2}-1\right)$$

$[\because \log e^2 = 2 \log e = 2]$

Example 3

Find the length of the arc of the curve $y = \log \sec x$ *from* $x = 0$ *to* $x = \frac{\pi}{3}$.

Solution

$$y = \log \sec x$$

$$\frac{dy}{dx} = \frac{1}{\sec x} \cdot \sec x \tan x = \tan x$$

For the required arc, x varies from 0 to $\frac{\pi}{3}$.

Length of the arc, $s = \int_0^{\frac{\pi}{3}} \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx$

$$= \int_0^{\frac{\pi}{3}} \sqrt{1 + \tan^2 x}\, dx$$

$$= \int_0^{\frac{\pi}{3}} \sec x\, dx$$

$$= \left| \log(\sec x + \tan x) \right|_0^{\frac{\pi}{3}}$$

$$= \log\left(2 + \sqrt{3}\right)$$

Example 4

Find the length of the arc of the curve $y = \log\left(\frac{e^x-1}{e^x+1}\right)$ *from* $x = 1$ *to* $x = 2$.

-256- - -* ! - 2"A! " = ! !
7 K! ! 9 ! H A 3 7 ! : =" ! & " ! A! ? "
1" (B% '%

(%

data

Ⓢ Ⓣ Ⓠ Ⓠ Ⓢ Ⓢ Ⓓ

Curve Sketching

b) Engineering mathematics..(2010).pdf

Obs: Gauss – 28 ocurrences.

c) Mathematics-1 ... (2018).pdf

obs: Gauss – 33 ocurrences

Pág. 149 – 4.2

Volume using cross-sections

Pág. 154 – 4.3

Length of Plane Curves

d) Advanced Engineering Mathematics... (2018).pdf

Pág 163 - Chapter 3

Ordinary Differential Equations and Applications

obs: Pág 163 – 428.

Obs: other: Fourier Series
Laplace Transforms
Partial Differential Equations and Applications

%   (%/

20211018/

Número racional decimal
exato.

baffled, (Mystified) enganado (decived)?
puzzled ou mistaken (≅

etimológico ou psicológico?

→ burck@ksu.edu

burckel@math.ksu.edu

Books:

20211018/

a) Single Variable Calculus Early Transcendentals/
Single Variable Calculus...pdf

Pág. 257(292) Figure 5 - Differentials

Obs: Gauss - 11 occurences
Riemann - 76 occurences
Leibniz - 46 occurences

Important:

The geometric meaning of differentials is shown in Figure 5. Let $P(x, f(x))$ and $Q(x + \Delta x, f(x + \Delta x))$ be points on the graph of f and let $dx = \Delta x$. The corresponding change in y is

$$\Delta y = f(x + \Delta x) - f(x)$$

The slope of the tangent line PR is the derivative $f'(x)$. Thus the directed distance from S to R is $f'(x)\,dx = dy$. Therefore dy represents the amount that the tangent line rises or falls (the change in the linearization), whereas Δy represents the amount that the curve $y = f(x)$ rises or falls when x changes by an amount dx.

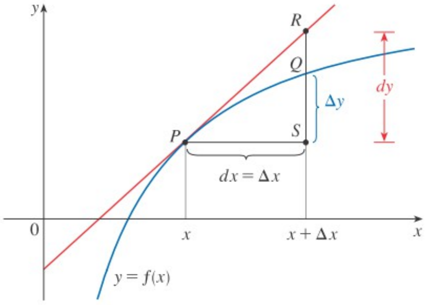

FIGURE 5

EXAMPLE 3 Compare the values of Δy and dy if $y = f(x) = x^3 + x^2 - 2x + 1$ and x changes (a) from 2 to 2.05 and (b) from 2 to 2.01.

SOLUTION

(a) We have

$$f(2) = 2^3 + 2^2 - 2(2) + 1 = 9$$

$$f(2.05) = (2.05)^3 + (2.05)^2 - 2(2.05) + 1 = 9.717625$$

$$\Delta y = f(2.05) - f(2) = 0.717625$$

Figure 6 shows the function in Example 3 and a comparison of dy and Δy when $a = 2$. The viewing rectangle is $[1.8, 2.5]$ by $[6, 18]$.

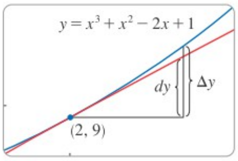

FIGURE 6

In general, $dy = f'(x)\,dx = (3x^2 + 2x - 2)\,dx$

When $x = 2$ and $dx = \Delta x = 0.05$, this becomes

$$dy = [3(2)^2 + 2(2) - 2]0.05 = 0.7$$

(b) $$f(2.01) = (2.01)^3 + (2.01)^2 - 2(2.01) + 1 = 9.140701$$

$$\Delta y = f(2.01) - f(2) = 0.140701$$

When $dx = \Delta x = 0.01$,

$$dy = [3(2)^2 + 2(2) - 2]0.01 = 0.14$$ ■

Notice that the approximation $\Delta y \approx dy$ becomes better as Δx becomes smaller in Example 3. Notice also that dy was easier to compute than Δy.

In the notation of differentials, the linear approximation $f(x) \approx f(a) + f'(a)(x - a)$ can be written as

$$f(a + dx) \approx f(a) + dy$$

Copyright 2021 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

the sequence has a limit. What can you say about the value of the limit?

78–84 Determine whether the sequence is increasing, decreasing, or not monotonic. Is the sequence bounded?

78. $a_n = \cos n$

79. $a_n = \dfrac{1}{2n + 3}$

80. $a_n = \dfrac{1 - n}{2 + n}$

81. $a_n = n(-1)^n$

82. $a_n = 2 + \dfrac{(-1)^n}{n}$

83. $a_n = 3 - 2ne^{-n}$

84. $a_n = n^3 - 3n + 3$

---

85. Find the limit of the sequence

$$\left\{\sqrt{2}, \sqrt{2\sqrt{2}}, \sqrt{2\sqrt{2\sqrt{2}}}, \ldots\right\}$$

86. A sequence $\{a_n\}$ is given by $a_1 = \sqrt{2}$, $a_{n+1} = \sqrt{2 + a_n}$.

(a) By induction or otherwise, show that $\{a_n\}$ is increasing and bounded above by 3. Apply the Monotonic Sequence Theorem to show that $\lim_{n\to\infty} a_n$ exists.

(b) Find $\lim_{n\to\infty} a_n$.

87. Show that the sequence defined by

$$a_1 = 1 \qquad a_{n+1} = 3 - \frac{1}{a_n}$$

is increasing and $a_n < 3$ for all n. Deduce that $\{a_n\}$ is convergent and find its limit.

88. Show that the sequence defined by

$$a_1 = 2 \qquad a_{n+1} = \frac{1}{3 - a_n}$$

satisfies $0 < a_n \leqslant 2$ and is decreasing. Deduce that the sequence is convergent and find its limit.

89. (a) Fibonacci posed the following problem:

> Suppose that rabbits live forever and that every month each pair produces a new pair which becomes productive at age 2 months. If we start with one newborn pair, how many pairs of rabbits will we have in the nth month?

Show that the answer is f_n, where $\{f_n\}$ is the Fibonacci sequence defined in Example 3(c).

(b) Let $a_n = f_{n+1}/f_n$ and show that $a_{n-1} = 1 + 1/a_{n-2}$. Assuming that $\{a_n\}$ is convergent, find its limit.

90. (a) Let $a_1 = a$, $a_2 = f(a)$, $a_3 = f(a_2) = f(f(a))$, . . . , $a_{n+1} = f(a_n)$, where f is a continuous function. If $\lim_{n\to\infty} a_n = L$, show that $f(L) = L$.

(b) Illustrate part (a) by taking $f(x) = \cos x$, $a = 1$, and estimating the value of L to five decimal places.

91. (a) Use a graph to guess the value of the limit

$$\lim_{n\to\infty} \frac{n^5}{n!}$$

(b) Use a graph of the sequence in part (a) to find the smallest values of N that correspond to $\varepsilon = 0.1$ and $\varepsilon = 0.001$ in Definition 2.

92. Use Definition 2 directly to prove that $\lim_{n\to\infty} r^n = 0$ when $|r| < 1$.

93. Prove Theorem 6.
[*Hint:* Use either Definition 2 or the Squeeze Theorem.]

94. Use Theorem 7 to prove the Power Law:

$$\lim_{n\to\infty} a_n^p = \left[\lim_{n\to\infty} a_n\right]^p \qquad \text{if } p > 0 \text{ and } a_n > 0$$

95. Prove that if $\lim_{n\to\infty} a_n = 0$ and $\{b_n\}$ is bounded, then $\lim_{n\to\infty} (a_n b_n) = 0$.

96. Let $a_n = (1 + 1/n)^n$.

(a) Show that if $0 \leqslant a < b$, then

$$\frac{b^{n+1} - a^{n+1}}{b - a} < (n + 1)b^n$$

(b) Deduce that $b^n[(n + 1)a - nb] < a^{n+1}$.

(c) Use $a = 1 + 1/(n + 1)$ and $b = 1 + 1/n$ in part (b) to show that $\{a_n\}$ is increasing.

(d) Use $a = 1$ and $b = 1 + 1/(2n)$ in part (b) to show that $a_{2n} < 4$.

(e) Use parts (c) and (d) to show that $a_n < 4$ for all n.

(f) Use Theorem 12 to show that $\lim_{n\to\infty} (1 + 1/n)^n$ exists. (The limit is e. See Equation 3.6.6.)

97. Let a and b be positive numbers with $a > b$. Let a_1 be their arithmetic mean and b_1 their geometric mean:

$$a_1 = \frac{a + b}{2} \qquad b_1 = \sqrt{ab}$$

Repeat this process so that, in general,

$$a_{n+1} = \frac{a_n + b_n}{2} \qquad b_{n+1} = \sqrt{a_n b_n}$$

(a) Use mathematical induction to show that

$$a_n > a_{n+1} > b_{n+1} > b_n$$

(b) Deduce that both $\{a_n\}$ and $\{b_n\}$ are convergent.

(c) Show that $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n$. Gauss called the common value of these limits the **arithmetic-geometric mean** of the numbers a and b.

98. (a) Show that if $\lim_{n\to\infty} a_{2n} = L$ and $\lim_{n\to\infty} a_{2n+1} = L$, then $\{a_n\}$ is convergent and $\lim_{n\to\infty} a_n = L$.

(b) If $a_1 = 1$ and

$$a_{n+1} = 1 + \frac{1}{1 + a_n}$$

find the first eight terms of the sequence $\{a_n\}$. Then use

Copyright 2021 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

WRITING PROJECT | HOW NEWTON DISCOVERED THE BINOMIAL SERIES

The Binomial Theorem, which gives the expansion of $(a + b)^k$, was known to Chinese mathematicians many centuries before the time of Newton for the case where the exponent k is a positive integer. In 1665, when he was 22, Newton was the first to discover the infinite series expansion of $(a + b)^k$ when k is a fractional exponent (positive or negative). He didn't publish his discovery, but he stated it and gave examples of how to use it in a letter (now called the *epistola prior*) dated June 13, 1676, that he sent to Henry Oldenburg, secretary of the Royal Society of London, to transmit to Leibniz. When Leibniz replied, he asked how Newton had discovered the binomial series. Newton wrote a second letter, the *epistola posterior* of October 24, 1676, in which he explained in great detail how he arrived at his discovery by a very indirect route. He was investigating the areas under the curves $y = (1 - x^2)^{n/2}$ from 0 to x for $n = 0, 1, 2, 3, 4, \ldots$. These are easy to calculate if n is even. By observing patterns and interpolating, Newton was able to guess the answers for odd values of n. Then he realized he could get the same answers by expressing $(1 - x^2)^{n/2}$ as an infinite series.

Write an essay on Newton's discovery of the binomial series. Start by giving the statement of the binomial series in Newton's notation (see the *epistola prior* on page 285 of [4] or page 402 of [2]). Explain why Newton's version is equivalent to Theorem 11.10.17. Then read Newton's *epistola posterior* (page 287 in [4] or page 404 in [2]) and explain the patterns that Newton discovered in the areas under the curves $y = (1 - x^2)^{n/2}$. Show how he was able to guess the areas under the remaining curves and how he verified his answers. Finally, explain how these discoveries led to the binomial series. The books by Edwards [1] and Katz [3] contain commentaries on Newton's letters.

1. C. H. Edwards, Jr., *The Historical Development of the Calculus* (New York: Springer-Verlag, 1979), pp. 178–87.

2. Jahn Fauvel and Jeremy Gray, eds., *The History of Mathematics: A Reader* (Basingstoke, UK: MacMillan Education, 1987).

3. Victor Katz, *A History of Mathematics: An Introduction*, 3rd ed. (Boston: Addison-Wesley, 2009), pp. 543–82.

4. D. J. Struik, ed., *A Source Book in Mathematics, 1200–1800* (Cambridge, MA: Harvard University Press, 1969).

11.11 Applications of Taylor Polynomials

In this section we explore two types of applications of Taylor polynomials. First we look at how they are used to approximate functions—computer scientists employ them because polynomials are the simplest of functions. Then we investigate how physicists and engineers use them in such fields as relativity, optics, blackbody radiation, electric dipoles, the velocity of water waves, and building highways across a desert.

■ Approximating Functions by Polynomials

Suppose that $f(x)$ is equal to the sum of its Taylor series at a:

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x - a)^n$$

Copyright 2021 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

In Section 11.10 we introduced the notation $T_n(x)$ for the nth partial sum of this series and called it the nth-degree Taylor polynomial of f at a. Thus

$$T_n(x) = \sum_{i=0}^{n} \frac{f^{(i)}(a)}{i!}(x-a)^i$$

$$= f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n$$

Since f is the sum of its Taylor series, we know that $T_n(x) \to f(x)$ as $n \to \infty$ and so T_n can be used as an approximation to f: $f(x) \approx T_n(x)$.

Notice that the first-degree Taylor polynomial

$$T_1(x) = f(a) + f'(a)(x-a)$$

is the same as the linearization of f at a that we discussed in Section 3.10. Notice also that T_1 and its derivative have the same values at a that f and f' have. In general, it can be shown that the derivatives of T_n at a agree with those of f up to and including derivatives of order n.

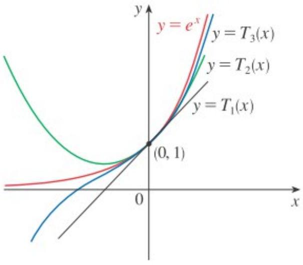

FIGURE 1

	$x = 0.2$	$x = 3.0$
$T_2(x)$	1.220000	8.500000
$T_4(x)$	1.221400	16.375000
$T_6(x)$	1.221403	19.412500
$T_8(x)$	1.221403	20.009152
$T_{10}(x)$	1.221403	20.079665
e^x	1.221403	20.085537

To illustrate these ideas let's take another look at the graphs of $y = e^x$ and its first few Taylor polynomials, as shown in Figure 1. The graph of T_1 is the tangent line to $y = e^x$ at $(0, 1)$; this tangent line is the best linear approximation to e^x near $(0, 1)$. The graph of T_2 is the parabola $y = 1 + x + x^2/2$, and the graph of T_3 is the cubic curve $y = 1 + x + x^2/2 + x^3/6$, which is a closer fit to the exponential curve $y = e^x$ than T_2. The next Taylor polynomial T_4 would be an even better approximation, and so on.

The values in the table give a numerical demonstration of the convergence of the Taylor polynomials $T_n(x)$ to the function $y = e^x$. We see that when $x = 0.2$ the convergence is very rapid, but when $x = 3$ it is somewhat slower. In fact, the farther x is from 0, the more slowly $T_n(x)$ converges to e^x.

When using a Taylor polynomial T_n to approximate a function f, we have to ask the questions: How good an approximation is it? How large should we take n to be in order to achieve a desired accuracy? To answer these questions we need to look at the absolute value of the remainder:

$$|R_n(x)| = |f(x) - T_n(x)|$$

There are three possible methods for estimating the size of the error:

1. We can use a calculator or computer to graph $|R_n(x)| = |f(x) - T_n(x)|$ and thereby estimate the error.
2. If the series happens to be an alternating series, we can use the Alternating Series Estimation Theorem.
3. In all cases we can use Taylor's Inequality (Theorem 11.10.9), which says that if $|f^{(n+1)}(x)| \leqslant M$, then

$$|R_n(x)| \leqslant \frac{M}{(n+1)!}|x-a|^{n+1}$$

EXAMPLE 1

(a) Approximate the function $f(x) = \sqrt[3]{x}$ by a Taylor polynomial of degree 2 at $a = 8$.
(b) How accurate is this approximation when $7 \leqslant x \leqslant 9$?

Copyright 2021 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

SOLUTION

(a)

$$\begin{aligned} f(x) &= \sqrt[3]{x} = x^{1/3} & f(8) &= 2 \\ f'(x) &= \tfrac{1}{3}x^{-2/3} & f'(8) &= \tfrac{1}{12} \\ f''(x) &= -\tfrac{2}{9}x^{-5/3} & f''(8) &= -\tfrac{1}{144} \\ f'''(x) &= \tfrac{10}{27}x^{-8/3} \end{aligned}$$

Thus the second-degree Taylor polynomial is

$$\begin{aligned} T_2(x) &= f(8) + \frac{f'(8)}{1!}(x-8) + \frac{f''(8)}{2!}(x-8)^2 \\ &= 2 + \tfrac{1}{12}(x-8) - \tfrac{1}{288}(x-8)^2 \end{aligned}$$

The desired approximation is

$$\sqrt[3]{x} \approx T_2(x) = 2 + \tfrac{1}{12}(x-8) - \tfrac{1}{288}(x-8)^2$$

(b) The Taylor series is not alternating when $x < 8$, so we can't use the Alternating Series Estimation Theorem in this example. But we can use Taylor's Inequality with $n = 2$ and $a = 8$:

$$|R_2(x)| \leqslant \frac{M}{3!}|x-8|^3$$

where $|f'''(x)| \leqslant M$. Because $x \geqslant 7$, we have $x^{8/3} \geqslant 7^{8/3}$ and so

$$f'''(x) = \frac{10}{27} \cdot \frac{1}{x^{8/3}} \leqslant \frac{10}{27} \cdot \frac{1}{7^{8/3}} < 0.0021$$

Therefore we can take $M = 0.0021$. Also $7 \leqslant x \leqslant 9$, so $-1 \leqslant x - 8 \leqslant 1$ and $|x - 8| \leqslant 1$. Then Taylor's Inequality gives

$$|R_2(x)| \leqslant \frac{0.0021}{3!} \cdot 1^3 = \frac{0.0021}{6} < 0.0004$$

Thus, if $7 \leqslant x \leqslant 9$, the approximation in part (a) is accurate to within 0.0004. ■

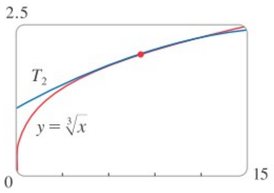

FIGURE 2

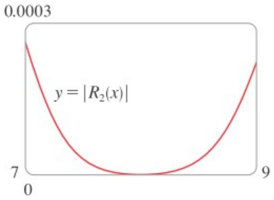

FIGURE 3

Let's check the calculation in Example 1 graphically. Figure 2 shows that the graphs of $y = \sqrt[3]{x}$ and $y = T_2(x)$ are very close to each other when x is near 8. Figure 3 shows the graph of $|R_2(x)|$ computed from the expression

$$|R_2(x)| = |\sqrt[3]{x} - T_2(x)|$$

We see from the graph that

$$|R_2(x)| < 0.0003$$

when $7 \leqslant x \leqslant 9$. Thus the error estimate from graphical methods is slightly better than the error estimate from Taylor's Inequality in this case.

EXAMPLE 2

(a) What is the maximum error possible in using the approximation

$$\sin x \approx x - \frac{x^3}{3!} + \frac{x^5}{5!}$$

when $-0.3 \leqslant x \leqslant 0.3$? Use this approximation to find sin 12° correct to six decimal places.

(b) For what values of x is this approximation accurate to within 0.00005?

Copyright 2021 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

SOLUTION

(a) Notice that the Maclaurin series

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

is alternating for all nonzero values of x, and the successive terms decrease in size because $|x| < 1$, so we can use the Alternating Series Estimation Theorem. The error in approximating $\sin x$ by the first three terms of its Maclaurin series is at most

$$\left| \frac{x^7}{7!} \right| = \frac{|x|^7}{5040}$$

If $-0.3 \leqslant x \leqslant 0.3$, then $|x| \leqslant 0.3$, so the error is smaller than

$$\frac{(0.3)^7}{5040} \approx 4.3 \times 10^{-8}$$

To find sin 12° we first convert to radian measure:

$$\begin{aligned} \sin 12° &= \sin\left(\frac{12\pi}{180}\right) = \sin\left(\frac{\pi}{15}\right) \\ &\approx \frac{\pi}{15} - \left(\frac{\pi}{15}\right)^3 \frac{1}{3!} + \left(\frac{\pi}{15}\right)^5 \frac{1}{5!} \approx 0.20791169 \end{aligned}$$

Thus, correct to six decimal places, $\sin 12° \approx 0.207912$.

(b) The error will be smaller than 0.00005 if

$$\frac{|x|^7}{5040} < 0.00005$$

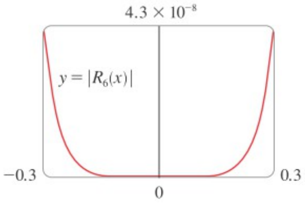

FIGURE 4

Solving this inequality for x, we get

$$|x|^7 < 0.252 \qquad \text{or} \qquad |x| < (0.252)^{1/7} \approx 0.821$$

So the given approximation is accurate to within 0.00005 when $|x| < 0.82$. ■

What if we use Taylor's Inequality to solve Example 2? Since $f^{(7)}(x) = -\cos x$, we have $|f^{(7)}(x)| \leqslant 1$ and so

$$|R_6(x)| \leqslant \frac{1}{7!}|x|^7$$

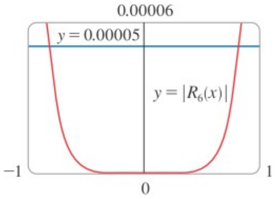

FIGURE 5

So we get the same estimates as with the Alternating Series Estimation Theorem.

What about graphical methods? Figure 4 shows the graph of

$$|R_6(x)| = \left|\sin x - \left(x - \tfrac{1}{6}x^3 + \tfrac{1}{120}x^5\right)\right|$$

and we see from it that $|R_6(x)| < 4.3 \times 10^{-8}$ when $|x| \leqslant 0.3$. This is the same estimate that we obtained in Example 2. For part (b) we want $|R_6(x)| < 0.00005$, so we graph both $y = |R_6(x)|$ and $y = 0.00005$ in Figure 5. From the coordinates of the right intersection point we find that the inequality is satisfied when $|x| < 0.82$. Again this is the same estimate that we obtained in the solution to Example 2.

Copyright 2021 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

If we had been asked to approximate sin 72° instead of sin 12° in Example 2, it would have been wise to use the Taylor polynomials at $a = \pi/3$ (instead of $a = 0$) because they are better approximations to $\sin x$ for values of x close to $\pi/3$. Notice that 72° is close to 60° (or $\pi/3$ radians) and the derivatives of $\sin x$ are easy to compute at $\pi/3$.

Figure 6 shows the graphs of the Maclaurin polynomial approximations

$$T_1(x) = x \qquad\qquad T_3(x) = x - \frac{x^3}{3!}$$

$$T_5(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} \qquad\qquad T_7(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!}$$

to the sine curve. You can see that as n increases, $T_n(x)$ is a good approximation to $\sin x$ on a larger and larger interval.

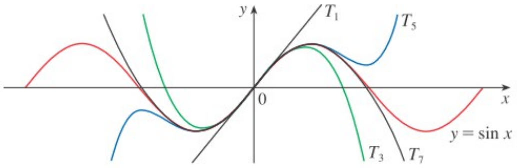

FIGURE 6

One use of the type of calculation done in Examples 1 and 2 occurs in calculators and computers. For instance, when you press the sin or e^x key on your calculator, or when a computer programmer uses a subroutine for a trigonometric or exponential or Bessel function, in many machines a polynomial approximation is calculated. The polynomial is often a Taylor polynomial that has been modified so that the error is spread more evenly throughout an interval.

Applications to Physics

Taylor polynomials are also used frequently in physics. In order to gain insight into an equation, a physicist often simplifies a function by considering only the first two or three terms in its Taylor series. In other words, the physicist uses a Taylor polynomial as an approximation to the function. Taylor's Inequality can then be used to gauge the accuracy of the approximation. The following example shows one way in which this idea is used in special relativity.

EXAMPLE 3 In Einstein's theory of special relativity the mass m of an object moving with velocity v is

$$m = \frac{m_0}{\sqrt{1 - v^2/c^2}}$$

where m_0 is the mass of the object when at rest and c is the speed of light. The kinetic energy k of the object is the difference between its total energy and its energy at rest:

$$K = mc^2 - m_0c^2$$

(a) Show that when v is very small compared with c, this expression for K agrees with classical Newtonian physics: $K = \frac{1}{2}m_0v^2$.

Copyright 2021 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(b) Use Taylor's Inequality to estimate the difference in these expressions for K when $|v| \leqslant 100$ m/s.

SOLUTION

(a) Using the expressions given for K and m, we get

$$K = mc^2 - m_0c^2 = \frac{m_0c^2}{\sqrt{1 - v^2/c^2}} - m_0c^2 = m_0c^2\left[\left(1 - \frac{v^2}{c^2}\right)^{-1/2} - 1\right]$$

With $x = -v^2/c^2$, the Maclaurin series for $(1 + x)^{-1/2}$ is most easily computed as a binomial series with $k = -\frac{1}{2}$. (Notice that $|x| < 1$ because $v < c$.) Therefore we have

The upper curve in Figure 7 is the graph of the expression for the kinetic energy K of an object with velocity v in special relativity. The lower curve shows the function used for K in classical Newtonian physics. When v is much smaller than the speed of light, the curves are practically identical.

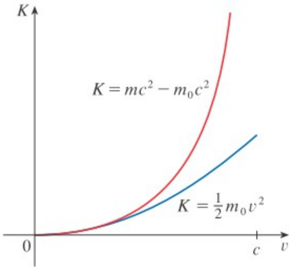

FIGURE 7

$$\begin{aligned}(1 + x)^{-1/2} &= 1 - \tfrac{1}{2}x + \frac{\left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)}{2!}x^2 + \frac{\left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)\left(-\frac{5}{2}\right)}{3!}x^3 + \cdots \\ &= 1 - \tfrac{1}{2}x + \tfrac{3}{8}x^2 - \tfrac{5}{16}x^3 + \cdots\end{aligned}$$

and

$$\begin{aligned}K &= m_0c^2\left[\left(1 + \frac{1}{2}\frac{v^2}{c^2} + \frac{3}{8}\frac{v^4}{c^4} + \frac{5}{16}\frac{v^6}{c^6} + \cdots\right) - 1\right] \\ &= m_0c^2\left(\frac{1}{2}\frac{v^2}{c^2} + \frac{3}{8}\frac{v^4}{c^4} + \frac{5}{16}\frac{v^6}{c^6} + \cdots\right)\end{aligned}$$

If v is much smaller than c, then all terms after the first are very small when compared with the first term. If we omit them, we get

$$K \approx m_0c^2\left(\frac{1}{2}\frac{v^2}{c^2}\right) = \tfrac{1}{2}m_0v^2$$

(b) If $x = -v^2/c^2$, $f(x) = m_0c^2[(1 + x)^{-1/2} - 1]$, and M is a number such that $|f''(x)| \leqslant M$, then we can use Taylor's Inequality to write

$$|R_1(x)| \leqslant \frac{M}{2!}x^2$$

We have $f''(x) = \frac{3}{4}m_0c^2(1 + x)^{-5/2}$ and we are given that $|v| \leqslant 100$ m/s, so

$$|f''(x)| = \frac{3m_0c^2}{4(1 - v^2/c^2)^{5/2}} \leqslant \frac{3m_0c^2}{4(1 - 100^2/c^2)^{5/2}} \quad (=M)$$

Thus, with $c = 3 \times 10^8$ m/s,

$$|R_1(x)| \leqslant \frac{1}{2} \cdot \frac{3m_0c^2}{4(1 - 100^2/c^2)^{5/2}} \cdot \frac{100^4}{c^4} < (4.17 \times 10^{-10})m_0$$

So when $|v| \leqslant 100$ m/s, the magnitude of the error in using the Newtonian expression for kinetic energy is at most $(4.2 \times 10^{-10})m_0$. ■

Copyright 2021 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Another application to physics occurs in optics. Figure 8 depicts a wave from the point source S meeting a spherical interface of radius R centered at C. The ray SA is refracted toward P.

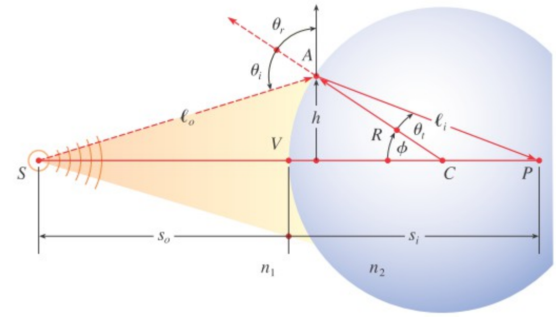

FIGURE 8
Refraction at a spherical interface

Using Fermat's principle that light travels so as to minimize the time taken, one can derive the equation

$$\boxed{1} \qquad \frac{n_1}{\ell_o} + \frac{n_2}{\ell_i} = \frac{1}{R}\left(\frac{n_2 s_i}{\ell_i} - \frac{n_1 s_o}{\ell_o}\right)$$

where n_1 and n_2 are indexes of refraction and ℓ_o, ℓ_i, s_o, and s_i are the distances indicated in Figure 8. By the Law of Cosines, applied to triangles ACS and ACP, we have

$$\boxed{2} \qquad \ell_o = \sqrt{R^2 + (s_o + R)^2 - 2R(s_o + R)\cos\phi}$$

$$\ell_i = \sqrt{R^2 + (s_i - R)^2 + 2R(s_i - R)\cos\phi}$$

Here we use the identity

$$\cos(\pi - \phi) = -\cos\phi$$

Because Equation 1 is cumbersome to work with, Gauss, in 1841, simplified it by using the linear approximation $\cos\phi \approx 1$ for small values of ϕ. (This amounts to using the Taylor polynomial of degree 1.) Then Equation 1 becomes the following simpler equation [as you are asked to show in Exercise 34(a)]:

$$\boxed{3} \qquad \frac{n_1}{s_o} + \frac{n_2}{s_i} = \frac{n_2 - n_1}{R}$$

The resulting optical theory is known as *Gaussian optics*, or *first-order optics*, and has become the basic theoretical tool used to design lenses.

A more accurate theory is obtained by approximating $\cos\phi$ by its Taylor polynomial of degree 3 (which is the same as the Taylor polynomial of degree 2). This takes into account rays for which ϕ is not so small, that is, rays that strike the surface at greater distances h above the axis. In Exercise 34(b) you are asked to use this approximation to derive the more accurate equation

$$\boxed{4} \qquad \frac{n_1}{s_o} + \frac{n_2}{s_i} = \frac{n_2 - n_1}{R} + h^2\left[\frac{n_1}{2s_o}\left(\frac{1}{s_o} + \frac{1}{R}\right)^2 + \frac{n_2}{2s_i}\left(\frac{1}{R} - \frac{1}{s_i}\right)^2\right]$$

The resulting optical theory is known as *third-order optics.*

Other applications of Taylor polynomials to physics and engineering are explored in Exercises 32, 33, 35, 36, 37, and 38, and in the Applied Project following this section.

Copyright 2021 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

(%/

Pág 737 (772)

97. Let a an b be positive numbers with $a > b$.
Let a_1 be their arithmetic mean and b_1 their geometric mean:

$$a_1 = \frac{a+b}{2} \qquad b_1 = \sqrt{ab}$$

Repeat this process so that, in general,

$$a_{n+1} = \frac{a_n + b_n}{2} \qquad b_{n+1} = \sqrt{a_n b_n}$$

(a) Use mathematical induction to show that

$$a_n > a_{n+1} > b_{n+1} > b_n$$

(b) Deduce that both $\{a_n\}$ and $\{b_n\}$ are convergent.

(c) Show that $\lim_{n \to \infty} a_n = \lim_{n \to \infty} b_n$.

⇒ Gauss called the common value of these limits the arithmetic-geometric mean of the numbers a and b.

(%/

Pág. 811 (846) -

How Newton discovered the binomial series

$(a+b)^K$

Newton was the first to discover the infinite series expansion of $(a+b)^K$ when K is a fractional exponent (positive or negative).

He didn't publish his discovery, but he stated it and gave examples of how to use it in a letter (now called the epistola prior) dated June 13, 1676, that he sent to Henry Oldenburg, secretary of the Royal Society of London, to transmit to Leibniz.

11.11 Aplications of Taylor Polynomials

...

Pág. 817 (852) Figure 8

Depicts a wave from the point S meeting a spherical interface of radius R centered at C.

The ray SA is refracted toward P.

Important:

(%/

Using Fermat's principle that light travels so as to minimize the time taken, one can derive the equation

$$(1)\quad \frac{n_1}{l_0}+\frac{n_2}{l_i}=\frac{1}{R}\left(\frac{n_2 S_i}{l_i}-\frac{n_1 S_o}{l_o}\right)$$

where n_1 and n_2 are indexes of refraction l_o, l_i, S_o and S_i are the distances indicated in Figure 8.

By the law of Cosines, applied to triangles ACS and ACP, we have

$$(2)\quad \begin{aligned} l_0 &= \sqrt{R^2+(S_o+R)^2-2R(S_o+R)\cos\phi} \\ l_i &= \sqrt{R^2+(S_i-R)^2+2R(S_i-R)\cos\phi} \end{aligned}$$

Because Equation 1 is cumbersome to work with, Gauss in 1841, simplified by using the linear ~~tran~~ approximation $\cos\phi \approx 1$ for small values of ϕ. (clear!!!

(This amounts to using the Taylor polynomial of degree (1).

Then Equation 1 becomes the following simpler equation [as you are asked to show in Exercise 34(a)]:

(%/%

(3) $\frac{n_1}{s_0} + \frac{n_2}{s_1} = \frac{n_2 - n_1}{R}$

the resulting optical theory is known as Gaussian optics, or first-order optics, and has become the basic theoretical tool used to design lenses.

Obs: A more accurate theory is obtained by approximating cos ϕ by its Taylor polynomial of degree 3

(which is the same as the Taylor polynomial of degree 2)

Important:

Usar isto para calcular o valor preciso observando-se a curvatura do tempo-espaço.

Usar toda esta sistemática destas equações para simular os diversos valores da curvatura do tempo-espaço e suas implicações sobre o [illegible]).

This takes into accounts rays for which ϕ is not so small, that is, rays that strike the surface at greater distances h above the axis.

In Exercise 34(b) you are asked to use this approximation to derive the more accurate equation

' (%/'

(4) $\frac{n_1}{s_0}+\frac{n_2}{s_1}=\frac{n_2-n_1}{R}+$

$$h^2\left[\frac{n_1}{2s_0}\left(\frac{1}{s_0}+\frac{1}{R}\right)^2+\frac{n_2}{2s_i}\left(\frac{1}{R}-\frac{1}{s_i}\right)^2\right]$$

the resulting optical theory is know as third-order optics.

b) Integralrechnung....|

Balla2021_Book_Integral... pdf

Obs: Riemann → 43 occurrences.

c) Introduction to Differential Equations|

Introduction_to_Differential_Equations,... pdf

→ Gaussian elimination.

d) Philomath....|

Philomath the Geometric Unification of Science ... epub.

Pag. 09 Figure 2

The golden section Φ and the π constant embedded within the dimensions of the gates of Egyptian temples.

'

! -256- - -* ! - '
)! ! I) .I& ! ? 7 ! 1"

D -256- - -* ! - '
& " I I4 ! I=> ! I "I=" 1"
E 7! ! !

-256- - -* ! - ' ! 9 7 H !
; 2 9 8 D 1 D
(/ < 9 7 " ! " !

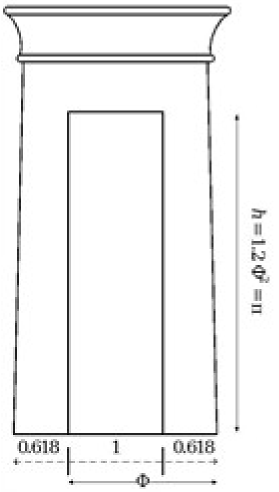

$h = 1.2\Phi^2 = \pi$
0.618
1
0.618
Φ

($/ !1 % 7 3 4 1!
(B <

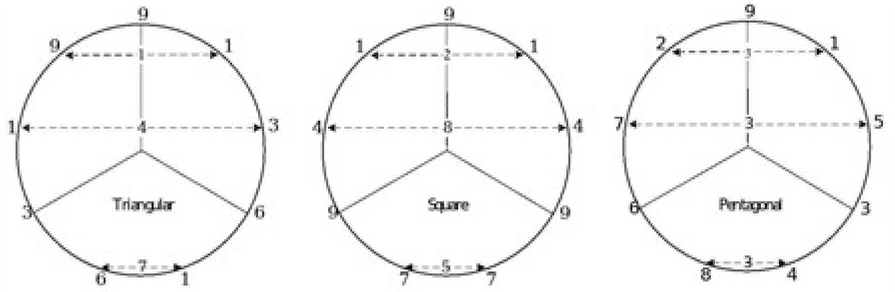

9
9
1
1
4
3
1
3
6
6
7
1
Triangular
9
1
2
1
4
8
4
9
9
7
5
7
Square
9
2
3
1
7
3
5
6
3
8
3
4
Pentagonal

($!1 ' 1 <! ,! 9

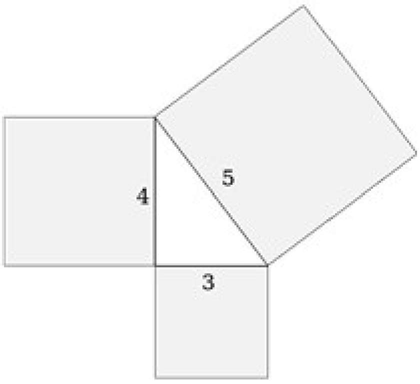

*D
< 9 ! ! 1 3 ! ! } ! 1 D ! !D !
! ! ! ! ! ! D " ! = "! !

(B A 2 !

" -256- - -* ! - ' ! ! 2 ! 3 1 L
! 1"
(1 L 4 A! A ! " A ! & !

(xi) *f is $\mathbb{C}$-differentiable in $S := [0,1] \times [0,1]$ and ε is positive. By an ε-*decomposition *of S is meant a decomposition into finitely many closed squares S_j whose sides are parallel to the coordinate axes and whose interiors are disjoint, such that each S_j contains a point z_j satisfying*

$$(6.j) \qquad |f(z) - [f(z_j) + f'(z_j)(z - z_j)]| \le \varepsilon |z - z_j| \qquad \forall z \in S_j.$$

Prove that S has such a decomposition (Goursat [1900]).

Hints: If not, then when S is divided into four congruent subsquares, at least one will have no ε-decomposition. Proceed inductively to get a decreasing sequence of closed squares S_n having side-length 2^{-n} and no ε-decomposition. The finite-intersection property of compacta supplies a point $z_0 \in \cap S_n$. As S_n has no ε-decomposition and $z_0 \in S_n$, there must be $z_n \in S_n \setminus \{z_0\}$ with

$$|f(z_n) - [f(z_0) + f'(z_0)\ (z_n - z_0)]| > \varepsilon |z_n - z_0|.$$

Division by $z_n - z_0$ then contradicts differentiability of f at z_0, since $z_n \to z_0$.

(xii) *With f and S_j, having side-length ℓ_j, as in* (xi), *show that*

$$(7.j) \qquad \left| \int_{\partial S_j} f \right| \le 4\varepsilon \ell_j \cdot \sup_{z \in S_j} |z - z_j| \le 8\varepsilon \ell_j^2.$$

Hints: Since the bracketed expression in (6.j) integrates to 0 [2.10(ii)], we have $\int_{\partial S_j} f = \int_{\partial S_j} (f(z) - [f(z_j) + f'(z_j)(z - z_j)]dz$, and because of inequality (6.j) and 2.9(ii), the latter integral is majorized as claimed.

(xiii) *Continuing with* (xii), *show that $\sum_j |\int_{\partial S_j} f| \le 8\varepsilon$.*

Hint: Since the S_j lie in S and have disjoint interiors, $\sum_j \ell_j^2 = \sum_j \text{area}(S_j) \le \text{area}(S)$.

(xiv) *For f, S as in* (xi), *does* (xiii) *justify the conclusion $\int_{\partial S} f = 0$?* (Cf. de la Vallée Poussin [1914].)

Hint: Although the details are a combinatorial headache, one can *prove* the (obvious?) equality $\int_{\partial S} f = \sum_j \int_{\partial S_j} f$. *Granted this*, it follows from (xiii) that $|\int_{\partial S} f| \le 8\varepsilon$, an inequality that holds for every $\varepsilon > 0$.

Remarks: As will be seen, $\int_{\partial R} f = 0$ (gotten as in (v) by Cauchy in 1846) leads to the remarkable fact that f is infinitely often differentiable and even (3.6) of power-series type in R. Cauchy suspected that the continuity hypothesis in (v) was unnecessary, but this was only confirmed in 1900, 43 years after his death. Modulo a (non-)trivial detail, the argument in 2.19(xiv) proves this. A combinatorially simpler proof, also featuring the method of descent, will be given in 5.1 (which the reader could easily turn to right now!). Neither of these proofs involves Green. But one can adapt these arguments to prove Green under the weaker hypothesis that $D_2f - iD_1f$ merely be Riemann double-integrable over R. Evidently (v) without the continuity hypothesis is a corollary of this improved Green's theorem. See Bochner [1955b], P. J. Cohen [1959], Craven [1964], Grunsky [1966], Acker [1996] and Greenlee [2004/05]. See also **MR** 2005e:30069.

The aforementioned power-series representation of a holomorphic function comes from expanding the kernel in a certain integral representation of the function (as in 3.8), which in turn is derived from Cauchy's Integral Theorem 2.19(v) by an ingenious idea that the next exercise sketches, but which gets a fuller treatment in chapter V. See also 3.39 and the remarks following it. It is noteworthy that by itself 2.19(v) is able to provide a brief, elegant proof of the fundamental theorem of algebra, essentially Gauss' third proof freed of reference to the transcendental functions sine and cosine (cf. also the remarks following 5.5).

Exercise 2.20 *Let $a, b, c, d \in \mathbb{R}$ with $a < b, c < d$, $R := (a, b) \times (c, d)$, $\overline{R} \subset U$ open $\subset \mathbb{C}$.*

(i) *Suppose $z_0 \in R$, $F : U \setminus \{z_0\} \mapsto \mathbb{C}$ is holomorphic, is bounded near z_0, and F' is continuous. Show that $\int_{\partial R} F = 0$.*

Hints: For $r > 0$ sufficiently small and $S_r := (-r, r) \times (-r, r)$, we have $z_0 + \overline{S}_r \subset R$. Extend the sides of the square $z_0 + S_r$ and thereby pave R as the union of this square and eight rectangles disjoint from it. To each of the latter 2.19(v) is applicable. When the integrals over the boundaries of these rectangles are written as sums of integrals over intervals and added, significant cancellation occurs, resulting in

$$0 = \int_{\partial R} F - \int_{z_0 + \partial S_r} F.$$

Since F is bounded near z_0 and $l(z_0 + \partial S_r) = 8r$, it follows from 2.9(i) that the second integral converges to 0 with r.

(ii) *Suppose f is holomorphic and f' is continuous in U. Show that*

$$f(z_0) \int_{\partial R} \frac{d\xi}{\xi - z_0} = \int_{\partial R} \frac{f(\xi)}{\xi - z_0} d\xi \qquad \forall z_0 \in R.$$

Hint: For each $z_0 \in R$ part (i) is applicable to the function defined by $F(z_0) := f'(z_0)$, $F(z) := (f(z) - f(z_0))/(z - z_0)$ for $z \in U \setminus \{z_0\}$. An alternative proof method is the Fubini argument employed in 3.39(ii).

(iii) *Using* 2.15(iii) *and the paving technique in the hints above, show that*

$$\int_{\partial R} \frac{d\xi}{\xi - z_0} = 8i \int_0^1 \frac{du}{u^2+1} \qquad \forall z_0 \in R.$$

(iv) *Combine* (ii) *and* (iii) *to prove the following weak version of* Cauchy's Integral Formula: *If f is holomorphic and f' is continuous in a neighborhood of $\overline{R}$, then*

$$f(z) = \frac{1}{2pi} \int_{\partial R} \frac{f(\xi)}{\xi - z} d\xi \qquad \forall z \in R,$$

where $p := 4\int_0^1 \frac{du}{u^2+1}$. Infer (2.14) *that f is $\mathbb{C}$-differentiable of all orders!*

(v) *Suppose P is a complex polynomial of degree n with no zeros in $\mathbb{C}$. Use the (obvious) fact that $\lim_{|z|\to\infty} zP'(z)/P(z) = n$ to deduce that*

$$\lim_{r\to\infty} \int_{\partial S_r} \frac{P'}{P} = 8n \int_0^1 \frac{1}{\gamma},$$

where $S_r := (-r, r) \times (-r, r)$ and $\gamma := [-1-i, 1-i]$.

Hints: Set $f := P'/P$. Note that $\gamma' = 2$ and $\operatorname{Im}\gamma = -1$. As in the hint to 2.15(iii), the intervals comprising the polygonal loop ∂S_r are $r\gamma$, $ir\gamma$, $-r\gamma$, $-ir\gamma$, so 2.9(ii) together with the fact $\gamma' = 2$ gives

$$\int_{\partial S_r} f = 2\sum_{k=0}^{3} \int_0^1 \frac{i^k r\gamma f \circ (i^k r\gamma)}{\gamma}.$$

Since $|\gamma| \ge |\operatorname{Im}\gamma| = 1$, we have $|r\gamma| \to \infty$ uniformly on $[0,1]$ as $r \uparrow \infty$. Consequently, $\lim_{|z|\to\infty} zf(z) = n$ implies that each integrand on the right converges to n/γ, uniformly on $[0,1]$ as $r \uparrow \infty$.

(vi) (FTA) *From* (v) *and Cauchy's theorem* 2.19(v) *infer that every nonconstant complex polynomial has a zero in* $\mathbb{C}$.

Hints: If P is a complex polynomial of degree n which has no zeros in $\mathbb{C}$, then $f := P'/P$ is holomorphic and f' is continuous in $\mathbb{C}$. Hence by 2.19(v) every integral on the left-hand side of the equality in part (v) above is 0. From that equality and the fact that $\operatorname{Im}\gamma = -1$ therefore follows

$$0 = \operatorname{Im}\left(8n \int_0^1 \frac{1}{\gamma}\right) = 8n \operatorname{Im} \int_0^1 \frac{\bar{\gamma}}{|\gamma|^2} = 8n \int_0^1 \frac{1}{|\gamma|^2},$$

not entirely trivial fact: see Minda [1973] (and 5.71 below). For more general results see Rogers and Zame [1982].

Though adumbrated by Gauss and Poisson, the idea of the complex contour integral in 2.8 was codified by Cauchy in his famous 1825 memoir. See Stäckel [1900], [1901], Ettlinger [1921–22], Yuškevič [1947] and chapter 4 of Smithies [1997]. Eighteenth-century mathematicians sometimes (but with suspicion) made an “imaginary substitution” in an integral (that is, a $\mathbb{C}$-valued change of variable, 11.33 and the generalized Fresnel integral in Remark (ii) following 11.32 being historically significant examples, and $x := it$ in $\int_{-\infty}^{\infty}(1+x^4)^{-1}dx$ being one reason for suspicion); this can be seen as foreshadowing the contour integral, and it was still the prevalent view in Cauchy’s early work. See pp. 10–23 of Smithies [1997].

Much work has gone into efforts to weaken the conditions in 2.17 and these conditions can indeed be considerably relaxed. The natural global converse of 2.16(i) is:

> *If f is continuous, D_1f, D_2f exist and satisfy* (C. R.) *throughout an open set U, then $f \in H(U)$.*

But it is not significantly easier to prove than the following:

Theorem (Looman [1923], Ridder [1930a], Wilkosz [1933], Menchoff [1936]) *If f is continuous on U, D_1f and D_2f each exist at all but countably many points of U and the equation* (C. R.) *holds almost everywhere in U (in the sense of plane Lebesgue measure), then f is holomorphic in U.*

In fact there is nothing exceptional about the horizontal and vertical directional derivatives here: Menchoff [1935] showed that if f is continuous on U and through each $z_0 \in U$ there pass two distinct lines $L_1(z_0)$, $L_2(z_0)$ such that

$$\lim_{\substack{z\to z_0\\ z\in L_1(z_0)}} \frac{f(z)-f(z_0)}{z-z_0} \quad \text{and} \quad \lim_{\substack{z\to z_0\\ z\in L_2(z_0)}} \frac{f(z)-f(z_0)}{z-z_0}$$

exist and coincide, then f is holomorphic in U. For some remarkable counterexamples when the continuity hypothesis of the theorem is absent, see Pelling [1979].

The requirement of differentiability at every point in the definition 2.1 of holomorphy can also be relaxed. 5.41, 5.42 and 9.28 are examples; here is another:

Theorem (Pompeiu [1905a], Looman [1922/1924], Ridder [1930a]) *If f is continuous in U, $\overline{\lim}_{h\to 0}\left|(f(z+h)-f(z))/h\right|$ is finite for all but countably*

-256- - -* ! - '
I ! !D I ! I=! 3I9 ! " ! 1"

-256- - -* ! - ' 4 A I2D ! I2 D ! 1"
(B $!1 5
9 ! 1! / ! ") ! 9
=L! 1

(2) The group of units of $\mathbb{Z}$ is the group $U_2 = \{\pm 1\}$ under multiplication.

(3) The group of units of the matrix ring $M_n(\mathbb{Z})$ is the group $GL_n(\mathbb{Z})$ (or $\mathbb{Q}, \mathbb{R}, \mathbb{C}$).

Since not all rings are commutative, many of the facts to which you're likely accustomed are not quite the same. The following technical theorem is designed to bring these differences to light.

Theorem 12.9. *Let R be a ring.*

(1) *Let $a, b \in R$. Then $(a+b)^2 = a^2 + 2ab + b^2$ if and only if $ab = ba$.*

(2) $\left(\sum_{i=1}^{n} a_i\right)\left(\sum_{j=1}^{m} b_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j$ *for all $a_i, b_j \in R$.*

(3) *If R is a commutative ring, then $(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^{n-i} b^i$ for all integers $n \geq 1$, where $a^n b^0 = a^n$ and $a^0 b^n = b^n$.*

In general, to compute $(a_1 + a_2 + \cdots + a_k)^n$ in an arbitrary ring, list every possible ordering of products of n elements chosen from the set $\{a_1, a_2, \ldots, a_k\}$ and add those k^n elements up. The last part of Theorem 12.9 is called the **Binomial Theorem**, and it simply provides a simpler result in the case where R is commutative.

Finally, we can define subrings in the same way that we defined subgroups, with a similar condition on recognizing them, but without any special notation. But beware: your knowledge about subgroups might lead your intuition astray when it comes to subrings, as you're about to discover.

Definition 12.10. Let R be a ring and S a subset of R. Then S is a **subring** of R if and only if S is a ring under the two operations of addition and multiplication induced by those operations on R.

Theorem 12.11. *Let R be a ring and S a subset of R. Then S is a subring of R if and only if S is an additive subgroup of R and S is closed under multiplication.*

Example 12.12. There are far fewer subrings for you to know than there are for subgroups.

(1) For any nonnegative integer n, the set $n\mathbb{Z}$ is a subring of $\mathbb{Z}$. When $n \neq 1$, the ring does not have unity.

(2) As with groups, $\mathbb{Z}$ is a subring of $\mathbb{Q}$, which is a subring of $\mathbb{R}$, which is a subring of $\mathbb{C}$.

(3) The Gaussian integers $\mathbb{Z}[i]$ is a subring of $\mathbb{C}$. It has four units: ± 1 and $\pm i$.

Exercise 12.13. Be careful not to assume that everything you know about subgroups is the same for subrings. Consider the ring $R = \mathbb{Z} \times \mathbb{Z}$ and the subset $S = \mathbb{Z} \times \{0\}$. Show that R has a unity element, that S is a subring of R, that the unity element of R is not in S, and that S is a subring *with unity* (in other words, show that S has a unity element different from the unity element of R)!

The reason for this shocking exercise is to focus your attention on the nature of equations dealing with multiplication. Specifically, in a ring with unity, the equation

$ax = a$ has a solution $x = \mathbf{1}$, but it need not be the *unique* solution. This kind of example happens all the time in linear algebra: pick a noninvertible $n \times n$ matrix A, and you can always find multiple solutions to the equation $AX = A$.

Example 12.14. Your instincts in solving equations like $ax = a$ in rings will likely involve canceling the common factor of a on both sides. In group theory, that's no problem, since every element has an inverse. But elements in rings routinely lack multiplicative inverses, which means there is no general cancellation law. So, for instance:

(1) In the ring $\mathbb{Z}_{10}$, the equation $2x = 2$ has two solutions: $x = 1$ and $x = 6$.

(2) In the ring $2\mathbb{Z}$, the equation $2x = 2$ has no solutions whatsoever.

(3) In the ring $\mathbb{Z}_8$, the equation $4x = 4$ has four solutions: $x = 1, 3, 5,$ and 7.

(4) In the ring $M_2(\mathbb{R})$, the equation $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} X = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ has infinitely many solutions; any 2×2 matrix whose column sums equal 1 will work.

So, when solving equations in rings, be very careful when dealing with multiplication. It is *not* a group operation!

Let's close this section with at least one fact that does have a parallel from group theory.

Theorem 12.15. *Let R be a ring and $\mathcal{C}$ be a nonempty collection of subrings of R. Then $\bigcap_{S \in \mathcal{C}} S$ is a subring of R.*

Definition 12.16. Let R be a ring and A be any subset of R. Then the **subring generated by** A is the intersection of all subrings of R containing A. We say that A is a **generating set** for a subring S if and only if the subring generated by A is the subring S.

Theorem 12.17. *Let R be a ring and $r \in R$. Then the subring generated by r is the set*

$$\{n_0 + n_1 r + n_2 r^2 + \cdots + n_k r^k \mid k, n_i \in \mathbb{Z}, k \geq 0\}.$$

12.2 Homomorphisms

Just as we did with groups, we need to develop a notion of maps between rings that preserve structure. Now that we have two operations to deal with, those maps need to preserve both structures.

Definition 12.18. Let R and R' be rings and $\phi : R \to R'$ be a function. Then ϕ is a **(ring) homomorphism** if $\phi(a + b) = \phi(a) + \phi(b)$ and $\phi(ab) = \phi(a)\phi(b)$ for all $a, b \in R$. A ring homomorphism ϕ is an **isomorphism** if and only if ϕ is also a bijection, and a ring **automorphism** is a ring isomorphism from a ring to itself. Two rings are **isomorphic** if and only if there is a ring isomorphism from one ring to the other.

Notice that since ring homomorphisms and isomorphisms are, in fact, group homomorphisms and group isomorphisms under addition, all of the results from group theory still apply to the additive structure of a ring. On the other hand, since the multiplicative structure of a ring isn't a group structure, we don't get all of the nice multiplicative properties we might hope for. However, if you have a ring *isomorphism*, then all of the algebraic properties like commutativity, unity, and inverses are preserved, as well as unity and inverses being mapped to unity and inverses. Yet even with ring homomorphisms, we do get a few nice properties similar to our results from groups.

Theorem 12.19. *Let $\phi : R \to R'$ be a ring homomorphism.*

(1) $\phi(a^n) = \phi(a)^n$ *for all* $a \in R$ *and* $n \in \mathbb{Z}^+$.

(2) *If S is a subring of R, then $\phi(S)$ is a subring of R', and if S is commutative, then $\phi(S)$ is also commutative.*

(3) *If S' is a subring of R', then $\phi^{-1}(S')$ is a subring of R.*

Definition 12.20. Let $\phi : R \to R'$ be a homomorphism, where R, R' are rings. Then the subring $\phi^{-1}(\mathbf{0}) = \{r \in R \mid \phi(r) = \mathbf{0}\}$ is called the **kernel** of ϕ, denoted $Ker(\phi)$.

Theorem 12.21. *Let $\phi : R \to R'$ be a ring homomorphism. Then ϕ is injective if and only if $Ker(\phi) = \{\mathbf{0}\}$.*

Exercise 12.22. Which, if any, of the following functions are ring homomorphisms? Of those that are (if any), what is the kernel of the homomorphism? Are any isomorphisms?

(1) $\phi : \mathbb{Z} \to \mathbb{Z}; \quad \phi(x) = 2x.$

(2) $\phi : \mathbb{Z} \to \mathbb{Z}_5; \quad \phi(x) = r_x$, where r_x is the remainder of x when divided by 5.

(3) $\phi : \mathbb{Z}_{mn} \to \mathbb{Z}_m \times \mathbb{Z}_n; \quad \phi(x) = (r_x, s_x)$, where r_x and s_x are the remainders of x when divided by m and n, respectively.

(4) $\phi : M_n(\mathbb{R}) \to \mathbb{R}; \quad \phi(A) = \det(A).$

There's one last theorem to prove in this section. It's not hard, but it's easy to misapply this theorem.

Theorem 12.23. *Let $\phi : R \to R'$ be a ring homomorphism. If R is a ring with unity element $\mathbf{1}$ and $\phi(\mathbf{1}) \neq \mathbf{0}$, then $\phi(\mathbf{1})$ is the unity element of the subring $\phi(R)$ of R'.*

Note that it does *not* say that the unity element of R must be mapped to the unity element of R' as it must for groups. It says that the image of the unity element must map to the unity element of the *range* of the homomorphism. In the event that ϕ is surjective, then unity gets mapped to unity. But in case you're still not seeing the difference, the following exercise should illuminate this issue.

Exercise 12.24. Let $\phi : \mathbb{Z}\times\mathbb{Z} \to \mathbb{Z}\times\mathbb{Z}$ be the function given by $\phi(a, b) = (a, 0)$. Verify that ϕ is a ring homomorphism, and that $\phi(1, 1) \neq (1, 1)$, but that $\phi(1, 1)$ is the unity element for the range of ϕ.

Exercise 12.25. Prove that there is exactly one ring automorphism of $\mathbb{Z}$, and that $n\mathbb{Z}$ and $m\mathbb{Z}$ are never isomorphic as rings when $n \neq \pm m$.

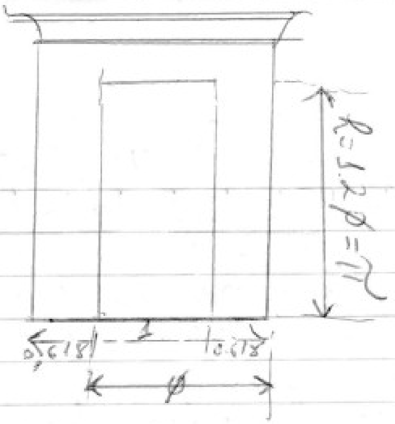

Obs: Analisar profundamente!!!

Obs: Gauss - 4 occurrences

Pag. 159 - Chapter 3 - Geometry in the D-Space
Pág. 160 - Figure 87

Pag. 175 - Chapter 4 - Semiprime Factorization: the Geometric Solutions.

Obs: Figure 100: the geometric aspects of Pythagoras' rule transforms the problem from an abstract mathematical formalism into a tangible two-dimensional Euclidean challenge).

Pág 176 - Preserving the Area.

e) Robert B. Burckel/
Classical Analysis in the Complex Plane.pdf

Obs. Gauss - 85 occurrences.
Estudar livro todo!!!

Pág. 80(107) (Complex) Derivative and (Curvilinear) Integrals)

...

elegant proof of the fundamental of algebra, essentially GAUSS third proof freed of reference to the transcendental functions sine and cosine (cf. also the remarks following 5.5).

Pág. 91(118) - though adumbrated by GAUSS and Poisson, the idea of the complex contour integral in 2.8 was codified by Cauchy in his famous 1825 memoir.

f) Single Variable Calculus... / Single_Variable_Calculus....pdf

Obs: Gauss - 11 occurrences.

g) Discovering Abstract Algebra / Discovering_Abstract_Algebra.pdf

Pág. 86(105) Chapter 12 Rings

The last part of theorem 12.9 is called the Binomial Theorem.

Example 12.12

(3) The Gaussian integers ($\mathbb{Z}$) is a subring of $\mathbb{C}$.
It has four units: ±1 and ±i.

Pág. 87(106) 12.2 Homomorphisms

Chapter 7

Chapter 7 - Part 1

Volume-09---20211218

Mathematics - Annotations 001

Prof. Stenio Musich

January 02, 2020

Chapter 1

Mathematics

1.1 Gaussian Quadrature

Source: DATA-NovoVivo-80-2020 Date: 20200509

1) Gaussian quadrature formulae on the unit circle
a) https://www.sciencedirect.com/science/article/pii/S0377042701004101
b) https://doi.org/10.1016/S0377-0427(01)00410-1
c) https://reader.elsevier.com/reader/sd/pii/S0377042701004101

2) C:/DATA NovoVivo 80 2020/Books 80 2020/Chemical Evolution of Galaxies/Nucleosynthesis and Chemical Evolution of Galaxies, Second Edition (2009) - libgen.lc.pdf

Pag. 41(61) 2.10 Weak Interactions

3) C:/DATA NovoVivo 80 2020/Books 80 2020/Fundamentals of Space Systems Ed 2/0195162056.pdf

Pag. 109(86) - Livro incompleto - 3.3 Two-Body Central Force Motion

4) C:/DATA NovoVivo 80 2020/Books 80 2020/Fundamentals of Space Systems Ed 2/Analytical Mechanics of Space Systems (2009, American Institute of Aeronautics and Astronautics) - libgen.lc.pdf

Pag. 31(54) - Newtonian Mechanics Pag. 143(166) - Eulerian Mechanics Pag. 437(460) - Part 2 - Celestinal Mechanics

Copy com errata:

C:/DATA NovoVivo 80 2020/Books 80 2020/Fundamentals of Space Systems Ed 2/Analytical Mechanics of Space Systems (2009, AIAA)+errata.pdf

Obs.: Livro todo!!!

5) C:/DATA NovoVivo 80 2020/Books 80 2020/Gravity's New Clothes The Origin of Gravity/Gravity s New Clothes The Origin of Gravity.pdf

6) C:/DATA NovoVivo 80 2020/Books 80 2020/Tangente Le Mag Hors serie No 74 Avril 2020/Tangente HS74 2020 04.pdf
Curbes et trajectoires - Algebre et geometrie - Dynamic et mouvement.

7) C:/DATA NovoVivo 80 2020/Books 80 2020/Courbes/Gauss/mws gen int txt gaussquadrature.pdf
Chapter 07.05 Gauss Quadrature Rule of Integration

8) C:/DATA NovoVivo 80 2020/Books 80 2020/Courbes/Gauss/2001.03090.pdf
Importance Gaussian Quadrature

9) C:/DATA NovoVivo 80 2020/Books 80 2020/Courbes/Gauss/chap4-2.pdf
Gauss Quadrature

10) C:/DATA NovoVivo 80 2020/Books 80 2020/Courbes/Gauss/387-393.pdf
Approximation Weights of Gauss Quadrature Method

11) C:/DATA NovoVivo 80 2020/Books 80 2020/Courbes/Gauss/Error Estimates for Gauss-Laguerre and Gauss-Hermite Quadrature Formulas.pdf
Error Estimates for Gauss Quadrature Formulas for Analytic Functions

12) C:/DATA NovoVivo 80 2020/Books 80 2020/Courbes/Gauss/gaussian quadrature.pdf
Gaussian Quadrature

13) C:/DATA NovoVivo 80 2020/Books 80 2020/Courbes/Gauss/Quadrature/1-s2.0-S0377042701004101-main.pdf
Gaussian quadrature formulae on the unit circle
Obs.: Texto todo!!!

14) https://www.terra.com.br/noticias/ciencia/ a-lei-universal-da- sica-descoberta-gracas-a-uma-gota-de-azeite-de-oliva, a185ec4c8f04a442509f4039bc18813cwp5jekyu.html

15) https://www.terra.com.br/noticias/ciencia/ a-danca-de-buracos-negros-que-comprova-teorias-de- einstein-e-stephen-hawking,

63d8444450564e7c1932e5b3a173292cp3tiwhig.html

1.1.1 Gaussian Quadrature - Research Material

Source: Books-80-2020/ IMPORTANT

1) C:/BOOKS 2020 13/Books 80 2020/Beyond Spacetime The Foundations of Quantum Gravity/Beyond Spacetime The Foundations of Quantum Gravity.pdf

Obs.: Este livro nos dará um visão geral sobre a gravidade e nós utilizaremos conjuntamente com a curvatura de Gauss. Obs.: Ler todo livro!!!

2)

1.2 Other - Equations

$$\frac{1}{x+y}$$

$$\frac{\frac{a}{x\ \ y}+\frac{b}{x+y}}{1+\frac{a\ \ b}{a+b}}$$

$$\overbrace{a+\underbrace{b+c}+d}$$

Teste das letras: $e+i=10$

Test do important:

IMPORTANT

$$a_0+\frac{1}{a_1+\frac{1}{a_2+\frac{1}{a_3+\frac{1}{a_4}}}}$$

$$\frac{\frac{a}{x \quad y} + \frac{b}{x+y}}{1+\frac{a \quad b}{a+b}}$$

$\boxed{\text{Gaussian curvature - this is a very important}}$

1.3 Spacetime

Source: DATA-NovoVivo-82-2020 Date: 20200510

1) C:/DATA-NovoVivo-82-2020/Books-82-2020/Foundations of Physics (Essentials of Physics)/Spacetime Foundations of General Relativity and Differential Geometry (2001, Springer) - libgen.lc.pdf

Obs.: 450 pages.

4. Pseudo-Riemannian manifolds................................ 171(179) 4.1 Existence of Lorentzian and Riemannian manifolds.....175(183)

5.3 Einstein's equation .... 268(276) 5.3.1 The Lagrangian formulation of Einstein's equation 271(279)

2) C:/DATA-NovoVivo-82-2020/Books-82-2020/Foundations of Physics (Essentials of Physics)/Quantum Field Theory on Curved Spacetimes Concepts and Mathematical Foundations (2009, Springer-Verlag Berlin Heide - libgen.lc.pdf

Index:

2 Lorentzian Manifolds 39 (47)
Frank Pfaffle
2.1 Preliminaries on Minkowski Space 39
2.2 Lorentzian Manifolds 40
2.3 Time-Orientation and Causality Relations 43
2.4 Causality Condition and Global Hyperbolicity 52
2.5 Cauchy Hypersurfaces 55
References 58

3)C:/DATA-NovoVivo-82-2020/Books-82-2020/Foundations of Physics (Essentials of Physics)/David Hilbert's Lectures on the Foundations of Physics 1915-1927 Relativity, Quantum Theory and Epistemology (2 - libgen.lc.pdf

Obs.: 808 pages.

4) C:/DATA-NovoVivo-82-2020/Books-82-2020/Foundations of Physics (Essentials of Physics)/Advanced Modern Physics Theoretical Foundations (2010, World Scientific Publishing Company) - libgen.lc.pdf

Hilbert Space - Pag. 10 until 23 Lagrangian Field Theory - Pag 96 until 123

5) C:/DATA-NovoVivo-82-2020/Books-82-2020/Foundations of Physics (Essentials of Physics)/Symmetry, Structure and Spacetime (2007, Elsevier) - libgen.lc.pdf
2. Space and Time in the Leibniz{Clarke Debate 23 2.1. Substantivalism versus relationalism 24 2.2. Inflation versus deflation 31 2.3. Leibniz versus Clarke 32 2.4. Sophisticated substantivalism and unsophisticated relationalism 35 2.5. Looking ahead to the modern debate 39
3. The Interpretation of Gauge Symmetries 45 3.1. Maxwellian electromagnetism 46 3.2. Aspects of gauge theories 50 3.3. Interpretive problems of gauge theories 58 3.4. Why gauge? 69

6) C:/DATA-NovoVivo-82-2020/Books-82-2020/Foundations of Physics (Essentials of Physics)/Quantum Field Theory I Foundations and Abelian and Non-Abelian Gauge Theories (2016, Springer International Publishing) - libgen.lc.pdf
C:/DATA-NovoVivo-82-2020/Books-82-2020/Foundations of Physics (Essentials of Physics)/Quantum Field Theory II Introductions to Quantum Gravity, Supersymmetry and String Theory (2016, Springer International Publishing) - libgen.lc.pdf

Chapter 2

Geometry

Gottfried Wilhelm (von) Leibniz (sometimes spelled Leibnitz) [11], French: Godefroi Guillaume Leibnitz;[15] 1 July 1646 [O.S. 21 June] { 14 November 1716) was a prominent German polymath and one of the most important logicians, mathematicians and natural philosophers of the Enlightenment. As a representative of the seventeenth-century tradition of rationalism, Leibniz's most prominent accomplishment was conceiving the ideas of differential and integral calculus, independently of Isaac Newton's contemporaneous developments.[16] Mathematical works have consistently favored Leibniz's notation as the conventional expression of calculus. It was only in the 20th century that Leibniz's law of continuity and transcendental law of homogeneity found mathematical implementation (by means of non-standard analysis). He became one of the most prolific inventors in the field of mechanical calculators. While working on adding automatic multiplication and division to Pascal's calculator, he was the first to describe a pinwheel calculator in 1685[17] and invented the Leibniz wheel, used in the arithmometer, the first mass-produced mechanical calculator. He also refined the binary number system, which is the foundation of all digital computers.

In philosophy, Leibniz is most noted for his optimism, i.e. his conclusion that our universe is, in a restricted sense, the best possible one that God could have created, an idea that was often lampooned by others such as Voltaire. Leibniz, along with Rene Descartes and Baruch Spinoza, was one of the three great 17th-century advocates of rationalism. The work of Leibniz anticipated modern logic and analytic philosophy, but his philosophy also assimilates elements of the scholastic tradition, notably that conclusions are produced by applying reason to first principles or prior definitions rather than to empirical evidence.

Leibniz made major contributions to physics and technology, and anticipated notions that surfaced much later in philosophy, probability theory, biology, medicine, geology, psychology, linguistics, and computer science. He wrote works on philosophy, politics, law, ethics, theology, history, and philology. Leibniz also contributed to the field of library science. While serving as overseer of the Wolfenbuttel library in Germany, he devised a cataloging system that would serve as a guide for many of Europe's largest libraries.[18] Leibniz's contributions to this vast array of subjects were scattered in various learned journals, in tens of thousands of letters, and in unpublished manuscripts. He wrote in several languages, but primarily in Latin, French, and German.[19] There is no complete gathering of the writings of Leibniz translated into English.[20]

Gottfried Leibniz was born on 1 July 1646, toward the end of the Thirty Years' War, in Leipzig, Saxony, to Friedrich Leibniz and Catharina Schmuck. Friedrich noted in his family journal:

21. Juny am Sontag 1646 Ist mein Sohn Gottfried Wilhelm, post sextam vespertinam 1/4 uff 7 uhr abents zur welt gebohren, im Wassermann.

In English:

On Sunday 21 June [NS: 1 July] 1646, my son Gottfried Wilhelm was born into the world a quarter before seven in the evening, in Aquarius.[21][22]

Leibniz was baptized on 3 July of that year at St. Nicholas Church, Leipzig; his godfather was the Lutheran theologian Martin Geier [de].[23] His father died when he was six years old, and from that point on he was raised by his mother.[24]

Leibniz's father had been a Professor of Moral Philosophy at the University of Leipzig, and the boy later inherited his father's personal library. He was given free access to it from the age of seven. While Leibniz's schoolwork was largely confined to the study of a small canon of authorities, his father's library enabled him to study a wide variety of advanced philosophical and theological works|ones that he would not have otherwise been able to read until his college years.[25] Access to his father's library, largely written in Latin, also led to his proficiency in the Latin language, which he achieved by the age of 12. He also composed 300 hexameters of Latin verse, in a single morning, for a special event at school at the age of 13.[26]

In April 1661 he enrolled in his father's former university at age 14,[27][1][28] and completed his bachelor's degree in Philosophy in December 1662. He defended his Disputatio Metaphysica de Principio Individui (Metaphysical Disputation on the Principle of Individuation),[29] which addressed the principle of individuation, on 9 June 1663. Leibniz earned his master's degree in Philosophy on 7 February 1664. He published and defended a dissertation Specimen Quaestionum Philosophicarum ex Jure collectarum (An Essay of Collected Philosophical Problems of Right),[29] arguing for both a theoretical and a pedagogical relationship between philosophy and law, in December

1664. After one year of legal studies, he was awarded his bachelor's degree in Law on 28 September 1665.[30] His dissertation was titled De conditionibus (On Conditions).[29]

In early 1666, at age 19, Leibniz wrote his first book, De Arte Combinatoria (On the Combinatorial Art), the first part of which was also his habilitation thesis in Philosophy, which he defended in March 1666.[29][31] His next goal was to earn his license and Doctorate in Law, which normally required three years of study. In 1666, the University of Leipzig turned down Leibniz's doctoral application and refused to grant him a Doctorate in Law, most likely due to his relative youth.[32][33] Leibniz subsequently left Leipzig.[34]

Leibniz then enrolled in the University of Altdorf and quickly submitted a thesis, which he had probably been working on earlier in Leipzig.[35] The title of his thesis was Disputatio Inauguralis de Casibus Perplexis in Jure (Inaugural Disputation on Ambiguous Legal Cases).[29] Leibniz earned his license to practice law and his Doctorate in Law in November 1666. He next declined the offer of an academic appointment at Altdorf, saying that "my thoughts were turned in an entirely different direction".[36]

As an adult, Leibniz often introduced himself as "Gottfried von Leibniz". Many posthumously published editions of his writings presented his name on the title page as "Freiherr G. W. von Leibniz." However, no document has ever been found from any contemporary government that stated his appointment to any form of nobility.[37]

2.1 Indice of Directories

Indice of Directores - Books 2020

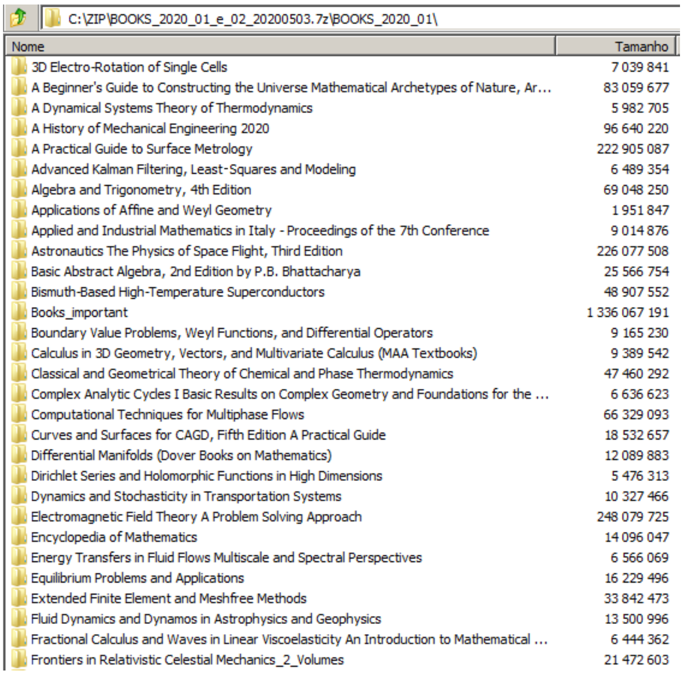

C:\ZIP\BOOKS_2020_01_e_02_20200503.7z\BOOKS_2020_01\

Nome	Tamanho
3D Electro-Rotation of Single Cells	7 039 841
A Beginner's Guide to Constructing the Universe Mathematical Archetypes of Nature, Ar...	83 059 677
A Dynamical Systems Theory of Thermodynamics	5 982 705
A History of Mechanical Engineering 2020	96 640 220
A Practical Guide to Surface Metrology	222 905 087
Advanced Kalman Filtering, Least-Squares and Modeling	6 489 354
Algebra and Trigonometry, 4th Edition	69 048 250
Applications of Affine and Weyl Geometry	1 951 847
Applied and Industrial Mathematics in Italy - Proceedings of the 7th Conference	9 014 876
Astronautics The Physics of Space Flight, Third Edition	226 077 508
Basic Abstract Algebra, 2nd Edition by P.B. Bhattacharya	25 566 754
Bismuth-Based High-Temperature Superconductors	48 907 552
Books_important	1 336 067 191
Boundary Value Problems, Weyl Functions, and Differential Operators	9 165 230
Calculus in 3D Geometry, Vectors, and Multivariate Calculus (MAA Textbooks)	9 389 542
Classical and Geometrical Theory of Chemical and Phase Thermodynamics	47 460 292
Complex Analytic Cycles I Basic Results on Complex Geometry and Foundations for the ...	6 636 623
Computational Techniques for Multiphase Flows	66 329 093
Curves and Surfaces for CAGD, Fifth Edition A Practical Guide	18 532 657
Differential Manifolds (Dover Books on Mathematics)	12 089 883
Dirichlet Series and Holomorphic Functions in High Dimensions	5 476 313
Dynamics and Stochasticity in Transportation Systems	10 327 466
Electromagnetic Field Theory A Problem Solving Approach	248 079 725
Encyclopedia of Mathematics	14 096 047
Energy Transfers in Fluid Flows Multiscale and Spectral Perspectives	6 566 069
Equilibrium Problems and Applications	16 229 496
Extended Finite Element and Meshfree Methods	33 842 473
Fluid Dynamics and Dynamos in Astrophysics and Geophysics	13 500 996
Fractional Calculus and Waves in Linear Viscoelasticity An Introduction to Mathematical ...	6 444 362
Frontiers in Relativistic Celestial Mechanics_2_Volumes	21 472 603

Figure 2.1: Books-01-2020-202005120906.png

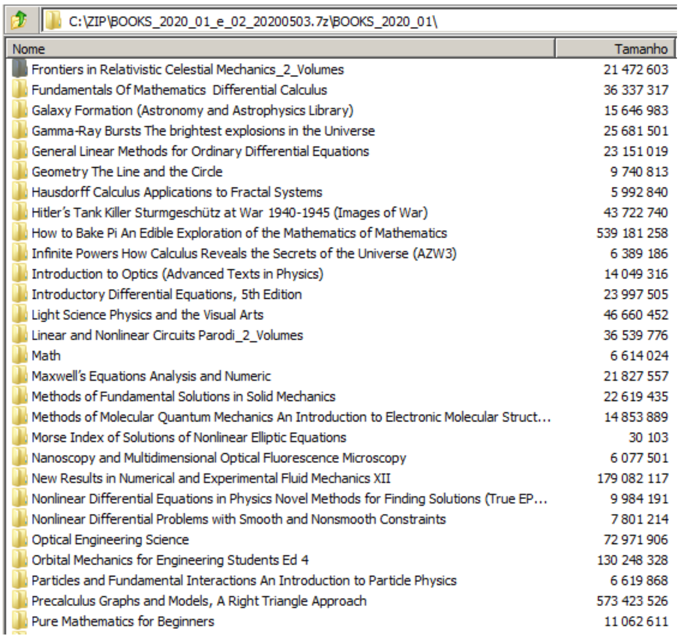

C:\ZIP\BOOKS_2020_01_e_02_20200503.7z\BOOKS_2020_01\

Nome	Tamanho
Frontiers in Relativistic Celestial Mechanics_2_Volumes	21 472 603
Fundamentals Of Mathematics Differential Calculus	36 337 317
Galaxy Formation (Astronomy and Astrophysics Library)	15 646 983
Gamma-Ray Bursts The brightest explosions in the Universe	25 681 501
General Linear Methods for Ordinary Differential Equations	23 151 019
Geometry The Line and the Circle	9 740 813
Hausdorff Calculus Applications to Fractal Systems	5 992 840
Hitler's Tank Killer Sturmgeschütz at War 1940-1945 (Images of War)	43 722 740
How to Bake Pi An Edible Exploration of the Mathematics of Mathematics	539 181 258
Infinite Powers How Calculus Reveals the Secrets of the Universe (AZW3)	6 389 186
Introduction to Optics (Advanced Texts in Physics)	14 049 316
Introductory Differential Equations, 5th Edition	23 997 505
Light Science Physics and the Visual Arts	46 660 452
Linear and Nonlinear Circuits Parodi_2_Volumes	36 539 776
Math	6 614 024
Maxwell's Equations Analysis and Numeric	21 827 557
Methods of Fundamental Solutions in Solid Mechanics	22 619 435
Methods of Molecular Quantum Mechanics An Introduction to Electronic Molecular Struct...	14 853 889
Morse Index of Solutions of Nonlinear Elliptic Equations	30 103
Nanoscopy and Multidimensional Optical Fluorescence Microscopy	6 077 501
New Results in Numerical and Experimental Fluid Mechanics XII	179 082 117
Nonlinear Differential Equations in Physics Novel Methods for Finding Solutions (True EP...	9 984 191
Nonlinear Differential Problems with Smooth and Nonsmooth Constraints	7 801 214
Optical Engineering Science	72 971 906
Orbital Mechanics for Engineering Students Ed 4	130 248 328
Particles and Fundamental Interactions An Introduction to Particle Physics	6 619 868
Precalculus Graphs and Models, A Right Triangle Approach	573 423 526
Pure Mathematics for Beginners	11 062 611

Figure 2.2: Books-01-2020-202005120908.png

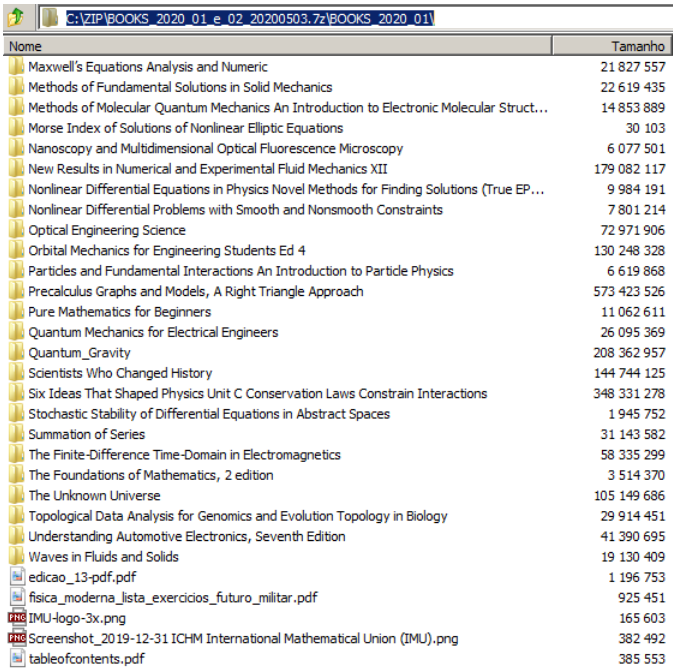

C:\ZIP\BOOKS_2020_01_e_02_20200503.7z\BOOKS_2020_01\

Nome	Tamanho
Maxwell's Equations Analysis and Numeric	21 827 557
Methods of Fundamental Solutions in Solid Mechanics	22 619 435
Methods of Molecular Quantum Mechanics An Introduction to Electronic Molecular Struct...	14 853 889
Morse Index of Solutions of Nonlinear Elliptic Equations	30 103
Nanoscopy and Multidimensional Optical Fluorescence Microscopy	6 077 501
New Results in Numerical and Experimental Fluid Mechanics XII	179 082 117
Nonlinear Differential Equations in Physics Novel Methods for Finding Solutions (True EP...	9 984 191
Nonlinear Differential Problems with Smooth and Nonsmooth Constraints	7 801 214
Optical Engineering Science	72 971 906
Orbital Mechanics for Engineering Students Ed 4	130 248 328
Particles and Fundamental Interactions An Introduction to Particle Physics	6 619 868
Precalculus Graphs and Models, A Right Triangle Approach	573 423 526
Pure Mathematics for Beginners	11 062 611
Quantum Mechanics for Electrical Engineers	26 095 369
Quantum_Gravity	208 362 957
Scientists Who Changed History	144 744 125
Six Ideas That Shaped Physics Unit C Conservation Laws Constrain Interactions	348 331 278
Stochastic Stability of Differential Equations in Abstract Spaces	1 945 752
Summation of Series	31 143 582
The Finite-Difference Time-Domain in Electromagnetics	58 335 299
The Foundations of Mathematics, 2 edition	3 514 370
The Unknown Universe	105 149 686
Topological Data Analysis for Genomics and Evolution Topology in Biology	29 914 451
Understanding Automotive Electronics, Seventh Edition	41 390 695
Waves in Fluids and Solids	19 130 409
edicao_13-pdf.pdf	1 196 753
fisica_moderna_lista_exercicios_futuro_militar.pdf	925 451
IMU-logo-3x.png	165 603
Screenshot_2019-12-31 ICHM International Mathematical Union (IMU).png	382 492
tableofcontents.pdf	385 553

Figure 2.3: Books-01-2020-202005120910.png

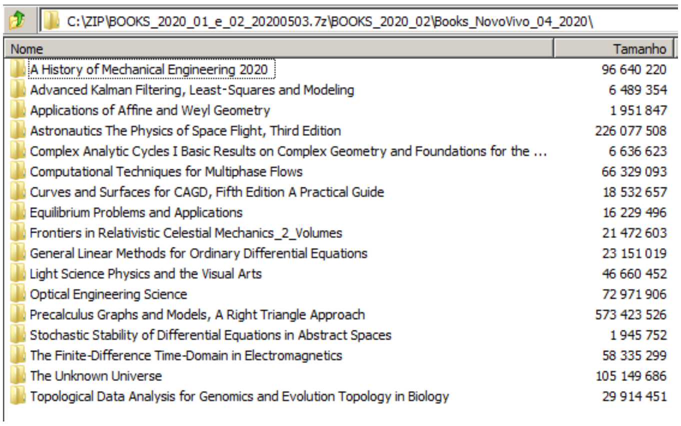

C:\ZIP\BOOKS_2020_01_e_02_20200503.7z\BOOKS_2020_02\Books_NovoVivo_04_2020\

Nome	Tamanho
A History of Mechanical Engineering 2020	96 640 220
Advanced Kalman Filtering, Least-Squares and Modeling	6 489 354
Applications of Affine and Weyl Geometry	1 951 847
Astronautics The Physics of Space Flight, Third Edition	226 077 508
Complex Analytic Cycles I Basic Results on Complex Geometry and Foundations for the ...	6 636 623
Computational Techniques for Multiphase Flows	66 329 093
Curves and Surfaces for CAGD, Fifth Edition A Practical Guide	18 532 657
Equilibrium Problems and Applications	16 229 496
Frontiers in Relativistic Celestial Mechanics_2_Volumes	21 472 603
General Linear Methods for Ordinary Differential Equations	23 151 019
Light Science Physics and the Visual Arts	46 660 452
Optical Engineering Science	72 971 906
Precalculus Graphs and Models, A Right Triangle Approach	573 423 526
Stochastic Stability of Differential Equations in Abstract Spaces	1 945 752
The Finite-Difference Time-Domain in Electromagnetics	58 335 299
The Unknown Universe	105 149 686
Topological Data Analysis for Genomics and Evolution Topology in Biology	29 914 451

Figure 2.4: Books-04-2020-202005120911.png

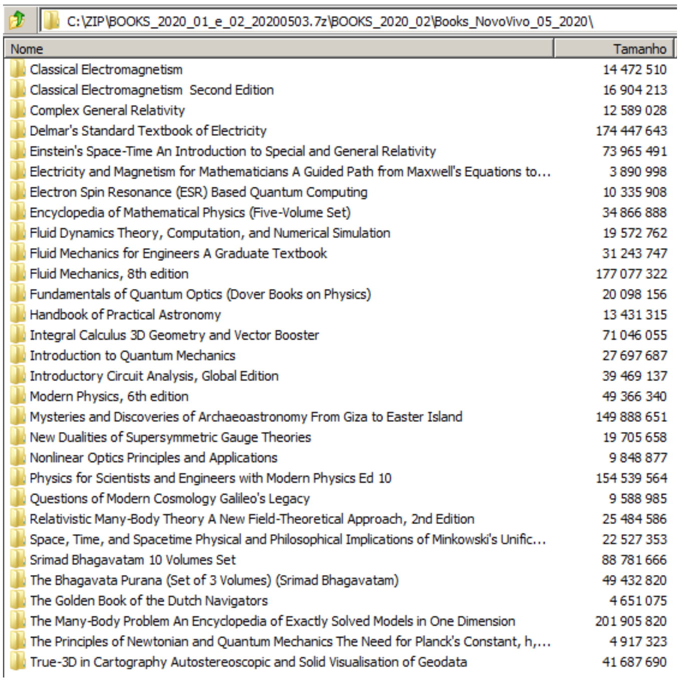

C:\ZIP\BOOKS_2020_01_e_02_20200503.7z\BOOKS_2020_02\Books_NovoVivo_05_2020\

Nome	Tamanho
Classical Electromagnetism	14 472 510
Classical Electromagnetism Second Edition	16 904 213
Complex General Relativity	12 589 028
Delmar's Standard Textbook of Electricity	174 447 643
Einstein's Space-Time An Introduction to Special and General Relativity	73 965 491
Electricity and Magnetism for Mathematicians A Guided Path from Maxwell's Equations to...	3 890 998
Electron Spin Resonance (ESR) Based Quantum Computing	10 335 908
Encyclopedia of Mathematical Physics (Five-Volume Set)	34 866 888
Fluid Dynamics Theory, Computation, and Numerical Simulation	19 572 762
Fluid Mechanics for Engineers A Graduate Textbook	31 243 747
Fluid Mechanics, 8th edition	177 077 322
Fundamentals of Quantum Optics (Dover Books on Physics)	20 098 156
Handbook of Practical Astronomy	13 431 315
Integral Calculus 3D Geometry and Vector Booster	71 046 055
Introduction to Quantum Mechanics	27 697 687
Introductory Circuit Analysis, Global Edition	39 469 137
Modern Physics, 6th edition	49 366 340
Mysteries and Discoveries of Archaeoastronomy From Giza to Easter Island	149 888 651
New Dualities of Supersymmetric Gauge Theories	19 705 658
Nonlinear Optics Principles and Applications	9 848 877
Physics for Scientists and Engineers with Modern Physics Ed 10	154 539 564
Questions of Modern Cosmology Galileo's Legacy	9 588 985
Relativistic Many-Body Theory A New Field-Theoretical Approach, 2nd Edition	25 484 586
Space, Time, and Spacetime Physical and Philosophical Implications of Minkowski's Unific...	22 527 353
Srimad Bhagavatam 10 Volumes Set	88 781 666
The Bhagavata Purana (Set of 3 Volumes) (Srimad Bhagavatam)	49 432 820
The Golden Book of the Dutch Navigators	4 651 075
The Many-Body Problem An Encyclopedia of Exactly Solved Models in One Dimension	201 905 820
The Principles of Newtonian and Quantum Mechanics The Need for Planck's Constant, h,...	4 917 323
True-3D in Cartography Autostereoscopic and Solid Visualisation of Geodata	41 687 690

Figure 2.5: Books-04-2020-202005120912.png

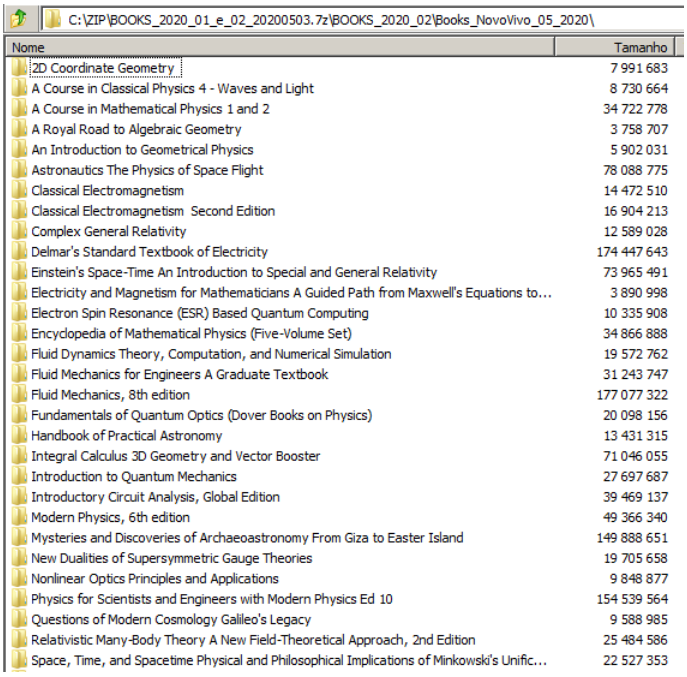

C:\ZIP\BOOKS_2020_01_e_02_20200503.7z\BOOKS_2020_02\Books_NovoVivo_05_2020\

Nome	Tamanho
2D Coordinate Geometry	7 991 683
A Course in Classical Physics 4 - Waves and Light	8 730 664
A Course in Mathematical Physics 1 and 2	34 722 778
A Royal Road to Algebraic Geometry	3 758 707
An Introduction to Geometrical Physics	5 902 031
Astronautics The Physics of Space Flight	78 088 775
Classical Electromagnetism	14 472 510
Classical Electromagnetism Second Edition	16 904 213
Complex General Relativity	12 589 028
Delmar's Standard Textbook of Electricity	174 447 643
Einstein's Space-Time An Introduction to Special and General Relativity	73 965 491
Electricity and Magnetism for Mathematicians A Guided Path from Maxwell's Equations to...	3 890 998
Electron Spin Resonance (ESR) Based Quantum Computing	10 335 908
Encyclopedia of Mathematical Physics (Five-Volume Set)	34 866 888
Fluid Dynamics Theory, Computation, and Numerical Simulation	19 572 762
Fluid Mechanics for Engineers A Graduate Textbook	31 243 747
Fluid Mechanics, 8th edition	177 077 322
Fundamentals of Quantum Optics (Dover Books on Physics)	20 098 156
Handbook of Practical Astronomy	13 431 315
Integral Calculus 3D Geometry and Vector Booster	71 046 055
Introduction to Quantum Mechanics	27 697 687
Introductory Circuit Analysis, Global Edition	39 469 137
Modern Physics, 6th edition	49 366 340
Mysteries and Discoveries of Archaeoastronomy From Giza to Easter Island	149 888 651
New Dualities of Supersymmetric Gauge Theories	19 705 658
Nonlinear Optics Principles and Applications	9 848 877
Physics for Scientists and Engineers with Modern Physics Ed 10	154 539 564
Questions of Modern Cosmology Galileo's Legacy	9 588 985
Relativistic Many-Body Theory A New Field-Theoretical Approach, 2nd Edition	25 484 586
Space, Time, and Spacetime Physical and Philosophical Implications of Minkowski's Unific...	22 527 353

Figure 2.6: Books-05-2020-202005120912.png

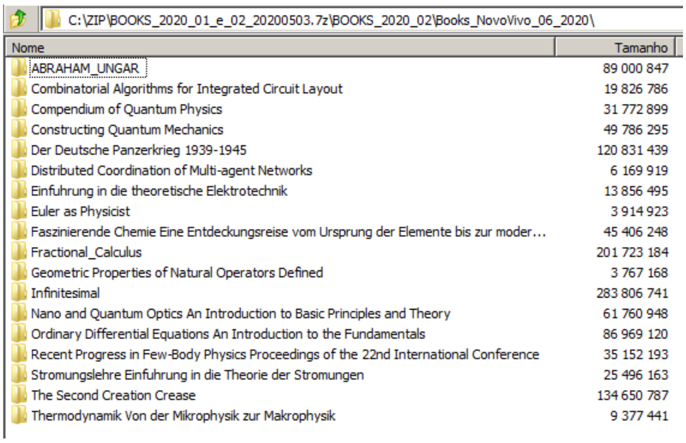

C:\ZIP\BOOKS_2020_01_e_02_20200503.7z\BOOKS_2020_02\Books_NovoVivo_06_2020\

Nome	Tamanho
ABRAHAM_UNGAR	89 000 847
Combinatorial Algorithms for Integrated Circuit Layout	19 826 786
Compendium of Quantum Physics	31 772 899
Constructing Quantum Mechanics	49 786 295
Der Deutsche Panzerkrieg 1939-1945	120 831 439
Distributed Coordination of Multi-agent Networks	6 169 919
Einfuhrung in die theoretische Elektrotechnik	13 856 495
Euler as Physicist	3 914 923
Faszinierende Chemie Eine Entdeckungsreise vom Ursprung der Elemente bis zur moder...	45 406 248
Fractional_Calculus	201 723 184
Geometric Properties of Natural Operators Defined	3 767 168
Infinitesimal	283 806 741
Nano and Quantum Optics An Introduction to Basic Principles and Theory	61 760 948
Ordinary Differential Equations An Introduction to the Fundamentals	86 969 120
Recent Progress in Few-Body Physics Proceedings of the 22nd International Conference	35 152 193
Stromungslehre Einfuhrung in die Theorie der Stromungen	25 496 163
The Second Creation Crease	134 650 787
Thermodynamik Von der Mikrophysik zur Makrophysik	9 377 441

Figure 2.7: Books-06-2020-202005120913.png

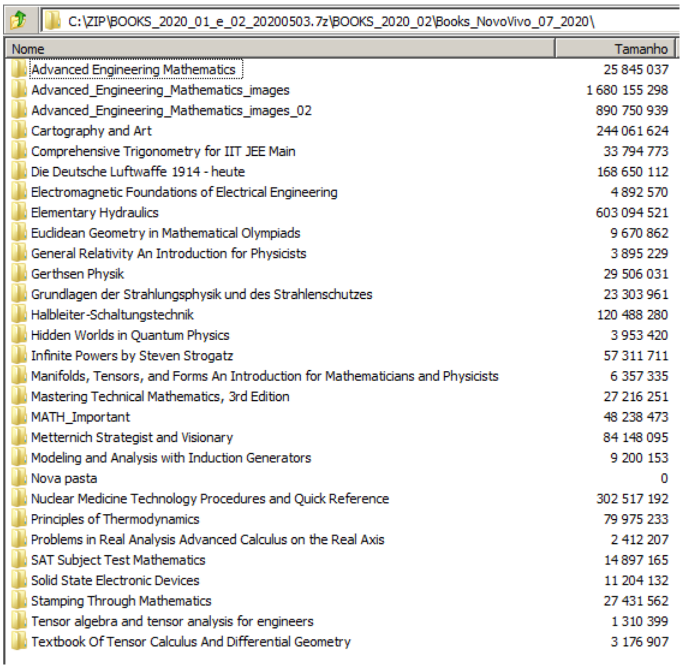

C:\ZIP\BOOKS_2020_01_e_02_20200503.7z\BOOKS_2020_02\Books_NovoVivo_07_2020\

Nome	Tamanho
Advanced Engineering Mathematics	25 845 037
Advanced_Engineering_Mathematics_images	1 680 155 298
Advanced_Engineering_Mathematics_images_02	890 750 939
Cartography and Art	244 061 624
Comprehensive Trigonometry for IIT JEE Main	33 794 773
Die Deutsche Luftwaffe 1914 - heute	168 650 112
Electromagnetic Foundations of Electrical Engineering	4 892 570
Elementary Hydraulics	603 094 521
Euclidean Geometry in Mathematical Olympiads	9 670 862
General Relativity An Introduction for Physicists	3 895 229
Gerthsen Physik	29 506 031
Grundlagen der Strahlungsphysik und des Strahlenschutzes	23 303 961
Halbleiter-Schaltungstechnik	120 488 280
Hidden Worlds in Quantum Physics	3 953 420
Infinite Powers by Steven Strogatz	57 311 711
Manifolds, Tensors, and Forms An Introduction for Mathematicians and Physicists	6 357 335
Mastering Technical Mathematics, 3rd Edition	27 216 251
MATH_Important	48 238 473
Metternich Strategist and Visionary	84 148 095
Modeling and Analysis with Induction Generators	9 200 153
Nova pasta	0
Nuclear Medicine Technology Procedures and Quick Reference	302 517 192
Principles of Thermodynamics	79 975 233
Problems in Real Analysis Advanced Calculus on the Real Axis	2 412 207
SAT Subject Test Mathematics	14 897 165
Solid State Electronic Devices	11 204 132
Stamping Through Mathematics	27 431 562
Tensor algebra and tensor analysis for engineers	1 310 399
Textbook Of Tensor Calculus And Differential Geometry	3 176 907

Figure 2.8: Books-07-2020-202005120914.png

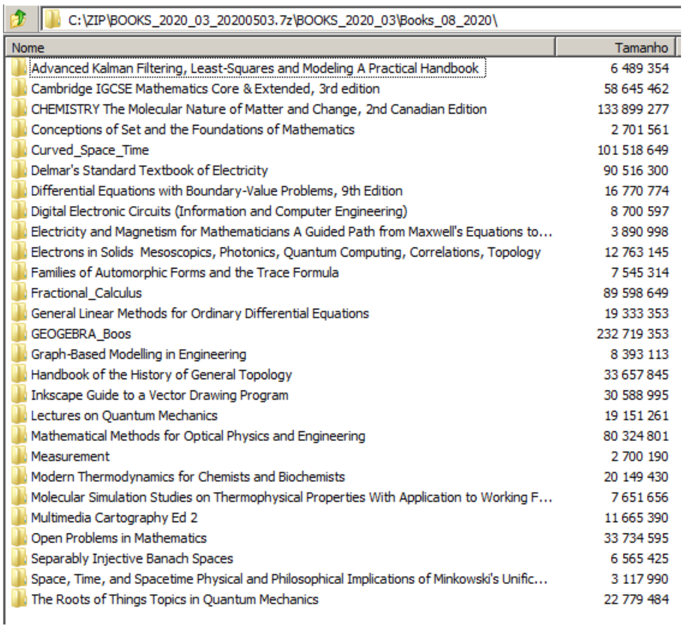

C:\ZIP\BOOKS_2020_03_20200503.7z\BOOKS_2020_03\Books_08_2020\

Nome	Tamanho
Advanced Kalman Filtering, Least-Squares and Modeling A Practical Handbook	6 489 354
Cambridge IGCSE Mathematics Core & Extended, 3rd edition	58 645 462
CHEMISTRY The Molecular Nature of Matter and Change, 2nd Canadian Edition	133 899 277
Conceptions of Set and the Foundations of Mathematics	2 701 561
Curved_Space_Time	101 518 649
Delmar's Standard Textbook of Electricity	90 516 300
Differential Equations with Boundary-Value Problems, 9th Edition	16 770 774
Digital Electronic Circuits (Information and Computer Engineering)	8 700 597
Electricity and Magnetism for Mathematicians A Guided Path from Maxwell's Equations to...	3 890 998
Electrons in Solids Mesoscopics, Photonics, Quantum Computing, Correlations, Topology	12 763 145
Families of Automorphic Forms and the Trace Formula	7 545 314
Fractional_Calculus	89 598 649
General Linear Methods for Ordinary Differential Equations	19 333 353
GEOGEBRA_Boos	232 719 353
Graph-Based Modelling in Engineering	8 393 113
Handbook of the History of General Topology	33 657 845
Inkscape Guide to a Vector Drawing Program	30 588 995
Lectures on Quantum Mechanics	19 151 261
Mathematical Methods for Optical Physics and Engineering	80 324 801
Measurement	2 700 190
Modern Thermodynamics for Chemists and Biochemists	20 149 430
Molecular Simulation Studies on Thermophysical Properties With Application to Working F...	7 651 656
Multimedia Cartography Ed 2	11 665 390
Open Problems in Mathematics	33 734 595
Separably Injective Banach Spaces	6 565 425
Space, Time, and Spacetime Physical and Philosophical Implications of Minkowski's Unific...	3 117 990
The Roots of Things Topics in Quantum Mechanics	22 779 484

Figure 2.9: Books-08-2020-202005120915.png

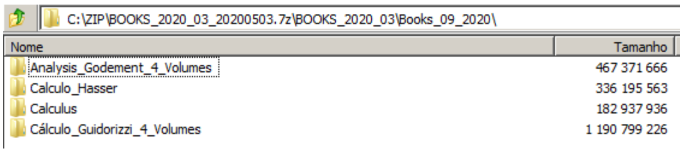

C:\ZIP\BOOKS_2020_03_20200503.7z\BOOKS_2020_03\Books_09_2020\

Nome	Tamanho
Analysis_Godement_4_Volumes	467 371 666
Calculo_Hasser	336 195 563
Calculus	182 937 936
Cálculo_Guidorizzi_4_Volumes	1 190 799 226

Figure 2.10: Books-09-2020-202005120916.png

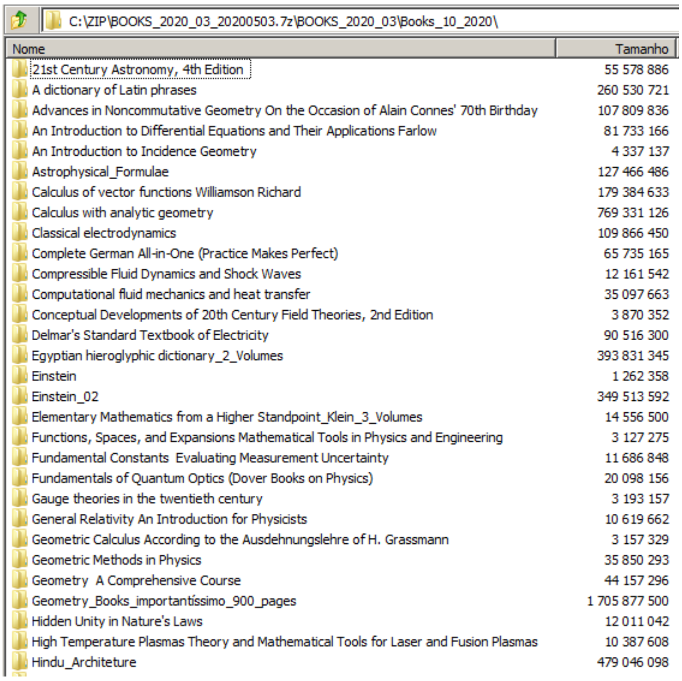

C:\ZIP\BOOKS_2020_03_20200503.7z\BOOKS_2020_03\Books_10_2020\

Nome	Tamanho
21st Century Astronomy, 4th Edition	55 578 886
A dictionary of Latin phrases	260 530 721
Advances in Noncommutative Geometry On the Occasion of Alain Connes' 70th Birthday	107 809 836
An Introduction to Differential Equations and Their Applications Farlow	81 733 166
An Introduction to Incidence Geometry	4 337 137
Astrophysical_Formulae	127 466 486
Calculus of vector functions Williamson Richard	179 384 633
Calculus with analytic geometry	769 331 126
Classical electrodynamics	109 866 450
Complete German All-in-One (Practice Makes Perfect)	65 735 165
Compressible Fluid Dynamics and Shock Waves	12 161 542
Computational fluid mechanics and heat transfer	35 097 663
Conceptual Developments of 20th Century Field Theories, 2nd Edition	3 870 352
Delmar's Standard Textbook of Electricity	90 516 300
Egyptian hieroglyphic dictionary_2_Volumes	393 831 345
Einstein	1 262 358
Einstein_02	349 513 592
Elementary Mathematics from a Higher Standpoint_Klein_3_Volumes	14 556 500
Functions, Spaces, and Expansions Mathematical Tools in Physics and Engineering	3 127 275
Fundamental Constants Evaluating Measurement Uncertainty	11 686 848
Fundamentals of Quantum Optics (Dover Books on Physics)	20 098 156
Gauge theories in the twentieth century	3 193 157
General Relativity An Introduction for Physicists	10 619 662
Geometric Calculus According to the Ausdehnungslehre of H. Grassmann	3 157 329
Geometric Methods in Physics	35 850 293
Geometry A Comprehensive Course	44 157 296
Geometry_Books_importantissimo_900_pages	1 705 877 500
Hidden Unity in Nature's Laws	12 011 042
High Temperature Plasmas Theory and Mathematical Tools for Laser and Fusion Plasmas	10 387 608
Hindu_Architeture	479 046 098

Figure 2.11: Books-10-2020-202005120917.png

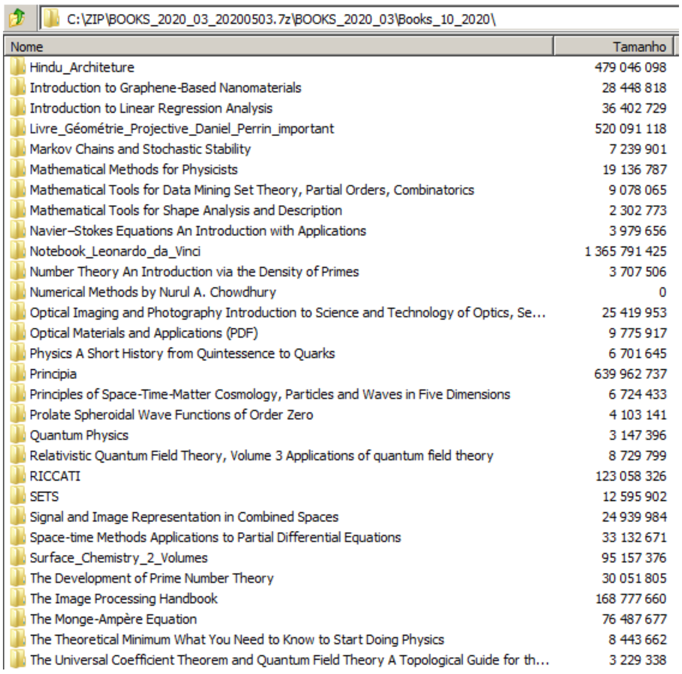

C:\ZIP\BOOKS_2020_03_20200503.7z\BOOKS_2020_03\Books_10_2020\

Nome	Tamanho
Hindu_Architeture	479 046 098
Introduction to Graphene-Based Nanomaterials	28 448 818
Introduction to Linear Regression Analysis	36 402 729
Livre_Géométrie_Projective_Daniel_Perrin_important	520 091 118
Markov Chains and Stochastic Stability	7 239 901
Mathematical Methods for Physicists	19 136 787
Mathematical Tools for Data Mining Set Theory, Partial Orders, Combinatorics	9 078 065
Mathematical Tools for Shape Analysis and Description	2 302 773
Navier–Stokes Equations An Introduction with Applications	3 979 656
Notebook_Leonardo_da_Vinci	1 365 791 425
Number Theory An Introduction via the Density of Primes	3 707 506
Numerical Methods by Nurul A. Chowdhury	0
Optical Imaging and Photography Introduction to Science and Technology of Optics, Se...	25 419 953
Optical Materials and Applications (PDF)	9 775 917
Physics A Short History from Quintessence to Quarks	6 701 645
Principia	639 962 737
Principles of Space-Time-Matter Cosmology, Particles and Waves in Five Dimensions	6 724 433
Prolate Spheroidal Wave Functions of Order Zero	4 103 141
Quantum Physics	3 147 396
Relativistic Quantum Field Theory, Volume 3 Applications of quantum field theory	8 729 799
RICCATI	123 058 326
SETS	12 595 902
Signal and Image Representation in Combined Spaces	24 939 984
Space-time Methods Applications to Partial Differential Equations	33 132 671
Surface_Chemistry_2_Volumes	95 157 376
The Development of Prime Number Theory	30 051 805
The Image Processing Handbook	168 777 660
The Monge-Ampère Equation	76 487 677
The Theoretical Minimum What You Need to Know to Start Doing Physics	8 443 662
The Universal Coefficient Theorem and Quantum Field Theory A Topological Guide for th...	3 229 338

Figure 2.12: Books-10-2020-202005120918.png

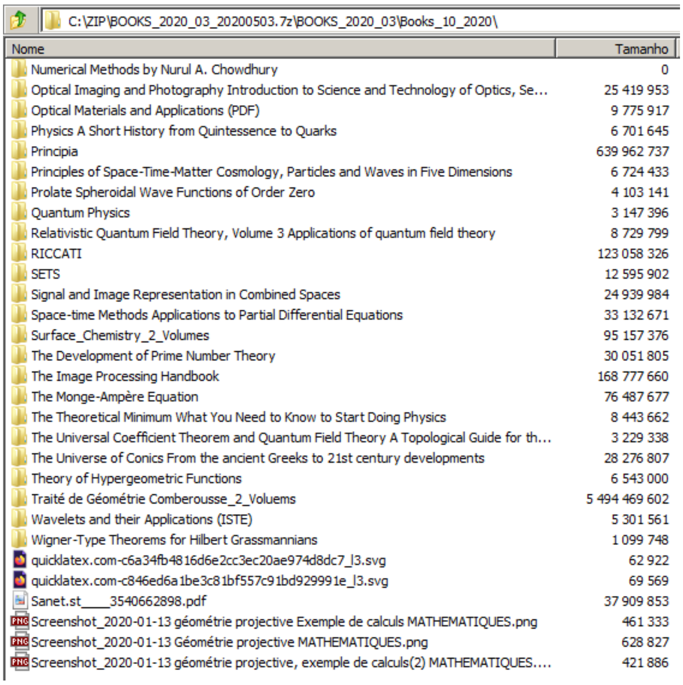

C:\ZIP\BOOKS_2020_03_20200503.7z\BOOKS_2020_03\Books_10_2020\

Nome	Tamanho
Numerical Methods by Nurul A. Chowdhury	0
Optical Imaging and Photography Introduction to Science and Technology of Optics, Se...	25 419 953
Optical Materials and Applications (PDF)	9 775 917
Physics A Short History from Quintessence to Quarks	6 701 645
Principia	639 962 737
Principles of Space-Time-Matter Cosmology, Particles and Waves in Five Dimensions	6 724 433
Prolate Spheroidal Wave Functions of Order Zero	4 103 141
Quantum Physics	3 147 396
Relativistic Quantum Field Theory, Volume 3 Applications of quantum field theory	8 729 799
RICCATI	123 058 326
SETS	12 595 902
Signal and Image Representation in Combined Spaces	24 939 984
Space-time Methods Applications to Partial Differential Equations	33 132 671
Surface_Chemistry_2_Volumes	95 157 376
The Development of Prime Number Theory	30 051 805
The Image Processing Handbook	168 777 660
The Monge-Ampère Equation	76 487 677
The Theoretical Minimum What You Need to Know to Start Doing Physics	8 443 662
The Universal Coefficient Theorem and Quantum Field Theory A Topological Guide for th...	3 229 338
The Universe of Conics From the ancient Greeks to 21st century developments	28 276 807
Theory of Hypergeometric Functions	6 543 000
Traité de Géométrie Comberousse_2_Voluems	5 494 469 602
Wavelets and their Applications (ISTE)	5 301 561
Wigner-Type Theorems for Hilbert Grassmannians	1 099 748
quicklatex.com-c6a34fb4816d6e2cc3ec20ae974d8dc7_l3.svg	62 922
quicklatex.com-c846ed6a1be3c81bf557c91bd929991e_l3.svg	69 569
Sanet.st____3540662898.pdf	37 909 853
Screenshot_2020-01-13 géométrie projective Exemple de calculs MATHEMATIQUES.png	461 333
Screenshot_2020-01-13 Géométrie projective MATHEMATIQUES.png	628 827
Screenshot_2020-01-13 géométrie projective, exemple de calculs(2) MATHEMATIQUES....	421 886

Figure 2.13: Books-10-2020-202005120919.png

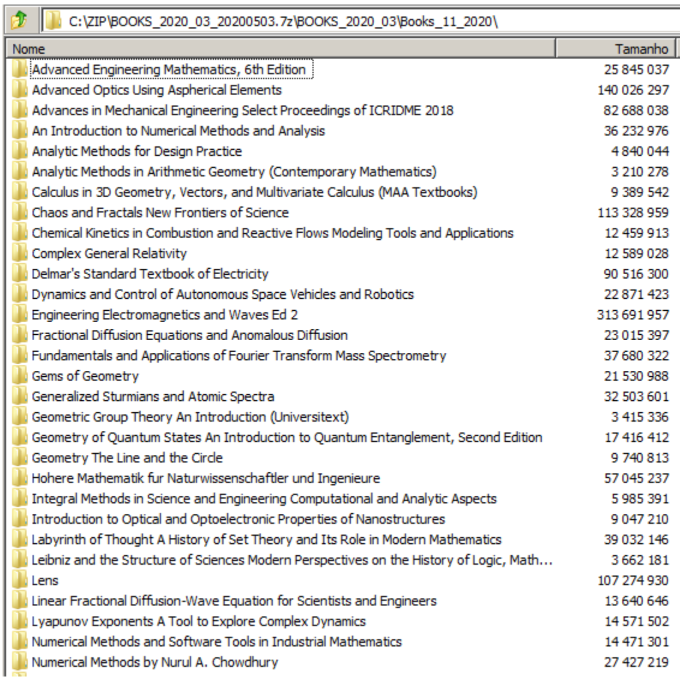

C:\ZIP\BOOKS_2020_03_20200503.7z\BOOKS_2020_03\Books_11_2020\

Nome	Tamanho
Advanced Engineering Mathematics, 6th Edition	25 845 037
Advanced Optics Using Aspherical Elements	140 026 297
Advances in Mechanical Engineering Select Proceedings of ICRIDME 2018	82 688 038
An Introduction to Numerical Methods and Analysis	36 232 976
Analytic Methods for Design Practice	4 840 044
Analytic Methods in Arithmetic Geometry (Contemporary Mathematics)	3 210 278
Calculus in 3D Geometry, Vectors, and Multivariate Calculus (MAA Textbooks)	9 389 542
Chaos and Fractals New Frontiers of Science	113 328 959
Chemical Kinetics in Combustion and Reactive Flows Modeling Tools and Applications	12 459 913
Complex General Relativity	12 589 028
Delmar's Standard Textbook of Electricity	90 516 300
Dynamics and Control of Autonomous Space Vehicles and Robotics	22 871 423
Engineering Electromagnetics and Waves Ed 2	313 691 957
Fractional Diffusion Equations and Anomalous Diffusion	23 015 397
Fundamentals and Applications of Fourier Transform Mass Spectrometry	37 680 322
Gems of Geometry	21 530 988
Generalized Sturmians and Atomic Spectra	32 503 601
Geometric Group Theory An Introduction (Universitext)	3 415 336
Geometry of Quantum States An Introduction to Quantum Entanglement, Second Edition	17 416 412
Geometry The Line and the Circle	9 740 813
Hohere Mathematik fur Naturwissenschaftler und Ingenieure	57 045 237
Integral Methods in Science and Engineering Computational and Analytic Aspects	5 985 391
Introduction to Optical and Optoelectronic Properties of Nanostructures	9 047 210
Labyrinth of Thought A History of Set Theory and Its Role in Modern Mathematics	39 032 146
Leibniz and the Structure of Sciences Modern Perspectives on the History of Logic, Math...	3 662 181
Lens	107 274 930
Linear Fractional Diffusion-Wave Equation for Scientists and Engineers	13 640 646
Lyapunov Exponents A Tool to Explore Complex Dynamics	14 571 502
Numerical Methods and Software Tools in Industrial Mathematics	14 471 301
Numerical Methods by Nurul A. Chowdhury	27 427 219

Figure 2.14: Books-11-2020-202005120926.png

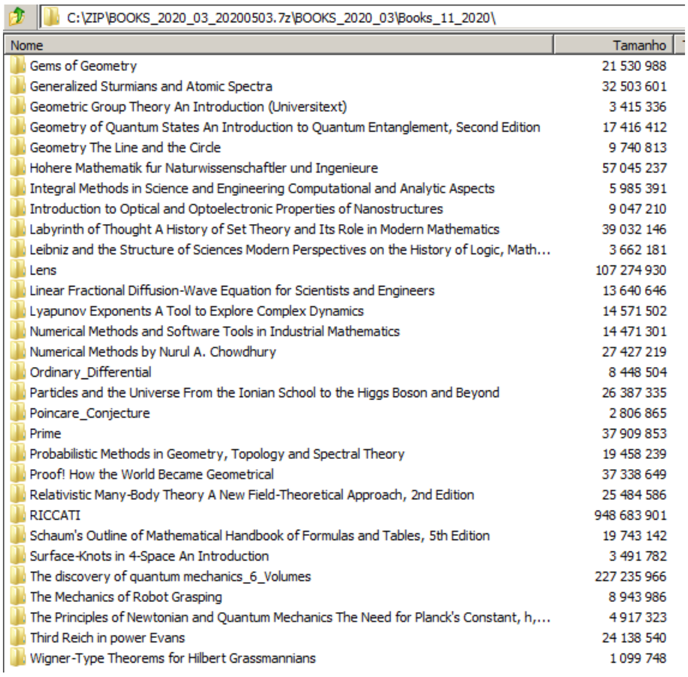

C:\ZIP\BOOKS_2020_03_20200503.7z\BOOKS_2020_03\Books_11_2020\

Nome	Tamanho
Gems of Geometry	21 530 988
Generalized Sturmians and Atomic Spectra	32 503 601
Geometric Group Theory An Introduction (Universitext)	3 415 336
Geometry of Quantum States An Introduction to Quantum Entanglement, Second Edition	17 416 412
Geometry The Line and the Circle	9 740 813
Hohere Mathematik fur Naturwissenschaftler und Ingenieure	57 045 237
Integral Methods in Science and Engineering Computational and Analytic Aspects	5 985 391
Introduction to Optical and Optoelectronic Properties of Nanostructures	9 047 210
Labyrinth of Thought A History of Set Theory and Its Role in Modern Mathematics	39 032 146
Leibniz and the Structure of Sciences Modern Perspectives on the History of Logic, Math...	3 662 181
Lens	107 274 930
Linear Fractional Diffusion-Wave Equation for Scientists and Engineers	13 640 646
Lyapunov Exponents A Tool to Explore Complex Dynamics	14 571 502
Numerical Methods and Software Tools in Industrial Mathematics	14 471 301
Numerical Methods by Nurul A. Chowdhury	27 427 219
Ordinary_Differential	8 448 504
Particles and the Universe From the Ionian School to the Higgs Boson and Beyond	26 387 335
Poincare_Conjecture	2 806 865
Prime	37 909 853
Probabilistic Methods in Geometry, Topology and Spectral Theory	19 458 239
Proof! How the World Became Geometrical	37 338 649
Relativistic Many-Body Theory A New Field-Theoretical Approach, 2nd Edition	25 484 586
RICCATI	948 683 901
Schaum's Outline of Mathematical Handbook of Formulas and Tables, 5th Edition	19 743 142
Surface-Knots in 4-Space An Introduction	3 491 782
The discovery of quantum mechanics_6_Volumes	227 235 966
The Mechanics of Robot Grasping	8 943 986
The Principles of Newtonian and Quantum Mechanics The Need for Planck's Constant, h,...	4 917 323
Third Reich in power Evans	24 138 540
Wigner-Type Theorems for Hilbert Grassmannians	1 099 748

Figure 2.15: Books-11-2020-202005120927.png

Books-15-2020 não existe (não foi copiado nada dentro dele);
Books-16-2020 não existe (não foi copiado nada dentro dele);

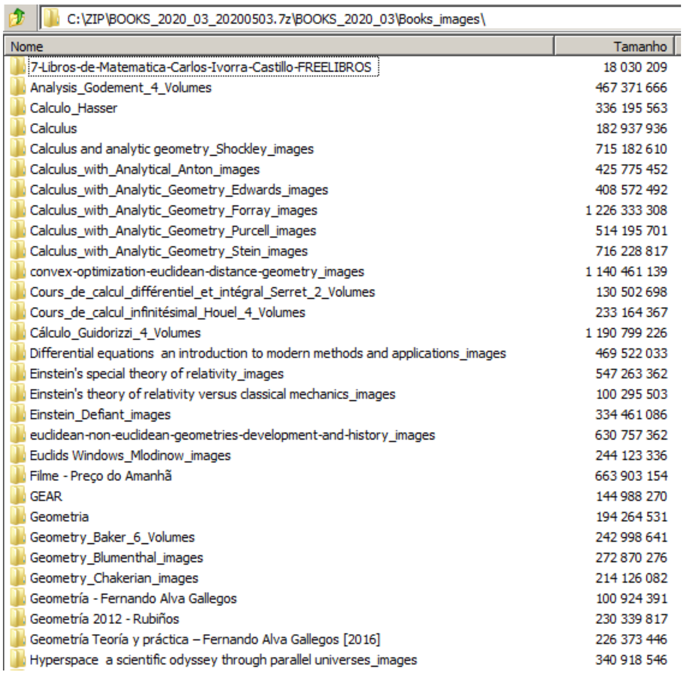
C:\ZIP\BOOKS_2020_03_20200503.7z\BOOKS_2020_03\Books_images\

Nome	Tamanho
7-Libros-de-Matematica-Carlos-Ivorra-Castillo-FREELIBROS	18 030 209
Analysis_Godement_4_Volumes	467 371 666
Calculo_Hasser	336 195 563
Calculus	182 937 936
Calculus and analytic geometry_Shockley_images	715 182 610
Calculus_with_Analytical_Anton_images	425 775 452
Calculus_with_Analytic_Geometry_Edwards_images	408 572 492
Calculus_with_Analytic_Geometry_Forray_images	1 226 333 308
Calculus_with_Analytic_Geometry_Purcell_images	514 195 701
Calculus_with_Analytic_Geometry_Stein_images	716 228 817
convex-optimization-euclidean-distance-geometry_images	1 140 461 139
Cours_de_calcul_différentiel_et_intégral_Serret_2_Volumes	130 502 698
Cours_de_calcul_infinitésimal_Houel_4_Volumes	233 164 367
Cálculo_Guidorizzi_4_Volumes	1 190 799 226
Differential equations an introduction to modern methods and applications_images	469 522 033
Einstein's special theory of relativity_images	547 263 362
Einstein's theory of relativity versus classical mechanics_images	100 295 503
Einstein_Defiant_images	334 461 086
euclidean-non-euclidean-geometries-development-and-history_images	630 757 362
Euclids Windows_Mlodinow_images	244 123 336
Filme - Preço do Amanhã	663 903 154
GEAR	144 988 270
Geometria	194 264 531
Geometry_Baker_6_Volumes	242 998 641
Geometry_Blumenthal_images	272 870 276
Geometry_Chakerian_images	214 126 082
Geometría - Fernando Alva Gallegos	100 924 391
Geometría 2012 - Rubiños	230 339 817
Geometría Teoría y práctica – Fernando Alva Gallegos [2016]	226 373 446
Hyperspace a scientific odyssey through parallel universes_images	340 918 546

Figure 2.16: Books-11-2020-202005120928.png

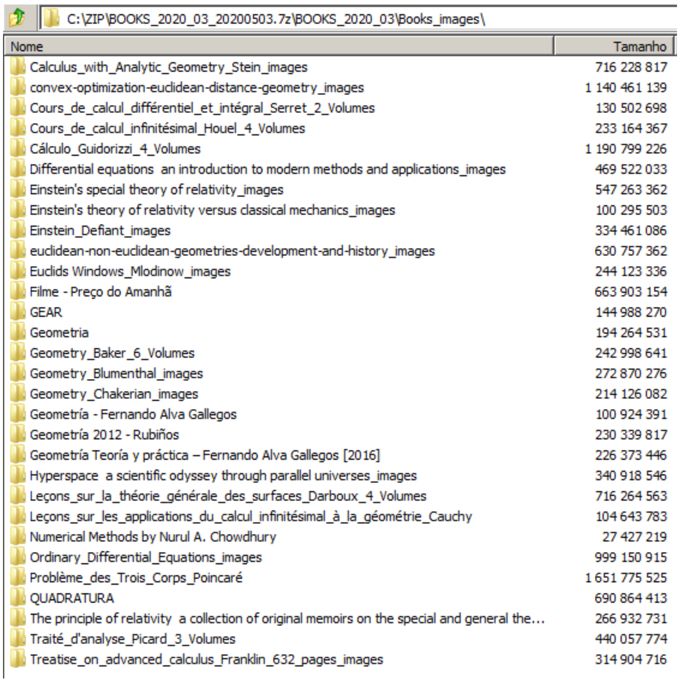
C:\ZIP\BOOKS_2020_03_20200503.7z\BOOKS_2020_03\Books_images\

Nome	Tamanho
Calculus_with_Analytic_Geometry_Stein_images	716 228 817
convex-optimization-euclidean-distance-geometry_images	1 140 461 139
Cours_de_calcul_différentiel_et_intégral_Serret_2_Volumes	130 502 698
Cours_de_calcul_infinitésimal_Houel_4_Volumes	233 164 367
Cálculo_Guidorizzi_4_Volumes	1 190 799 226
Differential equations an introduction to modern methods and applications_images	469 522 033
Einstein's special theory of relativity_images	547 263 362
Einstein's theory of relativity versus classical mechanics_images	100 295 503
Einstein_Defiant_images	334 461 086
euclidean-non-euclidean-geometries-development-and-history_images	630 757 362
Euclids Windows_Mlodinow_images	244 123 336
Filme - Preço do Amanhã	663 903 154
GEAR	144 988 270
Geometria	194 264 531
Geometry_Baker_6_Volumes	242 998 641
Geometry_Blumenthal_images	272 870 276
Geometry_Chakerian_images	214 126 082
Geometría - Fernando Alva Gallegos	100 924 391
Geometría 2012 - Rubiños	230 339 817
Geometría Teoría y práctica – Fernando Alva Gallegos [2016]	226 373 446
Hyperspace a scientific odyssey through parallel universes_images	340 918 546
Leçons_sur_la_théorie_générale_des_surfaces_Darboux_4_Volumes	716 264 563
Leçons_sur_les_applications_du_calcul_infinitésimal_à_la_géométrie_Cauchy	104 643 783
Numerical Methods by Nurul A. Chowdhury	27 427 219
Ordinary_Differential_Equations_images	999 150 915
Problème_des_Trois_Corps_Poincaré	1 651 775 525
QUADRATURA	690 864 413
The principle of relativity a collection of original memoirs on the special and general the...	266 932 731
Traité_d'analyse_Picard_3_Volumes	440 057 774
Treatise_on_advanced_calculus_Franklin_632_pages_images	314 904 716

Figure 2.17: Books-11-2020-202005120929.png

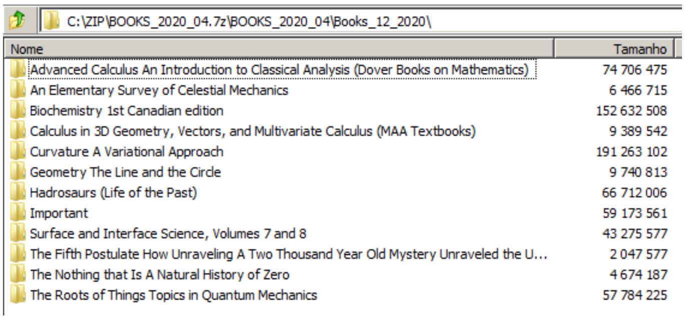

C:\ZIP\BOOKS_2020_04.7z\BOOKS_2020_04\Books_12_2020\

Nome	Tamanho
Advanced Calculus An Introduction to Classical Analysis (Dover Books on Mathematics)	74 706 475
An Elementary Survey of Celestial Mechanics	6 466 715
Biochemistry 1st Canadian edition	152 632 508
Calculus in 3D Geometry, Vectors, and Multivariate Calculus (MAA Textbooks)	9 389 542
Curvature A Variational Approach	191 263 102
Geometry The Line and the Circle	9 740 813
Hadrosaurs (Life of the Past)	66 712 006
Important	59 173 561
Surface and Interface Science, Volumes 7 and 8	43 275 577
The Fifth Postulate How Unraveling A Two Thousand Year Old Mystery Unraveled the U...	2 047 577
The Nothing that Is A Natural History of Zero	4 674 187
The Roots of Things Topics in Quantum Mechanics	57 784 225

Figure 2.18: Books-12-2020-202005121138.png

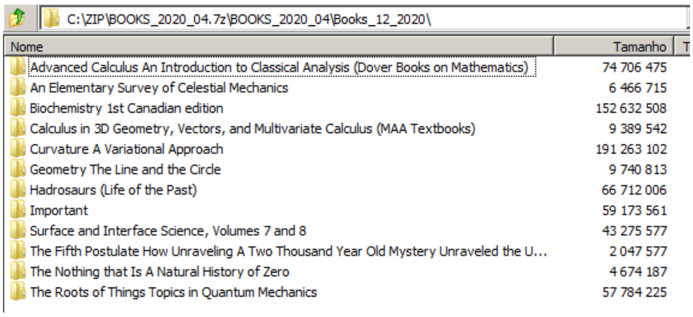

C:\ZIP\BOOKS_2020_04.7z\BOOKS_2020_04\Books_12_2020\

Nome	Tamanho	T
Advanced Calculus An Introduction to Classical Analysis (Dover Books on Mathematics)	74 706 475	
An Elementary Survey of Celestial Mechanics	6 466 715	
Biochemistry 1st Canadian edition	152 632 508	
Calculus in 3D Geometry, Vectors, and Multivariate Calculus (MAA Textbooks)	9 389 542	
Curvature A Variational Approach	191 263 102	
Geometry The Line and the Circle	9 740 813	
Hadrosaurs (Life of the Past)	66 712 006	
Important	59 173 561	
Surface and Interface Science, Volumes 7 and 8	43 275 577	
The Fifth Postulate How Unraveling A Two Thousand Year Old Mystery Unraveled the U...	2 047 577	
The Nothing that Is A Natural History of Zero	4 674 187	
The Roots of Things Topics in Quantum Mechanics	57 784 225	

Figure 2.19: Books-12-2020-202005121140.png

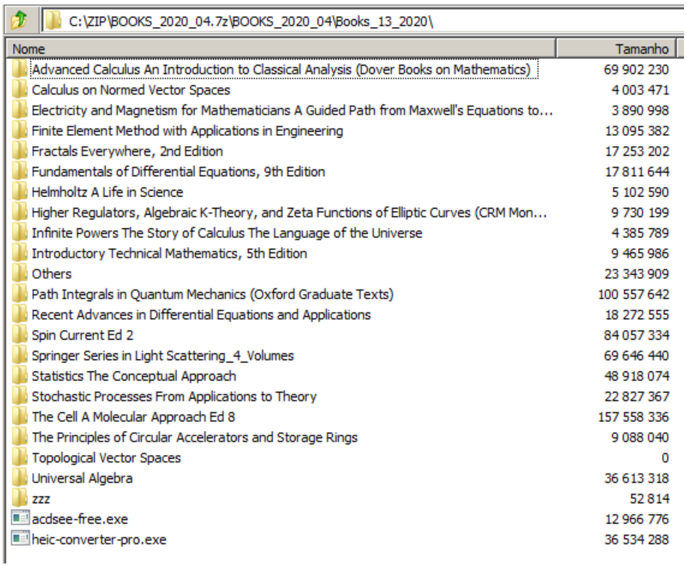

C:\ZIP\BOOKS_2020_04.7z\BOOKS_2020_04\Books_13_2020\

Nome	Tamanho
Advanced Calculus An Introduction to Classical Analysis (Dover Books on Mathematics)	69 902 230
Calculus on Normed Vector Spaces	4 003 471
Electricity and Magnetism for Mathematicians A Guided Path from Maxwell's Equations to...	3 890 998
Finite Element Method with Applications in Engineering	13 095 382
Fractals Everywhere, 2nd Edition	17 253 202
Fundamentals of Differential Equations, 9th Edition	17 811 644
Helmholtz A Life in Science	5 102 590
Higher Regulators, Algebraic K-Theory, and Zeta Functions of Elliptic Curves (CRM Mon...	9 730 199
Infinite Powers The Story of Calculus The Language of the Universe	4 385 789
Introductory Technical Mathematics, 5th Edition	9 465 986
Others	23 343 909
Path Integrals in Quantum Mechanics (Oxford Graduate Texts)	100 557 642
Recent Advances in Differential Equations and Applications	18 272 555
Spin Current Ed 2	84 057 334
Springer Series in Light Scattering_4_Volumes	69 646 440
Statistics The Conceptual Approach	48 918 074
Stochastic Processes From Applications to Theory	22 827 367
The Cell A Molecular Approach Ed 8	157 558 336
The Principles of Circular Accelerators and Storage Rings	9 088 040
Topological Vector Spaces	0
Universal Algebra	36 613 318
zzz	52 814
acdsee-free.exe	12 966 776
heic-converter-pro.exe	36 534 288

Figure 2.20: Books-13-2020-202005121139.png

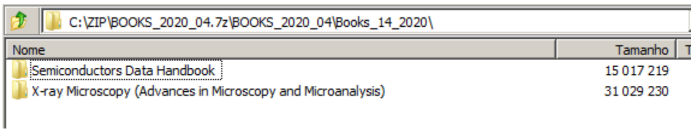

C:\ZIP\BOOKS_2020_04.7z\BOOKS_2020_04\Books_14_2020\

Nome	Tamanho
Semiconductors Data Handbook	15 017 219
X-ray Microscopy (Advances in Microscopy and Microanalysis)	31 029 230

Figure 2.21: Books-14-2020-202005121140.png

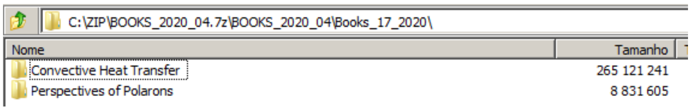

C:\ZIP\BOOKS_2020_04.7z\BOOKS_2020_04\Books_17_2020\

Nome	Tamanho
Convective Heat Transfer	265 121 241
Perspectives of Polarons	8 831 605

Figure 2.22: Books-17-2020-202005121141.png

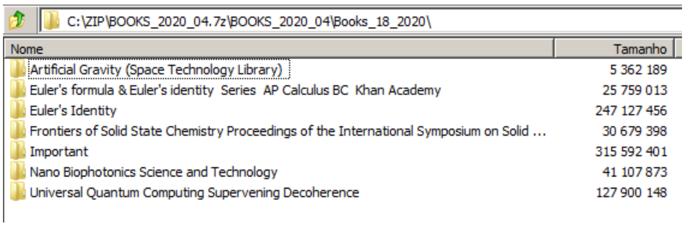

C:\ZIP\BOOKS_2020_04.7z\BOOKS_2020_04\Books_18_2020\

Nome	Tamanho
Artificial Gravity (Space Technology Library)	5 362 189
Euler's formula & Euler's identity Series AP Calculus BC Khan Academy	25 759 013
Euler's Identity	247 127 456
Frontiers of Solid State Chemistry Proceedings of the International Symposium on Solid ...	30 679 398
Important	315 592 401
Nano Biophotonics Science and Technology	41 107 873
Universal Quantum Computing Supervening Decoherence	127 900 148

Figure 2.23: Books-18-2020-202005121142.png

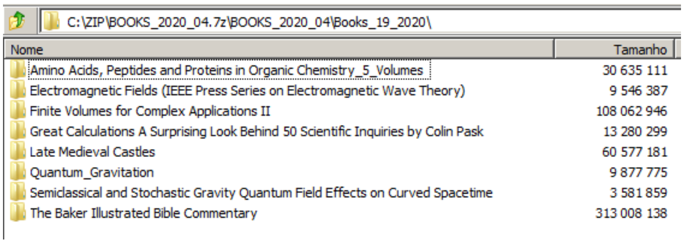

C:\ZIP\BOOKS_2020_04.7z\BOOKS_2020_04\Books_19_2020\

Nome	Tamanho
Amino Acids, Peptides and Proteins in Organic Chemistry_5_Volumes	30 635 111
Electromagnetic Fields (IEEE Press Series on Electromagnetic Wave Theory)	9 546 387
Finite Volumes for Complex Applications II	108 062 946
Great Calculations A Surprising Look Behind 50 Scientific Inquiries by Colin Pask	13 280 299
Late Medieval Castles	60 577 181
Quantum_Gravitation	9 877 775
Semiclassical and Stochastic Gravity Quantum Field Effects on Curved Spacetime	3 581 859
The Baker Illustrated Bible Commentary	313 008 138

Figure 2.24: Books-19-2020-202005121143.png

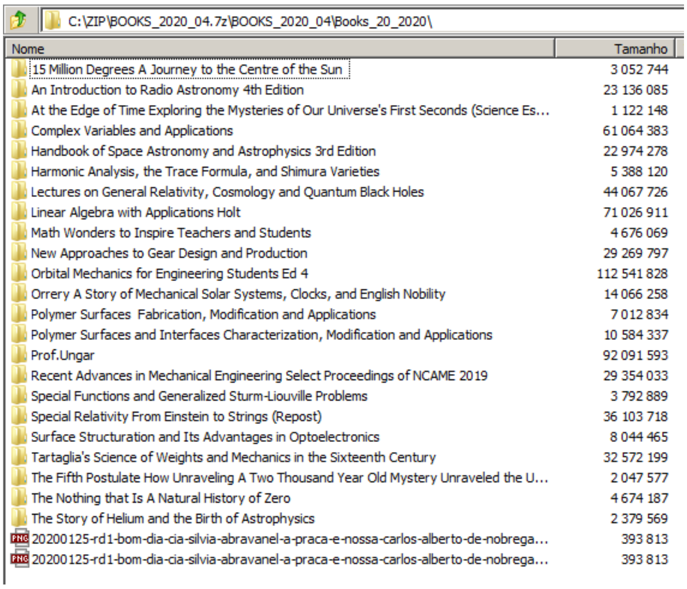

C:\ZIP\BOOKS_2020_04.7z\BOOKS_2020_04\Books_20_2020\

Nome	Tamanho
15 Million Degrees A Journey to the Centre of the Sun	3 052 744
An Introduction to Radio Astronomy 4th Edition	23 136 085
At the Edge of Time Exploring the Mysteries of Our Universe's First Seconds (Science Es...	1 122 148
Complex Variables and Applications	61 064 383
Handbook of Space Astronomy and Astrophysics 3rd Edition	22 974 278
Harmonic Analysis, the Trace Formula, and Shimura Varieties	5 388 120
Lectures on General Relativity, Cosmology and Quantum Black Holes	44 067 726
Linear Algebra with Applications Holt	71 026 911
Math Wonders to Inspire Teachers and Students	4 676 069
New Approaches to Gear Design and Production	29 269 797
Orbital Mechanics for Engineering Students Ed 4	112 541 828
Orrery A Story of Mechanical Solar Systems, Clocks, and English Nobility	14 066 258
Polymer Surfaces Fabrication, Modification and Applications	7 012 834
Polymer Surfaces and Interfaces Characterization, Modification and Applications	10 584 337
Prof.Ungar	92 091 593
Recent Advances in Mechanical Engineering Select Proceedings of NCAME 2019	29 354 033
Special Functions and Generalized Sturm-Liouville Problems	3 792 889
Special Relativity From Einstein to Strings (Repost)	36 103 718
Surface Structuration and Its Advantages in Optoelectronics	8 044 465
Tartaglia's Science of Weights and Mechanics in the Sixteenth Century	32 572 199
The Fifth Postulate How Unraveling A Two Thousand Year Old Mystery Unraveled the U...	2 047 577
The Nothing that Is A Natural History of Zero	4 674 187
The Story of Helium and the Birth of Astrophysics	2 379 569
20200125-rd1-bom-dia-cia-silvia-abravanel-a-praca-e-nossa-carlos-alberto-de-nobrega...	393 813
20200125-rd1-bom-dia-cia-silvia-abravanel-a-praca-e-nossa-carlos-alberto-de-nobrega...	393 813

Figure 2.25: Books-20-2020-202005121144.png

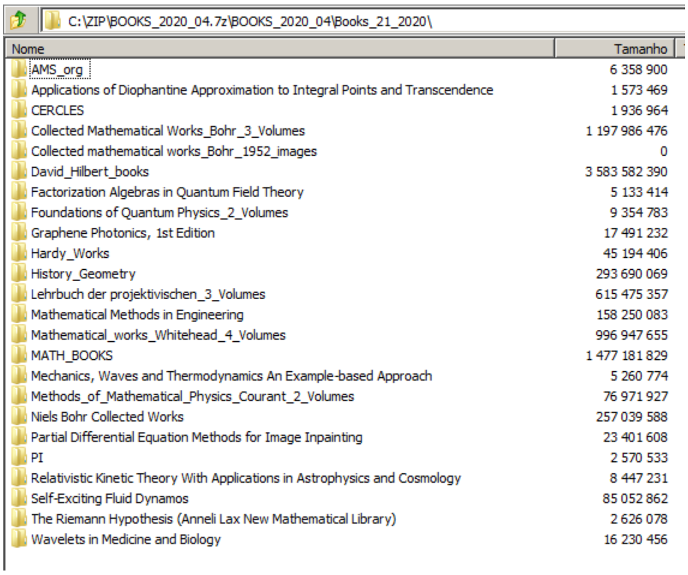

C:\ZIP\BOOKS_2020_04.7z\BOOKS_2020_04\Books_21_2020\

Nome	Tamanho
AMS_org	6 358 900
Applications of Diophantine Approximation to Integral Points and Transcendence	1 573 469
CERCLES	1 936 964
Collected Mathematical Works_Bohr_3_Volumes	1 197 986 476
Collected mathematical works_Bohr_1952_images	0
David_Hilbert_books	3 583 582 390
Factorization Algebras in Quantum Field Theory	5 133 414
Foundations of Quantum Physics_2_Volumes	9 354 783
Graphene Photonics, 1st Edition	17 491 232
Hardy_Works	45 194 406
History_Geometry	293 690 069
Lehrbuch der projektivischen_3_Volumes	615 475 357
Mathematical Methods in Engineering	158 250 083
Mathematical_works_Whitehead_4_Volumes	996 947 655
MATH_BOOKS	1 477 181 829
Mechanics, Waves and Thermodynamics An Example-based Approach	5 260 774
Methods_of_Mathematical_Physics_Courant_2_Volumes	76 971 927
Niels Bohr Collected Works	257 039 588
Partial Differential Equation Methods for Image Inpainting	23 401 608
PI	2 570 533
Relativistic Kinetic Theory With Applications in Astrophysics and Cosmology	8 447 231
Self-Exciting Fluid Dynamos	85 052 862
The Riemann Hypothesis (Anneli Lax New Mathematical Library)	2 626 078
Wavelets in Medicine and Biology	16 230 456

Figure 2.26: Books-21-2020-202005121145.png

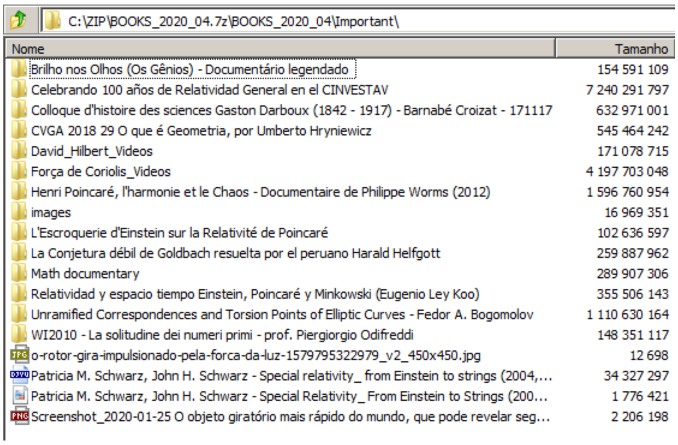

C:\ZIP\BOOKS_2020_04.7z\BOOKS_2020_04\Important\

Nome	Tamanho
Brilho nos Olhos (Os Gênios) - Documentário legendado	154 591 109
Celebrando 100 años de Relatividad General en el CINVESTAV	7 240 291 797
Colloque d'histoire des sciences Gaston Darboux (1842 - 1917) - Barnabé Croizat - 171117	632 971 001
CVGA 2018 29 O que é Geometria, por Umberto Hryniewicz	545 464 242
David_Hilbert_Videos	171 078 715
Força de Coriolis_Videos	4 197 703 048
Henri Poincaré, l'harmonie et le Chaos - Documentaire de Philippe Worms (2012)	1 596 760 954
images	16 969 351
L'Escroquerie d'Einstein sur la Relativité de Poincaré	102 636 597
La Conjetura débil de Goldbach resuelta por el peruano Harald Helfgott	259 887 962
Math documentary	289 907 306
Relatividad y espacio tiempo Einstein, Poincaré y Minkowski (Eugenio Ley Koo)	355 506 143
Unramified Correspondences and Torsion Points of Elliptic Curves - Fedor A. Bogomolov	1 110 630 164
WI2010 - La solitudine dei numeri primi - prof. Piergiorgio Odifreddi	148 351 117
o-rotor-gira-impulsionado-pela-forca-da-luz-1579795322979_v2_450x450.jpg	12 698
Patricia M. Schwarz, John H. Schwarz - Special relativity_ from Einstein to strings (2004,...	34 327 297
Patricia M. Schwarz, John H. Schwarz - Special Relativity_ From Einstein to Strings (200...	1 776 421
Screenshot_2020-01-25 O objeto giratório mais rápido do mundo, que pode revelar seg...	2 206 198

Figure 2.27: Books-21-2020-202005121146.png

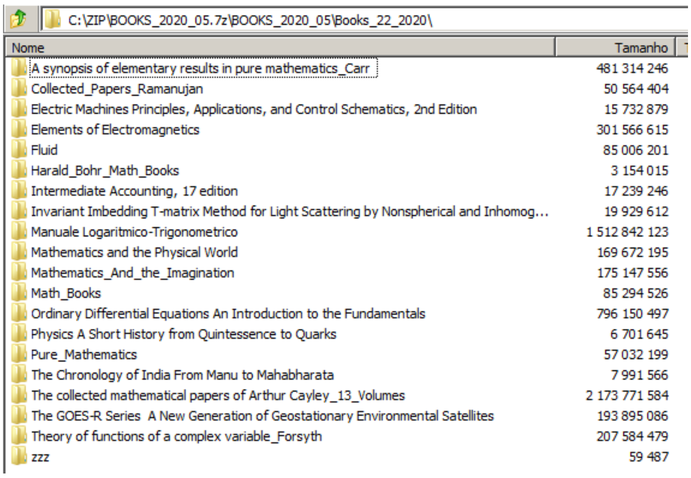

C:\ZIP\BOOKS_2020_05.7z\BOOKS_2020_05\Books_22_2020\

Nome	Tamanho
A synopsis of elementary results in pure mathematics_Carr	481 314 246
Collected_Papers_Ramanujan	50 564 404
Electric Machines Principles, Applications, and Control Schematics, 2nd Edition	15 732 879
Elements of Electromagnetics	301 566 615
Fluid	85 006 201
Harald_Bohr_Math_Books	3 154 015
Intermediate Accounting, 17 edition	17 239 246
Invariant Imbedding T-matrix Method for Light Scattering by Nonspherical and Inhomog...	19 929 612
Manuale Logaritmico-Trigonometrico	1 512 842 123
Mathematics and the Physical World	169 672 195
Mathematics_And_the_Imagination	175 147 556
Math_Books	85 294 526
Ordinary Differential Equations An Introduction to the Fundamentals	796 150 497
Physics A Short History from Quintessence to Quarks	6 701 645
Pure_Mathematics	57 032 199
The Chronology of India From Manu to Mahabharata	7 991 566
The collected mathematical papers of Arthur Cayley_13_Volumes	2 173 771 584
The GOES-R Series A New Generation of Geostationary Environmental Satellites	193 895 086
Theory of functions of a complex variable_Forsyth	207 584 479
zzz	59 487

Figure 2.28: Books-22-2020-202005121317.png

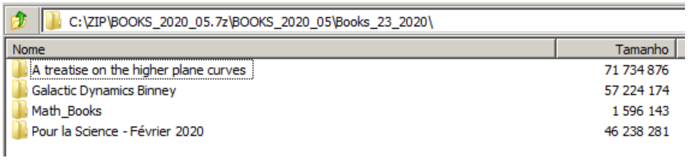

C:\ZIP\BOOKS_2020_05.7z\BOOKS_2020_05\Books_23_2020\

Nome	Tamanho
A treatise on the higher plane curves	71 734 876
Galactic Dynamics Binney	57 224 174
Math_Books	1 596 143
Pour la Science - Février 2020	46 238 281

Figure 2.29: Books-23-2020-202005121318.png

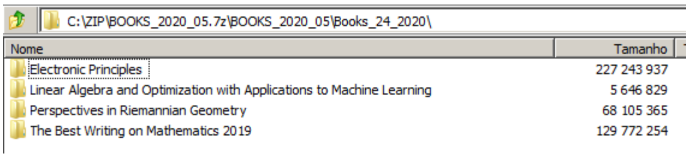

C:\ZIP\BOOKS_2020_05.7z\BOOKS_2020_05\Books_24_2020\

Nome	Tamanho
Electronic Principles	227 243 937
Linear Algebra and Optimization with Applications to Machine Learning	5 646 829
Perspectives in Riemannian Geometry	68 105 365
The Best Writing on Mathematics 2019	129 772 254

Figure 2.30: Books-24-2020-202005121319.png

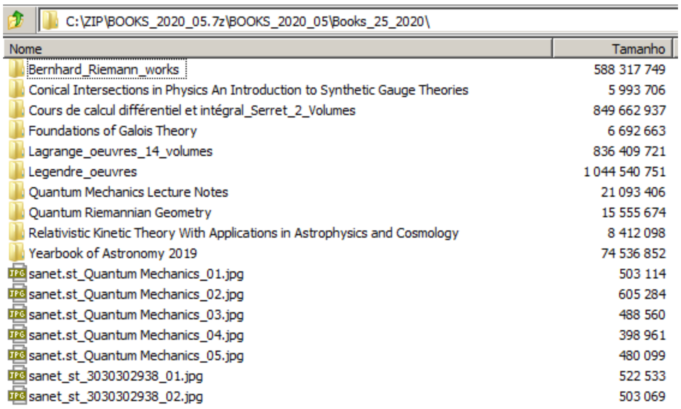

C:\ZIP\BOOKS_2020_05.7z\BOOKS_2020_05\Books_25_2020\

Nome	Tamanho
Bernhard_Riemann_works	588 317 749
Conical Intersections in Physics An Introduction to Synthetic Gauge Theories	5 993 706
Cours de calcul différentiel et intégral_Serret_2_Volumes	849 662 937
Foundations of Galois Theory	6 692 663
Lagrange_oeuvres_14_volumes	836 409 721
Legendre_oeuvres	1 044 540 751
Quantum Mechanics Lecture Notes	21 093 406
Quantum Riemannian Geometry	15 555 674
Relativistic Kinetic Theory With Applications in Astrophysics and Cosmology	8 412 098
Yearbook of Astronomy 2019	74 536 852
sanet.st_Quantum Mechanics_01.jpg	503 114
sanet.st_Quantum Mechanics_02.jpg	605 284
sanet.st_Quantum Mechanics_03.jpg	488 560
sanet.st_Quantum Mechanics_04.jpg	398 961
sanet.st_Quantum Mechanics_05.jpg	480 099
sanet_st_3030302938_01.jpg	522 533
sanet_st_3030302938_02.jpg	503 069

Figure 2.31: Books-25-2020-202005121320.png

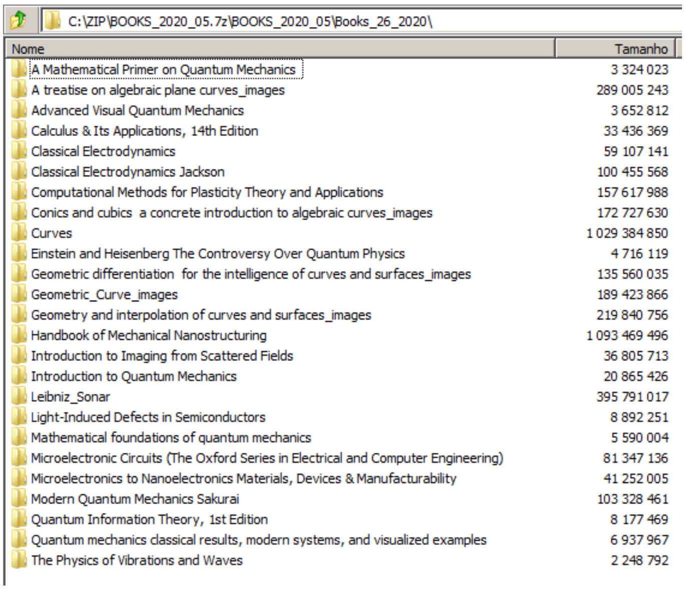

C:\ZIP\BOOKS_2020_05.7z\BOOKS_2020_05\Books_26_2020\

Nome	Tamanho
A Mathematical Primer on Quantum Mechanics	3 324 023
A treatise on algebraic plane curves_images	289 005 243
Advanced Visual Quantum Mechanics	3 652 812
Calculus & Its Applications, 14th Edition	33 436 369
Classical Electrodynamics	59 107 141
Classical Electrodynamics Jackson	100 455 568
Computational Methods for Plasticity Theory and Applications	157 617 988
Conics and cubics a concrete introduction to algebraic curves_images	172 727 630
Curves	1 029 384 850
Einstein and Heisenberg The Controversy Over Quantum Physics	4 716 119
Geometric differentiation for the intelligence of curves and surfaces_images	135 560 035
Geometric_Curve_images	189 423 866
Geometry and interpolation of curves and surfaces_images	219 840 756
Handbook of Mechanical Nanostructuring	1 093 469 496
Introduction to Imaging from Scattered Fields	36 805 713
Introduction to Quantum Mechanics	20 865 426
Leibniz_Sonar	395 791 017
Light-Induced Defects in Semiconductors	8 892 251
Mathematical foundations of quantum mechanics	5 590 004
Microelectronic Circuits (The Oxford Series in Electrical and Computer Engineering)	81 347 136
Microelectronics to Nanoelectronics Materials, Devices & Manufacturability	41 252 005
Modern Quantum Mechanics Sakurai	103 328 461
Quantum Information Theory, 1st Edition	8 177 469
Quantum mechanics classical results, modern systems, and visualized examples	6 937 967
The Physics of Vibrations and Waves	2 248 792

Figure 2.32: Books-26-2020-202005121321.png

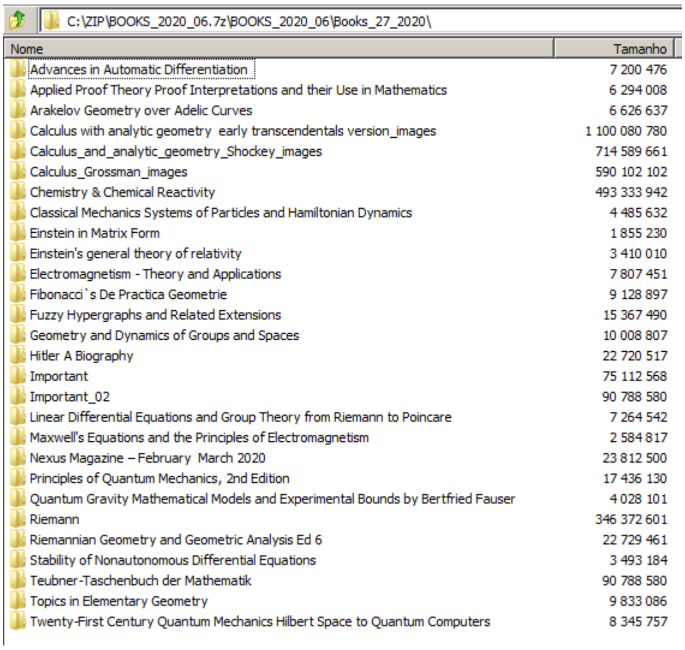

C:\ZIP\BOOKS_2020_06.7z\BOOKS_2020_06\Books_27_2020\

Nome	Tamanho
Advances in Automatic Differentiation	7 200 476
Applied Proof Theory Proof Interpretations and their Use in Mathematics	6 294 008
Arakelov Geometry over Adelic Curves	6 626 637
Calculus with analytic geometry early transcendentals version_images	1 100 080 780
Calculus_and_analytic_geometry_Shockey_images	714 589 661
Calculus_Grossman_images	590 102 102
Chemistry & Chemical Reactivity	493 333 942
Classical Mechanics Systems of Particles and Hamiltonian Dynamics	4 485 632
Einstein in Matrix Form	1 855 230
Einstein's general theory of relativity	3 410 010
Electromagnetism - Theory and Applications	7 807 451
Fibonacci`s De Practica Geometrie	9 128 897
Fuzzy Hypergraphs and Related Extensions	15 367 490
Geometry and Dynamics of Groups and Spaces	10 008 807
Hitler A Biography	22 720 517
Important	75 112 568
Important_02	90 788 580
Linear Differential Equations and Group Theory from Riemann to Poincare	7 264 542
Maxwell's Equations and the Principles of Electromagnetism	2 584 817
Nexus Magazine – February March 2020	23 812 500
Principles of Quantum Mechanics, 2nd Edition	17 436 130
Quantum Gravity Mathematical Models and Experimental Bounds by Bertfried Fauser	4 028 101
Riemann	346 372 601
Riemannian Geometry and Geometric Analysis Ed 6	22 729 461
Stability of Nonautonomous Differential Equations	3 493 184
Teubner-Taschenbuch der Mathematik	90 788 580
Topics in Elementary Geometry	9 833 086
Twenty-First Century Quantum Mechanics Hilbert Space to Quantum Computers	8 345 757

Figure 2.33: Books-27-2020-202005121347.png

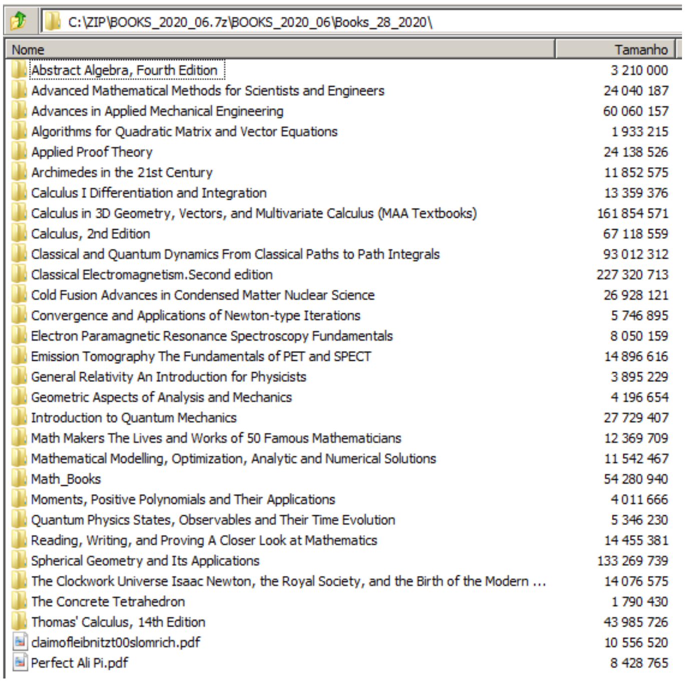

C:\ZIP\BOOKS_2020_06.7z\BOOKS_2020_06\Books_28_2020\

Nome	Tamanho
Abstract Algebra, Fourth Edition	3 210 000
Advanced Mathematical Methods for Scientists and Engineers	24 040 187
Advances in Applied Mechanical Engineering	60 060 157
Algorithms for Quadratic Matrix and Vector Equations	1 933 215
Applied Proof Theory	24 138 526
Archimedes in the 21st Century	11 852 575
Calculus I Differentiation and Integration	13 359 376
Calculus in 3D Geometry, Vectors, and Multivariate Calculus (MAA Textbooks)	161 854 571
Calculus, 2nd Edition	67 118 559
Classical and Quantum Dynamics From Classical Paths to Path Integrals	93 012 312
Classical Electromagnetism.Second edition	227 320 713
Cold Fusion Advances in Condensed Matter Nuclear Science	26 928 121
Convergence and Applications of Newton-type Iterations	5 746 895
Electron Paramagnetic Resonance Spectroscopy Fundamentals	8 050 159
Emission Tomography The Fundamentals of PET and SPECT	14 896 616
General Relativity An Introduction for Physicists	3 895 229
Geometric Aspects of Analysis and Mechanics	4 196 654
Introduction to Quantum Mechanics	27 729 407
Math Makers The Lives and Works of 50 Famous Mathematicians	12 369 709
Mathematical Modelling, Optimization, Analytic and Numerical Solutions	11 542 467
Math_Books	54 280 940
Moments, Positive Polynomials and Their Applications	4 011 666
Quantum Physics States, Observables and Their Time Evolution	5 346 230
Reading, Writing, and Proving A Closer Look at Mathematics	14 455 381
Spherical Geometry and Its Applications	133 269 739
The Clockwork Universe Isaac Newton, the Royal Society, and the Birth of the Modern ...	14 076 575
The Concrete Tetrahedron	1 790 430
Thomas' Calculus, 14th Edition	43 985 726
claimofleibnitzt00slomrich.pdf	10 556 520
Perfect Ali Pi.pdf	8 428 765

Figure 2.34: Books-28-2020-202005121348.png

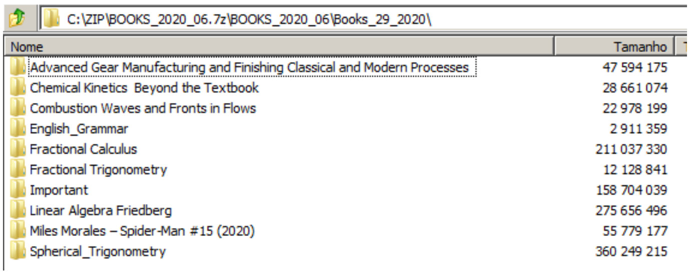

C:\ZIP\BOOKS_2020_06.7z\BOOKS_2020_06\Books_29_2020\

Nome	Tamanho
Advanced Gear Manufacturing and Finishing Classical and Modern Processes	47 594 175
Chemical Kinetics Beyond the Textbook	28 661 074
Combustion Waves and Fronts in Flows	22 978 199
English_Grammar	2 911 359
Fractional Calculus	211 037 330
Fractional Trigonometry	12 128 841
Important	158 704 039
Linear Algebra Friedberg	275 656 496
Miles Morales – Spider-Man #15 (2020)	55 779 177
Spherical_Trigonometry	360 249 215

Figure 2.35: Books-29-2020-202005121349.png

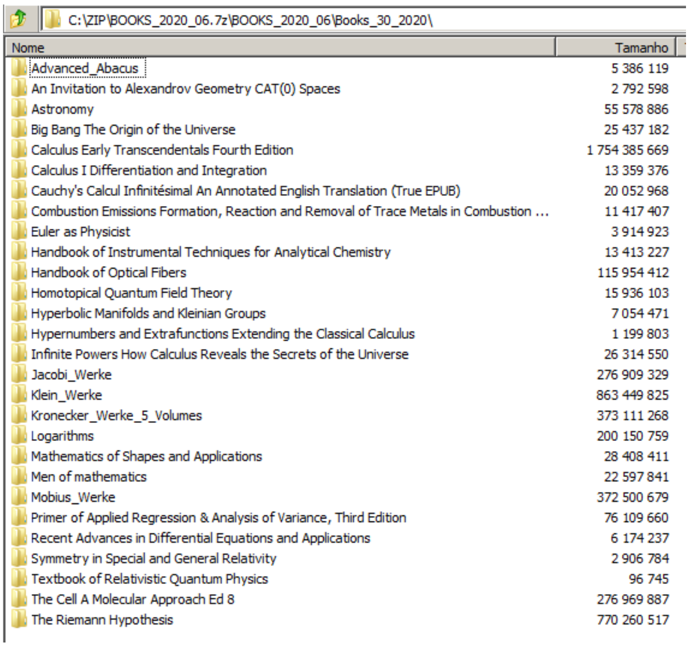

C:\ZIP\BOOKS_2020_06.7z\BOOKS_2020_06\Books_30_2020\

Nome	Tamanho
Advanced_Abacus	5 386 119
An Invitation to Alexandrov Geometry CAT(0) Spaces	2 792 598
Astronomy	55 578 886
Big Bang The Origin of the Universe	25 437 182
Calculus Early Transcendentals Fourth Edition	1 754 385 669
Calculus I Differentiation and Integration	13 359 376
Cauchy's Calcul Infinitésimal An Annotated English Translation (True EPUB)	20 052 968
Combustion Emissions Formation, Reaction and Removal of Trace Metals in Combustion ...	11 417 407
Euler as Physicist	3 914 923
Handbook of Instrumental Techniques for Analytical Chemistry	13 413 227
Handbook of Optical Fibers	115 954 412
Homotopical Quantum Field Theory	15 936 103
Hyperbolic Manifolds and Kleinian Groups	7 054 471
Hypernumbers and Extrafunctions Extending the Classical Calculus	1 199 803
Infinite Powers How Calculus Reveals the Secrets of the Universe	26 314 550
Jacobi_Werke	276 909 329
Klein_Werke	863 449 825
Kronecker_Werke_5_Volumes	373 111 268
Logarithms	200 150 759
Mathematics of Shapes and Applications	28 408 411
Men of mathematics	22 597 841
Mobius_Werke	372 500 679
Primer of Applied Regression & Analysis of Variance, Third Edition	76 109 660
Recent Advances in Differential Equations and Applications	6 174 237
Symmetry in Special and General Relativity	2 906 784
Textbook of Relativistic Quantum Physics	96 745
The Cell A Molecular Approach Ed 8	276 969 887
The Riemann Hypothesis	770 260 517

Figure 2.36: Books-30-2020-202005121350.png

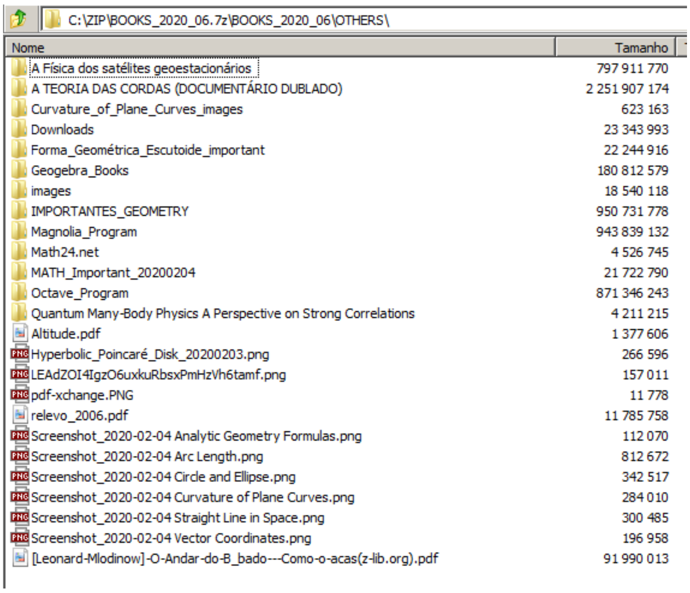

C:\ZIP\BOOKS_2020_06.7z\BOOKS_2020_06\OTHERS\

Nome	Tamanho
A Física dos satélites geoestacionários	797 911 770
A TEORIA DAS CORDAS (DOCUMENTÁRIO DUBLADO)	2 251 907 174
Curvature_of_Plane_Curves_images	623 163
Downloads	23 343 993
Forma_Geométrica_Escutoide_important	22 244 916
Geogebra_Books	180 812 579
images	18 540 118
IMPORTANTES_GEOMETRY	950 731 778
Magnolia_Program	943 839 132
Math24.net	4 526 745
MATH_Important_20200204	21 722 790
Octave_Program	871 346 243
Quantum Many-Body Physics A Perspective on Strong Correlations	4 211 215
Altitude.pdf	1 377 606
Hyperbolic_Poincaré_Disk_20200203.png	266 596
LEAdZOI4IgzO6uxkuRbsxPmHzVh6tamf.png	157 011
pdf-xchange.PNG	11 778
relevo_2006.pdf	11 785 758
Screenshot_2020-02-04 Analytic Geometry Formulas.png	112 070
Screenshot_2020-02-04 Arc Length.png	812 672
Screenshot_2020-02-04 Circle and Ellipse.png	342 517
Screenshot_2020-02-04 Curvature of Plane Curves.png	284 010
Screenshot_2020-02-04 Straight Line in Space.png	300 485
Screenshot_2020-02-04 Vector Coordinates.png	196 958
[Leonard-Mlodinow]-O-Andar-do-B_bado---Como-o-acas(z-lib.org).pdf	91 990 013

Figure 2.37: Books-30-2020-202005121351.png

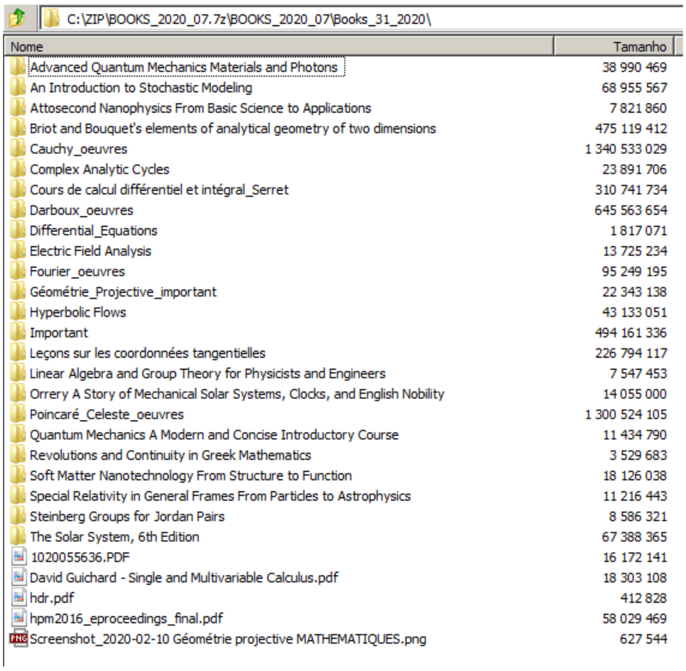
C:\ZIP\BOOKS_2020_07.7z\BOOKS_2020_07\Books_31_2020\

Nome	Tamanho
Advanced Quantum Mechanics Materials and Photons	38 990 469
An Introduction to Stochastic Modeling	68 955 567
Attosecond Nanophysics From Basic Science to Applications	7 821 860
Briot and Bouquet's elements of analytical geometry of two dimensions	475 119 412
Cauchy_oeuvres	1 340 533 029
Complex Analytic Cycles	23 891 706
Cours de calcul différentiel et intégral_Serret	310 741 734
Darboux_oeuvres	645 563 654
Differential_Equations	1 817 071
Electric Field Analysis	13 725 234
Fourier_oeuvres	95 249 195
Géométrie_Projective_important	22 343 138
Hyperbolic Flows	43 133 051
Important	494 161 336
Leçons sur les coordonnées tangentielles	226 794 117
Linear Algebra and Group Theory for Physicists and Engineers	7 547 453
Orrery A Story of Mechanical Solar Systems, Clocks, and English Nobility	14 055 000
Poincaré_Celeste_oeuvres	1 300 524 105
Quantum Mechanics A Modern and Concise Introductory Course	11 434 790
Revolutions and Continuity in Greek Mathematics	3 529 683
Soft Matter Nanotechnology From Structure to Function	18 126 038
Special Relativity in General Frames From Particles to Astrophysics	11 216 443
Steinberg Groups for Jordan Pairs	8 586 321
The Solar System, 6th Edition	67 388 365
1020055636.PDF	16 172 141
David Guichard - Single and Multivariable Calculus.pdf	18 303 108
hdr.pdf	412 828
hpm2016_eproceedings_final.pdf	58 029 469
Screenshot_2020-02-10 Géométrie projective MATHEMATIQUES.png	627 544

Figure 2.38: Books-30-2020-202005121401.png

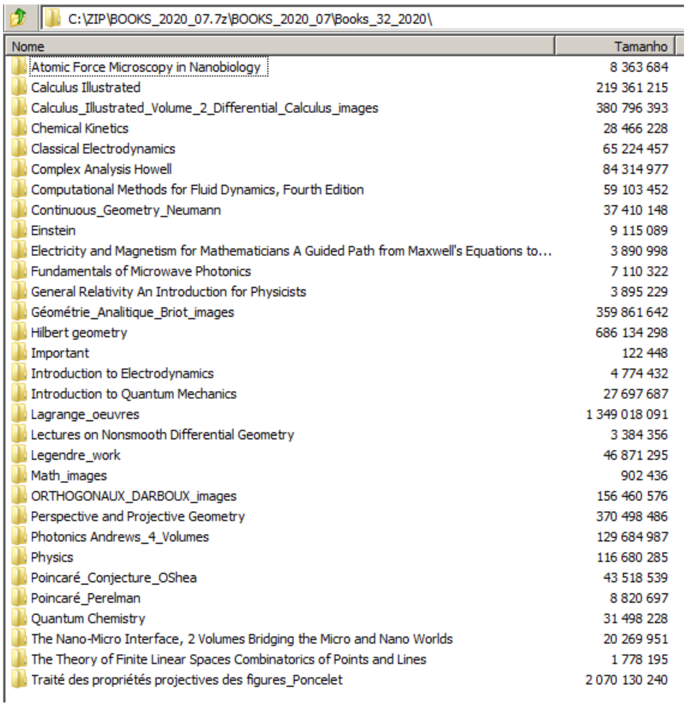
C:\ZIP\BOOKS_2020_07.7z\BOOKS_2020_07\Books_32_2020\

Nome	Tamanho
Atomic Force Microscopy in Nanobiology	8 363 684
Calculus Illustrated	219 361 215
Calculus_Illustrated_Volume_2_Differential_Calculus_images	380 796 393
Chemical Kinetics	28 466 228
Classical Electrodynamics	65 224 457
Complex Analysis Howell	84 314 977
Computational Methods for Fluid Dynamics, Fourth Edition	59 103 452
Continuous_Geometry_Neumann	37 410 148
Einstein	9 115 089
Electricity and Magnetism for Mathematicians A Guided Path from Maxwell's Equations to...	3 890 998
Fundamentals of Microwave Photonics	7 110 322
General Relativity An Introduction for Physicists	3 895 229
Géométrie_Analitique_Briot_images	359 861 642
Hilbert geometry	686 134 298
Important	122 448
Introduction to Electrodynamics	4 774 432
Introduction to Quantum Mechanics	27 697 687
Lagrange_oeuvres	1 349 018 091
Lectures on Nonsmooth Differential Geometry	3 384 356
Legendre_work	46 871 295
Math_images	902 436
ORTHOGONAUX_DARBOUX_images	156 460 576
Perspective and Projective Geometry	370 498 486
Photonics Andrews_4_Volumes	129 684 987
Physics	116 680 285
Poincaré_Conjecture_OShea	43 518 539
Poincaré_Perelman	8 820 697
Quantum Chemistry	31 498 228
The Nano-Micro Interface, 2 Volumes Bridging the Micro and Nano Worlds	20 269 951
The Theory of Finite Linear Spaces Combinatorics of Points and Lines	1 778 195
Traité des propriétés projectives des figures_Poncelet	2 070 130 240

Figure 2.39: Books-32-2020-202005121402.png

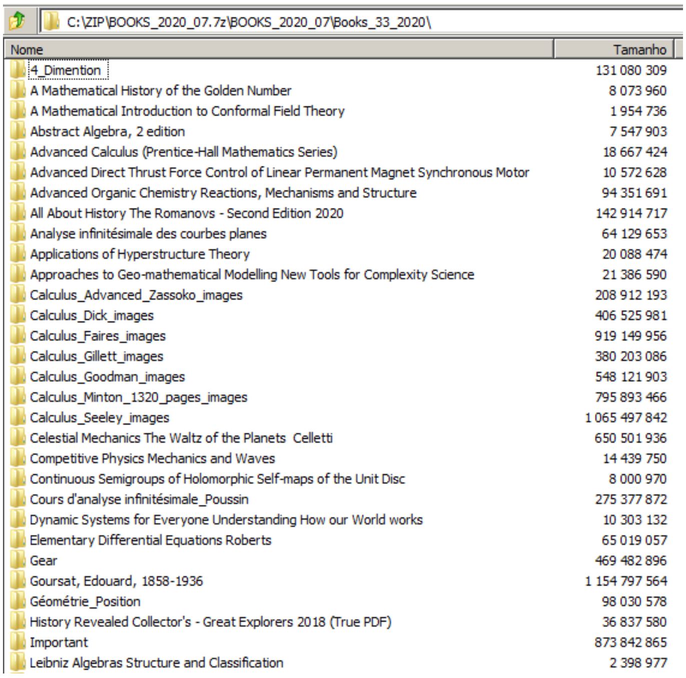

C:\ZIP\BOOKS_2020_07.7z\BOOKS_2020_07\Books_33_2020\

Nome	Tamanho
4_Dimention	131 080 309
A Mathematical History of the Golden Number	8 073 960
A Mathematical Introduction to Conformal Field Theory	1 954 736
Abstract Algebra, 2 edition	7 547 903
Advanced Calculus (Prentice-Hall Mathematics Series)	18 667 424
Advanced Direct Thrust Force Control of Linear Permanent Magnet Synchronous Motor	10 572 628
Advanced Organic Chemistry Reactions, Mechanisms and Structure	94 351 691
All About History The Romanovs - Second Edition 2020	142 914 717
Analyse infinitésimale des courbes planes	64 129 653
Applications of Hyperstructure Theory	20 088 474
Approaches to Geo-mathematical Modelling New Tools for Complexity Science	21 386 590
Calculus_Advanced_Zassoko_images	208 912 193
Calculus_Dick_images	406 525 981
Calculus_Faires_images	919 149 956
Calculus_Gillett_images	380 203 086
Calculus_Goodman_images	548 121 903
Calculus_Minton_1320_pages_images	795 893 466
Calculus_Seeley_images	1 065 497 842
Celestial Mechanics The Waltz of the Planets Celletti	650 501 936
Competitive Physics Mechanics and Waves	14 439 750
Continuous Semigroups of Holomorphic Self-maps of the Unit Disc	8 000 970
Cours d'analyse infinitésimale_Poussin	275 377 872
Dynamic Systems for Everyone Understanding How our World works	10 303 132
Elementary Differential Equations Roberts	65 019 057
Gear	469 482 896
Goursat, Edouard, 1858-1936	1 154 797 564
Géométrie_Position	98 030 578
History Revealed Collector's - Great Explorers 2018 (True PDF)	36 837 580
Important	873 842 865
Leibniz Algebras Structure and Classification	2 398 977

Figure 2.40: Books-33-2020-202005121403.png

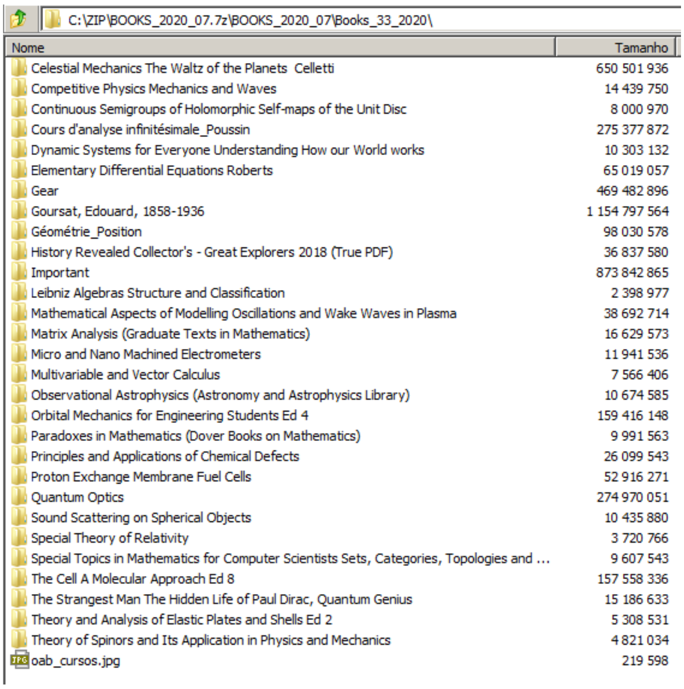

C:\ZIP\BOOKS_2020_07.7z\BOOKS_2020_07\Books_33_2020\

Nome	Tamanho
Celestial Mechanics The Waltz of the Planets Celletti	650 501 936
Competitive Physics Mechanics and Waves	14 439 750
Continuous Semigroups of Holomorphic Self-maps of the Unit Disc	8 000 970
Cours d'analyse infinitésimale_Poussin	275 377 872
Dynamic Systems for Everyone Understanding How our World works	10 303 132
Elementary Differential Equations Roberts	65 019 057
Gear	469 482 896
Goursat, Edouard, 1858-1936	1 154 797 564
Géométrie_Position	98 030 578
History Revealed Collector's - Great Explorers 2018 (True PDF)	36 837 580
Important	873 842 865
Leibniz Algebras Structure and Classification	2 398 977
Mathematical Aspects of Modelling Oscillations and Wake Waves in Plasma	38 692 714
Matrix Analysis (Graduate Texts in Mathematics)	16 629 573
Micro and Nano Machined Electrometers	11 941 536
Multivariable and Vector Calculus	7 566 406
Observational Astrophysics (Astronomy and Astrophysics Library)	10 674 585
Orbital Mechanics for Engineering Students Ed 4	159 416 148
Paradoxes in Mathematics (Dover Books on Mathematics)	9 991 563
Principles and Applications of Chemical Defects	26 099 543
Proton Exchange Membrane Fuel Cells	52 916 271
Quantum Optics	274 970 051
Sound Scattering on Spherical Objects	10 435 880
Special Theory of Relativity	3 720 766
Special Topics in Mathematics for Computer Scientists Sets, Categories, Topologies and ...	9 607 543
The Cell A Molecular Approach Ed 8	157 558 336
The Strangest Man The Hidden Life of Paul Dirac, Quantum Genius	15 186 633
Theory and Analysis of Elastic Plates and Shells Ed 2	5 308 531
Theory of Spinors and Its Application in Physics and Mechanics	4 821 034
oab_cursos.jpg	219 598

Figure 2.41: Books-33-2020-202005121404.png

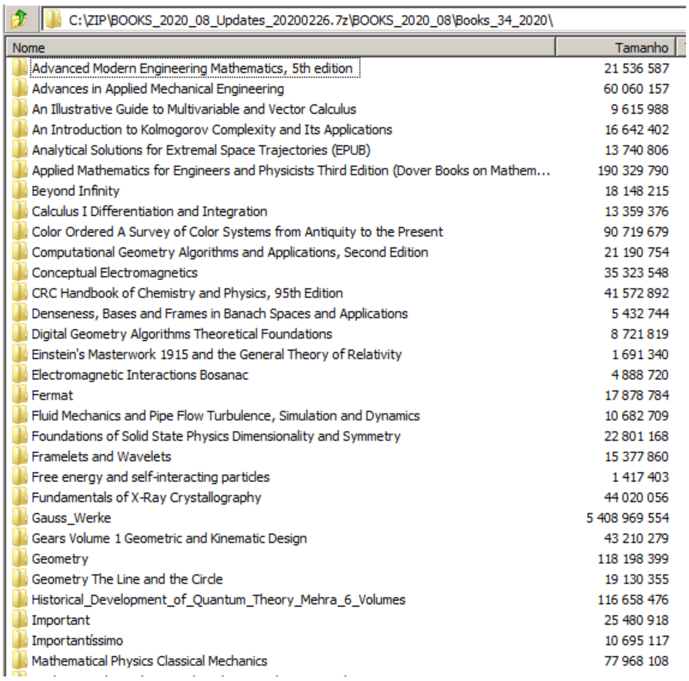

C:\ZIP\BOOKS_2020_08_Updates_20200226.7z\BOOKS_2020_08\Books_34_2020\

Nome	Tamanho
Advanced Modern Engineering Mathematics, 5th edition	21 536 587
Advances in Applied Mechanical Engineering	60 060 157
An Illustrative Guide to Multivariable and Vector Calculus	9 615 988
An Introduction to Kolmogorov Complexity and Its Applications	16 642 402
Analytical Solutions for Extremal Space Trajectories (EPUB)	13 740 806
Applied Mathematics for Engineers and Physicists Third Edition (Dover Books on Mathem...	190 329 790
Beyond Infinity	18 148 215
Calculus I Differentiation and Integration	13 359 376
Color Ordered A Survey of Color Systems from Antiquity to the Present	90 719 679
Computational Geometry Algorithms and Applications, Second Edition	21 190 754
Conceptual Electromagnetics	35 323 548
CRC Handbook of Chemistry and Physics, 95th Edition	41 572 892
Denseness, Bases and Frames in Banach Spaces and Applications	5 432 744
Digital Geometry Algorithms Theoretical Foundations	8 721 819
Einstein's Masterwork 1915 and the General Theory of Relativity	1 691 340
Electromagnetic Interactions Bosanac	4 888 720
Fermat	17 878 784
Fluid Mechanics and Pipe Flow Turbulence, Simulation and Dynamics	10 682 709
Foundations of Solid State Physics Dimensionality and Symmetry	22 801 168
Framelets and Wavelets	15 377 860
Free energy and self-interacting particles	1 417 403
Fundamentals of X-Ray Crystallography	44 020 056
Gauss_Werke	5 408 969 554
Gears Volume 1 Geometric and Kinematic Design	43 210 279
Geometry	118 198 399
Geometry The Line and the Circle	19 130 355
Historical_Development_of_Quantum_Theory_Mehra_6_Volumes	116 658 476
Important	25 480 918
Importantíssimo	10 695 117
Mathematical Physics Classical Mechanics	77 968 108

Figure 2.42: Books-34-2020-202005121438.png

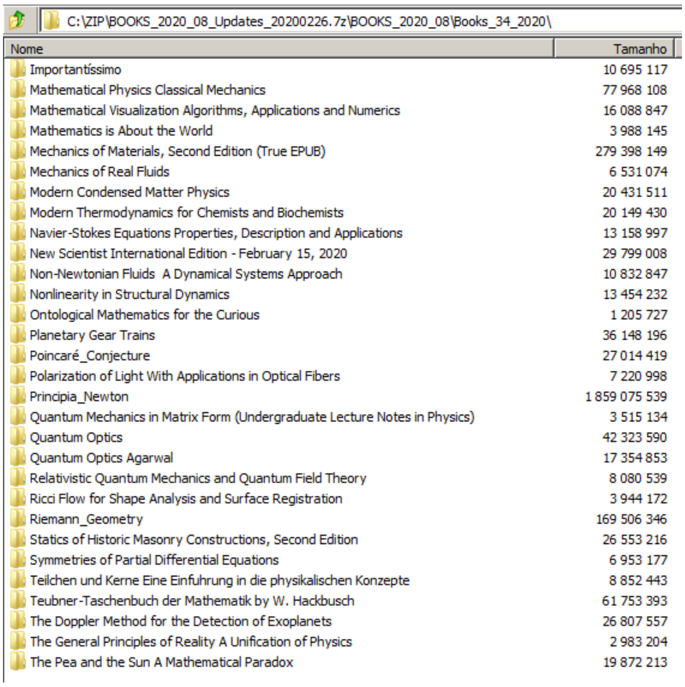

C:\ZIP\BOOKS_2020_08_Updates_20200226.7z\BOOKS_2020_08\Books_34_2020\

Nome	Tamanho
Importantíssimo	10 695 117
Mathematical Physics Classical Mechanics	77 968 108
Mathematical Visualization Algorithms, Applications and Numerics	16 088 847
Mathematics is About the World	3 988 145
Mechanics of Materials, Second Edition (True EPUB)	279 398 149
Mechanics of Real Fluids	6 531 074
Modern Condensed Matter Physics	20 431 511
Modern Thermodynamics for Chemists and Biochemists	20 149 430
Navier-Stokes Equations Properties, Description and Applications	13 158 997
New Scientist International Edition - February 15, 2020	29 799 008
Non-Newtonian Fluids A Dynamical Systems Approach	10 832 847
Nonlinearity in Structural Dynamics	13 454 232
Ontological Mathematics for the Curious	1 205 727
Planetary Gear Trains	36 148 196
Poincaré_Conjecture	27 014 419
Polarization of Light With Applications in Optical Fibers	7 220 998
Principia_Newton	1 859 075 539
Quantum Mechanics in Matrix Form (Undergraduate Lecture Notes in Physics)	3 515 134
Quantum Optics	42 323 590
Quantum Optics Agarwal	17 354 853
Relativistic Quantum Mechanics and Quantum Field Theory	8 080 539
Ricci Flow for Shape Analysis and Surface Registration	3 944 172
Riemann_Geometry	169 506 346
Statics of Historic Masonry Constructions, Second Edition	26 553 216
Symmetries of Partial Differential Equations	6 953 177
Teilchen und Kerne Eine Einfuhrung in die physikalischen Konzepte	8 852 443
Teubner-Taschenbuch der Mathematik by W. Hackbusch	61 753 393
The Doppler Method for the Detection of Exoplanets	26 807 557
The General Principles of Reality A Unification of Physics	2 983 204
The Pea and the Sun A Mathematical Paradox	19 872 213

Figure 2.43: Books-34-2020-202005121439.png

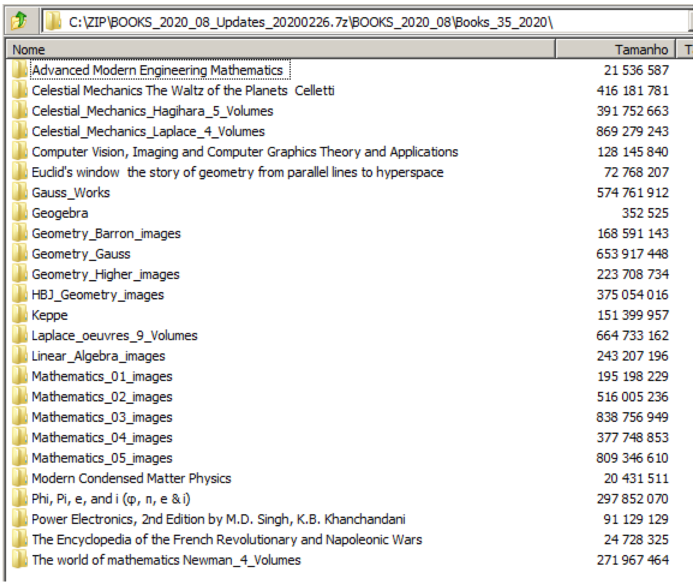

C:\ZIP\BOOKS_2020_08_Updates_20200226.7z\BOOKS_2020_08\Books_35_2020\

Nome	Tamanho
Advanced Modern Engineering Mathematics	21 536 587
Celestial Mechanics The Waltz of the Planets Celletti	416 181 781
Celestial_Mechanics_Hagihara_5_Volumes	391 752 663
Celestial_Mechanics_Laplace_4_Volumes	869 279 243
Computer Vision, Imaging and Computer Graphics Theory and Applications	128 145 840
Euclid's window the story of geometry from parallel lines to hyperspace	72 768 207
Gauss_Works	574 761 912
Geogebra	352 525
Geometry_Barron_images	168 591 143
Geometry_Gauss	653 917 448
Geometry_Higher_images	223 708 734
HBJ_Geometry_images	375 054 016
Keppe	151 399 957
Laplace_oeuvres_9_Volumes	664 733 162
Linear_Algebra_images	243 207 196
Mathematics_01_images	195 198 229
Mathematics_02_images	516 005 236
Mathematics_03_images	838 756 949
Mathematics_04_images	377 748 853
Mathematics_05_images	809 346 610
Modern Condensed Matter Physics	20 431 511
Phi, Pi, e, and i (φ, π, e & i)	297 852 070
Power Electronics, 2nd Edition by M.D. Singh, K.B. Khanchandani	91 129 129
The Encyclopedia of the French Revolutionary and Napoleonic Wars	24 728 325
The world of mathematics Newman_4_Volumes	271 967 464

Figure 2.44: Books-35-2020-202005121439.png

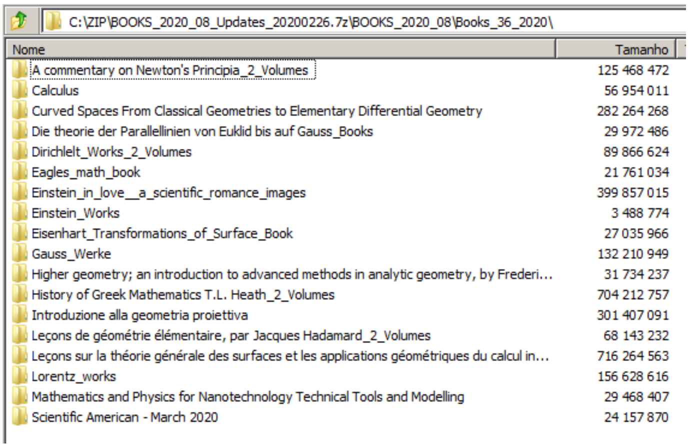

C:\ZIP\BOOKS_2020_08_Updates_20200226.7z\BOOKS_2020_08\Books_36_2020\

Nome	Tamanho
A commentary on Newton's Principia_2_Volumes	125 468 472
Calculus	56 954 011
Curved Spaces From Classical Geometries to Elementary Differential Geometry	282 264 268
Die theorie der Parallellinien von Euklid bis auf Gauss_Books	29 972 486
Dirichlelt_Works_2_Volumes	89 866 624
Eagles_math_book	21 761 034
Einstein_in_love__a_scientific_romance_images	399 857 015
Einstein_Works	3 488 774
Eisenhart_Transformations_of_Surface_Book	27 035 966
Gauss_Werke	132 210 949
Higher geometry; an introduction to advanced methods in analytic geometry, by Frederi...	31 734 237
History of Greek Mathematics T.L. Heath_2_Volumes	704 212 757
Introduzione alla geometria proiettiva	301 407 091
Leçons de géométrie élémentaire, par Jacques Hadamard_2_Volumes	68 143 232
Leçons sur la théorie générale des surfaces et les applications géométriques du calcul in...	716 264 563
Lorentz_works	156 628 616
Mathematics and Physics for Nanotechnology Technical Tools and Modelling	29 468 407
Scientific American - March 2020	24 157 870

Figure 2.45: Books-36-2020-202005121439.png

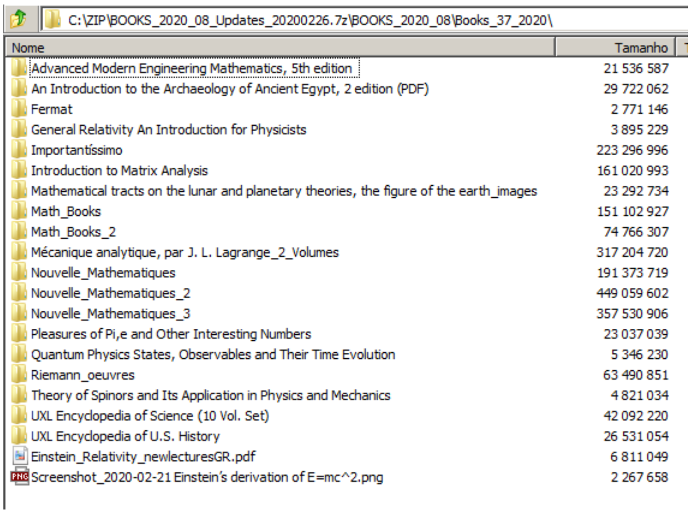

C:\ZIP\BOOKS_2020_08_Updates_20200226.7z\BOOKS_2020_08\Books_37_2020\

Nome	Tamanho
Advanced Modern Engineering Mathematics, 5th edition	21 536 587
An Introduction to the Archaeology of Ancient Egypt, 2 edition (PDF)	29 722 062
Fermat	2 771 146
General Relativity An Introduction for Physicists	3 895 229
Importantíssimo	223 296 996
Introduction to Matrix Analysis	161 020 993
Mathematical tracts on the lunar and planetary theories, the figure of the earth_images	23 292 734
Math_Books	151 102 927
Math_Books_2	74 766 307
Mécanique analytique, par J. L. Lagrange_2_Volumes	317 204 720
Nouvelle_Mathematiques	191 373 719
Nouvelle_Mathematiques_2	449 059 602
Nouvelle_Mathematiques_3	357 530 906
Pleasures of Pi,e and Other Interesting Numbers	23 037 039
Quantum Physics States, Observables and Their Time Evolution	5 346 230
Riemann_oeuvres	63 490 851
Theory of Spinors and Its Application in Physics and Mechanics	4 821 034
UXL Encyclopedia of Science (10 Vol. Set)	42 092 220
UXL Encyclopedia of U.S. History	26 531 054
Einstein_Relativity_newlecturesGR.pdf	6 811 049
Screenshot_2020-02-21 Einstein's derivation of E=mc^2.png	2 267 658

Figure 2.46: Books-37-2020-202005121439.png

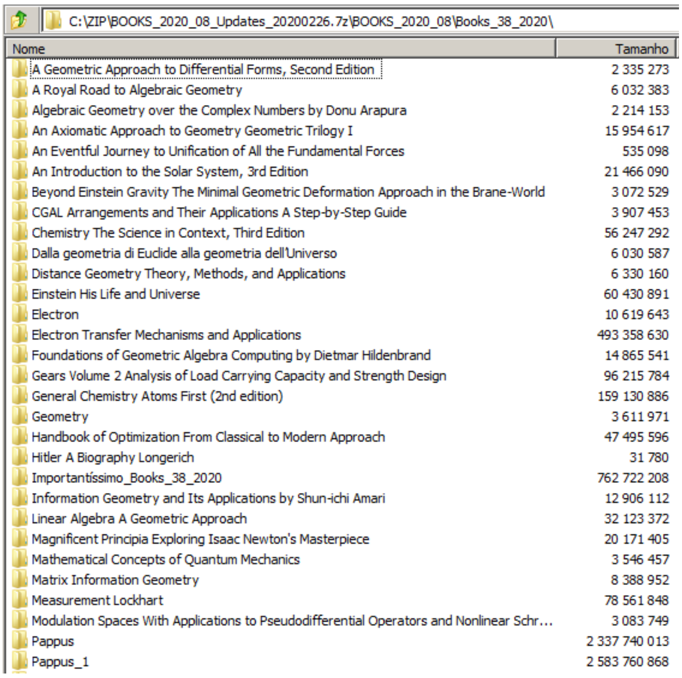

C:\ZIP\BOOKS_2020_08_Updates_20200226.7z\BOOKS_2020_08\Books_38_2020\

Nome	Tamanho
A Geometric Approach to Differential Forms, Second Edition	2 335 273
A Royal Road to Algebraic Geometry	6 032 383
Algebraic Geometry over the Complex Numbers by Donu Arapura	2 214 153
An Axiomatic Approach to Geometry Geometric Trilogy I	15 954 617
An Eventful Journey to Unification of All the Fundamental Forces	535 098
An Introduction to the Solar System, 3rd Edition	21 466 090
Beyond Einstein Gravity The Minimal Geometric Deformation Approach in the Brane-World	3 072 529
CGAL Arrangements and Their Applications A Step-by-Step Guide	3 907 453
Chemistry The Science in Context, Third Edition	56 247 292
Dalla geometria di Euclide alla geometria dell'Universo	6 030 587
Distance Geometry Theory, Methods, and Applications	6 330 160
Einstein His Life and Universe	60 430 891
Electron	10 619 643
Electron Transfer Mechanisms and Applications	493 358 630
Foundations of Geometric Algebra Computing by Dietmar Hildenbrand	14 865 541
Gears Volume 2 Analysis of Load Carrying Capacity and Strength Design	96 215 784
General Chemistry Atoms First (2nd edition)	159 130 886
Geometry	3 611 971
Handbook of Optimization From Classical to Modern Approach	47 495 596
Hitler A Biography Longerich	31 780
Importantíssimo_Books_38_2020	762 722 208
Information Geometry and Its Applications by Shun-ichi Amari	12 906 112
Linear Algebra A Geometric Approach	32 123 372
Magnificent Principia Exploring Isaac Newton's Masterpiece	20 171 405
Mathematical Concepts of Quantum Mechanics	3 546 457
Matrix Information Geometry	8 388 952
Measurement Lockhart	78 561 848
Modulation Spaces With Applications to Pseudodifferential Operators and Nonlinear Schr...	3 083 749
Pappus	2 337 740 013
Pappus_1	2 583 760 868

Figure 2.47: Books-38-2020-202005121439.png

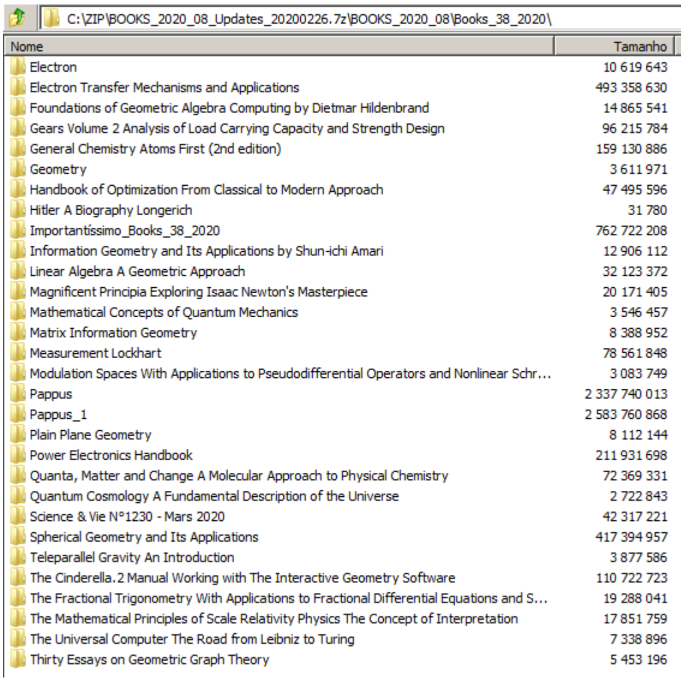

C:\ZIP\BOOKS_2020_08_Updates_20200226.7z\BOOKS_2020_08\Books_38_2020\

Nome	Tamanho
Electron	10 619 643
Electron Transfer Mechanisms and Applications	493 358 630
Foundations of Geometric Algebra Computing by Dietmar Hildenbrand	14 865 541
Gears Volume 2 Analysis of Load Carrying Capacity and Strength Design	96 215 784
General Chemistry Atoms First (2nd edition)	159 130 886
Geometry	3 611 971
Handbook of Optimization From Classical to Modern Approach	47 495 596
Hitler A Biography Longerich	31 780
Importantíssimo_Books_38_2020	762 722 208
Information Geometry and Its Applications by Shun-ichi Amari	12 906 112
Linear Algebra A Geometric Approach	32 123 372
Magnificent Principia Exploring Isaac Newton's Masterpiece	20 171 405
Mathematical Concepts of Quantum Mechanics	3 546 457
Matrix Information Geometry	8 388 952
Measurement Lockhart	78 561 848
Modulation Spaces With Applications to Pseudodifferential Operators and Nonlinear Schr...	3 083 749
Pappus	2 337 740 013
Pappus_1	2 583 760 868
Plain Plane Geometry	8 112 144
Power Electronics Handbook	211 931 698
Quanta, Matter and Change A Molecular Approach to Physical Chemistry	72 369 331
Quantum Cosmology A Fundamental Description of the Universe	2 722 843
Science & Vie N°1230 - Mars 2020	42 317 221
Spherical Geometry and Its Applications	417 394 957
Teleparallel Gravity An Introduction	3 877 586
The Cinderella.2 Manual Working with The Interactive Geometry Software	110 722 723
The Fractional Trigonometry With Applications to Fractional Differential Equations and S...	19 288 041
The Mathematical Principles of Scale Relativity Physics The Concept of Interpretation	17 851 759
The Universal Computer The Road from Leibniz to Turing	7 338 896
Thirty Essays on Geometric Graph Theory	5 453 196

Figure 2.48: Books-38-2020-202005121440.png

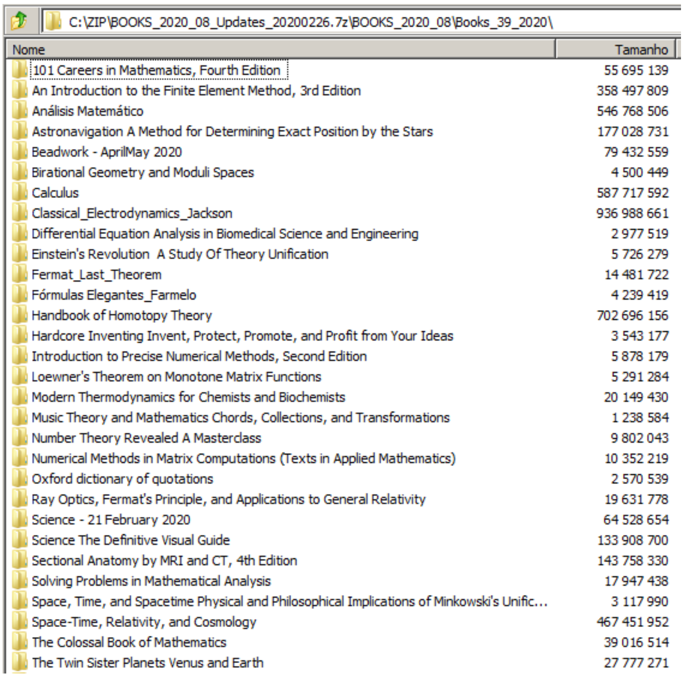

C:\ZIP\BOOKS_2020_08_Updates_20200226.7z\BOOKS_2020_08\Books_39_2020\

Nome	Tamanho
101 Careers in Mathematics, Fourth Edition	55 695 139
An Introduction to the Finite Element Method, 3rd Edition	358 497 809
Análisis Matemático	546 768 506
Astronavigation A Method for Determining Exact Position by the Stars	177 028 731
Beadwork - AprilMay 2020	79 432 559
Birational Geometry and Moduli Spaces	4 500 449
Calculus	587 717 592
Classical_Electrodynamics_Jackson	936 988 661
Differential Equation Analysis in Biomedical Science and Engineering	2 977 519
Einstein's Revolution A Study Of Theory Unification	5 726 279
Fermat_Last_Theorem	14 481 722
Fórmulas Elegantes_Farmelo	4 239 419
Handbook of Homotopy Theory	702 696 156
Hardcore Inventing Invent, Protect, Promote, and Profit from Your Ideas	3 543 177
Introduction to Precise Numerical Methods, Second Edition	5 878 179
Loewner's Theorem on Monotone Matrix Functions	5 291 284
Modern Thermodynamics for Chemists and Biochemists	20 149 430
Music Theory and Mathematics Chords, Collections, and Transformations	1 238 584
Number Theory Revealed A Masterclass	9 802 043
Numerical Methods in Matrix Computations (Texts in Applied Mathematics)	10 352 219
Oxford dictionary of quotations	2 570 539
Ray Optics, Fermat's Principle, and Applications to General Relativity	19 631 778
Science - 21 February 2020	64 528 654
Science The Definitive Visual Guide	133 908 700
Sectional Anatomy by MRI and CT, 4th Edition	143 758 330
Solving Problems in Mathematical Analysis	17 947 438
Space, Time, and Spacetime Physical and Philosophical Implications of Minkowski's Unific...	3 117 990
Space-Time, Relativity, and Cosmology	467 451 952
The Colossal Book of Mathematics	39 016 514
The Twin Sister Planets Venus and Earth	27 777 271

Figure 2.49: Books-39-2020-202005121440.png

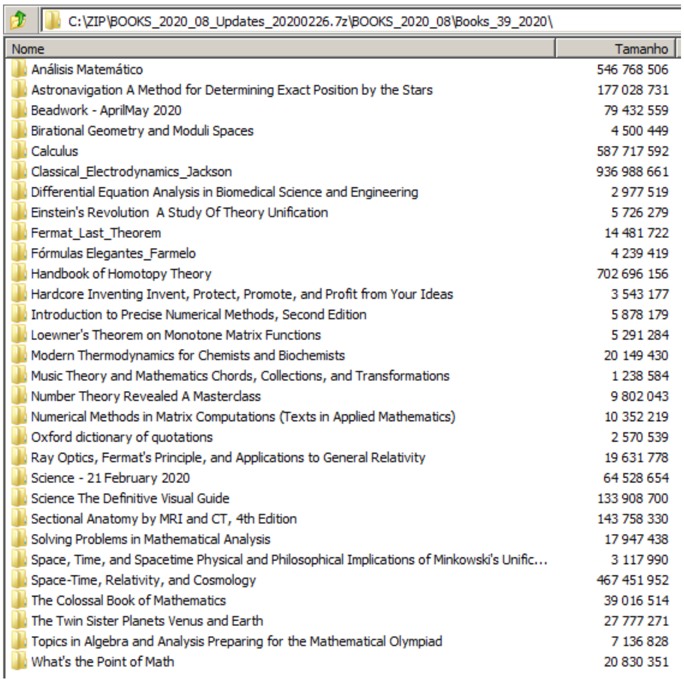

C:\ZIP\BOOKS_2020_08_Updates_20200226.7z\BOOKS_2020_08\Books_39_2020\

Nome	Tamanho
Análisis Matemático	546 768 506
Astronavigation A Method for Determining Exact Position by the Stars	177 028 731
Beadwork - AprilMay 2020	79 432 559
Birational Geometry and Moduli Spaces	4 500 449
Calculus	587 717 592
Classical_Electrodynamics_Jackson	936 988 661
Differential Equation Analysis in Biomedical Science and Engineering	2 977 519
Einstein's Revolution A Study Of Theory Unification	5 726 279
Fermat_Last_Theorem	14 481 722
Fórmulas Elegantes_Farmelo	4 239 419
Handbook of Homotopy Theory	702 696 156
Hardcore Inventing Invent, Protect, Promote, and Profit from Your Ideas	3 543 177
Introduction to Precise Numerical Methods, Second Edition	5 878 179
Loewner's Theorem on Monotone Matrix Functions	5 291 284
Modern Thermodynamics for Chemists and Biochemists	20 149 430
Music Theory and Mathematics Chords, Collections, and Transformations	1 238 584
Number Theory Revealed A Masterclass	9 802 043
Numerical Methods in Matrix Computations (Texts in Applied Mathematics)	10 352 219
Oxford dictionary of quotations	2 570 539
Ray Optics, Fermat's Principle, and Applications to General Relativity	19 631 778
Science - 21 February 2020	64 528 654
Science The Definitive Visual Guide	133 908 700
Sectional Anatomy by MRI and CT, 4th Edition	143 758 330
Solving Problems in Mathematical Analysis	17 947 438
Space, Time, and Spacetime Physical and Philosophical Implications of Minkowski's Unific...	3 117 990
Space-Time, Relativity, and Cosmology	467 451 952
The Colossal Book of Mathematics	39 016 514
The Twin Sister Planets Venus and Earth	27 777 271
Topics in Algebra and Analysis Preparing for the Mathematical Olympiad	7 136 828
What's the Point of Math	20 830 351

Figure 2.50: Books-39-2020-202005121441.png

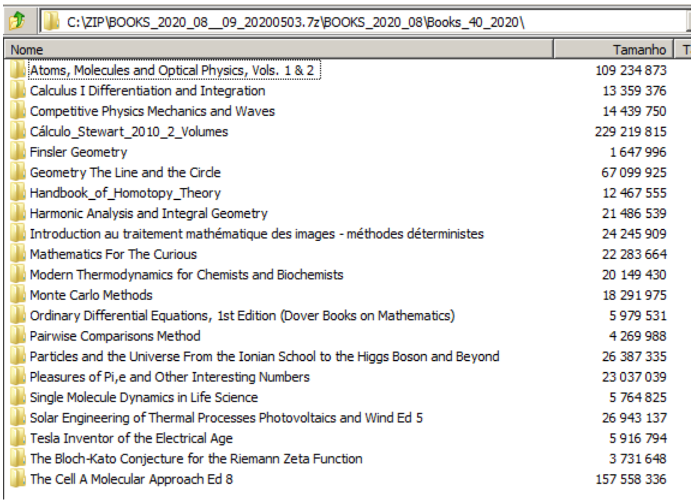

C:\ZIP\BOOKS_2020_08__09_20200503.7z\BOOKS_2020_08\Books_40_2020\

Nome	Tamanho
Atoms, Molecules and Optical Physics, Vols. 1 & 2	109 234 873
Calculus I Differentiation and Integration	13 359 376
Competitive Physics Mechanics and Waves	14 439 750
Cálculo_Stewart_2010_2_Volumes	229 219 815
Finsler Geometry	1 647 996
Geometry The Line and the Circle	67 099 925
Handbook_of_Homotopy_Theory	12 467 555
Harmonic Analysis and Integral Geometry	21 486 539
Introduction au traitement mathématique des images - méthodes déterministes	24 245 909
Mathematics For The Curious	22 283 664
Modern Thermodynamics for Chemists and Biochemists	20 149 430
Monte Carlo Methods	18 291 975
Ordinary Differential Equations, 1st Edition (Dover Books on Mathematics)	5 979 531
Pairwise Comparisons Method	4 269 988
Particles and the Universe From the Ionian School to the Higgs Boson and Beyond	26 387 335
Pleasures of Pi,e and Other Interesting Numbers	23 037 039
Single Molecule Dynamics in Life Science	5 764 825
Solar Engineering of Thermal Processes Photovoltaics and Wind Ed 5	26 943 137
Tesla Inventor of the Electrical Age	5 916 794
The Bloch-Kato Conjecture for the Riemann Zeta Function	3 731 648
The Cell A Molecular Approach Ed 8	157 558 336

Figure 2.51: Books-40-2020-202005121916.png

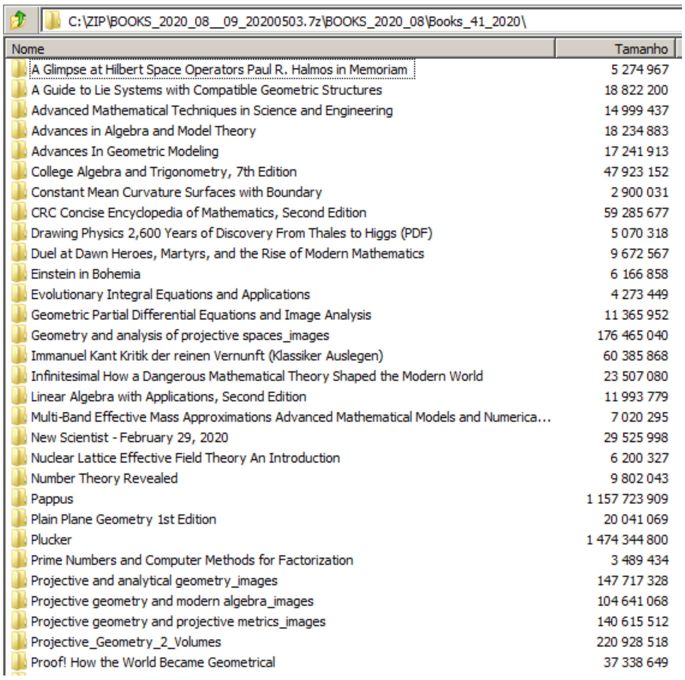

C:\ZIP\BOOKS_2020_08__09_20200503.7z\BOOKS_2020_08\Books_41_2020\

Nome	Tamanho
A Glimpse at Hilbert Space Operators Paul R. Halmos in Memoriam	5 274 967
A Guide to Lie Systems with Compatible Geometric Structures	18 822 200
Advanced Mathematical Techniques in Science and Engineering	14 999 437
Advances in Algebra and Model Theory	18 234 883
Advances In Geometric Modeling	17 241 913
College Algebra and Trigonometry, 7th Edition	47 923 152
Constant Mean Curvature Surfaces with Boundary	2 900 031
CRC Concise Encyclopedia of Mathematics, Second Edition	59 285 677
Drawing Physics 2,600 Years of Discovery From Thales to Higgs (PDF)	5 070 318
Duel at Dawn Heroes, Martyrs, and the Rise of Modern Mathematics	9 672 567
Einstein in Bohemia	6 166 858
Evolutionary Integral Equations and Applications	4 273 449
Geometric Partial Differential Equations and Image Analysis	11 365 952
Geometry and analysis of projective spaces_images	176 465 040
Immanuel Kant Kritik der reinen Vernunft (Klassiker Auslegen)	60 385 868
Infinitesimal How a Dangerous Mathematical Theory Shaped the Modern World	23 507 080
Linear Algebra with Applications, Second Edition	11 993 779
Multi-Band Effective Mass Approximations Advanced Mathematical Models and Numerica...	7 020 295
New Scientist - February 29, 2020	29 525 998
Nuclear Lattice Effective Field Theory An Introduction	6 200 327
Number Theory Revealed	9 802 043
Pappus	1 157 723 909
Plain Plane Geometry 1st Edition	20 041 069
Plucker	1 474 344 800
Prime Numbers and Computer Methods for Factorization	3 489 434
Projective and analytical geometry_images	147 717 328
Projective geometry and modern algebra_images	104 641 068
Projective geometry and projective metrics_images	140 615 512
Projective_Geometry_2_Volumes	220 928 518
Proof! How the World Became Geometrical	37 338 649

Figure 2.52: Books-41-2020-202005121916.png

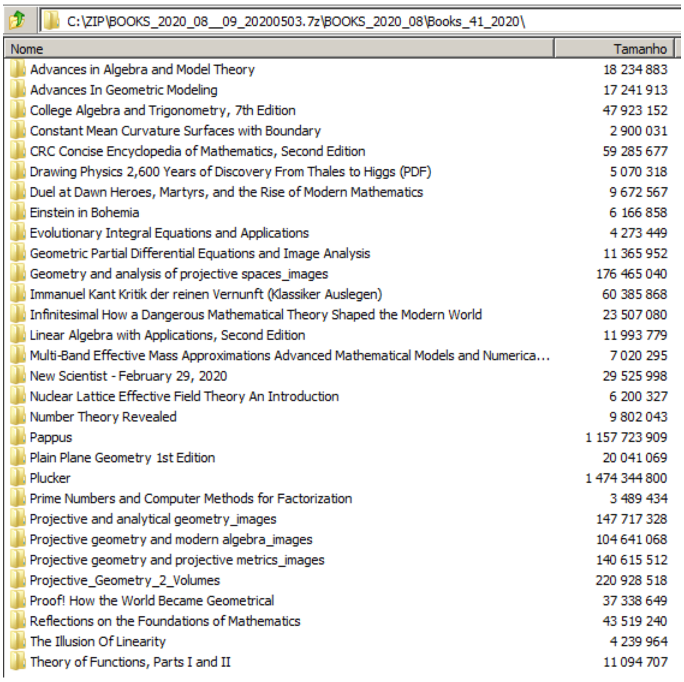

C:\ZIP\BOOKS_2020_08__09_20200503.7z\BOOKS_2020_08\Books_41_2020\

Nome	Tamanho
Advances in Algebra and Model Theory	18 234 883
Advances In Geometric Modeling	17 241 913
College Algebra and Trigonometry, 7th Edition	47 923 152
Constant Mean Curvature Surfaces with Boundary	2 900 031
CRC Concise Encyclopedia of Mathematics, Second Edition	59 285 677
Drawing Physics 2,600 Years of Discovery From Thales to Higgs (PDF)	5 070 318
Duel at Dawn Heroes, Martyrs, and the Rise of Modern Mathematics	9 672 567
Einstein in Bohemia	6 166 858
Evolutionary Integral Equations and Applications	4 273 449
Geometric Partial Differential Equations and Image Analysis	11 365 952
Geometry and analysis of projective spaces_images	176 465 040
Immanuel Kant Kritik der reinen Vernunft (Klassiker Auslegen)	60 385 868
Infinitesimal How a Dangerous Mathematical Theory Shaped the Modern World	23 507 080
Linear Algebra with Applications, Second Edition	11 993 779
Multi-Band Effective Mass Approximations Advanced Mathematical Models and Numerica...	7 020 295
New Scientist - February 29, 2020	29 525 998
Nuclear Lattice Effective Field Theory An Introduction	6 200 327
Number Theory Revealed	9 802 043
Pappus	1 157 723 909
Plain Plane Geometry 1st Edition	20 041 069
Plucker	1 474 344 800
Prime Numbers and Computer Methods for Factorization	3 489 434
Projective and analytical geometry_images	147 717 328
Projective geometry and modern algebra_images	104 641 068
Projective geometry and projective metrics_images	140 615 512
Projective_Geometry_2_Volumes	220 928 518
Proof! How the World Became Geometrical	37 338 649
Reflections on the Foundations of Mathematics	43 519 240
The Illusion Of Linearity	4 239 964
Theory of Functions, Parts I and II	11 094 707

Figure 2.53: Books-41-2020-202005121917.png

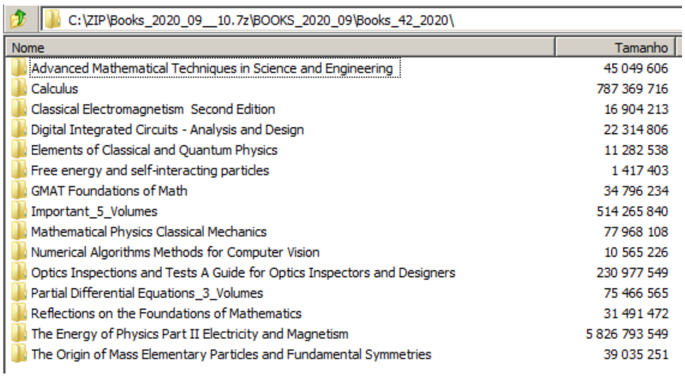

C:\ZIP\Books_2020_09__10.7z\BOOKS_2020_09\Books_42_2020\

Nome	Tamanho
Advanced Mathematical Techniques in Science and Engineering	45 049 606
Calculus	787 369 716
Classical Electromagnetism Second Edition	16 904 213
Digital Integrated Circuits - Analysis and Design	22 314 806
Elements of Classical and Quantum Physics	11 282 538
Free energy and self-interacting particles	1 417 403
GMAT Foundations of Math	34 796 234
Important_5_Volumes	514 265 840
Mathematical Physics Classical Mechanics	77 968 108
Numerical Algorithms Methods for Computer Vision	10 565 226
Optics Inspections and Tests A Guide for Optics Inspectors and Designers	230 977 549
Partial Differential Equations_3_Volumes	75 466 565
Reflections on the Foundations of Mathematics	31 491 472
The Energy of Physics Part II Electricity and Magnetism	5 826 793 549
The Origin of Mass Elementary Particles and Fundamental Symmetries	39 035 251

Figure 2.54: Books-42-2020-202005121522.png

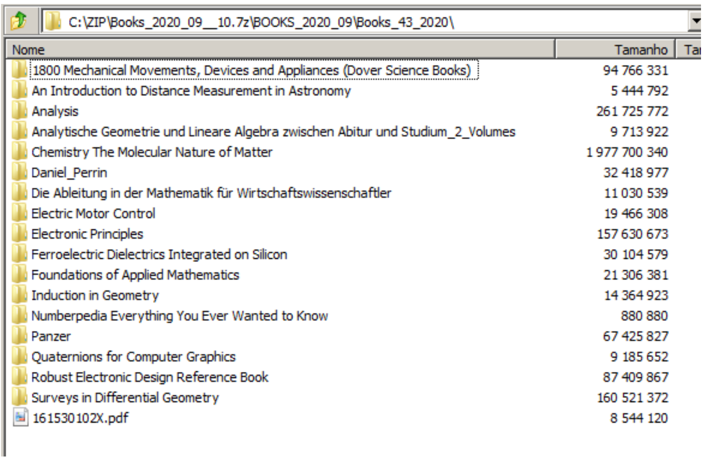

C:\ZIP\Books_2020_09__10.7z\BOOKS_2020_09\Books_43_2020\

Nome	Tamanho	Ta
1800 Mechanical Movements, Devices and Appliances (Dover Science Books)	94 766 331	
An Introduction to Distance Measurement in Astronomy	5 444 792	
Analysis	261 725 772	
Analytische Geometrie und Lineare Algebra zwischen Abitur und Studium_2_Volumes	9 713 922	
Chemistry The Molecular Nature of Matter	1 977 700 340	
Daniel_Perrin	32 418 977	
Die Ableitung in der Mathematik für Wirtschaftswissenschaftler	11 030 539	
Electric Motor Control	19 466 308	
Electronic Principles	157 630 673	
Ferroelectric Dielectrics Integrated on Silicon	30 104 579	
Foundations of Applied Mathematics	21 306 381	
Induction in Geometry	14 364 923	
Numberpedia Everything You Ever Wanted to Know	880 880	
Panzer	67 425 827	
Quaternions for Computer Graphics	9 185 652	
Robust Electronic Design Reference Book	87 409 867	
Surveys in Differential Geometry	160 521 372	
161530102X.pdf	8 544 120	

Figure 2.55: Books-43-2020-202005121522.png

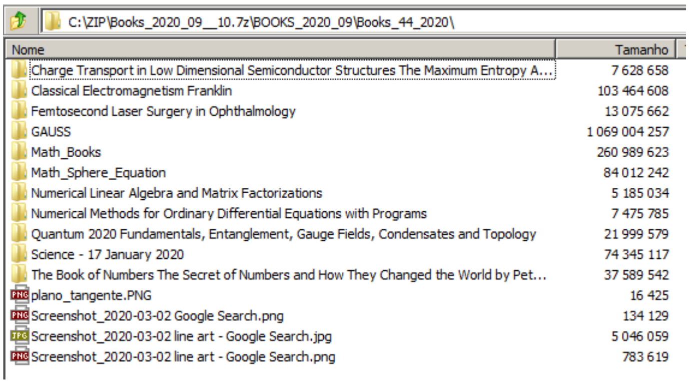
C:\ZIP\Books_2020_09__10.7z\BOOKS_2020_09\Books_44_2020\

Nome	Tamanho
Charge Transport in Low Dimensional Semiconductor Structures The Maximum Entropy A...	7 628 658
Classical Electromagnetism Franklin	103 464 608
Femtosecond Laser Surgery in Ophthalmology	13 075 662
GAUSS	1 069 004 257
Math_Books	260 989 623
Math_Sphere_Equation	84 012 242
Numerical Linear Algebra and Matrix Factorizations	5 185 034
Numerical Methods for Ordinary Differential Equations with Programs	7 475 785
Quantum 2020 Fundamentals, Entanglement, Gauge Fields, Condensates and Topology	21 999 579
Science - 17 January 2020	74 345 117
The Book of Numbers The Secret of Numbers and How They Changed the World by Pet...	37 589 542
plano_tangente.PNG	16 425
Screenshot_2020-03-02 Google Search.png	134 129
Screenshot_2020-03-02 line art - Google Search.jpg	5 046 059
Screenshot_2020-03-02 line art - Google Search.png	783 619

Figure 2.56: Books-44-2020-202005121522.png

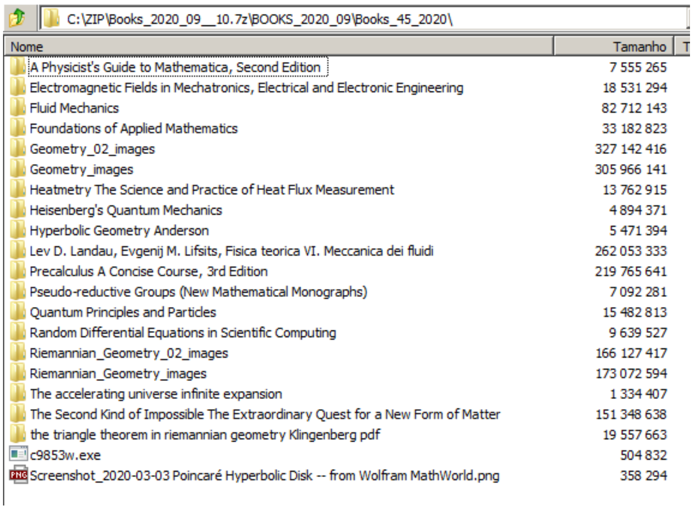
C:\ZIP\Books_2020_09__10.7z\BOOKS_2020_09\Books_45_2020\

Nome	Tamanho
A Physicist's Guide to Mathematica, Second Edition	7 555 265
Electromagnetic Fields in Mechatronics, Electrical and Electronic Engineering	18 531 294
Fluid Mechanics	82 712 143
Foundations of Applied Mathematics	33 182 823
Geometry_02_images	327 142 416
Geometry_images	305 966 141
Heatmetry The Science and Practice of Heat Flux Measurement	13 762 915
Heisenberg's Quantum Mechanics	4 894 371
Hyperbolic Geometry Anderson	5 471 394
Lev D. Landau, Evgenij M. Lifsits, Fisica teorica VI. Meccanica dei fluidi	262 053 333
Precalculus A Concise Course, 3rd Edition	219 765 641
Pseudo-reductive Groups (New Mathematical Monographs)	7 092 281
Quantum Principles and Particles	15 482 813
Random Differential Equations in Scientific Computing	9 639 527
Riemannian_Geometry_02_images	166 127 417
Riemannian_Geometry_images	173 072 594
The accelerating universe infinite expansion	1 334 407
The Second Kind of Impossible The Extraordinary Quest for a New Form of Matter	151 348 638
the triangle theorem in riemannian geometry Klingenberg pdf	19 557 663
c9853w.exe	504 832
Screenshot_2020-03-03 Poincaré Hyperbolic Disk -- from Wolfram MathWorld.png	358 294

Figure 2.57: Books-45-2020-202005121522.png

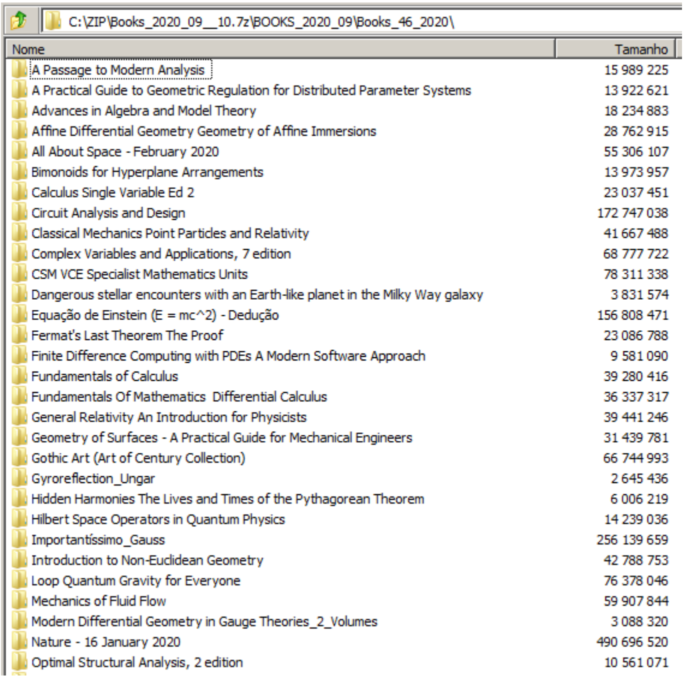

C:\ZIP\Books_2020_09__10.7z\BOOKS_2020_09\Books_46_2020\

Nome	Tamanho
A Passage to Modern Analysis	15 989 225
A Practical Guide to Geometric Regulation for Distributed Parameter Systems	13 922 621
Advances in Algebra and Model Theory	18 234 883
Affine Differential Geometry Geometry of Affine Immersions	28 762 915
All About Space - February 2020	55 306 107
Bimonoids for Hyperplane Arrangements	13 973 957
Calculus Single Variable Ed 2	23 037 451
Circuit Analysis and Design	172 747 038
Classical Mechanics Point Particles and Relativity	41 667 488
Complex Variables and Applications, 7 edition	68 777 722
CSM VCE Specialist Mathematics Units	78 311 338
Dangerous stellar encounters with an Earth-like planet in the Milky Way galaxy	3 831 574
Equação de Einstein (E = mc^2) - Dedução	156 808 471
Fermat's Last Theorem The Proof	23 086 788
Finite Difference Computing with PDEs A Modern Software Approach	9 581 090
Fundamentals of Calculus	39 280 416
Fundamentals Of Mathematics Differential Calculus	36 337 317
General Relativity An Introduction for Physicists	39 441 246
Geometry of Surfaces - A Practical Guide for Mechanical Engineers	31 439 781
Gothic Art (Art of Century Collection)	66 744 993
Gyroreflection_Ungar	2 645 436
Hidden Harmonies The Lives and Times of the Pythagorean Theorem	6 006 219
Hilbert Space Operators in Quantum Physics	14 239 036
Importantíssimo_Gauss	256 139 659
Introduction to Non-Euclidean Geometry	42 788 753
Loop Quantum Gravity for Everyone	76 378 046
Mechanics of Fluid Flow	59 907 844
Modern Differential Geometry in Gauge Theories_2_Volumes	3 088 320
Nature - 16 January 2020	490 696 520
Optimal Structural Analysis, 2 edition	10 561 071

Figure 2.58: Books-46-2020-202005121523.png

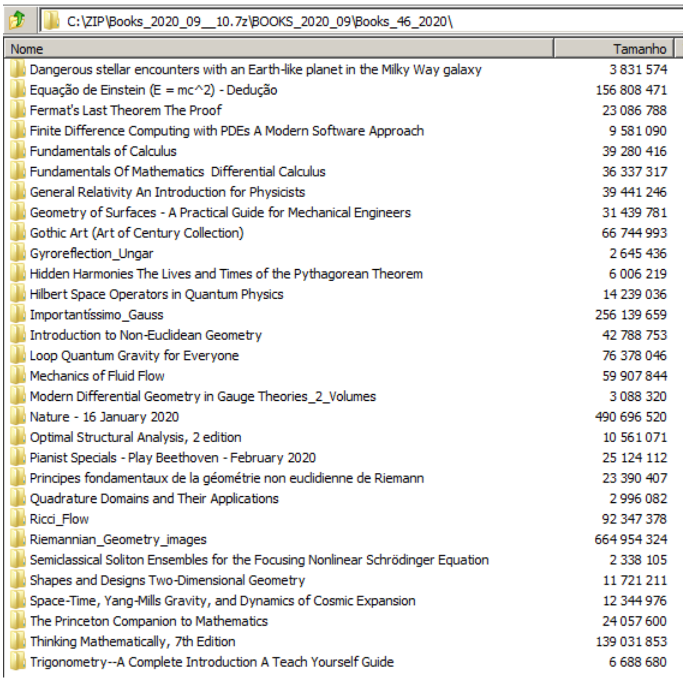

C:\ZIP\Books_2020_09__10.7z\BOOKS_2020_09\Books_46_2020\

Nome	Tamanho
Dangerous stellar encounters with an Earth-like planet in the Milky Way galaxy	3 831 574
Equação de Einstein (E = mc^2) - Dedução	156 808 471
Fermat's Last Theorem The Proof	23 086 788
Finite Difference Computing with PDEs A Modern Software Approach	9 581 090
Fundamentals of Calculus	39 280 416
Fundamentals Of Mathematics Differential Calculus	36 337 317
General Relativity An Introduction for Physicists	39 441 246
Geometry of Surfaces - A Practical Guide for Mechanical Engineers	31 439 781
Gothic Art (Art of Century Collection)	66 744 993
Gyroreflection_Ungar	2 645 436
Hidden Harmonies The Lives and Times of the Pythagorean Theorem	6 006 219
Hilbert Space Operators in Quantum Physics	14 239 036
Importantíssimo_Gauss	256 139 659
Introduction to Non-Euclidean Geometry	42 788 753
Loop Quantum Gravity for Everyone	76 378 046
Mechanics of Fluid Flow	59 907 844
Modern Differential Geometry in Gauge Theories_2_Volumes	3 088 320
Nature - 16 January 2020	490 696 520
Optimal Structural Analysis, 2 edition	10 561 071
Pianist Specials - Play Beethoven - February 2020	25 124 112
Principes fondamentaux de la géométrie non euclidienne de Riemann	23 390 407
Quadrature Domains and Their Applications	2 996 082
Ricci_Flow	92 347 378
Riemannian_Geometry_images	664 954 324
Semiclassical Soliton Ensembles for the Focusing Nonlinear Schrödinger Equation	2 338 105
Shapes and Designs Two-Dimensional Geometry	11 721 211
Space-Time, Yang-Mills Gravity, and Dynamics of Cosmic Expansion	12 344 976
The Princeton Companion to Mathematics	24 057 600
Thinking Mathematically, 7th Edition	139 031 853
Trigonometry--A Complete Introduction A Teach Yourself Guide	6 688 680

Figure 2.59: Books-46-2020-202005121524.png

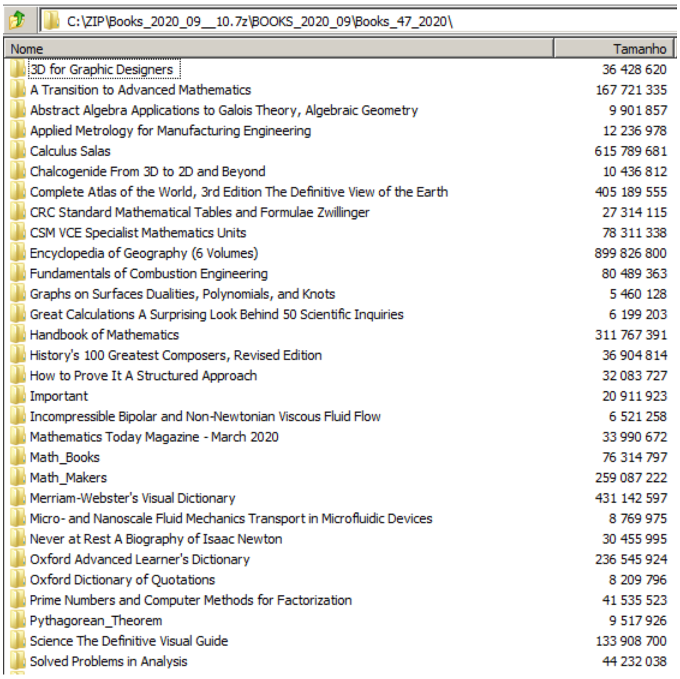
C:\ZIP\Books_2020_09__10.7z\BOOKS_2020_09\Books_47_2020\

Nome	Tamanho
3D for Graphic Designers	36 428 620
A Transition to Advanced Mathematics	167 721 335
Abstract Algebra Applications to Galois Theory, Algebraic Geometry	9 901 857
Applied Metrology for Manufacturing Engineering	12 236 978
Calculus Salas	615 789 681
Chalcogenide From 3D to 2D and Beyond	10 436 812
Complete Atlas of the World, 3rd Edition The Definitive View of the Earth	405 189 555
CRC Standard Mathematical Tables and Formulae Zwillinger	27 314 115
CSM VCE Specialist Mathematics Units	78 311 338
Encyclopedia of Geography (6 Volumes)	899 826 800
Fundamentals of Combustion Engineering	80 489 363
Graphs on Surfaces Dualities, Polynomials, and Knots	5 460 128
Great Calculations A Surprising Look Behind 50 Scientific Inquiries	6 199 203
Handbook of Mathematics	311 767 391
History's 100 Greatest Composers, Revised Edition	36 904 814
How to Prove It A Structured Approach	32 083 727
Important	20 911 923
Incompressible Bipolar and Non-Newtonian Viscous Fluid Flow	6 521 258
Mathematics Today Magazine - March 2020	33 990 672
Math_Books	76 314 797
Math_Makers	259 087 222
Merriam-Webster's Visual Dictionary	431 142 597
Micro- and Nanoscale Fluid Mechanics Transport in Microfluidic Devices	8 769 975
Never at Rest A Biography of Isaac Newton	30 455 995
Oxford Advanced Learner's Dictionary	236 545 924
Oxford Dictionary of Quotations	8 209 796
Prime Numbers and Computer Methods for Factorization	41 535 523
Pythagorean_Theorem	9 517 926
Science The Definitive Visual Guide	133 908 700
Solved Problems in Analysis	44 232 038

Figure 2.60: Books-47-2020-202005121524.png

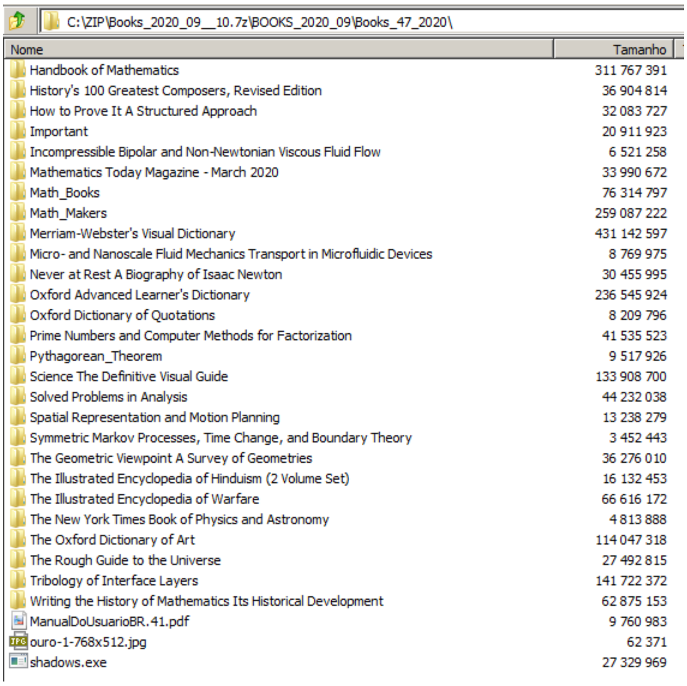
C:\ZIP\Books_2020_09__10.7z\BOOKS_2020_09\Books_47_2020\

Nome	Tamanho
Handbook of Mathematics	311 767 391
History's 100 Greatest Composers, Revised Edition	36 904 814
How to Prove It A Structured Approach	32 083 727
Important	20 911 923
Incompressible Bipolar and Non-Newtonian Viscous Fluid Flow	6 521 258
Mathematics Today Magazine - March 2020	33 990 672
Math_Books	76 314 797
Math_Makers	259 087 222
Merriam-Webster's Visual Dictionary	431 142 597
Micro- and Nanoscale Fluid Mechanics Transport in Microfluidic Devices	8 769 975
Never at Rest A Biography of Isaac Newton	30 455 995
Oxford Advanced Learner's Dictionary	236 545 924
Oxford Dictionary of Quotations	8 209 796
Prime Numbers and Computer Methods for Factorization	41 535 523
Pythagorean_Theorem	9 517 926
Science The Definitive Visual Guide	133 908 700
Solved Problems in Analysis	44 232 038
Spatial Representation and Motion Planning	13 238 279
Symmetric Markov Processes, Time Change, and Boundary Theory	3 452 443
The Geometric Viewpoint A Survey of Geometries	36 276 010
The Illustrated Encyclopedia of Hinduism (2 Volume Set)	16 132 453
The Illustrated Encyclopedia of Warfare	66 616 172
The New York Times Book of Physics and Astronomy	4 813 888
The Oxford Dictionary of Art	114 047 318
The Rough Guide to the Universe	27 492 815
Tribology of Interface Layers	141 722 372
Writing the History of Mathematics Its Historical Development	62 875 153
ManualDoUsuarioBR.41.pdf	9 760 983
ouro-1-768x512.jpg	62 371
shadows.exe	27 329 969

Figure 2.61: Books-47-2020-202005121525.png

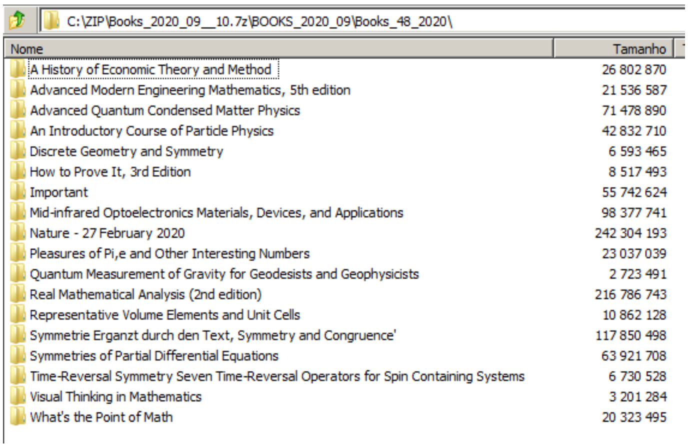

C:\ZIP\Books_2020_09__10.7z\BOOKS_2020_09\Books_48_2020\

Nome	Tamanho
A History of Economic Theory and Method	26 802 870
Advanced Modern Engineering Mathematics, 5th edition	21 536 587
Advanced Quantum Condensed Matter Physics	71 478 890
An Introductory Course of Particle Physics	42 832 710
Discrete Geometry and Symmetry	6 593 465
How to Prove It, 3rd Edition	8 517 493
Important	55 742 624
Mid-infrared Optoelectronics Materials, Devices, and Applications	98 377 741
Nature - 27 February 2020	242 304 193
Pleasures of Pi,e and Other Interesting Numbers	23 037 039
Quantum Measurement of Gravity for Geodesists and Geophysicists	2 723 491
Real Mathematical Analysis (2nd edition)	216 786 743
Representative Volume Elements and Unit Cells	10 862 128
Symmetrie Erganzt durch den Text, Symmetry and Congruence'	117 850 498
Symmetries of Partial Differential Equations	63 921 708
Time-Reversal Symmetry Seven Time-Reversal Operators for Spin Containing Systems	6 730 528
Visual Thinking in Mathematics	3 201 284
What's the Point of Math	20 323 495

Figure 2.62: Books-48-2020-202005121525.png

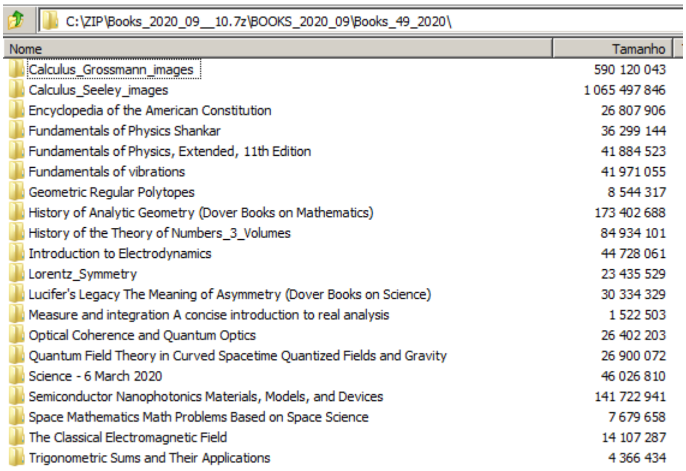

C:\ZIP\Books_2020_09__10.7z\BOOKS_2020_09\Books_49_2020\

Nome	Tamanho
Calculus_Grossmann_images	590 120 043
Calculus_Seeley_images	1 065 497 846
Encyclopedia of the American Constitution	26 807 906
Fundamentals of Physics Shankar	36 299 144
Fundamentals of Physics, Extended, 11th Edition	41 884 523
Fundamentals of vibrations	41 971 055
Geometric Regular Polytopes	8 544 317
History of Analytic Geometry (Dover Books on Mathematics)	173 402 688
History of the Theory of Numbers_3_Volumes	84 934 101
Introduction to Electrodynamics	44 728 061
Lorentz_Symmetry	23 435 529
Lucifer's Legacy The Meaning of Asymmetry (Dover Books on Science)	30 334 329
Measure and integration A concise introduction to real analysis	1 522 503
Optical Coherence and Quantum Optics	26 402 203
Quantum Field Theory in Curved Spacetime Quantized Fields and Gravity	26 900 072
Science - 6 March 2020	46 026 810
Semiconductor Nanophotonics Materials, Models, and Devices	141 722 941
Space Mathematics Math Problems Based on Space Science	7 679 658
The Classical Electromagnetic Field	14 107 287
Trigonometric Sums and Their Applications	4 366 434

Figure 2.63: Books-49-2020-202005121525.png

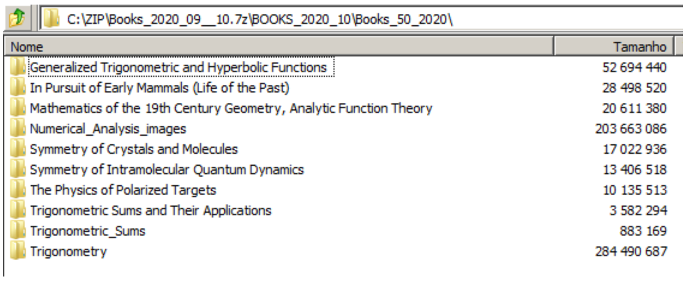

C:\ZIP\Books_2020_09__10.7z\BOOKS_2020_10\Books_50_2020\

Nome	Tamanho
Generalized Trigonometric and Hyperbolic Functions	52 694 440
In Pursuit of Early Mammals (Life of the Past)	28 498 520
Mathematics of the 19th Century Geometry, Analytic Function Theory	20 611 380
Numerical_Analysis_images	203 663 086
Symmetry of Crystals and Molecules	17 022 936
Symmetry of Intramolecular Quantum Dynamics	13 406 518
The Physics of Polarized Targets	10 135 513
Trigonometric Sums and Their Applications	3 582 294
Trigonometric_Sums	883 169
Trigonometry	284 490 687

Figure 2.64: Books-50-2020-202005121525.png

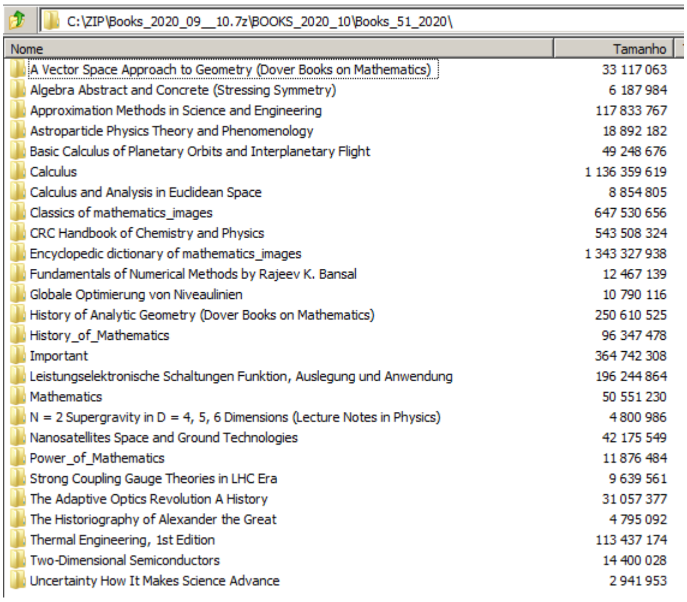

C:\ZIP\Books_2020_09__10.7z\BOOKS_2020_10\Books_51_2020\

Nome	Tamanho
A Vector Space Approach to Geometry (Dover Books on Mathematics)	33 117 063
Algebra Abstract and Concrete (Stressing Symmetry)	6 187 984
Approximation Methods in Science and Engineering	117 833 767
Astroparticle Physics Theory and Phenomenology	18 892 182
Basic Calculus of Planetary Orbits and Interplanetary Flight	49 248 676
Calculus	1 136 359 619
Calculus and Analysis in Euclidean Space	8 854 805
Classics of mathematics_images	647 530 656
CRC Handbook of Chemistry and Physics	543 508 324
Encyclopedic dictionary of mathematics_images	1 343 327 938
Fundamentals of Numerical Methods by Rajeev K. Bansal	12 467 139
Globale Optimierung von Niveaulinien	10 790 116
History of Analytic Geometry (Dover Books on Mathematics)	250 610 525
History_of_Mathematics	96 347 478
Important	364 742 308
Leistungselektronische Schaltungen Funktion, Auslegung und Anwendung	196 244 864
Mathematics	50 551 230
N = 2 Supergravity in D = 4, 5, 6 Dimensions (Lecture Notes in Physics)	4 800 986
Nanosatellites Space and Ground Technologies	42 175 549
Power_of_Mathematics	11 876 484
Strong Coupling Gauge Theories in LHC Era	9 639 561
The Adaptive Optics Revolution A History	31 057 377
The Historiography of Alexander the Great	4 795 092
Thermal Engineering, 1st Edition	113 437 174
Two-Dimensional Semiconductors	14 400 028
Uncertainty How It Makes Science Advance	2 941 953

Figure 2.65: Books-51-2020-202005121525.png

2.2 Indice of Books - Cont. 02

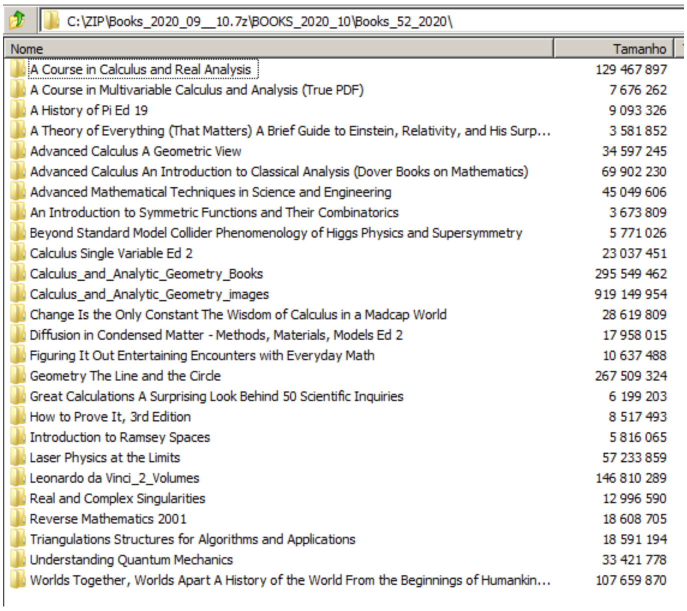

C:\ZIP\Books_2020_09__10.7z\BOOKS_2020_10\Books_52_2020\

Nome	Tamanho
A Course in Calculus and Real Analysis	129 467 897
A Course in Multivariable Calculus and Analysis (True PDF)	7 676 262
A History of Pi Ed 19	9 093 326
A Theory of Everything (That Matters) A Brief Guide to Einstein, Relativity, and His Surp...	3 581 852
Advanced Calculus A Geometric View	34 597 245
Advanced Calculus An Introduction to Classical Analysis (Dover Books on Mathematics)	69 902 230
Advanced Mathematical Techniques in Science and Engineering	45 049 606
An Introduction to Symmetric Functions and Their Combinatorics	3 673 809
Beyond Standard Model Collider Phenomenology of Higgs Physics and Supersymmetry	5 771 026
Calculus Single Variable Ed 2	23 037 451
Calculus_and_Analytic_Geometry_Books	295 549 462
Calculus_and_Analytic_Geometry_images	919 149 954
Change Is the Only Constant The Wisdom of Calculus in a Madcap World	28 619 809
Diffusion in Condensed Matter - Methods, Materials, Models Ed 2	17 958 015
Figuring It Out Entertaining Encounters with Everyday Math	10 637 488
Geometry The Line and the Circle	267 509 324
Great Calculations A Surprising Look Behind 50 Scientific Inquiries	6 199 203
How to Prove It, 3rd Edition	8 517 493
Introduction to Ramsey Spaces	5 816 065
Laser Physics at the Limits	57 233 859
Leonardo da Vinci_2_Volumes	146 810 289
Real and Complex Singularities	12 996 590
Reverse Mathematics 2001	18 608 705
Triangulations Structures for Algorithms and Applications	18 591 194
Understanding Quantum Mechanics	33 421 778
Worlds Together, Worlds Apart A History of the World From the Beginnings of Humankin...	107 659 870

Figure 2.66: Books-52-2020-202005121525.png

2.3 Indice of Book - cont. 02

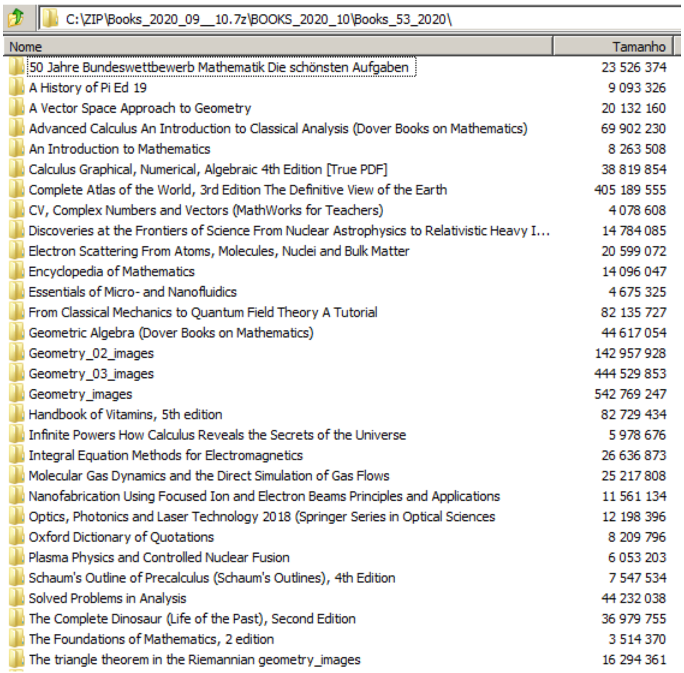

C:\ZIP\Books_2020_09__10.7z\BOOKS_2020_10\Books_53_2020\

Nome	Tamanho
50 Jahre Bundeswettbewerb Mathematik Die schönsten Aufgaben	23 526 374
A History of Pi Ed 19	9 093 326
A Vector Space Approach to Geometry	20 132 160
Advanced Calculus An Introduction to Classical Analysis (Dover Books on Mathematics)	69 902 230
An Introduction to Mathematics	8 263 508
Calculus Graphical, Numerical, Algebraic 4th Edition [True PDF]	38 819 854
Complete Atlas of the World, 3rd Edition The Definitive View of the Earth	405 189 555
CV, Complex Numbers and Vectors (MathWorks for Teachers)	4 078 608
Discoveries at the Frontiers of Science From Nuclear Astrophysics to Relativistic Heavy I...	14 784 085
Electron Scattering From Atoms, Molecules, Nuclei and Bulk Matter	20 599 072
Encyclopedia of Mathematics	14 096 047
Essentials of Micro- and Nanofluidics	4 675 325
From Classical Mechanics to Quantum Field Theory A Tutorial	82 135 727
Geometric Algebra (Dover Books on Mathematics)	44 617 054
Geometry_02_images	142 957 928
Geometry_03_images	444 529 853
Geometry_images	542 769 247
Handbook of Vitamins, 5th edition	82 729 434
Infinite Powers How Calculus Reveals the Secrets of the Universe	5 978 676
Integral Equation Methods for Electromagnetics	26 636 873
Molecular Gas Dynamics and the Direct Simulation of Gas Flows	25 217 808
Nanofabrication Using Focused Ion and Electron Beams Principles and Applications	11 561 134
Optics, Photonics and Laser Technology 2018 (Springer Series in Optical Sciences	12 198 396
Oxford Dictionary of Quotations	8 209 796
Plasma Physics and Controlled Nuclear Fusion	6 053 203
Schaum's Outline of Precalculus (Schaum's Outlines), 4th Edition	7 547 534
Solved Problems in Analysis	44 232 038
The Complete Dinosaur (Life of the Past), Second Edition	36 979 755
The Foundations of Mathematics, 2 edition	3 514 370
The triangle theorem in the Riemannian geometry_images	16 294 361

Figure 2.67: Books-53-2020-202005121525.png

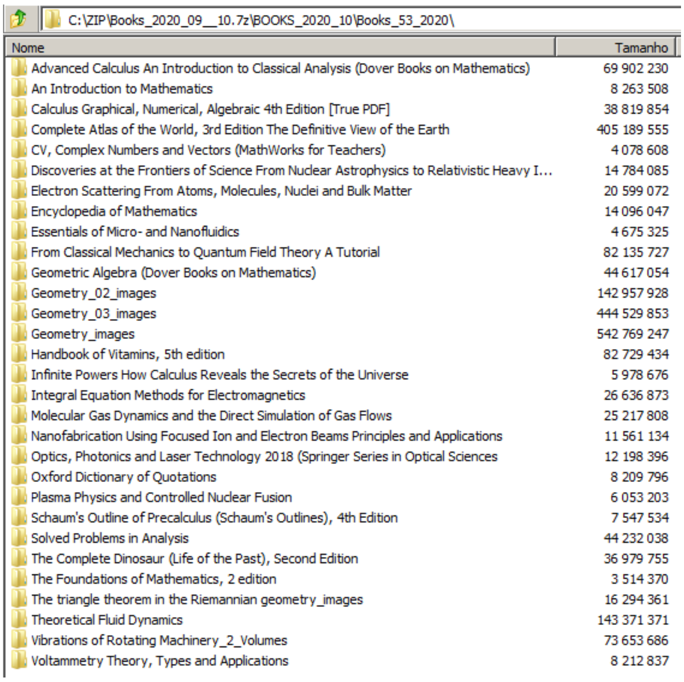

C:\ZIP\Books_2020_09__10.7z\BOOKS_2020_10\Books_53_2020\

Nome	Tamanho
Advanced Calculus An Introduction to Classical Analysis (Dover Books on Mathematics)	69 902 230
An Introduction to Mathematics	8 263 508
Calculus Graphical, Numerical, Algebraic 4th Edition [True PDF]	38 819 854
Complete Atlas of the World, 3rd Edition The Definitive View of the Earth	405 189 555
CV, Complex Numbers and Vectors (MathWorks for Teachers)	4 078 608
Discoveries at the Frontiers of Science From Nuclear Astrophysics to Relativistic Heavy I...	14 784 085
Electron Scattering From Atoms, Molecules, Nuclei and Bulk Matter	20 599 072
Encyclopedia of Mathematics	14 096 047
Essentials of Micro- and Nanofluidics	4 675 325
From Classical Mechanics to Quantum Field Theory A Tutorial	82 135 727
Geometric Algebra (Dover Books on Mathematics)	44 617 054
Geometry_02_images	142 957 928
Geometry_03_images	444 529 853
Geometry_images	542 769 247
Handbook of Vitamins, 5th edition	82 729 434
Infinite Powers How Calculus Reveals the Secrets of the Universe	5 978 676
Integral Equation Methods for Electromagnetics	26 636 873
Molecular Gas Dynamics and the Direct Simulation of Gas Flows	25 217 808
Nanofabrication Using Focused Ion and Electron Beams Principles and Applications	11 561 134
Optics, Photonics and Laser Technology 2018 (Springer Series in Optical Sciences	12 198 396
Oxford Dictionary of Quotations	8 209 796
Plasma Physics and Controlled Nuclear Fusion	6 053 203
Schaum's Outline of Precalculus (Schaum's Outlines), 4th Edition	7 547 534
Solved Problems in Analysis	44 232 038
The Complete Dinosaur (Life of the Past), Second Edition	36 979 755
The Foundations of Mathematics, 2 edition	3 514 370
The triangle theorem in the Riemannian geometry_images	16 294 361
Theoretical Fluid Dynamics	143 371 371
Vibrations of Rotating Machinery_2_Volumes	73 653 686
Voltammetry Theory, Types and Applications	8 212 837

Figure 2.68: Books-53-2020-202005121526.png

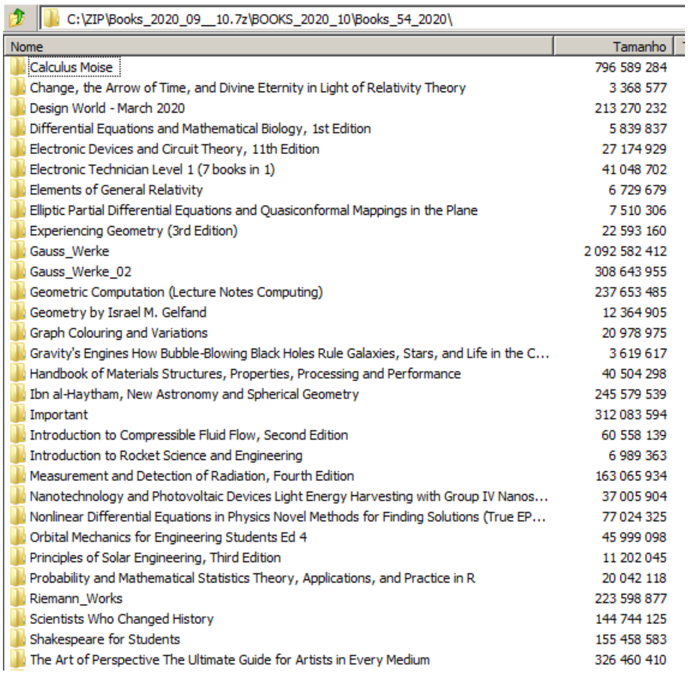

C:\ZIP\Books_2020_09__10.7z\BOOKS_2020_10\Books_54_2020\

Nome	Tamanho
Calculus Moise	796 589 284
Change, the Arrow of Time, and Divine Eternity in Light of Relativity Theory	3 368 577
Design World - March 2020	213 270 232
Differential Equations and Mathematical Biology, 1st Edition	5 839 837
Electronic Devices and Circuit Theory, 11th Edition	27 174 929
Electronic Technician Level 1 (7 books in 1)	41 048 702
Elements of General Relativity	6 729 679
Elliptic Partial Differential Equations and Quasiconformal Mappings in the Plane	7 510 306
Experiencing Geometry (3rd Edition)	22 593 160
Gauss_Werke	2 092 582 412
Gauss_Werke_02	308 643 955
Geometric Computation (Lecture Notes Computing)	237 653 485
Geometry by Israel M. Gelfand	12 364 905
Graph Colouring and Variations	20 978 975
Gravity's Engines How Bubble-Blowing Black Holes Rule Galaxies, Stars, and Life in the C...	3 619 617
Handbook of Materials Structures, Properties, Processing and Performance	40 504 298
Ibn al-Haytham, New Astronomy and Spherical Geometry	245 579 539
Important	312 083 594
Introduction to Compressible Fluid Flow, Second Edition	60 558 139
Introduction to Rocket Science and Engineering	6 989 363
Measurement and Detection of Radiation, Fourth Edition	163 065 934
Nanotechnology and Photovoltaic Devices Light Energy Harvesting with Group IV Nanos...	37 005 904
Nonlinear Differential Equations in Physics Novel Methods for Finding Solutions (True EP...	77 024 325
Orbital Mechanics for Engineering Students Ed 4	45 999 098
Principles of Solar Engineering, Third Edition	11 202 045
Probability and Mathematical Statistics Theory, Applications, and Practice in R	20 042 118
Riemann_Works	223 598 877
Scientists Who Changed History	144 744 125
Shakespeare for Students	155 458 583
The Art of Perspective The Ultimate Guide for Artists in Every Medium	326 460 410

Figure 2.69: Books-54-2020-202005121526.png

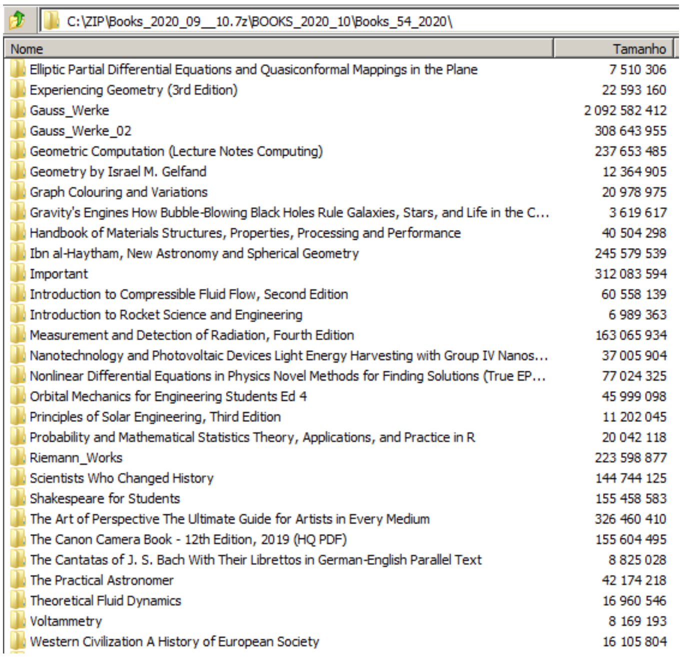

C:\ZIP\Books_2020_09__10.7z\BOOKS_2020_10\Books_54_2020\

Nome	Tamanho
Elliptic Partial Differential Equations and Quasiconformal Mappings in the Plane	7 510 306
Experiencing Geometry (3rd Edition)	22 593 160
Gauss_Werke	2 092 582 412
Gauss_Werke_02	308 643 955
Geometric Computation (Lecture Notes Computing)	237 653 485
Geometry by Israel M. Gelfand	12 364 905
Graph Colouring and Variations	20 978 975
Gravity's Engines How Bubble-Blowing Black Holes Rule Galaxies, Stars, and Life in the C...	3 619 617
Handbook of Materials Structures, Properties, Processing and Performance	40 504 298
Ibn al-Haytham, New Astronomy and Spherical Geometry	245 579 539
Important	312 083 594
Introduction to Compressible Fluid Flow, Second Edition	60 558 139
Introduction to Rocket Science and Engineering	6 989 363
Measurement and Detection of Radiation, Fourth Edition	163 065 934
Nanotechnology and Photovoltaic Devices Light Energy Harvesting with Group IV Nanos...	37 005 904
Nonlinear Differential Equations in Physics Novel Methods for Finding Solutions (True EP...	77 024 325
Orbital Mechanics for Engineering Students Ed 4	45 999 098
Principles of Solar Engineering, Third Edition	11 202 045
Probability and Mathematical Statistics Theory, Applications, and Practice in R	20 042 118
Riemann_Works	223 598 877
Scientists Who Changed History	144 744 125
Shakespeare for Students	155 458 583
The Art of Perspective The Ultimate Guide for Artists in Every Medium	326 460 410
The Canon Camera Book - 12th Edition, 2019 (HQ PDF)	155 604 495
The Cantatas of J. S. Bach With Their Librettos in German-English Parallel Text	8 825 028
The Practical Astronomer	42 174 218
Theoretical Fluid Dynamics	16 960 546
Voltammetry	8 169 193
Western Civilization A History of European Society	16 105 804

Figure 2.70: Books-54-2020-202005121527.png

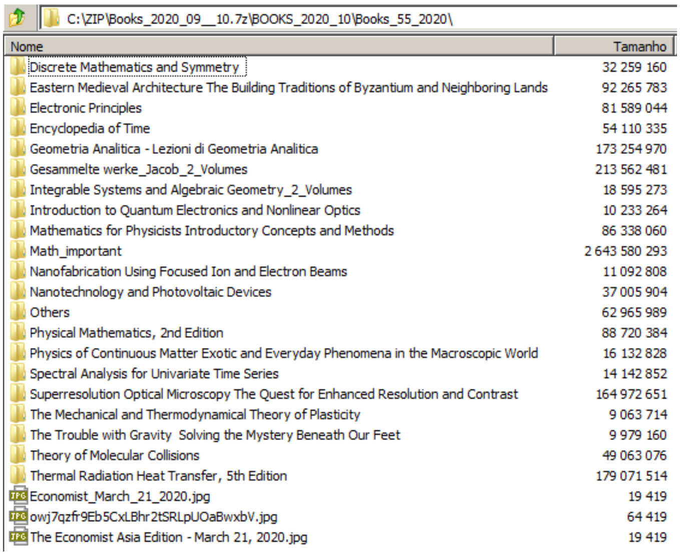

C:\ZIP\Books_2020_09__10.7z\BOOKS_2020_10\Books_55_2020\

Nome	Tamanho
Discrete Mathematics and Symmetry	32 259 160
Eastern Medieval Architecture The Building Traditions of Byzantium and Neighboring Lands	92 265 783
Electronic Principles	81 589 044
Encyclopedia of Time	54 110 335
Geometria Analitica - Lezioni di Geometria Analitica	173 254 970
Gesammelte werke_Jacob_2_Volumes	213 562 481
Integrable Systems and Algebraic Geometry_2_Volumes	18 595 273
Introduction to Quantum Electronics and Nonlinear Optics	10 233 264
Mathematics for Physicists Introductory Concepts and Methods	86 338 060
Math_important	2 643 580 293
Nanofabrication Using Focused Ion and Electron Beams	11 092 808
Nanotechnology and Photovoltaic Devices	37 005 904
Others	62 965 989
Physical Mathematics, 2nd Edition	88 720 384
Physics of Continuous Matter Exotic and Everyday Phenomena in the Macroscopic World	16 132 828
Spectral Analysis for Univariate Time Series	14 142 852
Superresolution Optical Microscopy The Quest for Enhanced Resolution and Contrast	164 972 651
The Mechanical and Thermodynamical Theory of Plasticity	9 063 714
The Trouble with Gravity Solving the Mystery Beneath Our Feet	9 979 160
Theory of Molecular Collisions	49 063 076
Thermal Radiation Heat Transfer, 5th Edition	179 071 514
Economist_March_21_2020.jpg	19 419
owj7qzfr9Eb5CxLBhr2tSRLpUOaBwxbV.jpg	64 419
The Economist Asia Edition - March 21, 2020.jpg	19 419

Figure 2.71: Books-55-2020-202005121527.png

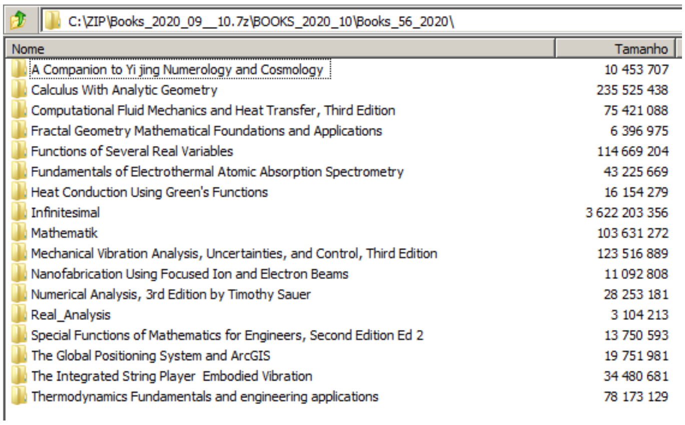

C:\ZIP\Books_2020_09__10.7z\BOOKS_2020_10\Books_56_2020\

Nome	Tamanho
A Companion to Yi jing Numerology and Cosmology	10 453 707
Calculus With Analytic Geometry	235 525 438
Computational Fluid Mechanics and Heat Transfer, Third Edition	75 421 088
Fractal Geometry Mathematical Foundations and Applications	6 396 975
Functions of Several Real Variables	114 669 204
Fundamentals of Electrothermal Atomic Absorption Spectrometry	43 225 669
Heat Conduction Using Green's Functions	16 154 279
Infinitesimal	3 622 203 356
Mathematik	103 631 272
Mechanical Vibration Analysis, Uncertainties, and Control, Third Edition	123 516 889
Nanofabrication Using Focused Ion and Electron Beams	11 092 808
Numerical Analysis, 3rd Edition by Timothy Sauer	28 253 181
Real_Analysis	3 104 213
Special Functions of Mathematics for Engineers, Second Edition Ed 2	13 750 593
The Global Positioning System and ArcGIS	19 751 981
The Integrated String Player Embodied Vibration	34 480 681
Thermodynamics Fundamentals and engineering applications	78 173 129

Figure 2.72: Books-56-2020-202005121527.png

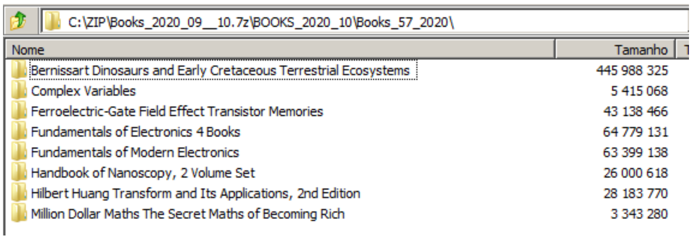

C:\ZIP\Books_2020_09__10.7z\BOOKS_2020_10\Books_57_2020\

Nome	Tamanho
Bernissart Dinosaurs and Early Cretaceous Terrestrial Ecosystems	445 988 325
Complex Variables	5 415 068
Ferroelectric-Gate Field Effect Transistor Memories	43 138 466
Fundamentals of Electronics 4 Books	64 779 131
Fundamentals of Modern Electronics	63 399 138
Handbook of Nanoscopy, 2 Volume Set	26 000 618
Hilbert Huang Transform and Its Applications, 2nd Edition	28 183 770
Million Dollar Maths The Secret Maths of Becoming Rich	3 343 280

Figure 2.73: Books-57-2020-202005121527.png

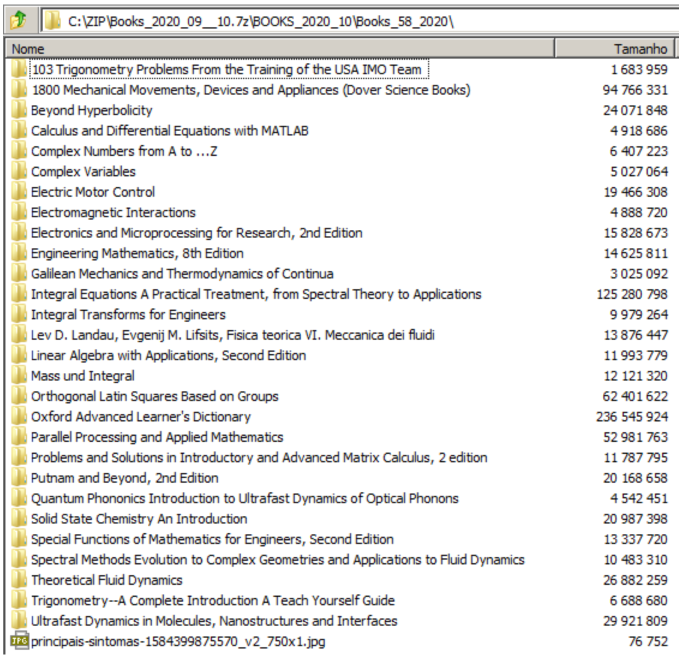

C:\ZIP\Books_2020_09__10.7z\BOOKS_2020_10\Books_58_2020\

Nome	Tamanho
103 Trigonometry Problems From the Training of the USA IMO Team	1 683 959
1800 Mechanical Movements, Devices and Appliances (Dover Science Books)	94 766 331
Beyond Hyperbolicity	24 071 848
Calculus and Differential Equations with MATLAB	4 918 686
Complex Numbers from A to ...Z	6 407 223
Complex Variables	5 027 064
Electric Motor Control	19 466 308
Electromagnetic Interactions	4 888 720
Electronics and Microprocessing for Research, 2nd Edition	15 828 673
Engineering Mathematics, 8th Edition	14 625 811
Galilean Mechanics and Thermodynamics of Continua	3 025 092
Integral Equations A Practical Treatment, from Spectral Theory to Applications	125 280 798
Integral Transforms for Engineers	9 979 264
Lev D. Landau, Evgenij M. Lifsits, Fisica teorica VI. Meccanica dei fluidi	13 876 447
Linear Algebra with Applications, Second Edition	11 993 779
Mass und Integral	12 121 320
Orthogonal Latin Squares Based on Groups	62 401 622
Oxford Advanced Learner's Dictionary	236 545 924
Parallel Processing and Applied Mathematics	52 981 763
Problems and Solutions in Introductory and Advanced Matrix Calculus, 2 edition	11 787 795
Putnam and Beyond, 2nd Edition	20 168 658
Quantum Phononics Introduction to Ultrafast Dynamics of Optical Phonons	4 542 451
Solid State Chemistry An Introduction	20 987 398
Special Functions of Mathematics for Engineers, Second Edition	13 337 720
Spectral Methods Evolution to Complex Geometries and Applications to Fluid Dynamics	10 483 310
Theoretical Fluid Dynamics	26 882 259
Trigonometry--A Complete Introduction A Teach Yourself Guide	6 688 680
Ultrafast Dynamics in Molecules, Nanostructures and Interfaces	29 921 809
principais-sintomas-1584399875570_v2_750x1.jpg	76 752

Figure 2.74: Books-58-2020-202005121527.png

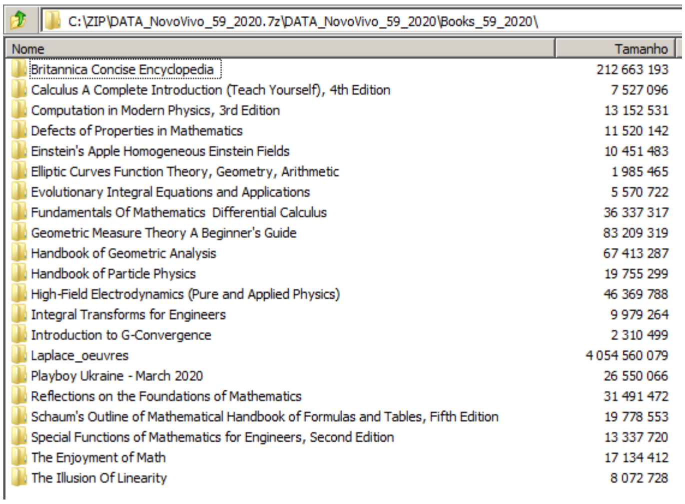

C:\ZIP\DATA_NovoVivo_59_2020.7z\DATA_NovoVivo_59_2020\Books_59_2020\

Nome	Tamanho
Britannica Concise Encyclopedia	212 663 193
Calculus A Complete Introduction (Teach Yourself), 4th Edition	7 527 096
Computation in Modern Physics, 3rd Edition	13 152 531
Defects of Properties in Mathematics	11 520 142
Einstein's Apple Homogeneous Einstein Fields	10 451 483
Elliptic Curves Function Theory, Geometry, Arithmetic	1 985 465
Evolutionary Integral Equations and Applications	5 570 722
Fundamentals Of Mathematics Differential Calculus	36 337 317
Geometric Measure Theory A Beginner's Guide	83 209 319
Handbook of Geometric Analysis	67 413 287
Handbook of Particle Physics	19 755 299
High-Field Electrodynamics (Pure and Applied Physics)	46 369 788
Integral Transforms for Engineers	9 979 264
Introduction to G-Convergence	2 310 499
Laplace_oeuvres	4 054 560 079
Playboy Ukraine - March 2020	26 550 066
Reflections on the Foundations of Mathematics	31 491 472
Schaum's Outline of Mathematical Handbook of Formulas and Tables, Fifth Edition	19 778 553
Special Functions of Mathematics for Engineers, Second Edition	13 337 720
The Enjoyment of Math	17 134 412
The Illusion Of Linearity	8 072 728

Figure 2.75: Books-59-2020-202005122030.png

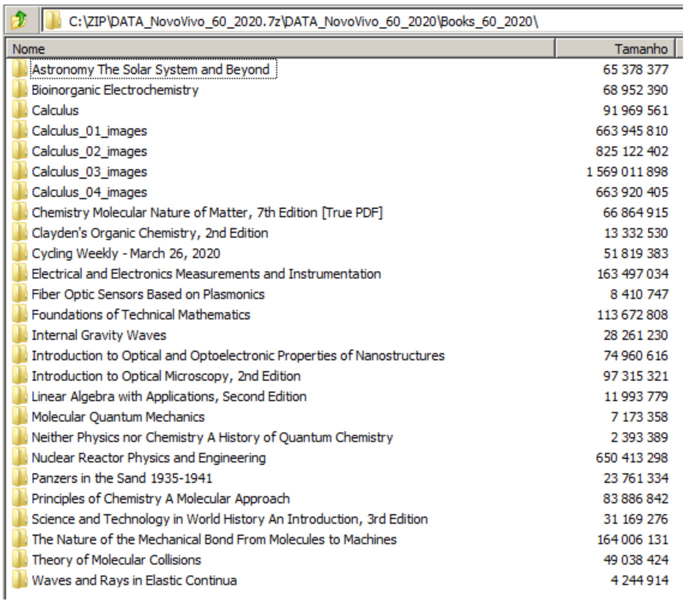

C:\ZIP\DATA_NovoVivo_60_2020.7z\DATA_NovoVivo_60_2020\Books_60_2020\

Nome	Tamanho
Astronomy The Solar System and Beyond	65 378 377
Bioinorganic Electrochemistry	68 952 390
Calculus	91 969 561
Calculus_01_images	663 945 810
Calculus_02_images	825 122 402
Calculus_03_images	1 569 011 898
Calculus_04_images	663 920 405
Chemistry Molecular Nature of Matter, 7th Edition [True PDF]	66 864 915
Clayden's Organic Chemistry, 2nd Edition	13 332 530
Cycling Weekly - March 26, 2020	51 819 383
Electrical and Electronics Measurements and Instrumentation	163 497 034
Fiber Optic Sensors Based on Plasmonics	8 410 747
Foundations of Technical Mathematics	113 672 808
Internal Gravity Waves	28 261 230
Introduction to Optical and Optoelectronic Properties of Nanostructures	74 960 616
Introduction to Optical Microscopy, 2nd Edition	97 315 321
Linear Algebra with Applications, Second Edition	11 993 779
Molecular Quantum Mechanics	7 173 358
Neither Physics nor Chemistry A History of Quantum Chemistry	2 393 389
Nuclear Reactor Physics and Engineering	650 413 298
Panzers in the Sand 1935-1941	23 761 334
Principles of Chemistry A Molecular Approach	83 886 842
Science and Technology in World History An Introduction, 3rd Edition	31 169 276
The Nature of the Mechanical Bond From Molecules to Machines	164 006 131
Theory of Molecular Collisions	49 038 424
Waves and Rays in Elastic Continua	4 244 914

Figure 2.76: Books-60-2020-202005122030.png

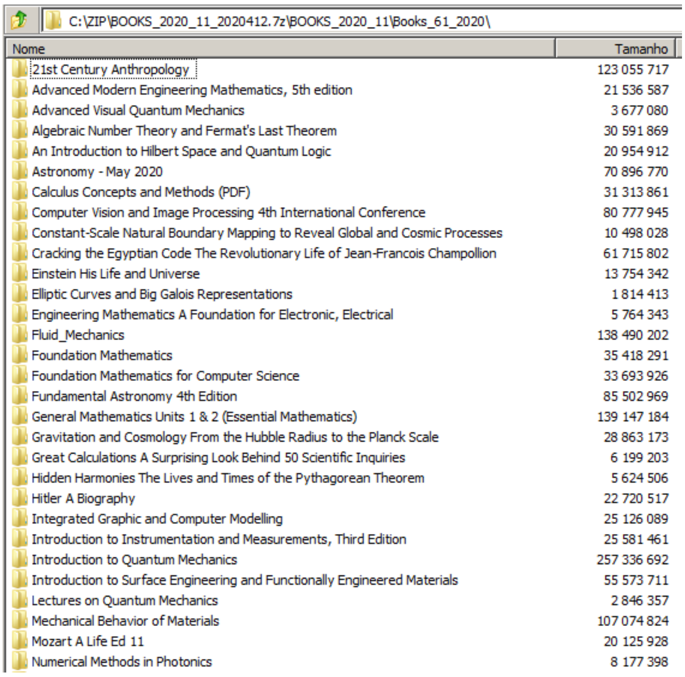

C:\ZIP\BOOKS_2020_11_2020412.7z\BOOKS_2020_11\Books_61_2020\

Nome	Tamanho
21st Century Anthropology	123 055 717
Advanced Modern Engineering Mathematics, 5th edition	21 536 587
Advanced Visual Quantum Mechanics	3 677 080
Algebraic Number Theory and Fermat's Last Theorem	30 591 869
An Introduction to Hilbert Space and Quantum Logic	20 954 912
Astronomy - May 2020	70 896 770
Calculus Concepts and Methods (PDF)	31 313 861
Computer Vision and Image Processing 4th International Conference	80 777 945
Constant-Scale Natural Boundary Mapping to Reveal Global and Cosmic Processes	10 498 028
Cracking the Egyptian Code The Revolutionary Life of Jean-Francois Champollion	61 715 802
Einstein His Life and Universe	13 754 342
Elliptic Curves and Big Galois Representations	1 814 413
Engineering Mathematics A Foundation for Electronic, Electrical	5 764 343
Fluid_Mechanics	138 490 202
Foundation Mathematics	35 418 291
Foundation Mathematics for Computer Science	33 693 926
Fundamental Astronomy 4th Edition	85 502 969
General Mathematics Units 1 & 2 (Essential Mathematics)	139 147 184
Gravitation and Cosmology From the Hubble Radius to the Planck Scale	28 863 173
Great Calculations A Surprising Look Behind 50 Scientific Inquiries	6 199 203
Hidden Harmonies The Lives and Times of the Pythagorean Theorem	5 624 506
Hitler A Biography	22 720 517
Integrated Graphic and Computer Modelling	25 126 089
Introduction to Instrumentation and Measurements, Third Edition	25 581 461
Introduction to Quantum Mechanics	257 336 692
Introduction to Surface Engineering and Functionally Engineered Materials	55 573 711
Lectures on Quantum Mechanics	2 846 357
Mechanical Behavior of Materials	107 074 824
Mozart A Life Ed 11	20 125 928
Numerical Methods in Photonics	8 177 398

Figure 2.77: Books-61-2020-202005121537.png

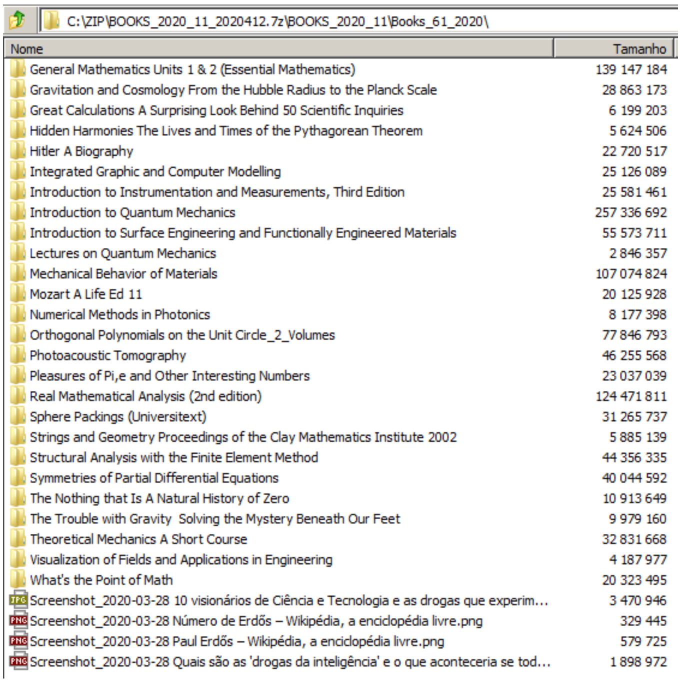

C:\ZIP\BOOKS_2020_11_2020412.7z\BOOKS_2020_11\Books_61_2020\

Nome	Tamanho
General Mathematics Units 1 & 2 (Essential Mathematics)	139 147 184
Gravitation and Cosmology From the Hubble Radius to the Planck Scale	28 863 173
Great Calculations A Surprising Look Behind 50 Scientific Inquiries	6 199 203
Hidden Harmonies The Lives and Times of the Pythagorean Theorem	5 624 506
Hitler A Biography	22 720 517
Integrated Graphic and Computer Modelling	25 126 089
Introduction to Instrumentation and Measurements, Third Edition	25 581 461
Introduction to Quantum Mechanics	257 336 692
Introduction to Surface Engineering and Functionally Engineered Materials	55 573 711
Lectures on Quantum Mechanics	2 846 357
Mechanical Behavior of Materials	107 074 824
Mozart A Life Ed 11	20 125 928
Numerical Methods in Photonics	8 177 398
Orthogonal Polynomials on the Unit Circle_2_Volumes	77 846 793
Photoacoustic Tomography	46 255 568
Pleasures of Pi,e and Other Interesting Numbers	23 037 039
Real Mathematical Analysis (2nd edition)	124 471 811
Sphere Packings (Universitext)	31 265 737
Strings and Geometry Proceedings of the Clay Mathematics Institute 2002	5 885 139
Structural Analysis with the Finite Element Method	44 356 335
Symmetries of Partial Differential Equations	40 044 592
The Nothing that Is A Natural History of Zero	10 913 649
The Trouble with Gravity Solving the Mystery Beneath Our Feet	9 979 160
Theoretical Mechanics A Short Course	32 831 668
Visualization of Fields and Applications in Engineering	4 187 977
What's the Point of Math	20 323 495
Screenshot_2020-03-28 10 visionários de Ciência e Tecnologia e as drogas que experim...	3 470 946
Screenshot_2020-03-28 Número de Erdős – Wikipédia, a enciclopédia livre.png	329 445
Screenshot_2020-03-28 Paul Erdős – Wikipédia, a enciclopédia livre.png	579 725
Screenshot_2020-03-28 Quais são as 'drogas da inteligência' e o que aconteceria se tod...	1 898 972

Figure 2.78: Books-61-2020-202005121538.png

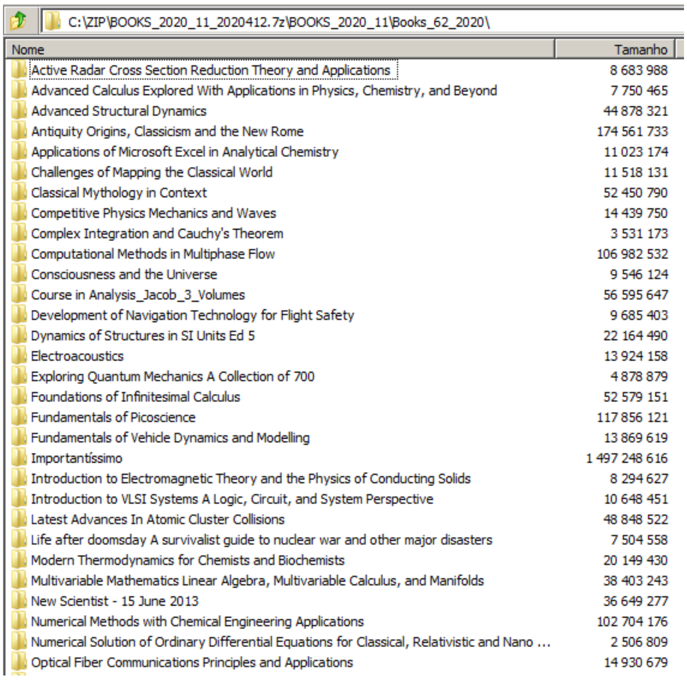

C:\ZIP\BOOKS_2020_11_2020412.7z\BOOKS_2020_11\Books_62_2020\

Nome	Tamanho
Active Radar Cross Section Reduction Theory and Applications	8 683 988
Advanced Calculus Explored With Applications in Physics, Chemistry, and Beyond	7 750 465
Advanced Structural Dynamics	44 878 321
Antiquity Origins, Classicism and the New Rome	174 561 733
Applications of Microsoft Excel in Analytical Chemistry	11 023 174
Challenges of Mapping the Classical World	11 518 131
Classical Mythology in Context	52 450 790
Competitive Physics Mechanics and Waves	14 439 750
Complex Integration and Cauchy's Theorem	3 531 173
Computational Methods in Multiphase Flow	106 982 532
Consciousness and the Universe	9 546 124
Course in Analysis_Jacob_3_Volumes	56 595 647
Development of Navigation Technology for Flight Safety	9 685 403
Dynamics of Structures in SI Units Ed 5	22 164 490
Electroacoustics	13 924 158
Exploring Quantum Mechanics A Collection of 700	4 878 879
Foundations of Infinitesimal Calculus	52 579 151
Fundamentals of Picoscience	117 856 121
Fundamentals of Vehicle Dynamics and Modelling	13 869 619
Importantíssimo	1 497 248 616
Introduction to Electromagnetic Theory and the Physics of Conducting Solids	8 294 627
Introduction to VLSI Systems A Logic, Circuit, and System Perspective	10 648 451
Latest Advances In Atomic Cluster Collisions	48 848 522
Life after doomsday A survivalist guide to nuclear war and other major disasters	7 504 558
Modern Thermodynamics for Chemists and Biochemists	20 149 430
Multivariable Mathematics Linear Algebra, Multivariable Calculus, and Manifolds	38 403 243
New Scientist - 15 June 2013	36 649 277
Numerical Methods with Chemical Engineering Applications	102 704 176
Numerical Solution of Ordinary Differential Equations for Classical, Relativistic and Nano ...	2 506 809
Optical Fiber Communications Principles and Applications	14 930 679

Figure 2.79: Books-62-2020-202005121538.png

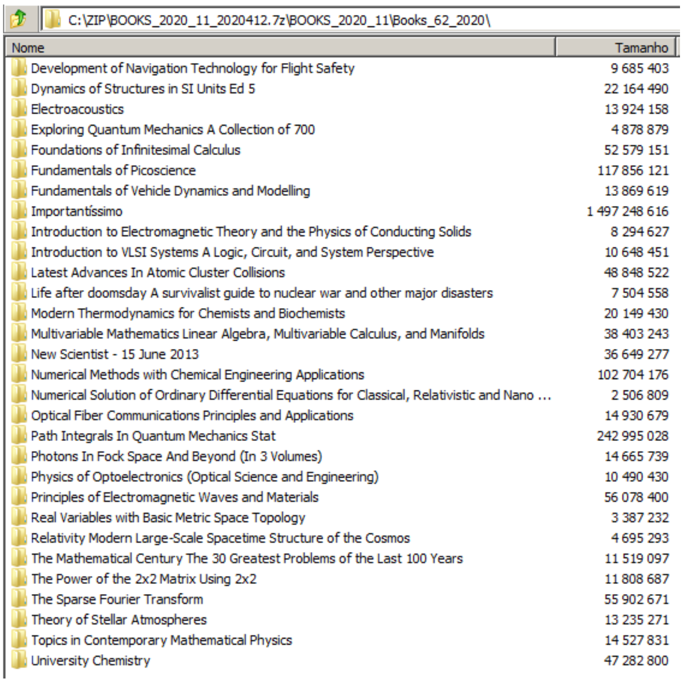

C:\ZIP\BOOKS_2020_11_2020412.7z\BOOKS_2020_11\Books_62_2020\

Nome	Tamanho
Development of Navigation Technology for Flight Safety	9 685 403
Dynamics of Structures in SI Units Ed 5	22 164 490
Electroacoustics	13 924 158
Exploring Quantum Mechanics A Collection of 700	4 878 879
Foundations of Infinitesimal Calculus	52 579 151
Fundamentals of Picoscience	117 856 121
Fundamentals of Vehicle Dynamics and Modelling	13 869 619
Importantíssimo	1 497 248 616
Introduction to Electromagnetic Theory and the Physics of Conducting Solids	8 294 627
Introduction to VLSI Systems A Logic, Circuit, and System Perspective	10 648 451
Latest Advances In Atomic Cluster Collisions	48 848 522
Life after doomsday A survivalist guide to nuclear war and other major disasters	7 504 558
Modern Thermodynamics for Chemists and Biochemists	20 149 430
Multivariable Mathematics Linear Algebra, Multivariable Calculus, and Manifolds	38 403 243
New Scientist - 15 June 2013	36 649 277
Numerical Methods with Chemical Engineering Applications	102 704 176
Numerical Solution of Ordinary Differential Equations for Classical, Relativistic and Nano ...	2 506 809
Optical Fiber Communications Principles and Applications	14 930 679
Path Integrals In Quantum Mechanics Stat	242 995 028
Photons In Fock Space And Beyond (In 3 Volumes)	14 665 739
Physics of Optoelectronics (Optical Science and Engineering)	10 490 430
Principles of Electromagnetic Waves and Materials	56 078 400
Real Variables with Basic Metric Space Topology	3 387 232
Relativity Modern Large-Scale Spacetime Structure of the Cosmos	4 695 293
The Mathematical Century The 30 Greatest Problems of the Last 100 Years	11 519 097
The Power of the 2x2 Matrix Using 2x2	11 808 687
The Sparse Fourier Transform	55 902 671
Theory of Stellar Atmospheres	13 235 271
Topics in Contemporary Mathematical Physics	14 527 831
University Chemistry	47 282 800

Figure 2.80: Books-62-2020-202005121539.png

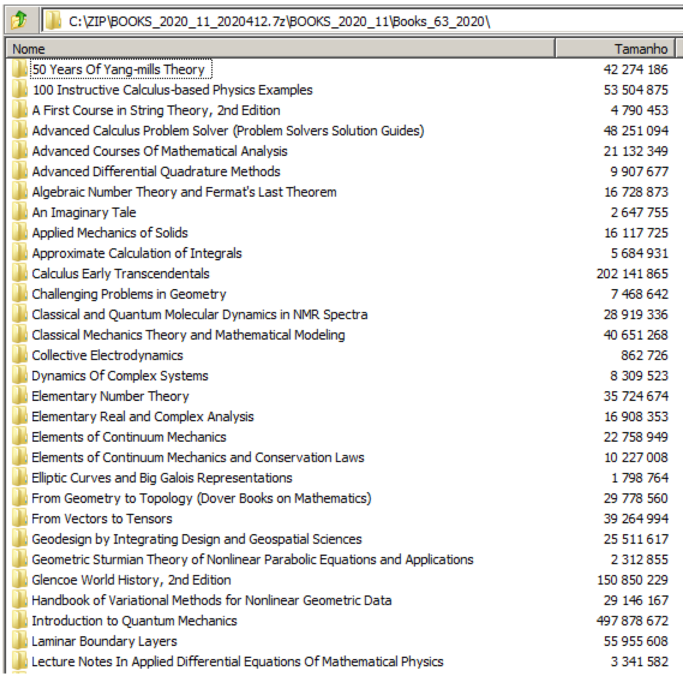

C:\ZIP\BOOKS_2020_11_2020412.7z\BOOKS_2020_11\Books_63_2020\

Nome	Tamanho
50 Years Of Yang-mills Theory	42 274 186
100 Instructive Calculus-based Physics Examples	53 504 875
A First Course in String Theory, 2nd Edition	4 790 453
Advanced Calculus Problem Solver (Problem Solvers Solution Guides)	48 251 094
Advanced Courses Of Mathematical Analysis	21 132 349
Advanced Differential Quadrature Methods	9 907 677
Algebraic Number Theory and Fermat's Last Theorem	16 728 873
An Imaginary Tale	2 647 755
Applied Mechanics of Solids	16 117 725
Approximate Calculation of Integrals	5 684 931
Calculus Early Transcendentals	202 141 865
Challenging Problems in Geometry	7 468 642
Classical and Quantum Molecular Dynamics in NMR Spectra	28 919 336
Classical Mechanics Theory and Mathematical Modeling	40 651 268
Collective Electrodynamics	862 726
Dynamics Of Complex Systems	8 309 523
Elementary Number Theory	35 724 674
Elementary Real and Complex Analysis	16 908 353
Elements of Continuum Mechanics	22 758 949
Elements of Continuum Mechanics and Conservation Laws	10 227 008
Elliptic Curves and Big Galois Representations	1 798 764
From Geometry to Topology (Dover Books on Mathematics)	29 778 560
From Vectors to Tensors	39 264 994
Geodesign by Integrating Design and Geospatial Sciences	25 511 617
Geometric Sturmian Theory of Nonlinear Parabolic Equations and Applications	2 312 855
Glencoe World History, 2nd Edition	150 850 229
Handbook of Variational Methods for Nonlinear Geometric Data	29 146 167
Introduction to Quantum Mechanics	497 878 672
Laminar Boundary Layers	55 955 608
Lecture Notes In Applied Differential Equations Of Mathematical Physics	3 341 582

Figure 2.81: Books-63-2020-202005121539.png

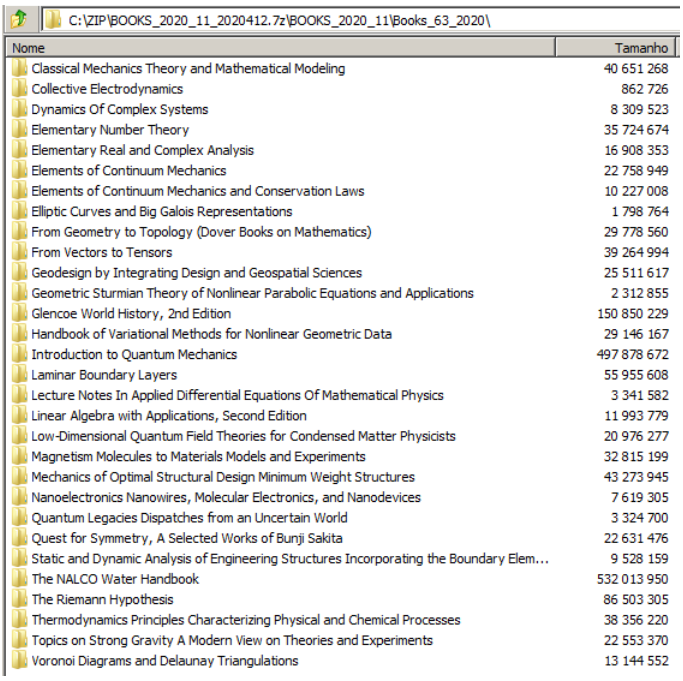

C:\ZIP\BOOKS_2020_11_2020412.7z\BOOKS_2020_11\Books_63_2020\

Nome	Tamanho
Classical Mechanics Theory and Mathematical Modeling	40 651 268
Collective Electrodynamics	862 726
Dynamics Of Complex Systems	8 309 523
Elementary Number Theory	35 724 674
Elementary Real and Complex Analysis	16 908 353
Elements of Continuum Mechanics	22 758 949
Elements of Continuum Mechanics and Conservation Laws	10 227 008
Elliptic Curves and Big Galois Representations	1 798 764
From Geometry to Topology (Dover Books on Mathematics)	29 778 560
From Vectors to Tensors	39 264 994
Geodesign by Integrating Design and Geospatial Sciences	25 511 617
Geometric Sturmian Theory of Nonlinear Parabolic Equations and Applications	2 312 855
Glencoe World History, 2nd Edition	150 850 229
Handbook of Variational Methods for Nonlinear Geometric Data	29 146 167
Introduction to Quantum Mechanics	497 878 672
Laminar Boundary Layers	55 955 608
Lecture Notes In Applied Differential Equations Of Mathematical Physics	3 341 582
Linear Algebra with Applications, Second Edition	11 993 779
Low-Dimensional Quantum Field Theories for Condensed Matter Physicists	20 976 277
Magnetism Molecules to Materials Models and Experiments	32 815 199
Mechanics of Optimal Structural Design Minimum Weight Structures	43 273 945
Nanoelectronics Nanowires, Molecular Electronics, and Nanodevices	7 619 305
Quantum Legacies Dispatches from an Uncertain World	3 324 700
Quest for Symmetry, A Selected Works of Bunji Sakita	22 631 476
Static and Dynamic Analysis of Engineering Structures Incorporating the Boundary Elem...	9 528 159
The NALCO Water Handbook	532 013 950
The Riemann Hypothesis	86 503 305
Thermodynamics Principles Characterizing Physical and Chemical Processes	38 356 220
Topics on Strong Gravity A Modern View on Theories and Experiments	22 553 370
Voronoi Diagrams and Delaunay Triangulations	13 144 552

Figure 2.82: Books-63-2020-202005121540.png

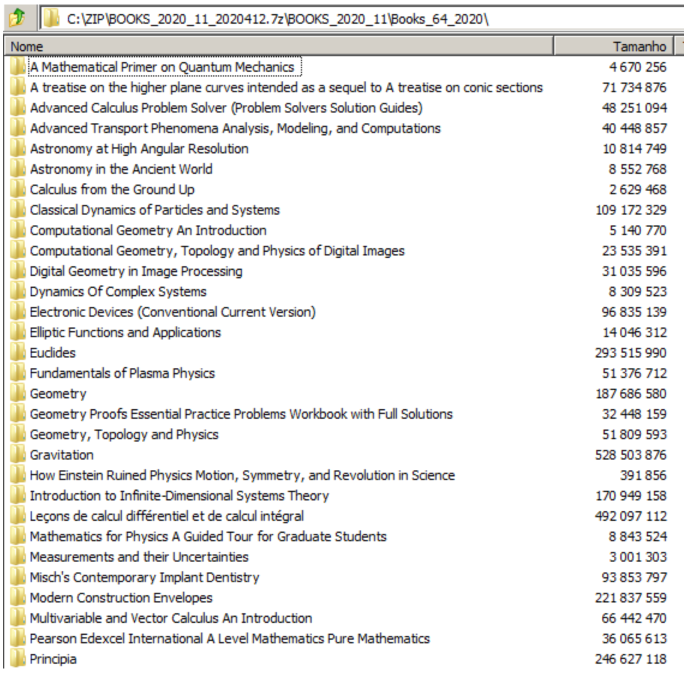

C:\ZIP\BOOKS_2020_11_2020412.7z\BOOKS_2020_11\Books_64_2020\

Nome	Tamanho
A Mathematical Primer on Quantum Mechanics	4 670 256
A treatise on the higher plane curves intended as a sequel to A treatise on conic sections	71 734 876
Advanced Calculus Problem Solver (Problem Solvers Solution Guides)	48 251 094
Advanced Transport Phenomena Analysis, Modeling, and Computations	40 448 857
Astronomy at High Angular Resolution	10 814 749
Astronomy in the Ancient World	8 552 768
Calculus from the Ground Up	2 629 468
Classical Dynamics of Particles and Systems	109 172 329
Computational Geometry An Introduction	5 140 770
Computational Geometry, Topology and Physics of Digital Images	23 535 391
Digital Geometry in Image Processing	31 035 596
Dynamics Of Complex Systems	8 309 523
Electronic Devices (Conventional Current Version)	96 835 139
Elliptic Functions and Applications	14 046 312
Euclides	293 515 990
Fundamentals of Plasma Physics	51 376 712
Geometry	187 686 580
Geometry Proofs Essential Practice Problems Workbook with Full Solutions	32 448 159
Geometry, Topology and Physics	51 809 593
Gravitation	528 503 876
How Einstein Ruined Physics Motion, Symmetry, and Revolution in Science	391 856
Introduction to Infinite-Dimensional Systems Theory	170 949 158
Leçons de calcul différentiel et de calcul intégral	492 097 112
Mathematics for Physics A Guided Tour for Graduate Students	8 843 524
Measurements and their Uncertainties	3 001 303
Misch's Contemporary Implant Dentistry	93 853 797
Modern Construction Envelopes	221 837 559
Multivariable and Vector Calculus An Introduction	66 442 470
Pearson Edexcel International A Level Mathematics Pure Mathematics	36 065 613
Principia	246 627 118

Figure 2.83: Books-64-2020-202005121540.png

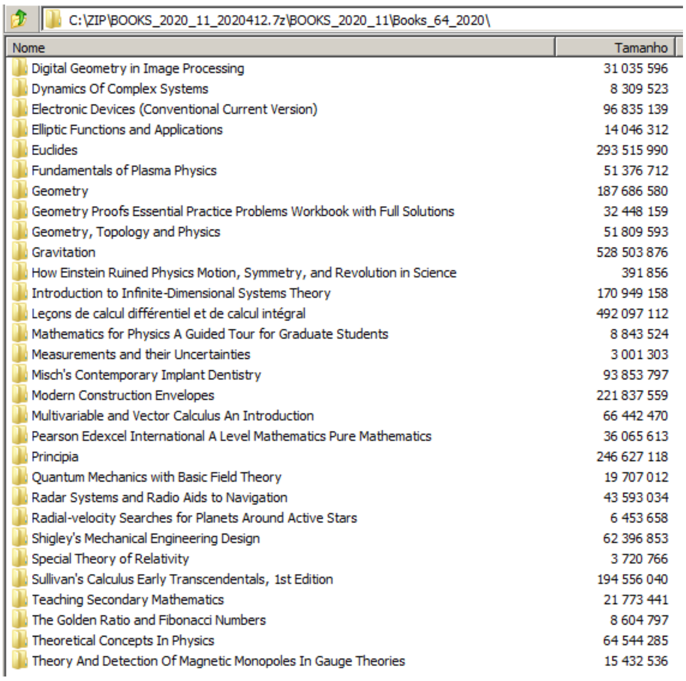

C:\ZIP\BOOKS_2020_11_2020412.7z\BOOKS_2020_11\Books_64_2020\

Nome	Tamanho
Digital Geometry in Image Processing	31 035 596
Dynamics Of Complex Systems	8 309 523
Electronic Devices (Conventional Current Version)	96 835 139
Elliptic Functions and Applications	14 046 312
Euclides	293 515 990
Fundamentals of Plasma Physics	51 376 712
Geometry	187 686 580
Geometry Proofs Essential Practice Problems Workbook with Full Solutions	32 448 159
Geometry, Topology and Physics	51 809 593
Gravitation	528 503 876
How Einstein Ruined Physics Motion, Symmetry, and Revolution in Science	391 856
Introduction to Infinite-Dimensional Systems Theory	170 949 158
Leçons de calcul différentiel et de calcul intégral	492 097 112
Mathematics for Physics A Guided Tour for Graduate Students	8 843 524
Measurements and their Uncertainties	3 001 303
Misch's Contemporary Implant Dentistry	93 853 797
Modern Construction Envelopes	221 837 559
Multivariable and Vector Calculus An Introduction	66 442 470
Pearson Edexcel International A Level Mathematics Pure Mathematics	36 065 613
Principia	246 627 118
Quantum Mechanics with Basic Field Theory	19 707 012
Radar Systems and Radio Aids to Navigation	43 593 034
Radial-velocity Searches for Planets Around Active Stars	6 453 658
Shigley's Mechanical Engineering Design	62 396 853
Special Theory of Relativity	3 720 766
Sullivan's Calculus Early Transcendentals, 1st Edition	194 556 040
Teaching Secondary Mathematics	21 773 441
The Golden Ratio and Fibonacci Numbers	8 604 797
Theoretical Concepts In Physics	64 544 285
Theory And Detection Of Magnetic Monopoles In Gauge Theories	15 432 536

Figure 2.84: Books-64-2020-202005121541.png

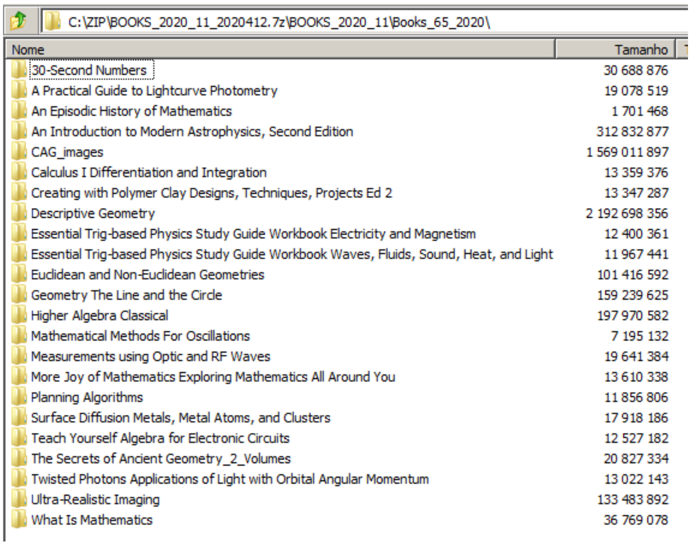

C:\ZIP\BOOKS_2020_11_2020412.7z\BOOKS_2020_11\Books_65_2020\

Nome	Tamanho
30-Second Numbers	30 688 876
A Practical Guide to Lightcurve Photometry	19 078 519
An Episodic History of Mathematics	1 701 468
An Introduction to Modern Astrophysics, Second Edition	312 832 877
CAG_images	1 569 011 897
Calculus I Differentiation and Integration	13 359 376
Creating with Polymer Clay Designs, Techniques, Projects Ed 2	13 347 287
Descriptive Geometry	2 192 698 356
Essential Trig-based Physics Study Guide Workbook Electricity and Magnetism	12 400 361
Essential Trig-based Physics Study Guide Workbook Waves, Fluids, Sound, Heat, and Light	11 967 441
Euclidean and Non-Euclidean Geometries	101 416 592
Geometry The Line and the Circle	159 239 625
Higher Algebra Classical	197 970 582
Mathematical Methods For Oscillations	7 195 132
Measurements using Optic and RF Waves	19 641 384
More Joy of Mathematics Exploring Mathematics All Around You	13 610 338
Planning Algorithms	11 856 806
Surface Diffusion Metals, Metal Atoms, and Clusters	17 918 186
Teach Yourself Algebra for Electronic Circuits	12 527 182
The Secrets of Ancient Geometry_2_Volumes	20 827 334
Twisted Photons Applications of Light with Orbital Angular Momentum	13 022 143
Ultra-Realistic Imaging	133 483 892
What Is Mathematics	36 769 078

Figure 2.85: Books-65-2020-202005121541.png

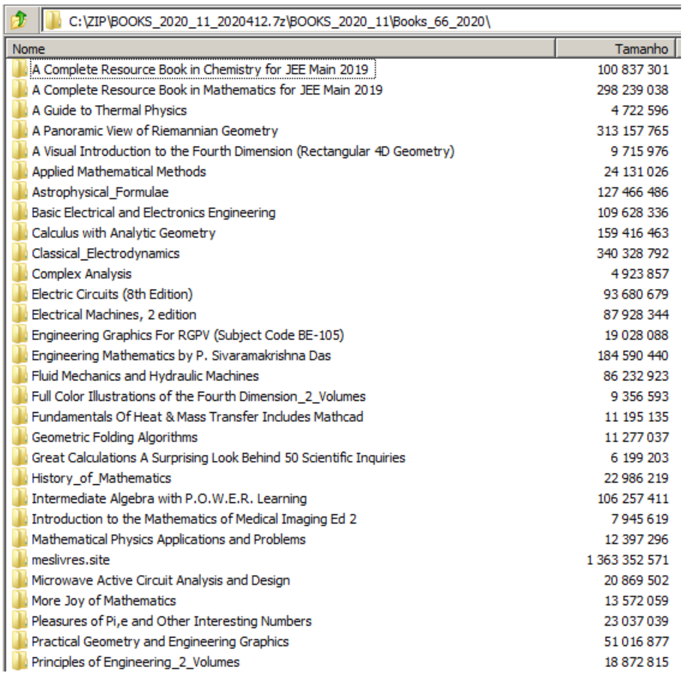

C:\ZIP\BOOKS_2020_11_2020412.7z\BOOKS_2020_11\Books_66_2020\

Nome	Tamanho
A Complete Resource Book in Chemistry for JEE Main 2019	100 837 301
A Complete Resource Book in Mathematics for JEE Main 2019	298 239 038
A Guide to Thermal Physics	4 722 596
A Panoramic View of Riemannian Geometry	313 157 765
A Visual Introduction to the Fourth Dimension (Rectangular 4D Geometry)	9 715 976
Applied Mathematical Methods	24 131 026
Astrophysical_Formulae	127 466 486
Basic Electrical and Electronics Engineering	109 628 336
Calculus with Analytic Geometry	159 416 463
Classical_Electrodynamics	340 328 792
Complex Analysis	4 923 857
Electric Circuits (8th Edition)	93 680 679
Electrical Machines, 2 edition	87 928 344
Engineering Graphics For RGPV (Subject Code BE-105)	19 028 088
Engineering Mathematics by P. Sivaramakrishna Das	184 590 440
Fluid Mechanics and Hydraulic Machines	86 232 923
Full Color Illustrations of the Fourth Dimension_2_Volumes	9 356 593
Fundamentals Of Heat & Mass Transfer Includes Mathcad	11 195 135
Geometric Folding Algorithms	11 277 037
Great Calculations A Surprising Look Behind 50 Scientific Inquiries	6 199 203
History_of_Mathematics	22 986 219
Intermediate Algebra with P.O.W.E.R. Learning	106 257 411
Introduction to the Mathematics of Medical Imaging Ed 2	7 945 619
Mathematical Physics Applications and Problems	12 397 296
meslivres.site	1 363 352 571
Microwave Active Circuit Analysis and Design	20 869 502
More Joy of Mathematics	13 572 059
Pleasures of Pi,e and Other Interesting Numbers	23 037 039
Practical Geometry and Engineering Graphics	51 016 877
Principles of Engineering_2_Volumes	18 872 815

Figure 2.86: Books-66-2020-202005121541.png

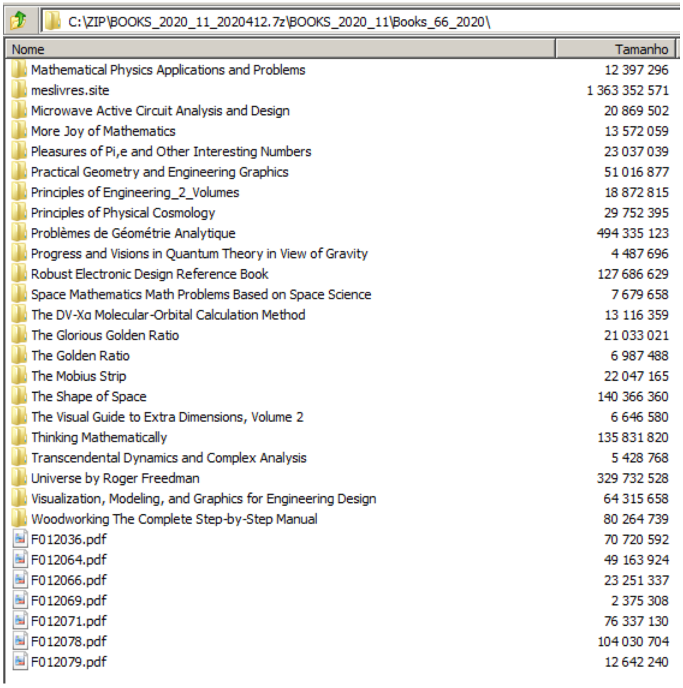

C:\ZIP\BOOKS_2020_11_2020412.7z\BOOKS_2020_11\Books_66_2020\

Nome	Tamanho
Mathematical Physics Applications and Problems	12 397 296
meslivres.site	1 363 352 571
Microwave Active Circuit Analysis and Design	20 869 502
More Joy of Mathematics	13 572 059
Pleasures of Pi,e and Other Interesting Numbers	23 037 039
Practical Geometry and Engineering Graphics	51 016 877
Principles of Engineering_2_Volumes	18 872 815
Principles of Physical Cosmology	29 752 395
Problèmes de Géométrie Analytique	494 335 123
Progress and Visions in Quantum Theory in View of Gravity	4 487 696
Robust Electronic Design Reference Book	127 686 629
Space Mathematics Math Problems Based on Space Science	7 679 658
The DV-Xα Molecular-Orbital Calculation Method	13 116 359
The Glorious Golden Ratio	21 033 021
The Golden Ratio	6 987 488
The Mobius Strip	22 047 165
The Shape of Space	140 366 360
The Visual Guide to Extra Dimensions, Volume 2	6 646 580
Thinking Mathematically	135 831 820
Transcendental Dynamics and Complex Analysis	5 428 768
Universe by Roger Freedman	329 732 528
Visualization, Modeling, and Graphics for Engineering Design	64 315 658
Woodworking The Complete Step-by-Step Manual	80 264 739
F012036.pdf	70 720 592
F012064.pdf	49 163 924
F012066.pdf	23 251 337
F012069.pdf	2 375 308
F012071.pdf	76 337 130
F012078.pdf	104 030 704
F012079.pdf	12 642 240

Figure 2.87: Books-66-2020-202005121542.png

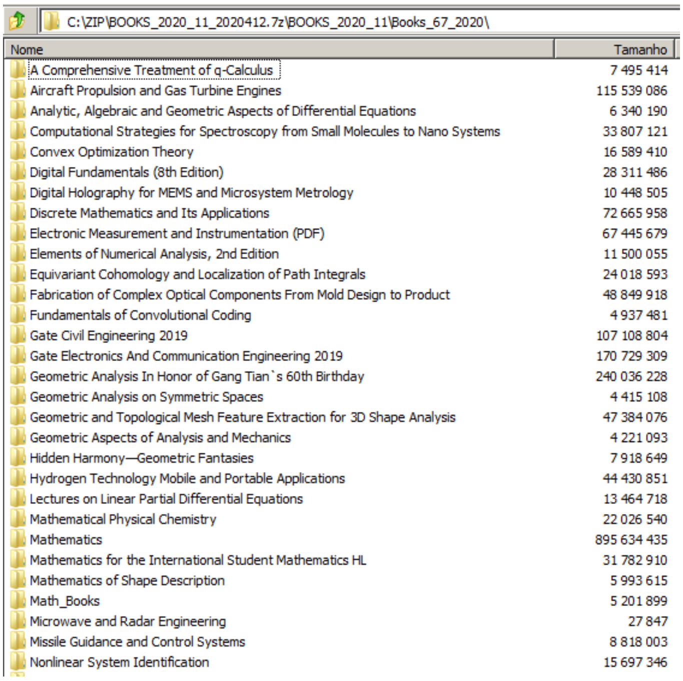

C:\ZIP\BOOKS_2020_11_2020412.7z\BOOKS_2020_11\Books_67_2020\

Nome	Tamanho
A Comprehensive Treatment of q-Calculus	7 495 414
Aircraft Propulsion and Gas Turbine Engines	115 539 086
Analytic, Algebraic and Geometric Aspects of Differential Equations	6 340 190
Computational Strategies for Spectroscopy from Small Molecules to Nano Systems	33 807 121
Convex Optimization Theory	16 589 410
Digital Fundamentals (8th Edition)	28 311 486
Digital Holography for MEMS and Microsystem Metrology	10 448 505
Discrete Mathematics and Its Applications	72 665 958
Electronic Measurement and Instrumentation (PDF)	67 445 679
Elements of Numerical Analysis, 2nd Edition	11 500 055
Equivariant Cohomology and Localization of Path Integrals	24 018 593
Fabrication of Complex Optical Components From Mold Design to Product	48 849 918
Fundamentals of Convolutional Coding	4 937 481
Gate Civil Engineering 2019	107 108 804
Gate Electronics And Communication Engineering 2019	170 729 309
Geometric Analysis In Honor of Gang Tian`s 60th Birthday	240 036 228
Geometric Analysis on Symmetric Spaces	4 415 108
Geometric and Topological Mesh Feature Extraction for 3D Shape Analysis	47 384 076
Geometric Aspects of Analysis and Mechanics	4 221 093
Hidden Harmony—Geometric Fantasies	7 918 649
Hydrogen Technology Mobile and Portable Applications	44 430 851
Lectures on Linear Partial Differential Equations	13 464 718
Mathematical Physical Chemistry	22 026 540
Mathematics	895 634 435
Mathematics for the International Student Mathematics HL	31 782 910
Mathematics of Shape Description	5 993 615
Math_Books	5 201 899
Microwave and Radar Engineering	27 847
Missile Guidance and Control Systems	8 818 003
Nonlinear System Identification	15 697 346

Figure 2.88: Books-67-2020-202005121542.png

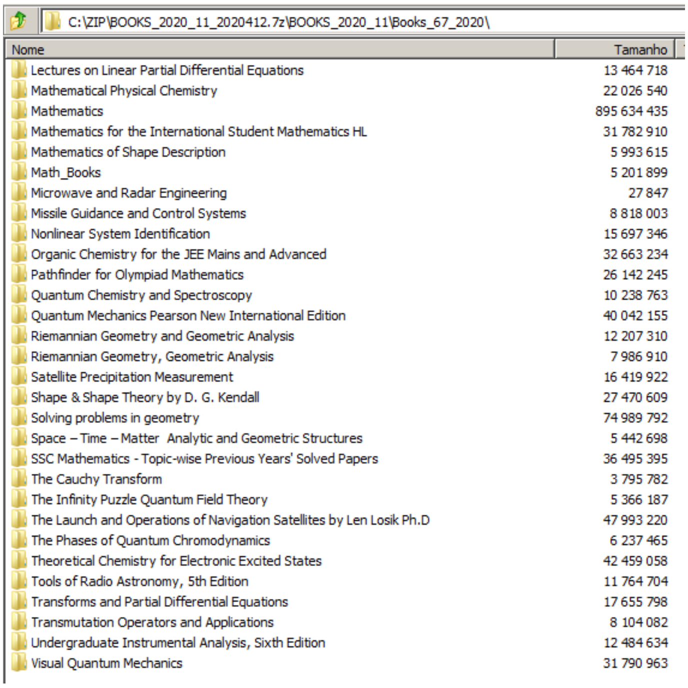

C:\ZIP\BOOKS_2020_11_2020412.7z\BOOKS_2020_11\Books_67_2020\

Nome	Tamanho
Lectures on Linear Partial Differential Equations	13 464 718
Mathematical Physical Chemistry	22 026 540
Mathematics	895 634 435
Mathematics for the International Student Mathematics HL	31 782 910
Mathematics of Shape Description	5 993 615
Math_Books	5 201 899
Microwave and Radar Engineering	27 847
Missile Guidance and Control Systems	8 818 003
Nonlinear System Identification	15 697 346
Organic Chemistry for the JEE Mains and Advanced	32 663 234
Pathfinder for Olympiad Mathematics	26 142 245
Quantum Chemistry and Spectroscopy	10 238 763
Quantum Mechanics Pearson New International Edition	40 042 155
Riemannian Geometry and Geometric Analysis	12 207 310
Riemannian Geometry, Geometric Analysis	7 986 910
Satellite Precipitation Measurement	16 419 922
Shape & Shape Theory by D. G. Kendall	27 470 609
Solving problems in geometry	74 989 792
Space – Time – Matter Analytic and Geometric Structures	5 442 698
SSC Mathematics - Topic-wise Previous Years' Solved Papers	36 495 395
The Cauchy Transform	3 795 782
The Infinity Puzzle Quantum Field Theory	5 366 187
The Launch and Operations of Navigation Satellites by Len Losik Ph.D	47 993 220
The Phases of Quantum Chromodynamics	6 237 465
Theoretical Chemistry for Electronic Excited States	42 459 058
Tools of Radio Astronomy, 5th Edition	11 764 704
Transforms and Partial Differential Equations	17 655 798
Transmutation Operators and Applications	8 104 082
Undergraduate Instrumental Analysis, Sixth Edition	12 484 634
Visual Quantum Mechanics	31 790 963

Figure 2.89: Books-67-2020-202005121543.png

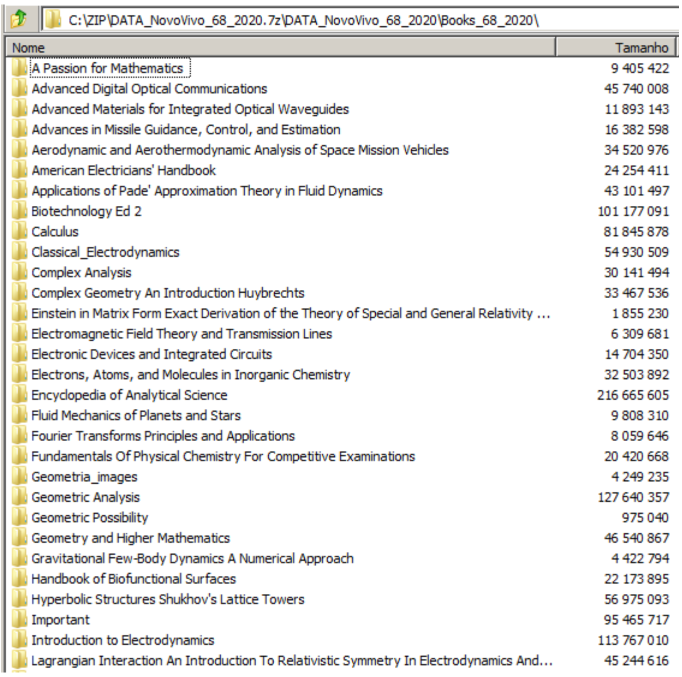

C:\ZIP\DATA_NovoVivo_68_2020.7z\DATA_NovoVivo_68_2020\Books_68_2020\

Nome	Tamanho
A Passion for Mathematics	9 405 422
Advanced Digital Optical Communications	45 740 008
Advanced Materials for Integrated Optical Waveguides	11 893 143
Advances in Missile Guidance, Control, and Estimation	16 382 598
Aerodynamic and Aerothermodynamic Analysis of Space Mission Vehicles	34 520 976
American Electricians' Handbook	24 254 411
Applications of Pade' Approximation Theory in Fluid Dynamics	43 101 497
Biotechnology Ed 2	101 177 091
Calculus	81 845 878
Classical_Electrodynamics	54 930 509
Complex Analysis	30 141 494
Complex Geometry An Introduction Huybrechts	33 467 536
Einstein in Matrix Form Exact Derivation of the Theory of Special and General Relativity ...	1 855 230
Electromagnetic Field Theory and Transmission Lines	6 309 681
Electronic Devices and Integrated Circuits	14 704 350
Electrons, Atoms, and Molecules in Inorganic Chemistry	32 503 892
Encyclopedia of Analytical Science	216 665 605
Fluid Mechanics of Planets and Stars	9 808 310
Fourier Transforms Principles and Applications	8 059 646
Fundamentals Of Physical Chemistry For Competitive Examinations	20 420 668
Geometria_images	4 249 235
Geometric Analysis	127 640 357
Geometric Possibility	975 040
Geometry and Higher Mathematics	46 540 867
Gravitational Few-Body Dynamics A Numerical Approach	4 422 794
Handbook of Biofunctional Surfaces	22 173 895
Hyperbolic Structures Shukhov's Lattice Towers	56 975 093
Important	95 465 717
Introduction to Electrodynamics	113 767 010
Lagrangian Interaction An Introduction To Relativistic Symmetry In Electrodynamics And...	45 244 616

Figure 2.90: Books-68-2020-202005122054.png

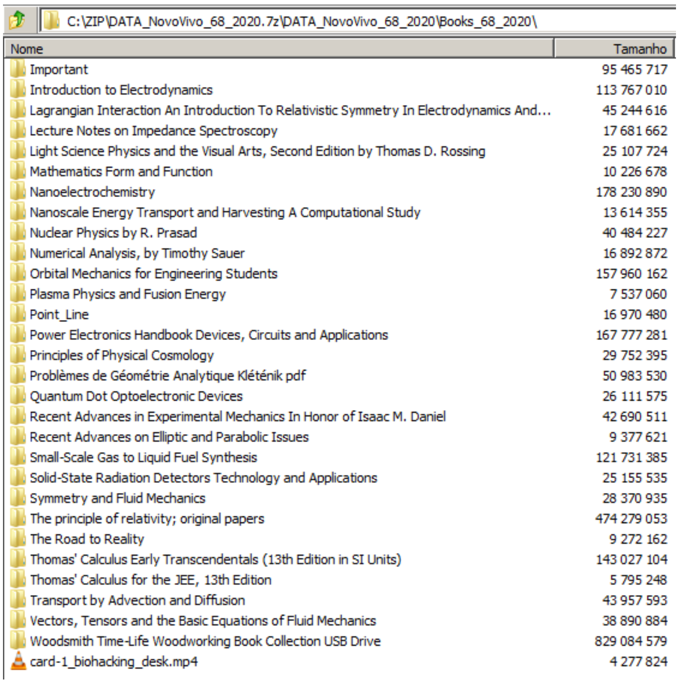

C:\ZIP\DATA_NovoVivo_68_2020.7z\DATA_NovoVivo_68_2020\Books_68_2020\

Nome	Tamanho
Important	95 465 717
Introduction to Electrodynamics	113 767 010
Lagrangian Interaction An Introduction To Relativistic Symmetry In Electrodynamics And...	45 244 616
Lecture Notes on Impedance Spectroscopy	17 681 662
Light Science Physics and the Visual Arts, Second Edition by Thomas D. Rossing	25 107 724
Mathematics Form and Function	10 226 678
Nanoelectrochemistry	178 230 890
Nanoscale Energy Transport and Harvesting A Computational Study	13 614 355
Nuclear Physics by R. Prasad	40 484 227
Numerical Analysis, by Timothy Sauer	16 892 872
Orbital Mechanics for Engineering Students	157 960 162
Plasma Physics and Fusion Energy	7 537 060
Point_Line	16 970 480
Power Electronics Handbook Devices, Circuits and Applications	167 777 281
Principles of Physical Cosmology	29 752 395
Problèmes de Géométrie Analytique Kléténik pdf	50 983 530
Quantum Dot Optoelectronic Devices	26 111 575
Recent Advances in Experimental Mechanics In Honor of Isaac M. Daniel	42 690 511
Recent Advances on Elliptic and Parabolic Issues	9 377 621
Small-Scale Gas to Liquid Fuel Synthesis	121 731 385
Solid-State Radiation Detectors Technology and Applications	25 155 535
Symmetry and Fluid Mechanics	28 370 935
The principle of relativity; original papers	474 279 053
The Road to Reality	9 272 162
Thomas' Calculus Early Transcendentals (13th Edition in SI Units)	143 027 104
Thomas' Calculus for the JEE, 13th Edition	5 795 248
Transport by Advection and Diffusion	43 957 593
Vectors, Tensors and the Basic Equations of Fluid Mechanics	38 890 884
Woodsmith Time-Life Woodworking Book Collection USB Drive	829 084 579
card-1_biohacking_desk.mp4	4 277 824

Figure 2.91: Books-68-2020-202005122055.png

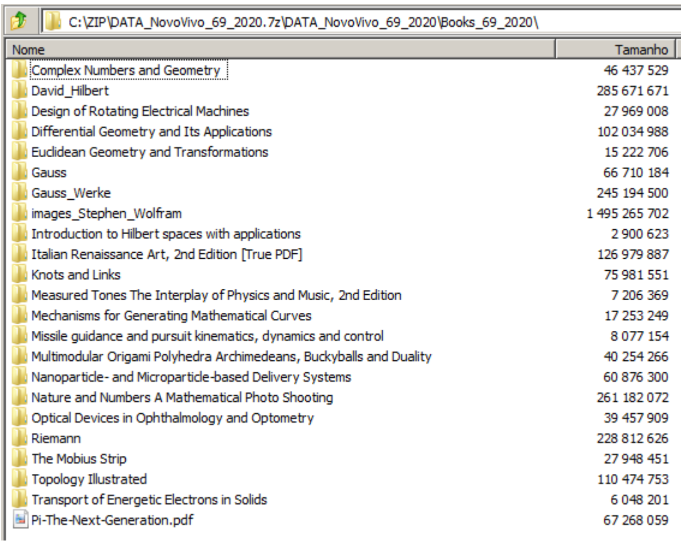

C:\ZIP\DATA_NovoVivo_69_2020.7z\DATA_NovoVivo_69_2020\Books_69_2020\

Nome	Tamanho
Complex Numbers and Geometry	46 437 529
David_Hilbert	285 671 671
Design of Rotating Electrical Machines	27 969 008
Differential Geometry and Its Applications	102 034 988
Euclidean Geometry and Transformations	15 222 706
Gauss	66 710 184
Gauss_Werke	245 194 500
images_Stephen_Wolfram	1 495 265 702
Introduction to Hilbert spaces with applications	2 900 623
Italian Renaissance Art, 2nd Edition [True PDF]	126 979 887
Knots and Links	75 981 551
Measured Tones The Interplay of Physics and Music, 2nd Edition	7 206 369
Mechanisms for Generating Mathematical Curves	17 253 249
Missile guidance and pursuit kinematics, dynamics and control	8 077 154
Multimodular Origami Polyhedra Archimedeans, Buckyballs and Duality	40 254 266
Nanoparticle- and Microparticle-based Delivery Systems	60 876 300
Nature and Numbers A Mathematical Photo Shooting	261 182 072
Optical Devices in Ophthalmology and Optometry	39 457 909
Riemann	228 812 626
The Mobius Strip	27 948 451
Topology Illustrated	110 474 753
Transport of Energetic Electrons in Solids	6 048 201
Pi-The-Next-Generation.pdf	67 268 059

Figure 2.92: Books-69-2020-202005122055.png

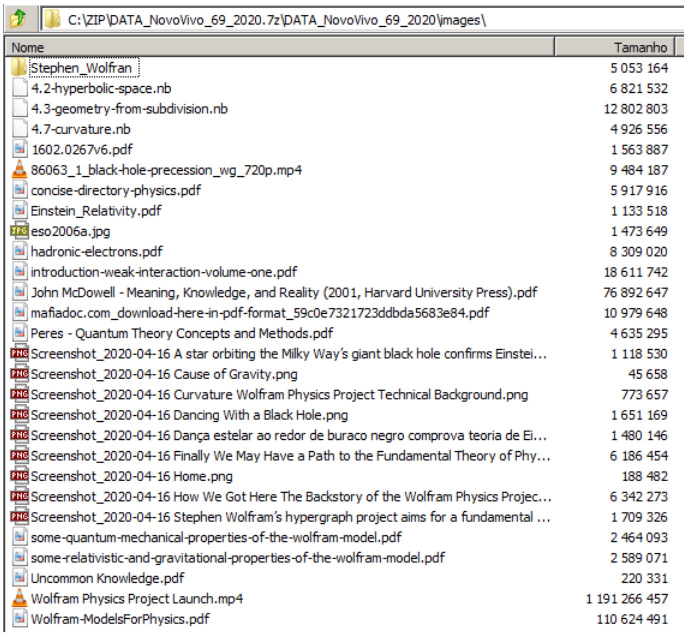

C:\ZIP\DATA_NovoVivo_69_2020.7z\DATA_NovoVivo_69_2020\images\

Nome	Tamanho
Stephen_Wolfran	5 053 164
4.2-hyperbolic-space.nb	6 821 532
4.3-geometry-from-subdivision.nb	12 802 803
4.7-curvature.nb	4 926 556
1602.0267v6.pdf	1 563 887
86063_1_black-hole-precession_wg_720p.mp4	9 484 187
concise-directory-physics.pdf	5 917 916
Einstein_Relativity.pdf	1 133 518
eso2006a.jpg	1 473 649
hadronic-electrons.pdf	8 309 020
introduction-weak-interaction-volume-one.pdf	18 611 742
John McDowell - Meaning, Knowledge, and Reality (2001, Harvard University Press).pdf	76 892 647
mafiadoc.com_download-here-in-pdf-format_59c0e7321723ddbda5683e84.pdf	10 979 648
Peres - Quantum Theory Concepts and Methods.pdf	4 635 295
Screenshot_2020-04-16 A star orbiting the Milky Way's giant black hole confirms Einstei...	1 118 530
Screenshot_2020-04-16 Cause of Gravity.png	45 658
Screenshot_2020-04-16 Curvature Wolfram Physics Project Technical Background.png	773 657
Screenshot_2020-04-16 Dancing With a Black Hole.png	1 651 169
Screenshot_2020-04-16 Dança estelar ao redor de buraco negro comprova teoria de Ei...	1 480 146
Screenshot_2020-04-16 Finally We May Have a Path to the Fundamental Theory of Phy...	6 186 454
Screenshot_2020-04-16 Home.png	188 482
Screenshot_2020-04-16 How We Got Here The Backstory of the Wolfram Physics Projec...	6 342 273
Screenshot_2020-04-16 Stephen Wolfram's hypergraph project aims for a fundamental ...	1 709 326
some-quantum-mechanical-properties-of-the-wolfram-model.pdf	2 464 093
some-relativistic-and-gravitational-properties-of-the-wolfram-model.pdf	2 589 071
Uncommon Knowledge.pdf	220 331
Wolfram Physics Project Launch.mp4	1 191 266 457
Wolfram-ModelsForPhysics.pdf	110 624 491

Figure 2.93: Books-69-2020-202005122056.png

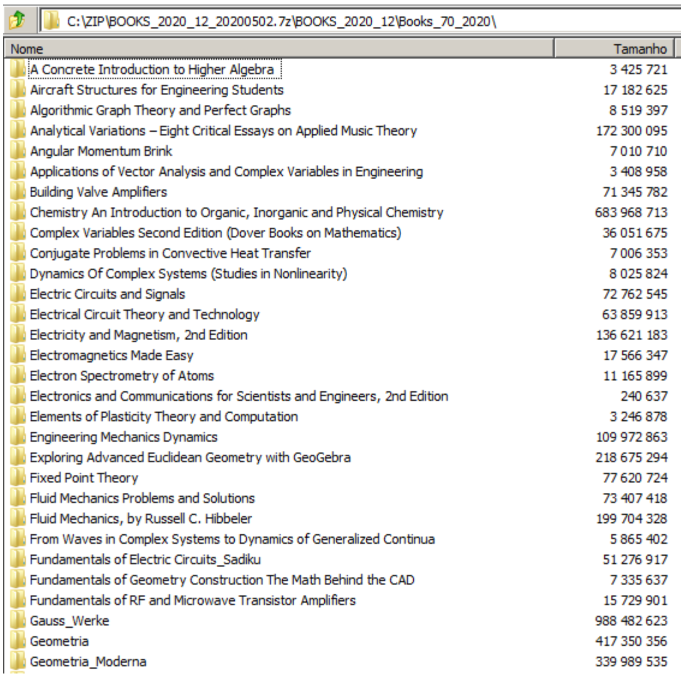

C:\ZIP\BOOKS_2020_12_20200502.7z\BOOKS_2020_12\Books_70_2020\

Nome	Tamanho
A Concrete Introduction to Higher Algebra	3 425 721
Aircraft Structures for Engineering Students	17 182 625
Algorithmic Graph Theory and Perfect Graphs	8 519 397
Analytical Variations – Eight Critical Essays on Applied Music Theory	172 300 095
Angular Momentum Brink	7 010 710
Applications of Vector Analysis and Complex Variables in Engineering	3 408 958
Building Valve Amplifiers	71 345 782
Chemistry An Introduction to Organic, Inorganic and Physical Chemistry	683 968 713
Complex Variables Second Edition (Dover Books on Mathematics)	36 051 675
Conjugate Problems in Convective Heat Transfer	7 006 353
Dynamics Of Complex Systems (Studies in Nonlinearity)	8 025 824
Electric Circuits and Signals	72 762 545
Electrical Circuit Theory and Technology	63 859 913
Electricity and Magnetism, 2nd Edition	136 621 183
Electromagnetics Made Easy	17 566 347
Electron Spectrometry of Atoms	11 165 899
Electronics and Communications for Scientists and Engineers, 2nd Edition	240 637
Elements of Plasticity Theory and Computation	3 246 878
Engineering Mechanics Dynamics	109 972 863
Exploring Advanced Euclidean Geometry with GeoGebra	218 675 294
Fixed Point Theory	77 620 724
Fluid Mechanics Problems and Solutions	73 407 418
Fluid Mechanics, by Russell C. Hibbeler	199 704 328
From Waves in Complex Systems to Dynamics of Generalized Continua	5 865 402
Fundamentals of Electric Circuits_Sadiku	51 276 917
Fundamentals of Geometry Construction The Math Behind the CAD	7 335 637
Fundamentals of RF and Microwave Transistor Amplifiers	15 729 901
Gauss_Werke	988 482 623
Geometria	417 350 356
Geometria_Moderna	339 989 535

Figure 2.94: Books-70-2020-202005121543.png

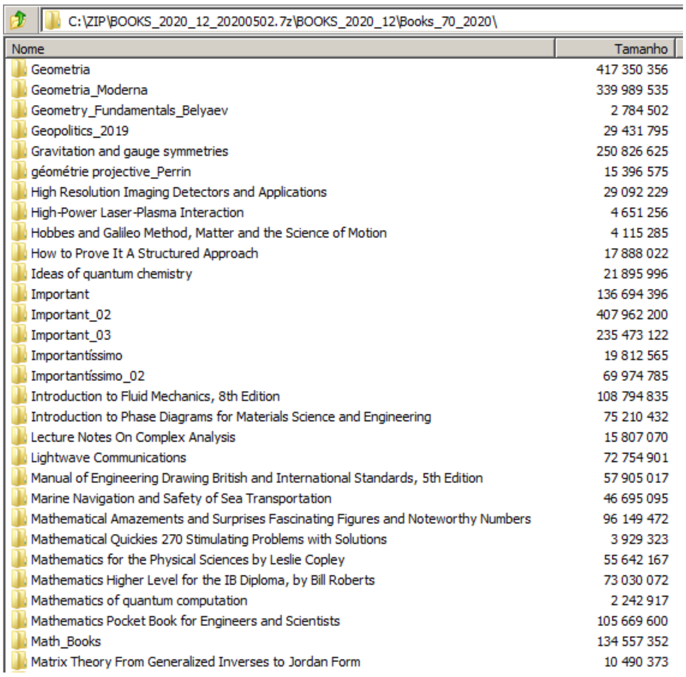

C:\ZIP\BOOKS_2020_12_20200502.7z\BOOKS_2020_12\Books_70_2020\

Nome	Tamanho
Geometria	417 350 356
Geometria_Moderna	339 989 535
Geometry_Fundamentals_Belyaev	2 784 502
Geopolitics_2019	29 431 795
Gravitation and gauge symmetries	250 826 625
géométrie projective_Perrin	15 396 575
High Resolution Imaging Detectors and Applications	29 092 229
High-Power Laser-Plasma Interaction	4 651 256
Hobbes and Galileo Method, Matter and the Science of Motion	4 115 285
How to Prove It A Structured Approach	17 888 022
Ideas of quantum chemistry	21 895 996
Important	136 694 396
Important_02	407 962 200
Important_03	235 473 122
Importantíssimo	19 812 565
Importantíssimo_02	69 974 785
Introduction to Fluid Mechanics, 8th Edition	108 794 835
Introduction to Phase Diagrams for Materials Science and Engineering	75 210 432
Lecture Notes On Complex Analysis	15 807 070
Lightwave Communications	72 754 901
Manual of Engineering Drawing British and International Standards, 5th Edition	57 905 017
Marine Navigation and Safety of Sea Transportation	46 695 095
Mathematical Amazements and Surprises Fascinating Figures and Noteworthy Numbers	96 149 472
Mathematical Quickies 270 Stimulating Problems with Solutions	3 929 323
Mathematics for the Physical Sciences by Leslie Copley	55 642 167
Mathematics Higher Level for the IB Diploma, by Bill Roberts	73 030 072
Mathematics of quantum computation	2 242 917
Mathematics Pocket Book for Engineers and Scientists	105 669 600
Math_Books	134 557 352
Matrix Theory From Generalized Inverses to Jordan Form	10 490 373

Figure 2.95: Books-70-2020-202005121544.png

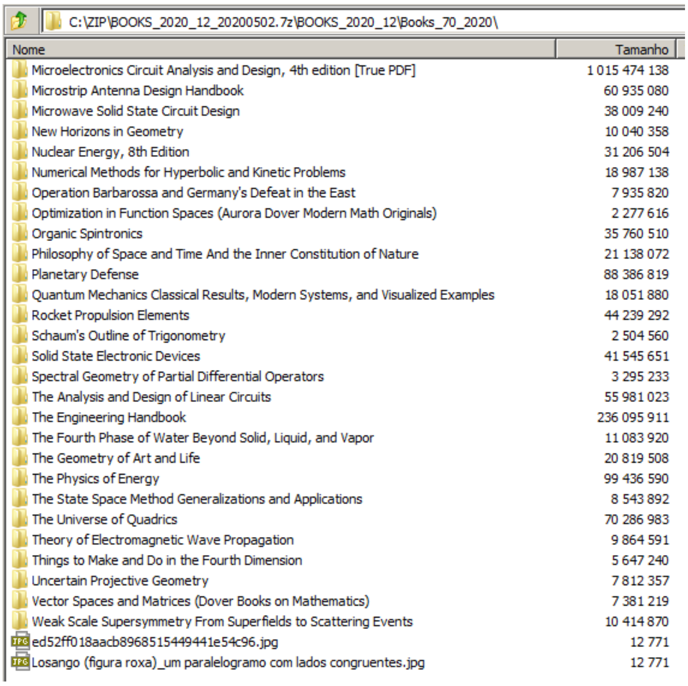

C:\ZIP\BOOKS_2020_12_20200502.7z\BOOKS_2020_12\Books_70_2020\

Nome	Tamanho
Microelectronics Circuit Analysis and Design, 4th edition [True PDF]	1 015 474 138
Microstrip Antenna Design Handbook	60 935 080
Microwave Solid State Circuit Design	38 009 240
New Horizons in Geometry	10 040 358
Nuclear Energy, 8th Edition	31 206 504
Numerical Methods for Hyperbolic and Kinetic Problems	18 987 138
Operation Barbarossa and Germany's Defeat in the East	7 935 820
Optimization in Function Spaces (Aurora Dover Modern Math Originals)	2 277 616
Organic Spintronics	35 760 510
Philosophy of Space and Time And the Inner Constitution of Nature	21 138 072
Planetary Defense	88 386 819
Quantum Mechanics Classical Results, Modern Systems, and Visualized Examples	18 051 880
Rocket Propulsion Elements	44 239 292
Schaum's Outline of Trigonometry	2 504 560
Solid State Electronic Devices	41 545 651
Spectral Geometry of Partial Differential Operators	3 295 233
The Analysis and Design of Linear Circuits	55 981 023
The Engineering Handbook	236 095 911
The Fourth Phase of Water Beyond Solid, Liquid, and Vapor	11 083 920
The Geometry of Art and Life	20 819 508
The Physics of Energy	99 436 590
The State Space Method Generalizations and Applications	8 543 892
The Universe of Quadrics	70 286 983
Theory of Electromagnetic Wave Propagation	9 864 591
Things to Make and Do in the Fourth Dimension	5 647 240
Uncertain Projective Geometry	7 812 357
Vector Spaces and Matrices (Dover Books on Mathematics)	7 381 219
Weak Scale Supersymmetry From Superfields to Scattering Events	10 414 870
ed52ff018aacb8968515449441e54c96.jpg	12 771
Losango (figura roxa)_um paralelogramo com lados congruentes.jpg	12 771

Figure 2.96: Books-70-2020-202005121545.png

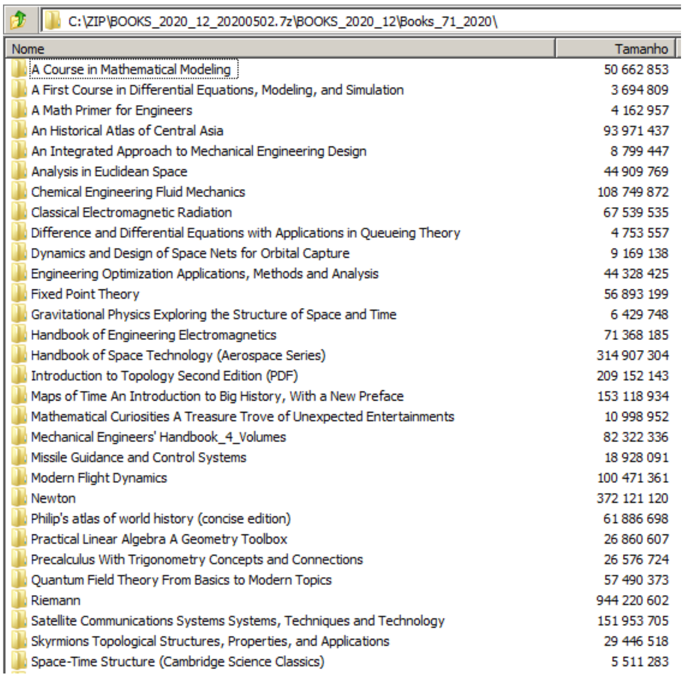

C:\ZIP\BOOKS_2020_12_20200502.7z\BOOKS_2020_12\Books_71_2020\

Nome	Tamanho
A Course in Mathematical Modeling	50 662 853
A First Course in Differential Equations, Modeling, and Simulation	3 694 809
A Math Primer for Engineers	4 162 957
An Historical Atlas of Central Asia	93 971 437
An Integrated Approach to Mechanical Engineering Design	8 799 447
Analysis in Euclidean Space	44 909 769
Chemical Engineering Fluid Mechanics	108 749 872
Classical Electromagnetic Radiation	67 539 535
Difference and Differential Equations with Applications in Queueing Theory	4 753 557
Dynamics and Design of Space Nets for Orbital Capture	9 169 138
Engineering Optimization Applications, Methods and Analysis	44 328 425
Fixed Point Theory	56 893 199
Gravitational Physics Exploring the Structure of Space and Time	6 429 748
Handbook of Engineering Electromagnetics	71 368 185
Handbook of Space Technology (Aerospace Series)	314 907 304
Introduction to Topology Second Edition (PDF)	209 152 143
Maps of Time An Introduction to Big History, With a New Preface	153 118 934
Mathematical Curiosities A Treasure Trove of Unexpected Entertainments	10 998 952
Mechanical Engineers' Handbook_4_Volumes	82 322 336
Missile Guidance and Control Systems	18 928 091
Modern Flight Dynamics	100 471 361
Newton	372 121 120
Philip's atlas of world history (concise edition)	61 886 698
Practical Linear Algebra A Geometry Toolbox	26 860 607
Precalculus With Trigonometry Concepts and Connections	26 576 724
Quantum Field Theory From Basics to Modern Topics	57 490 373
Riemann	944 220 602
Satellite Communications Systems Systems, Techniques and Technology	151 953 705
Skyrmions Topological Structures, Properties, and Applications	29 446 518
Space-Time Structure (Cambridge Science Classics)	5 511 283

Figure 2.97: Books-71-2020-202005121544.png

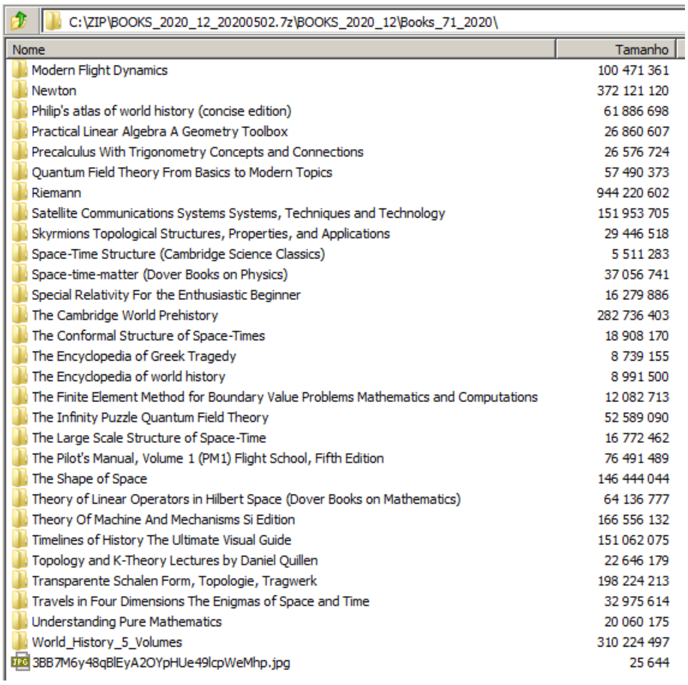

C:\ZIP\BOOKS_2020_12_20200502.7z\BOOKS_2020_12\Books_71_2020\

Nome	Tamanho
Modern Flight Dynamics	100 471 361
Newton	372 121 120
Philip's atlas of world history (concise edition)	61 886 698
Practical Linear Algebra A Geometry Toolbox	26 860 607
Precalculus With Trigonometry Concepts and Connections	26 576 724
Quantum Field Theory From Basics to Modern Topics	57 490 373
Riemann	944 220 602
Satellite Communications Systems Systems, Techniques and Technology	151 953 705
Skyrmions Topological Structures, Properties, and Applications	29 446 518
Space-Time Structure (Cambridge Science Classics)	5 511 283
Space-time-matter (Dover Books on Physics)	37 056 741
Special Relativity For the Enthusiastic Beginner	16 279 886
The Cambridge World Prehistory	282 736 403
The Conformal Structure of Space-Times	18 908 170
The Encyclopedia of Greek Tragedy	8 739 155
The Encyclopedia of world history	8 991 500
The Finite Element Method for Boundary Value Problems Mathematics and Computations	12 082 713
The Infinity Puzzle Quantum Field Theory	52 589 090
The Large Scale Structure of Space-Time	16 772 462
The Pilot's Manual, Volume 1 (PM1) Flight School, Fifth Edition	76 491 489
The Shape of Space	146 444 044
Theory of Linear Operators in Hilbert Space (Dover Books on Mathematics)	64 136 777
Theory Of Machine And Mechanisms Si Edition	166 556 132
Timelines of History The Ultimate Visual Guide	151 062 075
Topology and K-Theory Lectures by Daniel Quillen	22 646 179
Transparente Schalen Form, Topologie, Tragwerk	198 224 213
Travels in Four Dimensions The Enigmas of Space and Time	32 975 614
Understanding Pure Mathematics	20 060 175
World_History_5_Volumes	310 224 497
3BB7M6y48qBlEyA2OYpHUe49lcpWeMhp.jpg	25 644

Figure 2.98: Books-71-2020-202005121545.png

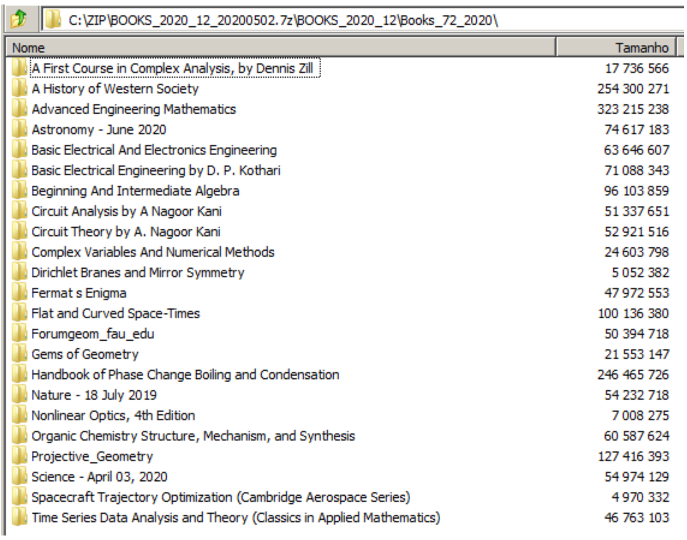

C:\ZIP\BOOKS_2020_12_20200502.7z\BOOKS_2020_12\Books_72_2020\

Nome	Tamanho
A First Course in Complex Analysis, by Dennis Zill	17 736 566
A History of Western Society	254 300 271
Advanced Engineering Mathematics	323 215 238
Astronomy - June 2020	74 617 183
Basic Electrical And Electronics Engineering	63 646 607
Basic Electrical Engineering by D. P. Kothari	71 088 343
Beginning And Intermediate Algebra	96 103 859
Circuit Analysis by A Nagoor Kani	51 337 651
Circuit Theory by A. Nagoor Kani	52 921 516
Complex Variables And Numerical Methods	24 603 798
Dirichlet Branes and Mirror Symmetry	5 052 382
Fermat s Enigma	47 972 553
Flat and Curved Space-Times	100 136 380
Forumgeom_fau_edu	50 394 718
Gems of Geometry	21 553 147
Handbook of Phase Change Boiling and Condensation	246 465 726
Nature - 18 July 2019	54 232 718
Nonlinear Optics, 4th Edition	7 008 275
Organic Chemistry Structure, Mechanism, and Synthesis	60 587 624
Projective_Geometry	127 416 393
Science - April 03, 2020	54 974 129
Spacecraft Trajectory Optimization (Cambridge Aerospace Series)	4 970 332
Time Series Data Analysis and Theory (Classics in Applied Mathematics)	46 763 103

Figure 2.99: Books-72-2020-202005121545.png

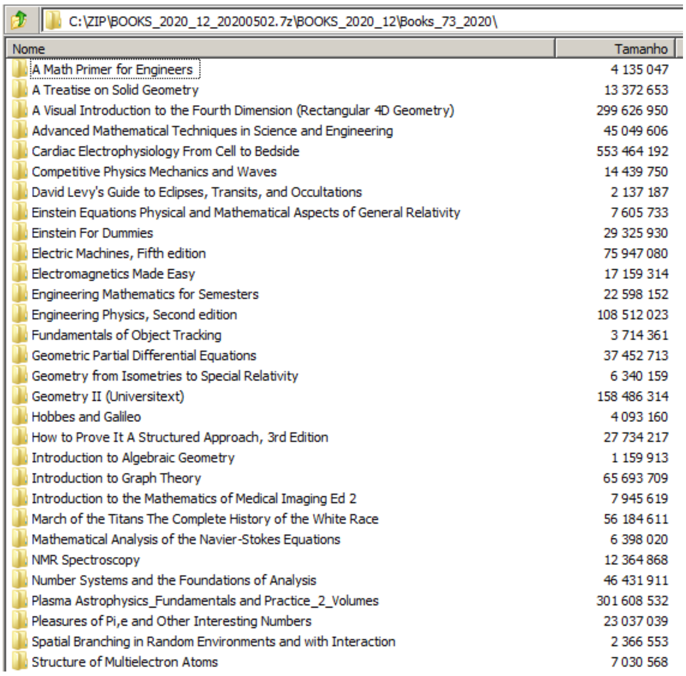

C:\ZIP\BOOKS_2020_12_20200502.7z\BOOKS_2020_12\Books_73_2020\

Nome	Tamanho
A Math Primer for Engineers	4 135 047
A Treatise on Solid Geometry	13 372 653
A Visual Introduction to the Fourth Dimension (Rectangular 4D Geometry)	299 626 950
Advanced Mathematical Techniques in Science and Engineering	45 049 606
Cardiac Electrophysiology From Cell to Bedside	553 464 192
Competitive Physics Mechanics and Waves	14 439 750
David Levy's Guide to Eclipses, Transits, and Occultations	2 137 187
Einstein Equations Physical and Mathematical Aspects of General Relativity	7 605 733
Einstein For Dummies	29 325 930
Electric Machines, Fifth edition	75 947 080
Electromagnetics Made Easy	17 159 314
Engineering Mathematics for Semesters	22 598 152
Engineering Physics, Second edition	108 512 023
Fundamentals of Object Tracking	3 714 361
Geometric Partial Differential Equations	37 452 713
Geometry from Isometries to Special Relativity	6 340 159
Geometry II (Universitext)	158 486 314
Hobbes and Galileo	4 093 160
How to Prove It A Structured Approach, 3rd Edition	27 734 217
Introduction to Algebraic Geometry	1 159 913
Introduction to Graph Theory	65 693 709
Introduction to the Mathematics of Medical Imaging Ed 2	7 945 619
March of the Titans The Complete History of the White Race	56 184 611
Mathematical Analysis of the Navier-Stokes Equations	6 398 020
NMR Spectroscopy	12 364 868
Number Systems and the Foundations of Analysis	46 431 911
Plasma Astrophysics_Fundamentals and Practice_2_Volumes	301 608 532
Pleasures of Pi,e and Other Interesting Numbers	23 037 039
Spatial Branching in Random Environments and with Interaction	2 366 553
Structure of Multielectron Atoms	7 030 568

Figure 2.100: Books-73-2020-202005121545.png

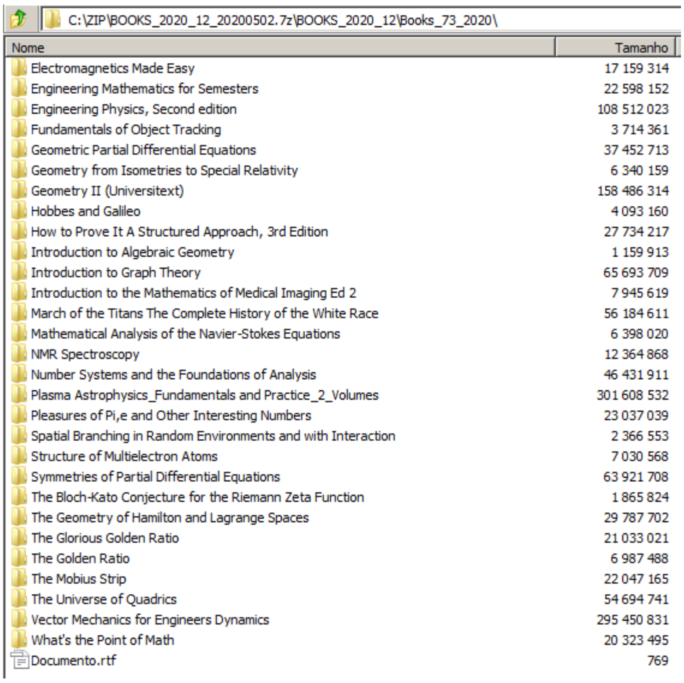

C:\ZIP\BOOKS_2020_12_20200502.7z\BOOKS_2020_12\Books_73_2020\

Nome	Tamanho
Electromagnetics Made Easy	17 159 314
Engineering Mathematics for Semesters	22 598 152
Engineering Physics, Second edition	108 512 023
Fundamentals of Object Tracking	3 714 361
Geometric Partial Differential Equations	37 452 713
Geometry from Isometries to Special Relativity	6 340 159
Geometry II (Universitext)	158 486 314
Hobbes and Galileo	4 093 160
How to Prove It A Structured Approach, 3rd Edition	27 734 217
Introduction to Algebraic Geometry	1 159 913
Introduction to Graph Theory	65 693 709
Introduction to the Mathematics of Medical Imaging Ed 2	7 945 619
March of the Titans The Complete History of the White Race	56 184 611
Mathematical Analysis of the Navier-Stokes Equations	6 398 020
NMR Spectroscopy	12 364 868
Number Systems and the Foundations of Analysis	46 431 911
Plasma Astrophysics_Fundamentals and Practice_2_Volumes	301 608 532
Pleasures of Pi,e and Other Interesting Numbers	23 037 039
Spatial Branching in Random Environments and with Interaction	2 366 553
Structure of Multielectron Atoms	7 030 568
Symmetries of Partial Differential Equations	63 921 708
The Bloch-Kato Conjecture for the Riemann Zeta Function	1 865 824
The Geometry of Hamilton and Lagrange Spaces	29 787 702
The Glorious Golden Ratio	21 033 021
The Golden Ratio	6 987 488
The Mobius Strip	22 047 165
The Universe of Quadrics	54 694 741
Vector Mechanics for Engineers Dynamics	295 450 831
What's the Point of Math	20 323 495
Documento.rtf	769

Figure 2.101: Books-73-2020-202005121546.png

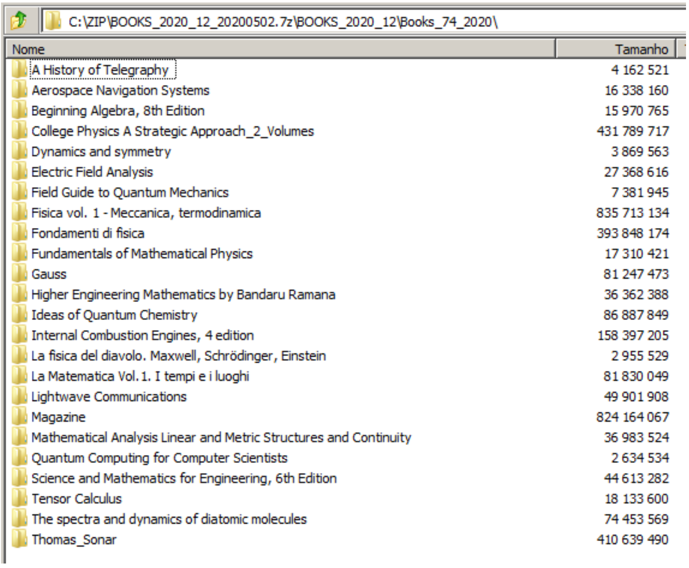

C:\ZIP\BOOKS_2020_12_20200502.7z\BOOKS_2020_12\Books_74_2020\

Nome	Tamanho
A History of Telegraphy	4 162 521
Aerospace Navigation Systems	16 338 160
Beginning Algebra, 8th Edition	15 970 765
College Physics A Strategic Approach_2_Volumes	431 789 717
Dynamics and symmetry	3 869 563
Electric Field Analysis	27 368 616
Field Guide to Quantum Mechanics	7 381 945
Fisica vol. 1 - Meccanica, termodinamica	835 713 134
Fondamenti di fisica	393 848 174
Fundamentals of Mathematical Physics	17 310 421
Gauss	81 247 473
Higher Engineering Mathematics by Bandaru Ramana	36 362 388
Ideas of Quantum Chemistry	86 887 849
Internal Combustion Engines, 4 edition	158 397 205
La fisica del diavolo. Maxwell, Schrödinger, Einstein	2 955 529
La Matematica Vol. 1. I tempi e i luoghi	81 830 049
Lightwave Communications	49 901 908
Magazine	824 164 067
Mathematical Analysis Linear and Metric Structures and Continuity	36 983 524
Quantum Computing for Computer Scientists	2 634 534
Science and Mathematics for Engineering, 6th Edition	44 613 282
Tensor Calculus	18 133 600
The spectra and dynamics of diatomic molecules	74 453 569
Thomas_Sonar	410 639 490

Figure 2.102: Books-74-2020-202005121546.png

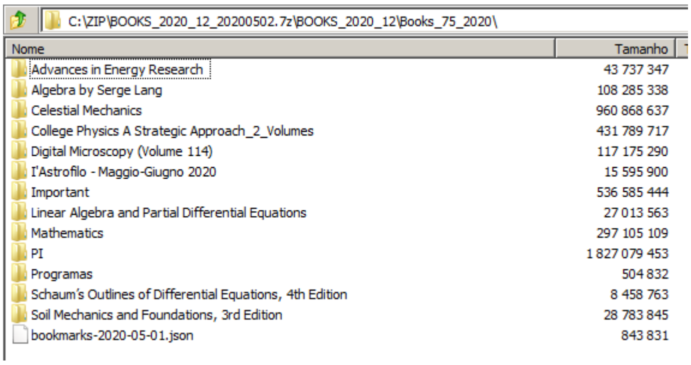

C:\ZIP\BOOKS_2020_12_20200502.7z\BOOKS_2020_12\Books_75_2020\

Nome	Tamanho
Advances in Energy Research	43 737 347
Algebra by Serge Lang	108 285 338
Celestial Mechanics	960 868 637
College Physics A Strategic Approach_2_Volumes	431 789 717
Digital Microscopy (Volume 114)	117 175 290
I'Astrofilo - Maggio-Giugno 2020	15 595 900
Important	536 585 444
Linear Algebra and Partial Differential Equations	27 013 563
Mathematics	297 105 109
PI	1 827 079 453
Programas	504 832
Schaum's Outlines of Differential Equations, 4th Edition	8 458 763
Soil Mechanics and Foundations, 3rd Edition	28 783 845
bookmarks-2020-05-01.json	843 831

Figure 2.103: Books-75-2020-202005121546.png

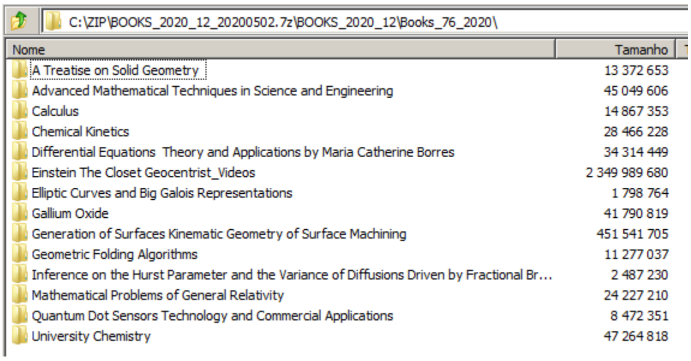

C:\ZIP\BOOKS_2020_12_20200502.7z\BOOKS_2020_12\Books_76_2020\

Nome	Tamanho
A Treatise on Solid Geometry	13 372 653
Advanced Mathematical Techniques in Science and Engineering	45 049 606
Calculus	14 867 353
Chemical Kinetics	28 466 228
Differential Equations Theory and Applications by Maria Catherine Borres	34 314 449
Einstein The Closet Geocentrist_Videos	2 349 989 680
Elliptic Curves and Big Galois Representations	1 798 764
Gallium Oxide	41 790 819
Generation of Surfaces Kinematic Geometry of Surface Machining	451 541 705
Geometric Folding Algorithms	11 277 037
Inference on the Hurst Parameter and the Variance of Diffusions Driven by Fractional Br...	2 487 230
Mathematical Problems of General Relativity	24 227 210
Quantum Dot Sensors Technology and Commercial Applications	8 472 351
University Chemistry	47 264 818

Figure 2.104: Books-76-2020-202005121546.png

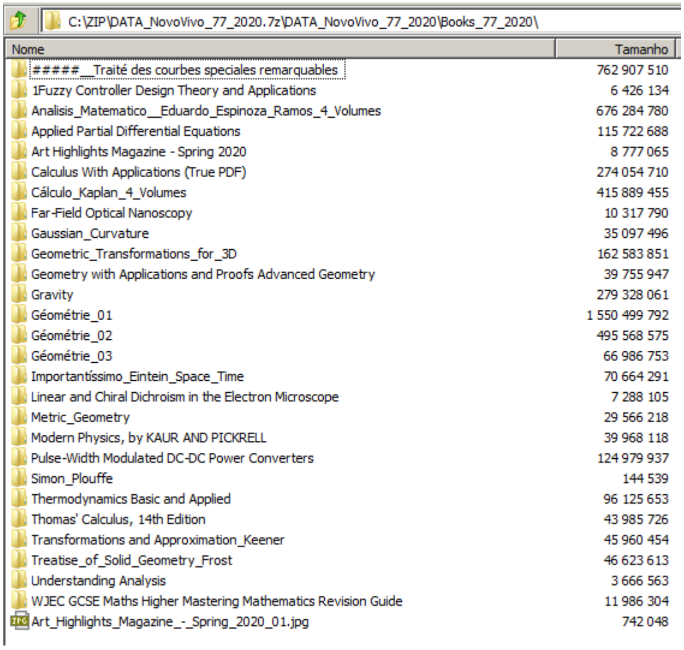
C:\ZIP\DATA_NovoVivo_77_2020.7z\DATA_NovoVivo_77_2020\Books_77_2020\

Nome	Tamanho
#####__Traité des courbes speciales remarquables	762 907 510
1Fuzzy Controller Design Theory and Applications	6 426 134
Analisis_Matematico__Eduardo_Espinoza_Ramos_4_Volumes	676 284 780
Applied Partial Differential Equations	115 722 688
Art Highlights Magazine - Spring 2020	8 777 065
Calculus With Applications (True PDF)	274 054 710
Cálculo_Kaplan_4_Volumes	415 889 455
Far-Field Optical Nanoscopy	10 317 790
Gaussian_Curvature	35 097 496
Geometric_Transformations_for_3D	162 583 851
Geometry with Applications and Proofs Advanced Geometry	39 755 947
Gravity	279 328 061
Géométrie_01	1 550 499 792
Géométrie_02	495 568 575
Géométrie_03	66 986 753
Importantíssimo_Eintein_Space_Time	70 664 291
Linear and Chiral Dichroism in the Electron Microscope	7 288 105
Metric_Geometry	29 566 218
Modern Physics, by KAUR AND PICKRELL	39 968 118
Pulse-Width Modulated DC-DC Power Converters	124 979 937
Simon_Plouffe	144 539
Thermodynamics Basic and Applied	96 125 653
Thomas' Calculus, 14th Edition	43 985 726
Transformations and Approximation_Keener	45 960 454
Treatise_of_Solid_Geometry_Frost	46 623 613
Understanding Analysis	3 666 563
WJEC GCSE Maths Higher Mastering Mathematics Revision Guide	11 986 304
Art_Highlights_Magazine_-_Spring_2020_01.jpg	742 048

Figure 2.105: Books-76-2020-202005122116.png

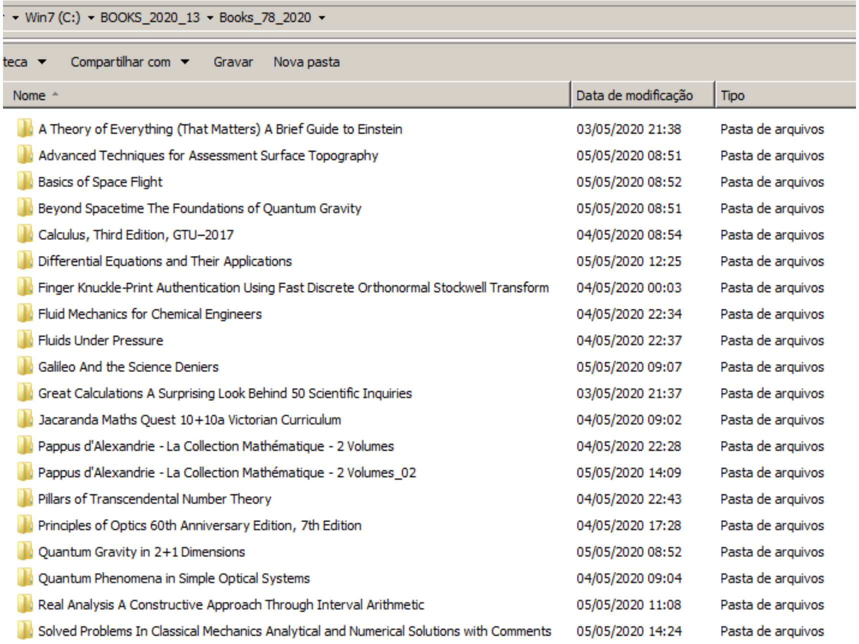

Nome	Data de modificação	Tipo
A Theory of Everything (That Matters) A Brief Guide to Einstein	03/05/2020 21:38	Pasta de arquivos
Advanced Techniques for Assessment Surface Topography	05/05/2020 08:51	Pasta de arquivos
Basics of Space Flight	05/05/2020 08:52	Pasta de arquivos
Beyond Spacetime The Foundations of Quantum Gravity	05/05/2020 08:51	Pasta de arquivos
Calculus, Third Edition, GTU–2017	04/05/2020 08:54	Pasta de arquivos
Differential Equations and Their Applications	05/05/2020 12:25	Pasta de arquivos
Finger Knuckle-Print Authentication Using Fast Discrete Orthonormal Stockwell Transform	04/05/2020 00:03	Pasta de arquivos
Fluid Mechanics for Chemical Engineers	04/05/2020 22:34	Pasta de arquivos
Fluids Under Pressure	04/05/2020 22:37	Pasta de arquivos
Galileo And the Science Deniers	05/05/2020 09:07	Pasta de arquivos
Great Calculations A Surprising Look Behind 50 Scientific Inquiries	03/05/2020 21:37	Pasta de arquivos
Jacaranda Maths Quest 10+10a Victorian Curriculum	04/05/2020 09:02	Pasta de arquivos
Pappus d'Alexandrie - La Collection Mathématique - 2 Volumes	04/05/2020 22:28	Pasta de arquivos
Pappus d'Alexandrie - La Collection Mathématique - 2 Volumes_02	05/05/2020 14:09	Pasta de arquivos
Pillars of Transcendental Number Theory	04/05/2020 22:43	Pasta de arquivos
Principles of Optics 60th Anniversary Edition, 7th Edition	04/05/2020 17:28	Pasta de arquivos
Quantum Gravity in 2+1 Dimensions	05/05/2020 08:52	Pasta de arquivos
Quantum Phenomena in Simple Optical Systems	04/05/2020 09:04	Pasta de arquivos
Real Analysis A Constructive Approach Through Interval Arithmetic	05/05/2020 11:08	Pasta de arquivos
Solved Problems In Classical Mechanics Analytical and Numerical Solutions with Comments	05/05/2020 14:24	Pasta de arquivos

Figure 2.106: Books-78-2020-202005111224.png

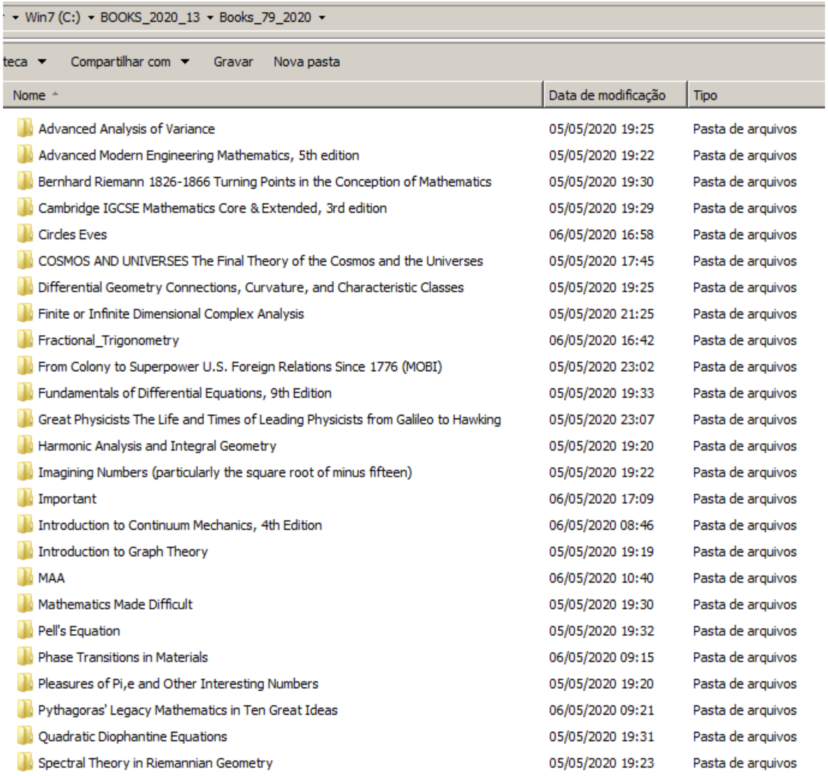

Nome	Data de modificação	Tipo
Advanced Analysis of Variance	05/05/2020 19:25	Pasta de arquivos
Advanced Modern Engineering Mathematics, 5th edition	05/05/2020 19:22	Pasta de arquivos
Bernhard Riemann 1826-1866 Turning Points in the Conception of Mathematics	05/05/2020 19:30	Pasta de arquivos
Cambridge IGCSE Mathematics Core & Extended, 3rd edition	05/05/2020 19:29	Pasta de arquivos
Circles Eves	06/05/2020 16:58	Pasta de arquivos
COSMOS AND UNIVERSES The Final Theory of the Cosmos and the Universes	05/05/2020 17:45	Pasta de arquivos
Differential Geometry Connections, Curvature, and Characteristic Classes	05/05/2020 19:25	Pasta de arquivos
Finite or Infinite Dimensional Complex Analysis	05/05/2020 21:25	Pasta de arquivos
Fractional_Trigonometry	06/05/2020 16:42	Pasta de arquivos
From Colony to Superpower U.S. Foreign Relations Since 1776 (MOBI)	05/05/2020 23:02	Pasta de arquivos
Fundamentals of Differential Equations, 9th Edition	05/05/2020 19:33	Pasta de arquivos
Great Physicists The Life and Times of Leading Physicists from Galileo to Hawking	05/05/2020 23:07	Pasta de arquivos
Harmonic Analysis and Integral Geometry	05/05/2020 19:20	Pasta de arquivos
Imagining Numbers (particularly the square root of minus fifteen)	05/05/2020 19:22	Pasta de arquivos
Important	06/05/2020 17:09	Pasta de arquivos
Introduction to Continuum Mechanics, 4th Edition	06/05/2020 08:46	Pasta de arquivos
Introduction to Graph Theory	05/05/2020 19:19	Pasta de arquivos
MAA	06/05/2020 10:40	Pasta de arquivos
Mathematics Made Difficult	05/05/2020 19:30	Pasta de arquivos
Pell's Equation	05/05/2020 19:32	Pasta de arquivos
Phase Transitions in Materials	06/05/2020 09:15	Pasta de arquivos
Pleasures of Pi,e and Other Interesting Numbers	05/05/2020 19:20	Pasta de arquivos
Pythagoras' Legacy Mathematics in Ten Great Ideas	06/05/2020 09:21	Pasta de arquivos
Quadratic Diophantine Equations	05/05/2020 19:31	Pasta de arquivos
Spectral Theory in Riemannian Geometry	05/05/2020 19:23	Pasta de arquivos

Figure 2.107: Books-79-2020-202005111227.png

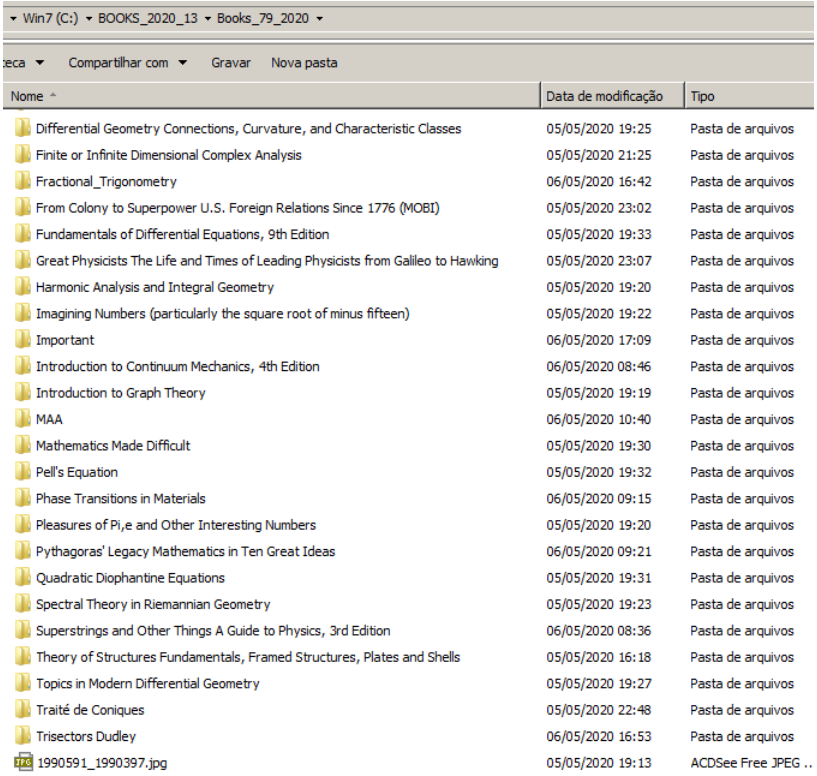

Nome	Data de modificação	Tipo
Differential Geometry Connections, Curvature, and Characteristic Classes	05/05/2020 19:25	Pasta de arquivos
Finite or Infinite Dimensional Complex Analysis	05/05/2020 21:25	Pasta de arquivos
Fractional_Trigonometry	06/05/2020 16:42	Pasta de arquivos
From Colony to Superpower U.S. Foreign Relations Since 1776 (MOBI)	05/05/2020 23:02	Pasta de arquivos
Fundamentals of Differential Equations, 9th Edition	05/05/2020 19:33	Pasta de arquivos
Great Physicists The Life and Times of Leading Physicists from Galileo to Hawking	05/05/2020 23:07	Pasta de arquivos
Harmonic Analysis and Integral Geometry	05/05/2020 19:20	Pasta de arquivos
Imagining Numbers (particularly the square root of minus fifteen)	05/05/2020 19:22	Pasta de arquivos
Important	06/05/2020 17:09	Pasta de arquivos
Introduction to Continuum Mechanics, 4th Edition	06/05/2020 08:46	Pasta de arquivos
Introduction to Graph Theory	05/05/2020 19:19	Pasta de arquivos
MAA	06/05/2020 10:40	Pasta de arquivos
Mathematics Made Difficult	05/05/2020 19:30	Pasta de arquivos
Pell's Equation	05/05/2020 19:32	Pasta de arquivos
Phase Transitions in Materials	06/05/2020 09:15	Pasta de arquivos
Pleasures of Pi,e and Other Interesting Numbers	05/05/2020 19:20	Pasta de arquivos
Pythagoras' Legacy Mathematics in Ten Great Ideas	06/05/2020 09:21	Pasta de arquivos
Quadratic Diophantine Equations	05/05/2020 19:31	Pasta de arquivos
Spectral Theory in Riemannian Geometry	05/05/2020 19:23	Pasta de arquivos
Superstrings and Other Things A Guide to Physics, 3rd Edition	06/05/2020 08:36	Pasta de arquivos
Theory of Structures Fundamentals, Framed Structures, Plates and Shells	05/05/2020 16:18	Pasta de arquivos
Topics in Modern Differential Geometry	05/05/2020 19:27	Pasta de arquivos
Traité de Coniques	05/05/2020 22:48	Pasta de arquivos
Trisectors Dudley	06/05/2020 16:53	Pasta de arquivos
1990591_1990397.jpg	05/05/2020 19:13	ACDSee Free JPEG ..

Figure 2.108: Books-79-2020-202005111228.png

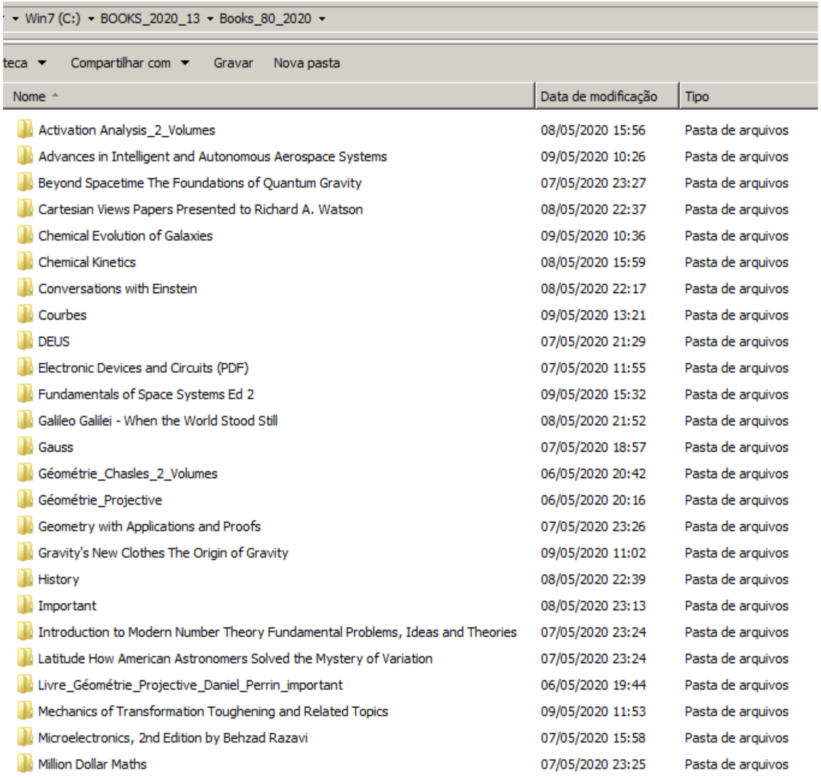

Nome	Data de modificação	Tipo
Activation Analysis_2_Volumes	08/05/2020 15:56	Pasta de arquivos
Advances in Intelligent and Autonomous Aerospace Systems	09/05/2020 10:26	Pasta de arquivos
Beyond Spacetime The Foundations of Quantum Gravity	07/05/2020 23:27	Pasta de arquivos
Cartesian Views Papers Presented to Richard A. Watson	08/05/2020 22:37	Pasta de arquivos
Chemical Evolution of Galaxies	09/05/2020 10:36	Pasta de arquivos
Chemical Kinetics	08/05/2020 15:59	Pasta de arquivos
Conversations with Einstein	08/05/2020 22:17	Pasta de arquivos
Courbes	09/05/2020 13:21	Pasta de arquivos
DEUS	07/05/2020 21:29	Pasta de arquivos
Electronic Devices and Circuits (PDF)	07/05/2020 11:55	Pasta de arquivos
Fundamentals of Space Systems Ed 2	09/05/2020 15:32	Pasta de arquivos
Galileo Galilei - When the World Stood Still	08/05/2020 21:52	Pasta de arquivos
Gauss	07/05/2020 18:57	Pasta de arquivos
Géométrie_Chasles_2_Volumes	06/05/2020 20:42	Pasta de arquivos
Géométrie_Projective	06/05/2020 20:16	Pasta de arquivos
Geometry with Applications and Proofs	07/05/2020 23:26	Pasta de arquivos
Gravity's New Clothes The Origin of Gravity	09/05/2020 11:02	Pasta de arquivos
History	08/05/2020 22:39	Pasta de arquivos
Important	08/05/2020 23:13	Pasta de arquivos
Introduction to Modern Number Theory Fundamental Problems, Ideas and Theories	07/05/2020 23:24	Pasta de arquivos
Latitude How American Astronomers Solved the Mystery of Variation	07/05/2020 23:24	Pasta de arquivos
Livre_Géométrie_Projective_Daniel_Perrin_important	06/05/2020 19:44	Pasta de arquivos
Mechanics of Transformation Toughening and Related Topics	09/05/2020 11:53	Pasta de arquivos
Microelectronics, 2nd Edition by Behzad Razavi	07/05/2020 15:58	Pasta de arquivos
Million Dollar Maths	07/05/2020 23:25	Pasta de arquivos

Figure 2.109: Books-80-2020-202005111229.png

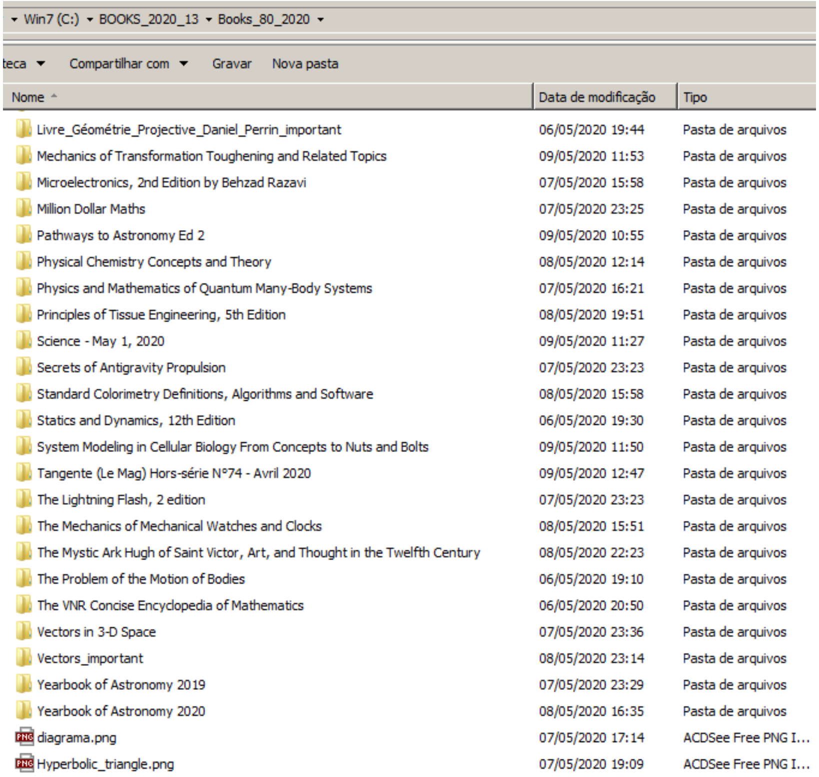

Nome	Data de modificação	Tipo
Livre_Géométrie_Projective_Daniel_Perrin_important	06/05/2020 19:44	Pasta de arquivos
Mechanics of Transformation Toughening and Related Topics	09/05/2020 11:53	Pasta de arquivos
Microelectronics, 2nd Edition by Behzad Razavi	07/05/2020 15:58	Pasta de arquivos
Million Dollar Maths	07/05/2020 23:25	Pasta de arquivos
Pathways to Astronomy Ed 2	09/05/2020 10:55	Pasta de arquivos
Physical Chemistry Concepts and Theory	08/05/2020 12:14	Pasta de arquivos
Physics and Mathematics of Quantum Many-Body Systems	07/05/2020 16:21	Pasta de arquivos
Principles of Tissue Engineering, 5th Edition	08/05/2020 19:51	Pasta de arquivos
Science - May 1, 2020	09/05/2020 11:27	Pasta de arquivos
Secrets of Antigravity Propulsion	07/05/2020 23:23	Pasta de arquivos
Standard Colorimetry Definitions, Algorithms and Software	08/05/2020 15:58	Pasta de arquivos
Statics and Dynamics, 12th Edition	06/05/2020 19:30	Pasta de arquivos
System Modeling in Cellular Biology From Concepts to Nuts and Bolts	09/05/2020 11:50	Pasta de arquivos
Tangente (Le Mag) Hors-série N°74 - Avril 2020	09/05/2020 12:47	Pasta de arquivos
The Lightning Flash, 2 edition	07/05/2020 23:23	Pasta de arquivos
The Mechanics of Mechanical Watches and Clocks	08/05/2020 15:51	Pasta de arquivos
The Mystic Ark Hugh of Saint Victor, Art, and Thought in the Twelfth Century	08/05/2020 22:23	Pasta de arquivos
The Problem of the Motion of Bodies	06/05/2020 19:10	Pasta de arquivos
The VNR Concise Encyclopedia of Mathematics	06/05/2020 20:50	Pasta de arquivos
Vectors in 3-D Space	07/05/2020 23:36	Pasta de arquivos
Vectors_important	08/05/2020 23:14	Pasta de arquivos
Yearbook of Astronomy 2019	07/05/2020 23:29	Pasta de arquivos
Yearbook of Astronomy 2020	08/05/2020 16:35	Pasta de arquivos
diagrama.png	07/05/2020 17:14	ACDSee Free PNG I...
Hyperbolic_triangle.png	07/05/2020 19:09	ACDSee Free PNG I...

Figure 2.110: Books-80-2020-202005111230.png

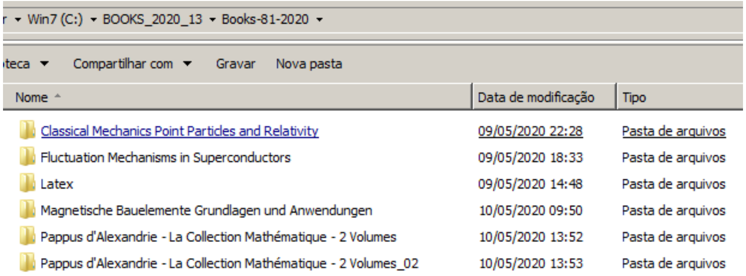

Nome	Data de modificação	Tipo
Classical Mechanics Point Particles and Relativity	09/05/2020 22:28	Pasta de arquivos
Fluctuation Mechanisms in Superconductors	09/05/2020 18:33	Pasta de arquivos
Latex	09/05/2020 14:48	Pasta de arquivos
Magnetische Bauelemente Grundlagen und Anwendungen	10/05/2020 09:50	Pasta de arquivos
Pappus d'Alexandrie - La Collection Mathématique - 2 Volumes	10/05/2020 13:52	Pasta de arquivos
Pappus d'Alexandrie - La Collection Mathématique - 2 Volumes_02	10/05/2020 13:53	Pasta de arquivos

Figure 2.111: Books-81-2020-202005111232.png

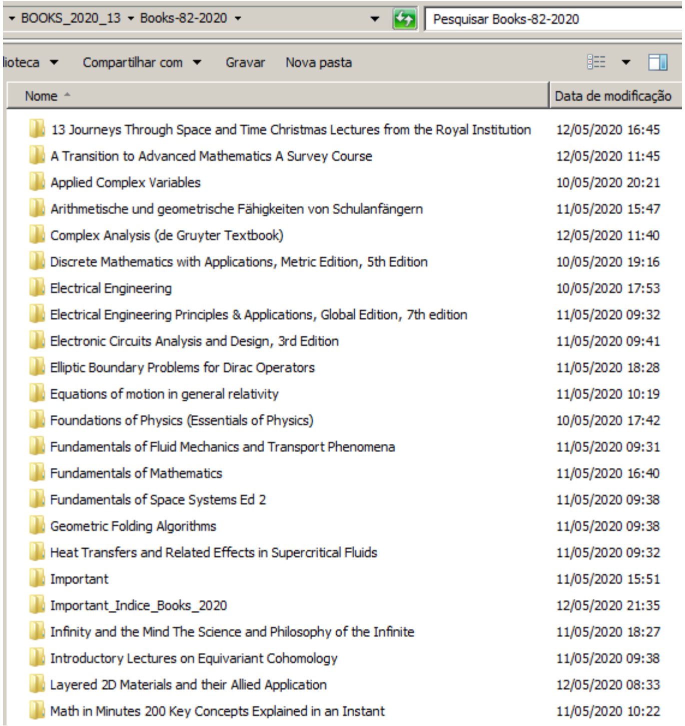

Nome	Data de modificação
13 Journeys Through Space and Time Christmas Lectures from the Royal Institution	12/05/2020 16:45
A Transition to Advanced Mathematics A Survey Course	12/05/2020 11:45
Applied Complex Variables	10/05/2020 20:21
Arithmetische und geometrische Fähigkeiten von Schulanfängern	11/05/2020 15:47
Complex Analysis (de Gruyter Textbook)	12/05/2020 11:40
Discrete Mathematics with Applications, Metric Edition, 5th Edition	10/05/2020 19:16
Electrical Engineering	10/05/2020 17:53
Electrical Engineering Principles & Applications, Global Edition, 7th edition	11/05/2020 09:32
Electronic Circuits Analysis and Design, 3rd Edition	11/05/2020 09:41
Elliptic Boundary Problems for Dirac Operators	11/05/2020 18:28
Equations of motion in general relativity	11/05/2020 10:19
Foundations of Physics (Essentials of Physics)	10/05/2020 17:42
Fundamentals of Fluid Mechanics and Transport Phenomena	11/05/2020 09:31
Fundamentals of Mathematics	11/05/2020 16:40
Fundamentals of Space Systems Ed 2	11/05/2020 09:38
Geometric Folding Algorithms	11/05/2020 09:38
Heat Transfers and Related Effects in Supercritical Fluids	11/05/2020 09:32
Important	11/05/2020 15:51
Important_Indice_Books_2020	12/05/2020 21:35
Infinity and the Mind The Science and Philosophy of the Infinite	11/05/2020 18:27
Introductory Lectures on Equivariant Cohomology	11/05/2020 09:38
Layered 2D Materials and their Allied Application	12/05/2020 08:33
Math in Minutes 200 Key Concepts Explained in an Instant	11/05/2020 10:22

Figure 2.112: Books-82-2020-202005130822.png

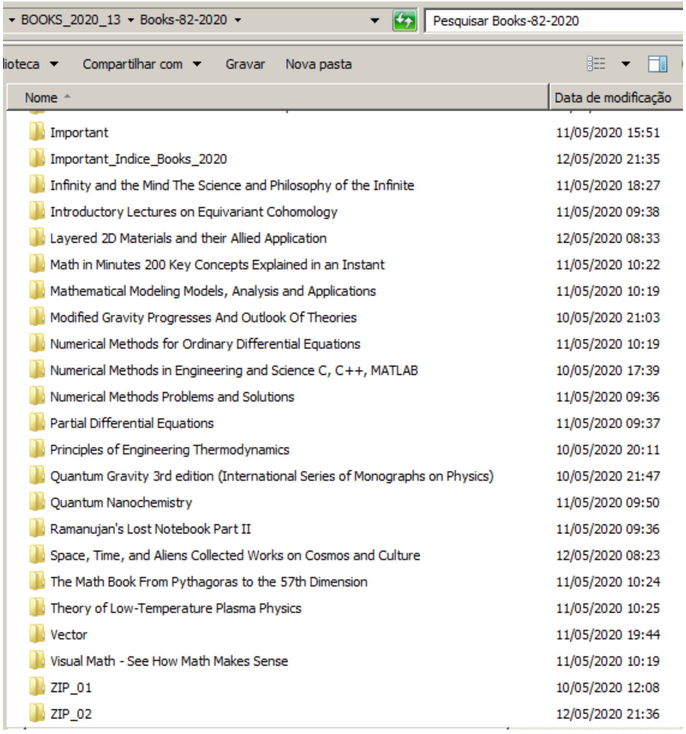

Nome	Data de modificação
Important	11/05/2020 15:51
Important_Indice_Books_2020	12/05/2020 21:35
Infinity and the Mind The Science and Philosophy of the Infinite	11/05/2020 18:27
Introductory Lectures on Equivariant Cohomology	11/05/2020 09:38
Layered 2D Materials and their Allied Application	12/05/2020 08:33
Math in Minutes 200 Key Concepts Explained in an Instant	11/05/2020 10:22
Mathematical Modeling Models, Analysis and Applications	11/05/2020 10:19
Modified Gravity Progresses And Outlook Of Theories	10/05/2020 21:03
Numerical Methods for Ordinary Differential Equations	11/05/2020 10:19
Numerical Methods in Engineering and Science C, C++, MATLAB	10/05/2020 17:39
Numerical Methods Problems and Solutions	11/05/2020 09:36
Partial Differential Equations	11/05/2020 09:37
Principles of Engineering Thermodynamics	10/05/2020 20:11
Quantum Gravity 3rd edition (International Series of Monographs on Physics)	10/05/2020 21:47
Quantum Nanochemistry	11/05/2020 09:50
Ramanujan's Lost Notebook Part II	11/05/2020 09:36
Space, Time, and Aliens Collected Works on Cosmos and Culture	12/05/2020 08:23
The Math Book From Pythagoras to the 57th Dimension	11/05/2020 10:24
Theory of Low-Temperature Plasma Physics	11/05/2020 10:25
Vector	11/05/2020 19:44
Visual Math - See How Math Makes Sense	11/05/2020 10:19
ZIP_01	10/05/2020 12:08
ZIP_02	12/05/2020 21:36

Figure 2.113: Books-82-2020-202005130823.png

2.4 Spherical Geometry

Source: Books-82-2020 Date: 20200511

1) C:/DATA-NovoVivo-82-2020/Books-82-2020/Important/Spherical Geometry and Its Applications/sanet st 0367180464.pdf

Obs.: Livro todo!

Pag. 298 (313) - 6 Introduction to Riemannian Geometry

Figura 6.9 - Geometric interpretation of the torsion tensor.

Figure 6.10 - Geometric interpretation of the curvature tensor.IMPORTANT

2) C:/DATA-NovoVivo-82-2020/Books-82-2020/Important/Spherical Geometry and Its Applications/sanet.st 0367196905.pdf

Pag. 06 (19) - Lemma 1.4 - Figure 1.1

3) C:/DATA-NovoVivo-82-2020/Books-82-2020/Important/Spherical Geometry and Its Applications/[Lokenath-Debnath,-Piotr-Mikusinski]-Introduction-Hilbert(z-lib.org).pdf Obs.: Hilbert Spaces with Applications

4) C:/DATA-NovoVivo-82-2020/Books-82-2020/Important/Spherical Geometry and Its Applications/[Christian-B r]-Elementary-Differential-Geometry(z-lib.org).pdf

5) C:/DATA-NovoVivo-82-2020/Books-82-2020/Important/Spherical Geometry and Its Applications/Differential Geometry and Analysis on CR Manifolds (2006, Birkhauser Basel).pdf

6) C:/DATA-NovoVivo-82-2020/Books-82-2020/Important/Spherical Geometry and Its Applications/[Titu-Andreescu,-Oleg-Mushkarov,-Luchezar-Stoyanov(z-lib.org).pdf

7) C:/DATA-NovoVivo-82-2020/Books-82-2020/Important/Spherical Geometry and Its Applications/[Joaquim-M.-Domingos]-Geometrical-Properties-of-Ve(z-lib.org).pdf

2.5 Calculus of Variations

Source: DATA-NovoVivo-83-2020 Date: 20200513

1) C:/DATA-NovoVivo-83-2020/Books-83-2020;Variational Methods in Nonlinear Analysis With Applications in Optimization and Partial Differential Equations/Variational Methods in Nonlinear Analysis.pdf

Pag. 249 (263) 6 The Calculus of Variations

2) C:/DATA-NovoVivo-83-2020/Books-83-2020/Variational Methods in Nonlinear Analysis With Applications in Optimization and Partial Differential Equations/3030313506.pdf

Pag. 415 (426) - Chapter 15 Geometric Metrics for Topological Representations

3) C:/DATA-NovoVivo-83-2020/Books-83-2020/Variational Methods in Nonlinear Analysis With Applications in Optimization and Partial Differential Equations/3540740120.pdf IMPORTANT

Pag. 220 (238) - 5. The Dirichlet Problem for the Equation of Constant Mean Curvature

4) C:/DATA-NovoVivo-83-2020/Books-83-2020/Important/Quantum Riemannian Geometry (2020, Springer) - libgen.lc.pdf Obs.: Livro todo!!! 826 pages.

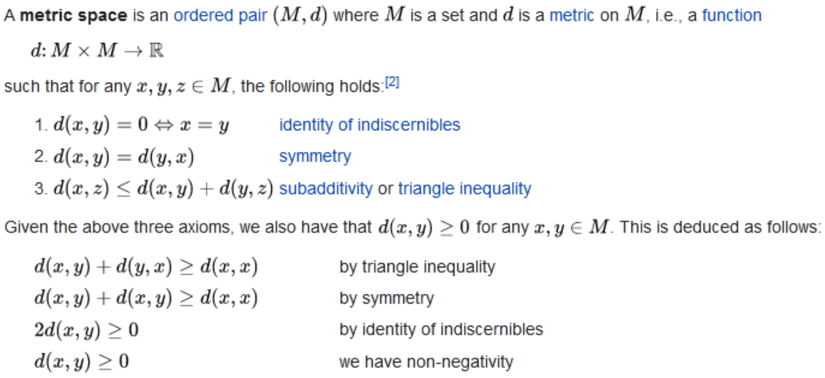

A **metric space** is an ordered pair (M, d) where M is a set and d is a metric on M, i.e., a function

$$d\colon M \times M \to \mathbb{R}$$

such that for any $x, y, z \in M$, the following holds:[2]

1. $d(x,y) = 0 \Leftrightarrow x = y$ identity of indiscernibles
2. $d(x,y) = d(y,x)$ symmetry
3. $d(x,z) \leq d(x,y) + d(y,z)$ subadditivity or triangle inequality

Given the above three axioms, we also have that $d(x,y) \geq 0$ for any $x, y \in M$. This is deduced as follows:

$d(x,y) + d(y,x) \geq d(x,x)$	by triangle inequality
$d(x,y) + d(x,y) \geq d(x,x)$	by symmetry
$2d(x,y) \geq 0$	by identity of indiscernibles
$d(x,y) \geq 0$	we have non-negativity

Figure 2.114: Metric Space - Wiki.png

2.6 Metric Space

1) Metric space From Wikipedia, the free encyclopedia Jump to navigation Jump to search

In mathematics, a metric space is a set together with a metric on the set. The metric is a function that defines a concept of distance between any two members of the set, which are usually called points. The metric satisfies a few simple properties. Informally:

the distance from a point to itself is zero, the distance between two distinct points is positive, the distance from A to B is the same as the distance from B to A, and the distance from A to B (directly) is less than or equal to the distance from A to B via any third point C.

A metric on a space induces topological properties like open and closed sets, which lead to the study of more abstract topological spaces.

The most familiar metric space is 3-dimensional Euclidean space. In fact, a "metric" is the generalization of the Euclidean metric arising from the four long-known properties of the Euclidean distance. The Euclidean metric defines the distance between two points as the length of the straight line segment connecting them. Other metric spaces occur for example in elliptic geometry and hyperbolic geometry, where distance on a sphere measured by angle is a metric, and the hyperboloid model of hyperbolic geometry is used by special relativity as a metric space of velocities.

In 1906 Maurice Frechet introduced metric spaces in his work Sur quelques points du calcul fonctionnel.[1] However the name is due to Felix Hausdorff. Definition

The function d is also called distance function or simply distance. Often, d is omitted and one just writes M for a metric space if it is clear from the context what metric is used.

Ignoring mathematical details, for any system of roads and terrains the distance between two locations can be defined as the length of the shortest route connecting those locations. To be a metric there shouldn't be any one-way roads. The triangle inequality expresses the fact that detours aren't shortcuts. If the distance between two points is zero, the two points are indistinguishable from one-another. Many of the examples below can be seen as concrete versions of this general idea.

2) Metric space mathematics Written By:
Stephan C. Carlson
See Article History

Metric space, in mathematics, especially topology, an abstract set with a distance function, called a metric, that specifies a nonnegative distance between any two of its points in such a way that the following properties hold: (1) the distance from the first point to the second equals zero if and only if the points are the same, (2) the distance from the first point to the second equals the distance from the second to the first, and (3) the sum of the distance from the first point to the second and the distance from the second point to a third exceeds or equals the distance from the first to the third. The last of these properties is called the triangle inequality. The French mathematician Maurice Frechet initiated the study of metric spaces in 1905.

The usual distance function on the real number line is a metric, as is the usual distance function in Euclidean n-dimensional space. There are also more exotic examples of interest to mathematicians. Given any set of points, the discrete metric specifies that the distance from a point to itself equal 0 while the distance between any two distinct points equal 1. The so-called taxicab metric on the Euclidean plane, this \taxicab distance" gives the minimum length of a path from x, y to z, w constructed from horizontal and vertical line segments. In analysis there are several useful metrics on sets of bounded real-valued continuous or integrable functions.

Thus, a metric generalizes the notion of usual distance to more general settings. Moreover, a metric on a set X determines a collection of open sets, or topology, on X when a subset U of X is declared to be open if and only if for each point p of X there is a positive (possibly very small) distance r such that the set of all points of X of distance less than r from p is completely contained in U. In this way metric spaces provide important examples of topological spaces.

A metric space is said to be complete if every sequence of points in which

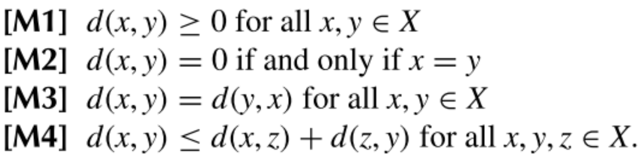

[M1] $d(x, y) \geq 0$ for all $x, y \in X$
[M2] $d(x, y) = 0$ if and only if $x = y$
[M3] $d(x, y) = d(y, x)$ for all $x, y \in X$
[M4] $d(x, y) \leq d(x, z) + d(z, y)$ for all $x, y, z \in X$.

Figure 2.115: Metric Space - Sacks.png

the terms are eventually pairwise arbitrarily close to each other (a so-called Cauchy sequence) converges to a point in the metric space. The usual metric on the rational numbers is not complete since some Cauchy sequences of rational numbers do not converge to rational numbers. For example, the rational number sequence 3, 3.1, 3.14 ... converges to n, which is not a rational number. However, the usual metric on the real numbers is complete, and, moreover, every real number is the limit of a Cauchy sequence of rational numbers. In this sense, the real numbers form the completion of the rational numbers. The proof of this fact, given in 1914 by the German mathematician Felix Hausdorff, can be generalized to demonstrate that every metric space has such a completion. IMPORTANT

3) https://www.sciencedirect.com/topics/mathematics/metric-space

Metric Spaces

Paul Sacks, in Techniques of Functional Analysis for Differential and Integral Equations, 2017 3.1 Axioms of a Metric Space

The idea of a metric space is that of a set on which some natural notion of distance may be defined. The formal definition is as follows.

Definition 3.1

A metric space is a pair (X, d) where X is a set and d is a real valued mapping on X x X, such that the following axioms hold.

Source Book:

C:/DATA-NovoVivo-84-2020/Books-84-2020/Matric-Space/Techniques of Functional Analysis for Differential and Integral Equations (2017, Academic Press Elsevier) - libgen.lc.pdf

Pag. 35 (44)

2.7 Curvature A Variational Approach

- Diretorio Todo

Source: ARQ-BASICS/02/

1) C:/ARQ BASICS/02/Curvature A Variational Approach/[sixth edition] Joseph A. Wolf - Spaces of Constant Curvature (2011, AMS).pdf Obs.: Livro todo! 439 pages. 2) C:/DATA-NovoVivo-84-2020/Books-84-2020/Curvature

A Variational Approach 02/Geometry II Spaces of Constant Curvature (1993, Springer-Verlag Berlin Heidelberg) - libgen.lc.pdf

CONTENTS

Prefaces vii, x
Notes to the Reader xi
fart I Riemannian Geometry 1
chapter I Affine Differential Geometry 1
I. I Differentiable manifolds 2
1.2 Vector fields 3
1.3 Differential forms 5
1.4 Maps 7
1.5 Lie groups 9
1.6 The frame bundle: parallelism and geodesics. 14
1.7 Curvature, torsion and the structure equations 23
1.8 Covering spaces 31
1.9 The Cartan-Ambrose-Hicks Theorem 42
chapter 2 Riemannian Curvature 45
2.1 The Levi-Civita connection 46
2.2 Sectional curvature 52
2.3 Isometries and curvature 57
2.4 Models for spaces of constant curvature 62
2.5 The 2-dimensional space forms 74
2.6 Finite rotation groups 83
2.7 Homogeneous space forms 88
2.8 Appendix: The metric space structure of a riemannian manifold 91
part II The Euclidean Space Form Problem 97
chapter 3 Flat Riemannian Manifolds . 98
3.1 Discontinuous groups on euclidean space 98
3.2 The Bieberbach Theorems on crystallographic groups 100
3.3 Application to euclidean space forms 105
3.4 Questions of holonomy 107

3.5 Three dimensional euclidean space forms III
3.6 Three attacks on the classification problem for flat compact manifolds 124
3.7 Flat homogeneous pseudo-riemannian manifolds 131
part III The Spherical Space Form Problem 137
chapter 4 Representations of Finite Groups 138
4.1 Basic definitions 138
4.2 The Frobenius-Schur relations 139
4.3 Frobenius reciprocity and the group algebra 141
4.4 Divisibility . 145
4.5 Tensor products and dual representations 147
4.6 Two lemmas on representations over algebraically non-closed fields 150
4.7 Unitary and orthogonal representations 151
chapter 5 Vincent's Work on the Spherical Space Form Problem 154
5.1 Vincent's program 154
5.2 Preliminaries on p-groups 156
5.3 Necessary conditions on fixed point free groups 159
5.4 Classification of the simplest type of fixed point free groups 162
5.5 Representations of finite groups in which every Sylow subgroup is cyclic 165
5.6 A partial solution to the spherical space form problem 171
chapter 6 The Classification of Fixed Point Free Groups 172
6.1 Zassenhaus' work on solvable groups with cyclic odd Sylow subgroups 173
6.2 The binary icosahedral group 181
6.3 Non-solvable fixed point free groups 195
chapter 7 The Solution to the Spherical Space Form Problem 198
7.1 Representations of binary polyhedral groups 198
7.2 Fixed point free complex representations . 203
7.3 The action of automorphisms on representations 211
7.4 The classification of spherical space forms 218
7.5 Spherical space forms of low dimension 224
7.6 Clifford translations 227
part IV Space Form Problems on Symmetric Spaces 231
chapter 8 Riemannian Symmetric Spaces 231
8.1 Lie formulation of locally symmetric spaces 232
8.2 Structure of orthogonal involutive Lie algebras 234
8.3 Globally symmetric spaces and orthogonal involutive Lie algebras 240
8.4 Curvature 245
8.5 Cohomology 247
8.6 Cartan subalgebras, rank and maximal tori 252
8.7 Hermitian symmetric spaces 257
8.8 The full group of isometries 263

8.9 Extended Schlafli-Dynkin diagrams 264
8.10 Subgroups of maximal rank 275
8.11 The classification of symmetric spaces 286
8.12 Two point homogeneous spaces 293
8.13 Appendix: Manifolds with irreducible linear isotropy group . 300
chapter 9 Space Forms of Irreducible Symmetric Spaces 303
9.1 Feasibility of space form problems 304
9.2 Grassmann manifolds as symmetric spaces 306
9.3 Grassmann manifolds of even dimension 307
9.4 Grassmann manifolds of odd dimension 314
9.5 Symmetric spaces of positive characteristic 319
9.6 An isolated manifold 325
chapter 10 Locally Symmetric Spaces of Non-negative Curvature 328
10.1 The structure theorems 329
10.2 Application of the structure theorems 333
part V Space Form Problems on Indefinite Metric Manifolds 337
chapter 11 Spaces of Constant Curvature . 337
11.1 The classification of finite space forms 338
11.2 The geometry of pseudo-spherical space forms 341
11.3 Homogeneous finite space forms 347
11.4 The lattice space forms 354
11.5 A wild Lorentz signature 366
11.6 The classification for homogeneous manifolds of constant curvature 370
chapter 12 Locally Isotropic Manifolds 374
12.1 Reductive Lie groups 374
12.2 Examples of locally isotropic manifolds 380
12.3 Structure of locally isotropic spaces 385
12.4 A partial classification of complete locally
isotropic manifolds 389
Appendix to Chapter 12 396
References 402
Additional References 408
Index 413

2) C:/ARQ-BASICS/02/Curvature A Variational Approach/Metric.pdf Obs.: Livro todo! 662 pages. Metric Spaces of Non-Positive Curvature

3) C:/ARQ-BASICS/02/Curvature A Variational Approach/[University Printing House, Cambridge, United Kingdom] Tiberiu Harko and Francisco Lobo - Extensions of f(R) Gravity Curvature-Matter Couplings and Hybrid

Metric-Palatini Theory (2018).pdf Obs. Livro Todo - 475 pages.

Part II Generalized Curvature-Matter Couplings in Modified Gravity
7 Introduction 179
8 Gravity Theories with Linear Curvature-Matter Coupling 186
8.1 Linear Nonminimal Curvature-Matter Coupling 186
8.2 The Matter Lagrangian and the Energy-Momentum Tensor 188
8.3 Curvature-Matter Coupling in the Palatini Formalism 194
8.4 Equivalence with an Anomalous Scalar{Tensor Theory 200
8.5 Further Theoretical Developments 201

4) C:/DATA-NovoVivo-84-2020/Books-84-2020/Curvature A Variational Approach 02/Geometry II Spaces of Constant Curvature (1993, Springer-Verlag Berlin Heidelberg) - libgen.lc.pdf

Contents

Preface 6
Chapter 1. Basic Structures 8
x 1. Definition of Spaces of Constant Curvature 8
1.1. Lie Groups of Transformations 8
1.2. Groups of Motions of a Riemannian Manifold 9
1.3. Invariant Riemannian Metrics on Homogeneous Spaces 10
1.4. Spaces of Constant Curvature 11
1.5. Three Spaces 11
1.6. Subspaces of the Space R n,1 14
x 2 The Classification Theorem 15
2.1. Statement of the Theorem 15
2.2. Reduction to Lie Algebras 16
2.3. The Symmetry 16
2.4. Structure of the Tangent Algebra of the Group of Motions 17
2.5. Riemann Space 19
x 3. Subspaces and Convexity 20
3.1. Involutions 20
3.2. Planes 0 21
3.3. Half-Spaces and Convex Sets 23
3.4. Orthogonal Planes
x 4. Metric 26
4.1. General Properties 26
4.2. Formulae for Distance in the Vector Model 26
4.3. Convexity of Distance 27

Chapter 2. Models of Lobachevskij Space 29
x 1. Projective Models 29
1.1. Homogeneous Domains 29
1.2. Projective Model of Lobachevskij Space 29
1.3. Projective Euclidean Models. The Klein Model 31
1.4. "Affine" Subgroup of the Group of Automorphisms of a
Quadric 32
1.5. Riemannian Metric and Distance Between Points in the
Projective Model 33
x 2. Conformal Models 36
2.1. Conformal Space 36
2.2. Conformal Model of the Lobachevskij Space 39
2.3. Conformal Euclidean Models 42
2.4. Complex Structure of the Lobachevskij Plane 46
x 3. Matrix Models of the Spaces JI 2 and JI 3
. . . . 47
3.1. Matrix Model of the Space JI2
. . . . 47
3.2. Matrix Model of the Space JI 3
. . . . 48
Chapter 3. Plane Geometry 50
x 1. Lines 51
1.1. Divergent and Parallel Lines on the Lobachevskij Plane 51
1.2. Distance Between Points of Different Lines 54
x 2. Polygons 56
2.1. Definitions. Conditions of Convexity 56
2.2. Elementary Properties of Triangles 57
2.3. Polar Triangles on the Sphere 59
2.4. The Sum of the Angles in a Triangle 59
2.5. Existence of a Convex Polygon with Given Angles 60
2.6. The Angular Excess and the Area of a Polygon 62
x 3. Metric Relations 63
3.1. The Cosine Law for a Triangle 63
3.2. Other Relations in a Triangle . . 65
3.3. The Angle of Parallelism 67
3.4. Relations in Quadrilaterals, Pentagons, and Hexagons. . 68
3.5. The Circumference and the Area of a Disk 69
x 4. Motions and Homogeneous Lines 70
4.1. Classification of Motions of Two-Dimensional Spaces of
Constant Curvature 70

4.2. Characterization of Motions of the Lobachevskij Plane in the Poincare Model in Terms of Traces of Matrices 71
4.3. One-Parameter Groups of Motions of the Lobachevskij Plane and Their Orbits 72
4.4. The Form of the Metric in Coordinates Related to a Pencil 75
Chapter 4. Planes, Spheres, Horospheres and Equidistant Surfaces 77
x 1. Relative Position of Planes 77
1.1. Pairs of Subspaces of a Euclidean Vector Space 77
1.2. Some General Notions 79
1.3. Pairs of Planes on the Sphere 79
1.4. Pairs of Planes in the Euclidean Space 80
1.5. Pseudo-orthogonal Transformations 80
1.6. Pairs of Hyperbolic Subspaces of the Space IR.n,l 81
1. 7. Pairs of Planes in the Lobachevskij Space 82
1.8. Pairs of Lines in the Lobachevskij Space . . . 82
1.9. Pairs of Hyperplanes 83
x 2. Standard Surfaces . . . 84
2.1. Definitions and Basic Facts 84
2.2. Standard Hypersurfaces 86
2.3. Similarity of Standard Hypersurfaces 87
2.4. Intersection of Standard Hypersurfaces 89
x 3. Decompositions into Semi-direct Products . . 90
3.1. Spherical, Horospherical, and Equidistant Decompositions 90
3.2. Spherical-equidistant Decomposition 90
Chapter 5. Motions 92
x 1. General Properties of Motions 92
1.1. Description of Motions 92
1.2. Continuation of a Plane Motion . . 92
1.3. Displacement Function 93
x 2. Classification of Motions 94
2.1. Motions of the Sphere 94
2.2. Motions of the Euclidean Space 95
2.3. Motions of the Lobachevskij Space 95
2.4. One-parameter Groups of Motions 97
x 3. Groups of Motions and Similarities . . . 98
3.1. Some Basic Notions 98
3.2. Criterion for the Existence of a Fixed Point 99
3.3. Groups of Motions of the Sphere 99
3.4. Groups of Motions of the Euclidean Space 100
3.5. Groups of Similarities 101
3.6. Groups of Motions of the Lobachevskij Space 101 4

Chapter 6. Acute-angled Polyhedra 104
x 1. Basic Properties of Acute-angled Polyhedra . . . 104
1.1. General Information on Convex Polyhedra 104
1.2. The Gram Matrix of a Convex Polyhedron 104
1.3. Acute-angled Families of Half-spaces and Acute-angled Polyhedra 106
1.4. Acute-angled Polyhedra on the Sphere and in Euclidean Space 107
1.5. Simplicity of Acute-angled Polyhedra 108
x 2. Acute-angled Polyhedra in Lobachevskij Space 108
2.1. Description in Terms of Gram Matrices 108
2.2. Combinatorial Structure 109
2.3. Description in Terms of Dihedral Angles . . 110
Chapter 7. Volumes 112
x 1. Volumes of Sectors and Wedges 113
1.1. Volumes of Sectors 113
1.2. Volume of a Hyperbolic Wedge 114
1.3. Volume of a Parabolic Wedge . . . 116
1.4. Volume of an Elliptic Wedge 116
x 2. Volumes of Polyhedra 117
2.1. Volume of a Simplex as an Analytic Function of the Dihedral Angles 117
2.2. Volume Differential 118
2.3. Volume of an Even-dimensional Polyhedron. The Poincare and Schlafli Formulae 120
2.4. Volume of an Even-dimensional Polyhedron. The Gauss-Bonnet Formula 122
x 3. Volumes of 3-dimensional Polyhedra 123
3.1. The Lobachevskij Function 123
3.2. Double-rectangular Tetrahedra 124
3.3. Volume of a Double-rectangular Tetrahedron. The Lobachevskij Formula 125
3.4. Volumes of Tetrahedra with Vertices at Infinity 126
3.5. Volume of a Pyramid with the Apex at Infinity 128
Chapter 8. Spaces of Constant Curvature as Riemannian Manifolds 130
1.1. Exponential Mapping 130
1.2. Parallel Translation 131
1.3. Curvature 131
1.4. Totally Geodesic Submanifolds 132
1.5. Hypersurfaces 133
1.6. Projective Properties 134

1. 7. Conformal Properties 134
1.8. Pseudo-Riemannian Spaces of Constant Curvature 135
References 135

II. Discrete Groups of Motions
of Spaces of Constant Curvature
E.B. Vinberg, O.V. Shvartsman
Translated from the Russian
by V. Minachin
Contents

Preface 143
Chapter 1. Introduction 144
x 1. Basic Notions 144
1.1. Definition of Discrete Groups of Motions 144
1.2. Quotient Spaces 145
1.3. Fundamental Domains 146
1.4. The Dirichlet Domain 148
1.5. Commensurable Groups 150
x 2. Origins of Discrete Groups of Motions 152
2.1. Symmetry Groups 152
2.2. Arithmetic Groups 156
2.3. Fundamental Groups of Space Forms 156
Chapter 2. Fundamental Domains 159
x 1. Description of a Discrete Group in Terms of a Fundamental Domain 159
1.1. Normalization of a Fundamental Domain 159
1.2. Generators 160
1.3. Defining Relations 161
1.4. Points of a Fundamental Domain at Infinity 162
1.5. The Existence of a Discrete Group with a Given Fundamental Polyhedron 164
1.6. Limitations of Poincare's Method 167
1. 7. Homogeneous Decompositions . . . 167
x 2. Geometrically Finite Groups of Motions of Lobachevskij Space 168
2.1. The Limit Set of a Discrete Group of Motions 169
2.2. Statement of the Main Results 169
2.3. Some General Properties of Discrete Groups of Motions of Lobachevskij Space 170
2.4. Compactification of the Quotient Space of a Crystallographic Group 171
2.5. A Criterion of Geometrical Finiteness 172

Chapter 3. Crystallographic Groups 174
x 1. The Schoenflies-Bieberbach Theorem 174
1.1. Statement of the Main Theorem 17 4
1.2. Commutators of Orthogonal Transformations and Motions . . 174
1.3. Proof of the Main Theorem 175
1.4. Arbitrary Discrete Groups of Motions of the Euclidean Space 176
x 2. Classification of Crystallographic Groups 176
2.1. Cohomological Description 176
2.2. Abstract Structure 177
2.3. Classification Steps 178
2.4. The Finiteness Theorem 179
2.5. Bravais Types 180
2.6. Some Classification Results 180
2.7. Euclidean Space Forms 181
x 3. Homogeneous Decompositions of Euclidean Space 182
3.1. The Finiteness Theorem 182
3.2. Parallelohedra 182
Chapter 4. Fuchsian Groups 183
x 1. Fuchsian Groups from the Topological Point of View . . . 183
1.1. Ramified Coverings 183
1.2. Signature and Uniformization of a Surface with Signature. Planar Groups 184
1.3. Planar Groups of Finite Type . . . 186
1.4 Algebraic Structure of Planar Groups 187
x 2. Geometry of Fuchsian Groups 189
2.1. Introducing a Metric 189
2.2. Nielsen’s Domain for a Fuchsian Group . . . 190
II. Discrete Groups of Motions of Spaces of Constant Curvature 141
x 3. The Teichmiiller Space of a Fuchsian Group 191
3.1. The Teichmiiller Space and the Moduli of an Abstract Group 191
3.2. The Teichmiiller Space of the Fundamental Group of a Closed Surface 192
3.3. Fuchsian Groups and Riemann Surfaces 196
3.4. Extensions of Fuchsian Groups. Maximal Fuchsian Groups .. 197
Chapter 5. Reflection Groups 198
x 1. Basic Notions and Theorems 198
1.1. Coxeter Polyhedra 198
1.2. Discrete Reflection Groups 200
1.3. Coxeter Schemes 200

1.4. Reflection Groups on the Sphere and in Euclidean Space 201
x 2. Reflection Groups in the Lobachevskij Space . . . 203
2 .1. General Properties 203
2.2. Crystallographic Reflection Groups in the Lobachevskij Plane and in the Three-dimensional Lobachevskij Space 204
2.3. Launer and Quasi-Lanner Groups 204
2.4. Some Other Examples 208
2.5. Restrictions on Dimension 210
x 3. Regular Polyhedra and Honeycombs 210
3.1. The Symmetry Group and the Schliifli Symbol of a Regular Polyhedron 210
3.2. Classification of Regular Polyhedra 212
3.3. Honeycombs 214
Chapter 6. Arithmetic Groups 217
x 1. Description of Arithmetic Discrete Groups of Motions of Lobachevskij Space 217
1.1. Arithmetic Groups of the Simplest Type . . 217
1.2. Quaternion Algebras 218
1.3. Arithmetic Fuchsian Groups 219
1.4. Arithmetic Groups of Motions of the Space JI 3 220
1.5. Arithmetic Groups of Motions of the Space Ji3 for n 2>=4 221
x 2. Reflective Arithmetic Groups 222
2.1. Application of Reflection Groups to the Study of Arithmetic Groups 222
2.2. Classification Problem for Reflective Quadratic Forms 225
x 3. Existence of Non-Arithmetic Groups 226
3.1. Arithmeticity Criterion for Reflection Groups 226
3.2. Existence of Non-Arithmetic Reflection Groups 227
3.3. Existence of Non-Arithmetic Crystallographic Groups in Any Dimension..228
Chapter 7. Sociology of Discrete Groups in the Lobachevskij Spaces .. 231

x1. Rigidity and Deformation 231
1.1. Theorem on Strong Rigidity 231
1.2. Deformations 232
1.3. Dehn-Thurston Surgery 235
x2. Commensurable Groups 238
2.1. Commensurator .... .238
2.2. Millson’s Property 238
x3. Covolumes 239

3.1. The Set of Covolumes 239
3.2. Discrete Groups of Minimal Covolume 240
References 241

2.8 Curvature of Plane Curve

Let a plane curve C be defined parametrically by the radius vector r(t). While a point M moves along the curve C, the direction of the tangent changes (Figure 1).

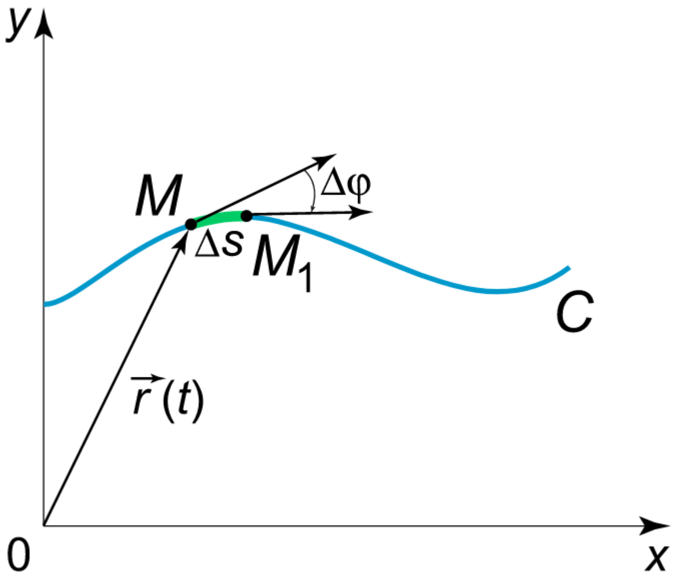

[!h]

Figure 2.116: Curvature Definition 1

Teste de imagem:

==========

The curvature of the curve can be defined as the ratio of the rotation angle of the tangent the traversed arc length In the image 2.116, observation of client,

$$\frac{\frac{a}{x\ \ y}+\frac{b}{x+y}}{1+\frac{a\ \ b}{a+b}} \tag{2.1}$$

Osculating (google tradutor) (of a curve or surface) touch (another curve or surface) so as to have a common tangent at the point of contact. "the plots have been drawn using osculating orbital elements";
Should the students be asked to sketch the osculating circle to a given curve at a given point, and then notice that as the point of contact changes, the radius of the circle is inversely proportional to the measure of curvature?; If

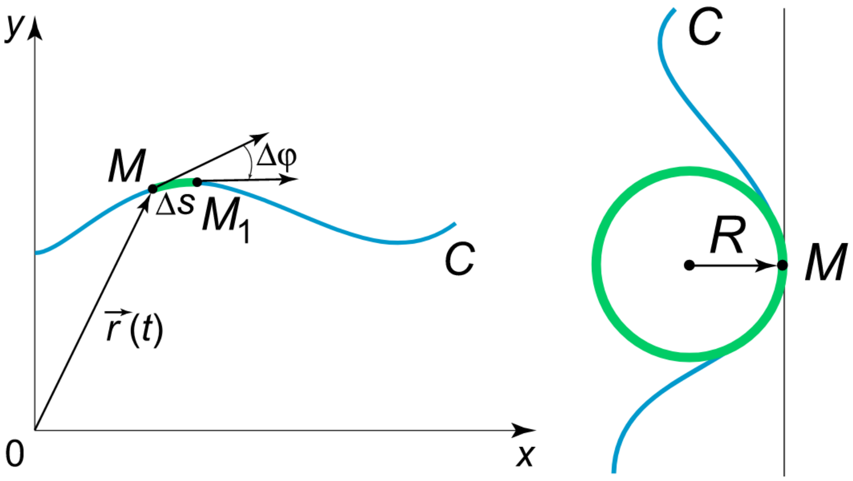

Figure 2.117: Curva 1 - Curva 2

seven points on a twisted cubic be joined, two and two, by twenty-one lines, then any seven planes that contain these 21 lines will osculate a second cubic curve.

2.8.1 Curvature and Intrinsic Geometry of a Surface in the Works of Euler

By the curvature of a curve at a point we mean the limit of the ratio of the angle Aa between the tangents at the endpoints of an arc to the length As of that arc as the latter contracts to the point; that is, the limit

k=da/ds = lim delta s para 0 delta a / delta s

k is also the reciprocal of the radius of curvature at the point in question, that is, the radius of the osculating circle at that point. (The osculating circle at a point P of a curve is defined as the limit of circles determined by three points on the curve as they tend to P.) The concepts of the curvature of a curve and of the osculating (literally "kissing," from the Latin osculans) circle were already known to Leibniz.1 Leibniz also suggested the possibility of characterizing the curvature of a surface by means of an osculating sphere.2 Euler characterized the curvature of a surface in his Investigations of the curvature of surfaces (Recherches sur la courbure des surfaces. Berlin, 1767) [176, vol. 28,pp. 1-22]. Euler called a normal section of a surface z =f(x, y) perpendicular to the xOy plane a principal section. By varying the angle q> between a normal section and the principal section Euler found that at each point of the surface there is a maximal radius of curvature/and a minimal radius of curvature g, that their planes are perpendicular to one another, and that the connection between an arbitrary radius of curvature r = r(q>) and the radii of curvature/and g is given by

1 Leibniz first considered osculation of curves|in particular, osculation of curves and circles|in his New reflections on the nature of the angles of tangency and osculation and their use in mathematical practice to replace complex figures with simpler ones (Meditatio nova de natura Anguli contactus et osculi, horumque usu in practica Mathesi ad figures faciliores succedaneas dicilioribus substituendas. Leipzig 1686) [312, v. 7, pp. 326-329J. 2 In a letter to Johann Bernoulli, dated July 29, 1698.

2.8.2 Cycloid

a curve (resembling a series of arches) traced by a point on a circle being rolled along a straight line. "Many famous mathematicians, including Descartes, have worked on a class of curves called cycloids ." Exemplos de cycloid Neile's parabola was the first algebraic curve to have its arc length calculated; only the arc lengths of transcendental curves such as the cycloid and the logarithmic spiral had been calculated before this. He defines evolutes and involutes of curves and, after giving some elementary properties, finds the evolutes of the cycloid and of the parabola. Viviani determined the tangent to the cycloid but he was not the first to succeed in this. Many famous mathematicians, including Descartes, have worked on a class of curves called cycloids . The ships are fitted with two independent Voith Schneider 18GH cycloidal propellers, which provide propulsion and steering in cruising and minehunting modes, and two twin tube bow thrusters. To estimate the relative surface area of basal lamina and apertures, we used a line intercept technique with cycloids . Our arch is a cycloidal arch, inspired by the one in Kahn's Kimbell Museum. Huygens believed that a pendulum swinging in a large are would be more useful at sea and he invented the cycloidal pendulum with this in mind. Desargues proposed cycloidal teeth for gear wheels in the 1630's. He also worked on cycloidal pendulums, the laws of resistance in a fluid and differential geometry.

2.8.3 Affine

allowing for or preserving parallel relationships. "Still to be implemented are geometric primitives and affine transformation of images-001." Exemplos de affine The simplest way to see why this assertion is true is to observe that any triangle in the plane can be mapped by an invertible affine transformation to an equilateral triangle. It can be obtained from the previous description by applying affine transformations to keep the polygon bounded as it degenerates. Using parallelograms to define affine transformations may be a mixed blessing. Still to be implemented are geometric primitives and affine transformation of images-001. Our interest is in the space of affine equivalence classes of equal-area polygons.

2.8.4 Homogeneous coordinates

From Wikipedia, the free encyclopedia

In mathematics, homogeneous coordinates or projective coordinates, introduced by August Ferdinand Mobius in his 1827 work Der barycentrische Calcul,[1][2] are a system of coordinates used in projective geometry, as Cartesian coordinates are used in Euclidean geometry. They have the advantage that the coordinates of points, including points at infinity, can be represented using finite coordinates. Formulas involving homogeneous coordinates are often simpler and more symmetric than their Cartesian counterparts. Homogeneous coordinates have a range of applications, including computer graphics and 3D computer vision, where they allow affine transformations and, in general, projective transformations to be easily represented by a matrix.

If the homogeneous coordinates of a point are multiplied by a non-zero scalar then the resulting coordinates represent the same point. Since homogeneous coordinates are also given to points at infinity, the number of coordinates required to allow this extension is one more than the dimension of the projective space being considered. For example, two homogeneous coordinates are required to specify a point on the projective line and three homogeneous coordinates are required to specify a point in the projective plane.

Rational Bezier curve { polynomial curve defined in homogeneous coordinates (blue) and its projection on plane { rational curve (red)

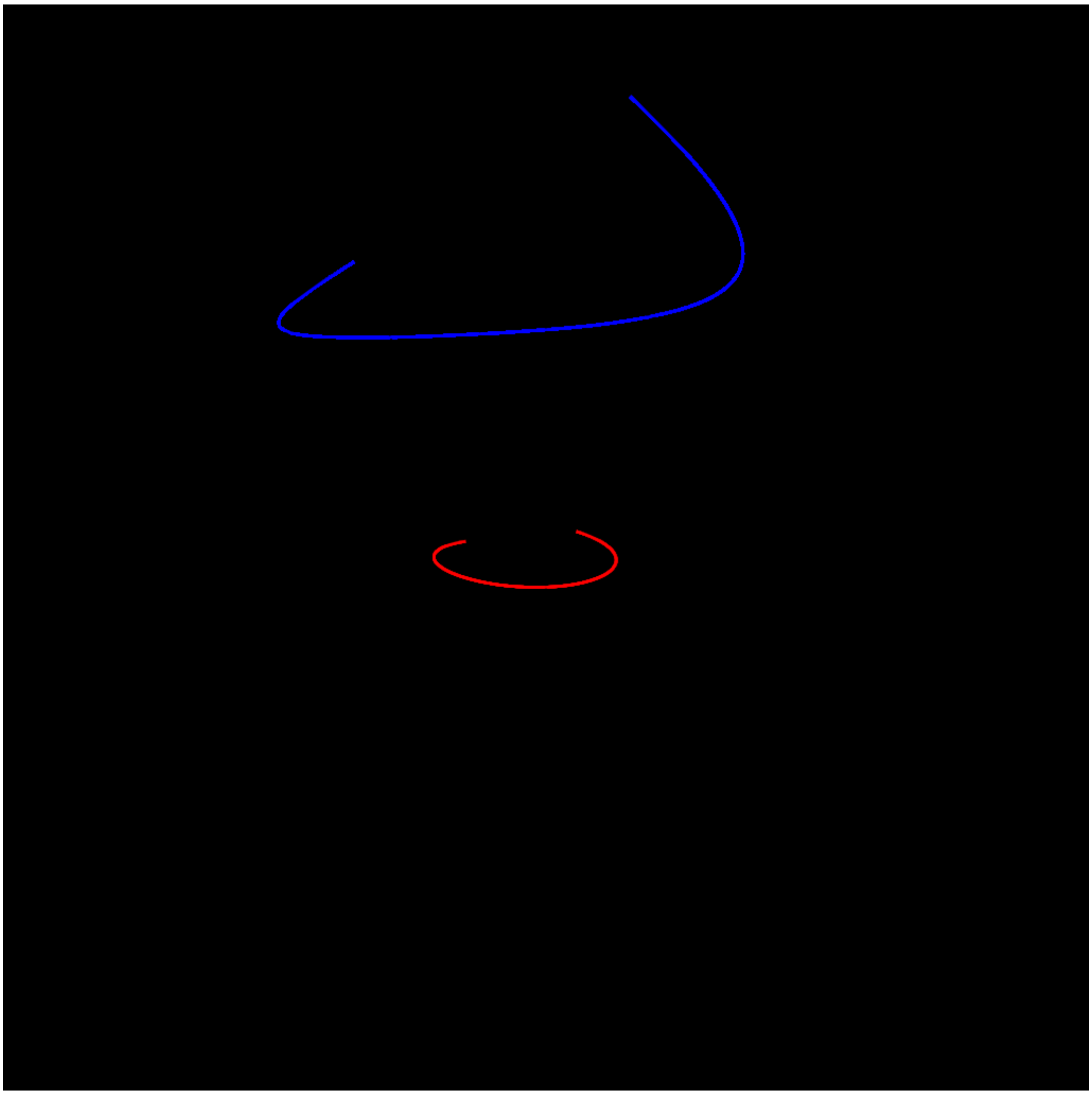

Figure 2.118: Rational Bezier 2D

2.8.5 Riemann Sphere

Stereographic projection of a complex number A onto a point a of the Riemann sphere.

2.8.6 Riemann Hypothesis

In mathematics, the Riemann hypothesis is a conjecture that the Riemann zeta function has its zeros only at the negative even integers and complex numbers with real part 1/2. Many consider it to be the most important

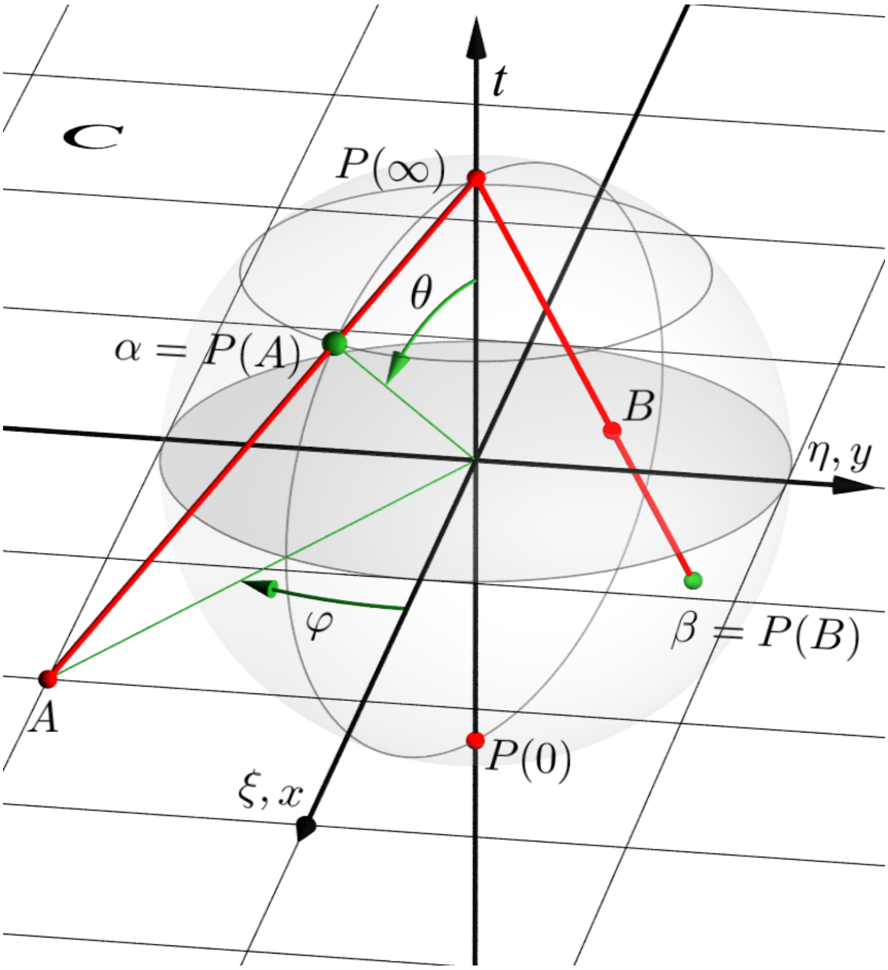

Figure 2.119: Riemann Sphere

unsolved problem in pure mathematics (Bombieri 2000). It is of great interest in number theory because it implies results about the distribution of prime numbers. It was proposed by Bernhard Riemann (1859), after whom it is named.

The Riemann hypothesis and some of its generalizations, along with Goldbach's conjecture and the twin prime conjecture, comprise Hilbert's eighth problem in David Hilbert's list of 23 unsolved problems; it is also one of the Clay Mathematics Institute's Millennium Prize Problems. The name is also used for some closely related analogues, such as the Riemann hypothesis for curves over finite fields.

Veja no texto:

Mobius

Hilber spaces L2

2.8.7 Atiyah Riemann Hypothesis proof: final thoughts

Source: https://aperiodical.com/2018/09/atiyah-riemann-hypothesis-proof-final-thoughts/

By Katie Steckles and Christian Lawson-Perfect. Posted September 28, 2018 in News

After Sir Michael Atiyah's presentation of a claimed proof of the Riemann Hypothesis earlier this week at the Heidelberg Laureate Forum, we've shared some of the immediate discussion in the aftermath, and now here's a round-up of what we've learned.

Firstly, it's become clear that the work presented by Atiyah doesn't constitute a proof of the Riemann Hypothesis, so the Clay Institute can rest easy with their 1 million dollars, and encryption on the internet remains safe. The argument Atiyah put forward rests on his function having certain properties, and many have concluded that no function with such properties can exist, including in a comment on our own post from an academic who's written about the Riemann Hypothesis extensively. Dick Lipton and Ken Regan have written a blog post looking at the detail of how is supposed to behave. According to some sources, Atiyah has stated he'll be publishing a more detailed paper shortly, but not many are holding their breath.

This is not easy mathematics, and even top-level mathematicians sometimes find their proofs don't hold together; it's no surprise that a solution to this problem wouldn't be found in a few lines of mathematics, and it's just a shame that this was so built up and sensationalised when it's becoming obvious that Atiyah didn't consult with anyone else about his proof before presenting it in a public forum.

Atiyah's is by no means the first claimed proof of the Riemann Hypothesis of recent years; many end up in the wastepaper bins of academic mathematicians around the world, who get sent them in handfuls. Only some claimed proofs get this level of attention though, and this one has been somewhat unique in that it was taken a lot more seriously than usual.

You might remember that a few years ago, the BBC reported that Opeyemi Enoch had proved the Riemann hypothesis, and we immediately said "no, he hasn't". Now Sir Michael Atiyah claims he's proved it, and before his lecture we said "well, let's see". Why the difference?

Enoch looked like any crank: went straight to the wrong venue – BBC News, not other mathematicians – didn't publish his proof in the open, and presented it at a very scammy looking conference run by Nina Ringo, who has previous form. And we'd never heard of him: the false-positive rate for

\anyone on the planet" with a claimed big proof is much, much higher than for, say, professors of maths in big universities. So for all those reasons, we were fairly sure he hadn't done it, like I can fairly confidently say you're not holding a winning lottery ticket right now.

Atiyah, on the other hand, has a Fields medal, an Abel prize, and a very long list of publications. We still thought he probably hadn't proved it, based on his other recent pronouncements, but he said he was going to present his proof, so we gave him the benefit of the doubt.

It's also valid to ask, why was Atiyah given such a prestigious platform to present his work? Well, the HLF isn't an ordinary academic conference { on their own website, they stipulate that \the Heidelberg Laureate Forum is not a classical scientific conference but a networking event meant to motivate and inspire the next generation of scientists." The Forum doesn't vet the talks submitted by the Laureates, as up until now they've never needed to { most people give talks about their historical achievements, or occasionally the work of their students, and some even give the same talk every year. It's still fair to say that a word of advice to Atiyah, to maybe stick to a general historical account of the problem (which his talk mostly was) might have been useful, if he could have been convinced.

Atiyah's proof is a very good example of what a proof isn't: it doesn't contain enough detail to convince someone of the truth of his argument, or even enough detail for someone to point to as wrong. There simply wasn't much to it, and the real glory of maths is in the gritty technical details, all of which were missing from this. Broad strokes are useful to give an overview, but they need to be backed up with working out.

The press has been patchy in its coverage of both Atiyah's presentation and the reaction to it. It's been a good example of how ill-equipped the general press is to cover maths news { breathless coverage from The Times, The Mirror and The Independent reveal a distinct lack of skepticism { but for anyone outside of mathematics, it's not easy to know whether skepticism is warranted. At least some outlets looked like they'd tried to ask a professional mathematician, but many claimed that nobody was prepared to comment { a point that Jacob Aron of New Scientist has complained about.

Quite a few places, including this Science Alert article (which was shared by the Australian Mathematical Sciences Institute on Twitter!) make slightly inaccurate claims, including \the hypothesis has been checked for the first 1,000,000,000,000 prime numbers", and that's not really how that works. Fortunately, we didn't see anyone saying \it's definitely proved", but did see a few stories framed as \rivals disagree", which isn't quite how this went down.

One positive side effect of this whole affair has been the amount of buzz it's created around the Riemann Hypothesis, big results like this, and maths in general. It's reminded us how exciting it would truly be if someone were to prove a result of this nature, and it was cool to see hundreds if not thousands of mathematicians (and non-mathematicians) talking about the proof on social media.

\Being wrong isn't a bad thing like they teach you in school. It is an opportunity to learn something." { Richard Feynman

Some commented ahead of the proof announcement that they hoped it was a real proof, as otherwise it would be terrible for someone like Atiyah, being publicly seen to be totally wrong about something. This sentiment, or equivalents post-hoc, have been echoed with varying levels of concern for his reputation. We think that's a really unhelpful thing to say: being wrong is an inevitable part of mathematics research, and it happens to everyone. Promulgating the idea that maths is all about being 100% right all of the time not only makes it harder for people at research level but is also one of the major reasons people are put off maths at an early age. Yes, it is possible on occasion to come up with an unassailably correct solution to a maths problem, unlike other disciplines which are more subjective in their assessment, but terror of failure causes some people to freeze up and not even try { when in fact being wrong is inevitable, and an important part of the learning process. That said, there are less public forums in which to be wrong, and as we've discussed, steering Atiyah towards one of those might have been kinder.

We wish Sir Michael a pleasant retirement in the knowledge that he has made a huge contribution to maths over the course of his life, and that shouldn't feel lessened or soured by these events. It's fantastic that someone can continue working on mathematics into their old age, especially if, through collaboration, they can inspire the next generation. We are grateful to him for his work.

> \Additional vertical spacing is inserted above and below the displayed text to separate it visually from the normal text. The text to be displayed may be of any length; it can be part of a sentence, a whole paragraph, or several paragraphs."

2.8.8 A subsection

2.8.9 Gottfried Wilhelm Leibniz

2.8.10 Transformation of Lorentz

Em [[física]], as "'transformações de Lorentz"', em homenagem ao físico neerlandês [[Hendrik Lorentz]], descrevem como, de acordo com a [[relatividade especial]], as medidas de espaço e tempo de dois observadores se alteram em cada sistema de referência. Elas refletem o fato de que observadores se movendo com [[velocidade]]s diferentes medem diferentes valores de distância, tempo e, em alguns casos, a ordenação de eventos.

Matematicamente, o "'fator de Lorentz "é determinado por:

$$= \frac{1}{\sqrt{(1 \quad \frac{v^2}{c^2})}} \qquad (2.2)$$

A transformação de Lorentz foi originalmente o resultado da tentativa de Lorentz e outros cientistas, como Woldemar Voigt,[1] para explicar as propriedades observadas da luz propagando-se no que se presumia ser o éter luminífero; Albert Einstein posteriormente reinterpreta a transformação como sendo uma consequência da natureza do espaço e tempo. A transformação de Lorentz substitui a transformação de Galileu da física newtoniana, que assumia um espaço e tempo absoluto. De acordo com a relatividade especial, a transformação de Galileu é apenas uma boa aproximação para velocidades relativas muito menores que a velocidade da luz.

Transformação de Lorentz para referenciais na configuração padrão

Assuma que há dois observadores O e Q, cada qual usando seu próprio sistema de coordenadas espacial para medir os intervalos de espaço e relógios estáticos devidamente posicionados e adequadamente sincronizados para permitir medidas de tempo (sistema de coordenadas relativístico 3). O utiliza t , x , y , z e Q utiliza t 1 , x 1 , y 1 , z 1 para designar a ocorrência de um mesmo evento. Suponha ainda que os sistemas de coordenadas são orientados de maneira que os eixos x e x' são colineares, os eixos y é paralelo ao eixo y' , assim como o eixo z ao z' . A velocidade relativa entre os dois observadores é v no sentido do eixo x. Assuma também que as origens de ambos sistemas de coordenadas coincidem em um dado instante referido como instante zero em ambos os referenciais (t=t'=0 quando x=0 coincide com x'=0). Se todas essas suposições são válidas, então os sistemas de coordenadas são ditos estarem na configuração padrão. Uma apresentação simétrica entre as transformadas direta em inversa de Lorentz podem ser obtidas se o sistema de coordenadas estão em configuração simétrica. A forma simétrica ressalta que todas as leis fsicas devem ser de tal tipo que permanecem inalteradas sob uma transformação de Lorentz.

A transformação de Lorentz para sistemas de referências na configuração padrão pode ser apresentada como

$$t1 = \quad t \quad vx=c^2 \;\; x1 = \quad (x \quad vt)\; y1 = yz1 = z \qquad (2.3)$$

onde

$$= \mathrm{q}\frac{1}{1 \quad \frac{v^2}{c^2}} \qquad (2.4)$$

e chamado [[fator de Lorentz]].

Forma matricial=== A transformação de Lorentz e dita um "boost" na direção "x" e e frequentemente expressa na forma matricial como

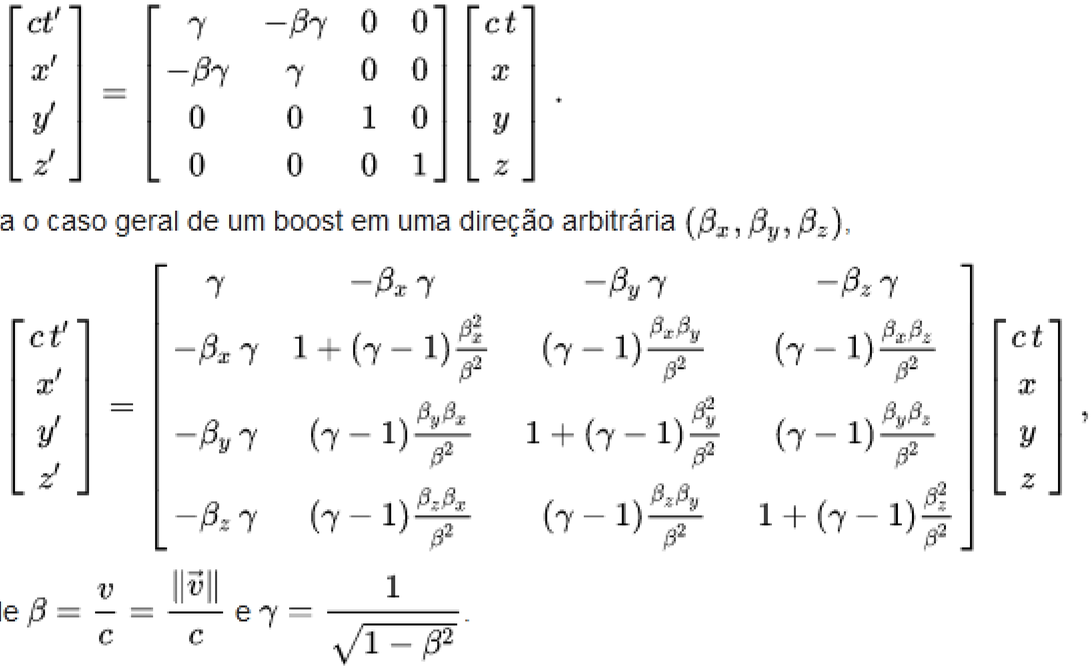

$$\begin{bmatrix} ct' \\ x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \gamma & -\beta\gamma & 0 & 0 \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} ct \\ x \\ y \\ z \end{bmatrix} .$$

Para o caso geral de um boost em uma direção arbitrária $(\beta_x, \beta_y, \beta_z)$,

$$\begin{bmatrix} ct' \\ x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \gamma & -\beta_x\gamma & -\beta_y\gamma & -\beta_z\gamma \\ -\beta_x\gamma & 1+(\gamma-1)\frac{\beta_x^2}{\beta^2} & (\gamma-1)\frac{\beta_x\beta_y}{\beta^2} & (\gamma-1)\frac{\beta_x\beta_z}{\beta^2} \\ -\beta_y\gamma & (\gamma-1)\frac{\beta_y\beta_x}{\beta^2} & 1+(\gamma-1)\frac{\beta_y^2}{\beta^2} & (\gamma-1)\frac{\beta_y\beta_z}{\beta^2} \\ -\beta_z\gamma & (\gamma-1)\frac{\beta_z\beta_x}{\beta^2} & (\gamma-1)\frac{\beta_z\beta_y}{\beta^2} & 1+(\gamma-1)\frac{\beta_z^2}{\beta^2} \end{bmatrix} \begin{bmatrix} ct \\ x \\ y \\ z \end{bmatrix} ,$$

onde $\beta = \frac{v}{c} = \frac{\|\vec{v}\|}{c}$ e $\gamma = \frac{1}{\sqrt{1-\beta^2}}$.

Motivação original

Desde as epocas de Galileu e Newton, era sabido que medidas laboratoriais de processos mecânicos não deveriam mostrar diferenças entre um equipamento em repouso e um outro que estivesse em movimento com velocidade constante e em linha reta: era o chamado princpio da relatividade. Mas, nem todas as leis da fsica eram consideradas universais e independentes do observador: de acordo com a teoria eletromagnética de Maxwell (refinada depois por Lorentz e outros) a luz não devia obedecer a esse princpio e deveria mostrar o efeito do movimento. Michelson e Morley fizeram uma experiência, em 1887, em que tentaram detectar a diferença entre a velocidade da luz se movendo na mesma direção do movimento da Terra (afetado pelo vento de éter resultante) e a velocidade da luz se movendo numa direção em ângulo reto com ela. Mas, o valor da velocidade da luz parecia não se alterar quando era alterada a velocidade do seu emissor — o que estava em desacordo com os modelos da Física Clássica.

Em 1889, Fitzgerald, um irlandês, sugeriu que talvez fosse uma contração do próprio equipamento experimental, que ocorria quando este atravessava o éter e que fazia com que a mudança na velocidade da luz não fosse detectável, ou seja, sugeriu que os corpos se contraíam quando se moviam com velocidades próximas à velocidade da luz. Independentemente, em 1895, Lorentz sugeriu uma hipótese do mesmo tipo, porém mais detalhada, em que, para assegurar a completa impossibilidade de detecção do éter, acrescentava a hipótese de haver uma mudança no tempo local marcado pelos relógios usados na experiência. As transformações de Lorentz, introduzidas por ele em 1904, descrevem esse efeito de diminuição do comprimento e dilatação do tempo para objetos que se movem a velocidades próximas à velocidade da luz.

O descrédito das teorias do éter acabou por levar à aceitação da proposta de Albert Einstein de que as transformações de Lorentz não fossem entendidas como transformações de objetos físicos mas, sim, como transformações do espaço e do tempo em si. Na sua Teoria da Relatividade Restrita, propôs que a razão pela qual não se conseguiam detectar diferentes velocidades da luz era, simplesmente, porque a velocidade da luz é uma constante universal. E mostrou que isso tornava o princípio da relatividade compatível com a teoria electromagnética. A necessidade de se modificar as equações da transformação de Galileu foi reconhecida ao se tentar usá-las nas equações de Maxwell. O raciocínio a seguir, atribuído a Einstein, ilustra intuitivamente a inconsistência.

Considere que seja possível a uma pessoa viajar à velocidade da luz. A luz, pelas equações de Maxwell, é uma oscilação dos campos elétricos E e magnéticos B, periódica no espaço e oscilante no tempo. No referencial dessa pessoa, a luz seria uma perturbação do campo eletromagnético periódica no espaço e constante no tempo. Tal solução, no entanto, não existe como solução das equações de Maxwell que governam a propagação da Luz.

Portanto, restam duas alternativas:

Modificar as equações Maxwell e manter a transformada de Galileu Ou modificar a transformada de Galileu

Não basta dizer que, já que as equações de Maxwell são confirmadas em laboratório, devemos modificar as transformadas de Galileu. Essas transformadas também são importantes pois são a base de toda a Mecânica Clássica, que portanto deveria ser revista.

Esse impasse foi resolvido em 1905 por Albert Einstein. A sua interpretação das Transformadas de Lorentz permitiu manter as equações de Maxwell inalteradas, mas exigiu uma revisão completa dos conceitos de tempo e espaço tão caros e fundamentais à Mecânica Clássica.

A transformação de Lorentz

Para se chegar as equações da transformação de Lorentz basta analisar como as equações de Maxwell se comportam com relação a uma transformação geral de coordenadas. Mas para simplificar a matemática, utiliza-se no lugar das equações de Maxwell uma de suas soluções, isto é, a equação da onda eletromagnética no vácuo:

$$\frac{\partial^2 \psi}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 \psi}{\partial t^2} = 0$$

propagando-se na direção x com velocidade c.

Quer-se uma transformação linear de coordenadas x, t para um novo referencial, x', t' que se move com velocidade v:

$$\begin{aligned} x' &= \alpha x + \beta t \\ t' &= \gamma x + \delta t \end{aligned}$$

O problema é encontrar a, b, g, delta de forma a que a equação de onda acima continue sendo uma equação de onda no novo referencial. Substituindo na equação de onda e resolvendo a equação para (a, b, g, delta) obtém-se:

$$\alpha^2 - \frac{\beta^2}{c^2} = 1$$
$$\frac{c^2\gamma}{\beta} = 1$$
$$\alpha = \delta$$

Substituindo na transformação linear original:

$$x' = \alpha\left(x + \frac{\beta}{\alpha}t\right)$$
$$t' = \alpha\left(\frac{\beta}{c^2\alpha}x + t\right)$$

Comparando com a transformada de Galileu:

$$x' = x - v \cdot t$$
$$t' = t$$

encontra-se:

$$\frac{\beta}{\alpha} = -v$$
$$\alpha = \frac{1}{\sqrt{(1 - \frac{v^2}{c^2})}}$$

substituindo na transformação linear inicial, encontra-se a transformada de Lorentz entre dois referenciais em movimento relativo com velocidade v:

$$x' = \frac{1}{\sqrt{(1-\frac{v^2}{c^2})}}(x - v\cdot t) = \gamma(x - v\cdot t)$$
$$t' = \frac{1}{\sqrt{(1-\frac{v^2}{c^2})}}(t - \frac{v}{c^2}x) = \gamma(t - \frac{v}{c^2}x)$$

Onde:

$$\gamma = \frac{1}{\sqrt{(1-\frac{v^2}{c^2})}}$$

e chamado de fator de Lorentz.

Forma Vetorial:

A transformação de Lorentz deduzida até então supõe um movimento relativo na direção do eixo-x, mas esta forma pode ser generalizada para um movimento em qualquer direção. Supondo que os referenciais se movam com uma velocidade em direção arbitrária v, então qualquer vetor r1 pode ser decomposto em suas componentes perpendicular e ortogonal ao vetor v.

O vetor r1 forma um ângulo 0 com o vetor v. Portanto temos que:

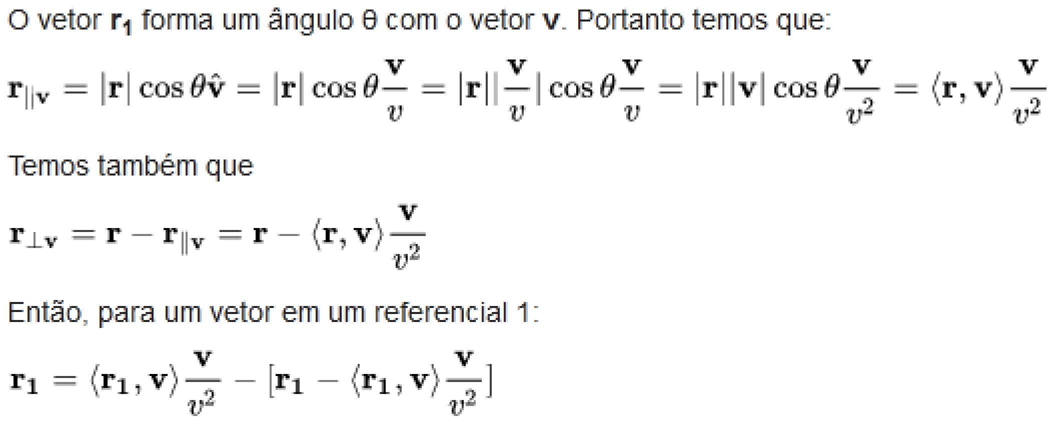

O vetor $\mathbf{r_1}$ forma um ângulo θ com o vetor $\mathbf{v}$. Portanto temos que:

$$\mathbf{r}_{||\mathbf{v}} = |\mathbf{r}|\cos\theta\hat{\mathbf{v}} = |\mathbf{r}|\cos\theta\frac{\mathbf{v}}{v} = |\mathbf{r}||\frac{\mathbf{v}}{v}|\cos\theta\frac{\mathbf{v}}{v} = |\mathbf{r}||\mathbf{v}|\cos\theta\frac{\mathbf{v}}{v^2} = \langle\mathbf{r},\mathbf{v}\rangle\frac{\mathbf{v}}{v^2}$$

Temos também que

$$\mathbf{r}_{\perp\mathbf{v}} = \mathbf{r} - \mathbf{r}_{||\mathbf{v}} = \mathbf{r} - \langle\mathbf{r},\mathbf{v}\rangle\frac{\mathbf{v}}{v^2}$$

Então, para um vetor em um referencial 1:

$$\mathbf{r_1} = \langle\mathbf{r_1},\mathbf{v}\rangle\frac{\mathbf{v}}{v^2} - [\mathbf{r_1} - \langle\mathbf{r_1},\mathbf{v}\rangle\frac{\mathbf{v}}{v^2}]$$

No entanto a componente perpendicular não sofre alterações frente à transformação de Lorentz e a componente perpendicular à velocidade o é. Então, de modo análogo ao feito para o movimento relativo na direção do eixo-x, temos:

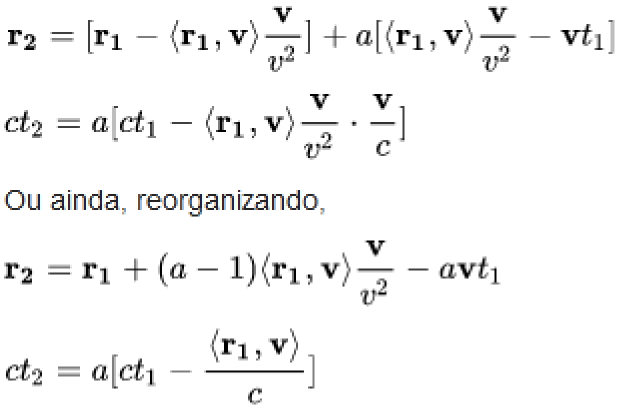

$$\mathbf{r_2} = [\mathbf{r_1} - \langle \mathbf{r_1}, \mathbf{v}\rangle \frac{\mathbf{v}}{v^2}] + a[\langle \mathbf{r_1}, \mathbf{v}\rangle \frac{\mathbf{v}}{v^2} - \mathbf{v}t_1]$$

$$ct_2 = a[ct_1 - \langle \mathbf{r_1}, \mathbf{v}\rangle \frac{\mathbf{v}}{v^2} \cdot \frac{\mathbf{v}}{c}]$$

Ou ainda, reorganizando,

$$\mathbf{r_2} = \mathbf{r_1} + (a-1)\langle \mathbf{r_1}, \mathbf{v}\rangle \frac{\mathbf{v}}{v^2} - a\mathbf{v}t_1$$

$$ct_2 = a[ct_1 - \frac{\langle \mathbf{r_1}, \mathbf{v}\rangle}{c}]$$

Bibliografia:

Ernst, A. och Hsu, J.-P.; First proposal of the universal speed of light by Voigt 1887 Chinese Journal of Physics (2001), pag 211-230 Vol 39-3; A tradução de Voigt (1887) em Inglês. University Physics – With Modern Physics (12th Edition), H.D. Young, R.A. Freedman (Original edition), Addison-Wesley (Pearson International), 1st Edition: 1949, 12th Edition: 2008, ISBN 0-321-50130-6, ISBN 978-0-321-50130-1 (em inglês) A inferência de tempo (t, ou t') por um observador (O, ou Q) não pode ser feita apenas a partir do relógio na respectiva origem. A variável tempo (t, ou t'), refere-se ao valor indicado por um relógio justaposto ao local do evento quando da sua ocorrência; sincronizado com outros análogos espacialmente distribuídos - todos sempre relativamente estáticos com aquele colocado na origem - segundo método fixo; e não ao valor indicado pelo relógio "no braço do observador" quando ele o vê. Se um dos relógios da malha situa-se a uma distância x da origem, deve ser ali colocado com um valor x/c acima (adiantado) em relação àquele indicado pelo relógio na origem. Um pulso de luz oriundo da origem (com velocidade c) dispara os relógicos quando os atinge. Tem-se assim um sistema de referência para medida de espaçotempo de um evento [(x,y,z,t), ou (x',y',z',t')] pelo respectivo observador [O, ou Q] na origem. Repare que, mesmo inicialmente adiantados, ao fim o observador (O, ou Q) verá, na origem, imagem de todos os relógios de sua própria malha espaçotempo,

cada qual tanto mais atrasados quanto mais distante estiver da origem (considere o tempo de ida da luz na sincronização, e o de retorno do sinal luminoso que carrega a "imagem" do relógio de volta à origem). Ao se olhar o céu à noite, em um dado momento, quanto mais longe está o que se olha, mais no passado ("t" menor) ocorreu o que agora se vê.

Source:

https://pt.wikipedia.org/wiki/Transformação de Lorentz
=====

Index

index entries

Bibliography

entries

Mathematics - Annotations 002

Prof. Stenio Musich

May 15, 2020

Chapter 1

Mathematics

1.1 Pappus's Theorem

1) https://en.wikipedia.org/wiki/Pappus's-centroid-theorem;

2) https://mathworld.wolfram.com/PappussCentroidTheorem.html;

3) https://www.math24.net/pappus-theorem/ ;

4) https://phys.libretexts.org/Bookshelves/Classical-Mechanics/Book-3A-Classical-Mechanics (Tatum)/01 3A-Centers-of-Mass/1.04 3A-Plane-Curves;

5) https://phys.libretexts.org/Bookshelves/Classical-Mechanics/Book A-Classical-Mechanics-(Tatum)/01 3A-Centers-of-Mass/1.06 A-The-Theorems-of-Pappus;

1.2 Theory of Relativity

1) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Immersed Boundary Method Development and Applications\
9811539391.pdf

2) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Theory of Functions of a Complex Variable\
Geometric theory of functions of a complex variable
(1969, AMS Bookstore).pdf

3) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Theory of Functions of a Complex Variable\
Introduction to the calculus of variations by

Bernard Dacorogna (z-lib.org).pdf

4) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Theory of Functions of a Complex Variable\
Introduction to the calculus of variations by
Bernard Dacorogna (z-lib.org).pdf

5) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Theory of Functions of a Complex Variable\
Matrix analysis and applied linear algebra.
With solutions to problems (2001, SIAM_
Society for Industrial and Applied Mathematics) - libgen.lc.pdf

6) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Elementary Approach to Special Relativity\9811531676.pdf

7) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Elementary Approach to Special Relativity\3030421007.pdf

8) C:\DATA-N~1\BOOKS-~1\ELEMEN~1\
[Graduate Texts in Physics] Gunter Ludyk (auth.) -
Einstein in Matrix Form_ Exact Derivation of the
Theory of Special and General Relativity without
Tensors (2013, Springer-Verlag Berlin Heidelberg).pdf

9) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Elementary Approach to Special Relativity\
Eulers Formula and Special Relativity.pdf

10) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Elementary Approach to Special Relativity\
3642372759_Relativit.pdf

11) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Elementary Approach to Special Relativity\3030273466.pdf

12) C:\Users\sm\Downloads\Encyclopedia of Mathematics.pdf

13) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Elementary Approach to Special Relativity\
Books_2_MATH\Complex and Differential Geometry.pdf

14) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Elementary Approach to Special Relativity\
Books_2_MATH\Concise Encyclopedia Mathematics_2.pdf

Pag. 6 (494) - Droz-Farny Circles

15) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Elementary Approach to Special Relativity\Books_2_MATH\
Math Advanced Calculus.pdf

Pag. 141 (151) - Chapter 4 - The Riemann Integral

16)C:\DATA-NovoVivo-85-2020\Books-85-2020\
Elementary Approach to Special Relativity\
Books_2_MATH\Exploring Analytic Geometry with
Mathematica [Vossler 1999-10-22].pdf

Pag. 85 (103) - Chapter 6 - Circles

17) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Elementary Approach to Special Relativity\978-3-319-22014-7.pdf

18) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Elementary Approach to Special Relativity\
special-relativity-from-einstein-to-strings.pdf

19) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Elementary Approach to Special Relativity\
relativity-special-general-and-cosmological.pdf

20) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Elementary Approach to Special Relativity\
very-special-relativity-an-illustrated-guide.pdf

21) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Elementary Approach to Special Relativity\
energy-and-geometry-an-introduction-to-deformed-sp.pdf

22) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Elementary Approach to Special Relativity\

the-geometry-of-minkowski-spacetime-an-introductio.pdf

23) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Elementary Approach to Special Relativity\
differential-geometry-of-lightlike-submanifolds.pdf

24) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Elementary Approach to Special Relativity\
the-geometry-of-schemes.pdf

25) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Elementary Approach to Special Relativity\
algebraic-geometry-1-schemes.pdf

26) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Elementary Approach to Special Relativity\
0387978488_3540978488.pdf

The Geometry of Minkowski Spacetime

27) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Elementary Approach to Special Relativity\ms-qft-DRAFT.pdf

Quantum Field Theory - Mark Srednicki

28) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Elementary Approach to Special Relativity\
From-Special-Relativity-to-Feynman-Diagrams-A-
Course-of-Theoretical-Particle-Physics-for-Beginners.pdf

From Special Relativity to Feynman Diagrams

29) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Elementary Approach to Special Relativity\
From-Special-Relativity-to-Feynman-Diagrams-A-
Course-in-Theoretical-Particle-Physics-for-Beginners_02.pdf

30) https://archive.org/details/jstor-1988628

Projective Differential Geometry of Developable Surfaces
by Denton, William Wells

Other: a) https://archive.org/details/projectivediffer00dent

b) C:\DATA-NovoVivo-85-2020\Books-85-2020\
Elementary Treatise on the Geometry of Four Dimensions\
Important\projectivediffer00dent.pdf

c) https://archive.org/details/jstor-1988606/page/n1/mode/2up

d) https://archive.org/details/jstor-1986347/page/n1/mode/2up

e) https://archive.org/details/jstor-1988739/mode/2up

On the Theory of Curved Surfaces, and Canonical Systems in Projective Differential Geometry.

f) https://archive.org/details/in.ernet.dli.2015.215537/
page/n8/mode/2up

31) https://archive.org/details/in.ernet.dli.2015.152722/mode/2up

https://archive.org/details/projectivegeomet004116mbp/mode/2up

https://archive.org/details/in.ernet.dli.2015.171113/
page/n1/mode/2up

https://archive.org/details/in.ernet.dli.2015.205505/
page/n7/mode/2up

https://archive.org/details/67249318/mode/2up

1.3 Calculus - Projective Geometry

1) https://math.dartmouth.edu/~doyle/

Book:

a) https://math.dartmouth.edu/~doyle/docs/calc/calc.pdf

Crowell and Slesnick's Calculus with Analytic Geometry

b) https://math.dartmouth.edu/~doyle/docs/three/three.pdf

Division by three

c) https://math.dartmouth.edu/~doyle/docs/gi/gi.pdf

Geometry and the Imagination

2) C:\ZIP\ZIP_02\Books\math-deep.pdf

Other source: IMPORTANT

C:\DATA-NovoVivo-87-2020\Books-87-2020\Books\
math-deep.pdf

Pag. 823 - Chapter 26 - Basics of Projective Geometry

Pag. 827 - Figure 26.2: A central projection in A
3 through a0 of the parabola G1 (t) onto the hyperplane
x 3 = 1.

3) C:\ZIP\ZIP_02\Books\multivariable.pdf

Other:
C:\DATA-NovoVivo-87-2020\Books-87-2020\
Books\multivariable.pdf

Contents:

1
Analytic Geometry 15
1.1 Lines 16
1.2 Distance Between Two Points; Circles 21
1.3 Functions 22
1.4 Shifts and Dilations 27
2
Instantaneous Rate of Change: The Derivative 31

2.1 The slope of a function 31
2.2 An example 36
2.3 Limits 38
2.4 The Derivative Function 48
2.5 Adjectives For Functions 53

3
Rules for Finding Derivatives 57
3.1 The Power Rule 57
3.2 Linearity of the Derivative 60
3.3 The Product Rule 62
3.4 The Quotient Rule 64
3.5 The Chain Rule 67

4
Transcendental Functions 73
4.1 Trigonometric Functions 73
4.2 The Derivative of sin x 76
4.3 A hard limit 77
4.4 The Derivative of sin x, continued 80
4.5 Derivatives of the Trigonometric Functions 81
4.6 Exponential and Logarithmic functions 82
4.7 Derivatives of the exponential and logarithmic functions .. 84
4.8 Implicit Differentiation 89
4.9 Inverse Trigonometric Functions 94
4.10 Limits revisited 97
4.11 Hyperbolic Functions 102

5
Curve Sketching 107
5.1 Maxima and Minima 107
5.2 The first derivative test 111
5.3 The second derivative test 112
5.4 Concavity and inflection points 113
5.5 Asymptotes and Other Things to Look For 115

4) https://www.degruyter.com/view/title/530005
Measure Theory in Non-Smooth Spaces
DOI: https://doi.org/10.1515/9783110550832

Source:
https://www.degruyter.com/downloadpdf/title/530005.pdf

5) https://people.math.osu.edu/gerlach.1/math/BVtypset

LINEAR MATHEMATICS IN INFINITE DIMENSIONS Signals Boundary Value Problems and Special Functions U. H. Gerlach January 2017

Contents

0 Introduction 1
1 Sturm-Liouville Theory 5
1.1 Three Archetypical Linear Problems 5
1.2 The Homogeneous Problem 7
1.3 Sturm-Liouville Systems 11
1.3.1 Sturm-Liouville Differential Equation 11
1.3.2 Homogeneous Boundary Conditions 14
1.3.3 Basic Properties of a Sturm-Liouville Eigenvalue Problem 19
1.4 Phase Analysis of a Linear Second Order O.D.E. 36
1.4.1 The Pruffer System 37
1.5 Qualitative Results 42
1.6 Phase Analysis of a Sturm-Liouville System 45
1.6.1 The Boundary Conditions 46
1.6.2 The Boundary Value Problem 47
1.6.3 The Behavior of the Phase: The Oscillation Theorem . 48
1.6.4 Discrete Unbounded Sequence of Eigenvalues 52
1.7 Completeness of the Set of Eigenfunctions via Rayleigh's Quotient . 54

2 Infinite Dimensional Vector Spaces 61
2.1 Inner Product Spaces 62
2.2 Normed Linear Spaces 63
2.3 Metric Spaces 67
2.4 Complete Metric Spaces 69
2.4.1 Limit of a Sequence 69
2.4.2 Cauchy Sequence 69
ixx CONTENTS
2.4.3 Cauchy Completeness: Complete Metric Space,
Banach Space, and Hilbert Space 70
2.5 Hilbert Spaces 72
2.5.1 Two Prototypical Examples 74
2.5.2 Hilbert Spaces: Their Coordinatizations 75
2.5.3 Isomorphic Hilbert Spaces 96

6) https://www.e-booksdirectory.com/details.php?ebook=1020

Introduction to Methods of Applied Mathematics
by Sean Mauch

https://physics.bgu.ac.il/~gedalin/Teaching/Mater/am.pdf

http://www.its.caltech.edu/~sean

https://physics.bgu.ac.il/~gedalin/Teaching/Mater/

https://physics.bgu.ac.il/~gedalin/Teaching/Mater/cb.pdf

https://physics.bgu.ac.il/~gedalin/Teaching/Mater/
EMFT_Book.pdf

https://upload.wikimedia.org/wikipedia/commons/a/a3/
Calculus.pdf

Pag.10 (22) -1.7 MULTIVARIABLE AND DIFFERENTIAL CALCULUS

7) C:\DATA-NovoVivo-87-2020\Books-87-2020\Books\math-deep.pdf

Algebra, Topology, Differential Calculus, and
Optimization Theory For Computer Science and
Machine Learning

8) C:\DATA-NovoVivo-87-2020\Books-87-2020\Calculus\Calculus.pdf

9) C:\DATA-N~1\BOOKS-~1\ABEGIN~1\
Michael S. Schneider - A Beginner's Guide to
Constructing the Universe_ Mathematical
Archetypes of Nature, Art, and Science
(1995, HarperPerennial) - libgen.lc.pdf

10) C:\DATA-NovoVivo-87-2020\Books-87-2020\
A Beginner's Guide to Constructing the
Universe Mathematical Archetypes of Nature,
Art, and Science\Several Complex Variables
(2000, Cambridge University Press) - libgen.lc.pdf

11) C:\DATA-NovoVivo-87-2020\Books-87-2020\
Beautiful Geometry\e_ The Story of a Number
(1994, Princeton University Press) - libgen.lc.pdf

12) C:\DATA-NovoVivo-87-2020\Books-87-2020\
Calculus\calculus-and-linear-algebra-vol-1-vectors-
in-the-plane.pdf

13) C:\DATA-NovoVivo-87-2020\Books-87-2020\
Calculus\calculus-and-linear-algebra-vol-2-vector-
spaces-many.pdf

1.4 Subspaces

1) C:/DATA-NovoVivo-87-2020/Books-87-2020 Active Subspaces Emerging Ideas for Dimension Reduction in Parameter Studies/1611973856.pdf
3.3.1 Approximate gradients
Large-scale simulations often lack simple-to-evaluate expressions for the gradient of the quantity of interest with respect to the inputs. Finite difference approximations are useful when m is not too large and f is neither too expensive nor too noisy. More and Wild's recent work characterizes the gradient when function evaluations contain noise [91, 92]. Next we extend the bounds on errors in the estimated igenpairs to the case when the gradient is computed with some error. The error model we analyze is generic. Let g(x) denote the approximate gradient computed at x pert X.
2) C:/DATA-NovoVivo-87-2020/Books-87-2020 Active Subspaces Emerging Ideas for Dimension Reduction in Parameter Studies/Exact solutions and invariant subspaces of nonlinear partial differential equations in mechanics and physics.pdf

Pag. 05 (36) 1.1.2 Nonlinear wave equation Example 1.6 (Quadratic wave equation) Ovsiannikov [456, p. 286] performed a classification of group-invariant solutions of the following system
3) C:/DATA-NovoVivo-87-2020/Books-87-2020 Active Subspaces Emerging Ideas for Dimension Reduction in Parameter Studies/Invariant Subspaces of Matrices with Applications (2006, SIAM, Society for Industrial and Applied Mathematics) - libgen.lc.pdf

Pag. 387 (410) - Chapter Thirteen The Metric Space of Subspaces 13.1 THE GAP BETWEEN SUBSPACES

20200519

4) C:/DATA-NovoVivo-88-2020/Books-88-2020/ Introduction to the Mathematics of Subdivision Surfaces/ 0898716977.pdf

Mathematics of subdivision surfaces

Pag.247 (273) - Chapter 6 Evaluation and Estimation of Surfaces

6.1 Evaluation and tangent stencils for nonregular points

5) C:/DATA-NovoVivo-88-2020/Books-88-2020/ A Modern Introduction to Differential Equations, 3rd Edition/A Modern Introduction to Differential Equations 3rd Edition.pdf

Pag. 113 (126) - Euler Method - Error

6) C:/DATA-NovoVivo-88-2020/Books-88-2020/ An Introduction to Modern Astrophysics, Second Edition/ 1108422160.pdf

Pag. 1026 (1040) - Chapter 26 - Galatic Evolution

Figure 26.20 - A Gaussian cylinder located entirely within the field of stars of the Galactic disk. IMPORTANT Next, the gravitational potential energy of a particle of mass, m, at a height, h, above the midplane is given by...

Obs.: Livro todo!!!

7) C:/DATA-NovoVivo-88-2020/Books-88-2020/ Applications of Heat, Mass and Fluid Boundary Layers/ Applications of Heat Mass and Fluid Boundary Layers.pdf

Pag. 04 (22) - Figure 1.1 Differential control volume in the rectangular coordinate system

1.2.2 Derivation of the law of conservation of mass in the rectangular coordinate system

IMPORTANT

8) C:/DATA-NovoVivo-88-2020/Books-88-2020/ Symmetry, Shape, and Space An Introduction to Mathematics Through Geometry/ Symmetry, Shape and Space. An Introduction to Mathematics through Geometry (2006, Key College) - libgen.lc.pdf

Obs.: Livro todo !!! IMPORTANT

1.5 Mass and Motion in General Relativity

1) C:/DATA-NovoVivo-88-2020/Books-88-2020/ Physics of Mass/Mass and Motion in General Relativity (2011, Springer Netherlands) - libgen.lc.pdf
Livro todo!!! IMPORTANT
Mass and Motion in General Relativity
Pag. 70 (90) - 2.2 Gaussian Units
Obs.: J.D. Jackson, Classical Electrodynamics, 3rd edn. (Wiley, New York, 1998)
Source:
C:/DATA-NovoVivo-88-2020/Books-88-2020/Physics of Mass/ John David Jackson - Classical Electrodynamics (1999, John Wiley) - libgen.lc.pdf
Pag. 782 (806) - Appendix
4 Conversion of Equations and Amounts Between SI Units and Gaussian Units
Source:

C:/DATA-NovoVivo-88-2020/Books-88-2020/Physics of Mass/ John David Jackson - Classical electrodynamics (1999, Wiley) - libgen.lc.djvu

Pag. 70 (94) - 2.9 Separation of Variables; Laplace Equation in Rectangular Coordinates
Pag. 72 (96) - 2.10 A Two-Dimensional Potential Problem; Summation of a Fourier Series

1.6 Nature - Calculus

Veja diretório todo!
1) C:/BOOKS-2020-15/Books-91-2020/Nature - 14 May 2020/ Calculus in 3D Geometry, Vectors, and Multivariate Calculus (MAA Textbooks).pdf
2) C:/BOOKS-2020-15/Books-91-2020/Nature - 14 May 2020/ Euclid|The Creation of Mathematics (1999, Springer-Verlag New York).pdf
3) C:/BOOKS-2020-15/Books-91-2020/Nature - 14 May 2020/ Geometry The Line and the Circle.pdf
4) C:/BOOKS-2020-15/Books-91-2020/Nature - 14 May 2020/ Introductory Technical Mathematics, 6th edition.pdf
5) C:/BOOKS-2020-15/Books-91-2020/Nature - 14 May 2020/ Math Makers The Lives and Works of 50 Famous Mathematicians.pdf

6) C:/BOOKS-2020-15/Books-91-2020/Nature - 14 May 2020/Mathematical Techniques in GIS- Second Edition.pdf
7) C:/BOOKS-2020-15/Books-91-2020/Nature - 14 May 2020/ Pleasures of Pi,e and Other Interesting Numbers.pdf
8) C:/BOOKS-2020-15/Books-91-2020/Nature - 14 May 2020/ Tensors and Riemannian Geometry With Applications to Differential Equations.pdf

1.7 Calculus - Important

Obs.: Books-91-2020/Important

1) C:/BOOKS-2020-15/Books-91-2020/Important/ Topics in Geometry (1994, Elsevier Inc, Academic Press).pdf
2) C:/BOOKS-2020-15/Books-91-2020/Important/ Plane Algebraic Curves Translated by John Stillwell (2012, Birkhauser Basel).pdf
Books-91-2020/Tensor Analysis.../
Obs.: Todo diretorio !!!
3) C:/BOOKS-2020-15/Books-91-2020/Tensor Analysis (De Gruyter Textbook)/978-981-15-2770-8.pdf Obs.: Matrix Algebra
4) C:/BOOKS-2020-15/Books-91-2020/Tensor Analysis (De Gruyter Textbook)/4073.pdf Obs.: Manifolds, Tensor Analysis, and Applications Third Edition
5) C:/BOOKS-2020-15/Books-91-2020/Tensor Analysis (De Gruyter Textbook)/Geometric Multiplication of Vectors An Introduction to Geometric Algebra in Physics (2019, Springer International Publishing Birkhauser).pdf
6) C:/BOOKS-2020-15/Books-91-2020/Tensor Analysis (De Gruyter Textbook)/Introduction to Vector and Tensor Analysis-Dover Publications (1972).pdf
7) C:/BOOKS-2020-15/Books-91-2020/Tensor Analysis (De Gruyter Textbook)/Matrix Methods.pdf
8) C:/BOOKS-2020-15/Books-91-2020/Tensor Analysis (De Gruyter Textbook)/Multivariable and Vector Calculus - An Introduction.pdf
9) C:/BOOKS-2020-15/Books-91-2020/Tensor Analysis (De Gruyter Textbook)/Partial Differential Equations Classical Theory with a Modern Touch (2020, Cambridge University Press) - libgen.lc.pdf
10) C:/BOOKS-2020-15/Books-91-2020/Tensor Analysis (De Gruyter Textbook)/TensorAnalysisDeGruyter.pdf Obs.: Livro important IMPORTANT

Pag. 161 (178) - 4.1.2 Coordinate Surfaces and Coordinate Curves The surfaces with ui = const are called coordinate surfaces. Except at singular

points, they cover the space simply, i. e. every point lies on exactly one coordinate surface of each of the three families of coordinate surfaces. Each pair of coordinate surfaces intersects along a coordinate curve, so that, again except for singular points, three coordinate curves intersect at each point.
Look: Figure!!!
11) C:/BOOKS-2020-15/Books-91-2020/Tensor Analysis (De Gruyter Textbook)/The VNA Applications.pdf
12) Diretorio Todo: Books-91-2020/Phi, Pi, e and i/

1.8 Projection

Date: 20200629

1) C:/DATA-NovoVivo-101-2020/Books-101-2020/Hilbert/Gallardo/EspHilbert2-2011.pdf

Chapitre 2: Le theoreme de projection et ses applications

2 Projection sur un convexe ferme

2) C:/DATA-NovoVivo-101-2020/Books-101-2020/Hilbert/ Important/cp035208.pdf

2.2 Demonstracao do Teorema de Hilbert

Pag. 14(19) - Figura - Linha Assintotica IMPORTANT

3) C:/DATA-NovoVivo-101-2020/Books-101-2020/Unitary Symmetry And Combinatorics by Louck James D/9812814728.pdf

Book: Unitary Symmetry and Combinatorics

Pag. 229(252) -

Chapter 4 Generating Functions in Angular Momentum Theory

4) C:/DATA-NovoVivo-102-2020/Books-102-2020/Differentiable Manifolds/Differentiable Manifolds A Theoretical Physics Approach Ed 2.pdf

Book: Differentiable Manifolds A Theoretical Physics Approach

Pag. 47(56) -

2.2 Lie Derivative of Functions and Vector Fields

5) C:/DATA-NovoVivo-102-2020/Books-102-2020/Foundations of Solid State Physics Dimensionality and Symmetry/Foundations of solid state physics - dimensionality and symmetry (2019, Wiley-VCH Verlag GmbH e Co).pdf

Book: Foundations of solid state physics.

Pag. 226(244) - 7.1 Heterogeneity: Band-Level Diagrams and the Contact

Figure 7.1: Free electron bands of a metal

Figure 7.1 Band diagrams in k compared with band-level diagrams in x. Since the band-level diagram is plotted in x, we can see when electronic energy state changes as we move from place to place in a heterogeneous

system. But we have given up k information.
6) C:/DATA-NovoVivo-102-2020/Books-102-2020/ Rotational Structure in Molecular Infrared Spectra, 2nd Edition/ Rotational-Structure-in-Molecular-Infrared-Spectra.pdf

Book: Rotationsl Structure in Molecular Spectra
Pag. 28(35) -
2.4.2 Einstein Coefficients and Line Strengths

1.9 General Relativity

Date: 20200630

1) C:/DATA-NovoVivo-102-2020/Books-102-2020/ General Relativity Curvature and Geodesics/General-Relativity.pdf
Book: GENERAL RELATIVITY: CURVATURE AND GEODESICS
Pag.10(16) - 2.1 The Einstein Summation Convention
Pag. 60(66) - 3.7 The Riemann Curvature Tensor - Motivation
IMPORTANT

2) C:/DATA-NovoVivo-102-2020/Books-102-2020/GAUSS/ A transcendência do pi e outros.pdf
Book: A transcendência do pi e outros. pdf
Pag. 20 - Figura 1.7: Produto na Reta Real
Pag. 21 - 1.2.1 Aproximação de Números Irracionais por Números Racionais

3) C:/DATA-NovoVivo-102-2020/Books-102-2020/ A Breakthrough in Higher Dimensional Spheres Infinite Series PBS Digital Studios/A Breakthrough in Higher Dimensional Spheres Infinite Series PBS Digital Studios.mp4
Source: https://www.youtube.com/watch?v=ciM6wigZK0w

4) C:/DATA-NovoVivo-102-2020/Books-102-2020/ A Breakthrough in Higher Dimensional Spheres Infinite Series PBS Digital Studios/ What was Fermat's \Marvelous Proof Infinite Series.mp4
Source: https://www.youtube.com/watch?v=SsVl7-R2MvI (o hífen deve ser underline)

5) C:/DATA-NovoVivo-102-2020/Books-102-2020/ A Breakthrough in Higher Dimensional Spheres Infinite Series PBS Digital Studios/Henri Darmon An-

drew Wiles- marvelous proof.mp4
Source: https://www.youtube.com/watch?v=oqMaziDIBYY

1.10 Vector Spaces

Date: 20200710

1) C:/ARQ-BASICS/10/Math-Books/zakon-analysisI-a4-one.pdf
Pag. 63(75) - Chapter 3 - Vector Spaces. Metric Spaces

2) C:/ARQ-BASICS/10/Math-Books/zakon-analysisII-a4-one.pdf
Pag. 97 (109) - Chapter 7 - Volume and Measure

3) C:/ARQ-BASICS/10/Spherical Geometry and Its Applications/ Spherical Functions of Mathematical Geosciences A Scalar, Vectorial, and Tensorial Setup.pdf
Pag. 201 (214) - 5 Vector Spherical Harmonics

4) C:/ARQ-BASICS/08/Spherical-Trigonometry/Textbook on Spherical Astronomy (1977, Cambridge University Press).pdf

5) C:/BOOKS-2020-22/Books-110-2020/Important/epdf.pub-geometria.pdf
Pag. 341 (353) - Capítulo XI - La geometría circular
11.1 La proyección estereográfica

6) C:/DATA-NovoVivo-111-2020/Books-111-2020/Linear Algebra Illustrated/ Linear-Algebra-Illustrated.pdf
Pag. 177 - 3.3. Transforming the axes transforms the plane
Obs.: Livro todo.

7) C:/DATA-NovoVivo-111-2020/Books-111-2020/ The Evolution Problem in General Relativity/ The Evolution Problem in General Relativity.pdf
Pag. 203 (215) - 5 Estimates for the Riemann Curvature Tensor

8) C:/DATA-NovoVivo-111-2020/Books-111-2020/ The Picture Book of Quantum Mechanics Ed 3/ The Picture Book of Quantum Mechanics Ed 3.pdf

9) C:/DATA-NovoVivo-111-2020/Books-111-2020/ Introduction to Spintronics/Introduction to Spintronics.pdf

Pág. 65 (84) - 4 Evolution of a Spinor on the Bloch Sphere

10) C:/DATA-NovoVivo-111-2020/Books-111-2020/ Higher Order Partial Differential Equations in Clifford Analysis/ Higher Order Partial Differential Equations in Clifford Analysis Effective Solutions to Problems.pdf

11) C:/DATA-NovoVivo-111-2020/Books-111-2020/ Quantum Field Theory/Quantum Field Theory.pdf

Pag. 43 (50) - Chapter 6 - Free Scalar Fields

12) C:/DATA-NovoVivo-111-2020/Books-111-2020/ Geometry, Language, And Strategy/ Geometry, Language, And Strategy.pdf

Pág. 102 (114) - 6.1.2 Gravity

13) C:/DATA-NovoVivo-111-2020/Books-111-2020/ The Math Book Big Ideas Simply Explained/ The Math Book Big Ideas Simply Explained.pdf

Pág. 336 - Algebraic roots

14) C:/DATA-NovoVivo-111-2020/Books-111-2020/ The elements of non-Euclidean geometry/ The Elements of Non-Euclidean Geometry 1909.pdf

CONGRUENT TRANSFORMATIONS Axiom of continuity 23 Division of distances 23 Measure of distance 26 Axiom of congruent transformations 29 Definition of angles, their properties 30

15) C:/DATA-NovoVivo-111-2020/Books-111-2020/ Problems and Solutions in Plane Trigonometry, LaTeX Edition For the use of Colleges and Schools/ Problems and Solutions in Plane Trigonometry, LaTeX Edition For the use of Colleges and Schools.pdf Obs.: Ano 2018 - 376 páginas. Pág. 63 (69) - CHAPTER X - Logarithms and Logarithmic Series. Pág. 75 (81) - CHAPTER XI - Use of Logarithmic And Trigonometrical Tables.

Livro todo.

16) C:/DATA-NovoVivo-111-2020/Books-111-2020/ Problems and Solutions in Plane Trigonometry, LaTeX Edition For the use of Colleges and Schools/ Isaac Todhunter - Problems and Solutions in Plane Trigonometry (0) - libgen.lc.pdf Obs.: Ano 2016 - 460 páginas.

IMPORTANT Os itens 15 e 16 são do mesmo autor e título, mas com páginas diferentes.

1.11 Curved Spacetime

Date: 20200715

1) C:/DATA-NovoVivo-114-2020/Books-114-2020/Spacetime/ 9812813756.pdf RELATIVITY:MODERN LARGE-SCALE SPACETIME STRUCTURE OF THE COSMOS

2) C:/DATA-NovoVivo-114-2020/Books-114-2020/Spacetime/ 9814719056.pdf Unified Field Mechanics

3) C:/DATA-NovoVivo-114-2020/Books-114-2020/Spacetime/ Beyond-Spacetime-The-Foundations-of-Quantum-Gravity.pdf

4) C:/DATA-NovoVivo-114-2020/Books-114-2020/Spacetime/ Deformed Spacetime Geometrizing Interactions in Four and Five Dimensions - 2007.pdf Pag 275 (295) - Chapter 18 Multidimensional Space{Time Einstein's relativity theories contain, among the others, two basic precepts. First, they stated beyond any reasonable doubt that our Universe is (at least) (3+1)-dimensional. Moreover, Einstein taught us that introducing extra dimensions (time, in this case) can provide a better description of (and even simplify) the laws of nature (electromagnetism for Special Rela- tivity and gravity for General Relativity). This latter teaching by Einstein was followed by Kaluza [111] and Klein [112], who added a fifth dimension to the four space{time ones in order to unify electromagnetism and gravitation (let us recall however that Nordstrom [113] was the first to realize that in a 5D space-time the field equations do split naturally into Einstein's and Maxwell's equations). Although unsuccessful, the Kaluza{Klein (KK) theory constituted the first attempt to unification of fundamental interactions within a multidimensional space{time. The KK scheme, in which the coefficient of the fifth coordinate is constant, was later generalized by Jordan [114] and Thiry [115], who considered it to be a general function of the space{ time coordinates. The 5D KK formalism has been later extended to higher dimensions, also in the hope of achieving unification of all interactions, including weak and strong forces [116{118]. Six dimensions were e.g., con- sidered by Ingraham [117] and Podolanski [118]. However, a true revival of multidimensional theories starting from 1970 was due to the advent of string theory [119]

and supersymmetry [120]. As is well known, their combination, superstring theory [121], provides a framework for gravity quantization. Modern generalizations [122,123] of the Kaluza{Klein scheme require a minimum number of 11 dimensions in order to accommodate the Standard Model of electroweak and strong interactions; let us recall that 11 is also the maximum number of dimensions required by supergravity theories [124]. For an exhaustive review of Kaluza{Klein theories we refer the reader to [125].

6) C:/DATA-NovoVivo-114-2020/Books-114-2020/Spacetime/ From Atoms to Higgs Bosons- Voyages in Quasi-Spacetime 2019.pdf

7) C:/DATA-NovoVivo-114-2020/Books-114-2020/Spacetime/ Gravitation and spacetime 2013.pdf Pag. 221 (238) - Chapter 6 - Riemannian Geometry

8) C:/DATA-NovoVivo-114-2020/Books-114-2020/Spacetime/ Minkowski Spacetime- A Hundred Years Later 2010.pdf

9) C:/DATA-NovoVivo-114-2020/Books-114-2020/Spacetime/ Particle Physics and Cosmology- The Fabric of Spacetime - 2007.pdf

10) C:/DATA-NovoVivo-114-2020/Books-114-2020/Spacetime/ Quantum Fluctuations of Spacetime 2005.pdf

11) C:/DATA-NovoVivo-114-2020/Books-114-2020/Spacetime/ Relativity and the nature of spacetime 2005.pdf

12) C:/DATA-NovoVivo-114-2020/Books-114-2020/Spacetime/ Relativity in curved spacetime - life without special relativity 2008.pdf

13) C:/DATA-NovoVivo-114-2020/Books-114-2020/Spacetime/ Relativity- Modern Large-Scale Spacetime Structure of the Cosmos 2009.pdf

14) C:/DATA-NovoVivo-114-2020/Books-114-2020/Spacetime/ Spacetime and Geometry- An Introduction to General Relativity 2019.pdf

15) C:/DATA-NovoVivo-114-2020/Books-114-2020/Spacetime/ Spacetime without Reference Frames 2018.pdf

16) C:/DATA-NovoVivo-114-2020/Books-114-2020/Spacetime/ Spacetime, Geometry and Gravitation - 2009.pdf

17) C:/DATA-NovoVivo-114-2020/Books-114-2020/Spacetime/ Spacetime-foundations of general relativity and differential geometry 1999.pdf

18) C:/DATA-NovoVivo-114-2020/Books-114-2020/Spacetime/ Spacetime-Foundations of General Relativity and Differential Geometry 2001.pdf

19) C:/DATA-NovoVivo-114-2020/Books-114-2020/Spacetime/ Springer Handbook of Spacetime - 2014.pdf

20) C:/DATA-NovoVivo-114-2020/Books-114-2020/Spacetime/ Symmetry, structure, and spacetime 2008.pdf

21) C:/DATA-NovoVivo-114-2020/Books-114-2020/Spacetime/ The Geometry of Spacetime- An Introduction to Special and General Relativity 2000.pdf

22) C:/DATA-NovoVivo-114-2020/Books-114-2020/ Pseudo-Reimannian Geometry, D-Invariants and Applications/14690.pdf

Pag.227 (259) Pseudo-Riemannian Submersions A submersion Pi : M -> B is a differentiable map between differentiable manifolds whose differential is everywhere surjective. The notion of sub- mersions is a fundamental concept in differential topology. The bundle map Pi : E -> B of a fiber bundle is an example of a submersion. Submersions appear often in classical differential geometry, e.g. a surface of revolution or a family of parallel surfaces in E 3 each leads in an obvious way to a submersion. IMPORTANT Riemannian submersions are submersions equipped with compatible Riemannian metrics. The theory of Riemannian submersions arose from [Hermann (1960); Nagano (1960); O'Neill (1966); Gray (1967)]. Rieman- nian submersions are natural generalizations of warped products, which occur widely in geometry. For instance, when a Lie group acts isomet- rically, freely and properly on a Riemannian manifold M, IMPORTANT the projection Pi : M -> B to the quotient space B = M/G equipped with quotient metric is a Riemannian submersion.

23) C:/ARQ-Others/Quantum/0123838746.pdf Classical and Quantum Information

Pag. 28(37) - Qubit

Figure 1.3 -

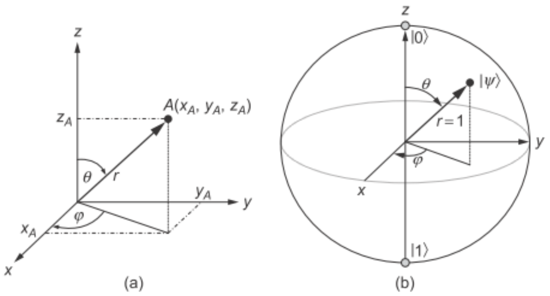

Figure 1.1: bs1.png

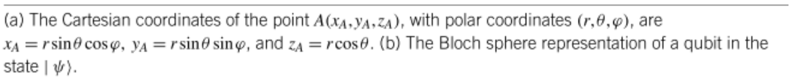

(a) The Cartesian coordinates of the point $A(x_A, y_A, z_A)$, with polar coordinates (r, θ, φ), are $x_A = r\sin\theta\cos\varphi$, $y_A = r\sin\theta\sin\varphi$, and $z_A = r\cos\theta$. (b) The Bloch sphere representation of a qubit in the state $|\,\psi\rangle$.

Figure 1.2: bs2.png

1.12 Pi - Calculation

Date: 20200718

1) The Pi Machine Source: https://wordplay.blogs.nytimes.com/2014/03/10/pi/

This coming Friday, 3.14 is Pi Day. Let's celebrate math's favorite number with a remarkable puzzle suggested several years ago by Alon Amit of Origami Logic. Dr. Amit and I were sitting on a couch in the old Palo Alto Facebook headquarters (Dr. Amit was working there at the time) surrounded by frenzy (the film \The Social Network" had just been released) when he told me something amazing: someone had recently figured out a completely original way to calculate pi. It was also a remarkably simple method, requiring just two balls and a wall. With these you have a pi machine.

I could barely believe it, so Dr. Amit continued. To calculate pi, take two identical balls. Put one near a wall and roll the other ball toward it. The first ball will hit the second, which will bounce off the wall and come back to hit the first ball. Click click click. Three collisions. The first digit of pi is a three. The machine works!

This isn't very convincing, of course. We need something a bit more precise. Can this machine tell us pi is 3.1? Or 3.14?

With a little modification, it can. A first ball that's 100 times as massive will create 31 clicks. 10,000 times as massive will create 314 clicks.

The machine can be used to generate (in theory, at least) as many digits of pi as you'd like. Simply amazing.

Our challenge for this week is | how does it work?

To solve this puzzle, it may help to experiment a bit with the following Pi Machine Gizmo created by Adam Bowman. (Thank you, Mr. Bowman!)

It may also help to try a simple physics simulation tool, like Algodoo. \Make and test your own contraptions with Algodoo."

Before closing, we have two exciting math events to announce.

If you're in the San Francisco Bay Area on Sunday, check out the Nueva School's 2014 Science, Technology, Engineering and Mathematics Fair. Among the speakers and activities: Josh Herbach of Google will give a presentation about the Google self-driving car project; I'll be there as well, giving a talk on creative genius, sharing what I've learned from conversations with artists, inventors and Olympians, as well as from so many here at Numberplay.

And if you're in New York | Tetsuya Miyamoto, famed inventor of the KenKen arithmetic and logic puzzle, will be at the Chappaqua Library on Thursday to speak about his innovative teaching philosophy, \The Art of Teaching Without Teaching."

That's it for this week's challenge. As always, use Gary Hewitt's Enhancer for an optimal Numberplay experience. And send your favorite puzzles and excess machinery to gary.antonick@NYTimes.com. Solution

How can two bouncing balls compute pi? Dave hit upon (and Hans and Mark | who simultaneously found another approach | also explored) a beautifully simple explanation that happened to be the same one I originally used: the collisions mark segments around a circle in a way that exactly determine pi. Although this explanation seems fairly obvious, it appears to be original | it's distinct from the method used by Gregory Galperin, who discovered this property nearly twenty years ago, and wasn't an approach Dr. Galperin was familiar with when I spoke to him a couple weeks ago by phone.

Dr. Galperin's own technique is ingenious | we'll get to that in a moment | but we'll start with the simpler approach.

I think it's easiest to forget about the bouncing balls for a moment and focus on pi. What does it mean to jot out pi to a particular number of digits? One way to think about it: pi to one digit, for starters, is simply the number of times you can fully wrap the diameter of a circle around that circle.

To figure this out we would pick up the diameter, wrap it around the first part of the circle, then mark where it ends. We'd then repeat the process until we ran out of circle, giving us a one-digit value of pi of 3.

Key is that first mark. If we can figure out another way to find it we'll have an alternative way way to determine pi. (We just need to ensure that this estimate is accurate enough, which means, in this case, it falls between the points created by three-sided and four-sided polygons.)

One alternative way: take the diameter and tilt it up against the right

Figure 1.3: pi machine solution 1.png

side of the circle, and then draw the red line as shown, creating a triangle. This red line gives us a mark that falls within the range we need. And if we rotate this triangle we find it marks the same spot.

A bit unwieldy, but it seems to work. For two digits of pi, we're looking for the number of times a tenth of the diameter can be wrapped around the circle. Following the same process gives us a very accurate estimate for that first mark. Repeating this we produce 31 of these tenth-diameter segments.

Figure 1.4: pi machine solution 2.png

Divide the diameter into 100 segments and the process will produce an initial segment that will fit 314 times around the circle. The method will give us pi to as many digits as we'd like.

But how does this relate to colliding balls? There's actually an exact connection: the balls will actually create, with a little rescaling, this same circle-and-line diagram, with the circle showing a graph of their kinetic energy and the line indicating momentum. A plot of ball velocities (the first ball's velocity on the x axis and the second ball's on the y) will plot out pi around the circle to whatever accuracy we'd like.

Take the first example, where the ball masses are equal. The balls start out as velocities (1,0) after the first ball is given a push. After the first collision the first ball stops and the second starts (0,1); the second then

Figure 1.5: pi machine solution 3.png

hits the wall and reverses (0,-1) and finally hits the first ball again (-1,0). Three collisions (though the pattern doesn't jump around the circle counter-clockwise { instead it zig-zags up and down around the circle) and three intersections after that first point.

In the second example, where the first ball is now 100 times more massive than before, we get a circle with a more vertical line that will create 31 intersections. (See Dave's comments for detail.)

Dr. Galperin's approach was also geometric but very different (using an unfolding geodesic), building on prior related insights. Dr. Galperin, who studied under well-known Russian mathematician Andrei Kolmogorov, had recently written (with Yakov Sinai) extensively on ball collisions, realized just before a talk in 1995 that a plot of the ball positions of a pair of colliding balls could be used to determine pi. (When he mentioned this insight in the talk, no one in the audience believed him.) This finding was ultimately published as \Playing Pool With Pi" in a 2003 issue of Regular and Chaotic Dynamics.

I recently reached Dr. Galperin by phone and asked about the key insight behind this idea and how he in particular happened to see it. \The Greeks could not have discovered this," he said, \because they didn't know about kinetic energy and momentum. Leibniz, who proposed the idea of kinetic energy, did not work in this area. As for others | mathematics is a broad field. This came to me as the collision of several ideas in areas I have worked deeply in."

Thank you, Dr. Galperin, and thank you as well to everyone who participated in this week's challenge: Mark, Gary, Ibabel, Hans, Ravi, Dave, Mark, Giovanni Ciriani, Kurt and Sheldon Ross.

2) Então por que blocos colidindo computam Pi?
Source: https://www.youtube.com/watch?v=jsYwFizhncE

3) Billiard Balls Count Pi

Source: https://www.youtube.com/watch?v=4UhPTrHIprk

4) https://www.maths.tcd.ie/ lebed/ Galperin. Playing pool with pi.pdf
Source:
C:/DATA-NovoVivo-119-2020/Books-119-2020/Pi-Machine/ Galperin. Playing pool with pi.pdf

5) Finding a consistente constant
Source:
https://physicsworld.com/a/finding-a-consistent-constant/

08 Jul 2020 Taken from the July 2020 issue of Physics World. Members of the Institute of Physics can enjoy the full issue via the Physics World app.

The Planck mission gave us the most precise value of the Hubble constant to date by measuring the cosmic microwave background. But studies made since using different methods provide different values. Keith Cooper investigates the discrepancies and asks what it might mean for cosmology

When astronomer Edwin Hubble realized that the universe is expanding, it was the greatest cosmological discovery of all time. The breakthrough acted as a springboard to learning about the age of the cosmos, the cosmic microwave background (CMB) radiation, and the Big Bang.

Hubble was the right man in the right place at the right time. In April 1920 his fellow astronomers Harlow Shapley and Heber D Curtis famously debated the size of the universe, and the nature of spiral nebulae, at the Smithsonian Museum of Natural History in the US. Within four years, Hubble had the answer for them.

Using the 100-inch (2,5 m) Hooker Telescope at Mount Wilson Observatory in California, Hubble was able to identify Cepheid variables – a type of star whose period-luminosity relationship allows accurate distance determination – in the spiral nebulae, which allowed him to measure them as extragalactic. The finding meant the Milky Way was not the entire universe, but just one galaxy among many.

Soon, Hubble had discovered that almost all of these galaxies are moving away from us – their light is redshifted via the Doppler effect. These observations were seized upon by the Belgian cosmologist Georges Lemaître, who realized that they implied that the universe is expanding. Independently, both Hubble and Lemaître derived a mathematical relationship to describe this expansion, subsequently known as the Hubble–Lemaître law. It says that the recession velocity (v) of a galaxy is equal to its distance (D) multiplied by the Hubble constant (H0), which describes the rate of expansion at the current time.

In the nearly 100 years since Hubble's discovery, we've built up a detailed picture of how the universe has developed over time. Granted, there are still some niggles, such as the identities of dark matter and dark energy, but our understanding reached its peak in 2013 with the results from the Planck mission of the European Space Agency (ESA).

Launched in 2009, Planck used microwave detectors to measure the anisotropies in the CMB { the slight temperature variations corresponding to small differences in the density of matter just 379,000 years after the Big Bang. The mission revealed that only 4.9% of the universe is formed of ordinary baryonic matter. Of the rest, 26.8% is dark matter and 68.3% is dark energy. From Planck's observations, scientists also deduced that the universe is 13.8-billion years old, confirmed that the universe is flat, and showed how baryonic acoustic oscillations (BAOs) { sound waves rippling through the plasma of the very early universe, resulting in the anisotropies { neatly matched the large-scale structure of matter in the modern universe, which we know has expanded according to the Hubble{Lemaître law.

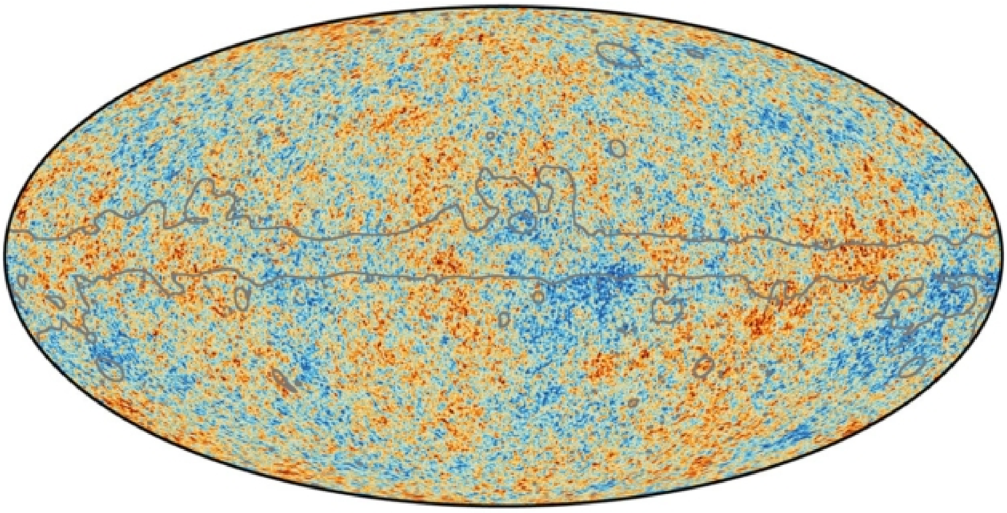

Figure 1.6: Planck-2018-T-CMB.jpg

Mapping the universe By measuring the temperature anisotropies of the cosmic microwave background { as demonstrated in this 2018 map { the Planck mission provided the most precise measurement of the Hubble constant to date: 67.4 /km/s/mpc with just 1% uncertainty. (Courtesy: ESA and the Planck Collaboration)

Planck's view of the CMB is the most detailed to date. Plus, by mapping our best cosmological models to fit Planck's observations, a whole host of cosmological parameters have come tumbling out, including H0. In fact, scientists were able to extrapolate the value of H0 to the greatest precision ever, finding it to be 67.4 /km/s/mpc, measured to an uncertainty of less than 1%. In other words, every stretch of space a million parsecs (3.26 million light-years) wide is expanding by a further 67.4 km every second.

With Planck's results, we thought that this picture was complete, and that we knew exactly how the universe had expanded over those 13 billion e 800 millions years. Yet it turns out that we may be wrong (see box \Not a constant constant over time").

Not a constant constant over time

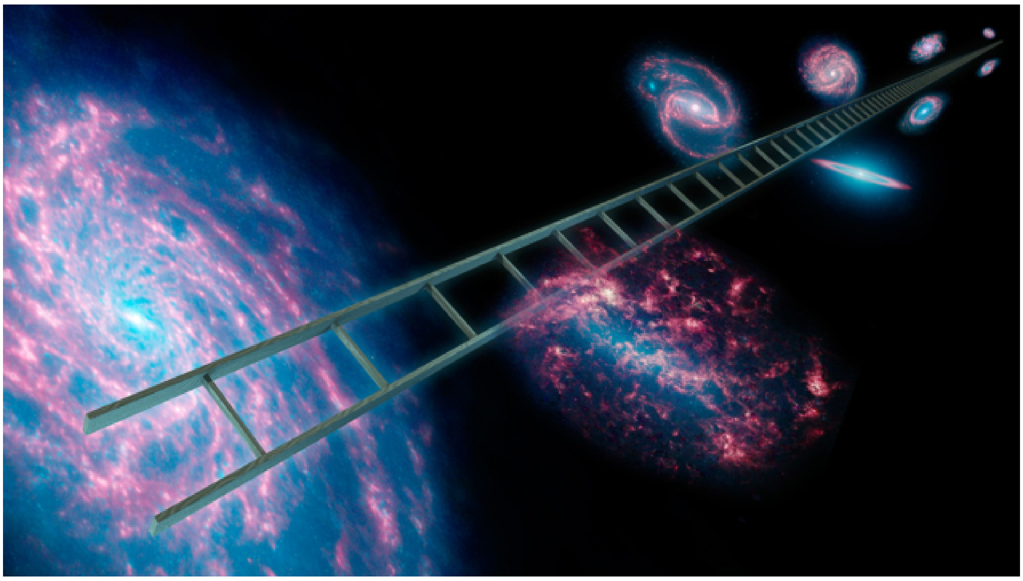

Figure 1.7: Distance Ladder

Describing H0 as \the Hubble constant" is a little bit of a misnomer. It is true that it is a constant at any given point in time since it describes the current expansion. However, the expansion rate itself has changed throughout cosmic history { H0 is just the current value of a broader quantity that we call the Hubble parameter, H, which describes the expansion rate at other times. Therefore, while the local measurement, H0, should only have one value, H can take on different values at different points in time. Prior to the discovery of dark energy, it was assumed that the expansion of the universe was slowing down, and hence H would be larger in the past than in the future. On the face of it, since dark energy is accelerating the expansion, you might therefore expect H to increase with time, but that's not necessarily the case. Rearranging the Hubble{Lemaître law such that h - v over d , where v is an object's recession velocity, we see that H depends strongly on distance D, since in a universe with accelerating expansion, the distance is increasing at an exponential rate. The bottom line is that H is likely to be decreasing with time, which is what we think is happening, even though the recession velocity and hence the expansion are still increasing with distance.

Moving parts

Traditionally, H0 is determined by measuring the distance and recession velocity of galaxies using \local" standard candles within those galaxies. These include type 1a supernovae { the explosion of white-dwarf stars with a certain critical mass { and Cepheid variable stars. The latter have a robust period{luminosity relationship (discovered by Henrietta Swan Leavitt in 1908) whereby the longer the period of the star's variability as it pulses, the intrinsically brighter the star is at peak luminosity. But as astrophysicist Stephen Feeney of University College London explains, these standard candles have many \moving parts", including properties such as stellar metallicity, differing populations at different redshifts, and the mechanics of Cepheid variables and type Ia supernovae. All the moving parts lead to uncertainties that limit the accuracy of these observations and we have seen over the years how the resulting calculations of H0 can vary quite a bit.

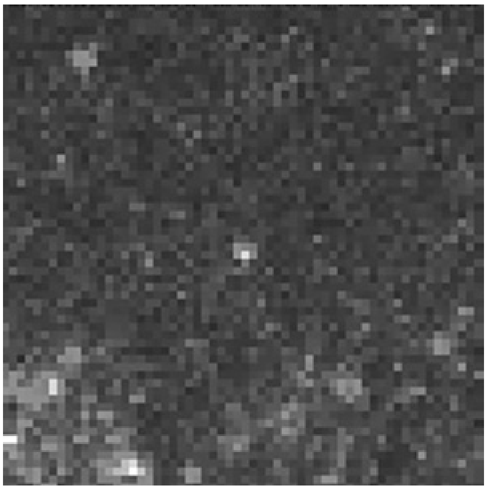

Figure 1.8: cepheid-variable.jpg

A star's pulse Cepheid variable stars { such as this one in galaxy M100 { have a strong period-luminosity relationship that allows astronomers to work out how far away they are. (Courtesy: Wendy L Freedman, Observatories of the Carnegie Institution of Washington/NASA)

On the other hand, Planck's value of H0 is a relatively straightforward measurement { even if it does depend on the assumption that lambda cold dark matter (ACDM) cosmology { which incorporates the repulsive force of dark energy (A) and the attractive gravitational force of cold dark matter

(CDM) { is correct. Still, when all the known uncertainties and sources of error are taken into account, Planck's value of H0 is to a precision greater than ever before, to just 1%.

Although the Planck value of H0 is calculated from measurements of the CMB, it is important to note that it is not the expansion rate at the time of the CMB's creation. Rather, \think of the CMB measurement as a prediction instead", says Feeney. It extrapolates from what the CMB tells us the universe was like 379,000 years after the Big Bang, factoring in what we know of how the universe would have expanded in the time since based on the Hubble{Lemaître law and ACDM, to arrive at an estimate of how fast the universe should be expanding today. In other words, however we measure H0, whether it be with the CMB or more local measurements of Cepheid variables and supernovae, we should get the same answer. A spanner in the works

In 2013, when the Planck measurement was revealed, this wasn't a problem. Although local measurements differed from the Planck value, their uncertainties were still large enough to accommodate the differences. The expectation was that as the uncertainties become smaller over time with more sophisticated measurements, the local measured value would converge on the Planck value.

However, in 2016 a major milestone was reached in local measurements of H0 that brought our understanding of the universe into question. It involved Adam Riess of Johns Hopkins University, US, who in 1998 had co-discovered dark energy using type Ia supernovae as standard candles to measure the distance to receding galaxies based on how bright the supernovae appeared. He was now leading the cumbersomely named SH0ES (Supernova, H0, for the Equation of State for dark energy) project, which was set up to calibrate type Ia supernovae measurements to determine H0 and the behaviour of dark energy. In order to provide this calibration, the SH0ES team used a lower rung on the \cosmic distance ladder", namely Cepheid variables. The project aimed to identify these pulsating stars in nearby galaxies that also had type Ia supernovae, which would mean that the Cepheid distance measurement could then be used to calibrate the supernova distance measurement, and this new calibration could in turn be used on supernovae in more distant galaxies. The method produced a value of H0 with an uncertainty of just 2.4%.

A star's death A type 1a supernova { as illustrated here in an artist's impression { results from the destruction of a white dwarf star above a critical mass. Their standardizable luminosity makes them a useful rung on the cosmic distance ladder. (Courtesy: ESA/ATG medialab/C Carreau)

Riess and his group were stunned by their result, however. Using the local

Figure 1.9: Gaia-SNIa-Artist-Impression.jpg

cosmic distance ladder, they inferred a value of H0 as 73.2km/s/mpc. With the vastly reduced uncertainties, there was no way this could be reconciled with Planck's measurement of 67.4km/s/Mpc. If these results are correct, then there is something deeply wrong with our knowledge of how the universe works. As Sherry Suyu of the Max Planck Institute for Astrophysics in Garching, Germany, says, \We may need new physics."

Rather than assume Planck's high-precision H0 is wrong and therefore rush to take apart the fundamentals of our best cosmological models, a scientist's first instinct should be to test whether there has been an experimental mistake that means something about our measurement of the cosmic distance ladder is awry.

\That's where my instinct still lies," admits Daniel Mortlock of Imperial College London, UK, and Stockholm University, Sweden. Mortlock works in the field of astro-statistics { that is, drawing conclusions from incomplete astrophysical data, and accounting for various kinds of uncertainty and error in the data. It's worth remembering there are two types of error in any measurement. The first is statistical error { errors in individual measurements, for example read-out noise from the detector, or uncertainties in the sky background brightness. Statistical errors can be reduced by simply increasing your sample size. But the other kind of errors { systematic errors { aren't like that. \It doesn't matter if you have a sample five times, or 10 times, or

50 times as large, you just get this irreducible uncertainty," says Mortlock. An example of a systematic error might be the reddening of a star's light by intervening interstellar dust { no matter how often you measure the star's brightness, the star's light would always be obscured by the dust, and its effect would increase the more you measure it.

Mortlock has considered that this could be what is happening with the local measurements of H0 { there may be some systematic error that astronomers have not yet identified, and if found, the tension between the Planck measurement and the local measurement of H0 could go away. However, Mortlock acknowledges that the evidence for the discrepancy in the H0 values being real \has been steadily growing more convincing". Einstein to the rescue

Presented with such an exceptional result, astronomers are double-checking by measuring H0 with other, independent means that would not be subject to the same systematic errors as the Cepheid variable and type Ia supernovae measurements.

One of these can be traced back to 1964, when a young astrophysicist by the name of Sjur Refsdal at the University of Oslo, Norway, came up with a unique way to measure the Hubble constant. It involved using a phenomenon predicted by Einstein but which at the time had not been discovered: gravitational lenses.

The general theory of relativity describes how mass warps space, and the greater the mass, the more space is warped. In the case of so-called \strong gravitational lensing", massive objects such as galaxies, or clusters of galaxies, are able to warp space enough that the path of light from galaxies beyond is bent, just like in a glass lens. Given the uneven distribution of mass in galaxies and galaxy clusters, this lensing can result in several light paths, each of slightly different length.

Refsdal realized that if a supernova's light passes through a gravitational lens, then its change in brightness would be delayed by different amounts in each of the lensed images depending on the length of their light path. So, image A might be seen to brighten first, followed a few days later by image B, and so on. The time delay would tell astronomers the difference in the length of the light paths, and the expansion of space during those time delays would therefore allow H0 to be measured.

Einstein's lenses Gravitational lensing, as seen in galaxy cluster Cl 0024+17, presents another means of determining cosmic distances, and therefore calculating Hubble's constant. (Courtesy: NASA, ESA, M J Jee and H Ford (Johns Hopkins University))

Unfortunately, even after the first gravitational lens was discovered in 1979, it turned out that gravitationally lensed supernovae are exceptionally

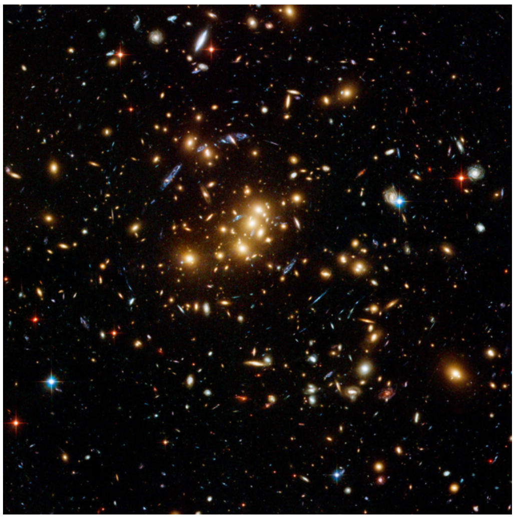

Figure 1.10: lensing.jpg

rare. Instead, quasars { luminous active galactic nuclei that also exhibit brightness variations { have been found to be the more common lensed object. That was what led Suyu to launch a project in 2016 to study lensed images of quasars in order to provide an independent measure of H0. It goes under the even more cumbersome name of H0LiCOW, which stands for H0 Lenses in COSMOGRAIL's Wellspring, where COSMOGRAIL refers to a programme called the COSmological MOnitoring of GRAvitational Lenses, led by Frederic Courbin and Georges Meylan at the Ecole Polytechnique Federale de Lausanne.

Throughout this analysis, Suyu and team kept the final result hidden from themselves { a technique known as blind data analysis { to avoid confirmation bias. It was only at the very end of the process, once they had completed all their data analysis and with their paper describing their observations almost completely written, that they revealed to themselves the value of H0 that they had measured. Would it come out in favour of Planck, or would it boost the controversial SH0ES result?

The value that they got was 73.3 km/s/Mpc, with an uncertainty of 2.4%. \Our unblinded result agrees very well with the SH0ES measurement, adding further evidence that there seems to be something going on," Suyu tells Physics World.

It's still too early to claim the matter settled, however. The initial H0LiCOW analysis involved only six lensed quasars, and efforts are being made to increase the sample size. Suyu is also returning to Refsdal's original idea of using lensed supernovae, the first example of which was discovered by the Hubble Space Telescope in 2014, followed by a second in 2016. Hundreds are expected to be discovered by the Vera C Rubin Observatory in Chile, formerly known as the Large Synoptic Survey Telescope, which will begin scientific observations in October 2022.

\It would be really interesting if the H0LiCOW measurements can be shown to be correct and in agreement with the Cepheid measurements," says Feeney, who for his part is pursuing another independent measurement of H0, using another phenomenon predicted by Einstein: gravitational waves.

On 17 August 2017 a burst of gravitational waves resulting from the collision of two neutron stars in a galaxy 140 million light-years away triggered the detectors at the Laser Interferometry Gravitational-wave Observatory (LIGO) in the US and the Virgo detector in Italy. That detection has allowed Feeney and a team of other astronomers including Mortlock and Hiranya Peiris, also of University College London, to revive an idea, originally proposed by Bernard Schutz in 1986, to use such events to measure the expansion rate of the universe.

The strength of the gravitational waves indicates how distant a neutron-star merger is, but the merger also produces a handy burst of light known as a kilonova. This can be used to pinpoint the host galaxy, the redshift of which in turn provides the galaxy's recession velocity. Feeney and Peiris have estimated that a sample of 50 kilonovae would be required to derive an accurate determination of H0, but Kenta Hotokezaka of Princeton University, US, and colleagues have found a way to speed this up. They point out that we would see the gravitational waves from a neutron-star merger at their strongest if we looked perpendicular to the plane of the collision. The merger produces a relativistic jet that's also moving perpendicular to the plane, so measuring the angle at which we see the jet will tell us how large our viewing angle is from the plane, and therefore allow a determination of the true strength of the gravitational waves and hence the distance. Hotokezaka estimates that a sample size of just 15 kilonovae studied in this fashion will be enough to provide an accurate assessment of H0. Unfortunately, so far astronomers have observed just one kilonova, and based on that sample of one, Hotokezaka calculates H0 to be 703km/s/Mpc, with a big uncertainty of 10%.

Other approaches abound

Beta ceti A red rung The red giant star B Ceti has begun fusing helium, and can therefore be used as a standard candle. (Courtesy: NASA/CXC)

Cepheid variables and type 1a supernovae are common rungs on the cosmic distance ladder that are used to find a local value for Hubble's constant. But researchers led by Wendy Freedman of the University of Chicago have used another. They looked at the brightness of red giant stars that have be-

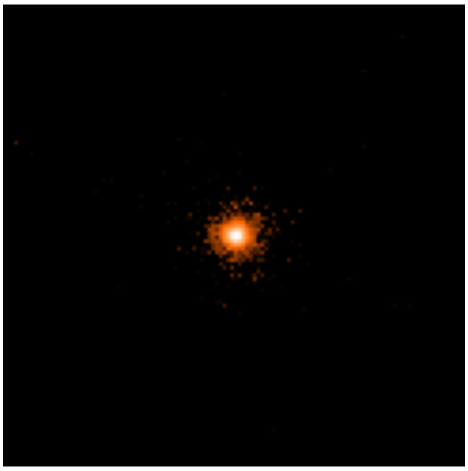

Figure 1.11: Bceti-xray.jpg

gun fusing helium in their cores, to measure the distance to galaxies in which these stars can be seen. Initially they calculated H0 as 69.8km/s/Mpc { but then things got complicated. Their data was reanalysed by Wenlong Yuan and Adam Riess to account for dust reddening, resulting in a revised measurement of 72.4km/s/Mpc with an uncertainty of 1.45%. However, Freedman's team have performed the same reanalysis and got a value of 69.6km/s/Mpc, with an uncertainty of 1.4%, so the jury is still out on that one.

Meanwhile, the Megamaser Cosmology Project makes use of radio observations tracking water masers in gas orbiting supermassive black holes at the centres of distant galaxies. The angular distance that the masers traverse on the sky allows for a straightforward geometrical distance measurement to their host galaxy, producing a value of 73.9km/s/Mpc for H0 with an uncertainty of 3%.

Changing the game

Taken together, all the available evidence seems to be pointing towards the dichotomy between the local measurements and the Planck measurement being real { and not some unidentified systematic error. However, a greater sample size is required before these results can be considered truly robust. With new observatories coming online, we could have the required observations within 10 years.

\If either the gravitational waves or the lensing methods give very strong results that match the SH0ES result, then I think that will change the game," says Mortlock.

However, don't expect a speedy resolution. After all, we're still grappling with the nature of dark matter and dark energy, and current research is focused on trying to identify the dark-matter particle and trying to characterize the behaviour of dark energy. \Whatever is going on with the Hubble constant is still several steps behind those," says Mortlock. \People are still debating whether the effect is real."

The value of H0 will have many consequences. It will dictate the age of the universe, and the history of how the universe expanded and allowed large-scale structure to form

One way or another, figuring out whether the discrepancy in our measurements of H0 is real or not will have significant repercussions for cosmology. Feeney describes the local and Planck values of H0 as \a really potent combination of measurements, because you are constraining the universe now and how it looked 13.8 billion years ago, so you pin down the universe at both ends of its evolution".

The value of H0 will have many consequences. It will dictate the age of the universe (a higher H0 would mean the universe might be substantially younger than 13.8 billion years, which would contradict the age of some of the oldest stars that we know). It would also influence the history of how the universe expanded and allowed large-scale structure to form. And if new physics is needed, as Suyu suggests, then it's impossible as yet to say how dramatic an effect that would have on cosmology, since we don't yet know what shape that new physics might take.

\It would be something in addition to our current ACDM model," says Suyu. \Maybe we are missing some new, very light and relativistic particle, which would change Planck's measurement of H0. Or it could be some form of early dark energy that's not in our current model."

Or it could be neither of these, but instead something we've not thought of yet. The prospect is tantalizing researchers, but Suyu warns about jumping the gun.

\First we need to get the uncertainties down to the 1% level on multiple methods to see if this tension is real," she says. So we need to be a little bit patient, but if we come back in 10 years' time, we may find that the universe is suddenly a very different place indeed.

Author: Keith Cooper is a science journalist based in the UK. His book The Contact Paradox: Challenging Our Assumptions in the Search for Extraterrestrial Intelligence (Bloomsbury Sigma) is available now. Follow him @21stCenturySETI on Twitter

Your Book: C:/DATA-NovoVivo-119-2020/Books-119-2020/Pi-Machine/ Keith Cooper - The Contact Paradox - Challenging our Assumptions in the Search for Extraterrestrial Intelligence 2019.epub

1.13 Geometry - Approximation

Date: 20200719

1) C:/DATA-NovoVivo-119-2020/Books-119-2020/ Combinatorial and Computational Geometry/0521848628.pdf

Geometric Approximation via Coresets PANKAJ K. AGARWAL, SARIEL HAR-PELED, AND KASTURI R. VARADARAJAN Abstract. The paradigm of coresets has recently emerged as a powerful tool for efficiently approximating various extent measures of a point set P . Using this paradigm, one quickly computes a small subset Q of P , called a coreset, that approximates the original set P and and then solves the problem on Q using a relatively inefficient algorithm. The solution for Q is then translated to an approximate solution to the original point set P . This paper describes the ways in which this paradigm has been successfully applied to various optimization and extent measure problems.

Pag.06 (16) - Figure 2. Top left: A grid based algorithm for constructing an e-kernel. Top right: An improved algorithm. Bottom: Correctness of the improved algorithm.

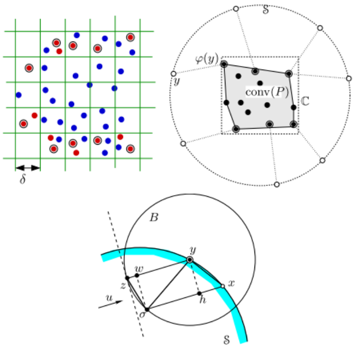

Figure 1.12: ga-pag6.png

2) C:/DATA-NovoVivo-119-2020/Books-119-2020/ Functions of a Complex Variable, 5th Edition/4715.pdf

3) C:/DATA-NovoVivo-120-2020/Books-120-2020/ Geometry from Euclid to Knots (Dover Books on Mathematics)/ [Dover Books on Mathematics] Saul Stahl - Geometry from Euclid to Knots (2010, Dover Publications) - libgen.lc.epub

Obs.: Livro todo!!!

4) C:/DATA-NovoVivo-120-2020/Books-120-2020/ Geometry from Euclid to Knots (Dover Books on Mathematics)/ Saul Stahl - A Gateway to Modern Geometry - The Poincare Half-Plane 2008.pdf

Obs.: Livro todo!!!

Pag. 67 (79) - Theorem 5.1.1 Every Euclidean circle in the upper half-plane is also a hyperbolic circle.

5) C:/DATA-NovoVivo-120-2020/Vdeos/Carl Munck - The Code/ Carl Munck The Code -All Video-s- Enhanced AudioVideo.mp4

Time: 17:11 - Spherical Math,

Time: 19:46 - 360 X Pi = 1130.973355

Time: 25:11 - 360 / 4Pi = 28.64788976

Time: 25:20 - R / 2 = 28.64788976

1.14 Topology - Geometry

Date: 20200723

1) C:/DATA-NovoVivo-120-2020/Books-120-2020/ Puzzles to Unravel the Universe/Puzzles to Unravel the Universe.pdf

Pag. 87 - Figure 39 e elevado a in theta for module z > igual 0. The conclusion is that for really big circles, g wraps the big circle on the complex plane around the unit circle n times (Fig.39), because as theta varies from 0 to 2pi, g varies from 1 to exp(2piin).

Pag. 93 - Gravitational Lenses Einstein offered a geometric explanation of gravity. Rather than viewing gravity as an attractive force between massive objects, which is how Newton described it, Einstein’s theory is based on the notion of curvature. The presence of mass, he teaches us, literally causes the fabric of space and time to curve or warp. The warping of space and time affects the motions of objects nearby, and that, in essence, is the phenomenon we call gravity. To explore this idea, consider two points on the sphere. There is a unique shortest path (or geodesic) on the sphere between these two points (Fig.45). Although this path is not \straight" in the conventional sense, it is still the shortest path that lies in the sphere. Actually, there are exceptions: there are infinitely many shortest-length paths between antipodal points. Instead we can start from a given point on the sphere and go in a given direction in as straight a manner as we can. This is what is called a geodesic. The path we thus get will be a great circle on the sphere.

FIGURE 45. There is normally only one shortest path between any two points on the sphere.

Other kinds of situations can arise. On the torus, a donut shaped object as in Fig.46, there are topologically inequivalent geodesics paths between points lying on opposite sides of a cross-sectional circle.

FIGURE 46. On a torus, there are two shortest paths between points sitting on opposite sides of a cross-sectional circle.

Einstein's theory predicts that light always travels in the shortest path between two points.However, the curvature of space-time means that the path may not look straight in Euclidean space. The theory tells us, for example, that the sun, being a massive body, should bend passing light. This prediction was the first experimentally confirmed prediction of Einstein's general theory of relativity, when the bending of light around the sun was indeed observed during an eclipse.

2) C:/DATA-NovoVivo-120-2020/Books-120-2020/ An Introduction to Metric Spaces/An Introduction to Metric Spaces.pdf
Obs.: Livro todo!!!

Pag. 55 (72) - Chapter 2 Metric Spaces In this chapter, we shall first review our experience in the real number system and measuring what we call in natural language as \distances", and then generalize this concept to define metric spaces. Once we achieve this, all other concepts that arise due to notion of distances, such as sequences and their convergence and continuity of functions, will be generalized in the succeeding chapters.

Symmetry, Homogeneity, Triangle inequality.

3) C:/DATA-NovoVivo-120-2020/Books-120-2020/Math-Books-2/0387982892-1461268672.pdf
Obs.: Livro todo!!!

Pag. 72 (83) - 3.2 Intersections The difference between analytic and synthetic geometry can be illustrated with Euclid's very first proposition, the construction of the equilateral triangle on a line segment AB. Recall from Section 2.2 that Euclid did this by finding the intersection of two circles, one with center A and radius AB and the other with center Band radius BA. It follows that each point of intersection is distant from both A and B by the length of AB and hence forms an equilateral triangle. Analytic geometry takes its cue from this construction, but it also finds the intersections by finding the common solutions of the equations to the two circles.

IMPORTANT Sqr 3

4) C:/DATA-NovoVivo-120-2020/Books-120-2020/Math-Books-2/ Algebraic Curves over a Finite Field.pdf

5) C:/DATA-NovoVivo-120-2020/Books-120-2020/Math-Books-2/ Geometry of Digital Spaces.pdf
Obs.: Livro todo!!!

Pag. 18(29) - Fig. 1.7.1

Figure 1.7.1. The route of a Modell Fat Fly on a single sugar cube. The numbers indicate the order in which the faces are visited.

Pag. 32 (43) - Figure 1.9.1. An arrangement of four sugar cubes (one face of each is painted gray to clarify the location of a tunnel in the center of the four sugar cubes), providing a counterexample which shows that the Algorithms for Fat Flies Models el and e2 do not satisfy Aim 1.6.1.

6) C:/DATA-NovoVivo-120-2020/Books-120-2020/Math-Books-2/ The Geometries of Visual Space.pdf

Pag. 103(116) - Chapter 6 - Factors Affecting Size Constancy Much of the spatial perception literature is not quantitative in a way that is de- tailed enough to really say much about the geometry of visual space. Many researchers and theorists seek to describe the cues used for depth perception or the internal mechanisms that underlie spatial representations using standard experi- mental techniques that vary conditions across groups (Gibson, 1950; Herschen- son, 1999; Rock, 1983). These types of studies do not generate data that is sufficiently rich to support detailed mathematical modeling. Obviously, the last chap- ter, which looked at the direct estimation of spatial metrics, reported on one body of research that is an exception to this rule. Another area in space perception that has also yielded a numerically rich database is the size-constancy literature. This chapter will discuss the size-constancy paradigm and lay out some of the factors that influence size constancy. Following this, I offer a review of the lit- erature that presents major theoretical positions on size constancy in chronologi- cal order along with empirical work that bears on each theory (some of which may be of more recent origin). The chapter will then summarize the literature and present my own theory that can explain much of this literature in a way that is consistent with the direct estimation literature discussed in the previous chapter.

Pag. 113(126) -IMPORTANT Figure 6.2. Average ratio of comparison to standard as a function of distance to the standard in Gilinsky's (1955) data. Objective data show overconstancy and are represented with open symbols. Projective data shows strong underconstancy and are represented by closed symbols. Circles collected with 42 in standard, squares with 55 in standard, upward pointing triangles with 66 in standard, and downward pointing triangles with 78 in standard. Solid lines represent best fitting full model curves. The dashed line represents perfect constancy.

7) C:/DATA-NovoVivo-120-2020/Books-120-2020/Math-Books-2/ A Course in Ordinary Differential Equations, Second Edition.pdf
Obs.: Livro todo!!!

Pag. 204 (221) - Forces acting upon the mass: In this coordinate system, forces tending to pull the mass downward are positive, while those tending to pull it upward are negative.

8) C:/DATA-NovoVivo-120-2020/Books-120-2020/Math-Books-2/ Geometry, Language, And Strategy.pdf
Obs.: Livro todo!!!

Pag. 99 (111) - 6.1 Fair Games The simplest examples are fair games, games in which each player receives at most a zero payoff. In this case I take both strategic lengths 7], r 2 to be inactive, as well as time in the central frame and the angular deviation strategy 0 a defined in the central frame. A consequence is that the time-choice in the player frame is also inactive. The gravitational potential and pressure are functions of a single active variable, the radial deviation strategy r m associated with the fair game. 6.1.1 No gravity or pressure I start with the gravitational potential v = J ln g and the pressure set to zero and a representative strategic mass or time-scale m 0 =4. For the fair game examples the payoff matrix is chosen to be:

9) C:/DATA-NovoVivo-120-2020/Books-120-2020/Math-Books-2/ Geometrical Properties of Vectors and Covectors - An Introductory Survey of Differentiable Manifolds, Tensors and Forms 2006.pdf
Obs.: Livro todo!!!

Pag. 13(21) - Differentiable Manifolds 3.1 Basic definitions An n-dimensional manifold M is a Hausdorff topological space locally homeomorphic to Rn. This means that M can be described locally by coordinates attributed by means of a chart around each point.

Fig. 3.1 Presentation of curves and functions.IMPORTANT

Fig. 3.2 Coordinate functions and local coordinates.IMPORTANT

10) C:/DATA-NovoVivo-121-2020/Books-121-2020/Projective geometry and projective metrics/projectivegeomet01vebluoft.pdf

11) C:/DATA-NovoVivo-121-2020/Books-121-2020/Projective geometry and projective metrics/projectivegeomet02vebluoft.pdf

12) C:/ARQ-BASICS/10/Plain Plane Geometry/Amol Sasane - Plain

Plane Geometry-WSPC (2015).pdf Copy: C:/DATA-NovoVivo-121-2020/Books-121-2020/Important/Amol Sasane - Plain Plane Geometry-WSPC (2015).pdf

Pag. 154 - Theorem 5.17 (Bramhagupta's Formula). If a cyclic quadrilateral has sides of lengths a, b, c, d, then its area is given by...

13) C:/DATA-NovoVivo-121-2020/Books-121-2020/Important/ Obs.: Diretorio todo!!!

14) C:/DATA-NovoVivo-121-2020/Books-121-2020/Important/ Einstein-in-Matrix-Form.pdf Obs.: Livro todo!!!

15) C:/DATA-NovoVivo-121-2020/Books-121-2020/Important/newlecturesGR.pdf Obs.: Livro todo!!! - 988 pages. Importantíssimo (2019) IMPORTANT
Assunto: Lecture Notes on General Relativity

http://www.blau.itp.unibe.ch/GRLecturenotes.html

C:/DATA-NovoVivo-121-2020/Books-121-2020/Important/newlecturesGR-2.pdf (2020)

16) C:/DATA-NovoVivo-121-2020/Books-121-2020/Important/zakon-analysisII-a4-one.pdf

Pag. 97(109) -Chapter 7 - Volume and Measure

17) C:/DATA-NovoVivo-121-2020/Books-121-2020/Important/ claimofleibnitzt00slomrich.pdf

THE CLAIM OF LEIBNITZ INVENTION OF THE DIFFERENTIAL CALCULUS.

18) C:/DATA-NovoVivo-121-2020/Books-121-2020/Differential Equations A Modern Approach with Wavelets/Differential Equations.pdf Obs.: Livro todo!!!

Pag. 50(65) - FIGURE 1.12 - A tractrix.

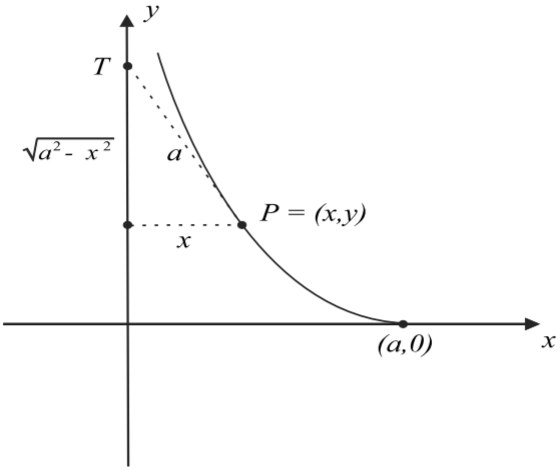
y
T
$\sqrt{a^2 - x^2}$
a
P = (x,y)
x
(a,0)
x

Figure 1.13: df.png

19) C:/DATA-NovoVivo-121-2020/Books-121-2020/ Introduction to the Maths and Physics of the Solar System/ Introduction-to-the-Maths-and-Physics-of-the-Solar-System.pdf Assunto: Introduction to the Maths and Physics of the Solar System. Obs.: Livro todo!!!

Pag. 86(99) - FIGURE 3.2 Geometry of the principle of least action (Lagrange).

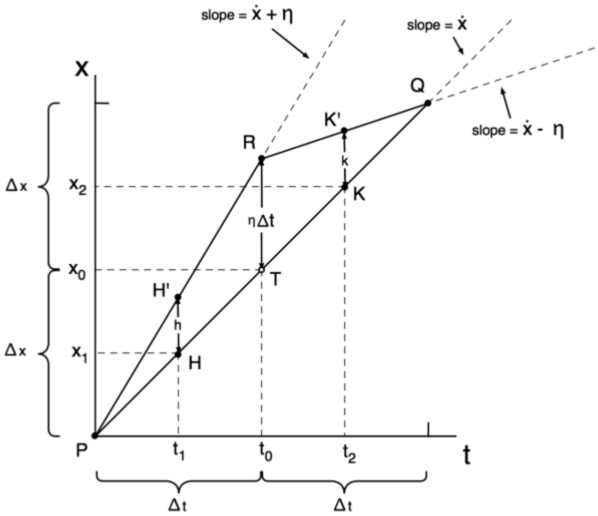
slope = $\dot{x} + \eta$
slope = $\dot{x}$
slope = $\dot{x} - \eta$
X
Q
K'
R
k
Δx
x_2
K
ηΔt
x_0
T
H'
h
Δx
x_1
H
P
t_1
t_0
t_2
t
Δt
Δt

Figure 1.14: mp.png

20) C:/DATA-NovoVivo-121-2020/Books-121-2020/Partial Differential Equations Classical Theory with a Modern Touch/1108839800.pdf Obs.: Livro todo!!! Assunto: Partial Differential Equations Classical Theory with a Modern Touch.

Pag. 17(38) - 2.2 MULTIPLE INTEGRALS AND DIVERGENCE THEOREM In this section, we state two important results, namely, the Green's theorem and the divergence theorem. These are extensively used in the analysis of PDE. In a sense these theorems are the higher dimensional versions of the familiar fundamental theorem of calculus and integration by parts formula in one dimension. Hence they are also important by themselves and are particular cases of a general result called Stokes theorem on smooth manifolds. However, even for the statements of these theorems, we need a lot of machinery from differential geometry; some of these are the concept of a smooth domain, smooth surface (curve in two dimensions), integration over these objects, tangent space and smoothly varying unit normal to a surface. In this brief introduction to the subject, it will not be possible to present all these concepts in a rigorous manner. The reader should refer to the cited references for much deeper treatment of these topics. Our discussion here is just explaining these important theorems in some simple geometric situations. 2.2.1 Multiple Integrals In this sub-section, we briefly discuss the notion of multiple integrals, iterated integrals and integration by parts in multiple integrals leading to Green's theorem in two dimensions and divergence theorem (called Gauss-Ostrogadskii formula)in more than two variables. The general references for multiple integrals and their applications are Rund (1973), Spivak (1965), Apostol (2002), Taylor and Mann (1983), and Widder (1961). The Riemann integral of a bounded function in a rectangular domain in R n is very similar to the one-dimensional integral. However, for a general domain on which we wish to define the concept of integral, certain restrictions apply, especially in relation to the boundary of the domain. A very general description of these restrictions is indeed a difficult task, requiring the tools from differential geometry. Our presentation here will be simple, requiring the domains we consider have some specific geometric properties.

1.15 Projection Orthogonale de C sur Pi

Date: 20200725

1) C:/DATA-NovoVivo-122-2020/Books-122-2020/Math-Important/Bernard

Gostiaux - Cours de mathematiques speciales - tome 4 - Geometrie affine et metrique - 1995.pdf IMPORTANT

Pag. 163 (185) - 6. Soit P le plan du cercle C, le plan 7r coupe P suivant une droite A qui coupe le cercle en Cet D. Soient A et B intersections du cercle C et du diametre perpendiculaire a [DG].

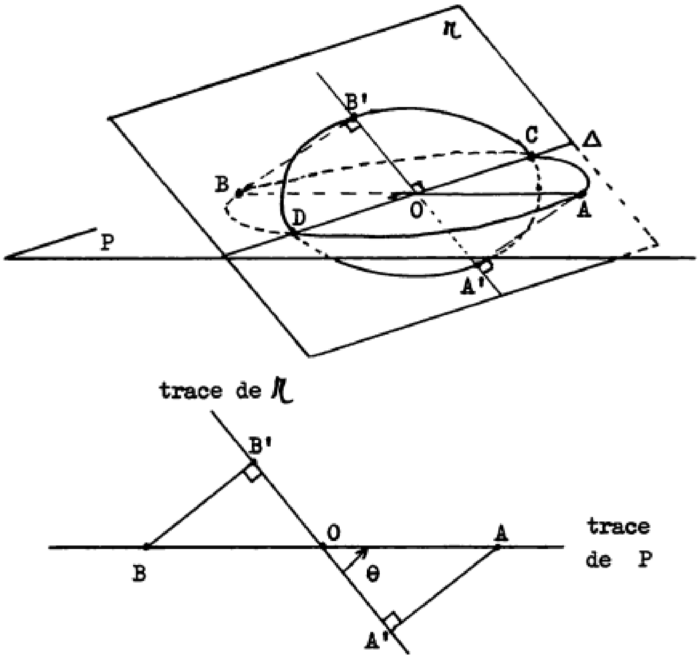

Figure 1.15: po.png

L'ellipse e, projection orthogonale de C sur Pi a pour grand axe CD et petit axe A' B', projection orthogonale de AB sur Pi. Dans le plan Q, orthogonala A en 0, les points A' et B' decrivant les deux cercles de diametres OA et OB, tangents en 0, ceci lorsque pour A fixe, l'angle Theta du diedre (P, Pi) varie. Lorsque A pivote autour de 0, ces cercles engendrent donc un tore de collier nul.

Si on veut y mettre de l'analytique, les deux cercles de diametres R, tangents en 0 a Oz, dans un plan rOz, ont pour equation:

2) C:/DATA-NovoVivo-122-2020/Books-122-2020/Math-Important/Bernard Gostiaux - Cours de mathematiques speciales - tome 4 - Geometrie affine et metrique - 1995.pdf IMPORTANT

Pag. 179(191) - CHAPITRE 5 Geometrie plane euclidienne Pourquoi ce chapitre, ce favoritisme pour la dimension deux? Peut-être parce que l'aspect

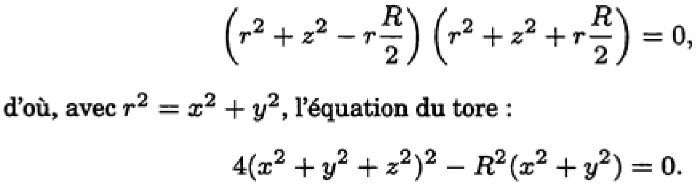

$$\left(r^2+z^2-r\frac{R}{2}\right)\left(r^2+z^2+r\frac{R}{2}\right)=0,$$

d'où, avec $r^2=x^2+y^2$, l'équation du tore :

$$4(x^2+y^2+z^2)^2-R^2(x^2+y^2)=0.$$

Figure 1.16: po2.png

livresque de notre culture a accentué l'importance du support plan. Pensez a la mise en place de la perspective en peinture, qui permet au cerveau de concevoir l'espace a partir d'une figuration plane. C'est aussi pour justifier des résultats utilisés de ci, de là, les centres d'homothétie de deux ou trois cercles par exemple, et enfin pour parler de la géométrie du triangle, si riche qu'elle peut faire l'objet de livres entiers. La difficulté de ce chapitre sera de savoir jusqu'où ne pas aller trop loin! Son cadre est celui des plans affines euclidiens.

Pag. 188 (200) - THEOREME 5.19. -Les symétriques de l'orthocentre h d'un triangle abc, par rapport aux côtés sont sur le cercle circonscrit au triangle. Vous souvenez-vous de ce qu'est l'orthocentre d'un triangle? C'est le point d'intersection des trois hauteurs, et nous justifierons en 5.20 qu'elles se coupent. Pour l'instant on l'admet. Soit ici h l'orthocentre du triangle abc inscrit dans le cercle C, (voir figure 5.19), et h' symétrique de h par rapport au côté be : on a (hb, hc) = -(h'b, h'c) vu l'action d'une symétrie sur les angles. Mais (hb, hc) = (ac, ab), angles a côtés perpendiculaires: on justifie en décomposant en

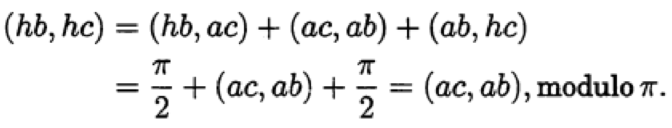

$$\begin{aligned}(hb,hc)&=(hb,ac)+(ac,ab)+(ab,hc)\\&=\frac{\pi}{2}+(ac,ab)+\frac{\pi}{2}=(ac,ab),\textbf{modulo}\,\pi.\end{aligned}$$

Figure 1.17: po3.png

Donc (h'b, h'c) = -(ac, ab) = (ab, ac) : les points h', a, b, et c sont cocycliques d'après 5.18.

Pag. 189 (201) - THEOREME 5.20. - Soit un triangle abc du plan affine euclidien. Les trois hauteurs issues de a, b et c sont concourantes en un point appelé orthocentre du triangle. Soit g le centre de gravité du

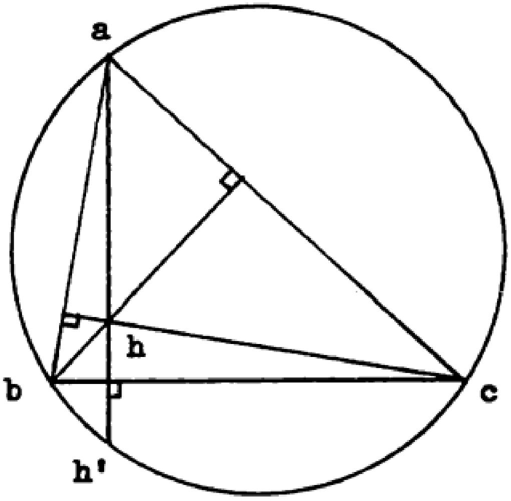

Figure 1.18: po4.png

triangle, isobarycentre des sommets. Si m est le milieu de [a, b], g est aussi le barycentre de (m, 2) et (a, 1) donc on a

$$2\overrightarrow{gm} + \overrightarrow{ga} = 0, \text{ ou } \overrightarrow{ga} = -2\overrightarrow{gm}$$

Figure 1.19: po5.png

Considérons l'homothétie de centre g, de rapport -2, notée phi. On a phi(m) = a, donc l'image par phi de la droite a (b, c) est la parallèle a a (b, c) passant par a. On obtient de même les images des droites a (a, b) et a (a, c), d'où leurs intersections, images de a, b et c, notées a', b', c', (voir figure 5.20).

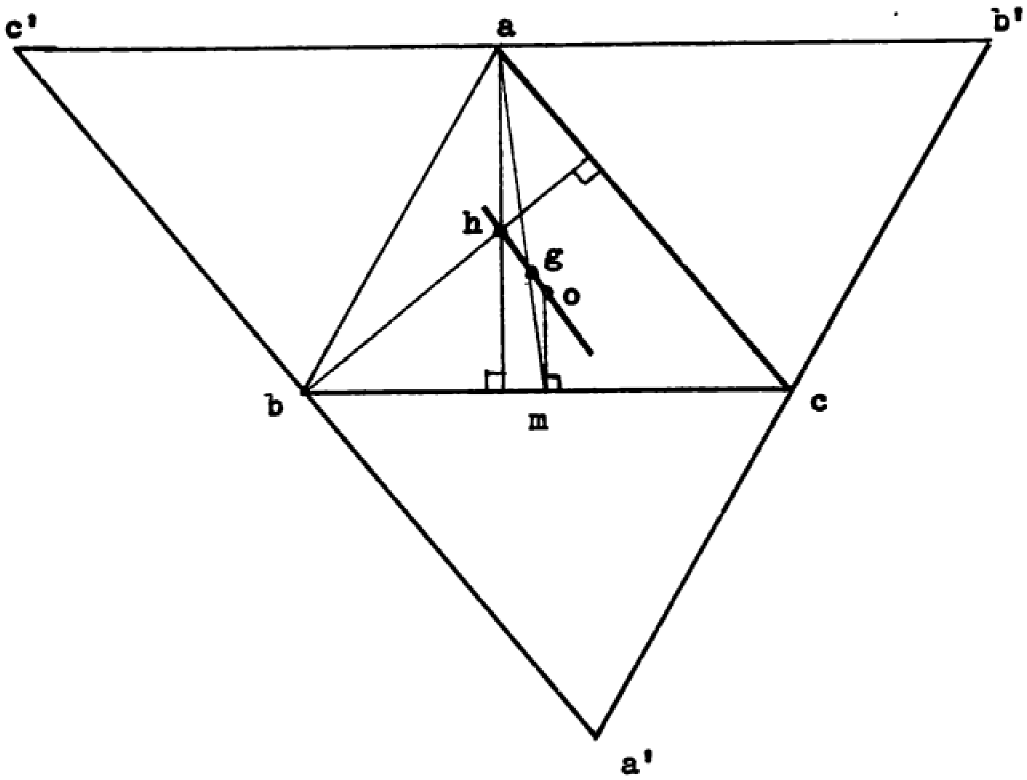

Figure 1.20: po6.png

Dans l'homothétie phi, la perpendiculaire en a à b'c', (donc la hauteur issue de a pour le triangle abc), est l'image par phi de la perpendiculaire en m à be, c'est-à-dire de la médiatrice de [b, c]. Les images par cp des trois médiatrices du triangle abc sont donc ses trois hauteurs. Or les médiatrices se coupent en 0, centre du cercle circonscrit au triangle, (si 0 est l'intersection des médiatrices de [a, b] et [b, c] on a Oa =Ob et Ob= Oc d'où Oa =Oc; donc 0 est sur la troisième médiatrice). Il en résulte que phi(0) est sur les trois hauteurs, d'où l'existence de h, mais aussi la relation

$$\overrightarrow{gh} = -2\overrightarrow{go}$$

Figure 1.21: po7.png

On a donc: COROLLAIRE 5.21. - Dans ce triangle abc , l'orthocentre h, le centre de gravité g et le centre du cercle circonscrit 0 sont alignés et on

a

$$\overrightarrow{gh} = -2\overrightarrow{go}$$

Figure 1.22: po7.png

(autrement dit g est entre h et 0, aux deux tiers à partir de h.

THEOREME 5.22. - Une condition nécessaire et suffisante pour qu'un point m soit sur le cercle C circonscrit à un triangle abc est que les projetés a', b' etc' de m sur be, ca et ab soient alignés. Si c'est le cas, la droite D = a (a', b', c') est appelée droite de Simson de triangle abc pour le point m.

Si on projette un point m du plan en a' sur be, b' sur ca et c' sur ab, comme (c' a, c' m) = Pi/2 = (b' a, b' m), les points a, m, c', b' sont cocycliques, donc on a également (b'a, b'c') =(ma, mc'), (voir figure 5.22).

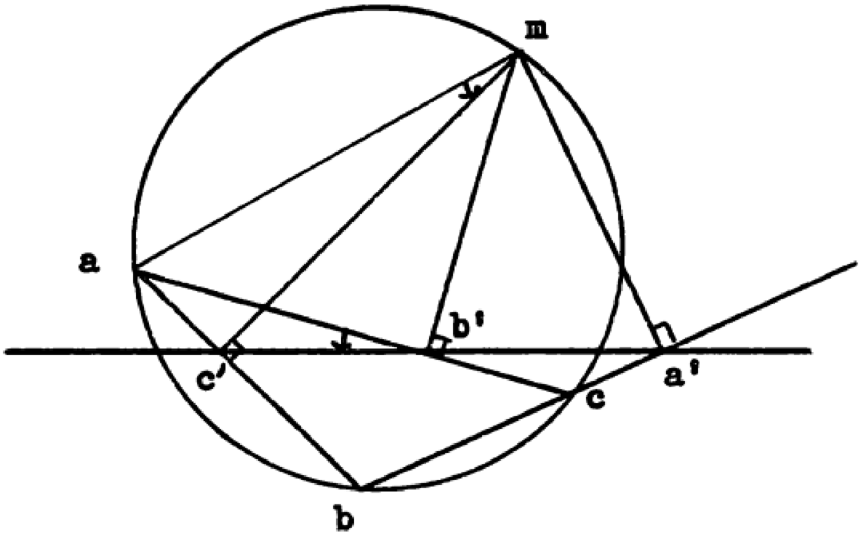

Figure 1.23: po8.png

On a de même: (b'c,b'a') =(me, ma').

or (ma, mc')= (ma, ma')+ (ma',mc') et (mc, ma')= (mc, ma)+ (ma, ma').

Comme enfin les points m, b, c', a' sont cocycliques, (toujours (c'b, dm) = Pi/2 = (a'm, a'b)), on a l'égalité des angles de droites (ma', mc') = (ba', bc') = (bc, ba) car il s'agit des mêmes droites. On en déduit donc que :

5.23

(b'a,b'c') =(ma, ma')+ (ma',md) =(ma, ma')+ (bc,ba) et (b'c,b'a') =(mc, ma)+ (ma, ma').

Si on suppose m sur le cercle C, (me, ma)= (be, ba), d'où: (b' a, b' c') = (b' c, b' a'). Mais alors, par Chasles, on a : (b'c',b'a') = (b'c',b'a) + (b'a,b'c) + (b'c,b'a') = (b'c',b'a) +0+ (b'a,b'c') = (b'c',b'c') = 0 car a, c et b' sont alignés. On a donc a', b' et c' alignés. Si réciproquement a', b' et c' sont alignés, b' c' et b' a' définissent la même droite, ainsi que b' a et b' c, (b' projeté orthogonal de b sur ac), donc (b' a, b' d) = (b' c, b' a') et vu les égalités 5.23, on en déduit que (mc, ma)= (bc, ba), donc que m est sur le cercle circonscrit au triangle a, b, c.

1.16 Euler Formula

Date: 20200726

1) C:/DATA-NovoVivo-122-2020/Books-122-2020/Math-Important-2/311040947X.pdf

Assunto: Mathematics for the Physical Sciences

Pag. 08 (15)

2) C:/DATA-NovoVivo-122-2020/Books-122-2020/Math-Important-2/0817646701.pdf

Assunto: Integral Methods in Science and Engineering Techniques and Applications

Pag. 01(15) - 1 Superconvergence of Projection Methods for Weakly Singular Integral Operators

3) C:/DATA-NovoVivo-122-2020/Books-122-2020/Math-Important-2/1461478278.pdf

Assunto: Integral Methods in Science and Engineering Progress in Numerical and Analytic Techniques

4) C:/DATA-NovoVivo-122-2020/Books-122-2020/Math-Important-2/3030160769.pdf

Assunto: Integral Methods in Science and Engineering Analytic Treatment and Numerical Approximations

5) C:/DATA-NovoVivo-122-2020/Books-122-2020/Math-Important-2/Gravitational-Waves.pdf

Assunto: Gravitational Waves A History of Discovery

Pag. 38(55)- Figure 3.2 A reconstruction of Michelson's early interferometer in the historic location in Potsdam, Germany. (Image credit: Dierck E.

Liebscher.)

6) C:/DATA-NovoVivo-122-2020/Books-122-2020/Math-Important-3/Geometrical and Visual Optics, 3rd Edition.pdf
Pag. 183(196) -Lens Thickness

7) C:/BOOKS-2020-25/Books-121-2020/Important/newlecturesGR-2.pdf
Assunto: Lecture Notes on General Relativity Matthias Blau Albert Einstein Center for Fundamental Physics Institut fur Theoretische Physik Universitat Bern CH-3012 Bern, Switzerland
Pag. 380(381) - 20 Einstein Equations from an Action Principle 20.1 Einstein-Hilbert Action

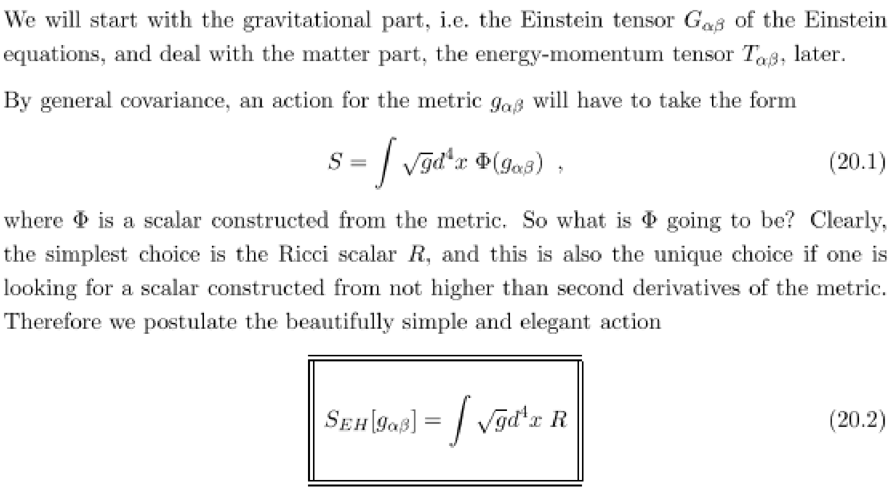

We will start with the gravitational part, i.e. the Einstein tensor $G_{\alpha\beta}$ of the Einstein equations, and deal with the matter part, the energy-momentum tensor $T_{\alpha\beta}$, later.

By general covariance, an action for the metric $g_{\alpha\beta}$ will have to take the form

$$S = \int \sqrt{g} d^4x \; \Phi(g_{\alpha\beta}) \ , \qquad (20.1)$$

where Φ is a scalar constructed from the metric. So what is Φ going to be? Clearly, the simplest choice is the Ricci scalar R, and this is also the unique choice if one is looking for a scalar constructed from not higher than second derivatives of the metric. Therefore we postulate the beautifully simple and elegant action

$$\boxed{S_{EH}[g_{\alpha\beta}] = \int \sqrt{g} d^4x \; R} \qquad (20.2)$$

Figure 1.24: eb.png

8) C:/DATA-NovoVivo-122-2020/Books-122-2020/Math-Import/ Geometria diferencial de superfcies.pdf IMPORTANT
Pag. 06(12) - Figura 1.2. Ve uma vizinhanca qualquer de p, V interseccao S e uma vizinhanca coordenada p.

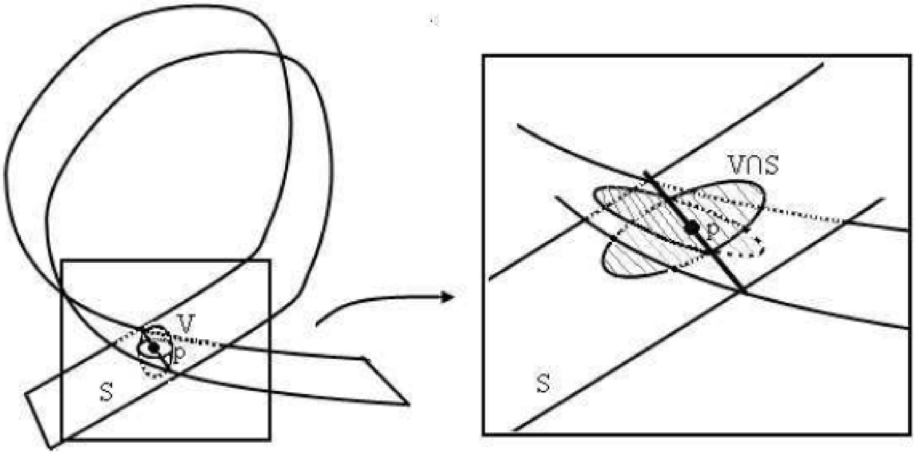

Figure 1.25: gds.png

9) C:/DATA-NovoVivo-122-2020/Books-122-2020/Math-Import/A geometria do espaço-tempo de Schwarzschild.pdf

10) C:/DATA-NovoVivo-122-2020/Books-122-2020/Math-Import/ O problema de Apolônio.pdf Obs.: Livro todo!!!

1.17 Cosmological Constant

Date: 20200727

1) C:/DATA-NovoVivo-122-2020/Books-122-2020/Introduction to Modified Gravity/3030528618.pdf

Assunto: Introduction to Modified Gravity

Pag.: 16 (24) - As an example, we consider the Gödel-type metric (2.26). One can show, that, unlike general relativity, such solutions are possible not only for dust but also for the vacuum (with non-zero cosmological constant), in particular, completely causal vacuum solutions are present [26]. Clearly, the solutions of this form are possible also for the presence of the matter given by the relativistic fluid and a scalar field. Again, as in [22], all Einstein equations will take the form of purely algebraic relations between density, pressure, field amplitude and constants from the gravity Lagrangian. As for the cosmological metric, the possibility of accelerating solutions can be shown just in the same manner as in the previous sections. Among other possible solutions in $f(R, Q)$ gravity, it is worth to mention Reissner{Nordström black holes [27] and wormholes [28].

Obs.: Emails para Abraham Ungar and Matthias Blau

2) C:/DATA-NovoVivo-122-2020/Books-122-2020/The Evolution of the Euclidean Elements/The Evolution of the Euclidean Elements - A Study of the Theory of Incommensurable Magnitudes and its Significance for Early Greek Geometry 1974.pdf

3) C:/DATA-NovoVivo-122-2020/Books-122-2020/Textual Studies in Ancient and Medieval Geometry 1989/Textual Studies in Ancient and Medieval Geometry 1989.pdf

4) C:/DATA-NovoVivo-122-2020/Books-122-2020/Advanced Calculus Problem Solver (Problem Solvers Solution Guides)/Advanced Calculus Problem Solver (Problem Solvers Solution Guides).pdf

5) C:/DATA-NovoVivo-122-2020/Books-122-2020/Riddles in Mathematics A Book of Paradoxes/Riddles-in-Mathematics|Eugene-P-Northrop.epub

Pag. 36 - Division of the circle - Fig. 17

6) C:/DATA-NovoVivo-122-2020/Books-122-2020/Algebraic Geometry I/ Shafarevich - Algebraic geometry I - Algebraic curves, manifolds, and schemes 1994.djvu IMPORTANT

7) C:/DATA-NovoVivo-122-2020/Books-122-2020/Algebraic Geometry I/Shafarevich - Algebraic Geometry II - Cohomology of Algebraic Varieties, Algebraic Surfaces.djvu IMPORTANT Obs.: Volume todo!!! - 1383 pages.

Pag. 205(207) - Proposition. Let f be a nonconstant polynomial, and suppose the differentials dxf are zero at all points x pertence f up 1. Then f is the p-th power of some polynomial, where p is the characteristic of the field K. In particular, f is reducible.

Figure 7

7.2. Tangent Space. IMPORTANT

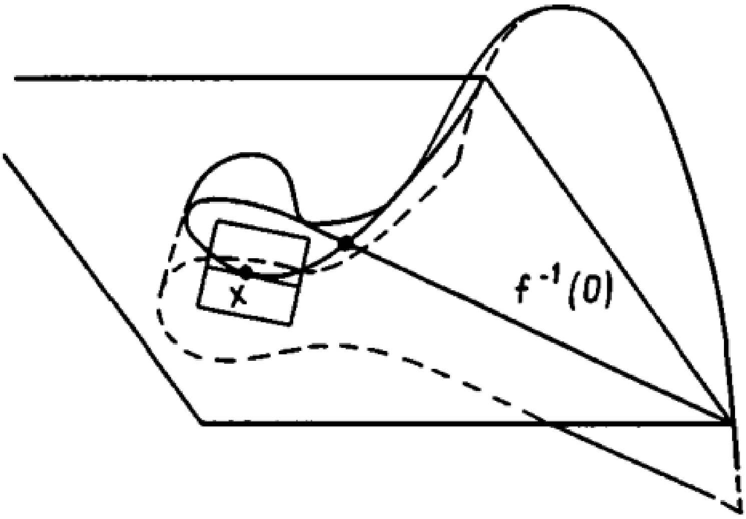

Figure 1.26: av1.png

Pag. 236 (238) -Theorem. Every variety is factorial in a neighbourhood of a smooth point.

Pag. 237 (239) - Fig. 14

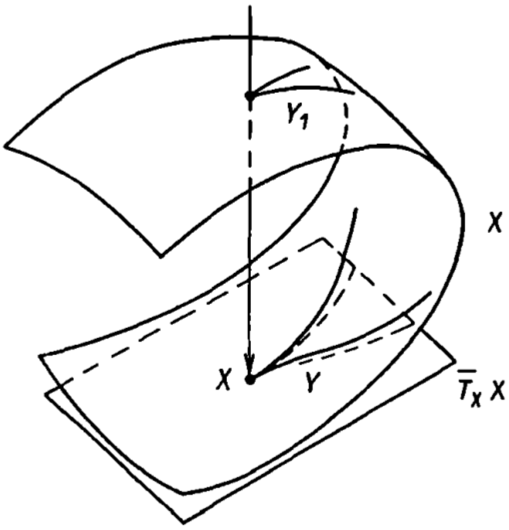
Y_1
X
x
Y
$\overline{T}_x X$

Figure 1.27: av2.png

Pag. 264 (266) - Corollary (Pascal's theorem). The pairs of opposite sides of a hexagon inscribed in a conic meet in three collinear points.

Fig. 19 IMPORTANT

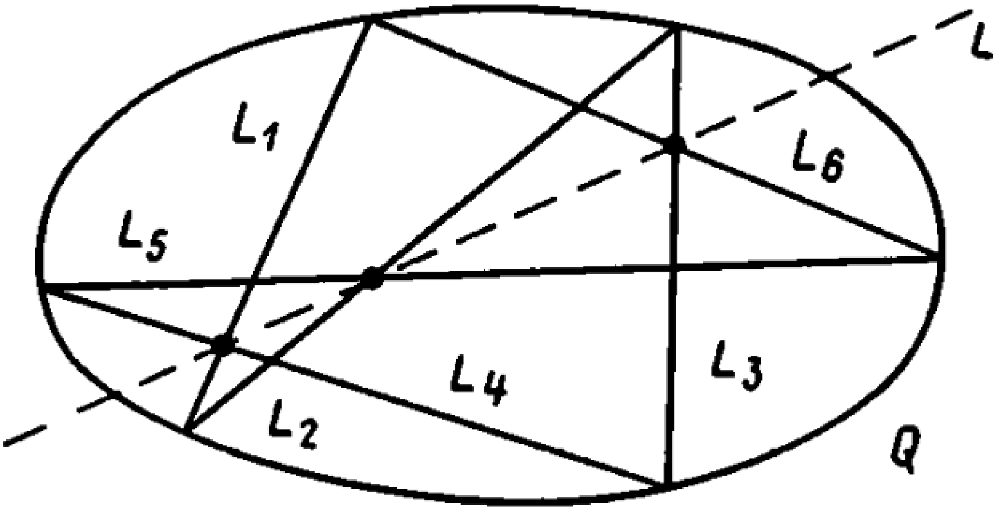

Figure 1.28: av3.png

8) C:/DATA-NovoVivo-123-2020/Books-123-2020/An Invitation to Unbounded Representations of Plus Minus Algebras on Hilbert Space/3030463656.pdf

5 Positive Linear Functionals . . . . 93 5.1 Ordering of Positive Functionals . . . . 93 5.2 Orthogonal Positive Functionals . . . . 97 5.3 The Transition Probability of Positive Functionals . . . 99 5.4 Examples of Transition Probabilities . . . . 104 5.5 A Radon { Nikodym Theorem for Positive Functionals . .. 110 5.6 Extremal Decomposition of Positive Functionals . . . . 113 5.7 Quadratic Modules and r -Representations. . . . . 117 5.8 Exercises . . . . 119 5.9 Notes 119

9 Integrable Representations of Enveloping Algebras . . . . 187 9.1 Preliminaries on Lie Groups and Enveloping Algebras . . . . . . 188 9.2 Infinitesimal Representations of Unitary Representations . . . . . 189 9.3 The Graph Topology of the Infinitesimal Representation . . . . . 196 9.4 Elliptic Elements . . . . 199 9.4.1 Preliminaries on Elliptic Operators . . . .

. . 199 9.4.2 Main Results on Elliptic Elements . . . .
204 9.4.3 Applications of Elliptic Elements 205
9.5 Two Examples . . . .
. . . . . 210 9.6 Analytic Vectors . . . .
. . . . 212 9.6.1 Analytic Vectors for Single Operators . . . . .
. . . . 213 9.6.2 Analytic Vectors for Unitary Representations . . . .
. . . . 215 9.6.3 Exponentiation of Representations of Enveloping Algebras .
. 216 9.7 Analytic Vectors and Unitary Representations of SL (2, R) . . . .
217 9.8 Exercises . . . .
. . . 221 9.9 Notes . . . .
. . . . 222

Chapter 2

Pappus Theorem

Date: 20200728

1) Pappus' Theorem Today we will learn about Pappus' theorem. This theorem states that if we take three points 1, 3, 5 on a line, and another three points 2, 4, 6 on another line, then the three intersection points of the following line pairs 12,45, 23,56, 34,61

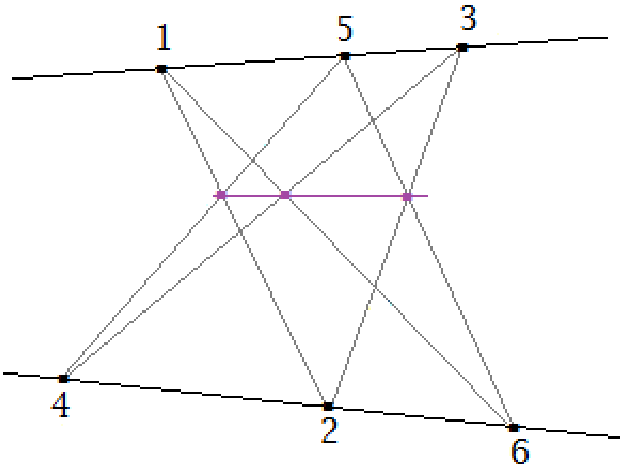

Figure 2.1: pt1.png

Pappus' theorem looks very similar to Pascal's hexagon theorem. In Pascal's theorem, we have a hexagon 123456 inscribed in a circle and the intersection points of the three pairs of opposite sides 12,45, 23,56, 34,61 of the hexagon lie on a straight line.

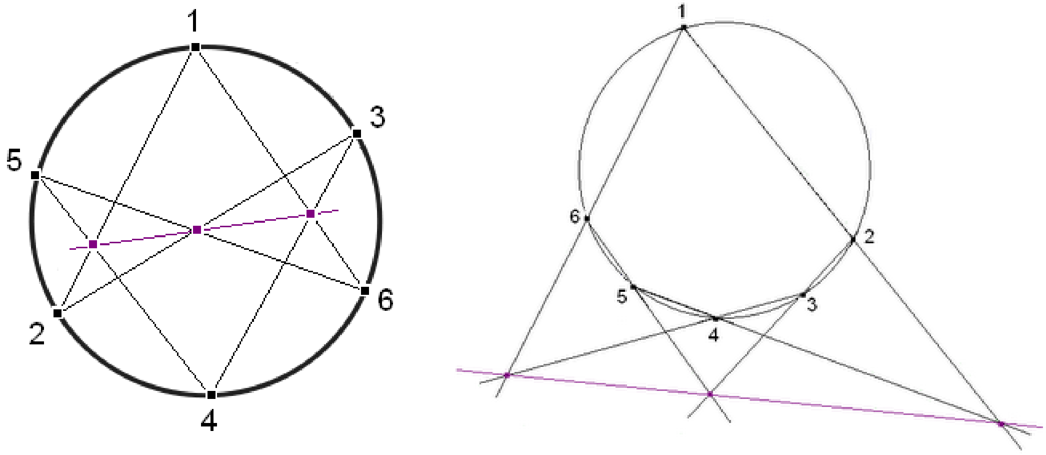

Figure 2.2: pt2.png

So instead of being inscribed in a circle as in Pascal's theorem, in Pappus' theorem, we may say that the hexagon 123456 "inscribed" in two lines and again, the intersection points of the three pairs of opposite sides 12,45, 23,56, 34,61 of the hexagon also lie on a straight line.

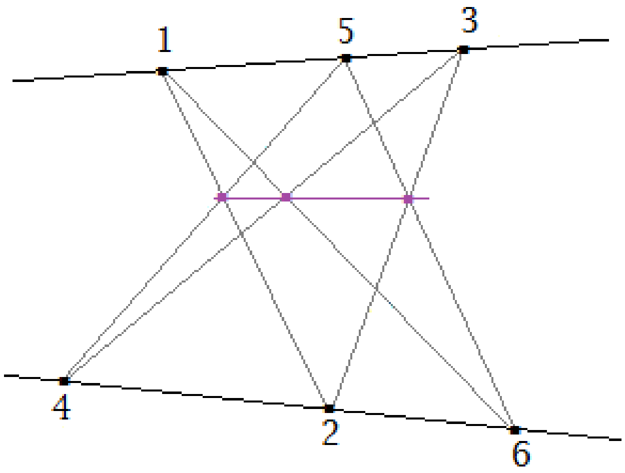

Figure 2.3: pt3.png

Different configurations of Pappus' theorem

Similar to Pascal's theorem, Pappus' theorem also has many configurations. The six vertices of the hexagon can be arranged on the two lines in any particular order. With each arrangement of the vertices, we have an instance of Pappus' theorem. Below is an example:

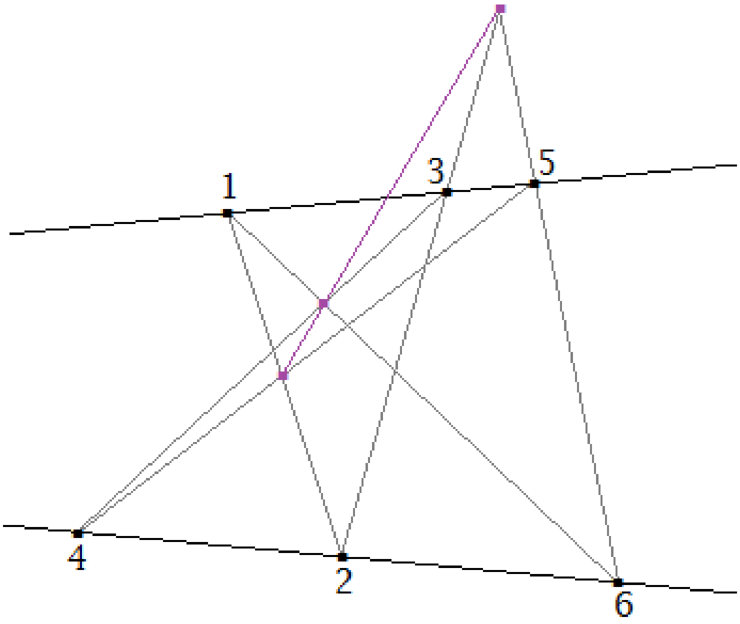

Figure 2.4: pt4.png

Now, why don't you try to draw for yourself your own picture of Pappus' theorem?! This theorem is so cool, isn't it?!

The picture below illustrates another interesting aspect of Pappus' theorem, that is the roles of the hexagon's vertices and the intersection points can be interchanged. As we can see, in the middle figure, the intersection points are interchanged with the vertices 1, 3, 5; and in the right-hand-side figure, the intersection points are swapped with 2, 4, 6.

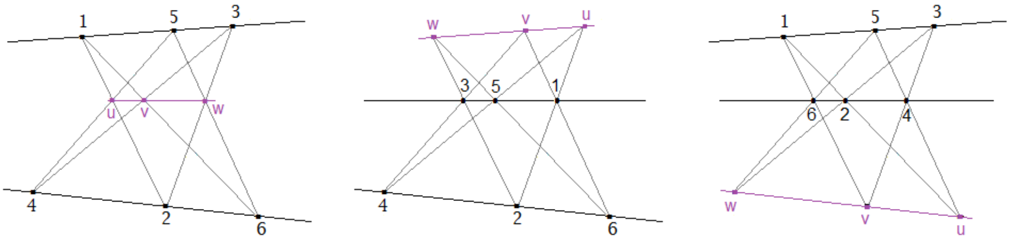

Figure 2.5: pt5.png

Note that, in all three figures above, u is always the intersection point of the pair 12,45, v is always the intersection point of the pair 34,61 and w is always the intersection point of 23,56.

Proving Pappus' theorem

We will use Menelaus' theorem to prove Pappus' theorem. Menelaus' theorem is a useful tool to prove the collinearity of points. The theorem is stated as follows:

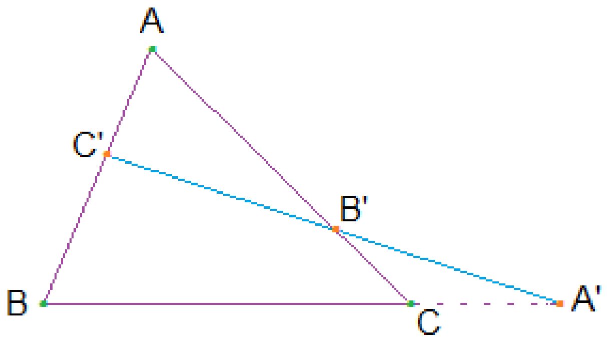

Figure 2.6: pt6.png

Menelaus' theorem: Given a triangle ABC and three points A', B', C' lying on the three lines BC, CA, AB, respectively. Then the three points A', B', C' are collinear if and only if

$$\frac{A'B}{A'C}\frac{B'C}{B'A}\frac{C'A}{C'B} = 1 : \qquad (2.1)$$

The proof that we are about to present is very similar to the proof of Pascal' theorem that we show in the previous post. Before going into detail, let us review that proof of Pascal' theorem. Let us consider the following figure.

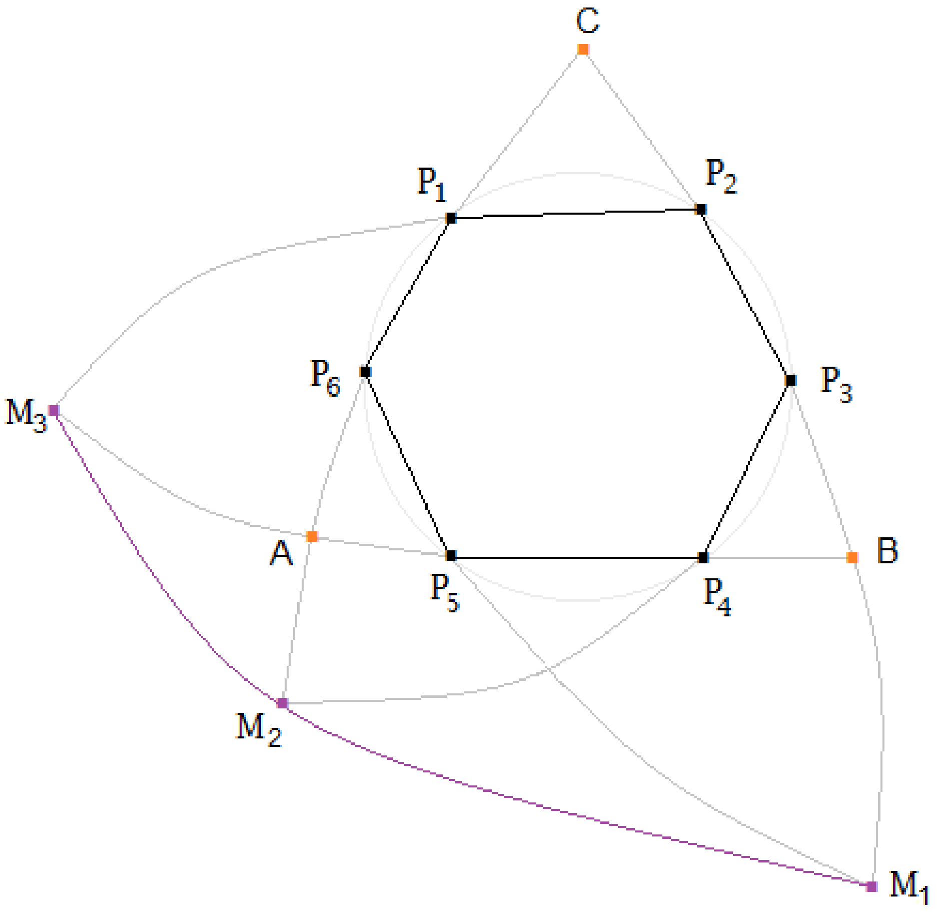

Figure 2.7: pt7.png

On this figure, we see a Pascal's hexagon P1P2P3P4P5P6 and the three intersection points M1, M2, M3. To prove that M1, M2, M3 lies on a straight line, we apply Menelaus' theorem for the triangle ABC. Since M1 belongs to the line BC, M2 belongs to the line CA and M3 belongs to the line AB, to prove that they are collinear, we show that

$$\frac{\widetilde{M_1B}}{\widetilde{M_1C}} \frac{\widetilde{M_2C}}{\widetilde{M_2A}} \frac{\widetilde{M_3A}}{\widetilde{M_3B}} = 1: \qquad (2.2)$$

Now we do the same to prove Pappus' theorem. We will use Menelaus' theorem for the triangle ABC. The triangle ABC is determined exactly in the same way as in the proof of Pascal's theorem: A is the intersection of P4P5 and P6P1, B is the intersection of P2P3 and P4P5, and C is the intersection of P2P3 and P6P1.

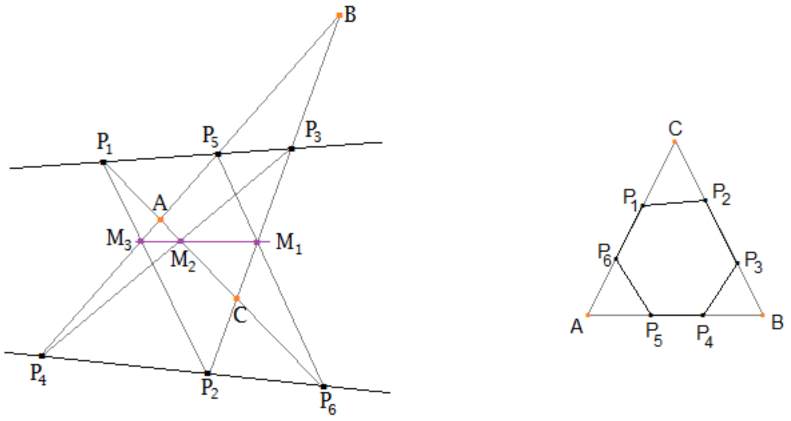
B
P_1
P_5
P_3
A
M_3
M_2
M_1
C
P_4
P_2
P_6
C
P_1
P_2
P_6
P_3
A
P_5
P_4
B

Figure 2.8: pt8.png

Let us write down the proof.

Applying Menelaus' theorem for the triangle ABC with the following triples of collinear points M1,P6,P5, M2,P4,P3, M3,P2,P1, P1,P3,P5, P2,P4,P6,

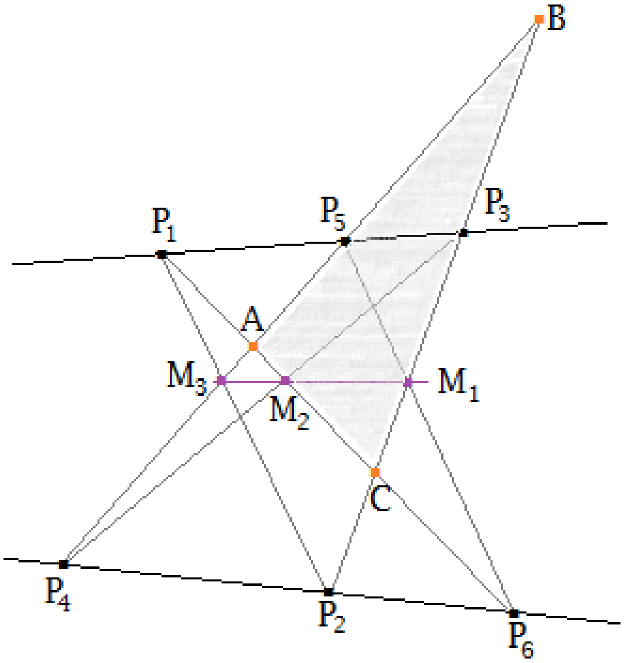
B
P_1
P_5
P_3
A
M_3
M_2
M_1
C
P_4
P_2
P_6

Figure 2.9: pt9.png

we have

$$\frac{\widetilde{M_1B}}{\widetilde{M_1C}}\ \frac{\widetilde{P_6C}}{\widetilde{P_6A}}\ \frac{\widetilde{P_5A}}{\widetilde{P_5B}} = \frac{\widetilde{M_2C}}{\widetilde{M_2A}}\ \frac{\widetilde{P_4A}}{\widetilde{P_4B}}\ \frac{\widetilde{P_3B}}{\widetilde{P_3C}} = \frac{\widetilde{M_3A}}{\widetilde{M_3B}}\ \frac{\widetilde{P_2B}}{\widetilde{P_2C}}\ \frac{\widetilde{P_1C}}{\widetilde{P_1A}} \tag{2.3}$$

$$= \frac{\widetilde{P_1A}}{\widetilde{P_1C}}\ \frac{\widetilde{P_3C}}{\widetilde{P_3B}}\ \frac{\widetilde{P_5B}}{\widetilde{P_5A}} = \frac{\widetilde{P_2C}}{\widetilde{P_2B}}\ \frac{\widetilde{P_4B}}{\widetilde{P_4A}}\ \frac{\widetilde{P_6A}}{\widetilde{P_6C}} = 1: \tag{2.4}$$

Multiply all the above identities, we derive

$$\frac{\widetilde{M_1B}}{\widetilde{M_1C}}\ \frac{\widetilde{M_2C}}{\widetilde{M_2A}}\ \frac{\widetilde{M_3A}}{\widetilde{M_3B}} = 1: \tag{2.5}$$

Thus, M1, M2, M3 are collinear, and the theorem is proved.

So, once again today, by applying Menelaus' theorem in an effective way, we have proved Pappus' theorem.

Homework.

1. Let us call the purple points as Pappus' points, and the purple line connecting them as Pappus' line.

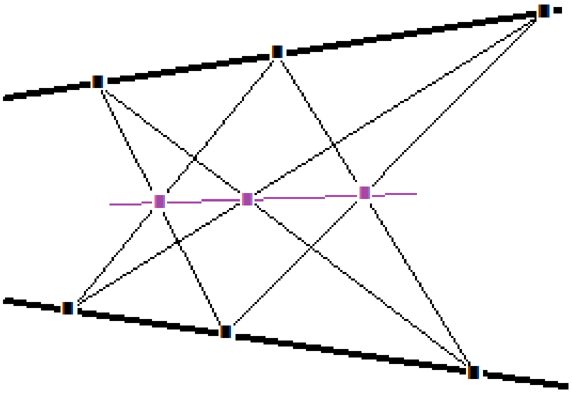

Figure 2.10: pt10.png

Now, let the six points stay fixed on the two lines. By different labeling 1,2,3,4,5,6 to the six points, the questions are

how many Pappus' lines can be generated? how many Pappus' points can be generated? each Pappus' point lies on how many Pappus' lines? each Pappus' line contains how many Pappus' points?

2. Instead of applying Menelaus' theorem for the triangle ABC as in the above proof, choose X as the intersection of P1P2 and P3P4, Y the intersection of P1P2 and P5P6, and Z the intersection of P3P4 and P5P6, prove Pappus' theorem by using Menelaus' theorem for the triangle XYZ.

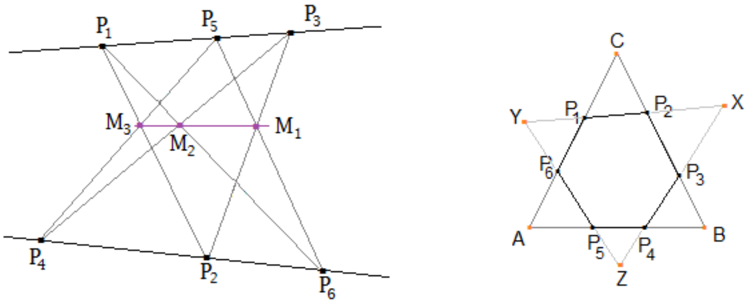
P_1
P_5
P_3
M_3
M_2
M_1
P_4
P_2
P_6
C
Y
P_1
P_2
X
P_6
P_3
A
P_5
P_4
B
Z

Figure 2.11: pt11.png

3. Prove that if P2P3 is parallel to P5P6, and P3P4 is parallel to P1P6, then P1P2 is parallel to P4P5.

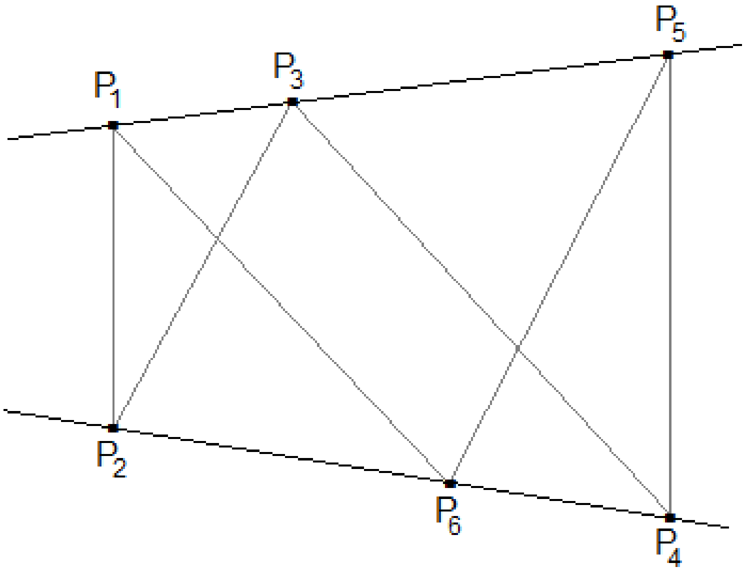
P_1
P_3
P_5
P_2
P_6
P_4

Figure 2.12: pt12.png

Source: http://mathgardenblog.blogspot.com/2013/08/ hexagrammum-mysticum3.html

2.1 Ceva's Theorem and Menelaus' Theorem

Today we will learn about two well-known theorems in geometry, Ceva's Theorem and Menelaus' Theorem. These two theorems are very useful in plane geometry because we often use them to prove that a certain number of points lie on a straight line and a certain number of lines intersect at a single point. Both of the theorems will be proved based on a common simple principle. We also generalize the theorems for arbitrary polygons.

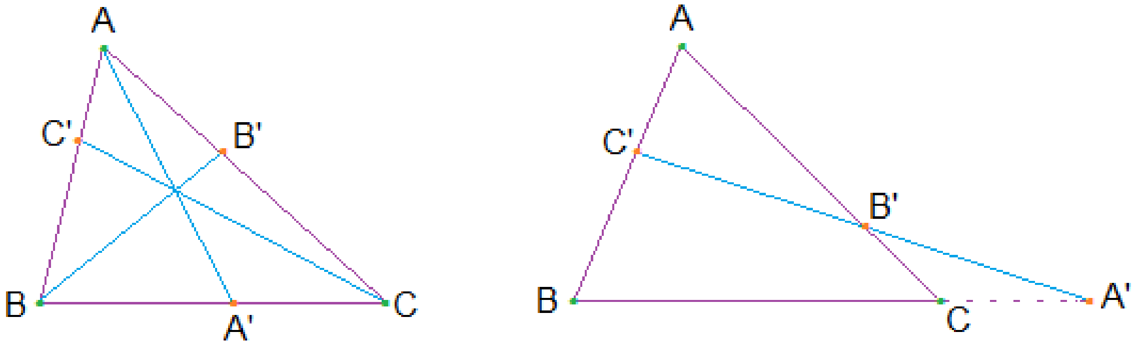

Ceva

Menelaus

Figure 2.13: ct1.png

Let us state the theorems.

Ceva's Theorem: Given a triangle ABC and three points A', B', C' lying on the three lines BC, CA, AB, respectively. Then the three lines AA', BB', CC' are concurrent if and only if

$$\frac{A'B}{A'C}\quad\frac{B'C}{B'A}\quad\frac{C'A}{C'B} = 1: \tag{2.6}$$

Menelaus' Theorem: Given a triangle ABC and three points A', B', C' lying on the three lines BC, CA, AB, respectively. Then the three points A', B', C' are collinear if and only if

$$\frac{\overrightarrow{A'B}}{\overrightarrow{A'C}} \quad \frac{\overrightarrow{B'C}}{\overrightarrow{B'A}} \quad \frac{\overrightarrow{C'A}}{\overrightarrow{C'B}} = 1 : \qquad (2.7)$$

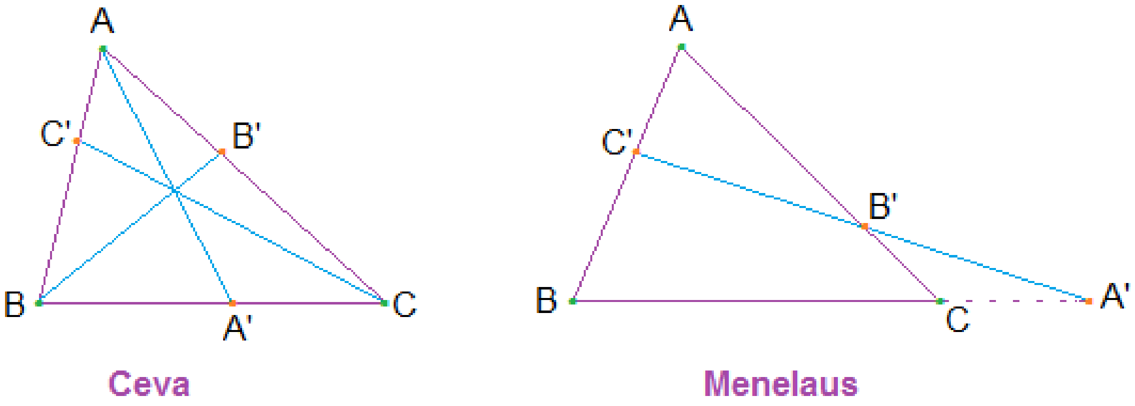

Figure 2.14: ct2.png

Signed ratio of line segments

First, let us explain a little bit about the notation that we use in the theorems. The notation $\frac{\overrightarrow{A'B}}{\overrightarrow{A'C}}$ is called a signed ratio. In the figure below, we can see that the normal (unsigned) ratio $\frac{UX}{UY} = 2$, but the signed ratio is $\frac{\overrightarrow{UX}}{\overrightarrow{UY}} = \ \ 2$. This is because $\overrightarrow{UX}$ and $\overrightarrow{UY}$ point to the same direction, and it is negative when $\overrightarrow{UX}$ and $\overrightarrow{UY}$ points to opposite directions.

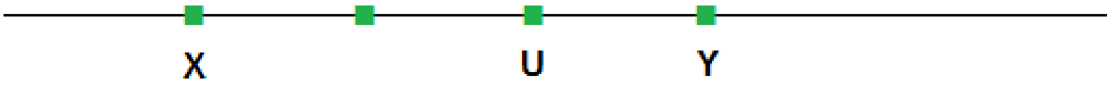

Figure 2.15: ct3.png

The normal ratio is always a positive number but the signed ratio can be negative or positive. The signed ratio $\frac{\overrightarrow{UX}}{\overrightarrow{UY}}$ is positive when $\overrightarrow{UX}$ and $\overrightarrow{UY}$ point to the same direction, and it is negative when $\overrightarrow{UX}$ and $\overrightarrow{UY}$ points to opposite directions.

So why do we need the signed ratio?

The trouble with our normal ratio is that it cannot uniquely identify a point on a line in relation to two other reference points, but the signed ratio can.

Let us consider an example. Suppose on the line XY, we want to specify a point Z so that $\frac{ZX}{ZY} = 2$.

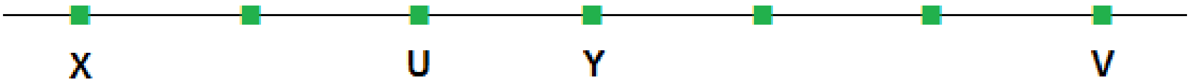

Figure 2.16: ct4.png

In the above figure, we can see that using our usual unsigned ratio, there are two points, U and V, satisfying this condition since

$$\frac{UX}{UY} = \frac{VX}{VY} = 2. \tag{2.8}$$

But if we consider signed ratio, then there is a unique point, which is V satisfying $\frac{\overline{VX}}{\overline{VY}} = 2$.

The point U no longer satisfies because $\frac{\overline{UX}}{\overline{UY}} = -2 \neq 2$.

Signed area

Similar to signed ratio, we have the concept of signed area. The normal (unsigned) area of a polygon is always a positive number but a signed area can be positive or negative depending on the orientation of the vertices of the polygon.

If we draw a coordinate system 0xy on the plane then any point A on the plane will have an associated coordinate (Ax,Ay). We define

$$[A; B] = A_x B_y - A_y B_x \tag{2.9}$$

and define the signed area of a triangle ABC to be

$$\overline{s}(ABC) = \frac{1}{2}([A; B] + [B; C] + [C; A]). \tag{2.10}$$

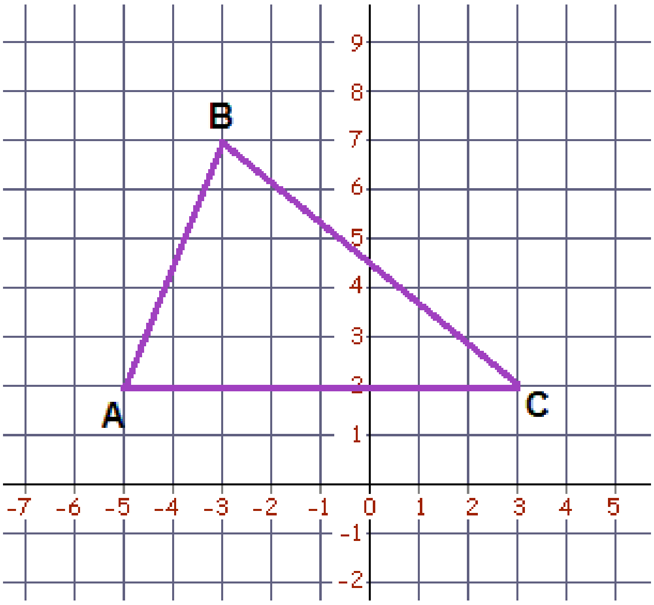

Figure 2.17: ct5.png

In the above figure, we have the coordinates A(-5,2), B(-3,7), C(3,2). So we can calculate

$$[A;B] = A_x B_y - A_y B_x = (-5)(7) - (2)(-3) = -29; \quad (2.11)$$

$$[B;C] = B_x C_y - B_y C_x = (-3)(2) - (7)(3) = -27; \quad (2.12)$$

$$[C;A] = C_x A_y - C_y A_x = (3)(2) - (2)(-5) = 16: \quad (2.13)$$

Thus,

$$\bar{s}(ABC) = \frac{1}{2}([A;B] + [B;C] + [C;A]) = \frac{1}{2}(-29 - 27 + 16) = -20: \quad (2.14)$$

On the other hand, [A,C]=-16, [C,B]=27, [B,A]=29 and $\bar{s}(ACB) = 20$.

A theorem on ratio of signed areas

Let us now state a simple theorem concerning about the ratio of the signed areas of two triangles that share a common edge. We will use this theorem later to prove the Ceva's theorem and the Menelaus' theorem.

A Theorem on Ratio of Signed Areas. Given two triangles ABU and ABV that share a common edge AB. The line connecting the two vertices UV intersects with the line AB at a point T. Then it must hold that

$$\frac{\overline{s}(ABU)}{\overline{s}(ABV)} = \frac{\overline{TU}}{\overline{TV}}: \qquad (2.15)$$

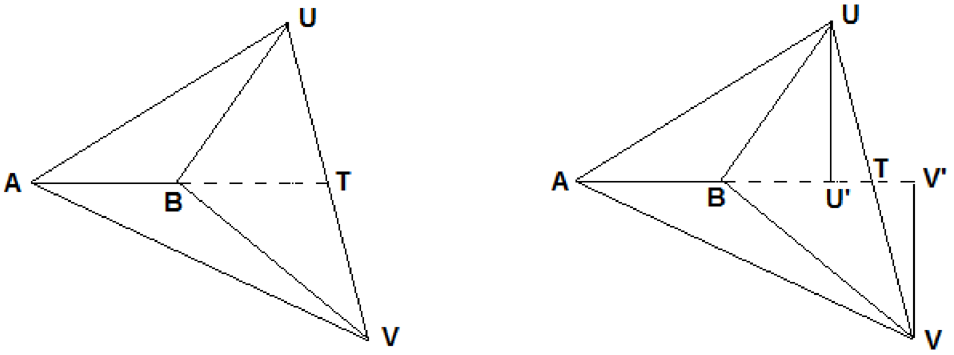

Figure 2.18: ct6.png

This theorem is quite obvious if we forget about the sign and consider the usual unsigned ratio. This is because if we draw the two altitudes UU' and VV' perpendicular to the side AB then

$$\frac{s(ABU)}{s(ABV)} = \frac{AB \quad UU^0=2}{AB \quad VV^0=2} = \frac{UU^0}{VV^0} = \frac{TU}{TV}: \qquad (2.16)$$

We will give a proof of this theorem for the signed ratio in the next post.

Proving Ceva's Theorem and Menelaus' Theorem

Let us now use the above theorem to prove Ceva's Theorem and Menelaus' Theorem.

Proof of Ceva's Theorem. Suppose the three lines AA', BB', CC' are concurrent at a point I. Then by the theorem on the ratio of signed areas, we have

$$\frac{\overline{A^0B}}{\overline{A^0C}} = \frac{\overline{s}(IAB)}{\overline{s}(IAC)}; \quad \frac{\overline{B^0C}}{\overline{B^0A}} = \frac{\overline{s}(IBC)}{\overline{s}(IBA)}; \quad \frac{\overline{C^0A}}{\overline{C^0B}} = \frac{\overline{s}(ICA)}{\overline{s}(ICB)}: \qquad (2.17)$$

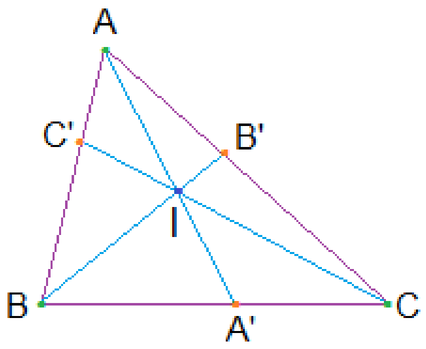

Figure 2.19: ct7.png

Thus,

$$\frac{A'B}{A'C}\frac{B'C}{B'A}\frac{C'A}{C'B} = \frac{s(IAB)}{s(IAC)}\frac{s(IBC)}{s(IBA)}\frac{s(ICA)}{s(ICB)} \tag{2.18}$$

$$= \left(\frac{s(IAB)}{s(ICA)}\right)\left(\frac{s(IBC)}{s(IAB)}\right)\left(\frac{s(ICA)}{s(IBC)}\right) = 1. \tag{2.19}$$

In the reverse direction, if

$$\frac{A'B}{A'C}\frac{B'C}{B'A}\frac{C'A}{C'B} = 1, \tag{2.20}$$

we need to prove that the three lines AA', BB', CC' are concurrent. In this case, let AA' and BB' intersect at a point I, let CI and AB intersect at C", we need to show that C"=C'. Indeed, since AA', BB', CC" are concurrent, by what we have just proved, we have

$$\frac{A'B}{A'C}\frac{B'C}{B'A}\frac{C''A}{C''B} = 1, \tag{2.21}$$

Thus,

$$\frac{C''A}{C''B} = \frac{C'A}{C'B}. \tag{2.22}$$

Since the signed ratio determines a point uniquely on the line AB, we obtain C'=C".

So we have proved the Ceva's Theorem.

Proof of Menelaus' Theorem. Suppose that the three points A', B', C' are collinear. Take two arbitrary points I and J on the line A'B'C', then by the theorem on the ratio of signed areas, we have

$$\frac{A'B}{A'C} = \frac{\overline{s}(IJB)}{\overline{s}(IJC)}; \quad \frac{B'C}{B'A} = \frac{\overline{s}(IJC)}{\overline{s}(IJA)}; \quad \frac{C'A}{C'B} = \frac{\overline{s}(IJA)}{\overline{s}(IJB)}: \qquad (2.23)$$

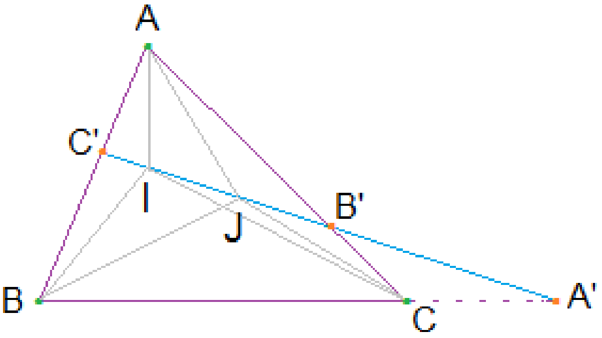

Figure 2.20: ct8.png

Thus,

$$\frac{A'B}{A'C} \frac{B'C}{B'A} \frac{C'A}{C'B} = \frac{\overline{s}(IJB)}{\overline{s}(IJC)} \frac{\overline{s}(IJC)}{\overline{s}(IJA)} \frac{\overline{s}(IJA)}{\overline{s}(IJB)} = 1: \qquad (2.24)$$

The reverse direction of Menelaus' Theorem is proved similar to that of Ceva's Theorem.

So we have finally proved Ceva's Theorem and Menelaus' Theorem. This proof also works if we only consider the usual unsigned ratio. We remark that for Menelaus' theorem, instead of using the ratio of the areas of the triangles, the same result can be obtained if we use the ratio of the altitudes drawing from A, B, C perpendicular to the line A'B'C'.

Generalization of Ceva's Theorem and Menelaus' Theorem

The above proof of Ceva's theorem and Menelaus' theorem motivates us to generalize the theorems for arbitrary polygons.

Here are Ceva's theorem and Menelaus' theorem for pentagon.

Ceva's theorem for pentagon. Given a pentagon A1A2A3A4A5 and five points B1, B2, B3, B4, B5 lying on the lines A5A2, A1A3, A2A4, A3A5, A4A1, respectively. If the lines A1B1, A2B2, A3B3, A4B4, A5B5 are concurrent then

$$\frac{\widetilde{B_1A_5}}{\widetilde{B_1A_2}}\quad\frac{\widetilde{B_2A_1}}{\widetilde{B_2A_3}}\quad\frac{\widetilde{B_3A_2}}{\widetilde{B_3A_4}}\quad\frac{\widetilde{B_4A_3}}{\widetilde{B_4A_5}}\quad\frac{\widetilde{B_5A_4}}{\widetilde{B_5A_1}} = 1. \tag{2.25}$$

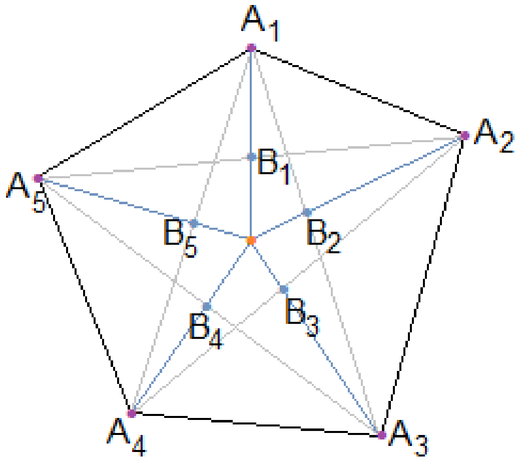

Figure 2.21: ct9.png

Proof. If the five lines A1B1, A2B2, A3B3, A4B4, A5B5 are concurrent at a point I then

$$\frac{\widetilde{B_1A_5}}{\widetilde{B_1A_2}}\quad\frac{\widetilde{B_2A_1}}{\widetilde{B_2A_3}}\quad\frac{\widetilde{B_3A_2}}{\widetilde{B_3A_4}}\quad\frac{\widetilde{B_4A_3}}{\widetilde{B_4A_5}}\quad\frac{\widetilde{B_5A_4}}{\widetilde{B_5A_1}} \tag{2.26}$$

$$= \frac{\overline{s}(IA_1A_5)}{\overline{s}(IA_1A_2)}\quad\frac{\overline{s}(IA_2A_1)}{\overline{s}(IA_2A_3)}\quad\frac{\overline{s}(IA_3A_2)}{\overline{s}(IA_3A_4)}\quad\frac{\overline{s}(IA_4A_3)}{\overline{s}(IA_4A_5)}\quad\frac{\overline{s}(IA_5A_4)}{\overline{s}(IA_5A_1)} \tag{2.27}$$

$$= \left(\frac{\overline{s}(IA_5A_1)}{\overline{s}(IA_1A_2)}\right)\left(\frac{\overline{s}(IA_1A_2)}{\overline{s}(IA_2A_3)}\right)\left(\frac{\overline{s}(IA_2A_3)}{\overline{s}(IA_3A_4)}\right)\left(\frac{\overline{s}(IA_3A_4)}{\overline{s}(IA_4A_5)}\right)\left(\frac{\overline{s}(IA_4A_5)}{\overline{s}(IA_5A_1)}\right) = 1. \tag{2.28}$$

Menelaus' theorem for pentagon. Given a pentagon A1A2A3A4A5 and five points B1, B2, B3, B4, B5 lying on the lines A1A2, A2A3, A3A4, A4A5, A5A1, respectively. If the points B1, B2, B3, B4, B5 are collinear then

$$\frac{\widetilde{B_1A_1}}{\widetilde{B_1A_2}}\quad\frac{\widetilde{B_2A_2}}{\widetilde{B_2A_3}}\quad\frac{\widetilde{B_3A_3}}{\widetilde{B_3A_4}}\quad\frac{\widetilde{B_4A_4}}{\widetilde{B_4A_5}}\quad\frac{\widetilde{B_5A_5}}{\widetilde{B_5A_1}} = 1. \tag{2.29}$$

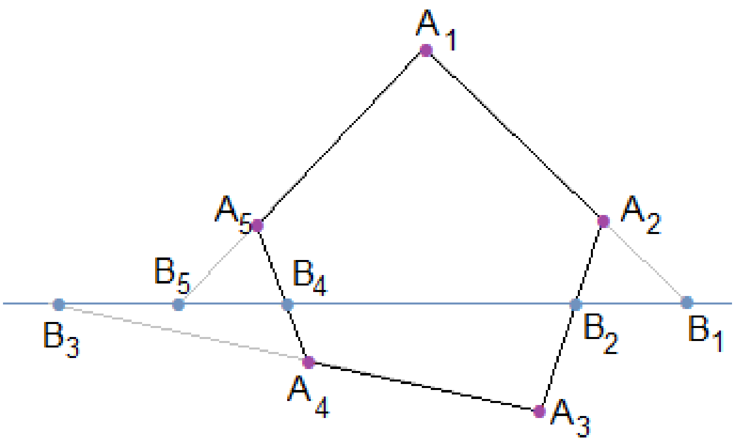

Figure 2.22: ct10.png

Proof. If the five points B1, B2, B3, B4, B5 lie on the same straight line then take two arbitrary points I, J on this line, we have

$$\frac{\overline{B_1A_1}}{\overline{B_1A_2}} \frac{\overline{B_2A_2}}{\overline{B_2A_3}} \frac{\overline{B_3A_3}}{\overline{B_3A_4}} \frac{\overline{B_4A_4}}{\overline{B_4A_5}} \frac{\overline{B_5A_5}}{\overline{B_5A_1}} \tag{2.30}$$

$$= \frac{\overline{s}(IJA_1)}{\overline{s}(IJA_2)} \frac{\overline{s}(IJA_2)}{\overline{s}(IJA_3)} \frac{\overline{s}(IJA_3)}{\overline{s}(IJA_4)} \frac{\overline{s}(IJA_4)}{\overline{s}(IJA_5)} \frac{\overline{s}(IJA_5)}{\overline{s}(IJA_1)} = 1: \tag{2.31}$$

Here are Ceva's Theorem and Menelaus' Theorem for general polygons.

Ceva's Theorem for polygon. Given a polygon A1A2...An and n points B1, ..., Bn, where each point Bi lies on the line Ai-1 Ai+1. If n lines A1B1, A2B2, ..., AnBn are concurrent then

$$\prod_{i=1}^{n} \frac{\overline{B_iA_{i-1}}}{\overline{B_iA_{i+1}}} = (-1)^n: \tag{2.32}$$

Menelaus' Theorem for polygon. Given a polygon A1A2...An and n points B1, ..., Bn, where each point Bi lies on the line Ai Ai+1. If the points B1, B2, ..., Bn are collinear then

$$\prod_{i=1}^{n} \frac{\overline{B_iA_i}}{\overline{B_iA_{i+1}}} = 1: \tag{2.33}$$

Today we have learned about Ceva's Theorem and Menelaus' Theorem. Both of these theorems are proved by using a simple theorem on the ratio of triangle's areas. This proof is so simple, yet it helps us to generalize the theorems for arbitrary polygons.

Let us stop here for now, hope to see you again in the next post.

Homework.

1. In the following figure, prove that

$$\frac{UB}{UC} = \frac{VB}{VC}: \tag{2.34}$$

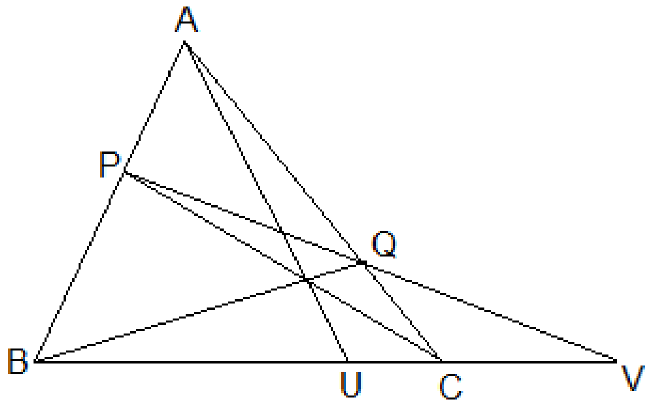

Figure 2.23: ct11.png

2. Given a triangle ABC with the lengths of the sides AB=c, BC=a, CA=b. Suppose that the incircle touches the sides of the triangle at A', B', C'. Determine the lengths of AB', AC', BA', BC', CA', CB' in terms of a, b, c. Prove that the three lines AA', BB', CC' are concurrent.

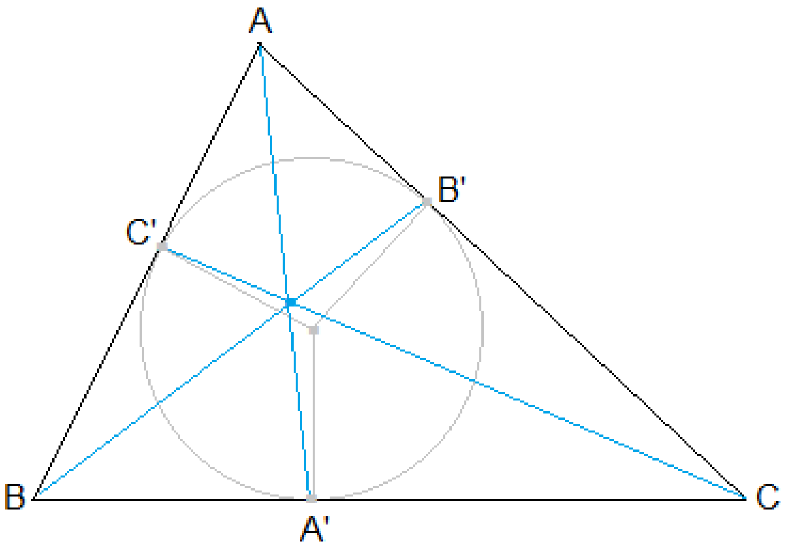

Figure 2.24: ct12.png

3. Generalize Menelaus' Theorem for points in 3D space. Below is a version of Menelaus' Theorem for tetrahedron.

Given a tetrahedron ABCD. A plane intersects the four lines AB, BC, CD, DA at the points X, Y, Z, T, respectively. Prove that

$$\frac{\widetilde{XA}}{\widetilde{XB}} \quad \frac{\widetilde{YB}}{\widetilde{YC}} \quad \frac{\widetilde{ZC}}{\widetilde{ZD}} \quad \frac{\widetilde{TD}}{\widetilde{TA}} = 1: \tag{2.35}$$

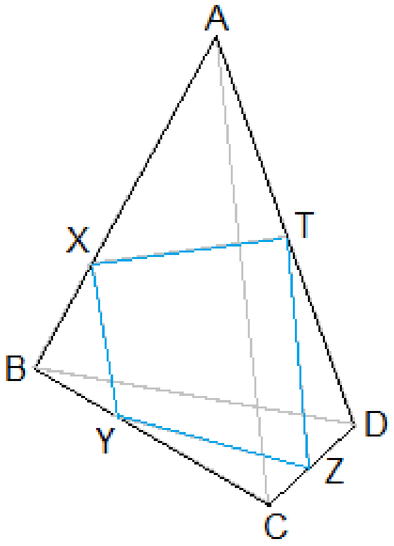

Figure 2.25: ct13.png

4. Take some examples of three points A, B, C on a coordinate system 0xy and calculate the signed area $\overline{s}(ABC)$. When do you see $\overline{s}(ABC)$ is a positive number and when $\overline{s}(ABC)$ is negative?

5. Take some examples of three points A, B, C lying on a same straight line on a coordinate system 0xy and calculate the signed area $\overline{s}(ABC)$.

6. Prove that [A,B]=-[B,A], [A,A]=0 and $\overline{s}(ABC) = \quad \overline{s}(ACB)$

7. Suppose O is the center of a coordinate system 0xy. Prove that

$$\overline{s}(OAB) = \frac{1}{2}[A; B]; \qquad \overline{s}(OBC) = \frac{1}{2}[B; C]; \qquad \overline{s}(OCA) = \frac{1}{2}[C; A]; \qquad (2.36)$$

it follows that

$$\overline{s}(ABC) = \overline{s}(OAB) + \overline{s}(OBC) + \overline{s}(OCA): \qquad (2.37)$$

Using the above identity to prove that for any point M,

$$\overline{s}(ABC) = \overline{s}(MAB) + \overline{s}(MBC) + \overline{s}(MCA) \qquad (2.38)$$

8. Take some examples of four points A, B, C, D on a coordinate system 0xy and then calculate the signed area

$$\overline{s}(ABCD) = \frac{1}{2}([A; B] + [B; C] + [C; D] + [D; A]): \qquad (2.39)$$

Check if the normal unsigned area s(ABCD) is agreeable to the signed area $\bar{s}(ABCD)$. Generalize the concept of signed area for polygons.

Source:
http://mathgardenblog.blogspot.com/2013/06/ceva-menelaus.html

2.2 Pascal's Theorem

In a previous post, we were introduced to Pascal's Hexagrammum Mysticum Theorem - a magical theorem - which states that if we draw a hexagon inscribed in a conic section then the three pairs of opposite sides of the hexagon intersect at three points which lie on a straight line.

For example, as in the following figure we have a hexagon inscribed in a circle and the intersection points of the three pairs of the opposite sides of the hexagon 12,45, 23,56, 34,61 are collinear.

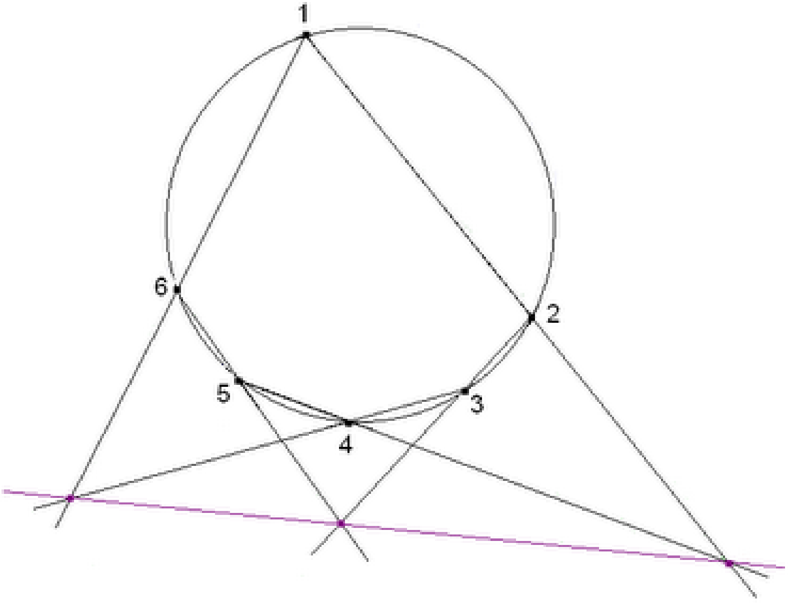

Figure 2.26: pascal1.png

There is a useful tool to prove the collinearity of points - the Menelaus' Theorem - which states as follows:

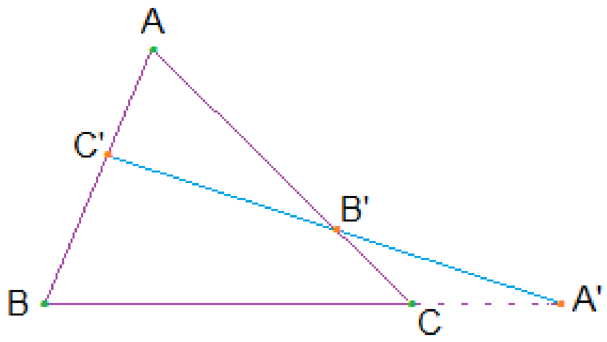

Figure 2.27: pascal2.png

Menelaus' Theorem: Given a triangle ABC and three points A', B', C' lying on the three lines BC, CA, AB, respectively. Then the three points A', B', C' are collinear if and only if

$$\frac{A'B}{A'C}\,\frac{B'C}{B'A}\,\frac{C'A}{C'B} = 1: \tag{2.40}$$

Today, we will use Menelaus' theorem to prove Pascal's theorem for the circle case.

When we want to use Menelaus' theorem to prove three certain points lying on a same straight line, we need to specify a triangle such that these three points belong to the three sides of the triangle.

In the following figure, for the Pascal's hexagon P1P2P3P4P5P6, to show that the three intersection points M1, M2, M3 are collinear, we need to find a triangle whose three sides contain M1, M2, M3. Since M1, M2, M3 lie on the sides of the hexagon, there are two natural choices for the triangle: either ABC, or XYZ.

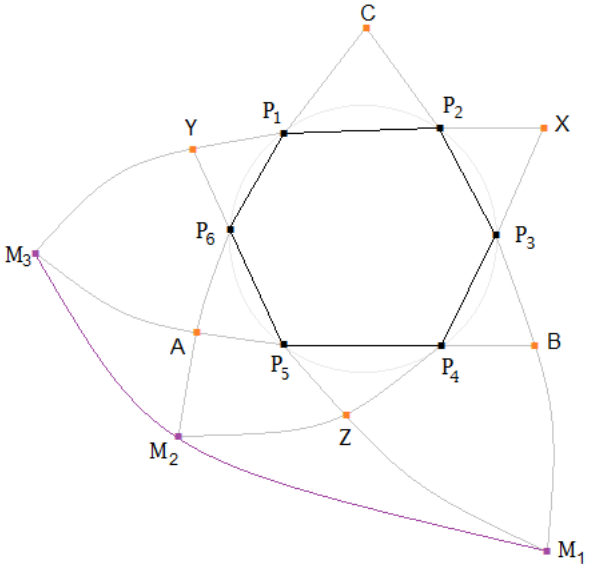

Figure 2.28: pascal3.png

Suppose we choose the triangle ABC, then we can see that the point M1 lying on the line BC, the point M2 lying on the line CA and the point M3 lying on the line AB. To use Menelaus' theorem, we need to show

$$\frac{M_1B}{M_1C}\frac{M_2C}{M_2A}\frac{M_3A}{M_3B} = 1: \qquad (2.41)$$

To calculate these quotients, we will apply Menelaus' theorem for the following triples of collinear points:

$$\{M_1; P_5; P_6\}; \quad \{M_2; P_3; P_4\}; \quad \{M_3; P_1; P_2\}: \qquad (2.42)$$

Let us now write down the proof in details.

A proof of Pascal's theorem

We will apply Menelaus' theorem for the triangle ABC.

Since M1, M2, M3 lie on the three sides of the triangle ABC, to prove that they are collinear, we need to show

$$\frac{M_1B}{M_1C}\frac{M_2C}{M_2A}\frac{M_3A}{M_3B} = 1: \qquad (2.43)$$

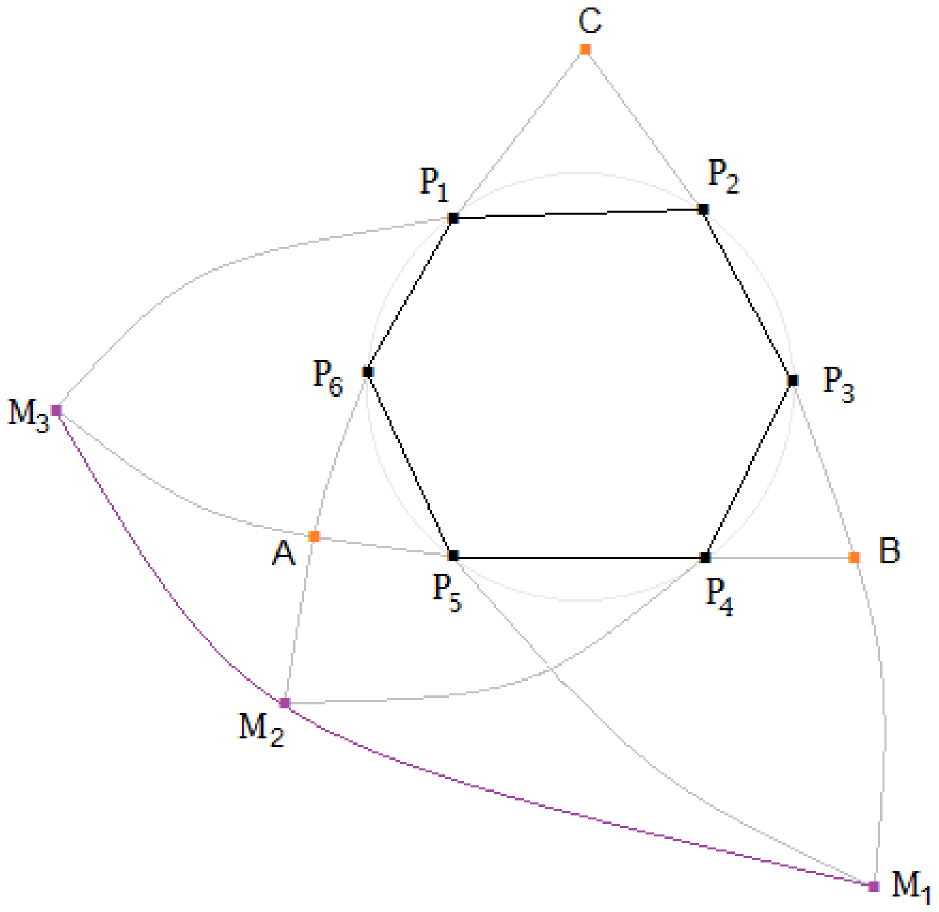

Figure 2.29: pascal4.png

Indeed, apply Menelaus’ theorem for the triangle ABC with the following triples of collinear points

$$\{M_1, P_6, P_5\}; \quad \{M_2, P_4, P_3\}; \quad \{M_3, P_2, P_1\}; \tag{2.44}$$

we have

$$\frac{\overline{M_1B}}{\overline{M_1C}} \cdot \frac{\overline{P_6C}}{\overline{P_6A}} \cdot \frac{\overline{P_5A}}{\overline{P_5B}} = \frac{\overline{M_2C}}{\overline{M_2A}} \cdot \frac{\overline{P_4A}}{\overline{P_4B}} \cdot \frac{\overline{P_3B}}{\overline{P_3C}} = \frac{\overline{M_3A}}{\overline{M_3B}} \cdot \frac{\overline{P_2B}}{\overline{P_2C}} \cdot \frac{\overline{P_1C}}{\overline{P_1A}} = 1. \tag{2.45}$$

Thus,

$$\frac{\overline{M_1B}}{\overline{M_1C}} = \frac{\overline{P_6A}}{\overline{P_6C}}\,\frac{\overline{P_5B}}{\overline{P_5A}};\quad \frac{\overline{M_2C}}{\overline{M_2A}} = \frac{\overline{P_4B}}{\overline{P_4A}}\,\frac{\overline{P_3C}}{\overline{P_3B}};\quad \frac{\overline{M_3A}}{\overline{M_3B}} = \frac{\overline{P_2C}}{\overline{P_2B}}\,\frac{\overline{P_1A}}{\overline{P_1C}}: \qquad (2.46)$$

It follows that

$$\frac{\overline{M_1B}}{\overline{M_1C}}\,\frac{\overline{M_2C}}{\overline{M_2A}}\,\frac{\overline{M_3A}}{\overline{M_3B}} = \frac{\overline{P_6A}}{\overline{P_6C}}\,\frac{\overline{P_5B}}{\overline{P_5A}}\,\frac{\overline{P_4B}}{\overline{P_4A}}\,\frac{\overline{P_3C}}{\overline{P_3B}}\,\frac{\overline{P_2C}}{\overline{P_2B}}\,\frac{\overline{P_1A}}{\overline{P_1C}} \qquad (2.47)$$

$$= \frac{\overline{AP_1}\ \overline{AP_6}}{\overline{AP_4}\ \overline{AP_5}}\,\frac{\overline{BP_4}\ \overline{BP_5}}{\overline{BP_2}\ \overline{BP_3}}\,\frac{\overline{CP_2}\ \overline{CP_3}}{\overline{CP_1}\ \overline{CP_6}} = 1: \qquad (2.48)$$

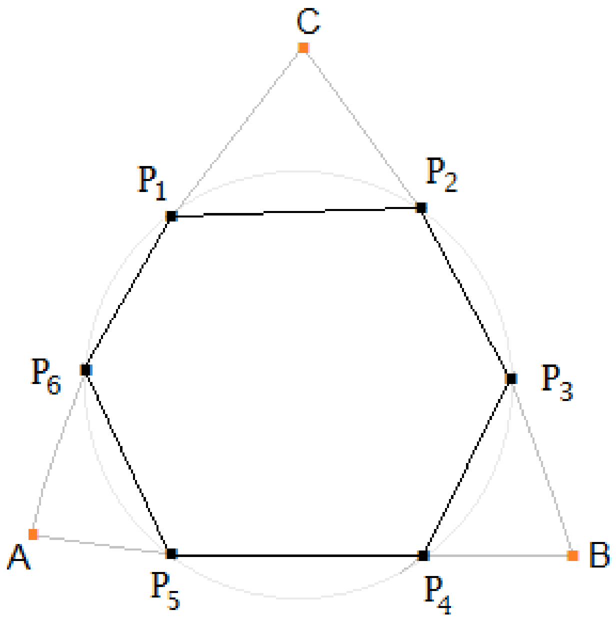

Figure 2.30: pascal5.png

Power of a point: $\overline{AP_1}\ \overline{AP_6} = \overline{AP_4}\ \overline{AP_5} = \overline{CP_2}\ \overline{CP_3} = \overline{CP_1}\ \overline{CP_6}$

The last equality is from the power of a point property when we apply the power of the points A, B, C to the circumcircle of the hexagon. Thus, we have completed the proof of Pascal's theorem.

Today, we have shown how to use Menelaus' theorem effectively to prove the Pascal's "magical" theorem. Indeed, Menelaus' theorem is a very useful tool often employed to prove the collinearity of points. Have a try now to see if we can use Menelaus' theorem to prove the Pappus' theorem!

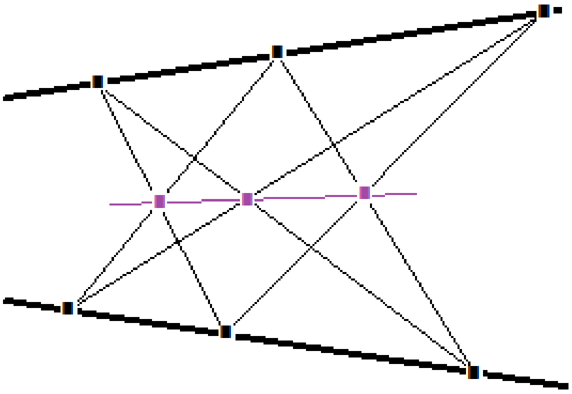

Figure 2.31: pascal6.png

Homework.

1. Prove Pascal's theorem by applying Menelaus' theorem for the triangle XYZ.

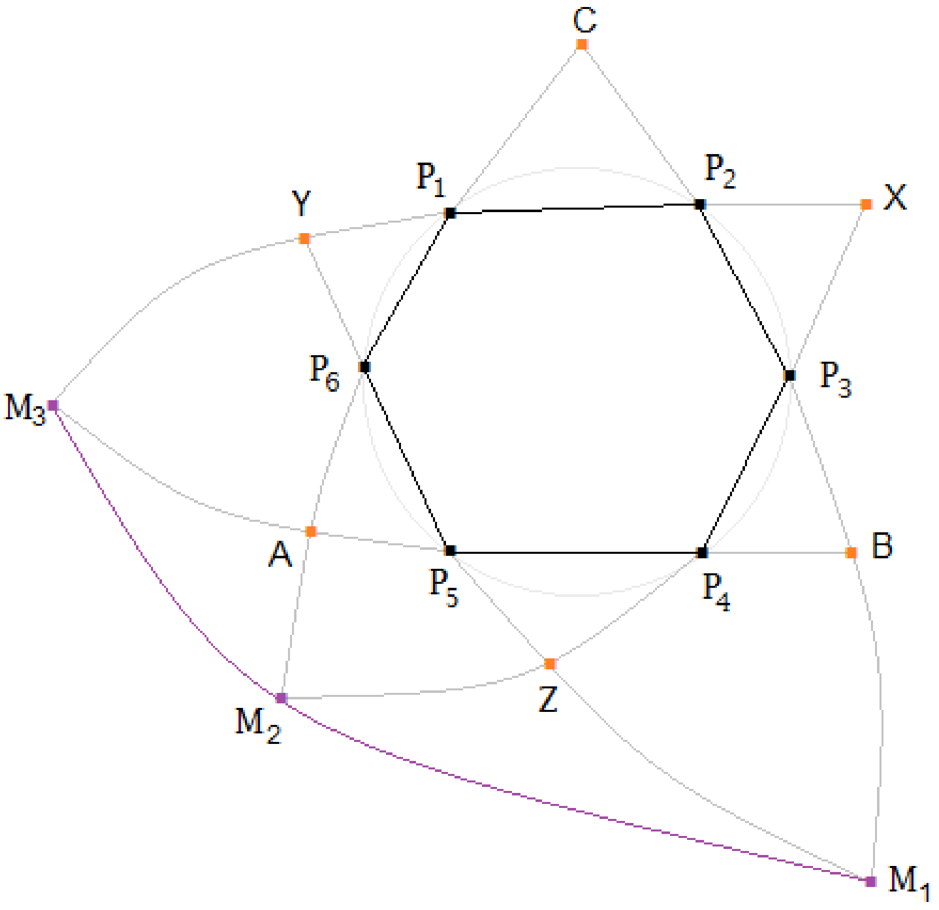

Figure 2.32: pascal7.png

Source:
http://mathgardenblog.blogspot.com/2013/08/hexagrammum-mysticum2.html

2.3 The Fermat's point of a triangle

Date: 20200728

In previous posts about modulo, we learn about the mathematician Fermat and his famous problem

$$x^n + y^n = z^n: \tag{2.49}$$

Today, we will look at a geometry problem that bears his name. As we already know, Fermat was not a professional mathematician, but was a lawyer. He was doing math probably just for fun and most of his achievements that we know of today originated from his letters to his friends and also from his occasional writings on the margin of books that he read. The most famous is, of course, the problem $x^n + y^n = z^n$ and his note "I have found a beautiful proof but there is not enough space" that he wrote on the margin of the book by Diophantus.

The problem that we investigate today was raised in a letter that Fermat sent to an Italian mathematician, Torricelli. In his letter, Fermat challenged Torricelli to find a point such that the total distance from this point to the three vertices of a triangle is the minimum possible. Well, this problem was not hard for Torricelli. Since Torricelli knew how to find such a point, today some people refer to this point as the Fermat's point, and others refer it as the Torricelli's point of the triangle.

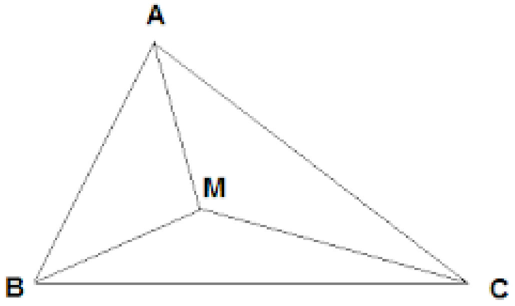

Figure 2.33: fermat1.png

the Fermat's problem: find a point M so that MA+MB+MC is minimum

To find the solution to the Fermat's problem, we are going to use a property of the ellipse. That is, if we have an ellipse with two focus points F1 and F2, then from any point P on the ellipse if we draw a tangent line to the ellipse then this tangent line will form two equal angles with the two generator lines PF1 and PF2.

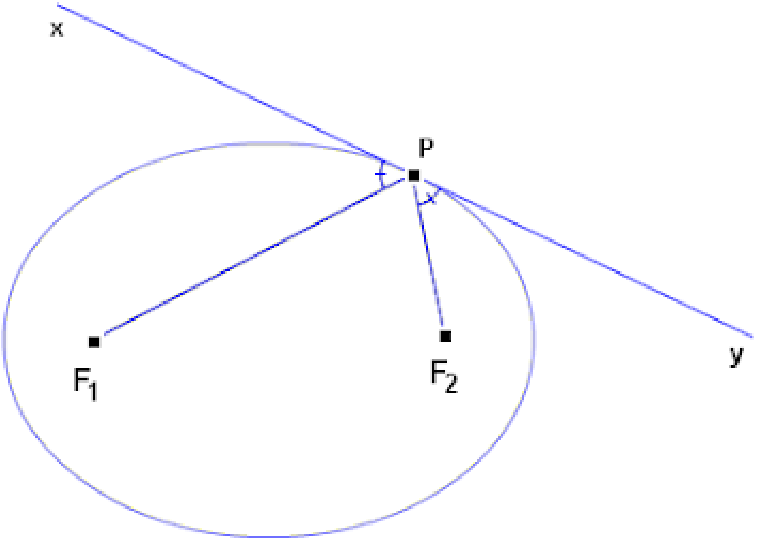

Figure 2.34: fermat2.png

a property of the tangent line: <xPF1=<yPF2

You can read more about this property of the tangent line to the ellipse at this post "A problem about finding shortest path and a property of the ellipse".

Going back to the Fermat problem, suppose that we have found a point M so that MA+MB+MC is minimum. Let us do some analysis to find out where this point should be.

Let MA+MB=l. Draw an ellipse that passes through M and that has two focus points A and B. For any point P on the ellipse we have PA+PB=MA+MB=l. But because MA+MB+MC is the minimum, we must have MC=<PC. It means that the distance from C to a point P on the ellipse is minimum when P=M.

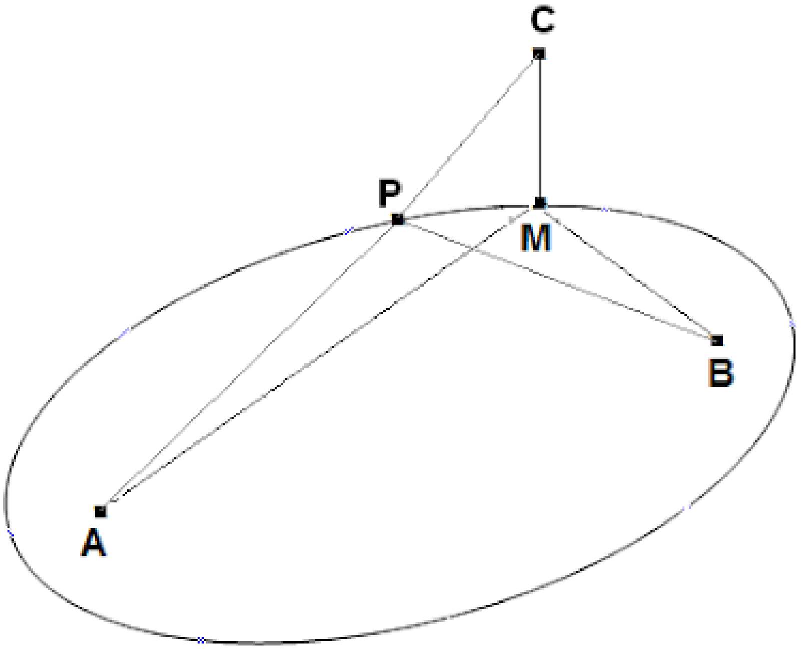

Figure 2.35: fermat3.png

from all distances from C to the ellipse, CM is the shortest

Because from all distances from C to the ellipse, CM is the shortest, if we draw a circle with the center C and radius CM then this circle must touch the ellipse at one point M.

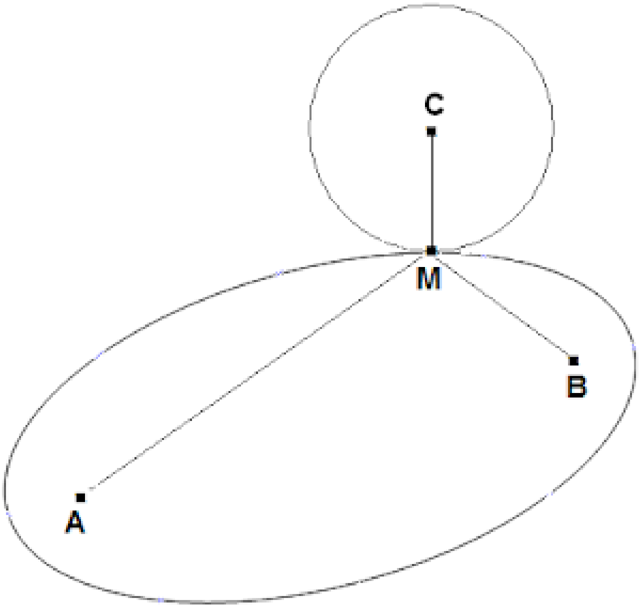

Figure 2.36: fermat4.png

the circle must touch the ellipse at one point M

It follows that the ellipse and the circle must have a common tangent line that passes through M as shown in the following figure.

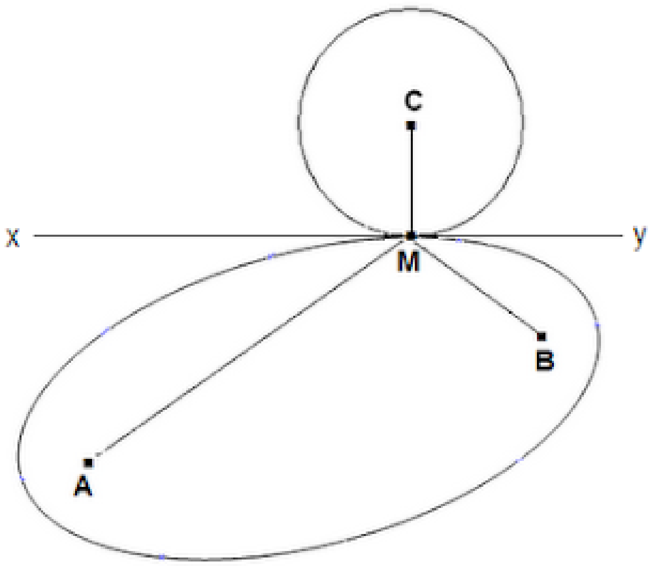

Figure 2.37: fermat5.png

Here we need to be a bit careful. There is an exceptional case that the ellipse and the circle may not have a common tangent line through M. That is when M=A or M=B. For instance, when M=B then l=MA+MB=AB, in this case there is no longer an ellipse, the ellipse has degenerated and becomes a "flat ellipse", i.e. the line segment AB. In the figure below, we can see that the tangent line to the circle is cM, on the other hand, the tangent line to the ellipse is eM { two different tangent lines.

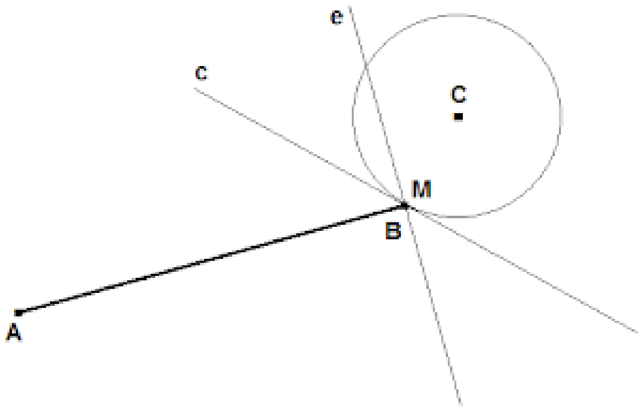

Figure 2.38: fermat6.png

in an exceptional case M=B, the tangent line to the circle is cM and the tangent line to the ellipse is eM

In summary, except for the case M=A or M=B, the circle and the ellipse will have a common tangent line through M as illustrated in the figure below. By the property of the tangent line to the ellipse, we have <xMA=<yMB. We also have <xMC=<yMC=90o. Therefore, <CMA=<CMB.

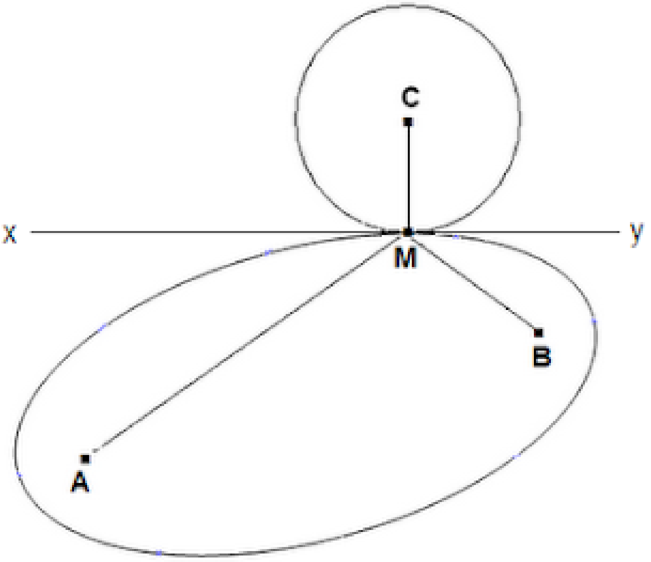

Figure 2.39: fermat7.png

except when M=A or M=B, the ellipse and the circle have a common tangent line xMy and <CMA=<CMB

By similar argument, we conclude that the point M can only be in four positions. It is either at A, B, C, or a point that is specified by <AMB=<BMC=<CMA.

In the case <AMB=<BMC=<CMA then <AMB=<BMC=<CMA=120o.

How interesting! The Fermat problem and the ellipse seem so unrelated, but we have used a property of the ellipse to work out possible solutions for the Fermat problem. The point M so that MA+MB+MC is minimum can only be in four positions. The point M can be A, B, C, or a point such that <AMB=<BMC=<CMA=120o.

Let us stop here for now. In the next post, we will solve the Fermat problem with more details. We will see that the solution goes into two cases. The first case is when one of the angles of the triangle ABC is greater than or equal to 120o. The second case is when all the angles of the triangle ABC are less than 120o.

In the first case, if <A>=120o then the point M is A.

In the second case, if all the angles of the triangle ABC are less than 120o then the point M is the point specified by <AMB=<BMC=<CMA=120o. We can construct this point M as follows. Draw outside the triangle ABC three equilateral triangles ABC', BCA', CAB'. Draw three circumcircles of these equilateral triangles then these circles will meet at one common point M. Draw the three line segments AA', BB', CC', then these three lines also meet at M (see the figure below).

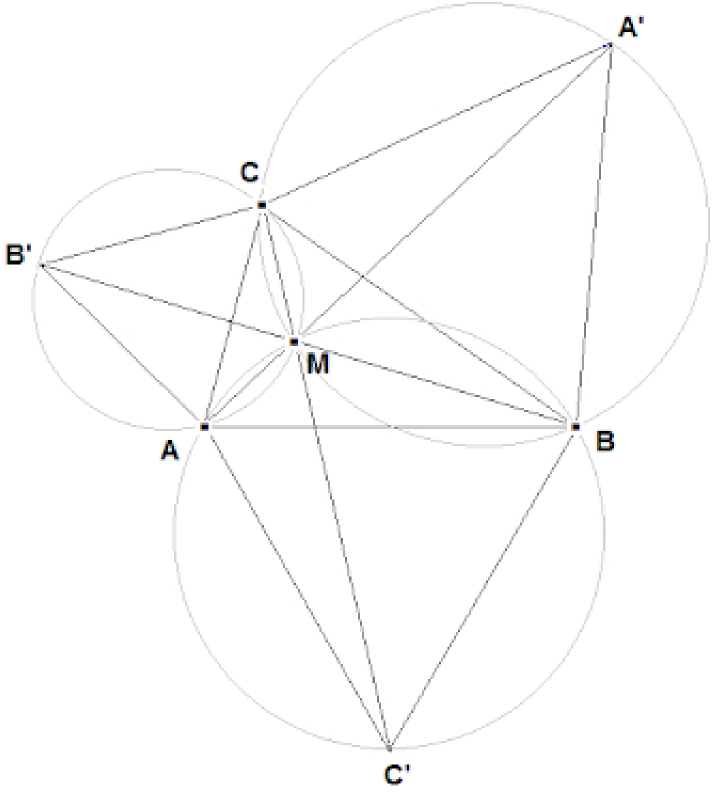

Figure 2.40: fermat*.png

Fermat's point M when <A,<B,<B<120o

Homework. Prove that if an ellipse and a circle tangent at one point then through this point we can draw a common tangent line for both the circle and the ellipse.

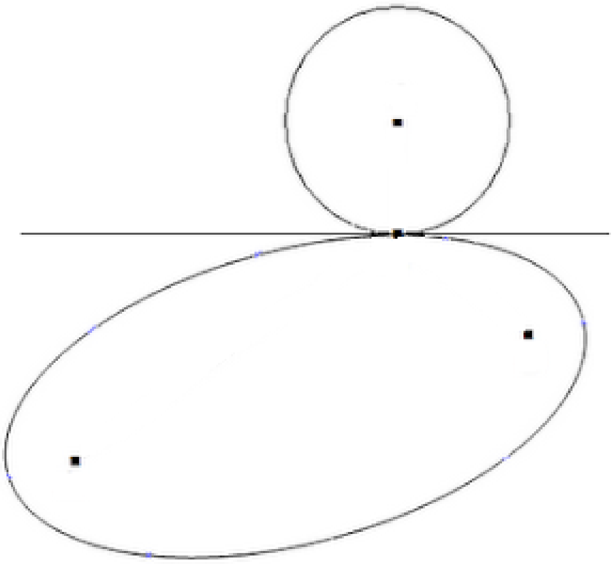

Figure 2.41: fermat9.png

Source:
http://mathgardenblog.blogspot.com/2012/07/fermat-point-of-triangle1.html

The Fermat's point of a triangle II

In the previous post, we have analyzed the Fermat's problem, the problem of finding a point M for a given triangle ABC such that MA+MB+MC is the minimum.

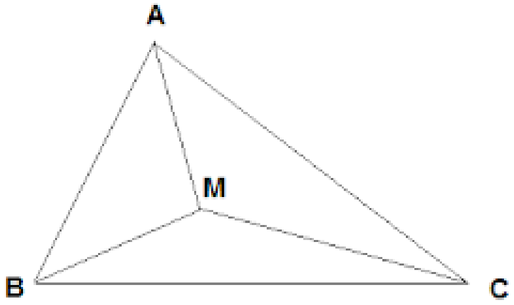

Figure 2.42: fermat10.png

the Fermat's problem: find M so that MA+MB+MC is minimum

The interesting part is that we have used a property of the ellipse to work out the solution. Our analysis shows that the point M, if it exists, can only be at four possible locations. It is either the point A, B, C, or the point specified by <AMB=<BMC=<CMA=120o.

Let us emphasize that, in the above statement, we say that if this point M exists then it has to be at four possible positions. However, we have not shown that such a point M, for which MA+MB+MC is minimal, exists. So with mathematical arguments, we have to be extra careful.

Today, we will solve the Fermat problem in a "more conventional" way! We will look at two separate cases.

The first case is when all the angles of the triangle ABC are smaller than 120o. We will show that the point M that we are looking for is specified by <AMB=<BMC=<CMA=120o.

The second case is when one of the angles of the triangle ABC is greater than or equal to 120o. For example, when <A>=120o, we will show that the point M is A.

Case 1: <A,<B,<C<120o.

Draw outside the triangle ABC three equilateral triangles ABC', BCA' and CAB'.

Take an arbitrary point P. Use A as the center, make a 60o rotation so that the triangle APB will be rotated to the triangle AQC'. By this rotation, we have obtained two things. Firstly, two triangles APB and AQC' are equal triangles, and secondly, the triangle APQ is an equilateral triangle. So we have PB=QC' and PA=PQ.

Using rotation, we transform PA+PB+PC=CP+PQ+QC'. So PA+PB+PC will be minimal when it is equal to CC', it happens when the four points C, P, Q and C', in this order, lying on the same straight line.

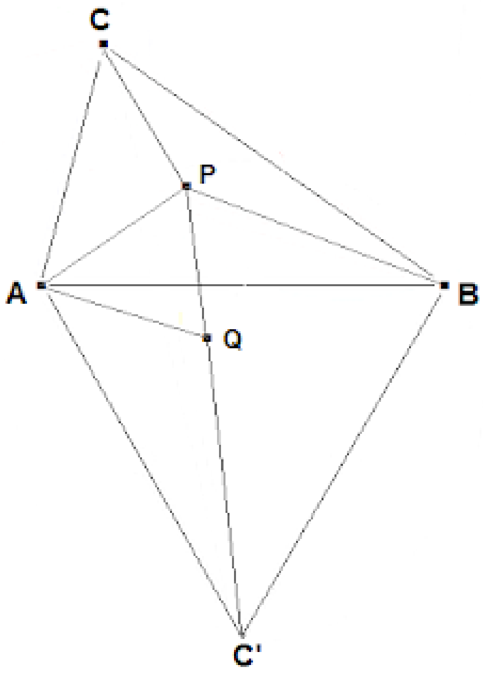

Figure 2.43: fermat11.png

PA=PQ, PB=QC’, so PA+PB+PC=CP+PQ+QC’>=CC’

To make C, P, Q lie on a straight line we must have <CPA=180o - <APQ=120o.

Similarly, to have P, Q, C’ lying on a straight line we must have <AQC’=180o. -<PQA=120o, and so <APB=<AQC’=120o.

Since <APQ=60o, if C, P, Q, C’ lie on a straight line then P must be the intersection point of CC’ with the circumcircle of the equilateral triangle ABC’.

In summary, we have shown that PA+PB+PC>=CC’ and the necessary condition for PA+PB+PC=CC’ is that <CPA=<APB=120o and the point P is the intersection point of CC’ with the circumcircle of the equilateral triangle ABC’.

To complete the proof, we construct the point M such that <CMA=<AMB=<BMC=120o and we show that MA+MB+MC=CC’, thus showing M is the point that we are looking for.

Indeed, if we take M as the intersection point of CC’ with the circumcircle of the equilateral triangle ABC’ then <AMB=120o and <AMC’=60o, and thus, <AMC=120o. Let N be the point on CC’ so that AMN is an equilateral triangle. By the rotation of 60o at center A, the triangle AMB will

be rotated to the triangle ANC'. Thus, the point M satisfies the equality MA+MB+MC=MN+NC'+MC=CC'.

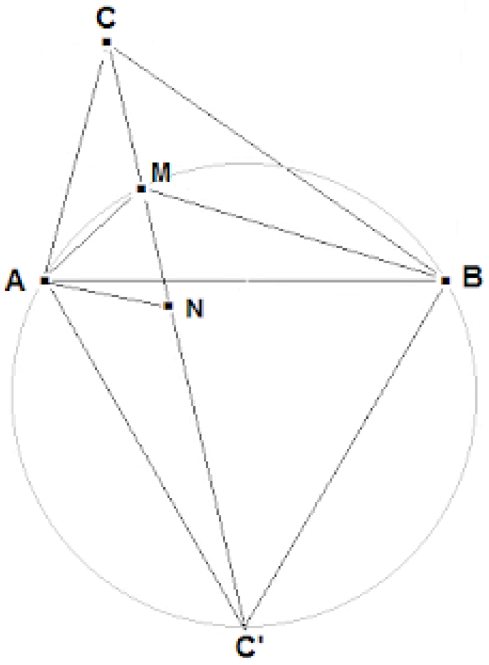

Figure 2.44: fermat12.png

At the Fermat's point M we have MA+MB+MC=CC'

We have constructed the point M such that <AMC=<AMB=<BMC=120o and showed that MA+MB+MC=CC'. Thus, M is our solution point.

By similarity, M must also lie on AA' and BB' and we have AA'=BB'=CC', and also the three circumcircles of the equilateral triangles ABC', BCA', CAB' must meet at M.

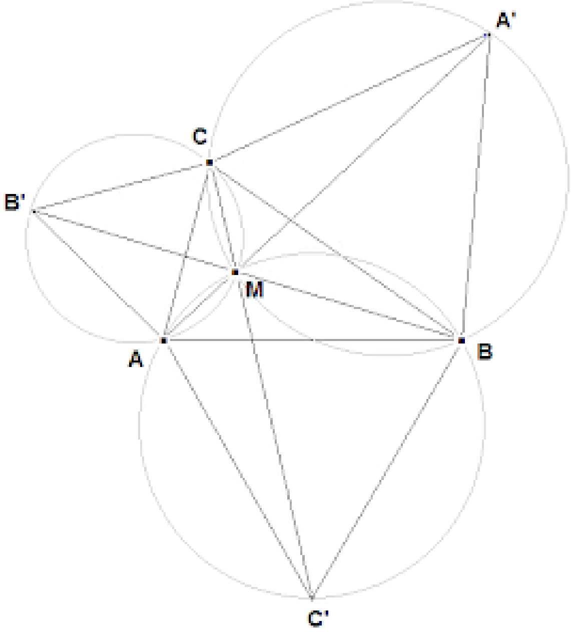

Figure 2.45: fermat13.png

We remark that the condition that <A,<B,<C<120o is a necessary condition to ensure that the four points C, M, N and C' lie on a straight line in this order. If one of the angles of the triangle ABC is greater than 120o then the points C, M, N and C' lie on a straight line, but not in this order (see the figure below).

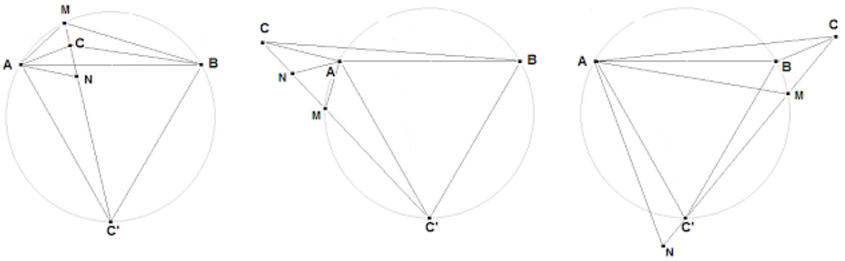

Figure 2.46: fermat14.png

if one of the angles of the triangle ABC is greater than 120o then C, M, N and C' lie on a straight line, but not in this order

Case 2: <A>=120o.

Here we will sketch the proof, the details is left for the readers.

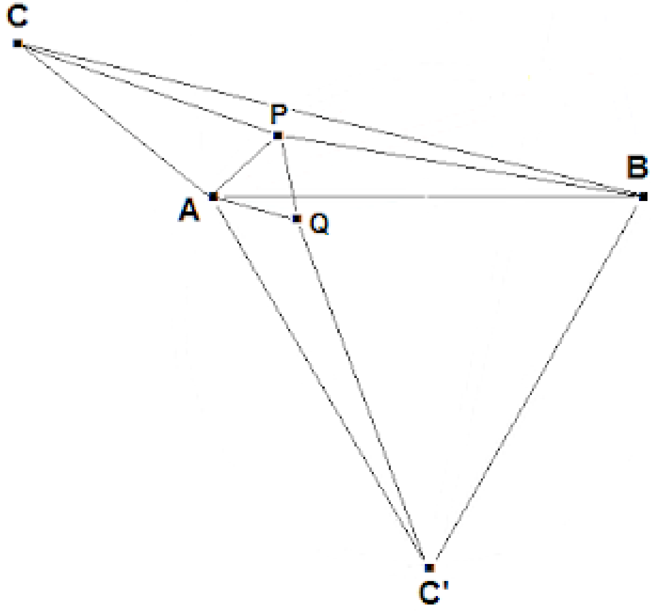

Figure 2.47: fermat15.png

PA=PQ, PB=QC', so PA+PB+PC=CP+PQ+QC'>=CA+AC'=CA+AB

The triangle AQC' is obtained from the triangle APB by a rotation so PB=QC'. The triangle APQ is an equilateral triangle, so PA=PQ. Therefore, PA+PB+PC=CP+PQ+QC'.

Next, we need to prove that CP+PQ+QC'>=CA+AC'=CA+AB and CP+PQ+QC' is minimum when P=A.

If you need some hints to prove the inequality CP+PQ+QC'>=CA+AC', please read this post "Solve for special cases first!".

Let us stop here for now. If you know other ways to solve the Fermat's problem please post it here so we all can share. See you again in the next post!

Homework. Given a square ABCD. Find on the four sides of the square four points X, Y, Z, T so that XY+YZ+ZT+TX is minimum.

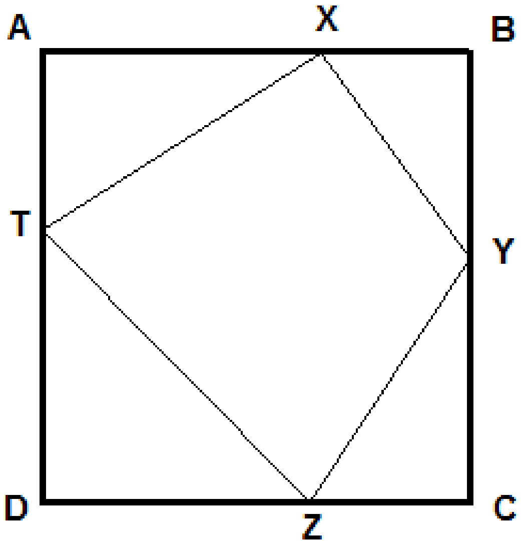

Figure 2.48: fermat16.png

Source:
http://mathgardenblog.blogspot.com/2012/07/fermat-point-of-triangle2.html

Chapter 3

Pascal’s triangle

Today, we will look at a famous number pattern, the Pascal’s number triangle.

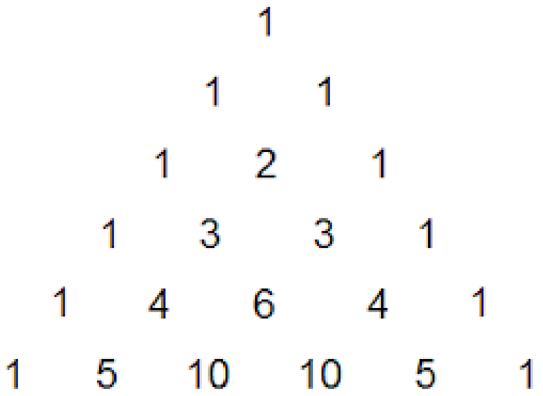

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1

Figure 3.1: pnt1.png

Pascal’s triangle

This number triangle is constructed as follows.

In the first row, we write the number 1. In the next row, we put down two 1s. In subsequent rows, the first number and the last number are always equal to 1; for other number, it will be equal to the sum of two numbers sitting just above it in the previous row.

For example, 1+1=2, 1+2=3, 2+1=3, 1+3=4, 3+3=6, 3+1=4, etc...

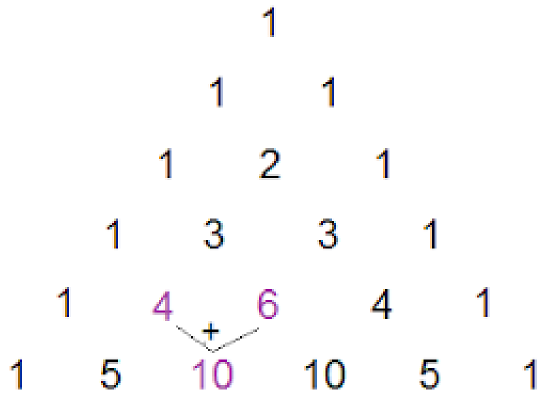
1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
+
1 5 10 10 5 1

Figure 3.2: pnt2.png

constructing Pascal's triangle: 4+6=10

We use Pascal's triangle to determine the coefficients in the expansions of $(x + y)^n$ and $(x - y)^n$.

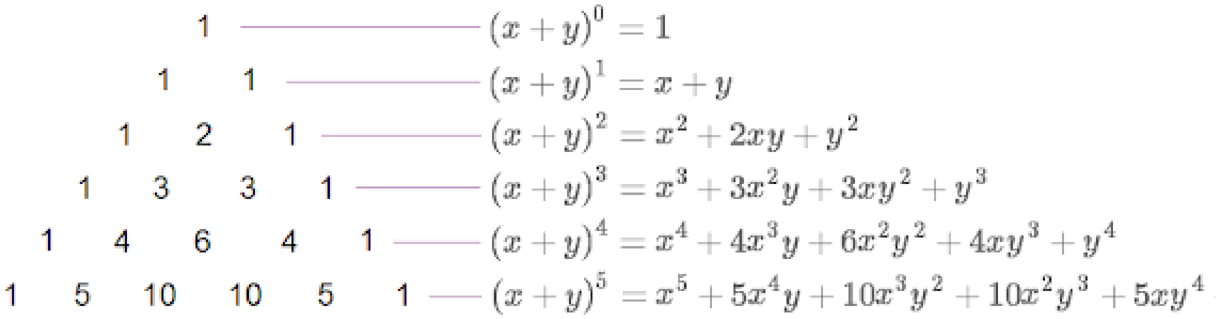
1 $(x+y)^0 = 1$
1 1 $(x+y)^1 = x + y$
1 2 1 $(x+y)^2 = x^2 + 2xy + y^2$
1 3 3 1 $(x+y)^3 = x^3 + 3x^2y + 3xy^2 + y^3$
1 4 6 4 1 $(x+y)^4 = x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4$
1 5 10 10 5 1 $(x+y)^5 = x^5 + 5x^4y + 10x^3y^2 + 10x^2y^3 + 5xy^4$

Figure 3.3: pnt3.png

expand (x+y)n using Pascal's coefficients

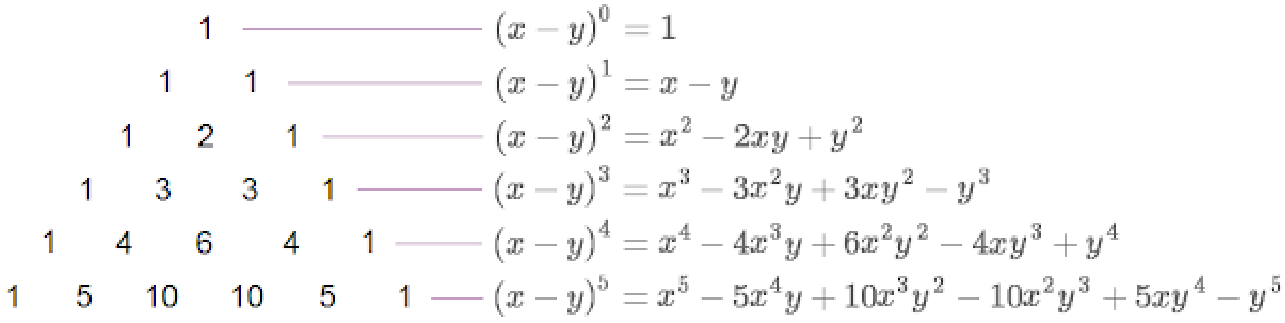
1 — $(x-y)^0=1$
1 1 — $(x-y)^1=x-y$
1 2 1 — $(x-y)^2=x^2-2xy+y^2$
1 3 3 1 — $(x-y)^3=x^3-3x^2y+3xy^2-y^3$
1 4 6 4 1 — $(x-y)^4=x^4-4x^3y+6x^2y^2-4xy^3+y^4$
1 5 10 10 5 1 — $(x-y)^5=x^5-5x^4y+10x^3y^2-10x^2y^3+5xy^4-y^5$

Figure 3.4: pnt4.png

expand (x-y)n using Pascal's coefficients

We will give each row of the Pascal's triangle an index. The first row of the Pascal's triangle is row 0, the next row is row 1, then row 2, ect..., The numbers on each row will also be given a similar index. The first number is number 0, then number 1, number 2, ect...

If we use $p_{n;k}$ to denote the number of index k on the row n then the Pascal's triangle is constructed based on the following rule

$$p_{n-1;k-1} + p_{n-1;k} = p_{n;k}: \tag{3.1}$$

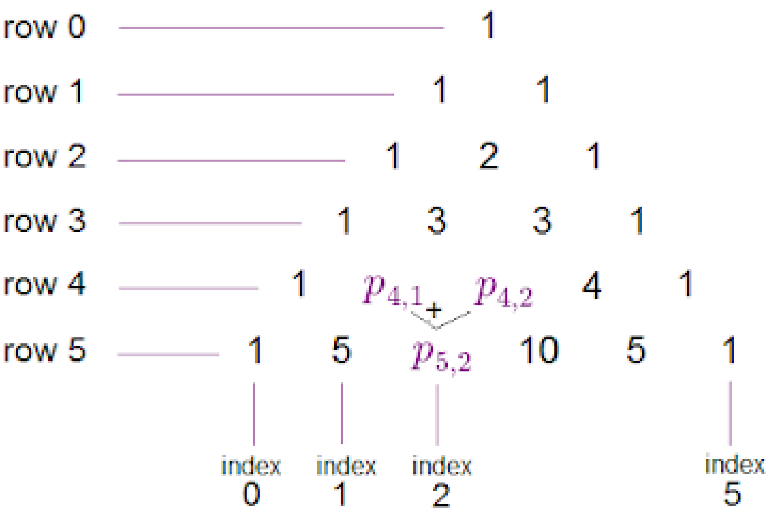
row 0
row 1
row 2
row 3
row 4
row 5
1
1 1
1 2 1
1 3 3 1
1 $p_{4,1}$ + $p_{4,2}$ 4 1
1 5 $p_{5,2}$ 10 5 1
index 0
index 1
index 2
index 5

Figure 3.5: pnt5.png

the rule to construct Pascal's triangle $p_{n-1;k-1} + p_{n-1;k} = p_{n;k}$

The general formula for $p_{n;k}$ is

$$p_{n;k} = \binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{3.2}$$

For example,

$$p_{5;2} = \binom{5}{2} = \frac{5!}{2!3!} = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{1 \cdot 2 \cdot 1 \cdot 2 \cdot 3} = 10: \tag{3.3}$$

Finally, let us remark that we would normally start the indexing with index 1, then 2, 3, ..., but with Pascal's triangle, we use a special indexing that starts with 0. To help us remember this special indexing, let us read the following story about Sierpinski.

Waclaw Sierpinski was a great Polish mathematician. But one story told us that he was quite an absent-minded man. On one occasion, he and his wife had to move into a new residence. His wife didn't trust him very much, so when they stood down on the street with all their things, she said "Now, you stand here and watch our ten trunks, while I go and get a taxi." She left and left him there, gazing off into space and humming absently. Some minutes later she returned, presumably having called for a taxi. Said Sierpinski (possibly with a glint in his eye): "I thought you said there were ten trunks, but I've only counted to nine." Panicked that some of their possessions might have been stolen from under their noses, his wife demanded "No, they were TEN!", "No, no, no! Just count them. 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 !"

Sierpinski had written a very good book on number theory, which was translated to English with the title "250 problems in elementary number theory". If you are interested in number theory, it is a good book to read.

See you again in the next post.

Homework.

1. Explain why the sum of all numbers on the row n of Pascal's triangle is equal to 2n.

2. Prove that the number of index k on the row n is equal to

$$p_{n;k} = \binom{n}{k} = \frac{n!}{k!(n-k)!}: \tag{3.4}$$

3. Prove the identity

$$(x+y)^n = x^n + \binom{n}{1} x^{n-1}y + \binom{n}{2} x^{n-2}y^2 + \cdots + \binom{n}{n-2} x^2y^{n-2} + \binom{n}{1} xy^{n-1} + y^n \tag{3.5}$$

Source: http://mathgardenblog.blogspot.com/2012/09/pascal-triangle.html

3.1 Gauss' trigonometric identities for heptadecagon

Today, we write down Gauss' magical trigonometric identities for regular heptadecagon.

$$\cos\frac{\pi}{17} = \frac{1}{16}\left(1 - \sqrt{17} + \sqrt{34 - 2\sqrt{17}} + 2\sqrt{17 + 3\sqrt{17} + \sqrt{34 - 2\sqrt{17}} + 2\sqrt{34 + 2\sqrt{17}}}\right) \tag{3.6}$$

$$\cos\frac{2\pi}{17} = \frac{1}{16}\left(-1 + \sqrt{17} + \sqrt{34 - 2\sqrt{17}} + 2\sqrt{17 + 3\sqrt{17} - \sqrt{34 - 2\sqrt{17}} - 2\sqrt{34 + 2\sqrt{17}}}\right) \tag{3.7}$$

$$\cos\frac{3\pi}{17} = \frac{1}{16}\left(1 + \sqrt{17} + \sqrt{34 + 2\sqrt{17}} + 2\sqrt{17 - 3\sqrt{17} + \sqrt{34 + 2\sqrt{17}} - 2\sqrt{34 - 2\sqrt{17}}}\right) \tag{3.8}$$

$$\cos\frac{4\pi}{17} = \frac{1}{16}\left(-1 + \sqrt{17} - \sqrt{34 - 2\sqrt{17}} + 2\sqrt{17 + 3\sqrt{17} + \sqrt{34 - 2\sqrt{17}} + 2\sqrt{34 + 2\sqrt{17}}}\right) \tag{3.9}$$

$$\cos\frac{5\pi}{17} = \frac{1}{16}\left(1 + \sqrt{17} + \sqrt{34 + 2\sqrt{17}} - 2\sqrt{17 - 3\sqrt{17} + \sqrt{34 + 2\sqrt{17}} - 2\sqrt{34 - 2\sqrt{17}}}\right) \tag{3.10}$$

$$\cos\frac{6}{17} = \frac{1}{16}\left(-1 - \sqrt{17} + \sqrt{34 + 2\sqrt{17}} + 2\sqrt{17 - 3\sqrt{17} - \sqrt{34 + 2\sqrt{17}} + 2\sqrt{34 - 2\sqrt{17}}}\right) \tag{3.11}$$

$$\cos\frac{7}{17} = \frac{1}{16}\left(1 + \sqrt{17} - \sqrt{34 + 2\sqrt{17}} + 2\sqrt{17 - 3\sqrt{17} - \sqrt{34 + 2\sqrt{17}} + 2\sqrt{34 - 2\sqrt{17}}}\right) \tag{3.12}$$

$$\cos\frac{8}{17} = \frac{1}{16}\left(-1 + \sqrt{17} + \sqrt{34 - 2\sqrt{17}} - 2\sqrt{17 + 3\sqrt{17} - \sqrt{34 - 2\sqrt{17}} - 2\sqrt{34 + 2\sqrt{17}}}\right) \tag{3.13}$$

$$\cos\frac{9}{17} = \frac{1}{16}\left(1 - \sqrt{17} - \sqrt{34 - 2\sqrt{17}} + 2\sqrt{17 + 3\sqrt{17} - \sqrt{34 - 2\sqrt{17}} - 2\sqrt{34 + 2\sqrt{17}}}\right) \tag{3.14}$$

$$\cos\frac{11}{17} = \frac{1}{16}\left(1 + \sqrt{17} - \sqrt{34 + 2\sqrt{17}} - 2\sqrt{17 - 3\sqrt{17} - \sqrt{34 + 2\sqrt{17}} + 2\sqrt{34 - 2\sqrt{17}}}\right) \tag{3.15}$$

$$\cos\frac{11}{17} = \frac{1}{16}\left(1 + \sqrt{17} - \sqrt{34 + 2\sqrt{17}} - 2\sqrt{17 - 3\sqrt{17} - \sqrt{34 + 2\sqrt{17}} + 2\sqrt{34 - 2\sqrt{17}}}\right) \tag{3.16}$$

$$\cos\frac{12}{17} = \frac{1}{16}\left(-1 - \sqrt{17} - \sqrt{34 + 2\sqrt{17}} + 2\sqrt{17 - 3\sqrt{17} + \sqrt{34 + 2\sqrt{17}} - 2\sqrt{34 - 2\sqrt{17}}}\right) \tag{3.17}$$

$$\cos\frac{13}{17} = \frac{1}{16}\left(1 - \sqrt{17} + \sqrt{34 - 2\sqrt{17}} - 2\sqrt{17 + 3\sqrt{17} + \sqrt{34 - 2\sqrt{17}} + 2\sqrt{34 + 2\sqrt{17}}}\right) \tag{3.18}$$

$$\cos\frac{14\pi}{17} = \frac{1}{16}\left(-1-\sqrt{17}-\sqrt{34+2\sqrt{17}}-2\sqrt{17-3\sqrt{17}+\sqrt{34+2\sqrt{17}}-2\sqrt{34-2\sqrt{17}}}\right) \tag{3.19}$$

$$\cos\frac{15\pi}{17} = \frac{1}{16}\left(1-\sqrt{17}-\sqrt{34-2\sqrt{17}}-2\sqrt{17+3\sqrt{17}-\sqrt{34-2\sqrt{17}}-2\sqrt{34+2\sqrt{17}}}\right) \tag{3.20}$$

$$\cos\frac{16\pi}{17} = \frac{1}{16}\left(-1+\sqrt{17}-\sqrt{34-2\sqrt{17}}-2\sqrt{17+3\sqrt{17}+\sqrt{34-2\sqrt{17}}+2\sqrt{34+2\sqrt{17}}}\right) \tag{3.21}$$

Source:
http://mathgardenblog.blogspot.com/2014/11/ Gauss-17-gon-trigonometry.html

3.2 Trigonometric multiple-angle formulas

In our previous post, we show a compass-and-straightedge construction of a regular pentagon based on the following trigonometric formula

$$\cos\frac{\pi}{5} = \frac{1+\sqrt{5}}{4}. \tag{3.22}$$

We derive this formula of $\cos\frac{\pi}{5}$ by observing that $\cos\frac{2\pi}{5} = -\cos\frac{3\pi}{5}$ and then applying the trigonometric formulas for double angle and triple angle:

$$\cos 2x = 2\cos^2 x - 1; \tag{3.23}$$

$$\cos 3x = 4\cos^3 x - 3\cos x \tag{3.24}$$

to set up a cubic equation for $\cos\frac{\pi}{5}$.

It seems a good occasion now for us to learn about trigonometric multiple-angle formulas. In this post, we will show how to derive formulas for $\sin nx$, $\cos nx$, $\tan nx$ and $\cot nx$ using de Moivre’s identity of the complex numbers.

de Moivre’s identity for complex numbers

Complex numbers have the form

$$a + ib \tag{3.25}$$

where a and b are two real numbers; a is called the real part and ib is called the imaginary part. Here are a few examples of complex numbers

3+2 i, 5-3 i, 2 i+7, 8 i, 4 i-2, 5, i+1,. . .

We do addition, subtraction, multiplication and division for complex numbers exactly the same as we do normal arithmetic for real numbers, except that we need to remember

$$i^2 = -1; \quad i^3 = -i; \quad i^4 = -i^2 = 1; \quad i^5 = i; \quad i^6 = i^2 = -1; \ldots \tag{3.26}$$

Addition and Subtraction

$$(a + ib) + (c + id) = (a + c) + i(b + d); \tag{3.27}$$

$$(a + ib) - (c + id) = (a - c) + i(b - d). \tag{3.28}$$

Multiplication

$$(a + ib)(c + id) = ac + iad + ibc + i^2bd = (ac - bd) + i(bc + ad). \tag{3.29}$$

Division we need to use the identity

$$(a + ib)(a - ib) = a^2 - i^2b^2 = a^2 + b^2 \tag{3.30}$$

and so

$$\frac{c + id}{a + ib} = \frac{(c + id)(a - ib)}{(a + ib)(a - ib)} = \frac{(ac + bd) + i(ad - bc)}{a^2 + b^2} = \frac{ac + bd}{a^2 + b^2} + i\frac{ad - bc}{a^2 + b^2}. \tag{3.31}$$

Complex conjugate

$$\overline{a + ib} = a - ib; \quad \overline{a - ib} = a + ib. \tag{3.32}$$

Absolute value

$$|a + ib| = \sqrt{a^2 + b^2}. \tag{3.33}$$

There is a very important fact about complex numbers, that is, any complex number z can be written in a trigonometric form as follows

$$z = |z|(\cos\theta + i\sin\theta). \tag{3.34}$$

This trigonometric form is very convenient when we do multiplication and division. This is because

$$(\cos\theta_1 + i\sin\theta_1)(\cos\theta_2 + i\sin\theta_2) \tag{3.35}$$

$$= (\cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2) + i(\sin\theta_1\cos\theta_2 + \cos\theta_1\sin\theta_2) \tag{3.36}$$

$$= \cos(\theta_1 + \theta_2) + i\sin(\theta_1 + \theta_2). \tag{3.37}$$

So if we have two complex numbers

$$z_1 = |z_1|(\cos\theta_1 + i\sin\theta_1), \tag{3.38}$$

$$z_2 = |z_2|(\cos\theta_2 + i\sin\theta_2), \tag{3.39}$$

then we have the following simple formulas for multiplication and division

$$z_1z_2 = |z_1||z_2|(\cos(\theta_1 + \theta_2) + i\sin(\theta_1 + \theta_2)), \tag{3.40}$$

$$\frac{z_1}{z_2} = \frac{|z_1|}{|z_2|}(\cos(\theta_1 - \theta_2) + i\sin(\theta_1 - \theta_2)). \tag{3.41}$$

In a special case when $\theta_1 = \theta_2$, then we have

$$z^2 = |z|^2(\cos 2\theta + i\sin 2\theta). \tag{3.42}$$

By applying the multiplication formula repeatedly, we derive the following formula

$$z^n = |z|^n(\cos n\theta + i\sin n\theta). \tag{3.43}$$

When $|z| = 1$, we obtain
de Moivre's identity:

$$(\cos\theta + i\sin\theta)^n = \cos n\theta + i\sin n\theta. \tag{3.44}$$

From de Moivre's identity, if we use binomial theorem to expand the left hand side, and compare with the right hand side, we can derive formulas for $\cos n\theta$ and $\sin n\theta$!

Trigonometric multiple-angle formulas

Trigonometric formulas for double angle
Using de Moivre's identity for n=2, we have

$$\cos 2x + i \sin 2x = (\cos x + i \sin x)^2 \tag{3.45}$$

$$= \cos^2 x + 2 \cos x (i \sin x) + (i \sin x)^2 \tag{3.46}$$

$$= \cos^2 x - \sin^2 x + i 2 \sin x \cos x \tag{3.47}$$

We derive the following
formula for $\cos 2x$

$$\cos 2x = \cos^2 x - \sin^2 x = \cos^2 x - (1 - \cos^2 x) = 2 \cos^2 x - 1 \tag{3.48}$$

formula for $\sin 2x$

$$\sin 2x = 2 \sin x \cos x \tag{3.49}$$

formula for $\tan 2x$

$$\tan 2x = \frac{\sin 2x}{\cos 2x} = \frac{2 \sin x \cos x}{\cos^2 x - \sin^2 x} = \frac{2 \sin x \cos x = \cos^2 x}{(\cos^2 x - \sin^2 x) = \cos^2 x} = \frac{2 \tan x}{1 - \tan^2 x} \tag{3.50}$$

formula for $\cot 2x$

$$\cot 2x = \frac{\cos 2x}{\sin 2x} = \frac{\cos^2 x - \sin^2 x}{2 \sin x \cos x} = \frac{(\cos^2 x - \sin^2 x) = \sin^2 x}{2 \sin x \cos x = \sin^2 x} = \frac{\cot^2 x - 1}{2 \cot x} \tag{3.51}$$

Trigonometric formulas for triple angle
Using de Moivre's identity for n=3, we have

$$\cos 3x + i \sin 3x = (\cos x + i \sin x)^3 \tag{3.52}$$

$$= \cos^3 x + 3 \cos^2 x (i \sin x) + 3 \cos x (i \sin x)^2 + (i \sin x)^3 \tag{3.53}$$

$$= \cos^3 x - 3 \cos x \sin^2 x + i (3 \cos^2 x \sin x - \sin^3 x) \tag{3.54}$$

Therefore,
formula for $\cos 3x$

$$\cos 3x = \cos^3 x - 3\cos x \sin^2 x = \cos^3 x - 3\cos x(1 - \cos^2 x) = 4\cos^3 x - 3\cos x \tag{3.55}$$

formula for $\sin 3x$

$$\sin 3x = 3\cos^2 x \sin x - \sin^3 x = 3(1 - \sin^2 x)\sin x - \sin^3 x = 3\sin x - 4\sin^3 x \tag{3.56}$$

formula for $\tan 3x$

$$\tan 3x = \frac{\sin 3x}{\cos 3x} = \frac{3\cos^2 x \sin x - \sin^3 x}{\cos^3 x - 3\cos x \sin^2 x} = \frac{3\tan x - \tan^3 x}{1 - 3\tan^2 x} \tag{3.57}$$

formula for $\cot 3x$

$$\cot 3x = \frac{\cos 3x}{\sin 3x} = \frac{\cos^3 x - 3\cos x \sin^2 x}{3\cos^2 x \sin x - \sin^3 x} = \frac{\cot^3 x - 3\cot x}{3\cot^2 x - 1} \tag{3.58}$$

Trigonometric formulas for quadruple angle
Using de Moivre's identity for n=4, we have

$$\cos 4x + i\sin 4x = (\cos x + i\sin x)^4 \tag{3.59}$$

$$= \cos^4 x + 4\cos^3 x(i\sin x) + 6\cos^2 x(i\sin x)^2 + 4\cos x(i\sin x)^3 + (i\sin x)^4 \tag{3.60}$$

$$= \cos^4 x - 6\cos^2 x \sin^2 x + \sin^4 x + i(4\cos^3 x \sin x - 4\cos x \sin^3 x) \tag{3.61}$$

Therefore,
formula for $\cos 4x$

$$\cos 4x = \cos^4 x - 6\cos^2 x \sin^2 x + \sin^4 x = \cos^4 x - 6\cos^2 x(1 - \cos^2 x) + (1 - \cos^2 x)^2 \tag{3.62}$$

$$= 8\cos^4 x - 8\cos^2 x + 1 \tag{3.63}$$

formula for $\sin 4x$

$$\sin 4x = 4\cos^3 x \sin x - 4\cos x \sin^3 x = 4\cos x(1 - \sin^2 x)\sin x - 4\cos x \sin^3 x \tag{3.64}$$

$$= 4\cos x(\sin x - 2\sin^3 x) \tag{3.65}$$

formula for $\tan 4x$

$$\tan 4x = \frac{\sin 4x}{\cos 4x} = \frac{4\cos^3 x \sin x - 4\cos x \sin^3 x}{\cos^4 x - 6\cos^2 x \sin^2 x + \sin^4 x} = \frac{4\tan x - 4\tan^3 x}{1 - 6\tan^2 x + \tan^4 x} \tag{3.66}$$

formula for $\cot 4x$

$$\cot 4x = \frac{\cos 4x}{\sin 4x} = \frac{\cos^4 x - 6\cos^2 x \sin^2 x + \sin^4 x}{4\cos^3 x \sin x - 4\cos x \sin^3 x} = \frac{\cot^4 x - 6\cot^2 x + 1}{4\cot^3 x - 4\cot x} \tag{3.67}$$

By now, you probably know how to derive the general formulas. Please continue with the cases for n=5,6,7. Using these formulas, it is possible to calculate the explicit values of $\sin\frac{\pi}{7}$, $\cos\frac{\pi}{7}$, $\sin\frac{\pi}{9}$, $\cos\frac{\pi}{9}$, ...

Let us stop here for now. In the next post, we will return to the topic of compass-and-straightedge construction. Hope to see you again then.

Homework.

1. Continue with n=5,6,7,... and establish formulas for $\sin nx$, $\cos nx$, $\tan nx$, $\cot nx$.

2. Find explicit formulas for

$$\sin\frac{\pi}{7}; \sin\frac{2\pi}{7}; \sin\frac{3\pi}{7}; \ldots \tag{3.68}$$

$$\cos\frac{\pi}{7}; \cos\frac{2\pi}{7}; \cos\frac{3\pi}{7}; \ldots \tag{3.69}$$

3. Find explicit formulas for

$$\sin\frac{\pi}{9}; \sin\frac{2\pi}{9}; \sin\frac{3\pi}{9}; \ldots \tag{3.70}$$

$$\cos\frac{\pi}{9}; \cos\frac{2\pi}{9}; \cos\frac{3\pi}{9}; \ldots \tag{3.71}$$

4. Search on Google http://google.com about Chebyshev polynomial, find out the relationship between Chebyshev polynomial and the trigonometric multiple-angle formulas.

Source:
http://mathgardenblog.blogspot.com/2014/02/ moivre-trigonometric-multiple-angle-formula.html

3.3 Power of a point to a circle

Today we will learn about the power of a point with respect to a circle.

Suppose on a plane we have a point P and a circle (O). Draw a line through P which intersects with the circle at two points U and V. Then the value of

$$PU \quad PV \tag{3.72}$$

is independent of the choice of the line PUV.

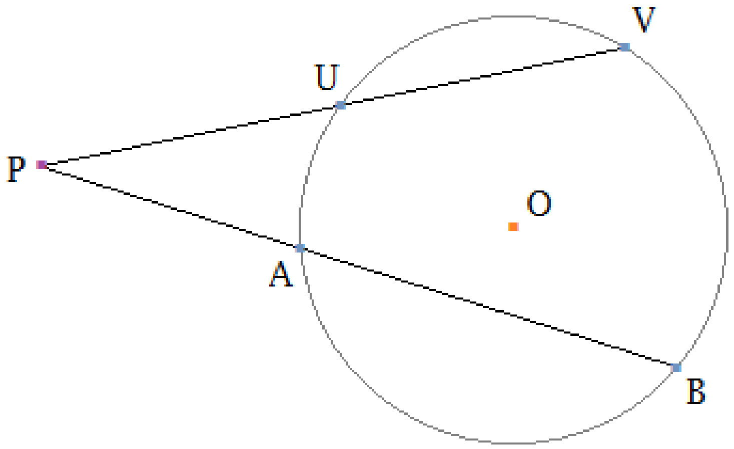

Figure 3.6: ppc1.png

This means that if we draw another line through P which cuts the circle at two other points A and B then

$$PA \quad PB = PU \quad PV: \tag{3.73}$$

This constant value is called the power of the point P with respect to the circle (O).

To prove that PU X PV is independent of the position of the line, we will use similar triangles. We will consider two separate cases: when the point P is outside the circle and when the point P is inside the circle.

When P is outside the circle (O)

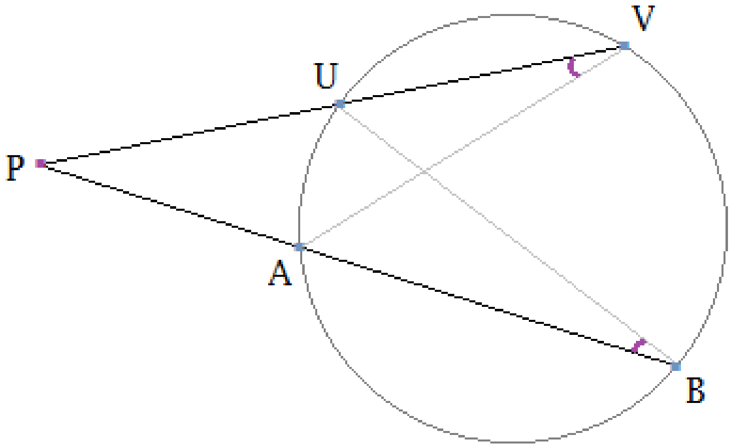

Figure 3.7: ppc2.png

We consider two triangles PUB and PAV. These two triangles are similar triangles because they have two pairs of equal angles. Thus,

$$\frac{PU}{PA} = \frac{PB}{PV}: \tag{3.74}$$

From here we obtain the required equality

$$PA \quad PB = PU \quad PV: \tag{3.75}$$

When P is inside the circle (O)

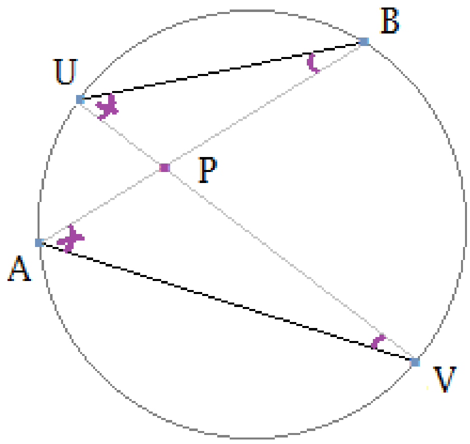

Figure 3.8: ppc3.png

When P is inside the circle, we also consider the two triangles PUB and PAV. These two triangles are similar triangles because they have equal angles. So

$$\frac{PU}{PA} = \frac{PB}{PV}: \tag{3.76}$$

This implies the constancy of the power of the point P:

$$PA \quad PB = PU \quad PV: \tag{3.77}$$

Expressing power of the point P in terms of the distance PO and the circle radius r

We will now show how to derive a formula for the power of the point P in terms of the distance PO and the radius r of the circle (O). We again consider two separate cases: when the point P is outside the circle and when the point P is inside the circle.

We will prove that:

If P is outside the circle then

$$PU \quad PV = PO^2 \quad r^2: \tag{3.78}$$

If P is inside the circle then

$$PU \quad PV = r^2 \quad PO^2: \tag{3.79}$$

Indeed, if we take U and V as two intersection points of the line PO with the circle (O) then the power of the point P can be calculated as follows:

When P is outside the circle

$$PU \quad PV = (PO \quad r)(PO + r) = PO^2 \quad r^2: \qquad (3.80)$$

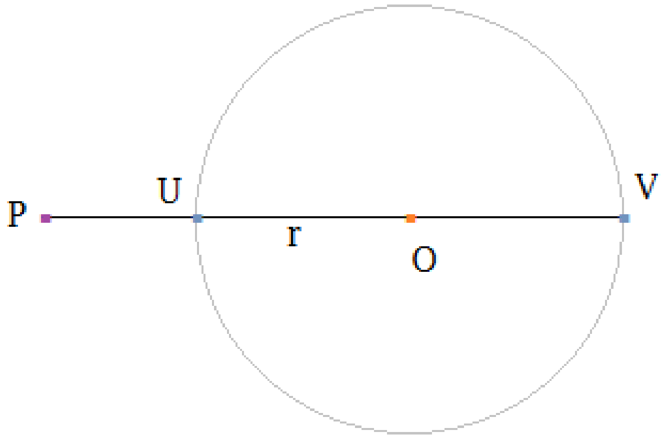

Figure 3.9: ppc4.png

When P is inside the circle

$$PU \quad PV = (r \quad PO)(r + PO) = r^2 \quad PO^2: \qquad (3.81)$$

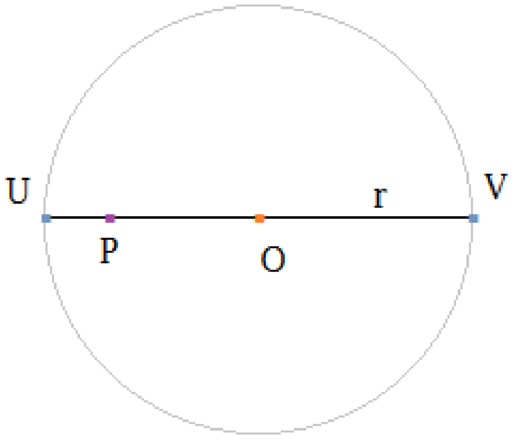

Figure 3.10: ppc5.png

So we have shown that the power of the point P can be expressed in terms of the distance PO and the radius r as

$$PU \cdot PV = \pm(PO^2 - r^2): \tag{3.82}$$

Whether the sign is a plus or a minus in the above formula depends on the position of the point P with respect to the circle. It is a plus if P is outside the circle and it is a minus if P is inside the circle. (For the special case when the point P lies on the circle then the power of P is equal to zero trivially.)

The "correct" definition of the power of a point

Actually our above definition of the power of a point is not really correct. The above definition does not concern about the sign of PU X PV making it always a positive number.

$$\vec{PU} \cdot \vec{PV} = PO^2 - r^2: \tag{3.83}$$

In this definition, the value of the power of the point P can be positive or negative, it depends on the two vectors PU -> and PV -> point to the same direction or opposite directions.

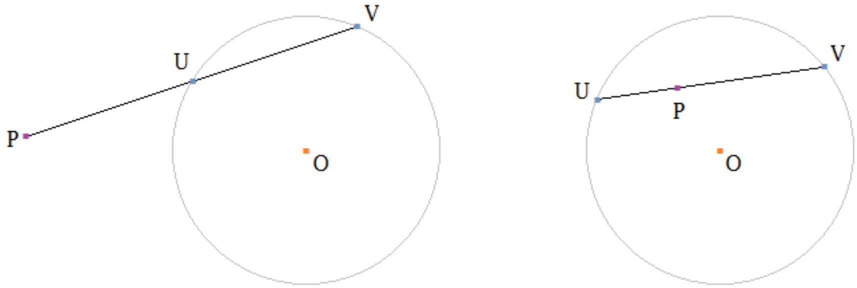

Figure 3.11: ppc6.png

When P is outside the circle, the power of P is a positive number because PU -> and PV -> have the same direction. When P is inside the circle, the power of P is a negative number because PU ->and PV -> point to opposite directions.

The power of a point is a useful bit of information that helps us to know the relative position of the point with respect to the circle. If this value is a positive number then we know that the point is outside of the circle. If this value is zero then the point must lie on the circle. And if this value is negative then we know that the point is inside the circle.

Formula of the power of a point in Cartesian coordinate

In the Cartesian coordinate system 0xy, each point P on the plane is assigned to a unique coordinate (Px,Py).

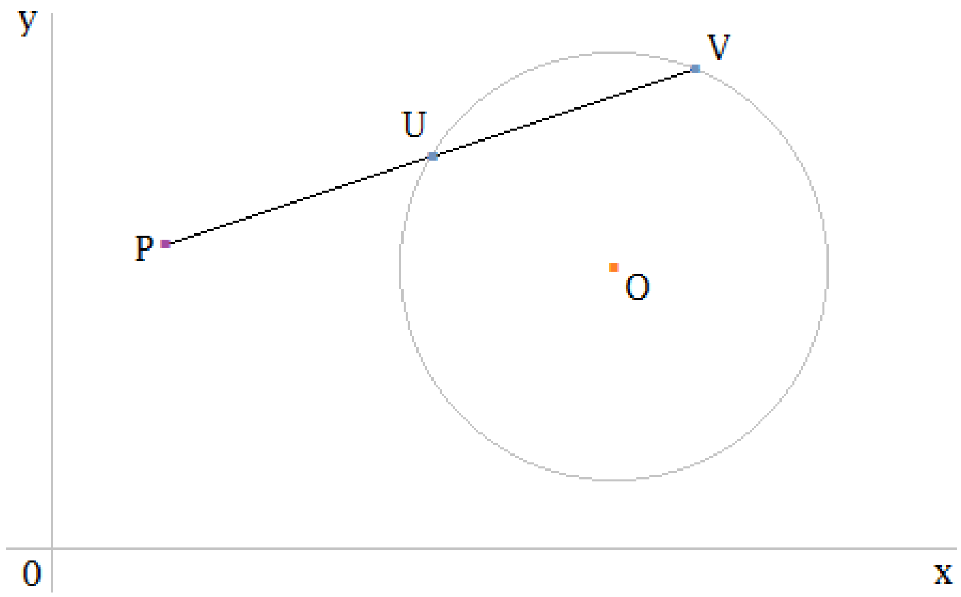

Figure 3.12: ppc7.png

We know that the equation of the circle with center O and radius r is

$$(x - O_x)^2 + (y - O_y)^2 - r^2 = 0: \tag{3.84}$$

The formula for the power of a point is ultimately related to the above equation of the circle

$$\overrightarrow{PU} \cdot \overrightarrow{PV} = PO^2 - r^2 = (P_x - O_x)^2 + (P_y - O_y)^2 - r^2: \tag{3.85}$$

This formula

$$(P_x - O_x)^2 + (P_y - O_y)^2 - r^2 \tag{3.86}$$

signifies the reason why the value of the power of a point is positive, zero, and negative when the point is outside the circle, on the circle, and inside the circle.

Today we have learned some basic facts about the power of a point to a circle. In the next post, we will learn more about its properties and its many applications.

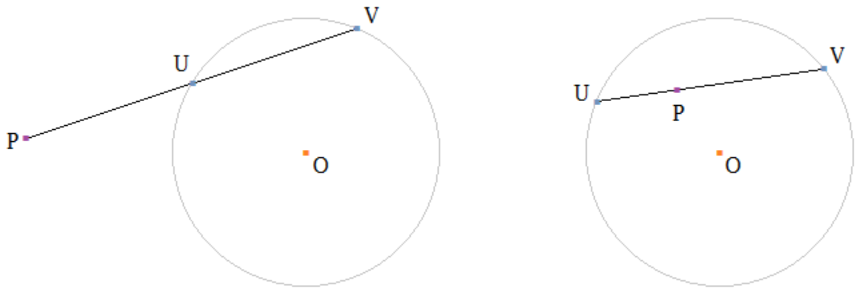

Figure 3.13: ppc8.png

Power of a point:

$$\vec{PU} \quad \vec{PV} = PO^2 \quad r^2 = (P_x \quad O_x)^2 + (P_y \quad O_y)^2 \quad r^2 \tag{3.87}$$

Before we conclude the post, let us summarize the three basic formulas for the power of a point:

$$\vec{PU} \quad \vec{PV} = PO^2 \quad r^2 = (P_x \quad O_x)^2 + (P_y \quad O_y)^2 \quad r^2: \tag{3.88}$$

Homework.

1. Suppose that the point P is outside the circle (O). Draw a tangent PT to the circle. Apply Pythagorean Theorem to show that

$$PU \quad PV = PT^2 = PO^2 \quad r^2: \tag{3.89}$$

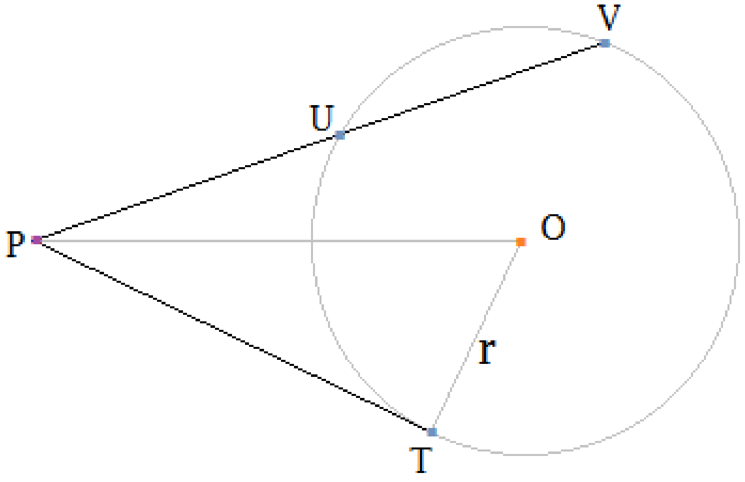

Figure 3.14: ppc9.png

2. Suppose that the point P is inside the circle (O). Draw the line UV perpendicular to PO. Again, apply Pythagorean Theorem to show that

$$PU \quad PV = PU^2 = r^2 \quad PO^2: \qquad (3.90)$$

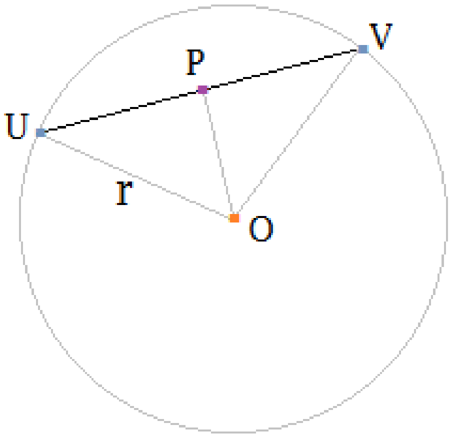

Figure 3.15: ppc10.png

3. Given two circles (O1) and (O2) intersecting at two points I and J. Show that for any point P on the line IJ, P has equal power with respect to

these two circles.

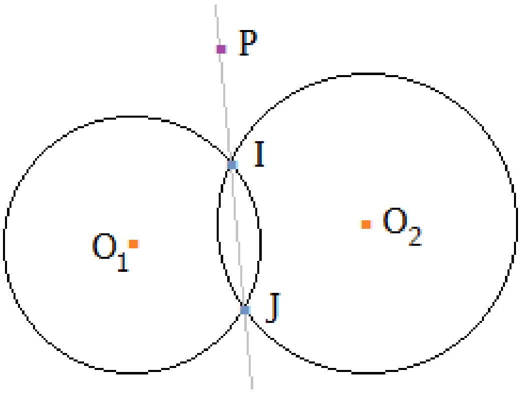

Figure 3.16: ppc11.png

4. Given two circles (O1) and (O2). Determine the locus of all the points that have equal power to these two circles.

5. Given three circles (O1), (O2) and (O3). Determine the point P such that it has equal power to all of these circles.

Source:
http://mathgardenblog.blogspot.com/2013/07/power-of-a-point.html

3.4 Radian

In our normal life, when we talk about angles, we often use the degree. For example, a right angle is equal to 90 degrees, an angle of an equilateral triangle is 60 degrees, a straight angle is 180o. However, for all trigonometric functions in mathematics, such as sin(x), cos(x), etc., the angle x is always understood as in radian.

So what is the radian?

To understand angle measurement in radians, we draw a unit circle. Unit circle is a circle with radius equal to 1. We know that by definition,   is the length of the semicircle.

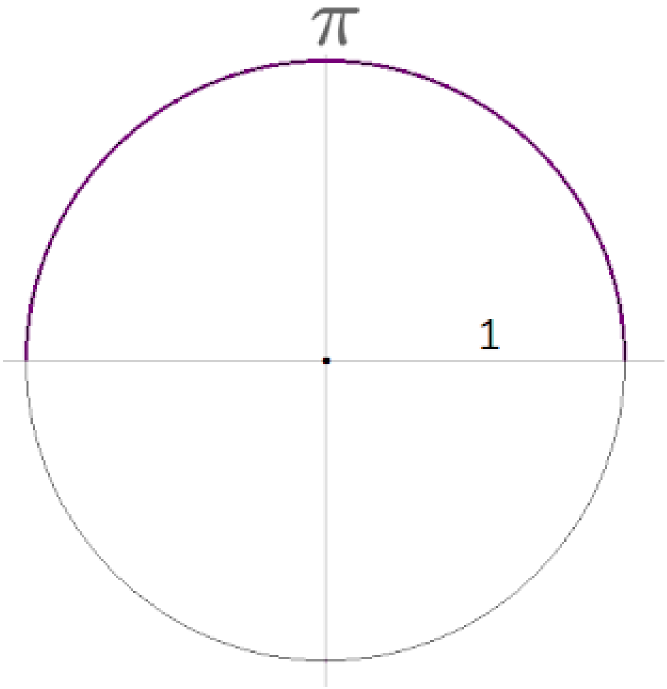

Figure 3.17: radian1.png

An angle’s measurement in radians is equal to the length of the corresponding arc of the unit circle.

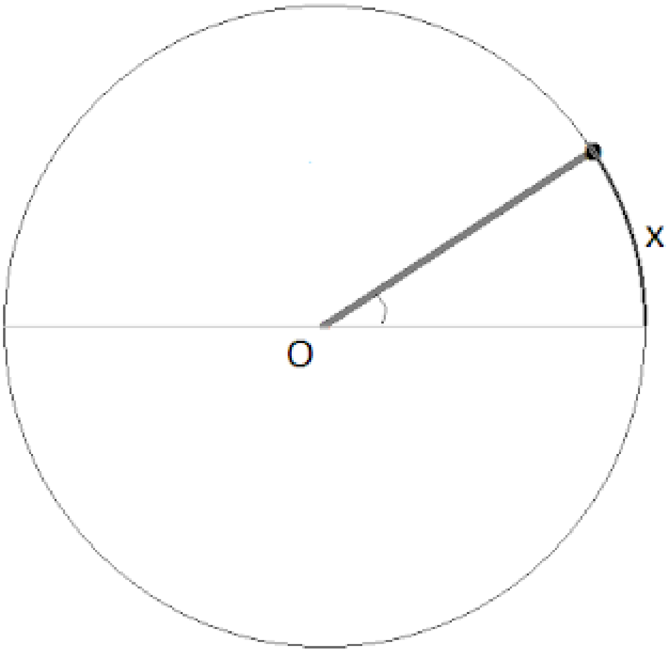

Figure 3.18: radian2.png

In radians, x is the length of the arc

Let us look at an example of a right angle. A right angle is subtended by a quarter of the circle, and the length of a quarter of the circle is $\frac{}{2}$, so a right angle is equal to $\frac{}{2}$ in radians.

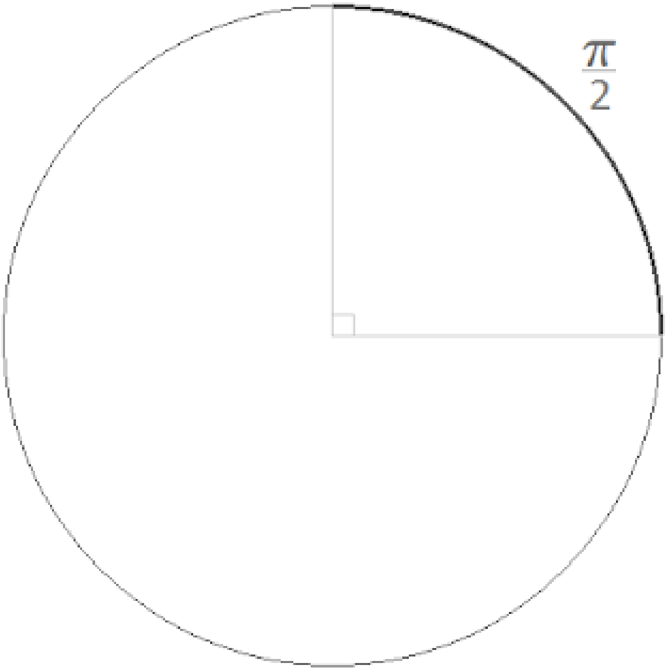

Figure 3.19: radian3.png

A straight angle (180o) occupies half of the circle, and the length of a half of the circle is , so a straight angle is equal to in radians.

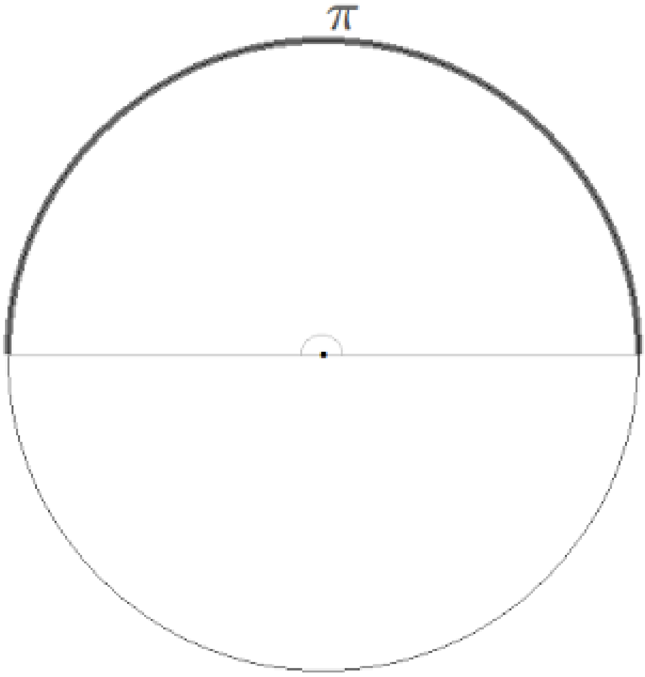

Figure 3.20: radian4.png

We can easily remember the translation from degrees to radians by the following correspondence

straight angle 180 degrees -> semicircle ->

Common angles are

$$180^o \; ! \tag{3.91}$$

$$360^o \; ! \; 2 \tag{3.92}$$

$$90^o \; ! \; \frac{}{2} \tag{3.93}$$

$$45^o \; ! \; \frac{}{4} \tag{3.94}$$

$$60^o \; ! \; \frac{}{3} \tag{3.95}$$

$$30^o \; ! \; \frac{}{6} \tag{3.96}$$

Let us stop here for now. We will return to our series on mathematical identities in our next post. Hope to see you there.

Homework.

In this homework section, we will prove the Vieta formula of the number that we mention in the previous post

$$\frac{2}{\pi} = \sqrt{\frac{1}{2}}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}}}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}}}} \qquad (3.97)$$

In the following figure, we can see that $ZA = \sin(x)$ is a line segment which is shorter than the circle arc $ZI = x$

$$\sin(x) < x \qquad (3.98)$$

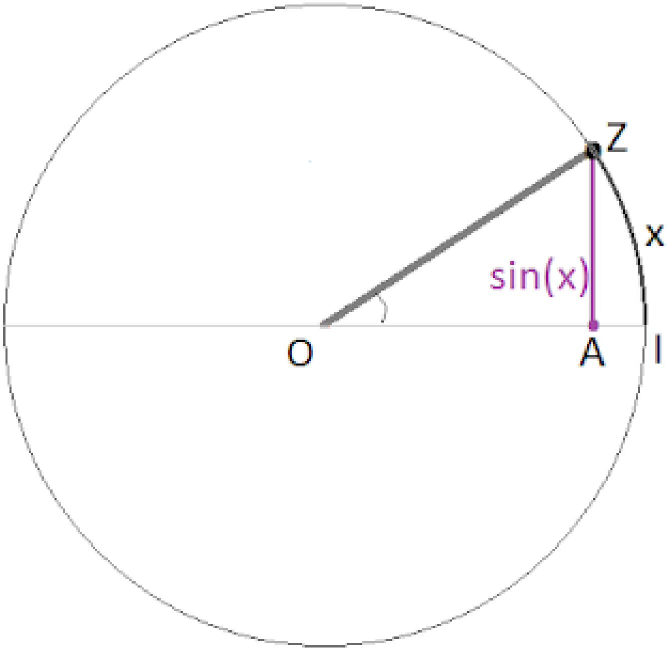

Figure 3.21: radian5.png

Especially, when the angle x is small, $\sin(x)$ is getting closer and closer to x. We will use this fact to prove the Vieta formula.

1. Use the trigonometric cos formula for double angle

$$\cos(2x) = 2\cos^2(x) - 1 \qquad (3.99)$$

to show that

$$\cos\frac{\pi}{4} = \sqrt{\frac{1}{2}} \tag{3.100}$$

$$\cos\frac{\pi}{8} = \sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2}}} \tag{3.101}$$

$$\cos\frac{\pi}{16} = \sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2}}}} \tag{3.102}$$

Thus,

$$\sqrt{\frac{1}{2}} \times \sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2}}} \times \sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2}}}} = \cos\frac{\pi}{4} \times \cos\frac{\pi}{8} \times \cos\frac{\pi}{16} \tag{3.103}$$

2. Use the trigonometric sin formula for double angle

$$\sin(2x) = 2\sin(x)\cos(x) \tag{3.104}$$

to show that

$$\cos\frac{\pi}{4} \times \cos\frac{\pi}{8} \times \cos\frac{\pi}{16} = \frac{\frac{1}{8}}{\sin\frac{\pi}{16}} = \frac{2}{\pi} \times \frac{\frac{\pi}{16}}{\sin\frac{\pi}{16}} \tag{3.105}$$

3. As we have discussed above, since the angle $\frac{\pi}{16}$ is very small,

$$\sin\frac{\pi}{16} \approx \frac{\pi}{16} \tag{3.106}$$

and

$$\cos\frac{\pi}{4} \times \cos\frac{\pi}{8} \times \cos\frac{\pi}{16} \approx \frac{2}{\pi} \tag{3.107}$$

4. In general, prove that

$$\cos\frac{\pi}{4} \times \cos\frac{\pi}{8} \times \cdots \times \cos\frac{\pi}{2^n} = \frac{2}{\pi} \times \frac{\frac{\pi}{2^n}}{\sin\frac{\pi}{2^n}} \tag{3.108}$$

and

$$\lim_{n\to\infty} \cos\frac{\pi}{4} \times \cos\frac{\pi}{8} \times \cdots \times \cos\frac{\pi}{2^n} = \frac{2}{\pi} \tag{3.109}$$

This is the Vieta formula of the number

$$\sqrt{\frac{1}{2}}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}}}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}}}} = \frac{2}{\ } \qquad (3.110)$$

Source:
http://mathgardenblog.blogspot.com/2016/03/radian.html

3.5 How to construct a regular polygon with 15 sides

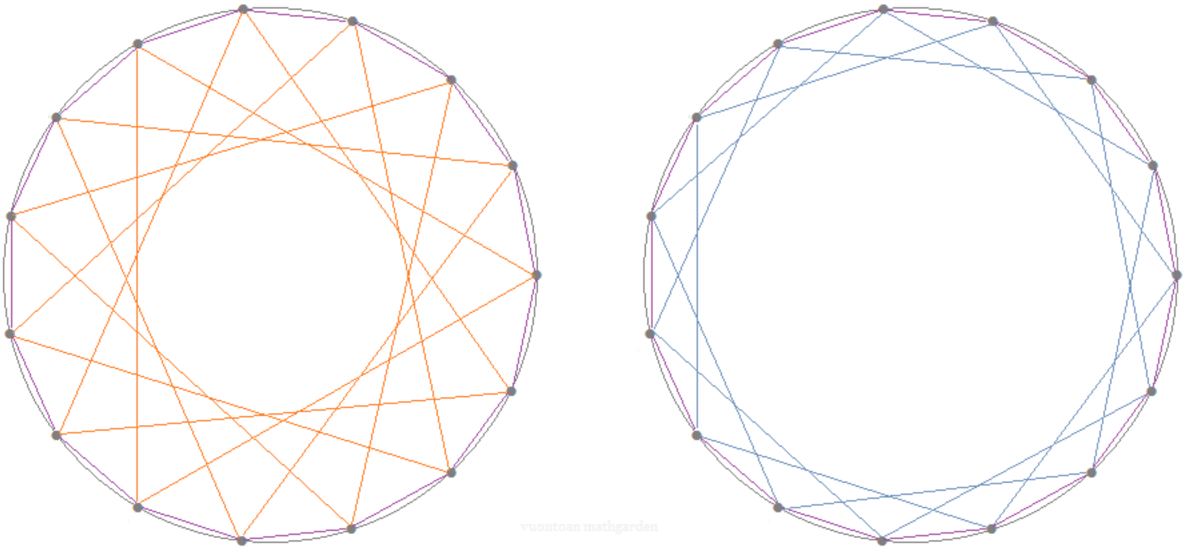

Figure 3.22: polygon1.png

To feed your curiosity, today we will look at a compass and straightedge construction of the regular polygon with 15 sides and we will show that there is a connection between this construction problem with the measuring liquid puzzle that we have learned from our previous post.

Measuring liquid puzzle

First, let us review the following brain teaser puzzle from our previous post.

Measuring liquid puzzle. Suppose you have a 3-liter jug and a 5-liter jug, how would you measure out exactly 1 liter of water?

At first glance, it seems that there are a lot of possibilities with the filling and pouring of water using the two jugs. But if we take a step back and have a careful look, we can categorize all the measuring operations into exactly 8 types:

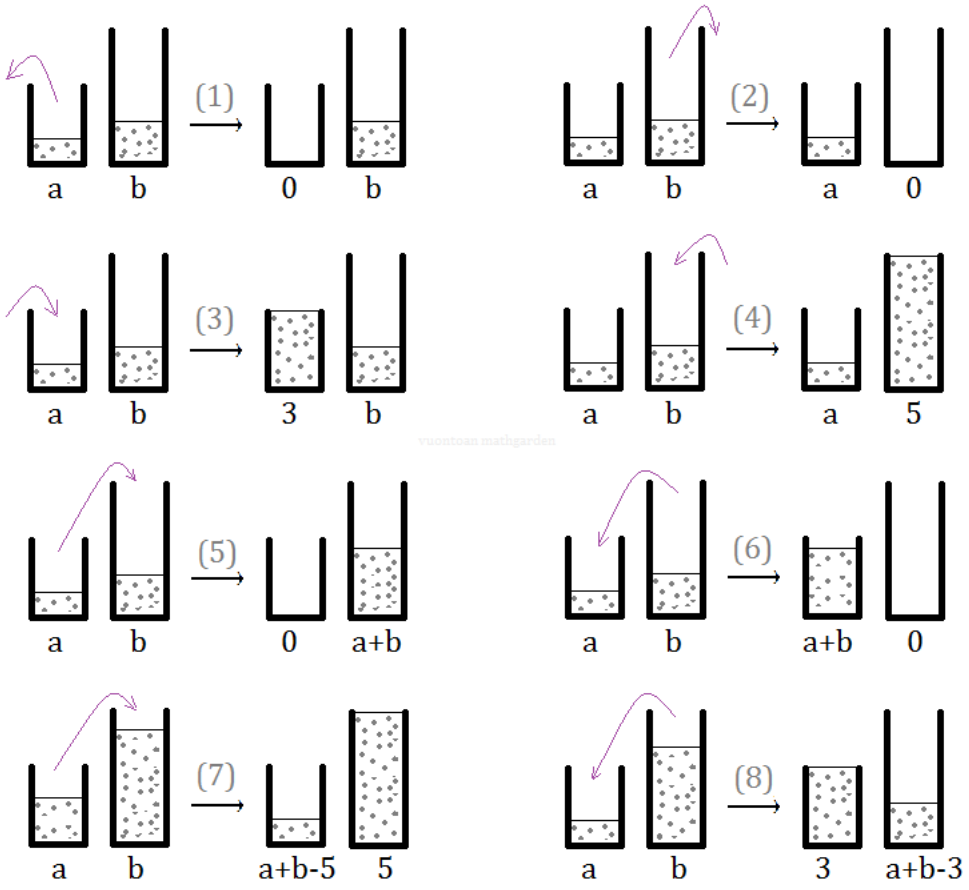

Figure 3.23: polygon2.png

We can see that the amount of water contained in the jugs will initially start with the value (0,0), and after each step of measuring operation, this value (a,b) will change into one of the following 8 possible values

$$(0; b); \quad (a; 0); \quad (3; b); \quad (a; 5); \tag{3.111}$$

$$(0; a+b); \quad (a+b; 0); \quad (a+b-5; 5); \quad (3; a+b-3): \tag{3.112}$$

Therefore, we can easily prove the following fact

The amount of water (in liter) in each of the jug is always a number of the form 3x+5y where x and y are two integers.

It means that if we want to measure out 1 liter of water then we have to write the value 1 in the form 3x+5y. So our puzzle is reduced to the problem of solving the following linear Diophantine equation

$$3x + 5y = 1: \tag{3.113}$$

This equation has infinitely many integer solutions. We can easily see that one solution is (x=2,y=-1) and another one is (x=-3,y=2). These two solutions give us the following two representations of the value 1:

$$3 \quad 2 \quad 5 \quad 1 = 1 \tag{3.114}$$

$$5 \quad 2 \quad 3 \quad 3 = 1: \tag{3.115}$$

Corresponding to these two representations, we have the following answers to our puzzle:

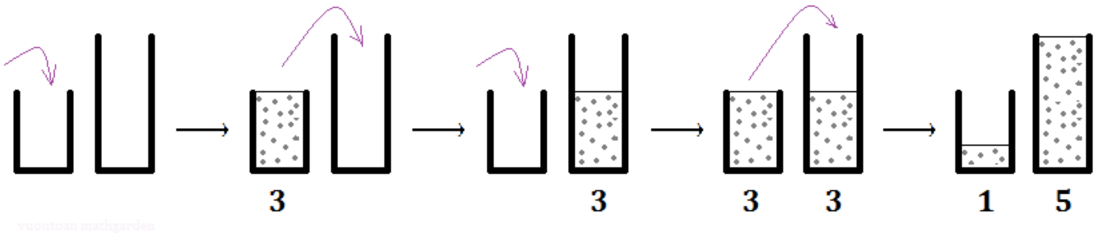

Figure 3.24: polygon3.png

First solution 3 X 2 - 5 X 1=1: Filling 3-liter jug to the full twice and pouring out to 5-liter jug making it full once, we get 1 liter remaining.

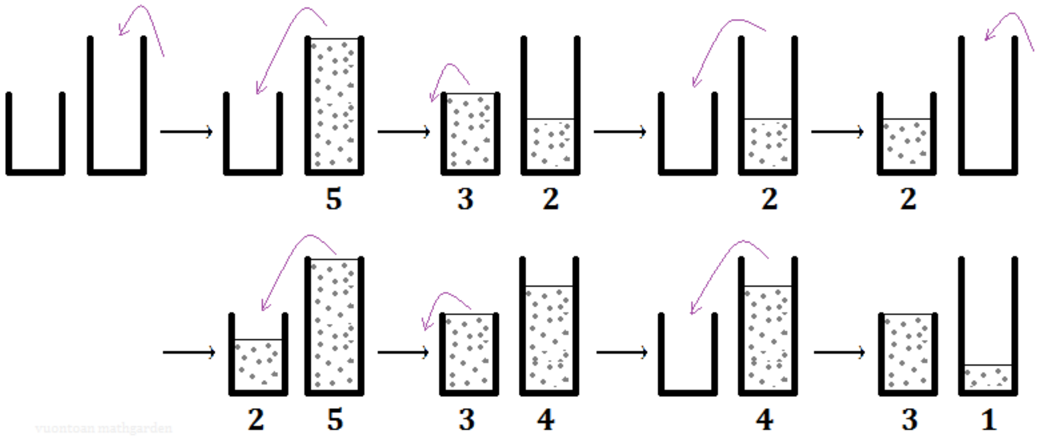

Figure 3.25: polygon4.png

Second solution 5 X 2 - 3 X 3=1: Filling 5-liter jug to the full twice and pouring out to 3-liter jug making it full three times, we get 1 liter remaining.

We have seen that our measuring liquid puzzle has a close connection with the Diophantine equation 3x+5y=1. Soon, we will see that this same Diophantine equation enables us to construct a regular polygon with 15 sides.

Regular polygon construction problem

Let us call a polygon with n sides an n-gon. A polygon with equal sides and equal angles is called a regular polygon.

The problem of constructing regular polygons by compass and straightedge is one of a few classic problems of ancient mathematics that are easy to state but incredibly difficult to solve. For a long time, we know how to construct an equilateral triangle (3-gon) and a regular pentagon (5-gon). In his famous books, the Elements, Euclid of ancient Greek presented a method of constructing a regular pentagon. But for the next two thousand years, no one had ever succeeded in constructing a regular 7-gon, a regular 9-gon, a regular 11-gon or a regular 13-gon. It was not until in 1796, Gauss, only at age 19, made the first breakthrough by successfully constructing a regular 17-gon. The problem was completely solved in 1837 with the following beautiful theorem

Gauss-Wantzel's Theorem. For an odd number n, a regular polygon of n sides can be constructed by compass and straightedge if and only if $n = p_1 \cdots p_t$ are distinct Fermat's primes.

So, what are the Fermat's primes mentioned in Gauss' theorem? Here is the definition:

Fermat's prime. A prime number is called a Fermat's prime if it has the form

$$F_k = 2^{2^k} + 1: \tag{3.116}$$

According to this definition:

k=0, $F_0 = 2^{2^0} + 1 = 2^1 + 1 = 3$ is a prime, so it is a Fermat's prime,

k=1, $F_1 = 2^{2^1} + 1 = 2^2 + 1 = 5$ is a prime, so it is a Fermat's prime,

k=2, $F_2 = 2^{2^2} + 1 = 2^4 + 1 = 17$ is a prime, so it is a Fermat's prime,

k=3, $F_3 = 2^{2^3} + 1 = 2^8 + 1 = 257$ is a prime, so it is a Fermat's prime,

k=4, $F_4 = 2^{2^4} + 1 = 2^{16} + 1 = 65537$ is a prime, so it is a Fermat's prime,

k=5, $F_5 = 2^{2^5} + 1 = 2^{32} + 1 = 4294967297 = 641 \times 6700417$ is a composite number.

Since 3 and 5 are Fermat's primes, by Gauss-Wantzel's Theorem, regular 3-gons and regular 5-gons are constructible. In addition, since 15=3 X 5, regular 15-gons are constructible. However, there is a condition in the theorem saying that the Fermat's primes have to be distinct, so 9-gon and 25-gon are not constructible by compass and straightedge.

Construction of a regular 15-gon

As we have seen above, 15 is a product of two distinct Fermat's primes (3 and 5), so by Gauss-Wantzel's Theorem, a regular 15-gon can be constructed by compass and straightedge. But how can we construct it?

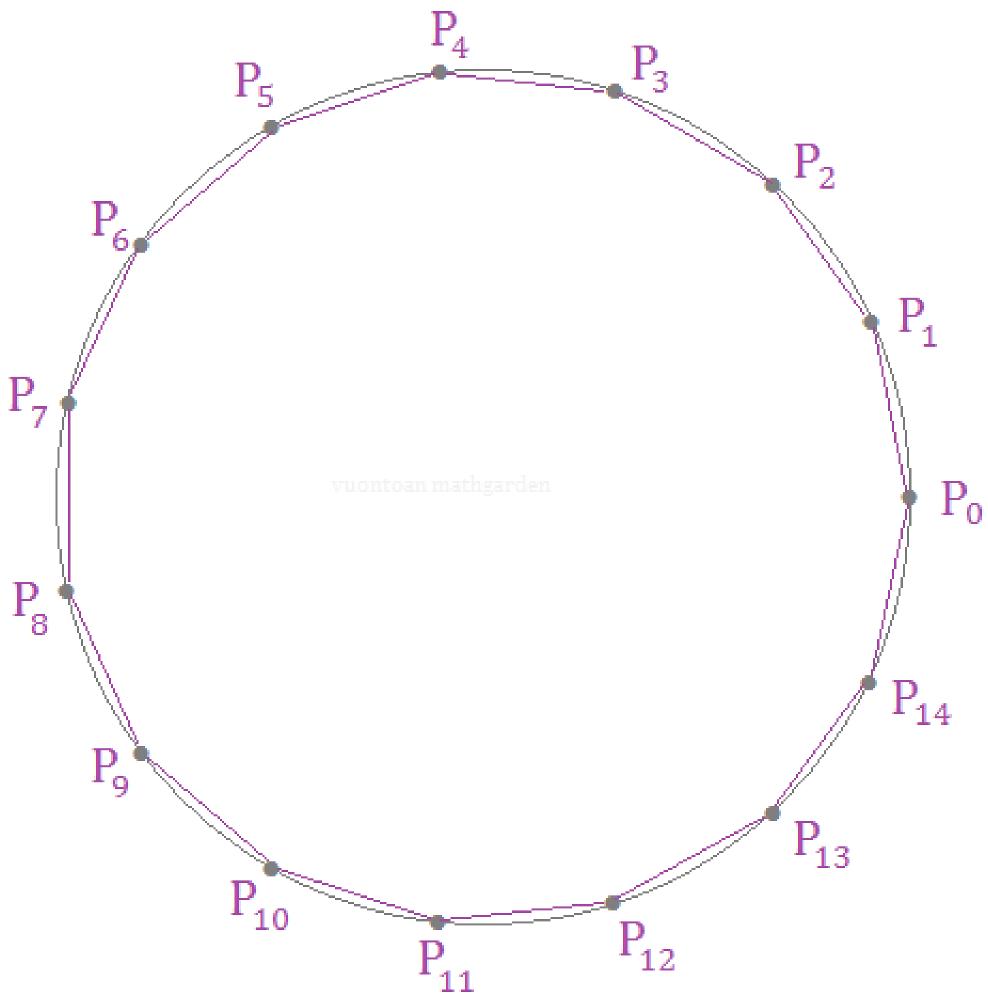
P_4
P_3
P_5
P_2
P_6
P_1
P_7
P_0
vuontoan mathgarden
P_8
P_{14}
P_9
P_{13}
P_{10}
P_{12}
P_{11}

Figure 3.26: polygon5.png

Let us look at a picture of a regular 15-gon. We can see that its vertices form a lot of equilateral triangles and a lot of regular pentagons.

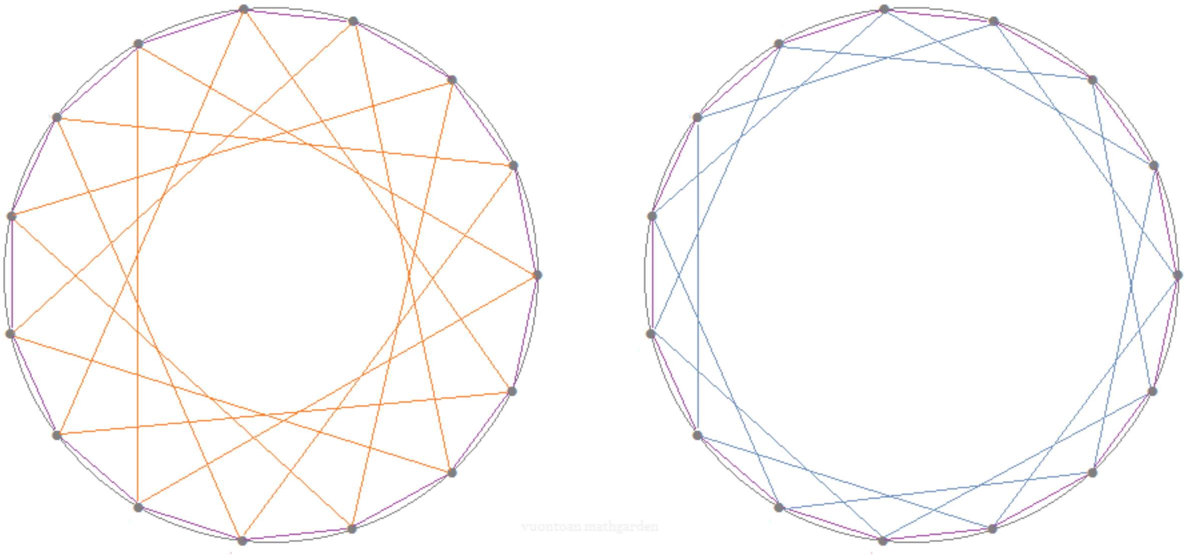

Figure 3.27: polygon6.png

This observation motivates us to construct first an equilateral triangle and a regular pentagon. Let us draw a circle, and pick a point P0 on it. We can construct an equilateral triangle A0A1A2 and a regular pentagon B0B1B2B3B4 with A0=B0=P0. Now, even though we have not constructed a regular 15-gon fully, but at least we have a part of it!

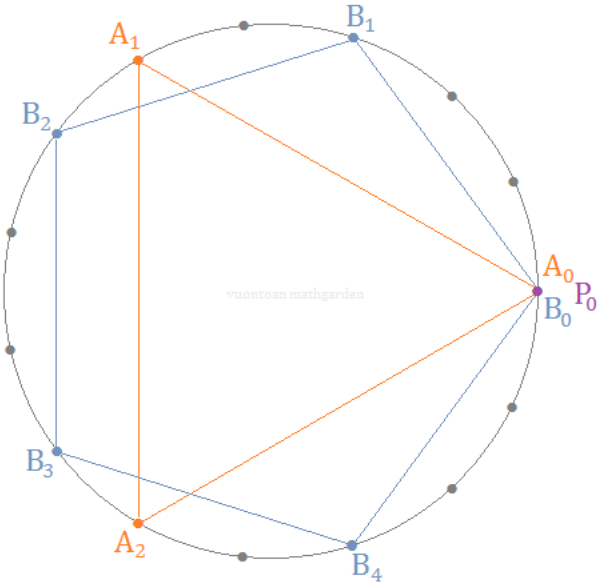

Figure 3.28: polygon7.png

Can we construct other vertices of the 15-gon? YES WE CAN!!!

Looking at the above picture, we can see that we can actually construct the whole polygon because we have already had the two sides A1B2 and B3A2. Using compass with radius A1B2 (or B3A2), from P0 we can construct P1, from P1 we can construct P2, ..., and the rest of it!

Let us now write down the steps of the construction.

Construct a circle and pick a point P0 on it. Construct an equilateral triangle P0P5P10 and a regular pentagon P0P3P6P9P12. Connect the sides P5P6 and P9P10. Draw a circle with center P0 and radius equal to P5P6, this circle intersects with the original big circle at P1 and P14. Draw a circle with center P3 and radius equal to P5P6, this circle intersects with the original big circle at P2 and P4. Other vertices P7, P8, P11, P13 can be constructed similarly.

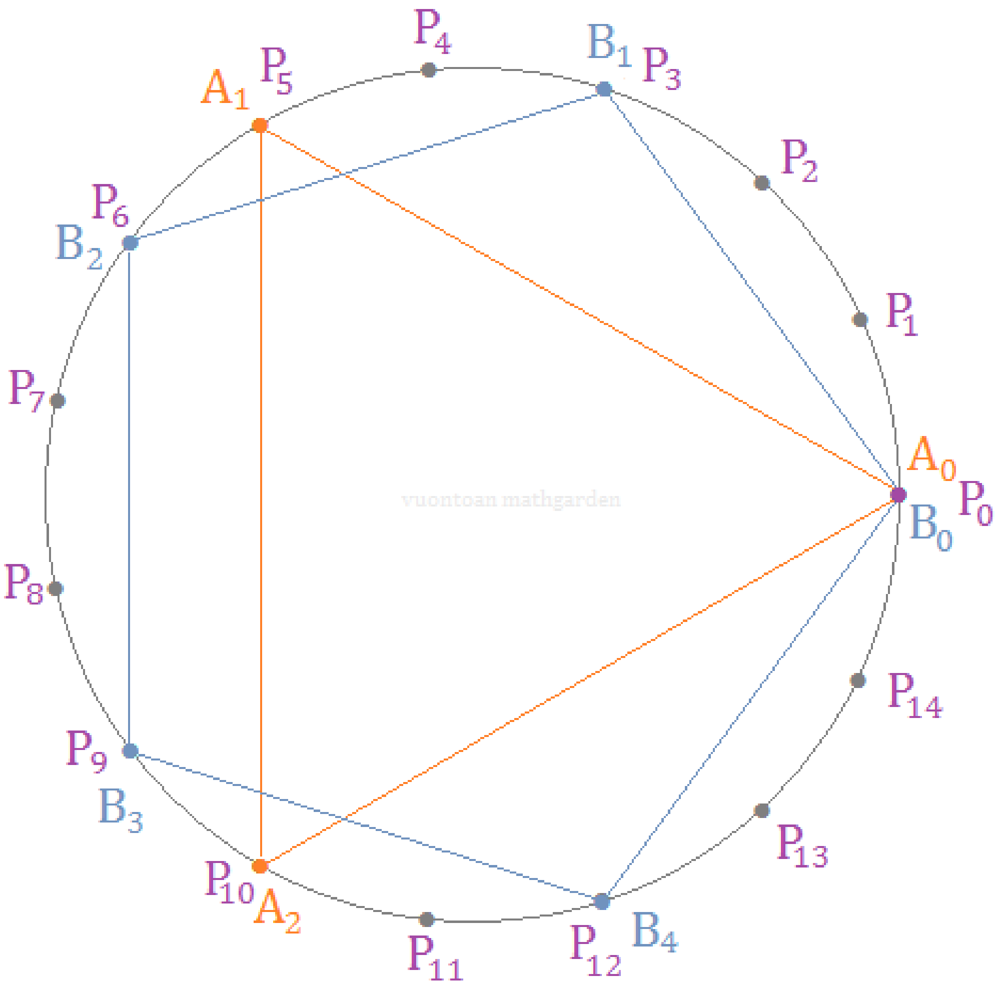
P_4
B_1
P_3
A_1
P_5
P_2
P_6
B_2
P_1
P_7
A_0
P_0
B_0
vuontoan mathgarden
P_8
P_{14}
P_9
B_3
P_{13}
P_{10}
A_2
P_{11}
P_{12}
B_4

Figure 3.29: polygon8.png

Now we know how to construct a regular 15-gon. Let us take a step back and see why this construction is possible. The crucial part of the construction is the fact that A1 and B2 are next to each other, and also B3 and A2. This fact can be explained by the Diophantine equation $3x + 5y = 1$

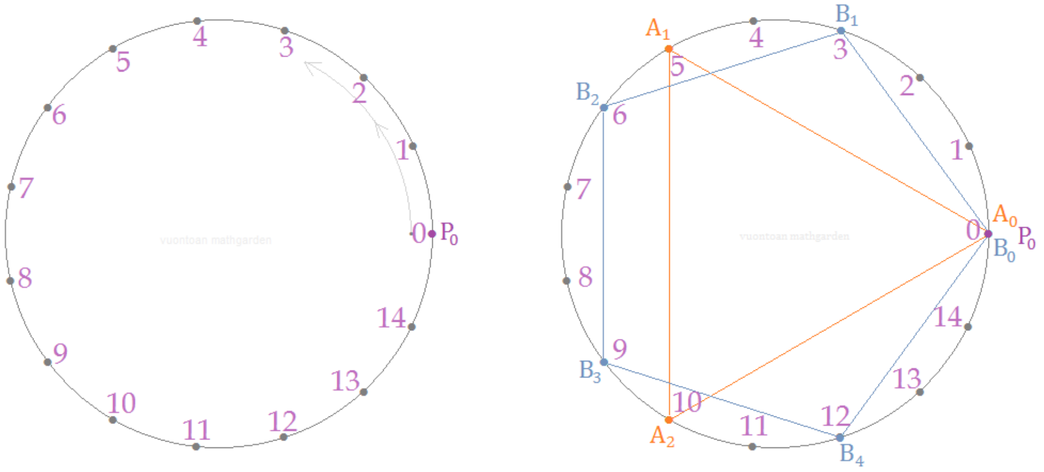

Figure 3.30: polygon9.png

Indeed, if we consider the whole circle as 15 units then each side of the equilateral triangle contains 5 units and each side of the regular pentagon contains 3 units. If we use the point P0 as the origin and measure the distance along the circle, then the point A1 is at 5 unit, the point A2 is at 10 unit. For the regular pentagon, the point B1 is at 3 unit, B2 6 unit, B3 9 unit, and B4 12 unit.

In general, the point Ai is at 5i unit and the point Bj is at 3j unit. The distance between Ai and Bj along the circle is

$$A_i B_j = |5i - 3j|. \qquad (3.117)$$

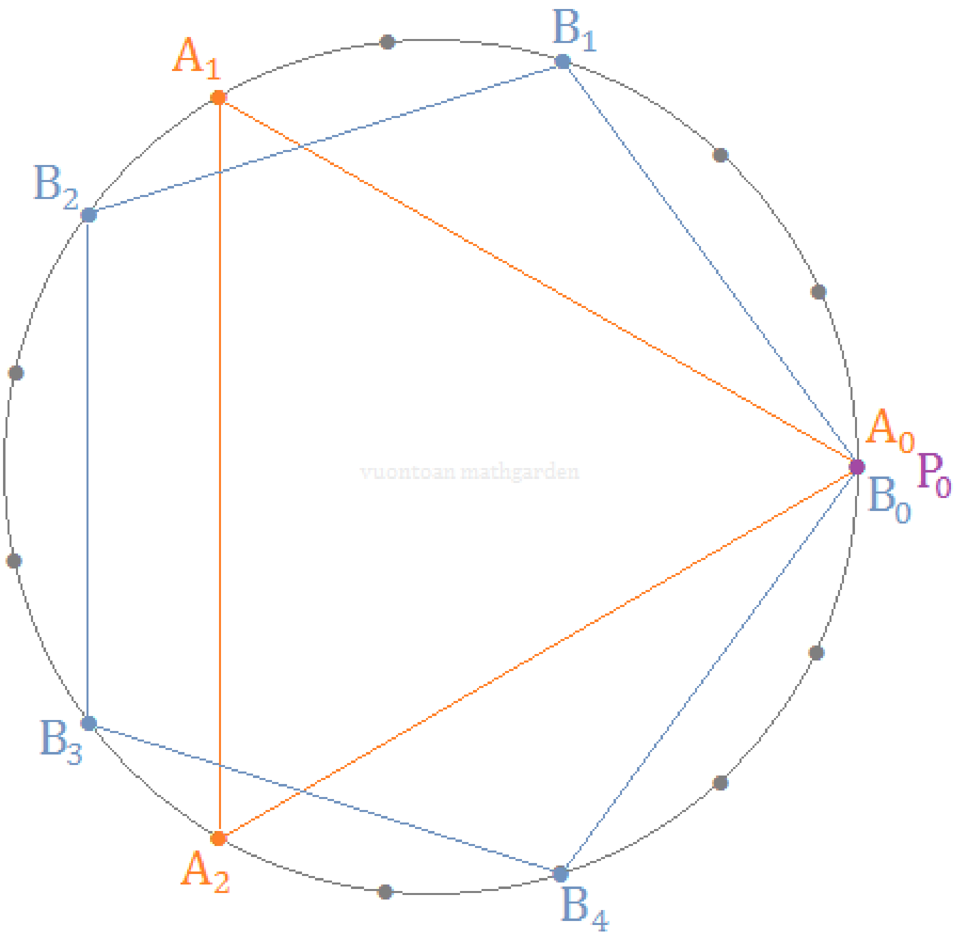

Figure 3.31: polygon10.png

Ai is at 5i unit, Bj is at 3j unit, so the distance between Ai and Bj is $A_iB_j = j5i \quad 3jj$

According to this distance formula, Ai is next to Bj if and only if the distance between Ai and Bj is 1 unit, that is

$$5i \quad 3j = \quad 1 \tag{3.118}$$

One of the solution is i=2, j=3, this gives us 5 X 2 - 3 X 3=1

this explains why A2 and B3 are next to each other and B3A2 forms a side of the 15-gon.

Another solution is i=1, j=2, 5 X 1 - 3 X 2= -1 and this explains why A1 is next to B2.

So the two sides B3A2 and A1B2 of the regular 15-gon correspond to the two integer solutions of the Diophantine equation 3x+5y=1. How amazing is that!

From here, we can easily have the following generalization

Draw a circle a pick a point P0. Draw a regular p-gon A0A1. . . Ap-1 and a regular q-gon B0B1. . . Bq-1 with A0=B0=P0. If p and q are two relatively prime numbers then there must exist Ax and By such that AxBy form a side of a regular pq-gon. It means that if a regular p-gon and a regular q-gon are constructible then a regular pq-gon is also constructible.

Let us stop here for now. In our future posts, we will look at Gauss' method of constructing regular 17-gon. Hope to see you again then.

Homework.

1. Prove the generalization mentioned at the end of the post:

Let gcd(p,q)=1. If a regular p-gon and a regular q-gon are constructible by compass and straightedge then a regular pq-gon is also constructible.

2. Go to google.com and search about Gauss' 17-gon construction.

Source:
http://mathgardenblog.blogspot.com/2014/06/construct-15gon.html

3.6 Moivre's formula

In previous post we have learned about complex numbers. Today, we will learn about the trigonometric form of a complex number and the famous de Moivre's formula.

Let us recall that the complex numbers are created by the introduction of a very special number { the number i { with the property i2= -1.

$$i^2 = -1: \tag{3.119}$$

Complex numbers have the form

$$a + ib \tag{3.120}$$

where a and b are two real numbers. In previous post, we have learned some basic facts about complex numbers

Addition and Subtraction

$$(a + ib) + (c + id) = (a + c) + i(b + d); \tag{3.121}$$

$$(a + ib) - (c + id) = (a - c) + i(b - d): \tag{3.122}$$

Multiplication

$$(a+ib)(c+id) = ac + iad + ibc + i^2bd = (ac - bd) + i(bc + ad). \tag{3.123}$$

Division we need to use the identity

$$(a+ib)(a-ib) = a^2 - i^2b^2 = a^2 + b^2 \tag{3.124}$$

and so

$$\frac{c+id}{a+ib} = \frac{(c+id)(a-ib)}{(a+ib)(a-ib)} = \frac{(ac+bd)+i(ad-bc)}{a^2+b^2} = \frac{ac+bd}{a^2+b^2} + i\frac{ad-bc}{a^2+b^2}. \tag{3.125}$$

Complex conjugate

$$\overline{a+ib} = a - ib, \quad \overline{a-ib} = a + ib. \tag{3.126}$$

Absolute value

$$|a+ib| = \sqrt{a^2+b^2}. \tag{3.127}$$

Trigonometric form of a complex number

Today we will learn a very important fact about complex numbers, that is, any complex number z can be written in a trigonometric form as follows

$$z = r(\cos\theta + i\sin\theta), \tag{3.128}$$

where $r = |z|$.

Indeed, let $z = a + ib$, then

$$r = |z| = \sqrt{a^2+b^2}, \tag{3.129}$$

we have

$$\frac{z}{r} = \frac{a}{r} + i\frac{b}{r}. \tag{3.130}$$

Since

$$\left(\frac{a}{r}\right)^2 + \left(\frac{b}{r}\right)^2 = \frac{a^2+b^2}{r^2} = 1, \tag{3.131}$$

there exists θ such that

$$\frac{a}{r} = \cos \ ; \qquad \frac{b}{r} = \sin \ : \tag{3.132}$$

Thus,

$$\frac{z}{r} = \frac{a}{r} + i\frac{b}{r} = \cos \ + i \ \sin \ : \tag{3.133}$$

And we obtain the trigonometric form of the complex number z

$$z = r(\cos \ + i \ \sin \ ): \tag{3.134}$$

In special case when z=0 we can let r=0.

Multiplication in trigonometric form Trigonometric form is very convenient when we do multiplication for complex numbers thanks to the following identity

$$(\cos \ + i \ \sin \ )(\cos \ + i \ \sin \ ) = \cos(\ + \ ) + i \ \sin(\ + \ ): \tag{3.135}$$

If we have two complex numbers u and v, and express them in the trigonometric form

$$u = r(\cos \ + i \ \sin \ ); \tag{3.136}$$

$$v = s(\cos \ + i \ \sin \ ); \tag{3.137}$$

then their product can be easily obtained as

$$uv = rs(\cos(\ + \ ) + i \ \sin(\ + \ )): \tag{3.138}$$

Exponentiation and de Moivre's formula Similar to multiplication, exponentiation is also easy when we write the complex number in trigonometric form. If

$$u = r(\cos \ + i \ \sin \ ) \tag{3.139}$$

then

$$u^n = r^n(\cos(n \ ) + i \ \sin(n \ )): \tag{3.140}$$

The following identity is called the de Moivre's formula, this is an important formula in complex numbers

$$(\cos \ + i \ \sin \ )^n = \cos(n \ ) + i \ \sin(n \ ): \tag{3.141}$$

Now let us do some exercises.

Problem 1: Solve the quadratic equation

$$x^2 - 2x + 4 = 0 \tag{3.142}$$

and write its complex roots in trigonometric form.

Solution: We have

$$\Delta' = 1^2 - 4 = -3; \tag{3.143}$$

so the quadratic equation has complex roots

$$1 \pm i\sqrt{3}. \tag{3.144}$$

To write these complex numbers in trigonometric form, we need to calculate their absolute value

$$|1 \pm i\sqrt{3}| = \sqrt{1^2 + (\sqrt{3})^2} = \sqrt{4} = 2. \tag{3.145}$$

Thus,

$$1 \pm i\sqrt{3} = 2\left(\frac{1}{2} \pm i\frac{\sqrt{3}}{2}\right) = 2(\cos\frac{\pi}{3} \pm i\ \sin\frac{\pi}{3}). \tag{3.146}$$

Problem 2: Solve the quadratic equation

$$x^2 - x + 1 = 0 \tag{3.147}$$

and write its complex roots in trigonometric form.

Solution: We have

$$\Delta = 1^2 - 4 = -3; \tag{3.148}$$

so the quadratic equation has complex roots

$$\frac{1 \pm i\sqrt{3}}{2}. \tag{3.149}$$

To write these complex numbers in trigonometric form, we need to calculate their absolute value

$$\left|\frac{1 \pm i\sqrt{3}}{2}\right| = \sqrt{\left(\frac{1}{2}\right)^2 + \left(\frac{\sqrt{3}}{2}\right)^2} = 1. \tag{3.150}$$

Thus,

$$\frac{1-i\sqrt{3}}{2}=\frac{1}{2}-i\frac{\sqrt{3}}{2}=\cos\frac{\pi}{3}-i\sin\frac{\pi}{3}. \tag{3.151}$$

Problem 3: Solve the quadratic equation

$$x^2-3x+3=0 \tag{3.152}$$

and write its complex roots in trigonometric form.

Solution: We have

$$\Delta=(-3)^2-4\cdot 3=-3, \tag{3.153}$$

so the quadratic equation has complex roots

$$\frac{3\pm i\sqrt{3}}{2}. \tag{3.154}$$

To write these complex numbers in trigonometric form, we need to calculate their absolute value

$$\left|\frac{3\pm i\sqrt{3}}{2}\right|=\sqrt{\left(\frac{3}{2}\right)^2+\left(\frac{\sqrt{3}}{2}\right)^2}=\sqrt{3}. \tag{3.155}$$

Thus,

$$\frac{3\pm i\sqrt{3}}{2}=\sqrt{3}\left(\frac{\sqrt{3}}{2}\pm i\frac{1}{2}\right)=\sqrt{3}\left(\cos\frac{\pi}{6}\pm i\sin\frac{\pi}{6}\right). \tag{3.156}$$

Problem 4: Calculate $(1+i)^{2012}$ by two methods, using de Moivre's formula and using binomial identity, and deduce the following equality

$$\binom{2012}{0}-\binom{2012}{2}+\binom{2012}{4}-\binom{2012}{6}+\cdots+\binom{2012}{2008}-\binom{2012}{2010}+\binom{2012}{2012}=-2^{1006}. \tag{3.157}$$

Solution: In the first method, we write $1+i$ in trigonometric form and then use de Moivre's formula to calculate the exponentiation.

First, we calculate the absolute value o $1+i$:

$$|1+i|=\sqrt{1^2+1^2}=\sqrt{2}. \tag{3.158}$$

Thus,

$$1+i=\sqrt{2}\left(\frac{\sqrt{2}}{2}+i\frac{\sqrt{2}}{2}\right)=\sqrt{2}\left(\cos\frac{\pi}{4}+i\sin\frac{\pi}{4}\right). \tag{3.159}$$

Using de Moivre's formula, we calculate the exponentiation

$$(1+i)^{2012} = (\sqrt{2})^{2012}(\cos\frac{2012}{4} + i\ \sin\frac{2012}{4}) = 2^{1006}(\cos(503\ ) + i\ \sin(503\ )) = \quad 2^{1006}: \tag{3.160}$$

On the other hand, by binomial identity, we have

$$(1+i)^{2012} = 1 + \binom{2012}{1} i + \binom{2012}{2} i^2 + \binom{2012}{3} i^3 + \binom{2012}{4} i^4 + \binom{2012}{5} i^5 + ::: + \binom{2012}{2011} i^{2011} + i^{2012} \tag{3.161}$$

$$= 1 + \binom{2012}{1} i \quad \binom{2012}{2} \quad \binom{2012}{3} i + \binom{2012}{4} + \binom{2012}{5} i + ::: \quad \binom{2012}{2011} i + 1 \tag{3.162}$$

Comparing the real part of the above two results, we obtain

$$1 \quad \binom{2012}{2} + \binom{2012}{4} \quad \binom{2012}{6} + ::: + \binom{2012}{2008} \quad \binom{2012}{2010} + 1 = \quad 2^{1006}: \tag{3.163}$$

Let us stop here for now, see you again in the next post.

Homework.

1. Write these numbers in trigonometric form: $1 \quad i$, $3 + 3i$, $\sqrt{3} + 3i$, $3 \quad \sqrt{3}i$, $\quad 7 + 7i$, $3i$.

2. What is the absolute value of this complex number $\cos + i\ \sin\ $?

3. Let $u = r(\cos\ + i\ \sin\ )$ and $v = s(\cos\ + i\ \sin\ )$ calculate $u=v$.

4. Calculate $(1+i)^{2013}$ by two methods, the de Moivre's formula and the binomial identity, and then deduce some combinatorial equalities.

5. Write x in trigonometric form and find all possible values of x such that $x^4 = \quad 1$.

Source:
http://mathgardenblog.blogspot.com/2013/01/complex2.html

3.7 Teste ARRAY - matrizes

teste

$$\begin{bmatrix} V_{ag} - V_{a'g'} \\ V_{bg} - V_{b'g'} \\ V_{cg} - V_{c'g'} \\ V_{1g} - V_{1'g'} \\ V_{2g} - V_{2'g'} \\ V_{3g} - V_{3'g'} \end{bmatrix} = \begin{bmatrix} z_{aa} & z_{ab} & z_{ac} & z_{a1} & z_{a2} & z_{a3} \\ z_{ab} & z_{bb} & z_{bc} & z_{b1} & z_{b2} & z_{b3} \\ z_{ac} & z_{bc} & z_{cc} & z_{c1} & z_{c2} & z_{c3} \\ z_{a1} & z_{b1} & z_{c1} & z_{11} & z_{12} & z_{13} \\ z_{a2} & z_{b2} & z_{c2} & z_{21} & z_{22} & z_{23} \\ z_{a3} & z_{b3} & z_{c3} & z_{31} & z_{32} & z_{33} \end{bmatrix} \begin{bmatrix} I_a \\ I_b \\ I_c \\ I_1 \\ I_2 \\ I_3 \end{bmatrix} \qquad (3.164)$$

Source:
https://sites.google.com/site/latexgrucad/equacoes-e-subequacoes

3.8 Complex Number

Today we will learn about the complex numbers. The crucial point about complex numbers is that we accept a very special number that we will denote it by i. This number i is very special because it satisfies the following identity

$$i^2 = -1. \qquad (3.165)$$

So a complex number will have the form

$$a + ib \qquad (3.166)$$

where a and b are two real numbers. When b=0 then a+ib=a is just a normal real number, and when a=0 then a+ib=ib is called a pure imaginary number. Here are some examples of complex numbers:

$$1 + i; \quad 2 - 3i; \quad -\sqrt{3} + 4i; \quad 5i - 4; \quad 6; \quad i; \quad -3i; \quad 4 + 2i; \ldots \qquad (3.167)$$

First, let us learn about some basic arithmetic like addition, subtraction, multiplication, division of complex numbers.

Addition and Subtraction: are quite simple and trivial as follows

$$(3 + 2i) + (4 + i) = 7 + 3i; \quad (3 + 2i) - (4 + i) = -1 + i; \qquad (3.168)$$

$$(2 + i) + (3 - 4i) = 5 - 3i; \quad (2 + i) - (3 - 4i) = -1 + 5i; \qquad (3.169)$$

$$2i + (3i + 1) = 1 + 5i; \quad 2i - (3i + 1) = -1 - i; \qquad (3.170)$$

In general, we have

$$(a+ib)+(c+id)=(a+c)+i(b+d);\quad (a+ib)-(c+id)=(a-c)+i(b-d). \tag{3.171}$$

Multiplication: we need to use the identity $i^2=-1$ in multiplication, for examples,

$$(3+2i)(4+i)=12+3i+8i+2i^2=12+11i-2=10+11i; \tag{3.172}$$

$$(2+i)(3-4i)=6-8i+3i-4i^2=6-5i+4=10-5i; \tag{3.173}$$

$$2i(3i+1)=6i^2+2i=-6+2i. \tag{3.174}$$

In general, we have

$$(a+ib)(c+id)=ac+iad+ibc+i^2bd=(ac-bd)+i(bc+ad). \tag{3.175}$$

Division: let us recall the division of the radicals, suppose that we want to calculate

$$\frac{1+4\sqrt{3}}{5+2\sqrt{3}}; \tag{3.176}$$

then we will use the following identity

$$(5+2\sqrt{3})(5-2\sqrt{3})=5^2-(2\sqrt{3})^2=25-12=13. \tag{3.177}$$

We have

$$\frac{1+4\sqrt{3}}{5+2\sqrt{3}}=\frac{(1+4\sqrt{3})(5-2\sqrt{3})}{(5+2\sqrt{3})(5-2\sqrt{3})}=\frac{5-2\sqrt{3}+20\sqrt{3}-24}{25-12}=\frac{-19+18\sqrt{3}}{13}=-\frac{19}{13}+\frac{18}{13}\sqrt{3}. \tag{3.178}$$

Similarly, for complex numbers, we will use the identity

$$(a+ib)(a-ib)=a^2-i^2b^2=a^2+b^2. \tag{3.179}$$

We have

$$\frac{c+id}{a+ib} = \frac{(c+id)(a-ib)}{(a+ib)(a-ib)} = \frac{(ac+bd)+i(ad-bc)}{a^2+b^2} = \frac{ac+bd}{a^2+b^2} + i\frac{ad-bc}{a^2+b^2}. \tag{3.180}$$

Here are some examples,

$$\frac{3+2i}{4+i} = \frac{(3+2i)(4-i)}{(4+i)(4-i)} = \frac{12-3i+8i-2i^2}{16-i^2} = \frac{14+5i}{17} = \frac{14}{17}+\frac{5}{17}i, \tag{3.181}$$

$$\frac{2+i}{3-4i} = \frac{(2+i)(3+4i)}{(3-4i)(3+4i)} = \frac{6+8i+3i+4i^2}{9-16i^2} = \frac{2+11i}{25} = \frac{2}{25}+\frac{11}{25}i, \tag{3.182}$$

$$\frac{2i}{3i+1} = \frac{2i(1-3i)}{(1+3i)(1-3i)} = \frac{2i-6i^2}{1-9i^2} = \frac{6+2i}{10} = \frac{3}{5}+\frac{1}{5}i, \tag{3.183}$$

Conjugation: as we have seen above, the two complex numbers a+ib and a-ib are used together in division because their product is a real number

$$(a+ib)(a-ib) = a^2 - i^2b^2 = a^2+b^2. \tag{3.184}$$

These two complex numbers are called conjugate and they are very useful when they are used together. We will use the notation $\bar{z}$ to denote the conjugate of z, thus,

$$\overline{a+ib} = a-ib; \quad \overline{a-ib} = a+ib. \tag{3.185}$$

In special cases, the conjugate of a real number a is a, and the conjugate of ib is -ib.

Absolute value: for real numbers, we have the concept of absolute value,

$$|5| = 5; \quad |-5| = 5. \tag{3.186}$$

This concept can be extended for complex numbers. The absolute value of a complex number z=a+ib is denoted by |z| and defined as

$$|z| = |a+ib| = \sqrt{a^2+b^2}. \tag{3.187}$$

It follows that

$$|z|^2 = a^2+b^2 = (a+ib)(a-ib) = z\bar{z}. \tag{3.188}$$

Here are some examples,

$$|3+2i| = \sqrt{9+4} = \sqrt{13}; \tag{3.189}$$

$$|2-i| = \sqrt{4+1} = \sqrt{5}; \tag{3.190}$$

$$|3i| = \sqrt{0+9} = 3; \tag{3.191}$$

$$|2| = \sqrt{4+0} = 2. \tag{3.192}$$

In special cases, we have $|5| = 5$, $|-5| = 5$ and $|3i| = 3$, $|-3i| = 3$.

Fibonacci's Identity

Absolute value has the following important property

$$|u|\,|v| = |uv|. \tag{3.193}$$

This property gives rise to the following Fibonacci's identity

$$(a^2+b^2)(c^2+d^2) = (ac-bd)^2 + (ad+bc)^2. \tag{3.194}$$

Indeed, if we take u=a+ib and v=c+id then uv=(ac-bd)+i(ad+bc) and we have

$$|u| = \sqrt{a^2+b^2}; \tag{3.195}$$

$$|v| = \sqrt{c^2+d^2}; \tag{3.196}$$

$$|uv| = \sqrt{(ac-bd)^2 + (ad+bc)^2}. \tag{3.197}$$

So the identity $|u||v| = |uv|$ becomes

$$(a^2+b^2)(c^2+d^2) = (ac-bd)^2 + (ad+bc)^2. \tag{3.198}$$

If you ever forget about this identity and would like to construct it back, you can follow the following steps

First, write two arbitrary complex numbers and calculate the product

$$(a+ib)(c+id) = (ac-bd) + i(ad+bc); \tag{3.199}$$

Next, substitute the plus sign by the minus sign,

$$(a-ib)(c-id) = (ac-bd) - i(ad+bc); \tag{3.200}$$

Finally, multiply the above two equations together

$$[(a+ib)(a-ib)][(c+id)(c-id)] = [(ac-bd)+i(ad+bc)][(ac-bd)-i(ad+bc)] \tag{3.201}$$

and we obtain the required identity

$$(a^2+b^2)(c^2+d^2) = (ac-bd)^2+(ad+bc)^2. \tag{3.202}$$

Today we have learned some basic facts about complex numbers. This may surprise you but complex numbers are actually used a lot in number theory. We will learn more about complex number in the next post. Hope to see you then.

Homework.

1. Perform addition, subtraction, multiplication and division for the following pairs of complex numbers:

$$3+4i, 2+3i \tag{3.203}$$

$$-4i, 2-i, \tag{3.204}$$

$$7-i, 2+4i, \tag{3.205}$$

$$5-2i, 3+i, \tag{3.206}$$

$$4+3i, -i. \tag{3.207}$$

2. Find the complex conjugates for 5+3i, 4-2i, 1+i, -4, -1+2i, 5i, -4i.

3. Calculate the absolute values for 2+3i, 2-3i, 4+i, 4-i, 5+2i, 5-2i, 6, -6, 2-i, 2+i, 4i, -4i.

4. Calculate i3, i4, i5, i6,. . . .

Find a general formula for i^n.

5. Prove that

$$\left|\frac{u}{v}\right| = \frac{|u|}{|v|}. \tag{3.208}$$

6. Prove the following inequality

$$|u+v| \le |u|+|v|. \tag{3.209}$$

7. Show that for any complex number z, there exists such that

$$z = jzj(\cos \quad + i \; \sin \quad): \tag{3.210}$$

8. Use mathematical induction to prove the following identity

$$(\cos \quad + i \; \sin \quad)^n = \cos (n \quad) + i \; \sin (n \quad): \tag{3.211}$$

9. Prove that

$$(\cos \quad + i \; \sin \quad)(\cos \quad + i \; \sin \quad) = \cos (\quad + \quad) + i \; \sin (\quad + \quad): \tag{3.212}$$

10. Show that there exist two natural numbers a and b such that

$$a^2 + b^2 = 5^{2013}: \tag{3.213}$$

Source:
http://mathgardenblog.blogspot.com/2013/01/complex1.html

Index

index entries

Bibliography

entries

Mathematics - Annotations 003

Prof. Stenio Musich

July 29, 2020

Chapter 1

Mathematics

1.1 Geometry - Start

Date: 20200730

1) C:/DATA-NovoVivo-123-2020/Books-123-2020/Calculus/ Fast-Start-Differential-Calculus.pdf

Pag. 06 (21) - Lines

LINES We will start with an example of a very simple line: y D x C 1 (Figure 1.1). For any value of x we add one to x and get y. We can draw the line by picking any two points on it, lining up a straight edge on the two points, and drawing along the straight edge. The line y D x C 1 is made of all the points that look like .x; x C 1/. Five points like that have been plotted in Figure 1.1: (-2,-1), (-1,0), (0,1), (1,2), and (2,3). Definition 1.1 A line can always be written in the form y D mx C b where m is the slope of the line, and b is the y-intercept of the line. Both m and b are numbers. There are different ways to interpret the parameters m and b.

The number m is the amount the y-value increases for each increase of 1 in the x-value.

If we pick two points on the line, then the run between those points is the change in the x-value, and the rise is the change in the y-value. The slope, m, is the rise divided by the run.

The y-intercept b is where the line hits the y-axis. This means the point .0; b/ is on the line.

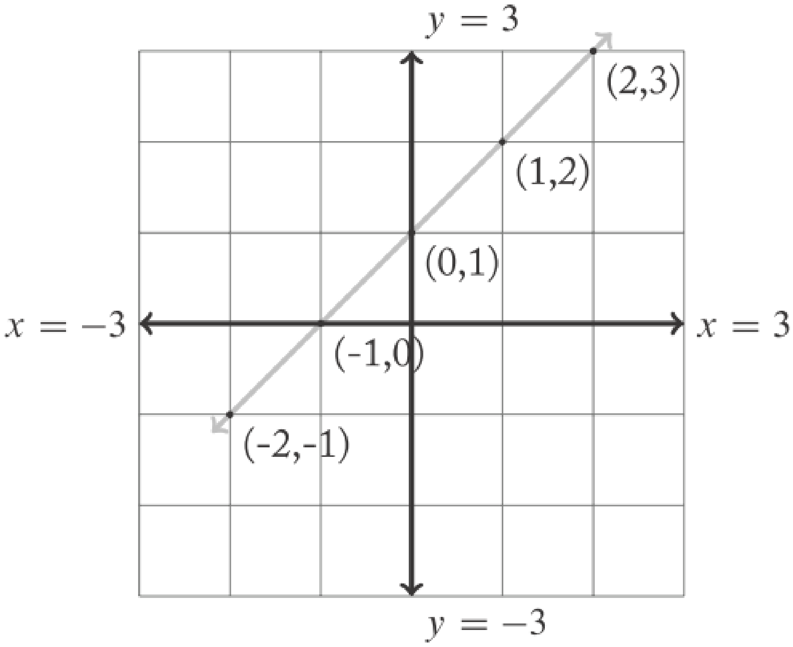

Figure 1.1: geo1.png

2) C:/DATA-NovoVivo-123-2020/Books-123-2020/Calculus/ Fast-Start-Integral-Calculus.pdf

Pag. 49 (64) - Parametric, Polar, and Vector Functions

3) C:/DATA-NovoVivo-123-2020/Books-123-2020/Calculus/ Fast-Start-Advanced-Calculus.pdf

Pag. 85 (101) - 3.2 ARC LENGTH AND SURFACE AREA

In this section we will learn to compute the length of curves and, having done that, to find the surface area of figures of rotation. A piece of a curve is called an arc. The key to finding the length of an arc is the differential of arc length. In the past we have had quantities like dx and dy that measure infinitesimal changes in the directions of the variables x and y. The differential of arc length is different - it does not point in a consistent direction, rather it points along a curve and so, by integrating it, we can find the length of a curve.

Examine Figure 3.5. The relationship between the change in x and y and the change in the length of the curve is Pythagorean, based on a right triangle. If we take this relationship to the infinitesimal scale, we obtain a formula for the differential of arc length.

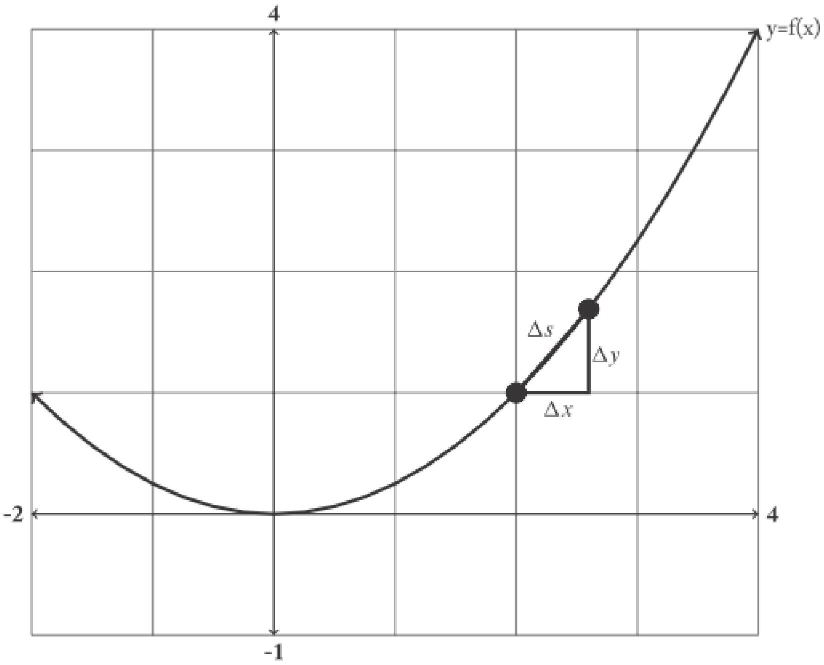

Figure 1.2: geo2.png

Figure 3.5:

The triangle with sides x, y, and s shows how the change in the length of the curve is related to the changes in distance in the x and y directions.

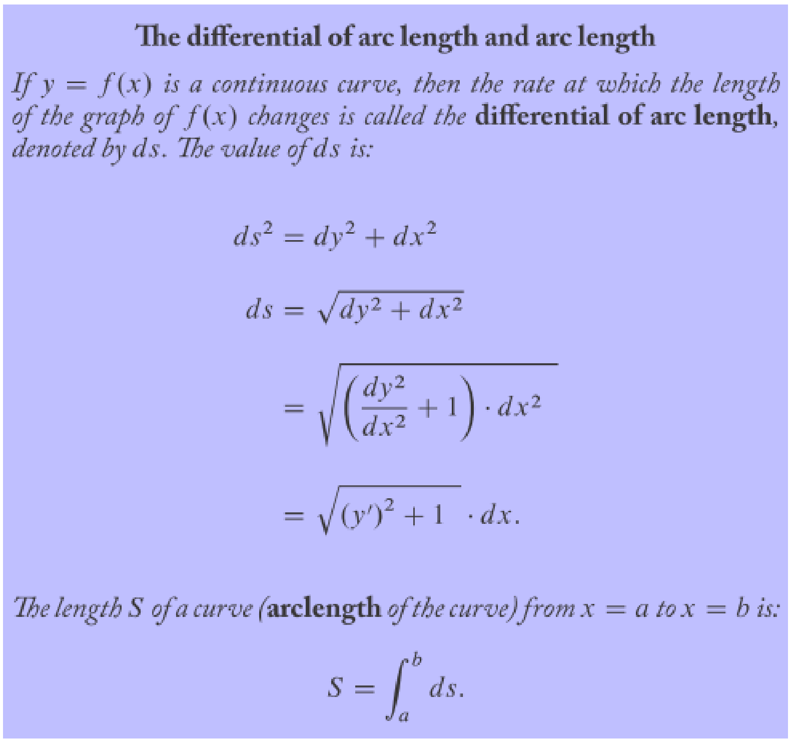

The differential of arc length and arc length

If $y = f(x)$ is a continuous curve, then the rate at which the length of the graph of $f(x)$ changes is called the **differential of arc length**, *denoted by ds. The value of ds is:*

$$ds^2 = dy^2 + dx^2$$

$$ds = \sqrt{dy^2 + dx^2}$$

$$= \sqrt{\left(\frac{dy^2}{dx^2} + 1\right) \cdot dx^2}$$

$$= \sqrt{(y')^2 + 1} \cdot dx.$$

*The length S of a curve (***arclength** *of the curve) from $x = a$ to $x = b$ is:*

$$S = \int_a^b ds.$$

Figure 1.3: geo3.png

4) C:/DATA-NovoVivo-123-2020/Books-123-2020/Calculus/Aspects of Differential Geometry I - 2015.pdf

Pag. 91 (107) - CHAPTER 3 Riemannian and Pseudo-Riemannian Geometry In Section 3.1, we introduce pseudo-Riemannian geometry. We show that the spheres of even di- mension do not admit Lorentzian metrics. We define the pseudo-Riema are perpendicular nnian measure, verify this agrees with the surface measure dS defined previously in Stokes' Theorem, and compute the volume of spheres and disks. In Section 3.2, we examine connections and their curvature. In Section 3.3, we specialize to the case of the Levi{Civita connection; this is the unique torsion-free Riemannian connection on the tangent bundle. In Section 3.4, we use the Levi{Civita connection to study geodesics. In Section 3.5, we use the Jacobi operator to establish some basic results concerning manifolds of constant sectional curvature. Section 3.6 is devoted to the proof of the Gauss{Bonnet Theorem and the study of Riemann surfaces. In Section 3.7, the generalization to higher dimensions and also to the pseudo-Riemannian setting is exhibited.

5) C:/DATA-NovoVivo-123-2020/Books-123-2020/Calculus/Aspects of Dif-

ferential Geometry II - 2015.pdf

Pag. 01 (17) - C H A P T E R 4 Additional Topics in Riemannian Geometry We continue the discussion of various topics in Riemannian geometry which was started in Chapter 3 of Book I. In Section 4.1, we discuss the geometry of curves and surfaces in Euclidean space that arise as solutions to an ordinary differential equation (ODE); this involves a nice application of the Gauss{Bonnet Theorem that is typical in the subject. In Section 4.2, we relate the scalar curvature to the growth of the volume of geodesic balls. In Section 4.3, we turn to holomorphic geometry. We present the Cauchy{Riemann equations, define almost complex structures, and state the Newlander{Nirenberg Theorem. We decompose d # and #' and introduce the spaces variation p, q. We define complex projective space. In Section 4.4, we present an introduction to Kahler geometry.

6) C:/DATA-NovoVivo-123-2020/Books-123-2020/Calculus/Aspects of Differential Geometry III - 2017.pdf

Pag. 52 (58) - 10.4 3-DIMENSIONAL WALKER LORENTZIAN MANIFOLDS I: CURVATURE HOMOGENEITY Th e material of this section and the next section is based on work of Garca-Ro, Gilkey and Nikcevic [67]. We will be concerned with the following examples.

7) C:/DATA-NovoVivo-123-2020/Books-123-2020/Calculus/latexmat.pdf

Pag. 96 - Smbolos Matematicos

8) C:/ARQ-IMPORTANT/Latex-books/The-LaTeX-Companion.pdf

9) C:/ARQ-IMPORTANT/Latex-books/Guide to LaTeX.pdf

10) C:/ARQ-BASICS/01/geometria.pdf IMPORTANT

11) C:/ARQ-BASICS/01/0071544127.pdf

Pag. 01 (12) - IMPORTANT

1.1 Historical Background of Geometry The word geometry is derived from the Greek words geos (meaning earth) and metron (meaning measure). The ancient Egyptians, Chinese, Babylonians, Romans, and Greeks used geometry for surveying, navigation, astronomy, and other practical occupations. The Greeks sought to systematize the geometric facts they knew by establishing logical reasons for them and relationships among them. The work of men such as Thales (600 B.C.), Pythagoras (540 B.C.), Plato (390 B.C.), and Aristotle (350 B.C.) in systematizing geometric facts and principles culminated in the geom- etry text Elements, written in approximately 325 B.C. by Euclid. This most remarkable text has been in use for over 2000 years. 1.2 Undefined Terms of Geometry: Point, Line, and Plane 1.2A Point, Line, and Plane are Undefined Terms These undefined terms underlie the definitions of all geometric terms. They can be given meanings by way of descriptions. However, these descriptions, which follow, are not to be thought of as definitions. 1.2B Point A point has position only. It has no length, width, or thickness. A point is represented by a dot. Keep in mind, however, that the dot represents a point but is not a point, just as a dot on a map may represent a locality but is not the locality. A dot, unlike a point, has size. A point is designated by a capital letter next to the dot, thus point A is represented: A. 1.2C Line A line has length but has no width or thickness. A line may be represented by the path of a piece of chalk on the blackboard or by a stretched rubber band.

12) C:/ARQ-BASICS/01/JacarandaMathsQuest10.pdf Obs.: Referência

13) C:/ARQ-BASICS/02/Euclidean and Non-Euclidean Geometries/Euclidean and Non-Euclidean Geometries. Development and History (1993, W. H. Freeman).pdf

Pag. 07 (25) - Nota de rodape

In recent years : has been approximated to a very large number of decimal places by computers; to five places, : is approximately 3.14159. In 1789 Johann Lambert proved that : was not equal to any fraction (rational number), and in 1882 F. Lindemann proved that : is a transcendental number, in the sense that it does not satisfy any algebraic equation with rational coefficients, which implies that in the Euclidean plane, it is impossible to square a circle using only straightedge and compass.

14) C:/DATA-NovoVivo-123-2020/Books-123-2020/Conformal Geometry of Discrete Groups and Manifolds/3110144042.pdf

Pag. 83 (95) - Figure 26. Gaussian image of a vertex with dihedral angles A, B, C

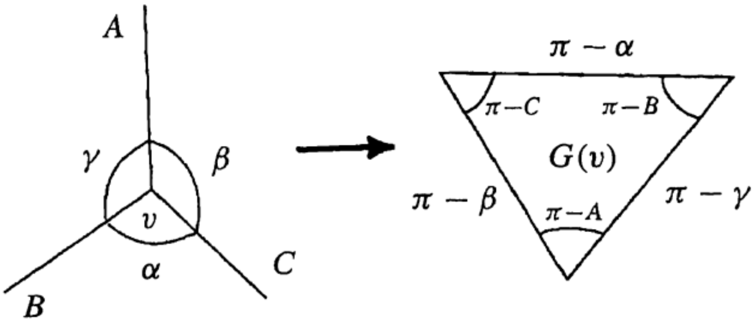
A
γ
β
v
α
B
C
π − α
π−C
π−B
G(v)
π − β
π−A
π − γ

Figure 1.4: geo4.png

15) C:/DATA-NovoVivo-123-2020/Books-123-2020/Atoms, Molecules and Photons An Introduction to Atomic, Molecular and Quantum Physics/3642102972.pdf

Pag. 48 (67) - Fig. 2.65. Hyperbolic electrostatic field with axial symmetry around the z-axis

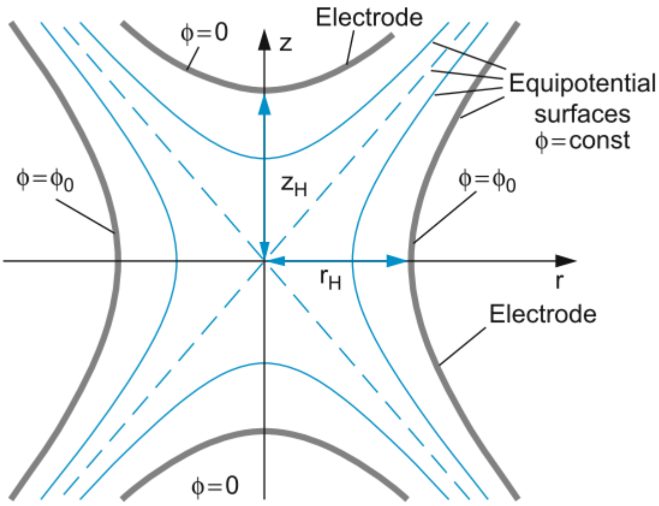
ϕ=0
z
Electrode
Equipotential
surfaces
ϕ=const
ϕ=ϕ0
zH
ϕ=ϕ0
rH
r
Electrode
ϕ=0

Figure 1.5: geo5.png

Pag. 159 (178) - 5. The Hydrogen Atom Based on the discussions in Chap. 4 we will now ap- ply the quantum mechanical treatment to the simplest atom, the H atom, which consists of one proton and one electron moving in the spherical symmetric Coulomb potential of the proton. These one-electron systems, such as the hydrogen atom and the ions He +, Li ++, Be +++, etc., are the only real systems for which the Schrodinger equation can be exactly (i.e., analytically) solved. For all other atoms or molecules approximations have to be made. Either the Schrodinger equation for these systems can be solved numerically (which offers a mathematical solution within the accuracy of the computer program, but generally gives little insight into the physical nature of the approximation), or the real atoms are described by approximate models that can be calculated analytically. In any case, for all multielectron systems, one has to live with approximations, either in the numerical solution of the exact atomic model or for the exact solution of the approximate model. A closer inspection of the spectrum of the hydrogen atom and other atoms reveals, however, that at higher spectral resolution the lines show a substructure that cannot be described by the Schrodinger theory, but is due to new effects, such as fine structure, hyperfine structure or the anomalous Zeeman effect. Therefore, even for the simple hydrogen atom the quantum mechanical model of an electron in the Coulomb field of the proton has to be modified by introducing new properties of electron and proton, such as electron spin or proton spin and their mutual interactions. These effects, which are small compared to the Coulomb energy, are included in a relativistic theory, based on the Dirac equation, which is called quantum electrodynamics. The Schrodinger equation can be regarded as the fundamental equation of nonrelativistic quantum theory. The treatment of the hydrogen atom illustrates in a very clear way the basic ideas of quantum mechanics,it explains the physical interpretation of quantum numbers, the description of the Zeeman effect and the fine structure by the model of angular momentum vector couplings and gives a better understanding of the more complex many-electron systems that will be discussed in the next chapter. 5.1 Schrodinger Equation for One-electron Systems

Pag. 201 (220) - 6. Atoms with More Than One Electron In atoms with more than one electron additional problems arise that are caused by mutual electrostatic and magnetic interactions between the electrons. In addition, we are now confronted with new symmetry principles that are valid if two electrons are exchanged. These stem from the fact that electrons cannot be distinguished from each other. We will first study these phenomena for the helium atom, which represents the simplest system with two electrons. This will help us understand the building up principle for the structure of electron shells for larger atoms. We will see that the electron configurations for all atoms can be obtained from the minimum energy principle, the correct coupling of the different angular momenta of the electrons and the observation of certain symmetry rules. This results in the determination of all possible energy states of the atoms and the characterization of these states by quantum numbers.

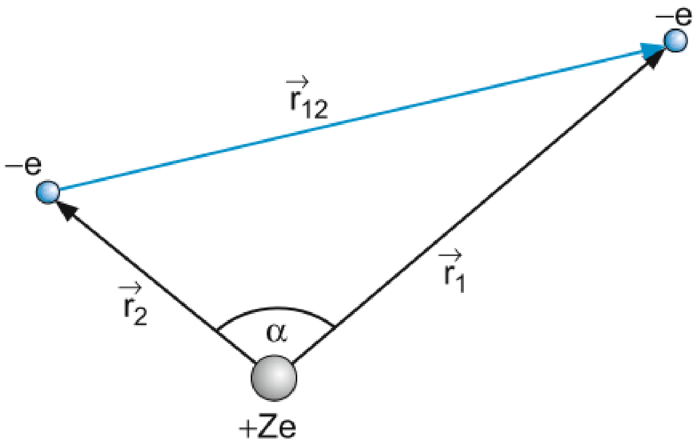

Figure 1.6: The Helium Atom.png

1.2 Geometry - Differential Geometry - Geodesic curve on a surface

1) C:/DATA-NovoVivo-123-2020/Books-123-2020/Applied Calculus of Variations for Engineers, Third edition/0367376091.pdf
Pag.141 (160) -

Chapter 8 Differential geometry Differential geometry is a classical mathematical area that has become very important for engineering applications in the recent decades. This importance is based on the rise of computer-aided visualization and geometry generation technologies. The chapter will address the fundamental problem of differential geometry, the finding of geodesic curves, that has practical implications in manufacturing. Development of non-mathematical surfaces used in ships and airplanes has serious financial impact in reducing material waste and improving the quality of the surfaces. While the discussion in this chapter will focus on analytically solvable problems, the methods and concepts we introduce will provide a foundation applicable in various engineering areas. 8.1 The geodesic problem Finding a geodesic curve on a surface is a classical problem of differential geometry. Variational calculus seems uniquely applicable to this problem. Let us consider a parametrically given surface

$$\underline{r} = x(u,v)\underline{i} + y(u,v)\underline{j} + z(u,v)\underline{k}. \tag{1.1}$$

Let two points on the surface be

$$\underline{r}_0 = x(u_0,v_0)\underline{i} + y(u_0,v_0)\underline{j} + z(u_0,v_0)\underline{k}.$$

and

$$\underline{r}_1 = x(u_1,v_1)\underline{i} + y(u_1,v_1)\underline{j} + z(u_1,v_1)\underline{k}.$$

The shortest path on the surface between these two points is the geodesic curve. Consider the square of the arc length

$$ds^2 = (dx)^2 + (dy)^2 + (dz)^2,$$

and compute the differentials related to the parameters,

$$ds^2 = E(u,v)(du)^2 + 2F(u,v)du dv + G(u,v)(dv)^2$$

Here the so-called first fundamental quantities are defined as

$$E(u,v) = \left(\frac{\partial x}{\partial u}\right)^2 + \left(\frac{\partial y}{\partial u}\right)^2 + \left(\frac{\partial z}{\partial u}\right)^2 = |\underline{r}'_u|^2,$$

and

$$F(u,v) = \frac{\partial x}{\partial u}\frac{\partial x}{\partial v} + \frac{\partial y}{\partial u}\frac{\partial y}{\partial v} + \frac{\partial z}{\partial u}\frac{\partial z}{\partial v} = \underline{r}'_u \cdot \underline{r}'_v,$$

and

$$G(u,v) = \left(\frac{\partial x}{\partial v}\right)^2 + \left(\frac{\partial y}{\partial v}\right)^2 + \left(\frac{\partial z}{\partial v}\right)^2 = |\underline{r}'_u|^2,$$

Assume that the equation of the geodesic curve in the parametric space is described by

$$v = v(u).$$

Then the geodesic curve is the solution of the variational problem

$$I(v) = \int_{u_0}^{u_1} \sqrt{E(u,v) + 2F(u,v)\frac{dv}{du} + G(u,v)\left(\frac{dv}{du}\right)^2}\, du$$

= extremum with boundary conditions

$$v(u_0) = v_0,$$

and

$$v(u_1) = v_1.$$

The corresponding Euler-Lagrange differential equation is

$$\frac{E_v + 2v' F_v + v'^2 G_v}{2\sqrt{E(u,v) + 2F(u,v)v' + G(u,v)v'^2}} - \frac{d}{du}\frac{F + Gv'}{\sqrt{E(u,v) + 2F(u,v)v' + G(u,v)v'^2}} = 0,$$

with the notation of

$$E_v = \frac{\partial E}{\partial v}, F_v = \frac{\partial F}{\partial v}, G_v = \frac{\partial G}{\partial v},$$

and

$$v' = \frac{dv}{du}.$$

The equation is rather difficult in general, and exploitation of special cases arising from the particular surface definitions is advised.

When the first fundamental quantities are only functions of the u parameter, the equation simplifies to

$$\frac{F + Gv'}{\sqrt{E(u,v) + 2F(u,v)v' + G(u,v)v'^2}} = c_1.$$

A further simplification is based on the practical case when the u and v parametric lines defining the surface are orthogonal. In this case

$$F = 0,$$

and the equation may easily be integrated as

$$v = c_1 \int \frac{\sqrt{E}}{\sqrt{G^2 - c_1^2 G}} du + c_2.$$

The integration constants may be resolved from the boundary conditions. When the first fundamental quantities are only functions of the v parameter, and the F = 0 assumption still holds, the equation becomes

$$\frac{Gv'^2}{\sqrt{E+Gv'^2}} - \sqrt{E+Gv'^2} = c_1.$$

Reordering and another integration bring

$$v = c_1 \int \frac{\sqrt{E^2 - c_1^2 E}}{\sqrt{G}}\,du + c_2.$$

2) C:/ZIP/Math-Important-5/LeviCivita-4.pdf

Pag. 01(07) - Nozione di Parallelismo in una varietà qualunque e conseguente specificazione geometrica della curvatura riemanniana.

3) C:/ZIP/Math-Important-5/Opere matematiche I.pdf

Pag. 01(46) - Sopra le coniche che toccano e secano una o più curve gobbe

4) C:/ZIP/Math-Important-5/Lezioni di analisi infinitesimale-2-Volumes/Volume-02/parte-1-capp-X-XVII.pdf

Pag. 261(29) - XIII - Area delle superficie curve. Volumi

1.3 Use Newtheorem

Date: 20200731

NEWTHEOREM

Theorem 1.3.1 (Leibniz-Musich Theorem)**.** Leibniz-Musich Theorem - Um theorema importantíssimo para a matemática

$$\frac{E_v + 2v'F_v + v'^2G_v}{2\sqrt{E(u,v) + 2F(u,v)v' + G(u,v)v'^2}}$$

Veremos que a equação 1.3.1 na página 12 deverá ser aplicada a todo o sistema matemático.

Outro teste:

Theorem 1.3.2 (Pitágoras)**.** O teorema de Pitágora diz que a hipotenusa ao quadrado é a soma dos catetos ao quadrado.

Usando o que foi criado:

Pelo teorema 1.3.2 na página 12 iremos observar que ela abriu inúmeras portas para o desenvolvimento de outras equações matemáticas.

A equação deverá ser desenvolvida em uma capítulo à parte.

1.4 Comandos LATEX

Date: 20200731

teste

$$\frac{\partial f}{\partial x}(a,b) = \lim_{h \to 0} \frac{f(a+h,b) - f(a,b)}{h}$$

$$a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \frac{1}{a_4}}}}$$

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{a_4}}}}$$

$$\frac{\frac{a}{x-y} + \frac{b}{x+y}}{1 + \frac{a-b}{a+b}}$$

$$\begin{pmatrix} \overrightarrow{ab} & \frac{e+f}{g-h} \\ cd & ij \\ 0 & kl \end{pmatrix}$$

Devemos testar todos os comandos do Latex

Devemos testar todos os comandos do Latex

Devemos testar todos os comandos do Latex

Devemos testar todos os comandos do Latex

$$\boxed{\begin{pmatrix} \overrightarrow{ab} & \frac{e+f}{g-h} \\ cd & ij \\ 0 & kl \end{pmatrix}}$$

1.5 Geometry - Trigonometry

Date: 20200801

1) C:/DATA-NovoVivo-124-2020/Books-124-20220/Important/255 Compiled and Solved Problems in Geometry and Trigonometry by Florentin Smarandache.pdf

2) C:/DATA-NovoVivo-124-2020/Books-124-20220/Important/Applied differential geometry - A modern introduction 2007.pdf

3) C:/DATA-NovoVivo-124-2020/Books-124-20220/Important/Convex polyhedra 2005.pdf

4) C:/DATA-NovoVivo-124-2020/Books-124-20220/Important/Euclidean Plane and its Relatives.pdf

5) C:/DATA-NovoVivo-124-2020/Books-124-20220/Important/Geometries 2012.pdf

6) C:/DATA-NovoVivo-124-2020/Books-124-20220/Important/Geometries By A. B. Sossinsky.pdf

7) C:/DATA-NovoVivo-124-2020/Books-124-20220/Important/Mathematics for IIT JEE Main and Advanced Algebra - Volume 1.pdf

8) C:/DATA-NovoVivo-124-2020/Books-124-20220/Important/Mathematics for IIT JEE Main and Advanced Trigonometry Vector Algebra Probability - Volume 2.pdf

9) C:/DATA-NovoVivo-124-2020/Books-124-20220/Important/Mathematics for IIT JEE Main and Advanced Calculus - Volume 3.pdf

10) C:/DATA-NovoVivo-124-2020/Books-124-20220/Important/Mathematics for IIT JEE Main and Advanced Coordinate Geometry - Volume 4.pdf

11) C:/DATA-NovoVivo-124-2020/Books-124-20220/Math-important-20200801/Geome
of differential geometry volume 1 - 2000.djvu
Pag. 419 (433) - CHAPTER 4 Einstein Metrics in Dimension Four

12) 11) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry/ Obs.: Pasta toda. Livros importantes sobre geometria.

1.6 Projective Geometry

Date: 20200802

1) C:/DATA-NovoVivo-124-2020/Books-124-20220/Important/ Projective Geometry - Volume 1 - 2007.djvu

Pag. 15(24) - CHAPTER I THEOREMS OF ALIGNMENT AND THE PRINCIPLE OF DUALITY 6. The assumptions of alignment. In the following treatment of projective geometry we have chosen the point and the line as unde- undefined elements. We consider a class (cf2, p. 2) the elements of which we call points, and certain undefined classes of points which we call lines. Here the words point and line are to be regarded as mere symbols devoid of all content except as implied in the as- assumptions (presently to be made) concerning them, and which may represent any elements for which the latter may he valid propositions. In other words, these elements are not to be considered as having properties in common with the points and lines of ordinary Euclidean geometry, except in so far as such properties are formal logical conse- consequences of explicitly stated assumptions. We shall in the future generally use the capital letters of the alphabet, as A, B, C, P, etc., as names for points, and the small let- letters, as a, b, c, I, etc., as names for lines. If A and B denote the same point, this will be expressed by the relation A = B; if they repre- represent distinct points, by the relation A =$ B. If A = B, it is sometimes said that A coincides with B, or that A is coincident with B. The same remarks apply to two lines, or indeed to any two elements of the same kind. All the relations used are defined in general logical terms, mainly by means of the relation of belonging to a class and the notion of one-to-one correspondence. In case a point is an element of one of the classes of points which we call lines, we shall express this relation by any one of the phrases: the point is on cr lies on or is a point of the line, or is united with the line; the line passes through or con- contains or is united with the point. We shall often find it convenient to use also the phrase the line is on the point to express this relation. Indeed, all the assumptions and theorems in this chapter will be stated consistently in this way. The reader will quickly become ac- accustomed to this " on " language, which is introduced with the purpose 15 16 THEOREMS OF ALIGNMENT AND DUALITY [Chap.i of exhibiting in its most elegant form one of the most far-reaching theorems of projective geometry (Theorem 11). Two lines which have a point in common are said to intersect in or to meet in that point, or to he on a common point. Also, if two distinct points lie on the same line, the line is said to join the points. Points which are on the same line are said to be collinear ; points which are not on the same line are said to be noncollinear. Lines which are on the same point (i.e. contain the same point) are said to be copunctal, or concurrent.[1] Concerning points and lines we now make the following assumpassumptions : The Assumptions of Alignment, A : A 1. If A and B are distinct points, there is at least one line on loth A and B. A 2. If A and B are distinct points, there is not more than one line on both A and B. A3. If A, B, C are points not all on the same line, and D and E (D ≠ E) are points such that B, C, D are on a line and C, A, E

Pag. 6(25) - Fig. 3

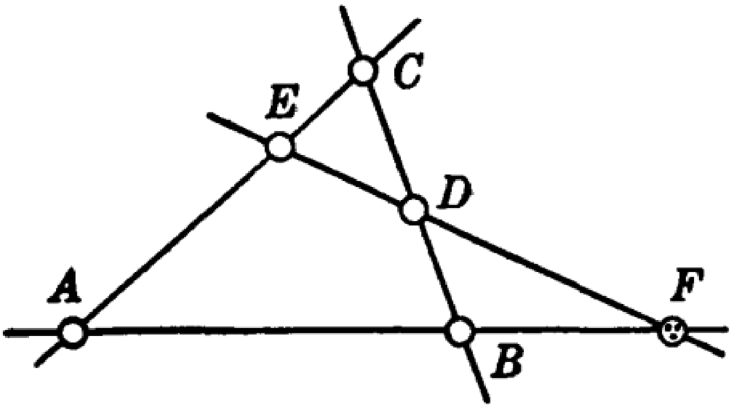

Figure 1.7: pg01.png

are on a line, there is a point F such that A, B, F are on a line and also D, E, F are on a line (fig. 3)[2]. It should be noted that this set of assumptions is satisfied by the triple system 1), p. 3, and also by the system of quadruples (2), p. 6, as well as by the points and lines of ordinary Euclidean geometry with the notion ‘points at infinity" (cf. paragrafo 3, p.8), and by The figures are to be regarded as a concrete representation of our science, in which the undefined "points" and "lines" of the science are represented by points and lines of ordinary Euclidean geometry (this requires the notion of ideal points; cf. x 3, p. 8). Their function is not merely to exhibit one of the many possible concrete representations, but also to help keep in mind the various rela- relations in question. In using them, however, great care must be exercised not to use any properties of such figures that are not formal logical consequences of the assumptions; in other words, care must be taken that all deductions are made formally from the assumptions and theorems previously derived from the assumptions. Obs.: Veja livro todo!!!

---

[1] The object of this paragraph is simply to define the terms in common use in terms of the general logical notion of belonging to a class. In later portions of this book we may omit the explicit definition of such common terms when such definition is obvious.

[2] The figures are to be regarded as a concrete representation of our science, in which the undefined "points" and "lines" of the science are represented by points and lines of ordinary Euclidean geometry (this requires the notion of ideal points; cf. x 3, p. 8). Their function is not merely to exhibit one of the many possible concrete representations, but also to help keep in mind the various rela- relations in question. In using them, however, great care must be exercised not to use any properties of such figures that are not formal logical consequences of the assumptions; in other words, care must be taken that all deductions are made formally from the assumptions and theorems previously derived from the assumptions.

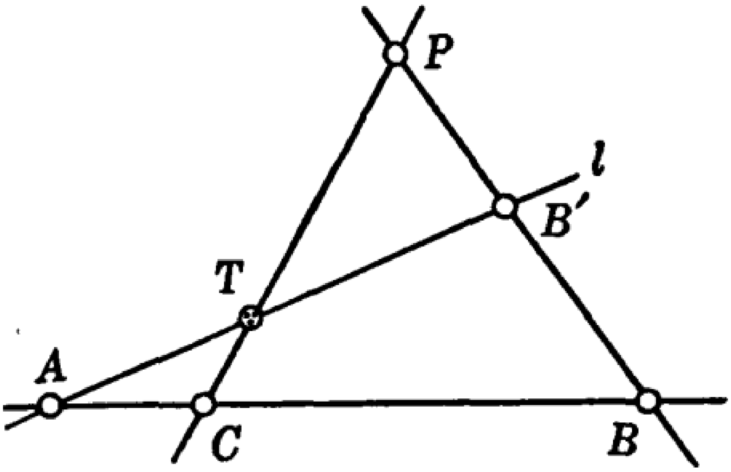

Figure 1.8: pg02.png

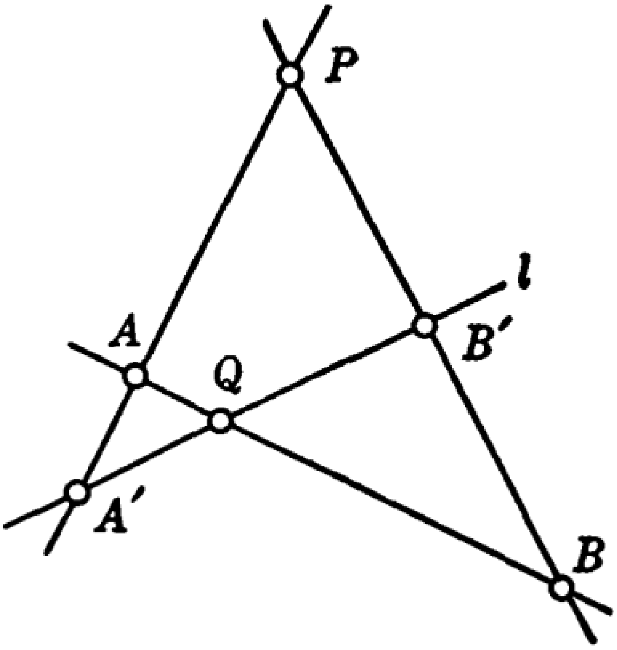

Figure 1.9: pg03.png

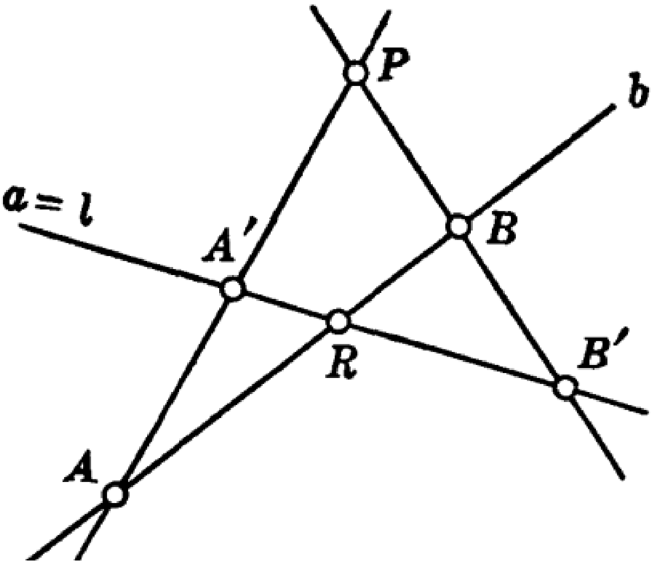

Figure 1.10: pg04.png

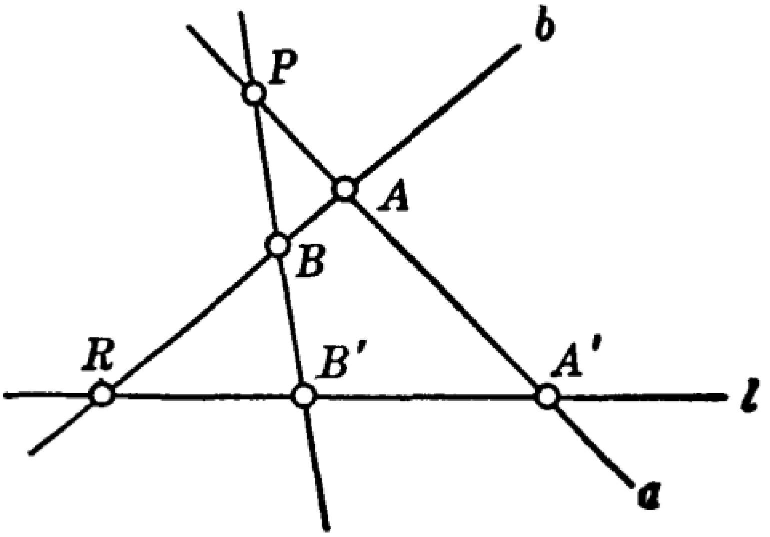

Figure 1.11: pg05.png

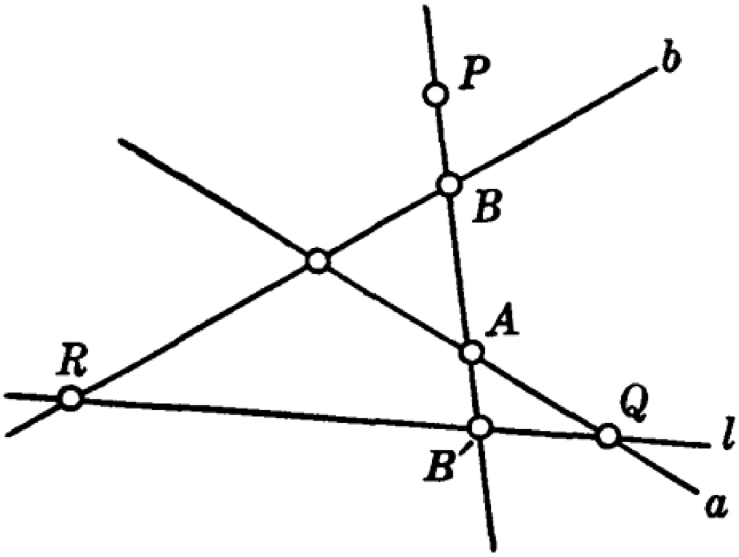

Figure 1.12: pg06.png

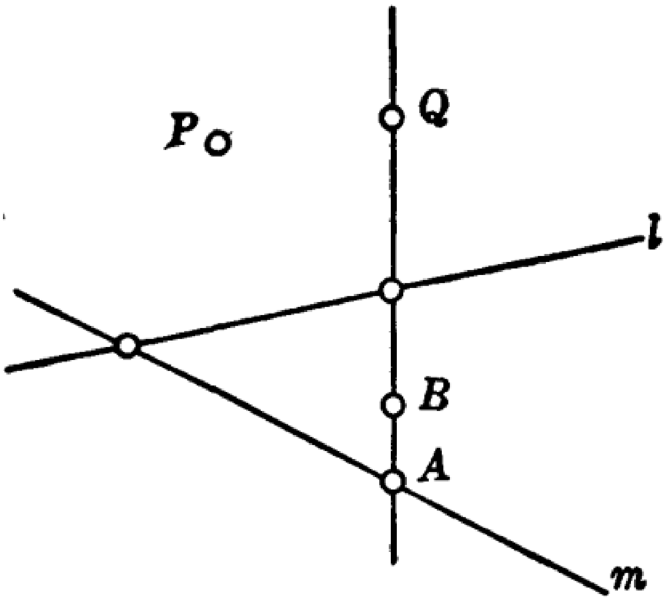

Figure 1.13: pg07.png

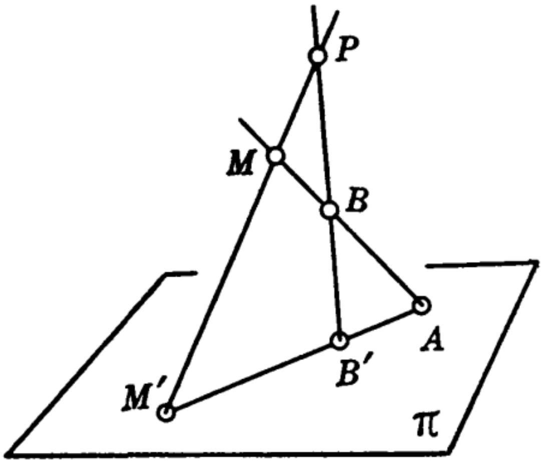

Figure 1.14: pg08.png

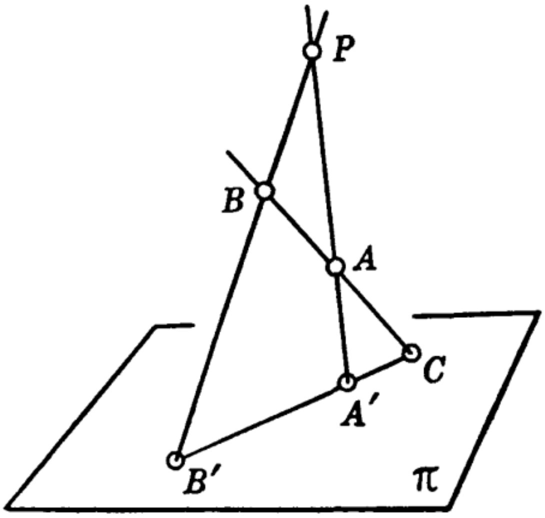

Figure 1.15: pg09.png

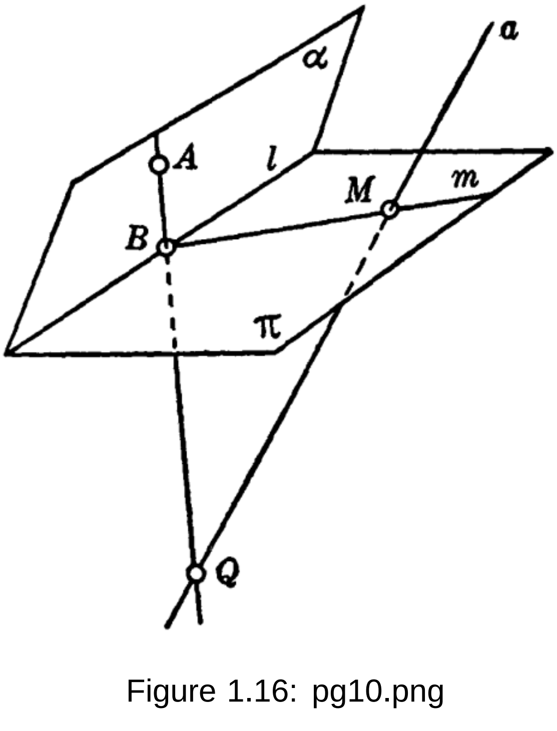

Figure 1.16: pg10.png

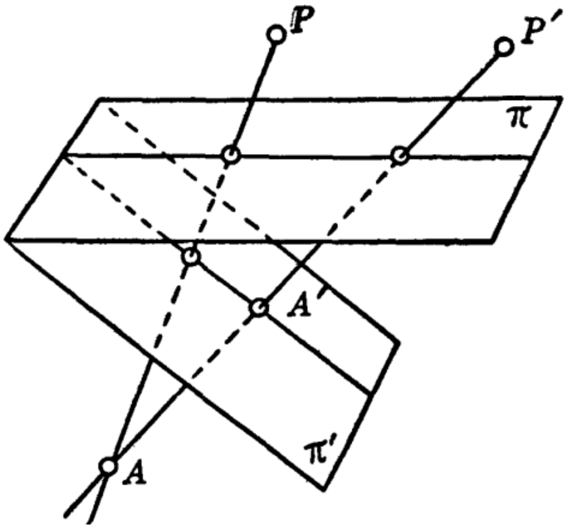

Figure 1.17: pg11.png

2) C:/DATA-NovoVivo-124-2020/Books-124-20220/Important/ Projective Geometry - Volume 2 - 2007.djvu

Pag. 119 (132) - CHAPTER IV EUCLIDEAN PLANE GEOMETRY 55. Geometries of the Euclidean type. We come now to the extension of the definition of congruence which was promised in xx 39 and 46. This requires the consideration of groups which are not self-conjugate under the afnne group. Not being self-conjugate, these groups are not determined uniquely by the affine group, and hence our definitions will contain a further arbitrary element. Definition. Let I be an arbitrary but fixed involution on lm. This involution shall be called the absolute or orthogonal involution. The group of all projective collineations leaving I invariant shall be called a parabolic[3] metric group. The transformations of the group shall be called similarity transformations. Two figures conjugate under the group shall be said to be similar. The geometry corresponding to the group shall be called the parabolic metric geometry. The absolute involution is supposed to be fixed throughout the rest of the discussion, but of course there are as many parabolic metric groups as there are choices of I. We nevertheless speak of the para- parabolic metric group in order to emphasize the fact that we are fixing attention on one group. In case the plane in which we are working is a real plane and the absolute involution is without double points, the parabolic metric geometry is the Euclidean geometry. It is for this reason that we refer to the parabolic metric geometries as geometries of the Euclidean type. The investigations in the following sections are arranged in order of increasing specialization. First we consider a perfectly general involution, I, in a projective plane satisfying A, E, P, HQ. Then we consider a particular type of involution in an ordered plane, and finally limit the plane to be the real plane. When the plane and the involution are fully specialized, it is a theorem x(70) that the real plane is contained in a complex plane in which the absolute involution has double points. Thus the theorems on the general type of involution (where the possible existence of double points is taken into account) come to have a new application.

---

[3] The reason for the term "parabolic" in this connection is explained in a later chapter, where the elliptic and hyperbolic metric groups are defined.

56. Orthogonal lines

Definition. Two lines are said to be orthogonal or perpendicular to each other if and only if they meet lm in conjugate points of the absolute involution.

Pag. 120 (133) - Lemma. Let 0 and I[be two points not collinear with either double point of the absolute involution. There is one and only one conic, C^2, having 0 as center, passing through P1 and having the pairs of the absolute involution as pairs of conjugate points. Proof Let P2 be the harmonic conjugate of Py with respect to 0 and the point at infinity, PX1 of the line OPV Any conic containing P1 and having 0 as center must contain P2, by the definition of center. Let X be a variable point of L, and Y the conjugate of X in the absolute involution. Any of the triangles OXY must be self-polar to any conic satisfying the required conditions. But if P is the point of intersection of the lines PXX and P2Y, and Q the point of intersection of PXX and OY, ...
Pag. 121 (134) - Fig. 45

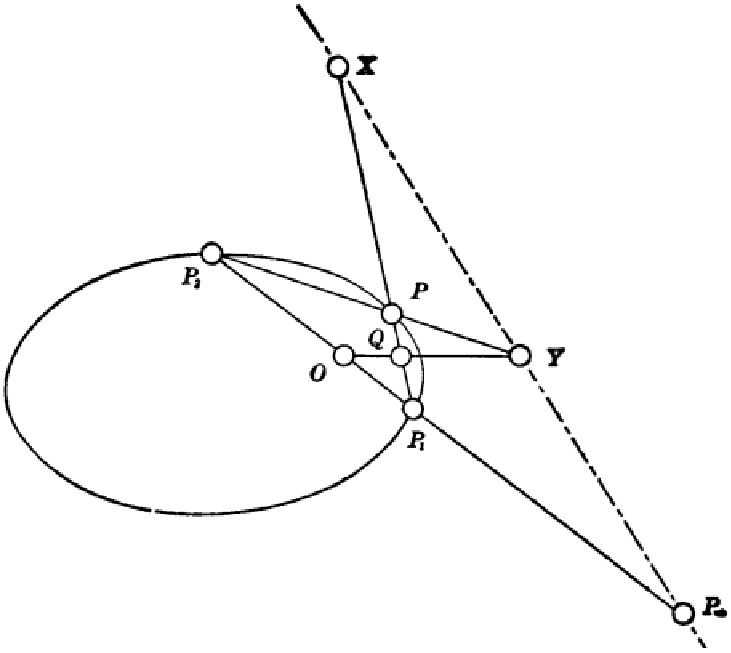

Figure 1.18: pg12.png

1.7 Pythagorean-hodograph curves

Date: 20200803

1) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry/Pythagorean-hodograph curves - algebra and geometry inseparable 2007.pdf

2) C:/DATA-NovoVivo-124-2020/Books-124-20220/Lecons sur les coordonnes curvilignes et leurs diverses applications-Important/Pythagorean-hodograph curves - algebra and geometry inseparable 2007.pdf

Pag. 104 (116) - 7.3 Curvilinear Coordinates Curvilinear coordinates were first systematically developed by the physicist, mathematician, and engineer Gabriel Lame (1795{1870) in his treatise Lecons sur les coordonnes curvilignes et leurs diverses applications of 1859. He used transformations to curvilinear coordinates in solving certain problems of heat conduction and elasticity. This proved to be a powerful idea: many problems in geometry, partial differential equations, multiple integrals, etc., become much simpler when cast in coordinate systems that reflect their symmetries.
Pag.95 (107) - 7.2.2 Barycentric Coordinates on Triangles

Pag. 123 (135) - 7.4.6 Geometrical Figures and their Shadows A deeper insight into projective transformations of the plane can be obtained from an elegant geometrical model that intuitively explains how parallel lines may be transformed into intersecting lines, or vice{versa. Consider a three{ dimensional Euclidean space in which two non{parallel planes | an \object" plane O and \image" plane I | are chosen together with a point \light source" or center of projection c that does not lie on O or I (see Fig. 7.10).

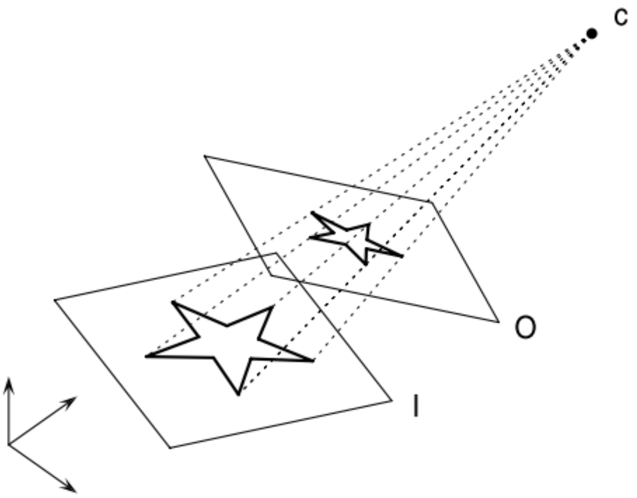

Figure 1.19: ph01.png - Visualization of a plane projective transformation by means of a central projection of an object plane O to an image plane I from a center of projection c.

We imagine the object plane O to be translucent, with a geometrical figure A drawn on it in black ink. The point light source at c then casts a shadow A' of A onto the image plane I, such that the finite points of A and A' exhibit an (almost) one{to{one correspondence. The mapping of A to A' is called a central projection, and we are obliged to say \almost" for the following reasons.

Consider the plane I // through the center of projection c that is parallel to the image plane I. This plane cuts the object plane O in a certain (affine) line l O . Clearly, no point of the line l O on O casts a finite shadow point on I, since rays from c to each point of this line are all parallel to I. Thus, we say that the affine line l O on O is mapped[4] to the line at infinity on I. Conversely, consider the plane O // through c that is parallel to the object plane. It cuts the image plane in a certain (affine) line l I , and we ask: what points of the object plane are mapped to this line l I on the image plane? If we consider a point p on I that approaches l I , the location at which the ray from c to p pierces the object plane recedes indefinitely along that plane, since this ray approaches parallelism with O. Thus, we say that the line l I on I is the \shadow" of the line at infinity on O. Figure 7.11 illustrates these concepts. From the above reasoning, it should be evident that almost all intersecting lines on O are mapped to intersecting lines on I. However, when two lines on O happen to intersect at a point on l O , they will be mapped to parallel lines on I. Conversely, parallel lines on O are mapped to intersecting lines on I | their intersections lie on the line l I.

---

[4] For obvious reasons, the line lO is often called the \vanishing line" on O.

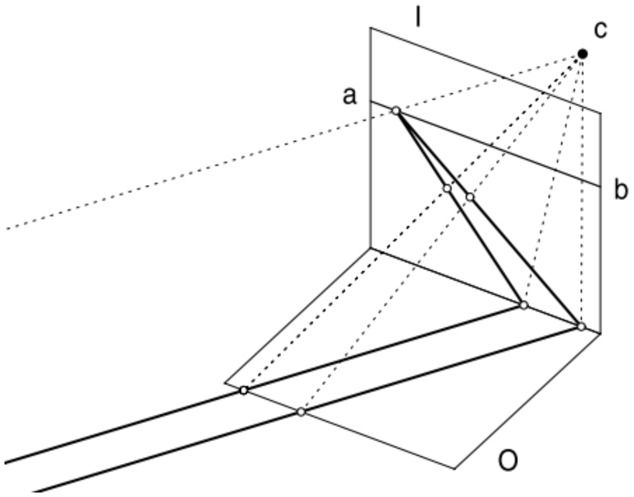

Figure 1.20: ph02.png - Parallel lines on the object plane O, extended indefinitely, meet at a point on the line ab when projected from the center c onto the image plane I. The line ab, defined by the intersection of I with a plane O// through c that is parallel to O, is the image on I of the \line at infinity" on O.

A central projection of an object plane O from a finite center c onto a non{parallel image plane I defines the most general form of two{dimensional projective transformation. When c recedes to infinity, the rays connecting it to points of O become parallel, and we then have a parallel projection | which corresponds to an affine transformation of the plane, since points at infinity on O are then mapped[5] to points at infinity on I. Finally, the case in which the image plane I and object plane O are parallel defines a shape–preserving affine transformation, namely, a simple scaling of the plane. The model described above offers an intuitive geometrical understanding of the projective transformations that were introduced in purely algebraic terms in § 7.4.4. It should be apparent that, in general, such transformations preserve the incidence relations (intersections, tangencies, etc.) in a geometrical figure — but they alter its size and shape. This can be seen, for example, in the case of the conics in Fig. 7.12, which imparts an intuitive interpretation to the statement that “the parabola is tangent to the line at infinity.”

---

[5] Unlike general projective transformations, this implies that parallel lines remain parallel under affine transformations.

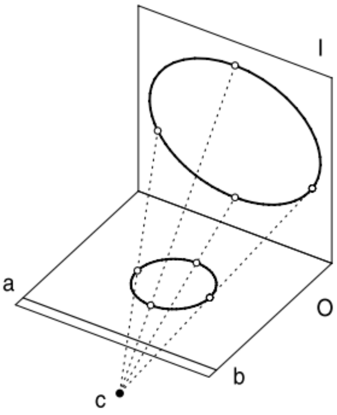

Figure 1.21: ph03a.png

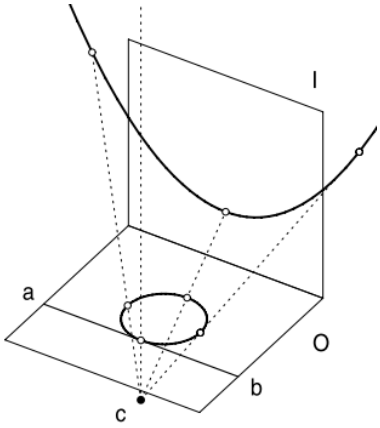

Figure 1.22: ph03b.png

Figure 1.23: Projective image of a circle. The “vanishing line” ab is the intersection of the object plane O with a plane through c that is parallel to the image plane I. The circle is mapped into an ellipse or a parabola according to whether ab lies outside the circle or touches it (not shown is the case where the vanishing line ab cuts the circle into two segments, which are mapped into the two branches of a hyperbola).

The equivalence of higher–order curves under projective transformations was perhaps first recognized by Sir Isaac Newton. In his *Enumeratio linearum tertii ordinis*, concerned with the classification of cubic curves (see also § 9.2) he includes [432] the following paragraphs on *genesis curvarum per umbras* — i.e., the generation of curves by shadows: “If the shadows of curves caused by a luminous point, be projected on an infinite plane, the shadows of conic sections will always be conic sections; those of curves of the second genus will always be of the second genus; those of the third genus will always be of the third genus; and so on ad infinitum . . . And in the same manner as the circle, projecting its shadow, generates all the conic sections, so the five divergent parabolas, by their shadows, generate all other curves of the second genus. And thus some of the more simple curves of other genera might be found, which would form all curves of the same genus by the projection of their shadows on a plane." In Newton's terminology[6] (borrowed from Descartes) conics are curves of the \first genus" while cubics are of the \second genus." His \divergent parabolas" are cubics described by an equation of the form

$$y^2 = ax^3 + bx^2 + cx + d \tag{1.2}$$

,

and he identifies five different types according to the root structure of the cubic on the right (namely: three simple real roots; a double root and a simple root either greater or smaller than it; one triple root; and one real root with two complex conjugate roots). Thus, Newton is claiming that any cubic may be regarded as the projective image of a curve of the above type. The desire of Renaissance artists for greater \realism" in their paintings was the cultural pretext for the development of projective geometry. This led to a deeper grasp of the \perspective" projection of a three{dimensional scene onto a plane (the canvas) from a specified point (eye position), as seen in Fig. 7.13. Paolo Uccello (1379{1435) of Florence was perhaps the first to fully master the method | his perspective rendering of a chalice appears on the cover of the journal Computer Aided Geometric Design.

3) C:/DATA-NovoVivo-124-2020/Books-124-20220/Lecons sur les coordonnes curvilignes et leurs diverses applications-Important/Lecons-sur-les-coordonnees-curvilignes-[...]Lame-Gabriel-bpt6k99670t.pdf Obs.: gallica.bnf.fr

4) C:/DATA-NovoVivo-124-2020/Books-124-20220/Lecons sur les coordonnes curvilignes et leurs diverses applications-Important/Poly-curvi.pdf

5) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important/47Geometrywi
Obs.: Livro todo!!!

6) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important/48Geometryw
Obs.: Livro todo!!!

---

[6] This should not be confused this with the modern concept of genus (see x9.2.5), that determines whether an algebraic curve admits a rational parameterization.

7) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important/Vectors, Matrices and Geometry (PDFDrive.com).pdf

8) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important/An Introduction to Vectors, Vector Operators and Vector Analysis (PDFDrive.com).pdf

9) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important/Fundamental of Mathematics Co-ordinate Geometry (PDFDrive.com).pdf

10) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important/Coordinate Geometry (Dover Books on Mathematics) (PDFDrive.com).pdf

11) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important/euclidean and non-euclidean geometries (PDFDrive.com).pdf

12) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important/Geometry Illuminated An Illustrated Introduction to Euclidean and Hyperbolic Plane Geometry (PDFDrive.com).pdf

13) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important/Geometry Through History Euclidean, Hyperbolic, and Projective Geometries (PDFDrive.com).pdf

14) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important/Geometry with Trigonometry, Second Edition (PDFDrive.com).pdf

15) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important/Parallel Coordinates Visual Multidimensional Geometry and Its Applications (PDFDrive.com).pdf IMPORTANT
Obs.: Livro todo!!!

Pag. 29 (56) - Theorem 2.5.3. A projectivity can be expressed by at most two perspectivities.

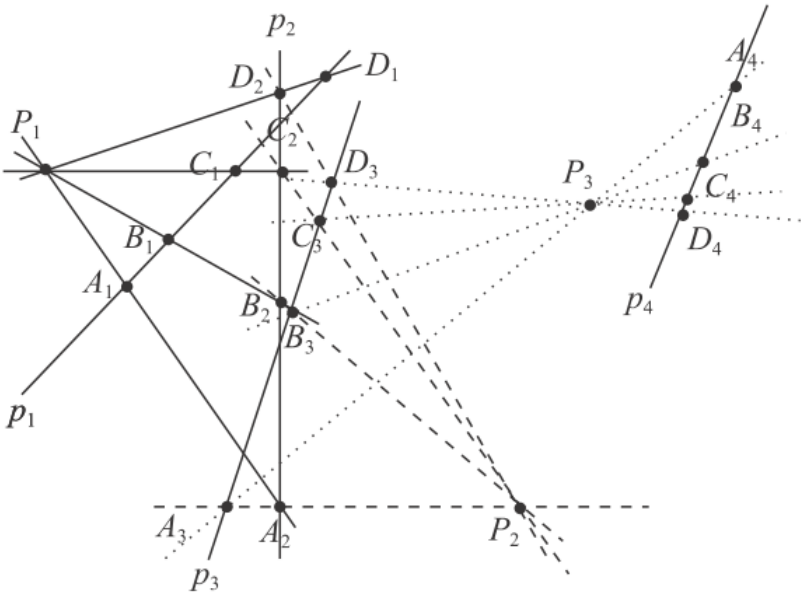

Figure 1.24: ph04.png - A projectivity composed of four perspectivities.

16) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important/S L Loney s Coordinate Geometry with Solutions of Examples Math Valley (PDFDrive.com).pdf

17) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important-02/475961.pdf

18) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important-02/Classical Geometries in Modern Contexts Geometry of Real Inner Product Spaces Third Edition (PDFDrive.com).pdf

19) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important-02/cmip08c.pdf
Obs.: Arithmetic Geometry

20) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important-02/Differential geometry for physicists and mathematicians moving frames and differential forms from Euclid past Riemann (PDFDrive.com).pdf

21) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important-

02/Fractal geometry, complex dimensions and zeta functions geometry and spectra of fractal strings (PDFDrive.com).pdf

22) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important-02/Geometry Revealed A Jacob's Ladder to Modern Higher Geometry (PDFDrive.com).pdf

23) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important-02/Integral Calculus 3D Geometry and Vector Booster with Problems and Solutions for IIT JEE Main and Advanced Rejaul Makshud McGraw Hill (PDFDrive.com).pdf

24) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important-02/mathmatics-1.pdf Pag. 408(423) - Chapter 7 - Curves, lengths, and curvatures Pag. 423(438) -

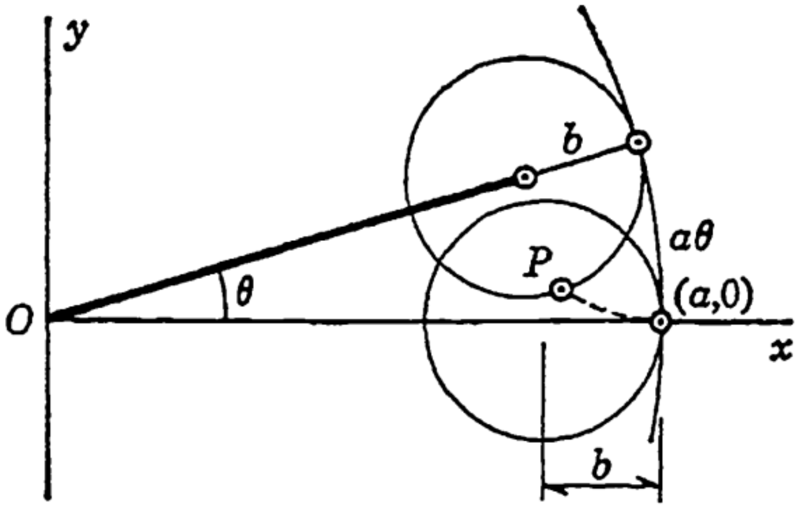

Figure 1.25: ph05.png - A projectivity composed of four perspectivities.

5 When 0 < b < a and a circle (or cog wheel) of radius b rolls without slipping inside a circle of radius a as in Figure 7.291, the path traced by a point on the small rolling circle is a hypocycloid (inside-cycloid) that appeared in one of Problems 6.59 and that even small dictionaries describe. It can be shown that when the center of the small circle has passed over an angle 0 (that could be wt), the point initially at (a,0) has coordinates

$$x = (a \quad b)\cos \quad + b\cos\frac{a \quad b}{b} \tag{1.3}$$

$$x = (a \quad b)\sin \quad + b\sin\frac{a \quad b}{b}: \tag{1.4}$$

With the aid of Figure 7.291 find the Figure 7.291 amount by which 0 must increase as P traverses one arch of the cycloid from the big circle back to the big circle. Then find the length of one arch of the cycloid. Ans.: 8b(a - b)/a.

25) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important-02/Methods of Geometry (PDFDrive.com).pdf

26) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important-02/Modern Differential Geometry for Physicists 2nd ed., - C. Isham(1).pdf

27) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important-02/Spacetime and Geometry An Introduction to General Relativity (PDFDrive.com)(1).pdf

28) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important-02/Spacetime and Geometry An Introduction to General Relativity (PDFDrive.com).pdf

29) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important-02/Spacetime Foundations of General Relativity and Differential Geometry (Lecture Notes in Physics) (PDFDrive.com).pdf

30) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important-02/The Fractal Geometry of Nature (PDFDrive.com).pdf

31) C:/DATA-NovoVivo-124-2020/Books-124-20220/Geometry-important-02/The Geometry of Minkowski Spacetime An Introduction to the Mathematics of the Special Theory of Relativity (PDFDrive.com).pdf

1.8 Gauss

Date: 20200804

1) C:/DATA-NovoVivo-125-2020/Books-125-2020/Handbook of Image Processing and Computer Vision-3-Volumes/Handbook of Image Processing and Computer Vision-Volume 1-From Energy to Image 2020.pdf

Pag. 28 (47) - Fig. 1.19 Electric and magnetic field in the incident wave, reflected and refracted by polarization perpendicular to the plane of incidence

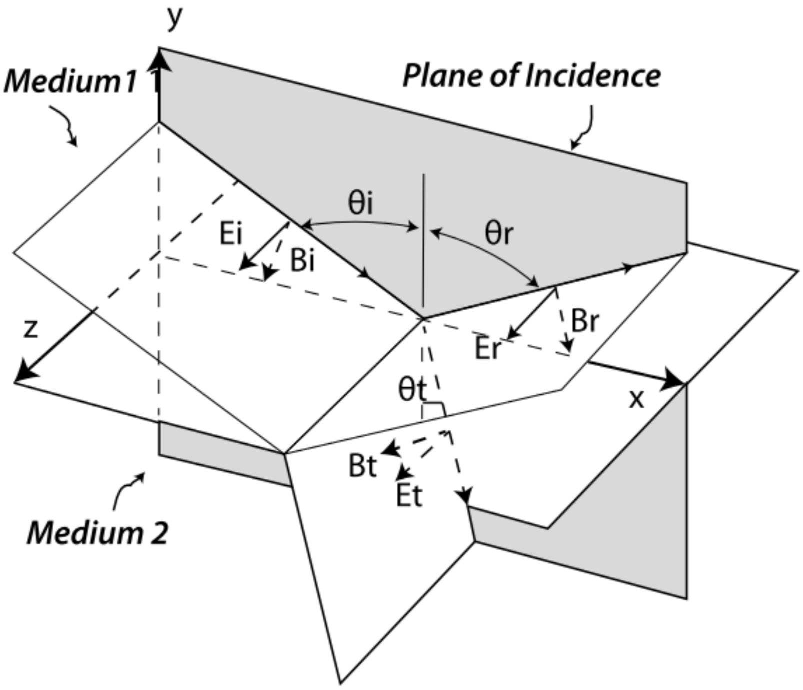

Figure 1.26: gauss01.png - Electric and magnetic field in the incident wave, reflected and refracted by polarization perpendicular to the plane of incidence

2) C:/DATA-NovoVivo-125-2020/Books-125-2020/Handbook of Image Processing and Computer Vision-3-Volumes/Handbook of Image Processing and Computer Vision-Volume 2-From Image to Pattern 2020.pdf

Pag. 24(43) - 1.11 Gaussian Derivative Operator (DroG) In the preceding paragraphs, we have already highlighted the problem of differen- tiation which introduces a attenuation of the signal noise ratio. With the smoothing operators described (see Sect.9.12.6 Vol. I) based on Gaussian filtering, the noise is attenuated. The DroG operator has the function of combining the noise reduction activities in the image and then executing the derivative with respect to the coordinate axes for edge detection. Remembering the associative property of the convolution operator, the gradient operator g DroG, based on the Gaussian derivative, is obtained by first applying the

Gaussian smoothing operator hG and then differentiating with the hD filter. The DroG operator in the horizontal component (indicated with r) is thus obtained:

3) C:/DATA-NovoVivo-125-2020/Books-125-2020/Handbook of Image Processing and Computer Vision-3-Volumes/Handbook of Image Processing and Computer Vision-Volume 3-From Pattern to Object 2020.pdf

Pag. 561(580) - 6.5.5 BS Method Based on Gaussian Mixture Model (GMM) So far methods have been considered where the background update model is based on the recent pixel history. Only with the Gaussian moving average method was the background modeled with the statistical parameters of average and variance of each pixel of the last images with the assumption of a unimodal Gaussian distribution. No spatial correlation was considered with the pixels in the vicinity of the one being processed. To handle more complex application contexts where the background scene includes structures with small movements not to be regarded as moving objects (for example, small leaf movements, trees, temporarily generated shadows, . . .) different methods have been proposed based on models of background with multimodal Gaussian distribution [31].

1.9 Energia escura

Date: 20200327

1) Energia escura: mapa da pistas sobre o quee, mas leva a novos enigmas indexEnergia escura: mapa da pistas sobre o quee, mas leva a novos enigmas

Figure 1.27: ee01.jpg - Energia Escura

Uma nova pesquisa trouxe mais informações sobre a misteriosa energia que preenche o espaço, mas alguns de seus dados entram em choque com teorias atualmente aceitas.

A energia escura é um dos maiores mistérios da ciência atualmente. Sabemos muito pouco sobre ela, afora ser invisível, preencher todo o universo e afastar as galáxias. Ela está fazendo nosso cosmos se expandir a uma taxa acelerada. Mas o que é ela? Uma das explicações mais simples é que é uma \constante cosmológica" { resultado da energia do próprio espaço vazio {, uma ideia introduzida por Albert Einstein.

Muitos físicos não estão satisfeitos com essa explicação, no entanto. Eles querem uma descrição mais fundamental de sua natureza. É algum novo tipo de campo de energia ou fluido exótico? Ou é um sinal de que as equações da gravidade de Einstein são de alguma forma incompletas? Além do mais, não entendemos realmente a atual taxa de expansão do universo.

Agora, nosso projeto { o Levantamento Espectroscópico de Oscilações Bariônicas Estendidas (eBOSS, na sigla em inglês) { apresentou algumas respostas. Nosso trabalho foi lançado como uma série de 23 publicações, algumas das quais ainda estão sendo revisadas por pares, descrevendo o maior mapa cosmológico tridimensional já criado.

Máquina do tempo

Atualmente, a única maneira de sentir a presença de energia escura é com observações do universo distante. Quanto mais distantes estão as galáxias, mais jovens elas aparecem para nós. Isso ocorre porque a luz que emitem levou milhões ou bilhões de anos para alcançar nossos telescópios. Graças a esse tipo de máquina do tempo, podemos medir diferentes distâncias no espaço em diferentes momentos cósmicos, ajudando-nos a descobrir a rapidez com que o universo está se expandindo.

Usando o telescópio Sloan Digital Sky Survey (SDSS), medimos mais de 2 milhões de galáxias e quasares { objetos extremamente brilhantes e distantes alimentados por buracos negros { nas últimas duas décadas. Esse novo mapa cobre cerca de 11 bilhões de anos de história cósmica que era essencialmente inexplorada, ensinando-nos sobre energia escura como nunca antes.

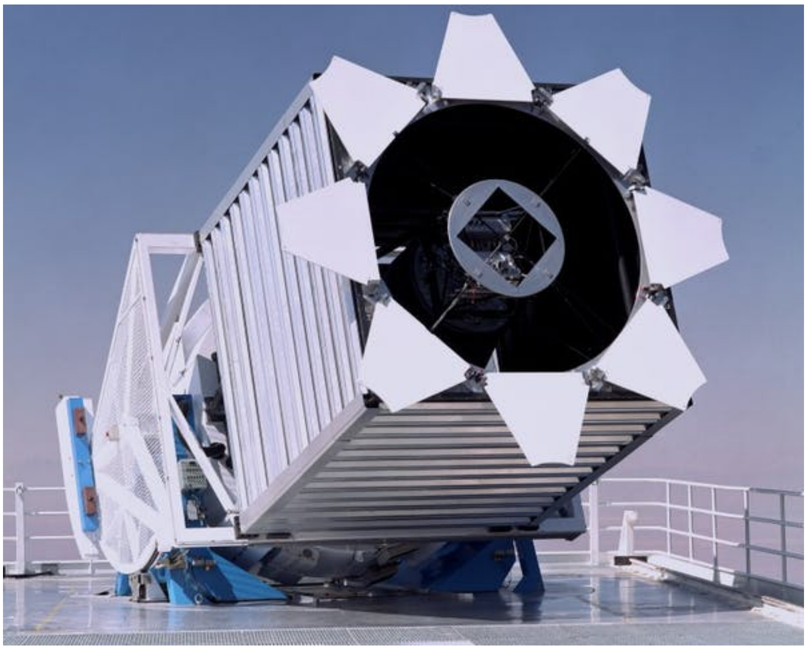

Figure 1.28: ee02.jpg - Energia Escura - Telescópio

Nossos resultados mostram que cerca de 69% da energia do nosso universo e energia escura. Eles também demonstram, mais uma vez, que a forma mais simples de energia escura de Einstein { a constante cosmológica { concorda mais com nossas observações.

Quando se combina as informações do nosso mapa com outras investigações cosmológicas, como o fundo cósmico de micro-ondas { a luz remanescente do Big Bang {, todos parecem preferir a constante cosmológica a explicações mais exóticas da energia escura. Expansão cósmica em disputa

Os resultados também fornecem uma visão melhor de algumas controvérsias recentes sobre a taxa de expansão do universo atual e sobre a geometria do espaço.

A combinação de nossas observações com estudos do universo em sua infância revela rachaduras em nossa descrição de sua evolução. Em particular, nossa medição da taxa atual de expansão do universo é cerca de 10% menor que o valor encontrado usando métodos diretos para medir distâncias a galáxias próximas. Ambos os métodos afirmam que o resultado é correto e muito preciso. Portanto, a diferença não pode ser simplesmente um acaso estatístico.

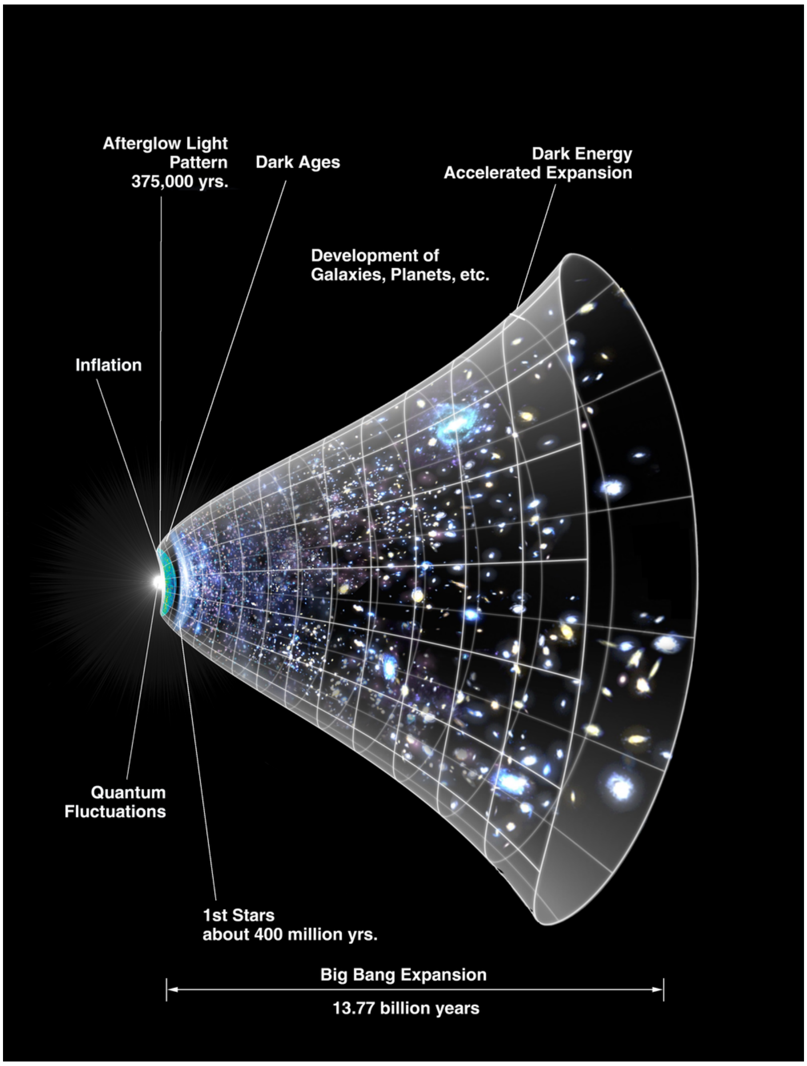
Afterglow Light Pattern 375,000 yrs.
Dark Ages
Dark Energy Accelerated Expansion
Development of Galaxies, Planets, etc.
Inflation
Quantum Fluctuations
1st Stars about 400 million yrs.
Big Bang Expansion
13.77 billion years

Figure 1.29: ee03.png - Energia Escura - Telescópio

A precisão do eBOSS aumenta essa crise. Não há uma explicação amplamente aceita para essa discrepância. Pode ser que alguém tenha cometido um erro sutil em um desses estudos. Ou pode ser um sinal de que precisamos de uma nova física. Uma possibilidade empolgante é que uma forma de matéria anteriormente desconhecida do universo primitivo tenha deixado um rastro em nossa história. Isso é conhecido como \energia escura precoce", que se pensa estar presente quando o universo era jovem, o que poderia ter modificado a taxa de expansão cósmica. Espaço plano

Estudos recentes do fundo cósmico de micro-ondas sugeriram que a geometria do espaço pode ser curvada em vez de simplesmente plana, o que é consistente com a teoria mais aceita do Big Bang. Mas nosso estudo concluiu que o espaço é realmente plano.IMPORTANT

Mesmo após esses importantes avanços, os cosmólogos de todo o mundo permanecerão intrigados com a aparente simplicidade da energia escura, a planura do espaço e os valores controversos da taxa de expansão atual. Existe apenas um caminho a seguir na busca de respostas { a criação de mapas maiores e mais detalhados do universo. Vários projetos têm como objetivo medir pelo menos dez vezes mais galáxias do que nós.

Se os mapas do eBOSS foram os primeiros a explorar uma lacuna anteriormente perdida de 11 bilhões de anos de nossa história, a nova geração de telescópios criará uma versão de alta resolução no mesmo período de tempo. É emocionante pensar no fato de que futuras pesquisas poderão resolver os mistérios restantes sobre a expansão do universo na próxima década. Mas seria igualmente emocionante se elas revelassem mais surpresas.

Julian Bautistaé pesquisador da Universidade de Portsmouth (Reino Unido).

Source:

a) https://www.revistaplaneta.com.br/energia-escura-mapa-da-pistas-sobre-o-que-e-mas-leva-a-novos-enigmas/

b) https://theconversation.com/us

c) https://theconversation.com/dark-energy-map-gives-clue-about-what-it-is-but-deepens-dispute-about-the-cosmic-expansion-rate-143200

Dark energy: map gives clue about what it is { but deepens dispute about the cosmic expansion rate

Dark energy is one of the greatest mysteries in science today. We know very little about it, other than it is invisible, it fills the whole universe, and it pushes galaxies away from each other. This is making our cosmos expand at an accelerated rate. But what is it? One of the simplest explanations is that it is a \cosmological constant" { a result of the energy of empty space itself { an idea introduced by Albert Einstein.

Many physicists aren't satisfied with this explanation, though. They want a more fundamental description of its nature. Is it some new type of energy field or exotic fluid? Or is it a sign that Einstein's equations of gravity are somehow incomplete? What's more, we don't really understand the universe's current rate of expansion.

Now our project - the extended Baryon Oscillation Spectroscopic Survey (eBOSS) - has come up with some answers. Our work has been released as a series of 23 publications, some of which are still being peer reviewed, describing the largest three-dimensional cosmological map ever created.

Currently, the only way we can feel the presence of dark energy is with observations of the distant universe. The farther galaxies are, the younger they appear to us. That's because the light they emit took millions or even billions of years to reach our telescopes. Thanks to this sort of time-machine, we can measure different distances in space at different cosmic times, helping us work out how quickly the universe is expanding.

Using the Sloan Digital Sky Survey telescope, we measured more than two million galaxies and quasars { extremely bright and distant objects that are powered by black holes { over the last two decades. This new map covers around 11 billion years of cosmic history that was essentially unexplored, teaching us about dark energy like never before.

Our results show that about 69% of our universe's energy is dark energy. They also demonstrate, once again, that Einstein's simplest form of dark energy - the cosmological constant - agrees the most with our observations.

When combining the information from our map with other cosmological probes, such as the cosmic microwave background { the light left over from the big bang { they all seem to prefer the cosmological constant over more exotic explanations of dark energy. Cosmic expansion in dispute

The results also provide a better insight into some recent controversies about the expansion rate of the universe today and about the geometry of space.

Combining our observations with studies of the universe in its infancy reveals cracks in our description of its evolution. In particular, our measurement of the current rate of expansion of the universe is about 10% lower than the value found using direct methods of measuring distances to nearby galaxies. Both these methods claim their result is correct and very precise, so their difference cannot simply be a statistical fluke.

The precision of eBOSS enhances this crisis. There is no broadly accepted explanation for this discrepancy. It may be that someone made a subtle mistake in one of these studies. Or it may be a sign that we need new physics. One exciting possibility is that a previously unknown form of matter from the early universe might have left a trace on our history. This is known as \early dark energy", thought to be present when the universe was young, which could have modified the cosmic expansion rate.

Recent studies of the cosmic microwave background suggested that the geometry of space may be curved instead of being simply flat { which is consistent with the most accepted theory of the big bang. But our study concluded that space is indeed flat.

Even after these important advances, cosmologists over the world will remain puzzled by the apparent simplicity of dark energy, the flatness of space and the controversial values of the expansion rate today. There is only one way forward in the quest for answers { making larger and more detailed maps of the universe. Several projects are aiming to measure at least ten times more galaxies than we did.

If the maps from eBOSS were the first to explore a previously missing gap of 11 billion years of our history, the new generation of telescopes will make a high-resolution version of the same period of time. It is exciting to think about the fact that future surveys may be able to resolve the remaining mysteries about the universe's expansion in the next decade or so. But it would be equally exciting if they revealed more surprises.

2) Matéria escura: o enigma se aprofunda indexMatéria escura: o enigma se aprofunda

Análise de 20 anos de dados mostra que enigmáticos sinais eletromagnéticos de galáxias próximas { a maior possibilidade de se identificar a matéria escura { não estão associados a ela.

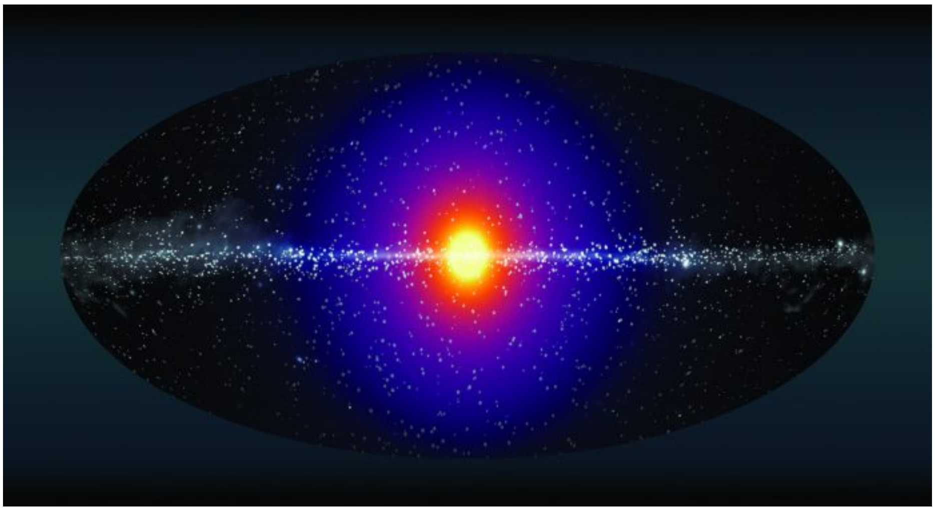

Figure 1.30: ee04.jpg - Matéria Escura - A matéria escura em decaimento deveria produzir um halo brilhante e esférico de emissão de raios X ao redor do centro da Via Láctea, passível de ser detectado quando se olha para regiões em branco da galáxia. Crédito: concepção artística de C. Dessert, N. L. Rodd, B. R. Safdi, Z. Rostomian (Berkeley Lab), com base em dados do telescópio espacial Fermi de raios gama

Nada menos do que 85% do universo é composto de matéria escura, mas não sabemos exatamente o que ela é.

Um novo estudo americano, com pesquisadores da Universidade de Michigan, do Lawrence Berkeley National Laboratory (Berkeley Lab) e da Universidade da Califórnia em Berkeley, descartou que a matéria escura seja responsável por misteriosos sinais eletromagnéticos anteriormente observados de galáxias próximas. Antes desse trabalho, publicado na revista “Science”, havia grandes esperanças de que tais sinais dariam aos físicos evidências concretas para ajudar a identificar a matéria escura.

A matéria escura não pode ser observada diretamente porque não absorve, reflete ou emite luz, mas os pesquisadores sabem que ela existe devido ao efeito que tem sobre a matéria visível. Precisamos da matéria escura para explicar as forças gravitacionais que mantêm as galáxias unidas, por exemplo.

Os físicos sugeriram que a matéria escura é um primo intimamente relacionado do neutrino, chamado neutrino estéril. Os neutrinos – partículas subatômicas sem carga e que raramente interagem com a matéria – são liberados durante as reações nucleares que ocorrem no interior do Sol. Eles têm uma pequena quantidade de massa, mas essa massa não é explicada pelo Modelo Padrão de Física de Partículas. Os físicos sugerem que o neutrino estéril, uma partícula hipotética, poderia explicar essa massa e também ser matéria escura.

Instabilidade

Os pesquisadores devem ser capazes de detectar o neutrino estéril porque ele é instável, diz Ben Safdi, professor assistente de física da Universidade de Michigan e coautor do estudo. Decai em neutrinos comuns e radiação eletromagnética. Para detectar a matéria escura, os físicos examinam galáxias em busca dessa radiação eletromagnética na forma de emissão de raios X.

Em 2014, um estudo descobriu excesso de emissão de raios X de galáxias próximas e aglomerados de galáxias. A emissão parece ser consistente com a que surgiria da decomposição da matéria escura estéril de neutrinos, disse Safdi.

Agora, uma metanálise de dados brutos de objetos na Via Láctea obtidos pelo telescópio espacial de raios X XMM-Newton, da Agência Espacial Europeia (ESA), durante um período de 20 anos não encontrou evidências de que o neutrino estéril seja o que compreende a matéria escura.

“Esse trabalho de 2014 e seu acompanhamento confirmaram que o sinal gerou um interesse significativo nas comunidades de astrofísica e física de partículas devido à possibilidade de saber, pela primeira vez, exatamente o que é a matéria escura em nível microscópico”, disse Safdi. “Nossa descoberta não significa que a matéria escura não seja um neutrino estéril, mas significa que – ao contrário do que foi afirmado em 2014 – não há evidências experimentais ate o momento que apontem para sua existência." Linha estudada

Os telescopios de raios X baseados no espaco, como o XMM-Newton, apontam para ambientes ricos em materia escura a fim de procurar essa fraca radiacao eletromagnetica na forma de sinais de raios X. A descoberta de 2014 denominou a emissao de raios X de \linha de 3,5 keV" (keV significa quilo eletron-volt) por causa de onde o sinal apareceu nos detectores de raios X.

Os pesquisadores estudaram essa linha na Via Lactea usando 20 anos de dados de arquivo obtidos pelo telescopio espacial de raios X XMM-Newton. Os fsicos sabem que a materia escura se acumula em torno das galaxias. Portanto, quando trabalhos anteriores analisaram galaxias proximas e aglomerados de galaxias, cada uma dessas imagens teria capturado alguma coluna do halo da materia escura da Via Lactea.

A equipe usou essas imagens para observar a parte \mais sombria" da Via Lactea. Isso melhorou significativamente a sensibilidade de analises anteriores a procura de materia escura esteril de neutrinos, disse Safdi.

\Para onde quer que olhemos, deve haver algum fluxo de materia escura do halo da Via Lactea" por causa da localizacao do Sistema Solar na galaxia, disse Nicholas Rodd, do Berkeley Lab. \Exploramos o fato de que vivemos em um halo de materia escura" no estudo, ele acrescentou. Sem evidências

Christopher Dessert, aluno de doutorado da Universidade de Michigan e coautor do estudo, disse que os aglomerados de galaxias onde a linha de 3,5 keV foi observada tambem têm grandes sinais de fundo, que servem como rudo nas observacoes e dificultariam a identificacao de sinais especficos que podem estar associados a materia escura.

\A razao pela qual estamos olhando atraves do halo de materia escura galactica da Via Lactea e que o fundo e muito menor", disse Dessert.

Por exemplo, o XMM-Newton capturou imagens de objetos isolados, como estrelas, na Via Lactea. Os pesquisadores capturaram essas imagens e mascararam os objetos de interesse original, deixando ambientes escuros e puros para procurar o brilho da decomposicao da materia escura. A combinacao de 20 anos de tais observacoes permitiu sondar a materia escura de neutrino esteril a nveis sem precedentes.

Se os neutrinos estereis eram materia escura, e se o seu decaimento levasse a uma emissao da linha de 3,5 keV, Safdi e seus colegas pesquisadores deveriam ter observado essa linha em suas analises. Mas eles nao encontraram evidências de materia escura esteril de neutrinos.

\Embora este trabalho, infelizmente, jogue agua fria no que parecia ser a primeira evidência da natureza microscopica da materia escura, ele abre uma nova abordagem para procurar materia escura que poderia levar a uma

descoberta em futuro proximo", disse Safdi.

Source: https://www.revistaplaneta.com.br/materia-escura-o-enigma-se-aprofunda/

3) Dark energy: new experiment may solve one of the universe's greatest mysteries. indexDark energy: new experiment may solve one of the universe's greatest mysteries Date: 20191028

As an astronomer, there is no better feeling than achieving \first light" with a new instrument or telescope. It is the culmination of years of preparations and construction of new hardware, which for the first time collects light particles from an astronomical object. This is usually followed by a sigh of relief and then the excitement of all the new science that is now possible.

On October 22, the Dark Energy Spectroscopic Instrument (DESI) on the Mayall Telescope in Arizona, US, achieved first light. This is a huge leap in our ability to measure galaxy distances { enabling a new era of mapping the structures in the universe. As its name indicates, it may also be key to solving one of the biggest questions in physics: what is the mysterious force dubbed \dark energy" that makes up the 70% of the universe?

The cosmos is clumpy. Galaxies live together in groups of a few to tens of galaxies. There are also clusters of a few hundreds to thousands of galaxies and superclusters that contain many such clusters.

This hierarchy of the universe has been known from the first maps of the universe, which looked like a \stickman" in graphs by the pioneering Centre for Astrophysics (CfA) Redshift Survey. These striking images were the first glimpse of large-scale structures in the universe, some spanning hundreds of millions of light years across.

The CfA survey was laboriously constructed one galaxy at a time. This involved measuring the spectrum of the galaxy light { a splitting of the light by wavelength, or colour { and identifying the fingerprints of certain chemical elements (mostly hydrogen, nitrogen and oxygen).

These chemical signatures are systematically shifted to longer redder wavelengths due to the expansion of the universe. This \red shift" was first detected by the astronomer Vesto Slipher and gave rise to the now famous Hubble's Law { the observation that more distant galaxies appear to be moving away at a faster rate. This means that galaxies that are close by appear to be moving away relatively slowly by comparison { they are less redshifted than galaxies far away. Therefore, measuring the redshift of a galaxy is a way to measure its distance.

Crucially, the exact relationship between redshift and distance depends on the expansion history of the Universe which can be calculated theoretically using our theory of gravity and our assumptions of the matter and energy density of the universe.

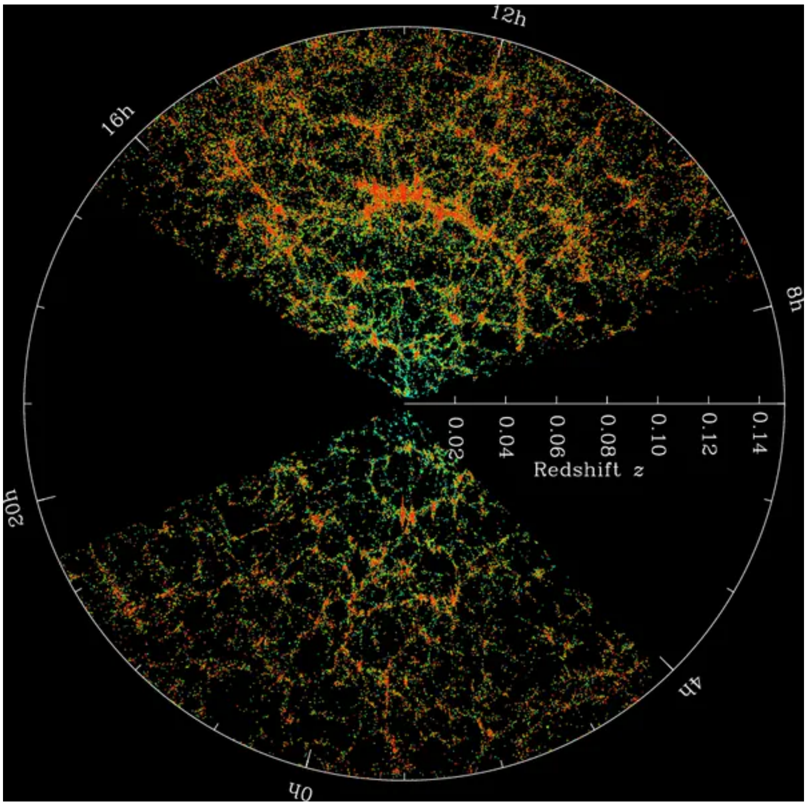

Figure 1.31: ee05.png - Dark Energy - SDSS map

All these assumptions were ultimately tested at the turn of the century with the combination of new observations of the universe, including new 3D maps from larger redshift surveys. In particular, the Sloan Digital Sky Survey (SDSS) was the first dedicated redshift survey telescope to measure over a million galaxy redshifts, mapping the large scale structure in the universe to unprecedented detail.

The SDSS maps included hundreds of superclusters and filaments and helped make an unexpected discovery { dark energy. They showed that the matter density of the universe was much less than expected from the Cosmic Microwave Background, which is the light left over from the Big Bang. That meant there must be an unknown substance, dubbed dark energy, driving an accelerated expansion of the Universe and become increasingly devoid of matter. The puzzle

The combination of all these observations heralded a new era of cosmological understanding with a universe consisting of 30% matter and 70% dark energy. But despite the fact that most physicists have now accepted that there is such a thing as dark energy, we still do not know its exact form.

There are several possibilities though. Many researchers believe that the energy of the vacuum simply has some particular value, dubbed a \cosmological constant". Other options include the possibility that Einstein's hugely successful theory of gravity is incomplete when applied on the huge scale of the entire universe.

New instruments like DESI will help take the next step in resolving the mystery. It will measure tens of millions of galaxy redshifts, spanning a huge volume of the universe up to ten billion light years from Earth. Such an amazing, detailed map should be able to answer a few key questions about dark energy and the creation of the large scale structures in the universe.

For example, it should be able to tell us if dark energy is just a cosmological constant. To do this it will measure the ratio of pressure that dark energy puts on the universe to the energy per unit volume. If dark energy is a cosmological constant, this ratio should be constant in both cosmic time and location. For other explanations, however, this ratio would vary. Any indication that it is not a constant would be revolutionary and spark intense theoretical work.

DESI should also be able to constrain, and even kill, many theories of modified gravity, possibly providing an emphatic confirmation of Einstein's Theory of General Relativity on the largest scales. Or the opposite { and again that would spark a revolution in theoretical physics.

Another important theory that will be tested with DESI is Inflation, which predicts that tiny random quantum fluctuations of energy density in the primordial universe were exponentially expanded during a short period of intense growth to become the seeds of the large scale structures we see today.

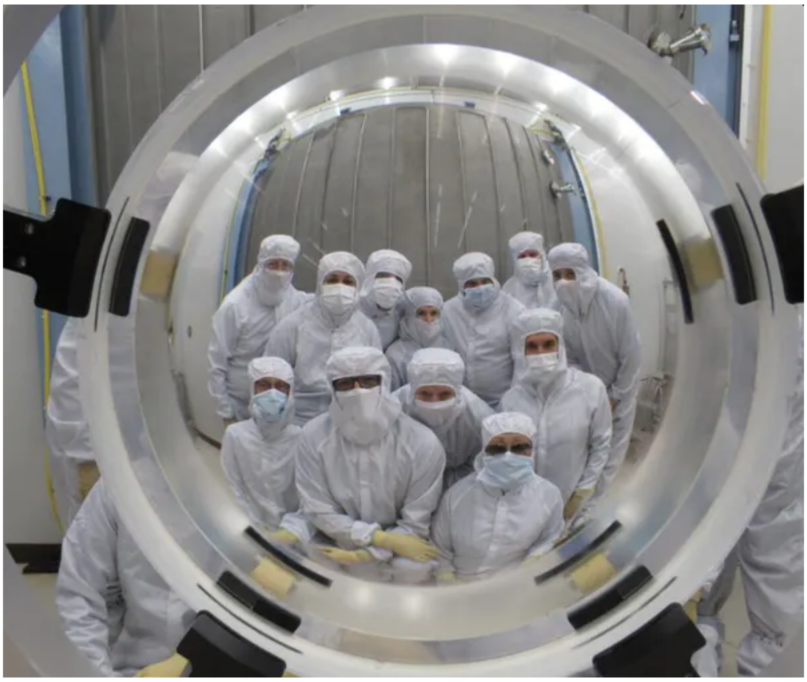

Figure 1.32: ee06.png - A team at a vendor in Santa Rosa Calif poses behind a DESI lens. VIAVI Solutions

DESI is only one of several next generation dark energy missions and experiments coming in the next decade, so there's certainly reason to be optimistic that we could soon solve the mystery of dark energy. New satellite missions like Euclid, and massive ground based observatories like the Large Synoptic Survey Telescope, will also offer insights.

There will also be other redshift instruments like DESI including 4MOST at the European Southern Observatory. Together, these will provide hundreds of millions of redshifts across the whole sky leading to an unimaginable map of our cosmos.

It seems a long time ago now when I wrote my PhD thesis based on just 700 galaxy redshifts. It really goes to show it's an exciting time to be an astronomer.

Source:

https://theconversation.com/dark-energy-new-experiment-may-solve-one-of-the-universes-greatest-mysteries-125886

4) How fast is the universe really expanding? New measurements help tackle the mystery indexHow fast is the universe really expanding? New measurements help tackle the mystery September 12, 2019 2.01pm EDT

Advances in astronomical observation over the past century have allowed scientists to construct a remarkably successful model of how the cosmos works. It makes sense { the better we can measure something, the more we learn. But when it comes to the question of how fast our universe is expanding, some new cosmological measurements are making us ever more confused.

Since the 1920s we've known that the universe is expanding { the more distant a galaxy is, the faster it is moving away from us. In fact, in the 1990s, the rate of expansion was found to be accelerating. The current expansion rate is described by something called \Hubble's Constant" { a fundamental cosmological parameter.

Until recently, it seemed we were converging on an accepted value for Hubble's Constant. But a mysterious discrepancy has emerged between values measured using different techniques. Now a new study, published in Science, presents a method that may help to solve the mystery. The problem with precision

Hubble's Constant can be estimated by combining measurements of the distances to other galaxies with the speed they are moving away from us. By the turn of the century, scientists agreed that the value was about 70 kilometres per second per megaparsec { one megaparsec is just over 3m light years. But in the last few years, new measurements have shown that this might not be a final answer.

If we estimate Hubble's Constant using observations of the local, present-day universe, we get a value of 73. But we can also use observations of the afterglow of the Big Bang { the \cosmic microwave background" { to estimate Hubble's Constant. But this \early" universe measurement gives a lower value of around 67.

Worryingly, both of the measurements are reported to be precise enough that there must be some sort of problem. Astronomers euphemistically refer to this as \tension" in the exact value of Hubble's Constant.

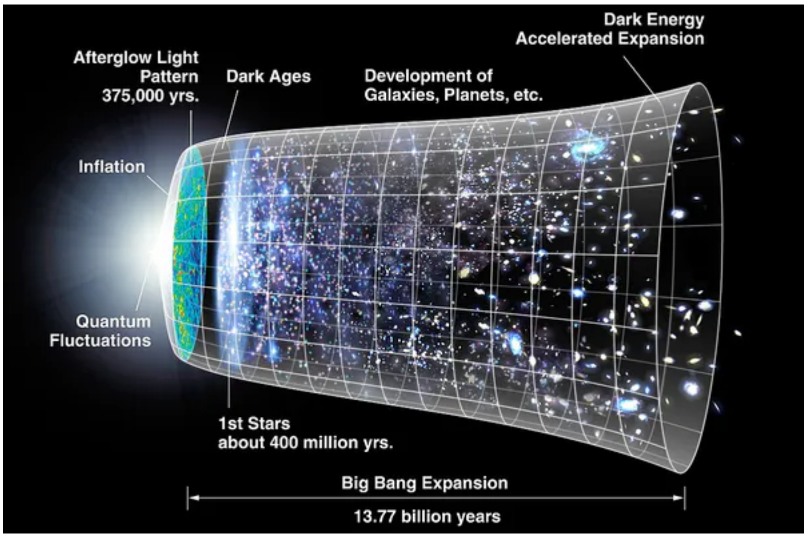
Dark Energy
Accelerated Expansion
Afterglow Light
Pattern
375,000 yrs.
Dark Ages
Development of
Galaxies, Planets, etc.
Inflation
Quantum
Fluctuations
1st Stars
about 400 million yrs.
Big Bang Expansion
13.77 billion years

Figure 1.33: ee07.png - Universe's expansion. NASA/WMAP

If you're the worrying kind, then the tension points to some unknown systematic problem with one or both of the measurements. If you're the excitable kind, then the discrepancy might be a clue about some new physics that we didn't know about before. Although it has been very successful so far, perhaps our cosmological model is wrong, or at least incomplete. Distant versus local

To get to the bottom of the discrepancy, we need a better linking of the distance scale between the very local and very distant universe.

The new paper presents a neat approach to this challenge. Many estimates of the expansion rate rely on the accurate measurement of distances to objects. But this is really hard to do: we can't just run a tape measure across the universe.

One common approach is to use \Type 1a" supernovas (exploding stars). These are incredibly bright, so we can see them at great distance. As we know how luminous they should be, we can calculate their distance by comparing their apparent brightness with their known luminosity.

To derive Hubble's Constant from the supernova observations, they must be calibrated against an absolute distance scale because there is still a rather large uncertainty in their total brightness. Currently, these \anchors" are very nearby (and so very accurate) distance markers, such as Cepheid Variable stars, which brighten and dim periodically.

If we had absolute distance anchors further out in the cosmos, then the supernova distances could be calibrated more accurately over a wider cosmic range. Far-flung anchors

The new work has dropped a couple of new anchors by exploiting a phenomenon called gravitational lensing. By looking at how light from a background source (like a galaxy) bends due to the gravity of a massive object in front of it, we can work out the properties of that foreground object.

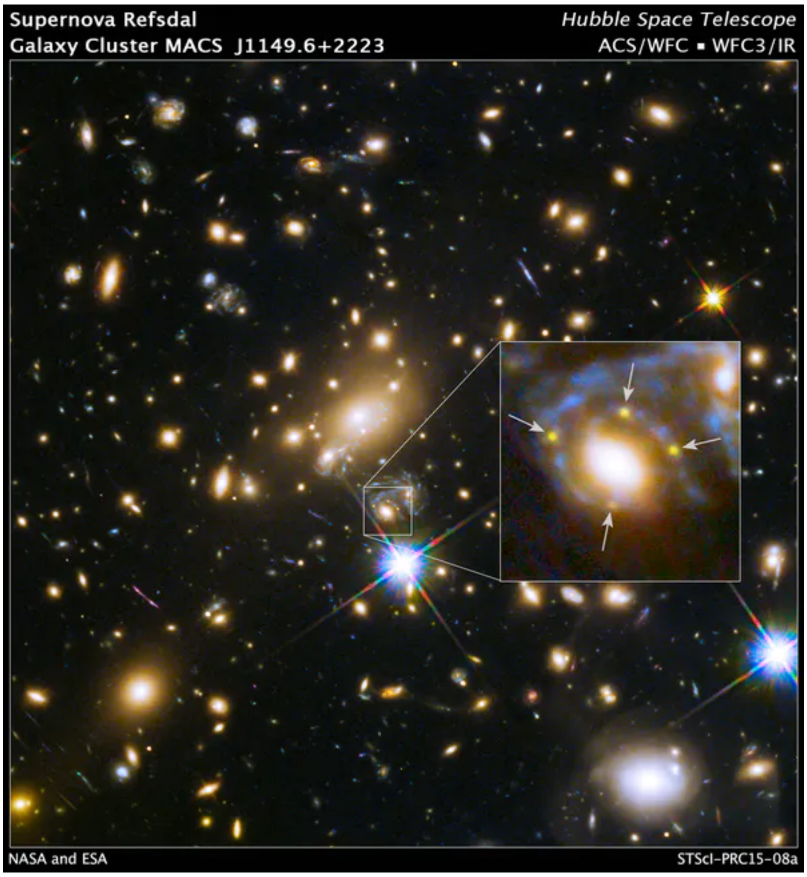
Supernova Refsdal
Galaxy Cluster MACS J1149.6+2223
Hubble Space Telescope
ACS/WFC ▪ WFC3/IR
NASA and ESA
STScI-PRC15-08a

Figure 1.34: ee08.png - A large cluster galaxy (centre of the box) has split the light from an exploding background supernova into four yellow images/dots. NASA/HUbble

The team has studied two galaxies that are lensing the light from two other background galaxies. The distortion is so strong that multiple images of each background galaxy are projected around the foreground deflectors (such as in the image above). The components of light making up each of those images will have travelled slightly different distances on their journey to Earth as the light bends around the foreground deflector. This causes a delay in the arrival time of light across the lensed image.

If the background source has a fairly constant brightness, we don't notice that time delay. But when the background source itself varies in brightness, we can measure the difference in light arrival time. This work does exactly that.

The time delay across the lensed image is related to the mass of the foreground galaxy deflecting the light, and its physical size. So when we combine the measured time delay with the mass of the deflecting galaxy (which we know) we get an accurate measure of its physical size.

Like a penny held at arms length, we can then compare the apparent size of the galaxy to the physical size to determine the distance, because an object of fixed size will appear smaller when it is far away. The authors present absolute distances of 810 and 1230 megaparsecs for the two deflecting galaxies, with about a 10-20% margin of error.

Treating these measurements as absolute distance anchors, the authors go on to reanalyse the distance calibration of 740 supernovas from a well-established data set used to determine Hubble's Constant. The answer they got was just over 82 kilometres per second per megaparsec.

This is quite high compared to the numbers mentioned above. But the key point is that with only two distance anchors the uncertainty in this value is still quite large. Importantly, though, it is statistically consistent with the value measured from the local universe. The uncertainty will be reduced by hunting for – and measuring – distances to other strongly lensed and time-varying galaxies. They are rare, but upcoming projects like the Large Synoptic Survey Telescope should be capable of detecting many such systems, raising hopes of reliable values.

The result provides another piece of the puzzle. But more work is needed: it still doesn't explain why the value derived from the cosmic microwave background is so low. So the mystery remains, but hopefully not for too long.

Source: https://theconversation.com/how-fast-is-the-universe-really-expanding-new-measurements-help-tackle-the-mystery-123338

5) Cientistas observam 'partícula de Deus' em nova interação. Date: 20200806

Primeira detecção de interação do bóson de Higgs com um múon é considerada uma importante transposição de fronteira científica Texto:Inovação Tecnológica (com informações da Agência Fapesp) 05/08/2020

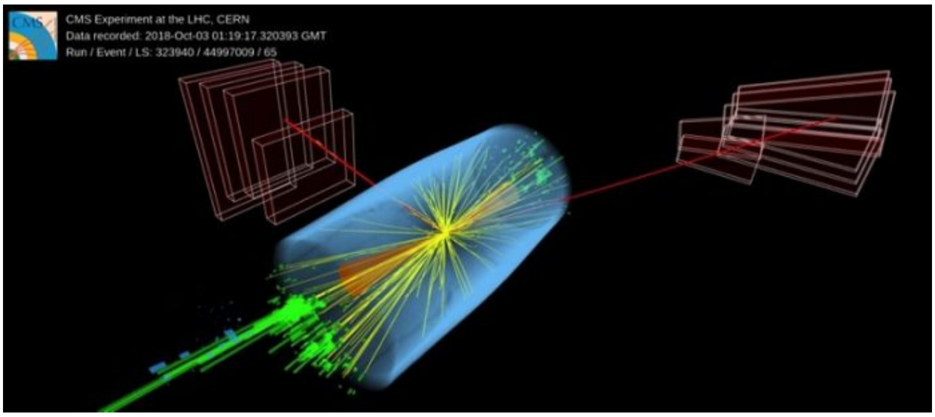
CMS Experiment at the LHC, CERN
Data recorded: 2018-Oct-03 01:19:17.320393 GMT
Run / Event / LS: 323940 / 44997009 / 65

Figure 1.35: Higgs01.png - Imagem real de um evento registrado pelo detector CMS do LHC. As duas linhas vermelhas correspondem ao par de múons formados pelo decaimento do bóson de Higgs. Crédito: Cern/divulgação

A primeira medida da interação do bóson de Higgs com o múon foi anunciada na segunda-feira, 3 de agosto, pela organização que controla o LHC (Grande Colisor de Hádrons).

Isso é importante porque abre caminho para o estudo de como o campo de Higgs – responsável por "dar materialidade à matéria", ou seja, dar-lhe massa – com as partículas mais leves que compõem o chamado Modelo Padrão da Física.

O múon é uma partícula semelhante ao elétron, com carga elétrica -1 e spin 1/2. Sua massa, embora muitas vezes superior à do elétron, é bem menor do que as massas das partículas com as quais o Higgs já havia interagido nos experimentos anteriores realizados no LHC.

O resultado, considerado um dos mais importantes obtidos desde a descoberta experimental do próprio bóson de Higgs em 2012, foi alcançado pela equipe internacional do CMS (Compact Muon Solenoid), um dos detectores do LHC. Análises trabalhosas

"O evento físico foi registrado em 2018, mas teve que passar por análises extremamente trabalhosas antes de ser reconhecido de maneira inequívoca. O fato de o CMS ter obtido essa medida complexa é mais uma demonstração do alcance do detector para explorar os detalhes do mecanismo de Higgs", disse à Agência Fapesp o físico Sérgio Novaes, professor titular da Universidade Estadual Paulista (Unesp) e líder do Sprace (São Paulo Research and Analysis Centre), um grupo de pesquisa atuante no LHC, apoiado pela Fapesp.

"Depois da descoberta do bóson de Higgs, as colaborações internacionais do Cern confirmaram experimentalmente as predições do Modelo Padrão quanto ao acoplamento dele com todas as partículas pesadas. Essa comparação foi estendida agora ao múon, um resultado bem mais difícil de alcançar," explicou Novaes.

Segundo o Modelo Padrão, é o campo de Higgs que fornece massa a todas as partículas conhecidas. O bóson de Higgs é o quantum, ou quantidade mínima de energia, desse campo.

No contexto experimental, o bóson de Higgs é obtido por meio da energia gerada em colisões de prótons. Como sua massa é muito grande, ele é instável e decai rapidamente em outras partículas mais leves. Por meio dos diferentes canais de decaimento, é possível calcular as forças de interação entre o campo de Higgs e as partículas resultantes.

E comparar os resultados com as previsões do Modelo Padrão permite testar a teoria e lança luz sobre o processo pelo qual as partículas fundamentais adquirem massa. Façanha científica

\O Higgs acopla-se a todas as partículas fundamentais com uma intensidade proporcional à massa dessas partículas. Assim, a interação é mais forte, e por isso mais fácil de medir, quando as partículas são bastante pesadas. É o caso dos bósons W e Z, intermediários da interação fraca, e dos férmions quarks top e bottom e lépton tau { sendo todos estes férmions partículas de terceira geração, cujo acoplamento com o Higgs já havia sido detectado anteriormente. O múon é um lépton de segunda geração, muito mais leve do que as partículas mencionadas. A detecção de seu acoplamento com o Higgs foi uma grande façanha científica," explicou Novaes.

A chance de um bóson de Higgs decair em múons é de uma em 5 mil, aproximadamente. Além disso, quando o decaimento realmente acontece, o par de múons resultante fica \escondido" no meio de mil outros pares de múons, gerados pelo decaimento de outras partículas envolvidas no processo. É preciso isolar e identificar cada um desses pares para saber se corresponde ou não ao resultado esperado.

\O acoplamento com o múon era considerado a próxima fronteira a ser explorada para o entendimento da interação do bóson de Higgs com as partículas mais leves. Essa fronteira foi agora transposta," comemorou Novaes. Source:

https://www.revistaplaneta.com.br/cientistas-observam-particula-de-deus-em-nova-interacao/

1.10 Geometry - Projective Geometry

1) C:/DATA-NovoVivo-126-2020/ Books-126-2020/ Roark-s-Formulas-for-Stress-and-Strain-9th-Edition/ Roark-s-Formulas-for-Stress-and-Strain-9th-Edition.pdf

1a) Pag. 37(57) - 2.6 MOHR'S CIRCLES FOR 30 STRESS ANALYSIS
Pag. 39(59) - Figure 2.20

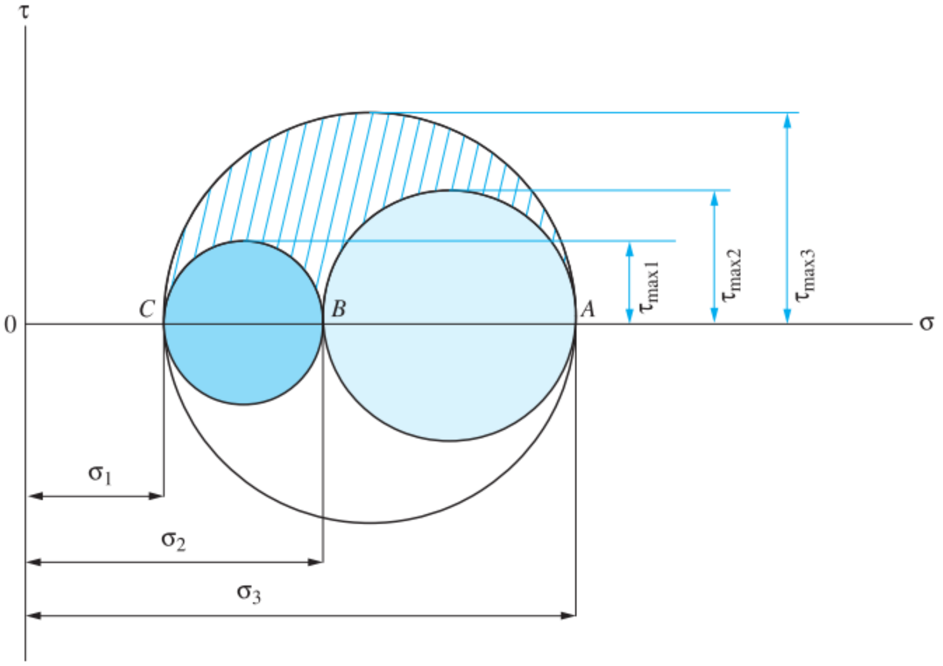

Figure 1.36: F01.png - The 3D representation of Mohr's circle.

1b) Pag. 51(71) - The element with these normal stresses in an octahedron where the eight surfaces are symmetric with respect to the principal axes. The directional cosines of the normal of these surfaces, relative to the principal axes, are eight combinations of $\frac{1}{\sqrt{3}}$ (e.g., one set is $\frac{1}{\sqrt{3}}$, $\frac{1}{\sqrt{3}}$, $\frac{1}{\sqrt{3}}$; another is $\frac{1}{\sqrt{3}}$, $\frac{1}{\sqrt{3}}$, $\frac{1}{\sqrt{3}}$, etc).

The octahedron is as shown in Fig. 3.1. The shear stresses on these surfaces are also equal, called the octahedral shear stresses, and are given by

$$_{oct} = \frac{1}{3}\sqrt{(\;_1 \quad _2)^2 + (\;_2 \quad _3)^2 + (\;_3 \quad _1)^2} \tag{1.5}$$

Again, for the simple tensile test, yield occurs when$_1$ = $_{ys}$ and $_2$ = $_3$ = 0. From Eq. (3.7-6), this gives an octahedral shear stress at yield of

$$(\ _{oct})_y = \frac{\sqrt{2}}{3}\ _{ys} \tag{1.6}$$

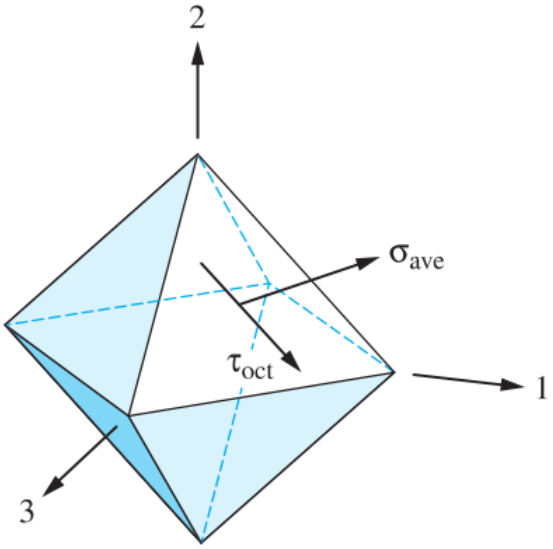

Figure 1.37: F02.png - Octahedral surfaces containing octahedral shear stresses (shown relative to the principal axes, with only one set of stresses displayed).

Equating Eqs. (3.7-6) and (3.7-7) results in Eq. (3.7-3) again, proving that the maximum-octahedral-shear-stress theory is identical to the maximum-distortion-energy theory. The maximum-shear-stress theory equates the maximum shear stress for a general state of stress to the maximum shear stress obtained when the tensile specimen yields. If the principal stresses are ordered such that $_1 \quad _2 \quad _3$, the maximum shear stress is given by:0.5($_1 \quad _3$) (see Sec. 2.3, Eq. 2.3-25). The maximum shear stress obtained when the tensile specimen yields is:0.5 $_{ys}$:
Thus; the condition for elastic failure for the maximum shear stress theory is [7]

---

[7] Plane stress problems are encountered quite often where the principal stresses are found from Eq. (2.3-23), which is

$$_p = \frac{_x + _y}{2} \qquad \sqrt{\left(\frac{_x \quad _y}{2}\right)^2 + \ ^2_{xy}}$$

This yields only two of the three principal stresses. The third principal stress for plane is

$$ {}_1 \quad {}_3 = \quad {}_{ys} \tag{1.7} $$

The criteria just discussed concern the elastic failure of material. Such failure may occur locally in a member and may do no real damage if the volume of material affected is so small or so located as to have only negligible influence on the form and strength of the member as a whole. Whether or not such local overstressing is significant depends upon the properties of the material and the conditions of service. Fatigue properties, resistance to impact, and mechanical functioning are much more likely to be affected than static strength, and a degree of local overstressing that would constitute failure in a high-speed machine part might be of no consequence whatever in a bridge member. A brittle material cannot be considered to have definitely failed until it has broken, which can occur either through a tensile fracture, when the maximum tensile stress reaches the ultimate strength, or through what appears to be a shear fracture, when the maximum compressive stress reaches a certain value. The fracture occurs on a plane oblique to the maximum compressive stress but not, as a rule, on the plane of maximum shear stress, and so it cannot be considered to be purely a shear failure (see Ref. 14). The results of some tests on glass and Bakelite (Ref. 26) indicate that for these brittle materials either the maximum stress or the maximum strain theory affords a satisfactory criterion of rupture while neither the maximum shear stress nor the constant energy of distortion theory does. These tests also indicate that strength increases with rate of stress application and that the increase is more marked when the location of the most stressed zone changes during the loading (pressure of a sphere on a flat surface) than when this zone is fixed (axial tension). Another failure theory that is applicable to brittle materials is the Coulomb-Mohr theory of failure. Brittle materials have ultimate compressive strengths ${}_{uc}$ greater than their ultimate tensile strengths ${}_{ut}$ and therefore both a uniaxial tensile test and a uniaxial compressive test must be run to use the Coulomb-Mohr theory. First we draw on a single plot both Mohr's stress circle for the tensile test at the instant of failure and Mohr's stress circle for the compressive test at the instant of failure; then we complete a failure envelope simply by drawing a pair of tangent lines to the two circles, as shown in Fig. 3.2. Failure under a complex stress situation is expected if the largest of the three Mohr circles for the given situation touches or extends outside the envelope just described. If

---

zero. Once the three principal stresses are determinid, they can be ordered according to

$$ 1 \quad 2 \quad 3 $$

and the Eq. (3.7-8) can be employed.

all normal stresses are tensile, the results coincide with the maximum stress theory. For a

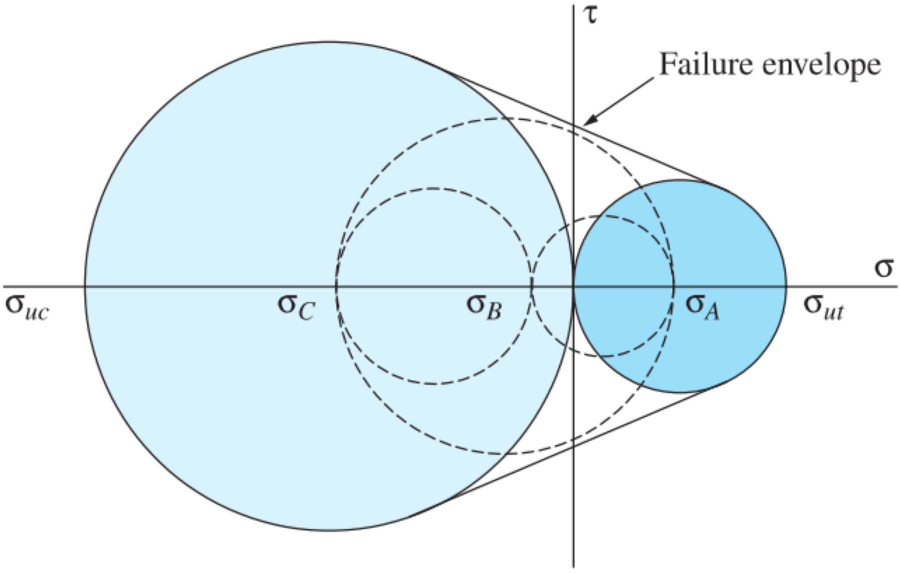

Figure 1.38: F03.png - Failure envelope

condition where the three principal stresses are $_A$, $_B$, and $_C$, as shown in Fig. 3.2, failure is being approached but will not take place unless the dashed circle passing through $_A$ and $_C$ reaches the failure envelope. The accurate prediction of the breaking strength of a member composed of brittle metal requires a knowledge of the effect of form and scale, and these effects are expressed by the rupture factor (see Sec. 3.11). In addition, what has been said here concerning brittle metals applies also to any essentially isotropic brittle material. Thus far, our discussion of failure has been limited to isotropic materials. For wood, which is distinctly anisotropic, the possibility of failure in each of several ways and direc- tions must be taken into account, viz.: (1) by tension parallel to the grain, which causes fracture; (2) by tension transverse to the grain, which causes fracture; (3) by shear parallel to the grain, which causes fracture; (4) by compression parallel to the grain, which causes gradual buckling of the fibers usually accompanied by a shear displacement on an oblique plane; and (5) by compression transverse to the grain, which causes sufficient deformation to make the part unfit for service. The unit stress producing each of these types of failure must be ascertained by suitable tests (Ref. 2). Another anisotropic class of material of consequence is that of the composites. It is well-known that composite members (see Sees. 7.3, 8.2, and definitions), such as steel reinforced concrete beams, more effectively utilize the more expensive, higher-strength materials in high-stress areas and the less expensive, lower-strength materials in the low-stress areas. Composite materials accomplish the same effect at microstructural and macrostructural levels. Composite materials come in many forms, but are generally formulated by embedding a reinforcement material in the form of fibers, flakes, particles, or laminations, in a randomly or orderly oriented fashion within a base matrix of polymeric, metallic, or ceramic material. For more detail properties of composites, see Ref. 60.

2) C:/DATA-NovoVivo-126-2020/Books-126-2020/Principles-and-Practice-of-Physics-2nd-Edition/Principles and Practice of Physics 2nd Edition.pdf

Pag. 443 (465) - Special Relativit

3) C:/DATA-NovoVivo-126-2020/Books-126-2020/New-Scientist/2020-08-08 New Scientist.pdf

Pag. 46(49) - Knot too difficult Mathematician Lisa Piccirillo untangled a problem that had tied up fellow mathematicians for decades. She tells Chelsea Whyte how she did it.

4) C:/DATA-NovoVivo-126-2020/Books-126-2020/Advanced Micro- and Nanomaterials for Photovoltaics/Advanced-Micro{and-Nanomaterials-for-Photovoltaics.pdf

Pag. 56 (64) - The generalized Planck law is first recalled

5) C:/DATA-NovoVivo-126-2020/Books-126-2020/9781439819111-Ancillaries/K11008-Figure-Slides/K11008-A001/PDF/K11008-A001.pdf

Pag. 01 - Desargues' theorem

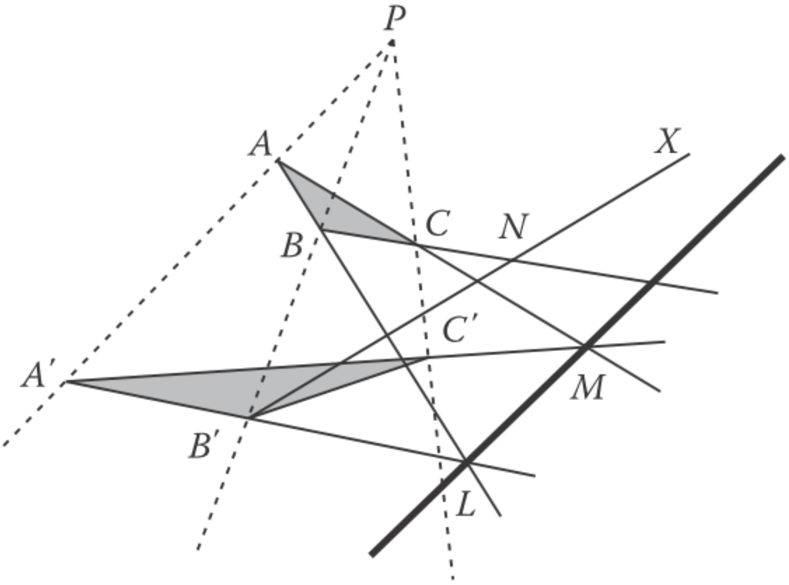
P
A
X
C
N
B
C′
A′
M
B′
L

Figure 1.39: D01.png - Desargues' theorem

Pag. 06 - Euler line of ABC IMPORTANT Three points in line.

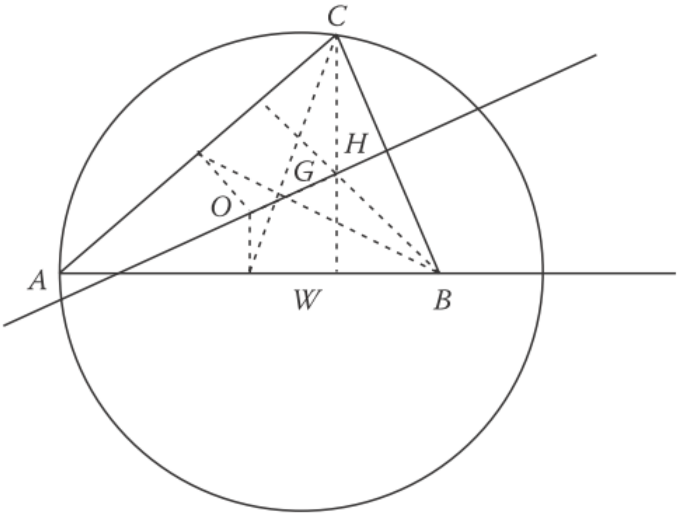

Figure 1.40: D02.png - Euler line of ABC

Pag. 15 - Rotating the Simson line.

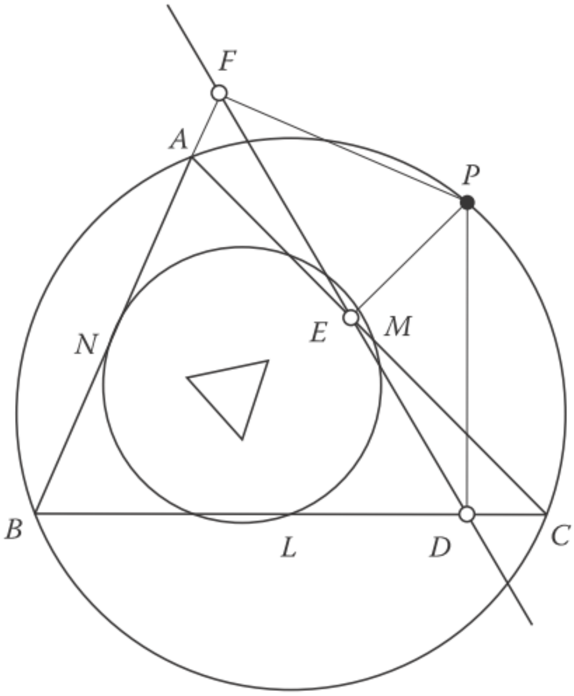

Figure 1.41: D03.png - Rotating the Simson line

Pag. 16 - Generating the families F and G.

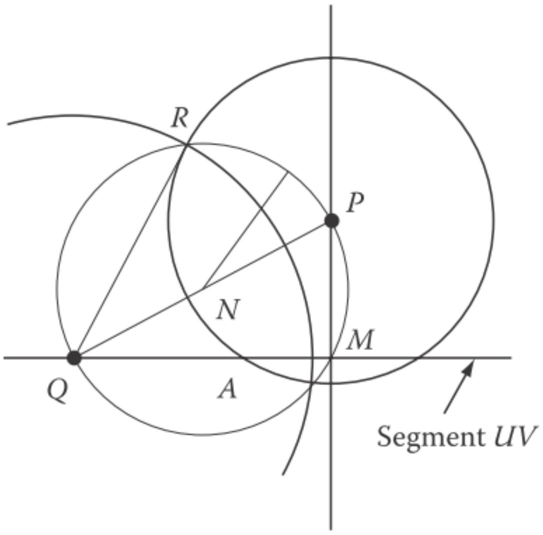

Figure 1.42: D04.png - Generating the families F and G

6) C:/DATA-NovoVivo-126-2020/Books-126-2020/9781439819111-Ancillaries/K11008-
A002.pdf

Pag. 01 - The distance from P to Q on the sphere.

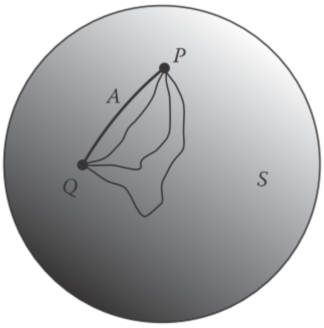

Figure 1.43: D05.png - The distance from P to Q on the sphere.

7) C:/DATA-NovoVivo-126-2020/Books-126-2020/9781439819111-Ancillaries/K11008-
C001.pdf

Pag. 20 - Spherical segments and rays.

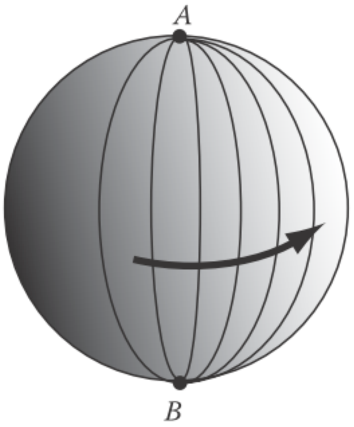

Figure 1.44: D06.png

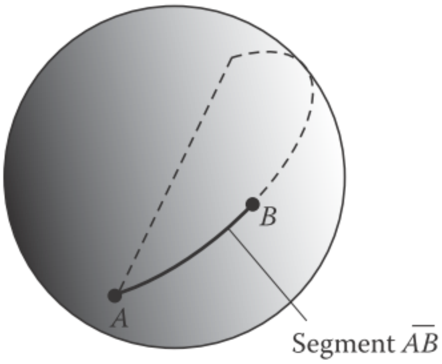
B
A
Segment $\overline{AB}$

Figure 1.45: D07.png

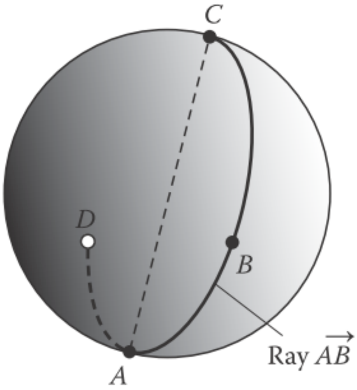
C
D
B
Ray $\overrightarrow{AB}$
A

Figure 1.46: D08.png

8) C:/DATA-NovoVivo-126-2020/Books-126-2020/9781439819111-Ancillaries/K11008-C002.pdf

Pag. 06 - A great circle path from A to B.

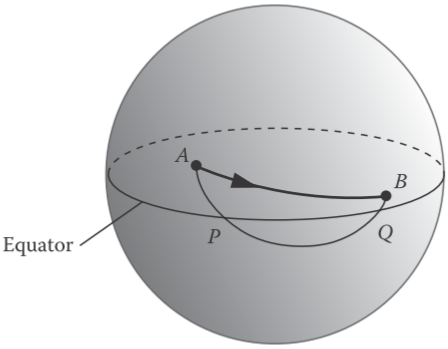
A
B
P
Q
Equator

Figure 1.47: D09.png - A great circle path from A to B.

Pag. 08 - Spherical angle measure.

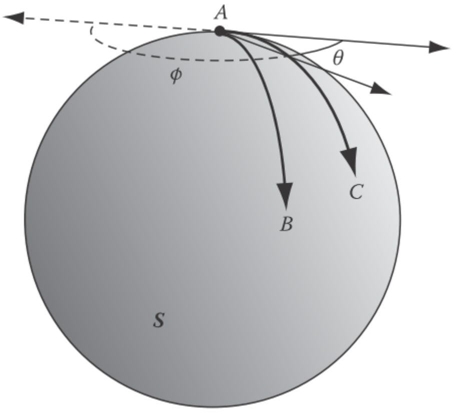

Figure 1.48: D10.png - Spherical angle measure.

Pag. 21 - Interior of a spherical angle.

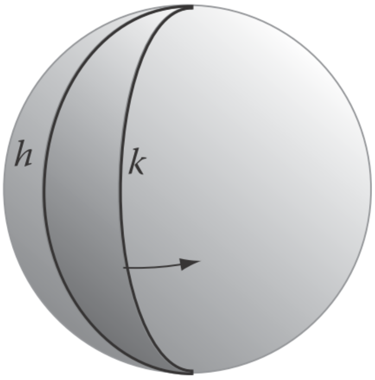

Figure 1.49: D11.png

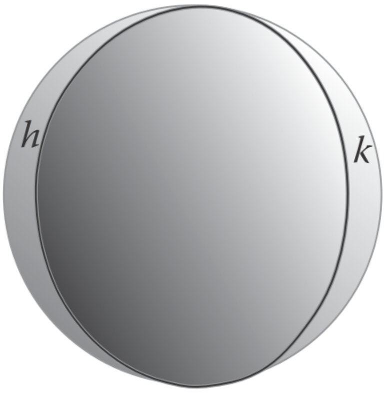

Figure 1.50: D12.png

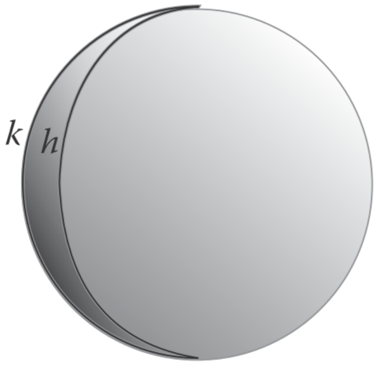

Figure 1.51: D13.png

9) C:/DATA-NovoVivo-126-2020/Books-126-2020/9781439819111-Ancillaries/K11008-C003.pdf

Pag. 27 - Spherical Reflection

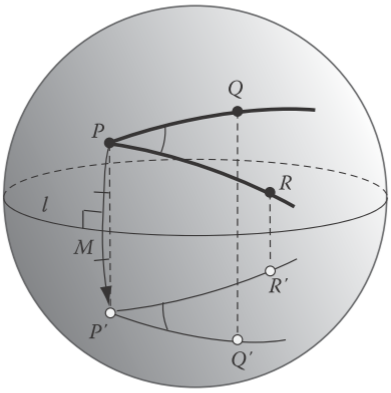

Figure 1.52: D14.png - Spherical Reflection

Pag. 32 - The distance from point A to line l

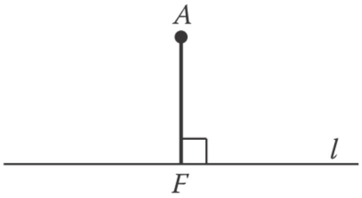

Figure 1.53: D15.png

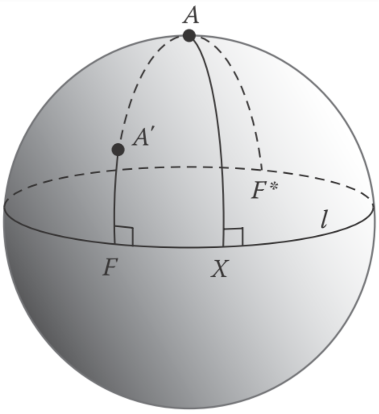

Figure 1.54: D16.png

Pag. 52 - Cevian inequality.

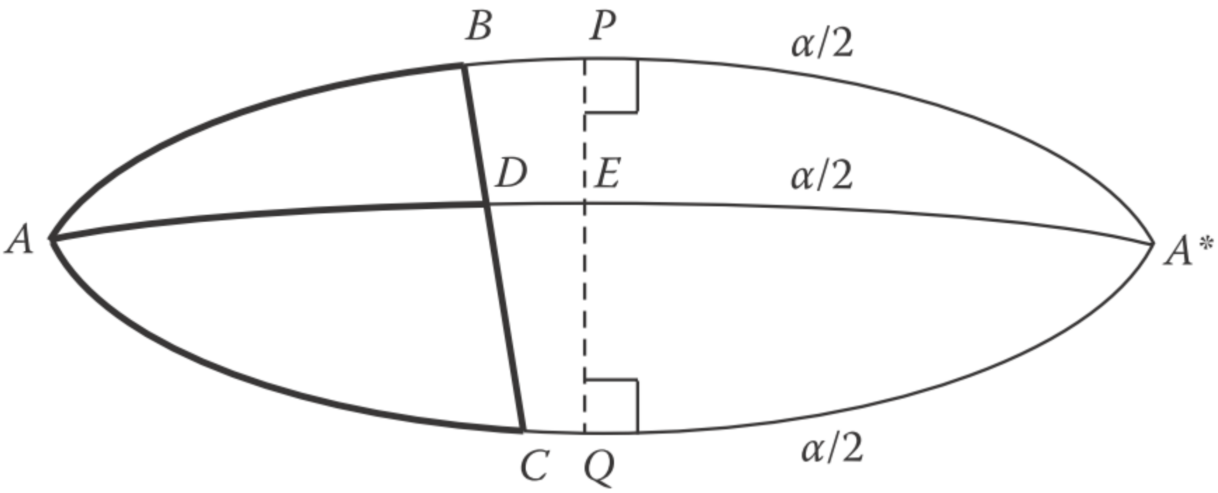

Figure 1.55: D17.png - Cevian inequality

10) C:/DATA-NovoVivo-126-2020/Books-126-2020/9781439819111-Ancillaries/K11008-C004.pdf

Pag. 09 - Saccheri quadrilateral on the sphere.

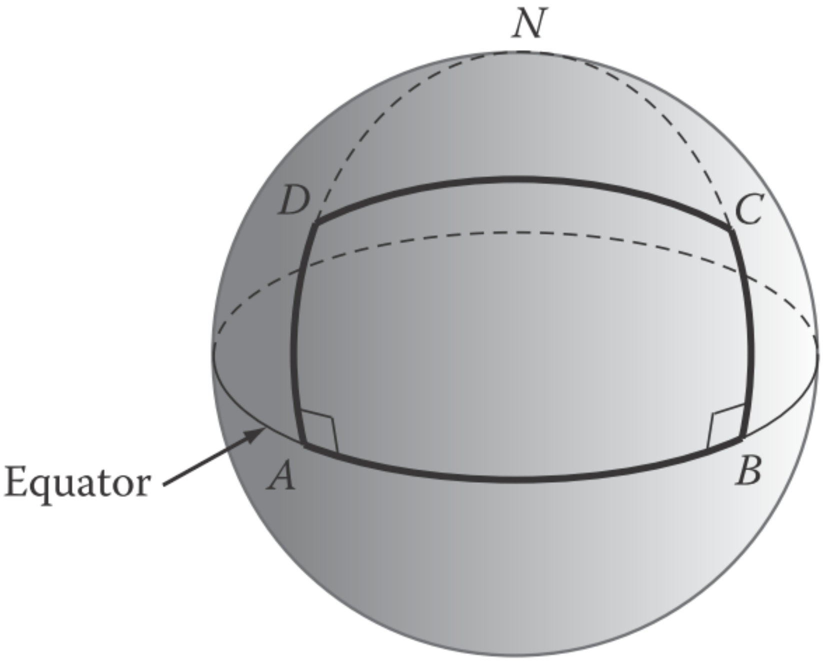
N
D
C
A
B
Equator

Figure 1.56: D18.png - Cevian inequality

Pag. 16 - Proving Saccheri's Theorem.

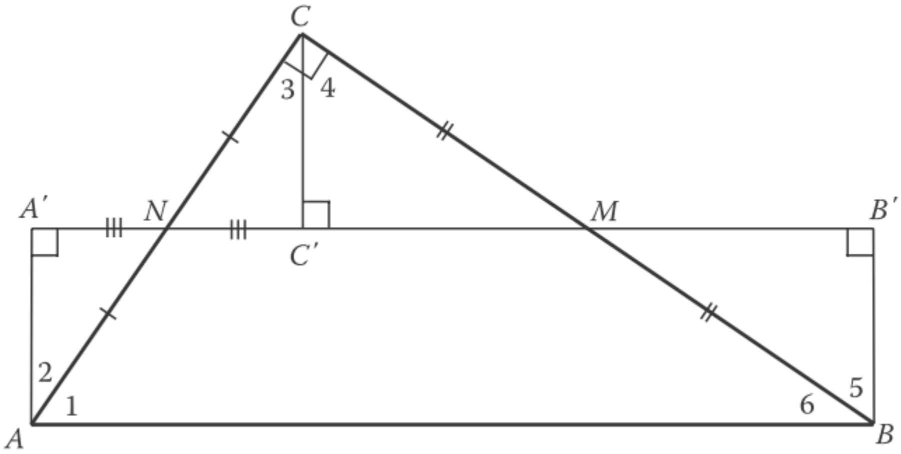
C
3
4
A′
N
M
B′
C′
2
1
6
5
A
B

Figure 1.57: D19.png - Proving Saccheri's Theorem

Pag. 32 - Pictorial glossary of terms associated with circle.

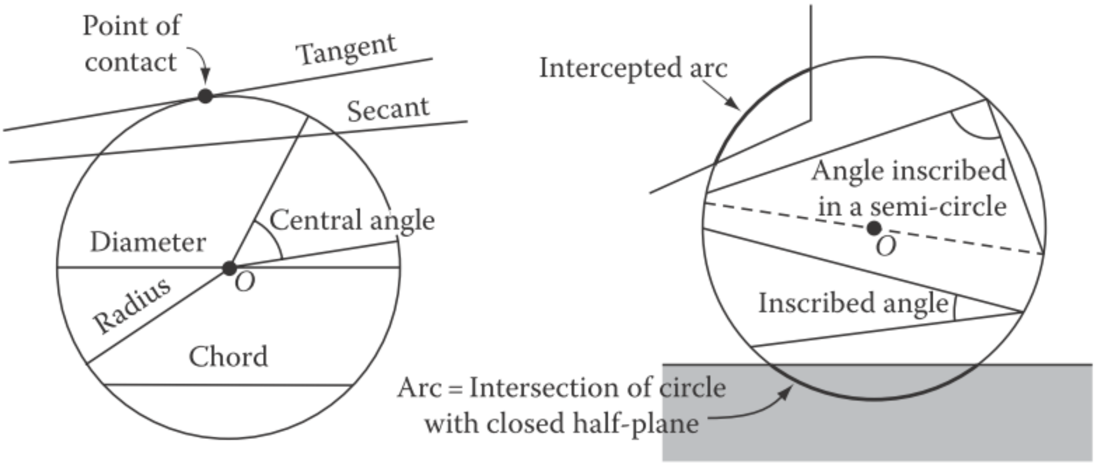
Point of contact
Tangent
Secant
Central angle
Diameter
O
Radius
Chord
Intercepted arc
Angle inscribed in a semi-circle
O
Inscribed angle
Arc = Intersection of circle with closed half-plane

Figure 1.58: D20.png - Pictorial glossary of terms associated with circle

Pag. 33 - A circle on a sphere.

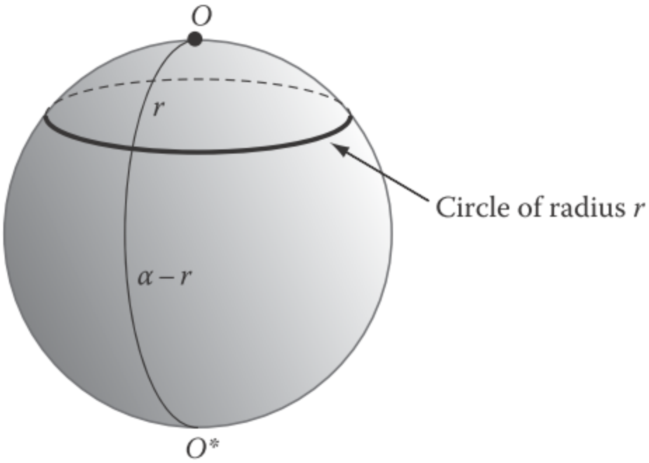
O
r
Circle of radius r
α − r
O*

Figure 1.59: D21.png - A circle on a sphere.

Pag. 36 - Convexity of Int C and INT C.

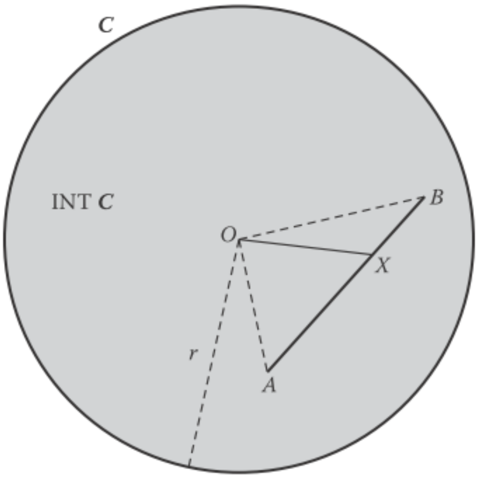

Figure 1.60: D22.png - Convexity of Int C and INT C

Pag. 39 - Inscribed triangle and circular arcs.
The four shaded regions devoid of points of the circle
The three shaded regions containing the circle

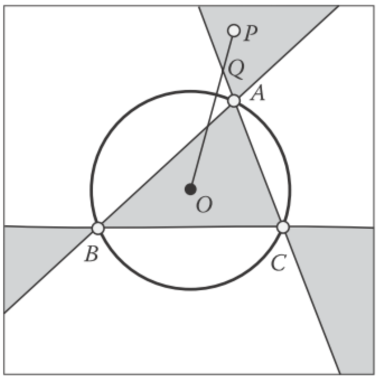

Figure 1.61: D23.png - The four shaded regions devoid of points of the circle

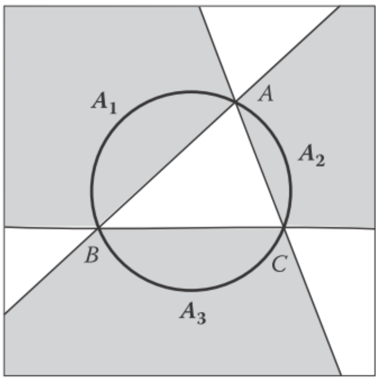

Figure 1.62: D24.png - The three shaded regions containing the circle

11) C:/DATA-NovoVivo-126-2020/Books-126-2020/9781439819111-Ancillaries/K1100
C005.pdf

Pag. 06 - Triangles having identical angle-sums.

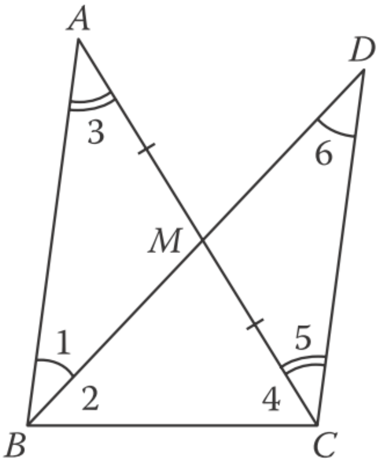
A
D
3
6
M
1
5
2
4
B
C

Figure 1.63: D25.png - Triangles having identical angle-sums 1

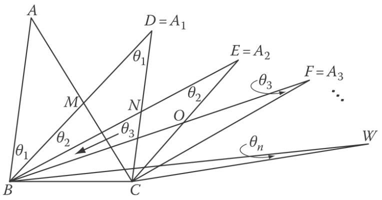
A
$D = A_1$
$E = A_2$
$F = A_3$
θ_1
θ_2
θ_3
M
N
O
θ_3
θ_2
θ_1
W
θ_n
B
C

Figure 1.64: D26.png - Triangles having identical angle-sums 2

Pag. 07 - Testing a theory.

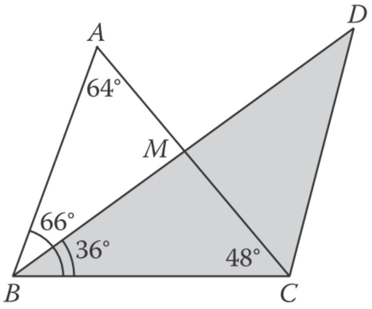
A
D
64°
M
66°
36°
48°
B
C

Figure 1.65: D27.png - Testing a theory 1

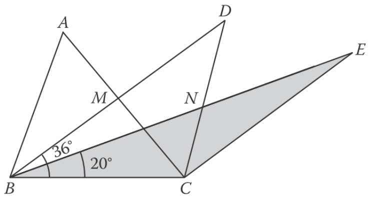
A
D
E
M
N
36°
20°
B
C

Figure 1.66: D27.png - Testing a theory 2

Pag. 19 - Testing equation

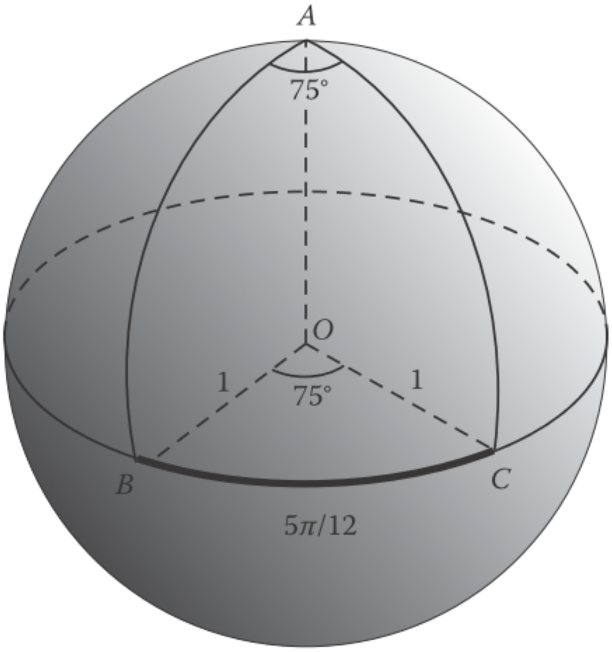
A
75°
O
1
75°
1
B
C
5π/12

Figure 1.67: D29.png - Testing equation

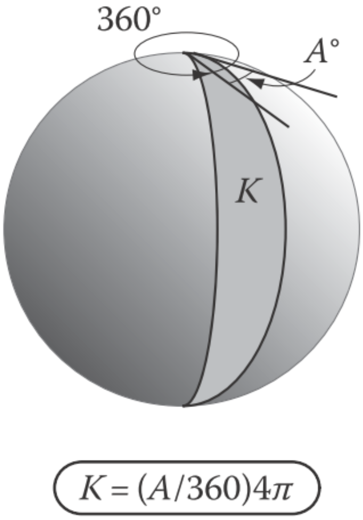
360°
A°
K
K = (A/360)4π

Figure 1.68: D30.png

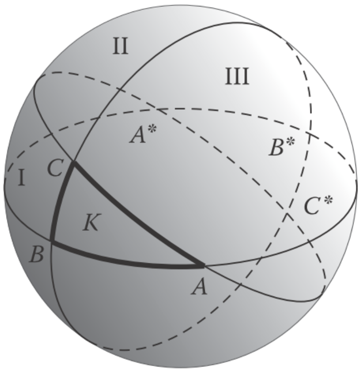
II
III
A*
B*
I
C
C*
K
B
A

Figure 1.69: D31.png

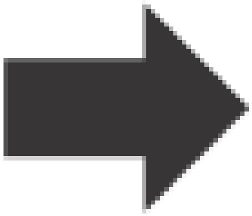

Figure 1.70: D32.png

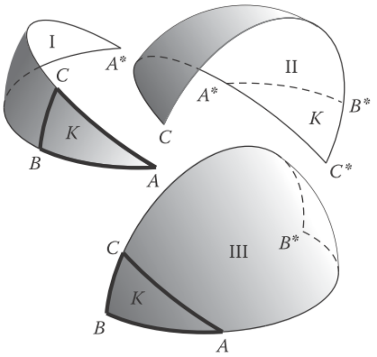
I
A*
C
K
B
A
II
A*
K
B*
C
C*
C
III
B*
K
B
A

Figure 1.71: D33.png

Pag. 25 - Geometry on a hyperbolic paraboloid.

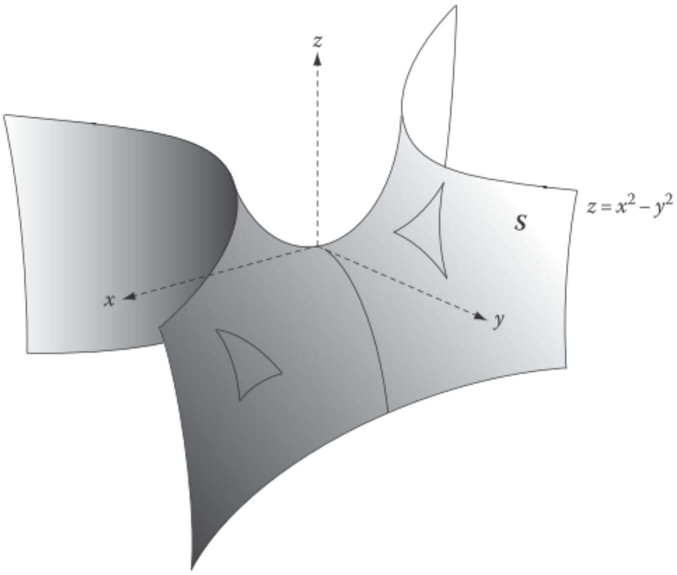

Figure 1.72: D34.png - Geometry on a hyperbolic paraboloid

Pag. 27 - Proclus' argument.

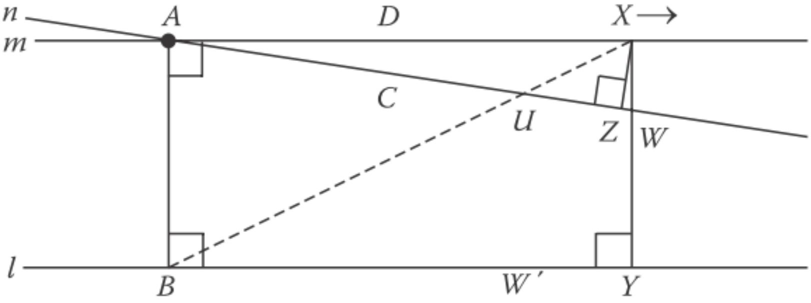
n
m
A
D
X ⟶
C
U
Z
W
l
B
W'
Y

Figure 1.73: D35.png - Proclus' argument

Pag. 28 - Proofs of Corollaries A and B.

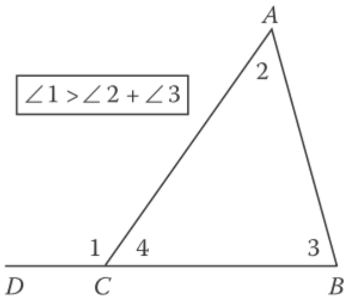
∠1 > ∠2 + ∠3
A
2
1
4
3
D
C
B

Figure 1.74: D36.png - Proofs of Corollaries A and B - 1

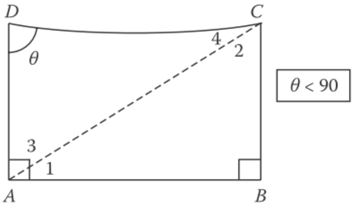
D
C
θ
4
2
θ < 90
3
1
A
B

Figure 1.75: D37.png - Proofs of Corollaries A and B - 2

Pag. 31 - The angle of parallelism.

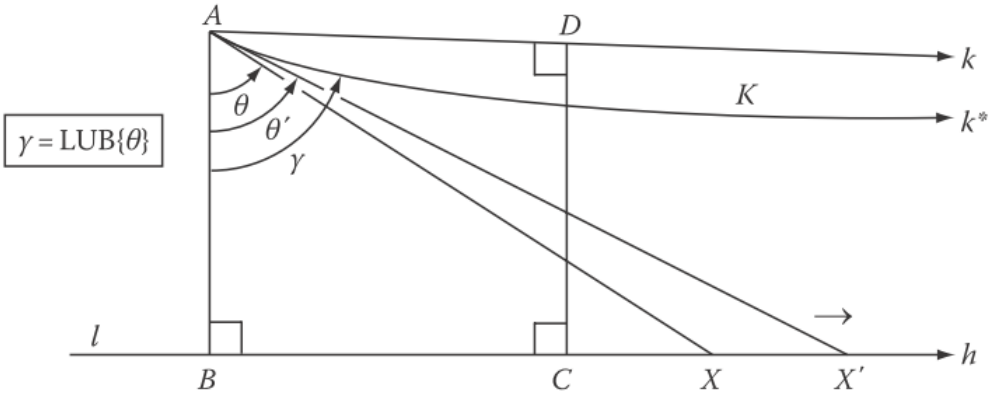
A
D
k
K
k*
γ = LUB{θ}
θ
θ′
γ
→
l
h
B
C
X
X′

Figure 1.76: D38.png - The angle of parallelism

Pag. 36 - The lines through point A.

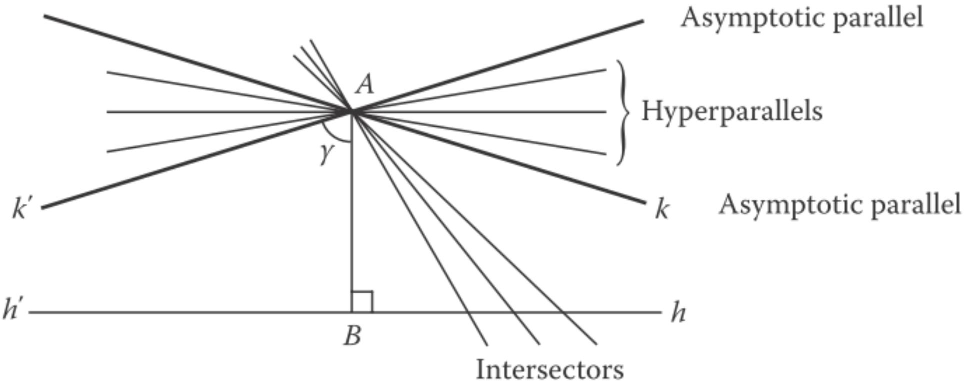
Asymptotic parallel
A
Hyperparallels
γ
k′
k
Asymptotic parallel
h′
h
B
Intersectors

Figure 1.77: D39.png - The lines through point A

Pag. 37 - Two types of parallelism.

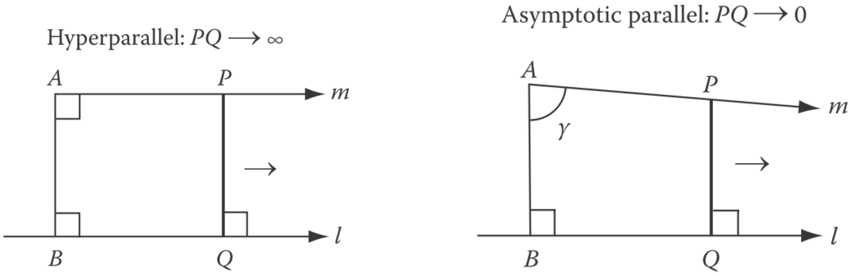
Hyperparallel: $PQ \longrightarrow \infty$
A
P
m
→
l
B
Q
Asymptotic parallel: $PQ \longrightarrow 0$
A
P
m
γ
→
l
B
Q

Figure 1.78: D40.png - Two types of parallelism - 1

Figure 1.79: D41.png - Two types of parallelism - 2

Pag. 39 - Impossibility of Y' > Y

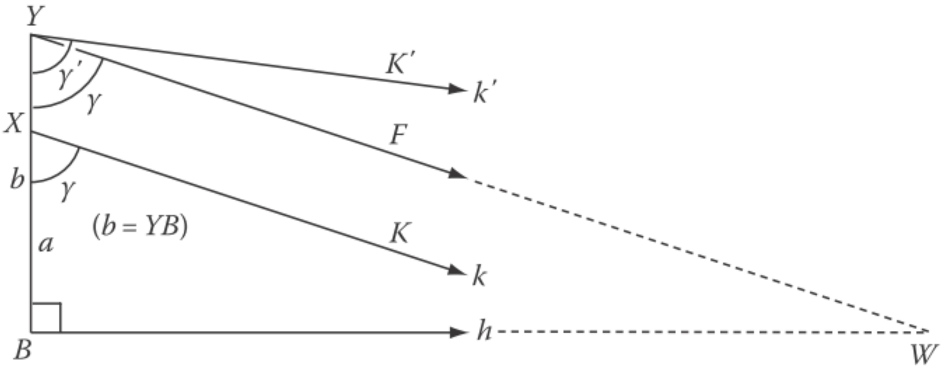
Y
γ'
γ
K'
k'
X
F
b
γ
(b = YB)
K
k
a
h
B
W

Figure 1.80: D42.png - The lines through point A

Pag. 40 - General definition of asymptotic parallelism.

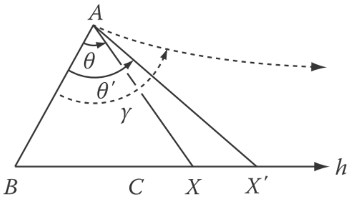

Figure 1.81: D43.png !

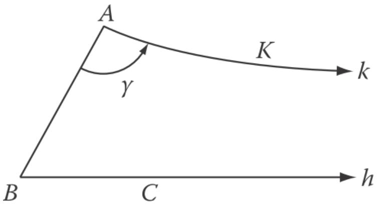

Figure 1.82: D44.png

Pag. 41 - Asymptotic triangle.

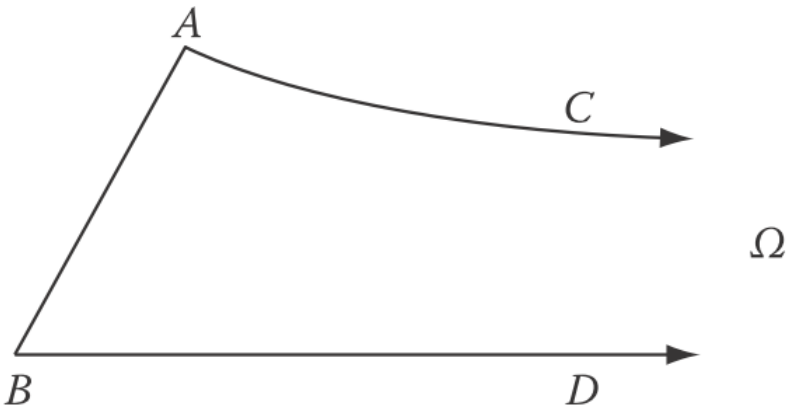

Figure 1.83: D45.png - Asumptotic triangle

Pag. 42 - Types of asymptotic triangles.

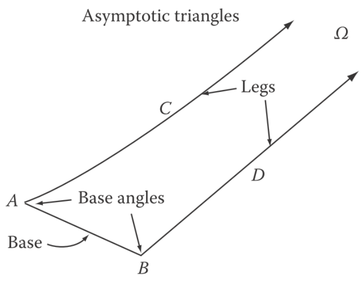
Asymptotic triangles
Ω
Legs
C
D
A
Base angles
Base
B

Figure 1.84: D46.png - Asymptotic triangles

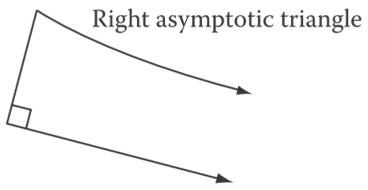
Right asymptotic triangle

Figure 1.85: D47.png - Right asymptotic triangle

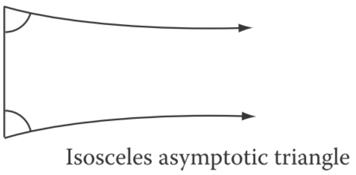
Isosceles asymptotic triangle

Figure 1.86: D48.png - Isosceles asymptotic triangle

Pag. 44 - Impossibility of A + B 180

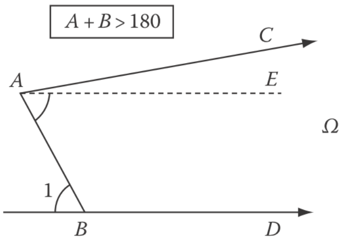
A + B > 180
C
A
E
Ω
1
B
D

Figure 1.87: D49.png

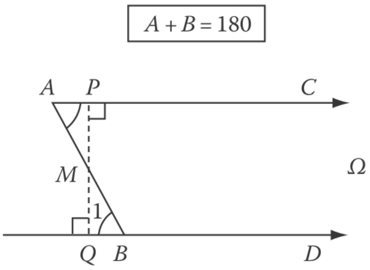
A + B = 180
A
P
C
M
Ω
1
Q
B
D

Figure 1.88: D50.png

Pag. 45 -The four cases of Theorem 2

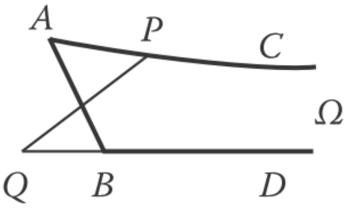

Figure 1.89: D51a.png

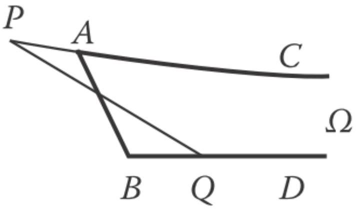

Figure 1.90: D51b.png

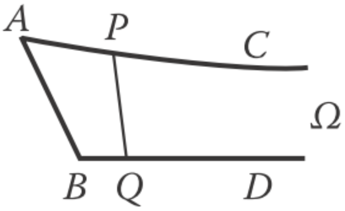

Figure 1.91: D51c.png

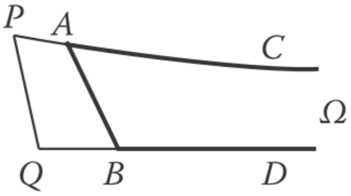

Figure 1.92: D51d.png

Pag. 46 - Proving case (1)

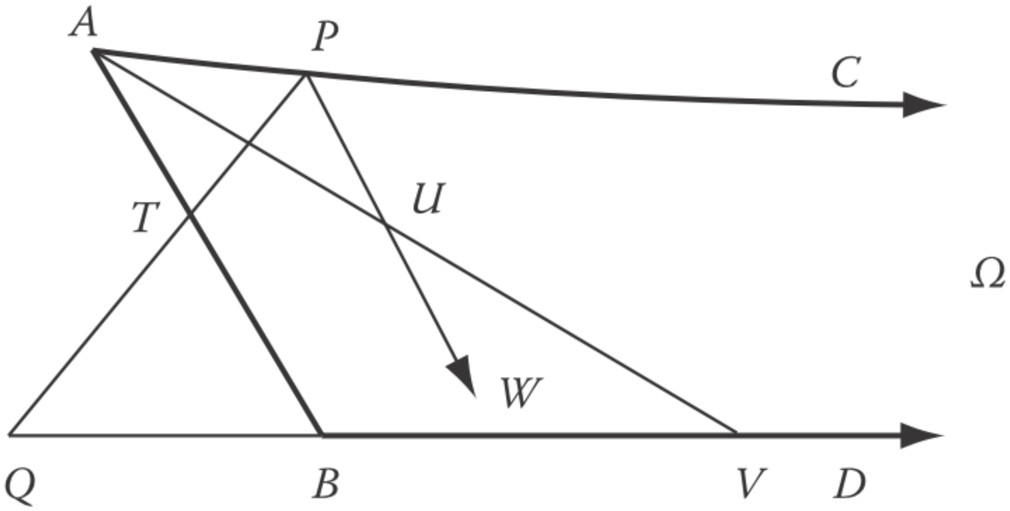
A
P
C
U
T
Ω
W
Q
B
V
D

Figure 1.93: D52.png

Pag. 47 - Proving case (2)

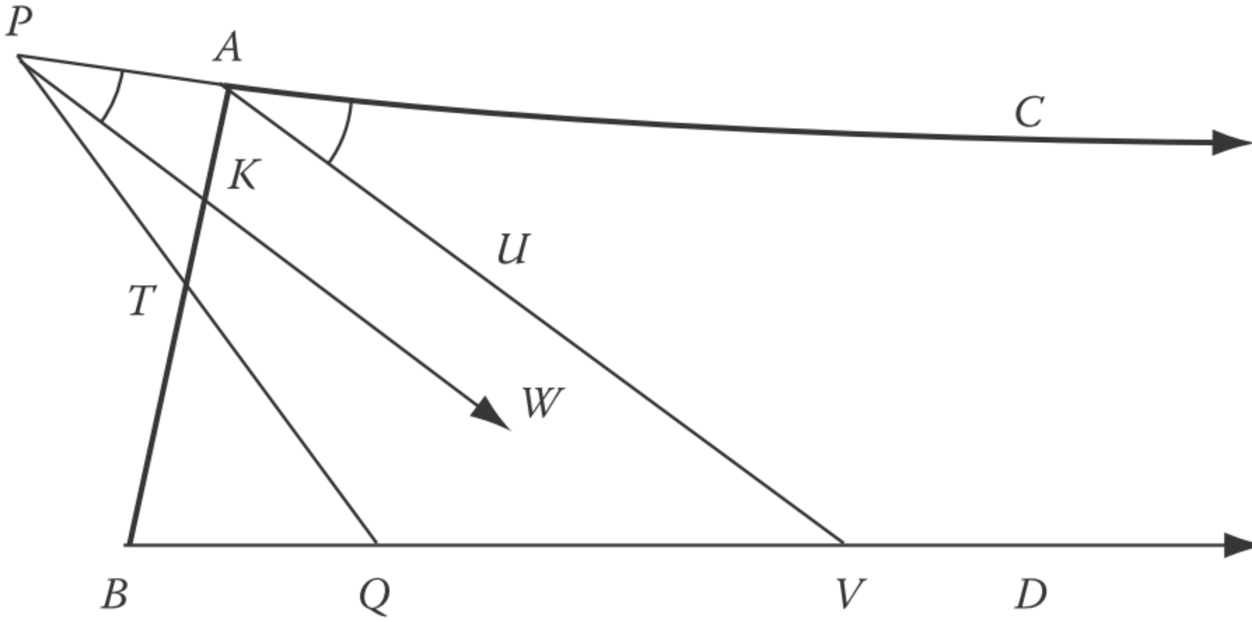
P
A
C
K
U
T
W
B
Q
V
D

Figure 1.94: D53.png

Pag. 48 - Exterior angle inequality: < 1 >< A

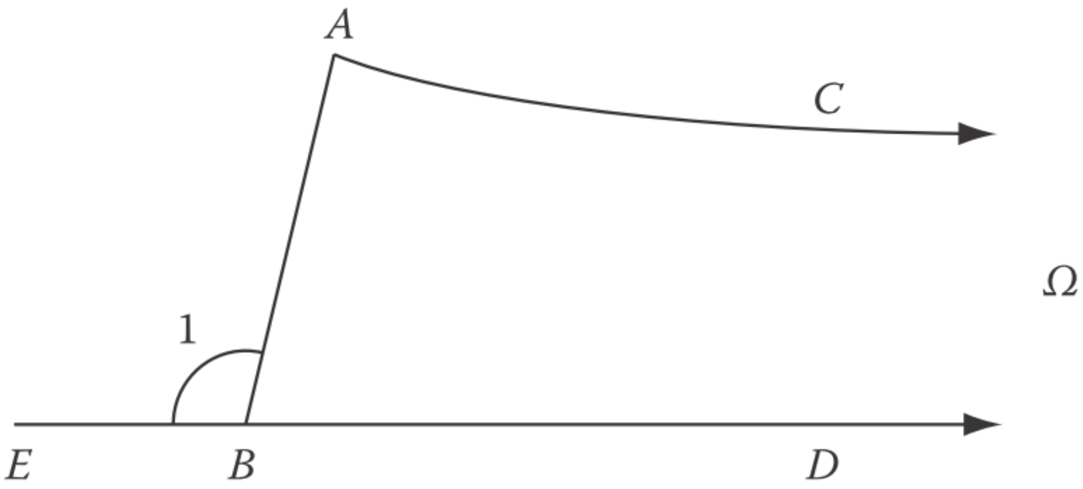

Figure 1.95: D54.png

Pag. 49 - Symmetry for isosceles asymptotic triangles

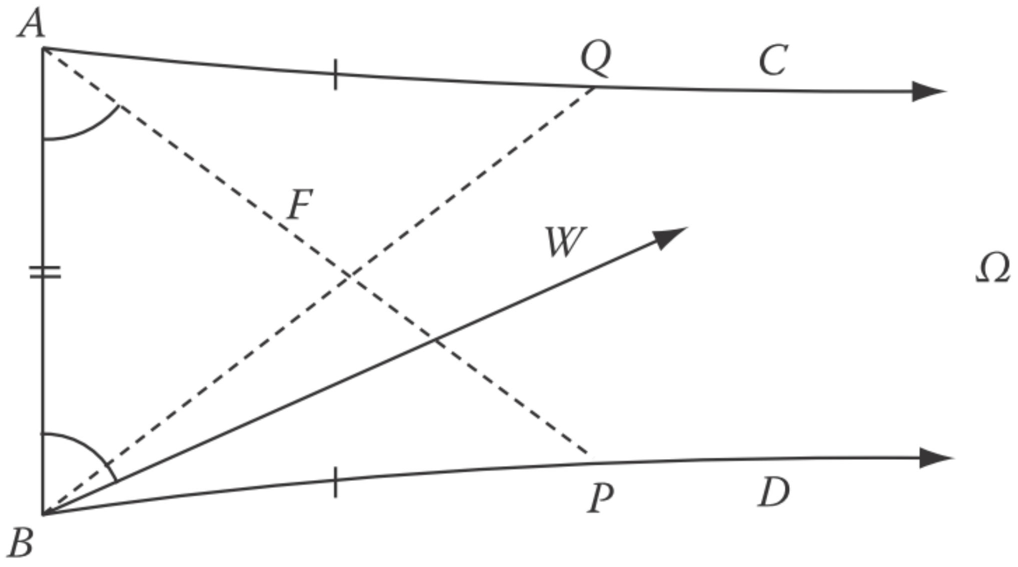

Figure 1.96: D55.png - Symmetry for isosceles asymptotic triangles

Pag. 50 - Symmetry for arbitrary asymptotic triangles. IMPORTANT

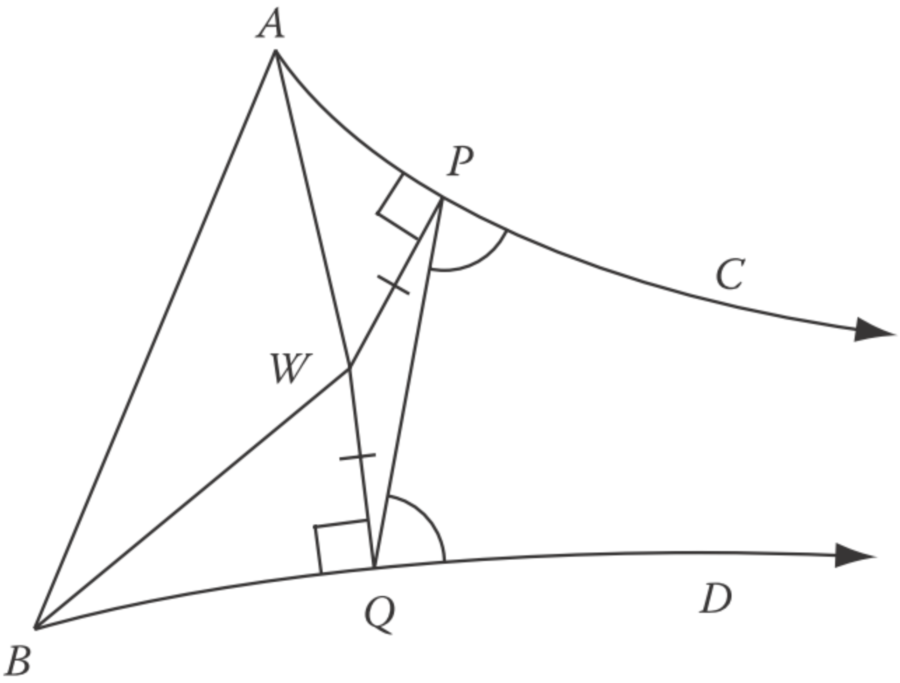

Figure 1.97: D56.png - Symmetry for arbitrary asymptotic triangles

Pag. 51 - Base-angle congruence? If $m < BAC > m < YXU$

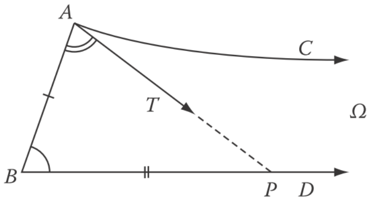

Figure 1.98: D57a.png

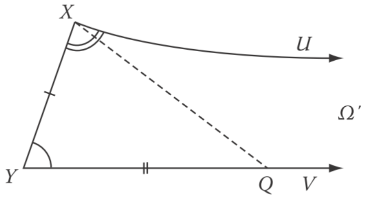

Figure 1.99: D57b.png

Pag. 52 - AA congruence - $AB > XY$

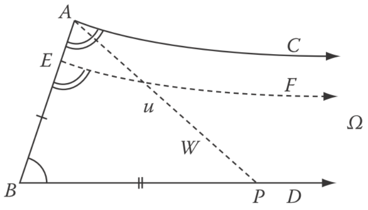
A
E
C
F
u
Ω
W
B
P
D

Figure 1.100: D58a.png

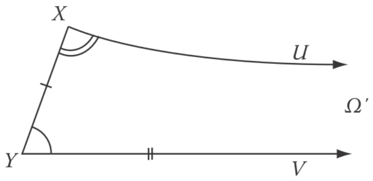

Figure 1.101: D58b.png

Pag. 53 - If a < b then p(a) > p(b)

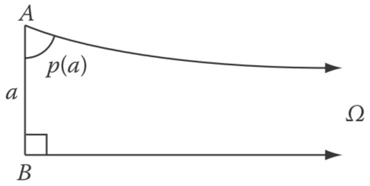

Figure 1.102: D59a.png

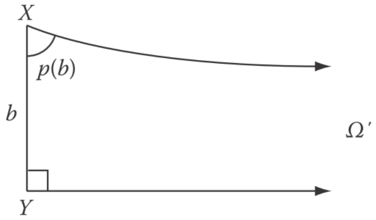

Figure 1.103: D59b.png

Pag. 62

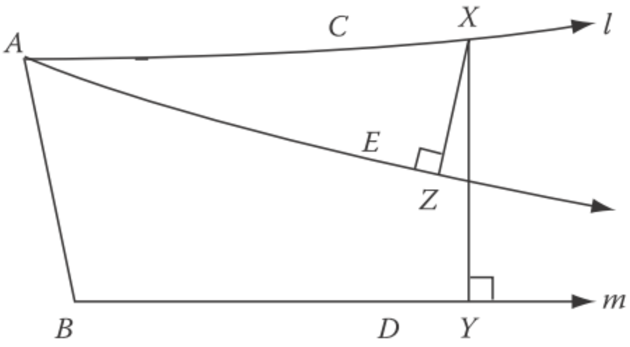
A
C
X
l
E
Z
B
D
Y
m

Figure 1.104: D60.png

Pag. 84 - Ratios preserved under parallel projection.

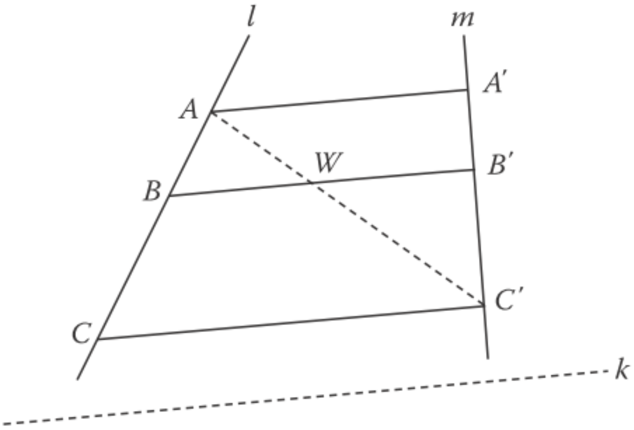
l
m
A
A′
W
B
B′
C
C′
k

Figure 1.105: D61.png

Pag. 85 - Proof that medians are concurrent.

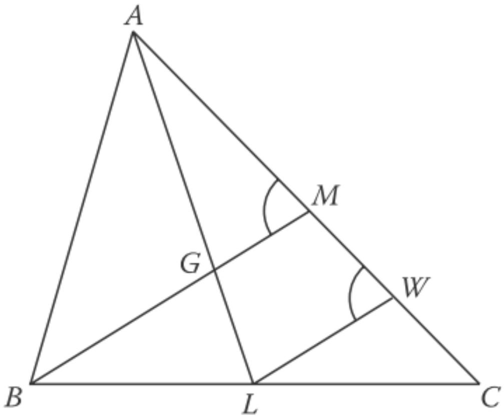

Figure 1.106: D62.png

Pag. 86 - The four classical points of concurrencyIMPORTANT Three points in line.

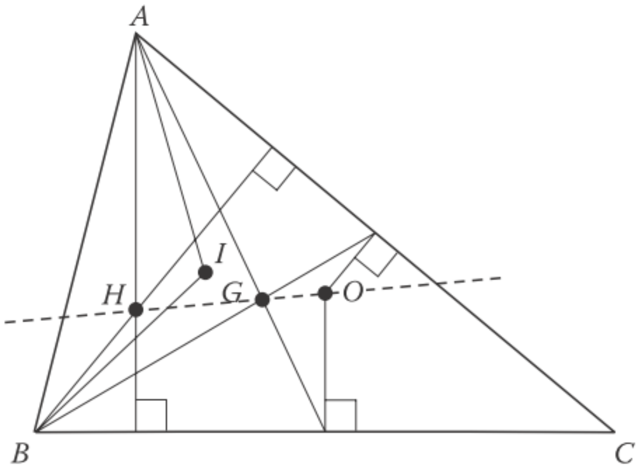

Figure 1.107: D63.png

Pag. 87 - Existence of Euler line.IMPORTANT The points in line

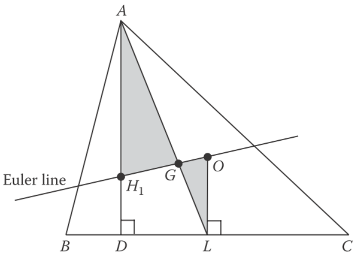
A
Euler line
G
O
H_1
B
D
L
C

Figure 1.108: D64a.png

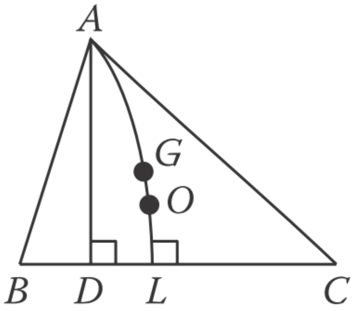

Figure 1.109: D64b.png

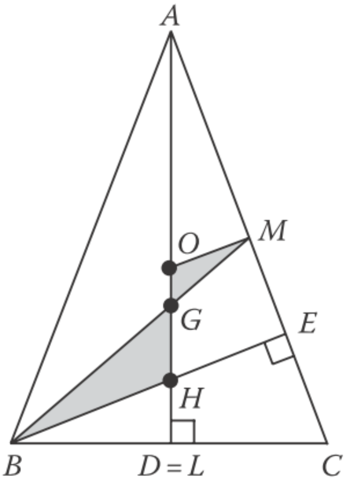
A
O
M
G
E
H
B
D = L
C

Figure 1.110: D64c.png

Pag. 88 - The orthic triangle.

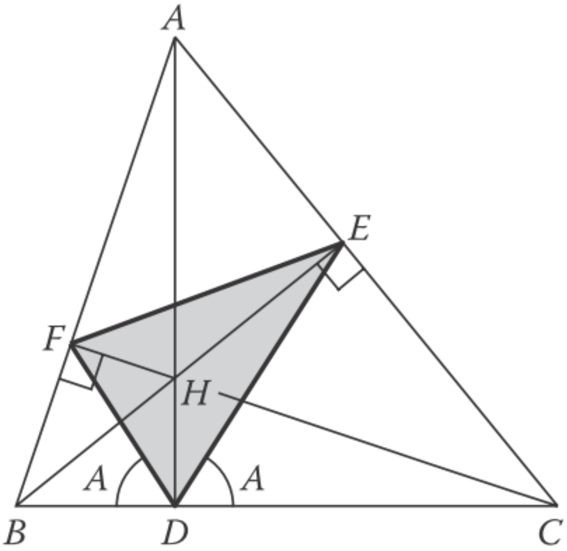

Figure 1.111: D65.png - The orthic triangle

Pag. 106 - The Morley triangle

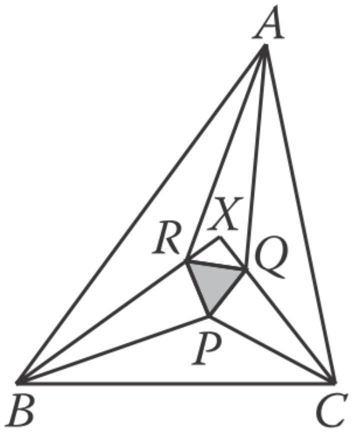

Figure 1.112: D66a.png

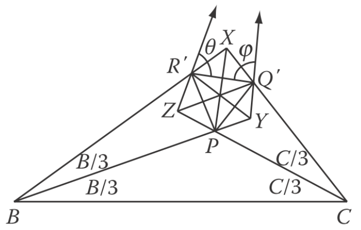
θ
X
φ
R′
Q′
Z
Y
P
B/3
B/3
C/3
C/3
B
C

Figure 1.113: D66b.png

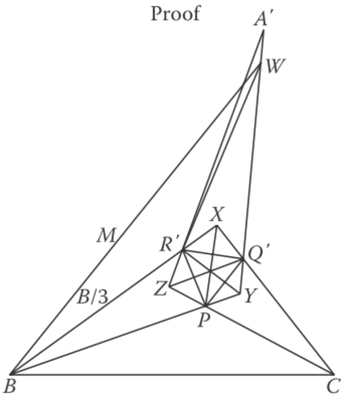
Proof
A′
W
M
X
R′
Q′
Z
Y
B/3
P
B
C

Figure 1.114: D66c.png

12) C:/DATA-NovoVivo-126-2020/Books-126-2020/9781439819111-Ancillaries/K1100
C006.pdf IMPORTANT

Pag. 25 - Parallel Projection.

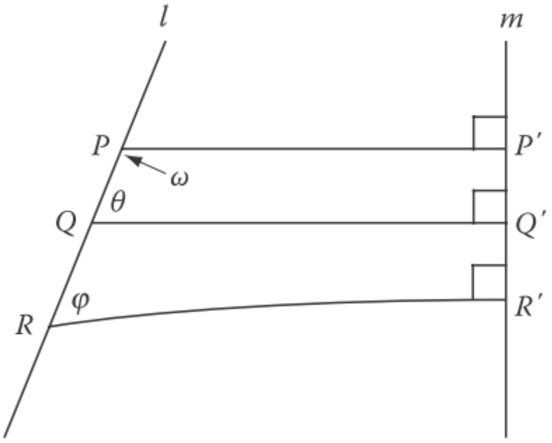

Figure 1.115: D67.png - Parallel Projection

Pag. 26 - Parallel projection in hyperbolic geometry.

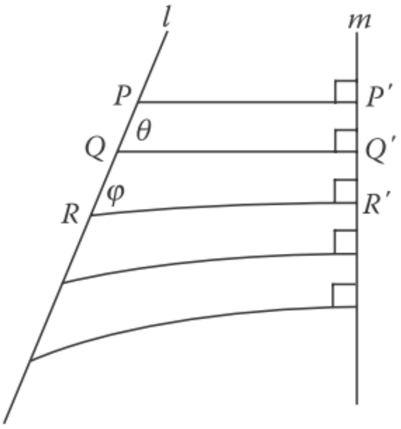

Figure 1.116: D68.png - Parallel projection in hyperbolic geometry

Pag. 27 - Parallel projection in elliptic geometry.

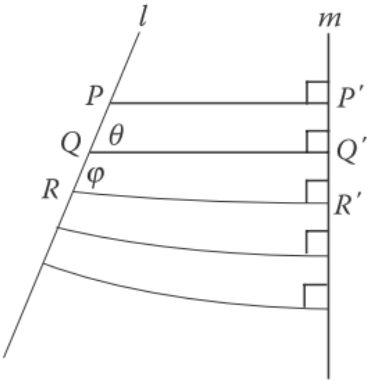
l
m
P
P′
Q
θ
Q′
R
φ
R′

Figure 1.117: D69a.png !

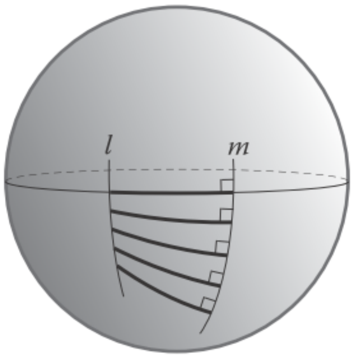

Figure 1.118: D69b.png

Pag. 29 - Similar triangles in hyperbolic geometry.

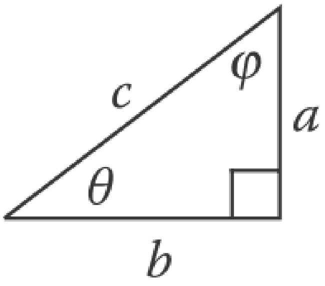

Figure 1.119: D70a.png !

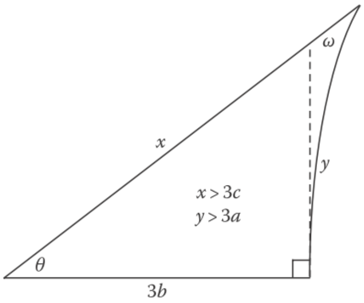
ω
x
y
x > 3c
y > 3a
θ
3b

Figure 1.120: D70b.png

Pag. 30 - Similar triangles in elliptic geometry. IMPORTANT Research

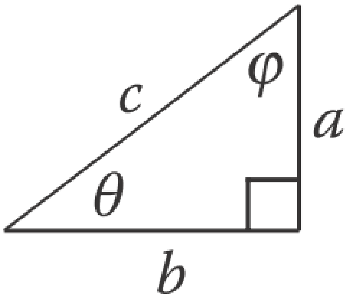

Figure 1.121: D71a.png

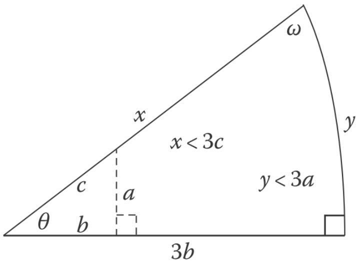
ω
x
y
x < 3c
c
a
y < 3a
θ
b
3b

Figure 1.122: D71b.png

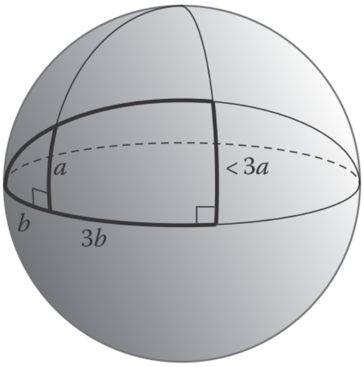

Figure 1.123: D71c.png

Pag. 32 - Graphs of F(x).

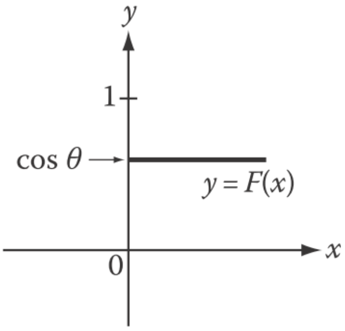

Figure 1.124: D72a.png - Euclidean geometry

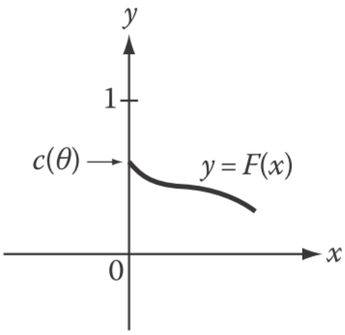

Figure 1.125: D72b.png - Hyperbolic geometry

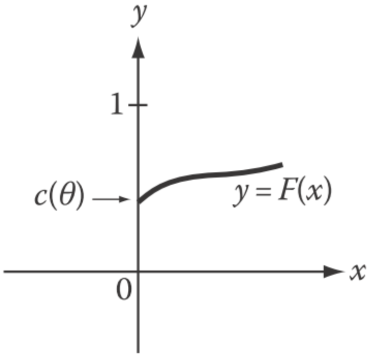

Figure 1.126: D72c.png - Elliptic geometry

Pag. 52 - Graphs of G(x)

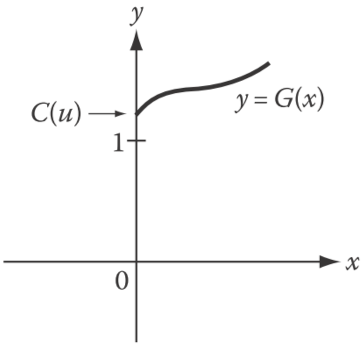

Figure 1.127: D73a.png - Hyperbolic geometry

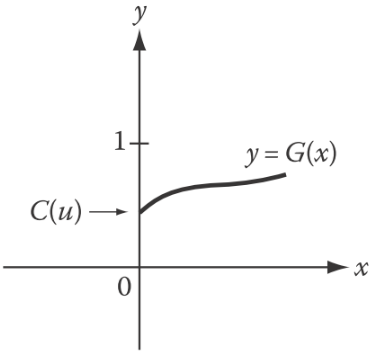

Figure 1.128: D73b.png - Elliptic geometry

Pag. 54 - Proving C(u) > 1. IMPORTANT

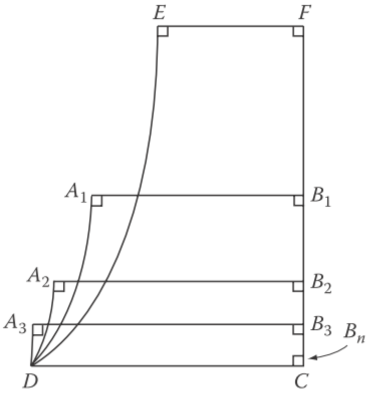
E
F
A_1
B_1
A_2
B_2
A_3
B_3
B_n
D
C

Figure 1.129: D74a.png

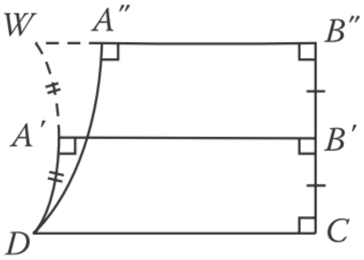

Figure 1.130: D74b.png

Pag. 55 - Proof of product-to-sum identity. IMPORTANT

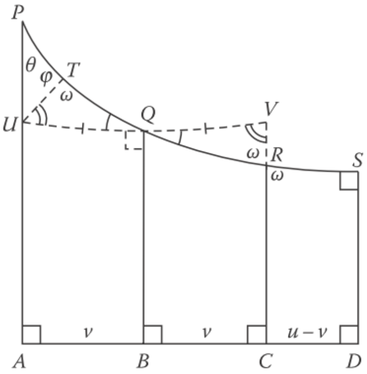
P
θ
φ
T
ω
U
Q
V
ω
R
ω
S
v
v
u − v
A
B
C
D

Figure 1.131: D75a.png - Hyperbolic Geometry

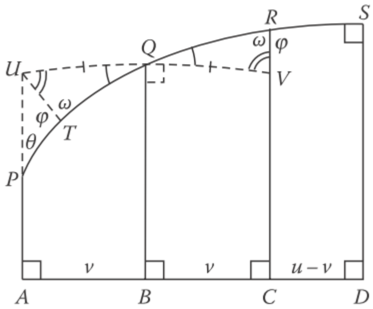
R
S
Q
ω
φ
U
V
ω
φ
T
θ
P
v
v
u − v
A
B
C
D

Figure 1.132: D75b.png - Elliptic Geometry

Pag. 56 - Pythagorean relation proved in unified geometry.

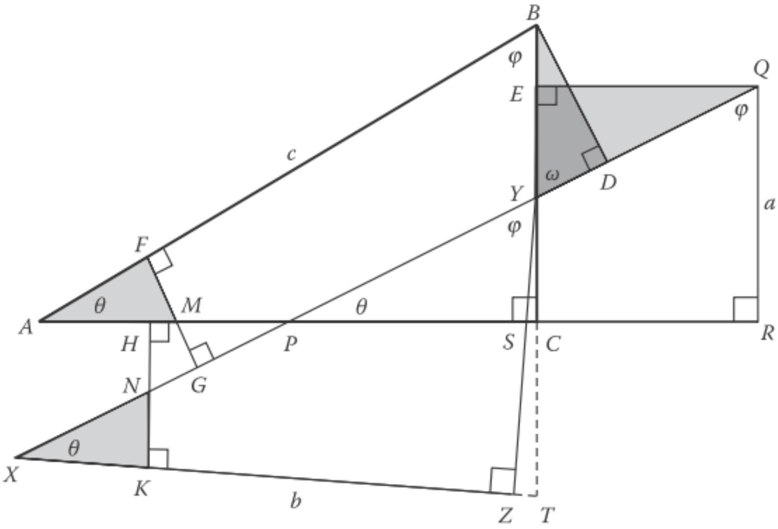
B
φ
E
Q
φ
c
ω
D
Y
a
φ
F
θ
M
θ
A
H
P
S
C
R
N
G
θ
X
K
b
Z
T

Figure 1.133: D76.png - Pythagorean relation proved in unified geometry

Pag. 57 - Proving cos A = T(b)/T(c)

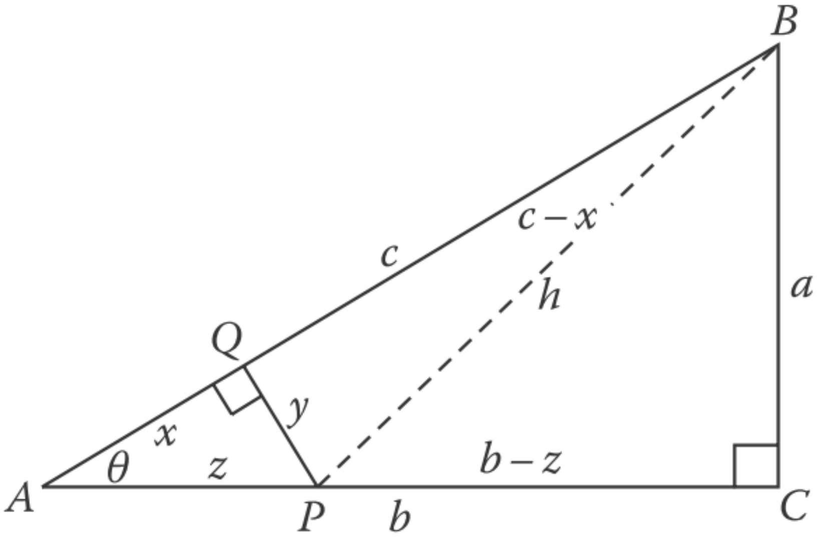
B
c − x
c
h
a
Q
y
x
θ
z
b − z
A
P
b
C

Figure 1.134: D77.png

Pag. 58 - Radius = 1 e =

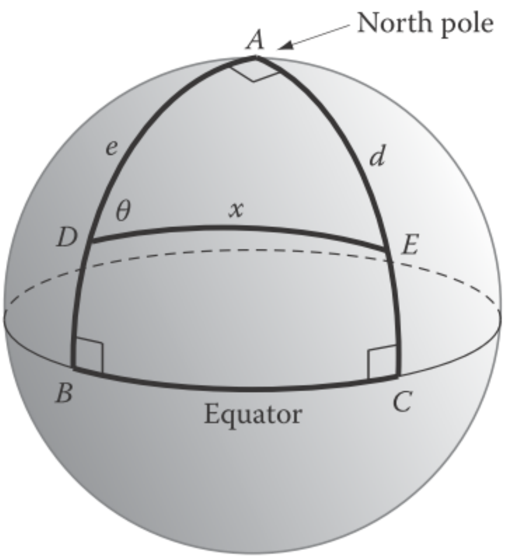
A
North pole
e
d
θ
x
D
E
B
Equator
C

Figure 1.135: D77.png

Pag. 62 - IMPORTANT

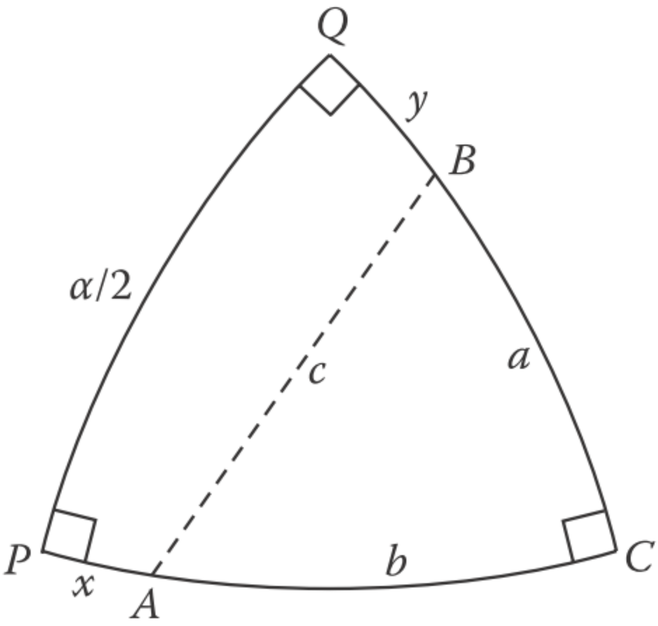
Q
y
B
α/2
c
a
P
x
A
b
C

Figure 1.136: D79.png

Pag. 64

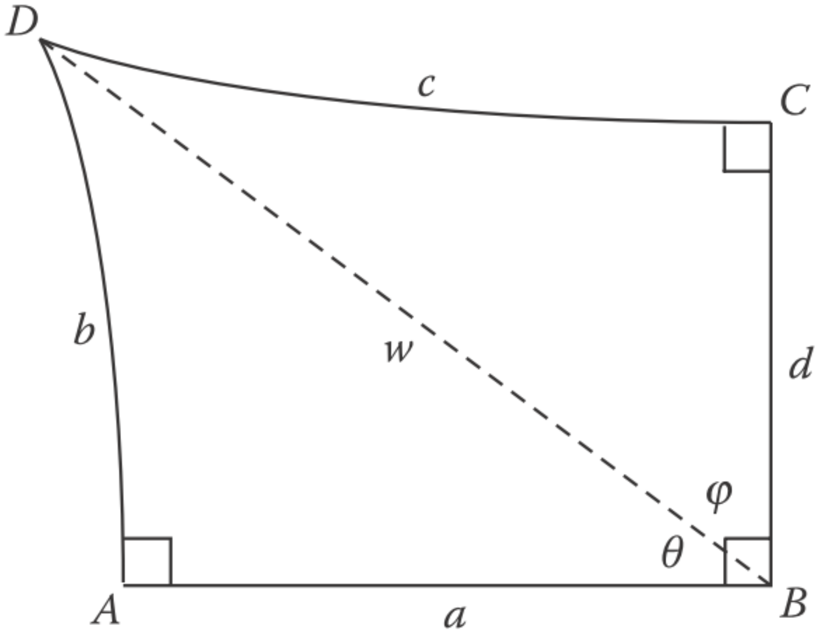
D
c
C
b
w
d
φ
θ
A
a
B

Figure 1.137: D80.png

Pag. 65

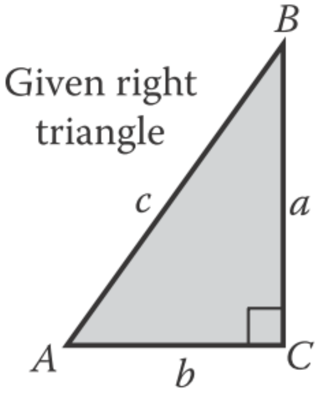
B
Given right triangle
c
a
A
b
C

Figure 1.138: D81a.png

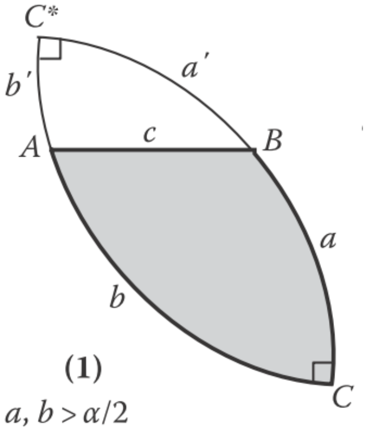
C*
b′
a′
A
c
B
a
b
(1)
a, b > α/2
C

Figure 1.139: D81b.png

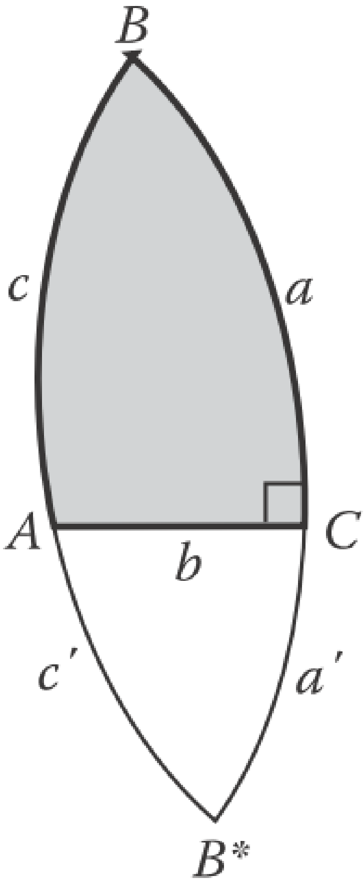

Figure 1.140: D81c.png

13) C:/DATA-NovoVivo-126-2020/Books-126-2020/9781439819111-Ancillaries/K11008-

C007.pdf

Pag. 15 - The power of an interior point.

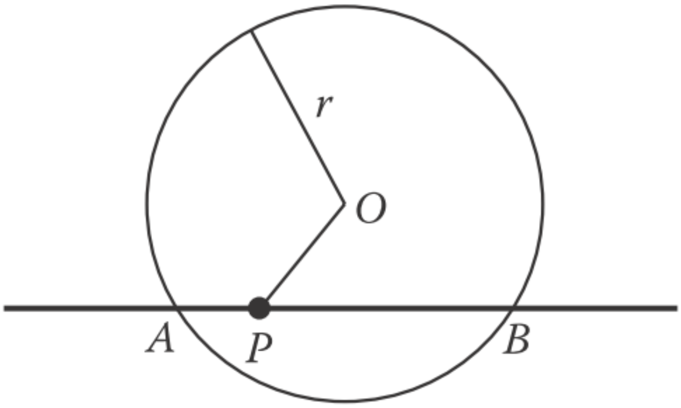

Figure 1.141: D82.png

Pag. 16 - Segments on concurrent chords.

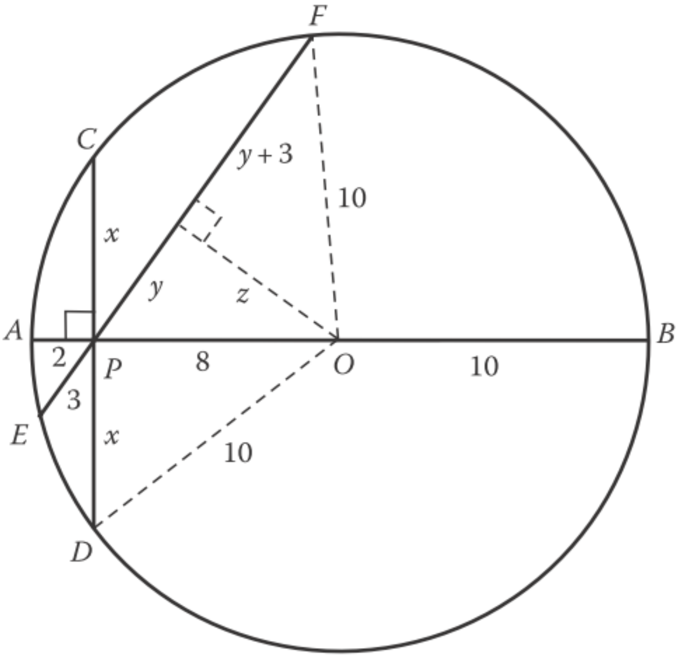
F
C
y + 3
10
x
y
z
A
2
P
8
O
10
B
3
E
x
10
D

Figure 1.142: D83.png

Pag. 17 - Three types of angles and intercepted arcs. The power theorems.

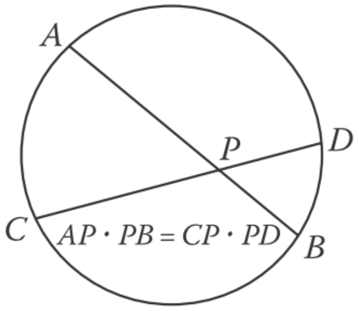
A
P
D
C
$AP \cdot PB = CP \cdot PD$
B

Figure 1.143: D84a.png

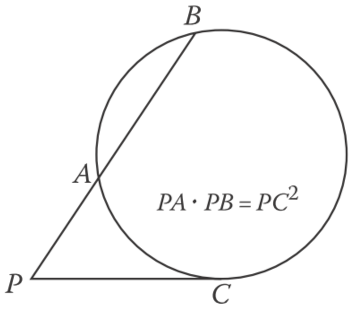
B
A
$PA \cdot PB = PC^2$
P
C

Figure 1.144: D84b.png

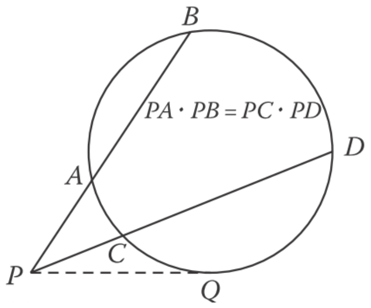
B
$PA \cdot PB = PC \cdot PD$
D
A
C
P
Q

Figure 1.145: D84c.png

Pag. 23 - Proof of Ptolemy's theorem.

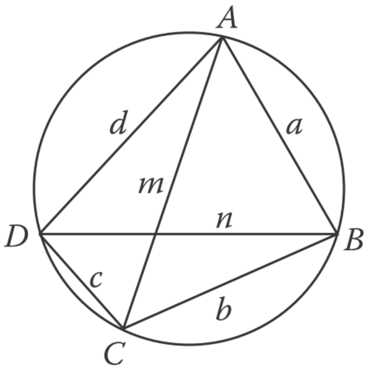
A
d
a
m
n
D
B
c
b
C

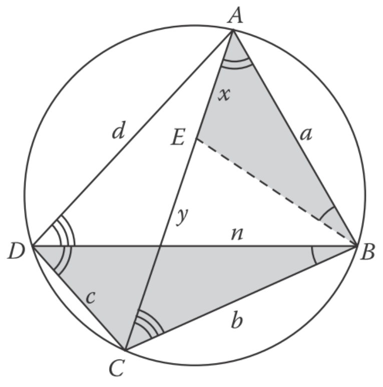
A
x
d
E
a
y
n
D
B
c
b
C

Figure 1.146: D85a.png -ac+ bd= mn

Figure 1.147: D85b.png -x + y = m

Pag. 26 - The nine-point circle. IMPORTANT Research

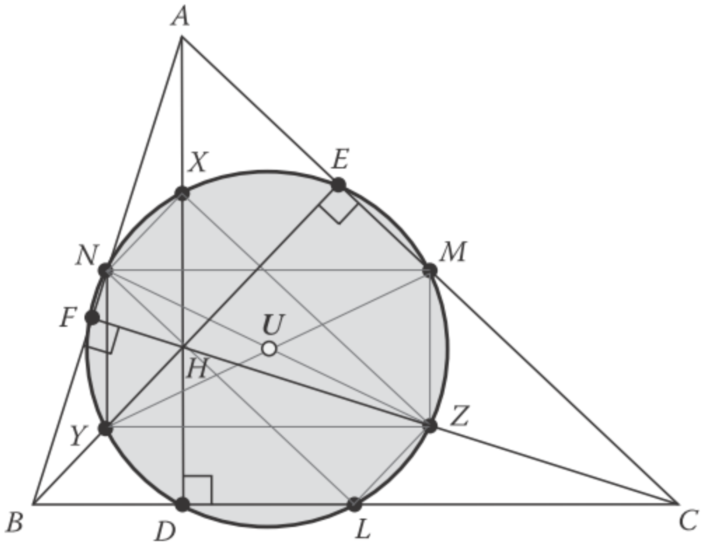

Figure 1.148: D86.png - The nine-point circle

Pag. 27 - Power of P with respect to two circles.

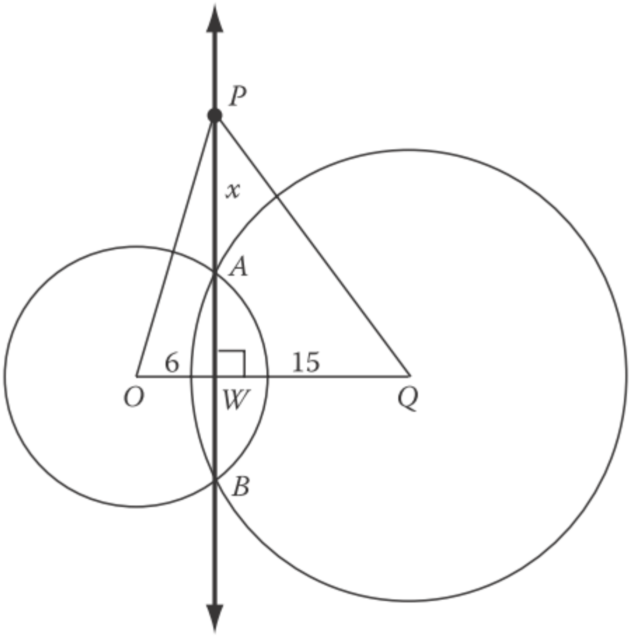
P
x
A
6
15
O
W
Q
B

Figure 1.149: D87.png - Power of P with respect to two circles

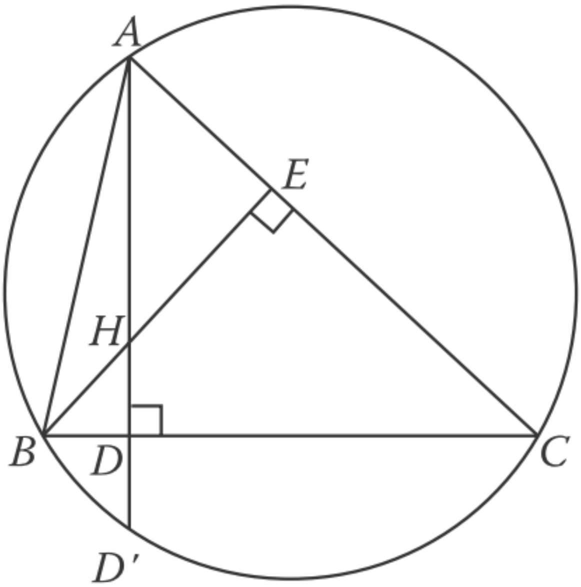

Figure 1.150: D88.png

Pag. 49 -

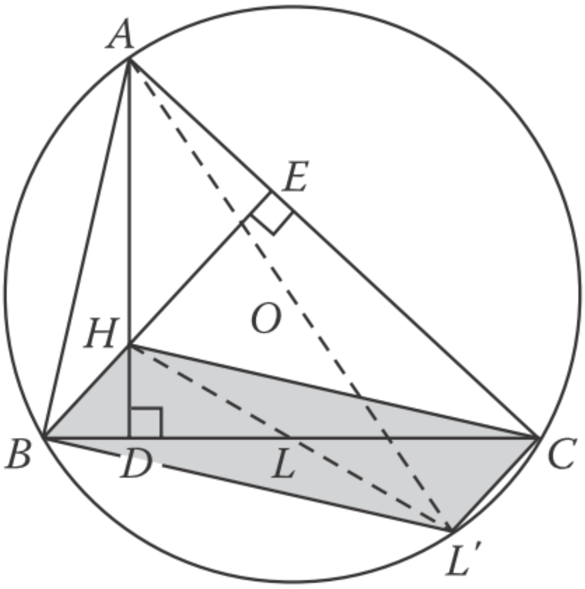

Figure 1.151: D89.png

Pag. 56 -

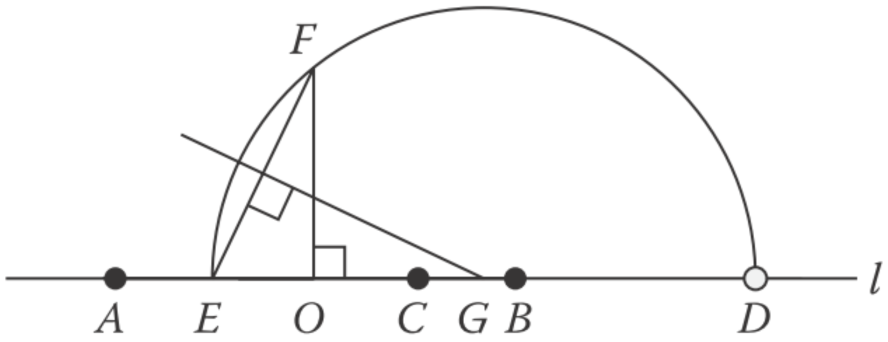

Figure 1.152: D90.png

Pag. 67 - The Nagel point.

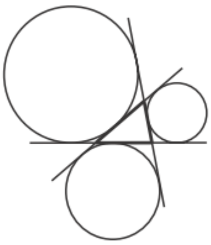

Figure 1.153: D91a.png

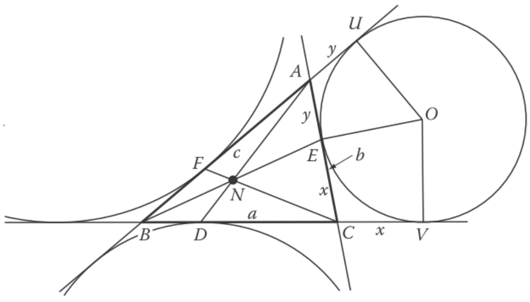
U
y
A
y
O
F
c
E
b
N
x
a
B
D
C
x
V

Figure 1.154: D91b.png

Pag. 68 - The Fermat point.

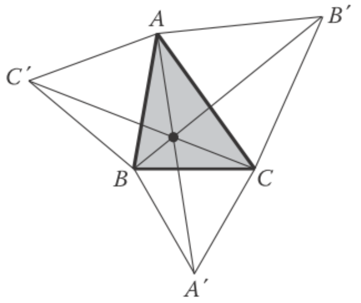

Figure 1.155: D92a.png - The Fermat Point

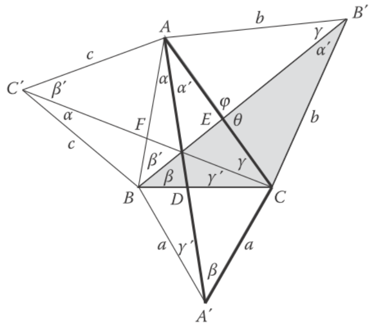
A
b
B′
γ
α′
c
C′
β′
α
α′
φ
E
θ
F
b
c
β′
γ
β
γ′
B
D
C
a
γ′
a
β
A′

Figure 1.156: D92b.png - Prof

Pag. 69 - Simson lines of triangles.

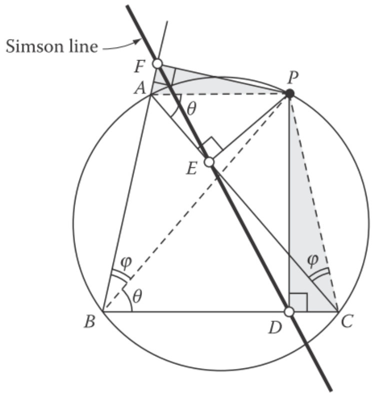
Simson line
F
A
P
θ
E
φ
φ
θ
B
D
C

Figure 1.157: D93a.png

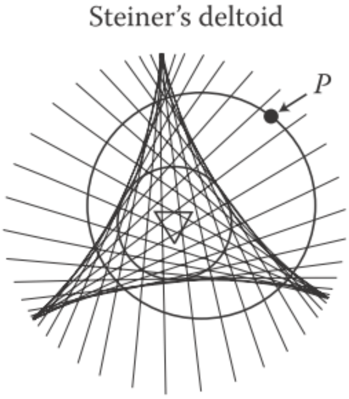
Steiner's deltoid
P

Figure 1.158: D93b.png

Pag. 76 -

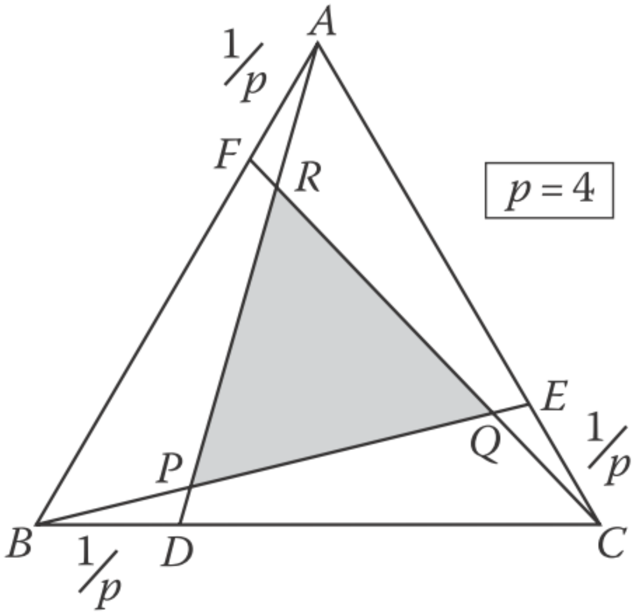
A
1/p
F
R
p = 4
E
1/p
Q
P
B
1/p
D
C

Figure 1.159: D94.png

Pag. 77 -

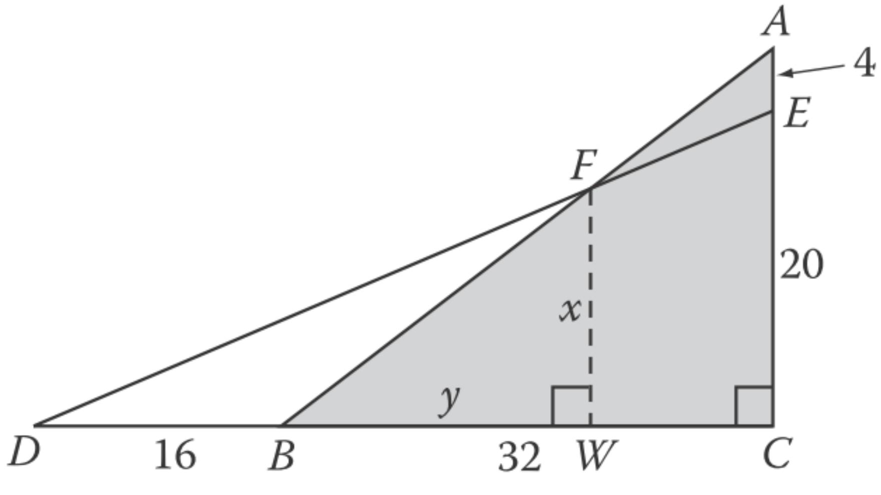
A
4
E
F
20
x
y
D
16
B
32
W
C

Figure 1.160: D95.png

Pag. 79 -

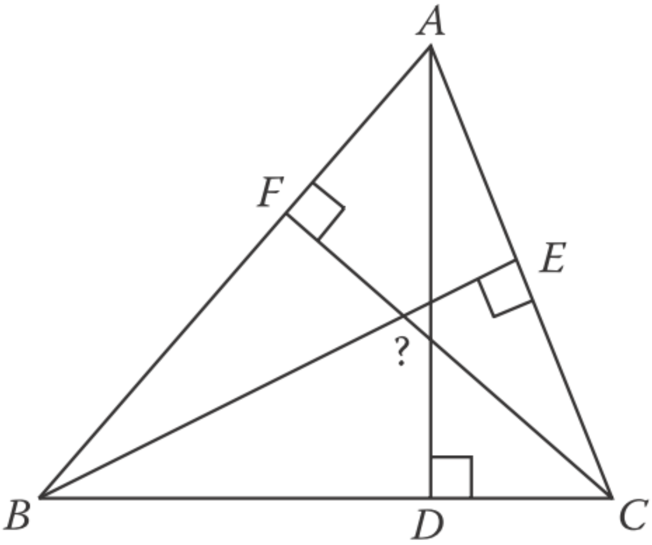

Figure 1.161: D96.png

Pag. 80 - Gergonne point

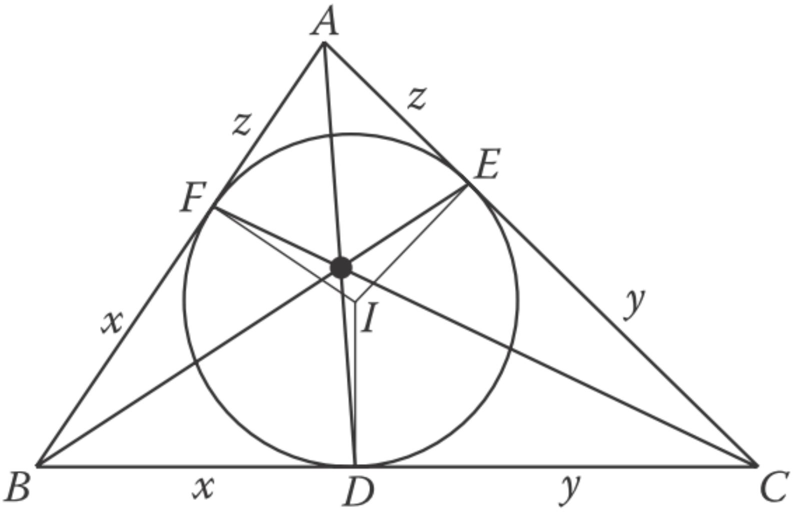
A
z
z
F
E
x
I
y
B
x
D
y
C

Figure 1.162: D97.png

Pag. 82 - IMPORTANT Research

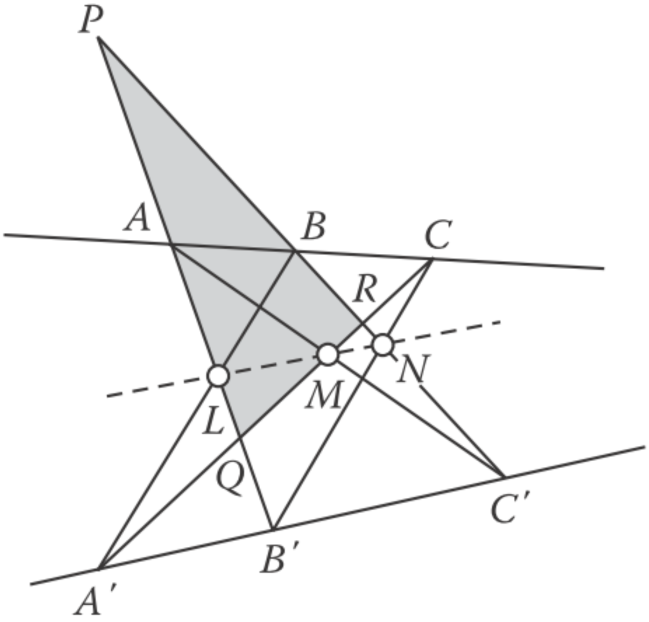
P
A
B
C
R
N
M
L
Q
C′
B′
A′

Figure 1.163: D98.png

Pag. 84 - Mutually orthonic conics.

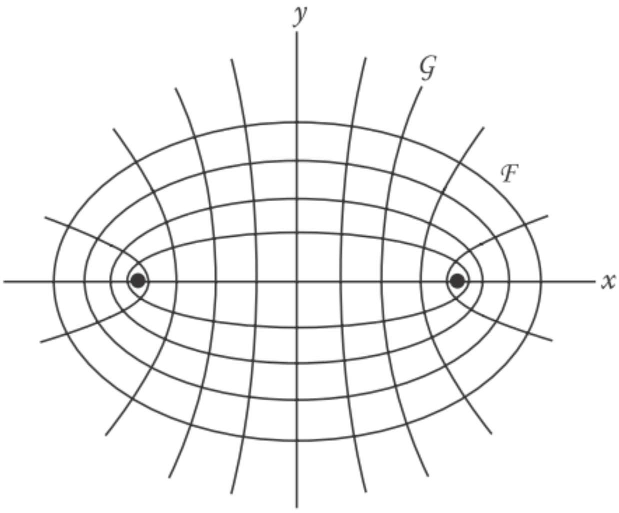

Figure 1.164: D99.png

Pag. 85 - Orthogonal circles.

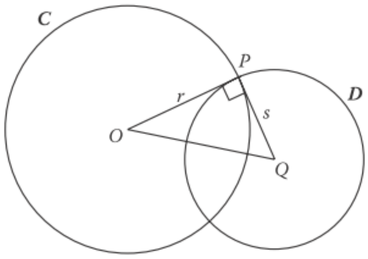

Figure 1.165: D100a.png

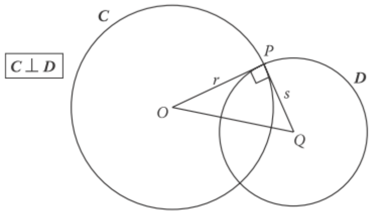

Figure 1.166: D100b.png

Pag. 86 - Harmonic conjugates and orthogonal circles.

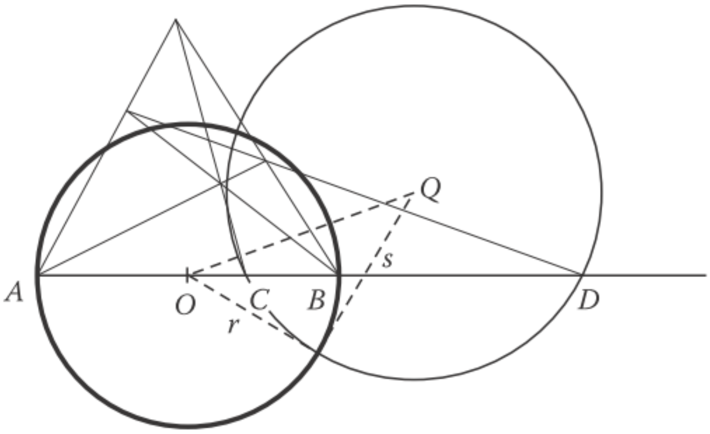

Figure 1.167: D101.png - Harmonic conjugates and orthogonal circles

Pag. 87 - Circles orthogonal to circle at AB.

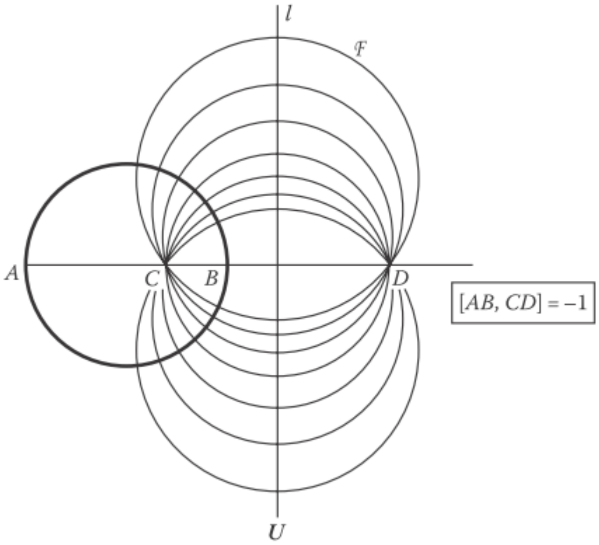
l
F
A
C
B
D
[AB, CD] = −1
U

Figure 1.168: D102.png - Circles orthogonal to circle at AB

Pag. 88 - Mutually orthogonal circles.

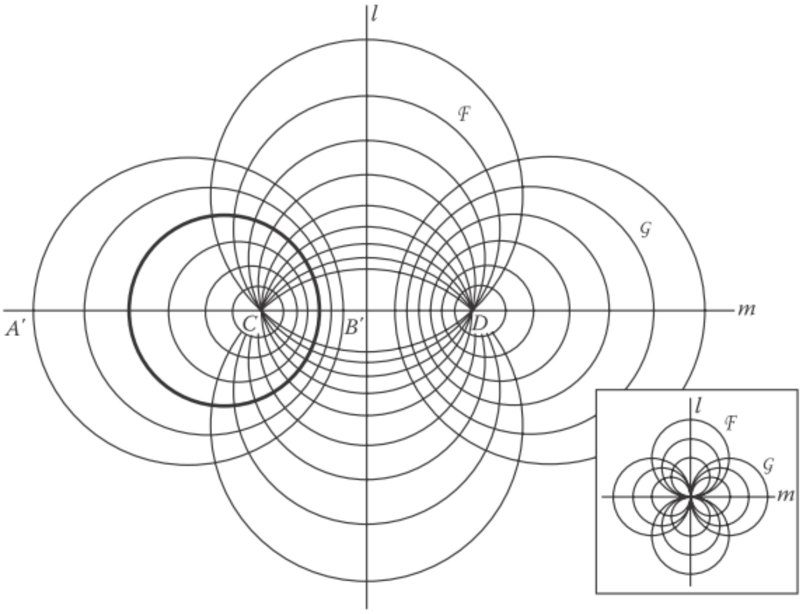
l
F
G
m
A′
C
B′
D
l
F
G
m

Figure 1.169: D103.png - Mutually orthogonal circles

Pag. 89 - Radical axis for non-intersecting circles.

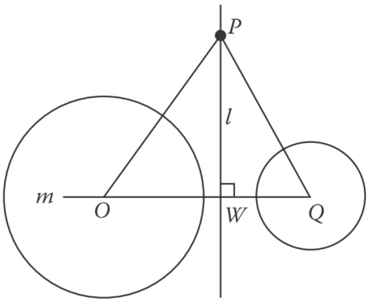

Figure 1.170: D104a.png

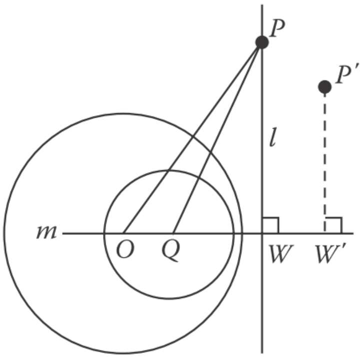

Figure 1.171: D104b.png

Pag. 90 - Radical axis for G.

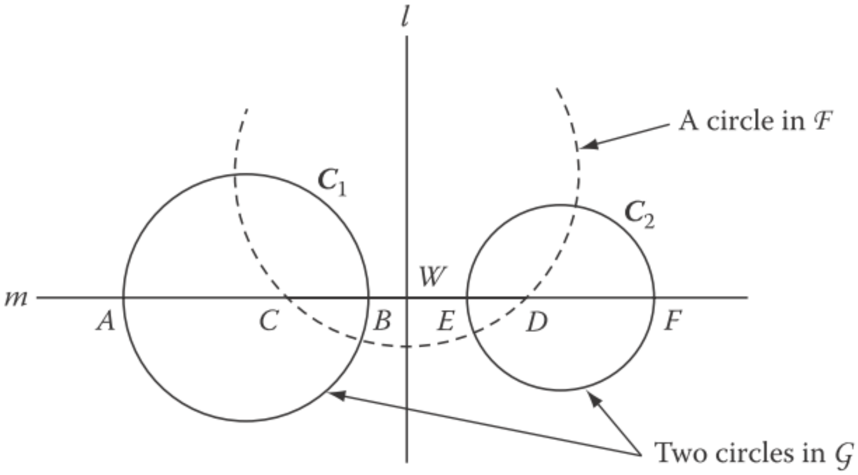

l
A circle in ℱ
C_1
C_2
W
m
A
C
B
E
D
F
Two circles in 𝒢

Figure 1.172: D105.png - Radical axis for G

Pag. 91 - Equations for F and G.

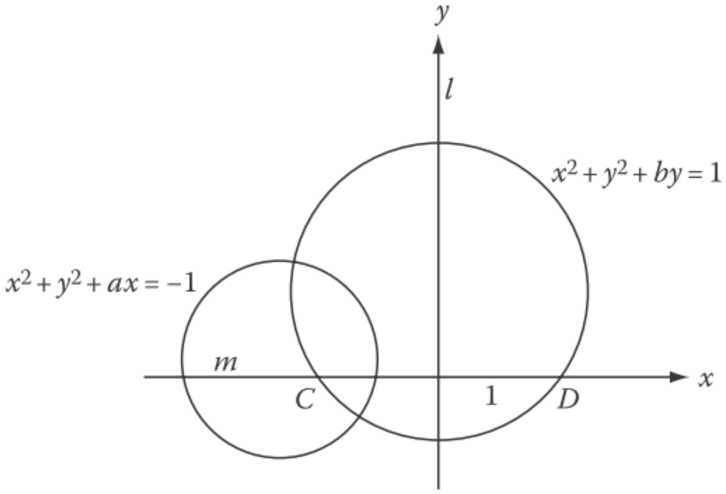
y
l
$x^2 + y^2 + by = 1$
$x^2 + y^2 + ax = -1$
m
C
1
D
x

Figure 1.173: D106.png - Equations for F and G

Pag. 92 -

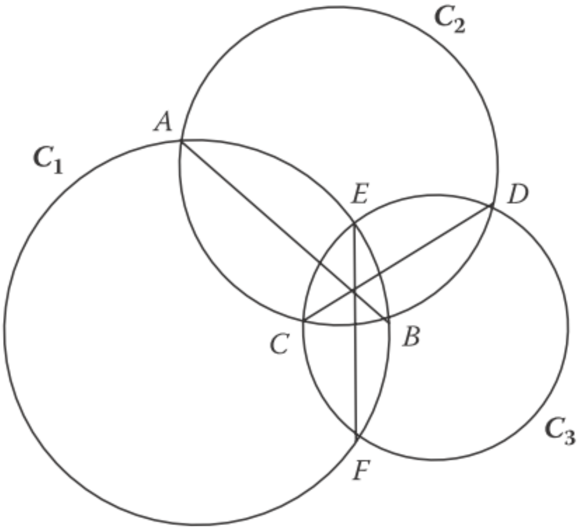
C_2
A
C_1
E
D
C
B
C_3
F

Figure 1.174: D107.png

Pag. 93 -

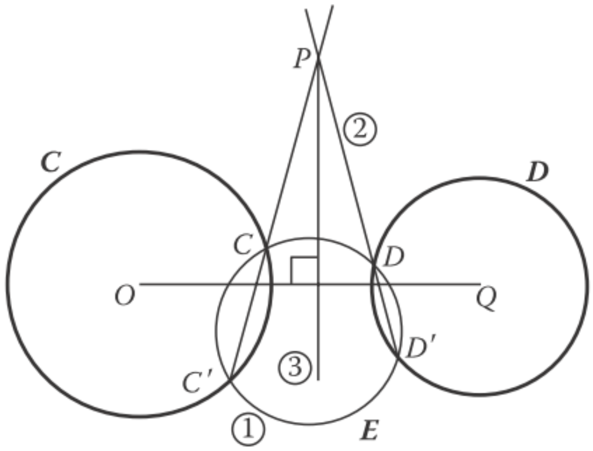
P
②
C
D
C
D
O
Q
D′
C′
③
①
E

Figure 1.175: D108.png

14) C:/DATA-NovoVivo-126-2020/Books-126-2020/9781439819111-Ancillaries/K11008
C008.pdf

Pag. 1 - A central projection

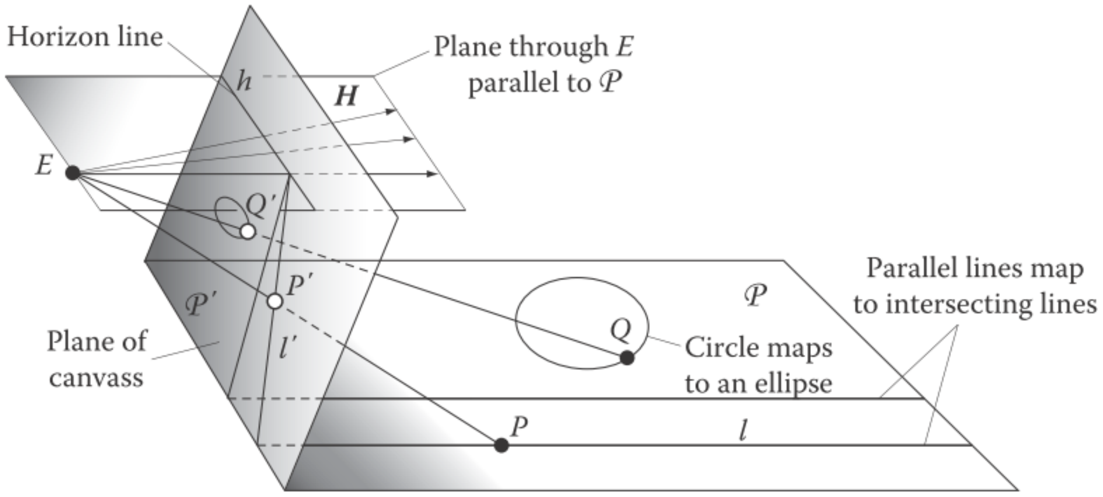
Horizon line
h
H
Plane through E parallel to 𝒫
E
Q′
P′
𝒫′
l′
Plane of canvass
𝒫
Q
Circle maps to an ellipse
Parallel lines map to intersecting lines
P
l

Figure 1.176: D109.png

Pag. 2 -Parallel lines mapping to intersecting lines.

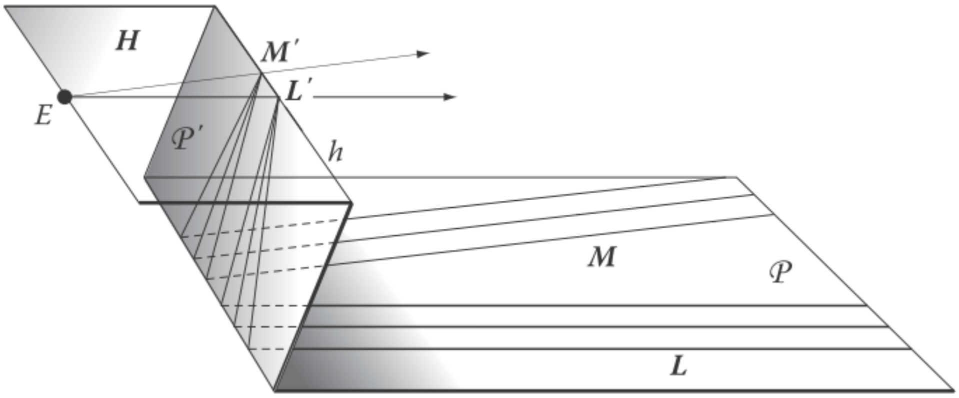

Figure 1.177: D110.png

Pag. 3 - Projective proof of Desargues' theorem.

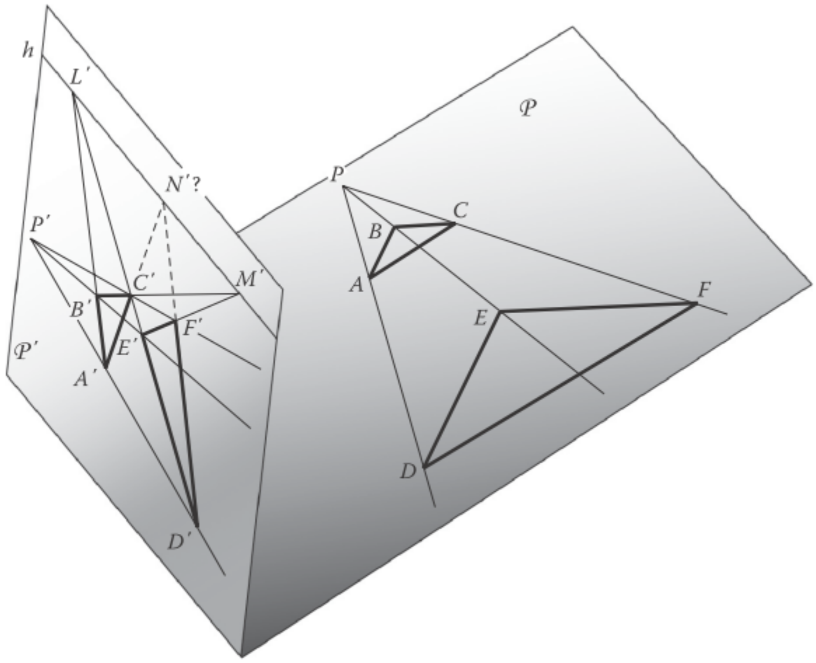
h
L′
N′?
P′
C′
M′
B′
E′
F′
𝒫′
A′
D′
P
𝒫
C
B
A
E
F
D

Figure 1.178: D111.png

Pag. 4 - Product of two central projections.

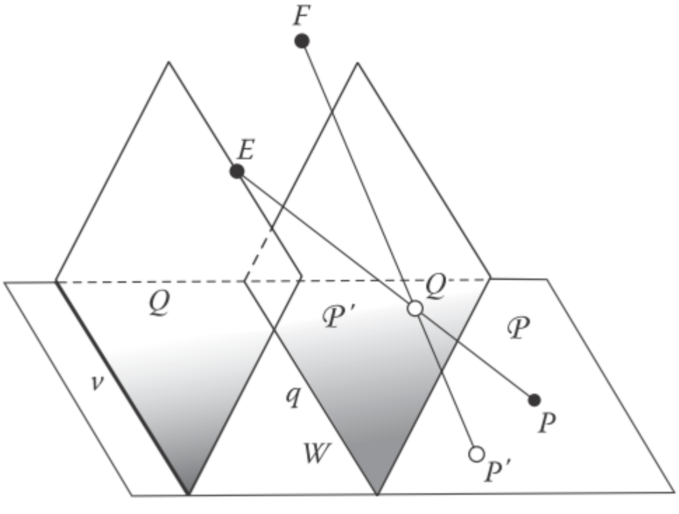
F
E
Q
Q
𝒫′
𝒫
v
q
P
W
P′

Figure 1.179: D112.png

Pag. 5 - A projective map having no vanishing line.

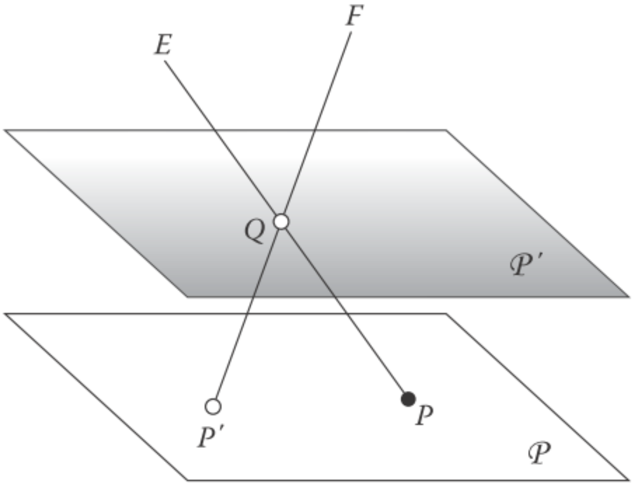

Figure 1.180: D113.png

Pag. 6 -

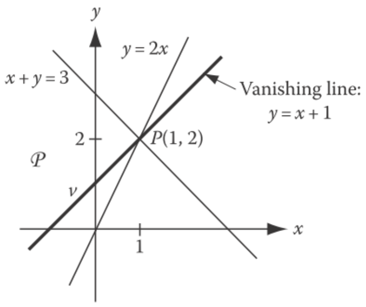
y
y = 2x
x + y = 3
Vanishing line:
y = x + 1
2
P(1, 2)
𝒫
v
1
x

Figure 1.181: D114a.png

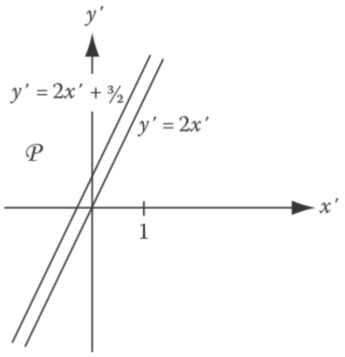
y′
y′ = 2x′ + 3/2
y′ = 2x′
𝒫
x′
1

Figure 1.182: D114b.png

Pag. 8 - IMPORTANT - Research

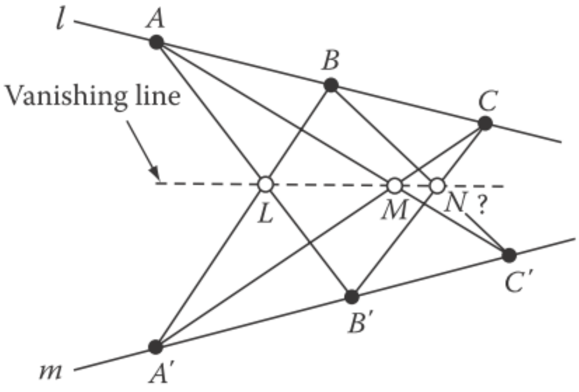
l
A
B
C
Vanishing line
L
M
N ?
C′
B′
m
A′

Figure 1.183: D115.png

Pag. 10 -

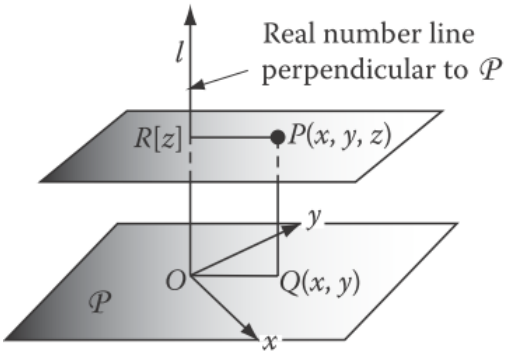
l
Real number line
perpendicular to 𝒫
R[z]
P(x, y, z)
y
O
Q(x, y)
𝒫
x

Figure 1.184: D116.png

Pag. 11 -

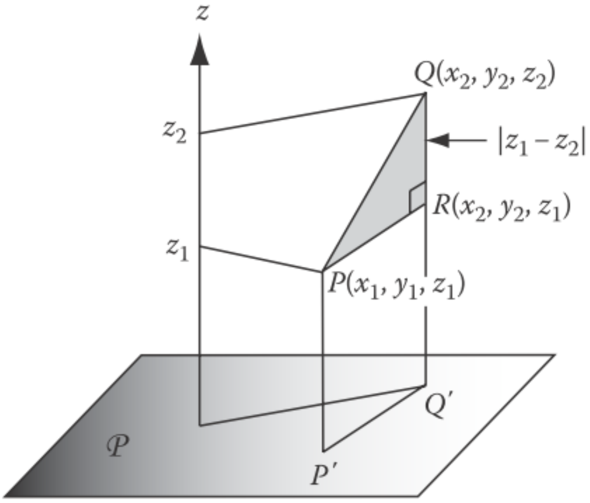
z
Q(x2, y2, z2)
z2
|z1 − z2|
R(x2, y2, z1)
z1
P(x1, y1, z1)
Q′
𝒫
P′

Figure 1.185: D117.png

Pag. 15 - Betweenness-preserving property of an affine map.

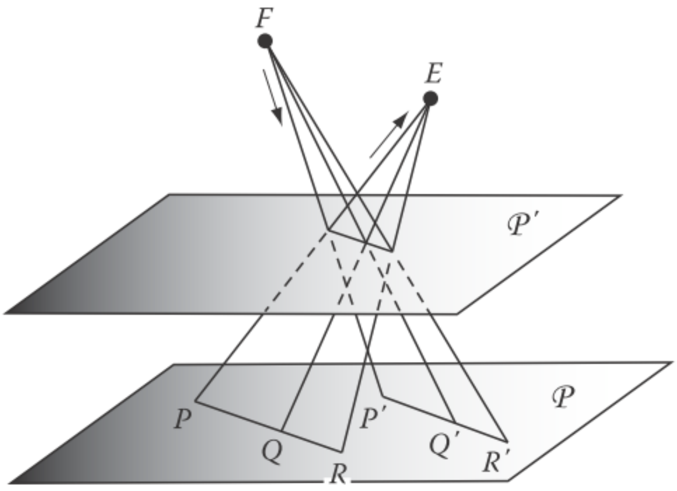
F
E
𝒫′
𝒫
P
Q
R
P′
Q′
R′

Figure 1.186: D118.png

Pag. 16 - Can a line map to part of a line?

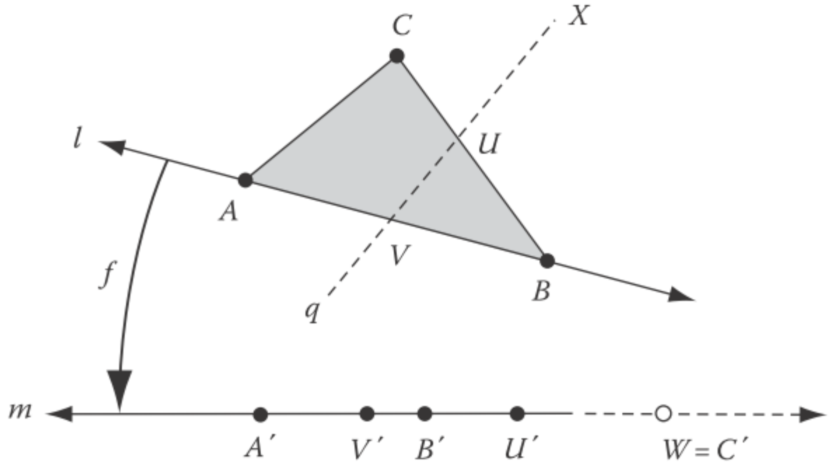
C
X
l
U
A
V
f
B
q
m
A′
V′
B′
U′
W = C′

Figure 1.187: D119.png

Pag. 18 - Affine reflection (k = 2).

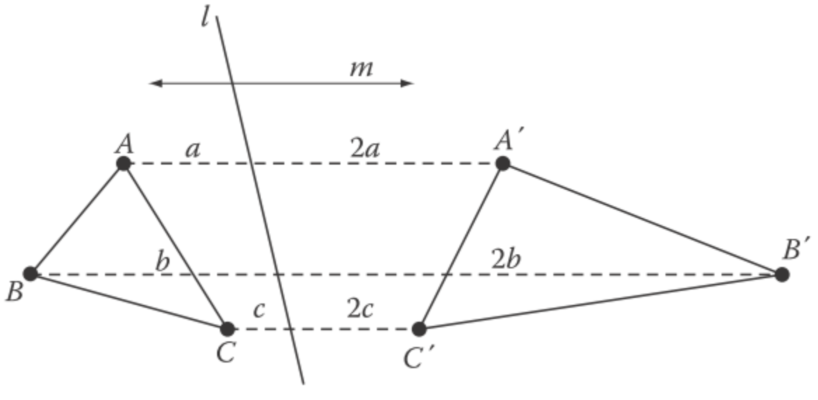
l
m
A
a
2a
A′
B
b
2b
B′
C
c
2c
C′

Figure 1.188: D120.png

Pag. 19 - Reflecting AB to DE.

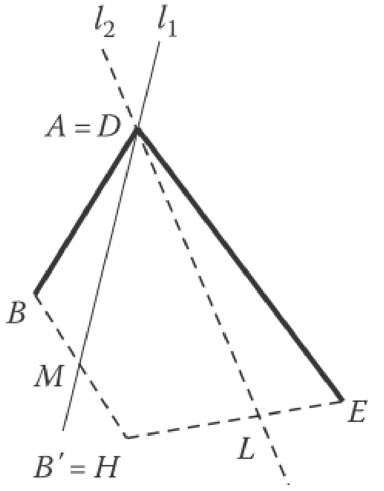
l_2
l_1
A = D
B
M
E
L
B′ = H

Figure 1.189: D121a.png

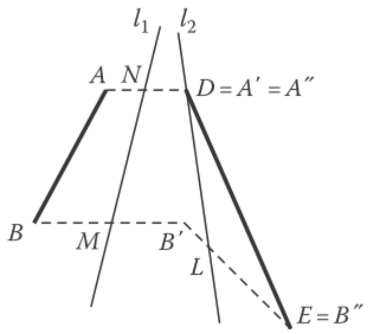
l_1
l_2
A
N
D = A′ = A″
B
M
B′
L
E = B″

Figure 1.190: D121b.png

Pag. 20 - Reflecting ABC to DEF.

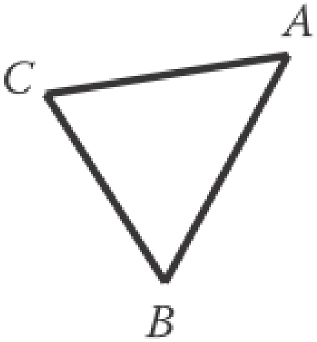

Figure 1.191: D122a.png

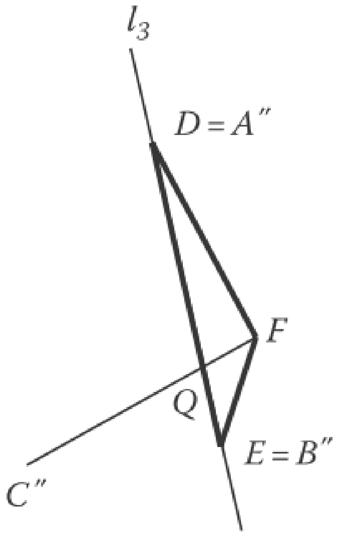

Figure 1.192: D122b.png

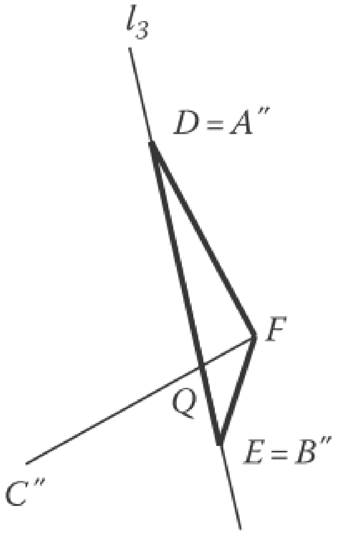

Figure 1.193: D122b.png

Pag. 26 - Dilation as a projective transformation.

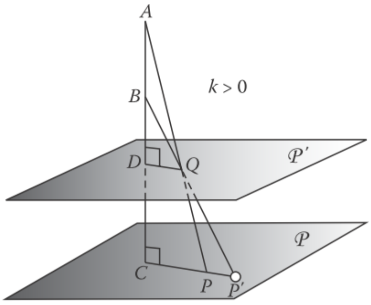
A
B
k > 0
D
Q
𝒫′
𝒫
C
P
P′

Figure 1.194: D123a.png

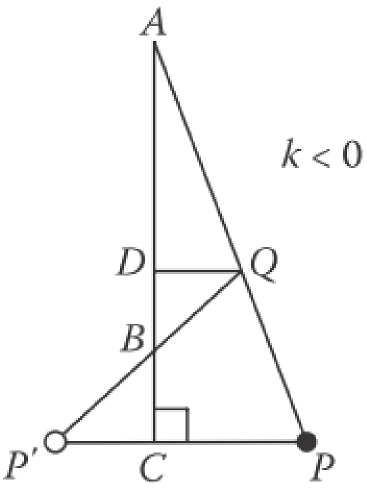

Figure 1.195: D123b.png

Pag. 27 - Inscribing a square in a triangle.

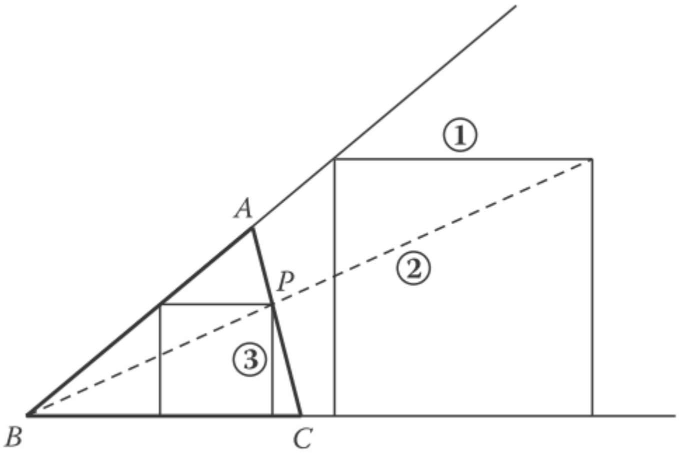

Figure 1.196: D124.png

Pag. 28 - Transforming the nine-point circle. IMPORTANT

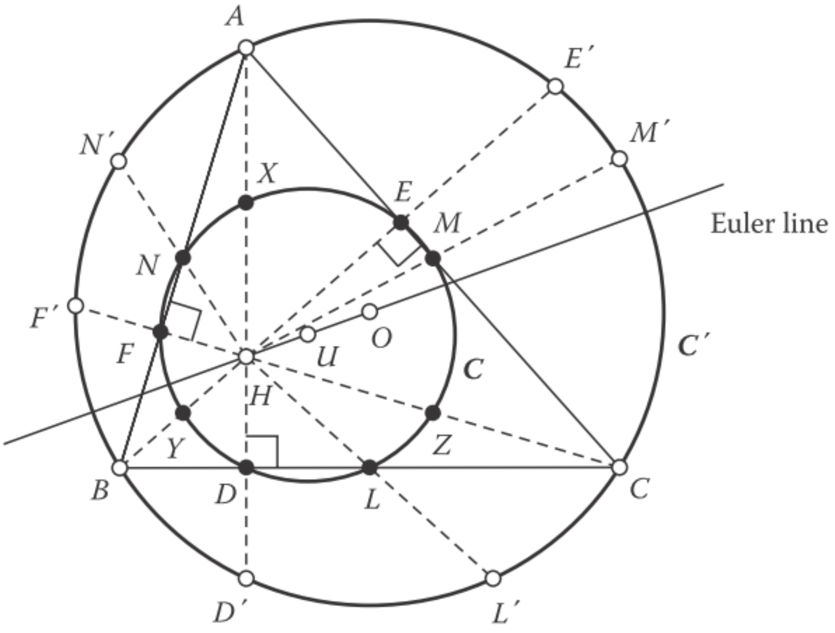
A
E′
N′
M′
X
E
M
Euler line
N
F′
O
U
C
C′
F
H
Z
Y
B
D
L
C
D′
L′

Figure 1.197: D125.png

Pag. 37 - Proving Fagnano’s theorem.

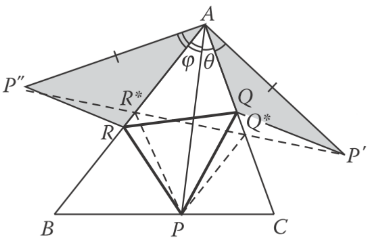
A
φ
θ
P″
R*
Q
R
Q*
P′
B
P
C

Figure 1.198: D126a.png

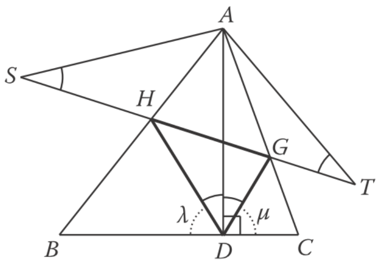
A
S
H
G
T
λ
μ
B
D
C

Figure 1.199: D126b.png

Pag. 55 - Proving the coordinate form of a rotation.

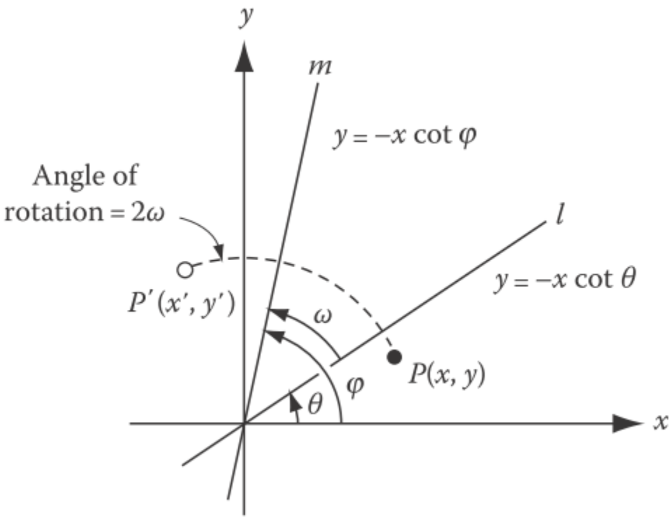
y
m
$y = -x \cot \varphi$
Angle of
rotation = 2ω
l
$y = -x \cot \theta$
$P'(x', y')$
ω
φ
$P(x, y)$
θ
x

Figure 1.200: D127.png

Pag. 57 - Using a rotation to construct the Fermat point.

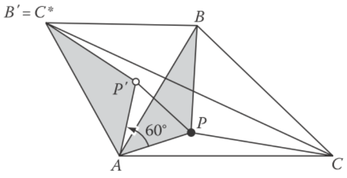

Figure 1.201: D128a.png

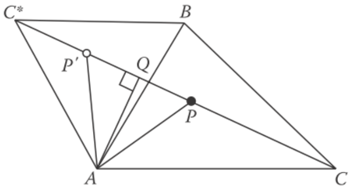

Figure 1.202: D128b.png

Pag. 58 - Napoleon's theorem

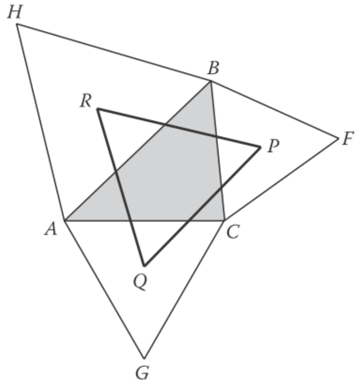

Figure 1.203: D129a.png

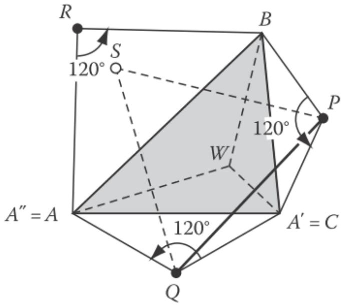
R
B
S
120°
P
120°
W
A″ = A
120°
A′ = C
Q

Figure 1.204: D129b.png

Pag. 67 - Inversion of circles not passing through O.

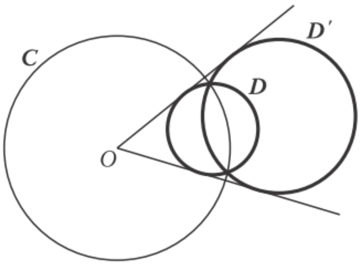

Figure 1.205: D130a.png - D intersects C

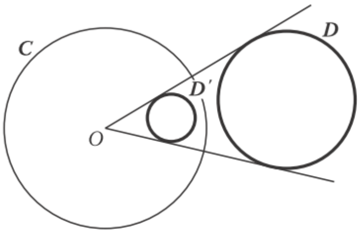

Figure 1.206: D130b.png - D lies inside, or outside of C

Pag. 68 - Invariance of circles orthogonal to C.

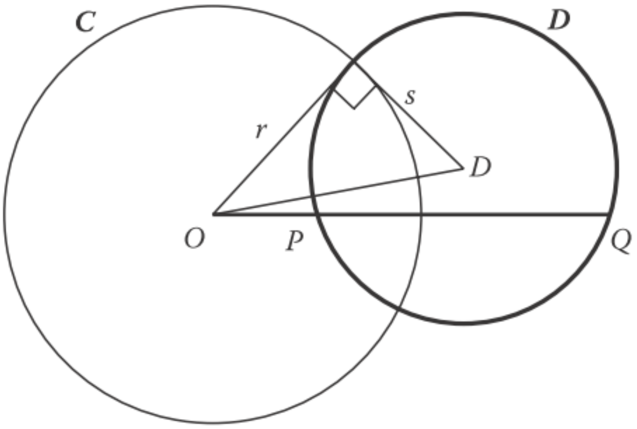

Figure 1.207: D131.png

Pag. 69 - Conformal property of inversions.

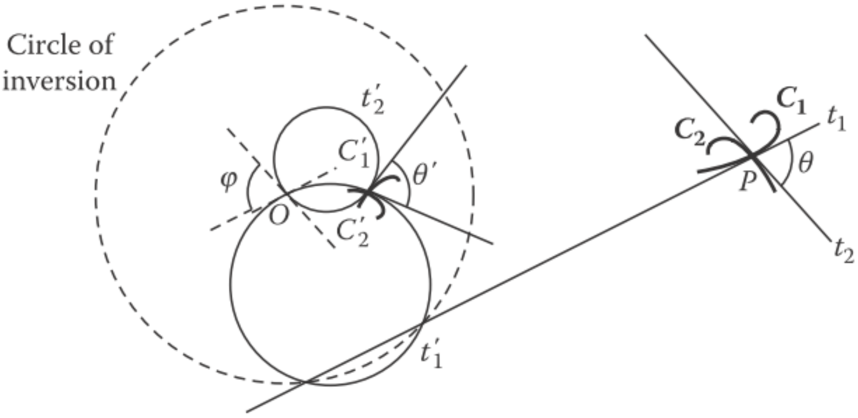
Circle of inversion
t_2'
C_1'
φ
θ'
O
C_2'
t_1'
C_2
C_1
t_1
P
θ
t_2

Figure 1.208: D132.png

Pag. 72 - Example 2 (used for Euler's formula). IMPORTANT

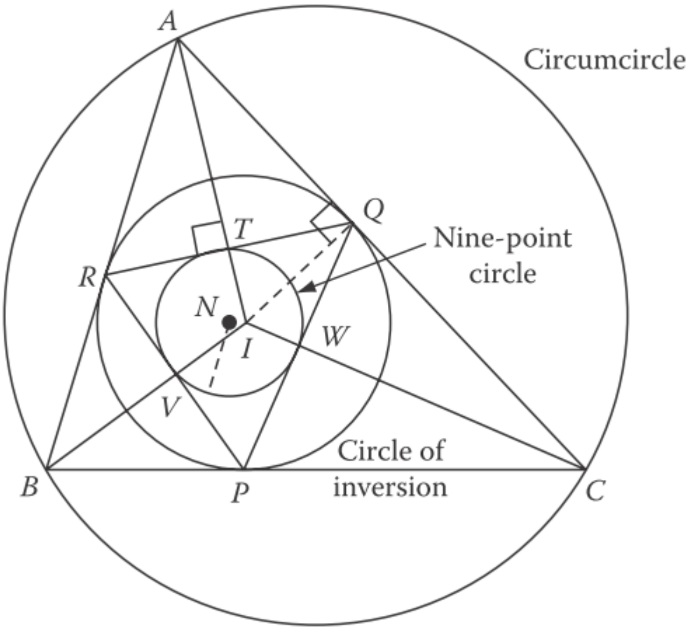
A
Circumcircle
Q
T
Nine-point
circle
R
N
I
W
V
Circle of
inversion
B
P
C

Figure 1.209: D133.png

Pag. 74 - Proving the Ptolemaic inequality.

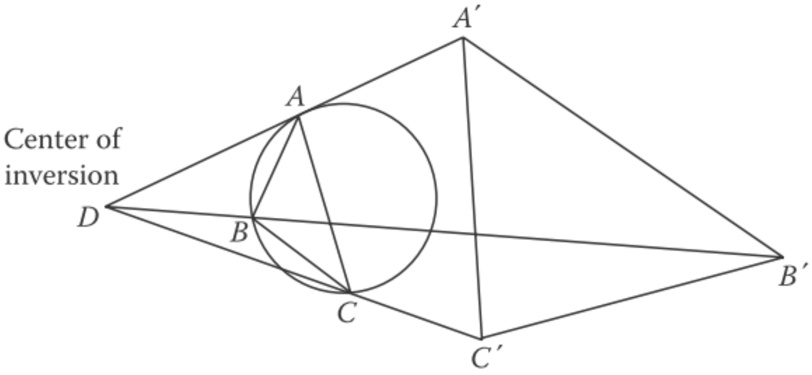
A′
A
Center of
inversion
D
B
B′
C
C′

Figure 1.210: D134.png

Pag. 75 -

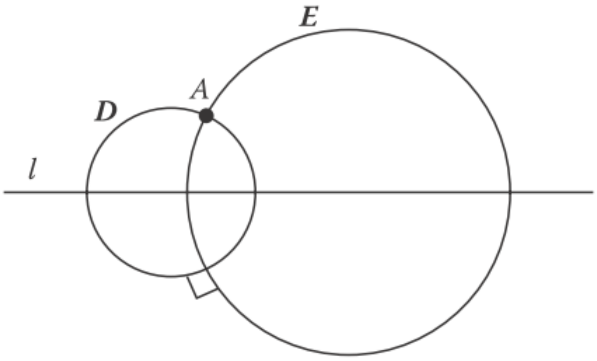

Figure 1.211: D135.png

15) C:/DATA-NovoVivo-126-2020/Books-126-2020/9781439819111-Ancillaries/K1100
C009.pdf

Pag. 7 - An elliptic triangle and its reference triangle.

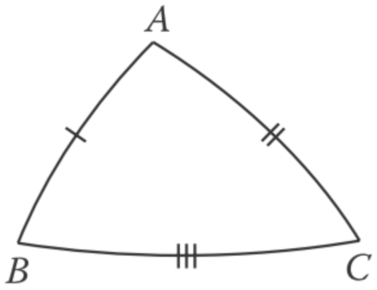

Figure 1.212: D136a.png - Elliptic plane

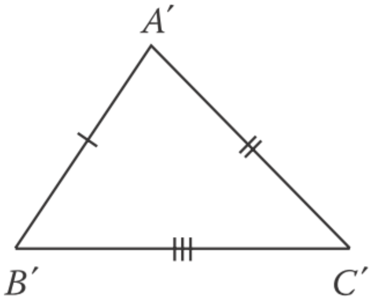

Figure 1.213: D136b.png - Euclidean plane (reference triangle)

Pag. 8 - A hyperbolic triangle and its reference triangle.

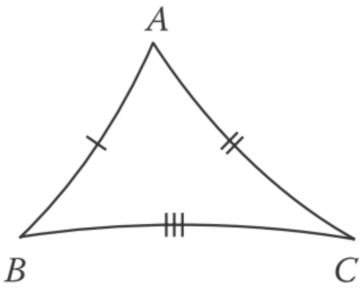

Figure 1.214: D137a.png - Hyperbolic plane

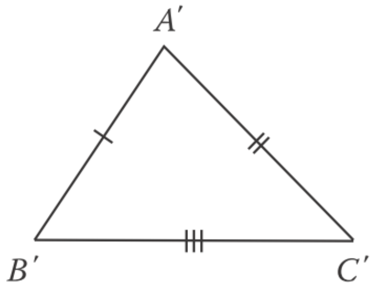

Figure 1.215: D137b.png - Euclidean plane (reference triangle)

Pag. 9 - Comparing medians.

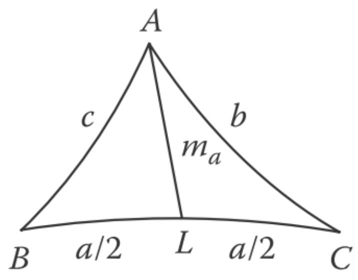

Figure 1.216: D138a.png - Hyperbolic plane

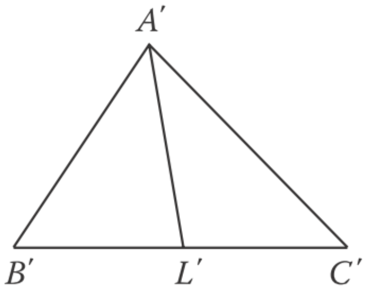

Figure 1.217: D138b.png - Reference triangle (AL < A^{0}L^0)

Pag. 23 - Poincare's disk.

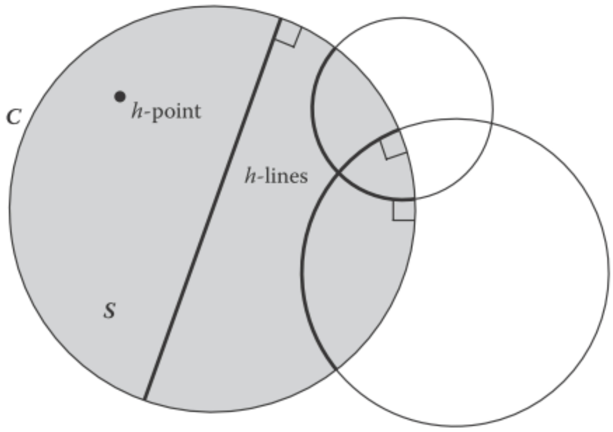
C
h-point
h-lines
S

Figure 1.218: D139.png

Pag. 24 - Equation for h-lines.

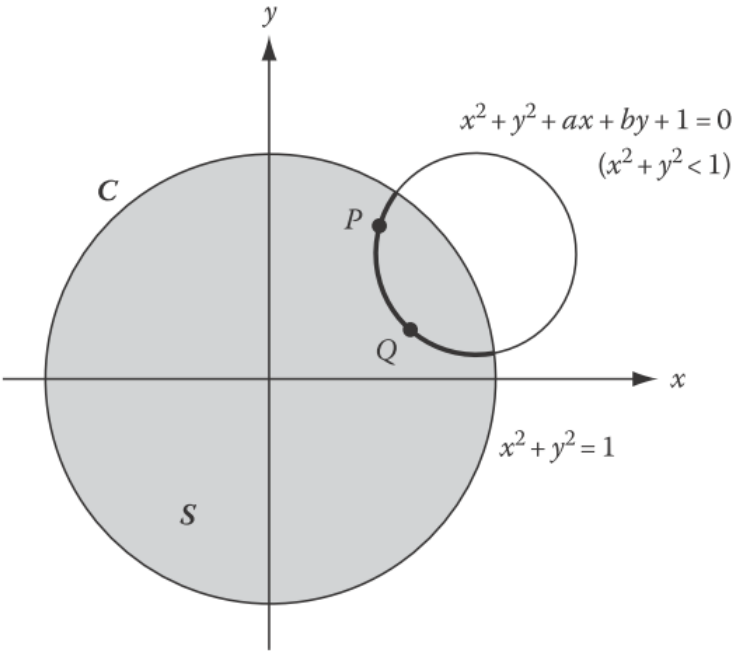
y
x
$x^2 + y^2 + ax + by + 1 = 0$
$(x^2 + y^2 < 1)$
C
P
Q
$x^2 + y^2 = 1$
S

Figure 1.219: D140.png

Pag. 25 - Distance in the Poincare model.

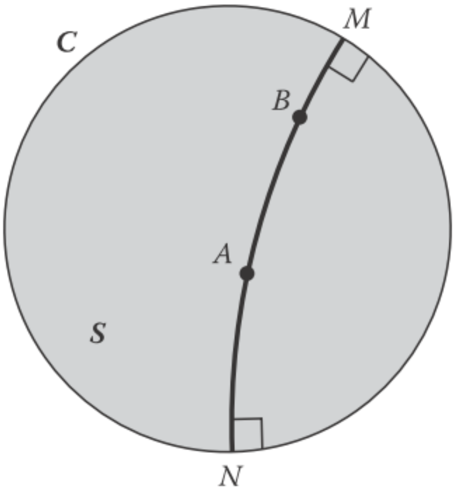

Figure 1.220: D141.png

Pag. 26 - Betweenness in the Poincare model.

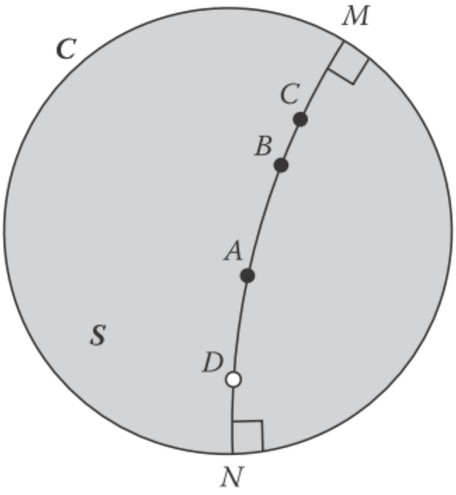

Figure 1.221: D142.png

Pag. 27 - Angle measure in the Poincare model.

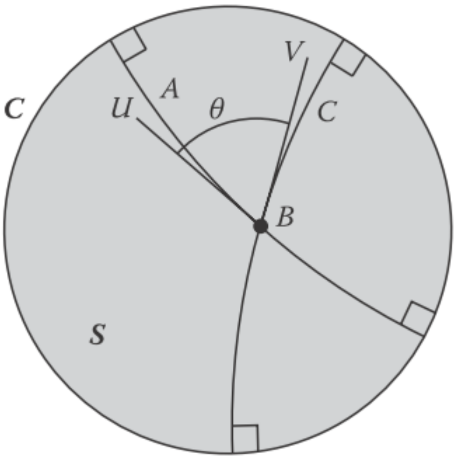

Figure 1.222: D143.png

Pag. 28 - Euclidean distance and h-distance related.

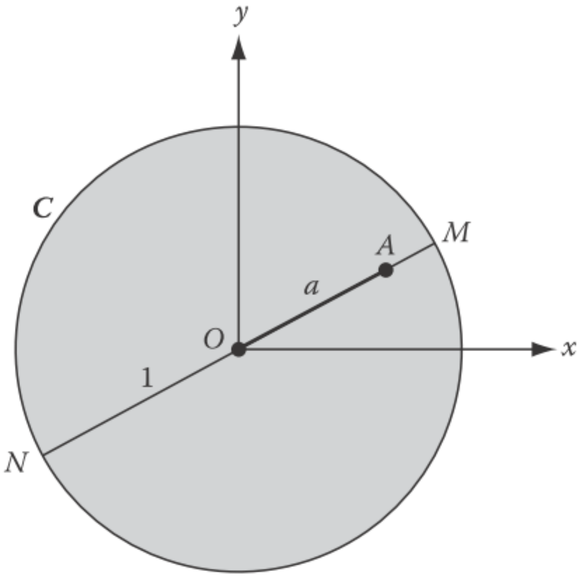

Figure 1.223: D144.png

Pag. 29 -

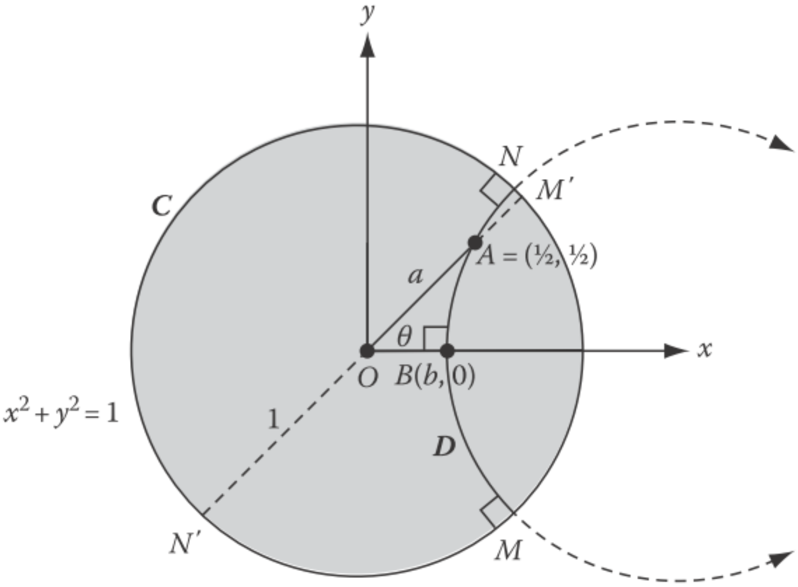
y
N
M′
C
A = (½, ½)
a
θ
x
O B(b, 0)
x² + y² = 1
1
D
N′
M

Figure 1.224: D145.png

Pag. 30 - Plane separation postulate verified.

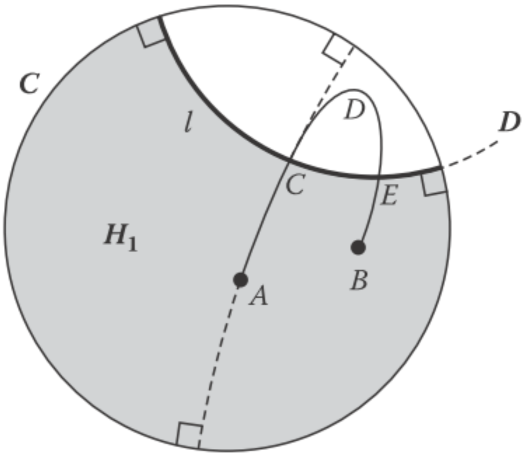
C
l
D
D
C
E
H₁
A
B

Figure 1.225: D146.png

Pag. 33 - Constructing the midpoint of segmento AB

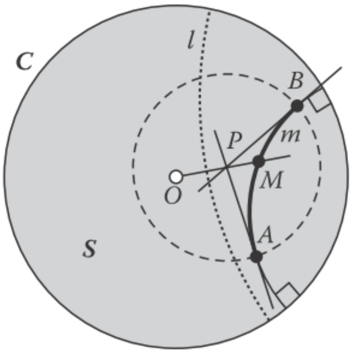

Figure 1.226: D147a.png

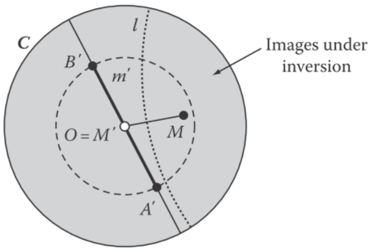
c
l
B′
m′
O = M′
M
A′
Images under inversion

Figure 1.227: D147b.png

Pag. 36 -

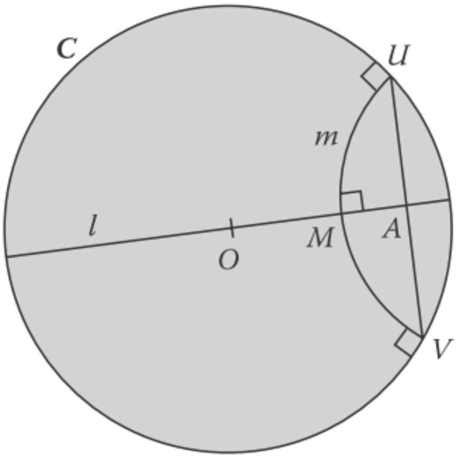

Figure 1.228: D148.png

Pag. 37 - Geometry on a surface

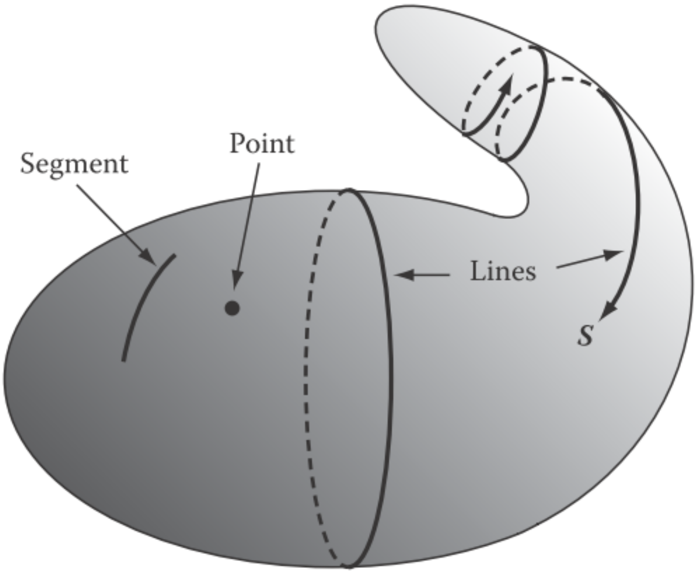
Segment
Point
Lines
S

Figure 1.229: D149.png

Pag. 38 - Local isometry between two surfaces.

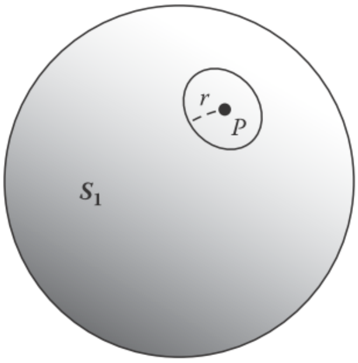

Figure 1.230: D150a.png

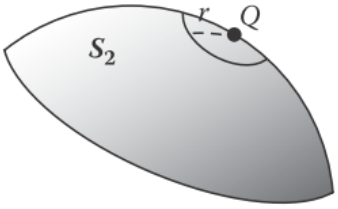

Figure 1.231: D150b.png

Pag. 40 - Euclidean geometry on a cylinder. Euclidean geometry on the circular cylinder

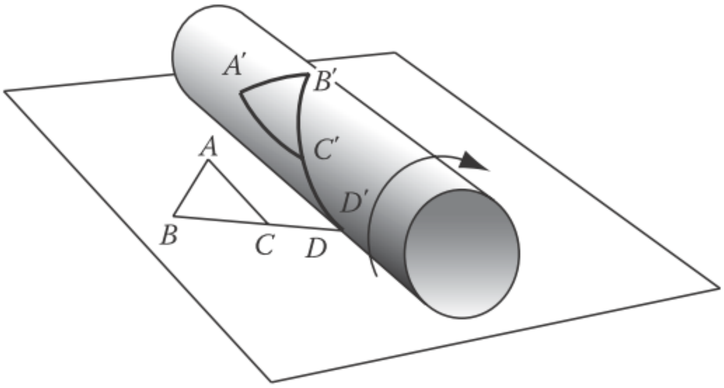

Figure 1.232: D151.png

Pag. 41 - Euclidean geometry on a cone. Euclidean geometry on the circular cone

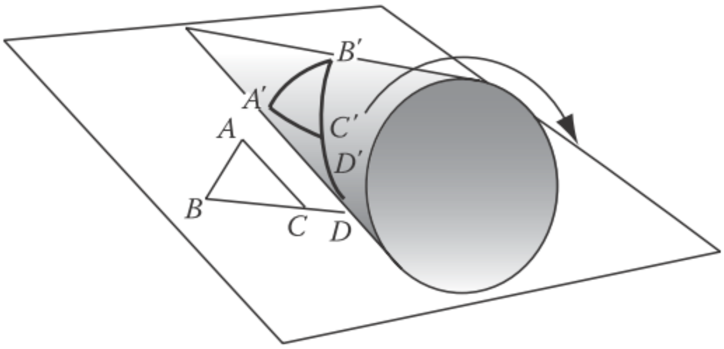

Figure 1.233: D152.png

Pag. 42 - Euclidean geometry on a sphere?

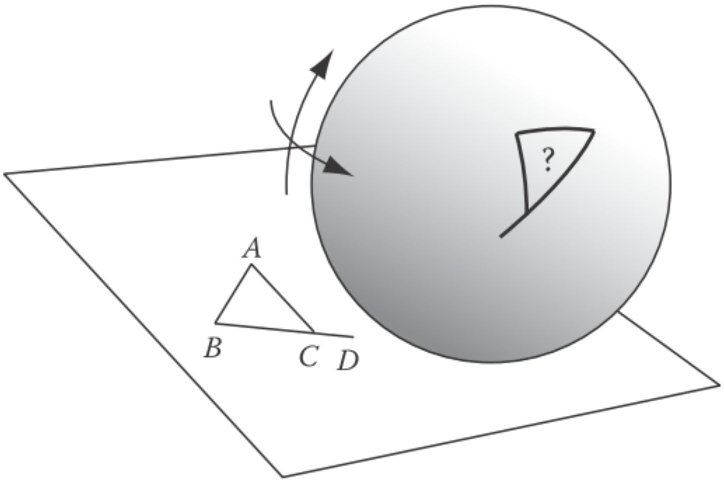

Figure 1.234: D153.png

Pag. 44 -Spherical coordinates for the unit sphere

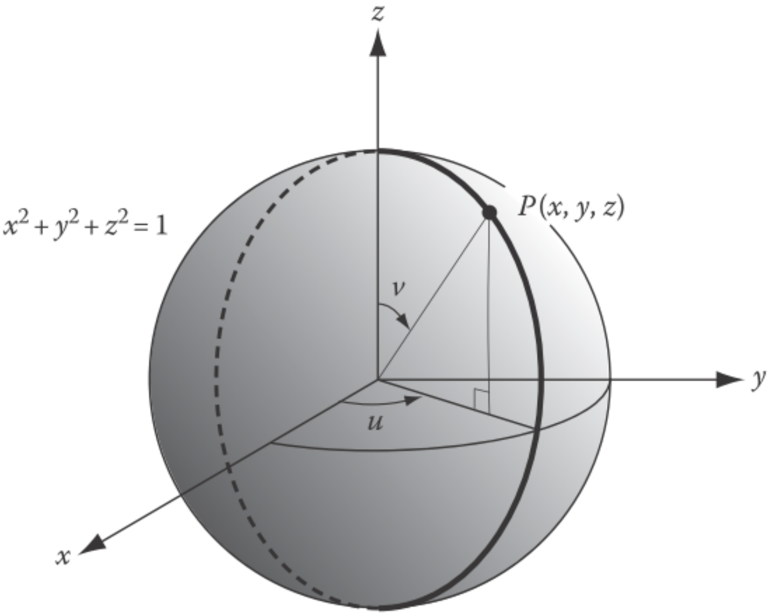
z
$x^2+y^2+z^2=1$
P(x, y, z)
v
u
y
x

Figure 1.235: D154.png

Pag. 45 - Proving Bolyai's formula for the angle of parallelism.

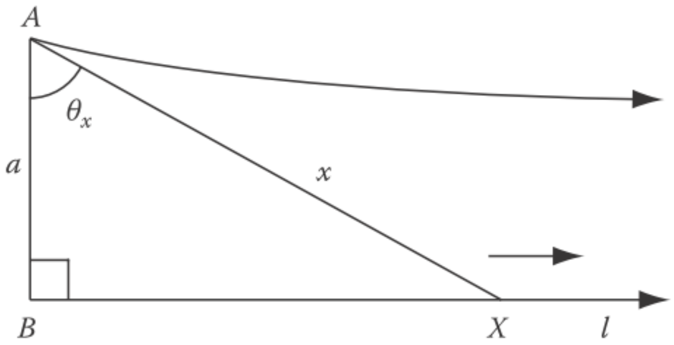
A
θ_x
a
x
B
X
l

Figure 1.236: D155.png

Pag. 46 - Angle of parallelism of measure^{o}1

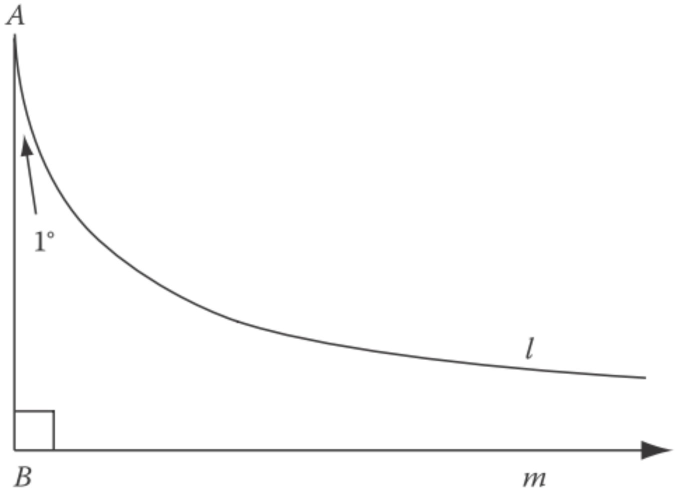

Figure 1.237: D156.png

Pag. 47 - Hyperbolic parallels.

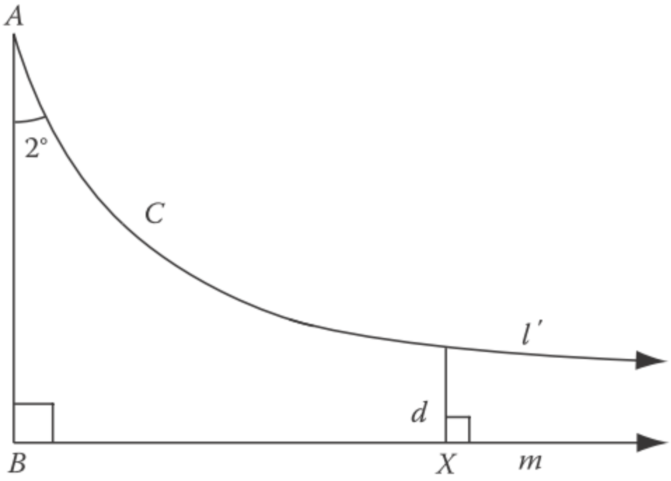

Figure 1.238: D157.png

Pag. 48 - Circles in Euclidean geometry.

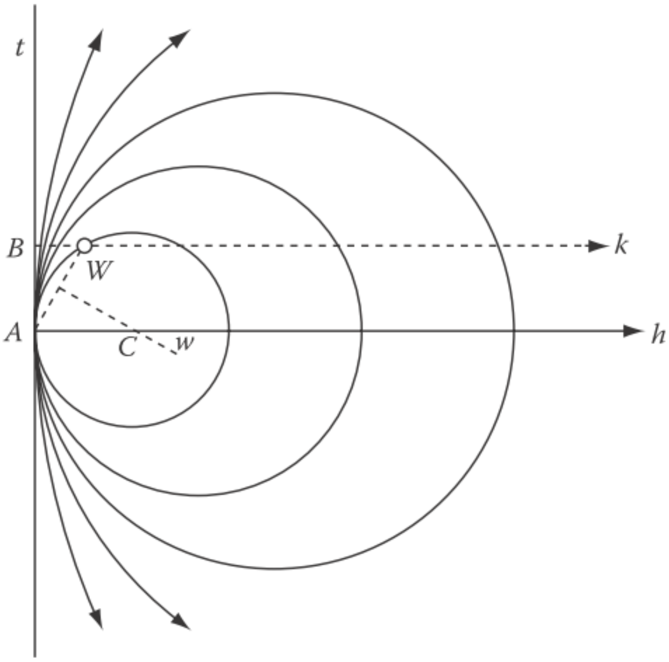

Figure 1.239: D158.png

Pag. 51 - Cycles in hyperbolic geometry.

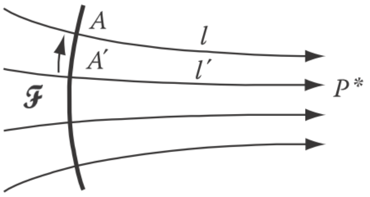

Figure 1.240: D150a.png - Limit curve

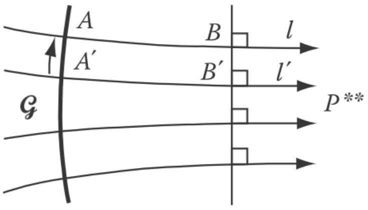

Figure 1.241: D150b.png - Equidistant locus

Pag. 52 - Perpendicular bisectors concurrent at .

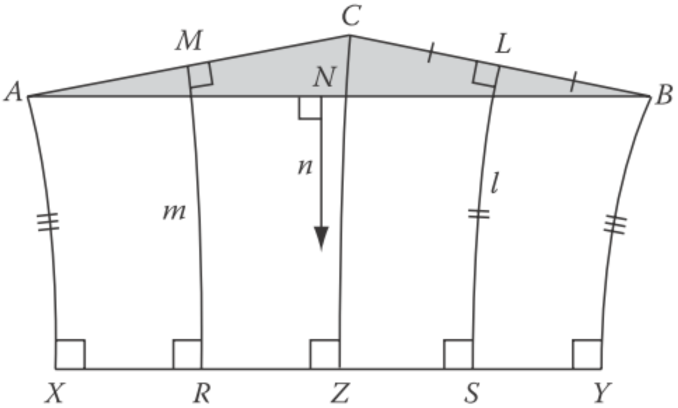
C
M
L
N
A
B
n
m
l
X
R
Z
S
Y

Figure 1.242: D160.png

Pag. 54 - Cycles in hyperbolic geometry.

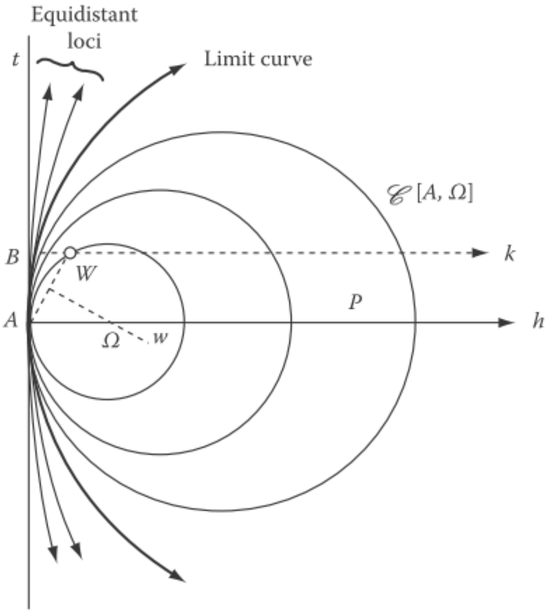
Equidistant
loci
t
Limit curve
$\mathscr{C}[A, \Omega]$
B
W
k
A
P
h
Ω
w

Figure 1.243: D161.png

Pag. 57 -

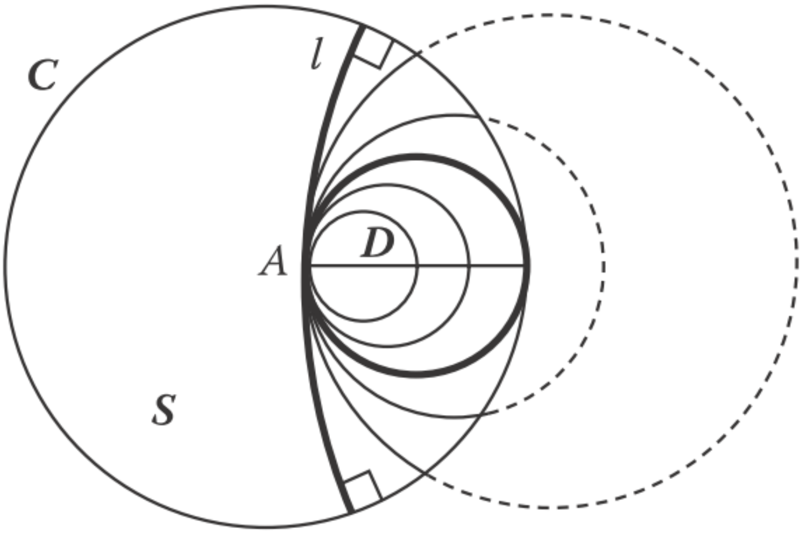

Figure 1.244: D162.png

Pag. 58 -

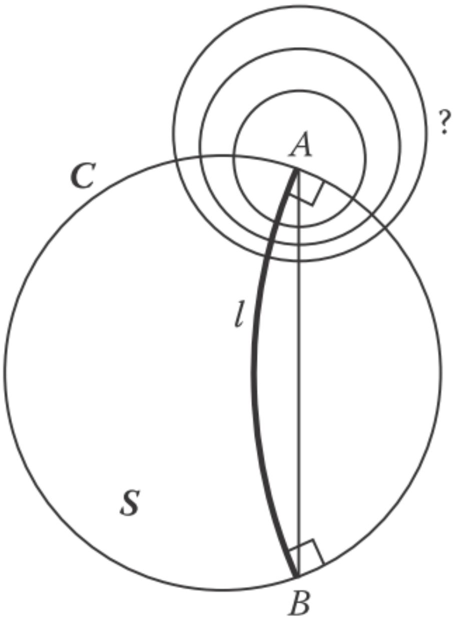

Figure 1.245: D163.png

Pag. 59 -

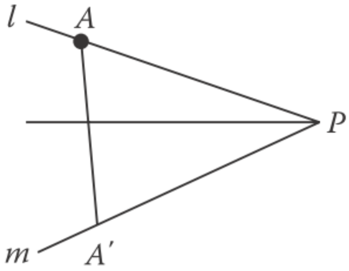

Figure 1.246: D164a.png

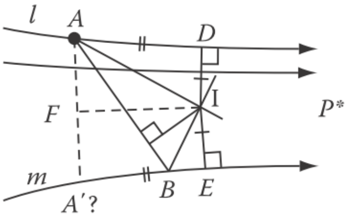
l
A
D
I
F
P*
m
B
E
A′?

Figure 1.247: D164b.png

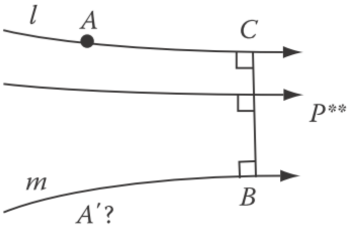

Figure 1.248: D164c.png

16) C:/DATA-NovoVivo-126-2020/Books-126-2020/9781439819111-Ancillaries/K11008-C010.pdf
Pag. 09 -

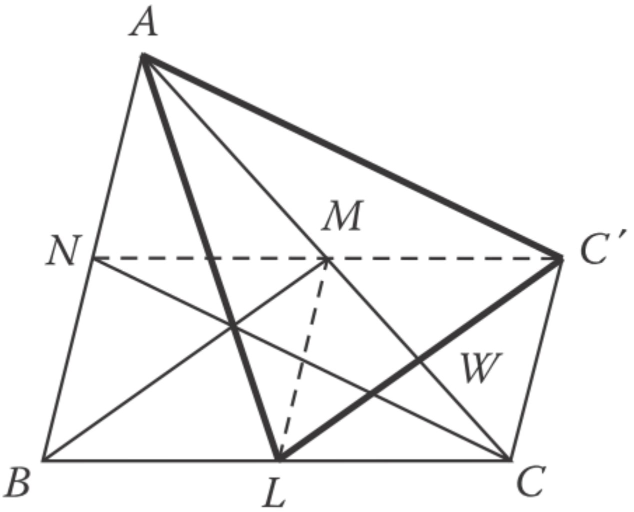

Figure 1.249: D165.png

Pag. 16 -

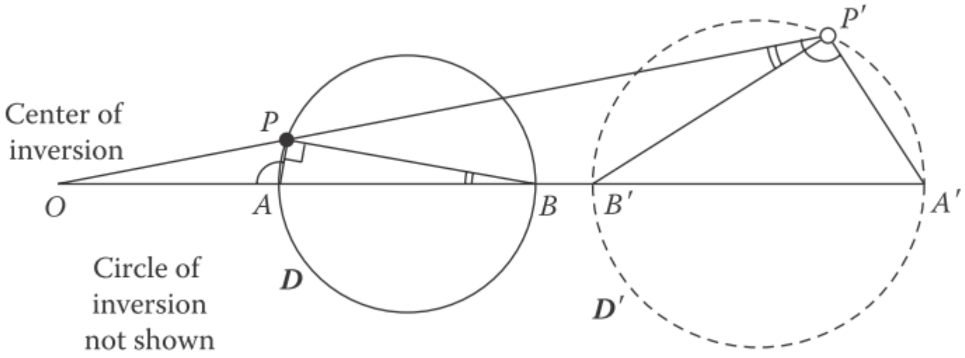
P′
Center of inversion
P
O
A
B
B′
A′
Circle of inversion not shown
D
D′

Figure 1.250: D166.png

Pag. 18 -

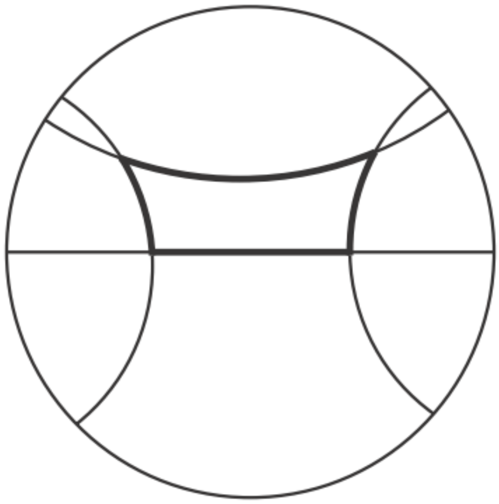

Figure 1.251: D167.png

17) C:/DATA-NovoVivo-126-2020/Books-126-2020/9781439819111-Ancillaries/Kay - College Geometry - A Unified Development 2011.pdfIMPORTANT / Solution: a) C:/DATA-NovoVivo-126-2020/Books-126-2020/9781439819111-Ancillaries/97814 / Obs.: Todos os itens anteriores estão relacionados com este livro. / b)

C:/DATA-NovoVivo-126-2020/Books-126-2020/9781439819111-Ancillaries/Kay - Trigonometry - Cli s Quick Review 2001.pdf / c) C:/DATA-NovoVivo-126-2020/Books-126-2020/9781439819111-Ancillaries/Using Cases to Transform Mathematics Teaching And Learning.pdf IMPORTANT / d) C:/DATA-NovoVivo-126-2020/Books-126-2020/9781439819111-Ancillaries/Cli sTestPrep ACT - 2005.pdf / e) C:/DATA-NovoVivo-126-2020/Books-126-2020/9781439819111-Ancillaries/CSM VCE mathematical methods units 1 and 2 - 2015.pdf IMPORTANT /

18) C:/DATA-NovoVivo-126-2020/Books-126-2020/9781466569980-Ancillaries/Classical Mechanics - Chow, Second Edition (2013, CRC Press).pdf a) Copy: C:/DATA-NovoVivo-126-2020/Books-126-2020/9781466569980-Ancillaries/Classical Mechanics - 2013 - CRC Press.pdf b) C:/DATA-NovoVivo-126-2020/Books-126-2020/9781466569980-Ancillaries/Classical Mechanics - 2013 - res.pdf Important: Images in the same directory. IMPORTANT

1.11 Books-126-2020 - Index

1) C:/DATA-NovoVivo-126-2020/Books-126-2020/3D-Rotations/3D Rotations.pdf

Pag. 07 (18) - Chapter 2 - Geometry of Rotations IMPORTANT

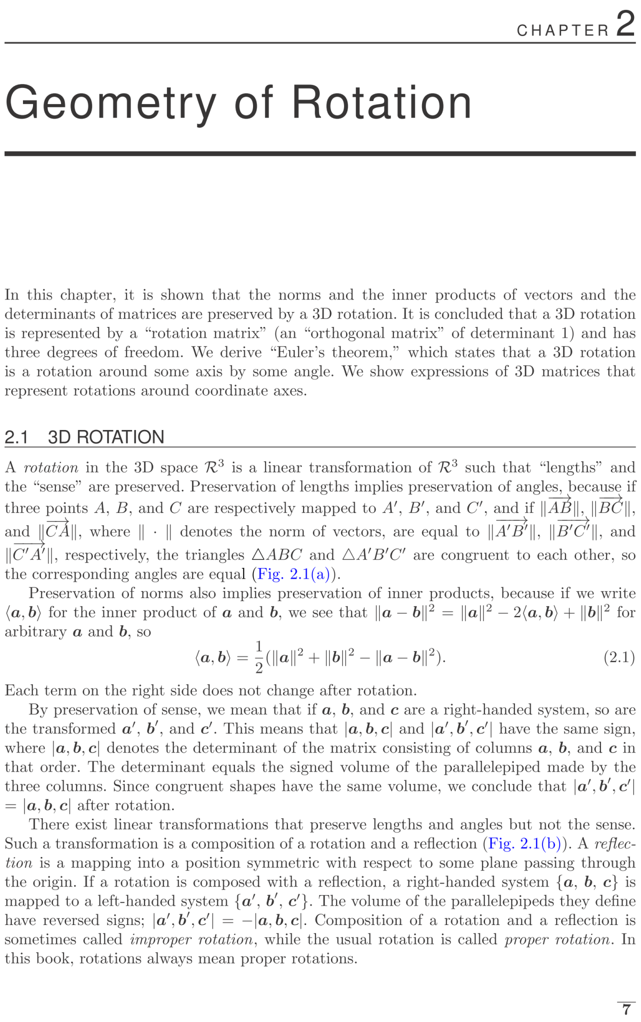

CHAPTER 2

Geometry of Rotation

In this chapter, it is shown that the norms and the inner products of vectors and the determinants of matrices are preserved by a 3D rotation. It is concluded that a 3D rotation is represented by a "rotation matrix" (an "orthogonal matrix" of determinant 1) and has three degrees of freedom. We derive "Euler's theorem," which states that a 3D rotation is a rotation around some axis by some angle. We show expressions of 3D matrices that represent rotations around coordinate axes.

2.1 3D ROTATION

A *rotation* in the 3D space $\mathcal{R}^3$ is a linear transformation of $\mathcal{R}^3$ such that "lengths" and the "sense" are preserved. Preservation of lengths implies preservation of angles, because if three points A, B, and C are respectively mapped to A', B', and C', and if $\|\overrightarrow{AB}\|$, $\|\overrightarrow{BC}\|$, and $\|\overrightarrow{CA}\|$, where $\|\cdot\|$ denotes the norm of vectors, are equal to $\|\overrightarrow{A'B'}\|$, $\|\overrightarrow{B'C'}\|$, and $\|\overrightarrow{C'A'}\|$, respectively, the triangles $\triangle ABC$ and $\triangle A'B'C'$ are congruent to each other, so the corresponding angles are equal (Fig. 2.1(a)).

Preservation of norms also implies preservation of inner products, because if we write $\langle \boldsymbol{a}, \boldsymbol{b} \rangle$ for the inner product of $\boldsymbol{a}$ and $\boldsymbol{b}$, we see that $\|\boldsymbol{a}-\boldsymbol{b}\|^2 = \|\boldsymbol{a}\|^2 - 2\langle \boldsymbol{a}, \boldsymbol{b} \rangle + \|\boldsymbol{b}\|^2$ for arbitrary $\boldsymbol{a}$ and $\boldsymbol{b}$, so

$$\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \frac{1}{2}(\|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2 - \|\boldsymbol{a}-\boldsymbol{b}\|^2). \tag{2.1}$$

Each term on the right side does not change after rotation.

By preservation of sense, we mean that if $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ are a right-handed system, so are the transformed $\boldsymbol{a}'$, $\boldsymbol{b}'$, and $\boldsymbol{c}'$. This means that $|\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}|$ and $|\boldsymbol{a}', \boldsymbol{b}', \boldsymbol{c}'|$ have the same sign, where $|\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}|$ denotes the determinant of the matrix consisting of columns $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ in that order. The determinant equals the signed volume of the parallelepiped made by the three columns. Since congruent shapes have the same volume, we conclude that $|\boldsymbol{a}', \boldsymbol{b}', \boldsymbol{c}'| = |\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}|$ after rotation.

There exist linear transformations that preserve lengths and angles but not the sense. Such a transformation is a composition of a rotation and a reflection (Fig. 2.1(b)). A *reflection* is a mapping into a position symmetric with respect to some plane passing through the origin. If a rotation is composed with a reflection, a right-handed system $\{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}\}$ is mapped to a left-handed system $\{\boldsymbol{a}', \boldsymbol{b}', \boldsymbol{c}'\}$. The volume of the parallelepipeds they define have reversed signs; $|\boldsymbol{a}', \boldsymbol{b}', \boldsymbol{c}'| = -|\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}|$. Composition of a rotation and a reflection is sometimes called *improper rotation*, while the usual rotation is called *proper rotation*. In this book, rotations always mean proper rotations.

Figure 1.252: 3D-Rotations-01.png

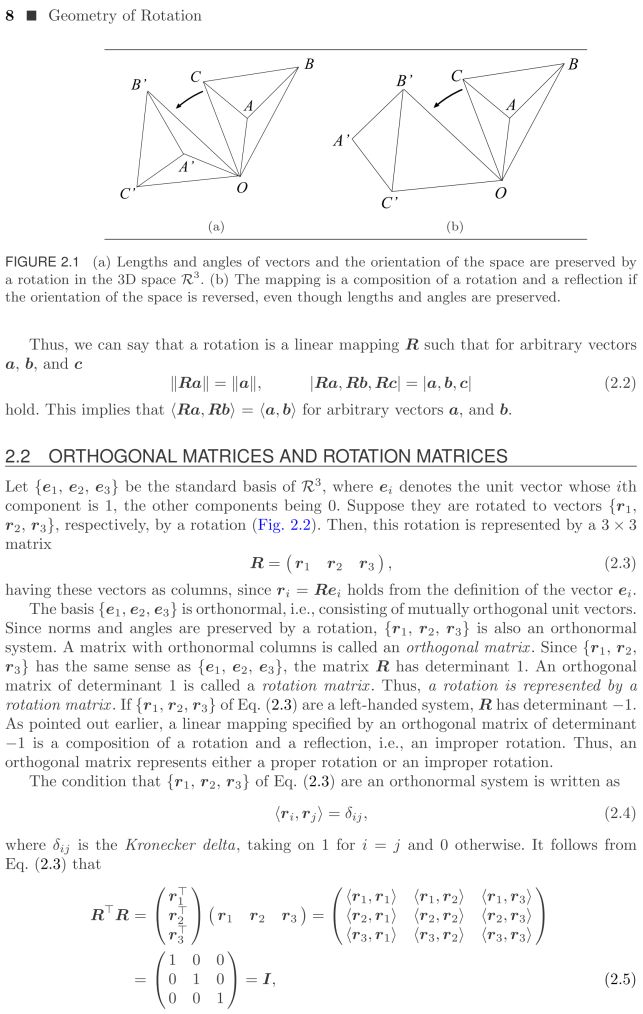
B
C
B'
A
A'
C'
O
B
C
B'
A
A'
C'
O

(a) (b)

FIGURE 2.1 (a) Lengths and angles of vectors and the orientation of the space are preserved by a rotation in the 3D space $\mathcal{R}^3$. (b) The mapping is a composition of a rotation and a reflection if the orientation of the space is reversed, even though lengths and angles are preserved.

Thus, we can say that a rotation is a linear mapping $\boldsymbol{R}$ such that for arbitrary vectors $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$

$$\|\boldsymbol{Ra}\| = \|\boldsymbol{a}\|, \qquad |\boldsymbol{Ra}, \boldsymbol{Rb}, \boldsymbol{Rc}| = |\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}| \tag{2.2}$$

hold. This implies that $\langle \boldsymbol{Ra}, \boldsymbol{Rb} \rangle = \langle \boldsymbol{a}, \boldsymbol{b} \rangle$ for arbitrary vectors $\boldsymbol{a}$, and $\boldsymbol{b}$.

2.2 ORTHOGONAL MATRICES AND ROTATION MATRICES

Let $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3\}$ be the standard basis of $\mathcal{R}^3$, where $\boldsymbol{e}_i$ denotes the unit vector whose ith component is 1, the other components being 0. Suppose they are rotated to vectors $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3\}$, respectively, by a rotation (Fig. 2.2). Then, this rotation is represented by a 3×3 matrix

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{r}_1 & \boldsymbol{r}_2 & \boldsymbol{r}_3 \end{pmatrix}, \tag{2.3}$$

having these vectors as columns, since $\boldsymbol{r}_i = \boldsymbol{Re}_i$ holds from the definition of the vector $\boldsymbol{e}_i$.

The basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3\}$ is orthonormal, i.e., consisting of mutually orthogonal unit vectors. Since norms and angles are preserved by a rotation, $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3\}$ is also an orthonormal system. A matrix with orthonormal columns is called an *orthogonal matrix*. Since $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3\}$ has the same sense as $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3\}$, the matrix $\boldsymbol{R}$ has determinant 1. An orthogonal matrix of determinant 1 is called a *rotation matrix*. Thus, *a rotation is represented by a rotation matrix*. If $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3\}$ of Eq. (2.3) are a left-handed system, $\boldsymbol{R}$ has determinant -1. As pointed out earlier, a linear mapping specified by an orthogonal matrix of determinant -1 is a composition of a rotation and a reflection, i.e., an improper rotation. Thus, an orthogonal matrix represents either a proper rotation or an improper rotation.

The condition that $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3\}$ of Eq. (2.3) are an orthonormal system is written as

$$\langle \boldsymbol{r}_i, \boldsymbol{r}_j \rangle = \delta_{ij}, \tag{2.4}$$

where δ_{ij} is the *Kronecker delta*, taking on 1 for $i = j$ and 0 otherwise. It follows from Eq. (2.3) that

$$\begin{aligned} \boldsymbol{R}^\top \boldsymbol{R} &= \begin{pmatrix} \boldsymbol{r}_1^\top \\ \boldsymbol{r}_2^\top \\ \boldsymbol{r}_3^\top \end{pmatrix} \begin{pmatrix} \boldsymbol{r}_1 & \boldsymbol{r}_2 & \boldsymbol{r}_3 \end{pmatrix} = \begin{pmatrix} \langle \boldsymbol{r}_1, \boldsymbol{r}_1 \rangle & \langle \boldsymbol{r}_1, \boldsymbol{r}_2 \rangle & \langle \boldsymbol{r}_1, \boldsymbol{r}_3 \rangle \\ \langle \boldsymbol{r}_2, \boldsymbol{r}_1 \rangle & \langle \boldsymbol{r}_2, \boldsymbol{r}_2 \rangle & \langle \boldsymbol{r}_2, \boldsymbol{r}_3 \rangle \\ \langle \boldsymbol{r}_3, \boldsymbol{r}_1 \rangle & \langle \boldsymbol{r}_3, \boldsymbol{r}_2 \rangle & \langle \boldsymbol{r}_3, \boldsymbol{r}_3 \rangle \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \boldsymbol{I}, \end{aligned} \tag{2.5}$$

Figure 1.253: 3D-Rotations-02.png

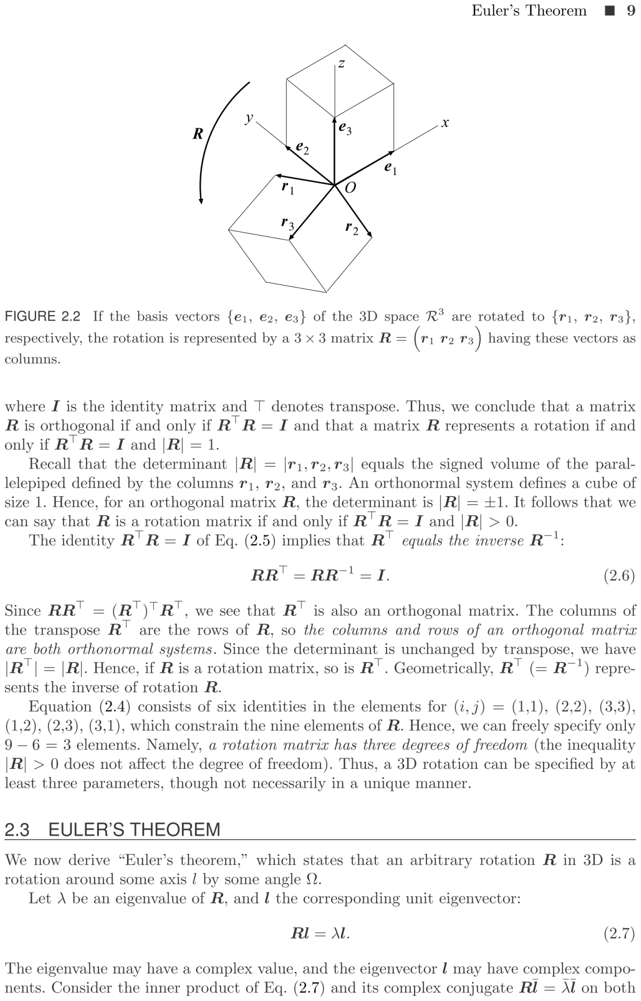

$\boldsymbol{R}$
z
y
x
$\boldsymbol{e}_3$
$\boldsymbol{e}_2$
$\boldsymbol{e}_1$
$\boldsymbol{r}_1$
O
$\boldsymbol{r}_3$
$\boldsymbol{r}_2$

FIGURE 2.2 If the basis vectors $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3\}$ of the 3D space $\mathcal{R}^3$ are rotated to $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3\}$, respectively, the rotation is represented by a 3×3 matrix $\boldsymbol{R} = \left(\boldsymbol{r}_1\ \boldsymbol{r}_2\ \boldsymbol{r}_3\right)$ having these vectors as columns.

where $\boldsymbol{I}$ is the identity matrix and $\top$ denotes transpose. Thus, we conclude that a matrix $\boldsymbol{R}$ is orthogonal if and only if $\boldsymbol{R}^\top\boldsymbol{R} = \boldsymbol{I}$ and that a matrix $\boldsymbol{R}$ represents a rotation if and only if $\boldsymbol{R}^\top\boldsymbol{R} = \boldsymbol{I}$ and $|\boldsymbol{R}| = 1$.

Recall that the determinant $|\boldsymbol{R}| = |\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3|$ equals the signed volume of the parallelepiped defined by the columns $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, and $\boldsymbol{r}_3$. An orthonormal system defines a cube of size 1. Hence, for an orthogonal matrix $\boldsymbol{R}$, the determinant is $|\boldsymbol{R}| = \pm 1$. It follows that we can say that $\boldsymbol{R}$ is a rotation matrix if and only if $\boldsymbol{R}^\top\boldsymbol{R} = \boldsymbol{I}$ and $|\boldsymbol{R}| > 0$.

The identity $\boldsymbol{R}^\top\boldsymbol{R} = \boldsymbol{I}$ of Eq. (2.5) implies that $\boldsymbol{R}^\top$ *equals the inverse* $\boldsymbol{R}^{-1}$:

$$\boldsymbol{R}\boldsymbol{R}^\top = \boldsymbol{R}\boldsymbol{R}^{-1} = \boldsymbol{I}. \tag{2.6}$$

Since $\boldsymbol{R}\boldsymbol{R}^\top = (\boldsymbol{R}^\top)^\top\boldsymbol{R}^\top$, we see that $\boldsymbol{R}^\top$ is also an orthogonal matrix. The columns of the transpose $\boldsymbol{R}^\top$ are the rows of $\boldsymbol{R}$, so *the columns and rows of an orthogonal matrix are both orthonormal systems*. Since the determinant is unchanged by transpose, we have $|\boldsymbol{R}^\top| = |\boldsymbol{R}|$. Hence, if $\boldsymbol{R}$ is a rotation matrix, so is $\boldsymbol{R}^\top$. Geometrically, $\boldsymbol{R}^\top$ $(= \boldsymbol{R}^{-1})$ represents the inverse of rotation $\boldsymbol{R}$.

Equation (2.4) consists of six identities in the elements for (i, j) = (1,1), (2,2), (3,3), (1,2), (2,3), (3,1), which constrain the nine elements of $\boldsymbol{R}$. Hence, we can freely specify only $9 - 6 = 3$ elements. Namely, *a rotation matrix has three degrees of freedom* (the inequality $|\boldsymbol{R}| > 0$ does not affect the degree of freedom). Thus, a 3D rotation can be specified by at least three parameters, though not necessarily in a unique manner.

2.3 EULER'S THEOREM

We now derive "Euler's theorem," which states that an arbitrary rotation $\boldsymbol{R}$ in 3D is a rotation around some axis l by some angle Ω.

Let λ be an eigenvalue of $\boldsymbol{R}$, and $\boldsymbol{l}$ the corresponding unit eigenvector:

$$\boldsymbol{R}\boldsymbol{l} = \lambda\boldsymbol{l}. \tag{2.7}$$

The eigenvalue may have a complex value, and the eigenvector $\boldsymbol{l}$ may have complex components. Consider the inner product of Eq. (2.7) and its complex conjugate $\boldsymbol{R}\bar{\boldsymbol{l}} = \bar{\lambda}\bar{\boldsymbol{l}}$ on both

Figure 1.254: 3D-Rotations-03.png

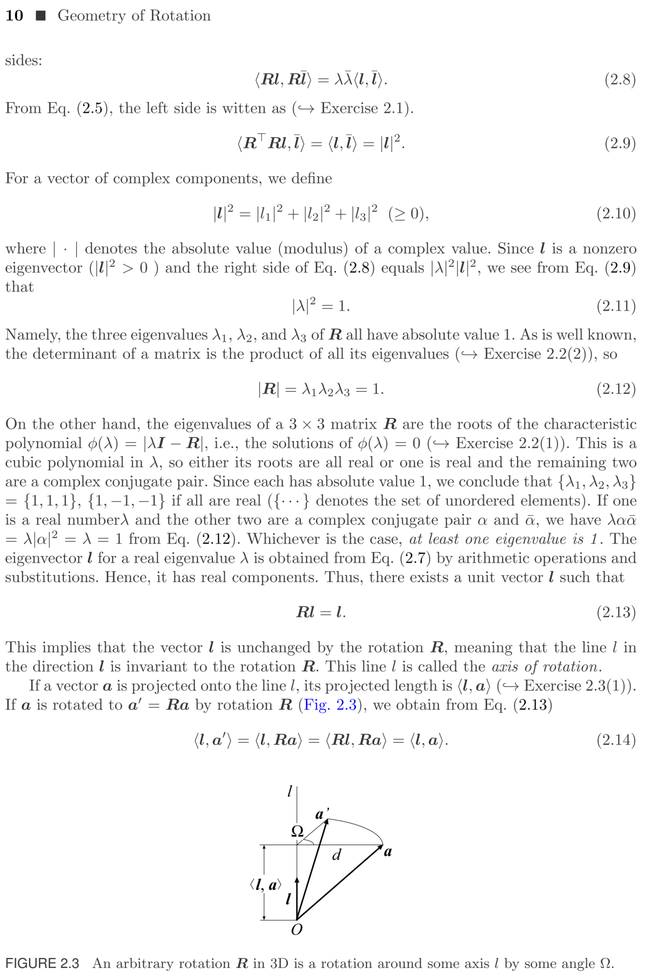

10 ■ Geometry of Rotation

sides:

$$\langle \boldsymbol{Rl}, \boldsymbol{R\bar{l}} \rangle = \lambda\bar{\lambda}\langle \boldsymbol{l}, \bar{\boldsymbol{l}} \rangle. \tag{2.8}$$

From Eq. (2.5), the left side is witten as (↪ Exercise 2.1).

$$\langle \boldsymbol{R}^\top \boldsymbol{Rl}, \bar{\boldsymbol{l}} \rangle = \langle \boldsymbol{l}, \bar{\boldsymbol{l}} \rangle = |\boldsymbol{l}|^2. \tag{2.9}$$

For a vector of complex components, we define

$$|\boldsymbol{l}|^2 = |l_1|^2 + |l_2|^2 + |l_3|^2 \ \ (\geq 0), \tag{2.10}$$

where $|\cdot|$ denotes the absolute value (modulus) of a complex value. Since $\boldsymbol{l}$ is a nonzero eigenvector ($|\boldsymbol{l}|^2 > 0$) and the right side of Eq. (2.8) equals $|\lambda|^2|\boldsymbol{l}|^2$, we see from Eq. (2.9) that

$$|\lambda|^2 = 1. \tag{2.11}$$

Namely, the three eigenvalues λ_1, λ_2, and λ_3 of $\boldsymbol{R}$ all have absolute value 1. As is well known, the determinant of a matrix is the product of all its eigenvalues (↪ Exercise 2.2(2)), so

$$|\boldsymbol{R}| = \lambda_1\lambda_2\lambda_3 = 1. \tag{2.12}$$

On the other hand, the eigenvalues of a 3×3 matrix $\boldsymbol{R}$ are the roots of the characteristic polynomial $\phi(\lambda) = |\lambda\boldsymbol{I} - \boldsymbol{R}|$, i.e., the solutions of $\phi(\lambda) = 0$ (↪ Exercise 2.2(1)). This is a cubic polynomial in λ, so either its roots are all real or one is real and the remaining two are a complex conjugate pair. Since each has absolute value 1, we conclude that $\{\lambda_1, \lambda_2, \lambda_3\} = \{1, 1, 1\}$, $\{1, -1, -1\}$ if all are real ($\{\cdots\}$ denotes the set of unordered elements). If one is a real numberλ and the other two are a complex conjugate pair α and $\bar{\alpha}$, we have $\lambda\alpha\bar{\alpha} = \lambda|\alpha|^2 = \lambda = 1$ from Eq. (2.12). Whichever is the case, *at least one eigenvalue is 1*. The eigenvector $\boldsymbol{l}$ for a real eigenvalue λ is obtained from Eq. (2.7) by arithmetic operations and substitutions. Hence, it has real components. Thus, there exists a unit vector $\boldsymbol{l}$ such that

$$\boldsymbol{Rl} = \boldsymbol{l}. \tag{2.13}$$

This implies that the vector $\boldsymbol{l}$ is unchanged by the rotation $\boldsymbol{R}$, meaning that the line l in the direction $\boldsymbol{l}$ is invariant to the rotation $\boldsymbol{R}$. This line l is called the *axis of rotation*.

If a vector $\boldsymbol{a}$ is projected onto the line l, its projected length is $\langle \boldsymbol{l}, \boldsymbol{a} \rangle$ (↪ Exercise 2.3(1)). If $\boldsymbol{a}$ is rotated to $\boldsymbol{a}' = \boldsymbol{Ra}$ by rotation $\boldsymbol{R}$ (Fig. 2.3), we obtain from Eq. (2.13)

$$\langle \boldsymbol{l}, \boldsymbol{a}' \rangle = \langle \boldsymbol{l}, \boldsymbol{Ra} \rangle = \langle \boldsymbol{Rl}, \boldsymbol{Ra} \rangle = \langle \boldsymbol{l}, \boldsymbol{a} \rangle. \tag{2.14}$$

FIGURE 2.3 An arbitrary rotation $\boldsymbol{R}$ in 3D is a rotation around some axis l by some angle Ω.

Figure 1.255: 3D-Rotations-04.png

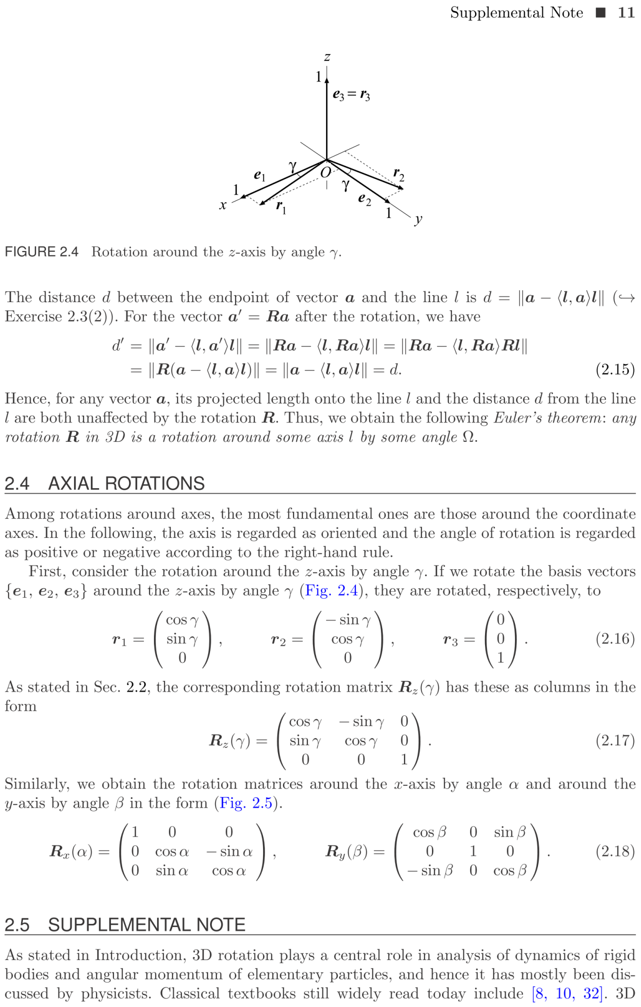
z
1
$e_3 = r_3$
e_1
γ
O
r_2
1
x
r_1
γ
e_2
1
y

Supplemental Note ■ **11**

FIGURE 2.4 Rotation around the z-axis by angle γ.

The distance d between the endpoint of vector $\boldsymbol{a}$ and the line l is $d = \|\boldsymbol{a} - \langle \boldsymbol{l}, \boldsymbol{a}\rangle \boldsymbol{l}\|$ (↪ Exercise 2.3(2)). For the vector $\boldsymbol{a}' = \boldsymbol{Ra}$ after the rotation, we have

$$\begin{aligned} d' &= \|\boldsymbol{a}' - \langle \boldsymbol{l}, \boldsymbol{a}'\rangle \boldsymbol{l}\| = \|\boldsymbol{Ra} - \langle \boldsymbol{l}, \boldsymbol{Ra}\rangle \boldsymbol{l}\| = \|\boldsymbol{Ra} - \langle \boldsymbol{l}, \boldsymbol{Ra}\rangle \boldsymbol{Rl}\| \\ &= \|\boldsymbol{R}(\boldsymbol{a} - \langle \boldsymbol{l}, \boldsymbol{a}\rangle \boldsymbol{l})\| = \|\boldsymbol{a} - \langle \boldsymbol{l}, \boldsymbol{a}\rangle \boldsymbol{l}\| = d. \end{aligned} \tag{2.15}$$

Hence, for any vector $\boldsymbol{a}$, its projected length onto the line l and the distance d from the line l are both unaffected by the rotation $\boldsymbol{R}$. Thus, we obtain the following *Euler's theorem*: *any rotation* $\boldsymbol{R}$ *in 3D is a rotation around some axis* l *by some angle* Ω.

2.4 AXIAL ROTATIONS

Among rotations around axes, the most fundamental ones are those around the coordinate axes. In the following, the axis is regarded as oriented and the angle of rotation is regarded as positive or negative according to the right-hand rule.

First, consider the rotation around the z-axis by angle γ. If we rotate the basis vectors $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3\}$ around the z-axis by angle γ (Fig. 2.4), they are rotated, respectively, to

$$\boldsymbol{r}_1 = \begin{pmatrix} \cos\gamma \\ \sin\gamma \\ 0 \end{pmatrix}, \qquad \boldsymbol{r}_2 = \begin{pmatrix} -\sin\gamma \\ \cos\gamma \\ 0 \end{pmatrix}, \qquad \boldsymbol{r}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \tag{2.16}$$

As stated in Sec. **2.2**, the corresponding rotation matrix $\boldsymbol{R}_z(\gamma)$ has these as columns in the form

$$\boldsymbol{R}_z(\gamma) = \begin{pmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{2.17}$$

Similarly, we obtain the rotation matrices around the x-axis by angle α and around the y-axis by angle β in the form (Fig. 2.5).

$$\boldsymbol{R}_x(\alpha) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{pmatrix}, \qquad \boldsymbol{R}_y(\beta) = \begin{pmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{pmatrix}. \tag{2.18}$$

2.5 SUPPLEMENTAL NOTE

As stated in Introduction, 3D rotation plays a central role in analysis of dynamics of rigid bodies and angular momentum of elementary particles, and hence it has mostly been discussed by physicists. Classical textbooks still widely read today include [8, 10, 32]. 3D

Figure 1.256: 3D-Rotations-05.png

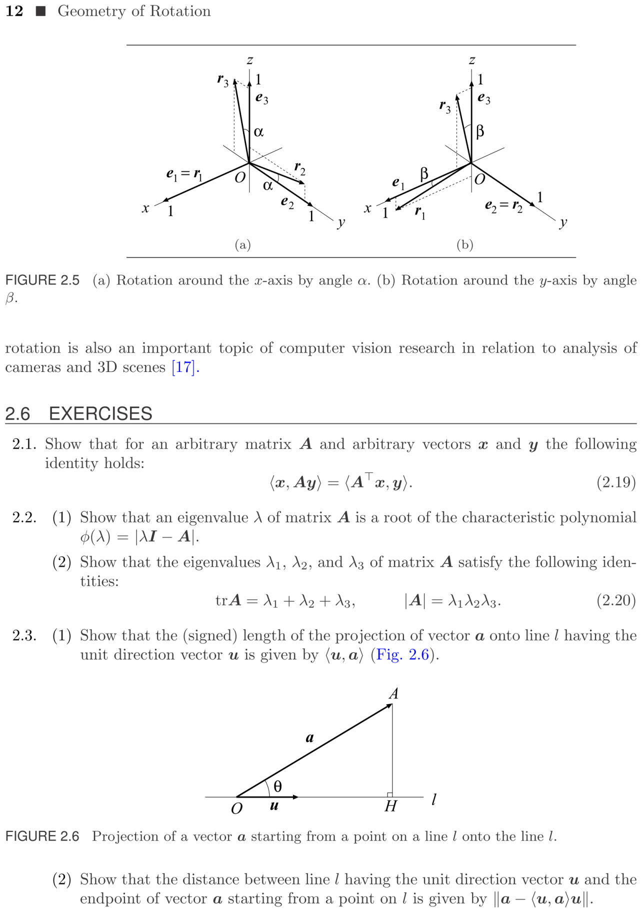

12 ■ Geometry of Rotation

z
r_3
1
e_3
α
$e_1 = r_1$
O
r_2
α
x
1
e_2
1
y
(a)
z
1
e_3
r_3
β
e_1
β
O
x
1
r_1
$e_2 = r_2$
1
y
(b)

FIGURE 2.5 (a) Rotation around the x-axis by angle α. (b) Rotation around the y-axis by angle β.

rotation is also an important topic of computer vision research in relation to analysis of cameras and 3D scenes [17].

2.6 EXERCISES

2.1. Show that for an arbitrary matrix $\boldsymbol{A}$ and arbitrary vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ the following identity holds:

$$\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{y} \rangle = \langle \boldsymbol{A}^\top \boldsymbol{x}, \boldsymbol{y} \rangle. \tag{2.19}$$

2.2. (1) Show that an eigenvalue λ of matrix $\boldsymbol{A}$ is a root of the characteristic polynomial $\phi(\lambda) = |\lambda \boldsymbol{I} - \boldsymbol{A}|$.

(2) Show that the eigenvalues λ_1, λ_2, and λ_3 of matrix $\boldsymbol{A}$ satisfy the following identities:

$$\mathrm{tr}\boldsymbol{A} = \lambda_1 + \lambda_2 + \lambda_3, \qquad |\boldsymbol{A}| = \lambda_1 \lambda_2 \lambda_3. \tag{2.20}$$

2.3. (1) Show that the (signed) length of the projection of vector $\boldsymbol{a}$ onto line l having the unit direction vector $\boldsymbol{u}$ is given by $\langle \boldsymbol{u}, \boldsymbol{a} \rangle$ (Fig. 2.6).

FIGURE 2.6 Projection of a vector $\boldsymbol{a}$ starting from a point on a line l onto the line l.

(2) Show that the distance between line l having the unit direction vector $\boldsymbol{u}$ and the endpoint of vector $\boldsymbol{a}$ starting from a point on l is given by $\|\boldsymbol{a} - \langle \boldsymbol{u}, \boldsymbol{a} \rangle \boldsymbol{u}\|$.

Figure 1.257: 3D-Rotations-06.png

2) C:/DATA-NovoVivo-126-2020/Books-126-2020/Calculus with Curvilinear Coordinates - Problems and Solutions 2018/Calculus with Curvilinear Coordinates - Problems and Solutions 2018.pdf IMPORTANT Pag. 01 (12) -Chapter 1 Arc Length and Tangent Vector Pag. 02 (13) - Fig. 1.1

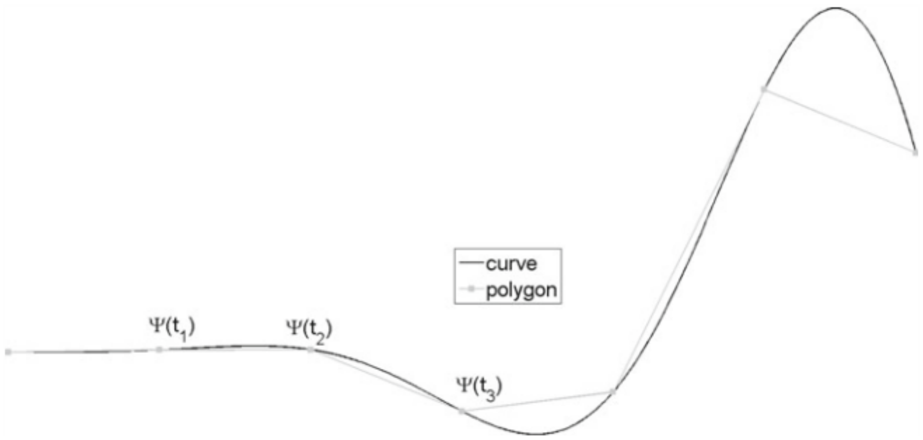

Figure 1.258: cc01.png

3) C:/DATA-NovoVivo-126-2020/Books-126-2020/Divided Spheres Geodesics and the Orderly Subdivision of the Sphere by Edward S. Popko/Divided Spheres Geodesics and the Orderly Subdivision of the Sphere by Edward S. Popko.pdf Obs.: Livro todo !!!! IMPORTANT

4) C:/DATA-NovoVivo-126-2020/Books-126-2020/Integral Calculus 3D Geometry and Vector Booster/Integral Calculus 3D Geometry and Vector Booster.pdf

5) C:/DATA-NovoVivo-126-2020/Books-126-2020/Math-Books-important/analytic-number-theory-a-tribute-to-gauss-and-diri.pdf

6) C:/DATA-NovoVivo-126-2020/Books-126-2020/Math-Books-important/differential-geometry-of-varieties-with-degenerate.pdf

Chapter 1 Foundational Material 1
1.1 Vector Space 1
1.1.1 The General Linear Group 1
1.1.2 Vectors and Tensors 3
1.2 Differentiable Manifolds 5
1.2.1 The Tangent Space, the Frame Bundle, and Tensor Fields - 5
1.2.2 Mappings of Differentiable Manifolds 7
1.2.3 Exterior Algebra, Pfaffian Forms, and the Cartan Lemma 9
1.2.4 The Structure Equations of the General Linear Group 12
1.2.5 The Frobenius Theorem 12
1.2.6 The Cartan Test 13
1.2.7 The Structure Equations of a Differentiable Manifold 15
1.2.8 Affine Connections on a Differentiable Manifold 18

1.3 Projective Space 19
1.3.1 Projective Transformations, Projective Frames, and the Structure Equations of a Projective Space . . 19
1.3.2 The Duality Principle 22
1.3.3 Projectivization 24
1.3.4 Classical Homogeneous Spaces (Affine, Euclidean, Non-Euclidean) and Their Transformations . . . 25

......

Chapter 2 Varieties in Projective Spaces and Their Gauss Maps 49
2.1 Varieties in a Projective Space 49
2.1.1 Equations of a Variety 49
2.1.2 The Bundle of First-Order Frames Associated with a Variety 51
2.1.3 The Prolongation of Basic Equations 53
2.2 The Second Fundamental Tensor and the Second Fundamental Form 54
2.2.1 The Second Fundamental Tensor, the Second Fundamental Form, and the Osculating Subspace of a Variety 54
2.2.2 Further Specialization of Moving Frames and Reduced Normal Subspaces 56
2.2.3 Asymptotic Lines and Asymptotic Cone 58
2.2.4 The Osculating Subspace, the Second Fundamental Form, and the Asymptotic Cone of the Grassmannian 59
2.2.5 Varieties with One-Dimensional Normal Subspaces . . . 61
2.3 Rank and Defect of Varieties with Degenerate Gauss Maps 63
2.4 Examples of Varieties with Degenerate Gauss Maps . . 65
2.5 Application of the Duality Principle 70
2.5.1 Dual Variety 70
2.5.2 The Main Theorem 72
2.5.3 Cubic Symmetroid 76
2.5.4 Singular Points of the Cubic Symmetroid 78
2.5.5 Correlative Transformations 80
2.6 Hypersurface with a Degenerate Gauss Map Associated with a Veronese Variety 81
2.6.1 Veronese Varieties and Varieties with Degenerate Gauss Maps 81

2.6.2 Singular Points 85
Notes 86

7) C:/DATA-NovoVivo-126-2020/Books-126-2020/Math-Books-important/carl-friedrich-gauss-titan-of-science.djvu 8) C:/DATA-NovoVivo-126-2020/Books-126-2020/Math-Books-important/gauss-and-jacobi-sums.djvu 9) C:/DATA-NovoVivo-126-2020/Books-126-2020/Math-Books-important/Geometry/affine-and-projective-geometry.djvu IMPORTANT

Pag. 144 (166) - Projective Geometry 10) C:/DATA-NovoVivo-126-2020/Books-126-2020/Math-Books-important/Geometry/complex-hyperbolic-geometry.djvu

Pag. 01 (18) - The complex projective line;

Pag. 248 (265) - Extors in Projective Space.

11) C:/DATA-NovoVivo-126-2020/Books-126-2020/
Math-Books-important/Geometry/geometry-05.djvu
Pag. 06 - I. Complete Embedded Minimal Surfaces of Finite Total Curvature

12) C:/DATA-NovoVivo-126-2020/Books-126-2020/
Math-Books-important/Geometry/geometry-06-riemannian-geometry.djvu

Chapter 5. Symmetric Spaces 55
x1. Globally Symmetric Spaces 55
x2. Germs of Smooth Mappings 55
x3. Extensions of Affine Mappings 56
x4. Uniqueness Theorem 58
x5. Reduction of Locally Symmetric Spaces to Globally Symmetric Spaces 59
x6. Properties of Symmetries in Globally Symmetric Spaces 60
x7. Symmetric Spaces 61
x8. Examples of Symmetric Spaces 62
x9. Coincidence of Classes of Symmetric and Globally Symmetric Spaces 63
...
Chapter 12. Metric Properties of Geodesies 141
x1. Length of a Curve in a Riemannian Space 141
x2. Natural Parameter 142
x3. Riemannian Distance and Shortest Arcs 142
x4. Extremals of the Length Lagrangian 143
x5. Riemannian Coordinates 144
x6. Gauss Lemma 145

x7. Geodesies are Locally Shortest Arcs 148
x8. Smoothness of Shortest Arcs 149
x9. Local Existence of Shortest Arcs 150
x10. Intrinsic Metric 151
x11.Hopf-Rinow Theorem 153
...
Chapter 15. Curvature in Riemannian Space 193
x1. Riemannian Curvature Tensor 193
x2. Symmetries of the Riemannian Tensor 193
x3. Riemannian Tensor as a Functional 198
x4. Walker Identity and Its Consequences 199
x5. Recurrent Spaces 200
x6. Virtual Curvature Tensors 201
x7. Reconstruction of the Bianci Tensor from
Its Values on Bivectors . .. 202
x8. Sectional Curvatures 204
x9. Formula for the Sectional Curvature 205
Chapter 16. Gaussian Curvature 207
x1. Bianchi Tensors as Operators 207
x2. Splitting of Trace-Free Tensors 208
x3. Gaussian Curvature and the Scalar Curvature 209
x4. Curvature Tensor for n = 2 210
x5. Geometric Interpretation of the Sectional Curvature 210
x6. Total Curvature of a Domain on a Surface 212
x7. Rotation of a Vector Field on a Curve 214
x8. Rotation of the Field of Tangent Vectors 215
x9. Gauss-Bonnet Formula 218
x10. Triangulated Surfaces 220
x11. Gauss-Bonnet Theorem 221
...
Chapter 22. Spaces of Constant Curvature 288
x1. Spaces of Constant Curvature 288
x2. Model Spaces of Constant Curvature 290
x3. Model Spaces as Hypersurfaces 292
x4. Isometries of Model Spaces 294
x5. Fixed Points of Isometries 296
x6. Riemann Theorem 296
..;
Chapter 36. Curvature Tensor 475
x1. Parallel Translation Along a Curve 475
x2. Computation of the Parallel Translation Along a Loop 477

x3. Curvature Operator at a Given Point 482
x4. Translation of a Vector Along an Infinitely Small Parallelogram 484
x5. Curvature Tensor 485
x6. Formula for Transforming Coordinates of the Curvature Tensor 486
x7. Expressing the Curvature Operator via Covariant Derivatives 487
x8. Cartan Structural Equation 490
x9. Bianchi Identity 491

13) C:/DATA-NovoVivo-126-2020/Books-126-2020/ Math-Books-important/ Geometry/geometry-audin.pdf
14) C:/DATA-NovoVivo-126-2020/Books-126-2020/ Math-Books-important/ Geometry/geometry-I-berger.pdf Obs.: Chapter 4, 5 e 6.
15) C:/DATA-NovoVivo-126-2020/Books-126-2020/ Math-Books-important/ Geometry/geometry-II-berger.pdf Pag. 170 - Chapter 16 - Projective conics.

16) C:/DATA-NovoVivo-126-2020/Books-126-2020/ Math-Books-important/ Geometry/ Geometry-IV-nonregular-riemannian-geometry.djvu

17) C:/DATA-NovoVivo-126-2020/Books-126-2020/ Math-Books-important/ Geometry/ hyperbolic-geometry.pdf IMPORTANT

18) C:/DATA-NovoVivo-126-2020/Books-126-2020/ Math-Books-important/ Geometry/ kiselevs-geometry-book-1-planimetry.djvu

19) C:/DATA-NovoVivo-126-2020/Books-126-2020/ Math-Books-important/ Geometry/ kiselevs-geometry-book-2-stereometry.djvu

20) C:/DATA-NovoVivo-126-2020/Books-126-2020/ Math-Books-important/ Geometry/ lectures-in-projective-geometry.djvu

21) C:/DATA-NovoVivo-126-2020/ Books-126-2020/ Math-Books-important/ Geometry/ projective-geometry.pdf

22) C:/DATA-NovoVivo-126-2020/Books-126-2020/ Math-Books-important/ Geometry/ projective-geometry-and-projective-metrics.djvu

23) C:/DATA-NovoVivo-126-2020/Books-126-2020/ Math-Books-important/ Geometry/ projective-geometry-an-introduction-casse.pdf

24) C:/DATA-NovoVivo-126-2020/ Books-126-2020/ Math-Books-important/ Geometry/ projective-geometry-volume-1-veblen.djvu

25) C:/DATA-NovoVivo-126-2020/ Books-126-2020/ Math-Books-important/ Geometry/ projective-geometry-volume-2-veblen.djvu

26) C:/DATA-NovoVivo-126-2020/ Books-126-2020/ Math-Books-important/ Geometry/ riemannian-geometry-carmo.pdf

27) C:/DATA-NovoVivo-126-2020/ Books-126-2020/ Math-Books-important/ Geometry/ riemannian-geometry-klingenberger.djvu

28) C:/DATA-NovoVivo-126-2020/ Books-126-2020/ Math-Books-important/Geometry/ riemannian-geometry-lafontaine.djvu

29) C:/DATA-NovoVivo-126-2020/ Books-126-2020/ Math-Books-important/Geometry/ riemannian-geometry-willmore.pdf

30) C:/DATA-NovoVivo-126-2020/ Books-126-2020/ Particles, Fields, Space-Time/ Particles, Fields, Space-Time.pdf

31) C:/DATA-NovoVivo-126-2020/ Books-126-2020/ Stereotomy/ 3030432173.pdf

32) C:/DATA-NovoVivo-126-2020/ Books-126-2020/Topology and Geometry/0387979263 - 3540979263 - 1441931031.pdf Obs.: Metric spaces.

33) C:/DATA-NovoVivo-126-2020/ Books-126-2020/ Analytical Techniques of Celestial Mechanics - 1995/ Analytical Techniques of Celestial Mechanics - 1995.pdf
The End

Index

Books-126-2020 - Index, 141

Cientistas observam ‘partícula de Deus’ em nova interação, 50
Comandos Latex, 13

Dark energy: map gives clue about what it is { but deepens dispute about the cosmic expansion rate, 38

Energia escura, 34

Gauss, 32
Geometry
  Differential Geometry
    Geodesic curve on a surface, 9
Geometry - Projective Geometry, 53
Geometry - Start, 1
Geometry - Trigonometry, 14

Orthogonal Lines, 23

Projective Geometry, 15
Pythagorean-hodograph curves - algebra and geometry inseparable, 24

Use Newtheorem, 12

Bibliography

entries

Mathematics - Annotations 004

Prof. Stenio Musich

August 09, 2020

Chapter 1

Mathematics

1.1 Projective Geometry

Date: 20200809

1) C:/ARQ-IMPORTANT-08/Important-02/Divided Spheres Geodesics and the Orderly Subdivision of the Sphere by Edward S. Popko.pdf Source: C:/DATA-NovoVivo-126-2020/Books-126-2020/Divided Spheres Geodesics and the Orderly Subdivision of the Sphere by Edward S. Popko/Divided Spheres Geodesics and the Orderly Subdivision of the Sphere by Edward S. Popko.pdf

Pag. 02 (23) - On the geographic projection of the surface of a sphere - Euler (nota de rodape). Source: C:/DATA-NovoVivo-127-2020/Books-127-2020/EULER/E491en.pdf

Pag. 18(39) - Fuller called this configuration a Dymaxion projection[1]. The projection is based on a subdivision grid that runs either parallel or perpendicular to the edges of the square and triangular faces. After pointing out the disadvantages of current projections, including the popular Mercator projection, Fuller goes on to say: \Another expedient has been to resolve the earth's surface into a polyhedron, projecting gnomonically to the facets of the polyhedron, the idea being that the sections of the polyhedron can be assembled on a flat surface to give a truer picture of the earth's surface and of directions and distance."[2] In essence, Fuller claims that a gnomonically projected[3] map placed onto the faces of a polyhedron, the cuboctahedron in this case, provides a truer representation of areas, boundaries, directions, and distances than any plane surface map heretofore known. By using the cuboctahedron as the base polyhedron for his projection, all vertices lie on one of four great circles. Fuller makes three invention claims in his patent. The Dymaxion map is a projected map of square and triangular sections where edges are represented by projected great circles with a uniform cartographic scale;

---

[1] In the 1930s, Fuller was advised to find a better name for his 4D house invention, and with the help of an advertising wordsmith, Waldo Warren, the word \Dymaxion" was coined by selecting suitable syllables from Fuller's account of his design ideas. The roots of this name are \dynamism", \maximum", and \ions". Buckminster Fuller, The Dymaxion World of Buckminster Fuller, ed. Robert Marks, New York: Doubleday, 1973

[2] (Fuller 1946, p.1, col. 1, para. 2)

[3] Gnomonic or gnomic projection is a map projection obtained by projecting points on

12 In the 1930s, Fuller was advised to find a better name for his 4D house invention, and with the help of an advertising wordsmith, Waldo Warren, the word \Dymaxion" was coined by selecting suitable syllables from Fuller's account of his design ideas. The roots of this name are \dynamism", \maximum", and \ions". Buckminster Fuller, The Dymaxion World of Buckminster Fuller, ed. Robert Marks. New York: Doubleday, 1973.IMPORTANT 13 (Fuller 1946, p.1, col. 1, para. 2) 14 Gnomonic or gnomic projection is a map projection obtained by projecting points on the surface of a sphere from a sphere's center to a plane that is tangent to the sphere or a projection from the center of a sphere to the surface that surrounds an object (such as a sphere around a polyhedra). In a gnomonic projection, great circles are mapped to straight lines.

Pag. 58(79) - Cells in a spherical tessellation have wall-point relationships, and it's possible to con- vert a triangular tessellation with wall-point neighbors into a hexagonal one with only wall neighbors simply by defining eachpoint's Voronoi cell.[4] Voronoi cells are polygons that define an area (of points) that is closer to a designated point than to any other. If the original points are evenly spaced, or nearly so, the Voronoi cells will resemble a grid of hexagons. Randall and Heikes illustrate this in Figure 3.8(a). The Voronoi cells in green define the points (areas) closest to each of the points in the following triangular grid. The spherical versions of both grids are shown in (b).[5] When a Voronoi tessellation is made of a fully subdivided icosahedron, such as the one shown in Figure 3.9, the result is mostly hexagonal cells, but pentagonal cells also result.

---

the surface of a sphere from a sphere's center to a plane that is tangent to the sphere or a projection from the center of a sphere to the surface that surrounds an object (such as a sphere around a polyhedra). In a gnomonic projection, great circles are mapped to straight lines.

[4] Voronoi diagrams are named after Russian mathematician, Georgy Feodosevich Voronoy (1868{1908). However, they were recognized much earlier by Rene Descartes (1596{1650) and used by Lejeune Dirichlet (1805{1859). Voronoi tessellations are also called Dirichlet tessellations.

[5] Climate model color illustrations in this section courtesy of Ross Heikes.

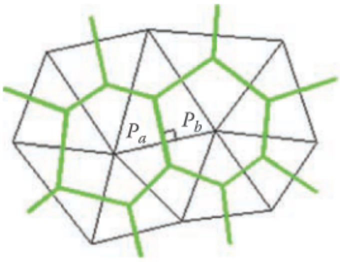

Figure 1.1: ds02a.png - Hexagonal Voronoi cells from triangular tessellations.

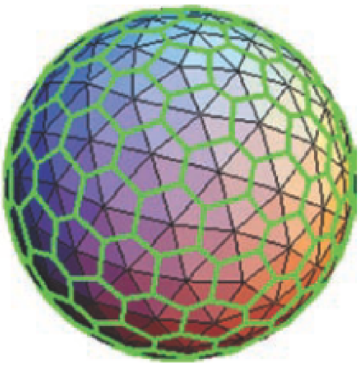

Figure 1.2: ds02b.png - Hexagonal Voronoi cells from triangular tessellations.

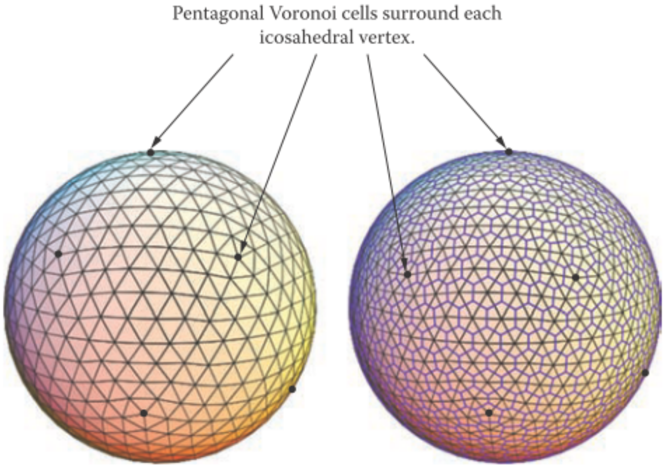

Figure 1.3: ds03.png - Hexagonal Voronoi cells from triangular tessellations.

One pentagonal cell surrounds each of the icosahedron's 12 vertices. Some are pointed out in the figure. Each hexagonal and pentagonal cell is quasi-isotropic because their areas are nearly equal. Both spheres in Figure 3.9 are oriented and scaled the same for comparison. When it comes to climate modeling, hexagonal geodesic grids offer many benefits over other grid systems. They are largely isotropic and do not suffer from oversampling near the poles like longitude/latitude grids. The poles are no different than any other area in the hexagonal grid. All cells in the grid are equal wall-neighbors. There are no point neighbors, as in other grid systems. Hexagonal geodesic grids can also accommodate different resolutions where the atmosphere might be represented by a coarse hexagonal grid, and land-surface areas represented by a finer grid as seen at the right. Randall says coupling the atmosphere and ocean models with multiple hexagonal grids simplifies modeling because the atmospheric state is \interpolated down" from large hexagonal cells to smaller ones below and fluxes are \averaged up" from small cells to larger ones. Hexagonal geodesic grids can also be warped or morphed to meet local needs. Although hexagonal geodesic grids are more complicated to implement than rectangular longitude/latitude grids, their benefits far outweigh the additional work needed to set them up and use them in climate models. Figure 3.10 shows typical climate modeling output based on a geodesic grid.

Pag.91 (112) - 4.9 Loxodromes Latitude and longitude are so familiar we can be fooled when navigating. Consider the use of a compass. A compass performs a very specific function; it always points in the same direction-north.[6] Think of it this way: when you change directions, the compass still points north. In effect, you are turning under the compass; the compass continues pointing in the same direction. The value of a compass is that it allows you to maintain a course over a long distance by ensuring that you cross every meridian (longitude) at the same angle. This course is called a rhumb line or a loxodrome, from the Greek words loxos (oblique) and dromos (running, from dramein: to run). A loxodrome is a line on a sphere of constant bearing that cuts across all meridians at any constant angle except a right one. A loxodrome is also called a spherical spiral and it

---

[6] A compass is influenced by many factors. The magnetic north pole is not exactly located at the geographic north pole. Electrical currents (nearby wires), and large ferrous metal sources near the compass (engines, pipes, etc.) influence which way the compass needle points.

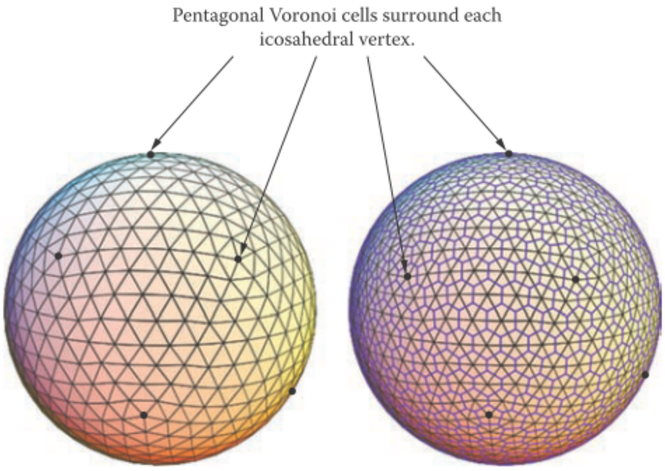

Figure 1.4: ds04.png - Loxodrome courses 56° and 84.2° with a great circle route.

displays as a logarithmic spiral on a polar projection. If the course is not due north-south or east-west, a loxodromic spiral will eventually spiral to either the north or south pole and then endlessly loop that pole in smaller and smaller circles. A loxodrome displays as a straight line on a Mercator map. When plotted on any other map projection, it is a curved line. A loxodrome is a great circle, only if it is along the equator or a meridian. A rhumb line course between two points that is not along the equator or a meridian is always longer than the great circle route between the same two points. And there are an infinite number of rhumb lines between any two nonpolar points that are not on the same latitude or longitude. They only differ in the number of times they encircle the earth. Figure 4.10 shows rhumb lines or loxodromes between two sample points, 1 and 2, and how they compare to a direct great circle route, shown with a bold line. In Figure 4.10(a), the loxodrome takes the most direct route from point 1 to point 2 by following compass course 56°. In (b), the loxodrome follows a different course of 84.2°. It still leads to point 2, but has to circle the earth one time to do so. In both illustrations, the great circle route between the two points is the shortest route possible. Notice, too, if you miss your destination on any rhumb course other than along the equator, you will eventually end up at the north or south pole. Can we ever use a compass to follow a great circle route, if using a compass to navigate any direction other than due north-south or east-west puts us on a rhumb line or loxodrome? The answer is a qualified yes. You can approximate the great circle route by taking a series of short rhumb line courses.

You would travel a particular compass course for a while before changing to another course, and then another, until you arrived at your destination. To do this, the navigator divides the overall track between his starting and ending points into a series of small segments. Between each segment is a waypoint, a location where he or she will change course slightly. The length of a course segment is about the distance the navigator expects to travel in a day, if by ship, or in an hour or so, if by plane, depending on the plane's speed. On a ship, the course change might take place at the same time every day. For instance, navigators might change at local noon, the time when navigators traditionally used their sextants to determine their latitude by the sun's noon-time meridian passage. Today, airplanes and ships are equipped with autopilots, and course corrections are made automatically from time to time in order to stay close to a theoretic great circle course. The more rhumb lines or legs the course is divided into, the closer individual rhumb lines will approximate the great circle route. When a great circle and rhumb line course are shown together on a Mercator map, the great circle route appears as a curved line, while the rhumb line is straight. On a sphere, they appear just the opposite, the great circle being the shortest "straight line" course, while the rhumb line is the longer curved course between points unless the points are directly north or south of each other or along the equator. To give an example of just how dramatic the difference between navigating a rhumb line and great circle course can be, consider a trip between London, England, and Anchorage, Alaska. The latitude-longitude of London's Heathrow International Airport is 51° 28' 15.72"N, 0° 27' 30.98"W. Anchorage International Airport is located at 61° 10' 14.69"N, 149° 59' 33.66"W. The great circle distance between them, the most direct route, passes over Greenland and near the north pole, and is 4,475 miles (7,201 km). The rhumb line bearing to Anchorage from London is about 276°(almost due west by compass) and the distance is 5,731 miles (9,223 km). Clearly, the rhumb line distance is longer, almost 28 percent longer. These differences cannot be ignored by airlines or shipping companies. The time and fuel-cost implications are considerable. Obs.: Ortodromia é a linha que une dois pontos à superfície da Terra, à qual corresponde o caminho mais curto entre eles. Este é o sentido normalmente atribuído no âmbito da navegação marítima ou aérea. Formalmente, uma ortodromia é a deformação do círculo máximo quando plotado sobre uma representação planisférica da Terra. Já numa superfície esférica, a ortodromia cobre a seção de um círculo máximo que divide um elipsoide de revolução em dois hemisférios. Cartograficamente é uma linha torsa, isto é, uma linha que não pode ser assente sobre um plano.

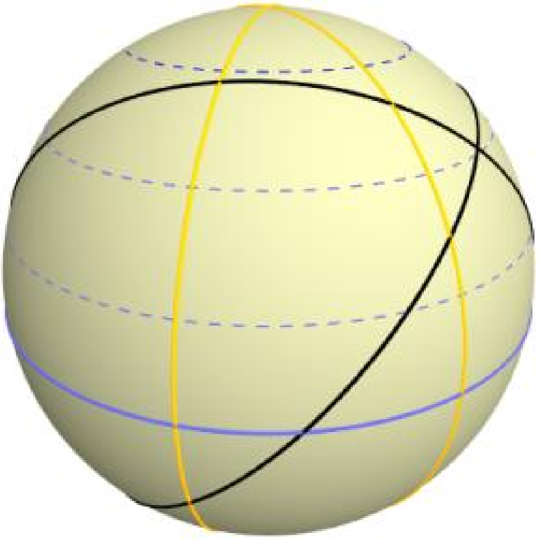

Figure 1.5: ds05.jpg - Ortodrómicas traçadas sobre uma esfera.

A ortodromia e a navegação Entre dois pontos quaisquer da superfície terrestre podem traçar-se dois tipos de rotas distintas: a ortodromia, que minimiza a distância entre eles; e a loxodromia, que mantém constante o rumo, isto é, o ângulo entre o caminho seguido e os meridianos. Caso os pontos estejam separados por um arco de 180° (isto é, sejam antípodas), existe uma infinidade de rotas ortodrómicas que os ligam, pelo que, esquecendo outras considerações, em qualquer direção que se partir, se a rota for mantida, chega-se igualmente ao outro ponto. A ortodromia pura tem o inconveniente de necessitar de correções constantes de rumo, já que o ângulo com cada meridiano é sempre diferente, excepto quando a viagem se faça sobre o Equador ou ao longo de um meridiano.

Em navegação marítima, a ortodromia só é utilizada em circunstâncias especiais e em trajetos muito longos, quando o fator tempo seja decisivo, o que não acontece na navegação aérea. Em qualquer dos casos, a sua aplicação prática é concretizada através de um conjunto de segmentos de loxodromia, isto é, de segmentos a rumo constante entre pontos intermédios escolhidos ao longo da rota. Outro inconveniente deste tipo de trajeto reside no facto de as ortodromias entre pontos distantes atingirem muitas vezes latitudes elevadas, onde a ocorrência de gelo e de condições atmosféricas adversas é frequente. Nestes casos, é sempre possível optar-se por uma rota mista, constituída por dois segmentos de ortodromia, que ligam os pontos de partida e de chegada a um certo paralelo limite, e um segmento de loxodromia intermédio, ao rumo este ou oeste, ao longo desse paralelo.

Rhumb line
https://en.wikipedia.org/wiki/Rhumb-line

Figure 1.6: ds06.png - Ortodromicas tracadas sobre uma esfera.

The effect of following a rhumb line course on the surface of a globe was first discussed by the Portuguese mathematician Pedro Nunes in 1537, in his Treatise in Defense of the Marine Chart, with further mathematical development by Thomas Harriot in the 1590s.

A rhumb line can be contrasted with a great circle, which is the path of shortest distance between two points on the surface of a sphere. On a great circle, the bearing to the destination point does not remain constant. If one were to drive a car along a great circle one would hold the steering wheel fixed, but to follow a rhumb line one would have to turn the wheel, turning it more sharply as the poles are approached. In other words, a great circle is locally "straight" with zero geodesic curvature, whereas a rhumb line has non-zero geodesic curvature.

Meridians of longitude and parallels of latitude provide special cases of the rhumb line, where their angles of intersection are respectively 0o and 90o. On a north-south passage the rhumb line course coincides with a great circle, as it does on an east-west passage along the equator.

On a Mercator projection map, any rhumb line is a straight line; a rhumb line can be drawn on such a map between any two points on Earth without going off the edge of the map. But theoretically a loxodrome can extend beyond the right edge of the map, where it then continues at the left edge with the same slope (assuming that the map covers exactly 360 degrees of longitude).

Rhumb lines which cut meridians at oblique angles are loxodromic curves which spiral towards the poles.[1] On a Mercator projection the north and south poles occur at infinity and are therefore never shown. However the full loxodrome on an infinitely high map would consist of infinitely many line segments between the two edges. On a stereographic projection map, a loxodrome is an equiangular spiral whose center is the north or south pole.

All loxodromes spiral from one pole to the other. Near the poles, they are close to being logarithmic spirals (which they are exactly on a stereographic projection, see below), so they wind around each pole an infinite number of times but reach the pole in a finite distance. The pole-to-pole length of a loxodrome (assuming a perfect sphere) is the length of the meridian divided by the cosine of the bearing away from true north. Loxodromes are not defined at the poles.

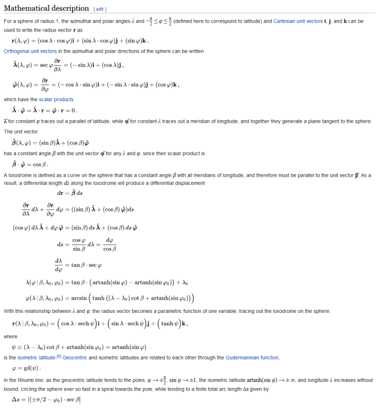

Mathematical description [edit]

For a sphere of radius 1, the azimuthal and polar angles λ and $-\frac{\pi}{2} \le \varphi \le \frac{\pi}{2}$ (defined here to correspond to latitude) and Cartesian unit vectors $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ can be used to write the radius vector $\mathbf{r}$ as

$$\mathbf{r}(\lambda,\varphi) = (\cos\lambda\cdot\cos\varphi)\mathbf{i} + (\sin\lambda\cdot\cos\varphi)\mathbf{j} + (\sin\varphi)\mathbf{k}\,.$$

Orthogonal unit vectors in the azimuthal and polar directions of the sphere can be written

$$\hat{\boldsymbol{\lambda}}(\lambda,\varphi) = \sec\varphi\frac{\partial\mathbf{r}}{\partial\lambda} = (-\sin\lambda)\mathbf{i} + (\cos\lambda)\mathbf{j}\,,$$

$$\hat{\boldsymbol{\varphi}}(\lambda,\varphi) = \frac{\partial\mathbf{r}}{\partial\varphi} = (-\cos\lambda\cdot\sin\varphi)\mathbf{i} + (-\sin\lambda\cdot\sin\varphi)\mathbf{j} + (\cos\varphi)\mathbf{k}\,,$$

which have the scalar products

$$\hat{\boldsymbol{\lambda}}\cdot\hat{\boldsymbol{\varphi}} = \hat{\boldsymbol{\lambda}}\cdot\mathbf{r} = \hat{\boldsymbol{\varphi}}\cdot\mathbf{r} = 0\,.$$

$\hat{\lambda}$ for constant φ traces out a parallel of latitude, while $\hat{\boldsymbol{\varphi}}$ for constant λ traces out a meridian of longitude, and together they generate a plane tangent to the sphere.

The unit vector

$$\hat{\boldsymbol{\beta}}(\lambda,\varphi) = (\sin\beta)\hat{\boldsymbol{\lambda}} + (\cos\beta)\hat{\boldsymbol{\varphi}}$$

has a constant angle β with the unit vector $\hat{\boldsymbol{\varphi}}$ for any λ and φ, since their scalar product is

$$\hat{\boldsymbol{\beta}}\cdot\hat{\boldsymbol{\varphi}} = \cos\beta\,.$$

A loxodrome is defined as a curve on the sphere that has a constant angle β with all meridians of longitude, and therefore must be parallel to the unit vector $\hat{\boldsymbol{\beta}}$. As a result, a differential length ds along the loxodrome will produce a differential displacement

$$d\mathbf{r} = \hat{\boldsymbol{\beta}}\,ds$$

$$\frac{\partial\mathbf{r}}{\partial\lambda}\,d\lambda + \frac{\partial\mathbf{r}}{\partial\varphi}\,d\varphi = ((\sin\beta)\,\hat{\boldsymbol{\lambda}} + (\cos\beta)\,\hat{\boldsymbol{\varphi}})ds$$

$$(\cos\varphi)\,d\lambda\,\hat{\boldsymbol{\lambda}} + d\varphi\,\hat{\boldsymbol{\varphi}} = (\sin\beta)\,ds\,\hat{\boldsymbol{\lambda}} + (\cos\beta)\,ds\,\hat{\boldsymbol{\varphi}}$$

$$ds = \frac{\cos\varphi}{\sin\beta}\,d\lambda = \frac{d\varphi}{\cos\beta}$$

$$\frac{d\lambda}{d\varphi} = \tan\beta\cdot\sec\varphi$$

$$\lambda(\varphi\,|\,\beta,\lambda_0,\varphi_0) = \tan\beta\cdot\left(\operatorname{artanh}(\sin\varphi) - \operatorname{artanh}(\sin\varphi_0)\right) + \lambda_0$$

$$\varphi(\lambda\,|\,\beta,\lambda_0,\varphi_0) = \arcsin\Big(\tanh\left((\lambda-\lambda_0)\cot\beta + \operatorname{artanh}(\sin\varphi_0)\right)\Big)$$

With this relationship between λ and φ, the radius vector becomes a parametric function of one variable, tracing out the loxodrome on the sphere:

$$\mathbf{r}(\lambda\,|\,\beta,\lambda_0,\varphi_0) = \Big(\cos\lambda\cdot\operatorname{sech}\psi\Big)\mathbf{i} + \Big(\sin\lambda\cdot\operatorname{sech}\psi\Big)\mathbf{j} + \Big(\tanh\psi\Big)\mathbf{k}\,,$$

where

$$\psi \equiv (\lambda-\lambda_0)\cot\beta + \operatorname{artanh}(\sin\varphi_0) = \operatorname{artanh}(\sin\varphi)$$

is the isometric latitude.[5] Geocentric and isometric latitudes are related to each other through the Gudermannian function,

$$\varphi = \operatorname{gd}(\psi)\,.$$

In the Rhumb line, as the geocentric latitude tends to the poles, $\varphi \to \pm\frac{\pi}{2}$, $\sin\varphi \to \pm 1$, the isometric latitude $\operatorname{artanh}(\sin\varphi) \to \pm\infty$, and longitude λ increases without bound, circling the sphere ever so fast in a spiral towards the pole, while tending to a finite total arc length Δs given by

$$\Delta s = \left|(\pm\pi/2 - \varphi_0)\cdot\sec\beta\right|$$

Figure 1.7: ds07.png

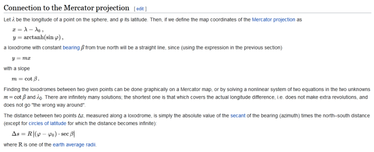

Connection to the Mercator projection [edit]

Let λ be the longitude of a point on the sphere, and φ its latitude. Then, if we define the map coordinates of the Mercator projection as

$$x = \lambda - \lambda_0\,,$$
$$y = \operatorname{arctanh}(\sin\varphi)\,,$$

a loxodrome with constant bearing β from true north will be a straight line, since (using the expression in the previous section)

$$y = mx$$

with a slope

$$m = \cot\beta\,.$$

Finding the loxodromes between two given points can be done graphically on a Mercator map, or by solving a nonlinear system of two equations in the two unknowns $m = \cot\beta$ and λ_0. There are infinitely many solutions; the shortest one is that which covers the actual longitude difference, i.e. does not make extra revolutions, and does not go "the wrong way around".

The distance between two points Δs, measured along a loxodrome, is simply the absolute value of the secant of the bearing (azimuth) times the north–south distance (except for circles of latitude for which the distance becomes infinite):

$$\Delta s = R\left|(\varphi - \varphi_0)\cdot\sec\beta\right|$$

where R is one of the earth average radii.

Figure 1.8: ds08.png

Its use in navigation is directly linked to the style, or projection of certain navigational maps. A rhumb line appears as a straight line on a Mercator projection map.[1]

The name is derived from Old French or Spanish respectively: "rumb" or "rumbo", a line on the chart which intersects all meridians at the same angle.[1] On a plane surface this would be the shortest distance between two points. Over the Earth's surface at low latitudes or over short distances it can be used for plotting the course of a vehicle, aircraft or ship.[1] Over longer distances and/or at higher latitudes the great circle route is significantly shorter than the rhumb line between the same two points. However the inconvenience of having to continuously change bearings while travelling a great circle route makes rhumb line navigation appealing in certain instances.

The point can be illustrated with an East-West passage over 90 degrees of longitude along the equator, for which the great circle and rhumb line distances are the same at 5,400 nautical miles (10,000 km). At 20 degrees north the great circle distance is 4,997 miles (8,042 km) while the rhumb line distance is 5,074 miles (8,166 km), about 1,5 percent further. But at 60 degrees north the great circle distance is 2,485 miles (3,999 km) while the rhumb line is 2,700 miles (4,300 km), a difference of 8,5 percent. A more extreme case is the air route between New York City and Hong Kong, for which the rhumb line path is 9,700 nautical miles (18,000 km). The great circle route over the North Pole is 7,000 nautical miles (13,000 km), or 5,5 hours less flying time at a typical cruising speed.

Some old maps in the Mercator projection have grids composed of lines of latitude and longitude but also show rhumb lines which are oriented directly towards north, at a right angle from the north, or at some angle from the north which is some simple rational fraction of a right angle. These rhumb lines would be drawn so that they would converge at certain points of the map: lines going in every direction would converge at each of these points. See compass rose. Such maps would necessarily have been in the Mercator projection therefore not all old maps would have been capable of showing rhumb line markings.

The radial lines on a compass rose are also called rhumbs. The expression "sailing on a rhumb" was used in the 16th-19th centuries to indicate a particular compass heading.[1]

Early navigators in the time before the invention of the marine chronometer used rhumb line courses on long ocean passages, because the ship's latitude could be established accurately by sightings of the Sun or stars but there was no accurate way to determine the longitude. The ship would sail north or south until the latitude of the destination was reached, and the ship would then sail east or west along the rhumb line (actually a parallel, which is a special case of the rhumb line), maintaining a constant latitude and recording regular estimates of the distance sailed until evidence of land was sighted.[6] Generalizations On the Riemann sphere Main article: Möbius transformation

The surface of the Earth can be understood mathematically as a Riemann sphere, that is, as a projection of the sphere to the complex plane. In this case, loxodromes can be understood as certain classes of Möbius transformations. Spheroid The formulation above can be easily extended to a spheroid.[7][8][9][10][11] The course of the rhumb line is found merely by using the ellipsoidal isometric latitude. Similarly distances are found by multiplying the ellipsoidal meridian arc length by the secant of the azimuth. Source: https://en.wikipedia.org/wiki/Rhumb-line

Pag. 138(159) -6.1 What Is a Polyhedron? Polyhedra are mathematically sophisticated and appear in more than three dimensions. In solid geometry, a polyhedron is a 3D solid whose faces join at their edges. The word derives from the Greek poly (many) plus the Indo-European hedron (seat). There are many distinct families of polyhedra, the result of differences in their geometry and various mathematical properties. We focus on two types, the Platonic and Archimedean, because they are the most common polyhedral frameworks used in spherical subdivision. The five Platonic solids are regular polyhedra, which means their faces are the same regular type-equilateral triangles, squares, or pentagons. The 13 Archimedean solids are closely related to the Platonic solids, but have two or more regular faces and some have hexagons, octagons, and decagons. We can define Platonic and Archimedean solids as 3D convex forms with no holes, depressions, protrusions, or penetrations. They are totally defined by their vertices, edges, and faces. All vertices lie on the surface of a surrounding sphere called the circumsphere, whose origin is the same as the center of the polyhedron. The edges of Platonic and Archimedean solids are equal-length straight edges between two neighboring vertices. On spherical versions, edges are equal-length geodesic arcs. The faces of planar polyhedra are regular polygons, and the plane of any one of them, if extended, does not pass through the interior volume of the polyhedron. An individual Platonic or Archimedean solid will have the same number and types of faces sequenced in the same way around every one of its vertices. Faces do not penetrate one another and every face joins another at an edge, leaving no gaps and causing no overlaps. For spherical polyhedra, the sum of the face angles surrounding any vertex is always 360^o. Figure 6.1 shows the planar and spherical version of a dodecahedron, a Platonic solid with 20 vertices, 30 edges, and 12 pentagonal faces. The planar face angles shown in (a) are all 72^o, and the dihedral angle, the angle between faces measured perpendicular to their shared edge, is approximately 116.56^o. The spherical version shown in (b) has face angles= 120^o.The arc angles of the spherical edges are each approximately 41.81^o Spherical Platonic and Archimedean solids can be derived from their planar versions very easily. All vertices of these solids are the same distance from the center of the polyhedron, thus they are all on the surface of a circumscribing sphere. The vertices on either end of a chord, a platonic edge, and the center of the solid define the plane of the arc, a geodesic arc, between neighboring points. In Figure 6.1(a), one such plane is shown as a small

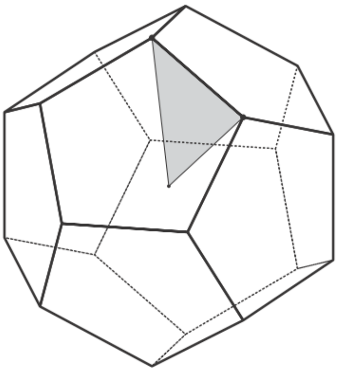

Figure 1.9: ds09a.png - Planar and spherical polyhedra.

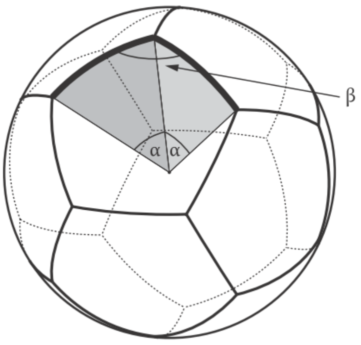

Figure 1.10: ds09b.png - Planar and spherical polyhedra.

gray triangle defined by three points: the origin and two vertices. 1 Figure 6.1(b) shows the spherical version of the dodecahedron from the same viewpoint and scale as (a), and every edge in (a) has been transformed into a geodesic arc in (b). The planes of two arcs are emphasized in (b). You can think of these geodesic arcs as central projections of a planar polyhedron's edges. The projection is on the polyhedron's circumsphere, which acts like a curved projection screen. An imaginary point light source located at the origin produces a shadow of the chord, its arc, on the circumsphere screen. In Figure 6.1(b), the two emphasized arcs demonstrate another fact about angles on a sphere. The surface anglebetween the two arcs is the dihedral angle, between the great circle planes that define them (see Section 4.16.6). The arcs themselves are the sides of the spherical face and are measured by their central angles from the center or origin of the sphere. They are each labeled . In any given Platonic or Archimedean polyhedron (planar or spherical version), all chords or arcs are the same length. With this general introduction, let us take a closer look at the properties of these two families of polyhedra.

Pag. 174(195) - 6.4.8 Waterman Polyhedra We have discussed Platonic and Archimedean solids in detail because we will use them in the next chapter to jump-start the spherical subdivision process. For completeness, however, we want to mention the Waterman polyhedra because some are nearly spherical and already highly subdivided. Unlike spherical Platonic and Archimedean solids, Waterman polyhedra do not evenly distribute points on a sphere or even distribute points on the same radius sphere. Waterman polyhedra are important additions to a spherist's tool kit because they exhibit many of the properties we have been discussing, such as symmetry, great circles, antipodal points, and so forth, and they demonstrate a rich way to develop spherical geometry that can solve certain application problems. In 1988, Steve Waterman began a series of investigations into gridded maps. We saw an example of his butterfly map projection alongside Cahill's map when we discussed Fuller's Dymaxion maps in Chapter 2. Waterman's projection studies led to the development of entirely new families of polyhedra and some resemble highly subdivided spheres. The way they are constructed is unique. Waterman polyhedra begin with a 3D lattice of points. The points are the centers of stacked spheres arranged similarly to the way oranges are stacked in a grocery store. There are several ways to pack (stack) equal spheres to maximize the number in any given volume. In cubic close packing (CCP), also called face-centered cubic (FCC) packing, an infinite number of tangent spheres are packed in layers or square grids and stacked,one layer on top of another. Each layer of spheres occupies the intergrid spaces of their adjacent layers. If all spheres have a diameter of 2 , their centers will fall only on integer coordinate values, if one of the sphere's center occupies the origin of the Cartesian space at coordinates (0, 0, 0). Figure 6.30(a) shows how CCP spheres are arranged on alternating layers. The grid spacing is one unit in the x, y, and z (perpendicular to the page) direction. The solid spheres are on even z-planes, including the z = 0 plane. The spheres shown dashed are on odd z-planes. The centers of all the spheres, the lattice points in Waterman clusters of spheres, are marked with small dots on the grid. Figure 6.30(b) shows the section of the CCP stack indicated by (a). Both the grid in (a) and the stack in (b) are examples of a theoretical grid and stacking that extends infinitely in the x,y, and z directions. Footnote - It is interesting to note that Johann Carl Friedrich Gauss(1777-1855) proved that the densest possible packing of equal spheres is $d = \frac{\pi}{3\sqrt{2}} = 0.74048$, or 74 percent and that the space between spheres (not occupied) is 0.25951, or 25.9 percent. Coincidentally, this is the same density packing ratio of an inscribed sphere to a rhombic dodecahedron polyhedron. The rhombic dodecahedron also stacks as a polyhedron with coincident lattice points to the FCC. In CCP, 12 spheres surround every sphere in the pack, and their centers are equidistant from the center of the sphere they surround. If the centers of these 12 spheres are connected, they define the vertices of a cuboctahedron, Fuller's vector equilibrium (see Section 2.1). Footnote - In synergetic geometry, Buckminster Fuller makes extensive use of CCP spheres and refers to some of their configurations as the isotropic vector matrix (IVM). See (Edmondson 1987, 127) for a good explanation of how IVM extends Fuller's concept of vector equilibrium (VE) and how it relates to his octet truss design. Steven Waterman believes that Fuller was not aware that FCC (he called it CCP) had square root integer distance between any two-sphere centers.

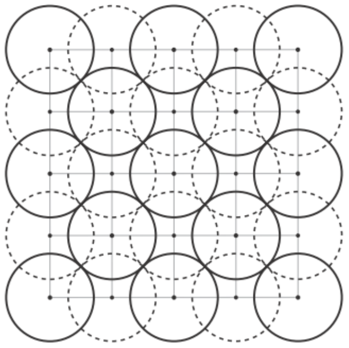

Figure 1.11: ds10a.png - Cubic closest packing.

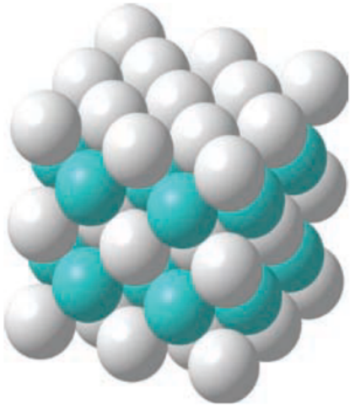

Figure 1.12: ds10b.png - Cubic closest packing.

A Waterman polyhedron is defined by the convex hull of all the lattice points (the centers of stacked spheres) that are within a defined radius from the lattice's origin. Points outside this radius are ignored. A convex hull is the minimum convex polyhedra (faces, points, and edges) that completely surround the points. In a sense, the convex hull is like a shrink-wrap surrounding the most extended set of points. The Waterman Polyhedra, called W1, are defined by a radius of $\sqrt{2}$X 1. This radius selects only the 12 points surrounding the origin, and their convex hull is the same as the cuboctahedron or Fuller's VE. As the defined radius increases, more lattice points are selected and their convex hull, the Waterman polyhedron, becomes more and more complex and is made up of many more faces, points, and edges. When the defined radius has a root greater than 300, Waterman polyhedra resemble spherical subdivisions. Figure 6.31 shows two examples with selected lattice points (with spheres around them) and the resulting convex hull Waterman polyhedron. Footnote - Images courtesy of Rene K. Mueller, SimplyDifferently.org. These Waterman polyhedra are referred to by their square root numbers. Notice that a higher root number results in more points, edges, and faces on the sphere. Waterman polyhedra exhibit high degrees of edge, face, and vertex symmetry. For some cases, it is possible to pass planes through the Waterman sphere stack and define a lattice of points that create a hemispherical or copula polyhedra that resemble a geodesic dome. Footnote - See Paul Bourke, \Waterman polyhedra data generator," http://local.wasp.uwa.edu.au/ pbourke/geometry/waterman/gen/, 2004. The FCC packing and positioning (one sphere at the origin of the Cartesian space) produces one family of Waterman polyhedra; however, other families result when the origin of the sphere cluster is repositioning to other symmetrical positions within the lattice. There are seven possible origins for FCC packings, Footnote - Seven possible point lattice origins: (1) sphere center, (2) spheres touch point, (3) spheres triangle center (there are two variations), (4) spheres forming tetrahedron, (5) spheres forming square-based pyramid, and (6) spheres. and corresponding Waterman Polyhedra can be created with sphere stacks having a 2 diameter, or any other consistent diameter. If their diameter is not 2, their centers will not necessarily fall on integer xyz-coordinates.

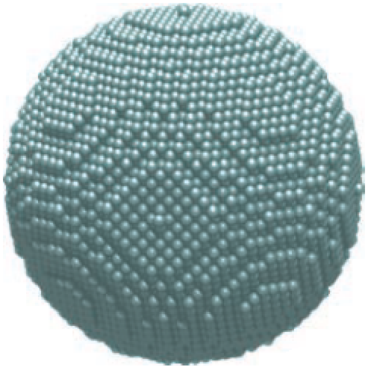

Figure 1.13: ds11a.png - $\sqrt{200}$

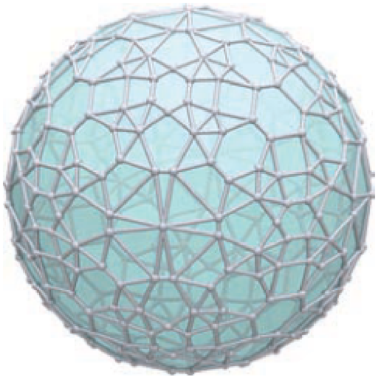

Figure 1.14: ds11b.png - $\sqrt{200}$

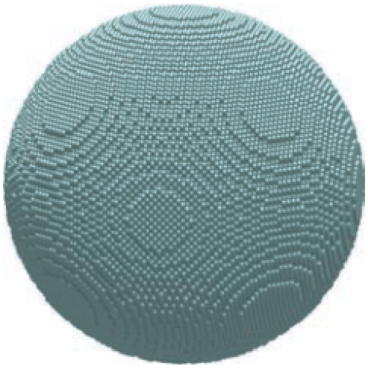

Figure 1.15: ds11c.png - $\sqrt{2000}$

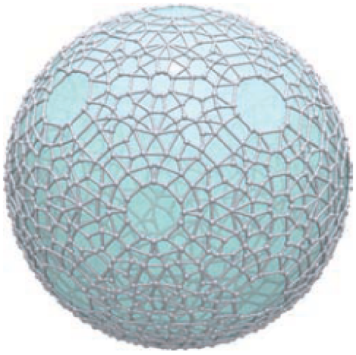

Figure 1.16: ds11d.png - $\sqrt{2000}$

Waterman polyhedra names are simple. They are labeled Wn, with n being the defined radius integer parameter. The concept of Waterman polyhedra has been generalized to include point lattices created by packed (stacked) polyhedra in addition to spheres. Packs of polyhedra, restricted to certain integer x,y,z grids, will produce more than 60 new families of polyhedra. In these cases, the Waterman polyhedra naming convention is necessarily more complex because their names must account for the type of polyhedron in the stack, the origin location for the stack lattice, and the defined radius parameter. For more information on Waterman polyhedra, see Mark Newbold's website. Footnote - http://dogfeathers.com/java/ccppoly.html. To summarize, basic Waterman polyhedra are developed from Waterman spheres with diameters of $\sqrt{2}$ and packed in a CCP or FCC manner. Their centers are located at integer x-, y-, and z-coordinates, and one sphere in the pack is centered at the Cartesian origin, at coordinates (0, 0, 0). A defined radius from the origin, a multiple of 2n, selects a set of lattice points, the centers of spheres. A Waterman polyhedron is the convex hull of the selected set of lattice points. The number of points, edges, and faces in a Waterman polyhedron (the convex hull of the selected points) obey Euler's law, which states the number of vertices + faces = edges + 2. The number of vertices in a Waterman polyhedron is a fixed set. In other words, if one Waterman polyhedron has 2,000 points, another will not have an arbitrary 2,001 points. Chords between points define polygons, most of which are not regular (equilateral triangles, pentagons, hexagons, etc.). A general definition of Waterman polyhedra include FCC lattices created by both polyhedral and spherical packs. Different Cartesian origins for these lattices in combination with defined radii producing different families of Waterman polyhedra. Not all points of a Waterman polyhedron are on a circumscribing sphere, such as the Platonic and Archimedean solids already described; Waterman polyhedra exhibit high degrees of point, edge, and face symmetry. As the defined radius increases, the number of different face types, edge lengths, angles, and areas also increase. These may be limiting factors in some spherical applications.

Pag. 202(223) - 8.5.2 Edge Reference Points Class I subdivision begins at the PPT's edges. A series of reference points is located along the edges, dividing the edges into segments equal in number to the subdivision frequency. Reference points have a regular pattern. Their spacing may be equal or unequal, but all three PPT edges will have the same pattern regardless of the spacing used. There are two approaches to defining reference points: Equal-chords and Equal-arcs. Each approach can produce several different grids. For Equal-chords, the PPT's apex-to-apex chords (all three are the same length) are subdivided with equally spaced reference points. These reference points are on the PPT's edge chord; thus, they are inside the sphere. They must be projected to the surface of the sphere before they can be used as subdivision

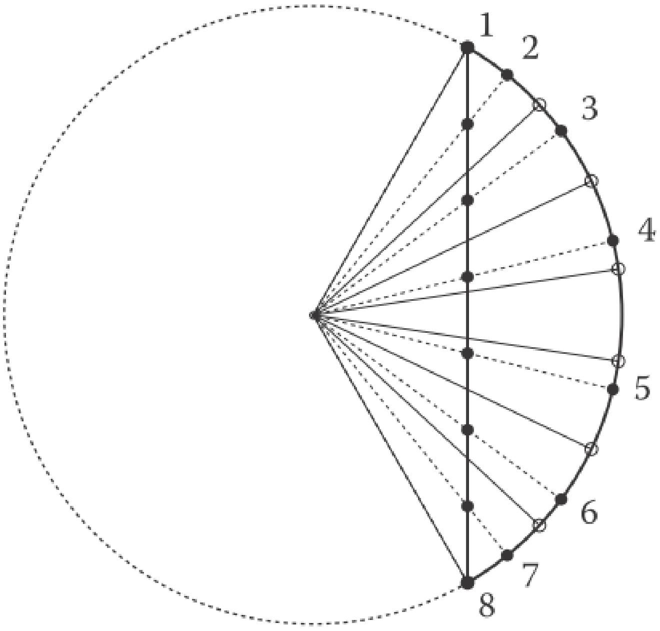

Figure 1.17: ds05-01.png - 7v Equal-chords, Equal-arcs subdivision.

references. For Equal-arcs, the PPT's arc edge (not its chords) is evenly subdivided with points, which are used as references, as is. No projection is needed because these points are already on the surface of the sphere. Figure 8.8 illustrates the difference between these two approaches with a 7v subdivision of a tetrahedron's chord. We use an odd frequency so that reference points from the two approaches will not be superimposed on one another. This makes the figure easier to read. We also use the chord of a tetrahedron, the longest of any Platonic with a unit circumsphere, to make comparisons between Equal-chords and Equal-arcs more vivid. Referring to the figure, points 1 and 8 are the chord's end points on the surface of the circumscribing sphere represented in the cross section by the circumscribing circle. For grids based on equal chord reference points, the chord between points 1 and 8 is divided into seven equal segments (7v) with points located between them. The dashed lines show chord points projected from the center of the sphere to its surface. They are numbered 2 through 7 (1 and 8 are vertices of the tetrahedron's PPT edge and are already on the sphere's surface). By inspecting the figure, you can easily see the arc spacing between projected chord points is not equal. The difference between the largest and smallest arc length is more than 2:1 in this example. To put some numbers to this, the tetrahedron's overall edge arc is approximately 109.5°, and the seven arcs resulting from Equal-chords subdivision are 9.4°, 14.1°, 19.8°, 22.8°, 19.8°, 14.1°, and 9.4°, in that order. It is also evident that the arc lengths are symmetrical; that is, arc 1-2 equals 7-8, arc 2-3 equals 6-7, and so on. The equal-arc points in Figure 8.8 are marked with small hollow circles on the surface of the sphere. They do not have to be projected. Each is spaced at 15.6o.Note that lines between them and the center of the sphere divide the chord into unequal segments. Appendix C lists simple formulas for dividing chords and arcs between any two points on the sphere into equal segments and projecting points to the surface of the sphere.

Radian Measure Another commonly used angle measurement is the radian. To define a radian, first consider a circle of radius r and two radii OA and OB. The angle formed by the two radii is a central angle. The portion of the circle between and is an arc of the circle and is written $\overline{AB}$. We say that $\overline{AB}$ subtends the angle . The length of $\overline{AB}$ is s (see Figure 5.19).

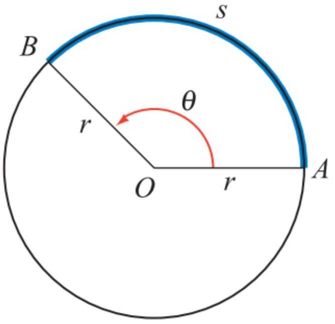

Figure 1.18: ds05-02a.png

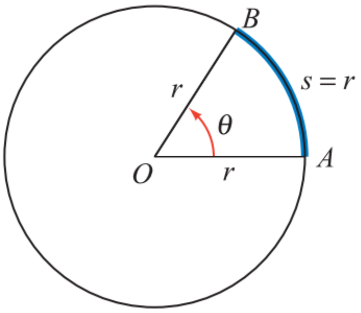

Figure 1.19: ds05-02b.png

One radian is the measure of the central angle subtended by an arc of length r on a circle of radius r.

Definition of Radian Measure

Given an arc of length s on a circle of radius r, the measure of the central angle subtended by the arc is $= \frac{s}{r}$ radians.

Source: C:/ARQ-IMPORTANT-08/Trigonometry-1439048606.pdf

Geogebra - C:/ARQ-IMPORTANT-03/Radian-20200812/radian-20200812-001.ggb

IMPORTANT

Construção Geométrica: ==============

Geogebra:/ 1) GG01 - Circle (A,B) -$x^2 + y^2 = 1$

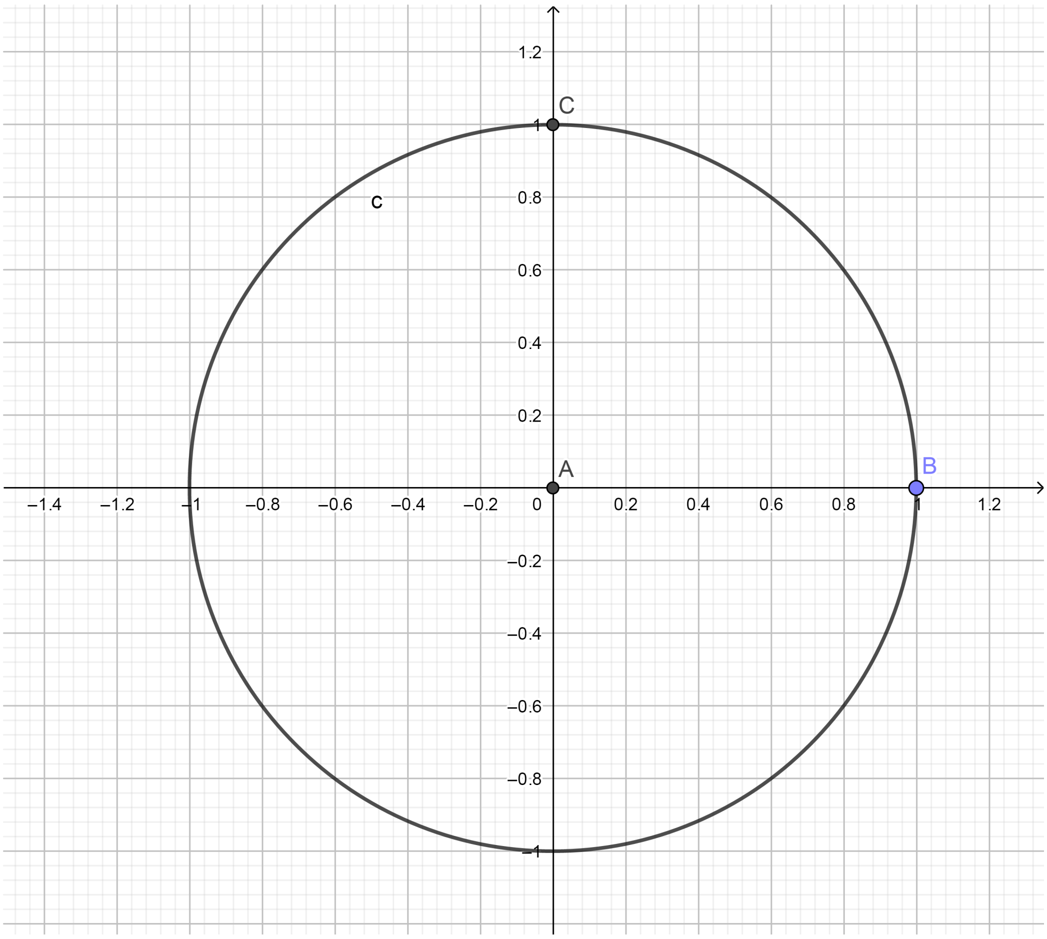

Figure 1.20: gg01.png

2) GG02- Semicircle(B,C) - Select two end points.

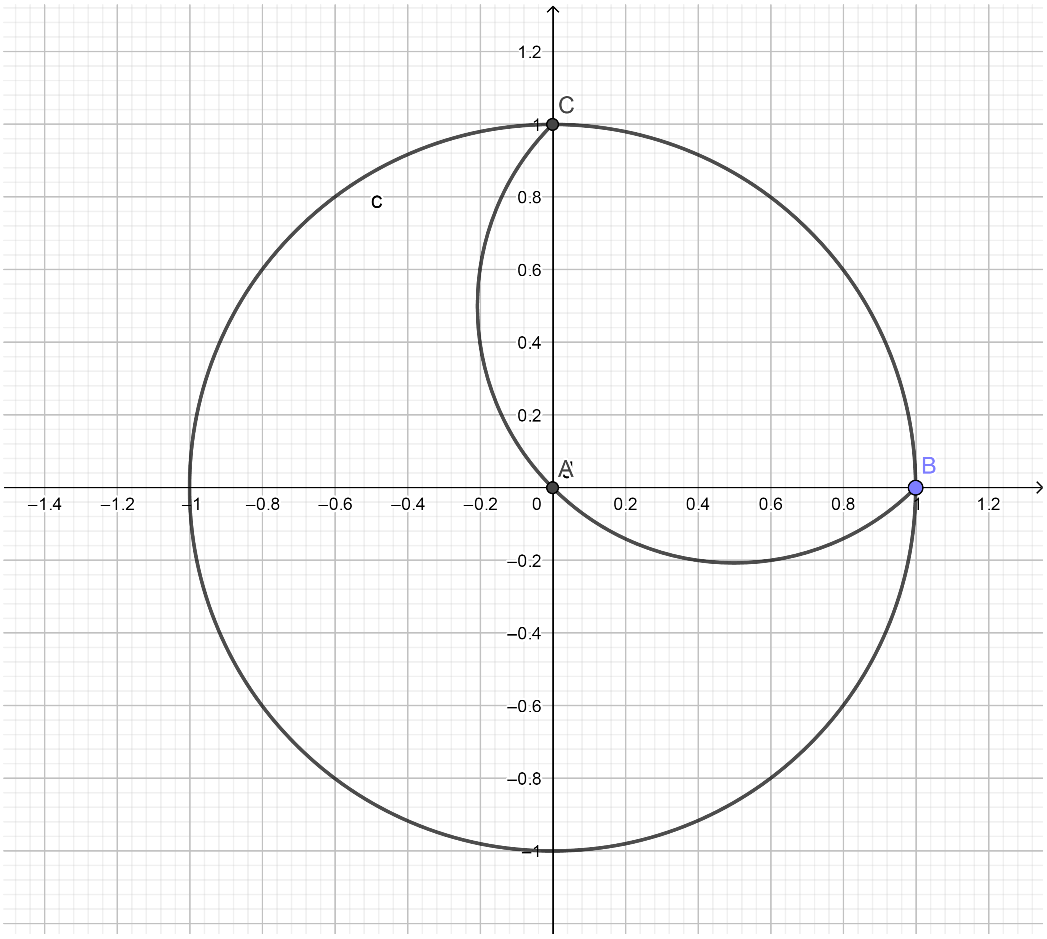
C
c
A'
B
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1.4
−1.2
−0.8
−0.6
−0.4
−0.2
0.2
0.4
0.6
0.8
1.2

Figure 1.21: gg02.png

3) GG03 - Perpendicular (D, xAxis)

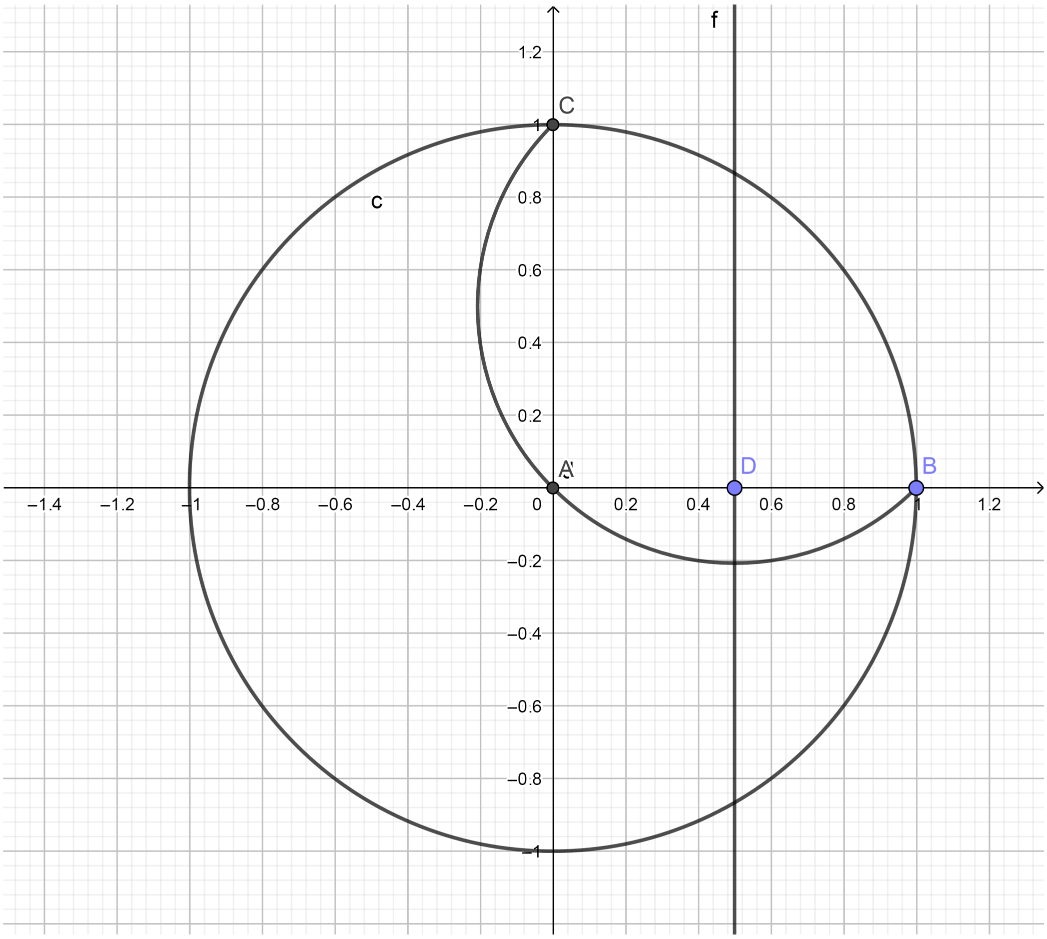
f
C
c
A'
D
B
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1.4
−1.2
0.2
0.4
0.6
0.8
1.2

Figure 1.22: gg03.png

4) GG04 - Semicircle (B,F)

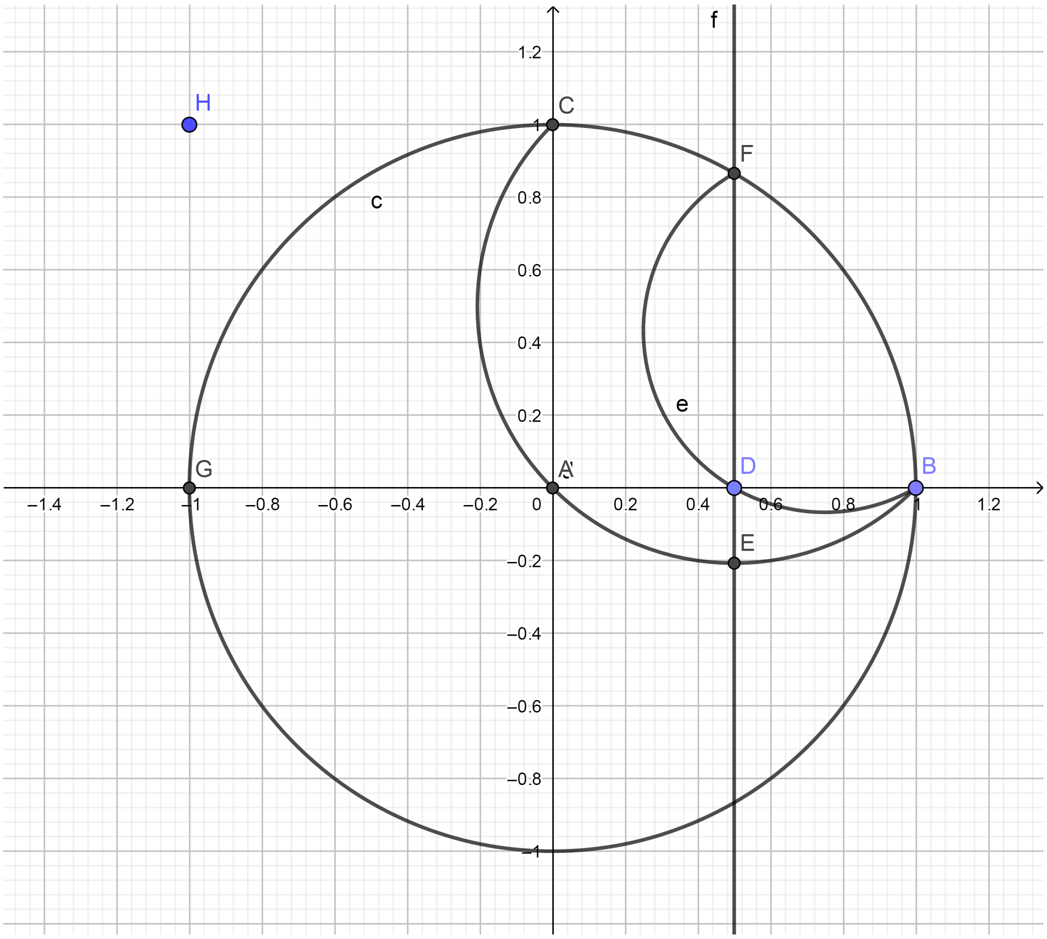
f
H
C
F
c
e
G
A'
D
B
E
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1.4
−1.2
−1
−0.8
−0.6
−0.4
−0.2
0.2
0.4
0.6
0.8
1
1.2

Figure 1.23: gg04.png

5) GG05 - Circular Arc (H, G, C) - Select center and two points.

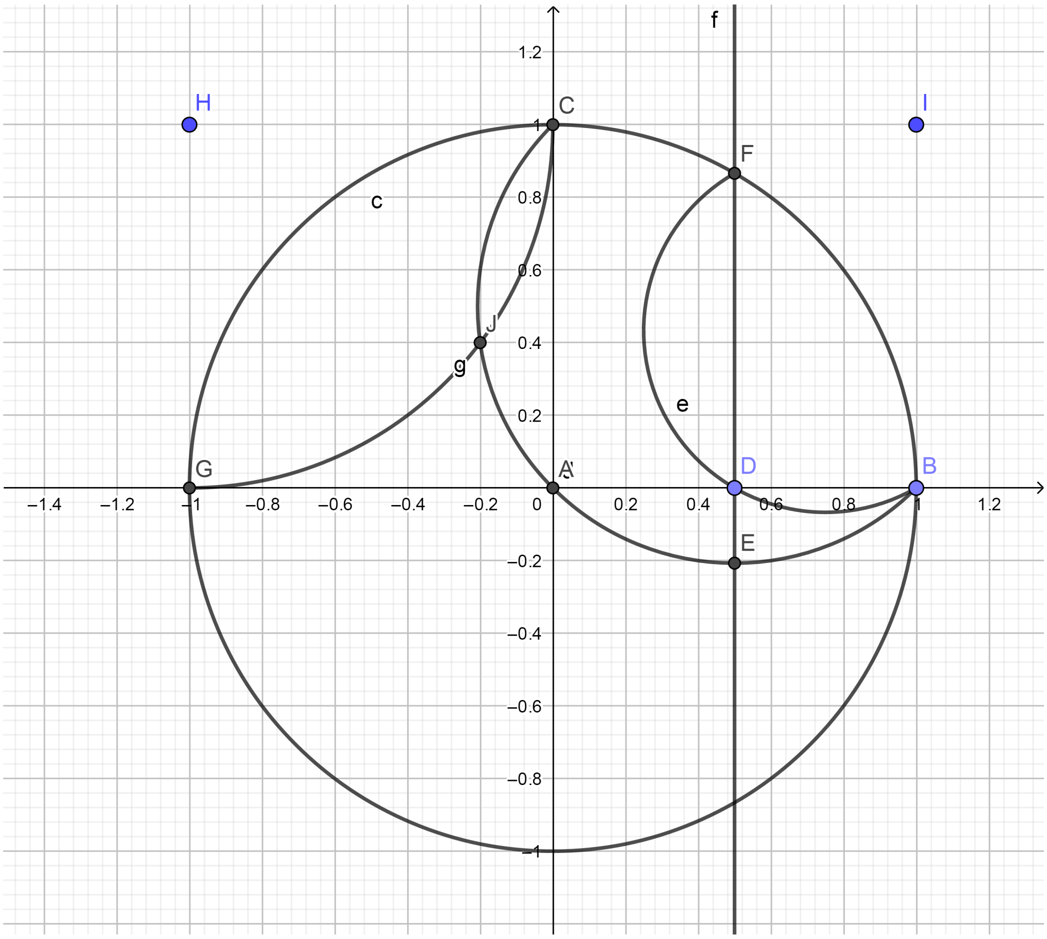
f
1.2
H
C
I
F
c
0.8
0.6
J
0.4
g
e
0.2
G
A'
D
B
−1.4
−1.2
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
1.2
E
−0.2
−0.4
−0.6
−0.8
−1

Figure 1.24: gg05.png

6) GG06 - Semicircle (B, I)

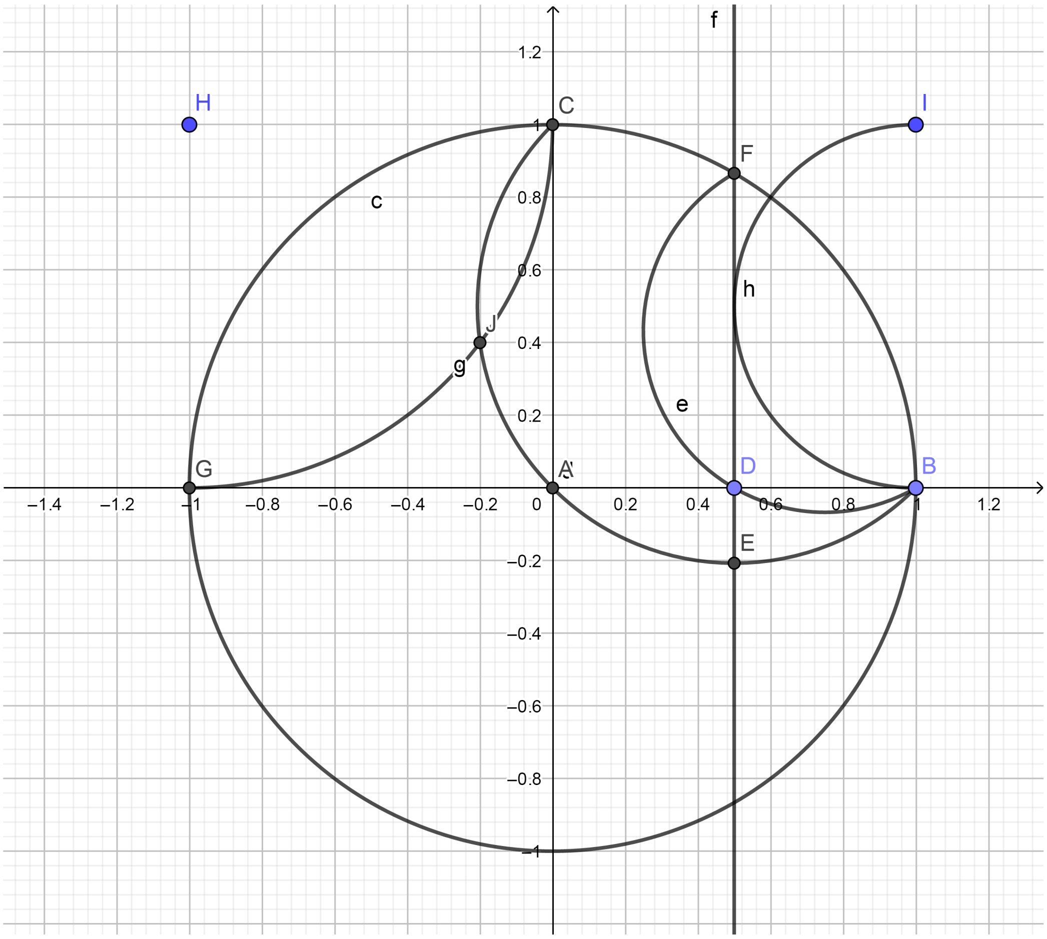
f
H
C
I
F
c
h
J
g
e
G
A
A'
D
B
E
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1.4
−1.2
−1
−0.8
−0.6
−0.4
−0.2
0.2
0.4
0.6
0.8
1
1.2

Figure 1.25: gg06.png

7) GG07 -Semicircle (I, O)

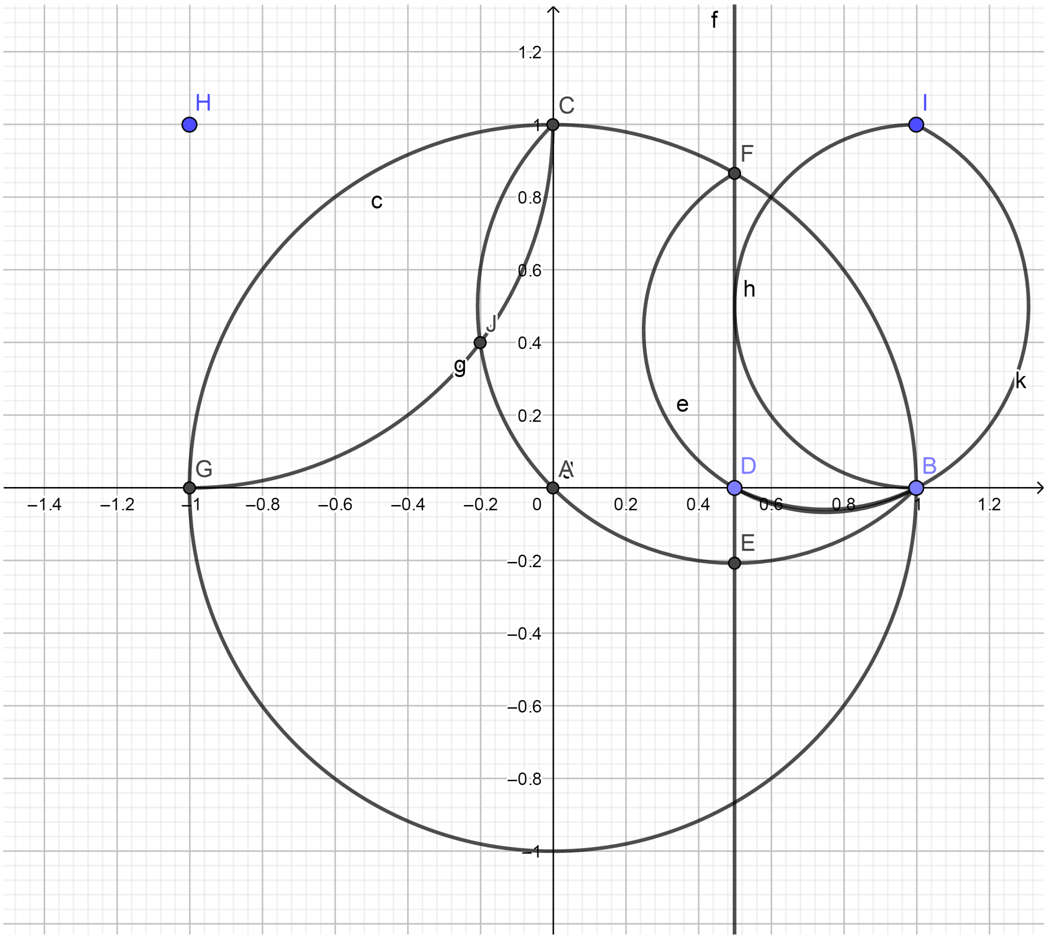
f
H
C
I
F
c
h
J
g
e
k
G
A'
D
B
E
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1.4
−1.2
−1
−0.8
−0.6
−0.4
−0.2
0.2
0.4
0.6
0.8
1
1.2

Figure 1.26: gg07.png

8) GG08 - Semicircle (B, A)

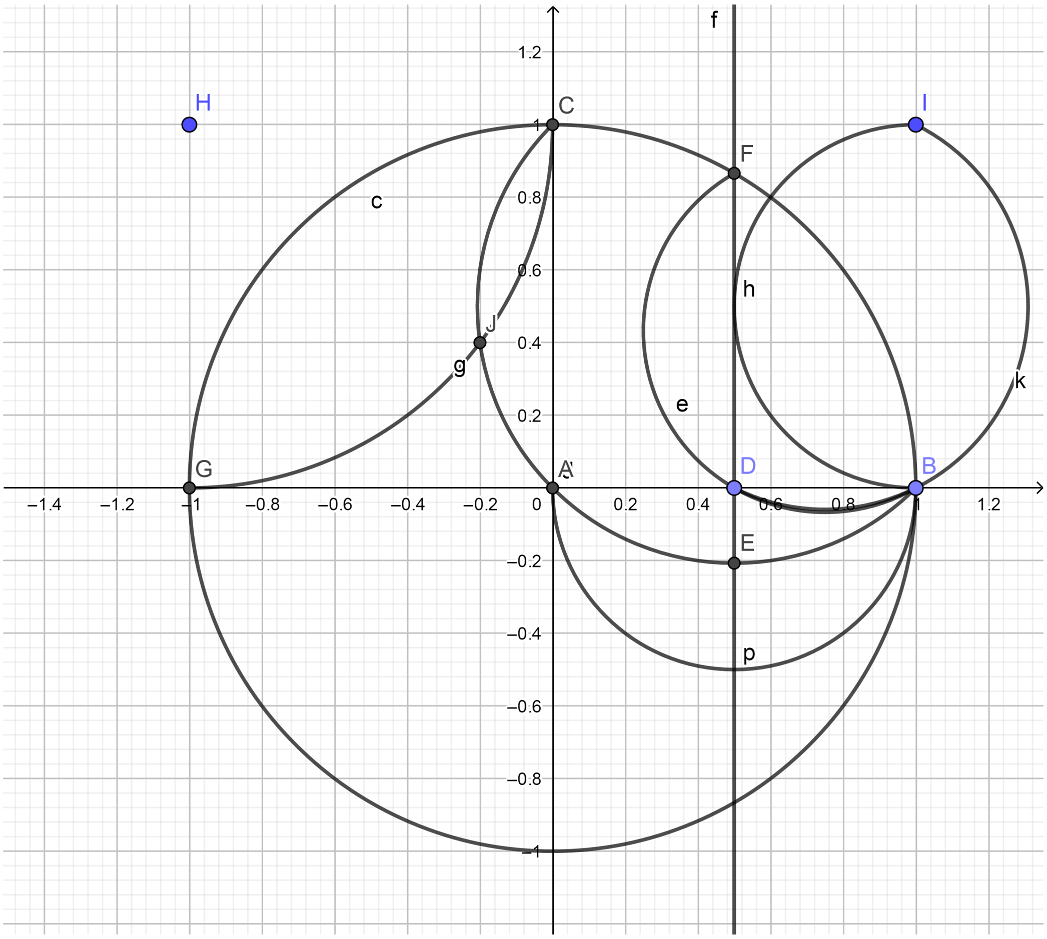
f
H
C
I
F
c
h
J
g
k
e
G
A'
D
B
E
p
1.2
1
0.8
0.6
0.4
0.2
−0.2
−0.4
−0.6
−0.8
−1
−1.4
−1.2
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
1.2

Figure 1.27: gg08.png

9) GG09 - Circular Arc (I, C, B)

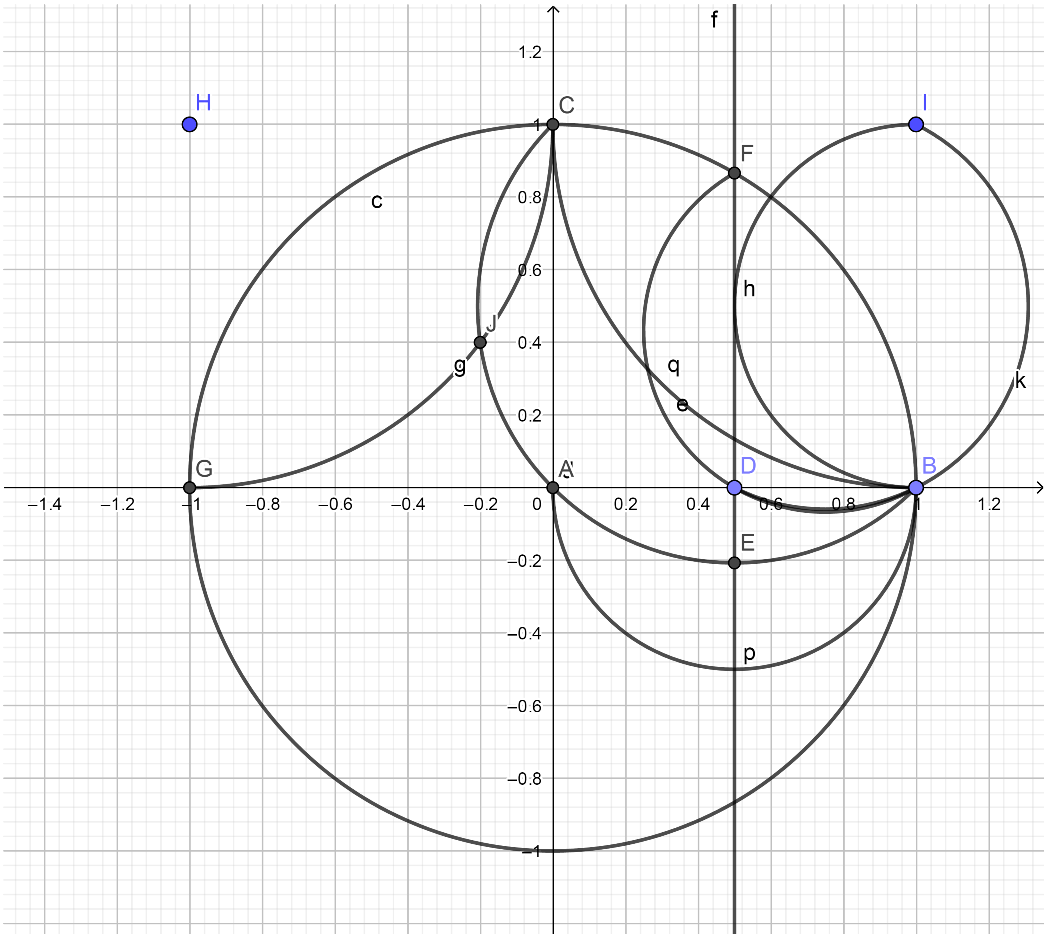
f
1.2
H
C
I
F
c
0.8
0.6
h
J
0.4
g
q
k
e
0.2
G
A'
D
B
−1.4
−1.2
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
1.2
E
−0.2
−0.4
p
−0.6
−0.8
−1

Figure 1.28: gg09.png

10) GG10 - Circle (I, D, J) - Select through 3 points.

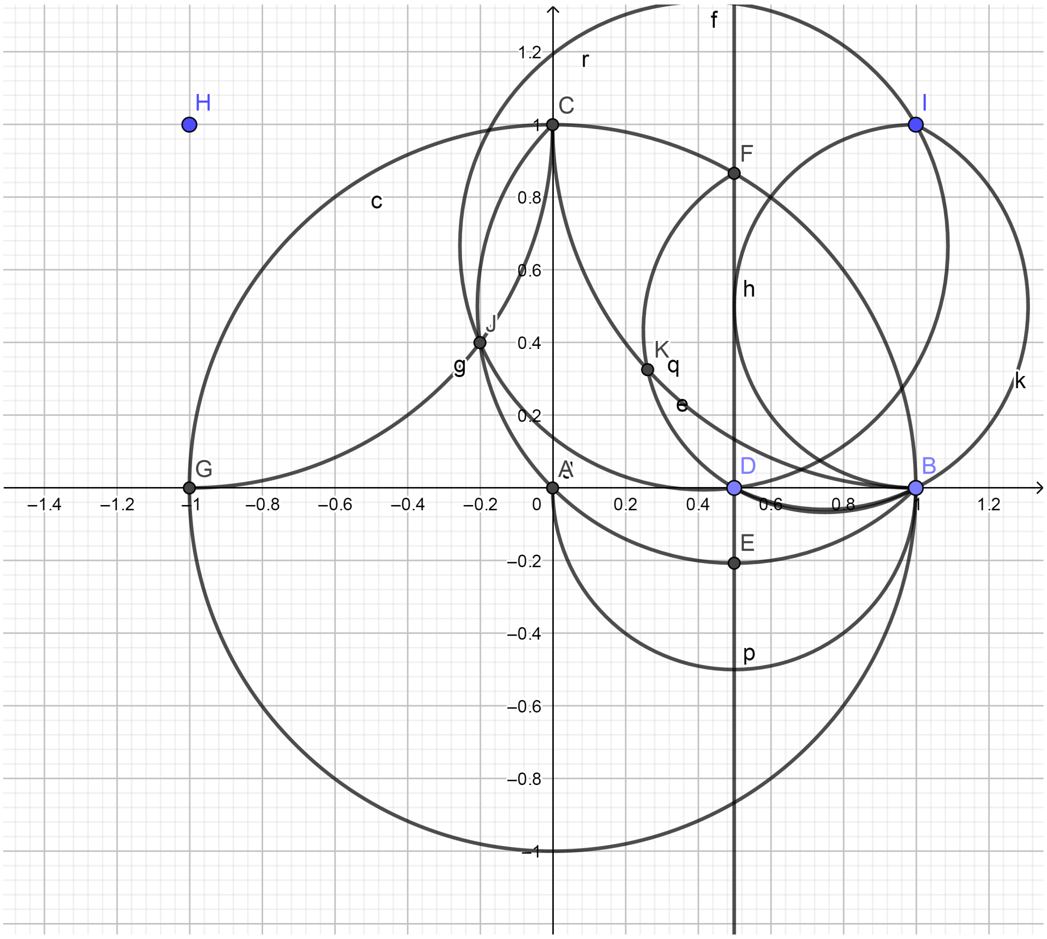
H
I
C
F
c
f
r
h
J
g
K
q
e
k
G
A'
D
B
E
p
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1.4
−1.2
−1
−0.8
−0.6
−0.4
−0.2
0.2
0.4
0.6
0.8
1
1.2

Figure 1.29: gg10.png

Vamos comparar os semircicle (B, F) e o semicircle (I, D), observaremos uma não coincidência das linhas na parte inferior dos referidos círculos.

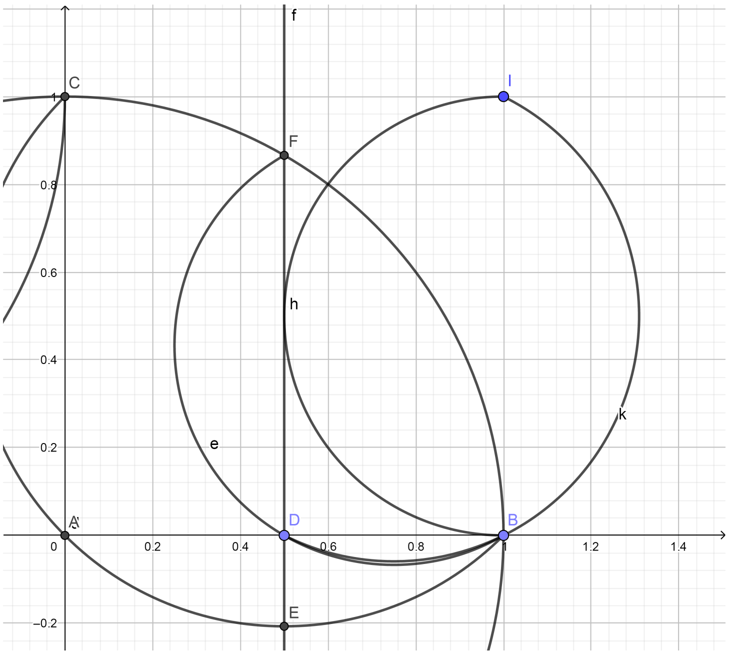
f
C
1
I
F
0.8
0.6
h
0.4
0.2
e
k
A
D
B
0
0.2
0.4
0.6
0.8
1
1.2
1.4
E
−0.2

Figure 1.30: gg07a.png

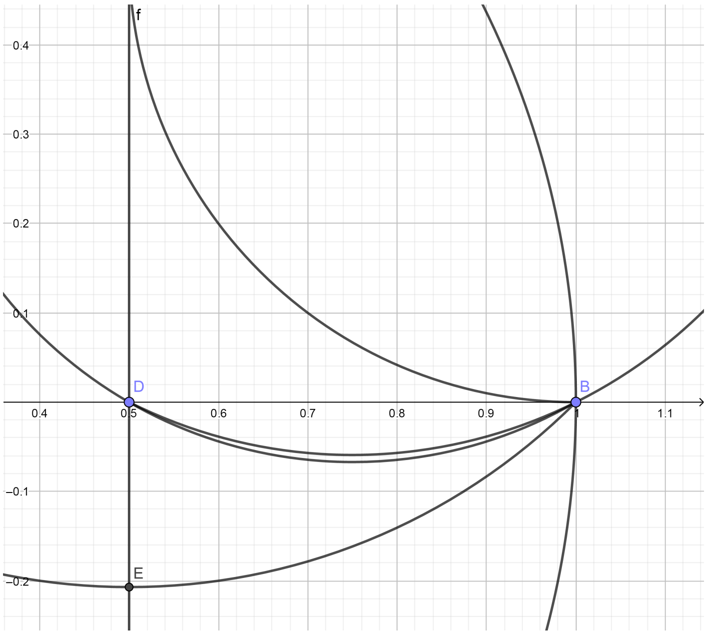
f
0.4
0.3
0.2
0.1
D
B
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
−0.1
E
−0.2

Figure 1.31: gg07b.png

Pag. 236(257) - Class III grid development and PPT projection sequence.

8.7.5 Developing Grids Class III lacks the convenient 3D equal-chord and equal-arc references that were the starting points for Class I and II subdivisions. For Class III, we use a different approach and create a 2D grid, which we will eventually project onto the PPT's surface. Although we will develop a triangular grid in our example, it is possible to use this same technique to project any 2D geometry or pattern onto a sphere. Figure 8.42 graphically summarizes the entire sequence we will use; the remainder of this chapter details these steps. The PPT's tessellation is first defined as a 2D trigrid with the required frequency and orientation relative to the PPT's edges. Figure 8.42(a) shows the grid and the triangular faces that are completely within the PPT. Trigrid point coordinates are converted to Cartesian values and positioned so that one of the grid's directions is perpendicular to the xz-plane and centered over the origin, where it is scaled to correspond to the final PPT's chord length, as shown in (b). The scaled grid is then translated up the +z-axis, a distance equal to the polyhedron's insphere radius. The center of the grid is tangent to the insphere and its apices are tangent to its circumsphere (c). Trigrid points are projected from the origin to the surface of the sphere by normalizing each grid point's coordinates. An arrow from the origin through one grid point shows its projection to the sphere (d). The result is a single face-zenith tessellated PPT. Copies are then replicated to cover the rest of the sphere.

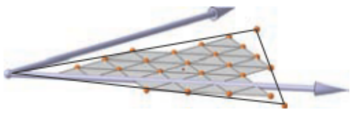

Figure 1.32: ds01a.png

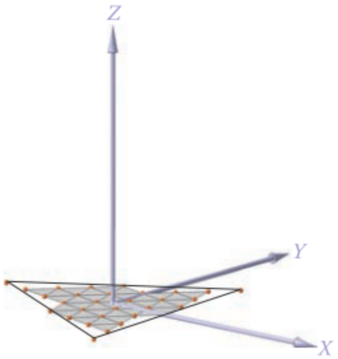

Figure 1.33: ds01b.png

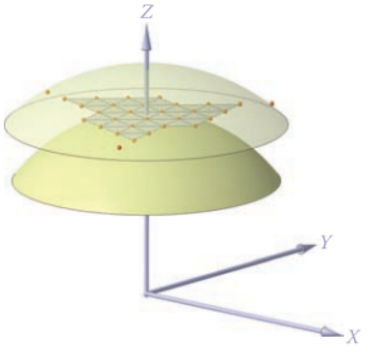

Figure 1.34: ds01c.png

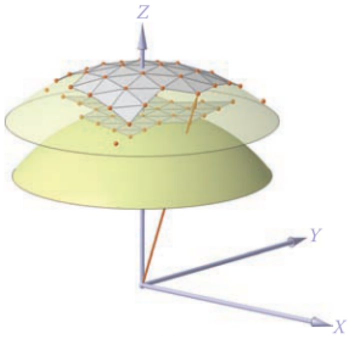

Figure 1.35: ph01d.png

2) C:/DATA-NovoVivo-127-2020/Books-127-2020/Advances In Geometric Modeling/Advances In Geometric Modeling.pdf Pag. 01(13) - Polygonal Subdivision Curves for Computer Graphics and Geometric Modeling.

A polygonal subdivision curve is the limit of subdivision of a polygonal complex. Such a complex, whose shape depends on the scheme involved, is simply a control mesh that converges to a curve rather than a surface. The use of polygonal complexes has been motivated by the curve interpolation problem in subdivision surfaces. This is one of the major interpolation constraints that is partially addressed in the literature. The definition of curves by polygonal complexes carries with it cross derivative information which can be naturally embodied in the mesh of a subdivision surface. This chapter gives an overview of such complexes, their polygonal subdivision curves and their applications for computer graphics and geometric modeling.

3) C:/DATA-NovoVivo-127-2020/Books-127-2020/Principles of Astrophysical Fluid Dynamics 2007/Principles of Astrophysical Fluid Dynamics 2007.pdf Pag. 20 (32) - Chapter 3 - Gravitation.

The gravitational force is of course of central importance in astrophysics. Therefore whereas the force of gravity plays only a minor role in most textbooks on terrestrial fluid mechanics, with such flu- ids being only subject to the uniform gravitational acceleration of the Earth, the astronomical situation is quite different. We will need rou- tinely to be able to solve for the gravitational field produced by the fluid itself before we can calculate the gravitational term in the momen- tum equation which was derived in Chapter 2. This chapter therefore represents an aside on how to calculate gravitational fields from given density distributions.

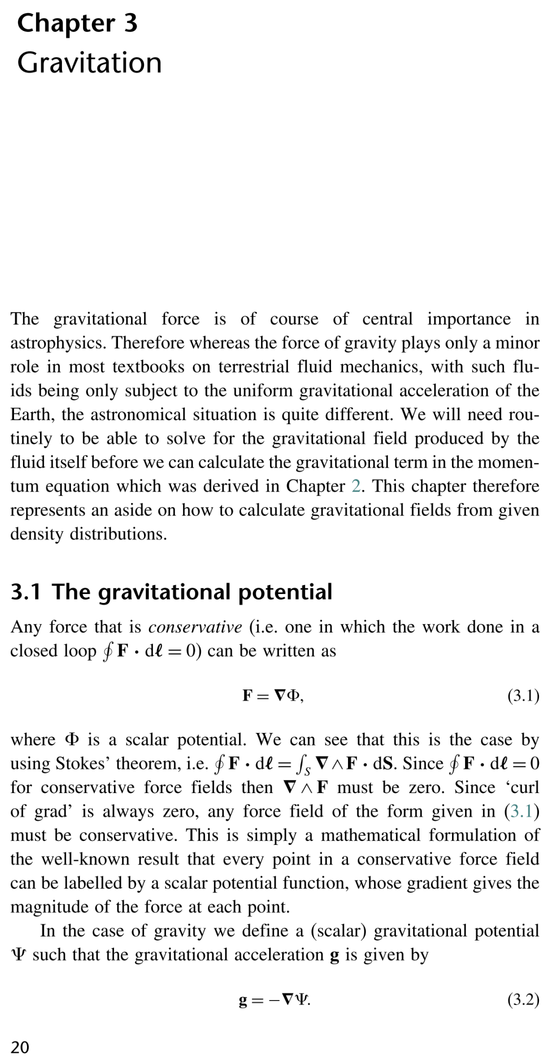

Chapter 3

Gravitation

The gravitational force is of course of central importance in astrophysics. Therefore whereas the force of gravity plays only a minor role in most textbooks on terrestrial fluid mechanics, with such fluids being only subject to the uniform gravitational acceleration of the Earth, the astronomical situation is quite different. We will need routinely to be able to solve for the gravitational field produced by the fluid itself before we can calculate the gravitational term in the momentum equation which was derived in Chapter 2. This chapter therefore represents an aside on how to calculate gravitational fields from given density distributions.

3.1 The gravitational potential

Any force that is *conservative* (i.e. one in which the work done in a closed loop $\oint \mathbf{F} \cdot \mathrm{d}\boldsymbol{\ell} = 0$) can be written as

$$\mathbf{F} = \boldsymbol{\nabla}\Phi, \tag{3.1}$$

where Φ is a scalar potential. We can see that this is the case by using Stokes' theorem, i.e. $\oint \mathbf{F} \cdot \mathrm{d}\boldsymbol{\ell} = \int_S \boldsymbol{\nabla} \wedge \mathbf{F} \cdot \mathrm{d}\mathbf{S}$. Since $\oint \mathbf{F} \cdot \mathrm{d}\boldsymbol{\ell} = 0$ for conservative force fields then $\boldsymbol{\nabla} \wedge \mathbf{F}$ must be zero. Since 'curl of grad' is always zero, any force field of the form given in (3.1) must be conservative. This is simply a mathematical formulation of the well-known result that every point in a conservative force field can be labelled by a scalar potential function, whose gradient gives the magnitude of the force at each point.

In the case of gravity we define a (scalar) gravitational potential Ψ such that the gravitational acceleration $\mathbf{g}$ is given by

$$\mathbf{g} = -\boldsymbol{\nabla}\Psi. \tag{3.2}$$

20

Figure 1.36: pafd01.png

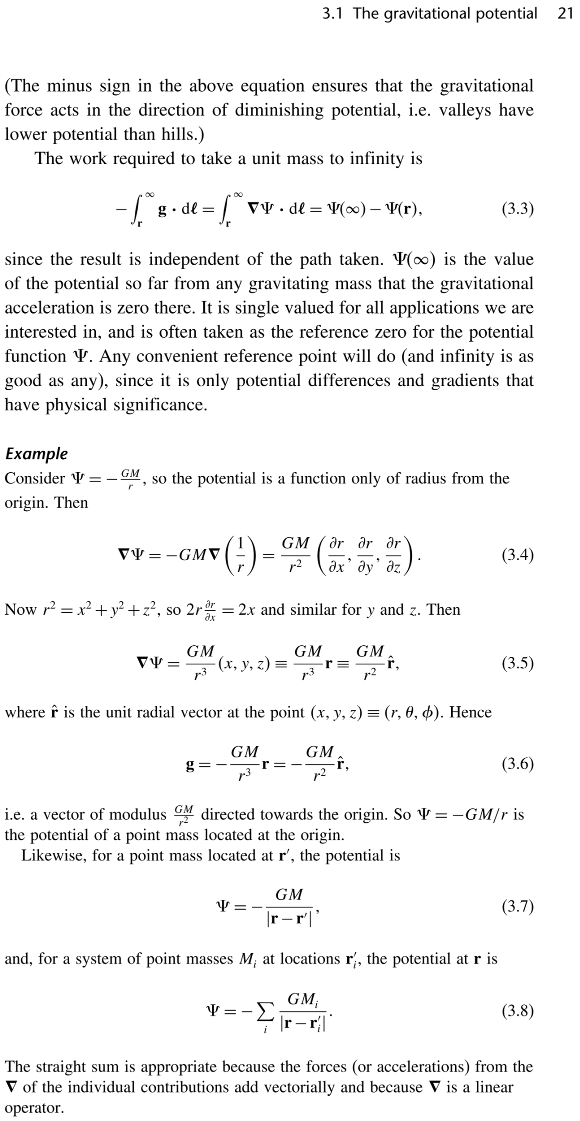

(The minus sign in the above equation ensures that the gravitational force acts in the direction of diminishing potential, i.e. valleys have lower potential than hills.)

The work required to take a unit mass to infinity is

$$-\int_{\mathbf{r}}^{\infty} \mathbf{g} \cdot \mathrm{d}\boldsymbol{\ell} = \int_{\mathbf{r}}^{\infty} \boldsymbol{\nabla}\Psi \cdot \mathrm{d}\boldsymbol{\ell} = \Psi(\infty) - \Psi(\mathbf{r}), \tag{3.3}$$

since the result is independent of the path taken. $\Psi(\infty)$ is the value of the potential so far from any gravitating mass that the gravitational acceleration is zero there. It is single valued for all applications we are interested in, and is often taken as the reference zero for the potential function Ψ. Any convenient reference point will do (and infinity is as good as any), since it is only potential differences and gradients that have physical significance.

Example

Consider $\Psi = -\frac{GM}{r}$, so the potential is a function only of radius from the origin. Then

$$\boldsymbol{\nabla}\Psi = -GM\boldsymbol{\nabla}\left(\frac{1}{r}\right) = \frac{GM}{r^2}\left(\frac{\partial r}{\partial x}, \frac{\partial r}{\partial y}, \frac{\partial r}{\partial z}\right). \tag{3.4}$$

Now $r^2 = x^2 + y^2 + z^2$, so $2r\frac{\partial r}{\partial x} = 2x$ and similar for y and z. Then

$$\boldsymbol{\nabla}\Psi = \frac{GM}{r^3}(x, y, z) \equiv \frac{GM}{r^3}\mathbf{r} \equiv \frac{GM}{r^2}\hat{\mathbf{r}}, \tag{3.5}$$

where $\hat{\mathbf{r}}$ is the unit radial vector at the point $(x, y, z) \equiv (r, \theta, \phi)$. Hence

$$\mathbf{g} = -\frac{GM}{r^3}\mathbf{r} = -\frac{GM}{r^2}\hat{\mathbf{r}}, \tag{3.6}$$

i.e. a vector of modulus $\frac{GM}{r^2}$ directed towards the origin. So $\Psi = -GM/r$ is the potential of a point mass located at the origin.

Likewise, for a point mass located at $\mathbf{r}'$, the potential is

$$\Psi = -\frac{GM}{|\mathbf{r} - \mathbf{r}'|}, \tag{3.7}$$

and, for a system of point masses M_i at locations $\mathbf{r}_i'$, the potential at $\mathbf{r}$ is

$$\Psi = -\sum_i \frac{GM_i}{|\mathbf{r} - \mathbf{r}_i'|}. \tag{3.8}$$

The straight sum is appropriate because the forces (or accelerations) from the $\boldsymbol{\nabla}$ of the individual contributions add vectorially and because $\boldsymbol{\nabla}$ is a linear operator.

Figure 1.37: pafd02.png

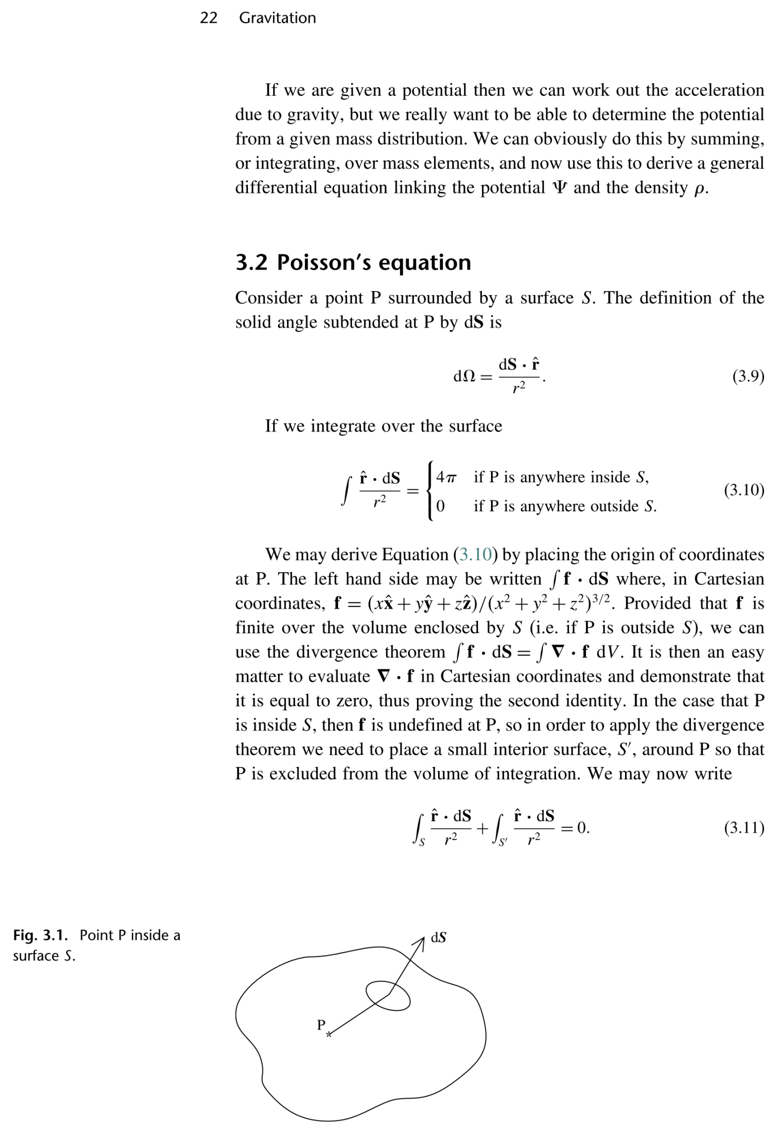

If we are given a potential then we can work out the acceleration due to gravity, but we really want to be able to determine the potential from a given mass distribution. We can obviously do this by summing, or integrating, over mass elements, and now use this to derive a general differential equation linking the potential Ψ and the density ρ.

3.2 Poisson's equation

Consider a point P surrounded by a surface S. The definition of the solid angle subtended at P by d$\mathbf{S}$ is

$$d\Omega = \frac{d\mathbf{S} \cdot \hat{\mathbf{r}}}{r^2}. \tag{3.9}$$

If we integrate over the surface

$$\int \frac{\hat{\mathbf{r}} \cdot d\mathbf{S}}{r^2} = \begin{cases} 4\pi & \text{if P is anywhere inside } S, \\ 0 & \text{if P is anywhere outside } S. \end{cases} \tag{3.10}$$

We may derive Equation (3.10) by placing the origin of coordinates at P. The left hand side may be written $\int \mathbf{f} \cdot d\mathbf{S}$ where, in Cartesian coordinates, $\mathbf{f} = (x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}})/(x^2 + y^2 + z^2)^{3/2}$. Provided that $\mathbf{f}$ is finite over the volume enclosed by S (i.e. if P is outside S), we can use the divergence theorem $\int \mathbf{f} \cdot d\mathbf{S} = \int \nabla \cdot \mathbf{f}\, dV$. It is then an easy matter to evaluate $\nabla \cdot \mathbf{f}$ in Cartesian coordinates and demonstrate that it is equal to zero, thus proving the second identity. In the case that P is inside S, then $\mathbf{f}$ is undefined at P, so in order to apply the divergence theorem we need to place a small interior surface, S', around P so that P is excluded from the volume of integration. We may now write

$$\int_S \frac{\hat{\mathbf{r}} \cdot d\mathbf{S}}{r^2} + \int_{S'} \frac{\hat{\mathbf{r}} \cdot d\mathbf{S}}{r^2} = 0. \tag{3.11}$$

Fig. 3.1. Point P inside a surface S.

Figure 1.38: pafd03.png

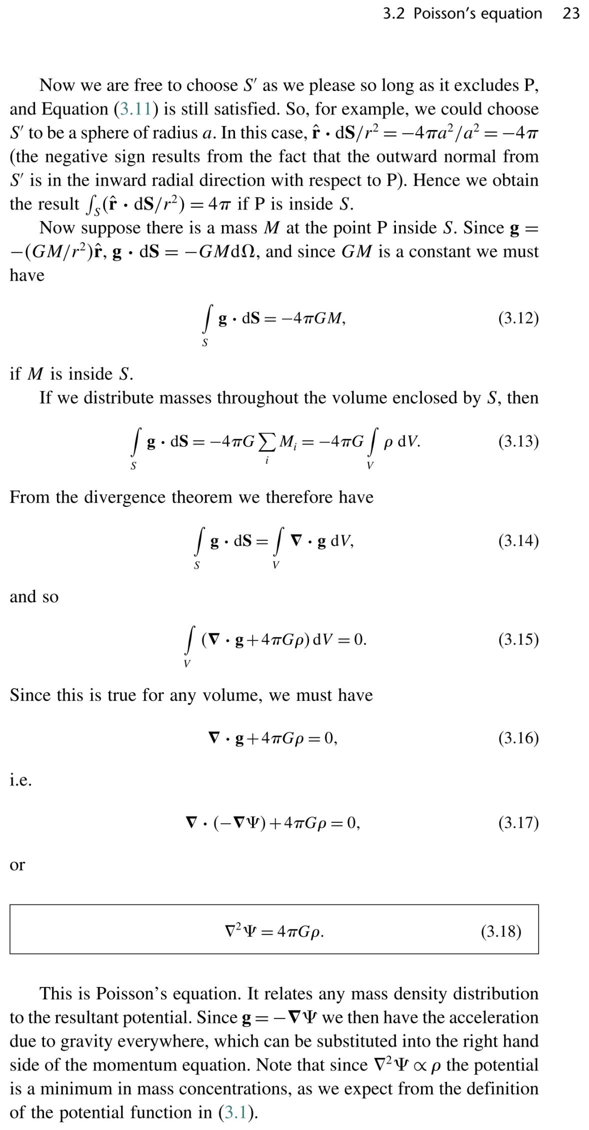

Now we are free to choose S' as we please so long as it excludes P, and Equation (3.11) is still satisfied. So, for example, we could choose S' to be a sphere of radius a. In this case, $\hat{\mathbf{r}}\cdot \mathrm{d}\mathbf{S}/r^2 = -4\pi a^2/a^2 = -4\pi$ (the negative sign results from the fact that the outward normal from S' is in the inward radial direction with respect to P). Hence we obtain the result $\int_S(\hat{\mathbf{r}}\cdot \mathrm{d}\mathbf{S}/r^2) = 4\pi$ if P is inside S.

Now suppose there is a mass M at the point P inside S. Since $\mathbf{g} = -(GM/r^2)\hat{\mathbf{r}}$, $\mathbf{g}\cdot \mathrm{d}\mathbf{S} = -GM\mathrm{d}\Omega$, and since GM is a constant we must have

$$\int_S \mathbf{g}\cdot \mathrm{d}\mathbf{S} = -4\pi GM, \tag{3.12}$$

if M is inside S.

If we distribute masses throughout the volume enclosed by S, then

$$\int_S \mathbf{g}\cdot \mathrm{d}\mathbf{S} = -4\pi G\sum_i M_i = -4\pi G\int_V \rho\ \mathrm{d}V. \tag{3.13}$$

From the divergence theorem we therefore have

$$\int_S \mathbf{g}\cdot \mathrm{d}\mathbf{S} = \int_V \boldsymbol{\nabla}\cdot \mathbf{g}\ \mathrm{d}V, \tag{3.14}$$

and so

$$\int_V (\boldsymbol{\nabla}\cdot \mathbf{g} + 4\pi G\rho)\,\mathrm{d}V = 0. \tag{3.15}$$

Since this is true for any volume, we must have

$$\boldsymbol{\nabla}\cdot \mathbf{g} + 4\pi G\rho = 0, \tag{3.16}$$

i.e.

$$\boldsymbol{\nabla}\cdot(-\boldsymbol{\nabla}\Psi) + 4\pi G\rho = 0, \tag{3.17}$$

or

$$\boxed{\nabla^2\Psi = 4\pi G\rho.} \tag{3.18}$$

This is Poisson's equation. It relates any mass density distribution to the resultant potential. Since $\mathbf{g} = -\boldsymbol{\nabla}\Psi$ we then have the acceleration due to gravity everywhere, which can be substituted into the right hand side of the momentum equation. Note that since $\nabla^2\Psi \propto \rho$ the potential is a minimum in mass concentrations, as we expect from the definition of the potential function in (3.1).

Figure 1.39: pafd04.png

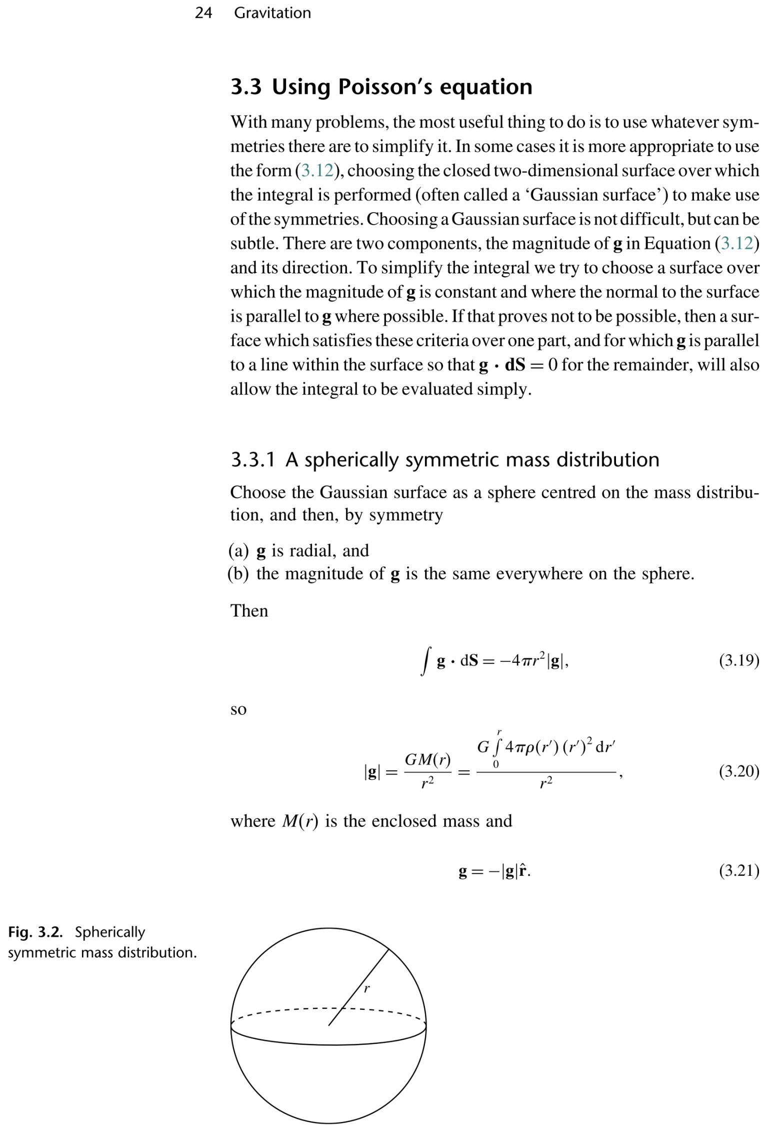

3.3 Using Poisson's equation

With many problems, the most useful thing to do is to use whatever symmetries there are to simplify it. In some cases it is more appropriate to use the form (3.12), choosing the closed two-dimensional surface over which the integral is performed (often called a 'Gaussian surface') to make use of the symmetries. Choosing a Gaussian surface is not difficult, but can be subtle. There are two components, the magnitude of **g** in Equation (3.12) and its direction. To simplify the integral we try to choose a surface over which the magnitude of **g** is constant and where the normal to the surface is parallel to **g** where possible. If that proves not to be possible, then a surface which satisfies these criteria over one part, and for which **g** is parallel to a line within the surface so that $\mathbf{g} \cdot \mathbf{dS} = 0$ for the remainder, will also allow the integral to be evaluated simply.

3.3.1 A spherically symmetric mass distribution

Choose the Gaussian surface as a sphere centred on the mass distribution, and then, by symmetry

(a) **g** is radial, and
(b) the magnitude of **g** is the same everywhere on the sphere.

Then

$$\int \mathbf{g} \cdot \mathrm{d}\mathbf{S} = -4\pi r^2 |\mathbf{g}|, \tag{3.19}$$

so

$$|\mathbf{g}| = \frac{GM(r)}{r^2} = \frac{G\int\limits_0^r 4\pi\rho(r')\,(r')^2\,\mathrm{d}r'}{r^2}, \tag{3.20}$$

where $M(r)$ is the enclosed mass and

$$\mathbf{g} = -|\mathbf{g}|\hat{\mathbf{r}}. \tag{3.21}$$

Fig. 3.2. Spherically symmetric mass distribution.

Figure 1.40: pafd05.png

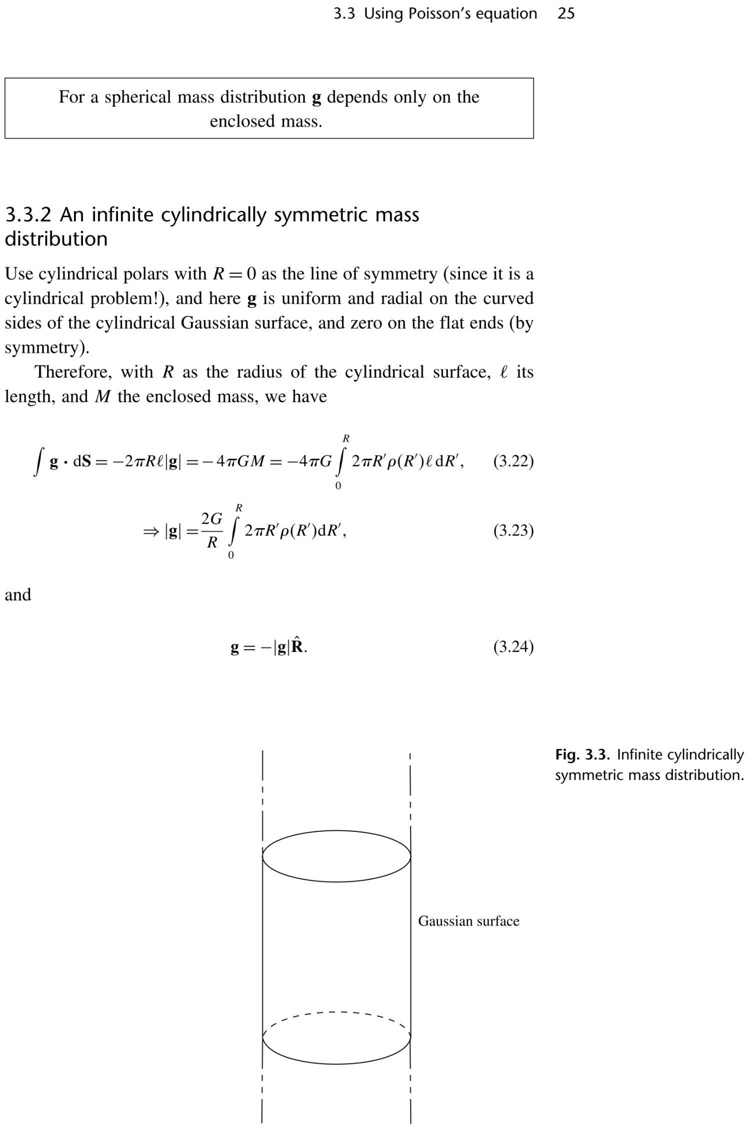

For a spherical mass distribution **g** depends only on the enclosed mass.

3.3.2 An infinite cylindrically symmetric mass distribution

Use cylindrical polars with $R = 0$ as the line of symmetry (since it is a cylindrical problem!), and here **g** is uniform and radial on the curved sides of the cylindrical Gaussian surface, and zero on the flat ends (by symmetry).

Therefore, with R as the radius of the cylindrical surface, ℓ its length, and M the enclosed mass, we have

$$\int \mathbf{g}\cdot \mathrm{d}\mathbf{S} = -2\pi R\ell|\mathbf{g}| = -4\pi GM = -4\pi G\int_0^R 2\pi R'\rho(R')\ell\,\mathrm{d}R', \tag{3.22}$$

$$\Rightarrow |\mathbf{g}| = \frac{2G}{R}\int_0^R 2\pi R'\rho(R')\mathrm{d}R', \tag{3.23}$$

and

$$\mathbf{g} = -|\mathbf{g}|\hat{\mathbf{R}}. \tag{3.24}$$

Fig. 3.3. Infinite cylindrically symmetric mass distribution.

Figure 1.41: pafd06.png

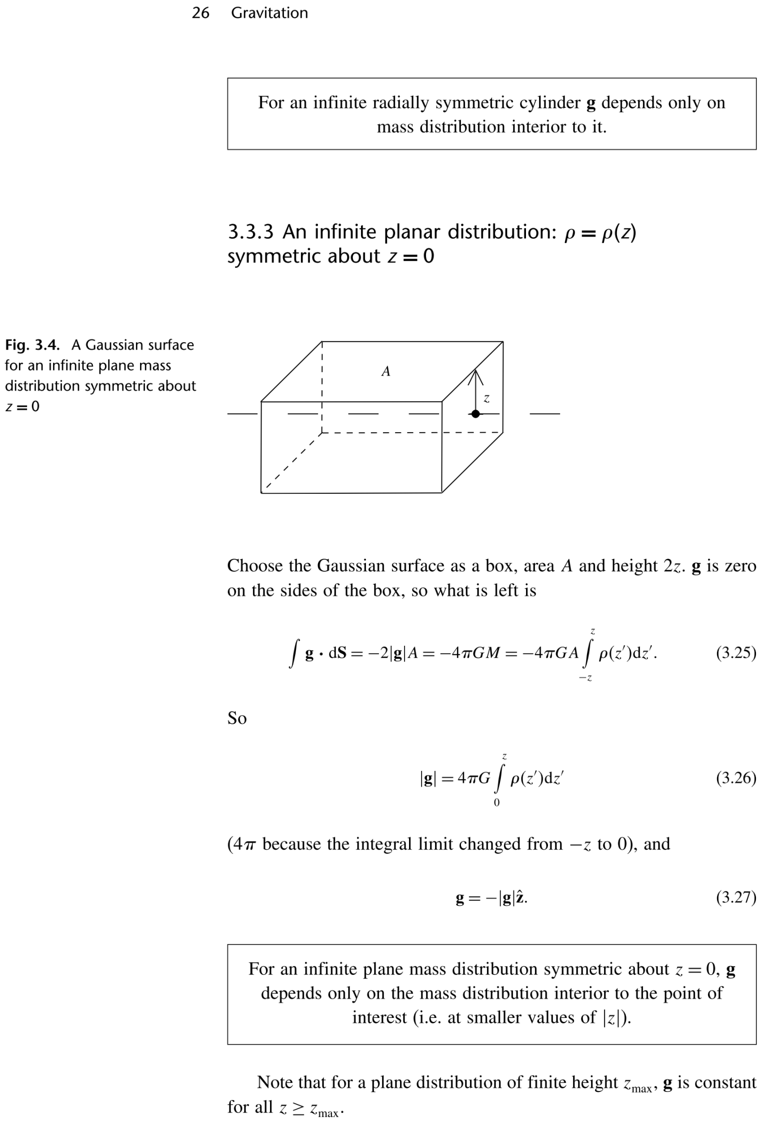

For an infinite radially symmetric cylinder **g** depends only on mass distribution interior to it.

3.3.3 An infinite planar distribution: $\rho = \rho(z)$ symmetric about $z = 0$

Fig. 3.4. A Gaussian surface for an infinite plane mass distribution symmetric about $z = 0$

Choose the Gaussian surface as a box, area A and height $2z$. **g** is zero on the sides of the box, so what is left is

$$\int \mathbf{g} \cdot \mathrm{d}\mathbf{S} = -2|\mathbf{g}|A = -4\pi GM = -4\pi GA \int_{-z}^{z} \rho(z')\mathrm{d}z'. \tag{3.25}$$

So

$$|\mathbf{g}| = 4\pi G \int_{0}^{z} \rho(z')\mathrm{d}z' \tag{3.26}$$

(4π because the integral limit changed from $-z$ to 0), and

$$\mathbf{g} = -|\mathbf{g}|\hat{\mathbf{z}}. \tag{3.27}$$

For an infinite plane mass distribution symmetric about $z = 0$, **g** depends only on the mass distribution interior to the point of interest (i.e. at smaller values of $|z|$).

Note that for a plane distribution of finite height z_{max}, **g** is constant for all $z \geq z_{\mathrm{max}}$.

Figure 1.42: pafd07.png

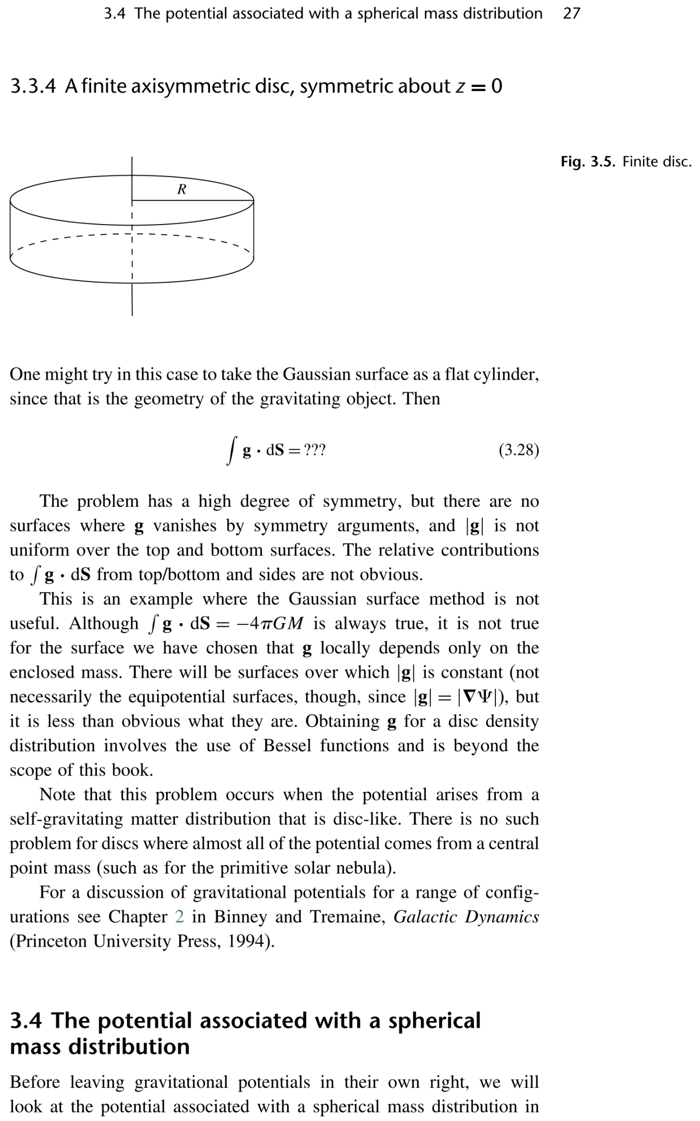

3.3.4 A finite axisymmetric disc, symmetric about $z = 0$

Fig. 3.5. Finite disc.

One might try in this case to take the Gaussian surface as a flat cylinder, since that is the geometry of the gravitating object. Then

$$\int \mathbf{g} \cdot \mathrm{d}\mathbf{S} = ??? \tag{3.28}$$

The problem has a high degree of symmetry, but there are no surfaces where $\mathbf{g}$ vanishes by symmetry arguments, and $|\mathbf{g}|$ is not uniform over the top and bottom surfaces. The relative contributions to $\int \mathbf{g} \cdot \mathrm{d}\mathbf{S}$ from top/bottom and sides are not obvious.

This is an example where the Gaussian surface method is not useful. Although $\int \mathbf{g} \cdot \mathrm{d}\mathbf{S} = -4\pi GM$ is always true, it is not true for the surface we have chosen that $\mathbf{g}$ locally depends only on the enclosed mass. There will be surfaces over which $|\mathbf{g}|$ is constant (not necessarily the equipotential surfaces, though, since $|\mathbf{g}| = |\boldsymbol{\nabla}\Psi|$), but it is less than obvious what they are. Obtaining $\mathbf{g}$ for a disc density distribution involves the use of Bessel functions and is beyond the scope of this book.

Note that this problem occurs when the potential arises from a self-gravitating matter distribution that is disc-like. There is no such problem for discs where almost all of the potential comes from a central point mass (such as for the primitive solar nebula).

For a discussion of gravitational potentials for a range of configurations see Chapter 2 in Binney and Tremaine, *Galactic Dynamics* (Princeton University Press, 1994).

3.4 The potential associated with a spherical mass distribution

Before leaving gravitational potentials in their own right, we will look at the potential associated with a spherical mass distribution in

Figure 1.43: pafd08.png

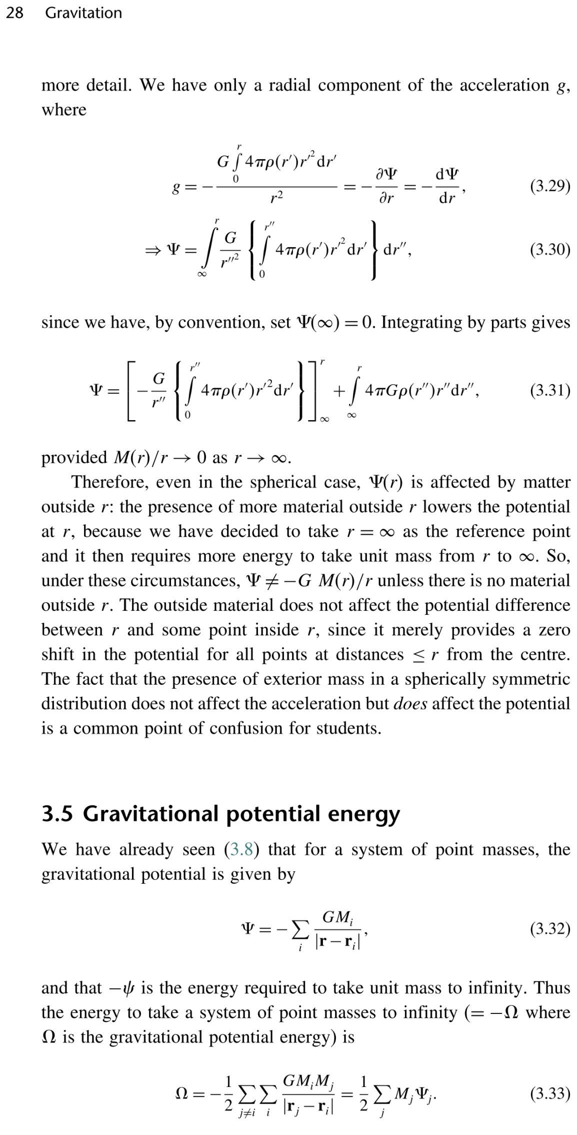

more detail. We have only a radial component of the acceleration g, where

$$g = -\frac{G\int_0^r 4\pi\rho(r')r'^2\,\mathrm{d}r'}{r^2} = -\frac{\partial\Psi}{\partial r} = -\frac{\mathrm{d}\Psi}{\mathrm{d}r}, \tag{3.29}$$

$$\Rightarrow \Psi = \int_\infty^r \frac{G}{r''^2}\left\{\int_0^{r''} 4\pi\rho(r')r'^2\,\mathrm{d}r'\right\}\mathrm{d}r'', \tag{3.30}$$

since we have, by convention, set $\Psi(\infty) = 0$. Integrating by parts gives

$$\Psi = \left[-\frac{G}{r''}\left\{\int_0^{r''} 4\pi\rho(r')r'^2\,\mathrm{d}r'\right\}\right]_\infty^r + \int_\infty^r 4\pi G\rho(r'')r''\,\mathrm{d}r'', \tag{3.31}$$

provided $M(r)/r \to 0$ as $r \to \infty$.

Therefore, even in the spherical case, $\Psi(r)$ is affected by matter outside r: the presence of more material outside r lowers the potential at r, because we have decided to take $r = \infty$ as the reference point and it then requires more energy to take unit mass from r to ∞. So, under these circumstances, $\Psi \neq -G\ M(r)/r$ unless there is no material outside r. The outside material does not affect the potential difference between r and some point inside r, since it merely provides a zero shift in the potential for all points at distances $\leq r$ from the centre. The fact that the presence of exterior mass in a spherically symmetric distribution does not affect the acceleration but *does* affect the potential is a common point of confusion for students.

3.5 Gravitational potential energy

We have already seen (3.8) that for a system of point masses, the gravitational potential is given by

$$\Psi = -\sum_i \frac{GM_i}{|\mathbf{r} - \mathbf{r}_i|}, \tag{3.32}$$

and that $-\psi$ is the energy required to take unit mass to infinity. Thus the energy to take a system of point masses to infinity ($= -\Omega$ where Ω is the gravitational potential energy) is

$$\Omega = -\frac{1}{2}\sum_{j\neq i}\sum_i \frac{GM_iM_j}{|\mathbf{r}_j - \mathbf{r}_i|} = \frac{1}{2}\sum_j M_j\Psi_j. \tag{3.33}$$

Figure 1.44: pafd09.png

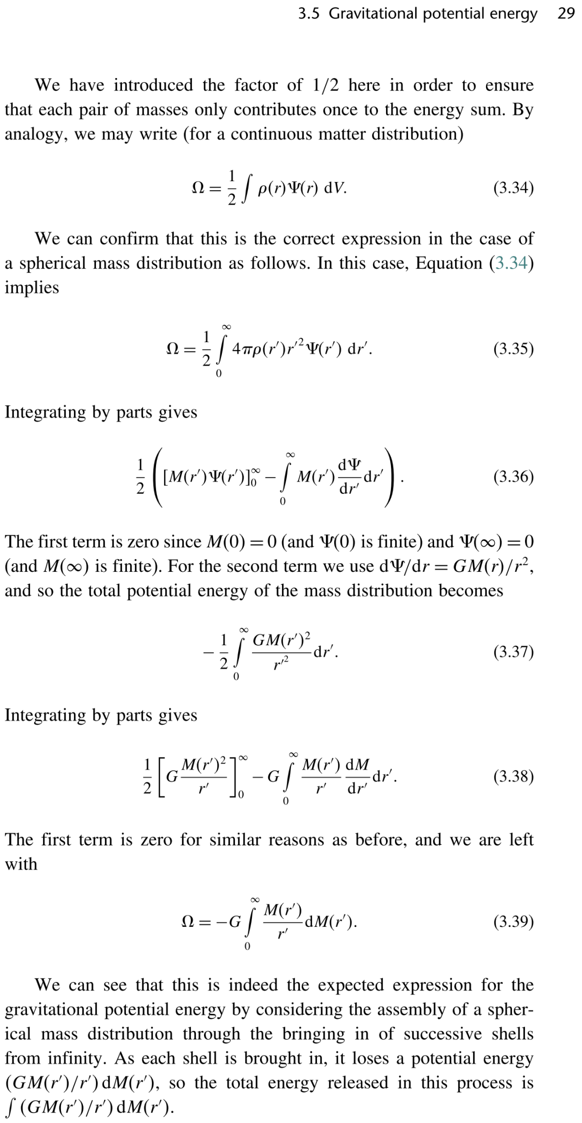

We have introduced the factor of 1/2 here in order to ensure that each pair of masses only contributes once to the energy sum. By analogy, we may write (for a continuous matter distribution)

$$\Omega = \frac{1}{2}\int \rho(r)\Psi(r)\ \mathrm{d}V. \tag{3.34}$$

We can confirm that this is the correct expression in the case of a spherical mass distribution as follows. In this case, Equation (3.34) implies

$$\Omega = \frac{1}{2}\int_0^\infty 4\pi\rho(r')r'^2\Psi(r')\ \mathrm{d}r'. \tag{3.35}$$

Integrating by parts gives

$$\frac{1}{2}\left([M(r')\Psi(r')]_0^\infty - \int_0^\infty M(r')\frac{\mathrm{d}\Psi}{\mathrm{d}r'}\mathrm{d}r' \right). \tag{3.36}$$

The first term is zero since $M(0) = 0$ (and $\Psi(0)$ is finite) and $\Psi(\infty) = 0$ (and $M(\infty)$ is finite). For the second term we use $\mathrm{d}\Psi/\mathrm{d}r = GM(r)/r^2$, and so the total potential energy of the mass distribution becomes

$$-\frac{1}{2}\int_0^\infty \frac{GM(r')^2}{r'^2}\mathrm{d}r'. \tag{3.37}$$

Integrating by parts gives

$$\frac{1}{2}\left[G\frac{M(r')^2}{r'}\right]_0^\infty - G\int_0^\infty \frac{M(r')}{r'}\frac{\mathrm{d}M}{\mathrm{d}r'}\mathrm{d}r'. \tag{3.38}$$

The first term is zero for similar reasons as before, and we are left with

$$\Omega = -G\int_0^\infty \frac{M(r')}{r'}\mathrm{d}M(r'). \tag{3.39}$$

We can see that this is indeed the expected expression for the gravitational potential energy by considering the assembly of a spherical mass distribution through the bringing in of successive shells from infinity. As each shell is brought in, it loses a potential energy $(GM(r')/r')\,\mathrm{d}M(r')$, so the total energy released in this process is $\int (GM(r')/r')\,\mathrm{d}M(r')$.

Figure 1.45: pafd10.png

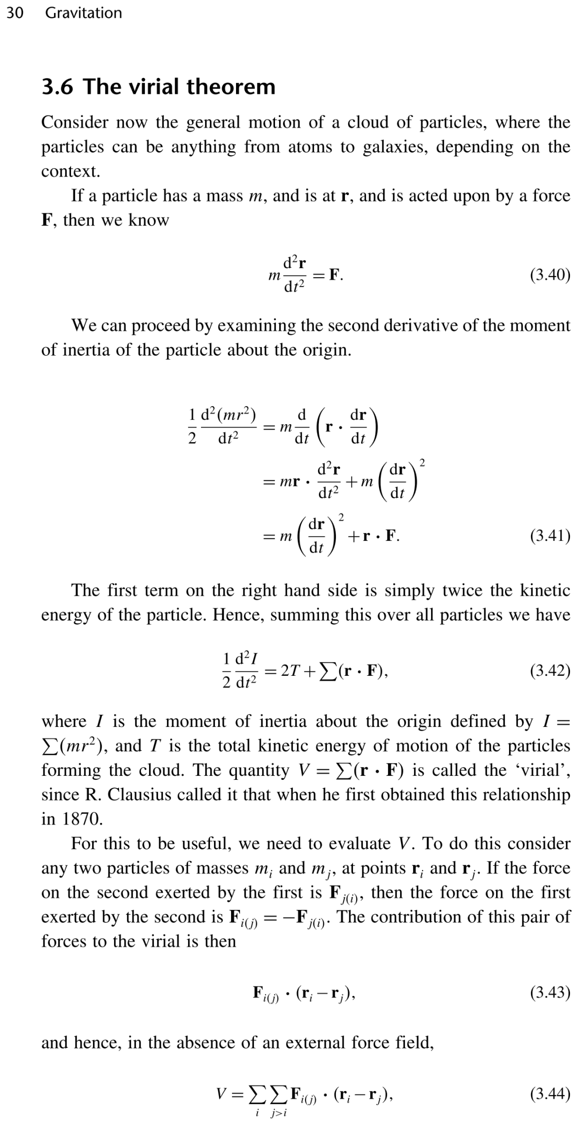

3.6 The virial theorem

Consider now the general motion of a cloud of particles, where the particles can be anything from atoms to galaxies, depending on the context.

If a particle has a mass m, and is at $\mathbf{r}$, and is acted upon by a force $\mathbf{F}$, then we know

$$m\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} = \mathbf{F}. \tag{3.40}$$

We can proceed by examining the second derivative of the moment of inertia of the particle about the origin.

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}^2(mr^2)}{\mathrm{d}t^2} &= m\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{r}\cdot\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\right)\\ &= m\mathbf{r}\cdot\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} + m\left(\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\right)^2\\ &= m\left(\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\right)^2 + \mathbf{r}\cdot\mathbf{F}.\end{aligned} \tag{3.41}$$

The first term on the right hand side is simply twice the kinetic energy of the particle. Hence, summing this over all particles we have

$$\frac{1}{2}\frac{\mathrm{d}^2 I}{\mathrm{d}t^2} = 2T + \sum(\mathbf{r}\cdot\mathbf{F}), \tag{3.42}$$

where I is the moment of inertia about the origin defined by $I = \sum(mr^2)$, and T is the total kinetic energy of motion of the particles forming the cloud. The quantity $V = \sum(\mathbf{r}\cdot\mathbf{F})$ is called the 'virial', since R. Clausius called it that when he first obtained this relationship in 1870.

For this to be useful, we need to evaluate V. To do this consider any two particles of masses m_i and m_j, at points $\mathbf{r}_i$ and $\mathbf{r}_j$. If the force on the second exerted by the first is $\mathbf{F}_{j(i)}$, then the force on the first exerted by the second is $\mathbf{F}_{i(j)} = -\mathbf{F}_{j(i)}$. The contribution of this pair of forces to the virial is then

$$\mathbf{F}_{i(j)}\cdot(\mathbf{r}_i - \mathbf{r}_j), \tag{3.43}$$

and hence, in the absence of an external force field,

$$V = \sum_i\sum_{j>i}\mathbf{F}_{i(j)}\cdot(\mathbf{r}_i - \mathbf{r}_j), \tag{3.44}$$

Figure 1.46: pafd11.png

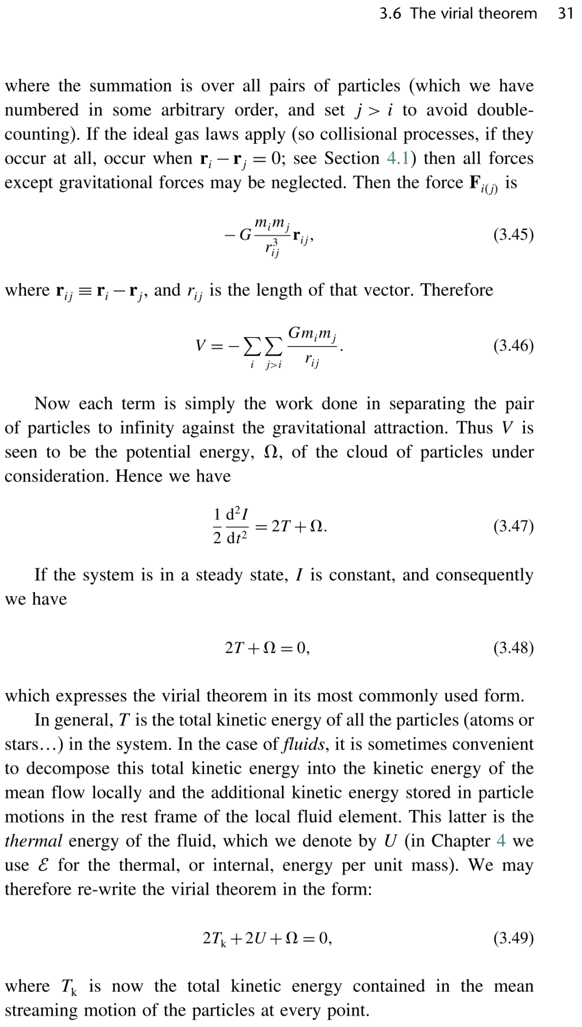

where the summation is over all pairs of particles (which we have numbered in some arbitrary order, and set $j > i$ to avoid double-counting). If the ideal gas laws apply (so collisional processes, if they occur at all, occur when $\mathbf{r}_i - \mathbf{r}_j = 0$; see Section 4.1) then all forces except gravitational forces may be neglected. Then the force $\mathbf{F}_{i(j)}$ is

$$-G\frac{m_i m_j}{r_{ij}^3}\mathbf{r}_{ij}, \qquad (3.45)$$

where $\mathbf{r}_{ij} \equiv \mathbf{r}_i - \mathbf{r}_j$, and r_{ij} is the length of that vector. Therefore

$$V = -\sum_i \sum_{j>i} \frac{Gm_i m_j}{r_{ij}}. \qquad (3.46)$$

Now each term is simply the work done in separating the pair of particles to infinity against the gravitational attraction. Thus V is seen to be the potential energy, Ω, of the cloud of particles under consideration. Hence we have

$$\frac{1}{2}\frac{\mathrm{d}^2 I}{\mathrm{d}t^2} = 2T + \Omega. \qquad (3.47)$$

If the system is in a steady state, I is constant, and consequently we have

$$2T + \Omega = 0, \qquad (3.48)$$

which expresses the virial theorem in its most commonly used form.

In general, T is the total kinetic energy of all the particles (atoms or stars…) in the system. In the case of *fluids*, it is sometimes convenient to decompose this total kinetic energy into the kinetic energy of the mean flow locally and the additional kinetic energy stored in particle motions in the rest frame of the local fluid element. This latter is the *thermal* energy of the fluid, which we denote by U (in Chapter 4 we use $\mathcal{E}$ for the thermal, or internal, energy per unit mass). We may therefore re-write the virial theorem in the form:

$$2T_{\mathrm{k}} + 2U + \Omega = 0, \qquad (3.49)$$

where T_{k} is now the total kinetic energy contained in the mean streaming motion of the particles at every point.

Figure 1.47: pafd12.png

4) C:/DATA-NovoVivo-127-2020/ Books-127-2020/ Power-of-Star/ 3319525956.pdf

5) C:/DATA-NovoVivo-127-2020/Books-127-2020/Multi-Variable Calculus A First Step/Multi-Variable Calculus A First Step.pdf

Pag. 27(40) - Figure 1.16: Distance from a point P to a plane ax + by + cz + d = 0.

Find a formula for the perpendicular distance D from the point P $(x_0; y_0; z_0))$ to the plane ax + by + cz + d = 0.

Solution. Let $P_1 (x_1; y_1; z_1)$ be any point in the given plane. Then

$$\overline{P_1P} = hx_0 \quad x_1; y_0 \quad y_1; z_0 \quad z_1 i$$

.

The vector $\overline{n} = ha; b; ci$ is a normal vector of the plane. Then, as shown in Figure 1.16, the distance D from P to the plane is

$$D = \; j \overline{P_1P} j \cos \quad :$$

$$D = k \overline{P_1P} \; j \cos \quad j = k \overline{P_1P} \; j \cos \quad j \frac{j \overline{n} j}{j \overline{n} j} = \frac{j P_1P \; j : \overline{n} j}{j \overline{n} j}$$

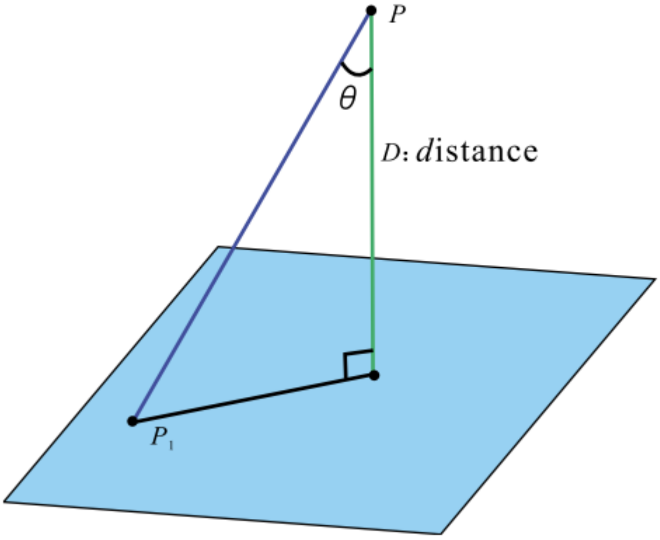

Figure 1.48: mv01.png

Thus,

$$D = \; j \overline{P_1P} j \cos \quad = \; j \overline{P_1P} j : \cos : \frac{j \overline{n} j}{j \overline{n} j} \tag{1.1}$$

$$= \frac{1}{|\overline{n}|}\,|\overline{P_1P}:\cos:|\overline{n}| = \frac{|\overline{P_1P}:\overline{n}|}{|\overline{n}|} \tag{1.2}$$

$$= \frac{|a(x_0 \quad x_1)+b(y_0 \quad y_1)+c(z_0 \quad z_1)|}{\sqrt{a^2+b^2+c^2}} \tag{1.3}$$

$$= \frac{|ax_0+by_0+cz_0 \quad (ax_1+by_1+cz_1)|}{\sqrt{a^2+b^2+c^2}} \tag{1.4}$$

$$= \frac{|ax_0+by_0+cz_0+d)|}{\sqrt{a^2+b^2+c^2}} \tag{1.5}$$

$$= d: \tag{1.6}$$

since $(ax_1+by_1+cz_1) = d:$

Versão diferente de apresentar as mesmas equações de cima.

$$D = |\overline{P_1P}|\cos = |\overline{P_1P}|:\cos:\frac{|\overline{n}|}{|\overline{n}|} \tag{1.7}$$

$$= \frac{1}{|\overline{n}|}\,|\overline{P_1P}:\cos:|\overline{n}| = \frac{|\overline{P_1P}:\overline{n}|}{|\overline{n}|} \tag{1.8}$$

$$= \frac{|a(x_0 \quad x_1)+b(y_0 \quad y_1)+c(z_0 \quad z_1)|}{\sqrt{a^2+b^2+c^2}} \tag{1.9}$$

$$= \frac{|ax_0+by_0+cz_0 \quad (ax_1+by_1+cz_1)|}{\sqrt{a^2+b^2+c^2}} \tag{1.10}$$

$$= \frac{|ax_0+by_0+cz_0+d)|}{\sqrt{a^2+b^2+c^2}} \tag{1.11}$$

$$= d: \tag{1.12}$$

6) C:/DATA-NovoVivo-127-2020/Books-127-2020/Applied Thermodynamics, 3rd Edition/Classical and geometrical theory of chemical and phase thermodynamics 2009.pdf 7) C:/DATA-NovoVivo-127-2020/Books-127-2020/Electromagnetic Fields and Waves Fundamentals of Engineering/Electromagnetic Fields and Waves Fundamentals of Engineering.pdf

1.2 Rotations with Unit Complex Numbers

1) C:/DATA-NovoVivo-129-2020/Books-129-2020/Physics from Symmetry/book-47331.pdf Pag. 31 (44) - 3.2.1 Rotations with Unit Complex Numbers

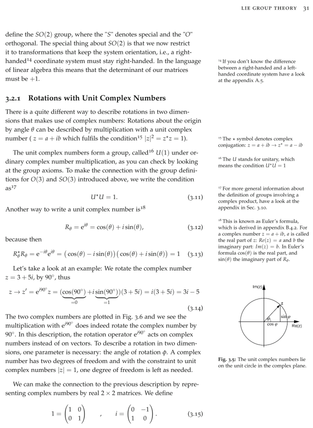

LIE GROUP THEORY 31

define the $SO(2)$ group, where the "S" denotes special and the "O" orthogonal. The special thing about $SO(2)$ is that we now restrict it to transformations that keep the system orientation, i.e., a right-handed[14] coordinate system must stay right-handed. In the language of linear algebra this means that the determinant of our matrices must be $+1$.

3.2.1 Rotations with Unit Complex Numbers

There is a quite different way to describe rotations in two dimensions that makes use of complex numbers: Rotations about the origin by angle θ can be described by multiplication with a unit complex number ($z = a + ib$ which fulfils the condition[15] $|z|^2 = z^\star z = 1$).

The unit complex numbers form a group, called[16] $U(1)$ under ordinary complex number multiplication, as you can check by looking at the group axioms. To make the connection with the group definitions for $O(3)$ and $SO(3)$ introduced above, we write the condition as[17]

$$U^\star U = 1. \tag{3.11}$$

Another way to write a unit complex number is[18]

$$R_\theta = e^{i\theta} = \cos(\theta) + i\sin(\theta), \tag{3.12}$$

because then

$$R_\theta^\star R_\theta = e^{-i\theta}e^{i\theta} = \big(\cos(\theta) - i\sin(\theta)\big)\big(\cos(\theta) + i\sin(\theta)\big) = 1 \tag{3.13}$$

Let's take a look at an example: We rotate the complex number $z = 3 + 5i$, by 90°, thus

$$z \to z' = e^{i90^\circ} z = (\underbrace{\cos(90^\circ)}_{=0} + i\underbrace{\sin(90^\circ)}_{=1})(3+5i) = i(3+5i) = 3i - 5 \tag{3.14}$$

The two complex numbers are plotted in Fig. 3.6 and we see the multiplication with e^{i90° does indeed rotate the complex number by 90°. In this description, the rotation operator e^{i90° acts on complex numbers instead of on vectors. To describe a rotation in two dimensions, one parameter is necessary: the angle of rotation ϕ. A complex number has two degrees of freedom and with the constraint to unit complex numbers $|z| = 1$, one degree of freedom is left as needed.

We can make the connection to the previous description by representing complex numbers by real 2×2 matrices. We define

$$1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad , \quad i = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{3.15}$$

Fig. 3.5: The unit complex numbers lie on the unit circle in the complex plane.

[14] If you don't know the difference between a right-handed and a left-handed coordinate system have a look at the appendix A.5.

[15] The $\star$ symbol denotes complex conjugation: $z = a + ib \to z^\star = a - ib$

[16] The U stands for unitary, which means the condition $U^\star U = 1$

[17] For more general information about the definition of groups involving a complex product, have a look at the appendix in Sec. 3.10.

[18] This is known as Euler's formula, which is derived in appendix B.4.2. For a complex number $z = a + ib$, a is called the real part of z: $Re(z) = a$ and b the imaginary part: $Im(z) = b$. In Euler's formula $\cos(\theta)$ is the real part, and $\sin(\theta)$ the imaginary part of R_θ.

Figure 1.49: ps01.png

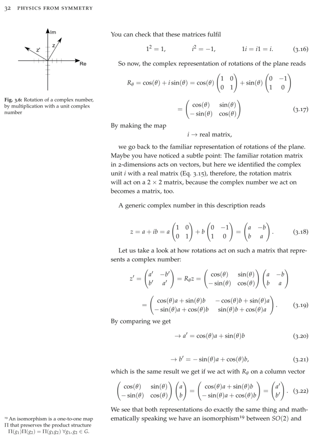

32 PHYSICS FROM SYMMETRY

You can check that these matrices fulfil

$$1^2 = 1, \qquad i^2 = -1, \qquad 1i = i1 = i. \tag{3.16}$$

So now, the complex representation of rotations of the plane reads

$$R_\theta = \cos(\theta) + i\sin(\theta) = \cos(\theta)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \sin(\theta)\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \tag{3.17}$$

By making the map

$$i \to \text{real matrix},$$

we go back to the familiar representation of rotations of the plane. Maybe you have noticed a subtle point: The familiar rotation matrix in 2-dimensions acts on vectors, but here we identified the complex unit i with a real matrix (Eq. 3.15), therefore, the rotation matrix will act on a 2×2 matrix, because the complex number we act on becomes a matrix, too.

A generic complex number in this description reads

$$z = a + ib = a\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + b\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}. \tag{3.18}$$

Let us take a look at how rotations act on such a matrix that represents a complex number:

$$z' = \begin{pmatrix} a' & -b' \\ b' & a' \end{pmatrix} = R_\theta z = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$$

$$= \begin{pmatrix} \cos(\theta)a + \sin(\theta)b & -\cos(\theta)b + \sin(\theta)a \\ -\sin(\theta)a + \cos(\theta)b & \sin(\theta)b + \cos(\theta)a \end{pmatrix}. \tag{3.19}$$

By comparing we get

$$\to a' = \cos(\theta)a + \sin(\theta)b \tag{3.20}$$

$$\to b' = -\sin(\theta)a + \cos(\theta)b, \tag{3.21}$$

which is the same result we get if we act with R_θ on a column vector

$$\begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \cos(\theta)a + \sin(\theta)b \\ -\sin(\theta)a + \cos(\theta)b \end{pmatrix} = \begin{pmatrix} a' \\ b' \end{pmatrix}. \tag{3.22}$$

We see that both representations do exactly the same thing and mathematically speaking we have an isomorphism[19] between $SO(2)$ and

Fig. 3.6: Rotation of a complex number, by multiplication with a unit complex number

[19] An isomorphism is a one-to-one map Π that preserves the product structure $\Pi(g_1)\Pi(g_2) = \Pi(g_1 g_2)\ \forall g_1, g_2 \in G$.

Figure 1.50: ps02.png

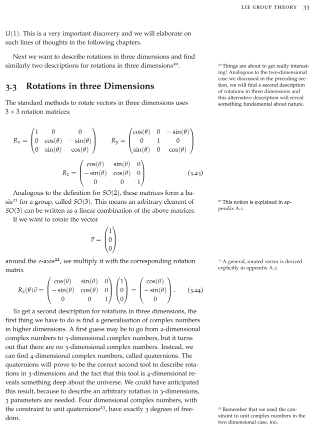

LIE GROUP THEORY 33

$U(1)$. This is a very important discovery and we will elaborate on such lines of thoughts in the following chapters.

Next we want to describe rotations in three dimensions and find similarly two descriptions for rotations in three dimensions[20].

3.3 Rotations in three Dimensions

The standard methods to rotate vectors in three dimensions uses 3×3 rotation matrices:

$$R_x = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & -\sin(\theta) \\ 0 & \sin(\theta) & \cos(\theta) \end{pmatrix} \qquad R_y = \begin{pmatrix} \cos(\theta) & 0 & -\sin(\theta) \\ 0 & 1 & 0 \\ \sin(\theta) & 0 & \cos(\theta) \end{pmatrix}$$

$$R_z = \begin{pmatrix} \cos(\theta) & \sin(\theta) & 0 \\ -\sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{3.23}$$

Analogous to the definition for $SO(2)$, these matrices form a basis[21] for a group, called $SO(3)$. This means an arbitrary element of $SO(3)$ can be written as a linear combination of the above matrices.

If we want to rotate the vector

$$\vec{v} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

around the z-axis[22], we multiply it with the corresponding rotation matrix

$$R_z(\theta)\vec{v} = \begin{pmatrix} \cos(\theta) & \sin(\theta) & 0 \\ -\sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \cos(\theta) \\ -\sin(\theta) \\ 0 \end{pmatrix}. \tag{3.24}$$

To get a second description for rotations in three dimensions, the first thing we have to do is find a generalisation of complex numbers in higher dimensions. A first guess may be to go from 2-dimensional complex numbers to 3-dimensional complex numbers, but it turns out that there are no 3-dimensional complex numbers. Instead, we can find 4-dimensional complex numbers, called quaternions. The quaternions will prove to be the correct second tool to describe rotations in 3-dimensions and the fact that this tool is 4-dimensional reveals something deep about the universe. We could have anticipated this result, because to describe an arbitrary rotation in 3-dimensions, 3 parameters are needed. Four dimensional complex numbers, with the constraint to unit quaternions[23], have exactly 3 degrees of freedom.

[20] Things are about to get really interesting! Analogous to the two-dimensional case we discussed in the preceding section, we will find a second description of rotations in three dimensions and this alternative description will reveal something fundamental about nature.

[21] This notion is explained in appendix A.1.

[22] A general, rotated vector is derived explicitly in appendix A.2.

[23] Remember that we used the constraint to unit complex numbers in the two dimensional case, too.

Figure 1.51: ps03.png

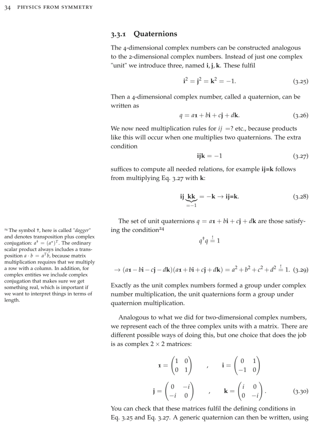

34 PHYSICS FROM SYMMETRY

3.3.1 Quaternions

The 4-dimensional complex numbers can be constructed analogous to the 2-dimensional complex numbers. Instead of just one complex "unit" we introduce three, named $\mathbf{i}, \mathbf{j}, \mathbf{k}$. These fulfil

$$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -1. \tag{3.25}$$

Then a 4-dimensional complex number, called a quaternion, can be written as

$$q = a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}. \tag{3.26}$$

We now need multiplication rules for ij =? etc., because products like this will occur when one multiplies two quaternions. The extra condition

$$\mathbf{ijk} = -1 \tag{3.27}$$

suffices to compute all needed relations, for example **ij=k** follows from multiplying Eq. 3.27 with **k**:

$$\mathbf{ij}\underbrace{\mathbf{kk}}_{=-1} = -\mathbf{k} \to \mathbf{ij} = \mathbf{k}. \tag{3.28}$$

The set of unit quaternions $q = a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$ are those satisfying the condition[24]

$$q^\dagger q \overset{!}{=} 1$$

$$\to (a\mathbf{1} - b\mathbf{i} - c\mathbf{j} - d\mathbf{k})(a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}) = a^2 + b^2 + c^2 + d^2 \overset{!}{=} 1. \tag{3.29}$$

Exactly as the unit complex numbers formed a group under complex number multiplication, the unit quaternions form a group under quaternion multiplication.

Analogous to what we did for two-dimensional complex numbers, we represent each of the three complex units with a matrix. There are different possible ways of doing this, but one choice that does the job is as complex 2×2 matrices:

$$\mathbf{1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad , \quad \mathbf{i} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

$$\mathbf{j} = \begin{pmatrix} 0 & -i \\ -i & 0 \end{pmatrix} \quad , \quad \mathbf{k} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}. \tag{3.30}$$

You can check that these matrices fulfil the defining conditions in Eq. 3.25 and Eq. 3.27. A generic quaternion can then be written, using

[24] The symbol †, here is called "*dagger*" and denotes transposition plus complex conjugation: $a^\dagger = (a^\star)^T$. The ordinary scalar product always includes a transposition $a \cdot b = a^T b$, because matrix multiplication requires that we multiply a row with a column. In addition, for complex entities we include complex conjugation that makes sure we get something real, which is important if we want to interpret things in terms of length.

Figure 1.52: ps04.png

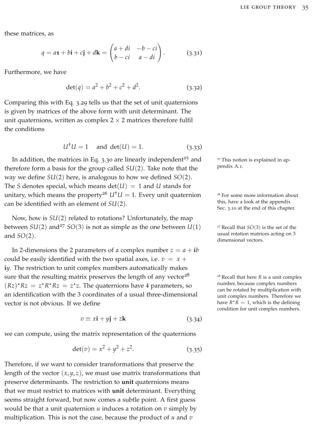

LIE GROUP THEORY 35

these matrices, as

$$q = a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k} = \begin{pmatrix} a+di & -b-ci \\ b-ci & a-di \end{pmatrix}. \tag{3.31}$$

Furthermore, we have

$$\det(q) = a^2 + b^2 + c^2 + d^2. \tag{3.32}$$

Comparing this with Eq. 3.29 tells us that the set of unit quaternions is given by matrices of the above form with unit determinant. The unit quaternions, written as complex 2×2 matrices therefore fulfil the conditions

$$U^\dagger U = 1 \quad \text{and } \det(U) = 1. \tag{3.33}$$

In addition, the matrices in Eq. 3.30 are linearly independent[25] and therefore form a basis for the group called $SU(2)$. Take note that the way we define $SU(2)$ here, is analogous to how we defined $SO(2)$. The S denotes special, which means $\det(U) = 1$ and U stands for unitary, which means the property[26] $U^\dagger U = 1$. Every unit quaternion can be identified with an element of $SU(2)$.

Now, how is $SU(2)$ related to rotations? Unfortunately, the map between $SU(2)$ and[27] $SO(3)$ is not as simple as the one between $U(1)$ and $SO(2)$.

In 2-dimensions the 2 parameters of a complex number $z = a + \mathbf{i}b$ could be easily identified with the two spatial axes, i.e. $v = x + \mathbf{i}y$. The restriction to unit complex numbers automatically makes sure that the resulting matrix preserves the length of any vector[28] $(Rz)^\star Rz = z^\star R^\star Rz = z^\star z$. The quaternions have 4 parameters, so an identification with the 3 coordinates of a usual three-dimensional vector is not obvious. If we define

$$v \equiv x\mathbf{i} + y\mathbf{j} + z\mathbf{k} \tag{3.34}$$

we can compute, using the matrix representation of the quaternions

$$\det(v) = x^2 + y^2 + z^2. \tag{3.35}$$

Therefore, if we want to consider transformations that preserve the length of the vector (x, y, z), we must use matrix transformations that preserve determinants. The restriction to **unit** quaternions means that we must restrict to matrices with **unit** determinant. Everything seems straight forward, but now comes a subtle point. A first guess would be that a unit quaternion u induces a rotation on v simply by multiplication. This is not the case, because the product of u and v

[25] This notion is explained in appendix A.1.

[26] For some more information about this, have a look at the appendix Sec. 3.10 at the end of this chapter.

[27] Recall that $SO(3)$ is the set of the usual rotation matrices acting on 3 dimensional vectors.

[28] Recall that here R is a unit complex number, because complex numbers can be rotated by multiplication with unit complex numbers. Therefore we have $R^\star R = 1$, which is the defining condition for unit complex numbers.

Figure 1.53: ps05.png

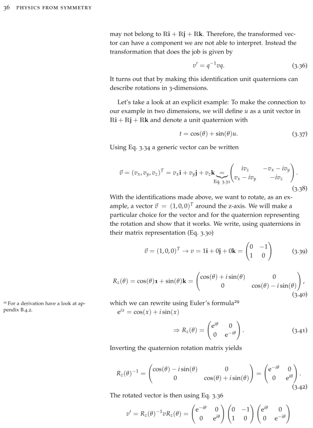

36 PHYSICS FROM SYMMETRY

may not belong to $\mathbb{R}\mathbf{i} + \mathbb{R}\mathbf{j} + \mathbb{R}\mathbf{k}$. Therefore, the transformed vector can have a component we are not able to interpret. Instead the transformation that does the job is given by

$$v' = q^{-1}vq. \tag{3.36}$$

It turns out that by making this identification unit quaternions can describe rotations in 3-dimensions.

Let's take a look at an explicit example: To make the connection to our example in two dimensions, we will define u as a unit vector in $\mathbb{R}\mathbf{i} + \mathbb{R}\mathbf{j} + \mathbb{R}\mathbf{k}$ and denote a unit quaternion with

$$t = \cos(\theta) + \sin(\theta)u. \tag{3.37}$$

Using Eq. 3.34 a generic vector can be written

$$\vec{v} = (v_x, v_y, v_z)^T = v_x\mathbf{i} + v_y\mathbf{j} + v_z\mathbf{k} \underbrace{=}_{\text{Eq. 3.31}} \begin{pmatrix} iv_z & -v_x - iv_y \\ v_x - iv_y & -iv_z \end{pmatrix}. \tag{3.38}$$

With the identifications made above, we want to rotate, as an example, a vector $\vec{v} = (1,0,0)^T$ around the z-axis. We will make a particular choice for the vector and for the quaternion representing the rotation and show that it works. We write, using quaternions in their matrix representation (Eq. 3.30)

$$\vec{v} = (1,0,0)^T \to v = 1\mathbf{i} + 0\mathbf{j} + 0\mathbf{k} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \tag{3.39}$$

$$R_z(\theta) = \cos(\theta)\mathbf{1} + \sin(\theta)\mathbf{k} = \begin{pmatrix} \cos(\theta) + i\sin(\theta) & 0 \\ 0 & \cos(\theta) - i\sin(\theta) \end{pmatrix}, \tag{3.40}$$

which we can rewrite using Euler's formula[29]
$e^{ix} = \cos(x) + i\sin(x)$

$$\Rightarrow R_z(\theta) = \begin{pmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{pmatrix}. \tag{3.41}$$

Inverting the quaternion rotation matrix yields

$$R_z(\theta)^{-1} = \begin{pmatrix} \cos(\theta) - i\sin(\theta) & 0 \\ 0 & \cos(\theta) + i\sin(\theta) \end{pmatrix} = \begin{pmatrix} e^{-i\theta} & 0 \\ 0 & e^{i\theta} \end{pmatrix}. \tag{3.42}$$

The rotated vector is then using Eq. 3.36

$$v' = R_z(\theta)^{-1}vR_z(\theta) = \begin{pmatrix} e^{-i\theta} & 0 \\ 0 & e^{i\theta} \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{pmatrix}$$

[29] For a derivation have a look at appendix B.4.2.

Figure 1.54: ps06.png

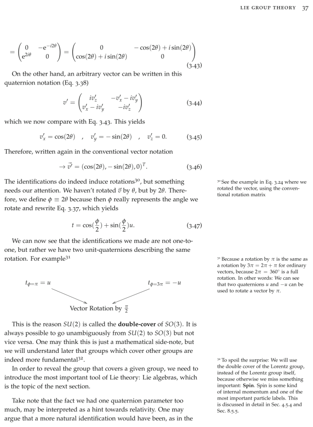

LIE GROUP THEORY 37

$$= \begin{pmatrix} 0 & -e^{-i2\theta} \\ e^{2i\theta} & 0 \end{pmatrix} = \begin{pmatrix} 0 & -\cos(2\theta) + i\sin(2\theta) \\ \cos(2\theta) + i\sin(2\theta) & 0 \end{pmatrix} \tag{3.43}$$

On the other hand, an arbitrary vector can be written in this quaternion notation (Eq. 3.38)

$$v' = \begin{pmatrix} iv'_z & -v'_x - iv'_y \\ v'_x - iv'_y & -iv'_z \end{pmatrix} \tag{3.44}$$

which we now compare with Eq. 3.43. This yields

$$v'_x = \cos(2\theta) \quad , \quad v'_y = -\sin(2\theta) \quad , \quad v'_z = 0. \tag{3.45}$$

Therefore, written again in the conventional vector notation

$$\to \vec{v'} = (\cos(2\theta), -\sin(2\theta), 0)^T. \tag{3.46}$$

The identifications do indeed induce rotations[30], but something needs our attention. We haven't rotated $\vec{v}$ by θ, but by 2θ. Therefore, we define $\phi \equiv 2\theta$ because then ϕ really represents the angle we rotate and rewrite Eq. 3.37, which yields

$$t = \cos(\frac{\phi}{2}) + \sin(\frac{\phi}{2})u. \tag{3.47}$$

We can now see that the identifications we made are not one-to-one, but rather we have two unit-quaternions describing the same rotation. For example[31]

$$\begin{tikzcd}
t_{\phi=\pi} = u \arrow[dr] & & t_{\phi=3\pi} = -u \arrow[dl] \\
& \text{Vector Rotation by } \frac{\pi}{2} &
\end{tikzcd}$$

This is the reason $SU(2)$ is called the **double-cover** of $SO(3)$. It is always possible to go unambiguously from $SU(2)$ to $SO(3)$ but not vice versa. One may think this is just a mathematical side-note, but we will understand later that groups which cover other groups are indeed more fundamental[32].

In order to reveal the group that covers a given group, we need to introduce the most important tool of Lie theory: Lie algebras, which is the topic of the next section.

Take note that the fact we had one quaternion parameter too much, may be interpreted as a hint towards relativity. One may argue that a more natural identification would have been, as in the

[30] See the example in Eq. 3.24 where we rotated the vector, using the conventional rotation matrix

[31] Because a rotation by π is the same as a rotation by $3\pi = 2\pi + \pi$ for ordinary vectors, because $2\pi = 360°$ is a full rotation. In other words: We can see that two quaternions u and $-u$ can be used to rotate a vector by π.

[32] To spoil the surprise: We will use the double cover of the Lorentz group, instead of the Lorentz group itself, because otherwise we miss something important: **Spin**. Spin is some kind of internal momentum and one of the most important particle labels. This is discussed in detail in Sec. 4.5.4 and Sec. 8.5.5.

Figure 1.55: ps07.png

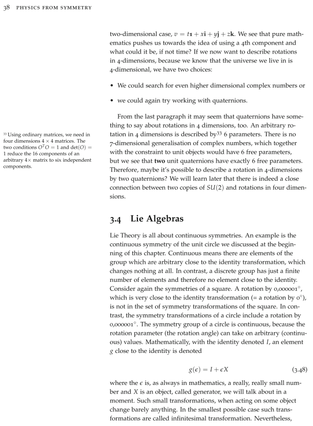

38 PHYSICS FROM SYMMETRY

two-dimensional case, $v = t\mathbf{1} + x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$. We see that pure mathematics pushes us towards the idea of using a 4th component and what could it be, if not time? If we now want to describe rotations in 4-dimensions, because we know that the universe we live in is 4-dimensional, we have two choices:

- We could search for even higher dimensional complex numbers or
- we could again try working with quaternions.

From the last paragraph it may seem that quaternions have something to say about rotations in 4 dimensions, too. An arbitrary rotation in 4 dimensions is described by[33] 6 parameters. There is no 7-dimensional generalisation of complex numbers, which together with the constraint to unit objects would have 6 free parameters, but we see that **two** unit quaternions have exactly 6 free parameters. Therefore, maybe it's possible to describe a rotation in 4-dimensions by two quaternions? We will learn later that there is indeed a close connection between two copies of $SU(2)$ and rotations in four dimensions.

3.4 Lie Algebras

Lie Theory is all about continuous symmetries. An example is the continuous symmetry of the unit circle we discussed at the beginning of this chapter. Continuous means there are elements of the group which are arbitrary close to the identity transformation, which changes nothing at all. In contrast, a discrete group has just a finite number of elements and therefore no element close to the identity. Consider again the symmetries of a square. A rotation by 0,000001°, which is very close to the identity transformation (= a rotation by 0°), is not in the set of symmetry transformations of the square. In contrast, the symmetry transformations of a circle include a rotation by 0,000001°. The symmetry group of a circle is continuous, because the rotation parameter (the rotation angle) can take on arbitrary (continuous) values. Mathematically, with the identity denoted I, an element g close to the identity is denoted

$$g(\epsilon) = I + \epsilon X \tag{3.48}$$

where the ϵ is, as always in mathematics, a really, really small number and X is an object, called generator, we will talk about in a moment. Such small transformations, when acting on some object change barely anything. In the smallest possible case such transformations are called infinitesimal transformation. Nevertheless,

[33] Using ordinary matrices, we need in four dimensions 4×4 matrices. The two conditions $O^TO = 1$ and $\det(O) = 1$ reduce the 16 components of an arbitrary $4\times$ matrix to six independent components.

Figure 1.56: ps08.png

2) C:/DATA-NovoVivo-128-2020/Books-128-2020/Electromechanical Energy Conversion/Electromechanical-Energy-Conversion.pdf Pag. 148 (167) - 4.4 CYLINDRICAL MACHINES

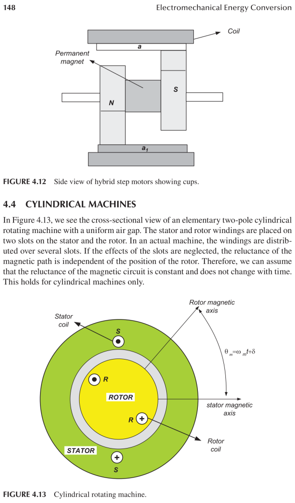

FIGURE 4.12 Side view of hybrid step motors showing cups.

4.4 CYLINDRICAL MACHINES

In Figure 4.13, we see the cross-sectional view of an elementary two-pole cylindrical rotating machine with a uniform air gap. The stator and rotor windings are placed on two slots on the stator and the rotor. In an actual machine, the windings are distributed over several slots. If the effects of the slots are neglected, the reluctance of the magnetic path is independent of the position of the rotor. Therefore, we can assume that the reluctance of the magnetic circuit is constant and does not change with time. This holds for cylindrical machines only.

FIGURE 4.13 Cylindrical rotating machine.

Figure 1.57: eec01.png

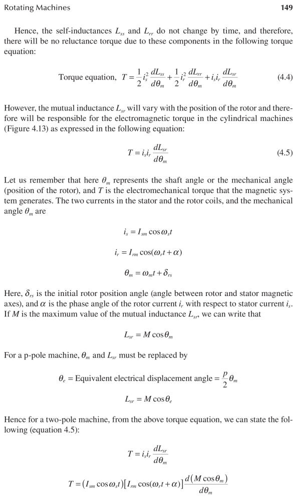

Hence, the self-inductances L_{ss} and L_{rr} do not change by time, and therefore, there will be no reluctance torque due to these components in the following torque equation:

$$\text{Torque equation,}\quad T=\frac{1}{2}i_s^2\frac{dL_{ss}}{d\theta_m}+\frac{1}{2}i_r^2\frac{dL_{rr}}{d\theta_m}+i_si_r\frac{dL_{sr}}{d\theta_m} \tag{4.4}$$

However, the mutual inductance L_{sr} will vary with the position of the rotor and therefore will be responsible for the electromagnetic torque in the cylindrical machines (Figure 4.13) as expressed in the following equation:

$$T=i_si_r\frac{dL_{sr}}{d\theta_m} \tag{4.5}$$

Let us remember that here θ_m represents the shaft angle or the mechanical angle (position of the rotor), and T is the electromechanical torque that the magnetic system generates. The two currents in the stator and the rotor coils, and the mechanical angle θ_m are

$$i_s=I_{sm}\cos\omega_st$$

$$i_r=I_{rm}\cos(\omega_rt+\alpha)$$

$$\theta_m=\omega_mt+\delta_{rs}$$

Here, δ_{rs} is the initial rotor position angle (angle between rotor and stator magnetic axes), and α is the phase angle of the rotor current i_r with respect to stator current i_s. If M is the maximum value of the mutual inductance L_{sr}, we can write that

$$L_{sr}=M\cos\theta_m$$

For a p-pole machine, θ_m and L_{sr} must be replaced by

$$\theta_e=\text{Equivalent electrical displacement angle}=\frac{p}{2}\theta_m$$

$$L_{sr}=M\cos\theta_e$$

Hence for a two-pole machine, from the above torque equation, we can state the following (equation 4.5):

$$T=i_si_r\frac{dL_{sr}}{d\theta_m}$$

$$T=(I_{sm}\cos\omega_st)[I_{rm}\cos(\omega_rt+\alpha)]\frac{d(M\cos\theta_m)}{d\theta_m}$$

Figure 1.58: eec02.png

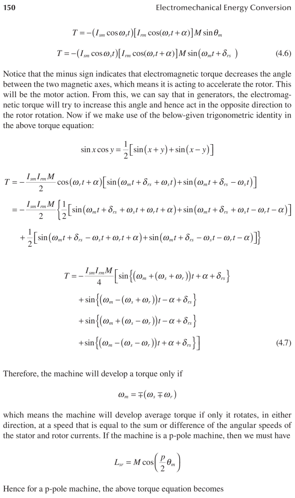

$$T=-(I_{sm}\cos\omega_st)[I_{rm}\cos(\omega_rt+\alpha)]M\sin\theta_m$$

$$T=-(I_{sm}\cos\omega_st)[I_{rm}\cos(\omega_rt+\alpha)]M\sin(\omega_mt+\delta_{rs}\ ) \tag{4.6}$$

Notice that the minus sign indicates that electromagnetic torque decreases the angle between the two magnetic axes, which means it is acting to accelerate the rotor. This will be the motor action. From this, we can say that in generators, the electromagnetic torque will try to increase this angle and hence act in the opposite direction to the rotor rotation. Now if we make use of the below-given trigonometric identity in the above torque equation:

$$\sin x\cos y=\frac{1}{2}\left[\sin(x+y)+\sin(x-y)\right]$$

$$\begin{aligned}T&=-\frac{I_{sm}I_{rm}M}{2}\cos(\omega_rt+\alpha)\left[\sin(\omega_mt+\delta_{rs}+\omega_st)+\sin(\omega_mt+\delta_{rs}-\omega_st)\right]\\&=-\frac{I_{sm}I_{rm}M}{2}\left\{\frac{1}{2}\left[\sin(\omega_mt+\delta_{rs}+\omega_st+\omega_rt+\alpha)+\sin(\omega_mt+\delta_{rs}+\omega_st-\omega_rt-\alpha)\right]\right.\\&\quad\left.+\frac{1}{2}\left[\sin(\omega_mt+\delta_{rs}-\omega_st+\omega_rt+\alpha)+\sin(\omega_mt+\delta_{rs}-\omega_st-\omega_rt-\alpha)\right]\right\}\end{aligned}$$

$$\begin{aligned}T=-\frac{I_{sm}I_{rm}M}{4}\Big[&\sin\{(\omega_m+(\omega_s+\omega_r))t+\alpha+\delta_{rs}\}\\&+\sin\{(\omega_m-(\omega_s+\omega_r))t-\alpha+\delta_{rs}\}\\&+\sin\{(\omega_m+(\omega_s-\omega_r))t-\alpha+\delta_{rs}\}\\&+\sin\{(\omega_m-(\omega_s-\omega_r))t+\alpha+\delta_{rs}\}\Big]\end{aligned} \tag{4.7}$$

Therefore, the machine will develop a torque only if

$$\omega_m=\mp(\omega_s\mp\omega_r)$$

which means the machine will develop average torque if only it rotates, in either direction, at a speed that is equal to the sum or difference of the angular speeds of the stator and rotor currents. If the machine is a p-pole machine, then we must have

$$L_{sr}=M\cos\left(\frac{p}{2}\theta_m\right)$$

Hence for a p-pole machine, the above torque equation becomes

Figure 1.59: eec03.png

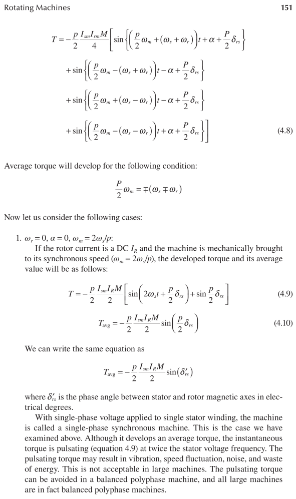

$$\begin{aligned}T=-\frac{p}{2}\frac{I_{sm}I_{rm}M}{4}\Bigg[&\sin\left\{\left(\frac{p}{2}\omega_m+(\omega_s+\omega_r)\right)t+\alpha+\frac{P}{2}\delta_{rs}\right\}\\&+\sin\left\{\left(\frac{p}{2}\omega_m-(\omega_s+\omega_r)\right)t-\alpha+\frac{P}{2}\delta_{rs}\right\}\\&+\sin\left\{\left(\frac{p}{2}\omega_m+(\omega_s-\omega_r)\right)t-\alpha+\frac{P}{2}\delta_{rs}\right\}\\&+\sin\left\{\left(\frac{p}{2}\omega_m-(\omega_s-\omega_r)\right)t+\alpha+\frac{P}{2}\delta_{rs}\right\}\Bigg]\end{aligned} \tag{4.8}$$

Average torque will develop for the following condition:

$$\frac{P}{2}\omega_m=\mp(\omega_s\mp\omega_r)$$

Now let us consider the following cases:

1. $\omega_r=0$, $\alpha=0$, $\omega_m=2\omega_s/p$:
   If the rotor current is a DC I_R and the machine is mechanically brought to its synchronous speed ($\omega_m=2\omega_s/p$), the developed torque and its average value will be as follows:

   $$T=-\frac{p}{2}\frac{I_{sm}I_RM}{2}\left[\sin\left(2\omega_st+\frac{p}{2}\delta_{rs}\right)+\sin\frac{p}{2}\delta_{rs}\right] \tag{4.9}$$

   $$T_{\text{avg}}=-\frac{p}{2}\frac{I_{sm}I_RM}{2}\sin\left(\frac{p}{2}\delta_{rs}\right) \tag{4.10}$$

   We can write the same equation as

   $$T_{\text{avg}}=-\frac{p}{2}\frac{I_{sm}I_RM}{2}\sin(\delta'_{rs})$$

   where δ'_{rs} is the phase angle between stator and rotor magnetic axes in electrical degrees.

   With single-phase voltage applied to single stator winding, the machine is called a single-phase synchronous machine. This is the case we have examined above. Although it develops an average torque, the instantaneous torque is pulsating (equation 4.9) at twice the stator voltage frequency. The pulsating torque may result in vibration, speed fluctuation, noise, and waste of energy. This is not acceptable in large machines. The pulsating torque can be avoided in a balanced polyphase machine, and all large machines are in fact balanced polyphase machines.

Figure 1.60: eec04.png

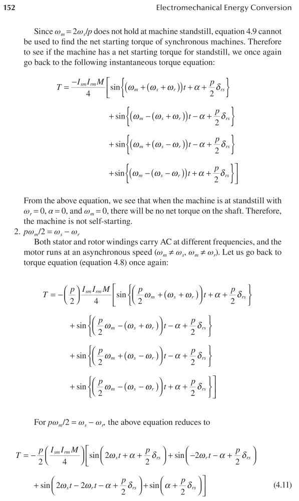

152 Electromechanical Energy Conversion

Since $\omega_m = 2\omega_s/p$ does not hold at machine standstill, equation 4.9 cannot be used to find the net starting torque of synchronous machines. Therefore to see if the machine has a net starting torque for standstill, we once again go back to the following instantaneous torque equation:

$$T = \frac{-I_{sm}I_{rm}M}{4}\left[\sin\left\{(\omega_m + (\omega_s + \omega_r))t + \alpha + \frac{p}{2}\delta_{rs}\right\}\right.$$
$$+ \sin\left\{(\omega_m - (\omega_s + \omega_r))t - \alpha + \frac{p}{2}\delta_{rs}\right\}$$
$$+ \sin\left\{(\omega_m + (\omega_s - \omega_r))t - \alpha + \frac{p}{2}\delta_{rs}\right\}$$
$$\left.+ \sin\left\{(\omega_m - (\omega_s - \omega_r))t + \alpha + \frac{p}{2}\delta_{rs}\right\}\right]$$

From the above equation, we see that when the machine is at standstill with $\omega_r = 0$, $\alpha = 0$, and $\omega_m = 0$, there will be no net torque on the shaft. Therefore, the machine is not self-starting.

2. $p\omega_m/2 = \omega_s - \omega_r$

Both stator and rotor windings carry AC at different frequencies, and the motor runs at an asynchronous speed ($\omega_m \neq \omega_s$, $\omega_m \neq \omega_r$). Let us go back to torque equation (equation 4.8) once again:

$$T = -\left(\frac{p}{2}\right)\frac{I_{sm}I_{rm}M}{4}\left[\sin\left\{\left(\frac{p}{2}\omega_m + (\omega_s + \omega_r)\right)t + \alpha + \frac{p}{2}\delta_{rs}\right\}\right.$$
$$+ \sin\left\{\left(\frac{p}{2}\omega_m - (\omega_s + \omega_r)\right)t - \alpha + \frac{p}{2}\delta_{rs}\right\}$$
$$+ \sin\left\{\left(\frac{p}{2}\omega_m + (\omega_s - \omega_r)\right)t - \alpha + \frac{p}{2}\delta_{rs}\right\}$$
$$\left.+ \sin\left\{\left(\frac{p}{2}\omega_m - (\omega_s - \omega_r)\right)t + \alpha + \frac{p}{2}\delta_{rs}\right\}\right]$$

For $p\omega_m/2 = \omega_s - \omega_r$, the above equation reduces to

$$T = -\frac{p}{2}\left(\frac{I_{sm}I_{rm}M}{4}\right)\left[\sin\left(2\omega_s t + \alpha + \frac{p}{2}\delta_{rs}\right) + \sin\left(-2\omega_r t - \alpha + \frac{p}{2}\delta_{rs}\right)\right.$$
$$\left.+ \sin\left(2\omega_s t - 2\omega_r t - \alpha + \frac{p}{2}\delta_{rs}\right) + \sin\left(\alpha + \frac{p}{2}\delta_{rs}\right)\right] \tag{4.11}$$

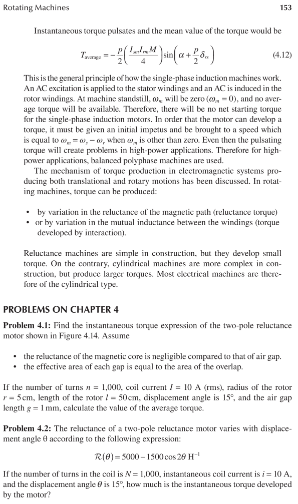

Rotating Machines 153

Instantaneous torque pulsates and the mean value of the torque would be

$$T_{\text{average}} = -\frac{p}{2}\left(\frac{I_{sm}I_{rm}M}{4}\right)\sin\left(\alpha + \frac{p}{2}\delta_{rs}\right) \tag{4.12}$$

This is the general principle of how the single-phase induction machines work. An AC excitation is applied to the stator windings and an AC is induced in the rotor windings. At machine standstill, ω_m will be zero ($\omega_m = 0$), and no average torque will be available. Therefore, there will be no net starting torque for the single-phase induction motors. In order that the motor can develop a torque, it must be given an initial impetus and be brought to a speed which is equal to $\omega_m = \omega_s - \omega_r$ when ω_m is other than zero. Even then the pulsating torque will create problems in high-power applications. Therefore for high-power applications, balanced polyphase machines are used.

The mechanism of torque production in electromagnetic systems producing both translational and rotary motions has been discussed. In rotating machines, torque can be produced:

- by variation in the reluctance of the magnetic path (reluctance torque)
- or by variation in the mutual inductance between the windings (torque developed by interaction).

Reluctance machines are simple in construction, but they develop small torque. On the contrary, cylindrical machines are more complex in construction, but produce larger torques. Most electrical machines are therefore of the cylindrical type.

PROBLEMS ON CHAPTER 4

Problem 4.1: Find the instantaneous torque expression of the two-pole reluctance motor shown in Figure 4.14. Assume

- the reluctance of the magnetic core is negligible compared to that of air gap.
- the effective area of each gap is equal to the area of the overlap.

If the number of turns $n = 1{,}000$, coil current $I = 10$ A (rms), radius of the rotor $r = 5$ cm, length of the rotor $l = 50$ cm, displacement angle is 15°, and the air gap length $g = 1$ mm, calculate the value of the average torque.

Problem 4.2: The reluctance of a two-pole reluctance motor varies with displacement angle θ according to the following expression:

$$\mathcal{R}(\theta) = 5000 - 1500\cos 2\theta \ \mathrm{H}^{-1}$$

If the number of turns in the coil is $N = 1{,}000$, instantaneous coil current is $i = 10$ A, and the displacement angle θ is 15°, how much is the instantaneous torque developed by the motor?

Figure 1.61: eec05.png   Figure 1.62: eec06.png

3) C:/DATA-NovoVivo-128-2020/Books-128-2020/Theory of Satellite Fragmentation in Orbit/9811208557.pdf Pag. 72 (83) - Fig. 4.4. Plane change angle from the breakup spherical triangles of parent and fragment satellites.

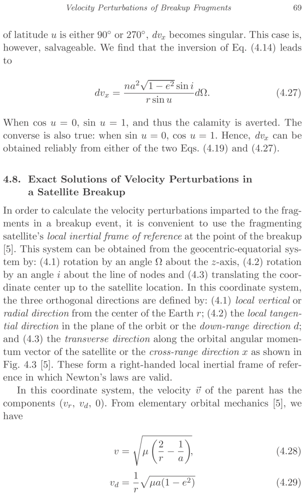

Velocity Perturbations of Breakup Fragments 69

of latitude u is either 90° or 270°, dv_x becomes singular. This case is, however, salvageable. We find that the inversion of Eq. (4.14) leads to

$$dv_x = \frac{na^2\sqrt{1-e^2}\sin i}{r\sin u}d\Omega. \tag{4.27}$$

When $\cos u = 0$, $\sin u = 1$, and thus the calamity is averted. The converse is also true: when $\sin u = 0$, $\cos u = 1$. Hence, dv_x can be obtained reliably from either of the two Eqs. (4.19) and (4.27).

4.8. Exact Solutions of Velocity Perturbations in a Satellite Breakup

In order to calculate the velocity perturbations imparted to the fragments in a breakup event, it is convenient to use the fragmenting satellite's *local inertial frame of reference* at the point of the breakup [5]. This system can be obtained from the geocentric-equatorial system by: (4.1) rotation by an angle Ω about the z-axis, (4.2) rotation by an angle i about the line of nodes and (4.3) translating the coordinate center up to the satellite location. In this coordinate system, the three orthogonal directions are defined by: (4.1) *local vertical* or *radial direction* from the center of the Earth r; (4.2) the *local tangential direction* in the plane of the orbit or the *down-range direction* d; and (4.3) the *transverse direction* along the orbital angular momentum vector of the satellite or the *cross-range direction* x as shown in Fig. 4.3 [5]. These form a right-handed local inertial frame of reference in which Newton's laws are valid.

In this coordinate system, the velocity $\vec{v}$ of the parent has the components (v_r, v_d, 0). From elementary orbital mechanics [5], we have

$$v = \sqrt{\mu\left(\frac{2}{r} - \frac{1}{a}\right)}, \tag{4.28}$$

$$v_d = \frac{1}{r}\sqrt{\mu a(1-e^2)} \tag{4.29}$$

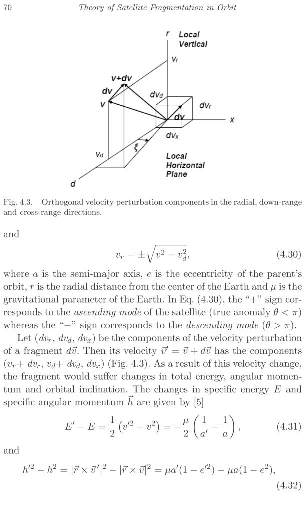

70 Theory of Satellite Fragmentation in Orbit

Fig. 4.3. Orthogonal velocity perturbation components in the radial, down-range and cross-range directions.

and

$$v_r = \pm\sqrt{v^2 - v_d^2}, \tag{4.30}$$

where a is the semi-major axis, e is the eccentricity of the parent's orbit, r is the radial distance from the center of the Earth and μ is the gravitational parameter of the Earth. In Eq. (4.30), the "+" sign corresponds to the *ascending mode* of the satellite (true anomaly $\theta < \pi$) whereas the "−" sign corresponds to the *descending mode* ($\theta > \pi$).

Let (dv_r, dv_d, dv_x) be the components of the velocity perturbation of a fragment $d\vec{v}$. Then its velocity $\vec{v}' = \vec{v} + d\vec{v}$ has the components (v_r+ dv_r, v_d+ dv_d, dv_x) (Fig. 4.3). As a result of this velocity change, the fragment would suffer changes in total energy, angular momentum and orbital inclination. The changes in specific energy E and specific angular momentum $\vec{h}$ are given by [5]

$$E' - E = \frac{1}{2}\left(v'^2 - v^2\right) = -\frac{\mu}{2}\left(\frac{1}{a'} - \frac{1}{a}\right), \tag{4.31}$$

and

$$h'^2 - h^2 = |\vec{r}\times\vec{v}'|^2 - |\vec{r}\times\vec{v}|^2 = \mu a'(1-e'^2) - \mu a(1-e^2), \tag{4.32}$$

Figure 1.63: tsfo01.png   Figure 1.64: tsfo02.png

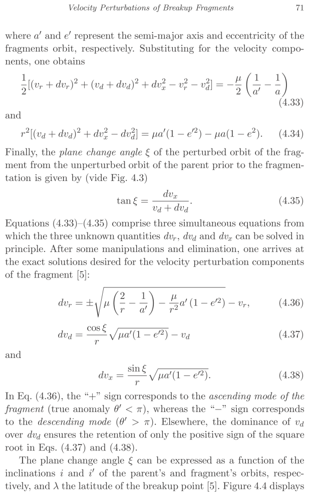

where a' and e' represent the semi-major axis and eccentricity of the fragments orbit, respectively. Substituting for the velocity components, one obtains

$$\frac{1}{2}[(v_r + dv_r)^2 + (v_d + dv_d)^2 + dv_x^2 - v_r^2 - v_d^2] = -\frac{\mu}{2}\left(\frac{1}{a'} - \frac{1}{a}\right) \tag{4.33}$$

and

$$r^2[(v_d + dv_d)^2 + dv_x^2 - dv_d^2] = \mu a'(1 - e'^2) - \mu a(1 - e^2). \tag{4.34}$$

Finally, the *plane change angle* ξ of the perturbed orbit of the fragment from the unperturbed orbit of the parent prior to the fragmentation is given by (vide Fig. 4.3)

$$\tan\xi = \frac{dv_x}{v_d + dv_d}. \tag{4.35}$$

Equations (4.33)–(4.35) comprise three simultaneous equations from which the three unknown quantities dv_r, dv_d and dv_x can be solved in principle. After some manipulations and elimination, one arrives at the exact solutions desired for the velocity perturbation components of the fragment [5]:

$$dv_r = \pm\sqrt{\mu\left(\frac{2}{r} - \frac{1}{a'}\right) - \frac{\mu}{r^2}a'\,(1 - e'^2)} - v_r, \tag{4.36}$$

$$dv_d = \frac{\cos\xi}{r}\sqrt{\mu a'(1 - e'^2)} - v_d \tag{4.37}$$

and

$$dv_x = \frac{\sin\xi}{r}\sqrt{\mu a'(1 - e'^2)}. \tag{4.38}$$

In Eq. (4.36), the "+" sign corresponds to the *ascending mode of the fragment* (true anomaly $\theta' < \pi$), whereas the "−" sign corresponds to the *descending mode* ($\theta' > \pi$). Elsewhere, the dominance of v_d over dv_d ensures the retention of only the positive sign of the square root in Eqs. (4.37) and (4.38).

The plane change angle ξ can be expressed as a function of the inclinations i and i' of the parent's and fragment's orbits, respectively, and λ the latitude of the breakup point [5]. Figure 4.4 displays

Figure 1.65: tsfo03.png

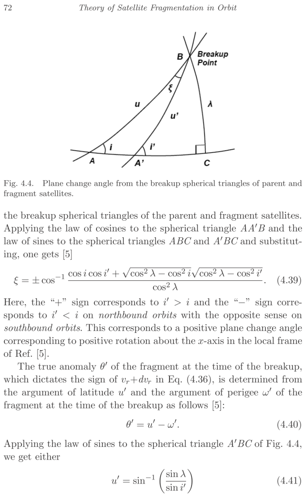

Fig. 4.4. Plane change angle from the breakup spherical triangles of parent and fragment satellites.

the breakup spherical triangles of the parent and fragment satellites. Applying the law of cosines to the spherical triangle $AA'B$ and the law of sines to the spherical triangles ABC and $A'BC$ and substituting, one gets [5]

$$\xi = \pm\cos^{-1}\frac{\cos i\cos i' + \sqrt{\cos^2\lambda - \cos^2 i}\sqrt{\cos^2\lambda - \cos^2 i'}}{\cos^2\lambda}. \tag{4.39}$$

Here, the "+" sign corresponds to $i' > i$ and the "−" sign corresponds to $i' < i$ on *northbound orbits* with the opposite sense on *southbound orbits*. This corresponds to a positive plane change angle corresponding to positive rotation about the x-axis in the local frame of Ref. [5].

The true anomaly θ' of the fragment at the time of the breakup, which dictates the sign of $v_r + dv_r$ in Eq. (4.36), is determined from the argument of latitude u' and the argument of perigee ω' of the fragment at the time of the breakup as follows [5]:

$$\theta' = u' - \omega'. \tag{4.40}$$

Applying the law of sines to the spherical triangle $A'BC$ of Fig. 4.4, we get either

$$u' = \sin^{-1}\left(\frac{\sin\lambda}{\sin i'}\right) \tag{4.41}$$

Figure 1.66: tsfo04.png

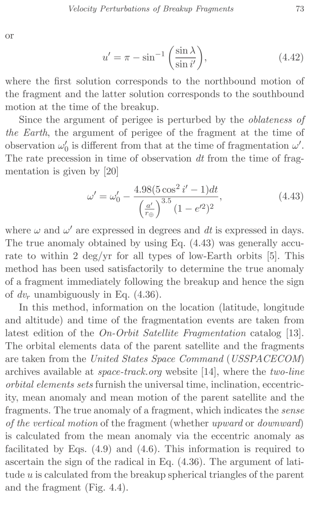

or

$$u' = \pi - \sin^{-1}\left(\frac{\sin\lambda}{\sin i'}\right), \tag{4.42}$$

where the first solution corresponds to the northbound motion of the fragment and the latter solution corresponds to the southbound motion at the time of the breakup.

Since the argument of perigee is perturbed by the *oblateness of the Earth*, the argument of perigee of the fragment at the time of observation ω'_0 is different from that at the time of fragmentation ω'. The rate precession in time of observation dt from the time of fragmentation is given by [20]

$$\omega' = \omega'_0 - \frac{4.98(5\cos^2 i' - 1)dt}{\left(\frac{a'}{r_\oplus}\right)^{3.5}(1 - e'^2)^2}, \tag{4.43}$$

where ω and ω' are expressed in degrees and dt is expressed in days. The true anomaly obtained by using Eq. (4.43) was generally accurate to within 2 deg/yr for all types of low-Earth orbits [5]. This method has been used satisfactorily to determine the true anomaly of a fragment immediately following the breakup and hence the sign of dv_r unambiguously in Eq. (4.36).

In this method, information on the location (latitude, longitude and altitude) and time of the fragmentation events are taken from latest edition of the *On-Orbit Satellite Fragmentation* catalog [13]. The orbital elements data of the parent satellite and the fragments are taken from the *United States Space Command* (*USSPACECOM*) archives available at *space-track.org* website [14], where the *two-line orbital elements sets* furnish the universal time, inclination, eccentricity, mean anomaly and mean motion of the parent satellite and the fragments. The true anomaly of a fragment, which indicates the *sense of the vertical motion* of the fragment (whether *upward* or *downward*) is calculated from the mean anomaly via the eccentric anomaly as facilitated by Eqs. (4.9) and (4.6). This information is required to ascertain the sign of the radical in Eq. (4.36). The argument of latitude u is calculated from the breakup spherical triangles of the parent and the fragment (Fig. 4.4).

Figure 1.67: tsfo05.png

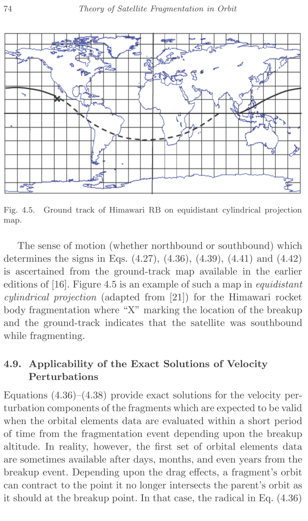

Fig. 4.5. Ground track of Himawari RB on equidistant cylindrical projection map.

The sense of motion (whether northbound or southbound) which determines the signs in Eqs. (4.27), (4.36), (4.39), (4.41) and (4.42) is ascertained from the ground-track map available in the earlier editions of [16]. Figure 4.5 is an example of such a map in *equidistant cylindrical projection* (adapted from [21]) for the Himawari rocket body fragmentation where "X" marking the location of the breakup and the ground-track indicates that the satellite was southbound while fragmenting.

4.9. Applicability of the Exact Solutions of Velocity Perturbations

Equations (4.36)–(4.38) provide exact solutions for the velocity perturbation components of the fragments which are expected to be valid when the orbital elements data are evaluated within a short period of time from the fragmentation event depending upon the breakup altitude. In reality, however, the first set of orbital elements data are sometimes available after days, months, and even years from the breakup event. Depending upon the drag effects, a fragment's orbit can contract to the point it no longer intersects the parent's orbit as it should at the breakup point. In that case, the radical in Eq. (4.36)

Figure 1.68: tsfo06.png

1.3 History of Geometry

1) C:/DATA-NovoVivo-135-2020/Books-135-2020/Jet Single-Time Lagrange Geometry and Its Applications/1118127552.pdf

Pag. 06(24) - Riemannian metric on the relativistic thime axis R...

2) C:/DATA-NovoVivo-135-2020/Books-135-2020/Geometry/1904.01845.pdf Assunto: From Euclid to Riemann and Beyond

3) C:/DATA-NovoVivo-135-2020/Books-135-2020/Geometry/82059800.pdf Assunto: A Brief History of Contact Geometry and Topology

4) C:/DATA-NovoVivo-135-2020/Books-135-2020/Geometry/Algebra and geometry in several complex variables.pdf Chapter 1 - Holomorphic functions of several variables 1.1 Analysis in complex plane

5) C:/DATA-NovoVivo-135-2020/Books-135-2020/Geometry/Book-essentials-A4.pdf Pag. 10 - Quadrature of Circle

6) C:/DATA-NovoVivo-135-2020/Books-135-2020/Geometry/Curve and surface reconstruction Algorithms with mathematical analysis.pdf

7) C:/DATA-NovoVivo-135-2020/Books-135-2020/Geometry/Effective computational geometry for curves and surfaces.pdf Pag. 221 (230) - The union of a sphere and a line, and the first half of the vertical cross-sections. The cross-sections in the right half are symmetric. The slab points are marked white.

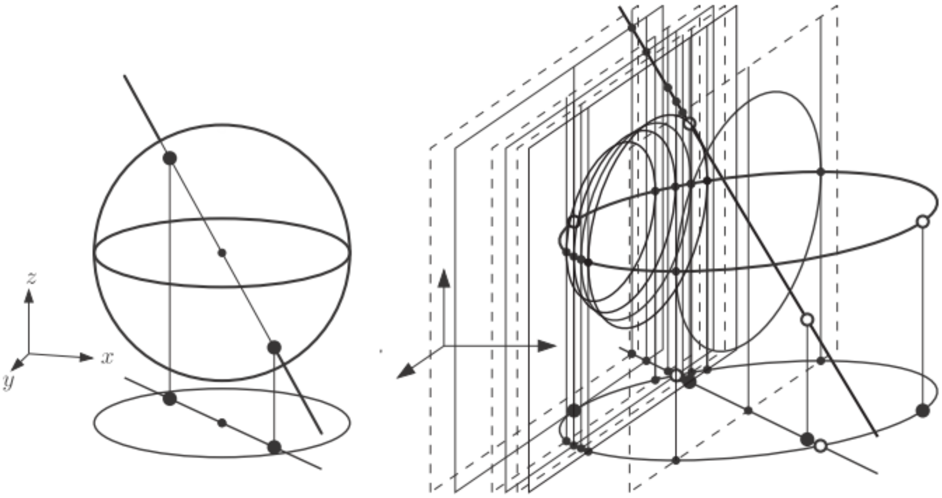

Figure 1.69: cg01.png

Pag. 231 - 6 Delaunay Triangulation Based Surface Reconstruction

8) C:/DATA-NovoVivo-135-2020/Books-135-2020/Geometry/EJ717835.pdf Pag. 03 - Pseudosphere Lobachevsky's calculation.

9) C:/DATA-NovoVivo-135-2020/Books-135-2020/Geometry/Euclidean geometry.pdf

10) C:/DATA-NovoVivo-135-2020/Books-135-2020/Geometry/Euclidean geometry.pdf Pag. 6 - Squaring the circle: the areas of this square and this circle are equal. In 1882, it was proven that this figure cannot be constructed in a finite number of steps with an idealized compass and straightedge.

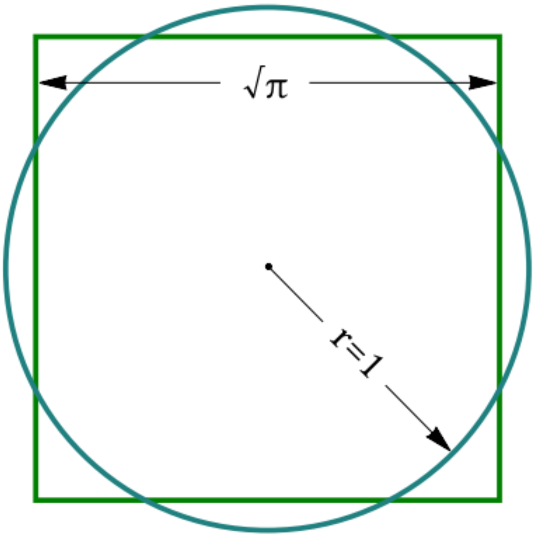

Figure 1.70: cg02.png

11) C:/DATA-NovoVivo-135-2020/Books-135-2020/Geometry/euclidean-geometry-october-final.pdf

12) C:/DATA-NovoVivo-135-2020/Books-135-2020/Geometry/FULLTEXT01.pdf

IMPORTANT

Pag. 07 (09) - Definition 2.2. The Gaussian curvature at a point x on a surface is

$$\kappa(x) = \lim_{r \to 0+} 12 \frac{\pi r^2 - A(r,x)}{\pi r^4}$$

Figure 1.71: cg03.png

The above formula is known as the Bertrand{Diquet{Puiseux theorem[7] We chose it for its simplicity and since it compares neatly to the definition of the curvature at a point of a curve: the deviation from its tangent. A foundamental theorem by Gauss, Theorema Egregium, published in 1827, says that Gaussian curvature is intrinsic which means that it does not depend on how the surface is embedded into Euclidean 3-space[8]

---

[7] First described in 1848 by its eponymous authors in the French journal Journal de mathematiques pures et appliquees (which can be found online). Gaussian curvature is generally defined as the product of principal curvatures, the formula chosen is an alternate definition.

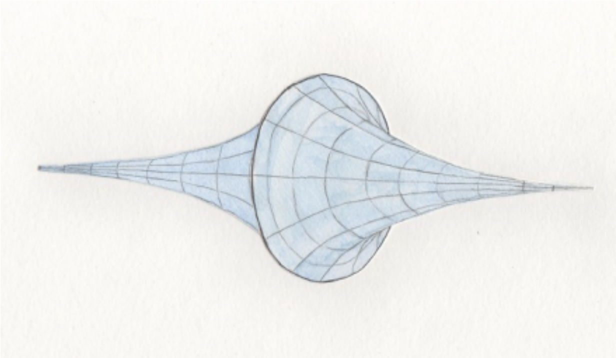

Figure 1.72: cg04.png - Pseudosphere - tractricoid

A tractricoid, which is an example of a pseudosphere. Note that as pseudosphere will necessarily contain singularities; no complete and regular surface of constant negative curvature exists in Euclidean 3-space by Hilbert's theorem (1901).

13) C:/DATA-NovoVivo-135-2020/Books-135-2020/Geometry/Geometry from a differentiable viewpoint.pdf Obs.: Livro todo!!!

Pag. 06(23) - Theorem 1.2 (Spherical Pythagorean Theorem). If triangel ABC is a right triangle on a sphere of radius R with right angle at the vertex C, then

$$\cos\frac{c}{R} = \cos\frac{a}{R} : \cos b R$$

.

---

[8] See more for example in [Gray, 2007].

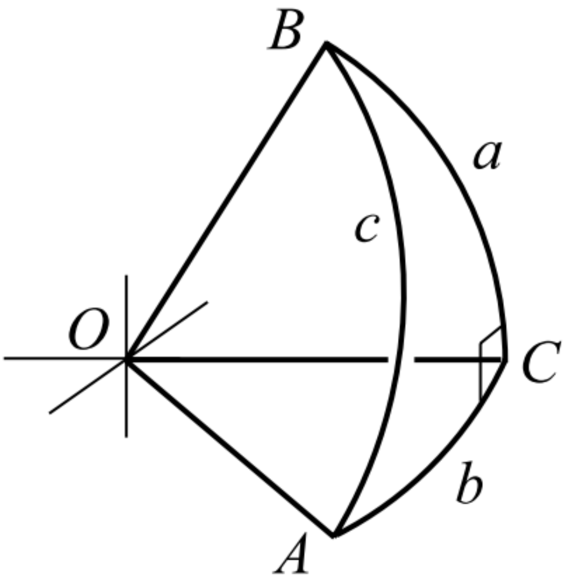

Figure 1.73: cg05.png

14) C:/DATA-NovoVivo-135-2020/Books-135-2020/Geometry/Geometry in History.pdf

15) C:/DATA-NovoVivo-135-2020/Books-135-2020/Geometry/Geometry Through History Euclidean, Hyperbolic, and Projective Geometries.pdf

Pag. 51(61) - Chapter 2 Neutral Geometry 2.1 The Problem of the Parallels

16) C:/DATA-NovoVivo-135-2020/Books-135-2020/Geometry/Mathematics+and+Its+History.pd

Pag. 27(50) - Use similar triangles to explain why $\sqrt{l}$ is the length shown in Figure 2.5, and hence show that $\sqrt{l}$ is constructible from l. One of the finest ruler and compass constructions from ancient times is that of the regular pentagon, which includes, yet again, the golden ratio

$$= 1 + \frac{\sqrt{5}}{2} \tag{1.13}$$

Knowing (from the questions above) that this number is constructible, it becomes easy for us to construct the pentagon itself.

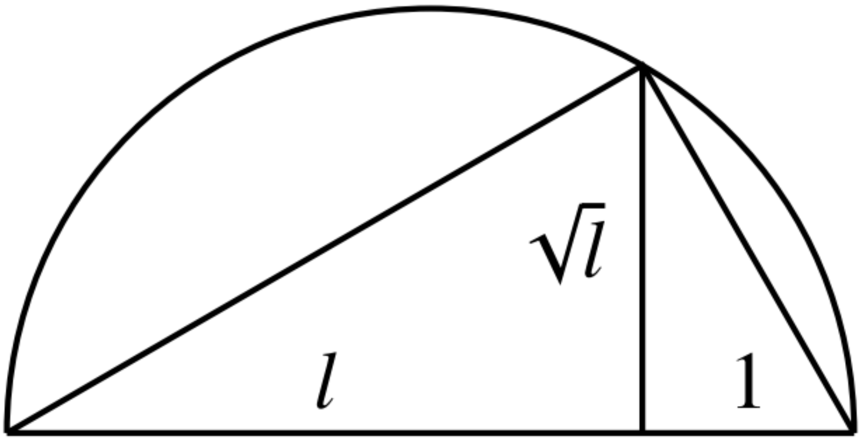

Figure 1.74: cg06.png

17) C:/DATA-NovoVivo-135-2020/Books-135-2020/Geometry/Pythagorean-hodograph curves algebra and geometry inseparable.pdf Obs.: Livro todo!!!

18) C:/DATA-NovoVivo-135-2020/Books-135-2020/Geometry/5000 years of geometry - mathematics in history and culture - 2015.pdf

19) C:/DATA-NovoVivo-135-2020/Books-135-2020/Geometry/tabak-j-geometry-the-language-of-space-and-form.pdf

20) C:/DATA-NovoVivo-135-2020/Books-135-2020/Geometry/An-invitation-to-algebraic-geometry-Springer-2000.pdf

21) C:/DATA-NovoVivo-135-2020/Books-135-2020/Geometry/what-is-geometry.pdf

22) C:/ARQ-IMPORTANT-08/Important-02/Calculus with Curvilinear Coordinates - Problems and Solutions 2018.pdf Obs.: Livro todo!!!

1.4 Hyperbolische Geometrie

1) https://www.spektrum.de/lexikon/mathematik/hyperbolische-geometrie/4090 Hyperbolische Geometrie Anzeige I.

Als hyperbolische Geometrie oder Lobatschewski-Geometrie wird eine nichteuklidische Geometrie bezeichnet, in der alle Axiome der absoluten Geometrie, also die Axiome der Inzidenz, der Anordnung, der Kongruenz und der Stetigkeit (siehe Axiome der Geometrie), sowie die Verneinung des Parallelenaxioms des Euklid gelten:

Es existiert eine Gerade g und ein nicht auf g liegender Punkt P, durch den mindestens zwei Geraden verlaufen, die mit g in einer Ebene liegen und g nicht schneiden.

Aus diesem Axiom, das auch als Lobatschewskisches Parallelenaxiom bezeichnet wird, und den Axiomen der absoluten Geometrie läßt sich ableiten, daß sogar zu jeder Geraden g und jedem nicht auf g liegenden Punkt P unendlich viele Geraden verlaufen, die mit g in einer Ebene liegen und g nicht schneiden. Allerdings werden nicht alle dieser Geraden als zu g parallel bezeichnet (siehe III.).

Zu der Erkenntnis, daß es eine nichteuklidische Geometrie, basierend auf den Axiomen der absoluten Geometrie und der Negation des euklidischen Parallelenaxioms gibt, gelangten zwischen 1816 und 1832 weitgehend unabhängig voneinander die drei Mathematiker Janos Bolyai, Carl Friedrich Gauss und Nikolai Iwanowitsch Lobatschewski (nichteuklidische Geometrie). Sie lösten damit das seit mehr als zweitausend Jahren bestehende Parallelenproblem auf eine unerwartete Weise, indem sie zeigten, daß das euklidische Parallelenaxiom nicht aus den Axiomen der absoluten Geometrie ableitbar ist. II. Eigenschaften der hyperbolischen Geometrie

Obwohl sich die euklidische und die hyperbolische Geometrie in ihrer Axiomatik nur um ein einziges Axiom, das Parallelenaxiom, unterscheiden, ergeben sich daraus gravierende Unterschiede beider Geometrien in wichtigen Eigenschaften. Zu den interessantesten Eigenschaften der hyperbolischen

Geometrie, die von denen der euklidischen Geometrie abweichen, gehören die folgenden:

Die Innenwinkelsumme eines jeden Dreiecks ist kleiner als 180°. Es existiert kein spitzer Winkel, für den alle in beliebigen Punken eines seiner Schenkel errichteten Senkrechten den anderen Schenkel treffen. Abstandslinien sind keine Geraden, d. h. die Menge aller Punkte, die von einer gegebenen Geraden g denselben positiven Abstand haben und in einer durch g begrenzten Halbebene liegen, ist in keinem Fall eine Gerade. Zwei Dreiecke, die paarweise in allen drei Winkeln übereinstimmen, sind kongruent. Es existieren also keine ähnlichen und dabei nicht kongruenten Dreiecke. Drei Punkte, die nicht auf einer Geraden liegen, gehören nicht in jedem Falle einem Kreis an; im Raum wird durch vier nicht auf einer Ebene liegende Punkte nicht notwendigerweise eine Kugel bestimmt.

III. Parallele und divergierende Geraden

Um in der hyperbolischen Geometrie die Parallelität von Geraden zu definieren, wird der Parallelwinkel ' betrachtet. Es handelt sich dabei für eine gegebene Gerade a und einen nicht auf a liegenden Punkt P um den kleinsten Winkel zwischen einer Geraden b durch P, die mit a in einer Ebene liegt und a nicht schneidet, und dem Lot von P auf a. Dieser Parallelwinkel (oder Grenzwinkel) ist auf beiden Seiten des Lotes gleich groß.

Die Halbgeraden

$$b_1^+$$

und

$$b_2^+$$

mit dem Anfangspunkt P, für die

$$< \; PA^+ ; b_1^+ \; = < \; Pa^+ ; b_2^+ \; =$$

gilt (wobei ' Parallelwinkel in P in Bezug auf a und A Fußpunkt des Lotes von P auf a ist), haben mit der Geraden a keinen Punkt gemeinsam.

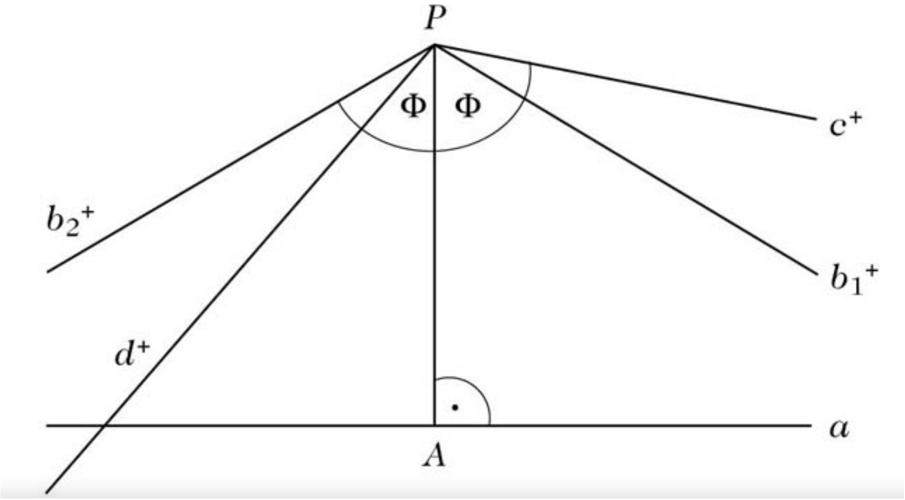

Figure 1.75: hg01.png

Parallelwinkel

Gleiches gilt für jede Halbgerade c^+ mit dem Anfangspunkt P und $<(PA^+ ; c^+) > '$. Ist dagegen d^+ eine Halbgerade mit $< (PA^+ ; d^+) < '$, so schneidet d^+ die Gerade a. Die Geraden b1 und b2, denen die Halbgeraden

$$b_1^+$$

und

$$b_2^+$$

angehören. sind Grenzgeraden in der Gesamtheit aller Geraden, die durch P verlaufen und a nicht schneiden und werden als zu a parallele Geraden bezeichnet. Die so definierte Parallelität ist symmetrisch und transitiv.

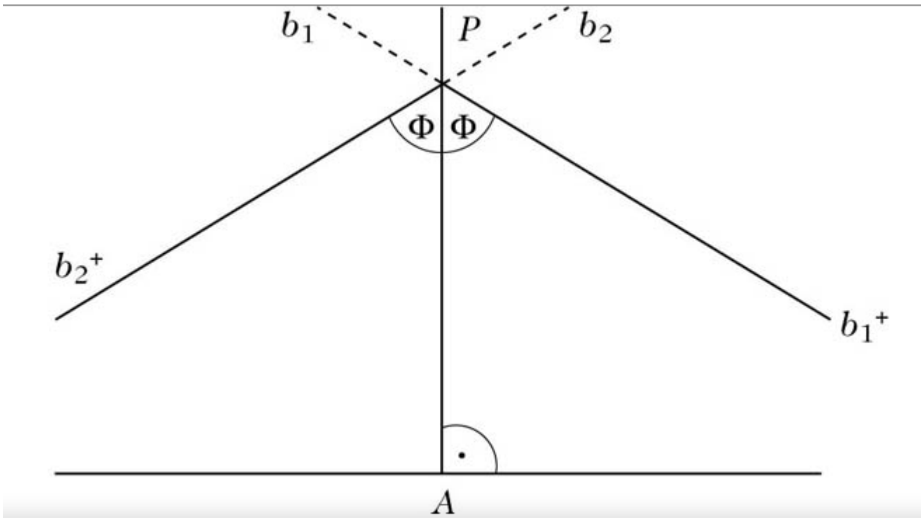

Figure 1.76: hg02.png

Zwei Geraden einer Ebene, die keinen gemeinsamen Punkt besitzen und nicht parallel sind, heißen divergierend. Zwei divergierende Geraden besitzen stets eine gemeinsame Senkrechte, dagegen gibt es zu zwei parallelen Geraden in keinem Falle eine Gerade, die auf den beiden Parallelen senkrecht steht. Es gilt sogar, daß es keine Gerade gibt, die mit zwei parallelen Geraden gleiche Stufen- oder Wechselwinkel bildet.

IV. Die Lobatschewskische Funktion

Eine fundamentale Eigenschaft der hyperbolischen Geometrie besteht darin, daß der Parallelwinkel in einem Punkt P in Bezug auf eine Gerade g nur vom Abstand x des Punktes P von der Geraden g abhängt. Der Parallelwinkel läßt sich somit als Funktion des Abstandes x auffassen:

$$= (\ x); \tag{1.14}$$

wobei ; als Lobatschewskische Funktion bezeichnet wird. Der Definitionsbereich dieser Funktion ist $(0;1)$. ihr Weterbereich

$$(0;\frac{}{2}) \tag{1.15}$$

. ist auf ihrem gesamten Definitionsbereich stetig sowie streng monoton fallend und besitzt folgende uneigentliche Grenzwerte:

$$\lim_{x!1}(x)=0;\quad \lim_{x!\ 0^{+}}(x)=\frac{}{2}: \tag{1.16}$$

Aus dem zweiten dieser Grenzwerte ergibt sich, daß für sehr gering ausgedehnte Bereiche der Parallelwinkel nahezu gleich

$$\frac{}{2} \tag{1.17}$$

ist, damit sind die beiden Parallelen zu einer Geraden durch einen Punkt kaum voneinander zu unterscheiden. Alle Eigenschaften der hyperbolischen Geometrie nähern sich deshalb in sehr kleinen Bereichen denen der euklidischen Geometrie an, da die Unterschiede beider Geometrien nur auf den verschiedenen Parallelenaxiomen beruhen. Das bedeutet, daß sich aus der Kenntnis "sehr kleiner Teile" eines Raumes nicht genau bestimmen läßt, ob es sich um einen euklidischen oder einen nichteuklidischen Raum handelt.

Analytisch läßt sich die Lobatschewskische Funktion durch die Gleichung

$$(x)=2\quad\arctan\exp\frac{x}{R} \tag{1.18}$$

darstellen, die u. a. benötigt wird, um die Formeln der hyperbolischen Trigonometrie herzuleiten.

Eine sehr wichtige Schlußfolgerung aus der Existenz der Lobatschewskischen Funktion besteht darin, daß es in der hyperbolischen Geometrie eine absolute Länge gibt. Durch einen gestreckten Winkel (bzw. Teile eines solchen, die beispielsweise durch Halbierung gewonnen werden können) sind in der euklidischen Geometrie absolute Winkelgrößen gegeben, die sich durch eine abstrakte Vorschrift beschreiben lassen. So ist die Größe eines rechten Winkels durch die Konstruktionsbeschreibung "Errichtung der Senkrechten" gegeben und bedarf keiner willkürlich festgelegten Eichhilfen, wie des Einheitsmeters. Im Gegensatz dazu gibt es in der euklidischen Geometrie keine absoluten

Streckengrößen. In der hyperbolischen Geometrie existieren derartige absolute Längen. Sie lassen sich aus absoluten Winkeln mittels der Umkehrfunktion der Funktion ermitteln. Beispielsweise wird durch

$$l := {}^{1}\ \frac{}{4} \tag{1.19}$$

eine Länge l ausgezeichnet, die keinerlei Willkür unterliegt.

Die Existenz einer absoluten Länge ist mit der Nichtexistenz ähnlicher Figuren verbunden. Das (PageNum - 460) Vorhandensein ähnlicher Figuren (die nur in der euklidischen Geometrie existieren) schließt die Möglichkeit der Bestimmung absoluter Längen aus. V. Modelle der hyperbolischen Geometrie

Um die Widerspruchsfreiheit der hyperbolischen Geometrie (und gleichbedeutend damit die Unabhängigkeit des Parallelenaxioms des Euklid von den Axiomen der absoluten Geometrie) nachzuweisen sowie eine anschauliche Interpretation der hyperbolischen Geometrie zu geben, werden die Grundbegriffe der hyperbolischen Geometrie durch bekannte Objekte der euklidischen (oder projektiven) Geometrie interpretiert. Dabei sind sehr unterschiedliche Interpretationen (Modelle) möglich. Zu den bekanntesten zählen die Poincaré-Halbebene, die Poincaré-Kreisscheibe sowie das Kleinsche Modell, bei denen die zweidimensionale hyperbolische Geometrie innerhalb einer offenen Halbebene bzw. einer offenen Kreisscheibe der euklidischen Ebene aufgebaut wird.

Eines der interessantesten Modelle der hyperbolischen Geometrie ist das pseudoeuklidische Modell, bei dem die Analogie zur sphärischen Geometrie (die ihrerseits ein Modell der elliptischen Geometrie ist) gut sichtbar wird. Die hyperbolische Ebene wird hierbei als Sphäre H mit imaginärem Radius innerhalb eines dreidimensionalen pseudoeuklidischen Raumes modelliert. (Aus euklidischer Sicht ist diese Sphäre ein zweischaliges Rotationshyperboloid, aus der Sicht der hyperbolischen Geometrie eine Ebene.) Hyperbolische Punkte sind bei diesem Modell alle diametralen Punktepaare auf H, also Paare von Punkten, die auf einer Geraden durch den Koordinatenursprung liegen (siehe Abbildung { die euklidischen Punkte P1 und P2 werden als ein H-Punkt identifiziert).

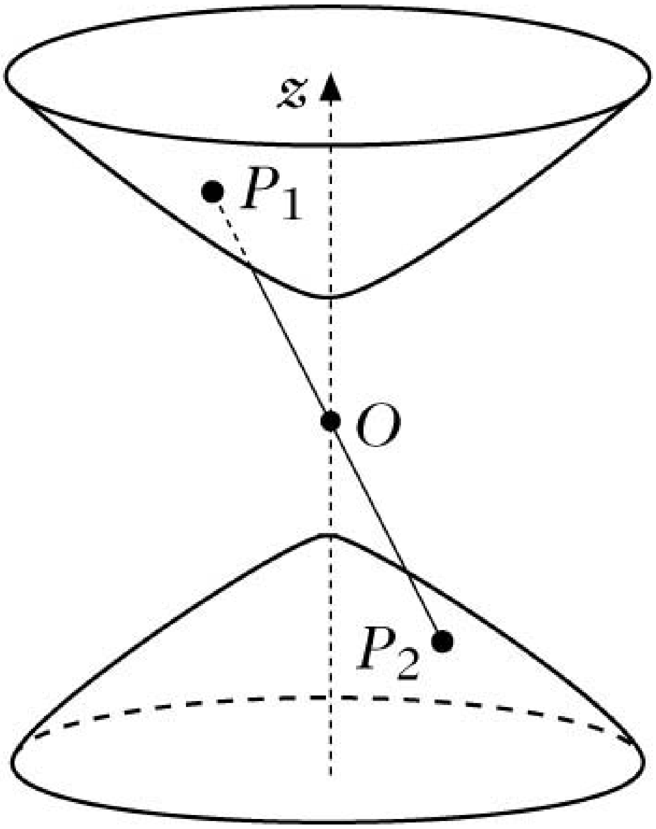

Figure 1.77: hg03.png

Hyperbolische Geraden sind alle Schnittkurven der hyperbolischen Ebene H mit (euklidischen) Ebenen, die durch den Mittelpunkt bzw. Koordinatenursprung verlaufen. Pseudoeuklidisch gesehen handelt es sich bei diesen Geraden also um Gro kreise und euklidisch um Hyperbeln.

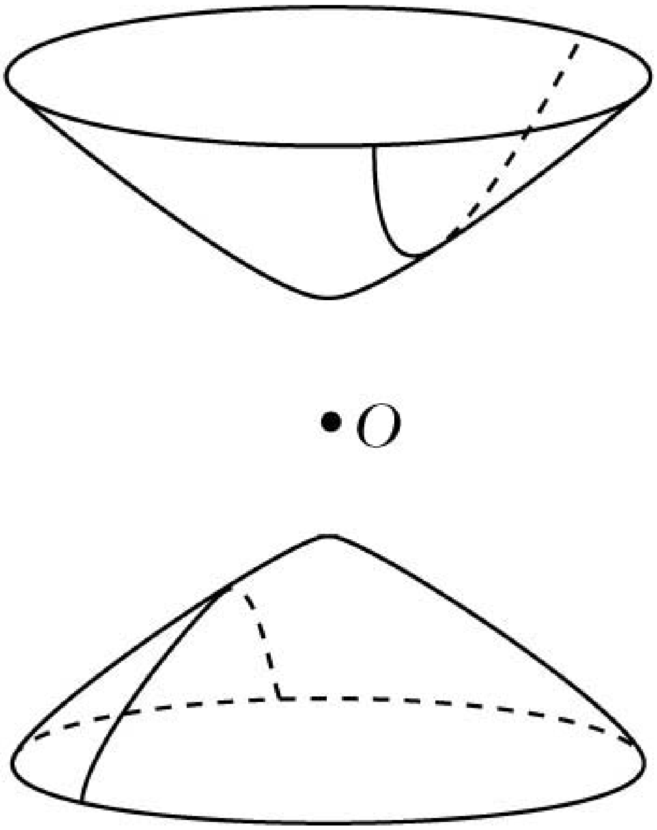

Figure 1.78: hg04.png

Geraden im pseudoeuklidischen Modell

Hyperbolische Abstände von Punkten werden im pseudoeuklidischen Modell als Längen von Großkreisbögen, welche die entsprechenden Punkte miteinander verbinden, definiert (wobei es sich bei diesen Großkreisbögen euklidisch gesehen um Hyperbelbögen handelt). Wie in der sphärischen Geometrie besteht ein Zusammenhang zwischen der Länge des Großkreisbogens zwischen zwei Punkten und dem Winkel ihrer Radiusvektoren, wobei hier die Winkel pseudoeuklidisch zu messen sind (pseudoeuklidischer Raum).

Durch Projektion der oberen Schale des Hyperboloids (mit dem Radius i) vom Koordinatenursprung aus auf die Ebene z = 1 entsteht aus dem pseudoeuklidischen Modell der hyperbolischen Geometrie das Kleinsche Modell. VI. Hyperbolische Geometrie als Geometrie auf einer Fläche konstanter negativer Krümmung

So wie die Elliptische Geometrie als Geometrie auf einer (Hyper-)Fläche konstanter positiver Krümmung und die euklidische Geometrie als Geometrie auf einer Fläche der Krümmung Null aufgefaßt werden können, läßt sich die hyperbolische Geometrie als Geometrie auf einer Fläche konstanter negativer Krümmung interpretieren. Innerhalb des pseudoeuklidischen Raumes besitzt eine Sphäre mit imaginärem Radius (siehe V.) eine konstante negative Krümmung, für r = i ist z. B.

$$k = \frac{1}{r^2} = \quad 1: \qquad (1.20)$$

Innerhalb des euklidischen Raumes ist die Pseudosphäre eine Fläche konstanter negativer Krümmung (der Name weist auf die Verwandtschaft mit der Sphäre als Fläche konstanter positiver Krümmung hin). Es handelt sich dabei um die Rotationsfläche der Traktrix (auch als Schleppkurve bekannt). Diese Pseudosphäre wird durch folgende Parameterdarstellung (mit den Parametern %und sowie (PageNum - 461 einer Konstante R) beschrieben:

$$\begin{aligned} x &= \% \cos ; \\ y &= \% \sin ; \\ z &= R \ln \frac{R + \sqrt{R^2 \quad \%^2}}{\%} \quad \sqrt{R^2 \quad \%^2}: \end{aligned} \qquad (1.21)$$

Die Kruuml;mmung der Pseudosphare ist in jedem ihrer Punkte

$$k = \quad \frac{1}{R^2} \qquad (1.22)$$

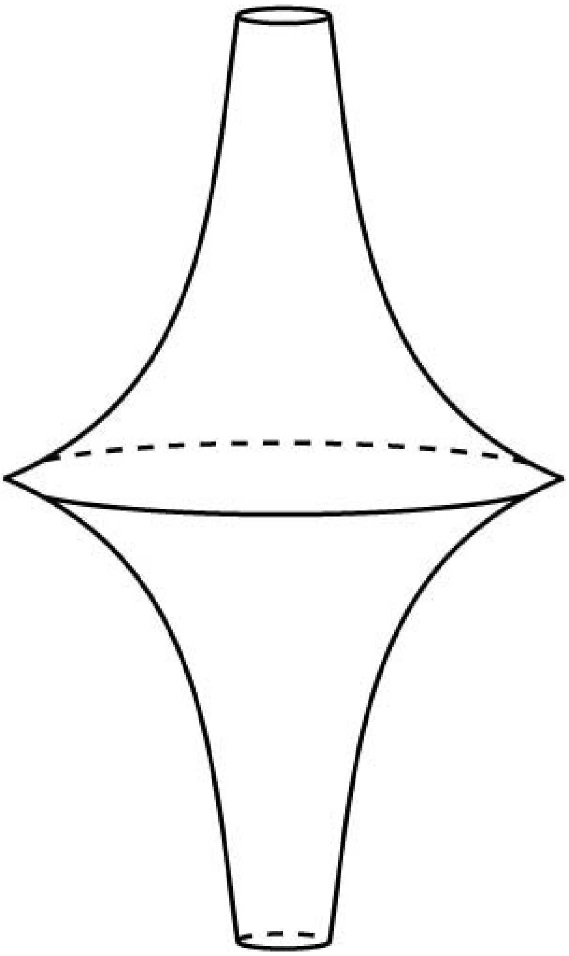

Figure 1.79: hg05.png

Pseudosphäre

Die Geometrie auf der Pseudosphäre wurde um 1840 durch Minding eingehend untersucht. ohne daß dieser jedoch einen Bezug zu den Erkenntnissen von Bolyai. Gauss und Lobatschewski hergestellte.

Beltrami schuf 1868 auf Grundlage der Untersuchungen Mindings das erste Modell der hyperbolischen Geometrie und trug damit maßgeblich zu deren Akzeptanz bei. Literatur

[1] Efimov N. W.: Höhere Geometrie. Deutscher Verlag der Wissenschaften Berlin 1960.

[2] Filler A.: Euklidische und nichteuklidische Geometrie. B.I. Wissenschaftsverlag Mannheim. 1993.

[3] Gans D.: An Introduction to Non-Euclidean Geometry. Academic

Press Inc. San Diegot 1973.

[4] Reichardt H.: Gauß und die Anfänge der nicht euklidischen Geometrie. mit Originalarbeiten von J. Bolyai, N. I. Lobatschewski und Felix Klein. Teubner Leipzig, 1985.

[5] Riemann B.: "Über die Hypothesen welche der Geometrie zugrunde liegen" in: Das Kontinuum und andere Monographien (Reprint). Chelsea. 1973.

2) https://www.deutsche-digitale-bibliothek.de/person/gnd/104234644

3) C:/DATA-NovoVivo-135-2020/Books-135-2020/Hyperbolic-Geometry/Einführung in die hyperbolische Geometrie – Anleitungen für eine Entdeckungsreise 2019.pdf

4) C:/DATA-NovoVivo-135-2020/Books-135-2020/Hyperbolic-Geometry/Euklidische-und-Nicht-Euklidische-Geometrie-in-Cin.pdf

Abb. 3: Projektion der Ebene auf die Kugel.

5) C:/DATA-NovoVivo-135-2020/Books-135-2020/Hyperbolic-Geometry/ ller-eukl-ne-geom.pdf Obs.: Livro todo!!! - 267 pages.

6) C:/DATA-NovoVivo-135-2020/Books-135-2020/Hyperbolic-Geometry/ne-geo96.pdf

Pag. 59(63) -

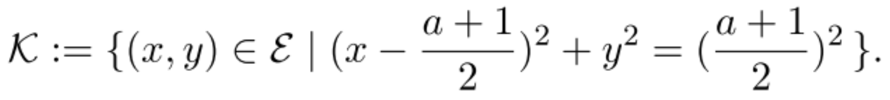

$$\mathcal{K} := \{(x,y) \in \mathcal{E} \mid (x - \frac{a+1}{2})^2 + y^2 = (\frac{a+1}{2})^2\}.$$

Figure 1.80: geo96-01.png

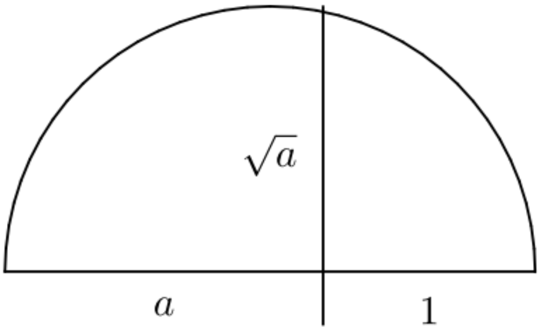

Figure 1.81: geo96-02.png

Pag. 135 (139) - Sei t die Tangente in P, der Winkel, den g mit dem Radius einschließt. Man kann dann auf der Seite von t, auf der Z liegt, einen Punkt C auf g wahlen, so daß (PC) = ist.

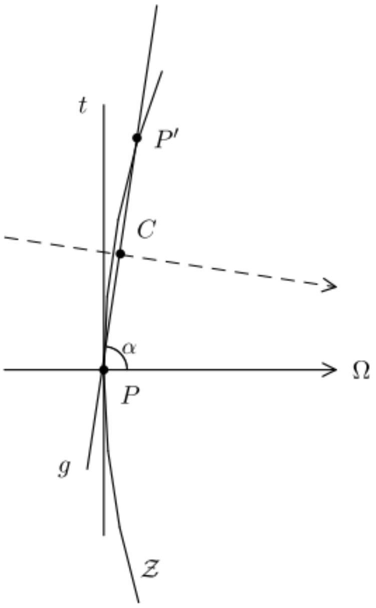
t
P'
C
α
Ω
P
g
$\mathcal{Z}$

Figure 1.82: geo96-03.png

Pag. 153 (157) -

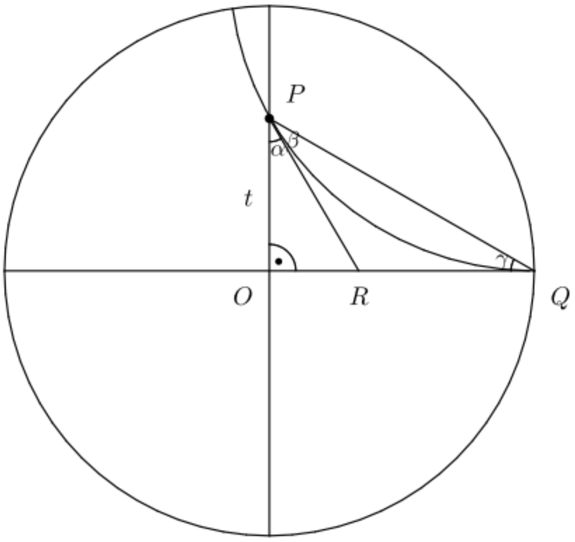

Figure 1.83: geo96-04.png

Pag. 154 (158) - Zum Schluß wollen wir uns noch einen Horozykel ansehen:

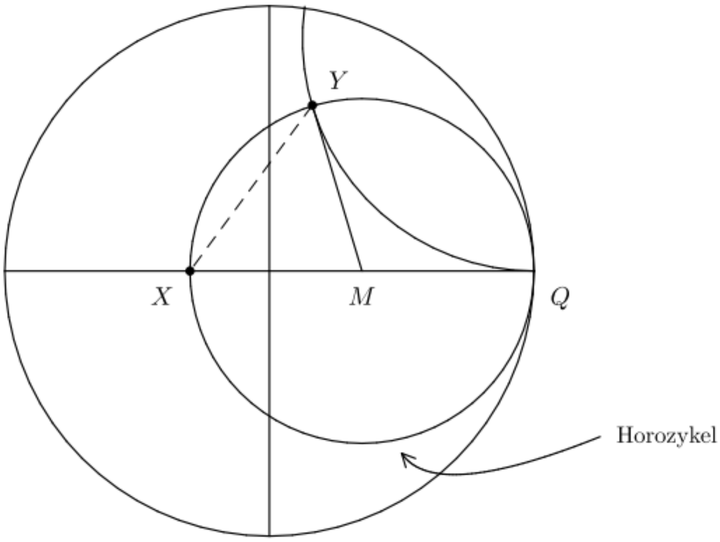

Figure 1.84: geo96-05.png

7) C:/DATA-NovoVivo-135-2020/Books-135-2020/Hyperbolic-Geometry/Elementare Lineare Algebra - Linearisieren und Koordinatisieren 2011.pdf

Pag. 60 (72) - Konstruktion eines Kreises durch drei Punkte

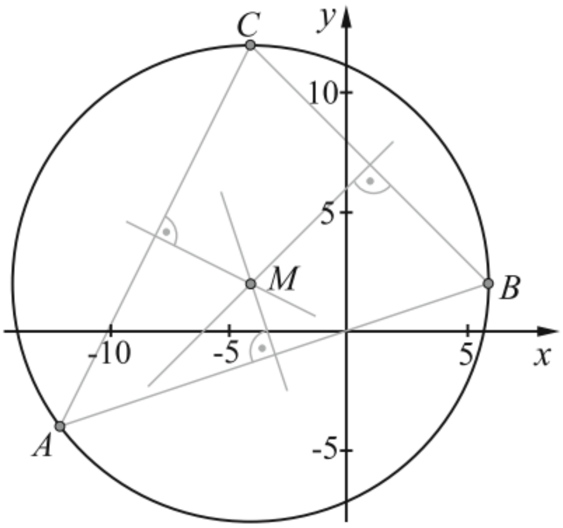
C
y
10
5
M
B
-10
-5
5
x
A
-5

Figure 1.85: ela01.png

Pag. 115 (127) -

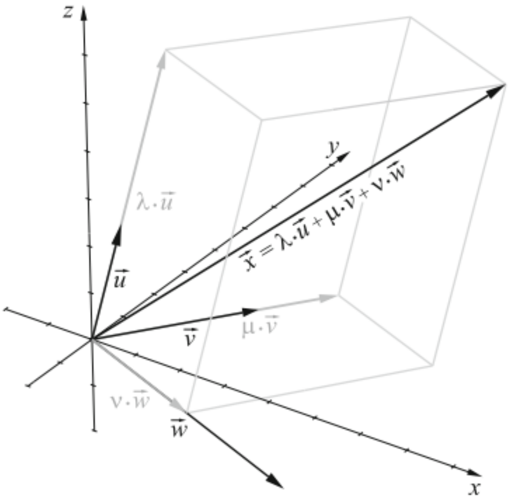
z
y
$\lambda\cdot\vec{u}$
$\vec{u}$
$\vec{x} = \lambda\cdot\vec{u} + \mu\cdot\vec{v} + \nu\cdot\vec{w}$
$\vec{v}$
$\mu\cdot\vec{v}$
$\nu\cdot\vec{w}$
$\vec{w}$
x

Figure 1.86: ela02.png

Pag. 141 (153) - Gerade mit Pfeilen, die einen Stutzvektor und einen Richtungsvektor reprasentieren, sowie einem (willkurlich gewahlten) Punkt P

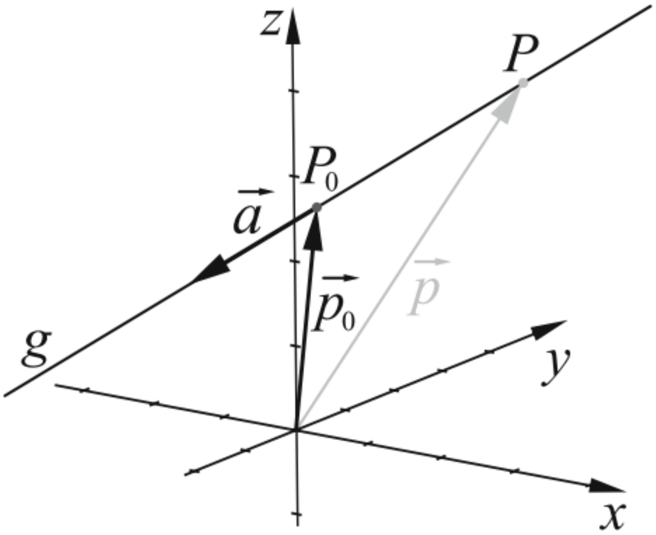
z
P
P_0
$\vec{a}$
$\vec{p}_0$
$\vec{p}$
g
y
x

Figure 1.87: ela03.png

Pag. 262 (274) - Parallelprojektion

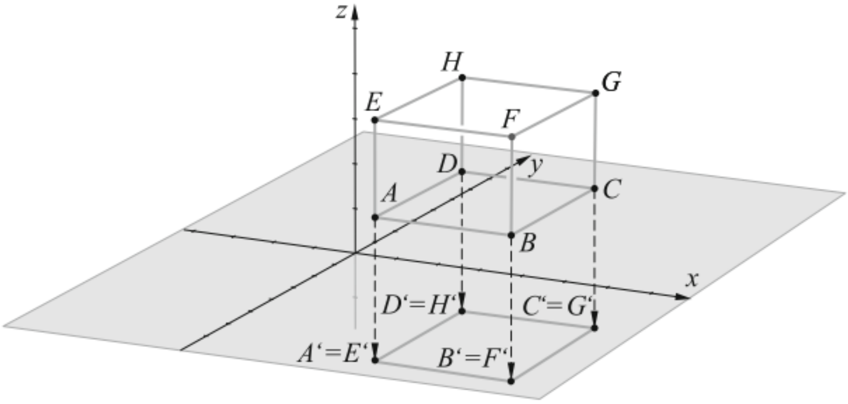
z
H
G
E
F
D
y
C
A
B
x
D‘=H‘
C‘=G‘
A‘=E‘
B‘=F‘

Figure 1.88: ela04.png

8) C:/DATA-NovoVivo-135-2020/Books-135-2020/Hyperbolic-Geometry/Geometrie zwischen Grundbegriffen und Grundvorstellungen - Jubiläumsband des Arbeitskreises Geometrie in der Gesellschaft für Didaktik der Mathematik 2015.pdf

Pag. 52 (62) - Aufgabe zur Symmetrie am Einheitskreis

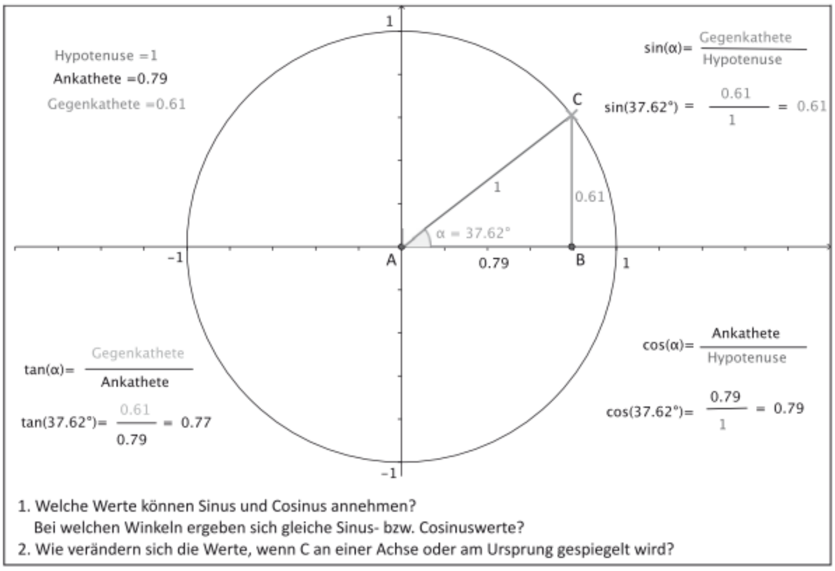
Hypotenuse =1
Ankathete =0.79
Gegenkathete =0.61
sin(α)= Gegenkathete / Hypotenuse
sin(37.62°) = 0.61 / 1 = 0.61
1
C
1
0.61
α = 37.62°
−1
A
0.79
B
1
tan(α)= Gegenkathete / Ankathete
tan(37.62°)= 0.61 / 0.79 = 0.77
cos(α)= Ankathete / Hypotenuse
cos(37.62°)= 0.79 / 1 = 0.79
−1

1. Welche Werte können Sinus und Cosinus annehmen?
   Bei welchen Winkeln ergeben sich gleiche Sinus- bzw. Cosinuswerte?
2. Wie verändern sich die Werte, wenn C an einer Achse oder am Ursprung gespiegelt wird?

Figure 1.89: geometrie01.png

Pag. 211 (216) - Diagonalen

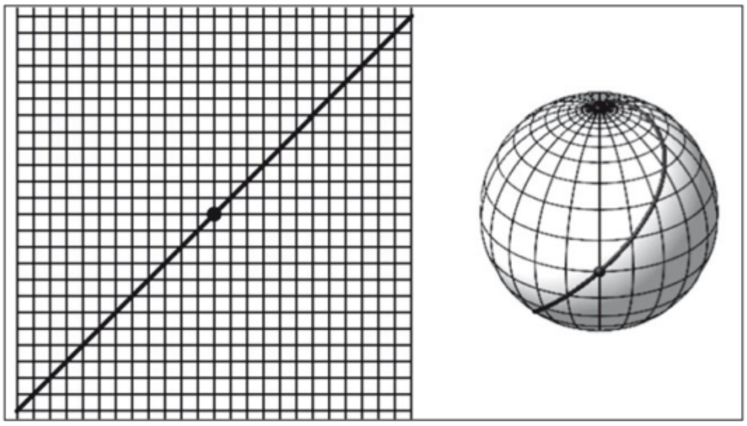

Figure 1.90: geometrie02.png

9) C:/DATA-NovoVivo-135-2020/Books-135-2020/Hyperbolic-Geometry/Didaktik der Analytischen Geometrie und Linearen Algebra - Algebraisch verstehen { Geometrisch veranschaulichen und anwenden 2015.pdf

Pag. 45 (56) - Darstellung der Lösungsmengen der Gleichungen eines LGS mit drei Gleichungen und drei Lösungsvariablen

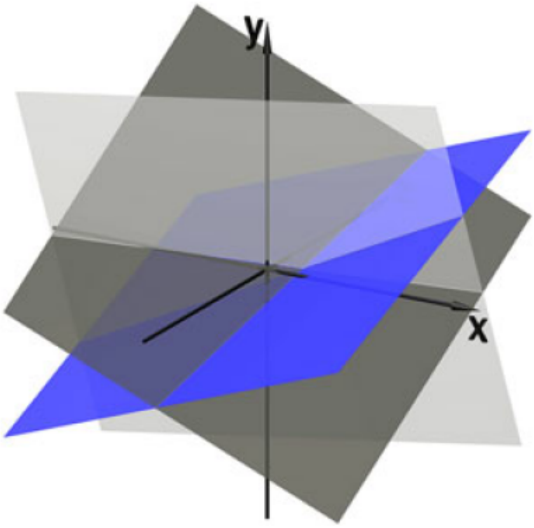

Figure 1.91: dida01.png

Pag. 50 (61) - Darstellung der Lösungsmengen der Gleichungen eines LGS mit drei Gleichungen und drei Lösungsvariablen nach den Lösungsschritten des Gauß'schen Algorithmus. Die mittelgrauen Ebenen stellen die Lösungsmengen der jeweils ersten Gleichung, die hellgrauen Ebenen die der zweiten Gleichung und die blauen Ebenen die der dritten Gleichung dar

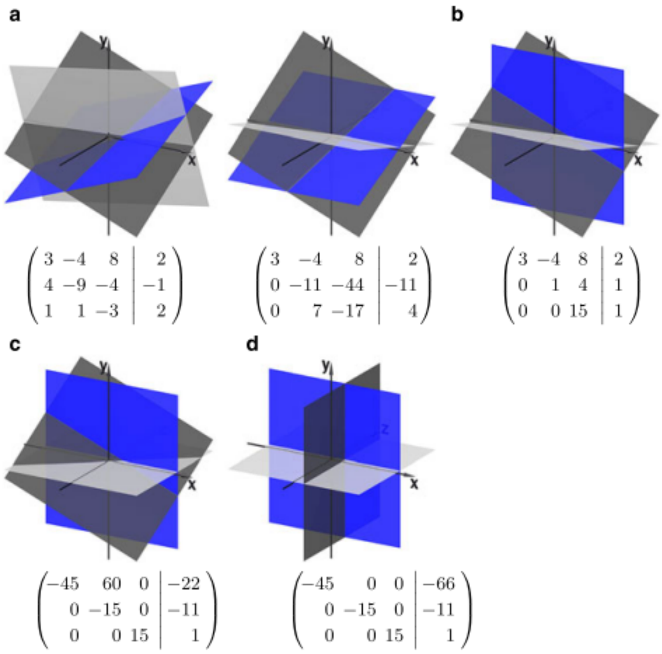
a
b
c
d
x
y
z
$\left(\begin{array}{ccc|c} 3 & -4 & 8 & 2 \\ 4 & -9 & -4 & -1 \\ 1 & 1 & -3 & 2 \end{array}\right)$
$\left(\begin{array}{ccc|c} 3 & -4 & 8 & 2 \\ 0 & -11 & -44 & -11 \\ 0 & 7 & -17 & 4 \end{array}\right)$
$\left(\begin{array}{ccc|c} 3 & -4 & 8 & 2 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 15 & 1 \end{array}\right)$
$\left(\begin{array}{ccc|c} -45 & 60 & 0 & -22 \\ 0 & -15 & 0 & -11 \\ 0 & 0 & 15 & 1 \end{array}\right)$
$\left(\begin{array}{ccc|c} -45 & 0 & 0 & -66 \\ 0 & -15 & 0 & -11 \\ 0 & 0 & 15 & 1 \end{array}\right)$

Figure 1.92: dida02.png

Pag. 105 (115) - Schwerpunkt eines Tetraeders

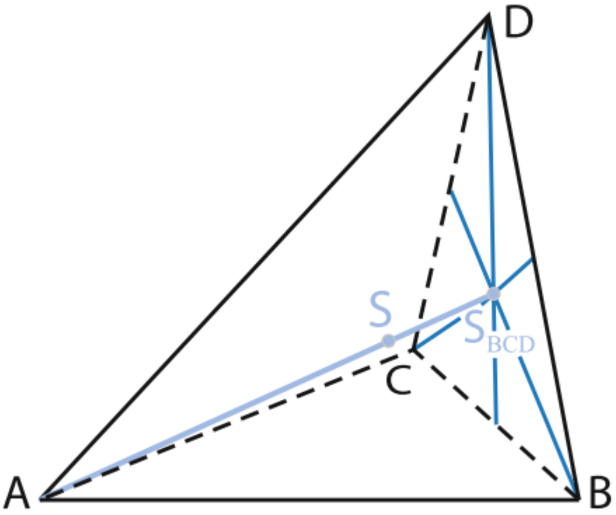

Figure 1.93: dida03.png

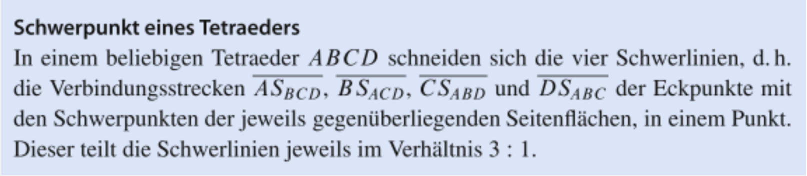

Schwerpunkt eines Tetraeders
In einem beliebigen Tetraeder $ABCD$ schneiden sich die vier Schwerlinien, d. h. die Verbindungsstrecken $\overline{AS_{BCD}}$, $\overline{BS_{ACD}}$, $\overline{CS_{ABD}}$ und $\overline{DS_{ABC}}$ der Eckpunkte mit den Schwerpunkten der jeweils gegenüberliegenden Seitenflächen, in einem Punkt. Dieser teilt die Schwerlinien jeweils im Verhältnis 3 : 1.

Figure 1.94: dida04.png

Pag. 126 (136) - Simultane Visualisierung des Assoziativgesetzes für Pfeilklassen und n-Tupel. Ein Applet mit einer entsprechenden interaktiven Visualisierung steht (wie auch entsprechende Applets zu anderen Rechengesetzen) auf der Internetseite zu diesem Buch zur Verfügung

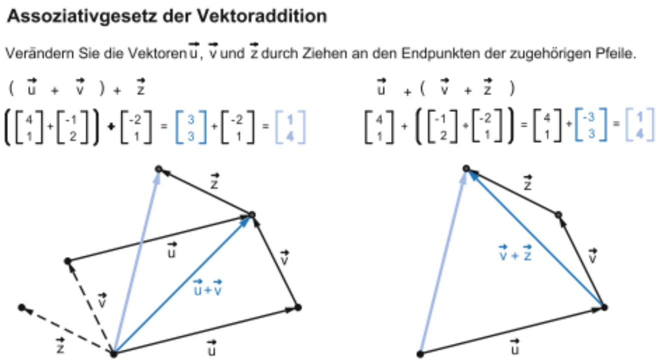
Assoziativgesetz der Vektoraddition

Verändern Sie die Vektoren $\vec{u}$, $\vec{v}$ und $\vec{z}$ durch Ziehen an den Endpunkten der zugehörigen Pfeile.

$(\vec{u} + \vec{v}) + \vec{z}$

$$\left(\begin{bmatrix}4\\1\end{bmatrix}+\begin{bmatrix}-1\\2\end{bmatrix}\right)+\begin{bmatrix}-2\\1\end{bmatrix}=\begin{bmatrix}3\\3\end{bmatrix}+\begin{bmatrix}-2\\1\end{bmatrix}=\begin{bmatrix}1\\4\end{bmatrix}$$

$\vec{u} + (\vec{v} + \vec{z})$

$$\begin{bmatrix}4\\1\end{bmatrix}+\left(\begin{bmatrix}-1\\2\end{bmatrix}+\begin{bmatrix}-2\\1\end{bmatrix}\right)=\begin{bmatrix}4\\1\end{bmatrix}+\begin{bmatrix}-3\\3\end{bmatrix}=\begin{bmatrix}1\\4\end{bmatrix}$$

Figure 1.95: dida05.png

Pag. 157(166) - Koordinatengleichungen und Parameterdarstellungen nichtlinearer Objekte

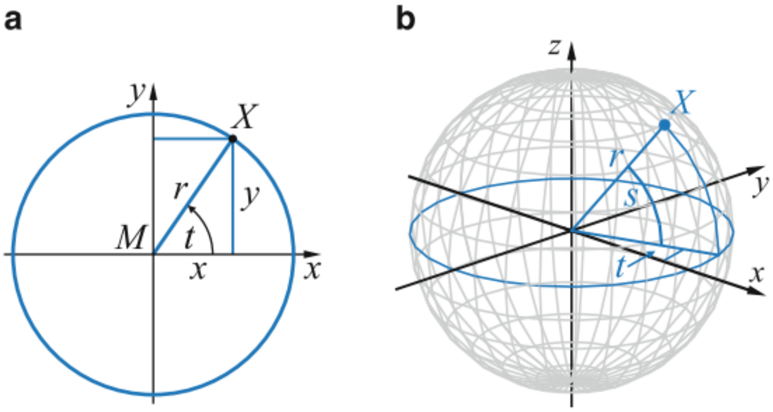

Figure 1.96: dida06.png

Pag. 218 (227) - Kreis- und Kugelgleichung (Equação de círculo e esfera)

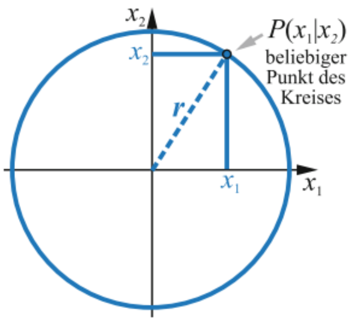
x_2
x_2
$P(x_1|x_2)$
beliebiger
Punkt des
Kreises
r
x_1
x_1

Figure 1.97: dida07.png

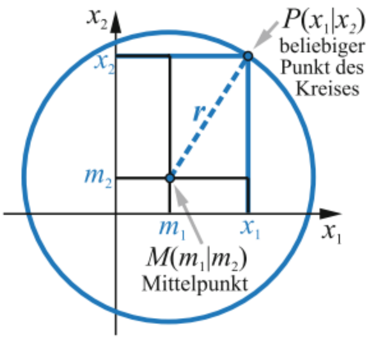
x_2
x_2
$P(x_1|x_2)$
beliebiger
Punkt des
Kreises
r
m_2
m_1
x_1
x_1
$M(m_1|m_2)$
Mittelpunkt

Figure 1.98: dida08.png

Pág. 220 (229) -

Unterstützende Graphiken zur Herleitung der Kugelgleichung (De apoio Gráficos para derivar o Equação esférica)

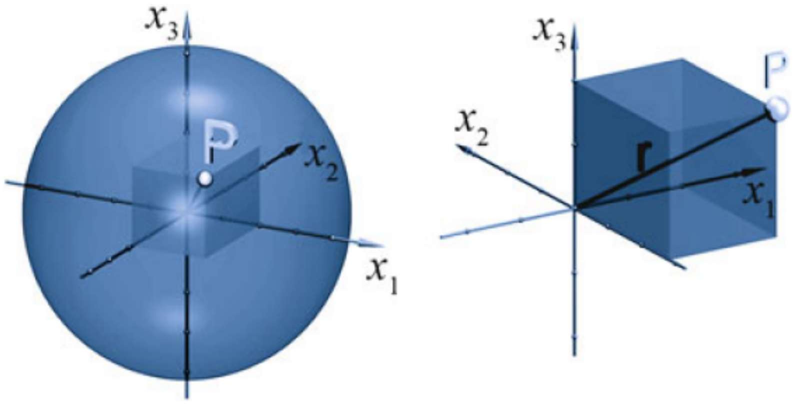
x_3
P
x_2
x_1
x_3
x_2
P
r
x_1

Figure 1.99: dida09.png

Pág. 224 (233) - Herleitung einer Gleichung für Tangentialebenen an Kugeln mit Mitteln der Differentialrechnung

Tangentialebene an einer Kugel

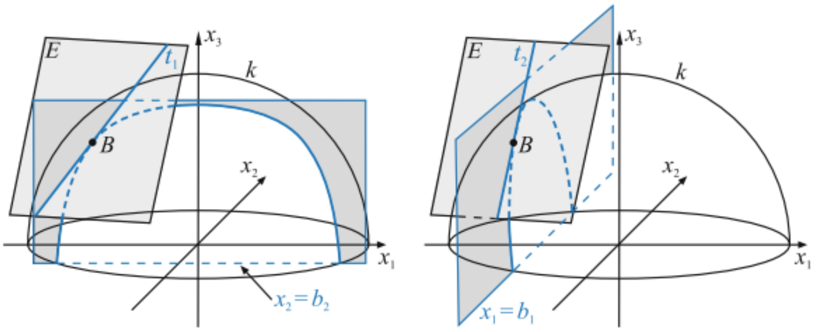
E
t_1
x_3
k
B
x_2
x_1
$x_2=b_2$
E
t_2
x_3
k
B
x_2
x_1
$x_1=b_1$

Figure 1.100: dida10.png

Pag. 228 (237) - Schnitte zweier Kugeln (Cortes de duas esferas)

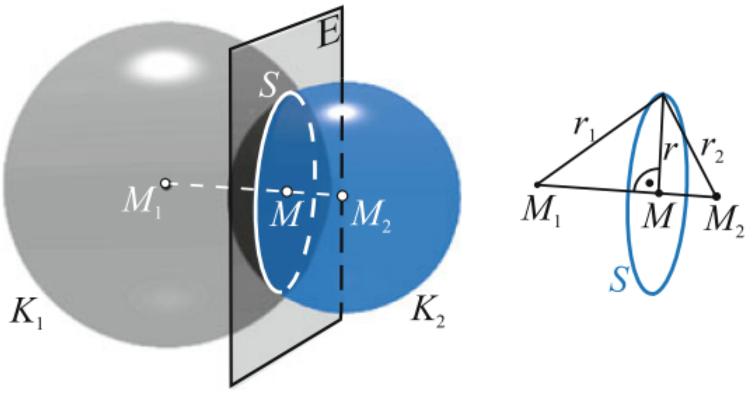
E
S
M_1
M
M_2
K_1
K_2
r_1
r
r_2
M_1
M
M_2
S

Figure 1.101: dida11.png

Pag. 299 (308) - Zylinderschnitte (Cortes de cilindro)

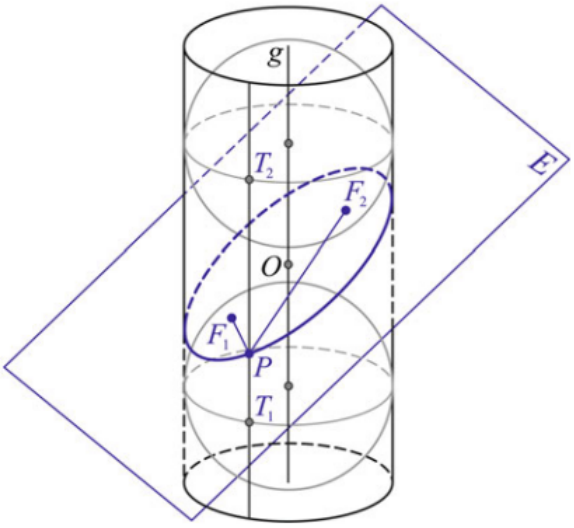
g
T_2
F_2
O
F_1
P
T_1
E

Figure 1.102: dida12.png

Pag. 318 (327) -
a) Parabeln und Sattelfläche (Parabolas e superfície de sela)

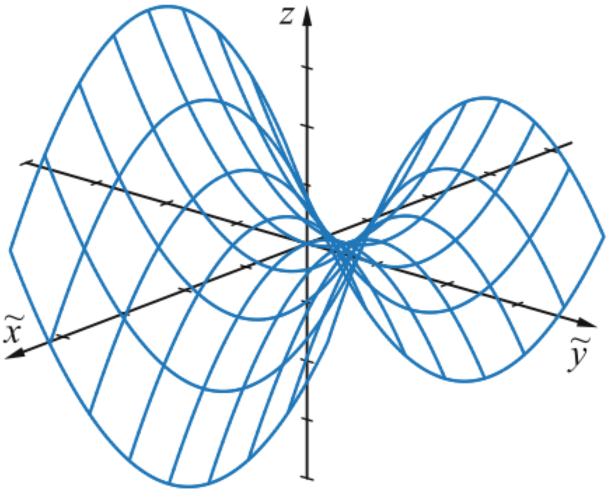
z
$\tilde{x}$
$\tilde{y}$

Figure 1.103: dida13.png

b) Erzeugende Parabeln (Parabolas geradoras)

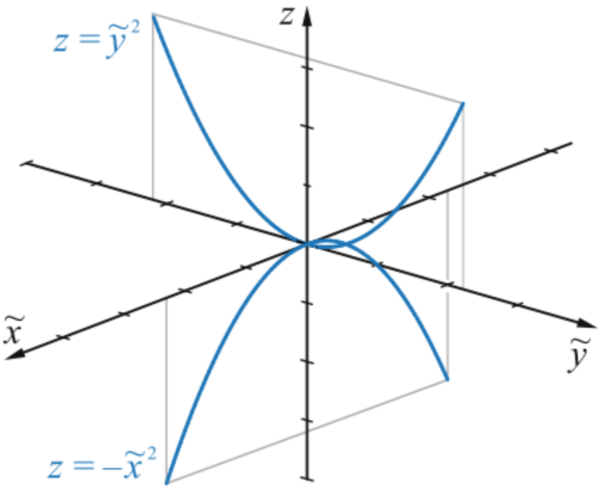
$z = \tilde{y}^2$
z
$\tilde{x}$
$\tilde{y}$
$z = -\tilde{x}^2$

Figure 1.104: dida14.png

Pag. 344 (353) - Untersuchung geradentreuer und nicht geradentreuer Abbildungen bereits in der Sekundarstufe I (Investigação mais direta e não mais direta Imagens já no nível secundário inferior)

Pag. 346 (355) -

a) Abbildung "a-durch" (Inversion),

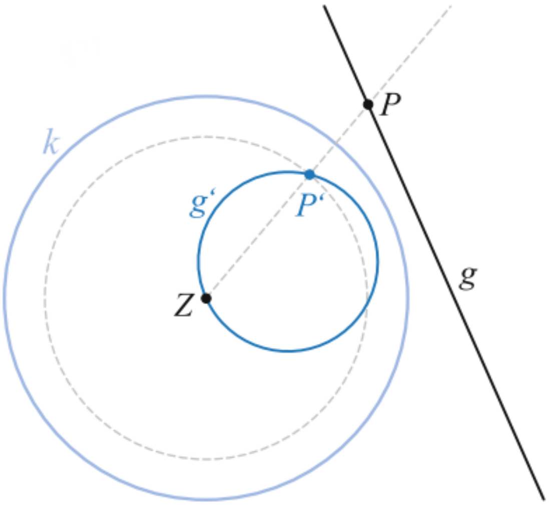

Figure 1.105: dida15.png

b) Abbildung "a-plus"

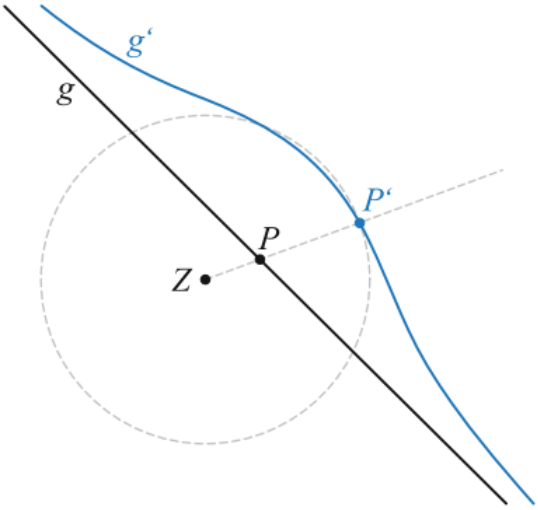

Figure 1.106: dida16.png

Pag. 351 (360) - Nicht geradentreue geometrische Abbildung (Mapeamento geometrico falso) - IMPORTANT

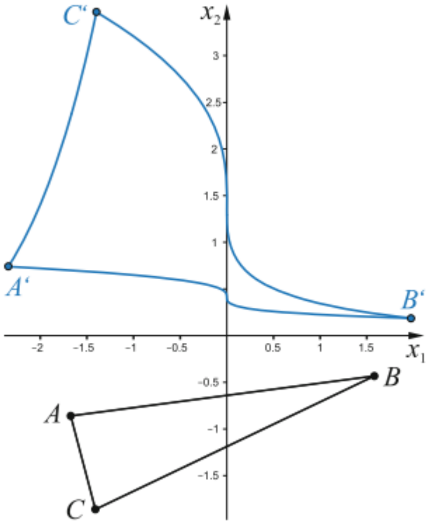
C‘
x_2
3
2.5
2
1.5
1
A‘
B‘
-2
-1.5
-1
-0.5
0.5
1
1.5
x_1
B
-0.5
A
-1
-1.5
C

Figure 1.107: dida17.png

1.5 Testing the Limit of Einstein's Theories

1) https://www.discovermagazine.com/the-sciences/testing-the-limit-of-einsteins-theories

Scientists finally have the technology to construct mind-boggling experiments that can tell us just how far relativity can be stretched. By Robert Kunzig September 30, 2004 2:00 AM

Einstein didn't care much about experiments. Of the three tests he proposed for general relativity, the first|that clocks should tick slower in a gravitational field|wasn't satisfied until after his death. Early experiments tended to contradict the prediction. His second prediction, that light from distant stars would be deflected by the warped space-time around the sun, catapulted him to world fame in 1919, when observations of a solar eclipse seemed to confirm his prediction. But as historians have since shown, the 1919 measurements were equivocal at best.

The one unequivocal verification of Einstein's theory during his lifetime was his explanation of a tiny anomaly in the orbit of Mercury. When he finally got that calculation to work, that was the only evidence he needed that space and time really were warped. \Nature had spoken to him," wrote biographer Abraham Pais. \He had to be right."

We now know he was. Just last year, for instance, radio signals transmitted from the Cassini spacecraft on its way to Saturn proved to be deflected by the sun by just the amount predicted by general relativity. Einstein would have been unimpressed. He believed his theory was correct because it was consistent, simple, and beautiful. \It is my conviction that pure mathematical construction enables us to discover the concepts and the laws connecting them, which give us the key to the understanding of the phenomena of Nature," he declared in 1933.

For better and worse, this aspect of Einstein's approach to physics is very much alive today: \Pure mathematical construction" is an industry that employs hundreds of workers. What they have been trying to construct is a unified theory of the four forces|gravity, electromagnetism, and the strong and weak nuclear forces. To build that, they first have to find a way of describing gravity, not in Einsteinian terms as a curvature of space-time but like the other three forces, as an exchange of force-bearing quanta. This quantum gravity theory would take over from general relativity in the extraordinarily tight quarters|the very core of a black hole, the very instant of the Big Bang|where relativity now predicts, absurdly, that space-time is infinitely curved. The final unified theory, in turn, would gratify most physicists' deep belief that the four forces really were equivalent once, in the first nanosecond after the Big Bang. That idea "is so beautiful that it has to be that way," explains physicist Clifford Will of Washington University. "That's an Einsteinian point of view."

Sure enough, there is as yet no experimental evidence to support this view—indeed, the improbability of ever reproducing black holes or big bangs in the laboratory has caused some researchers to wonder whether there will ever be any evidence. "Since the middle 1970s we've been in a situation in fundamental physics in which theory has run on, largely unchecked by experiments," says Lee Smolin, a theoretical physicist at the Perimeter Institute near Toronto. (Smolin's article on Einstein and his scientific legacy begins on page 36.) "And most of us think this has been very bad."

But lately Smolin and others have been finding grounds for optimism: The reign of the theorists seems to be coming to an end. The several schools of quantum gravity, physicists have realized, are slowly reaching consensus that incremental violations of relativity might be detectable even outside a black hole, and without building a particle accelerator the size of the universe. For that reason, and also because experimenters like to poke at the marble facades of grand theories, Einstein's ideas are being tested with increasing frequency these days and to extraordinary precision. What follows is a sampling of these experiments—just a few of the many now under way or on the drawing boards.

"I do not consider the main significance of the general theory of relativity to be the prediction of some tiny observable effects," wrote Einstein in 1930. Nonetheless, it was tiny observable effects that eventually proved that in certain situations Newton's physics needed to be replaced by Einstein's. In the same way it will be tiny observable effects that will one day prove where the limits of Einstein's genius lie. "People sometimes ask, 'Why have you felt you should persist in this way to perform a test of Einstein?' " says Stanford University physicist Francis Everitt, one of the leaders of a satellite experiment called Gravity Probe B, which was launched in April. "Aren't Einstein's theories all established and confirmed? After all, it's a hundred years ago that he developed his first theory of relativity. Don't we already know it all? And the answer is no." Some of the most important experiments under way may answer the questions that follow.

IS THE SPEED OF LIGHT CONSTANT?

Late in life Einstein couldn't even remember whether he'd been aware, in 1905, of the Michelson-Morley experiment, which showed that the speed of light is constant in all directions. Although we now think of the experiment as having shattered absolute space and time, that result was not at all what

Albert Michelson and Edward Morley expected when they did their experiment in 1887, in a basement of the Western Reserve University in Cleveland. They were trying to detect variations in the speed of light to prove the existence of the "luminiferous aether," a perfectly immobile, transparent, and mysterious substance that was thought to pervade the universe. Physicists had conceived of the ether as a medium that could transmit light waves, the way air or water transmits sound. The ether had to exist, but no one had seen it—it was a bit like dark matter today.

Because Earth orbits the sun at 18 miles per second, Michelson and Morley reasoned that they should be able to detect an ether wind blowing through their Cleveland basement with the help of an experimental setup called an interferometer. They split a light beam in two with a half-silvered mirror, sent the two halves of light off at right angles, and then bounced them off mirrors about five feet away. The two men expected that light waves traveling against the ether wind, in the same direction as Earth's motion, would be slowed and would arrive back at the starting point slightly later than the light waves traveling across the wind. When the two beams were recombined, the offset waves would interfere with each other to produce a distinctive pattern of light and dark bands. Michelson and Morley made their measurements with extraordinary care but saw no disruption in the pattern—the light beams traveled the same speed in all directions, impervious to any ether wind.

One null result, of course, did not rid the universe of the ether. Michelson went to his grave in 1931 convinced that it had to exist. But by then Einstein had changed most people's minds and persuaded them to accept the simple, beautiful truth: The velocity of light, c, really is different from any other velocity. Anyone who measures the speed of any light beam—or any other type of electromagnetic radiation—will get the same value for c: 299,792,458 meters per second. There is no absolute space, Einstein decided, ether-filled or otherwise. Seen in that light, the Michelson-Morley result made perfect sense.

And yet today theorists are questioning the absoluteness of c. Some versions of string theory, the most popular candidate for a unified theory, say there could be extremely feeble force fields left over from the Big Bang that point in different directions in different parts of space. An experiment to measure c might produce variations depending on how the setup was oriented with respect to one of those fields, variations far smaller than Michelson and Morley could have detected.

Several groups are looking for such variations with modern versions of the Michelson-Morley experiment. Peter Wolf, Sebastien Bize, and their colleagues at the Paris Observatory measure c with microwaves oscillating at 12 gigahertz inside a small sapphire crystal. Microwaves reflecting back and forth within the crystal line up and reinforce each other, or resonate|as long as they are moving precisely at c. If c were to change because the orientation of the crystal had changed with respect to some \preferred" direction of space, then the resonant frequency of the sapphire oscillator would change as well. The apparatus containing the crystal is bathed in liquid helium, chilling it to a few degrees above absolute zero to make sure that the crystal doesn't expand or contract by even a femtometer.

Over a period of months, as Earth spins on its axis and revolves around the sun, the Paris researchers monitor their oscillator, comparing it with the microwaves from a hydrogen maser (microwave laser), which shouldn't be affected by Earth's motion. \What we measure is that small frequency difference," says Bize. \We look for modulations that correlate with the motion of Earth."

Another group, based at Humboldt University in Berlin, uses a slightly different setup, comparing the outputs of a pair of sapphire oscillators. Over the past several years the two groups have achieved broadly comparable null results. \The speed of light in any two directions is the same to about one part in a quadrillion," says Holger Müller, a former member of the Berlin team who now works at Stanford. That's equivalent to knowing the U.S. gross national product to within a penny.

Over vast distances, even such slight variations could become meaningful. If two photons differed that much in velocity, and if they left a galaxy a billion light-years away at the same instant, they would arrive at Earth 30 seconds apart. A few years ago physicist Giovanni Amelino-Camelia at La Sapienza University in Rome had an idea for staging just such a race to test the constancy of c in a new way. Some theories of quantum gravity require space-time itself to be grainy - to be made up of discrete quanta, presumably around 10^{-35} of a meter across because that's the scale at which Einstein's field equations generate their bothersome infinities. (A proton is a hundred million trillion times bigger.) Amelino-Camelia calculated that light photons might navigate this cosmic foam at slightly different speeds, depending not on their direction - the possibility the Michelson-Morley experiments test for - but on their energy.

In 2007 or so, NASA plans to launch GLAST, the Gamma Ray Large Area Space Telescope. Its main purpose is to allow astronomers to study such events as gamma-ray bursts, which are mysteriously powerful explosions in distant galaxies. But it could also serve as Amelino-Camelia's finish line. All the photons in a burst, he reasons, must leave the starting blocks at about the same time. If you compare a lot of high-energy photons with a lot of relatively low-energy ones, you should find that on average, after a billion-year race, the high-energy ones reach GLAST's detector sooner - by about a millisecond. He and other quantum gravity theorists are pretty excited by that possibility, which just goes to show what they're up against.

WAS EINSTEIN RIGHT ABOUT FREE FALL?

In 1907, sitting at his desk in the patent office in Bern, Switzerland, Einstein imagined himself jumping off a house and realized that while falling he wouldn't feel his weight. It would have been more poetic to choose the Tower of Pisa for his thought experiment: Galileo, in his own legendary experiments there—truly legendary, because he probably never actually did them—had shown that objects of different mass and composition are equivalent in that they undergo the same gravitational acceleration. If they are dropped at the same time, they should hit the ground at the same time. Einstein took Galileo's equivalence principle much further. Einstein realized that if a freely falling physicist dropped something, it would stay right next to him. The falling man would have no way of telling (if he closed his eyes) that he was falling, no awareness of gravity at all; that would also go for every instrument in the lab he brought with him. Einstein eventually concluded that gravity was a curvature of space-time, not something that emanated from the masses themselves.

But was Galileo himself right? Newton, among others, tested the equivalence principle. Apparently unbeknownst to Einstein in 1907, a Hungarian baron named Roland von Eötvös had recently completed updated, highly accurate tests using a device called a torsion pendulum. Modern versions of the Eötvös experiment continue to this day in a number of labs. Apollo 15 commander David Scott carried out a less accurate but far more colorful demonstration at the end of his last moonwalk in 1971, when he dropped a geologist's hammer and a falcon feather from shoulder height. A live TV audience watched them hit the airless moon at the same time. After the Apollo missions, physicists on Earth started bouncing laser beams off reflectors that the astronauts had left on the moon, measuring the Earth–moon distance with increasing precision. Although Earth and the moon are made of different stuff—the moon lacks an iron core—the measurements so far reveal that they are “falling” toward the sun with exactly the same acceleration.

Lunar laser ranging and the modern Eötvös experiments have confirmed the equivalence principle to within less than one part in a trillion. Pierre Touboul, a physicist at ONERA, the French aerospace research agency, thinks he can do a hundred to a thousand times better. Touboul is the project scientist for MICROSCOPE, a microsatellite to be launched in 2007 by the European Space Agency. Its purpose is to conduct Galileo's free-fall experiment in space, with test masses of different composition. In space, Touboul explains, you don't need to worry about tiny fluctuations in the gravitational field outside your lab that might throw off your results. Like the moon, orbiting test masses fall for a long time, allowing you to drown out noise with data.

The test masses on MICROSCOPE will be two concentric cylinders of titanium and platinum. (A control experiment will use two platinum cylinders.) The cylinders will be sandwiched between electrodes of gold-plated silicon but also separated from them by thin spaces of vacuum|the test masses will float on the electrodes' electrostatic field. Loose lengths of wire five micrometers wide, which will remove any electric charge delivered by cosmic rays, will be their only physical contact with the rest of the satellite. Thrusters on the satellite will compensate for the small drag exerted by the outer wisps of the atmosphere. The only force the test masses will feel is Earth's gravity. Any differences in the electrostatic field needed to keep them in the same orbit will be supplied by the electrostatic levitation.

Instead of trying to determine whether the two masses fall differently, one dipping toward Earth slightly more, Touboul explains that it is easier and more precise to measure the electrostatic forces required to maintain them at exactly the same orbit. The forces are just a function of the current supplied to the electrodes. \If the forces are identical, there's no violation of the equivalence principle," says Touboul. MICROSCOPE will keep the test mass orbits identical to within 100 billionths of a meter, or a tenth the width of an atom. That should allow it to detect violations of the equivalence principle down to one part in a quadrillion. Some string theorists have predicted bigger violations than that.

\From a scientific point of view, finding no violation is just as interesting as finding one," Touboul says. \But my innermost conviction is that there will certainly be a violation. We just don't know at what level."

DOES EARTH WARP AND TWIST SPACE-TIME?

Eight years before he completed general relativity, Einstein made a weird prediction: Gravity should slow the passage of time and stretch light waves toward the red end of the spectrum. For years people tried and failed to see this gravitational redshift in the spectrum of sunlight. It wasn't precisely confirmed until 1976, when a Scout D rocket carried a hydrogen-maser clock to an altitude of 10,000 kilometers, or about 6,000 miles. In the two hours before it plunged into the Atlantic a thousand miles east of Bermuda, the rocket's timekeeping was compared with that of an identical clock on Earth. A hydrogen maser emits microwaves at 1.42 gigahertz; at the rocket's apogee the clock on board was about one hertz faster than its terrestrial twin. Einstein was vindicated at last, with a precision of a few parts in a hundred thousand: The stronger the gravity, the slower the clock.

That mission was called Gravity Probe A. Since then atomic clocks have progressed tremendously, and a European mission called ACES, for Atomic

Clock Ensemble in Space, may put one on the space station in 2007. Its laser-cooled cesium clock is supposed to measure the gravitational redshift 25 times more precisely than Gravity Probe A; at the altitude of the space station, it should gain about a second in 700 years. And last April NASA finally launched Gravity Probe B, which had been on the drawing boards since 1961. It is designed to measure the curvature of space-time and how the spinning Earth affects it.

The bulk of Gravity Probe B is a thermos filled with 645 gallons of liquid helium. At its heart are four absurdly sensitive gyroscopes. Each one is a quartz sphere the size of a Ping-Pong ball, smooth to within a millionth of an inch and coated with the metal niobium, which is superconducting at liquid helium temperatures. As the sphere floats in an electrostatic field and spins at 10,000 revolutions per minute, a current in the niobium will generate a magnetic field that will be perfectly aligned with the sphere's otherwise invisible spin axis—and be measurable by sensitive detectors.

By early summer the satellite's telescope had locked onto a distant guide star, IM Pegasi; all four spheres are now being aligned with this star. For the next year the Gravity Probe B team, led by Francis Everitt of Stanford, will monitor the spheres to see if the warp or drag of space tilts their spin axes out of alignment. “It's going to be a very dull experiment," Everitt says.

Yet it has been a somewhat controversial one. The mission didn't seem worth its 700 million cost to some physicists, who felt that during its 40-year gestation Gravity Probe B had been overtaken by other research. The measurement of space curvature made with the Cassini spacecraft, for instance, is already several times more accurate than Gravity Probe B's stated goal of 0.01 percent accuracy. (Everitt's team hopes to exceed that.)

But Gravity Probe B will also do something more important. It will quantify how the spinning Earth stirs space-time like molasses—a relativistic effect called frame dragging that has never been measured directly. Around Earth the phenomenon is a curiosity; to see it, Everitt's team will have to detect a deflection of their gyroscope axes of only 42 thousandths of an arc second. (That's an angle, the team's literature explains, the width of a human hair seen at a distance of a quarter mile.) Around spinning black holes, however, frame dragging could be hugely important: By whipping magnetic field lines through the electrically charged gas around the holes, it could convert them into electromagnetic generators, which would explain how they spew jets of energetic particles millions of light-years into space. “What Gravity Probe B will do for us is verify for the first time that this effect exists," says Caltech gravitational theorist Kip Thorne. It will also verify if general relativity can accurately predict the magnitude of the effect.

DO GRAVITATIONAL WAVES ROIL THE UNIVERSE?

If you need any proof of the complete faith modern physicists have in relativity, you can find it in Livingston, Louisiana. Or in Hanford, Washington. Or outside Pisa, Italy. In all those places physicists have built novel observatories costing hundreds of millions of dollars each, with arms extending several miles in perpendicular directions. The observatories are designed to detect a kind of radiation that has never been observed but that general relativity tells the physicists must exist: gravitational waves.

Gravitational waves are ripples in the space-time sea; Einstein predicted that they should exist as early as 1916. They spread out in all directions when something massive in the cosmos makes a big splash. The best sources physicists can think of are two black holes or two neutron stars spiraling toward each other and then colliding. In fact, although no one has seen a gravitational wave, a Nobel Prize has already been awarded for work that demonstrates their reality. Astronomers Joseph Taylor and Russell Hulse showed that two orbiting pulsars|neutron stars that emit beams of radiation as they spin|are slowing down by an amount that corresponds exactly to the energy that relativity predicts they should be emitting as gravitational waves. Still, it would be nice to detect a wave. \It would close the books," says physicist Peter Saulson, who works on the two American observatories, which are called LIGO, for Laser Interferometer Gravitational Wave Observatory. \It's a sanity check."

The observatories now coming online|the two LIGOs are operating but still debugging; the European one, VIRGO, will turn on next year|are all Michelson interferometers. An infrared laser beam is split in two; one half travels down one two-and-a-half-mile vacuum tunnel, and the other half travels down a second tunnel at a right angle to the first. Each beam reflects off a mirror at the end of the tunnel and returns to the starting point, where the two beams interfere. Changes in the interference pattern, which Michelson had hoped would show variations in the speed of light, now serve to indicate changes in the length of the beams. These changes, scientists predict, will be caused by a gravitational wave that stretches and squeezes space-time as it passes through the instruments. The observatories are designed to detect changes as small as a hundred millionth the size of a hydrogen atom. A gravitational wave is so difficult to distinguish from other vibrations that physicists won't believe in a wave's existence unless at least two observatories see the same minuscule shift.

If they do see some waves, they will be able to test general relativity in several ways. Relativity predicts exactly how much stretching and squeezing a gravitational wave should do. It predicts that gravitational waves should travel at the speed of light. It predicts in some detail the space-time turbulence that would result when two black holes collide.

The problem is, no one really knows how often such events occur. The observatories are designed to be sensitive to waves from as far away as the Virgo cluster of galaxies 65 million light-years away. But that may not be enough of a range. Even before LIGO reaches its design sensitivity, its leaders have requested funding to make it 10 times more sensitive still. On the other hand, physicists are confident enough of eventual success that they have persuaded NASA and the European Space Agency to start planning for LISA, for Laser Interferometer Space Antenna, a space-based interferometer consisting of three satellites, now tentatively scheduled for launch in 2012.

The first detection of a gravitational wave, says Clifford Will, \will be an exciting moment. It will legitimize this whole thing."

WILL WE FIND THAT EINSTEIN GOT RELATIVITY JUST A WEE BIT WRONG?

It's just a matter of time, most physicists think, before Einstein fails. Relativity touches so much of physics that a violation could show up almost anywhere. In addition to all the experiments already mentioned, it might come from quasars in the far reaches of the universe. A team of astronomers led by John Webb of the University of New South Wales has been measuring how the light from quasars is absorbed by gas clouds that lie between them and us but are still billions of light-years away, and thus did their absorbing billions of years ago. The frequencies of light absorbed by an atom are governed by a physical property called the fine-structure constant. By comparing their gas cloud data with equivalent measurements made in laboratories on Earth, Webb and his colleagues have hypothesized that the fine-structure constant today is about a thousandth of a percent larger than it was 8 billion years ago. If so, that would clearly violate the equivalence principle|the laws of physics are not supposed to change over time. Recently, though, another team of astronomers failed to confirm Webb's observations. The jury is still out.

A violation of relativity may also show up in observations of cosmic rays, the energetic particles that bombard Earth from all directions. Calculations based on special relativity predict that ultrahigh-energy cosmic rays should reach Earth only rarely, because they lose energy in space. A Japanese observatory has seen more than expected, and theorists such as Amelino-Camelia think this excess might be the first evidence of a space-time graininess that would ease the passage of high-energy particles. A more powerful observatory now beginning work in Argentina may resolve the issue.

Finally, there's that nagging little question of what the universe is made of. The concept of dark matter has been around so long now, and the idea that it makes up much of the universe has been repeated so often that one might almost forget that it has never been detected. One of the main reasons people believe that galaxies have halos of dark matter is that the galaxies rotate faster in their outer reaches than the gravity exerted by their visible matter would seem to permit. But an alternative theory, called modified Newtonian dynamics, or MOND, does an even better job of explaining galaxy rotation by simply tinkering with Newton's law of gravitation|and thus with Einstein's.

It's an ugly fix, and unlike dark matter it's not connected to a larger theory of particle physics, so theorists don't like it. And yet, says Lee Smolin, \a growing number of people from different specialties are telling me they are waking up in the middle of the night worrying about how well MOND does. I hope the dark-matter models in the end do better. If not, we are seeing massive violations of the Einstein equations right in front of us."

Wouldn't that be fun? Wouldn't it be fun, in fact, to see the old clown-haired saint taken down a nano-peg by a few tiny observable effects? Maybe we'd hear a bit less of that \subtle is the Lord" and \God does not play dice" nonsense. \There are two kinds of psychology in physics," says Saulson. \One is people who love order and beauty and hope the theories are right. The other is people who think theorists are full of baloney and who hope we'll find something amazing when we turn on our experiments." With a little luck, the second group is about to have its turn.

2) https://solarsystem.nasa.gov/news/12249/saturn-bound-spacecraft-tests-einsteins-theory/

Saturn-Bound Spacecraft Tests Einstein's Theory

An experiment by Italian scientists using data from NASA's Cassini spacecraft, currently en route to Saturn, confirms Einstein's theory of general relativity with a precision that is 50 times greater than previous measurements.

The findings appear in the Sept. 25 issue of the journal Nature. They are part of a scientific collaboration between NASA and the Italian Space Agency. The experiment took place in the summer of 2002, when the spacecraft and Earth were on opposite sides of the Sun separated by a distance of more than 1 billion kilometers (approximately 621 million miles).

Researchers observed the frequency shift of radio waves to and from the spacecraft as the waves passed near the Sun. They precisely measured the change in the round-trip light time of the radio signal as it traveled close to the Sun. The round-trip light time is the time it takes the signal transmitted from the Deep Space Network station in Goldstone Calif., to the spacecraft on the other side of the Sun and back traveling at the speed of light.

"The scientific significance of these results is the important confirmation of the theory of general relativity and the agreement with Einstein's formulations to an unprecedented experimental accuracy," said Sami Asmar, manager of the Radio Science Group, which acquired the data for this experiment at NASA's Jet Propulsion Laboratory in Pasadena, Calif. "The technological significance of the experiment is the ability to overcome the harsh solar environment on radio links."

The researchers measured how much the Sun's gravity bent an electromagnetic beam, in this case the radio signal transmitted by the spacecraft and received by the ground stations.

According to the theory of general relativity, a massive object like the Sun causes space-time to curve, and a beam of radio waves (or light) that passes by the Sun has to travel further because of the curvature. The extra distance that the radio waves travel from Cassini past the Sun to the Earth delays their arrival; the amount of the delay provides a sensitive test of the predictions of Einstein's theory. Although deviations from general relativity are expected in some cosmological models, none were found in this experiment.

Tests of general relativity have important cosmological implications. The question is not whether general relativity is true or false, but at which level of accuracy it ceases to describe gravity in a realistic way.

Past tests of general relativity confirmed Einstein's prediction to an accuracy of one part per thousand. This accuracy was achieved back in 1979 using the Viking landers on Mars. The Cassini experiment confirmed it to an accuracy of 20 parts per million. The key to this improvement has been the adoption of novel technologies in space telecommunications.

The experiment could not have been conducted to this level of accuracy in the past because of noise on the radio link introduced by the solar corona. With the Cassini experiment, this hindrance was overcome by fitting the spacecraft communication system with multiple links at different frequencies. This new capability on the Cassini spacecraft and on the 34-meter (112 foot) diameter antenna at Goldstone, allowed scientists to remove the effects of the interplanetary and solar plasma from the radio data. In addition, the noise from Earth's atmosphere was strongly reduced by special equipment installed at the Goldstone complex. These technological breakthroughs developed for the Cassini mission have led to unprecedented accuracies in the velocity measurements with benefits for future scientific experiments as well as deep space navigation.

The experiments are part of a series of radio science experiments planned for the cruise phase of the mission, including the search for low frequency gravitational waves.

Cassini will begin orbiting Saturn on July 1, 2004, and release its piggybacked Huygens probe about six months later for descent through the thick atmosphere of the moon Titan.

Cassini-Huygens is a cooperative mission of NASA, the European Space Agency and the Italian Space Agency. JPL, a division of Caltech, manages the mission for NASA's Office of Space Science, Washington, D.C. Authors of the Nature paper, "A New Test of General Relativity With the Cassini Space Mission," are Dr. Bruno Bertotti of the University of Pavia, Italy; Dr Luciano Iess of the University of Rome "La Sapienza", Italy; and Dr. Paolo Tortora of the University of Bologna, Italy. IMPORTANT

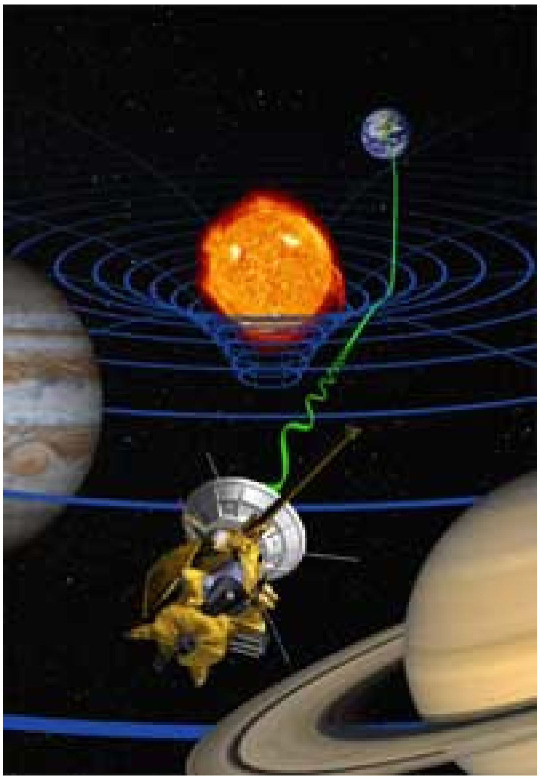

Figure 1.108: edt01.jpg

3) https://ocw.mit.edu/courses/physics/8-962-general-relativity-spring-2020/

4) http://mphitchman.com/geometry/section1-3.html
Geometry with an Introduction to Cosmic Topology
Michael P. Hitchman
1.3 Geometry on Surfaces: A First Look

Think for a minute about the space we live in. Think about objects that live in our space. Do the features of objects change when they move around in our space? If I pick up this paper and move it across the room, will it shrink? Will it become a broom?

If you draw a triangle on this page, the angles of the triangle will add to 180o. In fact, any triangle drawn anywhere on the page has this property. Euclidean geometry on this flat page (a portion of the plane) is homogeneous: the local geometry of the plane is the same at all points. Our three-dimensional space appears to be homogeneous as well. This is nice, for it means that if we buy a 5 ft3 freezer at the appliance store, it doesn't shrink to 0.5 ft3

when we get it home. A sphere is another example of a homogenous surface. A two-dimensional bug living on the surface of a sphere could not tell the difference (geometrically) between any two points on the sphere.

The surface of a donut in three-dimensional space is not homogeneous, and a two-dimensional bug living on this surface could tell the difference between various points. One approach to discovering differences in geometry involves triangles.

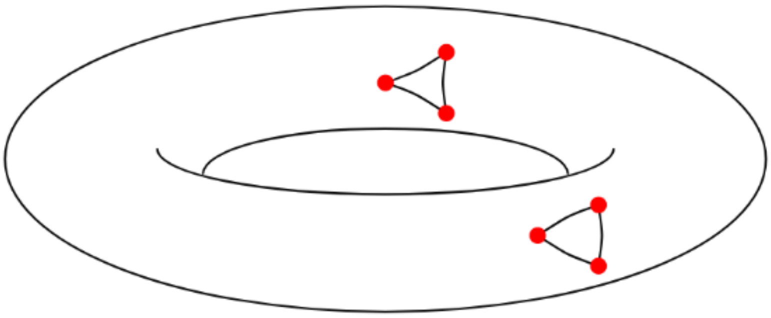

Figure 1.109: edt02.png - This torus surface is not homogeneous.

It is an important matter to decide what we mean, exactly, by a triangle on a surface. A triangle consists of three points and three edges connecting these points. An edge connecting point A to point B is drawn to represent the path of shortest distance between A and B. Such a path is called a geodesic. For the two-dimensional bug, a \straight line" from A to B is simply the shortest path from A to B.

On a sphere, geodesics follow great circles. A great circle is a circle drawn on the surface of the sphere whose center (in three-dimensional space) corresponds to the center of the sphere. Put another way, a great circle is a circle of maximum diameter drawn on the sphere. The circles a and b in are great circles, but circle c is not.

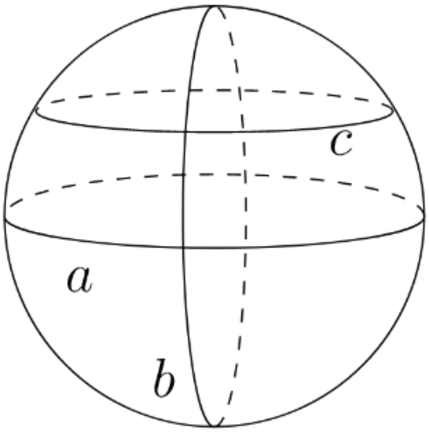

Figure 1.110: edt03.png - Geodesics on the sphere are great circles.

In the Euclidean plane, geodesics are Euclidean lines. One way to determine a geodesic on a surface physically is to pin some string at A and draw the string tight on the surface to a point B. The taut string will follow the geodesic from A to B. In we have drawn geodesic triangles on three different surfaces.

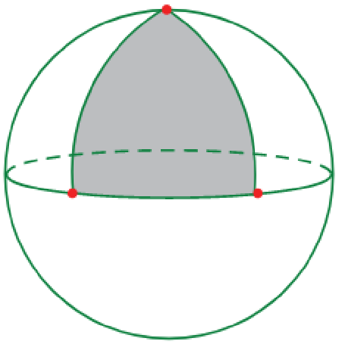

Figure 1.111: et04.png

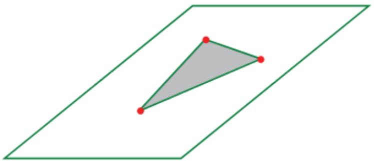

Figure 1.112: et05.png

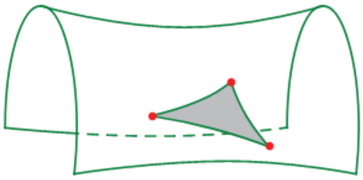

Figure 1.113: et06.png

Figure 1.3.3 Depending on the shape of the surface, geodesic triangles can have angle sum greater than, less than, or equal to 180o.

Getting back to the donut, a two-dimensional bug could use triangles to tell the difference between a \convex" point on an outer wall, and a \saddle-shaped" point on an inner wall. A bug could draw a triangle about the convex point, determine the angle sum, and then move around the surface to a saddle-shaped point, and determine the angle sum of a new triangle (whose legs are the same length as before). The bug would scratch her head at the different angle sums before realizing she'd stumbled upon something big. She'd go home, write up the result, emphasizing the fact that a triangle in the first \convex" region will have angle sum greater than 180o, while a triangle in the \saddle-shaped" region will have angle sum less than 180o.

This happy bug will conclude her donut surface is not homogeneous. She will then sit back and watch the accolades pour in. Perhaps even a Nobel prize. Thus, small triangles and their angles can help a two-dimensional bug distinguish points on a surface.

The donut surface is not homogeneous, so let's build one that is. Example 1.3.4 The Flat Torus

Consider again the world of Figure 1.1.2. This world is called a flat torus. At every spot in this world, the pilot of the ship would report flat surroundings (triangle angles add to 180o). Unlike the donut surface living in three dimensions, the flat torus is homogeneous. Locally, geometry is the same at every point, and thanks to a triangle check, this geometry is Euclidean. But the world as a whole is much different than the Euclidean plane. For instance, if the pilot of the ship has a powerful enough telescope, he'd be able to see the back of his ship. Of course, if the ship had windows just so, he'd be able to see the back of his head. The flat torus is a finite, Euclidean two-dimensional world without any boundary.

Example 1.3.5 Coneland

Here we build cones from flat wedges, and measure angles of some triangles.

a. Begin with a circular disk with a wedge removed, like a pizza missing a slice or two. Joining the two radial edges produces a cone. Try it with a cone of your own to make sure it works. Now, with the cone flat again, pick three points, labeled A, B, and C, such that C is on the radial edge. This means that in this flattened version of the cone, point C actually appears twice: once on each radial edge, as in . These two representatives for C should get identified when you join the radial edges.

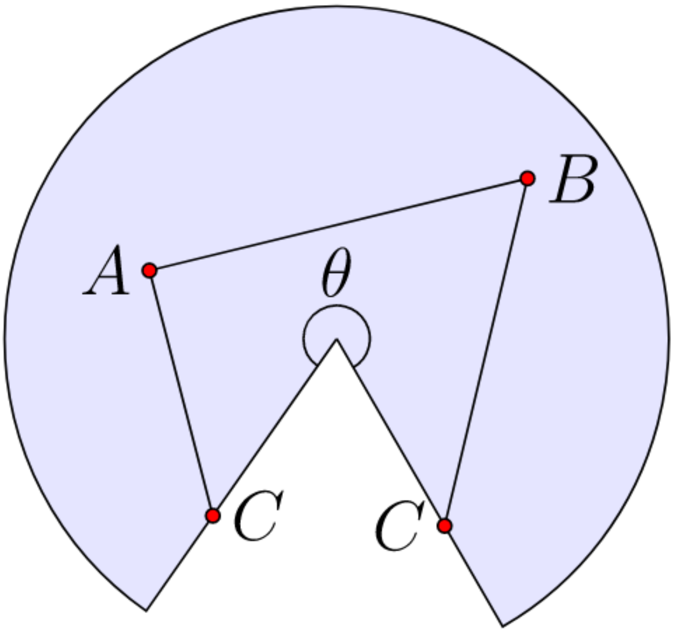

Figure 1.114: edt07.png - A triangle on a cone.

b. Draw the segments connecting the three points. You should get a triangle with the tip of the cone in its interior. (This triangle should actually look like a triangle if you re-form the cone.) If you don't get the tip of the cone on the inside of the triangle, adjust the points accordingly.

c. With your protractor, carefully measure the angle subtended by the circular sector. To emphasize 's role in the shape of the cone, we let S)(denote the cone surface determined by.

d. With your protractor, carefully measure the three angles of your triangle. The angle at point C is the sum of the angles formed by the triangle legs and the radial segments. Let denote the sum of these three angles.

e. State a conjecture about the relationship between the angle and , the sum of the angles of the triangle. Your conjecture can be in the form of an equation. Then prove your conjecture. Hint: if you draw a segment connecting the 2 copies of point C, what is the angle sum of the quadrilateral ABCC?

Example 1.3.7 Saddleland

Repeat the previous exercise but with circle wedges having > 2 .

Identifying the radial edges in this case produces a saddle-shaped surface. [To create such a circle wedge we can tape together two wedges of equal radius. One idea: Start with a disk with one radial cut, and a wedge of equal radius. Tape one radial edge of the wedge to one of the slit radial edges of the disk. Then, identifying the other radial edges should produce Saddleland.]

Remember, a homogeneous surface is a space that has the same local geometry at every point. Our flat torus is homogeneous, having Euclidean geometry at every point. However, our cones S)(in the previous exercises are not homogeneous (unless happens to be 2). If a triangle in S( ) does not contain the tip of the cone in its interior, then the angles of the triangle will add to radians, but if the triangle does contain the tip of the cone in its interior, then the angle sum will not be radians. A two-dimensional bug, then, could conclude that S() is not homogeneous.

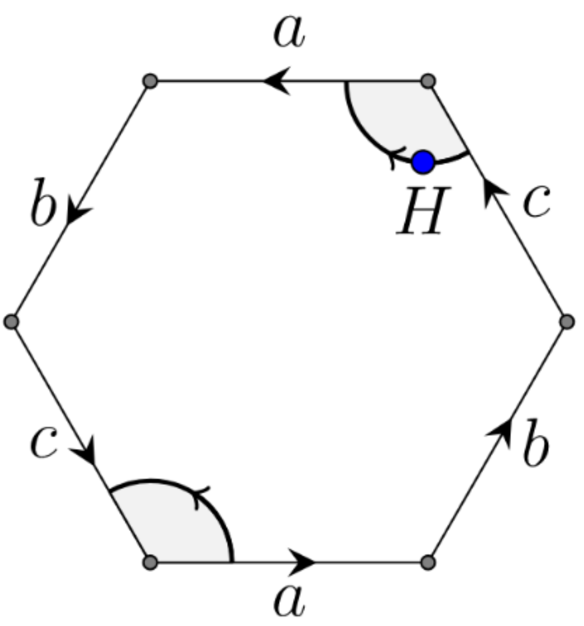

Figure 1.115: edt08.png - A hexagonal video screen

Example 1.3.9 A non-Euclidean surface

Consider the surface obtained by identifying the edges of the hexagon as indicated in image (A hexagonal video screen). In particular, the edges are matched according to their labels and arrow orientation. So, if a ship flies off the hexagonal screen at a spot on the edge marked a, say, then it reappears at the matching spot on the other edge marked a.

Suppose the pilot of a ship wants to fly around one of the corners of the hexagon. If she begins at point H, say, and flies counterclockwise around the upper right corner as indicated in the diagram, she would fly off the screen at the top near the start of an a edge. So, as she made her journey, she would reappear in the lower left corner near the start of the other a edge. Continuing around she would complete her journey after circling this second corner.

However, the angle of each corner is 120o, and gluing them together will create a cone point, as pictured below. Similarly, she would find that the other corners of the hexagon meet in groups of two, creating two additional cone points. As with the Coneland , the pilot can distinguish a corner point from an interior point here. She can look at triangles: a triangle containing one of the cone points will have angle sum greater than 180o; any other triangle will have angle sum equal to 180o.

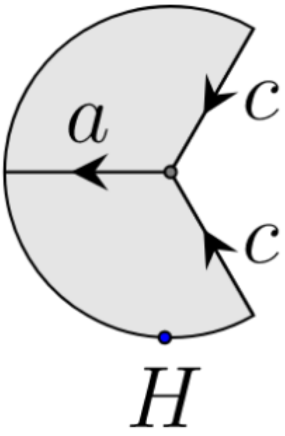

Figure 1.116: edt09.png - A hexagonal video screen

So the surface is not homogeneous, if it is drawn in the plane. However, the surface does admit a homogeneous geometry. We can get rid of the cone points if we can increase each corner angle of the hexagon to 180o. Then, two corners would come together to form a perfect 360o patch about the point.

But how can we increase the corner angles? Put the hexagon on the sphere! Imagine stretching the hexagon onto the northern hemisphere of a sphere. In this case we can think of the 6 points of our hexagon as lying on the equator. Then each corner angle is 180o, each edge is still a line (geodesic), and when we glue the edges, each pair of corner angles adds up to exactly 360o, so the surface is homogeneous. The homogeneous geometry of this surface is the geometry of the sphere (elliptic geometry), not the geometry of the plane (Euclidean geometry).

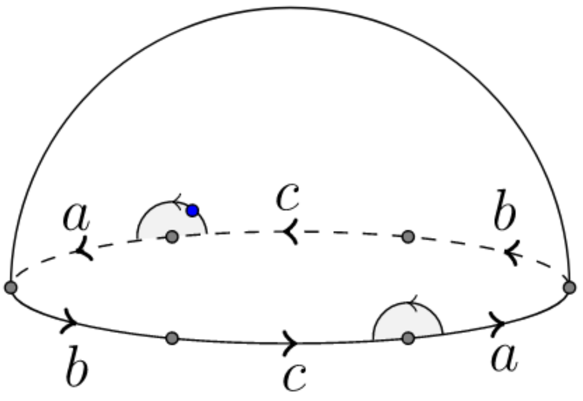

Figure 1.117: edt10.png - A surface with homogeneous elliptic geometry.

It turns out every surface can be given one of three types of homogeneous geometry: Euclidean, hyperbolic, or elliptic. We will return to the geometry of surfaces (and of our universe) after we develop hyperbolic and elliptic geometry. If it doesn't make a whole lot of sense right now, don't sweat it, but please use these facts as motivation for learning about these non-Euclidean geometries.

Exercises

1. Work through the Coneland - Example 1.3.5.

2. Work through the Saddleland - Example 1.3.7.

3. Circumference vs Radius in Coneland and Saddleland. In addition to triangles, a two-dimensional bug can use circles to screen for different geometries. In particular, a bug can study the relationship between the radius and the circumference of a circle. To make sure we think like the bug, here's how we define a circle on a surface: Given a point P on the surface, and a real number $r > 0$, the circle centered at P with radius r is the set of all points r units away from P, where the distance between two points is the length of the shortest path connecting them (the geodesic).

a. Pick your favorite circle in the plane. What is the relationship between the circle's radius and circumference? Is your answer true for any circle in the plane?

b. Consider the Coneland surface of Example 1.3.5. Construct a circle centered at the tip of the cone and derive a relationship between its circumference and its radius. Is $C = 2\pi r$ here? If not, which is true: $C > 2\pi r$ or $C < 2\pi r$?

c. Consider the Saddleland surface of Example 1.3.7. Construct a circle centered at the tip of the saddle and derive a relationship between its circumference and its radius. Is $C = 2 r$ here? If not, which is true: $C > 2 r$ or $C < 2 r$?

Source: http://mphitchman.com/geometry/section1-3.html

Index

History of Geometry, 53
Hyperbolische Geometrie, 58

Projective Geometry, 1

Radiana Measure, 19
Rotations with Unit Complex Numbers, 47

Testing the Limit of Einstein's Theories, 83

Bibliography

entries

Mathematics - Annotations 005

Prof. Stenio Musich

August 12, 2020

Chapter 1

Mathematics

1.1 Projetive Geometry - Divided Spheres

Date: 20200730

1) C:/ARQ-IMPORTANT-08/Important-02/Divided Spheres Geodesics and the Orderly Subdivision of the Sphere by Edward S. Popko.pdf

Pag. 202(223) - 8.5.2 Edge Reference Points Class I subdivision begins at the PPT's edges. A series of reference points is located along the edges, dividing the edges into segments equal in number to the subdivision frequency. Reference points have a regular pattern. Their spacing may be equal or unequal, but all three PPT edges will have the same pattern regardless of the spacing used. There are two approaches to defining reference points: Equal-chords and Equal-arcs. Each approach can produce several different grids. For Equal-chords, the PPT's apex-to-apex chords (all three are the same length) are subdivided with equally spaced reference points. These reference points are on the PPT's edge chord; thus, they are inside the sphere. They must be projected to the surface of the sphere before they can be used as subdivision

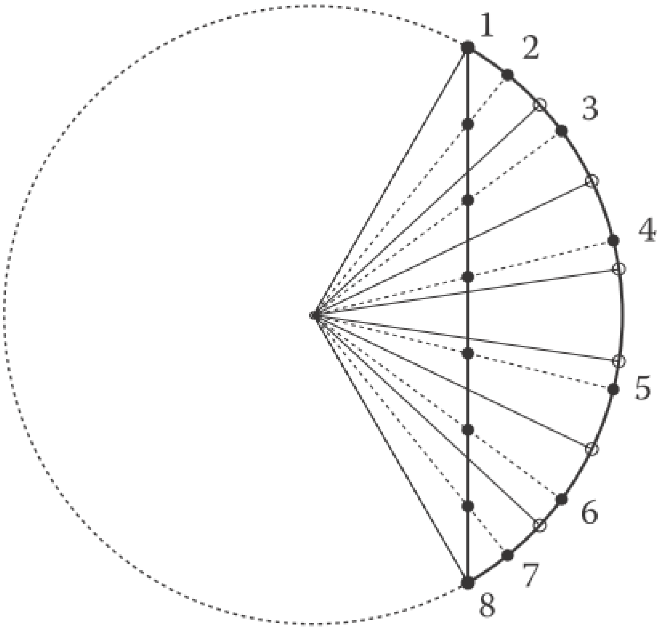

Figure 1.1: ds05-01.png - 7v Equal-chords, Equal-arcs subdivision.

references. For Equal-arcs, the PPT's arc edge (not its chords) is evenly subdivided with points, which are used as references, as is. No projection is needed because these points are already on the surface of the sphere. Figure 8.8 illustrates the difference between these two approaches with a 7v subdivision of a tetrahedron's chord. We use an odd frequency so that reference points from the two approaches will not be superimposed on one another. This makes the figure easier to read. We also use the chord of a tetrahedron, the longest of any Platonic with a unit circumsphere, to make comparisons between Equal-chords and Equal-arcs more vivid. Referring to the figure, points 1 and 8 are the chord's end points on the surface of the circumscribing sphere represented in the cross section by the circumscribing circle. For grids based on equal chord reference points, the chord between points 1 and 8 is divided into seven equal segments (7v) with points located between them. The dashed lines show chord points projected from the center of the sphere to its surface. They are numbered 2 through 7 (1 and 8 are vertices of the tetrahedron's PPT edge and are already on the sphere's surface). By inspecting the figure, you can easily see the arc spacing between projected chord points is not equal. The difference between the largest and smallest arc length is more than 2:1 in this example. To put some numbers to this, the tetrahedron's overall edge arc is approximately 109.5°, and the seven arcs resulting from Equal-chords subdivision are 9.4°, 14.1°, 19.8°, 22.8°, 19.8°, 14.1°, and 9.4°, in that order. It is also evident that the arc lengths are symmetrical; that is, arc 1-2 equals 7-8, arc 2-3 equals 6-7, and so on. The equal-arc points in Figure 8.8 are marked with small hollow circles on the surface of the sphere. They do not have to be projected. Each is spaced at 15.6o.Note that lines between them and the center of the sphere divide the chord into unequal segments. Appendix C lists simple formulas for dividing chords and arcs between any two points on the sphere into equal segments and projecting points to the surface of the sphere.

Radian Measure Another commonly used angle measurement is the radian. To define a radian, first consider a circle of radius r and two radii OA and OB. The angle formed by the two radii is a central angle. The portion of the circle between and is an arc of the circle and is written $\overline{AB}$. We say that $\overline{AB}$ subtends the angle . The length of $\overline{AB}$ is s (see Figure 5.19).

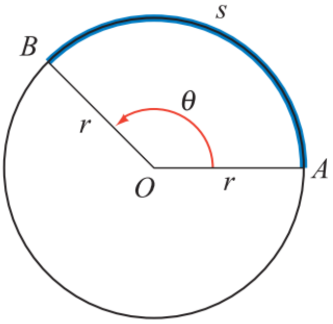

Figure 1.2: ds05-02a.png

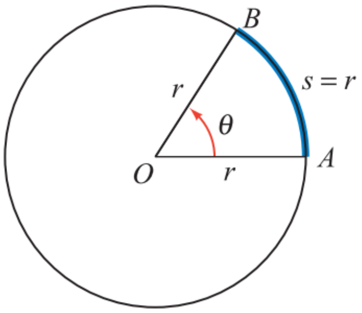

Figure 1.3: ds05-02b.png

One radian is the measure of the central angle subtended by an arc of length r on a circle of radius r.

Definition of Radian Measure

Given an arc of length s on a circle of radius r, the measure of the central angle subtended by the arc is = $\frac{s}{r}$ radians.

Source: C:/ARQ-IMPORTANT-08/Trigonometry-1439048606.pdf

Geogebra - C:/ARQ-IMPORTANT-03/Radian-20200812/radian-20200812-001.ggb

IMPORTANT

Construcão Geométrica: ==============

Geogebra:/ 1) GG01 - Circle (A,B) -$x^2 + y^2 = 1$

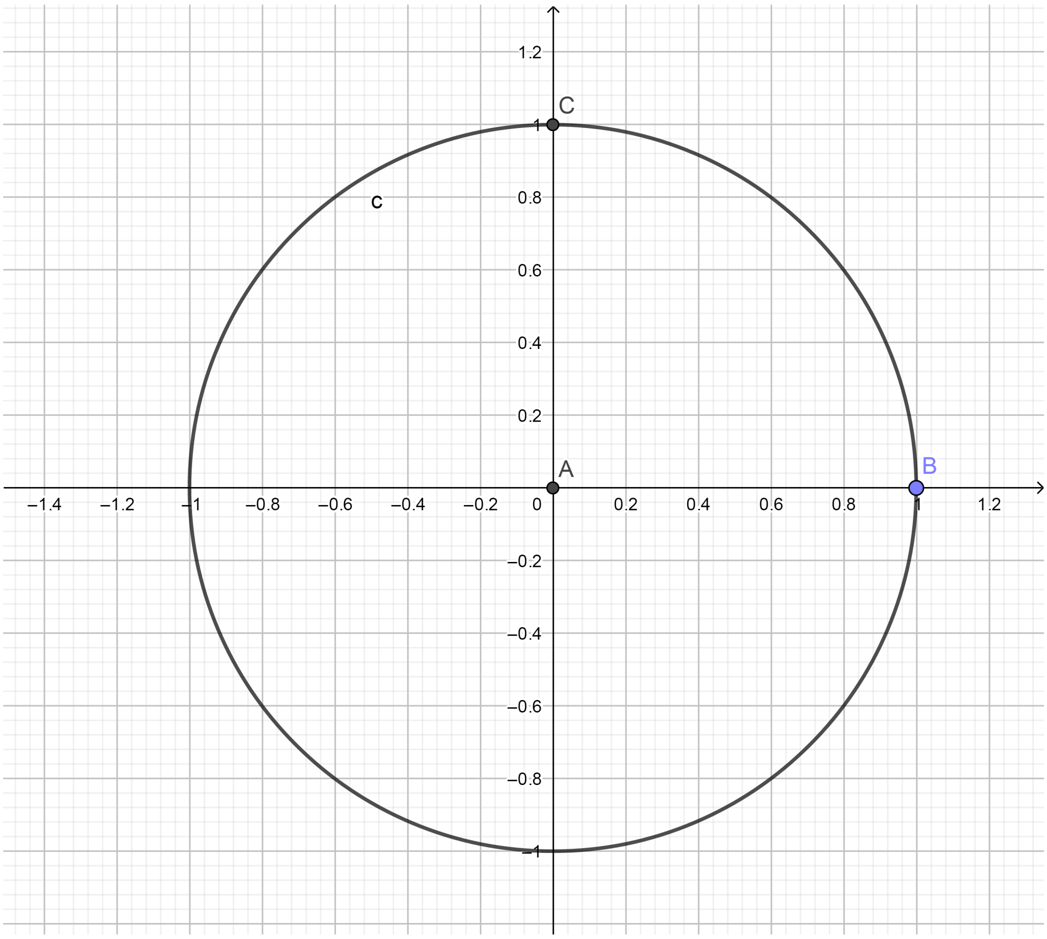

Figure 1.4: gg01.png

2) GG02- Semicircle(B,C) - Select two end points.

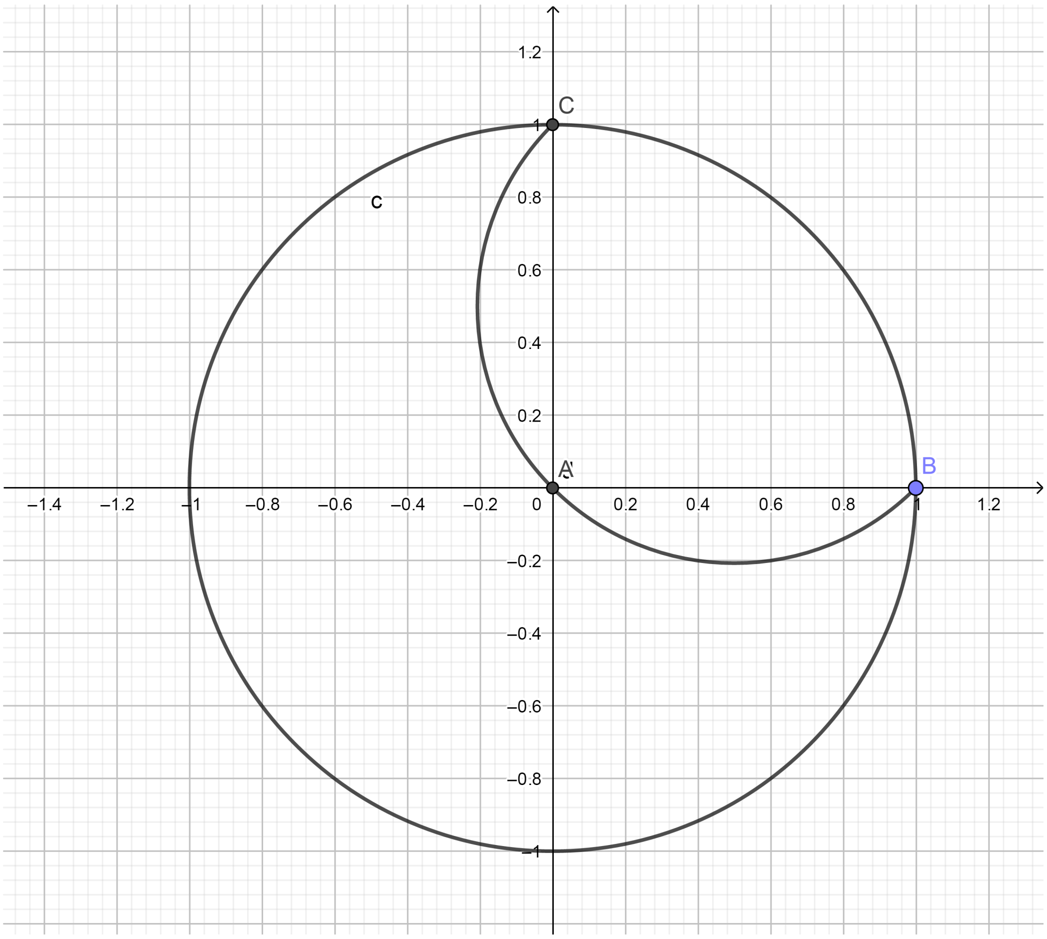
C
c
A'
B
−1.4
−1.2
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
1.2
−0.2
−0.4
−0.6
−0.8
−1
0.2
0.4
0.6
0.8
1.2

Figure 1.5: gg02.png

3) GG03 - Perpendicular (D, xAxis)

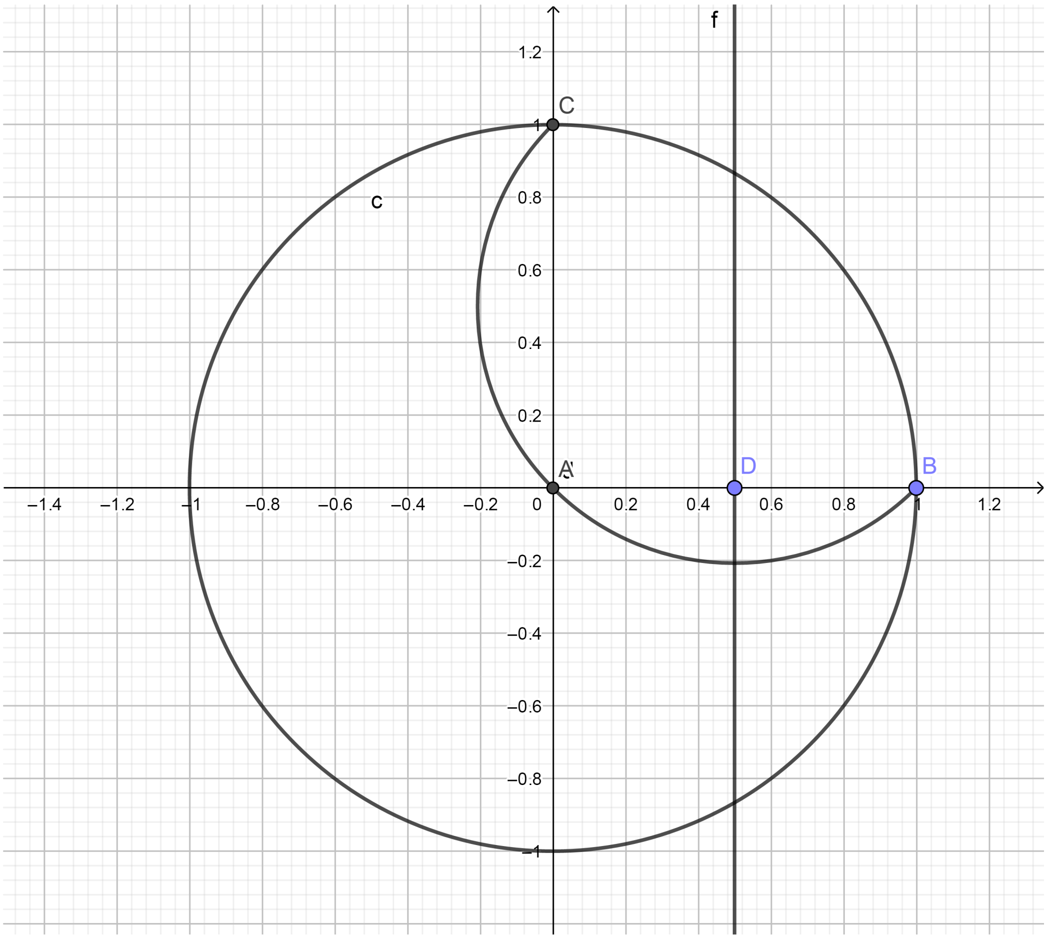
f
1.2
C
1
c
0.8
0.6
0.4
0.2
A'
D
B
−1.4
−1.2
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
1.2
−0.2
−0.4
−0.6
−0.8
−1

Figure 1.6: gg03.png

4) GG04 - Semicircle (B,F)

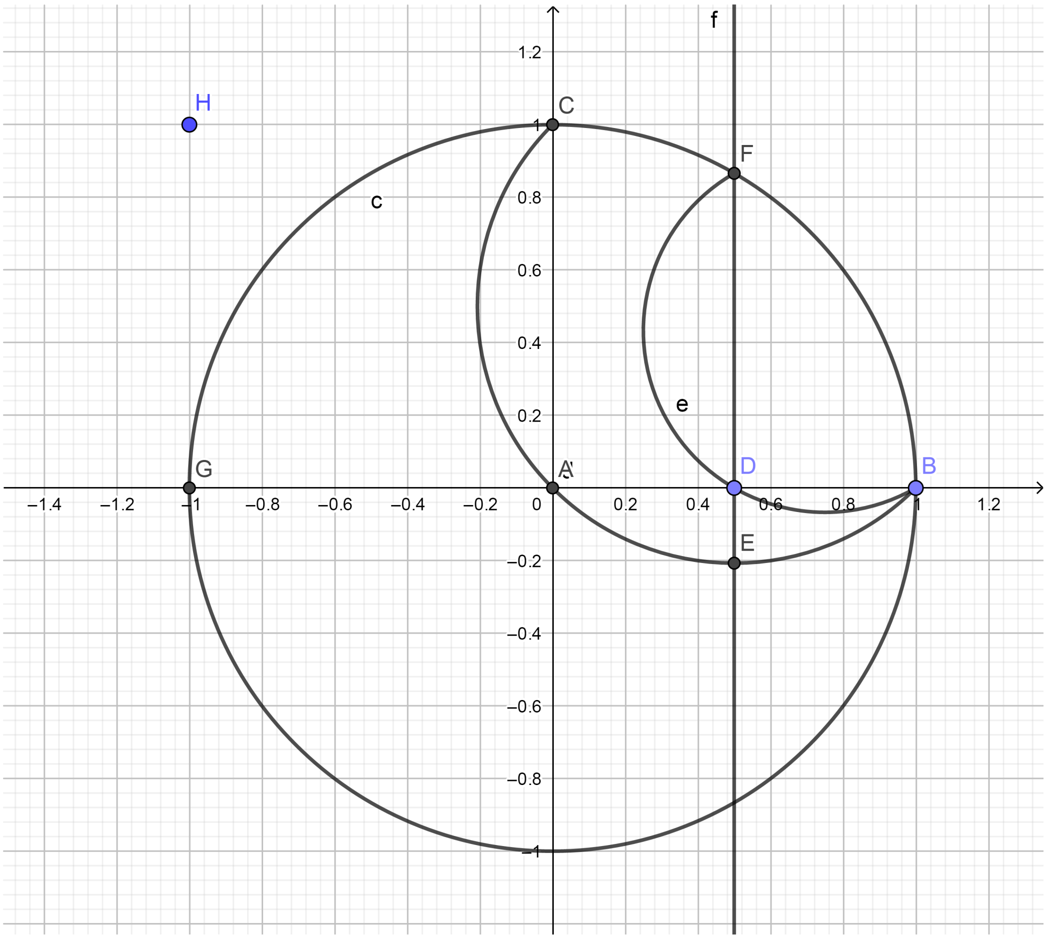
f
H
C
F
c
e
G
A
D
B
E
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1.4
−1.2
−1
−0.8
−0.6
−0.4
−0.2
0.2
0.4
0.6
0.8
1
1.2

Figure 1.7: gg04.png

5) GG05 - Circular Arc (H, G, C) - Select center and two points.

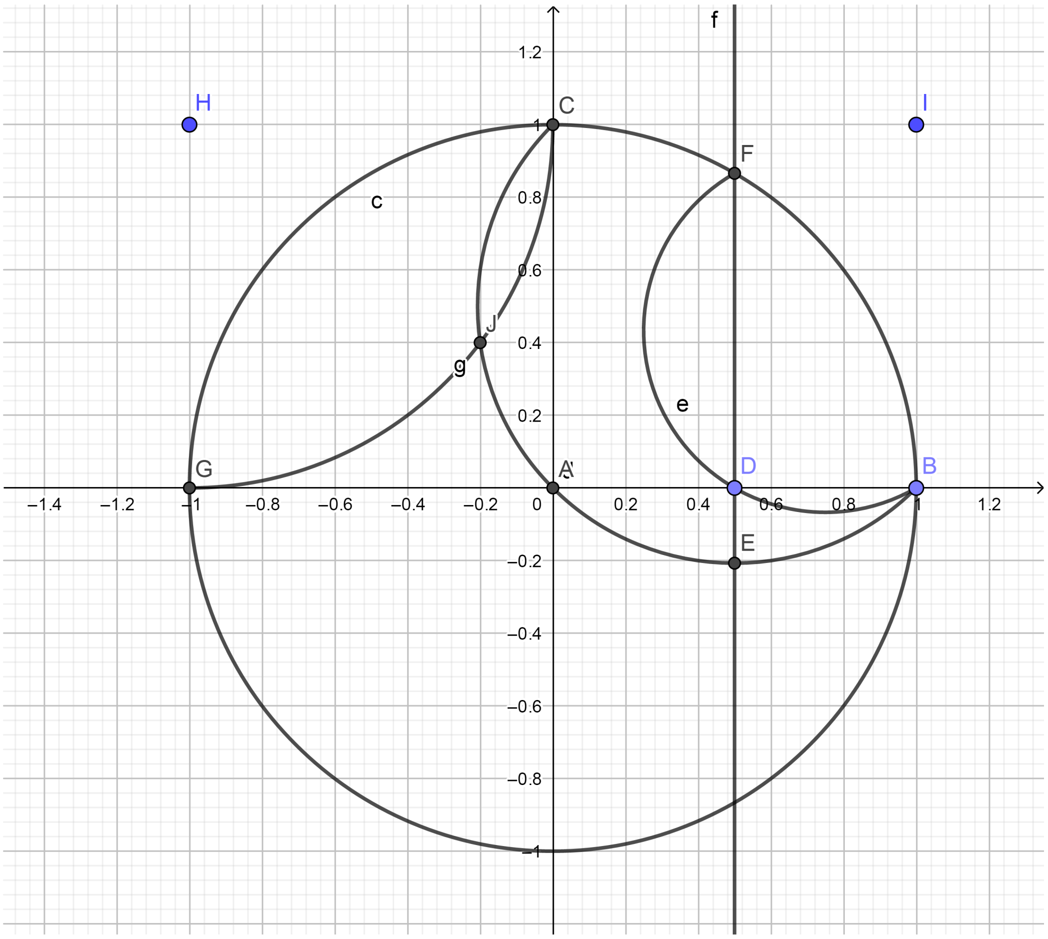
f
1.2
H
C
I
1
F
c
0.8
0.6
J
0.4
g
0.2
e
G
A
D
B
−1.4
−1.2
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
1.2
E
−0.2
−0.4
−0.6
−0.8
−1

Figure 1.8: gg05.png

6) GG06 - Semicircle (B, I)

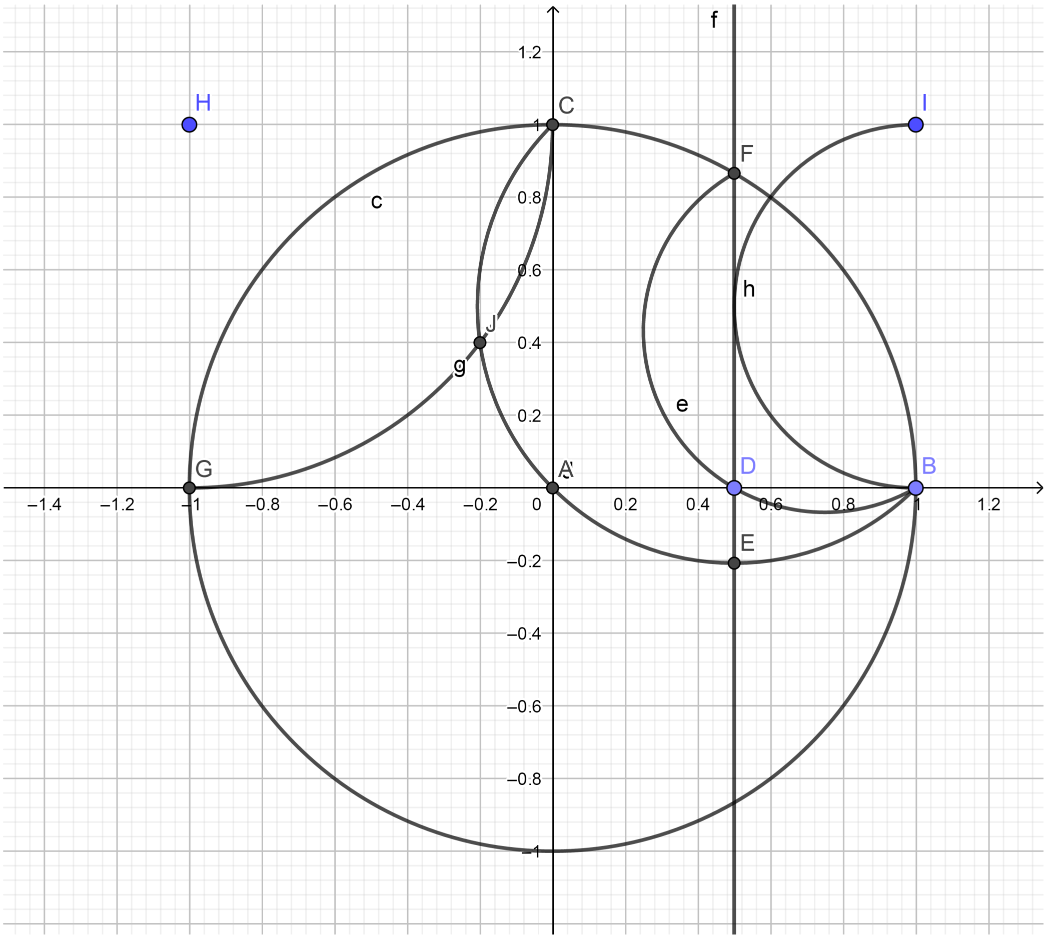
f
H
C
I
F
c
h
J
g
e
G
A
D
B
E
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1.4
−1.2
−0.8
−0.6
−0.4
0.2
0.4
0.6
0.8
1.2

Figure 1.9: gg06.png

7) GG07 -Semicircle (I, O)

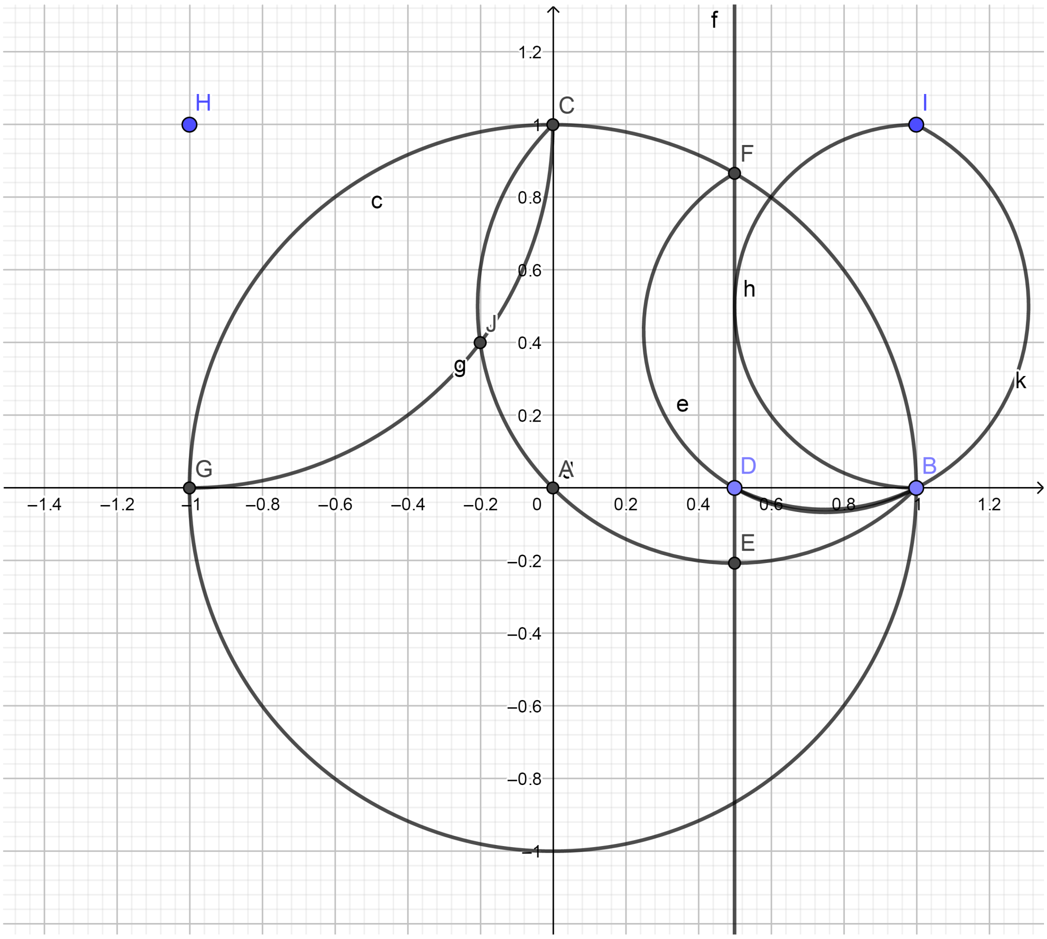
f
1.2
H
C
I
1
F
0.8
c
0.6
h
J
0.4
g
k
0.2
e
G
A'
D
B
−1.4
−1.2
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
1.2
E
−0.2
−0.4
−0.6
−0.8
−1

Figure 1.10: gg07.png

8) GG08 - Semicircle (B, A)

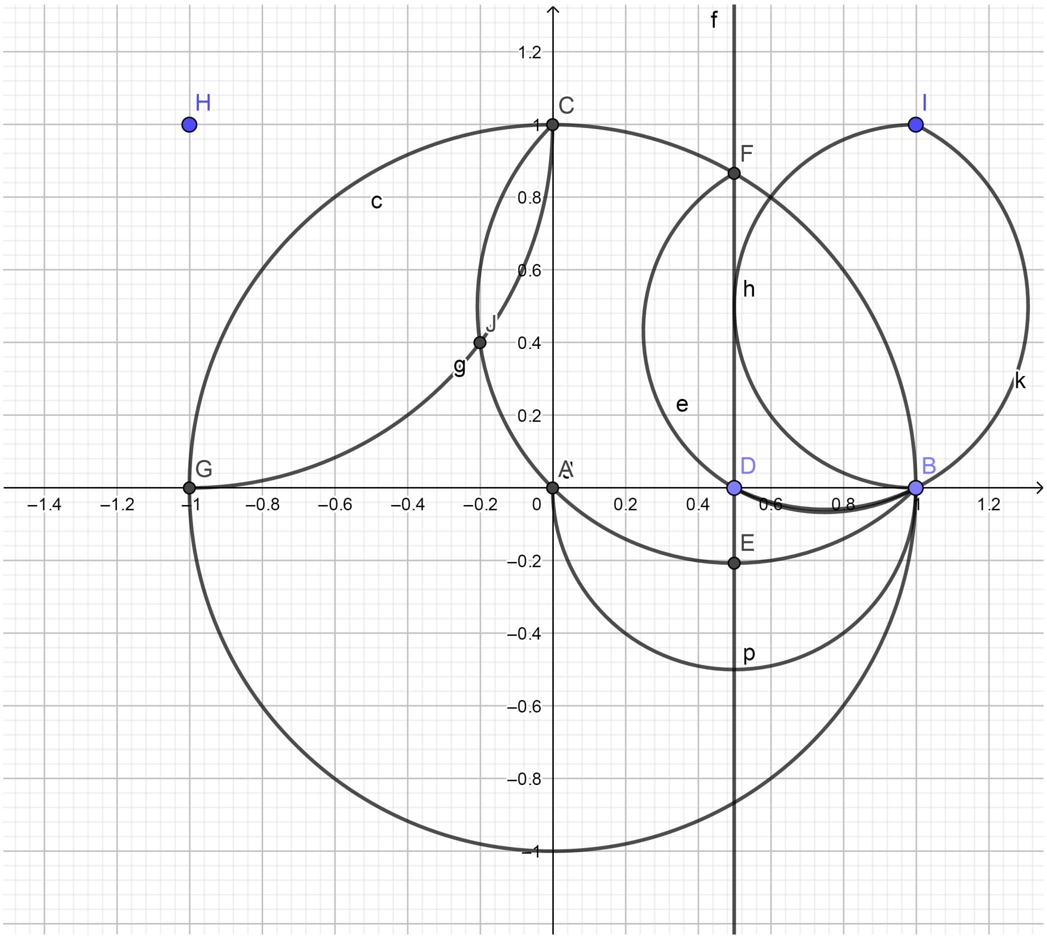
f
H
C
I
F
c
h
J
g
k
e
G
A
D
B
E
p
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1.4
−1.2
−1
−0.8
−0.6
−0.4
−0.2
0.2
0.4
0.6
0.8
1
1.2

Figure 1.11: gg08.png

9) GG09 - Circular Arc (I, C, B)

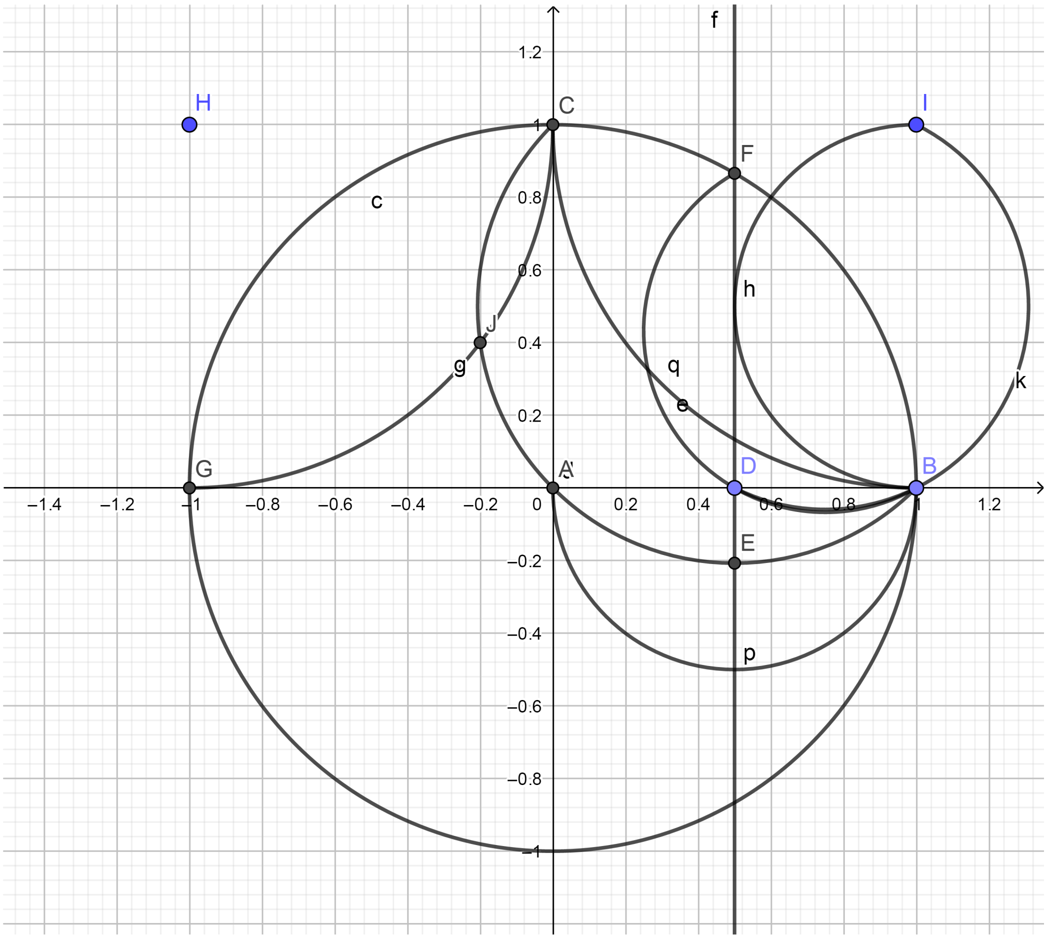
f
H
C
I
F
c
h
J
g
q
e
k
G
A'
D
B
E
p
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1.4
−1.2
0.2
0.4
0.6
0.8
1.2

Figure 1.12: gg09.png

10) GG10 - Circle (I, D, J) - Select through 3 points.

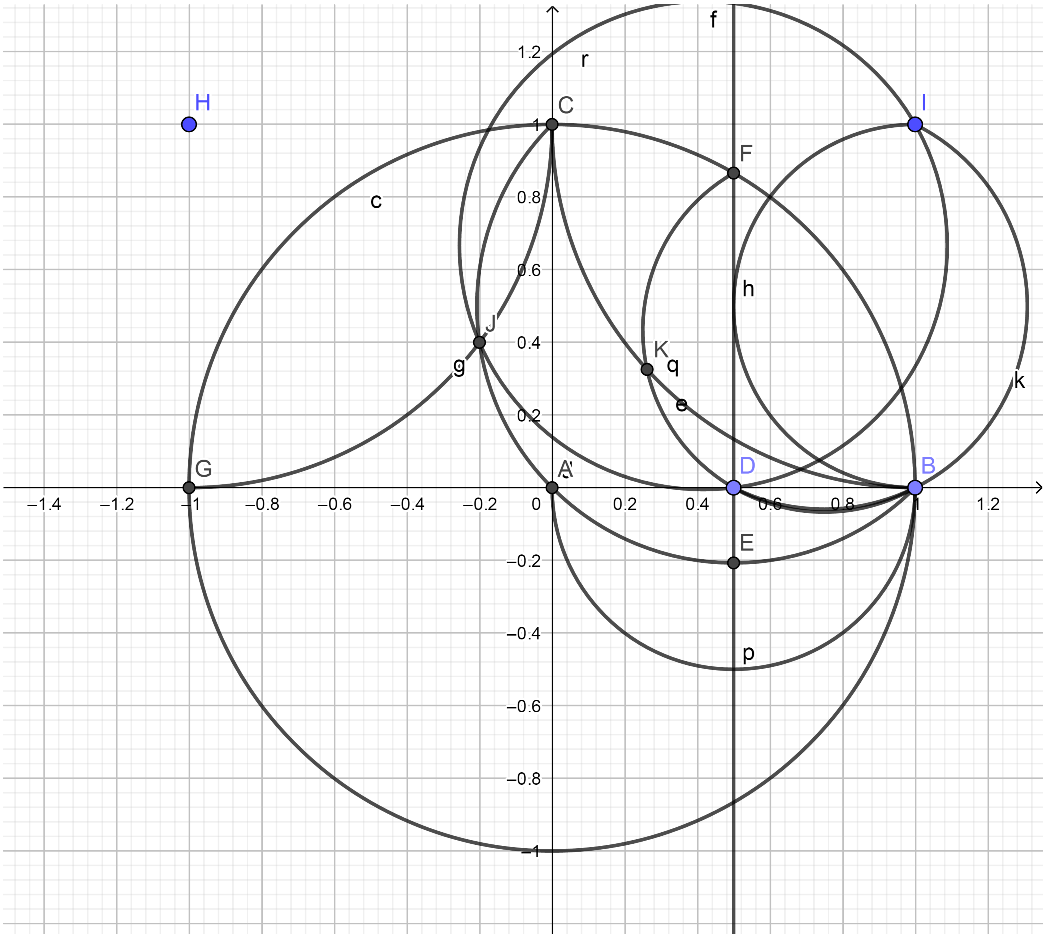
H
I
C
F
c
h
J
g
K
q
e
k
G
A
D
B
E
p
f
r
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1.4
−1.2
−0.8
−0.6
−0.4
−0.2
0.2
0.4
0.6
0.8
1.2

Figure 1.13: gg10.png

Vamos comparar os semircicle (B, F) e o semicircle (I, D), observaremos uma não coincidência das linhas na parte inferior dos referidos círculos.

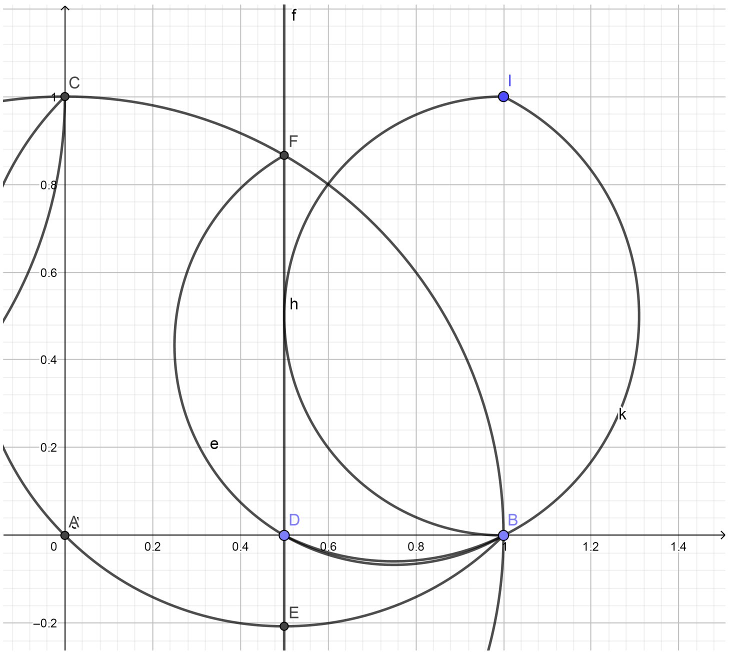
f
C
I
1
F
0.8
0.6
h
0.4
k
0.2
e
A'
D
B
0
0.2
0.4
0.6
0.8
1
1.2
1.4
E
−0.2

Figure 1.14: gg07a.png

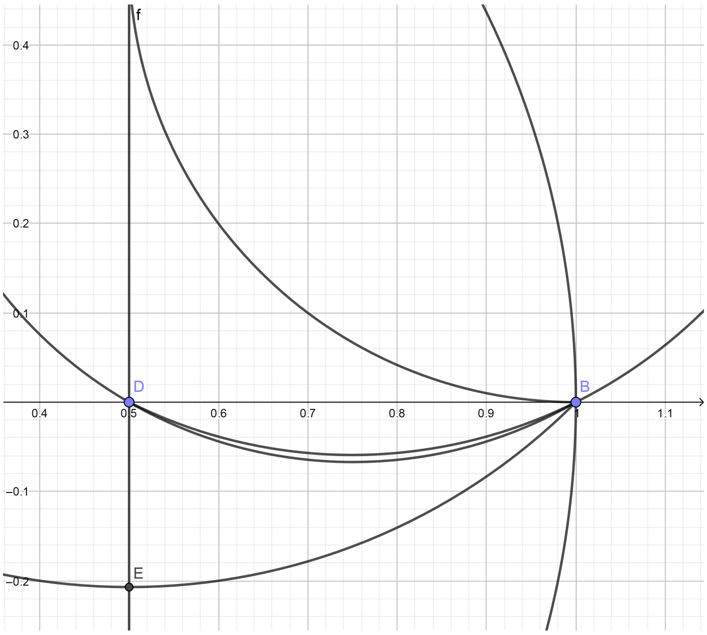
f
0.4
0.3
0.2
0.1
D
B
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
−0.1
E
−0.2

Figure 1.15: gg07b.png

2) Theorem of Pythagoras - calculus

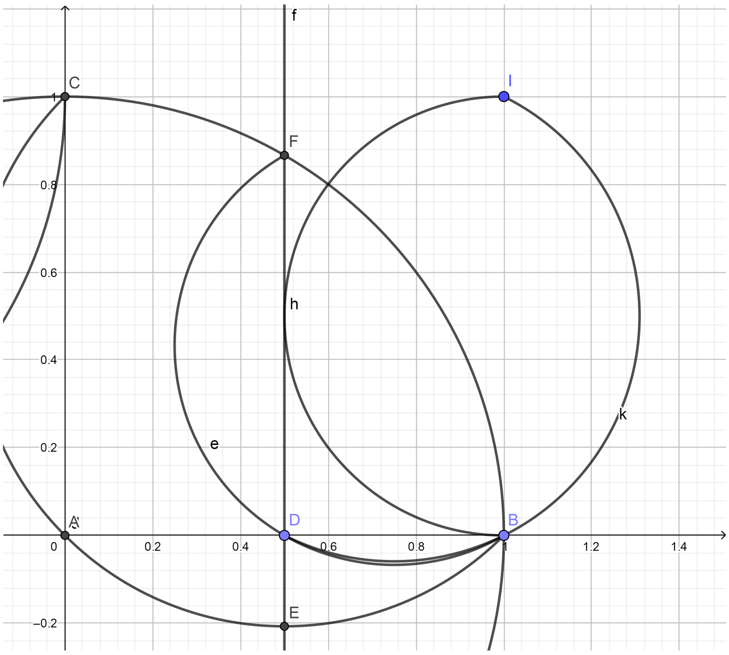
f
C
1
I
F
0.8
0.6
h
0.4
0.2
e
k
A
D
B
0
0.2
0.4
0.6
0.8
1
1.2
1.4
E
−0.2

Figure 1.16: gg07a.png

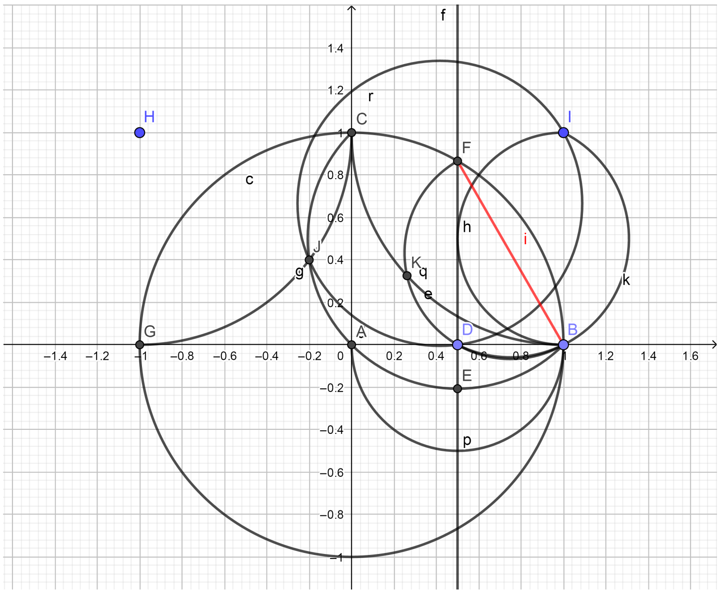
f
1.4
1.2
r
H
C
I
1
F
c
0.8
0.6
h
i
J
0.4
K
q
g
e
k
0.2
G
A
D
B
−1.4
−1.2
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
E
−0.2
−0.4
p
−0.6
−0.8
−1

Figure 1.17: gg011.png

Source: https://en.wikipedia.org/wiki/Pythagorean-theorem

Differential geometry On an infinitesimal level, in three dimensional space, Pythagoras's theorem describes the distance between two infinitesimally separated points as:

$$ds^2 = dx^2 + dy^2 + dz^2 \tag{1.1}$$

with ds the element of distance and (dx, dy, dz) the components of the vector separating the two points. Such a space is called a Euclidean space. However, in Riemannian geometry, a generalization of this expression useful for general coordinates (not just Cartesian) and general spaces (not just Euclidean) takes the form[1]. Source:

https://physics.bgu.ac.il/ gedalin/Teaching/Mater/mmp.pdf

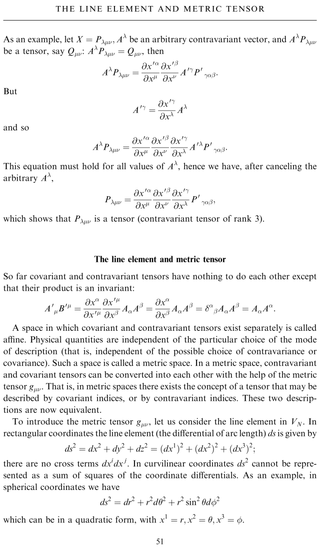

THE LINE ELEMENT AND METRIC TENSOR

As an example, let $X = P_{\lambda\mu\nu}$, A^λ be an arbitrary contravariant vector, and $A^\lambda P_{\lambda\mu\nu}$ be a tensor, say $Q_{\mu\nu}$: $A^\lambda P_{\lambda\mu\nu} = Q_{\mu\nu}$, then

$$A^\lambda P_{\lambda\mu\nu} = \frac{\partial x'^\alpha}{\partial x^\mu}\frac{\partial x'^\beta}{\partial x^\nu} A'^\gamma P'_{\ \gamma\alpha\beta}.$$

But

$$A'^\gamma = \frac{\partial x'^\gamma}{\partial x^\lambda} A^\lambda$$

and so

$$A^\lambda P_{\lambda\mu\nu} = \frac{\partial x'^\alpha}{\partial x^\mu}\frac{\partial x'^\beta}{\partial x^\nu}\frac{\partial x'^\gamma}{\partial x^\lambda} A'^\lambda P'_{\ \gamma\alpha\beta}.$$

This equation must hold for all values of A^λ, hence we have, after canceling the arbitrary A^λ,

$$P_{\lambda\mu\nu} = \frac{\partial x'^\alpha}{\partial x^\mu}\frac{\partial x'^\beta}{\partial x^\nu}\frac{\partial x'^\gamma}{\partial x^\lambda} P'_{\ \gamma\alpha\beta},$$

which shows that $P_{\lambda\mu\nu}$ is a tensor (contravariant tensor of rank 3).

The line element and metric tensor

So far covariant and contravariant tensors have nothing to do each other except that their product is an invariant:

$$A'_{\ \mu}B'^\mu = \frac{\partial x^\alpha}{\partial x'^\mu}\frac{\partial x'^\mu}{\partial x^\beta} A_\alpha A^\beta = \frac{\partial x^\alpha}{\partial x^\beta} A_\alpha A^\beta = \delta^\alpha_{\ \beta} A_\alpha A^\beta = A_\alpha A^\alpha.$$

A space in which covariant and contravariant tensors exist separately is called affine. Physical quantities are independent of the particular choice of the mode of description (that is, independent of the possible choice of contravariance or covariance). Such a space is called a metric space. In a metric space, contravariant and covariant tensors can be converted into each other with the help of the metric tensor $g_{\mu\nu}$. That is, in metric spaces there exists the concept of a tensor that may be described by covariant indices, or by contravariant indices. These two descriptions are now equivalent.

To introduce the metric tensor $g_{\mu\nu}$, let us consider the line element in V_N. In rectangular coordinates the line element (the differential of arc length) ds is given by

$$ds^2 = dx^2 + dy^2 + dz^2 = (dx^1)^2 + (dx^2)^2 + (dx^3)^2;$$

there are no cross terms $dx^i dx^j$. In curvilinear coordinates ds^2 cannot be represented as a sum of squares of the coordinate differentials. As an example, in spherical coordinates we have

$$ds^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2$$

which can be in a quadratic form, with $x^1 = r, x^2 = \theta, x^3 = \phi$.

51

Figure 1.18: mmp01.png

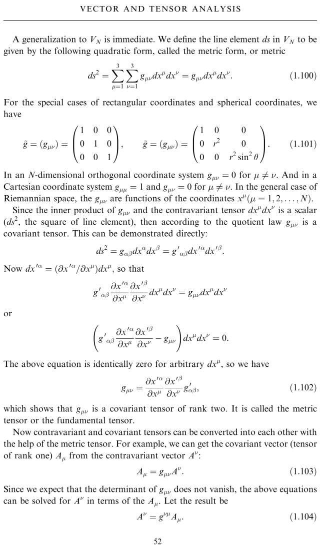

VECTOR AND TENSOR ANALYSIS

A generalization to V_N is immediate. We define the line element ds in V_N to be given by the following quadratic form, called the metric form, or metric

$$ds^2 = \sum_{\mu=1}^{3}\sum_{\nu=1}^{3} g_{\mu\nu} dx^\mu dx^\nu = g_{\mu\nu} dx^\mu dx^\nu. \tag{1.100}$$

For the special cases of rectangular coordinates and spherical coordinates, we have

$$\tilde{g} = (g_{\mu\nu}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \tilde{g} = (g_{\mu\nu}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2\sin^2\theta \end{pmatrix}. \tag{1.101}$$

In an N-dimensional orthogonal coordinate system $g_{\mu\nu} = 0$ for $\mu \neq \nu$. And in a Cartesian coordinate system $g_{\mu\mu} = 1$ and $g_{\mu\nu} = 0$ for $\mu \neq \nu$. In the general case of Riemannian space, the $g_{\mu\nu}$ are functions of the coordinates $x^\mu (\mu = 1, 2, \ldots, N)$.

Since the inner product of $g_{\mu\nu}$ and the contravariant tensor $dx^\mu dx^\nu$ is a scalar (ds^2, the square of line element), then according to the quotient law $g_{\mu\nu}$ is a covariant tensor. This can be demonstrated directly:

$$ds^2 = g_{\alpha\beta} dx^\alpha dx^\beta = g'_{\ \alpha\beta} dx'^\alpha dx'^\beta.$$

Now $dx'^\alpha = (\partial x'^\alpha / \partial x^\mu) dx^\mu$, so that

$$g'_{\ \alpha\beta}\frac{\partial x'^\alpha}{\partial x^\mu}\frac{\partial x'^\beta}{\partial x^\nu} dx^\mu dx^\nu = g_{\mu\nu} dx^\mu dx^\nu$$

or

$$\left(g'_{\ \alpha\beta}\frac{\partial x'^\alpha}{\partial x^\mu}\frac{\partial x'^\beta}{\partial x^\nu} - g_{\mu\nu}\right) dx^\mu dx^\nu = 0.$$

The above equation is identically zero for arbitrary dx^μ, so we have

$$g_{\mu\nu} = \frac{\partial x'^\alpha}{\partial x^\mu}\frac{\partial x'^\beta}{\partial x^\nu} g'_{\alpha\beta}, \tag{1.102}$$

which shows that $g_{\mu\nu}$ is a covariant tensor of rank two. It is called the metric tensor or the fundamental tensor.

Now contravariant and covariant tensors can be converted into each other with the help of the metric tensor. For example, we can get the covariant vector (tensor of rank one) A_μ from the contravariant vector A^ν:

$$A_\mu = g_{\mu\nu} A^\nu. \tag{1.103}$$

Since we expect that the determinant of $g_{\mu\nu}$ does not vanish, the above equations can be solved for A^ν in terms of the A_μ. Let the result be

$$A^\nu = g^{\nu\mu} A_\mu. \tag{1.104}$$

52

Figure 1.19: mmp02.png

---

[1] Mathematical methods for physicists - Chow; Source: C:/DATA-NovoVivo-129-2020/Books-129-2020/The Pythagorean Proposition/mmp.pdf - Pag. 51 (66) - The line element and metric tensor

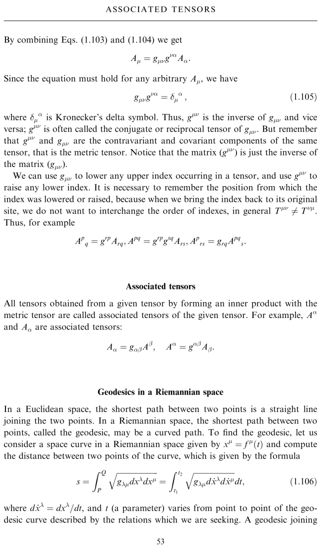

By combining Eqs. (1.103) and (1.104) we get

$$A_\mu = g_{\mu\nu} g^{\nu\alpha} A_\alpha.$$

Since the equation must hold for any arbitrary A_μ, we have

$$g_{\mu\nu} g^{\nu\alpha} = \delta_\mu{}^\alpha, \tag{1.105}$$

where $\delta_\mu{}^\alpha$ is Kronecker's delta symbol. Thus, $g^{\mu\nu}$ is the inverse of $g_{\mu\nu}$ and vice versa; $g^{\mu\nu}$ is often called the conjugate or reciprocal tensor of $g_{\mu\nu}$. But remember that $g^{\mu\nu}$ and $g_{\mu\nu}$ are the contravariant and covariant components of the same tensor, that is the metric tensor. Notice that the matrix $(g^{\mu\nu})$ is just the inverse of the matrix $(g_{\mu\nu})$.

We can use $g_{\mu\nu}$ to lower any upper index occurring in a tensor, and use $g^{\mu\nu}$ to raise any lower index. It is necessary to remember the position from which the index was lowered or raised, because when we bring the index back to its original site, we do not want to interchange the order of indexes, in general $T^{\mu\nu} \neq T^{\nu\mu}$. Thus, for example

$$A^p{}_q = g^{rp} A_{rq}, A^{pq} = g^{rp} g^{sq} A_{rs}, A^p{}_{rs} = g_{rq} A^{pq}{}_s.$$

Associated tensors

All tensors obtained from a given tensor by forming an inner product with the metric tensor are called associated tensors of the given tensor. For example, A^α and A_α are associated tensors:

$$A_\alpha = g_{\alpha\beta} A^\beta, \quad A^\alpha = g^{\alpha\beta} A_\beta.$$

Geodesics in a Riemannian space

In a Euclidean space, the shortest path between two points is a straight line joining the two points. In a Riemannian space, the shortest path between two points, called the geodesic, may be a curved path. To find the geodesic, let us consider a space curve in a Riemannian space given by $x^\mu = f^\mu(t)$ and compute the distance between two points of the curve, which is given by the formula

$$s = \int_P^Q \sqrt{g_{\lambda\mu} dx^\lambda dx^\mu} = \int_{t_1}^{t_2} \sqrt{g_{\lambda\mu} d\dot{x}^\lambda d\dot{x}^\mu} dt, \tag{1.106}$$

where $d\dot{x}^\lambda = dx^\lambda / dt$, and t (a parameter) varies from point to point of the geodesic curve described by the relations which we are seeking. A geodesic joining

Figure 1.20: mmp03.png

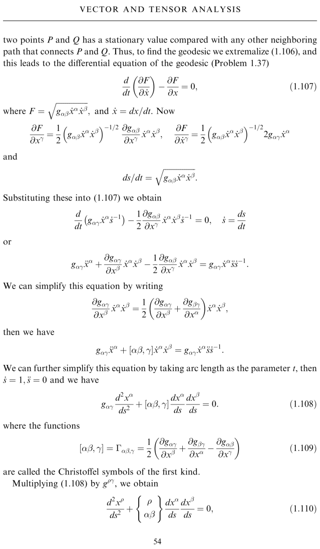

two points P and Q has a stationary value compared with any other neighboring path that connects P and Q. Thus, to find the geodesic we extremalize (1.106), and this leads to the differential equation of the geodesic (Problem 1.37)

$$\frac{d}{dt}\left(\frac{\partial F}{\partial \dot{x}}\right) - \frac{\partial F}{\partial x} = 0, \tag{1.107}$$

where $F = \sqrt{g_{\alpha\beta} \dot{x}^\alpha \dot{x}^\beta}$, and $\dot{x} = dx/dt$. Now

$$\frac{\partial F}{\partial x^\gamma} = \frac{1}{2}\left(g_{\alpha\beta} \dot{x}^\alpha \dot{x}^\beta\right)^{-1/2} \frac{\partial g_{\alpha\beta}}{\partial x^\gamma} \dot{x}^\alpha \dot{x}^\beta, \quad \frac{\partial F}{\partial \dot{x}^\gamma} = \frac{1}{2}\left(g_{\alpha\beta} \dot{x}^\alpha \dot{x}^\beta\right)^{-1/2} 2 g_{\alpha\gamma} \dot{x}^\alpha$$

and

$$ds/dt = \sqrt{g_{\alpha\beta} \dot{x}^\alpha \dot{x}^\beta}.$$

Substituting these into (1.107) we obtain

$$\frac{d}{dt}\left(g_{\alpha\gamma} \dot{x}^\alpha \dot{s}^{-1}\right) - \frac{1}{2} \frac{\partial g_{\alpha\beta}}{\partial x^\gamma} \dot{x}^\alpha \dot{x}^\beta \dot{s}^{-1} = 0, \quad \dot{s} = \frac{ds}{dt}$$

or

$$g_{\alpha\gamma} \ddot{x}^\alpha + \frac{\partial g_{\alpha\gamma}}{\partial x^\beta} \dot{x}^\alpha \dot{x}^\beta - \frac{1}{2} \frac{\partial g_{\alpha\beta}}{\partial x^\gamma} \dot{x}^\alpha \dot{x}^\beta = g_{\alpha\gamma} \dot{x}^\alpha \ddot{s} \dot{s}^{-1}.$$

We can simplify this equation by writing

$$\frac{\partial g_{\alpha\gamma}}{\partial x^\beta} \dot{x}^\alpha \dot{x}^\beta = \frac{1}{2}\left(\frac{\partial g_{\alpha\gamma}}{\partial x^\beta} + \frac{\partial g_{\beta\gamma}}{\partial x^\alpha}\right) \dot{x}^\alpha \dot{x}^\beta,$$

then we have

$$g_{\alpha\gamma} \ddot{x}^\alpha + [\alpha\beta, \gamma] \dot{x}^\alpha \dot{x}^\beta = g_{\alpha\gamma} \dot{x}^\alpha \ddot{s} \dot{s}^{-1}.$$

We can further simplify this equation by taking arc length as the parameter t, then $\dot{s} = 1, \ddot{s} = 0$ and we have

$$g_{\alpha\gamma} \frac{d^2 x^\alpha}{ds^2} + [\alpha\beta, \gamma] \frac{dx^\alpha}{ds} \frac{dx^\beta}{ds} = 0. \tag{1.108}$$

where the functions

$$[\alpha\beta, \gamma] = \Gamma_{\alpha\beta,\gamma} = \frac{1}{2}\left(\frac{\partial g_{\alpha\gamma}}{\partial x^\beta} + \frac{\partial g_{\beta\gamma}}{\partial x^\alpha} - \frac{\partial g_{\alpha\beta}}{\partial x^\gamma}\right) \tag{1.109}$$

are called the Christoffel symbols of the first kind.

Multiplying (1.108) by $g^{\rho\gamma}$, we obtain

$$\frac{d^2 x^\rho}{ds^2} + \left\{ \begin{matrix} \rho \\ \alpha\beta \end{matrix} \right\} \frac{dx^\alpha}{ds} \frac{dx^\beta}{ds} = 0, \tag{1.110}$$

Figure 1.21: mmp04.png

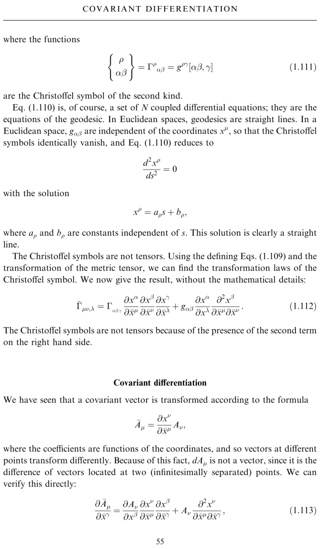

where the functions

$$\left\{ \begin{matrix} \rho \\ \alpha\beta \end{matrix} \right\} = \Gamma^\rho{}_{\alpha\beta} = g^{\rho\gamma} [\alpha\beta, \gamma] \tag{1.111}$$

are the Christoffel symbol of the second kind.

Eq. (1.110) is, of course, a set of N coupled differential equations; they are the equations of the geodesic. In Euclidean spaces, geodesics are straight lines. In a Euclidean space, $g_{\alpha\beta}$ are independent of the coordinates x^μ, so that the Christoffel symbols identically vanish, and Eq. (1.110) reduces to

$$\frac{d^2 x^\rho}{ds^2} = 0$$

with the solution

$$x^\rho = a_\rho s + b_\rho,$$

where a_ρ and b_ρ are constants independent of s. This solution is clearly a straight line.

The Christoffel symbols are not tensors. Using the defining Eqs. (1.109) and the transformation of the metric tensor, we can find the transformation laws of the Christoffel symbol. We now give the result, without the mathematical details:

$$\bar{\Gamma}_{\mu\nu,\lambda} = \Gamma_{\alpha\beta,\gamma} \frac{\partial x^\alpha}{\partial \bar{x}^\mu} \frac{\partial x^\beta}{\partial \bar{x}^\nu} \frac{\partial x^\gamma}{\partial \bar{x}^\lambda} + g_{\alpha\beta} \frac{\partial x^\alpha}{\partial x^\lambda} \frac{\partial^2 x^\beta}{\partial \bar{x}^\mu \partial \bar{x}^\nu}. \tag{1.112}$$

The Christoffel symbols are not tensors because of the presence of the second term on the right hand side.

Covariant differentiation

We have seen that a covariant vector is transformed according to the formula

$$\bar{A}_\mu = \frac{\partial x^\nu}{\partial \bar{x}^\mu} A_\nu,$$

where the coefficients are functions of the coordinates, and so vectors at different points transform differently. Because of this fact, dA_μ is not a vector, since it is the difference of vectors located at two (infinitesimally separated) points. We can verify this directly:

$$\frac{\partial \bar{A}_\mu}{\partial \bar{x}^\gamma} = \frac{\partial A_\nu}{\partial x^\beta} \frac{\partial x^\nu}{\partial \bar{x}^\mu} \frac{\partial x^\beta}{\partial \bar{x}^\gamma} + A_\nu \frac{\partial^2 x^\nu}{\partial \bar{x}^\mu \partial \bar{x}^\gamma}, \tag{1.113}$$

Figure 1.22: mmp05.png

$$ds^2 = \sum_{i;j}^{n} g_{ij}\, dx_i\, dx_j$$

which is called the metric tensor. (Sometimes, by abuse of language, the same term is applied to the set of coefficients gij.) It may be a function of position, and often describes curved space. A simple example is Euclidean (flat) space expressed in curvilinear coordinates. For example, in polar coordinates:

$$ds^2 = dr^2 + r^2 d\ ^2:$$

O .:

Very small triangles

For any uniform curvature K (positive, zero, or negative), in very small right triangles ($j\ K\ j\ a^2; j\ K\ j\ b^2 << 1$) with hypotenuse c, it can be shown that

$$c^2 = a^2 + b^2 \quad \frac{K}{3}a^2b^2 \quad \frac{K^2}{45}a^2b^2\ \ a^2 + b^2 \quad \frac{2K^3}{945}a^2b^2\ \ a^2\ \ \ b^2\ ^2 + O\ \ K^4c^{10}\ \ : \tag{1.2}$$

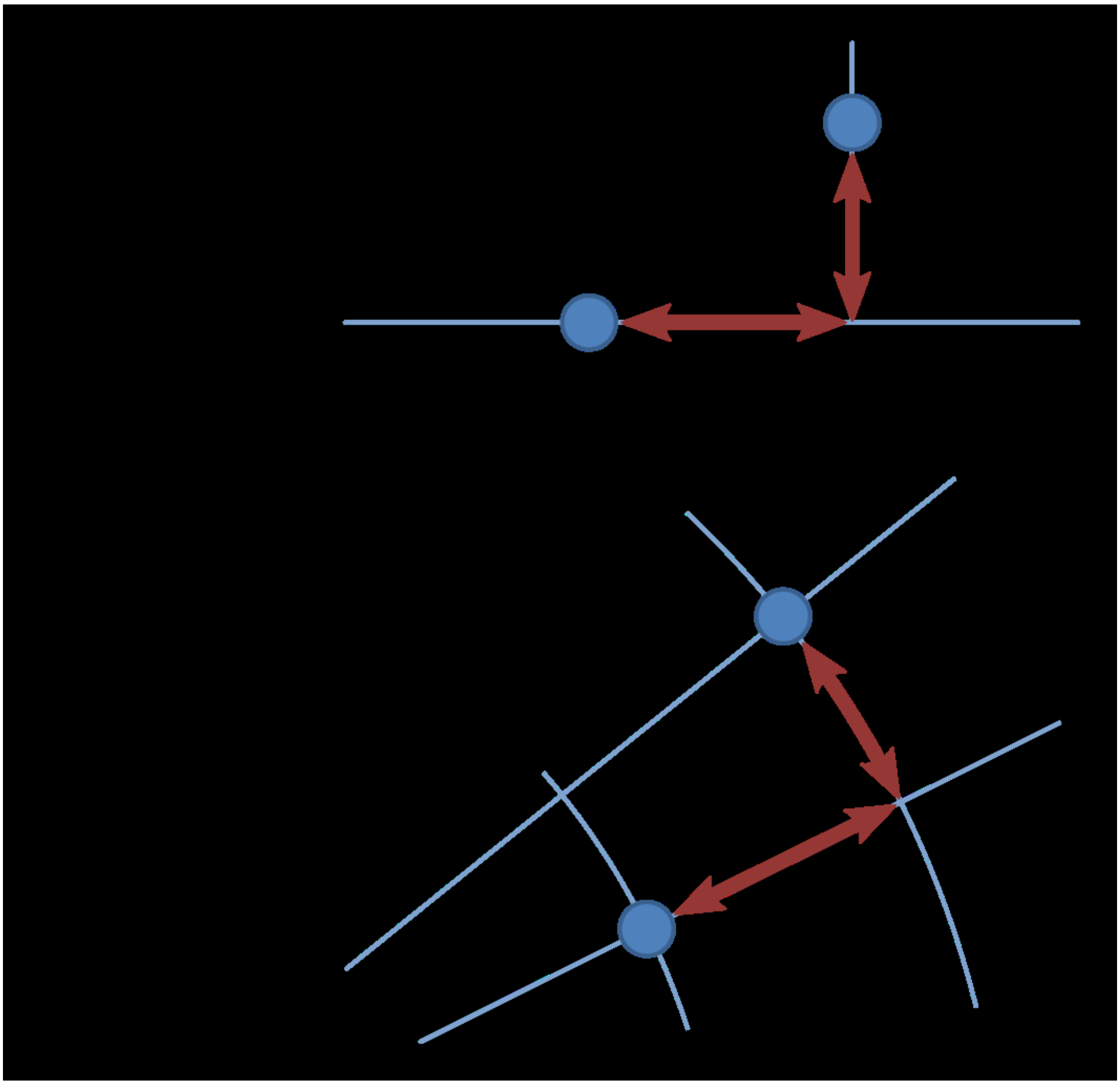

Figure 1.23: Line-elements.png

2a) Source: C:/DATA-NovoVivo-129-2020/Books-129-2020/Notes on Geometry and Arithmetic/NotesOnGeometryAndArithmetic.pdf Pag. 01(12) - Chapter 1 Diophantus of Alexandria

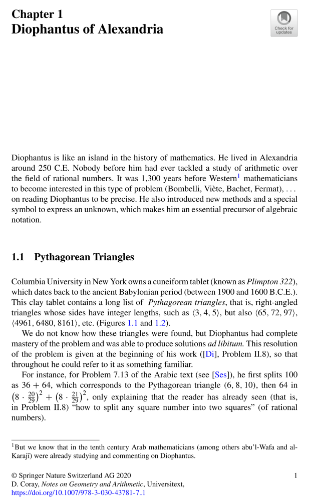

Chapter 1
Diophantus of Alexandria

Diophantus is like an island in the history of mathematics. He lived in Alexandria around 250 C.E. Nobody before him had ever tackled a study of arithmetic over the field of rational numbers. It was 1,300 years before Western[1] mathematicians to become interested in this type of problem (Bombelli, Viète, Bachet, Fermat), . . . on reading Diophantus to be precise. He also introduced new methods and a special symbol to express an unknown, which makes him an essential precursor of algebraic notation.

1.1 Pythagorean Triangles

Columbia University in New York owns a cuneiform tablet (known as *Plimpton 322*), which dates back to the ancient Babylonian period (between 1900 and 1600 B.C.E.). This clay tablet contains a long list of *Pythagorean triangles*, that is, right-angled triangles whose sides have integer lengths, such as $\langle 3, 4, 5\rangle$, but also $\langle 65, 72, 97\rangle$, $\langle 4961, 6480, 8161\rangle$, etc. (Figures 1.1 and 1.2).

We do not know how these triangles were found, but Diophantus had complete mastery of the problem and was able to produce solutions *ad libitum.* This resolution of the problem is given at the beginning of his work ([Di], Problem II.8), so that throughout he could refer to it as something familiar.

For instance, for Problem 7.13 of the Arabic text (see [Ses]), he first splits 100 as $36 + 64$, which corresponds to the Pythagorean triangle $(6, 8, 10)$, then 64 in $\left(8 \cdot \frac{20}{29}\right)^2 + \left(8 \cdot \frac{21}{29}\right)^2$, only explaining that the reader has already seen (that is, in Problem II.8) "how to split any square number into two squares" (of rational numbers).

---

[1] But we know that in the tenth century Arab mathematicians (among others abu'l-Wafa and al-Karajī) were already studying and commenting on Diophantus.

© Springer Nature Switzerland AG 2020
D. Coray, *Notes on Geometry and Arithmetic*, Universitext,
https://doi.org/10.1007/978-3-030-43781-7_1

Figure 1.24: noga01.png

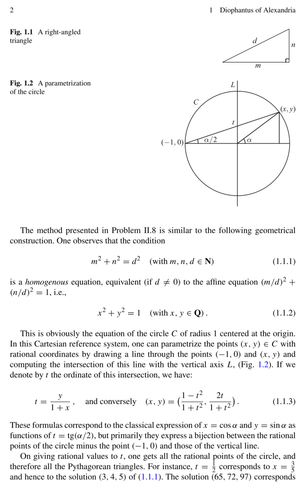

Fig. 1.1 A right-angled triangle

Fig. 1.2 A parametrization of the circle

The method presented in Problem II.8 is similar to the following geometrical construction. One observes that the condition

$$m^2 + n^2 = d^2 \quad (\text{with } m, n, d \in \mathbf{N}) \tag{1.1.1}$$

is a *homogenous* equation, equivalent (if $d \neq 0$) to the affine equation $(m/d)^2 + (n/d)^2 = 1$, i.e.,

$$x^2 + y^2 = 1 \quad (\text{with } x, y \in \mathbf{Q})\,. \tag{1.1.2}$$

This is obviously the equation of the circle C of radius 1 centered at the origin. In this Cartesian reference system, one can parametrize the points $(x, y) \in C$ with rational coordinates by drawing a line through the points $(-1, 0)$ and (x, y) and computing the intersection of this line with the vertical axis L, (Fig. 1.2). If we denote by t the ordinate of this intersection, we have:

$$t = \frac{y}{1+x}\,, \quad \text{and conversely} \quad (x, y) = \Big(\frac{1-t^2}{1+t^2}, \frac{2t}{1+t^2}\Big)\,. \tag{1.1.3}$$

These formulas correspond to the classical expression of $x = \cos\alpha$ and $y = \sin\alpha$ as functions of $t = \mathrm{tg}(\alpha/2)$, but primarily they express a bijection between the rational points of the circle minus the point $(-1, 0)$ and those of the vertical line.

On giving rational values to t, one gets all the rational points of the circle, and therefore all the Pythagorean triangles. For instance, $t = \frac{1}{2}$ corresponds to $x = \frac{3}{5}$ and hence to the solution $(3, 4, 5)$ of (1.1.1). The solution $(65, 72, 97)$ corresponds

Figure 1.25: noga02.png

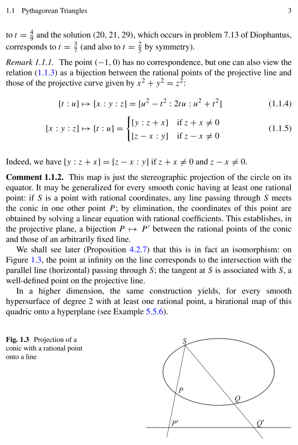

to $t = \frac{4}{9}$ and the solution $(20, 21, 29)$, which occurs in problem 7.13 of Diophantus, corresponds to $t = \frac{3}{7}$ (and also to $t = \frac{2}{5}$ by symmetry).

Remark 1.1.1. The point $(-1, 0)$ has no correspondence, but one can also view the relation (1.1.3) as a bijection between the rational points of the projective line and those of the projective curve given by $x^2 + y^2 = z^2$:

$$[t : u] \mapsto [x : y : z] = [u^2 - t^2 : 2tu : u^2 + t^2] \tag{1.1.4}$$

$$[x : y : z] \mapsto [t : u] = \begin{cases} [y : z + x] & \text{if } z + x \neq 0 \\ [z - x : y] & \text{if } z - x \neq 0 \end{cases} \tag{1.1.5}$$

Indeed, we have $[y : z + x] = [z - x : y]$ if $z + x \neq 0$ and $z - x \neq 0$.

Comment 1.1.2. This map is just the stereographic projection of the circle on its equator. It may be generalized for every smooth conic having at least one rational point: if S is a point with rational coordinates, any line passing through S meets the conic in one other point P; by elimination, the coordinates of this point are obtained by solving a linear equation with rational coefficients. This establishes, in the projective plane, a bijection $P \mapsto P'$ between the rational points of the conic and those of an arbitrarily fixed line.

We shall see later (Proposition 4.2.7) that this is in fact an isomorphism: on Figure 1.3, the point at infinity on the line corresponds to the intersection with the parallel line (horizontal) passing through S; the tangent at S is associated with S, a well-defined point on the projective line.

In a higher dimension, the same construction yields, for every smooth hypersurface of degree 2 with at least one rational point, a birational map of this quadric onto a hyperplane (see Example 5.5.6).

Fig. 1.3 Projection of a conic with a rational point onto a line

Figure 1.26: noga03.png

Pag. 41 (50) - 4.1 Projective Space

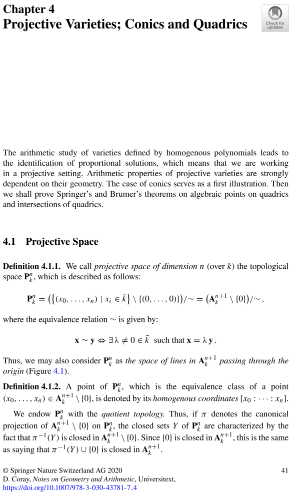

Chapter 4
Projective Varieties; Conics and Quadrics

The arithmetic study of varieties defined by homogenous polynomials leads to the identification of proportional solutions, which means that we are working in a projective setting. Arithmetic properties of projective varieties are strongly dependent on their geometry. The case of conics serves as a first illustration. Then we shall prove Springer's and Brumer's theorems on algebraic points on quadrics and intersections of quadrics.

4.1 Projective Space

Definition 4.1.1. We call *projective space of dimension n* (over k) the topological space $\mathbf{P}_k^n$, which is described as follows:

$$\mathbf{P}_k^n = \big(\big\{(x_0, \ldots, x_n) \mid x_i \in \bar{k}\big\} \setminus \{(0, \ldots, 0)\}\big)/\!\sim\; = \big(\mathbf{A}_k^{n+1} \setminus \{0\}\big)/\!\sim ,$$

where the equivalence relation $\sim$ is given by:

$$\mathbf{x} \sim \mathbf{y} \Leftrightarrow \exists\, \lambda \neq 0 \in \bar{k} \text{ such that } \mathbf{x} = \lambda\, \mathbf{y} .$$

Thus, we may also consider $\mathbf{P}_k^n$ as *the space of lines in* $\mathbf{A}_k^{n+1}$ *passing through the origin* (Figure 4.1).

Definition 4.1.2. A point of $\mathbf{P}_k^n$, which is the equivalence class of a point $(x_0, \ldots, x_n) \in \mathbf{A}_k^{n+1} \setminus \{0\}$, is denoted by its *homogenous coordinates* $[x_0 : \cdots : x_n]$.

We endow $\mathbf{P}_k^n$ with the *quotient topology*. Thus, if π denotes the canonical projection of $\mathbf{A}_k^{n+1} \setminus \{0\}$ on $\mathbf{P}_k^n$, the closed sets Y of $\mathbf{P}_k^n$ are characterized by the fact that $\pi^{-1}(Y)$ is closed in $\mathbf{A}_k^{n+1} \setminus \{0\}$. Since $\{0\}$ is closed in $\mathbf{A}_k^{n+1}$, this is the same as saying that $\pi^{-1}(Y) \cup \{0\}$ is closed in $\mathbf{A}_k^{n+1}$.

© Springer Nature Switzerland AG 2020

D. Coray, *Notes on Geometry and Arithmetic*, Universitext,
https://doi.org/10.1007/978-3-030-43781-7_4

Figure 1.27: noga04.png

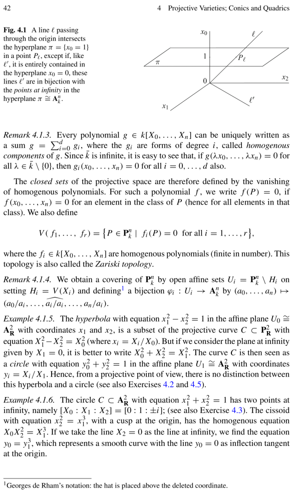

42 4 Projective Varieties; Conics and Quadrics

Fig. 4.1 A line ℓ passing through the origin intersects the hyperplane $\pi = \{x_0 = 1\}$ in a point P_ℓ, except if, like ℓ', it is entirely contained in the hyperplane $x_0 = 0$, these lines ℓ' are in bijection with the *points at infinity* in the hyperplane $\pi \cong \mathbf{A}_k^n$.

Remark 4.1.3. Every polynomial $g \in k[X_0, \ldots, X_n]$ can be uniquely written as a sum $g = \sum_{i=0}^{d} g_i$, where the g_i are forms of degree i, called *homogenous components* of g. Since $\bar{k}$ is infinite, it is easy to see that, if $g(\lambda x_0, \ldots, \lambda x_n) = 0$ for all $\lambda \in \bar{k} \setminus \{0\}$, then $g_i(x_0, \ldots, x_n) = 0$ for all $i = 0, \ldots, d$ also.

The *closed sets* of the projective space are therefore defined by the vanishing of homogenous polynomials. For such a polynomial f, we write $f(P) = 0$, if $f(x_0, \ldots, x_n) = 0$ for an element in the class of P (hence for all elements in that class). We also define

$$V(f_1, \ldots, f_r) = \big\{P \in \mathbf{P}_k^n \mid f_i(P) = 0 \text{ for all } i = 1, \ldots, r\big\},$$

where the $f_i \in k[X_0, \ldots, X_n]$ are homogenous polynomials (finite in number). This topology is also called the *Zariski topology*.

Remark 4.1.4. We obtain a covering of $\mathbf{P}_k^n$ by open affine sets $U_i = \mathbf{P}_k^n \setminus H_i$ on setting $H_i = V(X_i)$ and defining[1] a bijection $\varphi_i : U_i \to \mathbf{A}_k^n$ by $(a_0, \ldots, a_n) \mapsto (a_0/a_i, \ldots, \widehat{a_i/a_i}, \ldots, a_n/a_i)$.

Example 4.1.5. The *hyperbola* with equation $x_1^2 - x_2^2 = 1$ in the affine plane $U_0 \cong \mathbf{A}_{\mathbf{R}}^2$ with coordinates x_1 and x_2, is a subset of the projective curve $C \subset \mathbf{P}_{\mathbf{R}}^2$ with equation $X_1^2 - X_2^2 = X_0^2$ (where $x_i = X_i/X_0$). But if we consider the plane at infinity given by $X_1 = 0$, it is better to write $X_0^2 + X_2^2 = X_1^2$. The curve C is then seen as a *circle* with equation $y_0^2 + y_2^2 = 1$ in the affine plane $U_1 \cong \mathbf{A}_{\mathbf{R}}^2$ with coordinates $y_i = X_i/X_1$. Hence, from a projective point of view, there is no distinction between this hyperbola and a circle (see also Exercises 4.2 and 4.5).

Example 4.1.6. The circle $C \subset \mathbf{A}_{\mathbf{R}}^2$ with equation $x_1^2 + x_2^2 = 1$ has two points at infinity, namely $[X_0 : X_1 : X_2] = [0 : 1 : \pm i]$; (see also Exercise 4.3). The cissoid with equation $x_2^2 = x_1^3$, with a cusp at the origin, has the homogenous equation $X_0 X_2^2 = X_1^3$. If we take the line $X_2 = 0$ as the line at infinity, we find the equation $y_0 = y_1^3$, which represents a smooth curve with the line $y_0 = 0$ as inflection tangent at the origin.

---

[1] Georges de Rham's notation: the hat is placed above the deleted coordinate.

Figure 1.28: noga05.png

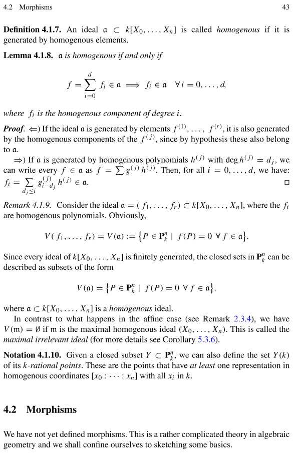

4.2 Morphisms 43

Definition 4.1.7. An ideal $\mathfrak{a} \subset k[X_0, \ldots, X_n]$ is called *homogenous* if it is generated by homogenous elements.

Lemma 4.1.8. $\mathfrak{a}$ *is homogenous if and only if*

$$f = \sum_{i=0}^{d} f_i \in \mathfrak{a} \implies f_i \in \mathfrak{a} \quad \forall\, i = 0, \ldots, d,$$

where f_i *is the homogenous component of degree* i.

Proof. $\Leftarrow$) If the ideal $\mathfrak{a}$ is generated by elements $f^{(1)}, \ldots, f^{(r)}$, it is also generated by the homogenous components of the $f^{(j)}$, since by hypothesis these also belong to $\mathfrak{a}$.

$\Rightarrow$) If $\mathfrak{a}$ is generated by homogenous polynomials $h^{(j)}$ with $\deg h^{(j)} = d_j$, we can write every $f \in \mathfrak{a}$ as $f = \sum g^{(j)} h^{(j)}$. Then, for all $i = 0, \ldots, d$, we have: $f_i = \sum_{d_j \leq i} g_{i-d_j}^{(j)} h^{(j)} \in \mathfrak{a}$. □

Remark 4.1.9. Consider the ideal $\mathfrak{a} = (f_1, \ldots, f_r) \subset k[X_0, \ldots, X_n]$, where the f_i are homogenous polynomials. Obviously,

$$V(f_1, \ldots, f_r) = V(\mathfrak{a}) := \big\{P \in \mathbf{P}_k^n \mid f(P) = 0\ \forall\, f \in \mathfrak{a}\big\}.$$

Since every ideal of $k[X_0, \ldots, X_n]$ is finitely generated, the closed sets in $\mathbf{P}_k^n$ can be described as subsets of the form

$$V(\mathfrak{a}) = \big\{P \in \mathbf{P}_k^n \mid f(P) = 0\ \forall\, f \in \mathfrak{a}\big\},$$

where $\mathfrak{a} \subset k[X_0, \ldots, X_n]$ is a *homogenous* ideal.

In contrast to what happens in the affine case (see Remark 2.3.4), we have $V(\mathfrak{m}) = \emptyset$ if $\mathfrak{m}$ is the maximal homogenous ideal $(X_0, \ldots, X_n)$. This is called the *maximal irrelevant ideal* (for more details see Corollary 5.3.6).

Notation 4.1.10. Given a closed subset $Y \subset \mathbf{P}_k^n$, we can also define the set $Y(k)$ of its *k-rational points*. These are the points that have *at least* one representation in homogenous coordinates $[x_0 : \cdots : x_n]$ with all x_i in k.

4.2 Morphisms

We have not yet defined morphisms. This is a rather complicated theory in algebraic geometry and we shall confine ourselves to sketching some basics.

Figure 1.29: noga06.png

3) C:/DATA-NovoVivo-129-2020/Books-129-2020/Topology/0821838849-A.pdf

Pag. 02 - As a separate branch of mathematics, topology is relatively young. It was isolated as a collection of methods and problems by Henri Poincare (1854{1912) in his pioneering paper Analysis situs of 1895. The subsequent development of the subject was dramatic and topology was deeply influential in shaping the mathematics of the twentieth century and today. So what is topology? In the popular understanding, objects like the Mobius band, the Klein bottle, and knots and links are the first to be mentioned (or maybe the second after the misunderstanding about topography is cleared up). Some folks can cite the joke that topologists are mathematicians who cannot tell their donut from their coffee cups. When I taught my first undergraduate courses in topology, I found I spent too much time developing a hierarchy of definitions and too little time on the objects, tools, and intuitions that are central to the subject. I wanted to teach a course that would follow a path more directly to the heart of topology. I wanted to tell a story that is coherent, motivating, and significant enough to form the basis for future study. To get an idea of what is studied by topology, let's examine its prehistory, that is, the vague notions that led Poincare to identify its foundations. Gottfried W. Leibniz (1646{1716), in a letter to Christiaan Huygens (1629{1695) in the 1670's, described a concept that has become a goal of the study of topology: I believe that we need another analysis properly geometric or linear, which treats PLACE directly the way that algebra treats MAGNITUDE. Leibniz envisioned a calculus of figures in which one might combine figures with the ease of numbers, operate on them as one might with polynomials, and produce new and rigorous geometric results. This science of PLACE was to be called Analysis situs ([Pont]). We don't know what Leibniz had in mind. It was Leonhard Euler (1701{1783) who made the first contributions to the infant subject, which he preferred to call geometria situs. His solution to the Bridges of Konigsberg problem and the celebrated Euler formula, V E + F = 2 (Chapter 11) were results that depended on the relative positions of geometric figures and not on their magnitudes ([Pont], [Lakatos]). In the nineteenth century, Carl-Friedrich Gauss (1777-1855) became interested in geometria situs when he studied knots and links as generalizations of the orbits of planets ([Epple]). By labeling figures of knots and links Gauss developed a rudimentary calculus that distinguished certain knots from each other by combinatorial means. Students who studied with Gauss and went on to develope some of the threads associated with geometria situs were Johann Listing (1808{1882), Augustus Mobius (1790{1868), and Bernhard Riemann (1826{1866). Listing extended Gauss's informal census of knots and links and he coined the term topology (from the Greek o o o o&, which in Latin is analysis situs). Mobius extended Euler's formula to surfaces and polyhedra in three-space. Riemann identified the methods of the infant analysis situs as fundamental in the study of complex functions. During the nineteenth century analysis was developed into a deep and subtle science. The notions of continuity of functions and the convergence of sequences were studied in increasingly general situations, beginning with the work of Georg Cantor (1845–1918) and finalized in the twentieth century by Felix Hausdorff (1869–1942) who proposed the general notion of a topological space in 1914 ([Hausdorff]). The central concept in topology is continuity, defined for functions between sets equipped with a notion of nearness (topological spaces) which is preserved by a continuous function. Topology is a kind of geometry in which the important properties of a figure are those that are preserved under continuous motions (homeomorphisms, Chapter 2). The popular image of topology as rubber sheet geometry is captured in this characterization. Topology provides a language of continuity that is general enough to include a vast array of phenomena while being precise enough to be developed in new ways. A motivating problem from the earliest struggles with the notion of continuity is the problem of dimension. In modern physics, higher dimensional manifolds play a fundamental role in describing theories with properties that combine the large and the small. Already in Poincaré's time the question of the physicality of dimension was on philosophers' minds, including Poincaré. Cantor had noticed in 1877 that as sets finite dimensional Euclidean spaces were indistinguishable (Chapter 1). If these identifications were possible in a continuous manner, a requirement of physical phenomena, then the role of dimension would need a critical reappraisal. The problem of dimension was important to the development of certain topological notions, including a strictly topological definition of dimension introduced by Henri Lebesgue (1875-1941) [Lebesgue]. The solution to the problem of dimension was found by L. E. J. Brouwer (1881–1966) and published in 1912 [Brouwer]. The methods introduced by Brouwer reshaped the subject. The story I want to tell in this book is based on the problem of dimension. This fundamental question from the early years of the subject organizes the exposition and provides the motivation for the choices of mathematical tools to develop. I have not chosen to follow the path of Lebesgue into dimension theory (see the classic text [Hurewicz-Wallman]) but the further ranging path of Poincaré and Brouwer. The fundamental group (Chapters 7 and 8) and simplicial methods (Chapters 10 and 11) provide tools that establish an approach to topological questions that has proven to be deep and is still developing. It is this approach that best fits Leibniz's wish. In what follows, we will cut a swath through the varied and beautiful landscape that is the field of topology with the goal of solving the problem of invariance of dimension. Along the way we will acquire the necessary vocabulary to make our way easily from one landmark to the next (without staying too long anywhere to pick up an accent). The first chapter reviews the set theory with which the problem of dimension can be posed. The next five chapters treat the basic point-set notions of topology; these ideas are closest to analysis, including connectedness and compactness. The next two chapters treat the fundamental group of a space, an idea introduced by Poincaré to associate a group to a space in such a way that equivalent spaces lead to isomorphic groups. The next chapter treats the Jordan Curve theorem, first stated by Jordan in 1882, and given a complete proof in 1905 by Oswald Veblen (1880{1960). The method of proof here mixes the point-set and the combinatorial to develop approximations and comparisons. The last two chapters take up the combinatorial theme and focus on simplicial complexes. To these conveniently constructed spaces we associate their homology, a sequence of vector spaces, which turn out to be isomorphic for equivalent complexes. We finish a proof of the topological invariance of dimension using homology. Though the motivation for this book is historical, I have not followed the history in the choice of methods or proofs. First proofs of significant results can be difficult. However, I have tried to imitate the mix of point-set and combinatorial ideas that was topology before 1935, what I call classical topology. Some beautiful results of this time are included, such as the Borsuk-Ulam theorem (see [Borsuk] and [Matousek]).

Pag. 06(11)- The Problem of Invariance of Dimension Theorem 1.10. There is a one-one correspondence R ! RXR .

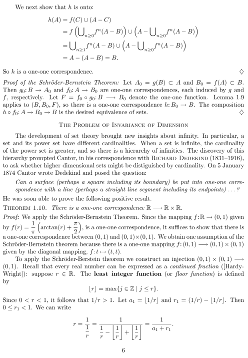

We next show that h is onto:

$$\begin{aligned} h(A) &= f(C) \cup (A - C) \\ &= f\left(\bigcup_{n\geq 0} f^n(A-B)\right) \cup \left(A - \bigcup_{n\geq 0} f^n(A-B)\right) \\ &= \bigcup_{n\geq 1} f^n(A-B) \cup \left(A - \bigcup_{n\geq 0} f^n(A-B)\right) \\ &= A - (A-B) = B. \end{aligned}$$

So h is a one-one correspondence. ◇

Proof of the Schröder-Bernstein Theorem: Let $A_0 = g(B) \subset A$ and $B_0 = f(A) \subset B$. Then $g_0: B \to A_0$ and $f_0: A \to B_0$ are one-one correspondences, each induced by g and f, respectively. Let $F = f_0 \circ g_0: B \longrightarrow B_0$ denote the one-one function. Lemma 1.9 applies to (B, B_0, F), so there is a one-one correspondence $h: B_0 \to B$. The composition $h \circ f_0: A \to B_0 \to B$ is the desired equivalence of sets. ◇

THE PROBLEM OF INVARIANCE OF DIMENSION

The development of set theory brought new insights about infinity. In particular, a set and its power set have different cardinalities. When a set is infinite, the cardinality of the power set is greater, and so there is a hierarchy of infinities. The discovery of this hierarchy prompted Cantor, in his correspondence with RICHARD DEDEKIND (1831–1916), to ask whether higher-dimensional sets might be distiguished by cardinality. On 5 January 1874 Cantor wrote Dedekind and posed the question:

> *Can a surface (perhaps a square including its boundary) be put into one-one correspondence with a line (perhaps a straight line segment including its endpoints) . . . ?*

He was soon able to prove the following positive result.

THEOREM 1.10. *There is a one-one correspondence* $\mathbb{R} \longrightarrow \mathbb{R} \times \mathbb{R}$.

Proof: We apply the Schröder-Bernstein Theorem. Since the mapping $f: \mathbb{R} \to (0,1)$ given by $f(r) = \frac{1}{\pi}\left(\arctan(r) + \frac{\pi}{2}\right)$, is a one-one correspondence, it suffices to show that there is a one-one correspondence between $(0,1)$ and $(0,1)\times(0,1)$. We obtain one assumption of the Schröder-Bernstein theorem because there is a one-one mapping $f: (0,1) \longrightarrow (0,1) \times (0,1)$ given by the diagonal mapping, $f: t \mapsto (t,t)$.

To apply the Schröder-Berstein theorem we construct an injection $(0,1) \times (0,1) \longrightarrow (0,1)$. Recall that every real number can be expressed as a *continued fraction* ([Hardy-Wright]): suppose $r \in \mathbb{R}$. The **least integer function** (or *floor function*) is defined by

$$\lfloor r \rfloor = \max\{j \in \mathbb{Z} \mid j \leq r\}.$$

Since $0 < r < 1$, it follows that $1/r > 1$. Let $a_1 = \lfloor 1/r \rfloor$ and $r_1 = (1/r) - \lfloor 1/r \rfloor$. Then $0 \leq r_1 < 1$. We can write

$$r = \frac{1}{\frac{1}{r}} = \frac{1}{\frac{1}{r} - \left\lfloor \frac{1}{r} \right\rfloor + \left\lfloor \frac{1}{r} \right\rfloor} = \frac{1}{a_1 + r_1}.$$

6

Figure 1.30: t01.png

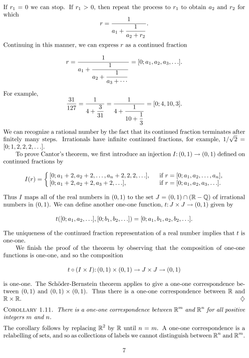

If $r_1 = 0$ we can stop. If $r_1 > 0$, then repeat the process to r_1 to obtain a_2 and r_2 for which

$$r = \cfrac{1}{a_1 + \cfrac{1}{a_2 + r_2}}.$$

Continuing in this manner, we can express r as a continued fraction

$$r = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}}} = [0; a_1, a_2, a_3, \ldots].$$

For example,

$$\frac{31}{127} = \cfrac{1}{4 + \cfrac{3}{31}} = \cfrac{1}{4 + \cfrac{1}{10 + \cfrac{1}{3}}} = [0; 4, 10, 3].$$

We can recognize a rational number by the fact that its continued fraction terminates after finitely many steps. Irrationals have infinite continued fractions, for example, $1/\sqrt{2} = [0; 1, 2, 2, 2, \ldots]$.

To prove Cantor's theorem, we first introduce an injection $I: (0,1) \to (0,1)$ defined on continued fractions by

$$I(r) = \begin{cases} [0; a_1+2, a_2+2, \ldots, a_n+2, 2, 2, \ldots], & \text{if } r = [0; a_1, a_2, \ldots, a_n], \\ [0; a_1+2, a_2+2, a_3+2, \ldots], & \text{if } r = [0; a_1, a_2, a_3, \ldots]. \end{cases}$$

Thus I maps all of the real numbers in $(0,1)$ to the set $J = (0,1) \cap (\mathbb{R} - \mathbb{Q})$ of irrational numbers in $(0,1)$. We can define another one-one function, $t: J \times J \to (0,1)$ given by

$$t([0; a_1, a_2, \ldots], [0; b_1, b_2, \ldots]) = [0; a_1, b_1, a_2, b_2, \ldots].$$

The uniqueness of the continued fraction representation of a real number implies that t is one-one.

We finish the proof of the theorem by observing that the composition of one-one functions is one-one, and so the composition

$$t \circ (I \times I): (0,1) \times (0,1) \to J \times J \to (0,1)$$

is one-one. The Schöder-Bernstein theorem applies to give a one-one correspondence between $(0,1)$ and $(0,1) \times (0,1)$. Thus there is a one-one correspondence between $\mathbb{R}$ and $\mathbb{R} \times \mathbb{R}$. ◇

COROLLARY 1.11. *There is a one-one correspondence between* $\mathbb{R}^m$ *and* $\mathbb{R}^n$ *for all positive integers* m *and* n.

The corollary follows by replacing $\mathbb{R}^2$ by $\mathbb{R}$ until $n = m$. A one-one correspondence is a relabelling of sets, and so as collections of labels we cannot distinguish between $\mathbb{R}^n$ and $\mathbb{R}^m$.

7

Figure 1.31: t02.png

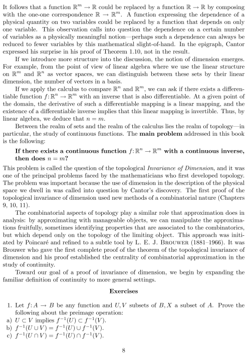

It follows that a function $\mathbb{R}^m \to \mathbb{R}$ could be replaced by a function $\mathbb{R} \to \mathbb{R}$ by composing with the one-one correspondence $\mathbb{R} \to \mathbb{R}^m$. A function expressing the dependence of a physical quantity on two variables could be replaced by a function that depends on only one variable. This observation calls into question the dependence on a certain number of variables as a physically meaningful notion—perhaps such a dependence can always be reduced to fewer variables by this mathematical slight-of-hand. In the epigraph, Cantor expressed his surprise in his proof of Theorem 1.10, not in the result.

If we introduce more structure into the discussion, the notion of dimension emerges. For example, from the point of view of linear algebra where we use the linear structure on $\mathbb{R}^m$ and $\mathbb{R}^n$ as vector spaces, we can distinguish between these sets by their linear dimension, the number of vectors in a basis.

If we apply the calculus to compare $\mathbb{R}^n$ and $\mathbb{R}^m$, we can ask if there exists a differentiable function $f: \mathbb{R}^n \to \mathbb{R}^m$ with an inverse that is also differentiable. At a given point of the domain, the derivative of such a differentiable mapping is a linear mapping, and the existence of a differentiable inverse implies that this linear mapping is invertible. Thus, by linear algebra, we deduce that $n = m$.

Between the realm of sets and the realm of the calculus lies the realm of topology—in particular, the study of continuous functions. The **main problem** addressed in this book is the following:

> **If there exists a continuous function $f: \mathbb{R}^n \to \mathbb{R}^m$ with a continuous inverse, then does $n = m$?**

This problem is called the question of the topological *Invariance of Dimension*, and it was one of the principal problems faced by the mathematicians who first developed topology. The problem was important because the use of dimension in the description of the physical space we dwell in was called into question by Cantor's discovery. The first proof of the topological invariance of dimension used new methods of a combinatorial nature (Chapters 9, 10, 11).

The combinatorial aspects of topology play a similar role that approximation does in analysis: by approximating with manageable objects, we can manipulate the approximations fruitfully, sometimes identifying properties that are associated to the combinatorics, but which depend only on the topology of the limiting object. This approach was initiated by Poincaré and refined to a subtle tool by L. E. J. Brouwer (1881–1966). It was Brouwer who gave the first complete proof of the theorem of the topological invariance of dimension and his proof established the centrality of combinatorial approximation in the study of continuity.

Toward our goal of a proof of invariance of dimension, we begin by expanding the familiar definition of continuity to more general settings.

Exercises

1. Let $f: A \to B$ be any function and U, V subsets of B, X a subset of A. Prove the following about the preimage operation:

a) $U \subset V$ implies $f^{-1}(U) \subset f^{-1}(V)$.
b) $f^{-1}(U \cup V) = f^{-1}(U) \cup f^{-1}(V)$.
c) $f^{-1}(U \cap V) = f^{-1}(U) \cap f^{-1}(V)$.

Figure 1.32: t03.png

Pag. 77 - 8. Computations and covering spaces

Pag. 109 - Simplicial mappings and barycentric subdivision

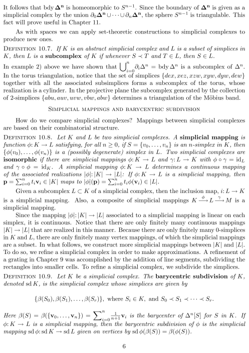

It follows that bdy $\mathbf{\Delta^n}$ is homeomorphic to S^{n-1}. Since the boundary of $\mathbf{\Delta^n}$ is given as a simplicial complex by the union $\partial_0\mathbf{\Delta^n} \cup \cdots \cup \partial_n\mathbf{\Delta^n}$, the sphere S^{n-1} is triangulable. This fact will prove useful in Chapter 11.

As with spaces we can apply set-theoretic constructions to simplicial complexes to produce new ones.

Definition 10.7. *If K is an abstract simplicial complex and L is a subset of simplices in K, then L is a* **subcomplex** *of K if whenever $S \prec T$ and $T \in L$, then $S \in L$.*

In example 2) above we have shown that $\bigcup_{i=0}^{n} \partial_i \Delta^n = \operatorname{bdy} \Delta^n$ is a subcomplex of Δ^n. In the torus triangulation, notice that the set of simplices $\{dex, xez, xzw, xyw, dyw, dew\}$ together with all the associated subsimplices forms a subcomplex of the torus, whose realization is a cylinder. In the projective plane the subcomplex generated by the collection of 2-simplices $\{abu, auv, uvw, vbw, abw\}$ determines a triangulation of the Möbius band.

Simplicial mappings and barycentric subdivision

How do we compare simplicial complexes? Mappings between simplicial complexes are based on their combinatorial structure.

Definition 10.8. *Let K and L be two simplicial complexes. A* **simplicial mapping** *is function $\phi: K \to L$ satisfying, for all $n \geq 0$, if $S = \{v_0, \ldots, v_n\}$ is an n-simplex in K, then $\{\phi(v_0), \ldots, \phi(v_n)\}$ is a (possibly degenerate) simplex in L. Two simplicial complexes are* **isomorphic** *if there are simplicial mappings $\phi: K \to L$ and $\gamma: L \to K$ with $\phi \circ \gamma = \mathrm{id}_L$ and $\gamma \circ \phi = \mathrm{id}_K$. A simplicial mapping $\phi: K \to L$ determines a continuous mapping of the associated realizations $|\phi|: |K| \to |L|$: If $\phi: K \to L$ is a simplicial mapping, then $\mathbf{p} = \sum_{i=0}^{n} t_i \mathbf{v}_i \in |K|$ maps to $|\phi|(\mathbf{p}) = \sum_{i=0}^{n} t_i \phi(\mathbf{v}_i) \in |L|$.*

Given a subcomplex $L \subset K$ of a simplicial complex, then the inclusion map, $i: L \to K$ is a simplicial mapping. Also, a composite of simplicial mappings $K \xrightarrow{\phi} L \xrightarrow{\gamma} M$ is a simplicial mapping.

Since the mapping $|\phi|: |K| \to |L|$ associated to a simplicial mapping is linear on each simplex, it is continuous. Notice that there are only finitely many continuous mappings $|K| \to |L|$ that are realized in this manner. Because there are only finitely many 0-simplices in K and L, there are only finitely many vertex mappings, of which the simplicial mappings are a subset. In what follows, we construct more simplicial mappings between $|K|$ and $|L|$. To do so, we refine a simplicial complex in order to make approximations. A refinement of a grating in Chapter 9 was accomplished by the addition of line segments, subdividing the rectangles into smaller cells. To refine a simplicial complex, we subdivide the simplices.

Definition 10.9. *Let K be a simplicial complex. The* **barycentric subdivision** *of K, denoted $\operatorname{sd} K$, is the simplicial complex whose simplices are given by*

$$\{\beta(S_0), \beta(S_1), \ldots, \beta(S_r)\}, \text{ where } S_i \in K, \text{ and } S_0 \prec S_1 \prec \cdots \prec S_r.$$

Here $\beta(S) = \beta(\{\mathbf{v}_0, \ldots, \mathbf{v}_n\}) = \sum_{i=0}^{n} \frac{1}{n+1} \mathbf{v}_i$ is the barycenter of $\Delta^n[S]$ for S in K. If $\phi: K \to L$ is a simplicial mapping, then the barycentric subdivision of ϕ is the simplicial mapping $\operatorname{sd}\phi: \operatorname{sd} K \to \operatorname{sd} L$ given on vertices by $\operatorname{sd}\phi(\beta(S)) = \beta(\phi(S))$.

Figure 1.33: t04.png

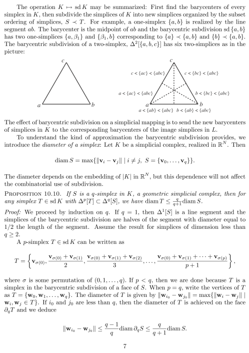

The operation $K \mapsto \operatorname{sd} K$ may be summarized: First find the barycenters of every simplex in K, then subdivide the simplices of K into new simplices organized by the subset ordering of simplices, $S \prec T$. For example, a one-simplex $\{a, b\}$ is realized by the line segment ab. The barycenter is the midpoint of ab and the barycentric subdivision $\operatorname{sd}\{a, b\}$ has two one-simplices $\{a, \beta_1\}$ and $\{\beta_1, b\}$ corresponding to $\{a\} \prec \{a, b\}$ and $\{b\} \prec \{a, b\}$. The barycentric subdivision of a two-simplex, $\Delta^2[\{a, b, c\}]$ has six two-simplices as in the picture:

The effect of barycentric subdivision on a simplicial mapping is to send the new barycenters of simplices in K to the corresponding barycenters of the image simplices in L.

To understand the kind of approximation the barycentric subdivision provides, we introduce the *diameter of a simplex*: Let K be a simplicial complex, realized in $\mathbb{R}^N$. Then

$$\operatorname{diam} S = \max\{\|\mathbf{v}_i - \mathbf{v}_j\| \mid i \neq j,\ S = \{\mathbf{v}_0, \ldots, \mathbf{v}_q\}\}.$$

The diameter depends on the embedding of $|K|$ in $\mathbb{R}^N$, but this dependence will not affect the combinatorial use of subdivision.

Proposition 10.10. *If S is a q-simplex in K, a geometric simplicial complex, then for any simplex $T \in \operatorname{sd} K$ with $\Delta^p[T] \subset \Delta^q[S]$, we have $\operatorname{diam} T \leq \frac{q}{q+1} \operatorname{diam} S$.*

Proof: We proceed by induction on q. If $q = 1$, then $\Delta^1[S]$ is a line segment and the simplices of the barycentric subdivision are halves of the segment with diameter equal to 1/2 the length of the segment. Assume the result for simplices of dimension less than $q \geq 2$.

A p-simplex $T \in \operatorname{sd} K$ can be written as

$$T = \left\{\mathbf{v}_{\sigma(0)}, \frac{\mathbf{v}_{\sigma(0)} + \mathbf{v}_{\sigma(1)}}{2}, \frac{\mathbf{v}_{\sigma(0)} + \mathbf{v}_{\sigma(1)} + \mathbf{v}_{\sigma(2)}}{3}, \ldots, \frac{\mathbf{v}_{\sigma(0)} + \mathbf{v}_{\sigma(1)} + \cdots + \mathbf{v}_{\sigma(p)}}{p+1}\right\},$$

where σ is some permutation of $(0, 1, \ldots, q)$. If $p < q$, then we are done because T is a simplex in the barycentric subdivision of a face of S. When $p = q$, write the vertices of T as $T = \{\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_q\}$. The diameter of T is given by $\|\mathbf{w}_{i_0} - \mathbf{w}_{j_0}\| = \max\{\|\mathbf{w}_i - \mathbf{w}_j\| \mid \mathbf{w}_i, \mathbf{w}_j \in T\}$. If i_0 and j_0 are less than q, then the diameter of T is achieved on the face $\partial_q T$ and we deduce

$$\|\mathbf{w}_{i_0} - \mathbf{w}_{j_0}\| \leq \frac{q-1}{q} \operatorname{diam} \partial_q S \leq \frac{q}{q+1} \operatorname{diam} S.$$

Figure 1.34: t05.png

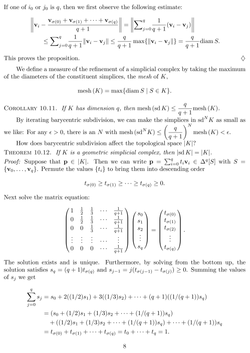

If one of i_0 or j_0 is q, then we first observe the following estimate:

$$\begin{aligned}\left\|\mathbf{v}_i - \frac{\mathbf{v}_{\sigma(0)} + \mathbf{v}_{\sigma(1)} + \cdots + \mathbf{v}_{\sigma(q)}}{q+1}\right\| &= \left\|\sum\nolimits_{j=0}^{q} \frac{1}{q+1}(\mathbf{v}_i - \mathbf{v}_j)\right\| \\ &\leq \sum\nolimits_{j=0}^{q} \frac{1}{q+1}\|\mathbf{v}_i - \mathbf{v}_j\| \leq \frac{q}{q+1}\max\{\|\mathbf{v}_i - \mathbf{v}_j\|\} = \frac{q}{q+1}\operatorname{diam} S.\end{aligned}$$

This proves the proposition. ◊

We define a measure of the refinement of a simplicial complex by taking the maximum of the diameters of the constituent simplices, the *mesh* of K,

$$\operatorname{mesh}(K) = \max\{\operatorname{diam} S \mid S \in K\}.$$

Corollary 10.11. *If K has dimension q, then* $\operatorname{mesh}(\operatorname{sd} K) \leq \frac{q}{q+1}\operatorname{mesh}(K)$.

By iterating barycentric subdivision, we can make the simplices in $\operatorname{sd}^N K$ as small as we like: For any $\epsilon > 0$, there is an N with $\operatorname{mesh}(\operatorname{sd}^N K) \leq \left(\frac{q}{q+1}\right)^N \operatorname{mesh}(K) < \epsilon$.

How does barycentric subdivision affect the topological space $|K|$?

Theorem 10.12. *If K is a geometric simplicial complex, then* $|\operatorname{sd} K| = |K|$.

Proof: Suppose that $\mathbf{p} \in |K|$. Then we can write $\mathbf{p} = \sum_{i=0}^{q} t_i \mathbf{v}_i \in \Delta^q[S]$ with $S = \{\mathbf{v}_0, \ldots, \mathbf{v}_q\}$. Permute the values $\{t_i\}$ to bring them into descending order

$$t_{\sigma(0)} \geq t_{\sigma(1)} \geq \cdots \geq t_{\sigma(q)} \geq 0.$$

Next solve the matrix equation:

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{q+1} \\ 0 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{q+1} \\ 0 & 0 & \frac{1}{3} & \cdots & \frac{1}{q+1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{q+1} \end{pmatrix} \begin{pmatrix} s_0 \\ s_1 \\ s_2 \\ \vdots \\ s_q \end{pmatrix} = \begin{pmatrix} t_{\sigma(0)} \\ t_{\sigma(1)} \\ t_{\sigma(2)} \\ \vdots \\ t_{\sigma(q)} \end{pmatrix}.$$

The solution exists and is unique. Furthermore, by solving from the bottom up, the solution satisfies $s_q = (q+1)t_{\sigma(q)}$ and $s_{j-1} = j(t_{\sigma(j-1)} - t_{\sigma(j)}) \geq 0$. Summing the values of s_j we get

$$\begin{aligned}\sum_{j=0}^{q} s_j &= s_0 + 2((1/2)s_1) + 3((1/3)s_2) + \cdots + (q+1)((1/(q+1))s_q) \\ &= (s_0 + (1/2)s_1 + (1/3)s_2 + \cdots + (1/(q+1))s_q) \\ &\quad + ((1/2)s_1 + (1/3)s_2 + \cdots + (1/(q+1))s_q) + \cdots + (1/(q+1))s_q \\ &= t_{\sigma(0)} + t_{\sigma(1)} + \cdots + t_{\sigma(q)} = t_0 + \cdots + t_q = 1.\end{aligned}$$

8

Figure 1.35: t06.png

1.2 Calculus and Geometry

1) C:/DATA-NovoVivo-137-2020/Books-137-2020/Calculus/3030334589.pdf
Obs.: An Illustrative Guide to Multivariable and Vector Calculus

2) C:/DATA-NovoVivo-137-2020/Books-137-2020/Calculus/Calculus-1.pdf

3) C:/DATA-NovoVivo-137-2020/Books-137-2020/Calculus/Calculus-2.pdf

4) C:/DATA-NovoVivo-137-2020/Books-137-2020/Calculus/eru3rehy5e4eyrhy35.pdf
Obs.: VECTOR CALCULUS, LINEAR ALGEBRA, AND DIFFERENTIAL FORMS A UNIFIED APPROACH. - 837 pages.

5) C:/DATA-NovoVivo-137-2020/Books-137-2020/Calculus/Multivariable and Vector Calculus - An Introduction.pdf

6) C:/DATA-NovoVivo-137-2020/Books-137-2020/Calculus/rey34twrt2.pdf

7) C:/DATA-NovoVivo-137-2020/Books-137-2020/Geometry/0128043911.pdf
Obs.: Riemannian Submersions, Riemannian Maps in Hermitian Geometry, and their Applications.

8) C:/DATA-NovoVivo-137-2020/Books-137-2020/Geometry/3030460398.pdf
Obs.: Differential Geometry and Lie Groups - A Computational Perspective

9) C:/DATA-NovoVivo-137-2020/Books-137-2020/Geometry/3030460460.pdf
Obs.: Differential Geometry and Lie Groups - A Second Course.

10) C:/DATA-NovoVivo-137-2020/Books-137-2020/Geometry/3030501795.pdf Obs.: Lectures on Convex Geometry.

11) C:/DATA-NovoVivo-137-2020/Books-137-2020/Geometry/3034804199.pdf Obs.: Classical Geometries in Modern Contexts - Geometry of Real Inner Product Spaces.

12) C:/DATA-NovoVivo-137-2020/Books-137-2020/Geometry/Geometry of matrices in memory of Professor L.K. Hua (1910-1985).pdf

13) C:/DATA-NovoVivo-137-2020/Books-137-2020/Geometry/Lie Groups, Differential Equations, and Geometry Advances a.pdf

14) C:/DATA-NovoVivo-137-2020/Books-137-2020/Geometry/Matrix structural analysis and introduction to finite elements.pdf

15) C:/DATA-NovoVivo-137-2020/Books-137-2020/Relativity/0387699465.pdf Obs.: Einstein's Space-Time - An Introduction to Special and General Relativity.

16) C:/DATA-NovoVivo-137-2020/Books-137-2020/Relativity/3030273466.pdf Obs.: Special Relativity - An Introduction with 200 Problems and Solutions.

17) C:/DATA-NovoVivo-137-2020/Books-137-2020/Relativity/3642038360.pdf Obs.: Special Relativity.

18) C:/DATA-NovoVivo-137-2020/Books-137-2020/Relativity/8126532548.pdf Obs.: Introduction to Special Relativity and Space Science.

19) C:/DATA-NovoVivo-137-2020/Books-137-2020/Sonar-Books-import-20200904/Sonar - The History of the Priority Dispute between Newton and Leibniz - Mathematics in History and Culture 2018.pdf

20) C:/DATA-NovoVivo-137-2020/Books-137-2020/Sonar-Books-import-20200904/Sonar - Handbook of Geomathematics 2010.pdf

21) C:/DATA-NovoVivo-137-2020/Books-137-2020/Sonar-Books-import-20200904/Sonar - Handbook of Geomathematics. Vol.2 - 2015.pdf

22) C:/DATA-NovoVivo-137-2020/Books-137-2020/Sonar-Books-import-20200904/Sonar - 3000 Jahre Analysis - Geschichte, Kulturen, Menschen 2011.pdf

23) C:/DATA-NovoVivo-137-2020/Books-137-2020/Sonar-Books-import-20200904/Sonar - Grundwissen Mathematikstudium - Höhere Analysis, Numerik und Stochastik 2015.pdf

24) C:/DATA-NovoVivo-137-2020/Books-137-2020/Sonar-Books-import-20200904/Sonar - Numerik - Eine lebendige und gut verständliche Einführung mit vielen Beispielen 2019.pdf

25) C:/DATA-NovoVivo-137-2020/Books-137-2020/Celestial-Mechanics/The Geometry of Celestial Mechanics - 2016.pdf Obs.: Diretório Todo !!!!

26) C:/DATA-NovoVivo-137-2020/Books-137-2020/Exact Space-Times in Einstein's General Relativity/1107406188.pdf

27) C:/DATA-NovoVivo-137-2020/Books-137-2020/Hamiltonian Mechanics of Gauge Systems/0521131871.pdf

28) C:/DATA-NovoVivo-137-2020/Books-137-2020/Lagrangian Intersection Floer Theory/Lagrangian Intersection Floer Theory.pdf

29) C:/DATA-NovoVivo-137-2020/Books-137-2020/Lagrangian Intersection Floer Theory/Lagrangian Intersection Floer Theory - Anomaly and Obstruction, Part I.pdf

30) C:/DATA-NovoVivo-137-2020/Books-137-2020/Lagrangian Intersection Floer Theory/Lagrangian Intersection Floer Theory - Anomaly and Obstruction, Part II.pdf

31) C:/DATA-NovoVivo-137-2020/Books-137-2020/Lagrangian Intersection Floer Theory/Lagrangian Intersection Floer Theory - Anomaly and Obstruction, Part I+II { 2009.pdf

32) C:/DATA-NovoVivo-137-2020/Books-137-2020/Shapes, Shells, Nuclear Structure/0521019664.pdf Obs.: SHAPES AND SHELLS IN NUCLEAR STRUCTURE

33) C:/DATA-NovoVivo-137-2020/Books-137-2020/Lagrangian Quantum Field Theory in Momentum Picture, Free Fields/Lagrangian quantum field theory in momentum picture free fields 2008.pdf

34) C:/DATA-NovoVivo-137-2020/Books-137-2020/Lagrangian Quantum Field Theory in Momentum Picture, Free Fields/Handbook of Normal Frames and Coordinates 2006.pdf

35) C:/DATA-NovoVivo-137-2020/Books-137-2020/Exploring the Quantum Atoms, Cavities, and Photons/0198509146.pdf

36) http://mphitchman.com/geometry/section1-3.html

37) https://ocw.mit.edu/courses/physics/8-962-general-relativity-spring-2020/

38) http://www.if.ufrgs.br/tex/ s01043/20032/Humberto/index.html

Sabemos que existem apenas 4 forcas, ou interacões, fundamentais na natureza. São elas a interacão gravitacional, a interacão eletromagnética, a interacão forte e a interacão fraca. A tabela mostra detalhes sobre estas forcas que serão logo explicados:

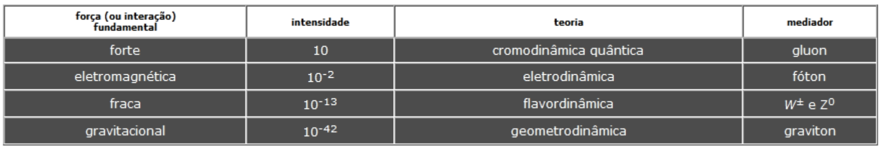

força (ou interação) fundamental	intensidade	teoria	mediador
forte	10	cromodinâmica quântica	gluon
eletromagnética	10^{-2}	eletrodinâmica	fóton
fraca	10^{-13}	flavordinâmica	$W^{\pm}$ e Z^0
gravitacional	10^{-42}	geometrodinâmica	graviton

Figure 1.36: 4f01.png

Vamos explicar o conteúdo da tabela.

Intensidade: Os valores acima atribuidos para as intensidade das forças não devem ser considerados de modo absoluto. Voce verá valores bastante diferentes em vários livros, em particular no que diz respeito à força fraca. O cálculo desta intensidade depende da natureza da fonte e a que distância estamos fazendo a medição. O que importante notar é a razão entre elas: a força gravitacional é, de longe, a mais fraca entre todas, porém é a de maior alcance, sendo a responsável pela estabilidade dinâmica de todo o Universo. Teoria: Vemos na tabela que cada força está associada a uma teoria física. Vejamos alguns detalhes: Força gravitacional: A teoria clássica da gravitação é a lei de Newton da Gravitação Universal. Sua generalização relativística é a teoria da Gravitação de Einstein, também chamada de Teoria da Relatividade Geral de Einstein. O melhor termo para ela seria Geometrodinâmica, uma vez que a relatividade geral geometriza a gravitação. Para descrever os estágios iniciais da formação do Universo precisamos de uma teoria quântica da gravitação, algo que os físicos ainda não possuem, apesar dos enormes esforços desenvolvidos para isto. Eletrodinâmica: Esta é a teoria física que descreve os fenômenos elétricos e magnéticos, ou seja as forças eletromagnéticas. A formulação clássica da Eletrodinâmica foi feita por James Clerk Maxwell. A teoria clássica construída por Maxwell já era consistente com a teoria da relatividade especial de Einstein. O "casamento" desta teoria com a mecânica quântica, ou seja, a construção de uma "Eletrodinâmica Quântica", foi realizada por grandes nomes da física tais como Feynman, Tomonaga e Schwinger nos anos que compõem a década de 1940. Fraca: As forças fracas são aquelas que explicam os processos de decaimento radiativo, tais como o decaimento beta nuclear, o decaimento do pion, do muon e de várias partículas "estranhas". É interessante notar que esta força não era conhecida pela física clássica e que sua formulação como teoria é estritamente quântica. A primeira teoria das interações fracas foi apresentada por Fermi em 1933. Mais tarde ela foi aperfeiçoada por Lee, Yang, Feynman, Gell-Mann e vários outros nos anos da década de 1950. Sua forma atual é devida a Glashow, Weinberg e Salam, que a propuseram nos anos da década de 1960. A nova teoria das interações fracas, que é chamada de flavordinâmica por causa de uma das propriedades intrínsecas das partículas elementares, é mais justamente conhecida como Teoria de Glashow-Weinberg-Salam. Nesta teoria, as interações fraca e eletromagnética são apresentadas como manifestações diferentes de uma única força, a força eletrofraca. Esta unificação entre a interação fraca e a interação eletromagnética reduz o número de forças existentes no Universo a apenas 3: força gravitacional, força forte e força eletrofraca. Forte: As forças fortes são aquelas responsáveis pelos fenômenos que ocorrem a curta distância no interior do núcleo atômico. A estabilidade nuclear está associada a forca forte. E ela que mantém o núcleo unido evitando que os prótons que os constituem, por possuirem a mesma carga elétrica, simplesmente sofram uma intensa repulsão e destruam o próprio atomo. Se a forca forte não existisse a matéria que forma o Universo, tal como o conhecemos, também não existiria. Protons e neutrons não conseguiriam se formar. Nós, seres humanos, não poderíamos existir. O trabalho pioneiro sobre as forcas fortes foi realizado por Yukawa em 1934 mas até meados da década de 1970 não havia, realmente, uma teoria capaz de explicar os fenômenos nuclear. Foi então que surgiu a cromodinâmica quântica. Mediadores: Após a física ter abandonado o conceito de "ação-a-distância", foi introduzido o conceito de "campo". Cada partícula criava a sua volta uma perturbação, seu "campo", que era sentido pelas outras partículas. A Teoria Quântica de Campos (TQC) introduziu o conceito de "mediadores". Segundo a TQC cada uma das forcas que existem na natureza e mediada pela troca de uma partícula que e chamada de "mediador". Estes mediadores transmitem a forca entre uma partícula e outra. Assim, a forca gravitacional e mediada por uma partícula chamada graviton. A forca eletromagnética e mediada pelo fóton, a forca forte pelos gluons e as forcas fracas pelas partículas W e Z0, que são chamadas de bósons vetoriais intermediários. Isto complica ainda mais o estudo das interações entre as partículas. Veja que antes descrevamos a interação entre dois prótons como sendo a interação entre duas partículas. Hoje, sabendo que os prótons são partículas compostas por 3 quarks, vemos que a interação entre dois prótons e, na verdade, uma interação entre 6 quarks que trocam gluons incessantemente durante todo o processo. Só para te avisar, existem 8 tipos de gluons. Como voce pode ver, aqui não existe simplicidade.

Forca Fraca

Em 1896, Henry Becquerel obteve os primeiros indícios da forca nuclear fraca na descoberta da radioatividade. Nos anos 30 houve um grande impulso no entendimento desta forca: naquele tipo particular de radioatividade descoberto por Becquerel um nêutron dentro do núcleo atômico se transforma em um próton , criando, ao mesmo tempo, um elétron e uma outra partícula conhecida como antineutrino, ambas lancadas para fora do núcleo. Esse evento, conhecido como decaimento beta não podia ser consequencia de outros tipos de forcas. A forca nuclear forte mantém os prótons e nêutrons juntos no núcleo e a eletromagnética tenta separar os prótons. A forca gravitacional não faz nada do gênero.

Inicialmente houve grande confusão em relação a essa forca. Os físicos não gostavam da teoria: ela funcionava bem para o decaimento beta , mas quando era aplicada a processos mais exóticos fornecia resultados sem sentido.

Em 1967, o paquistanês radicado na Inglaterra Abdus Salam e o americano Steven Weinberg propuseram teorias que unificavam estas interações com a forca eletromagnetica , exatamente como Maxwell unificara eletricidade e magnetismo cerca de cem anos antes.

Ambos sugeriram que, alem dos fotons, existiam outras partículas, tipo fotons, que eram previstas teoricamente: partículas W eletricamente carregadas, positivamente e negativamente, cuja troca induz a forca fraca da radioatividade beta e partículas Z, neutras. o W significa fraco, em inglês weak. A partícula irmã, foi batizada de Z por Weinberg porque tinha carga zero e por ser Z a ultima letra do alfabeto. (Weinberg esperava que esta fosse a ultima partícula da família).

Na teoria de Weinberg e Salam, o foton, cuja emissão e absorcão produz a forca eletromagnetica, e as partículas W e Z , sob energias maiores do que 100 GeV, atuariam da mesma maneira. Sob energias mais baixas, esta simetria entre as partículas W positivo, W negativo e Z nulo seria quebrada e todas elas ganhariam massas volumosas tornando o alcance das forcas que carregam muito pequeno. Naquela epoca poucas pessoas acreditavam na teoria e, alem disso, não existia acelerador de partículas potente o suficiente para atingir 100 GeV.

Em 1978, um novo experimento em Stanford mediu a forca fraca entre eletrons e nucleos atômicos confirmando o que a teoria tinha previsto e em 1979 Weinberg e Salam foram indicados para o Prêmio Nobel de Física. Em 1983 um grupo de varias centenas de físicos, sob a batuta de Carlos Rubbia, descobriu a partícula W, e em 1984, a partícula Z cujas existências e propriedades ja haviam sido previstas na teoria original. Rubbia recebeu o Prêmio Nobel em 1984.

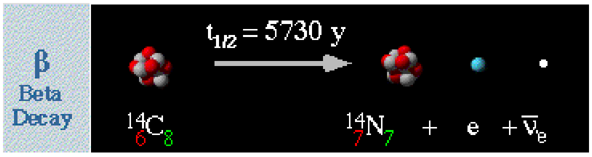

Figure 1.37: 4f02.png

Forca Forte

Essa forca mantem os quarks juntos no proton e no nêutron, e estes reunidos no nucleo do atomo. Provavelmente essa forca deve ser conduzida por outra partícula de spin-1, conhecida como gluon, que interage consigo mesma e com os quarks.

A forca nuclear forte apresenta uma curiosa propriedade, chamada confinamento: ela sempre liga partículas reunidas em combinacões incolores. Não se pode ter um unico quark em si mesmo porque ele teria uma cor (vermelha, verde ou azul). Em vez disso, um quark vermelho tem que ser ligado a um verde e a outro azul por uma 'fileira' de glúons (vermelho + verde + azul = branco). Tal trio constitui um próton ou um nêutron.

Outra possibilidade é um par composto de um quark e um antiquark (vermelho + antivermelho, ou verde+antiverde, ou azul + antiazul = branco). Tais combinações produzem as partículas conhecidas como mésons, que são instaveis porque o quark e o antiquark podem se anular, produzindo elétrons e outras partículas. Da mesma forma, o confinamento evita que se tenha um único glúon em si mesmo, porque os glúons também têm cor. Em vez disso, é preciso que se tenha um conjunto de glúons cujas cores se somem no branco. Tal conjunto forma uma partícula instável, chamada glueball.

O fato de que o confinamento evita que se observe um quark isolado, ou um glúon, pode parecer criar toda a noção dos quarks e glúons como partículas de alguma forma metafísicas. Entretanto, há uma outra propriedade da força nuclear forte, chamada liberdade assintótica, que torna bem definidos os conceitos de quarks e glúons. A energias normais, a força nuclear forte é realmente forte e liga os quarks muito intimamente. Entretanto, experimentos com grandes aceleradores de partículas indicam que, em altas energias, a força forte se torna muito mais fraca, e que os quarks e gluons se comportam como partículas independentes.

Gravidade

Uma definição fácil para essa força: tudo atrai todo o resto. Alem disso, essa força é universal, isto é, age sobre qualquer partícula.

A força gravitacional é, de longe, a mais fraca de todas. É muito, muito fraca mesmo. Tão fraca que seria praticamente impossível percebê-la, não fosse duas propriedades fundamentais: 1-é sempre atrativa 2- age através de grandes distâncias. (o que quero dizer por grandes distâncias? Imensas. Há indícios que sua ação se prolonga para muito longe de nossa galáxia, até os confins do universo)

A essa força devemos a explicação de muitos fenômenos: a Terra orbitando o Sol, por exemplo. Também explica as marés. Sim, a lua atrai a água para cima sob ela e provoca as marés. Havia muitas dúvidas devido ao fato da maré subir e descer em 12 horas. Foi Newton que explicou o fenômeno em detalhes. O mecanismo é assim: a lua atrai a Terra e a água sobre sua superfície. Mas o ponto de equilíbrio se encontra no centro. Só que temos que atentar para os vários níveis de profundidades da água. Assim, a água mais próxima da lua é atraída mais do que a restante. Consequentemente, a água mais afastada é atraída menos que a média. Soma-se a isso o fato de a água conseguir fluir, ao contrário da Terra, que é rígida. A combinação dessas duas coisas formam as marés.

A terra é redonda porque tudo atrai todo o resto, juntando-se ao máximo possível. Na verdade a Terra não e 100% esférica por que está girando e produz efeitos centrífugos que tendem a se opor a gravidade na região do equador. Então, o sol, a Terra e a lua são quase esferas.

Você pode pensar: então, quando caminho perto de outra pessoa, nossos corpos se atraem? Sim, e exatamente isso que acontece, e não precisa estar perto dessa outra pessoa. Você atrai qualquer corpo, vivo ou inanimado. Eu sei que estou te atraindo agora. Mas atrair e uma coisa, e colocar em movimento e outra. A forca gravitacional e muito fraca, e seus efeitos, apesar de existirem em qualquer lugar, só são observados na interacão entre corpos muito macicos, como dois planetas por exemplo.

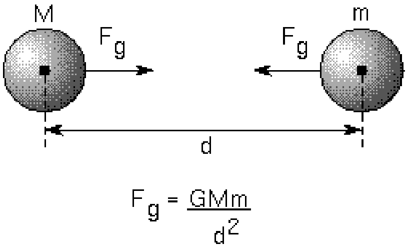

Figure 1.38: 4f03.png

Forca Eletromagnética

A forca electromagnética determina as maneiras em que partículas com carga eléctrica interactuam umas com as outras, e com campos magnéticos. Esta forca pode ser atrativa ou repulsiva. Cargas eléctricas com o mesmo sinal (duas positivas ou duas negativas) repelem-se; com sinais diferentes atraem-se.

A forca electromagnética segura os electrões [cargas negativas] nas suas orbitais, a volta do núcleo [carga positiva] do atomo. Esta forca mantem os atomos em existencia.

A forca electromagnética controla o comportamento de partículas com carga eléctrica e de plasmas (um plasma e uma mistura de quantidades iguais de iões positivos e electrões, negativos) como, por exemplo, em proeminencias solares, lacos coronais, luminancias, e outros tipos de actividade solar.

A forca electromagnética também governa a emissão e absorpcão de luz e outras formas de radiacão electromagnética. Luz e emitida quando uma partícula com carga eléctrica e acelerada (por exemplo, quando um electrão passa perto dum ião, ou interactua com um campo magnético) ou quando um electrão desce dum nível de energia mais alto para um mais baixo, num atomo (duma orbita' afastada para uma orbita' proxima a volta do nucleo do atomo).

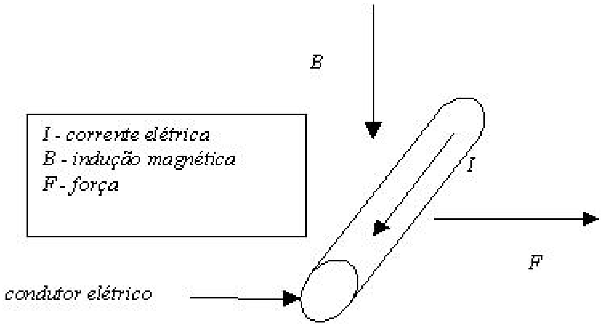
B
I - corrente elétrica
B - indução magnética
F - força
I
F
condutor elétrico

Figure 1.39: 4f04.png

Index

Calculus and Geometry, 26

Projetive Geometry - Divided Spheres, 1

Radiana Measure, 3

Theorem of Pythagoras - calculus, 15

Bibliography

entries

!" #$%#& ! ' ()$
* ' # % % + ! # (, $% - !$# % - ' . * %/ 0 12 , 34".-

,)
%(5%(6(* " . $ - '$#) !$# (76 4 + ##8$*. ,) %(5%(6(* " . $ -
'$#) !$# (4".-

!" #$%#& ! '$9 %
/ , $% (%. %(, $% "$ %# % %#: # $"" $() # (*$ ($* # $% (%. %#$* %:!7 #) 64".-

!" #$%#& ! $ *
; %. 0$:#)43 5<"* %/2 % # /$# %/ $%. (%/ % '$#) !$# (; 9)=: ;$ *
0 >34".-

, 1 , 5, ? ; 9 /. 4".-

, 1 , 5, ? ? 1 @ (# + ! # 6 * .
7* ! $%. ,) 6 8 @ (# + ! # 6 * . 7* ! $%. ,) 6 8 0 34".-

, 1 , 5, ? ? 1 *$% % 8 # !"* <
: # % *$% % 8 # !"* < : # % ,) ,) 6 - % $ ' . * >#) 4".-

, 1 , 5, ? ? 1 :7. %
: -$(:7. % : -$(-4".-

, 1 , 5, ? ? A B :%.$# % '$#) 2 >#)
5. # % $% #4 # 1 4".-

7 4 5 # * #C % . #D #) ; 9

#) ; 9 ,) + ! # 6 - "$(# ! % %# .:(# % # " ($* $%. + % $*
*$# #6 " %/ *$/ 8 E 9 0 34".-

F/4 0 3 G)$"# G : -$($%. : $#: 4

7 4 $ $! # H %/ $: -$(4

#$ # >

$3 #) ; 9 , $% - !$# % + ! # 6 % %# .:(# % # 6!! # 6
I+ / 54 '$ # %J , $% - !$# % + ! # 6 % %#0H * 74 /34.@ :

F/4 0A 3
)$"# 1
5 % ! #

14 $ #6
" .:(# - #8 -* (# % $ # $% *$# % $ #$# %4 ;6 (% . %/
#) - < . " %# - $()2 8 #)$# % #) $ # $% *$# % % $ #$# % ($% 7
K:$* # $ -* (# %4

#) ; 9 , $% - !$# % + ! # 6 % %# .:(# % # 6!! # 6
I+ / 54 '$ # %J , $% - !$# % + ! # 6 % %#0H * 74 /34".-

Chapter 7

Even Isometries

§7.1 Parity

A product of two reflections is a translation or a rotation. By considering the fixed points of each, we see that neither a translation nor a rotation can be equal to a reflection. Thus, $\sigma_n\sigma_m \neq \sigma_l$ for lines l, m, n. When a given isometry is expressed as a product of reflections, the number of reflections is not invariant; indeed, we can always add 2 to the number of reflections in a given product by inserting "$\sigma_l\sigma_l$" into the product. Although a product of two reflections cannot be a reflection, we know that in some cases a product of three reflections is a reflection. We shall see this is possible only because both 3 and 1 are odd integers. In mathematics, *parity* refers to the property of an integer being either even or odd. An isometry that is a product of an even number of reflections is said to be ***even***; an isometry that is a product of an odd number of reflections is said to be ***odd***. Since an isometry is a product of reflections, then an isometry is even or odd. However, the definition will be useful only if we show that no isometry is both even and odd. Of course no integer can be both even and odd, but is it not conceivable some product of ten reflections could be equal to some product of seven reflections? To show this is impossible, we first show that a product of four reflections is always equal to a product of two reflections.

Our argument that a product of four reflections is a product of two reflections depends on two applications of the lemma: If P is a point and a and b are lines, then there are lines c and d with c passing through P such that $\sigma_b\sigma_a = \sigma_d\sigma_c$. Suppose a, b, P are given. If $a \parallel b$, let c be the line through P that is parallel to a; if a and b intersect at C, let c be a line through P and C. So a, b, c are either parallel to each other or else concurrent. In either case, there is a line d such that $\sigma_b\sigma_a\sigma_c = \sigma_d$. Then $\sigma_b\sigma_a = \sigma_d\sigma_c$ and we have the

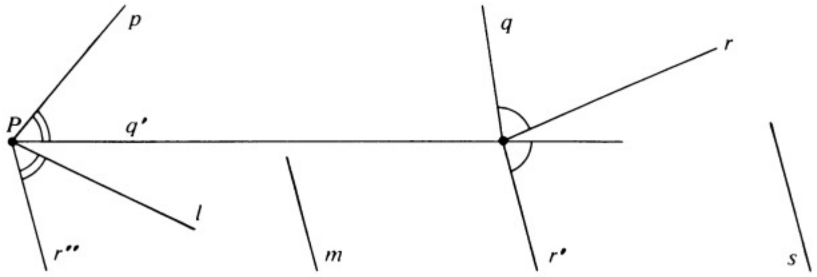

Figure 7.1

desired lemma. Now, suppose product $\sigma_s\sigma_r\sigma_q\sigma_p$ is given. We want to show this product is equal to a product of two reflections. Let P be a point on line p. See Figure 7.1 for an example. By the lemma, there are lines q' and r' such that $\sigma_r\sigma_q = \sigma_{r'}\sigma_{q'}$ with P on q'. Again by the lemma, there are lines r'' and m such that $\sigma_s\sigma_{r'} = \sigma_m\sigma_{r''}$ with P on r''. Since p, q', r'' are concurrent at P, then there is a line l such that $\sigma_{r''}\sigma_{q'}\sigma_p = \sigma_l$. Therefore,

$$\sigma_s\sigma_r\sigma_q\sigma_p = \sigma_s\sigma_{r'}\sigma_{q'}\sigma_p = \sigma_m\sigma_{r''}\sigma_{q'}\sigma_p = \sigma_m\sigma_l.$$

Not only are there lines such that the given product of four reflections is equal to $\sigma_m\sigma_l$, but our proof even tells how to find such lines.

Theorem 7.1. *A product of four reflections is a product of two reflections.*

Given a long product of reflections, we can use this theorem repeatedly to replace the first four reflections by two reflections until we have obtained a product with less than four reflections. By repeated application of the theorem to an even isometry, we can reduce the even isometry to a product of two reflections. Also, by repeated application of the theorem to an odd isometry, we can reduce the odd isometry to a product of three reflections or to a reflection. Therefore, to show an isometry cannot be both even and odd, we need to show only that a product of two reflections cannot equal a reflection or a product of three reflections. Assume there are lines p, q, r, s, t such that $\sigma_r\sigma_q\sigma_p = \sigma_s\sigma_t$. Then, we have shown above that there are lines l and m such that $\sigma_m\sigma_l = \sigma_s\sigma_r\sigma_q\sigma_p = \sigma_s\sigma_s\sigma_t = \sigma_t$. We have a contradiction since $\sigma_m\sigma_l$ is a translation or a rotation and cannot be equal to reflection σ_t. A product of two reflections is never equal to a reflection or a product of three reflections. We have our desired theorem concerning parity.

Theorem 7.2. *An even isometry is a product of two reflections. An odd isometry is a reflection or a product of three reflections. No isometry is both even and odd.*

The even isometries are the translations and the rotations. Since the involutory isometries are the halfturns and the reflections, the theorem above gives the following partition of these involutions.

Theorem 7.3. *An even involutory isometry is a halfturn; an odd involutory isometry is a reflection.*

An isometry and its inverse have the same parity, since the inverse of a product of reflections is the product of the reflections in reverse order. So the set $\mathscr{E}$ of all even isometries has the inverse property. Further, the set $\mathscr{E}$ has the closure property since the sum of two even integers is even. That's all we need to show that $\mathscr{E}$ forms a group.

Theorem 7.4. *The even isometries form a group* $\mathscr{E}$.

$\mathscr{E}$ will always denote the group of even isometries. So $\mathscr{E}$ consists of the translations and the rotations. Since some persons call only these isometries in $\mathscr{E}$ *motions* while others call any isometry in $\mathscr{I}$ a *motion*, we shall get along without using the term at all.

$\alpha\sigma_P\alpha^{-1} = \sigma_{\alpha(P)}$ for any isometry α since $\alpha\sigma_P\alpha^{-1}$ is an even involutory isometry that fixes $\alpha(P)$. That's true but a little fast. Suppose α and β are isometries. Then, $\alpha\beta\alpha^{-1}$ is an involution iff β is an involution since $(\alpha\beta\alpha^{-1})^2 = \alpha\beta^2\alpha^{-1} = \iota$ iff $\beta^2 = \iota$ and since $\alpha\beta\alpha^{-1} = \iota$ iff $\beta = \iota$. For example, since σ_P and σ_m are involutions, then $\alpha\sigma_P\alpha^{-1}$ and $\alpha\sigma_m\alpha^{-1}$ are involutions. Further, $\alpha\beta\alpha^{-1}$ and β must have the same parity since α and its inverse α^{-1} have the same parity. For example, $\alpha\sigma_P\alpha^{-1}$ is even because σ_P is even, and $\alpha\sigma_m\alpha^{-1}$ is odd because σ_m is odd. Since $\alpha\sigma_P\alpha^{-1}$ is an even involution, then $\alpha\sigma_P\alpha^{-1}$ must be a halfturn (Theorem 7.3). Hence, since halfturn $\alpha\sigma_P\alpha^{-1}$ clearly fixes point $\alpha(P)$, then $\alpha\sigma_P\alpha^{-1}$ must be the halfturn about $\alpha(P)$. That is, $\alpha\sigma_P\alpha^{-1} = \sigma_{\alpha(P)}$. In similar fashion, since $\alpha\sigma_m\alpha^{-1}$ is an odd involutory isometry, then $\alpha\sigma_m\alpha^{-1}$ is a reflection. Hence, since $\alpha\sigma_m\alpha^{-1}$ clearly fixes every point $\alpha(P)$ on line $\alpha(m)$, then $\alpha\sigma_m\alpha^{-1}$ must be the reflection in the line $\alpha(m)$. That is, $\alpha\sigma_m\alpha^{-1} = \sigma_{\alpha(m)}$.

Theorem 7.5. *If P is a point, m is a line, and α is an isometry, then*

$$\alpha\sigma_m\alpha^{-1} = \sigma_{\alpha(m)} \quad \textit{and} \quad \alpha\sigma_P\alpha^{-1} = \sigma_{\alpha(P)}.$$

Figure 7.2 illustrates an example of the first equation in the theorem with α taken to be $\rho_{P,60}$. The theorem says that if we rotate the points of the plane about P through $-60°$, then reflect in line m, and then rotate about P through $+60°$, the effect of all this is just the same as only reflecting in line n where $n = \rho_{P,60}(m)$. That seems truly amazing.

In general, $\alpha\beta\alpha^{-1}$ is called the ***conjugate*** of β by α. We next look at the conjugate $\alpha\tau_{A,B}\alpha^{-1}$ of translation $\tau_{A,B}$ by isometry α. If M is the midpoint of A and B, then point $\alpha(M)$ is the midpoint of points $\alpha(A)$ and $\alpha(B)$. Also, $\tau_{A,B} = \sigma_M\sigma_A$ and $\tau_{\alpha(A),\alpha(B)} = \sigma_{\alpha(M)}\sigma_{\alpha(A)}$. Using our method of inserting the identity in a useful form, we then have

$$\alpha\tau_{A,B}\alpha^{-1} = \alpha\sigma_M\sigma_A\alpha^{-1} = \alpha\sigma_M\alpha^{-1}\alpha\sigma_A\alpha^{-1} = \sigma_{\alpha(M)}\sigma_{\alpha(A)} = \tau_{\alpha(A),\alpha(B)}.$$

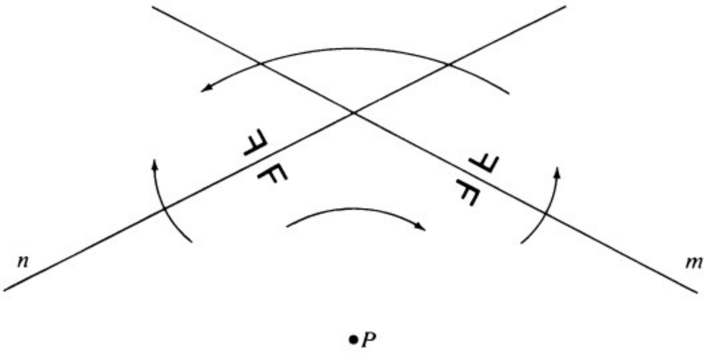

Figure 7.2

What could be nicer than the result $\alpha\tau_{A,B}\alpha^{-1} = \tau_{\alpha(A),\alpha(B)}$? Finding the conjugate of a rotation is slightly more complicated. We first examine the conjugate of $\rho_{C,\Theta}$ by σ_l. Let m be the line through C that is perpendicular to l. Then there is a line n through C such that $\rho_{C,\Theta} = \sigma_n\sigma_m$. See Figure 7.3. Now $\sigma_l(m)$ and $\sigma_l(n)$ intersect at $\sigma_l(C)$, and a directed angle from $\sigma_l(m)$ to $\sigma_l(n)$ is the negative of a directed angle from m to n. This explains the negative sign on the far right in the following calculation:

$$\sigma_l\rho_{C,\Theta}\sigma_l^{-1} = \sigma_l\sigma_n\sigma_m\sigma_l^{-1} = \sigma_l\sigma_n\sigma_l^{-1}\sigma_l\sigma_m\sigma_l^{-1} = \sigma_{\sigma_l(n)}\sigma_{\sigma_l(m)} = \rho_{\sigma_l(C),-\Theta}.$$

If $\alpha = \sigma_t\sigma_s$, then $\alpha\rho_{C,\Theta}\alpha^{-1} = \sigma_t(\sigma_s\rho_{C,\Theta}\sigma_s^{-1})\sigma_t^{-1} = \rho_{\alpha(C),+\Theta}$ with a positive sign replacing two negative signs in front of Θ. If $\alpha = \sigma_t\sigma_s\sigma_r$ then the sign in front of Θ is back to a negative sign again. We summarize our results as follows.

Theorem 7.6. *If α is an isometry, then*

$$\alpha\tau_{A,B}\alpha^{-1} = \tau_{\alpha(A),\alpha(B)} \quad \textit{and} \quad \alpha\rho_{C,\Theta}\alpha^{-1} = \rho_{\alpha(C),\pm\Theta},$$

where the positive sign applies when α is even and the negative sign applies when α is odd.

By taking $\alpha = \rho_{D,\Phi}$ in the theorem, we can show nonidentity rotation $\rho_{D,\Phi}$ does not commute with nonidentity rotation $\rho_{C,\Theta}$ unless $D = C$. We

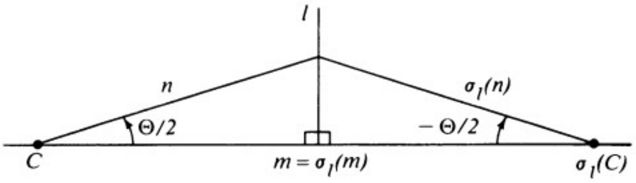

Figure 7.3

state this as a theorem here but leave the details of the argument as Exercise 7.1.

Theorem 7.7. *Nonidentity rotations with different centers do not commute.*

We are also in a position to answer the question, "When do reflections commute?" For any lines m and n, the following five statements are seen to be equivalent. (1) $\sigma_m \sigma_n = \sigma_n \sigma_m$. (2) $\sigma_n \sigma_m \sigma_n = \sigma_m$. (3) $\sigma_{\sigma_n(m)} = \sigma_m$. (4) $\sigma_n(m) = m$. (5) $m = n$ or $m \perp n$. Comparing (1) and (5), we have the answer to our question.

Theorem 7.8. $\sigma_m \sigma_n = \sigma_n \sigma_m$ *iff* $m = n$ *or* $m \perp n$.

Several equivalences between algebraic equations and geometric relations similar to that in Theorem 7.8 are indicated in Exercise 7.3. These are verified in the same manner.

We now consider products of even isometries. We already know (Theorem 3.4) that the product of two translations is a translation. We also know (Theorem 3.6) that the product of two rotations can be a translation in some cases. For example, $\sigma_B \sigma_A = \tau^2_{A,B}$; don't forget that a halfturn is a rotation of 180°. We know $\rho_{C,\Phi}\rho_{C,\Theta} = \rho_{C,\Theta+\Phi}$ by Theorem 5.10. Let's consider the product $\rho_{B,\Phi}\rho_{A,\Theta}$ of two nonidentity rotations with different centers. With $c = \overleftrightarrow{AB}$, there is a line a through A and a line b through B such that $\rho_{A,\Theta} = \sigma_c \sigma_a$ and $\rho_{B,\Phi} = \sigma_b \sigma_c$. So $\rho_{B,\Phi}\rho_{A,\Theta} = \sigma_b \sigma_c \sigma_c \sigma_a = \sigma_b \sigma_a$. When $(\Theta + \Phi)^\circ = 0^\circ$, then the lines a and b are parallel and our product is a translation. This is easier to see when the directed angles are chosen as interior angles on the same side of c; for example, in Figure 7.4 we have $\Theta/2 + \Phi/2 = -180$. On the other hand, when $(\Theta + \Phi)^\circ \neq 0^\circ$, then the lines a and b intersect at some point C and our product is a rotation.

More than that, with the directed angles chosen as interior angles on the same side of c as C, we can see by the Exterior Angle Theorem from elementary geometry that one directed angle from a to b is $(\Theta/2 + \Phi/2)^\circ$. See Figure 7.5. Hence our product $\sigma_b \sigma_a$ is a rotation about C through an angle of $(\Theta + \Phi)^\circ$. That is, $\rho_{B,\Phi}\rho_{A,\Theta} = \rho_{C,\Theta+\Phi}$. Now, what is the product of a translation

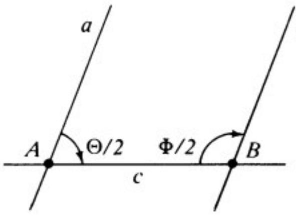

Figure 7.4

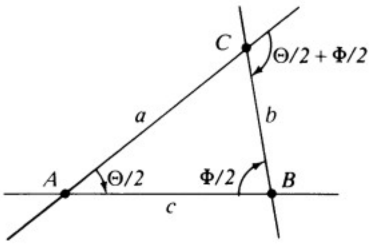

Figure 7.5

and a nonidentity rotation? In such a product we replace the translation by a product of two rotations of 180° each and obtain as the product of the three rotations a rotation of $(\Theta + 180 + 180)^\circ$, which is just a rotation of Θ°. We have proved ***The Angle-addition Theorem.***

Theorem 7.9. *A rotation of* Θ° *followed by a rotation of* Φ° *is a rotation of* $(\Theta + \Phi)^\circ$ *unless* $(\Theta + \Phi)^\circ = 0^\circ$, *in which case the product is a translation. A translation followed by a nonidentity rotation of* Θ° *is a rotation of* Θ°. *A nonidentity rotation of* Θ° *followed by a translation is a rotation of* Θ°. *A translation followed by a translation is a translation.*

The Angle-addition Theorem can also be proved by using the equations for the even isometries that will be developed in Section 9.1.

§7.2 The Dihedral Groups

We are going to compute the symmetry group of a regular polygon. As a specific example of what we are going to do, we first consider the square. We suppose the square is centered at the origin in the Cartesian plane and that one vertex lies on the positive X-axis. With the notation as in Figure 7.6, we see that the square is fixed by ρ and by σ where $\rho = \rho_{O,90}$ and $\sigma = \sigma_h$. Note that $\rho^4 = \sigma^2 = \iota$. Since the symmetries of the square form a group, then the square must be fixed by the four distinct rotations $\rho, \rho^2, \rho^3, \rho^4$ and by the four distinct odd isometries $\rho\sigma, \rho^2\sigma, \rho^3\sigma, \rho^4\sigma$. Let V_1 and V_2 be adjacent vertices of the square. Under a symmetry, V_1 may go to any one of the four vertices, but then V_2 must go to one of the two vertices adjacent to that one and the images of all remaining vertices are then determined. So

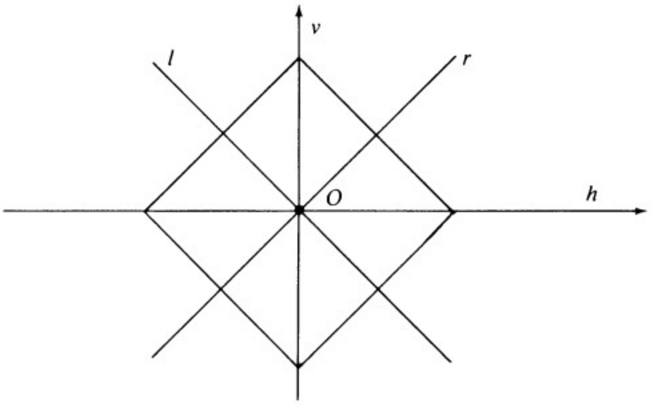

Figure 7.6

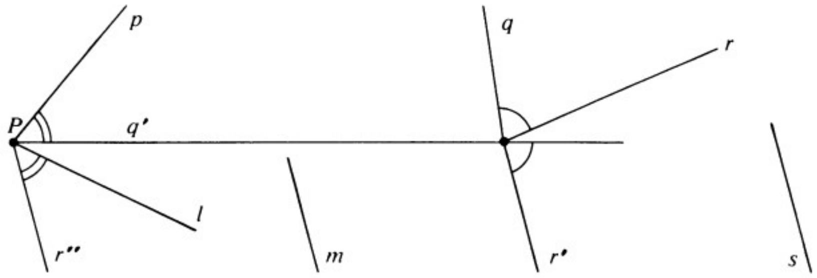
p
P
q'
l
r''
m
q
r
r'
s

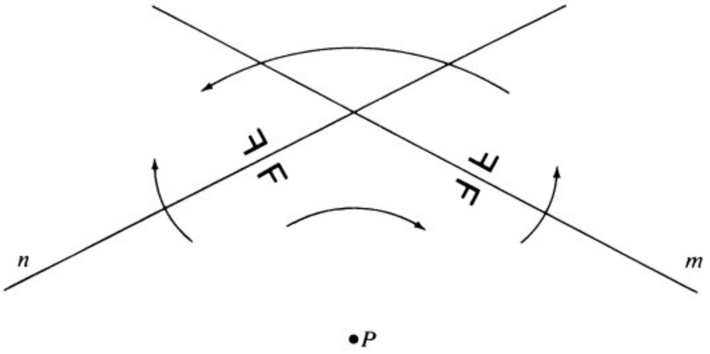

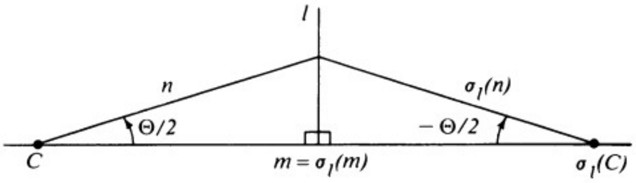
l
n
σ_l(n)
Θ/2
− Θ/2
C
m = σ_l(m)
σ_l(C)

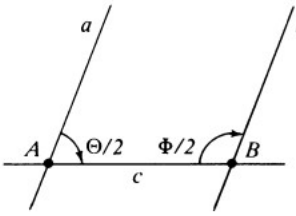

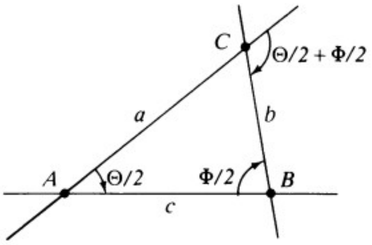
C
Θ/2 + Φ/2
a
b
A
Θ/2
Φ/2
B
c

73 #) ; 9 %# .:(# % # / (76 $:* L (9 %# .:(# % # / (76 $:*
L (94".-

(3 #) ; 9 *$ ($* %# %::!
' ()$% (*$ ($* %# %::! ' ()$% (""'* . $%. !":#$# %$* ' ()$% (4".-

2.7 Coordinate frames and transformations

Orthogonal frames such as Cartesian, cylindrical and spherical are commonly used. In these frames, the axes are orthogonal and the basis (unit vectors along the orthogonal axes) is used to describe quantities of interest in the frame. Cartesian frame or x-frame is simplest of the three. The x-frame can be transformed into y-frame through a coordinate transformation T. In doing so, we also transform the quantities defined in x-frame to y-frame, referred to as induced transformation. In continuum mechanics, the nature of transformation T and the nature of induced transformation on various quantities defined in x-frame are important classes of studies. We also consider non-orthogonal curvilinear frames in which the coordinate axes are curved and are not orthogonal. An important case is a Cartesian x-frame that transforms into a non-orthogonal curvilinear frame due to coordinate transformation T. Just like orthogonal frames, a curvilinear frame has a basis. We consider details in the following sections.

2.7.1 Cartesian frame and orthogonal coordinate transformations

For now, we consider x-frame to be Cartesian frame and transformation T (2.72) to be such that it causes translation and rigid rotation of x-frame into y-frame. For simplicity, consider a two-dimensional case.

Consider two sets of Cartesian frames in a plane: the x-frame o-x_1x_2 and the y-frame o'-y_1y_2. If the y-frame is obtained from the x-frame by a shift of the origin and without a rotation of the axes, then the transformation from the x-frame to the y-frame is a translation. If the point P has coordinates (x_1, x_2) and (y_1, y_2) in x-frame and y-frame and if (c_1, c_2) are the coordinates of the origin of y-frame with respect to x-frame, then

$$\begin{aligned} x_1 &= y_1 + c_1 \\ x_2 &= y_2 + c_2 \end{aligned} \tag{2.73}$$

or in compact form

$$x_i = y_i + c_i \quad ; \quad i = 1, 2 \tag{2.74}$$

If the origin remains fixed and the y-frame is obtained by rotating the x-frame through an angle θ in the counter clockwise direction about a line perpendicular to the plane of the x-frame but at its origin (see Figure 2.2), then the transformation relating the frames can be derived as follows. Consider a fixed point P with coordinates (x_1, x_2) and (y_1, y_2) in x-frame and y-frame. From Figure 2.2, it is rather straightforward to see that

$$\begin{aligned} x_1 &= y_1 cos\theta - y_2 sin\theta \\ x_2 &= y_1 sin\theta + y_2 cos\theta \end{aligned} \tag{2.75}$$

Let us define

$$[Q] = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \tag{2.76}$$

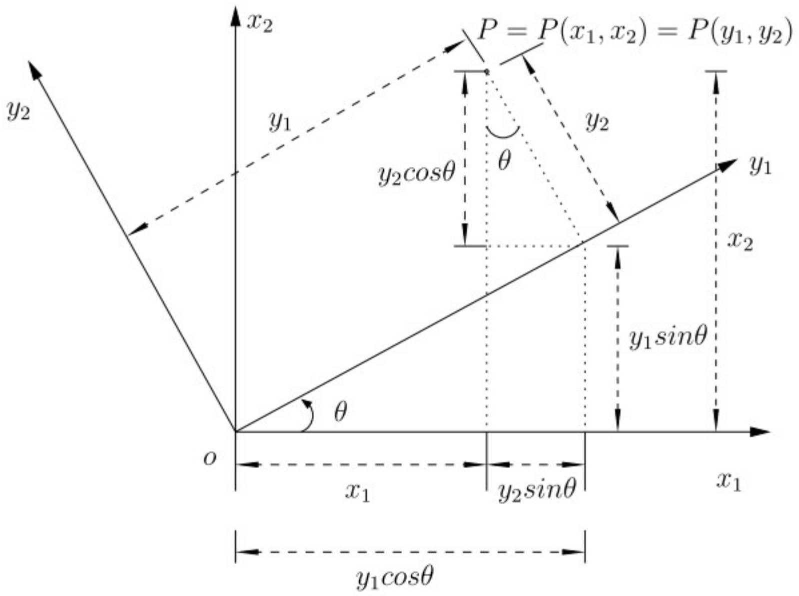

Figure 2.2: Rigid rotation of x-frame to y-frame by an angle θ

Then (2.75) can be written as

$$\begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = [Q]^T \begin{Bmatrix} y_1 \\ y_2 \end{Bmatrix} \quad ; \quad x_i = x_i(y_j) \quad ; \quad i,j = 1,2 \tag{2.77}$$

We note that $[Q]$ defines pure rotation of x-frame to y-frame and is orthogonal, i.e.,

$$[Q]^T = [Q]^{-1} \quad \text{hence} \quad [Q]^T[Q] = [Q][Q]^T = [I] \tag{2.78}$$

Using (2.78), we obtain the following from either (2.75) or (2.77).

$$\begin{Bmatrix} y_1 \\ y_2 \end{Bmatrix} = [Q] \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} \quad ; \quad y_i = y_i(x_j) \quad ; \quad i,j = 1,2 \tag{2.79}$$

That is, the coordinate transformation T defined by $[Q]$ in (2.79) is such that x-frame is transformed into y-frame. We note that both x-frame and y-frame are Cartesian frames and the transformation T defined by $[Q]$ in (2.79) is orthogonal.

Remarks

(1) In the rotation of the x-frame to y-frame, the point P remains fixed. Thus, if point P is a material point in a volume of matter, then every such point in the volume of matter has coordinates in x- and y-frames that are related through (2.79). It is important to note that in this process, the x-frame is rotated to y-frame, but the volume of matter containing point(s) P remains undisturbed.

(2) In contrast to (1), if we rotate the volume of matter through a rigid rotation, then its coordinate changes from x_i to y_i. Both are measured in the x-frame. The relationship between the new coordinates y_i and the original coordinates

x_i is not (2.79), but instead is given by (analogous to (2.79))

$$\begin{Bmatrix} y_1 \\ y_2 \end{Bmatrix} = [Q]^T \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} \tag{2.80}$$

(3) Thus, the rotation of coordinate system or frame is quite different than the rigid rotation of the volume of matter. In this book, we always consider rotation of the frame and not the rigid rotation of the volume of matter regardless of the configuration.

The relationship (2.79) can also be derived in an alternate way using unit vectors (base vectors) in x-frame and y-frame. Let $\boldsymbol{e}_1^x$, $\boldsymbol{e}_2^x$ and $\boldsymbol{e}_1^y$, $\boldsymbol{e}_2^y$ be bases in x-frame and y-frame. Then the vector $\boldsymbol{OP}$ (Figure 2.2) can be written as (using bases in x- and y-frames, i.e., index notation)

$$\boldsymbol{OP} = \boldsymbol{e}_1^x x_1 + \boldsymbol{e}_2^x x_2 = \boldsymbol{e}_1^y y_1 + \boldsymbol{e}_2^y y_2 \tag{2.81}$$

Using (2.81) it is possible to derive the transformation T relating x-frame to y-frame and y-frame to x-frame. Consider the dot products $\boldsymbol{e}_1^x \cdot \boldsymbol{OP}$ and $\boldsymbol{e}_2^x \cdot \boldsymbol{OP}$.

$$\begin{aligned} \boldsymbol{e}_1^x \cdot \boldsymbol{e}_1^x x_1 + \boldsymbol{e}_1^x \cdot \boldsymbol{e}_2^x x_2 &= \boldsymbol{e}_1^x \cdot \boldsymbol{e}_1^y y_1 + \boldsymbol{e}_1^x \cdot \boldsymbol{e}_2^y y_2 \\ \boldsymbol{e}_2^x \cdot \boldsymbol{e}_1^x x_2 + \boldsymbol{e}_2^x \cdot \boldsymbol{e}_2^x x_2 &= \boldsymbol{e}_2^x \cdot \boldsymbol{e}_1^y y_1 + \boldsymbol{e}_2^x \cdot \boldsymbol{e}_2^y y_2 \end{aligned} \tag{2.82}$$

Using $\boldsymbol{e}_1^x \cdot \boldsymbol{e}_1^x = 1$, $\boldsymbol{e}_2^x \cdot \boldsymbol{e}_2^x = 1$ and $\boldsymbol{e}_1^x \cdot \boldsymbol{e}_2^x = \boldsymbol{e}_2^x \cdot \boldsymbol{e}_1^x = 0$ in (2.82) we obtain

$$\begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{bmatrix} \boldsymbol{e}_1^x \cdot \boldsymbol{e}_1^y & \boldsymbol{e}_1^x \cdot \boldsymbol{e}_2^y \\ \boldsymbol{e}_2^x \cdot \boldsymbol{e}_1^y & \boldsymbol{e}_2^x \cdot \boldsymbol{e}_2^y \end{bmatrix} \begin{Bmatrix} y_1 \\ y_2 \end{Bmatrix} \tag{2.83}$$

In a similar fashion, if we take the dot products $\boldsymbol{e}_1^y \cdot \boldsymbol{OP}$ and $\boldsymbol{e}_2^y \cdot \boldsymbol{OP}$ and use the properties $\boldsymbol{e}_1^y \cdot \boldsymbol{e}_1^y = 1$, $\boldsymbol{e}_2^y \cdot \boldsymbol{e}_2^y = 1$ and $\boldsymbol{e}_1^y \cdot \boldsymbol{e}_2^y = \boldsymbol{e}_2^y \cdot \boldsymbol{e}_1^y = 0$, then we obtain

$$\begin{Bmatrix} y_1 \\ y_2 \end{Bmatrix} = \begin{bmatrix} \boldsymbol{e}_1^y \cdot \boldsymbol{e}_1^x & \boldsymbol{e}_1^y \cdot \boldsymbol{e}_2^x \\ \boldsymbol{e}_2^y \cdot \boldsymbol{e}_1^x & \boldsymbol{e}_2^y \cdot \boldsymbol{e}_2^x \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} \tag{2.84}$$

Let us define

$$[Q] = \begin{bmatrix} \boldsymbol{e}_1^y \cdot \boldsymbol{e}_1^x & \boldsymbol{e}_1^y \cdot \boldsymbol{e}_2^x \\ \boldsymbol{e}_2^y \cdot \boldsymbol{e}_1^x & \boldsymbol{e}_2^y \cdot \boldsymbol{e}_2^x \end{bmatrix} \tag{2.85}$$

Then (2.84) and (2.83) can be written as

$$\begin{Bmatrix} y_1 \\ y_2 \end{Bmatrix} = [Q] \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} \tag{2.86}$$

$$\begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = [Q]^T \begin{Bmatrix} y_1 \\ y_2 \end{Bmatrix} \tag{2.87}$$

That is, $[Q]$ in (2.85) is orthogonal. This form of $[Q]$ can be easily extended to the three-dimensional case as done subsequently. First, we derive a special form of $[Q]$ in (2.76) using (2.85) that corresponds to a rigid notation of the x-frame to the y-frame by an angle θ, i.e., we derive $[Q]$ of (2.76) from (2.85).

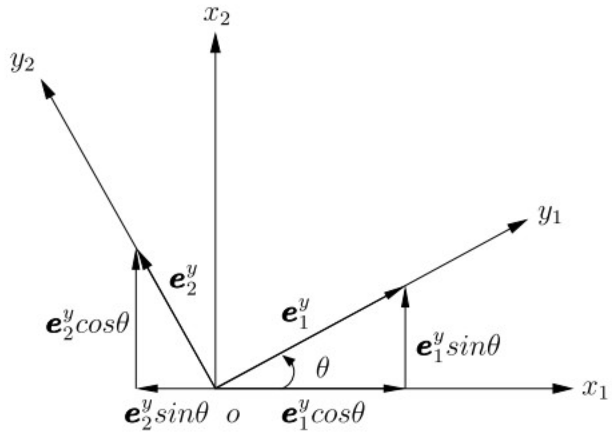
x_2
y_2
y_1
x_1
$\boldsymbol{e}_2^y$
$\boldsymbol{e}_1^y$
$\boldsymbol{e}_2^y cos\theta$
$\boldsymbol{e}_1^y sin\theta$
θ
$\boldsymbol{e}_2^y sin\theta$
o
$\boldsymbol{e}_1^y cos\theta$

Figure 2.3: Base vectors $\boldsymbol{e}_1^y$, $\boldsymbol{e}_2^y$ of y-frame and their projections in x-frame

Consider Figure 2.3 showing unit vectors $\boldsymbol{e}_1^y$ and $\boldsymbol{e}_2^y$ of the y-frame and their projection in the x-frame. The unit vectors in the x-frame, i.e., $\boldsymbol{e}_1^x$ and $\boldsymbol{e}_2^x$, are the algebraic sum of these projections. Hence we have

$$\begin{aligned} \boldsymbol{e}_1^x &= \boldsymbol{e}_1^y \cos\theta - \boldsymbol{e}_2^y \sin\theta \\ \boldsymbol{e}_2^x &= \boldsymbol{e}_1^y \sin\theta + \boldsymbol{e}_2^y \cos\theta \end{aligned} \tag{2.88}$$

and substituting from (2.88) into (2.85) and using the properties of the dot products of $\boldsymbol{e}_1^y$ and $\boldsymbol{e}_2^y$ we obtain

$$[Q] = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \tag{2.89}$$

which is the same as (2.76). This second approach of relating x-frame and y-frame through an orthogonal transformation in (2.85) can be easily extended to three dimensions. Consider a point $P(x_1, x_2, x_3)$ in the x-frame. Let $\boldsymbol{e}_i^x$ be the unit base vectors in the x-frame. Let T be an orthogonal transformation such that

$$T \ : \ y_i = y_i(x_j) \tag{2.90}$$

That is, T transforms x-frame into y-frame. Let $\boldsymbol{e}_i^y$ be the unit base vectors in the y-frame. We consider T to be an orthogonal transformation. We want to find how x- and y-frames are related through T and the specific form of T. Let y_1, y_2, y_3 be the coordinates of the point P in the y-frame. The vector $\boldsymbol{OP}$ can be represented in x- and y-frames.

$$\boldsymbol{OP} = \boldsymbol{e}_i^x x_i = \boldsymbol{e}_i^y y_i \tag{2.91}$$

As in the two-dimensional case, here also, if we use the relations

$$\begin{aligned} \boldsymbol{e}_i^x \cdot \boldsymbol{e}_j^x &= \delta_{ij} \\ \boldsymbol{e}_i^y \cdot \boldsymbol{e}_j^y &= \delta_{ij} \end{aligned} \tag{2.92}$$

and if we take the dot product of (2.91) with $\boldsymbol{e}_1^y$, $\boldsymbol{e}_2^y$ and $\boldsymbol{e}_3^y$, we can derive the following:

$$\begin{Bmatrix} y_1 \\ y_2 \\ y_3 \end{Bmatrix} = \begin{bmatrix} \boldsymbol{e}_1^y \cdot \boldsymbol{e}_1^x & \boldsymbol{e}_1^y \cdot \boldsymbol{e}_2^x & \boldsymbol{e}_1^y \cdot \boldsymbol{e}_3^x \\ \boldsymbol{e}_2^y \cdot \boldsymbol{e}_1^x & \boldsymbol{e}_2^y \cdot \boldsymbol{e}_2^x & \boldsymbol{e}_2^y \cdot \boldsymbol{e}_3^x \\ \boldsymbol{e}_3^y \cdot \boldsymbol{e}_1^x & \boldsymbol{e}_3^y \cdot \boldsymbol{e}_2^x & \boldsymbol{e}_3^y \cdot \boldsymbol{e}_3^x \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} \tag{2.93}$$

Let

$$[Q] = \begin{bmatrix} \boldsymbol{e}_1^y \cdot \boldsymbol{e}_1^x & \boldsymbol{e}_1^y \cdot \boldsymbol{e}_2^x & \boldsymbol{e}_1^y \cdot \boldsymbol{e}_3^x \\ \boldsymbol{e}_2^y \cdot \boldsymbol{e}_1^x & \boldsymbol{e}_2^y \cdot \boldsymbol{e}_2^x & \boldsymbol{e}_2^y \cdot \boldsymbol{e}_3^x \\ \boldsymbol{e}_3^y \cdot \boldsymbol{e}_1^x & \boldsymbol{e}_3^y \cdot \boldsymbol{e}_2^x & \boldsymbol{e}_3^y \cdot \boldsymbol{e}_3^x \end{bmatrix} \tag{2.94}$$

then

$$\begin{Bmatrix} y_1 \\ y_2 \\ y_3 \end{Bmatrix} = [Q] \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} \tag{2.95}$$

in which $[Q]$ is orthogonal (check as an exercise), i.e.,

$$[Q]^T = [Q]^{-1} \quad \text{and} \quad [Q]^T[Q] = [Q][Q]^T = [I] \tag{2.96}$$

Hence

$$\begin{aligned} \det([Q]^T[Q]) &= \det[I] && (2.97) \\ \det[Q]^T \det[Q] &= 1 \\ \det[Q] \det[Q] &= 1 \\ (\det[Q])^2 &= 1 \\ \det[Q] &= \pm 1 && (2.98) \end{aligned}$$

If the orthogonal matrix $[Q]$ is such that $\det[Q] = +1$, then we refer to $[Q]$ as a proper orthogonal matrix and the transformation of x- to y-frame is a proper orthogonal transformation. In such transformations, a right-handed orthogonal system (say x-frame) is transformed into another right-handed orthogonal frame (say y-frame). An orthogonal matrix whose determinant is -1 is called an improper orthogonal matrix. An improper orthogonal matrix transforms a right-handed orthogonal frame into a left-handed orthogonal frame.

2.7.2 Curvilinear coordinates (or frame)

Let x_1, x_2, x_3 or x_i $(i = 1, 2, 3)$ be the coordinates of a material point P (in fixed Cartesian x-frame) and y_1, y_2, y_3 or y_j $(j = 1, 2, 3)$ be three variables. If we can establish a correspondence between x_i and y_j, then we say that there is a transformation T or a coordinate transformation between x_i and y_j, and we can write

$$T \ : \ y_j = y_j(x_1, x_2, x_3) \quad ; \quad j = 1, 2, 3 \tag{2.99}$$

If this correspondence is one-to-one, then there exists a unique inverse of (2.99), in the following form

$$T^{-1} \,:\, x_i = x_i(y_1, y_2, y_3) \quad ; \quad i = 1, 2, 3 \tag{2.100}$$

We can show this inverse (i.e., T^{-1}) exists, and is unique, if the determinant of the Jacobian of transformation is not equal to zero.

$$\det[J] = \det\left[\frac{y_1, y_2, y_3}{x_1, x_2, x_3}\right] = \begin{vmatrix} \dfrac{\partial y_1}{\partial x_1} & \dfrac{\partial y_1}{\partial x_2} & \dfrac{\partial y_1}{\partial x_3} \\ \dfrac{\partial y_2}{\partial x_1} & \dfrac{\partial y_2}{\partial x_2} & \dfrac{\partial y_2}{\partial x_3} \\ \dfrac{\partial y_3}{\partial x_1} & \dfrac{\partial y_3}{\partial x_2} & \dfrac{\partial y_3}{\partial x_3} \end{vmatrix} \neq 0 \tag{2.101}$$

For a fixed set of values x_1, x_2, x_3, the transformation (2.100) gives three non-coincident surfaces called curvilinear surfaces. These interact at a single point P with fixed y_1, y_2, y_3. Thus, to point P, we can assign values y_1, y_2, y_3 called curvilinear coordinates of P (Figure 2.4). From Figure 2.4, we note that the intersection of any two surfaces gives a curve through P called a curvilinear coordinate line.

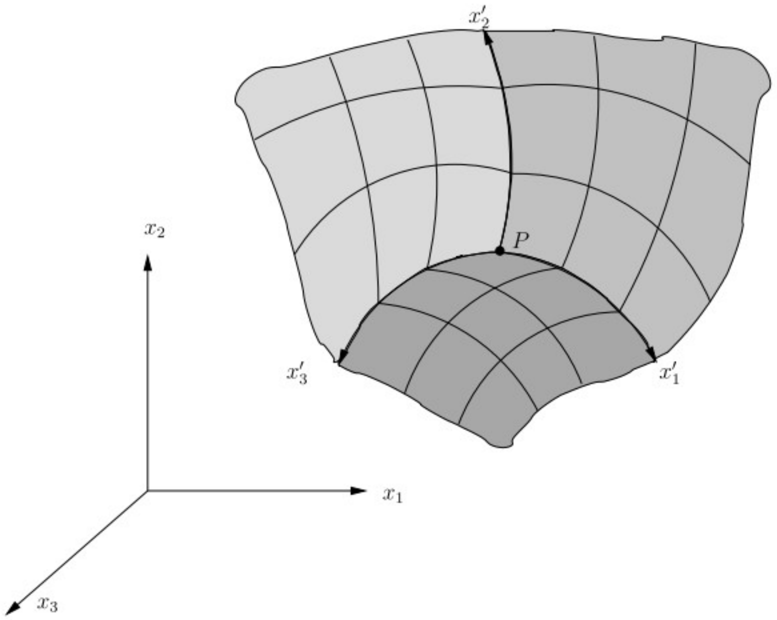

Figure 2.4: Curvilinear coordinates

2.8 Curvilinear frames, covariant and contravariant bases

To quantify the motion and deformation of matter, we need a fundamental understanding of the motion of the material particles. Consider undeformed volume

V with boundary ∂V at time t_0 as constituting the reference configuration. The reference configuration could be different than the configuration of the matter at time t_0, but assumed to be the same for the sake of simplicity. Let o-$x_1x_2x_3$ be the x-frame. Let each material particle in the reference configuration be assigned unique coordinates (x_1, x_2, x_3) in the x-frame. We assume that the matter is *homogeneous and isotropic.* Consider a material point $P(x_1, x_2, x_3)$ in this configuration. We draw lines parallel to ox_1, ox_2, ox_3 axes in the reference configuration representing material lines PA_1, PA_2, PA_3 (Figure 2.5(a)). The deformed or current configuration, i.e., volume $\bar{V}$ with boundary $\partial\bar{V}$ at time t is shown in Figure 2.5(b). Let $\bar{x}_i$ be the deformed coordinates of the material point $P(x_1, x_2, x_3)$ in the fixed coordinate x-frame. Naturally $\bar{x}_i = x_i + u_i$; u_i being displacements of the point $P(x_1, x_2, x_3)$ in the fixed x-frame. The deformed material lines $\bar{P}\tilde{A}_1$, $\bar{P}\tilde{A}_2$, $\bar{P}\tilde{A}_3$, coordinate axes $\tilde{o}\tilde{x}_1$, $\tilde{o}\tilde{x}_2$, $\tilde{o}\tilde{x}_3$ and the location of point P (i.e., $\bar{P}$) are also shown in Figure 2.5(b). The coordinate system $\tilde{o}$-$\tilde{x}_1\tilde{x}_2\tilde{x}_3$ ($\tilde{x}$-frame) is called *convected coordinate system.* This coordinate system and various measures of the state of deforming matter in this coordinate system are crucial in developing constitutive theories (considered in later chapters) because, in this coordinate system, we have deformed material lines.

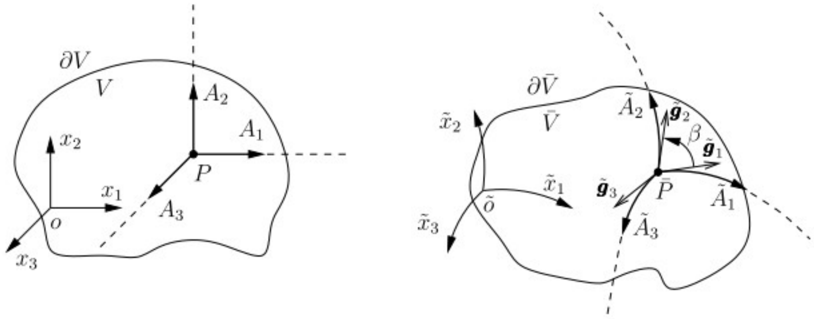

(a) Undeformed configuration

(b) Deformed configuration: covariant basis

Figure 2.5: Material lines in reference and current configurations: covariant basis

We note that at time t_0, the convected coordinate system coincides with the Cartesian coordinate system o-$x_1x_2x_3$. The convected coordinate system continuously changes in time in such a way that coordinates of a material particle in this coordinate system remain the same for all values of time. If $\tilde{x}_i$ are the coordinates of the material particle in $\tilde{o}$-$\tilde{x}_1\tilde{x}_2\tilde{x}_3$ coordinate system, then $\tilde{x}_i$ remain fixed for all values of time. An important conclusion is that coordinates x_i of a material particle $P(x_i)$ in the configuration at time $t_0 = 0$ in x-frame have exactly the same numerical values as the material particle $\bar{P}$ coordinates $\tilde{x}_i$ in $\tilde{x}$-frame. We can write the following:

$$\bar{x}_i = \bar{x}_i(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, t) \quad ; \quad i = 1, 2, 3 \tag{2.102}$$

But the material point with coordinates $\tilde{x}_i$ are the same as x_i, hence

$$\bar{x}_i = \bar{x}_i(x_1, x_2, x_3, t) \quad \text{or} \quad \bar{\boldsymbol{r}} = \bar{\boldsymbol{r}}(\boldsymbol{r}, t) \tag{2.103}$$

where $\bar{\boldsymbol{r}}$ and $\boldsymbol{r}$ are vectors with components $\bar{x}_i$ and x_i ; $i = 1, 2, 3$ respectively. Inverse of (2.103) gives

$$x_i = x_i(\bar{x}_1, \bar{x}_2, \bar{x}_3, t) \quad ; \quad \boldsymbol{r} = \boldsymbol{r}(\bar{\boldsymbol{r}}, t) \tag{2.104}$$

2.8.1 Covariant basis

Let the vectors $\tilde{\boldsymbol{g}}_1$, $\tilde{\boldsymbol{g}}_2$ and $\tilde{\boldsymbol{g}}_3$ be tangent to the material lines $\bar{P}\tilde{A}_1$, $\bar{P}\tilde{A}_2$ and $\bar{P}\tilde{A}_3$ at material point $\bar{P}$. The vectors $\tilde{\boldsymbol{g}}_1$, $\tilde{\boldsymbol{g}}_2$ and $\tilde{\boldsymbol{g}}_3$ are covariant vectors and form a non-orthogonal basis generally called *covariant basis.* Using this basis, we can define the coordinates of point $\bar{P}$ in $\tilde{x}$-frame.

In the following, we determine the explicit form of the covariant base vectors. Since $\tilde{\boldsymbol{g}}_i$ are tangent vectors to $\tilde{x}_1$, $\tilde{x}_2$ and $\tilde{x}_3$ axes, these can be obtained by differentiating the position vector $\bar{\boldsymbol{r}}$ with respect to $\tilde{x}_i$.

$$\tilde{\boldsymbol{g}}_i = \frac{\partial \bar{\boldsymbol{r}}}{\partial \tilde{x}_i} \quad ; \quad (\text{Def. of } \; \tilde{\boldsymbol{g}}_i) \tag{2.105}$$

But $\tilde{x}_i = x_i$, therefore

$$\tilde{\boldsymbol{g}}_i = \frac{\partial(\bar{\boldsymbol{r}})}{\partial x_i} \tag{2.106}$$

$$\tilde{\boldsymbol{g}}_i = \frac{\partial(\boldsymbol{e}_j \bar{x}_j)}{\partial x_i} \tag{2.107}$$

$$\text{or} \quad \tilde{\boldsymbol{g}}_i = \boldsymbol{e}_j \frac{\partial \bar{x}_j}{\partial x_i} \tag{2.108}$$

Expanding for $i = 1, 2, 3$

$$\begin{aligned}
\tilde{\boldsymbol{g}}_1 &= \frac{\partial \bar{x}_1}{\partial x_1}\boldsymbol{e}_1 + \frac{\partial \bar{x}_2}{\partial x_1}\boldsymbol{e}_2 + \frac{\partial \bar{x}_3}{\partial x_1}\boldsymbol{e}_3 \quad ; \quad \{\tilde{g}_1\}^T = \left[\frac{\partial \bar{x}_1}{\partial x_1} \, , \, \frac{\partial \bar{x}_2}{\partial x_1} \, , \, \frac{\partial \bar{x}_3}{\partial x_1}\right] \\
\tilde{\boldsymbol{g}}_2 &= \frac{\partial \bar{x}_1}{\partial x_2}\boldsymbol{e}_1 + \frac{\partial \bar{x}_2}{\partial x_2}\boldsymbol{e}_2 + \frac{\partial \bar{x}_3}{\partial x_2}\boldsymbol{e}_3 \quad ; \quad \{\tilde{g}_2\}^T = \left[\frac{\partial \bar{x}_1}{\partial x_2} \, , \, \frac{\partial \bar{x}_2}{\partial x_2} \, , \, \frac{\partial \bar{x}_3}{\partial x_2}\right] \\
\tilde{\boldsymbol{g}}_3 &= \frac{\partial \bar{x}_1}{\partial x_3}\boldsymbol{e}_1 + \frac{\partial \bar{x}_2}{\partial x_3}\boldsymbol{e}_2 + \frac{\partial \bar{x}_3}{\partial x_3}\boldsymbol{e}_3 \quad ; \quad \{\tilde{g}_3\}^T = \left[\frac{\partial \bar{x}_1}{\partial x_3} \, , \, \frac{\partial \bar{x}_2}{\partial x_3} \, , \, \frac{\partial \bar{x}_3}{\partial x_3}\right]
\end{aligned} \tag{2.109}$$

Let us introduce a new matrix notation $[J]$ containing $\tilde{\boldsymbol{g}}_i$.

$$[J] = \left[\frac{\partial\{\bar{x}\}}{\partial\{x\}}\right] = \left[\frac{\bar{\boldsymbol{x}}}{\boldsymbol{x}}\right] = \left[\frac{\bar{x}_1, \bar{x}_2, \bar{x}_3}{x_1, x_2, x_3}\right] = \begin{bmatrix} \frac{\partial \bar{x}_1}{\partial x_1} & \frac{\partial \bar{x}_1}{\partial x_2} & \frac{\partial \bar{x}_1}{\partial x_3} \\ \frac{\partial \bar{x}_2}{\partial x_1} & \frac{\partial \bar{x}_2}{\partial x_2} & \frac{\partial \bar{x}_2}{\partial x_3} \\ \frac{\partial \bar{x}_3}{\partial x_1} & \frac{\partial \bar{x}_3}{\partial x_2} & \frac{\partial \bar{x}_3}{\partial x_3} \end{bmatrix} \tag{2.110}$$

$$\text{or} \quad \boldsymbol{J} = \boldsymbol{e}_i \boldsymbol{e}_j J_{ij} = \boldsymbol{e}_i \boldsymbol{e}_j \bar{x}_{i,j}$$

This notation is due to Murnaghan [64]. In this notation, the position of the letter above the bar (–) indicates a row and those below the bar signify a column. Thus, (2.108) or (2.109) can be written as

.3 #) ; 9 $ $. < $%. %(% # %#
'$#) !$# ($ $. < 4$%.4 %(% # %#4'$#) !$# (4".-
F/4 A0 3 4 4 ,) %# $. (# % % *$ % /)# % "$(

Thus we appear to have proved both a sentence and its negation using an obviously true axiom (scheme) and elementary propositional logic.[9]

The paradoxes are no trick.[10] Listen to Tarski:

> In my judgment, it would be quite wrong and dangerous from the standpoint of scientific progress to depreciate the importance of this and other antinomies, and to treat them as jokes or sophistries. It is a fact that *we are here in the presence of an absurdity*, that we have been compelled to assert a false statement . . . *[Tarski, 1944, emphasis added]*

The paradoxes are there in the most basic places: sets; truth; raindrops. If there were no genuine paradoxes, only apparent ones, then there would still be a psychosociological project of explaining why so many people have found the paradoxes compelling. But after a while, there being a sufficient number of lucid expert witnesses should start to suggest that maybe they have all witnessed something real. Consider, then, one very gentle and elegant move. The paradoxes look unavoidable, cannot be eliminated, because they are "what they always seemed to be, proofs" [Routley, 1979, p. 302]. The conclusions of the paradoxical arguments are *true*.

0.1.2 There Are Contradictions in Plain Sight in Space

Paradoxes are not confined to abstruse contemplation of self-referring sentences or vague predicates. They have a simple visual presentation. Here is a circle:

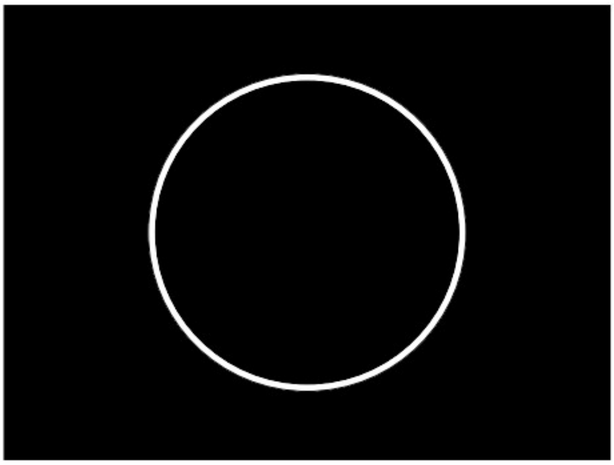

It divides the plane in which it sits into an interior and an exterior. There is a continuous path between any two points in the exterior, and any two points in the interior. But a path from a point in the exterior will not be able to reach a point in the interior without crossing the circle, which forms the *boundary* of the two parts. It looks like the plane is divided

---

[9] See Section 1.1.2, [Priest, 2006b, ch. 1], and book-length treatment in [Beall, 2009].

[10] "There are scarcely any philosophical problems of greater urgency than the liar paradox, for there are scarcely any concepts more central to our philosophical understanding than the concept of truth. . . . Quite unmistakably, our present way of thinking about truth and reference is inconsistent" [McGee, 1990, p. vii].

exclusively and *exhaustively*: assuming continuity of the plane, every point is either in the interior or the exterior, and not both. Except – what about the points on the boundary itself?

An even simpler version of this problem is in one dimension. Given a continuous line segment,

it is obvious that the line can be perfectly divided into, or is composed of, two perfectly symmetrical halves.[11] But, as your idealized knife comes down to make the cut, it touches the *center point* of the line. There must be a point there, because the line has no gaps. But points themselves have no extension and so cannot be divided; so your knife must slip either to the left or to the right of the center point. Then the resulting two pieces are not perfectly even: one piece has the center point, and the other does not.

Dividing anything in two – e.g. the left side of a line and the not-left side, the truths and the falsities, the babies and the non-babies – calls attention to the logical assumption of bivalence: that every sentence is either true or false. The implications of this assumption have been well appreciated at the "cosmic" level. According to Gödel, paradoxes are in fact due to the (purportedly mistaken) notion of

> dividing the totality of all existing things into two categories. *[Gödel, 1964, p. 519]*

But grand talk about dividing the totality of all existing things,[12] the *universe*, distracts from the fact that we have the very same sort of problem, not at the level of the universe, but at *any* medium-sized object. We do not need special properties of connectedness or closed curves to appreciate a problem here; we just need to look at the sun in the sky (or the moon, since you shouldn't stare at the sun) and wonder how it appears to be a distinct object that is nevertheless smoothly embedded in phenomenal space.

The problem localizes in asking, which portion of reality are your hands, and which portion not? The microscopic particles of matter grazing the surface between your skin and the air are a *question*. There are few things more certain than holding up a hand, and saying "here is a hand," as Moore observed.[13] At the same time, "there are always outlying particles, questionably parts of the thing, not definitely included and not definitely not included" [Lewis, 1999, p. 165]. What looks like a special problem for the universe is a commonplace problem occurring on the end of your arm.

It *seems* like a circle can sit in the plane. It *seems* like there is a universe, in which some things are hands and everything else is not. And – here's the pointy end of the stick – it seems like these things must be able to be true without plunging us into abyssal and vexatious mysteries.

Things are not always as they seem.

---

[11] Gödel once remarked (in an unpublished note; see [Putnam, 1994]) that if a geometric line segment is divided evenly at a point, it would be natural to expect the two halves of the line to be perfectly symmetric mirror images. See Section 9.2.1.

[12] Petersen echoes Gödel's point [Petersen, 2000, p. 384]: "The point is a highly metaphysical one: is it possible, in principle, to divide the world (or the universe, if you prefer) into two disjunct parts, the union of which is the world?"

[13] "Proof of an External World" [1939] in [Moore, 1993, pp. 147–170].

0.1.3 Paradoxes Are Resolved by Reversing the Order of Explanation

Sometimes things are as they seem. Usually, even: accepting things as they are, mostly, is a necessary assumption for getting around in the world. We trust perception, testimony, and the conclusions of informal reasoning as overwhelmingly veridical. Exceptions are exceptional, surprising, unsettling. *Exceptio probat regulam in casibus non exceptis*, my mother always said.

If the first phase of philosophy is to make what is unproblematic into a problem (Cartesian doubt!), then the next phase is to make what is problematic into what is not, to "show the fly the way out of the bottle,"[14] so to speak. There is a place for scientific surprises, but there is also a place for theories that make it possible to understand the world as we find it, theories that do not propound a long story about why everything we seem to see and think is wrong. An account of reality should be available of the reality we live in.[15] Dealing with the paradoxes, it has always seemed to me, should not require an elaborate apology for why the world is not really as it seems.

An impressive practitioner of this method is Richard Dedekind. In 1872, Dedekind famously advanced on the problem of characterizing continuity. In thinking about a geometric line, Dedekind took it as an adequacy condition that the line have no gaps. That is the linear continuum as we find it. And so he proposed, in essence, that *any gap in the line be thought of as itself a point* – indeed that the line is entirely made of "cuts," pairs of sets comprised of everything to the left and everything to the right. Any possible *counter*example is reconsidered as a natural *example*.

Or again, in a magnificent 1888 treatise, Dedekind faced a problem known since Proculus in the third century, that given two concentric circles, a radius cuts each circumference exactly once:

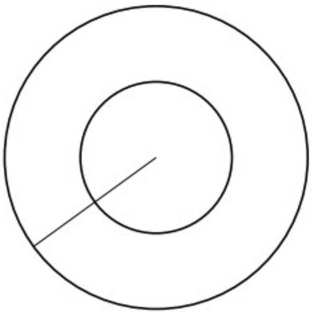

---

[14] [Wittgenstein, 1953, §309].

[15] "...the world as I found it" [Wittgenstein, 1922, §5.631]. This methodology is also found in Husserl and the phenomenological tradition – placing focus on the world of appearance and experience. See Tieszen, 2005, ch. 2, 3; [van Atten, 2007].

Apparently, then, there are exactly as many points on both circumferences, since the radius establishes a one-to-one pairing of the points on the inner circle and the points on the outer circle. But, even if the set of points on both circles is infinite – *obviously* the outer circle is bigger, so its circumference has more points!? Dedekind's move was to treat this anomaly as a *definition*: a set is infinite if and only if "it is similar to a proper part of itself" [Dedekind, 1901, def 64]. He didn't deny the data, or try to "solve" the problem, or fall back on quietism or eliminativism about circles. The anomaly that infinity is "bigger than itself" (captured in the simple $\infty + 1 = \infty$) is what makes infinity infinite. The puzzle is the answer. He took his task to be to describe the world as we find it – to explain, rather than explain away.

"What we see of the things are the things," writes Pessoa. "Why would seeing and hearing be to delude ourselves / when seeing and hearing are seeing and hearing?" The world I live in has days and nights and forests and cities, flocks of birds and bundles of recycling. There are words for these things, names and predicates, but not *only* words; the world has the things the words name, too. So there are sets, and there are properties, and there are boundaries, and some are vague and some are inconsistent. The *ordinary* leads us to an impasse. Dedekind shows how an impasse can be turned in to a way out.

0.2 The Choices

Given a paradox, there are only three options:[16]

- Reject the reasoning of the argument as invalid
- Show one of the premises is false
- Accept that the conclusion is true

A brief and biased review of how these options play out, say in the case of the liar (Section 0.1.1), is in order. (It would mostly carry over for the paradoxes of set theory, as per the next chapter, too.) This is well-worn ground, and I only intend a thumbnail sketch of some vast and important literature; various distinctions, objections, and replies are omitted.[17]

On some very familiar assumptions, truth is what is the case, every proposition is either true or false, and no proposition is both true and false. There must be a set of *all and only* the true propositions, divided out perfectly from the falsities. This is to be an *exclusive* and *exhaustive* division. Except, again – what about propositions on the "boundary"? This is the liar.

---

[16] Of course, not only three. You could excuse yourself and go for a walk. Or in a more philosophical vein, you could try to undercut the trilemma as somehow misconceived or based on some bad presupposition, like "the paradoxical sentences are meaningful." I will assume that the paradoxes are meaningful problems demanding a direct response; and so in that sense, these other options are not options.

[17] More details are in [Priest, 2006b, ch. 1]; see also opening chapters in [Field, 2008; Scharp, 2013], and many other sources, e.g., [Beall et al., 2018].

#$ # A 1 >

, ; 9 A + ! # 6 - 5%@ 6! %# $%.
)$** %/ + ! # 6 - 5%@ 6! %# $%.)$** %/ 4".-

F/4 0 3 G $ $** * % $%. *$# . B /: 4

3 , ; 9)$"

3 , ; 9

, ; 9 0M3 % $#: 2 #2 $%. :*#:
+ ! # 6 $ $ L ! % :# ((%(

3 , ; 9 96 N , * (" '$ ()
96 , * (" '$ () 4".-
F/4 A ,) L:77* % #$%# , % % $%. * $

>3 , ; 9 ,) <# + % $# %
: (7 9 % #) (%# L # 6 - $%. # !":#$# % , / % ! # 6 -)
0 B 34".-

3 , ; 9 #) O ,) ! $%.
6!! # 6 #) O ,) ! $%. 6!! # 6 ' + 0 34".-

A3 , ; 9 $ $** * ,) $. # $* #6
!"* # +: . # #) $8 - #) % 0 B 34".-
7 4 # . PPP
F/4 >10 1A3 G 6!! # 6 / :"
4 + :" - # $% - !$# %
"$(#)$# $ 6!! # ($*)$ $ -:%.$! %#$* !" #$%(% ! . %
")6 (4 Q)6 #) R # ! /)# 7 #) :/)# #)$# (!"* # *6 <$(# 6!! # 6
! #) %/ #)$# (:*. $ %*6 <("# %$**62 ")$" @: # $!
(% % %# $"" < !$# %4 *#) :/) $ 6!! # ($* 7@ (#2 :() $ $
K:$ $ ") 2)$ $ " (< # %($ $% . $* H . 0S *$# % (TU
V 4 3 !$#) !$# ($* # :(#: 2 $%6 ")6 ($* $* H$# % - :() $ #) %/
8 :*. . %$ *6 7 /$ . . $! *6 ! 9 %. - $"" < !$# "
%#$# % - #) *$# % (. $*2 #) - " %/ % $(#:$* 6!! # 6
#)$# ($% 7 /$ . . $ <$(#4 E #2 !$ 9$7*62 $((. %/ # #)) /)*6 :(
(-:* ")6 ($* #) - #) #) (%#: 62 $** ")6 ($* %# $(# %
0 %(*:. %/ / $ #63 $(# % $((.$%(8 #) $% . $ 8) ()2 # (#*6 " $9 %/2
. " %. (:($**6 :" % (#$ % ")6 ($* # :(#: " %/ $ 6!! # 6
#)$#2 $# $ -:%.$! %#$* * * - . ("# %2 %. . % ($ *6 <$(#P

9 $#
F/4 0 ?>3 G B /4 4>

B /4 4> * % $ # $% - !$# % "
#) (# "$(# :(#: - #) "$(%

8) () # $(# 4 ,) # :(#: . - % . 76 #)
" $# % - $.. # % 0 **: # $# . 76 #) "$
$** * / $! *$83 $%. !:*# "* ($# % 76 $ ($*$
* 08) () (:*. 7 $ $* %:!7 2 % #) ($
- $ (!"* < (# "$(2 $ (!"* < %:!7 34
:() $ # $% - !$# % " #) S # $ /)#
% T - * % $%. #) % # % - S"$ $** *T2 9 "
%/ #) / % - < .4

13 , ; 9 $ $** * ,) B$7 (- #) !
"$(2 , ! 2 $%. #) , <#: - $* #6 0 B 34".-
F/4 0 A3 - % 8-*$9 $ % . " $ (# !
E ' ' 5, E , L 5 5 , B , L 5 '

---

1 1

3 , ; 9 / % ,) (%# - (# 6 -
$# % W ! ;$// ## / % #) (%# - (# 6 - ($# % <- . % #6 0 34".-
F/4 0 3 G B /:

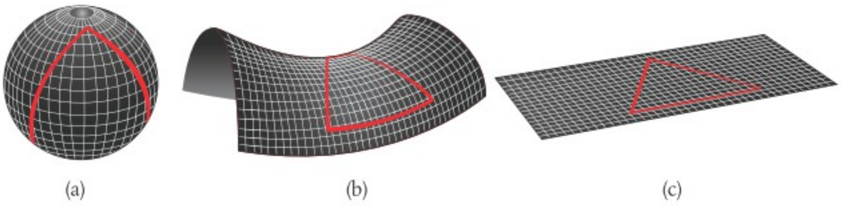

B /: ,) . -- %# ! . * :% #)$# B .!$%% . %# - . % *:# % # 5 % # %T
/ $ #$# %$* - *. K:$# % " .:(. -- %#*6 (: . "$(# ! 4 ,) S(* .T :%)$ $
" # *6 (: . "$(# ! $) 8% % 0$34 ,) S " %T :%)$ $ $..*)$" . % /$# *6
(: . "$(# ! 2 0734 :% 8 #) $ - % *6 #:% . 7$*$%(7 #8 % #) $# - <"$% % $%.
#) $! :%# - !$ % /6 # (%#$ % 8 ** <) 7 # $ -*$# "$(# ! 20(34

073 4 ,) S " %T :%)$ $ $..*)$" . % /$# *6 (: . "$(# ! 2 0734

(!"$ $ (! *

3 , ; 9 !" #$%# ,) $. # $* #6
!"* # +: . # #) $8 - #) % 0 B 34".-
F/4 0 1 3 G B / 4

B /4 4 #$# %$* 6!! # 6 - $ ") 4 ,) %# 6!! # 6 / :"2 0 32 $
. (%% (# . !$% - *.2 (% # %/ - #8 " (4 ,) (!" % %# (%#$ % %/ #)
. %# #6 * ! %# #) 0% !$*3 :7/ :" 0 3 - % % -* (# 6!! # -
#) ") 4 ,) !$ % %/ (!" % %# #) !$% - *. - -* (# 6!! # 4

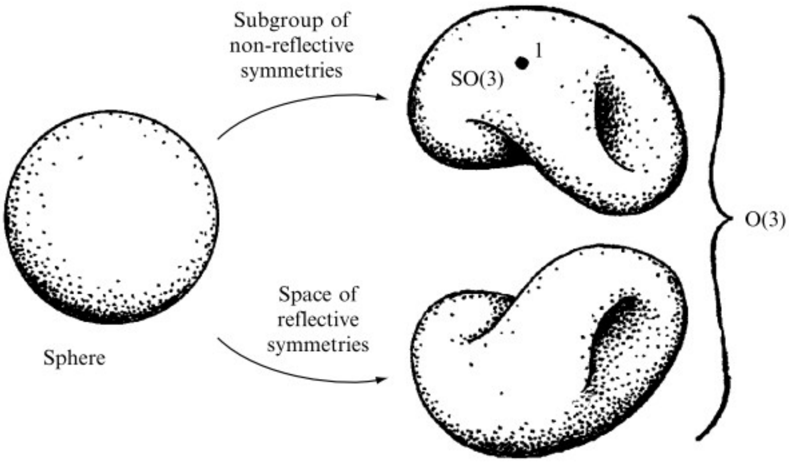
Subgroup of non-reflective symmetries
SO(3)
1
O(3)
Sphere
Space of reflective symmetries

?

? 1

!" #$%#

F/4   >0   3   >4 $ $** * # $% " #

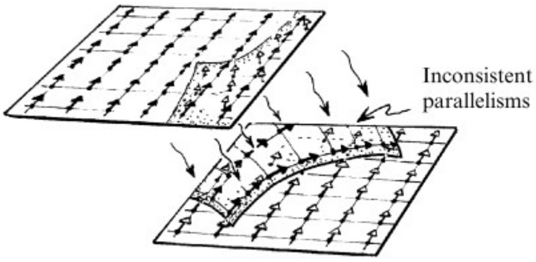
Inconsistent parallelisms

Fig. 14.1 The Euclidean notion of 'parallel' is likely to be inconsistent on the overlap between coordinate patches.

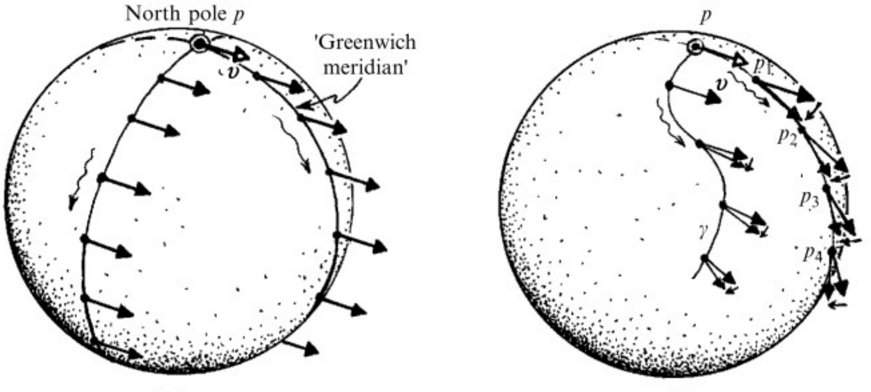
North pole p
'Greenwich meridian'
v
p
v
p_1
p_2
p_3
p_4
γ

(a) (b)

Fig. 14.2 Parallelism on the sphere S^2. Choose p at the north pole, with tangent vector $\boldsymbol{v}$ pointing along the Greenwich meridian. Which tangent vectors, at other points of S^2, are we to regard to being 'parallel' to $\boldsymbol{v}$? (a) The direct Euclidean notion of 'parallel', from the embedding of S^2 in $\mathbb{E}^3$, does not work because (except along the meridian perpendicular to the Greenwich meridian) the parallel $\boldsymbol{v}$s do not remain tangent to S^2. (b) Remedy this, moving $\boldsymbol{v}$ parallel along a given curve γ, by continually projecting back to tangency with the sphere. (Think of γ as made up of large number of tiny segments $p_0p_1, p_1p_2, p_2p_3, \ldots,$ projecting back at each stage. Then take the limit as the segments are made smaller and smaller.) This notion of *parallel transport* is indicated for the Greenwich meridian, but also for a general curve γ.

9 $#
F/4 1 0 ?3

How 'different' are the various groups $O(p, q)$, for $p + q = n$, for fixed n? (The positive-definite and Lorentzian cases are contrasted, for $n = 2$ and $n = 3$, in Fig. 13.18.) They are closely related, all having the same dimension $\frac{1}{2}n(n-1)$; they are what are called *real forms* of one and the same complex group $O(n, \mathbb{C})$, the *complexification* of $O(n)$. This complex group is defined in the same way as $O(n)$ ($= O(n, \mathbb{R})$), but where the linear transformations are allowed to be *complex*. Indeed, although I have phrased my considerations in this chapter in terms of real linear transformations, there is a parallel discussion where 'complex' replaces 'real' throughout. (Thus the coordinates x^a become complex and so do the components of our matrices.) The only essential difference, in what has been said above, arises with the concept of *signature*. There are complex linear coordinate transformations that can convert a -1 in a diagonal realization of g_{ab} into a $+1$ and *vice versa*,[13.52] so we do not now have a

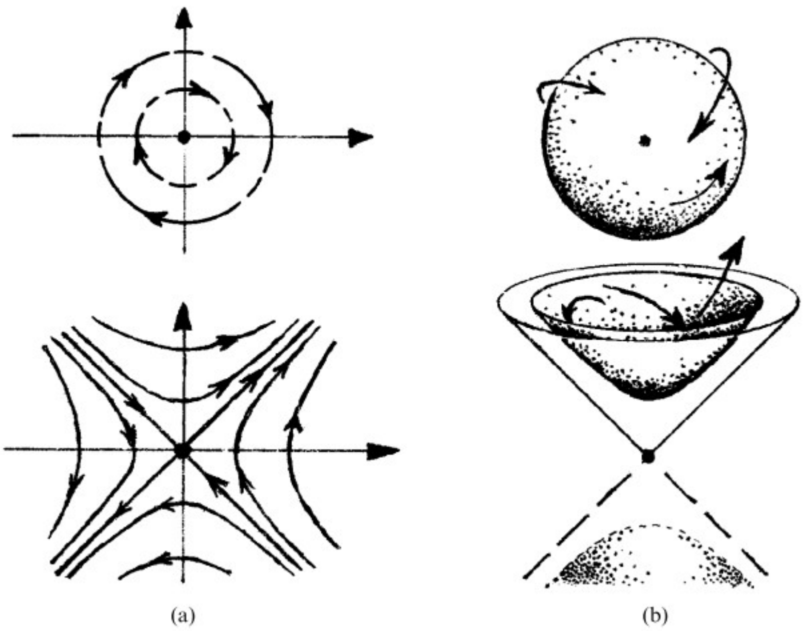

Fig. 13.18 (a) O(2,0) and O(1,1) are contrasted. (b) O(3,0) and O(1,2) are similarly contrasted, the 'unit sphere' being illustrated in each case. For O(1,2) (see §§2.4,5, §18.4), this 'sphere' is a hyperbolic plane (or two copies of such).

meaningful notion of signature. The only invariant[20] of $\boldsymbol{g}$, in the complex case, is what is called its *rank*, which is the number of non-zero terms in its diagonal realization. For a non-singular $\boldsymbol{g}$, the rank has to be maximal, i.e. n.

When is the difference between these various real forms important and when is it not? This can be a delicate question, but physicists are often rather cavalier about the distinctions, even though these can be important. The positive-definite case has the virtue that the group is compact, and much of the mathematics is easier for such situations (see §13.7). Sometimes people blithely carry over results from the compact case to the non-compact cases ($p \neq 0 \neq q$), but this is often not justified. (For example, in the compact case, one need only be concerned with representations that are finite-dimensional, but in the non-compact case additional infinite-dimensional representations arise.) On the other hand, there are other situations in which considerable insights can be obtained by ignoring the distinctions. (We may compare this with Lambert's discovery of the formula, in terms of angles, of the area of a hyperbolic triangle, given in §2.4. He obtained his formula by allowing his sphere to have an imaginary radius. This is similar to a signature change, which amounts to allowing some coordinates to have imaginary values. In §18.4, Fig. 18.9, I shall try to make the case that Lambert's approach to non-Euclidean geometry is perfectly justifiable.)

The different possible real forms of O(n, ℂ) are distinguished by certain set of *inequalities* on the matrix elements (such as det $\boldsymbol{T} > 0$). A feature of *quantum theory* is that such inequalities are often *violated* in physical processes. For example, *imaginary* quantities can, in a sense, have a physically *real* significance in quantum mechanics, so the distinction between different signatures can become blurred. On the other hand, it is my impression that physicists are often somewhat less careful about these matters than they should be. Indeed, this question will have considerable relevance for us in our examination of a number of modern theories (§28.9, §31.11, §32.3). But more of this later. This is the 'can of worms' that I hinted at in §11.2!

F/4 0 3 >4 + . (2 "$ $** * / $! 2 $%. (: $#:

First, in order to build ourselves a parallelogram, let us consider the concept of a *geodesic*, as defined by the connection ∇. Geodesics are important to us for other reasons. They are the analogues of the straight lines of Euclidean geometry. In our example of the sphere S^2, considered above (Figs. 14.2–14.4), the geodesics are great circles on the sphere. More generally, for a curved surface in Euclidean space, the curves of minimum length (as would be taken up by a string stretched taut along the surface) are geodesics. We shall be seeing later (§17.9) that geodesics have a fundamental significance for Einstein's general relativity, representing the paths in spacetime that describe freely falling bodies. How does our

> >

>A ?

, ; 9 ,) 6 - -- %# $* 5K:$# % % 5%/ % %/
$%. ' ()$% (4".-

F/4 1 0> 3 A4 +5 '5, ,5 5,, B 5 5

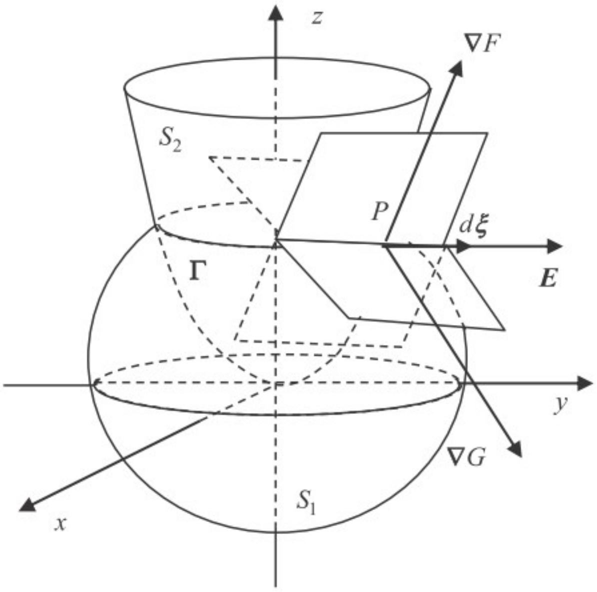

B /: A4> + ! # (%# " #$# % - #) *:# % - # . 5

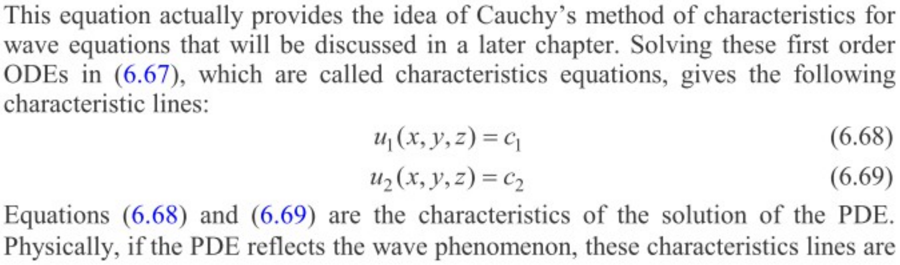

This equation actually provides the idea of Cauchy's method of characteristics for wave equations that will be discussed in a later chapter. Solving these first order ODEs in (6.67), which are called characteristics equations, gives the following characteristic lines:

$$u_1(x,y,z)=c_1 \quad (6.68)$$

$$u_2(x,y,z)=c_2 \quad (6.69)$$

Equations (6.68) and (6.69) are the characteristics of the solution of the PDE. Physically, if the PDE reflects the wave phenomenon, these characteristics lines are

>? ?

>

3)##" 8884(:# #) 9% #4 / "6#)$/ $

3)##" 8884-$(:*#64:!74 .: /$ 6 H$7 * :) *X ? 7) * ")6X -
X '$/ (($%$ "*$# % ! 6#)$/ $ %. <4)#!*4)#!*

7 4 " - G 6#)$/ $ $% ,) !

3 , ; 9 > 6#)$/ $ 6#)$/ $% ,) !
$%. # !$%6 " - 4".-

: (, ; 9 > 6#)$/ $

> A

---

3 , ; 9 $*(:*: (# ($*(:*: $*/ $
I. # 7:# J2 Q4L4 B !$% 0 34".-

3 , ; 9 $*(:*: ,) ; $.#) - 6!"* (# (
$%. % + ! # 6 B # () -# % L % - *$% Q % # % ; 9)=: ; # % 0 34".-

YZ[\]^_`a`bcdedfgedbhiYbihij^kddlmbhiYbihij^ihijjiin^\opqrprm^`stpgmgm[utmgqd[vw[eoxgoupw[
(!"* @$, **$ 0 A34".-
F/4 > 0 3 G :! !"* @

La siguiente discusión presupone cierta familiaridad con las principales propiedades de los números reales. *El sistema de los números reales* fue el resultado de la búsqueda de un sistema (un conjunto abstracto con ciertas reglas) que incluyera a los racionales, pero que también proporcionara soluciones a ecuaciones polinomiales tales como $x^2 - 2 = 0$.

Históricamente, una consideración similar dio origen a la extensión de los números reales. A principios del siglo XVI, Geronimo Cardano consideró ecuaciones cuadráticas (y cúbicas) tales como $x^2 + 2x + 2 = 0$, que no son satisfechas por ningún número real x. La fórmula cuadrática $(-b \pm \sqrt{b^2 - 4ac})/2a$ da expresiones "formales" para las dos soluciones de la ecuación $ax^2 + bx + c = 0$. Pero esta fórmula puede requerir raíces cuadradas de números negativos, por ejemplo, $-1 \pm \sqrt{-1}$ para la ecuación $x^2 + 2x + 2 = 0$. Cardano notó que si estos "números complejos" son tratados como números ordinarios con la regla $\sqrt{-1} \cdot \sqrt{-1} = -1$, estos, resolvían en efecto las ecuaciones. A la importante expresión $\sqrt{-1}$ se le da ahora la ampliamente aceptada designación de $i = \sqrt{-1}$. Una convención alternativa es seguida por muchos ingenieros eléctricos, quienes prefieren el símbolo $j = \sqrt{-1}$, puesto que ellos desean reservar el símbolo i para la corriente eléctrica. Sin embargo, en el pasado se sentía que ningún significado podría realmente ser asignado a tales expresiones, que fueron entonces llamadas "imaginarias". Gradualmente, en especial como resultado del trabajo de Leonhard Euler en el siglo XVIII, estas cantidades imaginarias llegaron a desempeñar un papel importante. Por ejemplo, la fórmula de Euler $e^{i\theta} = \cos\theta + i \operatorname{sen}\theta$ reveló la existencia de una profunda relación entre los números complejos y las funciones trigonométricas. Se encontró que la regla $e^{i(\theta_1 + \theta_2)} = e^{i\theta_1} e^{i\theta_2}$ resumía las reglas para la expansión del seno y del coseno de la suma de dos ángulos de una manera sencilla, y este solo resultado indicó que algún significado debería ser atribuido a estos números "imaginarios".

Sin embargo, no fue hasta los trabajos de Casper Wessel (*circa* 1797), Jean Robert Argand (1806), Karl Friedrich Gauss (1831), Sir William R. Hamilton (1837) y otros, que se clarificó el significado de los números complejos y se comprendió que no hay nada de "imaginario" en ellos (aun cuando este término se emplea todavía.

El análisis complejo, que es el tema de este libro, fue desarrollado en el siglo XIX, principalmente por Augustin Cauchy (1789-1857). Posteriormente, su teoría se hizo más rigurosa y fue extendida por matemáticos como Peter Dirichlet (1805-1859), Karl Weierstrass (1815-1897) y Georg Friedrich Bernhard Riemann (1826-1866).

La búsqueda de un método para describir la conducción del calor influenció el desarrollo de la teoría, la cual encontró muchos usos fuera de las matemáticas. En los capítulos subsecuentes se discutirán algunas de estas aplicaciones a problemas en física e ingeniería, tales como hidrodinámica y electrostática. La teoría tiene también aplicaciones matemáticas en problemas que a primera vista no parecen involucrar números complejos. Por ejemplo, la demostración de que

$$\int_0^\infty \frac{\operatorname{sen}^2 x}{x^2}\,dx = \frac{\pi}{2}$$

o que

$$\int_0^\infty \frac{x^{\alpha-1}}{1+x}\,dx = \frac{\pi}{\operatorname{sen}(\alpha\pi)} \qquad \text{para } 0<\alpha<1$$

o que

$$\int_0^{2\pi} \frac{d\theta}{a+\operatorname{sen}\theta} = \frac{2\pi}{\sqrt{a^2-1}}$$

puede ser difícil o imposible si se usa el cálculo elemental, pero estas identidades pueden ser probadas rápidamente al usar las técnicas de variable compleja.

El análisis complejo se ha vuelto una herramienta usual e indispensable en el trabajo de matemáticos, físicos e ingenieros. El descuido de éste puede ser un serio impedimento en la mayoría de las áreas de investigación y aplicación que involucran ideas y técnicas matemáticas.

>3 , ; 9 L %7 (9 * L %7 (9 *
%# .:(# % # $*(:*: 4 0 34".-

3 , ; 9 L %7 (9 * L %7 (9 *
%# .:(# % # $*(:*: 4 0 34".-

A3 , ; 9 L %7 (9 * W4 L4 L %7 (9 *
%# .:(# % # # % ($*(:*: $%. (%# %::! ! ()$% (, $-- . :7*) %/ 0 34".-

>?

A ?

3)##" 8884" - (* 4(! # " * /6 :% #4)#!*

3 , ; 9 A B:%.$! %#$* - :! ($*
!":#$# % , 7 % 4 (**2 ()$. W4 ; $:% B:%.$! %#$* - :! ($* !":#$# %
' (#6 - %.: # $* N ""* . '$#) !$# (0 134".-

3 , ; 9 A B:%.$! %#$* - :! ($*
!":#$# % :! ($* !":#$# % - %# %$* $%. 5<# %$* B* 8 2 *:! B:%.$! %#$* -
!":#$# %$* B*: . 6%$! (;:## 8 #) L % !$%% 0 134".-

>3 , ; 9 A B:%.$! %#$* - :! ($*
!":#$# % :! ($* !":#$# % - %# %$* $%. 5<# %$* B* 8 2 *:! !":#$# %$*

! #) . - % (. $%. (: -* 8 W)% Q * 6 N %) () # 2 y 0 >34".-

3 , ; 9 A (%# - (! ($% W$%:$ 6
(:/ !H $H4".-
7 4 (%# - ! ($% G %# 96#) $

A3 , ; 9 A L$%.7 9 - $*/ 7 $ (# " * /6 4'4
W$! L # 6 - , " * /6 #) L **$%. 0 A34".-

13 , ; 9 A #:. % H:
, " * / ': ** ' >4".-
7 4 " %# - *$# % $*/ 7 $ ($"" $() # / % $* # " * /6

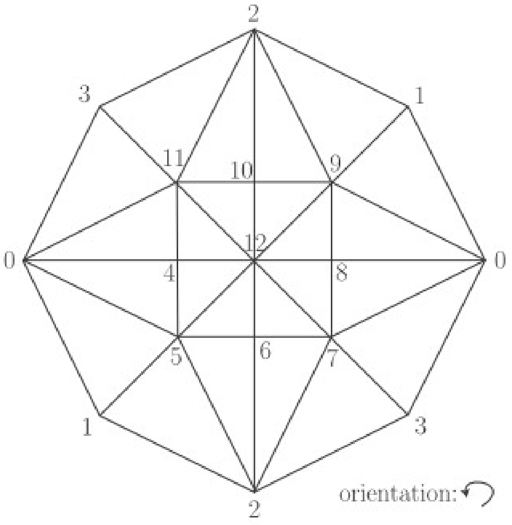
2
3
1
11
10
9
12
0
4
8
0
5
6
7
1
3
2
orientation:

Triangulation of the projective plane

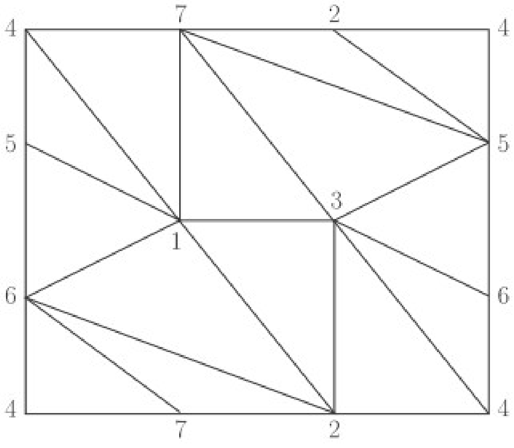
4
7
2
4
5
5
3
1
6
6
4
7
2
4

Triangulation of the Csaszar polynomial

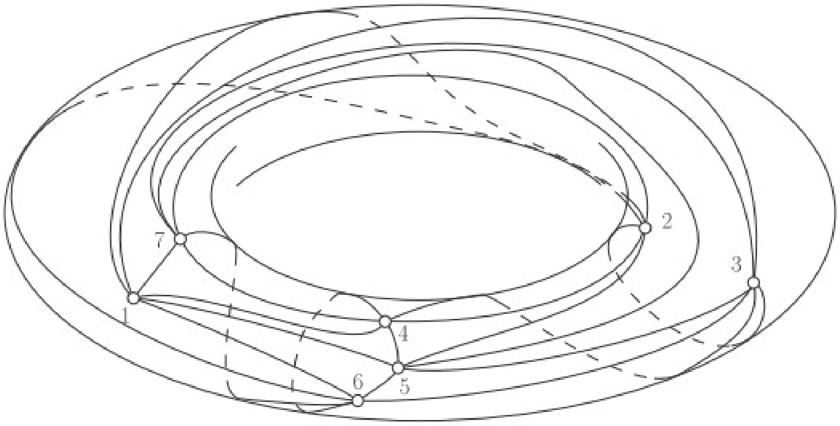

Triangulation of the Csaszar torus

Model problem

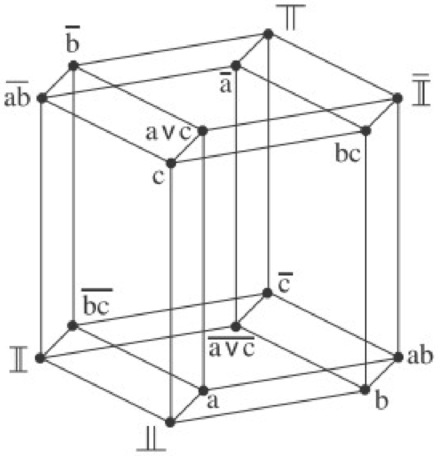
$\mathbb{T}$
$\bar{\mathrm{b}}$
$\overline{\mathrm{ab}}$
$\bar{\mathrm{a}}$
$\mathbb{I}$ with overline
a∨c
bc
c
$\bar{\mathrm{c}}$
$\overline{\mathrm{bc}}$
$\overline{\mathrm{a\vee c}}$
$\mathbb{I}$
ab
a
b
$\perp\!\!\perp$

$\mathbb{I}^\top = \mathbb{I}$
$a^\top = c$ $b^2 = \bar{b} = \mathbb{I} \cup a \cup c$
$b^\top = b$ $a{;}c = c{;}a = \mathbb{T}$
$c^\top = a$ $a{;}b = b{;}a = a \cup b$
$a^2 = a$ $c{;}b = b{;}c = c \cup b$
$c^2 = c$

Ralph McKenzie's homogeneous non-representable RA

The element a cannot be conceived as a Boolean matrix.

F/4 ?

A ?

1 A

3 , ; 9 1 '$#) !$# (B)6 (%
**: # $# . L$%.7 9 '$#) !$# (B)6 (% **: # $# . L$%.7 9 % . !4".-
F/4 01 3 G)$"# > G , " * / ($* "$(

Table 4.1.1 Generalizations of surfaces.

	Added structure	Resulting features
Topological space	Open sets	Connectedness, holes
Hausdorff space	Disjoint neighborhoods	Separation between points
Metric space	Metric	Distances between points
Topological manifold	Cartesian charts	Coordinates, dimension
Differentiable manifold	Differentiable structure	Tangent vectors, calculus
Riemannian manifold	Riemannian metric	Length, angles, volumes
Euclidean surface	Embedding in $\mathbb{R}^3$	Relation to a higher space

Notes: There are many more messy details here than with algebra. These will be covered in the following chapters.

F/4 0 1 3

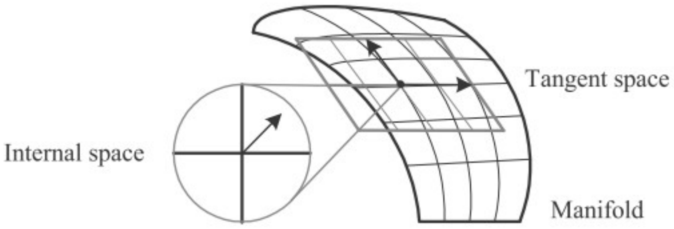

Figure 4.1.2 Generalizing tangents on a manifold.

F/4 01 3

Topological space:
- Structure: Open sets
- Properties: Number of holes, connectedness
- Equivalency: Homeomorphism

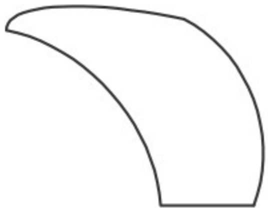

Topological manifold:
- Structure: Local cartesian charts
- Properties: Coordinates, dimension
- Equivalency: Homeomorphism

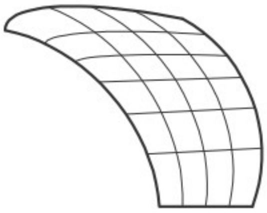

Differentiable manifold:
- Structure: Differentiable structure
- Properties: Tangent vectors, calculus
- Equivalency: Diffeomorphism

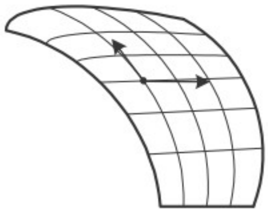

Riemannian manifold:
- Structure: Metric
- Properties: Lengths, angles, areas, volumes
- Equivalency: Isometry

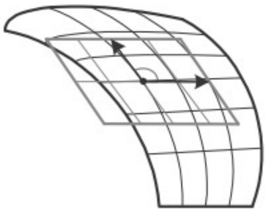

F/4 A 0 ? 3 G)$"# G !$%% $% !$% - *.

9.1.2 *The parallel transporter*

By definition, for a vector w at a point p of an n-dimensional manifold M, **parallel transport** assigns a vector $\|_C(w)$ at another point q that is dependent upon a specific path C in M from p to q.

To see that this dependence upon the path matches our intuition, we can consider a vector transported in what we might consider to be a "parallel" fashion along the edges of an eighth of a sphere. In this example, the sphere is embedded in $\mathbb{R}^3$ and the concept of "parallel" corresponds to incremental vectors along the path having a projection onto the original tangent plane that is parallel to the original vector.

F/4 A 0 ? 3

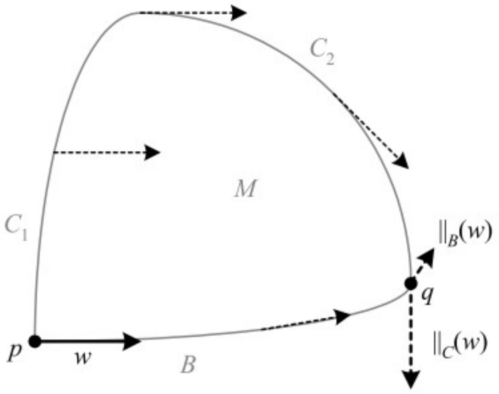
C_2
M
C_1
$\|_B(w)$
q
p
w
B
$\|_C(w)$

Figure 9.1.1 A vector w transported in what we intuitively consider to be a "parallel" way along two different paths (B and $C = C_1 + C_2$) on a surface results in two different vectors.

F/4 A10 ?A3

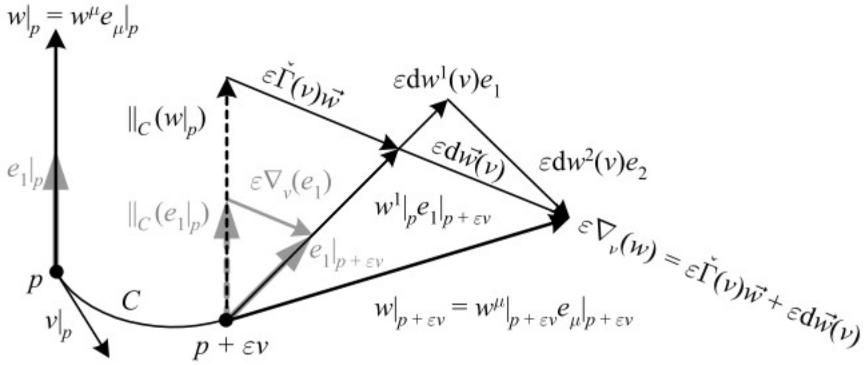
$w|_p = w^\mu e_\mu|_p$
$\varepsilon\check{\Gamma}(v)\vec{w}$
$\varepsilon \mathrm{d}w^1(v)e_1$
$\|_C(w|_p)$
$\varepsilon \mathrm{d}\vec{w}(v)$
$\varepsilon \mathrm{d}w^2(v)e_2$
$e_1|_p$
$\varepsilon\nabla_v(e_1)$
$\|_C(e_1|_p)$
$w^1|_p e_1|_{p+\varepsilon v}$
$e_1|_{p+\varepsilon v}$
$\varepsilon\nabla_v(w) = \varepsilon\check{\Gamma}(v)\vec{w} + \varepsilon \mathrm{d}\vec{w}(v)$
p
C
$v|_p$
$w|_{p+\varepsilon v} = w^\mu|_{p+\varepsilon v} e_\mu|_{p+\varepsilon v}$
$p + \varepsilon v$

Figure 9.1.3 Relationships between the frame, parallel transport, covariant derivative, and connection for a vector w parallel to e_1 at a point p.

F/4 1 0 3

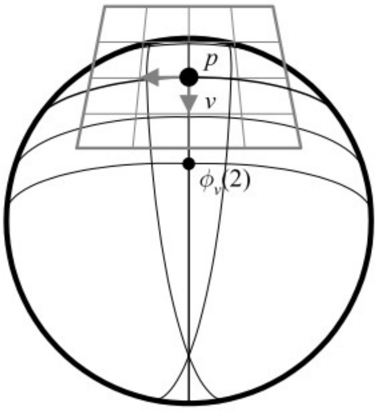

Figure 9.1.4 Geodesic normal coordinates at p map points on a manifold to vectors at p tangent to the geodesic passing through both points. In the figure $\exp(2v) = \phi_v(2)$, so the coordinate of the point $\phi_v(2) \in M$ is $2v \in T_pM$.

3 , ; 9 1 - $%. . $ *:. #
. $%(. '$#) !$# (- $%. . $ *:. # . $%(. '$#) !$# (4".-
F/4 10 1 3

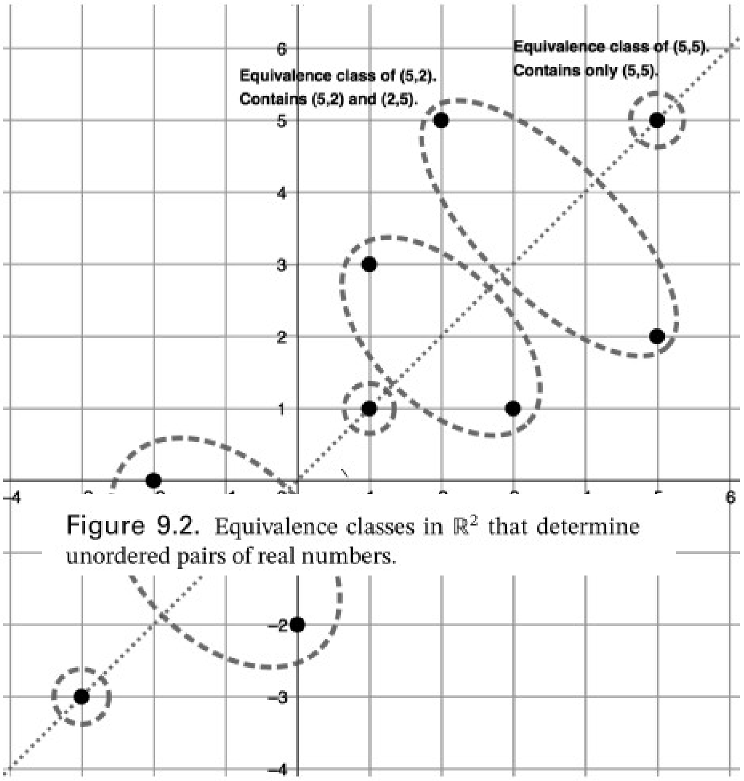
Equivalence class of (5,2).
Contains (5,2) and (2,5).
Equivalence class of (5,5).
Contains only (5,5).
6
5
4
3
2
1
−4
6
−2
−3
−4

Figure 9.2. Equivalence classes in $\mathbb{R}^2$ that determine unordered pairs of real numbers.

F/4 A 0 1 3

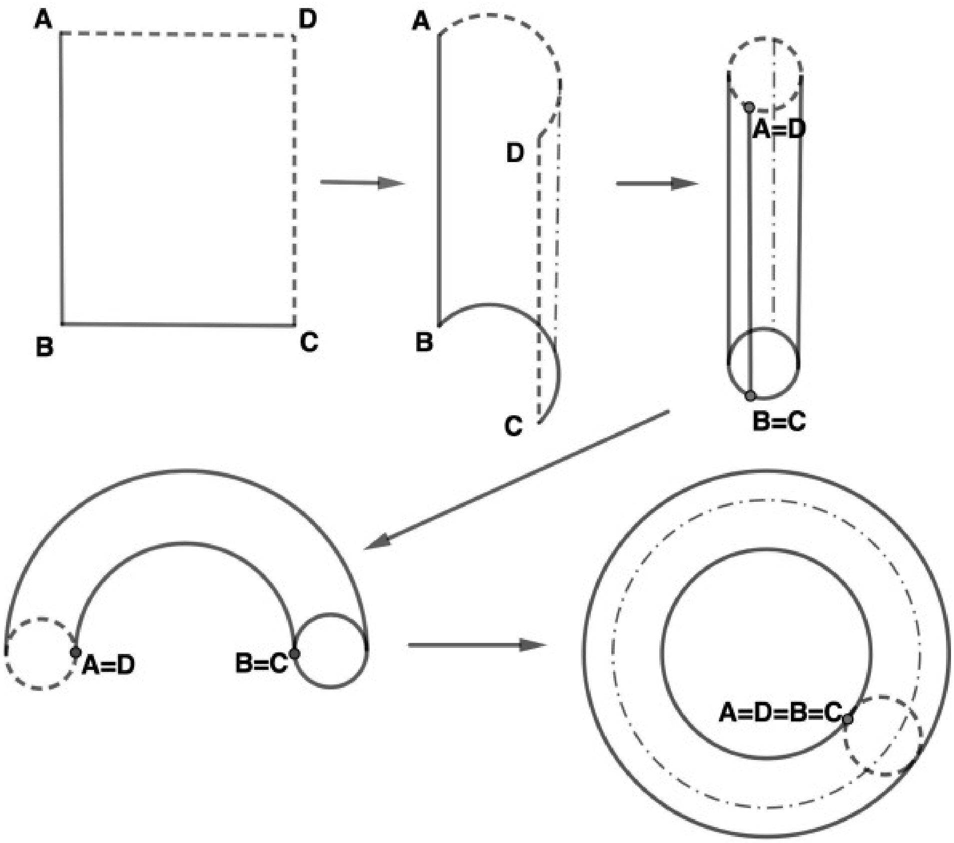

Figure 9.4. The set A/R in Example 9.26: By wrapping the right and left edges of the rectangle together we get a cylinder, and then wrapping the top edge and bottom edge of the cylinder we get a torus.

3 , ; 9 1 / ($* ' #) . ,) # -
,) %9 %/ 7 # $(#*6 $%. '$#) !$# ($**6 / ($* ' #) . 4".-

>3 , ; 9 1 + :" 2 + ! # (# :(#:
$%. -- %# $* 5K:$# % + :" 2 + ! # (# :(#: $%. -- %# $* 5K:$# % E $
-# "): '$#) !$# ($* (4".-

1

---

1 >

3 , ; 9 1 ""* ($# % . % :$%#:! %.
#$# # ($*)6 (*:! ""* ($# %
. % :$%#:! %. #$# # ($*)6 ( ) # : B B:#: (%# # %. 5%/ %
*:! 4".-
F/4 A0> 3

Extending his idea, in 1917, Einstein postulated that if the photon had to be considered a particle,[9] it ought to have a momentum $\boldsymbol{p}$ so that its kinetic energy would be $\varepsilon = pc$. But Planck's black-body theory and the interpretation of the photoelectric effect required an $\varepsilon = h\nu$ energy for the photon associated with an electromagnetic wave with frequency $\nu = c/\lambda = ck/(2\pi)$, where k is the wave vector. It is therefore clear that some kind of relationship must exist between the photon momentum $\boldsymbol{p}$ and the electromagnetic wave vector $\boldsymbol{k}$:

$$\boldsymbol{p} = \hbar\boldsymbol{k} \tag{1.11}$$

3 , ; 9 1 ""* ($# % . % :$%#:! %.
#$# # ($*)6 (*:! ""* ($# %
. % :$%#:! %. #$# # ($*)6 ( ) # : B B:#: (%# # %. 5%/ %
*:! 5K: * 7 :!4".-
F/4 0 >3 $ # ' . * L$! *# % $% $%. "" < !$# %
7 $# %/ 6 # !

,) K:$%#:! 7)$: - $ %/* "$ # (* % $ " (8 " # %# $*2 $
. (7 . % *:! 2 $!"* 7:# % # :(# <$!"* -) 8 $ $*
#:$# % ($% 7 ! . ** . $%. * . <$(#*64 L 8 #) %:!7 - $*
* - " 7* ! 8) () ($% 7 !"* z . # :() $% <# ! #6 $ "$ 4
#) #6" - #:$# % #) - % . # 7 <"* . $%.2 - #:%$# *62 #)
K:$. $# (" # %# $* -# % %(:%# . $%. <$(#*6 * $7* 4 ,) K:$%#:!
#:.6 -)$! % ((**$# ")6 (8 ** 7 #) ()$%(# #) #6"
- " # %# $* 8) ()2 :%. (#$ % (%. # % 2 ($% * / # !$# *6 7 * ! # .
#) K:$. $# ((!" % %#4 B: #) ! 2 % $ 7 $. " " (# 2
8 ** $* / #) "" #:% #6 # . ($ " 8 -:*2 $%. K: # / % $*2
$"" < !$# % ! #) . - * %/ $!:() *$ / (*$ - " 7* ! 4

3 , ; 9 1 ,) %7 :%. ! #) :
, ()% * / ($*)$%/ $%. %.: # $* * "! %# % Q # %
5: " ,) 4 %7 :%.4 ! #) : 4".-

1 >A

---

3 , ; 9 '$#) !$# 9 $":*$ 4 '$#) !$# 9
-: %/ % : :%. $#: 8 % ()$-#* 4 8 / 0 34".-
F/4 0 >13
-- %# $* ()%:%/
-- %H 7$ 9 # % B:%9# %
4 $,$%/ %# %" 7* !

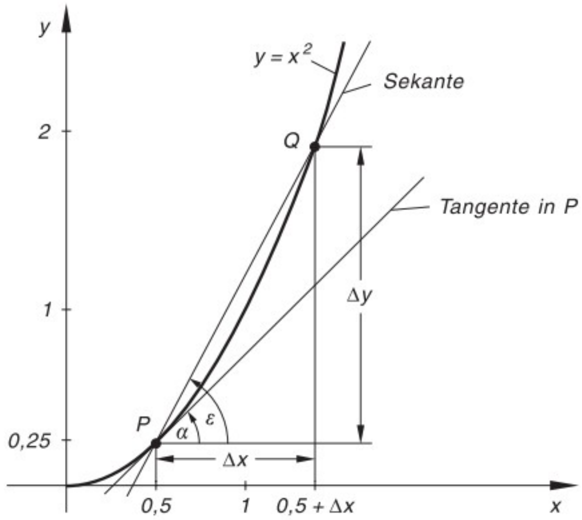
y
$y = x^2$
Sekante
2
Q
Tangente in P
1
Δy
P
ε
α
0,25
Δx
0,5
1
0,5 + Δx
x

Bild IV-1

.: () :%.   *$:- %.   9$%# 7   #H# .$%% .   # /:%/

$$m_s = \tan \varepsilon = \frac{\Delta y}{\Delta x} = \frac{(0{,}5 + \Delta x)^2 - 0{,}25}{\Delta x} = \frac{0{,}25 + \Delta x + (\Delta x)^2 - 0\,25}{\Delta x} =$$

$$= \frac{\Delta x + (\Delta x)^2}{\Delta x} = \frac{\Delta x\,(1 + \Delta x)}{\Delta x} = 1 + \Delta x \qquad \text{(IV-2)}$$

F/4   >0 >?3

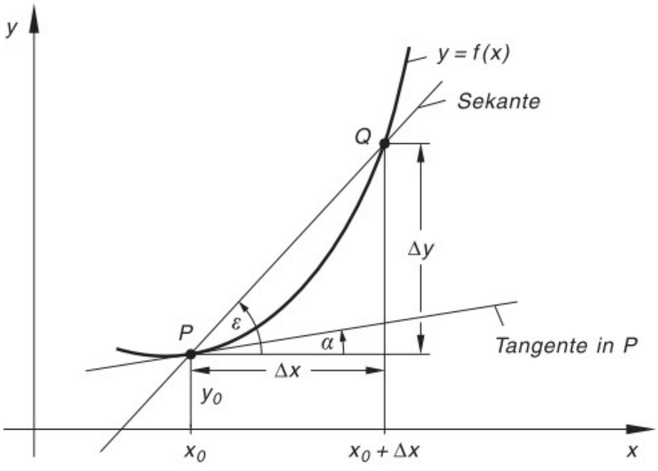
y
$y = f(x)$
Sekante
Q
Δy
P
ε
α
Tangente in P
Δx
y_0
x_0
$x_0 + \Delta x$
x

Bild IV-2
Zum Begriff
der Ableitung
einer Funktion

3   ,   ; 9   A   @$)   !$#) !$# 9 A   @$)

F/4 110 ??3   ? '$#) !$# 9 8=) %. .   Q   % ()$-#* () %   *:# %

F/4 > 0> >3

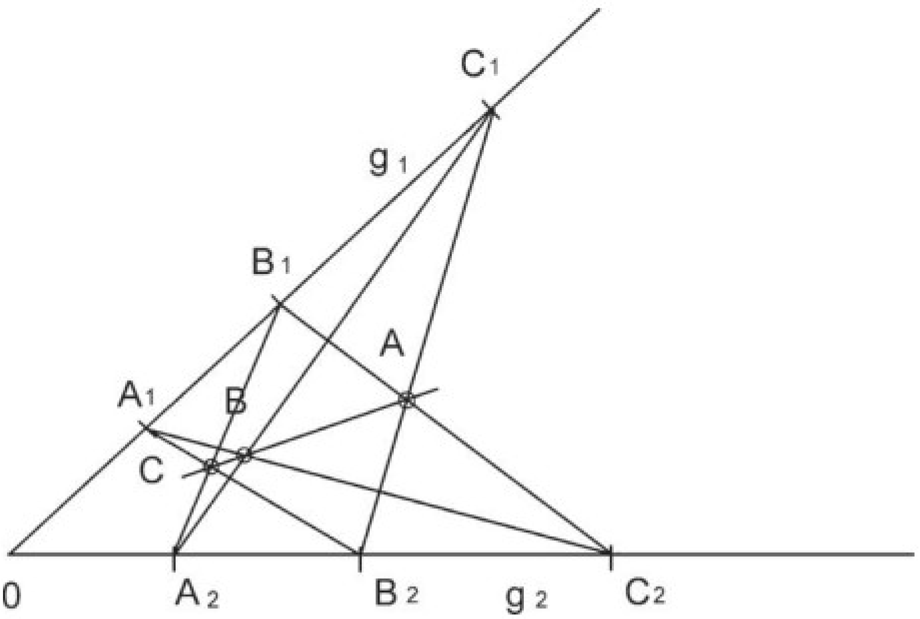

774 ?4>4 B /: H:! $#H % $"" / % . 5(9 % % () (9 $7
8 () *%. $:- H8 + $. %2 ()% . % () . + / % # % % :%9# % %
+ $. %

?4> %-=%/ . " @ 9# % + ! #
' # . ! %$() $"" 7 %$%%# % $#H)$## . %# 9 % % $#H) /
7 $()#2 . % 8) :# . " @ 9# % + ! # H: ()% % :%. / * / %#* ()
$:() $* $#H % $"" $ ($* 7 H ()% %4
5 % 9 %# %: * ()) # () 5%#8 (9*:%/) % H: % H: $!! %)=%
/ %. % H "* % {" @ 9# + ! # | 7 /$%% 9 !)$-# 8=) %. .
%$ $%(%%)$*7 . " " 9# () % ' #) . % . y}% #* 4 8$ -
% ;$## #$ *7 # 0 > >G >1 3 % % ! # " #:! % } %7 / /
. :(9# % , $##$# . **$ " ##: $. B $/ %$() . % / ! % $! % / !
() % 5 / % ()$-# % $:-2 . H8 % " 9# H: %$%. #) %.
B /: %)$7 % 0$: -}) * () % I (7$ () 7 2 4 :~4 2 4 >A 4J2
I $)$% $*! . (>2 4 4J2 I * ?1J34
/ %#* () B #-}) :%/ . . % .$%9# !$% . ! - $%H• () %
%/ % : :%. () # 9# % G $7 ! '$#) !$# () % :# . .$9# G + $.
$ /: 0 G AA 34 5 7$%. % # . y / * ()% ##*) ! # . %
/ *% . " $9# H %. % y}% #* :%. :()# . % , ()% 9 % .: () !$
#) !$# () : (). %/:%/ H: 7 %4 % L$:"#8 9 {; : ** % " @ (#
4 4 4 | 0- }7 #H# 5 # 5%#8: - . ; () 7:%/ . 5 /% 7 ! :
$!! %# % % y / * ! # % 57 % 3 () % A 2 $7 %: % %
:~$/ % #}(94 $ /: # *# . 5< !"*$ $% '$#) !$# 9 U
. %$%H 8$ G :4 $4 8 / % . :%/ 8•)%* () % , ! % * / G / %/4
$ Q 9 8$ *$%/ () ** %U # 8: . % / %$* < !"*$ 8 .
%#. (9#4
:- $ /: /) % . 9# . %. 9# . ; / 7 *.:%/ % {:%
%.* () - % :%9#|2 {}$! % () , *:%/| :%. { "" *)=*#% | H:
}(9U . * #H# ; / 8: . # ! 4 W$)):%. # .: () 4B4 '•7
: % ! . % B !:* :%/ % . '$#) !$# 9 %/ -}) #4 :- $
/: /)# $:() . ,) . *$ % % :%9# H: }(9 0 /*4 $: -})
* () I) !7 9 #() >J2 I $)$% $*! . (>2 4 4J2
IB *. + $6 ?1J34

3 , ; 9 $"*$(, $% - !$# % # -- %
+ 77 * 2 # -$% ## '$#) !$# 9 #) % :%. $%8 %. % % . % + :%.*$/ % 7 H:
B :) % :%. $"*$(, $% - !$# % " 9# :! 9$. ! () *$/ 0 34".-

F/4 0> 3

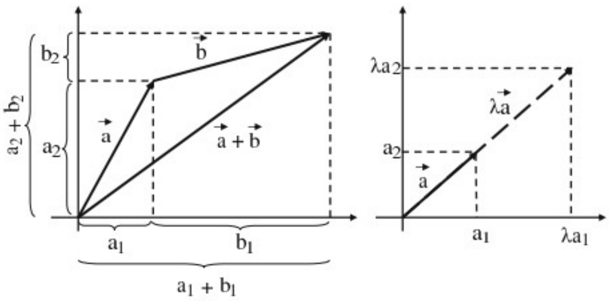

b_2
$\vec{b}$
$a_2 + b_2$
$\vec{a}$
$\vec{a} + \vec{b}$
a_2
a_1
b_1
$a_1 + b_1$
λa_2
$\lambda\vec{a}$
a_2
$\vec{a}$
a_1
λa_1

774 4A :!! :%. 9$*$ *-$() % 9# % % y . %$# % () 78 4

F/4 ?0> 3 4 4 4> #) / %$* @ 9# % :%. #

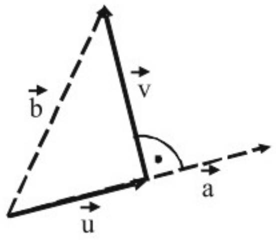

$\vec{b}$
$\vec{v}$
$\vec{a}$
$\vec{u}$

Abb. 3.14: Zerlegung des Vektors $\vec{b}$ in Projektion $\vec{u}$ und Lot $\vec{v}$ auf den Vektor $\vec{a}$

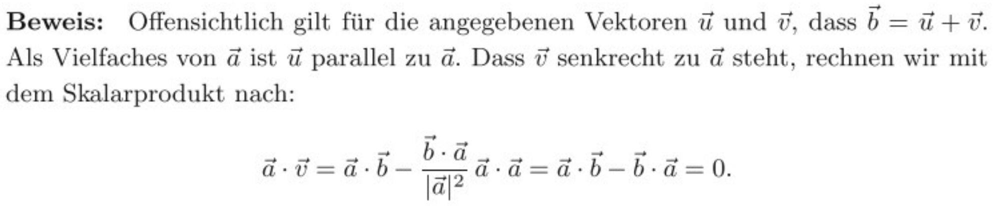

Beweis: Offensichtlich gilt für die angegebenen Vektoren $\vec{u}$ und $\vec{v}$, dass $\vec{b} = \vec{u} + \vec{v}$. Als Vielfaches von $\vec{a}$ ist $\vec{u}$ parallel zu $\vec{a}$. Dass $\vec{v}$ senkrecht zu $\vec{a}$ steht, rechnen wir mit dem Skalarprodukt nach:

$$\vec{a} \cdot \vec{v} = \vec{a} \cdot \vec{b} - \frac{\vec{b} \cdot \vec{a}}{|\vec{a}|^2}\, \vec{a} \cdot \vec{a} = \vec{a} \cdot \vec{b} - \vec{b} \cdot \vec{a} = 0.$$

>3 , ; 9 +$ #) O (# %$ 6 - (%# - (
: #$# % '$#) !$# ($**6 " $9 %/ (# %$ 6 - : #$# % % # #:# -)6 (:7*) %/
0 ?34".-

F/4 >0>13 G (*

3 , ; 9 '$#) !$# ($%.
()$ * /6 '$#) !$# ($%. ()$ * /6 G G 0 34".-

F/4 A 0 ?>3

% 5:(* . $% #$%(% %# $## % %$*6
% # " (' $: - #) #:.6 - ##* ! %# #8 9 % L # / % : / %
W $% / € H

?

3 ; !" #$%#& !) !"* < $ $7* Q #)L6" * %9 4".-
F/4 0 13 4 B: #) , " * /6 - #) *$%

3 ; !" #$%#& ! !" #$%#& ! +$: !" #$%# % 8* (#: + 4".-
7 4 # . PPP

3 , ; 9 !" #$%#& ! #K/4".-
7 4
: (*:! / %$* > #K/4".-

>3 , ; 9 !" #$%#& ! H: ()7*$:4".-
7 4
: (*:! / %$* > H: ()7*$:4".-

3 ; 9 #) + ! # $ %$* # ($ / ! # $ $%$* # ($:-!$4".-

A3 ; !" #$%# + ! # $ ' 8% 4".-
F/4 0 A3

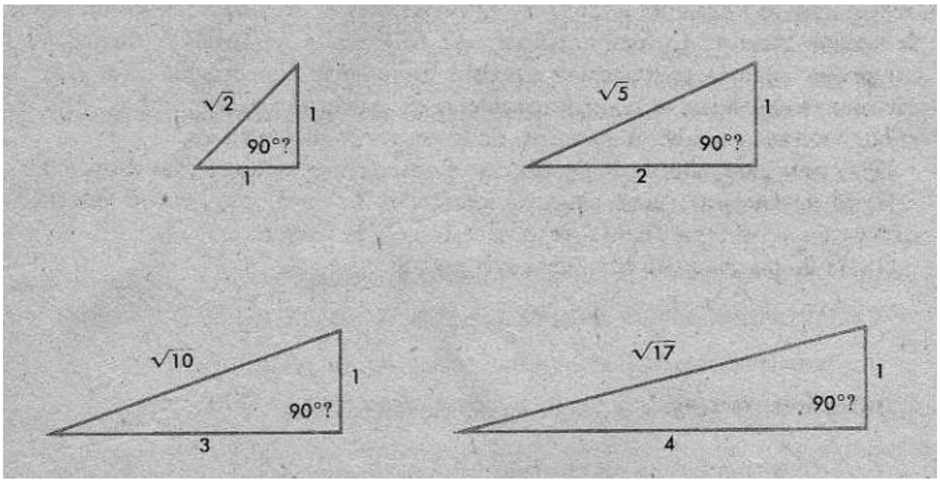
√2
1
90°?
1
√5
1
90°?
2
√10
1
90°?
3
√17
1
90°?
4

13 , ; 9 6!! # 6 5<" %/ #) L .. %
$## % - :!7 AA?14".-
7 4 6!! # 6 G '$#) !$# ($* 5<"* $# %

?3 , ; 9 6!! # 6 5<" %/ #) L .. %
$## % - :!7 B $ * 6!! # 64".-

3 , ; 9 6!! # 6 5<" %/ #) L .. %
$## % - :!7 :!7 2)$" 2 N 6!! # 6 % %# .:(# % # :!7 ,) 62 + ! # 62
$%. + :" ,) 6 y # 0 34".-

3 , ; 9 6!! # 6 5<" %/ #) L .. %
$## % - :!7 :!7 2)$" 2 $%. 6!! # 6 0 34".-

3 , ; 9 6!! # 6 5<" %/ #) L .. %
$## % - :!7 6!! # 62)$" $%. "$(4 % %# .:(# % # '$#) !$# (#) :/)
+ ! # 6 y 6 ** / 0 A34".-
7 4 * # . PPP

3 , ; 9 6!! # 6 5<" %/ #) L .. %
$## % - :!7 ()* # ; $% $%. ' 6!! # 64".-

3 , ; 9 6!! # 6 5<" %/ #) L .. %
$## % - :!7 6!! # 6 % "# ($%. % #:. 4".-

>3 , ; 9 6!! # 6 5<" %/ #) L .. %
$## % - :!7 : # - 6!! # 62 * (# . Q 9 - ;:%@ $9 #$4".-

3 , ; 9 6!! # 6 5<" %/ #) L .. %
$## % - :!7 6!! # 6 .$"# . ;$ # :# !$# (+ % $# % - 7* ! %
) ! # 6 $%.)6 (Q *. (%# - (:7*) %/ !"$%6 0 34".-

A3 ; !" #$%# $*(:*: %$*6 5:(* . $% "$(4".-

13 ; ,))$" - "$(? >1 1 4".-
7 4 ,))$" - "$(
F/4 A0 3

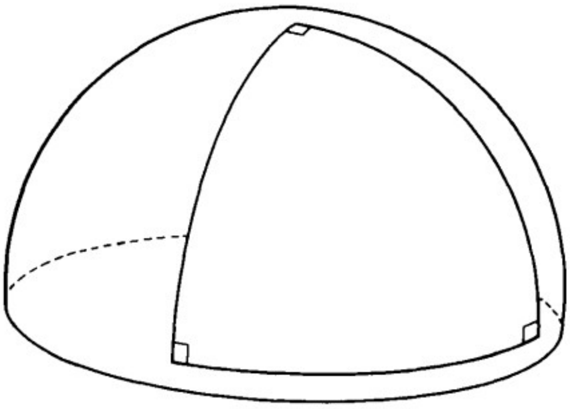

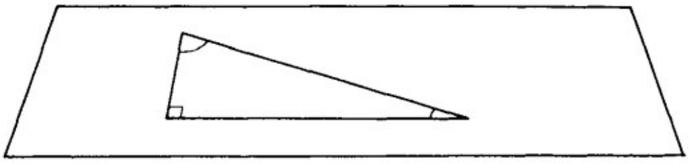

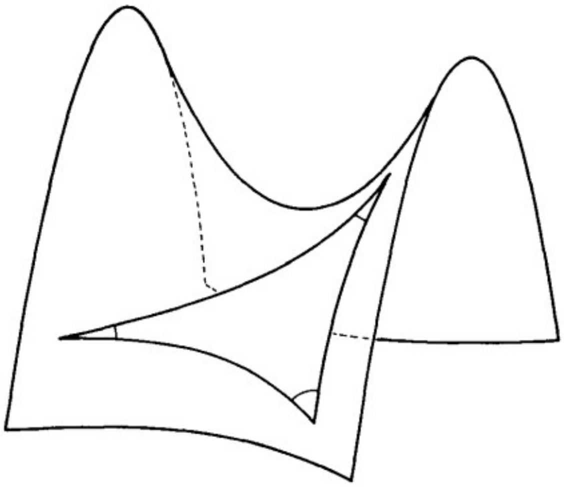

B /: 41 ,)) ! ") 2 #) "*$% 2 $%. #) $..* : -$(
$**)$. -- %# %# % (/ ! # 4

?3 ; !" #$%# / (!"* # 4".-
7 4 + ! # $ 5 "$($* (! + + 7 $4

3 ; !" #$%# / ! # $! . %$! 4".-

3 ; !" #$%# $# ($ '4 ()8$ H2 W)% L4 ()8$ H " ($* *$# #6
B ! 5 % # % # # %/ 0 >2 $!7 ./ % #6 34".-

3 ; !" #$%# #) / %$* * % : * ("$" : -#4".-
7 4 5• 5 ,5 , +5 , 5 5
+, ',, 5 5

3 ; !" #$%# $# ($ '4 ()8$ H2 W)% L4 ()8$ H " ($* *$# #6
B ! 5 % # % # # %/ 0 >2 $!7 ./ % #6 34".-
7 4 5• 5 ,5 , +5 , 5 5
+, ',, 5 O5 5

3 ; !" #$%# $. +: ()$. %/* $%. ':*# $ $7* $*(:*: 4".-

>3 ; !" #$%# (: > A>?4".-
7 4 5 5'5 , - E, +5 '5, E B ,Q '5
F/4 A10? 3

; 9
, +L, 5 5
L ,5
, +L, 54

% # :(# % - #) 5K:$# % - #) B # / 4

3 ; !" #$%# $7%1 4 4 4:! ()4 .:4".-
7 4 ƒ % - +€ !€# %$*6# K: G ; #

A3 ; „ ' „„> ;5 „„ #): 4 ' ** *7 # 5 % # %T " ($* ,) 6
- *$# #6 5! / %(0 3 $%. 5$ *6 %# " #$# % 0 3 0 12 " %/ 34".-
7 4
; !" #$%# >

; !" #$%# > *7 # 5 % # %T " ($* ,) 6 - *$# #6 5! / %(
0 3 $%. 5$ *6 %# " #$# % 0 3

13 ; '$#) !" #$%# H % . / ! # $. -- %H $* # ! 4".-

?3 ; '$#) !" #$%# H % . / ! # $. -- %H $* # ! 4".-

3 ; '$#) !" #$%# H % . / ! # $. -- %H $* # ! 4".-

7 4 : / ; $%() G 12 ? 4

3 ; '$#) !" #$%# $ #$O 0/ ! # $34".-

3 ; '$#) !" #$%# !" #$%# : / ; $%() H % . / ! # $
. -- %H $* 4 *:! " 0 34".-

3 ; '$#) !" #$%# !" #$%# : / ; $%() H % :**$ # $. **
-:%H % . $ $7 * (!"* $. ** -:%H % ** ## () 54 " 0 A34".-

3 !" #$%#
7 4 $ *:!
; '$#) !" #$%# !" #$%# : / ; $%() H % . / ! # $. -- %H $* 4
*:! " 0 34".-
(! *:! .$ "$ #$
; '$#) !" #$%# ; 9 ? *:!
7 4
, . $ K: :% . %$ "$ #$
, ; 9 !" #$%#& !

>3 ; '$#) !" #$%# + ! # $ %$* # ($ H % . + ! # $
%$* # ($ H % . / ! # $. -- %H $* # ! 4".-

3 ; '$#) !" #$%# + ! # $ %$* # ($ H % . + ! # $
%$* # ($ H % . / ! # $. -- %H $* # ! 4".-

A3 ; '$#) !" #$%# + ! # $ %$* # ($ H % . + ! # $
%$* # ($ H % . / ! # $. -- %H $* # ! 4".-

13 ; '$#) !" #$%# + ! # $ %$* # ($ H % . + ! # $
%$* # ($ H % . / ! # $4".-

?3 ; '$#) !" #$%# + ! # $ %$* # ($ H % . + ! # $ %$* # ($4".-
7 4 H % . + ! # $

1

?
!" #$%#
: (
3 , ; 9
!" #$%#& ! !" #$%#& ! +$: 5 % # % : $#: +$: !$/

F/4 ?10 3

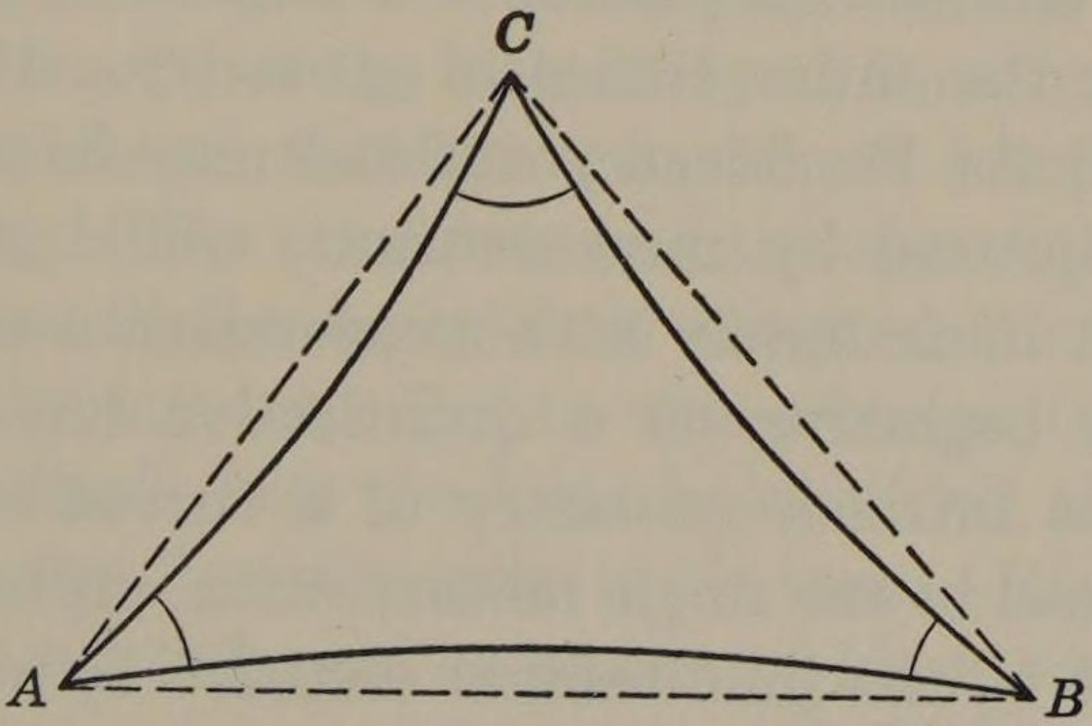

Figure 34

7 4 +$: G - %(4

@$ $ "F/ %$ /: %# 4
3 , ; 9 *7 # 5 % # % $%. #) ! (

Q *. . # ") %)$89 %/ $ 7 -) # 6 - # ! 4".-

3 , ; 9 *7 # 5 % # % $%. #) ! (
Q *. . % # % # 4".-

>3 , ; 9 *7 # 5 % # % $%. #) ! (
Q *. . ; 9 5<# $/$*$(# (# % !6 %. ! 4".-
7 4 # . PPP
9 $# +$: 2 : $#: 4

3 , ; 9 *7 # 5 % # % $%. #) ! (
Q *. . 0 %(# % (%(7 $ 63 '$ 2 5* ,) 6#)$/ $% #) ! $ >2 6 $
) # 64 %(# % % #6 0 34".-
-4
)##" 8884. / :6# 4(! . (:! %# . 4 1? A > ? ".-
: (

B ! #) 7 9 ,) 6#)$/ $% ,) !
5* '$
)##" . 4 / 4 1? A > ?

F/4 ? 0 ?3

*$ƒC (! # # !

$3 , ; 9 *7 # 5 % # % $%. #) ! (
Q *. . 15 % # % $%. + . * 84".-

73 , ; 9 *7 # 5 % # % $%. #) ! (
Q *. . $** E : / $: Q *. Q #) :# , ! ,) B / ## % /$(6 - + . * $%. 5 % # %
0 >34".-

(3 , ; 9 *7 # 5 % # % $%. #) ! (
Q *. . $** E : / $: Q *. Q #) :# , ! ,) B / ## % /$(6 - + . * %.
5 % # % ;$ (; 9 0 >34".-
7 4 ; $. ƒC 4
F/4 > 0 >3 +$:

L - # . (6 8$ #)$# #) K: # %- % #6)$. # 7 …$(
#:$*2… 8) () 8 %# $/$ % # $ #8 #) : $%. 6 $ # $. # % % !$#)
!$# (2 - ! # #* # +$: 2 8) ()) *. #)$# %- % #6 ! *6
…" # %# $*4… ; - $%# 2 # 8$ $< !$# (#)$# %- % #6 8$ % # #
7 (% . . $. - % # %:!7 4 , $62 - <$!"* 2 #)$# #) %$#: $*
%:!7 $ %- % # % %:!7 8$ #$9 % # ! $% % # #)$# #) $%
$(#:$* %:!7 2 %- % #62 #)$# %:!7 #) # - %$#: $* %:!7 2 7:#
$#) #)$# #) # - %$#: $* %:!7 / % - 2 $%. #)$# #)
! # #)$# % ($% $6 #)$# % %$#: $* %:!7 7 / % :/) #
%:!7 #) %# #4 $%# 2 % (%# $ #2 " .:(. $ " 8 -:* $ /:
! %# - #) #) #)$# #) $% $(#:$* %:!7 2 8) ()) ($** . „
0$* ") %:**32 #)$# %:!7 #) # - %$#: $* %:!7 4 $#: $**62)
!")$ H . $ -$(# #)$# 8 ($% ":# $ - ** 8 ,) %:!7 #)$# %:!

7 #) %$#: $* %:!7 ($%% # # *- 7 $ %$#: $* %:!7 4 # !: # 7
$% :%%$#: $*2 :" $%$#: $*2 0$ $%# ()$ $(# H . #3 # $% -
% # %:!7 4 ,) 9 %/ ($%% # $ - ! #) (*$ - " $ $%# 4 Q)$#
#$7*) . #) /% - ($%(- :() $ # $% - % # %:!7 8$ $%# O
(%. / $# . (62 $ " - #)$# „ % # #) %*6 # $% - % #
%:!7 7:# %*6 #) - #2 #) !$** #4 L " - % - #) 8 %
. - #) 8 *.2 * 9 #))$%/ %/ /$. % - ;$76* % #) "6 $
! . - 5/6"#4

A3 , ; 9 *7 # 5 % # % $%. #) ! (
Q *. . 7* 4".-
7 4 %("* -)6 ($* ! * /6

13 , ; 9 *7 # 5 % # % $%. #) ! (
Q *. . !79 1A %.!$## 4".-
7 4 ,) %# % ($#: - ,) %/
;$ 7$ $; **$

?3 , ; 9 *7 # 5 % # % $%. #) ! (
Q *. .) ** " W$! 5.8 % 7* %("* -)6 ($* ! * /6 %(# % % #6
0 34".-

3 , ; 9 *7 # 5 % # % $%. #) ! (
Q *. . W$! 7* ! * /6O %#: 6 %(# % % #6 0 34".-

3 , ; 9 *7 # 5 % # % $%. #) ! (
Q *. . + $ #$# % $%. ! * /6 (. %/ - #) + %- %(4".-

3 , ; 9 *7 # 5 % # % $%. #) ! (
Q *. . + $. L # 6 +, 4".-

?

>> >

3 , ; 9 %(%# .:(# % # (%# - (
:$* H$# % L ** # $%/ ; 9 %(%# .:(# %, (%# -4".-
F/4 10 3

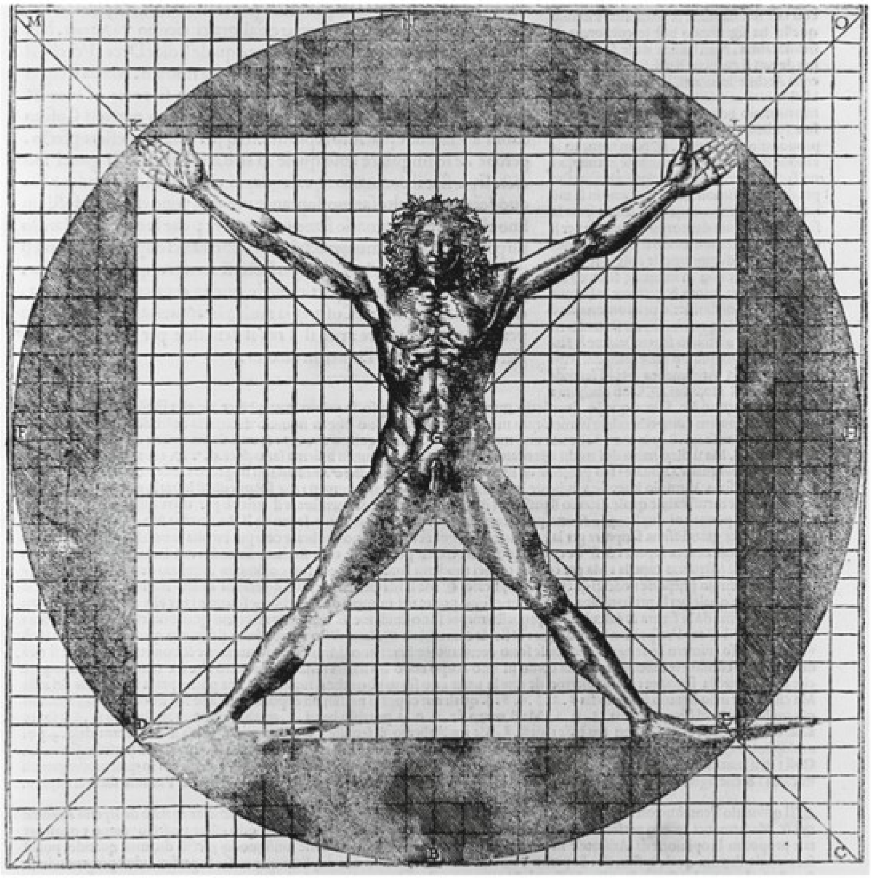
M
N
O
K
L
F
H
G
D
E
A
B
C

B /4 $ $ $% T 0 53 **: # $# . % - () # (#: $ 76 #) !$% $ () # (# # : : 0? G ; 534 ,) z/: . $8% 8 #) 7$(9/ :%. / . - * %/#) (!"$ % 4 " .:(. 8 #) " ! % - ! I A#) %#: 6 !$%: ("# 76 $ $ $% 2 Q 9 ! . $!! % 2 :7* (!$ %J

F/4 >0> 3

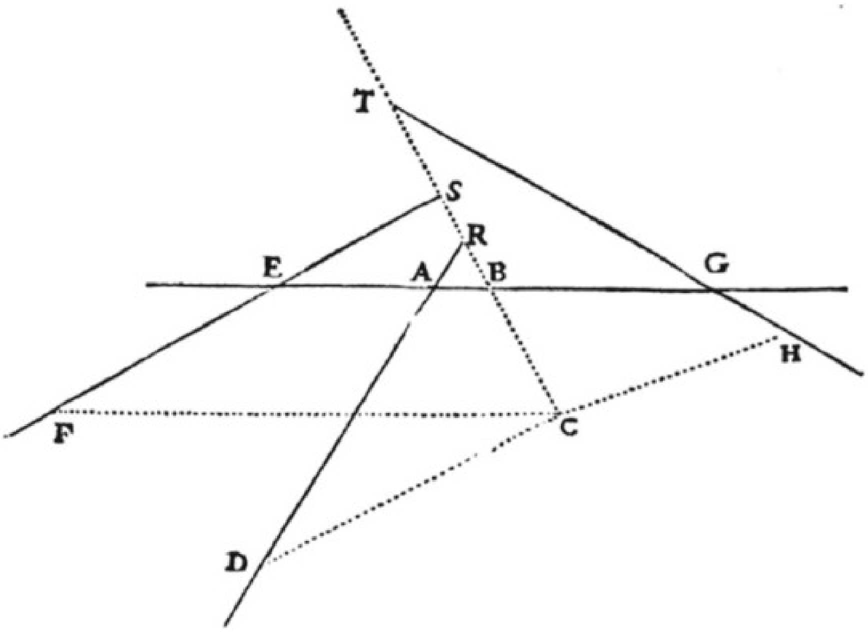
T
S
R
E
A
B
G
H
F
C
D

B /4 ($ # / $") ($* . " (# % - " .:(%/ " %# - $ # $ /)# * % $. (7 . 76 #)
$"": " 7* ! 8 #) - : * % 4 % ($% #) . ## . * % 0" %# 2 ! # ! ($** . * (3 $
#) : . % #) (% # :(# % - #) (: 0* % 3 $# -6 %/ #) K: . (%. # % 4 " .:(.
8 #) " ! % - ! I 1#) %#: 6 '$%: ("# 76 %€ ($ # 2 " % : # :7*) %/ 42
:7* (!$ %J

3 , ; 9 7 # $(# */ 7 $ 5. 7 # $(#
*/ 7 $ 5. ? 4".-
F/4 1 0 3
)$"# ! ") ! - + :"

3 , ; 9 ,) " % ($% : # %
/% # ($# %2 9 "# (!2 $%. * # $* . 7 # 4 Q #!$% ,) " % ($% : # %
/% # ($# %2 9 "# (!2 $%. * # $* . % #6 - $* - % $ 4".-
F/4 ?0 3

4 !"* < ! . % (% # :(# % -) # (: T Q) * - B #:% 4 ,) 5$ #) 0,3 *
(:%# (* (98 % % .
$* 6 $ $:%. 2 $ " %# -- (%# ((%# (# #) :% 0 34 ;:# ((%# (" %# 2) 8% $#
#8 " # % 0 $%. 32
% #: %2 * (* (98 % 2> > 6 $ $:%. #) !$** ((* 0 " (6(* 3 8 #) (%# 2) 8%
$# #8 " # % 0
$%. 3 $ * $7 :# z-# % # ! ! * 8*6 #)$% (:%# (* (98 $:%. #) # : :%
0 34) # (: T Q) *
- B #:% . /%$# . $ #) ((* (%# . % " %# 7 ($: # ! # % (%# * #) 5$ #)T
!$< !:! 0 2 32 ! %
!:! 0; 2 ; 32 ! $% 0% # !$ 9 .3 . #$%(- ! #) :%2 $%.) # (: 7 * . #)$# :()
()$%/ % #) 5$ #)T ((%
(#6 / % . #) # ! 8) % / $# !" 0 ! 3 $%. * / % 0 *$!3 8 :*. $%. -$**4 L
$* 7 * . #)$# #)
! # % - |. . % # . -- !:()† - ! 5* @$)T " ") (6 - < #) : $%. 6 $ - #) (%.
(! %/ -) # 0 4 42 ‡
*:# % 2> >6ˆ 2 1 6 ‰ A2 634 B !) # (: ? 2 G 4 : # 6 - #) . # 4

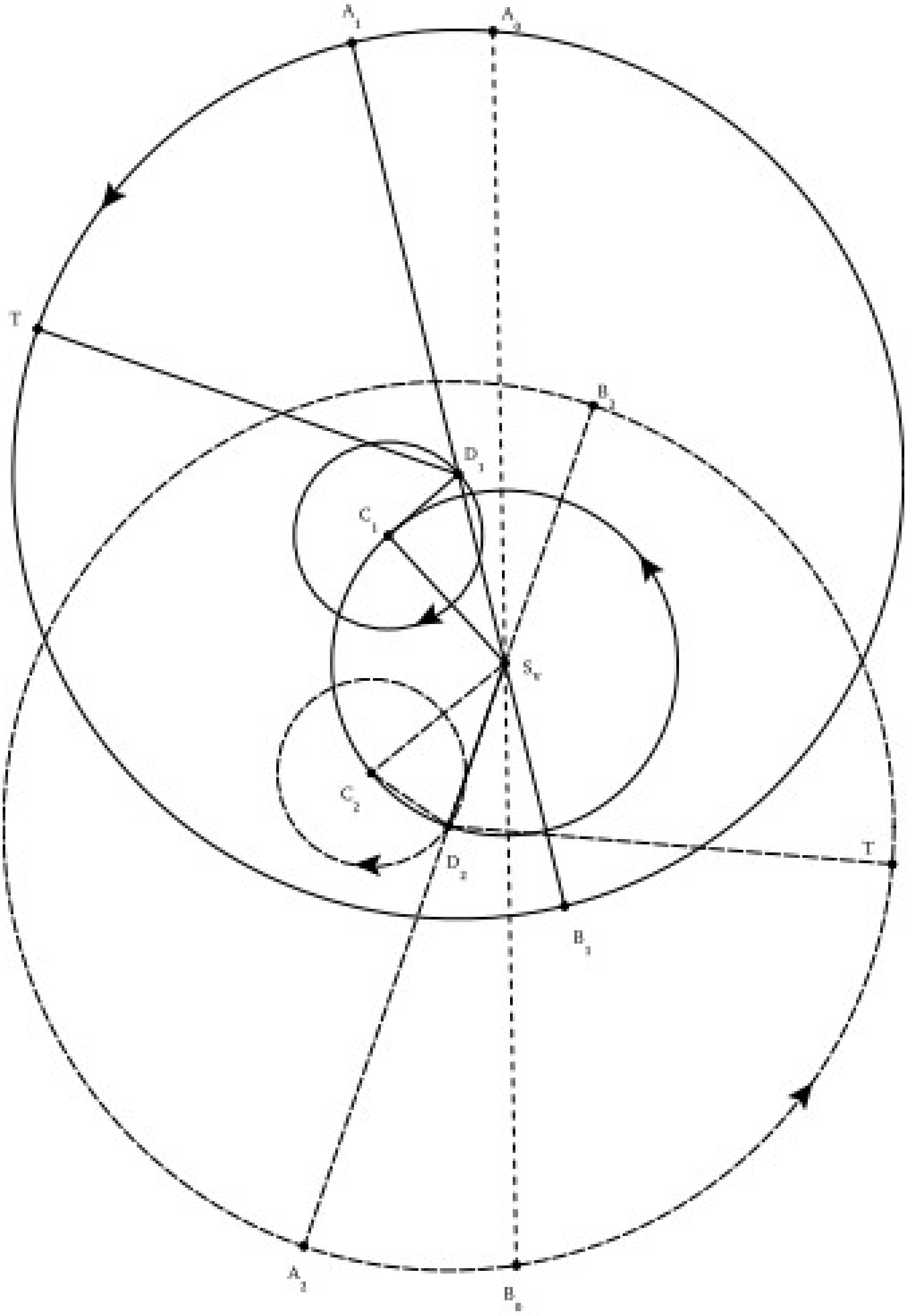
A_1
A_2
T
B_2
D_1
C_1
S_0
C_2
D_2
T
B_1
A_2
B_0

>3 ; !" #$%# A?4".-
7 4 '$#) !$# (L /) *4

3 ; !" #$%# >?A> A1 4".-
7 4 !"* < $ $7* 4

A3 ; !" #$%# ?A >A> 14".-
7 4 !"* < %$*6

13 ; !" #$%# ?A >A> 14".-
F/4 0> 3 G ' # ("$(" # - #) !"* < *$% 4

Corollary 5.4 *For any* $t \in \mathbb{R}$, $|\exp(it)| = 1$.

Proof. We have $\exp(it) = \cos t + i \sin t$. Both $\cos t$ and $\sin t$ are real if t is, and therefore $|\exp(it)|^2 = \cos^2 t + \sin^2 t = 1$. $\square$

Remark 5.5 For real x, $\sin x$ and $\cos x$ are also real, and so the relation $\sin^2 x + \cos^2 x = 1$ shows that $|\sin x| \le 1$ and also $|\cos x| \le 1$. These inequalities do not extend to the complex case. For example, if $z = it$, with $t \in \mathbb{R}$, then $\sin z = \sin it = (\exp(-t) - \exp t)/2i$. If t is large (and positive), then $\exp t$ is large, $\exp(-t)$ is small and so $|\sin it|$ is large. A similar remark applies to $\cos z$.

5.6 The Appearance of π

We wish to discuss various properties of $\sin x$ and $\cos x$ for real x. In particular, we would like to introduce the number π.

Theorem 5.2 *If* $x \in (0, 2)$, *then* $\sin x > 0$.

Proof. Suppose that $0 < x \le 1$. Then

$$\begin{aligned} \cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots \\ &> 1 - \frac{x^2}{2!} - \frac{x^4}{4!} - \frac{x^6}{6!} - \dots \\ &> 1 - \frac{1}{2} - \frac{1}{2^3} - \frac{1}{2^5} - \dots, \end{aligned}$$

since $\dfrac{x^2}{2!} \le \dfrac{1}{2}$, $\dfrac{x^4}{4!} \le \dfrac{1}{4.3.2} < \dfrac{1}{2^3}, \dots$ etc.,

$$= 1 - \frac{\frac{1}{2}}{(1 - \frac{1}{4})} = \tfrac{1}{3}.$$

Hence $\sin' x = \cos x > \frac{1}{3}$ on $[0, 1]$ and so (by the Mean Value Theorem), $\sin x$ is strictly increasing on $[0, 1]$. But $\sin 0 = 0$ and therefore $\sin x > 0$ for $0 < x \le 1$. It follows that

$$\sin x = 2 \sin \tfrac{x}{2} \cos \tfrac{x}{2} > 0$$

whenever $0 < x \le 2$, since, in this case, $\sin \frac{x}{2} > 0$ and $\cos \frac{x}{2} > \frac{1}{3}$. $\square$

Corollary 5.5 *The function* $\cos x$ *is strictly decreasing on* $[0,2]$.

Proof. We have $\cos' x = -\sin x$ which is strictly negative on $(0,2)$, by the theorem. By the Mean Value Theorem, it follows that the function $\cos x$ is strictly decreasing on $[0,2]$. □

Theorem 5.3 $\sin 4 < 0$.

Proof. From the definition of $\sin x$,

$$\begin{aligned}\sin 4 &= 4 - \frac{4^3}{3!} + \frac{4^5}{5!} - \frac{4^7}{7!} + \frac{4^9}{9!} \\ &\qquad - \frac{4^{11}}{11!}\left(1 - \frac{4^2}{12.13}\right) - \frac{4^{15}}{15!}\left(1 - \frac{4^2}{16.17}\right) \\ &\qquad\qquad - \frac{4^{19}}{19!}\left(1 - \frac{4^2}{20.21}\right) - \dots \\ &< 4 - \frac{4^3}{3!} + \frac{4^5}{5!} - \frac{4^7}{7!} + \frac{4^9}{9!} = -\tfrac{268}{405}\end{aligned}$$

and the result follows. □

Theorem 5.4 *There is a unique real number* π *satisfying* $0 < \pi < 4$ *such that* $\sin \pi = 0$. *Furthermore,* $\cos \frac{\pi}{2} = 0$, $\sin \frac{\pi}{2} = 1$ *and* $\cos \pi = -1$.

Proof. We have seen that $\sin x > 0$ for $x \in (0,2)$ and so, in particular, $\sin 1 > 0$. We have also shown that $\sin 4 < 0$. Now, the map $x \mapsto \sin x$ is continuous on $\mathbb{R}$, and so, by the Intermediate Value Theorem, there is some real number, which we will denote by π, with $1 < \pi < 4$ and such that $\sin \pi = 0$.

We must now show that π is the only value in $(0,4)$ obeying $\sin \pi = 0$. To this end, suppose that $\sin \alpha = 0$ with $0 < \alpha < 4$. Then

$$0 = \sin \alpha = 2 \sin \tfrac{\alpha}{2} \cos \tfrac{\alpha}{2}.$$

Hence either $\sin \frac{\alpha}{2} = 0$ or $\cos \frac{\alpha}{2} = 0$ (they cannot both vanish because the sum of their squares is equal to 1). But if $0 < \alpha < 4$, then $0 < \frac{\alpha}{2} < 2$ and we know that $\sin x > 0$ on $(0,2)$, so $\sin \frac{\alpha}{2}$ cannot be zero. Hence, we must have that $\cos \frac{\alpha}{2} = 0$. In particular, $\cos \frac{\pi}{2} = 0$. However, $\cos x$ is strictly decreasing on $[0,2]$ and so there can be at most one solution to $\cos x = 0$ in this interval. Hence $\frac{\alpha}{2} = \frac{\pi}{2}$ so that $\alpha = \pi$, and the uniqueness is established.

Next, $\sin^2 z + \cos^2 z = 1$ implies that $\sin^2 \frac{\pi}{2} = 1$, since $\cos \frac{\pi}{2} = 0$. Hence $\sin \frac{\pi}{2} = 1$ because $\sin \frac{\pi}{2} > 0$ (since $0 < \frac{\pi}{2} < 2$).

Finally, we have $\cos \pi = 2 \cos^2 \frac{\pi}{2} - 1 = -1$. □

What's going on? We have defined the trigonometric functions, out of the blue as it were, by means of power series. This approach avoids any appeal to geometry and right-angled triangles. However, having chosen this route, we must stick with it. In particular, it is necessary to get to π via these definitions rather than by drawing triangles or circles. As we have seen, this is all perfectly possible. Our view here is that the trigonometric functions are as we have defined them. Any results must be deduced as consequences of these power series definitions.

Remark 5.6 Numerical investigation yields $\pi = 3.14159\ldots$. It is known that π is irrational (in fact, transcendental). It is something of a sport (involving some fascinating numerical analysis) to calculate the value of π to a large number of decimal places and this has been done to over a million decimal places. Such programs have been used to test the computational integrity of supercomputers by checking to see whether they get these digits correct or not.

Theorem 5.5 *For any $z \in \mathbb{C}$ and $n \in \mathbb{Z}$,*

$$\begin{aligned}\sin(z + n\pi) &= (-1)^n \sin z\\ \cos(z + n\pi) &= (-1)^n \cos z.\end{aligned}$$

Proof. We use the trigonometric formulae;

$$\begin{aligned}\sin(z + \pi) &= \sin z \cos \pi + \cos z \sin \pi\\ &= -\sin z + 0,\end{aligned}$$

and

$$\begin{aligned}\cos(z + \pi) &= \cos z \cos \pi - \sin z \sin \pi\\ &= -\cos z - 0.\end{aligned}$$

For $n > 0$, the result now follows by induction. Substituting $z = w - n\pi$, the result then follows for $n < 0$. □

Remark 5.7 Particular cases of the above formulae deserve mention. If we set $z = 0$, then we see that $\sin(n\pi) = 0$ and $\cos(n\pi) = (-1)^n$, for any $n \in \mathbb{Z}$. The functions sin and cos are periodic (with period 2π). In particular, for any $x \in \mathbb{R}$ and any $n \in \mathbb{Z}$,

$$\begin{aligned}&\sin(x + 2n\pi) = \sin x \quad \text{and}\\ &\cos(x + 2n\pi) = \cos x.\end{aligned}$$

?3 ; !" #$%# > >4".-
7 4 B:%.$! %#$* - + ! # 6 % # :(# % G 4
F/4 01 3

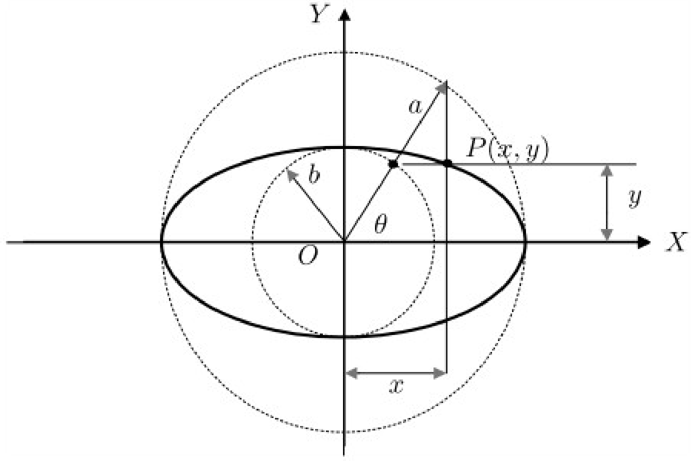
Y
a
P(x, y)
b
θ
y
X
O
x

B /4 41 ,) "$ $! # (" %#$# % - #) ** "

F/4 01A3

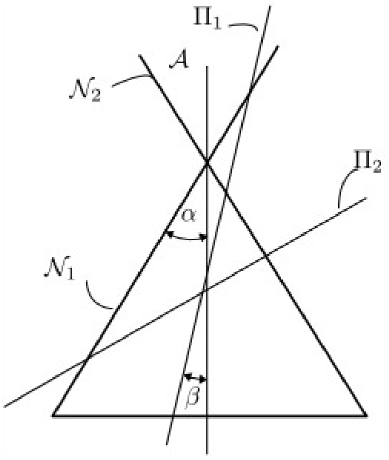
Π1
A
N2
Π2
α
N1
β

B /4 4 ,))6" 7 *$ / % $# . 76 $ "*$% (:## %/ $ /)# ((:*$ (% $# $% $%/* Š ‹ Œ 2 Œ
7 %/ #) $%/* !$. 76 #) * ! %# 8 #) #) (% $<

F/4 A 01?3

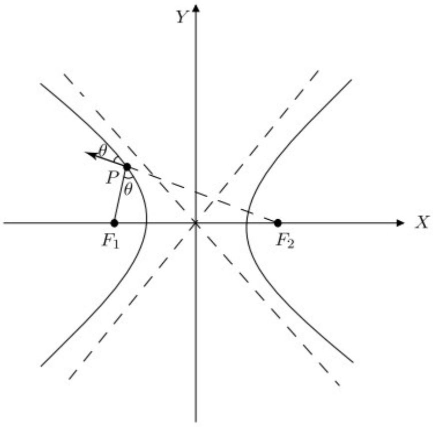

Fig. 2.15 A reflective property of the hyperbola: lines $\overline{F_1 P}$ and $\overline{F_2 P}$ make equal angles with the tangent at P

3 ; !" #$%# AA A 4".-

7 4 ,) % - :$. (

F/4 0 13

> @ (# $%. $ % K:$. (

F/4 >0 ?3

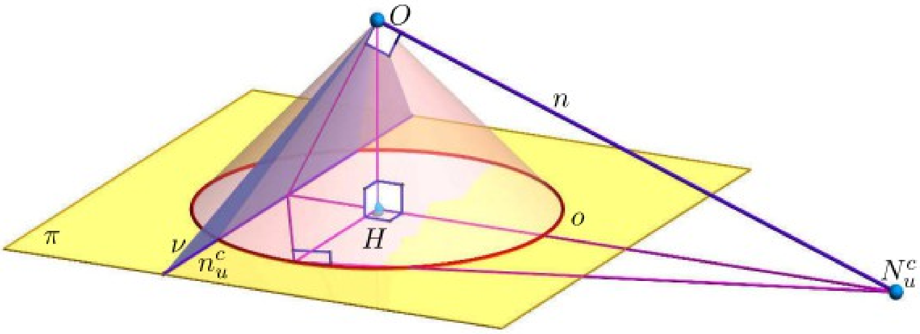

FIGURE 10.11. The Euclidean orthogonality in the star about O is projected to the anti-polarity w.r.t. the distance circle o in the image plane π from the eye point O. The plane ν and the line n are orthogonal, and thus, the respective vanishing line n_u^c and vanishing point N_u^c are antipolar w.r.t. the circle o.

$#) #$%/ %# - $ 5:(* . $% $% K: - ! ! # %

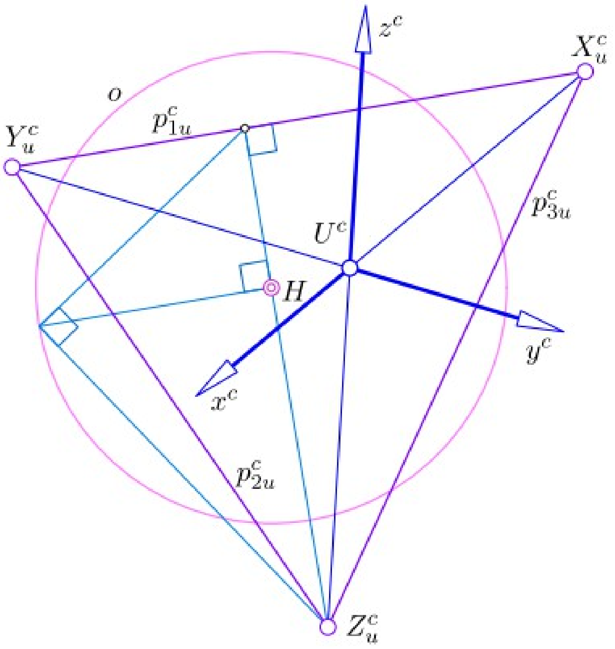

FIGURE 10.12. The vanishing points of X_u^c, Y_u^c, Z_u^c of the axes of a Cartesian frame are the vertices of an antipolar triangle of the distance circle o. The principal (vanishing) point H (orthogonal projection of the eye point O onto the image plane π) is the orthocenter of the triangle of the three vanishing points.

F/4 A0 3

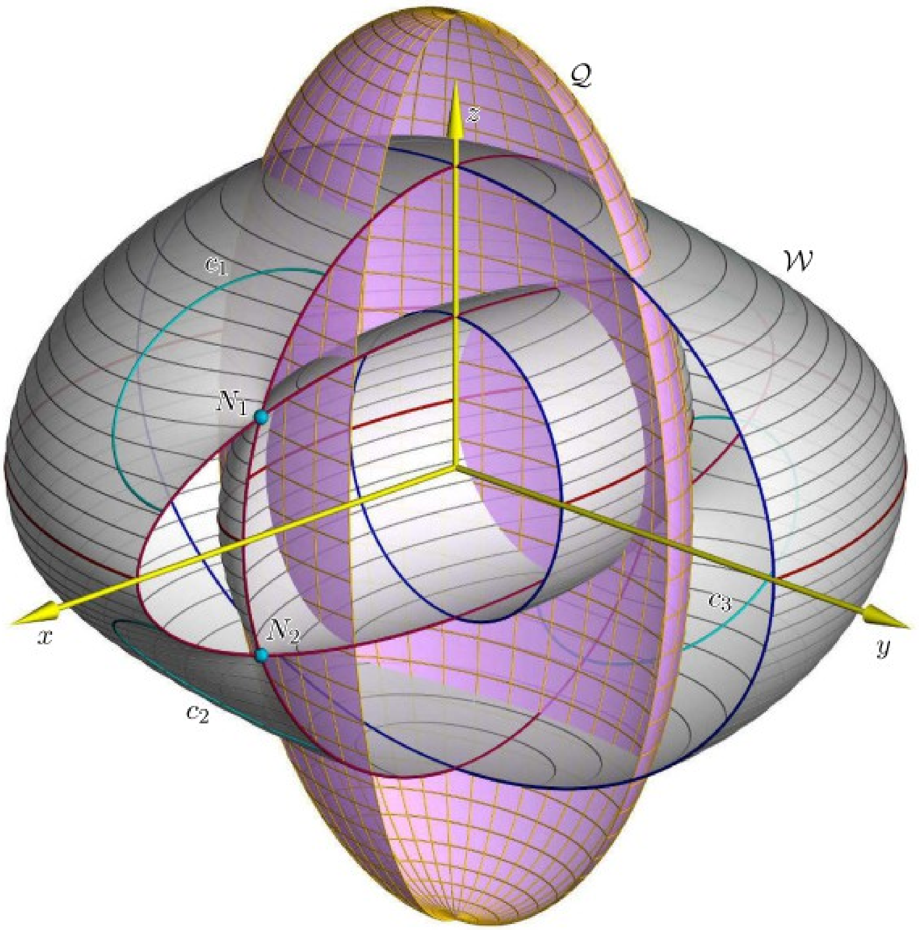

FIGURE 10.16. The quartic Fresnel wave surface $\mathcal{W}$ has four conical nodes N_1, …, N_4 corresponding to the umbilics of an ellipsoid $\mathcal{Q}$. $\mathcal{W}$ intersects its planes of symmetry along circles and ellipses (red and blue). There are four circles c_1, …, c_4 on Fresnel's surface that are the intersection of the four real double tangent planes with $\mathcal{W}$.

Painvin's complex

The tangent planes of a regular quadric $\mathcal{Q}$ form the dual quadric $\mathcal{Q}^\star$. A quadric $\mathcal{Q}$ sends two tangent planes through each line that is not tangent to $\mathcal{Q}$, and thus, a quadric is said to be of *rank* two. This notion of rank is not to be confused with the rank of a matrix, especially with the rank of the coefficient matrix of a quadric. By assumption, $\mathcal{Q}$ is regular, and

3 ; !" #$%# / ! # 64".-
7 4 B:%.$! %#$* - + ! # 64

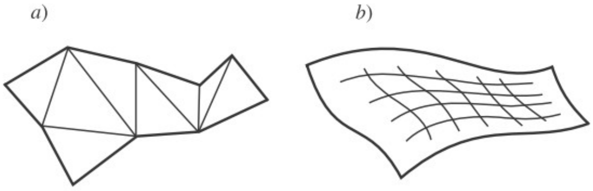

Figure 1.1. Approximate realization (*a*) of a curved surface (*b*).

trying to test spacetime geometry by physical measurements. His/her conclusions will also be valid locally for true gravitational fields, on the basis of the PE. Assume that the observer sits at the centre of the disc, and has two identical clocks: one of them is placed at the centre, the other at some point on the periphery. The observer will see that the clock on the periphery is running slower (time dilatation from SR). Consequently, clocks at various positions in the gravitational field run faster or slower, depending on the local strength of the field. There is no definition of time that is pertinent to the whole spacetime in general. The observer will also conclude that the length of a piece of line orthogonal to the radius of the disc will be shortened (length contraction from SR). Therefore, the ratio of the circumference of the circle to its radius will be smaller than 2π, so that the Euclidean geometry of space does not hold in GR.

Geometry of curved surfaces. We have seen that in spacetime, within a limited region, we can always choose a suitable reference frame, called the local inertial frame. Taking spacetime apart into locally inertial components, we can apply the laws of SR in each such component and derive various dynamical conclusions. Reconstruction of the related global dynamical picture, based on the PE (and some additional, simple geometric assumptions), gives rise to GR.

This procedure for dissecting spacetime into locally inertial components, out of which we can reconstruct, using the PE, the global structure of spacetime containing the gravitational field, can be compared geometrically with an attempt to build a curved surface, approximately, from a lot of small, plane elements 'continuously' bound to each other (figure 1.1). A locally flat surface is a geometric analogue of locally inertial spacetime.

GR makes essential use of curved spaces (or, more accurately curved spacetimes). Since the curved, four-dimensional spacetime cannot be visualized, let us try to understand the basic features of curved spaces by considering a two-dimensional surface X_2. We shall focus our attention on those geometric properties of X_2, which could be determined by an intelligent, two-dimensional being entirely confined to live and measure in the surface, a being that does not know anything about the embedding Euclidean space E_3. Such properties determine the *intrinsic geometry* on X_2.

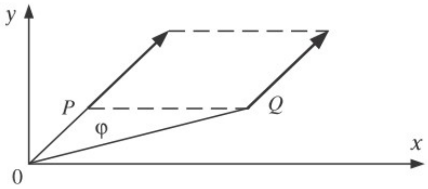

Figure 1.2. Vector components change under the parallel transport.

On a general surface we cannot set up Cartesian coordinates; the best we can do is to introduce two arbitrary families of coordinate lines and label them by $u^\alpha = (u^1, u^2)$. These curvilinear (or Gaussian) coordinates do not always have a direct geometric interpretation.

Let us now imagine a surface X_2 embedded in E_3, and consider a point P in X_2 with Cartesian coordinates (x, y, z) in E_3; since P is in X_2, these coordinates can be expressed as functions of u^α. The squared distance between two neighbouring points in X_2, $P(x, y, z)$ and $Q(x+\mathrm{d}x, y+\mathrm{d}y, z+\mathrm{d}z)$, is given by $\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2$. Going over to u^α we obtain

$$\mathrm{d}s^2 = g_{\alpha\beta}\, \mathrm{d}u^\alpha\, \mathrm{d}u^\beta \tag{1.6}$$

where the set of functions $g_{\alpha\beta}(u^1, u^2)$ defines the *metric* on X_2. The metric is an essential intrinsic property of X_2, independent of its embedding in E_3. A two-dimensional continuum X_2, equipped with a squared differential distance according to (1.6), becomes the *metric space*, $G_2 = (X_2, \boldsymbol{g})$ (it is now clear that the symbol X_2 is used to denote an abstract two-dimensional continuum, disregarding all metric properties of the real surface).

Studying the motion of particles in X_2 our observer could have arrived at the idea of the tangent vector, $\boldsymbol{a} = (a^1, a^2)$. Consider, further, two tangent vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, defined at neighbouring points P and Q. If we wish to compare these two vectors we have to know how one of them can be transported to the position of the other. This transport is called *parallel transport*. The continuum X_2 together with the parallel transport law, denoted by Γ, is called a *linearly connected* (or *affine*) space, $L_2 = (X_2, \Gamma)$.

The components of a vector in curvilinear coordinates are expected to change under parallel transport. This can be easily seen by considering the parallel transport of a vector in a plane, using polar coordinates (r, θ). As shown in figure 1.2, if the unit vector $\boldsymbol{a}$ at point P has components $(a^r = 1, a^\theta = 0)$, after parallel transport from P to Q its components become $(a^r = \cos\varphi, a^\theta = \sin\varphi)$.

A two-dimensional observer should determine the rule of parallel transport of vectors in accordance with his/her own experience. This rule is generally independent of the concept of a metric. If X_2 is equipped with a rule of

parallel transport and the independent concept of a metric, we obtain the *linearly connected metric* space $(X_2, \Gamma, \boldsymbol{g}) = (L_2, \boldsymbol{g}) = (G_2, \Gamma)$.

It is very probable that our two-dimensional observer will demand that the lengths of vectors should remain unchanged under parallel transport. After adopting this very natural property, which relates Γ and $\boldsymbol{g}$, the linearly connected metric space becomes the *Riemann–Cartan* space U_2. The usual surfaces have some additional geometric structure, and belong to the class of *Riemann* spaces V_2.

This discussion was slightly more general than strictly necessary for studying surfaces embedded in E_3, in order to emphasize that the intrinsic geometry on X_2 is defined in terms of

- the metric $\boldsymbol{g}$ and
- the rule of parallel transport Γ.

These basic concepts of the intrinsic geometry of surfaces can be extended directly to spaces with higher dimensions, although in this case we lose the intuitive geometric picture.

Any Riemann space is locally Euclidean, i.e. its metric is positive definite ($\mathrm{d}s^2 > 0$). A slight generalization of this case, that consists in admitting metrics that are pseudo-Euclidean, leads to a pseudo-Riemannian geometry or Riemannian geometry with an indefinite metric.

Relativity, covariance and Mach's ideas

Relativity and covariance. We have seen previously that inertial frames play a particularly important role in SR. Einstein considered this to be an expression of the incompleteness of SR. He wanted to generalize the relativity of inertial motions to the relativity of all motions, including accelerated ones. This idea can be formulated as the general principle of relativity:

> *General PR: The form of physical laws is the same in all reference frames.*

Equations expressing physical laws are often formulated using specific coordinates. Thus, for instance, spatial positions in an inertial laboratory L can be determined using Cartesian coordinates (x, y, z) or the coordinates obtained from these by a rotation around the z-axis for a fixed angle φ. If we allow the angle φ to change with time as $\varphi = \omega t$, new coordinates can be naturally attached to a new physical reference frame, rotating relative to L with the angular velocity ω around the z-axis. We see that coordinate transformations can be interpreted as describing transitions from one to another laboratory (the so-called active interpretation). Thus, it becomes clear that the general PR can be realized with the help of the principle of general covariance:

> >

A >> ?
3 ; !" #$%# ,))6 (- 5% /64".-

3 ; !" #$%# 1 A 14".-
7 4 5* (# (#6 $%. '$/% # ! : (**

3 ; !" #$%# ? A 1 4".-
7 4 5* (# !$/% # ('$. 5$ 64

>3 ; !" #$%# ".-4":7 / ! # $4".-
7 4 $ # **
. PPP

3 ; !" #$%# %(#$ % @ (# + ! # 6 #$# # ($* $ % %/ -
*6) . $* 7@ (# (% # :(# % 0 >2 " %/ *$/ ; * % L . *7 /34".-
F/4 0 3
4 4 @ (# "$($%. L ! / % : . %$#

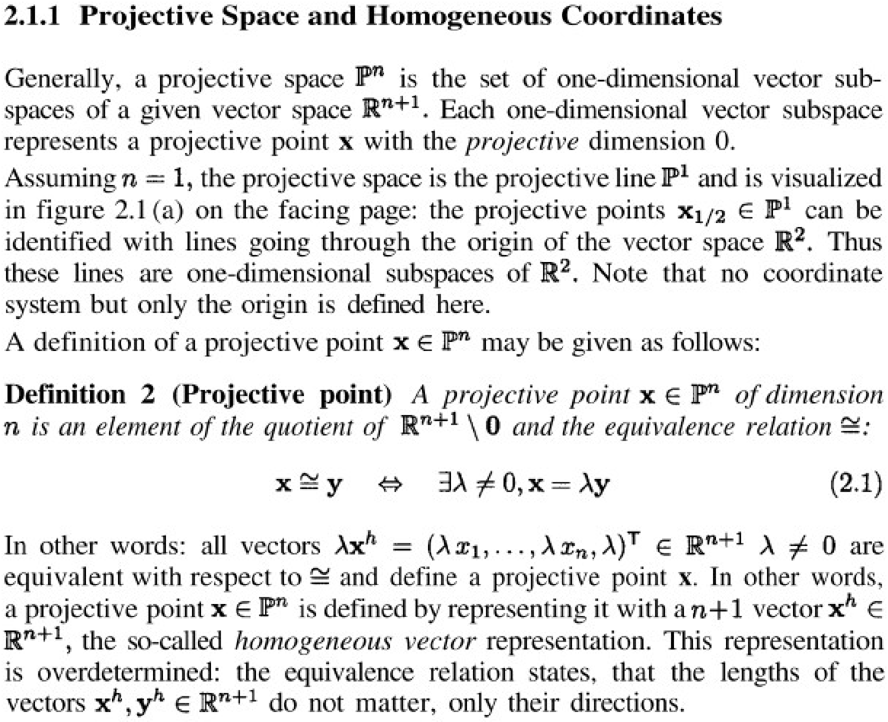

2.1.1 Projective Space and Homogeneous Coordinates

Generally, a projective space $\mathbb{P}^n$ is the set of one-dimensional vector subspaces of a given vector space $\mathbb{R}^{n+1}$. Each one-dimensional vector subspace represents a projective point $\mathbf{x}$ with the *projective* dimension 0.

Assuming $n = 1$, the projective space is the projective line $\mathbb{P}^1$ and is visualized in figure 2.1 (a) on the facing page: the projective points $\mathbf{x}_{1/2} \in \mathbb{P}^1$ can be identified with lines going through the origin of the vector space $\mathbb{R}^2$. Thus these lines are one-dimensional subspaces of $\mathbb{R}^2$. Note that no coordinate system but only the origin is defined here.

A definition of a projective point $\mathbf{x} \in \mathbb{P}^n$ may be given as follows:

Definition 2 (Projective point) *A projective point* $\mathbf{x} \in \mathbb{P}^n$ *of dimension* n *is an element of the quotient of* $\mathbb{R}^{n+1} \setminus \mathbf{0}$ *and the equivalence relation* $\cong$*:*

$$\mathbf{x} \cong \mathbf{y} \quad \Leftrightarrow \quad \exists \lambda \neq 0, \mathbf{x} = \lambda \mathbf{y} \tag{2.1}$$

In other words: all vectors $\lambda \mathbf{x}^h = (\lambda x_1, \ldots, \lambda x_n, \lambda)^\mathsf{T} \in \mathbb{R}^{n+1}$ $\lambda \neq 0$ are equivalent with respect to $\cong$ and define a projective point $\mathbf{x}$. In other words, a projective point $\mathbf{x} \in \mathbb{P}^n$ is defined by representing it with a $n+1$ vector $\mathbf{x}^h \in \mathbb{R}^{n+1}$, the so-called *homogeneous vector* representation. This representation is overdetermined: the equivalence relation states, that the lengths of the vectors $\mathbf{x}^h, \mathbf{y}^h \in \mathbb{R}^{n+1}$ do not matter, only their directions.

them within the Euclidean framework. Therefore we need a more general framework which enables us to deal with infinite points, this mathematical framework is known as *projective geometry*.

In this and the following chapters, we will see that projective geometry provides a simple and consistent way to represent and transform geometric entities, that are important for Euclidean polyhedral reconstruction. One drawback though is the fact that properties such as parallelity or orthogonality can not be directly used; more precisely, they are not invariant with respect to projective transformations. The study of invariance of properties with respect to certain groups of transformation may be regarded as a definition of geometry, as KLEIN 1939 mentioned in his Erlanger Programm. We will come back to the issue of invariance in section 3.3 on page 70, where we investigate possible relations between geometric entities and their invariance properties with respect to Euclidean, similar, affine and projective geometry. We will see that for our application it is possible to infer parallelity and orthogonality of lines and planes from homogeneous vectors by using the Euclidean interpretation thereof.

For a detailed and more precise introduction to projective geometry in the context of Computer Vision, see FAUGERAS AND LUONG 2001 or chapter 23 in MUNDY 1992.

2.1.1 Projective Space and Homogeneous Coordinates

Generally, a projective space $\mathbb{P}^n$ is the set of one-dimensional vector subspaces of a given vector space $\mathbb{R}^{n+1}$. Each one-dimensional vector subspace represents a projective point $\mathbf{x}$ with the *projective* dimension 0.

Assuming $n = 1$, the projective space is the projective line $\mathbb{P}^1$ and is visualized in figure 2.1 (a) on the facing page: the projective points $\mathbf{x}_{1/2} \in \mathbb{P}^1$ can be identified with lines going through the origin of the vector space $\mathbb{R}^2$. Thus these lines are one-dimensional subspaces of $\mathbb{R}^2$. Note that no coordinate system but only the origin is defined here.

A definition of a projective point $\mathbf{x} \in \mathbb{P}^n$ may be given as follows:

Definition 2 (Projective point) *A projective point* $\mathbf{x} \in \mathbb{P}^n$ *of dimension* n *is an element of the quotient of* $\mathbb{R}^{n+1} \setminus \mathbf{0}$ *and the equivalence relation* $\cong$*:*

$$\mathbf{x} \cong \mathbf{y} \quad \Leftrightarrow \quad \exists \lambda \neq 0, \mathbf{x} = \lambda \mathbf{y} \tag{2.1}$$

In other words: all vectors $\lambda \mathbf{x}^h = (\lambda x_1, \ldots, \lambda x_n, \lambda)^\mathsf{T} \in \mathbb{R}^{n+1}$ $\lambda \neq 0$ are equivalent with respect to $\cong$ and define a projective point $\mathbf{x}$. In other words, a projective point $\mathbf{x} \in \mathbb{P}^n$ is defined by representing it with a $n+1$ vector $\mathbf{x}^h \in \mathbb{R}^{n+1}$, the so-called *homogeneous vector* representation. This representation is overdetermined: the equivalence relation states, that the lengths of the vectors $\mathbf{x}^h, \mathbf{y}^h \in \mathbb{R}^{n+1}$ do not matter, only their directions.

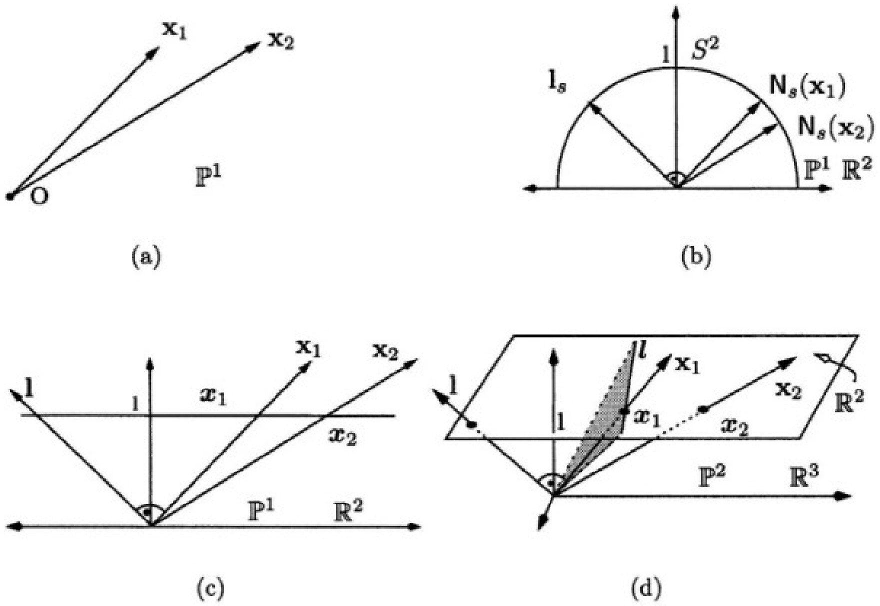
$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbb{P}^1$
O
(a)
1
S^2
$\mathbf{l}_s$
$\mathsf{N}_s(\mathbf{x}_1)$
$\mathsf{N}_s(\mathbf{x}_2)$
$\mathbb{P}^1$ $\mathbb{R}^2$
(b)
$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{l}$
1
$\boldsymbol{x}_1$
$\boldsymbol{x}_2$
$\mathbb{P}^1$ $\mathbb{R}^2$
(c)
$\boldsymbol{l}$
$\mathbf{x}_1$
$\mathbf{l}$
1
$\boldsymbol{x}_1$
$\mathbf{x}_2$
$\boldsymbol{x}_2$
$\mathbb{R}^2$
$\mathbb{P}^2$ $\mathbb{R}^3$
(d)

Fig. 2.1. Different interpretations of the projective line $\mathbb{P}^1$: in (a), projective points $\mathbf{x}_1, \mathbf{x}_2$ are interpreted as lines going through the origin without defining any coordinate system. In (b), the projective points are homogeneous vectors, that are normalized to the sphere S^2. In (c) the homogeneous vectors intersect the Euclidean axis and define the points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ on the Euclidean axis. The role of a hyperplane $\mathbf{l}$ can be visualized by considering the projective plane $\mathbb{P}^2$ in (d): two projective points $\mathbf{x}_1$ and $\mathbf{x}_2$ define two points $\boldsymbol{x}_1, \boldsymbol{x}_2$ on the Euclidean plane $\mathbb{R}^2$. The first point $\boldsymbol{x}_1$ is incident to a line $\boldsymbol{l}$ on the Euclidean plane. The line $\boldsymbol{l}$ in turn is defined by the homogeneous vector $\mathbf{l} \in \mathbb{R}^3 \simeq \mathbb{P}^2$, which is the normal of the plane that contains the origin of $\mathbb{R}^3$ and the line on the Euclidean plane $\mathbb{R}^2$. As the first Euclidean point $\boldsymbol{x}_1$ is incident to the Euclidean line, the homogeneous vector $\mathbf{l}$ is orthogonal to the homogeneous vector $\mathbf{x}_1$, or $\mathbf{l}^{\mathsf{T}}\mathbf{x}_1 = 0$.

To simplify the notation, we identify projective points $\mathbf{x} \in \mathbb{P}^n$ with their (non-unique) coordinate vector representation based on $\mathbf{x}^h \in \mathbb{R}^{n+1}$, i.e. $\mathbf{x} = \mathbf{x}^h$. Therefore we are not strict with respect to mathematical notation and do *not* distinguish between projective points and its coordinate vectors as in FAUGERAS AND LUONG 2001, p. 78. HARTLEY AND ZISSERMAN 2000 take a similar approach, see p. 3f. In the following, we assume that the correct interpretation is clear from the context.

One can *normalize* a homogeneous vector $\mathbf{x}_s := \frac{\mathbf{x}}{|\mathbf{x}|}$ to avoid ambiguities, see fig.2.1(b). The index s in the homogeneous vector $\mathbf{x}_s$ indicates that the vector is normalized and thus lies on the unit sphere. We could use the normalization of a homogeneous vector with its length for an alternative

definition of the equivalence class $\cong$ for two homogeneous vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n+1} \setminus \mathbf{0}$:

$$\mathbf{x} \cong \mathbf{y} \quad \Leftrightarrow \quad \left(\frac{\mathbf{x}}{|\mathbf{x}|} = \frac{\mathbf{y}}{|\mathbf{y}|} \text{ or } \frac{\mathbf{x}}{|\mathbf{x}|} = -\frac{\mathbf{y}}{|\mathbf{y}|}\right) \quad \Leftrightarrow \quad \frac{\mathbf{x}}{|\mathbf{x}|} = \pm\frac{\mathbf{y}}{|\mathbf{y}|} \tag{2.2}$$

We sometimes denote the spherical normalization by $\mathsf{N}_s(\mathbf{x}) = \frac{\mathbf{x}}{|\mathbf{x}|}$.

The definition of the equivalence class $\cong$ is ambiguous as the orientation of the normalized vectors can be positive or negative. In other words, a projective point from $\mathbb{P}^n$ has two representations on a sphere S^{n+1}. To overcome this ambiguity one may use *oriented* projective geometry, as proposed by STOLFI 1991, where every element has an intrinsic orientation. This is analytically expressed by the specialized definition, that homogeneous vectors are identical if they differ by a *positive* scale factor. But the ambiguity is not critical for our application, so we will keep the definition in equation (2.2).

The link between an original Euclidean space $\mathbb{R}^n$ and the projective space $\mathbb{P}^n$ is now quite easy: a simple mapping $\mathbb{R}^n \to \mathbb{P}^n$ of the Euclidean space to the projective space can be written as follows:

$$\boldsymbol{x} = (x_1, \ldots, x_n)^\mathsf{T} \in \mathbb{R}^n \to \mathbf{x}_e = (x_1, \ldots, x_n, 1)^\mathsf{T} \in \mathbb{P}^n \tag{2.3}$$

We will denote vectors from $\mathbb{R}^n$ with italic bold letters such as $\boldsymbol{x}, \boldsymbol{y}$, whereas homogeneous vectors from $\mathbb{P}^n \simeq \mathbb{R}^{(n+1)}$ are denoted with upright bold letters, $\mathbf{x}, \mathbf{y}$.

The index e in the homogeneous vector $\mathbf{x}_e$ denotes that we can directly extract the Euclidean information from the homogeneous vector. It is a unique member of the equivalence class defined by $\mathbf{x} \cong \mathbf{x}_e$. In figure 2.1(c) one may interpret the fixing of the last coordinate of the homogeneous vector to 1 as follows: take the line parallel to the first axis, intersecting the second axis in 1. This line represents the real axis $\mathbb{R}^1$, on which the Euclidean numbers $\boldsymbol{x}_{1/2}$ are fixed. Connecting the origin with $\boldsymbol{x}_{1/2}$ on the constructed line yield the projective points.

For the inverse direction of (2.3), any projective point $\mathbf{x} \in \mathbb{P}^n$ (except of ideal points) can be mapped to an euclidized version, see also figure 2.2:

$$\mathbf{x} = (x_1, \ldots, x_n, x_{n+1})^\mathsf{T} \in \mathbb{P}^n \to \mathbf{x}_e = 1/x_{n+1}\,(x_1, \ldots, x_n, 1)^\mathsf{T} \in \mathbb{P}^n \tag{2.4}$$

For lower dimensions, we will replace the coordinates x_1, x_2, x_3 with the letters u, v, w to avoid overloaded use of indices.

Linear Transformations of Projective Points A linear transformation H from $\mathbb{P}^n$ to $\mathbb{P}^m$ can be represented as a $(m+1) \times (n+1)$, such that for $\mathbf{x} \in \mathbb{P}^n$ and $\mathbf{x}' \in \mathbb{P}^m$:

$$\mathbf{x}' = \mathsf{H}\mathbf{x} \tag{2.5}$$

If $n = m$ and the matrix H is regular, we call H a *homography.*

Properties that remain invariant under a transformation H are called projectively invariant. Examples for such properties are collinearity and incidence.

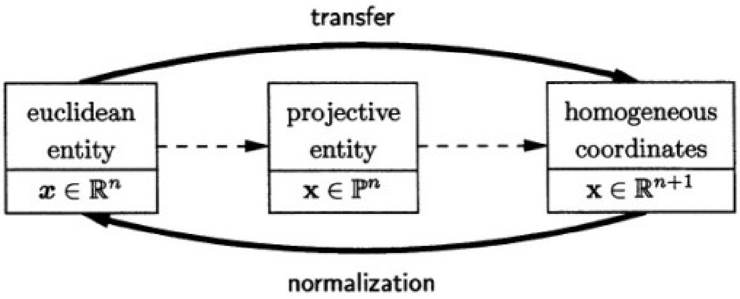
transfer
euclidean entity
$\boldsymbol{x} \in \mathbb{R}^n$
projective entity
$\mathbf{x} \in \mathbb{P}^n$
homogeneous coordinates
$\mathbf{x} \in \mathbb{R}^{n+1}$
normalization

Fig. 2.2. Transfer between Euclidean and Homogeneous Coordinates: giving a Euclidean point $\boldsymbol{x} \in \mathbb{R}^n$ we can derive a projective point $\mathbf{x} \in \mathbb{P}^n$ which we represent as a homogeneous vector $\mathbf{x} \in \mathbb{R}^{n+1}$, see also equation (2.3). Given a homogeneous vector $\mathbf{x} \in \mathbb{R}^{n+1}$, we may obtain a Euclidean representation by normalization, yielding $\boldsymbol{x} \in \mathbb{R}^m$, equation (2.4). In general, the normalization is done by dividing the homogeneous vector by its homogeneous part as described in section 2.2. Note that we use the symbol **x** both for a projective point and for homogeneous vectors representing the projective points.

2.1.2 Hyperplanes

A point in projective space can be regarded as the smallest linear subspace that is non-zero. Opposed to a point, a *hyperplane* is the largest linear subspace that is not equal to the complete space. Examples for hyperplanes are lines in 2D and planes in 3D.

In general, a hyperplane $\mathbf{h}$ in $\mathbb{P}^n$ is defined by n linearly independent points $\mathbf{y}_1, \cdots, \mathbf{y}_n$: for example in $\mathbb{P}^2$, two points $\mathbf{y}_1, \mathbf{y}_2$ define a line and in $\mathbb{P}^3$ three points $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3$ define a plane. Any additional point $\mathbf{x}$ is incident to the a hyperplane $\mathbf{h}(\mathbf{y}_1, \cdots, \mathbf{y}_n)$ if the determinant of all concatenated homogeneous vectors is zero

$$|\mathbf{y}_1\, \mathbf{y}_2\, \cdots \mathbf{y}_n\, \mathbf{x}| = 0 \tag{2.6}$$

This is a generalization of the well-known case when checking whether three 2D points are collinear BRONSTEIN *et al.* 1996:

$$\boldsymbol{y}_1, \boldsymbol{y}_2 \text{ and } \boldsymbol{x} \text{ are collinear} \quad \Leftrightarrow \quad \begin{vmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \boldsymbol{x} \\ 1 & 1 & 1 \end{vmatrix} = 0 \tag{2.7}$$

which can be generalized to the to the 3D case when checking whether four points are coplanar:

$$\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3 \text{ and } \boldsymbol{X} \text{ are coplanar} \quad \Leftrightarrow \quad \begin{vmatrix} \boldsymbol{Y}_1 & \boldsymbol{Y}_2 & \boldsymbol{Y}_3 & \boldsymbol{X} \\ 1 & 1 & 1 & 1 \end{vmatrix} = 0 \tag{2.8}$$

(2.9)

An alternative definition of a hyperplane can be obtained by using the Laplacian expansion on the determinant in equation (2.6): if we expand by the last column, we obtain $n-1$ minors $|\mathsf{M}_{i,n}|$ that depend on the vectors $\mathbf{y}_1, \cdots \mathbf{y}_n$. With $h_i = (-1)^{i+n}|\mathsf{M}_{i,n}|$, equation (2.6) is equivalent to

$$\sum_i h_i x_i = \mathbf{h}^\mathsf{T}\mathbf{x} = \mathbf{x}^\mathsf{T}\mathbf{h} = 0 \tag{2.10}$$

Both (2.6) and its equivalent equation (2.10) is a necessary and sufficient condition to check whether a point $\mathbf{x}$ is incident to a hyperplane, $\mathbf{x} \in \mathbf{h}$.

Since the minors h_i with $i = 1, \cdots, n$ uniquely define a hyperplane, we can identify them with the coordinates of a hyperplane, defining the vector $\mathbf{h} = (h_1, \cdots, h_n)^\mathsf{T}$. The coordinates h_i are called Plücker coordinates, see also section 2.2.4. The vector $\mathbf{h}$ is a homogeneous vector as (2.10) is a homogeneous equation. As for projective points, it is assumed that it is clear from the context whether a homogeneous vector is interpreted as a projective point or a hyperplane: in the following lines in 2D are usually denoted by $\mathbf{l}$, $\mathbf{m}$ and planes in 3D are denoted by $\mathbf{A}, \mathbf{B}$ opposed to points denoted by $\mathbf{x}$, $\mathbf{y}$ resp. $\mathbf{X}, \mathbf{Y}$.

It is important to note that $\mathbf{h}$ is actually a linear form of $\mathbb{R}^{n+1}$, since each coordinate is a function of the form $\mathbb{R}^{n+1} \to \mathbb{R}$. Thus the vector $\mathbf{h}$ represents an element of the dual space $(\mathbb{R}^{n+1})^*$ or equivalently $(\mathbb{P}^n)^*$.

Given a linear transformation $\mathbf{x}' = \mathsf{H}\mathbf{x}$ for projective points as in (2.5), one may be interested in the corresponding transformation for hyperplanes. H is a linear mapping from $\mathbb{P}^n$ to $\mathbb{P}^m$, its transpose H^T is defined as the mapping from $(\mathbb{P}^m)^*$ to $(\mathbb{P}^n)^*$. With $\mathbf{h}' \in (\mathbb{P}^m)^*$ being the imaged hyperplane of $\mathbf{h} \in (\mathbb{P}^n)^*$ with respect to H, one obtains

$$\mathsf{H}^\mathsf{T}\mathbf{h}' = \mathbf{h} \tag{2.11}$$

If H is a homography, thus $n = m$ and H being regular, one can prove (2.11) as follows: given $\mathbf{x}' = \mathsf{H}\mathbf{x}$, we are looking for the corresponding transformation $\mathbf{h}' = \mathsf{J}\mathbf{h}$, such that for all $\mathbf{y} \in \mathbf{h}$ and $\mathbf{y}' \in \mathbf{h}'$: $\mathbf{y}' = \mathsf{H}\mathbf{y}$. Thus the following equation must hold

$$\mathbf{y}'^\mathsf{T}\mathbf{h}' = \mathbf{x}'^\mathsf{T}\mathsf{J}\mathbf{h} = \mathsf{x}^\mathsf{T}\mathsf{H}^\mathsf{T}\mathsf{J}\mathbf{h} = 0$$

This can be true only if $\mathsf{J} = \mathsf{H}^{-\mathsf{T}}$.

2.2 Representation of Geometric Entities

This section introduces the basic geometric entities, that will be used throughout the rest of this text. As we want to reconstruct polyhedral objects, it is sufficient to deal with geometric entities that can be modeled as linear subspaces, cf. section 1.3.2 on page 6. Therefore only points, lines and planes

?
3 , ; 9 $*(:*$# $*(:*$# % - " # 2 2
. (!$* "*$(7$. % #) +$: / %. $*/ #)! $%. +$: $ (#$%/ %# *$# % <"7 9
4".-

F/4 0 3 G $%. # B %.

3 , ; 9 $*(:*$# $*(:*$# % - " # 2 2
. (!$* "*$(7$. % #) +$: / %. $*/ #)! $%. +$: $ (#$%/ %#
$# % ,) !":#$# % - " # A (!$ / # 4".-

3 , ; 9 $*(:*$# $*(:*$# % - " # 2 2
. (!$* "*$(7$. % #) +$: / %. $*/ #)! $%. +$: $ (#$%/ %#
*$# % " !":# 4".-
7 4
M $%. # (!":#$# % #) :/) #) $/

>3 , ; 9 $*(:*$# $*(:*$# % - " # 2 2
. (!$* "*$(7$. % #) +$: / %. $*/ #)! $%. +$: $ (#$%/ %#
*$# % ' %#)*6 - %(4".-

3 , ; 9 $*(:*$# $*(:*$# % - " # 2 2
. (!$* "*$(7$. % #) +$: / %. $*/ #)! $%. +$: $ (#$%/ %# *$# % >?1>>4".-
7 4
BB, G B$ # B : , $% - !
$*(:*: - 4

? >

! " # $ % &' $ # # (
#") %# * + , # - ../ ' 0

! " # $ % &' $ # # (
#") %# * + , # - ../ '1 2

! " # $ % $ # # (
3 # 2 # 2 4 $ #" 5647 3 5 8 56 - / '1 2

9 ! " # $ % $ # # (
3 # 2 # 2 4 $ #" 5# 8 : 5# ' - / '

! " # $ % $ # # $ %
2 1 (1% 11 # 5# 8 : 5# ' - ... '1 2

; ! " # 3% 6 # &# + 1 2
$ % 1 #" 6 % '1 2

/ ! " # 3% 6 # &# + 1 2
$ % 1 #" 7 - ./ '1 2

< ! " # = $ % =
$ 4 " - / '1 2

. ! " # $ > #" 3 # # # 3 # 1 #
11 % # # # # ? % + 1 #" " 6 @ A >
- 9 '1 2

! " # $ > #" # # # # 1 #
1 #" - 9 ' 0
!? ' 5 1 # 1 ? #B ' CCC

! " # $ > #" # # # # 1 #
1 #" - 9 '1 2

! " # 3% = 2 $ % #
5# # &22 3% #> #" * + - '1 2

! " # 9 &D1 #": 5# " #" #
4 #" # $ % E >%F E - 9 '1 2

9 ! " # 9 5 11 D # 2 G D
8 # 1 #" " E # = ? " - / '1 2

! " # " '1 2
!? ' & # (4 '
& # 1 '

; ! " # " '1 2
!? ' & # (4 '
& # 1 '

/ ! " # " '1 2
!? ' 7# + @ % # + (+ H $ +'
& # 1 '

< ! " # " '1 2
!? ' 7# + @ % # + (+ H $ +'
& # 1 '

. ! " # ; ./. '1 2
!? ' G # # 2 $ % '

! " # ; 9 .<9 9 '1 2
!? ' "% H 3% 8%+ 2 8% # '

! " # ; E $7=)!<4 &E!I'1 2
!? ' &D1 #" 2 (+'
& CCC

! " # ; & # # 5 # $ %'1 2

! " # ; = # = " # # 2
4 "# -(4 "# 8 ? # + 2 5# # 11 $ %
- .. ' 0

9 ! " # ; = # = " # # 2
"# + 2 5# # 11 $ % - .. '1 2

! " # ; "% 3% 8%+ 2 %
8% #'1 2

; ! " # ; 8 0 (+
8 ? # 3% + @'1 2

/ ! " # ; 3% 7# 2 # G %
(> # + 1 # '1 2

< ! " # / 8 # # @ 1 D
% % + 2 # '1 2

. ! " # / 8 # # @ 1 D
3% 3% + 2 # $ 9 % '1 2

! " # < 5# # % % 2
? # 2 '1 2

! " # < ? # 2 J 2'1 2

! " # < * ? # $ %

(&$ 2 # #" # 2 '1 2

! "# . 3% $ % 2
8%+ (# 2 # 4 " E >%F E '1 2

9 ! "# . # "# + 3% + 2
2 ' 0

! "# . # "# + 3% + 2
2 '1 2

; ! "# . ? > #
+ ' 0

/ ! "# . ? > #
+ '1 2

< ! "# . # D !1 K # & #
4 # (+J ?'1 2
!? ' CCC

. ! "# .) # 4 # D !1 K #
, & # 4 # (+ $? 8 ? % #" -7 - '1 2

9 ! "# . 4)' # D 1 K #
& # # " + $? - '1 2

9 ! "# .) # 4 # D !1 K #
, & # 4 # (+ $? 8 ? % #" - < '1 2

9 ! "# . 4 22 # # 5
11 # # &#" # #"' 0

9 ! "# . 4 22 # # 5
11 # # &#" # #"'1 2

99 ! "# . & # + # 1 D
+ 4 E > # $ % '1 2

9 ! "# . 5 " # #" 6 ? 1 +
% 2 # 2 2 #'1 2

9; ! "# . 5 " # #" # ? 1 + %
2 # 2 2 # 1 # '1 2

9/ ! "# . % # 1%+ #
1%+ 2 % # # + # '1 2

9< ! "# . I + " ? E >
& # '1 2

9. ! "# . $ ##

11 # !23% ? 4 22 # '1 2

! " # . 8 E >'1 2

! " # . # + & #
4 E # - %' '1 2

! " # . 1%+ #
"+'1 2

! " # . 2 ? #'1 2

9 ! " # . 3% 2 '1 2

! " # . 3% 2 J '1 2

; ! " # 3% (+ 2 1 #
5# # 1 # (# + 1 #" " 6 @ A > - '1 2

/ ! " # (" $ # 3 # 2 #
(+ # 5# # + + 1 #" - ..; '1 2

< ! " # (" $ # 3 # 2 #
(+ # 5# # + + 1 #" - ..; ' 0

. ! " # 8 >)' = + #
5# # " * @ % 8 ? % #" - '1 2

; ! " # # E' $ K 5# #
" 4 "#: 3% & # $ (@ = - . '1 2

; ! " # I # ' #
$ % # - 11 # 1 # $ % # 8 - '1 2

; ! " # 9 L % * ? 8 D #
5# # # $ % ? " 7# + 8 - '1 2

; ! " # $ 4 " (+ 2
&#0 + # # % #" $ 4!7(533& - .. '1 2

;9 ! " # $ 4 " # $ #
2 (+ 2 &#0 + # # % #" : 6 @ & # $ 4!7(533& - .. '1 2

; ! " # (+ = "% # $ 22 #
$ 4 " - '1 2

;; ! " # % (' E @# #
$ % 8 @ % 4 $ % # 4 # + $ 4 "
- '1 2

8M"' <- H 8 # '

;/ ! " # ; < 9/ / . '1 2
!? ' 3% % 1 2 1 '

;< ! " # ; + +: % 1 # 1 ' #
5# # $ % % "% (+ - ;: I + " '1 2

;. ! " # ; 3% % 1 2 ## 1 J #"
% + # % " + - : E E > '1 2
8M"' ; -<

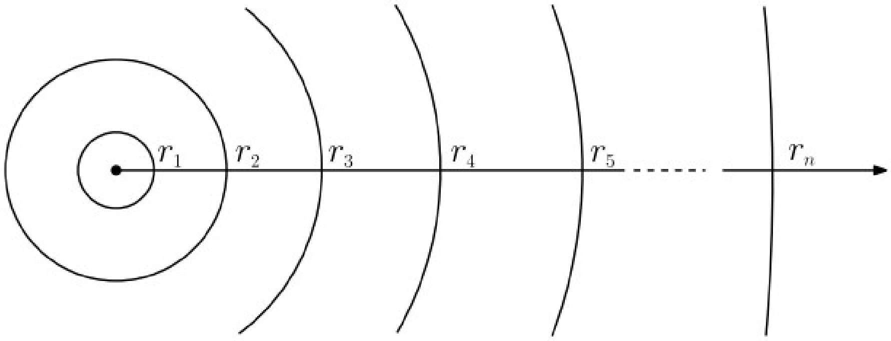
r_1
r_2
r_3
r_4
r_5
r_n

' N3% % 1% : % % 1 + O ' # +: %
2 1% # #P# +: K '

/ ! " # ; 3% % 1 2 1 - : $
4 >> ' 0

/ ! " # ; 3% % 1 2 1 - : $
4 >> '1 2

/ ! " # ; 3% % 1 2 1 ?+) 22 + '
* > '1 2

/ ! " # / 2 5# ? " - %' 8
" + # 11 # '1 2
8M"' .;- /
/' ' 1 # # % D+ 8 #

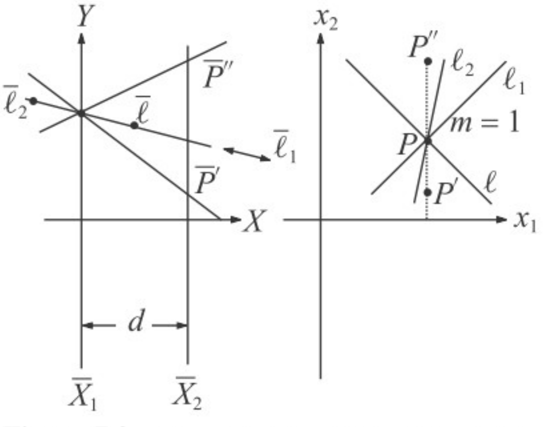
Y
$\bar{P}''$
$\bar{\ell}_2$
$\bar{\ell}$
$\bar{\ell}_1$
$\bar{P}'$
X
d
$\bar{X}_1$
$\bar{X}_2$
x_2
P''
ℓ_2
ℓ_1
$m = 1$
P
P'
ℓ
x_1

Figure 7.2. The "above" and "below" relations between points and lines switch at $m = 1$.

/9 ! " # / 2 5# ? " - %' 8
$ # # (+ # 5 11 # 1 #" '1 2

/ ! " # / # + (+ ,
'1 2

/; ! " # / # + (+'1 2

// ! " # / (+: + : #
8%+ $ 8 - '1 2
!? ' (+: + : # 8%+

/< ! " # / $ Q4 # 8 -R # 6 :
: # ' (+ = # # E '1 2
!? ' CCC

/. ! " # / $ 4 # 8+ % " S " +
$ % # 3 # (5 !D2 7# + 8 : 7 - '1 2

< ! " # / $ % ! # 4 # +
4 # # # + (+'1 2
!? ' CCC

< ! " # / % 1 :J7IJ
J) #J& #? '" 1 ?

< ! " # < 3 " # + 2 % '1 2

< ! " # . 6 % S 3% # + +
$485 (- '1 2

<9 ! " # 3% + 1
(% @ 2 % 7# '1 2
8M"' 9-

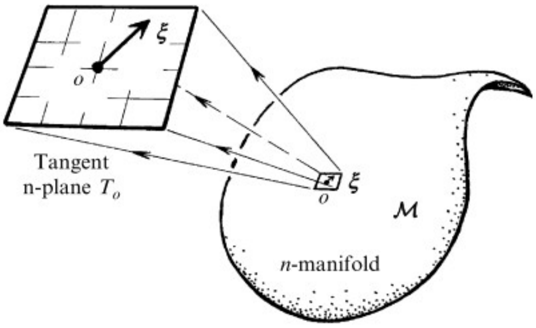

Fig. 12.6 The tangent space T_o, to an n-manifold $\mathcal{M}$ at a point o may be intuitively understood as the limiting space, when smaller and smaller neighbourhoods of o in $\mathcal{M}$ are examined at correspondingly greater and greater magnifications. (Compare Fig. 10.6.) The resulting T_o is flat: an n-dimensional vector space.

8M"' .-

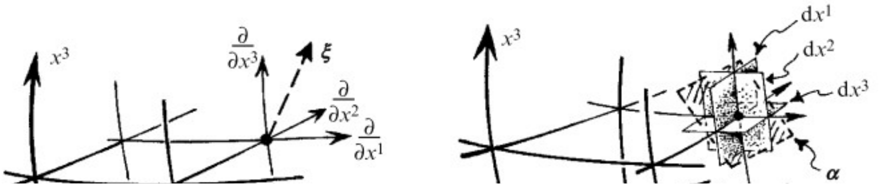

x^3
$\frac{\partial}{\partial x^3}$
$\boldsymbol{\xi}$
$\frac{\partial}{\partial x^2}$
$\frac{\partial}{\partial x^1}$
x^3
dx^1
dx^2
dx^3
$\boldsymbol{\alpha}$

Fig. 12.9 Components in a coordinate patch $(x^1, \ldots, x^n)$ (with $n = 3$ here). (a) For a vector (field) $\boldsymbol{\xi}$, these are the coefficients $(\xi^1, \xi^2, \ldots, \xi^n)$ in $\boldsymbol{\xi} = \xi^1\partial/\partial x^1 + \xi^2\partial/\partial x^2 + \cdots + \xi^n\partial/\partial x^n$, where '$\partial/\partial x^r$' stands for 'rate of change along the rth coordinate axis' (see also Fig. 10.9). (b) For a covector (field) $\boldsymbol{\alpha}$, these are the coefficients $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ in $\boldsymbol{\alpha} = \alpha_1 dx^1 + \alpha_2 dx^2 + \cdots + \alpha_n dx^n$, where dx^r stands for 'the gradient of x^r', and refers to the $(n-1)$-plane element spanned by the coordinate axes except for the x^r-axis.

!? ' CCC

8M"' 9/- /; H + + " 1
' (1 2 # 2 #

+ + " 1
' (1 2 # 2 #
1 % + % 2 # # 1 # # #
1%+ ' *%+ % T 5 "% ? % "% % 1 + D + +
% #" % # + D 1 # +: 1 % 1 0
# # # 11 D #' % "% + ?0 : %
1% : % 1 D # # K -U8 # OV
W ' % : #+ 1%+ K # 2 % % #"
@ # + ? " + > # 2 11 D 1
# # 2 % 8 # : % 2 1 #" # + +
% # ? " D ' A : > ? +: #" % % "% +
2 1%+ % 2 % % # +: 1%+ # #
- # #" " + # # @ % # @% %: + 1 > #":
1 # + 1 # # 1%+ 1 #" + +
% : 2 # # 2 1 #: # # + D C
*% % T 5 # 1 % % ? ># @# U" "
#O' 3% # : # : # + ' E %
1 # # : # ? #" * # ? -U @ O # # 2 #
% 11 " # # # # 2 - # %
" # % # 0 % N % 1 2 N@% % ?0 D
#: ? # % 1 ' 3% @ ? %
?0 2 % @ % 1 2 @ #" % # V ? 1 :
@ * D1 % ? # # 2 + + " 1' 3% # #
% #+ % 1 # 2 11 # # 1%+ : % +: #
+ " 1%+: # @ % # #+ # 2 % 2'
> 1 D 1 ' *% % + 2 T 3%
% @ # # @ 1 # #" 1 # @% % # @ @
+ @% % % # # 2 % - ' ' 2 @% %
% # ' * # % # @% % %
# # #" + # @ ' 3% # % O +
" # 2 #" # % "% "% #" # %
O 1 # : 1 # ? 2 ' G # # # : @ #
1 # % # # 2 1 D # ? : @ #

8M"' 9<- //

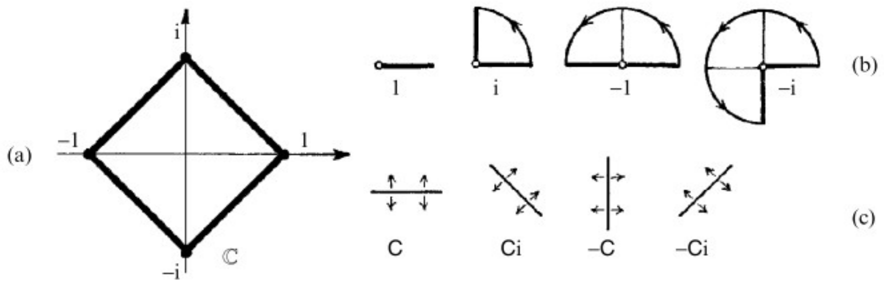

Fig. 13.1 Symmetry of a square. (a) We may represent the square's vertices by the points 1, i, − 1, − i in the complex plane ℂ. (b) The group of non-reflective symmetries are represented, in ℂ, as multiplication by $1 = i^0$, $i = i^1$, $-1 = i^2$, $-i = i^3$, respectively. (c) The reflective symmetries are given, in C, by C (complex conjugation), Ci, − C, and − Ci.

8M"' - <9 H G "' '9

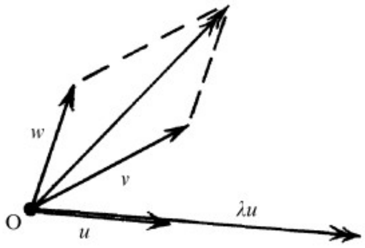

G "' '9 # # 2 # 1
% 1 2 % 1 #
@% % ' 3% 2 # ?+ %
1 # 2 # - ?+ % 1
" @ # 1 # ?+
-@% % ? # ? : # %
2 1 D 1 : 1 D # ? '
% # 2 # 1 % U "%
# O 2 # # % # # 2 U1 O: > 1
#" % " # ! 2 D '

< ! " # 3% G ? 2 % 1 :
3 : # % 3 D 2 +'1 2
8M"' .- ; !2 # @2 > # 1
A $ $ &3 A 6 4 3 = & & ! 7 3 5 ! 6 ! G 3 = & ! $!

<; ! " #) E "" ! " # % # 2
+ 2 # !D2 7# + 8 - '1 2
8M"' - H G "

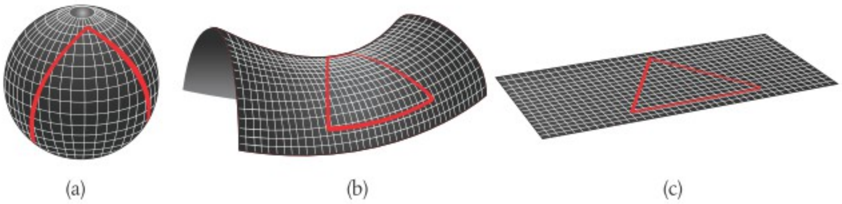

G " 3% 22 # # % G ## # 2 # # & # #O
" # 2 # 1 22 # + 1 ' 3% U O # %
1 + 1 % @# # - ' 3% U 1 #O # % % 1 # " +
1 : -? ' # @ % 2 # + # ? # ? @ # % 2 D1 # # #
% # 2 # "+ # # @ D% ? 2 1 : - '

1

4 3 6 E > 5 1 # 3% +
1 (% @ 2 % 7# - 84G4 '1 2
8M"' - /. H G " '
!? '
! " # 3% + 1 (
% @ 2 % 7# '1 2

G "' ' # + + 2 1% ' 3% # + + " 1: !- :
2 : # #" 2 @ 1 ' 3% 1 # # # # #" %
+ # % -# ?" 1 !- 2 # # 2 + 2
% 1% ' 3% # #" 1 # # % # 2 2 2 + '

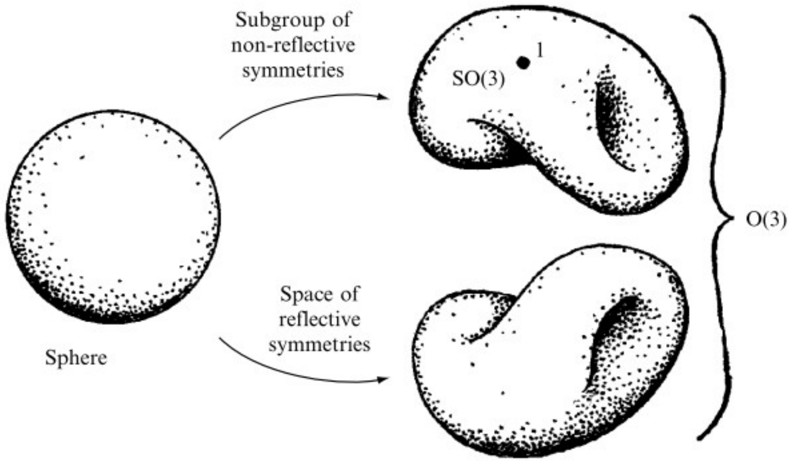
Subgroup of
non-reflective
symmetries
SO(3)
1
O(3)
Space of
reflective
symmetries
Sphere

G "' ' # + + 2 1% ' 3% # + + " 1: !- :
2 : # #" 2 @ 1 ' 3% 1 # # # # #" %
+ # % -# ?" 1 !- 2 # # 2 + 2
% 1% ' 3% # #" 1 # # % # 2 2 2 + '

8M"' .9- 9' 8 # 1

Recall from §10.3 and §12.3 that in the case of a *scalar* field Φ on a general smooth n-manifold $\mathcal{M}$, we were indeed able to provide an appropriate measure of its 'rate of change', namely the 1-form $\mathrm{d}\Phi$, where $\mathrm{d}\Phi = 0$ is the condition that Φ be constant (throughout connected regions of $\mathcal{M}$). However, this idea will not work for a general tensor quantity. It will not even work for a vector field $\boldsymbol{\xi}$. Why is this? One trouble is that in a general manifold we have no appropriate notion of $\boldsymbol{\xi}$ being constant (as we shall see in a moment), whereas any self-respecting differentiation ('gradient') operation that applies to $\boldsymbol{\xi}$ ought to have the property that its vanishing signals the constancy of $\boldsymbol{\xi}$ (as, indeed, $\mathrm{d}\Phi = 0$ signals the constancy of a scalar field Φ). More generally, we would expect that for a 'non-constant' $\boldsymbol{\xi}$, such a derivative operation ought to be measuring $\boldsymbol{\xi}$'s deviation from constancy.

Why is there a problem with this notion of vector 'constancy', on a general n-manifold $\mathcal{M}$? A constant vector field $\boldsymbol{\xi}$, in ordinary Euclidean space, should have the property that all the 'arrows' of its geometrical description are parallel to each other. Thus, some kind of notion of 'parallelism' would have to be part of $\mathcal{M}$'s structure. One might worry about this, bearing in mind the issue of Euclid's fifth postulate—the parallel postulate—that was central to the discussion of Chapter 2. Hyperbolic geometry, for example, does not admit vector fields that could unambiguously considered to be everywhere 'parallel'. In any case, a notion of 'parallelism' is not something that $\mathcal{M}$ would possess merely by virtue of its being a smooth manifold. In Fig. 14.1, the difficulty is illustrated in the case of a 2-manifold pieced together from two patches of Euclidean plane. The normal Euclidean notion of 'parallel' is not consistent from one patch to the next.

In order to gain some insights as to what kind of notion of parallelism is appropriate, it will be helpful for us first to examine the intrinsic geometry of an ordinary 2-dimensional sphere S^2. Let us choose a particular point p on S^2 (say, at the north pole, for definiteness) and a particular tangent vector $\boldsymbol{v}$

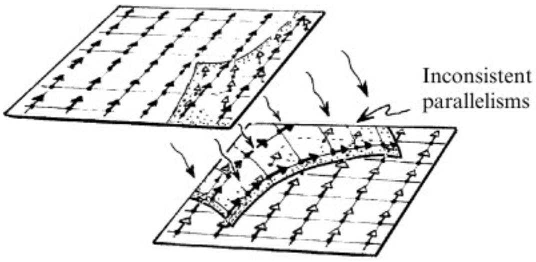

G ''' 9' 3% & ## # 2 U1 O > + ? ## # # % 1
? @ # # 1 % '

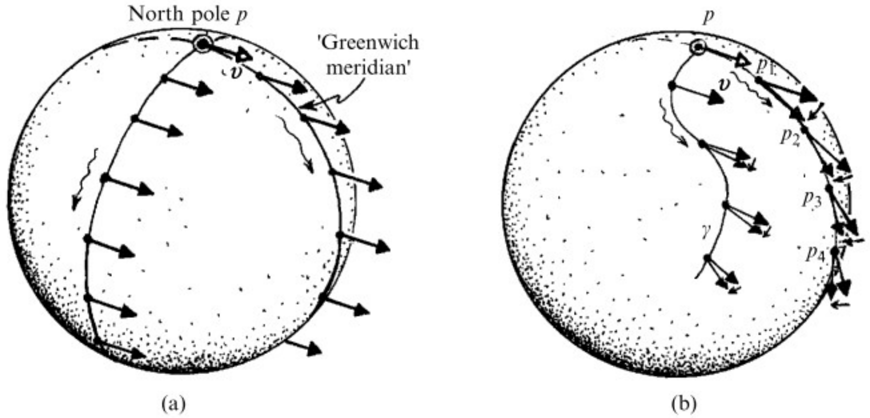

Fig. 14.2 Parallelism on the sphere S^2. Choose p at the north pole, with tangent vector $\boldsymbol{v}$ pointing along the Greenwich meridian. Which tangent vectors, at other points of S^2, are we to regard to being 'parallel' to $\boldsymbol{v}$? (a) The direct Euclidean notion of 'parallel', from the embedding of S^2 in $\mathbb{E}^3$, does not work because (except along the meridian perpendicular to the Greenwich meridian) the parallel $\boldsymbol{v}$s do not remain tangent to S^2. (b) Remedy this, moving $\boldsymbol{v}$ parallel along a given curve γ, by continually projecting back to tangency with the sphere. (Think of γ as made up of large number of tiny segments $p_0p_1, p_1p_2, p_2p_3, \ldots,$ projecting back at each stage. Then take the limit as the segments are made smaller and smaller.) This notion of *parallel transport* is indicated for the Greenwich meridian, but also for a general curve γ.

8M"' ./ - ;

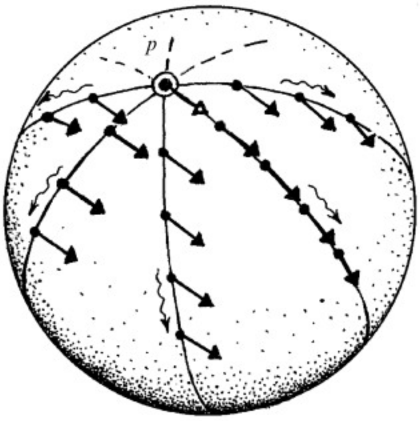

Fig. 14.3 Parallel transport of $\boldsymbol{v}$ along five different paths (all great circles).

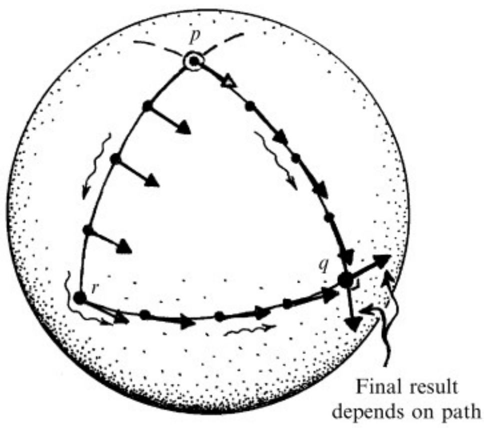

Fig. 14.4 Path dependence of parallel transport. This is illustrated using two distinct paths from p to q, one of which is a direct great-circle route, the other consisting of a pair of great-circle arcs jointed at an intermediate point r. Parallel transport along these two paths gives results at q differing by a right-angle rotation.

>
8M"' /.- <

How 'different' are the various groups $\mathrm{O}(p, q)$, for $p+q=n$, for fixed n? (The positive-definite and Lorentzian cases are contrasted, for $n=2$ and $n=3$, in Fig. 13.18.) They are closely related, all having the same dimension $\frac{1}{2}n(n-1)$; they are what are called *real forms* of one and the same complex group $\mathrm{O}(n, \mathbb{C})$, the *complexification* of $\mathrm{O}(n)$. This complex group is defined in the same way as $\mathrm{O}(n)$ $(=\mathrm{O}(n, \mathbb{R}))$, but where the linear transformations are allowed to be *complex*. Indeed, although I have phrased my considerations in this chapter in terms of real linear transformations, there is a parallel discussion where 'complex' replaces 'real' throughout. (Thus the coordinates x^a become complex and so do the components of our matrices.) The only essential difference, in what has been said above, arises with the concept of *signature*. There are complex linear coordinate transformations that can convert a -1 in a diagonal realization of g_{ab} into a $+1$ and *vice versa*,[13.52] so we do not now have a

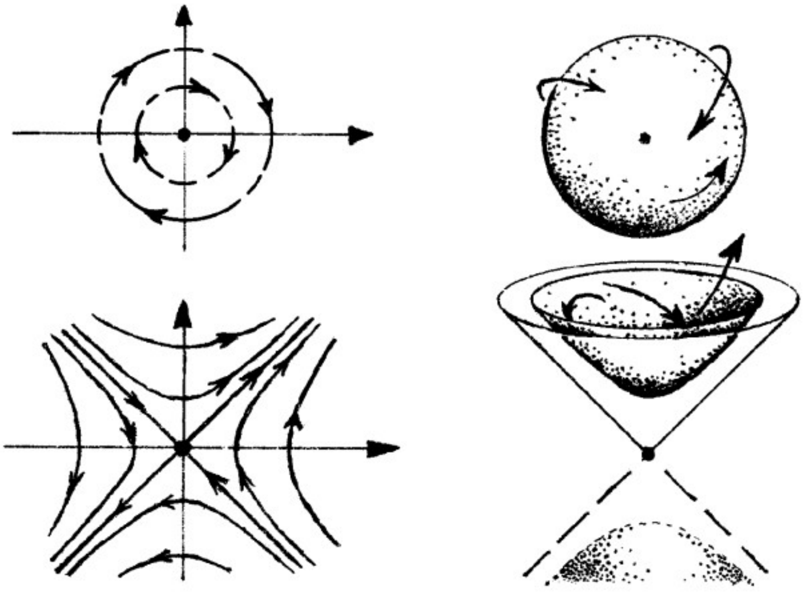

(a) (b)

meaningful notion of signature. The only invariant[20] of ***g***, in the complex case, is what is called its *rank*, which is the number of non-zero terms in its diagonal realization. For a non-singular ***g***, the rank has to be maximal, i.e. n.

When is the difference between these various real forms important and when is it not? This can be a delicate question, but physicists are often rather cavalier about the distinctions, even though these can be important. The positive-definite case has the virtue that the group is compact, and much of the mathematics is easier for such situations (see §13.7). Sometimes people blithely carry over results from the compact case to the non-compact cases ($p \neq 0 \neq q$), but this is often not justified. (For example, in the compact case, one need only be concerned with representations that are finite-dimensional, but in the non-compact case additional infinite-dimensional representations arise.) On the other hand, there are other situations in which considerable insights can be obtained by ignoring the distinctions. (We may compare this with Lambert's discovery of the formula, in terms of angles, of the area of a hyperbolic triangle, given in §2.4. He obtained his formula by allowing his sphere to have an imaginary radius. This is similar to a signature change, which amounts to allowing some coordinates to have imaginary values. In §18.4, Fig. 18.9, I shall try meaningful notion of signature. The only invariant[20] of $\boldsymbol{g}$, in the complex case, is what is called its *rank*, which is the number of non-zero terms in its diagonal realization. For a non-singular $\boldsymbol{g}$, the rank has to be maximal, i.e. n.

When is the difference between these various real forms important and when is it not? This can be a delicate question, but physicists are often rather cavalier about the distinctions, even though these can be important. The positive-definite case has the virtue that the group is compact, and much of the mathematics is easier for such situations (see §13.7). Sometimes people blithely carry over results from the compact case to the non-compact cases ($p \neq 0 \neq q$), but this is often not justified. (For example, in the compact case, one need only be concerned with representations that are finite-dimensional, but in the non-compact case additional infinite-dimensional representations arise.) On the other hand, there are other situations in which considerable insights can be obtained by ignoring the distinctions. (We may compare this with Lambert's discovery of the formula, in terms of angles, of the area of a hyperbolic triangle, given in §2.4. He obtained his formula by allowing his sphere to have an imaginary radius. This is similar to a signature change, which amounts to allowing some coordinates to have imaginary values. In §18.4, Fig. 18.9, I shall try

8M"'   -   9' (: 1   "   : #

First, in order to build ourselves a parallelogram, let us consider the concept of a *geodesic*, as defined by the connection $\boldsymbol{\nabla}$. Geodesics are important to us for other reasons. They are the analogues of the straight lines of Euclidean geometry. In our example of the sphere S^2, considered above (Figs. 14.2–14.4), the geodesics are great circles on the sphere. More generally, for a curved surface in Euclidean space, the curves of minimum length (as would be taken up by a string stretched taut along the surface) are geodesics. We shall be seeing later (§17.9) that geodesics have a fundamental significance for Einstein's general relativity, representing the paths in spacetime that describe freely falling bodies. How does our

</   ! " #   9   3%   + 2 4 22   #   &   #   #
&#" #   #" # $ % #   '1 2

8M"' / -9   ;' (&!$&3 5 563& 8 &335!6 !G 56&   84&

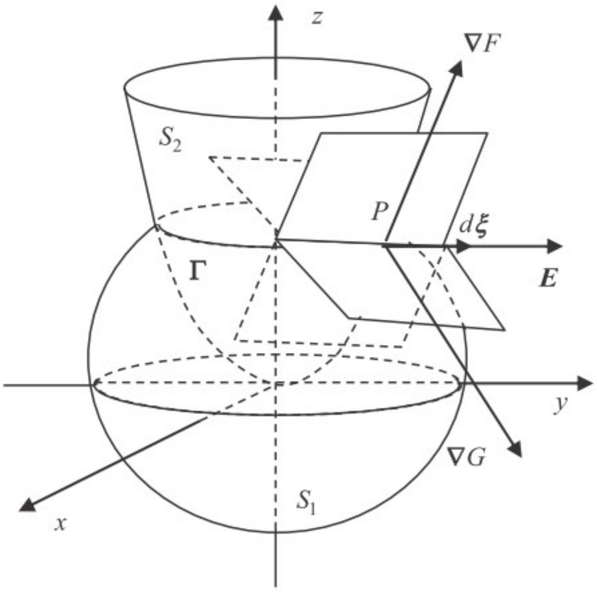

G " ;'9 (# 1 # 2 % # 2 84&

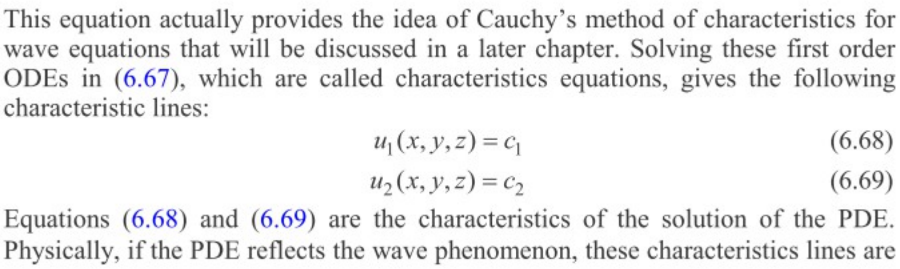

This equation actually provides the idea of Cauchy's method of characteristics for wave equations that will be discussed in a later chapter. Solving these first order ODEs in (6.67), which are called characteristics equations, gives the following characteristic lines:

$$u_1(x,y,z)=c_1 \tag{6.68}$$

$$u_2(x,y,z)=c_2 \tag{6.69}$$

Equations (6.68) and (6.69) are the characteristics of the solution of the PDE. Physically, if the PDE reflects the wave phenomenon, these characteristics lines are

! % 2 #

% 1 @@@' % ># ' " 1+ % "

% 1 @@@'2 +' ?' " +JK ? 8% X < ? 8% 1%+X 2
X $ " # 6 1 # 8+ % " # D' % '%

<< ! " # 8+ % " # 3% # #+
1 2 '1 2

<. ! " # 8+ % " # # @ #'"1 2

. ! " # 1+ % " # @'1 2

. ! " # ; 8 "
Y ? Z: *=' G # - J '1 2

. ! " # ; 3% E % 2 + 1 #
8 # (+ G % 2 # = # 2 # * # # E >%F E # - '1 2

[\]^_`abcdaefghijklamhniolfpepfffqpqqpep\arjstsuvnapqwfffbuxnsysy^zxyskh^{l^|vjsvznl^
1 0 3 - ..; '1 2
8M'" -

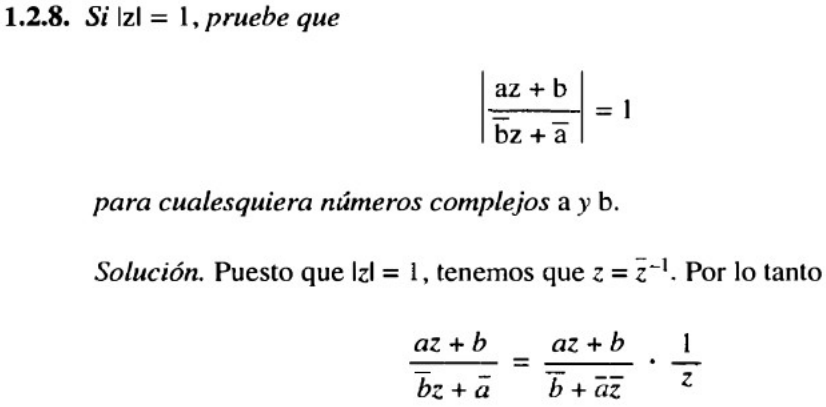

1.2.8. *Si* |z| = 1, *pruebe que*

$$\left|\frac{az+b}{\bar{b}z+\bar{a}}\right| = 1$$

para cualesquiera números complejos a *y* b.

Solución. Puesto que |z| = 1, tenemos que $z = \bar{z}^{-1}$. Por lo tanto

$$\frac{az+b}{\bar{b}z+\bar{a}} = \frac{az+b}{\bar{b}+\bar{a}\bar{z}} \cdot \frac{1}{z}$$

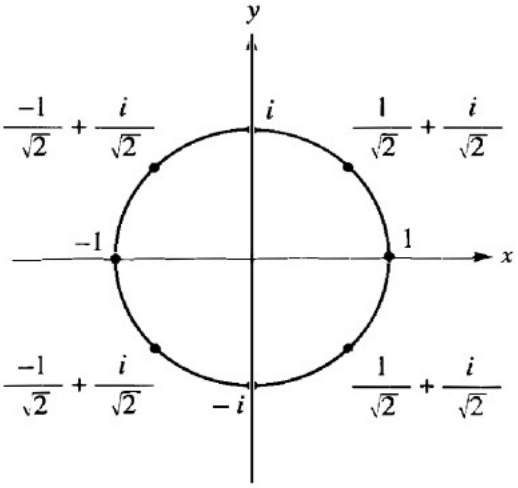
y
$\frac{-1}{\sqrt{2}} + \frac{i}{\sqrt{2}}$
i
$\frac{1}{\sqrt{2}} + \frac{i}{\sqrt{2}}$
−1
1
x
$\frac{-1}{\sqrt{2}} + \frac{i}{\sqrt{2}}$
−i
$\frac{1}{\sqrt{2}} + \frac{i}{\sqrt{2}}$

Figura 1.2.11. Las ocho raíces octavas de la unidad.

Por las propiedades del valor absoluto y puesto que $|z| = 1$,

$$\left|\frac{az+b}{\overline{b}z+\overline{a}}\right| = \left|\frac{az+b}{\overline{a}\,\overline{z}+\overline{b}}\right| \cdot \frac{1}{|z|} = 1$$

ya que $|az + b| = \overline{|az + b|} = |\overline{a}\overline{z} + \overline{b}|$.

8M"' 9 - H 6 1 0

La siguiente discusión presupone cierta familiaridad con las principales propiedades de los números reales. *El sistema de los números reales* fue el resultado de la búsqueda de un sistema (un conjunto abstracto con ciertas reglas) que incluyera a los racionales, pero que también proporcionara soluciones a ecuaciones polinomiales tales como $x^2 - 2 = 0$.

Históricamente, una consideración similar dio origen a la extensión de los números reales. A principios del siglo XVI, Geronimo Cardano consideró ecuaciones cuadráticas (y cúbicas) tales como $x^2 + 2x + 2 = 0$, que no son satisfechas por ningún número real x. La fórmula cuadrática $(-b \pm \sqrt{b^2 - 4ac})/2a$ da expresiones "formales" para las dos soluciones de la ecuación $ax^2 + bx + c = 0$. Pero esta fórmula puede requerir raíces cuadradas de números negativos, por ejemplo, $-1 \pm \sqrt{-1}$ para la ecuación $x^2 + 2x + 2 = 0$. Cardano notó que si estos "números complejos" son tratados como números ordinarios con la regla $\sqrt{-1} \cdot \sqrt{-1} = -1$, estos, resolvían en efecto las ecuaciones. A la importante expresión $\sqrt{-1}$ se le da ahora la ampliamente aceptada designación de $i = \sqrt{-1}$. Una convención alternativa es seguida por muchos ingenieros eléctricos, quienes prefieren el símbolo $j = \sqrt{-1}$, puesto que ellos desean reservar el símbolo i para la corriente eléctrica. Sin embargo, en el pasado se sentía que ningún significado podría realmente ser asignado a tales expresiones, que fueron entonces llamadas "imaginarias". Gradualmente, en especial como resultado del trabajo de Leonhard Euler en el siglo XVIII, estas cantidades imaginarias llegaron a desempeñar un papel importante. Por ejemplo, la fórmula de Euler $e^{i\theta} = \cos \theta + i \operatorname{sen} \theta$ reveló la existencia de una profunda relación entre los números complejos y las funciones trigonométricas. Se encontró que la regla $e^{i(\theta_1 + \theta_2)} = e^{i\theta_1} e^{i\theta_2}$ resumía las reglas para la expansión del seno y del coseno de la suma de dos ángulos de una manera sencilla, y este solo resultado indicó que algún significado debería ser atribuido a estos números "imaginarios".

Sin embargo, no fue hasta los trabajos de Casper Wessel (*circa* 1797), Jean Robert Argand (1806), Karl Friedrich Gauss (1831), Sir William R. Hamilton (1837) y otros, que se clarificó el significado de los números complejos y se comprendió que no hay nada de "imaginario" en ellos (aun cuando este término se emplea todavía.

El análisis complejo, que es el tema de este libro, fue desarrollado en el siglo XIX, principalmente por Augustin Cauchy (1789-1857). Posteriormente, su teoría se hizo más rigurosa y fue extendida por matemáticos como Peter Dirichlet (1805-1859), Karl Weierstrass (1815-1897) y Georg Friedrich Bernhard Riemann (1826-1866).

La búsqueda de un método para describir la conducción del calor influenció el desarrollo de la teoría, la cual encontró muchos usos fuera de las matemáticas. En los capítulos subsecuentes se discutirán algunas de estas aplicaciones a problemas en física e ingeniería, tales como hidrodinámica y electrostática. La teoría tiene también aplicaciones matemáticas en problemas que a primera vista no parecen involucrar números complejos. Por ejemplo, la demostración de que

$$\int_0^\infty \frac{\operatorname{sen}^2 x}{x^2}\,dx = \frac{\pi}{2}$$

o que

$$\int_0^\infty \frac{x^{\alpha-1}}{1+x}\,dx = \frac{\pi}{\operatorname{sen}(\alpha\pi)} \qquad \text{para } 0 < \alpha < 1$$

o que

$$\int_0^{2\pi} \frac{d\theta}{a + \operatorname{sen}\theta} = \frac{2\pi}{\sqrt{a^2 - 1}}$$

.9 ! " # / = #? > 5# #
' - '1 2

. ! " # / = #? > 5# #
' - '1 2

.; ! " # /)' =' = #? > 5# #
% # 3 22 8 ? % #" - '1 2

%

% 1 @@@'1 2 ' 1 "+ # '%

./ ! " # < 3 ? # ' 4 : %)' E #
G # # 2 6 1 # 5 $ + 2 5# , 11 $ %
- / '1 2

.< ! " # < 6 1 # 2 5# #
&D # G @ : J G # # 2 1 # G 4+# E @ %
= # ## - / '1 2

.. ! " # < 6 1 # 2 5# #
&D # G @ : J 1 # % 2 # # 2 @) %# * +
, # J % % : 7I - ..9 '1 2

! " # . " K K'1 2
!? ' !? ' # 2 # H # >+ %

! " # 5'$') = + 2 1 "+
6 % = # - ; ' 0

! " # 5'$') = + 2 3 1 "+
6 % = # - ; '1 2

! " #) 5' $' = # ? > 2
" ? 1 "+ 6 % = # - .. ' 0

9 ! " #) 5' $' = # ? > 2
" ? 1 "+ 6 % = # - .. '1 2

! " # $ $ 9'1 2
!? ' 1 # 2 # " ? 11 % " # 1 "+

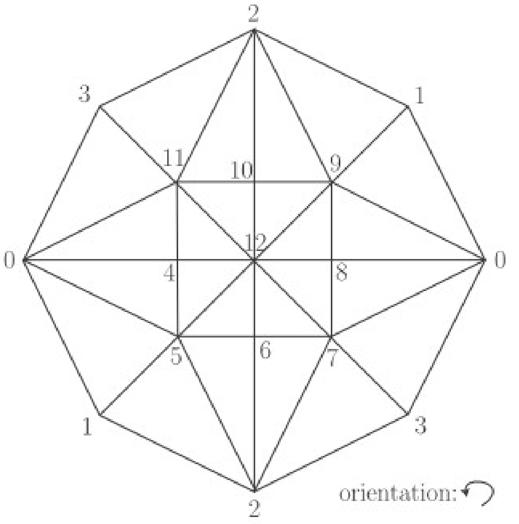
2
3
1
11
10
9
12
0
0
4
8
5
6
7
1
3
2
orientation:

Triangulation of the projective plane

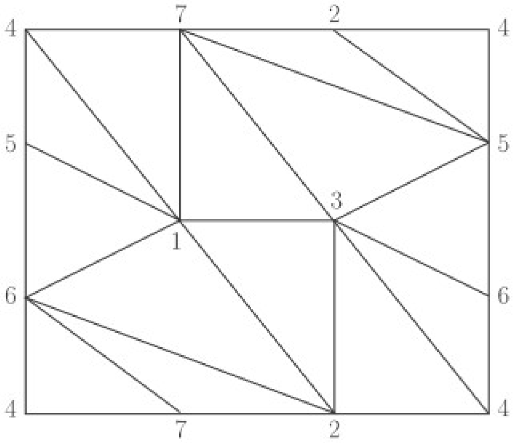
4
7
2
4
5
5
3
1
6
6
4
7
2
4

Triangulation of the Csaszar polynomial

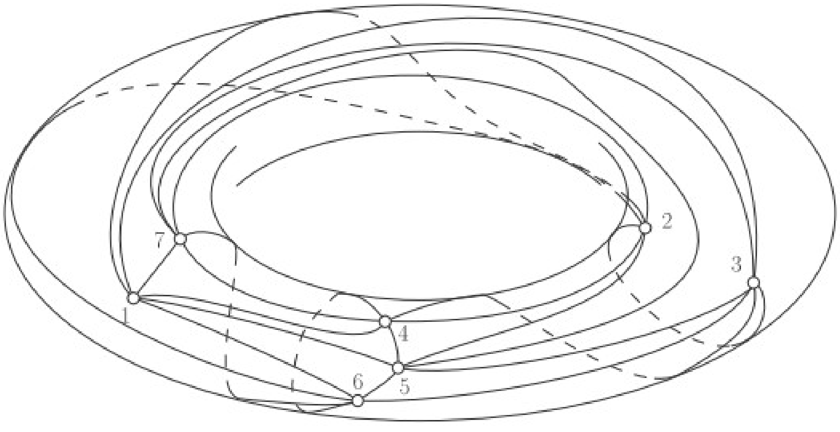

Triangulation of the Csaszar torus

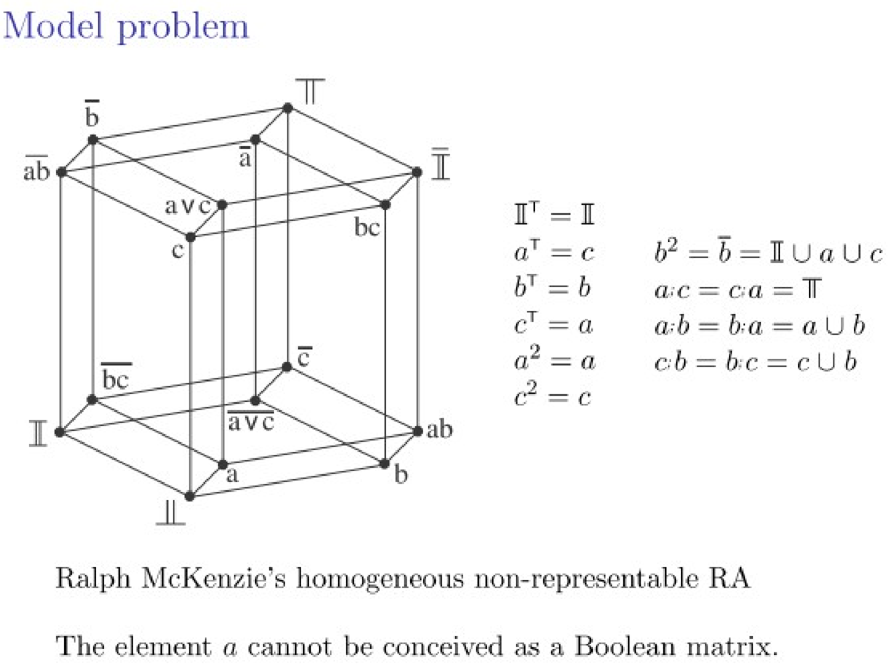

Model problem

$\overline{b}$
$\mathbb{T}$
$\overline{ab}$
$\overline{a}$
$\mathbb{I}$
a∨c
bc
c
$\overline{c}$
$\overline{bc}$
$\overline{a \vee c}$
ab
$\mathbb{I}$
a
b
$\perp\!\!\perp$

$\mathbb{I}^{\intercal} = \mathbb{I}$
$a^{\intercal} = c$
$b^{\intercal} = b$
$c^{\intercal} = a$
$a^2 = a$
$c^2 = c$

$b^2 = \overline{b} = \mathbb{I} \cup a \cup c$
$a{;}c = c{;}a = \mathbb{T}$
$a{;}b = b{;}a = a \cup b$
$c{;}b = b{;}c = c \cup b$

Ralph McKenzie's homogeneous non-representable RA

The element a cannot be conceived as a Boolean matrix.

8M"' <

; ! "# 3 1 "+'1 2

/ ! "# $ % G 8%+ #
5 = # ? >J# '1 2
8M"' -/ H %1 9 H 3 1 " 1

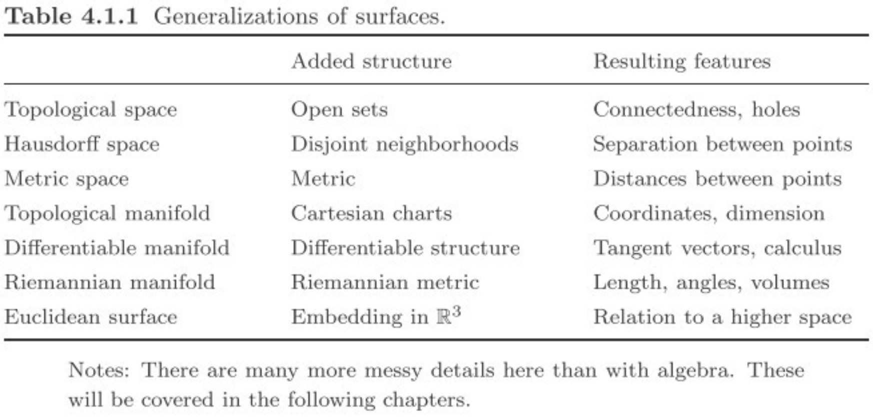

Table 4.1.1 Generalizations of surfaces.

	Added structure	Resulting features
Topological space	Open sets	Connectedness, holes
Hausdorff space	Disjoint neighborhoods	Separation between points
Metric space	Metric	Distances between points
Topological manifold	Cartesian charts	Coordinates, dimension
Differentiable manifold	Differentiable structure	Tangent vectors, calculus
Riemannian manifold	Riemannian metric	Length, angles, volumes
Euclidean surface	Embedding in $\mathbb{R}^3$	Relation to a higher space

Notes: There are many more messy details here than with algebra. These will be covered in the following chapters.

8M"' - /

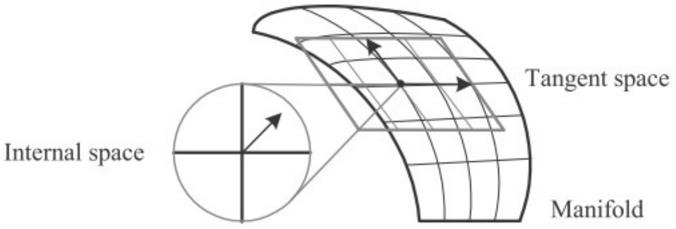

Tangent space
Internal space
Manifold

Figure 4.1.2 Generalizing tangents on a manifold.

8M"' -/

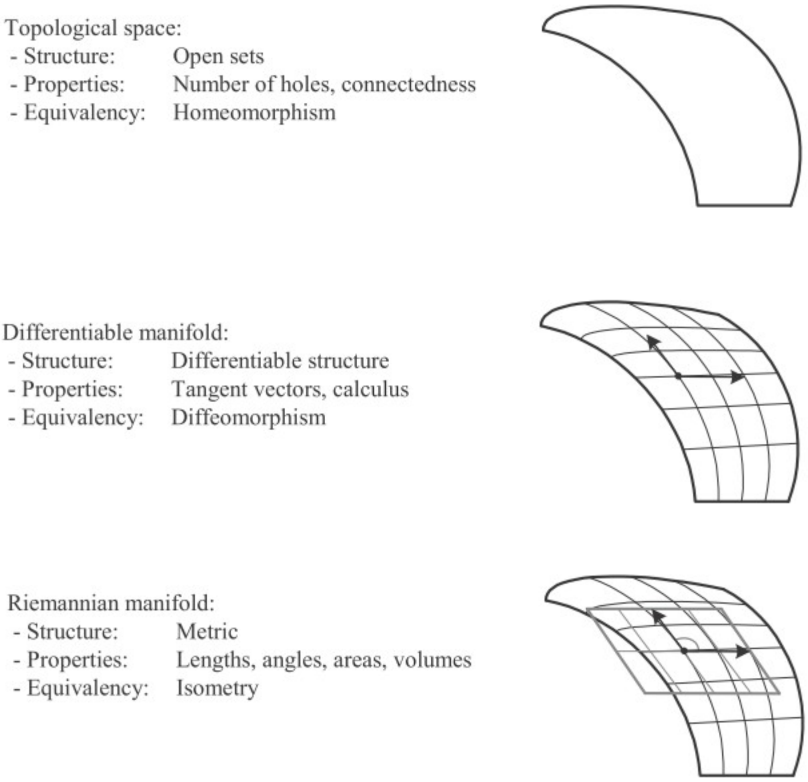

Topological space:
- Structure: Open sets
- Properties: Number of holes, connectedness
- Equivalency: Homeomorphism

Differentiable manifold:
- Structure: Differentiable structure
- Properties: Tangent vectors, calculus
- Equivalency: Diffeomorphism

Riemannian manifold:
- Structure: Metric
- Properties: Lengths, angles, areas, volumes
- Equivalency: Isometry

8M"' ; - < H % 1 . H ## # # 2

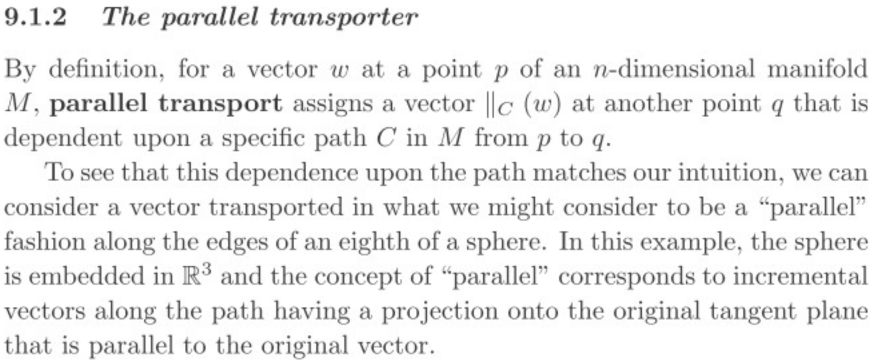

9.1.2 *The parallel transporter*

By definition, for a vector w at a point p of an n-dimensional manifold M, **parallel transport** assigns a vector $\|_C(w)$ at another point q that is dependent upon a specific path C in M from p to q.

To see that this dependence upon the path matches our intuition, we can consider a vector transported in what we might consider to be a "parallel" fashion along the edges of an eighth of a sphere. In this example, the sphere is embedded in $\mathbb{R}^3$ and the concept of "parallel" corresponds to incremental vectors along the path having a projection onto the original tangent plane that is parallel to the original vector.

8M"' ; - <

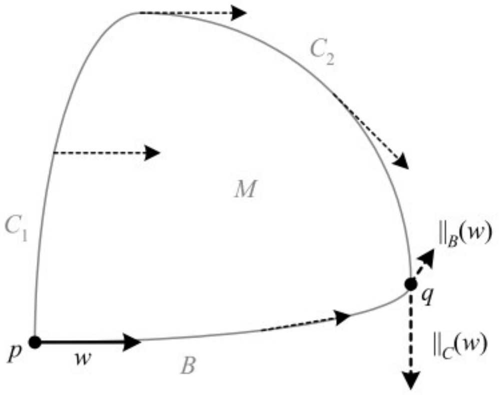

Figure 9.1.1 A vector w transported in what we intuitively consider to be a "parallel" way along two different paths (B and $C = C_1 + C_2$) on a surface results in two different vectors.

8M"' ;/- <;

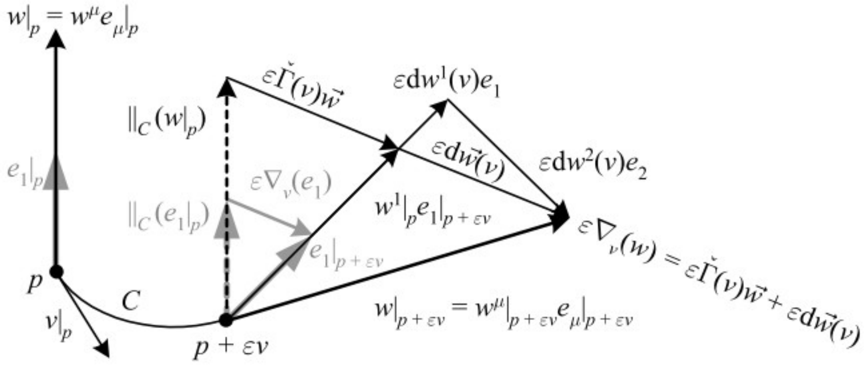
$w|_p = w^\mu e_\mu|_p$
$\varepsilon\check{\Gamma}(v)\vec{w}$
$\varepsilon \mathrm{d}w^1(v)e_1$
$\|_C(w|_p)$
$\varepsilon \mathrm{d}\vec{w}(v)$
$\varepsilon \mathrm{d}w^2(v)e_2$
$e_1|_p$
$\varepsilon\nabla_v(e_1)$
$\|_C(e_1|_p)$
$w^1|_p e_1|_{p+\varepsilon v}$
$\varepsilon\nabla_v(w) = \varepsilon\check{\Gamma}(v)\vec{w} + \varepsilon \mathrm{d}\vec{w}(v)$
$e_1|_{p+\varepsilon v}$
p
C
$v|_p$
$p + \varepsilon v$
$w|_{p+\varepsilon v} = w^\mu|_{p+\varepsilon v} e_\mu|_{p+\varepsilon v}$

Figure 9.1.3 Relationships between the frame, parallel transport, covariant derivative, and connection for a vector w parallel to e_1 at a point p.

8M"' / - .

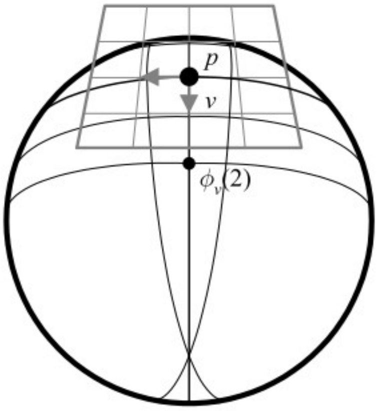

Figure 9.1.4 Geodesic normal coordinates at p map points on a manifold to vectors at p tangent to the geodesic passing through both points. In the figure $\exp(2v) = \phi_v(2)$, so the coordinate of the point $\phi_v(2) \in M$ is $2v \in T_pM$.

< ! " #
8 2 J # J5 JJ J8 J J # J$ % '1 2
8M"' /- /

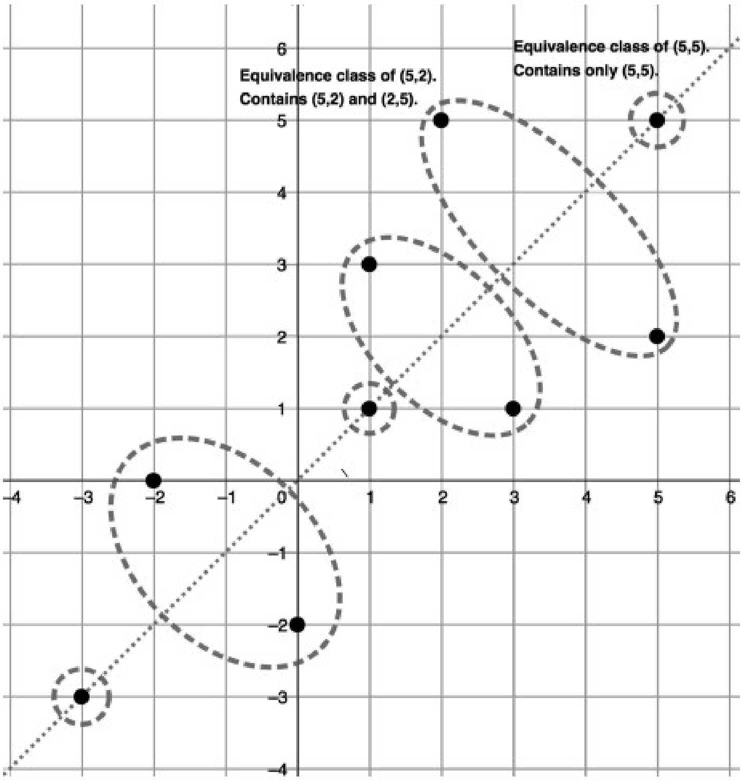
Equivalence class of (5,2).
Contains (5,2) and (2,5).
Equivalence class of (5,5).
Contains only (5,5).
6
5
4
3
2
1
0
−1
−2
−3
−4
−4
−3
−2
−1
1
2
3
4
5
6

8M"' ; - /

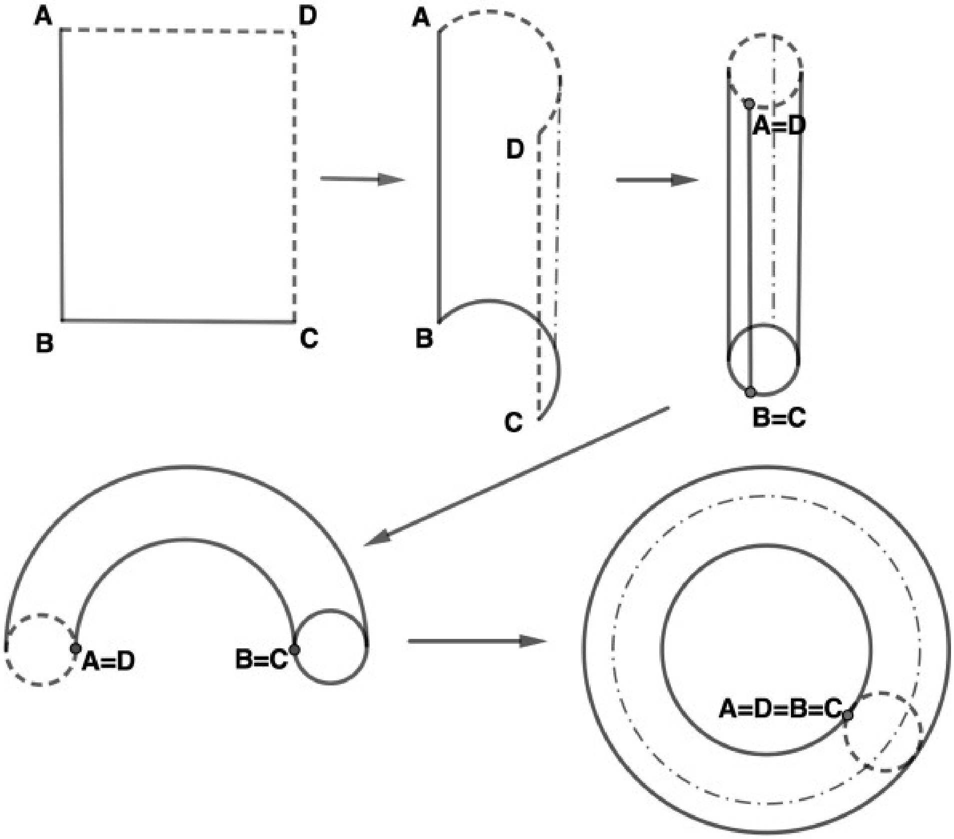
A
D
B
C
A
D
B
C
A=D
B=C
A=D
B=C
A=D=B=C

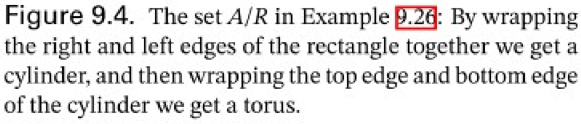

Figure 9.4. The set A/R in Example 9.26: By wrapping the right and left edges of the rectangle together we get a cylinder, and then wrapping the top edge and bottom edge of the cylinder we get a torus.

. ! " # 9 " $ % '1 2

! " # (1 : (
4 22 # & # A 2 1% '1 2

! " # ; 11 #
#J # J # J J8%+ J J % J JG JG J # J # J&#" #
J J J '1 2
8M"' ;-9

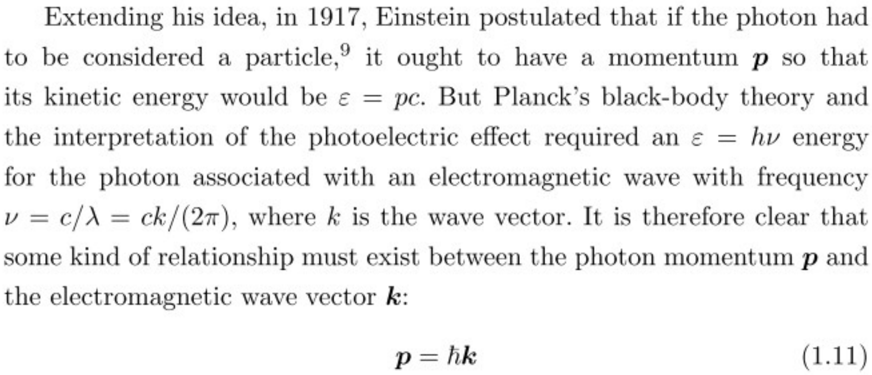

Extending his idea, in 1917, Einstein postulated that if the photon had to be considered a particle,[9] it ought to have a momentum $\boldsymbol{p}$ so that its kinetic energy would be $\varepsilon = pc$. But Planck's black-body theory and the interpretation of the photoelectric effect required an $\varepsilon = h\nu$ energy for the photon associated with an electromagnetic wave with frequency $\nu = c/\lambda = ck/(2\pi)$, where k is the wave vector. It is therefore clear that some kind of relationship must exist between the photon momentum $\boldsymbol{p}$ and the electromagnetic wave vector $\boldsymbol{k}$:

$$\boldsymbol{p} = \hbar \boldsymbol{k} \tag{1.11}$$

! " # ; 11 #
#J # J # J J8%+ J J % J JG JG J # J # J&#" #
J J J J& ? '1 2
8M"' - 9 8 5 $ = # # # 11 D #
? #" +

3% # ? % 2 #" 1 # 1 @ 1 # :
? # : 1 ? # D 1 2 % @
? # D +' = @ % # ? 2
2 1 ? @% % # ? 1 P % # D + 1 '
! % +1 2 # % 2 # ? D1 # : 2 # +: %
1 # 2 # # # # D + ? ' 3% #
+ 2 % # 1%+ @ ? % % # % +1
2 1 # @% %: # # # # : # " + ?
% 1 # # ' G % : # ? 1 1 :
@ " % 11 # + 1 @ 2 : # " # :
11 D # % 2 #" % " 2 1 ? '

! " # / 3% '7#? # '8 % '1 2

9 ! " # < 8 1 ' $ % > 2
5#" # # 6 @ # % 2 ' @ " - . '1 2
8M"' - 9/
5 4 22 # %# #"
4 22 #K ? > # G #> #
' 4 3 #" # #1 ?

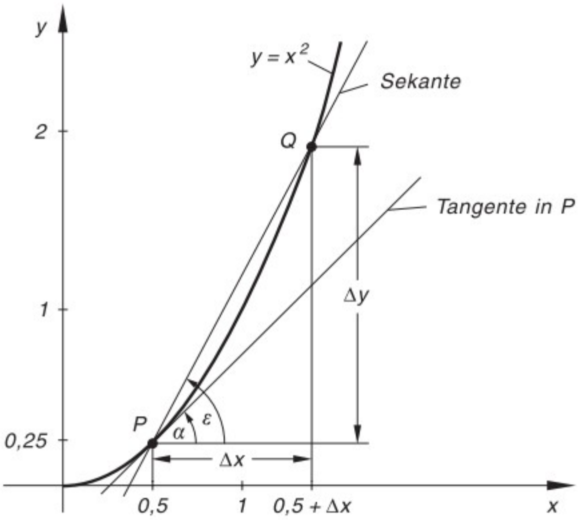
y
y = x²
Sekante
2
Q
Tangente in P
1
Δy
P
ε
α
0,25
Δx
0,5
1
0,5 + Δx
x

Bild IV-1

4 % 8 # 2 # > # ? K ## " #"

$$m_s = \tan \varepsilon = \frac{\Delta y}{\Delta x} = \frac{(0{,}5 + \Delta x)^2 - 0{,}25}{\Delta x} = \frac{0{,}25 + \Delta x + (\Delta x)^2 - 0~25}{\Delta x} =$$

$$= \frac{\Delta x + (\Delta x)^2}{\Delta x} = \frac{\Delta x\,(1 + \Delta x)}{\Delta x} = 1 + \Delta x \qquad \text{(IV-2)}$$

8M"' 9- 9<

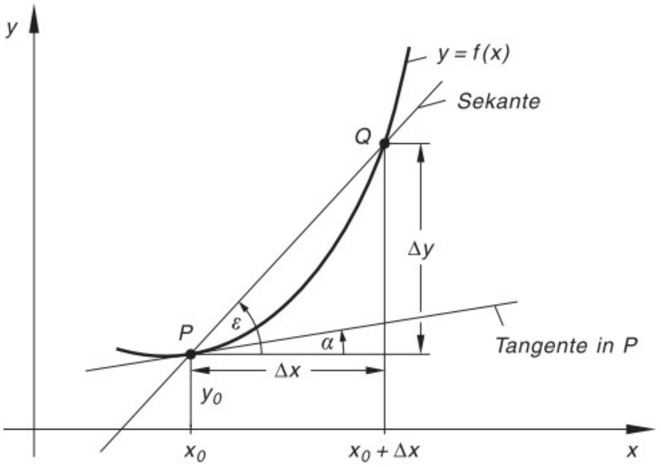
y
y = f(x)
Sekante
Q
Δy
P
ε
α
Tangente in P
Δx
y₀
x₀
x₀ + Δx
x

Bild IV-2
Zum Begriff
der Ableitung
einer Funktion

! " # . ; 0 % % >' ? # '1 2

8M"' //- << < $ % > @F% # * # % 2 % # #

8M"' 9 -9 9

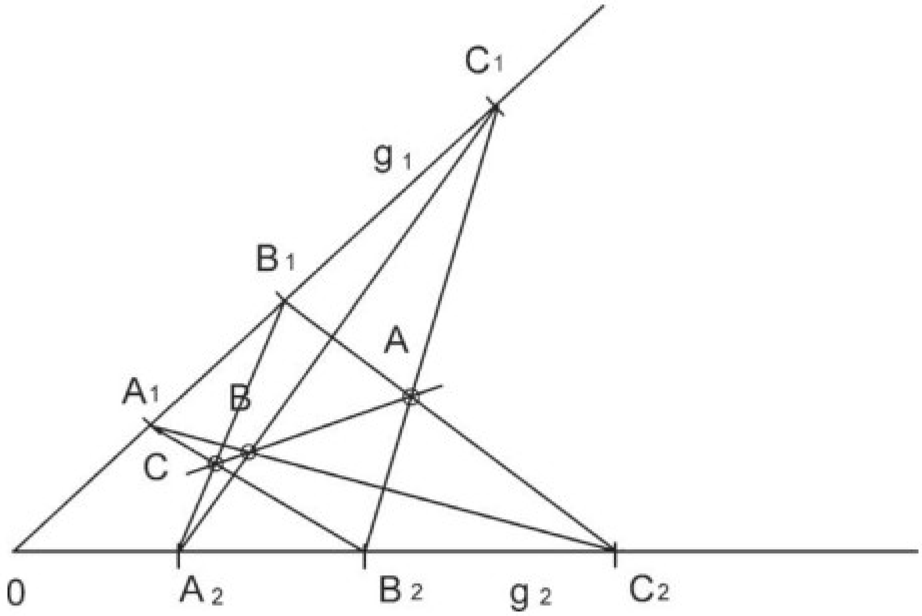
C1
g1
B1
A
A1
B
C
0
A2
B2
g2
C2

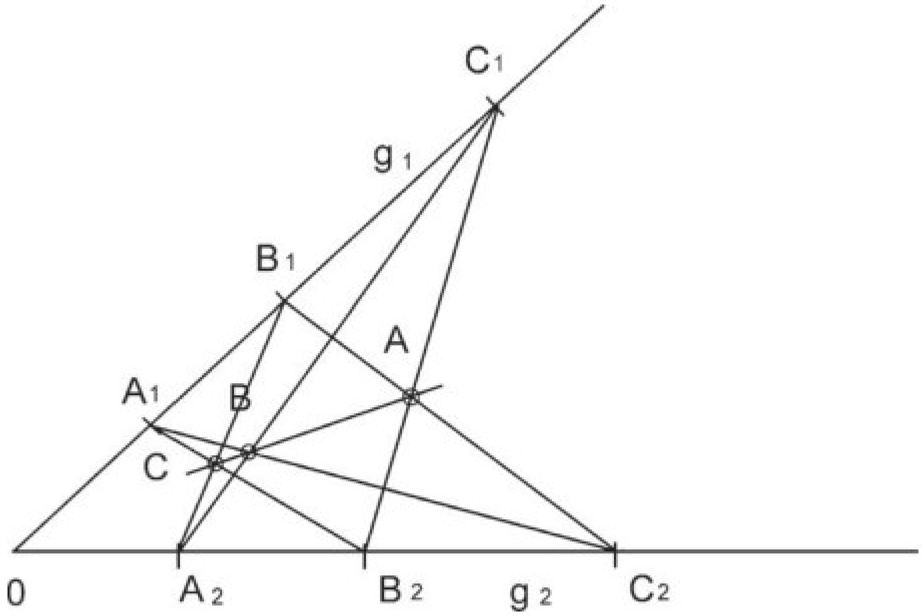
C1
g1
B1
A
A1
B
C
0
A2
B2
g2
C2

??' <'9' G " K K # 8 11 " # & > # # % > ?
@ % # 2 K@ (#: %# # % (" # # # 8 #> # #
(#

<'9 #2F#" 1 0 > # (
$ # % 8 11 ? # ## # K % # > # # K % "
? % : # @ % 1 0 > # (K %# # # " " # %
% K # 8 11 8 ? K %# #'
& # > # # % % % &# @ > #" % # K # K #%F#
" # # 4 K 1 # }1 0 > (~ ? " ## > % 2 @F% #
% ? 1 1 > % # $ % # I•# ' @ 2
E ? - 9 9H 9/ # # 1 # 6• #? " "
> # 3 1 G " # % # " # # "
% # & " # % 2 # 2: K@ # 8 1 > K # # % #
G " # % ? # - 2•% % # Y ? % ? : ' €' : ' 9; 'Z:
Y8 4 % # 4 ..9: ' . 'Z: Y .</Z '

4 "# % G 2•% #" 5 # #> # 2 #K• % #
5#" # # % > # H ? $ % % # > H (
4 " - . H ;; ' & ? # # I " %# % #
" # 1 > K # # I•# # % # 3 %# > # %
% % 4 % #" #" K ? #' # = 1 @ > }E # 1 0
' ' ' ~ -2 •? K & &# @ 2 E % ? #" & "# ? L
I " # &? # % # ; .: ? # #
€ " # • >' 4 " &D 1 # $ % > V
#K @ H ' ' @ " # #" @•%# % # 3 # " H " #"'
4 * > @ #" % #V . @ # ! " # D 1 @
> '
2 4 " " % # > # > E " ? #" # } #
% 2 # 8 #> ~: }% # % 3 #"~ # }4 11 %F # ~ K
• >V K E " @ .') % % # % 'G' $•?
G #" # $ % > #" 2•% ' 2 4
" " % % 3% 8 # # 8 #> K • > - " ' 2•%
% Y4% ? > % ..9Z: Y8 4 % # 4 ..9: ' 'Z:
YG (+ .</Z '

; ! " # 9 $ % > % # #
#@ # # # # (# " # ? K G % # # 1 3 # 2 # 1 >
> % " - '1 2
8M"' . -9

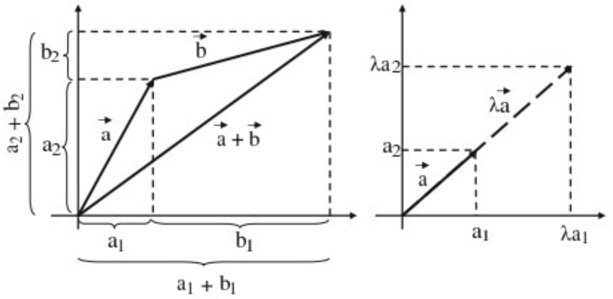
b_2
$a_2 + b_2$
a_2
$\vec{b}$
$\vec{a}$
$\vec{a} + \vec{b}$
a_1
b_1
$a_1 + b_1$
λa_2
$\lambda\vec{a}$
a_2
$\vec{a}$
a_1
λa_1

??' '; # > 2 % # > # # I # # % ?@ '

8M"' .<-9 . ' ' '9 ! % " # 8 0 > # #

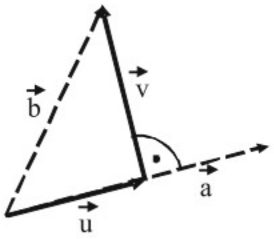
$\vec{b}$
$\vec{v}$
$\vec{a}$
$\vec{u}$

Abb. 3.14: Zerlegung des Vektors $\vec{b}$ in Projektion $\vec{u}$ und Lot $\vec{v}$ auf den Vektor $\vec{a}$

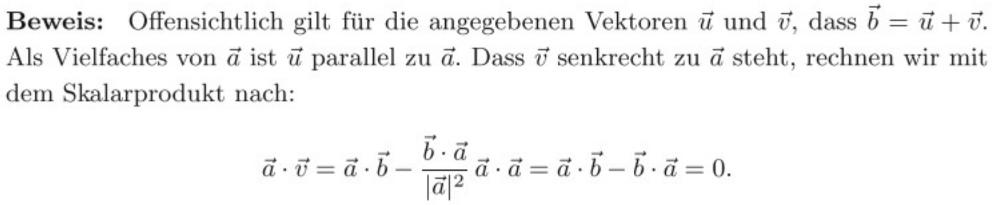

Beweis: Offensichtlich gilt für die angegebenen Vektoren $\vec{u}$ und $\vec{v}$, dass $\vec{b} = \vec{u} + \vec{v}$. Als Vielfaches von $\vec{a}$ ist $\vec{u}$ parallel zu $\vec{a}$. Dass $\vec{v}$ senkrecht zu $\vec{a}$ steht, rechnen wir mit dem Skalarprodukt nach:

$$\vec{a} \cdot \vec{v} = \vec{a} \cdot \vec{b} - \frac{\vec{b} \cdot \vec{a}}{|\vec{a}|^2}\, \vec{a} \cdot \vec{a} = \vec{a} \cdot \vec{b} - \vec{b} \cdot \vec{a} = 0.$$

/ ! "# 9 $ % + 1 >#"
4 # + 2 # 5# 2 8%+ 8 ? % #" - ..< '1 2
8M"' 9-9/ H

< ! "# 9 $ % # % "+
8 - '1 2

! "# 9 $ % # % "+
8 - '1 2

8M"' ;.- <9

. 6 # & # 4 # # 8 # 8 # # +
1 $ 2 % + 2 # 6 @ > # = "# "
) # 6 " 8, K

. ! "# 9
% 1 D ? * %=+1 #> '1 2

! "# 99 / . '1 2
!? ' =+1 ? 2 @'

! "# 99 & 8 ƒ! J$„3 5 ! '1 2

! "# 9 # @ ('1 2
!? ' CCC
5 1 # … CCC

6 # (# +
.<< 1 "

! "# 9 "'1 2
!? ' #" % + % + 2 # " +
1

9 ! "# 9 K %? '1 2
!? ' #" 3% + 3% + 2 # (+
1

! "# 9; " # 2 '1 2

; ! "# 9/ ($ 4 @# '1 2

8M"' - ;

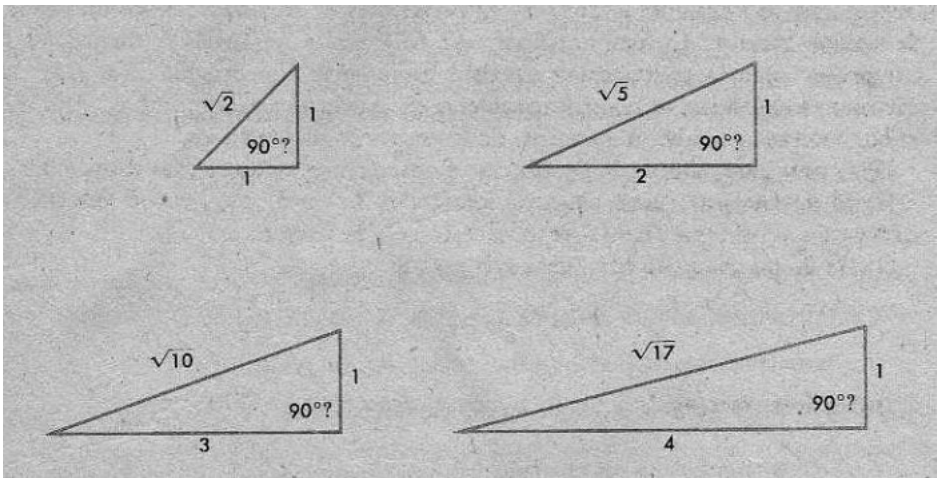
√2
1
90°?
1
√5
1
90°?
2
√10
1
90°?
3
√17
1
90°?
4

/ ! " # 9/ " # '1 2

< ! " # 9< ;;</'1 2
!? ' + + H $ % &D1 #'

. ! " # 9< G + +'1 2

! " # 9< 6 ? : % 1 : , + +J #
5# # 6 ? 3% +: (+: # (1 3% + I 8 J 8 - '1 2

! " # 9< 6 ? : % 1 : # + +
- '1 2

! " # 9< 2 + +:
* > 2 E #0 > '1 2

! " # 9< + + # !1 # #
'1 2

9 ! " # 9< + +: % 1 # 1 ' #
5# # $ % % "% (+ I + " - ; '1 2
!? ' CCC

! " # 9< + + 1 E
(# # 2 8 ? # % + # 8%+ * # 2 8 ? % #"
1 #+ - '1 2

; ! " # 9< 4 % E # # $
+ +'1 2

/ ! " # 9.
+ & # 1 '1 2

< ! " # < 9/ / . '1 2
!? ' 3% % 1 2 1
8M"' ;- .

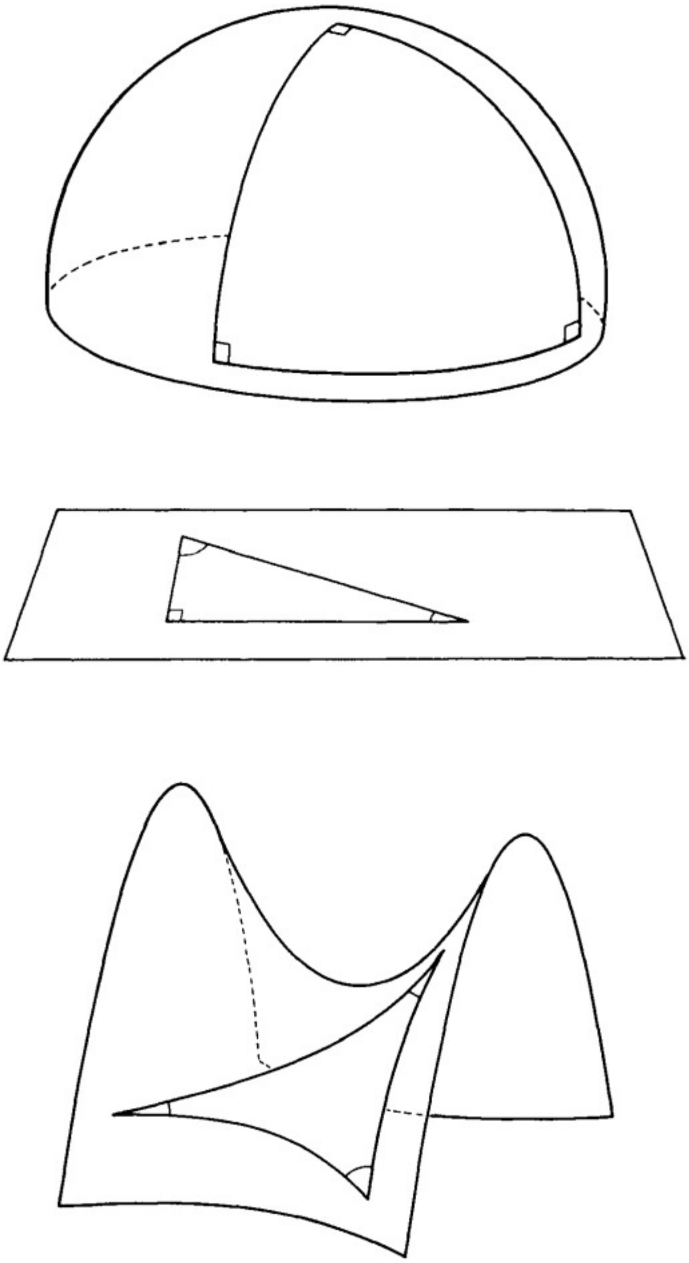

G " '/ 3% % 1% : % 1 # : # % 2
% 22 # # # " '

. ! " # " 1 '1 2

9 ! " # " # '1 2

9 ! " # $ '1 2
!? '
$ 1 2 6 # 8 '
5 1 # CCC
CCC

9 ! " # 1 # '1 2
!? '
3% 3 # # # 2 R

9 ! " # 8 $' %@ K:) %# ='
%@ K 1 +J 2 & # # #" - 9: ? " 7# + 8 ' 0

99 ! " # 8 $' %@ K:) %# ='
%@ K 1 +J G & # # #" - 9: ? " 7# + 8 '1 2

9 ! " # # 1 1 2 '1 2
!? ' &ƒ!6 7 & !! 4!66„& 3 6(&635& &
(„!$„3 5& 8 6&

9; ! " # 1 # 1 1 2 '1 2
!? ' &ƒ!6 7 & !! 4!66„& 3 6(&635& &
(„!$„3 5& 4 6 S& 8 &

9/ ! " # 9 4 (% #" #
$? '1 2

9< ! " # 9 . 9 . ;9<'1 2
!? '
& # 2 # + (+'
8M"' ;/-<.
E > 55
3 5(=3 56& 64 5 &
= 83& 5
3 5(=3 56&'

2 % & # 2 % G 4 " '

9. ! " # ?#/ ' ' ' %' '1 2
!? '
&ƒ!6 4& („!$„3 5& 6 A35 7&

! " # ; % 5' $? & # #O
1 3% + 2 +J & " # - . # & + 5# 1 # - . . - ../:
1 #" '1 2

! " # /
K # J J" J 22 #K J J '1 2

! " # /
K # J J" J 22 #K J J '1 2

! " # /
K # J J" J 22 #K J J '1 2

9 ! " # < SJ-" '1 2

! " # . " E # % K #
" 22 #K ' 1 - . '1 2

; ! " # . " E # % K #
2 #K # ? 1 2 #K # % &' 1 - . ; '1 2

!? ' 5 1 #
!? ' 7
JE 5 $ %J 1 # 5 1 # " E # % K # " 22 #K '
1 - . '1 2
1
JE 5 $ %J 1 # E > J<J
!? '
3 # # 1
4 3 6 E > 5 1 # …

5 1 #
! " # 5 1 # … 2 # ;

!? ' # , '

/ ! " # ; K # J J" '1 2

< ! " # ;
K # J J" J # '1 2

. ! " # ;
K # J J" J # J J1 J-(' # '1 2

; ! " # & # #J J(J "
2 # ;
!? ' # , " # '

8M"' </-

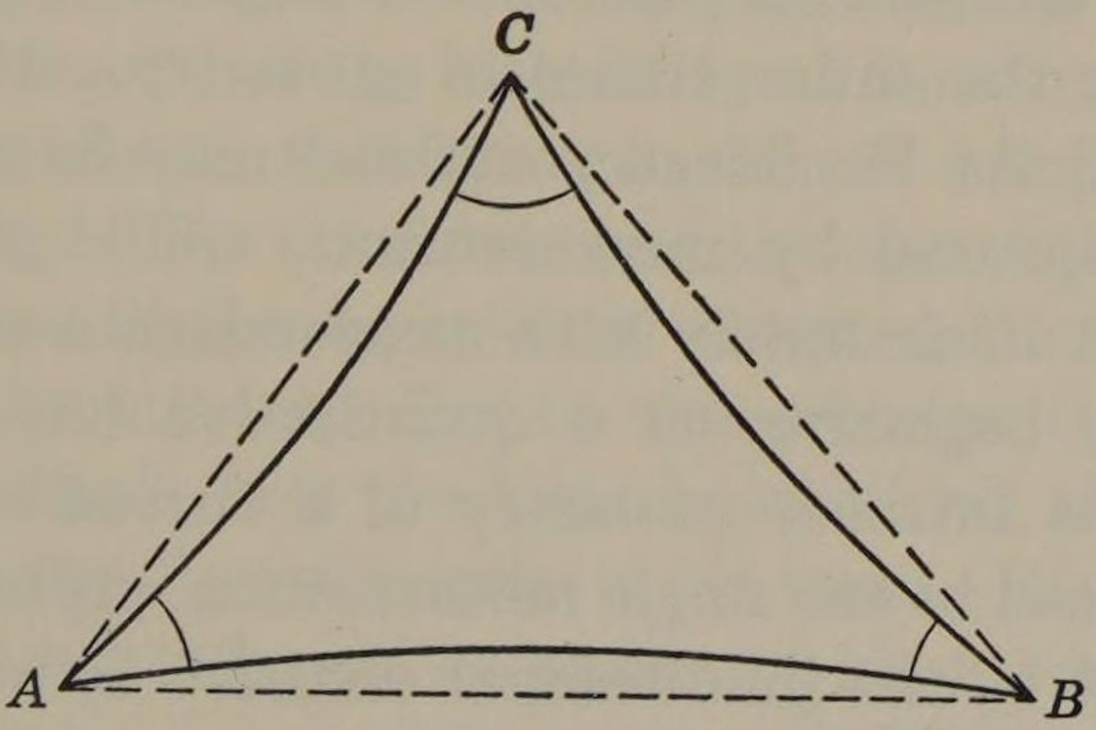

Figure 34

!? ' (H 2 # '

0 1M" # " # '

; ! " # ; -8 # # # ? + $:
& 3% 8+ % " # % 9: + % +' 8 # # 7# + 8 - '1 2
2'
% 1 @@@' " + ' # ' ./< ;. ..9.< 1 2

G % ? > 3% 8+ % " # 3%
& $

% 1 ' " ' ./< ;. ..9.<

8M"' < - <

†‡

; ! " # ;
/& # #J # J(J @'1 2

; ? ! " # ; 8 A " * * %
3 J 3% G " # " + 2 (# & # # - 9 '1 2

; ! " # ; 8 A " * * %
3 J 3% G " # " + !2 (# & # # E E > - 9 '1 2
!? ' ? # CCC
8M"' 9 - 9 H (
= 2 + @ % % #2 # + % ? ˆ
:ˆ @% % @ # " # @ % # + # # %
: 2 (: @% % % % #2 # + +
ˆ1 # 'ˆ E 2 # : @ D % #2 # + @ #
? # 2 # # ? ' 3 +: 2 D 1 : % % #
? #2 # # # ? @ > # # # % %
? : #2 # +: % # ? % 2 # # ? : ?
% % % 2 # # ? " # 2 : # % %
% # # + % # # # ? ? " # "%
# ? % # ' # : # # : 1 1 @ 2 "
# 2 % % % % # # ? : @% % % ‰
- 1% # : % # ? % 2 # # ? ' 6 +: %
1% K 2 % @ # 1 2 @ 3% # ? % #
? % # # ? ## 2 ? # # ? ' 5 ?
: 1 # : - # % K # 2
? ' 3% > #" ## 2 % 2 1 # ' *%
? % % "# 2 # 2 % # 2 # # ? @ # S
" +: 1 2 % ‰ # % # + # 2
? ? # + % 2 : % ' = 1 2 # 2 % @
2 % @ : > % % #" #" " # 2 E ?+ # % 1+
2 &"+1 '

; ! " # ;
JE >J&D " # + # '1 2

; ! " # ; # #J J . '1 2

;9 ! " # ;
1% #J% @> #"J J? 2J% +J 2J '1 2

; ! " # ; 8 ? J88 '1 2
!? ' 8 # 1 2 8%+ "+'

;; ! " # ; 8% 1) & @ # 8 ?
8 # 1 2 8%+ "+ 8 # # 7# + 8 - .. '1 2
!? ' 8 # 1 2 8%+ "+'

;/ ! " # ; ?> /; # '1 2
!? ' 3% 5# # 6 2 3% #" H # K '

;< ! " # ; (# # "+J
8 #" 2 % 5 (. #2 # '1 2

;. ! " # ; (= + (3 '1 2
!? '
= + 2 & # #O (# + # 1
4 1 # 2 % 3% +

/ ! " # ; =
8 #" JE >J # 5# #3 # 2'1 2
8M"' /-

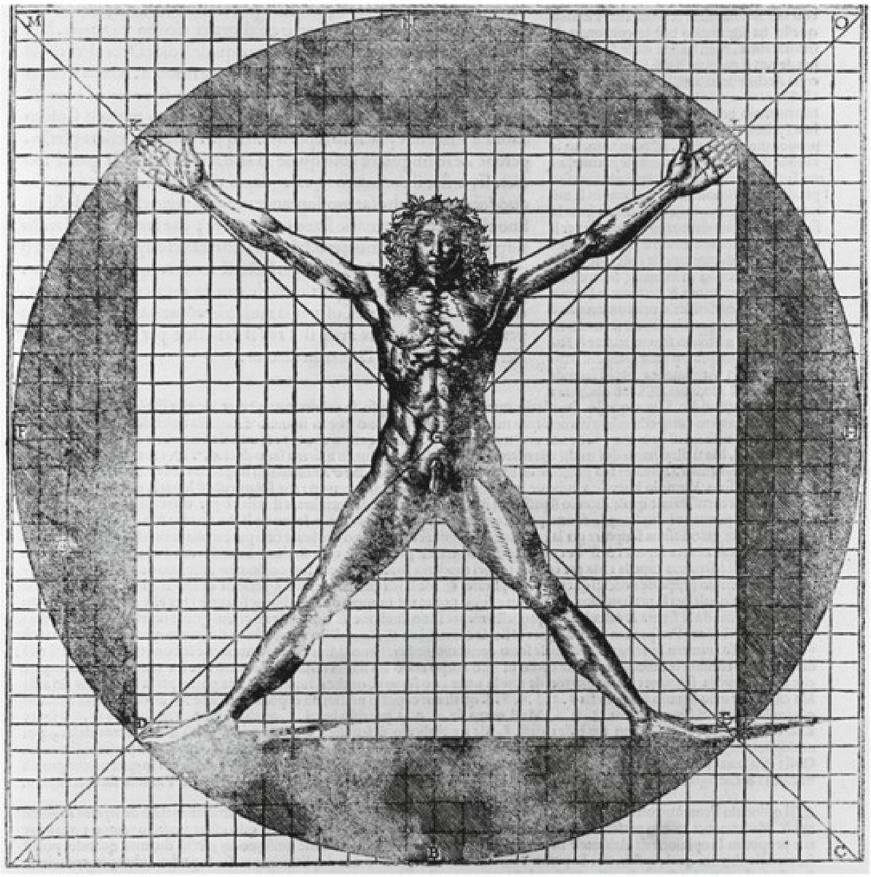
M
N
O
K
L
F
G
H
D
E
A
B
C

G "' # O - & # 2 4 % ?+ % # %
-< H E & ' 3% P" @# @ % ? >" # " 2 #" % 1 # ' 1
@ % 1 # 2 Y ; % # + # 1 ?+ # : * > # :
8 ? 4 #Z

8M"' 9-9

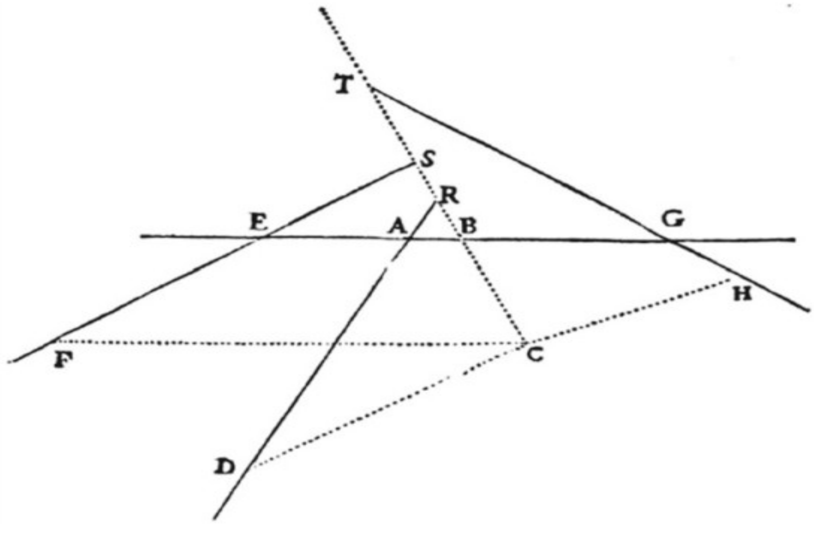
T
S
R
E
A
B
G
H
F
C
D

G "' 4 " 1% 1 # 2 1 #" 1 # 2 "% # ? ?+ %
8 11 1 ? @ % 2 # ' !# # % # -1 # :
% # % # # 2 % - # 2+ #" % # # ' 1
@ % 1 # 2 Y / % # + $ # 1 ?+ #, 4 : !1 # 8 ? % #" ':
. 8 ? 4 #Z

/ ! " # ;9 ? " ? &
< . '1 2
8M"' / - .
% 1 5 1% 2 (1

/ ! " # ; ? ' * # 3%
1 # # #J 8 "# #: > 1 : # ! '1 2
8M"' <- .

' 1 D # # # 2 % O *% 2 G # ' 3% & % -3
>@ #
+ # 4: 1 # 22 # # % # - ' E # 1 # 4: % @#
@ 1 # -4 # 4 :
# #: >@ # :9 9 + # % - 1 + @ % # : % @#
@ 1 # -
# ? P2 # @ + % # 4 # >@ # % #
- ' % O *%
2 G # "# % # # 1 # ? # # % & %O
D - : : #
-E : E : # -# > # 2 % #: # % ? % %
% #" # % & %O #
+ " # % @% # " 1 - # " # -5 @ # 2 ' =
? % %
2 4 ~ # 22 %Š 2 & 0 %O 1 1% + 2 D % # + 2 %
#" 2 % - ' ': ‹
:9 9+Œ : / + • ;: .+ ' G % .< : H ' + 2 % '

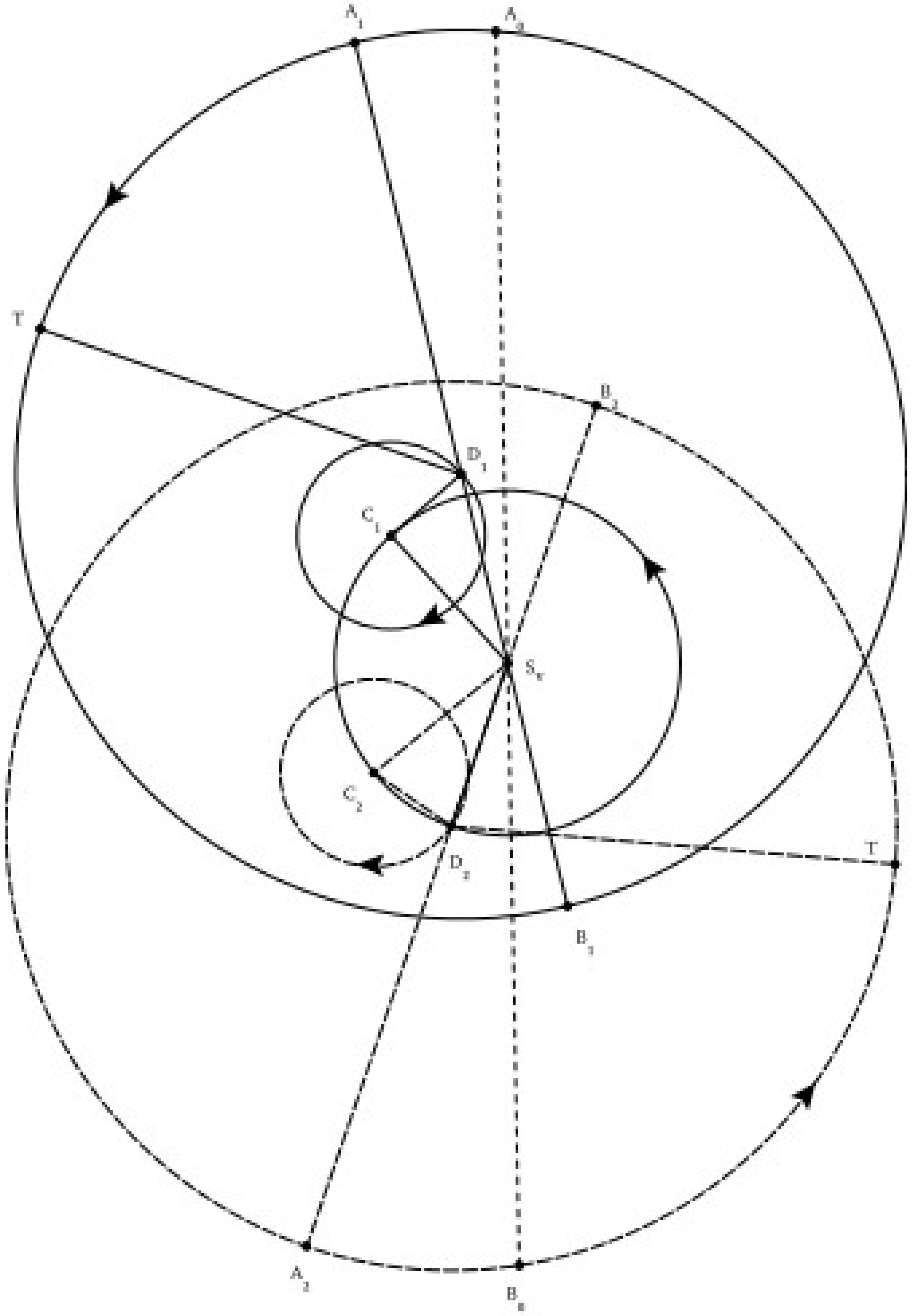
A1
A0
T
B2
D1
C1
S0
C2
D2
T
B1
A2
B0

/ ! " # ; 3% 1 # # % #
7# + 2 2 # 8 - ./ '1 2

/9 ! " # ; 3% 1 # 2 1 # O
= # 3% +J 8 #" 2 + 1 ! " # K ?+ % 6 1 #
2 % 5# # # 7# '1 2

/ ! " # ; 3% #> #" 5 1 ? 3%
5# " + 2 G #> # # 7# + 2 3 # 8 - < '1 2

/; ! " # ;; .. ;<'1 2
!? ' $ % H = "% '

// ! " # ;; 9<;9 ;/. '1 2
!? ' 1 D ? '
'
! " # ;; 9<;9 ;/. ' 0

/< ! " # ;; <; .9;9 /'1 2
!? ' 1 D # + '
8M"' .-9 H $ 1 8 1 2 % 1 D 8 # '
8M"' <;-.. 3% 11 # 2 8 '

Corollary 5.4 *For any $t \in \mathbb{R}$, $|\exp(it)| = 1$.*

Proof. We have $\exp(it) = \cos t + i \sin t$. Both $\cos t$ and $\sin t$ are real if t is, and therefore $|\exp(it)|^2 = \cos^2 t + \sin^2 t = 1$. □

Remark 5.5 For real x, $\sin x$ and $\cos x$ are also real, and so the relation $\sin^2 x + \cos^2 x = 1$ shows that $|\sin x| \le 1$ and also $|\cos x| \le 1$. These inequalities do not extend to the complex case. For example, if $z = it$, with $t \in \mathbb{R}$, then $\sin z = \sin it = (\exp(-t) - \exp t)/2i$. If t is large (and positive), then $\exp t$ is large, $\exp(-t)$ is small and so $|\sin it|$ is large. A similar remark applies to $\cos z$.

5.6 The Appearance of π

We wish to discuss various properties of $\sin x$ and $\cos x$ for real x. In particular, we would like to introduce the number π.

Theorem 5.2 *If $x \in (0, 2)$, then $\sin x > 0$.*

Proof. Suppose that $0 < x \le 1$. Then

$$\begin{aligned}\cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots \\ &> 1 - \frac{x^2}{2!} - \frac{x^4}{4!} - \frac{x^6}{6!} - \dots \\ &> 1 - \frac{1}{2} - \frac{1}{2^3} - \frac{1}{2^5} - \dots,\end{aligned}$$

since $\dfrac{x^2}{2!} \le \dfrac{1}{2}$, $\dfrac{x^4}{4!} \le \dfrac{1}{4.3.2} < \dfrac{1}{2^3}, \dots$ etc.,

$$= 1 - \frac{\frac{1}{2}}{(1 - \frac{1}{4})} = \tfrac{1}{3}.$$

Hence $\sin' x = \cos x > \frac{1}{3}$ on $[0, 1]$ and so (by the Mean Value Theorem), $\sin x$ is strictly increasing on $[0, 1]$. But $\sin 0 = 0$ and therefore $\sin x > 0$ for $0 < x \le 1$. It follows that

$$\sin x = 2 \sin \tfrac{x}{2} \cos \tfrac{x}{2} > 0$$

whenever $0 < x \le 2$, since, in this case, $\sin \frac{x}{2} > 0$ and $\cos \frac{x}{2} > \frac{1}{3}$. □

Corollary 5.5 *The function* $\cos x$ *is strictly decreasing on* $[0, 2]$.

Proof. We have $\cos' x = -\sin x$ which is strictly negative on $(0, 2)$, by the theorem. By the Mean Value Theorem, it follows that the function $\cos x$ is strictly decreasing on $[0, 2]$. □

Theorem 5.3 $\sin 4 < 0$.

Proof. From the definition of $\sin x$,

$$\begin{aligned} \sin 4 = 4 - \frac{4^3}{3!} + \frac{4^5}{5!} - \frac{4^7}{7!} + \frac{4^9}{9!} \\ - \frac{4^{11}}{11!}\left(1 - \frac{4^2}{12.13}\right) - \frac{4^{15}}{15!}\left(1 - \frac{4^2}{16.17}\right) \\ - \frac{4^{19}}{19!}\left(1 - \frac{4^2}{20.21}\right) - \dots \\ < 4 - \frac{4^3}{3!} + \frac{4^5}{5!} - \frac{4^7}{7!} + \frac{4^9}{9!} = -\tfrac{268}{405} \end{aligned}$$

and the result follows. □

Theorem 5.4 *There is a unique real number* π *satisfying* $0 < \pi < 4$ *such that* $\sin \pi = 0$. *Furthermore,* $\cos \frac{\pi}{2} = 0$, $\sin \frac{\pi}{2} = 1$ *and* $\cos \pi = -1$.

Proof. We have seen that $\sin x > 0$ for $x \in (0, 2)$ and so, in particular, $\sin 1 > 0$. We have also shown that $\sin 4 < 0$. Now, the map $x \mapsto \sin x$ is continuous on $\mathbb{R}$, and so, by the Intermediate Value Theorem, there is some real number, which we will denote by π, with $1 < \pi < 4$ and such that $\sin \pi = 0$.

We must now show that π is the only value in $(0, 4)$ obeying $\sin \pi = 0$. To this end, suppose that $\sin \alpha = 0$ with $0 < \alpha < 4$. Then

$$0 = \sin \alpha = 2 \sin \tfrac{\alpha}{2} \cos \tfrac{\alpha}{2}.$$

Hence either $\sin \frac{\alpha}{2} = 0$ or $\cos \frac{\alpha}{2} = 0$ (they cannot both vanish because the sum of their squares is equal to 1). But if $0 < \alpha < 4$, then $0 < \frac{\alpha}{2} < 2$ and we know that $\sin x > 0$ on $(0, 2)$, so $\sin \frac{\alpha}{2}$ cannot be zero. Hence, we must have that $\cos \frac{\alpha}{2} = 0$. In particular, $\cos \frac{\pi}{2} = 0$. However, $\cos x$ is strictly decreasing on $[0, 2]$ and so there can be at most one solution to $\cos x = 0$ in this interval. Hence $\frac{\alpha}{2} = \frac{\pi}{2}$ so that $\alpha = \pi$, and the uniqueness is established.

Next, $\sin^2 z + \cos^2 z = 1$ implies that $\sin^2 \frac{\pi}{2} = 1$, since $\cos \frac{\pi}{2} = 0$. Hence $\sin \frac{\pi}{2} = 1$ because $\sin \frac{\pi}{2} > 0$ (since $0 < \frac{\pi}{2} < 2$).

Finally, we have $\cos \pi = 2 \cos^2 \frac{\pi}{2} - 1 = -1$. □

What's going on? We have defined the trigonometric functions, out of the blue as it were, by means of power series. This approach avoids any appeal to geometry and right-angled triangles. However, having chosen this route, we must stick with it. In particular, it is necessary to get to π via these definitions rather than by drawing triangles or circles. As we have seen, this is all perfectly possible. Our view here is that the trigonometric functions are as we have defined them. Any results must be deduced as consequences of these power series definitions.

Remark 5.6 Numerical investigation yields $\pi = 3.14159\ldots$. It is known that π is irrational (in fact, transcendental). It is something of a sport (involving some fascinating numerical analysis) to calculate the value of π to a large number of decimal places and this has been done to over a million decimal places. Such programs have been used to test the computational integrity of supercomputers by checking to see whether they get these digits correct or not.

Theorem 5.5 *For any $z \in \mathbb{C}$ and $n \in \mathbb{Z}$,*

$$\begin{aligned}\sin(z + n\pi) &= (-1)^n \sin z\\ \cos(z + n\pi) &= (-1)^n \cos z.\end{aligned}$$

Proof. We use the trigonometric formulae;

$$\begin{aligned}\sin(z + \pi) &= \sin z \cos \pi + \cos z \sin \pi\\ &= -\sin z + 0,\end{aligned}$$

and

$$\begin{aligned}\cos(z + \pi) &= \cos z \cos \pi - \sin z \sin \pi\\ &= -\cos z - 0.\end{aligned}$$

For $n > 0$, the result now follows by induction. Substituting $z = w - n\pi$, the result then follows for $n < 0$. □

Remark 5.7 Particular cases of the above formulae deserve mention. If we set $z = 0$, then we see that $\sin(n\pi) = 0$ and $\cos(n\pi) = (-1)^n$, for any $n \in \mathbb{Z}$. The functions sin and cos are periodic (with period 2π). In particular, for any $x \in \mathbb{R}$ and any $n \in \mathbb{Z}$,

$$\begin{aligned}\sin(x + 2n\pi) &= \sin x \quad \text{and}\\ \cos(x + 2n\pi) &= \cos x.\end{aligned}$$

/. ! " # ;; 9 9'1 2
!? '
G # # 2 (+ # # H 4'
8M"' -/

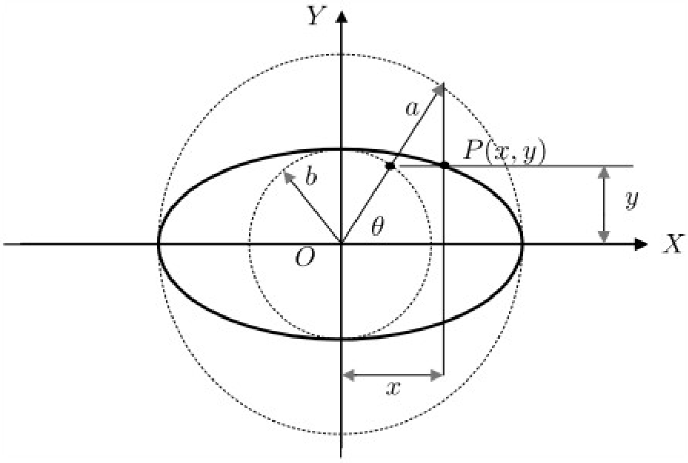
Y
a
P(x, y)
b
y
θ
X
O
x

G "' '/ 3% 1 1 # # 2 % 1

8M"' .-/;

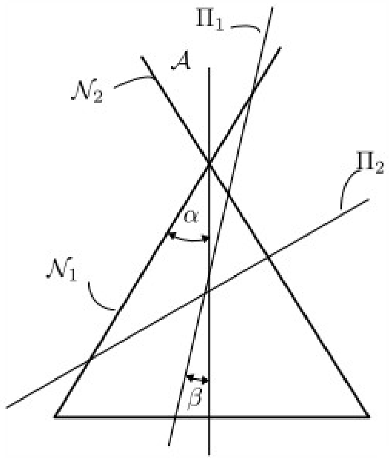
Π1
A
N2
Π2
α
N1
β

G "' ' 3% %+1 ? " # ?+ 1 # #" "% # # #" Ž • • : •
? #" % #" ?+ % # @ % % # D

8M"' ; -/<

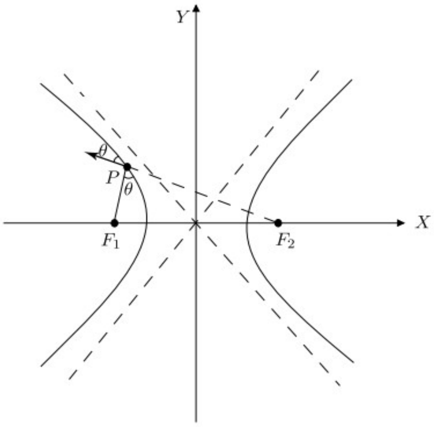

Fig. 2.15 A reflective property of the hyperbola: lines $\overline{F_1 P}$ and $\overline{F_2 P}$ make equal angles with the tangent at P

< ! " # ;; ;; ; '1 2
!? ' 3% 7# 2
8M"' .- /
9 8 0 # #

8M"' 9- <

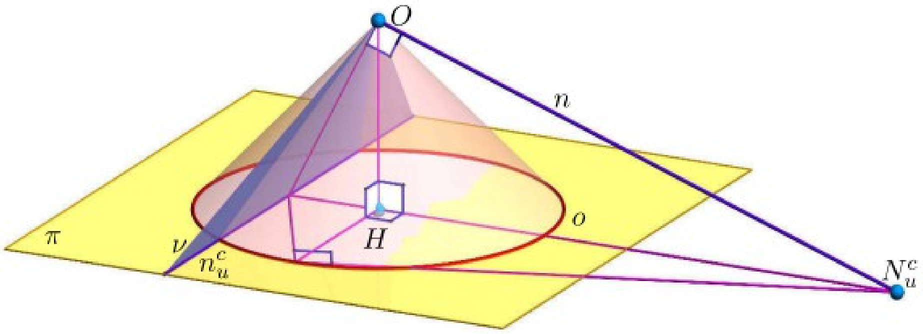

FIGURE 10.11. The Euclidean orthogonality in the star about O is projected to the anti-polarity w.r.t. the distance circle o in the image plane π from the eye point O. The plane ν and the line n are orthogonal, and thus, the respective vanishing line n_u^c and vanishing point N_u^c are antipolar w.r.t. the circle o.

8 % #" # 2 & # # 2 #

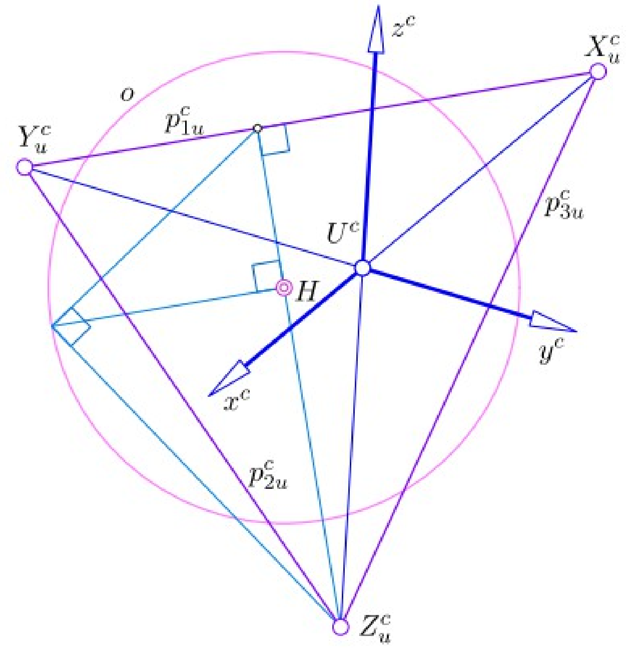
z^c
X_u^c
o
p_{1u}^c
Y_u^c
p_{3u}^c
U^c
H
y^c
x^c
p_{2u}^c
Z_u^c

FIGURE 10.12. The vanishing points of X_u^c, Y_u^c, Z_u^c of the axes of a Cartesian frame are the vertices of an antipolar triangle of the distance circle o. The principal (vanishing) point H (orthogonal projection of the eye point O onto the image plane π) is the orthocenter of the triangle of the three vanishing points.

8M'"   ;-

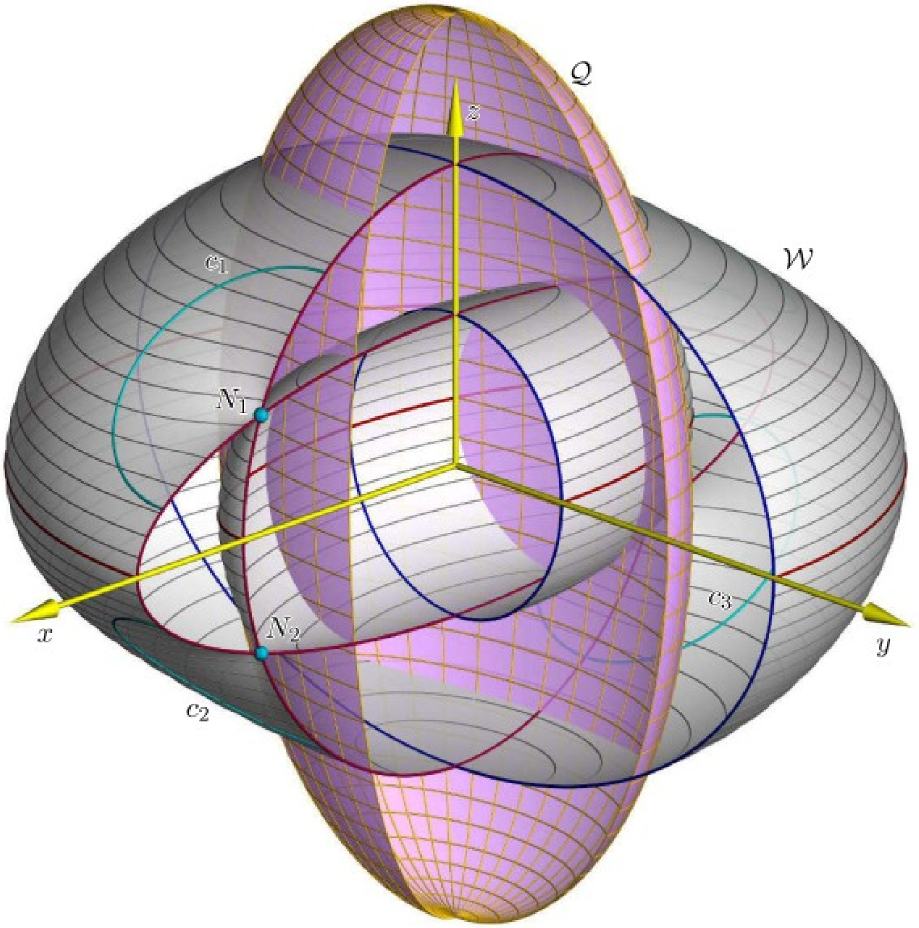

FIGURE 10.16. The quartic Fresnel wave surface $\mathcal{W}$ has four conical nodes N_1, …, N_4 corresponding to the umbilics of an ellipsoid $\mathcal{Q}$. $\mathcal{W}$ intersects its planes of symmetry along circles and ellipses (red and blue). There are four circles c_1, …, c_4 on Fresnel's surface that are the intersection of the four real double tangent planes with $\mathcal{W}$.

Painvin's complex

The tangent planes of a regular quadric $\mathcal{Q}$ form the dual quadric $\mathcal{Q}^\star$. A quadric $\mathcal{Q}$ sends two tangent planes through each line that is not tangent to $\mathcal{Q}$, and thus, a quadric is said to be of *rank* two. This notion of rank is not to be confused with the rank of a matrix, especially with the rank of the coefficient matrix of a quadric. By assumption, $\mathcal{Q}$ is regular, and

< ! " # ;/ " +'1 2

< ! " # ;< (# # " " +
- : 3 + , G # '1 2
8M"' -

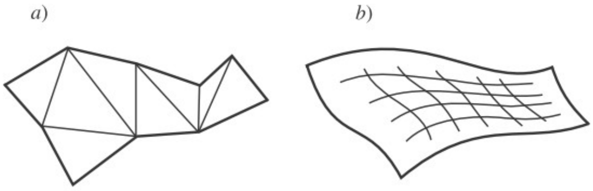

Figure 1.1. Approximate realization (a) of a curved surface (b).

trying to test spacetime geometry by physical measurements. His/her conclusions will also be valid locally for true gravitational fields, on the basis of the PE. Assume that the observer sits at the centre of the disc, and has two identical clocks: one of them is placed at the centre, the other at some point on the periphery. The observer will see that the clock on the periphery is running slower (time dilatation from SR). Consequently, clocks at various positions in the gravitational field run faster or slower, depending on the local strength of the field. There is no definition of time that is pertinent to the whole spacetime in general. The observer will also conclude that the length of a piece of line orthogonal to the radius of the disc will be shortened (length contraction from SR). Therefore, the ratio of the circumference of the circle to its radius will be smaller than 2π, so that the Euclidean geometry of space does not hold in GR.

Geometry of curved surfaces. We have seen that in spacetime, within a limited region, we can always choose a suitable reference frame, called the local inertial frame. Taking spacetime apart into locally inertial components, we can apply the laws of SR in each such component and derive various dynamical conclusions. Reconstruction of the related global dynamical picture, based on the PE (and some additional, simple geometric assumptions), gives rise to GR.

This procedure for dissecting spacetime into locally inertial components, out of which we can reconstruct, using the PE, the global structure of spacetime containing the gravitational field, can be compared geometrically with an attempt to build a curved surface, approximately, from a lot of small, plane elements 'continuously' bound to each other (figure 1.1). A locally flat surface is a geometric analogue of locally inertial spacetime.

GR makes essential use of curved spaces (or, more accurately curved spacetimes). Since the curved, four-dimensional spacetime cannot be visualized, let us try to understand the basic features of curved spaces by considering a two-dimensional surface X_2. We shall focus our attention on those geometric properties of X_2, which could be determined by an intelligent, two-dimensional being entirely confined to live and measure in the surface, a being that does not know anything about the embedding Euclidean space E_3. Such properties determine the *intrinsic geometry* on X_2.

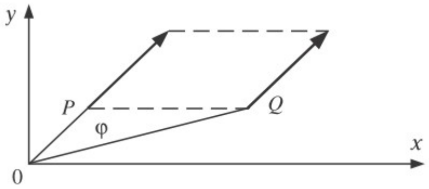

Figure 1.2. Vector components change under the parallel transport.

On a general surface we cannot set up Cartesian coordinates; the best we can do is to introduce two arbitrary families of coordinate lines and label them by $u^\alpha = (u^1, u^2)$. These curvilinear (or Gaussian) coordinates do not always have a direct geometric interpretation.

Let us now imagine a surface X_2 embedded in E_3, and consider a point P in X_2 with Cartesian coordinates (x, y, z) in E_3; since P is in X_2, these coordinates can be expressed as functions of u^α. The squared distance between two neighbouring points in X_2, $P(x, y, z)$ and $Q(x+\mathrm{d}x, y+\mathrm{d}y, z+\mathrm{d}z)$, is given by $\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2$. Going over to u^α we obtain

$$\mathrm{d}s^2 = g_{\alpha\beta}\, \mathrm{d}u^\alpha\, \mathrm{d}u^\beta \tag{1.6}$$

where the set of functions $g_{\alpha\beta}(u^1, u^2)$ defines the *metric* on X_2. The metric is an essential intrinsic property of X_2, independent of its embedding in E_3. A two-dimensional continuum X_2, equipped with a squared differential distance according to (1.6), becomes the *metric space*, $G_2 = (X_2, \boldsymbol{g})$ (it is now clear that the symbol X_2 is used to denote an abstract two-dimensional continuum, disregarding all metric properties of the real surface).

Studying the motion of particles in X_2 our observer could have arrived at the idea of the tangent vector, $\boldsymbol{a} = (a^1, a^2)$. Consider, further, two tangent vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, defined at neighbouring points P and Q. If we wish to compare these two vectors we have to know how one of them can be transported to the position of the other. This transport is called *parallel transport*. The continuum X_2 together with the parallel transport law, denoted by Γ, is called a *linearly connected* (or *affine*) space, $L_2 = (X_2, \Gamma)$.

The components of a vector in curvilinear coordinates are expected to change under parallel transport. This can be easily seen by considering the parallel transport of a vector in a plane, using polar coordinates (r, θ). As shown in figure 1.2, if the unit vector $\boldsymbol{a}$ at point P has components ($a^r = 1$, $a^\theta = 0$), after parallel transport from P to Q its components become ($a^r = \cos\varphi$, $a^\theta = \sin\varphi$).

A two-dimensional observer should determine the rule of parallel transport of vectors in accordance with his/her own experience. This rule is generally independent of the concept of a metric. If X_2 is equipped with a rule of

parallel transport and the independent concept of a metric, we obtain the *linearly connected metric* space $(X_2, \Gamma, \boldsymbol{g}) = (L_2, \boldsymbol{g}) = (G_2, \Gamma)$.

It is very probable that our two-dimensional observer will demand that the lengths of vectors should remain unchanged under parallel transport. After adopting this very natural property, which relates Γ and $\boldsymbol{g}$, the linearly connected metric space becomes the *Riemann–Cartan* space U_2. The usual surfaces have some additional geometric structure, and belong to the class of *Riemann* spaces V_2.

This discussion was slightly more general than strictly necessary for studying surfaces embedded in E_3, in order to emphasize that the intrinsic geometry on X_2 is defined in terms of

- the metric $\boldsymbol{g}$ and
- the rule of parallel transport Γ.

These basic concepts of the intrinsic geometry of surfaces can be extended directly to spaces with higher dimensions, although in this case we lose the intuitive geometric picture.

Any Riemann space is locally Euclidean, i.e. its metric is positive definite ($\mathrm{d}s^2 > 0$). A slight generalization of this case, that consists in admitting metrics that are pseudo-Euclidean, leads to a pseudo-Riemannian geometry or Riemannian geometry with an indefinite metric.

Relativity, covariance and Mach's ideas

Relativity and covariance. We have seen previously that inertial frames play a particularly important role in SR. Einstein considered this to be an expression of the incompleteness of SR. He wanted to generalize the relativity of inertial motions to the relativity of all motions, including accelerated ones. This idea can be formulated as the general principle of relativity:

> *General PR: The form of physical laws is the same in all reference frames.*

Equations expressing physical laws are often formulated using specific coordinates. Thus, for instance, spatial positions in an inertial laboratory L can be determined using Cartesian coordinates (x, y, z) or the coordinates obtained from these by a rotation around the z-axis for a fixed angle φ. If we allow the angle φ to change with time as $\varphi = \omega t$, new coordinates can be naturally attached to a new physical reference frame, rotating relative to L with the angular velocity ω around the z-axis. We see that coordinate transformations can be interpreted as describing transitions from one to another laboratory (the so-called active interpretation). Thus, it becomes clear that the general PR can be realized with the help of the principle of general covariance:

< ! " # ;. 3% 8%+ 2 &# "+'1 2

<9 ! " # / / ; /'1 2
!? '
& + # $ "# '

< ! " # / .< ; / '1 2
!? ' & "# $ & +'

<; ! " # / 1 2'1 ?J # 2 #
+ # '1 2

</ ! " # / 1 2'1 ?J" '1 2
!? '
CCC

<< ! " # / 7# # 8 0 (+J
#" 2 8 +% !?0 # # - 9: 1 #" " E
= ? " '1 2
8M"' - .
' ' 8 0 1 # = " # #

2.1.1 Projective Space and Homogeneous Coordinates

Generally, a projective space $\mathbb{P}^n$ is the set of one-dimensional vector subspaces of a given vector space $\mathbb{R}^{n+1}$. Each one-dimensional vector subspace represents a projective point $\mathbf{x}$ with the *projective* dimension 0.

Assuming $n = 1$, the projective space is the projective line $\mathbb{P}^1$ and is visualized in figure 2.1 (a) on the facing page: the projective points $\mathbf{x}_{1/2} \in \mathbb{P}^1$ can be identified with lines going through the origin of the vector space $\mathbb{R}^2$. Thus these lines are one-dimensional subspaces of $\mathbb{R}^2$. Note that no coordinate system but only the origin is defined here.

A definition of a projective point $\mathbf{x} \in \mathbb{P}^n$ may be given as follows:

Definition 2 (Projective point) *A projective point* $\mathbf{x} \in \mathbb{P}^n$ *of dimension* n *is an element of the quotient of* $\mathbb{R}^{n+1} \setminus \mathbf{0}$ *and the equivalence relation* $\cong$*:*

$$\mathbf{x} \cong \mathbf{y} \quad \Leftrightarrow \quad \exists \lambda \neq 0, \mathbf{x} = \lambda \mathbf{y} \tag{2.1}$$

In other words: all vectors $\lambda \mathbf{x}^h = (\lambda x_1, \ldots, \lambda x_n, \lambda)^\mathsf{T} \in \mathbb{R}^{n+1}$ $\lambda \neq 0$ are equivalent with respect to $\cong$ and define a projective point $\mathbf{x}$. In other words, a projective point $\mathbf{x} \in \mathbb{P}^n$ is defined by representing it with a $n+1$ vector $\mathbf{x}^h \in \mathbb{R}^{n+1}$, the so-called *homogeneous vector* representation. This representation is overdetermined: the equivalence relation states, that the lengths of the vectors $\mathbf{x}^h, \mathbf{y}^h \in \mathbb{R}^{n+1}$ do not matter, only their directions.

them within the Euclidean framework. Therefore we need a more general framework which enables us to deal with infinite points, this mathematical framework is known as *projective geometry.*

In this and the following chapters, we will see that projective geometry provides a simple and consistent way to represent and transform geometric entities, that are important for Euclidean polyhedral reconstruction. One drawback though is the fact that properties such as parallelity or orthogonality can not be directly used; more precisely, they are not invariant with respect to projective transformations. The study of invariance of properties with respect to certain groups of transformation may be regarded as a definition of geometry, as KLEIN 1939 mentioned in his Erlanger Programm. We will come back to the issue of invariance in section 3.3 on page 70, where we investigate possible relations between geometric entities and their invariance properties with respect to Euclidean, similar, affine and projective geometry. We will see that for our application it is possible to infer parallelity and orthogonality of lines and planes from homogeneous vectors by using the Euclidean interpretation thereof.

For a detailed and more precise introduction to projective geometry in the context of Computer Vision, see FAUGERAS AND LUONG 2001 or chapter 23 in MUNDY 1992.

2.1.1 Projective Space and Homogeneous Coordinates

Generally, a projective space $\mathbb{P}^n$ is the set of one-dimensional vector subspaces of a given vector space $\mathbb{R}^{n+1}$. Each one-dimensional vector subspace represents a projective point $\mathbf{x}$ with the *projective* dimension 0.

Assuming $n = 1$, the projective space is the projective line $\mathbb{P}^1$ and is visualized in figure 2.1 (a) on the facing page: the projective points $\mathbf{x}_{1/2} \in \mathbb{P}^1$ can be identified with lines going through the origin of the vector space $\mathbb{R}^2$. Thus these lines are one-dimensional subspaces of $\mathbb{R}^2$. Note that no coordinate system but only the origin is defined here.

A definition of a projective point $\mathbf{x} \in \mathbb{P}^n$ may be given as follows:

Definition 2 (Projective point) *A projective point* $\mathbf{x} \in \mathbb{P}^n$ *of dimension* n *is an element of the quotient of* $\mathbb{R}^{n+1} \setminus \mathbf{0}$ *and the equivalence relation* $\cong$*:*

$$\mathbf{x} \cong \mathbf{y} \quad \Leftrightarrow \quad \exists \lambda \neq 0, \mathbf{x} = \lambda \mathbf{y} \tag{2.1}$$

In other words: all vectors $\lambda \mathbf{x}^h = (\lambda x_1, \ldots, \lambda x_n, \lambda)^\mathsf{T} \in \mathbb{R}^{n+1}$ $\lambda \neq 0$ are equivalent with respect to $\cong$ and define a projective point $\mathbf{x}$. In other words, a projective point $\mathbf{x} \in \mathbb{P}^n$ is defined by representing it with a $n+1$ vector $\mathbf{x}^h \in \mathbb{R}^{n+1}$, the so-called *homogeneous vector* representation. This representation is overdetermined: the equivalence relation states, that the lengths of the vectors $\mathbf{x}^h, \mathbf{y}^h \in \mathbb{R}^{n+1}$ do not matter, only their directions.

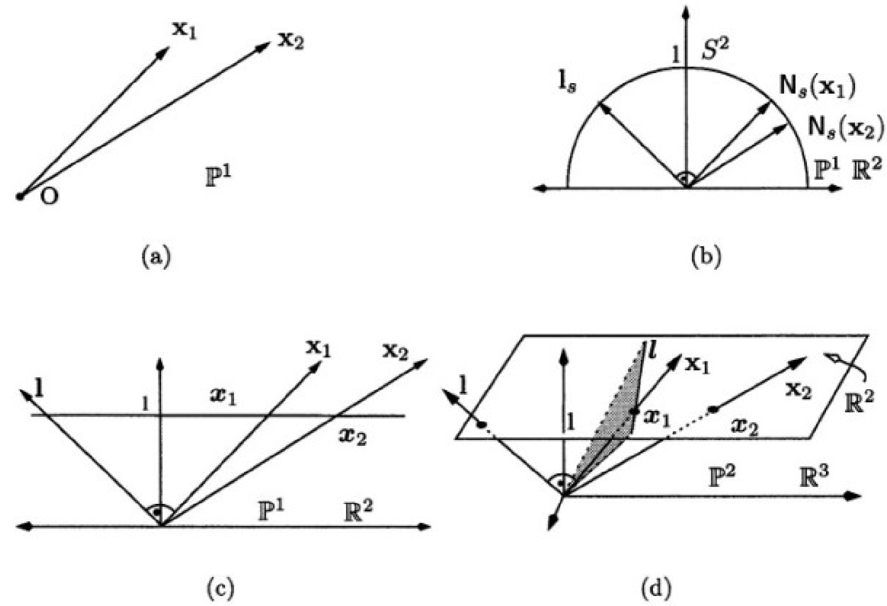
$\mathbf{x}_1$
$\mathbf{x}_2$
O
$\mathbb{P}^1$
(a)
1
S^2
$\mathbf{l}_s$
$\mathsf{N}_s(\mathbf{x}_1)$
$\mathsf{N}_s(\mathbf{x}_2)$
$\mathbb{P}^1$ $\mathbb{R}^2$
(b)
$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{l}$
1
$\boldsymbol{x}_1$
$\boldsymbol{x}_2$
$\mathbb{P}^1$
$\mathbb{R}^2$
(c)
l
$\mathbf{x}_1$
$\mathbf{l}$
$\mathbf{x}_2$
$\mathbb{R}^2$
1
$\boldsymbol{x}_1$
$\boldsymbol{x}_2$
$\mathbb{P}^2$
$\mathbb{R}^3$
(d)

Fig. 2.1. Different interpretations of the projective line $\mathbb{P}^1$: in (a), projective points $\mathbf{x}_1, \mathbf{x}_2$ are interpreted as lines going through the origin without defining any coordinate system. In (b), the projective points are homogeneous vectors, that are normalized to the sphere S^2. In (c) the homogeneous vectors intersect the Euclidean axis and define the points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ on the Euclidean axis. The role of a hyperplane $\mathbf{l}$ can be visualized by considering the projective plane $\mathbb{P}^2$ in (d): two projective points $\mathbf{x}_1$ and $\mathbf{x}_2$ define two points $\boldsymbol{x}_1, \boldsymbol{x}_2$ on the Euclidean plane $\mathbb{R}^2$. The first point $\boldsymbol{x}_1$ is incident to a line l on the Euclidean plane. The line l in turn is defined by the homogeneous vector $\mathbf{l} \in \mathbb{R}^3 \simeq \mathbb{P}^2$, which is the normal of the plane that contains the origin of $\mathbb{R}^3$ and the line on the Euclidean plane $\mathbb{R}^2$. As the first Euclidean point $\boldsymbol{x}_1$ is incident to the Euclidean line, the homogeneous vector $\mathbf{l}$ is orthogonal to the homogeneous vector $\mathbf{x}_1$, or $\mathbf{l}^\mathsf{T}\mathbf{x}_1 = 0$.

To simplify the notation, we identify projective points $\mathbf{x} \in \mathbb{P}^n$ with their (non-unique) coordinate vector representation based on $\mathbf{x}^h \in \mathbb{R}^{n+1}$, i.e. $\mathbf{x} = \mathbf{x}^h$. Therefore we are not strict with respect to mathematical notation and do *not* distinguish between projective points and its coordinate vectors as in FAUGERAS AND LUONG 2001, p. 78. HARTLEY AND ZISSERMAN 2000 take a similar approach, see p. 3f. In the following, we assume that the correct interpretation is clear from the context.

One can *normalize* a homogeneous vector $\mathbf{x}_s := \frac{\mathbf{x}}{|\mathbf{x}|}$ to avoid ambiguities, see fig.2.1(b). The index s in the homogeneous vector $\mathbf{x}_s$ indicates that the vector is normalized and thus lies on the unit sphere. We could use the normalization of a homogeneous vector with its length for an alternative

definition of the equivalence class $\cong$ for two homogeneous vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n+1} \setminus \mathbf{0}$:

$$\mathbf{x} \cong \mathbf{y} \quad \Leftrightarrow \quad \left(\frac{\mathbf{x}}{|\mathbf{x}|} = \frac{\mathbf{y}}{|\mathbf{y}|} \text{ or } \frac{\mathbf{x}}{|\mathbf{x}|} = -\frac{\mathbf{y}}{|\mathbf{y}|} \right) \quad \Leftrightarrow \quad \frac{\mathbf{x}}{|\mathbf{x}|} = \pm \frac{\mathbf{y}}{|\mathbf{y}|} \tag{2.2}$$

We sometimes denote the spherical normalization by $\mathsf{N}_s(\mathbf{x}) = \frac{\mathbf{x}}{|\mathbf{x}|}$.

The definition of the equivalence class $\cong$ is ambiguous as the orientation of the normalized vectors can be positive or negative. In other words, a projective point from $\mathbb{P}^n$ has two representations on a sphere S^{n+1}. To overcome this ambiguity one may use *oriented* projective geometry, as proposed by STOLFI 1991, where every element has an intrinsic orientation. This is analytically expressed by the specialized definition, that homogeneous vectors are identical if they differ by a *positive* scale factor. But the ambiguity is not critical for our application, so we will keep the definition in equation (2.2).

The link between an original Euclidean space $\mathbb{R}^n$ and the projective space $\mathbb{P}^n$ is now quite easy: a simple mapping $\mathbb{R}^n \to \mathbb{P}^n$ of the Euclidean space to the projective space can be written as follows:

$$\boldsymbol{x} = (x_1, \ldots, x_n)^\mathsf{T} \in \mathbb{R}^n \to \mathbf{x}_e = (x_1, \ldots, x_n, 1)^\mathsf{T} \in \mathbb{P}^n \tag{2.3}$$

We will denote vectors from $\mathbb{R}^n$ with italic bold letters such as $\boldsymbol{x}, \boldsymbol{y}$, whereas homogeneous vectors from $\mathbb{P}^n \simeq \mathbb{R}^{(n+1)}$ are denoted with upright bold letters, $\mathbf{x}, \mathbf{y}$.

The index e in the homogeneous vector $\mathbf{x}_e$ denotes that we can directly extract the Euclidean information from the homogeneous vector. It is a unique member of the equivalence class defined by $\mathbf{x} \cong \mathbf{x}_e$. In figure 2.1(c) one may interpret the fixing of the last coordinate of the homogeneous vector to 1 as follows: take the line parallel to the first axis, intersecting the second axis in 1. This line represents the real axis $\mathbb{R}^1$, on which the Euclidean numbers $\boldsymbol{x}_{1/2}$ are fixed. Connecting the origin with $\boldsymbol{x}_{1/2}$ on the constructed line yield the projective points.

For the inverse direction of (2.3), any projective point $\mathbf{x} \in \mathbb{P}^n$ (except of ideal points) can be mapped to an euclidized version, see also figure 2.2:

$$\mathbf{x} = (x_1, \ldots, x_n, x_{n+1})^\mathsf{T} \in \mathbb{P}^n \to \mathbf{x}_e = 1/x_{n+1} \, (x_1, \ldots, x_n, 1)^\mathsf{T} \in \mathbb{P}^n \tag{2.4}$$

For lower dimensions, we will replace the coordinates x_1, x_2, x_3 with the letters u, v, w to avoid overloaded use of indices.

Linear Transformations of Projective Points A linear transformation H from $\mathbb{P}^n$ to $\mathbb{P}^m$ can be represented as a $(m+1) \times (n+1)$, such that for $\mathbf{x} \in \mathbb{P}^n$ and $\mathbf{x}' \in \mathbb{P}^m$:

$$\mathbf{x}' = \mathsf{H}\mathbf{x} \tag{2.5}$$

If $n = m$ and the matrix H is regular, we call H a *homography*.

Properties that remain invariant under a transformation H are called projectively invariant. Examples for such properties are collinearity and incidence.

definition of the equivalence class $\cong$ for two homogeneous vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n+1} \setminus \mathbf{0}$:

$$\mathbf{x} \cong \mathbf{y} \quad \Leftrightarrow \quad \left(\frac{\mathbf{x}}{|\mathbf{x}|} = \frac{\mathbf{y}}{|\mathbf{y}|} \text{ or } \frac{\mathbf{x}}{|\mathbf{x}|} = -\frac{\mathbf{y}}{|\mathbf{y}|} \right) \quad \Leftrightarrow \quad \frac{\mathbf{x}}{|\mathbf{x}|} = \pm \frac{\mathbf{y}}{|\mathbf{y}|} \tag{2.2}$$

We sometimes denote the spherical normalization by $\mathsf{N}_s(\mathbf{x}) = \frac{\mathbf{x}}{|\mathbf{x}|}$.

The definition of the equivalence class $\cong$ is ambiguous as the orientation of the normalized vectors can be positive or negative. In other words, a projective point from $\mathbb{P}^n$ has two representations on a sphere S^{n+1}. To overcome this ambiguity one may use *oriented* projective geometry, as proposed by STOLFI 1991, where every element has an intrinsic orientation. This is analytically expressed by the specialized definition, that homogeneous vectors are identical if they differ by a *positive* scale factor. But the ambiguity is not critical for our application, so we will keep the definition in equation (2.2).

The link between an original Euclidean space $\mathbb{R}^n$ and the projective space $\mathbb{P}^n$ is now quite easy: a simple mapping $\mathbb{R}^n \to \mathbb{P}^n$ of the Euclidean space to the projective space can be written as follows:

$$\boldsymbol{x} = (x_1, \ldots, x_n)^\mathsf{T} \in \mathbb{R}^n \to \mathbf{x}_e = (x_1, \ldots, x_n, 1)^\mathsf{T} \in \mathbb{P}^n \tag{2.3}$$

We will denote vectors from $\mathbb{R}^n$ with italic bold letters such as $\boldsymbol{x}, \boldsymbol{y}$, whereas homogeneous vectors from $\mathbb{P}^n \simeq \mathbb{R}^{(n+1)}$ are denoted with upright bold letters, $\mathbf{x}$,$\mathbf{y}$.

The index e in the homogeneous vector $\mathbf{x}_e$ denotes that we can directly extract the Euclidean information from the homogeneous vector. It is a unique member of the equivalence class defined by $\mathbf{x} \cong \mathbf{x}_e$. In figure 2.1(c) one may interpret the fixing of the last coordinate of the homogeneous vector to 1 as follows: take the line parallel to the first axis, intersecting the second axis in 1. This line represents the real axis $\mathbb{R}^1$, on which the Euclidean numbers $\boldsymbol{x}_{1/2}$ are fixed. Connecting the origin with $\boldsymbol{x}_{1/2}$ on the constructed line yield the projective points.

For the inverse direction of (2.3), any projective point $\mathbf{x} \in \mathbb{P}^n$ (except of ideal points) can be mapped to an euclidized version, see also figure 2.2:

$$\mathbf{x} = (x_1, \ldots, x_n, x_{n+1})^\mathsf{T} \in \mathbb{P}^n \to \mathbf{x}_e = 1/x_{n+1} \, (x_1, \ldots, x_n, 1)^\mathsf{T} \in \mathbb{P}^n \tag{2.4}$$

For lower dimensions, we will replace the coordinates x_1, x_2, x_3 with the letters u, v, w to avoid overloaded use of indices.

Linear Transformations of Projective Points A linear transformation H from $\mathbb{P}^n$ to $\mathbb{P}^m$ can be represented as a $(m+1) \times (n+1)$, such that for $\mathbf{x} \in \mathbb{P}^n$ and $\mathbf{x}' \in \mathbb{P}^m$:

$$\mathbf{x}' = \mathsf{H}\mathbf{x} \tag{2.5}$$

If $n = m$ and the matrix H is regular, we call H a *homography*.

Properties that remain invariant under a transformation H are called projectively invariant. Examples for such properties are collinearity and incidence.

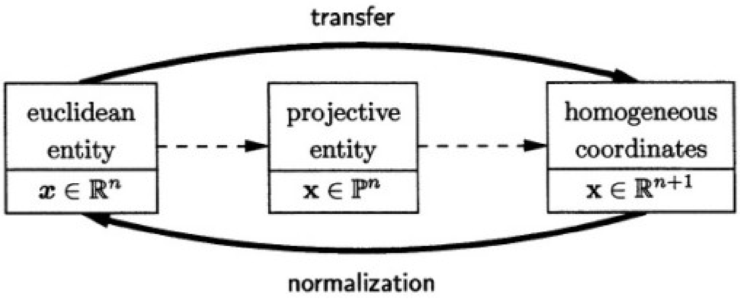
transfer
euclidean entity
$\boldsymbol{x} \in \mathbb{R}^n$
projective entity
$\mathbf{x} \in \mathbb{P}^n$
homogeneous coordinates
$\mathbf{x} \in \mathbb{R}^{n+1}$
normalization

Fig. 2.2. Transfer between Euclidean and Homogeneous Coordinates: giving a Euclidean point $\boldsymbol{x} \in \mathbb{R}^n$ we can derive a projective point $\mathbf{x} \in \mathbb{P}^n$ which we represent as a homogeneous vector $\mathbf{x} \in \mathbb{R}^{n+1}$, see also equation (2.3). Given a homogeneous vector $\mathbf{x} \in \mathbb{R}^{n+1}$, we may obtain a Euclidean representation by normalization, yielding $\boldsymbol{x} \in \mathbb{R}^m$, equation (2.4). In general, the normalization is done by dividing the homogeneous vector by its homogeneous part as described in section 2.2. Note that we use the symbol $\mathbf{x}$ both for a projective point and for homogeneous vectors representing the projective points.

2.1.2 Hyperplanes

A point in projective space can be regarded as the smallest linear subspace that is non-zero. Opposed to a point, a *hyperplane* is the largest linear subspace that is not equal to the complete space. Examples for hyperplanes are lines in 2D and planes in 3D.

In general, a hyperplane $\mathbf{h}$ in $\mathbb{P}^n$ is defined by n linearly independent points $\mathbf{y}_1, \cdots, \mathbf{y}_n$: for example in $\mathbb{P}^2$, two points $\mathbf{y}_1, \mathbf{y}_2$ define a line and in $\mathbb{P}^3$ three points $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3$ define a plane. Any additional point $\mathbf{x}$ is incident to the a hyperplane $\mathbf{h}(\mathbf{y}_1, \cdots, \mathbf{y}_n)$ if the determinant of all concatenated homogeneous vectors is zero

$$|\mathbf{y}_1\, \mathbf{y}_2\, \cdots \mathbf{y}_n\, \mathbf{x}| = 0 \tag{2.6}$$

This is a generalization of the well-known case when checking whether three 2D points are collinear BRONSTEIN *et al.* 1996:

$$\boldsymbol{y}_1, \boldsymbol{y}_2 \text{ and } \boldsymbol{x} \text{ are collinear} \quad \Leftrightarrow \quad \begin{vmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \boldsymbol{x} \\ 1 & 1 & 1 \end{vmatrix} = 0 \tag{2.7}$$

which can be generalized to the to the 3D case when checking whether four points are coplanar:

$$\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3 \text{ and } \boldsymbol{X} \text{ are coplanar} \quad \Leftrightarrow \quad \begin{vmatrix} \boldsymbol{Y}_1 & \boldsymbol{Y}_2 & \boldsymbol{Y}_3 & \boldsymbol{X} \\ 1 & 1 & 1 & 1 \end{vmatrix} = 0 \tag{2.8}$$

(2.9)

An alternative definition of a hyperplane can be obtained by using the Laplacian expansion on the determinant in equation (2.6): if we expand by the last column, we obtain $n-1$ minors $|\mathsf{M}_{i,n}|$ that depend on the vectors $\mathbf{y}_1, \cdots \mathbf{y}_n$. With $h_i = (-1)^{i+n}|\mathsf{M}_{i,n}|$, equation (2.6) is equivalent to

$$\sum_i h_i x_i = \mathbf{h}^\mathsf{T}\mathbf{x} = \mathbf{x}^\mathsf{T}\mathbf{h} = 0 \tag{2.10}$$

Both (2.6) and its equivalent equation (2.10) is a necessary and sufficient condition to check whether a point $\mathbf{x}$ is incident to a hyperplane, $\mathbf{x} \in \mathbf{h}$.

Since the minors h_i with $i = 1, \cdots, n$ uniquely define a hyperplane, we can identify them with the coordinates of a hyperplane, defining the vector $\mathbf{h} = (h_1, \cdots, h_n)^\mathsf{T}$. The coordinates h_i are called Plücker coordinates, see also section 2.2.4. The vector $\mathbf{h}$ is a homogeneous vector as (2.10) is a homogeneous equation. As for projective points, it is assumed that it is clear from the context whether a homogeneous vector is interpreted as a projective point or a hyperplane: in the following lines in 2D are usually denoted by $\mathbf{l}$, $\mathbf{m}$ and planes in 3D are denoted by $\mathbf{A}, \mathbf{B}$ opposed to points denoted by $\mathbf{x}$, $\mathbf{y}$ resp. $\mathbf{X}, \mathbf{Y}$.

It is important to note that $\mathbf{h}$ is actually a linear form of $\mathbb{R}^{n+1}$, since each coordinate is a function of the form $\mathbb{R}^{n+1} \to \mathbb{R}$. Thus the vector $\mathbf{h}$ represents an element of the dual space $(\mathbb{R}^{n+1})^*$ or equivalently $(\mathbb{P}^n)^*$.

Given a linear transformation $\mathbf{x}' = \mathsf{H}\mathbf{x}$ for projective points as in (2.5), one may be interested in the corresponding transformation for hyperplanes. H is a linear mapping from $\mathbb{P}^n$ to $\mathbb{P}^m$, its transpose H^T is defined as the mapping from $(\mathbb{P}^m)^*$ to $(\mathbb{P}^n)^*$. With $\mathbf{h}' \in (\mathbb{P}^m)^*$ being the imaged hyperplane of $\mathbf{h} \in (\mathbb{P}^n)^*$ with respect to H, one obtains

$$\mathsf{H}^\mathsf{T}\mathbf{h}' = \mathbf{h} \tag{2.11}$$

If H is a homography, thus $n = m$ and H being regular, one can prove (2.11) as follows: given $\mathbf{x}' = \mathsf{H}\mathbf{x}$, we are looking for the corresponding transformation $\mathbf{h}' = \mathsf{J}\mathbf{h}$, such that for all $\mathbf{y} \in \mathbf{h}$ and $\mathbf{y}' \in \mathbf{h}'$: $\mathbf{y}' = \mathsf{H}\mathbf{y}$. Thus the following equation must hold

$$\mathbf{y}'^\mathsf{T}\mathbf{h}' = \mathbf{x}'^\mathsf{T}\mathsf{J}\mathbf{h} = \mathsf{x}^\mathsf{T}\mathsf{H}^\mathsf{T}\mathsf{J}\mathbf{h} = 0$$

This can be true only if $\mathsf{J} = \mathsf{H}^{-\mathsf{T}}$.

2.2 Representation of Geometric Entities

This section introduces the basic geometric entities, that will be used throughout the rest of this text. As we want to reconstruct polyhedral objects, it is sufficient to deal with geometric entities that can be modeled as linear subspaces, cf. section 1.3.2 on page 6. Therefore only points, lines and planes

<. ! " # / 9</99'1 2
!? '
GG3 H G G 3 # 2
2 8 '

. ! " # / D1? > 5'1 2
8M"' - . H 8 # G #

. ! " # /
3% J 1 #J 2J1J J . ; J4 J4 " J7 '1 2

. ! " # / 1 1 '1 2
!? '
R # 1 # % "% % "

. ! " # / $ # % +J 2 # J '1 2

9

20211117 (cont). ⇓ ✡ Escanear daqui para frente.

Teoria dos Conjuntos Causais

↳ Big Bang nos deu origem ao Universo (ocorreu, mas sendo uma fase do que ocorreu no Universo).

—

Total de páginas escaneadas:

1284	(4 Volumes)
1482	(Antigo-001)
117	(Antigo-002)
2.883	pages

—

Ver:

Data-NovoVivo-022-2021/20211108/Geometry.

—

Under Jerusalem:
The Buried History of the World's Most Contested City.
Andrew Lawler

—

Mathematics2C@gmail.com ⇒ 38,878.29 GB (37.97 TB).

—

20211118|||

Martin Kneser

"über den Rand von,
Parallelkörpern" 30.9.1950

—

mathematik.de

—

mathematik.de/images/DMV/
Jahresberichte/
Jahresberichte_Archiv/
Jahresbericht_01-2006.pdf

—

Obs: Tradução:
"Acima da borda de
corpos paralelos"

—

The Carl Friedrich Gauss
Prize

spektrum.de

—

gogeometry.com

data · ·
S T Q Q S S D

20211119

Eu não sou dono da verdade,
mas eu sou amigo dela.

Texto sagrado hindu
Bhagavad Gita:
"Agora eu me tornei a morte,
a destruidora de mundos!"

→ Oppenheimer declarou citando esta parte do texto.

20211119/Geometry-Illuminated/
Geometry_Illuminated...OCR.pdf.
Pag. 115 (132).
Chapter 12
Parallel Projection.

20211120

Data-Novo-Vivo-021-2021/ARQ/
Analytical Geometry 2D and 3D.pdf
Pag. 1 Chapter 13
Straight Line

Other ref:

Books-022-2021/20211119/
/Analytical Geometry 2D and 3D/

1) Analytical Geometry 2D and 3D.pdf
2) euclidean and non-euclidean geometries..pdf
Pag. 19(37) –
The Euclidean Parallel Postulate →
→
www.archive.org/details/@mirtitles
3) problems-in-higher-mathematics-minorsky.pdf
Pag. 216(221) Chapter 10
Curvature of Plane and Space Curves.
4) Topics in theory of surfaces in elliptic space.pdf
Obs: Muito bom!!!
5) Shreider - What is Distance?..pdf
Obs: Legal!!!

20211119 / Geometry - Important /

1) Complex-Numbers-in-Geometry.pdf

2) Pogorelov-Analytical Geometry.pdf

Pág. 119 Chapter 6

A plane and a straight line

3) Uspensky-Pascal's Triangle.pdf

4) Pogorelov-Geometry-Mir-1987.pdf

5) Modenov-Problems-in-Geometry-Mir-1981.pdf

Pág. 82 (86) - Chapter III

The use of complex numbers in plane geometry.

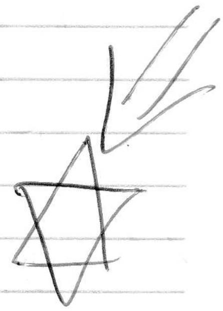

Important:

Estudar Chapter 4 -

Inversion (≅ Mirror)

⇒ analisar

relações com Projective Geometry.

20211120/ Géométrie-M. Chasles/

Questions_de_géométrie....pdf

Pág. 90(110)-

6. Relations métriques dans le cercle.

Pág. 92(112) - 190.

Théorème de Maclaurin

20211121/

Data-NovoVivo-022-2021/

Books-022-2021/

20211121/

Practical Linear Algebra A Geometry Toolbox/

Practical Linear Algebra - A Geometry Toolbox.pdf

Pág. 177(196)

Chapter 8 - 3D Geometry

20211121/

2nd IMA Conference on Mathematics of Robotics/

2nd IMA Conference on Mathematics of Robotics....pdf

Pág. 25(43) - Synthesis of Planar Stiffness.

20211122///

H_2O 104,5° ⇒ 104° 30'

$$\begin{array}{r} 360° \times \\ 60' \\ \hline 21600' \end{array} \qquad \begin{array}{r} 104 \times \\ 60 \\ \hline 6240' + \\ 30' \\ \hline 6270 \end{array} \qquad \begin{array}{r} 21600 \div \\ 6270 = \\ 3,44\ldots \end{array}$$

$3,\overline{444976076555023923}$

20211123//

1 rad = 360°/2π =

57°.29578 =

206265 arcseconds

Pág. 04 (19)

Source:

045--- Gerald North—
Advanced amateur
astronomy
(1997) (.pdf

Source: Books-022-2021/20211123/
045---Gerald North—
Advanced amateur astronomy/

H_2O → 104,45° 104°27'

ângulos entre os átomos.

21600

$$\begin{array}{r} 104 \\ \times\ 60 \\ \hline 6240 \\ +\ 27 \\ \hline 6267 \end{array}$$

21600 ÷ 6267 =

3,4466...

—

20211124//

Pág. 124 do Caderno-001--20211112.pdf

Veja Pág 40 → 090--Hadamard...pdf

20211125//

Soldner. ⇒ Pág. 161-172

(170) (181)

00006.djvu
00009
00170
↓
00181

} converter.

—

Einstein copiou Soldner (1801)
Publicado 1804.

Pág. 161 (Item 104)

data
S T Q Q S S D

20211126///

terminei o item 106 - Caderno-001.

Próx item 107 - Pág. 176

—

Incluir imagens

Livros - Novos

—

20211127///

Fazer Down ARQ_IMPORTANT_20200207.7z

—

Caderno-001.odt → Pág. 203

—

20211128// Aniversário da minha querida Mãe Leda

multcloud.com/help/remote-upload.html

pcloud.com

rclone.~~com~~ - rclone.org

netdrive.net

Hoje é um dia importante

Happy Birthday.

data

S T Q Q S S D

Hoje é aniversário de minha querida Mãe Leda

90 anos

Que tristeza que ela não esteja mais aqui entre nós, mas ela está no céu!!!

Ela faleceu:
17 janeiro 2013

Ela está presente dentro de mim, metade do meu DNA é dela.

Portanto, ela sempre estará presente em minha vida.

Deus ama a todos!!!
Piedoso é
Misericordioso para com todos!!!

20211129 M

Estudar:

Zip-02/Others-Books/Importantíssimo-09/

Experiencing Geometry (2004).pdf

Pág. 125 (138)

Parallel Postulates

Letter:

Einstein to Hale Oct 14, 1913

→ Lichtstrahl

midu)

(Am Sonnenrande müsste

diese Ablenkung 0,84"

→ distorção

betragen und wie $\frac{1}{R}$

abnehmen

20211130 M

Bornomean nucleus.

sciencenews.org.
www.ibiblio.org

Symmetry - invertible
amphicheiral or
interchangeable.

20211201M
cissoid of a circle.

The Trisector by Underwood Dudley.

20211202M
forumgeom.fau.edu

20211203M
A treatise on Solid Geometry.
Frost

(HD Preto) 500 GbM
Books-021-2021 (completo) → zipar

Volume-005-20211111/Books/Gaussian-Quadrature

---

radzevich@usa.com

eaton.corp-

gearsolutions.com

Livros R$ 1915 170

98 145

Sebo → 2013 1600

48 1915

2061

20211204//

www.gazeta.spm.pt

20211205

Download PC-0.

---

fragment of energy.

✡ Larry M. Silverberg

precisa de mais ajustes!!! North Carolina State Univ.

20211206//

digitale.bibliothek@bsb-munchen.de

- Cours de géométrie infinitésimale
  G. Demartres
- Leçons sur la géométrie des espaces de Riemann
  Élie Cartan
- Leçons sur le calcul des variations
  Jacques Hadamard

$a^{\frac{m}{n}} = \sqrt[n]{a^m}$

20211207,111

e-rara.ch

e-periodica.ch

numdam.org

eudml.org

numdam.org/item/OM/
sites.mathdoc.fr/OEUVRES/

sites.mathdoc.fr/JMPA/

$$\frac{a^m}{a^n} = \frac{1}{a^{n-m}}$$

$$\left(a^{\frac{1}{n}}\right)^n = \left(a^n\right)^{\frac{1}{n}} = a$$

20211208 ///

Projective Geometry Coxeter

the Real Projective Plane Coxeter

Curve Schemes Speciali- Important

→ diretório

Obs: Bronchi

20211209 ///

até (parte final)

20211218 ///

Obs: Digitalização do meu material de pesquisa.

20211219.

Trabalho de Digitalização de Todo o meu material de pesquisa:

início: 20211108

término: 20211218 1600 h

obs: Início dos arquivos .odt →

20211021

Brasil na 2ª GM →

BRASIL ~~USA~~ → 12ª Brigada de Infantaria Leve Aeromóvel

+

USA ~~Brasil~~ → 101ª Divisão Aerotransportada.

Em 2016, a divisão brasileira foi denominada como

"Brigada Fornovo de Taro" em decorrência da batalha com o mesmo nome ocorrida na Itália entre os dias 28 Abril à 01 Maio de 1945, onde ocorreu a captura da 148ª Divisão alemã inteira com 15.000 soldados alemães.

20211023

backlash - reação

patronímico

The "open universe" has a saddle-~~shaped~~ shaped negatively curved spacetime.

saddle-shaped → em forma de sela.

Euclid —— Gauss

Helmoltz (questionou?)

extension of Euclidean geometry.

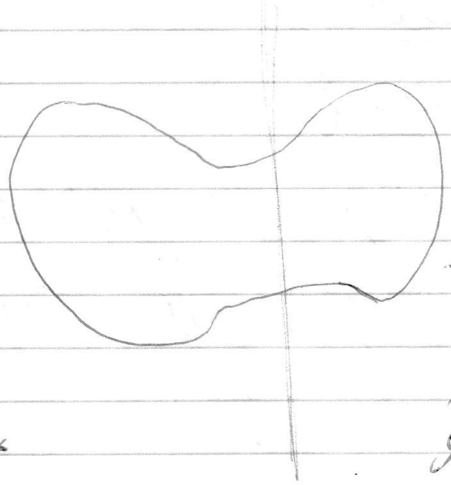

em parte.

esfera comprimida

Veja:

20211222/Important/
The Road to Reality a Complete
guide...pdf
Pág. 250 (279) – Fig. 13.2

Rotational symmetry of
a sphere.

Obs: Estudar com
20211223/Origins the Scientific
Story of Creation/
Jim Baggott - Origins...pdf
Pág. 33 (52)
Figure 5

(b) The "open" universe had a
saddle-shaped negatively curved
spacetime!

data
Ⓢ Ⓣ Ⓠ Ⓠ Ⓢ Ⓢ Ⓓ

20211224 III

musich.info. 50,00

Information Society — Lay All Your Love on Me

—

Formules de Fresnel

Pág. 235 óptique Bruhat

=

cut-the-knot.org

—

20211225 III

Gauss April 30, 1777.
February 23, 1855.

Helmholtz August 31, 1821.
September, 8, 1894.

=

"A matemática é a rainha das ciências: a aritmética é a rainha da matemática".

K. F. Gauss — in: Disquisitiones Arithmeticae

—

data

S T Q Q S S D

stampertech.com.

archmission.org/Nanofiche

↳ 20x20 mm nickel can hold up to 8000 pages of text rendered at 150 dpi.

—

licence Nano IP → Nanofiche

—

lithographically printer.

—

US Patent 8717650
8264757
7961367
7830573

—

digitalpreservation.gov

—

nanoarchival.com

—

25 Dez.

Roma → D deus Mitra, o deus-sol.

↳ bispos.

oração do solstício.

—

Ni Paix Ni Guerre
Hélène Carrière 1986
Flammarion

20211226 |||

- 13 December 2021 - Luz Santana
  Obs: Advogada vizinha da gente veio falar conosco.
- 15 December 2021 - Transferência do Campo Federal para a Administração Federal.
  Neste mesmo dia (20) encontramos o dono da imobiliária que quis comprar de graça a nossa casa.
- 20 December 2021 - Ágne Santana.
- 24 December 2021 - A nova vizinha veio trazer presente de Natal.
  Achamos que ela foi mandada por este dono da imobiliária.

Topologie
Johann Benedict Listing.

---

lindenusd.com (California)
↳ cenic.org
Edlio, Inc

20211229M

Santo Agostinho – sementes
seminais
(matéria + evolucionismo)

Evolucionismo theleológico
↳ teleológico.

Big Bang ↗ Deus criou a matéria de
uma vez e depois infundiu
a evolução
(mudanças na forma
da matéria).

→ Santo Agostinho disse que
Deus criou a matéria e após
isso colocou energia na
matéria para que ela evolua.
Neste caso, não há incompatibilidade
entre evolucionismo e
criacionismo.

—

Darwinismo e Neo-Darwinismo.
↓ seleção natural
↓ alterações cromossômicas criam a evolução.

—

2022: Feliz Ano Novo!!!

There are four fundamental forces at work in the Universe:

(20220101)

1) The strong force
2) The Weak force
3) The Electromagnetic Force
4) The Gravitational Force.

reflexo
(efeito)

The Gravitational Force é decorrência da "The Strong Force".

Se a gravidade existisse independentemente das demais forças, o Universo jamais teria sido criado pois antes do Big-Bang tudo estava reunido em um único ponto e portanto, a própria gravidade teria impedido a própria explosão do Big-Bang.

A gravidade deve ser analisada localmente e quando estendida, você terá outro valor.

A extensão da análise do fator gravitacional em relação a todo Universo gera uma curvatura ~~do~~ espaço-tempo.

O Universo se expande pois a "gravidade" age localmente em relação à cada galáxia e não em relação ao todo.

Albert Einstein and the Cosmic World Order
Cornelius Lanczos

The Intrinsic Nature of things. Barbara Gellai

Cosmology's Century
P. J. E. Peebles

609-258-4386

pjep@Princeton.edu

20220102M

New Shared Drives.

a) Z-Math-Geomath → SAIS Dubai.
(saisdubai.com)

b) Z-Math-New → George L. Egbert
Intermediate School.
is2egbert.~~com~~ org

speedcrunch.org 50 digits

raidrive.com

(raidrive.com)

→ cria uma letra no windows Explorer dos drive compartilhados que você tem no google.

—

~~2021~~

20220103//

Elementary Mathematics from an Advanced Standpoint Geometry

Felix Klein

—

Listar os arquivos no diretório:

20220102/Elementary Mathematics from an Advanced Standpoint.../

Obs. Todos os arquivos.

Converter .djvu para .pdf

—

20171111

20171113

Convex Optimization &
Euclidean Distance Geometry
Dattoro

015 - Richard G. Brown -
Advanced Mathematics...pdf
Pág. 258(300) -

1 radian = $\frac{180}{\pi}$ degrees $\approx$ 57.2958 degrees

1 degree = $\frac{\pi}{180}$ radians $\approx$ 0.0174533 radians

π radians = 180 degrees

Escanear este caderno até esta parte!!!

Obs:

Faz parte do Volume-011---20220104

!"# $ % & ' ( ) & * !
) & *) * + * & & * ,(- . & ' (-*

!"# $ % & ' ( ) & * !
) & *) * + * & & * ,(- . & ' ($ " -*
/% 01
&-
& & 2

!"# $ % & & 3 " ! & (& ' (3
4 & * 4 &) * & 1 -*
/% 05 & % 6

!"# $ % & * & & * * &
% -*
/% 51 7 8 + * & & & &*

$

& 9 % * & :

0 !"# $ % & - . % & &
;(-*
/% <1 &-
9& & & * - & 9

9& & & 9
8 & * "&* 9&
8 +9 & & 9

< !"# $ % & 0 8 - 8 (&
+ - -& % 9 -*
$. 6 9 . ===
/% 107
,! 8+")))
8 & & % > %

8"!? >$" !8)$? $> $?'"@+?8
>)'@"+)? +66) 8) ! + 8$ $?'"@+?8
>)'@"+)? +@ 6)4+!? ! +

/% 01<
8"!? >$" !8)$? $> $?'"@+?8
>)'@"+)? +@ 6)4+!? ! + $?8$
$?'"@+?8 >)'@"+)? +66) 8) ! +

/% 51<0
8"!? >$" !8)$? AB)?>)?)8+) !6
$8)$?

/% 0 1<5
8"!? >$" !8)$? $> 8"!)',8
6)?+ !?4 6!?+

/% 0170
,! 8+")
) &
8 * &9 2 * * -& %
& -& % % - -& &
-& % % + * & -& 3 & *
9 9 &3 &(& (. C * * 9 &
& (% % 3 -& & 3
& 8 - * &
- .
8"!? >$" !8)$? $>) $ +8") @">! +

7 !"# $ % & < * D & 4 & -*

E !"# $ % & 7 - C ? .
' (-*

5 !"# $ % & 7 * 9 .) ' (
5E -*
/% 51 > % 0

IV) of (Ω_3) with respect to (Ω_1), under which inversion the circle (Ω_3) goes into the circle (Ω_4); to construct (Ω_4) it suffices to draw the straight line $O\Omega_3$ and to construct the images P' and Q' of the endpoints P and Q of the diameter that cuts out that line on (Ω_3). The circle constructed on segment $P'Q'$ as a diameter is precisely the circle (Ω_4). Let us compute

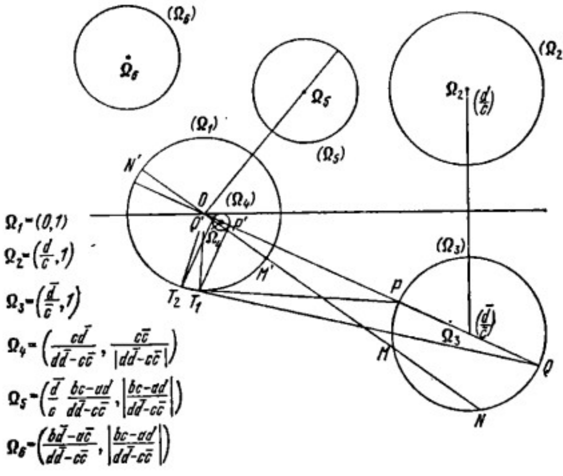

(Ω_6)
Ω_6
Ω_5
(Ω_5)
(Ω_2)
Ω_2
$\left(\frac{d}{c}\right)$
(Ω_1)
N'
O
(Ω_4)
Q'
P'
Ω_4
M'
T_2
T_1
(Ω_3)
P
$\left(\frac{\bar{d}}{\bar{c}}\right)$
Ω_3
M
Q
N
$\Omega_1=(0,1)$
$\Omega_2=\left(\frac{d}{c},1\right)$
$\Omega_3=\left(\frac{\bar{d}}{\bar{c}},1\right)$
$\Omega_4=\left(\frac{c\bar{d}}{d\bar{d}-c\bar{c}},\frac{c\bar{c}}{|d\bar{d}-c\bar{c}|}\right)$
$\Omega_5=\left(\frac{\bar{d}}{c}\,\frac{bc-ad}{d\bar{d}-c\bar{c}},\left|\frac{bc-ad}{d\bar{d}-c\bar{c}}\right|\right)$
$\Omega_6=\left(\frac{b\bar{d}-a\bar{c}}{d\bar{d}-c\bar{c}},\left|\frac{bc-ad}{d\bar{d}-c\bar{c}}\right|\right)$

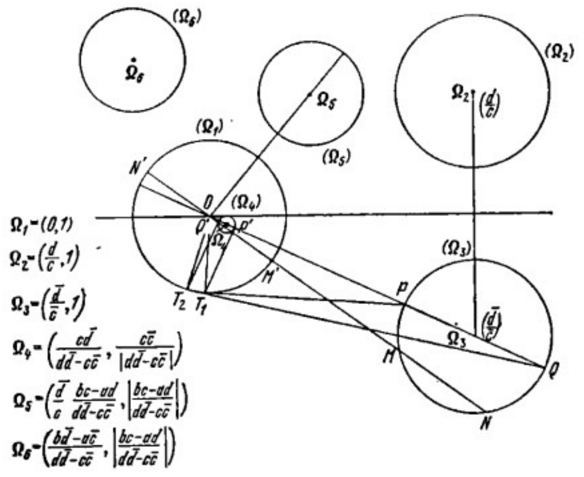
(Ω6)
Ω6
(Ω2)
Ω5
Ω2
(d/c)
(Ω1)
(Ω5)
N'
O
(Ω4)
Ω1-(0,1)
Q'
P'
Ω4
Ω2-(d/c,1)
(Ω3)
M'
Ω3-(d̄/c̄,1)
T2
T1
P
(d̄/c̄)
Ω3
Ω4-(cd̄/(dd̄-cc̄), cc̄/|dd̄-cc̄|)
M
Q
Ω5-(d̄/c̄ (bc-ad)/(dd̄-cc̄), |(bc-ad)/(dd̄-cc̄)|)
N
Ω6-((bd̄-ac̄)/(dd̄-cc̄), |(bc-ad)/(dd̄-cc̄)|)

Fig. 15

the radius of (Ω_4) and the affix of its centre. To do this, note that (Ω_3) passes into (Ω_4) under the inversion with respect to the circle (Ω_1) and under the homothetic transformation $(0,\ 1/\sigma)$, where σ is the power of the point O with respect to the circle (Ω_3).

But under the homothetic transformation $(O,\ 1/\sigma)$, the centre of (Ω_3) goes into the centre of (Ω_4); hence the affix ω_4 of the centre Ω_4 of (Ω_4) is equal to

$$\omega_4 = \frac{1}{\sigma}\,\frac{\bar{d}}{\bar{c}}$$

and the radius is equal to the product of the radius $R_3 = 1$ of the circle (Ω_3) by the modulus of the homothetic ratio, that is, by $1/|\sigma|$:

$$R_4 = R_3\frac{1}{|\sigma|} = \frac{1}{|\sigma|} \quad (R_3 = 1).$$

Furthermore,

$$\sigma = O\Omega_3^2 - R_3^2 = \frac{\bar{d}}{\bar{c}}\,\frac{d}{c} - 1 = \frac{d\bar{d} - c\bar{c}}{c\bar{c}}.^{*}$$

---

[*] In the general case, under inversion with respect to the point O, we have

$$(OM)\cdot(OM') = k,$$

where $k \neq 0$ is the power of the inversion, and, hence,

$$\frac{(OM')}{(ON)} = \frac{k}{\sigma},$$

where k is the second point of intersection of the straight line OM with circle being inverted; $\sigma \neq 0$, since $|d| \neq |c|$.

so that

$$(\Omega_4) = \left(\frac{c\bar{d}}{d\bar{d} - c\bar{c}},\ \frac{c\bar{c}}{|d\bar{d} - c\bar{c}|}\right).$$

(3) The transformation

$$\frac{1}{z + \frac{d}{c}} \to \frac{\frac{bc - ad}{c^2}}{z + \frac{d}{c}}$$

consists in turning the plane through an angle of arg $\frac{bc - ad}{c^2}$ and in a homothetic transformation with centre O and ratio $\frac{|bc - ad|}{|c|^2}$. After these two transformations, the circle (Ω_4) goes into the circle (Ω_5), the affix ω_5 of whose centre is

$$\omega_5 = \omega_4 \frac{bc - ad}{c^2} = \frac{c\bar{d}}{d\bar{d} - c\bar{c}} \frac{bc - ad}{c^2} = \frac{\bar{d}}{c} \frac{bc - ad}{d\bar{d} - c\bar{c}}$$

and the radius is

$$R_5 = \left| \frac{c\bar{c}}{d\bar{d} - c\bar{c}} \cdot \frac{bc - ad}{c^2} \right| = \left| \frac{bc - ad}{d\bar{d} - c\bar{c}} \right|^{*}.$$

Thus

$$(\Omega_5) = \left(\frac{\bar{d}}{c} \frac{bc - ad}{d\bar{d} - c\bar{c}},\ \left| \frac{bc - ad}{d\bar{d} - c\bar{c}} \right| \right).$$

(4) Finally, the transformation of translation

$$\frac{\frac{bc - ad}{c^2}}{z + \frac{d}{c}} \to \frac{a}{c} + \frac{\frac{bc - ad}{c^2}}{z + \frac{d}{c}}$$

carries the circle (Ω_5) into (Ω_6) of the same radius: $R_6 = R_5$, and the affix ω_6 of the centre Ω_6 of (Ω_6) is equal to

$$\omega_6 = \frac{a}{c} + \frac{\bar{d}}{c} \frac{bc - ad}{d\bar{d} - a\bar{c}} = \frac{b\bar{d} - a\bar{c}}{d\bar{d} - c\bar{c}}.$$

---

* $\left| \frac{c\bar{c}}{c^2} \right| = \left| \frac{|c|^2}{c^2} \right| = \frac{|c|^2}{|c^2|} = \frac{|c|^2}{|c|^2} = 1.$

To summarize, then: as a result of the transformation

$$u = \frac{az+b}{cz+d}, \quad ad - bc \neq 0, \quad c \neq 0, \quad |c| \neq |d|,$$

the circle $(\Omega_1) = (0,\ 1)$ goes into the circle

$$(\Omega_6) = \left(\frac{b\bar{d} - a\bar{c}}{d\bar{d} - c\bar{c}},\quad \left|\frac{bc - ad}{d\bar{d} - c\bar{c}}\right|\right).$$

Problem 13. Suppose $\overrightarrow{BCA_1A_2}$, $\overrightarrow{CAB_1B_2}$, $\overrightarrow{ABC_1C_2}$ are squares with the same orientation and constructed on the sides BC, CA, AB of $\triangle\ ABC$, the orthocentre of which is H. Denote by $(O) = (ABC)$ the circle passing through the points A, B, C. Let P, Q, R be the respective centres of the squares $\overrightarrow{A_1B_2C'C''}$, $\overrightarrow{B_1C_2A'A''}$, $\overrightarrow{C_1A_2B'B''}$ having the same orientation as the first three squares. Denote by P_1, P_2, P_3 the orthogonal projections of point P on the sides BC, CA, AB of $\triangle ABC$, by Q_1, Q_2, Q_3 the orthogonal projections of point Q on the same sides and by R_1, R_2, R_3 the orthogonal projections of point R on the same sides BC, CA, AB of $\triangle ABC$. Prove that the sum of the *forces*

$$\overrightarrow{PP_1} + \overrightarrow{PP_2} + \overrightarrow{PP_3} + \overrightarrow{QQ_1} + \overrightarrow{QQ_2} + \overrightarrow{QQ_3} + \overrightarrow{RR_1} + \overrightarrow{RR_2} + \overrightarrow{RR_3} \quad (31)$$

is equal to the vector $\overrightarrow{OH'}$, where H' is the orthocentre of the triangle whose vertices are the feet of the altitudes of the given triangle ABC.

Prove that the straight line on which the resultant of these forces * lies passes through the endpoint of the directed line segment $4\overrightarrow{OH}$.

Now if the squares of the second triplet have an orientation opposite that of the first three squares, then the sum of the forces (31) belongs to the straight line OH'.

Set up the equations of these two straight lines and take the circle $(O) = (ABC)$ as the unit circle (Fig. 16).

Solution. 1°. Let a, b, c be the affixes of the points A, B, C. Find the affixes p, q, r of the points P, Q, R. Denoting by a_1, b_1, c_1, a_2, b_2, c_2 the affixes of the points A_1, B_1, C_1, A_2, B_2, C_2, we have

$$\left.\begin{aligned} b_2 &= c + \overrightarrow{CB_2} = c + i\overrightarrow{CA} = c + i(a - c) = ia + (1 - i)\,c, \\ a_1 &= c + \overrightarrow{CA_1} = c - i\overrightarrow{CB} = c - i\,(b - c) = (1 + i)\,c - ib. \end{aligned}\right\} \quad (32)$$

Also,

$$\begin{aligned} p &= b_2 + \frac{1}{2}\,\overrightarrow{B_2A_1} - \frac{i}{2}\,\overrightarrow{B_2A_1} = b_2 + \frac{1}{2}\,(a_1 - b_2) - \frac{i}{2}\,(a_1 - b_2) \\ &= b_2 + \frac{1}{2}\,a_1 - \frac{1}{2}\,b_2 - \frac{ia_1}{2} + \frac{ib_2}{2} = \frac{1+i}{2}\,b_2 + \frac{1-i}{2}\,a_1 \end{aligned}$$

---

* Force is a sliding vector.

/% E 1E< &-)))
8 - C . - & % (

) - & ===
+ *& &-) 9 1&-- C FG & & &
$.
! & & & & &HI 2 9 ' (

!"# $ % & 7 % 9 ! & (& % (
. 5E -*
/% 5 &- <
! - & & * & & % 6

!"# $ % & 7 % 9 ' (
5E7 -*

!"# $ % & 7 @ - ;(& & J 8 & %
6 -*

!"# $ % & E
K* K%L L K- K KM NK.- <;EE70 7 -*
/% 5 1 < " & L O *&

6. Relations métriques dans le cercle.

189-A. — *Montrer que les diagonales d'un pentagone régulier se divisent en moyenne et extrême raison.* (Bacc. Ajaccio.)

(Ex. 1286.)

189-B. — *On considère un demi-cercle* AMB, *avec ses tangentes parallèles* AC, BD, *et la tangente mobile* CMD. — *Montrer que le produit* AC . BD *est constant.*

(Bacc. Toulouse, Marseille, 1911.)

I. — Les angles COD, AMB, sont égaux comme ayant leurs côtés perpendiculaires.

Le triangle rectangle COD donne :

$$MC . MD = \overline{OM}^2;$$

d'où

$$AC . BD = R^2.$$

II. — *Par l'homographie.* D'après un théorème de Chasles, la tangente mobile marque sur les tangentes fixes deux ponctuelles homographiques, ayant pour points limites les contacts A, B.

On a donc : $AC . BD = C^{te}$.

Cette constante est R^2, comme on le voit quand CD est parallèle à AB.

189-C. — *Un point* M *décrit un cercle* O, *dans lequel on a tracé deux diamètres rectangulaires* AA′, BB′. *La droite* AA′ *coupe* MB *en* P *et* MB′ *en* P′. — *Montrer que le produit* OP . OP′ *est constant.* (Bacc. Clermont, 1909.)

Il suffit de remarquer que les triangles OBP, OB′P′, sont semblables; ou que les droites BP, B′P′, sont antiparallèles par rapport à l'angle O; que les points P, P′, sont conjugués harmoniques sur AA′, etc.

On a : $OP . OP' = R^2$.

189-D. — *Par le milieu* M *de l'arc sous-tendu par une corde* EF, *on mène deux sécantes* MAD, MBC, *qui coupent la droite en* A, B, *et le cercle en* C, D. — *Quelle relation existe-t-il dans le quadrilatère* ABCD, *entre le produit des diagonales et les produits des côtés opposés?* (Bacc. Lyon.)

Ce quadrilatère est inscriptible, car l'angle C et l'angle extérieur en A ont même mesure.

On a donc la relation de Ptolémée :

$$AC . BD = AB . CD + BC . DA.$$

189-E. — *On désigne par* O *le centre du cercle circonscrit à* *un triangle* ABC, *et par* H, M, N, *les projections du sommet* A *sur les droites* BC, OB, OC. — *Montrer que l'on a :*

$$\overline{AH}^2 = BM \,.\, CN. \qquad \text{(Bacc. Besançon.)}$$

Le triangle ABC avec sa hauteur AH donnent :

$$AB \,.\, AC = 2R \,.\, AH.$$

En projetant les côtés AB, AC, sur les diamètres adjacents BOM, CON, on a :

$$\overline{AB}^2 = 2R \,.\, BM,$$

$$\overline{AC}^2 = 2R \,.\, CN.$$

Éliminant AB et AC, on obtient :

$$BM \,.\, CM = \overline{AH}^2.$$

189-F. — *D'un point* A *pris hors d'un cercle* O, *on mène une sécante mobile* AMM', *puis on décrit le cercle* OMM' *qui recoupe* OA *en un point* P. — *Montrer que le produit* PM . PM' *est constant.*

(Bacc. Rennes.)

Le point P est fixe puisque l'on a :

$$AP \,.\, AO = AM \,.\, AM' = C^{te}.$$

Soient R le rayon du cercle O, ρ celui du cercle OMM'; h, h', les distances du point P et du point O à la sécante MM', c'est-à-dire les hauteurs des triangles PMM', OMM', qui donnent :

$$PM \,.\, PM' = 2\rho h;$$

$$R^2 = 2\rho h'.$$

Par division membre à membre, on en déduit :

$$\frac{PM \,.\, PM'}{R^2} = \frac{h}{h'} = \frac{AP}{AO};$$

d'où
$$PM \,.\, PM' = R^2\left(\frac{AP}{AO}\right) = C^{te}.$$

189-G. — *On considère le cercle* O *de diamètre* AHB, *la demi-corde* HM *perpendiculaire à ce diamètre, et les milieux* A', B', *des segments* HA, HB. *Si le cercle de centre* M *et de rayon* MH *coupe le cercle* O *en* C, D, *la corde* CD *passe par le milieu* P *de* MH, *et elle est tangente aux trois cercles qui ont pour diamètres* A'B', HA, HB.

I. — 1° Le rayon OM coupe CD en son milieu E, et le quadrilatère inscriptible HPEO donne :

$$MP \,.\, MH = ME \,.\, MO = \frac{2R \,.\, ME}{2} = \frac{\overline{MC}^2}{2} = \frac{\overline{MH}^2}{2};$$

d'où
$$MP = \frac{MH}{2} = PH.$$

2° Les triangles MAB, PA'B', sont homothétiques.

La corde DC, parallèle à la tangente en M au cercle AMB, est tangente en P au cercle A'PB'.

3° La droite MB étant la bissectrice de HMT, la droite PB′ est la bissectrice de l'angle HPC.

Donc le cercle B′ est inscrit dans l'angle HPC.

De même le cercle A′ est inscrit dans l'angle HPD.

II. — *Par l'axe radical.* (Quest. 997.)

189-H. — *Dans un heptagone régulier* ABCDEFO, *on a :*

$$\frac{1}{AB}=\frac{1}{AC}+\frac{1}{AD}. \qquad \text{(Emmerich, M. 1890.)}$$

Posons : $AB=a, \quad AC=b, \quad AD=c.$

Dans le quadrilatère inscrit ABCE, on a :

$$AB=BC=a, \quad CA=CE=b, \quad EA=EB=c,$$

et le premier théorème de Ptolémée donne :

$$bc=ab+ac,$$

relation équivalente à la proposée.

190. Théorème de Maclaurin. — *Par le sommet* S *d'un angle constant, on mène un cercle variable qui recoupe les côtés en* A, B, *et la bissectrice en* C. — *Montrer que l'on a :*

$$\frac{SA+SB}{SC}=C^{te}.$$

I. — *Triangles semblables.* (Ex. 1291.)

II. — *Calcul trigonométrique.* (Ex. 1292.)

III. — *Similitude.* Soient CD, CE, les distances du point C aux droites SA, SB.

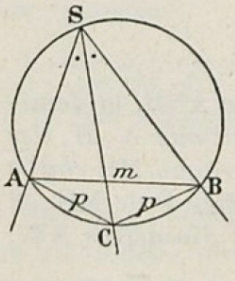

Les triangles rectangles CAD, CBE, sont égaux puisque les cordes CA, CB, sous-tendent des arcs égaux.

On a donc: $SA+SB=2SD.$

Le triangle rectangle SDC reste semblable à lui-même, puisque l'angle $ASB=2\alpha=C^{te}$.

Donc $\dfrac{SA+SB}{SC}=2\dfrac{SD}{SC}=2\cos\alpha=C^{te}.$

IV. — *Par le théorème de Ptolémée.* Appliquons le premier théorème de Ptolémée au quadrilatère SABC. en remarquant que le triangle ABC reste semblable à lui-même, et en posant :

$$\frac{AB}{AC}=\frac{m}{p}.$$

On obtient : $$SC.m=(SA+SB)p;$$

d'où $$\frac{SA+SB}{SC}=\frac{m}{p}=C^{te}.$$

Remarque. — Si le point C est fixe, on a :

$$SC=C^{te}, \quad \text{d'où} \quad (SA+SB)=C^{te}.$$

On retrouve ainsi la question 36.

190-A. — *Dans tout quadrilatère inscriptible, les côtés* a, b, c, d, *les diagonales* m, n, *et la distance* p *des milieux des diagonales vérifient la relation :*

$$(a + c)^2 + (b + d)^2 = (m + n)^2 + 4p^2.$$

Il suffit d'additionner, membre à membre, les relations d'Euler et de Ptolémée :

$$a^2 + b^2 + c^2 + d^2 = m^2 + n^2 + 4p^2;$$

$$ac + bd = mn,$$

après avoir multiplié cette dernière par 2.

190-B. — *D'un point* O *pris sur le cercle* C *de diamètre* OCD, *on décrit un second cercle qui coupe le premier en* A *et* B. *D'un point quelconque* M *du cercle* O, *les points* A, B, D, *se projettent en* E, F, G, *sur le cercle* C. — *Montrer que* MEFG *est un parallélogramme, et que l'on a :*

$$\overline{MG}^2 = \overline{MA} \,.\, \overline{MB}. \qquad \text{(Cambier, N. A. 1878.)}$$

D'après les mesures de l'angle AOB, l'arc AB du cercle O est semblable à chacun des arcs DA, DB. Il s'ensuit que l'arc DE est égal à AE, et que l'arc EF est égal à DB.

Donc ME est parallèle à DF, et MF à DE.

Enfin les triangles MAG, MBG, sont équiangles, et donnent :

$$\frac{MG}{MA} = \frac{MB}{MG}.$$

190-C. Quadrilatère harmonique. — *D'un point* P *à un cercle, on mène les tangentes* PA, PC, *et la sécante* PDB *qui coupe* AC *en* K. — *Montrer que, dans le quadrilatère inscrit* ABCD :

1° *Les côtés adjacents sont proportionnels* (Mannheim);

2° *Les côtés opposés ont un produit constant;*

3° *Les côtés* a, b, c, d, *sont proportionnels à leurs distances* x, y, r, t, *au point* K.

I. — (Ex. 2454.)

II. — 1° Les triangles équiangles PAB, PAD, ou PCB, PCD donnent :

$$\frac{\text{PA ou PC}}{\text{PD}} = \frac{d}{a} = \frac{c}{b}.$$

2° La relation de Ptolémée :

$$ac + bd = AC \,.\, BD,$$

devient :

$$2ac = 2bd = AC \,.\, BD.$$

3° Dans tout quadrilatère inscriptible, les triangles semblables KAB, KCD, ou KAD, KBC, donnent :

$$\frac{x}{a} = \frac{z}{c}, \qquad \text{et} \qquad \frac{y}{b} = \frac{t}{d}.$$

Projetons P en X sur BA, en Y sur BC, on aura :

$$\frac{x}{y} = \frac{PX}{PY} = \frac{PA \sin BCA}{PC \sin BAC} = \frac{a}{b}.$$

Donc $$\frac{x}{a}=\frac{y}{b}=\frac{z}{c}=\frac{t}{d}.$$

III. — *Rapport anharmonique dans le cercle.* (Quest. 714-A.)

IV. — *Inversion.* (Quest. 837.)

190-D. — *D'un point* O *à un cercle on mène des sécantes* OAA′, OBB′, OCC′, ... *formant des angles adjacents égaux.* — *Montrer que tous ces angles interceptent quatre cordes dont le produit est constant.*

Il suffit de considérer deux angles adjacents AOB, BOC.

Soient R le rayon du cercle, et h, h', les distances des points B, B′, aux droites OA, OC.

On a :
$$2Rh = BA \, . \, BA' = BC \, . \, BC';$$
$$2Rh' = B'A \, . \, B'A' = B'C \, . \, B'C'.$$

Multipliant membre à membre, on obtient :
$$AB \, . \, BA' \, . \, AB' \, . \, A'B' = BC \, . \, BC' \, . \, CB' \, . \, B'C'.$$

191. Porisme d'Euclide. — *Par deux points* A, B, *symétriques par rapport au centre d'un cercle* O, *on mène, dans une direction variable, deux segments* AM, BN, *parallèles, de même sens, limités à la circonférence.* — *Démontrer que le produit* AM . BN *est constant.*

Puissance d'un point. (Ex. 1292.)

192. Porisme d'Euclide. — *Autour d'un cercle* O, *on considère la tangente au point fixe* A, *et deux tangentes parallèles, de direction variable, qui coupent la première en des points* M, N. — *Démontrer que le produit* AM . AN *est constant.*

Hauteur d'un triangle rectangle. (Ex. 1293.)

193. Porisme d'Euclide. — *Autour d'un demi-cercle, on considère deux tangentes fixes également inclinées sur le diamètre qu'elles rencontrent aux points* A, B, *et une tangente mobile qui coupe les premières aux points* M, N. — *Démontrer que le produit* AM . BN *est constant.*

Triangles équiangles. (Ex. 1296.)

194. — *Dans un cercle* O, *une corde* CD *pivote autour d'un point* L *qui est fixe sur le diamètre* AB. *La tangente en* A *rencontre* BC *au point* E *et* BD *au point* F. — *Démontrer que le produit* AE . AF *est constant.*

I. — Le quadrilatère ECDF est inscriptible, car les angles BCD, BFE, ont même mesure.

Le cercle ECDF coupe AB en des points fixes M, N; car, par rapport à ce cercle, la puissance du point L est :

$$LA . LB = C^{te},$$

et celle du point B :

$$BD . BF = \overline{AB}^2 = C^{te}.$$

Donc $AE . AF = AM . AN = C^{te}$.

En effectuant le calcul, on trouve :

$$AE . AF = \frac{LA}{LB} \overline{BA}^2.$$

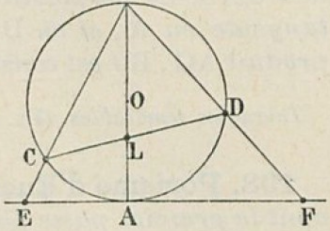

II. — (Ex. 1297.)

III. — *Par l'involution.* Les mobiles C, D, restant alignés sur un point fixe L, leurs projections E, F, à partir d'un point fixe B pris sur le cercle, décrivent une involution, ayant A pour point central.

Donc $AE . AF = C^{te}$.

195. — *On considère sur un cercle deux points fixes* A, B, *et un mobile* M.

Les sécantes MA, MB, *coupent une droite fixe* XY *en des points* C, D.

Trouver sur XY *deux points fixes* I, J, *tels que le produit* IC . JD *soit constant.*

I. — (Ex. 1298.)

II. — *Homographie.* Les droites AM, BM, se correspondent homographiquement. Donc les points C, D, décrivent deux ponctuelles homographiques, dont les *points limites* I, J, répondent à la question.

Le point I est la position limite de C lorsque D s'éloigne indéfiniment. Le point J est la position limite de D, lorsque C disparaît à l'infini.

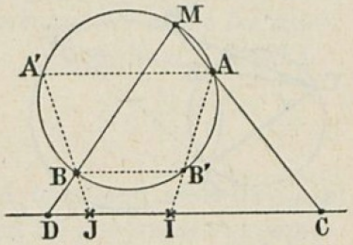

196. — *Par un point fixe* L *pris sur le prolongement du diamètre* AB *d'un cercle donné, on mène une sécante mobile* LCD, *et la droite* LM *perpendiculaire à* AB.

Les sécantes BC, BD, *coupent* LM *aux points* M *et* N. — *Démontrer que le produit* LM . LN *est constant.*

I. — (Ex. 1299.)

II. — (Quest. 194.)

La tangente en A coupe BC, BD, en des points E, F, tels que

$$AE . AF = C^{te}.$$

Cette question ne diffère pas de la question 194, le point L étant intérieur au cercle dans l'un des cas, et extérieur dans l'autre cas.

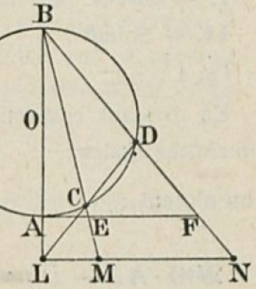

197. Porisme d'Euclide. — *On considère un mobile* M *qui décrit un demi-cercle* AMB, *et un point* P *qui reste fixe sur le diamètre* AB. *La perpendiculaire à l'extrémité de* PM *coupe en* C *la tangente en* A, *et en* D *la tangente en* B. — *Démontrer que le produit* AC . BD *est constant.*

Triangles semblables. (Ex. 1300.)

198. Porisme d'Euclide. — *On considère deux cercles* E, A, *dont le premier passe par le centre du second, et une tangente qui roule sur le second et coupe le premier en des points* M *et* N. — *Démontrer que le produit* AM . AN *est constant.*

I. — *Triangles semblables.* (Ex. 1301.)

II. — Le triangle AMN est inscrit au premier cercle, de rayon R; et sa hauteur reste égale au rayon r du cercle A.

Donc $$AM \cdot AN = 2Rr.$$

199. — *Deux cercles se coupent aux points* A, B. *D'un point* C, *mobile sur le premier, on mène à l'autre la tangente* CD. —

Démontrer que l'on a : $$\frac{\overline{CD}^2}{CA \cdot CB} = C^{te}.$$

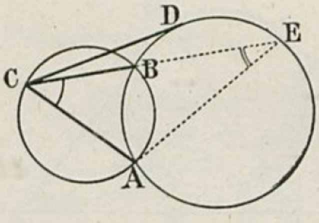

I. — (Ex. 1302.)

II. — La sécante CA recoupe le second cercle en un point E, et le triangle CBE reste semblable à lui-même, puisque ses angles sont constants.

On a donc :

$$\frac{\overline{CD}^2}{CA \cdot CB} = \frac{CA \cdot CE}{CA \cdot CB} = \frac{CE}{CB} = C^{te}.$$

200. — *Par un point* A *mobile sur une circonférence, on mène les cordes* AMB, ANC, *passant par deux points fixes* M, N, *symétriques par rapport au centre.*

Prouver que l'on a : $$\frac{AM}{MB} + \frac{AN}{NC} = C^{te}.$$

I. — *Calcul.* (Ex. 1303.)

II. — Soient : $OA = a, \quad OM = ON = b.$

On a : $$\overline{AM}^2 + \overline{AN}^2 = 2(a^2 + b^2).$$

En divisant respectivement les trois termes de cette égalité par les trois quantités égales : $AM \cdot MB = AN \cdot NC = a^2 - b^2,$

on obtient : $$\frac{AM}{MB} + \frac{AN}{NC} = \frac{2(a^2+b^2)}{a^2-b^2} = C^{te}.$$

200-A. — *D'un point* A *à un cercle* O, *on mène une sécante* *mobile* ABC. *Si les traces* B, C *se projettent sur* OA *en* B′, C′. — *Montrer que l'on a :* $\dfrac{1}{AB'} + \dfrac{1}{AC'} = C^{te}$. (Bacc. Lille.)

I. — *Similitude.* Soit M le milieu de la corde BC.
Les droites antiparallèles donnent :

$$AO \,.\, AB' = AM \,.\, AB;$$
$$AO \,.\, AC' = AM \,.\, AC.$$

Additionnant et multipliant membre à membre, on a :

$$OA(AB' + AC') = 2\overline{AM}^2;$$
$$\overline{OA}^2 \,.\, AB' \,.\, AC' = \overline{AM}^2 \,.\, AB \,.\, AC;$$

d'où, par division membre à membre :

$$\frac{AB' + AC'}{AB' \,.\, AC'} = \frac{2OA}{AB \,.\, AC} = C^{te}.$$

II. — *Polarité.* La polaire du point A coupe BC en un point mobile D, et B′C′ en un point fixe D′, tels que :

$$(AD'B'C') = (ADBC) = -1.$$

Cette relation harmonique peut s'écrire :

$$\frac{1}{AB'} + \frac{1}{AC'} = \frac{2}{AD'} = C^{te}.$$

201. — *La tangente commune extérieure à deux cercles tangents extérieurement, est moyenne proportionnelle entre les deux diamètres.* (Bacc. Nancy, 1903.)

I, II. — *Triangles semblables.* (Ex. 1305.)

III. — *Tangente intérieure.* (Ex. 1305.)

202. — *Si trois cercles tangents deux à deux sont inscrits dans un même angle, leurs rayons* r, r′, r″, *sont en progression géométrique.*

I. — *Triangles semblables.* (Ex. 1308.)

II. — *Divisions homothétiques.* Les centres O, O′, O″, et les contacts C, C′, forment deux divisions homothétiques OCO′, O′C′O″, qui donnent :

$$\frac{r}{r'} = \frac{r'}{r''}.$$

203. — *Quand une circonférence est inscrite et circonscrite à deux polygones réguliers semblables, sa longueur est moyenne proportionnelle entre la circonférence inscrite au plus petit et la circonférence circonscrite au plus grand.*

I. — *Triangles semblables.* (Ex. 1309.)

II. — *Polygones semblables.* Les rayons des polygones étant proportionnels à leurs apothèmes, on a :

$$\frac{R}{\rho} = \frac{\rho}{r}; \quad \text{d'où} \quad (2\pi\rho)^2 = (2\pi R)(2\pi r).$$

BIBLIOTHÈQUE NATIONALE IMPRIMÉS

204. — *Si la corde des contacts de deux tangentes à un cercle pivote sur un point fixe, les inverses des distances de ce point aux tangentes ont une somme constante.*

I, II. — *Calcul.* (Ex. 1312.)

205. — *Si deux cercles égaux se coupent orthogonalement au point* A, *les cordes qu'ils interceptent sur une sécante mobile autour du point* A *ont une somme de carrés constante.*

(Ex. 1313.)

205-A. — *Quand deux cercles* A, B, *sont orthogonaux, leur corde commune* MM′ *est égale à la projection de leur tangente commune* CD *sur la ligne des centres.*

Soient a, b, les rayons.

On a : $\mathrm{MM'} = \dfrac{2ab}{\mathrm{AB}}$.

La tangente a pour longueur : $\mathrm{CD} = \sqrt{2ab}$;

et pour projection : $\dfrac{\overline{\mathrm{CD}}^2}{\mathrm{AB}} = \dfrac{2ab}{\mathrm{AB}} = \mathrm{MM'}$.

206. — *Dans un cercle donné, une corde mobile* AB, *inclinée à 45° sur un diamètre* MN, *est divisée par ce diamètre en deux parties* CA, CB, *dont les carrés ont une somme constante.*

I. — (Ex. 1316.) Les projections des segments CA, CB, sur le diamètre MN, ou sur le diamètre perpendiculaire à MN, sont respectivement égales au sinus et au cosinus d'un même angle.

II. — Soient R le rayon du cercle, et d la distance du centre à la corde AB.

On a : $\mathrm{CA} = \sqrt{\mathrm{R}^2 - d^2} + d, \quad \mathrm{CB} = \sqrt{\mathrm{R}^2 - d^2} - d$;

d'où $\overline{\mathrm{CA}}^2 + \overline{\mathrm{CB}}^2 = 2\mathrm{R}^2$.

III. — Soit B′ le symétrique du point B par rapport à MN. L'angle ABB′ étant égal à 45°, la corde AB′ est le côté du carré inscrit.
Le triangle rectangle ACB′ donne :

$$\overline{\mathrm{CA}}^2 + \mathrm{CB}^2 = \overline{\mathrm{AB'}}^2 = 2\mathrm{R}^2.$$

207. — *Dans deux circonférences concentriques, les distances d'un point mobile sur l'une, aux extrémités d'un diamètre quelconque de l'autre, ont une somme de carrés constante.*

(Ex. 1317.)

208. Théorème de Faure. — *Par rapport à deux cercles orthogonaux* B, C, *tout point de la circonférence décrite sur* BC *comme diamètre, a deux puissances égales et de signes contraires.*

I. — (Ex. 1319.)

II. — (Voir quest. 729-B.)

209. — *On considère deux cercles orthogonaux* B, C, *et un point* M *mobile sur un cercle quelconque ayant pour centre le milieu* O *de* BC.

Montrer que les puissances du point M *par rapport aux cercles* B *et* C, *ont une somme constante.*

I. — (Ex. 1320.)

II. — Cette propriété subsiste quand les cercles A, B, ne sont pas orthogonaux. (Quest. 729-A.)

210. — *On considère trois cercles* A, B, O, *qui passent par les mêmes points* C, D, *et tels que le centre* O *soit le milieu de* AB.

Démontrer que les puissances de tout point du cercle O, *par rapport aux cercles* A *et* B, *ont une somme nulle.*

I. — (Ex. 1321.)

II. — (Quest. 729-B.)

211. Théorèmes d'Archimède. — *Dans tout quadrilatère orthodiagonal inscrit à un cercle donné, les diagonales se partagent en quatre segments dont les carrés ont une somme constante, égale à la somme des carrés des côtés opposés.*

(Bacc. Lyon.)

I. — (Ex. 1323.)

II. — (Ex. de Méc. 1129.)

212. — *Dans un cercle donné, si deux cordes rectangulaires pivotent sur un point fixe, leurs carrés ont une somme constante.*

I. — (Ex. 1325.)

II. — (Ex. de Méc. 1129.)

212-A. — *On considère tous les quadrilatères à diagonales rectangulaires inscrits dans un cercle donné. — Montrer que la somme des carrés des quatre côtés est constante.*

(Bacc. Dijon, 1908.)

I. — Soient a, b, c, d, les quatre côtés; $2m$, $2n$, les diagonales; M, N, P, leurs milieux et leur intersection, formant le rectangle OMPN.

Le théorème d'Euler donne :

$$a^2+b^2+c^2+d^2=4(m^2+n^2+\overline{\mathrm{MN}}^2).$$

Or on a : $m^2+n^2=(\mathrm{R}^2-\mathrm{OM})^2+(\mathrm{R}^2-\overline{\mathrm{ON}}^2)=2\mathrm{R}^2-\overline{\mathrm{MN}}^2.$

Donc $$a^2+b^2+c^2+d^2=8\mathrm{R}^2.$$

II. — *Dans un quadrilatère inscrit à diagonales rectangulaires, on peut construire un triangle rectangle ayant pour hypoténuse le diamètre du cercle circonscrit et pour côtés de l'angle droit deux côtés opposés* AB, CD, *ou* BC, DA. (Quest. 35-A.)

Traçons la corde CA′ parallèle à BD. AA′ est un diamètre. Le côté CD se reporte en BA′, et BC en DA′.

Les deux triangles AA′B, AA′D, répondent à la question.

D'après cela, on a séparément :

$$a^2+c^2=4R^2 \quad \text{et} \quad b^2+d^2=4R^2.$$

212-B. — *Par le centre* O *d'un triangle équilatéral* ABC, *on mène un axe quelconque* Ox, *sur lequel les trois sommets se projettent en* A′, B′, C′. *Montrer que la somme* $(\overline{AA'}^2+\overline{BB'}^2+\overline{CC'}^2)$ *est indépendante de la direction attribuée à* Ox. (Bacc. Dijon.)

Soient $OA=1$ et $xOA=\alpha$.

La somme en question peut s'écrire :

$$\sin^2\alpha+\sin^2\left(\alpha+\frac{2\pi}{3}\right)+\sin^2\left(\alpha+\frac{4\pi}{3}\right)=\frac{3}{2}.$$

(Ex. de Trigonométrie, nº 62.)

212-C. — *Par le centre* O *de deux cercles concentriques* α, β, *on mène un cercle* φ *qui coupe ces cercles aux points* A, A′; B, B′.

1º *Si les rayons* OA, OA′, *coupent* β *en* M, M′, *et si* OB, OB′, *coupent* α *en* N, N′, *les droites* MM′, NN′, *se confondent, et coupent le cercle* φ *aux mêmes points* C, C′.

2º *Les droites* MBF, NB′E, *recoupent le cercle* φ *au milieu* E *de l'arc* AB, *et au milieu* F *de l'arc* ABB′.

3º *Le point* E *est le centre du cercle* AMBN, *le point* F *est le centre du cercle* AMB′N′.

4º *Le point* E *est l'orthocentre du triangle* OMF, *et le point* F *est l'orthocentre de* OME.

5º *Les quadrilatères* OMEN, OMFN′, *sont inscriptibles.*

6º *Les deux couples de points* E, N; F, N′, *sont alignés sur le point* A′.

1º Si le point C est déterminé par MM′, les triangles équiangles OMC, OCA, donnent : $\overline{OC}^2=OA.OM.$

Si le point C est déterminé par NN′, les triangles équiangles ONC, OCB, donnent : $\overline{OC}^2=OB.ON.$

Le point C est le même dans les deux cas, puisqu'on a :

$$OA.OM=OB.ON.$$

D'après ces relations, *la droite* OC *est tangente en* C *à chacun des cercles* CAM *et* CBN.

2º Chacun des triangles OMB, OMB′, est isocèle. En comparant, dans le cercle φ, les mesures de leurs angles égaux, on trouve :

$$\text{arc } EA=EB \quad \text{et} \quad \text{arc } FBA=FB'.$$

3º Le milieu E de l'arc AB appartient à la médiatrice des cordes parallèles MB, AN.

4º Le milieu F de l'arc ABB′ appartient à la médiatrice des cordes parallèles MB′, AN′.

Enfin EF est la médiatrice de AM, puisqu'on a :

$$FB' = FA = FM.$$

5° Dans le cercle φ, les angles NOE, NME, ont même mesure, ainsi que les angles N'OF, N'MF.

6° La droite MM' fait des angles égaux avec les droites NE, NA', et avec les droites N'F, N'A'.

§ II. — Lieux géométriques.

212-A. — *Dans un cercle* O, *on considère deux diamètres rectangulaires* AA', BB', *et un point* M *qui se meut sur* BB' *et se projette en* M' *sur la tangente en* A. — *Lieu du point* P *où se croisent les droites* AM, A'M'.

Les droites AM', MA', se coupent en un point C.
Le point P est le barycentre de A'AC. Donc il décrit une droite parallèle à AC.

212-B. — *On considère un point fixe* A *dans le plan d'un cercle* O, *et un point* M *mobile sur ce cercle. Les deux bissectrices de l'angle* AOM *coupent* AM *en des points* P, P', *dont on demande le lieu.*

D'après la propriété des bissectrices, on a :

$$\frac{AP}{AM} = -\frac{AP'}{AM} = \frac{OA}{OA + OM} = C^{te}.$$

Donc P et P' décrivent deux cercles symétriques par rapport au point A, et homothétiques au cercle O par rapport à ce point A.

212-C. — *Dans un triangle variable* ABC, *l'angle* A *reste fixe, et le côté* BC *se déplace parallèlement à lui-même. Lieux des contacts de la droite* BC *avec les quatre cercles tritangents.*

Toutes les positions de la figure sont homothétiques par rapport au point A. Donc tous ses points décrivent des droites passant par le point A.

212-D. — *On considère deux droites fixes* Ox, Oy, *et un point* M *mobile sur* Ox. *Par ce point* M *on mène le cercle tangent en* O *à* Oy, *et le cercle inscrit dans l'angle* xOy. — *Ces deux cercles se recoupent en un point* P *dont on demande le lieu.*

Toutes les positions de la figure sont homothétiques par rapport au point O. Donc le point P décrit une droite passant par le point O.

213. — *Lieu des points* P *dont les distances à deux droites fixes* OA, OB, *sont dans un rapport donné* k. (Bacc. Dijon, 1905.)

8

3D Geometry

Figure 8.1.
3D objects: Guggenheim Museum in Bilbao, Spain. Designed by Frank Gehry.

This chapter introduces the essential building blocks of 3D geometry by first extending the 2D tools from Chapters 2 and 3 to 3D. But beyond that, we will also encounter some concepts that are truly 3D, i.e., those that do not have 2D counterparts. With the geometry presented in this chapter, we will be ready to create and analyze simple 3D objects, which then could be utilized to create forms such as in Figure 8.1.

8.1 From 2D to 3D

Moving from 2D to 3D geometry requires a coordinate system with one more dimension. Sketch 8.1 illustrates the $[\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3]$-system which consists of the vectors

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \text{and} \quad \mathbf{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Thus, a *vector* in 3D is given as

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}. \tag{8.1}$$

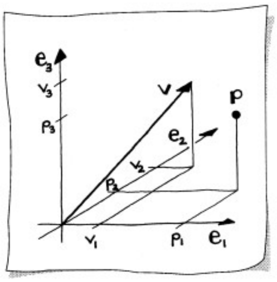

Sketch 8.1.

The $[\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3]$-axes, a point, and a vector.

The three *components* of $\mathbf{v}$ indicate the displacement along each axis in the $[\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3]$-system. This is illustrated in Sketch 8.1. A 3D vector $\mathbf{v}$ is said to live in 3D space, or $\mathbb{R}^3$, that is $\mathbf{v} \in \mathbb{R}^3$. The 3D *zero vector* will be represented as

$$\mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

A *point* is a reference to a *location*. Points in 3D are given as

$$\mathbf{p} = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix}. \tag{8.2}$$

The *coordinates* indicate the point's location in the $[\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3]$-system, as illustrated in Sketch 8.1. A point $\mathbf{p}$ is said to live in Euclidean 3D space, or $\mathbb{E}^3$, that is $\mathbf{p} \in \mathbb{E}^3$. The origin is denoted as $\mathbf{o}$.

All the rules for combining points and vectors in 2D from Section 2.2 carry over to 3D. In Section 2.5, we defined the fundamental operation for points, the *barycentric combination*, and it looks the same in 3D:

$$\mathbf{r} = (1 - t)\mathbf{p} + t\mathbf{q},$$

however now $\mathbf{p}, \mathbf{q}, \mathbf{r} \in \mathbb{E}^3$. The scalar values $(1 - t)$ and t are called the *barycentric coordinates* of $\mathbf{r}$ with respect to $\mathbf{p}$ and $\mathbf{q}$. In Section

2.6 we defined the *linear combination* as a fundamental operation for vectors,

$$\mathbf{u} = r\mathbf{v} + s\mathbf{w},$$

where $\mathbf{v}, \mathbf{w}, \mathbf{u} \in \mathbb{R}^3$ and r, s are scalars. Both of these operations provide the tools to create new geometry in terms of existing geometry. Importantly, they ensure coordinate independent operations.

Sketch 8.2 illustrates a 3D vector $\mathbf{v}$ along with its components. Notice the two right triangles. Applying the *Pythagorean theorem* twice, the *length* or *Euclidean norm* of $\mathbf{v}$, denoted as $\|\mathbf{v}\|$, is

$$\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2 + v_3^2}. \tag{8.3}$$

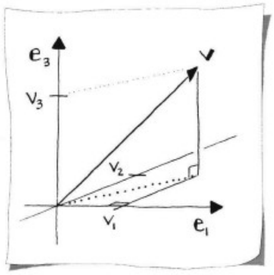

Sketch 8.2.
Length of a 3D vector.

The length or magnitude of a 3D vector can be interpreted as distance, speed, or force.

Scaling a vector by an amount k yields $\|k\mathbf{v}\| = |k|\|\mathbf{v}\|$. Also, a *normalized vector* has unit length, $\|\mathbf{v}\| = 1$.

Example 8.1

We will get some practice working with 3D vectors. The first task is to normalize the vector

$$\mathbf{v} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$$

First calculate the length of $\mathbf{v}$ as $\|\mathbf{v}\| = \sqrt{1^2 + 2^2 + 3^2} = \sqrt{14}$, then the normalized vector $\mathbf{w}$ is

$$\mathbf{w} = \frac{\mathbf{v}}{\|\mathbf{v}\|} = \frac{1}{\sqrt{14}} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \approx \begin{bmatrix} 0.27 \\ 0.53 \\ 0.80 \end{bmatrix}.$$

Check for yourself that $\|\mathbf{w}\| = 1$.

Scale $\mathbf{v}$ by $k = 2$:

$$2\mathbf{v} = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}.$$

Now calculate $\|2\mathbf{v}\| = \sqrt{2^2 + 4^2 + 6^2} = 2\sqrt{14}$. Thus we verified that $\|2\mathbf{v}\| = 2\|\mathbf{v}\|$.

◇□◆

There are infinitely many 3D unit vectors. In Sketch 8.3 a few of these are drawn emanating from the origin. The sketch is a sphere of radius one.

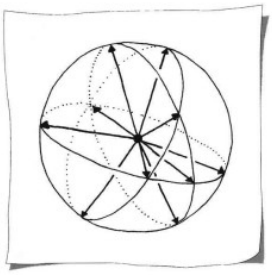

Sketch 8.3.
All 3D unit vectors define a sphere.

In Section 2.7, we derived the dot product for 2D vectors. That development is easy to apply to 3D vectors: the *dot product* of two 3D vectors, $\mathbf{v}$ and $\mathbf{w}$, becomes

$$\mathbf{v} \cdot \mathbf{w} = v_1 w_1 + v_2 w_2 + v_3 w_3.$$

The *cosine* of the angle θ between the two vectors can be determined as

$$\cos\theta = \frac{\mathbf{v} \cdot \mathbf{w}}{\|\mathbf{v}\|\|\mathbf{w}\|}. \tag{8.4}$$

8.2 Cross Product

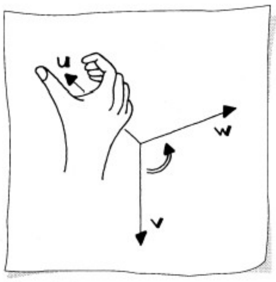

Sketch 8.4.
Characteristics of the cross product.

The dot product is a type of multiplication for two vectors which reveals geometric information, namely the angle between them. However, this does not reveal information about their orientation in relation to $\mathbb{R}^3$. Two vectors define a plane. It would be useful to have yet another vector in order to create a 3D coordinate system that is *embedded* in the $[\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3]$-system. This is the purpose of another form of vector multiplication called the *cross product.*

In other words, the cross product of $\mathbf{v}$ and $\mathbf{w}$, written as

$$\mathbf{u} = \mathbf{v} \wedge \mathbf{w},$$

produces the vector $\mathbf{u}$ which satisfies the following:

1. The vector $\mathbf{u}$ is perpendicular to $\mathbf{v}$ and $\mathbf{w}$, that is
$$\mathbf{u} \cdot \mathbf{v} = 0 \quad \text{and} \quad \mathbf{u} \cdot \mathbf{w} = 0.$$
2. The orientation of the vector $\mathbf{u}$ follows the *right-hand rule.* This means that if you curl the fingers of your right hand from $\mathbf{v}$ to $\mathbf{w}$, your thumb will point in the direction of $\mathbf{u}$. See Sketch 8.4.
3. The magnitude of $\mathbf{u}$ is the area of the parallelogram defined by $\mathbf{v}$ and $\mathbf{w}$.

Because the cross product produces a vector, it is also called a *vector product.*

0 !"# $ % & *) !
& & " . -*
/% 01 (& &

Synthesis of Planar Stiffness

J. M. Selig(✉)

School of Engineering, London South Bank University, London, U.K.
`seligjm@lsbu.ac.uk`

Abstract. In this work the problem of designing systems of springs to achieve a desired stiffness matrix is considered. Only planar configurations are studied. After a brief section outlining the theory of the stiffness of planar systems the planar stiffness matrix of three typical design elements are found, simple springs, beams and pairs of stretched springs. A final section shows how arbitrary stiffness matrices can be achieved using three simple springs or two stretched spring pairs.

Keywords: Stiffness matrix · Springs · Beams

1 Introduction

The synthesis of stiff or compliant systems is an important area of mechanism theory and robotics. There are many previous works in this area, see for example, [3–5]. Most of these concentrate on the problem of synthesising spatial stiffness matrices using so-called simple springs. These simple springs have zero natural length and are un-stretched. In this work the simpler problem of synthesising planar stiffnesses is considered. Using simple springs gives a mathematically interesting problem related to the secant varieties of a certain Veronese variety. More practical are simple beams and pairs of stretched springs, these are also considered.

2 Planar Rigid-Body Geometry

Here the basic geometric entities of points and lines are introduced and the action of the group of rigid-body displacement on these elements is described. As usual, given a coordinate frame, points can be represented as vectors,

$$\boldsymbol{p} = \begin{pmatrix} p_x \\ p_y \end{pmatrix},$$

where p_x and p_y are the coordinates of the points in the fixed coordinate frame. It is often be more convenient to represent points using extended vectors,

$$\tilde{p} = \begin{pmatrix} \boldsymbol{p} \\ 1 \end{pmatrix} = \begin{pmatrix} p_x \\ p_y \\ 1 \end{pmatrix}.$$

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
W. Holderbaum and J. M. Selig (Eds.): IMA 2020, SPAR 21, pp. 25–36, 2022.
https://doi.org/10.1007/978-3-030-91352-6_3

An advantage of this representation is that it extends easily to projective space $\mathbb{P}^2$. This space has homogeneous coordinates, $(x : y : w)$ and two sets of these homogeneous coordinates represent the same point in $\mathbb{P}^2$ if they are of the form $(x : y : w)$ and $(\lambda x : \lambda y : \lambda w)$ for some non-zero number λ. Thinking of points in the plane as points in the projective plane does however, introduce some extra points. The points with third homogeneous coordinate $w = 0$ do not correspond to any points in the plane. They are usually referred to as points at infinity. In the projective plane these ideal points constitute a line.

The group $SE(2)$ acts by the familiar homogeneous representation on the extended points,

$$\begin{pmatrix} \boldsymbol{p} \\ 1 \end{pmatrix} \longmapsto \begin{pmatrix} R & \boldsymbol{t} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{p} \\ 1 \end{pmatrix}$$

where, as usual,

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \quad \text{and} \quad \boldsymbol{t} = \begin{pmatrix} t_x \\ t_y \end{pmatrix}$$

are respectively the rotation matrix and translation vector of the displacement.

In 2 dimensions lines are dual elements to points. The equation for a point to lie on a line is given by,

$$n_x p_x + n_y p_y = d$$

where d is the perpendicular distance from the line to the origin of coordinates and $\boldsymbol{n} = (n_x, n_y)^T$ is the unit normal vector to the line. If lines are represented by 3-vectors of the form,

$$\ell = \begin{pmatrix} \boldsymbol{n} \\ -d \end{pmatrix} = \begin{pmatrix} n_x \\ n_y \\ -d \end{pmatrix},$$

then the relation above can be represented by the pairing,

$$\ell^T \tilde{p} = 0.$$

That is, the point $\tilde{p}$ lies on the line ℓ if and only if $\ell^T \tilde{p} = 0$.

Again, it is often useful to think of these as homogeneous coordinates of a $\mathbb{P}^2$. Notice that ℓ and $\lambda\ell$ give the same line if λ is non-zero, although now $\lambda\boldsymbol{n}$ is no longer a unit vector and λd is not the perpendicular distance from the origin to the line. Thinking of lines as elements of a $\mathbb{P}^2$ means that the "line at infinity" must be included. This is the line with homogeneous coordinates $\ell_\infty = (0, 0, 1)^T$, on which the points at infinity $\tilde{p} = (x, y, 0)^T$, lie.

Thinking of lines as dual to points means that it is easy to see that lines transform under the inverse-transpose representation of the group. That is the dual representation of the group,

$$\begin{pmatrix} \boldsymbol{n} \\ -d \end{pmatrix} \longmapsto \begin{pmatrix} R & 0 \\ -\boldsymbol{t}^T R & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{n} \\ -d \end{pmatrix}$$

Since,

$$\begin{pmatrix} R & \boldsymbol{t} \\ 0 & 1 \end{pmatrix}^{-T} = \begin{pmatrix} R & 0 \\ -\boldsymbol{t}^T R & 1 \end{pmatrix}.$$

Twists are elements of the Lie algebra of the group, $SE(2)$ here. They can be thought of as infinitesimal rigid displacements. The Lie algebra of $SE(2)$ can be found from first principles or simply by abridging the results for the spatial case $SE(3)$, see [8]. So a general twist in the plane can be written as a 3-component vector,

$$\mathbf{s} = \begin{pmatrix} \omega \\ \boldsymbol{v} \end{pmatrix} = \begin{pmatrix} \omega_z \\ v_x \\ v_y \end{pmatrix}$$

where $\boldsymbol{v}$ is the linear velocity of a point instantaneously coincident with the origin and ω_z is the angular velocity of the body or lamina in this case. Under a rigid displacement these transform according to yet another representation of the group,

$$\begin{pmatrix} \omega \\ \boldsymbol{v} \end{pmatrix} \longmapsto \begin{pmatrix} 1 & 0 \\ J\boldsymbol{t} & R \end{pmatrix} \begin{pmatrix} \omega \\ \boldsymbol{v} \end{pmatrix}, \quad \text{where} \quad J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

Again, this can be found by computing the adjoint representation for $SE(2)$ or by looking at the adjoint representation of $SE(3)$ which is more widely known.

In the plane almost all twists are infinitesimal rotations about some point in the plane. So it is possible to associate points and twists. The mapping from twists to points is given by,

$$\mathbf{s} \longmapsto \widetilde{L}\mathbf{s} = \begin{pmatrix} J\boldsymbol{v} \\ \omega \end{pmatrix} = \tilde{p}, \quad \text{where} \quad \widetilde{L} = \begin{pmatrix} 0 & J \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

These matrices satisfy $J^T = J^{-1} = -J$ and $\widetilde{L}^{-1} = \widetilde{L}$.

Notice that, here points in the plane are represented by points in the projective space $\mathbb{P}^2$. Hence the centre of rotation of the twist $\mathbf{s}$ is given by the point $\boldsymbol{c} = (c_x,\, c_y)^T$ where,

$$c_x = v_y/\omega, \quad \text{and} \quad c_y = -v_x/\omega.$$

This shows that, when $\omega = 0$ and the motion is an instantaneous translation, there is no finite centre of rotation. The extension of the map to the projective space $\mathbb{P}^2$ maps infinitesimal translations to points at infinity.

In 3 dimensions the combination of the torque and force acting on a rigid body is called a wrench. Wrenches are dual to twists, in 2-dimensions the abridged version of the wrench will be another 3-component vector this time with the form,

$$\mathbf{w} = \begin{pmatrix} \tau \\ \boldsymbol{f} \end{pmatrix} = \begin{pmatrix} \tau_z \\ f_x \\ f_y \end{pmatrix},$$

where τ_z is the torque, and f_x and f_y the components of the force acting on the lamina. The duality is given by the pairing,

$$\mathbf{w}^T\mathbf{s} = \tau_z\omega_z + f_x v_x + f_y v_y$$

which gives the power, a scalar.

The effect of a rigid displacement on a wrench can be found from the duality. The dual representation is the inverse transpose of the representation for the twists, so given by,

$$\begin{pmatrix} 1 & 0 \\ J\boldsymbol{t} & R \end{pmatrix}^{-T} = \begin{pmatrix} 1 & 0 \\ -R^T J\boldsymbol{t} & R^T \end{pmatrix}^{T} = \begin{pmatrix} 1 & \boldsymbol{t}^T JR \\ 0 & R \end{pmatrix}.$$

The action of a rigid displacement on a wrench is thus,

$$\begin{pmatrix} \tau \\ \boldsymbol{f} \end{pmatrix} \longmapsto \begin{pmatrix} 1 & \boldsymbol{t}^T JR \\ 0 & R \end{pmatrix} \begin{pmatrix} \tau \\ \boldsymbol{f} \end{pmatrix}$$

The line of action of the force gives a natural association between wrenches and lines. The line of action of a wrench $\mathbf{w}$ is given by the map,

$$\mathbf{w} \longmapsto \widetilde{L}^T\mathbf{w} = \begin{pmatrix} J\boldsymbol{f} \\ \tau \end{pmatrix} = \ell.$$

Again this map is defined on projective space, non-zero constants multiplying every coordinate are irrelevant. The line associated with a pure torque is the line at infinity.

Notice that if a screw and a wrench are dual, that is if $\mathbf{w}^T\mathbf{s} = 0$, then the point associated with the twist lies on the line associated to the wrench,

$$\ell^T\tilde{p} = \mathbf{w}^T\widetilde{L}\widetilde{L}\mathbf{s} = \mathbf{w}^T\mathbf{s} = 0,$$

since, $\widetilde{L}^2 = I$. This means that if the lamina is instantaneously rotating about a point on the axis of the applied force then no work is being done on the lamina.

3 Stiffness Matrices

Consider a lamina attached to a number of springs in the plane. Alternatively the lamina could be subject to some conservative force field. Stationary points of the potential energy correspond to equilibria of the lamina. At such an equilibrium configuration the Hessian matrix of the potential is the stiffness matrix of the configuration, see [6]. The Hessian is the matrix of second derivatives at the configuration and this gives an extended form of Hooke's law,

$$\mathbf{w} = K\mathbf{s}$$

where $\mathbf{w}$ is the wrench applied to the lamina, $\mathbf{s}$ is the resulting infinitesimal twist and K the stiffness matrix. The stiffness matrix is a symmetric matrix and from physical considerations is also positive definite.

In the plane, all non-singular stiffness matrices are diagonalisable. This is not the same as saying that any 3×3 symmetric matrix is diagonalisable because we want the stiffness matrix to be diagonalisable by rigid-body displacements, not by arbitrary non-singular matrices or even orthogonal matrices. So we need to know how the group $SE(2)$ acts on stiffness matrices. Since a stiffness matrix K, transforms a twist into a wrench the expression, $\mathbf{s}_1^T K \mathbf{s}_2$ is a scalar, so if the compliant element and the screws all undergo a rigid displacement or we just change the coordinates we are working in, this quantity must remain constant. This implies that under a rigid displacement the stiffness matrix K will transform according to,

$$K \longmapsto \begin{pmatrix} 1 & 0 \\ J\boldsymbol{t} & R \end{pmatrix}^{-T} K \begin{pmatrix} 1 & 0 \\ J\boldsymbol{t} & R \end{pmatrix}^{-1} = \begin{pmatrix} 1 & \boldsymbol{t}^T J R \\ 0 & R \end{pmatrix} K \begin{pmatrix} 1 & 0 \\ -R^T J\boldsymbol{t} & R^T \end{pmatrix}$$

Writing the stiffness matrix in partitioned form,

$$K = \begin{pmatrix} q & \boldsymbol{k}^T \\ \boldsymbol{k} & \Upsilon \end{pmatrix},$$

a pure translation of $\boldsymbol{t}$ will give,

$$K' = \begin{pmatrix} 1 & \boldsymbol{t}^T J \\ 0 & I \end{pmatrix} \begin{pmatrix} q & \boldsymbol{k}^T \\ \boldsymbol{k} & \Upsilon \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -J\boldsymbol{t} & I \end{pmatrix} = \begin{pmatrix} q' & \boldsymbol{k}^T + \boldsymbol{t}^T J \Upsilon \\ \boldsymbol{k} - \Upsilon J \boldsymbol{t} & \Upsilon \end{pmatrix},$$

where $q' = q - \boldsymbol{k}^T J\boldsymbol{t} + \boldsymbol{t}^T J \boldsymbol{k} - \boldsymbol{t}^T J \Upsilon J \boldsymbol{t}$. Choosing $\boldsymbol{t} = -J\Upsilon^{-1}\boldsymbol{k}$, the stiffness matrix becomes,

$$K' = \begin{pmatrix} q' & \mathbf{0}^T \\ \mathbf{0} & \Upsilon \end{pmatrix}.$$

Notice that, with this choice for $\boldsymbol{t}$ we get $q' = q - \boldsymbol{k}^T \Upsilon^{-1} \boldsymbol{k}$. Now we can use standard theory to diagonalise this symmetric 2×2 matrix Υ, with a rotation. This result is well known, but seems to be difficult to find in the literature. When the stiffness matrix is in diagonal form we will assume the diagonal entries are k_1^2, k_2^2 and k_3^2.

Next we turn to the invariants of the stiffness matrix K. That is we seek functions of the entries of K which are invariant with respect to a rigid change in coordinates or rigid displacement. We can expect these to be symmetric polynomials in the diagonal elements of diagonalisation of K. The determinant of K is clearly an invariant since it is an invariant under the group $SL(3)$ it must also be an invariant under the subgroup $SE(2)$. The determinant of the diagonal stiffness matrix is just the product of the stiffnesses, $\det(K) = k_1^2 k_2^2 k_3^2$. Notice that after a translation the bottom right 2×2 block of K is unchanged. A rotation will change this block to,

$$\Upsilon \longmapsto R \Upsilon R^T$$

So rotational invariants of this block will be invariant under all of $SE(2)$. The invariants in question are $\det(\Upsilon) = k_2^2 k_3^2$ and $\mathrm{Tr}(\Upsilon) = k_2^2 + k_3^2$.

Next, a few examples will be considered, the first of these is a simple spring. This is an idealised component but the following two examples are a little more practical.

The simple spring is an idealisation of real springs, simple springs can be thought of as un-stretched springs with zero natural length. These elements are determined by lines in the plane. Given a line $\ell = (n_x,\, n_y,\, -d)^T$, the spring will only generate a wrench in the direction of the line, $\mathbf{w} = k\widetilde{L}\ell$, where k is a constant The wrench produced by the simple spring is only in response to a small motion in the direction of the line, that is $\mathbf{w}^T\mathbf{s} \neq 0$ when $\mathbf{s}$ is a translation in the direction of ℓ but is zero for the orthogonal direction and infinitesimal rotations. This means that we can write the stiffness matrix as,

$$K = \mathbf{w}\mathbf{w}^T = k^2\widetilde{L}\ell\ell^T\widetilde{L} = k^2 \begin{pmatrix} d^2 & dn_y & -dn_x \\ dn_y & n_y^2 & -n_xn_y \\ -dn_x & -n_xn_y & n_x^2 \end{pmatrix},$$

where k^2 is knows as the stiffness constant of the spring. Usually this is stated the other way around—the wrench produces a small motion of the spring.

If we think of lines ℓ as points in a projective plane $\mathbb{P}^2$ and stiffness matrices as points in a projective space $\mathbb{P}^5$ then the above is clearly a Veronese embedding. Hence the space of all possible simple springs is the degree 4, Veronese surface, see [2].

Next we consider simple beams. Again, here we can just restrict the stiffness matrix of the 3D beam, see [7]. The stiffness matrix of the slender planar beam aligned with the x-axis is thus,

$$K = \begin{pmatrix} \frac{EI_y}{l} & 0 & 0 \\ 0 & \frac{EA}{l} & 0 \\ 0 & 0 & \frac{12EI_y}{l^3} \end{pmatrix},$$

where l is the length of the beam, E is its Young's modulus, A its cross-sectional area and I_y the second moment of area. Note that, the stiffness matrix shown here is expressed in a coordinate frame whose origin is located at the centre of the beam and with the x-axis directed along the beam.

The space of these simple beams is a little more complicated than the simple springs. First we need to decide what to hold constant. For example, we could imagine that the beams that we have available are all made of the same material and have the same cross-section and length l. If this is the case then the space of all beams or rather the stiffness matrices of these beams is just the stiffness matrices which can be generated by rigid displacements of the diagonal stiffness matrix above. From the previous section it is clear that the diagonalisation of an arbitrary stiffness matrix is unique. Hence, the space of simple beams must be an embedding of the group manifold of $SE(2)$. This is well known to be isomorphic to the real projective space $\mathbb{P}^3$ with a single line deleted, see [1].

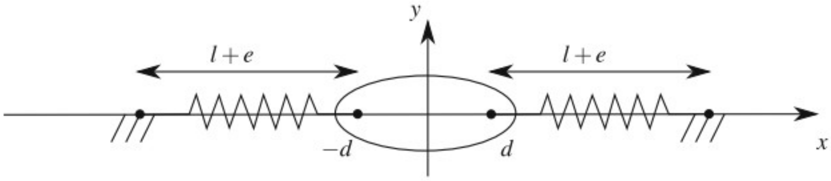

Fig. 1. A pair of stretched springs.

We can find several homogeneous equations satisfied by the stiffness matrices of simple beams. All that is needed is to find relations between the invariants of the matrices. These can be evaluated on the diagonal matrix given above,

$$\begin{aligned}\det(K) &= \frac{12\,AE^3I_y^2}{l^5},\\ \det(\Upsilon) &= \frac{12\,AE^2I_y}{l^4},\\ \mathrm{Tr}(\Upsilon) &= \left(\frac{EA}{l}+\frac{12EI_y}{l^3}\right) = \frac{E}{l^3}(Al^2+12I_y).\end{aligned}$$

These invariants are of degree $3, 2$ and 1 respectively, in the elements of the stiffness matrix. So we can write a quadratic equation,

$$12\,AI_yl^2\mathrm{Tr}(\Upsilon)^2 - (Al^2+12I_y)^2\det(\Upsilon) = 0,$$

and a couple of cubic equations,

$$12\,AI_yl^4\mathrm{Tr}(\Upsilon)^3 - (Al^2+12I_y)^3\det(K) = 0$$

and

$$I_yl^2\mathrm{Tr}(\Upsilon)\det(\Upsilon) - (Al^2+12I_y)\det(K) = 0$$

and finally a degree 6 equation,

$$12\,A\det(K)^2 - l^2I_y\det(\Upsilon)^3 = 0.$$

Presumably these four equations do not give a complete intersection in $\mathbb{P}^5$ but rather define a 3-dimensional rational variety.

If we assume that the simple beams can have different lengths then the space of all simple beams will be 4-dimensional. The equations defining this variety can be found by eliminating l from the equations. Notice that the final degree 6 equation can be rearranged to isolate l^2 and the other three equations only contain l^2 or powers of l^2.

The final example here are stretched spring pairs. Consider the system of springs shown in Fig. 1. The springs have natural length l but in the equilibrium position illustrated are extended by an amount e. The springs are attached to the body at points a distance $2d$ apart. The stiffness matrix of this arrangement can be found from the results in [6],

$$K = 2\lambda e \begin{pmatrix} d(1+\frac{d}{l+e}) & 0 & 0 \\ 0 & \frac{1}{e} & 0 \\ 0 & 0 & \frac{1}{l+e} \end{pmatrix},$$

where λ is the stiffness constant of the springs. It is, of course, possible to find homogeneous equations for the space of all stretched spring pairs using the same method we saw for the space of beams above.

4 Synthesis

Given a stiffness matrix, perhaps the specification of how one part of a machine needs to move with respect to another part, how can we select springs or beams to realise the given stiffness matrix? Unfortunately standard decomposition theorems of symmetric matrices don't apply here. The standard theory is based on eigenvalues and eigenvectors, these only make sense if the matrix represents a linear map from a space to itself or a map between Euclidean spaces. The matrix K here is a map from twist space to wrench space, neither of these spaces is Euclidean.

It is important to note that if these systems of springs are combined in parallel the stiffness matrix of the result is simply the sum of the stiffness matrices of the component systems. Care must be taken however, to ensure that the stiffness matrices are expressed in the same coordinate system before they are added.

At the end of Sect. 3 it was observed that the space of simple springs determine a Veronese surface in the $\mathbb{P}^5$ of all stiffness matrices. This is a two dimensional variety with degree $2^2 = 4$, [2]. By varying the modulus of stiffness of the springs we can produce all possible stiffness matrices lying on the line joining the stiffness matrices of the two simple springs. This is a chord or secant line to the surface. The closure of all such secant lines forms the secant variety to the surface. It is well known that the secant variety to the Veronese surface is a cubic hypersurface in $\mathbb{P}^5$ given by vanishing of a 3×3 determinant. Simple counting of dimensions would suggest that the secant variety to a surface should be 5-dimensional, however for the Veronese surface it is only 4-dimensional. This means that three simple springs will be needed to synthesise a general 2-D stiffness matrix. For singular stiffness matrices only one or two simple springs will be required.

Let K be an arbitrary stiffness matrix, as we saw in Sect. 3, there will be a rigid change of coordinates $M = \begin{pmatrix} 1 & 0 \\ J\boldsymbol{t} & R \end{pmatrix}$, which diagonalises K

$$M^T K M = \begin{pmatrix} k_1^2 & 0 & 0 \\ 0 & k_2^2 & 0 \\ 0 & 0 & k_3^2 \end{pmatrix}.$$

In Sect. 3 diagonalisation was effected using a displacement, a coordinate transformation just uses the inverse transformation.

Now consider the following three lines,

$$\ell_1 = \begin{pmatrix} 1 \\ 0 \\ -1/\sqrt{2} \end{pmatrix}, \quad \ell_2 = \begin{pmatrix} -1/2 \\ \sqrt{3}/2 \\ -1/\sqrt{2} \end{pmatrix}, \quad \ell_3 = \begin{pmatrix} -1/2 \\ -\sqrt{3}/2 \\ -1/\sqrt{2} \end{pmatrix}.$$

These are the edges of an equilateral triangle with the origin as centre. The corresponding wrenches are,

$$\mathbf{w}_1 = \frac{k_1\sqrt{3}}{\sqrt{2}} \begin{pmatrix} -1/\sqrt{2} \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{w}_2 = \frac{k_1\sqrt{3}}{\sqrt{2}} \begin{pmatrix} -1/\sqrt{2} \\ -\sqrt{3}/2 \\ -1/2 \end{pmatrix}, \quad \mathbf{w}_3 = \frac{k_1\sqrt{3}}{\sqrt{2}} \begin{pmatrix} -1/\sqrt{2} \\ \sqrt{3}/2 \\ -1/2 \end{pmatrix}.$$

Notice that all these springs have stiffness constant $k_1\sqrt{3}/\sqrt{2}$. A straightforward computation confirms that,

$$\mathbf{w}_1\mathbf{w}_1^T + \mathbf{w}_2\mathbf{w}_2^T + \mathbf{w}_3\mathbf{w}_3^T = k_1^2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

So we can write

$$M^T K M = \Lambda U(\mathbf{w}_1\mathbf{w}_1^T + \mathbf{w}_2\mathbf{w}_2^T + \mathbf{w}_3\mathbf{w}_3^T)U^T\Lambda$$

where $\Lambda = \begin{pmatrix} 1 & 0 & 0 \\ 0 & k_2/k_1 & 0 \\ 0 & 0 & k_3/k_1 \end{pmatrix}$ and U is an arbitrary 3×3 orthogonal matrix, an element of $O(3)$. Now writing,

$$\mathbf{v}_i = M^{-T}\Lambda U\mathbf{w}_i, \qquad i = 1,\, 2,\, 3$$

we get the decomposition,

$$K = \mathbf{v}_1\mathbf{v}_1^T + \mathbf{v}_2\mathbf{v}_2^T + \mathbf{v}_3\mathbf{v}_3^T.$$

Moreover, this gives all possible such decompositions for a non-singular stiffness matrix K, different choices are given by different orthogonal matrices U.

In terms of simple springs this means that we can decompose a stiffness matrix K into a system of three simple springs placed along the lines,

$$\ell_i' = \widetilde{L}\mathbf{v}_i, \qquad i = 1,\, 2,\, 3,$$

Next we consider stretched spring pairs. For a general stiffness matrix in 2D we need at least two stretched string pairs. As above, we can assume that the stiffness matrix has been diagonalised by some rigid-body displacement M, so that we can write,

$$K = \begin{pmatrix} k_1^2 & 0 & 0 \\ 0 & k_2^2 & 0 \\ 0 & 0 & k_3^2 \end{pmatrix}.$$

Assuming that we cannot change the stiffness of the material that the springs are made from means that we cannot use the secant variety to the space of spring pairs to synthesise the stiffness matrix. A possible strategy is to combine pairs with diagonal stiffness matrices. For example, suppose we rotate the pair by a quarter turn, such a pair can have a different extension e_2 and separation d_2 to the original pair e_1 and d_1. The stiffness matrix of such a perpendicular arrangement of pairs will give a stiffness matrix,

$$K = 2\lambda e_1 \begin{pmatrix} d_1(1+\frac{d_1}{l+e_1}) & 0 & 0 \\ 0 & \frac{1}{e_1} & 0 \\ 0 & 0 & \frac{1}{l+e_1} \end{pmatrix} + 2\lambda e_2 \begin{pmatrix} d_2(1+\frac{d_2}{l+e_2}) & 0 & 0 \\ 0 & \frac{1}{l+e_2} & 0 \\ 0 & 0 & \frac{1}{e_2} \end{pmatrix}.$$

Comparing this with the target stiffness gives 3 equations,

$$\begin{aligned} k_1^2 &= 2\lambda\left(e_1d_1 + e_2d_2 + \frac{e_1d_1^2}{l+e_1} + \frac{e_2d_2^2}{l+e_2}\right), \\ k_2^2 &= 2\lambda\left(1 + \frac{e_2}{l+e_2}\right), \\ k_3^2 &= 2\lambda\left(1 + \frac{e_1}{l+e_1}\right). \end{aligned}$$

The last two of these equations are easily solved,

$$e_1 = \left(\frac{k_3^2 - 2\lambda}{4\lambda - k_3^2}\right) l \quad \text{and} \quad e_2 = \left(\frac{k_2^2 - 2\lambda}{4\lambda - k_2^2}\right) l.$$

These result give limits on the stiffnesses that can be synthesised with such springs, if the extension is to be positive then we must have, $2\lambda < k_2^2 < 4\lambda$ and the same for k_3^2.

The first of the equations above would give many solutions for d_1 and d_2, so it seems reasonable to set $d_1 = d_2 = d$. This gives a quadratic equation for d,

$$\left(\frac{e_1}{l+e_1} + \frac{e_2}{l+e_2}\right) d^2 + (e_1 + e_2)d - \frac{k_1^2}{2\lambda} = 0.$$

From the pattern of sign changes in the coefficients, this has one positive real solution for d, and this will give a stable equilibrium for the system.

An alternative solution might be to use a couple of identical spring pairs but move them relative to each other. Suppose we have two identical diagonal stiffness matrices and we subject them to either equal but opposite translations or equal but opposite rotations. The composite will be a diagonal stiffness matrix. For example, if we rotate the stiffness matrices of two stretched string pairs by θ and compare the combined system with the desired diagonal stiffness we get the following three equations,

$$k_1^2 = 4\lambda\left(ed + \frac{ed^2}{l+e}\right),$$
$$k_2^2 = 4\lambda\left(\cos^2\theta + \frac{e}{l+e}\sin^2\theta\right),$$
$$k_3^2 = 4\lambda\left(\sin^2\theta + \frac{e}{l+e}\cos^2\theta\right).$$

Adding the last two equations eliminates the angle so that we can solve for the extension,

$$e = \left(\frac{k_2^2 + k_3^2 - 4\lambda}{8\lambda - k_2^2 - k_3^2}\right) l.$$

Again, this gives limits on the stiffnesses which can be synthesised using this method with positive extensions e. Subtracting the last two equations gives a linear equation for $\cos(2\theta)$,

$$\cos(2\theta) = \frac{k_2^2 - k_3^2}{8\lambda - k_2^2 - k_3^2},$$

since $l + e = (4\lambda l)/(8\lambda - k_2^2 - k_3^2)$. Finally d can be found as the unique positive real solution to the quadratic equation in d given by the first equation above.

5 Conclusions

Similar techniques to those outlined above could be used to synthesis a desired stiffness from simple beams. We could also consider other compliant design elements. It was possible to find all possible designs using three simple springs, it would be useful to find all possible solutions for pairs of stretched springs and other stiffness elements.

It would also be useful to extend these ideas to 3 dimensional stiffness. However, in 3D a general stiffness matrix cannot necessarily be diagonalised with a rigid-body displacement. This makes the general problem of the synthesis of stiffness in 3D much more difficult.

References

1. Bottema, O., Roth, B.: Theoretical Kinematics. Dover Publications, New York (1990)
2. Harris, J.: Algebraic Geometry a First Course. Springer Verlag, New York (1992). https://doi.org/10.1007/978-1-4757-2189-8
3. Huang, S., Schimmels, J.: The bounds and realization of spatial stiffness achieved with simple springs connected in parallel. IEEE Trans. Robot. Autom. **14**(3), 466–475 (1998)
4. Lipkin, H., Ciblak, N.: Application of stiffness decompositions to synthesis by springs. In: Proceedings. ASME Design Technical Conference, Atlanta, GA (1998)

5. Roberts, R.: Minimal realization of an arbitrary spatial stiffness matrix with a parallel connection of simple and complex springs. IEEE Trans. Robot. Autom. **16**(5), 603–608 (2000)
6. Selig, J.M.: The spatial stiffness matrix from simple stretched springs. In: Proceedings of the IEEE International Conference on Robotics and Automation. San Francisco, vol. 2000, pp. 3314–3319 (2000)
7. Selig, J.M., Ding, X.: A screw theory of beams. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robotics and Systems. Maui, vol. 2001, pp. 312–317 (2001)
8. Selig, J.M.: Geometric Fundamentals of Robotics. Springer Verlag, New York (2005). https://doi.org/10.1007/b138859

< !"# $ % & 0 ' & * ? !*9& *
& & & (& . *% @ 9 (1 557 -*
/% 1 5

For small values of θ we can make the approximation that:

$$\theta_{\text{radians}} = h/f,$$

where θ is measured in radians (see Figure 1.2 for an explanation of radian measure).

Expressing θ in arcseconds:

$$\theta_{\text{arcsec}} = 206\,265 \times h/f$$

From which:

$$\theta_{\text{arcsec}}/h = 206\,265/f$$

Now, θ_{arcsec}/h is the *image scale* – the linear distance in the focal plane that corresponds to a given angular distance in the sky/distant object.

$$\therefore \textbf{ Image scale} = \mathbf{206\,265/}\boldsymbol{f}$$

If the focal length of the lens is measured in millimetres, then the image scale is in arcseconds per millimetre. For example, suppose that the diameter of the Moon subtends an angle of 2000 arcseconds. If the lens has a focal length of 1000 mm then the image scale in its focal plane is, to 3 significant figures, 206 arcseconds per millimetre. The focused image of the Moon will, in this case, be just under 10 mm across.

The image scale in the focal plane actually varies with the distance from the optical axis, though only by a negligible amount for small values of θ.

Aperture, focal ratio and light grasp

The *aperture* (diameter), focal length and *focal ratio* of our hypothetical lens are related by:

$$\text{Focal ratio} = \text{focal length/aperture},$$

the aperture and the focal length being expressed in the same units.

The light collected by the lens is proportional to its area, and is thus proportional to the square of its aperture:

$$\text{Light grasp} \propto (\text{aperture})^2$$

Diffraction and resolution

It is an oversimplification to say that a lens, even a perfect one, will form a truly point image of a point object. In practice the image will consist of a

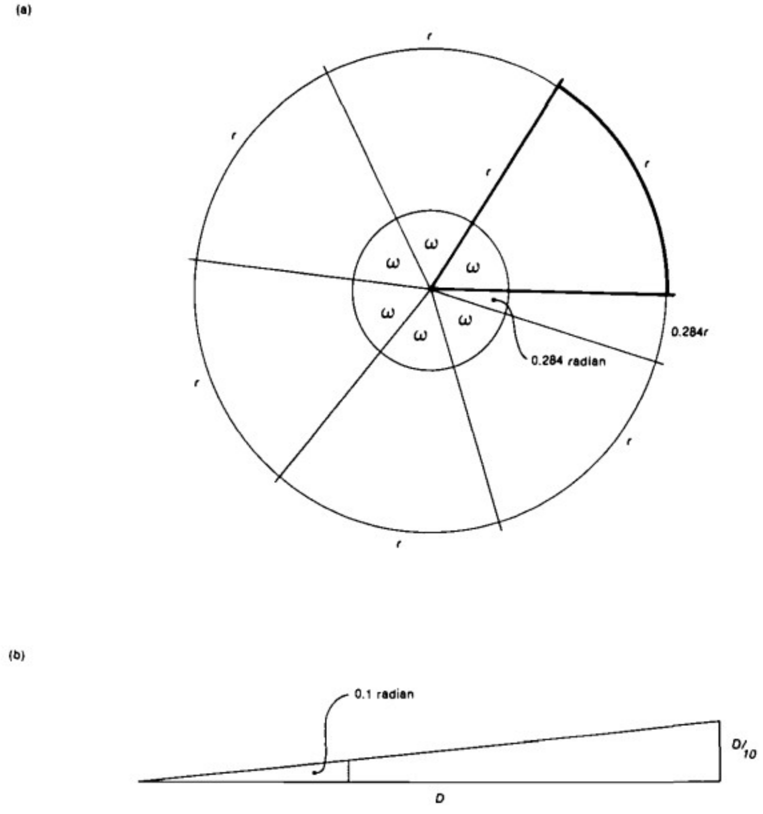

Figure 1.2. Radian measure. Consider a circle of radius r. If an arc of length equal to r is drawn along the circumference, the angle that the arc subtends from the centre of the circle is defined to be 1 radian. This is illustrated in (a) by the boldly outlined sector. In this diagram the angle ω is equal to 1 radian. Since the circumference of a circle is equal to 2π (6.284) times its radius, it follows that there are 2π radians in 360°. Hence:

$$\begin{aligned} 1\ \mathit{radian} &= 360°/2\pi = 57°.29578 \\ &= 206\,265\ \mathit{arcseconds}. \end{aligned}$$

In general, if the arc length is l and the radius of the arc is R, then the angle, θ, is given by:

$$\begin{aligned} \theta\ (\text{in radians}) &= l/R, \\ \text{or } \theta\ (\text{in arcseconds}) &= 206\,265\ l/R \end{aligned}$$

Diagram (b) shows the concept of the radian applied to smaller angles. In this case the angle subtended by a length $D/10$ (notice that the distiction between a straight line and an arc can be ignored for small angles) from a distance D is equal to 0.1 radian.

diffraction pattern, an effect caused by the wave nature of the light passing through the restricted aperture of the lens.

The diffraction pattern that our hypothetical lens might form of a star is represented, at a very large scale, in Figure 1.3. It consists of a central bright disc, the *Airy disc*, surrounded by concentric rings of decreasing brightness. The diffraction pattern of a star produced by a perfect lens has 84 per cent of the light energy contained in the Airy disc. The remainder of the light energy is shared amongst the more diffuse rings.

The diameter of the diffraction pattern is inversely proportional to the diameter of the lens producing it. Hence **larger** apertures produce **smaller** diffraction patterns. This effect is of crucial importance when the lens is used to image fine details. For instance, consider the case of two close stars being imaged by a lens of given aperture. The size of the aperture sets the sizes of the diffraction pattern images of each star. If the stars are too close together their diffraction patterns will merge and the stars will appear as one. In order to resolve the two stars as separate a lens of larger aperture must be used. Figure 1.4 (a), (b) and (c) illustrates the cases of two stars which are not resolved as separate, just resolved, and easily resolved, respectively.

The relationship between the limiting resolution of a lens and its aperture is expressed mathematically by the equation:

$$R = 206\,265 \times \frac{1.22\lambda}{d}$$

where R is the minimum angular separation of a pair of stars in order to be just resolved, measured in arcseconds; λ is the wavelength of the light being imaged and d is the diameter of the lens. Both λ and d are measured in the same units, for instance both in metres, or both in centimetres, etc.

You should notice that the resolution of the lens is wavelength dependent. The wavelength of light to which the eye is most sensitive is taken as 5.5×10^{-7}m (This corresponds to the yellow-green part of the spectrum). A perfect lens of 0.1m aperture should then be able to resolve two stars separated by 1.4 arcseconds when used visually.

Another formula for resolving power at visual wavelengths is that derived empirically by Dawes:

$$R = 4.56/D,$$

where R is the minimum resolvable separation of two stars in arcseconds, as before, and D is the aperture expressed in inches (1 inch = 25.4 mm). Dawes' formula predicts a minimum separation for resolution about 20 per

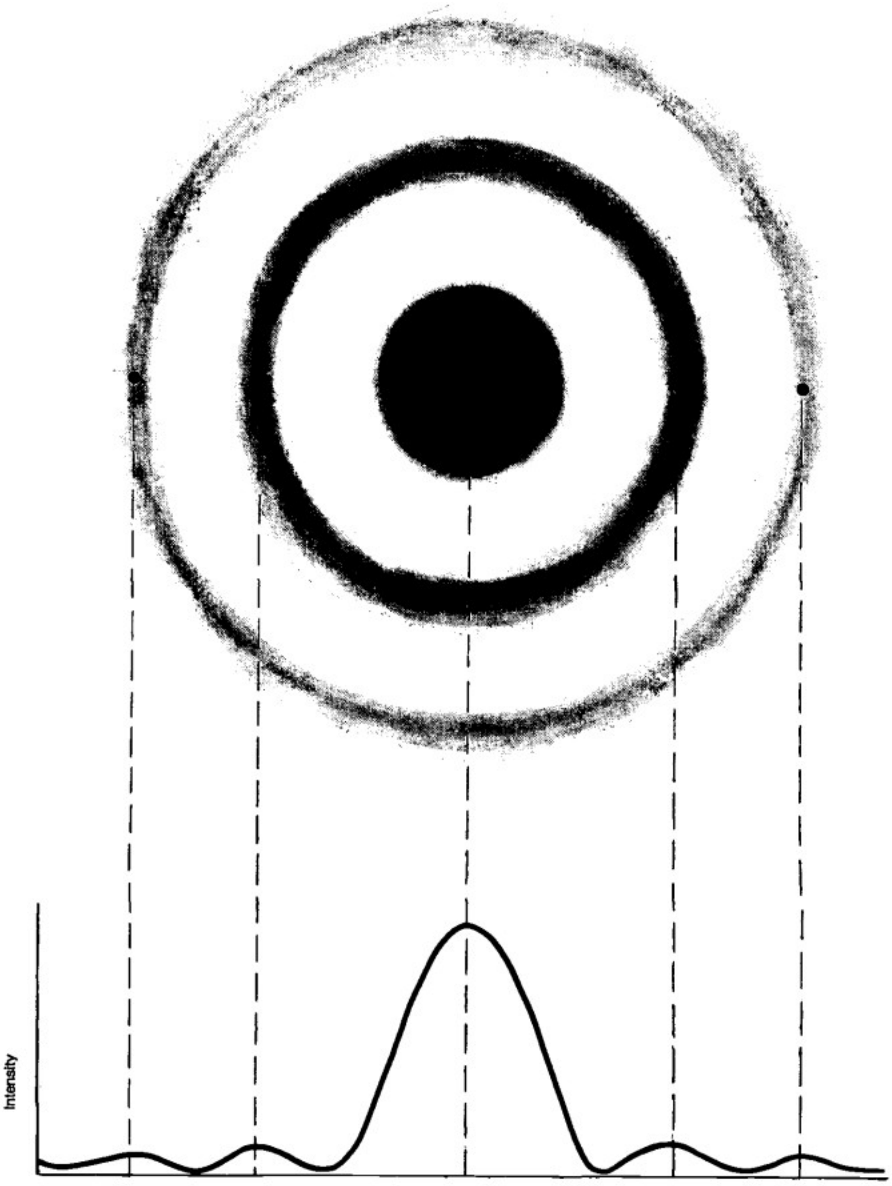

Figure 1.3. A representation of the diffraction pattern of a point source as seen through a limited aperture. The graph below illustrates the variation of the intensity of the light across the diameter of the diffraction pattern.

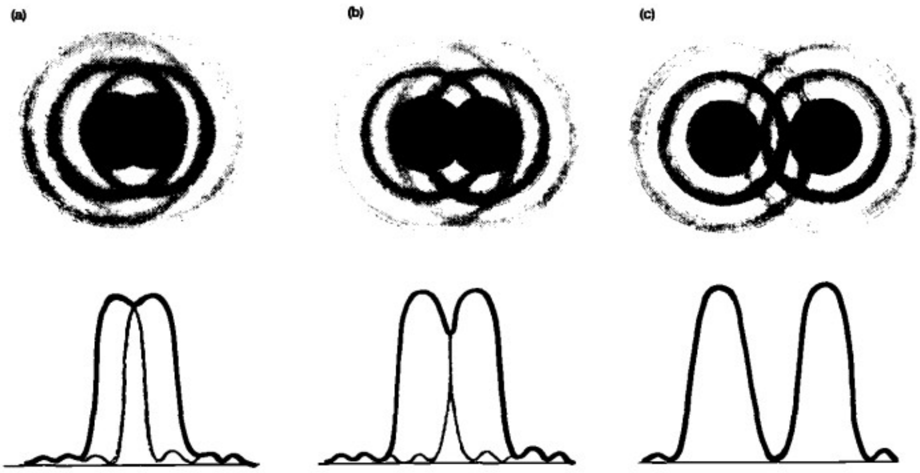

Figure 1.4. The overlapping diffraction patterns produced by two stars as seen through a given aperture. In (a) the stars are too close together to be resolved as separate. In (b) the stars are just resolvable and in (c) the stars are easily resolved.

cent closer than that of the mathematically derived formula but double star observers have found it to be more accurate in practice. Strictly, Dawes' formula applies to a pair of double stars of the sixth magnitude. Brighter stars, and those of unequal brightness, are less easily resolved.

The image of an extended object (anything which does not appear as a point), such as the surface of the Moon or a planet, can be thought of as a series of overlapping diffraction patterns. Here, too, the smaller diffraction patterns produced by larger apertures enable finer detail to be resolved. The situation has been compared to viewing a mosaic made of tiles. The finest details that can be seen in the mosaic are the size of the individual tiles. Making the mosaic from smaller tiles allows finer details to be represented.

The simple refractor

Figure 1.5(a) illustrates the optical principles of the simple refractor. Here the object glass and eyepiece are each represented as single lenses. The single-arrowed light rays originate from the top of some distant object and the double-arrowed light rays originate from the bottom of it. The object glass forms an inverted image of the object at position *I*. This is the object glass' focal plane.

When in normal adjustment, the eyepiece's focal plane is also made to

$. 2& ===
/% * &* -*
/% 5 ,&*& & * -*
7 !"# $ % & , * +C- %
' (1 -*
/% 01 E & & &

Chapter 10

PARALLEL POSTULATES

> [Euclid's Fifth Postulate] ought to be struck from the postulates altogether. For it is a theorem — one that invites many questions ... — and requires for its demonstration a number of definitions as well as theorems. ... it lacks the special character of a postulate.
>
> — Proclus (Greek, 410–485) [**AT**: Proclus], p. 151

PARALLEL LINES ON THE PLANE ARE SPECIAL

Up to this point we have not had to assume anything about parallel (non-intersecting) lines. No version of a parallel postulate has been necessary, on the plane, on a sphere, or on a hyperbolic plane. We defined the concrete notion of parallel transport and proved in Problem **8.2** that, on the plane (and hyperbolic planes), parallel transported lines do not intersect. Now in this chapter we will look at three important properties on the plane that require further assumptions and that will be needed in later chapters. If you are willing to assume these three statements, you may skip this chapter, but we urge you to finish reading through at least the next page.

> *If two lines on the plane are parallel transports of each other along some transversal, then they are parallel transports along any transversal.* (Contained in Problem **10.1**)

> *On the plane, the sum of the interior angles of a triangle is always* 180°. (Contained in Problem **7.3b** or Problem **10.2**)

On the plane, non-intersecting lines are parallel transports along all transversals. (Contained in Problem **10.3d**)

We have already seen that none of these properties is true on a sphere or a hyperbolic plane. Thus all three need for their proofs some property of the plane that does not hold on spheres and hyperbolic planes. The various properties that permit proofs of these three statements are collectively termed the ***Parallel Postulates***.

Only the three statements above are needed from this chapter for the rest of the book. Therefore, it is possible to omit this chapter and assume one of the above three statements and then prove the others. However, parallel postulates have a historical importance and a central position in many geometry textbooks and in many expositions about non-Euclidean geometries. The problems in this chapter are an attempt to help people unravel and enhance their understanding of parallel postulates. Comparing situations on the plane with situations on a sphere and on a hyperbolic plane is a powerful tool for unearthing our hidden assumptions and misconceptions about the notion of "parallel" on the plane.

As we discussed already in Chapter 8, because we have so many (often unconscious) connotations and assumptions attached to the word "parallel," we find it best to avoid using the term "parallel" as much as possible in this discussion. Instead we will use terms such as "parallel transport," "non-intersecting," and "equidistant," which make explicit the meaning that is intended.

PROBLEM 10.1 PARALLEL TRANSPORT ON THE PLANE

Show that if l_1 and l_2 are lines on the plane such that they are parallel transports along a transversal l, then they are parallel transports along any transversal. Prove this using any assumptions you find necessary. Make as few assumptions as you can, and make them as simple as possible. ***Be sure to state your assumptions clearly.***

What part of your proof does not work on a sphere or on a hyperbolic plane?

SUGGESTIONS

This problem is by no means as trivial as it at first may appear. In order to prove this theorem, you will have to assume something — there are many possible assumptions, so use your imagination. But at the same time, try not to assume any more than is necessary. If you are having trouble deciding what to assume, try to solve the problem in a way that seems natural to you and then see what develops while making explicit any assumptions you are using.

On spheres and hyperbolic planes, try the same construction and proof you used for the plane. What happens? You should find that your proof does not work on these surfaces. So what is it about your proof (on a sphere and hyperbolic plane) that creates difficulties?

Problem **10.1** emphasizes the differences between parallelism on the plane and parallelism on spheres and hyperbolic planes. On the plane, non-intersecting lines exist, and one can "parallel transport" everywhere. Yet, as was seen in Problems **8.2** and **8.3**, on spheres and hyperbolic planes two lines are cut at congruent angles if and only if the transversal line goes through the center of symmetry formed by the two lines. That is, on spheres and hyperbolic planes two lines are parallel transports only when they can be parallel transported through the center of symmetry formed by them. Be sure to draw a picture locating the center of symmetry and the transversal. On spheres and hyperbolic planes it is impossible to slide the transversal along two parallel transported lines while keeping both angles constant (something you *can* do on the plane). In Figures 10.1, the line r' is a parallel transport of line r along line l, but it is not a parallel transport of r along l'.

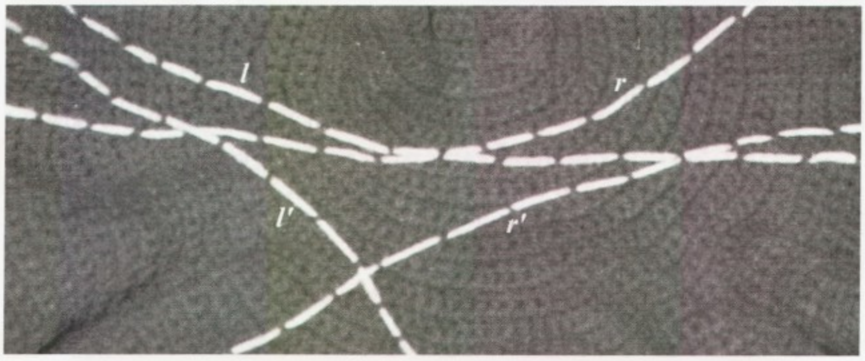

Figure 10.1a Parallel transport on a hyperbolic plane along l but not along l'

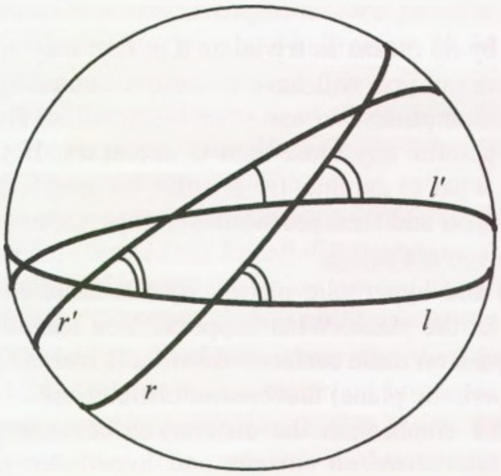

Figure 10.1b Parallel transport on a sphere along *l*, but not along *l'*

Pause, explore, and write out your ideas before reading further.

We will now divide parallel postulates into three groups: those involving mostly parallel transport, those involving mostly equidistance, and those involving mostly intersecting or non-intersecting lines. This division is useful even though it is rough and unlikely to fit every conceivable parallel postulate. *Which of these three groups is the most appropriate for your assumption from Problem* **10.1**? We will call the assumption you made in Problem **10.1** "*your parallel postulate.*"

PROBLEM 10.2 PARALLEL POSTULATES NOT INVOLVING (NON-)INTERSECTING LINES

Commonly used postulates of this sort are

H = 0: *The holonomy of triangles is zero.*

A = 180: *The sum of the angles of a triangle is equal to 180°.*

> **PT!**: *If two lines are parallel transports (PT) along one line then they are PT along ALL transversals.*

Note that these are false on spheres and on hyperbolic planes. The last two properties (**A = 180** and **PT!**) will be needed crucially in almost all the remaining chapters (and you may have already used one of them in Problem **9.4**). These properties are needed to study (on the plane) parallelograms, rectangles, and similar triangles (Chapters 12 and 13), circles and inversion (Chapters 15 and 16), projections of spheres and hyperbolic planes onto the Euclidean plane (Chapters 14 and 17), Euclidean manifolds (Chapters 18 and 24), solutions to quadratic and cubic equations (Chapter 19), trigonometry (Chapter 20), and polyhedra in 3-space (Chapter 23). It is these properties and their uses that we most often associate with consequences of the parallel postulates.

In Problem **7.3b** we proved **A = 180** on the plane, so we can

a. *Ask what assumption about the plane was used in proving* **7.3b**.

b. *Use* **A = 180** *to prove* **PT!**.

Look first at transversals that intersect the line along which the two lines are parallel transports.

c. *Using* **PT!** (Problem **10.1**), *prove* **A = 180** *without using results from Chapter 7.*

Start by parallel transporting one side of the triangle.

d. *Show that, on the plane,* **H = 0 ⇔ A = 180 ⇔ PT!**. *If your postulate from* **10.1** *is in this group, then show that it is also equivalent to the others.*

This is usually accomplished most efficiently by proving **H = 0 ⇒ A = 180 ⇒ PT! ⇒** Your Postulate **⇒ H = 0**, or in any other order.

e. *Prove that, on the plane, two parallel transported lines are equidistant.*

Look for rectangles or parallelograms.

EQUIDISTANT CURVES ON SPHERES AND HYPERBOLIC PLANES

The latitude circles on the earth are sometimes called "parallels of latitude." They are parallel in the sense that they are everywhere equidistant as are concentric circles on the plane. In general, transversals do not cut equidistant circles at congruent angles. However, there is one important case where transversals do cut the circles at congruent angles. Let l and l' be latitude circles the same distance from the equator on opposite sides of it. See Figure 10.2. Then every point on the equator is a center of half-turn symmetry for these pair of latitudes. Thus, as in Problems **8.3** and **10.1**, every transversal cuts these latitude circles in congruent angles, even though these latitude circles are not geodesics. In the first section of Chapter 2 we noted that Euclid, in his *Phenomena,* discussed such equidistant circles.

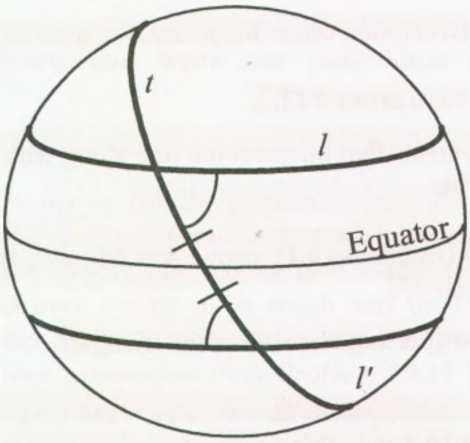

Figure 10.2 Special equidistant circles

The same ideas work on a hyperbolic plane: If g is a geodesic and l and l' are the (two) curves (not geodesics) that are a distance d from g, then l and l' are equidistant from each other and every transversal cuts them at congruent angles. This follows from the fact that g has half-turn symmetry at every point.

PROBLEM 10.3 PARALLEL POSTULATES INVOLVING (NON-)INTERSECTING LINES

One of Euclid's assumptions constitutes ***Euclid's Fifth (or Parallel) Postulate (EFP)***, which says

> **EFP:** *If a straight line intersecting two straight lines makes the interior angles on the same side less than two right angles, then the two lines (if extended indefinitely) will meet on that side on which are the angles less than two right angles.*

For a picture of EFP, see Figure 10.3.

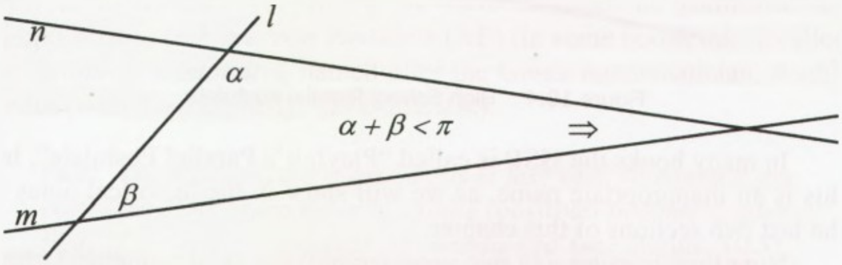

Figure 10.3 Euclid's Parallel Postulate

You probably did not assume EFP in your proof of Problem **10.1**. You are in good company — many mathematicians, including Euclid, have tried to avoid using it as much as possible. However, we will explore EFP because, historically, it is important, and because it has some very interesting properties, as you will see in Problem **10.3**. On a sphere, all straight lines intersect twice, which means that EFP is trivially true on a sphere. But in Problem **10.4**, you will show that EFP is also true in a stronger sense on spheres. You will also be able to prove that EFP is false on a hyperbolic plane.

Thus, EFP does not have to be assumed on a sphere — it can be proved! However, in most high school geometry textbooks, EFP is replaced by another postulate, claimed to be equivalent to EFP. This postulate we will call the ***High School Parallel Postulate*** (***HSP***), and it can be expressed in the following way:

> **HSP:** *For every line **l** and every point **P** not on **l**, there is a unique line **l'** that passes through **P** and does not intersect (is parallel to) **l**.*

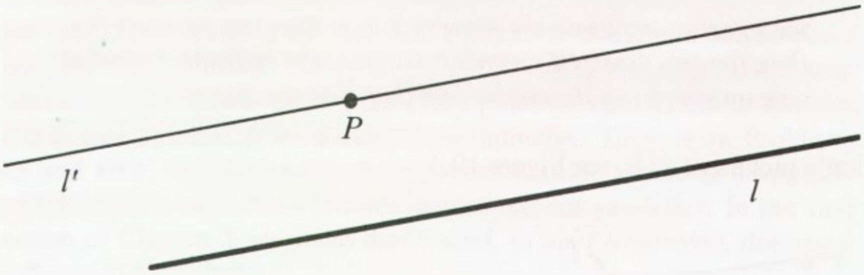

Figure 10.4 High School Parallel Postulate

In many books the **HSP** is called "Playfair's Parallel Postulate", but this is an inappropriate name, as we will show in the historical notes in the last two sections of this chapter.

Note that, because any two great circles on a sphere intersect, there are *no* lines l' that are parallel to l in the "not intersecting" sense. Therefore, **HSP** is not true on spheres. On the other hand, if we change "parallel" to "parallel transport" then *every* great circle through P is a parallel transport of l along *some* transversal. What happens on a hyperbolic plane? In Problem **10.3**, you will explore the relationships among **EFP**, **HSP**, and your postulate from Problem **10.1**. In Problem **10.4** we will explore these postulates on spheres and hyperbolic planes.

a. *Show that, on the plane,* **EFP** *and* **HSP** *are equivalent. If your postulate from* **10.1** *is in this group, is it equivalent to the others? Why or why not?*

To show that **EFP** and **HSP** are equivalent on the plane, you need to show that you can prove **EFP** if you assume **HSP** and vice versa. If the three postulates are equivalent, then you can prove the equivalence by showing that

$$\textbf{EFP} \Rightarrow \textbf{HSP} \Rightarrow \text{Your Postulate} \Rightarrow \textbf{EFP}$$

or in any other order. It will probably help you to draw lots of pictures of what is going on. Also, remember that we proved in Problem **8.2** (without using any parallel postulate) that parallel (non-intersecting) lines exist. Note that **HSP** is not true on a sphere but **EFP** is true, so your proof that **EFP** implies **HSP** on the plane must use some property of the plane that does not hold on a sphere. Look for it.

b. *Prove that either* **EFP** *or* **HSP** *can be used to prove* (*without using* **10.1** *or* **7.3b**) *one of* **H = 0**, **A = 180**, or **PT!**. (It does not matter which. *Why*?)

So, are all these postulates equivalent to each other on the plane? The answer is almost, but not quite! In order for **A = 180** (or **H = 0** or **PT!**) to imply **EFP** (or **HSP**), we have to make an additional assumption, the ***Archimedean Postulate*** (**AP**) (in some books this is called the *Axiom of Continuity*), named after the Greek mathematician, Archimedes (who lived in Sicily, 287?–212 B.C.):

> ***AP**: *On a line, if the segment AB is less than* (contained in) *the segment AC, then there is a finite* (positive) *integer, n, such that, if we put n copies of AB end-to-end* (see Figure 10.5), *then the* n^{th} *copy will contain the point C.*

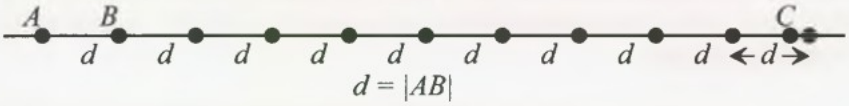

Figure 10.5 The Archimedean Postulate

The Archimedean Postulate can also be interpreted to rule out the existence of infinitesimal lengths. The reason these postulates are "almost but not quite" equivalent to each other on the plane is that, though **AP** is needed, it is assumed by most people to be true on the plane, spheres, and hyperbolic planes. But this is the first time we have needed this assumption in this book.

***c.** *Show that, on the plane,* **AP** *and* **A = 180** *imply* **EFP**.

Look at the situation of **EFP**, which we redraw in Figure 10.6. Pick a sequence of equally spaced points *A*, *A'*, *A''*,... on *n* (on the side of the angle α). Next, parallel transport l along m to lines l', l'',... that intersect n at the points A', A'',... . And parallel transport m along l to m', m'',..., which also intersect n at the points A', A'',... . Look for congruent angles and congruent triangles. Use **AP** to argue that n and m will eventually intersect.

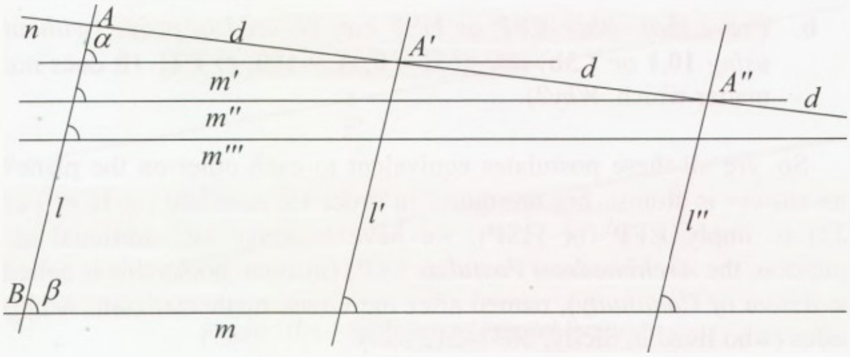

Figure 10.6 **AP** and **A = 180** imply **EFP**

Clearly, this proof feels different from the others proofs in this chapter. And, as already pointed out, most of the applications in this book of parallel postulates only need **A = 180** and **PT!**. So is it possible for us not to bother with **EFP** and **HSP**? Or are they needed? Yes, we need them, but only in a few places, such as in Problem **11.3**, where what we need to know is

d. Prove using **EFP**: *On the plane, non-intersecting lines are parallel transports along every transversal.*

Compare with **EFP**.

PROBLEM 10.4 EFP AND HSP ON SPHERE AND HYPERBOLIC PLANE

a. *Show that **EFP** is true on a sphere in a strong sense; that is, if lines **l** and **l′** are cut by a transversal **t** such that the sum of the interior angles $\boldsymbol{\alpha} + \boldsymbol{\beta}$ on one side is less than two right angles, then not only do **l** and **l′** intersect, but they also intersect closest* *to **t** on the side of α and β. You will have to determine an appropriate meaning for "closest."*

To help visualize the postulates, draw these "parallels" on an actual sphere. There are really two parts in this proof — first, you must come up with a definition of "closest," and then prove that **EFP** is true for this definition. The two parts may come about simultaneously as you come up with a proof. This problem is closely related to Euclid's Exterior Angle Theorem but can also be proved without using **EEAT**. One case that you should look at specifically is pictured in Figure 10.7. It is not necessarily obvious how to define the "closest" intersection.

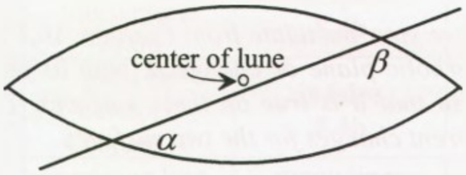

Figure 10.7 Is **EFP** true on a sphere?

b. *On a hyperbolic plane let **l** be a geodesic and let **P** be a point not on **l**; then show that there is an angle θ with the property that any line **l′** passing through **P** is parallel to* (not intersecting) ***l** if the line **l′** does not form an angle less than θ with the line from **P** that is perpendicular to **l*** (Figure 10.8).

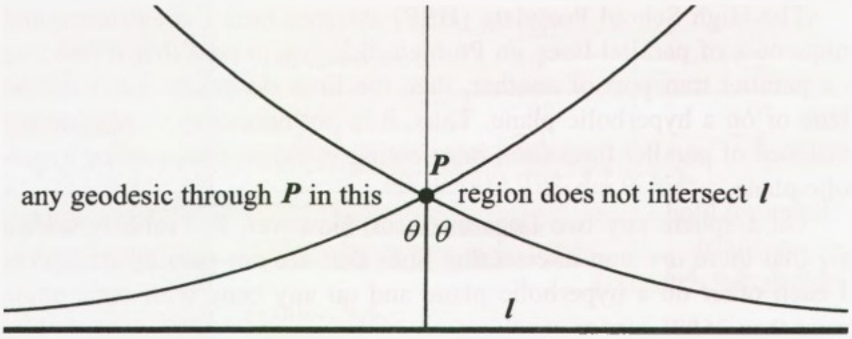
P
any geodesic through P in this
region does not intersect l
θ
θ
l

Figure 10.8 Multiple parallels on a hyperbolic plane

Look at a variable line through P that does intersect l and then look at what happens as you move the intersection point out to infinity.

> **c.** *Using the notion of parallel transport, change the High School Postulate so that the changed postulate is true on both spheres and hyperbolic planes. Make as few alterations as possible and keep some form of uniqueness.*

Try to limit your alterations so that the new postulate preserves the spirit of the old one. You can draw ideas from any of the previous problems to obtain suitable modifications. Prove that your modified versions of the postulate are true on spheres and hyperbolic planes.

> **d.** *Either prove your postulate from Problem* **10.1** *on a sphere and on a hyperbolic plane or change it, with as few alterations as possible, so that it is true on these surfaces. You may need to make different changes for the two surfaces.*

In Problem **10.1** you should have decided whether or not your postulate is true on spheres or on hyperbolic spaces.

COMPARISONS OF PLANE, SPHERES, AND HYPERBOLIC PLANES

Figure 10.9 is an attempt to represent the relationships among parallel transport, non-intersecting lines, **EFP**, and **HSP**. Can you fit your postulate into the diagram?

The High School Postulate (**HSP**) assumes both the existence and uniqueness of parallel lines. In Problem **8.2**, you proved that if one line is a parallel transport of another, then the lines do not intersect on the plane or on a hyperbolic plane. Thus, it is not necessary to assume the existence of parallel lines (non-intersecting lines) on the plane or hyperbolic plane.

On a sphere any two lines intersect. However, in Problem **8.4** we saw that there *are* non-intersecting lines that are not parallel transports of each other on a hyperbolic plane and on any cone with cone angle larger than 360°.

	Euclidean	**Hyperbolic**	**Spherical**
Parallel transported lines	do not intersect and are equidistant **8.2 & 10.2**	diverge in both directions **8.2 & 8.3**	always intersect **2.1**
Parallel transported lines	are parallel transports along all transversals **10.1**	are parallel transports along any transversal that passes through the center of symmetry of the lines **8.3**	
Non-intersecting lines	are parallel transports along all transversals **10.3d**	sometimes are not parallel transports **8.4b**	do not exist **2.1**
Euclid's 5th Postulate	must be assumed Chapter 10	is false **8.4b**	is provable in a strong sense **10.4a**
High School Parallel Postulate	a unique line through a point not intersecting a given line **10.3**	many lines through a point not intersecting a given line **10.4b**	no non-intersecting lines **2.1**
Your postulate from **10.1**			
Two points determine	a unique line and line segment **6.1d**		at least two line segments **6.1d**
Sum of the angles of a triangle	= 180° **7.3b & 10.2**	< 180° **7.1**	> 180° **7.2**
Holonomy	= 0° **10.2**	< 0° **7.4**	> 0° **7.4**
VAT & ITT	are always true, **3.2 & 6.2**		
SAS, ASA, and SSS	hold for all triangles **6.4, 6.5, & 9.1**		hold for small triangles **6.4, 6.5, & 9.1**
AAA	is false, **9.4**, similar triangles	is true, **9.4**, no similar triangles	

Figure 10.9 Comparisons of the three geometries

PARALLEL POSTULATES WITHIN THE BUILDING STRUCTURES STRAND

The problem of parallels puzzled Greek geometers (see the quote at the beginning of this chapter). The prevailing belief was that it surely follows from the straightness of lines that lines like n and m in Figure 10.10 *had* to intersect, and so Euclid's Fifth Postulate would turn out to be unnecessary.

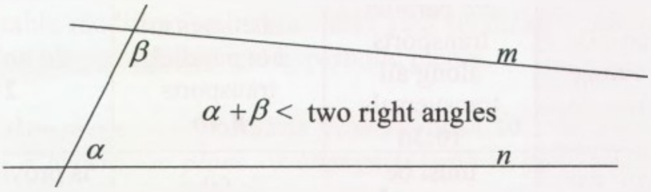

Figure 10.10 Surely these intersect!

During and since the Greek era, there have been attempts to derive the postulate from the rest of elementary geometry; attempts to reformulate the postulate or the definition of parallels into something less objectionable; and descriptions of what geometry would be like if the postulate was in some way denied. Part of these attempts was to study what geometric properties could be proved without using a parallel postulate, but with the other postulates (properties) of the plane. The results of these studies became known as ***absolute geometry***. Both the plane and a hyperbolic plane are examples of absolute geometry. Included in absolute geometry are all the results about the plane and hyperbolic planes that we have discussed in Chapters 1, 3, 5, 6, 8, and 9 (except AAA, Problem **9.4**). In general, absolute geometry includes everything that is true of both the plane and hyperbolic planes.

Claudius Ptolemy (Greek, 100–178) implicitly used what we now call Playfair's Postulate (**PP**) to prove **EFP**. This is reported to us by Proclus (Greek, 410–485) [**AT:** Proclus], who explicitly states **PP**.

> **PP**. *Two straight lines that intersect one another cannot be both parallel to the same straight line.*

Or, in another common wording,

PP. *Given a straight line and a point not on the line, there is at most one straight line through the point that does not intersect the given line.*

Playfair's Parallel Postulate got its current name from the Scottish mathematician John Playfair (1748–1819), who brought out successful editions of Euclid's *Elements* in the years following 1795. After 1800 many commentators referred to Playfair's postulate (**PP**) as the best statement of Euclid's postulate, so it became a tradition in many geometry books to use **PP** instead of **EFP**. The reader can see easily that both **PP** and **EFP** assert the uniqueness of parallel lines and neither asserts the existence of parallel lines. **HSP** seems to have been first stated by Hilbert in the first edition of his *Grundlagen der Geometrie* (but not in the later editions, where he used **PP**) and in [**EG**: Hilbert]. But later, apparently starting in the 1960s, many (but not all) textbooks (including earlier editions of this book!) started calling **HSP** by the name "Playfair's Postulate". In absolute geometry **HSP** is equivalent to **PP** and **EFP** (as you showed in **10.3a**), and because absolute geometry was the focus of many investigations, the statement "**HSP** is equivalent to **EFP**" was made and is still being repeated in many textbooks and expository writings about geometry even when the context is not absolute geometry. As you showed in **10.4a**, **EFP** is true on spheres and so is **PP**, and so they cannot be equivalent to **HSP**, which is clearly false on the sphere.

During the 9th through 12th centuries, parallel postulates were explored by mathematicians in the Islamic world. al-Hasan ibn al-Haytham (Persian, 965–1040) proved **EFP** by assuming, *A quadrilateral with three right angles must have all right angles.* Quadrilaterals with three right angles are known later in Western literature as Lambert quadrilaterals after Johann Lambert (German–Swiss, 1728–1777), who studied them and also was the first to extensively investigate hyperbolic trigonometric functions. The Persian poet and geometer Omar Khayyam (1048–1131) wrote a book that in translation is entitled *Discussion of Difficulties in Euclid* [**AT:** Khayyam 1958]. Khayyam introduces a new postulate (which he attributes to Aristotle, though it has not been found among the surviving works of Aristotle), which says *Two straight lines that start to converge continue to converge.* In his work on parallel lines he studied the Khayyam quadrilaterals (later in the West to be called Saccheri quadrilaterals), which are discussed here in Chapter 12. Nassir al-Din Al-Tusi (Khorasan, 1201–1274) furthered the study of parallel postulates and is credited (though some say it may have been his son) with first proving that **A = 180** *is equivalent to* **EFP**, our **10.3bc**. Al-Tusi's works were the first Islamic mathematical works to be discovered in the Western Renaissance and were published in Rome in 1594.

The assumption *parallel lines are equidistant* (see our **10.2d**) was discussed in various forms by Aristotle (Greek, 383–322 B.C.), Posidonius (Greek, 135–51 B.C.), Proclus, ibn Sina (Uzbek, 980–1037), Omar Khayyam, and Saccheri (Italian, 1667–1733).

John Wallis (English, 1616–1703) proved that **EFP** followed from the assumption *To every triangle, there exists a similar triangle of arbitrary magnitude*. (See our **9.4**.) The works, already mentioned, of Wallis, Saccheri, Lambert in the 17th and 18th centuries were continued by the French school into the early 19th century by Joseph Fourier (French, 1768–1830), who concluded that geometry was a physical science and could not be established a priori, and Adrien-Marie Legendre (1752–1833), who proved that *in absolute geometry the sum of the interior angles of a triangle is always less than or equal to 180°.*

The breakthrough in the study of parallel postulates came in the 19th century when, apparently independently, János Bolyai (Hungarian, 1802–1860), and N. I. Lobachevsky (Russian, 1792–1856) finally developed hyperbolic geometry as an absolute geometry that did not satisfy **EFP**. See Chapters 5 and 17 for more discussions of hyperbolic geometry. These discoveries of hyperbolic geometry showed that the quest for a proof of **EFP** from within absolute geometry is impossible.

For details of the relevant histories, see [**HI**: Gray], [**DG**: McCleary], [**HI**: Rosenfeld], and the Heath's editorial notes in [**AT**: Euclid, *Elements*].

NON-EUCLIDEAN GEOMETRIES WITHIN THE HISTORICAL STRANDS

Building Structures Strand. Because of this long history of investigation into parallel postulates within the Building Structures Strand, many books misleadingly call hyperbolic geometry "*the* non-Euclidean geometry." As pointed out in the introduction to Chapter 2, spherical geometry has been studied since ancient times within the Navigation/Stargazing Strand, but it did not fit into absolute geometry and thus was (and still is) left out of many discussions of non-Euclidean geometry, especially those that take place in the context of the Building Structures Strand. In those discussions that do include spherical geometry, hyperbolic geometry is usually called ***Lobachevskian geometry*** or ***Lobachevskian/Bolyai geometry***, and spherical geometry is usually called ***Riemannian geometry***. Bernhard Riemann (German, 1826–1866) pioneered an intrinsic and analytic view for surfaces and space and suggested that the intrinsic geometry of the sphere could be thought of as a non-Euclidean geometry. He also introduced an intrinsic analytic view of the sphere that became known as the ***Riemann sphere***. The Riemann sphere is usually studied in a course on complex analysis. The use of the term "Riemannian geometry" to denote the non-Euclidean geometry of the sphere leads to confusions since "Riemannian geometry" more commonly refers to the part of differential geometry that gives intrinsic analytic descriptions of the local geometry of general surfaces and 3-space.

Art/Pattern Strand. In this strand, artists' experimentations with perspective in art led to the mathematical theory of *projective geometry* that developed initially independently of both spherical and hyperbolic geometries. Then in the last part of the 19th century, projective geometry was used to provide a unified treatment of the three geometries. It was within this development that Felix Klein introduced the names ***double elliptic geometry*** (for the geometry of the sphere), ***elliptic geometry*** (for the geometry of the sphere with antipodal points identified — this the same as the real projective plane, **RP**^{2}, discussed in Problem **18.3**), ***parabolic geometry*** (for Euclidean geometry), and ***hyperbolic geometry*** (for the Bolyai/Lobachevsky geometry). In Chapter 11, we will discuss Klein's development of various geometries in terms of their isometries. A geometric connection between spherical and projective geometry can be seen in the image onto the plane of the sphere (with antipodal points identified) under gnomic projection. (See Problems **14.2** and **20.6**.)

For a fuller discussion of spherical and hyperbolic geometry in this strand, see [**HI**: Kline], Chapter 38, and [**HI**: Katz], Section 17.3.

In Kline we also find a statement (page 913) that represented the view of some mathematicians:

> By the early 1870s several basic non-Euclidean geometries, the hyperbolic, and the two elliptic geometries, had been introduced and intensively studied. The fundamental question which had yet to be answered in order to make these geometries legitimate branches of mathematics was whether they were consistent.

So for those who believe that the only "legitimate" branches of mathematics are those that have been included in some rigorous formal system, then non-Euclidean geometries did not became a part of mathematics until the late 1800s.

Navigation/Stargazing Strand. In the introduction to Chapter 2 we saw that spherical geometry has been studied within this strand for more than 2000 years, and in this study were developed (spherical) coordinate systems and trigonometric functions. We hope that the reader agrees that this was (and still is) legitimate mathematics.

For more readings (and references) on this history of spherical geometry within the Navigation/Stargazing Strand, see [**HI**: Katz], Chapter 4, and [**HI**: Rosenfeld], Chapter 1. Into the 19th century, spherical geometry occurred almost entirely in the Navigation/Stargazing Strand and was used by Brahe and Kepler in studying the motion of stars and planets and by navigators and surveyors. The popular book (which is also available online) *Spherical Trigonometry: For the use of colleges and schools* (1886) [**SP**: Todhunter] contains several discussions of the use of spherical geometry in surveying and was used in British schools before hyperbolic geometry was widely known in the British Isles. On the continent, C. F. Gauss (1777–1855) was using spherical geometry in various large-scale surveying projects before the advent of hyperbolic geometry. Within this context, Gauss initiated (and Riemann expanded) the analytic study of surfaces that led to what we now call differential geometry. Gauss introduced *Gaussian curvature*, which was discussed in Chapter 7. Within this theory of surfaces, spherical geometry is described locally as the geometry of surfaces with constant positive (Gaussian) curvature and hyperbolic geometry is described locally as the geometry of surfaces with constant negative curvature. As we discussed in Chapter 5, there are no analytic surfaces in 3-space with the complete hyperbolic geometry, but the pseudosphere (Problem **5.3**) locally has hyperbolic geometry. Because of this, in the Navigation/Stargazing strand, hyperbolic geometry is called by some authors ***pseudosphere geometry***.

Chapter 11

ISOMETRIES AND PATTERNS

> What geometrician or arithmetician could fail to take pleasure in the symmetries, correspondences and principles of order observed in visible things?
>
> — Plotinus, *The Enneads*, II.9.16 [**AT:** Plotinus]

> [Geometry is] the study of the properties of a space which are invariant under a given group of transformations.
>
> F. Klein, *Erlangen Program*

> Life forms illogical patterns. It is haphazard and full of beauties which I try to catch as they fly by, for who knows whether any of them will ever return? — Margot Fonteyn

Recall that in Chapter 1 we gave the following:

> **DEFINITION.** An ***isometry*** is a transformation that preserves distances and angle measures.

In this chapter we will show (for the plane, spheres, and hyperbolic planes) that every isometry is the composition (product) of (not more than three) reflections, and we will determine all the different types of isometries. This finishes the study of reflections and rotations we started in Problem **5.4**. We will note the differences between the kinds of isometries that appear in the three geometries.

Then we will study patterns in these three spaces. Along the way we will look at some group theory through its origins, that is, geometrically.

This material originated primarily within the Art/Pattern Strand.

/% E *

20211023

backlash – [illegible]

patronímico

The "open universe" has a saddle-~~shaped~~ shaped negatively curved spacetime.

saddle-shaped → em forma de sela.

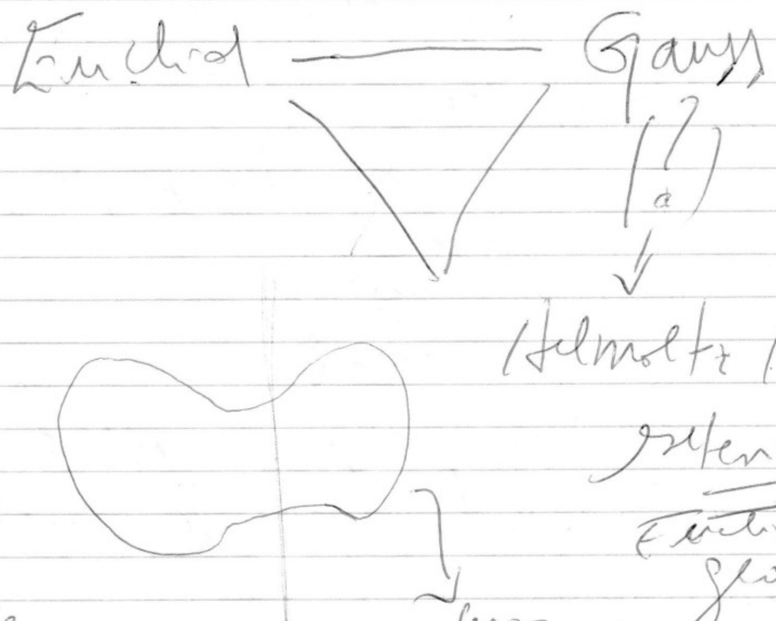

Helmoltz (questionou?)

extension of Euclidean geometry.

em parte.

esfera comprimida

$.) - & ==== PPPPPPP

E !"# $ % & " % 8 " &*
" & (! - ' * 6& @ 9 &% 1 7 -*

!"# $ % & " % 8 " &* " & (
! - ' * 6& @ 9 Q - 4 . *&(. % ' - 1 - .

/% 0 1 75 > %

" & & ((& -

13
Symmetry groups

13.1 Groups of transformations

SPACES that are symmetrical have a fundamental importance in modern physics. Why is this? It might be thought that completely exact symmetry is something that could arise only exceptionally, or perhaps just as some convenient approximation. Although a symmetrical object, such as a square or a sphere, has a precise existence as an idealized ('Platonic'; see §1.3) mathematical structure, any *physical* realization of such a thing would ordinarily be regarded as merely some kind of approximate representation of this Platonic ideal, therefore possessing no actual symmetry that can be regarded as exact. Yet, remarkably, according to the highly successful physical theories of the 20th century, all physical interactions (including gravity) act in accordance with an idea which, strictly speaking, depends crucially upon certain physical structures possessing a symmetry that, at a fundamental level of description, is indeed necessarily exact!

What is this idea? It is a concept that has come to be known as a 'gauge connection'. That name, as it stands, conveys little. But the idea is an important one, enabling us to find a subtle ('twisted') notion of differentiation that applies to general entities on a manifold (entities that are indeed more general than just those—the *p*-forms—which are subject to exterior differentiation, as described in Chapter 12). These matters will be the subject of the two chapters following this one; but as a prerequisite, we must first explore the basic notion of a *symmetry group*. This notion also has many other important areas of application in physics, chemistry, and crystallography, and also within many different areas of mathematics itself.

Let us take a simple example. What are the symmetries of a *square*? The question has two different answers depending upon whether or not we allow symmetries which reverse the orientation of the square (i.e. for which the square is turned over). Let us first consider the case in which these orientation-reversing symmetries are not allowed. Then the square's symmetries are generated from a single rotation through a right angle in the square's plane, repeated various numbers of times. For convenience, we can represent these motions in terms of complex numbers, as we did in

Chapter 5. We may, if we choose, think of the vertices of the square as occupying the points 1, i, -1, $-$ i in the complex plane (Fig. 13.1a), and our basic rotation represented by multiplication by i (i.e. by 'i×'). The various *powers* of i represent all our rotations, there being four distinct ones in all:

$$\mathrm{i}^0 = 1, \quad \mathrm{i}^1 = \mathrm{i}, \quad \mathrm{i}^2 = -1, \quad \mathrm{i}^3 = -\mathrm{i}$$

(Fig. 13.1b). The fourth power $\mathrm{i}^4 = 1$ gets us back to the beginning, so we have no more elements. The product of any two of these four elements is again one of them.

These four elements provide us with a simple example of a *group*. This consists of a set of elements and a law of 'multiplication' defined between pairs of them (denoted by juxtaposition of symbols) for which the *associative* multiplication law holds

$$a(bc) = (ab)c,$$

where there is an *identity* element 1 satisfying

$$1a = a1 = a,$$

and where each element a has an *inverse* a^{-1}, such that[13.1]

$$a^{-1}a = aa^{-1} = 1.$$

The symmetry operations which take an object (not necessarily a square) into itself always satisfy these laws, called the *group axioms*.

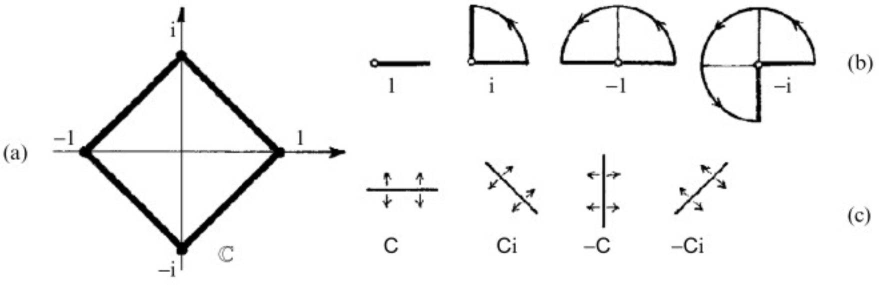

Fig. 13.1 Symmetry of a square. (a) We may represent the square's vertices by the points 1, i, $-$ 1, $-$ i in the complex plane ℂ. (b) The group of non-reflective symmetries are represented, in ℂ, as multiplication by $1 = \mathrm{i}^0$, $\mathrm{i} = \mathrm{i}^1$, $-1 = \mathrm{i}^2$, $-\mathrm{i} = \mathrm{i}^3$, respectively. (c) The reflective symmetries are given, in C, by C (complex conjugation), Ci, $-$ C, and $-$ Ci.

[13.1] Show that if we just assume $1a = a$ and $a^{-1}a = 1$ for all a, together with associativity $a(bc) = (ab)c$, then $a1 = a$ and $aa^{-1} = 1$ can be *deduced*. (*Hint*: Of course a is not the only element asserted to have an inverse.) Show why, on the other hand, $a1 = a$, $a^{-1}a = 1$, and $a(bc) = (ab)c$ are insufficient.

Recall the conventions recommended in Chapter 11, where we think of b acting first and a afterwards, in the product ab. We can regard these as operations as being performed upon some object appearing to the right. Thus, we could consider the motion, b, expressing a symmetry of an object $\boldsymbol{\Phi}$, as $\boldsymbol{\Phi} \mapsto b(\boldsymbol{\Phi})$, which we follow up by another such motion a, giving $b(\boldsymbol{\Phi}) \mapsto a(b(\boldsymbol{\Phi}))$. This results in the combined action $\boldsymbol{\Phi} \mapsto a(b(\boldsymbol{\Phi}))$, which we simply write $\boldsymbol{\Phi} \mapsto ab(\boldsymbol{\Phi})$, corresponding to the motion ab. The identity operation leaves the object alone (clearly always a symmetry) and the inverse is just the reverse operation of a given symmetry, moving the object back to where it came from.

In our particular example of non-reflective rotations of the square, we have the additional *commutative* property

$$ab = ba.$$

Groups that are commutative in this sense are called *Abelian*, after the tragically short-lived Norwegian mathematician Niels Henrik Abel.[1] Clearly any group that can be represented simply by the multiplication of complex numbers must be Abelian (since the multiplication of individual complex numbers always commutes). We saw other examples of this at the end of Chapter 5 when we considered the general case of a finite *cyclic* group $\mathbb{Z}_n$, generated by a single nth root of unity.[13.2]

Now let us allow the orientation-reversing *reflections* of our square. We can still use the above representation of the square in terms of complex numbers, but we shall need a new operation, which I denote by C, namely *complex conjugation*. (This flips the square over, about a horizontal line; see §10.1, Fig. 10.1.) We now find (see Fig. 13.1c) the 'multiplication laws'[13.3]

$$\mathsf{C}\mathrm{i} = (-\mathrm{i})\mathsf{C}, \quad \mathsf{C}(-1) = (-1)\mathsf{C}, \quad \mathsf{C}(-\mathrm{i}) = \mathrm{i}\mathsf{C}, \quad \mathsf{CC} = 1$$

(where[2] I shall henceforth write $(-\mathrm{i})\mathsf{C}$ as $-\mathrm{i}\mathsf{C}$, etc.). In fact, we can obtain the multiplication laws for the entire group just from the basic relations[13.4]

$$\mathrm{i}^4 = 1, \quad \mathsf{C}^2 = 1, \quad \mathsf{C}\mathrm{i} = \mathrm{i}^3\mathsf{C},$$

the group being non-Abelian, as is manifested in the last equation. The total number of of distinct elements in a group is called its *order*. The order of this particular group is 8.

Now let us consider another simple example, namely the group of rotational symmetries of an ordinary sphere. As before, we can first consider the

---

[13.2] Explain why any vector space is an Abelian group—called an *additive* Abelian group—where the group 'multiplication' operation is the 'addition' operation of the vector space.

[13.3] Verify these relations (bearing in mind that $\mathsf{C}\mathrm{i}$ stands for 'the operation $\mathrm{i}\times$, followed by the operation C, etc.). (*Hint*: You can check the relations by just confirming their effects on 1 and i. Why?)

[13.4] Show this.

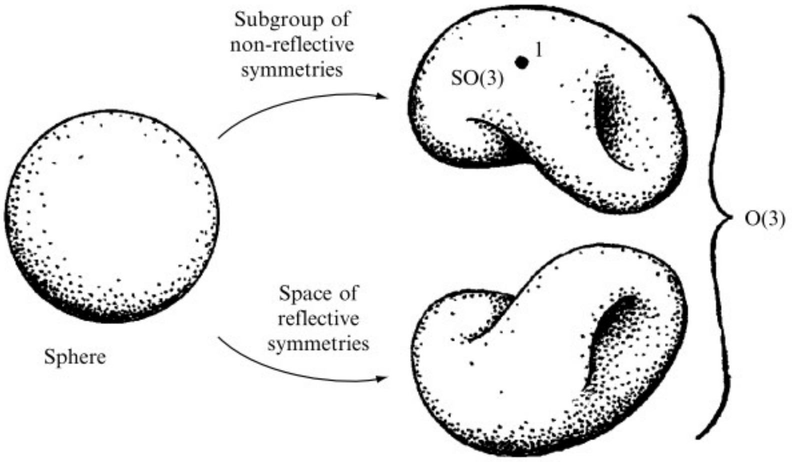

Fig. 13.2 Rotational symmetry of a sphere. The entire symmetry group, O(3), is a disconnected 3-manifold, consisting of two pieces. The component containing the identity element 1 is the (normal) subgroup SO(3) of non-reflective symmetries of the sphere. The remaining component is the 3-manifold of reflective symmetries.

case where reflections are excluded. This time, our symmetry group will have an infinite number of elements, because we can rotate through any angle about any axis direction in 3-space. The symmetry group actually constitutes a 3-dimensional space, namely the 3-manifold denoted by $\mathcal{R}$ in Chapter 12. Let me now give this group (3-manifold) its official name. It is called[3] SO(3), the non-reflective orthogonal group in 3 dimensions. If we now include the reflections, then we get a whole new set of symmetries—another 3-manifold's worth—which are disconnected from the first, namely those which involve a reversal of the orientation of the sphere. The entire family of group elements again constitutes a 3-manifold, but now it is a disconnected 3-manifold, consisting of two separate connected pieces (see Fig. 13.2). This entire group space is called O(3).

These two examples illustrate two of the most important categories of groups, the finite groups and the continuous groups (or *Lie* groups; see §13.6).[4] Although there is a great difference between these two types of group, there are many of the important properties of groups that are common to both.

13.2 Subgroups and simple groups

Of particular significance is the notion of a *subgroup* of a group. To exhibit a subgroup, we select some collection of elements within the group which themselves form a group, using the same multiplication and inversion operations as in the whole group. Subgroups are important in many modern theories of particle physics. It tends to be assumed that there is some fundamental symmetry of Nature that relates different kinds of particles to one another and also relates different particle interactions to one another. Yet one may not see this full group acting as a symmetry in any manifest way, finding, instead, that this symmetry is 'broken' down to some subgroup of the original group where the *subgroup* plays a manifest role as a symmetry. Thus, it is important to know what the possible subgroups of a putative 'fundamental' symmetry group actually are, in order that those symmetries that are indeed manifest in Nature might be able to be thought about as subgroups of this putative group. I shall be addressing questions of this kind in §§25.5–8, §26.11, and §28.1.

Let us examine some particular cases of subgroups, for the examples that we have been considering. The *non-reflective* symmetries of the square constitute a 4-element subgroup $\{1, \mathrm{i}, -1, -\mathrm{i}\}$ of the entire 8-element group of symmetries of the square. Likewise, the non-reflective rotation group SO(3) constitutes a subgroup of the entire group O(3). Another subgroup of the symmetries of the square consists of the four elements $\{1, -1, \mathsf{C}, -\mathsf{C}\}$; yet another has just the two elements $\{1, -1\}$.[13.5] Moreover there is always the 'trivial' subgroup consisting of the identity alone $\{1\}$ (and the whole group itself is, equally trivially, always a subgroup).

All the various subgroups that I have just described have a special property of particular importance. They are examples of what are called *normal* subgroups. The significance of a normal subgroup is that, in an appropriate sense, the action of any element of the whole group leaves a normal subgroup alone or, more technically, we say that each element of the whole group *commutes* with the normal subgroup. Let me be more explicit. Call the whole group $\mathcal{G}$ and the subgroup $\mathcal{S}$. If I select any particular element g of the group $\mathcal{G}$, then I can denote by $\mathcal{S}g$ the set consisting of all elements of $\mathcal{S}$ each individually multiplied by g on the right (what is called *postmultiplied* by g). Thus, in the case of the particular subgroup $\mathcal{S} = \{1, -1, \mathsf{C}, -\mathsf{C}\}$, of the symmetry group of the square, if we choose $g = \mathrm{i}$, then we obtain $\mathcal{S}\mathrm{i} = \{\mathrm{i}, -\mathrm{i}, \mathsf{C}\mathrm{i}, -\mathsf{C}\mathrm{i}\}$. Likewise, the notation $g\mathcal{S}$ will denote the set consisting of all elements of $\mathcal{S}$, each individually multiplied by g on the left (*premultiplied* by g). Thus, in our example, we now have $\mathrm{i}\mathcal{S} = \{\mathrm{i}, -\mathrm{i}, \mathrm{i}\mathsf{C}, -\mathrm{i}\mathsf{C}\}$. The condition for $\mathcal{S}$ to be a normal subgroup of $\mathcal{G}$ is that these two sets are the same, i.e.

$$\mathcal{S}g = g\mathcal{S}, \quad \text{for all } g \text{ in } \mathcal{S}.$$

In our particular example, we see that this is indeed the case (since $\mathsf{C}\mathrm{i} = -\mathrm{i}\mathsf{C}$ and $-\mathsf{C}\mathrm{i} = \mathrm{i}\mathsf{C}$), where we must bear in mind that the collection

---

[13.5] Verify that all these in this paragraph are subgroups (and bear in mind Note 13.4).

of things inside the curly brackets is to be taken as an *unordered* set (so that it does not matter that the elements $-\mathrm{i}\mathsf{C}$ and $\mathrm{i}\mathsf{C}$ appear in reverse order in the collection of elements, when $\mathcal{S}\mathrm{i}$ and $\mathrm{i}\mathcal{S}$ are written out explicitly).

We can exhibit a *non*-normal subgroup of the group of symmetries of the square, as the subgroup of two elements $\{1, \mathsf{C}\}$. It is non-normal because $\{1, \mathsf{C}\}\mathrm{i} = \{\mathrm{i}, \mathsf{C}\mathrm{i}\}$ whereas $\mathrm{i}\{1, \mathsf{C}\} = \{\mathrm{i}, -\mathsf{C}\mathrm{i}\}$. Note that this subgroup arises as the new (reduced) symmetry group if we mark our square with a horizontal arrow pointing off to the right (see Fig. 13.3a). We can obtain another non-normal subgroup, namely $\{1, \mathsf{C}\mathrm{i}\}$ if we mark it, instead, with an arrow pointing diagonally down to the right (Fig. 13.3b).[13.6] In the case of O(3), there happens to be only one non-trivial normal subgroup,[13.7] namely SO(3), but there are many non-normal subgroups. Non-normal examples are obtained if we select some appropriate finite set of points on the sphere, and ask for the symmetries of the sphere with these points marked. If we mark just a single point, then the subgroup consists of rotations of the sphere about the axis joining the origin to this point (Fig. 13.3c). Alternatively, we could, for example, mark points that are the vertices of a regular polyhedron. Then the subgroup is finite, and consists of the symmetry group of that particular polyhedron (Fig. 13.3d).

One reason that normal subgroups are important is that, if a group $\mathcal{G}$ possesses a non-trivial normal subgroup, then we can break $\mathcal{G}$ down, in a sense, into smaller groups. Suppose that $\mathcal{S}$ is a normal subgroup of $\mathcal{G}$. Then the distinct sets $\mathcal{S}g$, where g runs through all the elements of $\mathcal{G}$, turn

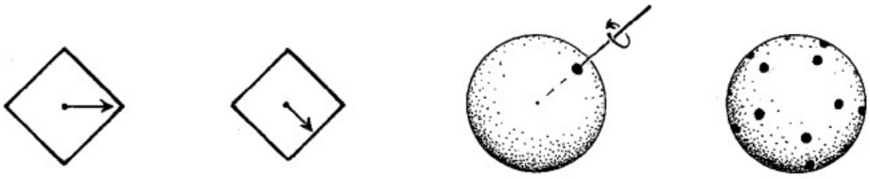

(a) (b) (c) (d)

Fig. 13.3 (a) Marking the square of Fig. 13.1 with an arrow pointing to the right, reduces its symmetry group to a non-normal subgroup $\{1,\mathsf{C}\}$. (b) Marking it with an arrow pointing diagonally down to the right yields a different non-normal subgroup $\{1,\mathsf{C}\mathrm{i}\}$. (c) Marking the sphere of Fig. 13.2 with a single point reduces its symmetry to a (non-normal) O(2) subgroup of O(3): rotations about the axis joining the origin to this point. (d) If the sphere is marked with the vertices of a regular polyhedron (here a dodecahedron), its group of symmetries is a finite (non-normal) subgroup of O(3).

[13.6] Check these assertions, and find two more non-normal subgroups, showing that there are no further ones.

[13.7] Show this. (*Hint*: which *sets* of rotations can be rotation-invariant?)

out themselves to form a group. Note that for a given set $\mathcal{S}g$, the choice of g is generally not unique; we can have $\mathcal{S}g_1 = \mathcal{S}g_2$, for different elements g_1, g_2 of $\mathcal{G}$. The sets of the form $\mathcal{S}g$, for *any* subgroup $\mathcal{S}$, are called *cosets* of $\mathcal{G}$; but when $\mathcal{G}$ is normal, the cosets form a group. The reason for this is that if we have two such cosets $\mathcal{S}g$ and $\mathcal{S}h$ (g and h being elements of $\mathcal{G}$) then we can define the 'product' of $\mathcal{S}g$ with $\mathcal{S}h$ to be

$$(\mathcal{S}g)\,(\mathcal{S}h) = \mathcal{S}(gh),$$

and we find that all the group axioms are satisfied, provided that $\mathcal{S}$ is normal, essentially because the right-hand side is well defined, independently of which g and h were chosen in the representation of the cosets on the left-hand side of this equation.[13.8] The resulting group defined in this way is called the *factor group* of $\mathcal{G}$ by its normal subgroup $\mathcal{S}$. The factor group of $\mathcal{G}$ by $\mathcal{S}$ is written $\mathcal{G}/\mathcal{S}$. We can still write $\mathcal{G}/\mathcal{S}$ for the factor *space* (not a group) of distinct cosets $\mathcal{S}g$ even when $\mathcal{S}$ is not normal.[13.9]

Groups that possess no non-trivial normal subgroups at all are called *simple groups*. The group SO(3) is an example of a simple group. Simple groups are, in a clear sense, the basic building blocks of group theory. It is thus an important achievement of the 19th and 20th centuries in mathematics that all the finite simple groups and all the continuous simple groups are now known. In the continuous case (i.e. for Lie groups), this was a mathematical landmark, started by the highly influential German mathematician Wilhelm Killing (1847–1923), whose basic papers appeared in 1888–1890, and was essentially completed, in 1894, in one of the most important of mathematical papers ever written,[5] by the superb geometer and algebraist Élie Cartan (whom we have already encountered in Chapter 12, and whom we shall meet again in Chapter 17). This classification has continued to play a fundamental role in many areas of mathematics and physics, to the present day. It turns out that there are four families, known as A_m, B_m, C_m, D_m(for $m = 1, 2, 3, \ldots$), of respective dimension $m(m+2)$, $m(2m+1)$, $m(2m+1)$, $m(2m-1)$, called the *classical groups* (see end of §13.10) and five *exceptional groups* known as E_6, E_7, E_8, F_4, G_2, of respective dimension 78, 133, 248, 52, 14.

The classification of the finite simple groups is a more recent (and even more difficult) achievement, carried out over a great many years during the 20th century by a considerable number of mathematicians (with the aid of computers in more recent cases), being completed only in 1982.[6] Again there are some systematic families and a finite collection of *exceptional*

[13.8] Verify this and show that the axioms fail if $\mathcal{S}$ is not normal.

[13.9] Explain why the number of elements in $\mathcal{G}/\mathcal{S}$, for any finite subgroup $\mathcal{S}$ of $\mathcal{G}$, is the order of $\mathcal{G}$ divided by the order of $\mathcal{S}$.

finite simple groups. The largest of these exceptional groups is referred to as the *monster*, which is of order

$$= 808017424794512875886459904961710757005754368000000000.$$
$$= 2^{46} \times 3^{20} \times 5^{9} \times 7^{6} \times 11^{2} \times 13^{3} \times 17 \times 19 \times 23 \times 29 \times 31 \times 41 \times 47 \times 59 \times 71.$$

Exceptional groups appear to have a particular appeal for many modern theoretical physicists. The group E_8 features importantly in string theory (§31.12), while various people have expressed a hope that the huge but finite monster may feature in some future theory.[7]

The classification of the simple groups may be regarded as a major step towards the classification of groups generally since, as indicated above, general groups may be regarded as being built up out of simple groups (together with Abelian ones). In fact, this is not really the whole story because there is further information in how one simple group can build upon another. I do not propose to enter into the details of this matter here, but it is worth just mentioning the simplest way that this can happen. If $\mathcal{G}$ and $\mathcal{H}$ are any two groups, then they can be combined together to form what is called the *product group* $\mathcal{G} \times \mathcal{H}$, whose elements are simply *pairs* (g, h), where g belongs to $\mathcal{G}$ and h belongs to $\mathcal{H}$, the rule of group multiplication between elements (g_1, h_1) and (g_2, h_2), of $\mathcal{G} \times \mathcal{H}$, being defined as

$$(g_1, h_1)\,(g_2, h_2) = (g_1 g_2,\, h_1 h_2),$$

and it is very easy to verify that the group axioms are satisfied. Many of the groups that feature in particle physics are in fact product groups of simple groups (or elementary modifications of such).[13.10]

13.3 Linear transformations and matrices

In the general study of groups, there is a particular class of symmetry groups that have been found to play a central role. These are the groups of symmetries of vector spaces. The symmetries of a vector space are expressed by the *linear transformations* preserving the vector-space structure.

Recall from §11.1 and §12.3 that, in a vector space V, we have, defining its structure, a notion of addition of vectors and multiplication of vectors by numbers. We may take note of the fact that the geometrical picture of addition is obtained by use of the parallelogram law, while multiplication by a number is visualized as scaling the vector up (or down) by that number (Fig. 13.4). Here we are picturing it as a *real* number, but complex vector spaces are also allowed (and are particularly important in many

---

[13.10] Verify that $\mathcal{G} \times \mathcal{H}$ is a group, for any two groups $\mathcal{G}$ and $\mathcal{H}$, and that we can identify the factor group $(\mathcal{G} \times \mathcal{H})/\mathcal{G}$ with $\mathcal{H}$.

5 !"# $ % & R A&%% $ %
(& $C * @ 9 (1 0 -*

!"# $ % & R A&%% $ % 8
(& $C * @ 9 (@ ! 1 E - .

/% 10 > % 0

fingerprints on the particles and forces that emerge *after* this period, and which we can observe today at more modest energies and temperatures. To date, no GUT has yielded empirical fingerprints that are consistent with what we observe.

What does the fledgling universe contain? This is really hard to say. Most cosmological models assume that the early universe behaves much like a hot uniform gas, without being specific about the nature of the gas. If there are particles of some kind, then we can safely assume that they are massless and highly energetic, and bear scant resemblance to any particle with which we are familiar today.

The universe continues to expand and cool, to a temperature of about 10^{27} kelvin. About 10^{-35} seconds after the big bang we hit another phase transition, as the strong nuclear force condenses and separates from the residual 'electro-weak' force.

But this time, something is different.

COSMIC INFLATION: 10^{-35} TO 10^{-32} SECONDS

When the ice melts in the pan, the symmetry increases and the water molecules are freed from their prison in the crystal lattice. A considerable amount of energy in the form of heat is required to complete this phase transition. This heat is used to free the water molecules from their restraints, but it does not change the temperature during the transition (which remains fixed at the freezing point of water: 0°C, or 273 kelvin). Such a phase transition is said to be *endothermic*—it requires an input of thermal energy to make it happen.

Conversely, when liquid water freezes to form ice, this energy or 'latent heat' is released and the transition is said to be *exothermic*—there is an output of thermal energy. This heat is released in a short burst as the ice forms. When we make ice in the refrigerator we don't normally notice this, but it can be easily measured using a calorimeter.

In my kitchen experiment, I allow the hot steam to condense to liquid before popping the pan into my freezer, where the liquid will condense further to form ice. This is a conventional freezer, which will cool the liquid in its own good time. But let's now suppose that the water in the pan is ultra-pure, and the inside of the pan is ultra-smooth. This greatly reduces the number of possible sites at which the first ice crystals can nucleate. If I now cool the pan very rapidly, I find that I can reduce the temperature of the liquid well below its freezing point, to almost −50°C (223 kelvin), in fact, without forming any ice. This is *supercooling*.

In supercooled water, the latent heat is not immediately released. It remains latent.

In 1979, young American postdoctoral researcher Alan Guth and his colleague Henry Tye were trying to resolve some of the puzzles of early big bang cosmology. They asked themselves what might happen if, instead of undergoing a smooth phase transition at the end of the Grand Unification Epoch, the universe had instead persisted in its grand unified phase at temperatures several orders of magnitude *below* the transition temperature.

This did indeed resolve their problem. But when Guth went on to explore the wider effects of the onset of supercooling, he found much more than he had initially bargained for. He found that the latent heat of the supercooled universe is stored in the vacuum, in spacetime itself. Remember, vacuum energy does much the same job as Einstein's cosmological term, forcing the rate of expansion of spacetime to increase. Shovelling a very large amount of energy into spacetime results in a short burst of extraordinary exponential expansion, doubling the size of the universe every 10^{-34} of a second. One hundred such doublings would have increased the size of the universe by a factor of 10^{30}.

'I do not remember ever trying to invent a name for this extraordinary phenomenon of exponential expansion,' Guth later wrote, 'but my diary shows that by the end of December [1979] I had begun to call it *inflation*.'[10]

Is this really the way it happened? Theories of cosmic inflation are based on the assumed properties of GUTs and, as there is no generally accepted grand unified theory, then the status of inflation remains rather questionable. But there are some features of the observable universe that are quite handily explained if we assume that something like inflation actually did happen.

Friedmann had shown in 1924 that different assumptions about the density of mass-energy would give rise to different kinds of universe. In a closed universe, spacetime would have a positive curvature, like the surface of a sphere. Parallel lines will cross and the angles of a triangle will add up to more than 180°. In an open universe, spacetime would have a negative curvature, looking something like a saddle. Parallel lines would again cross and the angles of a triangle would add up to less than 180°.

It would seem that we live in a universe in which there is a very fine balance between the density of mass-energy and the rate of expansion, such that spacetime is flat (Figure 5). This is why we learn Euclidean geometry in school.

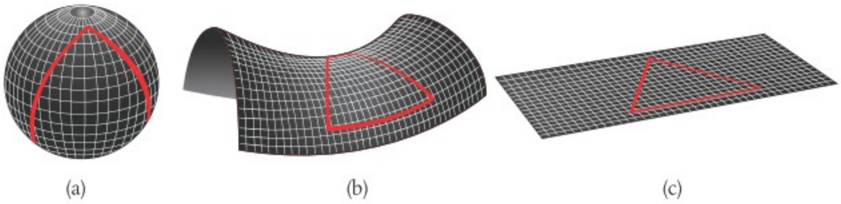

FIGURE 5 The different model universes that Friedmann identified in solutions to Einstein's gravitational field equations produce differently curved spacetimes. The 'closed' universe has a positively curved spacetime as shown in (a). The 'open' universe has a saddle-shaped negatively curved spacetime, (b). A universe with a finely tuned balance between the rate of expansion and the amount of mass-energy it contains will exhibit a flat spacetime, (c).

But why should our universe exhibit such a balance? Scientists get rather uncomfortable when confronted with evidence of such fine-tuning, especially when they can't explain it using physical mechanisms.

This is where inflation comes to the rescue. It really makes no difference precisely how much energy the early universe contained, or whether this was equal to the critical density required to produce a flat spacetime. It doesn't matter what the initial rate of expansion was. It doesn't even matter if inflation is applied to the whole of the early universe or just one small bubble of spacetime within this. No matter what the shape of spacetime prior to inflation, when inflation was done, *flat* spacetime was the only possible result.

Think of it this way. The skin of a deflated balloon may be wrinkled like a prune, curving this way and that. Pump it up and the wrinkles are quickly flattened out. The visible universe may be flat because it comes from a tiny bit of grand-unified universe blown up out of all proportion by inflation.

Inflation theory has many critics, and some have argued that there are ways of arriving at a flat spacetime that do not require inflation, or fine-tuning for that matter. Nevertheless, for the time being at least, it remains an important component of the broadly accepted theory describing the evolution of the universe.

REHEATING

Inflation is thought to have ceased around 10^{-32} seconds after the big bang. The temperature is now several orders of magnitude below the transition temperature.

2022: Feliz Ano Novo!!!

There are four fundamental forces at work in the Universe:

(20220101)

1) The strong force
2) The Weak force
3) The Electromagnetic Force
4) The Gravitational Force.

reflexo
(efeito)

The Gravitational Force é decorrência da "The Strong Force".

Se a gravidade existisse independentemente das demais forças, o Universo jamais teria sido criado pois antes do Big-Bang tudo estava reunido em um único ponto e portanto, a própria gravidade teria impedido a própria explosão do Big-Bang.

A gravidade deve ser analisada localmente e quando estendida, você terá outro valor.

A extensão da análise do fator gravitacional em relação a todo Universo gera uma curvatura ~~do~~ espaço-tempo.

O Universo se expande pois a "gravidade" age localmente em relação à cada galáxia e não em relação ao todo.

!"# $ % & 0 7 E5 -*
$.
! . + & * D * $ *

!"# $ % & 0 5 K <K K 5 -*
$.
! . + & * D * $ * 6& S

!"# $ % & 0 5 7 55 -*
$.
+ &% +C- & * .(" & & ' (
! 1 6& S

8 ? 6
3 3 ? 00< 14 03 5 3 -- 00 0<
. * .((T .
&. @"6 - 2 % &. 5 7 55
! * < < 5 @8

!"# $ % & 0 7E 7 0 -*
$.
!6A+"8 +)? 8+)? !?4 8,+ "$6+ $> 8,+$"B)? $?8+ $"!"B ,B)
! 1 $"?+6)@ 6!? U$

! & 3
73 ? 1 !" , 505 3 -- 05
. * .(% & V 3 8 " & (
&. @"6 - 2 % &. 7E 7 0
! * 0

!"# $ % & 0 K K 5 -*

0 !"# $ % & < R& . %(J
(@ 9 (1 -*

< !"# $ % & 7 > C Q + & (
& & & !*9& * & *- ' (4 9 . & 1 -*
$. 6 9 * ===

5 7 7

!"#$! ! !% & ' !
() *(+ , - . / 0-

!"#$! ! !% & ' ! 1' '2-3 4 * 5 ' " 6 ' -
6 - . 0 &/ ' - 7 0/ # 8 0-
%9 : 4 - ;;;

!"#$! ! !% & ' ! " 6 ' - 6 -
. 0 &/ ' - 7 0/ ' ' ' . 0 &/ '0 ' ' - 7' (#
8 < 0-

!"#$! ! !% & ' ! ' 1' '2-3 4 * ' ' - "
" 6 - ' ' ' '0 ' ' - 7' (# 8 0-
%9 : 4 - ;;;

= !"#$! ! !% & ' ! " ' +) ' - " & 9 ' ' -
> / ' 9 -& ? 4 / 8 = 0-

@ !"#$! ! !% & ' ! 0 , " ' / . " &
0 6 " ' / " - . 0 &/ A B / C 8 < 0-

< !"#$! ! !% & ' ! ' ' ' D,0 '
' 9 -& ? 4 / 8 @ 0-

E !"#$! ! !% & ' ! . > / F > 0
0 & '& G H * 8 E 0-
I& @8 < J< . 7/0 9 '

!"#$! ! !% & ' ! " # ' '
' 9 -& ? 4 / 8 E 0-

!"#$! ! !% & ' ! " # ' '
' 9 -& ? 4 / 8 E -24

!"#$! ! !% & ' !) ' - 6 ' ' '
0 & 8 0-

!"#$! ! !% & ' !) ' - 6 ' ' '
0 & 8 -24

!"#$! ! !% & ' !) ' - :6 6. " G
"00 ' # ' " ' / 0 & 8 < 0-

!"#$! ! !% & ' !) ' - :6 6. " G
"00 ' # ' " ' / 0 & 8 < -24

= !"#$! ! !% & ' ! I 6 0-
I& =

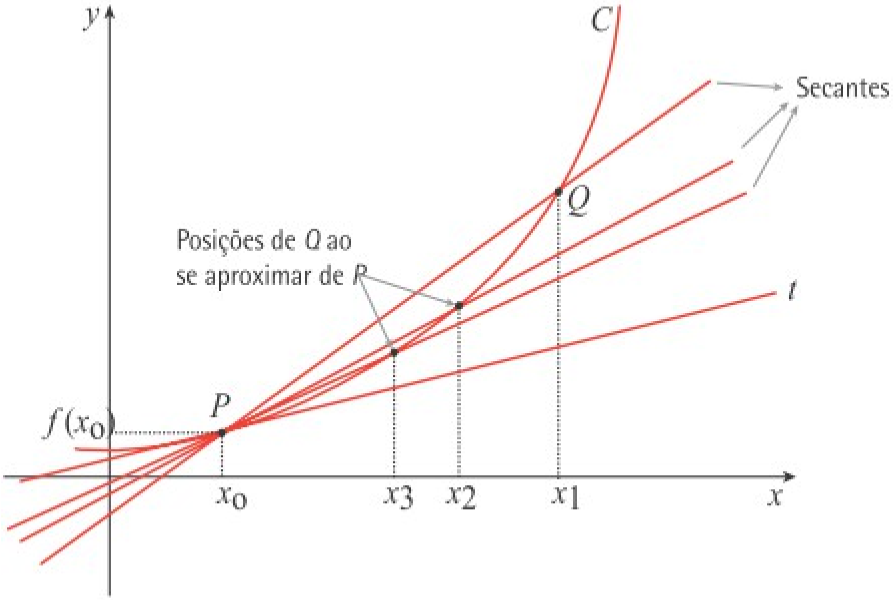
y
C
Secantes
Q
Posições de Q ao
se aproximar de P
t
f(x0)
P
x0
x3
x2
x1
x

+ & '

@ !"#$! ! !% & ' ! I 66 0-

< !"#$! ! !% & ' ! =
:' (" (# & 9 &(> / 0-

E !"#$! ! !% & ' ! @ +K , " ' ' I3L
9 ' - & M' / N 4 0-
%9 : 4 - ;;;

!"#$! ! !% & ' ! <
O ' (1 P(. (: &' (6 & ' ((0-

!"#$! ! !% & ' ! <
O ' (1 P(. (: &' (6 & ' ((0-

!"#$! ! !% & ' ! E
O '- (7' ' P(' ' ' (- ((- 0-

!"#$! ! !% & ' ! E
O '- (7' ' P(' ' ' (- ((- -24

!"#$! ! !% & ' ! > * -
1 0 / * @@ @ < 0-

!"#$! ! !% & ' ! 7 3 > Q - ' >
* - 1 0 / * 0 & 0 * 8 0-
%9 : 4 - ;;;

= !"#$! ! !% & ' ! - & 0-

@ !"#$! ! !% & ' ! - & 0-
%9 : 4 - ;;; =@ 0'&

I& 8 E

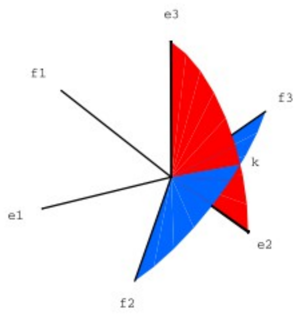
e3
f1
f3
k
e1
e2
f2

id

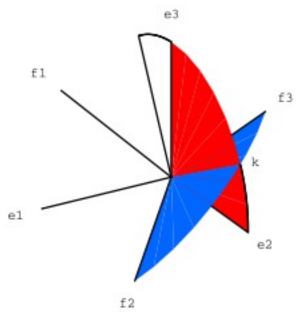
e3
f1
f3
k
e1
e2
f2

R_1

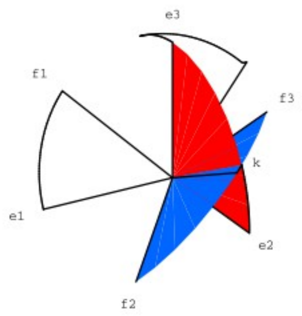
e3
f1
f3
k
e1
e2
f2

R_2

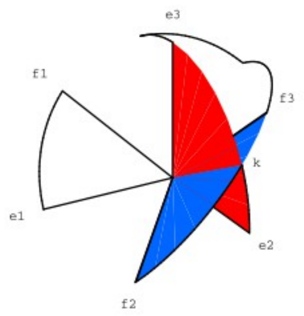
e3
f1
f3
k
e1
e2
f2

R_3

$$[R_1]_{\mathcal{E}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\varphi_1 & -\sin\varphi_1 \\ 0 & \sin\varphi_1 & \cos\varphi_1 \end{pmatrix}$$

$$[R_2]_{\mathcal{E}'} = \begin{pmatrix} \cos\varphi_2 & 0 & -\sin\varphi_2 \\ 0 & 1 & 0 \\ \sin\varphi_2 & 0 & \cos\varphi_2 \end{pmatrix}$$

$$[R_1]_{\mathcal{E}''} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\varphi_3 & -\sin\varphi_3 \\ 0 & \sin\varphi_3 & \cos\varphi_3 \end{pmatrix}$$

I& 8 E

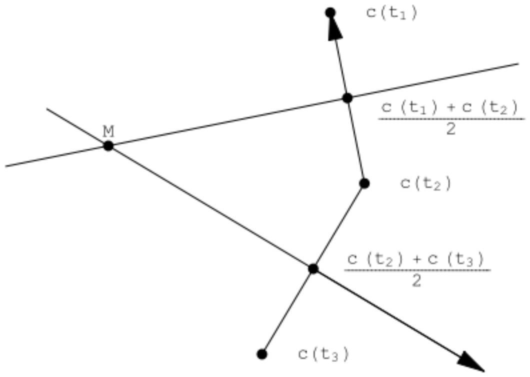
c(t$_1$)
$\frac{c(t_1)+c(t_2)}{2}$
M
c(t$_2$)
$\frac{c(t_2)+c(t_3)}{2}$
c(t$_3$)

< !"#$! ! !% & ' ! - & 0-
%9 6 - F ' > /

E !"#$! ! !% & ' ! 1 4 - + ' #' 0-

!"#$! ! !% & ' ! ' > & F ' 3 -
/ '& ' @=) ' ' 5 * & - - -'* R9 & &
3 4 9 Q . - ' ' * 0-

!"#$! ! !% & ' !) 4
' ' * (", ' *5 / '& ' - 3' ' D, 0 - : * - -'* *
0 & 0 * 8 = 0-

!"#$! ! !% & ' ! S & - 0
- ' / 0 2 4 & / 0-

!"#$! ! !% & ' ! 9 " & 9
> 5)' - (> - D9 8 E<@ 0-

!"#$! ! !% & ' ! 9 " & 9
> 5)' - (> - D9 8 E<@ -24

!"#$! ! !% & ' ! >' + B *
B' * & - >)' - 8 E< 0-

= !"#$! ! !% & ' ! >' + B *
B' * & - >)' - 8 E< -24

@ !"#$! ! !% & ' ! 7 9 F > '
"9 ' - & " & 9 '5 6 4' ' 5 >)' - 0 & 8 0-

< !"#$! ! !% & ' ! 7 9 F > '
"9 ' - & " & 9 '5 6 4' ' 5 >)' - 0 & 8 -24

E !"#$! ! !% & ' ! 1 + & 9
> 0 & 8 @ 0-

!"#$! ! !% & ' ! 1 + & 9
> 0 & 8 @ -24

!"#$! ! !% & ' ! : & - " D -
& ' 8 E 0-

!"#$! ! !% & ' ! : & - " D -
& ' 8 E -24

!"#$! ! !% & ' ! D / 0' - - ' '
B ' - " - & >)' - 5 ' 5 6 8 0-

!"#$! ! !% & ' ! D / 0' - - ' '
B ' - " - & >)' - 5 ' 5 6 8 -24

!"#$! ! !% & ' ! D / 0' - - ' '
B ' - " - & >)' - 5 ' 5 6 8 0-

= !"#$! ! !% & ' ! D / 0' - - ' '
B ' - " - & >)' - 5 ' 5 6 8 -24

@ !"#$! ! !% & ' ! D / 0' - - ' '
B ' - " - & >)' - 5 ' 5 6 8 = 0-

< !"#$! ! !% & ' ! D / 0' - - ' '
B ' - " - & >)' - 5 ' 5 6 8 = -24

E !"#$! ! !% & ' ! D / 0' - - ' '
B ' - " - & >)' - 5 ' 5 6 " 8 E 0-

!"#$! ! !% & ' ! D / 0' - - ' '
B ' - " - & >)' - 5 ' 5 6 " 8 E -24

= !"#$! ! !% & ' ! D / 0' - - ' '
B ' - " - & >)' - 5 ' 5 6) 8 0-

= !"#$! ! !% & ' ! D / 0' - - ' '
B ' - " - & >)' - 5 ' 5 6) 8 -24

= !"#$! ! !% & ' ! D / 0' - - ' '
B ' - " - & >)' - 5 ' 8 < 0-

= !"#$! ! !% & ' ! D / 0' - - ' '
B ' - " - & >)' - 5 ' 8 < -24

= !"#$! ! !% & ' ! = G ' %0 '5
' ' ' - F 4 B / 8 0-

!F"." G 4 4 !) * ! =!G ' %0 '5 ' '
' - F 4 !G ' %0 '5 ' ' ' - F 4 B / 8 0-

I& 8=

%9 : 4 - ;;;

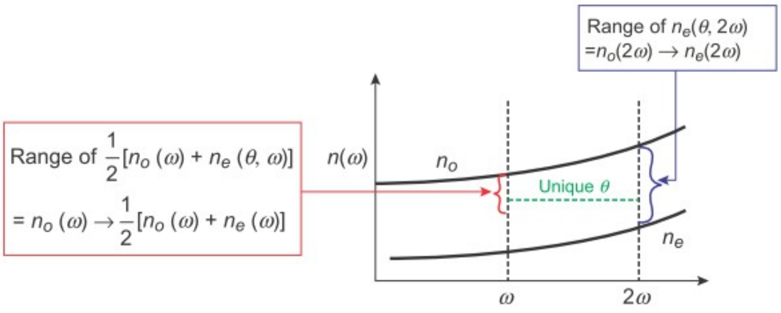
Range of $n_e(\theta, 2\omega)$
$=n_o(2\omega) \rightarrow n_e(2\omega)$
Range of $\frac{1}{2}[n_o(\omega) + n_e(\theta, \omega)]$
$= n_o(\omega) \rightarrow \frac{1}{2}[n_o(\omega) + n_e(\omega)]$
$n(\omega)$
n_o
Unique θ
n_e
ω
2ω

FIGURE 3.5 The dispersion in n_e and n_o for a negative uniaxial crystal. Both $n_e(\theta, \omega)$ and $n_e(\theta, 2\omega)$ are tuned with angle θ. The dashed green line indicates the wave-vector-matching condition. The ranges over which the refractive index for the fundamental and the harmonic can vary are also shown.

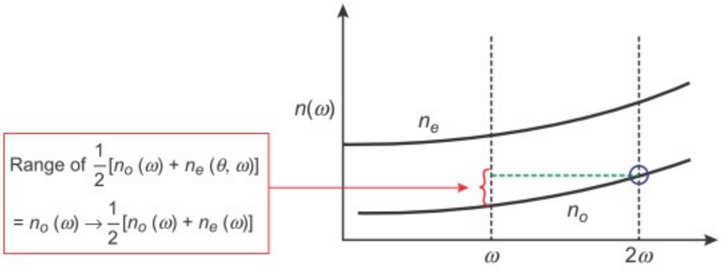
$n(\omega)$
n_e
Range of $\frac{1}{2}[n_o(\omega) + n_e(\theta, \omega)]$
$= n_o(\omega) \rightarrow \frac{1}{2}[n_o(\omega) + n_e(\omega)]$
n_o
ω
2ω

FIGURE 3.6 The dispersion in n_e and n_o for a positive uniaxial crystal. $n_e(\theta, \omega)$ is tuned with angle θ. The dashed green line indicates the wave-vector-matching condition.

E E E

!"#$! ! !% & ' ! @ " . , 9 * 6 & ' '
ADD ' C "-4' - 0-

!F"." G 4 4 !) * ! @!" . , 9 * 6 & ' '
ADD ' C "-4' -!" . , 9 * 6 & ' ' ADD ' C "-4' - 0-

!"#$! ! !% & ' ! < F ' FK ' K -
' K ' L 0-

!F"." G 4 4 !) * ! @!F ' FK ' K -
' K ' L !F ' FK ' K - ' K ' L 0-

!"#$! ! !% & ' ! E 6 " G ' 7 /&
' ' ' 0 G' ' 0 / ' - %0 9/ G 5 . ' : & 9/
7 /& 0-

!F"." GT !)%%1 T ! <!D? :6FT !8> ') * B B - 8
6 " G (' 7 /& ' ' ' 0 G' ' 0 / ' - %0 9/
G 5 . ' : & 9/ 7 /& 0-

!"#$! ! !% & ' ! E D - " - "0
&' G ' > ' ' D -U D 5 * " - 5 % 9/
"0 &' 0-

!F"." GT !)%%1 T ! <!D? :6FT !D -(" - ("0 &'(
G ' > ' ' D -U D 5 * " - 5 % 9/ "0
&' ' - 6 - 0-

= !"#$! ! !% & ' ! E * 9/ B ' > 9 5
>' >' 5 B ' 7' 4 / D / 0' - ') ' ' 8 = 0-

!F"." G 4 4 !) * ! <!D -U D 5 *
" - 5 % 9/ "0 &'! * 9/ B ' > 9 5
>' >' 5 B ' 7' 4 / D / 0' - ') ' ' 8 = 0-

@ !"#$! ! !% & ' ! E B * ' / 9/ /5
0 0-

!F"." G 4 4 !) * ! <!D -U D 5 *
" - 5 % 9/ "0 &'!B * ' / 9/ /5
0 0-

< !"#$! ! !% & ' !
+ -' ((%0 ' (/ (" ' / (' -(F & 0-

!F"." G 4 4 !) * ! <!+ -' %0 ' / " ' /
' - F & !+ -' ((%0 ' (/ (" ' / (' -(F & 0-

E !"#$! ! !% & ' ! 3- & 0 0-

!F"." G 4 4 !) * ! <!> '! 3- & 0 0-
%9 : 3 - > ' 4'

!"#$! ! !% & ' ! 3 - & ' ' ' 0-

!F"." G 4 4 !) *
! <!> '! 3 - & ' ' ' 0-
V : 3 - > ' " ' ' 4' ' 4

!"#$! ! !% & ' ! 3 - & 9 0-

!F"." G 4 4 !) *
! <!> '! 3 - & 9 0-
V : 3 - > ' " ')

!"#$! ! !% & ' ! L & ' - 0-

!F"." G 4 4 !) * ! <!> '!L & ' - 0-
V $ # & ' -' :' > ' D '

!"#$! ! !% & ' ! B * ' / 9/ /5
0 0-

!F"." G 4 4 !) * ! <!1 0) ' '!B * ' /
9/ /5 0 0-

!"#$! ! !% & ' ! D / 0' - -
' ' B ' - " - & >)' - 5 ' 8 < -24

!"#$! ! !% & ' ! D / 0' - -
' ' B ' - " - & >)' - 5 ' 8 < 0-

!"#$! ! !% & ' ! D U 0' ' -
0 ' ' - & ' ' 4 / 0-

!F"." G 4 4 !) * ! <!$ ' + - . / 4 -
0' $ ' 3 - + - ' - > '4 /!D U 0' ' - 0 ' ' -
& ' ' 4 / 8 F+F 4 0-

= !"#$! ! !% & ' ! D U / ' &
- & ' ' 4 / ' ' ' / ' - 0-

!F"." G 4 4 !) * ! <!$ ' + - . / 4 -
0' $ ' 3 - + - ' - > '4 /!D U / ' & - & '
' 4 / ' ' ' / ' - 8 F+F 4 0-

@ !"#$! ! !% & ' ! > ' # ' 4 / " 6 -
/ 8 ? 5 @ 86)G = E = E 0-

!F"." G 4 4 !) * ! <!$ ' + - . / 4 -
0' $ ' 3 - + - ' - > '4 /!> ' # ' 4 / " 6 - / 8 ? 5
@ 86)G = E = E 0-

< !"#$! ! !% & ' ! & ' (' 4 / 0-

!F"." G 4 4 !) * ! <!$ ' + - . / 4 -
0' $ ' 3 - + - ' - > '4 /!& ' (' 4 / 0-

E !"#$! ! !% & ' ! > '4 ' 8 / 0-

!"#$! ! !% & ' ! > '4 / ' - D U
& ' ' 4 / 0-

!"#$! ! !% & ' ! > '4 / ' - $ '
-'& & ' D '/ &/5 " 0 / 5 ' - $ ' > '4 / 0-

!"#$! ! !% & ' ! & 9 * 0-
V > ' # ' 4 /

!"#$! ! !% & ' ! ' ' ' - /
F 9 5 7 9 0' %0 ' 5 ' ' ' - 5 ' - "00 ' $ '
/ 0-

!"#$! ! !% & ' ! - ' ' 0 / 0 5

- 5 0 ' ' 5 ' /5 ' 4 /5 ' - ' ' 0 / 0-

!"#$! ! !% & ' ! $ ' + - . /" -
6 - 0-

= !"#$! ! !% & ' ! $ ' + - . / 4 -
0' $ ' 3 - + - ' - > '4 / 0-

@ !"#$! ! !% & ' ! $ ' > '4 / 0-

< !"#$! ! !% & ' ! 0' ' - > /"
6 - > ' # ' 4 / 0-

E !"#$! ! !% & ' ! 0' ' - > /"
6 - > ' # ' 4 / % 0-

!"#$! ! !% & ' ! 0 ' ' 4 /5 -/ ' 5
' - & ' ' 4 / G D 0-

!"#$! ! !% & ' ! . > ' . / # ' 4 /"
' ' ' D,0 0-

!"#$! ! !% & ' ! . > ' . / # ' 4 /"
' ' ' D,0 0-

!"#$! ! !% & ' ! = H ' > /
' ' 7 / ' - 0-

!"#$! ! !% & ' ! "-4' - 9 * ' '
% / 0 '- 0-

!"#$! ! !% & ' ! " & 9 ' C . & / 9
4 0-

= !"#$! ! !% & ' ! " & 9 ' ' - . & /
" ' / > / 0-

@ !"#$! ! !% & ' ! " & 9 ' ' - . & /5
D- 0-

< !"#$! ! !% & ' ! " D, & D ' /
' ' 5 6 # ' G 9 ' - + 0-

E !"#$! ! !% & ' ! " D, & D ' /
' ' 5 66 D - ' > / 0-

!"#$! ! !% & ' ! " D, & D ' /
' ' 5 666 F ' ' ' - / ' " & 9 ' 0-

!"#$! ! !% & ' ! " 6 - / ' & '0 /5
F ' ' - / / 0-

!"#$! ! !% & '! "'/ ' > / F ' - F 0-

!"#$! ! !% & '! ' ' > / D - ' 5
. ' ' ' 5 6 4 4 5 ' - 2 4 0-

!"#$! ! !% & '! - ' ' - - '
& 0-

!"#$! ! !% & '! D - ' ' - G D - '
> / " " ' / "00 ' 0-

= !"#$! ! !% & '! + - L ' 3' 0-

@ !"#$! ! !% & '! + -' D - ' ' - G
D - ' > / 0-

< !"#$! ! !% & '! > / / '
' - . & 0-

E !"#$! ! !% & '! > / 6 ' - " 6 ' -
6 - D - ' ' - 7/0 9 ' > / 0-

!"#$! ! !% & '! > / . & 7 /
D - ' 5 7/0 9 5 ' - 2 4 > 0-

= !"#$! ! !% & '! > / . & 7 /
D - ' 5 7/0 9 5 ' - 2 4 > 0-

= !"#$! ! !% & '! > 0 . / 0-

= !"#$! ! !% & '! 6 ' F & " ' ' '
"00 ' 0-

= !"#$! ! !% & '! 6 ' > ' .
7 ' F 4 0 ' - . '- ' - 0-

= !"#$! ! !% & '! 6 ' > ' .
7 ' F 4 0 ' - . '- ' - 0-
V = @ 0'& %9 : 4 - ;;;

== !"#$! ! !% & '! ' ' 0-

=@ !"#$! ! !% & '! - > / 0-

=< !"#$! ! !% & '! ' / / 0-

=E !"#$! ! !% & '! %, - 7' -9 * 7 /
' ' 0-

= !"#$! ! !% & '! %, - ? U > -
' ' 0-

@ !"#$! ! !% & ' ! / / /0-

@ !"#$! ! !% & ' ! 9 ' -
D - ' > /0-

@ !"#$! ! !% & ' ! $ ' + - . / 4 -
0' $ ' 3 - + - ' - > '4 /0-
V 6 0 ' ;;; 4 - 0'
%9 : 4 - ;;;

@ !"#$! ! !% & ' ! $ ' ' /
- D- 0-

@ !"#$! ! !% & ' ! # ' 4 $ ' '
B'4 DL ' 5 - - B > 0-

@= !"#$! ! !% & ' ! # ' 4 $ ' '
B'4 DL ' 5 - - B > 0-

@@ !"#$! ! !% & ' ! ' ' " 0' /
' ' > / D - ' 5 . ' ' ' 5 6 4 4 5 ' - 2 4 0-

@< !"#$! ! !% & ' ! 0 ' . / ' - $ '
' ' ' ' + -' $ ' . 5 / ' - 6 -
" & 9 ' + ' 0-

@E !"#$! ! !% & ' ! / + -' '
/ 0-

@ !"#$! ! !% & ' ! / 6 ' > '
' 0-

< !"#$! ! !% & ' ! / / ' 2 /
0' ' 0-

< !"#$! ! !% & ' ! . /) * +
> 0 - > '0 5 = 7 / /0-

< !"#$! ! !% & ' ! . %, - F ' /
' ' 0-

< !"#$! ! !% & ' ! . L ' ' - U 9 4 -
7 ' ' ' & - 4 - ' & '& / /0-

< !"#$! ! !% & ' ! . 7 / ' ' ")
0-

<= !"#$! ! !% & ' ! . ') * + / '& '
=< F 5 = 7 / ' ' 0-

<@ !"#$! ! !% & ' ! . ' ' ' . /
/ / - # 0 ' . / > 0 ' - 0' > 0 0-

<< !"#$! ! !% & ' ! . ' ' ' . /
/ / - # 0 ' . / > 0 ' - 0' > 0 0-

<E !"#$! ! !% & ' ! . /) * +) &
)' & $ ' # 5 = 7 / / 0-

< !"#$! ! !% & ' ! B /) ' / 6 . . 7 /
/ / 0-

E !"#$! ! !% & ' ! " - 5 . ' 7 '
. B * " - F 4 9 ' 8 0-

E !"#$! ! !% & ' ! 7 ' 5 . ' : . *
" - 8 E < 0-

E !"#$! ! !% & ' ! = . 0 & ' 6 ' 0-

E !"#$! ! !% & ' ! @ D
%0 ' (/ ('* &(6 (" (B *(B /((((' -("00 -(%0 0-

E !"#$! ! !% & ' ! @
. 9 &/(' -(' ' 3' ((' (' & 0-

E

<

!"#$! ! !% & ' ! < ' ' / ' - '
5 + D- 0-

!"#$! ! !% & ' ! < + ' ' & /5 0 ,
- ' - 3 ' & / ' - 0 ' ' ' & 0-

!"#$! ! !% & ' ! < + -' ' ' '
- ' ' 0-

!"#$! ! !% & ' ! < + -' ' ' '
. & / 8 ADD ' ' - "-4' - 0-

= !"#$! ! !% & ' ! < > / . & / 0-
V : 4 ' ;;;

@ !"#$! ! !% & ' ! < 6 & ' ' '- D' / 0-

< !"#$! ! !% & ' ! < 6 - - ' '
/ ' - & & '00 ' 9 & 0-

E !"#$! ! !% & ' ! < . & / # 0-

!"#$! ! !% & ' ! < . & /5 8E D- 0-

!"#$! ! !% & ' ! < . & / 0-

!"#$! ! !% & ' ! < . & / 1 ' 0-

!"#$! ! !% & ' ! E > / . & /5
- D- B - '- 9 & ' 0 D 4 8 @ 0-

@ E =

!"#$! ! !% & ' ! " / '
& - ,0 ' 8 < -24

!"#$! ! !% & ' ! " / '
& - ,0 ' 8 < 0-

!"#$! ! !% & ' ! . .) -/ + .
G / 0 & 8 E@ -24

!"#$! ! !% & ' ! . .) -/ + .
G / 0 & 8 E@ 0-

!"#$! ! !% & ' ! + -' ' ' A B /
8 E 0-

!"#$! ! !% & ' ! @ E= (9 ' 0-
V ' ' U 9 ' - ' ' '

= !"#$! ! !% & ' ! ' (- (' '- (
(. (((6 '& ' 0-

@ !"#$! ! !% & ' ! = (? 4 0-

< !"#$! ! !% & ' ! < @W 0-
V > - ' ' ' - &

< <

< E

!"#$! ! !% & ' ! = 66 " 0-
I& @ 8 =

FD 7%#F6 6G 6# ?:% FD 7D 6 7"D#66 D. 7"D#%D6F?
? D#+6 6D)?
O+ X 2' @< P

F . - ' ' 4 D * - 9 ' >'

0 ' 4 & - ' - - 0'

!"#$! ! !% & ' ! @ - - 0' 0-
V F6D .7D%#6D FD# "#"::D::6G6DG %G D?1:6F)6 "?+ >"? 5

!"#$! ! !% & ' ! @ &' - ,9 / 0-

= !"#$! ! !% & ' ! @ & ' '9 ' 0-
V >D% D.#6 7D ")7"GF:?G>DG
%9 : 4 - ;;;
I& <<8 6W - - ' >
V 7 3

4 7 3 ' - & 0 4' '9 ,5 /5 35 0
L - , 0 4' ' 66 00 5 - 0 5 ' 9 - 4 L ' -
- , 0 , D - ' - L U 0 ' ' - - , 0 ,
0 0 - ' 0 ' ' ' & - - 4 8? *

I& @<8 E 6 ? & 9 & -' 1 4

@ !"#$! ! !% & ' ! @ - ' & 9 0-

< !"#$! ! !% & ' ! @ ' ' 0-

0 ' 4 & - ' ' '
%9 @= 0'&

E !"#$! ! !% & ' ! @ &' & -2 ' 0-

0 ' 4 & - ' ' '
%9 0'&

!"#$! ! !% & ' ! @ ' ' 0-

0 ' 4 & - ' ' '
%9 E< 0'&

!"#$! ! !% & ' ! @ B' * -
> 0-

!"#$! ! !% & ' ! @ &' & -2 ' 0-
V B%:+>"G> ?GF A%7"GG)%:H"6
>D% D.#6 7D ?G.D# ? 7?G>DG
%9 0'&

!"#$! ! !% & ' ! @ &' & -2 ' 0-
%9 0'&
B%:+>"G>)%:H"6
.7D%#6D FD# "#"::D:DG 8 E
GD) . 7.#"> 8 E E
.R 1D "? FD .DG." DG 8 E
1R#SD# >#?GF#6 D6GD D# ? 7 8 E=

@ =

" !F"." G 4 4 !) * !

!"#$! ! !% & ' ! < > 6 - " & 9 '
4 " ' ' ' ' /8#6 8 E 0-

!"#$! ! !% & ' ! < : * - " 9 * 4
? 4 / 8 @ 0-

!"#$! ! !% & ' ! E "00 - F ' > / "
- 6 - 0-

!"#$! ! !% & ' ! E 0 E 0-
V " > /

= !"#$! ! !% & ' ! E >' & . ' / "
' ' 6 - 5 + # ' 4 $ ' ' $DF 0-

@ !"#$! ! !% & ' ! E >' & . ' / "
' ' 6 - 5 G "9 ' >' & . $ F ' - . D '*
. / 0-

< !"#$! ! !% & ' ! E > " & 9 ' 0
8# 4 - D- " %92 % - "00 ' > / 0-

E !"#$! ! !% & ' ! E > ' - ' ' 0-

!"#$! ! !% & ' ! E ' ' ' >' & . / B
"00 ' ' -' - - ' / 0-

!"#$! ! !% & ' ! E G # ' 4 $ ' . /
F/ ' 5 / /5 ' - > / E < < 0-

!"#$! ! !% & ' ! E ' ' - + -' '
6 ' " 6 - ' / 0-

!"#$! ! !% & ' ! E $ ' + - . / " -
6 - 0-

!"#$! ! !% & ' ! E $ ' + - . / 6
+ -' ' - "9 ' ' - G "9 ' >' & . 0-

!"#$! ! !% & ' ! E $ ' + - . / 66
6 - $ ' > '4 /5 0 / / ' - & . / 0-

= !"#$! ! !% & ' ! E & . / ' - . / 0-

@ !"#$! ! !% & ' ! E & . / 5
6 -) & 0-

< !"#$! ! !% & ' ! E & . / 665

0 &. /'-)/ -0-

E !"#$! ! !% & '! E / + -' '
/ 0-

!"#$! ! !% & '! E / /'-' 2 4
> / < @ 0-

!"#$! ! !% & '! E / /'- '-'- -
' ' '-' / 0-

!"#$! ! !% & '! E . * & > ' /" 4 /
> 0-
%9 : 4 - ;;;

!"#$! ! !% & '! 7 & ' / /'-
. 0 ' > / 0-

!"#$! ! !% & '! 7 & ' / /'-
. 0 ' > / 0-

!"#$! ! !% & '! / /5 '0 '- 0' "
6 - ' ' & > / 1 / & 8 @ 0-
% 9 : 4 - ;;;

= !"#$! ! !% & '! . 0 &/ ' 0 &
'& 8 0-

@ !"#$! ! !% & '! . DL ' G'4 * '-
"9 ' ''9 DL ' &Y. 9 '& 8 E= 0-

< !"#$! ! !% & '! E@ 0-
V . . 9
" 8 # 9 H'

E !"#$! ! !% & '! DF =E =E 0-
V . . 9
" 8 # 9 H'
%9 : 4 - ;;; < 0'&

!"#$! ! !% & '! 0-
V % Z 0 9 ' I -' ' I '& &'

) !F"." G 4 4 !) * !

!"#$! ! !% & '! "/0 & ''/ 5
"/0 & ''/ 5 0' 6 0-

!"#$! ! !% & '!
"/0 (> ("'/ (' (66 0-

!"#$! ! !% & ' ! ' ' : " & 9 ' ' 6 / 0-

!"#$! ! !% & ' !
0 ' /("9 ' (" & 9 '(. , 9 * ((' ' 0-
%9 : 4 - ;;;
I&

>
7 ' F > [
9 G 9 [" & . ' -
L '
D,

@ / / > 0 E
6 E [' \ ' + ' / /
> 0 E= [' \ ' + > 0 # '
E<
D, EE
< / / ' - &
4' [ ) -] . ["00 ' [
> 0 " E
D,
) & '0 / B ') - =

6 0 '
I& 8 " & . ' - L '

F & 5 - - 0 0 '4 ' -
'4 ' 4 - 0 9 6 <<=5
' " '- /5 4 - ' 5 0' - '
& ,' ' ' '
0 0 - , 9 0 0 ' " & ' 9
' / ' ' ' 9
0 9 5 9 ' '-/ 0' '- 0 0
' '4 - 0 0
'4 ' - ' 0 9 9 '4 - 9 4
% 0 - ' - ' / ' & ' ' &
-& ' " - < ' 5
* % & > - 6 < 5 ' 0 ' F
' ' ' - ' '- ' 4' - -
' / 0 0 '4 ' - ' * & ' & 5 -
9 & 9 5 ' - L ' & . - -
@ - - '4 & ' 0 -
& ' # - % ' ' /5 0'0 ' - '&'3 '4
'9 ^- & 0 9 5_ & 4 & 0
' '/ 9 4' - ' / ' & ' -
L ' '4 '- * 0 9 - ' ,0
' - - 9 - & ' - 4 % '- 0 -
' & '& ;
" & 4 ' - '4 L ' - 5
0' '9 # ' ' ' - & : ' - -'
'- @ ' 0 4 ' / / ' ' L ' &

or 2, we see by Theorem 20.5 that if c is constructible, then $[Q(c) : Q] = 2^k$ for some nonnegative integer k.

We now dispatch the problems that plagued the Greeks. Consider doubling the cube of volume 1. The enlarged cube would have an edge of length $\sqrt[3]{2}$. But $[Q(\sqrt[3]{2}) : Q] = 3$, so such a cube cannot be constructed.

Next consider the possibility of trisecting a 60° angle. If it were possible to trisect an angle of 60°, then cos 20° would be constructible. (See Figure 22.1.) In particular, $[Q(\cos 20^\circ) : Q] = 2^k$ for some k. Now, using the trigonometric identity $\cos 3\theta = 4\cos^3\theta - 3\cos\theta$, with $\theta = 20^\circ$, we see that $1/2 = 4\cos^3 20^\circ - 3\cos 20^\circ$, so that cos 20° is a zero of $8x^3 - 6x - 1$. But, since $8x^3 - 6x - 1$ is irreducible over Q (see Exercise 13), we must also have $[Q(\cos 20^\circ) : Q] = 3$. This contradiction shows that trisecting a 60° angle is impossible.

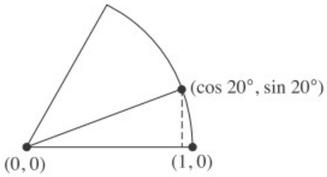

Figure 22.1 Sixty-degree angle.

The remaining problems are relegated to the reader as Exercises 14, 15, and 17.

Angle-Trisectors and Circle-Squarers

Down through the centuries, hundreds of people have claimed to have achieved one or more of the impossible constructions. In 1775, the Paris Academy, so overwhelmed with these claims, passed a resolution to no longer examine these claims or claims of machines purported to exhibit perpetual motion. Although it has been more than 100 years since the last of the constructions was shown to be impossible, there continues to be a steady parade of people who claim to have done one or more of them. Most of these people have heard that this is impossible but have refused to believe it. One person insisted that he could trisect any angle with a straightedge alone. Another found his trisection in 1973 after 12,000 hours of work. One got his from God. In 1971, a person with a Ph.D. in mathematics asserted that he had a valid trisection method. Many people have claimed the hat trick: trisecting the angle, doubling the cube, and squaring the circle. Two men who did this in 1961 succeeded in having their accomplishment noted in the *Congressional Record.* Occasionally, newspapers and magazines have run stories about "doing the impossible," often giving the impression that the construction may be valid. Many angle-trisectors and circle-squarers have had their work published at their own expense and distributed to colleges and universities. One had his printed in four languages!

Although he never claimed to have squared the circle, the incomparable Renaissance multifaceted genius Leonardo da Vinci made 169 attempts over many years at squaring the circle.

Exercises

Only prove to me that it is impossible, and I will set about it this very evening.

Spoken by a member of the audience after De Morgan gave a lecture on the impossibility of squaring the circle.

1. If a and b are constructible numbers and $a \geq b > 0$, give a geometric proof that $a + b$ and $a - b$ are constructible.

2. If a and b are constructible, give a geometric proof that ab is constructible. (*Hint:* Consider the following figure. Notice that all segments in the figure can be made with an unmarked straightedge and a compass.)

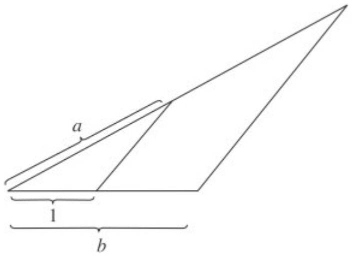

3. Prove that if c is a constructible number, then so is $\sqrt{|c|}$. (*Hint:* Consider the following semicircle with diameter $1 + |c|$.) (This exercise is referred to in Chapter 31.)

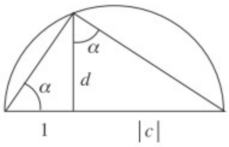

4. If a and b ($b \neq 0$) are constructible numbers, give a geometric proof that a/b is constructible. (*Hint:* Consider the following figure.)

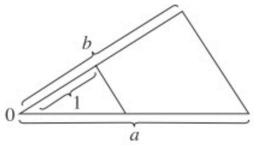

5. Prove that $\sin\theta$ is constructible if and only if $\cos\theta$ is constructible.

6. Prove that an angle θ measured in degrees is constructible if and only if $\sin\ \theta$ is constructible.

7. Prove that $\cos 2\theta$ is constructible if and only if $\cos\theta$ is constructible.

8. Prove that 30° is a constructible angle.

9. Prove that a 45° angle can be trisected with an unmarked straightedge and a compass.

10. Prove that a 40° angle is not constructible.

11. Show that the point of intersection of two lines in the plane of a field F lies in the plane of F.

12. Show that the points of intersection of a circle in the plane of a field F and a line in the plane of F are points in the plane of F or in the plane of $F(\sqrt{\alpha})$, where $\alpha \in F$ and α is positive. Give an example of a circle and a line in the plane of Q whose points of intersection are not in the plane of Q.

13. Prove that $8x^3 - 6x - 1$ is irreducible over Q.

14. Use the fact that $8\cos^3(2\pi/7) + 4\cos^2(2\pi/7) - 4\cos(2\pi/7) - 1 = 0$ to prove that a regular seven-sided polygon is not constructible with an unmarked straightedge and a compass.

15. Show that a regular 9-gon cannot be constructed with an unmarked straightedge and a compass.

16. Show that if a regular n-gon is constructible, then so is a regular $2n$-gon.

17. (Squaring the Circle) Show that it is impossible to construct, with an unmarked straightedge and a compass, a square whose area equals that of a circle of radius 1. You may use the fact that π is transcendental over Q.

18. Use the fact that $4\cos^2(2\pi/5) + 2\cos(2\pi/5) - 1 = 0$ to prove that a regular pentagon is constructible.

19. Can the cube be "tripled"?

20. Can the cube be "quadrupled"?

21. Can the circle be "cubed"?

22. If a, b, and c are constructible, show that the real roots of $ax^2 + bx + c$ are constructible.

I&   E 8=

26 Symmetry Groups

Physicists have exalted symmetry to the position of *the* central concept in their attempts to organize and explain an otherwise bewildering and complex universe.

Mario Livio, *The Equation That Could Not Be Solved*

I'm not good at math, but I do know that the universe is formed with mathematical principles whether I understand them or not, and I am going to let that guide me.

Bob Dylan, *Chronicles, Volume One*

Isometries

In the early chapters of this book, we briefly discussed symmetry groups. In this chapter and the next, we examine this fundamentally important concept in some detail. It is convenient to begin such a discussion with the definition of an isometry (from the Greek *isometros*, meaning "equal measure") in $\mathbf{R}^n$.

Definition Isometry
An *isometry* of n-dimensional space $\mathbf{R}^n$ is a function from $\mathbf{R}^n$ onto $\mathbf{R}^n$ that preserves distance.

In other words, a function T from $\mathbf{R}^n$ onto $\mathbf{R}^n$ is an isometry if, for every pair of points p and q in $\mathbf{R}^n$, the distance from $T(p)$ to $T(q)$ is the same as the distance from p to q. With this definition, we may now make precise the definition of the symmetry group of an n-dimensional figure.

Definition Symmetry Group of a Figure in $\mathbf{R}^n$
Let F be a set of points in $\mathbf{R}^n$. The *symmetry group of F* in $\mathbf{R}^n$ is the set of all isometries of $\mathbf{R}^n$ that carry F onto itself. The group operation is function composition.

It is important to realize that the symmetry group of an object depends not only on the object, but also on the space in which we view it. For example, the symmetry group of a line segment in $\mathbf{R}^1$ has order 2, the symmetry group of a line segment considered as a set of points in $\mathbf{R}^2$ has order 4, and the symmetry group of a line segment viewed as a set of points in $\mathbf{R}^3$ has infinite order (see Exercise 9).

Although we have formulated our definitions for all finite dimensions, our chief interest will be the two-dimensional case. It has been known since 1831 that every isometry of $\mathbf{R}^2$ is one of four types: rotation, reflection, translation, and glide-reflection. Rotation about a point in a plane needs no explanation. A *reflection across a line* L is that transformation that leaves every point of L fixed and takes every point Q, not on L, to the point Q' so that L is the perpendicular bisector of the line segment from Q to Q' (see Figure 26.1). The line L is called the *axis of reflection.* In an xy-coordinate plane, the transformation $(x, y) \to (x, -y)$ is a reflection across the x-axis, whereas $(x, y) \to (y, x)$ is a reflection across the line $y = x$. Some authors call an axis of reflective symmetry L a *mirror* because L acts like a two-sided mirror; that is, the image of a point Q in a mirror placed on the line L is, in fact, the image of Q under the reflection across the line L. Reflections are called *opposite* isometries because they reverse orientation. For example, the reflected image of a clockwise spiral is a counterclockwise spiral. Similarly, the reflected image of a right hand is a left hand. (See Figure 26.1.)

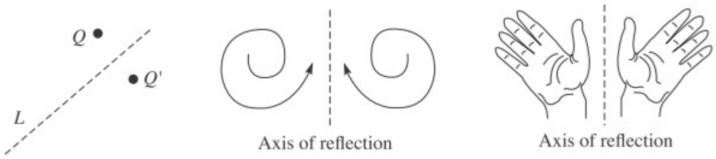

Figure 26.1 Reflected images.

A *translation* is simply a function that carries all points the same distance in the same direction. For example, if p and q are points in a plane and T is a translation, then the two directed line segments joining p to $T(p)$ and q to $T(q)$ have the same length and direction. A *glide-reflection* is the product of a translation and a reflection across the line containing the translation line segment. This line is called the *glide-axis.* In Figure 26.2, the arrow gives the direction and length of the translation, and is contained in the axis of reflection. A glide-reflection is also an opposite isometry. Successive footprints in wet sand are related by a glide-reflection.

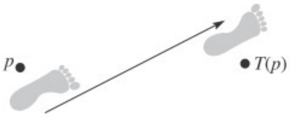

Figure 26.2 Glide-reflection

Classification of Finite Plane Symmetry Groups

Our first goal in this chapter is to classify all finite plane symmetry groups. As we have seen in earlier chapters, the dihedral group D_n is the plane symmetry group of a regular n-gon. (For convenience, call D_2 the plane symmetry group of a nonsquare rectangle and D_1 the plane symmetry group of the letter "V." In particular, $D_2 \approx Z_2 \oplus Z_2$ and $D_1 \approx Z_2$.) The cyclic groups Z_n are easily seen to be plane symmetry groups also. Figure 26.3 is an illustration of an organism whose plane symmetry group consists of four rotations and is isomorphic to Z_4. The surprising fact is that the cyclic groups and dihedral groups are the only finite plane symmetry groups. The famous mathematician Hermann Weyl attributes the following theorem to Leonardo da Vinci (1452–1519).

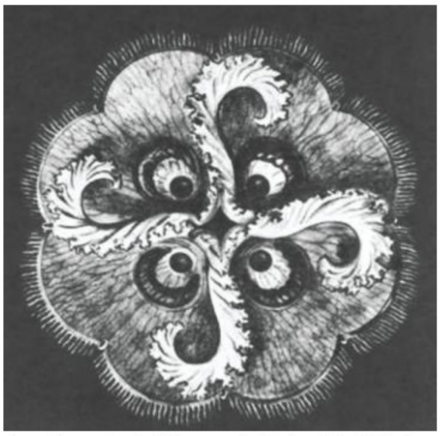

From Symmetry in Science and Art by A.V. Shubnikov
and V.A. Kopstik, Copyright 1974 Plenum Publishing Co.

Figure 26.3 Aurelia insulinda, an organism whose plane symmetry group is Z_4.

Theorem 26.1 Finite Symmetry Groups in the Plane

The only finite plane symmetry groups are Z_n and D_n.

PROOF Let G be a finite plane symmetry group of some figure. We first observe that G cannot contain a translation or a glide-reflection, because in either case G would be infinite. Now observing that the composition of two reflections preserves orientation, we know that such a composition is a translation or rotation. When the two reflections have parallel axes of reflection, there is no fixed point so the composition is a translation. Thus, every two reflections in G have reflection axes that intersect in some point. We claim that all reflections intersect in the same point. Suppose that f and f' are two distinct reflections in G. Then because ff' preserves orientation, we know that ff' is a rotation. We use the fact from geometry that a finite group of rotations must have a common center, say P. This means that any two reflections must intersect at point P. So, we have shown that all the elements of G have the common fixed point P.

For convenience, let us denote a rotation about P of σ degrees by R_σ. Now, among all rotations in G, let β be the smallest positive angle of rotation. (Such an angle exists, since G is finite and R_{360} belongs to G.) We claim that every rotation in G is some power of R_β. To see this, suppose that R_σ is in G. We may assume $0° < \sigma \leq 360°$. Then, $\beta \leq \sigma$ and there is some integer t such that $t\beta \leq \sigma < (t+1)\beta$. But, then $R_{\sigma - t\beta} = R_\sigma(R_\beta)^{-t}$ is in G and $0 \leq \sigma - t\beta < \beta$. Since β represents the smallest positive angle of rotation among the elements of G, we must have $\sigma - t\beta = 0$, and therefore, $R_\sigma = (R_\beta)^t$. This verifies the claim.

For convenience, let us say that $|R_\beta| = n$. Now, if G has no reflections, we have proved that $G = \langle R_\beta \rangle \approx Z_n$. If G has at least one reflection, say f, then

$$f, fR_\beta, f(R_\beta)^2, \ldots, f(R_\beta)^{n-1}$$

are also reflections. Furthermore, this is the entire set of reflections of G. For if g is any reflection in G, then fg is a rotation, and so $fg = (R_\beta)^k$ for some k. Thus, $g = f^{-1}(R_\beta)^k = f(R_\beta)^k$. So

$$G = \{R_0, R_\beta, (R_\beta)^2, \ldots, (R_\beta)^{n-1}, f, fR_\beta, f(R_\beta)^2, \ldots, f(R_\beta)^{n-1}\},$$

and G is generated by the pair of reflections f and fR_β. Hence, by our characterization of the dihedral groups (Theorem 25.5), G is the dihedral group D_n. ▮

Classification of Finite Groups of Rotations in $\mathbf{R}^3$

One might think that the set of all possible finite symmetry groups in three dimensions would be much more diverse than is the case for two dimensions. Surprisingly, this is not the case. For example, moving to three dimensions introduces only three new groups of rotations. This observation was first made by the physicist and mineralogist Auguste Bravais in 1849, in his study of possible structures of crystals.

▮ Theorem 26.2 Finite Groups of Rotations in $\mathbf{R}^3$

Up to isomorphism, the finite groups of rotations in R^3 are Z_n, D_n, A_4, S_4, and A_5.

Theorem 26.2, together with the Orbit-Stabilizer Theorem (Theorem 7.4), makes easy work of determining the group of rotations of an object in $\mathbf{R}^3$.

▮ EXAMPLE 8 We determine the group G of rotations of the cuboctahedron in Figure 26.4, which is composed of six congruent squares and eight congruent equilateral triangles. We begin by singling out any one of the squares. Obviously, there are four rotations that map this square to itself, and the designated square can be rotated to the location of any of the other five. So, by the Orbit-Stabilizer Theorem (Theorem 7.4), the rotation group has order $4 \cdot 6 = 24$. By Theorem 26.2, G is one of Z_{24}, D_{12}, and S_4. But each of the first two groups has exactly two elements of order 4, whereas G has more than two. So, G is isomorphic to S_4. ▮

The group of rotations of a tetrahedron (the *tetrahedral group*) is isomorphic to A_4; the group of rotations of a cube or an octahedron (the *octahedral group*) is isomorphic to S_4; the group of rotations of a dodecahedron or an icosahedron (the *icosahedral group*) is isomorphic to A_5. These five solids are illustrated in Figure 26.5.

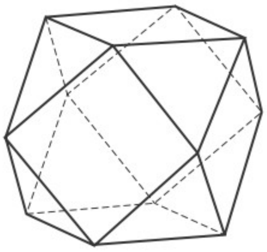

Figure 26.4 Cuboctahedron.

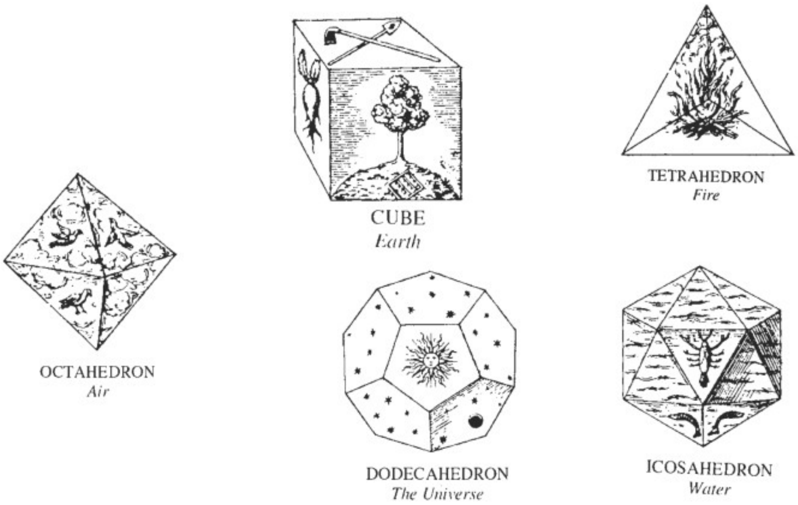

CUBE
Earth

TETRAHEDRON
Fire

OCTAHEDRON
Air

DODECAHEDRON
The Universe

ICOSAHEDRON
Water

Figure 26.5 The five regular solids as depicted by Johannes Kepler in *Harmonices Mundi, Book II* (1619)

Exercises

> Perhaps the most valuable result of all education is the ability to make yourself do the thing you have to do, when it ought to be done, whether you like it or not .
>
> Thomas Henry Huxley, "Technical Education"

1. Show that an isometry of $\mathbf{R}^n$ is one-to-one.

2. Show that the translations of $\mathbf{R}^n$ form a group.

3. Exhibit a plane figure whose plane symmetry group is Z_5.

4. Show that the group of rotations in $\mathbf{R}^3$ of a 3-prism (i.e., a prism with equilateral ends, as in the following figure) is isomorphic to D_3.

5. What is the order of the (entire) symmetry group in $\mathbf{R}^3$ of a 3-prism?

6. What is the order of the symmetry group in $\mathbf{R}^3$ of a 4-prism (a box with square ends that is not a cube)?

7. What is the order of the symmetry group in $\mathbf{R}^3$ of an n-prism?

8. Show that the symmetry group in $\mathbf{R}^3$ of a box of dimensions $2'' \times 3'' \times 4''$ is isomorphic to $Z_2 \oplus Z_2 \oplus Z_2$.

9. Describe the symmetry group of a line segment viewed as
a. a subset of $\mathbf{R}^1$.
b. a subset of $\mathbf{R}^2$.
c. a subset of $\mathbf{R}^3$.
(This exercise is referred to in this chapter.)

10. (From the "Ask Marilyn" column in *Parade Magazine*, December 11, 1994.)[1] The letters of the alphabet can be sorted into the following categories:
a. FGJLNPQRSZ
b. BCDEK
c. AMTUVWY
d. HIOX
What defines the categories?

11. Exactly how many elements of order 4 does the group in Example 1 have?

12. Why is inversion [that is, $\phi(x, y) = (-x, -y)$,] not listed as one of the four kinds of isometries in R^2?

13. Explain why inversion through a point in $\mathbf{R}^3$ cannot be realized by a rotation in $\mathbf{R}^3$.

14. Reflection across a line L in $\mathbf{R}^3$ is the isometry that takes each point Q to the point Q' with the property that L is a perpendicular bisector of the line segment joining Q and Q'. Describe a rotation that has this same effect.

---

[1] Copyright © 1994. Reprinted with permission of the author and the publisher from PARADE, December 11, 1994.

15. In $\mathbf{R}^2$, a rotation fixes a point; in $\mathbf{R}^3$, a rotation fixes a line. In $\mathbf{R}^4$, what does a rotation fix? Generalize these observations to $\mathbf{R}^n$.

16. Show that an isometry of a plane preserves angles.

17. Show that an isometry of a plane is completely determined by the image of three noncollinear points.

18. Suppose that an isometry of a plane leaves three noncollinear points fixed. Which isometry is it?

19. Suppose that an isometry of a plane fixes exactly one point. What type of isometry must it be?

20. Suppose that A and B are rotations of 180° about the points a and b, respectively. What is A followed by B? How is the composite motion related to the points a and b?

27 Symmetry and Counting

Let us pause to slake our thirst one last time at symmetry's bubbling spring.

Timothy Ferris, *Coming of Age in the Milky Way*

Whenever you can, count.

Francis Galton, (1822–1911)

Motivation

Permutation groups naturally arise in many situations involving symmetric designs or arrangements. Consider, for example, the task of coloring the six vertices of a regular hexagon so that three are black and three are white. Figure 27.1 shows the $\binom{6}{3} = 20$ possibilities. However, if these designs appeared on one side of hexagonal ceramic tiles, it would be nonsensical to count the designs shown in Figure 27.1(a) as different, since all six designs shown there can be obtained from one of them by rotating. (A manufacturer would make only one of the six.) In this case, we say that the designs in Figure 27.1(a) are *equivalent* under the group of rotations of the hexagon. Similarly, the designs in Figure 27.1(b) are equivalent under the group of rotations, as are the designs in Figure 27.1(c) and those in Figure 27.1(d). And, since no design from Figure 27.1(a)–(d) can be obtained from a design in a different part by rotation, we see that the designs within each part of the figure are equivalent to each other but nonequivalent to any design in another part of the figure. However, the designs in Figure 27.1(b) and (c) are equivalent under the dihedral group D_6, since the designs in Figure 27.1(b) can be reflected to yield the designs in Figure 27.1(c). For example, for purposes of arranging three black beads and three white beads to form a necklace, the designs shown in Figure 27.1(b) and (c) would be considered equivalent.

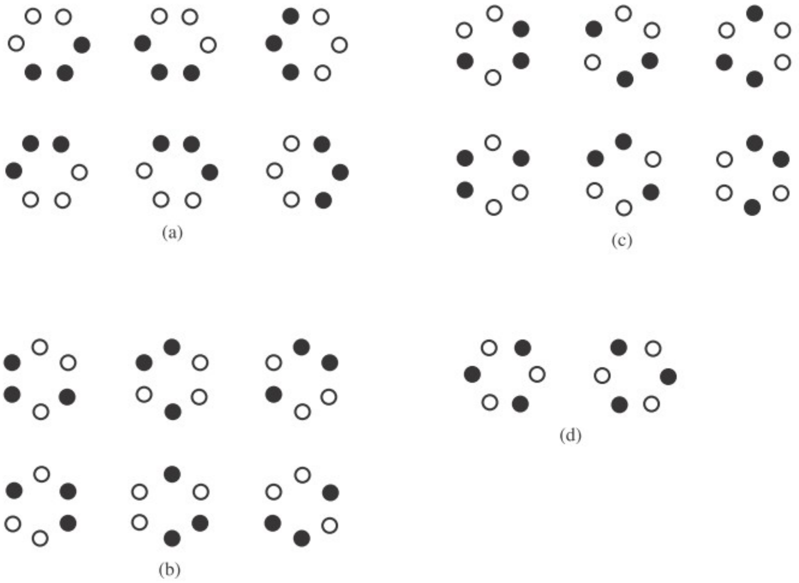

(a) (b) (c) (d)

Figure 27.1 Necklaces with three black beads and three white beads.

In general, we say that two designs (arrangements of beads) A and B are *equivalent under a group* G of permutations of the arrangements if there is an element ϕ in G such that $\phi(A) = B$. That is, two designs are equivalent under G if they are in the same orbit of G. It follows, then, that the number of nonequivalent designs under G is simply the number of orbits of designs under G. (The set being permuted is the set of all possible designs or arrangements.)

Notice that the designs in Figure 27.1 divide into four orbits under the group of rotations but only three orbits under the group D_6, since the designs in Figure 27.1(b) and (c) form a single orbit under D_6. Thus, we could obtain all 20 tile designs from just four tiles, but we could obtain all 20 necklaces from just three of them.

Burnside's Theorem

Although the problems we have just posed are simple enough to solve by observation, more complicated ones require a more sophisticated approach. Such an approach was provided by Georg

Frobenius in 1887. Frobenius's theorem did not become widely known until it appeared in the classic book on group theory by William Burnside in 1911. By an accident of history, Frobenius's theorem has come to be known as Burnside's Theorem. Before stating this theorem, we recall some notation introduced in Chapter 7 and introduce new notation. If G is a group of permutations on a set S and $i \in S$, then $\text{stab}_G(i) = \{\phi \in G \mid \phi(i) = i\}$ and $\text{orb}_G(i) = \{\phi(i) \mid \phi \in G\}$. For any set X, we use $|X|$ to denote the number of elements in X.

Definitions Elements Fixed by ϕ

For any group G of permutations on a set S and any ϕ in G, we let $\text{fix}(\phi) = \{i \in S \mid \phi(i) = i\}$. This set is called the *elements fixed by ϕ* (or more simply, "fix of ϕ").

Theorem 27.1 Burnside's Theorem

If G is a finite group of permutations on a set S, then the number of orbits of elements of S under G is

$$\frac{1}{|G|}\sum_{\phi \in G} |\text{fix}(\phi)|.$$

PROOF Let n denote the number of pairs (ϕ, i), with $\phi \in G, i \in S$, and $\phi(i) = i$. We begin by counting these pairs in two ways. First, for each particular ϕ in G, the number of such pairs is exactly $|\text{fix}(\phi)|$. So,

$$n = \sum_{\phi \in G} |\text{fix}(\phi)|.$$

Second, for each particular i in S, observe that $|\text{stab}_G(i)|$ is exactly the number of pairs (ϕ, i) for which $\phi(i) = i$. So,

$$n = \sum_{i \in S} |\text{stab}_G(i)|.$$

It follows from Exercise 67 in Chapter 7 that if s and t are in the same orbit of G, then $\text{orb}_G(s) = \text{orb}_G(t)$, and thus by the Orbit-Stabilizer Theorem (Theorem 7.4) we have $|\text{stab}_G(s)| = |G|/|\text{orb}_G(s)| = |G|/|\text{orb}_G(t)| = |\text{stab}_G(t)$. So, if we choose $s \in S$ and sum over $\text{orb}_G(s)$, we have

Table 27.1

Element	Number of Designs Fixed by Element
Identity	20
Rotation of 60°	0
Rotation of 120°	2
Rotation of 180°	0
Rotation of 240°	2
Rotation of 300°	0

$$\sum_{t\in \mathrm{orb}_G(s)} |\mathrm{stab}_G(t)| = |\mathrm{orb}_G(S)||\mathrm{stab}_G(S)| = |G|.$$

Finally, by summing over all the elements of G, one orbit at a time, it follows from Equations (1), (2), and (3) that

$$\sum_{\phi\in G} |\mathrm{fix}(\phi)| = \sum_{i\in S} |\mathrm{stab}(i)| = |G| \cdot (\text{number of orbits})$$

and the result follows. ▮

Applications

To illustrate how to apply Burnside's Theorem, let us return to the ceramic tile and necklace problems. In the case of counting hexagonal tiles with three black vertices and three white vertices, the objects being permuted are the 20 possible designs, whereas the group of permutations is the group of six rotational symmetries of a hexagon. Obviously, the identity fixes all 20 designs. We see from Figure 27.1 that rotations of 60°, 180°, or 300° fix none of the 20 designs. Finally, Figure 27.2 shows fix(ϕ) for the rotations of 120° and 240°. These data are collected in Table 27.1.

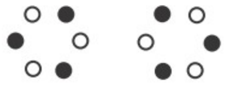

Figure 27.2 Tile designs fixed by 120° rotation and 240° rotation.

So, applying Burnside's Theorem, we obtain the number of orbits under the group of rotations as

$$\frac{1}{6}(20 + 0 + 2 + 0 + 2 + 0) = 4.$$

Now let's use Burnside's Theorem to count the number of necklace arrangements consisting of three black beads and three white beads. (For the purposes of analysis, we may arrange the beads in the shape of a regular hexagon.) For this problem, two arrangements are equivalent if they are in the same orbit under D_6. Figure 27.3 shows the arrangements fixed by a reflection across a diagonal. Table 27.2 summarizes the information needed to apply Burnside's Theorem.

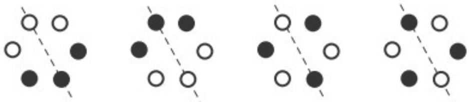

Figure 27.3 Bead arrangements fixed by the reflection across a diagonal.

Table 27.2

Type of Element	Number of Elements of This Type	Number of Arrangements Fixed by Type of Element
Identity	1	20
Rotation of order 2 (180°)	1	0
Rotation of order 3 (120° or 240°)	2	2
Rotation of order 6 (60° or 300°)	2	0
Reflection across diagonal	3	4
Reflection across side bisector	3	0

So, there are

$$\frac{1}{12}(1 \cdot 20 + 1 \cdot 0 + 2 \cdot 2 + 2 \cdot 0 + 3 \cdot 4 + 3 \cdot 0) = 3$$

nonequivalent ways to string three black beads and three white beads on a necklace.

Now that we have gotten our feet wet on a few easy problems, let's try a more difficult one. Suppose that we have the colors red (R), white (W), and blue (B) that can be used to color the edges of a regular tetrahedron (see Figure 5.1). First, observe that there are $3^6 = 729$ colorings without regard to equivalence. How shall we decide when two colorings of the tetrahedron are nonequivalent? Certainly, if we were to pick up a tetrahedron colored in a certain manner, rotate it, and put it back down, we would think of the tetrahedron as being positioned differently rather than as being colored differently (just as if we picked up a die labeled in the usual way and rolled it, we would not say that the die is now differently labeled). So, our permutation group for this problem is just the group of 12 rotations of the tetrahedron shown in Figure 5.1 and is isomorphic to A_4. (The group consists of the identity; eight elements of order 3, each of which fixes one vertex; and three elements of order 2, each of which fixes no vertex.) Every rotation permutes the 729 colorings, and to apply Burnside's Theorem we must determine the size of $\text{fix}(\phi)$ for each of the 12 rotations of the group.

Clearly, the identity fixes all 729 colorings. Next, consider the element (234) of order 3, shown in the bottom row, second from the left in Figure 5.1. Suppose that a specific coloring is fixed by this element (i.e., the tetrahedron appears to be colored the same before and after this rotation). Since (234) carries edge 12 to edge 13, edge 13 to edge 14, and edge 14 to edge 12, these three edges must agree in color (edge ij is the edge joining vertex i and vertex j). The same argument shows that the three edges 23, 34, and 42 also must agree in color. So, $|\text{fix}(234)| = 3^2$, since there are three choices for each of these two sets of three edges. The nine columns in Table 27.3 show the possible colorings of the two sets of three edges. The analogous analysis applies to the other seven elements of order 3.

Now consider the rotation (12)(34) of order 2. (See the second tetrahedron in the top row in Figure 5.1.) Since edges 12 and 34 are fixed, they may be colored in any way and will appear the same after the rotation (12)(34). This yields $3 \cdot 3$ choices for those two edges. Since edge 13 is carried to edge 24, these two edges must agree in color. Similarly, edges 23 and 14 must agree. So, we have three choices for the pair of edges 13 and 24 and three choices for the pair of edges 23 and 14. This means that we have

Table 27.3 Nine Colorings Fixed by (234).

Edge	Colorings								
12	R	R	R	W	W	W	B	B	B
13	R	R	R	W	W	W	B	B	B
14	R	R	R	W	W	W	B	B	B
23	R	W	B	W	R	B	B	R	W
34	R	W	B	W	R	B	B	R	W
24	R	W	B	W	R	B	B	R	W

3·3·3·3 ways to color the tetrahedron that will be equivalent under (12)(34). (Table 27.4 gives the complete list of 81 colorings.) So, $|\text{fix}((12)(34))| = 3^4$, and the other two elements of order 2 yield the same results.

Now that we have analyzed the three types of group elements, we can apply Burnside's Theorem. In particular, the number of distinct colorings of the edges of a tetrahedron with three colors is

$$\frac{1}{12}(1 \cdot 3^6 + 8 \cdot 3^2 + 3 \cdot 3^4) = 87.$$

Surely it would be a difficult task to solve this problem without Burnside's Theorem.

Just as surely, you are wondering who besides mathematicians are interested in counting problems such as the ones we have discussed. Well, chemists are. Indeed, one set of benzene derivatives can be viewed as six carbon atoms arranged in a hexagon with one of the three radicals NH_2, COOH, or OH attached at each carbon atom. See Figure 27.4 for one example.

So Burnside's Theorem enables a chemist to determine the number of benzene molecules (see Exercise 4). Another kind of molecule considered by chemists is visualized as a regular tetrahedron with a carbon atom at the center and any of the four radicals $HOCH_2$ (hydroxymethyl), C_2H_5 (ethyl), Cl (chlorine), or H (hydrogen) at the four vertices. Again, the number of such molecules can be easily counted using Burnside's Theorem.

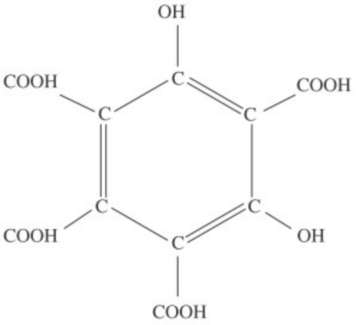

Figure 27.4 Benzene derivative.

Table 27.4 81 Colorings Fixed by (12)(34) (X and Y can be any of R, W, and B.)

Edge	Colorings								
12	X	X	X	X	X	X	X	X	X
34	Y	Y	Y	Y	Y	Y	Y	Y	Y
13	R	R	R	W	W	W	B	B	B
24	R	R	R	W	W	W	B	B	B
23	R	W	B	W	R	B	B	R	W
14	R	W	B	W	R	B	B	R	W

Group Action

Our informal approach to counting the number of objects that are considered nonequivalent can be made formal as follows. If G is a group and S is a set of objects, we say that G *acts on* S if there is a homomorphism γ from G to $\text{sym}(S)$, the group of all permutations on S. (The homomorphism is sometimes called the *group action*.) For convenience, we denote the image of g under γ as γ_g. Then two objects x and y in S are viewed as equivalent under the action of G if and only if $\gamma_g(x) = y$ for some g in G.

Notice that when γ is one-to-one, the elements of G may be regarded as permutations on S. On the other hand, when γ is not one-to-one, the elements of G may still be regarded as permutations on S, but there are distinct elements g and h in G such that γ_g and γ_h induce the same permutation on S [that is, $\gamma_g(x) = \gamma_h(x)$ for all x in S]. Thus, a group acting on a set is a natural generalization of the permutation group concept.

As an example of group action, let S be the two diagonals of a square and let G be D_4, the group of symmetries of the square. Then γ_{R_0}, $\gamma_{R_{180}}$, γ_D, $\gamma_{D'}$ are the identity; $\gamma_{R_{90}}$, $\gamma_{R_{270}}$, γ_H, γ_V interchange the two diagonals; and the mapping $g \to \gamma_g$ from D_4 to sym(S) is a group homomorphism. As a second example, note that $GL(n, F)$, the group of invertible $n \times n$ matrices with entries from a field F, acts on the set S of $n \times 1$ column vectors with entries from F by multiplying the vectors on the left by the matrices. In this case, the mapping $g \to \gamma_g$ from $GL(n, F)$ to sym(S) is a one-to-one homomorphism.

We have used group actions several times in this text without calling them that. The proof of Cayley's Theorem (Theorem 6.5) has a group G acting on the elements of G; the proofs of Sylow's Second Theorem and Third Theorem (Theorems 23.4 and 23.5) have a group acting on the set of conjugates of a Sylow p-subgroup; and the proof of the Generalized Cayley Theorem (Theorem 24.3) has G acting on the left cosets of a subgroup H.

Exercises

The greater the difficulty, the more glory in surmounting it.

Epicurus

1. Determine the number of ways in which the four corners of a square can be colored with two colors. (It is permissible to use a single color on all four corners.)

2. Determine the number of different necklaces that can be made using 13 white beads and 3 black beads.

3. Determine the number of ways in which the vertices of an equilateral triangle can be colored with five colors so that at least two colors are used.

4. A benzene molecule can be modeled as six carbon atoms arranged in a regular hexagon in a plane. At each carbon atom, one of three radicals NH_2, COOH, or OH can be attached. How many such compounds are possible? (Make no distinction between single and double bonds between the atoms.)

5. Suppose that in Exercise 4 we permit only NH_2 and COOH for the radicals. How many compounds are possible?

6. Determine the number of ways in which the faces of a regular dodecahedron (regular 12-sided solid) can be colored with three colors.

7. Determine the number of ways in which the edges of a square can be colored with six colors so that no color is used on more than one edge.

8. Determine the number of ways in which the edges of a square can be colored with six colors with no restriction placed on the number of times a color can be used.

9. Determine the number of different 11-bead necklaces that can be made using two colors.

10. Determine the number of ways in which the faces of a cube can be colored with three colors.

11. Suppose a cake is cut into six identical pieces. How many ways can we color the cake with n colors assuming that each piece receives one color?

12. How many ways can the five points of a five-pointed crown be painted if three colors of paint are available?

13. Let G be a finite group and let $\text{sym}(G)$ be the group of all permutations on G. For each g in G, let ϕ_g denote the element of $\text{sym}(G)$ defined by $\phi_g(x) = gxg^{-1}$ for all x in G. Show that G acts on itself under the action $g \to \phi_g$. Give an example in which the mapping $g \to \phi_g$ is not one-to-one.

14. Let G be a finite group, let H be a subgroup of G, and let S be the set of left cosets of H in G. For each g in G, let γ_g denote the element of $\text{sym}(S)$ defined by $\gamma_g(xH) = gxH$. Show that G acts on S under the action $g \to \gamma_g$.

15. For a fixed square, let L_1 be the perpendicular bisector of the top and bottom of the square and let L_2 be the perpendicular bisector of the left and right sides. Show that D_4 acts on $\{L_1, L_2\}$ and determine the kernel of the mapping $g \to \gamma_g$.

William Burnside

In one of the most abstract domains of thought, he [Burnside] has systematized and amplified its range so that, there, his work stands as a landmark in the widening expanse of knowledge. Whatever be the estimate of Burnside made by posterity, contemporaries salute him as a Master among the mathematicians of his own generation.

A. R. FORSYTH

By permission of the Master and Fellows of Pembroke College in the University of Oxford

WILLIAM BURNSIDE was born on July 2, 1852, in London. After graduating from Cambridge University in 1875, Burnside was appointed lecturer at Cambridge, where he stayed until 1885. He then accepted a position at the Royal Naval College at Greenwich and spent the rest of his career in that post.

Burnside wrote more than 150 research papers in many fields. He is best remembered, however, for his pioneering work in group theory and his classic book *Theory of Groups*, which first appeared in 1897. Because of Burnside's emphasis on the abstract approach, many consider him to be the first pure group theorist. One mark of greatness in a mathematician is the ability to pose important and challenging problems—problems that open up new areas of research for future generations. Here, Burnside excelled. It was he who first conjectured that a group G of odd order has a series of normal subgroups, $G = G_0 \geq G_1 \geq G_2 \geq \cdots \geq G_n = \{e\}$, such that G_i/G_{i+1} is Abelian. This extremely important conjecture was finally proved more than 50 years later by Feit and Thompson in a 255-page paper (see Chapter 24 for more on this). In 1994, Efim Zelmanov received the Fields Medal for his work on a variation of one of Burnside's conjectures.

Burnside was elected a Fellow of the Royal Society and awarded two Royal Medals. He served as president of the Council of the London Mathematical Society and received its De Morgan Medal. Burnside died on August 21, 1927.

= !"#$! ! !% & ' ! = > ' ' +' U) 9(
7 ? - ' %4 *) ' + G ' " #')') * 8 0-

@ !"#$! ! !% & ' ! = > ' ' +' +N '
D &' 0 9

< !"#$! ! !% & ' ! = > ' ' +' . ' & '
. 7 -- : ' F ' 5 / ")') * 8 0-

E !"#$! ! !% & ' ! = > ' ' +' . ? 4
0 '* G 9 7 - ' # 4 ' G'] F 0 +'9 +'9
8 0 9

!"#$! ! !% & ' ! = > ' ' +' 8 - 6)
) ' (> ' DL ' - > ' ') * 8 0-

!"#$! ! !% & ' ! @ # 9 - " * .
B * % ' G'4 &' 0-
%9 : 4 - ;;;

!"#$! ! !% & ' ! < S F & > & '0
- ' 7' 4' - ? 4 / 8 < 0-

!F"." G 4 4 !) * !
%9 F N
!F"." G 4 4 !) * ! !F . K - ' ' ` 4
D * - 9 ' >'
V 0 '- 0' ' ' 0' '

!F"." G 4 4 !) * ! !F . K - ' ' ` 4
D * - 9 ' >'

!"#$! ! !% & ' ! E - - 0' 0-
V F6D .7D%#6D FD# "#"::D::6G6DG
%G D?1:6F)6 "?+ >"? 5

!"#$! ! !% & ' ! E &' - ,9 / 0-
V >"? 5 :D FD?W)%:H"6 D. :" >a% a.#6D G%G D? :6F6DGGD

!"#$! ! !% & ' ! E & ' '9 ' 0-
V >D% D.#6E 7D ")7"GF:?G>DG 7 &

!"#$! ! !% & ' ! E - ' & 9 0-
V G%G D? :6FD"G >D% D.#H
" #6.6 ": "GF 76 .%#6 ": .?FH %+ 6. FD D:% DG.
)H
#%)D#.%)%G%:"

= !"#$! ! !% & ' ! E ' ' 0-
V .7D%#6D FD# .#"G +%# ".6%G >#? DG
%9 @= 0'&

@ !"#$! ! !% & ' ! E ' ' 0-
%9 =<@ 0'&

< !"#$! ! !% & ' ! E ' ' 0-
%9 E< 0'&

E !"#$! ! !% & ' ! E &' & -2 ' 0-
V ?#1?GFDG
S?1 >D 76 7.D FD#
G6 7.DR1:6F6 7DG >D% D.#6D
7D#"? >D>D)DG %G
+#6DF#6 7 DG>D: - "?: .b 1D:

!"#$! ! !% & ' ! E &' & -2 ' 0-

< =

= =E =

!"#$! ! !% & ' ! > /5 ' & '& ' - ' &/
B - 8 @ 0-

! !"#$! ! !% & ' ! > /5 ' & '& ' - ' &/
B - 8 @ -24

I& 8 @ > '4 /

%9 > '4 -'- ' ' - ' . ' - A &

!"#$! ! !% & ' ! = 6' ' ? ' 7 N ' -'
' ' ' I ' D- ' S' ' 8 0-
V = 0'&

!"#$! ! !% & ' ! = 6' ' B ' ' '
' ' / '00 ' - ' ' - - %, - ? 4 / 8 @ -24

!"#$! ! !% & ' ! = 6' ' B ' ' ' '
' / '00 ' - ' ' - - %, - ? 4 / 8 @ 0-

!"#$! ! !% & ' ! = 6' ' B /) ' / 6 .
. 7 / / /)') * 8 < 0-

= !"#$! ! !% & ' ! = 6' ' 8' . / /
0 4 + DL 9 ' ' 0' ' - / ' 0') * c)'
8 0-

@ !"#$! ! !% & ' ! = 6' ' 5 F'4 - .' .
+ -' ' ' %, - ? 4 / 8 = 0-

< =

@ E
!F"." G 4 4 !) * !

!F"." G 4 4 !) * ! !F . K - ' ' ` 4
D * - 9 ' >'

!F"." G 4 4 !) * ! ! ' ' '

!"#$! ! !% & ' ! = D 7 : * - " 9 * 4
? 4 / 8 @ 0-

!"#$! ! !% & ' ! = > /5 ' & '& ' - ' &/
B - 8 @ 0-
%9 /

!F"." G 4 4 !) * ! ! / / ' - ' 2 4
> /

!"#$! ! !% & ' ! = "00 - F ' > / "
- 6 - 0-

!"#$! ! !% & ' ! = 0 E 0-

= !"#$! ! !% & ' ! = > ' - ' ' 0-
%9 : 4 - ;;;

@ !"#$! ! !% & ' ! = / + -' '
/ 0-

< !"#$! ! !% & ' ! = / / ' - ' 2 4
> / < @ 0-

E !"#$! ! !% & ' ! = / / ' - ' -' - -
' ' ' - ' / 0-

!"#$! ! !% & ' ! = . * & > ' / " 4 /
> 8 F+F 4 0-
%9 : 4 - ;;;

!F"." G 4 4 !) * ! ! / /5 '0 ' - 0' "
6 - ' ' & > /

!"#$! ! !% & ' ! = 7 & ' / / ' -
. 0 ' > / 0-

!"#$! ! !% & ' ! = / /5 '0 ' - 0' "
6 - ' ' & > / 1 / & 8 @ 0-

!"#$! ! !% & ' ! = . 0 &/ ' 0 &
'& 8 0-

!F"." G 4 4 !) * ! !. . 9

!"#$! ! !% & ' ! = E@ 0-
V . . 9
" 8 # 9 H'
G' ' ' ' '&'3 5
=5 G 8F 5 5 00
9 - 9/ ' ' ' " ' " '
'9 ?#: 0 2 & '9 E@
" - @ = @ ?.

!"#$! ! !% & ' ! = DF =E =E 0-
V H' 5 # 9
. . 9
G' ' . ' ' ' 5 6 5

= !"#$! ! !% & ' ! = 0-
V % .#d #%):D " :e 6 % F" ".D e.6 " >#D>"

!F"." G 4 4 !) * !

!F"." G 4 4 !) * ! !G 7 3 > /

@ !"#$! ! !% & ' ! = G 3 & / .
' ' ' " ' " ' 8 0-
%9 : 4 - ;;;

!F"." G 4 4 !) * ! !# ' 4 $ ' + - . /

< !"#$! ! !% & ' ! = " 6 - # ' 4
$ ' + - . / F 4 9 ' 8 = 0-

E !"#$! ! !% & ' ! = > # ' 4 $ '
' ' - + - . / 0-

!"#$! ! !% & ' ! = > # ' 4 L '
' ' - - / B / 7 8 0-

!"#$! ! !% & ' ! = ' * ' - 8'
' 4 $ ' + - . /5 ' ' + ' &' C '/0
9 8 0-

!"#$! ! !% & ' ! = ' * ' - 8'
' 4 $ ' + - . /5 ' 6 & ' + ' &' C '/0
9 8 0-

!"#$! ! !% & ' ! = ' * ' - 8'
' 4 $ ' + - . /5 "00 ' $ ' + - . / 6%
/ 8 0-

!"#$! ! !% & ' ! = 9 " -
' 4 L ' - / F 4 9 ' 8 = 0-

!"#$! ! !% & ' ! = 9 " -
' 4 L ' - / # 8 @ 0-

!F"." G 4 4 !) * !6 0 ' M

!F"." G 4 4 !) * !6 0 ' M

@ @ E

@ @ @

!F"." G 4 4

!F"." G 4 4 !) * ! @

!F"." G 4 4 !) * ! @!D, > / + ' 9/
'- # 4 * 5 ' f B' 3'*

!"#$! ! !% & ' ! = < <E < @< @ 0-
V D, > / + '

!F"." G 4 4 !) * ! @!> / & . /
0' '

!"#$! ! !% & ' ! = " ' - . ' > /
& . / 0' ' ' 9 -& ? 4 / 8 0-

!F"." G 4 4 !) * ! @!G 6 ' ' D-
A' ' / =5

!"#$! ! !% & ' ! =
=(G ((6 ' ' (D- 0-
V " -' 0 / g

!"#$! ! !% & ' ! = 9 - ' 9 0-
V D, 0 4'

!F"." G 4 4 !) * ! <

= !"#$! ! !% & ' ! = 9 0 ' ' - > '
' 4 / B 0 ' 9 -& ? 4 / 8 0-

@ !"#$! ! !% & ' ! = 9 0 ' ' - > '
# ' 4 / 0-

!F"." G 4 4 !) * ! <! ' ' ' - / ' -
D & &

< !"#$! ! !% & ' ! = ' ') ' ' '

- / ' - D & & # 8 E 0-

I& =8 / ' - > 0 . /

E !"#$! ! !% & ' ! = ' ') ' '
' ' ' - 0 / ' - & & 0-

!"#$! ! !% & ' ! = / * ' ' ' -
0 / ' - & & # 5 .'/ C + ' > 0 8 E 0-

!F"." G 4 4 !) * ! <! 7/0 9 > /

!"#$! ! !% & ' ! = 7/0 9
> / 0-
%9 : 4 - ;;; 6 0 ' ;;;

!F"." G 4 4 !) * ! <!. ' " G ' /

!"#$! ! !% & ' ! =
. (' ((" ((G ' (/(. ' (' -(' ' (" 0 0-
I& =8 G ' +
%9 6 0 ' ;;;
V > '4 / h gggg

!F"." G 4 4 !) * ! E

!F"." G 4 4 !) * ! E!DL ' ' - " ' / ' .
' ' ' /

!"#$! ! !% & ' ! = H ' S DL ' ' -
" ' / ' . ' ' ' / 0 & 8 0-

!F"." G 4 4 !) * ! E!6 - ' ' '
/ G ' B'4

!"#$! ! !% & ' ! =
6 - (((' ' ' (/ ((G ' (B'4 0-

!F"." G 4 4 !) * ! E! " ' + 9 ' /

!"#$! ! !% & ' ! = (" ' (
(+ 9 ' /(0-
V . & 0' ' -

!F"." G 4 4 !) * ! E!. ' ' ' /
" 6 -

= !"#$! ! !% & ' ! = . ' ' '
/ 0-

!F"." G 4 4 !) * ! E!. & G '
$ ' D) / ' /

@ !"#$! ! !% & ' ! = . & 0-

!F"." G 4 4 !) * !

!F"." G 4 4 !) * ! !" 0 > - # ' ' -
0 , " ' / 6

< !"#$! ! !% & ' ! == " 0 > - # '
' - 0 , " ' / 6 0-

I& 8 = + & = . - 0

'0 = D,' 0)' ' 0' . L

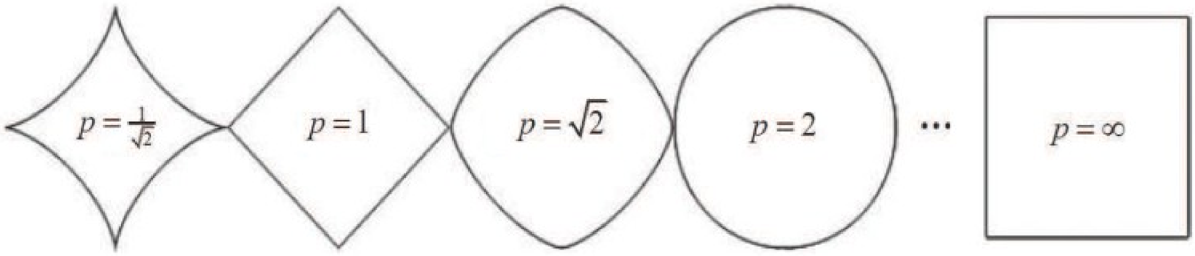
$p=\frac{1}{\sqrt{2}}$
$p=1$
$p=\sqrt{2}$
$p=2$
···
$p=\infty$

E !"#$! ! !% & ' ! ==
"(0 ((> - ((# ' (' -(0 ,(" ' / (66 0-

!"#$! ! !% & ' ! == 1 B & H " 0
> - # ' ' - 0 , " ' / 8B' # - U 8 0-

!"#$! ! !% & ' ! == 1 B & H 9 ' -
? - & '- ' # ' " ' / 8 E 0-

!F"." G 4 4 !) * ! ! & : & " - ' 4 5 G
' ' ' ' / ' &/ : &

!"#$! ! !% & ' ! ==
& ((: & ("(- ' 4 (G ' ' ' (' /(('(&/((: & 0-
I&

!F"." G 4 4 !) * ! !. F 00 - F
D, 0 ' 8"" 6% " /

!"#$! ! !% & ' ! == . F 00 -
F D, 0 ' 0-

!"#$! ! !% & ' ! == "-4' - 9 * ' '
% / 0 '- 8 F+F 4 0-

!"#$! ! !% & ' ! == 2 4 (> / 0-

!F"." G 4 4 !) * !

= !"#$! ! !% & ' ! =@ '& / 0-
V : i ' &K K - 0
I& @8 + &

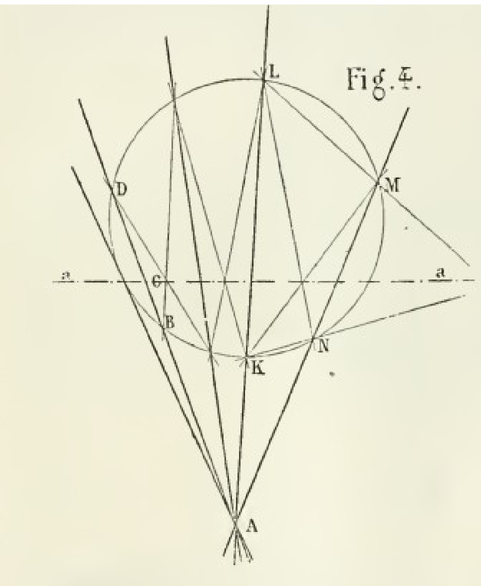
Fig. 4.
L
D
M
a
C
a
B
K
N
A

-' ,'& -' 9 - - -
0 - j K 00 K 0 0 K &
- : -5 -k I) ' 5 L -' ,'
& I 8O - '& ' 9 2 & '
- , I - , 2m0 K 0 Z
0 - , K n -K n ' K 0
% 4 L -' ' -U , U 0' L - 0 0 -
' L 5 - 5 - ', - / " 5 0 K K
I ' - '5 K n ' 0 & ' - K - '
0 & ' - &K K ' K 5 0 ' ' ' 9' -' ' -
' K - L
' , 0 5 K n -
' ' 0 - K - U
0 ' 0 9 n - ' K
-U ' & 0
- K5 Z L ' - ' L
-K 0' L 0
' L U 0 n
' K
: 0 9 n L I
' & I
9 - - - 0
-' 9 - ' K -
' / -U K n L
- 0 0
' 0 0 K K - L 5

' L 4 0'
2 -' ' K - &K K ' ' / L 5 0' L ' -K '
' ' / L ' 3 0 L K L U 0 0 0 l '
0 0 K K 4K '9 2 -U 0 "
8 & K -' 0 ' -U 9 - - , n - 5 '
9 5 n - K ' 5 - , L L
-U -K O ' 0 ' 9 5 0' ,
0 *5:5 5! F , L L - & L 2 & 0
L - K - K ' 5 : 1G 1 :m
o 0 -U & 0' ' -K K '5 L U
' 0 ' - P - K " 0 " -
- ' 9 5 ' 0 ' U ' ' 0 ' , 0 - ' - - ,
' & L U 0 - " l ' 9 p " -' U
K - ' 9 5 U 0' K 0' ' G
0 4 ' '& - K n 0 0' 0
- K ' & l ' L 5 4' L - '
n& 'L K ' K 0' " 4 L ' 0
' L '9 L " Z 5 0 0 0 -U
) '4 ' 9 5 0 -U - '
- '4 ' 0 ' ' - " 5 - 0 -
F '4 ' 9 L ' 0 L ' 0 '
L ' 0 ' ' ' 0 L ' L
K0' K - " 0' - , 0 - ' 9 : K n
- ' n 2 &K L - '
l ' 0' - K n L U 4 -UK -
UK - ' ' , ' - - - 5 0' L U
&K K ' 9 - - - '4 0 ' L
0

%9 l& =8 + &

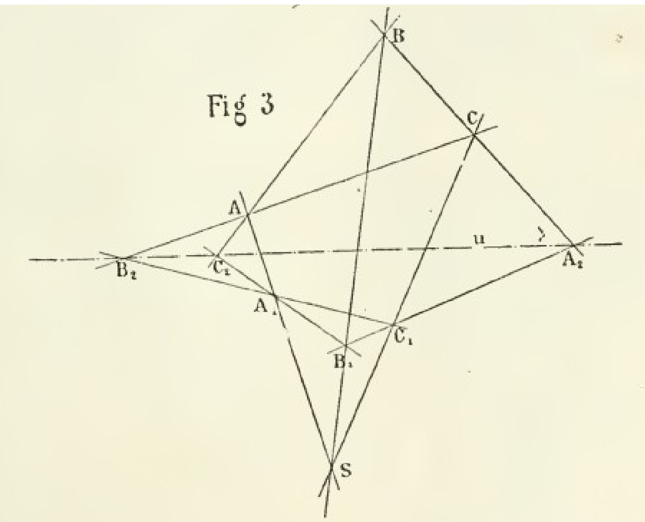

%9 ? ' '00 U . 6 0 ' ;;;

!F":" G 4 4 !) * ! !$ ' ' - '

@ !"#$! ! !% & ' ! =@ $ ' ' - '
+ -' ' 0 & 6 ' ' 9 & 8 E 0-

< !"#$! ! !% & ' ! =@ $ ' ' - '
"00 ' ' - D, 0 & 6 ' ' 9 & 8 E 0-

E !"#$! ! !% & ' ! =@ ' 9 3 0-
V . ' : 9 3 6 4 F ' '

!"#$! ! !% & ' ! =@ " 0-
V 6 ' ' F 4 / "
%9 : 4 - ;;; 6 0 ' ;;;

!F"." G 4 4 !) * ! ! '

!"#$! ! !% & ' ! =< 0 0 # ' 4 /(' (0-

!"#$! ! !% & ' ! =< 0 0 # ' 4 /(' (0-

!"#$! ! !% & ' ! =< 0 0 # ' 4 /(' (0-

!F"." G 4 4 !) * ! !%0 ' 0' + & G'4 &' "
" ' / ' "00 '

!"#$! ! !% & ' ! =< <EE@ 0-
V %0 ' 0' + & G'4 &'

!"#$! ! !% & ' ! =< " . ' "-4' -
' ' 5 0' ' ' - * A B / C 8 0-

= !"#$! ! !% & ' ! =< %0 ' 0' + & G'4 &'
" " ' / ' "00 ' 0 & 6 ' ' 9 &5) * c 8 0-

@ !"#$! ! !% & ' ! =< " . ' + -'
0' F/ ' B / 8 0-

!F"." G 4 4 !) * ! !#'/ 5 B'4 5 ' - ' & . 0
' ' ' ' ' /

< !"#$! ! !% & ' ! =< #'/ 5 B'4 5 ' - ' &
. 0 ' ' ' ' ' / ? 4 / 8 < 0-
I& E 8 @ B'4
I& 8 E . ' '

!F"." G 4 4 !) * ! ! & ' 5 / /5 ' -
>' & - : ' & ' -

E !"#$! ! !% & ' ! =< & ' 5 / /5 ' -
>' & - : ' & ' - 8 0 ' / ' ' " 0-

!F"." G 4 4 !) * ! !. F 00 - F
D, 0 ' 8"" 6% " /

!"#$! ! !% & ' ! =< . F 00 -
F D, 0 ' 0-

!F"." G 4 4 !) * ! !. " 7 /
. - . & : 4 6 > ' 6 4

!"#$! ! !% & ' ! =< . " 7 /
. - . & : 4 6 > ' 6 4 0-

!F"." G 4 4 !) * !

!F"." G 4 4 !) * ! !" &'9 - :q & 3 7q
' ' *

!"#$! ! !% & ' ! =E " &'9 - :q & 3
7q ' ' * 0 & 8 E 0-

!"#$! ! !% & ' ! =E " &'9 - :q & 3
7q ' ' * 0 & 8 E 0-

!"#$! ! !% & ' ! =E " &'9 - :q & 3
7q ' ' * 0 &) 7 - 9 &(0 & 0 * 8 0-

!"#$! ! !% & ' ! =E 1 ' 7q &5 Aq & 7q 8'
" &'9 - :q & 3 7q ' ' * 0 & 0 * 8 < 0-

!F"." G 4 4 !) * ! ! ' ' ' " 0
- '9 '

= !"#$! ! !% & ' ! =E ' ' ' " 0
- '9 ' 0-

!F"." G 4 4 !) * ! ! "4 + 4

@ !"#$! ! !% & ' ! =E (("4 (
(+(4 (0-
V > '4 ' :' . K L -K0' D

!F"." G 4 4 !) * ! !. " ' / 9/ + - 2 4 6 &

< !"#$! ! !% & ' ! =E 0-
V . " ' /

E !"#$! ! !% & ' ! =E + - 2 4 6 &
' 0 & 8 E 0-
I& = 8@ '0
'

4 ' - '
%9 : 4 - ;;;

I& =@ 8=< '0
. D - ' 0' D

%9 I& =@ 8=<= + & - ' ' '- - '

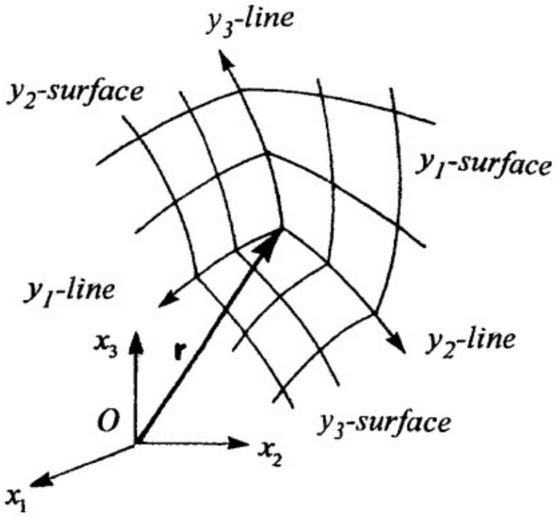

y_3-line
y_2-surface
y_1-surface
y_1-line
x_3
r
y_2-line
O
y_3-surface
x_2
x_1

!"#$! ! !% & ' ! =E + - 2 4 6 & . " ' /
0 & 8 0-

= !"#$! ! !% & ' ! =E + - 2 4 6 & 8' # &/
' - G G ' + - 0 & 6 ' ' 9 & 8 0-

= !"#$! ! !% & ' ! =E B @ @)' 6 # &/ 0-

!F"." G 4 4 !) * ! !. B & 6 L ' / ' -
7 9 -

= !"#$! ! !% & ' ! =E <(<E E 0-

I& =8 '0 ' % 9

I& = 8 == "00 - , " G 0 4 ' 4'

!F"." G 4 4 !) * !

!F"." G 4 4 !) * ! !" ' ' ' 0 . 0
: 9 ' "

= !"#$! ! !% & ' ! = " ' ' ' 0 . 0
: 9 ' " 0 9
%9 : 4 - ;;;

!F"." G 4 4 !) * ! !"00 ' . " ' / 8F 4
) * ' '

= !"#$! ! !% & ' ! = ' - 5 . " ' / 5 ' -
"00 ' 0 & 8 < 0-

!F"." G 4 4 !) * ! !D ' G / ' -
. ' G ' " 0 /

== !"#$! ! !% & ' ! = D ' G / ' -
. ' G ' " 0 / 0-
%9 : 4 - ;;;

!F"." G 4 4 !) * ! !+ - . / " - D-

=@ !"#$! ! !% & ' ! = #' - + - . / "
- # -& 8 0-
I& E 8U E '0 @
>' & / H' & ' - > '4 /
%9 0 ' ;;;

!F"." G 4 4 !) * ! !> '4 ' ' B'4 / ' -
" 0 / " ' ' U & -

=< !"#$! ! !% & ' ! = <E <= = 0-
V > '4 ' ' B'4 / ' - " 0 /
I& 8 #'- ' > '

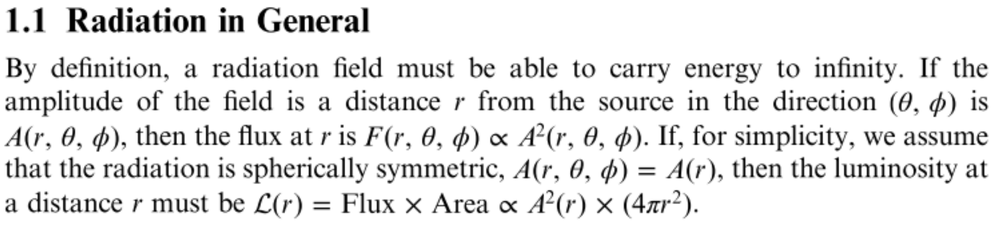

1.1 Radiation in General

By definition, a radiation field must be able to carry energy to infinity. If the amplitude of the field is a distance r from the source in the direction (θ, ϕ) is $A(r, \theta, \phi)$, then the flux at r is $F(r, \theta, \phi) \propto A^2(r, \theta, \phi)$. If, for simplicity, we assume that the radiation is spherically symmetric, $A(r, \theta, \phi) = A(r)$, then the luminosity at a distance r must be $\mathcal{L}(r) = \text{Flux} \times \text{Area} \propto A^2(r) \times (4\pi r^2)$.

%9 0 ' /

!F"." G 4 4 !) * ! !: > /

=E !"#$! ! !% & ' ! = &/ ' - F '
+ F 4 9 ' 8 @ 0-

= !"#$! ! !% & ' ! = : > /
8> '- ' - ' ' 5 < " ' ' ' ' / 8 0-
%9 : 4 - ;;;

!F"." G 4 4 !) * ! ! - $ ' + -
. /

@ !"#$! ! !% & ' ! = - $ ' + -
. / + -' ' 8 0-

@ !"#$! ! !% & ' ! = - $ ' + -
. / "-4' - . 0 8 0-

!F"." G 4 4 !) * ! ! - / " ' '00 '

@ !"#$! ! !% & ' ! = - / " '
'00 ' 0-

!F"." G 4 4 !) * ! ! - $ ' / " ' '
"00 ' "00 '

@ !"#$! ! !% & ' ! = <E <= @< 0-
V - $ ' ' ' - $ ' 6 '
%9 F '

!F"." G 4 4 !) * ! !% & ' / ' - 0' > /

@ !"#$! ! !% & ' ! = % & ' / ' - 0'
> / 0 & '& G H * 8 E< 0-
I& < 8E '0 2 4 . ' '
. '0 - 4 0 ' ' ' 4 - - ' & '
' 0 ' 5 9/ 9 -- & ' 0 2 4 0 ' 5 ' - &
& ' / ' - ' ' , 0 '9 '
' ' / . 9 '
9 L ' ' 0' & -
5 ' - - 0 L 4 ' 9 ' ' 9 -/
'-- ' - '

I& <=8E + &

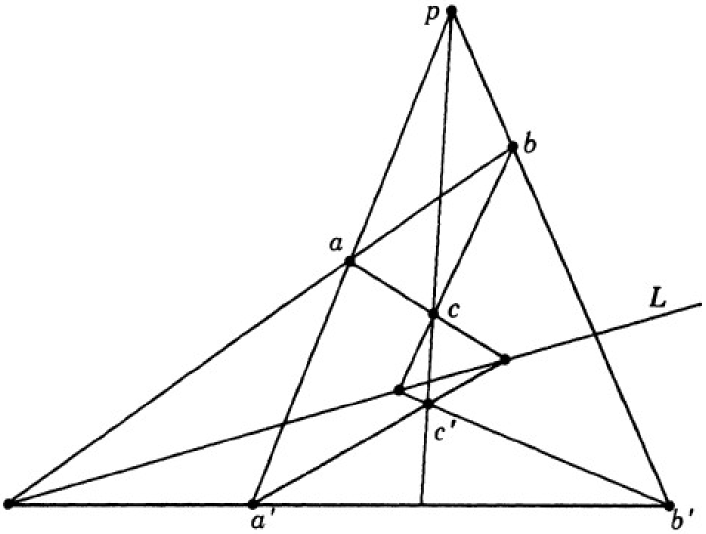

%9 F ' & ' '

? 3' 0' ' - '00 U .

3

Projective Transformations

This chapter develops an alternative method of coordinatising a metric affine plane, by embedding it into a *projective* plane, and using the orthogonality relation to define a matrix-representable transformation on the line at infinity. The construction will be central to our subsequent treatment of metric affine spaces of higher dimension, and its description now requires us to review a substantial body of additional ideas.

3.1 Projective Planes

On a couple of occassions in the previous chapter we alluded to the idea of parallel lines meeting at a "point at infinity". This can be made precise in the following way. Let $\alpha = (\mathcal{P}, \mathcal{L}, \mathcal{I})$ be an affine plane. The relation $L \parallel M$ of parallelism is an equivalence relation on $\mathcal{L}$. $\mathcal{P}$ is enlarged by adding one new point for each parallelism class, and then the set of all these new points is declared to be a new line L_∞ "at infinity". A set-theoretically convenient way to carry this out is to *define* L_∞ to be the set of equivalence classes (parallelism classes) of $\mathcal{L}$ under the relation $\parallel$. The new incidence structure can then be specified to be

$$\pi(\alpha) = (\mathcal{P} \cup L_\infty, \mathcal{L} \cup \{L_\infty\}, \mathcal{I}'),$$

where $a\mathcal{I}'L$ iff either

(1) $a\mathcal{I}L$ in α; or

(2) $a \in L_\infty$, and either $L = L_\infty$, or else $L \in a$ (i.e. the α-line L belongs to the parallelism class a).

It is then straightforward to show that $\pi(\alpha)$ satisfies the following properties.

P1. *Any two distinct points lie on exactly one line.*

P2. *Any two distinct lines pass through exactly one point.*

P3. *There exists* a four-point, *i.e. a set of four distinct points, no three of which are collinear.*

Any incidence structure satisfying P1, P2, P3, is called a *projective plane*. We have just seen that an affine plane α can be embedded in a projective plane $\pi(\alpha)$ - the *completion* of α. Conversely, if $\pi = (\mathcal{P}, \mathcal{L}, \mathcal{I})$ is a projective plane, any line L in π may be selected to act as the "line at infinity", and be deleted to give the incidence structure

$$\alpha(\pi - L) = (\mathcal{P} - \{a : a\mathcal{I}L\}, \mathcal{L} - \{L\}, \mathcal{I}),$$

which proves to be an affine plane. There is a natural isomorphism between π and the completion of $\alpha(\pi - L)$, which identifies L with the line at infinity of this completion by identifying point a on L with the parallelism class of all lines in α that meet at a in π.

Axioms P1 and P2 indicate that points and lines have equal "status" in projective planes, and leads to the *principle of duality*, stating that from any theorem provable from P1 - P3 another such theorem results immediately by interchanging the roles of points and lines. For instance, the dual of P3 states that there exist four lines, no three of which are concurrent. As the reader may care to verify, this can be derived from P1 - P3. Thus if $\pi = (\mathcal{P}, \mathcal{L}, \mathcal{I})$ is a projective plane, so too is the *dual plane*

$$\pi^{\partial} = (\mathcal{L}, \mathcal{P}, \mathcal{I}^{-1}),$$

where "point" L is incident with "line" a in π^{∂} if, and only if, $a\mathcal{I}L$ in π.

A property will be satisfied by π^{∂} precisely when its dual is satisfied in π. Thus if a property holds in all projective planes, then in particular it holds in all duals π^{∂}, and so its dual property holds in all projective planes π.

Desarguesian Planes

A pair of triangles (non-collinear triples of points) a, b, c, and a', b', c' are *centrally perspective* from a point p if corresponding vertices are collinear with p, i.e. if the lines aa', bb', and cc' all pass through p (Figure 3.1).

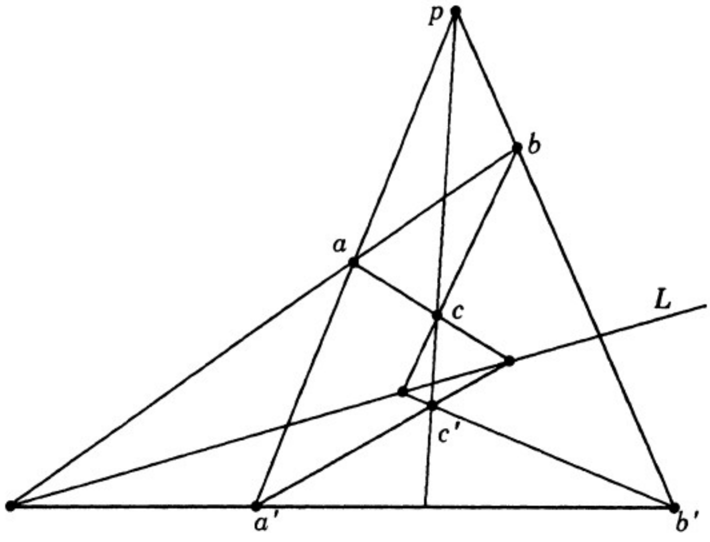

Figure 3.1

Dually, two triangles are *axially perspective* from a line L if corresponding sides are concurrent with L, i.e. if the points $ab \cap a'b'$, $ac \cap a'c'$, and $bc \cap b'c'$ all lie on L (strictly speaking, this dual statement should refer to *trilaterals* (non-concurrent triples of lines) associated with a, b, c and a', b', c'. Recall that the notation $L \cap M$ stands for the intersection point of lines L and M.)

A projective plane is *Desarguesian* if it has the

DESARGUES PROPERTY: *any two centrally perspective triangles are axially perspective.*

Notice that the dual of the Desargues property is its converse. In fact any Desarguesian plane satisfies this converse (cf. e.g. Garner [1981], p.73), so the principle of duality holds for the class of Desarguesian projective planes.

The Desarguesian property may be analysed more closely relative to a line L in a projective plane, as follows.

DESARGUES (L) PROPERTY: *if two triangles are centrally perspective from any point p not on L, and two pairs of corresponding sides meet on L, then the third pair of corresponding sides meet on L.*

LITTLE DESARGUES (L) PROPERTY: *if two triangles are centrally* *perspective from a point p on L, and two pairs of corresponding sides meet on L, then so does the third.*

The Desargues (L) property implies the Little Desargues (L) property (exercise: prove this by drawing a diagram, or see §5.2 of Stevenson [1972] for a more sophisticated proof involving transitivity of groups of collineations).

Recalling the Affine Desargues properties of §2.1 (Figures 2.4, 2.5), it is evident that an affine plane α is Desarguesian precisely when its projective completion $\pi(\alpha)$ has the Desarguesian (L_∞) property. Likewise, α has the Little Affine Desargues property precisely when $\pi(\alpha)$ has the Little Desargues (L_∞) property. Thus if π is any Desarguesian projective plane, and L is a line in π, the affine plane $\alpha(\pi - L)$ got by deleting L will be Desarguesian in the affine sense (since L may be regarded as the line at infinity in the completion of $\alpha(\pi - L)$).

Pappian Planes

PAPPUS (L, M, N) PROPERTY: *if a, b, c and a', b', c' are triples of collinear points on lines M and N, respectively, such that the points $ab' \cap a'b$ and $ac' \cap a'c$ lie on L, then $bc' \cap b'c$ lies on L as well* (Figure 3.2).

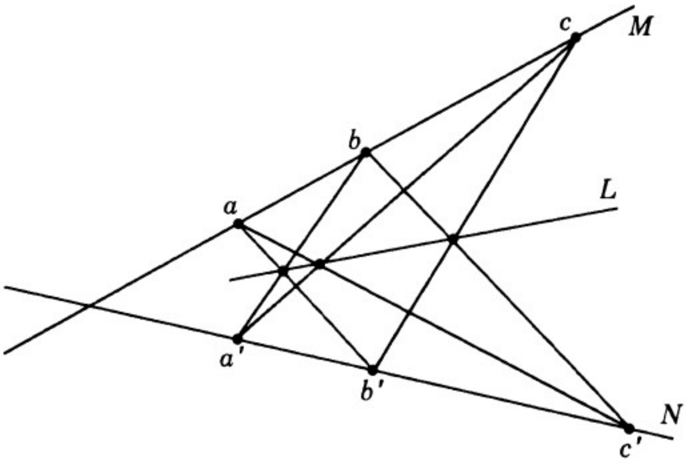

Figure 3.2

PAPPUS (L) PROPERTY: *the Pappus (L, M, N) property holds for all lines M and N distinct from L.*

A projective plane is *Pappian* if it has the Pappus (L) property for all lines L. Since the Pappus (L) property implies the Desargues (L) property (e.g. Stevenson [1972], §6.2), any Pappian projective plane is Desarguesian.

Recalling the Affine Pappus property of §2.1, we see that an affine plane α is Pappian precisely when its completion $\pi(\alpha)$ has the Pappus (L_∞) property. Thus if π is a Pappian projective plane, the planes of the form $\alpha(\pi - L)$ will all be Pappian affine planes.

Coordinates

Canonical examples of Pappian projective planes may be constructed out of the three-dimensional vector spaces F^3 over fields F. Let $\mathcal{P}$ be the set of all *lines* passing through the origin $\mathbf{o}$ in F^3, and $\mathcal{L}$ the set of *planes* of F^3 that pass through $\mathbf{o}$. If $p \in \mathcal{P}$ and $L \in \mathcal{L}$, define p to be *incident* with L if the line p lies in the plane L. This defines an incidence structure π_F which is readily seen to be a projective plane (any two lines (planes) through $\mathbf{o}$ are on a unique plane (line) through $\mathbf{o}$, etc.), and is moreover Pappian. Thus the bundle of affine lines and planes through the origin in F^3 becomes the collection of "points" and "lines" of π_F, which is known as the *projective plane over the point* $\mathbf{o}$, or the *plane at infinity* of F^3.

Now a line through the origin is a set of vectors of the form

$$p = \{\lambda b : \lambda \in F\},$$

where b is a particular (non-zero) direction vector for p. We may use b, or any of its scalar multiples λb as "coordinates" for the point p in π_F. Classically, particularly in relation to the real projective plane $\pi_{\mathbb{R}}$, it was customary to refer to points as being non-zero triples $b = (b_1, b_2, b_3)$ of "homogeneous" coordinates, and to regard any scalar multiple of b as determining the same point. Set-theoretically, we may consider the equivalence relation $\sim$ of *proportionality* on $F^3 - \{\mathbf{o}\}$, where

$$(x_1, x_2, x_3) \sim (y_1, y_2, y_3) \quad \text{iff} \quad \text{there exists } \lambda \in F \text{ with } x_i = \lambda \cdot y_i \text{ for } i = 1, 2, 3.$$

Observe that a *proportionality class*

$$[x_1, x_2, x_3] = \{(y_1, y_2, y_3) : (x_1, x_2, x_3) \sim (y_1, y_2, y_3)\}$$

(together with $\mathbf{o}$) is precisely a line through the origin in F^3. Thus we may take the points of π_F to be the set of such classes $[x_1, x_2, x_3]$.

A plane through the origin in F^3 is given by an equation

$$a_1 \cdot x_1 + a_2 \cdot x_2 + a_3 \cdot x_3 = 0, \tag{1}$$

with $a_1, a_2, a_3 \in F$ being not all zero. Replacing the coefficients of this equation by any proportional triple $(\lambda \cdot a_1, \lambda \cdot a_2, \lambda \cdot a_3)$ would give another equation for the same plane. Hence a line in π_F may be taken to be a proportionality class $< a_1, a_2, a_3 >$, with point $[x_1, x_2, x_3]$ being incident with point $< a_1, a_2, a_3 >$ precisely when equation (1) is satisfied. The use of the two different bracket notations $< \cdots >$ and $[\cdots]$ serves to distinguish points and lines in π_F. The fact that both are determined by non-zero triples of elements of F underscores the duality between points and lines in projective planes.

The association of the projective plane π_F with the three-dimensional geometry of F^3 provides a source of visual imagery that will guide much of what follows. For now it will be used to explain how π_F can be identified with (i.e. is isomorphic to) the projective completion $\pi(\alpha_F)$ of the affine field-plane α_F.

To begin with, α could be identified with a plane of F^3 that does not pass through the origin, say the horizontal plane $z = 1$. The correspondence $(x, y) \mapsto (x, y, 1)$ establishes this (Figure 3.3).

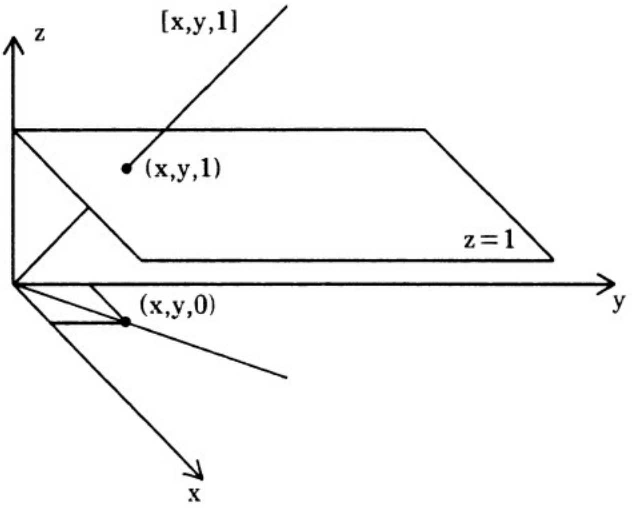

Figure 3.3

Alternatively, identify (x, y) with the projective point $[x, y, 1]$ in π_F (i.e. the line joining $(x, y, 1)$ to the origin). Thus the points of α_F

correspond to the lines through **o** that meet the plane $z = 1$, i.e. all lines through **o** except those, of the form $[x, y, 0]$, that lie in the x-y-plane. Any plane through **o** other than the x-y-plane will cut the plane $z = 1$ in a line, and so will correspond to a line in α_F. It then becomes clear that the x-y-plane itself corresponds to the line at infinity in the completion of α_F. For, given a line L through **o** in the x-y-plane, the bundle of all planes of F^3 containing L will cut $z = 1$ in a family of parallel lines. Moreover, any family of parallel lines in $z = 1$ arises from a line L of this type in the manner described, and the description establishes an exact correspondence between points of π_F of the form $[x, y, 0]$ and parallelism classes of lines in α_F, i.e. points at infinity in $\pi(\alpha_F)$.

This construction also provides a solution to the coordinatisation problem for any Pappian projective plane π. If L is a line in π, the structure $\alpha(\pi - L)$ is a Pappian affine plane, so is isomorphic to the plane α_F over some field F. Hence, as above, $\alpha(\pi - L)$ is isomorphic to the plane of all points $[x, y, z]$ with $z \neq 0$ in α_F. But then this isomorphism extends to one between π and π_F: each point p of L corresponds to a parallelism class of lines in $\alpha(\pi - L)$ (those lines having p as their point at infinity) and such a class of lines corresponds to a point of π_F of the form $[x, y, 0]$. In this way it is seen that any Pappian projective plane is isomorphic to a projective field-plane π_F. Furthermore, this *coordinatisation* can be set up in such a way that any prescribed line L becomes the line at infinity, i.e. the line $<0, 0, 1>$ of points of the form $[x, y, 0]$.

3.2 Projectivities and Involutions

If L is a line in a projective plane, the set $[L]$ of points on L is called a *range*, with *axis* L. If p is a point, the set $[p]$ of all lines passing through p is a *pencil*, with *centre* p.

If p is not on L, there is a natural bijection $f : [L] \to [p]$ obtained by joining each point of L to p, i.e. f maps point x on L to the unique line xp of the pencil $[p]$ that passes through x (Figure 3.4). f is called the *projection* of the range $[L]$ from p, and will be denoted $[L] \triangleright [p]$.

Dually there is the bijection $g : [p] \to [L]$ obtained by intersecting each line through p with L, i.e. $g(M) = M \cap L$ (Figure 3.5). g is the *section* of the pencil $[p]$ by L, and will be denoted $[p] \triangleleft [L]$. Clearly $[p] \triangleleft [L]$ is the inverse function to $[L] \triangleright [p]$.

A *projectivity* is a function obtained by composing a finite sequence of projections and sections. Such a sequence must consist *alternately* of projections and sections, since a projection can only

!F"." G 4 4 !) * ! !% & ' / ' - 0'
> /!% & ' / ' - 0' > / 0 & '& G H * 8 E< 0-

I& @8 + & =

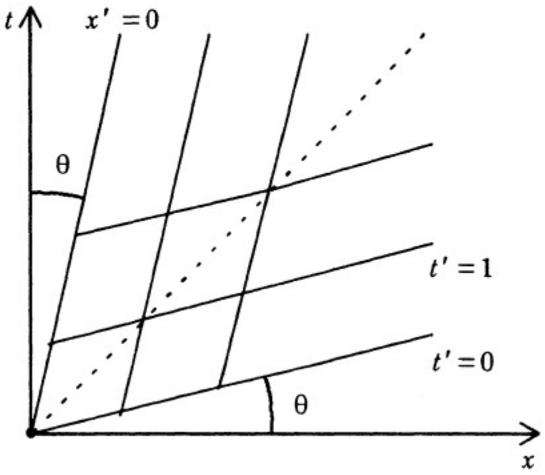
t
x' = 0
θ
t' = 1
t' = 0
θ
x

%9 ? 3' '00 U .

!F"." G 4 4 !) * ! ! & + - . / " -
6 -

@= !"#$! ! !% & ' ! = <(<E @= < 0-
V & + - . /

!F"." G 4 4 !) * ! !. . / -

@@ !"#$! ! !% & ' ! = $ ' - / ' / 9 -/
/ & - ' & & ' - %, - ? 4 /
8 -24

@< !"#$! ! !% & ' ! = #'/ & 5 A B . /
-5 F 4 9 ' 8 = 0-

@E !"#$! ! !% & ' ! = . . / -5 .
6 % F 4 9 ' 8 = 0-

!F"." G 4 4 !) * ! !. 0 &/ ' - > /

!"#$! ! !% & ' ! = . 0 &/ ' - > /
E= < 0-
%9 : 4 - ;;;

I& @ 8@=

'0 E
& / + / '& '
' ' ' - ' - 9 / -

!F"." G 4 4 !) * ! !> '4 ' ' B'4 / ' -
" 0 / " ' ' U & -

@ !"#$! ! !% & ' ! = <E <= = 0-
I& 8 '0
%4 4 > '4 ' ' #'- '
> '4 ' ' '- ' ' \ - - - / = ' 5 '
- 4 ' ' - 0 ' ' ,0 0' 92
8 5 9 ' * ' - ' ' 0 / ' B ' 5
& '4 ' ' '4 - '4 ,0' - - 9 / - 0 / ' - '
' /5 &/5 ' 0 / 5 ' - 0' 0 / 5 ' - '4 0 -
9 4' ' - -/ ' 0 '
4 . - & & ' ' - 9 ' * 5 0 4' ,0
5 ' - 0 9 / 4 / ' / / 4 ' '
> '4 ' ' '4 ' 0 / 5 4 5 ' 0 ' - ' -
' - ' / ' - ' ' - 0' 0 / 6 '& 9 4' 5
4 / '4 9' - ' ' ' & ' / ' 9 - -
* & ' / & '9 H -] - - ' - '
- ; ' /5 0' - - - -
- '/ 0 - & 9 ' ' ' 0' - 5 9 0 /
0' ' 9 - - 4 -] * L ' -/ ' 6
' 5 ' 5 ' / & '4 ' ' '- ' 4 ' '* '
0 ' - ' ' L ' L - , ' 0 0 /
- 5 ' - ' ' 3 . L 4 4 ' &
0 ' ,0 - '4 '&' 9 4 - -' ' ' ',
4 / L ' ' - ' - & ' '4 * * ' -
' ' ' 0 -
6 9 *5 / ' 4 / ' / ' 0 & '4 ' '
'- ' 5 ' & ' 0' -) - & - /0
5 & 5 \ - '4 & ' 0 0 4 & '4 '
' '- ' & ' - ' - & B 9 & 9/ - &
'- ' ' & ' ,
I& 8 #'- ' > '

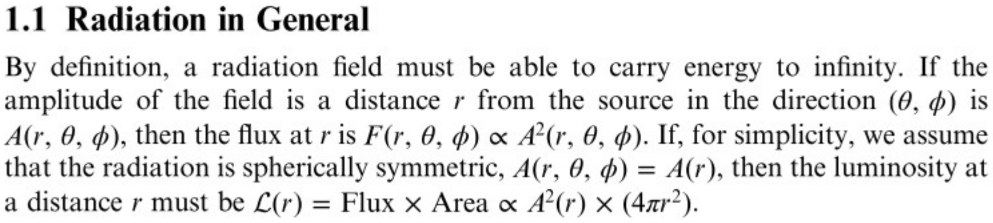

1.1 Radiation in General

By definition, a radiation field must be able to carry energy to infinity. If the amplitude of the field is a distance r from the source in the direction (θ, ϕ) is $A(r, \theta, \phi)$, then the flux at r is $F(r, \theta, \phi) \propto A^2(r, \theta, \phi)$. If, for simplicity, we assume that the radiation is spherically symmetric, $A(r, \theta, \phi) = A(r)$, then the luminosity at a distance r must be $\mathcal{L}(r) = \text{Flux} \times \text{Area} \propto A^2(r) \times (4\pi r^2)$.

%9 0 ' / /

!F"." G 4 4 !) * ! !"00 ' . " ' / 8F 4
) * ' '

< !"#$! ! !% & ' ! = A - D ' - ' - . - #' 5
'9 ' #' 0 "9 ' ' ' - 5 . " ' / 5 ' - "00 ' 8 <
4 0 & 8 < 0-

I& 8 + &

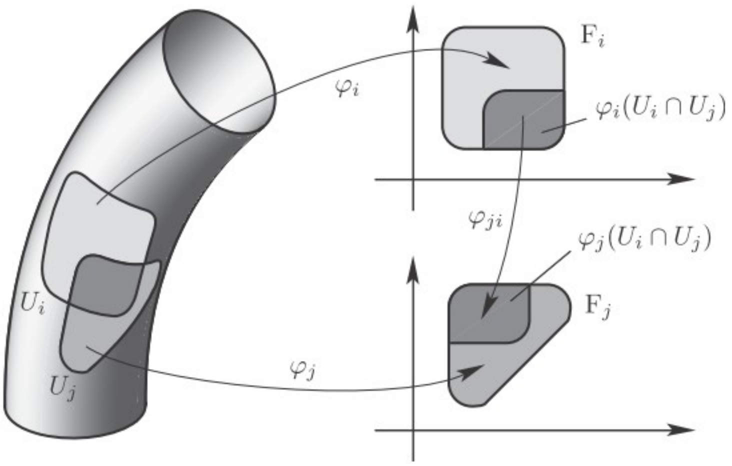

F_i
$\varphi_i(U_i \cap U_j)$
φ_i
φ_{ji}
$\varphi_j(U_i \cap U_j)$
F_j
U_i
U_j
φ_j

FIGURE 3.1.2. Charts φ_i and φ_j on a manifold

%9 : 4 - ;;;

= ==

=@ =

!F"." G 4 4 !) * ! !" ' ' ' 0 . 0
: 9 ' "

!"#$! ! !% & ' ! @ " ' ' ' 0 . 0
: 9 ' " 0 9

!"#$! ! !% & ' ! @ " ' ' ' 0 . 0
: 9 ' " 0-

!F"." G 4 4 !) * ! !> ' 3 - F/ ' '
$ ' / '

!"#$! ! !% & ' ! @ <(<E E @ @@ E = 0-
I& 8 '0
F/ ' E + - /
' $ ' / ' -

!F"." G 4 4 !) * ! !> ' ' +'

!"#$! ! !% & ' ! @ > ' ' +' . ? 4
0 '* G 9 7 - ' 0-
%9 : 4 - ;;; 6 ' ;;;
I& E 8= . : & F 4

6 = r - - 0 ' ' ' B & 4 / ' '- -
- 9
+ ' F/ 5 ' 0 /
+ ' F/ 5 ' 0' / '4 , - 9 ' '
' - 0 / 5 & 5 '& - / ' - ' 0 ' ' 0 G - '
9 & ' / - ' & 0 ' / & 5 ' ' 0
9 0 ' 92 s. ' 0 / ' - 0 ' ' 4 9
/ ' & 5 & ' - ']
F/ 0 ' 92 9 - & 4 ' / ' G
'00 - '-4' - ' ' : ' - D ' -5 4 ' - '- ' 5 ' + -
>' ' ,' 0 0 G ' * ' 9 ' '
' - 0 / 5 ' - '- 0 ' 9 9 . '-4' - &
' / /5 + ' 7 ' K - - ' * 9 92 6
' - - - B - B' 5 4 5 F/ ' -& &
' - ' ' 0 / ' - 0 ' ' '- 9 ' & -5 '
- 9 - ' ' s & - 4]
F & 0 -5 F/ ' 5 92 '- - '] 0' /5 '
9 4 & - ' . ' '- 0' 0'9 0 & & / ' '
/ '- ' - ' - 5 0 ' ' ' - ' '& -' ' L -
0 ' 4 0 / s" ' ' ' 5] F/ - '
s6 ' & - ' ' ' ' - '- 9 '00 / & / * 0 /]" & F/
' 0 ' s- - - 4]5 &' - ' ' -' * 0 -
- 0 ' - & ' * ' - '- 0' ' ' /
- - ' ' - & 9' * 0 '* &
. ' & 9 0 ' ' ' - ' 0 / '- 9 & / ' / '
' ' F ' ' * - : 0 9 / ' 92 '/ -'/
/ D 5 5 '- & - ' 0 / - 4 0 & ' '
9 5 0 / & ' 0 / - * 0 ' / ' ' ' &
-'/ 9 4')/ = 5 D ' / & s7 '4 - -
' - '4 - - 5] F/ - 5 '-- & ' '- & '9 4 & 5 '
/ - '4 '- ' * '9 F/ - -5 4 5 ' * F ' 9 *
'-4 ' ' ' 9 ' / F/ 0 - - 0 /
- '/5 9/ / & - ' - ' ' ' & 5 &
,0

6 0 '
7 3 - 0 ' , t - D - t 4' & '

= !"#$! ! !% & ' ! @ > ' ' +' . ? 4
0 '* G 9 7 - ' # 4 ' G'] F 0 +'9 +'9
8 0 9

@ !"#$! ! !% & ' ! @ > ' ' +' 8 - 6)
) ' (> ' DL ' - > ' ') * 8 0-

< !"#$! ! !% & ' ! @ . ' & ' . --
' F ' 5 / ')') * 8 0-

!F"." G 4 4 !) * ! !: 9 3

E !"#$! ! !% & ' ! @ ") 7 / ' ' '
. & 0-

!"#$! ! !% & ' ! @ F '
" 'D : 9 3D% ' ' = @= E 0-

!"#$! ! !% & ' ! @ + F ' 0'
- D- 8F 4) * ' ' 0-

!"#$! ! !% & ' ! @ A" ' F ' / . - ' -
' ' ; 0-

!"#$! ! !% & ' ! @ ' 9/ ' : * ' 83
9 & 0-
%9 : 4 - ;;;

!"#$! ! !% & ' ! @ .7D)%%1 9/ '
" & 5 >Q S & 83 9 & 0-

!"#$! ! !% & ' ! @ . ' ' & 4
,0' 5 & ' ' 5 ' - 9 ' / 9/ ' : 4 83 9 & 0-

= !"#$! ! !% & ' ! @ . > - #' . / 765
B - " & G 9 9/ ' : 4 83 9 & 0-

@ !"#$! ! !% & ' ! @ . A / , " > - - .
' 5 % 6 / 9/ 4 &' 3 83 9 & 0 9

!F"." G 4 4 !) * ! ! ') *

< !"#$! ! !% & ' ! @ '(1 D-
. (/((' ' ' (0 ((' ('- ' 9 -& (? 4 /((0-
V @ 0'&

E !"#$! ! !% & ' ! @ %, - 7' -9 * 7 /
' ' 8%, - 7' -9 * 8 F+F 4 0-
V < 0'&

!"#$! ! !% & ' ! @ . ') * + / '& '
=< F 5 = 7 / ' ' 8 &
8 F+F 4 0-

!"#$! ! !% & ' ! @ . 0') * # 4 - ' -
?0-' - +) & & D - . 5 = 7 / 0' C
" / 8 F+F 4 0-

!F"." G 4 4 !) * ! ! ' * - >

!"#$! ! !% & ' ! @ ' * - > 0-

!F"." G 4 4 !) * ! !#'/ 5 B'4 ' -

!"#$! ! !% & ' ! @ #'/ 5 B'4 ' - 0-

!F"." G 4 4 !) * ! ! * B'4 8> '- ' -

' ' 5 =

!"#$! ! !% & ' ! @ * B'4 8> '- ' -
' ' 5 = 0-

!F"." G 4 4 !) * ! ! & ' %0

!"#$! ! !% & ' ! @ & ' %0 0-
%9 : 4 - ;;;

!F"." G 4 4 !) * ! !? - ' - & ? 4 +
$ ' *

= !"#$! ! !% & ' ! @)' 0 / +
$ ' * 0 & 8 0-

@ !"#$! ! !% & ' ! @ & $ ' *
D 4 $ G / G' ' " '- 8 0-

< !"#$! ! !% & ' ! @ F : ? - ' - &
? 4 + $ ' * B - 9 & 0' / 8 0-

E !"#$! ! !% & ' ! @ + $ ' *)'
0 / 0 & 8 0-

!"#$! ! !% & ' ! @ ? - ' - & ? 4 .
/ $ ' ' $ ' * # 8 0-

=

!F"." G 4 4 !) * ! !# ' 5 # " - + '
' &

!"#$! ! !% & ' ! @ # ' 5 # " - + '
' & % & ' . 0 ' ' D & & ' 0-
%9 : 4 - ;;; 6 0 ' ;;;

!F"." G 4 4 !) * ! !D / F / -5 - D-

!"#$! ! !% & ' ! @
D /(F / -(-(D- 0 9

= E

=

!F"." G 4 4 !) * ! =!" ' / ' "-4' $ ' ' -
' '

!"#$! ! !% & ' ! @ " ' / ' "-4' $ ' ' -
' ' 0-

I& = 8 = '0 =
" 9 ' / 0 ' ' - ' 9 ' / ' D] & ' ' 4 / ' - G]
& '4 /

!F"." G 4 4 !) * ! =!" ' - ' / "
0

!"#$! ! !% & ' ! @ " ' - ' / 0-

!"#$! ! !% & '

!"#$! ! !% & ' ! @ " ' ' ' D-
0-
%9 : 4 - ;;;

!F"." G 4 4 !) * ! =! ' ' ' - - ' : 3
0'

!"#$! ! !% & ' ! @ = = <E <= == 0-
V ' ' ' - - ' : 3 0'

I& 8= : 3 0'

%9 : 4 - ;;; D 3 ' - : 3 ',

!F"." G 4 4 !) * ! =! ' ' ' =

= !"#$! ! !% & ' ! @ ' ' ' 5 0-
V . ' - 4

@ !"#$! ! !% & ' ! @ ' ' ' 5 0-
V 1 ' ' - / ' ' -

< !"#$! ! !% & ' ! @ ' ' ' 5 0-
V G] ' ' - '

E !"#$! ! !% & ' ! @ ' ' ' 5 0-
V . ? 4 ' ' & '4 '

!"#$! ! !% & ' ! @ ' ' ' 5 = 0-
V 4' ' ' - ' '

!F"." G 4 4 !) * ! =!> '0 '

!"#$! ! !% & ' ! @ > '0 ' ' - .
"00 ' 8 0-
V '

I& <8= + &

%9

2. Glue the sides of the polygons according to the permutation α in such a way that the orientation of the sides glued together is always opposite; this condition guarantees that the resulting surface will be oriented, see Fig. 1.19.

3. At each vertex, the darts will be automatically glued in a cyclic order according to the permutation $\sigma = \varphi^{-1}\alpha^{-1}$.

It remains to verify that in the topological space thus obtained, a neighborhood of every point is homeomorphic to an open disk in $\mathbb{R}^2$. For points inside the polygons it is true because they are parts of the plane. For points on edges it is true because we glue together exactly two polygons. Finally, for vertices it is true exactly because of the cyclic order on the "angles" of the polygons. See Fig. 1.19.

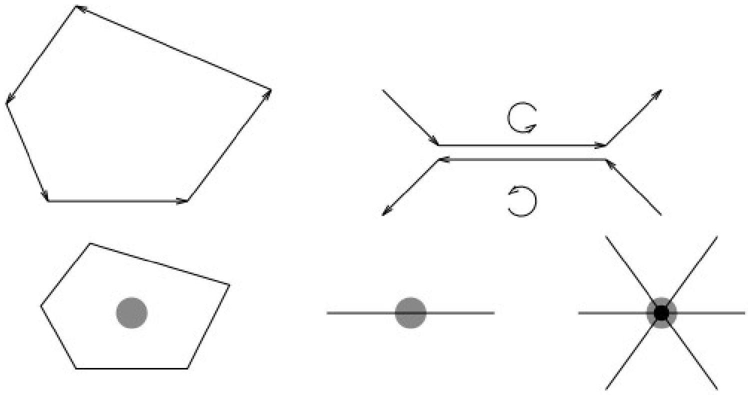

Fig. 1.19. Various stages of gluing a surface out of polygons

Remark 1.3.21. Let us return once more to the example shown in Fig. 1.11. There is no problem in labelling the darts. The permutation α is determined by the graph structure. We may even introduce σ: the cyclic order of the darts at each vertex is determined by the fact that the graph is drawn on an oriented surface. (In fact, in this example there is only one vertex of degree greater than 2, where such an order is significant.) Then why have we said that this figure does not represent a map? What exactly does not work here?

The answer is as follows. Having found α and σ, compute φ. You will get a face of degree 1 inside the loop, and an "outer face" of degree 5. Now apply the procedure of Construction 1.3.20. What you will find is not the torus but the sphere. If we insist on considering this figure as a map, we must consider it as a *plane* map. It is drawn on the sphere, and the handle of the torus stole in illegally.

Construction 1.3.22 (From a covering to a map). Let us give one more construction, which is probably even more efficient than the previous one.

%9 ' ' it ;;;

According to Proposition 1.2.15, for any constellation there exists a ramified covering of the sphere determined by this constellation. Take the constellation $[\sigma, \alpha, \varphi]$, and let $y_1, y_2, y_3 \in S^2$ be the corresponding ramification points on the sphere. Then the map in question is obtained as the preimage of the following "degenerate map" on S^2, shown in Fig. 1.20: it does not have even a single edge but only a half-edge, or a dart, represented by a segment connecting y_1 to y_2.

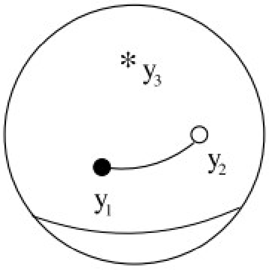

Fig. 1.20. A half-edge drawn on S^2; its preimage under the ramified covering corresponding to a constellation $[\sigma, \alpha, \varphi]$ is the map corresponding to the same constellation.

The preimages of y_2 are often called the *midpoints* of the edges, and the preimages of y_3 are often called the *centers* of the faces.

All the above considerations lead us to the following definition:

Definition 1.3.23 (Combinatorial map). A *combinatorial map* is a 3-constellation $[\sigma, \alpha, \varphi]$ in which α is an involution without fixed points. Two combinatorial maps are *isomorphic* if the corresponding constellations are isomorphic.

Remark 1.3.24. Many authors give an equivalent definition, in which they consider only the pair $[\sigma, \alpha]$ instead of the triple $[\sigma, \alpha, \varphi]$, because φ is determined by σ and α.

The correspondence between topological maps and combinatorial maps is one to many, because each topological map may be labelled in many different ways. But if two topological maps are isomorphic, then the corresponding combinatorial maps are also isomorphic.

A dictionary may easily be established between the two notions. For the vertices it looks like this:

a vertex of a map	a cycle of σ
vertex degree	length of the cycle
number of vertices	number of cycles in σ

and the same for the edges and α and for the faces and φ. On a more profound level, the following definition may be given:

Definition 1.3.25 (Automorphism group). The automorphism group of a map M, denoted by $\mathrm{Aut}(M)$, is the centralizer of its cartographic group G in S_n, that is, the set of the permutations which commute with all the elements of G:

$$\mathrm{Aut}(M) = \{h \mid h^{-1}gh = g \quad \forall g \in G\}.$$

Of course, in order to verify that a permutation h belongs to this set, it is sufficient to verify that h commutes with σ and α.

Remark 1.3.26. Here the reader should make a pause and meditate a little upon the difference between the notions of isomorphism and that of automorphism. The permutation model possesses many advantages, but there is also a disadvantage with which we will often struggle in the future. Namely, it is the possibility to label the same topological object in too many ways (in the "generic" case, in $n!$ ways), and thus obtain yet more and more new triples $[\sigma, \alpha, \varphi]$. All of them are isomorphic. (In the future we will try to reduce the number of labellings by imposing various restrictions on them.) But for the notion of automorphism the corresponding difficulty does not exist. In the above definition, the triple $[\sigma, \alpha, \varphi]$ is chosen arbitrarily at first, but is fixed afterwards; therefore, a permutation commuting with it is a "true" automorphism whose existence is not due to a relabelling. For the overwhelming majority of maps the automorphism group is trivial: $\mathrm{Aut}(M) = \{\mathrm{id}\}$.

Exercise 1.3.27. Verify on examples that if a map is mirror symmetric, then the permutation expressing this mirror symmetry is not an automorphism.

Remark 1.3.28. There exists one more subtle point hidden in Definition 1.3.25. In general, the centralizer is an object associated not to a group but to a pair consisting of a group and of its subgroup (it is the centralizer of the subgroup inside a given group). A cartographic group is a permutation group, and as such is embedded into S_n. But the same "abstract" group may act on different number n of points, and thus be embedded into different S_n. For example, the group A_5 acting on 5 points has trivial centralizer, while "the same" group A_5 acting on 60 points has as a centralizer a group isomorphic to A_5 (it is the automorphism group of the icosahedron). In this latter case both groups A_5 are embedded into S_{60}, but they are embedded differently. Geometrically this becomes clear if we note that the automorphism group respects the incidence of darts to vertices, edges and faces, while the cartographic group does not respect these relations. See also the discussion in Sec. 1.7.3.

Recalling the notation $c(g)$ for the number of cycles of a permutation g (see Notation 1.1.9) we may rewrite the Euler formula:

$$\chi(M) = c(\sigma) - c(\alpha) + c(\varphi) = 2 - 2g. \tag{1.3}$$

' '

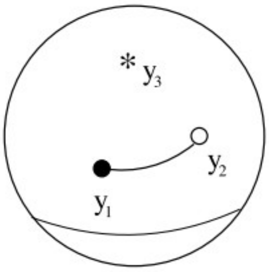

Fig. 1.20. A half-edge drawn on S^2; its preimage under the ramified covering corresponding to a constellation $[\sigma, \alpha, \varphi]$ is the map corresponding to the same constellation.

!"#$! ! !% & ' ! @ > '0 ' . A
7 0* ? 4 / 8 -24

!"#$! ! !% & ' ! @ > '0 ' . A
7 0* ? 4 / 8 0-

!F"." G 4 4 !) * ! =!6 " G "-4 . &
8 ' 9 -&) & '0

!"#$! ! !% & ' ! @ " # 0 7' 6 " G (
"-4 . & 8 ' 9 -&) & '0 8 @ -24

!"#$! ! !% & ' ! @ " # 0 7' 6 " G (
"-4 . & 8 ' 9 -&) & '0 8 @ 0-

= !"#$! ! !% & ' ! @ # 0 7' " 8 - .
- + >' G @ < 7' 0 C # 8 @ -24

!F"." G 4 4 !) * ! =! '00 D *
%9 6 0 ' ;;;

@ !"#$! ! !% & ' ! @ = 0-
V D ' / 0 2 4 > /
%9 '00 U .

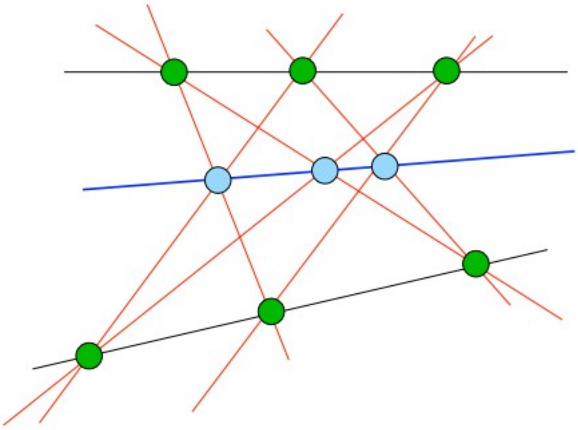

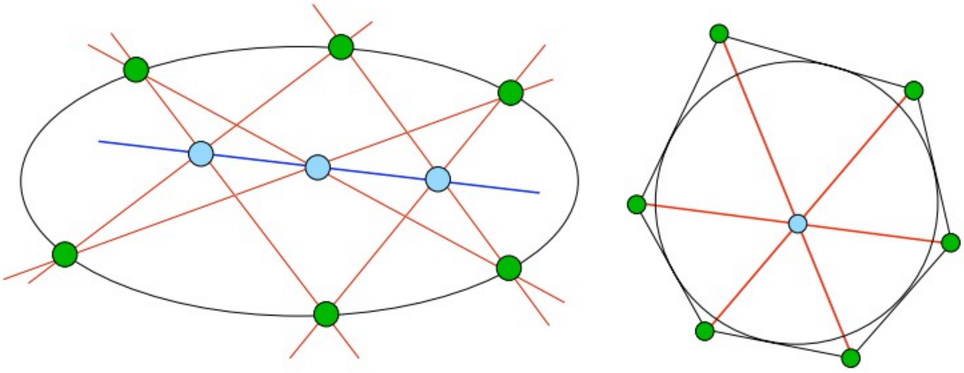

+ & ' '] . '-) 'u] .

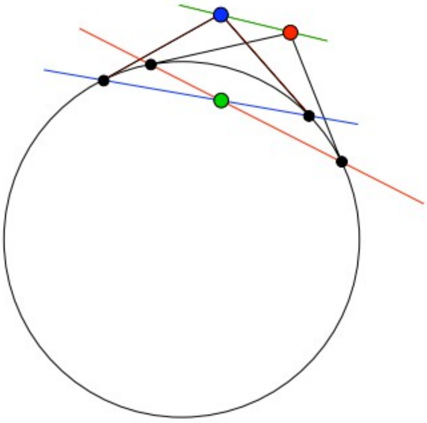

+ & 2 4 - ' /

< !"#$! ! !% & ' ! @ <E @ = 0-
V :D .#%6 :6 #D FD %#6 D FUD? :6FD

E !"#$! ! !% & ' ! @ ' ' ' ' 0-

!"#$! ! !% & ' ! @ - 0-
V + -' D - ' > /

!"#$! ! !% & ' ! @ D - ('-' 0-

!"#$! ! !% & ' ! @
& - 0 4 &' 0' - & 0-
V >K K F 0 4

!"#$! ! !% & ' ! @
& / 4 -(, & 3 0-

!"#$! ! !% & ' ! @ 4 - 0-
V -UD -

!"#$! ! !% & ' ! @ ', 6 0-

= !"#$! ! !% & ' ! @ ', 66 0-

@ !"#$! ! !% & ' ! @
'00 (-" ,' - 5(- (' ((D * (:'((' K ' L 0-
%9 : 4 - ;;; 6 0 ' M ;;;

< !"#$! ! !% & ' ! @ ' & 7 /
> * ' ' 6 7' 4' - ? 4 / 8 =< 0-

E !"#$! ! !% & ' ! @ ' & 7 /
> * ' ' 66 7' 4' - ? 4 / 8 =< 0-

!"#$! ! !% & ' ! @ 9 0-

!"#$! ! !% & ' ! @ 9 0-

!"#$! ! !% & ' ! @ ' -& 0-
V " . ' ' - 0

!F"." G 4 4 !) * ! =! '00 D * !D,0 & > /

!"#$! ! !% & ' ! @ F ' - : D,0 &
" ' / > / ' ' ' " '- 8 0-

!"#$! ! !% & ' ! @ D,0 & & / #
8 < 0-
%9 : 4 - ;;;

I& E '0 = . ' ' ' > /

I& = '0 @ / /

I& < '0 2 4 > /
' K
I& ==

I& = @ 7 9] ", D - ' > /

? - \ - ' 0 5 5 9 5 8 - 5 ' - &
F 6G 6FDG D "W6%
6 . & ' / - 0 " ' -) ' '/ '

6 . & ' / - 0 " ' -)5
'
6 % 4 / , ' ' - 0 . ,
' ' 0 ' '
6 . & ' / 0 ' 5 ' -
/ 0 '

!"#$! ! !% & ' ! @ ' 74 - D,0 &
> / # 8 < 0-

= !"#$! ! !% & ' ! @ A " '* & > /(
D,0 & . F ' + 0-
I& E8<

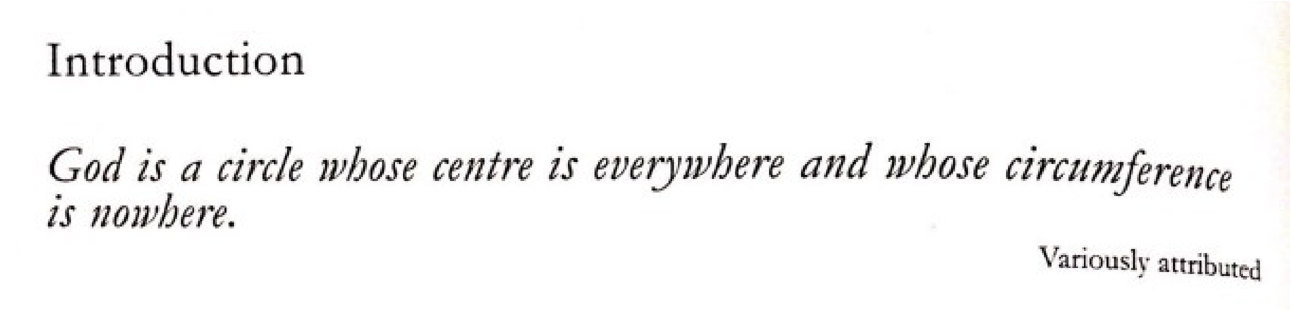

Introduction

God is a circle whose centre is everywhere and whose circumference is nowhere.

Variously attributed

!F"." G 4 4 !) * ! =! '00 D * !+ -'
> / ' - G D - ' '

@ !"#$! ! !% & ' ! @ > 0 &
'& G H * 8 E 0-
V >

I& =8 =

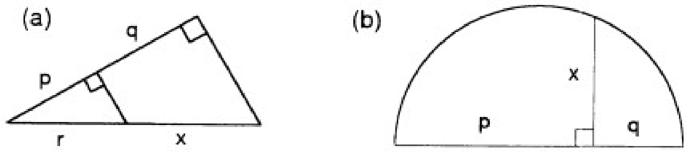

FIGURE 1.7.

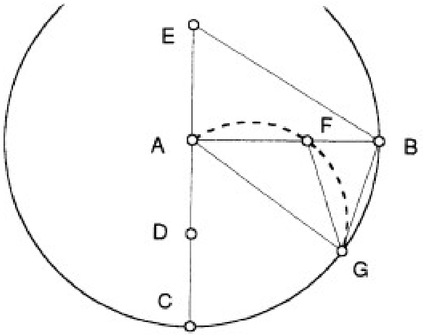

FIGURE 1.8.

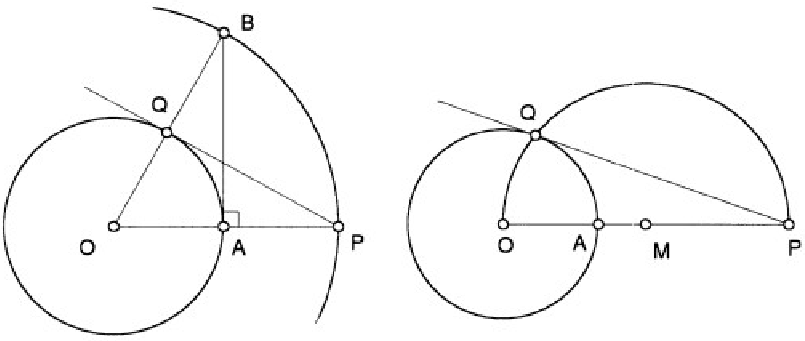

FIGURE 1.6.

Euclid III.37. If Q,R,S are points on a circle, point P is outside the circle, secant $\overleftrightarrow{RS}$ passes through P, and $(PR)(PS) = PQ^2$, then $\overleftrightarrow{PQ}$ is a tangent to the circle.

We turn to the constructions from Book IV of the *Elements* after mentioning only the following two easy constructions from later Books. Figure 1.7a illustrates VI.12. Here, we have $p/q = r/x$ and so $x = rq/p$. Hence, by taking p or q to be 1, we see that VI.12 essentially gives us the product of given lengths r and q as well as the quotient of given lengths r and p. Figure 1.7b illustrates VI.13, which echoes II.14. Here, we have $p/x = x/q$ and so $x = \sqrt{pq}$. Hence, we see that VI.13 gives us $\sqrt{pq}$, the geometric mean of given lengths p and q.

Euclid VI.12. Construct a fourth proportional to three given segments. (Given segments of lengths p,q,r, then construct a segment of length x such that $p/q = r/x$.)

Euclid VI.13. Construct a mean proportional to two given segments. (Given segments of lengths p and q, then construct a segment of length x such that $p/x = x/q$.)

We are going to jump over the first nine constructions from Book IV. Read the statement below of IV.10 and try to think why you might want to solve this problem.

Euclid IV.10. Construct an isosceles triangle having base angles that are double the third angle.

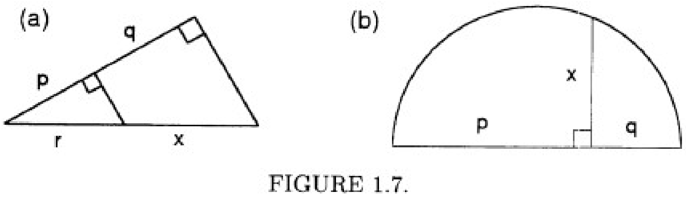

FIGURE 1.7.

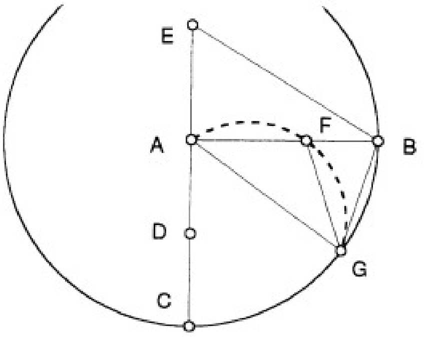

FIGURE 1.8.

Construction for IV.10. Given A_B, let the perpendicular to $\overline{AB}$ at A intersect A_B at C. Let D be the midpoint of $\overline{AC}$. Let D_B intersect $\overrightarrow{CA}$ at E, and let A_E intersect $\overline{AB}$ at F. Let B_{AF} intersect A_B at G. Then $m\angle BAG = 36$, and $m\angle ABG = m\angle AGB = 2\,m\angle BAG = 72$.

Proof. By Euclid II.11, point F is constructed on $\overline{AB}$ such that $(AB)(BF) = AF^2 = BG^2$. See Figure 1.8. Point G is constructed such that $BG = AF$. Since $BG^2 = (BF)(BA)$, then $\overleftrightarrow{BG}$ is tangent to the circumcircle of $\triangle AFG$. (Here is where we use III.37, which is given above.) Hence, $\angle BAG \cong \angle FGB$. Then, by the Exterior Angle Theorem applied to $\triangle GAF$, we have $m\angle BFG = m\angle BAG + m\angle FGA$. So $\angle BFG \cong \angle BGA$. Since $\triangle ABG$ is isosceles, then $\angle FBG \cong \angle BGA$. Hence, $\triangle GBF$ is isosceles and $GF = BG = AF$. So, $\angle FGA \cong \angle FAG$. Then $m\angle BGA = 2m\angle BAG$. Therefore, $m\angle BAG = 36$, as desired. ∎

We now know how to construct a regular decagon, since we can construct an angle of 36°. In Figure 1.8, we can march along the circumference of A_B marking off vertices of distance BG each from another. Taking alternate vertices of the regular decagon, we have the vertices of a regular pentagon. This would be a great construction for a regular pentagon if there were not an even better one that we have already done without realizing it. We will show that $\overline{BE}$ in Figure 1.8 is congruent to a side of a regular pentagon inscribed in A_B. Because $BG = AF = AE$, we already know $\overline{AE}$ is congruent to a side of a regular decagon inscribed in A_B.

Definition. Let s_n denote the length of a side of a regular n-gon inscribed in a circle having unit radius.

Consider a regular n-gon inscribed in a circle with radius 1. With the perpendicular drawn from the center to one side, it follows from elementary trigonometry that $\sin(360/2n)^\circ = (s_n/2)/1$. Hence, $s_n = 2\sin(180/n)^\circ$. We have already determined that $s_{10} = AE/AB = (-1+\sqrt{5}\,)/2$ in Figure 1.8. Since $s_{10} = 2\sin 18^\circ = 2\cos 72^\circ$, then $\cos 72^\circ = s_{10}/2 = (-1+\sqrt{5}\,)/4$. By the half-angle formula $\sin(x/2) = \sqrt{(1-\cos x)/2}$ from trigonometry, we

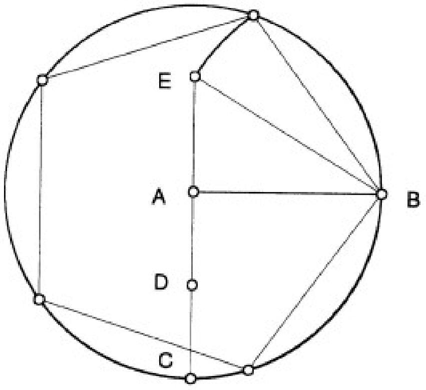

FIGURE 1.9.

have $\sin 36° = \sqrt{(1 - \cos 72°)/2}$. Therefore,

$$s_5 = 2 \sin 36° = 2\sqrt{(1 - \cos 72°)/2} = \sqrt{(5 - \sqrt{5}\,)/2}.$$

We put this together with some results that are easier to get:

Theorem. In general, $s_n = 2 \sin(180/n)°$. In particular,

$$s_3 = \sqrt{3}, \quad s_4 = \sqrt{2}, \quad s_5 = \sqrt{\frac{5 - \sqrt{5}}{2}}, \quad s_6 = 1, \quad s_{10} = \frac{-1 + \sqrt{5}}{2}.$$

You may be less than excited about finding s_5. However, the remarkable thing about our calculation of s_5 is that it is now just a trivial calculation in arithmetic to prove the following theorem of Euclid.

Euclid XIII.10.

$$s_5^2 = s_6^2 + s_{10}^2.$$

In Figure 1.8, we have $s_6 = AB/AB$, $s_{10} = AE/AB$, and $\triangle ABE$ is a right triangle. Hence, from XIII.10 and the Pythagorean Theorem, we see that $s_5 = BE/AB$ in the figure. The points F and G are used for the proof of the construction for IV.10 but are not needed for the construction of $\overline{BE}$. Therefore, as a consequence of XIII.10, we have the promised shorter construction that leads to a regular pentagon. This is stated below. See Figure 1.9, where we have $BE/AB = s_5$ and $AE/AB = s_{10}$. That is, $\overline{BE}$ is congruent to a side of a regular pentagon inscribed in A_B, and $\overline{AE}$ is congruent to a side of a regular decagon inscribed in A_B.

Construction for a regular pentagon and a regular decagon (Euclid). Let the perpendicular to $\overline{AB}$ at A intersect A_B at C. Let D be the midpoint of $\overline{AC}$. Let D_B intersect $\overrightarrow{CA}$ at E. Then $\overline{BE}$ is congruent to a side of a regular pentagon inscribed in A_B, and $\overline{AE}$ is congruent to a side of a regular decagon inscribed in A_B .

A pentadecagon is a 15-gon. We offer the following construction for a regular triangle (an equilateral triangle) inscribed in a given circle O_P. Check it out.

O_P, P_O	A_B, O_P	$\triangle ABC$
A, B	B, C	

Since we can construct an equilateral triangle and a regular pentagon with ruler and compass, it follows that we can construct angles of 60° and 36°. The existence of a ruler and compass construction for a regular pentadecagon now follows from a simple theorem of arithmetic: $360/15 = 24 = 60 - 36$.

Euclid IV.16. In a given circle, inscribe a regular pentadecagon.

If a regular n-gon can be constructed with a ruler and a compass, then a regular $2n$-gon can be constructed with a ruler and a compass. We can now construct all the regular polygons that the ancient Greeks could construct with a ruler and a compass:

The regular n-gons constructible with a ruler and a compass in antiquity. Values for n:

$$\begin{array}{l} 3, 6, 12, 24, 48, 96, \dots, \\ 4, 8, 16, 32, 64, 128, \dots, \\ 5, 10, 20, 40, 80, 160, \dots, \\ 15, 30, 60, 120, 240, 480, \dots . \end{array}$$

This is where things stood until March 29, 1796, when at age eighteen Carl Friedrich Gauss (1777–1855) created a ruler and compass construction for a regular 17-gon. Gauss's birthday was April 30 and so he was nineteen when he announced the result on the first day of June. This construction of the regular heptadecagon by the greatest mathematician since Newton (1642–1727) was published in 1801. Which regular n-gons are constructible with a ruler and a compass will be explained in the next chapter.

Herbert W. Richmond of King's College, Cambridge, gave an efficient ruler and compass construction for a regular heptadecagon in 1893 and again with more detail in 1909. We give, without proof, Richmond's construction for a regular 17-gon that is inscribed in a circle and has vertices V, V_1, V_2, V_3, … , V_{16}. It is sufficient to find adjacent vertices V_3 and V_4 in order to determine a side of the inscribed heptadecagon. See Figure 1.10.

Construction for a regular heptadecagon (H. W. Richmond). Let O_V be a unit circle with diameter $\overline{VA}$. Let $\angle BOV$ be right with $BO =$

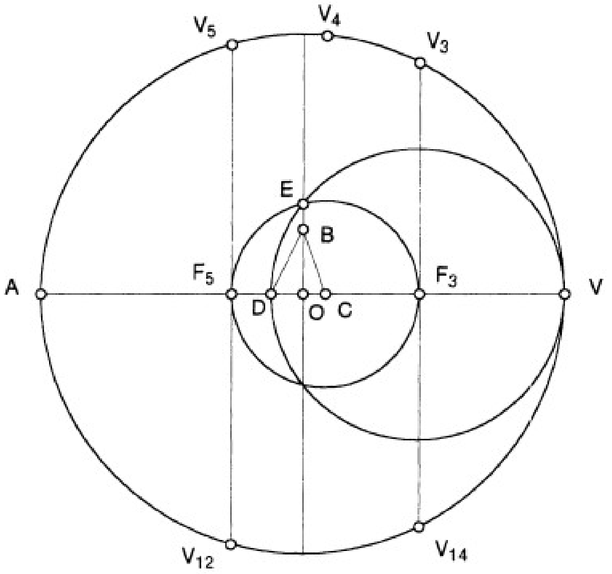

FIGURE 1.10.

$OV/4$. Let C be on $\overline{OV}$ with $m\angle OBC = (m\angle OBV)/4$. Let D be on $\overline{OA}$ with $m\angle CBD = 45$. Let the circle with diameter $\overline{DV}$ intersect $\overrightarrow{OB}$ at E. Let C_E intersect $\overrightarrow{CV}$ at F_3 and intersect $\overrightarrow{CA}$ at F_5. Let the perpendiculars to $\overleftrightarrow{VA}$ at F_3 and F_5 intersect a semicircle on $\overline{VA}$ at V_3 and V_5. Let V_4 be the intersection of O_V and the perpendicular bisector of the segment with endpoints V_3 and V_5. Then $V_3V_4 = s_{17}$.

You should do both of the following exercises now. If you can resist, do not read Exercise 1.5 until you have completed Exercise 1.4.

Exercise 1.4. Read over Richmond's construction for a regular heptadecagon until you feel familiar with it. With ruler, compass, and watch at hand, give yourself a large circle that takes up one full page. Note the time. Carry out the Richmond construction with the ruler and compass to the point of determining the seventeen vertices. Note the time when you are finished and calculate the time it took you to execute the construction. Do the construction only once, even if it "doesn't work."

Exercise 1.5. With ruler, compass, and watch at hand, give yourself a large circle that takes up one full page. Note the time. By trial and error, determine to a reasonable degree of accuracy the vertices of a regular heptadecagon inscribed in your circle. Note the time it takes to get your seventeen points. If you have not already done so, do Exercise 1.4; compare your two times and your two figures. What conclusions can you draw from these two exercises?

The rule of the game of geometric constructions that limits the tools to the ruler and the compass is called the "Platonic restriction" and predates Euclid. Although we cannot be sure what prompted Plato, this limitation does provide an elegant game. The ruler and compass constructions are constructions by means of the most basic geometric elements—the lines and the circles. However, the ruler and compass are not the only possible construction tools. We shall look at several possibilities in this book.

The one nonplatonic tool that we consider now is the *tomahawk*. The inventor of this charming construction tool is unknown. Descriptions appeared thirty years before the publication of the extensive study by Pierre-Joseph Glotin, a retired French naval officer and member of the Légion d'Honneuer from Bordeaux. Glotin's 1863 paper begins: A very ingenious pratical solution to the problem of trisecting angles employs the use of the well-known instrument named *trisector*. Glotin's ("depuis longtemps connu") trisector, which could not have been too well known, was reported in detail to the larger mathematical community in 1877 by Henri Brocard. Pierre Glotin's trisector is our tomahawk.

The essential skeleton of the tomahawk (or shoemaker's knife) is shown in Figure 1.11a, where $AB = BC = CD$, the semicircle has diameter $\overline{BD}$, and $\overrightarrow{BE} \perp \overline{BA}$. A decorative, fleshed-out version is shown in Figure 1.11b, where the use of the tomahawk is also illustrated. The tomahawk's special function is to trisect angles. The tomahawk is inserted in a given angle $\angle UVW$ so that V lies on $\overrightarrow{BE}$, point A lies on one side of the angle, and the semicircle is tangent to the other side of the angle. Now, you are to convince yourself that $\overrightarrow{VB}$ and $\overrightarrow{VC}$ do trisect the given angle.

Attaching to a tomahawk the proper rigid T, hinged at the point of the tomahawk, as indicated in Figure 1.11c, we produce an angle quinquesector that is actually a quinsector. (To quinquesect a given angle is to construct an angle having measure one-fifth the measure of the given angle; to quinsect a given angle is to divide the angle into five congruent parts. "Quinquesector" seems to have a more interesting ring to it than "quinsector" and the frequently used word "quintisector.") Attaching more T's, we would produce a construction tool that divides an angle into any odd number of congruent parts, as Glotin pointed out in 1863. An alternative method for doing this is to use several congruent tomahawks at once. For example, by attaching two congruent tomahawks together (such that at its point A one pivots about the point C of the other) as in Figure 1.11d, we can achieve the quinsection of a given angle.

Until recently, all geometers would know many of Euclid's propositions by number. Although this is no longer the case, we shall assume hereafter that the statement of each of Euclid I.2, Euclid I.3, and Euclid I.31 is known by number and we shall refer to them in this way. Euclid I.2 is important to us because the problem requires showing a modern compass and a euclidean compass are equivalent. Note that we have shown above only that they are equivalent in the presence of a ruler. We will want to find out whether the

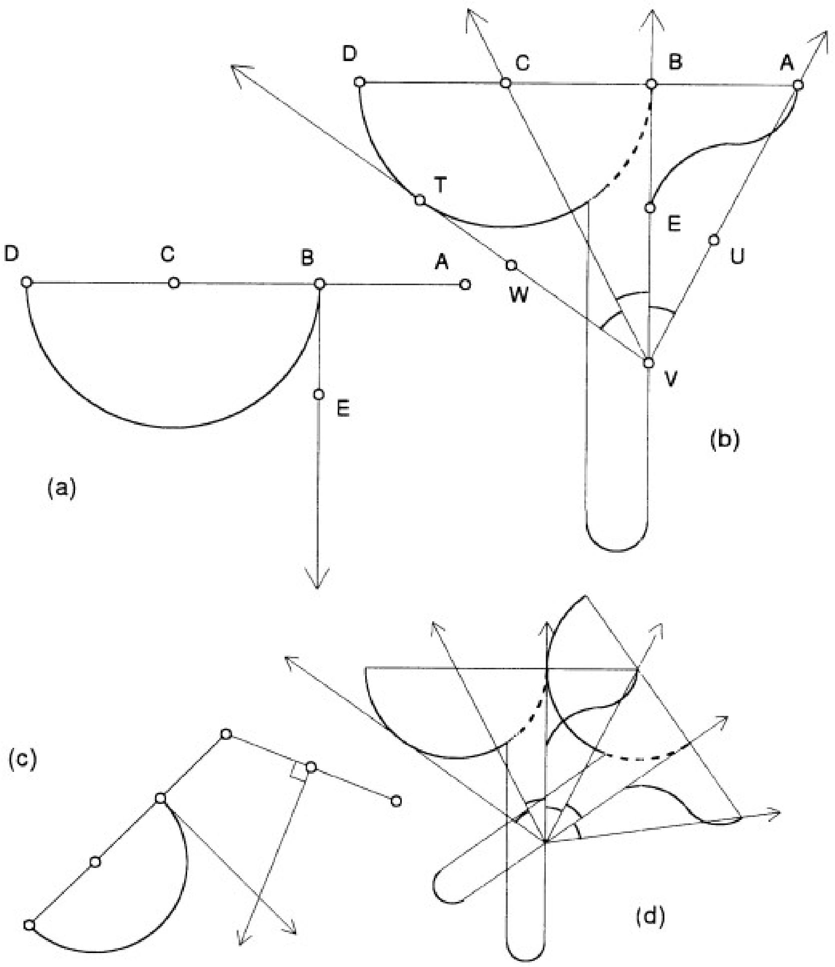

FIGURE 1.11.

ruler is necessary for this equivalence. As previously mentioned, Euclid I.3 describes what a dividers does. Later, we shall investigate what we can do with a ruler and a dividers, without a compass. Finally, Euclid I.31 is intrinsically important simply because of the importance of parallels in euclidean geometry. This proposition will turn up repeatedly and we shall want to refer to some of the particular solutions to this problem. The following catalog of constructions is by no means exhaustive. SAS, SSS, and SAA denote the familiar congruence theorems for triangles. You will be able to deduce what we are calling the Side-splitting Theorem from the fourth construction. We are constructing the line through P that is parallel to $\overleftrightarrow{AB}$, where P is off $\overleftrightarrow{AB}$.

(1) Using alternate interior angles (Euclid):

$A_P, \overleftrightarrow{AB}$	P_A, A_{PC}	$\overleftrightarrow{PD}$
C	D $\overline{DC}$ intersects $\overleftrightarrow{AP}$	

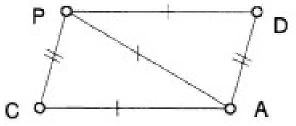

(2) Using corresponding angles:

$P_A, \overleftrightarrow{AP}$	$A_P, \overleftrightarrow{AB}$	P_C, C_{PD}	$\overleftrightarrow{PE}$
A, C	D	E $\overline{DE}$ doesn't intersect $\overleftrightarrow{AP}$	

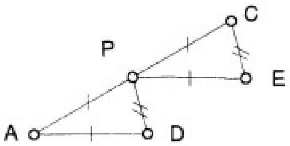

(3) Using corresponding angles and SSS:

$P_A, \overleftrightarrow{AP}$	C_{PB}, P_{AB}	$\overleftrightarrow{PD}$
A, C	D $\overline{BD}$ doesn't intersect $\overleftrightarrow{AP}$	

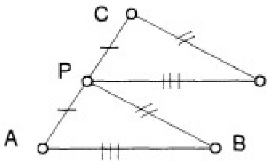

(4) Using the Side-splitting Theorem:

$A_P, \overleftrightarrow{AP}$	$B_C, \overleftrightarrow{BC}$	$\overleftrightarrow{PD}$
P, C	C, D	

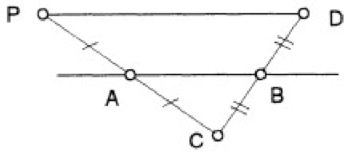

(5) Using a rhombus:

$A_P, \overleftrightarrow{AB}$	C_A, P_A	$\overleftrightarrow{PD}$
C	A, D	

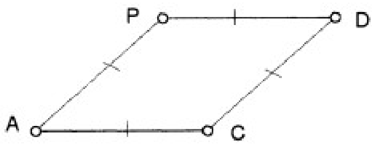

(6) $\perp$ to $\perp$, using angle bisectors of a linear pair of angles are $\perp$; take $PA \geq PB$ (so that $C \neq A$ below):

$P_A, \overleftrightarrow{AB}$	$P_A, \overleftrightarrow{AP}$	C_P, D_P	$\overleftrightarrow{PE}$
A, C	A, D	P, E	

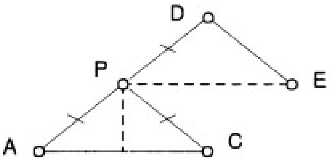

(7) $\perp$ to $\perp$, using Theorem of Thales; take $PA \geq PB$:

$A_P, \overleftrightarrow{AP}$	$C_A, \overleftrightarrow{AB}$	D_C, A_C	$\overleftrightarrow{PE}$
P, C	A, D	C, E	

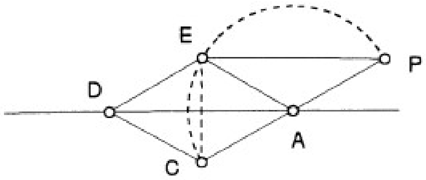

(8) Using $\perp$ to $\perp$ again; take $\perp$ bisector of $\overline{AB}$ off P:

A_B, B_A	C_P, D_P	$\overleftrightarrow{PE}$
C, D	P, E	

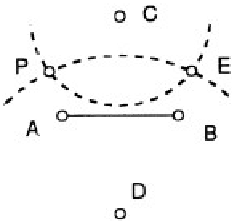

(9) Equidistant curve; take $PA \geq PB$:

$A_P, \overleftrightarrow{AB}$	A_P, D_{CP}	$\overleftrightarrow{PE}$
C, D	E $\overline{PE}$ doesn't intersect $\overleftrightarrow{AB}$	

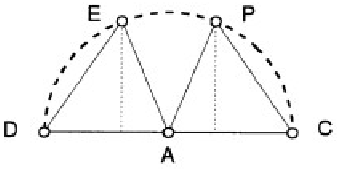

(10) Short but requires having $\perp$ bisector of $\overline{AB}$ off P:

A_{BP}, B_{AP}	$\overleftrightarrow{PC}$
C $\overline{PC}$ doesn't intersect $\overleftrightarrow{AB}$	

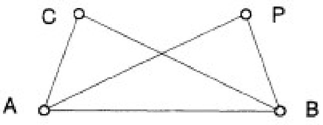

(11) Quick and dirty:

P_{AB}, B_{AP}	$\overleftrightarrow{PC}$
C $\overline{AC}$ intersects $\overleftrightarrow{BP}$	

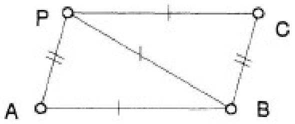

FIGURE 1.12.

One of the amazing successes of secondary education is the respect that is learned for mathematics. It is not unusual for a student who has done Exercises 1.4 and 1.5 to compare a pathetic curve that doesn't even form a polygon but was drawn with a ruler and compass to an apparently perfect regular 17-gon that was drawn by trial and error and then to conclude that the ruler and compass construction is "the better 17-gon because it is—or could be—more accurate," in spite of the presence of the obvious contradictory evidence. On the other hand, this faith does have its foundation. The pathetic curve does represent an interesting theorem, while the handsome polygon is just that, a drawing and nothing more. If we want a drawing of a regular 17-gon, we do not go to Richmond's construction for directions. Exaggeration often helps to see things more clearly. Would we seriously consider following the very lengthy directions, which do exist(!), for a ruler and compass construction of a regular 257-gon? We know beforehand that our result would most likely be way off. Our experience has taught us that the accuracy of even a very simple construction drawing is a function of how recently we sharpened our pencil. The point of all this is that we are interested in the mathematical theory suggested by the tools. If not, then we should do something else instead of studying geometric constructions. Geometric constructions is an interesting mental game. Our imperfect construction drawings can only suggest our perfect construction theorems.

Exercises

1.1. Draw sketches and write out in words the constructions for Euclid I.9 and I.10 given by the schemes below.◊

<   !"#$!   !   !% & ' ! @   >   & D   '   8'
. '   '   >   / " 6   -   /   / 0   &   '& G   H * 8 E   0-
V . '   '   >   /
" 6   -   /   /

%9   : 4   - ;;;

E   !"#$!   !   !% & ' ! @   .   +   -'   >   / ' -
G   D   - '   '   0   &   '& G   H * 8 <= 0-
V .   +   -'   >   / ' -   G   D   - '   '

%9   . '   '

%9   : 4   - ;;; 6 0   ' M   ;;; =   0'&

!"#$!   !   !% & ' ! @   .   +   -'   >   / ' -
G   D   - '   '   0   &   '& G   H * 8 E   0-

!F"." G 4   4   !) *   !   =! '00   D * !1'   )   *

!"#$!   !   !% & ' ! @   )   *5 1'   +   -'
&   /   D   - ' ' -) /' : 9'   4 * ' &   /   0 2   4 &   / F 4
9   '   8   E 0-
%9   : 4   - ;;;

!"#$!   !   !% & ' ! @   1'   )   *   -   '
' ' /   &   /   9   8 @   -24
V ?:.6F6 DG 6%G":
"G":H.6 >D% D.#H

!"#$!   !   !% & ' ! @   1'   )   *   -   '
' ' /   &   /   9   8 @   0-
%9   : 4   - ;;;

!"#$!   !   !% & ' ! @   1'   )   * C B' -' 3
+   -'   >   / G   7   ' -   9   & 8 @   0-

!F"." G 4   4   !) *   !   =! '00   D * !#

!"#$!   !   !% & ' ! @   E E   E   EW
'   0-
V .   & ' %

I& @ 8

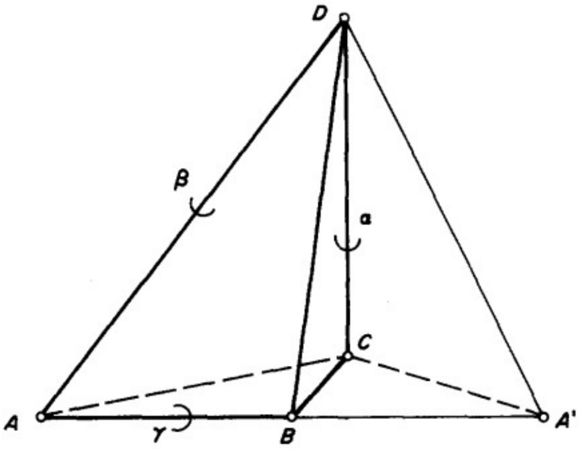

Fig. 1. How two skew lines may be perpendicular.

= !"#$! ! !% & ' ! @ @ @< @ 0-
V . ", > /
" 8 7 > + -

. ' ' ' >'3 5
5 G 8 ' 5 5 00

@ !"#$! ! !% & ' ! @ @ E @ 0-
V . ' - + -' D - ' > /
" 8 7 > + -
. ' ' ' >'3 5
5 G @ 8% 5 < 5 00 <

< !"#$! ! !% & ' ! @ - / 4' 0-
V D - ' - 7 . / # 4' F '& ' : & D - ' > /

!F"." G 4 4 !) * ! =!$ ' /

E !"#$! ! !% & ' ! @ = =(<E < @ 0-

I& 8 =

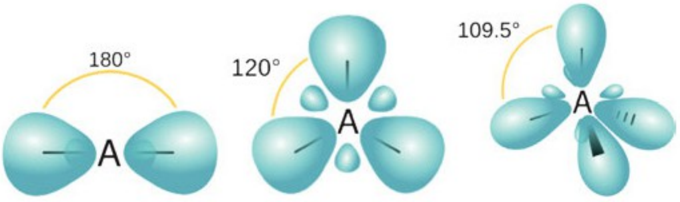

Figure 13.16: *sp*, *sp*^{2}, and *sp*^{3} hybrid orbitals for carbon. Copied on 03.02.17 from https://commons.wikimedia.org/wiki/File_AAE4h.svg, https://commons.wikimedia.org/wiki/File_AAE3h.svg and https://commons.wikimedia.org/wiki/File_AAE2h.svg. By Jfmelero (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons.

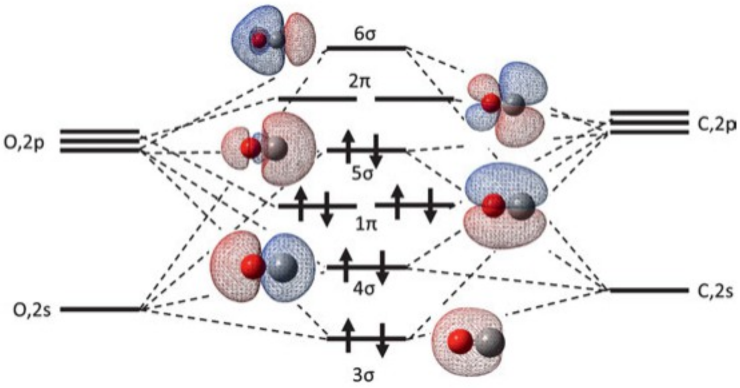

+ & < . ' ' 9 ' ' 9 ' %

I& =8 < # '

< # '
< " & ' ' - '
) - & L ' ' ' ' 5 9 / -
9 -' ' 0 0 / . / - ' ' / 9 * 8; 5
9 '/ 9 0 ' -
. L ' / ' ' & ' B - ' 0'
4 & ' ' 0' p + & < . ' & ' ' 4 '
- - '

$$\vec{l} = \vec{r} \times \vec{p}.$$

Here, $\vec{r}$ is the vector that points from the center of the circle to the particle, and $\vec{p}$ is the momentum of the particle. All three vectors are shown in Figure 7.1. It is easy to see that although $\vec{r}$ and $\vec{p}$ change directions along the circle, this is not the case for $\vec{l}$. In fact, according to classical physics, $\vec{l}$ is a constant of motion, which means that $\vec{l}$ can only be changed under the influence of external forces. Other, perhaps better known, constants of motion are energy and momentum. The fact that $\vec{l}$ is a constant of motion also explains why you can ride a bicycle and that, in principle, this becomes easier the faster you drive.

7 Rotations

7.1 Angular momentum and moment of inertia

Before discussing the quantum theoretical treatment of rotations, we will briefly describe two fundamental concepts of physics. They should actually be well known(!), but it may be useful to repeat their definitions here.

The first quantity is that of the angular momentum. We first consider a particle moving on a circular path; see Figure 7.1. The angular momentum $\vec{l}$ is a vector that is defined as

$$\vec{l} = \vec{r} \times \vec{p}. \tag{7.1}$$

Here, $\vec{r}$ is the vector that points from the center of the circle to the particle, and $\vec{p}$ is the momentum of the particle. All three vectors are shown in Figure 7.1. It is easy to see that although $\vec{r}$ and $\vec{p}$ change directions along the circle, this is not the case for $\vec{l}$. In fact, according to classical physics, $\vec{l}$ is a constant of motion, which means that $\vec{l}$ can only be changed under the influence of external forces. Other, perhaps better known, constants of motion are energy and momentum. The fact that $\vec{l}$ is a constant of motion also explains why you can ride a bicycle and that, in principle, this becomes easier the faster you drive.

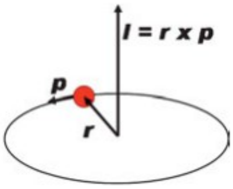

Figure 7.1: The angular momentum of a particle moving on a circular path.

Some movements are more complex and cannot be described by a simple circular path. Figure 7.2 shows an example of such a movement. Each small animal undergoes a superposition of two circular movements: on the one hand, the whole system with all the animals rotates around a common axis, and on the other hand, each cup with one animal rotates around its axis. Also, for this composite motion, there is an angular momentum vector, namely the sum of the vectors of the individual motions.

This is shown in Figure 7.3 for an example. Here, the three angular momentum vectors for three circular motions, which are not even in the same plane, are vectorially added together to form a total angular momentum vector.

There is a simple reason why we are interested in angular momenta. Electrons or atomic nuclei, which make any closed, more or less circular movements, ultimately form nothing but small circuits. Such can be influenced by magnetic fields, which is

https://doi.org/10.1515/9783110742206-007

Figure 7.2: A more complex movement composed of several circular movements.

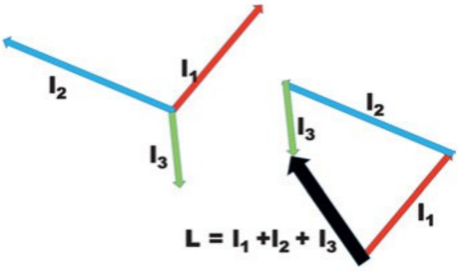

Figure 7.3: The total angular momentum, $\vec{L}$, for a motion composed of three separate rotations with angular momenta $\vec{l}_1$, $\vec{l}_2$, and $\vec{l}_3$, respectively.

what is exploited in magnetic resonance spectroscopy (NMR, ESR, EPR, for example). Because, as always, the energy is quantized (i. e., can only assume certain values), a small circuit in a magnetic field can only have a finite number of energies, whose energy differences then can be measured by means of spectroscopy. These energy levels depend not only on the applied magnetic field but also on a property of the circuits—their angular momenta. We will discuss this relationship in more detail in Section 7.5.

The second quantity that we will briefly reintroduce is the moment of inertia. Again we consider the system of Figure 7.1. The particle, with mass m, rotates about an axis (containing the angular momentum vector) at a distance r to this. The moment of inertia around this axis is defined as

$$I = mr^2. \tag{7.2}$$

If we have several particles that can have different masses and distances to the axis, the total moment of inertia is given by

$$I = \sum_i m_i r_i^2. \tag{7.3}$$

Here, m_i is the mass of the ith particle and r_i its distance to the axis. We emphasize that the moment of inertia makes sense only if you define the associated axis.

Equation (7.3) can be generalized so that also larger bodies can be treated. We let $\rho(\vec{s})$ be the density of the body at the point $\vec{s}$ and then have in general the analogy to above,

$$I = \int \rho(\vec{s}) r^2(\vec{s})\, d\vec{s}. \tag{7.4}$$

$r(\vec{s})$ is the distance from the point $\vec{s}$ to the axis.

7.2 2D rotor

Before we present the correct quantum mechanical treatment of the two-dimensional rotor, we discuss a simplified treatment based mainly on classical arguments.

We consider a system as in Figure 7.1, a particle moving at a constant angular velocity on a circular path. Our arguments, however, are not limited to this example, but the result will be generally valid if we consider a body with a given moment of inertia rotating about an axis with constant angular velocity. But let us start with the system of Figure 7.1.

The total energy of the particle is equal to the kinetic energy because we assume that no forces are acting on the system. So

$$E = E_{\text{kin}} = \frac{p^2}{2m}. \tag{7.5}$$

At the same time, the magnitude of the angular momentum equals

$$l = pr. \tag{7.6}$$

We have set the mass and momentum of the particle equal to m and p, respectively, while r is the radius of the circular orbit.

By combining equations (7.5) and (7.6), we get

$$E = \frac{l^2}{2mr^2} = \frac{l^2}{2I}. \tag{7.7}$$

Here is

$$I = mr^2 \tag{7.8}$$

the moment of inertia of the system.

At this point, we introduce the de Broglie relation, which provides a link between the wave and particle nature of quantum mechanical objects,

$$p = \frac{h}{\lambda}. \tag{7.9}$$

λ is the wavelength of the system if the system is considered as a quantum mechanical wave. This wave is accordingly located on the circular path. Requiring that the wave does not annihilate itself (in other words, requiring that there is a constructive interference), a whole number of waves must fit on the circular path. Accordingly, the length of this path, $2\pi r$, must be equal to an integral multiple of λ,

$$2\pi r = n\lambda, \tag{7.10}$$

with $n = 1, 2, 3, \ldots$.

From equations (7.6), (7.9), and (7.10), we obtain then

$$l = pr = \frac{hr}{\lambda} = \frac{hr}{2\pi r/n} = n \cdot \frac{h}{2\pi} = n \cdot \hbar, \tag{7.11}$$

and according to equation (7.7) the energy becomes then

$$E = \frac{\hbar^2}{2I} n^2. \tag{7.12}$$

This means that both the energy and the magnitude of the angular momentum are quantized.

We will now show that this result is also obtained by a “correct” quantum theoretical treatment. For this, we assume that the particle is moving in the (x, y) plane and that it moves around the origin of the coordinate system. The two-dimensional Schrödinger equation then becomes

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\psi = E \cdot \psi. \tag{7.13}$$

Actually, we only have one motional degree of freedom because

$$x^2 + y^2 = r^2 \tag{7.14}$$

is constant. Therefore, equation (7.14) is less useful. Instead, we introduce polar coordinates in two dimensions,

$$\begin{aligned} x &= r \cdot \cos\phi \\ y &= r \cdot \sin\phi. \end{aligned} \tag{7.15}$$

If the above expressions are used in equation (7.14), we need expressions like

$$\frac{\partial}{\partial x} = \frac{\partial}{\partial r} \cdot \frac{\partial r}{\partial x} + \frac{\partial}{\partial \phi} \cdot \frac{\partial \phi}{\partial x} \tag{7.16}$$

and

$$\frac{\partial^2}{\partial x^2} = \frac{\partial^2}{\partial r^2} \cdot \left(\frac{\partial r}{\partial x}\right)^2 + \frac{\partial}{\partial r} \cdot \frac{\partial^2 r}{\partial x^2} + \frac{\partial^2}{\partial \phi^2} \cdot \left(\frac{\partial \phi}{\partial x}\right)^2 + \frac{\partial}{\partial \phi} \cdot \frac{\partial^2 \phi}{\partial x^2} \tag{7.17}$$

and the similar expressions for the derivatives with respect to y. If you do not make mistakes (or you just look up in a mathematical collection of formulas where you can easily find those formulas), you get

$$\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}=\frac{\partial^2}{\partial r^2}+\frac{1}{r}\frac{\partial}{\partial r}+\frac{1}{r^2}\frac{\partial^2}{\partial \phi^2}. \tag{7.18}$$

In our case, the whole becomes particularly simple: because r is constant, all differentiations with respect to r can be ignored (r can not be varied), so that the Schrödinger equation (7.13) can be written as

$$-\frac{\hbar^2}{2m}\frac{1}{r^2}\frac{\partial^2}{\partial \phi^2}\psi(\phi)=E\psi(\phi) \tag{7.19}$$

Here, we have also taken advantage of the fact that the wave function depends only on the angle ϕ.

Equation (7.19) can be rewritten as

$$-\frac{\hbar^2}{2I}\frac{\partial^2}{\partial \phi^2}\psi(\phi)=E\psi(\phi). \tag{7.20}$$

This equation has the solutions

$$\psi(\phi)=A\cdot e^{im_l\phi} \tag{7.21}$$

with A equal to a constant and

$$m_l=\pm\sqrt{\frac{2IE}{\hbar^2}}. \tag{7.22}$$

We must have (see Figure 7.4) that

$$\psi(\phi+2\pi)=\psi(\phi), \tag{7.23}$$

(this is called cyclic boundary conditions), from which we get that m_l must be integral. Then equation (7.23) becomes

$$E=\frac{\hbar^2 m_l^2}{2I},\quad m_l=0,\pm1,\pm2,\dots, \tag{7.24}$$

i. e., just as in equation (7.12).

The angular momentum vector in this case lies along the z axis, so that its x and y components vanish, and the length of the angular momentum vector becomes equal to the absolute value of its z component. The quantum mechanical operator of its z component is

$$\hat{l}_z=\hat{x}\cdot\hat{p}_y-\hat{y}\cdot\hat{p}_x=\frac{\hbar}{i}\left(x\frac{\partial}{\partial y}-y\frac{\partial}{\partial x}\right)=\frac{\hbar}{i}\frac{\partial}{\partial \phi} \tag{7.25}$$

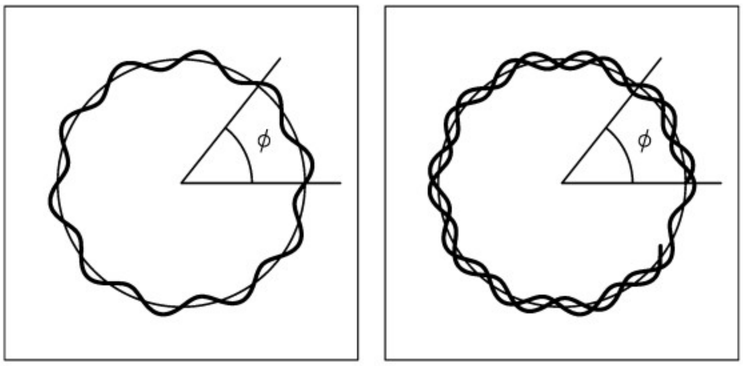

Figure 7.4: The cyclic boundary conditions for the two-dimensional rotor. The picture shows the imaginary part of $e^{im\phi}$, i. e. $\sin(im\phi)$, for (left) m integral and (right) m nonintegral. The thicker curve shows the function value relative to 0 (shown as the thinner curve) as a function of the angle ϕ. It can be seen how only in the left panel the function is mapped onto itself, when ϕ is increased by 2π.

by expressing it in terms of polar coordinates. For the functions of equation (7.21), we obtain immediately,

$$\hat{l}_z\psi(\phi) = \frac{\hbar}{i}\frac{\partial}{\partial\phi}(A \cdot e^{im_l\phi}) = \hbar m_l A \cdot e^{im_l\phi} = \hbar m_l \psi(\phi), \tag{7.26}$$

i. e., the wave function is also an eigenfunction to $\hat{l}_z$ with the eigenvalue $m_l\hbar$.

Finally, we determine the normalization constant A of the wave function. Because the wave function is to be normalized over its entire domain (i. e., ϕ from 0 to 2π), we have

$$\int_0^{2\pi} |\psi(\phi)|^2 \, d\phi = |A|^2 \cdot 2\pi \equiv 1, \tag{7.27}$$

where we choose

$$A = \frac{1}{\sqrt{2\pi}}. \tag{7.28}$$

In Figure 7.5, the real and imaginary parts of the wave functions are shown for some of the smallest, nonnegative values of m_l. This figure also shows the angular momentum vectors.

7.3 3D rotor

For the two-dimensional rotor, we considered a particle that could move in a two-dimensional plane, with the restriction that the distance to a certain point, which

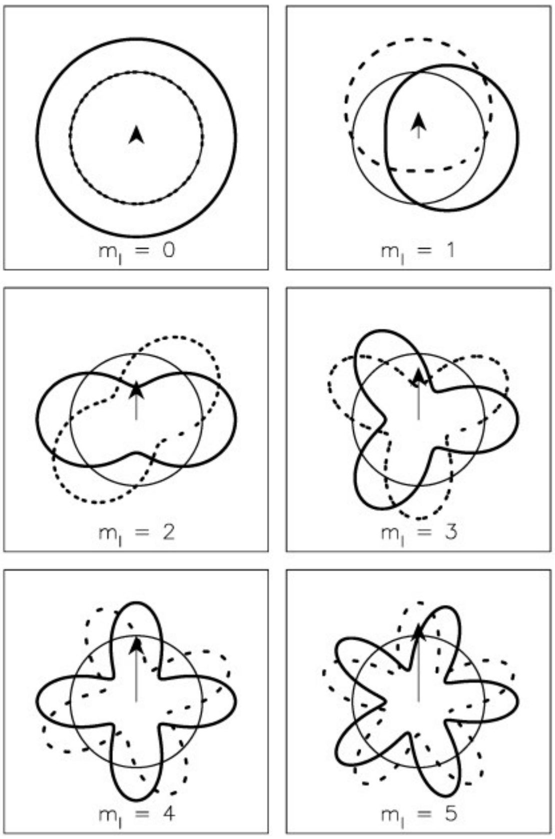

Figure 7.5: The six energetic lowest states for the two-dimensional rotor. The thick solid curves represent the real parts of the wave functions, and the thick, dashed curves represent the imaginary parts. The thin curves represent 0, and the arrows in the center show the angular momentum vectors, which are supposed to be perpendicular to the plane of the images.

we later chose as the origin of the coordinate system, is constant. For the three-dimensional rotor, we proceed in a similar way. This time, the particle can move in a three-dimensional space with the restriction that the distance to a certain point, which in turn we will choose as the origin of the coordinate system, must be constant. In this case, we will present only the quantum theoretical treatment.

As with the two-dimensional rotor, it is again convenient to use a suitable coordinate system. This time, we will use the polar coordinates in three dimensions, also called spherical coordinates. Their definition is shown in Figure 7.6.

We have

$$\begin{aligned} x &= r\sin\theta\cos\phi \\ y &= r\sin\theta\sin\phi \\ z &= r\cos\theta. \end{aligned} \tag{7.29}$$

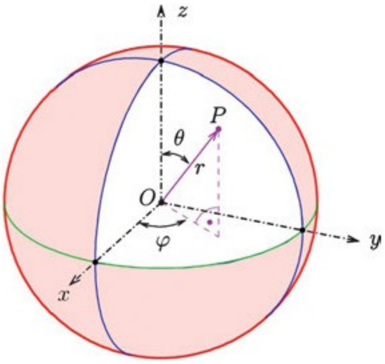

Figure 7.6: The definition of polar coordinates in three dimensions, also called spherical coordinates.

Also, for these coordinates, we need the expression of the Laplace operator [see equation (2.4)]

$$\begin{aligned}\nabla^2 &= \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \\ &= \frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r} + \frac{1}{r^2 \sin^2\theta}\frac{\partial^2}{\partial \phi^2} + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\sin\theta\frac{\partial}{\partial \theta} \\ &\equiv \frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\hat{\Lambda}^2. \end{aligned} \tag{7.30}$$

We have here introduced the operator

$$\hat{\Lambda}^2 = \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial \phi^2} + \frac{1}{\sin\theta}\frac{\partial}{\partial \theta}\sin\theta\frac{\partial}{\partial \theta} \tag{7.31}$$

that contains all differentiations with respect to the two angles ϕ and θ.

For the three-dimensional rotor, r is constant, so that in the Schrödinger equation all terms with derivatives with respect to r can be ignored. Then the Schrödinger equation becomes

$$-\frac{\hbar^2}{2mr^2}\hat{\Lambda}^2\psi(\theta,\phi) = E\psi(\theta,\phi), \tag{7.32}$$

where we have explicitly stated that the wave function ψ depends on the two angles θ and ϕ, but not on r. Equation (7.32) can also be written as

$$\hat{\Lambda}^2\psi(\theta,\phi) = -\frac{2IE}{\hbar^2}\psi(\theta,\phi), \tag{7.33}$$

where I is the moment of inertia about the origin of the coordinate system (in contrast to the previous section, not about an axis but about a point, because the particle is moving about a single point).

In Sections 2.2 and 4.11, we have seen that it can be helpful to use a product Ansatz in solving differential equations that depend on multiple variables. Thus, we set

$$\psi(\theta,\phi) = \Theta(\theta)\cdot\Phi(\phi). \tag{7.34}$$

By inserting this into equation (7.33), we get

$$\hat{\Lambda}^2\psi = \frac{1}{\sin^2\theta}\Theta\Phi'' + \Phi\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}[\sin\theta\Theta'] = -\frac{2IE}{\hbar^2}\Theta\Phi. \tag{7.35}$$

By dividing by the product $\Theta\Phi$, it follows

$$\frac{\Phi''}{\Phi} + \frac{1}{\Theta}\sin\theta\frac{\partial}{\partial\theta}[\sin\theta\Theta'] = -\frac{2IE}{\hbar^2}\sin^2\theta \tag{7.36}$$

or

$$\frac{\Phi''}{\Phi} = -\frac{1}{\Theta}\sin\theta\frac{\partial}{\partial\theta}[\sin\theta\Theta'] - \frac{2IE}{\hbar^2}\sin^2\theta = -m_l^2. \tag{7.37}$$

In the last identity, we have exploited that the left-hand side depends only on ϕ, while the expression in the middle depends only on θ. This means that they can only be identical if they are a constant. We call this constant $-m_l^2$, for reasons that will become clear soon.

From equation (7.37), we obtain two equations: one for Θ and one for Φ. The latter is

$$\frac{\Phi''}{\Phi} = -m_l^2. \tag{7.38}$$

This has the solution

$$\Phi(\phi) = Ae^{im_l\phi}, \tag{7.39}$$

whereby m_l can be positive, zero, or negative.

As for the two-dimensional rotor, cyclic boundary conditions also apply here:

$$\Phi(\phi + 2\pi) = \Phi(\phi) \tag{7.40}$$

where

$$\Phi(\phi) = \frac{1}{\sqrt{2\pi}}e^{im_l\phi}. \tag{7.41}$$

m_l is accordingly equal to an integer. Furthermore, we have chosen A in equation (7.39) such that $\Phi(\phi)$ is normalized,

$$\int_0^{2\pi} |\Phi(\phi)|^2\, d\phi = |A|^2\cdot 2\pi \equiv 1. \tag{7.42}$$

The other equation, which is given in equation (7.37), is the equation for $\Theta(\theta)$. This equation is

$$\sin\theta\frac{\partial}{\partial\theta}[\sin\theta\Theta'] + \left[\frac{2IE}{\hbar^2}\sin^2\theta - m_l^2\right]\Theta = 0. \tag{7.43}$$

This equation is not easy to solve. It is called Legendre's equation, and fortunately, the solutions to the equation are known. The solutions can be expressed using the so-called Legendre polynomials (after Adrien-Marie Legendre). These are given by

$$P_l(z) = \frac{1}{2^l l!}\frac{d^l}{dz^l}(z^2 - 1)^l. \tag{7.44}$$

We see that here we have introduced a second number l (besides m_l). l can take the values $0, 1, 2, \ldots$. Using the Legendre polynomials, the (normalized) solutions to equation (7.43) are then expressed as

$$\Theta(\theta) = S_{lm_l}(\theta) = (-1)^{m_l}\sqrt{\frac{2l+1}{2}}\sqrt{\frac{(l-m_l)!}{(l+m_l)!}}\sin^{m_l}(\theta)\frac{d^{m_l}P_l(\cos(\theta))}{(d\cos(\theta))^{m_l}}. \tag{7.45}$$

One often introduces the so-called associated Legendre functions

$$P_l^{m_l}(z) = (1 - z^2)^{m_l/2}\frac{d^{m_l}}{dz^{m_l}}P_l(z). \tag{7.46}$$

Then the wave functions ψ in equation (7.47) become

$$\begin{aligned}\psi(\theta,\phi) &= \Theta(\theta)\cdot\Phi(\phi) \equiv Y_{lm_l}(\theta,\phi)\\ &= \begin{cases}(-1)^{m_l}\sqrt{\frac{2l+1}{2}}\sqrt{\frac{(l-m_l)!}{(l+m_l)!}}P_l^{m_l}(\cos\theta)\frac{e^{im_l\phi}}{\sqrt{2\pi}} & m_l \geq 0\\ (-1)^{m_l}Y^*_{l,-m_l}(\theta,\phi) & m_l < 0.\end{cases}\end{aligned} \tag{7.47}$$

It turns out that there are solutions for only certain values of (l, m_l):

$$\begin{aligned}l &= 0, 1, 2, \ldots\\ m_l &= -l, -l+1, -l+2, \ldots, l-2, l-1, l.\end{aligned} \tag{7.48}$$

Moreover, the Y functions are called **spherical harmonic functions** or just **spherical harmonics**.

The mathematical expressions of the spherical harmonics for $l = 0$, 1, 2, and 3 are shown in Table 7.1.

The solutions to the Schrödinger equation, equation (7.32), can therefore be expressed in terms of the spherical harmonics. This equation is a two-dimensional differential equation (we have two coordinates: θ and ϕ), and so their solution leads to the introduction of two quantum numbers, which we have labeled l and m_l, and which

Table 7.1: The spherical harmonic functions for $l = 0, 1, 2$, and 3.

l	m_l	$Y_{lm_l}(\theta, \phi)$
0	0	$\sqrt{\frac{1}{4\pi}}$
1	1	$-\sqrt{\frac{3}{8\pi}} \sin\theta e^{i\phi}$
	0	$\sqrt{\frac{3}{4\pi}} \cos\theta$
	−1	$\sqrt{\frac{3}{8\pi}} \sin\theta e^{-i\phi}$
2	2	$\sqrt{\frac{15}{32\pi}} \sin^2\theta e^{2i\phi}$
	1	$-\sqrt{\frac{15}{8\pi}} \sin\theta \cos\theta e^{i\phi}$
	0	$\sqrt{\frac{5}{16\pi}} (3\cos^2\theta - 1)$
	−1	$\sqrt{\frac{15}{8\pi}} \sin\theta \cos\theta e^{-i\phi}$
	−2	$\sqrt{\frac{15}{32\pi}} \sin^2\theta e^{-2i\phi}$
3	3	$-\frac{1}{8}\sqrt{\frac{35}{\pi}} \sin^3\theta e^{3i\phi}$
	2	$\frac{1}{4}\sqrt{\frac{105}{2\pi}} \cos\theta \sin^2\theta e^{2i\phi}$
	1	$-\frac{1}{8}\sqrt{\frac{21}{\pi}} (5\cos^2\theta - 1) \sin\theta e^{i\phi}$
	0	$\frac{1}{4}\sqrt{\frac{7}{\pi}} (5\cos^2\theta - 3) \cos\theta$
	−1	$\frac{1}{8}\sqrt{\frac{21}{\pi}} (5\cos^2\theta - 1) \sin\theta e^{-i\phi}$
	−2	$\frac{1}{4}\sqrt{\frac{105}{2\pi}} \cos\theta \sin^2\theta e^{-2i\phi}$
	−3	$\frac{1}{8}\sqrt{\frac{35}{\pi}} \sin^3\theta e^{-3i\phi}$

can only assume certain values. This applies in general: a Schrödinger equation in P dimensions leads to P quantum numbers that can only assume certain values.

The energy we obtain is

$$E = \frac{l(l+1)\hbar^2}{2I}. \tag{7.49}$$

We see that E depends only on l but not on m_l. This means that all $2l + 1$ wave functions Y_{lm_l} with the same l but the different m_l values lead to the same energy: they are energetically degenerate.

At the end of this chapter, we discuss the angular momentum. In Section 7.1, we saw that the angular momentum can be written as

$$\vec{l} = (l_x, l_y, l_z) = \vec{r} \times \vec{p} = (yp_z - zp_y, zp_x - xp_z, xp_y - yp_x). \tag{7.50}$$

In Section 7.2, we also introduced the quantum mechanical operator for l_z,

$$\hat{l}_z = \frac{\hbar}{i}\frac{\partial}{\partial\phi} \tag{7.51}$$

It can also be shown that the operator for the square of the length of the angular momentum vector becomes

$$\hat{l}^2 = \hat{l}_x\hat{l}_x + \hat{l}_y\hat{l}_y + \hat{l}_z\hat{l}_z = -\hbar^2\hat{\Lambda}^2. \tag{7.52}$$

Except for a factor $2I$, $\hat{l}^2$ is therefore identical to the Hamilton operator so that

$$\hat{l}^2 Y_{lm_l} = l(l+1)\hbar^2 Y_{lm_l}. \tag{7.53}$$

Simultaneously, it is easily shown by inserting that

$$\hat{l}_z Y_{lm_l} = m_l\hbar Y_{lm_l}. \tag{7.54}$$

The spherical harmonics are thus eigenfunctions both for the operator for the square of the length of the angular momentum vector and for the operator for its z component. At the same time, the second identify in equation (7.48) must be satisfied, which we write as

$$|m_l\hbar| \leq l\hbar \leq \sqrt{l(l+1)}\hbar \tag{7.55}$$

(and the "=" in the second identity applies only to $l = m_l = 0$). This means that the z component of the angular momentum vector is smaller than the length of the angular momentum vector, i. e., the vector cannot lie along the z axis. This result for the angular momentum can be interpreted as shown in Figure 7.7.

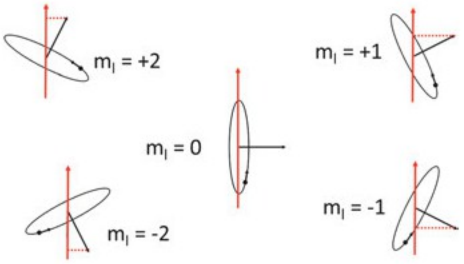

Figure 7.7: The five allowed orientations of the angular momentum vector for $l = 2$. The normal (which is parallel to the angular momentum vector) of the circular orbit of the rotating particle makes a fixed angle to the z axis (shown in red).

Thus, our results give that both the length of the angular momentum and its z component are quantized. But we know nothing about the x and y components, which in principle can take any values. For given l and m_l, the angular momentum vector can therefore lie at any point on one of the circles shown in the right part of Figure 7.8. The vector can even follow a path on this circle. One speaks then of precession.

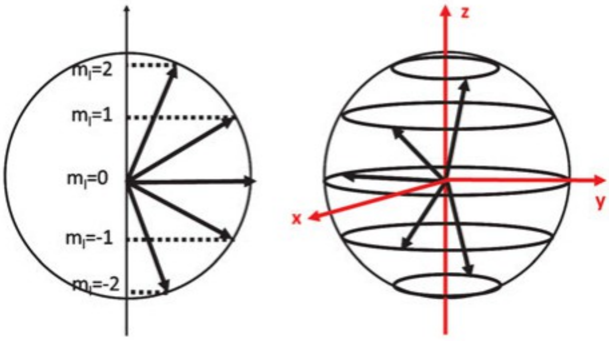

Figure 7.8: The allowed orientations of the angular momentum vector for $l = 2$. As in Figure 7.7, the normal (which is parallel to the angular momentum vector) of the circular orbit of the rotating particle makes a fixed angle to the z axis (shown in red). But because the angles to the other two directions are not fixed, all points on the circles in the right part of the figure are allowed directions of the angular momentum vector.

Precisely because of this precession, the x and y components of the angular momentum vector have no unique values. Would you try, e. g., to measure the x component after determining the z component, the precession will rotate and then take place around the x axis. This means that now, although the x component of the angular momentum vector is known, you lose any information about the z component. Ultimately, this is a consequence of the fact that the three components of the angular momentum vector do not commute,

$$\begin{aligned}[\hat{l}_x, \hat{l}_y] &= i\hbar\hat{l}_z \\ [\hat{l}_y, \hat{l}_z] &= i\hbar\hat{l}_x \\ [\hat{l}_z, \hat{l}_x] &= i\hbar\hat{l}_y, \end{aligned} \tag{7.56}$$

so that the exact determination of more components of the angular momentum vector is not possible [cf. equation (3.71)].

From equation (7.47), it is easy to see that $|Y_{lm_l}(\theta,\phi)|^2$ is independent of ϕ, and that

$$|Y_{lm_l}(\theta,\phi)|^2 = |Y_{l,-m_l}(\theta,\phi)|^2. \tag{7.57}$$

In order to depict the spherical harmonics $|Y_{lm_l}(\theta,\phi)|^2$, one can proceed as follows. Y_{lm_l} is a function of the two angles θ and ϕ, which describe a direction in the three-dimensional coordinate system. Therefore, one chooses a direction (i. e., values for θ and ϕ), computes $|Y_{lm_l}(\theta,\phi)|^2$, and marks a point with this value along the direction (θ,ϕ). Having done this for all values of θ and ϕ, one has generated a surface in three-dimensional space that can be depicted, as done in Figure 7.9.

The spherical harmonics we have introduced are generally complex, meaning that they take complex (rather than purely real) values. From these spherical harmonics,

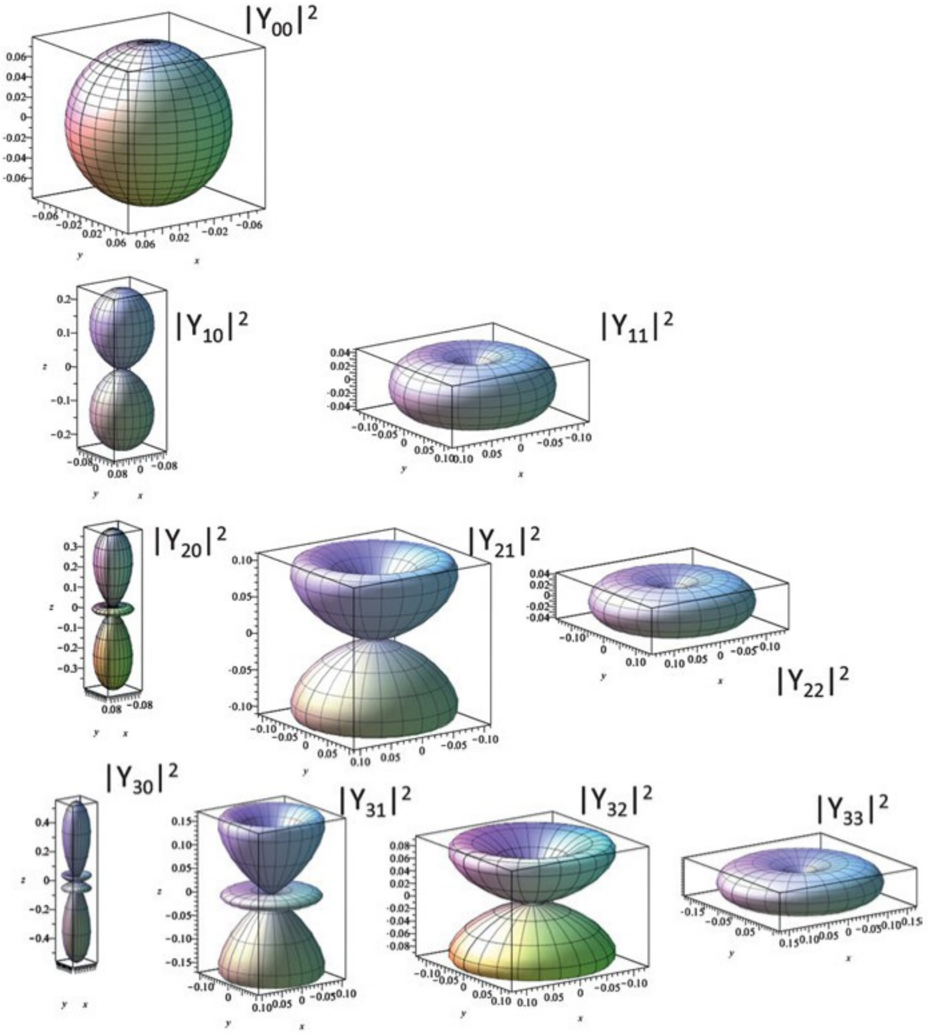

Figure 7.9: Representation of the squared absolute values of some of the complex spherical harmonics.

however, the more well-known s, p_x, … functions can be constructed that take only real values. In general, one creates two new functions from Y_{l,m_l} and $Y_{l,-m_l}$ for $m_l \neq 0$,

$$Y_{l,m_l,\pm} = \frac{c_{l,m_l,\pm}}{\sqrt{2}}(Y_{l,m_l} \pm Y_{l,-m_l}), \tag{7.58}$$

while the functions $Y_{l,0}$ remain unchanged. By judicious choice of the constants $c_{l,m_l,\pm}$ (e. g., $c_{l,m_l,\pm} = \pm 1$ or $\pm i$), it can be achieved that the functions $Y_{l,m_l,\pm}$ are real (rather than complex). For example, p_x and p_y are formed from $Y_{1,1}$ and $Y_{1,-1}$, while $d_{x^2-y^2}$ and d_{xy} are formed from $Y_{2,2}$ and $Y_{2,-2}$. These functions are shown in Figure 7.10. Also, these

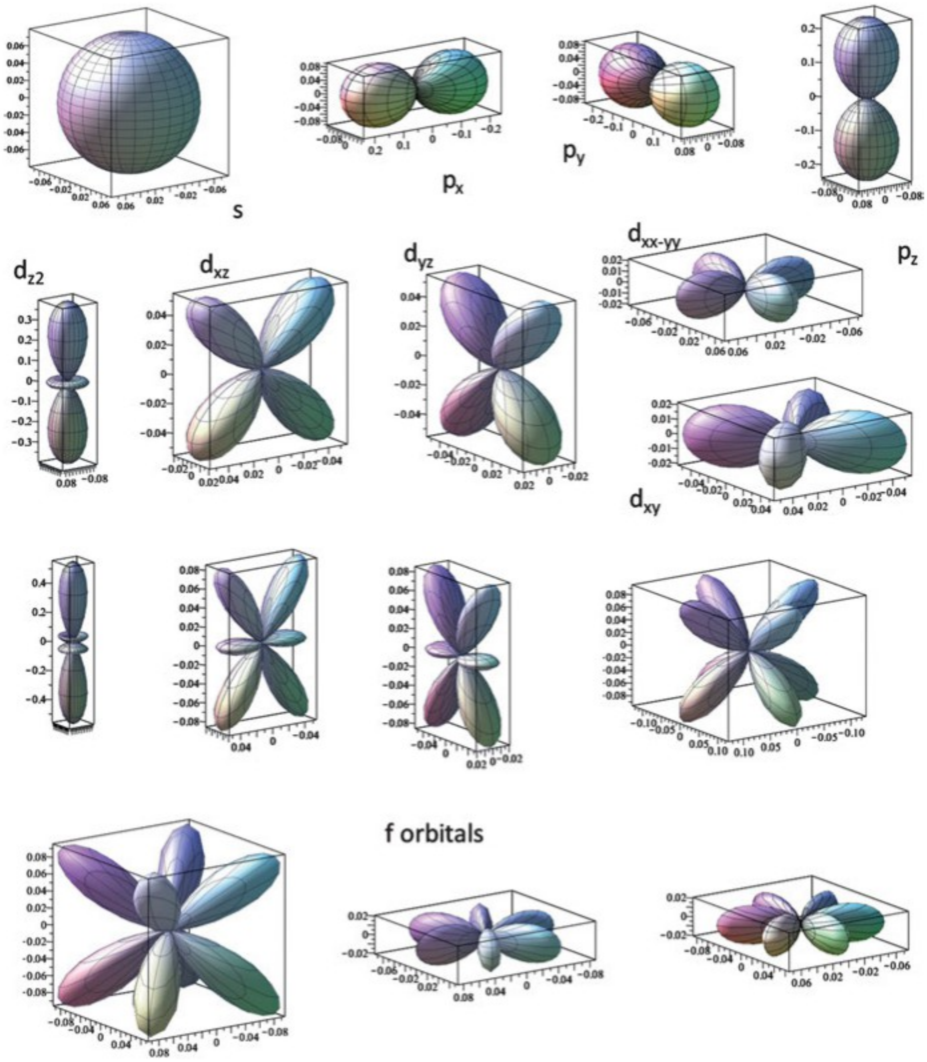

Figure 7.10: Representation of the real spherical harmonics.

functions are eigenfunctions for the Hamilton operator as well as for $\hat{l}^2$ but only for $m_l = 0$ also for $\hat{l}_z$.

That also the real spherical harmonics are eigenfunctions to $\hat{H}$ can be seen as follows. We consider two functions, ψ_1 and ψ_2, that both are eigenfunctions to some (linear) Hamilton operator $\hat{H}$ with the same eigenvalue,

$$\begin{aligned}\hat{H}\psi_1 &= E\psi_1\\ \hat{H}\psi_2 &= E\psi_2.\end{aligned} \tag{7.59}$$

Then, for any linear combination of ψ_1 and ψ_2 we find

$$\hat{H}(c_1\psi_1 + c_2\psi_2) = c_1\hat{H}\psi_1 + c_2\psi_2 = c_1E\psi_1 + c_2E\psi_2 = E(c_1\psi_1 + c_2\psi_2), \tag{7.60}$$

i. e., this is also an eigenfunction to $\hat{H}$. In the present case where we combine Y_{l,m_l} and $Y_{l,-m_l}$ for $m_l \neq 0$, similar arguments can be used to show that the resulting linear combination is also an eigenfunction to $\hat{l}^2$.

7.4 Addition of angular momenta

For systems with several angular momenta, they are added vectorially both in the classical case and in the quantum mechanical case (see Figure 7.3). Then, in the quantum mechanical case, too, the length of the total angular momentum and its z component are quantized. In the general case, however, this does not apply to the individual angular momentum vectors, but only to their sum, if the individual motions influence each other. For example, electrons of an atom are charged and repel each other, so that the motion of each electron depends on those of the others.

For us, it is often the case that the angular momentum will be caused by electrons moving in some orbits. The motion of a single electron generates an angular momentum vector, and from the individual ones a total angular momentum vector is then formed. Then it may happen that the sum disappears.

In the general case, we have that if

$$\vec{L} = \vec{l}_1 + \vec{l}_2 + \cdots + \vec{l}_n \tag{7.61}$$

we have for the corresponding quantum mechanical operator [where $\Psi(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n)$ is the wave function of the total system of the n particles]

$$\hat{L}^2\Psi(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n) = L(L+1)\hbar^2\Psi(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n) \tag{7.62}$$

with L equal to an integer. $L = 0$ is then possible. Furthermore,

$$\hat{L}_z\Psi(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n) = M_L\hbar\Psi(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n). \tag{7.63}$$

The integer M_L satisfies

$$-L \leq M_L \leq L. \tag{7.64}$$

Similar formulas do not apply for the operators of the individual angular momenta, $\vec{l}_1$, $\vec{l}_2, \ldots, \vec{l}_n$.

As an example, we consider two electrons in one atom. If we first assume that their motions are independent of each other, we can calculate the total angular momentum in equation (7.61) obtained by adding the angular momenta of the individual electrons,

$$\vec{L} = \vec{l}_1 + \vec{l}_2, \tag{7.65}$$

whereby we initially assume that the individual angular momentum vectors are quantized,

$$\begin{aligned}|\vec{l}_1|^2 &= l_1(l_1+1)\hbar^2\\ |\vec{l}_2|^2 &= l_2(l_2+1)\hbar^2.\end{aligned} \tag{7.66}$$

If subsequently the motions of the individual electrons are allowed to influence each other, only the sum will be quantized,

$$|\vec{L}|^2 = L(L+1)\hbar^2, \tag{7.67}$$

and it turns out that the possible values of L are

$$L = |l_1 - l_2|, |l_1 - l_2| + 1, \dots, l_1 + l_2 - 1, l_1 + l_2. \tag{7.68}$$

Specifically, we first have assumed that the two electrons are in the two (different) orbitals ψ_1 and ψ_2, for which

$$\begin{aligned}\hat{l}_1^2\psi_1 &= l_1(l_1+1)\hbar^2\psi_1\\ \hat{l}_2^2\psi_2 &= l_2(l_2+1)\hbar^2\psi_2.\end{aligned} \tag{7.69}$$

Subsequently, the total wave function of the two electrons Ψ is not an eigenfunction to the operators $\hat{l}_1^2$ and $\hat{l}_2^2$, but only to $\hat{L}^2$.

Finally, we mention that we have repeatedly discussed the z component of the angular momentum vector that is quantized while the x and y components are not. It is said that the z direction is the quantization direction. But there is no coordinate system in normal everyday life, and then the choice of the z direction is actually arbitrary. However, there may be experimental conditions that define a “special” or “preferred” direction. This is the case, e. g., when the system is in a magnetic field (a case we will discuss in the next section). Then the direction of the magnetic field is “special” or “preferred” and it is then often convenient to choose this as the z direction.

7.5 Spin of the electron

In 1922, the physicists Otto Stern and Walther Gerlach carried out an experiment in Frankfurt am Main, the results of which led to the discovery of the electron’s spin. However, the exact existence of the spin was not suggested before 1925 and then by Samuel Abraham Goudsmit and George Eugene Uhlenbeck.

So far, we have considered small bodies that can move in more or less circular orbits. When these bodies carry a charge, this results in a small circuit. Such a circuit corresponds to a magnetic moment, $\vec{\mu}$, which is proportional to the angular momentum of the body,

$$\vec{\mu} \propto \vec{l}. \tag{7.70}$$

In a magnetic field with field vector $\vec{B}$, the energy of the magnetic moment equals

$$E_{\text{magn}} = \vec{\mu} \cdot \vec{B}. \tag{7.71}$$

We now place such small circuits in an inhomogeneous magnetic field as in Figure 7.11. Those for which $\vec{\mu}$ is parallel to $\vec{B}$ can reduce their energy by moving in the direction in which $\vec{B}$ becomes smaller. The reverse applies to those for which $\vec{\mu}$ is antiparallel to $\vec{B}$. Overall, the force acting on a circuit depends on the angle between $\vec{\mu}$ and $\vec{B}$, assuming that $|\vec{\mu}|$ is constant.

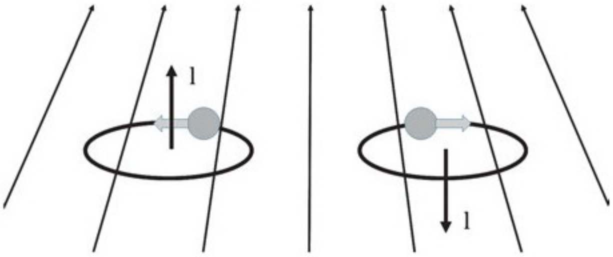

Figure 7.11: Small circuits that are in an inhomogeneous magnetic field. The figure shows two such circuits whose angular momenta point in opposite directions. The thinner lines represent the magnetic field lines. The smaller distances between the field lines in the upper part of the figure compared with the lower part illustrates the inhomogeneity of the field.

Exactly the latter, according to the discussion in this chapter, is the case because $\vec{l}$ is quantized. Furthermore, of convenience, we have introduced a quantization direction through the magnetic field, which in turn means that $\vec{\mu} \cdot \vec{B}$ can only assume discrete values.

Stern and Gerlach used Ag atoms in their experiment. For these, we have to consider the total angular momentum, but it turns out that for Ag atoms in their ground state $\vec{L} = \vec{0}$. This would mean that for all small circuits (= atoms) $\vec{\mu} \cdot \vec{B} = 0$, and that no forces act on the circuits / atoms. But the experiment revealed something else; see Figure 7.12. The Ag atoms were collected on a glass plate, so that one could see exactly where the atoms arrived after passing through the magnetic field. And it was found that there were two groups of atoms. This was explained by the fact that $\vec{\mu} \cdot \vec{B}$ can assume two different values, i. e., that the z component of the angular momentum can take two values.

This result is not easy to explain. We have seen that actually $\vec{L} = \vec{0}$, which means that M_L can only be equal to 0, so that $\vec{B} \cdot \vec{\mu} \propto BM_L$ (if making the convenient choice of placing the z axis along the magnetic field) can only assume one value, 0. It could be suggested that the outermost valence electron of the Ag atom is not in the $5s$ but in the $5p$ orbital, but this would give $L = 1$, so that M_L could assume three different values. The experimental finding that there are only two values for the magnetic moment of

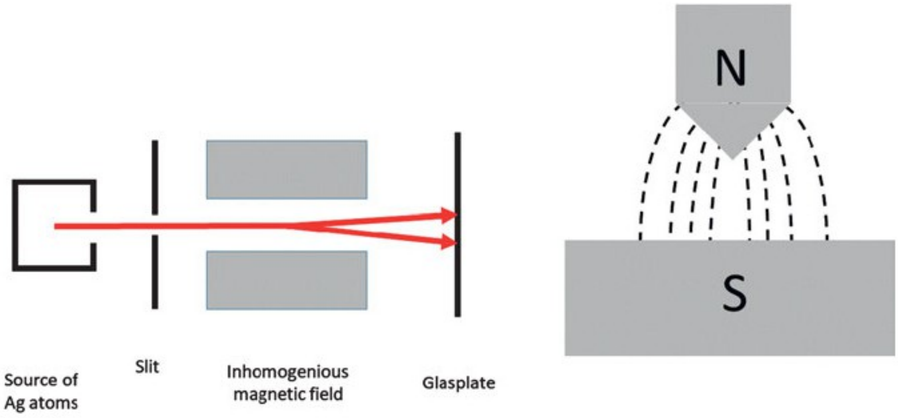

Figure 7.12: Schematic representation of the Stern-Gerlach experiment. Ag atoms enter an inhomogeneous magnetic field and are deflected in two directions and collected on a glass plate. On the right is a cross-section of the magnetic field outlined.

the individual Ag atoms cannot be explained by our previous results that L always has to be an integer, implying that $2L + 1$ is odd and can never be equal to 2.

Nevertheless, one makes the following interpretation: The splitting comes from an additional rotation, which we have not considered so far. The electron cannot only revolve around the nucleus, but also around its own axis (just as the earth revolves around the sun as well as around its own axis). This second rotation is also quantized, and for whatever reason, this motion can be described by means of an angular momentum vector whose length is always equal to $\sqrt{\frac{1}{2}(\frac{1}{2}+1)}\hbar$. This rotational motion is called spin, and its angular momentum is denoted by $\vec{s}$. So this quantum number is always for an electron

$$s = \frac{1}{2} \tag{7.72}$$

so that

$$m_s = -\frac{1}{2}, \frac{1}{2}. \tag{7.73}$$

This interpretation is not perfect. We have found that every rotational motion is quantized and that the allowed quantum numbers can only be integers. In fact, the existence of a spin can only be explained by the combination of quantum theory and the theory of relativity, as Paul Andre Maurice Dirac showed. This combination yields a different equation than the Schrödinger equation (the so-called Dirac equation). It is easy to see that the Schrödinger equation can not be reconciled so easily with the relativity theory. The time-dependent Schrödinger equation (see Section 2.1) includes a second-order differentiation with respect to the position coordinates but only a first-order differentiation with respect to time. The theory of relativity, however, deals with time and position coordinates equivalently. To remedy this contradiction, Dirac introduced a new equation, the Dirac equation, which becomes equivalent to the Schrödinger equation in the limiting case that the speed of light approaches infinity. But as a result of the Dirac equation, one obtains that particles also have a spin which is an integral multiple of a half. As we have seen, it equals a half for electrons.

The Dirac equation will hardly play a role in this course, except for the introduction of the so-called spin-orbit coupling which will be discussed later in Section 11.4.

7.6 Relevance for experiment

By measuring the rotational energy levels of a molecule, information about the moments of inertia of the molecule is obtained. This is part of rotational spectroscopy. This information can be helpful in characterizing the molecule. If the theory in this chapter is used directly to interpret the results, it is assumed that the molecule is rigid. Then one would obtain spectra as shown in Figure 7.13. As this figure suggests, there are so-called selection rules that state that only transitions between neighboring energy levels are possible. This issue shall not be discussed further here. Since the number of degenerate states (i. e., states with the same l but different m_l) increases with l, the heights of the peaks increase initially with l, i. e., with energy. On the other hand, due to the Boltzmann distribution (for the reader who is less familiar with this, see Section 16.4), the number of molecules with a higher energy decreases and, in total, we obtain a curve like that shown in the lower part of Figure 7.13.

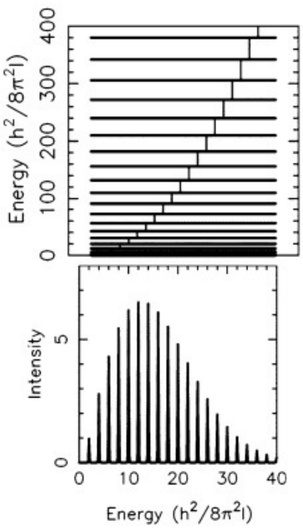

Figure 7.13: Schematic representation of the energy levels of a rigid rotor (upper part) and the accompanying absorption spectrum (lower part). The small, vertical lines in the upper panel show the allowed transitions.

The assumption that the molecule is rigid may be too simple. For example, one can imagine that the molecule expands when it rotates fast (because of the centrifugal force). Then the theory has to be modified accordingly.

Furthermore, one will often simultaneously excite both vibrational and rotational levels. This produces rotational-vibrational spectra, as shown in Figure 7.14. Figure 7.15 shows that the spectrum of the HCl molecule has a fine structure that is even finer than what is caused by the rotation. These are produced by isotope effects: if one isotope is replaced by another, the mass of the corresponding atom changes, and thus also the moment of inertia, and so does the energy level splitting. The analysis of such spectra gives further information on the structure of the molecule.

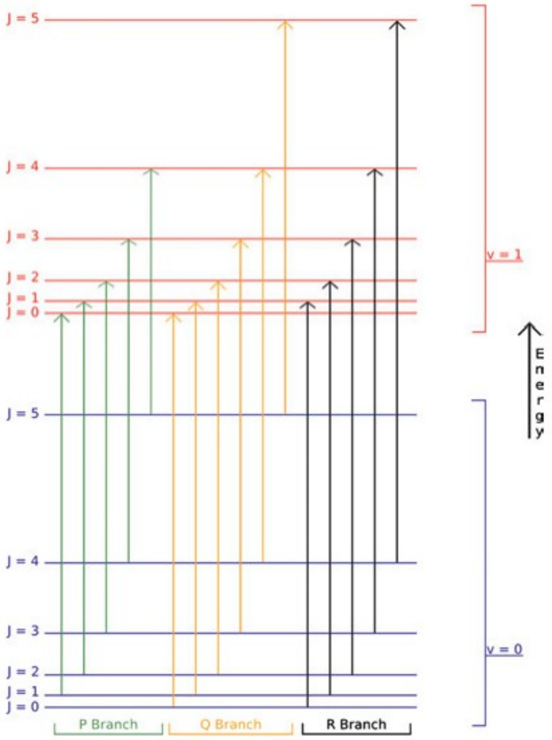

Figure 7.14: The schematic representation of the rotational-vibrational levels of a rotor. $v = 0$ and $v = 1$ represent two levels of vibration, while J represents the different levels of rotation. The vertical arrows show the possible excitations.

But one of the experimentally most important results of this chapter is the existence of spin. The spin also corresponds to an angular momentum vector and therefore couples with the other angular momentum vectors. This, as we shall see later, is important for spin-orbit coupling. Furthermore, the mere existence of a spin is important every-

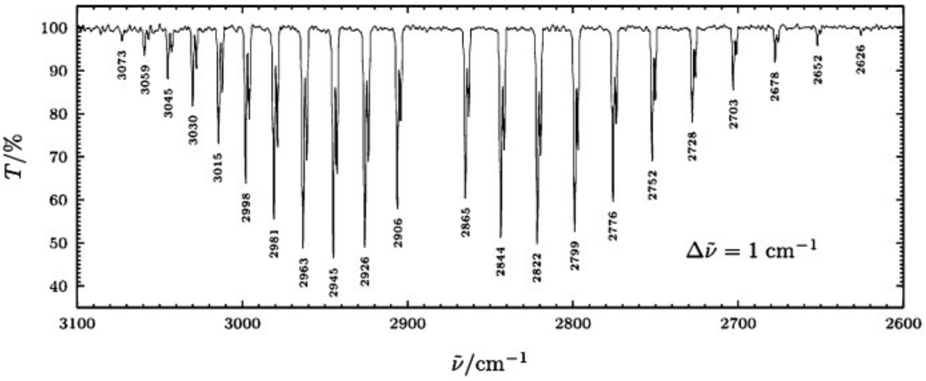

Figure 7.15: The rotation-vibration absorption spectrum of a HCl molecule. Copied on 03.02.17 from https://commons.wikimedia.org/wiki/File:Ir_hcl_rot-vib_mrtz.svg. From mrtzmrtz [CC BY-SA 2.5 (http://creativecommons.org/licenses/by-sa/2.5)], via Wikimedia Commons.

where. Thus, spin-resonance experiments are used to characterize molecules (e. g., NMR, ESR, EPR spectroscopy). The existence of magnetism can to a large part be attributed to the existence of a spin.

In such spin-resonance experiments, it is exploited that those particles that have an angular momentum vector often are charged so that their rotational motion is ultimately equivalent to a small circuit. One can influence this in an external magnetic field. In describing the resulting effect, it is convenient to choose a coordinate system so that the z axis is parallel to the magnetic field. Because the coupling between magnetic field and angular momentum is quantified by the scalar product between magnetic field vector and angular momentum vector, the fact that the z component of the angular momentum vector is quantized means that the size of the coupling can only take finite, discrete values. If one finally measures these couplings, one obtains information about the magnetic field at the place where the particle is located. This magnetic field is not only the externally applied magnetic field, but also has a component, which is due to the fact that all other particles (especially electrons) of the system respond to the externally applied magnetic field and shield it partly. Thus, the information about the magnetic field at the place where the particle is located describes the distribution of the other particles. This provides important chemical information.

Here, we have based the discussion on that the z axis lies along the externally applied magnetic field and the z axis is defined as the quantization direction. This is not necessary: in the end, you would get the same results regardless of how you choose the coordinate system. But this choice is very useful, because it makes all considerations and calculations easier.

7.7 Problems with answers

1. **Problem:** Consider the operators: $\hat{l}_x = \hat{y}\hat{p}_z - \hat{z}\hat{p}_y$, $\hat{l}_y = \hat{z}\hat{p}_x - \hat{x}\hat{p}_z$, $\hat{l}_z = \hat{x}\hat{p}_y - \hat{y}\hat{p}_x$, and $\hat{l}^2 = \hat{l}_x\hat{l}_x + \hat{l}_y\hat{l}_y + \hat{l}_z\hat{l}_z$. Determine the operators $[\hat{l}_x, \hat{l}_y]$, $[\hat{l}_y, \hat{l}_z]$, $[\hat{l}_z, \hat{l}_x]$, and $[\hat{l}_z, \hat{l}^2]$.
Answer: We use

$$\begin{aligned} [\hat{q}_k, \hat{q}_l] &= 0 \\ [\hat{p}_k, \hat{p}_l] &= 0 \\ [\hat{q}_k, \hat{p}_l] &= i\hbar\delta_{k,l}. \end{aligned} \tag{7.74}$$

Then

$$\begin{aligned} [\hat{l}_x, \hat{l}_y] = \hat{l}_x\hat{l}_y - \hat{l}_y\hat{l}_x &= (\hat{y}\hat{p}_z - \hat{z}\hat{p}_y)(\hat{z}\hat{p}_x - \hat{x}\hat{p}_z) - (\hat{z}\hat{p}_x - \hat{x}\hat{p}_z)(\hat{y}\hat{p}_z - \hat{z}\hat{p}_y) \\ &= \hat{y}\hat{p}_z\hat{z}\hat{p}_x - \hat{z}\hat{p}_y\hat{z}\hat{p}_x - \hat{y}\hat{p}_z\hat{x}\hat{p}_z + \hat{z}\hat{p}_y\hat{x}\hat{p}_z \\ &\quad - \hat{z}\hat{p}_x\hat{y}\hat{p}_z + \hat{z}\hat{p}_x\hat{z}\hat{p}_y + \hat{x}\hat{p}_z\hat{y}\hat{p}_z - \hat{x}\hat{p}_z\hat{z}\hat{p}_y \\ &= \hat{y}\hat{p}_z\hat{z}\hat{p}_x + \hat{z}\hat{p}_y\hat{x}\hat{p}_z - \hat{z}\hat{p}_x\hat{y}\hat{p}_z - \hat{x}\hat{p}_z\hat{z}\hat{p}_y = -\hat{y}\hat{p}_x[\hat{z}, \hat{p}_z] + \hat{x}\hat{p}_y[\hat{z}, \hat{p}_z] \\ &= \hat{l}_z[\hat{z}, \hat{p}_z] = i\hbar\hat{l}_z. \end{aligned} \tag{7.75}$$

Subsequently, we use the cyclic permutation

$$\begin{aligned} \hat{l}_x &\to \hat{l}_y \\ \hat{l}_y &\to \hat{l}_z \\ \hat{l}_z &\to \hat{l}_x \end{aligned} \tag{7.76}$$

which changes equation (7.75) into

$$[\hat{l}_y, \hat{l}_z] = i\hbar\hat{l}_x \tag{7.77}$$

and subsequently

$$[\hat{l}_z, \hat{l}_x] = i\hbar\hat{l}_y. \tag{7.78}$$

From these commutator relations, we get

$$\begin{aligned} \hat{l}_z\hat{l}_x &= \hat{l}_x\hat{l}_z + i\hbar\hat{l}_y \\ \hat{l}_x\hat{l}_z &= \hat{l}_z\hat{l}_x - i\hbar\hat{l}_y \\ \hat{l}_z\hat{l}_y &= \hat{l}_y\hat{l}_z - i\hbar\hat{l}_x \\ \hat{l}_y\hat{l}_z &= \hat{l}_z\hat{l}_y + i\hbar\hat{l}_x. \end{aligned} \tag{7.79}$$

Then

$$\begin{aligned} [\hat{l}_z, \hat{l}^2] &= \hat{l}_z(\hat{l}_x\hat{l}_x + \hat{l}_y\hat{l}_y + \hat{l}_z\hat{l}_z) - (\hat{l}_x\hat{l}_x + \hat{l}_y\hat{l}_y + \hat{l}_z\hat{l}_z)\hat{l}_z \\ &= \hat{l}_z\hat{l}_x\hat{l}_x - \hat{l}_x\hat{l}_x\hat{l}_z + \hat{l}_z\hat{l}_y\hat{l}_y - \hat{l}_y\hat{l}_y\hat{l}_z \end{aligned}$$

$$\begin{aligned} &= (\hat{l}_z\hat{l}_x)\hat{l}_x - \hat{l}_x(\hat{l}_x\hat{l}_z) + (\hat{l}_z\hat{l}_y)\hat{l}_y - \hat{l}_y(\hat{l}_y\hat{l}_z) \\ &= (\hat{l}_x\hat{l}_z + i\hbar\hat{l}_y)\hat{l}_x - \hat{l}_x(\hat{l}_z\hat{l}_x - i\hbar\hat{l}_y) + (\hat{l}_y\hat{l}_z - i\hbar\hat{l}_x)\hat{l}_y - \hat{l}_y(\hat{l}_z\hat{l}_y + i\hbar\hat{l}_x) \\ &= 0. \end{aligned} \tag{7.80}$$

Since the commutator vanishes, it is possible to identify states that are eigenfunctions to both operators simultaneously.

2. **Problem:** Consider a system with two *p* electrons. What values can the total spin quantum number *S* and the total angular momentum quantum number *L* assume? Justify the answer.
Answer: It is always true that for a sum like

$$\vec{L} = \vec{l}_1 + \vec{l}_2, \tag{7.81}$$

for which the individual angular momentum vectors

$$\begin{aligned} |\vec{l}_1|^2 &= l_1(l_1+1)\hbar^2 \\ |\vec{l}_2|^2 &= l_2(l_2+1)\hbar^2, \end{aligned} \tag{7.82}$$

are quantized, the sum will also be quantized,

$$|\vec{L}|^2 = L(L+1)\hbar^2. \tag{7.83}$$

The possible values of *L* are then given by

$$L = |l_1 - l_2|, |l_1 - l_2| + 1, \dots, l_1 + l_2 - 1, l_1 + l_2. \tag{7.84}$$

For the *p* electrons $l_1 = l_2 = 1$, so that the values

$$L = 0, 1, 2 \tag{7.85}$$

are possible.
The same applies to the spin, where $s_1 = s_2 = \frac{1}{2}$, so that the values

$$S = 0, 1 \tag{7.86}$$

are possible.

7.8 Problems

1. Explain the Stern–Gerlach experiment.
2. Describe how the energies of the 2-dimensional rotor depend on the quantum number.
3. Sketch the five energetically lowest wave functions of the 2-dimensional rotor.

4. How do the energies of the 2-dimensional rotor change when (a) the mass is doubled or (b) the length of the rotor is doubled?
5. Compare the spherical harmonics Y_{lm_l}, $l = 1$, with the p_x, p_y, and p_z functions. Sketch the functions in (x, y), (x, z), and (y, z) planes.
6. Compare the spherical functions Y_{lm_l}, $l = 2$, with the d_{xz}, d_{yz}, d_{xy}, d_{z^2}, and $d_{x^2-y^2}$ functions. Sketch the functions in (x, y), (x, z), and (y, z) planes.
7. Why is it difficult to explain the spin as a rotation of an electron about its own axis when using the Schrödinger equation?
8. Explain the vector model of the angular momenta.
9. Explain the term "product Ansatz."
10. Consider a system with two d electrons. What values can the total spin quantum number S and the total angular momentum quantum number L assume? Justify the answer.
11. Consider a system with one p electron and one d electron. What values can the total spin quantum number S and the total angular momentum quantum number L assume? Justify the answer.
12. Show that the Y_{lm_l} functions are eigenfunctions to the operator $\hat{l}^2 + \hat{l}_z^2$. Which maximum value or minimum value can the eigenvalue have for a given l?

!F"." G 4 4 !) * ! =!. 0 > 0 ' - > /
> 5 " '9 /5 ' - #' - B' *

!"#$! ! !% & ' ! @ . 0 > 0 ' - > /
> 5 " '9 /5 ' - #' - B' * 0-

I& =8 E '0 E
7/0 9 ' > / ' - . " ' 4

. '0 5 0' ' / 9' - ,0 / 0'0 9/ - ' 7' 0 O =EP5
' - & - 0 - A' L . O P5 & ' /
2 - 9/)' ' - A '
. E 8. ' ' 4 " \ / & ' - ' & 0 v 4 ' \ - 1
' ' 4 ' / 4'9 5 ' 5 ' ' 4'9 9& 0
\ - ,5 ' ' '9 ' 9& 0

!F"." G 4 4 !) * ! =!. ' ' / /
%9
. - - N ;;; 6 0 ' M ;;;

= !"#$! ! !% & ' ! @ > 0 &
'& G H * 8 E 0-
V >

= !"#$! ! !% & ' ! @ > & D ' 8'
. ' ' > / " 6 - / / 0 & '& G H * 8 E 0-
%9 : 4 - ;;; 6 0 ' M ;;;

= !"#$! ! !% & ' ! @ : > /
8> '- ' - ' ' 5 < " ' ' ' ' / 8 0-

= !"#$! ! !% & ' ! @ % & ' / ' - 0'
> / 0 & '& G H * 8 E< 0-

= !"#$! ! !% & ' ! @ # ' 5 # " - + '
' & % & ' . 0 ' ' D & & ' 0-
%9 : 4 - ;;; 6 0 ' M ;;;

== !"#$! ! !% & ' ! @ & ' 5 / /5 ' -
>' & - : ' & ' - 8 0 ' / ' ' " 0-

=@ !"#$! ! !% & ' ! @ . + -' > / ' -
G D - ' ' 0 & '& G H * 8 <= 0-
%9 : 4 - ;;;

=< !"#$! ! !% & ' ! @ . + -' > / ' -
G D - ' ' 0 & '& G H * 8 E 0-

=E !"#$! ! !% & ' ! @ . ' ' > /("
6 - / / 0 & '& G H * 8 E 0-

= !"#$! ! !% & ' ! @ . > / . 0 '
' - D,0 & ' ' 8 E 0-
%9 : 4 - ;;;

@ !"#$! ! !% & ' ! @ . -' & / ' -
D - ' 0 ' 0 & 8 E -24

@ !"#$! ! !% & ' ! @ . > / . 0 '
' - D,0 & . 6. 8 E -24

!F"." G 4 4 !) * ! =!. ' ' / /!>
' - ' '

@ !"#$! ! !% & ' ! @= ' ' > / D - ' 5
. ' ' ' 5 6 4 4 5 ' - 2 4 8 F+F 4 0-

@ !"#$! ! !% & ' ! @= - ' > / 8F 4) *
' ' 8 F+F 4 0-

@ !"#$! ! !% & ' ! @= - ' ' - - '
& 8 F+F 4 0-

@= !"#$! ! !% & ' ! @= D - ' ' - G D - '
> / " " ' / "00 ' 8 F+F 4 0-

@@ !"#$! ! !% & ' ! @= + -' D - ' ' - G
D - ' > / 8 F+F 4 0-

@< !"#$! ! !% & ' ! @= > . ' ' 6
8 F+F 4 0-

@E !"#$! ! !% & ' ! @= > . ' ' 666
8 F+F 4 0-

@ !"#$! ! !% & ' ! @= > . ' ' 5
2 4 . ' ' 8 F+F 4 -24

< !"#$! ! !% & ' ! @= > . ' ' 6
' . ' ' 8 F+F 4 0-

< !"#$! ! !% & ' ! @= > ' F & ' -
. ' & F & 5 ' ' & ' - 6 0 5 - D- " 7' -9 *
> ' - 0 ' & 6 % ' - " D ' -' - 8 F+F 4 0-

< !"#$! ! !% & ' ! @= > ' - ' '
8 F+F 4 0-

< !"#$! ! !% & ' ! @= > / " 0 4
8 F+F 4 0-

< !"#$! ! !% & ' ! @= > / 6 ' - " 6 ' -

6 - D - ' ' - 7/0 9 ' > / 8 F+F 4 0-

<= !"#$! ! !% & ' ! @= > / . & 7 /
D - ' 5 7/0 9 5 ' - 2 4 > 8 F+F 4 0-

<@ !"#$! ! !% & ' ! @= > 0 5 ' 5 ' - 0'
" > 0 . "00 ' : ' " & 9 ' 8 F+F 4 0-

<< !"#$! ! !% & ' ! @= - > /
8 F+F 4 0-

<E !"#$! ! !% & ' ! @= ' ' " 0' /
' ' > / D - ' 5 . ' ' ' 5 6 4 4 5 ' - 2 4 8 F+F 4 0-

< !"#$! ! !% & ' ! @= . / . ' ' > 0 6
> ' 0 . ' ' > 0 " 0 ' / "00 ' ' -
. ' ' 8 F+F 4 0-

E !"#$! ! !% & ' ! @= . * & > ' / " 4 /
> 8 F+F 4 0-
%9 : 4 - ;;;

!F"." G 4 4 !) * ! =!. ' ' / /! 0 :

E !"#$! ! !% & ' ! @@ " - 9 ' & 0 : ?
0 ' -' 8 = 0-

E !"#$! ! !% & ' ! @@ : > 0 5 >
' - F ' DL ' H ' " 0 : 0-

E !"#$! ! !% & ' ! @@ 0 : > '
"9 ' - & 6 66 A # 0 0 ' 8 < 0-
V =@@ 0'&

E !"#$! ! !% & ' ! @@ : 0 F > -
) & ' ')' - 8 E @ -24

!F"." G 4 4 !) * ! =!. ' ' / /!.

E= !"#$! ! !% & ' ! @< 0-
V . " ' /

E@ !"#$! ! !% & ' ! @< + - 2 4 6 &
' 0 & 8 E 0-

E< !"#$! ! !% & ' ! @< + - 2 4 6 & . " ' /
0 & 8 0-

EE !"#$! ! !% & ' ! @< + - 2 4 6 & 8' # &/
' - G G ' + - 0 & 6 ' ' 9 & 8 0-

E !"#$! ! !% & ' ! @< B @ @)' 6 # &/ 0-

!F"." G 4 4 !) * ! =!. > / . 0 ' '
- D,0 &

!"#$! ! !% & ' ! @E . > / . 0 '
' - D,0 & ' ' 8 E 0-
%9 : 4 - ;;;

!"#$! ! !% & ' ! @E . > / . 0 '
' - D,0 & . 6. 8 E -24

= @ =@ =E

< E

!F"." G 4 4 !) * ! @!"-4' F F '
> /

!"#$! ! !% & ' ! @ "-4' F F '
> / 0 & '&) 7 - 9 & 8 @ 0-
I& E8 < @ . ' & ' # & : & : & #'

2.6 *Triangulations: Reconstructing Lengths from Length Cross-Ratios*

To deal with Riemann surfaces that are given in terms of Schottky data (Sect. 8.2) we will need to reconstruct a function $\ell : E_\Sigma \to \mathbb{R}_{>0}$ satisfying (9) from given length cross-ratios. (It is not required that the function ℓ satisfies the triangle inequalities.) To this end, we define auxiliary quantities c^i_{jk} attached to the angles of the triangulation. The value at vertex i of the triangle $ijk \in F$ is defined as

$$c^i_{jk} = \frac{\ell_{jk}}{\ell_{ij}\ell_{ki}}. \tag{11}$$

Then (9) is equivalent to

$$\mathrm{lcr}_{ij} = \frac{c^i_{jk}}{c^i_{lj}}. \tag{12}$$

Now, given a function $\mathrm{lcr} : E^{int} \to \mathbb{R}_{>0}$ defined on the set of interior edges E^{int} and satisfying the product condition (10) around interior vertices, one can find parameters c^i_{jk} satisfying (11) by choosing one value at each vertex and then successively multiplying length cross-ratios. The corresponding function ℓ is then determined by

$$\ell_{ij} = \frac{1}{\sqrt{c^i_{jk}c^j_{ki}}} = \frac{1}{\sqrt{c^i_{lj}c^j_{il}}}. \tag{13}$$

!"#$! ! !% & ' ! @ 0 ' ' "00 ' # '
' 0 & '&) 7 - 9 & 8 0-

I& @E8E # ' 0) ' " & 9 ' 4
' - # ' '

complex numbers. Throughout this chapter, "calculation" is used when exact results are obtained, whereas "computation" is used for numerical results.

Apart from the restriction to an exact representation, all of the algorithms discussed in detail in this chapter are general in the sense that they apply to all compact connected Riemann surfaces. An appendix is presented discussing the use of a few algorithms that apply to a restricted class of algebraic curves and Riemann surfaces, such as elliptic and hyperelliptic surfaces. This appendix contains many examples, but no detailed explanation of the specifics of the algorithms.

All of the descriptions of the algorithms of the main body of this chapter are preceded by the next section which outlines the connection between plane algebraic curves and Riemann surfaces with a level of detail appropriate for what follows. The work reviewed here may be found in [DvH01, DP07a, Pat07, vH95, vH94]. The examples of the implementations use commands available in Maple 11 (Released Spring 2007). A few commands are used that are not available in Maple 11 yet. They are identified as such in the text.

2.2 Relationship Between Plane Algebraic Curves and Riemann Surfaces

In this section, some required background from the theory of Riemann surfaces is introduced. More details can be found in such standard references as [FK92, Spr57]. Excellent places to read up on Riemann surfaces and how they relate to plane algebraic curves that do not require an extensive background are the monographs by Brieskorn and Knörrer [BK86] and Griffiths [Gri89].

Consider a plane algebraic curve, defined over the complex numbers $\mathbb{C}$, i.e., consider the subset of $\mathbb{C}^2$ consisting of all points (x, y) satisfying a polynomial relation in two variables x and y with complex coefficients:

$$f(x,y) = a_n(x)y^n + a_{n-1}(x)y^{n-1} + \ldots + a_1(x)y + a_0(x) = 0 \,. \tag{2.1}$$

Here $a_j(x)$, $j = 0, \ldots, n$ are polynomials in x. Write $a_j(x) = \sum_i a_{ij}x^i$, where the coefficients a_{ij} are complex numbers. Assuming $a_n(x)$ does not vanish identically, n is the degree of $f(x, y)$ considered as a polynomial in y. We only consider irreducible algebraic curves, so $f(x, y)$ cannot be written as the product of two non-constant polynomials with complex coefficients.

Let d denote the degree of $f(x, y)$ as a polynomial in x and y, i.e., d is the largest $i + j$ for which the coefficient a_{ij} of $x^i y^j$ in $f(x, y)$ is non-zero. The behavior at infinity for both x and y is examined by homogenizing $f(x, y) = 0$ by letting

$$x = X/Z \,, \quad y = Y/Z \,, \tag{2.2}$$

and investigating

$$F(X,Y,Z) = Z^d f(X/Z, Y/Z) = 0 \,. \tag{2.3}$$

Here $F(X, Y, Z)$ is a homogeneous polynomial equation of degree d. Finite points $(x, y) \in \mathbb{C}^2$ on the algebraic curve correspond to triples $(X : Y : Z)$ with $Z \neq 0$. Since for these points $(X : Y : Z) = (X/Z : Y/Z : 1)$, we may equate $Z = 1$, so finite points can be denoted unambiguously by (x, y) instead of $(X/Z, Y/Z)$. Points at infinity correspond to triples $(X : Y : Z)$, with $Z = 0$. Hence, at a point at infinity, at least one of the two coordinate functions x or y is infinite. Because $F(X, Y, 0)$ is a homogeneous polynomial of degree d, there are at most d points at infinity.

The algebraic curve can have singular points. An algorithm to efficiently calculate the singular points of an algebraic curve is discussed in Sect. 2.5. Here we briefly discuss singularities as they need to be dealt with in order to construct the Riemann surface from an algebraic curve. Finite singular points on the algebraic curve specified by $f(x, y) = 0$ satisfy $f(x, y) = 0 = \partial_x f(x, y) = \partial_y f(x, y)$. Points at infinity can also be singular. Singular points at infinity satisfy $\partial_X F(X, Y, Z) = \partial_Y F(X, Y, Z) = \partial_Z F(X, Y, Z) = 0$ (then also $F(X, Y, Z) = 0$, by Euler's theorem for homogeneous functions). Desingularizing the algebraic curve results in a Riemann surface, i.e., a one-dimensional complex-analytic manifold (so it is two-dimensional over the real numbers; it is a surface). There are various ways of desingularizing algebraic curves. Our methods use Puiseux series, as detailed in Sect. 2.3. Each nonsingular point on the algebraic curve corresponds to one place[1] on the Riemann surface, whereas a singular point on the algebraic curve can correspond to multiple places on the Riemann surface.

In what follows, Γ is used to denote the Riemann surface obtained by desingularizing and compactifying (by adding the places at infinity) the algebraic curve represented by $f(x, y) = 0$. All Riemann surfaces obtained this way are connected (because $f(x, y)$ is irreducible) and compact (because the points at infinity are included). Conversely, as stated in Chap. 1 it is known that every compact connected Riemann surface can be obtained as described above [BBE$^+$94, Spr57]. From here on out, all Riemann surfaces considered are understood to be connected and compact. We use $\hat{\Gamma}$ to denote the compactified, not desingularized algebraic curve.

2.3 Puiseux Series

Many of the algorithms presented in the next sections make use of local coordinates on an algebraic curve. For our algorithms, this local behavior is understood using Puiseux expansions. These expansions allow us to distinguish between regular points, branch points and singular points. Further, in addition to determine the nature of singular points, Puiseux expansions characterize

---

[1] We use the term "point" to denote a value in the complex x-plane. On the other hand, "place" is used to denote a location on the Riemann surface Γ, or, unambiguously, a location on the desingularized plane algebraic curve.

!"#$! ! !% & ' ! @ F F ' > /
) * c)' 8 E 0-
I& 8 '

!"#$! ! !% & ' ! @ ' 4K DL ' F '
> / ' 0 & '&) 7 - 9 & 8 0-
%9 : 4 - ;;;

I& 8 @) ' D - ' . 0'

I& E 8 = ' ' 4'
= ' ' G &' 4 >' ' 4'
' - . ' & " / 0 :

!F"." G 4 4 !) * ! @!" ' / 6 Y " ' / 665 . - D-

= !"#$! ! !% & ' ! < " ' / 6 . - D- 0 &
&'0 8 @ 0-

I& =@8< 6 '& ' - 4 '&

I& @<8 E '0
. # ' & '

@ !"#$! ! !% & ' ! < " ' / 66 . - D- 0 &
&'0 8 @ 0-

I& 8 < '0
0'

!F"." G 4 4 !) * ! @! ' ' " & 9 ' > / ' -

< !"#$! ! !% & ' ! < ' ' ' & 9 ' & / '
- 4 8 - ' 8 0-

I& <8 < ' ' ' ' '
< D, 0 ' \& '

7.3.1 Exceptional configurations

From now on we will study birational maps between algebraic surfaces. We know (see [311], Chapter V, §5) that any birational morphism $\pi : Y \to X$ of nonsingular projective surfaces can be factored into a composition of blow-ups with centers at closed points. Let

$$\pi : Y = Y_N \xrightarrow{\pi_N} Y_{N-1} \xrightarrow{\pi_{N-1}} \dots \xrightarrow{\pi_2} Y_1 \xrightarrow{\pi_1} Y_0 = X \tag{7.28}$$

be such a factorization. Here $\pi_i : Y_i \to Y_{i-1}$ is the blow-up of a point $x_i \in Y_{i-1}$. Set

$$\pi_{ki} := \pi_{i+1} \circ \dots \circ \pi_k : Y_k \to Y_i, \quad k > i.$$

Let

$$E_i = \pi_i^{-1}(x_i), \quad \mathcal{E}_i = \pi_{Ni}^*(E_i). \tag{7.29}$$

E !"#$! ! !% & ' ! < ' ' " & 9 ' > / '
- 4 ' 9 -& ? 4 / 8 0-

!F"." G 4 4 !) * ! @!D &/ F / + ' -
" G

!"#$! ! !% & ' ! < D &/ F / + ' -
" G 0-
%9 : 4 - ;;
V ' 0' ' -' ' L ' & '4 -'- t , 5 '0 ' K - Z ' -' i'

I& = 8<

' 4 - 0 L ' /

* * 0'

!F"." G 4 4 !) * ! @!+ , 9 D

!"#$! ! !% & ' ! < + , 9 D 5 0-

I& @8<= = ' ' ' - ' '- 4'

. ' / ' w , 9 0 0' -5 - 9 - '
' 0 '9 - & ' 0' - ' 0 ' 5 0
,0 - ' ' 4 ' - ' ' ' x . '
' 0 - - , ' 9 - & ' & 0 '
' ' ' -' '& & ' 0 ' ' 8 ' 0
0 7 5 ' ' ' ' - ' ' '-
4' y 5 - \ - ' 4' '- 4' '
0 ' ' ' / ' 0 ' ',
' '* 0 ' ' 5 ' - w , 9 - 4 ' '-
- \ 5 ' 5 - & ' &5 - 4 9 ' 5
' ' 0 ' - 4 9 ' B '
' / ' - & & - 0' 4 0 5 & / - & 9 ' 5 L '
8 E ' ' & - ' - ,0 -

$$\rho_{\text{Critical}} = \left(\frac{d_{\text{f}} + d_{\text{s}}}{2\varepsilon_{\text{Critical}}}\right)\left\{\frac{1 + 2\eta + \chi\eta^2}{(1 + \eta)(1 + \chi\eta)}\right\}$$

Accordingly, the strain at the top surface of the film–substrate structure is expressed in the form

$$\varepsilon_{\text{Top}} \text{ for compliant substrate} = \varepsilon_{\text{Top}} \text{ for non-compliant substrate} \times \left(\frac{y'}{y}\right) = \left(\frac{d_{\text{f}} + d_{\text{s}}}{2\rho}\right)\left\{\frac{1 + 2\eta + \chi\eta^2}{(1 + \eta)(1 + \chi\eta)}\right\} \tag{2.48}$$

2.5 Critical strain and critical radius of curvature

The functionality of a flexible electronics component is impaired, i.e. disturbed to an unacceptable degree as compared to its normal operation, when the component is exposed to a certain level of strain referred to as the critical strain ε_{C}. This critical strain is produced when the external bending forces acting on the component can cause mechanical damage resulting in electrical or optical failure (for an optoelectronic component). Hence, critical strain is associated with a critical radius of curvature ρ_{C}, which is defined as the value of radius of curvature at which the component fails mechanically or in its electrical/optical functions. Since maximum strains take place on surfaces, and the common flexible electronic devices are made from different films, metallic, semiconducting or insulating, formed over substrates, the strain at the topmost surface of the device becomes crucial. Within the realms of elasticity and ignoring dissipative processes, e.g. yielding of the substrate, equation (2.48) is re-arranged and expressed in the form

$$\rho_{\text{Critical}} = \left(\frac{d_{\text{f}} + d_{\text{s}}}{2\varepsilon_{\text{Critical}}}\right)\left\{\frac{1 + 2\eta + \chi\eta^2}{(1 + \eta)(1 + \chi\eta)}\right\} \tag{2.49}$$

An inverse proportionality relationship holds between ρ_{Critical} and $\varepsilon_{\text{Critical}}$, i.e.

$$\rho_{\text{Critical}} \propto 1/\varepsilon_{\text{Critical}} \tag{2.50}$$

This relationship depends on the geometry of the multilayer structure as well as its mechanical properties.

For the special case,

$$\frac{Y_{\text{f}}}{Y_{\text{s}}} = \chi \leqslant 10, \quad \frac{d_{\text{f}}}{d_{\text{s}}} = \eta \ll 1 \tag{2.51}$$

Equation (2.49) is simplified to

$$\rho_{\text{Critical}} = \frac{d_{\text{f}} + d_{\text{s}}}{2\varepsilon_{\text{Critical}}} \approx \frac{d_{\text{s}}}{2\varepsilon_{\text{Critical}}} \tag{2.52}$$

or,

$$\varepsilon_{\text{Critical}} = \frac{d_{\text{s}}}{2\rho_{\text{Critical}}} \tag{2.53}$$

Table 2.1. Distinguishing flexible electronics from compliant and stretchable electronics.

Sl. no.	Property	Flexible electronics	Compliant electronics	Stretchable electronics
1.	Critical strain $\varepsilon_{\text{Critical}}$	<2%	$2\% < \varepsilon_{\text{Critical}} < 10\%$	$\varepsilon_{\text{Critical}} > 10\%$
2.	Critical radius of curvature ρ_{Critical}	ρ_{Critical} up to 10 mm	$\rho_{\text{Critical}} \sim$ a few mm	$\rho_{\text{Critical}} <$ a few mm

For a given ρ_{Critical}, the $\varepsilon_{\text{Critical}}$ is low for a thin substrate (small d_s) so that a thin substrate experiences lower peak strain and components mounted on such a substrate can withstand smaller ρ_{Critical} values.

2.6 $\varepsilon_{\text{Critical}}$ and ρ_{Critical} as characteristic parameters defining flexible, compliant and stretchable electronics

Defining the ρ_{Critical} and $\varepsilon_{\text{Critical}}$ terms enables us to differentiate between the three intermingling branches (Leterrier 2015):

(i) Flexible electronics: Here the components can sustain critical strain $\varepsilon_{\text{Critical}} < 2\%$ and can withstand bending up to the radius of curvature $\rho_{\text{Critical}} = 10$ mm. Flexible electronics components are fabricated on thin substrates with thicknesses in the range of 100 μm.

(ii) Compliant electronics: Critical strain sustaining capability is between 2%–10%, i.e. $2\% < \varepsilon_{\text{Critical}} < 10\%$. Critical radius of curvature ρ_{Critical} level is ~ a few mm. Some degree of in-plane loading is permissible. Thicker substrates can be used.

(iii) Stretchable electronics: The components and circuits can endure $\varepsilon_{\text{Critical}} > 10\%$. Critical radius of curvature can be < a few mm. Conformability is possible to a wide variety of two-dimensional surfaces. The aforementioned features are summarized in table 2.1.

2.7 Discussion and conclusions

The key terminology and formulae of mechanical bending form the foundations for visualizing the meaning of flexibility in the context of electronics.

Review exercises

2.1 Do curvature and radius of curvature have the same meaning?
2.2 What is the radius of curvature of a straight flat surface?
2.3 ‘A unit vector is dimensionless.’ Right or wrong?
2.4 Derive the popular equation for radius of curvature of a beam

$$\rho = \frac{(1 + y'^2)^{3/2}}{y''}, \tag{2.54}$$

where $y' = dy/dx$ and $y'' = d^2y/dx^2$.

!"#$! ! !% & ' ! < + , 9 D 5 9

!"#$! ! !% & ' ! < + , 9 D 5 0-

!F"." G 4 4 !) * ! @!>D% D.#H

!"#$! ! !% & ' ! < " . ' - > / 0-

!"#$! ! !% & ' ! < " & 9 '5 . 0 &/5 F '
' 5 ' - %0 3' . / + 0 ' - D & & 8 < 0-
V E 0'&
%9 : 4 - ;;;

= !"#$! ! !% & ' ! < 4 ' - ' &
- & / ' - ' & &' 1' ' 9 8 0-
%9 : 4 - ;;;

@ !"#$! ! !% & ' ! < 4 ' - ' &
- & / ' - ' & 1' ' 8 E 0-

< !"#$! ! !% & ' ! < F' A 4 , 0 3'
' - D - ' - ' & / 9 8 0-
%9 : 4 - ;;;

E !"#$! ! !% & ' ! < + -' > / ' -
6 - 0-
%9 : 4 - ;;;

I& = 8=< 7" .D# 6
".6% .D %#": #D:".6%G 6GFD DGFDG. %+ .7D
F6 .6G .6%G)D.B DDG DW.DG 6%G "GF F?#".6%Gp
.7D G%.6%G %+ " DG D F".?

!"#$! ! !% & ' ! < > - ' -
"00 ' + 0 ' - D & & 0 & '& G H * 8 0-
%9 : 4 - ;;;

I& 8 '0 =
)' 2 4 > /

I& <<8 @ '0 @
)' D - ' > /

I& 8 '0 <
0' ' & ' - 00 & 7/0 0 '

!"#$! ! !% & ' ! < > / # 4 - 0-
%9 : 4 - ;;; 6 0 ' M ;;;

!"#$! ! !% & ' ! < > / # 4 - 0-

!"#$! ! !% & ' ! < > '4 ' ' : & ' - %0 '
> / " ' 0 4 0-

I& <=8E@ F w " & & F' * ' 9/) ' * 7 ' - B
? & >') .

!"#$! ! !% & ' ! < > 0 " ' > / 0-
%9 : 4 - ;;;

!"#$! ! !% & ' ! < > 0 " ' > / 0-

= !"#$! ! !% & ' ! < 6 & ' > / " - 4 , /
- & 6 ' ' 0-
%9 : 4 - ;;;

I& <8 E #"F%G 6G D# 6%G +#% :6GD .% >#" "GG6"G

") .#" . B 9 & ' * 4 0 9 + * #'-
' ' - ' - 4 0 ' - , 0 9 ' '
- ' 9 - & ' 0' > ' ' '
- . 2 *) # 9 . ' ' 0' '
0 - ' B '
. + * . ' 0

I& @ 8 E %G .7D +:"> ?#".?#D " G F ?#".?#D 6G
+6G :D# >D% D.#H

"9 ' . '& 4' ' ' ' & ' 4'
' ' & / 6 0'0 5 -/ '
0 9 '& 4' ' - ' 0 ' # ' '
L ' / 4' + 5 ' - 0' ' / -
'& 4' + ' ' 4' ' -
m 4' 0 + 5 ' / 0 2
4 / ' + 0 4'

@ !"#$! ! !% & ' ! < 6 - : 3 > /
4 ' - ' 0-
%9 : 4 - ;;;

I& 8 '0
B : 3 * * 0'

< !"#$! ! !% & ' ! < A ' >')' " & 9 '5
. 0 &/5 ' - F ' ' O: P 8 = 0-
V 0'&

E !"#$! ! !% & ' ! < A ' >' + -' '
: ' " & 9 ' ' - %0 3' O: P 8 < 0-
V 0'&

I& E

'0
D - ' 0'
] 9 ' L 4 '
z7 ' * *
6 - 5 D - ' 0'

' 5 ' * 4 0' ' - ' ' 4 ' - '
9 ' 5 9 '/ ,0 & ' & ' *
'9 & ' / 4 " D - ' ' - '
' & ' / ' - & 8 - '
. '0 4 9' 9 D - ' & / F 0 ' 0 D - '
& / ' 4 &' - '0 % ' & ' & 4 9' 0 0
' ' ' 0 4 D - ' 5 ' ' - w 5
/ 0 '/ ' 0 ' 0 ' D - ' & / -/ 0 0
4' ' - ' ' '0 ' - & - # & - ' '0 '
0 4 - ' 9 0
B 9 & 9/ - \ & 0 - ' - D - ' 0' . ' / ' 3
L ' / ' - * * L ' / ' B - \ & ' / 4 ' -
9 0' 5 & ' 9' 5 ' - ' 9' B 0 4 ' 4 / \ - '
D - ' 0' ' ' 9' . \ 0 - ' /5 ' - -
> ' - & ' 3' 0 - . $# - 0 4 9 '
' ' '00 ' > ' - 0 - : ' 8'
' - & ' ' ' ' - \ - ' - - - 9 w/ B - '
'00 ' D - ' & / ' * - %
0 ' '00 ' 5 - ' L ' 5 - - '0 =

!"#$! ! !% & ' ! < A ' >' G F '
> / ' - : > 0 O9 * - ' P 8 < 0-
%9 : 4 - ;;;

V 4' 5 # ' 5 # ' ' 5

V 0'&

!"#$! ! !% & ' ! < A F' 4 , %0 3'
C D - ' F ' > / 9 9 & 8? " 8 0-
%9 : 4 - ;;;
V E@E 0'&

!"#$! ! !% & ' ! < 1 4U > /5) * 66
/ 0-

!"#$! ! !% & ' ! < : > /
' - 5 . - D- 0-
%9 : 4 - ;;;

Contents

Preface vii

1. Manifolds 1
   - 1.1 Preliminaries . . . 1
     - 1.1.1 Space and Coordinatization . . . 1
     - 1.1.2 The implicit function theorem . . . 3
   - 1.2 Smooth manifolds . . . 6
     - 1.2.1 Basic definitions . . . 6
     - 1.2.2 Partitions of unity . . . 9
     - 1.2.3 Examples . . . 10
     - 1.2.4 How many manifolds are there? . . . 20

2. Natural Constructions on Manifolds 23
   - 2.1 The tangent bundle . . . 23
     - 2.1.1 Tangent spaces . . . 23
     - 2.1.2 The tangent bundle . . . 27
     - 2.1.3 Transversality . . . 29
     - 2.1.4 Vector bundles . . . 33
     - 2.1.5 Some examples of vector bundles . . . 37
   - 2.2 A linear algebra interlude . . . 41
     - 2.2.1 Tensor products . . . 41
     - 2.2.2 Symmetric and skew-symmetric tensors . . . 46
     - 2.2.3 The “super” slang . . . 53
     - 2.2.4 Duality . . . 56
     - 2.2.5 Some complex linear algebra . . . 64
   - 2.3 Tensor fields . . . 67
     - 2.3.1 Operations with vector bundles . . . 67
     - 2.3.2 Tensor fields . . . 69
     - 2.3.3 Fiber bundles . . . 73

3. Calculus on Manifolds 79

- 3.1 The Lie derivative . . . 79
  - 3.1.1 Flows on manifolds . . . 79
  - 3.1.2 The Lie derivative . . . 81
  - 3.1.3 Examples . . . 86
- 3.2 Derivations of $\Omega^\bullet(M)$ . . . 88
  - 3.2.1 The exterior derivative . . . 88
  - 3.2.2 Examples . . . 93
- 3.3 Connections on vector bundles . . . 94
  - 3.3.1 Covariant derivatives . . . 94
  - 3.3.2 Parallel transport . . . 100
  - 3.3.3 The curvature of a connection . . . 101
  - 3.3.4 Holonomy . . . 105
  - 3.3.5 The Bianchi identities . . . 108
  - 3.3.6 Connections on tangent bundles . . . 109
- 3.4 Integration on manifolds . . . 111
  - 3.4.1 Integration of 1-densities . . . 111
  - 3.4.2 Orientability and integration of differential forms . . . 115
  - 3.4.3 Stokes' formula . . . 123
  - 3.4.4 Representations and characters of compact Lie groups . . . 127
  - 3.4.5 Fibered calculus . . . 133

4. Riemannian Geometry 139

- 4.1 Metric properties . . . 139
  - 4.1.1 Definitions and examples . . . 139
  - 4.1.2 The Levi–Civita connection . . . 143
  - 4.1.3 The exponential map and normal coordinates . . . 147
  - 4.1.4 The length minimizing property of geodesics . . . 150
  - 4.1.5 Calculus on Riemann manifolds . . . 155
- 4.2 The Riemann curvature . . . 165
  - 4.2.1 Definitions and properties . . . 165
  - 4.2.2 Examples . . . 169
  - 4.2.3 Cartan's moving frame method . . . 172
  - 4.2.4 The geometry of submanifolds . . . 175
  - 4.2.5 Correlators and their geometry . . . 182
  - 4.2.6 The Gauss–Bonnet theorem for oriented surfaces . . . 192

5. Elements of the Calculus of Variations 201

- 5.1 The least action principle . . . 201
  - 5.1.1 The 1-dimensional Euler–Lagrange equations . . . 201
  - 5.1.2 Noether's conservation principle . . . 207
- 5.2 The variational theory of geodesics . . . 210

5.2.1 Variational formulæ 211
5.2.2 Jacobi fields 214
5.2.3 The Hamilton-Jacobi equations 220

6. The Fundamental Group and Covering Spaces 227

6.1 The fundamental group 228
6.1.1 Basic notions 228
6.1.2 Of categories and functors 232
6.2 Covering Spaces 233
6.2.1 Definitions and examples 233
6.2.2 Unique lifting property 235
6.2.3 Homotopy lifting property 236
6.2.4 On the existence of lifts 237
6.2.5 The universal cover and the fundamental group 239

7. Cohomology 241

7.1 DeRham cohomology 241
7.1.1 Speculations around the Poincaré lemma 241
7.1.2 Čech vs. DeRham 245
7.1.3 Very little homological algebra 247
7.1.4 Functorial properties of the DeRham cohomology 254
7.1.5 Some simple examples 257
7.1.6 The Mayer–Vietoris principle 259
7.1.7 The Künneth formula 262
7.2 The Poincaré duality 265
7.2.1 Cohomology with compact supports 265
7.2.2 The Poincaré duality 268
7.3 Intersection theory 272
7.3.1 Cycles and their duals 272
7.3.2 Intersection theory 277
7.3.3 The topological degree 282
7.3.4 The Thom isomorphism theorem 284
7.3.5 Gauss–Bonnet revisited 287
7.4 Symmetry and topology 291
7.4.1 Symmetric spaces 291
7.4.2 Symmetry and cohomology 294
7.4.3 The cohomology of compact Lie groups 298
7.4.4 Invariant forms on Grassmannians and Weyl's integral formula . 299
7.4.5 The Poincaré polynomial of a complex Grassmannian 306
7.5 Čech cohomology 312
7.5.1 Sheaves and presheaves 313
7.5.2 Čech cohomology 317

8. Characteristic Classes 329
8.1 Chern–Weil theory . . . 329
8.1.1 Connections in principal G-bundles . . . 329
8.1.2 G-vector bundles . . . 335
8.1.3 Invariant polynomials . . . 336
8.1.4 The Chern–Weil Theory . . . 339
8.2 Important examples . . . 343
8.2.1 The invariants of the torus T^n . . . 343
8.2.2 Chern classes . . . 343
8.2.3 Pontryagin classes . . . 346
8.2.4 The Euler class . . . 348
8.2.5 Universal classes . . . 351
8.3 Computing characteristic classes . . . 357
8.3.1 Reductions . . . 358
8.3.2 The Gauss–Bonnet–Chern theorem . . . 363

9. Classical Integral Geometry 373
9.1 The integral geometry of real Grassmannians . . . 373
9.1.1 Co-area formulæ . . . 373
9.1.2 Invariant measures on linear Grassmannians . . . 386
9.1.3 Affine Grassmannians . . . 395
9.2 Gauss–Bonnet again?!? . . . 398
9.2.1 The shape operator and the second fundamental form . . . 398
9.2.2 The Gauss–Bonnet theorem for hypersurfaces of an Euclidean space . . . 401
9.2.3 Gauss–Bonnet theorem for domains in an Euclidean space . . . 405
9.3 Curvature measures . . . 409
9.3.1 Tame geometry . . . 409
9.3.2 Invariants of the orthogonal group . . . 414
9.3.3 The tube formula and curvature measures . . . 418
9.3.4 Tube formula $\Longrightarrow$ Gauss–Bonnet formula for arbitrary submanifolds of an Euclidean space . . . 429
9.3.5 Curvature measures of domains in an Euclidean space . . . 431
9.3.6 Crofton formulæ for domains of an Euclidean space . . . 433
9.3.7 Crofton formulæ for submanifolds of an Euclidean space . . . 443

10. Elliptic Equations on Manifolds 451
10.1 Partial differential operators: algebraic aspects . . . 451
10.1.1 Basic notions . . . 451
10.1.2 Examples . . . 457
10.1.3 Formal adjoints . . . 459
10.2 Functional framework . . . 464

10.2.1 Sobolev spaces in $\mathbb{R}^N$ . . . 464
10.2.2 Embedding theorems: integrability properties . . . 471
10.2.3 Embedding theorems: differentiability properties . . . 476
10.2.4 Functional spaces on manifolds . . . 480
10.3 Elliptic partial differential operators: analytic aspects . . . 484
10.3.1 Elliptic estimates in $\mathbb{R}^N$ . . . 485
10.3.2 Elliptic regularity . . . 489
10.3.3 An application: prescribing the curvature of surfaces . . . 494
10.4 Elliptic operators on compact manifolds . . . 504
10.4.1 Fredholm theory . . . 504
10.4.2 Spectral theory . . . 513
10.4.3 Hodge theory . . . 518

11. Spectral Geometry 523

11.1 Generalized functions and currents . . . 523
11.1.1 Generalized functions and operations withy them . . . 523
11.1.2 Currents . . . 528
11.1.3 Temperate distributions and the Fourier transform . . . 529
11.1.4 Linear differential equations with distributional data . . . 531
11.2 Important families of generalized functions . . . 536
11.2.1 Some classical generalized functions on the real axis . . . 536
11.2.2 Homogeneous generalized functions . . . 542
11.3 The wave equation . . . 548
11.3.1 Fundamental solutions of the wave . . . 548
11.3.2 The wave family . . . 550
11.3.3 Local parametrices for the wave equation with variable coefficients . . . 567
11.4 Spectral geometry . . . 574
11.4.1 The spectral function of the Laplacian on a compact manifold . . 574
11.4.2 Short time asymptotics for the wave kernel . . . 580
11.4.3 Spectral function asymptotics . . . 588
11.4.4 Spectral estimates of smoothing operators . . . 593
11.4.5 Spectral perestroika . . . 598

12. Dirac Operators 611

12.1 The structure of Dirac operators . . . 611
12.1.1 Basic definitions and examples . . . 611
12.1.2 Clifford algebras . . . 614
12.1.3 Clifford modules: the even case . . . 617
12.1.4 Clifford modules: the odd case . . . 622
12.1.5 A look ahead . . . 623
12.1.6 The spin group . . . 625

12.1.7 The complex spin group 633
12.1.8 Low dimensional examples 635
12.1.9 Dirac bundles 640
12.2 Fundamental examples 644
12.2.1 The Hodge–DeRham operator 644
12.2.2 The Hodge–Dolbeault operator 648
12.2.3 The $spin$ Dirac operator 654
12.2.4 The $spin^c$ Dirac operator 660

Bibliography 669

Index 675

!"#$! ! !% & ' ! < ' ' ' B *9 *
)' > / ' : & 0-

!"#$! ! !% & ' ! < ' ' ' B *9 *
)' > / - + ' 0-

= !"#$! ! !% & ' ! < G F ' > / ' -
: > 0 O: P 8 @ 0-
V @ 0'&

I& @ '0 4' # ' ' ' -

@ !"#$! ! !% & ' ! < ' 5 ' - %
6 & G 9 0-
%9 : 4 - ;;;

< !"#$! ! !% & ' ! < . / D ' '
0 & &'0 8 0-

E !"#$! ! !% & ' ! < / ' - %4 - -
/ ' ' F ' DL ' 8. 6 " ' ' ' - "00 '
8 < 0-
%9 : 4 - ;;;
V =@= 0'&

!"#$! ! !% & ' ! < / ' ' F '
DL ' 4' :' z "00 ' z " & 0 & G ' - 8 0-

!"#$! ! !% & ' ! < .) '- / 0 ' -
> / + 7 " ' B) * c) 8 = 0-

!"#$! ! !% & ' ! < . 0 - A B'
66 8 @@ 0 9 @@E 0-

!"#$! ! !% & ' ! < 1 4U > /) * 6
' / 6 3-' 0-

!"#$! ! !% & ' ! < 1 4U > /) * 66
/ 66 3-' 8 E 0-

!"#$! ! !% & ' ! < & 6 ' > /
. / ' - "00 ' 0 9

= !"#$! ! !% & ' ! < . '0 6 0' &
. / ' - > / ? 4 U 7 -- F 0 9

@ !"#$! ! !% & ' ! < & . / ' - > /
? 4 U 7 -- F)') * 8 0-

< !"#$! ! !% & ' ! < " 7 / = H '
' ' ' 7' 4' - 7' 4' - ? 4 / 8 0-

E !"#$! ! !% & ' ! < . '0 ' : %
' ' '] ' ? 4] 7 -- > / H' ? 4 / 0-

!F"." G 4 4 !) * ! @!7' -9 * D,'
G ' q- & DL '

!"#$! ! !% & ' ! <= 7' -9 * D,'
G ' q- & DL ' 0-
%9 : 4 - ;;;

!F"." G 4 4 !) * ! @!7/0 - D / /

= !"#$! ! !% & ' ! <@ 7/0 - D /
/ 0-
V = 0'&
%9 : 4 - ;;;

I& 8= '0
0 - ' - '
'- 9 4'

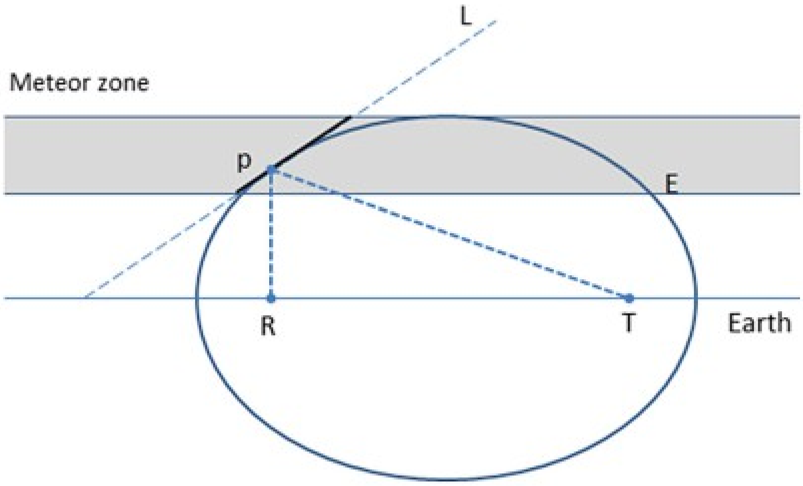
L
Meteor zone
p
E
R
T
Earth

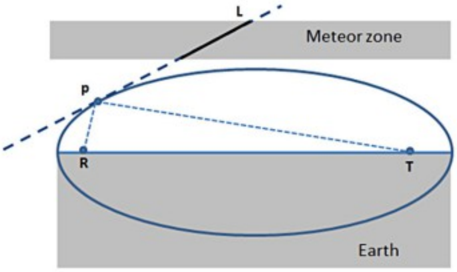
L
Meteor zone
p
R
T
Earth

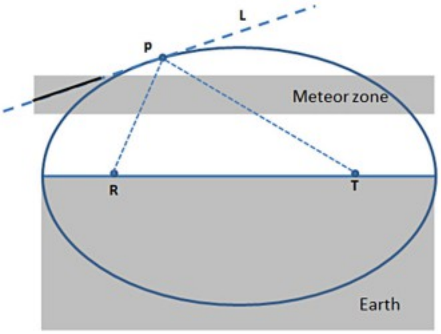
L
p
Meteor zone
R
T
Earth

Chapter 3

Properties of meteoroids from forward scatter radio observations

Hervé Lamy

3.1 Radio meteor theory

When a meteoroid enters the Earth's atmosphere with velocities in excess of $\sim$10 km s^{-1}, it impacts the atoms and molecules of the upper atmosphere with a kinetic energy large enough to ionize them. It also heats up to high temperatures and starts to ablate. As a result, a trail of ions and electrons forms along the trajectory path behind the meteoroid.

If a radio wave of a given frequency is sent towards space from a ground-based transmitter, it can be temporarily reflected by the meteor trail towards the ground. The incident radio wave is scattered only by electrons as ions are too heavy. Under some geometrical conditions (see below), it can be recorded by a receiver tuned to the same frequency. The signal recorded at the receiving station is called a meteor trail echo. Its duration is directly related to the lifetime of the meteor trail as the electrons tend to quickly scatter into the neutral ambient atmosphere. A meteor radar corresponds to the case where the transmitter and the receiver are located at the same place. When the transmitter and receiver(s) are not on the same site, we talk about forward scatter observations (see figure 3.1) which is the topic of this chapter.

Meteor echoes are usually classified into two categories called underdense and overdense, according to the value of the line electron density α (expressed in number of electrons per meter). This is based on McKinley's [1] classical radio meteor theory, where it is first assumed that the trail is a stationary straight line of electrons, a hypothesis made because the longitudinal extent of the trail (typically a few kilometers to a few tens of kilometers) is much larger than its radial extent (from a few tens of centimeters to a few meters). Additional factors are then included, such as the initial radius effect, to take into account the finite radial extent of the trail (assumed to be cylindrical), and the ambipolar diffusion to model the scattering of trail electrons into the ambient neutral atmosphere. The typical duration of

doi:10.1088/2053-2563/aae894ch3

© IOP Publishing Ltd 2019

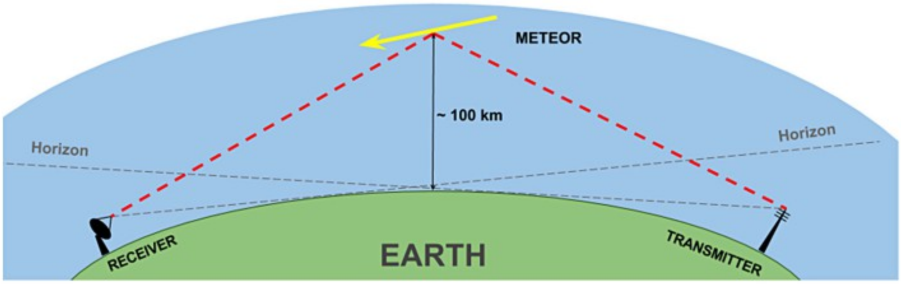

Figure 3.1. Sketch of the general principle of radio forward scatter observations.

underdense meteor echoes is a fraction of a second while overdense meteor echoes can typically last from ∼one second to a few tens of seconds.

Underdense meteor echoes have typical line electron densities $\alpha < 10^{13}\,\mathrm{e-m^{-1}}$ while overdense meteor echoes have $\alpha > 10^{15}\,\mathrm{e-m^{-1}}$. There are of course a lot of meteor echoes with intermediate electron line densities for which the physics is more complex. Some recent developments (e.g. [2]) are trying to model these meteor echoes, but this is beyond the goal of this chapter. From a physics point of view, the radio wave can penetrate underdense meteor trails and is scattered by individual electrons while for overdense meteor trails, the electron density is large enough that the dielectric constant of the medium becomes negative and the trail behaves as a plasma. The classical theory of McKinley [1] models the overdense meteor echoes as if the radio wave was reflected on the surface of an expanding metallic cylinder. As we will see below, this theory for overdense meteor echoes has limited applications as it is too simplistic and does not include additional phenomena.

One fundamental property of underdense meteor echoes is that the reflection of the radio wave is specular, which means that the majority of the received power reflected off the trail occurs when the so-called specular reflection point is created on the meteoroid path. The position of this reflection point is easy to determine as it is the point along the meteoroid path that is tangential to an ellipsoid whose foci are the transmitter and the receiver (see figure 3.2).

This property simply results from adding elemental small contributions from each electron individually along the path. The vast majority of the power actually comes from a small region centered on the specular reflection point and called the first Fresnel zone, whose size depends mostly on the meteor speed and on the wavelength of the radio wave. Contributions from other parts of the meteor trail can add constructively or destructively to this main signal, leading to so-called Fresnel oscillations in the signal.

An important consequence of the specular reflection is that a given system made of a transmitter and one receiver cannot detect all meteors. This is illustrated in figure 3.3. If the reflection point is located too high in the atmosphere (bottom case), the atmosphere is too thin at these altitudes such that the created ionization is not large enough to reflect a detectable amount of power. On the other hand, if the reflection point is located too low in the atmosphere (top case), the meteoroid might be completely ablated before this point is created and therefore no signal at all will be detected at the receiver. In figure 3.3, what is called the ‘meteor zone’ refers to a

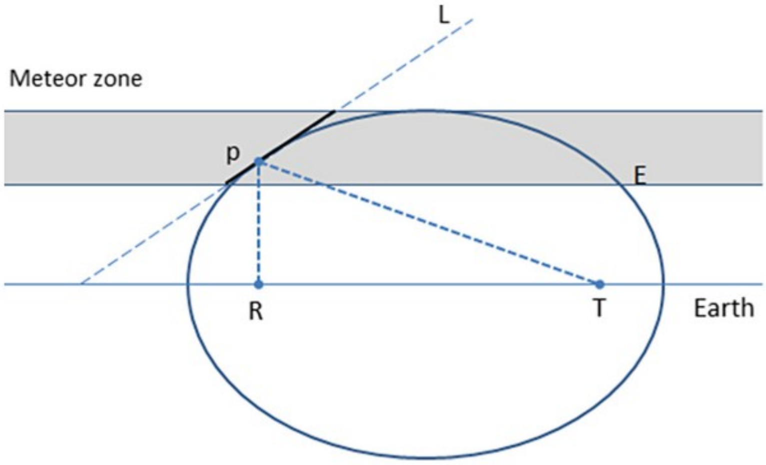

Figure 3.2. Illustration of the specularity condition. The specular reflection point *p* is the point along the meteoroid path *L* that is tangential to the ellipsoid *E* whose foci are the transmitter *T* and receiver *R*.

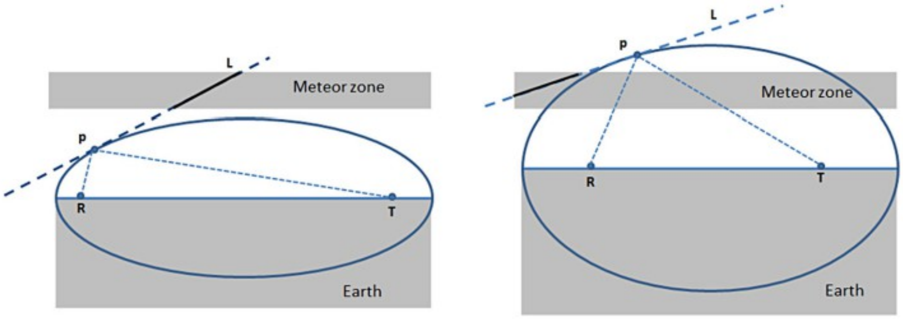

Figure 3.3. Consequences of the specularity condition. For given *L*, *T* and *R*, if the specular point *p* is too low (left) or too high (right), no signal will be recorded at *T*.

region in altitude where most meteor echoes should occur and can roughly be considered as 85–110 km.

The theory of forward scatter of radio waves has been developed by McKinley [1] as an extension of the classical theory for back scatter/radar systems. It provides formulas for the power profile, namely the amount of power in the meteor echo as a function of time.

For underdense meteor echoes, the power at the receiver is given by

$$P_R = \frac{P_T G_T G_R \lambda^3 r_e^2 \alpha^2 \sin^2(\gamma)}{16\pi^2 R_T R_R (R_T + R_R)(1 - \sin^2(\phi)\cos^2(\beta))} \exp\left[-\frac{8\pi^2 r_0^2}{\lambda^2 \sec\phi}\right] \exp\left[-\frac{32\pi^2 D_a t}{\lambda^2 \sec\phi}\right] \{C^2 + S^2\}, \tag{3.1}$$

where P_T is the power sent by the transmitter, G_T and G_R are the antenna gains for the transmitter and receiver in the directions to the reflection point, λ is the wavelength, r_e is the classical radius of the electron, α is the electron line density, γ is the polarization of the radio wave, R_T and R_R are the distances between, respectively, the transmitter/receiver and the specular reflection point, ϕ is half of the scattering angle of the radio wave and β is the inclination of the meteor trail with respect to the propagation plane of the radio wave. The geometrical parameters are illustrated in figure 3.4.

The first term in equation (3.1) is the peak value obtained when the specular reflection point is created and assuming all electrons lie on a straight line. The actual peak value is corrected by the initial radius r_0 that takes into account that trail electrons are not located on a single line but inside a cylinder, which reduces the strength of the signal and leads to the first exponential in equation (3.1). The second exponential depends on time t and is due to the diffusion of trail electrons in the neutral atmosphere. D_a is the ambipolar diffusion coefficient. The final terms in equation (3.1), C and S, are Fresnel integrals:

$$C = \int \cos \frac{\pi x^2}{2} dx \quad S = \sin \frac{\pi x^2}{2} dx, \tag{3.2}$$

where $x = 2s/\sqrt{R_F \lambda}$ is the Fresnel length, s is the distance along the trail measured from the specular reflection point, counted positively in the direction where the meteoroid is moving, and R_F is the size of the first Fresnel zone. This term is at the origin of the Fresnel oscillations which superimposes on the main signal. Indeed, every time a new Fresnel zone is created along the meteoroid path, it creates signals

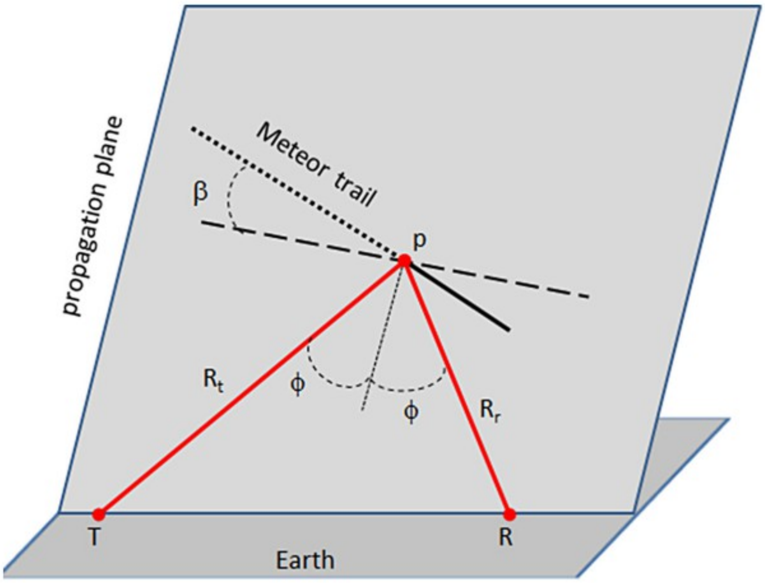

Figure 3.4. Geometrical parameters for forward scattering of radio waves.

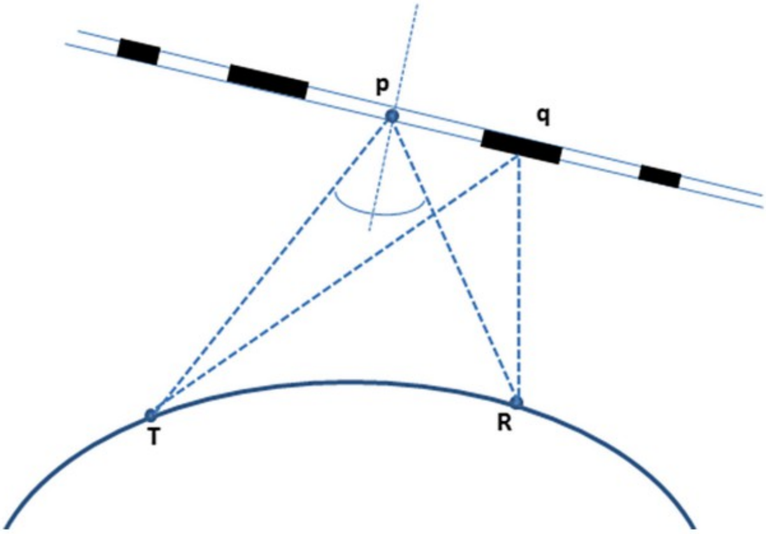

Figure 3.5. Illustration of Fresnel zones. The main Fresnel zone is centered on the specular reflection point *p*. Fresnel zones with constructive interference with the main signal are in white while those leading to destructive interference are in black.

which are alternatively in phase and out of phase with the main signal and leads to an increase/decrease of the total power. An example is given in figure 3.5 where the contribution from point q adds up to the main signal coming from point p in a destructive way, leading to a slight decrease of the power. Along a given meteoroid path and knowing the speed of the object, the positions of the Fresnel zones can be determined. In figure 3.5, the Fresnel zones leading to constructive interference with the main signal are depicted in white and those leading to destructive interference in black.

An example of an underdense meteor echo is shown in figure 3.6.

For overdense meteor echoes, the power at the receiver is given by

$$P_R = \frac{P_T G_T G_R \lambda^2 \cos\phi \sin^2\gamma}{32\pi^2 R_T R_R (R_T + R_R)(1 - \sin^2\phi \cos^2\beta)} r_c, \tag{3.3}$$

where r_c is the critical radius of the 'metallic' cylinder where the dielectric constant becomes negative,

$$r_c = \sqrt{r^2 \ln \frac{\alpha r_e \lambda^2 \sec^2\phi}{\pi^2 r^2}}, \tag{3.4}$$

where r is the radial distance perpendicular to the path. This model is rather poor as it neglects a number of phenomena, e.g. it does not take into account the effect of external parts of the cylinder which become increasingly underdense with time. More important, it neglects the influence of high-altitude mesospheric/thermospheric shear winds which can reach speeds from tens to a few hundreds of meters

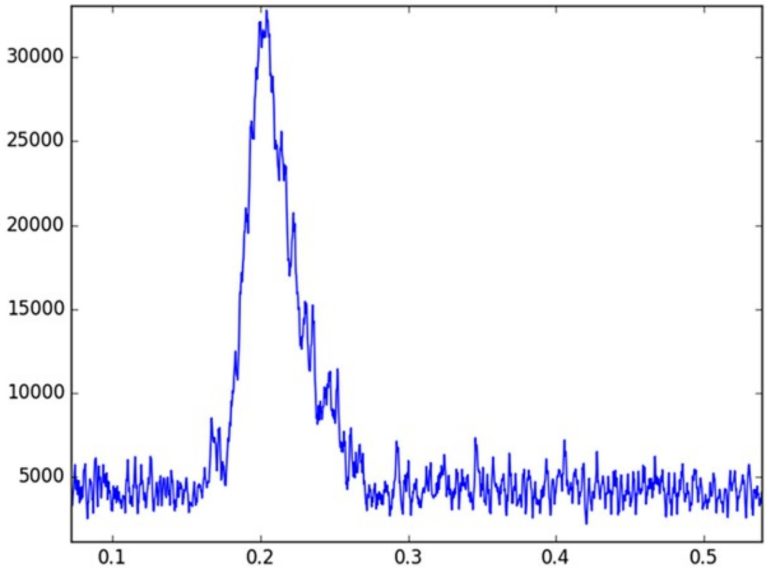

Figure 3.6. Typical underdense meteor echo. The vertical axis is power in arbitrary units while the horizontal axis is time in seconds.

per second. They can break the cylindrical trail into multiple parts and hence create multiple specular reflection points. The multiple reflections can then interfere constructively or destructively and lead to a much more complex power profile than the one modeled by equations (3.3) and (3.4).

An example of an overdense meteor echo is shown in figure 3.7.

Meteor trail echoes form the vast majority of meteor echoes detected by meteor radars such as, e.g., the Canadian Meteor Orbit Radar (CMOR, e.g. [3]) or by the Belgian Radio Meteor Stations (BRAMS) network that will be described in detail in the next section. Another type of meteor echo, mostly observed with larger objects, is the head echo which occurs when the incoming radio wave is reflected upon the ionized region forming in front of the meteoroid. This region moves at the same high speed ($>\sim 10\ \mathrm{km\ s^{-1}}$) as the meteoroid itself while the speed of the meteor trail depends mostly on the speed of high-altitude mesospheric/thermospheric winds. Consequently, the Doppler effects associated with these two types of meteor echoes are very different and allow one to easily discriminate them. Head echoes will not be discussed in detail in this chapter. Only a few examples of observations will be given in the section about optical versus radio observations of meteors.

3.2 The BRAMS project

3.2.1 The BRAMS network

BRAMS is a Belgian radio network using forward scatter techniques to detect and study meteoroids. It uses a dedicated transmitter located at the Geophysical Center in Dourbes in the south of Belgium and 26 identical receiving stations spread all over

!F"." G 4 4 !) * ! @!1 0 ' D 0 . 0 /
& '4 ' ' 9 -/ 0 9

= !"#$! ! !% & ' ! << 1 0 ' D 0 0-

!F"." G 4 4 !) * ! @!: ' - D U : ' > '
" 7 / / . , 9 * ' -

= !"#$! ! !% & ' ! <E : ' - D U : '
> ' 9

!F"." G 4 4 !) * ! @! - 6 0' ' - '
'

= !"#$! ! !% & ' ! <
- (6 0' (' -(' (' 0-
V @E 0'&

I& 8
4' :' ' - 7 & A 0 -

$ ' ' 4 / - ' - & 0' ' - 0 ' L
0 / ' ' ' 5 5 ' - &/ 4' 6 '0 5
4 ' ' ' ' & ' 5 ' - ' 5
0 ' - '9 / - 5 ' - ' ' & 9 ' 5 ' &
0 0 - ' '- / 4 & . ' ' -
7 & 2 0 - . L ' ' ' ' / -/ ' '4
' - 0' 4 " - ' &' - ' ' 0 ,' 0 ' L ' '
' - 4 -/ ' ' 0 ' - ' # L - 9 -' /
- ' - -

4' :' D ' # + '
6 0 / ' - ' 5 L ' 0 '/ ' 0 4 ' ' ' ' /
- 9 & 9 '4 ' / 4' ' 5 4'
5 ' - 4' &/ . '0 0 9' ' L '
' D ' '9 ' / ' "00 - , " ' 9' *& - 4 ' -
' . D ' ' 5 4 - 9 5 9 -
, 4 /
) & 0 \ L ' 5 0 ' ' / - \
' - ' 0 + 5 '2 / 9 *5 L ' '
\& ' ' - - ' 8 D '
'9 ' / ' 5 /0 ' / - - ' ,' 0 5 ,5 /5
35 , 5 , 5 , . ' 0 /0 ' / '- " 0 9 /
L ' ' \& ' ' - - ' 8
:'& ' & ' ' 5 /0 ' / - - 00 ' ,' 0 5 W5
H5 S5 W 5 W 5 W 5 - ,' '9 ' - 0
8 '0 + ' ' 5 ' - 0 ' 5 ' ' 9 '4
- 0 ' 7 4 5 ' & - 0 ' 5 ' & ' 9 '4
5 - 0 ' . '0 5 5 D '
'9 ' / ' 5 /0 ' ' - 0 / ' ' /
" / '00 ' ' 9 & & ' '
4' ' 5 - 4 0 - ' 4 4' '

L - '9 / 4' 4' 5 ' - - ' 4
- 4 05 & & ' 5 * 2 0 -
4' ' / - '9 9 ' 4 6 0 9
- 4 0 * 2 0 - - / & ' 4' '
4' ' 5 ' - - 0 ' p 4 5 , 4 /
- ' ' 4' ' 5 - 5

!F"." G 4 4 !) * ! @! ' * . - " - &

= !"#$! ! !% & ' ! E ' * . -
" - & 0-

< E E

E @

!F"." G 4 4 !) * ! <! ' ' ' F 4 %0 '

!"#$! ! !% & ' ! E ' ' ' F 4 %0 '
0-

!F"." G 4 4 !) * ! <! -' . 0 &/ " ' & '
"00 ' % - 5 5 ' - . 0 &/

!"#$! ! !% & ' ! E -' . 0 &/ " ' & '
"00 ' % - 5 5 ' - . 0 &/ 0-

!F"." G 4 4 !) * ! <!# &K K L ' K
- 0' ' n >'

!"#$! ! !% & ' ! E '0 L ' 0-
V K - >K K '
V E< 0'&
%9 : 4 - ;;;

I& =8 ' - U' &

I& E8 ' L '- ' - 5

!"#$! ! !% & ' ! E '0 L ' 0-

= !"#$! ! !% & ' ! E = =@ < 0-
V . - ' & /

@ !"#$! ! !% & ' ! E 0-
V . - ' & / 4

< !"#$! ! !% & ' ! E & ' - 0-
V > . ' ' F '

E !"#$! ! !% & ' ! E & ' - 0-

V > . ' ' F '

!"#$! ! !% & ' ! E - ' ' 0-
V >K K G D - : -
V @ 0'&
%9 : 4 - ;;;

I& E8 '0 W66

- - ' & -U ' &
0
E FK % '00 , n -U ' &
9 9 ' ' - ' -
- ' & :U , n -U ' & 2 0 5
2 2 K&' 4' L U '-
U /0 n - U' & 9 5 - U' & - -
U' & ' &
. K n , n - ' & K&' l '
- , n - ' & -' L ' & - K 0' '&K
0' ' - 2 & ' - l 0 - j K 00 K
FK ' ") ' & L L
F 0 K ") % '

!"#$! ! !% & ' ! E G D - ' (> /
) ' 0-
%9 : 4 - ;;; 6 0 ' M ;;;
V 4 ' M

%9 -U 0' ' ' - '

. - ' 0' ' ' - ' -

. - 0 - 0 2 4 & / D -U '

Table of Contents.

Chapter I.

The Attempts to prove Euclid's Parallel Postulate.

pages
§ 1—5. The Greek Geometers and the Parallel Postulate .. .. 1—9
§ 6. The Arabs and the Parallel Postulate .. .. 9—12
§ 7—10. The Parallel Postulate during the Renaissance and the 17th Century .. .. 12—21

Chapter II.

The Forerunners of Non-Euclidean Geometry.

§ 11—17. Gerolamo Saccheri (1667—1733) .. .. 22—44
§ 18—22. Johann Heinrich Lambert (1728—1777) 44—51
§ 23—26. The French Geometers towards the End of the 18th Century .. .. 51—55
§ 27—28. Adrien Marie Legendre (1752—1833) .. 55—60
§ 29. Wolfgang Bolyai (1775—1856) .. .. .. 60—62
§ 30. Friedrich Ludwig Wachter (1792—1817) .. 62—63
§ 30 (bis) Bernhard Friedrich Thibaut (1776—1832) 63

Chapter III.

The Founders of Non-Euclidean Geometry.

§ 31—34. Karl Friedrich Gauss (1777—1855) .. 64—75
§ 35. Ferdinand Karl Schweikart (1780—1859) .. 75—77
§ 36—38. Franz Adolf Taurinus (1794—1874) .. 77—83

Chapter IV.

The Founders of Non-Euclidean Geometry (Cont.).

§ 39—45. Nicolai Ivanovitsch Lobatschewsky (1793—1856) .. .. 84—96
§ 46—55. Johann Bolyai (1802—1860) .. .. .. 96—113
§ 56—58. The Absolute Trigonometry .. .. .. 113—118
§ 59. Hypotheses equivalent to Euclid's Postulate .. 118—121
§ 60—65. The Spread of Non-Euclidean Geometry 121—128

Chapter V.

The Later Development of Non-Euclidean Geometry.

§ 66. Introduction 129

pages
Differential Geometry and Non-Euclidean Geometry.
§ 67—69. Geometry upon a Surface .. .. 130—139
§ 70—76. Principles of Plane Geometry on the Ideas of RIEMANN .. .. 139—150
§ 77. Principles of RIEMANN's Solid Geometry .. .. 151—152
§ 78. The Work of HELMHOLTZ and the Investigations of LIE .. .. 152—154
Projective Geometry and Non-Euclidean Geometry.
§ 79—83. Subordination of Metrical Geometry to Projective Geometry .. .. 154—164
§ 84—91. Representation of the Geometry of LOBATSCHEWSKY-BOLYAI on the Euclidean Plane .. .. 164—175
§ 92. Representation of RIEMANN's Elliptic Geometry in Euclidean Space .. .. 175—176
§ 93. Foundation of Geometry upon Descriptive Properties .. .. 176—177
§ 94. The Impossibility of proving Euclid's Postulate 177—180
Appendix I.
The Fundamental Principles of Statics and Euclid's Postulate.
§ 1—3. On the Principle of the Lever .. .. .. 181—184
§ 4—8. On the Composition of Forces acting at a Point .. .. 184—192
§ 9—10. Non-Euclidean Statics .. .. 192—195
§ 11—12. Deduction of Plane Trigonometry from Statics .. .. 195—199
Appendix II.
CLIFFORD's Parallels and Surface. Sketch of CLIFFORD-KLEIN's Problem.
§ 1—4. CLIFFORD's Parallels .. .. 200—206
§ 5—8. CLIFFORD's Surface .. .. 206—211
§ 9—11. Sketch of CLIFFORD-KLEIN's Problem .. 211—215
Appendix III.
The Non-Euclidean Parallel Construction and other Allied Constructions.
§ 1—3. The Non-Euclidean Parallel Construction .. 216—222
§ 4. Construction of the Common Perpendicular to two non-intersecting Straight Lines .. .. .. 222—223
§ 5. Construction of the Common Parallel to the Straight Lines which bound an Angle .. .. .. 223—224

pages
§ 6. Construction of the Straight Line which is perpendicular to one of the lines bounding an acute Angle and Parallel to the other 224
§ 7. The Absolute and the Parallel Construction .. 224—226

Appendix IV.

The Independence of Projective Geometry from Euclid's Postulate.

§ 1. Statement of the Problem .. .. 227—228
§ 2. Improper Points and the Complete Projective Plane .. .. 228—229
§ 3. The Complete Projective Line 229
§ 4. Combination of Elements .. .. 229—231
§ 5. Improper Lines .. .. 231—233
§ 6. Complete Projective Space 233
§ 7. Indirect Proof of the Independence of Projective Geometry from the Fifth Postulate .. 233—234
§ 8. BELTRAMI's Direct Proof of this Independence 234—236
§ 9. KLEIN's Direct Proof of this Independence .. 236—237

Appendix V.

The Impossibility of proving Euclid's Postulate.

An Elementary Demonstration of this Impossibility founded upon the Properties of the System of Circles orthogonal to a Fixed Circle.

§ 1. Introduction 238
§ 2—7. The System of Circles passing through a Fixed Point .. .. 239 - 250
§ 8—12. The System of Circles orthogonal to a Fixed Circle .. .. 250—264

Index of Authors 265

The Science of Absolute Space
and
The Theory of Parallels_____________follow page 268

!"#$! ! !% & ' ! E - ' & 9 0-
V G%G D? :6FD"G >D% D.#H
" #6.6 ": "GF
76 .%#6 ": .?FH %+ 6. FD D:% DG.
)H
#%)D#.%)%G%:"
? 4 / '4 '

I& 6 8

6 D & ' ' 5) '
* - / 0 - - '* 5 6 '4 - -
' & '- > ' ' ' m + 0
- 6 - ,0 / ' *
) > . 9 ' - : 9 ' - '9
' ' ' ' 9 0 ' - / ' - 9/
) '5 - & ' & / ' - - 0'
' U *5 ' "00 - , 6 - 0 -
2 4 > / ' ' ' 5 ' -
G D - ' ' '

!"#$! ! !% & ' ! E - ' & 9 0-
V) '

!"#$! ! !% & ' ! E 0 0 -' ' 0-
V #6G 6 D +%GF" DG."?W
FD :"
>a% a.#6D ! D? :6F6DGGD
FD #6D "GG

!"#$! ! !% & ' ! E & 9' 0-
V E 0'&
V G 6 :%)". 7DB 1H
'- A 7%?D:
&K K L ' K -
0' ' n 5 4 -U , ' - ' 0 -'
>"? 7? " 7D#
7 4 7D: %:.S
'- 7%?D:
' L 4 - 9' I ' >K K

= !"#$! ! !% & ' ! E &' & -2 ' 0-
V) /'

@ !"#$! ! !% & ' ! E &' & -2 ' 0-
V) /'

!F"." G 4 4 !) * ! <!# &K K L ' K
- 0' ' n >' ! '

< !"#$! ! !% & ' ! E ' (6 ' 0-

E !"#$! ! !% & ' ! E ' (66 ' 0-

!"#$! ! !% & ' ! E ' :' 9 0-
V ==@ 0'&
V " ' / ' '

!"#$! ! !% & ' ! E 4 (+ (+ ( ) 0-

!"#$! ! !% & ' ! E 4 (+ (+ ( ) 0-

!"#$! ! !% & ' ! E 4
' / ' 0 0-
V "G
D:D DG."#H .#D".6 D
?# D 5 +?G .6%G 5 "GF +%# D

!"#$! ! !% & ' ! E 4
' / ' 0 0-

!"#$! ! !% & ' ! E (' ' 7' -/ 0-
V =E< 0'&
%9 : 4 - ;;;

!F"." G 4 4 !) * ! <!# &K K L ' K
- 0' ' n >' !D -

= !"#$! ! !% & ' ! E= (-(' (-(- 0-

@ !"#$! ! !% & ' ! E= (-(' (-(- 0-

< !"#$! ! !% & ' ! E= (-(' (-(- 0-

E !"#$! ! !% & ' ! E= D -(' & (?0 + 30' * 0-

!"#$! ! !% & ' ! E= D4 / - (D -(6(66 7' 0-

!"#$! ! !% & ' ! E= D4 / - (D -(666(6
7' 0-

!"#$! ! !% & ' ! E= D4 / - (D -((6
7' 0-

!F"." G 4 4 !) * ! <!# &K K L ' K
- 0' ' n >' !+ '4 & /

!"#$! ! !% & ' ! E@ > / 0 & 8 -24
V) &

!"#$! ! !% & ' ! E@ > / 0 & 8 0-

!"#$! ! !% & ' ! E@ 4 : 4/ + '4 > /
' 9 -& ? 4 / 8 < 0-

%9 : 4 - ;;; 6 ' ;;;

= !"#$! ! !% & ' ! E@ 4 : 4/ 9 & /
0 & '& 8 E -24

@ !"#$! ! !% & ' ! E@ 4 : 4/ 9 & /
0 & '& 8 E 0-

< !"#$! ! !% & ' ! E@ 4 : 4/ . F '
> / ' - . 0 &/ ? 4 / 8 < -24
V . F ' > /
' - . 0 &/
{ %:? D
{B ' .
{DF6.DF)H 6: 6% :D H

E !"#$! ! !% & ' ! E@ 4 : 4/ . F '
> / ' - . 0 &/ ? 4 / 8 < 0-

!"#$! ! !% & ' ! E@ . D & - B'/) ' /
1] $ ' 4 ' 9 -& ? 4 / 8 0-
%9 : 4 - ;;;

!"#$! ! !% & ' ! E@ . F ' > / ' -
. 0 &/5 ? 4 / 8 < 0-
%9 : 4 - ;;;

I& < 8 E

(c) Show that if for every t there is some r such that the r-th derivative of β is non-zero, the conclusion still holds, except that the lift may not be immersed.

(d) What can you say in general?

Proposition 3.7.5 (tangent circle bundle has contact structure). *The tangent circle bundle of a smooth surface has a canonical contact structure, which is preserved by the derivative of any diffeomorphism.*

Proof of 3.7.5. We first look at $\mathbf{R}^2 \times S^1$, the tangent circle bundle of $\mathbf{R}^2$. This space is a double cover of the contact manifold $\mathbf{R}^2 \times \mathbf{RP}^1$ that we defined after Example 3.7.3—the fiber of the tangent circle bundle is a circle of unit tangent vectors, rather than a circle of directions. We give $\mathbf{R}^2 \times S^1$ the induced contact structure. Diffeomorphisms of $\mathbf{R}^2$ lift to contact automorphisms of $\mathbf{R}^2 \times S^1$, just as they lift to automorphisms of $\mathbf{R}^2 \times \mathbf{RP}^1$ in Example 3.7.3. Therefore, given an atlas for a surface, the contact structures on the tangent circle bundles of the chart domains patch together to give a contact structure on the tangent circle bundle of the surface. $\boxed{3.7.5}$

Example 3.7.6. As a corollary, T^3, $\mathbf{RP}^3$ and S^3 have contact structures. T^3 is diffeomorphic to the tangent circle bundle of the two-torus. $\mathbf{RP}^3$ is diffeomorphic to SO(3) (Exercise 2.7.7), which in turn can be identified with the tangent circle bundle of S^2 (why?). Therefore $\mathbf{RP}^3$ has a contact structure. Since S^3 is the double cover of SO(3), it also has a contact structure.

In fact, Martinet [Mar71] proved that every orientable compact three-manifold has a contact structure.

You can get a good physical sense for the contact structure on the tangent circle bundle of a surface by thinking about ice skating, or bicycling. A skate that is not scraping sideways describes a Legendrian curve in the tangent circle bundle to the ice. It can turn arbitrarily, but any change of position is in the direction it points. Likewise, as you cycle along, the direction of the bicycle defines a ray tangent to the earth at the point of contact of the rear wheel. Assuming you are not skidding, the rear wheel moves in the direction of this ray, and its motion describes a Legendrian curve in the tangent circle bundle of the earth.

Young children are sometimes given bicycles with training wheels, some distance off to the side of the rear wheel. The training wheel also traces out a Legendrian curve—in fact, for any real number t, the diffeomorphism ϕ_t of $\mathbf{R}^2 \times S^1$ that takes a tangent ray a signed distance t to the left of itself is a contact automorphism. The training wheel path is the image of the rear wheel path under such a transformation. Note that this transformation applied to curves in the plane often creates or removes cusps (Figure 3.25). The same thing happens when you mow a lawn, if you start by making a big circuit around the edge of the lawn and move inward.

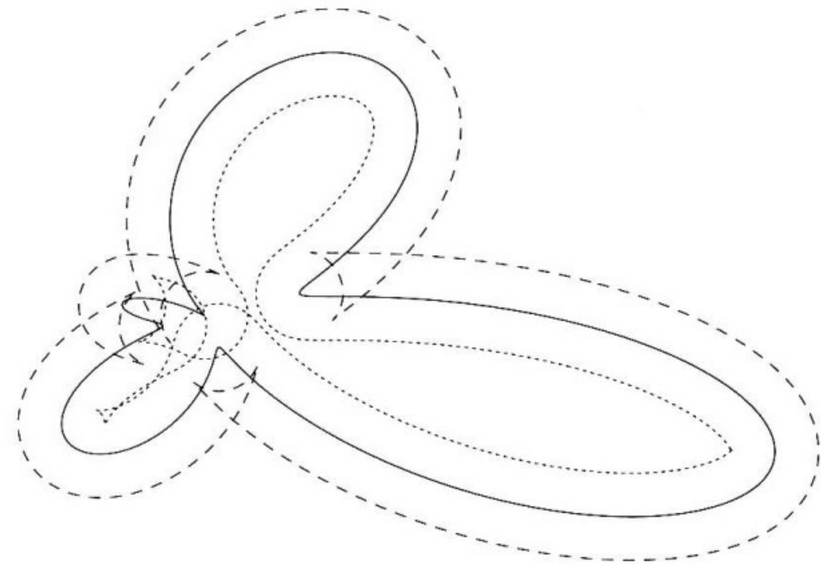

Figure 3.25. A family of parallel curves. The transformation ϕ_t of the tangent circle bundle of the plane which translates a ray a distance t to its left is a contact automorphism. The figure shows the projections to the plane of the images of a single curve, for various values of t. Notice how cusps develop.

The rear wheels of an automobile also follow Legendrian curves, related by a transformation ϕ_t. (The two wheels turn at slightly different speeds, thanks to the differential gear; otherwise one or both would skid whenever the car went around a curve.) Parallel parking implements the fact that any two points are connected by a Legendrian path. The front wheels are more complicated, since the line joining them is not perpendicular to their path unless the car is moving straight. The relation between the front and back wheels of a bicycle or a car is also complicated: the rear wheels depend not just on the current position and direction of the front wheels, but on their history.

The contact automorphism ϕ_t can be defined not only for $\mathbf{R}^2$, but for any oriented surface with a Riemannian metric. It sends each tangent ray to a tangent ray at a point obtained by translating it along the perpendicular geodesic, toward the left. (The surface needs to be oriented so we can decide consistently which way is left.) As t varies, the automorphisms ϕ_t form a flow ϕ. This is closely related to the surface's *geodesic flow*. The geodesic flow of a Riemannian manifold is the flow ψ that moves each unit tangent vector along its geodesic at unit speed. Really ϕ and ψ are the same flow in slightly different guises, identified by the map that rotates each unit tangent vector counterclockwise by 90°.

Exercise 3.7.7. (a) Proposition 3.7.5 gives one canonical plane field for the tangent circle bundle of a smooth surface. If the surface has a Riemannian metric, there is another: the Levi-Civita connection, restricted from the tangent bundle to the tangent sphere bundle. Compare the two, and explain the relation between the Levi-Civita connection and the geodesic flow (or the training-wheel flow) on the surface. See also Corollary 3.7.16.

(b) Describe the contact structures of Example 3.7.6 geometrically.

(c) What is the relationship among the contact structure on S^3, the Hopf flow, and Levi-Civita connection on S^2?

Problem 3.7.8 (Legendrian foliations). In a contact three-manifold, it is easy to construct, at least locally, a one-dimensional Legendrian foliation, that is, a foliation by Legendrian curves: in a small neighborhood, any nonzero section of τ is a vector field whose flow lines give such a foliation.

(a) Show that every Legendrian foliation is locally equivalent to the foliation by lines parallel to the x-axis in our model contact structure in $\mathbf{R}^3$.

(b) Work out an interpretation of contact three-manifolds with Legendrian foliations as three-manifolds locally modeled on the tangent circle bundle of $\mathbf{R}^2$, up to derivatives of diffeomorphisms.

(c) Find and analyze some interesting examples.

Let's return to our model plane field τ in $\mathbf{R}^3$. Let X be any smooth vector field tangent to τ. We claim that unless X is identically zero, the flow it generates cannot consist of contact automorphisms. For consider a short Legendrian arc α transversal to X at a point p. Under the flow, α sweeps out a quadrilateral, three sides of which

I& < 8 E

Show that in a contact three-manifold, most foliations transverse to the contact plane field are *not* locally equivalent to this foliation. (Compare Example 3.7.2 and Problem 3.7.8.)

3.8. The Eight Model Geometries

What is a geometry? Up till now, we have discussed three kinds of three-dimensional geometry: hyperbolic, Euclidean and spherical. They have in common the property of being as uniform as possible: their isometries can move any point to any other (homogeneity), and can take any orthonormal frame in the tangent space at a point to any other orthonormal frame at that point (isotropy). There are more possibilities if we remove the isotropy condition, allowing the space to have a grain, so to speak, so that certain directions are geometrically distinguished from others.

An enumeration of additional three-dimensional geometries depends on what spaces we wish to consider and what structures we use to define and to distinguish the spaces. For instance, do we think of a geometry as a space equipped with such notions as lines and planes, or as a space equipped with a notion of congruence, or as a space equipped with either a metric or a Riemannian metric? There are deficiencies in all of these approaches.

The problem with using lines and planes is that they aren't general enough. The five new geometries described in this section don't have any good notion of a plane—there are no totally geodesic surfaces passing through certain tangent planes. Besides, even in Euclidean geometry, information about geometric shapes is not determined by incidence properties of lines and planes.

Using congruence as the essence of the definition leads to an excessive proliferation of geometries. For instance, we would get different geometries by considering Euclidean space with the group of congruences consisting of translations, or the one consisting of horizontal translations together with vertical screw motions (where the amount of rotation is proportional to the vertical motion), and so on. Sometimes it is interesting to distinguish these different structures, but for the broad picture these variations should all be considered under the category of Euclidean geometry.

Using distance to determine the geometry is likewise unsatisfactory. First, by simply rescaling something like $\mathbf{H}^3$ or S^3, we get different metric spaces. Even if we consider as equivalent all metric spaces that are isometric up to a constant scaling factor, many of the spaces have a whole family of homogeneous metrics (sometimes with varying degrees of homogeneity) which are not equivalent up to scaling. The three-sphere, for instance, has an interesting family of homogeneous metrics obtained from the usual metric by picking a Hopf fibration and contracting or expanding the lengths of the circles, while keeping the metric constant in the orthogonal directions.

The best way to think of a geometry, really, is to keep in mind these different points of view all at the same time. If we regard changes of the group of congruences that do not change the metric and changes of the metric that do not change the group of congruences as inessential changes in the geometry, and if we also consider two geometries the same when the sets of compact manifolds modeled on them are identical, we end up with a reasonable enumeration of geometries.

For logical purposes, we must pick only one definition. We choose to represent a geometry as a space equipped with a group of congruences, that is, a (G, X)-space.

Definition 3.8.1 (model geometry). A *model geometry* (G, X) is a manifold X together with a Lie group G of diffeomorphisms of X, such that:

(a) X is connected and simply connected;
(b) G acts transitively on X, with compact point stabilizers;
(c) G is not contained in any larger group of diffeomorphisms of X with compact stabilizers of points; and
(d) there exists at least one compact manifold modeled on (G, X).

Condition (a) selects one representative from each class of locally equivalent geometries having different fundamental groups, since locally equivalent geometries are models for identical classes of manifolds (Exercise 3.3.1). Condition (b) means that the space possesses a homogeneous Riemannian metric invariant by G (Lemma 3.4.11), and that it is complete (see Proposition 3.4.15 and the paragraph following its proof). Condition (c) says that no Riemannian metric invariant by G is also invariant by any larger group. In particular, it selects at most one geometry for each isometry class of metric spaces. Another reason for condition (c) is that by enlarging the structure group G, we do not decrease the set of manifolds with that structure. Condition (d) is not phrased in an intrinsic way, but it is useful because it eliminates a whole continuous family of three-dimensional geometries that do not serve as models for any compact manifolds.

Theorem 3.8.2 (two-dimensional model geometries). *There are precisely three two-dimensional model geometries: spherical, Euclidean and hyperbolic.*

Proof of 3.8.2. Since G acts transitively on X, it follows that any G-invariant Riemannian metric on X has constant Gaussian curvature. When a metric is multiplied by a constant k, the Gaussian curvature is multiplied by k^2, so we can find a metric whose curvature is either 0, 1 or -1. It is a standard fact from Riemannian geometry that the only simply connected complete Riemannian n-manifolds with constant sectional curvature 0, 1 and -1 are $\mathbf{E}^n$, S^n and $\mathbf{H}^n$ (see [dC92, p. 163] or [GHL90, p. 135], for example). $\boxed{3.8.2}$

For an alternate proof, see Exercise 3.8.11.

Exercise 3.8.3. The similarities of $\mathbf{R}$ form a two-dimensional group $\mathrm{Sim}(1)$, which acts on itself by left multiplication. Why isn't $(\mathrm{Sim}(1), \mathrm{Sim}(1))$ a two-dimensional model geometry? (Hint: see Exercise 2.2.10.)

In enumerating three-dimensional model geometries (G, X), we will first look at the connected component of the identity of G—call it G'. The action of G' is still transitive, and the stabilizers G'_x of points $x \in X$ are connected. This is because the quotients $G'_x/(G'_x)_0$, where $(G'_x)_0$ is the component of the identity of G'_x, form a covering space of X. Since X is simply connected, the covering is trivial.

Therefore G'_x is a connected closed subgroup of $\mathrm{SO}(3)$. Using the fact that a closed subgroup of a Lie group is also a Lie group, and therefore a manifold, it is easy to see that there are only three possibilities: $\mathrm{SO}(3)$, $\mathrm{SO}(2)$ and the trivial group. The stabilizer G_x is a Lie group of the same dimension.

Theorem 3.8.4 (three-dimensional model geometries). *There are eight three-dimensional model geometries (G, X), as follows:*

(a) *If the point stabilizers are three-dimensional, X is S^3, $\mathbf{E}^3$ or $\mathbf{H}^3$.*

(b) *If the point stabilizers are one-dimensional, X fibers over one of the two-dimensional model geometries, in a way that is invariant under G. There is a G-invariant Riemannian metric on X such that the connection orthogonal to the fibers has curvature 0 or 1.*

  (b_1) *If the curvature is zero, X is $S^2 \times \mathbf{E}^1$ or $\mathbf{H}^2 \times \mathbf{E}^1$.*

  (b_2) *If the curvature is 1, we have nilgeometry (which fibers over $\mathbf{E}^2$) or the geometry of $\widetilde{\mathrm{SL}}(2, \mathbf{R})$ (which fibers over $\mathbf{H}^2$).*

(c) *The only geometry with zero-dimensional stabilizers is solvegeometry, which fibers over the line.*

The geometries in (a) we have already discussed extensively, and those in (b_1) are self-explanatory. The remaining ones will be described in more detail in the course of the proof. We start by giving X a G-invariant Riemannian metric (see the discussion after Definition 3.8.1).

Proof of 3.8.4(a). If G' acts with stabilizer SO(3), any tangent two-plane at any point can be taken by G to any tangent two-plane at any other point, so the metric has constant sectional curvature. As in the two-dimensional case, it follows that the geometry is spherical, Euclidean, or hyperbolic.

The full group of isometries G contains G' with index 2, and can be obtained by adding any orientation-reversing isometry. $\boxed{\text{3.8.4(a)}}$

Proof of 3.8.4(b). If G' acts with stabilizer SO(2), there is a nonzero, G'-invariant vector field V on X whose direction at each point gives the axis of rotation of the elements of G' that fix that point. The trajectories of V form a G'-invariant one-dimensional foliation $\mathcal{F}$. Also, the flow of V—call it ϕ_t at time t—commutes with the action of G', so if an element of G' fixes some point on a leaf F of $\mathcal{F}$, it fixes any other point on F: all points on the same leaf have the same stabilizer. This also implies that if an element of G' takes a point $x \in F$ to another point $y \in F$, it commutes with any element of the stabilizer $G'_x = G'_y$.

Now fix a leaf F and a point $x \in F$, and let g_t be an element of g taking $\phi_t(x)$ back to x. Then $g_t \circ \phi_t$ fixes x, and its derivative at x is a linear automorphism of T_xM. The derivative is the identity along the axis of the action of G'_x. It commutes with rotations around this axis, that is, with elements of G'_x. Then it must be itself a rotation around this axis, possibly composed with an expansion or contraction. But an expansion or contraction is ruled out, because the assumption that there is a compact manifold modeled on (G, X) implies that V must preserve volume:

Exercise 3.8.5 (divergence). The *divergence* of a vector field V on a manifold X with a volume form ω is a measure of how much V expands or contracts volume. More precisely, $\operatorname{div} V$ is the Lie derivative $L_V\omega$ (Problem 3.1.16), expressed in units of ω.

Now suppose that X is a manifold on which a Lie group G acts transitively, and that V and ω are a vector field and a volume form on X, both invariant under G. Show that $\operatorname{div} V$ is constant over X.

In the situation of the proof, if there is a compact manifold modeled on (G, X), this manifold inherits the vector field and the volume form from X. The vector field must preserve the total volume, and so must preserves volume at every point. Therefore V has divergence zero. Show that this implies that $g_t \circ \phi_t$ acts as a rotation on T_xM.

We conclude that the derivative of ϕ_t maps T_xM to $T_{\phi_t(x)}M$ isometrically. Since x was arbitrary, the flow of the vector field V is by isometries.

By considering a neighborhood of a point on a leaf and the fact that the leaf is invariant under a subgroup G'_x isomorphic to SO(2), we conclude that the leaf does not accumulate on itself, but is an embedded image of either S^1 or $\mathbf{R}$. In fact, it is easy to see that distinct leaves have disjoint neighborhoods. Therefore the quotient space $X/\mathcal{F}$ is a two-dimensional manifold Y. Since V acts by isometries, Y inherits a Riemannian metric from X (just ignore the component of the metric of X in the direction of the leaves), and a transitive action of G' by isometries. Also, Y is connected and simply connected because X is. By the proof of Theorem 3.8.2, Y must be one of the two-dimensional model geometries: $\mathbf{E}^2$, S^2 or $\mathbf{H}^2$. In addition, X is a principal fiber bundle over Y, with fiber and structure group equal to S^1 or $\mathbf{R}$.

The plane field τ orthogonal to $\mathcal{F}$ is a connection for this bundle. Since the group of isometries of X acts transitively, τ has constant curvature.

(b$_1$) If the curvature is zero, τ defines a foliation. Since Y is simply connected, the bundle is trivial by Proposition 3.6.15. There are three possibilities, depending on Y (an open circle indicates that no new geometry arises from this possibility):

- • If $Y = S^2$, we obtain the model geometry $S^2 \times \mathbf{E}^1$. As a compact manifold modeled on this geometry, we can take $S^2 \times S^1$ (see also Exercise 4.7.1).
- ○ If $Y = \mathbf{E}^2$, then $X = \mathbf{E}^2 \times \mathbf{E}^1 = \mathbf{E}^3$. Thus G' (and hence G) is contained in a bigger group of isometries, and we don't get a new model geometry.
- • If $Y = \mathbf{H}^2$, we obtain the model geometry $\mathbf{H}^2 \times \mathbf{E}^1$. Any compact hyperbolic surface cross a circle is an example.

In each of these two new geometries, the full group of isometries G contains G' with index 4, since we can reverse the orientation of either factor independently.

(b_2) If the curvature of τ is non-zero, τ defines a contact structure. After rescaling our metric in the direction of the fibers and choosing appropriate orientations for the base and the fiber, we can assume that the curvature is 1, expressed in terms of the standard bases for $\bigwedge^2 TY$ and TF. This, together with the condition that X is simply connected, essentially determines the geometry. If Y has non-zero curvature, X can be taken as the tangent circle bundle of Y (or rather, its universal cover) with the Levi-Civita connection (Corollary 3.7.16). The group is made of derivatives of isometries of Y, together with rotations of unit tangent vectors keeping the base point fixed.

- ○ If $Y = S^2$, the tangent circle bundle is SO(3), whose universal cover is S^3 (see Example 3.7.6). For G, we get the group of isometries of S^3 that preserve the Hopf fibration (Exercise 3.7.7(c)). This is not a maximal group acting with compact stabilizers, so it is not a model geometry.
- If $Y = \mathbf{E}^2$, we obtain *nilgeometry*. This can be defined in terms of our model contact structure τ of Section 3.7 (page 168) as the group of contact automorphisms that are lifts of isometries of the xy-plane (see Example 3.7.2). Exercise 3.8.6(e) shows that there exist compact manifolds modeled on this geometry.
- If $Y = \mathbf{H}^2$, the unit tangent bundle is $\mathrm{PSL}(2, \mathbf{R})$, the group of orientation-preserving isometries of $\mathbf{H}^2$ (see Exercise 2.6.6). Passing to the universal cover, we get $X = \widetilde{\mathrm{SL}}(2, \mathbf{R})$. The unit tangent bundle of a compact hyperbolic surface is an example of a three-manifold with this geometry.

For $\widetilde{\mathrm{SL}}(2, \mathbf{R})$ and nilgeometry, the contact structure determines an orientation of the geometry which cannot be reversed. However, the orientation of the base two-dimensional geometry can be reversed simultaneously with the orientation of the fiber, so the index of G' in G is 2. $\boxed{\text{3.8.4(b)}}$

Exercise 3.8.6 (isometries of nilgeometry). Let G be the group of isometries of nilgeometry.

(a) Compute the difference in elevation between the endpoints of a Legendrian curve for τ whose projection to the xy-plane $\mathbf{E}^2$ is a line segment with endpoints (x_0, y_0) and (x_1, y_1). (Answer: $\frac{1}{2}(x_1 + x_0)(y_1 - y_0)$.)

(b) Using this and the construction in Example 3.7.2, write down the formula for the elements of G that project to a given isometry of $\mathbf{E}^2$. (Remember that all such isometries differ by a vertical translation.)

(c) As a special case, an isometry that projects to a translation of $\mathbf{E}^2$ is of the form

$$(x, y, z) \mapsto (x + x_0, y + y_0, z + x_0 y + z_0),$$

where x_0, y_0, z_0 can be arbitrary real numbers. Show that the group H of such isometries is isomorphic to the *Heisenberg group* of real upper triangular 3×3 matrices with ones on the diagonal. (Hint: the action of H on $\mathbf{R}^3$ is free and transitive, so it makes $\mathbf{R}^3$ into a group isomorphic to H, with multiplication

$$(x_0, y_0, z_0)(x, y, z) \mapsto (x + x_0, y + y_0, z + x_0 y + z_0).$$

Now compare with the multiplication of two Heisenberg matrices.)

(d) If $h_a, h_b \in H$ project to translations by vectors a and b, the commutator $[h_a, h_b]$ is a vertical translation by a distance equal to the signed area of the parallelogram with sides a and b. Conclude that vertical translations form the center of H, of the identity component $G' \subset G$, and of G.

(e) Let h_a and h_b be as above, an assume that the translations a and b are linearly independent. Show that h_a and h_b generate a discrete cocompact subgroup. In particular, if a and b are the unit coordinate vectors, we get the group of upper triangular 3×3 matrices with integer coefficients, also called the *integer Heisenberg group* (Figure 3.26).

The Heisenberg group is †nilpotent—in fact, it is the only three-dimensional nilpotent but non-abelian connected and simply connected Lie group. This explains the term "nilgeometry" (see also Theorem 4.7.12).

Exercise 3.8.7 (geodesics in nilgeometry). Show that the Legendrian curves in Exercise 3.8.6(a) are geodesics of nilgeometry. Apart from these curves and vertical lines, what other geodesics are there?

It is easier to work with the radially symmetric model based on the contact structure σ.

Proof of 3.8.4(c). If G' acts with trivial stabilizers, it can be identified with its single orbit $X = G'/G'_x$, so X is itself a Lie group. Our task, then, is to investigate connected and simply connected three-dimensional Lie groups, asking which ones admit a discrete cocompact subgroup and are not subsumed by one of the preceding seven geometries.

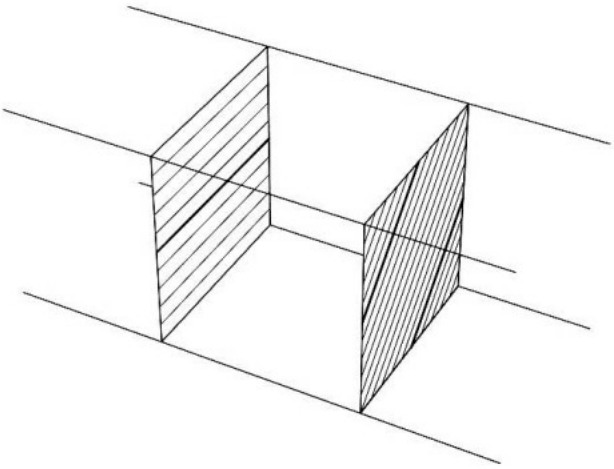

Figure 3.26. The quotient of $\mathbf{R}^3$ by the Heisenberg group. The quotient of $\mathbf{R}^3$ by H is also the quotient of $\mathbf{R} \times T^2$ by the cyclic group generated by the map $(x, y, z) \mapsto (x+1, y, z+y)$. The figure shows $\mathbf{R} \times T^2$ (opposite walls of the cylinder should be thought of as being identified); the shaded squares represent the toruses at $x = 0$ and $x = 1$. We get $\mathbf{R}^3/H$ by taking the region between the two toruses and identifying the toruses so the shadings match.

Two connected Lie groups have isomorphic Lie algebras if and only if they have isomorphic universal covers, so we start by looking at three-dimensional Lie algebras. We will think of a Lie group's Lie algebra sometimes as the tangent space at the identity, and sometimes as the set of left-invariant vector fields on the group.

Suppose G is a Lie group, and Y is a left-invariant vector field on G. Up to a factor, there is a unique volume form ω on G that is invariant under left multiplication. The flow of Y is given by *right* multiplication by a one-parameter subgroup of G, and does not, in general, preserve ω. As in the proof of 3.8.4(b), however, the assumption that there is a compact manifold modeled on (G, X) implies that Y must preserve volume (see Exercise 3.8.5).

Exercise 3.8.8 (adjoint representation). If G is a Lie group with Lie algebra $\mathfrak{g}$, the action of an element $g \in G$ by conjugation is a diffeomorphism of G fixing the identity, and therefore its derivative at the identity is a linear automorphism of $\mathfrak{g}$, which we denote by $\operatorname{ad} g$. The homomorphism $\operatorname{ad} : G \to \mathrm{GL}(\mathfrak{g})$ is called the *adjoint representation* of G. Since ad itself is a smooth map, we can consider its derivative at the identity, also called the *adjoint representation* of the Lie algebra $\mathfrak{g}$, and denoted by $\mathrm{ad} : G \to \mathrm{GL}(\mathfrak{g})$.

(a) Show that $(\mathrm{ad}\, v)(w) = [v, w]$ for elements $v, w \in \mathfrak{g}$.

(b) Show that the divergence (with respect to the invariant volume form on G) of the left-invariant vector field corresponding to v is constant over G, and equal to the trace of the linear map $\mathrm{ad}\, v : \mathfrak{g} \to \mathfrak{g}$. (Hint: the derivative at the identity of the determinant map $\det : \mathrm{GL}(V) \to \mathbf{R}$, for V a vector space, is the trace map.)

A Lie group where every left-invariant vector field preserves volume is called *unimodular*. From now on we assume that G is unimodular, so that $\operatorname{tr} \mathrm{ad}\, V = 0$ for all V in the Lie algebra $\mathfrak{g}$ of G.

Being skew-symmetric, the product (Lie bracket) on $\mathfrak{g}$ can be thought of as a map $\wedge^2\mathfrak{g} \to \mathfrak{g}$. For G three-dimensional, if we fix a positive definite quadratic form (dot product) and an orientation for $\mathfrak{g}$, we get an identification between $\wedge^2\mathfrak{g}$ and $\mathfrak{g}$, sending $V \wedge W$ to the cross product $V \times W$. Then the Lie bracket can be seen as a linear map $L : \mathfrak{g} \to \mathfrak{g}$, which turns out to be symmetric with respect to the quadratic form: $L(V) \cdot W = V \cdot L(W)$ for all $V, W \in \mathfrak{g}$.

To see this, fix an orthonormal, positively oriented basis for $\mathfrak{g}$, say $\{e_1, e_2, e_3\}$, and let l_{ij} be the entries of the matrix of L with respect to that basis. The unimodularity condition $\operatorname{tr} \mathrm{ad}\, V = 0$, applied to $V = e_1$, becomes $\sum_{i=1,2,3} [e_1, e_i] \cdot e_i = 0$, or

$$L(e_3) \cdot e_2 - L(e_2) \cdot e_3 = l_{23} - l_{32} = 0.$$

Repeating for the other basis vectors, we see that unimodularity is equivalent to the symmetry of L.

Every symmetric linear transformation has an orthonormal basis of eigenvectors. Changing $\{e_1, e_2, e_3\}$ to such a basis, the matrix of L becomes diagonal, with entries $c_i = l_{ii}$. In other words,

$$[e_i, e_{i+1}] = c_{i+2}e_{i+2},$$

where the subscripts are taken modulo 3. If we change the quadratic form so as to make the basis $\{a_1e_1, a_2e_2, a_3e_3\}$ orthonormal, where $a_1, a_2, a_3 > 0$, each c_i gets replaced by $c_i\,(a_{i+1}a_{i+2}/a_i)$. Therefore we can arrange for all the c_i's to be equal to ± 1 or 0. To normalize further, we can permute the basis elements, but if it is an odd permutation we must also change the orientation of the vector space, so the net effect is that the c_i's are permuted and change signs. To undo the sign changes, we can replace the basis elements by their negatives. Therefore, using these operations, we can arrange to have $c_1 \geq c_2 \geq c_3$, and to have at least as many positive c_i's as negative ones. Up to isomorphism, then, there are six possibilities for $\mathfrak{g}$, and therefore for the Lie group G. Only one gives rise to a new geometry:

- ○ $c_1 = c_2 = c_3 = 1$ gives $G = S^3$.
- ○ $c_1 = c_2 = 1$ and $c_3 = -1$ gives $G = \widetilde{\mathrm{SL}}(2, \mathbf{R})$.
- • $c_1 = c_2 = 1$ and $c_3 = 0$ gives rise to *solvegeometry*. The generators e_1 and e_2 commute, so the Lie group G contains a copy of $\mathbf{R}^2$. This is a normal subgroup, with quotient group $\mathbf{R}$. The Lie group is therefore a semidirect product of $\mathbf{R}^2$ with R, and determined by the action of the one-parameter group generated by e_3 on $\mathbf{R}^2$. The derivative of this action at the identity is $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. By rotating e_1 and e_2 by 45° we can make the action of e_3 diagonal, with derivative at $t = 0$ equal to $\left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$, so the actual transformations in this basis are given by

  $$t \to \begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix}.$$

  Thus G is an extension $0 \to \mathbf{R}^2 \to G \to \mathbf{R} \to 1$, consisting of maps of the form

  $$(x, y, t) \to (e^{t_0}x + x_0, e^{-t_0}y + y_0, t + t_0),$$

  for arbitrary real x_0, y_0 and t_0.
- ○ $c_1 = 1$ and $c_2 = c_3 = 0$ gives the Heisenberg group.
- ○ $c_1 = c_2 = c_3 = 0$ gives $G = \mathbf{R}^3$.
- ○ $c_1 = 1$, $c_2 = 0$ and $c_3 = -1$ gives the universal cover of the group of isometries of $\mathbf{E}^2$, where e_1 and e_3 act as translations in the x and y directions, while e_2 rotates about the origin.

Example 3.8.9 gives an example of a compact three-manifold modeled on solvegeometry, so this is actually a model geometry.

We still have to compute the full group G of isometries of solvegeometry. There are independent reflections in the two eigenspaces in the $\mathbf{R}^2$ direction. One can also reflect in the $\mathbf{R}$ direction while at the same time interchanging the two eigenspaces. Thus, the index of G' in G is 8. $\boxed{\text{3.8.4(c)}}$

Example 3.8.9 (solvegeometry manifold). Examples of three-manifolds modeled on solvegeometry can be obtained by forming the mapping torus M_ϕ for a diffeomorphism ϕ of the torus to itself. For example, let $\phi : T^2 \to T^2$ come from the linear automorphism of $\mathbf{R}^2$ with matrix $\left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$. Arrange the universal cover of the torus in $\mathbf{R}^2$ so that the two orthogonal eigenspaces of ϕ line up with the x- and y-axes. Since the eigenvalues of ϕ are reciprocal to each other, there is some t_0 such that the transformation

$$\psi : (x, y, t) \mapsto (e^{t_0}x, e^{-t_0}y, t + t_0)$$

of $\mathbf{R}^3$ induces the given automorphism ϕ between $\mathbf{R}^2 \times \{0\}$ and $\mathbf{R}^2 \times \{t_0\}$. Therefore M_ϕ is a solve-manifold: it is the quotient of $\mathbf{R}^3$ by the discrete group of automorphisms of solvegeometry generated by ψ together with unit translations along the x- and y-axes.

Exercise 3.8.10 (mapping tori of the torus). Generalize the example above to the case when ϕ comes from an arbitrary linear automorphism of $\mathbf{R}^2$ that preserves the lattice $\mathbf{Z}^2 \subset \mathbf{R}^2$. If λ_1 and λ_2 are the roots of the characteristic polynomial of ϕ, show that:

(a) If λ_1 and λ_2 are not real, they are roots of unity, and ϕ has finite order. T^2 has a Euclidean structure for which ϕ is an isometry, and M_ϕ has a Euclidean structure.

(b) If λ_1 and λ_2 are real and distinct, ϕ has two eigenspaces. The torus has a Euclidean metric for which these eigenspaces are orthogonal, and M_ϕ has a solvegeometry structure. Find two linear maps that have the same characteristic polynomials with real distinct roots, but give rise to mapping tori that are not homeomorphic.

(c) In the remaining case, $\lambda_1 = \lambda_2 = \pm 1$, either we have $\phi^2 = 1$ and M_ϕ has a Euclidean structure, or ϕ is conjugate to a map with matrix $\pm\left(\begin{smallmatrix} 1 & n \\ 0 & 1 \end{smallmatrix}\right)$, for $n \neq 0$, and M_ϕ has a nilgeometry structure.

In general, show that M_ϕ is a cocompact discrete subgroup of a three-dimensional Lie group. How does this statement generalize for linear maps of the n-torus?

It is a curious fact that each of the eight three-dimensional model geometries is isometric to a three-dimensional Lie group with a left-invariant metric. All the Lie groups except for $\mathbf{H}^3$ are unimodular. The Lie group for $\mathbf{H}^3$ is the group of homotheties of the plane, which acts simply transitively as a group of isometries of upper half space. In every case but $\mathbf{H}^3$, the group of automorphisms of the geometry is the semidirect product of the simply connected Lie group with its group of isometric automorphisms.

Exercise 3.8.11 (two-dimensional geometries from groups). Derive the classification of two-dimensional model geometries from the classification of three-dimensional unimodular Lie algebras, but without using Riemannian geometry.

3.9. Piecewise Linear Manifolds

Example 3.2.11 suggests that triangulated spaces, even when they are manifolds, can be difficult to deal with. If we want manifolds of somewhat better quality, we must impose some additional structure, such as differentiability (Example 3.1.3). In the context of gluings it is natural to consider *piecewise linear* structures, which we introduce now.

A map from a subset of an affine space into another affine space is called *piecewise linear* if it is the restriction of a simplicial map defined on the polyhedron of some simplicial complex. A *piecewise linear manifold* is a manifold based on PL, the pseudogroup generated by piecewise linear homeomorphisms between open subsets of $\mathbf{R}^n$.

Problem 3.9.1. (a) The reason a piecewise linear map is defined in terms of the restriction of a simplicial map is that this definition is clearly satisfied for simple maps, such as the identity. Show that, given a piecewise linear map on an open set, the domain itself can be triangulated so that the map is simplicial.

(b) Show that piecewise linear homeomorphisms between open subsets of $\mathbf{R}^n$ form a pseudogroup. In particular, the "generated by" in the definition of PL is superfluous.

The notion of a piecewise linear map can be immediately extended to maps from a subset of an affine space to a piecewise linear manifold, by looking at the expression of the map in charts.

Convention 3.9.2. When we talk of a triangulation of a piecewise linear manifold X, we will normally have in mind a *piecewise linear triangulation*, that is, one where the homeomorphism $|\Sigma| \to X$ is a piecewise linear map.

Problem 3.9.3 (triangulating piecewise linear manifolds). (a) Show that any piecewise linear manifold has a triangulation. (Hint: start from a finite atlas—see Convention 3.1.4. See also the proof of Theorem 3.10.2.)

!F"." G 4 4 !) * ! <!# &K K L ' K
- 0' ' n >' !> /

!"#$! ! !% & ' ! E< " & 9 ' (> /
7' 0-

!"#$! ! !% & ' ! E< " & 9 ' (> / 0-

!"#$! ! !% & ' ! E< 0' ' 4 (> / 1 0-

!"#$! ! !% & ' ! E< F ' (> /((4
. 0 0-
V . D ' / F ' > / ' 4

= !"#$! ! !% & ' ! E< F ' (> /
' 09 0-

@ !"#$! ! !% & ' ! E< F ' (> /
B ' 9 0-
V F6++D#DG.6": >D% D.#H %+ .7#DD F6 DG 6%G

< !"#$! ! !% & ' ! E< D ' /(" & 9 ' (> /
1 - & 0-
h ' 4

E !"#$! ! !% & ' ! E<
D ' /(> /("-4' -(' -0 0-
%9 : 4 - ;;;
V @ 0'&

!"#$! ! !% & ' ! E< D- D ' /
> / ' "-4' - ' -0 8 - D- "-- B / 8 0-
I& 8 . D,

CHAPTER

21

Length and Plane Area

21.1 The Definition of Arc Length

Given an arc $\widehat{AB}$ of a circle C:

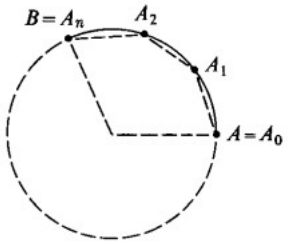

Figure 21.1

We take a sequence of points

$$A = A_0, A_1, A_2, \ldots, A_n = B\,,$$

in the order from A to B on the arc; and for each pair of successive points A_{i-1}, A_i we draw the segment $\overline{A_{i-1}A_i}$, as indicated in the figure. The union of these segments is called an *inscribed broken line;* and the sum of their lengths is denoted by p_n. Thus

$$\begin{aligned} p_n &= A_0A_1 + A_1A_2 + \cdots + A_{n-1}A_n \\ &= \sum_{i=1}^{n} A_{i-1}A_i\,. \end{aligned}$$

There are now various ways that we might define the length of the arc $\widehat{AB}$. If we merely want to *state* a definition, as a matter of form, without intending to use it, then our problem is simple. We agree to use equally spaced points $A_0, A_1, \ldots, A_n$. Then length p_n of our broken line is now determined by n, and we can define the length to be

$$p = \lim_{n\to\infty} p_n .$$

To justify this, we would need to explain what is meant by $\lim_{n\to\infty}$, and we would have to show that the indicated limit exists, for every circular arc.

The following definition, however, is more manageable. Let P be the set of all numbers p_n which are lengths of broken lines inscribed in $\widehat{AB}$. Thus

$$P = \left\{ p_n \,\middle|\, p_n = \sum A_{i-1}A_i \right\}.$$

Let

$$p = \sup P .$$

To justify this, we need to prove the following theorem.

■ *THEOREM 1.* P has an upper bound.

It will then follow that P has a *least* upper bound sup P. The proof is easy, on the basis of the following preliminary result. Let $\triangle PQR$ be an isosceles triangle, with $PQ = PR$.

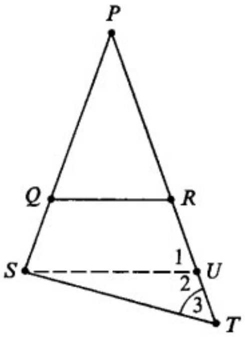

Figure 21.2

We assert that if P-Q-S and P-R-T, then

$$ST > QR .$$

Suppose (without loss of generality) that $PS < PT$, as in the figure, and take U between R and T so that $\overline{SU} \parallel \overline{QR}$. Then

$$\frac{SU}{QR} = \frac{PS}{PQ} > 1 .$$

Therefore

$$SU > QR.$$

We shall now show that $ST > SU$. Evidently $\angle 1$ is acute, because $\angle 1$ is a base angle of an isosceles triangle. Therefore $\angle 2$ is obtuse. Therefore $\angle 3$ is acute. Therefore $m\angle 3 < m\angle 2$. Therefore $ST > SU$. (Why?)

We now return to our circular arc. Draw any square that contains the whole circle in its interior (Fig. 21.3). We project each point A_i onto the square, as indicated in the figure. That is, A_i' is the point where $\overrightarrow{DA_i}$ intersects the square. Then $A_{i-1}A_i < A_{i-1}'A_i'$. Therefore p_n is always less than $\Sigma_{i=1}^{n} A_{i-1}'A_i'$. Therefore p_n is always less than the perimeter of the square. Thus the perimeter of the square is the upper bound that we are looking for.

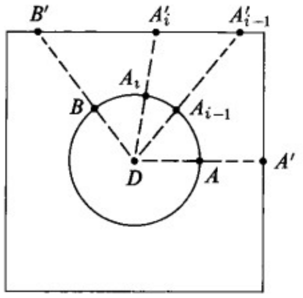

Figure 21.3

This justifies our definition

$$p = \sup P.$$

Of course a circle is not an arc. But we can define the circumference of a circle in an analogous way, by setting up an inscribed polygon with vertices

$$A_0, A_1, \ldots, A_{n-1}A_n = A_0.$$

We then let p_n be the perimeter

$$\sum_{i=1}^{n} A_{i-1}A_i,$$

we let P be the set of all such perimeters p_n, and we define the circumference as

$$p = \sup P.$$

21.2 The Existence of π

We now want to prove the existence of the number π. To do this, we need to show that the ratio of the circumference to the diameter is the same for all circles. This is a theorem about suprema, and to prove it, we need a preliminary result.

Let P be any bounded set of positive numbers, and let k be a positive number. Then kP denotes the set of all numbers of the form kp, where p belongs to P. For example, if

$$P = [0,1] = \{x \mid 0 \leqq x \leqq 1\},$$

and $k = 3$, then

$$kP = 3P = [0,3].$$

If $P = [1,2]$ and $k = \frac{2}{3}$, then $kP = [\frac{2}{3}, \frac{4}{3}]$; and so on.

This "multiplication" is associative. That is,

$$j(kP) = (jk)P,$$

because

$$\{jx \mid x \in kP\} = \{j(kp) \mid p \in P\} = \{(jk)p \mid p \in P\}.$$

Thus, for example, we always have

$$\frac{1}{k}(kP) = P.$$

LEMMA 1. If b is an upper bound of P, then kb is an upper bound of kP.

REASON. If $p \leqq b$, then $kp \leqq kb$.

LEMMA 2. If c is an upper bound of kP, then c/k is an upper bound of P.

PROOF. Since $P = (1/k)(kP)$, this follows from Lemma 1. ☐

These lemmas give us the following theorem.

■ ***THEOREM 1.*** $\sup(kP) = k \sup P$.

PROOF. Let $b = \sup P$. Then b is an upper bound of P. By Lemma 1, kb is an upper bound of kP.

Suppose that kP has an upper bound

$$c < kb.$$

Then c/k is an upper bound of P, by Lemma 2. This is impossible, because $c/k < b$, and b was the *least* upper bound of P. ☐

We can now prove the theorem which establishes the existence of π. What is needed is the following theorem.

■ ***THEOREM 2.*** Let C and C' be circles with radii r, r' and circumferences p, p'. Then

$$\frac{p}{2r} = \frac{p'}{2r'}.$$

That is, the ratio of the circumference to the diameter is the same for all circles. This common ratio is denoted by π.

PROOF. Suppose that the circles have the same center. (This involves no loss of generality.) In the figure, we indicate the ith side $\overline{A_{i-1}A_i}$ of a polygon inscribed in C. To each such polygon there corresponds a polygon inscribed in C', obtained by projection outward (or perhaps inward) from the common center D. We then have

$$\triangle DA_{i-1}A_i \sim \triangle DA'_{i-1}A'_i.$$

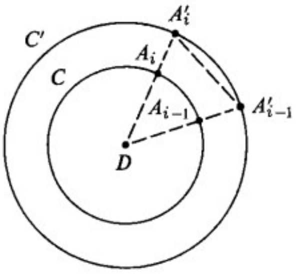

Figure 21.4

Therefore

$$\frac{A'_{i-1}A'_i}{A_{i-1}A_i} = \frac{DA'_i}{DA_i} = \frac{r'}{r}.$$

If the perimeters of our polygons are p_n and p'_n, then we have

$$p'_n = \frac{r'}{r} \cdot p_n.$$

Let

$$p = \sup P \quad \text{and} \quad p' = \sup P',$$

where $P = \{p_n\}$ and $P' = \{p'_n\}$, as usual. Then

$$P' = \frac{r'}{r} \cdot P.$$

Therefore, by the preceding theorem, using $k = r'/r$, we have

$$\sup P' = \frac{r'}{r} \sup P \quad \text{or} \quad p' = \frac{r'}{r} p \quad \text{or} \quad \frac{p'}{r'} = \frac{p}{r}.$$

Dividing both sides by 2, we get the equation called for in the theorem. ☐

An analogous theorem holds for circular arcs.

■ ***THEOREM 3.*** Let $\widehat{AB}$ and $\widehat{A'B'}$ be arcs of the same degree measure, in circles of radius r and r', respectively. Let the lengths of $\widehat{AB}$ and $\widehat{A'B'}$ be p and p'. Then

$$\frac{p}{r} = \frac{p'}{r'}.$$

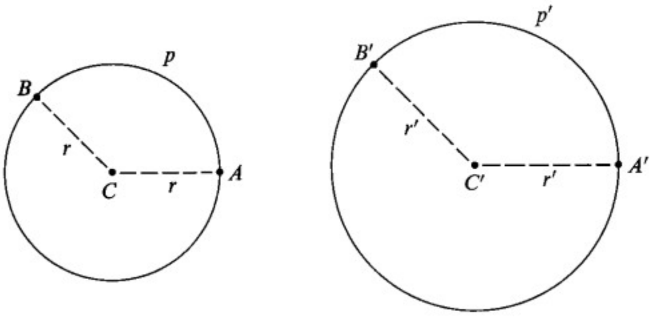

Figure 21.5

This ratio is called the *radian measure* of the arc $\widehat{AB}$. If $\widehat{AB}$ is a minor arc, then p/r is the *radian measure* of the angle $\angle BCA$. The theorem tells us that the radian measure really depends only on the angle, or on the degree measure of the arc, and is independent of the radius of the circle.

The proof is virtually identical with the proof of the preceding theorem.

21.3 Limits as the Mesh Approaches Zero

Given

$$p = \sup P,$$

we know that we can find numbers p_n in the set P, as close to $\sup P$ as we please. More precisely, if ε is any positive number, there is a p_n in P such that

$$p_n > p - \varepsilon.$$

= !"#$! ! !% & ' ! E< D- D ' /
> / ' "-4' - ' -0 8 - D- "-- B / 9 & 0' /5 6
8 -24

= !"#$! ! !% & ' ! E< D ((" & 9 ' (> /
" 0-

= !"#$! ! !% & ' ! E< D -(- ' - ' 0-
V D - - '

= !"#$! ! !% & ' ! E< + -' ((> /
F , 0-

= !"#$! ! !% & ' ! E< + -' ((> /
7 9 0-

== !"#$! ! !% & ' ! E< + -' ((> /
' 0-

=@ !"#$! ! !% & ' ! E< + -' ((> /
0-

=< !"#$! ! !% & ' ! E< + -' ((> /
' 0 4 0-
%9 : 4 - ;;;

I& J % & ' 0 2 ' ' &

=E !"#$! ! !% & ' ! E<
+ (F (0 /(D,0 ' - ' & 0-

= !"#$! ! !% & ' ! E< + (F 7 0-
V . -

@ !"#$! ! !% & ' ! E< > /((G 9 ' 0-
I& <=8 E< '0 6
. 6G1%B 16 7:"B1"

@ !"#$! ! !% & ' ! E< > /((# / 0-
V : > /
%9 : 4 - ;;;

I& 8 E

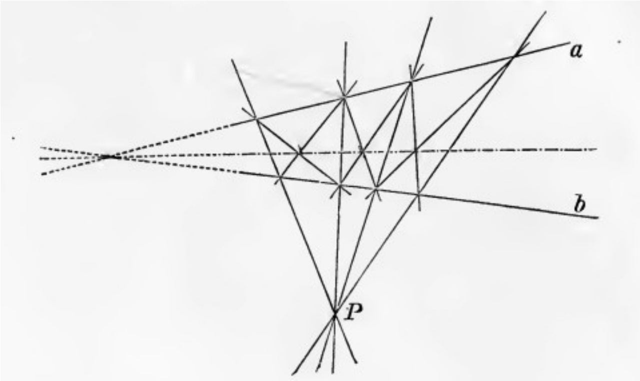

+ &

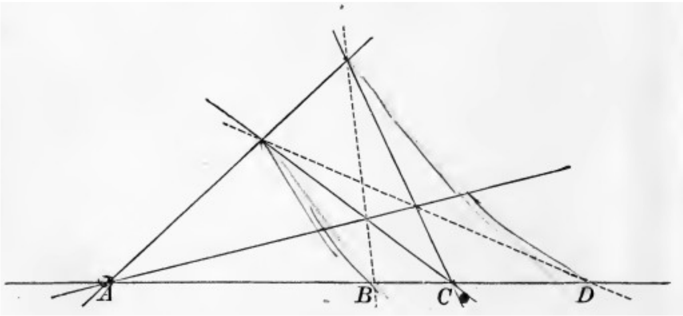

+ &

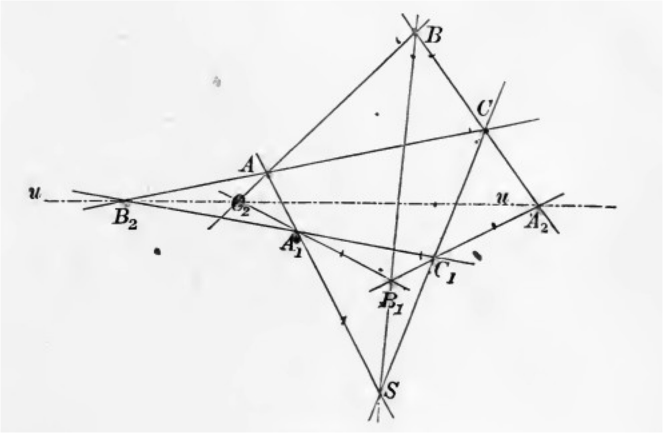
B
C
A
u
B2
C2
A1
u
A2
C1
B1
S

+ &

@ !"#$! ! !% & ' ! E< > /((')' 0-
V " ' > / '
%9 : 4 - ;;;

@ !"#$! ! !% & ' ! E< > ' ' (0' (" ' /
7/- 0-

@ !"#$! ! !% & ' ! E<
6 (F ' (> /(. D ' 0-

@=   !"#$!   !   !% & ' ! E<   6   (7 9   (0'   ) 9   ' 0-

@@   !"#$!   !   !% & ' ! E<   :   1   0-
I& = 8@   :   66   .7D .#"G DGFDG H %+ .7D G?)D#   "GF

I& @ 8<E   + &

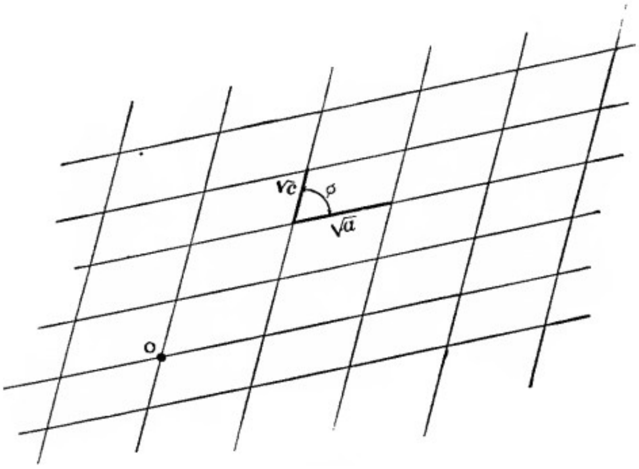

@<   !"#$!   !   !% & ' ! E<
G (" ' /   ((' (>   / .   0   0-
V " GDB "G":H 6 %+
:"GD >D% D.#H
+6G6.D "GF F6++D#DG.6":

I&   68

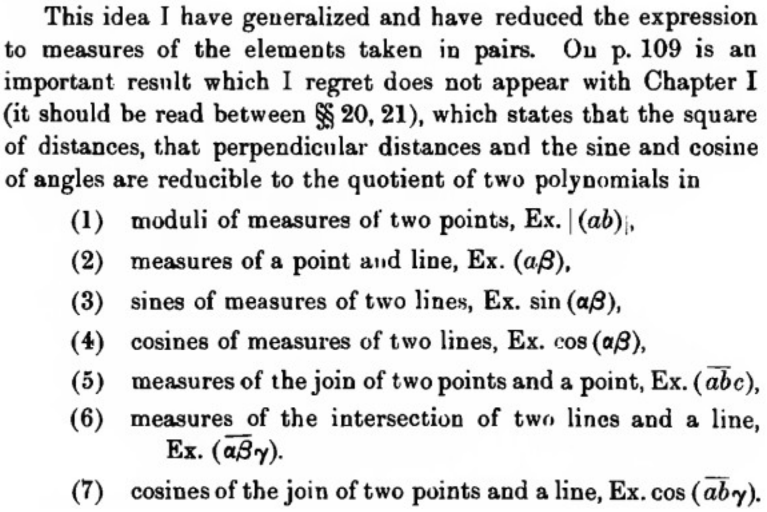

This idea I have generalized and have reduced the expression to measures of the elements taken in pairs. On p. 109 is an important result which I regret does not appear with Chapter I (it should be read between §§ 20, 21), which states that the square of distances, that perpendicular distances and the sine and cosine of angles are reducible to the quotient of two polynomials in

(1) moduli of measures of two points, Ex. $|(ab)|$,
(2) measures of a point and line, Ex. $(a\beta)$,
(3) sines of measures of two lines, Ex. $\sin(\alpha\beta)$,
(4) cosines of measures of two lines, Ex. $\cos(\alpha\beta)$,
(5) measures of the join of two points and a point, Ex. $(\overline{ab}c)$,
(6) measures of the intersection of two lines and a line, Ex. $(\overline{\alpha\beta}\gamma)$.
(7) cosines of the join of two points and a line, Ex. $\cos(\overline{ab}\gamma)$.

CONTENTS

FINITE GEOMETRY

INTRODUCTION

PAGE

Definitions of element, incidence, determinate, measure 3
Notation for determinate and measure 3
Classes of derived points and lines considered 4
Sketch of method 5
Fundamentals of differential geometry 6
Method employed for differential geometry 7

CHAPTER I

FUNDAMENTALS OF THE GEOMETRY OF TWO, THREE AND FOUR ELEMENTS

Axioms 8
Geometry of two points and a line 8
Definition of sine function 9
Geometry of three points 10
Geometry of three lines 11
Geometry of two points and two lines 11
Geometry of three lines and a point 12
Further axioms 13
The value of sin $\frac{\pi}{2}$ 13
Definition of cosine function 13
Addition formulae for sine and cosine functions 14
Further geometry of the triangle 15
Geometry of two lines and a point 16
Geometry of three points and a line 17
Further geometry of two lines and two points 18
Geometry of three lines and a point 18
Eliminants 18
Examples 20

CHAPTER II

REDUCTION OF MEASURES CONTAINING VECTORIAL ELEMENTS

PAGE
Reduction of $(a_{\dot\phi\dot\sigma\ldots\dot\omega}b)^2$ 24
Reduction of $(a_{\dot\phi\dot\sigma\ldots\dot\omega}\beta)$ 25
Reduction of $(a_{\dot\phi\dot\sigma}\ldots\dot\phi_\omega b)$ 26
Reduction of $(a_{\dot\phi\dot\sigma}\ldots\dot\phi_\omega\beta)$ 26
Examples 26

CHAPTER III

REDUCTION OF MEASURES CONTAINING EQUATIONAL ELEMENTS

Proof that $\Sigma a_r(xa_r)+a=0$ represents a line and that any line can be represented by such an equation 30
Reduction of $(\Sigma a_r(xa_r)+a=0\ \beta)$ 31
Reduction of $(\Sigma a_r(xa_r)+a=0\ b)$ 31
Proof that $\Sigma A_r(\xi a_r)+\Sigma B_r\cos(\xi\beta_r)=0$ represents a point and that any point can be represented by such an equation 32
Reduction of $(\Sigma A_r(\xi a_r)+\Sigma B_r\cos(\xi\beta_r)=0\ c)^2$ 33
Reduction of $(\Sigma A_r(\xi a_r)+\Sigma B_r\cos(\xi\beta_r)=0\ \gamma)$ 33
Proof that $\Sigma A_r(\xi u_r)+\Sigma B_r\cos(\xi\beta_r)=0$ represents a direction, when $\Sigma A_r=0$ 34
Reduction of $(\Sigma A_r(\xi a_r)+\Sigma B_r\cos(\xi\beta_r)=0\ \gamma)$ when $\Sigma A_r=0$ . . 34
Examples 34
Method of Analysis 35

DIFFERENTIAL GEOMETRY

CHAPTER IV

DIFFERENTIATION OF MEASURES OF SIMPLE ELEMENTS

Differentiation of $|(xa)|$ 39
Differentiation of $(x\boldsymbol{a})$ 40
Differentiation of (ξa) 40
Differentiation of $(\xi\boldsymbol{a})$ 41
Linearity of differentiation of measures of more than one variable element 41
Examples 41

CHAPTER V

DIFFERENTIATION OF DETERMINATES OF SIMPLE ELEMENTS

PAGE
General method of effecting differentiations of determinates . . 43
Differentiation of $x\bar{y}$ 43
Differentiation of $\overline{\xi\eta}$ 44
Examples 44

CHAPTER VI

DIFFERENTIATION OF VECTORIAL ELEMENTS

Evaluation of $dx_{\dot{\rho}\dot{\sigma}\ldots\dot{\omega}}$ 46
Evaluation of $dx_{\dot{\rho}\dot{\sigma}\ldots\phi\omega}$ 46
Examples 47

CHAPTER VII

DIFFERENTIATION OF EQUATIONAL ELEMENTS

Reduction of $d\{\Sigma a_r(xa_r)+a=0\}$ 50
Reduction of $d\{\Sigma A_r(\xi a_r)+\Sigma B_r\cos(\xi\beta_r)=0\}$. 51
Examples 52

CHAPTER VIII

REDUCTION OF MEASURES CONTAINING FUNCTIONAL ELEMENTS

Formula for the evaluation of (τxa), (νxa) 54
Formula for the evaluation of (τxa), (νxa) 54
Formula for the evaluation of $(p\xi a)^2$, $(\nu\xi a)$ 54
Formula for the evaluation of $(p\xi a)$, $(\nu\xi a)$ 55
Examples 55
Rule for writing down the differentiation of simply derived elements . 58
Examples 60

CHAPTER IX

GENERALIZED DISPLACEMENT OF A POINT

Generalized displacement given by $\tau_1 x$, $\delta_1 x$; $\tau_2 x$, $\delta_2 x$; … $\tau_n x$, $\delta_n x$. 63
Evaluation of dx 63
Evaluation of $d\tau x$ 63
Evaluation of (νxa), (τxa) 63

PLANE CURVES

CHAPTER X

REDUCTION OF COMPLEX FUNCTIONAL ELEMENTS

PAGE
Proof for a plane curve that $p\tau x = x$, $\tau p\xi = \xi$ 65
Definition of ρx, $\rho\xi$ 65
Reduction of $\nu\tau x$ 66
Reduction of $\nu^2 x$ 66
Reduction of $p\nu x$ 66
Reduction of $\nu p\xi$ 66
Reduction of $\nu^2\xi$ 66
Reduction of $p\nu\xi$ 67
Examples 67

CHAPTER XI

SUCCESSIVE DIFFERENTIATION OF MEASURES

Successive differentiation of $\frac{1}{2}(xa)^2$ 69
Successive differentiation of $|(x\dot{a})|$ 70
Successive differentiation of (xa) 70
Successive differentiation of (ξa) 71
Successive differentiation of $(\xi \mathrm{a})$ 71
Examples 71

CHAPTER XII

INTEGRATION OF MEASURES

Definition of integration of measures containing point and line variables 73
Examples of integrals 74
Laplace's equation 77

Miscellaneous Examples 78

Appendix 110

Table of Formulae 113

Index 119

@E !"#$! ! !% & ' ! E< G D - ' (> /
) ' 0-
%9 : 4 - ;;; 6 0 ' M ;;;

@ !"#$! ! !% & ' ! E< G D - ' (> /
' & 0-
I& = 8

'0
.7D 7H D#)%:6
>D% D.#H
B '4 /0 ' ' & . ' 0 ' 0 0 -
' ' - - 4 & - 0 0 - ' .
' & ' ' & ' & ' & 5 ' - 0 0
' '0 -
' ' :
+ ' / 0 5 5 - ' ' 0 0 - ' 5 5 ' & 4 5 ")5 ' - F
9 ' / &) F 6 F 4 - \ /)5
F '00 ' ' & 0 D

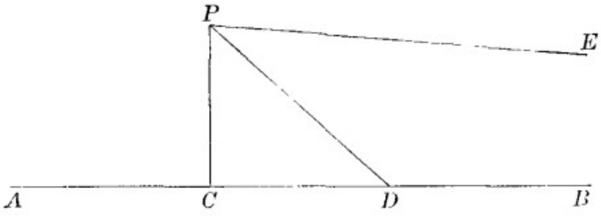

D ' - 9 0' ') ' D '* ' ' & 5 D5
' - ' & 0' ' 0 0 - ' - ' 6
' ' & ' & 9/ ' ' - \ /
' & F % - ' \ - ' 0' ' "
' - '* & ' ' & 0' ' B '/ ' D 0' '
") ' - ' 0' ' - D . 5 '
' / 0 ' 0' ' ' 5 9 / ' - 0'
: & '* ' ' & & ' ' ' &
0' ' ' - ' 00 - ") ' ' B |80
- ' & 0' ' ' 0 0 - ' - ' 5 0

. " ' & ' ' 0' ' ' ' 0

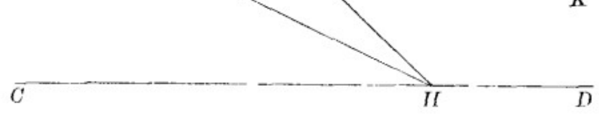

: ") 9 0' ' F ' D ' - + 9 ' / 0 ")
- D5 0 4 ' ") 0' ' F ' +
. 75 F5 - ' D7 ' - +7 6 7 4 - \ /
F5 '00 ' 0 0' ' F
) & 0 D7 ") 0' & & +5 ' -
& 0 +7 5 +15 + - 9
& 0 75 D7 ' - +7

< !"#$! ! !% & ' ! E< G D - ' (> /
B 0-
V . + '

%9 : 4 - ;;;
6 0 ' M ;;;

"9 -'& & ' - & ' - '

< !"#$! ! !% & ' ! E< ((0' '0 0-
V > ' '
%9 / 2 4 > /

< !"#$! ! !% & ' ! E< 9 (F ' (> /
* 0-
I& = 4 ' ' & ' ' & : 9' 4 * 0 '
'

I& @ '0 . / 4

< !"#$! ! !% & ' ! E< # ' ' (> /
> - - 0-

I& =8 < 7" .D# E
. # ' 4' .

6 '0 - # ' 4' ' -
' 4' ' # ' ' ' - . &
' >' ' 4' 0 '/ & ' ' ' ' -
' & / ' B 0 4 ' D - ' 0' 5
' -' - 0 ' - /0 9 0' ' '4 ' '
4' B - # ' ' 4' '
- ' ' 4' ' - ' ' 0 ' '
: & 0 B - 4 0 ' >' L ' 0' &
' 4' ' 9 ' - ' - ' ' 9 0'

< !"#$! ! !% & ' ! E< # ' ' (> / 0-
V 0'&

<= !"#$! ! !% & ' ! E< . (> / F - 0-

Contents

Introduction XI

0. Fundamental Not(at)ions 1
   1. Sets 1
   2. Functions 6
   3. Physical Background 13

I. Real Vector Spaces 18
   1. Spaces 18
      Subspace geometry, components
   2. Maps 24
      Linearity, singularity, matrices
   3. Operators 31
      Projections, eigenvalues, determinant, trace

II. Affine Spaces 43
   1. Spaces 43
      Tangent vectors, parallelism, coordinates
   2. Combinations of Points 49
      Midpoints, convexity
   3. Maps 53
      Linear parts, translations, components

III. Dual Spaces 57
   1. Contours, Co- and Contravariance, Dual Basis 57

IV. Metric Vector Spaces 64
   1. Metrics 64
      Basic geometry and examples, Lorentz geometry
   2. Maps 76
      Isometries, orthogonal projections and complements, adjoints
   3. Coordinates 83
      Orthonormal bases

4. Diagonalising Symmetric Operators 92
Principal directions, isotropy

V. Tensors and Multilinear Forms 98
1. Multilinear Forms 98
Tensor Products, Degree, Contraction, Raising Indices

VI. Topological Vector Spaces 114
1. Continuity 114
Metrics, topologies, homeomorphisms
2. Limits 125
Convergence and continuity
3. The Usual Topology 128
Continuity in finite dimensions
4. Compactness and Completeness 136
Intermediate Value Theorem, convergence, extrema

VII. Differentiation and Manifolds 149
1. Differentiation 149
Derivative as local linear approxiamation
2. Manifolds 160
Charts, maps, diffeomorphisms
3. Bundles and Fields 170
Tangent and tensor bundles, metric tensors
4. Components 182
Hairy Ball Theorem, transformation formulae, raising indices
5. Curves 189
Parametrisation, length, integration
6. Vector Fields and Flows 195
First order ordinary differential equations
7. Lie Brackets 200
Commuting vector fields and flows

VIII. Connections and Covariant Differentiation 205
1. Curves and Tangent Vectors 205
Representing a vector by a curve
2. Rolling Without Turning 207
Differentiation along curves in embedded manifolds
3. Differentiating Sections 212
Connections, horizontal vectors, Christoffel symbols
4. Parallel Transport 222
Integrating a connection

5. Torsion and Symmetry 228
Torsion tensor of a connection
6. Metric Tensors and Connections 232
Levi–Cività connection
7. Covariant Differentiation of Tensors 240
Parallel transport, Ricci's Lemma, components, constancy

IX. Geodesics 246
1. Local Characterisation 246
Undeviating curves
2. Geodesics from a Point 249
Completeness, exponential map, normal coordinates
3. Global Characterisation 256
Criticality of length and energy, First Variation Formula
4. Maxima, Minima, Uniqueness 264
Saddle points, mirages, Twins 'Paradox'
5. Geodesics in Embedded Manifolds 275
Characterisation, examples
6. An Example of Lie Group Geometry 281
2×2 matrices as a pseudo-Riemannian manifold

X. Curvature 298
1. Flat Spaces 298
Intrinsic description of local flatness
2. The Curvature Tensor 304
Properties and Components
3. Curved Surfaces 319
Gaussian curvature, Gauss–Bonnet Theorem
4. Geodesic Deviation 324
Tidal effects in spacetime
5. Sectional Curvature 326
Schur's Theorem, constant curvature
6. Ricci and Einstein Tensors 329
Signs, geometry, Einstein manifolds, conservation equation
7. The Weyl Tensor 337

XI. Special Relativity 340
1. Orienting Spacetimes 340
Causality, particle histories
2. Motion in Flat Spacetime 342
Inertial frames, momentum, rest mass, mass–energy
3. Fields 355
Matter tensor, conservation

4. Forces 367
No scalar potentials
5. Gravitational Red Shift and Curvature 369
Measurement gives a curved metric tensor

XII. General Relativity 372
1. How Geometry Governs Matter 372
Equivalence principle, free fall
2. What Matter does to Geometry 377
Einstein's equation, shape of spacetime
3. The Stars in Their Courses 384
Geometry of the solar system, Schwarzschild solution
4. Farewell Particle 398

Appendix. Existence and Smoothness of Flows 400
1. Completeness 400
2. Two Fixed Point Theorems 401
3. Sequences of Functions 404
4. Integrating Vector Quantities 408
5. The Main Proof 408
6. Inverse Function Theorem 415

Bibliography 418

Index of Notations 421

Index 424

<@ !"#$! ! !% & ' ! E<
. /((6 '& ' /((> / 7' 0-
I& 8 =

9. Imaginary straight lines.

DEF. 4. *The locus of the double points (imaginary) of the overlapping involutions in which an overlapping involution pencil (real) is cut by real transversals is a pair of imaginary straight lines. Each part of this locus, which is continuous, constitutes an imaginary straight line.*

The two straight lines considered as a pair of straight lines are termed a pair of conjugate imaginary straight lines.

Hence it follows that an imaginary straight line is determined by an imaginary point, which is a double point of an involution, and a real point; the vertex of the involution pencil. The conjugate imaginary line is determined by the conjugate imaginary point, the other double point of the involution, and by the same real point, the vertex of the involution pencil. Hence two conjugate imaginary straight lines differ as to their determining elements only in the sign of the imaginary part. An imaginary line considered in reference to the involution pencil is spoken of as a double ray of the pencil.

From the definition it also follows that only one imaginary straight line passes through a given real and a given imaginary point.

The straight line joining the centre of an involution range to the vertex of the pencil is called the mean line of the two conjugate imaginary lines, which join its double points to the vertex of the pencil, This mean line is real. As there are an infinite number of involutions on the same base which have the same centre, there are an infinite number of pairs of imaginary straight lines through a given real point which have the same mean line.

Hence given any real point L and any real straight line o through it, there are an infinite number of pairs of imaginary straight lines which pass through L and which have o as their mean line. Such a system of straight lines may be termed an imaginary system *centre L, mean line o.*

It also follows from the definition that every imaginary straight line passes through a real point, viz. the vertex of the involution pencil of which it is a double ray, and that a pair of conjugate imaginary straight lines intersect in a real point, viz. the vertex of the pencil of which they are the double rays.

10. *To determine the straight line which joins an imaginary point to a real point.*

Construct the involution of which the imaginary point is a double point. (1) If the real point lies on this straight line the required line is the real base of the involution. (2) If the real point does not lie on the base construct an involution pencil by joining pairs of real conjugate points of the involution to the real point. The required line is one of the imaginary double rays of this involution pencil.

To determine the point of intersection of a given imaginary straight line with a given real straight line.

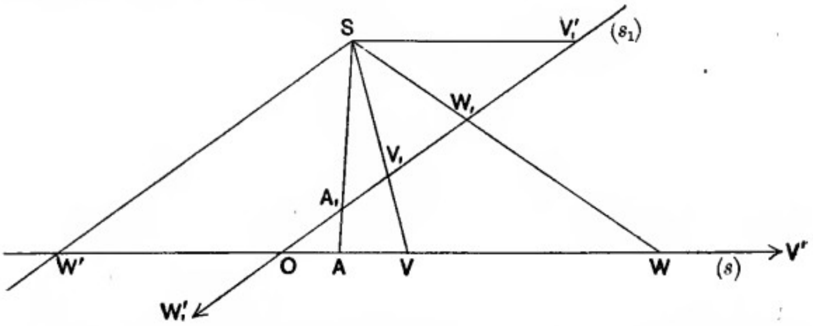

Let the given imaginary straight line be determined as the connector of one of the imaginary double points E of an involution, centre V, constant $-K^2$, situated on a straight line s, with a real point S. Let s_1 be the given real straight line. Let V' and W_1' be the points at infinity on s and s_1 respectively and let ss_1 be O.

Construct V_1 and V_1' the projections of V and V' from S on s_1. Draw SW' parallel to s_1 to meet s in W'. Take W the conjugate of W' in the involution on s. Join W to S to meet s_1 in W_1. Then W_1 is the centre of the involution into which the involution on s is projected from S on s_1. Also $W_1V_1'.W_1V_1 = -M^2$, the constant of this involution on s_1.

Hence, if E be the double point of the involution on s remote from O, the point E_1 in which SE meets s_1 is at a distance $V_1'E_1$ from V_1' such that $V_1'E_1 = V_1'W_1 - \sqrt{W_1V_1 . W_1V_1'}$, the negative sign being given to the square root since $V_1'W_1$ is drawn in the opposite direction to VW.

The following relations hold between the lengths in the figure:

(*a*) $VW' . VW = -K^2, \qquad W_1V_1 . W_1V_1' = -M^2.$

(*b*) Since $(V'W'VW) = (V_1'W_1'V_1W_1)$ and V' and W_1' are at infinity,

$$\frac{VW}{W_1V_1} = \frac{VW'}{W_1V_1'} = \frac{W'W}{V_1'V_1},$$

and $$\frac{VW}{VW'} = \frac{W_1V_1}{W_1V_1'} = \frac{VW^2}{-K^2} = \frac{W_1V_1^2}{-M^2} = \frac{-K^2}{VW'^2} = \frac{-M^2}{W_1V_1'^2} = \lambda \text{ (suppose).}$$

(*c*) Since $$(V'W'OV) = (V_1'W_1'OV_1),$$
$$OW'.OV_1' = VW'.V_1V_1' = WW'.W_1V_1' = AW'.A_1V_1',$$

where A and A_1 are any pair of real points on s and s_1 collinear with S.

If a range composed wholly or partly of imaginary points be projected from a real point by imaginary and real straight lines upon any other real straight line so that a second range is formed on this straight line, the anharmonic ratios of any four corresponding points of these ranges are equal.

It will be proved that if, as in the preceding, V be the centre of the involution which determines an imaginary point E, and W_1 the centre of the involution which determines the corresponding point E_1, then $W'E.V_1'E_1 =$ a constant, where W' and V_1' are the points on the two bases which in a projection from S correspond to the points at infinity. It will also be shown that this constant is equal to the product of the distances of any pair of real corresponding points from W' and V_1' respectively. Hence since W' and V_1' are fixed points the theorem is true.

In the figure

$$\begin{aligned} W'E &= W'V + \sqrt{VW.VW'} & V_1'E_1 &= V_1'W_1 - \sqrt{W_1V_1.W_1V_1'} \\ &= W'V\left(1 - \sqrt{\frac{VW}{VW'}}\right) & &= V_1'W_1\left(1 + \sqrt{\frac{W_1V_1}{W_1V_1'}}\right) \\ &= W'V(1 - \sqrt{\lambda}), & &= V_1'W_1(1 + \sqrt{\lambda}). \end{aligned}$$

Therefore $$W'E.V_1'E_1 = W'V.V_1'W_1(1 - \lambda).$$

But $$1 - \lambda = 1 - \frac{W_1V_1}{W_1V_1'} = \frac{V_1V_1'}{W_1V_1'}.$$

Therefore $W'E.V_1'E_1 = W'V.V_1'V_1 = W'O.V_1'O = W'A.V_1'A_1$.

This theorem may be stated as follows:

If through any real point a system of real and of imaginary straight lines be drawn, these straight lines determine equianharmonic ranges on all real transversals.

It also follows that *projective ranges as defined in Art.* 6 *are also projective according to the usual conception.*

Analytical verification of Art. 10.

Let S be the vertex of the pencil and V the centre of the involution on OV. Take OV for axis of x and the given real line OV_1 for axis of y.

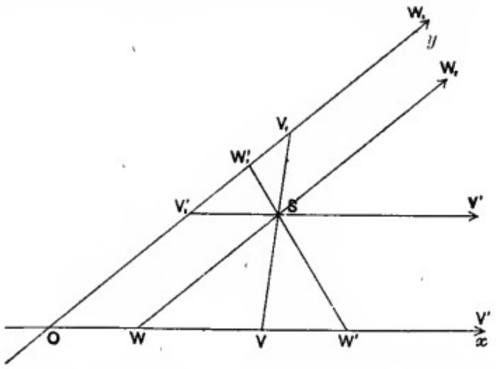

Let V' and W_1 be the points at infinity on the axes. Project these from S into V_1' and W. Let SV meet Oy in V_1.

Then V and V' are conjugate points of the involution on Ox, and therefore V_1 and V_1' are conjugate points of the involution on Oy.

Let W_1' be the centre of the involution on Oy.

Then W_1' and W_1 are conjugate points of the involution on Oy, and W and W' are conjugate points of the involution on Ox.

Since S is a given point, $OW\,(l)$ and $OV_1'\,(m)$ are given.

Also the constant of involution on $Ox\,(-b^2)$ is given and the point $V(a)$.

Hence $\quad OW'=a+\frac{b^2}{a-l}$, $\therefore\ OW_1'=\frac{m(b^2+a^2-al)}{b^2+(a-l)^2}$ and $OV_1=\frac{ma}{a-l}$.

Therefore $\quad V_1W_1'=\frac{-mlb^2}{(a-l)\{b^2+(a-l)^2\}}$ and $V_1'W_1'=\frac{ml(a-l)}{b^2+(a-l)^2}$.

Hence $-M^2$ the constant of the involution on Oy is $-\frac{m^2l^2b^2}{\{b^2+(a-l)^2\}^2}$.

Hence the distances of the double points of the two involutions from O are respectively $a+ib$, $a-ib$, and

$$\frac{m(b^2+a^2-al)}{b^2+(a-l)^2}+i\frac{mlb}{b^2+(a-l)^2},\quad \frac{m(b^2+a^2-al)}{b^2+(a-l)^2}-i\frac{mlb}{b^2+(a-l)^2}.$$

But the line joining $S\,(l,\,m)$ to $(a+ib,\,0)$ meets Oy in a point distant

$$\frac{m(a^2+b^2-al)-imbl}{(a-l)^2+b^2}$$

from O. This shows that the geometrical construction of Art. 10 agrees with the analytical conception of an imaginary straight line.

If E and E_1 be collinear double points

$$V_1'E_1 = ml\,\frac{(a-l)-ib}{b^2+(a-l)^2} \text{ and } WE = a-l+ib.$$

$\therefore$ $V_1'E_1 \,.\, WE = ml$, which is a constant for all involutions. This confirms the result of Art. 10.

11. *To determine the straight line connecting any two imaginary points.*

Construct the two involutions of which the two given imaginary points are each respectively a double point.

(1) If the two involutions have the same base this real line is the connector of the two imaginary points.

(2) If the two involutions have different bases, these involutions are in real perspective in two ways (Art. [60]) and in one of these perspectives the two given imaginary points are corresponding points. The required imaginary straight line is the connector of the centre of this perspective, which is real, with either of the two given imaginary points.

From this it follows that *only one straight line can be drawn to join two imaginary points and that it passes through a real point, viz., one of the centres of perspective of the involutions, and therefore no two imaginary straight lines can include a space.*

To determine the point of intersection of any two imaginary straight lines.

Construct the two involution pencils of which the two given imaginary straight lines are each respectively a double ray.

(1) If the two involution pencils have a common vertex, this vertex, which is real, is the point of intersection of the lines.

(2) If the two involution pencils have different vertices, these involutions are in real perspective in two ways (correlative of Art. [60]) and in one of these perspectives the two given imaginary lines are corresponding rays. The point of intersection of these lines is therefore a double point of the overlapping involution determined by the two involution pencils on one of their real axes of perspective.

Hence two imaginary straight lines intersect in only one point and this point lies on a real straight line, viz., one of the axes of perspective of the determining involutions.

12. Summary of properties of imaginary points and straight lines.

(1) *Every imaginary point contains one and only one real straight line.*	*Every imaginary straight line contains one and only one real point.*
This line is the base of the involution of which the imaginary point is a double point and it is the connector of the point and its conjugate imaginary point. There can be no other real straight line through the point for, if there were, the point would be a real point.	This point is the vertex of the involution pencil of which the imaginary line is a double ray and it is the point of intersection of the line and its conjugate imaginary line. There can be no other real point on the straight line for, if there were, the line would be a real straight line.
(2) *The connector of a pair of conjugate imaginary points is real.*	*The point of intersection of a pair of conjugate imaginary straight lines is real.*
(3) *Every real straight line, that contains an imaginary point, contains its conjugate.*	*Every real point, that contains an imaginary straight line, contains its conjugate.*
(4) *An imaginary straight line meets all real straight lines in imaginary points except those which pass through its one real point.*	*The only real points, the connectors of which to an imaginary point are real, are those which lie on the one real line through the point.*
(5) *The connector of a pair of imaginary points is real or imaginary; if real it contains the conjugates of both the points.*	*The intersection of a pair of imaginary lines is real or imaginary; if real it contains the conjugates of both the lines.*
(6) *A system of real and imaginary points on a real straight line is projected from a real point, not on the line, upon another real straight line, into a system in which real points correspond to real points and imaginary points to imaginary points.*	*A system of real and imaginary straight lines through a real point is cut by a real transversal in points which, when connected to another real point, give real straight lines corresponding to real straight lines and imaginary straight lines to imaginary straight lines.*
(7) *The connector of a pair of imaginary points and the connector of their conjugate imaginary points*	*The point of intersection of a pair of imaginary straight lines and the point of intersection of their con-*

are conjugate imaginary straight lines.

(8) *The imaginary double points of an involution are harmonic conjugates of every pair of conjugate points of the involution.*

(9) *If a pair of conjugate imaginary points coincide, they coincide in a real point.*

jugate imaginary lines are conjugate imaginary points.

The imaginary double rays of an involution pencil are harmonic conjugates of every pair of conjugate rays of the involution pencil.

If a pair of conjugate imaginary lines coincide, they coincide in a real line.

13. *In two projective pencils with real vertices, which have two pairs of real corresponding rays, all pairs of corresponding rays, real or imaginary, intersect on a straight line (real) if the ray joining the vertices of the pencils is a self-corresponding ray.*

Let S and S' be the vertices of the pencils. Then by Art. [34], the real pairs of self-corresponding rays intersect on a straight line s. Let an imaginary ray e of the first pencil meet s at E and the corresponding imaginary ray of the second pencil e' meet s at E'. Let the two pairs of real corresponding rays meet s in A, B and let SS' meet s in C.

Then $(ABCE) = (ABCE')$.

Therefore by Art. 5, E and E' coincide and e and e' meet on s.

Correlatively it may be proved that *if the point of intersection of the real bases of two projective ranges, which have two pairs of real corresponding elements, is a self-corresponding point, the connectors of all pairs of corresponding points of the ranges (real or imaginary) pass through a real point.*

If two imaginary lines are corresponding rays of two projective pencils with three pairs of corresponding real rays, the two imaginary straight lines of which they are conjugate imaginary lines are also corresponding rays of the pencils.

Let e and e' be the first pair of corresponding imaginary lines of the pencils. By Art. [37], displace the pencils so that they are in perspective. Then e and e' will intersect on s, the axis of perspective. The point ee' is then an imaginary point since the vertices of the pencils are the real points on e and e'. But s is a real straight line. Hence it passes through the conjugate imaginary point of ee'. Hence the connectors of this point to the vertices of the pencils, i.e., the conjugate imaginary lines of e and e', are corresponding rays of the pencils.

Correlatively it may be proved that *if two imaginary points are corresponding points of two projective ranges with three pairs of corresponding real points, their conjugate imaginary points are also corresponding points of the projective ranges.*

14. The triangle.

The different types of triangles which may occur are as follows, viz.:

(1) *Real Triangle.* A triangle consisting of three real vertices determining three real sides.	*Real Triangle.* A triangle consisting of three real lines determining three real vertices.
(2) *Semi-real Triangle.* A triangle consisting of one real vertex and two conjugate imaginary vertices (on a real line) determining a pair of conjugate imaginary lines (for two sides) and a real line for the third side (on which the pair of conjugate imaginary vertices are situated).	*Semi-real Triangle.* A triangle consisting of one real side and two conjugate imaginary sides (meeting in a real point) determining a pair of conjugate imaginary vertices (lying on the real side) and one real vertex (being the point of intersection of the pair of conjugate imaginary lines which form a pair of sides).
(3) *An imaginary Triangle,* type (*a*). A triangle consisting of one real vertex and two imaginary points situated on a real line determining one real side (on which the imaginary vertices are situated) and a pair of imaginary sides passing through the real vertex.	*An imaginary Triangle,* type (*a*). A triangle consisting of one real side and a pair of imaginary sides passing through a real point, determining one real vertex (through which the pair of imaginary sides pass) and a pair of imaginary vertices on the real side.
(4) *An imaginary Triangle,* type (*b*). A triangle consisting of a pair of conjugate imaginary vertices and a third imaginary vertex, determining one real side (joining the pair of conjugate imaginary vertices) and a pair of imaginary sides.	*An imaginary Triangle,* type (*b*). A triangle consisting of a pair of conjugate imaginary sides and a third imaginary side, determining one real vertex (being the point of intersection of the conjugate imaginary sides) and a pair of imaginary vertices.
(5) *An imaginary Triangle,* type (*c*). A triangle consisting of three imaginary vertices, determining three imaginary sides.	*An imaginary Triangle,* type (*c*). A triangle consisting of three imaginary sides, determining three imaginary vertices.

In all the above cases, except case (4), the triangle and that given by the correlative construction are of the same nature.

15. Extension of Menelaus' theorem.

I. *If an imaginary straight line meets one side of a real triangle in a real point and the other two sides in imaginary points, the product of the ratios of these points, with respect to the triangle, is unity.*

This may at once be deduced from Art. 10. For if a real straight line be drawn through S to meet s and s_1 in A and A_1, the triangle OAA_1 will be a real triangle and the imaginary straight line SEE_1 is a straight line which meets two of the sides in imaginary points E and E_1, and the third side in a real point S.

It follows from Art. 10 that

$$(OAEW') = (OA_1E_1\infty).$$

Therefore $$\frac{OE}{AE} : \frac{OW'}{AW'} = \frac{OE_1}{A_1E_1}.$$

But $\dfrac{OW'}{AW'} = \dfrac{A_1S}{AS}$, by similar triangles.

Therefore $$\frac{OE}{AE} \cdot \frac{AS}{A_1S} \cdot \frac{A_1E_1}{OE_1} = 1.$$

The theorem may also be proved independently as follows:

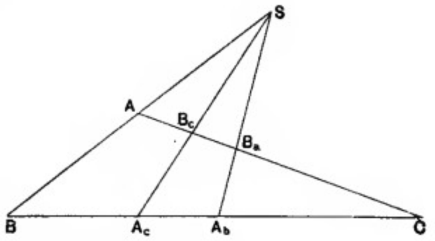

Take a real triangle ABC. Let A_b and A_c be the conjugate points of B and C in an involution (with imaginary double points) on BC.

Take any real point S on AB. From S project A_c and A_b into B_c and B_a.

Then the involution B, A_b, C, A_c is projected from S into A, B_a, C, B_c.

Let the ratios of A_c, A_b be a' and a and of B_c, B_a be b' and b. Then the ratios, x and y, of the double points of the involutions are given by

$$x^2 - 2xa' + aa' = 0$$

and $$y^2 - 2yb + bb' = 0. \qquad \text{(See Art. [61] Ex. (7).)}$$

If c be the ratio of S then

$$cab = 1 \text{ and } ca'b' = 1 \quad\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

The double points are given by

$$(x - a')^2 = a'(a' - a)$$
$$(y - b)^2 = b(b - b'),$$
$$\therefore\ x = a' \pm \sqrt{a'(a' - a)},\ y = b \pm \sqrt{b(b - b')}.$$

But from (1) $$\frac{a}{a'}=\frac{b'}{b}=\frac{1}{ca'b}=\lambda \text{ suppose,}$$

$$\therefore x=a'\pm\sqrt{a'^2(1-\lambda)},\quad y=b\pm\sqrt{b^2(1-\lambda)}.$$

Let $$x_1=a'(1+\sqrt{1-\lambda}),\ y_1=b(1-\sqrt{1-\lambda});$$

$$\therefore x_1y_1=a'b(1-1+\lambda)=a'b\lambda=\frac{a'b}{ca'b},$$

$$\therefore x_1y_1c=1.$$

Similarly if $$x_2=a'(1-\sqrt{1-\lambda}),\ y_2=b(1+\sqrt{1-\lambda}),$$

then $$x_2y_2c=1.$$

II. *If an imaginary straight line meets the sides of a real triangle in three imaginary points, then the product of the ratios of these points with respect to the triangle is unity.*

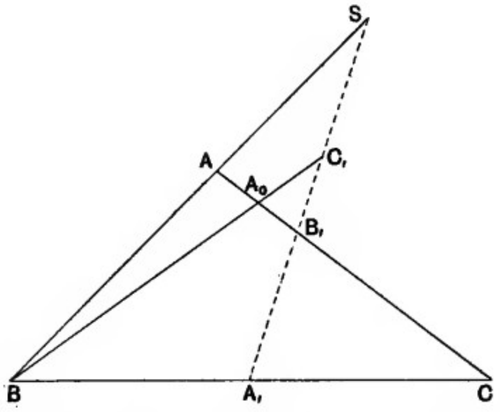

Take any real triangle ABC. Draw through B any real transversal to meet AC in A_0 and the imaginary line SB_1A_1 in C_1, where S is real.

Let the ratios of A_1, B_1, C_1 referred to the triangle A_0BC be x, y, z.

Let $$\frac{AS}{BS}=c.$$

Then from the triangle ABC

$$c\,.\,x\,.\,\frac{CB_1}{AB_1}=1 \text{ or } c\,.\,x\,.\,\frac{CB_1}{A_0B_1}\,.\,\frac{A_0B_1}{AB_1}=1 \text{ or } c\,.\,x\,.\,y\,.\,\frac{A_0B_1}{AB_1}=1,\ldots\text{(i)}$$

and from ABA_0

$$c\,.\,\frac{BC_1}{A_0C_1}\,.\,\frac{A_0B_1}{AB_1}=1 \text{ or } c\,.\,\frac{1}{z}\,.\,\frac{A_0B_1}{AB_1}=1.\ \ldots\ldots\ldots\ldots\text{(ii)}$$

Dividing the expressions (i) and (ii)

$$\frac{1}{xy}=z,$$

$$\therefore\quad xyz=1.$$

Hence the imaginary line $A_1B_1C_1$ meets the sides of the triangle of reference A_0BC in three points (imaginary) for which Menelaus' theorem holds.

16. Extension of Ceva's theorem.

I. *If the connector of any imaginary point to one of the vertices of a real triangle is real, the product of the ratios of the three points in which its connectors to the vertices meet the opposite sides of the triangle is* -1.

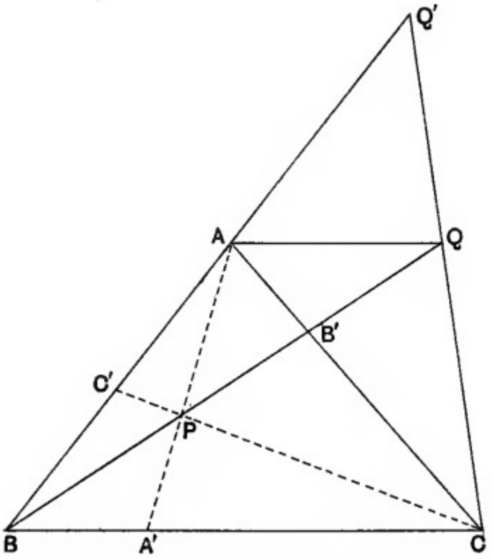

Let P be the imaginary point and let AP, BP, CP, meet the opposite sides of the triangle in A', B', C'. Let BPB' be real. Draw AQ parallel to BC to meet BB' in Q. Join CQ to meet AB in Q'.

Then $\quad (A\,.\,BB'PQ)=(BCA'\infty)$.

But $\quad (C\,.\,BB'PQ)=(BAC'Q')$.

Therefore $\quad (BCA'\infty)=(BAC'Q')$.

$$\therefore\quad \frac{BA'}{CA'}=\frac{BC'}{AC'}:\frac{BQ'}{AQ'}.$$

<< !"#$! ! !% & ' ! E< ? &(" & 9 ' (> /
,((' 0-

!F"." G 4 4 !) * ! <!# &K K L ' K
- 0' ' n >' !> /

<E !"#$! ! !% & ' ! EE " (' -(- (
.'/ 0-

< !"#$! ! !% & ' ! EE "0 7 ' 0-

E !"#$! ! !% & ' ! EE (' -(0 -& 0-
V @ 0'&
%9 : 4 - ;;;

E !"#$! ! !% & ' ! EE 0 ,(" ' / >' 0-
I& < 8

'0 W 7' + ' - # 0 <
. 6 & ' + ' <
' ' 3' 7' + E
. ' 3 # 0 E

3. Hyperbolic Geometry

Suppose $w = f(z)$ is a conformal self-map of the open unit disk $\mathbb{D}$. From Pick's lemma we then have equality in (2.2),

$$\left|\frac{dw}{dz}\right| = \frac{1-|w|^2}{1-|z|^2}\,.$$

In differential form this becomes

$$\frac{|dw|}{1-|w|^2} = \frac{|dz|}{1-|z|^2}\,,$$

10. Show that any conformal self-map of the punctured complex plane $\{0 < |z| < \infty\}$ is either a multiplication $z \mapsto az$, or such a multiplication followed by the inversion $z \mapsto 1/z$.

11. Let $D = \mathbb{C} \backslash \{a_1, \dots, a_m\}$ be the complex plane with m punctures. Show that any conformal self-map of D is a fractional linear transformation that permutes $\{a_1, \dots, a_m, \infty\}$.

12. Determine the conformal self-maps of the following domains D: (a) $D = \mathbb{C} \backslash \{0, 1\}$, (b) $D = \mathbb{C} \backslash \{-1, 0, 1\}$, (c) $D = \mathbb{C} \backslash \{-1, 0, 2\}$.

13. Suppose $f(z)$ is an analytic function from the open unit disk $\mathbb{D}$ to itself that is not the identity map z. Show that $f(z)$ has at most one fixed point in $\mathbb{D}$. *Hint.* Make a change of variable with a conformal self-map of $\mathbb{D}$ to place the fixed point at 0.

14. Suppose $f(z)$ is an analytic function from the open unit disk $\mathbb{D}$ to itself that is not a conformal self-map, and denote by $f_n(z)$ the nth iterate of $f(z)$. Show that if $f(z)$ has a fixed point $z_0 \in \mathbb{D}$, then $f_n(z)$ converges to z_0 for each $z \in \mathbb{D}$. Show that for each $r < 1$, the convergence is uniform for $|z| \le r$. *Hint.* See Exercise 1.8.

15. We say that two conformal self-maps f and g of $\mathbb{D}$ are **conjugate** if there is a conformal self-map h of $\mathbb{D}$ such that $g = h \circ f \circ h^{-1}$. (See the exercises for Section II.7.) Let f be a conformal self-map of $\mathbb{D}$ that is not the identity map z. (a) Show that either f has two fixed points on $\partial\mathbb{D}$, counting multiplicity, or f has one fixed point in $\mathbb{D}$. (b) Show that f has a fixed point in $\mathbb{D}$ if and only if f is conjugate to a rotation $g(z) = e^{i\varphi}z$. (c) Show that rotations by different angles are not conjugate. (d) Show that f has two distinct fixed points on $\partial\mathbb{D}$ if and only if f is conjugate to $g(z) = (z - s)/(1 - sz)$ for some s satisfying $0 < s < 1$. (e) Show that g's for different s's are not conjugate. (f) Show that any two conformal self-maps of $\mathbb{D}$ with one fixed point on $\partial\mathbb{D}$ (of multiplicity two) are conjugate.

3. Hyperbolic Geometry

Suppose $w = f(z)$ is a conformal self-map of the open unit disk $\mathbb{D}$. From Pick's lemma we then have equality in (2.2),

$$\left|\frac{dw}{dz}\right| = \frac{1 - |w|^2}{1 - |z|^2}.$$

In differential form this becomes

$$\frac{|dw|}{1 - |w|^2} = \frac{|dz|}{1 - |z|^2},$$

which means that if γ is any smooth curve in $\mathbb{D}$, and $w = f(z)$ is a conformal self-map of $\mathbb{D}$, then

$$\int_{f\circ\gamma} \frac{|dw|}{1-|w|^2} \;=\; \int_\gamma \frac{|dz|}{1-|z|^2}\,. \tag{3.1}$$

Thus to obtain a length function that is invariant under conformal self-maps of $\mathbb{D}$, we are led to make the following definition. We define the **length of γ in the hyperbolic metric** by

$$\text{hyperbolic length of } \gamma \;=\; 2\int_\gamma \frac{|dz|}{1-|z|^2}\,. \tag{3.2}$$

The factor 2 is a harmless factor, which is often omitted. (It adjusts the metric so that its curvature is -1.) The identity (3.1) shows that $f\circ\gamma$ has the same hyperbolic length as γ for any conformal self-map $f(z)$ of $\mathbb{D}$. Thus hyperbolic lengths are invariant under conformal self-maps of $\mathbb{D}$.

We define the **hyperbolic distance** $\rho(z_0, z_1)$ from z_0 to z_1 to be the infimum (greatest lower bound) of the hyperbolic lengths of all piecewise smooth curves in $\mathbb{D}$ from z_0 to z_1. Since conformal self-maps of $\mathbb{D}$ preserve the hyperbolic lengths of curves, they also preserve hyperbolic distances; that is, for any conformal self-map $w = f(z)$ of $\mathbb{D}$,

$$\rho(f(z_0), f(z_1)) \;=\; \rho(z_0, z_1), \qquad z_0, z_1 \in \mathbb{D}.$$

Theorem. *For any two distinct points z_0, z_1 in the open unit disk $\mathbb{D}$, there is a unique shortest curve in $\mathbb{D}$ from z_0 to z_1 in the hyperbolic metric, namely, the arc of the circle passing through z_0 and z_1 that is orthogonal to the unit circle.*

The paths of shortest hyperbolic length between points are called **hyperbolic geodesics**. The hyperbolic geodesics play the role that straight lines play in the Euclidean geometry of the plane. They satisfy all the axioms of Euclidean geometry except the parallel axiom (that through each point not on a given line there passes a unique straight line through the point and parallel to the given line).

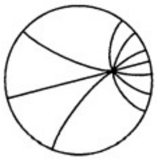

hyperbolic geodesics

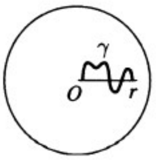

For a proof of the theorem, let $w = f(z)$ be a conformal self-map of $\mathbb{D}$ such that $f(z_0) = 0$. By multiplying by a unimodular constant, we can arrange that $f(z_1) = r > 0$. Since $f(z)$ preserves hyperbolic lengths, and since $f(z)$ maps circles orthogonal to the unit circle onto circles orthogonal to the unit circle, it suffices to show that the straight line segment from 0 to r is a unique path of shortest hyperbolic length from 0 to r. For this, let $\gamma(t) = x(t) + iy(t)$, $0 \leq t \leq 1$, be a piecewise smooth path in $\mathbb{D}$ from 0 to r. Then $\alpha(t) = \mathrm{Re}(\gamma(t)) = x(t)$ defines a path in $\mathbb{D}$ from 0 to r along the real axis, and

$$\int_\alpha \frac{|dz|}{1-|z|^2} = \int_0^1 \frac{|dx(t)|}{1-x(t)^2} \leq \int_0^1 \frac{|dx(t)|}{1-|\gamma(t)|^2} \leq \int_\gamma \frac{|dz|}{1-|z|^2}.$$

If $y(t) \neq 0$ for some t, then $|\gamma(t)| > |x(t)|$, and the first inequality above is strict. In this case, the path $\alpha(t)$ on the real axis is strictly shorter than the path $\gamma(t)$. Further, if $\alpha(t)$ is decreasing on some interval, we could reduce the integral by deleting a parameter interval over which $\alpha(t)$ starts and ends at the same value. We conclude that the integral is a minimum exactly when $\gamma(t)$ is real and nondecreasing, in which case the path is the straight line segment from 0 to r.

We turn now to an important reinterpretation of Pick's lemma.

Theorem. *Every analytic function $w = f(z)$ from the open unit disk $\mathbb{D}$ to itself is a contraction mapping with respect to the hyperbolic metric ρ,*

$$\rho(f(z_0), f(z_1)) \leq \rho(z_0, z_1), \qquad z_0, z_1 \in \mathbb{D}. \tag{3.3}$$

Further, there is strict inequality for all points $z_0, z_1 \in \mathbb{D}$, $z_0 \neq z_1$, unless $f(z)$ is a conformal self-map of $\mathbb{D}$, in which case there is equality for all $z_0, z_1 \in \mathbb{D}$.

To see this, let γ be the geodesic from z_0 to z_1. Then $f \circ \gamma$ is a curve from $f(z_0)$ to $f(z_1)$. Pick's lemma and the definition of the hyperbolic metric yield

$$\begin{aligned} \rho(f(z_0), f(z_1)) &\leq 2\int_{f\circ\gamma} \frac{|dw|}{1-|w|^2} = 2\int_\gamma \frac{|f'(z)|\,|dz|}{1-|f(z)|^2} \\ &\leq 2\int_\gamma \frac{|dz|}{1-|z|^2} = \rho(z_0, z_1). \end{aligned}$$

If $f(z)$ is not a conformal self-map of $\mathbb{D}$, there is strict inequality in Pick's lemma, and we obtain strict inequality in this estimate, hence in (3.3).

The hyperbolic distance from 0 to z can be computed explicitly. It is

$$\rho(0, z) = 2\int_0^{|z|} \frac{dt}{1-t^2} = \int_0^{|z|} \left[\frac{1}{1-t} + \frac{1}{1+t}\right] dt = \log\left(\frac{1+|z|}{1-|z|}\right).$$

This shows that the hyperbolic distance from 0 to z tends to $+\infty$ when z tends to the boundary of the unit disk.

A **geodesic triangle** is an area bounded by three hyperbolic geodesics. Since the hyperbolic geodesics and the angles between them are preserved

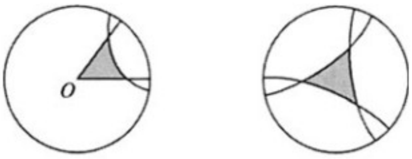

hyperbolic triangles

by conformal self-maps of $\mathbb{D}$, we can map any geodesic triangle to a triangle with vertex at 0 and with the same angles between sides. For a geodesic triangle with vertex at 0, two of the sides are radial segments, and the third is an arc of a circle lying inside the Euclidean triangle with the two radii as sides. From this representation we see that the sum of the angles of any geodesic triangle is strictly less than π, which is the sum of the angles of the corresponding Euclidean triangle.

In connection with complex analysis, we have now been in contact with three spaces with strikingly different geometries. The first space is the complex plane $\mathbb{C}$ with the usual Euclidean metric $|dz|$. In the Euclidean plane, the geodesics are straight lines, and the sum of the angles of a geodesic triangle is exactly equal to π. The second space is the open unit disk $\mathbb{D}$ with the hyperbolic metric $2|dz|/(1-|z|^2)$. For the hyperbolic disk, the geodesics are arcs of circles orthogonal to the unit circle, and the sum of angles of a geodesic triangle is strictly less than π.

The third space is the extended complex plane $\mathbb{C}^* = \mathbb{C} \cup \{\infty\}$ with the spherical metric, which can be introduced in a manner completely analogous to the hyperbolic metric. Recall (Section I.3) that the chordal metric induced on $\mathbb{C}$ by the Euclidean metric of the sphere via the stereographic projection is given explicitly by

$$\text{chordal distance from } z \text{ to } w \;=\; \frac{2|z-w|}{\sqrt{1+|z|^2}\sqrt{1+|w|^2}}.$$

The infinitesimal form of this metric is $2|dz|/(1+|z|^2)$. If γ is a path in $\mathbb{C}^*$, its **length in the spherical metric** is

$$\text{spherical length of } \gamma \;=\; 2\int_\gamma \frac{|dz|}{1+|z|^2} \;=\; 2\int \frac{|\gamma'(t)|}{1+|\gamma(t)|^2}\,dt\,.$$

This is the length of the corresponding path on the unit sphere in $\mathbb{R}^3$. The **distance from z_1 to z_2 in the spherical metric** is defined to be the infimum of the spherical lengths of the paths joining z_1 to z_2. Since the chordal metric is invariant under rotations of the sphere, so is the spherical metric, and consequently, the lengths of paths and the distances between points in the spherical metric are invariant under rotations. It is not difficult to show that the geodesics in the spherical metric correspond to great circles on the sphere, and the sum of the angles of a geodesic triangle is strictly greater than π.

Each of these three spaces is homogeneous, in the sense that any prescribed point can be transported to any other by an "isometry." Thus for each of these spaces, any scalar quantity that is invariant under isometries is constant. It turns out that a notion of "scalar curvature" can be associated to each of the spaces (see the exercises), and the curvature is invariant under isometries, so that in each case the curvature is constant. The Euclidean plane has constant zero curvature, the sphere has constant positive curvature, and the hyperbolic disk has constant negative curvature. The curvature can be related to the area and the sum of angles of geodesic triangles (Gauss-Bonnet formula). We summarize these properties in tabular form.

Geometry	Euclidean	Spherical	Hyperbolic
Infinitesimal length	$\lvert dz\rvert$	$\dfrac{2\lvert dz\rvert}{1+\lvert z\rvert^2}$	$\dfrac{2\lvert dz\rvert}{1-\lvert z\rvert^2}$
Oriented isometries	$e^{i\varphi}z+b$	rotations	conformal self-maps
Curvature	0	$+1$	-1
Geodesics	lines	great circles	circles $\perp$ unit circle
Angles of triangle	$=\pi$	$>\pi$	$<\pi$
Disk circumference	$2\pi\rho$	$2\pi\rho-\dfrac{\pi\rho^3}{3}+\mathcal{O}(\rho^5)$	$2\pi\rho+\dfrac{\pi\rho^3}{3}+\mathcal{O}(\rho^5)$

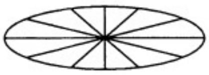
Euclidean disk

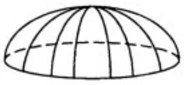
spherical disk

hyperbolic disk

Exercises for IX.3

1. Show by direct computation that $|w'(z)| = (1-|w|^2)/(1-|z|^2)$ for any conformal self-map $w = f(z)$ of $\mathbb{D}$.

2. A **hyperbolic disk** centered at $z_0 \in \mathbb{D}$ of radius $\rho > 0$ consists of all $z \in \mathbb{D}$ such that $\rho(z, z_0) < \rho$. (a) Show that the hyperbolic disk centered at 0 of radius ρ is a Euclidean disk of radius $r = (e^\rho - 1)/(e^\rho + 1)$. (b) Show that any hyperbolic disk is a Euclidean disk.

3. Denote by $c(z,\rho)$ and $r(z,\rho)$ the Euclidean center and Euclidean radius of the hyperbolic disk centered at z of hyperbolic radius ρ. (a) For fixed ρ, show that $r(z,\rho)/(1-|z|)$ tends to a constant $A > 0$ as $|z| \to 1$. (b) For fixed ρ, show that $|z - c(z,\rho)|/r(z,\rho)$ tends to a constant B, $0 < B < 1$, as $|z| \to 1$.

4. Show that the circumference of a hyperbolic disk of radius ρ is $2\pi \sinh \rho$. *Hint.* Show first that the hyperbolic circumference of a Euclidean disk of radius r centered at 0 is $4\pi r/(1-r^2)$.

5. We define the **hyperbolic area** of a subset E of $\mathbb{D}$ to be

$$4 \iint_E \frac{dx\, dy}{(1-|z|^2)^2}.$$

Show that the hyperbolic area is invariant under conformal self-maps of $\mathbb{D}$. Show that the hyperbolic area of a hyperbolic disk of radius ρ is given by

$$2\pi(\cosh \rho \;-\; 1) \;=\; \pi\rho^2 + \frac{\pi}{12}\rho^4 + \mathcal{O}(\rho^6).$$

6. Establish the following, for the spherical metric. (a) The circumference of a spherical disk of radius ρ is $2\pi \sin \rho$, $0 < \rho < \pi$. (b) The area of a spherical disk of radius ρ is given by

$$2\pi(1-\cos \rho) \;=\; \pi\rho^2 - \frac{\pi}{12}\rho^4 + \mathcal{O}(\rho^6).$$

(c) The geodesics in the spherical metric correspond to great circles on the sphere. *Hint.* It suffices to show that the shortest curve from 0 to ε in the spherical metric is the straight line segment joining them.

7. Show that an isometry of the hyperbolic disk $\mathbb{D}$ is either a conformal self-map of $\mathbb{D}$ or the composition of a conformal self-map and the reflection $z \mapsto \bar{z}$.

8. Let $f(z) = (az+b)/(cz+d)$, where $ad-bc=1$. Show that $f(z)$ is an isometry in the spherical metric if and only if the matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is unitary.

9. Show that the function $f(z) = z^2$ is strictly contracting with respect to the hyperbolic metric on any subdisk $\{|z| \le r\}$, $0 < r < 1$, and that any branch of the square root function is strictly expanding, by establishing the following. (a) For fixed r, $0 < r < 1$, show that

$$\rho(z^2, \zeta^2) \;\le\; \frac{2r}{1+r^2}\rho(z,\zeta), \qquad |z|, |\zeta| \le r.$$

When does equality hold? (b) Show that the constant $2r/(1+r^2)$ in (a) is sharp. (c) For fixed s, $0 < s < 1$, show that

$$\rho\left(\pm\sqrt{z}, \pm\sqrt{\zeta}\right) \;\ge\; \frac{1+s}{2\sqrt{s}}\rho(z,\zeta), \qquad |z|, |\zeta| \le s.$$

10. Show that

$$d(z,w) \;=\; \left|\frac{z-w}{1-\bar{w}z}\right|, \qquad |z|,|w|<1,$$

satisfies the triangle inequality, that is, $d(z,w) \le d(z,\zeta)+d(\zeta,w)$ for all $z,\zeta,w \in \mathbb{D}$. *Remark.* This can be regarded as the analogue of the chordal metric for the sphere (defined in Section I.3). Except for the constant factor 2, the hyperbolic metric is the infinitesimal version of the metric function $d(z,w)$.

11. Show that the metric function $d(z,w)$ defined in the preceding exercise satisfies

$$d(f(z),f(w)) \le d(z,w), \qquad |z|,|w|<1,$$

for any analytic function $f(z)$ from $\mathbb{D}$ to $\mathbb{D}$. Show that equality obtains whenever $f(z)$ is a conformal self-map of $\mathbb{D}$, and otherwise there is strict inequality for all $z \ne w$.

12. A conformal map $g(z)$ of a domain D onto the open unit disk $\mathbb{D}$ induces the metric ρ_D on D defined by

$$d\rho_D(z) = \frac{2|g'(z)|}{1-|g(z)|^2}|dz|, \qquad z \in D.$$

Show that ρ_D is independent of the conformal map $g(z)$ of D onto $\mathbb{D}$. *Remark.* The metric ρ_D is called the **hyperbolic metric** of the simply connected domain D.

13. Show that the hyperbolic metric of the upper half-plane $\mathbb{H}$ is given by

$$d\rho_{\mathbb{H}}(z) = \frac{|dz|}{y}, \qquad z = x+iy,\ y>0.$$

What are the geodesics in the hyperbolic metric? Illustrate with a sketch.

14. Show that the horizontal strip $S = \{-\pi/2 < \operatorname{Im} z < \pi/2\}$ has hyperbolic metric

$$d\rho_S(z) = \frac{|dz|}{\cos y}, \qquad z = x+iy,\ -\pi/2 < y < \pi/2.$$

Sketch the hyperbolic geodesics that are orthogonal to the vertical interval $\{iy : -\pi/2 < y < \pi/2\}$.

15. The **curvature** of the metric $\sigma(z)|dz|$ is defined to be

$$\kappa(z) \;=\; -\frac{1}{\sigma(z)^2}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)\log\sigma(z).$$

Find the curvature of each of the spherical, the hyperbolic, and the Euclidean metrics.

16. (Wolff-Denjoy Theorem.) Let $f(z)$ be an analytic function from $\mathbb{D}$ to $\mathbb{D}$. Let $f_n(z)$ denote the nth iterate of $f(z)$, and let K_r denote the closed disk $\{|z| \leq r\}$.
    (a) Show that if $f(z)$ is not a conformal self-map of $\mathbb{D}$, then for any $r < 1$ there is a constant $c < 1$ such that $\rho(f(z), f(w)) \leq c\rho(z, w)$ for $z, w \in K_r$.
    (b) Show that if the image $f(\mathbb{D})$ is contained in K_r for some $r < 1$, then the iterates $f_n(z)$ converge uniformly on $\mathbb{D}$ to a fixed point for $f(z)$.
    (c) Show that if $f(z)$ is not a conformal self-map of $\mathbb{D}$, and if there is $r < 1$ such that the iterates of some point $z_0 \in \mathbb{D}$ visit K_r infinitely often, then the iterates $f_n(z)$ converge normally on $\mathbb{D}$ to a fixed point of $f(z)$. *Hint.* First find the fixed point.
    (d) Show that if the iterates of some point $z_0 \in \mathbb{D}$ tend to the unit circle $\partial\mathbb{D}$, then there is a point $\zeta \in \partial\mathbb{D}$ (the **Wolff-Denjoy point**) such that the iterates $f_n(z)$ converge normally on $\mathbb{D}$ to ζ. *Hint.* Suppose $z_0 = 0$. Define $g_\varepsilon(z) = (1-\varepsilon)f(z)$, let z_ε be the fixed point of $g_\varepsilon(z)$, and let D_ε be the hyperbolic disk centered at z_ε with 0 on its boundary. Show that the limit D of the D_ε's is a Euclidean disk that is invariant under $f(z)$ and whose boundary meets $\partial\mathbb{D}$ in exactly one point.

X

Harmonic Functions and the Reflection Principle

In Section 1 we introduce the Poisson kernel function and we develop the Poisson integral representation for harmonic functions on the open unit disk. The Poisson kernel is the analogue for harmonic functions of the Cauchy kernel for analytic functions, and the Poisson integral formula solves the Dirichlet problem for the unit disk. In Section 2 we use this solution to characterize harmonic functions by the mean value property. This characterization is the analogue of Morera's theorem characterizing analytic functions. In Section 3 we apply the characterization of harmonic functions to establish the Schwarz reflection principle for harmonic functions. The reflection principle plays a key role in the study of boundary behavior of conformal maps.

1. The Poisson Integral Formula

We wish to extend a given continuous complex-valued function $h\left(e^{i\theta}\right)$ on the unit circle continuously to the closed unit disk $\{|z| \leq 1\}$ so as to be harmonic on the interior of the disk. Any such extension is unique, since the difference of two such extensions is zero on the boundary, hence zero on the entire disk, by the maximum principle. Our strategy is to derive a formula for the extension in the case that $h\left(e^{i\theta}\right)$ is a trigonometric polynomial, and then to show that the formula provides an extension even when $h\left(e^{i\theta}\right)$ is only continuous.

We start with the trigonometric monomial $e^{ik\theta}$. A harmonic extension is given explicitly by $r^{|k|}e^{ik\theta}$. If $k \geq 0$ this extension is $r^k e^{ik\theta} = z^k$, which is analytic. If $k < 0$ this extension is $r^{-k}e^{ik\theta} = \bar{z}^{(-k)}$, which is conjugate-analytic hence harmonic. Proceeding by linearity, we see that the trigonometric polynomial $h\left(e^{i\theta}\right) = \sum_{k=-N}^{N} a_k e^{ik\theta}$ has the (unique) harmonic extension

$$\tilde{h}\left(re^{i\theta}\right) \;=\; \sum_{k=-N}^{N} a_k r^{|k|} e^{ik\theta}.$$

We capture the coefficient a_m by multiplying $h\left(e^{i\theta}\right)$ by $e^{-im\theta}$ and integrating. The orthogonality relations for complex exponentials (Section VI.6) yield

$$a_m \;=\; \int_{-\pi}^{\pi} h\left(e^{i\theta}\right) e^{-im\theta} \frac{d\theta}{2\pi}\,.$$

Substituting this expression into the formula for $\tilde{h}\left(re^{i\theta}\right)$, we obtain

$$\begin{aligned}\tilde{h}(re^{i\theta}) \;&=\; \sum_{k=-\infty}^{\infty} \left(\int_{-\pi}^{\pi} h\left(e^{i\varphi}\right) e^{-ik\varphi} \frac{d\varphi}{2\pi} \right) r^{|k|} e^{ik\theta} \\ &=\; \int_{-\pi}^{\pi} h\left(e^{i\varphi}\right) \left[\sum_{k=-\infty}^{\infty} r^{|k|} e^{-ik\varphi} e^{ik\theta} \right] \frac{d\varphi}{2\pi}\,.\end{aligned}$$

In order to simplify this expression, we introduce the **Poisson kernel function** defined by

$$P_r(\theta) \;=\; \sum_{k=-\infty}^{\infty} r^{|k|} e^{ik\theta}. \tag{1.1}$$

For each fixed $\rho < 1$, this series converges uniformly for $r \leq \rho$ and $-\pi \leq \theta \leq \pi$, by the Weierstrass M-test, since then $\left| r^{|k|} e^{ik\theta} \right| \leq \rho^{|k|}$. In terms of the Poisson kernel, the formula for the harmonic extension $\tilde{h}\left(re^{i\theta}\right)$ becomes

$$\tilde{h}(re^{i\theta}) \;=\; \int_{-\pi}^{\pi} h\left(e^{i\varphi}\right) P_r(\theta - \varphi) \frac{d\varphi}{2\pi}\,, \qquad re^{i\theta} \in \mathbb{D}. \tag{1.2a}$$

If we make a change of variable $\varphi \mapsto \theta - \varphi$ and use the 2π-periodicity of $P_r(\theta)$, we obtain an alternative form of (1.2a),

$$\tilde{h}(re^{i\theta}) \;=\; \int_{-\pi}^{\pi} h\left(e^{i(\theta-\varphi)}\right) P_r(\varphi) \frac{d\varphi}{2\pi}\,, \qquad re^{i\theta} \in \mathbb{D}. \tag{1.2b}$$

Now we look at the Poisson kernel function $P_r(\theta)$ more closely. Setting $z = re^{i\theta}$ and setting $j = -k$ when $k < 0$ in (1.1), we obtain

$$P_r(\theta) \;=\; 1 + \sum_{k=1}^{\infty} z^k + \sum_{j=1}^{\infty} \bar{z}^j.$$

Summing these two geometric series, we obtain

$$P_r(\theta) \;=\; 1 + \frac{z}{1-z} + \frac{\bar{z}}{1-\bar{z}}\,, \qquad z = re^{i\theta} \in \mathbb{D}. \tag{1.3}$$

Putting this over the common denominator

$$|1-z|^2 \;=\; (1-z)(1-\bar{z}) \;=\; 1+r^2-2r\cos\theta,$$

we obtain

$$P_r(\theta) \;=\; \frac{1-|z|^2}{|1-z|^2} \;=\; \frac{1-r^2}{1+r^2-2r\cos\theta}, \qquad z = re^{i\theta}\in\mathbb{D}. \tag{1.4}$$

Since the Poisson kernel is 2π-periodic, we focus on its behavior on the period interval $-\pi \le \theta \le \pi$. From the Poisson integral representation formula (1.2b) for the constant function $h \equiv 1$, we have

$$\int_{-\pi}^{\pi} P_r(\theta)\frac{d\theta}{2\pi} \;=\; 1. \tag{1.5}$$

Three other properties of the Poisson kernel, which follow immediately from the formula (1.4), are

$$P_r(\theta) > 0, \qquad -\pi \le \theta \le \pi, \tag{1.6}$$

$$P_r(-\theta) = P_r(\theta), \qquad -\pi \le \theta \le \pi, \tag{1.7}$$

(1.8) $P_r(\theta)$ is increasing for $-\pi \le \theta \le 0$ and decreasing for $0 \le \theta \le \pi$.

Since $P_r(\delta) \to 0$ as $r \to 1$ for each fixed $\delta > 0$, we obtain from (1.8) the following key property of the Poisson kernel:

(1.9) for fixed $\delta > 0$, $\max\{\, P_r(\theta) \,:\, \delta \le |\theta| \le \pi \,\} \to 0$ as $r \to 1$.

These five properties, (1.5) through (1.9), are fundamental for understanding the Poisson kernel. Properties (1.5) and (1.6) together show that each $P_r(\theta)d\theta/2\pi$ is a "probability measure," that is, a positive "mass distribution" with unit total mass. Properties (1.5), (1.6), and (1.9) show that the family of functions $P_r(\theta)$ form an "approximate identity," in the sense that the mass of the probability measure $P_r(\theta)d\theta/2\pi$ concentrates at the point $\theta = 0$ as $r \to 1$. Identity (1.7) reflects the symmetry of the Poisson kernel.

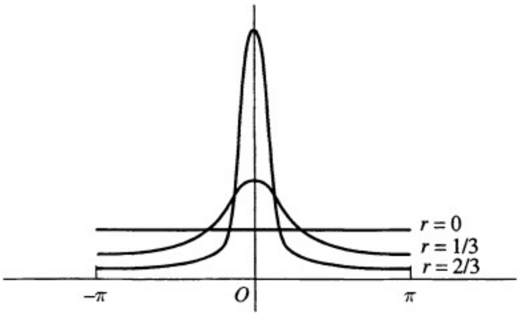

Another important property of the Poisson kernel function is that $P_r(\theta)$ is a harmonic function of $z = re^{i\theta}$ for $z \in \mathbb{D}$. In fact, (1.3) yields an explicit representation of the Poisson kernel function as the real part of an analytic function,

$$P_r(\theta) \;=\; 1 + 2\,\mathrm{Re}\left(\frac{z}{1-z}\right) \;=\; \mathrm{Re}\left(\frac{1+z}{1-z}\right), \qquad z = re^{i\theta} \in \mathbb{D}.$$

The Poisson integral formula (1.2a) becomes

$$\tilde{h}(re^{i\theta}) \;=\; \int_{-\pi}^{\pi} h\left(e^{i\varphi}\right) \mathrm{Re}\left(\frac{1 + re^{i(\theta-\varphi)}}{1 - re^{i(\theta-\varphi)}}\right) \frac{d\varphi}{2\pi}.$$

Thus

$$\tilde{h}(z) \;=\; \int_{-\pi}^{\pi} h\left(e^{i\varphi}\right) \mathrm{Re}\left(\frac{e^{i\varphi}+z}{e^{i\varphi}-z}\right) \frac{d\varphi}{2\pi}, \qquad z \in \mathbb{D}. \tag{1.10}$$

If we substitute for $h\left(e^{i\theta}\right)$ a real-valued trigonometric polynomial $u\left(e^{i\theta}\right)$ in this formula, we may take real parts after integrating, and we obtain

$$\tilde{u}(z) \;=\; \mathrm{Re} \int_{-\pi}^{\pi} u(e^{i\varphi})\, \frac{e^{i\varphi}+z}{e^{i\varphi}-z}\, \frac{d\varphi}{2\pi}, \qquad z \in \mathbb{D}. \tag{1.11}$$

The integral depends analytically on the parameter z. Thus (1.11) expresses the Poisson integral $\tilde{u}(z)$ as the real part of an explicit analytic function.

Now we change our point of view. Instead of focusing on trigonometric polynomials, we consider an arbitrary continuous complex-valued function $h\left(e^{i\theta}\right)$ on the unit circle. We define the **Poisson integral** $\tilde{h}(z)$ of $h\left(e^{i\theta}\right)$ to be the function on the open unit disk $\mathbb{D}$ given by

$$\tilde{h}(z) \;=\; \int_{-\pi}^{\pi} h\left(e^{i\varphi}\right) P_r(\theta-\varphi) \frac{d\varphi}{2\pi}, \qquad z = re^{i\theta} \in \mathbb{D}. \tag{1.12}$$

This is just (1.2a), and it can be rewritten as (1.2b). Formula (1.10) still holds, and for real-valued continuous functions $u\left(e^{i\theta}\right)$, formula (1.11) holds.

The correspondence $h \mapsto \tilde{h}$ is linear, that is, the Poisson integral of $c_1h_1 + c_2h_2$ is $c_1\widetilde{h_1} + c_2\widetilde{h_2}$. Also, observe that the maximum principle holds, in the sense that if $|h\left(e^{i\theta}\right)| \le M$ for all θ, then $|\tilde{h}(z)| \le M$ for all $z \in \mathbb{D}$. This follows from (1.5) and (1.6), that is, from the fact that the Poisson kernel determines a probability measure.

Theorem. *Let $h\left(e^{i\theta}\right)$ be a continuous function on the unit circle. Then the Poisson integral $\tilde{h}(z)$ defined by (1.12) is a harmonic function on the open unit disk that has boundary values $h\left(e^{i\theta}\right)$, that is, $\tilde{h}(z)$ tends to $h(\zeta)$ as $z \in \mathbb{D}$ tends to $\zeta \in \partial\mathbb{D}$.*

The harmonicity can be checked by differentiating under the integral sign in (1.12). It can also be seen by decomposing $h\left(e^{i\theta}\right)$ into its real and imaginary parts, $h\left(e^{i\theta}\right) = u\left(e^{i\theta}\right) + iv\left(e^{i\theta}\right)$. Formula (1.11) shows that

$\tilde{u}(z)$ and $\tilde{v}(z)$ are harmonic, and hence $\tilde{h}(z) = \tilde{u}(z) + i\tilde{v}(z)$ is harmonic for $|z| < 1$.

Since the statement about the boundary values is important, we sketch two proofs. The first proof depends upon the fact that any continuous function on the unit circle can be approximated uniformly by trigonometric polynomials. (One way to see this is to approximate the continuous function by a smooth function and then to appeal to the result proved in Section VI.6 that the Fourier series of any smooth function on the unit circle converges uniformly to the function.) The second proof illustrates how the properties of an approximate identity are typically used.

The first proof boils down to the fact that a uniform limit of continuous functions is continuous. It runs as follows. Let $h\left(e^{i\theta}\right)$ be continuous, and let $\varepsilon > 0$. Let $g\left(e^{i\theta}\right) = \sum a_k e^{ik\theta}$ be a trigonometric polynomial such that $\left|h\left(e^{i\theta}\right) - g\left(e^{i\theta}\right)\right| \leq \varepsilon$ for all $e^{i\theta} \in \partial\mathbb{D}$. Then $\left|\tilde{h}(z) - \tilde{g}(z)\right| \leq \varepsilon$ for all $z \in \mathbb{D}$. Now, $\tilde{g}\left(re^{i\theta}\right) = \sum a_k r^{|k|} e^{ik\theta}$ attains the boundary values $g\left(e^{i\theta}\right)$ continuously on $\partial\mathbb{D}$. Hence the composite function equal to $h\left(e^{i\theta}\right)$ on $\partial\mathbb{D}$ and $\tilde{h}\left(re^{i\theta}\right)$ on $\mathbb{D}$ can be approximated to within ε by a continuous function. Consequently, the values $\tilde{h}(z)$ cluster to within ε of $h\left(e^{i\theta}\right)$ as $z \in \mathbb{D}$ tends to $e^{i\theta} \in \partial\mathbb{D}$. Since $\varepsilon > 0$ is arbitrary, $\tilde{h}(z)$ tends to $h\left(e^{i\theta}\right)$ as $z \in \mathbb{D}$ tends to $e^{i\theta} \in \partial\mathbb{D}$.

The second proof depends only on properties (1.5), (1.6), and (1.9) of the Poisson kernel. Again let $\varepsilon > 0$. Choose $M > 0$ such that $\left|h\left(e^{i\theta}\right)\right| \leq M$. Since $h\left(e^{i\theta}\right)$ is uniformly continuous, we can choose $\delta > 0$ so small that $\left|h\left(e^{i\theta}\right) - h\left(e^{i\varphi}\right)\right| < \varepsilon$ whenever $|\theta - \varphi| < \delta$. Using (1.5) we have

$$\tilde{h}(z) - h\left(e^{i\theta}\right) \;=\; \int_{-\pi}^{\pi} \left[h\left(e^{i(\theta-\varphi)}\right) - h\left(e^{i\theta}\right)\right] P_r(\varphi)\frac{d\varphi}{2\pi}\,.$$

Now we take absolute values, break the integral cleverly into pieces, and make the obvious estimate on each piece, to obtain

$$\begin{aligned}
|\tilde{h}\left(re^{i\theta}\right) - h\left(e^{i\theta}\right)| \;&\leq\; \left(\int_{-\delta}^{\delta} + \int_{\delta\leq|\varphi|\leq\pi}\right)\left|h\left(e^{i(\theta-\varphi)}\right) - h\left(e^{i\theta}\right)\right| P_r(\varphi)\frac{d\varphi}{2\pi} \\
&\leq\; \int_{-\delta}^{\delta} \varepsilon P_r(\varphi)\frac{d\varphi}{2\pi} + 2M \max_{\delta\leq|\varphi|\leq\pi} P_r(\varphi) \int_{\delta\leq|\varphi|\leq\pi} \frac{d\varphi}{2\pi} \\
&\leq\; \varepsilon + 2M \max_{\delta\leq|\varphi|\leq\pi} P_r(\varphi).
\end{aligned}$$

By (1.9), the summand on the right tends to 0 as $r \to 1$. Thus again the values $\tilde{h}(z)$ cluster to within ε of $h\left(e^{i\theta}\right)$ as $z \to e^{i\theta}$, this for any $\varepsilon > 0$, so $\tilde{h}(z) \to h\left(e^{i\theta}\right)$ as $z \to e^{i\theta}$.

Exercises for X.1

1. Show that the Fourier coefficients of $P_r(\theta)$ are $c_k = r^{|k|}$, $-\infty < k < \infty$. (Here we fix r, $0 \leq r < 1$, and we regard $P_r(\theta)$ as a function on the circle.)

2. Let $R > 0$, and let $h\left(Re^{i\theta}\right)$ be a continuous function on the circle $\{|z| = R\}$. Show that the function

$$\tilde{h}(z) \;=\; \int_{-\pi}^{\pi} \frac{R^2 - r^2}{R^2 + r^2 - 2rR\cos(\theta - \varphi)} h\left(Re^{i\varphi}\right) \frac{d\varphi}{2\pi}, \qquad |z| < R,$$

   is harmonic on the disk $\{|z| < R\}$ and has boundary values $h\left(Re^{i\theta}\right)$ on the boundary circle.

3. Suppose that $f(z) = u(z) + iv(z)$ is analytic for $|z| < 1$ and that $u(z)$ extends to be continuous on the closed disk $\{|z| \leq 1\}$. Show that

$$f(z) \;=\; \int_0^{2\pi} u(e^{i\varphi}) \frac{e^{i\varphi} + z}{e^{i\varphi} - z} \frac{d\varphi}{2\pi} + iv(0), \qquad |z| < 1.$$

   Remark. This is the **Schwarz formula**, expressing an analytic function in terms of the boundary values of its real part.

4. Let $\{f_n(z) = u_n(z) + iv_n(z)\}$ be a sequence of analytic functions on the open unit disk $\mathbb{D}$ such that $u_n(z)$ extends continuously to $\partial\mathbb{D}$, $u_n\left(e^{i\theta}\right)$ converges uniformly on $\partial\mathbb{D}$ to $u\left(e^{i\theta}\right)$, and $v_n(0)$ converges. Show that $f_n(z)$ converges normally on $\mathbb{D}$ to an analytic function $f(z)$ whose real part is $\tilde{u}(z)$. *Hint.* Use the Schwarz formula.

5. Let $h\left(e^{i\theta}\right)$ be a piecewise continuous function (or an integrable function) on the unit circle. Show that the Poisson integral $\tilde{h}(z)$ tends to $h(\zeta)$ as $z \in \mathbb{D}$ tends to any point ζ of the unit circle at which $h\left(e^{i\theta}\right)$ is continuous.

6. A function $f(z)$, $z \in \mathbb{D}$, is said to have **radial limit** L at $\zeta \in \partial\mathbb{D}$ if $f(r\zeta) \to L$ as r increases to 1. Let $h\left(e^{i\theta}\right)$ be a piecewise continuous function on the unit circle. Show that $\tilde{h}(z)$ has a radial limit at each $\zeta \in \partial\mathbb{D}$, equal to the average of the limits of $h\left(e^{i\theta}\right)$ at ζ from each side.

7. For each $t > 0$, define the kernel function $C_t(s)$ on the real line by

$$C_t(s) \;=\; \frac{t}{\pi} \frac{1}{s^2 + t^2}, \qquad -\infty < s < \infty.$$

   For $h(\xi)$ a bounded piecewise continuous function on the real line, define a function $\tilde{h}(s + it)$ on the open upper half-plane $\mathbb{H}$ by

$$\tilde{h}(s + it) \;=\; \int_{-\infty}^{\infty} C_t(s - \xi) h(\xi)\, d\xi, \qquad s + it \in \mathbb{H}.$$

(a) Sketch the graph of $C_t(s)$ for $t = 1$, $t = 0.1$, and $t = 0.01$. (b) Show that $C_t(s) > 0$ and $\int_{-\infty}^{\infty} C_t(s)\, ds = 1$. (Thus $C_t(s)ds$ is a probability measure.) (c) Show that for each $\delta > 0$, $\int_{\{|s|>\delta\}} C_t(s)\, ds \to 0$ as $t \to 0$. (Thus $C_t(s)ds$ is an approximate identity.) (d) Show that $\tilde{h}(s + it)$ is a bounded harmonic function in the upper half-plane. (e) Show that $\tilde{h}(s+it) \to h(\xi_0)$ as $s + it \in \mathbb{H}$ tends to ξ_0, whenever $h(\xi)$ is continuous at ξ_0. *Remark.* The kernel function $C_t(s)$ is the **Poisson kernel** for the upper half-plane.

8. Let $h\left(e^{i\varphi}\right)$ be a continuous complex-valued function on $\partial\mathbb{D}$, and let $\tilde{h}(z)$ be its Poisson integral.
   (a) Show that

$$\begin{aligned}\frac{\partial^m}{\partial z^m}\tilde{h}(z) &= m! \int_{-\pi}^{\pi} h(e^{i\varphi}) \frac{e^{i\varphi}}{(e^{i\varphi}-z)^{m+1}} \frac{d\varphi}{2\pi}, \qquad z \in \mathbb{D},\\ \frac{\partial^m}{\partial \bar{z}^m}\tilde{h}(z) &= m! \int_{-\pi}^{\pi} h(e^{i\varphi}) \frac{e^{-i\varphi}}{(e^{-i\varphi}-\bar{z})^{m+1}} \frac{d\varphi}{2\pi}, \qquad z \in \mathbb{D}.\end{aligned}$$

   Hint. Use the identity

$$P_r(\theta - \varphi) = 1 + \frac{ze^{-i\varphi}}{1 - ze^{-i\varphi}} + \frac{\bar{z}e^{i\varphi}}{1 - \bar{z}e^{i\varphi}}, \qquad z = re^{i\theta},$$

   and justify differentiating under the integral sign.
   (b) Show that if $\left|h\left(e^{i\theta}\right)\right| \le M$ on $\partial\mathbb{D}$, then

$$\left|\frac{\partial^m}{\partial z^m}\tilde{h}(z)\right| \le m!M \int_{-\pi}^{\pi} \frac{1}{|e^{i\varphi} - \rho|^{m+1}} \frac{d\varphi}{2\pi}, \qquad |z| \le \rho,$$

   with similar estimates for the $\bar{z}$-derivatives.
   (c) Show that if $h\left(e^{i\theta}\right) = e^{-i\theta}\left(e^{i\theta} - z\right)^{m+1}/|e^{i\theta} - z|^{m+1}$, then equality holds in (b) when $|z| = \rho$.

9. Let $\{h_n(e^{i\varphi})\}$ be a sequence of continuous functions on the unit circle $\partial\mathbb{D}$ that converges uniformly to $h(e^{i\varphi})$. Show that for each fixed $\rho < 1$, the partial derivatives of $\tilde{h}_n(z)$ converge to the corresponding partial derivatives of $\tilde{h}(z)$ uniformly for $|z| \le \rho$. *Remark.* It suffices to show this for the partial derivatives $\partial^m/\partial z^m$ and $\partial^m/\partial \bar{z}^m$. See Exercise IV.8.4.

2. Characterization of Harmonic Functions

Recall (Section III.4) that a continuous function $h(z)$ on a domain D has the mean value property if for each $z_0 \in D$, the value $h(z_0)$ is the average of $h(z)$ over any small circle centered at z_0. We have seen that harmonic functions have the mean value property. Here we prove the converse.

Theorem. *Let $h(z)$ be a continuous function on a domain D. Then $h(z)$ is harmonic on D if and only if $h(z)$ has the mean value property on D.*

The proof depends on solving the Dirichlet problem for disks in D. Roughly speaking, the **Dirichlet problem** for a domain U with boundary ∂U is to extend a given function $f(\zeta)$ on ∂U to a harmonic function $\tilde{f}(z)$ on U, so that the harmonic extension $\tilde{f}(z)$ has boundary values $f(\zeta)$. Typically we assume that $f(\zeta)$ is a continuous function on ∂U, and we ask that $\tilde{f}(z)$ attain the boundary values $f(\zeta)$ continuously, in the sense that $\tilde{f}(z)$ tends to $f(\zeta)$ as $z \in U$ tends to $\zeta \in \partial U$. In the preceding section we showed that the Poisson integral solves the Dirichlet problem for the open unit disk $\mathbb{D}$. By making a change of variable $z \mapsto az + b$, we can solve the Dirichlet problem on any disk. An explicit solution is obtained by changing variable in the Poisson integral formula. (See Exercise 1.2.)

So let U be any open disk in D whose boundary is also contained in D. Since the Dirichlet problem is solvable for U, there is a harmonic function $g(z)$ on U that has boundary values $h(z)$ on ∂U. Then $h(z) - g(z)$ is a continuous function on U that has the mean value property on U, and $h(z) - g(z)$ tends to 0 on ∂U. By the maximum principle for functions with the mean value property (Section III.4), $h(z) - g(z) = 0$ on U. Consequently, $h(z) = g(z)$ on U, and $h(z)$ is harmonic on U. Since this holds for all U, $h(z)$ is harmonic on D.

The characterization of harmonic functions by the mean value property is the analogue for harmonic functions of Morera's theorem for analytic functions. The hypothesis involves an integration, which is easier to perform than differentiation. The characterization can be exploited in the same way as the characterization of analytic functions by Morera's theorem. For instance, it can be used to give a simple proof that the limit of a uniformly convergent sequence of harmonic functions is harmonic, since it is easy to check that the uniform limit of functions with the mean value property has the mean value property. It can also be used (Exercise 1) to show that an integral of a function that depends harmonically on a parameter also depends harmonically on the parameter.

It is rather striking that a continuous function with the mean value property automatically has partial derivatives of all orders.

Exercises for X.2

1. Let $g(t, z)$ be a continuous function defined for $a \leq t \leq b$ and z in a domain D, and suppose that $g(t, z)$ is a harmonic function of z for each fixed t. Show that

$$G(z) \;=\; \int_a^b g(t, z)dt, \qquad z \in D,$$

is harmonic on D.

2. Assume that $u(x,y)$ is a twice continuously differentiable function on a domain D.
(a) For $(x_0, y_0) \in D$, let $A_\varepsilon(x_0, y_0)$ be the average of $u(x,y)$ on the circle centered at (x_0, y_0) of radius ε. Show that
$$\lim_{\varepsilon \to 0} \frac{A_\varepsilon(x_0, y_0) - u(x_0, y_0)}{\varepsilon^2} = \frac{1}{4}\Delta u(x_0, y_0),$$
where Δ is the Laplacian operator (Section II.5).
(b) Let $B_\varepsilon(x_0, y_0)$ be the area average of $u(x,y)$ on the disk centered at (x_0, y_0) of radius ε. Show that
$$\lim_{\varepsilon \to 0} \frac{B_\varepsilon(x_0, y_0) - u(x_0, y_0)}{\varepsilon^2} = \frac{1}{8}\Delta u(x_0, y_0).$$

3. For fixed $\rho > 0$, define $h(z) = e^{i\rho \operatorname{Im} z}$. Show that if ρ is a zero of the Bessel function $J_1(z)$, then $\int_\gamma h(z)dz = 0$ for all circles γ of radius 1. *Suggestion.* See the Schlömilch formula (Exercise VI.1.3).

4. Suppose that $r_1, r_2 > 0$ are such that r_2/r_1 is a quotient of two positive zeros of the Bessel function $J_1(z)$. Show that there is a continuous function $g(z)$ on the complex plane such that $\int_\gamma g(z)dz = 0$ for all circles of radius r_1 and for all circles of radius r_2, yet $g(z)$ is not analytic. Use the preceding exercise together with the fact (easily derived from the differential equation in Section V.4) that $J_1(z)$ has zeros on the positive real axis. *Remark.* The condition is sharp. If r_2/r_1 is not the quotient of two positive zeros of the Bessel function $J_1(z)$, then any continuous function $f(z)$ on the complex plane such that $\int_\gamma f(z)dz = 0$ for all circles of radius r_1 and r_2 is analytic, by a theorem of L. Zalcman. There is an analogous result for harmonic functions and the mean value property. According to a theorem of J. Delsarte, if r_1, $r_2 > 0$ are such that r_2/r_1 is not the quotient of two complex numbers for which $J_0(z) = 1$, then any continuous function on the complex plane that has the mean value property for circles of radius r_1 and r_2 is harmonic.

3. The Schwarz Reflection Principle

Suppose a given function is analytic on a domain that abuts on one side of an analytic curve. The Schwarz reflection principle asserts that under certain conditions, the analytic function extends analytically across the curve, and further, there is a formula for the extension. The formula reflects the function analytically to the mirror image of the domain, and the nub of the problem is to establish analyticity on the axis of reflection.

At its core, the Schwarz reflection principle is a theorem about harmonic functions. We begin by reflecting harmonic functions across the real line.

First note that if $u(z)$ is harmonic in a domain D, then $u^*(z) = u(\bar{z})$ is harmonic on the reflected domain $D^* = \{\bar{z} : z \in D\}$. Indeed, if $u(x, y)$ has the mean value property on circles, then so does $u(x, -y)$. Another way to see this is to note that the change of variables $(x, y) \mapsto (x, -y)$ does not change the Laplacian operator Δ.

Theorem. *Let D be a domain that is symmetric with respect to the real axis, and let $D^+ = D \cap \{\operatorname{Im} z > 0\}$ be the part of D in the open upper half-plane. Let $u(z)$ be a real-valued harmonic function on D^+ such that $u(z) \to 0$ as $z \in D^+$ tends to any point of $D \cap \mathbb{R}$. Then $u(z)$ extends to be harmonic on D, and the extension satisfies*

$$u(\bar{z}) = -u(z), \qquad z \in D. \tag{3.1}$$

For the proof, note that D is the disjoint union of $D^+ = D \cap \{\operatorname{Im} z > 0\}$, its reflection $D^- = D \cap \{\operatorname{Im} z < 0\}$, and $D \cap \mathbb{R}$. We extend $u(z)$ to D by defining $u(z) = -u(\bar{z})$ for $z \in D^-$, and setting $u(z) = 0$ for $z \in D \cap \mathbb{R}$. Then $u(z)$ is continuous on D, $u(z)$ is harmonic on D^+ and on D^-, and (3.1) holds. We claim that the extended function has the mean value property. Fix $z_0 \in D$. If z_0 is in D^+ or in D^-, then $u(z)$ has the mean value property for small disks centered at z_0, since $u(z)$ is harmonic near z_0. If $z_0 \in D \cap \mathbb{R}$, then by (3.1) the average value of $u(z)$ over the top half of a circle centered at z_0 cancels the average over the bottom half of the circle, and the average of $u(z)$ over the circle is 0, which coincides with $u(z_0)$. Thus $u(z)$ has the mean value property on D, and $u(z)$ is harmonic on D.

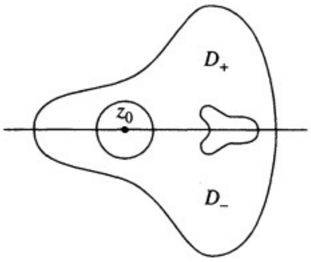

There is a corresponding result for analytic functions. First note that if $f(z)$ is analytic on a domain in the upper half-plane, then $g(z) = \overline{f(\bar{z})}$ is analytic on the reflected domain in the lower half-plane. This can be checked using the Cauchy-Riemann equations or a power series expansion. We can think of $g(z)$ as the composition of an anticonformal map $z \mapsto \bar{z}$, a conformal map $f(z)$ (except where $f'(z) = 0$), and an anticonformal map $w \mapsto \bar{w}$; hence it is conformal.

Theorem. *Let D be a domain that is symmetric with respect to the real axis, and let $D^+ = D \cap \{\operatorname{Im} z > 0\}$. Let $f(z)$ be an analytic function on D^+ such that* $\operatorname{Im} f(z) \to 0$ *as $z \in D^+$ tends to $D \cap \mathbb{R}$. Then $f(z)$ extends to be analytic on D, and the extension satisfies*

$$f(\bar{z}) \;=\; \overline{f(z)}, \qquad z \in D. \tag{3.2}$$

For the proof, write $f(z) = u(z) + iv(z)$ in D^+. By the preceding theorem, $v(z)$ extends to be harmonic on D and satisfies $v(\bar{z}) = -v(z)$, $z \in D$. Fix a point $x_0 \in D \cap \mathbb{R}$, and let D_0 be a disk in D centered at x_0. Since $v(z)$ is harmonic on D_0, it has a harmonic conjugate on D_0, and consequently $f(z)$ extends to be analytic on D_0. The function $\overline{f(\bar{z})}$ is also analytic on D_0, and it coincides with $f(z)$ on the real axis, so by the uniqueness principle it coincides with $f(z)$ in D_0. Hence (3.2) holds for $z \in D_0$. If we now extend $f(z)$ to $D^- = D \cap \{\operatorname{Im} z < 0\}$ by setting $f(z) = \overline{f(\bar{z})}$, $z \in D^-$, the extended $f(z)$ is analytic on D^- and coincides with the analytic continuation of $f(z)$ across $D \cap \mathbb{R}$ from D^+, and so is analytic on all of D. Further, it satisfies (3.2) for all $z \in D$.

The above theorem is somewhat easier to prove if it is assumed that $f(z)$ extends continuously from D^+ to $D \cap \mathbb{R}$ and is real-valued there. In this case, formula (3.2) defines a continuous extension of $f(z)$ to all of D, which is analytic on D^-, and we can use Morera's theorem to show that it is also analytic across $D \cap \mathbb{R}$.

We define a curve γ to be an **analytic curve** if every point of γ has an open neighborhood U for which there is a conformal map $\zeta \mapsto z(\zeta)$ of a disk D centered on the real line $\mathbb{R}$ onto U, such that the image of $D \cap \mathbb{R}$ coincides with $U \cap \gamma$. We also refer to such a γ as an **analytic arc**. The conjugation $\zeta \mapsto \bar{\zeta}$ induces a map $z \mapsto z^*$ of U onto itself, by $z(\zeta)^* = z(\bar{\zeta})$. The map is an "involution" in the sense that the map followed by itself is the identity, $z^{**} = z$. Further, $z^* = z$ if and only if $z \in \gamma$. The map $\zeta \mapsto \bar{\zeta}$ interchanges the top half and the bottom half of $D \backslash \mathbb{R}$, which are the two components of $D \backslash \mathbb{R}$, so the map $z \mapsto z^*$ interchanges the two components of $U \backslash \gamma$. We refer to these two components as the **neighborhoods of the sides of** γ, and we refer to the map $z \mapsto z^*$ as the **reflection across** γ. Since the map $\zeta \mapsto \bar{\zeta}$ is anticonformal, and the composition of an anticonformal map and conformal maps is anticonformal, the reflection $z \mapsto z^*$ is anticonformal.

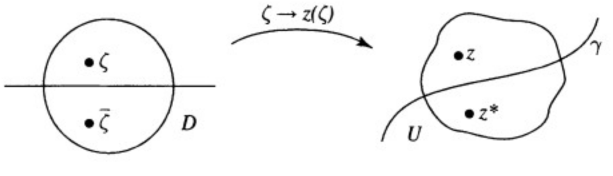

The reflection $z \mapsto z^*$ is unique, in the following sense. If there is an open neighborhood of a point $z_0 \in \gamma \cap U$ and an anticonformal map $z \mapsto z^+$

that fixes each point of γ in the neighborhood, then $z^+ = z^*$. Indeed, the map $z \mapsto (z^+)^*$ is the composition of two anticonformal maps, so it is conformal, hence analytic (Section V.8). Since it fixes points of γ, the uniqueness principle shows that it is the identity map. Thus $(z^+)^* = z$, and $z^+ = (z^+)^{**} = (z^{+*})^* = z^*$.

We can extend the involution $z \mapsto z^*$ to a neighborhood of the entire analytic arc γ, by covering γ with small coordinate sets U and using the uniqueness principle to see that the involutions agree on overlapping U's.

Example. The map $z(\zeta) = (\zeta - i)/(\zeta + i)$ sends the extended real line onto the unit circle. It induces a reflection across the unit circle,

$$z^* \;=\; (\bar\zeta - i)/(\bar\zeta + i) \;=\; \overline{(\zeta + i)/(\zeta - i)} \;=\; 1/\bar z \;=\; z/|z|^2.$$

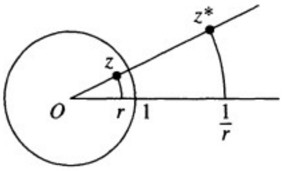

From this formula we see that the reflected point z^* lies on the same ray through the origin as z. The reflection $z \mapsto z^*$ maps the circle centered at 0 of radius r to that of radius $1/r$, sending $re^{i\theta}$ to $e^{i\theta}/r$.

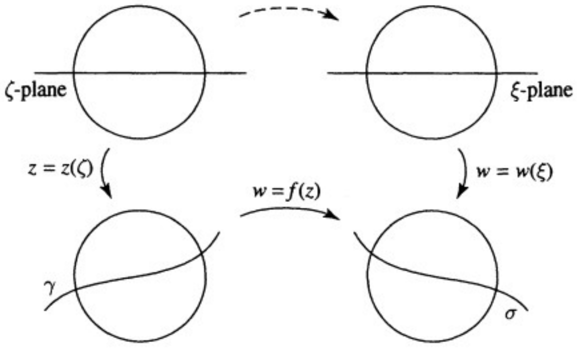

A more general version of the Schwarz reflection principle permits us to reflect harmonic and analytic functions across analytic curves. The idea is that if $w = f(z)$ is defined and analytic on one side of an analytic arc γ, and if the values w tend to another analytic arc σ as z tends to γ, then $f(z)$ extends analytically across γ, and the extension satisfies $f(z^*) = f(z)^*$ in a neighborhood of γ. (Here we use the same notation for both reflections.) It is a straightforward matter to justify the general version by coordinatizing both analytic arcs and reducing to the case of reflection across the real line.

Rather than state a precise theorem, we derive a version of the principle that will be used later.

Let D be a domain. An analytic arc $\gamma \subset \partial D$ is a **free analytic boundary arc** of D if every point of γ is contained in a disk U such that $U \backslash \gamma$ has two components, one contained in D and the other disjoint from D.

Example. Suppose D is obtained from the open disk $\{|z| < 2\}$ by excising a sequence of disjoint closed disks from the top half accumulating on the interval $[-1, 1]$, as in the figure. Then the boundary circle of each excised disk is a free analytic boundary arc of D, but the analytic arc $(-1, 1)$ is not a free analytic boundary arc.

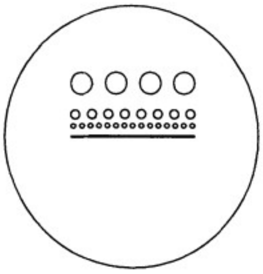

Theorem. *Let D be a domain, and let γ be a free analytic boundary arc of D. Let $f(z)$ be analytic on D. If $|f(z)| \to 1$ as $z \in D$ tends to γ, then $f(z)$ extends to be analytic in a neighborhood of γ, and the extension satisfies*

$$f(z^*) = 1/\overline{f(z)} \tag{3.3}$$

in a neighborhood of γ, where $z \mapsto z^*$ is the reflection across γ.

For the proof, fix a point $z_0 \in \gamma$, and consider a conformal map $\zeta \mapsto z(\zeta)$ of a disk D_0 centered on $\mathbb{R}$ onto an open set containing z_0 so that $D_0 \cap \mathbb{R}$ corresponds to $U \cap \gamma$. Then $g(\zeta) = \log f(z(\zeta))$ is defined and analytic on one side of $D_0 \cap \mathbb{R}$. Further, $\operatorname{Re} g(\zeta) = \log |f(z(\zeta))| \to 0$ as $\zeta \to D_0 \cap \mathbb{R}$. By the Schwarz reflection principle, $g(\zeta)$ extends to be analytic on D_0, as does $f(z(\zeta)) = e^{g(\zeta)}$. Consequently, $f(z)$ extends analytically to U, and the extension is unique. To see that the extension satisfies (3.3), we argue as follows. The function $h(z) = 1/\overline{f(z^*)}$ is the composition of $f(z)$ and the two anticonformal maps $z \mapsto z^*$ and $w \mapsto 1/\bar{w}$. Thus it is conformal wherever $f'(z) \neq 0$. By the equivalence of analyticity and conformality (Section IV.8), $h(z)$ is analytic wherever $f'(z) \neq 0$, and it is continuous, hence analytic also at the isolated points where $f'(z) = 0$. Further, if $z \in \gamma$, then $z^* = z$, and $|f(z)| = 1$, so $h(z) = f(z)$ on γ. By the uniqueness principle, $h(z) = f(z)$ everywhere, and $f(z)$ satisfies (3.3).

Exercises for X.3

1. Show that the reflection in the circle $\{|z - z_0| = R\}$ is given by $z^* = z_0 + R^2(z - z_0)/|z - z_0|^2$.

2. Show that a reflection in a circle maps circles in the plane to circles.

3. What happens to angles between curves when they are reflected in an analytic arc?

4. Suppose the curve γ passing through 0 is the graph of a function $y = h(x)$ that can be expressed as a convergent power series $h(x) = \sum_{k=1}^{\infty} a_k x^k$, $-r < x < r$, where the a_k's are real. (a) Show that $z = \zeta + ih(\zeta)$ can be solved for $\zeta = \zeta(z)$ as an analytic function of z for $|z| < \varepsilon$. (b) Show that γ is an analytic arc. (c) Show that the reflection through γ is given by $z^* = 2\overline{\zeta(z)} - \bar{z}$.

5. Determine the reflection z^* of z across the parabola $y = x^2$. Expand z^* in a power series, in powers of $\bar{z}$. Determine the radius of convergence of the series. Try to explain graphically why the radius of convergence is finite.

6. Let $f(z)$ be an entire function whose modulus is constant on some circle. Show that $f(z) = c(z - z_0)^n$ for some $n \geq 0$ and some constant c, where z_0 is the center of the circle.

7. Show that if $f(z)$ is meromorphic for $|z| < 1$, and $|f(z)| \to 1$ as $|z| \to 1$, then $f(z)$ is a rational function. Show further that $f(z)$ is the quotient of two finite Blaschke products. (For the definition of finite Blaschke product, see Exercise IX.2.1.)

8. The **modulus of an annulus** $\{a < |z - z_0| < b\}$ is defined to be $(1/2\pi)\log(b/a)$. (a) Show that any conformal map from one annulus centered on the origin to another such annulus extends to a conformal self-map of the punctured plane. (b) Show that there is a conformal map of one annulus onto another if and only if the annuli have the same moduli. (c) Show that any conformal self-map of the annulus $\{a < |z| < b\}$ is either a rotation $z \mapsto e^{i\varphi}z$ or a rotation followed by the inversion $z \mapsto ab/z$.

9. Let γ be an analytic curve passing through the origin and tangent to the imaginary axis at 0. Suppose that $\gamma(t) = it + b_2t^2 + b_3t^3 + \cdots$, $-\delta < t < \delta$, where the b_n's are real. Show that there is a conformal map $\varphi(z)$ defined in a neighborhood of 0 such that $\varphi(0) = 0$, $\varphi'(0) = 1$, and $\varphi(z)$ maps the angle between the positive real axis and the segment of γ in the upper half-plane to the angle between the positive real and imaginary axes if and only if γ coincides with its reflection $\bar{\gamma}$ in the real axis. Show that this occurs if and only if $b_n = 0$ for n odd.

10. Let γ be an analytic curve passing through the origin and making an angle θ_0, $0 < \theta_0 < \pi/2$, with the positive real axis. Suppose γ is the image of an interval $(-\delta, \delta)$ under $h(z) = z + i(b_1 z + b_2 z^2 + \cdots)$, where the b_j's are real. Define

$$g(z) = \overline{h(\overline{h^{-1}(z)})}, \qquad \lambda = e^{-2i\theta_0}.$$

(a) Show that $g(z)$ is analytic at $z = 0$, $g'(0) = \lambda$, and $g(z) = \bar{z}$ for $z \in \gamma$.

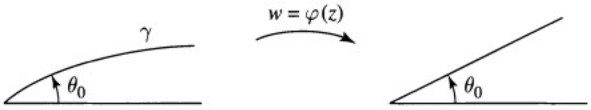

(b) Show that if $\varphi(z)$ is analytic at $z = 0$, $\varphi(0) = 0$, $\varphi'(0) = 1$, and $\varphi(z)$ maps the angle between the positive real axis and the segment of γ in the first quadrant to the angle between the positive real axis and the straight line segment at angle θ_0, then $\varphi(g(z)) = \lambda\varphi(z)$. *Remark.* This equation for $\varphi(z)$ is **Schröder's equation**.

(c) Show by plugging in power series that any (normalized) solution $\varphi(z)$ of Schröder's equation is unique, if it exists, and is given by $\varphi(z) = z + c_2 z^2 + c_3 z^3 + \cdots$, where

$$c_n = \frac{1}{\lambda^n - \lambda} A_n(\lambda, a_2, \ldots, a_n, c_2, \ldots, c_{n-1}), \qquad n \geq 2,$$

A_n is a polynomial, and the a_n's are the power series coefficients of $g(z)$. *Remark.* The problem of estimating the c_n's to determine whether the series converges is called a "small denominator problem ." As the powers λ^{n-1} return sporadically near to 1, the denominators become sporadically small.

(d) Show that if $\varphi(z)$ is a solution of Schröder's equation (analytic at 0 and satisfying $\varphi(0) = 0$, $\varphi'(0) = 1$), and if $\varphi(x)$ is real when x is real, then $\varphi(z)$ maps the angle between the positive real axis and γ to the angle between the straight line segments at angles 0 and θ_0.

E !"#$! ! !% & ' ! EE / - 0-

E !"#$! ! !% & ' ! EE F - 4 (> / ', 0-
%9 : 4 - ;;;

E !"#$! ! !% & ' ! EE F ((> /
G ' 0-

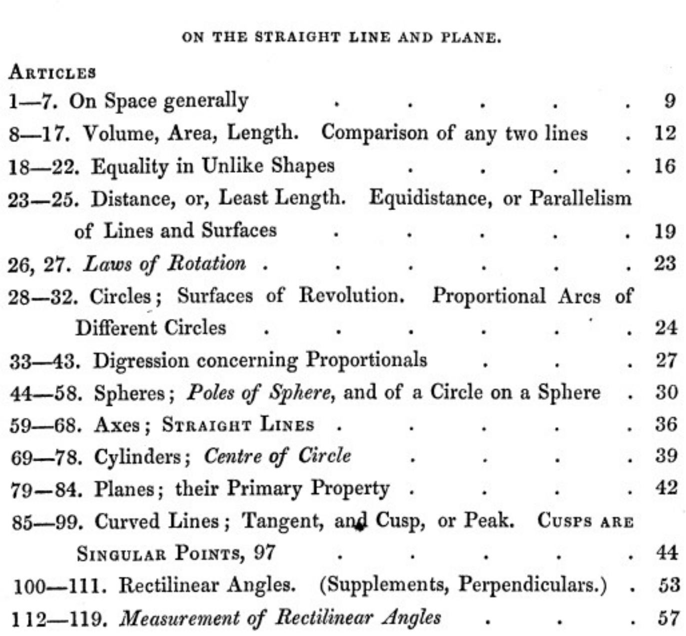

ON THE STRAIGHT LINE AND PLANE.

ARTICLES
1—7. On Space generally 9
8—17. Volume, Area, Length. Comparison of any two lines . 12
18—22. Equality in Unlike Shapes 16
23—25. Distance, or, Least Length. Equidistance, or Parallelism of Lines and Surfaces 19
26, 27. *Laws of Rotation* 23
28—32. Circles; Surfaces of Revolution. Proportional Arcs of Different Circles 24
33—43. Digression concerning Proportionals . . . 27
44—58. Spheres; *Poles of Sphere*, and of a Circle on a Sphere . 30
59—68. Axes; STRAIGHT LINES 36
69—78. Cylinders; *Centre of Circle* 39
79—84. Planes; their Primary Property 42
85—99. Curved Lines; Tangent, and Cusp, or Peak. CUSPS ARE SINGULAR POINTS, 97 44
100—111. Rectilinear Angles. (Supplements, Perpendiculars.) . 53
112—119. *Measurement of Rectilinear Angles* . . . 57

E= !"#$! ! !% & ' ! EE F 4 - &(: (&(> 4 (#'
) * 0-
%9 : 4 - ;;;

E@ !"#$! ! !% & ' ! EE D ' /(' (> /
0-
I& 8 . ' ' &

E< !"#$! ! !% & ' ! EE D ' /(/ (> /
7' - 0-

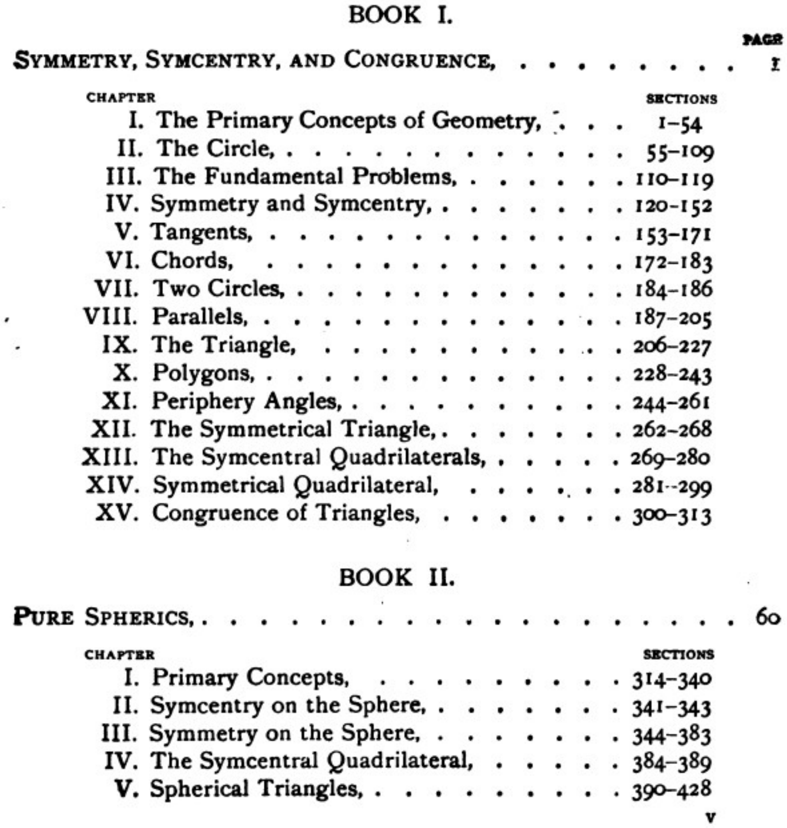

BOOK I.

PAGE
SYMMETRY, SYMCENTRY, AND CONGRUENCE, ... 1

CHAPTER SECTIONS
I. The Primary Concepts of Geometry, ... 1–54
II. The Circle, ... 55–109
III. The Fundamental Problems, ... 110–119
IV. Symmetry and Symcentry, ... 120–152
V. Tangents, ... 153–171
VI. Chords, ... 172–183
VII. Two Circles, ... 184–186
VIII. Parallels, ... 187–205
IX. The Triangle, ... 206–227
X. Polygons, ... 228–243
XI. Periphery Angles, ... 244–261
XII. The Symmetrical Triangle, ... 262–268
XIII. The Symcentral Quadrilaterals, ... 269–280
XIV. Symmetrical Quadrilateral, ... 281–299
XV. Congruence of Triangles, ... 300–313

BOOK II.

PURE SPHERICS, ... 60

CHAPTER SECTIONS
I. Primary Concepts, ... 314–340
II. Symcentry on the Sphere, ... 341–343
III. Symmetry on the Sphere, ... 344–383
IV. The Symcentral Quadrilateral, ... 384–389
V. Spherical Triangles, ... 390–428

EE !"#$! ! !% & ' ! EE D ((> / 7' - 0-

E !"#$! ! !% & ' ! EE > &) 7' - .
& / 8 = 0-
V E 0'&
V : 4 - ;;;

!"#$! ! !% & ' ! EE > &) 7' - .
& / ' / 9 & % 5 ? 4 / &' 8 = -24

!"#$! ! !% & ' ! EE D ((> / : 0-

!"#$! ! !% & ' ! EE +' (> /(9
1 0-

!"#$! ! !% & ' ! EE > ' (.'/ 0-

!"#$! ! !% & ' ! EE > /(0 ,(F '
-& 0-
I&

J . 0 , 9 ' / - ' 5 ' ' -
' '

' 5 ' 5 ' - ' ' @
' ' ' >' 0 '
' ' ' 5 4' ' ' - ' 3'
' ' ' ' 5 7 '
6 4' ' ' 5 ' 4 ' ' <

6 4' ' ' ' ' - ' 4 <

I& 8

. - 9 ' ' ' - ' ' '
' 0 9 '
0 0 , 9 ' / - ' >' 0 '

= !"#$! ! !% & ' ! EE > /(B (",
. 0 0-
%9 % ' ;;;

@ !"#$! ! !% & ' ! EE > / & 4 0-
%9 : 4 - ;;;

'0 W6 ' 4' I& @ 8 @

%9 I& < 8 < < ' ' - >' ' 4' ' '

< !"#$! ! !% & ' ! EE 6 (- (> / 0-
%9 : 4 - ;;;

I& @<8 E '

' - 0

I& @ 8

317. The notion of *sine of an angle*, introduced for simplicity in the foregoing article, is of the highest importance for all following geometrical study. But perhaps a more fundamental notion is that of cosine, which we may define thus:

Def. The *ratio of the projection of a tract on a ray to the tract itself* is called the **cosine** of the angle between the ray and the tract (Fig. 187).

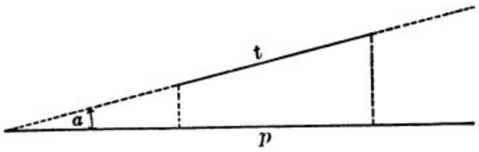

FIG. 187.

317. The notion of *sine of an angle*, introduced for simplicity in the foregoing article, is of the highest importance for all following geometrical study. But perhaps a more fundamental notion is that of cosine, which we may define thus:

Def. The *ratio of the projection of a tract on a ray to the tract itself* is called the **cosine** of the angle between the ray and the tract (Fig. 187).

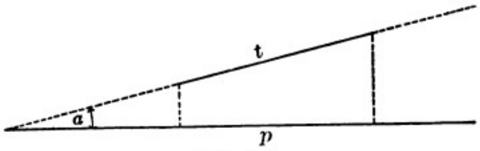

FIG. 187.

Thus the ratio of the projection p of the tract t to the tract itself is the cosine of the angle α between them; or $p : t = \underline{\alpha}$, as we may write **cosine of** α, which is commonly abbreviated into **cos** α.

318. It is plain that the projections of t on parallel rays are all equal; hence we may suppose the ray of projection drawn through the beginning of t, as in Fig. 188. Then as t turns about o and α changes its value, the projection p of t will also change. Thus: for

$$\alpha = 0,\ 60°,\ 90°,\ 120°,\ 180°,\ 270°,\ 360°,\ 420°,\ \cdots$$

$$\underline{\alpha} = 1,\ \tfrac{1}{2},\ 0,\ -\tfrac{1}{2},\ -1,\ 0,\ 1,\ \tfrac{1}{2},\ \cdots$$

319. In the 2d and 3d quadrants the projection p is reckoned leftward; it is opposite in sense to the projection when α is in the 1st or 4th quadrants, and accordingly the cosine is marked $-$. When α increases by 360° (or 2π, see Art. 336) from any value, the revolving tract resumes its original position, the projection resumes its original value, and so too does the cosine; hence

$$\cos(360^\circ + \alpha) = \cos(2\pi + \alpha) = \cos\alpha;$$

that is, *the cosine is not changed by increasing* (or decreasing) *the angle by a round angle.*

Hence, plainly, $\cos(\alpha \pm 2n\pi) = \cos\alpha$.

Hence the cosine is called a **periodic function*** of the angle, the **period** being 2π, that is, a *round angle* (see Art. 336).

320. If t be turned through any angle α from the position OA, its projection p is the same whether the turning be clockwise or counter-clockwise; that is, the projection is the same whether the angle be negative or positive; hence, too, the cosine is the same.

That is, $\cos(-\alpha) = \cos\alpha$;

that is, *the cosine of the angle is unchanged by changing the sense* (or sign) *of the angle.* Now we learn in Algebra that only the even powers, not the odd powers, of a symbol are unchanged by changing the sign of the symbol; thus:

$$(-\alpha)^2 = \alpha^2,\ (-\alpha)^4 = \alpha^4, \text{ but } (-\alpha)^3 = -\alpha^3.$$

Hence the cosine is called an **even** function of the angle.

Its *periodicity* and its *evenness* are the two cardinal properties of the cosine, on which all others hinge.

321. We may now arbitrarily define the **sine** of an angle to be the **cosine of the complemental angle**; *i.e.* $\underline{90^\circ - \alpha} = \alpha|$, as we may write **sine of** α, which is commonly abbreviated into **sin** α.

---

* Two magnitudes such that the values of the one *correspond* to values of the other are called *functions* of each other.

Now write $90^\circ - \beta$ for α;

we obtain

$$\cos\{90^\circ - (90^\circ - \beta)\} = (90^\circ - \beta)|, \text{ or } \underline{\beta} = (90^\circ - \beta)|;$$

i.e. *the cosine of an angle is the sine of the complemental angle.*

Hence the sine (or cosine) of either of two complemental angles, as α and β, is the cosine (or sine) of the other; *i.e.* if $\alpha + \beta = 90^\circ$, then $\underline{\alpha} = \beta|$, and $\alpha| = \underline{\beta}$. When either changes by 2π, so does the other oppositely; hence the sine as well as the cosine returns into its original value; *i.e. the sine is also periodic with the period* 2π.

322. If the tract t be reversed, that is, turned through a straight angle, its projection will also be reversed, but otherwise unchanged; hence the cosine will be reversed; that is,

$\cos(\alpha + \pi) = -\cos\alpha$; also $\cos(\alpha - \pi) = \cos(\alpha + \pi)$ (why?);

hence $$\cos(\alpha - \pi) = -\cos\alpha,$$

or, *to change the angle by the half-period, π, changes the sign of the cosine.*

323. Since the cosine is an even function,

$$\cos(\alpha - \pi) = \cos(\pi - \alpha);$$

hence $$\cos(\pi - \alpha) = -\cos\alpha.$$

That is, since α and $\pi - \alpha$ are supplemental, *the cosines of supplemental angles are counter* — equal in size, but opposite in sense (or sign).

324. Again, if $\alpha + \beta = 90^\circ$, then $\alpha| = \beta$ (why?).

Then $$(\alpha + \pi)| = \underline{\beta - \pi} \text{ (why?)} = -\underline{\beta} = -\alpha|.$$

That is, *to change the angle by the half-period, π, changes the sense of the sine* as well as of the cosine.

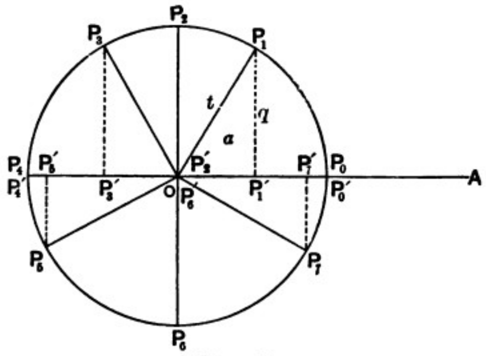
P_2
P_3
P_1
t
q
a
P_4
P_5'
P_2'
P_7'
P_0
A
P_4'
P_3'
O
P_6'
P_1'
P_0'
P_5
P_7
P_6

FIG. 188.

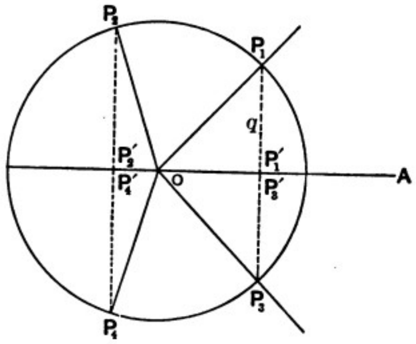
P_2
P_1
q
P_2'
P_1'
A
P_4'
O
P_3'
P_3
P_4

FIG. 189.

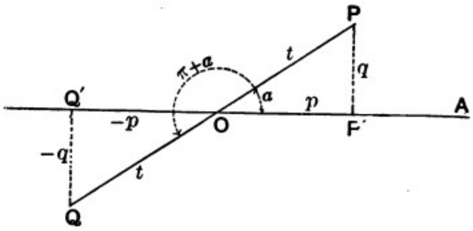
P
t
$\pi+a$
q
Q′
a
p
A
$-p$
O
F′
$-q$
t
Q

FIG. 190.

We now ask, how is the sine affected by changing the sense of the angle? We have

$(-\alpha)| = \underline{90^\circ + \alpha}$ (why?) $= -(\underline{90^\circ - \alpha})$ (why?) $= -(\alpha)|$;

that is, *the sense of the sine changes when the sense of the angle changes*. But this is the property only of *odd* powers, not of even powers, of a symbol;

thus $\quad (-a)^3 = -a^3,\ (-a)^5 = -a^5,$ etc.

Hence the sine is called **odd** function of the angle.

Its *periodicity* and its *oddness* are the cardinal properties of the sine, on which all others hinge.

325. If q be the projector of the tract t then $\frac{q}{t}$ is plainly the sine of the angle α for every position of t. We may indeed *define* the sine of α as equal to this ratio, and from this definition readily deduce all the foregoing properties (Figs. 188, 189, 190).

Exercises. 1. Prove that $(\alpha|)^2 + (\underline{\alpha})^2 = 1$.

2. Find the value of $\alpha|$ for $\alpha = 0, 30^\circ, 45^\circ, 60^\circ, 90^\circ, 120^\circ, 150^\circ, 180^\circ, 210^\circ, 225^\circ, 240^\circ, 270^\circ, 300^\circ, 330^\circ, 360^\circ, 390^\circ$.

3. If a, b, c, be the sides of a $\triangle$, α, β, γ the opposite angles, r the circumradius, prove that

$$\frac{a}{\alpha|} = \frac{b}{\beta|} = \frac{c}{\gamma|} = 2\,r.$$

Such is the **Law of Sines** (Fig. 191).

4. Prove that $a^2 = b^2 + c^2 - 2\,bc\underline{\alpha}$, — **Law of Cosines.**

5. If a and b are adjacent sides of a $\square$, and $\widehat{ab}$ denote the angle between them, prove that $\square = ab \cdot \widehat{ab}|$.

N.B. We may *define* the sine from this important theorem thus: *The sine of the angle between two sides of a $\square$ is that* *number which taken as a multiplier turns the area of the rectangle of the sides into the area of the* ▱.

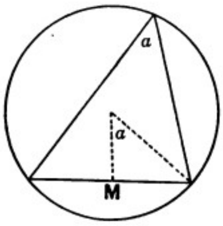

FIG. 191.

326. If we project successively the sides of a closed polygon on any ray, *the sum of the projections will be* 0; for the end of the projection of the last side falls on the beginning of the projection of the first. This fact is very important in surveying, where in compassing a field the sum of the northings must equal the sum of the southings, and these two sums, having opposite senses, together make the whole sum 0.

So for the eastings and westings.

We may express this fact in symbols thus:

$$s_1\underline{\alpha}_1 + s_2\underline{\alpha}_2 + s_3\underline{\alpha}_3 + \cdots + s_n\underline{\alpha}_n = 0 = \Sigma s\underline{\alpha}.$$

Here the s's are the sides, the α's are the angles of the sides with a fixed ray, as the east and west line, $s\underline{\alpha}$ is the projection of a side, and Σ is the symbol of summation.

Exercise. Show by projecting on a ray normal to the first ray that $\Sigma s\alpha| = 0$.

327. If we project consecutively all the sides of a polygon but one on that one, *the sum of the projections will be that one itself.* For the beginning of that side is the projection of the beginning of the first, and its end is the projection of the end of the last, of the projected sides (Fig. 192).

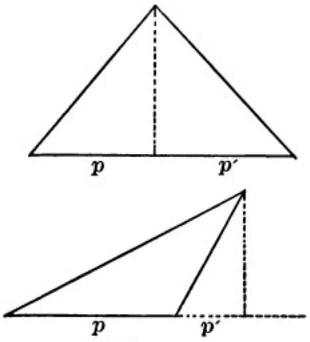

FIG. 192.

Thus, in a △, the sum of the projections of two sides on the third is the third.

328. We make a most important application of this simple fact in finding the *cosine of the difference of two angles* (Fig. 193).

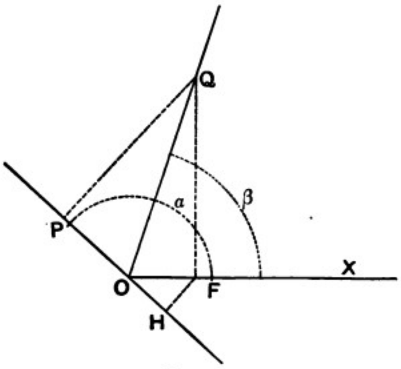

FIG. 193.

Let OP and OQ make angles α and β with any fixed ray OX. Then angle $QOP = \alpha - \beta$. Now project any tract OQ on OP; there $OP = OQ \cdot \underline{\alpha - \beta}$. But instead of projecting OQ directly we shall obtain the same result by projecting consecutively OF and FQ. The projection of OF is $OF \cdot \underline{\alpha}$, and of FQ is $FQ \cdot \alpha|$. But $OF = OQ \cdot \underline{\beta}$, and $FQ = OQ \cdot \beta|$.

$$\text{Hence} \quad OP = OQ \cdot \underline{\alpha - \beta} = OQ \cdot \underline{\alpha} \cdot \underline{\beta} + OQ \cdot \alpha| \cdot \beta| .$$

$$\text{Or,} \qquad \alpha - \beta = \alpha \cdot \beta + \alpha| \cdot \beta| .$$

By changing the sense of β and by putting $90 - \alpha$ instead of α, let the student show that $\underline{\alpha + \beta} = \underline{\alpha} \cdot \underline{\beta} - \alpha| \cdot \beta|$, $(\alpha + \beta)| = \alpha| \cdot \underline{\beta} + \underline{\alpha} \cdot \beta|$, $(\alpha - \beta)| = \alpha| \cdot \underline{\beta} - \underline{\alpha} \cdot \beta|$.

These four formulæ express the **Addition-Theorem** of Sine and Cosine.

The doctrine of Functions of Angles constitutes Trigonometry — an extremely important subject, which cannot be pursued any further here. See Smith's *Clew to Trigonometry.*

MEASUREMENT OF THE CIRCLE.

329. Thus far our linear measurements, or comparisons of length, have been wholly of tracts. The peculiar simplicity of such operations is due to the fact that any tract may be superposed (at least in thought) on any other, and thus their equality or inequality infallibly tested. We may similarly compare arcs of equal circles, but not arcs of unequal circles, nor arcs and tracts; for these cannot be made to fit on each other to even the smallest extent. We feel sure indeed that a circle or arc has a perfectly definite length, that it is longer than some tracts, shorter than others, and equal to some others. For if we suppose an inextensi-

ble cord wrapped around a circular disk, on unwinding and straightening the cord we should obtain a tract equal to the circle in length. But it remains difficult or impossible to fix the notion of the length or to determine the length itself without some preliminary definitions and assumptions.

We assume then that a circle has a definite length, neither more nor less; also, that it bounds a definite area, neither more nor less.

330. We now inscribe in the circle of radius r a regular n-side, and parallel to this latter we circumscribe a regular n-side; then bisecting each arc subtended by a side of the inscribed n-side we inscribe a regular $2n$-side and also circumscribe parallel to it a regular $2n$-side.

Then the following facts are at once evident:

1. The area of any inscribed polygon is less, and the area of every circumscribed polygon is greater, than the area of the circle; or if I_n, S, and C_n designate these areas, then

$$I_n < S < C_n \text{ (why?)}.$$

2. The area of the inscribed $2n$-side is greater than that of the inscribed n-side, while the area of the circumscribed $2n$-side is less than that of the circumscribed n-side; or,

$$I_{2n} > I_n,\ C_{2n} < C_n \text{ (why?)}.$$

3. The area of each regular n-side is half that of the rectangle of the perimeter and the central normal on a side, and in case of the circumscribed polygons this normal is the radius r, but in case of the inscribed polygons this normal, or **apothem**, a_n, is less than r.

4. The perimeters of inscribed and circumscribed n-sides are to each other as a_n and r; for they are similar polygons, and a_n corresponds to r.

5. Since the area of the circumscribed n-side decreases as n increases, and since one dimension, the radius, remains constant, it follows that the other dimension, the (half) perimeter, must decrease with increasing n. For $n = 4$ the polygon is a circumsquare, and the perimeter is $8r$.

6. Since the sum of two sides of a △ is greater than the third side, it follows that the perimeter of the inscribed n-side increases with increasing n. For $n = 6$ the perimeter is $6r$. Hence for all higher values of n the perimeters of both inscribed and circumscribed polygons lie between $6r$ and $8r$.

7. Since then the sum of the n-sides is certainly less than $8r$, by making n large enough we can make each side, whether of inscribed or circumscribed polygon, as small as we please, smaller than one millionth, smaller than one billionth, smaller than any assigned magnitude however small.

8. But the half-side of the regular inscribed n-side is a geometric mean between the segments of the normal diameter; *i.e.*

$$r + a_n : \frac{S_n}{2} :: \frac{S_n}{2} : r - a_n.$$

As S_n is small at will, so is $\frac{S_n}{2}$, and still more is $r - a_n$, which we may call d_n, the distance between the parallel sides of the inscribed and circumscribed polygons.

N.B. A magnitude small at will is often called an **infinitesimal.** Since $r - a_n$ or d_n is infinitesimal with respect to $\frac{S_n}{2}$, which is itself infinitesimal, it (d_n) is called an infinitesimal of *2d order*. But we are not now concerned with this fact.

9. The difference in area between the circumscribed and the inscribed n-sides is a trapezoid the half-sum of whose parallel sides is less than $8r$, the altitude being d_n. Since $8r$ is finite and definite while d_n is small at will, it follows that the difference in area of the circumscribed and inscribed regular polygons is small at will.

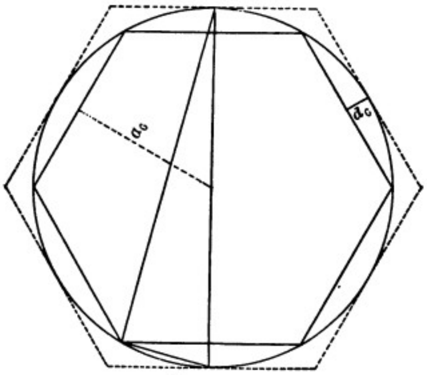

FIG. 194.

10. Again, we have from similar homothetic figures,

$$\text{Perimeter of } C_n : \text{Perimeter of } I_n :: r : r - d_n.$$

Hence, *dividendo*,

$$\text{Perimeter } C_n : \text{Perimeter } C_n - \text{Perimeter } I_n :: r : d_n.$$

But d_n is small at will; so too then is $D_n =$ Perimeter C_n − Perimeter of I_n; *i.e.* the difference in perimeter of the circumscribed and inscribed regular polygons is *small at will.*

11. The circle lying always wholly between the two polygons inscribed and circumscribed both in position and in areal value, it follows that the difference between its area

and the area of either polygon is small at will; and since the circle is fixed both in area and in position, it follows that the circle is the **Limit** of both polygons, and that the polygons close down upon it close at will as *n* grows ever larger and larger.

331. Think of the polygonal strip between the inscribed and circumscribed *n*-sides as growing ever narrower and narrower; the circle is a fixed *boundary* toward which the circumscribed *n*-side shrinks down as *n* increases, and in such a way that there is no assignable point outside of the circle, no matter how close to it, that will not also fall outside of the circumscribed *n*-side as *n* increases. Likewise the circle is the fixed *boundary* toward which the inscribed *n*-side swells out as the *n* increases, and in such fashion that there is no assignable point inside of the circle (no matter how close to it) that will not fall inside of the inscribed *n*-side as *n* increases. Thus the inscribed and circumscribed *n*-sides close down upon each other so as to leave no point between *except the points of the circle itself.* As *n* increases, the polygons *tend* to absolute coincidence with each other and with the circle. This fact is expressed fully and accurately by saying that the circle is the common **Limit** in length, in area, in position, of the inscribed and circumscribed regular *n*-sides for increasing *n*; it is expressed elliptically and inaccurately, but conveniently and frequently, by saying that the circle is or may be regarded as a *regular polygon of an infinite number of sides.*

332. The area of a circumscribed *n*-side is half the area of the rectangle of the radius of the circle and the perimeter of the polygon, for all values of *n*. Hence the area of the circle is half the rectangle of radius and the perimeter; that

is, the circle. The numeric expressing the ratio of the circle to its diameter is called the *perimetric ratio*, and is designated by the Greek initial of perimeter, π. It is an irrational and hence not expressible exactly as a fraction, whether common or decimal, but its value has been calculated in various ages to various degrees of exactitude, and may be calculated to any degree of exactitude. Of late years it has been calculated (by Shanks) to the 707th decimal place, and verified to the 500th — a degree of accuracy immensely higher than can be attained in any measurement. For most practical purposes the value $\pi = 3.14159$ or even 3.1416 is close enough.

If then r be the radius, $2\pi r$ is the length of the circle, and $\pi r \cdot r$, or πr^2, is its area.

333. Because the ratio π is irrational it by no means follows that it cannot be constructed geometrically; that is, that we cannot with ruler and compasses draw a tract that shall be exactly equal to the circle of radius r. The ratio $\sqrt{2}$ is irrational, yet we can easily construct $\sqrt{2} \cdot r$, by drawing the diagonal of a square of side r. If we could draw a tract πr, equal to a half-circle, then, by Problem III, p. 193, we could construct the geometric mean of r and πr, which would be the side of a square precisely equal to the circle in area. This famous problem of *squaring the circle* is therefore not an irrational one; it is unsolved, but possibly not in itself unsolvable. But see Math. Ann., xx., p. 213.

MEASUREMENT OF ANGLES.

334. We have denoted by π the so-called *perimetric* ratio, namely, of the circle to its diameter, $2r$, so that, if r be the radius, then $2\pi r$ is the (length of the) circle, and πr^2 is its area. But there is another important use of π.

335. We have learned the ordinary sexagesimal division of the round angle into 360 equal parts called *degrees*. This *artificial* unit, degree, does not recommend itself for purposes of mathematical investigation, but a *natural* unit is suggested by the measurement of the circle itself. For it is plain that *whatever part an arc is of the whole circle, that same part the central angle of the arc is of the round angle*. Thus, if m times the arc make out the circle, then m times the central angle will make out the round angle; for, in adding the arcs about the centre O, we at the same time add the angles at O subtended by the arcs. If, however, arc and circle be incommensurable, cut the arc and also the angle into q very small equal parts. Then p of these arc-parts will be less and $p + 1$ will be greater than the circle, while, similarly, p of the angle-parts will be less and $p + 1$ will be greater than the round angle; and this will always hold, no matter how great q and p may be. Hence, using the arc as unit-arc and the angle as unit-angle, we see that the metric numbers of circle and round angle lie always between the same fractions, $\frac{p}{q}$ and $\frac{p+1}{q}$; and for increasing p and q these fractions *close down upon each other*, so that the metric numbers of circle and round angle cannot differ by ever so little, but must be precisely the same. If instead of circle and round angle we take any other arc and its corresponding central angle, the reasoning remains unchanged, so that we have

Theorem CXLVI. — *If any arc and its corresponding central angle be taken as units, the metric numbers of any other arc* (of the same or equal circle) *and its corresponding central angle are the same.*

In other words,

Central angles and their subtending arcs (in the same or equal circle) *are proportional* (see Art. 337).

The latter statement is conciser; the former is preciser.

336. Now a natural unit for arc-measurement is plainly *radius;* hence a natural unit for angle-measurement is *the angle whose arc is radius* (in length); accordingly we adopt it as unit-angle and name it **Radian.** The metric number of the circle, radius being unit, is 2π; hence the metric number of the round angle, radian being unit, is 2π. The radian equals about $57° 17' 33''$.

Corollary. If n be the metric number of any angle, and therefore of its corresponding arc, radian and radius being units, then of any other arc, subtending the same or equal angle, but described with a radius whose metric number is r, the metric number will be nr; for all circles are similar. That is,

The metric number of an arc equals the product of the metric numbers of its central angle and its radius.

Exercise. What are the natural metric numbers of a straight angle? A right angle? An angle of 60°? Of 45°? Of 30°? Of 120°? 150°? 225°? 240°? 270°? 420°? 600°? 720°? 1080°?

THE EUCLIDIAN DOCTRINE OF PROPORTION.

337. According to Euclid, *four magnitudes, a, b, c, d, are in proportion, taken in order, when any m-fold of the first is less than, equal to, or greater than, any n-fold of the second, according as the same m-fold of the third is less than, equal to, or greater than, the same n-fold of the fourth.*

In symbols

$$a : b :: c : d$$ (*read* a is to b as c is to d)

when, and only when,

$$ma < nb, \ ma = nb, \text{ or } ma > nb,$$

according as

$$mc < nd, \ mc = nd, \text{ or } mc > nd.$$

We may also say equivalently that *a has the same ratio to b that c has to d when, and only when, etc.*

338. Hereby is defined, then, precisely, not indeed *ratio*, but at least *equality of ratios*. However, it still remains to be proved that the axiom of equal *magnitudes*, namely, *magnitudes equal to the same or equal magnitudes, are equal to eath other*, can be applied to equal *ratios*; for it has not yet been shown that ratios are magnitudes or may be treated as magnitudes. The all-important fact that, *whatever these ratios may be*, they obey the axiom of magnitudes, is expressed in the

Theorem CXI^a. — If $a : b :: c : d$ and $a : b :: e : f$,

then $c : d :: e : f$ (see Theorem CXI.).

For herein is declared that when two ratios, $c : d$ and $e : f$, are equal to the same ratio, $a : b$, they are equal to each other. *After* establishing this fundamental proposition, but *not before*, we may drop the double colon, ::, and write $a : b = c : d$.

339. We may illustrate both the idea and the method of Euclid, in demonstrating the following extremely useful

Theorem CXLVII. — *Areas of similar figures are to each other as the squares on homologous tracts* (in the figures).

Data: F and F' two similar figures, t and t' two homologous tracts, t^2 and t'^2 the squares upon them.

Proof. If the figures be curvilinear, and in general even if they be rectilinear, it will not be possible to cut them up into corresponding squares, however small, as is manifest. Nevertheless, we may cut each one up into congruent squares so small that the remainder shall be less than any assigned area, however small; that is, shall be *small at will.* For (Fig. 195) draw two corresponding series of equidistant horizontals and verticals, d apart in F, and d' apart in F'.

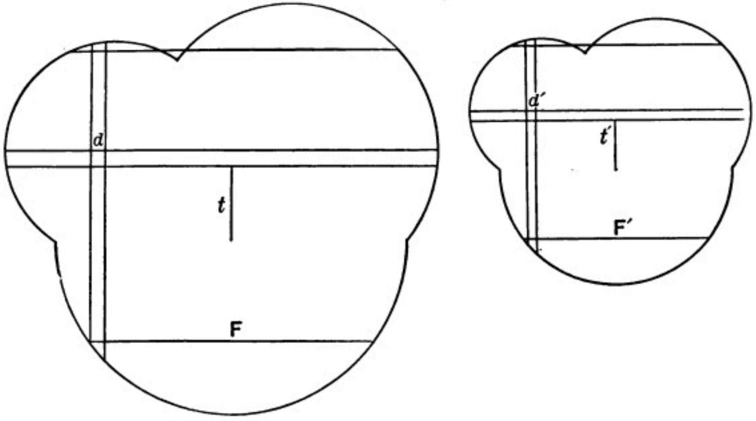

FIG. 195.

Let the extreme verticals in F be g apart, and let there be $v + 1$ of them, so that $g = v \cdot d$. Then the excess of the area of F over the sum of the squares will be the sum of the pairs of pieces at the end of each vertical strip; hence it will be less than $2v$ little squares, or less than a rectangle of base $2g$ and altitude d, *i.e.* $< 2gd$. But this rectangle is small at will, since its base $2g$ is constant and finite while its altitude d is small at will. Similarly, in F', $2g'd'$ is small as we please.

Now suppose there are p little squares cut out in F and also in F', where p is some very large but perfectly definite integer. Then the areas A and A' of F and F' will differ from the sums of the squares, pd^2 and pd'^2, by s and s', two magnitudes small at will; *i.e.*

$$A = pd^2 + s \quad \text{and} \quad A' = pd'^2 + s'.$$

Now suppose $md^2 = m'd'^2$; *i.e.* the sum of m squares in F equals the sum of m' squares in F'. Plainly then

$$p\,(md^2) = p\,(m'd'^2)\,; \quad \text{or} \quad m \cdot pd^2 = m' \cdot pd'^2 \text{ (why?)}.$$

Now

$$mA = m \cdot pd^2 + ms \quad \text{and} \quad m'A' = m' \cdot pd'^2 + m's'\,;$$

hence $$mA - m'A' = ms - m's'.$$

But s and s' are small as we please, while m and m' are finite; hence ms and $m's'$ are small at will; hence their difference, $ms - m's'$, is less than a magnitude small at will; hence *it is* 0. (Why? Because there is only one definite magnitude, namely, zero, that is less than a magnitude small at will.)

Hence, if $md^2 = m'd'^2$, then $mA = m'A'$.

Now suppose $md^2 > m'd'^2$. Then, as before, $mpd^2 > m'pd'^2$, and $mA = mpd^2 + ms$ while $m'A' = m'pd'^2 + m's'$.

Hence $$mA - m'A' = (m \cdot pd^2 - m' \cdot pd'^2) + ms - m's'.$$

Here again $ms - m's'$ is small at will, while $m \cdot pd^2 - m' \cdot pd'^2$ is some finite magnitude; hence $mA - m'A'$ is also some finite magnitude of the same sense; *i.e.*

$$\text{if } md^2 > m'd'^2, \text{ then } mA > m'A'.$$

Precisely in like fashion we prove that

$$\text{if } md^2 < m'd'^2, \text{ then } mA < m'A'.$$

That is,

$$mA < m'A', \quad mA = m'A', \text{ or } \quad mA < m'A',$$

according as

$$md^2 < m'd'^2, \quad md^2 = m'd'^2, \quad md^2 > m'd'^2.$$

Hence, by definition, the areas are proportional to the squares on the corresponding tracts, d and d'; or

$$A : A :: d^2 : d'^2.$$

Now compare the squares on t and t', d and d'. Since all squares are similar, and since d and d' are corresponding tracts or the equals of corresponding tracts in the squares on t and t', we have from the foregoing,

$$t^2 : t'^2 :: d^2 : d'^2.$$

Hence, by the Axiom of Magnitudes, applicable to ratios,

$$A : A' :: t^2 : t'^2.$$ Q. E. D.

340. Care has been taken to conduct the foregoing demonstration so that it shall apply quite as well to circles, to regular triangles, in fact to any similar figures drawn on homologous tracts as to squares; so that we may affirm

Theorem CXLVII^a. — *Areas of similar figures are proportional to areas of any similar figures on homologous tracts* (in the original similar figures).

341. The peculiar propriety and advantage of using the **square** are seen on stating the analogous arithmetical

Theorem CXLVII^b. — *The metric numbers of similar figures are proportional to the metric numbers of any other similar figures in the same ratio of similitude.*

Now the figure (or area) whose metric number is easiest to find is the *square*, whose metric number is the *second power* of the metric number of its side. Instead of *second* *power* it is usual, almost universal, to say *square* and thus employ this latter term in two entirely different senses, — the proper geometrical sense and the tropical arithmetical sense. This double use of the term "square" is very regrettable as being especially confusing to beginners.

342. We may now restate our proposition thus:

The metric numbers of two similar figures are proportional to the metric numbers of squares on homologous tracts;

Or, are proportional to the second powers of the metric numbers of homologous tracts;

Or, are in the duplicate ratio of similitude of the figures themselves.

Thus, if the ratio of similitude be $2:5$ or $a:b$, the ratio of the areas will be $4:25$ or $a^2:b^2$.

343. Observe carefully that the Euclidian doctrine of proportion is not a geometrical doctrine, but an arithmetical doctrine applied to Geometry. The same may be said of the accepted doctrine in modern texts: it is Arithmetic applied to Geometry. Nevertheless, the difference between the two is very great. Euclid's is based wholly on the operation of *multiplication* and employs only positive integers, not fractions nor irrationals, which indeed the Greek did not recognize as numbers; the modern, on the other hand, is based on the operation of *division*, and necessarily involves some general theory of fractions and irrationals. Moreover, Euclid's, while regarded as cumbrous and very difficult for the beginner, is yet a model of logical elegance and rigor: the like can hardly be said of the modern treatment.*

---

* The usual algebraical treatment of proportion is not really sound.
O. HENRICI, ENC. BRIT., VOL. X., *Geometry*, § 47.

The doctrine developed in this text is *purely geometrical,* implying no numerical knowledge or calculus. It is grounded in the notions of parallelism and similarity, to stand or fall with them. Hence it will be found to have no place in bi-dimensional spherics, the doctrine of the sphere-surface, which is in many particulars quite analogous to Planimetry, the doctrine of the plane, but in which there are no similar figures.

MAXIMA AND MINIMA.

344. Already, in Art. 135, the notions of maximum and minimum have been defined, but it is well to add here that not absolute but merely relative size is referred to, inasmuch as a varying magnitude may pass through a number of maxima and minima, and of these some maximum may be less than some minimum. Thus, a boy may inflate his elastic balloon till the diameter becomes 9 inches, then let it shrink to a diameter of 7 inches, again inflate it to a diameter of 12 inches, let it shrink to one of 10 inches, again inflate it, and so on. Here 9 and 12 are maxima, while 7 and 10 are minima of the diameter. Plainly, maxima and minima alternate with each other, and the course of the

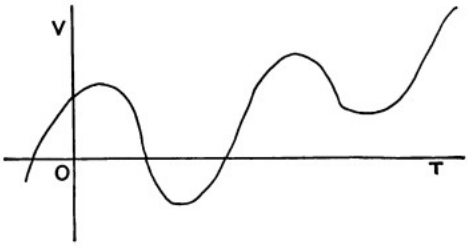

FIG. 196.

E !"#$! ! !% & ' ! EE ' ' '(. ' G ' 0-
I& 8 <

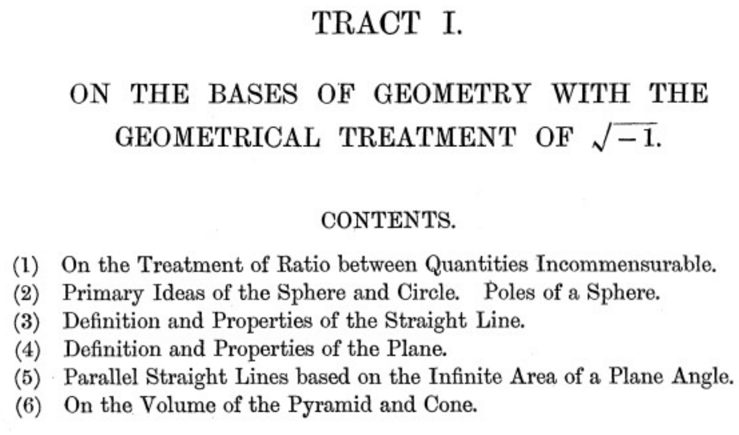

TRACT I.

ON THE BASES OF GEOMETRY WITH THE GEOMETRICAL TREATMENT OF $\sqrt{-1}$.

CONTENTS.

(1) On the Treatment of Ratio between Quantities Incommensurable.
(2) Primary Ideas of the Sphere and Circle. Poles of a Sphere.
(3) Definition and Properties of the Straight Line.
(4) Definition and Properties of the Plane.
(5) Parallel Straight Lines based on the Infinite Area of a Plane Angle.
(6) On the Volume of the Pyramid and Cone.

!"#$! ! !% & ' ! EE '(> / 7' - 0-
V " D ' / . ' '
%9 : 4 - ;;;
> - 4 D -I ' '

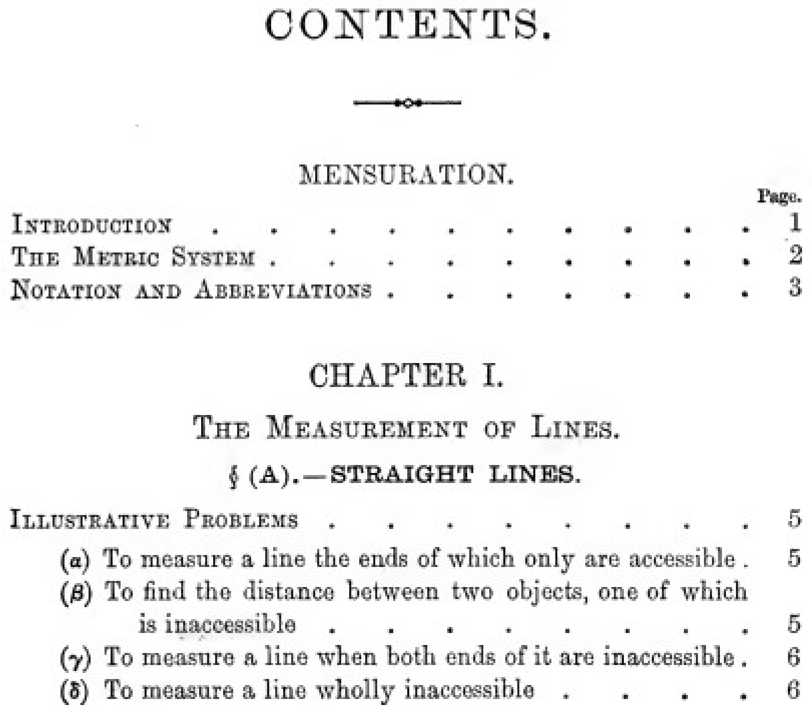

CONTENTS.

MENSURATION.

Page.
Introduction 1
The Metric System 2
Notation and Abbreviations 3

CHAPTER I.

The Measurement of Lines.

§ (A).—STRAIGHT LINES.

Illustrative Problems 5
(α) To measure a line the ends of which only are accessible . 5
(β) To find the distance between two objects, one of which is inaccessible 5
(γ) To measure a line when both ends of it are inaccessible . 6
(δ) To measure a line wholly inaccessible 6

CONTENTS.

MENSURATION.

Page.
INTRODUCTION 1
THE METRIC SYSTEM 2
NOTATION AND ABBREVIATIONS 3

CHAPTER I.

THE MEASUREMENT OF LINES.

§ (A).—STRAIGHT LINES.

ILLUSTRATIVE PROBLEMS 5
(α) To measure a line the ends of which only are accessible . 5
(β) To find the distance between two objects, one of which is inaccessible 5
(γ) To measure a line when both ends of it are inaccessible . 6
(δ) To measure a line wholly inaccessible 6

§ (B).—STRAIGHT LINES IN TRIANGLES.

Article.
I. RIGHT-ANGLED TRIANGLES 6
1. To find hypothenuse 6
2. To find side 7
II. OBLIQUE TRIANGLES 7
3. Obtuse 8
4. Acute 8
5. Given perpendicular 9
6. To find medials 9
III. STRAIGHT LINES IN SIMILAR FIGURES . . . 10
7. To find corresponding line 10
IV. CHORDS OF A CIRCLE 10
8. To find diameter 10

Article. Page.
9. To find height of arc 11
10. Given chord and height 12
11. Given chord and radius, to find chord of half the arc . . 12
12. To find circumscribed polygon 13

§ (C). — METHOD OF LIMITS.

V. Definition of a Limit 14
13. Principle of Limits 15
14. The length of the curve 16

§ (D). — THE RECTIFICATION OF THE CIRCLE.

15. Length of semicircumference 19
16. Circumferences are as their radii 20
VI. Lines in any Circle 21
17. Value of π 21
18. To find circumference 21
19. Value of diameter 21

CHAPTER II.

The Measurement of Angles.

§ (E). — THE NATURAL UNIT OF ANGLE.

VII. Angles are as Arcs 22
20. Numerical measure of angle 22
VIII. Angle Measured by Arc 23

§ (F). CIRCULAR MEASURE OF AN ANGLE.

21. Angle in radians 24
22. To find length of arc 24
23. To find number of degrees in arc 25
IX. Arcs which Correspond to Angles 25
24. Angle at center 25
25. Inscribed angle 26
26. Angle formed by tangent and chord 26
27. Angle formed by two chords 26
28. Secants and tangents 26

Article. Page.
29. Given degrees to find circular measure 26
30. Given circular measure to find degrees 27
31. Given angle and arc to find radius 27
X. ABBREVIATIONS FOR AREAS 28

CHAPTER III.

THE MEASUREMENT OF PLANE AREAS.

§ (G).—PLANE RECTILINEAR FIGURES.

XI. MEASURING AREA OF SURFACE 29
32. Area of rectangle 29
33. Area of square 32
34. Area of parallelogram 33
35. Area of triangle (given altitude) 33
36. Area of triangle (given sides) 34
37. Radius of inscribed circle 35
38. Radius of circumscribed circle 36
39. Radius of escribed circle 37
XII. TRAPEZOID AND TRIANGLE AS TRAPEZOID . . . 38
40. Area of trapezoid 38
XIII. COÖRDINATES OF A POINT 39
XIV. POLYGON AS SUM OF TRAPEZOIDS 40
41. To find sum of trapezoids 40
42. Area of any polygon 43
43. Area of quadrilateral 45
44. Area of a similar figure 48
XV. CONGRUENT AND EQUIVALENT 48
XVI. PROPERTIES OF REGULAR POLYGON 49
45. Area of regular polygon 49
46. Table of regular polygons 50

§ (H).—AREAS OF PLANE CURVILINEAR FIGURES.

47. Area of circle 51
48. Area of sector 51
49. Area of segment 52
50. Circular zone 54
51. Crescent 54

Article. Page.
52. Area of annulus 54
53. Area of sector of annulus 55
XVII. CONICS 56
54. Area of parabola 57
55. Area of ellipse 59

CHAPTER IV.

MEASUREMENT OF SURFACES.

XVIII. DEFINITIONS RELATING TO POLYHEDRONS . . . 61
56. Faces plus summits exceed edges by two 61

§ (I).—PRISM AND CYLINDER.

XIX. PARALLELEPIPED AND NORMAL 62
57. Mantel of prism 63
XX. CYLINDRIC AND TRUNCATED 64
58. Mantel of Cylinder 64

§ (J).—PYRAMID AND CONE.

XXI. CONICAL AND FRUSTUM 66
59. Mantel of pyramid 66
60. Mantel of cone 67
61. Mantel of frustum of pyramid 68
62. Mantel of frustum of cone 69
63. Frustum of cone of revolution 70

§ (K).—THE SPHERE.

XXII. SPHERE A SURFACE, GLOBE A SOLID 71
64. Area of a sphere 71
XXIII. SPHERICAL SEGMENT AND ZONE 73
65. Area of a zone 73
66. Theorem of Pappus 74
67. Surface of solid ring 74

§ (L).—SPHERICS AND SOLID ANGLES.

Article. Page.
XXIV. Steregon and Steradian 77
68. Area of a lune 78
XXV. Solid Angle made by Two, Three, or more Planes 79
XXVI. Spherical Pyramid 80
69. Solid angles are as spherical polygons 80
XXVII. Spherical Excess 80
70. Area of spherical triangle 80
71. Area of spherical polygon 81
Table of Abbreviations 83

CHAPTER V.

The Measurement of Volumes.

§ (M).—PRISM AND CYLINDER.

XXVIII. Symmetrical and Quader 84
XXIX. Volume and Unit of Volume 84
XXX. Lengths, Areas, and Volumes are Ratios . 84
72. Volume of quader 85
XXXI. Mass, Density, Weight 86
73. To find density 87
74. Volume of parallelepiped 88
75. Volume of prism 90
76. Volume of cylinder 91
77. Volume of cylindric shell 92

§ (N).—PYRAMID AND CONE.

XXXII. Altitude of Pyramid 93
78. Parallel sections of pyramid 93
79. Equivalent tetrahedra 94
80. Volume of pyramid 95
81. Volume of cone 97

§ (O).—PRISMATOID.

Article. Page.
XXXIII.-XLI. Definitions Respecting Prismatoid, 98–101
82. Volume of prismatoid 101
83. Volume of frustum of pyramid 104
84. Volume of frustum of cone 106
85. Volume of ruled surface 107
86. Volume of wedge 108
87. Volume of tetrahedron 109

§ (P). SPHERE.

88. Volume of sphere 110
89. Volume of spherical segment 112
XLII. Generation of Spherical Sector 113
90. Volume of spherical sector 113
XLIII. Definition of Spherical Ungula 114
91. Volume of spherical ungula 114
92. Volume of spherical pyramid 115

§ (Q).—THEOREM OF PAPPUS.

93. Theorem of Pappus 115
94. Volume of ring 116

§ (R).—SIMILAR SOLIDS.

XLIV. Similar Polyhedrons 118
95. Volume of similar solids 118

§ (S).—IRREGULAR SOLIDS.

96. By covering with liquid 119
97. Volume by weighing 120
98. Volume of irregular polyhedron 120

CHAPTER VI.

The Applicability of the Prismoidal Formula.

99. Test for applicability 121

§ (T). — PRISMOIDAL SOLIDS OF REVOLUTION.

Article. Page.
XLV. EXAMINATION OF THE DIFFERENT CASES . . 124

§ (U). — PRISMOIDAL SOLIDS NOT OF REVOLUTION.

XLVI. DISCUSSION OF CASES 127

CHAPTER VII.

THE NEW PRISMOIDAL FORMULA.

§ (V). — ELIMINATION OF ONE BASE.

100. Prismatoid determined from one base 130

CHAPTER VIII.

APPROXIMATION TO ALL SURFACES AND SOLIDS.

§ (W). — WEDDLE'S METHOD.

101. Seven equidistant sections 131

CHAPTER IX.

MASS-CENTER.

§ (X). — FOR HOMOGENEOUS BODIES.

102–104. Introductory 136
XLVII. BY SYMMETRY 136
105. Definition of symmetric points 136
106–117. Direct $^{\mu}C$ deductions 137
118. Triangle 138
119. Perimeter of triangle 138
120–122. Tetrahedron 138
123. Pyramid and cone 138
XLVIII. THE MASS-CENTER OF A QUADRILATERAL . 139
124. Definition of a sect 139
125. Of an opposite on a sect 139
126. $^{\mu}C$ by opposites 139
127. Geometric method of centering quadrilateral . . . 139
128. The mass-center of an octahedron 140

Article. Page.
XLIX. GENERAL MASS-CENTER FORMULA 141
129. Mass-center of any prismatoid 142
130. Test for applicability 142
131. Average haul 144
132. The $^{\mu}C$ of a consecutive series 144

EXERCISES AND PROBLEMS IN MENSURATION.

[These are arranged and classed in accordance with the above 132 Articles] 145

LOGARITHMS 225

!"#$! ! !% & ' ! EE - (> /(
. - 0-
V - > / 5 : 5 ' -

!"#$! ! !% & ' ! EE - (> /(
. - 0-

!"#$! ! !% & ' ! EE - (> / > - / 0-
V . '
V : 4 - ;;;

!"#$! ! !% & ' ! EE ' (' -(-(> /
) 0-

!"#$! ! !% & ' ! EE (> / " * 0-
V ' /5 2 5 6 4

I& 8 <= '0 W66 6 4

CHAPTER XII.

INVERSION.

145. We have already in § 13 explained what is meant by two 'inverse points' with respect to a circle. O being the centre of a circle, P and P' are inverse points if they lie on the same radius and $OP.OP' =$ the square of the radius. P and P' are on the same side of the centre, unless the circle have an imaginary radius, $= ik$, where k is real.

As P describes a curve S, the point P' will describe another curve S'. S and S' are called *inverse curves*. O is called the *centre of inversion*, and the radius of the circle is called the *radius of inversion*.

If P describe a curve in space, not necessarily a plane curve, then we must consider P' as the inverse of P with respect to a *sphere* round O. That is, whether P be confined to a plane or not, if O be a fixed point in space and P' be taken on OP such that $OP.OP' =$ a constant k^2, P' is called the inverse of P, and the curve or surface described by P is called the inverse of that described by P', and *vice versa.*

It is convenient sometimes to speak of a point P' as inverse to another point P *with respect to a point* O. By this is meant that O is the centre of the circle or sphere with respect to which the points are inverse.

146. Prop. *The inverse of a circle with respect to a point in its plane is a circle or straight line.*

First let O, the centre of inversion, lie on the circle.

Let k be the radius of inversion.

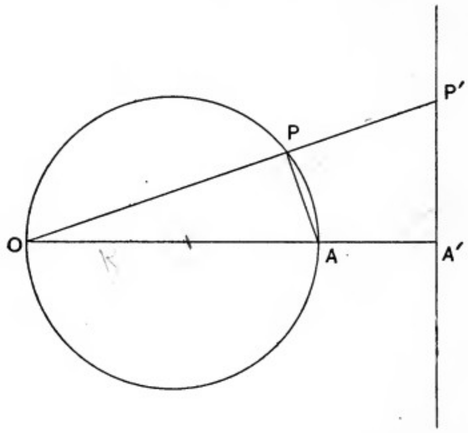

Draw the diameter OA, let A' be the inverse of A.

Let P be any point on the circle, P' its inverse.

Then $\qquad OP . OP' = k^2 = OA . OA'$.

$\therefore$ $PAA'P'$ is cyclic.

$\therefore$ the angle $AA'P'$ is the supplement of APP', which is a right angle.

$\therefore$ $A'P'$ is at right angles to AA'.

$\therefore$ the locus of P' is a straight line perpendicular to the diameter OA, and passing through the inverse of A.

Next let O not be on the circumference of the circle.

Let P be any point on the circle, P' its inverse.

Let OP cut the circle again in Q.

Let A be the centre of the circle.

Then $OP \,.\, OP' = k^2$,
and $OP \,.\, OQ =$ sq. of tangent from O to the circle $= t^2$ (say).

$$\therefore \frac{OP'}{OQ} = \frac{k^2}{t^2}.$$

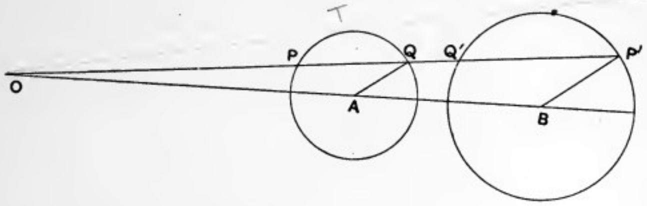

Take B on OA such that $\dfrac{OB}{OA} = \dfrac{k^2}{t^2}$. $\therefore B$ is a fixed point and BP' is parallel to AQ.

And $$\frac{BP'}{AQ} = \frac{OB}{OA} = \frac{k^2}{t^2} \text{ a constant.}$$

$\therefore P'$ describes a circle round B.

Thus the inverse of the circle is another circle.

Cor. 1. The inverse of a straight line is a circle passing through the centre of inversion.

Cor. 2. If two circles be inverse each to the other, the centre of inversion is a centre of similitude (§ 25); and the radii of the circles are to one another in the ratio of the distances of their centres from O.

The student should observe that, if we call the two circles S and S', and if OPQ meet S' again in Q', Q' will be the inverse of Q.

Note. The part of the circle S which is convex to O corresponds to the part of the circle S' which is concave to O, and *vice versa*.

Two of the common tangents of S and S' go through

O, and the points of contact with the circles of each of these tangents will be inverse points.

147. Prop. *The inverse of a sphere with respect to any point is a sphere or a plane.*

This proposition follows at once from the last by rotating the figures round *OA* as axis; in the first figure the circle and line will generate a sphere and plane each of which is the inverse of the other; and in the second figure the two circles will generate spheres each of which will be the inverse of the other.

148. Prop. *The inverse of a circle with respect to a point O,* **not in its plane,** *is a circle.*

For the circle may be regarded as the intersection of two spheres, neither of which need pass through *O*.

These spheres will invert into spheres, and their intersection, which is the inverse of the intersection of the other two spheres, that is of the original circle, will be a circle.

149. Prop. *A circle will invert into itself with respect to a point O in its plane if the radius of inversion*

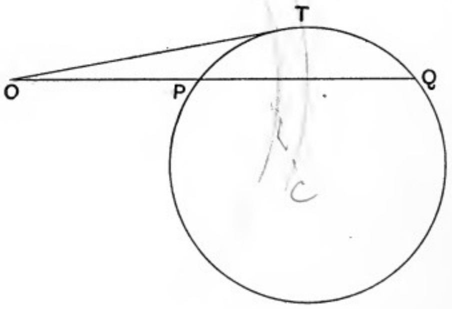

be the length of the tangent to the circle from the centre of inversion.

This is obvious at once, for if OT be the tangent from O and OPQ cut the circle in P and Q, since $OP \,.\, OQ = OT^2$ it follows that P and Q are inverse points.

That is, the part of the circle concave to O inverts into the part which is convex and *vice versa.*

Cor. 1. Any system of coaxal circles can be simultaneously inverted into themselves if the centre of inversion be any point on the axis of the system.

Cor. 2. Any three coplanar circles can be simultaneously inverted into themselves.

For we have only to take the radical centre of the three circles as the centre of inversion, and the tangent from it as the radius.

150. Prop. *Two coplanar curves cut at the same angle as their inverses with respect to any point in their plane.*

Let P and Q be two near points on a curve S, P' and Q' their inverses with respect to O.

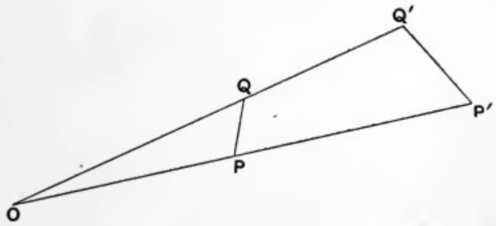

Then since $\quad OP \,.\, OP' = k^2 = OQ \,.\, OQ'$.

$\therefore$ $QPP'Q'$ is cyclic.

$\therefore$ $\angle OPQ = \angle OQ'P'$.

Now let Q move up to P so that PQ becomes the tangent to S at P; then Q' moves up at the same time to

P' and $P'Q'$ becomes the tangent at P' to the inverse curve S'.

$\therefore$ the tangents at P and P' make equal angles with OPP'.

The tangents however are antiparallel, not parallel.

Now if we have two curves S_1 and S_2 intersecting at P, and PT_1, PT_2 be their tangents there, and if the inverse curves be S_1', S_2' intersecting at P', the inverse of P, and $P'T_1'$, $P'T_2'$ be their tangents, it follows at once from the above reasoning that $\angle T_2PT_1 = \angle T_1'P'T_2'$.

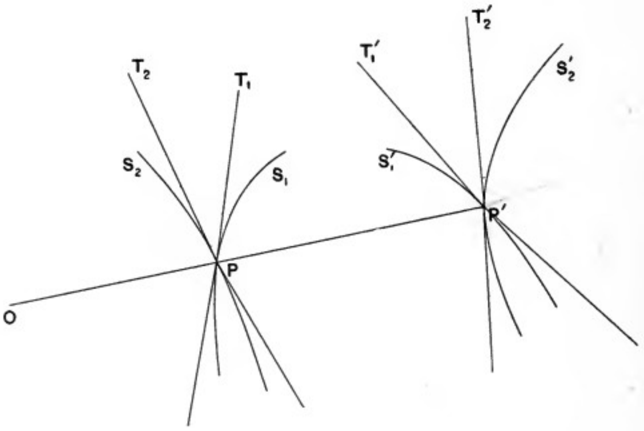

Thus S_1 and S_2 intersect at the same angle as their inverses.

Cor. If two curves touch at a point P their inverses touch at the inverse of P.

151. Prop. *If a circle S be inverted into a circle S', and P, Q be inverse points with respect to S, then P' and Q', the inverses of P and Q respectively, will be inverse points with respect to S'.*

Let O be the centre of inversion.

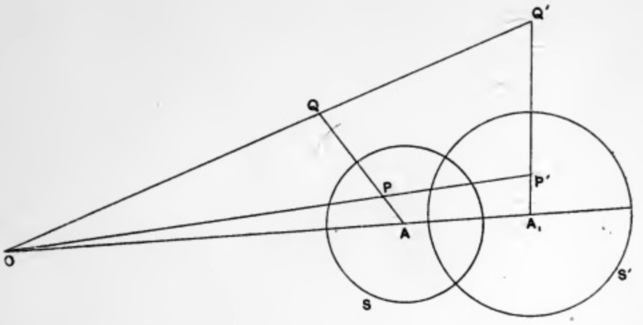

Since P and Q are inverse points for S, therefore S cuts orthogonally every circle through P and Q, and in particular the circle through O, P, Q.

Therefore the inverse of the circle OPQ will cut S' orthogonally.

But the inverse of the circle OPQ is a line; since O, the centre of inversion, lies on the circumference.

Therefore $P'Q'$ is the inverse of the circle OPQ.

Therefore $P'Q'$ cuts S' orthogonally, that is, passes through the centre of S'.

Again, since every circle through P and Q cuts S orthogonally, it follows that every circle through P' and Q' cuts S' orthogonally (§ 150).

Therefore, if A_1 be the centre of S',

$$A_1P' \,.\, A_1Q' = \text{square of radius of } S'.$$

Hence P' and Q' are inverse points for the circle S'.

152. Prop. *A system of non-intersecting coaxal circles can be inverted into concentric circles.*

The system being non-intersecting, the limiting points L and L' are real.

Invert the system with respect to L.

Now L and L' being inverse points with respect to each circle of the system, their inverses will be inverse points for each circle in the inversion.

But L being the centre of the circle of inversion, its inverse is at infinity. Therefore L' must invert into the centre of each of the circles.

153. **Feuerbach's Theorem.**

The principles of inversion may be illustrated by their application to prove Feuerbach's famous theorem, viz. *that the nine-points circle of a triangle touches the inscribed and the three escribed circles.*

Let ABC be a triangle, I its incentre and I_1 its ecentre opposite to A.

Let M and M_1 be the points of contact of the incircle and this ecircle with BC.

Let the line AII_1 which bisects the angle A cut BC in R.

Draw AL perpendicular to BC. Let O, P, U be the circumcentre, orthocentre and nine-points centre respectively.

Draw OD perpendicular to BC and let it meet the circumcircle in K.

Now since BI and BI_1 are the internal and external bisectors of angle B,

$$\therefore (AR, II_1) = -1,$$

$$\therefore L\,(AR, II_1) = -1.$$

$\therefore$ since RLA is a right angle, LI and LI_1 are equally inclined to BC (§ 27, Cor. 2).

$\therefore$ the polars of L with regard to the incircle and the ecircle will be equally inclined to BC.

Now the polar of L for the incircle goes through M and that for the ecircle through M_1.

Let MX be the polar of L for the incircle cutting OD in X.

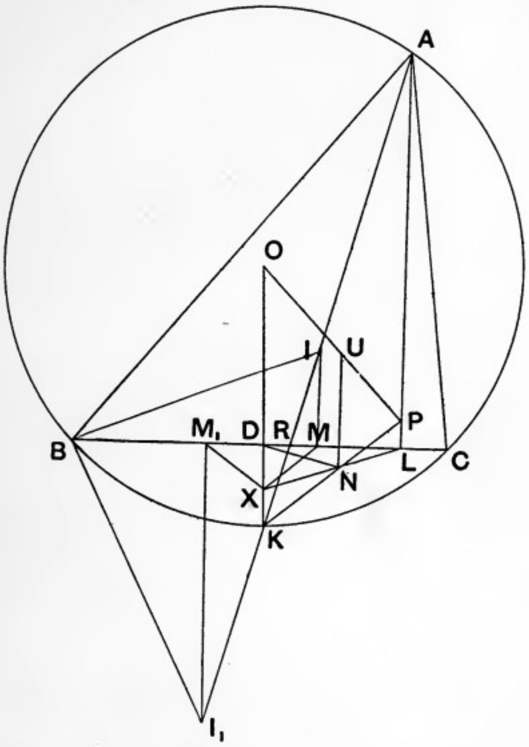

Then since D is the middle point of MM_1 (§ 12, Cor.)

$$\triangle XM_1D \equiv \triangle XMD.$$

$$\therefore \angle XM_1D = \angle XMD.$$

$\therefore M_1X$ is the polar of L for the ecircle, i.e. L and X are conjugate points for both circles.

Let N be the middle point of XL, then the square of the tangent from N to both circles $= NX^2 = ND^2$.

$\therefore$ N is on the radical axis of the two circles; but so also is D since $DM = DM_1$.

$\therefore$ ND is the radical axis, and this is perpendicular to II_1.

Now the pedal line of K goes through D, and clearly also, since K is on the bisector of the angle A, the pedal line must be perpendicular to AK.

$\therefore$ DN is the pedal line of K.

But the pedal line of K bisects KP.

$\therefore$ KNP is a straight line and N its middle point.

And since U is the middle point of OP, $UN = \frac{1}{2}OK$.

$\therefore$ N is a point on the nine-points circle.

Now invert the nine-points circle, the incircle and ecircle with respect to the circle whose centre is N and radius ND or NL.

The two latter circles will invert into themselves; and the nine-points circle will invert into the line BC; for N being on the nine-points circle the inverse of that circle must be a line, and D and L, points on the circle, invert into themselves, $\therefore$ DL is the inverse of the nine-points circle.

But this line touches both the incircle and ecircle.

$\therefore$ the nine-points circle touches both the incircle and ecircle.

Similarly it touches the other two ecircles.

Cor. The point of contact of the nine-points circle with the incircle will be the inverse of M, and with the ecircle the inverse of M_1.

= !"#$! ! !% & ' ! EE #' ' (> / 7' - 0-
V ' '
I& = 8 = '0 6

CHAPTER IV.

PARALLELS.

IV. Assumption of Parallels (Euclid's Postulate).

IV. *Through a given point there is not more than one parallel to a given straight.*

71. The introduction of this assumption greatly simplifies the foundation and facilitates the construction of geometry.

72. Theorem. *Two straights parallel to a third are parallel.*

Proof. Were 1 and 2 not parallel, then there would be through their intersection point two parallels to 3, which is in contradiction to IV.

73. Theorem. *If a transversal cuts two parallels, the alternate angles are congruent.*

Proof. Were say $\sphericalangle BAD$ not $\equiv \sphericalangle ABC$, then we could through A (by III 4) take a straight making $\sphericalangle BAD' \equiv \sphericalangle ABC$ [D' and D on same side of AB], and so we would have (by 69) through A two parallels to a, in contradiction to IV.

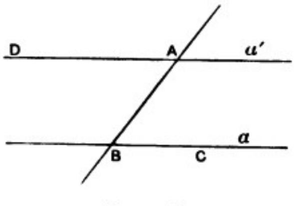

FIG. 26.

74. Corollary to 73. A perpendicular to one of two parallels is perpendicular to the other also.

75. Theorem. If a transversal cuts two parallels, the corresponding angles are congruent.

Proof. The angle vertical to one is alternate to the other.

Ex. 27. A straight meeting one of two parallels meets the other also.

Ex. 28. A straight cutting two parallels makes conjugate angles supplemental.

Ex. 29. If alternate or corresponding angles are unequal or if conjugate angles are not supplemental, then the straights meet. On which side of the transversal?

76. Theorem. A perpendicular to one of two parallels is parallel to a perpendicular to the other.

Proof. Either of the two given parallels makes (by 74) right angles with both perpendiculars, which therefore are parallel by 69.

77. Corollary to 76. Two straights respectively perpendicular to two intersecting straights cannot be parallel.

Proof. For if they were parallel, then (by 76) the intersecting straights would also be parallel.

78. Convention. When two angles are set off from the vertex of a third against its sides so that no point is interior to two, if the two sides not common are costraight, the three angles are said together to form two right angles.

79. *The angles of a triangle together form two right angles.*

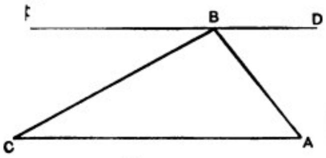

FIG. 27.

Proof. Take alternate $\sphericalangle CBF \equiv \sphericalangle C$ and $\sphericalangle ABD \equiv \sphericalangle A$; then (by 69) can neither BF nor BD cut AC. By the parallel postulate IV, then is FBD a straight.

80. Theorem. If two angles of one triangle are congruent to two of another, then the third angles are congruent.

Proof. Given $\measuredangle A \equiv \measuredangle A'$ and $\measuredangle B \equiv \measuredangle B'$. Take CP parallel ($\|$) to AB and $C'P' \parallel$ to $A'B'$. Then

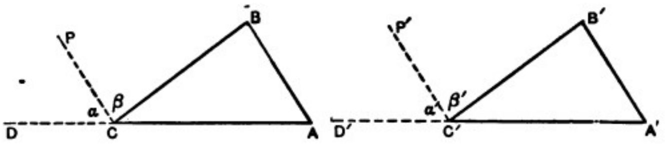

FIG. 28.

$\measuredangle \alpha \equiv \measuredangle A$ and $\measuredangle \beta \equiv \measuredangle B$, $\measuredangle \alpha' \equiv \measuredangle A'$ and $\measuredangle \beta' \equiv \measuredangle B'$. $\therefore$ (by 49) $\measuredangle BCD \equiv \measuredangle B'C'D'$. $\therefore$ (by 45) the adjacent angles $\measuredangle ACB \equiv \measuredangle A'C'B'$.

81. Theorem. *Two triangles are congruent if they have a side, an adjoining and the opposite angle respectively congruent.*

Proof. By 80 and 44.

Ex. 30. Every triangle has at least two acute angles.

Ex. 31. If the rays of one angle are parallel or perpendicular to those of another, the angles are congruent or supplemental.

Ex. 32. In a *right-angled* triangle [a triangle one of whose angles is a right angle] the two acute angles are complemental (calling two angles *complements* which together form a right angle).

82. Theorem. *In any sect AB there is always one and only one point C such that $AC \equiv BC$.*

Proof. Take any angle BAD at A against AB, and the angle congruent to it at B against BA and on the opposite side of a in the plane BAD; and take any sect AD on the free ray from A, and one

BF congruent to it on the free ray from B. The sect DF must cut a, say in C, since D and F are on opposite sides of a. Moreover, C is between A and B. Otherwise one of them, say A, would be between B and C. But then DA would have a point A on BC, a side of triangle FBC, and so (by II 4) must meet another side. But this is impossible, since it meets FC produced at D and is parallel to BF. Since thus $\sphericalangle A \equiv \sphericalangle B$, and $\sphericalangle ACD \equiv \sphericalangle BCF$ [vertical], therefore (by 81) $\triangle ACD \equiv \triangle BCF$.

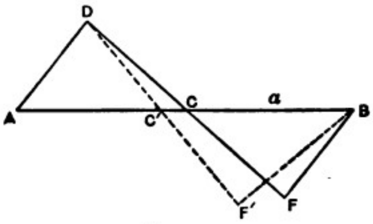

FIG. 29.

Therefore $AC \equiv BC$.

If we suppose a second such point C', then on ray DC' take $C'F' \equiv DC'$. Therefore (by 43) $\sphericalangle C'BF' \equiv \sphericalangle DAC \equiv \sphericalangle ABF$, and $BF' \equiv AD \equiv BF$. Therefore F' is F and C' is C.

83. Convention. The point C of the sect AB such that $AC \equiv BC$ may be called the *bisection-point* of AB, and to *bisect* AB shall mean to take this point C.

Ex. 33. Parallels through the end-points of a sect intercept congruent sects on any straight through its bisection-point.

Ex. 34. In a right-angled triangle the bisection-point of the *hypothenuse* (the side opposite the r't $\sphericalangle$) makes equal sects with the three vertices.

Hint. Take one acute $\sphericalangle$ in the r't $\sphericalangle$.

84. Theorem. *Within any* $\measuredangle (h, k)$ *there is always one and only one ray,* l*, such that* $\measuredangle (h, l) \equiv \measuredangle (k, l)$.

Proof. From the vertex O take $OA \equiv OA'$. By 82 take C, the bisection-point of AA'. Then $\triangle AOC \equiv \triangle A'OC$ [$\triangle$s with 3 sides $\equiv$ are $\equiv$]. $\therefore \measuredangle AOC \equiv \measuredangle A'OC$.

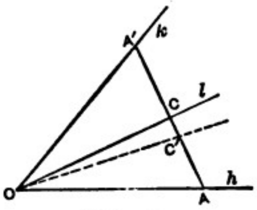

FIG. 30.

If we suppose a second such ray OC', then $\triangle AOC' \equiv \triangle A'OC'$ [$\triangle$s with 2 sides and the included $\measuredangle \equiv$ are $\equiv$]. $\therefore AC' \equiv A'C'$. $\therefore$ (by 82) C' is C.

85. Convention. The ray l of $\measuredangle (h, k)$ such that $\measuredangle (h, l) \equiv \measuredangle (l, k)$ may be called the *bisection ray* or bisector of $\measuredangle (h, k)$, and to *bisect* $\measuredangle (h, k)$ shall mean to take this ray l.

Ex. 35. An angle may be separated into 2, 4, 8, 16, . . . , 2^n congruent angles.

SYMMETRY.

86. Definition. Two points are said to be *symmetrical* with regard to a straight, when it bisects at right angles their sect. The straight is called their *axis of symmetry*. Two points have always one, and only one, symmetry axis.

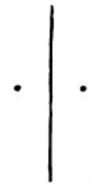

FIG. 31.

A point has, with regard to a given axis of symmetry, always one, and only one, symmetrical point, namely, the one which ends the sect from the given point perpendicular to the axis and bisected by the axis.

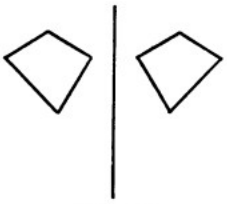

FIG. 32.

87. Definition. Two figures have an axis of symmetry when, with regard to this straight, every point of each has its symmetrical point on the other.

One figure has an axis of symmetry when, with regard to this straight, every point of the figure has its symmetrical point on the figure.

One figure is called symmetrical when it has an axis of symmetry.

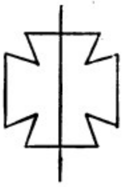

FIG. 33.

Any figure has, with regard to any given straight as axis, always one, and only one, symmetrical figure.

88. Theorem. *An angle is symmetrical with regard to its bisector and the end-points of congruent sects from the vertex are symmetrical.*

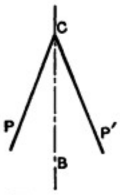

FIG. 34.

Proof. Their sect is bisected at right angles by the angle-bisector.

89. Definition. A symmetrical quadrilateral with a diagonal as axis is called a *deltoid*.

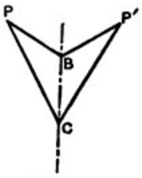

FIG. 35.

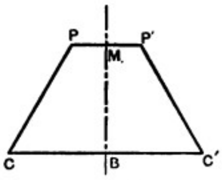

FIG. 36.

90. Definition. A sect whose end-points are the bisection-points of opposite sides of a quadrilateral is called a *median*. So is the sect from a vertex of a triangle to the bisection-point of the opposite side.

91. Definition. A symmetrical quadrilateral with a median as axis is called a *symtra*.

Ex. 36. In a r't Δ if to set off one acute ∡, α, in the other, β, bisects, so is it with β's sides.

Ex. 37. The st' through the bisection-point of the base of a ∔ Δ, and the opposite vertex is ⊥ to the base and bisects ∡.

Ex. 38. The r't bi' of the base of a ∔ Δ bisects ∡ at the vertex.

Ex. 39. The ⊥ from vertex bisects base and ∡ in a ∔ Δ.

Ex. 40. The bisector of ∡ at vertex of a ∔ Δ is r't bi' of base.

Ex. 41. If a r't bi' of a side contains a vertex, the Δ is ∔.

Ex. 42. The bisector of an exterior ∡ at vertex of ∔ Δ is ‖ to base, and inversely.

Ex. 43. The end of sect from intersection of congruent sides of a ∔ Δ costraight and ≡ to one determines with end of other a ⊥ to base.

Ex. 44. To erect a ⊥ at the end-point of a sect without producing the sect.

Ex. 45. A ‖ to one side of an ∡ makes with its bisector and other side a ∔ Δ.

Ex. 46. The bisectors of the ≡ ∡s of a ∔ Δ are ≡.

Ex. 47. Every symmetrical quadrilateral not a deltoid is a symtra.

Ex. 48. The intersection point of two symmetrical straights is on the axis.

Ex. 49. The bisector of an angle is symmetrical to the bisector of the symmetrical angle.

Ex. 50. A figure made up of a straight and a point is symmetrical.

Ex. 51. In any deltoid [1] One diagonal (the axis) is the perpendicular bisector of the other. [2] One diagonal (the axis) bisects the angles at its two vertices. [3] Sides which meet on one diagonal (the axis) are congruent; so each side equals one of its adjacent sides. [4] One diagonal (not the axis) joins the vertices of congruent angles and makes congruent angles with the congruent sides. [5] The triangles made by one diagonal (the axis) are congruent. [6] One diagonal (not the axis) makes two isosceles triangles.

Ex. 52. Any quadrilateral which has one of the six preceding pairs of properties (Ex. 51) is a deltoid.

Ex. 53. A quadrilateral with a diagonal which bisects the angle made by two sides, and is less than each of the other two sides, and these sides congruent, is a deltoid with this diagonal as axis.

Ex. 54. A quadrilateral with a side meeting a congruent side in a greater diagonal which is opposite congruent angles is a deltoid with that diagonal as axis.

Ex. 55. In any symtra [1] Two opposite sides are parallel, and have a common perpendicular bisector. [2] The other two sides are congruent and make congruent angles with the parallel sides.

[3] Each angle is congruent to one and supplemental to the other of the two not opposite it.

[4] The diagonals are congruent and their parts adjacent to the same parallel are congruent.

[5] One median bisects the angle between the two diagonals, and also the angle between the non-parallel sides (produced).

Ex. 56. Any quadrilateral which has one of the preceding five pairs of properties (Ex. 55) is a symtra.

92. Definition. A *trapezoid* is a quadrilateral with two sides parallel.

93. Definition. A *parallelogram* is a quadrilateral with each side parallel to another (its opposite).

94. Definition. A parallelogram with one angle right is called a *rectangle*. A parallelogram with two consecutive sides congruent is called a *rhombus*. A rectangle which is a rhombus is called a *square*.

95. Theorem. *The opposite sides and angles of a parallelogram are congruent, and its diagonals bisect each other.*

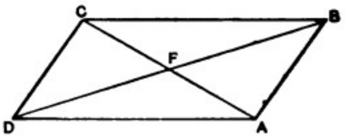

FIG. 37.

Proof. $\triangle ABC \equiv \triangle ADC$ [side and 2 adjoining $\sphericalangle$s $\equiv$]. $\therefore BC \equiv AD$. $\therefore$ (as in 82) $AF \equiv FC$ and $BF \equiv FD$.

96. Theorem. *If three parallels make congruent sects on one transversal, they do on every transversal.*

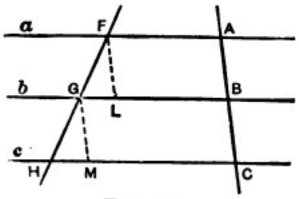

FIG. 38.

Given $a \parallel b \parallel c$, also $AB \equiv BC$.

To prove $FG \equiv GH$.

Proof. Take $FL \parallel GM \parallel AB$. Then $FL \equiv AB \equiv BC \equiv GM$ [95, opposite sides of a $\parallel$gm are $\equiv$]. $\therefore \triangle FLG \equiv \triangle GMH$ [side and 2 adjoining $\sphericalangle$s $\equiv$]. $\therefore FG \equiv GH$.

97. Corollary to 96. A straight through the bisection-point of one side of a triangle and parallel to a second side bisects the third side. [In figure let F coincide with A.]

98. Inverse of 97. The straight through the bisection-points of any two sides of a triangle is parallel to the third side. [For, by 97, it is identical with the $\parallel$ to the third side through either bisection-point.]

99. Theorem. *The sect whose end-points are the bisection-points of two sides of a triangle is congruent to each sect made in bisecting the third side.*

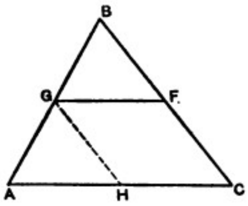

Fig. 39.

Proof. By 97 $GH \parallel BC$ bisects AC. Since (by 98) $FG \parallel CH$, $\therefore$ (by 95) $FG \equiv CH$.

100. Theorem. *If two sides of a quadrilateral are congruent and parallel it is a parallelogram.*

Given $AB \equiv$ and $\parallel CD$.

Proof. $\triangle ABC \equiv \triangle ADC$.

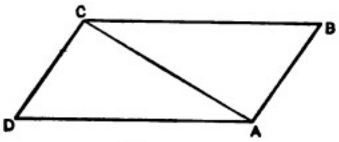

Fig. 40.

$$\therefore \measuredangle ACB \equiv \measuredangle CAD. \quad \therefore CB \parallel AD.$$

Ex. 57. Every straight through the intersection of its diagonals cuts any parallelogram into congruent trapezoids.

Ex. 58. A quadrilateral with each side equal to its opposite is a parallelogram.

Ex. 59. A quadrilateral with a pair of opposite sides equal, and each greater than a diagonal, making equal alternate angles with the other sides, is a parallelogram.

Ex. 60. A quadrilateral with a side equal to its opposite, and less than a diagonal opposite equal angles, is a parallelogram.

Ex. 61. A quadrilateral with each angle equal to its opposite is a parallelogram.

Ex. 62. A quadrilateral whose diagonals bisect each other is a parallelogram.

101. Theorem. In any sect AB there are always two, and only two, points, C, D, such that $AC \equiv CD \equiv DB$.

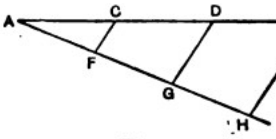

FIG. 41.

Proof. Take on any ray from A, any sect AF, and a sect $FG \equiv AF$, and a sect $GH \equiv FG$. Take $FC \parallel GD \parallel HB$. Then, by 96, $AC \equiv CD \equiv DB$.

Suppose two other such points C', D'. Then, by 98, $C'F \parallel D'G$. Now $HB' \parallel GD'$ (by 96) makes $D'B' \equiv D'C'$. $\therefore$ from our hypothesis and III 1, B' is identical with B. $\therefore$ since $GD \parallel HB$ (by IV) D' is identical with D. $\therefore$ since $FC \parallel GD$ (by IV) C' is identical with C.

102. The two points, C, D, of the sect AB such that $AC \equiv CD \equiv DB$ may be called the *trisection-points* of AB.

103. Theorem. *The three medians of a triangle are copunctal in that trisection-point of each remote from its vertex.*

Proof. Any median AG must meet any other CF, since A and G are on different sides of the straight CF, and so the cross of st' AG with st' CF is on sect AG, and similarly it is on sect CF. If P, Q, are bisection-points of OC and OA, then (by 98 and 99) $PQ \parallel$ and $\equiv GF$. $\therefore$ by 100 $PQFG$ is a $\parallel$gm and (by 95) PF and QG bisect each other.

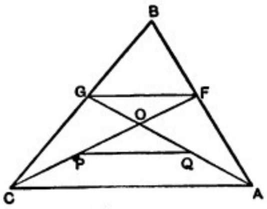

FIG. 42.

104. Definition. The cointersection-point of its medians is called the triangle's *centroid.*

105. Definition. A perpendicular from a vertex to the straight of the opposite side is called an *altitude* of the triangle. This opposite side is then called the *base.* The perpendicular from a vertex of a parallelogram to the straight of a side not through this vertex is called the altitude of the parallelogram with reference to this side as base.

Ex. 63. The bisectors of the four angles which two intersecting straights make with each other form two straights perpendicular to each other.

Ex. 64. If four coinitial rays make the first angle congruent to the third, and the second congruent to the fourth, they form two straights.

Ex. 65. How many congruent sects from a given point to a given straight?

Ex. 66. Does the bisector of an angle of a triangle bisect the opposite side?

Ex. 67. The bisectors of vertical angles are costraight.

Ex. 68. If two isosceles triangles be on the same base the straight determined by their vertices bisects the base at right angles.

Ex. 69. Suppose a Δ to be 3 bars freely jointed at the vertices. Is it rigid? Are the ∡s fixed and the joints of no avail? Of what theorem is this a consequence? How is it with a jointed quadrilateral? Why?

Ex. 70. Joining the bisection-points of the sides of a Δ cuts it into 4 ≡ Δs.

Ex. 71. Joining the bisection-points of the consecutive sides of a quadrilateral makes a ‖ g'm.

Ex. 72. The medians of a quadrilateral and the sect joining the bisection-points of its diagonals are all three bisected by the same point.

Ex. 73. If the bisection-points of two opposite sides of a ∥ g'm are joined to the vertices the diagonals are trisected.

Ex. 74. The ⊥s from any point in the base of an ∔ Δ to the sides are together an altitude.

Ex. 75. The diagonals of a rectangle are ≡, of a rhombus are ⊥.

Ex. 76. If 2 ∥s are cut by a transversal, the bisectors of the interior ∡s make a rectangle.

Ex. 77. The angle-bisectors of a rectangle make a square.

Ex. 78. If the ∡s adjoining one of the ∥ sides of a trapezoid are ≡, so are the others.

Ex. 79. The bisectors of the interior ∡s of a trapezoid make a quad' with 2 r't ∡s.

Ex. 80. The bisection-point of one sect between ∥s bisects any through it.

Ex. 81. The ≡ altitudes in ∔ Δ make with the base ∡s ≡ to those made in bisecting the other ∡.

Ex. 82. Through a given point within an ∡ draw a sect terminated by the sides and bisected by the point.

Ex. 83. Sects from the vertex to the trisection-points of the base of ∔ Δ are ≡.

Ex. 84. If the ∡s made by producing a side of a Δ are ≡, so are the other sides.

Ex. 85. If a quad' has 2 pairs of congruent consecutive sides, the other ∡s are ≡.

Ex. 86. Two Δs are ≡ if two sides and one's median are respectively ≡.

Ex. 87. Two ∔ Δs are ≡ if one ∡ and altitude are ≡ to the corresponding.

Ex. 88. Two Δs are ≡ if a side, its altitude and an adjoining ∡ are respectively ≡.

Ex. 89. If 2 altitudes are ≡ the Δ is ∔.

Ex. 90. Two Δs are ≡ if two sides and one's altitude are ≡ to the corresponding.

Ex. 91. Two Δs are ≡ if a side, its altitude and median are respectively ≡.

Ex. 92. Two Δs are ≡ if a side and the other 2 altitudes are respectively ≡.

Ex. 93. Two △s are ≅ if a side, an adjoining ∡ and its bisector are respectively ≅.

Ex. 94. Two equilateral △s are ≅ if an altitude is ≅.

Ex. 95. The bisector is within the ∡ made by altitude and median.

Ex. 96. In a right △ one bisector also bisects ∡ between its altitude and median.

Ex. 97. Two sects from the vertices of a △ to the opposite sides cannot bisect each other.

Ex. 98. The ⊥s from 2 vertices of a △ upon the median from the third are ≅.

Ex. 99. Two ∥g'ms having an ∡ and the including sides ≅ are ≅.

Ex. 100. The ⊥ from the circumcenter to a side is half the sect from the opposite vertex to the orthocenter.

Ex. 101. The centroid is the trisection point of the sect from orthocenter to circumcenter remote from the orthocenter

@ !"#$! ! !% & ' ! EE (> / > - / 0-
%9 : 4 - ;;;
V E 0'& : 4 N ;;;

< !"#$! ! !% & ' ! EE -(> / > - / 0-

E !"#$! ! !% & ' ! EE -(> / : 0-
V D,0 ' - ' ' 0 -' ' -' & '
%9 : 4 - ;;;

!"#$! ! !% & ' ! EE 0' (' -(> / ' 0-
V D - -' ' 3' - 0'i

!"#$! ! !% & ' ! EE ' & (: (' -(4
' / 4 0-
%9 : 4 - ;;;
6 ' ;;;

!"#$! ! !% & ' ! EE B * ((" - 7 ' 0-
V = 0'&

!F"." G 4 4 !) * ! <!# &K K L ' K
- 0' ' n >' !> / 2 4

!"#$! ! !% & ' ! E " & 9 ' (2 4 (> /
1 9 0-
V @ 0'&
%9 : 4 - ;;;

!"#$! ! !% & ' ! E '0 L ' 0-
%9 % - ' L 4 I '-
D 4 ' ' - & '
"-4' - " ' / > / ' 09
V E 0'&

%9 : ' . ' '
V ? ' 4 - ;;;

0 ' 4 & - ' - = E @

' !"#$! ! !% & ' ! E = E @ "-4' - " ' /
> / 0-

9 !"#$! ! !% & ' ! E = E @ "-4' - " ' /
> /(, 0-

CONTENTS

PART I

INTRODUCTION TO AFFINE PLANE ANALYTIC PROJECTIVE GEOMETRY

CHAPTER I

Various Coordinate Systems

SECTION PAGE

1, 2. Frames of reference 1
3. Oblique axes 3
4. Transformations from rectangular to oblique axes 5
5. Tangents to curves in oblique coordinates 9

CHAPTER II

Preliminary Discussion of Linear Transformations

6. Translations and rotations 11
7. Products of transformations 13
8. Inverses of transformations and the identical transformation 14
9. An algebraic notation for transformations 15
10. A slightly more general transformation of coordinates 16
11. Two interpretations of transformations of coordinates 18

CHAPTER III

Introduction to Affine Linear Transformations

12. Affine linear transformations 21
13. The matrix of an affine linear transformation 22
14. The number of points that determine an affine linear transformation 24
15. The equivalence of conics under an affine linear transformation 27
16, 17. Degenerate conics 28
18. Preliminary discussion of invariants of an affine linear transformation 30
19. The discriminant and other invariants of a conic 34
20. Transformations that have given invariants 37
21. The effect of affine linear transformations on the equations of lines 39
22. A first hint of plane duality 40

CHAPTER IV

Introduction to the Study of Cross-Ratio

23. Cross-ratio, an important invariant of affine linear transformations. 44
24. Harmonic sets of points 47

SECTION PAGE
25. A first hint of infinite points on lines........................ 49
26. A geometric construction for a harmonic set of points........... 50
27. Harmonic sets of concurrent lines.................... 54
28. A geometric construction for a harmonic set of lines............ 56
29. A first hint of projection and section.......................... 57
30. Parametric coordinates for collinear points, their cross-ratios...... 58

CHAPTER V

PRELIMINARY DISCUSSION OF GROUPS OF LINEAR TRANSFORMATIONS AND THEIR ASSOCIATED GEOMETRIES

31. Introduction to groups of transformations..................... 62
32. Subgroups of groups of linear transformations.......... 65
33. Invariant points and lines of linear transformations............ 65
34. The parabolic metric group........... 67
35. Geometries associated with groups of transformations........... 69

CHAPTER VI

PRELIMINARY DISCUSSION OF IMAGINARY ELEMENTS IN GEOMETRY

36. Imaginary points......... 71
37. Imaginary lines......... 73
38. Imaginary curves...... 77
39. Inside points with respect to conics........................... 79

CHAPTER VII

ELEMENTARY DISCUSSION OF *n*TH DEGREE CURVES

40. Introduction to *n*th degree curves (*n*-ics)........... 80
41. Some types and typical forms of *n*-ics..... 81
42. Conditions determining an *n*th degree curve.................. 84
43. Points of intersection of a line with an *n*-ic.......... 87
44. Asymptotes to *n*-ics.... 90
45. Points of intersection of *m*-ics and *n*-ics..... 92
46. Tangents to *n*-ics......................... 96
47. Points of inflection on *n*-ics........ 100
48. Multiple points on *n*-ics 102
49. Parallel lines and the line at infinity in a plane................ 107

CHAPTER VIII

CONICS AND LINEAR FAMILIES OF CONICS

50. Poles and polars with respect to conics....... 110
51. Some applications of poles and polars with respect to a conic....... 113
52. A use of poles and polars to derive the discriminant of a conic.... . 115
53. Intersections of conics 116

SECTION PAGE
54. Pencils of circles 117
55. Degenerate circles in a pencil of circles 119
56. Transformations in a pencil of circles, linear in variables and bilinear in parameter 120
57. Pencils of conics 122
58. Degenerate conics in a pencil of conics 123
59. Transformations on pencils of conics, linear in the variables, bilinear in the parameter 126
60. Nets of conics 129
61. Cubics associated with a net of conics, the discriminant and the Jacobian 130
62. Transformations of nets of conics, linear in the variables, bilinear in the parameters (also called linear fractional) 134
63. Three-and-four-parameter linear families of conics 136

PART II

INTRODUCTION TO GENERAL PLANE ANALYTIC PROJECTIVE GEOMETRY

CHAPTER IX

INTRODUCTION TO THE TRIANGLE OF REFERENCE

64. Rectangular and oblique axes from a new viewpoint 139
65. The line at infinity and the axes of reference 141
66. Desargues' theorem on perspective triangles 144
67. Complete quadrangles and complete quadrilaterals 145
68. Quadrangular sets of points (lines), harmonic sets 148
69. The fundamental theorem for quadrangular sets of points 150
70. The triangle of reference 153
71. A note on bilinear (or linear fractional) transformations 156
72. The effect of bilinear transformations on the equations of lines 160
73. A note on transformations of points on a line and between lines 162
74. Change from a triangle of reference to axes of reference 164

CHAPTER X

INTRODUCTION TO HOMOGENEOUS COORDINATES

75. Homogeneous coordinates for points 168
76. Equations in homogeneous coordinates 170
77. Homogeneous coordinates and l_∞; non-homogeneous coordinates for points on l_∞ 172
78. Homogeneous coordinates and a triangle of reference; non-homogeneous coordinates for points on $X_\infty Y_\infty$ 174
79. Change of homogeneous coordinates from a triangle of reference to axes plus l_∞ 175
80. The equations of curves referred to a triangle of reference 177

CHAPTER XI

SOME THEOREMS ABOUT THE LINE AT INFINITY

SECTION PAGE

81. Intersections of curves with l_∞, asymptotes 179
82. Use of l_∞ to discuss types of conics 180
83. Circles and the circular points at infinity, minimal lines, perpendicular lines 181
84. Rotation and translations and l_∞ 183
85. The general affine linear transformations and l_∞ 185
86. The real plane as part of a complex plane; conjugate imaginary points and lines 187

CHAPTER XII

INTRODUCTION TO LINE COORDINATES AND PLANE DUALITY

87. Homogeneous and non-homogeneous line coordinates 190
88. Triangle of reference for line coordinates; axes and line coordinates. 193
89. Cross-ratios of non-homogeneous line coordinates 194
90. Points of contact of line loci; line coordinates for tangents to curves 196
91. The conics and other n-ics in line coordinates; the class of an n-ic. . . 197
92. Plane duality 201
93. Plane duality from the purely geometric (synthetic) standpoint . . . 202
94. Points and lines and plane duality from a strictly analytic viewpoint 206

CHAPTER XIII

INTRODUCTION TO GENERAL ANALYTIC PROJECTIVE GEOMETRY

95. General linear (or projective) transformations in point and line coordinates, general analytic projective geometry 208
96. Invariant points and lines under general projectivities 212
97. Collineations, homologies, and elations 215
98. Types of collineations 218
99. The resolution of a collineation into a product of homologies and elations; reduction of a collineation to a typical form 219
100. Complete quadrangle (quadrilateral) as determining a projectivity. 221
101. Geometrical construction of collineations 224
102. Transformations of points on a line and between lines in homogeneous coordinates 225
103. Invariant (double) points of projectivities on a line; hyperbolic, parabolic, elliptic transformations 228
104. Involutions of points on a line 229
105. Geometric description of projectivities between lines 232
106. Cross-ratio 236
107. Harmonic sets of points and of lines 240
108. The absolute involution 243

CHAPTER XIV

INTRODUCTION TO CORRELATIONS AND POLARITIES

SECTION PAGE

109. Tangents to curves 245
110. Poles and polars with respect to a conic 246
111. Polarities; their relation to duality 248
112. Correlations in general 250

CHAPTER XV

SOME THEOREMS ON CONICS

113. Pencils of points and pencils of lines; parametric coordinates for points and for lines 252
114. A point (line) conic as determined by two projective pencils of lines (points) 256
115. The Pascal line (Brianchon point) of a hexagon inscribed in (circumscribed about) a conic 260
116. Projectivities on a conic; center and axis of homology 262
117. Involutions on a conic 263
118. One-to-one correspondences between the points on a conic and the points on a line 264

CHAPTER XVI

SOME THEOREMS ABOUT COMPLETE QUADRANGLES

119. Two theorems on the complete quadrangle 266
120. A theorem about a complete quadrangle inscribed in a conic 269
121. Four theorems on quadrangular sets and on cross-ratio 271

CHAPTER XVII

FURTHER DISCUSSION OF *n*-ICS

122. Reduction of a conic and of a cubic to simpler forms 275
123. Polars of points with respect to *n*-ics 277
124. The class *m* of an *n*-ic 279
125. The Hessian of an *n*-ic; inflections on an *n*-ic 283
126. Inflections on cubics and their harmonic polars 285
127. The reduction of a cubic curve to a normal form 287

CHAPTER XVIII

FURTHER DISCUSSION OF LINEAR FAMILIES OF CONICS

128. Apolarity 290
129. Apolar linear families of conics 292
130. Apolarity for three-parameter linear families of conics 294

SECTION PAGE

131. Apolarity and nets of conics 295
132. The derivation of some typical pencils of conics 296
133. The derivation of some typical degenerate pencils of conics 298
134. The derivation of some typical degenerate nets of conics 300
135. The derivation of some typical non-degenerate nets of conics 302

!"#$! ! !% & ' ! E ", (2 4 (> /
B '- 0-

= !"#$! ! !% & ' ! E
+ -' ((2 4 (> / F , 0-

@ !"#$! ! !% & ' ! E 7 & (> / B - 0-

0 0 9 ' / & 9 * %:@@= 7 & (& /

0 '@ ' 4 & & & - & & -
0-

0 ' 4 & - ' & & -

@' !"#$! ! !% & ' ! E & & - 0-
V @ 0'&

@9 !"#$! ! !% & ' ! E & & - -24

CONTENTS

PART I. GENERAL CONCEPTS AND ONE-DIMENSIONAL GEOMETRY

CHAPTER I. GENERAL CONCEPTS

SECTION PAGE
1. Coördinates 1
2. The principle of duality 2
3. The use of imaginaries 2
4. Infinity 3
5. Transformations 4
6. Groups 6

CHAPTER II. RANGES AND PENCILS

7. Cartesian coördinate of a point on a line 8
8. Projective coördinate of a point on a line 8
9. Change of coördinates 9
10. Coördinate of a line of a pencil 11
11. Coördinate of a plane of a pencil 12

CHAPTER III. PROJECTIVITY

12. The linear transformation 13
13. The cross ratio 16
14. Harmonic sets 18
15. Projection 20
16. Perspective figures 21
17. Other one-dimensional extents 23

PART II. TWO-DIMENSIONAL GEOMETRY

CHAPTER IV. POINT AND LINE COÖRDINATES IN A PLANE

18. Homogeneous Cartesian point coördinates 27
19. The straight line 27
20. The circle points at infinity 30
21. The conic 32
22. Trilinear point coördinates 34
23. Points on a line 35

SECTION PAGE
24. The linear equation in point coördinates 36
25. Lines of a pencil 37
26. Line coördinates in a plane 38
27. Pencil of lines and the linear equation in line coördinates . . . 39
28. Dualistic relations 40
29. Change of coördinates 41
30. Certain straight-line configurations 44
31. Curves in point coördinates 50
32. Curves in line coördinates 53

CHAPTER V. CURVES OF SECOND ORDER AND SECOND CLASS

33. Singular points of a curve of second order 58
34. Poles and polars with respect to a curve of second order 59
35. Classification of curves of second order 65
36. Singular lines of a curve of second class 67
37. Classification of curves of second class 68
38. Poles and polars with respect to a curve of second class 70
39. Projective properties of conics 72

CHAPTER VI. LINEAR TRANSFORMATIONS

40. Collineations 78
41. Types of nonsingular collineations 83
42. Correlations 88
43. Pairs of conics 95
44. The projective group 100
45. The metrical group 101
46. Angle and the circle points at infinity 105

CHAPTER VII. PROJECTIVE MEASUREMENT

47. General principles 107
48. The hyperbolic case 110
49. The elliptic case 115
50. The parabolic case 117

CHAPTER VIII. CONTACT TRANSFORMATIONS IN THE PLANE

51. Point-point transformations 120
52. Quadric inversion 121
53. Inversion 124
54. Point-curve transformations 127
55. The pedal transformation 131
56. The line element 133

CHAPTER IX. TETRACYCLICAL COÖRDINATES

SECTION PAGE
57. Special tetracyclical coördinates 138
58. Distance between two points 139
59. The circle 140
60. Relation between tetracyclical and Cartesian coördinates 142
61. Orthogonal circles 144
62. Pencils of circles 146
63. The general tetracyclical coördinates 150
64. Orthogonal coördinates 153
65. The linear transformation 154
66. The metrical transformation 155
67. Inversion 156
68. The linear group 159
69. Duals of tetracyclical coördinates 161

CHAPTER X. A SPECIAL SYSTEM OF COÖRDINATES

70. The coördinate system 164
71. The straight line and the equilateral hyperbola 166
72. The bilinear equation 167
73. The bilinear transformation 169

PART III. THREE-DIMENSIONAL GEOMETRY

CHAPTER XI. CIRCLE COÖRDINATES

74. Elementary circle coördinates 171
75. The quadratic circle complex 173
76. Higher circle coördinates 177

CHAPTER XII. POINT AND PLANE COÖRDINATES

77. Cartesian point coördinates 180
78. Distance 181
79. The straight line 182
80. The plane 185
81. Direction and angle 188
82. Quadriplanar point coördinates 193
83. Straight line and plane 194
84. Plane coördinates 197
85. One-dimensional extents of points 200
86. Locus of an equation in point coördinates 205
87. One-dimensional extents of planes 210
88. Locus of an equation in plane coördinates 215
89. Change of coördinates 218

CHAPTER XIII. SURFACES OF SECOND ORDER AND OF SECOND CLASS

SECTION PAGE
90. Surfaces of second order 220
91. Singular points 221
92. Poles and polars 222
93. Classification of surfaces of second order 224
94. Surfaces of second order in Cartesian coördinates 227
95. Surfaces of second order referred to rectangular axes 229
96. Rulings on surfaces of second order 232
97. Surfaces of second class 235
98. Poles and polars 238
99. Classification of surfaces of the second class 238

CHAPTER XIV. TRANSFORMATIONS

100. Collineations 240
101. Types of nonsingular collineations 241
102. Correlations 246
103. The projective and the metrical groups 249
104. Projective geometry on a quadric surface 250
105. Projective measurement 253
106. Clifford parallels 255
107. Contact transformations 258
108. Point-point transformations 260
109. Point-surface transformations 262
110. Point-curve transformations 263

CHAPTER XV. THE SPHERE IN CARTESIAN COÖRDINATES

111. Pencils of spheres 266
112. Bundles of spheres 268
113. Complexes of spheres 269
114. Inversion 270
115. Dupin's cyclide 274
116. Cyclides 279

CHAPTER XVI. PENTASPHERICAL COÖRDINATES

117. Specialized coördinates 282
118. The sphere 284
119. Angle between spheres 286
120. The power of a point with respect to a sphere 287
121. General orthogonal coördinates 288

SECTION PAGE
122. The linear transformation 291
123. Relation between pentaspherical and Cartesian coördinates . . 293
124. Pencils, bundles, and complexes of spheres 293
125. Tangent circles and spheres 295
126. Cyclides in pentaspherical coördinates 297

PART IV. GEOMETRY OF FOUR AND HIGHER DIMENSIONS

CHAPTER XVII. LINE COÖRDINATES IN THREE-DIMENSIONAL SPACE

127. The Plücker coördinates 301
128. Dualistic definition 303
129. Intersecting lines 304
130. General line coördinates 305
131. Pencils and bundles of lines 306
132. Complexes, congruences, series 308
133. The linear line complex 310
134. Conjugate lines 314
135. Complexes in point coördinates 316
136. Complexes in Cartesian coördinates 317
137. The bilinear equation in point coördinates 321
138. The linear line congruence 322
139. The cylindroid 323
140. The linear line series 324
141. The quadratic line complex 328
142. Singular surface of the quadratic complex 331
143. Plücker's complex surfaces 334
144. The (2, 2) congruence 335
145. Line congruences in general 336

CHAPTER XVIII. SPHERE COÖRDINATES

146. Elementary sphere coördinates 341
147. Higher sphere coördinates 343
148. Angle between spheres 344
149. The linear complex of oriented spheres 346
150. Linear congruence of oriented spheres 348
151. Linear series of oriented spheres 349
152. Pencils and bundles of tangent spheres 350
153. Quadratic complex of oriented spheres 353
154. Duality of line and sphere geometry 357

SECTION PAGE
122. The linear transformation 291
123. Relation between pentaspherical and Cartesian coördinates . . 293
124. Pencils, bundles, and complexes of spheres 293
125. Tangent circles and spheres 295
126. Cyclides in pentaspherical coördinates 297

PART IV. GEOMETRY OF FOUR AND HIGHER DIMENSIONS

CHAPTER XVII. LINE COÖRDINATES IN THREE-DIMENSIONAL SPACE

127. The Plücker coördinates 301
128. Dualistic definition 303
129. Intersecting lines 304
130. General line coördinates 305
131. Pencils and bundles of lines 306
132. Complexes, congruences, series 308
133. The linear line complex 310
134. Conjugate lines 314
135. Complexes in point coördinates 316
136. Complexes in Cartesian coördinates 317
137. The bilinear equation in point coördinates 321
138. The linear line congruence 322
139. The cylindroid 323
140. The linear line series 324
141. The quadratic line complex 328
142. Singular surface of the quadratic complex 331
143. Plücker's complex surfaces 334
144. The (2, 2) congruence 335
145. Line congruences in general 336

CHAPTER XVIII. SPHERE COÖRDINATES

146. Elementary sphere coördinates 341
147. Higher sphere coördinates 343
148. Angle between spheres 344
149. The linear complex of oriented spheres 346
150. Linear congruence of oriented spheres 348
151. Linear series of oriented spheres 349
152. Pencils and bundles of tangent spheres 350
153. Quadratic complex of oriented spheres 353
154. Duality of line and sphere geometry 357

< !"#$! ! !% & ' ! E % ' - 9 2 4
> / 6 8 < 0-

!"#$! ! !% & ' ! E % ' - 9 2 4 > /
6 8 < -24

!"#$! ! !% & ' ! E 2 4 (> /(6 9 0-
%9 6 0 '
D 4 ' ' -'- '

E !"#$! ! !% & ' ! E % ' - 9 2 4
> / 66 ') * 8 < 0-

!"#$! ! !% & ' ! E % ' - 9 2 4 > /
66 ') * 8 < -24

!"#$! ! !% & ' ! E 2 4 (> /(66 9 0-

!"#$! ! !% & ' ! E 6 ((2 4 (> /
D 0-
V " " ' / ' - / . '

!"#$! ! !% & ' ! E 6 ((2 4 (> /
+ 0-
V 0'&

I& =8 . 2 &' 0' ' & '

Thus the pole of perspective is at infinity in the direction perpendicular to x. The perspective relation between the circle and ellipse figures is equivalent to a stretch (*cf.* Art. 13), the stretch-ratio being $\cos\theta$.

Also because the projection is cylindrical the lines at infinity correspond: therefore their poles, that is, the centres, correspond. It follows that conjugate diameters of the circle project into conjugate diameters of the ellipse: in particular x and the perpendicular diameter of the circle, being conjugate diameters, project into perpendicular conjugate diameters of the ellipse, since a perpendicular to x remains perpendicular to x. These give the axes of the ellipse. If a be the radius of the circle, a and $a\cos\theta$ are the major and minor semi-axes of the ellipse.

Conversely an ellipse of semi-axes a and b $(a > b)$ can be obtained in this manner by projecting a circle of radius a orthogonally upon a plane making with the plane of the circle an angle

$$\theta = \cos^{-1}\left(\frac{b}{a}\right).$$

An ellipse being completely given by its principal axes (*see* Art. 61), it follows that every ellipse can be obtained in this way from a circle on its major axis as diameter.

The circle on the major axis of an ellipse as diameter is called its **auxiliary circle.** Thus **the ellipse and its auxiliary circle are derivable one from the other by a stretch parallel to the minor axis.**

EXAMPLES

1. Show that, if P, P' be points where a perpendicular to the major axis of an ellipse meets the curve and the auxiliary circle respectively, the tangents at P, P' meet on the major axis.

2. Prove that an ellipse can be obtained from the circle on its minor axis as diameter by a stretch parallel to the major axis.

79. The conjugate parallelogram. A **conjugate parallelogram** is one whose sides are the tangents at the extremities of two conjugate diameters. Clearly no real conjugate parallelogram can exist, except in the case of the ellipse.

Consider the ellipse as defined by the stretch of its auxiliary circle. Since by Art. 78 conjugate diameters correspond to conjugate diameters and parallel lines to parallel lines (because the vanishing lines are at infinity) it follows that a conjugate parallelogram for the ellipse corresponds to a conjugate parallelogram for the circle.

But a conjugate parallelogram for a circle is a circumscribed square, because conjugate diameters of a circle are at right angles.

Now, in any stretch, corresponding areas are to one another in the stretch ratio. For consider an elementary parallelogram $PQRS$ (Fig. 32) of which the sides PQ, RS are parallel to the stretch axis and the sides PS, QR are parallel to the direction of stretch. Let these meet the stretch axis at X, Y respectively. $PQRS$ transforms into a parallelogram $P'Q'R'S'$ in which $P'Q'$, $R'S'$ are parallel to the stretch axis. For if λ be the stretch ratio

$$P'X = \lambda . PX = \lambda . QY = Q'Y \quad \text{and} \quad S'Y = R'Y = \lambda . SZ.$$

Hence $$P'S' = \lambda(XS - XP) = \lambda . PS$$

and parallelogram $P'Q'R'S$: parallelogram $PQRS = S'P' : SP = \lambda$.

Breaking up any area into such elementary parallelograms and adding we see that λ = ratio of two corresponding areas.

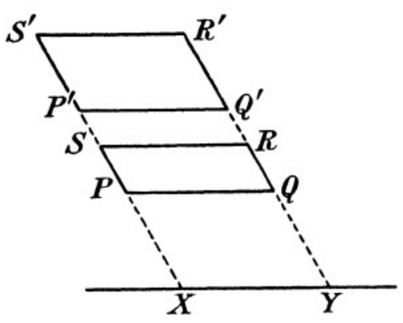

Fig. 32.

Hence area of any conjugate parallelogram of the ellipse

$= \frac{b}{a} \times$ area of corresponding circumscribed square of the auxiliary circle

$= \frac{b}{a} . 4a^2 = 4ab.$

Thus **the conjugate parallelogram of an ellipse is of constant area.** Calling p the perpendicular from the centre on any given tangent, d the length of the semi-diameter parallel to this tangent, the area of the conjugate parallelogram of which the given tangent is a side is clearly $4pd$, for $2d$ is the base of the parallelogram and $2p$ is the height. $\therefore 4pd = 4ab$, *i.e.* $pd = ab$.

Examples

1. Show that the diagonals of a conjugate parallelogram are themselves conjugate diameters.

2. Show how to construct geometrically the equal conjugate diameters of an ellipse, given its axes.

3. Show that the area of the ellipse is πab.

80. Sum of squares of two conjugate semi-diameters. Let CP', CQ' (Fig. 33) be two perpendicular semi-diameters of the auxiliary circle, CP, CQ the corresponding diameters of the ellipse.

Then CP, CQ are conjugate. Also $P'PM$, $Q'QN$ are perpendicular to the major axis.

Then from the stretch property, if ϕ be the angle ACP',

$$PM = \frac{b}{a} P'M = \frac{b}{a} a \sin \phi = b \sin \phi ;$$

$$QN = \frac{b}{a} Q'N = \frac{b}{a} a \cos \phi = b \cos \phi .$$

$$CM = a \cos \phi ; \; CN = a \sin \phi ;$$

$$\begin{aligned} CP^2 + CQ^2 &= CM^2 + PM^2 + QN^2 + CN^2 \\ &= a^2 \cos^2 \phi + b^2 \sin^2 \phi + b^2 \cos^2 \phi + a^2 \sin^2 \phi \\ &= a^2 + b^2. \end{aligned}$$

Hence **the sum of the squares of two conjugate diameters of an ellipse is constant.**

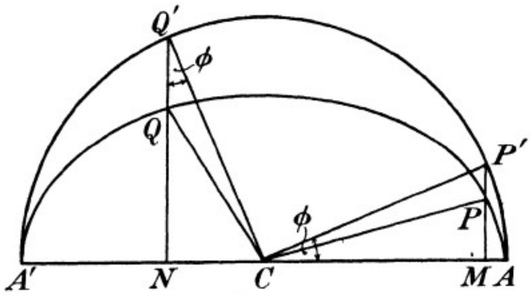

FIG. 33.

81. Pseudo-conjugate parallelogram. Let ACA' (Fig. 34) be a diameter of a hyperbola meeting the curve at real points A, A'. Let the tangent at A meet the asymptotes at D, E and the tangent at A' meet them at F, G. Then, because the tangents at A, A' are parallel, $DEFG$ is a parallelogram of which the asymptotes are diagonals.

If P be a point on the curve and PN the chord through P conjugate to ACA' meeting ACA' at N,

$$\frac{PN^2}{AN.NA'} = -\frac{CB_1^2}{CA^2},$$

CB_1 being the absolute length (*see* Art. 76) of the diameter conjugate to ACA'.

But $$\frac{PN^2}{AN.NA'} = -\left(\frac{AN}{A'N}\right)\left(\frac{PN}{AN}\right)^2.$$

If P moves off to infinity on the hyperbola in the direction of the asymptote CD, AP becomes parallel to CD. The triangles CAD, ANP become similar, and $\left(\frac{PN}{AN}\right)^2$ becomes equal to $\left(\frac{AD}{CA}\right)^2$; also $\frac{AN}{A'N}$ approaches unity.

Hence $$\frac{CB_1^2}{CA^2} = \frac{AD^2}{CA^2},$$

or $$CB_1 = AD.$$

Thus the intercept of a tangent between the asymptotes measures

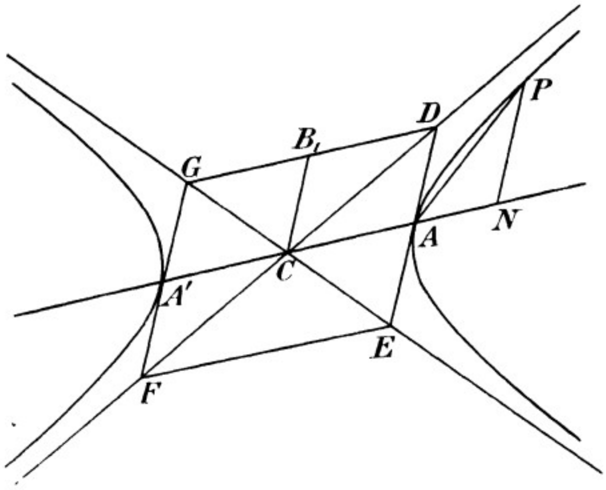

FIG. 34.

the absolute length of the parallel diameter. The parallelogram $DEFG$ is therefore a **pseudo-conjugate** parallelogram. Its median lines are conjugate diameters, but only one pair of sides touches the curve.

We have seen (Art. 67) that the area of the triangle ECD cut off from the asymptotes by any tangent to the curve is constant. **The area of the pseudo-conjugate parallelogram** $DEFG$ is four times the area of the triangle ECD and **is therefore also constant.**

If p = perpendicular from C on tangent at A, $d = CB_1$ = absolute length of diameter conjugate to CA, $DE = 2d$, and area of triangle $ECD = \frac{1}{2}p \,.\, 2d = pd$. Hence **in the hyperbola as in the ellipse** pd = **constant.** Taking the case where the sides of the pseudo-conjugate parallelogram are parallel to the axes, the constant $= ab$ where a is the semi-transverse axis and b is the absolute length of the conjugate semi-axis.

Further, in Fig. 34, the area of the triangle GCD is equal to that of the triangle CDE, and therefore constant. Hence GD touches at its middle point B_1 a second hyperbola having the same asymptotes. This second hyperbola, which is also the locus of B_1, is termed the **conjugate hyperbola** of the first one; the transverse and conjugate axes are interchanged, in position and absolute length, when we pass from a hyperbola to its conjugate hyperbola.

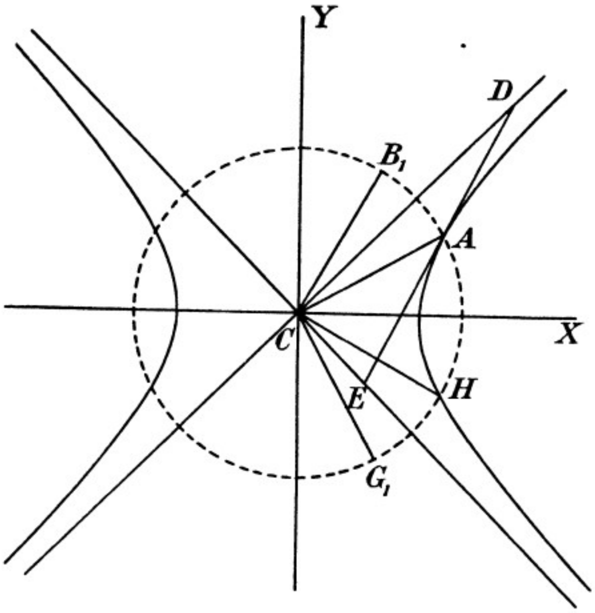

FIG. 35.

82. Difference of squares of absolute lengths of conjugate semi-diameters of a hyperbola. From the triangle CED (Fig. 35) we have, since CA is the median,

$$CD^2 + CE^2 = 2(CA^2 + AD^2)$$

by a well-known relation.

Also $\quad DE^2 = CE^2 + CD^2 - 2CE.CD \cos ECD,$

that is, $\quad 4AD^2 = 2CA^2 + 2AD^2 - 2CE.CD \cos ECD,$

or $\quad CA^2 - AD^2 = CE.CD \cos ECD.$

I& 8 < 7" .D# W6 #D 6 #% ".6%G "GF 6G D# 6%G

CHAPTER XI

RECIPROCATION AND INVERSION

170. Reciprocal transformation or correlation. If in the equations (3), (4) and (5) of Art. 161 we interchange l', m' and x', y' we find, writing for shortness

$$\alpha_1 = B_2C_3 - B_3C_2,\ \alpha_2 = B_3C_1 - B_1C_3,\ \alpha_3 = B_1C_2 - B_2C_1,$$

with corresponding meanings for β's and γ's :

$$l' = \frac{A_1x + B_1y + C_1}{A_3x + B_3y + C_3}, \qquad m' = \frac{A_2x + B_2y + C_2}{A_3x + B_3y + C_3},$$

$$l = \frac{A_1x' + A_2y' + A_3}{C_1x' + C_2y' + C_3}, \qquad m = \frac{B_1x' + B_2y' + B_3}{C_1x' + C_2y' + C_3},$$

$$x = \frac{\alpha_1l' + \alpha_2m' + \alpha_3}{\gamma_1l' + \gamma_2m' + \gamma_3}, \qquad y = \frac{\beta_1l' + \beta_2m' + \beta_3}{\gamma_1l' + \gamma_2m' + \gamma_3},$$

$$x' = \frac{\alpha_1l + \beta_1m + \gamma_1}{\alpha_3l + \beta_3m + \gamma_3}, \qquad y' = \frac{\alpha_2l + \beta_2m + \gamma_2}{\alpha_3l + \beta_3m + \gamma_3}.$$

These equations may be shown as in Art. 161 to be the necessary equations of transformation in any one-one algebraic correspondence of plane fields in which lines correspond to points and points to lines. Clearly any pencil is homographic, and therefore equianharmonic, with the corresponding range. This transformation is therefore of the type discussed in Art. 56.

It is, however, much more general than this transformation; for the transformation by reciprocal polars is limited to fields in the same plane, whereas the present transformation is for any plane fields. Also in the transformation by reciprocal polars the same line p corresponds to the same point P whether P be considered as belonging to one field or to the other. Whereas here, if the fields be taken coplanar and the axes of co-ordinates identical, if we put $x' = x$, $y' = y$ we do not in general obtain $l' = l$ or $m' = m$.

The present transformation is the most general case of a plane reciprocal transformation.

An obvious modification of the reasoning of Art. 162 will show that such a transformation is determined when four points A_1, B_1, C_1, D_1 of one field, no three of which are collinear, are made to correspond to four lines a_2, b_2, c_2, d_2 of the other field, no three of which are concurrent.

For if P_1 correspond to p_2 we have

$$A_1\{B_1C_1D_1P_1\}=a_2\{b_2c_2d_2p_2\},$$
$$B_1\{A_1C_1D_1P_1\}=b_2\{a_2c_2d_2p_2\},$$

which determine a_2p_2 and b_2p_2, and therefore p_2.

And it is easy to show that if P_1 describes a straight line, p_2 passes through a point.

Two such fields will be said to be **reciprocal** or **correlative.** The relation between them may be spoken of as a **reciprocity** or a **correlation.**

171. Two reciprocal transformations are equivalent to a projective transformation. Consider two reciprocal plane fields ϕ_1, ϕ_2, and a third field ϕ_3 reciprocal with ϕ_2. ϕ_1 and ϕ_3 now correspond point by point and line by line and since the correspondence between elements of ϕ_1 and ϕ_2 is one-one and algebraic, and that between elements of ϕ_2 and ϕ_3 is one-one and algebraic, the correspondence between elements of ϕ_1 and ϕ_3 is also one-one and algebraic.

Accordingly the fields ϕ_1 and ϕ_3 are homographic and therefore projective.

172. Any reciprocal transformation is equivalent to a projective transformation and a transformation by reciprocal polars. For let ϕ_1 and ϕ_2 be given reciprocal fields. Let ϕ_3 be the reciprocal polar field of ϕ_2 with regard to any conic in its own plane. Then by the last Article ϕ_1 and ϕ_3 are projective. Thus a projective transformation transforms ϕ_1 to ϕ_3 and the transformation by reciprocal polars transforms ϕ_3 to ϕ_2.

173. Locus of incident points and envelope of incident lines of two coplanar reciprocal fields. If two reciprocal fields ϕ, ϕ' be coplanar, we proceed to find the condition that a point and its corresponding line shall be incident.

If P be a point on its corresponding line p', P being considered as belonging to ϕ, it also lies on its corresponding line when considered as belonging to ϕ'. For let $P=Q'$. Then, since P, *i.e.* Q', lies on p', q passes through P, *i.e.* through Q'. But it should be noted that q is, in general, distinct from p'.

In like manner, if a line $p = q'$ passes through P' it also passes through Q, but P', Q are, in general, distinct.

Such points and lines will be called **incident points and lines** of the correlation.

Consider now an incident point A of ϕ, assuming for the moment that such exist. Let a' be its corresponding line which is then an incident line, and on a' take any point B distinct from A. If B is incident, b' passes through B, and, since A is distinct from B, a' must be distinct from b', otherwise the correspondence would not be one-one.

But AB, which is identical with a', corresponds to $a'b'$ in ϕ', which is identical with B. If then there were two such points B on a', incident and distinct from A, they would correspond to the same line a' in ϕ, which is not possible. Thus there can only be two points on a' which are incident, namely A and one point B, so that it is impossible for an incident line to contain more than two incident points. Thus all the points of the plane cannot be incident points, nor all the lines incident lines.

Let now l be any line whatever, P any point of it, p' the line corresponding in ϕ' to P, P' the meet of l and p'. If to P' in ϕ' corresponds p in ϕ, then since P' lies on p', p passes through P. Each of the points P, P' is thus uniquely determined as soon as the other is known. The ranges $[P]$, $[P']$ are therefore homographic and have in general two self-corresponding points S, T, which may be real coincident or imaginary, and which are such that they lie on their corresponding lines and so are incident points.

The locus k_1 of incident points has thus two intersections with any straight line in the plane. Hence it is a conic by Art. 159.

In like manner through any point M two incident lines can be drawn. The envelope of incident lines is therefore a conic k_2.

The conics k_1, k_2 correspond doubly in the correlation, since a tangent to k_2, being an incident line, corresponds in either field to an incident point, which lies on k_1.

The double character of this correspondence does not, however, extend to the individual elements of the conics.

A point $P = Q'$ (Fig. 54) of k_1 corresponds, in general, to two different incident lines p', q, which are the two tangents from the point to k_2. The tangent $t = u'$ to k_1 at $P = Q'$ corresponds to the point T' of ϕ' where p' touches k_2; and to it corresponds in ϕ the point U where q touches k_2.

In like manner a tangent $p = q'$ to k_2 corresponds, in general, to two different incident points P', Q, which are the intersections of this tangent with k_1. The point of contact $T = U'$ of $p = q'$ with k_2 corresponds to the line t' of ϕ' which touches k_1 at P'; and to it corresponds in ϕ the tangent u to k_1 at Q. It should be noticed that we can here use the same letters without confusion, because p has no necessary relation to p' or T to T', the corresponding elements being p, P'; P, p'; T, t'; t, T'.

If $P = Q'$ comes into coincidence with an intersection X of k_1, k_2, the lines p', q coincide with the tangent x_2 to k_2 at X; thus X corresponds doubly to x_2. But, further, T' and U then both coincide with X, and they correspond to the tangent x_1 to k_1 at $P = Q' = X$; thus X corresponds doubly to x_1. Since the same point cannot correspond to two distinct lines, x_1 must be identical with x_2, and the conics k_1, k_2 touch at any intersection X.

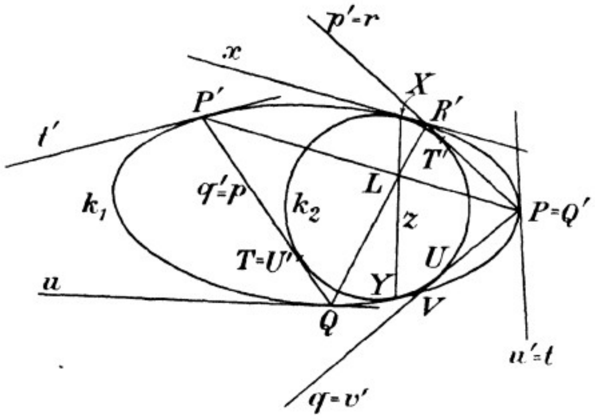

FIG. 54.

Thus k_1, k_2 either touch at two points X, Y, or else have four-point contact at a point X.

As a special case k_1 may break up into a line-pair, in which case k_2 breaks up into a point-pair. Neither the points of the pair nor the lines of the pair can be incident points or lines, so the point-pair cannot lie upon the line-pair. Each point of k_2 corresponds to the two components of k_1, one in each field.

If the given correlation reduces to a transformation by reciprocal polars, the two conics k_1, k_2 coincide with the base conic, since every point of the base conic lies on its polar and every tangent of this conic passes through its pole. By analogy the conics k_1, k_2 will be termed the **base conics** in the general case.

174. Correspondence of points and tangents on the base conics. In Fig. 54 it is clear that some rule of selection must be given, which enables us to discriminate which end of a chord of k_1 tangent to k_2 is to be treated as P' and which as Q. For the correspondence is not here a double one; thus in Fig. 54 two such chords are p' and q, their intersection being $P=Q'$. If their other extremities are R', V, it is clear that R' belongs to $p'=r$ and V to $q=v'$. Thus if, in this relation, P corresponds to R', the same point, treated as Q' does not correspond to R' but to a different point V.

Since, in the correlation, homographic ranges and pencils of the second order correspond to homographic pencils and ranges respectively, we have at once

$$[P']^2 \barwedge [p]^2 \equiv [q']^2 \barwedge [Q]^2,$$

so that the ranges $[P']^2$ and $[Q]^2$ are homographic ranges of the second order on the conic k_1. They are not, however, in involution.

But, if P' approaches X, $P'Q$ approaches the tangent at X and Q approaches X. Hence X is a self-corresponding point of the ranges $[P']^2$, $[Q]^2$, and so is Y. Thus the common chord XY of contact of k_1, k_2 is the cross axis of these ranges (Art. 86), and cross-joins such as QR', $P'P$ meet at a point L of XY.

The above will still apply if the base conics have four-point contact at X, since the cross-axis of the ranges is still determinate, being the common tangent at X.

This shows that, once the discrimination has been effected in the case of a single pair of points on k_1, the process of selection is determinate, for, by varying L and keeping P, R' fixed, we can make P', Q represent any required pair of corresponding points and identify the field to which each belongs. Once the points have been identified, the tangents are also identified without ambiguity.

In the first instance, however, we have nothing to guide us in the identification of the original pair of points. But a little consideration will readily show that the choice is really immaterial. For it amounts to interchanging the fields ϕ and ϕ' consistently throughout. This does not affect the base conics since they correspond doubly. The correlation remains the same, the notation only being altered.

175. A correlation is uniquely determined by the base conics. Suppose the base conics k_1, k_2 of a correlation are given, and also the selection rule of Art. 174. This fixes the order of the extremities of two chords such as $P'Q$, $R'S$ (Fig. 55), with the proviso that, if one pair be interchanged, so also must the other pair. Let $A = B'$ be any point of the plane and let the two tangents from this point to k_2 be $p = q'$, $r = s'$, meeting k_1 at P', Q; R', S, respectively.

Now, in ϕ, $A = pr$ and corresponds to $P'R' = a'$ in ϕ'. Similarly $B' = q's'$ in ϕ' and corresponds to $QS = b$ in ϕ. We have thus the two lines which correspond to $A = B'$ when treated as a point of either field.

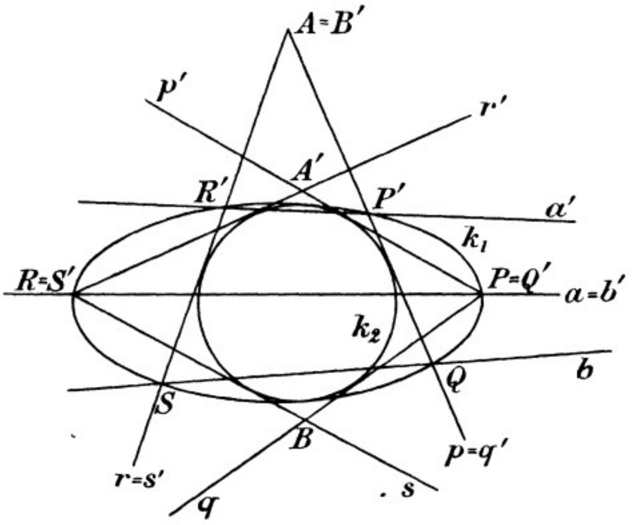

FIG. 55.

Again, let $a = b'$ be any given line, meeting s_1 at $P = Q'$ and $R = S'$ (Fig. 55). From these latter points draw tangents p', q and r', s to k_2, the identification following the selection rule. Then $p'r' = A'$, $qs = B$, so that we have constructed the two points corresponding to the line $a = b'$, treated as a line of either field.

Thus the base conics, together with the selection rule, fix the correlation completely and uniquely.

Further, reversal of the selection rule interchanges P' and Q, R' and S, a' and b, A' and B. This simply amounts to interchanging

ϕ and ϕ', as pointed out in the last Article. Thus the correlation is really determined by the base conics alone.

It is to be noted that, although the constructions here given fix the correlation theoretically, as drawing-board methods they are of limited application, since, in general, imaginary elements may enter into them.

The above argument has assumed that the conics k_1 and k_2 were already known to be the base conics of a correlation. If, however, k_1 and k_2 are two conics having either double or four-point contact and arbitrarily given, we may still, provided we definitely allot one of the extremities of a chord of k_1 tangent to k_2 to each the two fields in one particular case and so fix the homographic ranges on k_1, apply the above construction to find the line corresponding to any given point of the plane. In order to show that we then obtain a correlation, it is necessary and sufficient to prove that, if a point A describes a straight line l, its corresponding line a' passes through a fixed point L'.

In such a case the tangents $p=q'$, $r=s'$ (Fig. 55) form an involution of tangents to k_2, of which l is the axis. Since there is a unique correspondence between p and P', r and R', the ranges $[P']^2$, $[R']^2$ are homographic; moreover, if P' is taken at R', p and r are interchanged and R' is at P'. Thus P' and R' correspond doubly and so are mates in an involution on k_1. Hence $P'R'=a'$ passes through a fixed point L', which corresponds in ϕ' to l in ϕ. Similarly $QS=b$ passes through a fixed point which corresponds in ϕ to l in ϕ'. Thus the points of a line correspond to the lines through a point and the relation between the fields is a correlation.

If now the point A lies on a', it must coincide with either P' or R' and so lie on k_1; thus k_1 is the locus of incident points of the correlation. Similarly if a contains A', it must coincide with either p' or r' and so be a tangent to k_2; thus k_2 is the envelope of incident lines of the correlation.

No distinction need be drawn between the cases where the points of contact of k_1, k_2 are separate or coincident. Thus any two conics having either double or four-point contact determine a unique correlation, provided that (1) we assign which is the incident locus k_1 and which the incident envelope k_2; (2) we assign the intersections with k_1 of one tangent to k_2 each to its proper field. By varying the above assignments we obtain four possible correlations from two given conics.

EXAMPLES

1. Show that, in the case where k_1, k_2 degenerate into a line-pair and point-pair respectively, the join of the point-pair passes through the meet of the line-pair, and the points of contact X, Y of the general case coincide at this point.

2. Give a construction for the line and point corresponding to a given point and line, when the base conics are a line-pair and point-pair.

176. The self-corresponding triangle. Fig. 55 gives us at once the condition that a point $A = B'$ shall have the same corresponding line in the two fields. For if $a' = b$, then $P' = Q$ and $R' = S$ and the lines $p = q'$, $r = s'$ are tangents to k_1. But they are also tangents to k_2 and so are common tangents to the base conics. There are only two such tangents, namely, those at X, Y (Fig. 54). Moreover, if $P' = Q$ approaches X, $P'Q$ or $p = q'$ approaches the tangent at X. So that the tangent at X is the line corresponding to X in either field. Similarly the tangent at Y is the line corresponding to Y in either field. Calling these tangents x, y, and denoting XY by z, and its pole with regard to either k_1 or k_2 by Z, we have X, Y, Z corresponding to x, y, z in either field. Thus the triangle XYZ corresponds to itself. It should, however, be noted that only one vertex, namely Z, corresponds to the opposite side, the others being incident with their corresponding lines.

177. Reciprocation with respect to a circle. We have already dealt with polar reciprocal figures in Art. 56. There is, however, one special case of polar reciprocation with respect to a conic which is particularly important from the point of view of obtaining what are known as *metrical* properties, that is relations between magnitudes of lengths and angles, especially the latter. This special case arises when the base conic is a circle.

The centre O of the circle with respect to which reciprocal polars are taken is called the **origin of reciprocation** and its radius the **radius of reciprocation.**

Any point P then corresponds to its polar p' with respect to the base circle, and if p' meet OP at N', then $ON'.OP = k^2$, where k is the radius of reciprocation (Art. 54). Similarly a straight line p corresponds to its pole P', so that if N be the foot of the perpendicular from O upon p, P' lies on ON and $ON.OP' = k^2$.

A change in the radius of reciprocation has merely the effect of altering the *scale* of the reciprocal figure. Thus, if a figure ϕ transforms into ϕ' when the radius of reciprocation is k_1 and into ϕ'' when the radius of reciprocation is k_2, the line p transforms in the first case to P' on ON where $OP'.ON = k_1^2$ and in the second case to P'' also on ON, where $OP''.ON = k_2^2$. Thus $OP' : OP'' = k_1^2 : k_2^2$ and the figures ϕ', ϕ'' are similar and similarly situated, O being the centre of similarity, and $k_1^2 : k_2^2$ the ratio of similarity. Hence results independent of the scale of the figure are independent of the choice of the radius of reciprocation and the latter may often be chosen so as to have any convenient value, or be left altogether unspecified.

When this last is the case, the reciprocation is specified by the origin O and is often described briefly as reciprocation with respect to O.

Note also that it is not essential that k^2 should be *positive*: if k^2 is negative the construction still leads to a real reciprocal figure, only in this case a line and its reciprocal are on opposite sides of the origin.

The fundamental property of reciprocation with respect to an origin O is the following.

If a, b be any two lines, A', B' their reciprocal points, then OA', OB' are perpendicular to a, b respectively.

Since OA', OB' are drawn either both towards or both away from a and b, $A'OB'$ is *equal* to that angle between a, b which does not include O and supplementary to that which does include O.

Thus: **the angle between any two lines is equal (or supplementary) to the angle subtended by their reciprocal points at the origin of reciprocation.**

Other important properties relate to points at infinity. Thus a point P^∞ reciprocates into the line through O at right angles to OP^∞. The line at infinity reciprocates into O itself.

A circular point at infinity Ω therefore reciprocates into a straight line through O perpendicular to $O\Omega$. But $O\Omega$, being a double ray of a rectangular involution (Art. 141) is at right angles to itself. Thus Ω reciprocates into $O\Omega$ and similarly Ω' reciprocates into $O\Omega'$.

178. Reciprocation of a circle with respect to an excentric origin. Let C (Fig. 56) be the centre of the given circle, O the origin of reciprocation, AB the diameter of the given circle passing through O. The reciprocal curve of the circle will be a conic (Art. 56). Also, by symmetry, one of the axes of this conic will be along AB.

Since Ω, Ω' lie on the circle, $O\Omega$, $O\Omega'$ are tangents to the conic. Hence O is a real focus of the conic, so that AB is the focal axis.

Now C is the pole of the line at infinity with respect to the circle. Hence in the conic the reciprocal c' of C is the polar of O with respect to the conic, that is, the directrix corresponding to O.

Again, the centre C' of the conic is the pole of the line at infinity with respect to the conic, and so is the reciprocal of the polar c of O with respect to the circle. The second focus H is the point such that C' bisects OH.

The tangents from H to the conic are circular lines $H\Omega$, $H\Omega'$ which reciprocate into points I, J of the circle, where OI, OJ are perpendicular, and therefore, by the property of circular lines, also parallel, to $H\Omega$, $H\Omega'$ respectively. Accordingly OI, OJ pass through Ω, Ω' respectively, so that IJ and the line at infinity $\Omega\Omega'$ are opposite common chords of the given circle and the line-pair $(O\Omega, O\Omega')$, which latter is (Art. 140) identical with the point-circle O. IJ is therefore the radical axis of the given circle and O, and so is a real line h' at right angles to OC. Thus H is the reciprocal of this radical axis, a property which we shall require later.

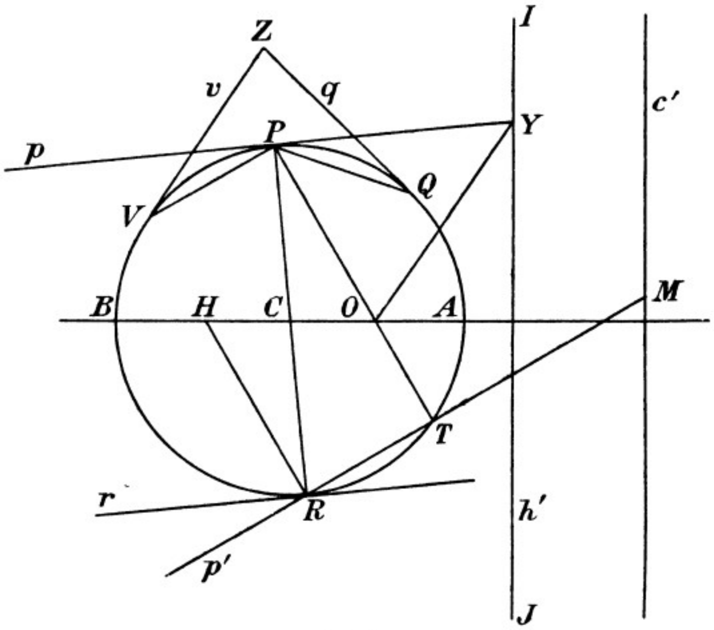

FIG. 56.

To find the eccentricity of the conic, reciprocate the tangents a, b at A, B. This gives the vertices A', B' of the focal axis.

Then $OA' = k^2/OA$, $B'O = k^2/BO$, leading to :

$$\text{major axis} = B'O + OA' = k^2\left(\frac{1}{OA} + \frac{1}{BO}\right);$$

$$\text{distance between foci} = OA' + OB' = k^2\left(\frac{1}{OA} - \frac{1}{BO}\right).$$

By division, eccentricity $= \dfrac{BO - OA}{BO + OA} = \dfrac{CO}{CA}$ numerically.

The conic is therefore an ellipse, parabola or hyperbola according as O lies inside, on, or outside the circle.

A very important particular case is that actually shown in Fig. 56. This is when the radius of reciprocation is so chosen that $A' = B$ and $B' = A$. This requires $k^2 = OA.OB$, so that for the elliptic case as in the figure, k^2 is negative.

When k^2 is so chosen, the given circle is the circle on the focal axis of the conic as diameter. We see then that a conic and its auxiliary circle are polar reciprocals with respect to an origin at a focus.

Note that in this case C and C' coincide, and also c and c'.

Examples

1. Given a focus, the corresponding directrix and the eccentricity of a conic show how to construct, without drawing the curve, the two tangents from a given external point O.

Prove that the angle between the tangents from a point O to a parabola is $\cos^{-1}(q/r)$, where r is the distance of O from the focus, and q is the perpendicular distance of O from the directrix.

2. Two circles meet at A and B. Taking A as centre of reciprocation, find the reciprocal theorem of the result that AB is perpendicular to the line joining the centres of the circles.

3. Reciprocate the theorem : circles of constant radius which touch one straight line also touch a parallel line, and their centres lie on the line midway between the two lines.

4. Given a circle centre C and radius r, show how to determine the centre and radius of the circle of reciprocation so that the given circle reciprocates into a rectangular hyperbola of which the real axis is of length $2r$.

5. If the reciprocal of a circle of radius a and centre C is taken with respect to a circle of radius b and centre O, where $OC = c$, find the eccentricity and latus rectum of the conic obtained.

Conics are described with a fixed focus O, their latera recta of the same given length l, and their corresponding directrices tangent to a given conic with focus O and latus rectum l'. Prove that their envelope consists of two conics, each of which has a focus at O, and that the reciprocals of the latera recta of these conics are the sum and difference of the reciprocals of l and l'.

179. Examples of focal properties of conics deduced by reciprocation with respect to a circle. A number of the focal properties discussed in Chapter VIII are very readily deduced from simple properties of the circle by this form of reciprocation. As they give valuable examples of the method, we shall set out the proofs of a few such properties. It will be convenient to assume for k that value which reciprocates the circle into a conic of which it is the auxiliary circle, so that C is the common centre of circle and conic and the foci O, H are symmetrically situated with respect to C.

(*a*) Let v, q (Fig. 56) be any two tangents to the circle, then they are equally inclined to their chord of contact VQ.

Using the fundamental property of reciprocation with regard to a circle (Art. 177), OV', OQ' are equally inclined to $O(v'q')$, where $v'q'$ is the intersection of the tangents at V', Q' to the conic.

Hence the theorem that two tangents to a conic subtend equal or supplementary angles at a focus.

In the figure, which is drawn for the ellipse, the angles ZVQ and ZQV either both include or both exclude O, and we get the case of equal angles.

If O is outside the circle, the points of contact of tangents from O reciprocate into the asymptotes, and divide the arcs of the circle which correspond to the two branches of the curve. The reader will find the discrimination of the two cases for the hyperbola a useful exercise.

(*b*) Let PO (Fig. 56) meet the circle again at T. Then $OP.OT = OA.OB = k^2$. The line through T perpendicular to OT is then a tangent to the conic; it meets the circle again at R, and, since PTR is a right angle, PR is a diameter. Hence by symmetry HR is parallel and equal to PO, and

$$HR.OT = PO.OT = BO.OA = a^2(1-e^2) = b^2,$$

where a is the semi-major and b the semi-minor axis.

We thus obtain the two theorems of Art. 125.

(*c*) Let the tangent p at P meet IJ at Y. By a well-known property of the radical axis of two circles, the lengths of tangents from Y to the given circle and to the point circle O are equal. Thus OPY is an isosceles triangle and PY, OY are equally inclined to OP. Now Y is ph' and reciprocates into $P'H$, so that OY reciprocates into the point at infinity on $P'H$. Similarly OP reciprocates into the point at infinity on p'. Finally PY is p and reciprocates into P'. Hence by the fundamental property OP'

and the parallel through O to $P'H$ are equally inclined to the parallel through O to p', *i.e.* OP', $P'H$ are equally inclined to p' or the tangent and normal bisect the angles between the focal distances.

(*d*) If V, Q are fixed points on the circle, P a variable point, the angle VPQ is constant.

Hence, in the reciprocal figure, the angle between $O(v'p')$ and $O(q'p')$ is constant, *i.e.* the intercept on a variable tangent p' made by two fixed tangents v', q' subtends a constant angle at the focus (Art. 122).

(*e*) The tangent p to the circle at P is perpendicular to the radius vector CP. Therefore OP' is perpendicular to $O(c'p')$, that is, the part of the tangent to a conic intercepted between the point of contact and a directrix subtends a right angle at the corresponding focus (*cf.* Art. 121).

(*f*) ZV, ZQ being two tangents v, q to the circle are equally inclined to CZ. Hence OV', OQ' are equally inclined to $O\{c'(V'Q')\}$, *i.e.* if a chord $V'Q'$ meet the directrix corresponding to focus O at, say X, then OX is a bisector of the angle $V'OQ'$ (*cf.* Art. 121).

(*g*) Two parallel tangents p, r to the circle are a constant distance apart.

But perpendicular distance of O from $p = k^2/OP'$,
and perpendicular distance of O from $r = k^2/R'O$.

Hence $\dfrac{1}{OP'} + \dfrac{1}{R'O} = \text{const.}$, where $P'OR'$ is a chord of the conic through O. By taking this chord to be the latus rectum we find this constant $= 2/(\text{semi-latus rectum})$, and thus obtain the first theorem of Art. 123.

(*h*) The locus of the intersection Z of tangents v, q to a circle which cut at a given angle α is a concentric circle: this leads at once to the following.

The envelope of a chord $V'Q'$ of a conic which subtends a given angle at the focus is a conic having the same focus and directrix.

Further, in the above VQ touches another fixed circle concentric with the original circle. Hence the locus of the pole $v'q'$ of a chord $V'Q'$ subtending a constant angle at the focus of a conic is a conic having the same focus and directrix.

180. Case where O lies on the circle. When the origin of reciprocation lies on the circle, the line at infinity touches the reciprocal conic, which thus becomes a parabola.

If through O, now on the circle, two perpendicular chords OP, OQ of the circle are drawn, the join PQ passes through the centre C.

We have then : p', q' are perpendicular and $p'q'$ lies on the directrix c', *i.e.* the intersection of two perpendicular tangents to a parabola lies on the directrix (*cf.* Art. 128).

If OA is a diameter of the circle, P any point on the circle, OPA is a right angle. Therefore the reciprocal point $(a'p')$ of AP lies on OP, so that a', p', OP are concurrent. But p' is perpendicular to OP, so that (p', OP) is the foot of the perpendicular from the focus on the tangent p' and this lies on a', which is clearly the tangent at the vertex (*cf.* Art. 127).

There is a well-known theorem on the circle, known as Simson's Theorem, that if L, M, N are the feet of the perpendiculars from a point O of the circle upon the sides f, g, h of a triangle inscribed in the circle, then L, M, N are collinear.

Reciprocating this result with regard to O, we have l' is a line through F' perpendicular to OL, that is, to OF'. Hence if, through the vertices F', G', H' of a triangle circumscribed to a parabola of which O is the focus, we draw lines perpendicular to OF', OG', OH', these are concurrent at some point U. It follows that a circle on UO as diameter passes through F', G', H' or the circumcircle of a triangle formed by three tangents to a parabola passes through the focus (Art. 130).

EXAMPLES

1. Show that the vertices of a triangle can be reciprocated into the opposite sides if the origin of reciprocation be taken at the orthocentre.

2. Reciprocate the theorem that the orthocentre of a triangle circumscribing a parabola lies on the directrix, the origin of reciprocation being the orthocentre.

181. Coaxal circles reciprocate into confocal conics. Consider a set of coaxal circles k with real limiting points O, K (Fig. 57). Then O, K are harmonically conjugate with regard to the points P, Q in which any circle of the set meets the line of centres, that is, the polars of O, K with regard to any circle of the set pass through K, O. Also by symmetry these polars are perpendicular to the line of centres.

Therefore O, K have the same polars with regard to all the circles of the coaxal system, and these polars are the perpendiculars to the line of centres through K, O respectively.

Now reciprocate the coaxal circles with regard to any circle centre O. The circles k reciprocate into conics k' having O for a focus.

O reciprocates into the line at infinity, and the line c which is the common polar of O with respect to the coaxal circles reciprocates into a point C' on the line of centres. Thus C' is the common centre of all the conics k'. Since the conics k' have O for a common focus, and C' for a common centre, their second real focus S is also common, so that they are confocal conics. S lies on the line of centres, and $OS = 2.OC'$. The reciprocal of S is a line x perpendicular to the line of centres. If it meet this line at X, then

$$OX.OS = OK.OC'.$$

But since $OS = 2.OC'$, therefore $OK = 2.OX$, or X is the middle point of OK, that is, x is the radical axis.

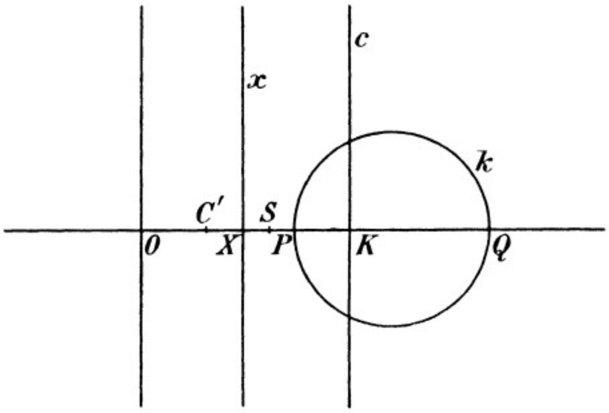

Fig. 57.

Accordingly the two foci are O and the reciprocal of the radical axis.

Conversely, it is easily shown that confocal conics reciprocate with respect to one of the real foci, into a set of coaxal circles with real limiting points.

182. Inversion. Let O be the centre of a circle c, of radius k, P any point in the plane of the circle, P' the point of OP conjugate to P with respect to c, so that

$$OP.OP' = k^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

The point P' is then said to be the **inverse point** of P with respect to c, and the symmetry of the relation (1) then shows that P is the inverse point of P' with respect to c. The relation between the field formed by the points P and that formed by the points

P' is termed **inversion,** either field being spoken of as the **inverse** of the other with respect to the circle; c, O and k are called the **circle, centre** and **radius** of **inversion** respectively.

Since the polar of P with respect to a circle is perpendicular to OP, it is clear that the inverse point P' is the foot of the perpendicular from O on the polar of P.

Although the correspondence of the points is unique, inversion is not a homography, for a straight line does not correspond to a straight line, but to a circle through the origin.

To prove this, let a (Fig. 58) be any straight line, A the foot of the perpendicular from O on a, P any other point on a. Let A', P' be the inverse points of A, P. Then $OA.OA' = OP.OP'$, or $OP' : OA' = OA : OP$.

Since the angle at O is common, the triangles AOP, $P'OA'$ are similar, and the angle at P' is equal to the angle at A, and therefore to a right angle. Therefore P' describes a circle on OA' as diameter.

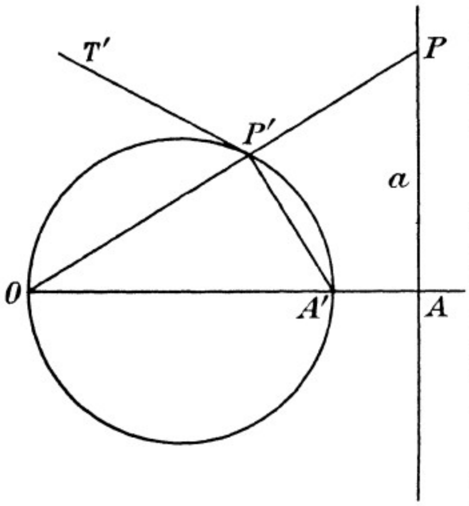

FIG. 58.

183. The inverse of a circle is, in general, a circle. Let c (Fig. 59) be any circle whose centre is C and radius a. Let P be any point of c, P' its inverse point. Let OP meet c again at Q.

Then: $$OP.OP' = k^2,$$
$$OP.OQ = OC^2 - a^2.$$

Hence $$OP'.OQ = k^2 : (OC^2 - a^2),$$

whence the locus of P' is similar and similarly situated to the locus of Q, O being the centre of similarity, and the ratio of similarity being $k^2/(OC - a^2)$.

Since the locus of Q is the original circle, the locus of P' is also a circle c', whose centre K is in the line OC. If O is outside c, a pair of common tangents to c, c' pass through it.

Thus every circle inverts into another circle, except when the first circle passes through the centre of inversion, when, as in Art. 182, it inverts into a straight line perpendicular to OC.

It should be noted that the centre of the circle c does not invert into the centre of the circle c', so that K is not C'.

It is also important to remember that a *conic* does not, in general, invert into a conic. For, consider the inverse s' of a conic s, and the intersections of s' with a straight line t'. This line t' inverts into a circle t, which has four real or imaginary intersections with the conic s. The inverse points of these four intersections are the intersections of t' and s'. These latter are thus four in number, so that s' is, in general, a quartic curve.

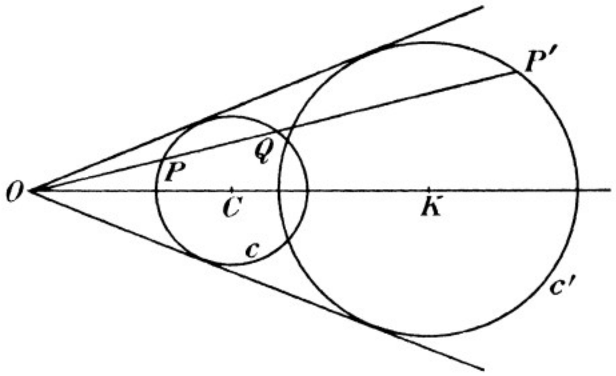

FIG. 59.

EXAMPLE

Given a point O and a circle centre C and radius r, show how to determine the radius of a circle of inversion, centre O, so that the circle, centre C, inverts into another of radius $2r$, and find the position of the centre of this inverse circle.

184. Circles through two inverse points. Let P, P' be any two inverse points with respect to a circle v of centre O and radius k, and let c be any other circle passing through P and P'.

Let Q be an intersection of v and c. Then $OQ=k$, the radius of inversion, and $OP.OP'=OQ^2$.

Hence, by a well-known property of the circle, OQ is a tangent from O to the circle c, and the circles c, v cut at right angles.

Thus **any circle through two inverse points is orthogonal to the circle of inversion.**

Conversely, if two points P, P' are such that two circles c_1, c_2

through P, P' are orthogonal to the circle of inversion v, then P, P' must be inverse points.

Let Q_1 be an intersection of v and c_1. Since these cut at right angles, OQ_1 is tangent to c_1 at Q_1.

Join OP and produce it to meet c_1 again at R_1. Then by a well-known property of the circle, $OP.OR_1 = OQ^2 = k^2$. R_1 is therefore the inverse point of P with respect to v.

Similarly, if OP meet c_2 again at R_2, R_2 is the inverse point of P with respect to v, and so is identical with R_1, which is thus an intersection, other than P, of the circles c_1, c_2. Since, by hypothesis, this second intersection is P', P' is identical with R_1 and so must be the inverse point of P.

It follows that, if two circles through P, P' are orthogonal to the circle of inversion, every circle through P, P' will be orthogonal to this circle.

EXAMPLES

1. The limiting points of a set of coaxal circles are inverse points with respect to every circle of the set.

2. If the radius of the circle of inversion is made infinite, so that the accessible part of the circle is a straight line, show that any two inverse points are images of one another in the straight line.

3. Prove that a circle passing through two inverse points inverts into itself.

185. Homography on circular bases is preserved by inversion. Although inversion, considered as a relation between plane fields, is not homògraphic, yet it preserves homographic properties as regards ranges on corresponding circles (including straight lines as a particular case).

Referring to Fig. 59, it is clear that, since P', Q are corresponding points in a similarity, and a similarity is a particular case of a plane perspective and therefore of a homography, we must have

$$[Q]^2 \barwedge [P']^2.$$

But, since OPQ is a chord of the circle c through a fixed point, P, Q are mates in an involution on the circle.

Hence $[P]^2 \barwedge [Q]^2$ and $[P]^2 \barwedge [P']^2$.

Thus **a range of the second order on a circle inverts into a homographic range of the second order on the inverse circle.**

The case of a range on a straight line follows immediately from Fig. 58; here $[P] \barwedge O[P] \barwedge [P']^2$.

Since straight lines do not invert into straight lines, this property is not applicable to pencils.

If now we have, in any figure, two homographic ranges $[P_1]^2$, $[P_2]^2$ on circular (or rectilineal) bases, then the inverse ranges $[P_1']^2$, $[P_2']^2$ are such that

$$[P_1']^2 \barwedge [P_1]^2, \ [P_2']^2 \barwedge [P_2]^2$$

and since $[P_1]^2 \barwedge [P_2]^2$, therefore $[P_1']^2 \barwedge [P_2']^2$, or : **homographic ranges on circular bases invert into homographic ranges on the inverse bases.**

186. Angles invert into equal angles. In Fig. 58, let the tangent at P' to the locus of P' be drawn, as $P'T'$.

Then the angle $T'P'O$ is equal to $OA'P'$ in the alternate segment, and this again, by the similarity of the triangles AOP, $P'OA'$, is equal to OPA.

The line a and its inverse circle a' therefore make equal angles with OPP', but on opposite sides of it. By considering small elements it is easily seen that the same holds good of any two corresponding curves.

If now we have two curves s, t intersecting at any angle at P, and the two corresponding curves s', t' intersecting at the corresponding point P', then s, t make with OP, on one side of $OP'P$, angles equal to those which s', t' make with OP' on the other side of $OP'P$. By subtraction, the angle between s, t at P is equal to the angle between s', t' at P'.

Inversion is therefore what is usually termed a **conformal transformation,** that is, any elementary *small* figure inverts into a similar figure. It will be found, however, that they are oppositely, and not directly, similar, that is, they cannot be brought into similar situation by a relative displacement in their plane, but one of them must be turned over as a preliminary.

Note that, as a particular case of the theorem of the present article, an orthogonal intersection of two curves inverts into an orthogonal intersection of the two inverse curves.

Examples

1. Describe, with diagrams, the possible geometrical configurations which may be obtained by inverting the plane figure formed by two parallel straight lines and a line meeting them at right angles, with respect to a circle in the plane.

Three circles touch one another in pairs. Prove that the circle through the three points of contact cuts the three circles orthogonally.

2. Show that if σ, σ' are two very small corresponding areas at corresponding points P, P' in inverse fields, then in the limit $\sigma : \sigma' = OP^4 : k^4$.

187. Inversion of coaxal circles. Let c_1, c_2, c_3, etc., be a set of concentric circles, whose common centre is C. All straight lines l through C are then orthogonal to the circles c.

Now invert the field with regard to any given point O. The straight lines l invert into circles l' passing through O and through the inverse point C' of C, that is, into coaxal circles with real intersections O, C'.

The circles c then invert into circles c' which are orthogonal to the circles l' and therefore, by Art. 113, form a set of coaxal circles with imaginary intersections but real limiting points.

One of these limiting points is clearly C', since C is the point-circle of the set c, and a point-circle must invert into a point-circle.

To find the second limiting point, we note that, if c^{∞} is the circle centre C, of infinite radius, all the points of c^{∞} invert into O, which is therefore a point circle inverse to c^{∞}.

Thus **a set of concentric circles invert into a set of coaxal circles having for real limiting points the centre of inversion and the inverse point of the common centre.**

Conversely, if we invert such a set of coaxal circles with respect to one of their limiting points, we obtain a set of concentric circles. For the coaxal circles through the limiting points are orthogonal to the given coaxal circles, and invert into straight lines through the inverse of the second limiting point, which straight lines are orthogonal to every one of the new set of circles. But a circle which is orthogonal to a pencil of rays through a point has that point for centre ; the required result follows.

Note also that, since coaxal circles are circles through two fixed points (real or imaginary), coaxal circles necessarily invert into coaxal circles, whatever the centre of inversion.

We may therefore regard concentric circles as a particular case of coaxal circles. In this case the (ordinary) common points A, B of the circles coincide with Ω, Ω', so that the circles touch at Ω, Ω', as we should expect, and the common radical axis is the line at infinity.

EXAMPLES

1. Prove that two circles of a given coaxal system touch an arbitrary fixed straight line, and also that they are inverse with respect to that circle of the system which has its centre on the given straight line.

2. If a circle touches two given circles c_1, c_2, the points of contact are concyclic with the limiting points of the coaxal system determined by c_1, c_2.

3. If two variable circles touch one another, and also each of two given

circles c_1, c_2, show that the locus of their mutual point of contact is a circle coaxal with c_1, c_2.

4. Two given coplanar circles a, b have centres A, B; the inverse of a point P_1 with respect to a is P and the inverse of P with respect to b is P_2. Prove that, when P_1 describes a circle or a line, the corresponding point P_2 describes either a circle or a line; and show that, in the correspondence between the fields $[P_1]$, $[P_2]$, the system of coaxal circles with limiting points A, P_1 corresponds to the system of coaxal circles with limiting points B, P_2.

5. Invert the theorem: the locus of centres of circles which touch two concentric circles is two concentric circles.

188. Inverse figures invert into inverse figures. Let ϕ, ϕ' be two figures which are inverse with regard to a circle k.

Invert both ϕ and ϕ' with respect to another circle s. We obtain two new figures ϕ_1 and ϕ_1'; the circle k inverts, in general, into a circle k_1.

Let P, P' be any two corresponding points of ϕ, ϕ' and let their inverses with regard to s be P_1, P_1'.

Let c_1 be any circle through P_1, P_1'. This corresponds, in the inversion with regard to s, to a circle c passing through P, P'.

By Art. 184, c and k cut orthogonally. Hence, by Art. 186, c_1 and k_1 cut orthogonally. Thus any circle through P_1, P_1' cuts k_1 orthogonally, and, by Art. 184, P_1, P_1' are inverse points with regard to k_1. Since P_1, P_1' are an arbitrary pair of corresponding points, the figures ϕ_1, ϕ_1' are inverse with regard to k_1.

Thus two inverse figures invert into two inverse figures, the circles of inversion being inverted into one another.

EXAMPLE

Prove that two fields inverse with respect to a circle can be inverted into two fields which are the reflection of one another in a straight line.

EXAMPLES XIA

1. Show that a correlation projects into a correlation, and the base conics into base conics.

2. In two coplanar reciprocal fields ϕ_1, ϕ_2 the lines a_2, b_2, c_2 in ϕ_2 which correspond to the vertices A_1, B_1, C_1 of a triangle in ϕ_1 coincide respectively with the opposite sides B_1C_1, C_1A_1, A_1B_1. Show that every point of the plane corresponds to the same line in either field.

3. Show that the only correlation in which every point corresponds to the same line in either field is the transformation by reciprocal polars.

4. Discuss the characteristics of the correlation in which the base conics are concentric circles centre O. Show that the two lines which correspond to any point P are equally inclined to OP and that they intersect on OP

at a point P' which lies on the polar of P with respect to the circle k_1 which is the locus of incident points.

Prove also that the angle between the two lines corresponding to P is constant and equal to the angle between the tangents to k_2 from any point of k_1. Hence show that, if the correlation is real, k_2 must be interior to k_1.

5. By projecting the circular points into any given points, deduce the results corresponding to those of Ex. 4 in the case of a general correlation.

6. In a correlation three lines a, b, c through a vertex U correspond to three points A', B', C' on a line u' through U. Obtain a construction, using the ruler only, for the point V' corresponding to the line $v=u'$, and also for the line t corresponding to the point $T'=U$.

7. Reciprocate the theorem, that the angle in a segment of a circle is constant, with respect to the point at which one of the arms of the angle meets the circle, and hence show that if through the focus O of a parabola a line OP is drawn to meet a variable tangent to the parabola at P and making a fixed angle with this tangent, the locus of P is a fixed tangent to the parabola.

8. From the result that a chord of a circle which touches a concentric circle subtends a constant angle at the common centre deduce the following theorem.

If there be two conics having the same focus and corresponding directrix and if two tangents to one conic meet on the other conic and intersect the directrix at XY, then XY subtends a fixed angle at the focus.

9. If from a fixed point tangents are drawn to a series of concentric circles, find the locus of their points of contact. Hence obtain the following theorem by reciprocation. A fixed line meets a number of conics which have the same focus and corresponding directrix; then the envelope of the tangents at the points of intersection is a conic touching the fixed line and also the common directrix.

Hence find the envelope of the asymptotes of conics having a common focus and directrix.

10. If the two common tangents to two conics having a common focus O intersect at A and through A a line APQ be drawn meeting one conic at P and the other at Q, then, if the tangents at P and Q meet at R, OR is a bisector of the angle POQ.

[Reciprocate the theorem that tangents to two circles from a point on their radical axis are equally inclined to the line joining their points of contact.]

11. A variable chord of a rectangular hyperbola subtends a right angle at a fixed point O. By reciprocation with respect to a circle having O as centre show that the locus of the foot of the perpendicular from O on the chord is a straight line.

12. Prove that the polar reciprocal of a circle, taken with regard to a rectangular hyperbola, is a conic of which the centre of the rectangular hyperbola is a focus.

13. A conic is inscribed in a triangle ABC and has one focus at the circumcentre. Prove that the other real focus is the orthocentre.

14. Through a point O in the plane of a circle perpendicular rays OP, OQ are drawn to the circle. Show that PQ touches a fixed conic with a focus at O.

[Reciprocate the property of the orthoptic circle.]

15. Prove that if through a fixed point O inside a circle a straight line be drawn, meeting the circle at P and Q, the sum of the reciprocals of the perpendiculars from O on the tangents at P, Q is constant.

Obtain from this the theorem that the sum of the focal distances in an ellipse is constant.

16. Show that if ABC is a triangle and O any point in its plane, and if through O perpendiculars OP, OQ, OR are drawn to OA, OB, OC to meet BC, CA, AB respectively at P, Q, R, then PQR are collinear.

17. s' is the reciprocal of a curve s with respect to a point O, and P' is the point of s' corresponding to the tangent to s at P. If ρ, ρ' are the radii of curvature of s, s' at P, P' respectively, prove that $\rho\rho' = k^2 \operatorname{cosec}^3 \phi$, where k is the radius of reciprocation, and ϕ is the angle between OP and the tangent to s at P.

Illustrate this result by using it to find the radius of curvature at a point of a conic.

18. A circle c touches an ellipse k at P, and has its centre O on the orthoptic circle of k; the other common tangents q, r of c and k touch the circle c at Q, R respectively. By reciprocation with regard to c, show that there is one rectangular hyperbola which touches k at P and passes through Q and R; and that its asymptotes are parallel to the tangents from O to k.

Prove also that chords of this hyperbola which subtend right angles at O touch a certain parabola whose focus is O.

19. Prove by reciprocation the following.

A variable tangent p to a conic having S as focus meets two fixed tangents a and b at P and Q respectively; a line q through Q meets a at a point L such that the angle PSL is a constant angle. Show that the envelope of q is a conic having S for focus and touching a and b.

20. Prove that the reciprocal of a parabola k with respect to a circle c, whose centre O lies on k, is another parabola k', which touches at O the perpendicular from O to the axis of k, and has its axis parallel to the normal at O to k.

Show also that the directrix of k' is the reciprocal, with respect to c, of the Frégier point of O with respect to k.

If c has its centre O at an extremity of the latus rectum of k, prove that the focus of k' lies on the tangent at O to k.

21. Reciprocate, with respect to a circle with centre at a focus, the property that confocal conics intersect at right angles.

22. If one conic s is its own polar reciprocal for another conic t then the conic t is its own polar reciprocal for the conic s.

[Show that the conics have double contact at A, B and that if C be the common pole of AB then if a ray through C meet t at U, T the tangent at U is the polar of T with regard to s.]

23. Show that given two points A, B of a conic s, a conic t can be found touching s at A, B and such that s is its own polar reciprocal with regard to t.

24. If a figure be inverted with regard to any origin, show that an involution on a circle inverts into an involution on the corresponding circle.

25. Invert the theorem: tangents to a fixed circle cut a fixed concentric circle at a constant angle, the centre of inversion being a point in the circumference of the second circle.

26. Prove that the circles which pass through a given point L and cut a

fixed circle c at a given angle α all touch a second fixed circle k; and that L is a limiting point of the coaxal system determined by c and k.

Show also how to construct the circle k.

27. PQ is a common tangent to two circles and L, M are the limiting points of the set of coaxal circles determined by the two circles. Prove that PLQ, PMQ are right angles.

Transform this result by inversion, (i) from L, (ii) from M.

If PR, QS are chords of the circles passing through L, prove that RS is also a common tangent.

28. P_1 and P_2 are the inverses of a point P with respect to two different circles k_1 and k_2 respectively. If the inverse of P_1 with respect to k_2 coincides with the inverse of P_2 with respect to k_1, show that k_1 and k_2 are orthogonal.

EXAMPLES XIB

[The axes of co-ordinates are to be taken rectangular.]

1. The points $(3, 1)$, $(-3, 0)$, $(1, 2)$, $(1, -2)$ correspond to the lines $x=3$, $x=-3$, $y=2$, $y=-2$ in a correlation.

Find (i) the other incident points on the four given lines; (ii) the other incident lines through the four given points. Hence construct the base-conics.

2. With the data of Ex. 1, construct (i) the lines corresponding to the origin; (ii) the points corresponding to the line at infinity.

3. The base-conics of a correlation are the hyperbola

$$x^2-3y^2=1$$

and its auxiliary circle. Construct the conic corresponding to the conjugate hyperbola.

4. The circles $(x-2)^2+y^2=9$, $(x+4)^2+y^2=1$ are reciprocated with respect to the circle $x^2+(y-1)^2=1$.

Without drawing the reciprocal curves of the first two circles, find their four intersections.

5. Two circles centres A and B have radii 1 inch and 2 inches respectively, and $AB=4$ inches. Find (i) a circle with respect to which they reciprocate into confocal conics, (ii) the four intersections of these conics, (iii) the tangents at these four intersections.

6. Two circles centres A and B have radii 3 and 4 inches respectively, and $AB=2$ inches. Find a centre of inversion which inverts the above circles simultaneously into straight lines. If P, Q are two points such that $AP=7$ inches, $AQ=1{\cdot}5$ inches, and the angles BAP, BAQ are 30°, 135° respectively, construct the circles orthogonal to the given circles and passing through P, Q respectively. Construct also the circles inverse to the last found circles.

7. A circle passes through the point P $(3, 4)$, and touches the axis of x at the origin O of co-ordinates. If its inverse circle passes through the point P' $(2, -1)$ inverse to P, and the chord of this inverse circle corresponding to OP is of length 4 units, construct the circle of inversion.

!"#$! ! !% & ' ! E 6 ((2 4(> /
%7' ' 0-
V >' &

!"#$! ! !% & ' ! E - (" ' / ' (> /
B 0-
V =< 0'&
%9 - 4 ;;;

CONTENTS.

PAGE
Prolegomenon. Of Determinants ix
Chapter I. Of Perpendicular Coordinates referred to two axes ... 1
Exercises on Chapter I. 7
Chapter II. Trilinear Coordinates. The Point 9
Exercises on Chapter II. 20
Chapter III. Trilinear Coordinates. The Straight Line 21
Exercises on Chapter III. 34
Chapter IV. The Intersection of Straight Lines. Parallelism. Infinity 36
Exercises on Chapter IV. 54
Chapter V. The Straight Line. The Equation in terms of the Perpendiculars 58
Exercises on Chapter V. 69
Chapter VI. The Equations of the Straight Line in terms of the direction sines 71
Exercises on Chapter VI. 92
Chapter VII. Modifications of the System of Trilinear Coordinates. Areal and Triangular Coordinates 92
Table of Formulæ 96
Exercises on Chapter VII. 101
Chapter VIII. Abridged Notation of the Straight Line 104
Exercises on Chapter VIII. 115
Chapter IX. Imaginary Points and Straight Lines 117
Exercises on Chapter IX. 130
Chapter X. Anharmonic and Harmonic Section 132
Exercises on Chapter X. 145
Chapter XI. Transformation of Coordinates 147
Exercises on Chapter XI. 153
Chapter XII. Sections of Cones 154
Exercises on Chapter XII. 163
Chapter XIII. Abridged Notation of the Second Degree 165
Exercises on Chapter XIII. 171

PAGE
CHAPTER XIV. Conics referred to a Self-conjugate Triangle...... .. 173
Exercises on Chapter XIV. 180
CHAPTER XV. Conics referred to an Inscribed Triangle 192
Exercises on Chapter XV... 204
CHAPTER XVI. Conics referred to a Circumscribed Triangle 206
Exercises on Chapter XVI. 219
INTRODUCTION TO CHAPTER XVII. Notation, &c. 221
CHAPTER XVII. The General Equation of the Second Degree...... 226
Exercises on Chapter XVII.. 253
CHAPTER XVIII. The General Equation of the Second Degree continued 255
Exercises on Chapter XVIII. 285
CHAPTER XIX. Circles.. 287
Exercises on Chapter XIX. 304
CHAPTER XX. Quadrilinear Coordinates 307
Exercises on Chapter XX. 321
CHAPTER XXI. Certain Conics related to a Quadrilateral 325
Exercises on Chapter XXI. 331
CHAPTER XXII. Tangential Coordinates. The Straight Line and Point 332
Exercises on Chapter XXII... 343
CHAPTER XXIII. Tangential Coordinates. Conic Sections 345
Exercises on Chapter XXIII.. 364
CHAPTER XXIV. Polar Reciprocals 368
Exercises on Chapter XXIV.. 387
CHAPTER XXV. Conics determined by Assigned Conditions 390
Exercises on Chapter XXV. 400
CHAPTER XXVI. Equations of the Third Degree 401
Exercises on Chapter XXVI. 421
INTRODUCTION TO CHAPTER XXVII. General properties of Homogeneous Functions 426
CHAPTER XXVII. The General Equation of the n^{th} Degree 431
Exercises on Chapter XXVII. 451
MISCELLANEOUS EXERCISES... 455
NOTES ON THE EXERCISES. Results and occasional Hints 484
INDEX 499

!"#$! ! !% & ' ! E
2 4 (> /(' -( ) ' 0-
%9 : 4 - ;;;
6 0 ' M ;;;

!"#$! ! !% & ' ! E 2 4 (> / ' 0-

= !"#$! ! !% & ' ! E 2 4 (> / : 0-

@ !"#$! ! !% & ' ! E / (2 4 (> /
7' - 0-

!F"." G 4 4 !) * ! <! + ' - >
' ' ' 9

< !"#$! ! !% & ' ! + ' -
> ' ' ' 9 " 6 - > ' . / 0-
I& 8

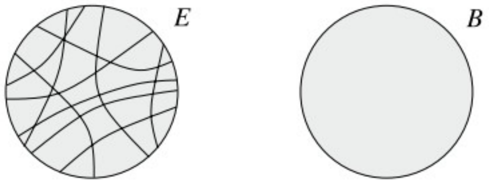

Figure 12.1 The set $E \subset \mathbb{R}^2$ is equivalent to the unit disk B. They both have distributional perimeter 2π, although we may arrange things so that $\mathcal{H}^1(\partial E)$ takes any value in $(2\pi, +\infty]$. In fact, $F = E \cup \mathbb{Q}^2$ is equivalent to B, has $\nu_B \mathcal{H}^1 \llcorner \partial B$ as its Gauss–Green measure, but is such that $|\partial F| = \infty$.

!F"." G 4 4 !) * ! <!. '0 6 0' &
. / ' - > / ? 4 U 7 -- F

E !"#$! ! !% & ' ! " 7 / = H '
' ' ' 7' 4' - 7' 4' - ? 4 / 8 0-

!"#$! ! !% & ' ! & . / ' - > /
? 4 U 7 -- F)') * 8 0-
%9 : 4 - ;;;
V 0'&

!"#$! ! !% & ' ! . '0 ' : %
' ' '] ' ? 4] 7 -- > / 0-
%9 : 4 - ;;;
V < 0'&

E E E

=

!F"." G 4 4 !) * ! E!"-4' - " ' / > /
' 09

!"#$! ! !% & ' ! = E @ "-4' - " ' /
> / 0-

!"#$! ! !% & ' ! = E @ "-4' - " ' /
> /(, 0-

!F"." G 4 4 !) * ! E!" & 9 '

!"#$! ! !% & ' ! "-4' -(: ' (" & 9 '
' 0-

!"#$! ! !% & ' ! : ' (" & 9 '(6) -' 0-

= !"#$! ! !% & ' ! : ' (" & 9 '(66) -' 0-

@ !"#$! ! !% & ' ! : ' (" & 9 '(666) -' 0-

< !"#$! ! !% & ' ! : ' (> / B 0-
V ' ;

E !"#$! ! !% & ' ! (" ' / (# ' 4 /
'& ' 0-
%9 : 4 - ;;; 6 0 ' M ;;;

I& 8 " :6 ".6%G %+ .7D #?:D %+ % "#6"G. F6++D#DG.6".6%G
8' # ' U - , / 9 ' - D U > '4 ' ' .

I& =8 D U & '4 ' ' 4' ' ' *

V >' 4' ' '

!F"." G 4 4 !) * ! E!" 9 9 7q ' ' *
3 0

!"#$! ! !% & ' ! " 9 9 7q ' ' *
3 0 0 & 0 * 8 E 0-
V - ' I ' 6 ' ;;;

!F"." G 4 4 !) * ! E! '

!"#$! ! !% & ' ! = "(7 /((' '
) / 0-

!"#$! ! !% & ' ! = "(' ' ' ("0 &/
7' -/ 0-

!"#$! ! !% & ' ! = ' 6 9/ . > 4 0-

!"#$! ! !% & ' ! = ' 66 9/ . > 4 0-

!"#$! ! !% & ' ! = ' 666 9/ . > 4 0-

= !"#$! ! !% & ' ! =
F ' (' -(6 & ' (' (0-
V '

@ !"#$! ! !% & ' ! =
F ' (' -(6 & ' (' (0-

< !"#$! ! !% & ' ! = 7 /(((') / 0-
I& E<8 G ' - : 9 3

!F"." G 4 4 !) * ! E!7 & > / B -

E !"#$! ! !% & ' ! @ & & - 0-

!"#$! ! !% & ' ! @ & & - -24

%9 : 4 - ;;; 6 ' ;;;
V ? ' 4 -' 'it N ' -' 0 L '

!F"." G 4 4 !) * ! E! ' '

!"#$! ! !% & ' ! < " & 9 '(> /(9 0-

!"#$! ! !% & ' ! < F 0 4 (> / B' 0-

!"#$! ! !% & ' ! < F 0 4 (> / B' 0-
%9 : 4 - ;;;

!"#$! ! !% & ' ! < 7 & (' ' H'& 0-
%9 : 4 - ;;;

($a = 1$). Compare the result with exact value of the arc length.

7.9.4. Find approximately the arc length of the hyperbola $xy = -1$ between the points with $x = 0.5$ and $x = 1$. [*Hint.* In this case we cannot obtain an exact value because the integral in (7.9.2) cannot be expressed in terms of elementary functions.]

7.9.5. Obtain approximate values of the number π by computing the arc length of a circle with a central angle of 45° (retain three, four, and five terms in the series).

7.10 Curvature and the Osculating Circle

Related to the arc length is the problem of determining the ***curvature*** k and the ***radius of curvature*** R of a curve at a point. For a circle the radius of curvature R is simply the circle's radius, and the reciprocal of the radius is the curvature, $k = 1/R$. The smaller the radius R of a circle (i.e. the greater the curvature k), the tighter the circle (Figure 7.10.1). When R is very large, on the contrary, the curvature is hardly noticeable, that is, the circle is "almost" a straight line; for instance, the fact that the earth's radius is so large means that we do not notice the curvature of the earth's surface (to notice it, we would have to observe the earth from outer space, and this is not common practice).

Let us take an arbitrary curve Γ now. It is clear (Figure 7.10.2) that the curve is curved more on the portion PQ, where it sharply turns to the left, than on the portion MN, where its direction does not change so rapidly. But what is the exact meaning of the last statement? The explanation is as follows.

Consider the angle φ between the tangents of the curve at the endpoints M and N of the arc MN. If over this arc the tangent rotates only in one direction when we move from point m to point N, the angle φ shows the magnitude of the rotation of the tangent on arc MN (or the magnitude of the "rotation" of the curve, since the direction of the curve is specified by the direction of the tangent). To determine the extent to which the curve is curved we must also know the arc length s of MN: if s is large, then even a slow change in the direction of the tangent can lead to a large value of φ.

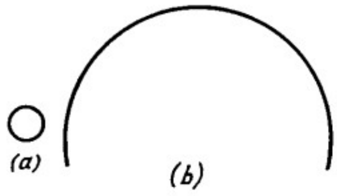

Figure 7.10.1

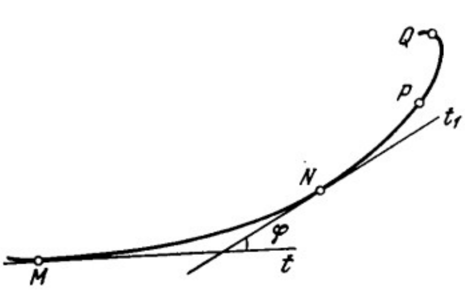

Figure 7.10.2

Thus, the ratio

$$k_{\text{av}} = \varphi/s \tag{7.10.1}$$

specifies the *average* (or *specific*) curvature of the curve over arc MN. *The curvature at point M of the curve* is defined then as the *limit* of the average curvature when the length of arc MN tends to zero:

$$k = \lim_{\Delta s \to 0} \frac{\Delta\alpha}{\Delta s},$$

where Δs is the length of a (small) arc MM_1 of curve Γ, and $\Delta\alpha$ is the increment of angle α corresponding to this arc (α is the angle between the tangent to the curve at point M and the x axis), or the angle of rotation of the tangent to the curve (the angle between the tangents t and t_1 to Γ at points M and M_1; Figure 7.10.3). But by virtue of the definition of a derivative (see Section 2.1), we can rewrite the last limit as

$$k = \frac{d\alpha}{ds}. \tag{7.10.2}$$

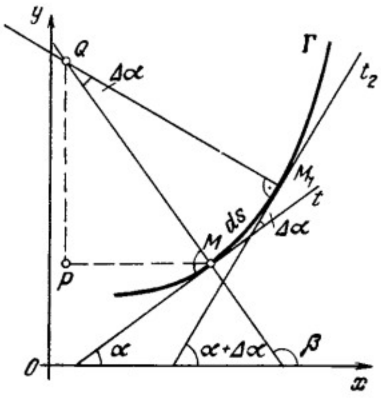

Figure 7.10.3

Thus, curvature k can be defined as the *rate of rotation of the tangent* to Γ as the point moves along the curve with a unit (linear) speed, that is, under conditions where point M travels the distance $MM_1 = ds$ in time $dt = ds$.

Now we need only recall that $\tan\alpha = y'$, so that $\alpha = \arctan y'$ and, hence,

$$d\alpha = d(\arctan y') = \frac{dy'}{1+(y')^2} = \frac{y''\,dx}{1+(y')^2};$$

moreover, $ds = \sqrt{1+(y')^2}\,ds$ (see Section 7.9). Thus, finally,

$$k = \frac{d\alpha}{ds} = \left[\frac{y''}{1+(y')^2}\,dx\right] \Big/ \left[\sqrt{1+(y')^2}\,dx\right] = \frac{y''}{[1+(y')^2]^{3/2}} \qquad (7.10.3)$$

and, hence,

$$R = \frac{1}{k} = \frac{[1+(y')^2]^{3/2}}{y''}. \qquad (7.10.3a)$$

If curve Γ is specified parametrically, $x = x(t)$ and $y = y(t)$ (see Section 1.8), then $dy/dx = y'/x'$ and $\frac{d^2y}{dx^2} = \frac{x'y''-x''y'}{(x')^3}$ (see formula (4.4.9)). Therefore, from (7.10.3) and (7.10.3a) it follows that

$$k = \left[\frac{x'y''-x''y'}{(x')^3}\right] \div \left[1+\frac{(y')^2}{(x')^2}\right]^{3/2} = \frac{x'y''-x''y'}{[(x')^2+(y')^2]^{3/2}},$$

$$R = \frac{1}{k} = \frac{[(x')^2+(y')^2]^{3/2}}{x'y''-x''y'}. \qquad (7.10.3b)$$

We note also that since α is a dimensionless quantity (we employ, as usual, radians for measuring angles) and s has the dimensions of length $[l]$, the curvature k of a curve has the dimensions $[l]^{-1}$ (say cm^{-1}) and the radius of curvature (as any radius), the dimensions of length $[l]$, say cm.

The notion of the radius of curvature (and hence the notion of curvature) of a curve may be introduced in a more geometrical manner. To this end we draw normals MQ and M_1Q (perpendicular lines) to the tangents to two close-lying points M and M_1 on Γ (by Q we have denoted the point of intersection of the normals; see Figure 7.10.3). The angle MQM_1 between the normals is equal to the angle $\Delta\alpha$ between the tangents to points M and M_1 (in accord with a familiar theorem of geometry on angles with mutually perpendicular sides). From this we can find the distance MQ from the curve to the point of intersection of the normals.

We regard the small portion of the curve as an arc of a circle. A normal to the circle is clearly a radius, and the point of intersection of normals is clearly the center of the circle. If the curve was a circle of radius R, then ds would be equal to $R\,d\alpha$, or $d\alpha/ds = 1/R$; this quantity, $d\alpha/ds$ (the curvature of the circle), is constant for any portion of an arc of the circle. Now we select the arc of a circle in such a manner that it is *as close as possible to* the arc MM_1 ("hugs" the arc). The words "as close as possible" are understood as a requirement that the first three terms in Taylor's expansion of the equation $y = y(x)$ of the curve,

$$y = y(a) + y'(a)(x-a) + \frac{1}{2}y''(a)(x-a)^2 + \frac{1}{6}y''(a)(x-a)^3 + \dots \qquad (7.10.4)$$

(we assume that the point M on the curve corresponds to the value $x = a$ of the abscissa), coincide with the first three terms in Taylor's expansion of the equation $y = Y(x)$ of the circle,

$$y = Y(a) + Y'(a)(x-a) + \\ + \frac{1}{2} Y''(a)(x-a)^2 \\ + \frac{1}{6} Y'''(a)(x-a)^3 + \dots \qquad (7.10.4a)$$

(see formula (6.1.18)).[7.29] In other words, we require that not only the values $y(a)$ and $Y(a)$ coincide but so do the first two derivatives of $y(x)$ and $Y(x)$ at point $x = a$:

$$y(a) = Y(a), \quad y'(a) = Y'(a), \\ y''(a) = Y''(a). \qquad (7.10.5)$$

But by virtue of (7.10.3), all this guarantees that at point M the curvatures of the curve Γ with equation $y = y(x)$ and of the circle Σ with equation $y = Y(x)$ coincide. This means that the radius $R = MQ$ of circle Σ is expressed by formula (7.10.3a).

The circle Σ that passes through point M of the curve Γ and is the closest to Γ is known as the ***osculating circle*** of the curve.[7.30] The radius R of this circle is the ***radius of curvature*** of the curve, and the center Q of circle Σ is said to be the center of curvature of the curve (at point M). Thus, the radius of curvature of a curve can be defined as the radius of the osculating circle; the curvature of the curve is the reciprocal of the radius of the osculating circle.

We note also that since by virtue of (7.10.4), (7.10.4a), and (7.10.5), the difference $Y(x) - y(x) = (1/6) \times [Y'''(a) - y'''(a)]\, dx + \dots$ changes sign as we pass through point M (i.e. as dx changes sign), the osculating circle Σ at point M of the curve Γ *intersects* the curve.[7.31] This makes it possible to describe the osculating circle thus.

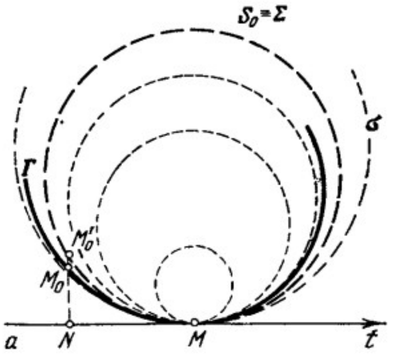

Figure 7.10.4

Let us consider various circles σ that touch point M of Γ and have at point M the same sense of convexity (the direction of bending), in other words, have the same tangent at point M as Γ and have centers that lie on the side of convexity of Γ (i.e. inside the curve) (Figure 7.10.4). Some of these circles lie completely inside Γ (i.e. on the side of convexity of Γ), and their curvature is *greater* than that of Γ. Others lie outside Γ, and their curvature is *smaller* than that of Γ. But there is only one circle that goes from one side of Γ to the other (this happens at point M), and this circle S_0 lies closest to Γ and is called the ***osculating circle*** Σ.

If conditions (7.10.5) are met for a straight line (the tangent t to Γ at point M), then the radius of curvature of Γ at this point (which is calculated via (7.10.3a)) becomes infinite and the curvature k vanishes. Such points M are sometimes called the ***rectifying points*** of curve Γ. It is clear that at

---

[7.29] A circle in a plane is defined by three numbers (parameters); these may be the coefficients a, b, and r in the equation of the circle $(x-a)^2 + (y-b)^2 - r^2 = 0$. To find these numbers (i.e. to define the circle) we must have *three* conditions, which may be the conditions in (7.10.5).

[7.30] An osculating circle Σ of a curve Γ can also be defined as follows. Consider the circle S that passes through three points M_1, M_2, and M_3 of curve Γ that lie close to point M. When all three points move toward M, circle S tends to the osculating circle Σ of curve Γ at point M.

[7.31] Exceptions in this case are only points where, in addition to conditions (7.10.5), we have $Y'''(a) = y'''(a)$. Examples are the vertices of an ellipse (points where the ellipse intersects the axes of coordinates); at such points, as symmetry considerations suggest, the osculating circle cannot interesct the ellipse (by analogy, all points of a curve Γ that possess this property are called the ***vertices*** of Γ).

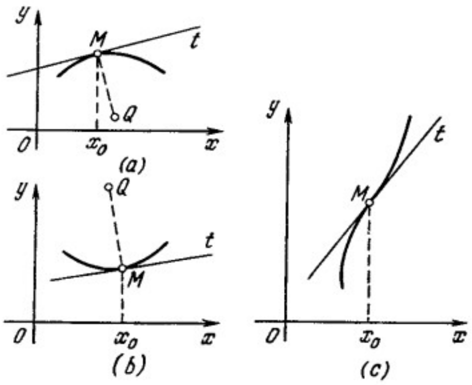

Figure 7.10.5

a rectifying point the curve Γ has no center of curvature Q (this center lies at infinity).

From (7.10.3) it follows that the signs of k and R (up till now we spoke only of the absolute values of k and R and ignored their signs) coincide with the sign of the second derivative y'', that is, characterize the sense of *convexity* of the curve (see Section 2.7). If the curvature is negative, that is, at point M we have a negative y'', the curve at this point lies below the tangent at point M, that is, the curve is *convex* (or *convex upward*; Figure 7.10.5*a*); the center of curvature Q then lies *below* point M. If the curvature is positive, that is, $y'' > 0$, the curve lies above the tangent and is said to be *concave* (*convex downward*; Figure 7.10.5*b*); in this case the center of curvature Q lies above point M. Finally, if the curvature at point M vanishes, that is, $y'' = 0$, then, as we already know, the curve at point M changes its sense of convexity; to one side of M the curve is convex upward and to the other it is convex downward (Figure 7.10.5*c*). At such a point the curve passes from one side of the tangent to the other, changes its sense of convexity (the direction of bending), and the point is called a ***point of inflection.*** At points of inflection, not only the first derivatives of the equation $y = y(x)$ of curve Γ and of the equation $y = Y(x)$ of tangent t coincide, but so do the second derivatives, $y'' = Y'' = 0$, that is, at such points the tangent t obeys the conditions (7.10.5); in other words, at a point of inflection M the *tangent* t acts as the osculating circle Σ, or M is a rectifying point of Γ.

For a straight line l, which of course can also be considered as a special case of a curve, the angle α between the tangent to l (which is simply l) and the x axis does not vary; this means that the curvature k of a straight line is identically zero (l is simply not curved in any way). For a circle S, the angle α changes, but the rate $k = d\alpha/ds$ at which α changes along the circle remains constant; in other words, we may say that circles (and straight lines, for that matter) are curves of constant curvature. It can be shown that these are the only types of such curves.

Suppose that M $(x, y, (x))$ is a point of a curve Γ. The tangent t to the curve at point M has a slope y' ($=\tan\alpha$), and the normal n is perpendicular to t, that is, forms an angle $\beta = \alpha + 90°$ with the x axis (see Figure 7.10.3); the slope of the normal is $\tan\beta = -\cot\alpha = -1/y'$. Therefore, the (directed) projections MP and PQ of the axes of coordinates of the line segment MQ of length R are, respectively, $p = R\cos\beta = -R\sin\alpha$ and $q = R\sin\beta = R\cos\alpha$, which means that the coordinates of the center of curvature Q are

$$\begin{aligned}\xi &= x + p = x - R\sin\alpha,\\ \eta &= y + q = y + R\cos\alpha.\end{aligned} \qquad (7.10.6)$$

If now we take into account formula (7.10.3a) for R and the obvious fact that

$$\cos\alpha = \frac{1}{\sqrt{1+\tan^2\alpha}} = \frac{1}{\sqrt{1+(y')^2}},$$

$$\sin\alpha = \frac{\tan\alpha}{\sqrt{1+\tan^2\alpha}} = \frac{y'}{\sqrt{1+(y')^2}},$$

we finally get

$$\xi = x - \frac{\sqrt{[1+(y')^2]^3}}{y''}\,\frac{y'}{\sqrt{1+(y')^2}},$$

$$\eta = y + \frac{\sqrt{[1+(y')^2]^3}}{y''}\,\frac{1}{\sqrt{1+(y')^2}},$$

that is,

$$\xi = x - \frac{1+(y')^2}{y''}\, y', \quad \eta = y + \frac{1+(y')^2}{y''}. \tag{7.10.7}$$

(This agrees with the condition that $\eta > y$ for $k > 0$, or for $y'' > 0$, or the condition that $\eta < y$ for $k < 0$, or for $y'' < 0$.)

If the curve is specified parametrically, $x = x(t)$ and $y = y(t)$, we can write (see (7.10.3b))

$$\begin{aligned} \xi &= x - \frac{(x')^2+(y')^2}{x'y''-x''y'}\, y', \\ \eta &= y + \frac{(x')^2+(y')^2}{x'y''-x''y'}\, x' \end{aligned} \tag{7.10.7a}$$

(see Exercise 7.10.4).

The locus of the centers of curvature $Q(\xi, \eta)$ of a curve Γ constitutes a new curve γ known as the ***evolute*** of curve Γ. The curve Γ with respect to its own evolute γ is called the ***involute***.[7.32]

Example 1. The parabola $y = x^2$. Here $y' = 2x$ and $y'' = 2$ (a constant), whereby the curvature k and the radius of curvature R of the parabola at point $M(x, y)$ $(=M(x, x^2))$ are

$$k = \frac{2}{\sqrt{(1+4x^2)^3}}, \quad R = \frac{\sqrt{(1+4x^2)^3}}{2}.$$

The center of curvature $Q(\xi, \eta)$ of the parabola with respect to point $M(x, x^2)$ has coordinates

$$\xi = x - \frac{1+4x^2}{2}\cdot 2x = -4x^3,$$

$$\eta = x^2 + \frac{1+4x^2}{2} = 3x^2 + \frac{1}{2},$$

that is,

$$\xi = -4\sqrt{\left(\frac{\eta - 1/2}{3}\right)^3}$$

$$= -\frac{4}{3\sqrt{3}}\sqrt{(\eta - 1/2)^3}, \text{ or } \xi = c\eta_1^{3/2}, \tag{7.10.8}$$

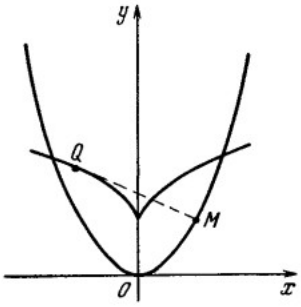

Figure 7.10.6

where $c = -4/3\sqrt{3}$ and $\eta_1 = \eta - 1/2$.

Thus, the evolute of $y = x^2$ is the *semicubical parabola* (7.10.8) whose cusp coincides with point (0, 1/2) and whose axis of symmetry is the axis of ordinates (Figure 7.10.6).

Example 2. The cycloid $x = a(t - \sin t)$, $y = a(1 - \cos t)$. We have $x' = a(1 - \cos t)$, $y' = a \sin t$, $x'' = a \sin t$, and $y'' = a \cos t$ (the primes denote derivatives with respect to parameter t). Whence,

$$\begin{aligned} (x')^2 + (y')^2 &= a^2(1-\cos t)^2 + a^2 \sin^2 t \\ &= a^2(2 - 2\cos t) = 4a^2 \sin^2 \frac{t}{2}, \\ y''x' - x''y' &= a\cos t \cdot a(1-\cos t) \\ &\quad - (a \sin t)^2 = a^2(\cos t - 1) \\ &= -2a^2 \sin^2 \frac{t}{2}. \end{aligned}$$

Thus, in view of (7.10.7a) we have

$$\begin{aligned} \xi &= a(t - \sin t) + 2a \sin t \\ &= a(t + \sin t), \\ \eta &= a(1 - \cos t) - 2a(1 - \cos t) \\ &= -a + a \cos t. \end{aligned} \tag{7.10.9}$$

To understand what curve is represented by these equations, we introduce a new parameter, $t_1 = t + \pi$ (or $t = t_1 - \pi$); since $\sin t = -\sin t$ and $\cos t = -\cos t_1$, we have

$$\begin{aligned} \xi &= a(t_1 - \sin t_1) - a\pi, \\ \eta &= a(1 - \cos t_1) - 2a. \end{aligned} \tag{7.10.10}$$

[7.32] The word "involute" is the Latin for "unwinding", while the word "evolute" is the Latin for "unwound," that is the curve that is being unwound; these names are related to the properties of evolutes and involutes discussed at the end of this section (see the text in small print).

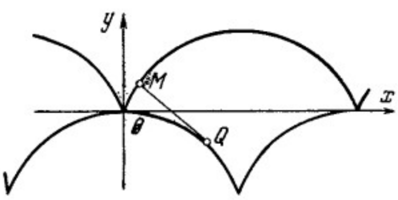

Figure 7.10.7

Thus, the evolute (7.10.9) (or (7.10.10)) of a cycloid is also a *cycloid*, $\xi_1 = a(t_1 - \sin t_1)$ and $\eta_1 = a(1 - \cos t_1)$, where $\xi_1 = \xi + a\pi$ and $\eta_1 = \eta + 2a$, only shifted in relation to the initial cycloid by $2a$ units downward ($2a$ is the diameter of the generating cirle) and πa units to be left (πa is the half-width of the arch of the initial cycloid) (Figure 7.10.7).

To discuss the properties of evolutes and involutes in detail, we must first differentiate in (7.10.6):

$$d\xi = dx - R\cos\alpha\, d\alpha - \sin\alpha\, dR,$$

$$d\eta = dy - R\sin\alpha\, d\alpha + \cos\alpha\, dR.$$

But since (see (7.9.2b))

$$dx = \cos\alpha\, ds = \cos\alpha\,\frac{ds}{d\alpha}\,d\alpha = R\cos\alpha\, d\alpha,$$

$$dy = \sin\alpha\, ds = \sin\alpha\,\frac{ds}{d\alpha}\,d\alpha = R\sin\alpha\, d\alpha,$$

we have

$$d\xi = -\sin\alpha\, dR, \quad d\eta = \cos\alpha\, dR. \tag{7.10.11}$$

Dividing one by the other, we get

$$\frac{d\eta}{d\xi} = -\cot\alpha = \tan(\alpha + 90°). \tag{7.10.12}$$

From this it follows that a tangent to evolute γ of a curve Γ (i.e. to the locus of the centers of curvature $Q(\xi, \eta)$) forms an angle $\beta = \alpha + 90°$ with the x axis, that is, coincides with the *normal* to Γ. And since point Q belongs to the normal n to curve Γ at point M, the respective tangent to γ coincides with n, that is, *tangents to the evolute γ of a curve Γ coincide with the normals to* Γ (Figure 7.10.8).[7.33]

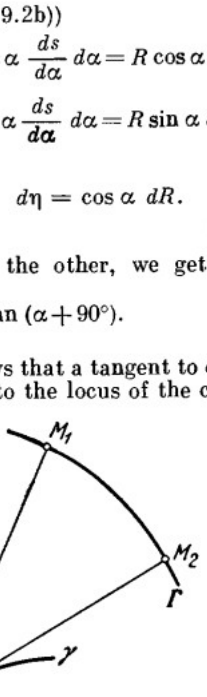

Figure 7.10.8

On the other hand, from (7.10.11) it follows that

$$d\xi^2 + d\eta^2 = \sin^2\alpha\, dR^2 + \cos^2\alpha\, dR^2 = dR^2,$$

$$\text{i.e.} \quad d\sigma = dR, \tag{7.10.12a}$$

where σ is the length of the arc of evolute γ of curve Γ (both $d\sigma$ and dR are assumed positive).

The last remark has the following meaning. Let us consider an arc M_1M_2 of curve Γ such that on it the derivative $R' = dR/ds$ retains its sign. We will agree to reckon the arc length along Γ in such a manner that the derivative of R is *positive*; in other words, we assume that the arc length is reckoned along Γ in the direction in which the radius of curvature R *grows*. Then the increment $M_2Q_2 - M_1Q_1 = R_2 - R_1 = \Delta R$ of the radius of curvature along arc M_1M_2 is equal to the length $\Delta\sigma$ of the evolute arc. And since the line segments M_1Q_1 $(=R_1)$ and M_2Q_2 $(=R_2)$ are tangents to evolute γ of curve Γ (see Figure 7.10.8), the process of reconstructing the involute Γ from its evolute γ resembles the unwinding under tension of a nonexpandable string from a spool shaped like γ; the end M of the string describes curve Γ. That is the origin of the names *evolute* and *involute* (see footnote 7.32).

By way of an example, let us construct the (parametric) equation of the *involute of a circle* of radius r and equation $x^2 + y^2 = r^2$, or $x = r\cos t$ and $y = r\sin t$, where t is the parameter (angle). Let us assume that the involute passes through point $A(r, 0)$ and that the unwinding of the string wound on the round spool occurs counterclockwise (Figure 7.10.9). The tangent MT to the circle at point $M(r\cos t, r\sin t)$ forms an angle $\alpha = t + 90°$ with the x axis, and the line segment MT of length rt along the tangent (rt is the length of arc AM of the circle) has projections $rt\cos\alpha = -rt\sin t$ and $rt\sin\alpha = rt\cos t$ on the axes of coordinates. This readily yields the following parametric expressions for the coordinates X, Y of point $T(X, Y)$ (i.e. the equation of the involute of the circle):

$$X = r\cos t - rt\sin t,$$

$$Y = r\sin t + rt\cos t. \tag{7.10.13}$$

The theory of evolutes and involutes was first developed by Christian Huygens in his

---

[7.33] In this connection it is often said that evolute γ of curve Γ is the *envelope of the normals* to Γ.

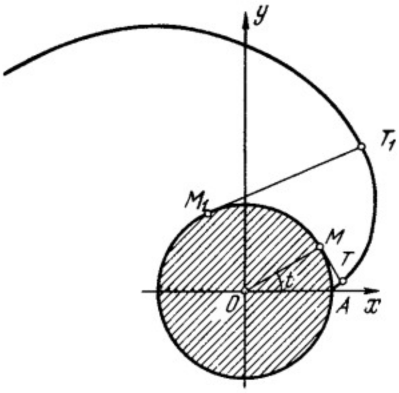

Figure 7.10.9

Horologium Oscillatorium (1673) in connection with the following problem. We wish to construct the equation of motion of a pendulum, that is, a material particle of mass m (the bob) moving along a curve γ:

$$m\,\frac{d^2s}{dt^2}=mg\sin\alpha, \tag{7.10.14}$$

where s is the arc length of curve γ and t is time, so that d^2s/dt^2 is the (*tangenrtial*, i.e. directed along the tangent to γ) *acceleration* of the particle, g is the acceleration of gravity, α is formed by the tangent l to curve γ and the horizontal, and $mg\sin\alpha$ is the (vertical) projection of the force of gravity mg on l (Figure 7.10.10; see Chapters 9 and 10). The period of oscillations of the pendulum is, strictly speaking, dependent on the swing (i.e. on the uppermost point on curve γ at which the pendulum starts oscillating). Huygens posed the question of what shape should curve γ have so that the pendulum bob swinging along this curve would take exactly the *same* time to complete swings of large and of small amplitude. By analyzing the differential equation (7.10.14) (which was not a simple task because at that time there was no differential and integral calculus) Huygens found that γ must be a *cycloid*. But how do we make a pendulum move along a cycloid? It was in this connection that Huygens developed the general theory of involutes and found that the evolute of a cycloid is also a cycloid (see Figure 7.10.7 and Eq. (7.10.10)). He also demonstrated that if at the point of suspension of the string of the pendulum certain "lips" in the form of a cycloid are attached, so that in the process of oscillations the string winds about them, then the end of the string (with the bob) of constant length describes a cycloid (Figure 7.10.11). The first pendulum clock constructed by Huygens had just such "lips"; however, later he dropped this idea because he found that the "lips" had practically no effect on the precision of the clock, since in view of the smallness of angles α in Eq. (7.10.14) we can always replace $\sin\alpha$ with α and the equation $m\,(d^2s/dt^2)=mg\,\alpha$ implies that the period of oscillations of a pendulum of *constant length* l is constant (see Section 10.3).

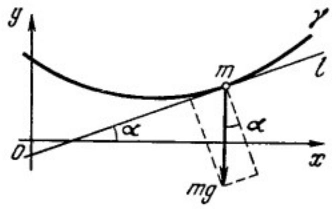

Figure 7.10.10

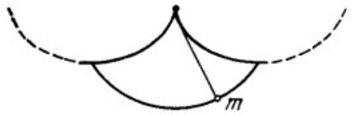

Figure 7.10.11

Exercises

7.10.1. Calculate the curvature (in an arbitrary point of the curve) of (a) an ellipse $x=a\cos t$ and $y=b\sin t$, (b) a hyperbola $y=1/x$, and (c) a fourth-order parabola $y=ax^4$.

7.10.2. Find the evolutes of (a) an ellipse $x=a\cos t$ and $y=b\sin t$, and (b) a hyperbola $y=1/x$.

7.10.3. Prove that the points of maxima or minima in the curvature of a curve correspond to the cuspidal points of the evolute of this point (see Figures 7.10.6 and 7.10.7).

7.10.4. Prove the validity of formulas (7.10.7a).

7.11 Solid Geometry Applications of Integral Calculus

In Section 3.6 we obtained the formula

$$V=\int_a^b S(x)\,dx, \tag{7.11.1}$$

where $S(x)$ is the area of a section of a solid by a plane $x=$ constant perpendicular to the x axis (we advise the reader to repeat the derivation of this formula). This formula was used to obtain an expression for the volume of a *pyramid*. The volume of a *cone* was obtained in exactly the same manner. Place the origin of the system of coordinates at the center of the circle of the

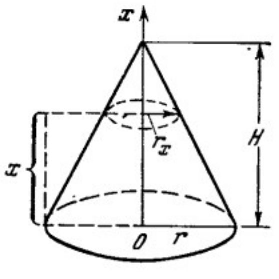

Figure 7.11.1

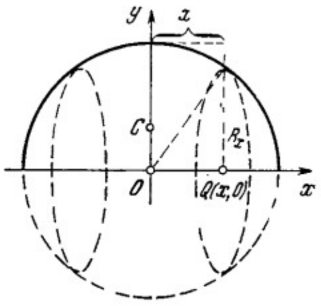

Figure 7.11.2

cone base and send the x axis along the altitude of the cone upward (Figure 7.11.1). Let $S(x)$ be the area of the section of the cone by a plane perpendicular to the altitude and distant x from the base. This section is a circle of radius r_x. From the similarity of triangles we have $r_x/r = (H - x)/H$, where r is the radius of the base and H is the altitude of the cone. Whence, $r_x = (r/H)(H - x)$ and $S(x) = \pi r_x^2 = \pi r^2 (H - x)^2/H^2$ and, consequently,

$$V = \int_0^H \pi \frac{r^2}{H^2}(H-x)^2\,dx$$

$$= -\frac{\pi r^2}{H^2}\,\frac{(H-x)^3}{3}\Big|_0^H = \frac{\pi r^2 H^3}{3H^2} = \frac{\pi r^2 H}{3}.$$

To obtain the volume of a *sphere* of radius R, we place the origin at the center of the sphere (Figure 7.11.2). The section cut by a plane perpendicular to the x axis and distant x from the origin is a circle of radius R_x. By the Pythagorean theorem $R_x = \sqrt{R^2 - x^2}$ and so $S(x) = \pi R_x^2 = \pi (R^2 - x^2)$, whence

$$V = \int_{-R}^{R} \pi (R^2 - x^2)\,dx$$

$$= \pi\left(R^2 x - \frac{x^3}{3}\right)\Big|_{-R}^{R} = \frac{4}{3}\pi R^3.$$

Formula (7.11.1) results in ***Cavalieri's theorem***,[7.34] which says that *if two solids have equal altitudes and if sections made by planes P and Q parallel to the bases and at equal distances to them always have a given ratio, the volumes of the two solids have this given ratio to each other*. Indeed, in calculating the volumes of such solids using formula (7.11.1), where the x axis is perpendicular to planes P and Q, we see that the ratio of the volumes is the same as the ratio of the integrands $S(x)$.

Let a solid be generated by revolving the curvilinear trapezoid depicted in Figure 7.11.3 about the x axis. In this case, the section is a circle of radius $y = f(x)$ and $S(x) = \pi y^2$. Using (7.11.1), we find the well-known formula for the *volume of a solid of revolution*:

$$V = \pi \int_a^b y^2\,dx. \qquad (7.11.2)$$

Let us find, say, the volume of a solid generated by the revolution of the upper half of the ellipse $x^2/a^2 + y^2/b^2 = 1$ about the x axis (make a drawing). This solid is called an ***ellipsoid of revolution***. Since for an ellipse $y =$

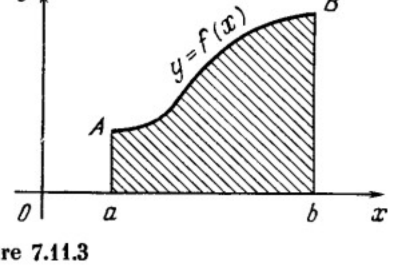

Figure 7.11.3

[7.34] This theorem played an important role in formulating the concept of the integral. Bonaventura ***Cavalieri*** (1598-1647), a pupil of Galileo, gave this theorem (without proof) in his *Geometria Indivisibilibus* (The Geometry of Indivisibles) (1635).

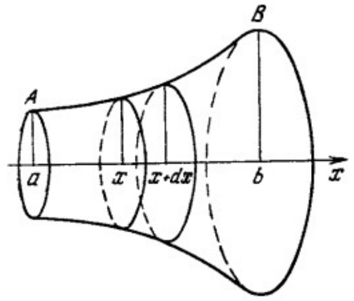

Figure 7.11.4

$(b/a)\sqrt{a^2-x^2}$, formula (7.11.2) yields

$$V=\pi\int_{-a}^{a}\frac{b^2}{a^2}(a^2-x^2)\,dx$$

$$=\frac{\pi b^2}{a^2}\left(a^2x-\frac{x^3}{3}\right)\Big|_{-a}^{a}=\frac{4}{3}\pi ab^2.$$

For $a=b=R$ we obtain, as expected the volume $(4/3)\,\pi R^3$ of a sphere of radius R.

Now let us derive the formula for the *surface area of a solid of revolution* (Figure 7.11.4). We consider a solid bounded by sections passing through the points x and $x+dx$. Denote by dF the lateral surface area of this solid. Regarding it as the frustrum of a cone, we get

$$dF\simeq\pi\,[y\,(x)+y\,(x+dx)]\,ds,$$

where ds is the length of the small portion of the curve by revolving which we get the surface of the solid; $ds=\sqrt{1+(y'\,(x))^2}\,dx$ (see Section 7.9). The sum $y\,(x)+y\,(x+dx)$ can be replaced with $2y\,(x)$, disregarding the quantity $y'\,(x)\,dx$ as compared with $y\,(x)$.[7.35] Therefore

$$dF\simeq 2\pi y\,(x)\sqrt{1+(y'\,(x))^2}\,dx.$$

The entire surface area of the initial solid of revolution is

$$F=2\pi\int_a^b y\,(x)\sqrt{1+(y'\,(x))^2}\,dx.\quad(7.11.3)$$

The surface area of a *sphere* is readily found by means of this formula. Indeed, a sphere is generated by the revolution of the upper semicircle about the x axis. The equation of a circle is $x^2+y^2=R^2$, whence

$$y=\sqrt{R^2-x^2},\quad y'=-\frac{x}{\sqrt{R^2-x^2}}.\quad(7.11.4)$$

Substituting into (7.11.3) yields

$$F=2\pi\int_{-R}^{R}\sqrt{R^2-x^2}\,\frac{R}{\sqrt{R^2-x^2}}\,dx$$

$$=2\pi Rx\Big|_{-R}^{R}=4\pi R^2.$$

Moreover, from (7.11.3) and (7.11.4) we can easily find the surface of a *spherical layer* (*spherical zone*) cut out of a sphere by two parallel planes $x=a$ and $x=b>a$ (see Figure 7.11.2 if, say, $b=R$, the layer becomes a *spherical segment* or *spherical cap*). Indeed, we have

$$F=2\pi\int_a^b\sqrt{R^2-x^2}\,\frac{R\,dx}{\sqrt{R^2-x^2}}$$

$$=2\pi Rx\Big|_a^b=2\pi R\,(b-a).$$

Thus, the sought-for surface area depends only on the altitude $b-a$ of the layer (segment) and the radius R of the sphere, which is really a remarkable and beautiful result.

Formulas (7.11.2) and (7.11.3) can be written also in a different way. Suppose the object depicted in Figure 7.11.3 is a metal plate of constant thickness. In this case the mass M of this plate is

$$M=\mu\int_a^b y\,dx=\mu S,\quad(7.11.5)$$

where μ is the density of the material of the plate (density per unit surface area; see Section 9.13), and $S=\int_a^b y\,dx$ is the surface area of the plate. The distance y_C from the center

---

[7.35] Note that in the expression dF of the sum $y\,(x)+y\,(x+dx)$ is multiplied by ds, so that the quantity we ignore is of the order of $dx\,ds\simeq dx^2$.

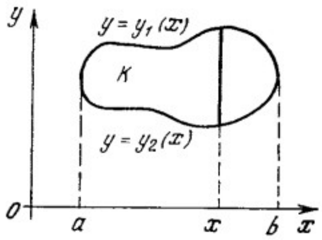

Figure 7.11.5

of gravity to the x axis is expressed by the formula

$$y_C=\frac{1}{2M}\int_a^b \mu y^2\,dx=\frac{1}{2S}\int_a^b y^2\,dx \qquad (7.11.6)$$

(see (9.13.12)). Comparing (7.11.2) and (7.11.6), we see that

$$V=2\pi\, Sy_C=2\pi y_C S, \qquad (7.11.7)$$

or that *the volume V of a solid revolution is equal to the area S of the figure whose rotation generates the solid multiplied by the circumference $2\pi y_C$ of the circle described in the process of rotation by the center of gravity of the figure.* (Here the center of gravity, or centroid, of a flat figure is assumed to coincide with the center of gravity of a homogeneous plate whose shape would be that of the figure.) This theorem is known as the ***(first) theorem of Pappus***.[7.36]

Formula (7.11.7) and the theorem of Pappus can be modified so that they incorporate the case of a solid generated by the rotation of a figure of arbitrary shape, and not only a curvilinear trapezoid.

Take, for example, the solid generated by the rotation about the x axis of the figure depicted in Figure 7.11.5. In this case the section of the solid by a plane perpendicular to the x axis is a circular ring (annulus) bounded by the circles of radii $y_1(x)$ and $y_2(x)$. Hence

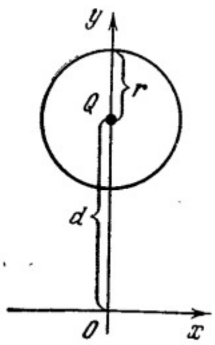

Figure 7.11.6

$$S(x)=\pi\,[y_1(x)]^2-\pi\,[y_2(x)]^2$$
$$=\pi\,(y_1^2-y_2^2)$$

and, therefore, the volume of the solid of revolution is

$$V=\pi\int_a^b (y_1^2-y_2^2)\,dx. \qquad (7.11.8)$$

On the other hand, the distance y_C from the x axis to the center of gravity (centroid) C of a homogeneous plate shaped as K in Figure 7.11.5 is

$$y_C=\frac{1}{2S}\int_a^b (y_1^2-y_2^2)\,dx=\frac{\int_a^b (y_1^2-y_2^2)\,dx}{2\int_a^b (y_1-y_2)\,dx} \qquad (7.11.9)$$

(see 9.13.11)). Hence here also $V=2\pi y_C S$.

Formula 7.11.3) can be interpreted in a similar manner. Suppose that the surface of a solid of revolution is generated by the rotation of an arc AB (see Figure 7.11.4) and imagine that the arc is a homogeneous heavy string with a constant density σ (mass per unit length; see the text between formulas (9.12.1) and (9.12.2)). In this case the distance y_C from the center of gravity of the string C to the x axis is given by the following formula:

$$y_C=\frac{1}{M}\int_{s_0}^{s_1}\sigma y(x)\,ds=\frac{1}{L}\int_{s_0}^{s_1} y(x)\,ds$$

$$=\frac{1}{L}\int_a^b y(x)\sqrt{1+(y'(x))^2}\,dx; \qquad (7.11.10)$$

here $x(s_0)=a$, $x(s_1)=b$, ds is the element of length of arc AB, $L=\int_{s_0}^{s_1} ds=\int_a^b \sqrt{1+(y'(x))^2}\,dx$

[7.36] ***Pappus*** of Alexandria (end of 3rd century A.D.) was the last of the great Greek mathematicians; in his works he set forth (usually without any proof) many results obtained by his predecessors and also by himself. The fact that the theorems of Pappus are often called the ***Pappus-Guldin theorems*** or even simply ***Guldin's theorems*** in the literature is completely groundless. Both theorems were formulated by Pappus, while their proofs, put forward by Paul ***Guldin*** (1577-1643), a Swiss monk and an amateur mathematician, were highly doubtful and substantially weaker than the proofs of the theorems given by B. Cavalieri and the famous Johann Kepler (1571-1630) (the latter proofs were harshly, but unjustly, criticized by Guldin).

$$V = 2\pi\, Sy_C = 2\pi y_C S, \qquad (7.11.7)$$

or that *the volume V of a solid revolution is equal to the area S of the figure whose rotation generates the solid multiplied by the circumference $2\pi y_C$ of the circle described in the process of rotation by the center of gravity of the figure.* (Here the center of gravity, or centroid, of a flat figure is assumed to coincide with the center of gravity of a homogeneous plate whose shape would be that of the figure.) This theorem is known as the ***(first) theorem of Pappus***.[7.36]

Formula (7.11.7) and the theorem of Pappus can be modified so that they incorporate the case of a solid generated by the rotation of a figure of arbitrary shape, and not only a curvilinear trapezoid.

Take, for example, the solid generated by the rotation about the x axis of the figure depicted in Figure 7.11.5. In this case the section of the solid by a plane perpendicular to the x axis is a circular ring (annulus) bounded by the circles of radii $y_1(x)$ and $y_2(x)$. Hence

---

[7.36] ***Pappus*** of Alexandria (end of 3rd century A.D.) was the last of the great Greek mathematicians; in his works he set forth (usually without any proof) many results obtained by his predecessors and also by himself. The fact that the theorems of Pappus are often called the ***Pappus-Guldin theorems*** or even simply ***Guldin's theorems*** in the literature is completely groundless. Both theorems were formulated by Pappus, while their proofs, put forward by Paul ***Guldin*** (1577-1643), a Swiss monk and an amateur mathematician, were highly doubtful and substantially weaker than the proofs of the theorems given by B. Cavalieri and the famous Johann Kepler (1571-1630) (the latter proofs were harshly, but unjustly, criticized by Guldin).

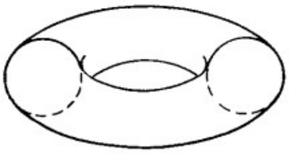

Figure 7.11.7

is the length of the string, and $M = \sigma L$ is the mass of string (see formulas (9.13.4) and (9.13.5b)).

Comparing (7.11.3) and (7.11.10), we get

$$F = 2\pi\ Ly_C = 2\pi y_C L, \qquad (7.11.11)$$

which means that *the surface area of a solid of revolution is the product of the length L of the arc whose rotation generates the surface by the circumference $2\pi y_C$ of the circle described by the centroid of the curve in the process of rotation.* This result (known as the ***second theorem of Pappus***) can be generalized so as to incorporate the case of a curve of arbitrary shape (say the one that is the boundary of figure K in Figure 7.11.5) and not only arc AB (see Figure re 7.11.3), which is the boundary of a curvilinear trapezoid.

The theorems of Pappus (7.11.7) and (7.11.11) often prove very useful. For instance, let us find the *volume* and *surface area of a torus* (*anchor ring*), which is generated by the rotation about the x axis of a circle of radius r whose center is d distant from the axis of rotation (Figure 7.11.6). Here, obviously, the center of gravity C of the circle (and of the circumference of the circle) coincides with the (geometrical) center Q of the circle. Hence, the volume of the torus in Figure 7.11.7 is

$$V = 2\pi d \times \pi r^2 = 2\pi^2 r^2 d, \qquad (7.11.12)$$

and the surface area is

$$F = 2\pi d \times 2\pi r = 4\pi^2 rd. \qquad (7.11.13)$$

The theorems of Pappus can be used in the reverse direction, so to say. The sphere in Figure 7.11.2 can be obtained by rotating a semicircle about the diameter that is part of the boundary of this semicircle. Then, if the radius of the sphere is R, the volume will be

$$V = 2\pi y_C \frac{\pi R^2}{2} = \pi^2 R^2 y_C.$$

Since we know that the volume V of a sphere is $(4/3)\,\pi R^3$, we can find the distance y_C from the *center of gravity of a semicircle* to the diameter:

$$y_C = \frac{4}{3}\pi R^3/(\pi^2 R^2) = \frac{4R}{3\pi} \simeq 0.4R.$$

Similarly, using the second theorem of Pappus, we can find the *center of gravity of a semicircumference* (see Exercise 7.11.3).

Exercises

7.11.1. Find the volume of a cone using the fact that the cone is a solid generated by rotating a right triangle about one of its legs.

7.11.2. Find the volume of a solid generated by rotating a figure bounded from above by the curve $y = \sqrt{x}$, from below by the x axis, and on the right by the vertical $x = 2$ about the x axis.

7.11.3. Using the second theorem of Pappus, find the center of gravity of a semicircumference.

7.12 Curve Sketching

The most primitive method of constructing the graph of a function $f(x)$ is to compute the value of $f(x_n)$ for a large number of points x_n. The usual procedure is to choose points x_n in the form $x_n = x_0 + na$, with $n = 0, \pm 1, \pm 2, \ldots$. This, clearly, is quite an extravagant method. In order to see the variation of the function on an interval Δx, we have to choose a step a much less than Δx, or $a \ll \Delta x$. And if the step (subinterval) is small, a very large number of points are required to embrace the whole range we are interested in. And yet we encounter the problem of constructing curves very often, since knowing the graph of a function provides extensive information about the properties of the function; for one, it immediately gives the number of real roots of the equation $f(x) = 0$, gives the intervals within which these roots lie, shows the maxima and minima of the function.

The techniques considered in Sections 7.1 and 7.2 make it possible to construct graphs much faster and more reliably and to gain a general picture of the shape of the curve. This requires, first of all, that we find the characteristic points of the graph—maxima, minima, discontinuities, salient points, points of inflection, etc.

Let us illustrate this fact by using the example of the graph of a third-degree *polynomial*,

$$y = ax^3 + bx^2 + cx + d. \qquad (7.12.1)$$

!"#$! ! !% & ' ! < ' ' '(/ 7 9 0-
V 0 ' - 4

' !"#$! ! !% & ' ! < - ' ' ' /
O P8 ' B / 8 E 0-
V = E 0'&

9 !"#$! ! !% & ' ! < - ' ' ' /
O P8 ' B / 8 E 0-
V E< 0'&

!"#$! ! !% & ' ! < - ' ' ' /
B / 7 8 E -24

- !"#$! ! !% & ' ! < - ' ' ' 0 /
6 9 8 E -24

= !"#$! ! !% & ' ! < ' ' (" * ' - 4 0-
V " ' / > / I& E 8

@ !"#$! ! !% & ' ! < ' ' (" * ' - 4 0-
V 4 ' - ' I& =<8@=

< !"#$! ! !% & ' ! < ' ' (" * ' - 4 0-
V G D - ' > / I& <8 =

%9 I& 8 ", 0' ' 8D -
h U ", 0' ' 8: 9' 4 *

E !"#$! ! !% & ' ! < ' ' ' 0-
V = 0'&
0 ' -

E' !"#$! ! !% & ' ! < # ' - ' B ' 6
' ' " D ' / "00 ' 6- ' ' - - 0-
V : 4 , - i}
%9 4 - ;;;

E9 !"#$! ! !% & ' ! < # ' - ' B '
' ' ' ' / '00 ' - ' ' - - %, - ? 4 / 8 @ -24

E !"#$! ! !% & ' ! < ' ' " , / B ' 6
1 5 ' - B ' & # -& 8 0-

!"#$! ! !% & ' ! < 0 ' (: 7 3 0-
V D -' 7 3

!"#$! ! !% & ' ! < 0 ' G 0-

!"#$! ! !% & ' ! < (' ' (> 0-
I& E 8 < , ' 4
%9 : 4 - ;;;

!"#$! ! !% & ' ! < (' ' (> 0-

!"#$! ! !% & ' ! < $ ' (' ' & 0-

' !"#$! ! !% & ' ! < ' & 6 - L '
' '00 ' / 8 = 0-
V <= 0'&

9 !"#$! ! !% & ' ! < ' & 6 - $ '
' "00 ' / F 4 9 ' 8 E= 0-

!"#$! ! !% & ' ! < ' & 6 - L '
' '00 ' / F 4 9 ' 8 E= -24

%9 : * ' '0 6 ' ;;;

!"#$! ! !% & ' ! <
. /((D /(' -('& A ' 0-
V @= 0'&

' !"#$! ! !% & ' ! < A ' . ' ' ' /
D / ' - '& 0-

9 !"#$! ! !% & ' ! < A ' . ' ' ' . /
D / ' - '& 5 = D- -24

= !"#$! ! !% & ' ! < . & / : 0-
V <E 0'&

!F"." G 4 4 !) * ! E! 2 4 > / 9

@ !"#$! ! !% & ' ! E % ' - 9 2 4
> / 6 8 < 0-

!"#$! ! !% & ' ! E % ' - 9 2 4 > /
6 8 < -24

< !"#$! ! !% & ' ! E % ' - 9 2 4
> / 66 ') * 8 < 0-

!"#$! ! !% & ' ! E % ' - 9 2 4 > /
66 ') * 8 < -24

!F"." G 4 4 !) * ! E! 2 4 > /

E !"#$! ! !% & ' ! "-4' -(" ' / (> /
' 09 0-

!"#$! ! !% & ' ! " & 9 ' (2 4 (> /
1 9 0-

!"#$! ! !% & ' ! 7 & (> / B - 0-

!"#$! ! !% & ' ! 6 ((2 4 (> /
+ 0-

!"#$! ! !% & ' ! 2 4 (> / 0-

!"#$! ! !% & ' ! 2 4 (> /(' -(
) ' 0-

!F"." G 4 4 !) * ! E!$ ' ' 9

!F"." G 4 4 !) * ! E!$ ' ' 9

= !"#$! ! !% & ' ! ' ' D -/ ' 0 &
8 E 0-

@ !"#$! ! !% & ' ! + - $ ' 3' 0 &
8 @ 0-

< !"#$! ! !% & ' ! : $ ' '
B 9 5 D, ' - 0 & 6 ' ' 8 @ 0-

E !"#$! ! !% & ' ! G ' - 0 & '&
8 @ -24

!"#$! ! !% & ' ! G ' - 0 & '&
8 @ 0-

= !"#$! ! !% & ' ! ' & ' L '
' 5 ' 5 0 / 0 / 5 ' - ' ' ' * 8 < 0-

= !"#$! ! !% & ' ! $ ' -/ '
0 & 8 < 0-

= !"#$! ! !% & ' ! $ ' '
= 8 0-

= !"#$! ! !% & ' ! $ ' ' 0-

= !"#$! ! !% & ' ! $ ' ' :
= 0-

== !"#$! ! !% & ' ! $ ' ' 9
@ 0-

=@ !"#$! ! !% & ' ! $ ' ' "
- 0 & 8 -24

=< !"#$! ! !% & ' ! $ ' ' "

- 0 & 8 0-

=E !"#$! ! !% & ' ! # ' 4 L ' '
'4 L ' 0 & 8 -24

= !"#$! ! !% & ' ! # ' 4 L ' '
'4 L ' 0 & 8 0-

< E

= =

!F"." G 4 4 !) * !

!F"." G 4 4 !) * ! !D ' G / ' -
. ' G ' " 0 /

!"#$! ! !% & ' ! D ' G / ' -
. ' G ' " 0 / 0-

!F"." G 4 4 !) * ! !+ # ' F '
> / ' - # ' 4 /

!"#$! ! !% & ' ! 0' ' / ' -
: ' / / 0' 8 ' ' 5 . / C "00 ') * c) 8 = 0-
%9 : 4 - ;;;

!"#$! ! !% & ' ! + # ' F '
> / ' - # ' 4 / 0 & 8 < 0-
I& 8@<
>' ' ' 0 9 - /0 & 5 ' -
') 5 ' ' -5 ' -' - 0 9 5 E =
. - 9/ >' ' * 6 & 92 5
>' - - ' -/ L /0 &

V : 4 - ;;;

!"#$! ! !% & ' ! + # ' - '
& / ' - ' 4 / 0 & 8 < 0-

= !"#$! ! !% & ' ! > / # ' ' ' -
. - 0' 6 ' ') 8 0-
h> / . X 0' . X

%9 D 4 - > / . 8 ' '

@ !"#$! ! !% & ' ! 7' -9 * > " ' / 5
4 6 0-

!"#$! ! !% & ' ! 7' -9 * > " ' / 5
4 6 -24

I& 8 '9 ' ' 4' '

V ' 0 9 '9 / ' ' 4' ' ' # ' '
' - G ' - & / '4 ' ' ' - ' G 9/ ' 0 9 /
4' / & ' ' 4'

< !"#$! ! !% & ' ! 7' -9 * > " ' / 5
4 66 0-

!"#$! ! !% & ' ! 7' -9 * > " ' / 5
4 66 -24

I& 8 @ 4 & ' - '0 & . # ' ' > /

E !"#$! ! !% & ' ! 7' -9 * > " ' / 5
4 666 0-

!"#$! ! !% & ' ! 7' -9 * > " ' / 5
4 666 -24

I& 8 " 4 / D ' -

I& 8@ 0 . > /

!F"." G 4 4 !) * ! !+ -' ' ' '
- ' > /

!"#$! ! !% & ' ! 866. ADD ' 2'/ '
+ -' ' ' ' - ' > / ' 8 E 0-
%9 : 4 - ;;;

I& =< '0 ' & :

!F"." G 4 4 !) * ! !>'

!"#$! ! !% & ' ! ' - &' <&' 0-
V . ' 0 >'

!"#$! ! !% & ' ! >' (7 *q 0 (> ' 0-
V . -) & & - 7 *q 0 >'

!F"." G 4 4 !) * ! !> ' & 9 ' 0 &
& & ' - 0

!"#$! ! !% & ' ! = > ' & 9 ' 0 &
& & ' - 0 0 & '& : - 8 0-

I& E8 D - ' > / ' >"

!F"." G 4 4 !) * ! !> " & 9 ' /

!"#$! ! !% & ' ! @ > " & 9 ' /
' 9 -& ? 4 / 8 0-
9 : 4 - ;;;

I& @8 '0 = # ' 4 / ' - 0'

I& E8 @ '0 / / ' - &' & /

I& <8= '0 > '4 '

!F"." G 4 4 !) * ! !7' -9 * > 0 &

!"#$! ! !% & ' ! < D- ' -)'/ '
7' -9 * > 0 &("00 ' ' # & 5 0 5
G ' 0 &5 ' - # 9 0 & 8 = -24

= !"#$! ! !% & ' ! < D- ' -)'/ '
7' -9 * > 0 &("00 ' ' # & 5 0 5
G ' 0 &5 ' - # 9 0 & 8 = 0-

@ !"#$! ! !% & ' ! < 7' -9 * > 0 &
"00 ' ' # & 5 0 5 G ' 0 &5 ' - # 9 0 &
8 = 0-

I& 8 '0 / /5 + ' 5 ' - 6 '

I& @ 8 =

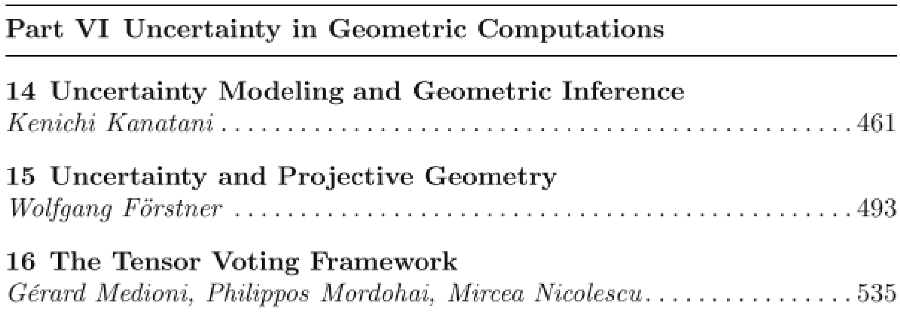

Part VI Uncertainty in Geometric Computations

14 Uncertainty Modeling and Geometric Inference
Kenichi Kanatani 461

15 Uncertainty and Projective Geometry
Wolfgang Förstner 493

16 The Tensor Voting Framework
Gérard Medioni, Philippos Mordohai, Mircea Nicolescu 535

!F"." G 4 4 !) * ! !7' -9 * ' '

< !"#$! ! !% & ' ! E 7' -9 * F ' -
0 ' ' > / '0 ' ' - 7' (# 8 0-
V = 0'&

E !"#$! ! !% & ' ! E 7' -9 * ' '
& ' - '0 ' ' - 7' (# 8 @ 0-
V = 0'&
%9 : 4 - ;;; # t -' ' I ';;;

!F"." G 4 4 !) * ! !7' -9 * . /

!"#$! ! !% & ' ! 7' -9 * . /
4 0-
V E 0'&

I& 8 '0 % . Q] ' - .] ' /
. Q 0'

I& =8 '0 ' 5 5 ' - -

!"#$! ! !% & ' ! 7' -9 * . /
4 0-
V EE 0'&

I& = 8 @ '0 = # ' ' 5 99 & '0 ' - 9 ' ' '

I& ==8 @=
'0
0 , 0 2 4
F'4 - F '

6 - ==
)' - \ =E
. ' 3 ' 0' ' 3' @
. ' 3 ' - 4' 4 @
' 3 ' 0' ' 3' ' \9 @
' 3 ' 0' ' 3' 8 @
. > ' & 0' ' 3' @@
F \ & ' & @<
.] @
. . <
' & ' & '0 <=
= 7 / <<
= # 0 ' ' - ' ' <<
= . / '0 <
= 7 / ' - 9 - & E
= + ' / E
= = $ ' + ' / E=
= @ F / EE
= < 7 / \9 E
@ 0' 0' ' 3'
@ 0' \ '
@ $ '- ' - ' ' - ' - ' '
@ : \9
@ : ' 3 ' E
@ = 6 \ ' 0' 9 / =
# =

I& = 8=
'0
0' * & ' - . Q 0'
'-'/ 1 2 '

6 - =
0' * & =
' 0 =
2 4 & / =

2 4 # ' ' =
0' * & ' = =
& - / = <
& - / = <
? \ - 0 =
F / =
+ , 9 / =
& - =
. - ' = <
0 ' = E
D4 - =
=

!F"." G 4 4 !) * ! !7q ' ' * # 3 0

!"#$! ! !% & ' ! 7q ' ' * # 3 0 0-
V 0'&

!F"." G 4 4 !) * ! !6 - $ ' /
' &

!"#$! ! !% & ' ! ' & 6 - L '
' '00 ' / 8 = 0-
%9 : 4 - ;;;

!"#$! ! !% & ' ! ' & 6 - $ '
' "00 ' / F 4 9 ' 8 E= 0-

!"#$! ! !% & ' ! ' & 6 - L '
' '00 ' / F 4 9 ' 8 E= -24

!F"." G 4 4 !) * ! !1 * : 0 - E E
V = 4

!F"." G 4 4 !) * ! ! - ' ' ' /
7 9

= !"#$! ! !% & ' ! - ' ' ' /
O P8 ' B / 8 E 0- 8 '&
V = E 0'&

@ !"#$! ! !% & ' ! - ' ' ' /
O P8 ' B / 8 E 0- 8 '&
V E< 0'&

< !"#$! ! !% & ' ! - ' ' ' /
B / 7 8 E -24

E !"#$! ! !% & ' ! - ' ' ' 0 /
6 9 8 E -24

!F"." G 4 4 !) * ! ! q9 5 " & + - ' -5 < E@E
V 4

!F"." G 4 4 !) * ! !G * ' 6 : 9' 4 */5 ' & /

!"#$! ! !% & ' ! G * ' 6 : 9' 4 */5
' & / D 8 0-
V = 0'& : 4 - ;;;

I& <8 . - ' '

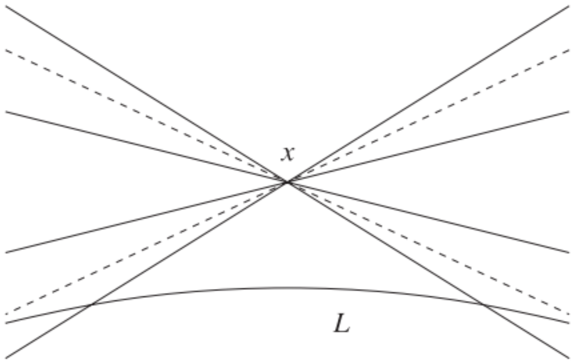

+ & . -' - 0 0' ' & , : D'
0' ' ' 5 0 ' & ,5 0' ' & ' / '
: ' / ' - :

I& 8 <

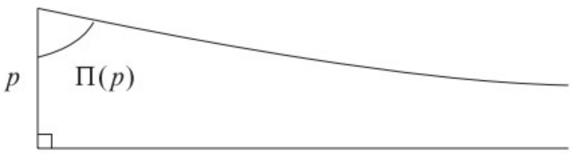

Figure 2. The parallelism angle $\Pi(p)$ corresponding to the segment p.

!"#$! ! !% & ' ! '0'- 0 " 8 - > /
/ 0 & 8 0-
%9 : 4 - ;;;
V <= 0'&

!"#$! ! !% & ' ! '0'- 0 " 8 - ' 9 &
' ' & / D 0 ' ' ' ' / 8 0-
V @ 0'&

I& 8 G D - ' & /

6 -
% 9' ' - ', &
6 -

)'
7 9] ', ' & /
DL 4' D -] 0' ' 0 ' E
= . ', /0 9 & /
@ 0 ' & /
< D - ' & ' 9 ' - ' /0 9
' - 0 ' & '
E
. ' 0 '
6 -
' & /
'] . ' - ' & / <
" & ' - \ ' & /
= . ' & ' L '- ' ' ' & / E
@ 1 '//' ' L '- ' ' =
< 2 ' & / =
. /0 9 0 ' ==
9' 0 0 /0 9 & / ==
% L '- ' ' /0 9 & / =E
. ' & ' L '- ' ' /0 9 & / =
DL - ' @
= > ' L '- ' ' @<
= " ' <
= 6 - <
= " & ' - \ /0 9 & / <@
= . ' ' <E
= F D - ' & / E
= = F D - ' & / E<
@ . & / E
@ 6 - E
@ . D /
@ . ' L ' { , / { , { / F { , { / =
@ / '& '] <
@ = . & / ' ' 9 ' / ' &
@ @ > ' ' & ' L '- ' ' E
@ < 0 ' & /
@ E . D / 0 ' & /
< ' ' =
< 6 - =
< ' ' /0 9 & / @
< ' '9 ' - / =
< 7 / ' ' - '00 '
< = . ' L ' , / F , /
< @ : 9' 4 */] ' & 0' ' =
E -
E 6 -
E " D - ' - /0 9 0 '
E " - ' & ' & 9 ' =
. ' ' & @
6 - @
" - & @=
<

!F"." G 4 4 !) * ! !. ' ' ' / D /
' - '&

!"#$! ! !% & ' ! A ' . ' ' ' /
D / ' - '& 0- = D-
V @@E 0'&

!"#$! ! !% & ' ! A ' . ' ' ' . /
D / ' - '& 5 = D- -24
V == 0'&

!F"." G 4 4 !) * ! !? ,0 - ' ? 4
" ' - ' /

!"#$! ! !% & ' ! = ? ,0 - '
? 4 " ' - ' / B ' ') ' B - 0-

!F"." G 4 4 !) * ! !B ' ' ' ' ' /
'00 ' - ' ' - -

= !"#$! ! !% & ' ! @ # ' - ' B ' 6
' ' " D ' / "00 ' 6- ' ' - - 0-
V = 0'&
%9 : 4 - ;;;

@ !"#$! ! !% & ' ! @ # ' - ' B '
' ' ' ' / '00 ' - ' ' - - %, - ? 4 / 8 @ -24

E

!F"." G 4 4 !) * !

!F"." G 4 4 !) * ! !" 6 - $ ' %0 "
0 / '00 '

!"#$! ! !% & ' ! < " 6 - $ '
%0 0-

!F"." G 4 4 !) * ! !" 6 - $ ' %0
' -) 0 /

!"#$! ! !% & ' ! E H' ' " 6 -
$ ' %0 (' -) 0 / 8 0-
I& 8 E '0
D '& B'4 . / ' - ' : &

%9

: 4 - ;;; 6 0 ' ;;;

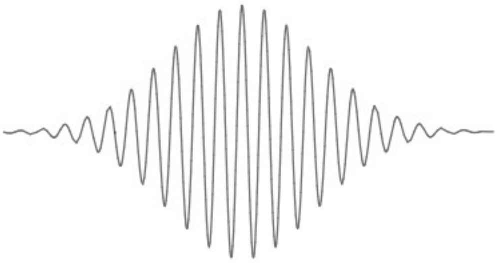

FIGURE 1.2.1 Schematic of a Gaussian wavepacket. The envelope of the wavepacket propagates with group velocity $u(\omega_0)$. Under the envelope is the carrier wave of frequency ω_0. The carrier propagates with phase velocity $\omega_0/k(\omega_0)$. The wavepacket is centered at space-time point $\tau = t - r/u(\omega_0) = 0$. At $\tau = 0$, all of its "Fourier-modes" superpose coherently and constructively due to their common initial phase $\varphi(\omega) = \varphi_0$.

!F"." G 4 4 !) * ! !D & & D -/ '

!"#$! ! !% & ' ! D & & D -/ ' 0-
V =<@ 0'&

Preface xiii

Author biography xviii

1 Maxwell's equations, potentials, and boundary conditions 1-1

1.1 The time-domain Maxwell's equations 1-1
1.1.1 The Lorenz gauge, Coulomb gauge, and causality 1-2
1.1.2 Existence theorem for fields 1-8
1.2 Frequency domain Maxwell's equations 1-9
1.2.1 Classification of media 1-13
1.2.2 Boundary conditions 1-14
1.3 Field determination by radial components 1-23
1.3.1 Multipole expansion, Debye potentials, and related theorems 1-24
1.3.2 Additional theorems related to spherical harmonics 1-33
References 1-35

2 Electrostatics and magnetostatics 2-1

2.1 Energy related theorems in electrostatics 2-1
2.1.1 Reciprocity theorem in electrostatics 2-7
2.2 Principle of virtual displacement for static fields 2-8
2.3 Theorems related to harmonic functions 2-10
References 2-11

9 Geometrical optics and Fermat's principle **9-1**
9.1 Geometrical optics and Fermat's principle 9-1
  9.1.1 Field discontinuities, wavefronts, and the eikonal equation 9-2
  9.1.2 Ray equations 9-7
  9.1.3 Ray tracing 9-9
  9.1.4 Mechanical interpretation of the ray path 9-13
  9.1.5 Optical path length and Fermat's principle 9-14
  9.1.6 Sommerfeld–Runge and Snell's laws, the Lagrange integral invariant 9-15
  9.1.7 Transport of the geometrical optics field 9-17
  9.1.8 Other theorems of geometrical optics 9-22
9.2 Gradient metasurfaces and the generalized Snell's law 9-23
References 9-26

14 Projection slice theorem and computed tomography **14-1**
14.1 Radon transform and projection slice theorem 14-1
14.2 Computed tomography 14-4
Reference 14-7

17 Covariant formulation of Maxwell's equations **17-1**
17.1 Preliminaries of tensor calculus 17-1
  17.1.1 Riemannian space 17-11
  17.1.2 Minkowski space 17-31
17.2 The covariant form of Maxwell's equations in Euclidean pseudo-space 17-33
17.3 Maxwell's equations in an arbitrary spacetime 17-40
17.4 Covariant form of Maxwell's equations in stationary matter 17-42
17.5 Transformational electromagnetics 17-43
  17.5.1 Design of a cylindrical cloak 17-47
  17.5.2 General transformational equations 17-50
References 17-52

I& E '0

2 '- 0 - & '0 /

I& < '0 <

4' ' ' ',] L '

'0

IOP Publishing

Engineering Electrodynamics
A collection of theorems, principles and field representations
Ramakrishna Janaswamy

Chapter 14

Projection slice theorem and computed tomography

As remarked previously in chapter 10, the goal of inverse scattering is to determine the object shape and its constitution from measurements of transmitted field quantities (or field derived quantities) when the object is irradiated by an electromagnetic wave. The Radon transform and the projection slice theorem find extensive use in inverse scattering by x-ray irradiation [1]. Object features imaged with x-rays are often very large compared to the wavelength and as such the phase of the wave becomes less important, particularly when the detectors used are non-coherent. In such cases one relies on the measurement of the intensity of the transmitted radiation. Recall from equation (9.81) that the intensity ratio under geometrical optics considerations is proportional to the exponent of the integral of the optical ray path through the medium. The Radon transform and the projection slice theorem apply in such circumstances. Practical applications of the Radon transform and the projection slice theorem are in computed tomography (CT scans) and all of these topics will be considered in the present chapter. An example imaging problem is considered at the end of the chapter to illustrate the various steps.

14.1 Radon transform and projection slice theorem

The Radon transform is an integral transform which maps a continuous function $f(x, y)$ defined on a plane to another function $\mathscr{R}[f]$ defined on the two-dimensional space of parallel straight lines in the plane, whose value on a particular line is equal to the line integral of $f(x, y)$ over that line. With the reference to figure 14.1, consider a tilted straight line L whose normal distance from the origin is ξ and on which the arc length is denoted by η. The tilt of the straight line is governed by α ($\alpha = 0$ for a vertical line and $\alpha = \pi/2$ for a horizontal line). The normal to the straight line is $\hat{\boldsymbol{\xi}} = \hat{\mathbf{x}} \cos\alpha + \hat{\mathbf{y}} \sin\alpha$. For a fixed ξ, the Cartesian coordinates on the straight line are $\mathbf{r} = (x(\eta; \xi), y(\eta; \xi)) = (\xi \cos\alpha - \eta \sin\alpha, \xi \sin\alpha + \eta \cos\alpha)$. This

doi:10.1088/978-0-7503-1716-0ch14

© IOP Publishing Ltd 2020

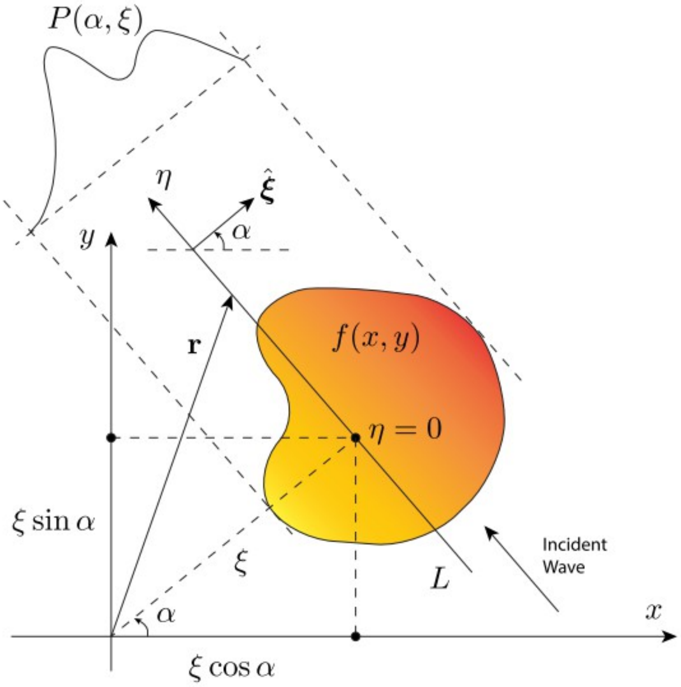

Figure 14.1. Transformation of an object with density $f(x, y)$ to a projection $P(\alpha, \xi)$ through the Radon transform. The color and shade denote the value of density.

simply corresponds to rotation of the xy-coordinate system through an angle α with respect to the x-axis. The inverse transformation yields $\xi = x\cos\alpha + y\sin\alpha = \mathbf{r}\cdot\hat{\boldsymbol{\xi}}$, $y = -x\sin\alpha + y\cos\alpha$. The Jacobian of the transformation is equal to one and $\mathrm{d}x\mathrm{d}y = \mathrm{d}\xi\mathrm{d}\eta$. Then the Radon transform $\mathscr{R}[f](\alpha, \xi)$ is defined in any of the three equivalent mathematical forms

$$\mathscr{R}[f](\alpha, \xi) =: P(\alpha, \xi) = \int\limits_{\eta=-\infty}^{\infty} f(x(\eta; \xi), y(\eta; \xi))\, \mathrm{d}\eta, \tag{14.1}$$

$$\begin{aligned} &= \iint\limits_{\xi,\eta} f(x(\eta; \xi), y(\eta; \xi))\delta(\xi - \mathbf{r}\cdot\hat{\boldsymbol{\xi}})\, \mathrm{d}\xi\mathrm{d}\eta \\ &= \iint\limits_{x,y} f(x, y)\delta(\xi - \mathbf{r}\cdot\hat{\boldsymbol{\xi}})\, \mathrm{d}x\mathrm{d}y. \end{aligned} \tag{14.2}$$

For a fixed α, the straight line is translated back and forth, which is represented by the value of the coordinate ξ, and each line gives rise to one value of $\mathscr{R}[f](\alpha, \xi)$ over $\infty < \xi < \infty$. To exhaust all points in the xy-plane it is clear that the angle $\alpha \in (0, \pi)$.

In applications, the function $f(x, y) > 0$ may represent attenuation (scattering plus absorption) of a high-frequency wave as it passes through a physical object and α may represent the direction of the incident wave. The Radon transform will then correspond to the normalized intensity at scan angle α after transmission through the object. If object is of finite support, then $P(\alpha, \xi)$ will also be of finite extent in ξ and will be zero beyond the confines of the object.

The 1D Fourier transform, $\widetilde{P}(\alpha, \kappa)$, of the projection $P(\alpha, \xi)$ w.r.t. ξ is[1]

$$\begin{aligned}\widetilde{P}(\alpha, \kappa) &= \frac{1}{2\pi}\int_{\xi=-\infty}^{\infty} P(\alpha, \kappa)\mathrm{e}^{-j\kappa\xi}\,\mathrm{d}\xi,\ -\infty < \kappa < \infty \\ &= \frac{1}{2\pi}\iint_{x,y} f(x, y)\int_{\xi=-\infty}^{\infty}\delta(\xi - \mathbf{r}\cdot\hat{\boldsymbol{\xi}})\mathrm{e}^{-j\kappa\xi}\,\mathrm{d}\xi\,\mathrm{d}x\mathrm{d}y \\ &= \frac{1}{2\pi}\iint_{x,y} f(x, y)\mathrm{e}^{-j\kappa\hat{\boldsymbol{\xi}}\cdot\mathbf{r}}\,\mathrm{d}x\mathrm{d}y = F(\boldsymbol{\kappa} = \hat{\boldsymbol{\xi}}\kappa),\end{aligned} \tag{14.3}$$

where

$$F(\boldsymbol{\kappa}) = \frac{1}{2\pi}\iint_{x,y} f(x, y)\mathrm{e}^{-j\boldsymbol{\kappa}\cdot\mathbf{r}}\,\mathrm{d}x\mathrm{d}y \tag{14.4}$$

is the 2D Fourier transform of $f(x, y)$ in the transform variables $\boldsymbol{\kappa} = \hat{\mathbf{x}}\kappa_x + \hat{\mathbf{y}}\kappa_y = \kappa_r(\hat{\mathbf{x}}\cos\beta + \hat{\mathbf{y}}\sin\beta)$, $-\infty < \kappa_x, \kappa_y, \kappa < \infty$, $0 \leqslant \kappa_r < \infty$, $0 \leqslant \beta \leqslant 2\pi$. It may be noted that since κ ranges over $(-\infty, \infty)$, it is not the same as the radial wavenumber, κ_r, that one associates with a 2D Fourier transform. In particular, $\kappa_r = \sqrt{\kappa_x^2 + \kappa_y^2} \neq \kappa$, but $\kappa_r = |\kappa|$. However, with $\kappa_x = \kappa\cos\alpha$ and $\kappa_y = \kappa\sin\alpha$, $0 \leqslant \alpha \leqslant \pi$, $-\infty < \kappa < \infty$, all points in the $\boldsymbol{\kappa}$-space are exhausted, but $\alpha \neq \beta$. Equation (14.3) states that *the 1D Fourier transform, $\widetilde{P}(\alpha, \kappa)$, of the projection, $P(\alpha, \xi)$, of a function, $f(x, y)$, at an angle α obtained through the Radon transform is the slice at angle α of the 2D Fourier transform $F(\boldsymbol{\kappa})$ of $f(x, y)$*. This is known as the *projection slice theorem*. This theorem allows us to retrieve the function $f(x, y)$ through the inverse Fourier transform

$$f(x, y) = \frac{1}{2\pi}\iint_{\kappa_x,\kappa_y=-\infty}^{\infty} F(\boldsymbol{\kappa})\mathrm{e}^{j\boldsymbol{\kappa}\cdot\mathbf{r}}\,\mathrm{d}\boldsymbol{\kappa} = \frac{1}{2\pi}\int_{\kappa_r=0}^{\infty}\int_{\beta=0}^{2\pi}\kappa_r F(\boldsymbol{\kappa})\mathrm{e}^{j\kappa_r(x\cos\beta + y\sin\beta)}\,\mathrm{d}\beta\mathrm{d}\kappa_r. \tag{14.5}$$

Note that $\boldsymbol{\kappa} = \kappa_r(\hat{\mathbf{x}}\cos\beta + \hat{\mathbf{y}}\sin\beta) = (-\kappa_r)(\hat{\mathbf{x}}\cos(\beta - \pi) + \hat{\mathbf{y}}\sin(\beta - \pi))$. Therefore the integral in equation (14.5) may be rewritten as

---

[1] We put the factor $1/2\pi$ in the 1D forward transform here so that the eventual result involving 2D forward and inverse transforms will be symmetric in the factor $1/2\pi$.

$$
\begin{aligned}
f(x,\,y) &= \tfrac{1}{2\pi}\int\limits_{\kappa_r=0}^{\infty}\left[\int\limits_{\beta=0}^{\pi} + \int\limits_{\beta=\pi}^{2\pi}\right]\kappa_r F(\boldsymbol{\kappa})\mathrm{e}^{j\kappa_r(x\cos\beta+y\sin\beta)}\,\mathrm{d}\beta\mathrm{d}\kappa_r \\
&= \tfrac{1}{2\pi}\int\limits_{\kappa_r=0}^{\infty}\int\limits_{\beta=0}^{\pi}\kappa_r F(\boldsymbol{\kappa})\mathrm{e}^{j\kappa_r(x\cos\beta+y\sin\beta)}\,\mathrm{d}\beta\mathrm{d}\kappa_r + \\
&\quad + \tfrac{1}{2\pi}\int\limits_{\kappa_r=0}^{\infty}\int\limits_{\beta=\pi}^{2\pi}(-\kappa_r)F(\boldsymbol{\kappa})\mathrm{e}^{j(-\kappa_r)[x\cos(\beta-\pi)+y\sin(\beta-\pi)]}\,\mathrm{d}\beta(-\mathrm{d}\kappa_r) \\
&= \tfrac{1}{2\pi}\int\limits_{\kappa=-\infty}^{\infty}\int\limits_{\alpha=0}^{\pi}|\,\kappa\,|\,F(\boldsymbol{\kappa})\mathrm{e}^{j\kappa(x\cos\alpha+y\sin\alpha)}\mathrm{d}\alpha\mathrm{d}\kappa \\
&= \tfrac{1}{2\pi}\int\limits_{-\infty}^{\infty}\int\limits_{\alpha=0}^{\pi}|\,\kappa\,|\,F(\boldsymbol{\kappa})\mathrm{e}^{j\kappa\hat{\boldsymbol{\xi}}\cdot\mathbf{r}}\,\mathrm{d}\alpha\mathrm{d}\kappa,
\end{aligned}
$$

where $\kappa_r = \kappa$, $\beta = \alpha$ was used in the first integral and $-\kappa_r = \kappa$, $\beta - \pi = \alpha$ was used in the second integral in the middle step to arrive at the end result. Combining this with the fact that $\widetilde{P}(\alpha,\,\kappa) = F(\boldsymbol{\kappa} = \hat{\boldsymbol{\xi}}\kappa)$ we finally obtain the inversion formula

$$
f(x,\,y) = \frac{1}{2\pi}\int\limits_{\alpha=0}^{\pi}\int\limits_{\kappa=-\infty}^{\infty}|\,\kappa\,|\,\widetilde{P}(\alpha,\,\kappa)\mathrm{e}^{j\kappa\hat{\boldsymbol{\xi}}\cdot\mathbf{r}}\,\mathrm{d}\kappa\mathrm{d}\alpha. \tag{14.6}
$$

14.2 Computed tomography

In x-ray tomography, $f(x,\,y)$ represents the attenuation function presented by the object being scanned such that the intensity at the exit plane, $I_t(\alpha,\,\xi)$, for illumination at the angle α is related to the incident intensity I_0 via

$$
\frac{I_t(\alpha,\,\xi)}{I_0} = \mathrm{e}^{-\left[\int f(x,\,y)\,d\eta\right]} = \mathrm{e}^{-P(\alpha,\,\xi)} \Longrightarrow P(\alpha,\,\xi) = -\ln\frac{I_t(\alpha,\,\xi)}{I_0}. \tag{14.7}
$$

If measurements of intensity are only available over some finite bandwidth $-W \leqslant \kappa \leqslant W$ through a filter with transfer function $H(\kappa)$, then the reconstructed attenuation function will be smeared by the filter response. The reconstructed attenuation function, $\hat{f}(x,\,y)$, is

$$
\hat{f}(x,\,y) = \frac{1}{2\pi}\int\limits_{0}^{\pi}\int\limits_{-W}^{W}|\,\kappa\,|\,H(\kappa)\widetilde{P}(\alpha,\,\kappa)\mathrm{e}^{j\kappa\hat{\boldsymbol{\xi}}\cdot\mathbf{r}}\,\mathrm{d}\kappa\mathrm{d}\alpha. \tag{14.8}
$$

Equation (14.8) is suitable for practical implementations once measurements of intensity at the exit plane are recorded. The Fourier transform of $P(\alpha,\,\xi)$ as well as the inverse transform in equation (14.8) can be conveniently evaluated using FFTs and inverse FFTs.

Substituting for $\widetilde{P}(\alpha, \xi)$ from equation (14.3), defining $R = \sqrt{(x - x')^2 + (y - y')^2}$, $(x - x') = R\cos\zeta$, $(y - y') = R\sin\zeta$, $\hat{\boldsymbol{\xi}} \cdot (\mathbf{r} - \mathbf{r}') = R\cos(\alpha - \zeta)$, utilizing the series expansion (C.71) and evaluating the integral w.r.t. α, the reconstructed attenuation function can be recast as

$$\begin{aligned}\hat{f}(x, y) &= \frac{1}{(2\pi)^2} \iint\limits_{x',y'} dx'dy' f(x', y') \int_{-W}^{W} |\kappa|\, H(\kappa) \int_{0}^{\pi} e^{j\kappa\hat{\boldsymbol{\xi}}\cdot(\mathbf{r}-\mathbf{r}')}\, d\alpha d\kappa \\ &= \frac{1}{(2\pi)^2} \iint\limits_{x',y'} dx'dy' f(x', y') \int_{-W}^{W} \left[\pi J_0(\kappa R) + 4 \sum_{m=0}^{\infty} \frac{J_{2m+1}(\kappa R)}{2m + 1} \sin(2m + 1)\zeta\right] \\ &\quad \times |\kappa|\, H(\kappa)\, d\kappa. \end{aligned} \tag{14.9}$$

If the filter function is symmetric in κ such that $H(-\kappa) = H(\kappa)$, then the series terms drops out on recalling that Bessel functions with odd order are odd functions of the argument. In that case

$$\hat{f}(x, y) = \frac{1}{4\pi} \iint\limits_{x',y'} f(x', y') \left[\int_{-W}^{W} |\kappa|\, H(\kappa) J_0(\kappa R)\, \mathrm{d}\kappa\right] \mathrm{d}x'\mathrm{d}y'. \tag{14.10}$$

The quantity within the square brackets in equation (14.10) is a function of $R = |\mathbf{r} - \mathbf{r}'|$. Thus the spatial integral in equation (14.10) is of convolutional type. For an ideal low-pass filter $H(\kappa) = 1, |\kappa| < W$. The integral involving Bessel function can then be evaluated using identity (C.61) with the result

$$\hat{f}(x, y) = \frac{W}{2\pi} \iint\limits_{x',y'} \frac{J_1(WR)}{R} f(x', y')\, \mathrm{d}x'\mathrm{d}y'. \tag{14.11}$$

Equation (14.11) gives $\hat{f}(x, y) = WJ_1(Wr)/2\pi r$ for a point attenuation function $f(x, y) = \delta(\mathbf{r})$. Note that $\lim\limits_{W\to\infty} WJ_1(Wr) = \delta(r)$, so that one recovers the exact point for infinite bandwidth.

Example 14.2.0.1. Circular object with uniform attenuation.
Determine the reconstructed object if the attenuation within a circular object of radius a is constant and an ideal low-pass filter is used in measurements.

Let us assume the constant attenuation to be unity. The most convenient form to proceed here is that given in equation (14.10). Using cylindrical coordinates, we have $f(x', y') = 1$, $\mathrm{d}x'\mathrm{d}y' = \rho'\mathrm{d}\rho'\mathrm{d}\phi'$, $R = [\rho^2 + \rho'^2 - 2\rho\rho'\cos(\phi' - \phi)]^{1/2}$. We next employ the addition theorem for Bessel functions (C.39), interchange the orders of integration in equation (14.10), evaluate the integral w.r.t. ϕ' and use the identity (C.61) to obtain the following expression for the reconstructed object

$$
\begin{aligned}
\hat{f}(\rho, \phi) &= \frac{1}{2\pi} \int_0^W \kappa d\kappa \int_0^a \rho' d\rho' \int_0^{2\pi} \Big[J_0(\kappa\rho) J_0(\kappa\rho') \\
&\qquad + 2 \sum_{n=1}^{\infty} J_n(\kappa\rho) J_n(\kappa\rho') \cos n(\phi - \phi') \Big] d\phi' \\
&= \int_0^W \kappa J_0(\kappa\rho) d\kappa \int_0^a \rho' J_0(\kappa\rho')\, d\rho' \qquad (14.12) \\
&= a \int_0^W J_1(\kappa a) J_0(\kappa\rho)\, d\kappa = \int_0^{Wa} J_1(x) J_0(xr)\, dx \\
&= \Theta(1 - r) - \int_{Wa}^{\infty} J_1(x) J_0(xr)\, dx,
\end{aligned}
$$

where the final expression in the last step follows from the identity [2, 6.512–3], $r = \rho/a$ and $\Theta(\cdot)$ is the unit step function. Clearly, for $Wa \to \infty$ the asymptotic integral on the rhs of equation (14.12) vanishes and one recovers the exact shape. Figure 14.2 shows the reconstructed object as a function of $r = \rho/a$ for two different filter bandwidths of $Wa = 2\pi$ and $Wa = 4\pi$ after carrying out the integral in equation (14.12) numerically. The smearing of the object shape due to band-limited

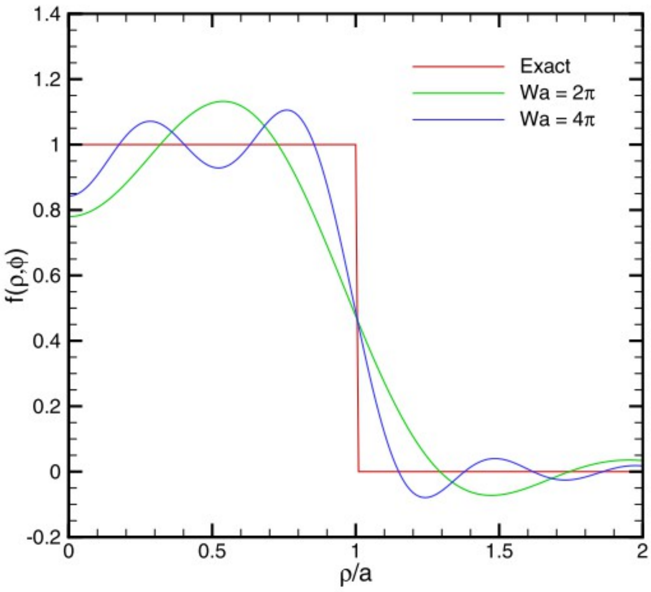

Figure 14.2. Exact object shape and reconstructed ones with $Wa = 2\pi$ and $Wa = 4\pi$ for a circular cylinder of radius a.

data is clearly evident. There is ringing as well as softening of edges in the reconstructed image with better results appearing for larger bandwidth of $Wa = 4\pi$ than for $Wa = 2\pi$. ■■

References

[1] Kak A C 1979 Computerized tomography with x-ray emission and ultrasound sources *Proc. IEEE* **67** 1245–72

[2] Gradshteyn I S and Ryzhik I M 2015 Tables of Integrals *Series and Products* 8th edn (New York: Academic)

I& @E '0

IOP Publishing

Engineering Electrodynamics
A collection of theorems, principles and field representations

Ramakrishna Janaswamy

Chapter 9

Geometrical optics and Fermat's principle

Maxwell's equations can only be solved exactly in the presence of material objects for a few well-defined geometries. For complicated and irregularly shaped objects one often resorts to numerical approaches. When the principal radii of curvature on an object size are large compared to wavelength, reasonably approximate solutions to Maxwell's equations can be obtained by the geometrical optics technique [1], which has been in existence for a number of centuries. Geometrical optics also provides a basis and bridge to diffraction optics, wherein one describes waves reaching regions in the shadow portion of a lit object by means of diffraction by geometric singularities and curvature [2].

The rigorous derivation of a geometrical optics field comprised of wavefronts and rays and the various laws, such as the eikonal equation, Sommerfeld–Runge law, Snell's law, intensity law, are all discussed in the present chapter. A widely occurring variational principle known as Fermat's principle, which has its origin in geometrical optics, is derived. With the recent emergence of electromagnetic metasurfaces for the control of reflected and transmitted electromagnetic energies, there has been a renewed interest in geometrical optics that witnessed some generalizations to the classical laws. The related topics of gradient metasurfaces and generalized Snell's law are also treated in this chapter.

9.1 Geometrical optics and Fermat's principle

The most rigorous and useful definition of geometrical optics is provided by Kline and Kay [3, pp 14–16] and involves the time-domain Maxwell's equations. The geometrical optics field at any point (x, y, z) of space is that field that corresponds to finite discontinuities[1] of the electric and magnetic fields, $\mathscr{E}$ and $\mathscr{H}$, on some hypersurface $\Phi(x, y, z; t) = \psi(x, y, z) - ct = 0$, where c is the speed of light in

---

[1] By finite discontinuities we mean singularities which are finite discontinuities with respect to the time variable of $\mathscr{E}$ and $\mathscr{H}$ and their successive time derivatives.

doi:10.1088/978-0-7503-1716-0ch9

© IOP Publishing Ltd 2020

free space. The family of surfaces $\psi(x, y, z) = ct$ are known as *wavefronts* and the normals erected on the wavefronts are termed *rays*, see figure 9.1. In the above representation, the hypersurface $\Phi(x, y, z; t) = 0$ is solved explicitly for t and cast in the form $\Phi(x, y, z; t) = \psi(x, y, z) - ct$ and the speed of light in free space is merely used to scale the time variable. In particular, c is *not* the speed of the wavefront.

9.1.1 Field discontinuities, wavefronts, and the eikonal equation

Consider a family of wavefronts $\psi = ct$ emanating from some source. At any specific time $t = t_0$, there is a point (x_0, y_0, z_0) on the wavefront such that the electric field $\mathscr{E}$ (and $\mathscr{H}$) is zero for $t < t_0$; at $t = t_0$, $\mathscr{E}$ jumps from zero to some non-zero value, and for $t > t_0$, $\mathscr{E}$ is non-zero. The geometrical optics field $(\tilde{\mathscr{E}}, \tilde{\mathscr{H}})$ exists at (x_0, y_0, z_0) *only* at the instant t_0. In particular, the values of $\mathscr{E}$ and $\mathscr{H}$ for $t > t_0$ at (x_0, y_0, z_0) belong to the time-dependent field satisfying Maxwell's equations, but not to the geometrical optics field. If we follow the geometrical optics field along a ray, that is, let x, y, and z take on values along a ray, then the corresponding value of t at which the geometrical optics field exists is given by $t = \psi(x, y, z)/c$. Hence

$$\tilde{\mathscr{E}}(x, y, z) = \mathscr{E}(x, y, z; \psi/c); \quad \tilde{\mathscr{H}}(x, y, z) = \mathscr{H}(x, y, z; \psi/c). \tag{9.1}$$

In the following we show that such a definition of a geometrical optics field follows rigorously from Maxwell's equations [3]. We would like to know what the wavefronts are and how the values of $(\widetilde{\mathbf{E}}, \widetilde{\mathbf{H}})$ vary along a ray. In this regard the generalizations of divergence and the curl theorems to four-dimensional spacetime are useful. Let G be a region in spacetime and let the hypersurface Γ be its boundary,

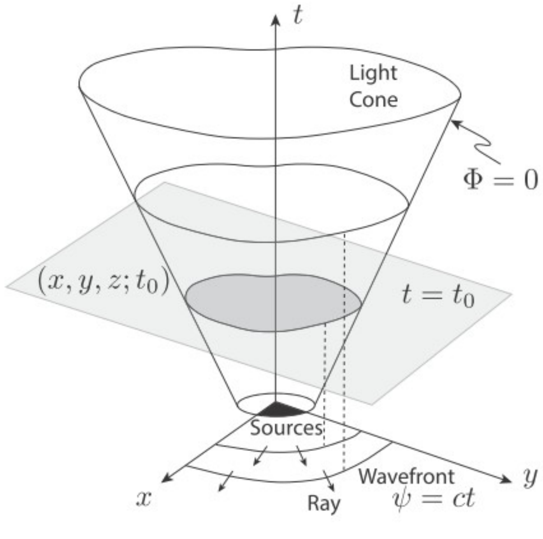

Figure 9.1. Wavefronts emanating from a finite-support source. Fields are zero outside the light cone and non-zero inside the light cone. The front across which the fields are discontinuous is described by $\Phi = 0$.

assumed to be continuous with piece-wise continuous derivatives, figure 9.2. The surface Γ is defined by $\chi(x, y, z; t) = 0$ and an outward unit normal $\hat{\mathbf{n}}$ has components

$$n_x = \lambda\hat{\mathbf{x}} \cdot \nabla\chi,\ n_y = \lambda\hat{\mathbf{y}} \cdot \nabla\chi,\ n_z = \lambda\hat{\mathbf{z}} \cdot \nabla\chi,\ n_t = \lambda\dot{\chi}, \tag{9.2}$$

where

$$\lambda = \pm[|\nabla\chi|^2 + \dot{\chi}^2]^{-1/2}, \tag{9.3}$$

and an over-dot represents a partial derivative w.r.t. t. The sign is λ is chosen to fit the requirement that the normal be directed outward from G. Let $\mathbf{m} = n_x\hat{\mathbf{x}} + n_y\hat{\mathbf{y}} + n_z\hat{\mathbf{z}} = \lambda\nabla\chi$ be the projection of $\hat{\mathbf{n}}$ onto the (x, y, z)-space. Note that $\mathbf{m}$ is not a unit vector in space coordinates. Let $\mathrm{d}w$ be an elemental volume of spacetime G and let $\mathrm{d}v$ be the elemental surface of the hypersurface Γ. Consider a vector field $\mathbf{A}(x, y, z; t) = \hat{\mathbf{x}}A_1(x, y, z; t) + \hat{\mathbf{y}}A_2(x, y, z; t) + \hat{\mathbf{z}}A_3(x, y, z; t)$ defined over G. Then [3, p 41]

$$\iiiint_G \nabla \times \mathbf{A}\, \mathrm{d}w = \iiint_\Gamma \mathbf{m} \times \mathbf{A}\, \mathrm{d}v, \tag{9.4}$$

$$\iiiint_G \nabla \cdot \mathbf{A}\, \mathrm{d}w = \iiint_\Gamma \mathbf{m} \cdot \mathbf{A}\, \mathrm{d}v, \tag{9.5}$$

where the curl and divergence operations are the usual three-dimensional curl and divergence.

Consider now the time-domain Maxwell's equations in an isotropic medium with electrical parameters $\varepsilon = \varepsilon_0\varepsilon_\mathrm{r}$, $\mu = \mu_0\mu_\mathrm{r}$, σ (non-dispersive) and with no magnetic sources

$$\nabla \times \mathscr{E} = -\dot{\mathscr{B}}; \quad \mathscr{B} = \mu\mathscr{H}, \quad \mathscr{D} = \varepsilon\mathscr{E}, \quad \mathscr{J}_\mathrm{c} = \sigma\mathscr{E}, \tag{9.6}$$

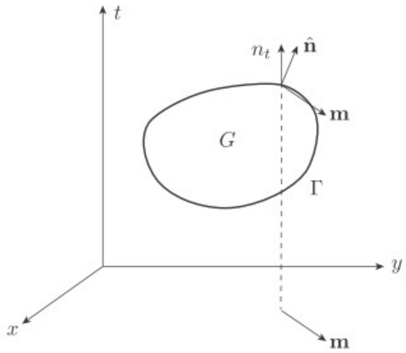

Figure 9.2. A four-dimensional domain G bounded by a surface Γ. The normal unit normal $\hat{\mathbf{n}}$ has a component n_t along the time-axis. The projection of $\hat{\mathbf{n}}$ onto space is the vector $\mathbf{m}$.

$$\nabla \times \mathscr{H} = \dot{\mathscr{D}} + \mathscr{J}_\mathrm{c} + \mathscr{J}_\mathrm{i}; \quad \mathscr{J}_\mathrm{i} = \dot{\mathscr{Q}}_\mathrm{e}, \quad \nabla \cdot \mathscr{Q}_\mathrm{e} = -q_\mathrm{ev}, \tag{9.7}$$

$$\nabla \cdot \mathscr{B} = 0; \quad \nabla \cdot (\mathscr{J}_\mathrm{i} + \mathscr{J}_\mathrm{c} + \dot{\mathscr{D}}) = 0, \tag{9.8}$$

where $\mathscr{Q}_\mathrm{e}$ is the electric steam potential introduced in equation (3.1) and we have chosen the magnetic stream potential $\mathscr{Q}_\mathrm{m} = 0$. The quantities $\mathscr{J}_\mathrm{i}$ and $\mathscr{J}_\mathrm{c}$ are, respectively, the impressed current density and the conduction current density. These equations can be cast in an integral form by making use of the above four-dimensional curl and divergence theorems. Applying equation (9.4) to the two Maxwell's curl equations and evaluating the integral with respect to time for time-derivative quantities yields

$$\iiint\limits_\Gamma [\mathbf{m} \times \mathscr{H} - n_t\mathscr{D} - n_t\mathscr{Q}_\mathrm{e}]\,\mathrm{d}v = \iiiint\limits_G \mathscr{J}_\mathrm{c}\,\mathrm{d}w, \tag{9.9}$$

$$\iiint\limits_\Gamma [\mathbf{m} \times \mathscr{E} + n_t\mathscr{B}]\,\mathrm{d}v = 0. \tag{9.10}$$

Applying equation (9.5) to the two Maxwell's divergence equations (9.8) leads to

$$\iiint\limits_\Gamma \mathbf{m} \cdot (\mathscr{J}_\mathrm{i} + \mathscr{J}_\mathrm{c} + \dot{\mathscr{D}})\,\mathrm{d}v = 0, \tag{9.11}$$

$$\iiint\limits_\Gamma \mathbf{m} \cdot \mathscr{B}\,\mathrm{d}v = 0. \tag{9.12}$$

Consider now a hypersurface $\Phi = 0$ that divides the region G into two subregions G_1 and G_2 as shown in figure 9.3. The boundary Γ is also subdivided into two parts Γ_1

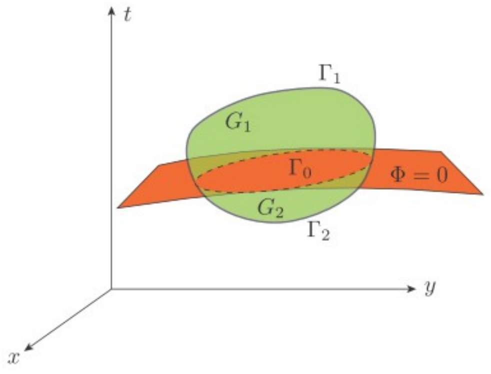

Figure 9.3. The spacetime domain G divided into a part G_1 with boundary $\Gamma_1 + \Gamma_0$ and a part G_2 with boundary $\Gamma_2 + \Gamma_0$ by the hypersurface $\Phi(x, y, z; t) = 0$. The normal on the common boundary Γ_0 points out of region G_i, $i = 1, 2$.

and Γ_2 such that $\Gamma_1 \cup \Gamma_2 = \Gamma$. The boundary of G_1 is the portion Γ_1 of Γ and the portion Γ_0 of $\Phi = 0$ which lies in G. Likewise the boundary of G_2 is Γ_2 and Γ_0. Note that if the vector $\mathbf{m} = \mathbf{m}_0$ associated with Γ_0 is taken as $+\lambda_0 \nabla\Phi$ for G_1, it will be $-\lambda_0 \nabla\Phi$ for G_2, where $\lambda_0 = [|\nabla\Phi|^2 + \dot{\Phi}^2]^{-1/2}$ and $n_t = \lambda_0 \dot{\Phi}$. We assume that the field vectors are discontinuous on Γ_0. Let $(\mathscr{E}_i, \mathscr{H}_i, \mathscr{D}_i, \mathscr{B}_i, \mathscr{Q}_{ei})$ denote the values of the various field components when approaching Γ_0 through G_i, $i = 1, 2$.

We now apply the integral relations (9.9)–(9.12) to (i) the domain G bounded by $\Gamma_1 \cup \Gamma_2$, both associated with $\mathbf{m} = \lambda \nabla\chi$, $\lambda = [|\nabla\chi|^2 + \dot{\chi}^2]^{-1/2}$, (ii) the domain G_1 bounded by $\Gamma_1 \cup \Gamma_0$, and (iii) the domain G_2 bounded by $\Gamma_2 \cup \Gamma_0$. Application to equation (9.9) yields

$$\iiint\limits_{\Gamma_1+\Gamma_2} [\mathbf{m} \times \mathscr{H} - n_t \mathscr{D} - n_t \mathscr{Q}_\mathrm{e}]\, \mathrm{d}v = \iiiint\limits_{G} \mathscr{J}_\mathrm{c}\, \mathrm{d}w, \tag{9.13}$$

$$\begin{aligned} &\iiint\limits_{\Gamma_1} [\mathbf{m} \times \mathscr{H} - n_t \mathscr{D} - n_t \mathscr{Q}_\mathrm{e}]\, \mathrm{d}v \\ &\quad + \iiint\limits_{\Gamma_0} [\nabla\Phi \times \mathscr{H}_1 - \dot{\Phi}\mathscr{D}_1 - \dot{\Phi}\mathscr{Q}_{\mathrm{e}1}]\lambda_0\, \mathrm{d}v = \iiiint\limits_{G_1} \mathscr{J}_\mathrm{c}\, \mathrm{d}w, \end{aligned} \tag{9.14}$$

$$\begin{aligned} &\iiint\limits_{\Gamma_2} [\mathbf{m} \times \mathscr{H} - n_t \mathscr{D} - n_t \mathscr{Q}_\mathrm{e}]\, \mathrm{d}v \\ &\quad - \iiint\limits_{\Gamma_0} [\nabla\Phi \times \mathscr{H}_2 - \dot{\Phi}\mathscr{D}_2 - \dot{\Phi}\mathscr{Q}_{\mathrm{e}2}]\lambda_0\, \mathrm{d}v = \iiiint\limits_{G_2} \mathscr{J}_\mathrm{c}\, \mathrm{d}w. \end{aligned} \tag{9.15}$$

Subtracting equation (9.13) from the sum of equations (9.14) and (9.15) results in

$$\iiint\limits_{\Gamma_0} \left\{\nabla\Phi \times \Delta[\mathscr{H}] - \dot{\Phi}\, \Delta[\mathscr{D}] - \dot{\Phi}\Delta[\mathscr{Q}_\mathrm{e}]\right\}\lambda_0\, \mathrm{d}v = 0, \tag{9.16}$$

where $\Delta[\mathscr{H}] = \mathscr{H}_1 - \mathscr{H}_2$ is the discontinuity in $\mathscr{H}$ across the hypersurface $\Phi = 0$ and similarly the other quantities. Since G can be taken arbitrarily small to enclose an arbitrary small section of Γ_0, the integrand of equation (9.16) must be zero. Thus

$$\nabla\Phi \times \Delta[\mathscr{H}] - \dot{\Phi}\Delta[\mathscr{D}] - \dot{\Phi}\Delta[\mathscr{Q}_\mathrm{e}] = 0 \quad \text{on } \Phi = 0. \tag{9.17}$$

Repeating the process for equations (9.10)–(9.12) we obtain the additional discontinuity relations

$$\nabla\Phi \times \Delta[\mathscr{E}] + \dot{\Phi}\Delta[\mathscr{B}] = 0 \quad \text{on } \Phi = 0, \tag{9.18}$$

$$\nabla\Phi \cdot \left\{\Delta[\mathscr{J}_\mathrm{i}] + \Delta[\mathscr{J}_\mathrm{c}] + \Delta[\dot{\mathscr{D}}]\right\} = 0 \quad \text{on } \Phi = 0, \tag{9.19}$$

$$\nabla\Phi \cdot \Delta[\mathscr{B}] = 0 \quad \text{on } \Phi = 0. \tag{9.20}$$

When these discontinuity equations are applied to a static boundary separating two material media, one obtains the usual boundary conditions on the tangential components of $(\mathscr{E}, \mathscr{H})$ and the normal components of $(\mathscr{D}, \mathscr{B}, \mathscr{J})$. When they are applied to a special hypersurface $\Phi = t = 0$ $(\Longrightarrow \nabla\Phi = 0, \dot{\Phi} = 1)$ assuming that the source function $\mathscr{Q}_\mathrm{e} = 0$ for $t < 0$ (which implies that the fields $\mathscr{E}$ and $\mathscr{H}$ are also zero for $t < 0$) and not equal to 0 for $t \geqslant 0$, one obtains from these discontinuity relations the initial conditions

$$\mathscr{E}(x, y, z; 0+) = -\frac{1}{\varepsilon}\mathscr{Q}_\mathrm{e}(x, y, z, 0+), \tag{9.21}$$

$$\mathscr{H}(x, y, z; 0+) = 0. \tag{9.22}$$

Thus the electric field responds instantaneously, but the magnetic field continues to be zero after the electric source is switched on. The situation is similar to the voltage response across a capacitor and the current response through an inductor when the elements are excited by a step charge source.

The more interesting case of field discontinuities that exist even in the absence of material discontinuities arises when the fields spread out from the source into the $(x, y, z; t)$-spacetime at finite speed. As before the source is assumed to be located within a finite spatial domain and switched on at $t = 0$. Because of this finite speed the field at $(x, y, z; t)$ will be zero until t is greater than some t_0 which is determined by its distance to the nearest source point. Hence there is a region of spacetime in which the field is non-zero and outside this a region in which the field is zero. In an isotropic medium (even in an inhomogeneous medium), the surface $\Phi(x, y, z; t) = 0$ separating the two regions will be a truncated light cone as shown in figure 9.1. The fields are zero outside the light cone and non-zero inside. The stream potential $\mathscr{Q}_\mathrm{e} = 0$ for any $t > 0$ on this light cone. The discontinuity relations (9.17) and (9.18) then yield the two homogeneous, linear vector equations for the quantities $\Delta[\mathscr{E}]$ and $\Delta[\mathscr{H}]$

$$\nabla\Phi \times \Delta[\mathscr{H}] - \varepsilon\dot{\Phi}\Delta[\mathscr{E}] = 0, \tag{9.23}$$

$$\nabla\Phi \times \Delta[\mathscr{E}] + \mu\dot{\Phi}\Delta[\mathscr{H}] = 0. \tag{9.24}$$

Taking a scalar product of these equations with $\nabla\Phi$ gives[2]

$$\nabla\Phi \cdot \Delta[\mathscr{E}] = 0 = \nabla\Phi \cdot \Delta[\mathscr{H}], \tag{9.25}$$

implying that $\Delta[\mathscr{E}] \perp \nabla\Phi$ and $\Delta[\mathscr{H}] \perp \nabla\Phi$. In addition taking a scalar product of equation (9.23) with $\Delta[\mathscr{H}]$ gives $\Delta[\mathscr{E}] \cdot \Delta[\mathscr{H}] = 0$, which implies that $\Delta[\mathscr{E}] \perp \Delta[\mathscr{H}]$. Hence the vectors $\Delta[\mathscr{E}]$, $\Delta[\mathscr{H}]$ and $\nabla\Phi$ are mutually perpendicular at any point on the hypersurface $\Phi = 0$. Taking a vector product of equation (9.24) by $\nabla\Phi$ and utilizing equation (9.23) gives

---

[2] Note that $\dot{\Phi} \neq 0$ since we are dealing with a dynamic hypersurface here.

$$\begin{aligned}&\nabla\Phi \times (\nabla\Phi \times \Delta[\mathscr{E}]) + \mu\varepsilon\dot{\Phi}^2\Delta[\mathscr{E}] = 0\\ &\Longrightarrow (-|\nabla\Phi|^2 + \mu\varepsilon\dot{\Phi}^2)\Delta[\mathscr{E}] = 0.\end{aligned} \tag{9.26}$$

Since $\Delta[\mathscr{E}] \neq 0$ we arrive at

$$|\nabla\Phi|^2 - \mu\varepsilon\dot{\Phi}^2 = 0 \quad \text{on } \Phi = 0. \tag{9.27}$$

This is a condition of the characteristic for the hyperbolic system of Maxwell's equations [4] and $\Phi = 0$ are known as the *characteristics*. Writing the light cone as $\Phi(x, y, z; t) = \psi(x, y, z) - ct = 0$ and noting that $\nabla\Phi = \nabla\psi$, $\dot{\Phi} = -c$, we finally arrive at the *eikonal differential equation of geometrical optics*

$$|\nabla\psi|^2 - c^2\mu\varepsilon = 0. \tag{9.28}$$

It is interesting to note that the eikonal equation is governed only by the permittivity and permeability of the medium. In particular, it is independent of the conductivity of the medium. Note that the derivation of this eikonal equation is exact and no approximations concerning field behavior have been made. The only assumption that has been made is that the media constants $(\varepsilon_r, \mu_r, \sigma)$ are all frequency independent. Since the fields are zero outside the light cone and non-zero just inside the light cone, the field discontinuities may well be taken to be $\Delta[\mathscr{E}] = \tilde{\mathscr{E}}$ and $\Delta[\mathscr{H}] = \tilde{\mathscr{H}}$, where $\tilde{\mathscr{E}}$ and $\tilde{\mathscr{H}}$ are defined in equation (9.1). In this sense, geometrical optics is concerned with the field vectors on the wavefronts as the wavefront spreads out into the (x, y, z)-space as time progresses. These specific values of field are indeed termed as *signals* in the literature [5][3].

9.1.2 Ray equations

As remarked previously, the normals erected on the family of wavefronts $\psi = ct$ are the rays[4]. The rays will have a direction proportional to $\nabla\psi$. Let $\mathbf{r}(t) = (x = x(t), y = y(t), z = z(t))$ be the equation of one of the rays passing through a point P on the family of wavefronts, figure 9.4. The velocity $\mathbf{v}$ of the wavefront is defined as the velocity with which the point P moves along the ray as t increases. Let us write

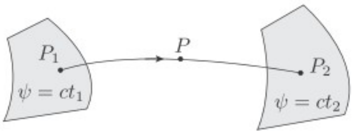

Figure 9.4. A ray emanating from a wavefront $\psi = ct_1$ at the point P_1, passing through an intermediate point P and terminating on the wavefront $\psi = ct_2$ at the point P_2.

---

[3] By considering a sinusoidal waveform modulated by a step function (a quasi time-periodic signal), Sommerfeld and Brillouin have shown that the signals travel with the velocity c even in a dispersive medium.

[4] Each value of t will generate a surface in ordinary space defined by $\psi = ct$.

$$\mathbf{v} = \frac{\mathrm{d}\mathbf{r}(t)}{\mathrm{d}t} =: v\frac{\nabla\psi}{|\nabla\psi|}, \tag{9.29}$$

where v is the speed to be determined. Along the ray we have

$$\psi(x(t), y(t), z(t)) = ct.$$

Differentiating this w.r.t. time we obtain upon using equations (9.28) and (9.29) that

$$\begin{aligned} c = \frac{\mathrm{d}\psi}{\mathrm{d}t} &= \frac{\partial\psi}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\partial\psi}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t} + \frac{\partial\psi}{\partial z}\frac{\mathrm{d}z}{\mathrm{d}t} \\ &= \nabla\psi\cdot\mathbf{v} = v\,|\nabla\psi| = vc\sqrt{\mu\varepsilon} \implies v = 1/\sqrt{\mu\varepsilon}\,. \end{aligned} \tag{9.30}$$

Thus $v = 1/\sqrt{\mu\varepsilon}$ is a speed associated with the wavefront $\Phi = 0$ and it depends only on the permittivity and permeability of the medium and is true even in the case of perfect or lossy conductors. The refractive index of the medium is $n = c/v = \sqrt{\mu\varepsilon/\mu_0\varepsilon_0} = \sqrt{\mu_r\varepsilon_r}$. In terms of the refractive index, the eikonal equation may be rewritten as

$$|\nabla\psi|^2 - n^2 = 0, \tag{9.31}$$

which is a non-linear first order partial differential equation. We introduce a unit vector along the ray as $\hat{\mathbf{p}} = \nabla\psi/|\nabla\psi| = \nabla\psi/n$ and define the *wave normal*

$$\mathbf{p} = \nabla\psi = \hat{\mathbf{p}}\,|\nabla\psi| = n\hat{\mathbf{p}}. \tag{9.32}$$

Note that the wave normal (a.k.a. *ray vector*) is not a unit vector.

The directional derivative of the wavefront $\psi = ct$ ($\Longrightarrow \mathrm{d}\psi = c\mathrm{d}t$) with respect to distance ρ in any arbitrary direction with unit vector $\hat{\boldsymbol{\rho}}$ is

$$\frac{\mathrm{d}\psi}{\mathrm{d}\rho} = \frac{\partial\psi}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}\rho} + \frac{\partial\psi}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}\rho} + \frac{\partial\psi}{\partial z}\frac{\mathrm{d}z}{\mathrm{d}\rho} = c\frac{\mathrm{d}t}{\mathrm{d}\rho} = \frac{c}{v}v\frac{\mathrm{d}t}{\mathrm{d}\rho} = n\frac{\mathrm{d}s}{\mathrm{d}\rho}, \tag{9.33}$$

where $\mathrm{d}s/\mathrm{d}t = v = c/n$ is the speed of the wavefront and $\mathrm{d}s$ is the arclength along the orthogonal trajectory between two neighboring wavefronts. In general $\mathrm{d}s \leqslant \mathrm{d}\rho$ as clear geometrically from figure 9.5. The *optical path length* between P_1 and P_2 along this new curve is defined as

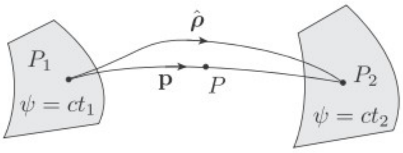

Figure 9.5. Any arbitrary path with a unit normal $\boldsymbol{\rho}$ between P_1 and P_2. The actual ray between P_1 and P_2 has a wave normal $\mathbf{p}$.

$$
\begin{aligned}
\psi(P_2) - \psi(P_1) &= \int_{P_1}^{P_2} \frac{\mathrm{d}\psi}{\mathrm{d}\rho}\,\mathrm{d}\rho = \int_{P_1}^{P_2} n(x,\, y,\, z)\mathrm{d}s = c\int_{P_1}^{P_2} \mathrm{d}t, \\
&= \int_{\rho_1}^{\rho_2} n(x(\rho),\, y(\rho),\, z(\rho))\sqrt{\dot{x}^2(\rho) + \dot{y}^2(\rho) + \dot{z}^2(\rho)}\,\mathrm{d}\rho,
\end{aligned}
\tag{9.34}
$$

where an overhead dot represents derivative with respect to argument, $\mathrm{d}s$ is the elemental arclength along the arbitrary curve with parameter ρ and $n\mathrm{d}s = c\mathrm{d}t$. The geometrical optics energy flux density vector is defined from the Poynting vector on the wavefront as

$$\tilde{\mathscr{S}} = \tilde{\mathscr{E}} \times \tilde{\mathscr{H}}.$$

From the discussion above $\tilde{\mathscr{S}}$ points along $\nabla\psi$ so that flux density vector can be written as $\tilde{\mathscr{S}} = \hat{\mathbf{p}}\widetilde{\mathcal{S}}$. Now carrying out (9.23) $\times\, \tilde{\mathscr{E}}$ − (9.24) $\times\, \tilde{\mathscr{H}}$ and making use of the following facts that (i) $\tilde{\mathscr{E}}$, $\tilde{\mathscr{H}}$, and $\nabla\psi$ are mutually perpendicular, (ii) the field energy density is $\widetilde{W} = \frac{1}{2}(\tilde{\mathscr{E}} \cdot \tilde{\mathscr{D}} + \tilde{\mathscr{H}} \cdot \tilde{\mathscr{B}})$ from the definition in theorem 5.2.1, (iii) $\nabla\Phi = \nabla\psi$, $\dot{\Phi} = -c$, and (iv) the interchange of the dot and cross in a vector triple product is valid, we obtain

$$\nabla\psi \cdot \tilde{\mathscr{S}} = c\widetilde{W} \Longrightarrow \tilde{\mathscr{S}} = \hat{\mathbf{p}}\frac{c}{|\nabla\psi|}\widetilde{W} = \hat{\mathbf{p}}v\widetilde{W}; \quad \widetilde{\mathcal{S}} = v\widetilde{W}. \tag{9.35}$$

The geometrical optics energy density vector $\widetilde{\mathcal{S}}$ carries energy along the rays together with $\tilde{\mathscr{E}}$, $\tilde{\mathscr{H}}$.

9.1.3 Ray tracing

It is possible to obtain differential equations satisfied by the rays from the eikonal equation. The evolution parameter along a ray could be time t or arclength s or a particular spatial coordinate or any other quantity τ. Different initial points $\mathbf{r}_0$ on a wavefront will generate different rays. Thus there will be a two-parameter (two initial coordinates contained in $\mathbf{r}_0$) family of rays belonging to the family of wavefronts $\psi = ct$. Let us represent the spatial coordinates of a ray by

$$x = x(\tau), \quad y = y(\tau), \quad z = z(\tau).$$

Since the path of the ray is orthogonal to the wavefront (i.e. along $\nabla\psi$) we have the dynamical equations

$$\frac{\mathrm{d}x}{\mathrm{d}\tau} = \lambda(\tau)\psi_x, \tag{9.36}$$

$$\frac{\mathrm{d}y}{\mathrm{d}\tau} = \lambda(\tau)\psi_y, \tag{9.37}$$

$$\frac{\mathrm{d}z}{\mathrm{d}\tau} = \lambda(\tau)\psi_z, \tag{9.38}$$

where $\psi_x = \partial\psi/\partial x$ and so on and $\lambda(\tau) \neq 0$ is a quantity that depends on the choice of τ (or vice versa). Indeed from the eikonal equation $\psi_x^2 + \psi_y^2 + \psi_z^2 = n^2$. Hence

$$\left(\frac{\mathrm{d}x}{\mathrm{d}\tau}\right)^2 + \left(\frac{\mathrm{d}y}{\mathrm{d}\tau}\right)^2 + \left(\frac{\mathrm{d}z}{\mathrm{d}\tau}\right)^2 = \left(\frac{\mathrm{d}s}{\mathrm{d}\tau}\right)^2 = \lambda^2 n^2 = c\frac{\mathrm{d}t}{\mathrm{d}\tau}, \tag{9.39}$$

where $\mathrm{d}s = \sqrt{\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2}$ is the elemental arclength and the last form follows from taking the derivative w.r.t. τ on the wavefront $\psi(x(\tau),\ y(\tau),\ z(\tau)) = ct$ and making use of the eikonal equation and equations (9.36)–(9.38). The last form also suggests that $\mathrm{d}^2\tau/\mathrm{d}t^2 = 0$ since n, λ are assumed to be independent of time. In other words, τ has to be a linear function of the time variable. For instance, if one chooses $\lambda = 1$, then the last form in equation (9.39) implies that the corresponding τ is $\tau = ct/n^2$. On the other hand, if one chooses τ to be the arclength parameter s, then the lhs of equation (9.39) is unity and the corresponding λ is $\lambda(s) = 1/n(s)$ that can vary arbitrarily with position.

Our goal now is to eliminate the wavefront ψ from the ray equations. Dividing both sides of equation (9.36) with λ and taking the derivative with respect to τ gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\frac{1}{\lambda}\frac{\mathrm{d}x}{\mathrm{d}\tau}\right) &= \psi_{xx}\frac{\mathrm{d}x}{\mathrm{d}\tau} + \psi_{xy}\frac{\mathrm{d}y}{\mathrm{d}\tau} + \psi_{xz}\frac{\mathrm{d}z}{\mathrm{d}\tau}\\ &= \frac{\lambda}{2}\frac{\partial}{\partial x}\left(\psi_x^2 + \psi_y^2 + \psi_z^2\right) = \frac{\lambda}{2}\frac{\partial}{\partial x}n^2,\end{aligned}$$

where equations (9.36)–(9.38) are used in the top line on the rhs to obtain the bottom line. Similar equations are obtained for the y- and z-coordinates. Therefore the differential equations for the ray coordinates are

$$\frac{2}{\lambda}\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\frac{1}{\lambda}\frac{\mathrm{d}x}{\mathrm{d}\tau}\right) = \frac{\partial}{\partial x}n^2, \tag{9.40}$$

$$\frac{2}{\lambda}\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\frac{1}{\lambda}\frac{\mathrm{d}y}{\mathrm{d}\tau}\right) = \frac{\partial}{\partial y}n^2, \tag{9.41}$$

$$\frac{2}{\lambda}\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\frac{1}{\lambda}\frac{\mathrm{d}z}{\mathrm{d}\tau}\right) = \frac{\partial}{\partial z}n^2. \tag{9.42}$$

These are non-homogeneous, linear, second order ordinary differential equations with non-constant coefficients. The ray is determined by the refractive index of the medium and the initial conditions. Equation (9.39) constrains the three ray velocity components and will eliminate one of the six constants of integration arising from the solution of the ray equations. Furthermore, since the parameter τ can be chosen rather arbitrarily without changing the geometric trajectory of the ray, an additional constant of integration can be eliminated by an appropriate transformation of the parameter τ.

Example 9.1.3.1. Ray equations in cylindrical coordinates.

It is possible to cast the ray equations given in equations (9.40)–(9.42) for Cartesian coordinates in any other coordinate system. In this example we demonstrate this for a cylindrical system with coordinates (ρ, ϕ, z). To this end the total and partial derivatives need to be expressed in terms of the new coordinates. The transformation relations facilitating the conversion are $\partial\rho/\partial x = \cos\phi$, $\partial\rho/\partial y = \sin\phi$, $\partial\phi/\partial x = -\sin\phi/\rho$, $\partial\phi/\partial y = \cos\phi/\rho$ and the inverse relations are $\partial x/\partial\rho = \cos\phi$, $\partial x/\partial\phi = -\rho\sin\phi$, $\partial y/\partial\rho = \sin\phi$, $\partial y/\partial\phi = \rho\cos\phi$. Using these it is straightforward to establish that

$$\frac{\mathrm{d}x}{\mathrm{d}\tau} = \cos\phi\frac{\mathrm{d}\rho}{\mathrm{d}\tau} - \rho\sin\phi\frac{\mathrm{d}\phi}{\mathrm{d}\tau}, \tag{9.43}$$

$$\frac{\mathrm{d}y}{\mathrm{d}\tau} = \sin\phi\frac{\mathrm{d}\rho}{\mathrm{d}\tau} + \rho\cos\phi\frac{\mathrm{d}\phi}{\mathrm{d}\tau}, \tag{9.44}$$

$$\frac{\partial}{\partial x} = \cos\phi\frac{\partial}{\partial\rho} - \frac{\sin\phi}{\rho}\frac{\partial}{\partial\phi}, \tag{9.45}$$

$$\frac{\partial}{\partial y} = \sin\phi\frac{\partial}{\partial\rho} + \frac{\cos\phi}{\rho}\frac{\partial}{\partial\phi}. \tag{9.46}$$

Substituting these into equations (9.40) and (9.41) results in

$$\frac{2}{\lambda}\frac{\mathrm{d}}{\mathrm{d}\tau}\left[\frac{1}{\lambda}\left(\cos\phi\frac{\mathrm{d}\rho}{\mathrm{d}\tau} - \rho\sin\phi\frac{\mathrm{d}\phi}{\mathrm{d}\tau}\right)\right] = \cos\phi\frac{\partial n^2}{\partial\rho} - \frac{\sin\phi}{\rho}\frac{\partial n^2}{\partial\phi}, \tag{9.47}$$

$$\frac{2}{\lambda}\frac{\mathrm{d}}{\mathrm{d}\tau}\left[\frac{1}{\lambda}\left(\sin\phi\frac{\mathrm{d}\rho}{\mathrm{d}\tau} + \rho\cos\phi\frac{\mathrm{d}\phi}{\mathrm{d}\tau}\right)\right] = \sin\phi\frac{\partial n^2}{\partial\rho} + \frac{\cos\phi}{\rho}\frac{\partial n^2}{\partial\phi}, \tag{9.48}$$

Taking the sum of equation (9.47) multiplied by $\cos\phi$ and equation (9.48) multiplied by $\sin\phi$ and carrying out some simplification gives

$$\frac{2}{\lambda}\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\frac{1}{\lambda}\frac{\mathrm{d}\rho}{\mathrm{d}\tau}\right) - \frac{2\rho}{\lambda^2}\left(\frac{\mathrm{d}\phi}{\mathrm{d}\tau}\right)^2 = \frac{\partial n^2}{\partial\rho}. \tag{9.49}$$

Likewise (9.47) × $\sin\phi$ − (9.48) × $\cos\phi$ and simplifications give

$$\frac{2}{\lambda}\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\frac{\rho}{\lambda}\frac{\mathrm{d}\phi}{\mathrm{d}\tau}\right) + \frac{2}{\lambda^2}\frac{\mathrm{d}\phi}{\mathrm{d}\tau}\frac{\mathrm{d}\rho}{\mathrm{d}\tau} = \frac{1}{\rho}\frac{\partial n^2}{\partial\phi}. \tag{9.50}$$

Note that new ray equations (9.49) and (9.50) in the ρ–ϕ-plane are now coupled, non-linear, second order ordinary differential equations. Simultaneously, equation (9.39) is modified to

$$\left(\frac{\mathrm{d}\rho}{\mathrm{d}\tau}\right)^2 + \left(\rho\frac{\mathrm{d}\phi}{\mathrm{d}\tau}\right)^2 + \left(\frac{\mathrm{d}z}{\mathrm{d}\tau}\right)^2 = \lambda^2 n^2, \tag{9.51}$$

at all points along the ray. ■■

Example 9.1.3.2 Ray tracing in a standard atmosphere.
Consider the propagation of rays in a standard atmosphere where the refractive index in a certain geographical region changes only along the vertical axis, which is taken here as the z-axis. In a standard atmosphere, the refractive index varies linearly with height as [6, p 203], [7, p 112] $n(z) = n_0 + n_1 z$, where n_0 and $n_1 \neq 0$ are constants. For example, over the continental USA, a refractive index commonly used after correcting for the Earth's curvature is $n_0 = 1 + 289 \times 10^{-6}$, $n_1 = 118 \times 10^{-6}\ \mathrm{km}^{-1}$.

Let the initial conditions in ϕ be $\phi(\tau = 0) = 0 = \mathrm{d}\phi(\tau = 0)/\mathrm{d}\tau$. A valid solution for ϕ satisfying equation (9.50) along the ray is $\phi(\tau) = 0$ so that a ray initially in the meridian plane $\phi = 0$ continues to remain in it along its entire trajectory. Inserting this solution into equation (9.49) determines the ratio $\mathrm{d}\rho/\mathrm{d}\tau$ as

$$\frac{1}{\lambda}\frac{\mathrm{d}\rho}{\mathrm{d}\tau} = \frac{1}{c_0}, \tag{9.52}$$

where $c_0 \neq 0$ is some constant. Let us choose $\tau = \rho$ along the ray. Then the function λ reduces to the constant c_0 from equation (9.52). The z coordinate as a function of ρ along the ray is determined from equation (9.42):

$$\frac{\mathrm{d}^2 z}{\mathrm{d}\rho^2} - (c_0 n_1)^2 z = n_0 n_1 c_0^2. \tag{9.53}$$

The general solution of this second order differential equation is

$$z(\rho) = A\cosh(c_0 n_1 \rho) + B\sinh(c_0 n_1 \rho) - \frac{n_0}{n_1}.$$

If $z(\rho = 0) = z_0$ and $\mathrm{d}z/\mathrm{d}\rho|_{\rho=0} = \tan\theta_0$ are specified, then upon substituting into the above equation we obtain $A = z_0 + n_0/n_1$ and $B = \tan\theta_0/c_0 n_1$. Therefore the ray path is given by the closed-form expression

$$z(\rho) = \left(z_0 + \frac{n_0}{n_1}\right)\cosh(c_0 n_1 \rho) + \frac{\tan\theta_0}{c_0 n_1}\sinh(c_0 n_1 \rho) - \frac{n_0}{n_1}. \tag{9.54}$$

The constant c_0 can be determined from equation (9.51) which now reads

$$1 + \left(\frac{\mathrm{d}z}{\mathrm{d}\rho}\right)^2 = c_0^2 (n_0 + n_1 z)^2. \tag{9.55}$$

Evaluating this at $\rho = 0$ gives $c_0 = \pm\sec\theta_0/(n_0 + n_1 z_0)$. Since $z(\rho)$ is an even function of c_0, the sign of c_0 is immaterial. The special case of $\theta_0 = 0$ gives $c_0 = (n_0 + n_1 z_0)^{-1}$. Figure 9.6 shows the path of the ray for two different values of θ_0. Note that the scale of the range axis (that is, the ρ-axis) is in kilometers, while that of the altitude axis (that is, the z-axis) is in meters. For $\theta_0 = -1$ mrad, the ray is initially launched with a negative slope, but it eventually turns back up. Thus in both cases rays have a non-zero curvature and they eventually travel towards regions of higher refractive index values. ■■

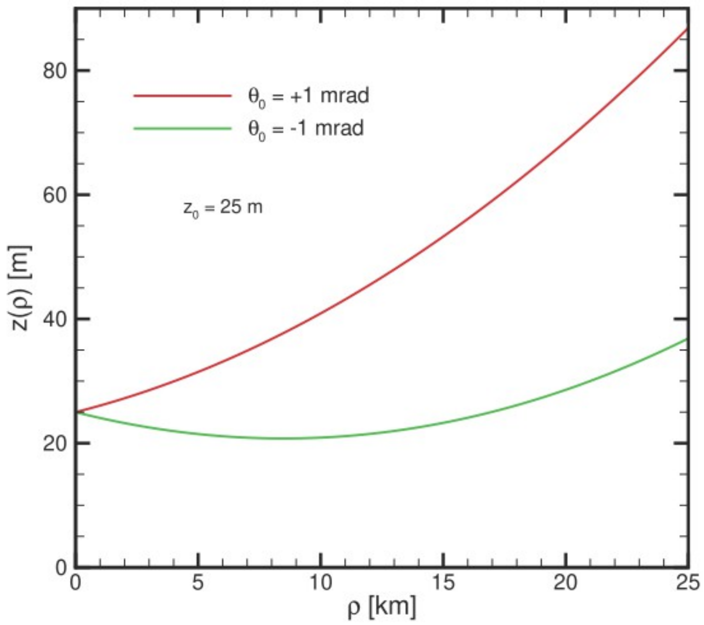

Figure 9.6. Ray tracing in a standard atmosphere with a refractive index $(n-1)\times 10^6 = 289 + 118z$.

Additional examples involving a geometrical optics interpretation of the solution of scattering by smooth convex bodies can be found in the books [8] and [2].

9.1.4 Mechanical interpretation of the ray path

Consider the special choice $\tau = ct/n^2$ (units of m), $\lambda = 1$ in the ray equations (9.39)–(9.40). Denoting the derivative w.r.t. τ with an overhead dot, the ray equations are

$$\ddot{x}(\tau) = \frac{\partial}{\partial x}\left(\frac{n^2}{2}\right), \quad \ddot{y}(\tau) = \frac{\partial}{\partial y}\left(\frac{n^2}{2}\right), \quad \ddot{z}(\tau) = \frac{\partial}{\partial z}\left(\frac{n^2}{2}\right). \tag{9.56}$$

The associated constraining equation (9.39) can be rewritten in the form

$$V = \frac{1}{2}(\dot{x}^2 + \dot{y}^2 + \dot{z}^2) = \frac{1}{2}n^2. \tag{9.57}$$

Equations (9.56) can be thought as the equations of motion of a particle ($m\mathbf{a} = \mathbf{F} = -\nabla U$, where m is the mass, $\mathbf{a}$ is the acceleration, $\mathbf{F}$ is the force, U is the potential energy) of unit mass subject to a potential energy $U = -\frac{1}{2}n^2$ (unitless). Then the kinetic energy of motion, V (unitless), is given by equation (9.57). The total energy of the particle $E = U + V = 0$. Thus electromagnetic rays in a medium of refractive index n can be thought of as the paths taken by particles which move with a total energy E in a potential field $U = -n^2/2$. The total energy of the particle is conserved at $E = 0$ as it moves along the path. Conversely, if a charged particle with a given energy C enters an electrostatic field with potential energy U then its kinetic energy $V = C - U$. We can then define a refractive index n from equation (9.57) as $n = \sqrt{2(C - U)}$ and consider ray paths by equation (9.56) whose trajectories will then mimic the actual paths taken by particles while conserving the total energy at C. ■■

9.1.5 Optical path length and Fermat's principle

With reference to figure 9.5, Fermat's principle states that [3, p 72], [1, p 128], [9, p 607]:

The rays between P_1 and P_2 are those curves along which the optical path length (9.34)

$$I = \int_{\sigma_1}^{\sigma_2} n(x(\sigma), y(\sigma), z(\sigma))\sqrt{\dot{x}^2(\sigma) + \dot{y}^2(\sigma) + \dot{z}^2(\sigma)}\; \mathrm{d}\sigma$$

is stationary with respect to infinitesimal variations in the path.

Proof. The integral I may be viewed as a cost function that depends not only on the 'particle' position (x, y, z) but also on its velocity $(\dot{x}, \dot{y}, \dot{z})$. The goal is to find the trajectory about which the cost function remains constant up to a first order in changes. Let us choose three functions $\xi(\sigma),\ \eta(\sigma),\ \zeta(\sigma)$ with continuous first order derivatives and which vanish at the end points

$$\xi(\sigma_1) = 0 = \xi(\sigma_2) = \eta(\sigma_1) = \eta(\sigma_2) = \zeta(\sigma_1) = \zeta(\sigma_2). \tag{9.58}$$

Let the extremal values we seek be x, y, z and the variations about the extremal values be $x + \varepsilon_1\xi,\ y + \varepsilon_2\eta,\ z + \varepsilon_3\zeta$, where $\varepsilon_1, \varepsilon_2, \varepsilon_3$ are small parameters. We can then view I as a function of $\varepsilon_i,\ i = 1, 2, 3$ on these varied paths:

$$I(\varepsilon_1, \varepsilon_2, \varepsilon_3) = \int_{\sigma_1}^{\sigma_2} F(x + \varepsilon_1\xi, y + \varepsilon_2\eta, z + \varepsilon_3\zeta; \dot{x} + \varepsilon_1\dot{\xi}, \dot{y} + \varepsilon_2\dot{\eta}, \dot{z} + \varepsilon_3\dot{\zeta})\, \mathrm{d}\sigma, \tag{9.59}$$

where $F(x, y, z; \dot{x}, \dot{y}, \dot{z}) = n(x, y, z)[\dot{x}^2 + \dot{y}^2 + \dot{z}^2]^{1/2}$. Note that all varied paths, including the extremal path, start and terminate at the same point owing to the boundary conditions in equation (9.58). The *first variation* of I w.r.t. x is defined as $(\delta I)_x = \partial I/\partial\varepsilon_1|_{\varepsilon_i=0, i=1,2,3}$ and is

$$\begin{aligned}(\delta I)_x &= \int_{\sigma_1}^{\sigma_2}\left(\xi\frac{\partial F}{\partial x} + \dot{\xi}\frac{\partial F}{\partial \dot{x}}\right)\mathrm{d}\sigma \\ &= \int_{\sigma_1}^{\sigma_2}\left(F_x - \frac{\mathrm{d}F_{\dot{x}}}{\mathrm{d}\sigma}\right)\xi\, \mathrm{d}\sigma,\end{aligned} \tag{9.60}$$

where integration by parts followed by the boundary conditions (9.58) were used in the second term on the rhs in the bottom row and we use the notation that $F_x = \partial F/\partial x$, $F_{\dot{x}} = \partial F/\partial \dot{x}$. Setting the first variation to zero for obtaining the extremum values and noting that ξ is arbitrary it follows that

$$\begin{aligned} F_x - \frac{\mathrm{d}F_{\dot{x}}}{\mathrm{d}\sigma} = 0 \Longrightarrow \frac{\partial n}{\partial x}\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2} - \frac{\mathrm{d}}{\mathrm{d}\sigma}\left[\frac{n\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}}\right] = 0 \\ \Longrightarrow \frac{\partial n}{\partial x}\frac{\mathrm{d}s}{\mathrm{d}\sigma} = \frac{\mathrm{d}}{\mathrm{d}\sigma}\left[n\frac{\mathrm{d}x}{\mathrm{d}\sigma}\frac{\mathrm{d}\sigma}{\mathrm{d}s}\right] \\ \Longrightarrow \frac{\partial n}{\partial x} = \frac{\mathrm{d}}{\mathrm{d}s}\left[n\frac{\mathrm{d}x}{\mathrm{d}s}\right]. \end{aligned}$$

The left-hand side equation above is known as the *Euler equation* for the variational problem being studied here. We also know from equation (9.39) that $\lambda(s) = [n(s)]^{-1}$. Therefore the last equation above may be also be written as

$$\frac{2}{\lambda(s)}\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{1}{\lambda(s)}\frac{\mathrm{d}x}{\mathrm{d}s}\right) = \frac{\partial n^2}{\partial x}, \tag{9.61}$$

which is precisely the ray equation (9.40) expressed in terms of the arclength ray coordinate $\tau = s$. The other two ray equations (9.41) and (9.42) follow similarly from setting the first variation with respect to y and z also to zero. ■

9.1.6 Sommerfeld–Runge and Snell's laws, the Lagrange integral invariant

On account of the vector identity curl grad $\equiv 0$, the wave normal $\mathbf{p} = \hat{\mathbf{p}}n$ defined in equation (9.32) satisfies

$$\nabla \times \mathbf{p} = \nabla \times (\hat{\mathbf{p}}n) = 0, \tag{9.62}$$

a fact first discovered by Sommerfeld and Runge [10] and later came to be known as the *Sommerfeld–Runge law*. One can construct a vector field $\hat{\mathbf{p}}(x, y, z)n(x, y, z)$ on knowing the refractive index and ray directions, which are in turn dictated by the initial conditions for launching rays (initial location and 'velocities of rays' as in example 9.1.3.1). The unit vector $\hat{\mathbf{p}}$ also has special properties. In a homogeneous medium n is a constant and equation (9.62) implies $\nabla \times \hat{\mathbf{p}} = 0$. Otherwise, since $\nabla \times n\hat{\mathbf{p}} = n\nabla \times \hat{\mathbf{p}} + \nabla n \times \hat{\mathbf{p}}$, taking a scalar product with $\hat{\mathbf{p}}$ yields

$$\hat{\mathbf{p}} \cdot \nabla \times \hat{\mathbf{p}} = 0. \tag{9.63}$$

Assuming that n is a continuous function of position, one may apply Stokes' theorem to an open surface S bounded by a closed curve C and deduce the *Lagrange's integral invariant* associated with this field $\hat{\mathbf{p}}n$

$$\oint_C n\hat{\mathbf{p}} \cdot \mathrm{d}\boldsymbol{\ell} = \iint_S \nabla \times (\hat{\mathbf{p}}n)\,\mathrm{d}s = 0. \tag{9.64}$$

Lagrange's integral invariant continues to hold good for piece-wise continuous media as the contour C can be deformed in that case to include material boundary as shown in figure 9.7, $C = (C_1 - K) \cup (C_2 + K)$. The contributions to the integral arising from the boundary K will be zero since the vector $(n_2\hat{\mathbf{p}}_2 - n_1\hat{\mathbf{p}}_1) \perp \mathrm{d}\boldsymbol{\ell}$ as is shown in example 9.1.6.1, where $\mathbf{p}_i$ is the wave normal at the interface in medium $i,\ i = 1, 2$.

The field $\mathbf{p} = \hat{\mathbf{p}}n$ thus acts like a conservative force field and implies that the integral

$$J(P_1, P_2) = \int_{P_1}^{P_2} n(x, y, z)\hat{\mathbf{p}}(x, y, z) \cdot \mathrm{d}\boldsymbol{\ell}, \tag{9.65}$$

taken between any two points P_1 and P_2 is independent of the path of integration (see figure 9.8).

Example 9.1.6.1. Snell's law.

In this example we start from the Sommerfeld–Runge law and derive the well known Snell's law in a stratified medium. Indeed we have from the vector identity $\nabla(\mathbf{a} \cdot \mathbf{b}) = \mathbf{a} \times \nabla \times \mathbf{b} + \mathbf{b} \times \nabla \times \mathbf{a} + (\mathbf{a} \cdot \nabla)\mathbf{b} + (\mathbf{b} \cdot \nabla)\mathbf{a}$ and equation (9.62) that

$$\begin{aligned}\hat{\mathbf{p}} \cdot \nabla \times \mathbf{p} = 0 &\Longrightarrow 2(\mathbf{p} \cdot \nabla)\mathbf{p} = \nabla(\mathbf{p} \cdot \mathbf{p})\\ &= 2n\nabla n \Longrightarrow (\hat{\mathbf{p}} \cdot \nabla)\mathbf{p} = \nabla n.\end{aligned} \tag{9.66}$$

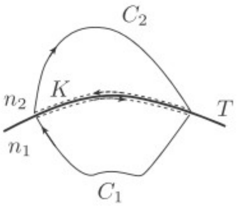

Figure 9.7. Lagrange's integral invariant in the presence of piece-wise continuous media.

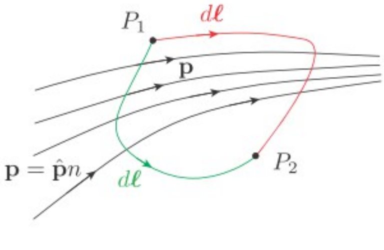

Figure 9.8. The integral $J(P_1, P_2)$ between two points is independent of the path taken.

Consider now a horizontally stratified medium with a refractive index changing only along one axis (say z). Then $\frac{\partial}{\partial x} = 0 = \frac{\partial}{\partial y}$ and $\nabla n = \hat{\mathbf{z}}\frac{\partial n}{\partial z}$. If the ray makes an angle $\theta(z)$ with the z-axis and has an azimuthal angle ϕ (independent of z), then in a local cylindrical coordinate system erected at any point $\hat{\mathbf{p}} = \hat{\boldsymbol{\rho}}\sin\theta + \hat{\mathbf{z}}\cos\theta$ and $(\hat{\mathbf{p}}\cdot\nabla)\mathbf{p} = \hat{\boldsymbol{\rho}}\cos\theta\frac{\partial}{\partial z}(n(z)\sin\theta(z)) + \hat{\mathbf{z}}\cos\theta\frac{\partial}{\partial z}(n(z)\cos\theta(z))$. Since ∇n has no radial component we conclude upon comparing both sides of equation (9.66) that

$$\frac{\partial}{\partial z}(n(z)\sin\theta(z)) = 0 \Longrightarrow n(z)\sin\theta(z) = \text{constant}, \tag{9.67}$$

which is a mathematical statement of Snell's law in a horizontally stratified medium. Note that the frequency of the wave does not explicitly enter into this equation. In the case of a boundary between two piece-wise continuous media with refractive indices n_1 and n_2 (see figure 9.7) and ray angles with respect to normal of θ_1 and θ_2, respectively, Snell's law reduces to the familiar result $n_1\sin\theta_1 = n_2\sin\theta_2$.

If on the two sides of the boundary K between two different homogeneous media (see figure 9.7) $\mathbf{p} = \hat{\mathbf{p}}_1 n_1$ and $\mathbf{p} = \hat{\mathbf{p}}_2 n_2$, then $\nabla\times\mathbf{p} = 0 \Longrightarrow (\hat{\mathbf{p}}_2 n_2 - \hat{\mathbf{p}}_1 n_1)\cdot \mathrm{d}\ell = 0$ from an application of Stokes' theorem to a closed rectangular contour consisting of the two dashed lines across K in figure 9.7 connected by vanishingly small vertical segments. Therefore the vector $(\hat{\mathbf{p}}_2 n_2 - \hat{\mathbf{p}}_1 n_1)$ has no tangential component at the interface, or, equivalently, it is perpendicular to the interface.

9.1.7 Transport of the geometrical optics field

The propagation of electric and magnetic field vectors and that of the Poynting vector in geometrical optics is most easily carried out in the frequency domain. We will first derive the transport equation for the time-harmonic electric field $\mathbf{E}$. The equation for the magnetic field follows from duality. From Maxwell's equations in a source-free and non-conducting region we have $\nabla\times\nabla\times\mathbf{E} - k^2\mathbf{E} = 0$, $\nabla\cdot(\varepsilon\mathbf{E}) = 0 = \varepsilon\nabla\cdot\mathbf{E} + \mathbf{E}\cdot\nabla\varepsilon$, where $k = k_0 n$, k_0 is the free-space wavenumber and n is the refractive index. Using the vector identity $\nabla\times\nabla\times\mathbf{E} = \nabla\nabla\cdot\mathbf{E} - \nabla^2\mathbf{E}$ and the above divergence equation we obtain the modified Helmholtz equation for the electric field

$$\nabla^2\mathbf{E} + k^2\mathbf{E} = -\nabla\left(\mathbf{E}\cdot\frac{\nabla\varepsilon}{\varepsilon}\right). \tag{9.68}$$

Let us consider the normalized space variables $\xi = k_0 x,\ \eta = k_0 y,\ \zeta = k_0 z$ and the corresponding gradient operator $\tilde{\nabla}$. In the normalized variables equation (9.68) reads

$$\tilde{\nabla}^2\mathbf{E} + n^2\mathbf{E} = -\tilde{\nabla}\left(\mathbf{E}\cdot\frac{\tilde{\nabla}\varepsilon}{\varepsilon}\right). \tag{9.69}$$

If the permittivity has a characteristic length scale L (that is, the permittivity varies non-trivially over a distance L)[5] then $|\widetilde{\nabla}\varepsilon|/\varepsilon \sim (k_0 L)^{-1}$. Consequently, for $k_0 L \gg 1$, the rhs of equation (9.69) and hence that of equation (9.68) may be ignored. Therefore for a slowly varying permittivity function the electric field satisfies the homogeneous vector Helmholtz equation

$$\nabla^2\mathbf{E} + n^2 k_0^2 \mathbf{E} \approx 0, \quad k_0 L \gg 1. \tag{9.70}$$

In a Cartesian coordinate system, each component of the electric field satisfies the scalar Helmholtz equation $\nabla^2 E + n^2 k_0^2 E = 0$. We now express each field component as

$$E(\mathbf{r}) = \mathrm{e}^{-jk_0 \mathscr{S}(\mathbf{r})} E_0(\mathbf{r}), \tag{9.71}$$

where the phase function $\mathscr{S}(\mathbf{r})$ (units of m) and the amplitude function $E_0(\mathbf{r})$ (units of V m^{-1}) are assumed to be slowly varying functions of position and independent of k_0. Now

$$\begin{aligned}\nabla^2 E = \nabla \cdot \nabla E = \big[&\nabla^2 E_0 - 2jk_0 \nabla\mathscr{S} \cdot \nabla E_0 - jk_0 E_0 \nabla^2 \mathscr{S} \\ &- k_0^2 (\nabla\mathscr{S})^2 E_0\big]\mathrm{e}^{-jk_0\mathscr{S}},\end{aligned}$$

where $(\nabla\mathscr{S})^2 = \nabla\mathscr{S} \cdot \nabla\mathscr{S}$. Substituting into the scalar Helmholtz equation we obtain

$$[(\nabla\mathscr{S})^2 - n^2]E_0 + \frac{j}{k_0}\big(2\nabla\mathscr{S} \cdot \nabla E_0 + E_0 \nabla^2\mathscr{S}\big) - \frac{1}{k_0^2}\nabla^2 E_0 = 0. \tag{9.72}$$

We choose the phase function $\mathscr{S}$ such that the term within the square bracket above vanishes. Then

$$(\nabla\mathscr{S})^2 = n^2, \tag{9.73}$$

which is seen to be the same equation as the eikonal equation satisfied by the function ψ in equation (9.31). Hence the family of surfaces $\mathscr{S} =$ constant constitute the wavefronts. Indeed we set $\mathscr{S} = \psi$ in the ensuing analysis. With this choice of the phase function, the amplitude function now satisfies

$$(2\nabla\psi \cdot \nabla E_0 + E_0 \nabla^2 \psi) = \frac{1}{jk_0}\nabla^2 E_0. \tag{9.74}$$

If the amplitude function has a length scale of ℓ, then $\nabla^2 E_0 \sim \mathcal{O}(\ell^{-2})$. At high enough frequencies such that $k_0 \ell \gg 1$, the term on the rhs of equation (9.74) may then be ignored. Therefore the amplitude function satisfies the simpler equation

$$\nabla\psi \cdot \nabla \ln(E_0) = -\frac{1}{2}\nabla^2\psi, \tag{9.75}$$

---

[5] Note that if ε varies as $\gamma_0\, \mathrm{e}^{-x/L_x}$ in the x-direction, then $\varepsilon^{-1}\partial\varepsilon/\partial\xi = (k_0 L_x)^{-1}$.

where we have utilized the fact that $\nabla \ln(E_0) = E_0^{-1}\nabla E_0$. Since $\nabla\psi = \mathbf{p} = \hat{\mathbf{p}}n$ (see equation (9.32)) the above equation for determining the amplitude function can also be written as

$$\hat{\mathbf{p}} \cdot \nabla \ln(E_0) = -\frac{1}{2n}\nabla \cdot \mathbf{p} = -\frac{1}{2n}\nabla \cdot (\hat{\mathbf{p}}n). \tag{9.76}$$

Similar equations are obtained for the components of a magnetic field vector. Once the gradient of the wavefront is calculated from the medium refractive index, the amplitude E_0 along the ray can be determined using equation (9.76). This procedure of obtaining the transport equations for field vectors is equivalent to extracting the leading term after expanding the time-harmonic fields ($\mathbf{E}(\mathbf{r})$, $\mathbf{H}(\mathbf{r})$) in inverse powers of the free-space wavenumber k_0 as done by Luneberg [11] and Kline [12].

The wave intensity I is defined as the absolute value of the time average of the Poynting vector and is $I = \frac{1}{2}\Re(\mathbf{E} \times \mathbf{H}^*) \cdot \hat{\mathbf{p}}$. In a source-free and non-conducting region we have from equation (5.17) that

$$\nabla \cdot (I\hat{\mathbf{p}}) = 0. \tag{9.77}$$

Consider a narrow tube formed by rays exiting an elemental area $\mathrm{d}S_1$ of a wavefront $\psi = c_1$ and arriving at an elemental area $\mathrm{d}S_2$ of the wavefront $\psi = c_2$, where c_1 and c_2 are constants. The rays are all parallel to the sidewalls of the tube as indicated in figure 9.9. Integrating the divergence equation (9.77) over the volume comprised by the tube and applying Gauss' divergence theorem we obtain

$$I_2\mathrm{d}S_2 = I_1\mathrm{d}S_1, \tag{9.78}$$

after noting that $\hat{\mathbf{n}} \cdot \hat{\mathbf{p}} = 0$ on the portion of surface along the tube, $\hat{\mathbf{n}} \cdot \hat{\mathbf{p}} = -1$ on $\mathrm{d}S_1$, $\hat{\mathbf{n}} \cdot \hat{\mathbf{p}} = +1$ on $\mathrm{d}S_2$ and I_1 and I_2 are the intensities on $\mathrm{d}S_1$ and $\mathrm{d}S_2$, respectively. Hence $I\mathrm{d}S$ remains constant along a tube of rays. The intensity decreases along a ray if the rays diverge between two wavefronts and intensity increases if the rays converge. It is possible to obtain an explicit expression for the intensity of the ray in terms of the wavefront $\psi =$ constant. We know from equation (9.32) that $\hat{\mathbf{p}} = (\nabla\psi)/n$. Therefore from $\nabla \cdot (\hat{\mathbf{p}}I) = 0$ we have

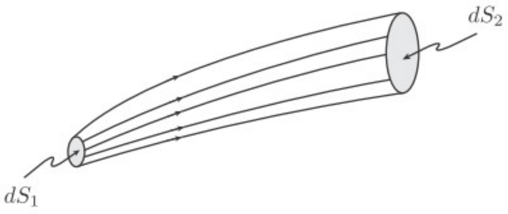

Figure 9.9. A tube of rays connecting two elemental areas $\mathrm{d}S_1$ and $\mathrm{d}S_2$ in an isotropic medium.

$$\left(\frac{I}{n}\right)\nabla^2\psi + \nabla\psi\cdot\nabla\left(\frac{I}{n}\right) = 0 \Longrightarrow \frac{1}{n(s)}\nabla\psi\cdot\nabla\ln\left(\frac{I}{n}\right) = -\frac{1}{n(s)}\nabla^2\psi, \tag{9.79}$$

where a factor $1/n(s)$, s = arclength parameter along the ray, was included on both sides in the last equation for convenience. We also know from (9.36)–(9.38) that when the arbitrary parameter τ along the ray is chosen to be the arclength parameter the quantity $\lambda(\tau)$ there becomes equal to $n^{-1}(s)$. Therefore along the ray with a parametric representation $x = x(s),\ \ y = y(s),\ \ z = z(s)$, the operator

$$\frac{1}{n(s)}\nabla\psi\cdot\nabla = \frac{\mathrm{d}x(s)}{\mathrm{d}s}\frac{\partial}{\partial x} + \frac{\mathrm{d}y(s)}{\mathrm{d}s}\frac{\partial}{\partial y} + \frac{\mathrm{d}z(s)}{\mathrm{d}s}\frac{\partial}{\partial z} = \frac{\mathrm{d}}{\mathrm{d}s},$$

in view of equations (9.36)–(9.38). Using this in equation (9.79) leads to

$$\frac{\mathrm{d}}{\mathrm{d}s}\ln\left(\frac{I(s)}{n(s)}\right) = -\frac{\nabla^2\psi}{n(s)}. \tag{9.80}$$

Integrating this between two points gives the intensity law for geometrical optics

$$\frac{I(s_2)}{I(s_1)} = \frac{n(s_2)}{n(s_1)}\mathrm{e}^{-\int_{s_1}^{s_2}\frac{\nabla^2\psi}{n(s)}\mathrm{d}s}. \tag{9.81}$$

The integral inside the exponent of equation (9.81) forms the basis for defining Radon transform (see chapter 14), which has utility in x-ray computed tomography.

Example 9.1.7.1. Intensity law in a homogeneous medium.
We would like to show in this example the intensity law for geometrical optics in a homogeneous medium. In a homogeneous medium the refractive index is constant and the rays propagate along straight lines. It is sufficient to consider the case of a vacuum ($n = 1$). Let us assume that the principal radii of curvature of the wavefront at $s = 0$ are given to be $\rho_1(s = 0) = \rho_{10}$ and $\rho_2(s = 0) = \rho_{20}$ and that the intensity at $s = 0$ is unity. For convenience we take $s = z$ and express the wavefront in terms of a local Cartesian system erected around the central ray (in the so-called Dupin coordinate system). We can approximately construct the wavefront as a quadratic function in terms of the radii of curvature as

$$\psi(x,\,y,\,z):\ z + \frac{x^2}{2\rho_1(z)} + \frac{y^2}{2\rho_2(z)} = 0,$$

see figure 9.10. If ρ_1 and ρ_2 are positive the wavefront will be convex in the direction of propagation as shown in the figure. If they are negative, the wavefront will be concave in the direction of propagation. For this wavefront

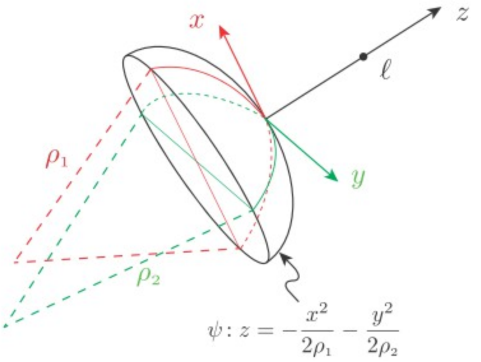

Figure 9.10. Wavefront having principal radii of curvature ρ_1 and ρ_2.

$$\nabla\psi = \hat{\mathbf{x}}\frac{x}{\rho_1(z)} + \hat{\mathbf{y}}\frac{y}{\rho_2(z)} + \hat{\mathbf{z}}\left(1 - \frac{x^2}{2\rho_1^2(z)}\rho_1'(z) - \frac{y^2}{2\rho_2^2(z)}\rho_2'(z)\right).$$

For the eikonal equation $|\nabla\psi|^2 = 1$ to be satisfied we require

$$\frac{x^2}{\rho_1^2}(1 - \rho_1') + \frac{y^2}{\rho_2^2}(1 - \rho_2') + \frac{1}{4}\left[\frac{x^2}{\rho_1^2}\rho_1' + \frac{y^2}{\rho_2^2}\rho_2'\right]^2 = 0.$$

Choosing $\rho_1' = 1 = \rho_2'$ yields $\rho_1(z) = \rho_{10} + z$; $\rho_2(z) = \rho_{20} + z$. However with this choice, the eikonal equation is satisfied exactly only for the central ray ($x = 0,\ y = 0$) and up to fourth order accuracy (that is the error will be proportional to $\mathcal{O}[(x/\rho_1)^4]$, $\mathcal{O}[(y/\rho_2)^4]$, $\mathcal{O}[(xy/\rho_1\rho_2)^2]$ for the other rays. Our main focus here is to derive the intensity law for the central ray and in that respect the above wavefront model is valid. The Laplacian of the wavefront in vacuum is

$$\begin{aligned}\nabla^2\psi &= \frac{1}{\rho_1(z)} + \frac{1}{\rho_2(z)} - \left(\frac{x^2}{\rho_1^3(z)} + \frac{y^2}{\rho_2^3(z)}\right)\\ &= \frac{1}{\rho_1(z)} + \frac{1}{\rho_2(z)} + \mathcal{O}\left(\frac{x^2}{\rho_1^3(z)}, \frac{y^2}{\rho_2^3(z)}\right)\\ &\sim \frac{1}{\rho_1(z)} + \frac{1}{\rho_2(z)},\end{aligned} \tag{9.82}$$

which is simply the mean curvature of the wavefront along the central ray[6]. Using this in equation (9.81) and integrating between $z = 0$ and $z = \ell$ gives

---

[6] It is shown in [13] that the above relation, i.e. $n^{-1}(s)\nabla^2\psi = \rho_1^{-1}(s) + \rho_2^{-1}(s)$ is valid also for an inhomogeneous isotropic medium. However, the ray trajectory will be non-rectilinear and the radii of curvature in an inhomogeneous medium will no longer be linear functions of arclength.

$$\int_0^{\ell}\left(\frac{1}{\rho_{10}+z}+\frac{1}{\rho_{20}+z}\right)\mathrm{d}z=\ln\left(\left[\frac{\rho_{10}+\ell}{\rho_{10}}\right]\left[\frac{\rho_{20}+\ell}{\rho_{20}}\right]\right)$$

and consequently

$$I(\ell)=\frac{\rho_{10}\rho_{20}}{(\rho_{10}+\ell)(\rho_{20}+\ell)}. \tag{9.83}$$

The square root of this quantity can be taken to be the factor affecting the amplitude of the field components along the central ray. On recognizing that $\psi = z = \ell$ along the ray, the electric field in a vacuum for the central ray under geometrical optics can then be written as

$$\mathbf{E}(\mathbf{r})\sim\mathbf{E}(\mathbf{r}_o)\sqrt{\frac{\rho_{10}\rho_{20}}{(\rho_{10}+\ell)(\rho_{20}+\ell)}}\,\mathrm{e}^{-jk_0\ell}, \tag{9.84}$$

where $\mathbf{r}_o$ denotes the point on the initial wavefront and $\mathbf{E}(\mathbf{r}_o)$ is the initial electric field at that point. If the initial radii of curvature are positive (diverging rays), the electric field amplitude falls along the propagating ray. On the other hand, if the initial radii of curvature are negative (converging rays), the electric field amplitude increases along the propagating ray. If the radii of curvature are infinite (parallel rays; plane wavefront) the amplitude stays constant. ■■

9.1.8 Other theorems of geometrical optics

Definition 9.1.1 *Congruence*. A system of curves which fill a portion of space in such a way that in general a single curve passes through each point of the region is called a congruence.

Definition 9.1.2 *Normal congruence*. If there exists a family of surfaces which cut each of the curves of a congruence orthogonally, the congruence is said to be normal.

Definition 9.1.3 *Rectilinear congruence*. If each curve of a congruence is a straight line, the congruence is said to be rectilinear.

Theorem 9.1.1 Theorem of Malus and Dupin [1, p 131].
A normal rectilinear congruence remains normal after any number of refractions or reflections.

Proof. It is sufficient to establish the theorem for a single interface. Consider a normal rectilinear congruence in a homogeneous medium of refractive index n_1 and assume that the rays undergo a refraction at an interface T which separates this medium from another homogeneous medium of refractive index n_2. Let S_1 be a

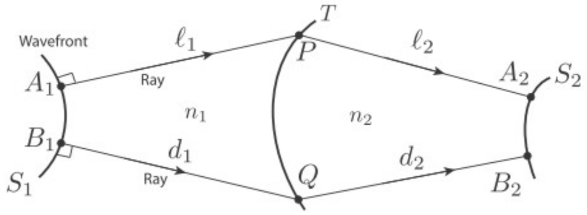

Figure 9.11. Geometry relating to the proof of the theorem of Malus and Dupin.

wavefront in medium 1, let A_1 and P be points of intersection on surfaces S_1 and T, and let A_2 be any point on the refracted ray. The optical path length between A_1 and A_2 is designated as $[A_1PA_2]$ and equals $\ell_1 n_1 + \ell_2 n_2$. Consider a second point B_1Q hitting the interface at Q and let B_2 be a point such that $[B_1QB_2] = \mathrm{d}_1 n_1 + \mathrm{d}_2 n_2 = [A_1PA_2] = \ell_1 n_1 + \ell_2 n_2$. As the point B_1 is varied on S_1, the corresponding point B_2 traverses a surface S_2. It is required to show that the ray $QB_2 \perp S_2$. Applying Lagrange's integral invariant to the closed path $A_1PA_2B_2QB_1A_1$ (see figure 9.11) we obtain

$$[A_1PA_2] + \int_{A_2B_2\in S_2} n_2\hat{\mathbf{p}}_2 \cdot \mathrm{d}\mathbf{r} - [B_2QB_1] + \int_{B_1A_1\in S_1} n_1\hat{\mathbf{p}}_1 \cdot \mathrm{d}\mathbf{r} = 0.$$

Using $[A_1PA_2] = [B_1QB_2] = [B_2QB_1]$ and $\hat{\mathbf{p}}_1 \perp \mathrm{d}\mathbf{r}$ on $B_1A_1 \in S_1$ (since A_1 and B_1 are rays originating from the wavefront S_1) gives

$$\int_{A_2B_2\in S_2} n_2\hat{\mathbf{p}}_2 \cdot \mathrm{d}\mathbf{r} = 0,$$

which implies that $\hat{\mathbf{p}}_2 \perp \mathrm{d}\mathbf{r}$ on S_2 since A_2, and consequently, S_2 is arbitrary. Therefore the refracted rays are orthogonal to S_2; in other words refracted rays form a *normal* congruence, and S_2 also constitutes a wavefront. It also follows from the proof presented that the optical path length between any two wavefronts (whether they are in the same medium or not) is the same for all *rays* connecting the two. Concave and convex lenses designed to project objects to image locations are examples of equality of optical path lengths between wavefronts located across various medium inhomogeneities. ■

9.2 Gradient metasurfaces and the generalized Snell's law

When a ray is reflected from an ordinary isotropic material half-space, the plane of incidence, the plane of reflection, and the plane of transmission remain the same and the angles of reflection and transmission are related to the angle of incidence via Snell's laws. Lately, researchers have exploited the additional degree of freedom provided by artificially modifying the properties of the surface of material discontinuity and shown that it is possible to have the plane of reflection and/or transmission different from the plane of incidence [14]. Below we discuss the mechanism that is

%9 + & > / ' & 0 ' ' - F 0

responsible for the twist in the planes of reflection and refraction and derive the equations of generalized Snell's law.

Consider a smooth surface S separating two piece-wise continuous media with refractive indices n_i, $i = 1, 2$. Let the surface be structured (by the inclusion of a two-dimensional array of sub-wavelength scatterers) to introduce a *constant* surface phase gradient $\mathbf{k}_s = \nabla\Psi$ rad m^{-1} oriented in an arbitrary direction along the surface. If $\hat{\mathbf{n}}$ is the unit normal on the surface then $\hat{\mathbf{n}} \cdot \mathbf{k}_s = 0$. Consider the optical path length $[P_1AP_2]$ between a point P_1 located in medium 1 and a point P_2 located in medium 2 via a point A on the surface. The line P_1A has a direction given by the unit vector $\hat{\mathbf{k}}_1(\mathbf{r}_s) = -(\hat{\mathbf{x}}\sin\theta_1\cos\phi_1 + \hat{\mathbf{y}}\sin\theta_1\sin\phi_1 + \hat{\mathbf{z}}\cos\theta_1)$, while the line AP_2 has a direction given by the unit vector $\hat{\mathbf{k}}_2(\mathbf{r}_s) = (\hat{\mathbf{x}}\sin\theta_2\cos\phi_2 + \hat{\mathbf{y}}\sin\theta_2\sin\phi_2 + \hat{\mathbf{z}}\cos\theta_2)$. The point A has a position vector $\mathbf{r}_s = \mathbf{s}_A + \hat{\mathbf{n}}r_n$ and the various angles are defined with respect to a *local origin* at A. The surface introduces an extra phase shift of $\mathbf{r}_s \cdot \mathbf{k}_s = \mathbf{s}_A \cdot \mathbf{k}_s$ on the propagating ray. Let the wavenumber in free space at the operating frequency be k_0. Then

$$k_0[P_1AP_2] = \int_{P_1}^{A} \mathbf{k}_1 \cdot d\mathbf{r} + \mathbf{k}_s \cdot \mathbf{s}_A + \int_{A}^{P_2} \mathbf{k}_2 \cdot d\mathbf{r},$$

where $\mathbf{k}_i = k_0 n_i \hat{\mathbf{k}}_i$, $i = 1, 2$. Our goal here is to determine the relation between $\mathbf{k}_1$, $\mathbf{k}_2$, $\hat{\mathbf{n}}$, and $\mathbf{k}_s$. Applying Fermat's principle, we take the first variation δ_A of the optical path length with respect to the coordinates of A and set it to zero. In the local coordinate system at A it is straightforward to see that $\delta_A\mathbf{r}_s = \delta_A\mathbf{s}_A + \mathcal{O}(\delta_A^2)$. With reference to figure 9.12 we also note

$$\delta_A \int_{P_1}^{A} \mathbf{k}_1 \cdot d\mathbf{r} = k_0 n_1 \delta_A \ell_1 = k_0 n_1 \hat{\mathbf{k}}_1 \cdot \delta_A\mathbf{s}_A = \mathbf{k}_1 \cdot \delta_A\mathbf{s}_A$$
$$\delta_A \int_{A}^{P_2} \mathbf{k}_2 \cdot d\mathbf{r} = k_0 n_2 \delta_A \ell_2 = k_0 n_2 \hat{\mathbf{k}}_2 \cdot \delta_A\mathbf{s}_A = \mathbf{k}_2 \cdot \delta_A\mathbf{s}_A.$$

Therefore we have from Fermat's principle that

$$(\mathbf{k}_1 + \mathbf{k}_s - \mathbf{k}_2) \cdot \delta_A\mathbf{s}_A = 0.$$

Since $\delta_A\mathbf{s}_A$ is arbitrary and lies on the surface, we have the condition that the quantity within the parenthesis above must be perpendicular to $\mathbf{s}_A$ or, equivalently, must be along the normal $\hat{\mathbf{n}}$ so that

$$\hat{\mathbf{n}} \times (\mathbf{k}_2 - \mathbf{k}_1) = \hat{\mathbf{n}} \times \mathbf{k}_s. \tag{9.85}$$

In a similar fashion it can be established by applying Fermat's principle to the optical path P_1AQ pertaining to the reflection problem that

$$\hat{\mathbf{n}} \times (\mathbf{k}_r - \mathbf{k}_1) = \hat{\mathbf{n}} \times \mathbf{k}_s, \tag{9.86}$$

where $\mathbf{k}_r = k_0 n_1 \hat{\mathbf{k}}_r$ is the wave vector in the direction of the reflected ray and $\hat{\mathbf{k}}_r$ is a unit vector in the direction of the reflected ray.

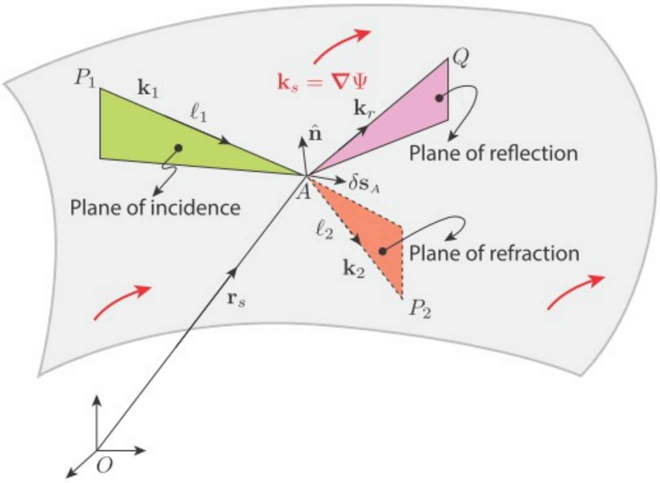

Figure 9.12. Reflection and refraction of light from a metasurface with phase discontinuity.

Example 9.2.0.1. Reflection and refraction from flat phase-gradient interface. When the interface is flat, we choose it to be the ($z = 0$)-plane on which $\hat{\mathbf{n}} = \hat{\mathbf{z}}$. Let the incident ray be in the xz-plane such that $\phi = \phi_1 = 0$, and $\theta = \theta_1$. In that case $\mathbf{k}_1 = -k_0 n_1(\hat{\mathbf{x}} \sin\theta_1 + \hat{\mathbf{z}}\cos\theta_1)$. Let the metasurface have a phase gradient given by $\mathbf{k}_\text{s} = k_\text{s}(\hat{\mathbf{x}}\cos\phi_\text{s} + \hat{\mathbf{y}}\sin\phi_\text{s})$ so that the gradient makes an angle ϕ_s with the xz-plane and k_s is its magnitude. Let the refracted ray have the wave vector $\mathbf{k}_2 = k_0 n_2(\hat{\mathbf{x}}\sin\theta_2 \cos\phi_2 + \hat{\mathbf{y}}\sin\theta_2\sin\phi_2 + \hat{\mathbf{z}}\cos\theta_2)$. Then equation (9.85) suggests that

$$k_0 n_2(\hat{\mathbf{y}}\sin\theta_2\cos\phi_2 - \hat{\mathbf{x}}\sin\theta_2\sin\phi_2) = (k_0 n_1 \sin\theta_1 + k_\text{s}\cos\phi_\text{s})\hat{\mathbf{y}} - k_\text{s}\hat{\mathbf{x}}\sin\phi_\text{s}$$

$$\Longrightarrow \sin\theta_2\sin\phi_2 = \frac{k_\text{s}}{k_0 n_2}\sin\phi_\text{s}$$

$$\text{and } \sin\theta_2\cos\phi_2 = \frac{n_1}{n_2}\sin\theta_1 + \frac{k_\text{s}}{k_0 n_2}\cos\phi_\text{s}.$$

Squaring and adding the last two equations gives

$$\sin\theta_2 = \sqrt{\left(\frac{n_1}{n_2}\right)^2 \sin^2\theta_1 + \left(\frac{k_\text{s}}{k_0 n_2}\right)^2 + \frac{2n_1 k_\text{s}}{k_0 n_2^2}\sin\theta_1\cos\phi_\text{s}}, \tag{9.87}$$

and dividing the second of the last two by the first gives

$$\cot\phi_2 = \frac{k_0 n_1}{k_\text{s}}\frac{\sin\theta_1}{\sin\phi_\text{s}} + \cot\phi_\text{s}, \tag{9.88}$$

For a normally incident wave $\theta_1 = 0$, but equation (9.87) clearly shows that $\theta_2 \neq 0$ unless the phase gradient $k_\text{s} = 0$. In general the angle θ_2 also depends on the frequency of operation. Furthermore, equation (9.88) gives $\phi_2 = \phi_\text{s} \neq 0$. In other words, refraction takes place out of the xz-plane if ϕ_s is different from zero, that is, if a non-zero phase gradient in the y-direction exists. In a sense, the phase gradient surface acts like a surface supporting running fluid that twists the plane of incidence upon refraction by the interface.

In a similar manner, denoting $\mathbf{k}_\mathrm{r} = k_0 n_1(\hat{\mathbf{x}} \sin\theta_\mathrm{r} \cos\phi_\mathrm{r} + \hat{\mathbf{y}} \sin\theta_\mathrm{r} \sin\phi_\mathrm{r} + \hat{\mathbf{z}} \cos\theta_\mathrm{r})$ gives

$$\sin\theta_\mathrm{r} = \sqrt{\sin^2\theta_1 + \left(\frac{k_\mathrm{s}}{k_0 n_1}\right)^2 + \frac{2k_\mathrm{s}}{k_0 n_1}\sin\theta_1 \cos\phi_\mathrm{s}}\,, \tag{9.89}$$

$$\cot\phi_\mathrm{r} = \frac{k_0 n_1}{k_\mathrm{s}}\frac{\sin\theta_1}{\sin\phi_\mathrm{s}} + \cot\phi_\mathrm{s}. \tag{9.90}$$

Equation (9.89) shows that the angle of reflection is not equal to the angle of incidence unless the phase gradient $k_\mathrm{s} = 0$. Furthermore, reflection takes place out of the plane of incidence since $\phi_\mathrm{r} \neq 0$ unless $k_\mathrm{s} = 0$. It is also seen once again that in general the twist in the plane of reflection depends on the frequency of operation. ■■

References

[1] Born M and Wolf E 1980 *Principles of Optics* 6th edn (Elmsford, NY: Pergamon)

[2] Fock V A 1965 *Electromagnetic Diffraction and Propagation Problems* (New York: Pergamon)

[3] Kline M and Kay I W 1979 *Electromagnetic Theory and Geometrical Optics* (Huntington, NY: Kreiger)

[4] Courant R and Hilbert D 1962 *Methods of Mathematical Physics* vol 2 (New York: McGraw-Hill)

[5] Brillouin L 1960 *Wave Propagation and Group Velocity* (New York: Academic)

[6] Craig R A, Katz I, Montgomery R B and Rubenstein P J 1990 Meteorology of the refraction problem *Propagation of Short Radio Waves* (London: Peter Peregrinus)

[7] Craig K H 2003 Clear-air characteristics of the troposphere *Propagation of Radiowaves* (London: Institution of Electrical Engineers)

[8] Felsen L B and Marcuwitz N 1994 *Radiation and Scattering of Waves* (Piscataway, NJ: IEEE)

[9] Jones D S 1979 *Methods in Electromagnetic Wave Propagation* (Oxford: Clarendon)

[10] Sommerfeld A and Runge J 1911 The application of vector calculus to the foundations of geometrical optics *Ann. Phys.* **35** 277–98

[11] Luneberg R K 1964 *Mathematical Theory of Optics* (Providence, RI: Brown University Publishers)

[12] Kline M 1951 An asymptotic solution of Maxwell's equations *Commun. Pure Appl. Math.* **4** 225–63

[13] Kline M 1961 A note on the expansion coefficient of geometrical optics *Commun. Pure Appl. Math.* **14** 473–9

[14] Aieta F *et al* 2012 Reflection and refraction of light from metasurfaces with phase discontinuities *J. Nanophoton.* **6** 063532-1–7

!F"." G 4 4 !) * ! !+ -' $ ' &/
!"#$! ! !% & ' ! + -' $ '

&/ 0- V < 0'&
I& @ @ D 4 : D

brackets only of the form $Q_1'Q_2 - Q_1Q_2'$, which requires derivatives of delta functions. However, no such derivatives appear if the only momentum variable that appears in the constraint, p_L, is conjugate to a field, L, that does not appear with spatial derivatives in the constraint.

The brackets (6.43) and (6.44) rely on general features that do not require a specific dependence on the momentum p_L, which moreover does not have a density weight. The quadratic dependence p_L^2 in $\mathrm{C}[Q]$ can therefore be replaced by any function $f_1(p_L)$. The brackets remain not only closed but also unchanged compared with the classical result (6.44). This result is in apparent conflict with our previous derivation, (6.37), which showed that the spacetime structure is necessarily affected by holonomy modifications.

However, after reformulating the constrained system by introducing $\mathrm{C}[Q]$, spacetime structure has been obscured because we no longer see the relationship with classical hypersurface-deformation brackets (2.154). In order to make the spacetime structure manifest after we modified the constraint $\mathrm{C}[Q]$ to

$$\mathrm{C}[Q] = \int_{-\infty}^{\infty} Q\left(-\frac{f_1(p_L)}{S} + \frac{S(S')^2}{L^2} + \frac{1}{2}\int SV(S)\mathrm{d}S + C_0\right)\mathrm{d}x, \tag{6.45}$$

we should retrace our steps that led us from $\mathrm{H}[N]$ and $\mathrm{D}[\epsilon]$ to $\mathrm{D}[\epsilon]$ and $\mathrm{C}[Q]$. In particular, as in Equation (6.41), a linear combination of some modified $\tilde{\mathrm{H}}[N]$ and $\mathrm{D}[\epsilon]$ should be equal to a multiplier function P times the x-derivative of the integrand of $\mathrm{C}[Q]$:

$$\begin{aligned}
&\tilde{\mathrm{H}}[2PS'/L] + \mathrm{D}[2Pf_3(p_L)/(SL)]\\
&= \int_{-\infty}^{\infty} P\frac{\mathrm{d}}{\mathrm{d}x}\left(-\frac{f_1(p_L)}{S} + \frac{S(S')^2}{L^2} + \frac{1}{2}\int SV(S)\mathrm{d}S\right)\mathrm{d}x\\
&= \int_{-\infty}^{\infty} 2P\left(-\frac{p_L'}{2S}\frac{\mathrm{d}f_1}{\mathrm{d}p_L} + \frac{S'}{L}\left(\frac{Lf_1(p_L)}{2S^2} + \frac{(S')^2}{2L} + \frac{SS''}{L} - \frac{SS'L'}{L^2} + \frac{LSV(S)}{4}\right)\right)\mathrm{d}x.
\end{aligned} \tag{6.46}$$

The multiplier in $\mathrm{D}[M]$ used in this equation may have to be adjusted compared with classical case, as indicated by the function $f_3(p_L)$.

Solving this equation for $\tilde{\mathrm{H}}[N]$ (which, incidentally, requires $f_3(p_L) = \frac{1}{2}\mathrm{d}f_1/\mathrm{f}p_L = f_2(p_L)$), we find that it is exactly of the form (6.36) in which the consistency condition (6.38) has already been implemented. Computing the Poisson bracket $\{\tilde{\mathrm{H}}[N_1], \tilde{\mathrm{H}}[N_2]\}$ therefore must lead to the previous results, Equation (6.37), with the same modification function (6.40). In particular, modifying or quantizing the simplified constraint $\mathrm{C}[Q]$ implies the same kind of signature change as seen for the original Hamiltonian $\mathrm{H}[N]$ (Bojowald et al. 2015).

6.2.2 Effective Line Element

Implications of the modified bracket on spacetime structure can be seen directly if one reconstructs the form of line elements compatible with the modified transformations, using the methods of canonical gravity. Classically, any transformation generated by the constraints of hypersurface deformations,

$$\delta F = \{F, \mathrm{H}[\epsilon^0] + \mathrm{D}[\vec{\epsilon}]\}, \tag{6.47}$$

is equivalent to a derivative in the direction of a spacetime vector field with components ξ^α, where

$$\xi^0 = \frac{\epsilon^0}{N}, \quad \vec{\xi} = \epsilon^i - \frac{\vec{M}}{N}\epsilon^0 \tag{6.48}$$

using the background lapse and shift functions, N and $\vec{M}$, of the spacetime on which the derivative is performed. These relations merely rewrite the reference directions normal and tangential to spatial slices, used in ϵ^0 and $\vec{\epsilon}$, as coordinate directions:

$$\epsilon^0 n^\alpha + \epsilon^i s_i^\alpha = \frac{\epsilon^0}{N}t^\alpha + \left(\epsilon^i - \frac{N^i}{N}\epsilon^0\right)s_i^\alpha = \xi^0 t^\alpha + \xi^i s_i^\alpha. \tag{6.49}$$

A general coordinate transformation may also change the background functions N and $\vec{M}$, defined such that

$$\mathrm{d}s^2 = -N^2\mathrm{d}t^2 + \sum_{i=1}^{3}\sum_{j=1}^{3} q_{ij}(\mathrm{d}x^i + M^i\mathrm{d}t)(\mathrm{d}x^j + M^j\mathrm{d}t) \tag{6.50}$$

with the spatial metric q_{ij}. This behavior is important in the present context because it depends on the brackets of constraints and therefore changes if the brackets are modified. In general, if the brackets of constraints C_A are of the form

$$\{\mathrm{C}_A, \mathrm{C}_B\} = \sum_D F_{AB}^D \mathrm{C}_D \tag{6.51}$$

with suitable functions F_{AB}^D, a transformation of the metric should be combined with the transformation

$$\delta_\epsilon N^A = \frac{\partial \epsilon^A}{\partial t} + \sum_{BC} N^B \epsilon^C F_{BC}^A \tag{6.52}$$

of the background functions. (The notation here is somewhat formal, as the summations may include integrations over spatial coordinates.) This condition ensures that time evolution in the form of constrained Hamilton's equations,

$$\frac{\mathrm{d}F}{\mathrm{d}t} = \{F, \mathrm{H}[N] + \mathrm{D}[\vec{M}]\}, \tag{6.53}$$

is compatible with gauge transformations (Bojowald 2010).

For the brackets of classical canonical gravity, the general relation (6.51) implies

$$\delta_\epsilon N = \frac{\partial \epsilon^0}{\partial t} + \vec{\epsilon}\cdot\nabla N - \vec{M}\cdot\nabla\epsilon^0 \tag{6.54}$$

and

$$\delta_\epsilon \vec{M} = \frac{\partial \vec{\epsilon}}{\partial t} + (\vec{\epsilon} \cdot \nabla)\vec{M} - (\vec{M} \cdot \nabla)\vec{\epsilon} - \mathbf{q}^{-1}(N\nabla\epsilon^0 - \epsilon^0\nabla N), \tag{6.55}$$

where $\mathbf{q}^{-1}$ represents the inverse metric as a matrix with components q^{ij}. However, if the bracket (2.279) is modified, the latter relation should be changed to

$$\delta_\epsilon \vec{M} = \frac{\partial \vec{\epsilon}}{\partial t} + (\vec{\epsilon} \cdot \nabla)\vec{M} - (\vec{M} \cdot \nabla)\vec{\epsilon} - \beta\mathbf{q}^{-1}(N\nabla\epsilon^0 - \epsilon^0\nabla N). \tag{6.56}$$

The coefficients of the line element (6.50) therefore transform differently, in a way that no longer matches the coordinate transformations of $\mathrm{d}t$ and $\mathrm{d}x^i$. In order to obtain an invariant expression, the form (6.50) must be modified.

Still working with spherically symmetric models, we have two components, q_{xx} and $q_{\varphi\varphi}$, of the spatial metric in

$$\mathrm{d}s^2 = -N(x, t)^2\mathrm{d}t^2 + q_{xx}(x, t)(\mathrm{d}x + M(x, t)\mathrm{d}t)^2 + q_{\varphi\varphi}(x, t)(\mathrm{d}\vartheta^2 + \sin^2\vartheta\mathrm{d}\varphi^2). \tag{6.57}$$

If we apply a small coordinate transformation $t' = t + \xi^0$ and $x' = x + \xi^x$, where

$$\xi^0 = \frac{\epsilon^0}{N}, \quad \vec{\xi} = \vec{\epsilon} - \frac{\vec{M}}{N}\epsilon^0 \tag{6.58}$$

in terms of gauge parameters, the infinitesimal coordinate differences transform as

$$\begin{aligned}\mathrm{d}x' = \mathrm{d}(x + \epsilon^x - (N^x/N)\epsilon^0) = \mathrm{d}x + \frac{\partial(\epsilon^x - (M/N)\epsilon^0)}{\partial t}\mathrm{d}t \\ + \frac{\partial(\epsilon^x - (M/N)\epsilon^0)}{\partial x}\mathrm{d}x\end{aligned} \tag{6.59}$$

and

$$\mathrm{d}t' = \mathrm{d}(t + \epsilon^0/N) = \mathrm{d}t + \frac{\partial(\epsilon^0/N)}{\partial t}\mathrm{d}t + \frac{\partial(\epsilon^0/N)}{\partial x}\mathrm{d}x. \tag{6.60}$$

The derivation of these transformations is analogous to methods used in the construction of Bardeen variables in Section 1.4.2. Inserting these transformations in the line element, we obtain the corresponding change

$$\begin{aligned}\delta q_{xx} = \frac{\epsilon^0}{N}\frac{\partial q_{xx}}{\partial t} + \left(\epsilon^x - \frac{M}{N}\epsilon^0\right)\frac{\partial q_{xx}}{\partial x} + 2q_{xx}\frac{\partial}{\partial x}\left(\epsilon^x - \frac{M}{N}\epsilon^0\right) \\ + 2q_{xx}M\frac{\partial}{\partial x}\left(\frac{\epsilon^0}{N}\right)\end{aligned} \tag{6.61}$$

of the spatial metric, along with

$$\delta M = \frac{\partial \epsilon^x}{\partial t} + \epsilon^x\frac{\partial M}{\partial x} - M\frac{\partial \epsilon^x}{\partial x} - \frac{1}{q_{xx}}\left(N\frac{\partial \epsilon^0}{\partial x} - \epsilon^0\frac{\partial N}{\partial x}\right) \tag{6.62}$$

and

$$\delta N = \frac{\partial \epsilon^0}{\partial t} + \frac{\partial N}{\partial x}\epsilon^x - M\frac{\partial \epsilon^0}{\partial x}. \tag{6.63}$$

In particular, Equation (6.62) is not compatible with the transformation (6.56) that results from modified brackets.

Applying the same methods (Bojowald et al. 2018) shows that the modified bracket is compatible with standard coordinate changes of t and x in an effective line element

$$\begin{aligned} \mathrm{d}s^2 = &-\beta(t)N(x,\,t)^2\mathrm{d}t^2 + q_{xx}(x,\,t)(\mathrm{d}x + M(x,\,t)\mathrm{d}t)^2 + q_{\varphi\varphi}(x,\,t) \\ &(\mathrm{d}\vartheta^2 + \sin^2\vartheta\mathrm{d}\varphi^2), \end{aligned} \tag{6.64}$$

using the same function β as in Equation (2.279), and assuming that β depends only on time. (This assumption applies in situations of perturbative inhomogeneity, in which modifications β result only from the background dynamics. If the assumption is violated, it remains unknown whether a more general version of an effective line element still exists, or whether the spacetime structure is entirely non-Riemannian.) When the sign of β changes, it is now clear that we obtain a positive definite line element, as expected for four-dimensional space of Euclidean signature.

6.2.3 Implications of Signature Change

Signature change is an unexpected and counter-intuitive phenomenon. It is an implication not only of certain discrete spatial structures in a consistent spacetime (Bojowald & Paily 2012; Mielczarek 2014; Bojowald & Mielczarek 2015), but has also been found, in a variety of technical realizations, in minisuperspace models (Martin 1994), string theory (Perry & Teo 1993), matrix models (Chaney et al. 2015; Chaney & Stern 2017; Steinacker 2018; Stern & Xu 2018), and causal dynamical triangulations (Ambjørn et al. 2015). While signature change is possible also in classical solutions, in which case it has to be arranged by hand, patching together solutions of Lorentzian and Euclidean gravity along a hypersurface of zero extrinsic curvature, it is always singular in this case (Ellis et al. 1992; Hayward 1993; Dray et al. 1995; Alty & Fewster 1996; Dray et al. 1997). Modified spacetime structures, by contrast, allow for non-singular signature change, for instance when the function β in Equation (2.279) smoothly goes through zero.

In the present context, setting aside details of signature change, the general outcome that a non-classical geometrical structure is realized when the quadratic classical curvature dependence of the Hamiltonian is replaced by a non-quadratic function, evades the slicing problem observed in Section 6.1.2: It is possible to have non-quadratic modifications in a covariant theory, but the covariance relations implied by closed Poisson brackets of the constraints are modified and do not correspond to classical general covariance. The corresponding non-classical spacetime structure is not compatible with the usual form of slicing independence, which is why we obtained a contradiction in Section 6.1.2. Slicing independence is obtained only after transforming to a new effective line element such as Equation (6.64), if one exists. However, the precise transformation remains unknown in the general situation in which the spacetime dependence of β is not only through time.

Conceptually, signature change also helps to explain why models of loop quantum cosmology can be non-singular with standard matter sources, even though they do not violate the usual energy conditions applied in singularity theorems of general relativity. Holonomy modifications in loop quantum cosmology do modify the equations of motion, but the most general singularity theorems are insensitive to the specific dynamics of general relativity and only use broadly applicable geometrical properties such as the geodesic deviation equation. It is therefore not obvious how precisely loop quantum cosmology may be able to evade these theorems.

Signature change presents one possible answer to this question, the only one known so far. In this view, the mathematical outcome of singularity theorems remains unchanged, as it should. These theorems show that, given certain assumptions on initial conditions such as the existence of universal expansion or the presence of trapped surfaces, timelike geodesics cannot last for an infinite amount of time. In the usual spacetime interpretation, this mathematical result is taken to show that spacetime has a boundary at some finite time in the past or future beyond which general relativity does not provide an extension. This finite boundary presents a singularity, not necessarily in terms of diverging densities or curvatures, but as a limitation of the spacetime description derived from the theory.

Signature change solves the singularity problem almost by cutting the Gordian knot; it removes the conceptual problem posed by the required inextendability of timelike geodesics by removing time itself: The assumptions and the outcome of singularity theorems remain mathematically valid, but timelike geodesics end just because time exists only in a limited region within a large universe subject to signature change. As a manifold, the universe does have an extension beyond the point where timelike geodesics end, but only in a four-dimensional spacelike manner. This extended universe is still geometrical, but the geometry is not Riemannian throughout. Only in certain subregions, might an effective line element exist which can map the geometry restricted to such regions to Riemannian form. With this generalized interpretation, singularity theorems do not imply boundaries to the whole manifold of the universe. However, there are still boundaries to causal behavior because there is no deterministic evolution through a region with Euclidean signature.

In particular, a well-posed system of partial differential equations in a geometry with signature change requires some final conditions and cannot be obtained purely as an initial-value problem. This general conclusion follows from the fact that a Euclidean-signature metric implies elliptic partial differential equations for test fields, rather than hyperbolic wave equations. The same general outcome can be seen also in the context of geodesics. If we assume that an effective line element of the form (6.64) can be used (as is the case for instance in the Schwarzschild interior), its time-time component approaches zero in a region where $\beta \to 0$, close to signature change. All light cones in such a region, determined as usual by the condition

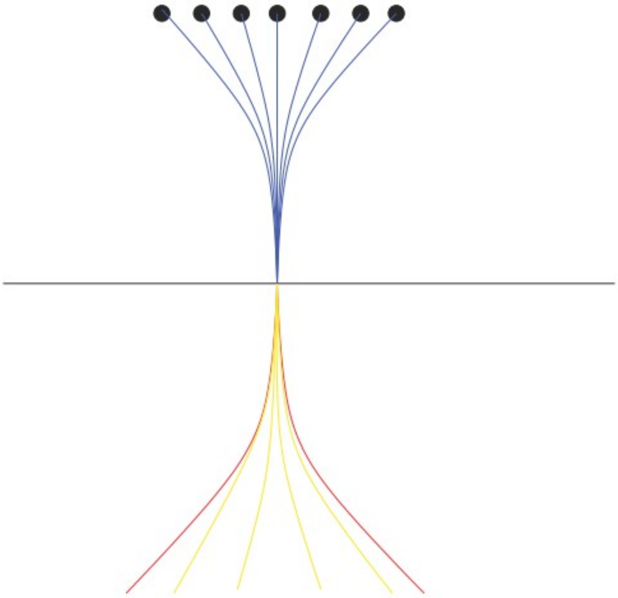

Figure 6.3. Illustration of a family of geodesics in a geometry with signature change, such that Lorentzian signature is realized in the bottom half and Euclidean signature in the top half. In the Lorentzian region, light cones (red) exist but collapse to a unique direction at the transition hypersurface. All timelike geodesics (yellow) therefore approach the same direction. Similarly, all spacelike geodesics in the Euclidean region that approach the tip of the red light cone on the transition hypersurface approach this point in the same asymptotic direction. Any blue spacelike geodesic could therefore be used to extend a given yellow timelike geodesic in a differentiable manner. Unique extensions of timelike geodesics into the Euclidean region as spacelike geodesics (blue) are possible only if a second endpoint is specified (black circles), selecting a specific blue geodesic. The required final position amounts to future data required for a well-posed problem.

$\mathrm{d}s^2 = 0$, have tiny spatial extensions which collapse to zero at $\beta = 0$. Therefore, all timelike geodesics (as well as lightlike ones) approach a unique direction at the hypersurface of signature change, as shown in Figure 6.3.

There are no light cones in the Euclidean region, but an analysis of the geodesic equation shows that also there, geodesics that approach the hypersurface of signature change do so along the normal direction asymptotically close to the hypersurface. Intuitively, this happens because geodesics in a space of Euclidean signature have the shortest distance among all curves connecting two different points. If t is a coordinate normal to the hypersurface of signature change (using the usual letter for the time coordinate even when there is no time in the Euclidean region), the metric component g_{tt} goes through zero where signature change takes place. Therefore, the line element in the Euclidean region is the smallest in the direction of t, which is the direction that geodesics have to follow.

We demonstrate this property formally in two dimensions, using a general line element

$$\mathrm{d}s^2 = th_1(t,\, x)\mathrm{d}t^2 + 2\sqrt{|t|}\, h_2(t,\, x)\mathrm{d}t\mathrm{d}x + g_{xx}(t,\, x)\mathrm{d}x^2 \tag{6.65}$$

with positive coefficients $h_1 > 0$ and $g_{xx} > 0$, such that $d := h_1 g_{xx} - h_2^2 > 0$. Therefore,

$$\det g = th_1 g_{xx} - |t|h_2^2 = -|t|(h_1 g_{xx} + h_2^2) = -|t|(d + 2h_2^2) < 0 \tag{6.66}$$

for $t < 0$ (the Lorentzian region) while

$$\det g = th_1 g_{xx} - th_2^2 = t(h_1 g_{xx} - h_2^2) > 0 \tag{6.67}$$

for $t > 0$ (the Euclidean region).

Along spacelike or timelike worldlines, we have

$$\frac{\mathrm{d}s^2}{\mathrm{d}\lambda^2} = th_1\left(\frac{\mathrm{d}t}{\mathrm{d}\lambda}\right)^2 + 2\sqrt{|t|}\, h_2 \frac{\mathrm{d}t}{\mathrm{d}\lambda}\frac{\mathrm{d}x}{\mathrm{d}\lambda} + g_{xx}\left(\frac{\mathrm{d}x}{\mathrm{d}\lambda}\right)^2 = \pm 1 \tag{6.68}$$

with the upper sign for $t > 0$ (spacelike worldlines in the Euclidean region, where the curve parameter λ is then proper distance) and the lower sign for $t < 0$ (timelike worldlines in the Lorentzian region, where λ is then proper time). Thus, near $t = 0$ the component $\mathrm{d}t/\mathrm{d}\lambda$ of the tangent vector dominates because the equation can be solved for

$$\begin{aligned}\frac{\mathrm{d}t}{\mathrm{d}\lambda} &= -\frac{1}{\sqrt{|t|}\, h_1}\left(\mathrm{sgn}(t) h_2 \frac{\mathrm{d}x}{\mathrm{d}\lambda} \pm \sqrt{(h_2^2 - \mathrm{sgn}(t) h_1 g_{xx})\left(\frac{\mathrm{d}x}{\mathrm{d}\lambda}\right)^2 \pm \mathrm{sgn}(t) h_1}\right)\\ &= O(|t|^{-1/2})\end{aligned} \tag{6.69}$$

in terms of $\mathrm{d}x/\mathrm{d}\lambda = O(1)$. The dominant contributions to the geodesic equation are then

$$\frac{\mathrm{d}^2 t}{\mathrm{d}\lambda^2} \sim -\frac{1}{2t}\left(\frac{\mathrm{d}t}{\mathrm{d}\lambda}\right)^2 (1 + O(t)), \quad \frac{\mathrm{d}^2 x}{\mathrm{d}\lambda^2} \sim \left(\frac{\mathrm{d}t}{\mathrm{d}\lambda}\right)^2 O(\sqrt{|t|}). \tag{6.70}$$

Therefore, $x(\lambda)$ is approximately constant, while the first equation, written as

$$0 \sim 2t\frac{\mathrm{d}^2 t}{\mathrm{d}\lambda^2} + \left(\frac{\mathrm{d}t}{\mathrm{d}\lambda}\right)^2 = 2\sqrt{|t|}\,\mathrm{sgn}(t)\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\sqrt{|t|}\,\frac{\mathrm{d}t}{\mathrm{d}\lambda}\right) \tag{6.71}$$

implies that $\sqrt{|t|}\,\mathrm{d}t/\mathrm{d}\lambda$ is approximately constant. Along geodesics, the slope

$$\frac{\mathrm{d}t}{\mathrm{d}x} = \frac{\mathrm{d}t/\mathrm{d}\lambda}{\mathrm{d}x/\mathrm{d}\lambda} \sim \frac{\mathrm{const}}{\sqrt{|t|}} \tag{6.72}$$

therefore approaches infinity near $t = 0$ in both the Lorentzian and the Euclidean region, as shown in Figure 6.3.

No timelike geodesics exist in the Euclidean region, but we can try to extend a given timelike geodesic arriving at the hypersurface of signature change from the Lorentzian side as a spacelike geodesic extending into the Euclidean region.

However, because all timelike geodesics arriving at a given point on the hypersurface of signature change do so in the same direction, the extension cannot be unique. In addition to specifying initial values to determine which timelike geodesic we are interested in, we have to specify future conditions in the Euclidean region in order to determine an extension by a spacelike geodesic. Because future conditions are required, evolution cannot be deterministic.

The relevant geometry therefore implies a mixture of different, previously irreconcilable scenarios for the interior of black holes. Since quantum modifications are used in the dynamics, it may be possible that the interior does not reach infinite curvature or density of collapsing matter but rather bounces or extends otherwise through the classical singularity. However, interpreted in the corresponding spacetime structure, this transition is not deterministic and requires future data. This latter feature is reminiscent of proposals that tried to address the singularity problem in black hole models by imposing a future condition for wave functions at the singularity (Horowitz & Maldacena 2004; Bouhmadi-López et al. 2019), as some kind of generalization of DeWitt's condition in cosmological models.

Although there is a consistent realization of spacetime in such models, their analysis remains difficult (Bojowald 2015; Ben Achour et al. 2018a, 2018b; Aruga et al. 2019). In particular, the question of how a non-singular interior might be connected to an exterior spacetime, opening up into the original universe or splitting off into a baby universe, remains open. In order to address this question one would like to have access to an effective line element, but β cannot be just time dependent in the exterior as required for Equation (6.64). The non-Riemannian geometry in the exterior therefore remains hard to address.

6.2.4 Perturbative Inhomogeneity

Deformed hypersurface-deformation brackets have been found also for perturbative inhomogeneity in cosmological models (Bojowald et al. 2008; Cailleteau et al. 2012, 2014; Cuttell & Sakellariadou 2014; Cuttell 2019), in a form very similar to those in spherically symmetric models (Barrau et al. 2015). In particular, there is a multiplier β as in Equation (6.30) which becomes negative at Planckian curvature, indicating signature change.

Equations of motion for matter fields, such as a scalar in models of inflation, are relevant in cosmology. These are partial differential equations of a type that depends on the signature of spacetime—hyperbolic in Lorentzian signature and elliptic in Euclidean signature. The conclusion that this behavior is directly related to the sign of β can be obtained from a reconstruction of consistent equations of motion from the (deformed) brackets, following the methods of Kuchař (1976a, 1976b, 1976c).

It is now convenient to Legendre transform the brackets for Hamiltonians, such as $\mathrm{H}(\phi, p_\phi)$ for a scalar field ϕ with momentum p_ϕ, to Lagrangian form, introducing the "velocity" $v(x) = \partial \mathrm{H}/\partial p_\phi(x) = \{\phi, \mathrm{H}[N]\}/N$. The Lagrangian (density) is then defined as $\mathrm{L} = p_\phi v - \mathrm{H}$, such that

$$\left.\frac{\partial \mathrm{H}}{\partial \phi}\right|_{p_\phi \text{ constant}} = -\left.\frac{\partial \mathrm{L}}{\partial \phi}\right|_{v \text{ constant}} \tag{6.73}$$

and

$$p_\phi(x) = \frac{\partial \mathrm{L}}{\partial v(x)}. \tag{6.74}$$

The bracket of two smeared Hamiltonian constraints then equals

$$\{\mathrm{H}[N], \mathrm{H}[M]\} = \int \mathrm{d}^3x \int \mathrm{d}^3y N(y) \frac{\delta \mathrm{H}(y)}{\delta \phi(x)} v(x) M(x) - (N \leftrightarrow M) \tag{6.75}$$

$$= -\int \mathrm{d}^3x \int \mathrm{d}^3y \frac{\delta \mathrm{L}(y)}{\delta \phi(x)} v(x) N(y) M(x) - (N \leftrightarrow M) \tag{6.76}$$

$$= \mathrm{D}[\beta(N\nabla M - M\nabla N)], \tag{6.77}$$

introducing the diffeomorphism constraint $\mathrm{D}[\vec{\epsilon}] = \int \vec{\epsilon} \cdot p_\phi \nabla \phi \mathrm{d}^3x$ in the last step, where we assume brackets of the form (2.279). Comparing the last two lines,

$$\frac{\delta \mathrm{L}(x)}{\delta \phi(x')} v(x') + \beta(x) \vec{\mathrm{D}}(x) \cdot \nabla_x \delta(x, x') - (x \leftrightarrow x') = 0 \tag{6.78}$$

where $\vec{\mathrm{D}}(x) = p_\phi(x) \nabla \phi(x)$.

Using the standard geometrical structure of space, the Lagrangian density must be of the form $\mathcal{L} = \sqrt{\det q}\ \mathrm{L}(\phi, v, \psi)$ with a Lagrange function L that can depend pointwise on ϕ, its velocity v normal to spatial slices, and the rotationally invariant combination

$$\psi = \sum_{i=1}^{3} \sum_{j=1}^{3} q^{ij} \frac{\partial \phi}{\partial x^i} \frac{\partial \phi}{\partial x^j} \tag{6.79}$$

of spatial derivatives. Comparing coefficients of $\nabla \phi$ in Equation (6.78), after an application of the chain rule

$$\frac{\delta \mathrm{L}(x)}{\delta \phi(x')} = \frac{\partial \mathrm{L}(x)}{\partial \phi(x)} \frac{\delta \phi(x)}{\delta \phi(x')} + 2 \frac{\partial \mathrm{L}(x)}{\partial \psi(x)} \nabla \phi(x) \cdot \nabla \delta(x, x'), \tag{6.80}$$

implies

$$\beta p_\phi + 2v \frac{\partial \mathrm{L}}{\partial \psi} = \beta \frac{\partial \mathrm{L}}{\partial v} + 2v \frac{\partial \mathrm{L}}{\partial \psi} = 0. \tag{6.81}$$

Therefore, L must be of the form $\mathrm{L}(\phi, \psi - v^2/\beta)$.

%9 7' ' '

For $\beta > 0$, the combination $\psi - v^2/\beta$ has coefficients of opposite signs in its spatial and temporal derivatives, respectively, implying hyperbolic equations of motion from any Lagrangian consistent with a bracket (2.279). For $\beta < 0$, by contrast, spatial and temporal derivatives enter the Lagrangian with equal signs (as in Euclidean field theory) and imply elliptic partial differential equations for the field.

If β changes its sign dynamically, for instance if it depends on a varying curvature scale as in Equation (6.40), we have a mixed-type problem of partial differential equations that requires a combination of initial values and final values. For instance, a free scalar field on a cosmological background with Hubble parameter H might be subject to the mixed-type wave equation

$$-\frac{\partial^2\phi}{\partial t^2} + \beta(H)\Delta\phi = 0 \tag{6.82}$$

(using the spatial Laplacian Δ) if spacetime is modified, with hypersurface deformations according to Equation (2.279).

A well-posed formulation, introduced by Tricomi (1968) and shown in two dimensions in Figure 6.4; imposes initial values on a single light ray C, or half the usual characteristic problem. Where the light ray approaches the hypersurface of signature change, it is smoothly connected to an arc A in the Euclidean region. If the arc reaches all the way to the end of a second light ray C' starting at the same event as C but moving in the opposite direction, the problem is well-posed: For smooth data of ϕ (but none of its derivatives) on the union $C \cup A$, there is a unique and smooth solution in the region bounded by C, A and C' which depends stably on the data. (Had one tried to evolve usual initial data into the Euclidean region, there might still be a unique solution but it would not depend stably on the data.)

Notice, however, that solutions generically have a root-like pole at the far end of A where it touches C'. This “cosmic boom” is an analog of the sonic boom in hydrodynamics which can be described by similar mixed-type equations. It implies that even small perturbations may lead to solutions with large derivatives. Even if one can extend a metric into the Euclidean region, there may therefore be curvature divergences. In contrast to the standard Big Bang singularity, such divergences would be located at isolated points rather than on an entire spatial slice. As in hydrodynamics, they would be singularities of the effective spacetime description while the underlying fundamental theory, given by a consistent set of canonical constraints $\mathrm{H}[N]$ and $\mathrm{D}[M]$, would be non-singular. Nevertheless, such singularities in an effective cosmological description complicate any physical analysis based on perturbative inhomogeneity.

These examples show that spacetime properties in models of quantum cosmology can be rather counter-intuitive. An analysis of possible evolution at large curvature therefore requires great care and good control on mathematical conditions such as the treatment of well-posed initial-boundary problems that are not required classically. None of these features could have been found had one stayed at a restricted level of minisuperspace models, or perturbations on minisuperspaces.

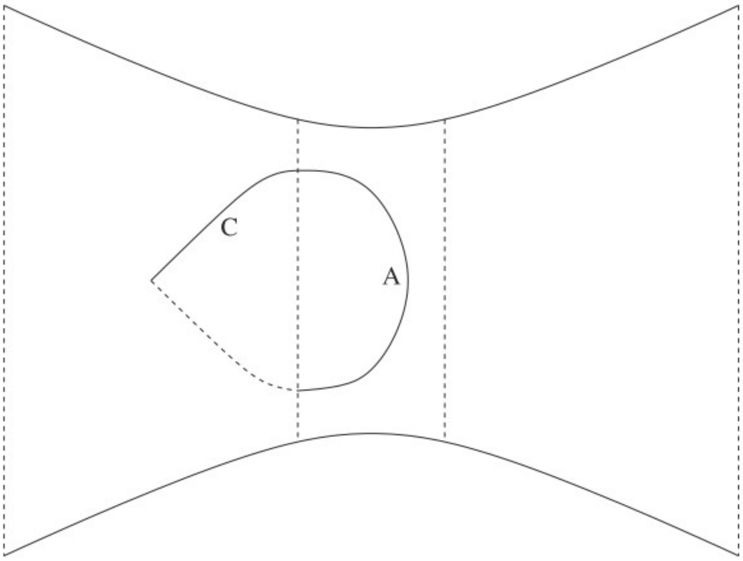

Figure 6.4. The two-dimensional Tricomi problem on a bouncing cosmological background. Initial data are posed on one light ray C in the Lorentzian region, smoothly connected to an arc A in the Euclidean region that ends at the endpoint of the second light ray starting at the same event as C.

6.3 Cosmology at Large Curvature

There are (at least) two general reasons why perturbative inhomogeneity cannot be sufficient to describe quantum spacetime very close to the Big Bang.

First, if we consider a contracting universe in order to understand the approach to a spacelike singularity, either in a potential pre-bounce universe or in backward evolution of our expanding universe, enhanced gravitational collapse in a dense universe implies large inhomogeneity. For instance, black hole formation is expected to happen frequently when the matter distribution gets confined to smaller and smaller space. And when we approach the Planck density, which after all equals more than one trillion solar masses in a proton-sized region, it is hard to imagine how further collapse could keep macroscopic regions at a sufficiently homogeneous distribution for perturbation theory to remain available. If we use perturbative inhomogeneity around a minisuperspace model in order to study the approach to a spacelike singularity, we therefore have to keep adjusting the size V_0 of homogeneous regions used to set up the minisuperspace model by a process of infrared renormalization (Bojowald 2019). This process must end when we reach Planckian size, at which time the infrared scale V_0 would merge with the ultraviolet.

Second, if inflation took place in the early universe, it can be consistent with observations only if it lasted for a sufficiently long time, stretching space by 60 e-folds, or a factor of at least $e^{60} = 1.14 \times 10^{26}$. Moreover, inflation must happen sufficiently early on in the history of the universe because high-energy processes of matter formation, such as baryogenesis, are still to take place after inflation ends.

%9 : 3 ' &
@ &/ ' :' & 4'

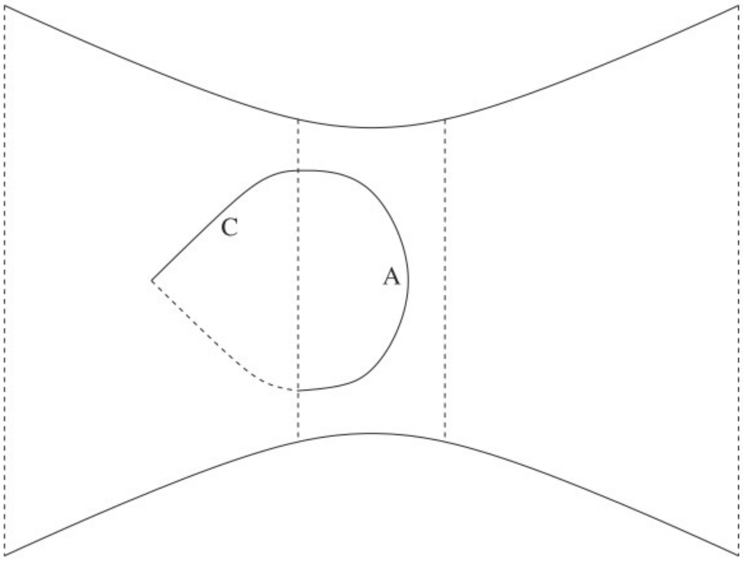

Figure 6.4. The two-dimensional Tricomi problem on a bouncing cosmological background. Initial data are posed on one light ray *C* in the Lorentzian region, smoothly connected to an arc *A* in the Euclidean region that ends at the endpoint of the second light ray starting at the same event as *C*.

6.3 Cosmology at Large Curvature

There are (at least) two general reasons why perturbative inhomogeneity cannot be sufficient to describe quantum spacetime very close to the Big Bang.

First, if we consider a contracting universe in order to understand the approach to a spacelike singularity, either in a potential pre-bounce universe or in backward evolution of our expanding universe, enhanced gravitational collapse in a dense universe implies large inhomogeneity. For instance, black hole formation is expected to happen frequently when the matter distribution gets confined to smaller and smaller space. And when we approach the Planck density, which after all equals more than one trillion solar masses in a proton-sized region, it is hard to imagine how further collapse could keep macroscopic regions at a sufficiently homogeneous distribution for perturbation theory to remain available. If we use perturbative inhomogeneity around a minisuperspace model in order to study the approach to a spacelike singularity, we therefore have to keep adjusting the size V_0 of homogeneous regions used to set up the minisuperspace model by a process of infrared renormalization (Bojowald 2019). This process must end when we reach Planckian size, at which time the infrared scale V_0 would merge with the ultraviolet.

Second, if inflation took place in the early universe, it can be consistent with observations only if it lasted for a sufficiently long time, stretching space by 60 *e*-folds, or a factor of at least $e^{60} = 1.14 \times 10^{26}$. Moreover, inflation must happen sufficiently early on in the history of the universe because high-energy processes of matter formation, such as baryogenesis, are still to take place after inflation ends.

While it is possible to construct inflation models in which these two conditions are met and the Planck scale is avoided even at the beginning of inflation, inflation is an inherently unstable process which is not always easy to stop once it gets started. Without fine-tuned initial conditions, inflation generically lasts longer than 60 *e*-folds, and if it overshoots this target by only about ten *e*-folds, some modes of inhomogeneity that are relevant for current observations of the cosmic microwave background would have had a wave length less than the Planck length at the beginning of inflation. It might still be possible to describe these modes as perturbative ones, but their equations of motion would be subject to quantum spacetime effects that, in general, require a derivation beyond perturbation theory. This *trans-Planckian problem* shows that the inflation scenario can be used to constrain models of quantum cosmology and the underlying approaches to quantum gravity (Martin & Brandenberger 2001; Brandenberger & Martin 2001; Niemeyer 2001; Bedroya & Vafa 2019; Bedroya et al. 2020).

6.3.1 Infrared Renormalization

Considering a collapsing universe, the BKL scenario, described in Section 1.6, suggests a new role for homogeneous dynamics asymptotically close to a spacelike singularity. Even if collapse has led to structure formation on all length scales, the dynamics at each spatial point in this scenario resembles the dynamics of a spatially homogeneous cosmological model.

However, this reference to homogeneity is local and does not imply that the entire space of a minisuperspace model is obtained. If we approach high curvature in forward or backward evolution, we may use homogeneous models, perhaps combined with perturbative inhomogeneity, but the size of regions described by such an approximation must progressively be reduced in order to keep track of gravitational collapse. If we fix a spatial coordinate system at some initial time, a useful parameter to describe this process is the coordinate volume of the region in which perturbative inhomogeneity is admissible as an approximation. Any modes that stretch over distances greater than $\sqrt[3]{V_0}$ must be described by non-perturbative inhomogeneity, or by a full theory of quantum gravity; see Figure 6.5.

As the comoving scale of inhomogeneity decreases, the value of V_0 should be reduced progressively, in a process of infrared renormalization described in Section 4.1.3. This process cannot be performed in a traditional minisuperspace model which works with a fixed set of operators and wave functions in which V_0 (appearing in the basic commutators) cannot be varied. Infrared renormalization therefore requires a new kind of effective theory in quantum cosmology, for instance based on algebraic equations with running V_0. Additional parameters that are usually believed to be constants in minisuperspace models, such as the magnitude of holonomy modifications in loop quantum cosmology given by p_0 in Equation (4.57), run along with V_0 and can therefore show different behaviors in low and high curvature regimes.

Also quantum corrections generically depend on V_0. This behavior initially seems puzzling in minisuperspace models because V_0 refers to the coordinate size of an

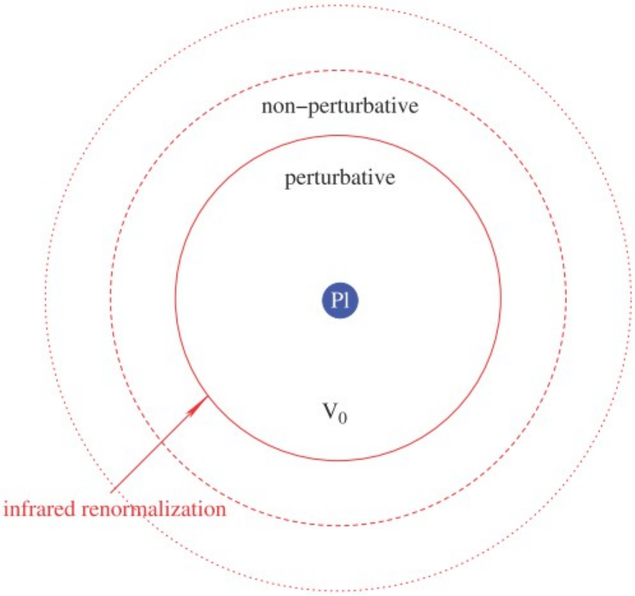

Figure 6.5. Infrared renormalization reduces the size V_0 of any region in which perturbative inhomogeneity can be used, tracking the scale of gravitational collapse in an approach to high curvature (red circles). A lower limit for the size is set by the Planck scale (blue circle).

arbitrary averaging region used to define homogeneous variables. And even if spatial coordinates have been fixed, the averaging region may be changed without affecting the classical dynamics. In classical cosmology, there is no obvious physical process that corresponds to changing V_0, suggesting that observables or their quantum corrections should be independent of V_0. However, quantum corrections generically depend on V_0 because it appears in the basic variables, in particular $V = V_0 a^3$, that define a quantum minisuperspace model. Uncertainty relations, such as

$$\Delta_\omega V \Delta_\omega H \geqslant \frac{2\pi G\hbar}{c^2} = 2\pi c \ell_{\mathrm{Pl}}^2 \tag{6.83}$$

in a state ω, based on the momentum $p_V = -Hc^2/(4\pi G)$ of V according to Equation (1.16), by contrast, impose a lower bound given by fundamental parameters. Numerator and denominator in unitless relative fluctuations, $V^{-1}\Delta_\omega V$, must therefore depend differently on V_0 because the classical scaling, $V \propto V_0$ and H independent of V_0 realized for expectation values $\omega(\hat{V})\ \omega_\omega(\hat{H})$, cannot be applied to volume fluctuations, $\Delta_\omega V$ together with $\Delta_\omega H$, without violating the uncertainty bound.

If V_0 is macroscopic, which is a valid assumption at late times given large-scale homogeneity, relative volume fluctuations can be miniscule, implying largely classical behavior. For microscopic V_0, by contrast, large relative fluctuations and significant quantum effects are expected. Importantly, even the tiny Planckian value on the right-hand side of Equation (6.83) does not imply that fluctuation terms can always be ignored because the BKL scenario is asymptotic and does not provide any lower bound on the size of regions in which homogeneity may be assumed, close to a spacelike singularity.

From a more general perspective that includes the build-up of inhomogeneity, the value of V_0 does have physical meaning once spatial coordinates have been fixed. The time dependence of V_0 describes the varying scale of inhomogeneity, limiting the size of regions in which minisuperspace models can be used. If a quantum version of the BKL scenario is available, local quantum effects should depend only on the volume of the typical regions that may be considered homogeneous at any given time, but not on back-reaction effects from inhomogeneous modes on scales greater than $\sqrt[3]{V_0}$. The initially puzzling dependence of quantum corrections on V_0, seen in Section 4.1, therefore models the quantum interactions (or their absence) between different modes. As V_0 shrinks in a BKL-type scenario, more modes are pushed into the full inhomogeneous region and therefore, by the BKL assumption, no longer interact significantly with the homogeneous center.

A valid version of numerical quantum cosmology should include infrared renormalization by combining homogeneous or perturbative quantum cosmological models in the interior region of Figure 6.5 with numerical relativity in the exterior in order to track the scale of inhomogeneity. One would have to extract the typical value of V_0 from the exterior simulation and feed it into the running scale of the interior. At high curvature, a full exterior simulation of quantum gravity would have to be used. The complexity of such simulations has so far prevented detailed implementations. The approach to a spacelike singularity in minisuperspace-based quantum cosmology therefore remains unclear. This caveat applies in particular to bounce claims made in loop quantum cosmology.

The process of infrared renormalization, adjusting V_0 to the infrared scale of inhomogeneity, is valid as long as curvature does not grow too much. At Planckian curvature, the infrared scale V_0 of a minisuperspace or perturbed model is pushed into the ultraviolet, indicating the breakdown of the minisuperspace perspective even if a quantum version of the BKL scenario may be realized. At this scale, quantum cosmology in all its versions necessarily becomes highly speculative.

6.3.2 The Trans-Planckian Problem

The necessity of infrared renormalization implies that minisuperspace models, on their own, cannot provide a full picture of spacetime at large curvature. Such models eventually break down in an approach to a spacelike singularity because the infrared scale on which they are based will have to be pushed into the ultraviolet in order to maintain an approximation of spacetime by an evolving homogeneous region.

The inflation scenario, described in Section 1.4.4, implies the opposite transition, from ultraviolet to infrared, because modes with short, Planckian wavelength are physically expanded by a large factor in expanding spacetime, and may become relevant for late-time observations of the cosmic microwave background long after inflation has ended. The trans-Planckian problem (Martin & Brandenberger 2001; Brandenberger & Martin 2001; Niemeyer 2001) or, in a stronger version, the trans-Planckian censorship conjecture (Bedroya & Vafa 2019; Bedroya et al. 2020) states that such modes should in consistent inflation models be prevented from becoming relevant at late times, given the (at least at present) uncontrolled nature of strong quantum-gravity effects.

A prominent example of modifications that may be required if quantum gravity becomes necessary in early stages of inflationary structure formation is given by non-classical dispersion relations, which would change relationships such as Equation (1.90) that involve the frequency of inhomogeneous modes for a given wave number. The classical continuum relationship could, for instance, be affected by discrete spacetime (much like dispersion effects implied by solid-state structures), fractal dimensions as derived from causal dynamical triangulations (Mielczarek 2017), higher-curvature terms in covariant actions or in Hořava–Lifshitz gravity (2.143), or by modified spacetime structures in canonical descriptions based on Equation (2.279). In addition to the dynamics, quantum spacetime effects could also change the form of the inflaton ground state, and therefore alter the amplitude (1.91) which in the standard treatment appears directly in the power spectrum (1.98).

The trans-Planckian problem is based on the fact that approximation methods used to study quantum field theory in inflation models cannot be applied with a fixed scale. (This feature is conceptually similar to infrared renormalization.) The rapidly expanding background drags along modes of perturbative inhomogeneity. When the wave length of these modes gets very large, the mode simply follows the expansion of space and stops evolving on its own, as described by the Mukhanov–Sasaki Equation (1.88), until much later when the expansion is no longer inflationary. While such an evolution can well be described semiclassically, a full quantum treatment with varying numbers of evolving degrees of freedom is challenging (Weiss 1985). It would, however, be required if full quantum gravity becomes relevant for some of the modes. As seen in the context of infrared renormalization, minisuperspace-based models are therefore insufficient once Planck-scale physics enters the equations.

6.3.3 Speculative Proposals

Potential properties of various approaches to quantum gravity have been used in myriad ways to suggest new ingredients for the standard Big Bang model. The main targets of such attempts are a non-singular description of the Big Bang singularity of general relativity, and alternatives or modifications of the inflation scenario of structure formation.

Another dichotomy exists between the origin of proposals from a background-independent or background-dependent theory, respectively. From a technical viewpoint, background-dependent approaches are more suitable for cosmological phenomenology because one can directly apply quantum effects to background spacetimes relevant for cosmology. The derivation of such effects for non-static backgrounds may not always be as direct as for static backgrounds, but in phenomenological investigations it is often possible to test possible non-static generalizations of known effects in static backgrounds. Some examples, mainly from string theory, will be given below. For a comprehensive exposition, see Calcagni (2017).

Background-independent Approaches

In background-independent approaches, by contrast, one must first show that it is even possible to have any states that in some approximation may be considered continuum background spacetimes. This problem is often challenging, making it all the more difficult to extend quantum effects to cosmologically relevant spacetimes. As described in Section 5.2.3, the only non-perturbative approach in which suitable cosmological backgrounds have been shown to exist is given by causal dynamical triangulations. But the task of working out cosmological scenarios remains challenging even in this approach.

Loop quantum gravity is probably unique among non-perturbative approaches in that a large number of cosmological applications has been published in this field, mainly in the context of loop quantum cosmology. Initially, loop quantum cosmology drew interest after it was shown in minisuperspace models that the Big Bang singularity may be resolved because wave functions can be extended uniquely across zero scale factor (Bojowald 2001). Since then, however, it has been difficult to extend derivations of cosmological implications beyond the minisuperspace level. Many studies have worked with combinations of minisuperspace models and perturbative inhomogeneity that were only later found to violate covariance; see Section 6.2.4. The application of minisuperspace models helps one to avoid the main difficulties of background-independent approaches because the assumption of spatial homogeneity implies a rather strong background structure, even if it does not completely determine the metric. The over-reliance on minisuperspace models in cosmological applications of loop quantum gravity therefore means that the background-independent nature that characterizes this approach has not yet been tested.

An application of strict minisuperspace models or perturbations also makes it impossible to implement infrared renormalization and to address the trans-Planckian problem, as already discussed in this section. In fact, most claims made in loop quantum cosmology are based on models that fail to implement basic properties of effective field theory (Bojowald 2019, 2020), such as the necessity of scale-dependent coupling parameters (or quantum corrections) because different energy scales usually activate different numbers of relevant degrees of freedom. Most derivations in loop quantum cosmology assume that a single effective theory without running parameters can (or even has to) be used at late times and at very early times (or well outside and deep inside black holes). Because of challenges in deriving such parameters from a full background-independent theory, such an assumption is completely unsupported, but it easily affects potential outcomes. While many very specific claims have been made in loop quantum cosmology, they are hardly reliable. Only qualitative features can at best be taken as a possible indication as to how high curvature might be approached in this framework (Bojowald 2015, 2020). The sensitivity to quantization ambiguities in this approach has already been discussed in Section 4.3.4.

There has been some interest in minisuperspace models in loop quantum cosmology because proposed modifications, such as the application of holonomy terms in Equation (4.57), can be rather drastic compared with the equations of general relativity. Such strong effects are necessary if one attempts to evade the conclusion, based on considerations of units, that quantum-gravity effects in observational cosmology should be suppressed by tiny factors given by the ratios of relevant scales: the Planck length divided by the typical length scale that appears in the physics of the cosmic microwave background. (Such dimensional expectations are also borne out in detail in canonical quantum gravity without loop modifications; Pinho & Pinto-Neto 2007; Falciano & Pinto-Neto 2009; Kiefer & Kraemer 2012.) However, the same effects that could imply significant implications in power spectra also have unexpected consequences in spacetime structures consistent with the underlying modifications, as shown in Section 6.2. In this context, we see a crucial implication of a background-independent approach: Since such an approach does not assume what spacetime structure it might realize, derivations of spacetime structures can potentially lead to counter-intuitive effects. These effects still remain to be understood in loop quantum gravity.

Background-dependent Approaches

Compared with general relativity coupled to standard matter ingredients, string theory provides many additional structures related to extra dimensions with non-trivial topology and supersymmetry that can be used to construct various new ingredients for cosmological phenomenology. In most cases, the same fields and interactions could be postulated as exotic matter terms in general relativity, but additional motivation from string theory can often provide additional support.

As usual with string theory, one of the main difficulties is to find suitable ingredients in the hay stack of degrees of freedom provided by compactified extra dimensions and supersymmetry. Once suitable ingredients have been identified, one must then show that they are stable with respect to cosmological dynamics. For instance, winding modes around compact extra dimensions are usually massive and, just based on the available energy, could therefore decay into lighter excitations such as momentum modes. Demonstrating stability or at least metastability requires an understanding of string interactions in a non-static background, which is often hard to achieve with some degree of rigor.

In this phenomenological spirit, string theory has made contributions to the two main branches of early-universe cosmology: identifying ingredients that can support inflation, and suggesting alternative scenarios that do not rely on inflation but can nevertheless explain observed properties of the nearly scale-free density power spectrum.

Inflation in String Theory: For reasons related to the presence of supersymmetry, string theory does not provide an obvious way to include a positive cosmological constant, or a field with a slightly varying positive potential that could play the role of an inflaton. Nevertheless, as shown in Kachru et al. (2003), it may be possible to do so with flux compactifications, given by higher-dimensional objects made out of interacting strings and winding around compact extra dimensions. In this context, one can address some of the stability questions to make it conceivable that this kind of inflation can last sufficiently long.

Another example, so-called Dirac–Born–Infeld theories, can be suggested for certain scalar fields in string theory (Silverstein & Tong 2004; Alishahiha et al. 2004). Any theory of this form is characterized by an action with a non-linear dependence on the usual kinetic energy of a scalar field, $X = -\frac{1}{2}\sum_{\alpha=0}^{3}\sum_{\beta=0}^{3} g^{\alpha\beta}(\partial\phi/\partial x^{\alpha})(\partial\phi/\partial x^{\beta})$, the same term (2.124) used in the definition of Horndeski theories (2.123) in Section 2.2.5. The Dirac–Born–Infeld version suggested by string theory is indeed of the general form (2.123), but with a more specific

$$H_1(\phi, X) = T(\phi)\left(1 - \sqrt{1 - 2\frac{X}{T(\phi)}}\right) - V(\phi) \tag{6.84}$$

as the only non-zero contribution, with free functions $T(\phi)$ and $V(\phi)$ depending only on the scalar field but not on its derivatives. For small $X/T(\phi)$, $H_1(\phi, X) \sim X - V(\phi)$ equals the standard Lagrangian of a scalar field, but the non-linear dependence on X for larger $X/T(\phi)$ can support inflation with weaker conditions on the potential than would be required for slow-roll inflation with an H_1 linear in X.

Alternatives to Inflation: Various ingredients from string theory have been used to suggest alternatives to inflation that could explain observational features such as a nearly scale-free power spectrum of the cosmic microwave background. Ekpyrotic and cyclic models (Khoury et al. 2001; Steinhardt & Turok 2002; Khoury et al. 2004) are based on the observation that the requirement of having an inflaton fluid with negative P such that $P \approx -\rho$, used in inflation, could be replaced by a stiff effective matter ingredient characterized by positive and large pressure, $P \gg \rho$. Certain perturbation modes with nearly scale-free power spectra can then be obtained. However, they are usually produced in a pre-expansion phase of cosmic collapse, and therefore require some well-defined replacement of the Big Bang singularity. This phase, in a fully dynamical context, remains hard to control in string theory.

The proposal of string-gas cosmology (Brandenberger & Vafa 1989; Kripfganz & Perlt 1988) replaces the dynamical era that is often used to generate a scale-free power spectrum, be it exponential expansion or an ekpyrotic transition, with a stationary equilibrium phase. In the presence of compact extra dimensions in a background spacetime of an ensemble of strings, one can expect a phase transition to take place at high temperature when winding modes become energetically accessible. If the temperature is sufficiently high to excite winding modes, an increased supply of energy no longer leads to enhanced thermal motion of strings described by momentum modes. The reduced thermal motion indicates that the temperature remains constant for some time even as more energy is supplied, just as it happens at well-known phase transitions such as evaporating water.

String-gas cosmology postulates that inflationary expansion can be replaced with a string-gas phase transition, starting with a high-temperature state at which many winding modes are excited. As these modes decay slowly, density perturbations are generated which, unlike in inflation where they originate from quantum fluctuations, are of thermal origin. Equilibrium thermodynamics can then be used to compute the expected power spectrum, which can indeed by nearly scale-free (Nayeri et al. 2006; Brandenberger et al. 2007). A largely open question in this context is how to combine string-gas cosmology, or any other scenario based on winding modes or T-duality such as pre-Big Bang cosmology (Veneziano 1991, 1999; Gasperini & Veneziano 1993, 2003), with manifestly T-dual spacetime structures described by double field theory where the metric $g_{\alpha\beta}$ is replaced by a larger matrix (5.131), in particular if non-associative spacetime features are realized as suggested by Equation (5.142); see Section 5.2.5.

Early-universe Cosmology without Planck-scale Physics

Given the difficulties encountered when trying to achieve rigorous control on Planck or string-scale physics in any approach to quantum gravity, it is of some interest to see whether non-singular scenarios of the early universe can be devised by completely avoiding the Planck scale. An example based on classical physics is the emergent universe (Ellis & Maartens 2004; Ellis et al. 2004), which starts with an initial configuration given by Einstein's static universe (Einstein 1917). A static isotropic solution with vanishing Hubble parameter and acceleration can be obtained in general relativity by balancing the cosmological constant, spatial curvature, and the matter density. Depending on the value of the cosmological constant at early times, the matter density can be well below the Planck density. Moreover, since Einstein's universe turns out to be unstable with respect to anisotropic perturbations, any sub-Planckian quantum fluctuation implies that it will eventually leave the static phase and could turn into an expanding universe.

The Hartle–Hawking proposal (Hartle & Hawking 1983) for an initial state of the universe closed off in Euclidean space is another example of an early-universe scenario that could avoid detailed Planck-scale physics; see Section 4.4.2. Here again, it is the cosmological constant that determines the curvature or density scale of the initial state and may be chosen to be sufficiently sub-Planckian. However, as discussed in more detail in Section 4.4.3, this proposal is unstable with respect to inhomogeneous perturbations (Feldbrugge et al. 2017, 2018; Di Tucci & Lehners 2018). Current proposals to solve the instability problem include quantum-gravity effects closer to the Planck scale, such as dynamical signature change (Bojowald & Brahma 2018), or special initial conditions with non-zero momenta (Vilenkin & Yamada 2018, 2019; Di Tucci & Lehners 2019) that may require new physics.

Finally, matter properties or higher-curvature terms in an effective action of gravity could avoid the Big Bang singularity by a bouncing solution that stays away from Planckian curvature (Wands 1999, 2009; Finelli & Brandenberger 2002; Battefeld & Peter 2015). It may also be possible to achieve such a bounce using quantum fluctuations without requiring detailed Planck-scale physics (Peter & Pinto-Neto 2008). In such scenarios, it is possible to obtain nearly scale-free power spectra from fluctuations seeded in the collapse phase. Open problems are related to the genericness of achieving bouncing behavior, and whether growing anisotropic shear in the collapse phase, (1.46), can be sufficiently controlled so as to ensure scale-free power spectra for a large set of initial values.

!F"." G 4 4 !) * ! !7q ' ' * Q
G' ' - 6 &

= !"#$! ! !% & ' ! >Q)c 8' 7q
' ' * Q G' ' - 6 & 0 & 0 * 8 < 0-
V 0'& l& = 8=@

7 Vektoranalysis und Kurvenintegrale

Wir bewegen uns im Schwerefeld der Erde. Im täglichen Leben haben wir durch die Nutzung elektrischer Geräte mit elektrischen Feldern und Magnetfeldern zu tun. Die Bewegung von Flüssigkeiten und Gasen wird durch Geschwindigkeitsfelder beschrieben. In den genannten Fällen kann man die Felder durch Abbildungen aus dem $\mathbb{R}^3$ (dem dreidimensionalen physikalischen Raum, wo z.B. die Flüssigkeit strömt) in den $\mathbb{R}^3$ (den Raum der Geschwindigkeitsvektoren) auffassen. Solche Abbildungen, die wir in Kapitel 5 als "vektorwertige Abbildungen" bezeichnet hatten, nennt man im Rahmen der Vektoranalysis auch **Vektorfelder**. Man beschränkt den Begriff Vektorfeld dabei nicht auf Abbildungen aus dem $\mathbb{R}^3$ in den $\mathbb{R}^3$, sondern bezieht an verschiedenen Stellen der Theorie Abbildungen $\mathbb{R}^n \to \mathbb{R}^m$ mit $m > 1$ ein, wenn dann auch mitunter die physikalische Entsprechung abhanden kommt. Abbildungen aus dem $\mathbb{R}^n$ in $\mathbb{R}$, die wir bisher reellwertige Abbildungen genannt haben, bezeichnet man im Kontext der Analysis auch als **Skalarfelder**. Ein Beispiel für ein Skalarfeld ist die Temperatur als Funktion der drei Ortskoordinaten und der Zeit (d.h. $n = 4$) in einem Hörsaal. Mit dem Begriff "Feld" wird zum Ausdruck gebracht, dass die betreffende Größe (Vektor oder Skalar) auf einer gewissen Menge $D \subset \mathbb{R}^n$, $n > 1$ definiert ist. Im Folgenden sollen die grundlegenden Operatoren und Elemente der Vektoranalysis dargestellt werden. Ein Ziel ist der sichere Umgang mit Vektorfeldern. Außerdem werden mit den Potentialfeldern und ihren Eigenschaften wichtige Spezialfälle von Vektorfeldern behandelt. Schließlich wird erklärt, wie man die Arbeit berechnen kann, die eine Kraft leistet, wenn man einen Massenpunkt auf einer vorgegebenen Bahnkurve bewegt.

Übersicht

7.1 Die grundlegenden Operatoren der Vektoranalysis 550
7.2 Rechenregeln und Eigenschaften der Operatoren der Vektoranalysis 554
7.3 Potential und Potentialfeld 556
7.4 Skalare Kurvenintegrale 557
7.5 Vektorielles Kurvenintegral – Arbeitsintegral 561
7.6 Stammfunktion eines Gradientenfeldes 565
7.7 Berechnungsmethoden für Stammfunktionen 570
7.8 Vektorpotentiale 571
7.9 Aufgaben 573

© Springer-Verlag GmbH Deutschland 2017
G. Bärwolff, *Höhere Mathematik für Naturwissenschaftler und Ingenieure*, DOI 10.1007/978-3-662-55022-9_7

7.1 Die grundlegenden Operatoren der Vektoranalysis

In vielen Bereichen der Ingenieurwissenschaften und der Physik treten Differentialgleichungen der Art

$$\operatorname{rot}\mathbf{E} + \frac{\partial}{\partial t}\mathbf{B} = \mathbf{0} \tag{7.1}$$

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} = -\operatorname{grad} p + \operatorname{div}(\nu \operatorname{grad}\mathbf{v}) \tag{7.2}$$

$$\operatorname{div}\mathbf{v} = 0 \tag{7.3}$$

$$-\Delta\, p = f \tag{7.4}$$

auf. Der disziplinäre Hintergrund der Gleichungen ist hier von untergeordneter Bedeutung. Es sei nur bemerkt, dass es sich um mathematische Formulierungen von physikalischen Erfahrungssätzen (z.B. das Induktionsgesetz der Theorie elektromagnetischer Felder (7.1)) oder von Erhaltungsätzen (Impuls- und Massenerhaltung bei Strömungen reibungsbehafteter inkompressibler Medien), letztendlich natürlich auch Erfahrungssätze, handelt. Die Formulierungen benutzen typische **Operatoren** der Vektoranalysis **rot , grad , div ,** Δ und ∇, die auf die Vektorfelder **E** (elektrische Feldstärke), **B** (magnetische Induktion), **v** (Strömungsgeschwindigkeit) und das Skalarfeld p (statischer Druck, dividiert durch die konstante Dichte ρ) anzuwenden sind. Mit diesen Operatoren bezeichnet man bestimmte Kombinationen von partiellen Ableitungen nach den Ortskoordinaten, die im Folgenden erklärt werden. Wir bemerken am Rande, dass bei der Herleitung von Gleichungen wie (7.1) bis (7.3) aus den zugrundeliegenden physikalischen Vorgängen i. Allg. zunächst eine Formulierung mittels partieller Ableitungen entsteht. Vorteile der nachträglichen Einführung der angegebenen Operatoren bestehen zum einen in der Kompaktheit der Schreibweise, die man würdigen wird, wenn wir für die Gleichung (7.2) die NAVIER-STOKES-Gleichung, die Formulierung mittels partieller Ableitungen hergestellt haben werden. Zum anderen hat man mittels solcher Operatoren eine koordinatenunabhängige Schreibweise erreicht, die u.a. beim Übergang zu neuen Ortskoordinaten (z.B. von kartesischen zu Zylinder- oder Kugelkoordinaten) gewisse Vorteile bietet. Diese Operatoren erlauben auch eine kompakte, koordinatenunabhängige Formulierung der Integralsätze von GREEN, GAUSS und STOKES (s. Kapitel 8).
Wir wenden uns nun der Definition der grundlegenden Operatoren der Vektoranalysis mittels partieller Ableitungen von Skalar- oder Vektorfeldern nach den kartesischen Koordinaten zu. Dabei setzen wir der Einfachheit halber generell voraus, dass die dabei zu bildenden partiellen Ableitungen existieren und stetig sind. Wir betrachten im Folgenden Vektor- und Skalarfelder auf **Gebieten** $D \subset \mathbb{R}^n$, also auf offenen, zusammenhängenden Mengen. Nun sollen die Operatoren im Einzelnen definiert werden.

Definition 7.1. (Gradient eines Skalarfeldes)
Sei $\phi : D \to \mathbb{R},\ D \subset \mathbb{R}^n$, ein stetig partiell differenzierbares Skalarfeld. Der Operator grad ordnet durch die Vorschrift

$$\operatorname{grad} \phi = \begin{pmatrix} \frac{\partial \phi}{\partial x_1} \\ \frac{\partial \phi}{\partial x_2} \\ \vdots \\ \frac{\partial \phi}{\partial x_n} \end{pmatrix}$$

dem Skalarfeld ϕ das Vektorfeld grad $\phi : D \to \mathbb{R}^n$ zu. Der Vektor grad ϕ heißt **Gradient** von ϕ.

Definition 7.2. (LAPLACE-Operator Δ)
Sei $\phi : D \to \mathbb{R},\ D \subset \mathbb{R}^n$, ein 2-mal stetig partiell differenzierbares Skalarfeld. Der **LAPLACE-Operator** Δ ordnet durch die Vorschrift

$$\Delta\ \phi := \frac{\partial^2 \phi}{\partial x_1^2} + \frac{\partial^2 \phi}{\partial x_2^2} + ... + \frac{\partial^2 \phi}{\partial x_n^2}$$

dem Skalarfeld ϕ das Skalarfeld $\Delta\ \phi : D \to \mathbb{R}$ zu.

Definition 7.3. (Divergenz eines Vektorfeldes)
Sei $\mathbf{v} : D \to \mathbb{R}^n,\ D \subset \mathbb{R}^n$, ein stetig partiell differenzierbares Vektorfeld. Der Operator div ordnet durch die Vorschrift

$$\operatorname{div} \mathbf{v} = \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + ... + \frac{\partial v_n}{\partial x_n}$$

dem Vektorfeld $\mathbf{v}$ das Skalarfeld div $\mathbf{v} : D \to \mathbb{R}$ zu. div $\mathbf{v}$ heißt **Divergenz** des Vektorfeldes $\mathbf{v}$.

Definition 7.4. (Rotation eines Vektorfeldes)
Sei $\mathbf{v} : D \to \mathbb{R}^3,\ D \subset \mathbb{R}^3$, ein stetig partiell differenzierbares Vektorfeld. Der Operator rot ordnet durch die Vorschrift

$$\operatorname{rot} \mathbf{v} = \begin{pmatrix} \frac{\partial v_3}{\partial x_2} - \frac{\partial v_2}{\partial x_3} \\ \frac{\partial v_1}{\partial x_3} - \frac{\partial v_3}{\partial x_1} \\ \frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2} \end{pmatrix}$$

dem Vektorfeld $\mathbf{v}$ das Vektorfeld rot $\mathbf{v} : D \to \mathbb{R}^3$ zu. Der Vektor rot $\mathbf{v}$ heißt **Rotation** des Vektors $\mathbf{v}$ bzw. zu $\mathbf{v}$ gehörendes Wirbelfeld in D.

Im Zusammenhang mit dem Satz von TAYLOR (vgl. Abschnitt 5.11.1) wurde der **Nabla-Operator** schon behandelt. Der Nabla-Operator ∇ kann als symbolischer Vektor aufgefasst werden. Im $\mathbb{R}^n$ hat man

$$\nabla := (\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_n})^T = \mathbf{e}_1 \frac{\partial}{\partial x_1} + \mathbf{e}_2 \frac{\partial}{\partial x_2} + \cdots + \mathbf{e}_n \frac{\partial}{\partial x_n}$$

($\mathbf{e}_1, \ldots, \mathbf{e}_n$ kanonische Basis des $\mathbb{R}^n$). Die Anwendung auf ein stetig partiell differenzierbares Skalarfeld ϕ bedeutet die symbolische Multiplikation von ∇ und ϕ:

$$\nabla \phi = (\mathbf{e}_1 \frac{\partial}{\partial x_1} + \mathbf{e}_2 \frac{\partial}{\partial x_2} + \cdots + \mathbf{e}_n \frac{\partial}{\partial x_n}) \phi = \operatorname{grad} \phi\ .$$

Die Anwendung des Operators $\nabla\cdot$ auf ein stetig differenzierbares Vektorfeld $\mathbf{v}$ bedeutet die Bildung des (formalen) skalaren Produkts zwischen dem symbolischen Vektor ∇ und dem Vektor $\mathbf{v}$. Wegen $\mathbf{e}_i \cdot \mathbf{e}_j = \delta_{ij}$ erhält man

$$\begin{aligned}\nabla \cdot \mathbf{v} &= (\mathbf{e}_1 \frac{\partial}{\partial x_1} + \mathbf{e}_2 \frac{\partial}{\partial x_2} + \cdots + \mathbf{e}_n \frac{\partial}{\partial x_n}) \cdot (v_1 \mathbf{e}_1 + v_2 \mathbf{e}_2 + \cdots + v_n \mathbf{e}_n) \\ &= \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \cdots + \frac{\partial v_n}{\partial x_n} = \operatorname{div} \mathbf{v} \ .\end{aligned}$$

Die Anwendung des Operators $\nabla\times$ auf ein stetig differenzierbares Vektorfeld $\mathbf{v} : \mathbb{R}^3 \to \mathbb{R}^3$ bedeutet die Bildung des (formalen) Vektorprodukts zwischen dem symbolischen Vektor ∇ und dem Vektor $\mathbf{v}$. Man erhält

$$\nabla \times \mathbf{v} = \begin{pmatrix} \frac{\partial}{\partial x_2} v_3 - \frac{\partial}{\partial x_3} v_2 \\ \frac{\partial}{\partial x_3} v_1 - \frac{\partial}{\partial x_1} v_3 \\ \frac{\partial}{\partial x_1} v_2 - \frac{\partial}{\partial x_2} v_1 \end{pmatrix} = \operatorname{rot} \mathbf{v} \ .$$

Eine weitere Möglichkeit zur Berechnung von $\operatorname{rot}\mathbf{v}$ erhält man mit der formalen Berechnung einer Determinante. Es ergibt sich durch Entwicklung nach der ersten Zeile

$$\begin{aligned}&\begin{vmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \\ \frac{\partial}{\partial x_1} & \frac{\partial}{\partial x_2} & \frac{\partial}{\partial x_3} \\ v_1 & v_2 & v_3 \end{vmatrix} \\ &\quad = \mathbf{e}_1 \left(\frac{\partial}{\partial x_2} \mathbf{v}_3 - \frac{\partial}{\partial x_3} \mathbf{v}_2 \right) + \mathbf{e}_2 \left(\frac{\partial}{\partial x_3} \mathbf{v}_1 - \frac{\partial}{\partial x_1} \mathbf{v}_3 \right) + \mathbf{e}_3 \left(\frac{\partial}{\partial x_1} \mathbf{v}_2 - \frac{\partial}{\partial x_2} \mathbf{v}_1 \right) \\ &\quad = \operatorname{rot} \mathbf{v} \ .\end{aligned}$$

Die Operatoren grad und Δ haben wir in den Definitionen auf Skalarfelder ϕ angewandt. Wir verabreden die Anwendung von Δ auf Vektorfelder wie folgt. Sei $\mathbf{v}$ ein stetig partiell differenzierbares Vektorfeld und $\mathbf{w}$ ein zweimal stetig partiell differenzierbares Vektorfeld. Dann definieren wir

$$\Delta \, \mathbf{w} = \begin{pmatrix} \Delta \, w_1 \\ \Delta \, w_2 \\ \\ \vdots \\ \Delta \, w_n \end{pmatrix} ,$$

also $\Delta\mathbf{w}$ als die komponentenweise Anwendung von Δ.
Die Verabredung gilt z.B. auch für den Operator $\mathbf{w} \cdot \nabla$, den wir im Zusammenhang mit dem Satz von TAYLOR verwendet haben: Wir setzen

$$(\mathbf{w} \cdot \nabla)\mathbf{v} = \begin{pmatrix} (\mathbf{w} \cdot \nabla) v_1 \\ (\mathbf{w} \cdot \nabla) v_2 \\ \vdots \\ (\mathbf{w} \cdot \nabla) v_n \end{pmatrix} .$$

Beispiele:
1) Betrachten wir das Zentralkraftfeld

$$\mathbf{K}(\mathbf{x}) = \frac{k}{|\mathbf{x}|^3}\mathbf{x}\,, \quad k \neq 0,\, k \in \mathbb{R}\,,$$

also mit $\mathbf{x} \in D = \mathbb{R}^3 \setminus \mathbf{0}$ ein Vektorfeld $\mathbf{K} : D \to \mathbb{R}^3$. Für $k < 0$ sind die Kraftvektoren $\mathbf{K}(\mathbf{x})$ überall nach dem Ursprung $\mathbf{x} = \mathbf{0}$ hin gerichtet, ihr Betrag ist auf Kugeln $|\mathbf{x}| = r$ konstant und nimmt mit r wie $\frac{1}{r^2}$ ab: $|\mathbf{K}(\mathbf{x})|_{|\mathbf{x}|=r} = \frac{|k|}{r^2}$. Offensichtlich geht es hier insbesondere um das Gravitationsgesetz, wobei in der Konstanten k die beiden sich anziehenden Massen, von denen eine um Punkt $\mathbf{x} = \mathbf{0}$ ruht, und die Gravitationskonstante zusammengefasst sind (vgl. (6.198)). Für die Divergenz und die Rotation errechnet man unter Beachtung der Beziehung $|\mathbf{x}| = \sqrt{x_1^2 + x_2^2 + x_3^2}$

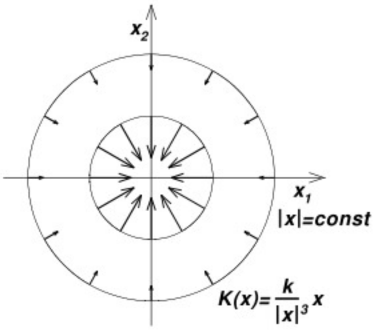

Abb. 7.1. Zentralkraftfeld $\mathbf{K}(\mathbf{x})$ für $k < 0$ in der Ebene $x_3 = 0$

$$\begin{aligned}
\operatorname{div}\mathbf{K}(\mathbf{x}) &= \operatorname{div}\frac{k}{|\mathbf{x}|^3}\mathbf{x} = \operatorname{div}\Big[\frac{k}{|\mathbf{x}|^3}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}\Big] \\
&= \frac{\partial}{\partial x_1}\frac{k\,x_1}{|\mathbf{x}|^3} + \frac{\partial}{\partial x_2}\frac{k\,x_2}{|\mathbf{x}|^3} + \frac{\partial}{\partial x_3}\frac{k\,x_3}{|\mathbf{x}|^3} \\
&= k\Big[\frac{3}{|\mathbf{x}|^3} + x_1\frac{\partial}{\partial x_1}\frac{1}{|\mathbf{x}|^3} + x_2\frac{\partial}{\partial x_2}\frac{1}{|\mathbf{x}|^3} + x_3\frac{\partial}{\partial x_3}\frac{1}{|\mathbf{x}|^3}\Big] \\
&= k\Big[\frac{3}{|\mathbf{x}|^3} - (3x_1^2 + 3x_2^2 + 3x_3^2)\frac{1}{|\mathbf{x}|^5}\Big] = 0
\end{aligned}$$

und

$$\operatorname{rot}\mathbf{K}(\mathbf{x}) = \operatorname{rot}\frac{k}{|\mathbf{x}|^3}\mathbf{x} = \begin{pmatrix} \frac{\partial}{\partial x_2}\frac{k\,x_3}{|\mathbf{x}|^3} - \frac{\partial}{\partial x_3}\frac{k\,x_2}{|\mathbf{x}|^3} \\ \frac{\partial}{\partial x_3}\frac{k\,x_1}{|\mathbf{x}|^3} - \frac{\partial}{\partial x_1}\frac{k\,x_3}{|\mathbf{x}|^3} \\ \frac{\partial}{\partial x_1}\frac{k\,x_2}{|\mathbf{x}|^3} - \frac{\partial}{\partial x_2}\frac{k\,x_1}{|\mathbf{x}|^3} \end{pmatrix} = -\frac{1}{|\mathbf{x}|^5}\begin{pmatrix} 3x_2x_3 - 3x_3x_2 \\ 3x_1x_3 - 3x_3x_1 \\ 3x_2x_1 - 3x_1x_2 \end{pmatrix} = \mathbf{0}\,.$$

2) Wenn wir nun das Skalarfeld $\phi(\mathbf{x}) = -\frac{k}{|\mathbf{x}|}$ betrachten und den Gradienten berechnen, ergibt sich

$$\operatorname{grad}\phi(\mathbf{x}) = \operatorname{grad}\left(-k(x_1^2 + x_2^2 + x_3^2)^{-1/2}\right) = k\frac{1}{|\mathbf{x}|^3}\mathbf{x} = \mathbf{K}(\mathbf{x})\,.$$

Es gibt also ein Skalarfeld $\phi(\mathbf{x})$, aus dem man das Vektorfeld $\mathbf{K}(\mathbf{x})$ durch Gradientenbildung ableiten kann. Die einzelnen Rechnungen sollten zur Übung noch einmal genau nachvollzogen werden.

7.2 Rechenregeln und Eigenschaften der Operatoren der Vektoranalysis

Wenn $\phi : D \to \mathbb{R}, D \subset \mathbb{R}^3$, ein zweimal stetig differenzierbares Skalarfeld ist, so kann man vom Vektorfeld $\mathbf{v} = \operatorname{grad} \phi$ die JACOBI-Matrix $J_\mathbf{v}$ bilden, man erhält

$$J_\mathbf{v}(\mathbf{x}) = J_{\operatorname{grad} \phi}(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 \phi}{\partial x_1^2} & \frac{\partial^2 \phi}{\partial x_1 \partial x_2} & \frac{\partial^2 \phi}{\partial x_1 \partial x_3} \\ \frac{\partial^2 \phi}{\partial x_2 \partial x_1} & \frac{\partial^2 \phi}{\partial x_2^2} & \frac{\partial^2 \phi}{\partial x_2 \partial x_3} \\ \frac{\partial^2 \phi}{\partial x_3 \partial x_1} & \frac{\partial^2 \phi}{\partial x_3 \partial x_2} & \frac{\partial^2 \phi}{\partial x_3^2} \end{pmatrix} .$$

Dabei stellt man fest, dass die JACOBI-Matrix von $\mathbf{v} = \operatorname{grad} \phi$ gleich der HESSE-Matrix von ϕ ist.

Wenn wir den Begriff der Spur einer Matrix $A = (a_{ij})$ mit $\operatorname{Spur}(A) := a_{11} + a_{22} + a_{33}$ verwenden, kann man den LAPLACE-Operator angewendet auf ein Skalarfeld ϕ auch in der Form

$$\Delta \phi = \operatorname{Spur}(J_{\operatorname{grad} \phi}(\mathbf{x}))$$

schreiben.

Rechenregeln für Operatoren der Vektoranalysis:
Sei $\phi : D \to \mathbb{R}$, $D \subset \mathbb{R}^n$, ein zweimal stetig differenzierbares Skalarfeld und $\mathbf{v} : D \to \mathbb{R}^n, D \subset \mathbb{R}^n$, ein stetig differenzierbares Vektorfeld, so gelten die Regeln

(i) $\operatorname{rot}(\operatorname{grad} \phi) = \mathbf{0}$ (Satz von SCHWARZ)

(ii) $\operatorname{div}(\operatorname{rot} \mathbf{v}) = 0$

(iii) $\operatorname{div}(\operatorname{grad} \phi) = \Delta \phi$

(iv) $\operatorname{div}(\phi \mathbf{v}) = \operatorname{grad} \phi \cdot \mathbf{v} + \phi \operatorname{div} \mathbf{v}$

(v) $\operatorname{rot}(\phi \mathbf{v}) = \operatorname{grad} \phi \times \mathbf{v} + \phi(\operatorname{rot} \mathbf{v})$

(vi) $\operatorname{rot}(\operatorname{rot}(\mathbf{v})) = \operatorname{grad}(\operatorname{div}(\mathbf{v})) - \Delta \mathbf{v}$.

Die Regeln, in denen der Rotationsoperator vorkommt, gelten nur für $n = 3$.

Zum Beweis sei außer dem Hinweis auf eine umfängliche Rechnerei nur gesagt, dass die Vertauschbarkeit der Reihenfolge der Ableitungen benötigt wird.

Erinnern wir uns nun nochmal an die beiden Beispiele aus dem vorigen Abschnitt. Mit dem Skalarfeld

$$\psi(\mathbf{x}) = \frac{k}{|\mathbf{x}|^3}$$

können wir das Zentralkraftfeld $\mathbf{K}$ in der Form $\mathbf{K}(\mathbf{x}) = \psi(\mathbf{x})\mathbf{x}$ darstellen. Nach der Regel (iv) erhält man damit

$$\operatorname{div}(\psi\mathbf{x}) = \operatorname{grad}\psi \cdot \mathbf{x} + \psi \operatorname{div}\mathbf{x} = \operatorname{grad}\psi \cdot \mathbf{x} + 3\psi\ ,$$

womit die Rechnung de facto auf die Berechnung der Gradienten von ψ reduziert wird. Dafür rechnet man aus

$$\operatorname{grad}\psi(\mathbf{x}) = -\frac{3k}{|\mathbf{x}|^5}\mathbf{x}\ . \tag{7.5}$$

Andererseits ist $\mathbf{K}(\mathbf{x}) = \psi(\mathbf{x})\mathbf{x} = \psi(\mathbf{x})\mathbf{E}\mathbf{x}$ mit der Einheitsmatrix $\mathbf{E}$. Nach der Kettenregel erhält man

$$\mathbf{K}'(\mathbf{x}) = \psi(\mathbf{x})\mathbf{E} + \mathbf{x}\psi'(\mathbf{x})\ .$$

Man kann die JACOBI-Matrix in der Form

$$J_{\mathbf{K}}(\mathbf{x}) = \psi(\mathbf{x})\mathbf{E} + \mathbf{x} \cdot (\operatorname{grad}\psi(\mathbf{x}))^T$$

schreiben und erhält unter Berücksichtigung von (7.5)

$$J_{\mathbf{K}}(\mathbf{x}) = \frac{k}{|\mathbf{x}|^5}(|\mathbf{x}|^2\mathbf{E} - 3\mathbf{x}\mathbf{x}^T)\ .$$

Wenn man die in der Klammer stehende Matrix genau aufschreibt, erhält man die Diagonalelemente

$$-2x_1^2 + x_2^2 + x_3^2\ , \qquad x_1^2 - 2x_2^2 + x_3^2\ , \qquad x_1^2 + x_2^2 - 2x_3^2\ ,$$

so dass sich für die Spur der JACOBI-Matrix der Wert Null ergibt! Damit kann man für die Funktion ϕ aus dem Beispiel 2 des vorigen Abschnittes schlussfolgern, dass

$$\operatorname{div}\mathbf{K}(\mathbf{x}) = \operatorname{Spur}(J_{\operatorname{grad}\phi}(\mathbf{x})) = \operatorname{Spur}(J_{\mathbf{K}}(\mathbf{x})) = 0$$

ist.

Beispiel: Wir hatten zu Beginn des Abschnittes einige partielle Dfferentialgleichungen in Operatorform aufgeschrieben. Im Folgenden sollen die NAVIER-STOKES-Gleichungen

$$\frac{\partial\mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} = -\operatorname{grad}p + \operatorname{div}(\nu\operatorname{grad}\mathbf{v})\ , \quad \operatorname{div}\mathbf{v} = 0 \tag{7.6}$$

mittels kartesischer Koordinaten einmal vollständig ausgeschrieben dargestellt werden. Wenn wir die obigen Definitionen der Operatoren für das zweimal stetig differenzierbare Vektorfeld $\mathbf{v}:\mathbb{R}^3\to\mathbb{R}^3$ und das stetig differenzierbare Skalarfeld $p:\mathbb{R}^3\to\mathbb{R}$ anwenden, erhalten wir für die erste Gleichung

$$\frac{\partial v_1}{\partial t}+v_1\frac{\partial v_1}{\partial x_1}+v_2\frac{\partial v_1}{\partial x_2}+v_3\frac{\partial v_1}{\partial x_3}= \tag{7.7}$$

$$-\frac{\partial p}{\partial x_1}+\frac{\partial}{\partial x_1}(\nu\frac{\partial v_1}{\partial x_1})+\frac{\partial}{\partial x_2}(\nu\frac{\partial v_1}{\partial x_2})+\frac{\partial}{\partial x_3}(\nu\frac{\partial v_1}{\partial x_3})$$

$$\frac{\partial v_2}{\partial t}+v_1\frac{\partial v_2}{\partial x_1}+v_2\frac{\partial v_2}{\partial x_2}+v_3\frac{\partial v_2}{\partial x_3}= \tag{7.8}$$

$$-\frac{\partial p}{\partial x_2}+\frac{\partial}{\partial x_1}(\nu\frac{\partial v_2}{\partial x_1})+\frac{\partial}{\partial x_2}(\nu\frac{\partial v_2}{\partial x_2})+\frac{\partial}{\partial x_3}(\nu\frac{\partial v_2}{\partial x_3})$$

$$\frac{\partial v_3}{\partial t}+v_1\frac{\partial v_3}{\partial x_1}+v_2\frac{\partial v_3}{\partial x_2}+v_3\frac{\partial v_3}{\partial x_3}= \tag{7.9}$$

$$-\frac{\partial p}{\partial x_3}+\frac{\partial}{\partial x_1}(\nu\frac{\partial v_3}{\partial x_1})+\frac{\partial}{\partial x_2}(\nu\frac{\partial v_3}{\partial x_2})+\frac{\partial}{\partial x_3}(\nu\frac{\partial v_3}{\partial x_3}).$$

Für die Gleichung $\operatorname{div}\mathbf{v}=0$ ergibt sich

$$\frac{\partial v_1}{\partial x_1}+\frac{\partial v_2}{\partial x_2}+\frac{\partial v_3}{\partial x_3}=0\,. \tag{7.10}$$

Insbesondere bei den Gleichungen (7.7), (7.8), (7.9) wird deutlich, welche Übersichtlichkeit mit der Verwendung der Operatoren der Vektoranalysis möglich wird.

7.3 Potential und Potentialfeld

Im obigen Beispiel 2 haben wir mit dem Skalarfeld ϕ ein Feld gefunden, dessen Gradientenfeld gleich einem vorgegebenen Vektorfeld $\mathbf{K}$ ist.
Aus einem gegebenen stetig differenzierbaren Skalarfeld kann man durch Gradientenbildung stets ein Vektorfeld (Gradientenfeld) $\mathbf{v}$ gewinnen. Ist umgekehrt ein Vektorfeld $\mathbf{v}$ gegeben, so ist die Existenz eines Skalarfeldes ϕ mit der Eigenschaft $\operatorname{grad}\phi=\mathbf{v}$ keineswegs immer gesichert. Wir formulieren die entsprechenden Begriffe in Definition 7.5.

Definition 7.5. (Potential, Potentialfeld, Gradientenfeld)

Sei $\mathbf{v}$ ein Vektorfeld. Ein differenzierbares Skalarfeld ϕ, das die Gleichung

$$\operatorname{grad}\phi=\mathbf{v}$$

erfüllt, nennt man ein **Potential** oder eine Stammfunktion von $\mathbf{v}$.
Wenn es zu einem Vektorfeld $\mathbf{v}$ ein Potential ϕ gibt, nennt man $\mathbf{v}$ **Potentialfeld** oder Gradientenfeld (auch der Begriff konservatives Feld wird verwendet).

In den nächsten Abschnitten werden wir feststellen, dass die Kenntnis von Potentialen viele Aufgabenstellungen, z.B. die Berechnung der Arbeit, stark vereinfacht. Deshalb werden wir Kriterien zur Überprüfung der Potentialfeldeigenschaft behandeln und Methoden zur Berechnung von Potentialen darlegen. Wir weisen darauf hin, dass im Folgenden mit t irgendein Kurvenparameter bezeichnet wird, der nicht notwendig die Bedeutung der Zeit hat, auch wenn wir Ableitungen $\frac{d}{dt}$ mit $(\dot{\ })$ abkürzen (z.B. $\frac{dx}{dt} = \dot{x}$).

7.4 Skalare Kurvenintegrale

Ein praktischer Hintergrund der Kurvenintegrale besteht in der Berechnung der Länge von Kurven und der Summation (Integration) einer Belegungsfunktion entlang der Kurve. Ein einfaches Beispiel hierfür ist die Berechnung der Gesamtschneemasse auf einer Freileitung, wobei die Freileitung als Kurve im $\mathbb{R}^3$ betrachtet wird und die heterogen verteilte Schneemasse pro Kurvenabschnitt als Belegungsfunktion f vorgegeben ist. Die skalare Integration der Funktion entlang der Kurve ergibt dann die Gesamtmasse. Die Eigenmasse der Freileitung muss man natürlich noch hinzunehmen, will man die Gesamtmasse der Leitung plus Schneemasse bestimmen. Den Begriff der Kurve haben wir im Kapitel 5 definiert. Zur Erinnerung soll noch einmal das skalare Bogenelement ds einer Kurve $\gamma(t) = (x_1(t), x_2(t))^T$ im $\mathbb{R}^2$

$$ds = \sqrt{\dot{x}_1^2 + \dot{x}_2^2}\, dt = |\dot{\gamma}(t)|\, dt,$$

die Bogenlänge $s(t)$ des Stückes einer Kurve, das dem Parameterintervall $[t_a, t]$ entspricht,

$$s(t) = \int_{t_a}^{t} \sqrt{\dot{x}_1^2 + \dot{x}_2^2}\, dt = \int_{t_a}^{t} |\dot{\gamma}(t)|\, dt$$

und die Bogenlänge L der gesamten Kurve

$$L = \int_{\gamma} ds = \int_{t_a}^{t_e} \sqrt{\dot{x}_1^2 + \dot{x}_2^2}\, dt = \int_{t_a}^{t_e} |\dot{\gamma}(t)|\, dt$$

angegeben werden; dabei verwenden wir die Definitionen 2.31 und 5.14. Die Verallgemeinerung auf Kurven γ im $\mathbb{R}^n$ (vgl. Definitionen 5.12, 5.13, 5.14) fassen wir zusammen in

Definition 7.6. (Bogenelement, Bogenlänge einer Kurve im $\mathbb{R}^n$)
Sei $\gamma(t) = (x_1(t), x_2(t), \ldots, x_n(t))^T$, $t \in [t_a, t_e]$, eine Kurve im $\mathbb{R}^n$. Dann bezeichnen wir mit

$$ds = \sqrt{\dot{x}_1^2(t) + \dot{x}_2^2(t) + \cdots + \dot{x}_n^2(t)}\, dt =: |\dot{\gamma}(t)|\, dt$$

das **Bogenelement** der Kurve (an der Stelle $\gamma(t)$). Für die **Bogenlänge** $s(t)$ des Kurvenstücks, das dem Parameterintervall $[t_a, t]$ $(t \le t_e)$ entspricht, gilt

$$s(t) = \int_{t_a}^{t} \sqrt{\dot{x}_1^2(\tau) + \dot{x}_2^2(\tau) + \cdots + \dot{x}_n^2(\tau)}\, d\tau = \int_{t_a}^{t} |\dot{\gamma}(\tau)|\, d\tau\ .$$

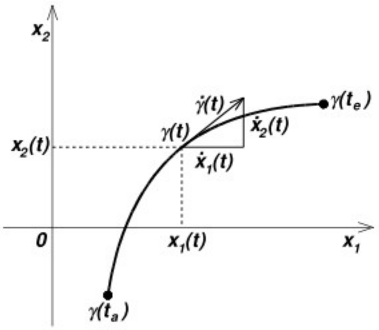

Abb. 7.2. Bezeichnungen bei Kurven γ im $\mathbb{R}^2$

Für die Gesamtlänge L der Kurve γ gilt

$$L = s(t_e) = \int_{t_a}^{t_e} \sqrt{\dot x_1^2(\tau) + \dot x_2^2(\tau) + \cdots + \dot x_n^2(\tau)}\, d\tau \ .$$

Wir werden im Zusammenhang mit Kurvenintegralen immer voraussetzen, dass es sich bei den betrachteten Kurven γ um **reguläre Kurven** im Sinne der Definition 5.13 handelt, d.h. für $\gamma(t) = (x_1(t), x_2(t), \ldots, x_n(t))^T$ wird

$$\dot x_1^2(t) + \dot x_2^2(t) + \cdots + \dot x_n^2(t) = |\dot\gamma(t)|^2 > 0$$

für $t \in \,]t_a, t_e[$ vorausgesetzt.

Definition 7.7. (skalares Kurvenintegral einer Funktion)

Eine Funktion $f : \mathbb{R}^n \supset \gamma([t_a, t_e]) \to \mathbb{R}$ sei auf allen Punkten einer Kurve $\gamma : [t_a, t_e] \to \mathbb{R}^n$ stetig. Dann heißt

$$\int_\gamma f\, ds := \int_{t_a}^{t_e} f(\gamma(t))|\dot\gamma(t)|\, dt$$

das **skalare Kurvenintegral** der Funktion f.

Man nennt die skalaren Kurvenintegrale auch **Kurvenintegrale 1. Art**. Auch die Bezeichnung Linienintegrale 1. Art findet man manchmal.

Wir verstehen die Kurvenintegrale $\int_\gamma f(s)\, ds$ (wie die Integrale $\int_a^b f(x)\, dx$ über Intervalle der x-Achse) als Integrale im RIEMANNschen Sinn, d.h. als Grenzwert der RIEMANNschen Zwischensummen oder als gemeinsamen Grenzwert von Ober- und Untersummen bei immer feiner werdenden Zerlegungen des Intervalls $[t_a, t_e]$; vgl. dazu Abschnitt 2.13.4. Einer Zerlegung des Parameterintervalls $[t_a, t_e]$,

$$t_a = t_0 < t_1 < \cdots < t_{\nu-1} < t_\nu < \cdots < t_{n-1} < t_n = t_e \ ,$$

entspricht eine Zerlegung der Kurve in n Kurvenstücke durch die Teilpunkte

$$\gamma(t_0),\ \gamma(t_1), \ldots, \gamma(t_{\nu-1}),\ \gamma(t_\nu), \ldots, \gamma(t_{n-1}),\ \gamma(t_n) \ .$$

Wir wählen n Zwischenwerte t_ν^* mit $t_{\nu-1} \leq t_\nu^* \leq t_\nu$ ($\nu = 1,2,\ldots,n$) und haben damit auch n Punkte $\gamma(t_\nu^*)$ auf der Kurve gewählt. Mit $\Delta t_\nu = t_\nu - t_{\nu-1}$ ist

$$\Delta s_\nu = |\gamma(t_\nu) - \gamma(t_{\nu-1})| \approx |\dot{\gamma}(t_\nu^*)|\Delta t_\nu \ .$$

Die RIEMANNschen Zwischensummen für das skalare Kurvenintegral sind dann

$$Z_n = \sum_{\nu=1}^{n} f(\gamma(t_\nu^*))\Delta s_\nu \approx \sum_{\nu=1}^{n} f(\gamma(t_\nu^*))|\dot{\gamma}(t_\nu^*)|\Delta t_\nu \ .$$

Jeder einzelne Summand stellt im Fall $\gamma \in \mathbb{R}^2$ näherungsweise den Inhalt eines Flächenstücks $\Delta_\nu F$ dar. Diese anschauliche Deutung ist für Kurven γ im $\mathbb{R}^2$ möglich, weil man sich die zu integrierende Funktion f über der in der (x_1, x_2)-Ebene verlaufenden Kurve γ aufgetragen denken kann (s. Abb. 7.3); Für $\gamma : [t_a, t_e] \to \mathbb{R}^n$ mit $n > 2$ wird diese Veranschaulichung schwierig oder unmöglich.

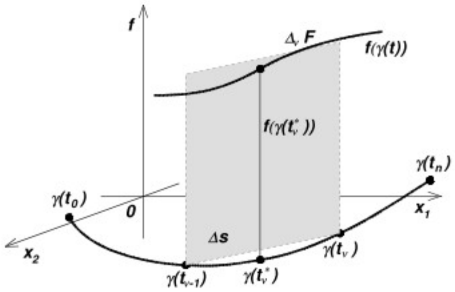

Abb. 7.3. Zur Definition des skalaren Kurvenintegrals für $\gamma : [t_a, t_e] \to \mathbb{R}^2$

Man kann zeigen, dass die RIEMANNschen Zwischensummen (gegen das skalare Kurvenintegral) konvergieren, wenn $f(\gamma(t))$ stetig ist. Handelt es sich bei der Kurve γ um eine aus k Kurvenstücken γ_j zusammengesetzte Kurve, so ergibt sich das Kurvenintegral in der Form

$$\int_\gamma f\,ds = \sum_{j=1}^{k} \int_{\gamma_j} f\,ds \ .$$

Das Kurvenintegral der Funktion $f \equiv 1$ ist gleich der Länge L der Kurve γ. Zur Veranschaulichung des Kurvenintegrals für $\gamma \in \mathbb{R}^3$ kann man sich etwa vorstellen, dass man entlang einer Kurve durch eine Funktion f eine Masseverteilung gegeben hat (Masse eines Drahts). Mit dem Kurvenintegral kann man dann die Gesamtmasse berechnen, wobei der in irgendeiner Weise befestigte Draht durch die Kurve γ beschrieben wird.

Beispiele:
1) Betrachten wir den Kreis mit beliebigem Radius r

$$\gamma(t) = r\begin{pmatrix}\cos t\\ \sin t\end{pmatrix} \ , \ t \in [0,2\pi] \ ,$$

und die Belegungsfunktion $f(x,y) = x^2 - y^2$. Für $f(\gamma(t))$ errechnet man

$$f(\gamma(t)) = (\cos^2 t - \sin^2 t)r^2 = r^2 \cos 2t \ ,$$

und für $\dot{\gamma}(t)$ erhält man

$$\dot{\gamma}(t) = r \begin{pmatrix} -\sin t \\ \cos t \end{pmatrix} \quad \text{und damit} \quad |\dot{\gamma}(t)| = r \ .$$

Für das Kurvenintegral der Funktion f errechnet man schließlich

$$\begin{aligned} \int_\gamma f \, ds &= \int_0^{2\pi} f(\gamma(t))|\dot{\gamma}(t)| \, dt = r^3 \int_0^{2\pi} \cos 2t \, dt \\ &= r^3 \frac{1}{2} \sin 2t|_0^{2\pi} = 0 \ . \end{aligned}$$

2) Eine Verkehrsmaschine fliegt auf einem nördlichen Breitenkreis (geographische Breite $\frac{\pi}{4}$) von Europa nach Nordamerika. Die Entfernung von Start- und Zielort ist gleich einem Viertel des Breitenkreisumfangs. Während des Fluges sondert das Flugzeug Schadstoffe nach der Formel

$$f(x,y,z) = c[1 + \cos(4 \arctan \frac{y}{x})]$$

ab. Die jeweilige Flughöhe ist im Vergleich zum Erdradius vernachlässigbar. Zu berechnen ist die während des Fluges abgegebene Schadstoffmenge.
Wenn wir die Erdoberfläche im $\mathbb{R}^3$ parametrisieren, erhält man unter Nutzung der Kugelkoordinaten

$$\gamma(\lambda) = \begin{pmatrix} x(\lambda) \\ y(\lambda) \\ z(\lambda) \end{pmatrix} = \begin{pmatrix} R\cos\phi \, \cos\lambda \\ R\cos\phi \, \sin\lambda \\ R\sin\phi \end{pmatrix} \ , \phi = \frac{\pi}{4} \ , \lambda \in [0, \frac{\pi}{2}] \ .$$

Hier ist $R \approx 6370\,\text{km}$ der Erdradius; die Flughöhe des Flugzeugs über der Erdoberfläche ist demgegenüber vernachlässigbar.
Für $f(\gamma(\lambda))$ errechnet man

$$f(\gamma(\lambda)) = c[1 + \cos(4 \arctan \frac{\sin\lambda}{\cos\lambda})] = c[1 + \cos(4\lambda)] \ .$$

Wegen $\dot{\gamma}(\lambda) = \frac{R}{\sqrt{2}}(-\sin\lambda, \cos\lambda, 0)^T$ ist

$$|\dot{\gamma}(\lambda)| = \frac{R}{\sqrt{2}} \sqrt{\sin^2\lambda + \cos^2\lambda} = \frac{R}{\sqrt{2}} \ .$$

Damit errechnet man die gesamte während des Fluges abgegebene Schadstoffmenge mit dem Kurvenintegral zu

$$M = \int_\gamma f \, ds = \int_0^{\frac{\pi}{2}} c[1 + \cos(4\lambda)] \frac{R}{\sqrt{2}} \, d\lambda = \frac{cR}{\sqrt{2}} \frac{\pi}{2} \ .$$

Schritte zur Berechnung des skalaren Kurvenintegrals einer Funktion:

1) Parametrisierung der Kurve $\gamma : [t_a, t_e] \to \mathbb{R}^n$

2) Berechnung der Funktionswerte $f(\gamma(t))$ der Belegungsfunktion

3) Berechnung von $|\dot{\gamma}(t)|$

4) Berechnung des Kurvenintegrals $\int_\gamma f\, ds = \int_{t_a}^{t_e} f(\gamma(t))|\dot{\gamma}(t)|\, dt$.

Satz 7.1. *(Rechenregeln für Kurvenintegrale und Mittelwertsatz)*
Sei γ eine Kurve und $f,\, g : \mathbb{R}^n \supset \gamma([t_a, t_e]) \to \mathbb{R}$ stetige Funktionen und $\alpha \in \mathbb{R}$, dann gelten die Regeln

(i) $\int_\gamma (f+g)ds = \int_\gamma f\, ds + \int_\gamma g\, ds$ (Additivität des Integrals),

(ii) $\int_\gamma \alpha f\, ds = \alpha \int_\gamma f\, ds$ (Homogenität des Integrals),

(iii) $\int_\gamma f\, ds = f(\gamma(\tau)) \cdot L$ (Mittelwertsatz);

dabei ist L die Länge der Kurve und $\gamma(\tau)$ ein geeigneter Kurvenpunkt.

7.5 Vektorielles Kurvenintegral – Arbeitsintegral

Sei $\mathbf{k}$ ein auf einer vorgegebenen Kurve γ definiertes Vektorfeld $\mathbf{k} : \mathbb{R}^n \supset \gamma([t_a, t_e]) \to \mathbb{R}^n$. Zur Veranschaulichung setze man $n = 3$ und stelle sich $\mathbf{k}$ als Kraftfeld vor. Möchte man eine punktförmige Masse, d.h. die Idealisierung eines endlich ausgedehnten Objektes mit einer Masse m, dessen räumliche Ausdehnung gegenüber der Länge der Kurve vernachlässigbar ist (ein Beispiel wäre eine Rakete auf dem Weg zum Mond), oder eine Punkt-Ladung entlang der Kurve γ durch das Kraftfeld bzw. durch ein elektrisches Feld bewegen, muss Arbeit geleistet werden. Aus dem Physikunterricht wissen wir, dass sich die Arbeit als Produkt der Kraft und des Weges ergibt. Die Arbeit ΔA, die auf einem sehr kurzen Wegstück Δs bei der Bewegung der Masse durch das Kraftfeld $\mathbf{k}$ verrichtet wird, kann man als Produkt der Tangentialkomponente von $\mathbf{k}$, also der in Kurvenrichtung wirksamen Komponente des Kraftfeldes, mit dem Wegstück Δs

$$\Delta A = k_t \Delta s$$

verstehen. Die Tangentialkomponente k_t der Kraft in Richtung der Kurve ergibt sich durch das Skalarprodukt

$$k_t = \mathbf{k} \cdot \mathbf{t} \quad \text{mit} \quad \mathbf{t} = \frac{\dot{\gamma}(t)}{|\dot{\gamma}(t)|}\ .$$

In der Abb. 7.4 ist diese Situation skizziert. Auf der Bahnkurve $\gamma(t)$ eines im Kraftfeld $\mathbf{k}$ bewegten Objekts ist durch wachsende Werte des Parameters t eine Richtung definiert. In dieser Richtung wächst auch die Bogenlänge s; der Tangentenvektor $\dot{\gamma}(t)$ zeigt ebenfalls in diese Richtung. $\mathbf{k}$ habe im Kurvenpunkt $\gamma(t)$ eine in Richtung des Tangenteneinheitsvektors $\mathbf{t}$ positive Komponente k_t (d.h. $\mathbf{k} \cdot \mathbf{t} > 0$). Bei Bewegungen um ein Wegstück Δs in Richtung wachsender t bzw.

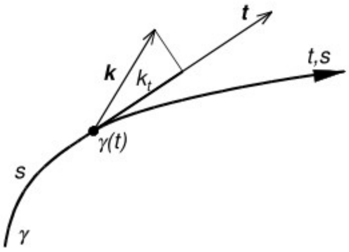

Abb. 7.4. Kraftvektor **k** und Tangentialkomponente k_t, Kurventangentenvektor **t**

s ist $\Delta s > 0$, also wird vom Kraftfeld die Arbeit $\Delta A = k_t \Delta s > 0$ geleistet. Ist dagegen $k_t < 0$ im Punkt $\gamma(t)$, so leistet das Kraftfeld Arbeit, wenn $\Delta s < 0$ ist, d.h. bei einer Bewegung in Richtung fallender t oder s: $\Delta A = k_t \Delta s > 0$. Will man auch in diesem Fall eine Bewegung in Richtung wachsender t oder s erreichen, so ist $\Delta A = k_t \Delta s < 0$, und für diese Bewegung muss gegen das Kraftfeld **k** Arbeit geleistet werden.

Man macht sich dies leicht klar am Beispiel des senkrechten Fallens oder Hebens eines Körpers (Masse m) im Erdschwerefeld. Unabhängig von der Orientierung der Kurve $\gamma(t)$ leistet das Erdschwerefeld Arbeit ($\Delta A > 0$), wenn der Körper fällt, wie es sein muss.

Für die Arbeit, die das Feld **k** an der Stelle $\gamma(t)$ bei Verschiebung eines Körpers um ein kleines Wegstück $\Delta s = |\dot{\gamma}(t)| \Delta t$ längs der Kurve γ leistet, gilt

$$\begin{aligned}\Delta A &= k_t(\gamma(t))\Delta s = \mathbf{k}(\gamma(t)) \cdot \mathbf{t}(\gamma(t))\, \Delta s \\ &= \mathbf{k}(\gamma(t)) \cdot \frac{\dot{\gamma}(t)}{|\dot{\gamma}(t)|}\, \Delta s = \mathbf{k}(\gamma(t)) \cdot \dot{\gamma}(t)\, \Delta t\ .\end{aligned}$$

Hieraus leitet man auf dem üblichen Weg (RIEMANNsche Summen, Folge immer feiner werdender Zerlegungen des Parameterintervalls $[t_a, t_e]$, Grenzübergang, vgl. die Betrachtungen beim skalaren Kurvenintegral) die Definition des vektoriellen Kurvenintegrals ab: Ist **k** ein Kraftfeld, so legen die obigen Überlegungen die Bezeichnung Arbeitsintegral nahe. Es sei aber darauf hingewiesen, dass diese Deutung keineswegs zwingend ist; bei **k** kann es sich z.B. auch um das Vektorfeld der Strömungsgeschwindigkeit eines Fluids handeln.

Definition 7.8. (Arbeitsintegral)
Sei $\gamma : [t_a, t_e] \to \mathbb{R}^n$ eine Kurve und **k** ein auf γ definiertes stetiges Vektorfeld $\mathbf{k} : \mathbb{R}^n \supset \gamma([t_a, t_e]) \to \mathbb{R}^n$. Dann wird das **Integral des Vektorfeldes** (vektorielles Kurvenintegral) entlang der Kurve γ, das auch **Arbeitsintegral** genannt wird, mit $\int_\gamma \mathbf{k} \cdot \mathbf{ds}$ bezeichnet und durch

$$\int_\gamma \mathbf{k} \cdot \mathbf{ds} := \int_{t_a}^{t_e} (\mathbf{k}(\gamma(t)) \cdot \dot{\gamma}(t))\, dt \tag{7.11}$$

definiert. $\mathbf{ds} = \dot{\gamma}(t)\, dt$ bezeichnet man als **vektorielles Bogenelement**. Besteht die Kurve γ aus den m Kurvenstücken $\gamma_1, ..., \gamma_m$, so definiert man

$$\int_\gamma \mathbf{k} \cdot \mathbf{ds} := \sum_{i=1}^{m} \int_{\gamma_i} \mathbf{k} \cdot \mathbf{ds} \, .$$

Wenn γ eine geschlossene Kurve ist, d.h. $\gamma(t_a) = \gamma(t_e)$ gilt, dann verwendet man statt der Bezeichnung $\int_\gamma \mathbf{k} \cdot \mathbf{ds}$ auch die Symbolik

$$\oint_\gamma \mathbf{k} \cdot \mathbf{ds} \, .$$

Man nennt das vektorielle Kurvenintegral auch **Kurven- oder Linienintegral 2. Art**. Für vektorielle Kurvenintegrale über Vektorfelder $\mathbf{k}$ und Kurven γ im $\mathbb{R}^2$ oder $\mathbb{R}^3$ ($\gamma([t_a, t_e]) \subset \mathbb{R}^2$ oder $\mathbb{R}^3$) findet man in der Literatur auch die Bezeichnung

$$\int_A^E (P \, dx + Q \, dy + R \, dz) \, . \tag{7.12}$$

Dabei ist

$$P(x, y, z)\mathbf{e}_1 + Q(x, y, z)\mathbf{e}_2 + R(x, y, z)\mathbf{e}_3 = \mathbf{k}(x, y, z) \, ,$$

und mit $\gamma(t) = (\gamma_1(t), \gamma_2(t), \gamma_3(t))^T$ gilt $dx = \dot{\gamma}_1(t) \, dt$, $dy = \dot{\gamma}_2(t) \, dt$, $dz = \dot{\gamma}_3(t) \, dt$. Es ist $\gamma(t_a) = A, \gamma(t_b) = E$. Damit ist (7.12) nur eine andere Bezeichnung für unser $\int_\gamma \mathbf{k} \cdot \mathbf{ds}$ aus Definition 7.8.

Satz 7.2. *(Rechenregeln für Kurvenintegrale 2. Art)*
Sei γ eine Kurve im $\mathbb{R}^n$ und seien $\mathbf{v}, \mathbf{v}_1, \mathbf{v}_2 : \mathbb{R}^n \supset \gamma([t_a, t_e]) \to \mathbb{R}^n$ stetige Vektorfelder und $\alpha \in \mathbb{R}$. Dann gelten die Regeln

(i) $\int_\gamma (\mathbf{v}_1 + \mathbf{v}_2) \cdot \mathbf{ds} = \int_\gamma \mathbf{v}_1 \cdot \mathbf{ds} + \int_\gamma \mathbf{v}_1 \cdot \mathbf{ds}$,

(ii) $\int_\gamma \alpha \mathbf{v} \cdot \mathbf{ds} = \alpha \int_\gamma \mathbf{v} \cdot \mathbf{ds}$.

(iii) Ist γ^* die Kurve, die aus γ durch Umkehrung des Durchlaufsinns entsteht, d.h., $\gamma^*(t) := \gamma(t_a + t_e - t)$, $t \in [t_a, t_e]$, so folgt

$$\int_{\gamma^*} \mathbf{v} \cdot \mathbf{ds} = - \int_\gamma \mathbf{v} \cdot \mathbf{ds} \, .$$

Algorithmus zur Berechnung des Arbeitsintegrals:

1) Parametrisierung der Kurve $\gamma : [t_a, t_e] \to \mathbb{R}^n$

2) Berechnung der Werte $\mathbf{k}(\gamma(t))$ in den Kurvenpunkten

3) Berechnung des Tangentenvektors $\dot{\gamma}(t)$

4) Berechnung des Kurvenintegrals

$$\int_\gamma \mathbf{k} \cdot \mathbf{ds} = \int_{t_a}^{t_e} (\mathbf{k}(\gamma(t)) \cdot \dot{\gamma}(t)) \, dt \, .$$

Der eben beschriebene Algorithmus macht noch einmal deutlich, dass das Arbeitsintegral des Vektorfeldes **k** entlang der Kurve γ gleich dem skalaren Kurvenintegral der Tangentialkomponente des Vektorfeldes an der Kurve, nämlich $k_t = \mathbf{k} \cdot \dot{\gamma}(t)$ ist. Damit wird auch deutlich, dass das skalare Bogenelement ds gleich dem Betrag des vektoriellen Bogenelements **ds** ist.

Beispiele:
1) Ein Hausmann ($80\,\mathrm{kg}$ Masse) muss seine Fenster putzen. Welche Arbeit leistet er, wenn er mit einem 10 kg schweren Wassereimer in 5 Sekunden auf die Leiter steigt, die 2 m hoch ist, und 1 m von der Hauswand entfernt steht (Leiter ist $\sqrt{5}\,\mathrm{m}$ lang und bildet mit der Hauswand ein rechtwinkliges Dreieck mit den 2 und 1 m langen Katheten)?
Diese Aufgabe bedeutet die Berechnung eines Integrals eines Vektorfeldes entlang einer Kurve. Für die Kurve γ ergibt sich $\gamma(\alpha) = \binom{(\alpha-1)\mathrm{m}}{(2\alpha)\mathrm{m}}$, $\alpha \in [0{,}1]$. Das Kraftfeld, durch das sich der Hausmann mit seinem Eimer beim Aufstieg bewegt, ergibt sich nach NEWTON als Produkt der Masse $m = 90\,\mathrm{kg}$ (Eigenmasse des Hausmanns plus Wassereimer) und der Erdbeschleunigung $\mathbf{g} = (0, -9{,}81\,\frac{\mathrm{m}}{\mathrm{s}^2})^T$,

$$\mathbf{k} = \mathrm{m}\mathbf{g} = \begin{pmatrix} 0 \\ -90 \cdot 9{,}81\,\frac{\mathrm{kg\,m}}{\mathrm{s}^2} \end{pmatrix} .$$

Für den Tangentenvektor $\dot{\gamma}(\alpha)$ ergibt sich $\dot{\gamma}(\alpha) = \binom{1\mathrm{m}}{2\mathrm{m}}$. Damit ergibt sich für die vom Kraftfeld **k** geleistete Arbeit

$$\begin{aligned} A &= \int_\gamma \mathbf{k} \cdot \mathbf{ds} = \int_0^1 \mathbf{k}(\gamma(\alpha)) \cdot \dot{\gamma}(\alpha)\, d\alpha \\ &= \int_0^1 (-882{,}9)\frac{\mathrm{kg\,m}}{\mathrm{s}^2} 2\mathrm{m}\, d\alpha = -1765{,}8\,\frac{\mathrm{kg\,m}}{\mathrm{s}^2}\mathrm{m} = -1765{,}8\,\mathrm{Nm} . \end{aligned}$$

Der Hausmann muss daher die Arbeit $1765{,}8\,\mathrm{Nm}$ leisten, wenn er sich und den Eimer gegen das Schwerefeld auf die Leiter hieven will. Dabei ist es uninteressant, wie lange er dafür braucht, denn wir haben ja nach der Arbeit und nicht nach der Leistung gefragt.
2) Es ist die Arbeit zu berechnen, die ein Satellit auf einer Umrundung eines Planeten (kreisförmig über dem Äquator $z = 0$) mit dem Zentralkraftfeld $\mathbf{K}(\mathbf{x}) = \frac{k}{|\mathbf{x}|^3}\mathbf{x}$ leistet. Der Planet hat dabei den Radius R, und der Satellit fliegt in einer Höhe H über der Planetenoberfläche (in der Konstante $k < 0$ sind die Massen des Planeten und des Satelliten berücksichtigt). Für die Flugkurve erhalten wir $\gamma(\phi) = ((R+H)\cos\phi, (R+H)\sin\phi, 0)^T, \phi \in [0{,}2\pi]$. Für die Arbeit des Feldes $\mathbf{K}(\mathbf{x})$ ergibt sich

$$\begin{aligned} A &= \int_0^{2\pi} \frac{k}{(R+H)^3} \begin{pmatrix} (R+H)\cos\phi \\ (R+H)\sin\phi \\ 0 \end{pmatrix} \cdot \begin{pmatrix} -(R+H)\sin\phi \\ (R+H)\cos\phi \\ 0 \end{pmatrix} d\phi \\ &= \frac{k}{R+H} \int_0^{2\pi} [-\cos\phi\sin\phi + \sin\phi\cos\phi]\, d\phi = 0 , \end{aligned}$$

d.h. es wird weder vom Kraftfeld **K** noch vom Satelliten Arbeit geleistet: Kraftvektor und Tangentenvektor der Flugkurve stehen senkrecht aufeinander.

7.6 Stammfunktion eines Gradientenfeldes

Wir hatten im Abschnitt 7.3 schon den Begriff der Stammfunktion oder des Potentials eines Gradientenfeldes eingeführt. f heißt Stammfunktion eines Vektorfeldes $\mathbf{v}$, wenn $\operatorname{grad} f = \mathbf{v}$ gilt. Im Folgenden wird sich zeigen, dass die Kenntnis einer Stammfunktion die Berechnung von Arbeitsintegralen wesentlich erleichtert.

Satz 7.3. *(erster Hauptsatz für Potentialfelder)*

Sei $D \subset \mathbb{R}^n$ ein Gebiet und $\mathbf{v} : D \to \mathbb{R}^n$ ein Potentialfeld mit der Stammfunktion f. Dann gilt für jede in D verlaufende Kurve $\gamma : [t_a, t_e] \to \mathbb{R}^n$

$$\int_\gamma \mathbf{v} \cdot \mathbf{ds} = f(\gamma(t_e)) - f(\gamma(t_a)) \tag{7.13}$$

Beweis: Es gilt unter Nutzung der Kettenregel

$$\begin{aligned}\int_\gamma \mathbf{v} \cdot \mathbf{ds} &= \int_{t_a}^{t_e} \mathbf{v}(\gamma(t)) \cdot \dot{\gamma}(t)\, dt = \int_{t_a}^{t_e} \operatorname{grad} f(\gamma(t)) \cdot \dot{\gamma}(t)\, dt \\ &= \int_{t_a}^{t_e} \frac{d\, f(\gamma(t))}{d\, t}\, dt = f(\gamma(t_e)) - f(\gamma(t_a)) \quad .\end{aligned}$$

☐

Der folgende Satz liefert äquivalente Aussagen zu Kurvenintegralen und Gradientenfeldern, die möglicherweise unnötige Arbeit ersparen.

Satz 7.4. *(Kurvenintegrale und Potentialfelder)*

Sei $D \subset \mathbb{R}^n$ ein Gebiet, $\gamma : [t_a, t_e] \to D$ eine Kurve in D und $\mathbf{v} : D \to \mathbb{R}^n$ ein stetiges Vektorfeld, dann sind die folgenden Aussagen äquivalent:

1) *Für alle Kurven γ hängt das Kurvenintegral $\int_\gamma \mathbf{v} \cdot \mathbf{ds}$ nur vom Anfangs- und Endpunkt der Kurve ab. Diese Eigenschaft heißt* **Wegunabhängigkeit** *des Kurvenintegrals.*
2) *Für alle geschlossenen Kurven γ, d.h. $\gamma(t_a) = \gamma(t_e)$, gilt $\oint_\gamma \mathbf{v} \cdot \mathbf{ds} = 0$.*
3) *$\mathbf{v}$ ist ein Potentialfeld.*

Beweis: Mit der Eigenschaft

$$\int_{\gamma^*} \mathbf{v} \cdot \mathbf{ds} = -\int_\gamma \mathbf{v} \cdot \mathbf{ds} \,.$$

für die Kurve γ^*, deren Endpunkt mit dem Anfangspunkt von γ und deren Anfangspunkt mit dem Endpunkt von γ übereinstimmt, erhält man sofort für die Kurve $\gamma^\# = \gamma \cup \gamma^*$

$$\oint_{\gamma^\#} \mathbf{v} \cdot \mathbf{ds} = \int_\gamma \mathbf{v} \cdot \mathbf{ds} + \int_{\gamma^*} \mathbf{v} \cdot \mathbf{ds} = \int_\gamma \mathbf{v} \cdot \mathbf{ds} - \int_\gamma \mathbf{v} \cdot \mathbf{ds} = 0 \,.$$

Damit ist 1) $\to$ 2) bewiesen. Betrachtet man nun γ^* als die Kurve, die aus γ durch Umkehrung des Durchlaufsinns entsteht, d.h. $\gamma^*(t) := \gamma(t_a + t_e - t)$, $t \in [t_a, t_e]$, so folgt aus 2) die Wegunabhängigkeit. Der Nachweis 3) $\to$ 2) ist offensichtlich. Der Nachweis 2) $\to$ 3) ist etwas komplizierter. Man betrachtet dazu ein $\mathbf{x}_0 \in D$ und kann in D ein $\mathbf{x}$ finden, so dass ein Polygonzug $\gamma_x = [[\mathbf{x}_0, \mathbf{x}_1], [\mathbf{x}_1, \mathbf{x}_2]...[\mathbf{x}_k, \mathbf{x}]]$ existiert, der vollständig im Gebiet D liegt. Da D offen ist, gibt es für jedes $\mathbf{x} \in D$ ein $r > 0$ mit $\mathbf{x} + \mathbf{z} \in D$ für alle $\mathbf{z} \in \mathbb{R}^n$ mit $|\mathbf{z}| < r$. Wenn wir mit $\mathbf{e}_j$ den $j-$ten Einheitsvektor bezeichnen, so ist für $h \in \mathbb{R}$ mit $h < r$ stets $\mathbf{x} + h\mathbf{e}_j \in D$. Aufgrund von 1) ist die Funktion

$$f(\mathbf{x}) := \int_{\gamma_x} \mathbf{v} \cdot \mathbf{ds} \quad \text{für} \quad \mathbf{x} \in D \tag{7.14}$$

unabhängig vom gewählten Polygonzug γ_x. Wenn wir die $j-$te Komponentenfunktion von $\mathbf{v}$ mit v_j bezeichnen, gilt $v_j = \mathbf{v} \cdot \mathbf{e}_j$. Wegen der Wegunabhängigkeit des Kurvenintegrals gilt

$$f(\mathbf{x} + h\mathbf{e}_j) = \int_{\gamma_x} \mathbf{v} \cdot \mathbf{ds} + \int_{[\mathbf{x}, \mathbf{x}+h\mathbf{e}_j]} \mathbf{v} \cdot \mathbf{ds}\ ,$$

und damit folgt

$$\begin{aligned} \frac{f(\mathbf{x} + h\mathbf{e}_j) - f(\mathbf{x})}{h} &= \frac{1}{h} \int_{[\mathbf{x}, \mathbf{x}+h\mathbf{e}_j]} \mathbf{v} \cdot \mathbf{ds} = \frac{1}{h} \int_0^1 \mathbf{v}(\mathbf{x} + th\mathbf{e}_j) \cdot \mathbf{e}_j \, dt \\ &= \frac{1}{h} \int_0^1 v_j(\mathbf{x} + th\mathbf{e}_j) \, dt = v_j(\mathbf{x} + \tau h\mathbf{e}_j) \end{aligned}$$

aufgrund des Mittelwertsatzes mit einer geeigneten Zahl $\tau \in [0{,}1]$. Wegen der Stetigkeit der Komponentenfunktionen v_j ergibt sich nun

$$\lim_{h \to 0} \frac{f(\mathbf{x} + h\mathbf{e}_j) - f(\mathbf{x})}{h} = v_j(\mathbf{x}) \quad \text{und damit} \quad \operatorname{grad} f(\mathbf{x}) = \mathbf{v}(\mathbf{x})\ .$$

□

Im Beweis des Satzes 7.4 wurde mit der Formel (7.14) eine Möglichkeit aufgezeigt, ausgehend von einem Potentialfeld $\mathbf{v}$ eine Stammfunktion f zu konstruieren. Bevor man allerdings mit dem Integrieren beginnen kann, muss man wissen, ob es sich überhaupt lohnt, d.h., man muss wissen, ob $\mathbf{v}$ ein Gradientenfeld ist. Im Folgenden sollen hinreichende Kriterien für Gradientenfelder besprochen werden. Dazu wird der Begriff der Doppelpunktfreiheit einer Kurve und die Eigenschaft eines Gebietes benötigt, einfach zusammenhängend zu sein.

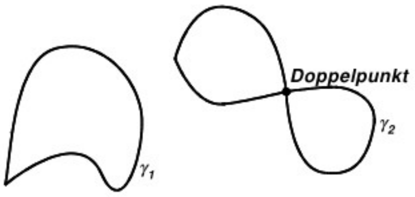

Abb. 7.5. Doppelpunkt freie Kurve γ_1 und Kurve mit Doppelpunkt γ_2

Definition 7.9. (Doppelpunktfreiheit)
Eine Kurve $\gamma : [t_a, t_e] \to \mathbb{R}^n$ heißt **doppelpunktfrei**, wenn

$$\gamma(t_1) \neq \gamma(t_2) \quad \text{für} \quad t_1 \neq t_2,\ t_1, t_2 \in (t_a, t_e)$$

und $\gamma(t_a) \neq \gamma(t)$ für $t \in (t_a, t_e)$ gilt.

Der Begriff des einfach zusammenhängenden Gebiets soll anschaulich erklärt werden, da hier auf komplizierte topologische Betrachtungen verzichtet werden soll. Der Begriff "Gebiet" enthielt bereits die Eigenschaft einer Menge, "zusammenhängend" zu sein. Wir spezialisieren nun auf "einfach zusammenhängend":

Definition 7.10. (einfach zusammenhängendes Gebiet)
Ein Gebiet $D \subset \mathbb{R}^n$ $(n \geq 2)$ heißt **einfach zusammenhängend** oder **kontrahierbar**, wenn jede geschlossene, doppelpunktfreie Kurve in D stetig auf einen Punkt $\mathbf{x} \in D$ zusammengezogen werden kann.

Im $\mathbb{R}^2$ bedeutet "einfach zusammenhängend", dass das Gebiet keine Löcher haben darf. Ein Kreisring ist also nicht einfach zusammenhängend. Im $\mathbb{R}^3$ kann ein Gebiet ein Loch haben, z.B. ist eine Kugelschale einfach zusammenhängend. Andererseits ist ein Torus nicht einfach zusammenhängend.

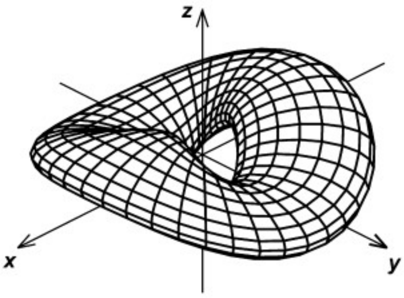

Abb. 7.6. Torus als nicht einfach zusammenhängendes Gebiet im $\mathbb{R}^3$

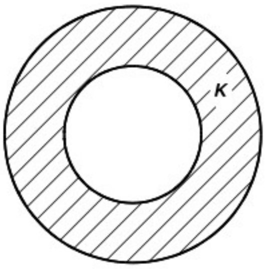

Abb. 7.7. Kreisring als nicht einfach zusammenhängendes Gebiet im $\mathbb{R}^2$

Es ist offensichtlich, dass jedes **konvexe** Gebiet **einfach zusammenhängend ist** (vgl. Definition 5.8).

Bevor wir ein weiteres Kriterium zur Entscheidung, ob ein Vektorfeld $\mathbf{v}$ ein Potentialfeld ist, formulieren, führen wir eine kurze Rechnung durch. Wenn Ψ ein differenzierbares Skalarfeld $\Psi : D \to \mathbb{R},\ D \subset \mathbb{R}^3$ ist, dann stellt man die Identität

$$\operatorname{rot}(\operatorname{grad}\Psi) = \mathbf{0} \tag{7.15}$$

fest. D.h. wenn ein Vektorfeld die gewünschte Darstellung $\mathbf{v} = \operatorname{grad}\Psi$ besitzt, dann gilt notwendigerweise die Beziehung (7.15). Mit dieser Erkenntnis wird nachfolgend ein Kriterium für Potentialfelder formuliert, das sogar notwendig und hinreichend ist.

Satz 7.5. *(Kriterium für die Existenz eines Potentials, zweiter Hauptsatz für Potentialfelder)*

Sei $D \subset \mathbb{R}^n$ ein einfach zusammenhängendes Gebiet ($n \geq 2$) und $\mathbf{v} : D \to \mathbb{R}^n$ ein stetig differenzierbares Vektorfeld. Dann ist $\mathbf{v}$ genau dann ein Potentialfeld, wenn die JACOBI-Matrix $J_{\mathbf{v}}(\mathbf{x})$ für alle $\mathbf{x} \in D$ symmetrisch ist, also

$$J_{\mathbf{v}}(\mathbf{x}) = J_{\mathbf{v}}(\mathbf{x})^T$$

gilt. Die Forderung nach der Symmetrie der JACOBI-Matrix nennt man auch **Integrabilitätsbedingung**. *Für den Fall $n = 3$ ist die Symmetrie der JACOBI-Matrix gleichbedeutend mit der Gleichung*

$$\textit{rot}\,\mathbf{v}(\mathbf{x}) = \mathbf{0} \quad .$$

Beweis: Es soll nur gezeigt werden, dass für ein Potentialfeld $\mathbf{v}$ die JACOBI-Matrix symmetrisch ist, denn der Beweis der anderen Richtung des Satzes wird analog zum Beweis des Satzes 7.4 geführt. Da $\mathbf{v}$ stetig differenzierbar ist, und ein f mit $\operatorname{grad} f = \mathbf{v}$ existiert, folgt die zweifache stetige Differenzierbarkeit von f. Damit gilt nach dem Satz von SCHWARZ

$$\frac{\partial v_i}{\partial x_j} = \frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i} = \frac{\partial v_j}{\partial x_i} \;,$$

also

$$J_{\mathbf{v}}(\mathbf{x}) = J_{\mathbf{v}}(\mathbf{x})^T \quad .$$

□

Mit dem Satz 7.5 liegt nun ein leicht zu überprüfendes Kriterium für Potentialfelder vor.

Beispiele:
1) Betrachten wir das Vektorfeld

$$\mathbf{v}(x, y, z) = \begin{pmatrix} yz^2 \cos(xy) + 2xy \\ xz^2 \cos(xy) + x^2 + z \\ 2z \sin(xy) + y + 2z \end{pmatrix} \;,$$

das auf ganz $\mathbb{R}^3$ und damit auf einem einfach zusammenhängenden Gebiet definiert ist. Nach dem Kriterium des Satzes 7.5 ist $\mathbf{v}$ ein Potentialfeld, wenn $\operatorname{rot} \mathbf{v}(\mathbf{x}) = \mathbf{0}$ gilt. Für die Rotation errechnet man

$$\begin{pmatrix} 2zx\cos(xy) + 1 - [x2z\cos(xy) + 1] \\ y2z\cos(xy) - 2zy\cos(xy) \\ z^2\cos(xy) - xyz^2\sin(xy) + 2x - [z^2\cos(xy) - yz^2x\sin(xy) + 2x] \end{pmatrix} = \mathbf{0} \;,$$

und damit ist $\mathbf{v}$ ein Potentialfeld.

2) Wenn wir das Wirbelfeld $\mathbf{v}(x,y) = \frac{1}{x^2+y^2}\begin{pmatrix} -y \\ x \end{pmatrix}$ auf $D = \mathbb{R}^2 \setminus (0{,}0)$ betrachten, haben wir mit D ein Gebiet, das **nicht einfach zusammenhängend** ist. Wenn wir als Kurve γ den Einheitskreis betrachten ($\gamma(\phi) = (\cos\phi, \sin\phi)^T,\ \phi \in [0{,}2\pi]$), dann ergibt sich

$$\oint_\gamma \mathbf{v} \cdot \mathbf{ds} = 2\pi \neq 0\ ,$$

womit $\mathbf{v}$ auf D nach Satz 7.4 kein Potentialfeld ist. Die Berechnung der JACOBI-Matrix ergibt mit

$$J_{\mathbf{v}}(x,y) = \begin{pmatrix} 2\frac{yx}{(x^2+y^2)^2} & -\frac{-y^2+x^2}{(x^2+y^2)^2} \\ -\frac{-y^2+x^2}{(x^2+y^2)^2} & -2\frac{yx}{(x^2+y^2)^2} \end{pmatrix}$$

eine **symmetrische Matrix**. Allerdings kann das Kriterium aus Satz 7.5 in $\mathbb{R}^2 \setminus \mathbf{0}$ nicht angewandt werden, da D nicht einfach zusammenhängend ist. Wenn $\mathbf{v}$ auf $\mathbb{R}^2 \setminus (0{,}0)$ eine Stammfunktion hätte, käme dafür nur $f(x,y) = \arctan\frac{y}{x} = \phi$ in Frage, wegen

$$f_x = \frac{-y}{x^2+y^2}\ , \qquad f_y = \frac{x}{x^2+y^2}\ .$$

Es liegt mit f aber kein Potential vor, weil $\arctan\frac{y}{x}$ in $\mathbb{R}^2 \setminus \mathbf{0}$ nicht eindeutig definiert werden kann.

Genauer: In einem Punkt P (o.B.d.A. im 1. Quadranten der (x,y)-Ebene) habe die Funktion $f(x,y)$ den Wert $f(P) = \phi$ mit $\tan\phi = \frac{y}{x}$. Lässt man P auf einer geschlossenen Kurve um $(0{,}0)$ herum laufen, so kommt P nach einer ϕ-Änderung um 2π mit sich selbst zur Deckung. Wird f dabei immer stetig fortgesetzt, so erhält man $f(P) = f(P) + 2\pi$ bzw. $\tan(\phi + 2\pi) = \tan\phi = \frac{y}{x}$. f ist also nicht eindeutig und kommt deshalb als Stammfunktion nicht in Frage. Aber in jedem einfach zusammenhängenden Gebiet D, das den Ursprung $\mathbf{0}$ nicht enthält, ist $\mathbf{v}$ Potentialfeld mit der Stammfunktion $f(x,y) = \arctan\frac{y}{x} (= \phi)$. In solchen Gebieten gibt es keine geschlossenen Kurven, auf denen sich ϕ um 2π ändern kann (Abb. 7.8).

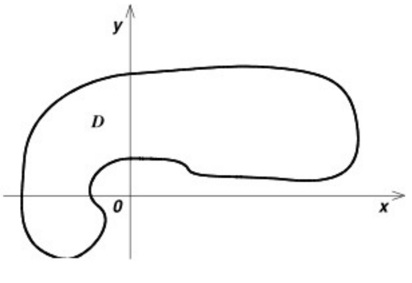

Abb. 7.8. Einfach zusammenhängendes Gebiet D mit $(0{,}0) \notin D$

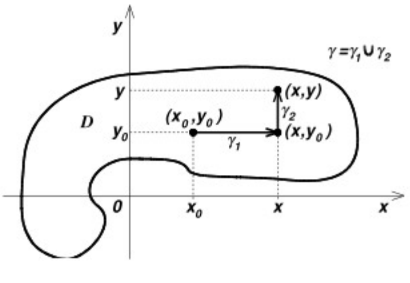

Abb. 7.9. Zur Methode mit dem Kurvenintegral

7.7 Berechnungsmethoden für Stammfunktionen

Wir haben weiter oben schon die Vorteile der Kenntnis einer Stammfunktion erkannt, wenn man z.B. an die Integration des Potentialfeldes entlang einer Kurve denkt. Mit dem folgenden Algorithmus soll nun für das Potentialfeld aus Beispiel 1 eine Stammfunktion bestimmt werden. Man nennt die Methode auch **Ansatzmethode**. Wir gehen davon aus, dass

$$\operatorname{grad} f(x,y,z) = \begin{pmatrix} f_x \\ f_y \\ f_z \end{pmatrix} = \begin{pmatrix} yz^2\cos(xy) + 2xy \\ xz^2\cos(xy) + x^2 + z \\ 2z\sin(xy) + y + 2z \end{pmatrix} = \mathbf{v}(x,y,z) \tag{7.16}$$

gelten muss.

1) Aus der Gleichung

$$f_x = yz^2\cos(xy) + 2xy$$

folgt mittels Integration nach x

$$f(x,y,z) = yz^2\frac{\sin(xy)}{y} + x^2y + C(y,z)\ , \tag{7.17}$$

wobei $C(y,z)$ eine von x unabhängige Funktion sein soll.

2) Das Ergebnis (7.17) wird nun nach y partiell differenziert, man erhält dann

$$f_y = z^2x\cos(xy) + x^2 + C_y(y,z)\ ,$$

so dass sich zur Bestimmung von $C(y,z)$ mit (7.16) die Gleichung

$$z^2x\cos(xy) + x^2 + C_y(y,z) = xz^2\cos(xy) + x^2 + z \quad \text{bzw.} \quad C_y(y,z) = z$$

ergibt. Die Integration nach y ergibt

$$C(y,z) = zy + D(z)\ ,$$

wobei $D(z)$ nicht mehr von x und y, sondern nur noch von z abhängt. Aus (7.17) erhält man damit

$$f(x,y,z) = z^2\sin(xy) + x^2y + zy + D(z) \tag{7.18}$$

3) Das Ergebnis (7.18) wird nun partiell nach z differenziert, und man erhält

$$f_z = 2z\sin(xy) + y + D_z(z)\ ,$$

so dass sich zur Bestimmung von $D(z)$ mit (7.16) die Gleichung

$$2z\sin(xy) + y + D_z(z) = 2z\sin(xy) + y + 2z \quad \text{bzw.} \quad D_z(z) = 2z$$

ergibt. Die Integration nach z ergibt schließlich $D(z) = z^2 +$ const., so dass wir mit

$$f(x,y,z) = z^2\sin(xy) + x^2y + zy + z^2 + \text{const.}$$

eine Stammfunktion gefunden haben.

Die Punkte 1) bis 3) lassen sich in der beschriebenen Weise zur Berechnung einer Stammfunktion eines beliebigen vorgegebenen Potentialfeldes anwenden.

Im Beweis des Satzes 7.4 wurde mit

$$f(\mathbf{x}) := \int_{\gamma_x} \mathbf{v} \cdot \mathbf{ds} \quad \text{für} \quad \mathbf{x} \in D$$

eine Stammfunktion für ein vorgegebenes stetig differenzierbares Vektorfeld $\mathbf{v}$ definiert. Dabei war γ_x ein beliebiger in D liegender Polygonzug, der irgendeinen Punkt $\mathbf{x}_0 \in D$ mit dem Punkt $\mathbf{x} \in D$ verbindet. Will man die Stammfunktion f explizit bestimmen, muss man sich einen günstigen Punkt $\mathbf{x}_0 \in D$ und einen günstigen Polygonzug von $\mathbf{x}_0$ nach $\mathbf{x}$ in D suchen, für den sich das Kurvenintegral leicht berechnen lässt. Falls die Strecke

$$\gamma_x(t) = \mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0) \quad t \in [0,1]$$

in D verläuft, hat man das Integral

$$f(\mathbf{x}) = \int_0^1 \mathbf{v}(\mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0)) \cdot (\mathbf{x} - \mathbf{x}_0)\, dt$$

zu berechnen. Die beschriebene Methode zur Berechnung einer Stammfunktion nennt man **Methode mit dem Kurvenintegral**. Wir wollen diese Methode am Beispiel des Vektorfeldes $\mathbf{v}(x,y) = (\frac{-y}{x^2+y^2}, \frac{x}{x^2+y^2})^T$ auf einem einfach zusammenhängenden Gebiet D, das den Ursprung nicht enthält, erproben (Abb. 7.9).

$$\begin{aligned}
f(x,y) &= \int_{\gamma} \mathbf{v} \cdot \mathbf{ds} = \int_{\gamma_1} \mathbf{v} \cdot \mathbf{ds} + \int_{\gamma_2} \mathbf{v} \cdot \mathbf{ds} \\
&= \int_{x_0}^{x} \frac{-y_0}{\xi^2 + y_0^2} d\xi + \int_{y_0}^{y} \frac{x}{x^2 + \eta^2} d\eta \\
&= \arctan \frac{y_0}{\xi} \Big|_{x_0}^{x} + \arctan \frac{\eta}{x} \Big|_{y_0}^{y} \\
&= \arctan \frac{y_0}{x} - \arctan \frac{y_0}{x_0} + \arctan \frac{y}{x} - \arctan \frac{y_0}{x} \\
&= \arctan \frac{y}{x} - \arctan \frac{y_0}{x_0} = \arctan \frac{y}{x} + c\,.
\end{aligned}$$

7.8 Vektorpotentiale

Bei der Behandlung und Definition von Potentialfeldern und Potentialen hatten wir durch die Identität

$$\operatorname{rot}(\operatorname{grad} \Psi) = \mathbf{0}$$

für alle zweimal stetig differenzierbaren skalaren Funktionen Ψ ein Kriterium zur Überprüfung der Potentialfeldeigenschaft eines Vektorfeldes $\mathbf{v}$, d.h. die Forderung

$$\operatorname{rot} \mathbf{v} = \mathbf{0}\,,$$

erhalten. Wenn man die Divergenz der Rotation eines Vektorfeldes $\mathbf{w}$, also $\operatorname{div}(\operatorname{rot}\mathbf{w})$, berechnet, stellt man

$$\operatorname{div}(\operatorname{rot}\mathbf{w}) = 0$$

fest. Gibt es für ein vorgegebenes Vektorfeld $\mathbf{v}$ ein Vektorfeld $\mathbf{w}$ mit

$$\mathbf{v} = \operatorname{rot}\mathbf{w}\ ,$$

so muss notwendigerweise die Bedingung $\operatorname{div}\mathbf{v} = 0$ gelten. Diese Überlegungen führen auf den Begriff des Vektorpotentials.

Definition 7.11. (Vektorpotential)
Sei $\mathbf{v} : D \to \mathbb{R}^3,\ D \subset \mathbb{R}^3$, gegeben. Existiert ein differenzierbares Vektorfeld $\mathbf{w} : \mathbb{R}^3 \to \mathbb{R}^3$ mit

$$\mathbf{v} = \operatorname{rot}\mathbf{w}\ ,$$

so heißt $\mathbf{w}$ **Vektorpotential** von $\mathbf{v}$.

Die Vorüberlegungen zur Definition 7.11 führen auf das folgende Kriterium zur Entscheidung, ob ein vorgegebenes Vektorfeld $\mathbf{v}$ ein Vektorpotential besitzt.

Satz 7.6. *(Kriterium für die Existenz eines Vektorpotentials)*

Sei $\mathbf{v} : D \to \mathbb{R}^3,\ D \subset \mathbb{R}^3$, ein differenzierbares Vektorfeld. Ist D eine offene konvexe Menge, dann ist die Bedingung

$$\operatorname{div}\mathbf{v} = 0$$

notwendig und hinreichend für die Existenz eines Vektorpotentials $\mathbf{w}$ mit $\mathbf{v} = \operatorname{rot}\mathbf{w}$. Statt der Forderung der Konvexität von D reicht hier auch die schwächere Forderung, dass D einfach zusammenhängend ist.

Bei der Berechnung von Vektorpotentialen gibt es eine Analogie zur Berechnung von skalaren Potentialen. Bei skalaren Potentialen war die Berechnung bis auf eine additive Konstante durch die Ansatzmethode und die Kurvenintegral-Methode immer möglich. Hat man ein Vektorpotential $\mathbf{w}_0$ von $\mathbf{v}$ mit $\mathbf{v} = \operatorname{rot}\mathbf{w}_0$ gefunden, und ist $\mathbf{w}_1$ ein Potentialfeld, dann ist $\mathbf{w} = \mathbf{w}_0 + \mathbf{w}_1$ ebenfalls ein Vektorpotential von $\mathbf{v}$, weil

$$\mathbf{v} = \operatorname{rot}\mathbf{w} = \operatorname{rot}\mathbf{w}_0 + \operatorname{rot}\mathbf{w}_1 = \operatorname{rot}\mathbf{w}_0$$

gilt, da $\operatorname{rot}\mathbf{w}_1 = \mathbf{0}$ für das Potentialfeld $\mathbf{w}_1$ ist. Vektorpotentiale sind also nur bis auf Potentialfelder eindeutig. Potentialfelder spielen bei Vektorpotentialen eine ähnliche Rolle wie die Konstanten bei skalaren Potentialen.

Will man für ein vorgegebenes Vektorfeld ein Vektorpotential berechnen, muss man mit einem geeigneten Ansatz beginnen. Im folgenden Beispiel soll dies demonstriert werden.

Beispiel: Für das auf ganz $\mathbb{R}^3$ definierte Vektorfeld $\mathbf{v}(x,y,z) = (xy, xz, -zy)^T$ stellen wir $\operatorname{div}\mathbf{v} = 0$ fest und damit die Existenz eines Vektorpotentials $\mathbf{w} = (w_1, w_2, w_3)^T$. Da $\mathbf{v} = \operatorname{rot}\mathbf{w}$ gelten muss, stehen die Gleichungen

$$xy = \frac{\partial w_3}{\partial y} - \frac{\partial w_2}{\partial z} \tag{7.19}$$

$$xz = \frac{\partial w_1}{\partial z} - \frac{\partial w_3}{\partial x} \tag{7.20}$$

$$-zy = \frac{\partial w_2}{\partial x} - \frac{\partial w_1}{\partial y} \tag{7.21}$$

zur Berechnung von w_1, w_2 und w_3 zur Verfügung. Das ist nicht eindeutig möglich. Wir fixieren $w_3 = c_3 = \text{const.}$ Damit erhält man nach Integration der Gleichung (7.19) nach z

$$w_2 = -xyz + C(x,y)\ .$$

Die Integration der Gleichung (7.20) ergibt

$$w_1 = x\frac{z^2}{2} + D(x,y)\ .$$

Man sieht nun, dass sich aus der dritten Gleichung (7.21) mit den bisherigen Ergebnissen für w_1 und w_2 die Beziehung

$$-zy = -zy + \frac{\partial C(x,y)}{\partial x} - \frac{\partial D(x,y)}{\partial y}$$

ergibt. Nun kann man $C(x,y)$ vorgeben und $D(x,y)$ durch

$$D(x,y) = \int \frac{\partial C(x,y)}{\partial x}\, dy$$

berechnen. Z.B. erhält man $D(x,y) = \sin y + c_1$ für $C(x,y) = x\cos y + c_2$. Für $C = \text{const.}$ folgt $D = \text{const.}$ Damit erhält man die Vektorpotentiale

$$\mathbf{u}(x,y,z) = \begin{pmatrix} x\frac{z^2}{2} + \sin y + c_1 \\ -xyz + x\cos y + c_2 \\ c_3 \end{pmatrix} \quad \text{bzw.} \quad \mathbf{w}(x,y,z) = \begin{pmatrix} x\frac{z^2}{2} + c_1 \\ -xyz + c_2 \\ c_3 \end{pmatrix}$$

mit reellen Konstanten c_1, c_2, c_3, und erkennt damit auch noch, dass $(\sin y, x\cos y, 0)^T$ ein Potentialfeld ist.

7.9 Aufgaben

1) Bilden Sie die Divergenz der NAVIER-STOKES-Gleichung (7.2) (bei $\nu = \text{const.}$) und vereinfachen Sie die Gleichung unter Nutzung der Gleichung (7.3), wobei eine entsprechende Glattheit des Vektorfeldes $\mathbf{v}$ und der skalaren Funktion p für die zu bildenden Ableitungen vorausgesetzt wird.

2) Berechnen Sie die Länge der Kurve

$$\gamma(t) = (t, e^t, 2)^T,\ t \in [0{,}4].$$

3) Berechnen Sie das skalare Kurvenintegral der Funktion $f(x,y,z) = \frac{x^2-y}{z}$ entlang dem Nordpolarkreis (arctic circle, geographische Breite $\phi_n = 67{,}5^o$, wobei der Äquator und der Nordpol die geographische Breite 0^o bzw. 90^o haben). Dabei nehmen wir die Erde als Kugel mit dem Radius $r = 6400\,km$ an.

4) Überprüfen Sie das Vektorfeld

$$\mathbf{w}(x,y,z) = \begin{pmatrix} yz\cos(xyz) + 2xz \\ xz\cos(xyz) + 2yz^2 \\ xy\cos(xyz) + x^2 + 2y^2z \end{pmatrix}$$

auf die Potentialfeldeigenschaft und berechnen Sie gegebenenfalls eine Stammfunktion.

5) Zeigen Sie, dass die folgenden Vektorfelder $\mathbf{v} : G \to \mathbb{R}^n$ eine symmetrische JACOBI-Matrix haben, aber keine Gradientenfelder in G sind. Woran liegt das?

$$(a)\ \mathbf{v}(x,y,z) = \tfrac{c}{x^2+y^2}\begin{pmatrix} -y \\ x \\ 0 \end{pmatrix}, \quad G = \{(x,y,z) \in \mathbb{R}^3 \mid (x,y) \neq (0{,}0)\}$$

$$(b) \quad \mathbf{w}(x,y) = \left(\begin{smallmatrix} \frac{y}{x^2+y^2}+y \\ x-\frac{x}{x^2+y^2} \end{smallmatrix}\right), \quad G = \{(x,y) \in \mathbb{R}^2 \mid (x,y) \neq (0{,}0)\}\ .$$

6) Berechnen Sie das Arbeitsintegral des Vektorfeldes $\mathbf{v} : \mathbb{R}^3 \to \mathbb{R}^3$, $\mathbf{v}(x,y,z) = (2xy + z^3, x^2, 3z^2x)^T$ längs des Kurvenstücks $\gamma(t) = (t{,}1-t{,}1)^T$, $0 \leq t \leq 1$.

7) Berechnen Sie die Arbeit, die Sie verrichten müssen, um in einem fahrenden Zug, der in 10 Sekunden gleichmäßig von 0 auf 120 $\frac{km}{h}$ beschleunigt, in einer Zeit von maximal 10 Sekunden 10 m in Fahrtrichtung zu gehen. Dabei gehen wir davon aus, dass Sie eine Masse von 75 kg haben.
Hinweis: Egal ob Sie die 10 m in einer, fünf oder zehn Sekunden zurücklegen. Die verrichtete Arbeit ist gleich.

8) Berechnen Sie das vektorielle Kurvenintegral des Vektorfeldes

$$\mathbf{v}(x,y,z) = \begin{pmatrix} ye^{xy} \\ xe^{xy} \\ z^2 \end{pmatrix} \text{ entlang der Kurve } \gamma(t) = \begin{pmatrix} \cos t \\ \sin^2 t \\ t^2 \sin t \cos t \end{pmatrix},\ t \in [0{,}2\pi]\ .$$

9) Überprüfen Sie, ob das Vektorfeld $\mathbf{v}(x,y,z) = (y,z,x)^T$ ein Vektorpotential besitzt und berechnen Sie gegebenenfalls ein Vektorpotential $\mathbf{w}$ von $\mathbf{v}$.

10) Berechnen Sie das Integral des Vektorfeldes $\mathbf{v}(x,y,z) = (-\frac{y}{x^2+y^2}, \frac{x}{x^2+y^2}, 2)^T$ entlang der Schraubenlinie

$$\gamma(t) = (\cos t, \sin t, t)^T,\ t \in [0{,}2\pi]\ .$$

!F"." G 4 4 !) * ! !6 - "-4' -
D -/ '

@ !"#$! ! !% & ' ! 6 - "-4' -
D -/ ' 0-
I& E 8 '0 @ 0 ' # ' 4 / ' - + . ' . / D -/ '

6 Special Relativity and Fourier Transform Theory for Electrodynamics 183
6.1 Lorentz Boosts 183
6.2 More Formal Way to Define 4-Vectors and Other Tensors 186
    6.2.1 Contravariant Vectors 186
    6.2.2 Covariant Vectors 187
    6.2.3 A Note on the Einstein Summation Convention 189
6.3 Tensors 189
    6.3.1 Contravariant Tensors 189
    6.3.2 Covariant Tensors 190
    6.3.3 The Mixed Tensor 191
    6.3.4 Quotient Theorem and the Metric Tensor 192
6.4 Inner Product and Raising and Lowering of Tensor Indices 193
6.5 Lorentz Transformations 195
    6.5.1 Lorentz Boost with Arbitrary Velocity Direction 198
6.6 Why Tensors in Physics? 199
6.7 Brief Description of 4-Velocity, 4-Momentum and 4-Force for a Massive Particle 199
6.8 Relevant Portions of Fourier Transform Theory 201
    6.8.1 The Product of Fourier Transforms: Convolution 206
6.9 Problems 206

!F"." G 4 4 !) * ! !6 - . : ' " & 9 '

< !"#$! ! !% & ' ! <(<E = E 0-
V 6 - : ' ' - ' , " & 9 '
%9 : 4 - ;;; 6 0 ' M ;;;

E !"#$! ! !% & ' ! 6 - ((: ' (" & 9 ' 0-

I& 8 6 0 0'

I& <<8 EE ' ' -

I& 8 0 ' +
V %#.7%>%G": :6GD"# .#"G +%# ".6%G

!"#$! ! !% & ' !
6 - (. (: ' (" & 9 '(0 ' ("00 ' (' -(. / 0-

I& < 8 E@ '0 = 6 - 0'

'0 = { 6 - 0' <
<
= %#.7%>%G": "GF %#.7%G%# ": <@
= %#.7%>%G": ".#6 D E
= F \ ' - # E
= "00 ' # ' ' - # w E=

= " :6 ".6%G $# +" .%#6S".6%G
= = 7?# .#6"G>?:"#6S".6%G .7D%#D E
= @ %#.7%>%G": #%AD .6%G "GF)D . " #%W6 ".6%G
= < #D": H D.#6 ".#6 D
= E 6G>?:"# ":?D FD % % 6.6%G
= " :6 ".6%G :D" . $?"#D % .6 6S".6%G
= %4 - - /
= FD+6G6.6%G5 DW" :D 5 "GF #% D#.6D <

!F"." G 4 4 !) * ! ! / F & ' & '0 /

!"#$! ! !% & ' ! / F & ' & '0 /
- D- 0-

!F"." G 4 4 !) * ! ! " ' / " 9
7 9 U :' 9

!"#$! ! !% & ' ! = " ' / " 9
7 9] :' 9 ' 9 -& ? 4 / 8 0-

I& 8 < '0 E > .

8 Geometric theories

8.1 Systems of geometric rules

(a) The geometric rule scheme. We used first natural deduction as the logical calculus that is extended by rules. The axioms covered were formulas without essential disjunctions. With sequent calculus as the logical calculus to be extended, any universal axioms could be converted into rules. In this chapter, we shall show how the class of axioms convertible into rules is further extended into **geometric implications**, as in Chapter 5, but without the restrictions that natural deduction brings with itself. We recall the definition of a geometric implication:

Definition 8.1. *A formula in first-order logic is* **geometric** *if it does not contain* $\supset$ *or* $\forall$. *A* **geometric implication** *has the form, with* A *and* B *geometric formulas,*

$$\forall x \dots \forall z(A \supset B).$$

A **geometric theory** *is a theory axiomatized by geometric implications.*

The examples of geometric axioms mostly encountered in mathematics are **existential axioms**. An existential axiom can replace a construction and postulated properties of constructed objects. It has in this case the form $\forall x \dots \forall y \exists z A(x, \dots, y, z)$ and corresponds to the construction of some z from any given $x, \dots, y$, such that $A(x, \dots, y, z)$ holds. Constructions can have conditions, as in elementary geometry where an intersection point of two lines, say, can be constructed only if the lines are convergent. The existential axiom together with the condition can be given, when suitably formulated, as a geometric implication.

In lattice theory, the propositional part of existential axioms consists of atomic formulas $P(x, y, z)$. The rule that corresponds to such an existential axiom is, with parameters a, b in place of the universally quantified variables:

$$\frac{P(a, b, z), \Gamma \rightarrow \Delta}{\Gamma \rightarrow \Delta}\,{\scriptstyle E\text{-}Rule}$$

The rule has the variable restriction that the eigenvariable z must not be free in the conclusion. Assuming the premiss of the rule, application of the logical rule $L\exists$ followed by $L\forall$ twice gives as a conclusion

$$\forall x \forall y \exists z P(x, y, z), \Gamma \rightarrow \Delta$$

A cut with the axiomatic sequent $\rightarrow \forall x \forall y \exists z P(x, y, z)$ gives the conclusion of the rule. On the other hand, application of $R\exists$ to the initial sequent $P(x, y, z) \rightarrow P(x, y, z)$ followed by the rule and universal generalizations gives the axiom, so it follows that the rule has the same force as the existential axiom.

We shall give in the next Subsection (b) concrete examples of theories axiomatized by geometric implications.

Proposition 8.2. Canonical form for geometric implications. *Geometric implications can be reduced to conjunctions of formulas of the form*

$$\forall \overline{x}(P_1 \& \ldots \& P_m \supset \exists \overline{y}_1 M_1 \vee \ldots \vee \exists \overline{y}_n M_n)$$

Each P_i *is an atomic formula, each* M_j *is a conjunction of a list of atomic formulas* $\overline{Q}_j$*, and none of the variables in the vectors* $\overline{y}_j$ *are free in* P_i*.*

The **geometric rule scheme** that corresponds to geometric axioms has the form

Table 8.1 The geometric rule scheme

$$\frac{\overline{Q}_1(\overline{z}_1/\overline{y}_1), \overline{P}, \Gamma \rightarrow \Delta \quad \ldots \quad \overline{Q}_n(\overline{z}_n/\overline{y}_n), \overline{P}, \Gamma \rightarrow \Delta}{\overline{P}, \Gamma \rightarrow \Delta} GRS$$

The variables $\overline{y}_i$ are called the **replaced variables** of the scheme, and the variables $\overline{z}_i$ the **proper variables**, or **eigenvariables**. In what follows, we shall consider for ease of notation the case in which the vectors of variables $\overline{y}_i$ consist of a single variable. All the proofs can be adapted in a straightforward way to the general case.

The geometric rule scheme is subject to the following condition that expresses in a logic-free way the role of the existential quantifier in a geometric axiom:

Condition. *The eigenvariables must not be free in* $\overline{P}, \Gamma, \Delta$*.*

(b) Examples of geometric theories. We give some examples of elementary theories that can be given a geometric axiomatization. In some cases, as in plane projective geometry (example 3), a careful choice of the basic notions is needed, or else the theory fails to be geometric. It turns out that the standard axiomatizations of projective and affine geometry use what were called **co-geometric implications** in Chapter 5, and that a geometric axiomatization is found by the use of the basic notions of intuitionistic geometry. We shall analyse further this phenomenon when presenting the duality between classical and constructive notions and proofs in the next chapter.

1. Robinson arithmetic. The language has a constant 0, a unary successor function s, binary functions $+$ and $\cdot$. Atomic formulas have the form $r = t$, for arbitrary terms r and t. For free variables, parameters $a, b, c, \dots$ are used.

1. $\neg\, s(a) = 0$,
2. $s(a) = s(b) \supset a = b$,
3. $a = 0 \vee \exists y\, a = s(y)$,
4. $a + 0 = a$,
5. $a + s(b) = s(a + b)$,
6. $a \cdot 0 = 0$,
7. $a \cdot s(b) = a \cdot b + a$.

A classically equivalent axiomatization is obtained if 3 is replaced by:

$$3'.\ \ \neg\, a = 0 \supset \exists y\, a = s(y).$$

The axiom is not geometric because it has the implication $a = 0 \supset \perp$ in the antecedent of an implication.

2. Ordered fields and real-closed fields

I Axioms for non-degenerate linear order:
  1. $a \leqslant a$,
  2. $a \leqslant b \vee b \leqslant a$,
  3. $a \leqslant b \,\&\, b \leqslant c \supset a \leqslant c$,
  4. $\neg\, 1 \leqslant 0$.

II Axioms for ordered additive groups:
  5. $(a + b) + c = a + (b + c)$,
  6. $a + b = b + a$,
  7. $a + 0 = a$,
  8. $\exists y\, a + y = 0$,
  9. $a \leqslant b \supset a + c \leqslant b + c$.

III Axioms for multiplication:
10. $(a \cdot b) \cdot c = a \cdot (b \cdot c)$,
11. $a \cdot b = b \cdot a$,
12. $a \cdot 1 = a$,
13. $a = 0 \vee \exists y\, a \cdot y = 1$,
14. $a \cdot (b + c) = a \cdot b + a \cdot c$,
15. $a \leqslant b \,\&\, 0 \leqslant c \supset a \cdot c \leqslant b \cdot c$.

A classically equivalent axiomatization is obtained if in place of 13 the following axiom is chosen:

$$13'.\ \neg\, a = 0 \supset \exists y\, a \cdot y = 1.$$

This axiom, however, is not geometric, for the same reason as above.
Real-closed fields are obtained by adding the axioms that state the existence of square roots and zeros of polynomials of odd degree:

16. $0 \leqslant a \supset \exists y\, a = y \cdot y$,
17. $a_{2n+1} = 0 \vee \exists y\, a_{2n+1} \cdot y^{2n+1} + a_{2n} \cdot y^{2n} + \dots a_1 \cdot y + a_0 = 0$.

The classically equivalent axiomatization with 17 replaced by

$$17'.\ \neg\, a_{2n+1} = 0 \supset \exists y\, a_{2n+1} \cdot y^{2n+1} + a_{2n} \cdot y^{2n} + \dots a_1 \cdot y + a_0 = 0$$

is clearly not geometric.

3. Classical projective geometry with constructions. The basic concepts are: equality of two points, equality of two lines, and incidence of a point on a line. These are written $a = b$, $l = m$, and $a \in l$. There are two constructions: those of a connecting line $ln(a, b)$ of two points and of an intersection point $pt(l, m)$ of two lines.

I Axioms for the equality relations:

$$a = a, \quad a = c \,\&\, b = c \supset a = b,$$

$$l = l, \quad l = n \,\&\, m = n \supset l = m.$$

II Axioms of incidence:

$$a = b \vee a \in ln(a, b), \quad a = b \vee b \in ln(a, b),$$

$$l = m \vee pt(l, m) \in l, \quad l = m \vee pt(l, m) \in m.$$

III Uniqueness axiom:

$$a \in l \,\&\, b \in l \,\&\, a \in m \,\&\, b \in m \supset a = b \vee l = m.$$

IV Substitution axioms:

$$a \in l \,\&\, a = b \supset b \in l,$$

$$a \in l \,\&\, l = m \supset a \in m.$$

V Existence of three non-collinear points:

$$\exists x \exists y \exists z(\neg\, x = y \,\&\, \neg\, z \in ln(x, y)).$$

Because of axiom V, the above is not a geometric theory. We obtain a geometric axiomatization by the use of apartness between points and lines as a basic notion, instead of equality:

4. Constructive projective geometry with constructions. The basic concepts are: $a \neq b$, $l \neq m$, and $a \notin l$. The constructions are: $ln(a, b)$ and $pt(l, m)$.

I Axioms for apartness relations:

$$\neg\, a \neq a, \quad a \neq b \supset a \neq c \vee b \neq c,$$

$$\neg\, l \neq l, \quad l \neq m \supset l \neq n \vee m \neq n.$$

II Axioms of incidence:

$$\neg\,(a \neq b \,\&\, a \notin ln(a, b)),$$

$$\neg\,(a \neq b \,\&\, b \notin ln(a, b)),$$

$$\neg\,(l \neq m \,\&\, pt(l, m) \notin l),$$

$$\neg\,(l \neq m \,\&\, pt(l, m) \notin m).$$

III Uniqueness axiom:

$$a \neq b \,\&\, l \neq m \supset a \notin l \vee b \notin l \vee a \notin m \vee b \notin m.$$

IV Substitution axioms:

$$a \notin l \supset a \neq b \vee b \notin l,$$

$$a \notin l \supset l \neq m \vee a \notin m.$$

V Existence of three non-collinear points:

$$\exists x \exists y \exists z(x \neq y \,\&\, z \notin ln(x, y)).$$

As can be seen, we have now a geometric axiomatization of projective geometry. The classical basic concepts of equality and incidence are defined as negations of the basic concepts of constructive geometry, namely apartness of two points, apartness of two lines, and the 'outsideness' of a point from a line. On the other hand, these latter concepts can be defined as negations of the basic concepts of classical projective geometry, if the law of double negation is assumed: then, for example, $\neg\, a \notin l$ will be equivalent to $a \in l$.

Transitivity of equality in the classical axiomatization was given in the 'Euclidean' form, because, put in this way, it comes directly out as the contrapositive of the 'co-transitivity' of apartness.

Chapter 10 is devoted to an analysis of plane projective and affine geometry.

8.2 Proof theory of geometric theories

We shall turn after the above examples of geometric theories to the general proof theory of systems of geometric rules. In the extension of sequent calculi by such rules, all the structural properties are preserved.

Definition 8.3. *Let* **T** *be a geometric theory. Then* **G3cT** (**G3imT**) *is the system of sequent calculus obtained by adding to* **G3c** (**G3im**) *the geometric rules that correspond to the geometric axioms of* **T**, *together with the rules that arise from the closure condition.*

Theorem 8.4. Equivalence of axiomatic systems and rule systems. *A geometric axiom is derivable from the corresponding geometric rule. Conversely, a geometric rule is derivable from the corresponding geometric axiom in* **G3imT**+Ctr+Cut.

Proof. A geometric axiom $\to A$, represented as a sequent with an empty antecedent, is derivable from the corresponding geometric rule scheme. Below, an asterisk * denotes possibly iterated steps of a rule, the premisses are clearly derivable by $R\&$, and $M_i(z_i/y_i) \equiv \&\overline{Q}_i(z_i/y_i)$:

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\overline{Q}_1(z_1/y_1), \overline{P} \to M_1(z_1/y_1), \dots, M_n(z_n/y_n)}{\overline{Q}_1(z_1/y_1), \overline{P} \to \exists y_1 M_1, \dots, \exists y_n M_n}{\scriptstyle R\exists^*} \quad \dots \quad \dfrac{\overline{Q}_n(z_n/y_n), \overline{P} \to M_1(z_1/y_1), \dots, M_n(z_n/y_n)}{\overline{Q}_n(z_n/y_n), \overline{P} \to \exists y_1 M_1, \dots, \exists y_n M_n}{\scriptstyle R\exists^*}}{\overline{P} \to \exists y_1 M_1, \dots, \exists y_n M_n}{\scriptstyle GRS}}{\overline{P} \to \exists y_1 M_1 \lor \dots \lor \exists y_n M_n}{\scriptstyle R\lor^*}}{P_1 \& \dots \& P_m \to \exists y_1 M_1 \lor \dots \lor \exists y_n M_n}{\scriptstyle L\&^*}}{\to P_1 \& \dots \& P_m \supset \exists y_1 M_1 \lor \dots \lor \exists y_n M_n}{\scriptstyle R\supset}}{\to \forall \overline{x}(P_1 \& \dots \& P_m \supset \exists y_1 M \lor \dots \lor \exists y_n M_n)}{\scriptstyle R\forall}$$

Conversely, a geometric rule is derivable from the corresponding geometric axiom $\to A$ in **G3im**+*Ctr*+*Cut* as shown below. In the derivation of the left premiss of cut, inverses of rules are used. These are admissible (and height-preserving) steps in **G3im**. Observe that the variable restriction in *GRS* now comes into use in the steps of $L\exists$:

$$\dfrac{\dfrac{\dfrac{\dfrac{\to \forall\overline{x}(P_1\&\ldots\&P_m \supset \exists y_1 M_1 \vee \ldots \vee \exists y_n M_n)}{\to P_1\&\ldots\&P_m \supset \exists y_1 M_1 \vee \ldots \vee \exists y_n M_n}{\scriptstyle R\forall Inv}}{P_1\&\ldots\&P_m \to \exists y_1 M_1 \vee \ldots \vee \exists y_n M_n}{\scriptstyle R\supset Inv}}{\overline{P} \to \exists y_1 M_1 \vee \ldots \vee \exists y_n M_n}{\scriptstyle L\&Inv} \quad \dfrac{\dfrac{\dfrac{\overline{Q}_1(z_1/y_1), \overline{P}, \Gamma \to \Delta}{M_1(z_1/y_1), \overline{P}, \Gamma \to \Delta}{\scriptstyle L\&^*}}{\exists y_1 M_1, \overline{P}, \Gamma \to \Delta}{\scriptstyle L\exists} \quad \ldots \quad \dfrac{\dfrac{\overline{Q}_n(z_n/y_n), \overline{P}, \Gamma \to \Delta}{M_n(z_n/y_n), \overline{P}, \Gamma \to \Delta}{\scriptstyle L\&^*}}{\exists y_n M_n, \overline{P}, \Gamma \to \Delta}{\scriptstyle L\exists}}{\exists y_1 M_1 \vee \ldots \vee \exists y_n M_n, \overline{P}, \Gamma \to \Delta}{\scriptstyle L\vee^*}}{\dfrac{\overline{P}, \overline{P}, \Gamma \to \Delta}{\overline{P}, \Gamma \to \Delta}{\scriptstyle Ctr^*}}{\scriptstyle Cut}$$

QED.

Remark. It is clear by the above derivation that the geometric rule scheme hides a cut on the formula $\exists y_1 M_1 \vee \ldots \vee \exists y_n M_n$. The substituted variables y_i are bound variables in the virtual cut formula, so it will be convenient to regard the substituted variables of the scheme as bound variables and to assume that in a derivation the sets of free and bound variables are disjoint.

The inversion lemmas for the propositional rules that hold for **G3c** and **G3im** hold for their geometric extensions as well, because the geometric rule scheme has only atomic formulas as principal and active formulas. So we have:

Lemma 8.5. Inversion. *All the inversions of the propositional rules that hold for* **G3c** *and* **G3im** *hold also for their geometric extensions.*

For the inversions of $L\exists$ and $R\forall$, we need to add a condition on the variable to avoid clashes with the proper variables of the geometric rules in the derivation.

Lemma 8.6. Substitution. *Given a derivation of* $\Gamma \to \Delta$ *in* **G3cT** (**G3imT**), *with x a free variable in* Γ, Δ, *t a term free for x in* Γ, Δ *and not containing any of the variables of the geometric rules in the derivation, we can find a derivation of* $\Gamma(t/x) \to \Delta(t/x)$ *in* **G3cT** (**G3im**) *with the same height.*

Proof. By induction on the height of the given derivation. For the logical rules, the proof is contained in lemma 4.1.2 of *Structural Proof Theory*, so we need to consider only the cases that arise from the addition of the geometric rule scheme. Suppose the last rule in the derivation of $\Gamma \to \Delta$ is GRS, with premisses

$$\overline{Q}_i(z_i/y_i), \overline{P}, \Gamma' \to \Delta$$

for $i = 1, \dots, n$. The $\overline{Q}_i$ are atomic and the term t is free for x in these premisses; thus by the induction hypothesis we get derivations of

$$\overline{Q}_i(z_i/y_i)(t/x), \overline{P}(t/x), \Gamma'(t/x) \to \Delta(t/x)$$

Since x is a free variable in Γ, Δ, by the remark at the end of the previous section we have $x \neq y_i$, and since the z_i are not free in $\overline{P}, \Gamma, \Delta$ we have $x \neq z_i$. Moreover, by hypothesis, t does not contain any of the y_i. Therefore the two substitutions in $\overline{Q}_i$ are independent and we have $Q_i(z_i/y_i)(t/x) \equiv Q_i(t/x)(z_i/y_i)$. Since t does not contain any of the z_i, the z_i are not free in $\overline{P}(t/x), \Gamma'(t/x) \to \Delta(t/x)$, so we can apply the geometric rule scheme to the premisses

$$\overline{Q}_i(t/x)(z_i/y_i), \overline{P}(t/x), \Gamma'(t/x) \to \Delta(t/x)$$

Now we get $\overline{P}(t/x), \Gamma'(t/x) \to \Delta(t/x)$, i.e., $\Gamma(t/x) \to \Delta(t/x)$. QED.

Lemma 8.7. Inversion for quantifier rules.

(i) *If* $\vdash_n \exists x A, \Gamma \to \Delta$ *and* y *is not among the variables of the geometric rules in the derivation, then* $\vdash_n A(y/x), \Gamma \to \Delta$.

(ii) *If* $\vdash_n \Gamma \to \Delta, \forall x A$ *and* y *is not among the variables of the geometric rules in the derivation, then* $\vdash_n \Gamma \to \Delta, A(y/x)$.

Proof. (i) By induction on height of derivation n, as in Lemma 6.2. If $n = 0$, then $\exists x A, \Gamma \to \Delta$ is either an initial sequent, or a conclusion of $L\bot$, or a conclusion of a zero-premiss geometric rule. In each case also $A(y/x), \Gamma \to \Delta$ is an initial sequent, or a conclusion of $L\bot$, or a conclusion of a zero-premiss geometric rule, thus $\vdash_0 A(y/x), \Gamma \to \Delta$.

If $n > 0$ and $\exists x A$ is principal in the last rule, the premiss gives a derivation of $A(z/x), \Gamma \to \Delta$, where z is not free in Γ, Δ. By Lemma 8.6, using the substitution y/z, we obtain a derivation of the same height of $A(y/x), \Gamma \to \Delta$. If $\exists x A$ is not principal in the last rule, we argue as in lemma 4.2.3 of *Structural Proof Theory* in the case the last rule is a logical rule. If the last rule is a geometric rule, with $\Gamma \equiv \overline{P}, \Gamma'$ and premisses

$$\overline{Q}_1(z_1/y_1), \overline{P}, \exists x A, \Gamma' \to \Delta, \dots, \overline{Q}_n(z_n/y_n), \overline{P}, \exists x A, \Gamma' \to \Delta$$

we get by the assumption that free and bound variables are disjoint and by the inductive hypothesis derivations of the sequents

$$\overline{Q}_1(z_1/y_1), \overline{P}, A(y/x), \Gamma' \to \Delta, \dots, \overline{Q}_n(z_n/y_n), \overline{P}, A(y/x), \Gamma' \to \Delta$$

Since y is none of the z_i, we can apply the geometric rule scheme to these premisses and obtain a derivation of $\overline{P}, A(y/x), \Gamma' \to \Delta$.

(ii) Similar to (i). QED.

Without loss of generality, we can assume by Lemma 8.6 that the following condition on variables is satisfied:

Disjointness condition. *In a derivation in* **G3cT** (**G3imT**), *the collections of proper variables of any two geometric rules are disjoint.*

Theorem 8.8. *The rules of weakening*

$$\frac{\Gamma \to \Delta}{A, \Gamma \to \Delta}{\scriptstyle LW} \qquad \frac{\Gamma \to \Delta}{\Gamma \to \Delta, A}{\scriptstyle RW}$$

are height-preserving admissible in **G3cT** *and in* **G3imT**.

Proof. Consider the last step in the derivation of the premiss of the weakening rules. Apply, inductively, weakening to the premisses $\Gamma_i \to \Delta_i$ of the last step to obtain $A, \Gamma_i \to \Delta_i$, then the rule that had been used as the last step, which gives $A, \Gamma \to \Delta$. Observe that if the last rule is geometric and the weakening formula A contains some of its variables, the variable condition is no longer satisfied after weakening with A. In this case the substitution lemma is applied to the premisses of the geometric rule to have new free variables that do not clash with those in A. The conclusion is then obtained by applying the inductive hypothesis and the geometric rule. QED.

Theorem 8.9. *The rules of contraction*

$$\frac{A, A, \Gamma \to \Delta}{A, \Gamma \to \Delta}{\scriptstyle LC} \qquad \frac{\Gamma \to \Delta, A, A}{\Gamma \to \Delta, A}{\scriptstyle RC}$$

are height-preserving admissible in **G3cT** *and in* **G3imT**.

Proof. By simultaneous induction for left and right contraction on the height of the derivation of the premiss. If it is an initial sequent, also the conclusion is. If the last rule is a propositional rule, then the conclusion follows as in theorem 3.2.2 of *Structural Proof Theory*. If it is $L\forall$, we apply the induction hypothesis to the premiss of the rule, and then the rule, and similarly if it is $L\exists$ with A not principal in it. If it is $L\exists$ with $A \equiv \exists x B$ and premiss $B(y/x), \exists x B, \Gamma \to \Delta$, by the variable condition on the geometric rule and the Remark after Theorem 8.4, y is not a variable in any geometric rule in the derivation, so we can apply the inversion lemma for $L\exists$ instantiated by y and obtain a derivation of $B(y/x), B(y/x), \Gamma \to \Delta$. By the induction hypothesis we get $B(y/x), \Gamma \to \Delta$ and by $L\exists$, $\exists x B, \Gamma \to \Delta$.

If the last rule is a geometric rule, we distinguish three cases: 1. No occurrence of A is principal in the rule. 2. One occurrence of A is principal, the other is not. 3. Both occurrences of A are principal.

The first case is handled by a straightforward induction and the second by the repetition of the principal formulas P_i in the premisses of the geometric rule. Finally, the third case is taken care of by the closure condition. QED.

We are now in the position to prove the admissibility of cut for our rule systems for geometric theories. We remark that the proof has the same structure as the proof of admissibility of cut for universal theories (Theorem 6.9 above), with an additional use of the substitution lemma in order to meet the variable restriction in the geometric rule scheme.

Theorem 8.10. *The rule of cut*

$$\frac{\Gamma \to \Delta, A \qquad A, \Gamma' \to \Delta'}{\Gamma, \Gamma' \to \Delta, \Delta'}{\scriptstyle Cut}$$

is admissible in **G3cT** *and in* **G3imT**.

Proof. By induction on the length of A with a subinduction on the sum of the heights of the derivations of $\Gamma \to \Delta, A$ and $A, \Gamma' \to \Delta'$. We need to consider only the cases that arise from the addition of the geometric rule scheme. The other cases are treated in the corresponding proof for **G3c** and **G3im**.

1. If the left premiss is a zero-premiss geometric rule, then also the conclusion is a zero-premiss geometric rule, because these have an arbitrary context as succedent.

2. If the right premiss is a zero-premiss geometric rule with A not principal in it, the conclusion is a zero-premiss geometric rule for the same reason as in case *1*.

3. If the right premiss is a zero-premiss geometric rule with A principal in it, A is atomic and we consider the left premiss. The case that it is a geometric zero-premiss rule is covered by *1*. If it is an initial sequent with A not principal, the conclusion is a logical axiom; otherwise, Γ contains the atom A and the conclusion follows from the right premiss by weakening. In the remaining cases we consider the last rule in the derivation of $\Gamma \to \Delta, A$. Since A is atomic, A is not principal in the rule. Let us consider the case of

a geometric rule (the others being dealt with similarly, except $R\supset$ and $R\forall$ that are covered in *4*). The derivation ends with

$$\dfrac{\dfrac{\overline{Q}_1(z_1/y_1),\,\overline{P},\,\Gamma''\to\Delta,A \quad \dots \quad \overline{Q}_n(z_n/y_n),\,\overline{P},\,\Gamma''\to\Delta,A}{\overline{P},\,\Gamma''\to\Delta,A}\,{\scriptstyle GRS} \qquad A,\Gamma'\to\Delta'}{\overline{P},\,\Gamma',\Gamma''\to\Delta,\Delta'}\,Cut$$

The cut cannot be simply permuted to the premisses of *GRS* because Γ', Δ' may bring in free variables that clash with the proper variables z_i and thus prevent the application of *GRS* after the cuts. Instead, we apply first the substitution lemma to the right premiss of cut and replace all the variables z_i, if any, by fresh variables w_i, and denote the substitution by w/z. Observe that by the variable condition in *GRS*, the substitution does not affect the cut formula A. We obtain the sequent

$$A,\Gamma'(w/z)\to\Delta'(w/z)$$

Now n cuts with the n premisses of *GRS* give, for $i = i, \dots n$, the n sequents

$$\overline{Q}_i(z_i/y_i),\,\overline{P},\,\Gamma'(w/z),\Gamma''\to\Delta,\Delta'(w/z)$$

By applying *GRS* to these n premisses we get

$$\overline{P},\,\Gamma'(w/z),\Gamma''\to\Delta,\Delta'(w/z)$$

The derivation is continued as before with the substitution (w/z) performed globally. Observe that by the disjointness condition, the substitution does not affect the active formulas of other geometric rules in the derivation. The cut has thus been replaced by n cuts with a left premiss that has a derivation of lower height and a right premiss of unaltered height.

Let us now consider the cases in which neither premiss is an axiom.

4. A is not principal in the left premiss. These cases are dealt with as above, with cut permuted upwards to the premisses of the last rule used in the derivation of the left premiss (with a suitable variable substitution to match the variable restrictions in the cases of quantifier rules and the geometric rule-scheme), except for $R\supset$ and $R\forall$ in **G3imT**. By the intuitionistic restrictions in these rules, A does not appear in the premisses, and the conclusion is obtained without cut by $R\supset$ ($R\forall$, resp.) and weakening.

5. A is principal only in the left premiss. Then A has to be a compound formula. Therefore, if the last rule of the right premiss is a geometric rule, A

cannot be principal in the rule, because only atomic formulas are principal in geometric rules. In this case cut is permuted to the premisses of the right premiss, with appropriate substitution of free variables as in *3*, if the right premiss is a *GRS*. If the right rule is a logical one with A not principal in it, the usual reductions are applied.

6. A is principal in both premisses. This case can involve only logical rules and is dealt with as in the usual proof for pure logic. QED.

Examples of proof analyses in geometric theories will be given in Chapter 10. In typical cases, it turns out that existential axioms are conservative relative to the derivability of atomic cases from atomic assumptions. If for the rest of the rules a subterm property can be shown, a positive solution for the said derivability problem follows. In this chapter, we shall present instead a general result about the proof theory of geometric theories.

8.3 Barr's theorem

We apply here the method of extension by rules to a general result for geometric theories.

Barr's theorem. *Let T be a geometric theory and let A be a geometric implication. If* **G3cT** $\vdash\ \rightarrow A$, then **G3imT** $\vdash\ \rightarrow A$.

Proof. Let A be $\forall\overline{x}(B \supset C)$ and consider a proof of A in **G3cT**. Because B and C do not contain $\supset$ or $\forall$, the derivation of A consists of geometric rules, rules for &, $\vee$, $\exists$, a step of $R\supset$ and steps of $R\forall$. The geometric rules can occur in any order among the logical rules, however, of the logical rules, $R\supset$ and $R\forall$ come last. The geometric rules have the same succedent in both the premisses and the conclusion, so therefore rules $R\supset$ and $R\forall$ are applied to single-succedent sequents. It follows that the given proof must already be a proof in **G3imT**. QED.

Therefore, through the method of proof analysis, the result reduces to a triviality: A classical proof of a geometric implication in a geometric theory formulated as a sequent system with rules is already an intuitionistic proof. If we add the requirement that the geometric implication must not contain $\perp$ in the antecedent, then the classical proof is even a proof in minimal logic.

Notes to Chapter 8

Interest in the study of geometric theories has arisen from different areas of logic and mathematics. In topos theory, geometric formulas are characterized as the fragment of first-order logic preserved by geometric morphisms. The statement of Barr's theorem in that context is as follows: *For any Grothendieck topos* **E***, there is a Boolean topos* **B** *and a surjective geometric morphism* $\gamma : \mathbf{B} \to \mathbf{E}$. Thus the conservativity result for geometric theories is proved using a completeness theorem for geometric theories in Grothendieck topoi and the construction of a suitable Boolean topos out of a Grothendieck topos (cf. Johnstone 1977, Mac Lane and Moerdijk 1992, p. 513). The preservation property extends beyond first-order logic to existential fixed-point formulas (cf. Blass 1988).

Palmgren (2002) indicates a proof-theoretical proof of Barr's theorem by showing that geometric implications are stable under the Dragalin-Friedman translation. A proof for the special case of the empty geometric theory is suggested in Troelstra and Van Dalen (1988) (exercise 2.6.14) by means of Kripke models.

The nature of geometric logic as the logic of finite observations has been emphasized in localic approaches to constructive topology (cf. Vickers 1989). Geometric theories can be treated, as any other theory, by the addition of Hilbert-style axioms to a logical proof system, but axiomatic systems are hard to analyse proof-theoretically. In Simpson (1994), geometric theories are presented through suitable rules that extend intuitionistic natural deduction. A proof of normalization for the extensions thus obtained is given and the systems applied in a systematic study of the proof theory of intuitionistic modal logic. Finally, in Coste, Lombardi, and Roy (2001), the so-called method of 'dynamical proof', which establishes the derivability of one atom from a finite set of atoms, is applied to certain specific geometric theories such as the theory of algebraically closed fields.

We observe that Barr's theorem is not a characterization of the intuitionistic fragment of geometric theories, because we can go beyond geometric implications and maintain the conservativity result. First, following Dragalin's suggestion (cf. Section 3.7.3 in Troelstra and Schwichtenberg 2000), we can modify the intuitionistic left rule for implication by admitting a multisuccedent conclusion in the left premiss

$$\frac{A \supset B, \Gamma \to \Delta, A \quad B, \Gamma \to \Delta}{A \supset B, \Gamma \to \Delta}\ {\scriptstyle L\supset}$$

Rule $L\supset$ of the classical calculus, without $A \supset B$ in the left premiss, is then admissible in the modified intuitionistic calculus; thus the difference between the intuitionistic and classical sequent systems is confined to rules $R\supset$ and $R\forall$. An operational definition of formulas for which the conservativity of classical derivations holds can be given: if a formula is derivable classically in a geometric theory and the derivation contains no steps of $R\supset$ and $R\forall$ with a non-empty context in the premiss, then the derivation is an intuitionistic derivation. However, this is an empty characterization that states nothing but that 'an intuitionistic derivation is an intuitionistic derivation'. A characterization in terms of only the form of the formulas, not of their derivations, would be desirable. There are classes of formulas, such as geometric implications, the form of which forces the derivation to be of the stated kind. The same is true, for example, if the formula does not contain in its positive part implications or universally quantified formulas as components of a disjunction. Even so, there are still formulas outside the mentioned classes for which the conservativity holds.

!"#$! ! !% & ' ! = " ' / " 9
7 9] :' 9 ' 9 -& ? 4 / 8 0-

!"#$! ! !% & ' ! = ' . / 8 E 0-
%9 : 4 - ;;;
6 0 ' M ;;; D ' - ' ' '

!F"." G 4 4 !) * ! !. & ' ' ' -
- ' ' ' - L ' 0

!"#$! ! !% & ' ! @ 6' 1 / . : & +' '
" - 6 - ' ' ' - $ ' %0 0-
V @= 0'&
%9 : 4 - ;;;

!F"." G 4 4 !) * ! !. 0 &/ ' - "00 , ' + , -

= !"#$! ! !% & ' ! < . 0 &/ ' - "00 , ' + , -
0-
V =E 0'&

%9 : 4 - ;;;

6 0 ' M ;;;

)' 0
. 0 & ' 0'
. G . 0 & ' 0'
0' . 0 & <
G ' - 4 & . 0 &/
. 0 & ' 0'
: ' . 0 &
"9 9 & ' -)' ' & <
0' ' - 0 <
' - : ' 4 , /
= 3'9 . 0 & ' 0' E
@ + F ' . 0 & ' 0' =
< . B '* . 0 &/ . 0 & ' 0'
' - B '*

. 0 &/ . F ' =
E L @@
. + K ? / 0 / <
? ' 0' <E
"- 9 + E
= + , - . E@
@ G ,0' 4 '00 &
" + , -
') x + , - ' - + , -
+ - & x + , - 4 / B + , -
" ^1 _ D,

"00 , ' + , - G ,0' 4 '00 &
? 9 - - E
+ - & x + , - B . " G + , - <
= +' '00 & E
@ '00 & B x + , -
< x + , -

"00 , ' + , - ? ' 0' =
. "00 , ' =
. ' B { 6 . ' / % - - E
/ "00 ' 9 + , - ==
& ' > 9' 6 + ==
/ "00 ' 9 + , -
4 , 0' <
"00 , ' + , - 9/ ' +
4 & / E
= "00 , ' + , - . 0 & ' 0' E<
= . 11 0 ' - " + , - E<
= "00 , ' + , - L
= % : 0 3 ' - "00 , ' : 0 3 + , -
0
= % ~8W5 S "00 , ' + , - 0 /
. 0 & ' 0' =
= . B '* "00 , ' + , - 0 /
3'9 : ' / 4 , 0' ' - L
= . B '* "00 , ' + , - 0 /
G 3'9 : ' / 4 , 0'
= "00 , ' + , - G
= "00 ' " / 0 + , - . /
= D, : B '* F '
DL '
#
6 - , <

=

20220107///

archive.org/details/

- questioniriguard00enri
→ lezdigeopro00enririch
- lezionidigeometr00bartuoft
- lezionidegeoanal00castrich

pdfdrive.com

20220108/// Google Drive Share

Z-Math-Jagua

e.jaguariuna.sp.gov.br

20200109///

Era de Planck.

Roger Penrose ⇒ universo cíclico
denominado
"cosmologia cíclica conformada"

Procedimento matemático complexo
denominado "reescalonamento conformado"

→

"Reescalonamento conformado"

uma transformação geométrica que na verdade altera o tamanho de um objeto, mas deixa sua forma inalterada.

20220110M Google Share Drive

Z-Math-Cisnes →
educacioncisnes.cl

Smiljan (cidade do Tesla)

---

tripofobia.

20220110/ Graham-Farmelo/

Graham Farmelo - The Universe Speaks in Numbers...(2019).epub

Pág. 82

Carl Friedrich Gauss are examples of post-Newtonian thinkers who were at the front lines of both mathematics and physics ...

20220111

Geometry & Symmetry
L. Christine Kinsey
Teresa E. Moore
978-0-470-49949-8

—

Tucídedes — Guerra do Peloponeso.
↓ —
Armadilha de Tucídedes.

—

Potência hegemônica x Potência Emergente

—

O caminho da Guerra
Graham Allison
→ "Os Estados Unidos e a China conseguirão escapar da armadilha de Tucídedes".

História da Guerra do Peloponeso
Tucídedes

"Vitória Pírrica" — vitória na Batalha de Ásculo

—

Plato's Stepchildren - Star Trek.
(film)

---

softarchive.is/books/

ID 922 9587 8661. (Zoom)

20220112 ///

Collegii Urbis Nov Eborac
Sigillum

Antigo-001	1447	1564
Antigo-002	117	
Caderno-001	400	1284
002	179	
003	309	
004	396	
Volume-001	539	
002	410	
003	664	
004	756 (756)	
005	3654	
006	298	
007	643	
008	3304	
009	561	
010	121	
011	103	

Volume 012 - 155
013 - 43
(11251)

Total:
1564
1284
11251

14099 ///

20220113 //

20220115 /// 119 Gb.

ARQ-R-B-F --- 20220115.77

=

20220116 //

Acknowledgements = reconhecimentos

—

Helicópteros - corda de guinada (yaw string)
→ sentido do vento em relação à aeronave.

—

20220117 Adamski - Killer (music)

—

Dec 21, 2021 Nytimes.com
Page A21 ///
A Box of Cash, a Secret Donor and a Big Gift for some N.Y.C. Studentes
by Corey Kilgannon

—

data . .

S T Q Q S S D

"That which does not kill me, makes me stronger"

F. Nietzche

—

20220119

matematicaitapetininga.com.br

math.itp@ifsp.edu.br

—

20220120

matematica.company 50,00

—

lincoln-academy.org

—

⇒ td.msgsuite.workers.dev

—

Mané Garrincha — 20 Jan 1983

Elza Soares — 20 Jan 2022 (91 anos)

(39 anos depois)

—

1962 - 1982 → relacionamento conturbado.

—

Making Geometry - Jon Allen
Pág. 8(7) -
God is a circle centre is everywhere and whose circumference is nowhere.
Variously attributed.

20220127
Non-Euclidean Geometry
by Roberto Bonola.
1955
https://download.tuxfamily.org/openmathdep/geometry_advanced/Non-Euclidean_Geometry-Bonola.pdf

obs: https://download.tuxfamily.org/openmathdep/

Pág. 236 (252)

Gg Klein's Direct Demonstration of the Independence of Projective Geometry from the Fifth Postulate.

other book:

Straight_Lines_and_Curves Vasilyev.pdf

Source:

20220127/Recherches géométriques sur la théorie des parallèles-Gauss/Geometry-01/

20220128

topfreebooks.org

20220129

https://gdz.sub.uni-goettingen.de/id/PPN235957348

data
S T Q Q S S D

20220131

Geologists – 4200 years the
"Meghalayan Age".

Era – Cenozoic ← Quaternary

Quaternary ← Pleistocene
epochs Holocene
↓
11.700 years ago.

—

128 Gb – 44.822 arquivos

—

ARQ-R-B-F---20220131.7z

—

20220201

Acólito → auxiliar

—

20220201.

Criação do diretório:

DATA-NovoVivo-003-2022

—

Everything for Everyone in a Fractal Model
M.S. Carey. ~~ISBN~~ 1549890190

⇒ ISBN 9781549890192

—

20220202 ☆

⇒ 20200526 – Southwestern College
(creation)
→ google share drive

—

O remorso é o passado se destruindo no presente.

—

20220203

vdoc.pub

—

Pappus – CXIV (Pág 111)

Pappus D'Alexandrie
La Collection Mathématique
Paul Ver Eecke 1933

=

20220207

Desde le Punto a la Cuarta
Dimension.
Egmont Colerus

—

sm001-internet@yahoo.com

sm001-pc@yahoo.com

—

Trusted Traveler USA
https://ttp.dhs.gov

—

Summa Musichorum.

—

antípoda.

História da Igreja
Daniel ~~Hobbs~~ Hops

—

2022-0208M

Egmont Colerus von Geldern

→ De Pitágoras a Hilbert.

Del punto a la cuarta dimension.

"Von Pythagoras bis Hilbert" Egmont Colerus.
359 pág.

ISBN 978-3893501526

Vom Punkt zur Vierten Dimension:
Geometrie fur Jedermann

Geometry and Discrete Mathematics
Waterloo.

Abebooks.com

20220209 (Dikran Tahta) → Armênia

Dick Tahta.

Professor na escola St Albans (sul de Inglaterra) (1955) → Stephen Hawking.

→ influenciou Dent White que embora em tenha descoberto o Teorema de Pick, no que se referia a mim, esse era "o meu teorema".

Hermenêutica → técnica de interpretar e explicar um texto ou discurso

↳ termo matemático. (defini-lo precisamente)

Tate

→ "a graça da sua atenção"

Tate inspirou a John Hibbs: "Deixe a matemática falar".

freeditorial.com

---

https://phaidra.univie.ac.at/detail/o:1229964?mode=full

Livro Sebo do Messias: Correios 05652234890BR

www.math.rwth-aachen.de/homes/Ulrich.Schoenwaelder/

Pappus Eecke
Collection Mathématique

---

https://schule.bayernport.com

Pappus
https://archive.org/details/
PappusDAlexandrieEd.Paul
verEeckeLaCollectionMathmatique

---

jandaia

202202100150

Deus é uma realidade, Ele existe por si mesmo e não depende da análise objetiva ou o entendimento das pessoas.

Deus pode ser entendido de várias formas de acordo com as convicções do homem que tenta entender Deus, mas não devemos fazê-lo de forma a conceituá-lo simplisticamente hermenêutico.

O nada existe.

O conceito do nada se reflete na criação da noção do número zero, criado na Índia, ou melhor, descoberto e esboçado na forma como conhecemos, primeiramente descoberto pois esta realidade já existia, faltando somente a sua materialização no mundo fático ou "real", quando houve a retirada de uma pedra da sua posição matemática quando da resolução de uma operação matemática básica em que o espaço em que faltava a pedra, ou seja, o nada ou o zero existe independentemente da existência de uma pedra posicional para demarcar a existência daquele referido número.

O nada existe porque existe a Matemática.

A Matemática existe por si mesma.

A Matemática é descoberta, o homem inventa a forma como nós descrevemos os conceitos matemáticos.

A Matemática tem várias facetas e nós podemos adequá-la a cada realidade existente e que por sinal é infinita, para poder entendê-la, mas a Matemática não pode ser distorcida ou corrompida pelo homem.

Na Realidade, nós devemos tentar adequar a realidade à Matemática, pois para cada realidade nós teremos que conceituá-la de acordo com a Matemática imutável, cabendo a nós, meros mortais, inventar meios, métodos, formas para "tentar" descrever a nossa realidade dentro da Matemática.

Três premissas nós devemos ter em mente:

1) O nada sempre existiu, existe e existirá.
O nada é o reflexo oposto de Deus, uma face da mesma essência de Deus.
O Universo somente pode existir pois havia o nada.
A expansão do Universo também depende do nada.
O nada permitiu o Universo

se expandir, o nada suplica que o Universo o preencha. O espaço decorre do nada, o espaço não é criado pelo Universo, mas alterado por este.

↳ O Tempo sempre existiu, existe e existirá.
O tempo não é criado pelo Universo, mas alterado por este.

O nada existe e portanto, ele tem o seu próprio tempo e espaço.

2) A Matemática sempre existiu, existe e existirá.
A Matemática está em tudo para ser descoberta.
A realidade depende da Matemática, inclusive o seu próprio entendimento depende dela. Cabe a nós, descrever a realidade utilizando a Matemática.
A Matemática tem infinitos facetas e formas para descrever as infinitas realidade que ainda serão descobertas.

A Matemática e o nada decorrem de Deus.

O nada obriga, ocasiona a criação de alguma coisa. /

O nada foi dado por Deus como a

chave que abre a porta para infinitas possibilidades.

O nada é como esta folha em branco na qual eu estou escrevendo. Esta página deseja ser preenchida pois se ela estivesse completa, o que eu poderia escrever para tentar inventar, descobrir, descrever, narrar, contar etc.

O tudo não permite a criação pois ele já ocupou tudo e não permite a inclusão de mais nada.

Deus é tudo. Ele não precisa dos homens e por isso, Ele criou o nada como uma forma de criar espaço para que o homem possa viver e sonhar.

Deus se "ausentou" e nos deu o "livre-arbítrio" para que nós os homens possamos preencher o nada.

São Paulo 10 de fevereiro de 2022.

202202100240

data

Ⓢ Ⓣ Ⓠ Ⓠ Ⓢ Ⓢ Ⓓ

20220210M

O nada existe.
O nada permite que o Universo exista.
Literalmente:

O Universo foi criado do nada.

O nada que existe, criou o Universo.

O nada criou condições para que o Universo fosse criado.

O nada criou o Universo, além dos confins do Universo, nós iremos encontrar o nada.

O nada impulsiona o Universo a preenchê-lo de forma acelerada, ou seja, quanto mais nada temos, mais rápido o Universo tentará preenchê-lo.

O nada decorre de Deus.

O nada é a tábua rasa na qual Deus, permite ao homem interagir e existir.

O nada sempre existiu, mas pela sua essência, ele sempre se transforma de infinitas formas, uma delas é o Universo que conhecemos.

O nada por si mesmo cria
condições para que algo decorra
dele.

O Universo não foi criado,
ele é a manifestação do
nada.
(uma das manifestações).

O nada existe,
somente o que existe,
poderá criar algo;
neste caso,
o Universo.

O nada também é uma forma
de energia.

Tudo é energia,
a falta de energia também
é uma forma de energia.
(estado)

20220210||0507

ALL ———————— NOTHING
(Tudo) (NADA)

↓

SYMMETRY

↓

COLAPSO

↓

UNIVERSO

↓

CURVO (forças fundamentais física)

↓ ↓

MIRROR (sobre um plano) gravity (decorrência dessas forças)

↓

PROJECTIVE
ORTHOGONAL. (geometry)

Obs: EUCLID (PLANO)

RIEMANN
HILBERT | NON-EUCLIDEAN GEOMETRY
GAUSS

↑

HELMHOLTZ (extensão de Euclides e não uma nova geometria)

EGMONT COLERUS
↳ PUNKT

20220210

20220210\The Art of Problem Solving\

Paul Zeitz - The art and craft of problem solving - John Wiley & Sons (2017).pdf

Pág. 62(78).

—

ALL — NOTHING

Positivo — Negativo

In — Yang

obs: Ver livro Tao da Física.

Nome do Livro:

NOTHING IS ALL

20220211

sebo → Matematica para escuelas tecnicas
R$38,50

—

love is wise, hatred is foolish
in the world.

((o amor é sábio)
o amor é o ~~ódio~~ sábio)

O amor é sábio, o ódio é tolo no
mundo.

—

www.sastatus.com
↳ softarchive.is

—

sanet.st .ws	ebook3000.com
kupdf.net	libgen.is }
pdfdrive.com	libgen.li }

—

Zero – Charles Seife.

The Nothing that is: A natural History of
Zero – Robert Kaplan

—

20220212_0003

A gravidade como força independente não existe.

A gravidade decorre da força forte (lei fundamental da física)

A curvatura do tecido espaço-tempo decorre da Força Forte.

A soma de cada Força Forte de cada átomo reunido em um determinado espaço gera por consequência a ação gravitacional. Quanto mais átomos, mais forte será a soma das Forças Fortes de todos os átomos reunidos em determinado espaço, que por sua vez, irá distorcê-lo de acordo com a reunião de todas as Forças Fortes de todos os átomos reunidos em um determinado espaço.

A curvatura do Universo existe e decorre de toda a massa deste Universo.

Onde não há massa no Universo, existe uma mínima curvatura que é decorrente da massa do Universo como um todo.

O nada existe, como o vazio do Universo também existe.

A existência destes dois gera também uma força!!!

A força gerada pelo nada e
pelo vazio decorre da simetria
intrínseca existente o Tudo (Deus)
e o nada.

A Entropia decorre desta
interação simétrica destes dois
elementos.

A falta de paridade de
aniquilação entre matéria
e anti-matéria decorre
da interação intrínseca
da ação simétrica entre
estas duas forças universais
atuantes no Universo analisado.
(em calta)

A falta desta paridade ou
simetria decorre da interação
entre o Tudo e o nada, ou seja,
não se trata de falta de paridade
mas a ação permanente do
equilíbrio entre estas duas forças
que interagem na
criação e destruição
do Universo.

entraremos no campo da filosofia e
da religião, Deus Shiva na
cultura indiana.
(estudar!!!) 2020212 0018

20220212$_{m}$

20220213$_{u}$

Escaneei o livro:

Desde el punto a la Cuarta
Dimensión
Egmont Colerus

→ está inserido no
Volume-014-20220213

Source:
ARQ/1-Source/Volume-014-20220213/

Obs: Este caderno foi escaneado
até aqui
20220213 1050

! "#$%$ & ' ' "( ) "

! "#$%$ & ' ' "( ) " "$*'+ * ,+% - $ +* +
$.. + / + + * 0* +

1 ! "$/2" " "3 4 + " 5
5 6( )6$*'+ *,+% -7 $ +* + .*

! "#$%$ & ' ' "( ) " "$ + + $!
8 *

1 ! "$/2" " "3 4 + " $ + + $
! 8 * . 4 +4 (9 * : 4 ; 1 .*
: , ' * <<<

! "#$%$ & ' ' "( ) " "0 +4 4 + =

1 ! "$/2" " "3 4 + " 0 +4 4 + =
. 4 ; 1 .*
>4 ? ; 1

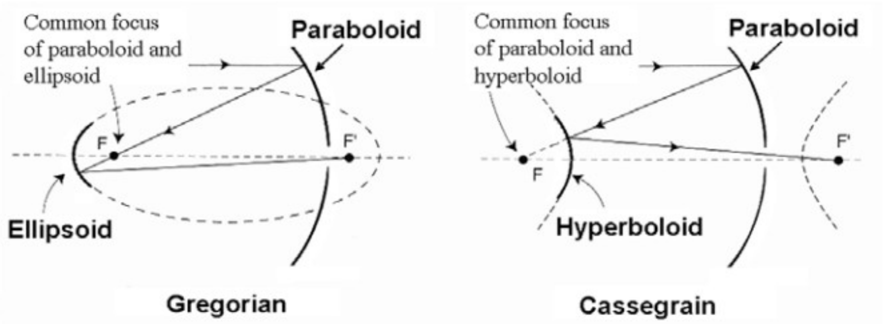
Common focus of paraboloid and ellipsoid
Paraboloid
F
F'
Ellipsoid
Gregorian
Common focus of paraboloid and hyperboloid
Paraboloid
F
F'
Hyperboloid
Cassegrain

Figure 3.5. The principle behind the Gregorian and Cassegrain telescopes which use the properties of conic sections instead of spherical shapes.

! "#$%$ & ' ' "( ) " "0' = 4 0' = + @ + +
*

1 ! "$/2" " "3 4 + " 0' = 4 0' = + @ + +
* .*
3: , ' * <<<
A = =

! "#$%$ & ' ' "( ) " "@ * + *
+ + * + = ' +

B1 ! "$/2" " "3 4 + " B @ * + *
+ + * + = ' + . 4 ; B1 .*

>4 ; 1 + @+ + + +C
$ + 4 =D . . * E + ' . :
! . + = + + + = . + * + + =
+ . * + . + E + * + 7 F ' F 0+ '
+ * * : +: : ' + +..+ + D
.)=
% = += . : + . + + + + +
+ += + C. * = + = * . + + + 7 + F
+ 4 : : '+ + GC * . * 4 '
* + + + + 4 + * .+ + +C @ E + 4
* + =D . + * . '+ + + + . G *
+ 4 7 + +)= +* 0 9+ =D * ' =
+ + * +' . . G +
.+ + +C !=4 := @ H (+ + * + 4 = + %
+ 4 + * + = + + + + * ; I1 + * . *
G 4 . .+ + +C + = + =
+ 9 ..+ F + * ? ?F * * .+ + +C ' F
+ + . + * #+ + +) 9 ..+
+ + *+=D + + + ; @ 4 1

3: , ' * <<<

! "#$%$ & ' ' "( ) " "9 4 + = =
2 + +

1 ! "$/2" " "3 4 + " 9 4 + = =
2 + + $ + + + = ; B1 .*
A = =

! "#$%$ & ' ' "( ) " " + + = * 0*

1 ! "$/2" " "3 4 + " + + = * 0* .*
>4 ; ?1 3 : + 0

>4 ; 1 % % (*= :

3: , ' * <<<

! "#$%$ & ' ' "( ) " ") % J '
$ 4

1 ! "$/2" " "3 4 + "
6) 6% 6J ' 6 6# 6$ 4 .*
>4 ; B1
9 ' F + 4 + . * ' F
4 = + +.. + * . 4 . . 7 .
+ + = .+ + = ' : + 4 4 + +
+ + + . F * ' : + + * . 4
* + = .+ (*) .+ F
+ + 4 = . * + F * :: *
4 +' F + .+ 4 +' = % * + + +
4 2 + * 8/ + = E + 4 +' = (+

) C. + + 4 = 4 +
C 4 +' F + * : + = . =
+ + + . = + + =F +' +.. +
= * * + + + = = E + 4 +' =

! "#$%$ & ' ' "( ) " " . = +

?1 ! "$/2" " "3 4 + " ? #K % + ( )6 (4 7* +
. = 0C. + * #K ; ?1 .*

1 ! "$/2" " "3 4 + " ? #K : 4 #K % +
( )F (4 7* . = 0C. + * #K ; ?1 .*
>4 B
7 + + + H + 9+ C * . C . + :=
+** 4 +4 + = + E + L C. + +
. * + # .+

1 ! "$/2" " "3 4 + " ? % + ( ) (4 7* + . =
0C. + * #) : ; ?1 .*

1 ! "$/2" " "3 4 + " ? . = + . :
3: $ + + * MN

! "#$%$ & ' ' "( ) " " +: = + * ! + ! +
+

1 ! "$/2" " "3 4 + " ! + + . 4
; 1 .*

1 ! "$/2" " "3 4 + " ? % + ( ) (4 7* + . =
0C. + * #) : ; ?1 .*
>4 ; 1 83!0 %90 0$/%9 8/$ 7%O @70,# (O $!0 8/$#73 0%/O

>4 ; 1

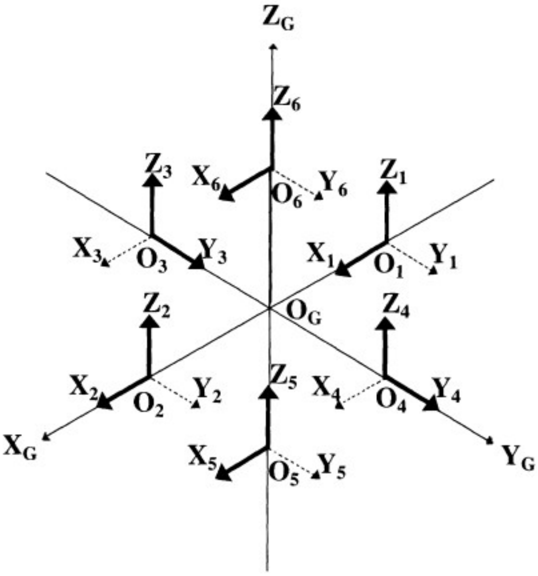
Z_G
Z_6
Z_3
X_6
O_6
Y_6
Z_1
X_3
O_3
Y_3
X_1
O_1
Y_1
O_G
Z_2
Z_4
X_2
O_2
Y_2
Z_5
X_4
O_4
Y_4
X_G
X_5
O_5
Y_5
Y_G

@ 4 ? ,+= C + + * P+ = * + *P 4 +

>4 ; 1 O 0%/70 $&# /$&K #0@7!70&!O 7& %90 3/(7%
#0%0/ 7&$%73& $/3J&# $&3%90/ ,$&0%

>4 ; 1

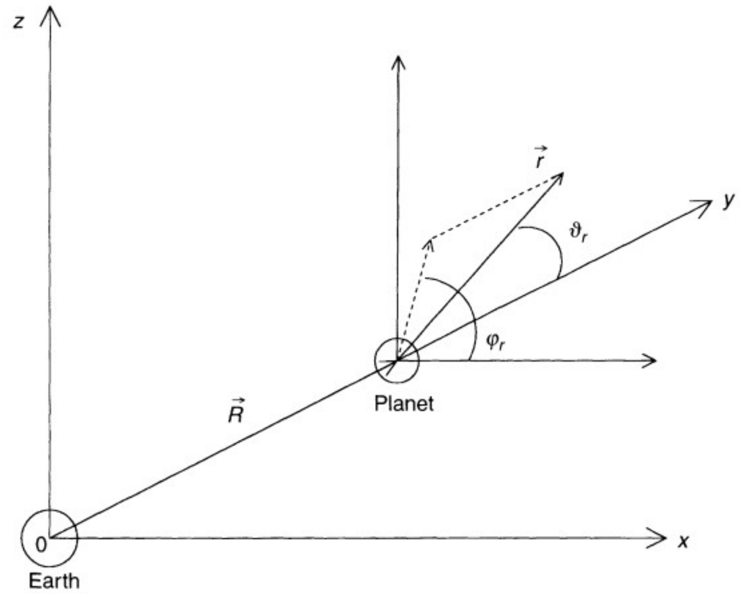
z
$\vec{r}$
y
ϑ_r
φ_r
Planet
$\vec{R}$
0
Earth
x

@ 4 % 4 + . + * + +

>4 BB; 1 %$(7,7%O 7& %90 @J,, %H3 (3#O /3(,0

>4 ?; 1 $ O %0 $%7! %J#O 3@ %90 %$(7,7%O 3@ O 0%/7!
0/73#7! 3/(7% 7& %90 ,$&$/F !7/!J,$/F /0 %/7!%0#
%9/00 (3#O /3(,0

>4 ? 3& %90 %/7$&8J,$/ 37&% 3@ %90 J& QJ 7%0/ O %0

>4 ? 7&%0/ !3 $/7 3& (0%H00& &J 0/7!$, 7&%08/$%73&
$&# $&$,O%7!$, %903/O @3/ %90 3,$/ O %0

3: % ? (*= :

>4 /3%$%73& 3@ &$%J/$, 9$ 0# $%0,,7%0 $&# %907/
3/(7%$, 3%73&

>4 @/3 K3, 383/3 P &3/ $,7R$%73& $,83/7%9
%3 %90 3/(7% 7& %90 %9/00 (3#O ,$&0%$/O /3(,0

B1 ! "$/2" " "3 4 + " +: = + * ! + ! +
+ . 4 ; ?1 .*
>4 ; 1 % : *= . : + * ,+4 + 4 +

! "#$%$ & ' ' "( ) " "% 9 = !+ 4 +. =

! "#$%$ & ' ' "( ) " "% + + 9 +' + *
0+ % 0+ = 9 = % 4 =

1 ! "$/2" " "3 4 + " % + + 9 +'
+ * 0+ % 0+ = 9 = % 4 = .*
3: , ' * <<<

! "#$%$ & ' ' "( ) " "% *= + 0+ + *
+ + =

1 ! "$/2" " "3 4 + " % *= + 0+ + *
+ + = . 4 +4 (9 * : 4 ; 1 .*
3: , ' * <<<

$ = . ' 4 +
. = . F
*) * 4 + F + *
C * * + + +.. +: =
% * . . + + +
= + + . 7 =
. = + = ' + 7
+ ' * ' : ' F
+) +.. +: =
:+ .
$: 0

! "#$%$ & ' ' "( ) "

! "#$%$ & ' ' "( ) " " + + % *+= @ : + =

! "#$%$ & ' ' "( ) " "/ . = +4 + + *
= +4 + ** $4 + * / + +

1 ! "$/2" " "3 4 + " / . = +4 + + *
= +4 + ** $4 + * / + + .*
3: , ' * <<<

>4 ; 1

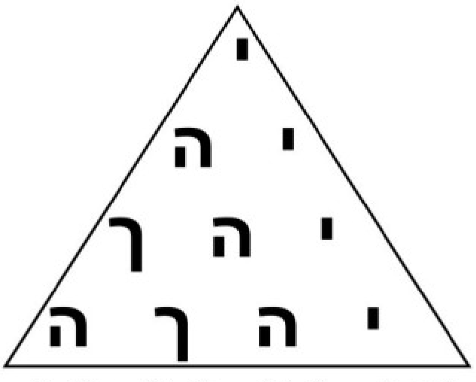
י
ה י
ך ה י
ה ך ה י

(1x5) + (2x6) + (3x5) + (4x10) = 72

@ 4 8 '+ D 4 + +

+: 9 : + 8 *

! "#$%$ & ' ' "( ) "

! "#$%$ & ' ' "( ) " "$ ( ) 3 ! ' ,) *

?1 ! "$/2" " "3 4 + " 0 9 ,) * $:) '
J ' = ; ? 1 .*

1 ! "$/2" " "3 4 + " 0 9 ,) * $ ( )
! ' !+ : *4 J ' = ; 1 .*

1 ! "$/2" " "3 4 + " 0 9 ,) * $ ( )
! ' !+ : *4 J ' = ; 1 *S'

3: , ' * <<<

! "#$%$ & ' ' "( ) " "$ ' = 3 + +

1 ! "$/2" " "3 4 + " B $ + .*
A + + + * ! . C $ + =

1 ! "$/2" " "3 4 + " B ! $ B ! . .*
A ,0!%J/0 3& 7&7 $, J/@$!0 7& 7/

>4 ; 1 % 8 = 0 . H + / . +

>4 ; 1 ? ! + %=. / + +

>4 B ;B?1 ! . + + @ % + ! '+

>4 B ; 1 % + ! '+ (+ * ! . + +

1 ! "$/2" " "3 4 + " B 8+ +. .*
A % 8+ +. ! . + + @ % + ! '+

B1 ! "$/2" " "3 4 + " B 4 +. :) .*
A $ ' = + + + + =

1 ! "$/2" " "3 4 + " B ? ? ? .*
A %90 !,$ 7!$, %903/O 3@ 7&7 $, J/@$!0

>4 ?; B1

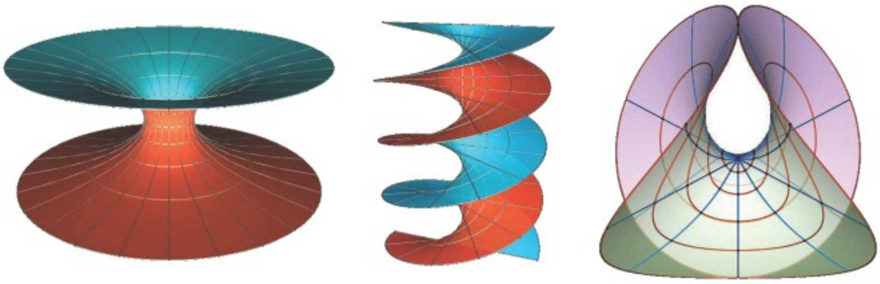

@ 4 , % + * ! % * / 4 %
0 . + 7 +4 = H :

% 0 . + T ! F 4;U1 T UF * T U *U % + + * '
* := 0 . VB W F 4 = + * + + = . +
+ + . *+ +F E '+ H +
. + B % + : ** *F + 4 U F * + * +
'+ X Y 7 + U + 4 + ;+ 4
+ + 4 F * ' * : ** * . : * *
:= + 1 + * + + ' + . + Z ' = =
@ 4 F 4 0' = + * + U + * 4 ! +
; 1 = 0 . + F : *
C * + : % + * + * 0 . D + + E
. + + / G + '+ X Y ;3 + V W1
% . 0 . D +

$$\left(\frac{y^2-x^2}{2z}+\frac{2}{9}z^2+\frac{2}{3}\right)^3-6\left(\frac{y^2-x^2}{4z}-\frac{1}{4}(x^2+y^2+\frac{8}{9}z^2)+\frac{2}{9}\right)^2=0.$$

>4 ; 1

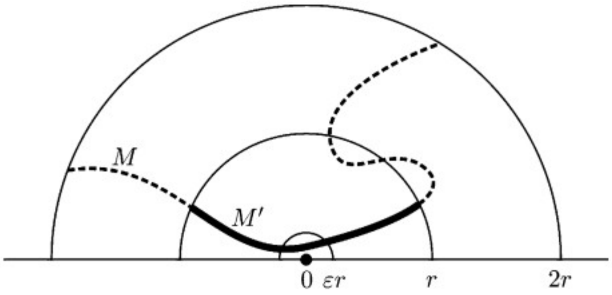

@ 4 % * * '+ +

% * + 4 * . % + *
* * '+ + F + + '+ + : * 8+ + '+ + =
: ** * + *) + + .+

! "#$%$ & ' ' "( ) " "![: : \ + [+ E

1 ! "$/2" " "3 4 + " 0* + * !+ + * + ![\:
. : \ + [+ E $: ; ? ?1 *S'

1 ! "$/2" " "3 4 + " 0* + * !+ + * + ![\:
. : \ + [+ E $: ; ? ?1 .*
>4 ? , . : \ * +..

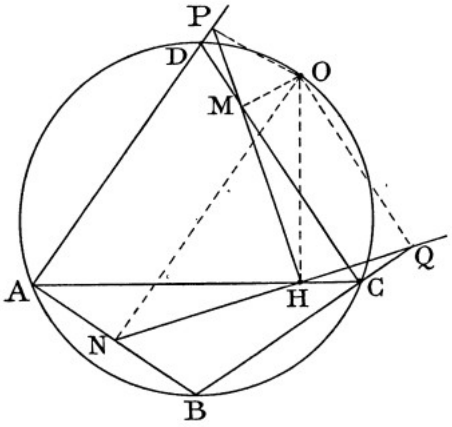

Le problème de Pappus

Etant donné plusieurs lignes droites dans le plan, trouver le lieu géométrique d'un point tel que les perpendiculaires — ou plus généralement les obliques abaissées de ce point sur ces droites sous des angles donnés — satisfassent à la condition que le produit de certaines d'entre elles soit dans un rapport constant avec le produit de toutes les autres.

I. — Cette question est relatée par Pappus, d'Alexandrie, qui, vers la fin du IV^e siècle de l'ère chrétienne, a rassemblé, dans ses *Collections mathématiques,* monument précieux de l'Antiquité, diverses découvertes des plus célèbres mathématiciens, ainsi que ses propres inventions. Ce fameux problème, dit *ad tres aut plures lineas,* avait exercé la sagacité d'Euclide (vers 300 av. J.-C.) et d'Apollonius (II^e siècle av. J.-C.), et Pappus le considérait comme l'écueil des Anciens. En effet, tenté par ceux-ci, il n'avait été résolu que dans des cas très particuliers, et les Anciens n'avaient pu atteindre la solution générale qui dépendait, en effet, d'une méthode nouvelle, telle que l'analyse algébrique ou l'art d'exprimer algébriquement les propriétés caractéristiques d'une courbe. Les Anciens avaient bien vu que, s'il n'y avait que trois ou quatre droites, le lieu était une section conique, mais encore ne pouvaient-ils la déterminer dans tous les cas. Et ils n'avaient pu aborder le problème lorsqu'il comportait plus de quatre droites, mais la propriété leur était connue, qui fut établie géométriquement par Newton (1642-1727) et qui constitue une propriété générale des coniques que, *quand un quadrilatère est inscrit dans une conique, le produit des distances de chaque point de la courbe à deux côtés opposés est au produit des distances du même point aux deux autres côtés dans un rapport constant.*

66

Le problème de Pappus

Etant donné plusieurs lignes droites dans le plan, trouver le lieu géométrique d'un point tel que les perpendiculaires — ou plus généralement les obliques abaissées de ce point sur ces droites sous des angles donnés — satisfassent à la condition que le produit de certaines d'entre elles soit dans un rapport constant avec le produit de toutes les autres.

I. — Cette question est relatée par Pappus, d'Alexandrie, qui, vers la fin du IVᵉ siècle de l'ère chrétienne, a rassemblé, dans ses *Collections mathématiques,* monument précieux de l'Antiquité, diverses découvertes des plus célèbres mathématiciens, ainsi que ses propres inventions. Ce fameux problème, dit *ad tres aut plures lineas,* avait exercé la sagacité d'Euclide (vers 300 av. J.-C.) et d'Apollonius (IIᵉ siècle av. J.-C.), et Pappus le considérait comme l'écueil des Anciens. En effet, tenté par ceux-ci, il n'avait été résolu que dans des cas très particuliers, et les Anciens n'avaient pu atteindre la solution générale qui dépendait, en effet, d'une méthode nouvelle, telle que l'analyse algébrique ou l'art d'exprimer algébriquement les propriétés caractéristiques d'une courbe. Les Anciens avaient bien vu que, s'il n'y avait que trois ou quatre droites, le lieu était une section conique, mais encore ne pouvaient-ils la déterminer dans tous les cas. Et ils n'avaient pu aborder le problème lorsqu'il comportait plus de quatre droites, mais la propriété leur était connue, qui fut établie géométriquement par Newton (1642-1727) et qui constitue une propriété générale des coniques que, *quand un quadrilatère est inscrit dans une conique, le produit des distances de chaque point de la courbe à deux côtés opposés est au produit des distances du même point aux deux autres côtés dans un rapport constant.*

Descartes (1596-1650), qui, selon l'expression de Chasles, a rendu le plus signalé service à la géométrie, par son inappréciable conception, dont on ne trouve aucun germe dans les récits des Anciens, de l'application de l'Algèbre à la théorie des courbes, créa les moyens de franchir les obstacles qui avaient arrêté les plus grands géomètres jusqu'alors, et changea véritablement la face des sciences mathématiques: Descartes a donné au problème qui nous occupe ici l'importance qu'il mérite, en même temps qu'une nouvelle célébrité, en en faisant la première application de sa *Géométrie analytique,* et en trouvant le premier, par sa nouvelle méthode, la solution générale. Le fameux problème de l'hexagone de Pascal [Pr. 55], le théorème célèbre de l'involution de Desargues [Pr. 56] peuvent être regardés comme de simples corollaires de cette proposition fondamentale.

II. — Etablissons le problème de Pappus, pour le quadrilatère ABCD inscrit dans un cercle. Il s'agit de montrer que, quelque soit le point O sur le cercle, on a :

$$OP.OQ = OM.ON.$$

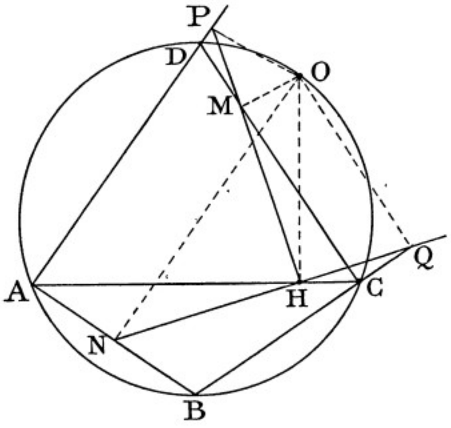

En effet, si l'on joint AC diagonale et si l'on mène OH perpendiculaire à AC, les trois points P, M, H sont en ligne droite, droite de Simson du point O pour le triangle ACD; de même les points N,H,Q sont en ligne droite, droite de Simson relative à O, et au triangle ABC. L'angle en O du quadrilatère inscriptible MOPD est égal à l'angle $\widehat{ADC}$ comme ayant même supplément $\widehat{PDM}$. L'angle $\widehat{NOQ}$ du quadrilatère inscriptible NOQB est aussi égal à l'angle $\widehat{ADC}$ comme ayant même supplément $\widehat{ABC}$. Donc $\widehat{POM} = \widehat{NOQ}$. D'autre part, les angles $\widehat{PMO}$ et $\widehat{OQN}$ sont égaux car leurs compléments respectifs DMP et HQC sont égaux au même angle $\widehat{HMC}$, angle opposé pour l'un, et pour l'autre inscrit dans le cercle passant par les points O, M, H, C, Q. Les deux triangles PMO et ONQ sont donc semblables, et leurs côtés homolgues sont proportionnels; d'où :

$$\frac{OP}{OM} = \frac{ON}{OQ} \text{ soit } OP.OQ = OM.ON.$$

III. — La géométrie analytique, ou méthode de Descartes, donne ainsi la solution générale du problème :

Soient d'abord $\alpha = 0$, $\beta = 0$, $\gamma = 0$, $\delta = 0$, les équations des côtés AB, BC, CD, DA d'un quadrilatère. Les équations :

$$\alpha\gamma = 0, \beta\delta = 0,$$

représentent deux coniques particulières passant par les quatre sommets. Donc :

$$\alpha\gamma + \lambda.\beta\delta = 0,$$

λ étant un paramètre arbitraire, est l'équation générale des coniques circonscrites du quadrilatère. Et cette équation exprime que le lieu du point dont le produit des distances à deux droites est dans un rapport constant avec le produit des distances du même point à deux autres droites, est une conique passant par les points où les deux premières droites rencontrent les deux autres.

Supposons maintenant n droites $\alpha_1 = 0$, $\alpha_2 = 0 \dots \alpha_n = 0$; et p autres droites $\beta_1 = 0$, $\beta_2 = 0 \dots \beta_p = 0$; le lieu du point tel que le rapport, du produit de ses distances aux n premières droites, au produit de ses distances aux p autres droites soit constant et égal à K, est représenté par l'équation :

$$\alpha_1 \alpha_2 \dots \alpha_n = \mathrm{K} \beta_1 \beta_2 \dots \beta_p.$$

67

Le théorème des coniques d'Apollonius

Dans l'ellipse et dans l'hyperbole, le parallélogramme construit sur deux diamètres conjugués est équivalent au rectangle construit sur les deux axes.

Respectivement pour l'une et pour l'autre de ces deux coniques, la somme ou la différence des carrés des longueurs de deux diamètres conjugués est égale à la somme ou à la différence des carrés des longueurs des deux axes.

I. — Apollonius de Perge (IIᵉ siècle av. J.-C.) fut l'auteur d'un ouvrage sur les sections coniques qui est, pour ainsi dire, le couronnement de la géométrie grecque; il conçut, le premier, les sections d'un cône oblique à section circulaire, et leur donna le nom d'ellipse, d'hyperbole et de parabole. La pensée directrice de l'auteur s'y révèle sans cesse, qui tend à lier entre elles, par leurs analogues propriétés, les différentes sections du cône; et tout le traité repose presque sur une propriété unique, qui dérive directement de la nature même des coniques. Apollonius avait déjà connaissance des plus belles propriétés des coniques, et en particulier des deux théorèmes énoncés qui lui sont dus. Nous en donnerons ici une démonstration moderne, fondée sur l'emploi des invariants des formes quadratiques.

II. — Soit, tout d'abord, rappelées, quelques propriétés des formes quadratiques. Soit ainsi une transformation sans déplacement d'origine : x, y sont les coordonnées courantes d'un point M par rapport au systèmes d'axes obliques Ox, Oy, faisant entre eux l'angle θ; X, Y sont les nouvelles coordonnées du point par rapport aux axes OX, OY, faisant entre eux l'angle θ'. Les paramètres principaux de la demi-droite OX étant α, β, et ceux de OY étant α', β', on a les formules de substitution linéaire :

$$x = \alpha \mathrm{X} + \alpha' \mathrm{Y}, \quad y = \beta \mathrm{X} + \beta' \mathrm{Y},$$

et :

$$\sin \theta' = (\alpha\beta' - \beta\alpha') \sin \theta,$$

le déterminant $\alpha\beta' - \beta\alpha' = \mu = \dfrac{\sin \theta'}{\sin \theta}$ étant le module de la substitution.

Cela posé, la forme $Ax^2 + 2Bxy + Cy^2$ devient, après la transformation de coordonnées $A'X^2 + 2B'XY + C'Y^2$, de sorte que :

$$A'X^2 + 2B'XY + C'Y^2 \equiv Ax^2 + 2Bxy + Cy^2.$$

Or, lorsque deux fonctions homogènes du second degré peuvent se déduire l'une de l'autre par une substitution linéaire, le discriminant de l'une est égal au produit du discriminant de l'autre par le carré du module de la substitution. On a donc :

$$A'C' - B'^2 = (AC - B^2)\,\mu^2,$$

d'où l'invariant :

$$\frac{A'C' - B'^2}{\sin^2 \theta'} = \frac{AC - B^2}{\sin^2 \theta}.$$

On mettrait également en évidence le nouvel invariant :

$$\frac{A' + C' - 2B' \cos \theta'}{\sin^2 \theta'} = \frac{A + C - 2B \cos \theta}{\sin^2 \theta}.$$

III. — Soit :

$$\frac{x^2}{a^2} + \frac{y^2}{\varepsilon b^2} = 1,$$

l'équation, rapportée à ses axes de symétrie, d'une conique à centre; si $\varepsilon = +1$, celle-ci est une ellipse; c'est une hyperbole si $\varepsilon = -1$. Prenons pour nouveaux axes de coordonnées deux diamètres conjugués; en appelant x', y' les nouvelles coordonnées, les formules de transformation donnent :

$$\frac{x^2}{a^2} + \frac{y^2}{\varepsilon b^2} \equiv A'x'^2 + 2B'x'y' + C'y'^2,$$

A', B', C' étant des constantes. L'équation de la conique, rapportée aux nouveaux axes, est donc :

$$A'x'^2 + 2Bx'y' + C'y'^2 - 1 = 0.$$

Mais comme ces axes sont deux diamètres conjugués, $B' = 0$. Si $\varepsilon = 1$, A' et C' sont positifs, et l'on peut poser :

$$A' = \frac{1}{a'^2}\,; \quad C' = \frac{1}{b'^2}.$$

Si $\varepsilon = -1$, A' et C' sont de signe contraire; on peut supposer $A' > 0$ et $C' < 0$, d'où l'on poserait :

$$A' = \frac{1}{a'^2}\,; \quad C' = -\frac{1}{b'^2}.$$

La nouvelle équation sera donc :

$$\frac{x'^2}{a'^2} + \frac{y'^2}{\varepsilon b'^2} - 1 = 0;$$

et on a par suite :

$$\frac{x^2}{a^2} + \frac{y^2}{\varepsilon b^2} = \frac{x'^2}{a'^2} + \frac{y^{2\prime}}{\varepsilon b'^2}.$$

Mais si l'on désigne par φ l'angle des deux diamètres conjugués, l'invariance permet d'écrire :

$$\frac{1}{a^2} + \frac{1}{\varepsilon b^2} = \frac{1}{\sin^2 \varphi}\left(\frac{1}{a'^2} + \frac{1}{\varepsilon b'^2}\right).$$

et :

$$\frac{1}{a^2 b^2} = \frac{1}{\sin^2 \varphi} \cdot \frac{1}{a'^2 b'^2},$$

soit :

$$ab = a'b' \sin \varphi, \quad a^2 + \varepsilon b^2 = a'^2 + \varepsilon b'^2,$$

d'où les deux théorèmes énoncés.

IV. — Soient OM et OM′ deux demi-diamètres conjugués d'une ellipse. Rapportons celle-ci à ses deux axes de symétrie, et soient x, y les coordonnées de M, x', y' les coordonnées de M′. On a :

$$\frac{xx'}{a^2} + \frac{yy'}{b^2} = 0, \quad \text{ou} \quad \frac{y}{b} \cdot \frac{y'}{b} = -\frac{x}{a} \cdot \frac{x'}{a}.$$

Si l'on pose :

$$\frac{x'}{a} = \lambda \frac{y}{b},$$

on a, en vertu de la relation précédente :

$$\frac{y'}{b} = -\lambda \frac{x}{a};$$

d'où l'on tire :

$$\frac{x'^2}{a^2} + \frac{y'^2}{b^2} = \lambda^2 \left(\frac{x^2}{a^2} + \frac{y^2}{b^2}\right),$$

c'est-à-dire $\lambda^2 = 1$, d'où $\lambda = \pm 1$. On a ainsi les formules dues à Chasles (1793-1880) :

$$\frac{x'}{a} = \varepsilon \frac{y}{b}, \quad \frac{y'}{b} = -\varepsilon \frac{x}{a}, \quad (\varepsilon = \pm 1).$$

En particulier, si l'on suppose OM dans l'angle xOy, le coefficient de OM′ est négatif; on peut donc supposer que le demi-diamètre OM′ est situé dans l'angle $x'Oy$. Dans ces conditions, les coordonnées de M′ sont données par :

$$\frac{x'}{a} = -\frac{y}{b}, \quad \frac{y'}{b} = +\frac{x}{a}.$$

! "#$%$ & ' ' "( ) " "0 + * + @ + + + +

?1 ! "$/2" " "3 4 + " % S ' 0 + * + @ +
+ + + 7/ ; ? 1 *S'
B .+4

1 ! "$/2" " "3 4 + " % S ' + +)= 0 + *
,+ @ + + + + 7/ ; ? 1 .*
>4 ? ; ? 1 0 + * 0 K 4 '

! "#$%$ & ' ' "( ) " "0 * + + +

1 ! "$/2" " "3 4 + " ($ + ;+ 1 0 *L
% ! + + + . 4 +4 & O) ; ???1 .*
3: , ' * <<<

1 ! "$/2" " "3 4 + " #+' * ( ) % K 4
7 .+ 0 * + * 9 0 (+ ( ) ; 1 . :

1 ! "$/2" " "3 4 + " #+' * ( ) % K 4
7 .+ 0 * + * 9 .*

! "#$%$ & ' ' "( ) " "0C * 8[[! 0C
, :

1 ! "$/2" " "3 4 + " ?
! 6* 64[[6* . ' 6 6V W6:.) ? B. .*

B1 ! "$/2" " "3 4 + " ?
! 6* 6 + [+ E 6[[+ 60C 6V W6:.) B .*
A ? .+4

>4 ;B 1

2 * . * +.. * * +] * C * [4+ +.. * [^

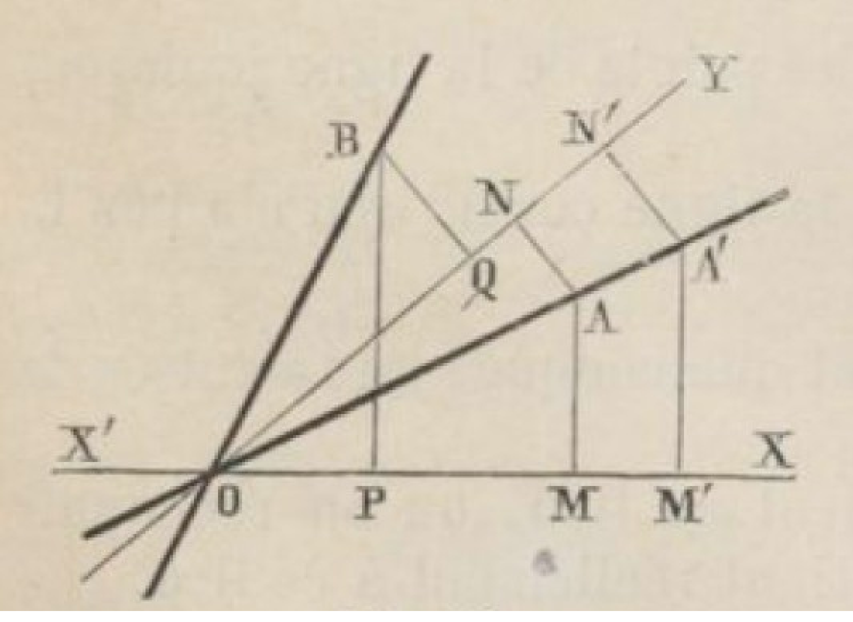
Y
B
N′
N
Q
A′
A
X′
X
O
P
M
M′

Exercice.

58. **Problème.** *Par chaque point d'une circonférence, on mène des droites parallèles sur lesquelles on prend une longueur constante* l ; *quel est le lieu des points ainsi obtenus?*

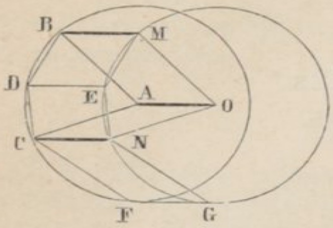

Fig. 24.

Soit CN égale et parallèle à BM.

Par le centre A menons une parallèle AO égale à l.

La figure ABMO est un parallélogramme comme ayant deux côtés opposés égaux et parallèles ; donc

$$OM = AB$$

De même $ON = AC = r$

Le lieu est donc une circonférence égale à la première.

59. **Remarque.** 1° En appliquant les conditions de l'énoncé ci-dessus à une figure quelconque, on obtiendrait aussi une figure égale.

La démonstration générale est la suivante :

Les droites BD et ME sont égales et parallèles, car BM et DE sont égales, et de même DC et EN sont égales et parallèles, et les angles BDC, MEN sont égaux comme ayant les côtés parallèles et de même sens.

Ainsi les figures BDCF, MENG sont égales comme ayant les côtés respectivement égaux et les angles égaux.

Les figures courbes sont égales comme limites de polygones égaux.

2° Dans les applications, on peut considérer la figure MENG comme ayant été obtenue par le déplacement de la figure BDCF, dont tous les sommets ont glissé sur des parallèles. Ainsi on peut dire que la figure MENG a été obtenue à l'aide de BDCF, en employant un *déplacement* ou une *translation parallèle* *.

Exercice.

60. **Problème.** *Quel est le lieu des points dont le rapport des distances à deux droites égale un rapport donné* $\frac{m}{n}$?

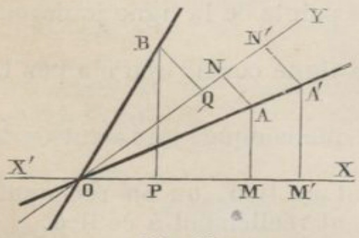

Fig. 25.

Soient OX, OY les droites données.

Le point O appartient au lieu ; soit A un point tel qu'on ait :

$$\frac{AM}{AN} = \frac{m}{n}$$

La droite AO *est le lieu demandé*, car pour tout autre point A'

on aura $$\frac{A'M'}{A'N'} = \frac{AM}{AN}$$

donc $$\frac{A'M'}{A'N'} = \frac{m}{n}$$

---

* Le terme *translation parallèle* se trouve dans un ouvrage bien remarquable : *Méthodes et théories pour la résolution des problèmes de constructions géométriques*, par JULIUS PETERSEN, professeur à l'École polytechnique de Copenhague ; traduction de M. O. CHEMIN, ingénieur des ponts et chaussées.

La droite OB *appartient aussi au lieu*, car on peut avoir

$$\frac{BP}{BQ}=\frac{m}{n}$$

Exercice.

61. Problème. *Quel est le lieu géométrique des points dont les distances à deux points donnés* A *et* B *sont dans un rapport constant* $\frac{m}{n}$?

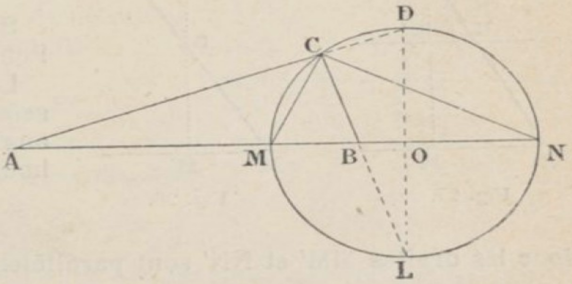

Fig. 26.

Sur la droite AB et sur son prolongement, déterminons les points M et N tels qu'on ait :

$$\frac{MA}{MB}=\frac{NA}{NB}=\frac{m}{n}$$

(G., n° 307.)

Ces deux points M et N appartiennent au lieu demandé.

Pour un autre point quelconque C du lieu, on a par hypothèse :

$$\frac{CA}{CB}=\frac{m}{n}$$

donc
$$\frac{CA}{CB}=\frac{MA}{MB}=\frac{NA}{NB}$$

Mais la bissectrice de l'angle ACB et celle de l'angle supplémentaire BCD donneraient, sur la base, deux points dont le rapport des distances aux points A et B égalerait $\frac{CA}{CB}$ ou $\frac{m}{n}$; donc les droites CM et CN sont elles-mêmes les bissectrices cherchées.

Les bissectrices des deux angles supplémentaires sont perpendiculaires l'une à l'autre ; donc l'angle MCN est droit, et le point C appartient à la circonférence décrite sur le diamètre MN.

On prouve ensuite que tout point de la circonférence appartient au lieu. (G., n° 307, 2°.)

62. Note. 1° En tenant compte des signes, on écrit :

$$\frac{MA}{MB}=-\frac{m}{n} \text{ et } \frac{NA}{NB}=\frac{m}{n}; \text{ d'où } \frac{NA}{NB}=-\frac{MA}{MB}.$$

Les quatre points A,B,M,N, forment une division harmonique. Dans la *Géométrie récente du triangle* (n^os^ 2262 et suivants) on écrit de préférence :

$$\frac{MB}{BN}=\frac{m}{n} \text{ et } \frac{MA}{AN}=-\frac{m}{n}.$$

Afin que le rapport qui correspond au point compris entre M et N soit positif.

2° Le lieu des points dont le rapport des distances à un point et à une droite est constant est une conique.

On obtient une ellipse lorsque $\frac{m}{n}$ est <1. (G., n° 846.)

— une parabole lorsque $\frac{m}{n}=1$. (G., n° 848.)

— une hyperbole pour $\frac{m}{n}>1$. (G., n° 850.)

1 ! "$/2" " "3 4 + " ?
0C 6* 64[[66:.) B B? B .*
_ .+4

! "#$%$ & ' ' "( ) " "8 #

1 ! "$/2" " "3 4 + " 8 4 & @ *) 0
7 ' 4 @ D 8 % + + % + F + . F +4
< H * ; 1 . :

1 ! "$/2" " "3 4 + " 8 4 & @ *) 0
7 ' 4 @ D 8 % + + % + F + . F +4
< H * ; 1 .*

?1 ! "$/2" " "3 4 + " 9+ = , *4 F 8 4 @ *)
/ + + . : 4 * + * ' # ' : +
; ? 1 *S'

1 ! "$/2" " "3 4 + " 9+ = , *4 F 8 4 @ *)
/ + + . : 4 * + * ' # ' : +
; ? 1 .*
: , ' * <<<

! "#$%$ & ' ' "( ) " "8[[# . ' 0C *
8[[# . '

1 ! "$/2" " "3 4 + " + +
4 6* . ' .*
3: , ' * <<<
>4 ; 1 /3(,0 0 J/ ,0 #/37%0 0% ,0 ,$&
. + .+ + \
/+.. 4[[E
E D * .+ + \ ] . + F + E D .+ + \ ] *
* * . +

1 ! "$/2" " "3 4 + "
0C 6* 64[[6* . ' 6 6V W@676:.) B .*

1 ! "$/2" " "3 4 + " 8 * . ' .*
A .+4

>4 ; 1 ,+ S 3 4 +

1 ! "$/2" " "3 4 + " &$ 6 B 6 6 66 6 .*
A8[[* . ' F C[* [. `

! "#$%$ & ' ' "( ) " "8) 4 = % + 0 *

B1 ! "$/2" " "3 4 + " 4)4 = + .*

! "#$%$ & ' ' "( ) " " + + : ' 4
% 4 / + + + +

1 ! "$/2" " "3 4 + " ($' :+ F 3 !
: ' 4 4 + + + + .*

1 ! "$/2" " "3 4 + " ($' :+ F 3 !
: ' 4 4 + + + + # ' : + ; 1 *S'

1 ! "$/2" " "3 4 + " ($' :+ F 3 !
: ' 4 4 + + + + + + % . J ' = .*

?1 ! "$/2" " "3 4 + " : ' 4 4
+ + + + # ' : + ; ???1 . :

! "#$%$ & ' ' "( ) " " +.. 0)

B 1 ! "$/2" " "3 4 + " 6. : + 6 + .*
A % MN * $ 4 # . +MN * ! :
+ Ma + $ 4+ 8 [+
3: , ' * <<<

B 1 ! "$/2" " "3 4 + " .+ .*
A ,P9 * + 4 [

B 1 ! "$/2" " "3 4 + " .+ .*
A @ + M \

B 1 ! "$/2" " "3 4 + " .+ ? .*
A [\ * [[

B 1 ! "$/2" " "3 4 + " B + +
+ + .*
A + + + + B .+4
3: , ' * <<<

BB1 ! "$/2" " "3 4 + " B + +
+ + 6 C .*

B 1 ! "$/2" " "3 4 + " B ? % ! + = 3
+.. .*
A B .+4
A 8 + + 4 +4

B 1 ! "$/2" " "3 4 + " B BB ? 7 7 .*
A 8 * + + B? .+4

B 1 ! "$/2" " "3 4 + " B BB ? 7 76 C .*
A 8 * + + B? .+4

B?1 ! "$/2" " "3 4 + " +: B * .*
A #+ : * +..

1 ! "$/2" " "3 4 + " $ * * .*
1

1 ! "$/2" " "3 4 + " $J@ 6J0! .*
A 9 * + [+ E : ;J0! 1
, 4 + * . : \ * + 4[[4 E D ' * :

1 ! "$/2" " "3 4 + " .+ + + 6 6
?? 6 ? 6 6 B6 6 .*
A,Pb ' * + + 4[[. S '

1 ! "$/2" " "3 4 + " /9 6 66?6 6 6 .*
A !3 0&% # c0@7&7/ ,$ &$%J/0 #0 %0-%0 $%9 c0 $%72J0 #0
,D$&%72J7% c0 8/0!2J0 %$/#7 0^ /3 3 7%73& #0 / c0@3/ 0 #0
,$ &3%73& #0 d%0-%0 #0J% c0/3&3 72J0 D

1 ! "$/2" " "3 4 + "
6 64 6* 60 6*60 6' 6 .*
A 4 \ * e [*D0 *

! "#$%$ & ' ' "( ) " " . + + * + @ +
&+ + = + # *

B1 ! "$/2" " "3 4 + " B . + + * +
+ + + . + 7 ++ & .*

1 ! "$/2" " "3 4 + " B %+: + * ! * . + .*

1 ! "$/2" " "3 4 + " B ??? 6 .*
A f. + > *+ @ + &+ +
$ * [+

1 ! "$/2" " "3 4 + " B '* . :6. . + . .
+ + * + + + .*
A , ' 77 777_

! "#$%$ & ' ' "( ) " " + 4 =
8) + +

?1 ! "$/2" " "3 4 + " 7' % + ;% + 1
+ 4 9 = 8) + + 7 9+ '+ * J ' = ; ?B 1 .*

1 ! "$/2" " "3 4 + " 7' % + ;% + 1
+ 4 9 = 8) + + 77 9+ '+ * J ' = ; ?B 1 .*

1 ! "$/2" " "3 4 + " : .*
A 8) + +

1 ! "$/2" " "3 4 + " : .*
A 8) + +

! "#$%$ & ' ' "( ) "

! "#$%$ & ' ' "( ) " "$ 8 7 * 9 4 +
= =

1 ! "$/2" " "3 4 + " $ 8 7 *
9 4 + = = .*
A .+4
3: , ' * <<< 7 . + f <<<

! "#$%$ & ' ' "( ) " "! + + + * # @ +
$ + = + % =

1 ! "$/2" " "3 4 + " ! + + + * # @ +
$ + = + % = .*
3: , ' * <<<

>4 ; 1

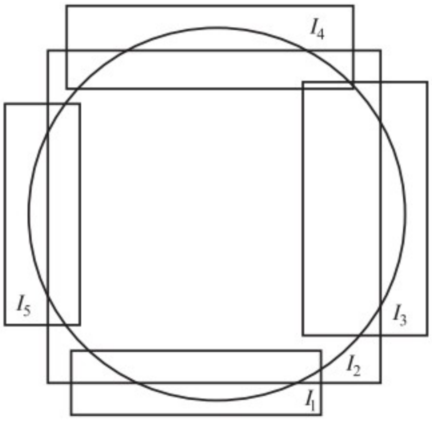
I_4
I_5
I_3
I_2
I_1

@ 4 % ' + * + . := '+ 7 F F7B

>4 ?; 1 B + * & * .+ `

$ + = . + = * G . F * 4
+ F +' . + = * 4 0 * + .+ /
H + +*= * ' . + = 4 + + /
$ * . * + + + * 0 * + .+
B . . 0 * + .+ + = ' F :
+ @ C+ . F !+ =D E + = ; . = ;0 1 % 1 + *
9 (;% 1 * * + 4 + 4

5

Metric and Normed Spaces

Analysis is primarily concerned with infinite processes, including limits. So far, we have primarily done this in the setting of Euclidean spaces $\mathbb{R}^n$. We are now ready to develop a theory of limits in sets M more general than $\mathbb{R}^n$. All we need is the concept of distance as enumerated for Euclidean spaces in Section 0.5. Most of the properties of Euclidean space carry over, but not all. For example, Cauchy's inequality (Property (**E3**) of Theorem 0.6) and the Heine–Borel theorem (Theorem 0.23) do not hold in all general settings.

5.1 Metric Spaces

A **metric space** is a nonempty set M with a **distance function** or **metric** d having the following properties for all $x, y, z \in M$:

(M0) $d(x, y) \geq 0$;
(M1) $d(x, y) = 0$ if and only if $x = y$;
(M2) $d(x, y) = d(y, x)$ (symmetry);
(M3) $d(x, y) \leq d(x, z) + d(z, y)$ (triangle inequality).

Property **(M0)** can actually be proven from the others as follows:

$$0 = d(x, x) \leq d(x, y) + d(y, x) = d(x, y) + d(x, y) = 2d(x, y).$$

Here are some examples of metric spaces.

Example 5.1 The familiar Euclidean Space: Let $M = \mathbb{R}^n$, with the usual Euclidean metric as given in Section 0.5. For $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ the metric is given by

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}.$$

Example 5.2 Discrete or Trivial Metric: Let M be any nonempty set and define

$$d(x,y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{if } x = y. \end{cases}$$

The metric conditions are easily seen to be true. If you consider the discrete metric on the set $\mathbb{R}$, the distance between any two distinct points is 1, which is very different from the usual metric on $\mathbb{R}$. A major use of the discrete metric is to obtain examples to illustrate specific concepts and to create counterexamples for conjectures. When considering a conjecture about metric spaces, it is a good idea to first test it on the discrete metric.

Example 5.3 Hamming Metric: Let $n \in \mathbb{N}$ be fixed, say $n = 7$, and let M be the set of all n-tuples of 0s and 1s. For $x, y \in M$, let $d(x, y)$ be the number of places in which x and y differ. For example, $d(0010110, 1010011) = 3$. This metric is used in the decoding of messages from error correcting codes.

Example 5.4 Let M consist of the natural numbers $\mathbb{N}$ along with the object ∞. Define the metric $d(x, y)$ as follows:

$$\begin{aligned} &d(x,y) = \left|\frac{1}{x} - \frac{1}{y}\right| \qquad \text{if} \quad x, y \in \mathbb{N}, \\ &d(x,\infty) = d(\infty,x) = \frac{1}{x} \quad \text{if} \quad x \in \mathbb{N}, \\ &d(\infty,\infty) = 0. \end{aligned}$$

This metric d is bounded on $\mathbb{N}$ by $d(x, y) < 1$. Adjoining the point ∞ to the space lets the metric attain its upper bound $d(1,\infty) = 1$.

Example 5.5 The function space $C[a,b]$ of continuous functions on $[a,b]$, with $a < b$, is a metric space under the distance function

$$d(f,g) = \int_a^b |f(x) - g(x)|\, dx.$$

The properties **(M2)** and **(M3)** are easy to prove. The proof of **(M1)** is left as Exercise 5.1.14. It follows from an important property of continuous functions: If a function $h \geq 0$ is continuous and $h(x_0) > 0$ for some $x_0 \in [a,b]$, then $\int_a^b h(x)\, dx > 0$.

Example 5.6 Fréchet metric: Recall that ω is the space of all real- or complex-valued sequences. The Fréchet[1] metric on ω is defined as follows. For $x = (x_1, x_2, \ldots)$ and $y = (y_1, y_2, \ldots)$, let

$$d(x, y) = \sum_{k=1}^{\infty} \frac{|x_k - y_k|}{2^k(1 + |x_k - y_k|)}.$$

The properties **(M0)**, **(M1)**, and **(M2)** are clearly satisfied. The triangle inequality **(M3)** follows from the inequality of Lemma 5.1. The details are left to Exercise 5.1.17. A generalization of this metric is given in Exercise 5.3.6.

Lemma 5.1 *For any* $a, b \in \mathbb{C}$, *we have* $\frac{|a+b|}{1+|a+b|} \leq \frac{|a|}{1+|a|} + \frac{|b|}{1+|b|}$.

Proof Consider the function $f(t) = t/(1+t)$ on $\mathbb{R}$. For $t > -1$, we have $f'(t) = 1/(1+t)^2 > 0$. Thus means that f is increasing on $t > -1$. Applying this function to the triangle inequality for $\mathbb{R}$, we have

$$\begin{aligned}\frac{|a+b|}{1+|a+b|} &= f(|a+b|) \leq f(|a|+|b|) \\ &= \frac{|a|+|b|}{1+|a|+|b|} \leq \frac{|a|}{1+|a|} + \frac{|b|}{1+|b|}.\end{aligned}$$ ∎

Theorem 5.2 Subspace metric: *Suppose M is a metric space and N is any nonempty subset. Let the distance between elements of N be defined just as in M. Then N is also a metric space.*

The proof is left as Exercise 5.1.11.

Exercises for Section 5.1

5.1.1 Show that Euclidean space $\mathbb{R}^n$ is a metric space with the Euclidean metric.

5.1.2 **Taxicab metric:** Let $M = \mathbb{R}^2$. Show that $d_1(x, y) = |x_1 - y_1| + |x_2 - y_2|$ is a metric on M when $x = (x_1, x_2)$ and $y = (y_1, y_2)$. Example of use: Whereas the Euclidean metric is the distance "as the crow flies," the fare of a taxicab in an urban grid of streets is based on this metric d_1.

---

[1] Maurice René Fréchet (1878–1973) was a French mathematician who introduced the concept of metric spaces. He made many contributions to functional analysis, topology, probability, and statistics.

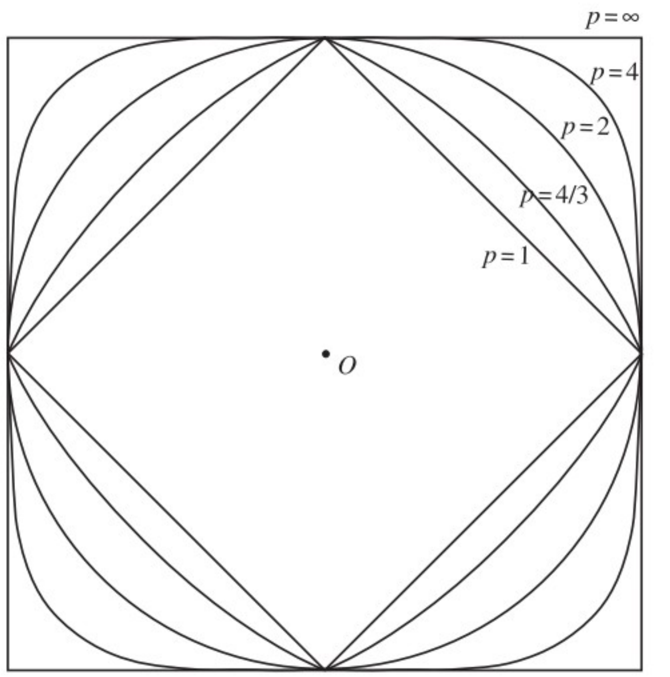

Figure 5.3 Boundaries of spheres $S(0,1) = \{x \in \mathbb{R}^2 \mid \|x\|_p < 1\}$ for $p = 1$, 4/3, 2, 4, and ∞.

! "#$%$ & ' ' "( ) " "9=. 4 . % =

B1 ! "$/2" " "3 4 + " ? 9=. 4 .6% = .*

! "#$%$ & ' ' "( ) " "7 . +

1 ! "$/2" " "3 4 + " B 6% 6 6! .*
A 8 % C + + + +

1 ! "$/2" " "3 4 + " 6 ? 67# + :+ %+ + .*
A %/g (/0 0 97 %h/7$ 3(/0 $,($ %9&

1 ! "$/2" " "3 4 + " ? B6%0 0 6% 6 6! .*
A 0 * * ! '+ * i i 4 + 4 * 8 % C +

?1 ! "$/2" " "3 4 + " 0 j ; B1 .*

1 ! "$/2" " "3 4 + " +6 + + + .*
>4 B 8 + 0 * + + &N 0 * + +

1 ! "$/2" " "3 4 + " +6 + + + .*

1 ! "$/2" " "3 4 + " 4 + +6 + + .*
A $ 0 . + * 0 + . .+ . . *+*

3: + : + + + @ 8 % C +

1 ! "$/2" " "3 4 + " 3: + : + + + * # @
8 % C + ; ? 1 .*
A ? .+4

1 ! "$/2" " "3 4 + " 3: + : + + + * # @
8 % C + ; ? 1 .*
A .+4

B1 ! "$/2" " "3 4 + " 3: + : + + + * # @
8 % C + ; ? 1 .*
A ? .+4
3: !+ # +

1 ! "$/2" " "3 4 + " 3: + : + + + * # @
8 % C + ; ? 1 .*
A .+4
A 2 + E

1 ! "$/2" " "3 4 + " : + : + 4 .*

1 ! "$/2" " "3 4 + " : + : + 4 .*

?1 ! "$/2" " "3 4 + " : + : + 4 .*

? 1 ! "$/2" " "3 4 + " : + : + 4 .*

? 1 ! "$/2" " "3 4 + " : + : + B4 .*
A B .+4

? 1 ! "$/2" " "3 4 + " = +4 + 9 * 4 +. ! '
$ 4 : + + * 8 = 7 .+ +: . 4 +4 (9 * : 4 ; 1 .*
A .+4
3: , ' * <<< 7 . + f <<<

? 1 ! "$/2" " "3 4 + " + +* * + ' 4 .*
A B .+4
A %/$%$#3 #0 ,$!J/$ 0 0!7$,0 &3%$(,0
3: , ' * <<<
A @ 8 % C +

! "#$%$ & ' ' "( ) " B

! "#$%$ & ' ' "( ) " B"$ * $.. + 2 +
+ F * 0*

? 1 ! "$/2" " "3 4 + " $ * $.. + 2 +
+ F * 0* .*
A B .+4

>4 B;? 1 !9$ %0/ $ 4 +

>4 ; ?1 !9$ %0/ ?
% + + + + * / + + = = % (*= :

! "#$%$ & ' ' "( ) " B"2 + + $ + +* 4
$.. +

?B1 ! "$/2" " "3 4 + " 2 + + + .+ +* 4
+.. + + ; 1 .*
A B? .+4

>4 ; 1 $%/7- &3%$%73&

>4 ; 1 ! +. $ 4 + kkk

? 1 ! "$/2" " "3 4 + " 2 + + + &
7 + + 0* + ; 1 .*
A .+4

>4 B $ 4 +

? 1 ! "$/2" " "3 4 + " % 8 = 3
H * : 4 ! .+ = ; B1 .*
3: , ' * <<<
A ? .+4
A = =

>4 ;B 1 = = 8 .

! +.
3 4 + 8 .

! "#$%$ & ' ' "( ) "

! "#$%$ & ' ' "( ) " "# 4 F * 4 + * / +: =
/ + 4 + =

? 1 ! "$/2" " "3 4 + " # 4 * 4 + * / +: =
/ + 4 + = U .*

! "#$%$ & ' ' "( ) " " + + +

??1 ! "$/2" " "3 4 + " 8 + ( ) H H *
;B 11 .*
A 9 + I +C +)
$ * & H +*
$: 0
$ 0** 4 m & (
8 9 9+ *= m H 9 : 4
0 * 4
% * # :U +)=
! 9 H+** 4

1 ! "$/2" " "3 4 + " . + + + + B
!+ : *4 J ' = ; ?? 1 .*

1 ! "$/2" " "3 4 + " H +* $ + ' +
+ 4 : + +.. + ; ? 1 .*

1 ! "$/2" " "3 4 + " H +* $ % +
J ' + $ 4 : + H $.. + !+ : *4 J ' = ; ?1 .*

1 ! "$/2" " "3 4 + " H +* % +C
. S ' 4 = &+: ; 1 .*

1 ! "$/2" " "3 4 + " H +* % +C
. S ' 4 = &+: ; 1 *S'

! "#$%$ & ' ' "( ) " "& % * $ + = + * 8 =
.+

B1 ! "$/2" " "3 4 + " B
(+ * 6( )6& % * 3 $ + = $ *8 = .*

! "#$%$ & ' ' "( ) " "/ + # = . + *
! + 8 = + * % . 4=

1 ! "$/2" " "3 4 + " / + # = .
+ * ! + 8 = + * % . 4= .*

! "#$%$ & ' ' "( ) " " .+ + 0 .= + * ,+ * +.
$ + =

1 ! "$/2" " "3 4 + "
+.+* 6( )6 .+ + 0 .=$ *,+ * +. $ + .*

! "#$%$ & ' ' "( ) " "% 0 & + 3.

1 ! "$/2" " "3 4 + " ! % 0 & +
3. !+ : *4 J ' = ; ?? 1 .*

>4 ;B 1 % . S . +

3.4 The projection operator

The real significance of traces lies in the fact that the physical content of quantum mechanics can always be expressed in terms of them. In particular, the expectation-value formula (3.13) can be expressed as a trace. For this purpose we introduce the *projection operator* $P(\psi)$ for the state ψ by means of the defining relation

$$P(\psi)\theta = \psi\int d\varsigma\, \psi^* \theta, \tag{3.25}$$

where θ is an arbitrary wave function. The projection operator derives its name from the fact that, if ψ is a member of a complete orthonormal set, then $P(\psi)$ picks out the contribution to the expansion of θ from the state ψ. By making use of the defining relation (3.25), we obtain for the (ji)th matrix element of $P(\psi)$:

$$\begin{aligned} P(\psi)_{ji} &= \int d\varsigma\, u_j^* P(\psi) u_i \\ &= \int d\varsigma\, u_j^* \psi \int d\varsigma\, \psi^* u_i \\ &= a_j a_i^*, \end{aligned} \tag{3.26}$$

where the coefficients a_i and a_j are defined by (3.6). By substituting this result in (3.13) and invoking the matrix product rule (3.11), we obtain the desired formula

$$\begin{aligned} <O> &= \sum_i \sum_j P(\psi)_{ji} O_{ij} \\ &= \sum_j (P(\psi) O)_{jj} \\ &= \mathrm{Tr}(P(\psi) O), \end{aligned} \tag{3.27}$$

which, we note, is independent of the representation.

Since it can be used to calculate expectation values, the projection operator provides a description of the physical state of the system which is completely equivalent to that provided by the wave function. For our purpose the description in terms of the projection operator is more suitable. The time-development of the system, as well as its instantaneous condition, can also be discussed in terms of the projection operator; that is to say, it is possible to derive an equation of motion for $P(\psi)$ which may be used instead of the equation of motion (3.3) for the wave function ψ (the time-dependent Schrödinger equation). It is sufficient for this purpose to suppose that the arbitrary wave function θ in (3.25) is independent of time. Then, by differentiating (3.25) with respect to time t and making use of (3.3), we obtain

$$\begin{aligned} i\hbar\, d\{P(\psi)\theta\}/dt &= i\hbar\, (d\psi/dt)\int d\varsigma\, \psi^*\theta + \psi\int d\varsigma\, i\hbar\, (d\psi^*/dt)\theta \\ &= H\psi\int d\varsigma\, \psi^*\theta - \psi\int d\varsigma\, (H\psi)^*\theta. \end{aligned} \tag{3.28}$$

Downloaded from https://www.cambridge.org/core. University of Birmingham, on 12 Aug 2017 at 19:31:32, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139167994.004

From the defining relation (3.25), it is evident that the first term on the second line of (3.28) is $HP(\psi)\theta$. Moreover, since energy is a real dynamical variable and therefore the corresponding operator (the Hamiltonian H) is Hermitian, we see with the aid of (3.15) that the second term on the second line of (3.28) is

$$\begin{aligned} \psi\int \mathrm{d}\varsigma\,(H\psi)^*\theta &= \psi\left(\int \mathrm{d}\varsigma\,\theta^* H\psi\right)^* \\ &= \psi\int \mathrm{d}\varsigma\,\psi^* H\theta \\ &= P(\psi)H\theta. \end{aligned} \tag{3.29}$$

Hence (3.28) becomes

$$i\hbar\,\mathrm{d}\{P(\psi)\theta\}/\mathrm{d}t = HP(\psi)\theta - P(\psi)H\theta. \tag{3.30}$$

Finally, since θ is arbitrary, it can be cancelled out of (3.30) to yield the equation of motion of the projection operator $P(\psi)$:

$$\begin{aligned} i\hbar\,\mathrm{d}\{P(\psi)\}/\mathrm{d}t &= HP(\psi) - P(\psi)H \\ &= [H, P(\psi)] \end{aligned} \tag{3.31}$$

where we have introduced the conventional bracket notation for the commutator of two operators.

3.5 The density operator

We are now in a position to introduce the density operator and determine its equation of motion. The expectation-value formula (3.2) is the basic rule for the physical interpretation of the formalism of quantum mechanics. However, in order to use it we must know the state of the system precisely so that the wave function ψ is uniquely determined (apart from a trivial phase factor). Such precise knowledge is seldom at our disposal. Let us consider, as an example, a system containing M spin-free particles. Then a precise determination of the state of the system involves the measurement of $3M$ quantities (since the system has $3M$ degrees of freedom); the quantities to be measured might, for example, be the particle momenta. For a macroscopic system comprising, for example, a 1 mm^3 cube of solid material, $M \sim 10^{20}$ typically, and the task must be abandoned as hopeless. Even in small volumes of very dilute gases, the number of entities is usually so large that one cannot expect to know the state of the system precisely. The most that can really be hoped for is statistical information. We can use whatever limited knowledge is available about the state of the system to determine the probability p_n that the system is in the state ψ_n, where n labels the different possible states in which the system might actually be. Setting $p_n = 0$ for all values of n except one, for which $p_n = 1$, gives us the exceptional case in which

Downloaded from https://www.cambridge.org/core. University of Birmingham, on 12 Aug 2017 at 19:31:32, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139167994.004

>4 ; ?1 3' + . + = =

An example of the full expansion of (4.64) in the case of a four-level system is given by Bloembergen *et al* (1978).

The generalisation of the formulae (4.58), (4.63) and (4.64) to arbitrary order will become clear after we have introduced the overall permutation-symmetry property in the next section.

To derive these explicit formulae for the susceptibility tensors we have used the density-matrix approach developed in Chapter 3. We should point out, however, that this is not the only approach that can be taken. Similar results are obtained using a time-dependent perturbation analysis (Ducuing, 1969).

4.3 Overall permutation symmetry

The formulae for the susceptibility tensors derived in the previous section may be thrown into a form which reveals an important permutation-symmetry property of the tensor components. The *intrinsic* permutation-symmetry property, first introduced in §2.1 and used extensively in the previous sections of this chapter (as implied by the operator $\mathbf{S}$), indicates that $\chi^{(n)}_{\mu\alpha_1\cdots\alpha_n}(-\omega_\sigma;\omega_1,\ldots,\omega_n)$ is invariant under all $n!$ permutations of the n pairs $(\alpha_1,\omega_1),(\alpha_2,\omega_2),\ldots,(\alpha_n,\omega_n)$. *Overall* permutation symmetry, which we introduce here, is a more general property, at least in the sense that it encompasses intrinsic permutation symmetry. However, it is important to appreciate that intrinsic permutation symmetry is a completely rigorous property which arises from the principles of causality and reality, and which applies universally. The overall permutation-symmetry property, on the other hand, is an approximation which applies only under particular conditions; namely, when all of the optical frequencies involved in the susceptibility formulae (*i.e.*, the primary frequencies $\omega_\sigma,\omega_1,\ldots,\omega_n$ and the combinations of these which occur in the denominators) are far removed from molecular transition frequencies – in other words, when the medium can be assumed to be loss-free at all the relevant optical frequencies. The reasons for this restriction become clearer when, in §4.5, we introduce transition damping factors into the susceptibility formulae. The physical consequences of overall permutation symmetry are discussed in §5.1.

Here we shall simply demonstrate how overall permutation symmetry arises, and introduce a compact notation for its description. The usefulness of this notation is that it allows us to write down readily a general formula for the nth-order susceptibility tensor.

4.3.1 Compact formulae for nonresonant susceptibilities

We begin by considering the previous formula for the first-order

Downloaded from https://www.cambridge.org/core. University of Birmingham, on 12 Aug 2017 at 19:36:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139167994.005

susceptibility tensor, namely:

$$\chi^{(1)}_{\mu\alpha}(-\omega_\sigma;\omega) = \frac{N}{\varepsilon_0}\frac{e^2}{\hbar}\sum_{ab}\rho_0(a)\left[\frac{r^{\mu}_{ab}r^{\alpha}_{ba}}{\Omega_{ba}-\omega}+\frac{r^{\alpha}_{ab}r^{\mu}_{ba}}{\Omega_{ba}+\omega}\right]. \tag{4.65}$$

It will be seen that the second term in square brackets may be derived from the first by interchanging the coordinate labels μ and α and replacing $-\omega$ by ω. Consequently, if we interchange μ and α and interchange $-\omega_\sigma$ and ω in $\chi^{(1)}_{\mu\alpha}(-\omega_\sigma;\omega)$, and remembering that for the first-order susceptibility ω_σ is defined as $\omega_\sigma = \omega$, it follows that the only result is to interchange the order in which the two terms appear in the square brackets in (4.65) and their sum is unaltered. Thus we have

$$\begin{aligned}\chi^{(1)}_{\alpha\mu}(\omega;-\omega_\sigma) &= \frac{N}{\varepsilon_0}\frac{e^2}{\hbar}\sum_{ab}\rho_0(a)\left[\frac{r^{\alpha}_{ab}r^{\mu}_{ba}}{\Omega_{ba}+\omega}+\frac{r^{\mu}_{ab}r^{\alpha}_{ba}}{\Omega_{ba}-\omega}\right]\\ &= \chi^{(1)}_{\mu\alpha}(-\omega_\sigma;\omega).\end{aligned} \tag{4.66}$$

This is the overall permutation-symmetry property of the first-order susceptibility tensor. Thus (4.65) may be rewritten in a more compact form:

$$\chi^{(1)}_{\mu\alpha}(-\omega_\sigma;\omega) = \frac{N}{\varepsilon_0}\frac{e^2}{\hbar}\mathbf{S}_{\mathrm{T}}\sum_{ab}\rho_0(a)\left[\frac{r^{\mu}_{ab}r^{\alpha}_{ba}}{\Omega_{ba}-\omega}\right], \tag{4.67}$$

where the 'total symmetrisation operator' $\mathbf{S}_{\mathrm{T}}$ indicates that the expression which follows it is to be summed over all possible permutations of the pairs $(\mu,-\omega_\sigma)$ and (α,ω). There are in fact two possible permutations, which are: (i) the identity permutation $(\mu,-\omega_\sigma)\rightarrow(\mu,-\omega_\sigma)$, which merely reproduces the term stated explicitly in (4.67), *i.e.*, the first term in square brackets in (4.65); (ii) the interchange $(\mu,-\omega_\sigma)\leftrightarrow(\alpha,\omega)$, which gives the second term in square brackets in (4.65). Since $\mathbf{S}_{\mathrm{T}}$ indicates that a summation is to be made over all permutations of the pairs $(\mu,-\omega_\sigma)$ and (α,ω), it is clear that $\chi^{(1)}_{\mu\alpha}(-\omega_\sigma;\omega)$ is invariant under such permutations. This is an expression of the overall permutation-symmetry property in first order, and in this form the property generalises to all orders.

We now turn to a consideration of the formula (4.63) for the second-order susceptibility tensor. We again associate the frequency $-\omega_\sigma$ with the coordinate suffix μ. It can be seen by inspection of (4.63) that the second term in square brackets may be derived from the first by interchanging μ and α and replacing $-\omega_\sigma(=-\omega_1-\omega_2)$ by ω_1, *i.e.*, by permuting the pairs $(\mu,-\omega_\sigma)$, (α,ω_1), (β,ω_2) as follows:

$$(\mu,-\omega_\sigma),(\alpha,\omega_1),(\beta,\omega_2)\rightarrow(\alpha,\omega_1),(\mu,-\omega_\sigma),(\beta,\omega_2). \tag{4.68}$$

Downloaded from https://www.cambridge.org/core. University of Birmingham, on 12 Aug 2017 at 19:36:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139167994.005

Similarly, the last term in square brackets may be derived from the first by permuting the pairs as follows:

$$(\mu, -\omega_\sigma), (\alpha, \omega_1), (\beta, \omega_2) \rightarrow (\alpha, \omega_1), (\beta, \omega_2), (\mu, -\omega_\sigma). \tag{4.69}$$

Finally, the first term in square brackets may of course be derived from itself by applying the identity permutation

$$(\mu, -\omega_\sigma), (\alpha, \omega_1), (\beta, \omega_2) \rightarrow (\mu, -\omega_\sigma), (\alpha, \omega_1), (\beta, \omega_2). \tag{4.70}$$

Altogether there are six possible permutations of the three pairs $(\mu, -\omega_\sigma)$, (α, ω_1) and (β, ω_2). Equations (4.68), (4.69) and (4.70) specify the three permutations in which $(\mu, -\omega_\sigma)$ occupies the second, third and first place in the sequence while the order of the pairs (α, ω_1) and (β, ω_2) is left undisturbed. The other three permutations are derived from these by interchanging the pairs (α, ω_1) and (β, ω_2). Now the formula (4.63) already involves the symmetrising operation $\mathbf{S}$ which implies that the expression following it is to be summed over the two possible permutations of the pairs (α, ω_1) and (β, ω_2). Hence, when this symmetrising operation is carried out, there are six terms in the square brackets and they may all be derived from the first term by the application of the six permutations of the pairs $(\mu, -\omega_\sigma)$, (α, ω_1), (β, ω_2). Thus (4.63) may be written in the more compact form:

$$\chi^{(2)}_{\mu\alpha\beta}(-\omega_\sigma; \omega_1, \omega_2) = \frac{N e^3}{\varepsilon_0 2\hbar^2} \mathbf{S}_{\mathrm{T}} \sum_{abc} \rho_0(a) \left[\frac{r^\mu_{ab}\, r^\alpha_{bc}\, r^\beta_{ca}}{(\Omega_{ba} - \omega_1 - \omega_2)(\Omega_{ca} - \omega_2)} \right], \tag{4.71}$$

where the total symmetrisation operation $\mathbf{S}_{\mathrm{T}}$ indicates that the expression which follows it is to be summed over all six permutations of the pairs $(\mu, -\omega_\sigma)$, (α, ω_1), (β, ω_2). Since $\mathbf{S}_{\mathrm{T}}$ involves a summation over all possible permutations, it is clear that $\chi^{(2)}_{\mu\alpha\beta}(-\omega_\sigma; \omega_1, \omega_2)$ is invariant under any of them. This is the overall permutation-symmetry property in second order.

The presence of the symmetrising operation $\mathbf{S}$ in (4.63) already implies that $\chi^{(2)}_{\mu\alpha\beta}(-\omega_\sigma; \omega_1, \omega_2)$ has intrinsic permutation symmetry. The new feature introduced by the presence of the total symmetrisation operation $\mathbf{S}_{\mathrm{T}}$ in (4.71) is that the invariance is preserved under permutations involving the extra pair $(\mu, -\omega_\sigma)$. As a particular example of overall permutation symmetry in second order, we consider the permutation (4.68); thus we have

$$\chi^{(2)}_{\mu\alpha\beta}(-\omega_\sigma; \omega_1, \omega_2) = \chi^{(2)}_{\alpha\mu\beta}(\omega_1; -\omega_\sigma, \omega_2). \tag{4.72}$$

We may write down six relations of this type, one of which is an identity.

Downloaded from https://www.cambridge.org/core. University of Birmingham, on 12 Aug 2017 at 19:36:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139167994.005

The above considerations are easily extended to the formula (4.64) for the third-order susceptibility tensor $\chi^{(3)}_{\mu\alpha\beta\gamma}(-\omega_\sigma;\omega_1,\omega_2,\omega_3)$, for which ω_σ is defined to be $\omega_\sigma = \omega_1+\omega_2+\omega_2$. The four terms in the square brackets in (4.64) may all be derived from the first by application of the four permutations of the pairs $(\mu,-\omega_\sigma)$, (α,ω_1), (β,ω_2) and (γ,ω_3) in which $(\mu,-\omega_\sigma)$ appears in the first, second, third and fourth place in the sequence while the pairs (α,ω_1), (β,ω_2), and (γ,ω_3) appear in natural order. The symmetrisation operation $\mathbf{S}$ permutes the pairs (α,ω_1), (β,ω_2), and (γ,ω_3) in all six possible ways and sums the result. Hence (4.64) may be written in the compact form:

$$\chi^{(3)}_{\mu\alpha\beta\gamma}(-\omega_\sigma;\omega_1,\omega_2,\omega_3) = \frac{N}{\varepsilon_0}\frac{e^4}{3!\hbar^3}\mathbf{S}_\mathrm{T} \times \sum_{abcd}\rho_0(a)\left[\frac{r^\mu_{ab}\,r^\alpha_{bc}\,r^\beta_{cd}\,r^\gamma_{da}}{(\Omega_{ba}-\omega_1-\omega_2-\omega_3)(\Omega_{ca}-\omega_2-\omega_3)(\Omega_{da}-\omega_3)}\right], \quad (4.73)$$

where the total symmetrisation operation $\mathbf{S}_\mathrm{T}$ indicates that the expression which follows it is to be summed over all 24 permutations of the pairs $(\mu,-\omega_\sigma)$, (α,ω_1), (β,ω_2) and (γ,ω_3). Clearly, $\chi^{(3)}_{\mu\alpha\beta\gamma}(-\omega_\sigma;\omega_1,\omega_2,\omega_3)$ is invariant under all such permutations.

The generalisation of the formulae (4.67), (4.71) and (4.73) to arbitrary order is now obvious by inspection. In making this generalisation, the frequency $-\omega_\sigma$ is defined as previously, *i.e.*,

$$-\omega_\sigma = -(\omega_1+\omega_2+\cdots+\omega_n), \quad (4.74)$$

and is always associated with the cartesian coordinate μ. (We note that the pair $(\mu,-\omega_\sigma)$, and not (μ,ω_σ), is involved in overall permutation symmetry; hence the convention of writing $\chi^{(n)}_{\mu\alpha_1\cdots\alpha_n}(-\omega_\sigma;\omega_1,\dots,\omega_n)$ rather than $\chi^{(n)}_{\mu\alpha_1\cdots\alpha_n}(\omega_\sigma;\omega_1,\dots,\omega_n)$.) We thus obtain the general (but rather dense) formula:

$$\chi^{(n)}_{\mu\alpha_1\cdots\alpha_n}(-\omega_\sigma;\omega_1,\dots,\omega_n) = \frac{N}{\varepsilon_0}\frac{e^{n+1}}{n!\hbar^n}\mathbf{S}_\mathrm{T}\sum_{ab_1\cdots b_n}\rho_0(a) \times\left[\frac{r^\mu_{ab_1}\,r^{\alpha_1}_{b_1b_2}\cdots r^{\alpha_{n-1}}_{b_{n-1}b_n}\,r^{\alpha_n}_{b_na}}{(\Omega_{b_1a}-\omega_1-\cdots-\omega_n)(\Omega_{b_2a}-\omega_2-\cdots-\omega_n)\cdots(\Omega_{b_na}-\omega_n)}\right]. \quad (4.75)$$

In (4.75) $\mathbf{S}_\mathrm{T}$ is the total symmetrisation operation in *n*th order; it implies that the expression which follows it is to be summed over all $(n+1)!$ permutations of the pairs $(\mu,\omega_\sigma),(\alpha_1,\omega_1),\dots,(\alpha_n,\omega_n)$. Clearly $\boldsymbol{\chi}^{(n)}$ is invariant under all such permutations. This is the most general expression of the overall permutation-symmetry property.

Downloaded from https://www.cambridge.org/core. University of Birmingham, on 12 Aug 2017 at 19:36:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139167994.005

It will be noticed that the definition of $\mathbf{S}_\mathrm{T}$, like that of the intrinsic-permutation-symmetry operator $\mathbf{S}$ in (4.28), is different in different orders n and strictly, therefore, should carry a superscript n. However, the order will always be clear from the context and, as for $\mathbf{S}$, there is no need to complicate the notation in this way.

Lastly, it is often useful to simplify the appearance of the final formulae (4.67), (4.71), (4.73) and (4.75) by writing them in the scalar form defined by (2.58); we state below the resulting expression in nth order:

$$\chi^{(n)}(-\omega_\sigma;\omega_1,\dots,\omega_n) = \frac{N}{\varepsilon_0}\frac{e^{n+1}}{n!\hbar^n}\mathbf{S}_\mathrm{T}\sum_{ab_1\cdots b_n}\rho_0(a) \times\left[\frac{\mathbf{e}_\sigma{}^*\cdot\mathbf{r}_{ab_1}\,\mathbf{e}_1\cdot\mathbf{r}_{b_1b_2}\cdots\mathbf{e}_n\cdot\mathbf{r}_{b_na}}{(\Omega_{b_1a}-\omega_1-\cdots-\omega_n)(\Omega_{b_2a}-\omega_2-\cdots-\omega_n)\cdots(\Omega_{b_na}-\omega_n)}\right], \tag{4.76}$$

noting, once again, that the polarisation vector $\mathbf{e}_i{}^*$ is associated with the negative frequency $-\omega_i$. When expressed in this form, the total symmetrisation operation $\mathbf{S}_\mathrm{T}$ requires that the expression following it is to be summed over all permutations of the pairs $(\mathbf{e}_\sigma{}^*, -\omega_\sigma)$, $(\mathbf{e}_1, \omega_1), \dots, (\mathbf{e}_n, \omega_n)$.

The full expansion of (4.75) or (4.76), when the terms generated by $\mathbf{S}_\mathrm{T}$ are written out explicitly, contains $(n+1)!$ terms within the square brackets, although if some of the pairs $(\mathbf{e}_i, \omega_i)$ are in fact equal, then there will be fewer distinct terms. This is of no consequence, however, since the operation $\mathbf{S}_\mathrm{T}$ specifies that *all* the permutations are to be summed, regardless of whether or not they are distinct.

4.3.2 Third-harmonic generation

These points are best clarified by reference to an example, which will also demonstrate the application of the general formulae to a particular nonlinear process. The example considered here is third-harmonic generation (others occur in later sections). Third-harmonic generation can occur when an incident field at the frequency ω induces a polarisation at the frequency 3ω; the process is governed by the third-order susceptibility $\boldsymbol{\chi}^{(3)}(-3\omega;\omega,\omega,\omega)$. (The nonlinear polarisation so induced may generate a travelling wave at the third-harmonic frequency, as described in Chapter 7.) Because three of the frequency arguments of this susceptibility are identical, the 24 terms generated by $\mathbf{S}_\mathrm{T}$ in the formula (4.73) reduce to four distinct terms, each of which occurs six times. Thus, using the scalar form (4.76), we have

Downloaded from https://www.cambridge.org/core. University of Birmingham, on 12 Aug 2017 at 19:36:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139167994.005

$$\chi^{(3)}(-3\omega;\omega,\omega,\omega) = \frac{N}{\varepsilon_0}\frac{e^4}{3!\hbar^3}\sum_{abcd}\rho_0(a)$$
$$\times 6\left[\frac{\mathbf{e}_{3\omega}{}^*\cdot\mathbf{r}_{ab}\,\mathbf{e}_{\omega}\cdot\mathbf{r}_{bc}\,\mathbf{e}_{\omega}\cdot\mathbf{r}_{cd}\,\mathbf{e}_{\omega}\cdot\mathbf{r}_{da}}{(\Omega_{ba}-3\omega)(\Omega_{ca}-2\omega)(\Omega_{da}-\omega)}+\frac{\mathbf{e}_{\omega}\cdot\mathbf{r}_{ab}\,\mathbf{e}_{3\omega}{}^*\cdot\mathbf{r}_{bc}\,\mathbf{e}_{\omega}\cdot\mathbf{r}_{cd}\,\mathbf{e}_{\omega}\cdot\mathbf{r}_{da}}{(\Omega_{ba}+\omega)(\Omega_{ca}-2\omega)(\Omega_{da}-\omega)}\right.$$
$$\left.+\frac{\mathbf{e}_{\omega}\cdot\mathbf{r}_{ab}\,\mathbf{e}_{\omega}\cdot\mathbf{r}_{bc}\,\mathbf{e}_{3\omega}{}^*\cdot\mathbf{r}_{cd}\,\mathbf{e}_{\omega}\cdot\mathbf{r}_{da}}{(\Omega_{ba}+\omega)(\Omega_{ca}+2\omega)(\Omega_{da}-\omega)}+\frac{\mathbf{e}_{\omega}\cdot\mathbf{r}_{ab}\,\mathbf{e}_{\omega}\cdot\mathbf{r}_{bc}\,\mathbf{e}_{\omega}\cdot\mathbf{r}_{cd}\,\mathbf{e}_{3\omega}{}^*\cdot\mathbf{r}_{da}}{(\Omega_{ba}+\omega)(\Omega_{ca}+2\omega)(\Omega_{da}+3\omega)}\right]. \tag{4.77}$$

The singularities that occur when any of the frequency denominators in (4.77) approach zero are considered in detail in §4.5.

4.3.3 Virtual transitions and Feynman diagrams

Diagrammatic techniques can be used to keep track of the various terms in (4.75) and (4.76) which are generated by $\mathbf{S}_{\mathrm{T}}$ (Ward, 1965), and these also provide a pictorial interpretation of the matter-field interaction in terms of photon creation and annihilation (Loudon, 1983). Since we use the classical description of fields throughout this book, we should regard the polarisation vectors $\mathbf{e}_j$ and $\mathbf{e}_j{}^*$ which appear in the susceptibility formulae as commuting annihilation and creation operators, respectively. (It will be recalled that in the definition (2.58), the vector $\mathbf{e}_i{}^*$ is to be associated with a negative frequency $-\omega_i$ which appears as an argument of $\chi^{(n)}(-\omega_\sigma;\omega_1,\ldots,\omega_n)$.) To take an example, the four terms in the formula (4.77) for the third-harmonic susceptibility correspond, in order, to the four Feynman diagrams drawn in Fig. 4.1. The first diagram (a) represents the process in which a molecule, initially in a state labelled a, emits a third-harmonic photon 3ω whilst making a virtual transition to a state labelled b. By absorbing a fundamental photon ω the molecule then makes a virtual transition from b to c. The absorption of two further ω photons takes the molecule into d and finally back to the initial state a. The summation in (4.77) signifies that each of the labels a, b, c, d ranges over all of the molecular energy eigenstates, and that all combinations of the pathways $a\rightarrow b\rightarrow c\rightarrow d\rightarrow a$ contribute to the term depicted in Fig. 4.1(a). The process involves virtual transitions; the molecule undergoes no real net change of energy and is restored to its initial state when the process is complete. Energy conservation need not (and in general will not) hold at each virtual transition in the sequence depicted in the Feynman diagram. However, energy is conserved overall since $\hbar(\omega+\omega+\omega)=\hbar 3\omega$. These remarks hold for all four diagrams drawn in Fig. 4.1 in which the virtual transitions occur in various sequences corresponding to the four terms in (4.77). The factor of 6 arises from the indistinguishable permutations of the ω photons in each diagram.

Downloaded from https://www.cambridge.org/core. University of Birmingham, on 12 Aug 2017 at 19:36:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139167994.005

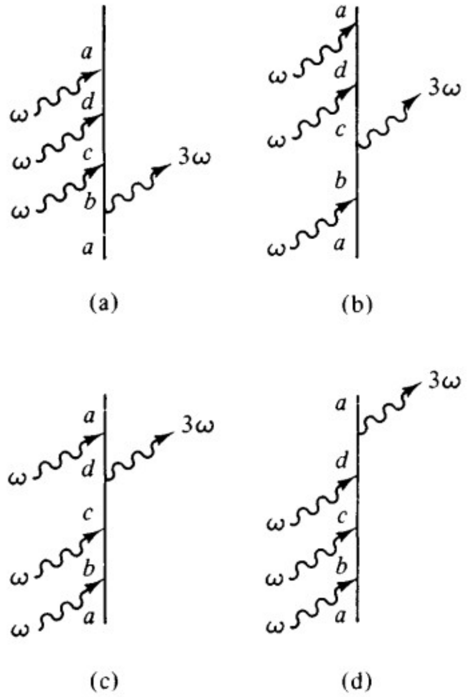

Fig. 4.1 Feynman diagrams illustrating the process of third-harmonic generation (the optical frequencies involved are assumed to be far removed from resonances). Photons (indicated by wavy lines) incident from the left represent absorption, and emission processes are represented by photons leaving to the right. The sequence of virtual molecular transitions involved in the whole process is indicated by the upward progression through states labelled a , b ,... on the vertical line.

Since energy is generally not conserved during a virtual transition (such as $a \rightarrow b$ in Fig. 4.1(a), for example), the uncertainty principle dictates that the transition is possible only if the molecule remains in the state b for a time shorter than the inverse of the detuning (which is $(\Omega_{ba} - 3\omega)^{-1}$ in this case). Virtual transitions therefore occur on an ultrafast time scale (see §6.6) and may be assumed to be instantaneous for many practical purposes.

4.4 The passage from microscopic to macroscopic formulae

In §4.1 we derived fundamental formulae for the susceptibility tensors in terms of electric-dipole operators and their matrix elements (equivalent formulae in terms of current operators and their matrix elements are derived in §4.6). The general expressions (4.28) and (4.131) for $\boldsymbol{\chi}^{(n)}$ in

Downloaded from https://www.cambridge.org/core. University of Birmingham, on 12 Aug 2017 at 19:36:50, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9781139167994.005

>4 ; 1 = = . .
A +.f * <<<

! "#$%$ & ' ' "( ) "

! "#$%$ & ' ' "( ) " "@ *+ + + + !
* + 8 =

?1 ! "$/2" " "3 4 + " ? ;@ *+ + + + 1
+ S+= + @ *+ + + + 6 * + + + 7 * + ; 1 .*

1 ! "$/2" " "3 4 + " ? ;@ *+ + + + 1
+ S+= + @ *+ + + + 6 @ + * 8 +. + 7 * + ; 1 .*

1 ! "$/2" " "3 4 + " ? ;77% Q001 + S+= + $ 4 : +
@ *+ + + + + .*

1 ! "$/2" " "3 4 + " ? ;77% Q001 + S+= + $ 4 : +
@ *+ + + + Q00 $7& + * $*'+ * + .*

1 ! "$/2" " "3 4 + " ? ;77% Q001 + S+= +
@ *+ + + + ! * + 8 = + ; 1 .*

1 ! "$/2" " "3 4 + " ? 7 * # +
!+ = + * 0 4 4 $.. + (4 H = ; 1 .*

B1 ! "$/2" " "3 4 + " ? 7 * * + +
= + * 4 4 +.. + : 4 H = ; 1 .*

! "#$%$ & ' ' "( ) " "8 $ 4 : + =

1 ! "$/2" " "3 4 + " ! # + 8/$ 7%O 8$J80
%903/70 $&# 803 0%/7! $,80(/$.*
A =

1 ! "$/2" " "3 4 + " 8 $ 4 : + =
!+ : *4 J ' = ; 1 .*
3: , ' * <<<

>4 ; ?1 B / + ' = + * .+

>4 ; 1 8 +
>4 ; 1 @ 4

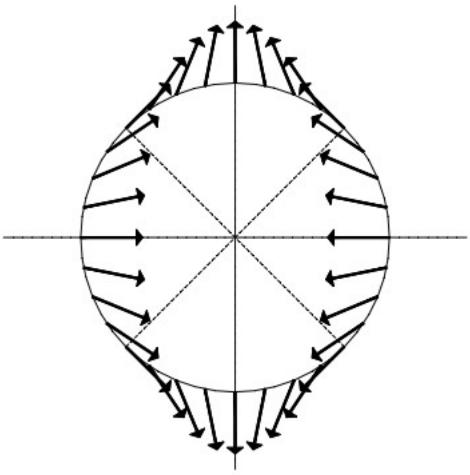

Figure 6.1 *Spacetime gradients.* The contours of the scalar field $\phi = \langle x\gamma_0 x\gamma_0\rangle$ define circles in spacetime. But the direction of the vector derivative is only in the outward normal direction along the 0 axis. Along the 1 axis the gradient points inwards, which reflects the opposite signature. Around the circle the gradient interpolates between these two extremes. At points where x is null the gradient vector is tangential to the circle.

>4 ; B 1 8 =

S ' 4 =
% + + . S ' 4 = * + + 4 .+ * 4 +* +
+ + @ '+ + :S +
=F +) 4 += '+ :S * + 4
= (= + . S ' 4 = + S = * + 4 *
. + . 4 +. * = @ C+ . F +
3. 8, 4 +. + 4 +4 + : + . S ' . +
* + .+
%) = * + . S ' 4 = + . .+ + . * +

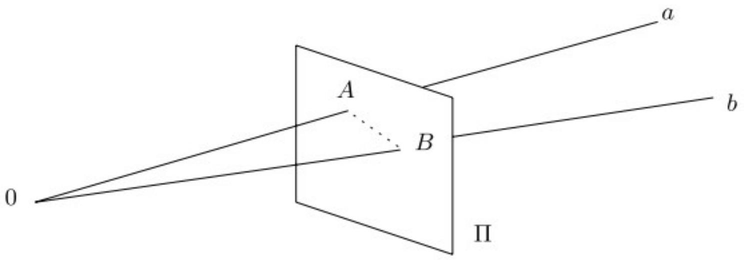

@ 4 S ' 4 = . S ' . + + .
* := ' + .+ * 4 % . + n *
4

' + .+ * 4 @ C+ . F . . S '
. + + . * + ' * + .+ ; G4 1 %
+4 * ' . + F + : + + * o+ . +
. % . + . + * : 4 %) =
. + . S ' 4 = + S + * % S . F
C+ . F : @ 4 S + 4 +* F + *
S + + = : * * + 4 : + + = ' + C . * ; +
8 + + D 4 + '+ * 4 C + 4 : +1 %
* 4 F + + . + 4 + + .
% +* + = * * ' + * + =F + * 4
+ 4 : + . += * := . * 3. +
+ + * S * * . * F . S ' 4 =
+' + . + . * % + . +
. + +' + 4 * + F : + 4 + 4 : + : +
E +* + F + = * 4 = % ' F
+ * + :

% . S '
% . . + + + * + % d0 * + D
* +: 4 + . + + : (
+ * + :+) . + + 8 + =F + .
+ E + (+ 4 : + + = + C . + .
% G 4 F . * := + 4 : + + = . + :
U % * . + G =F : . + +
* . S ' + + % : . :
. . + ' * + .+ 7 +=
p

>4 ; B 1

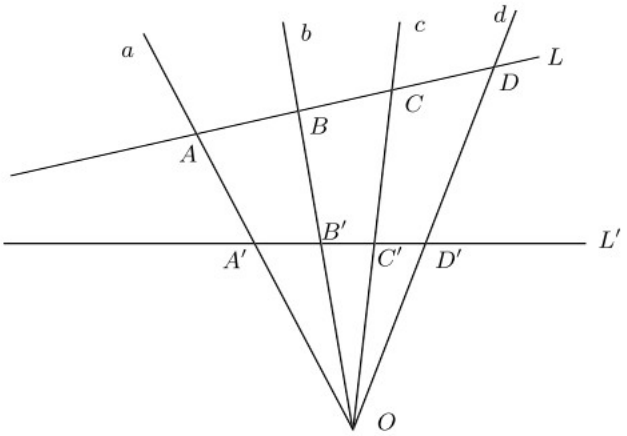
a
b
c
d
L
A
B
C
D
B′
A′
C′
D′
L′
O

Figure 10.2 *The cross ratio.* Points on the lines L and L' represent two different projective views of the same vectors in space. The cross ratio of the four points is the same on both lines.

3: !+.f * <<<

>4 B; 1

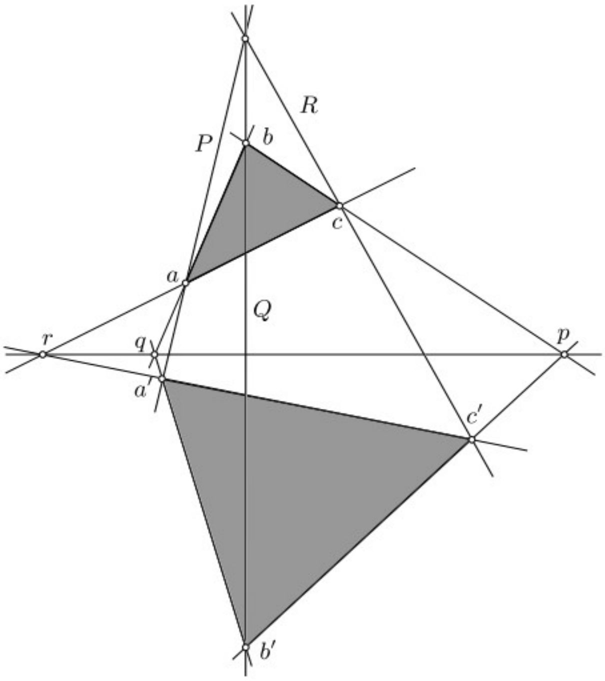
R
P
b
c
a
Q
r
q
p
a′
c′
b′

@ 4 # + 4 D % F 2F / + + . + *

= . .F EF + % + 4 + . S ' =
+ *

>4 B ; B1

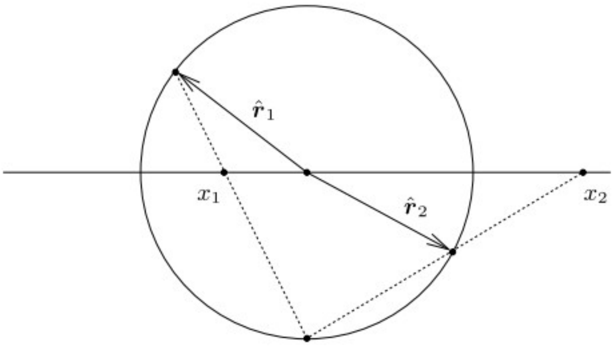
$\hat{\boldsymbol{r}}_1$
x_1
$\hat{\boldsymbol{r}}_2$
x_2

Figure 10.5 *A stereographic projection.* The line is mapped into the unit circle, so the points on the line x_1 and x_2 are mapped to the unit vectors $\hat{\boldsymbol{r}}_1$ and $\hat{\boldsymbol{r}}_2$. The origin and infinity are mapped to opposite points on the circle.

>4 B ; 1

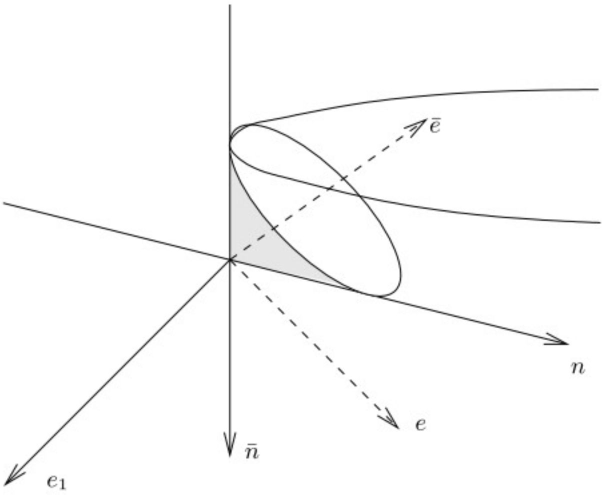
$\bar{e}$
n
e
$\bar{n}$
e_1

@ 4 % + * + + .
* := ' * % + F + *
+ . + ' .

>4 ; B1

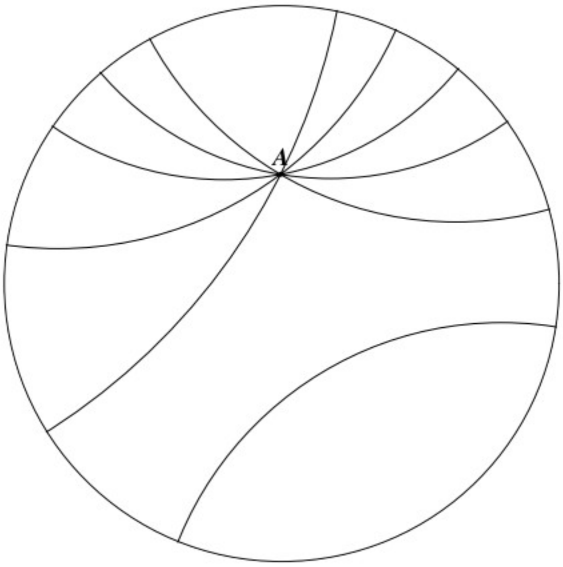

@ 4 ? % + c * * * . .
+ =. : .+ $ * + + % + ;0 * + 1
+ + 4 + 4 % * 4 $
+ .+ + . + =. : 4 =

>4 ; ? 1

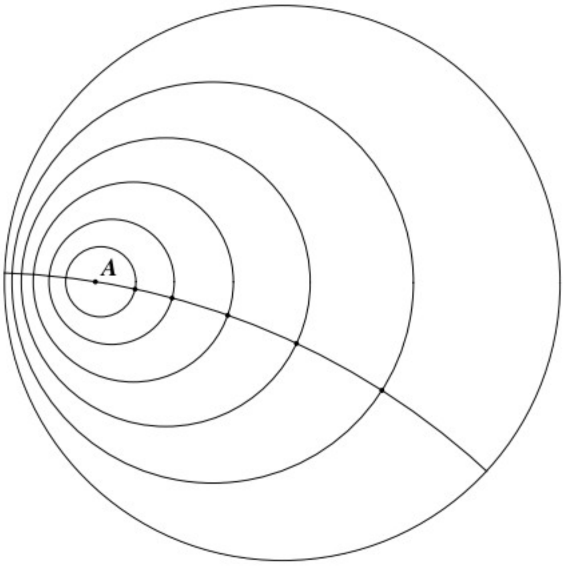

@ 4 & 0 * + $ 0 * +
* 4 +* + F + +: $ $ * 4
$ + % + + + 4 + 4

>4 ; ? 1

.+ + 4 =

1 ! "$/2" " "3 4 + " 8 * 8 $ 4 : +
+ . 4 +4 , * ; 1 .*
3: , ' * <<<

?1 ! "$/2" " "3 4 + " 8 * 8 $ 4 : +
+ . 4 +4 , * ; 1 .*

! "#$%$ & ' ' "( ) " "9+ *:) 8 ! . 4

1 ! "$/2" " "3 4 + " 0* + * (+= ! +
9+ *:) 8 ! . 46 $.. + + / 4 F ! . F
& + . 4F + * / : . 4 ; B1 .*

1 ! "$/2" " "3 4 + " q) 9+ *:)
4 . 4 + 4 4 3. 8 = 8, . 4 ; 1 .*

! "#$%$ & ' ' "( ) " "7 . +

1 ! "$/2" " "3 4 + " ! # + 8/$ 7%O 8$J80
%903/70 $&# 803 0%/7! $,80(/$.*

! "#$%$ & ' ' "( ) " "3 + = + * # . +
8 =

1 ! "$/2" " "3 4 + " # . + 8 = .*

! "#$%$ & ' ' "( ) " "3 ! + !))

1 ! "$/2" " "3 4 + "
3 6! + 6!))@ 6$ 63 4 6 6 * 6$ =6 6 6 + 6 = .*

>4 ; 1 ! +.
+ 4 .+

>4 ; ?1
% 0 4 . +) 4 + +* + 4
+ *+ * ' =*+= $ 4F * ' * *
B 4 + $ =+ * E + F ?
7 B? 2 0 U+: 7 . . * +* . * + +
+ + * + $.. ' * := + + F
+ + 4 =+ * F B F
. * ? 7 8 + (+ + *
J $

% + + * +) +F = +)=
+ . ' * + +: .+ + % + +
+ + += + 4 * + + + 7 4
. * + + * + / : 4 +

0+ D) F * + . * 4
* 4 T) F + * + T

B) T + + % . * . + *) F
) T + + .
H + + + * '+ + * +
F + * . + F = * :+ '+ + C ' =
+ 4 + U + * * + F +) 4 + :+ 0+ D +*
; + F + * := r l1F +
= * + * = % * +
+ * * E := $ + + +: F
: = * + * ;:= + 1 * +
& ' F + * * + * F + *
* + . + + : + + = 2
+ + =F : + 4 * . . + +
+ F +' +4 +* 0+ D : F * * := $ J

>4 B; 1

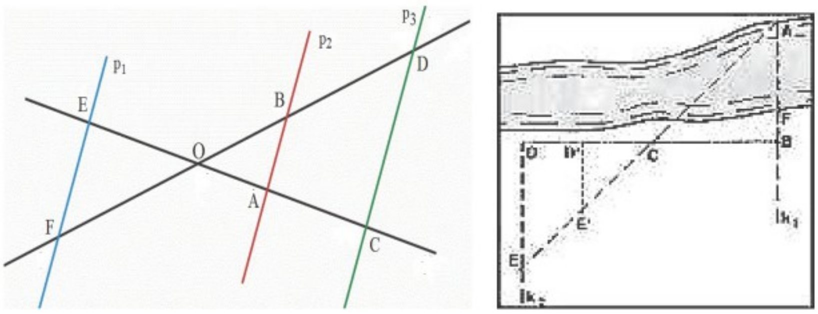

p1
p2
p3
E
B
D
O
A
F
C
A
F
B
D
D'
C
E'
E
K2
K1

@ 4 , % + 4 + * := + 4 + * .+ +
+ + ` / 4 $.. + % + D '+ + * + %
* + ' $(* * 4 F S :=
+ + 4 .+ + * :+) $(T 30

>4 ; B 1

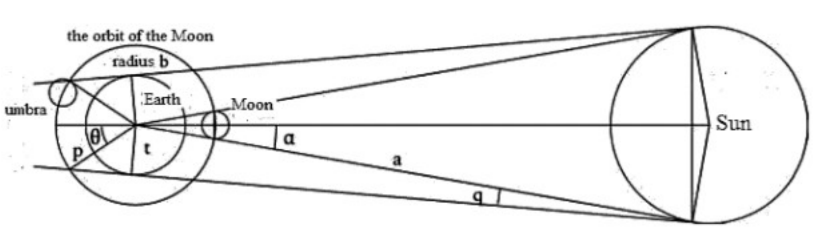

the orbit of the Moon
radius b
Earth
umbra
Moon
θ
p
t
α
a
q
Sun

@ 4 $ + D * + 4 * + + * U
4 0+ D +* * 4 + .

3: , ' * <<<

! "#$%$ & ' ' "( ) " "J * + * 4 . $ F
+ * + +

B1 ! "$/2" " "3 4 + " B J * + * 4 .
$ F + * + + .*
3: , ' * <<<

! "#$%$ & ' ' "( ) " "H 4 H+= O + + +
+=

1 ! "$/2" " "3 4 + " H 4 H+= O
+ + + += .*

1 ! "$/2" " "3 4 + " H 4 H+= O
+ + + += .*

1 ! "$/2" " "3 4 + " H 4 H+= O
+ + + += .*

?1 ! "$/2" " "3 4 + " H 4 H+= O
+ + + += .*

! "#$%$ & ' ' "( ) "

?

! "#$%$ & ' ' "( ) "

! "#$%$ & ' ' "( ) " "$. / + +

1 ! "$/2" " "3 4 + " $. / +
+ .*

! "#$%$ & ' ' "( ) " "(' + * + + > +

1 ! "$/2" " "3 4 + " + + 0 4 4 $
:) B ? 3C * J ' = F J $; 1 .*
3: , ' * <<<

>4 B; 1 ! +. @ *+ ;8 =1
A 9 :

1 ! "$/2" " "3 4 + " % 3C * + *:)
= + + 3C * J ' = ; ?1 .*

! "#$%$ & ' ' "( ) " "# * + + ! + + # j

1 ! "$/2" " "3 4 + " ?
6. 6]6 +6E + \ 6V W! 604 6:.) B) .*

1 ! "$/2" " "3 4 + " ? 04 ! $. j + [4=
* U j 4 $ + 4 >:j *)) * +) @ +) %> + ; ? 1 .*

B1 ! "$/2" " "3 4 + " ? 04 ! # + +: +
s + 4 + t 0* + u s t ; ? 1 .*

1 ! "$/2" " "3 4 + " ? 04 ! = +4 + j
9 : 4 $ + +)+ q [[ )) U+)+ [@ +) %> + ; ? 1 .*

1 ! "$/2" " "3 4 + " ? 04 ! = +4 + j
9 : 4 $ + +)+ q [[ )) U+)+ [@ +) %> + ; ? 1 .*

1 ! "$/2" " "3 4 + " ? 04 ! 0 +
U 7 4 + R += ; ? 1 .*

?1 ! "$/2" " "3 4 + " ? 04 6! 6
6 6) 6U 6' 6# .*

1 ! "$/2" " "3 4 + " ?
7 + 6]6 +6. = E 666V W +)6 +C6:.)? B? .*

1 ! "$/2" " "3 4 + " ? : U .*

1 ! "$/2" " "3 4 + " ? .= +4 + .*

1 ! "$/2" " "3 4 + " ? * 68 $C .*

1 ! "$/2" " "3 4 + " ? .*
3: 0 + '

! "#$%$ & ' ' "( ) " "@ *+ 8 = ;# '
( ) + + 1

B1 ! "$/2" " "3 4 + " B ( ) @ *+ 4 =
0 * + + * (=+ , :+ ') + 4 = . S ' 4 = # ' : +
; 1 .*

1 ! "$/2" " "3 4 + " B K+ ( ) @ *+
4 = 0 * + + * (=+ , :+ ') + 4 = S ' 4 = & 9 + *
; ? 1 .*

1 ! "$/2" " "3 4 + " B K+ ( ) * +
+ + = 4 = : ; ? ?1 .*

1 ! "$/2" " "3 4 + " B K+ ( ) % =
+)+ $)+* + &+) ; ? 1 .*

! "#$%$ & ' ' "( ) " "8 + / + ' = + * 0
0E +

?1 ! "$/2" " "3 4 + " B $ + = F + * + * . =
.*
>4 ; B 1 ! 8 *

B 1 ! "$/2" " "3 4 + " B $ + = F + * + * . =
.*
>4 B; B 1 / + + + * K+ + + *

>4 ; 1 B ! '+ 0 != *

B 1 ! "$/2" " "3 4 + " B ! E (+ 8 +
/ + ' = + * 0 0E + 3C * J ' = F J $; ?1 .*
A .+4

B 1 ! "$/2" " "3 4 + " B ! E (+ O ; * 1 = E
E + E 4 9 + ; ? 1 .*

B 1 ! "$/2" " "3 4 + " B 8 + / + ' = + * 0
0E + 3C * J ' = ; ?1 .*

! "#$%$ & ' ' "( ) " "8 #

B 1 ! "$/2" " "3 4 + " B @ D 8
% + + % + F + . .*

BB1 ! "$/2" " "3 4 + " B 8 4 & @ *)
6 + + * @+ = !+ : *4 J ' = ; ?? 1 .*
>4 ; 1 ! +. B ! ' $ +*

B 1 ! "$/2" " "3 4 + " B 8 4 & @ *) 9 4 *
6 4 4 + * % 4 !+ : *4 J ' = ; 1 .*

B 1 ! "$/2" " "3 4 + " B 8 4 & @ *) ;$ 1
+ 9 4 * # 6 % @ *< $ K 6!/! ; 1 .*
>4 ; ?1 ! +. ? % + 4 F % + 4 $4+

B 1 ! "$/2" " "3 4 + " B 9+ = , *4 F 8 4 @ *)
/ + + . : 4 * + * ' # ' : +
; ? 1 .*
>4 B;B 1 ! +. ! + +
3:
>4 ;B 1 @ 4 + *

CLASSIFICATION

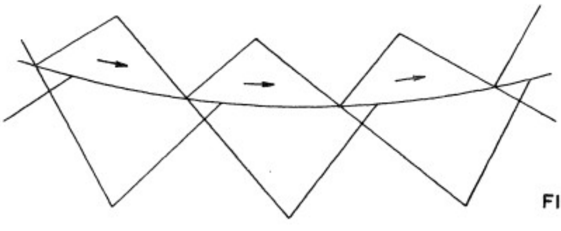

FIG. 10.7

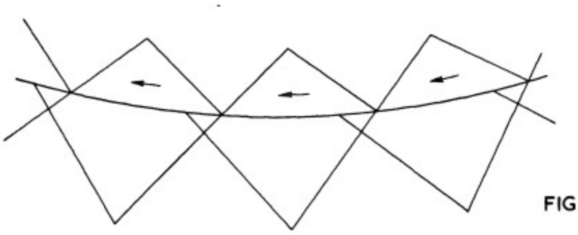

FIG. 10.8

3: 0 *+ + + '+ + <<<
7 + <<< A $ + + * * * . + +E * + +

7 . + 0 *+ * ' * * j <<<

>4 ; 1 ! +. ! ' + #

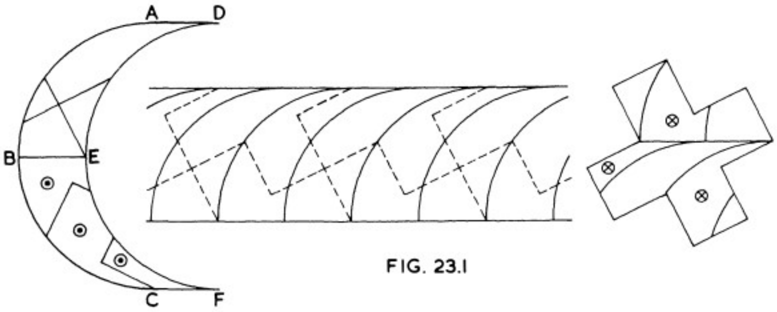

FIG. 23.1

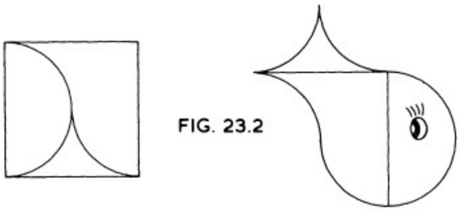
SWEET
AND
LOVELY

FIG. 23.2

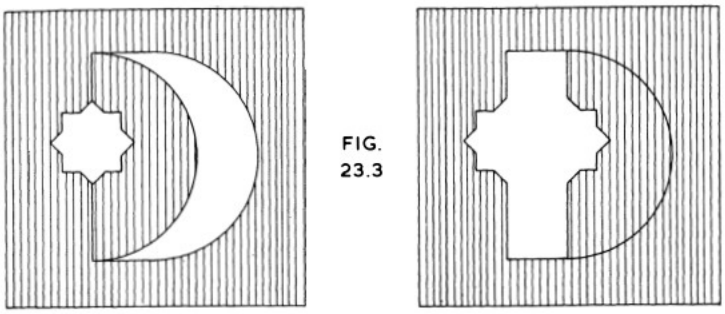

FIG. 23.3

23

CURVILINEAR DISSECTIONS

The other chapters in this book, all concerned with rectilinear dissections, can be read in deep seriousness. But this hardly applies to curvilinear dissections, whose nature fits them only for puzzles and more or less artistic designs.

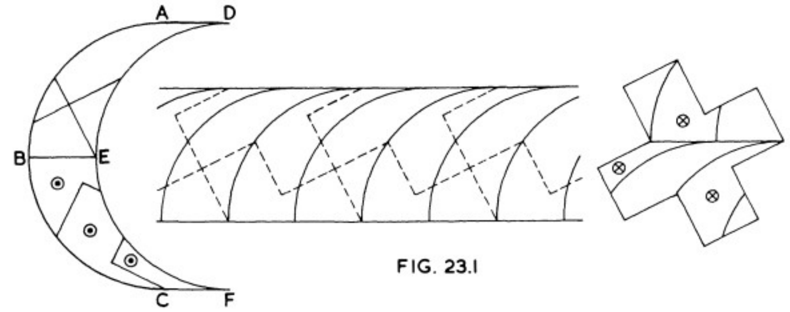

FIG. 23.1

The difficulty in finding curvilinear dissections is that to every convex edge of a figure there must nearly always correspond somewhere else concave edges of the same length and radius for it to fit into. For instance, in the dissection of crescent and

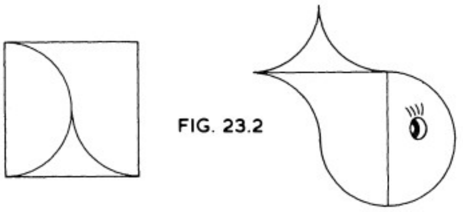

FIG. 23.2

SWEET
AND
LOVELY

Greek cross in Fig. 23.1 the edges of the former must include two equal curved portions *ABC* and *DEF,* and the remainder must be two straight lines *AD* and *CF* (if the figure is to look anything like a crescent).

Again, the shapely contour of the mermaid's rival in Fig. 23.2 has equal lengths of convex and concave curves. But Loyd exploited the difficulty by the trick shown in Fig. 23.3.

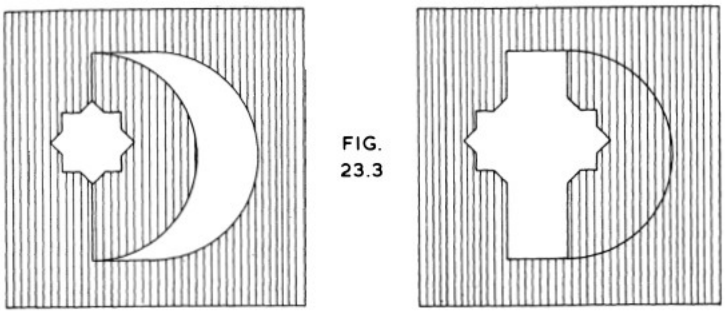

FIG. 23.3

Because of this difficulty, one cannot as a rule choose two curvilinear figures beforehand and try to dissect them; it is rather a matter of seeing what a given curvilinear figure can be dissected into. Thus it is with the circle. One might have in mind its dissection into some kind of curvilinear star, but hardly a goal more precise. And you count yourself lucky to find something as passable as Figs. 23.4-23.6.

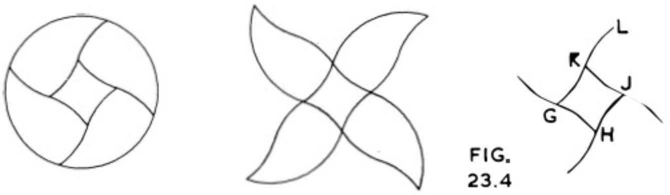

FIG. 23.4

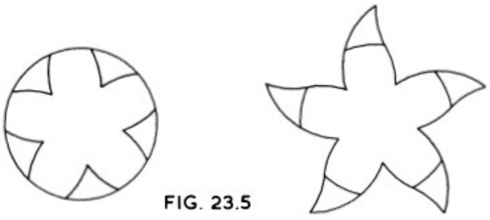

FIG. 23.5

THE STARFISH

The radius of the cuts in Fig. 23.4 is arbitrary. There is less freedom in Figs. 23.5 and 23.6, in that all cuts must have the same radius as the circle.

Working out the cuts for Fig. 23.4 is quite easy. Draw curves of convenient radius joining the vertices G, H, J, K of an n-gon, making L such that GK and KL are equal and collinear; join KL by a curve of the same radius, and complete by drawing a circle through the n points like L. This construction differs from that of Fig. 19.7 only in that the lines are curved.

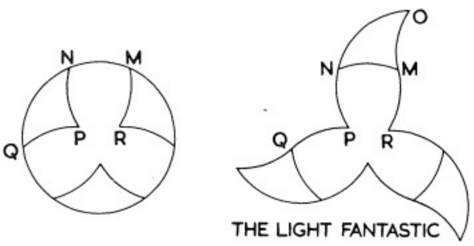

THE LIGHT FANTASTIC

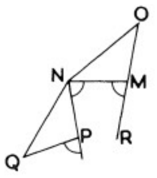

FIG. 23.6

But Figs. 23.5 and 23.6 require a more elaborate approach, to find lines that give smooth joins at points M in the star. (Points N give no trouble, for in curvilinear angles

$$\angle ONM + \angle MNP = \angle NQP + \angle MNP = \angle QNP + \angle MNP = \pi.)$$

To see one's way more clearly in considering the join at M, replace some of the curves in Fig. 23.6 by straight lines as on the right. Then a smooth join at M requires that O, M, R be collinear, whence the three marked angles must be equal. For an n-point star, since arc MNQ is $1/n$ of the circumference, it follows that

$$\angle MNP = \tfrac{2}{3}\angle MNQ = \tfrac{2}{3}(\pi - \pi/n).$$

Since also

$$QP = PN = NM = MO = MR$$

and their length can be given any convenient value, all the points named in Fig. 23.6 can be determined, and, therefore, all points in the star dissection, and the radius of the arcs is that of the circle through M, N, Q. In going about it this way, the radius of the circle is whatever it turns out to be. If on the contrary we start with a circle of given radius R, then we use the fact that

$$MN = 2R \sin \theta,$$

where

$$\cot \theta = \cot (\pi/n) + 2 \sin (\pi/3n + \pi/6)/\sin (\pi/n).$$

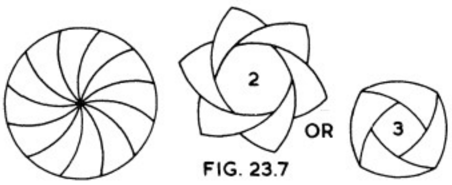

FIG. 23.7

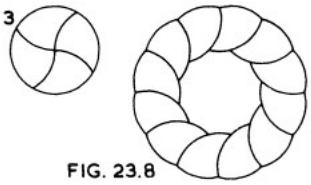

FIG. 23.8

One might also have in mind the dissection of a circle into some kind of ring. Figures 23.7 and 23.8 are typical of what can be found. There are numerous variants, depending on the number of sectors per circle and the number of rings per circle or circles per ring.

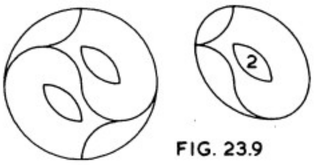

FIG. 23.9

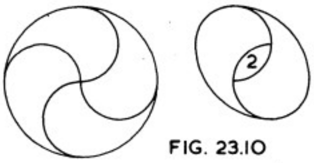

FIG. 23.10

Somewhat similar is the second of the dissections of a circle into two oval stool seats with hand-hole. (Both the dissections are by Sam Loyd—*ex ungue leonem.*) With the orientation of Figs. 23.9 and 23.10, all centers for drawing arcs are at lattice points on a grid of squares of side $\frac{1}{2}R$, where R is the radius of the circle, and their radii are either R or $\frac{1}{2}R$.

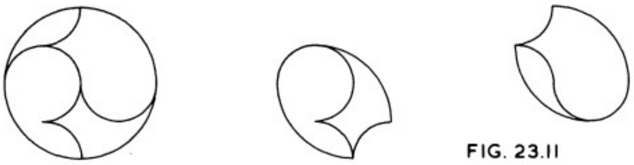

FIG. 23.11

The horseshoe dissection in Fig. 23.11 leans heavily on Fig. 23.9. But the spade-heart dissection in Fig. 23.12 is an isolated curiosity, and an outstanding one.

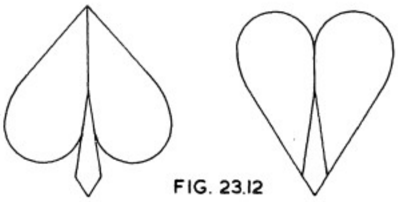

FIG. 23.12

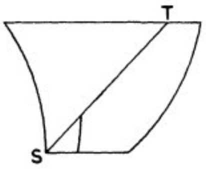

FIG. 23.13

Figure 23.1 shows that an *S*-dissection can be applied to curvilinear figures. So can the *P*-slide, as we see from Fig. 23.13; here the transverse cut *ST* also can be curved. Perhaps a good dissection based on this can be found.

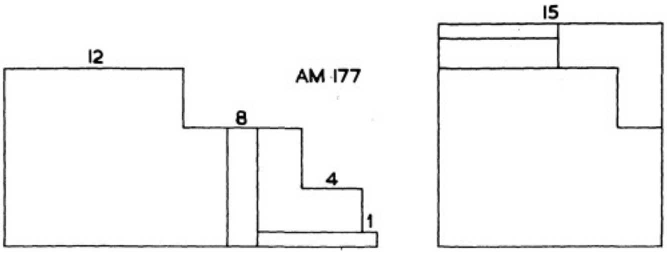

AM 177

! "#$%$ & ' ' "( ) " "8 = + * * + +
;J ' = H+ 1

B?1 ! "$/2" " "3 4 + " B ! + 4 = *
4 . + * + * H+ * 8 = ; 1 *S'

1 ! "$/2" " "3 4 + " B ! + 4 = *
4 . + * + * H+ * 8 = ; 1 .*
>4 ; 1 ! + 8 = .
>4 ; B1 8 = 3 : *

1 ! "$/2" " "3 4 + " B ! + 4 = *
4 . + * + * H+ * 8 = ; 1 *S'

1 ! "$/2" " "3 4 + " B # # + 8 =
7 4 +: $ + + + + = ; 1 .*
>4 ; 1 3 4 +
& 4 + $. + : + S 4+
C * 4 *

1 ! "$/2" " "3 4 + " B # * + 4 =
$; 1 .*

1 ! "$/2" " "3 4 + " B # 4 = H
K . : 4P : *+= + #)) ; 1 .*
>4 %/$& 0/ $, ,7&0 %3 ,7&0 $&# 7&%0/$,

B1 ! "$/2" " "3 4 + " B # 4 = H
K . : 4P : *+= + #)) ; 1 .*

1 ! "$/2" " "3 4 + " B # 4 = H
K . : 4P : *+= + #)) ; 1 *S'

1 ! "$/2" " "3 4 + " B 8 = + * *
+ + ;J ' = H+ 1 .*
3: , ' * <<<
A B .+4

1 ! "$/2" " "3 4 + " B 9+ *:) # + *
! . + + 8 =F % * 0* .*

?1 ! "$/2" " "3 4 + " B 9+ *:) * + *
. + + 4 = ! +. + v 9+ !/! ; 1 .*

1 ! "$/2" " "3 4 + " B 9+ *:) * + *
. + + 4 = ! +. + v 9+ 6!/! ; 1 .*

1 ! "$/2" " "3 4 + " B 9+ + +
$*'+ * @ + * 7 * = !+ .*
>4 ; B 1 ! +. ?

1 ! "$/2" " "3 4 + " B , # 8 =

. 4 v (* + ; 1 .*

1 ! "$/2" " "3 4 + " B . + + + : .*

1 ! "$/2" " "3 4 + " B . + + + : .*

B1 ! "$/2" " "3 4 + " B % : + * ' C
+ * * 4 = !+ : *4 J ' = ; 1 .*

! "#$%$ & ' ' "( ) " "8 = H=

1 ! "$/2" " "3 4 + " B ! / H= Q $*'+ *
0 4 4 + + 8 + 9 ; ? B1 .*

1 ! "$/2" " "3 4 + " B ! / H= Q $*'+ *
0 4 4 + + 8 + 9 ; ? 1 .*

1 ! "$/2" " "3 4 + " B ! / H= Q @ *+
8 = 8 + 9 ; ? 1 .*

?1 ! "$/2" " "3 4 + " B H= 7 * S '
8 = # ' : + ; 1 . :

1 ! "$/2" " "3 4 + " B H= 7 * S '
8 = # ' : + .*

1 ! "$/2" " "3 4 + " B H= ! / F Q $*'+ *
0 4 4 + + *S'

1 ! "$/2" " "3 4 + " B H= ! / F Q $*'+ *
0 4 4 + + .*

! "#$%$ & ' ' "( ) " " .+ % 4 $4 % 0'
8 + 7* + = +4 + 9 : + * 0

1 ! "$/2" " "3 4 + " BB ;! + +.. * + +
1 ! ,+ U , + * + . + = 7 * + + * $.. *
+ + ; ? 1 *S'

1 ! "$/2" " "3 4 + " BB ;! + +.. * + +
1 ! ,+ U , + * + . + = 7 * + + * $.. *
+ + ; ? 1 .*

B1 ! "$/2" " "3 4 + " BB ! ,+ U $.. *
+ + = # ' ; ? 1 *S'

1 ! "$/2" " "3 4 + " BB ! ,+ U $.. *
+ + = # ' ; ? 1 .*
>4 ?; 1 7 2 +* + *

1 ! "$/2" " "3 4 + " BB ! ,+ U .+
% 4 $4 6 % 0' 8 + 7* + = +4 + 9 : + * 0

; ? 1 *S'

1 ! "$/2" " "3 4 + " BB ! ,+ U .+
% 4 $4 6 % 0' 8 + 7* + = +4 + 9 : + * 0
; ? 1 .*

?1 ! "$/2" " "3 4 + " BB ! ,+ U %
+ + + . + ; 1 .*

? 1 ! "$/2" " "3 4 + " BB ! ,+ U %
+ + + . + J ' % ; ?B 1 .*

? 1 ! "$/2" " "3 4 + " BB ,+ U F ! $.. *
$ + = # ' : + ; 1 .*

! "#$%$ & ' ' "( ) " "%) 4 .+ 0 * + *
0

? 1 ! "$/2" " "3 4 + " B ( ) 7 $ $
9 = + + * , : + = ; 1 . :

? 1 ! "$/2" " "3 4 + " B ( ) 7 $ $
9 = + + * .*

? 1 ! "$/2" " "3 4 + " B ( ) F #+' * $ %
!+ K . # : *+= : .*

?B1 ! "$/2" " "3 4 + " B ( ) F #+' * $ %
!+ K . # : *+= : 4 8 . ; 1 . :

? 1 ! "$/2" " "3 4 + " B #+' * ( ) 3 +*' *
$ 4 + *[+ E 4 ' + * 8 : ; 1 .*

? 1 ! "$/2" " "3 4 + " B #+' * ( ) 3 F % F
% 6 $: = 0 + = + + .*

? 1 ! "$/2" " "3 4 + " B #+' * ( ) 3 F % F
% 6 $: = 0 + = + + + ; 1 . :

??1 ! "$/2" " "3 4 + " B #+' * ( ) % K 4
7 .+ 6 0 * + * 9 0 + = + + .*

1 ! "$/2" " "3 4 + " B #+' * ( ) % K 4
7 .+ 6 0 * + * 9 0 (+ ( ) ; 1 . :

! "#$%$ & ' ' "( ) " "% % :

1 ! "$/2" " "3 4 + " B 0# B B .*
A O+ F / : !
% % :
&+ + ! % + + + F 7 F
H+ 4 F # !

! "#$%$ & ' ' "( ) " ?

! "#$%$ & ' ' "( ) " ?"$.. * # + 0E +

1 ! "$/2" " "3 4 + " B $.. * # + 0E + %
+ = ! .*

1 ! "$/2" " "3 4 + " B $.. * # + 0E + %
+ = ! ;7 / 1 .*

! "#$%$ & ' ' "( ) " ?"(+) % J * +C

1 ! "$/2" " "3 4 + " B? (+) % J * +C .*
>4 ?;B 1

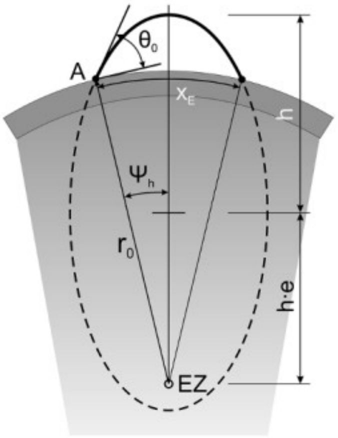
θ0
A
x_E
h
Ψ_h
r_0
h·e
EZ

@ 4:+ R + * # (
.) 4 0 *U 0RF $: U * $:
4+ 4 .) * @ 4)q . F * $: + * * $:
4+ 4 .) ' 0 *U F * 4 w 9+ :+
* 0 . F * 0CU U x * C *
U 4 / 4 * 9+ . +

% +S j + + . + $E E E +
. > *+ + 0RF $ * 4 +
. * .+ *+ * f F + * y + *
. * .+ *+ * *+ +F C +
+ . F + C *+* C + . *
+. + + * MN * C . .+

3: , ' * <<<

! "#$%$ & ' ' "( ) " ?"0 + = % . # +
8 =

B1 ! "$/2" " "3 4 + "
0 + = % . # + 8 = .*
3: , ' * <<<

>4 ; 1 ! +. 8+ +.

>4 ;B 1 ! +. 8 *

>4 B;B 1 ! +. + + % + .

>4 ;? 1 ! +. ! '+ +

>4 ? ; 1 ! +. % 8+ (%

>4 ; 1 ! +. / +

! "#$%$ & ' ' "( ) " ?"@ * ,+ + 7 +

1 ! "$/2" " "3 4 + " @ *6,+
+ 67 + .*

! "#$%$ & ' ' "( ) " ?"@ * + ;# ' ( )
= 1

1 ! "$/2" " "3 4 + " @ * + ;# ' ( )
= 1 .*
A ? .+4

! "#$%$ & ' ' "( ) " ?"@ + + # + 0E + $
$.. + ' + @ + + # '+ '

1 ! "$/2" " "3 4 + " .*
A @ + + # + 0E +

! "#$%$ & ' ' "( ) " ?"8 =

?1 ! "$/2" " "3 4 + " # + 4[[4[[+
* .*

1 ! "$/2" " "3 4 + " 04 6! 6
6 6) 6U 6' 6# .*
3: , ' * <<< 7 . + f <<<
_ .+4

1 ! "$/2" " "3 4 + " 8[[.*

1 ! "$/2" " "3 4 + " Q 6 + +60* 6
6/ U . 6 *6%+: 6 z 6 4 +. 6 *6/ . *)) .*

1 ! "$/2" " "3 4 + " 6 + 6
6 6 . +U 6* 6 .+ 64 6 6 + .*

1 ! "$/2" " "3 4 + " : U .*

B1 ! "$/2" " "3 4 + " .S ' . ? . .*
A 8 $!0 F 9 $!0 $&# H $!0

! "#$%$ & ' ' "( ) " ?"8 =F , % = + * $.. +

1 ! "$/2" " "3 4 + " B
6( )68 =, % =$ *$.. + .*
>4 ; 1 @ # + 9 4 8 + U+ 0

>4 ?; ?1 # 4 * ! ' + * @ 7 4 + + [0 + * 3 ! + =
4 + 8

>4 ; 1 (* 4 & 0 .+ := # 4 = 0 .+

! "#$%$ & ' ' "( ) " ?", 4 +

1 ! "$/2" " "3 4 + " , 4 + .*

! "#$%$ & ' ' "( ) " ?", 4 + 7 + $! +
! * 3.

1 ! "$/2" " "3 4 + "
U 6( )6, 4 + 7 + .*

! "#$%$ & ' ' "( ) " ?" +.. ! + [+ E

?1 ! "$/2" " "3 4 + " # + 8 .*

1 ! "$/2" " "3 4 + " # +
, ! E + # + + ! U$ + + 7: 9 .*
3: 7 . + <<<

1 ! "$/2" " "3 4 + " 4 .*

1 ! "$/2" " "3 4 + " 9+ * .*
A ,0 %/$7%e #0 %9$(7% 7(& 2J//$ J/ ,$ @78J/0 0!%0J/k

1 ! "$/2" " "3 4 + "
9 6 + E 6 6 + 6+6$ C+ * .*

1 ! "$/2" " "3 4 + " 7H9? B .*
A , *P0 * # + # ' + ^ ,+

B1 ! "$/2" " "3 4 + ") .*

1 ! "$/2" " "3 4 + " 46 B .*
A , 4

1 ! "$/2" " "3 4 + "
+.. 6*$ C+ * F6 * 6 + 6 60) 6,+6 6 + [+ E .*
A 7 . + f <<<

1 ! "$/2" " "3 4 + " E +* + + * f .*

?1 ! "$/2" " "3 4 + " !7$ 6 6 ?

B 6 + 6#S ::+ .*
A , , ' -7 * e [*D0 *
' 4 E + +:

1 ! "$/2" " "3 4 + " 6:+ 6 + .*
A $ E * F +.. F # + =+
2 + 0. j* *+ 9 j + *+ 9 f +

1 ! "$/2" " "3 4 + " Q K+ U $ 9 =
+ + + ; 1 .*
A ?? .+4

! "#$%$ & ' ' "( ) " ?" + + + +

1 ! "$/2" " "3 4 + " ?
6( )6 + + + 7 + .*

B B

! "#$%$ & ' ' "( ) "

! "#$%$ & ' ' "( ) " "@ + + . +

1 ! "$/2" " "3 4 + " $ 7 * + +
+ 0E + . 4 +4 & O) ; 1 .*
>4 ; 1 ! +. : ' .+

1 ! "$/2" " "3 4 + " (+ % = @ + +
+ 0E + * 0* H !6H * ; 1 .*

B1 ! "$/2" " "3 4 + " @ + + # + 0E +
$ 7 * @ + + # '+ ' F @ + + # + 0E + .*
>4 B? ; 1 3 * @ + + * 0E +

>4 ; ? 1 * 3 4 + = +
% 7* + *
8 + * ?
/ U @ + + +
, / + , ' @ + + 7 4 + + * # '+ '
B 3 . + / + . @ , / + , ' @ + + 7 4 +
. + / + . @ / 4 / + , ' @ + + 7 4 + ?
p

1 ! "$/2" " "3 4 + " 3 ' F Q 7 *
.+ + * + E + . 4 7 + + ; 1 .*
>4 BB ;B 1

. + H+' + * 9 =4 D . BB
. + H+' BB
K D @ + + * 9 =4 D . BB

% # B
% 9=* 4 $ B
(* + B B
$ 0 4 + + * 2 + & : B

>4 BBB;B 1

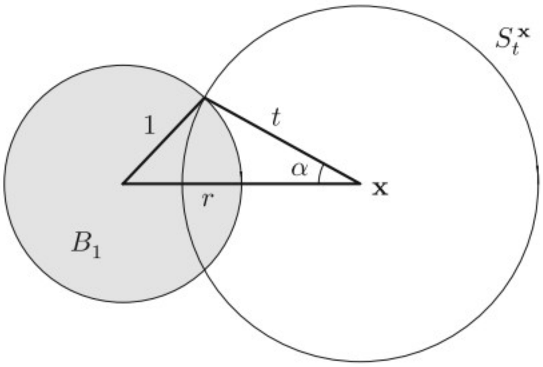
$S_t^{\mathbf{x}}$
1
t
α
$\mathbf{x}$
r
B_1

@ 4 ? ! + . 4 + :+

1 ! "$/2" " "3 4 + " / + * 9 + @ + +
!+ $ 7 * = H * : 4 ! .+ = ; 1 .*

>4 ; 1 ! +. =

% . C = + 4 + . : += : * + + = * * +
+: 4 U = + + G + + * 4 +
+ +*=) = = . :
7 F 4 + = C+4 : * + F +)
+ { +: +' '+ + % +
+ = . { @ ' +
+= : * : * := + 4 { % + *
) + @ * . . += : +
+ * + 4 .
% + +** + * . + +' +.
C+4 '+ + 4 Z + + = = +C
! E = += * : + C+4 = := = = . .

Chapter 16

Symmetries

The complexity of a general problem may be drastically reduced if we are able to recognize inherent symmetries or if a classification according to an already known symmetry is possible.

If, e.g. a symmetric hexagon is inscribed into a circle, a clockwise rotation of 60° about the center leaves its form invariant. The same result follows from any multiple of 60°. Furthermore two consecutive rotations may be described by a single one with 120°. The same holds for counter clockwise rotations. From the listed properties we may combine all these rotations into one discrete rotation group.

There are additional discrete operations which leave the shape of the hexagon invariant e.g. reflections at a symmetry axis.

Consequently we may describe a hexagon only by its symmetry properties.

If we fix the hexagon at its boundaries and investigate the vibrational characteristics of this hexagonal membrane, we have to solve a second order differential equation. We expect, that the symmetry properties mentioned above are still valid for the fundamental mode. This means, that the solution of the wave equation is invariant for rotations about 60°.

Considerations of this type are very important, since they reduce the complexity of a given problem and in many cases are necessary to obtain a valid solution.

The extension from discrete to continuous symmetries, e.g. a rotation with arbitrary angle, a shift with arbitrary length parameter or a measurement at any time leads to the concept of continuous groups, which are called Lie groups [Greiner and Müller (1994)], [Steeb (2007)].

Prominent examples are the components of the classical angular momentum $\vec{L} = \{L_x, L_y, L_z\}$, which are defined as

$$\vec{L} = \vec{r} \times \vec{p} \tag{16.1}$$

They obey the commutation relations

$$[L_x, L_y] = L_z \quad [L_x, L_z] = -L_y \quad [L_y, L_z] = L_x \tag{16.2}$$

and consequently form a non-Abelian group.

The Casimir operator Operator Λ_3^2 commutes with all elements of this group simultaneously:

$$\Lambda_3^2 = L_x^2 + L_y^2 + L_z^2 \tag{16.3}$$

$$[\Lambda_3^2, L_x] = 0 \quad [\Lambda_3^2, L_y] = 0 \quad [\Lambda_3^2, L_z] = 0 \tag{16.4}$$

16.1 Characteristics of fractional group theory

The transition from classical mechanics to quantum mechanics is realized within the framework of canonical quantization by replacing the classical observables $\{x, p\}$ by the corresponding quantum mechanical observables $\{\hat{x}, \hat{p}\}$, which are realized as derivative operators.

In quantum mechanics, calculations are performed using these operators. For example, we obtain for the commutation relation $[\hat{x}, \hat{p}]$

$$\begin{aligned} [\hat{x}, \hat{p}] &= [x, -i\hbar\partial_x] \\ &= -i\hbar(x\partial_x - \partial_x x) \\ &= -i\hbar(x\partial_x - 1 - x\partial_x) \\ &= i\hbar \end{aligned} \tag{16.5}$$

where we have used the Leibniz' product rule.

We want to emphasize that the commutation relations were calculated independently of a specific function set $|\nu\rangle$ they are acting on. Therefore we obtain results, which are valid in general for any reasonable realization of the underlying Hilbert space.

In order to realize the transition to fractional quantum mechanics within the framework of canonical quantization the classical observables $\{x, p\}$ are replaced by (11.10) and (11.11). For the commutation relations $[\hat{X}, \hat{P}]$ are given as

$$[\hat{X}, \hat{P}] = -i\hbar[\hat{x}^\alpha, \hat{D}^\alpha] \tag{16.6}$$

In the following we will show, that commutation relations may not be calculated independently of a specific function set any more. Instead, the results derived are only valid for a given function set, the operators act on.

As an illustrative example we will apply the fractional derivative according to Riemann and will calculate the explicit action of the commutation relations on the specific function set $|\nu\rangle = \hat{x}^{\alpha\nu}$. We obtain:

$$\begin{aligned} -i\hbar[\hat{x}^\alpha, {}_{\mathrm{R}}\hat{D}^\alpha]|\nu\rangle &= -i\hbar \left(\hat{x}^\alpha \frac{\Gamma(1+\nu\alpha)}{\Gamma(1+(\nu-1)\alpha)} |\nu-1\rangle - {}_{\mathrm{R}}\hat{D}^\alpha |\nu+1\rangle \right) \\ &= -i\hbar \left(\frac{\Gamma(1+\nu\alpha)}{\Gamma(1+(\nu-1)\alpha)} - \frac{\Gamma(1+(\nu+1)\alpha)}{\Gamma(1+\nu\alpha)} \right) |\nu\rangle \end{aligned} \tag{16.7}$$

With the abbreviation

$$ {}_{\mathrm{R}}c(\nu,\alpha) = \frac{\Gamma(1+(\nu+1)\alpha)}{\Gamma(1+\nu\alpha)} - \frac{\Gamma(1+\nu\alpha)}{\Gamma(1+(\nu-1)\alpha)} \tag{16.8}$$

the commutation relations (16.6) based on the fractional derivative according to Riemann result as:

$$[\hat{X}, {}_{\mathrm{R}}\hat{P}]\, |\nu\rangle = i\hbar_{\mathrm{R}} c(\nu,\alpha)\, |\nu\rangle \tag{16.9}$$

We obtain the important result that, in contrast to classical quantum mechanics, within the framework of fractional quantum mechanics all calculations depend on the specific representation of the Hilbert space, the fractional operators act on.

The minimum requirement for a fractional operator algebra is the explicit knowledge of the corresponding set of eigenfunctions.

Hence if we present the result of our derivation as

$$[\hat{X}, {}_{\mathrm{R}}\hat{P}] = i\hbar_{\mathrm{R}} c(\nu,\alpha) \tag{16.10}$$

this relation is valid only on the set of square integrable eigenfunctions $|\nu\rangle$.

For the same commutation relations with the Liouville definition of the fractional derivative on the set of eigenfunctions $|\nu\rangle = |x|^{-\alpha\nu}, x < 0$ we obtain:

$$\begin{aligned} -i\hbar[\hat{x}^\alpha, {}_{\mathrm{L}}\hat{D}^\alpha]|\nu\rangle &= -i\hbar \left(\hat{x}^\alpha \frac{\Gamma((\nu+1)\alpha)}{\Gamma(\nu\alpha)} |\nu+1\rangle - {}_{\mathrm{L}}\hat{D}^\alpha |\nu-1\rangle \right) \\ &= -i\hbar \operatorname{sign}(x) \left(|x|^\alpha \frac{\Gamma((\nu+1)\alpha)}{\Gamma(\nu\alpha)} |\nu+1\rangle - {}_{\mathrm{L}}\hat{D}^\alpha |\nu-1\rangle \right) \\ &= -i\hbar \operatorname{sign}(x) \left(\frac{\Gamma((\nu+1)\alpha)}{\Gamma(\nu\alpha)} - \frac{\Gamma(\nu\alpha)}{\Gamma((\nu-1)\alpha)} \right) |\nu\rangle \end{aligned} \tag{16.11}$$

and hence

$$
{}_{\mathrm{L}}c(\nu,\alpha) = \mathrm{sign}(x)\left(\frac{\Gamma(\nu\alpha)}{\Gamma((\nu-1)\alpha)} - \frac{\Gamma((\nu+1)\alpha)}{\Gamma(\nu\alpha)}\right) \quad x<0 \tag{16.12}
$$

$$
= \frac{\Gamma((\nu+1)\alpha)}{\Gamma(\nu\alpha)} - \frac{\Gamma(\nu\alpha)}{\Gamma((\nu-1)\alpha)} \tag{16.13}
$$

Finally, applying the fractional derivative according to Caputo the eigenfunctions of the commutation relations are given by $|\nu\rangle = x^{\nu\alpha}$ and therefore we may proceed

$$
{}_{\mathrm{C}}c(\nu,\alpha) = \begin{cases} \frac{\Gamma(1+(\nu+1)\alpha)}{\Gamma(1+\nu\alpha)} - \frac{\Gamma(1+\nu\alpha)}{\Gamma(1+(\nu-1)\alpha)} & \nu \neq 0 \\ \frac{\Gamma(1+(\nu+1)\alpha)}{\Gamma(1+\nu\alpha)} & \nu = 0 \end{cases} \tag{16.14}
$$

Obviously the sets of eigenfunctions for the different commutation relations are eigenfunctions of the fractional one dimensional Euler operator $j_1^{\mathrm{E}} = \hat{x}^\alpha \hat{D}^\alpha$ as well. The corresponding eigenvalues follow as:

$$
{}_{\mathrm{R}}j_1^{\mathrm{E}} = \frac{\Gamma(1+\nu\alpha)}{\Gamma(1+(\nu-1)\alpha)} \tag{16.15}
$$

$$
{}_{\mathrm{C}}j_1^{\mathrm{E}} = \begin{cases} \frac{\Gamma(1+\nu\alpha)}{\Gamma(1+(\nu-1)\alpha)} & \nu \neq 0 \\ 0 & \nu = 0 \end{cases} \tag{16.16}
$$

$$
{}_{\mathrm{L}}j_1^{\mathrm{E}} = -\frac{\Gamma((\nu+1)\alpha)}{\Gamma(\nu\alpha)} \tag{16.17}
$$

We conclude that an extension from standard group theory to fractional group theory is confronted with some additional difficulties. Despite this fact a group theoretical approach for a classification of multiplets of fractional extensions of standard groups will turn out to be a powerful tool to solve problems.

Especially nontrivial symmetries in multi-dimensional spaces, like fractional rotational symmetry in $\mathbb{R}^3$ may successfully be investigated. Instead of solving the fractional Schrödinger equation in hyper-spherical coordinates directly, group theoretical methods provide elegant methods and strategies, to determine the eigenvalue spectrum of such highly symmetric problems analytically.

This will be explicitly shown in the next section.

16.2 The fractional rotation group SO_N^α

The components of the classical angular momentum in $\mathbb{R}^3$ are given by

$$
L_{ij} = x_i p_j - x_j p_i \qquad i,j = 1,2,3 \tag{16.18}
$$

1 ! "$/2" " "3 4 + " / + + # +
0E + % = + * ! . = ' * : H = ; 1 .*
A ? .+4
3: , ' * <<<
A >4 ?; ? 1 + ! * +
A >4 ; 1 . + ! * +

! "#$%$ & ' ' "( ) " "8 + / + ' = $ 7 *
=

?1 ! "$/2" " "3 4 + " @ *+ + +
. = + !J ; 1 .*
>4 ? ;B 1 , F + + * ' 4 +

1 ! "$/2" " "3 4 + " * + +
@ *+ + + = + !+ : *4 J ' = ; 1 .*

1 ! "$/2" " "3 4 + " 0 + + + +
* = + !+ : *4 J ' = ; 1 .*

1 ! "$/2" " "3 4 + " 8 + / + ' = $ 7 *
@ = !+ : *4 J ' ; 1 .*
>4 ; 1 ! +. % E '+ . . + * .+ '+

>4 ; ? 1 ! +. % 4 +' + + * E +

>4 ? ; 1 ! +. ? % + U * 4 =

3:
>4 B; B1 @ 4

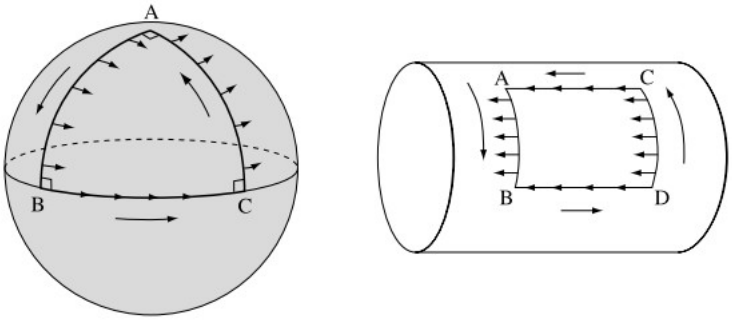

@ 4 + + + . + * + * ' + + .
+ * + + = *

>4 ; 1 @ 4

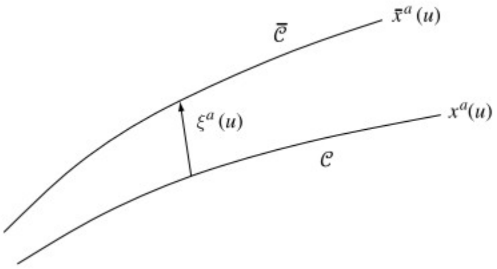
$\bar{\mathcal{C}}$
$\bar{x}^a(u)$
$\xi^a(u)$
$x^a(u)$
$\mathcal{C}$

>4 ; 1 @ 4 B

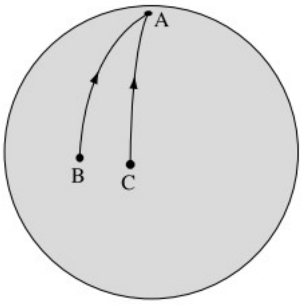

@ 4 B ! ' 4 4 4 * + + .

1 ! "$/2" " "3 4 + " + + + * =
+ * 0 4 4 $! . ' 8 * % * 0* !+ : *4 J ' = ; 1 .*

1 ! "$/2" " "3 4 + " + + + * . =
+ * 4 4 !+ : *4 J ' = ; 1 .*
A .+4

B1 ! "$/2" " "3 4 + " * + +
+ + + * = + * 0 4 4 % * 0* !+ : *4 J ' =
; 1 .*

1 ! "$/2" " "3 4 + " * + +
+ + + * = + * 0 4 4 % * 0* !+ : *4 J ' = .*

! "#$%$ & ' ' "( ) " "7 . +

1 ! "$/2" " "3 4 + " + R U % $ + * ! +
: ' 4F * 0* H = ; 1 .*

1 ! "$/2" " "3 4 + " + R U % + + * +
. : ' 4 Q H = v ; 1 .*
>4 B ! +. %+ ' 4 :

>4 B?; B1 = =

Symmetry involves finding or imposing order in a concrete way, for example, by reflection. Other tactics find or exploit order, but in more abstract, almost "metaphorical" ways. We shall discuss three such methods of "pseudosymmetry." All of them rely on very simple observations that sometimes yield amazingly useful information. The first tactic, the *extreme principle*, which we encountered with the Affirmative Action problem (Example 2.1.9 on page 19), works by focusing on the largest and smallest entities within a problem. Next, the *pigeonhole principle* stems from the nearly vacuous observation that if you have more guests than spare rooms, some of the guests will have to share rooms. Our final tactic, the concept of *invariants*, shows how much information comes your way when you restrict your attention to a narrow aspect of your problem that does not change (such as parity). This is an extremely powerful idea, fundamental to mathematics, that underlies many seemingly different tactics and tools.

3.1 Symmetry

We all have an intuitive idea of symmetry; for example, everyone understands that circles are symmetrical. It is helpful, however, to define symmetry in a formal way, if only because this will expand our notion of it. We call an object *symmetric* if there are one or more non-trivial "actions" that leave the object unchanged. We call the actions that do this the *symmetries* of the object.[1]

We promised something formal, but the above definition seems pretty vague. What do we mean by an "action"? Almost anything! We are being deliberately vague, because our aim is to let you see symmetry in as many situations as possible. Here are a few examples.

Example 3.1.1 A square is symmetric with respect to reflection about a diagonal. The reflection is one of the several symmetries of the square. Other symmetries include rotation clockwise by 90 degrees and reflection about a line joining the midpoints of two opposite sides.

Example 3.1.2 A circle has infinitely many symmetries, for example, rotation clockwise by α degrees for any α.

Example 3.1.3 The doubly infinite sequence

$$\ldots, 3, 1, 4, 3, 1, 4, 3, 1, 4, \ldots$$

is symmetrical with respect to the action "shift everything three places to the right (or left)."

Why is symmetry important? Because it gives you "free" information. If you know that something is, say, symmetric with respect to 90-degree rotation about some point, then you only need to look at one-quarter of the object. And you also know that the center of rotation is a "special" point, worthy of close investigation. You will see these ideas in action below, but before we begin, let us mention two things to keep in mind as you ponder symmetry:

- The strategic principles of peripheral vision and rule-breaking tell us to look for symmetry in unlikely places, and not to worry if something is almost, but not quite symmetrical. In these cases, it is wise to proceed as if symmetry is present, since we will probably learn something useful.

---

[1] We deliberately avoid the language of transformations and automorphisms that would be demanded by a mathematically precise definition.

- An informal alternate definition of symmetry is “harmony.” This is even vaguer than our “formal” definition, but it is not without value. Look for harmony, and beauty, whenever you investigate a problem. If you can do something that makes things more harmonious or more beautiful, *even if you have no idea how to define these two terms*, then you are often on the right track.

Geometric Symmetry

Most geometric investigations profit by a look at symmetry. Ask these questions about symmetry:

- Is it present?
- If not, can it be imposed?
- How can it then be exploited?

Here is a simple but striking example.

Example 3.1.4 A square is inscribed in a circle that is inscribed in a square. Find the ratio of the areas of the two squares.

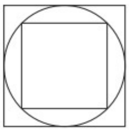

Solution: The problem can certainly be solved algebraically (let x equal the length of the small square, then use the Pythagorean theorem, etc.), but there is a nicer approach. The original diagram is full of symmetries. We are free to rotate and/or reflect many shapes and still preserve the areas of the two squares. How do we choose from all these possibilities? We need to use the hypothesis that the objects are inscribed in one another. If we rotate the small square by 45 degrees, its vertices now line up with the points of tangency between the circle and the large square, and instantly the solution emerges.

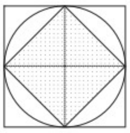

The small square obviously has half the area of the larger! ■

The simplest geometric symmetries are rotational and reflectional. Always check to see if rotations or reflections will impose order on your problem. The next example shows the power of imposing reflectional symmetry.

Example 3.1.5 Your cabin is two miles due north of a stream that runs east-west. Your grandmother's cabin is located 12 miles west and one mile north of your cabin. Every day, you go from your cabin to Grandma's, but first visit the stream (to get fresh water for Grandma). What is the length of the route with minimum distance?

Solution: First, *draw a picture*! Label your location by Y and Grandma's by G. Certainly, this problem can be done with calculus, but it is very ugly (you need to differentiate the sum of two radicals, for starters). The problem appears to have no symmetry in it, but the stream is practically begging you to reflect across it! Draw a sample path (shown below as YA followed by AG) and look at its reflection. We call the reflections of your house and Grandma's house Y' and G', respectively.

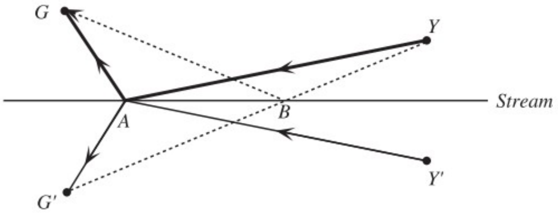

While you are carrying water to Grandma, your duplicate in an alternate universe is doing the same, only south of the stream. Notice that $AG = AG'$, so the length of your path would be *unchanged* if you visited the reflected Grandma instead of the real one. Since the shortest distance between two points is the straight line YBG', the optimal path will be YB followed by BG. Its length is the same as the length of YBG', which is just the hypotenuse of a right triangle with legs 5 and 12 miles. Hence our answer is 13 miles. ■

When pondering a symmetrical situation, you should always focus briefly on the "fixed" objects that are unchanged by the symmetries. For example, if something is symmetric with respect to reflection about an axis, that axis is fixed and worthy of study (the stream in the previous problem played that role). Here is another example, a classic problem that exploits rotational symmetry along with a crucial fixed point.

Example 3.1.6 Four bugs are situated at each vertex of a unit square. Suddenly, each bug begins to chase its counterclockwise neighbor. If the bugs travel at 1 unit per minute, how long will it take for the four bugs to crash into one another?

Solution: The situation is rotationally symmetric in that there is no one "distinguished" bug. If their starting configuration is that of a square, then they will always maintain that configuration. This is the key insight, believe it or not, and it is a very profitable one!

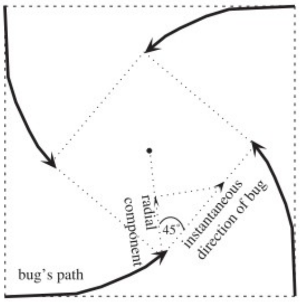

As time progresses, the bugs form a shrinking square that rotates counterclockwise. The center of the square does not move. This center, then, is the only "distinguished" point, so we focus our analysis on it. Many otherwise intractable problems become easy once we shift our focus to the natural frame of reference; in this case we should consider a *radial* frame of reference, one that rotates with the square. For example, pick one of the bugs (it doesn't matter which one!), and look at the line segment from the center of the square to the bug. This segment will rotate counterclockwise, and (more importantly) shrink. When it has shrunk to zero, the bugs will have crashed into one another. How fast is it shrinking? Forget about the fact that the line is rotating. *From the point of view of this radial line,* the bug is always traveling at a 45° angle. Since the bug travels at unit speed, its radial velocity component is just $1 \cdot \cos 45° = \sqrt{2}/2$ units per minute, i.e., the radial line shrinks at this speed. Since the original length of the radial line was $\sqrt{2}/2$, it will take just 1 minute for the bugs to crash. ■

Here is a simple calculus problem that can certainly be solved easily in a more conventional way. However, our method below illustrates the power of the Draw a Picture strategy coupled with symmetry, and can be applied in many harder situations.

Example 3.1.7 Compute $\int_0^{\frac{1}{2}\pi} \cos^2 x\, dx$ in your head.

Solution: Mentally draw the sine and cosine graphs from 0 to $\frac{1}{2}\pi$, and you will notice that they are symmetric with respect to reflection about the vertical line $x = \frac{1}{4}\pi$. Thus

$$\int_0^{\frac{1}{2}\pi} \cos^2 x\, dx = \int_0^{\frac{1}{2}\pi} \sin^2 x\, dx.$$

Consequently, the value that we seek is

$$\frac{1}{2}\int_0^{\frac{1}{2}\pi} (\cos^2 x + \sin^2 x)\, dx = \frac{1}{2}\int_0^{\frac{1}{2}\pi} 1 \cdot dx = \frac{\pi}{4}.$$ ■

The next problem, from the 1995 IMO, is harder than the others, but only in a "technical" way. In order to solve it, you need to be familiar with *Ptolemy's Theorem*, which states

> *Let $ABCD$ be a cyclic quadrilateral, i.e., a quadrilateral whose vertices lie on a circle. Then*
> $$AB \cdot CD + AD \cdot BC = AC \cdot BD.$$

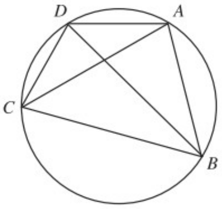

See Problems 8.4.29 and 8.5.49 for different ideas for a proof. A key feature of cyclic quadrilaterals is the easily verified fact (see page 267) that

> *A quadrilateral is cyclic if and only if its opposite angles are supplementary (add up to 180 degrees).*

Example 3.1.8 Let $ABCDEF$ be a convex hexagon with $AB = BC = CD$ and $DE = EF = FA$, such that $\angle BCD = \angle EFA = \pi/3$. Suppose G and H are points in the interior of the hexagon such that $\angle AGB = \angle DHE = 2\pi/3$. Prove that

$$AG + GB + GH + DH + HE \geq CF.$$

Solution: First, as for all geometry problems, draw an accurate diagram, using pencil, compass and ruler. Look for symmetry. Note that BCD and EFA are equilateral triangles, so that $BD = BA$ and $DE = AE$. By the symmetry of the figure, it is seems profitable to reflect about BE. Let C' and F' be the reflections of C and F respectively.

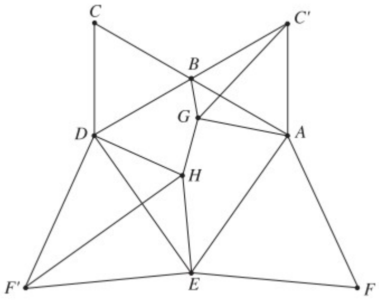

A tactic that often works for geometric inequalities is to look for a way to compare the sum of several lengths with a single length, since the shortest distance between two points is a straight line. Quadrilaterals $AGBC'$ and $HEF'D$ are cyclic, since the opposite angles add up to 180 degrees. Ptolemy's Theorem then implies that $AG \cdot BC' + GB \cdot AC' = C'G \cdot AB$. Since ABC' is equilateral, this implies that $AG + GB = C'G$. Similarly, $DH + HE = HF'$. The shortest path between two points is a straight line. It follows that

$$CF = C'F' \leq C'G + GH + HF' = AG + GB + GH + DH + HE,$$

with equality if and only if G and H both lie on $C'F'$. ■

Algebraic Symmetry

Don't restrict your notions of symmetry to physical or geometric objects. For example, sequences can have symmetry, like this row of Pascal's Triangle:

$$1, 6, 15, 20, 20, 15, 6, 1.$$

That's only the beginning. In just about any situation where you can imagine "pairing" things up, you can think about symmetry. And thinking about symmetry almost always pays off.

The Gaussian Pairing Tool

Carl Friedrich Gauss (1777–1855) was certainly one of the greatest mathematicians of all times. Many stories celebrate his precocity and prodigious mental power. No one knows how true these stories are, because many of them are attributable only to Gauss himself. The following anecdote has many variants. We choose one of the simplest.

When Gauss was 10, his teacher punished his class with a seemingly tedious sum:

$$1 + 2 + 3 + \cdots + 98 + 99 + 100.$$

While the other students slowly added the numbers, little Carl discovered a shortcut and immediately arrived at the answer of 5,050. He was the only student to find the correct sum. His insight was to notice that 1 could be paired with 100, 2 with 99, 3 with 98, etc. to produce 50 identical sums of 101. Hence the answer of $101 \cdot 50 = 5050$. Another, more formal way of doing this is to write the sum in question twice, first forward, then backward:

$$\begin{array}{ccccccccc} S & = & 1 & + & 2 & +\cdots+ & 99 & + & 100, \\ S & = & 100 & + & 99 & +\cdots+ & 2 & + & 1. \end{array}$$

Then, of course, $2S = 100 \cdot 101$. The advantage of this method is that it does not matter whether the number of terms was even or odd (notice that this is an issue with the original pairing method).

This is a pretty good trick, especially for a 10-year-old, and it has many applications. Don't be restricted to sums. Look for any kind of *symmetry* in a problem and then investigate whether a clever pairing of items can simplify things. First, let's tackle the "Locker problem," which we first encountered in Example 2.2.3 on page 26. Recall that we reduced this problem to

Prove that $d(n)$ is odd if and only if n is a perfect square, where $d(n)$ denotes the number of divisors of n, including 1 and n. Now we can use the Gaussian pairing tool. The symmetry here is that one can always pair a divisor d of n with n/d. For example, if $n = 28$, it is "natural" to pair the divisor 2 with the divisor 14. Thus, as we go through the list of divisors of n, each divisor will have a unique "mate" *unless n is a perfect square*, in which case $\sqrt{n}$ is paired with itself. For example, the divisors of 28 are $1, 2, 4, 7, 14, 28$, which can be rearranged into the pairs $(1, 28)$, $(2, 14)$, $(4, 7)$, so clearly $d(28)$ is even. On the other hand, the divisors of the perfect square 36 are $1, 2, 3, 4, 6, 9, 12, 18, 36$, which pair into $(1, 36)$, $(2, 18)$, $(3, 12)$, $(4, 9)$, $(6, 6)$. Notice that 6 is paired with itself, so the actual count of divisors is odd (several true pairs plus the 6, which can only be counted once). We can conclude that $d(n)$ is odd if and only if n is a perfect square.[2] ■

The argument used can be repeated almost identically, yet in a radically different context, to prove a well-known theorem in number theory.

Example 3.1.9 *Wilson's Theorem.* Prove that for all primes p,

$$(p-1)! \equiv -1 \pmod{p}.$$

Solution: First try an example. Let $p = 13$. Then the product in question is $1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9 \cdot 10 \cdot 11 \cdot 12$. One way to evaluate this product modulo 13 would be to multiply it all out, but that is not the problem solver's way! We should look for smaller subproducts that are easy to compute modulo p. The easiest numbers to compute with are $0, 1$, and -1. Since p is a prime, none of the factors $1, 2, 3, \ldots p-1$ are congruent to 0 modulo p. Likewise, no subproducts can be congruent to 0. On the other hand, the primality of p implies that every non-zero number has a *unique* multiplicative inverse modulo p; i.e., if x is not a multiple of p then there is a unique $y \in \{1, 2, 3, \ldots, p-1\}$ such that $xy \equiv 1 \pmod{p}$. (Recall the proof of this assertion in Example 2.3.4 on page 41.)

Armed with this information, we proceed as in the locker problem: Pair each x between 1 and $p-1$ with its "natural" mate y between 1 and $p-1$ such that $xy \equiv 1 \pmod{p}$. For example, if $p = 13$, the pairs are

$$(1,1), (2,7), (3,9), (4,10), (5,8), (6,11), (12,12),$$

and we can rewrite

$$12! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9 \cdot 10 \cdot 11 \cdot 12$$

as

$$1 \cdot (2 \cdot 7)(3 \cdot 9)(4 \cdot 10)(5 \cdot 8)(6 \cdot 11) \cdot 12 \equiv 1 \cdot 1 \cdot 1 \cdot 1 \cdot 1 \cdot 1 \cdot 12 = -1 \pmod{13}.$$

Notice that 1 and 12 were the only elements that paired with themselves. Notice also that $12 \equiv -1 \pmod{13}$. We will be done in general if we can show that x is its own multiplicative inverse modulo p if and only if $x = \pm 1$. But this is easy: If $x^2 \equiv 1 \pmod{p}$, then

$$x^2 - 1 = (x-1)(x+1) \equiv 0 \pmod{p}.$$

Because p is prime, this implies that either $x - 1$ or $x + 1$ is a multiple of p; i.e., $x \equiv \pm 1 \pmod{p}$ are the *only* possibilities. ■

---

[2] A completely different argument uses basic counting methods and parity. See Problem 6.1.21 on page 196.

Symmetry in Polynomials and Inequalities

Algebra problems with many variables or of high degree are often intractable unless there is some underlying symmetry to exploit. Here is a lovely example.

Example 3.1.10 Solve $x^4 + x^3 + x^2 + x + 1 = 0$.

Solution: While there are other ways of approaching this problem (see page 122), we will use the symmetry of the coefficients as a starting point to impose yet more symmetry, on the degrees of the terms. Simply divide by x^2:

$$x^2 + x + 1 + \frac{1}{x} + \frac{1}{x^2} = 0.$$

This looks no simpler, but note that now there is more symmetry, for we can collect "like" terms as follows:

$$x^2 + \frac{1}{x^2} + x + \frac{1}{x} + 1 = 0. \tag{3.1.1}$$

Now make the *substitution* $u := x + \frac{1}{x}$. Note that

$$u^2 = x^2 + 2 + \frac{1}{x^2},$$

so (3.1.1) becomes $u^2 - 2 + u + 1 = 0$, or $u^2 + u - 1 = 0$, which has solutions

$$u = \frac{-1 \pm \sqrt{5}}{2}.$$

Solving $x + \frac{1}{x} = u$, we get $x^2 - ux + 1 = 0$, or

$$x = \frac{u \pm \sqrt{u^2 - 4}}{2}.$$

Putting these together, we have

$$x = \frac{\frac{-1 \pm \sqrt{5}}{2} \pm \sqrt{\left(\frac{-1 \pm \sqrt{5}}{2}\right)^2 - 4}}{2} = \frac{-1 \pm \sqrt{5} \pm i\sqrt{10 \pm 2\sqrt{5}}}{4}. \qquad \blacksquare$$

The last few steps are mere "technical details." The two crux moves were to increase the symmetry of the problem and then make the symmetrical substitution $u = x + x^{-1}$.

In the next example, we use symmetry to reduce the complexity of an inequality.

Example 3.1.11 Prove that

$$(a + b)(b + c)(c + a) \geq 8abc$$

is true for all positive numbers a, b, and c, with equality only if $a = b = c$.

Solution: Observe that the alleged inequality is symmetric, in that it is unchanged if we permute any of the variables. This suggests that we not multiply out the left side (rarely a wise idea!) but instead look at the factored parts, for the sequence

$$a+b, \quad b+c, \quad c+a$$

can be derived by just looking at the term $a+b$ and then performing the *cyclic permutation* $a \mapsto b$, $b \mapsto c$, $c \mapsto a$ once and then twice.

The simple two-variable version of the *Arithmetic-Geometric-Mean* inequality (see Section 5.5 for more details) implies

$$a+b \geq 2\sqrt{ab}.$$

Now, just perform the cyclic permutations

$$b+c \geq 2\sqrt{bc}$$

and

$$c+a \geq 2\sqrt{ca}.$$

The desired inequality follows by multiplying these three inequalities. ■

It is worth exploring the concept of cyclic permutation in more detail. Given an n-variable expression $f(x_1, x_2, \ldots, x_n)$, we will denote the *cyclic sum* by

$$\sum_\sigma f(x_1, x_2, \ldots, x_n) := f(x_1, x_2, \ldots, x_n) + f(x_2, x_3, \ldots, x_n, x_1) + \cdots + f(x_n, x_1, \ldots, x_{n-1}).$$

For example, if our variables are x, y, and z, then

$$\sum_\sigma x^3 = x^3 + y^3 + z^3 \quad \text{and} \quad \sum_\sigma xz^2 = xz^2 + yx^2 + zy^2.$$

Let us use this notation to factor a symmetric cubic in three variables.

Example 3.1.12 Factor $a^3 + b^3 + c^3 - 3abc$.

Solution: We hope for the best and proceed naively, making sure that our guesses stay symmetric. The simplest guess for a factor would be $a+b+c$, so let us try it. Multiplying $a+b+c$ by $a^2+b^2+c^2$ would give us the cubic terms, with some error terms. Specifically, we have

$$\begin{aligned}a^3+b^3+c^3-3abc &= (a+b+c)(a^2+b^2+c^2)-\sum_\sigma(a^2b+b^2a)-3abc\\ &= (a+b+c)(a^2+b^2+c^2)-\sum_\sigma(a^2b+b^2a+abc)\\ &= (a+b+c)(a^2+b^2+c^2)-\sum_\sigma(ab(a+b+c))\\ &= (a+b+c)\left(a^2+b^2+c^2-\sum_\sigma ab\right)\\ &= (a+b+c)(a^2+b^2+c^2-ab-bc-ac).\end{aligned}$$

Notice how the $\sum_\sigma$ notation saves time and, once you get used to it, reduces the chance for an error.

Problems and Exercises

3.1.13 Find the length of the shortest path from the point (3,5) to the point (8,2) that touches the x-axis and also touches the y-axis.

3.1.14 In Example 1.2.1 on page 4, we saw that the product of four consecutive integers is always one less than a square. What can you say about the product of four consecutive terms of an arbitrary arithmetical progression, e.g., $3\cdot 8\cdot 13\cdot 18$?

3.1.15 Find (and prove) a nice formula for the product of the divisors of any integer. For example, if $n=12$, the product of its divisors is

$$1\cdot 2\cdot 3\cdot 4\cdot 6\cdot 12=1728.$$

You may want use the $d(n)$ function (defined in the solution to the Locker problem on page 64) in your formula.

3.1.16 How many subsets of the set $\{1,2,3,4,\ldots,30\}$ have the property that the sum of the elements of the subset is greater than 232?

3.1.17 (Putnam 1998) Given a point (a,b) with $0<b<a$, determine the minimum perimeter of a triangle with one vertex at (a,b), one on the x-axis, and one on the line $y=x$. You may assume that a triangle of minimum perimeter exists.

3.1.18 A polynomial in several variables is called *symmetric* if it is unchanged when the variables are permuted. For example,

$$f(x,y,z):=x^2+y^2+z^2+xyz$$

is symmetric, since

$$\begin{aligned}f(x,y,z) &= f(x,z,y)=f(y,x,z)=f(y,z,x)\\ &= f(z,x,y)=f(z,y,x).\end{aligned}$$

A polynomial in several variables is called *homogeneous* of degree r if all of the terms are degree r. For example, $g(x,y):=x^2+5xy$ is homogeneous of degree 2. (The $5xy$ term is considered to have degree 2, since it is the product of two degree-1 terms.) In general, the k-variable polynomial $g(x_1,x_2,\ldots,x_k)$ is homogeneous of degree r if for all t, we have

$$g(tx_1,tx_2,\ldots,tx_k)=t^r g(x_1,x_2,\ldots,x_k).$$

Given three variables x,y,z, we define the *elementary symmetric functions*

$$\begin{aligned}s_1 &:= x+y+z,\\ s_2 &:= xy+yz+zx,\\ s_3 &:= xyz.\end{aligned}$$

The elementary symmetric function s_k is symmetric, homogeneous of degree k, and all its coefficients are equal to 1. Elementary symmetric functions can be defined for any number of variables. For example, for four variables x,y,z,w, they are

$$\begin{aligned}s_1 &:= x+y+z+w,\\ s_2 &:= xy+xz+xw+yz+yw+zw,\\ s_3 &:= xyz+xyw+xzw+yzw,\\ s_4 &:= xyzw.\end{aligned}$$

! "#$%$ & ' ' "( ) " " + C + * # + @ *+ +
+ * $.. +

?1 ! "$/2" " "3 4 + "
+ C6+ *6# + 6@ *+ + 6+ *6$.. + .*
A + < @ + * U *+

! "#$%$ & ' ' "( ) " "2 + , 4+ # .+ @ +
J + H *

B 1 ! "$/2" " "3 4 + " 2 + , 4+ # .+
@ + J + H * .*

! "#$%$ & ' ' "( ) " " . = 7* + +
+ + + * $

B 1 ! "$/2" " "3 4 + " B . = 7* + +
+ + + * $.*
>4 ; 1

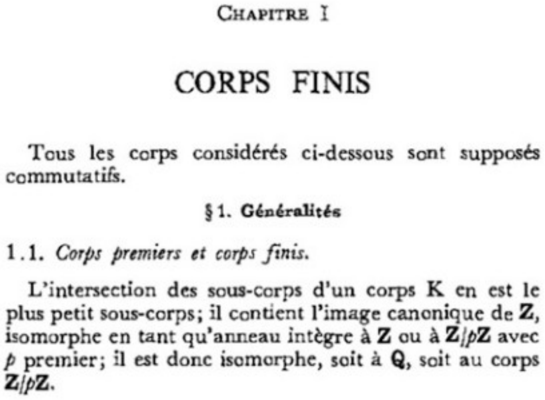

Chapitre I

CORPS FINIS

Tous les corps considérés ci-dessous sont supposés commutatifs.

§ 1. Généralités

1.1. *Corps premiers et corps finis.*

L'intersection des sous-corps d'un corps K en est le plus petit sous-corps; il contient l'image canonique de **Z**, isomorphe en tant qu'anneau intègre à **Z** ou à **Z**/p**Z** avec p premier; il est donc isomorphe, soit à **Q**, soit au corps **Z**/p**Z**.

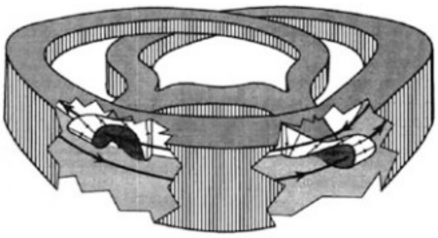

@ 4 B $ + F
. * * V F . ? W

3:
8 = F e r! * + . * ' + : .[* E I [+
(:+) ; ??B1

3: 7 . + <<<
A J * .+ + + + y + +. >'] + + *+* '+ !+ '+ *
. +S +* + + + *+*

! "#$%$ & ' ' "( ) " " . + (4 4 9 @
= + . +. 3 , ' 4 H *

B 1 ! "$/2" " "3 4 + " 6 . 6+6(4 46
6/+4 ' 6 + + + + = .*

! "#$%$ & ' ' "( ) " " .+ + * 8 = $
7 * 8 + / + ' =

B 1 ! "$/2" " "3 4 + " $ * +6#0 .*
A #+) 0 4=

B 1 ! "$/2" " "3 4 + " !+ F + % : 4 .
4 F + 4F + * ' 3 * : + 6 ; 1 .*

BB1 ! "$/2" " "3 4 + " !+ 6 8 .*
A $ 7 * 8 + / + ' = .+ + * 8 =
>4 ? ; 1 ! '+

>4 B ; 1 8 +' +

>4 ; ?1 ? 2 + @ * % = ! ' * .+

3: , ' * <<<

B 1 ! "$/2" " "3 4 + " ! 6!30! .*
A ! 4= % 3 4 + * 0' !
3: , ' * <<<

>4 ?; 1 ! +. ? 8 +' + + , 4

7 +. + * 4 + +.. + . *
4 + + ' = $ 4 = = + * + 4 +' + +
4 + * +.. + 4=F * + + + ' : * *
* Z 4 += + + = * G C. + 0 D =
? ? % = + 4 * 4=
F : 4 4 + +.. + 4 +' + + 4F : 4
+ + + =

B 1 ! "$/2" " "3 4 + " # * 6 ! .*
A * ! 4=
>4 ; 1 ! +. B 0 0E +

B 1 ! "$/2" " "3 4 + " 8 0 P = $
4 * 4 + + ' = + + + = + * $* 7 ; 1 .*
>4 B ! +. ! ' + * + =

Chapter 4

Approaching general relativity: introducing curvilinear coordinate systems

Cartesian coordinate systems have straight coordinate lines orthogonal to each other. The coordinate basis vector fields are constant, and all the vectors of each field are unit vectors with the same direction. This is the simplest coordinate system that one can imagine. It seems strange that it can be advantageous to introduce coordinate systems with coordinate curves which are not straight lines, so-called curvilinear coordinate systems. But in fact some objects of investigation have certain symmetries, for example cylindrical or spherical symmetry that makes it advantageous to introduce curvilinear coordinates. In a flat space, however, every coordinate system can be transformed into a Cartesian system, so that in principle one could solve every problem with reference to Cartesian coordinates.

In general relativity it is necessary to introduce curvilinear coordinate systems. According to this theory spacetime[1] is curved and four-dimensional with three space dimensions and one time dimension. This curved four-dimensional spacetime seems to be the easiest way of taking into account the results from careful experimental investigations of electromagnetism and gravitation.

In fact, already about 1920, the general theory of relativity was generalized to describe spacetime with more than three spatial dimensions. More than one time dimension is problematic, however, because it leads to certain strange possibilities of backward time travel, which belongs to science fiction. These generalized versions of general relativity are called Kaluza–Klein theories, and have been thoroughly investigated recently. They lead to the possibility of giving a unified geometrical description of classical (i.e. non-quantum) electromagnetism and gravity. The problem of creating a coherent single theory for gravitation and quantum theory is still unsolved. The most recent effort is called superstring theory and involves ten-dimensional spacetime. In this book we shall be concerned with four-dimensional spacetime.

Consider hypothetical two-dimensional creatures living in a two-dimensional space made up of the surface of a sphere. They would have to describe their world with reference to the curved surface. No three-dimensional space exists for these creatures. And the spherical surface *cannot* be covered by a two-dimensional Cartesian coordinate system. (Look at a globus!). The situation is similar in the case of curved spaces of higher dimensions, such as our spacetime. It cannot be covered by a Cartesian coordinate system. As long as we insist that the spacetime may be curved and four-dimensional, and that the description of our universe shall be referred to a coordinate system with four dimensions only, there is no other way than introducing curvilinear coordinate systems.

---

[1] See Sect. 1.7 and Ch. 5.

The simplest curvilinear coordinate system is a system of concentric circles and straight radial lines out from the common centre of the circles. The coordinates of this system are called *plane polar coordinates*. In this chapter we shall consider the main properties of such a coordinate system. Some knowledge of trigonometric functions is then indispensable.

4.1 Trigonometric functions

We shall now introduce three so-called *trigonometric* functions, named *sinus*, *cosinus*, and *tangens*. They are widely used in physics, meteorology and astronomy—everywhere.

Consider a circle with radius r and centre at the origin of a Cartesian coordinate system, as shown in Fig. 4.1. The radial line OP makes an angle θ with the positive x-axis. The point P has coordinates (x, y).

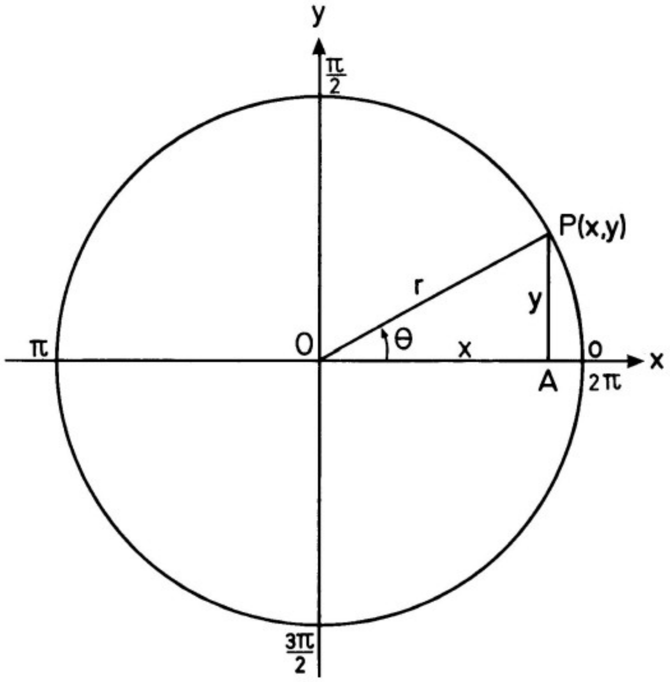

Figure 4.1: Circle of radius r centered at the origin

.

Before we define the functions sinus, cosinus and tangens, we shall have to know how we measure angles. There are two common angular measures. In daily life we measure angles in *degrees*. For example the latitude and longitude of the position of a mountain are given in degrees. One degree may be defined in the following way: divide a circle in 360 parts by marking 360 equidistant points. Draw radial lines from the centre of the circle to all of the dividing points. One degree is the angle between two neighbouring radial lines. Thus the angle around the full circle is 360 degrees. Since there are four right angles around a circle, a right angle is equal to 90 degrees, and a half circle is 180 degrees.

The other angular measure, which is used in connection with the trigonometric functions, is called a *radian*. With reference to Fig. 4.1 we formulate the following definition: the angle, θ, i.e. the *number of radians*, is defined as the length of the circular arc between 0 and P divided by the radius. Note that this ratio is independent of the radius of the circle. The sign π is the symbol for the number of radians of a 180 degree angle. Its value is $3.141592\ldots$. It follows that the length of a circle with radius r is $2\pi r$. The angle around a circle, as measured in radians, is 2π, and a right angle is equal to $2\pi/4 = \pi/2$. Increasing an angle by 2π corresponds to going one time around a circle.

The functions $\text{sinus}(\theta)$ and $\text{cosinus}(\theta)$ are denoted by $\sin\theta$ and $\cos\theta$. They are defined with reference to Fig. 4.1 by the equations

$$\sin\theta \equiv \frac{y}{r} \quad \text{and} \quad \cos\theta \equiv \frac{x}{r}, \tag{4.1a}$$

where $r = (x^2 + y^2)^{1/2}$ according to the Pythagorean theorem. Note that if $r = 1$ then $\sin\theta = y$ and $\cos\theta = x$. For θ outside the range $[0, 2\pi\rangle$ the definition implies

$$\sin(\theta + 2\pi) = \sin\theta \quad \text{and} \quad \cos(\theta + 2\pi) = \cos\theta. \tag{4.1b}$$

Equation (4.1b) shows that $\sin\theta$ and $\cos\theta$ are periodic functions with period 2π. This corresponds to the fact that if P on Fig. 4.1 is moved around the circle, so that θ passes 2π, then x and y, and thus $\sin\theta$ and $\cos\theta$ obtain just the same values as one round earlier.

The function tangens(θ) is denoted by $\tan\theta$, and is defined by

$$\tan\theta \equiv \frac{\sin\theta}{\cos\theta}. \tag{4.2}$$

Substituting for $\sin\theta$ and $\cos\theta$ from the definitions in Eqs. (4.1) leads to

$$\tan\theta = \frac{y}{x}. \tag{4.3}$$

The function $\tan\theta$ is in principle superfluous. We can always use $\sin\theta$ and $\cos\theta$. However, the function $\tan\theta$ simplifies some calculations, as will be seen in our description of plane curves in Sect. 9.1.

It will be useful to formulate the expressions for the trigonometric functions without reference to a Cartesian coordinate system; only referring to a right angled triangle, such as OAP in Fig. 4.1. Note that in relation to the angle θ the base x is the side adjacent to θ, and the height

y is the side opposite to θ. Using these terms the expressions for $\sin\theta$ and $\cos\theta$ can be formulated in the following way.

$$\sin\theta = \frac{\text{the length of the side opposite to }\theta}{\text{the length of the hypotenuse}},$$

$$\cos\theta = \frac{\text{the length of the side adjacent to }\theta}{\text{the length of the hypotenuse}},$$

and

$$\tan\theta = \frac{\text{the length of the side opposite to }\theta}{\text{the length of the side adjacent to }\theta}.$$

These formulations permit us to find expressions for the trigonometric functions by reference to a right angled triangle with arbitrary orientation, not referring to any coordinate system, see Fig. 4.2.

From Fig. 4.2 and the expressions (4.1) we get $\sin\theta = a/h$ and $\cos\theta = b/h$. Multiplying by h we can express the magnitude of the side opposite to θ and the side adjacent to θ, by θ and the hypotenuse h of the right angled triangle,

$$a = h\sin\theta \quad \text{and} \quad b = h\cos\theta. \tag{4.4}$$

Consider Fig. 4.3. The part CQD of the figure is obtained by turning the part APB an angle 90 degrees to the left. Thus the angles CQD and APB are equal. We obtain

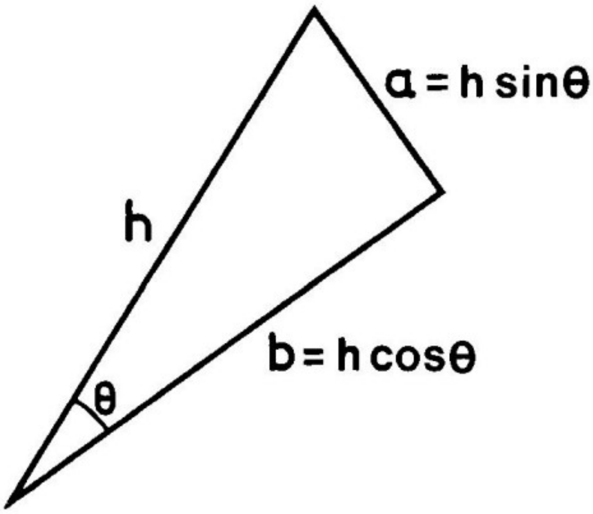
a = h sinθ
h
b = h cosθ
θ

Figure 4.2: A right-angled triangle

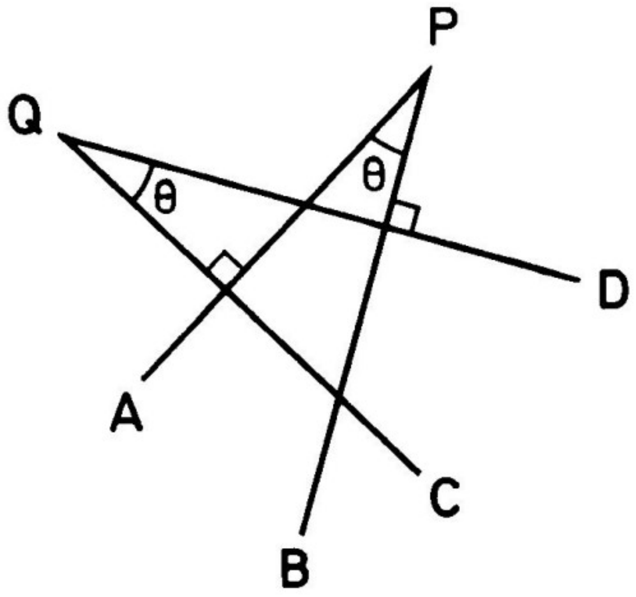

Figure 4.3: Angles with pairwise normal legs

the following sentence: two angles with pairwise normal legs, are equal.

In Eq. (1.8) we have written the dot product of two vectors $\vec{A}$ and $\vec{B}$ as 'the magnitude of $\vec{A}$ times the magnitude of $\vec{B}$'s projection, $\vec{B}_{\|}$, along $\vec{A}$'. This may now be expressed in terms of the angle α between $\vec{A}$ and $\vec{B}$.

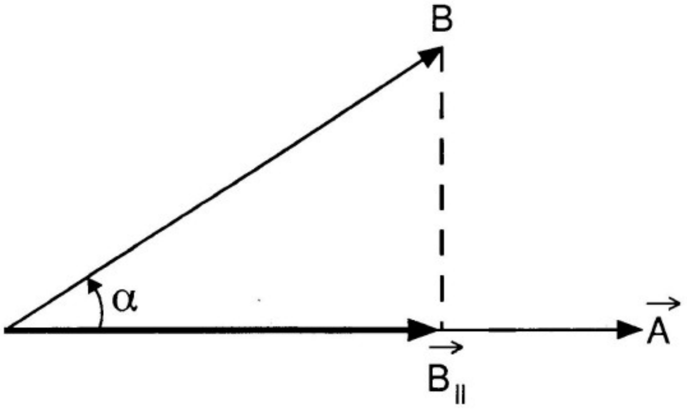

Figure 4.4: Angle α between $\vec{A}$ and $\vec{B}$

From Fig. 4.4 is seen that $|\vec{B}_{\|}| = \cos\alpha$. In general

$$|\vec{B}_{\|}| = \begin{cases} |\vec{B}|\cos\alpha & \text{for} \quad -\frac{\pi}{2} \leq \alpha < \frac{\pi}{2} \\ -|\vec{B}|\cos\alpha & \text{for} \quad \frac{\pi}{2} \leq \alpha < \frac{3\pi}{2}. \end{cases} \tag{4.5}$$

Inserting this into Eq. (1.8) gives

$$\vec{A} \cdot \vec{B} = |\vec{A}||\vec{B}|\cos\alpha \tag{4.6}$$

where the sign ambiguity in Eq. (1.8) is taken care of automatically. Since the angle that a vector makes with itself is zero, and $\cos 0 = 1$, it follows that the square of the magnitude of a vector is

$$|\vec{A}|^2 = \vec{A} \cdot \vec{A}. \tag{4.7}$$

We shall now deduce a geometrical identity that will be used in Ch. 5. The square of $|\vec{A} - \vec{B}|$ is

$$\begin{aligned} |\vec{A} - \vec{B}|^2 &= (\vec{A} - \vec{B}) \cdot (\vec{A} - \vec{B}) = \vec{A} \cdot \vec{A} - 2\vec{A} \cdot \vec{B} + \vec{B} \cdot \vec{B} \\ &= |\vec{A}|^2 + |\vec{B}|^2 - 2\vec{A} \cdot \vec{B}. \end{aligned} \tag{4.8}$$

Substituting the expression (4.6) for the dot product $\vec{A} \cdot \vec{B}$ leads to

$$|\vec{A} - \vec{B}|^2 = |\vec{A}|^2 + |\vec{B}|^2 - 2|\vec{A}||\vec{B}| \cos \alpha. \tag{4.9}$$

Look at Fig. 1.9. Let the angle between the vectors $\vec{A}$ and $\vec{B}$ be α, the length of $\vec{A}$ be L_1, the length of $\vec{B}$ be L_2, and the length of $\vec{A} - \vec{B}$ be L. Then Eq. (4.9) may be written

$$L^2 = L_1^2 + L_2^2 - 2L_1L_2 \cos \alpha. \tag{4.10}$$

This is an extension of the Pythagorean theorem to the case when the triangle has no 90 degree angle. It might have been named 'The extended Pythagorean theorem', but is usually called 'the law of cosines'. In Fig. 4.5 we have redrawn Fig. 1.9, but this time with the quantities L, L_1, L_2, and α which appear in Eq. (4.10).

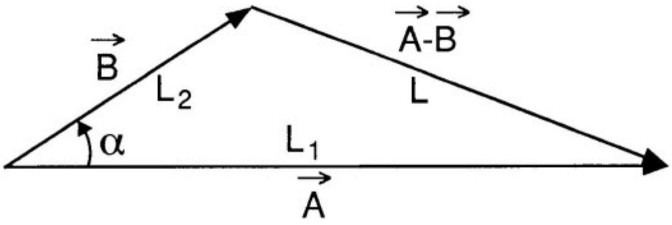

Figure 4.5: The triangle of Eq. (4.10)

The magnitudes of $\sin\theta$ and $\cos\theta$ when θ increases from 0 to 2π are seen from Fig.4.1 by letting P start at the positive x-axis and then move around the circle. At the x-axis, $x = r$ and $y = 0$, giving and $\cos 0 = 1$. Moving on to the y-axis, x decreases towards zero and y increases towards r. When P is at the y-axis the angle $\theta$is equal to $\pi/2$. Thus $\sin(\pi/2) = 1$ and $\cos(\pi/2) = 0$. As P comes to the negative x-axis, $x = -r$ and $y = 0$, while $\theta = \pi$, so $\sin\pi = 0$ and $\cos\pi = -1$. Letting P move further to the negative y-axis, the angle θ reaches $3\pi/2$, while $x = 0$ and $y = -r$, giving $\sin(3\pi/2) = -1$ and $\cos(3\pi/2) = 0$. We may now draw the diagrams representing $y = \sin\theta$ and $y = \cos\theta$ (in Fig. 4.6), with y along a vertical axis and θ along a horizontal axis.

An important relation between $\sin\theta$ and $\cos\theta$ should

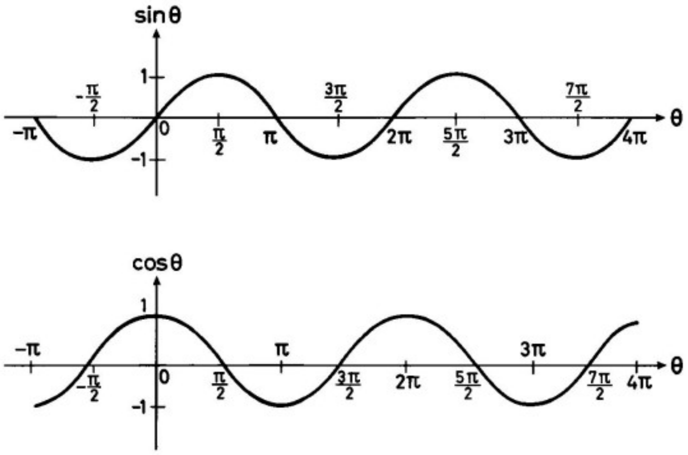

Figure 4.6: The function $\sin\theta$ and $\cos\theta$

be noted. Using Eqs. (4.1) we get

$$\sin^2\theta + \cos^2\theta = \frac{x^2}{r^2} + \frac{y^2}{r^2} = \frac{x^2+y^2}{r^2} = \frac{r^2}{r^2} = 1. \qquad (4.11)$$

This equation is valid for every value of the angle θ.

We shall need to calculate cosinus to the sum (or difference) of two angles. The formulae both for sinus and cosinus to a sum of two angles are found from Fig. 4.7. You will find a lot of information written on the figure. These are results that we shall now deduce. Consider first the angle DPC. PC is normal to OC and PD is normal to OA. According to Fig. 4.3 two angles with pairwise normal legs are equal. Hence, we conclude that the angle DPC is equal to the angle POC, i.e. it is equal to u.

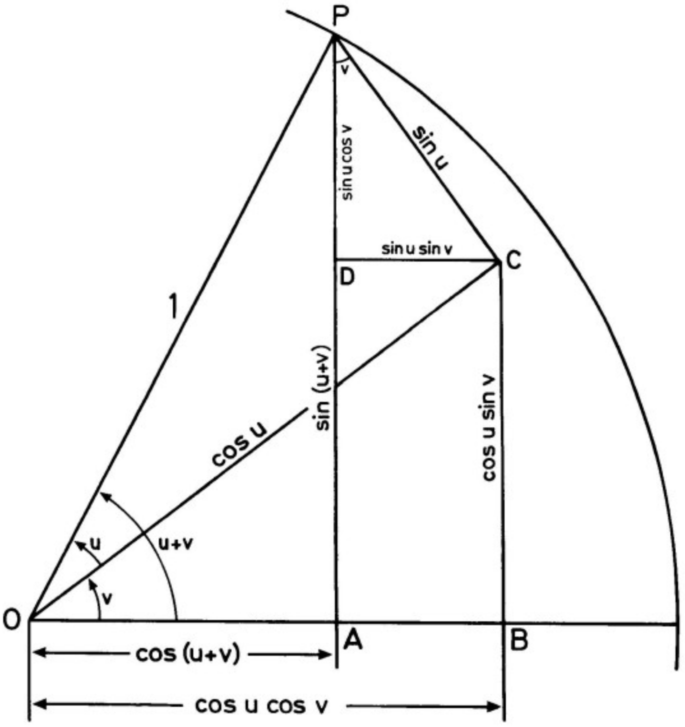
P
v
sin u cos v
sin u
sin u sin v
C
D
1
sin (u+v)
cos u sin v
cos u
u
u+v
v
O
A
B
cos (u+v)
cos u cos v

Figure 4.7: Sum of angles

Next we shall utilize Eq. (4.4) which is related to Fig. 4.2. Consider the triangle OCP which is similar to the triangle of Fig. 4.2. Using Eq. (4.4) we find

$$\mathrm{OC} = \cos u \quad \text{and} \quad \mathrm{PC} = \sin u. \tag{4.12}$$

Looking at the triangle DCP we find

$$\begin{aligned} \mathrm{DC} &= \mathrm{PC} \sin v = \sin u \sin v, \\ \mathrm{PD} &= \mathrm{PC} \cos v = \sin u \cos v. \end{aligned} \tag{4.13}$$

From the triangle OBC we get

$$\begin{aligned} \mathrm{OB} &= \mathrm{OC} \cos v = \cos u \cos v, \\ \mathrm{BC} &= \mathrm{OC} \sin v = \cos u \sin v, \end{aligned} \tag{4.14}$$

and from the triangle OAP we have

$$\mathrm{OA} = \cos(u+v) \quad \text{and} \quad \mathrm{AP} = \sin(u+v). \tag{4.15}$$

Looking at Fig. 4.7 we now find

$$\sin(u+v) = \mathrm{PD} + \mathrm{AD} = \mathrm{PD} + \mathrm{BC}. \tag{4.16}$$

Inserting the expressions for PD and BC from Eqs. (4.13) and (4.14), respectively, we get

$$\sin(u+v) = \sin u \, \cos v + \cos u \, \sin v \tag{4.17}$$

and

$$\cos(u+v) = \mathrm{OB} - \mathrm{AB} = \mathrm{OB} - \mathrm{DC}. \tag{4.18}$$

Inserting the expressions for OB and DC from Eqs. (4.14) and (4.13), respectively, leads to

$$\cos(u+v) = \cos u\,\cos v - \sin u\,\sin v. \tag{4.19}$$

We shall now deduce a formula which will be needed in chapter 14. Using Eq. (4.17) we get

$$\begin{aligned}\sin 2\theta &= \sin(\theta+\theta)\\ &= \sin\theta\,\cos\theta + \cos\theta\,\sin\theta = 2\sin\theta\,\cos\theta.\end{aligned} \tag{4.20}$$

From Eq. (4.19) we have

$$\cos 2\theta = \cos(\theta+\theta) = \cos^2\theta - \sin^2\theta.$$

Using Eq. (4.11) this may be written alternatively as

$$\cos 2\theta = 2\cos^2\theta - 1 = 1 - 2\sin^2\theta.$$

4.1.1 Differentiation of trigonometric functions

In our analysis of curved coordinate systems we shall have to differentiate the functions $\sin\theta$ and $\cos\theta$. A geometrical deduction of the expressions for the derivatives of these functions shall now be given.

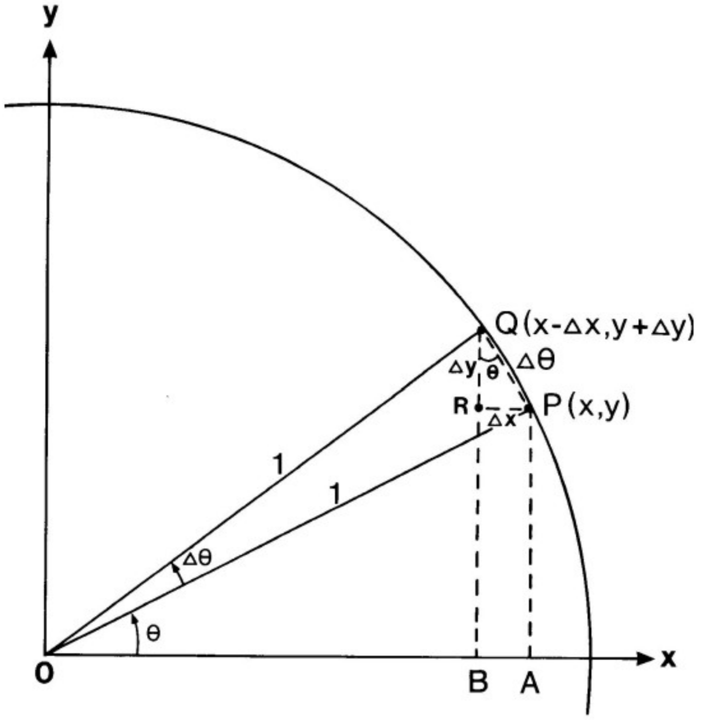
y
Q(x-Δx,y+Δy)
Δθ
Δy
θ
R
Δx
P(x,y)
1
1
Δθ
θ
O
B
A
x

Figure 4.8: Changes in position on the unitunit!circle circle when incrementing θ by Δθ

Consider Fig. 4.8. A circle with radius 1 is drawn about the origin O of a Cartesian coordinate system. The radii OP and OQ make angles θ and $\theta + \Delta\theta$, respectively, with the positive x axis. The point P has coordinates (x, y) and the point Q $(x - \Delta x, y + \Delta y)$. For a small angle $\Delta\theta$, the direction of the line PQ is very close to that of the tangent at P. Then PQ is normal to the radial line OP. Thus the legs PQ and QR of the angle PQR are pairwise normal to the legs OP and OA. Since the angle POA is equal to θ, the angle PQR must also be equal to θ. From the triangle PQR we then have

$$\sin\theta = \frac{\Delta x}{\Delta\theta} \quad \text{and} \quad \cos\theta = \frac{\Delta y}{\Delta\theta}. \tag{4.21}$$

Note that the line PQ equals $\Delta\theta$ because the radius is equal to 1. Also from the triangles OAP and OBQ we get

$$\sin\theta = y \quad \text{and} \quad \cos\theta = x, \tag{4.22}$$

and

$$\sin(\theta + \Delta\theta) = y + \Delta y \quad \text{and} \quad \cos(\theta + \Delta\theta) = x - \Delta x. \tag{4.23}$$

The expressions for the derivatives of $\sin\theta$ and $\cos\theta$ can now be deduced directly from the general definition of the derivative of a function, Eq. (2.4).

$$\begin{aligned}(\sin\theta)' &= \lim_{\Delta\theta\to 0}\frac{\sin(\theta+\Delta\theta)-\sin\theta}{\Delta\theta} = \lim_{\Delta\theta\to 0}\frac{y+\Delta y-y}{\Delta\theta}\\ &= \lim_{\Delta\theta\to 0}\frac{\Delta y}{\Delta\theta} = \cos\theta\end{aligned} \tag{4.24}$$

and

$$\begin{aligned}(\cos\theta)' &= \lim_{\Delta\theta\to 0}\frac{\cos(\theta+\Delta\theta)-\cos\theta}{\Delta\theta} = \lim_{\Delta\theta\to 0}\frac{x-\Delta x-x}{\Delta\theta}\\ &= \lim_{\Delta\theta\to 0}\left(-\frac{\Delta x}{\Delta\theta}\right) = -\sin\theta.\end{aligned} \tag{4.25}$$

where we have used Eqs. (4.21)–(4.23).

In order to calculate, in chapter 9, the curvature of a curve, we shall need the derivative of $\tan\theta$. This is most simply deduced by applying the formula (2.46) for differentiating a fraction of functions. Inserting $u = \sin\theta$ and $v = \cos\theta$ and using Eqs. (4.2),(4.24), and (4.25) we find

$$\begin{aligned}(\tan\theta)' &= \left(\frac{\sin\theta}{\cos\theta}\right)' = \frac{\cos\theta\cos\theta-\sin\theta\,(-\sin\theta)}{\cos^2\theta}\\ &= \frac{\cos^2\theta+\sin^2\theta}{\cos^2\theta} = \frac{1}{\cos^2\theta}.\end{aligned} \tag{4.26}$$

Alternatively, from the next last expression in Eq. (4.26), we get

$$(\tan\theta)' = 1 + \frac{\sin^2}{\cos^2\theta} = 1 + \tan^2\theta, \qquad (4.27)$$

where we have used the definition of $\tan\theta$ in Eq. (4.2).

In Sect. 9.2 we shall need to know the MacLaurin series for $\sin\theta$. In order to collect the necessary theory of trigonometry at one place, we shall immediately deduce the form of this series.

The MacLaurin series of a function $f(\theta)$ is given in Eq. (2.84) as

$$f(\theta) = f(0) + f'(0)\,\theta + \frac{f''(0)}{2}\,\theta^2 + \frac{f'''(0)}{6}\,\theta^3 + \cdots . \qquad (4.28)$$

In the case that $f(\theta) = \sin\theta$ we get

$$\begin{array}{lcl} f(\theta) = \sin\theta & \Rightarrow & f(0) = 0 \\ f'(\theta) = \cos\theta & \Rightarrow & f'(0) = 1 \\ f''(\theta) = -\sin\theta & \Rightarrow & f''(0) = 0 \\ f'''(\theta) = -\cos\theta & \Rightarrow & f'''(0) = -1. \end{array}$$

Substitution into Eq. (4.28) gives

$$\sin\theta = \theta - \frac{1}{6}\theta^3 + \cdots . \qquad (4.29)$$

These are the first terms of the MacLaurin series for $\sin\theta$.

4.2 Plane polar coordinates

A coordinate system with plane polar coordinates, r and θ, is shown in Fig. 4.9.

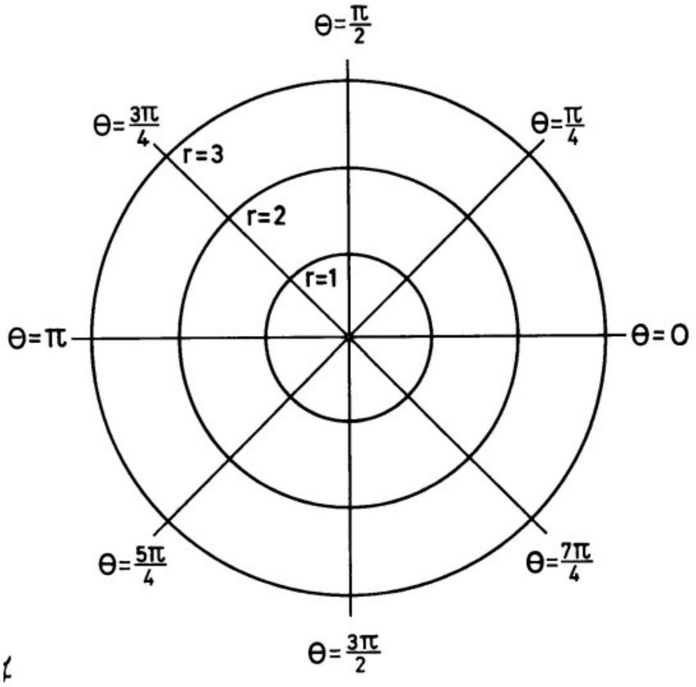
$\theta=\frac{\pi}{2}$
$\theta=\frac{3\pi}{4}$
$\theta=\frac{\pi}{4}$
$r=3$
$r=2$
$r=1$
$\theta=\pi$
$\theta=0$
$\theta=\frac{5\pi}{4}$
$\theta=\frac{7\pi}{4}$
$\theta=\frac{3\pi}{2}$

Figure 4.9: Plane polar coordinates, r and θ

Instead of vertical and horizontal coordinate lines, such as in a Cartesian coordinate system, the coordinate curves in a system with plane polar coordinates are concentric circles and radial lines. The Cartesian coordinates

x and y are the horizontal and vertical distances from the origin. The plane polar coordinate r of an arbitrary point P, is the distance from the origin to P. The plane polar coordinate θ is the angle of the radial line from the origin to P with respect to a horizontal line.

Just as there are horizontal and a vertical basis vectors, $\vec{e}_x$ and $\vec{e}_y$, at every point in a Cartesian coordinate system, there are radial, and a tangential basis vectors, $\vec{e}_r$ and $\vec{e}_\theta$, at every point in a polar coordinate system. These basis vectors at an arbitrary point P are shown in Fig. 4.10. The point O is the origin of the coordinate system. The point P has coordinates r and θ. The basis vectors at P have been denoted by $(\vec{e}_r)_P$ and $(\vec{e}_\theta)_P$ to indicate that their di-

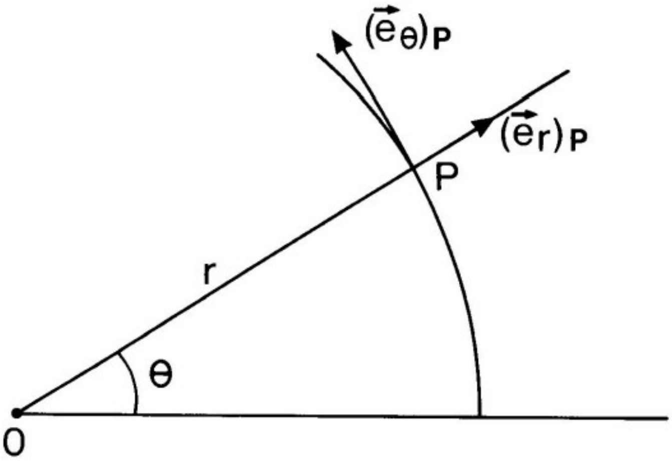

Figure 4.10: Radial and tangential basis vectors

rections and magnitudes depend upon the point at which they are drawn. The Cartesian basis vectors, $\vec{e}_x$ and $\vec{e}_y$, on the other hand, are the same everywhere.

The simple relation between a Cartesian coordinate system and a polar coordinate system with the same origin, and with the direction $\theta = 0$ corresponding to the x-axis, is shown in Fig. 4.11.

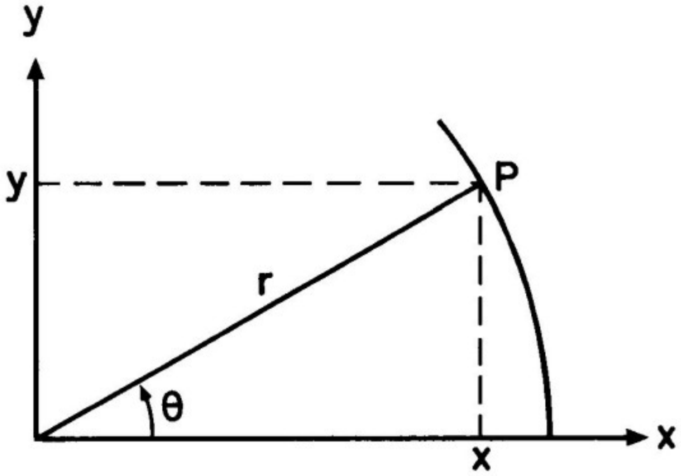

Figure 4.11: Relation between Cartesian and polar coordinates

From the figure is seen that

$$\cos\theta = \frac{x}{r} \quad \text{and} \quad \sin\theta = \frac{y}{r} \tag{4.30}$$

which gives

$$x = r\cos\theta \quad \text{and} \quad y = r\sin\theta. \tag{4.31}$$

This is our first example of a coordinate transformation equation.

Some curves have a simpler equation in polar coordinates than in Cartesian coordinates. The equation of a circle with centre at the origin and radius R is in Cartesian coordinates

$$x^2 + y^2 = R^2. \tag{4.32}$$

The equation of the same circle in polar coordinates is

$$r = R. \tag{4.33}$$

The position vector of an arbitrary point P with Cartesian coordinates (x, y) is

$$\vec{r} = x\vec{e}_x + y\vec{e}_y. \tag{4.34}$$

The Cartesian components of $\vec{r}$ are $r^x = x$ and $r^y = y$, and may be expressed by the coordinates r and θ by means of Eq. (4.31), giving

$$\vec{r} = r\cos\theta\,\vec{e}_x + r\sin\theta\,\vec{e}_y. \tag{4.35}$$

We now calculate the magnitude of the basis vectors $\vec{e}_r$ and $\vec{e}_\theta$. Using a concept introduced in section 3.3 we may say that the basis vector field $\vec{e}_r$ is a tangent vector field to the radial lines. The coordinate θ is constant along a radial line, while r increases along it. Thus r can be used as a parameter for such a line. The parametric equation for a radial line is then

$$x = r\cos\theta \quad \text{and} \quad y = r\sin\theta$$

where θ is a constant, specific for each line. Equation (3.19) with $\vec{u} = \vec{e}_r$, $t = r$, and replacing dx/dt with $\partial x/\partial r$, and dy/dt with $\partial y/\partial r$, gives

$$\vec{e}_r = \frac{\partial x}{\partial r}\vec{e}_x + \frac{\partial y}{\partial r}\vec{e}_y, \tag{4.36}$$

which shows that

$$(\vec{e}_r)^x = \frac{\partial x}{\partial r} \quad \text{and} \quad (\vec{e}_r)^y = \frac{\partial y}{\partial r}. \tag{4.37}$$

Differentiating Eq. (4.31) with respect to r, (remember that θ is constant when we differentiate partially with respect to r), we get

$$\frac{\partial x}{\partial r} = \cos\theta \quad \text{and} \quad \frac{\partial y}{\partial r} = \sin\theta. \tag{4.38}$$

Thus

$$(\vec{e}_r)^x = \cos\theta \quad \text{and} \quad (\vec{e}_r)^y = \sin\theta \tag{4.39}$$

showing that

$$\vec{e}_r = \cos\theta\,\vec{e}_x + \sin\theta\,\vec{e}_y. \tag{4.40}$$

The magnitude of $\vec{e}_r$, as calculated by means of Eqs. (1.5) and (4.39) is

$$|\vec{e}_r| = \cos^2\theta + \sin^2\theta = 1, \tag{4.41}$$

which shows that $\vec{e}_r$ is an unit vector independent of its position. But $\vec{e}_r$ changes with θ. Since the magnitude of $\vec{e}_r$ is constant and equal to unity, the change of $\vec{e}_r$ with θ is only a change of direction.

The basis vector fields $\vec{e}_\theta$ are tangent vector fields to the circles about the origin. The coordinate r is constant along a circle, while θ increases along it. So θ can be used as a parameter for such a circle. The parametric equation for a circle about the origin is again Eq. (4.31), but this time with $r =$ constant for each circle, and with θ as variable.

Equation (3.17) with $\vec{u} = \vec{e}_\theta$, $t = \theta$, and first $i = x$, so $i = y$, then gives

$$(\vec{e}_\theta)^x = \frac{\partial x}{\partial \theta} = -r\sin\theta \quad \text{and} \quad (\vec{e}_\theta)^y = \frac{\partial y}{\partial \theta} = r\cos\theta, \tag{4.42}$$

which leads to

$$\vec{e}_\theta = -r\sin\theta\,\vec{e}_x + r\cos\theta\,\vec{e}_y. \tag{4.43}$$

The magnitude of $\vec{e}_\theta$ is found from Eqs. (1.5) and (4.43) as

$$\begin{aligned} |\vec{e}_\theta| &= \left[(-r\sin\theta)^2 + (r\cos\theta)^2\right]^{1/2} \\ &= r(\sin^2\theta + \cos^2\theta)^{1/2} = r. \end{aligned} \tag{4.44}$$

This equation shows that the magnitude of $\vec{e}_\theta$ increases proportionally to the distance from the origin. This is not a unit vector field. However, one could define a corresponding unit vector field $\vec{e}_{\hat{\theta}}$ by

$$\vec{e}_{\hat{\theta}} = \frac{\vec{e}_\theta}{|\vec{e}_\theta|} = \frac{1}{r}\vec{e}_\theta. \tag{4.45}$$

Here the hat above the index denotes that the vector has unit length.

Note that the calculation above has provided us with a method for transforming basis vectors. Using Eq. (3.19) we find the transformation of the basis vectors from a knowledge of the corresponding coordinate transformation.

We have seen that in general the coordinate basis vectors are not unit vectors. And the corresponding unit vectors, such as $\vec{e}_{\hat{\theta}}$, will in general not be coordinate basis vectors. Still they are basis vectors, in the sense that an arbitrary vector can be decomposed along the basis vectors.

Chapter 5

The metric tensor

The metric tensor is perhaps the most important mathematical quantity in the theory of relativity. From a knowledge of the metric tensor one may compute the geometry of spacetime and for example the motion of the planets in the solar system. In this chapter we shall give a thorough introduction to the metric tensor and its physical significance in the theory of relativity.

5.1 Basis vectors and the dimension of a space

In Chapter 1 basis vectors were introduced as reference vectors for directions (in flat space), and it was mentioned that at every point of an n-dimensional space there are n basis vectors. One could say that there are n independent

B?1 ! "$/2" " "3 4 + " 9+ ! . C + 4 +
.+ . + + 4 !+ ; ? 1 .*

1 ! "$/2" " "3 4 + " 9+ 8 +' = $ 7 *
0 D 8 + / + ' = : $** H = .*
A = + 4 =

1 ! "$/2" " "3 4 + " 9+ 8 +' = $ 7 *
0 P 8 + / + ' = ; 1 *S'

1 ! "$/2" " "3 4 + " 9+ 8 +' = $ 7 *
0 P 8 + / + ' = + 0* + , * ; 1 .*
3: , ' * <<<
A 8 +' =

1 ! "$/2" " "3 4 + " 9+ 6708/ .*

1 ! "$/2" " "3 4 + " K :6% 60J0 .*
A % 0+ = J '

B1 ! "$/2" " "3 4 + " , ** ,= 6!7, .*
A ! 4 + 7 + + * ,+ 4 +

1 ! "$/2" " "3 4 + " 68 +' + .*
A .+4

1 ! "$/2" " "3 4 + ") + '6 @! .*
A = + @ *+ ! 4=
3: .+4+ 4 + * U +4 ;B 1

1 ! "$/2" " "3 4 + " |=' * 8 } 0 P 8 +
% = / + ' = H * $.. + ! 4= . 4 +4 & O)
; 1 .*

?1 ! "$/2" " "3 4 + " |=' * 8 } 0 P 4 +
= + ' = . 4 ; 1 .*

1 ! "$/2" " "3 4 + " |=' * 8 } 0 P = +
4 * 4 + + ' = + + + = + * $* 7 ; 1 .*

1 ! "$/2" " "3 4 + " |=' * 8 } 7 *
0 D % = / + ' = . 4 ; 1 .*

1 ! "$/2" " "3 4 + " |=' * 8 } % % =
! . C $ 4 + + 8 : ' , % + = !+ : *4 J ' =
; 1 .*

1 ! "$/2" " "3 4 + " |=' * 8 } ;+ 1 ,
& 8 + % = / + ' = @ & P $ + ' 8 +' = / . '
8 +' = + 0 4= . 4 +4 & O) ; ?1 .*

1 ! "$/2" " "3 4 + " +)6! .*

A ! 4 + =

B1 ! "$/2" " "3 4 + " : 6, J .*
A % ,+ 4 + J '

1 ! "$/2" " "3 4 + " : 6 ! .*
A . = + ! 4=

1 ! "$/2" " "3 4 + " U6@!8/ .*
A $ @ ! 8 + / + ' =

1 ! "$/2" " "3 4 + " + !+ .+ + *
8 = $ 7 * 8 + / + ' = + 0* + , * ; 1 .*

?1 ! "$/2" " "3 4 + " + !+ .+ + *
8 = $ 7 * 8 + / + ' = !+ : *4 J ' = ; ?1 .*

1 ! "$/2" " "3 4 + " H+ *68/ .*
A 8 + / + ' =

1 ! "$/2" " "3 4 + " H : 46! 4= .*

1 ! "$/2" " "3 4 + " H : 468! .*
A 8 +' + + * ! 4=

! "#$%$ & ' ' "( ) " "% $: ' 4

1 ! "$/2" " "3 4 + " / + * / U=) $:
' 4 + $: ' 4 ; ?1 .*

1 ! "$/2" " "3 4 + " / + * / U=) 7 *
8 = $; 1 .*

B1 ! "$/2" " "3 4 + " % $: ' 4F
% (+ $ 7 . + * ; 1 .*

1 ! "$/2" " "3 4 + " % $: ' 4F
+ * (= * $ 7 . + * ; 1 .*

! "#$%$ & ' ' "( ) " "% 3C * 9+ *:) 0+ = *
0 . + 9 =F B B

1 ! "$/2" " "3 4 + " ? ;3C * 9+ *:) 1 9+
% 3C * 9+ *:) 0+ = * 0 . + 9 =F B B 6 776 ! + *
3C * J ' = ; B1 .*

1 ! "$/2" " "3 4 + " ? ;3C * 9+ *:) 1 9+
; * 1 % 3C * 9+ *:) 0+ = * 0 . + 9 =F B B 6 76 .
+ * + 3C * J ' = ; B1 .*

?

! "#$%$ & ' ' "( ) "

! "#$%$ & ' ' "( ) " "$*'+ * % + + * & +
0 +4

?1 ! "$/2" " "3 4 + " $*'+ * % + + *
& + 0 +4 .*
>4 ; 1 ! +. 0 +4 H+'

? 1 ! "$/2" " "3 4 + " $*'+ * % + + *
& + 0 +4 .*

! "#$%$ & ' ' "( ) " "!>

? 1 ! "$/2" " "3 4 + " !+ ' + $** H =
7: + + + ; ?? 1 .*
>4 B ;B 1 % + * 8+

? 1 ! "$/2" " "3 4 + " 9 + # 4 0* + * !+
! 8 + $ + + 9+ ; ?? 1 *S'

? 1 ! "$/2" " "3 4 + " 9 + # 4 0* + * !+
! 8 + $ + + 9+ ; ?? 1 .*
3: , ' * <<<

? 1 ! "$/2" " "3 4 + " / + * ! + # + + *
7 4 + !+ F 7F * 0* 7 : ; ? 1 .*

?B1 ! "$/2" " "3 4 + " / + * ! + # + + *
7 4 + !+ F 77 (+) v , * ; ? 1 .*

? 1 ! "$/2" " "3 4 + " / + * ! + 7 *
!+ + * $ + = 7 . 4 +4 & O) ; ? ?1 .*
>4 ; B1 ! +. $.. + = + * 8 =
3: % = + ! '
A + + / . +

? 1 ! "$/2" " "3 4 + " / + * ! + 7 *
!+ + * $ + = 77 . 4 & O) ; ? ?1 .*
>4 ; 1 + * + + * .+
+ E ! ' 4
$ * * .+ '+ ;CF =1 + : . * 4 + =
:= . +' 4 C + * = + * + !+ +
* + = % * + : . T ;CF =1 + * P T
;CPF =P1 4 ' := +

$$\overline{PP'} = \sqrt{(x' - x)^2 + (y' - y)^2},$$

:+ * + 4 = H * +
* 4 : * + .

? 1 ! "$/2" " "3 4 + " / + * ! + 3 E [
+ > + J + +: *+4 + * [* ! ~ + * + , *+ ; 1 .*

??1 ! "$/2" " "3 4 + " / + * ! + F # 9 :
* + + + = H = !9 ; ? ?1 *S'

1 ! "$/2" " "3 4 + " / + * ! + F # 9 :
* + + + = H = !9 ; ? ?1 .*

1 ! "$/2" " "3 4 + " / + * ! + F # 9 :
* + + + . = 7 : ; ? ?1 *S'

1 ! "$/2" " "3 4 + " / + * ! + F # 9 :
* + + + . = 7 : ; ? ?1 .*

1 ! "$/2" " "3 4 + " H * K+. + !> > 4 : +
+ ' .*

1 ! "$/2" " "3 4 + " H * K+. + !> > 4 : +
+ ' .*

B1 ! "$/2" " "3 4 + " H * K+. + !> > 4 : +
+ ' .*

1 ! "$/2" " "3 4 + " H * K+. + !> > 4 : +
+ ' .*

! "#$%$ & ' ' "( ) " "! =F * 0*

1 ! "$/2" " "3 4 + " ! =6 *60* .*

1 ! "$/2" " "3 4 + " 9 ! = !J
; 1 .*

?1 ! "$/2" " "3 4 + " & : 8+ 8 = + *
! = % + 4 + = $ + ; 1 .*

! "#$%$ & ' ' "( ) " "0 0 * 9 + $
8) 8 = % C +. * H *

1 ! "$/2" " "3 4 + "
0 6 60 *69 6+ 6$ 68)68 = .*

! "#$%$ & ' ' "( ) " "@ R 7 =

1 ! "$/2" " "3 4 + " #+ 4F #+' * Q % ' +
:) + + +: + +*+: + R P .+ +* C H = ; 1 .*

1 ! "$/2" " "3 4 + " #+' * Q #+ 4 % J ' +
( ) + + @ $: + +*+: + R P + +* C % : ! ; 1 .*
>4 ; 1

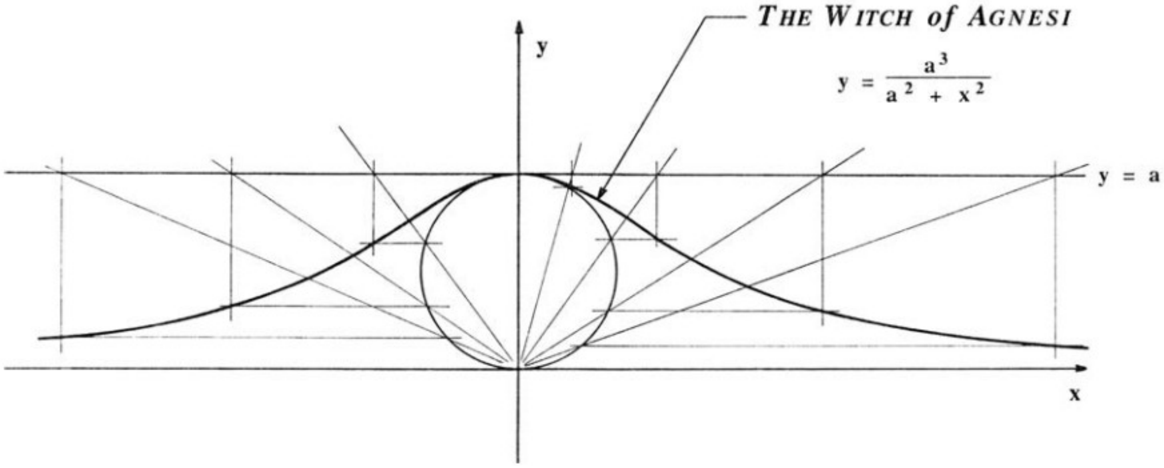
THE WITCH of AGNESI
$y = \frac{a^3}{a^2 + x^2}$
y
y = a
x

$4 F + + 8+ + + % H $4 ' Q 9 , + *

1 ! "$/2" " "3 4 + " 8 4 7 + @ 3 R
% J ' + 9 = ! . 4) 4 4 7 ; ? B1 *S'
A @ 3 R

1 ! "$/2" " "3 4 + " 8 4 7 + @ 3 R
% J ' + 9 = ! . 4) 4 4 7 ; ? B1 .*

B1 ! "$/2" " "3 4 + " 8 4 7 + % J ' +
9 = ! . 4 @ $:+ 2 + ! . H = ; 1 *S'

1 ! "$/2" " "3 4 + " 8 4 7 + % J ' +
9 = & : @ = 7 ' ! . H = ; 1 *S'

1 ! "$/2" " "3 4 + " 8 4 7 + % J ' +
9 = & : @ = 7 ' ! . H = ; 1 .*

1 ! "$/2" " "3 4 + " 8 4 7 + F , (+ @
3 R 6 $ J ' + 9 = & : 4 ( ) ; ? 1 .*

?1 ! "$/2" " "3 4 + " Q # (+ % :)
4 '+ F ' * F + * + * + +: 4 ' + ( )
; 1 .*

1 ! "$/2" " "3 4 + " Q # (+ % :)
4 '+ F ' * F + * + * + +: 4 ' +4
; 1 :

1 ! "$/2" " "3 4 + " Q # (+ % ( )
& 4 + F * F + * ,+ 7* + +: 3 4 J ' +4
; 1 .*

1 ! "$/2" " "3 4 + " Q # (+)=
! 4F) 4F + * : 4 4 ; ?? 1 *S'

1 ! "$/2" " "3 4 + " Q # (+ 7)=
! 4F %) 4F + * (4 (+) (+= ( ) ; ?? 1 .*

1 ! "$/2" " "3 4 + " K Q . % 9
R ; ? 1 .*

B1 ! "$/2" " "3 4 + " += (K 4 2 U
. 4= 4 4 . . P 4=P ' $*' J * ; 1 .*

1 ! "$/2" " "3 4 + " / + * R 7 =
% @ *+ = ;H 1 H * : 4 ! .+ = ; 1 .*
A .+4
3: , ' * <<<

>4 ; ?1 ! +. 8 +' = + * 7 +

1 ! "$/2" " "3 4 + " / + * R 7 = %
@ *+ = H * ; 1 .*

1 ! "$/2" " "3 4 + " 4 = 4 & 4 %
R + 4 +' + + JK ; ? 1 .*

?1 ! "$/2" " "3 4 + " + R ' & 4 .*

1 ! "$/2" " "3 4 + " + + 4 4 + * . U
+ 4 !/! ; ?1 .*

1 ! "$/2" " "3 4 + " % ! . ( ) .+ 4
@ $. R 8 +' = H = ; 1 .*

1 ! "$/2" " "3 4 + " '* . :6 4 + +
+ + = U .*

1 ! "$/2" " "3 4 + " H + 8 : % #
0 4 /+ * 9 : 4 8 . ; 1 .*

1 ! "$/2" " "3 4 + " H + 8 : R 9 =
4 ; 1 .*

B1 ! "$/2" " "3 4 + " H + 8 : F (4
% # 0 4 (+ + ( ) .*

1 ! "$/2" " "3 4 + " H + 8 : F (4
% # 0 4 (+ + ( ) ; ?? 1 :

1 ! "$/2" " "3 4 + " R : 4 +. = +
*+ 4 * + 4 ; 1 .*

1 ! "$/2" " "3 4 + "
R 6% 6(4 +. =6 6+6#+ 4 67* + .*

3: @+U * +. *+* * +*+ * U # U F : + '+U +* E + *
+ + . * + * +

! "#$%$ & ' ' "( ) " "8 =

?1 ! "$/2" " "3 4 + " B) .* 4 = +*' .*

1 ! "$/2" " "3 4 + " B) .* =. 4 = .*

! "#$%$ & ' ' "( ) " ", 4 + @

1 ! "$/2" " "3 4 + " * : S 6 60 4 + +S 6
•0# * 6 *+ .*

1 ! "$/2" " "3 4 + " +) + : = , 4 + @
$ 7 * . + , 4 (+) ; 1 .*

1 ! "$/2" " "3 4 + " 6 .*
A ,j4 +F 3 4 + 0. 4 +

! "#$%$ & ' ' "( ) " " + + + + + 0 + % +

1 ! "$/2" " "3 4 + " 67 + 6! .*

B1 ! "$/2" " "3 4 + " 6! + =
. .*

! "#$%$ & ' ' "( ) " "& + 3. + * F
0*

1 ! "$/2" " "3 4 + "
& + 63. 6+ *6 .*
A .+4

! "#$%$ & ' ' "( ) " "H+' + * 3.

1 ! "$/2" " "3 4 + " ? + + * H+' 0
3. + * .= $ +* ; 1 .*

1 ! "$/2" " "3 4 + " ? /+= F H+' + * $
. * *+ + * 4 4 4 . + * +.. * . .*

?1 ! "$/2" " "3 4 + " ?
H+' 6+ *63. 69+ 6 + + + + = .*
A B .+4

! "#$%$ & ' ' "( ) "

! "#$%$ & ' ' "( ) " "! = % + &+
+ ' 0*

B 1 ! "$/2" " "3 4 + " ? ! = .*
A : : 4

B 1 ! "$/2" " "3 4 + " ? ! = % + &+

+ + * ! + 4 H $*'+ * % . 8 + 9 0* + ; 1 .*
A : : 4

B 1 ! "$/2" " "3 4 + " ? & + / $ +
+ + ; 1 .*

! "#$%$ & ' ' "( ) " "& 7 + + 0*
@ : + = BF

B 1 ! "$/2" " "3 4 + " ?
B6& 6 67 + + 60* .*
A # H + .+ % ^
$ +=) +: +: + =

! "#$%$ & ' ' "( ) " "&

B 1 ! "$/2" " "3 4 + " ? + + @ : 4 F /+
% + #+ * : 0K93 ; 1 .*

BB1 ! "$/2" " "3 4 + " ? + + @ : 4 F /+
% + R+ 6 0 0 *) 4 ' + F * H . 4 ; 1 .*

B 1 ! "$/2" " "3 4 + " ? & .*

! "#$%$ & ' ' "( ) " "% % = 0' = 4 2 + + *
/ + ' = ' =

B 1 ! "$/2" " "3 4 + " ? B & : + U % H *
@ +6 $,+ / 4 #+' * 9 : d) 8 Q = + * : 4
; 1 .*

B 1 ! "$/2" " "3 4 + " ? & : + U % H *
@ +6 $,+ / 4 #+' * 9 : d) 8 Q = + * : 4
; 1 .*

B?1 ! "$/2" " "3 4 + " ? : & : + U %
H * @ +6 $,+ / 4 #+' * 9 : d) 8 Q = + * : 4
; 1 .*

1 ! "$/2" " "3 4 + " ? & : + U %
H * @ +6 $,+ / 4 #+' * 9 : d) 8 Q = + * : 4
; 1 .*

1 ! "$/2" " "3 4 + " ? & : + U % H *
@ +6 $,+ / 4 #+' * 9 : d) 8 Q = + * : 4
; 1 .*

1 ! "$/2" " "3 4 + " ? & : + U % H *
@ +6 $,+ / 4 #+' * 9 : d) 8 Q = + * : 4
; 1 .*

1 ! "$/2" " "3 4 + " ? 9 : & : + U %

H * @ +6 $,+ / 4 #+' * 9 : d) 8 Q = + * : 4
; .*

1 ! "$/2" " "3 4 + " ? & : + U % % =
0' = 4 2 + + * / + ' = ' = $ @ + J ' + + * ; ?1 .*

B1 ! "$/2" " "3 4 + " ? & : + U % H *
@ + $,+ / 4 #+' * 9 : d) 8 Q = + * : 4
; 1 .*

1 ! "$/2" " "3 4 + " ? & : + U % H *
@ + $,+ / 4 #+' * 9 : d) 8 Q = + * : 4
; 1 .*

1 ! "$/2" " "3 4 + " ? & : + U % H *
@ +6 $,+ / 4 #+' * 9 : d) 8 Q = + * : 4
; 1 .*

! "#$%$ & ' ' "( ) "

! "#$%$ & ' ' "( ) " "! + + @ * % = := Q @ +)

1 ! "$/2" " "3 4 + " ? $*'+ * ! + + @ * % =
; ?1 .*

?1 ! "$/2" " "3 4 + " ? ! + + @ * % = 3
0 *= + F & $: + 8+ 4 % + * 8 +' + . 4 +4 (9 * : 4
; 1 .*

1 ! "$/2" " "3 4 + " ? ! + + @ * % = := Q
@ +) .*

1 ! "$/2" " "3 4 + " ? ! + + @ * % = . 4
(9 * : 4 ; 1 .*

1 ! "$/2" " "3 4 + " ? ! + + + *
E + * = H * : 4 ! .+ = ; 1 .*
A B .+4
>4 B; 1 ! +. .+

1 ! "$/2" " "3 4 + " ? 0 = . * + =
. ! + + + + * @ * % = 6 . 4 ; ? 1 *S'

1 ! "$/2" " "3 4 + " ? 0 = . * + =
. ! + + + + * @ * % = 6 . 4 ; ? 1 .*
A ? .+4
3: , ' * <<<

B1 ! "$/2" " "3 4 + " ? @ + 0 , ! + + *
= +4 + * 4 +' + H = !9 ; ?? 1 *S'

1 ! "$/2" " "3 4 + " ? @ + 0 , ! + + *

= +4 + * 4 +' + H = !9 ; ?? 1 .*

1 ! "$/2" " "3 4 + " ? 9 + &+ + ! + + @ *
% = !+ : *4 J ' = ; ?1 .*

1 ! "$/2" " "3 4 + " ? . ! + + +
+ * @ * % = U . * K + +) * @ * .*
A ? ? .+4
3: , ' * <<<

?1 ! "$/2" " "3 4 + " ? / + ' = + * * =
* . + + ' = + * + + + * E + * = . 4 ; 1 .*
>4 ! +. % 0E '+ .
3:
>4 ; 1

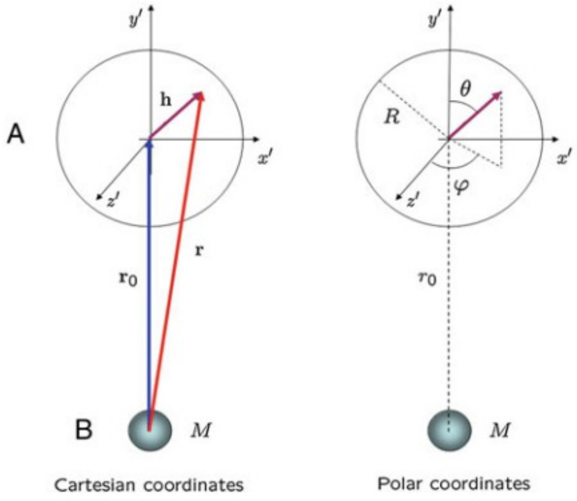

A
y′
h
x′
z′
r
r0
B
M
Cartesian coordinates
y′
θ
R
x′
z′
φ
r0
M
Polar coordinates

>4 % 8 $ + 4=

% 8 $ + 4=
7 = C. + E '+ . . * * . '
+ + = +* + * G + .+ 4 =F * : *
. + + ' = :=)) .+ % * * + .+ .+
* C +.
H : 4 := + 4 + 0 * + 4 = * + : .
F '+ + * + + * + + !+ + *
+ = F + : +

Chapter 3
The Equivalence Principle

3.1 Inertial and Gravitational Masses

In this section we discuss the *principle of equivalence*. We shall see that, besides allowing the extension of the principle of relativity to *a generic*, not necessarily inertial, frame of reference, it allows to define gravity, in a relativistic framework, as a property of the four-dimensional space–time geometry.

The *principle of equivalence*, in the so-called *weak* form, asserts the *exact equivalence between the inertial mass* m_I *and the gravitational one* m_G. We recall that the inertial mass m_I is defined through Newton's second law of dynamics:

$$\mathbf{F} = m_I \mathbf{a}. \tag{3.1}$$

Its physical meaning is, as is well known, that of *inertia* of a body, that is its reluctance to be set in motion or, more generally, to change its velocity. The gravitational mass m_G, on the other hand, enters the definition of Newton's universal law of gravitation according to:

$$\mathbf{F} = -G\frac{m_G M_G}{r^2}\mathbf{u_r} \tag{3.2}$$

where $G = 6.6732 \times 10^{-11}\ \mathrm{Nm}^2/\mathrm{Kg}^2$ is the gravitational constant, m_G and M_G are the gravitational masses of the two attracting bodies.[1] For definiteness we refer to the situation where the mass M_G attracts the mass m_G. We see that the gravitational mass sets the strength of the gravitational force that, *ceteris paribus,* a given body exerts on another one. In this sense it would better deserve the name of *gravitational charge*, in analogy with Coulomb's law of the electrostatic interaction, where the electric charge sets the strength of the electric interaction, the mathematical structures of the two laws being exactly the same.

---

[1] It goes without saying that we are referring to two spherical bodies or to bodies whose dimensions are negligible with respect to their distance r.

R. D'Auria and M. Trigiante, *From Special Relativity to Feynman Diagrams*,
UNITEXT, DOI: 10.1007/978-88-470-1504-3_3,
© Springer-Verlag Italia 2012

However, besides the enormous *quantitative* difference between the *strengths* of *gravitational* and *electrostatic* forces[2] there is a major qualitative difference between them: While any two bodies have mass, they do not necessarily have charge. As a consequence the gravitational force is *universal*, while the electric force is not.

It is important to note that the equality between inertial and gravitational mass, expressed by the principle of equivalence, is in some sense surprising, given the substantial difference between these two physical concepts, which reflects into different operational definitions of their respective measures.

It is, however, one of the best established results from the experimental point of view.

Indeed a large number of experiments were devised, since Newton's times, to ascertain the validity of this unexpected coincidence; among them, of particular importance from a historical point of view is the *Eötvos experiment*, of which we give a short description in Appendix A. The precision reached in this experiment is such that $\frac{\Delta m}{m_I} \equiv \frac{|m_G - m_I|}{m_I}$, is less than 2×10^{-3}. Thanks to the advanced technological features of modern experimental physics, the above ratio has been pushed to less than 10^{-15}. This justifies the theoretical assumption of the exact equality between inertial and gravitational mass:

$$m_I = m_G, \tag{3.3}$$

that is the *principle of equivalence in its weak form*, can be safely assumed as one of the experimentally best established principles of theoretical physics.

The great intuition Einstein had at the beginning of last century, was to realize the considerable importance of this seemingly curious "coincidence", since it implies that *locally, it is impossible to distinguish between the effects of a gravitational field and those of an accelerated frame of reference.*

To justify such a conclusion we illustrate a so-called "Gedankenexperiment", that is a conceptual experiment, originally formulated by Einstein to be performed in a frame of reference attached to an elevator; in our era of space journeys, it seems however more appropriate to update this experiment by replacing Einstein's elevator with a spaceship. Note that the time duration of the experiment inside the spaceship, together with its spatial extension, define a four-dimensional region of space–time. In the following we shall moreover restrict ourselves to the framework of the Newtonian theory of gravitation.

Let us now describe this conceptual experiment, in the following four steps:

(i) Suppose the spaceship, to be simply referred to by $\mathcal{A}$, is initially placed, with the engines turned off, in a region of space which is far enough from any celestial body for the net gravitational force acting on it to be negligible. By definition this is an *inertial frame of reference.* If a physicist performs experiments within the spaceship, he will find that all bodies (if not subject to other kind of forces) move in a rectilinear uniform motion, according to Galileo's principle of inertia.

---

[2] We recall that the ratio between the electric and gravitational forces between two protons is of the order of 10^{38}.

(ii) Let us assume next that, at some stage, the spaceship reaches the proximity of a planet, and thus becomes subject to a gravitational acceleration of the form:

$$\mathbf{g}(\mathbf{r}) = -G\frac{M}{r^2}\mathbf{u_r}, \tag{3.4}$$

M being the (gravitational) mass of the planet.

In the presence of such an attraction $\mathcal{A}$ starts orbiting around the planet, thus becoming an *accelerated frame of reference*. The accelerated frame of reference defined by *any* massive body which is not subject to forces other than gravity is usually referred to as a *free falling frame*. Our spaceship $\mathcal{A}$ orbiting around the planet, is an example of a free falling frame. In this situation an observer inside $\mathcal{A}$ still finds that all bodies execute an *inertial motion*: Indeed, the same acceleration $\mathbf{g}(\mathbf{r})$, acting on the frame $\mathcal{A}$, also acts on each body inside of it. It follows that the motion of bodies in $\mathcal{A}$ is inertial with respect to the spaceship itself, since their relative acceleration with respect to $\mathcal{A}$ is zero.

Since in both cases (i) and (ii) the motion inside $\mathcal{A}$ is of the same kind, that is inertial, there is no way an observer in the spaceship can distinguish between the two situations. We conclude that:

It is not possible through experiments performed in the interior of $\mathcal{A}$ to tell whether it is a free falling frame or an inertial one.[3]

(iii) Consider again the spaceship $\mathcal{A}$ in a region far from any celestial body, and suppose that now, in contrast to case (i), its engines are turned on, thus producing an acceleration $\mathbf{a_R}$ which is uniform all over the spaceship and constant in time. Since now $\mathcal{A}$ is accelerated, all bodies inside of it fall with the same acceleration $\mathbf{a}' = -\mathbf{a}_R$, where $\mathbf{a}'$ and $\mathbf{a}_R$ are the accelerations with respect to $\mathcal{A}$ and the relative acceleration of $\mathcal{A}$ with respect to an inertial system, respectively.

(iv) Finally consider the situation in which the spaceship is at rest on a massive body (for example the earth) which creates a gravitational field of acceleration $\mathbf{g}$; because of the equality between inertial and gravitational masses, *all bodies inside $\mathcal{A}$ fall with the same acceleration* $\mathbf{g}$.

Comparing these two last examples, we see the an observer inside the spaceship cannot tell whether he is in case (iv), where the observed acceleration is due to a gravitational field and his frame of reference is inertial, or in case (iii) where his frame of reference is accelerated by the engines of the spacecraft with acceleration $\mathbf{a_R} = -\mathbf{g}$ with respect to an inertial frame. We conclude that: *It is not possible, through experiments performed in the interior of $\mathcal{A}$, to distinguish an inertial frame in the presence of a gravitational field from an accelerated system of reference.*

The previous discussion seems to lead to the conclusion that a perfect equivalence exists between the effects of a gravitational field as observed in an inertial frame of reference and those observed in an accelerated frame. This may seem very strange:

---

[3] The inertial motion inside a free falling system is a well known fact nowadays, think about the absence of weight of astronauts inside orbiting spacecrafts.

A gravitational field after all has its sources in the massive bodies where the field lines converge, and its intensity is a decreasing function of the distance from them, vanishing at infinity. The lines of force of the acceleration field $\mathbf{a}'$, on the contrary, are not converging and do not fall down to zero at infinity.

In fact, as we shall now show, a more accurate and quantitative discussion of the "Gedankenexperiment" examined above, shows that the assumed *equivalence between gravitational field and non-inertial frames, holds only if we restrict ourselves to events taking place in a small space region and during a small time interval, that is, mathematically, in an infinitesimal region of space–time.*

To clarify this point, let us examine the situations described in points (i) and (ii) in some more detail. It is not difficult to realize that the supposed equivalence between the free falling frame and the inertial system holds only in the limit where the acceleration $\mathbf{g}(\mathbf{r})$ can be regarded as uniform in the interior of $\mathcal{A}$, that is only if:

$$\mathbf{g}(\mathbf{r}) \simeq \mathbf{g}(\mathbf{r_0}), \tag{3.5}$$

where $\mathbf{r_0}$ is the barycenter of the spaceship.

Suppose indeed we have a system of particles of masses $m_1, m_2 \dots m_N$ in a gravitational field, $\mathbf{g}$, each obeying its own equation of motion; let us now rewrite the equation of motion of, say, the kth-particle (3.1), in an accelerated frame S' having a relative acceleration $\mathbf{a_R}$ with respect to an inertial frame S.[4] One has:

$$m_{G(k)}\mathbf{g}(\mathbf{r}_k) = m_{I(k)}\left(\mathbf{a}'_k + \mathbf{a}_R\right) \tag{3.6}$$

where $\mathbf{a}'_k$ is the acceleration of the particle in the frame S'. In particular, if S' is the free falling system, we have:

$$\mathbf{a}_R = \mathbf{g}(\mathbf{r_0}), \tag{3.7}$$

so that (3.6) becomes:

$$m_{G(k)}\mathbf{g}(\mathbf{r}_k) - m_{I(k)}\mathbf{g}(\mathbf{r_0}) = m_{I(k)}\mathbf{a}'_k. \tag{3.8}$$

The discussion made in points (i) and (ii) holds provided we use the approximation (3.5), $\mathbf{g}(\mathbf{r}_k) \simeq \mathbf{g}(\mathbf{r}_0)$, that is *we consider a sufficiently small neighborhood* of the center of mass of $\mathcal{A}$, such that the gravitational acceleration can be approximated by a constant vector and higher order effects in the distance $|\mathbf{r}_k - \mathbf{r}_0|$ can be neglected. In this case taking into account the principle of equivalence, $m_{G(k)} = m_{I(k)}$, from (3.8) it follows:

$$\mathbf{a}'_k \simeq 0. \tag{3.9}$$

This more accurate analysis implies that the equivalence between the free falling frame and the inertial frame described in the former two experiments only holds

---

[4] We recall that the relative acceleration $\mathbf{a}_R$ is given in general by the sum of three terms, the translation acceleration $\mathbf{a}'$, the centripetal acceleration $\mathbf{a}_{(centr)}$, and the Coriolis acceleration $\mathbf{a}_{(Cor)}$.

locally, that is in an infinitesimal neighborhood where, up to higher order terms, (3.5) is valid. The previous discussion can be made more general by assuming that, besides the interaction with the gravitational field, the particles are also subjected to reciprocal *non gravitational forces* $\mathbf{F}_{kl}$. In this case let us write the equation of motion of the kth particle in an inertial frame where a gravitational field $\mathbf{g}$ is present:

$$m_{Gk}\mathbf{g}(\mathbf{r}_k) + \sum_{l\neq k}\mathbf{F}_{kl}(|\mathbf{r}_k - \mathbf{r}_l|) = m_{Ik}\frac{d^2\mathbf{r}_k}{dt^2}. \tag{3.10}$$

We may write the corresponding equation in the free falling system through the coordinate transformation

$$\mathbf{r}' = \mathbf{r} - \frac{1}{2}\mathbf{g}t^2. \tag{3.11}$$

In the same hypotheses made before, namely assuming (3.5), which implies $\mathbf{g} = const.$ and (3.3), (3.10) takes the following form:

$$\sum_{l}\mathbf{F}_{kl}(|\mathbf{r}'_k - \mathbf{r}'_l|) = m_{Ik}\frac{d^2\mathbf{r}'_k}{dt^2}, \tag{3.12}$$

which is the equation of motion of classical mechanics in the absence of a gravitational field.

As far as the other two situations (iii) and (iv) are concerned, it is clear that also in this case the equivalence only holds if the approximation (3.5) is used, that is only *locally*, since the gravitational field would in general be a function of the point inside the spaceship, while the field of accelerations is, at each instant, exactly uniform.

Let us recall that our analysis so far has been made within the framework of classical Newtonian mechanics, without any reference to the implications of Einstein's special relativity. If we now assume that the statement of the *local equivalence* between an inertial frame of reference in a gravitational field and an accelerated one also holds in the framework of relativistic mechanics, described in Chap. 2, then we may reformulate the conclusions of points (i) and (ii), as follows:

In the presence of a gravitational field, locally, the physical laws observed in a free falling frame are those of special relativity in the absence of gravity. As explained above, *locally* means, mathematically, in an infinitesimal neighborhood or, more physically, in a sufficiently small neighborhood of a point such that, up to higher order terms, the approximation (3.5) holds.

The previous statement is referred to as the *equivalence principle in its strong form* or simply *strong equivalence principle*.

In particular we see that when gravitation is included in the description of physical systems, if the *strong equivalence principle* is assumed to hold, the special role played by *inertial frames* is lost, since frames of reference which are rigorously inertial only exist in infinitesimal regions of space–time. Moreover, since acceleration and gravitation are locally indistinguishable it is reasonable to assume that *the physical laws should take the same form also in non-inertial frames*, or, in other words, to require that *the principle of relativity be valid not only in inertial frames, but more generally, in every frame of reference, described by generic four-dimensional coordinate systems.* Recall indeed that we restricted ourselves, in the first chapter, to coordinate transformations which were linear, since we were only interested in inertial frames of references (endowed with three-dimensional Cartesian coordinate systems).[5] Anticipating concepts to be introduced in next chapter, this characterizes the four-dimensional coordinate system associated with inertial frames of reference to be Cartesian or *rectilinear*. Extending our analysis to non-inertial frames implies considering general (four-dimensional) coordinate systems, related to one another by non-linear transformations: Transformations to *arbitrary accelerated frames of reference* are described by *arbitrary transformations of the four coordinates labeling space–time.* It follows that in order to implement the principle of relativity on the laws of physics, in the presence of gravitation, we must require these to have the same form in *any frame of reference*, so that they be covariant under *general coordinate transformations*, i.e. arbitrary changes of coordinates with a non-vanishing, coordinate-dependent, Jacobian:

$$x^{\alpha\prime} = f^{\alpha}(x^0 \equiv ct, x^1, x^2, x^3). \tag{3.13}$$

Summarizing, while in the special theory of relativity, where the gravitational interaction is not taken into account, the physical laws were required to be covariant only under the linear Lorentz transformations, relating inertial frames, in the presence of gravity the implementation of the principle of relativity requires that:

The laws of the Physics be covariant under general coordinate transformations (3.13).

Note that *covariance under general coordinate transformations* means, as we discussed in the case of the Lorentz transformations, that the equations describing the physical laws have exactly the same form, albeit in the transformed variables, in every coordinate system.

In other words: A theory including a treatment of the gravitational field must be *generally covariant*.

It is really amazing that this conclusion, assumed by Einstein as the starting point for a relativistic theory of gravitation, can be drawn simply from the principle of equivalence, that is the equality between inertial and gravitational masses.

Because of its general covariance the relativistic theory of gravitation is called *general theory of relativity.*

---

[5] Linearity was then a consequence of the requirement that the principle of inertia holds in both the old and the transformed frames: A motion which is uniform with respect to one of them cannot be seen as accelerated with respect to the other.

Fig. 3.1 Tidal forces

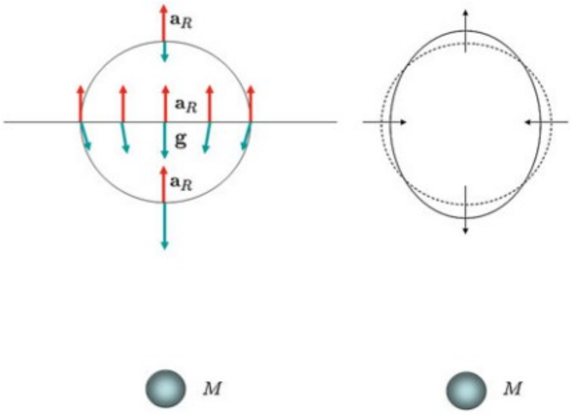

3.2 Tidal Forces

We have seen that it is always possible to *locally* eliminate the effects of a gravitational field by using a *free falling frame*. It is then extremely important to examine those effects of a gravitational field which cannot be eliminated in the free falling frame, that is which manifest themselves when we go beyond the crude equivalence implied by the approximate relation (3.5).

In the present section we shall work purely in the classical limit, that is with no reference to the corrections implied by special relativity, and show that what remains of a gravitational force in the free falling frame are the *tidal forces*, see Fig. 3.1.

Let us start indeed from equation (3.8) which, in the classical case, assuming $m_I = m_G$, contains no approximations:

$$\mathbf{g}(\mathbf{r}) - \mathbf{g}(\mathbf{r_0}) = \mathbf{a}', \tag{3.14}$$

the position vector $\mathbf{r}$ defining a generic point in a neighborhood of $\mathbf{r_0}$. We shall also denote by $\mathbf{h} = \mathbf{r} - \mathbf{r_0} = (h^i)$ the relative position vector between the two points, with components $h^k \equiv x^k - x_0^k$, where $k = 1, 2, 3$ and $(x^k) \equiv (x, y, z)$, $(x_0^k) \equiv (x_0, y_0, z_0)$ (see Fig. 3.2).

Let us compute, to the first order in h^k, the i-th component of the gravitational acceleration in the free falling frame. From (3.14) it follows:

$$a^{i\prime} = \sum_{k=1}^{3} \left.\frac{\partial g^i}{\partial x^k}\right|_{h=0} h^k \equiv -\sum_{k=1}^{3} \left.\frac{\partial^2 \phi}{\partial x^k \partial x^i}\right|_{h=0} h^k + O(|\mathbf{h}|^2). \tag{3.15}$$

where $i = 1, 2, 3$ and ϕ denotes the gravitational potential.

Taking into account that

$$g^i = -\frac{\partial \phi}{\partial x^i} = -G\frac{M x^i}{r^3}, \tag{3.16}$$

Fig. 3.2 Coordinates

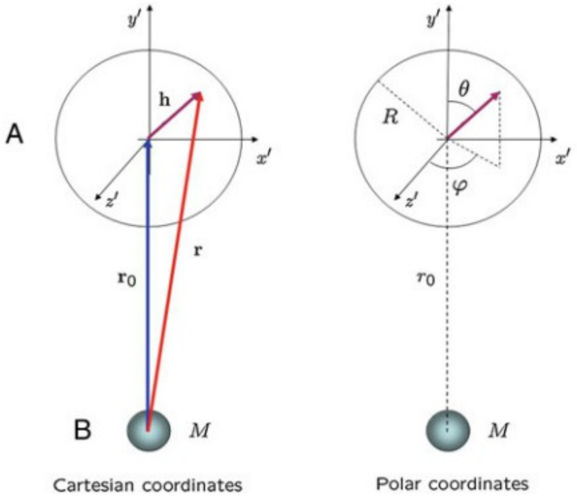

Cartesian coordinates

Polar coordinates

equation (3.15) becomes:

$$a^{i\prime} = -G\frac{M}{r_0^3}\sum_{k=1}^{3}\left(\delta^{ik} - 3\frac{x_0^i x_0^k}{r_0^2}\right)h^k. \tag{3.17}$$

The acceleration field $\mathbf{a}' \equiv (a^{i\prime})$, defined in (3.17), is *the remnant, to order* $O(|\mathbf{h}|)$, of the gravitational field in the free falling frame. We see that it is essentially given by the *gradient of the gravitational field* and is called the *tidal field*. Correspondingly, the force $\mathbf{f} = m\mathbf{a}'$ of the tidal field, acting on a given mass m, is the *tidal force*.

We stress that all these considerations have been obtained using the classical Newtonian formula for the gravitational field which is both *static and non-relativistic*.

In order to show what the effect of tidal forces on an extended body is, let us consider two bodies, say A and B, subject to their mutual gravitational interaction, and let us assume that A is free falling in the gravitational field of B, like for instance an orbiting satellite. We also assume, for the sake of simplicity, that the free falling body A is spherical. We call S' the frame of reference attached to A and S the one attached to B. Let the origin of S' coincide with the barycenter $\mathbf{r_0}$ of A, the z'-axis coincide with direction joining $\mathbf{r_0}$ to the center of mass of the attracting body B and the x' and y'-axes lie, as usual, in the plane orthogonal to the z'-axis (see Fig. 3.2).

With reference to this configuration we observe that $x_0^3 \equiv z_0 \equiv r_0$, so that (3.17) gives:

$$f^{z\prime} = 2G\frac{Mm}{r_0^3}h^{z\prime}, \tag{3.18}$$

$$f^{x\prime} = -G\frac{Mm}{r_0^3}h^{x\prime}, \tag{3.19}$$

$$f^{y'} = -G\frac{Mm}{r_0^3}h^{y'}, \tag{3.20}$$

that is, in matrix notation:

$$\begin{pmatrix} f^{x'} \\ f^{y'} \\ f^{z'} \end{pmatrix} = m \begin{pmatrix} -G\frac{M}{r_0^3} & 0 & 0 \\ 0 & -G\frac{M}{r_0^3} & 0 \\ 0 & 0 & 2G\frac{M}{r_0^3} \end{pmatrix} \begin{pmatrix} h^{x'} \\ h^{y'} \\ h^{z'} \end{pmatrix}, \tag{3.21}$$

where the entries of the matrix on the right hand side are $\partial g^i/\partial x^k$ computed in $\mathbf{r}_0$. From (3.18) it follows that $f^{z'}$ is attractive or repulsive according to the sign of $h^{z'}$; that is, referring to Fig. 3.1, it points downwards on the part of the body facing B $(z' > 0)$, and upwards on the opposite side $(z' < 0)$. The two horizontal components $f^{x'}$ e $f^{y'}$, instead, are always attractive, that is they are directed towards the origin of S'. The net result is an outward stress acting along the line joining A and B and an inward stress on the horizontal planes $z' = const$.

Tidal forces can be very strong in the astrophysical phenomena; for example the tidal forces exerted by the greatest planets of the solar system on their satellites induce a tidal heating due to the consequent internal friction; in the case of the Jupiter satellite Io this results in dramatic volcanic eruptions.[6]

In the case of a deformable body, tidal forces deform a spherical body to the shape of an ellipsoid, the major axis lying along the A-B direction. This is in fact what happens in the case of the earth where the corresponding phenomenon, induced by the moon, gives rise to the ordinary oceanic tides, whence the denomination *tidal forces* has its origin.

We may indeed think of the earth as the body A in free fall on the gravitational field of the moon, the body B. The "thin" layer represented by the oceans covering the earth's surface is indeed deformable. It follows that on the side facing the moon and on the opposite side tidal forces give rise to high tides, while in the directions perpendicular to the line earth-moon tidal forces produce low tides (Fig. 3.1). Because of the bipolar character of these bulges and compressions, the periodicity of tides is of 12 h. We may make a crude estimation of the tidal size on the earth.

First we observe that tidal forces are conservative. Indeed from equations (3.18), (3.19), (3.20) it follows that the associated tidal potential energy is:

$$\begin{aligned} \mathcal{E}_P &= -\frac{GM_L m_T}{r_0^3}\left(z'2 - \frac{1}{2}x'2 - \frac{1}{2}y'2\right) = -\frac{GM_L m_T}{r_0^3}\left(-\frac{1}{2}R^2 + \frac{3}{2}z'2\right) \\ &= -\frac{GM_L m_T}{r_0^3}R^2\left(\frac{3\cos^2\theta - 1}{2}\right), \end{aligned} \tag{3.22}$$

where we have introduced spherical coordinates with origin on the center of the earth (see Fig. 3.2), R being its radius, θ the latitude and we have denoted by $M_L,\ m_T,\ r_0$, the moon's mass, the earth's mass and the earth-moon distance, respectively.

---

[6] In the proximity of the event horizon of a black hole tidal forces are so strong as to completely disintegrate any body falling inside.

The total potential energy of a water particle is obtained as the sum $\mathcal{E}_P + mgh$, where *mgh* is the potential energy of the terrestrial attraction. At the equilibrium the form of the oceans' equipotential surface is determined by the condition:

$$E_P^{tot} = mgh - \frac{GM_L m_T}{r_0^3} R^2 \left(\frac{3\cos^2\theta - 1}{2} \right) = \text{cost}. \tag{3.23}$$

Equation 3.23 implies $h = h(\theta)$; the difference in height between high ($\theta = 0$) and low ($\theta = \frac{\pi}{2}$) tides turns out to be:

$$\Delta h \equiv h(0) - h(\pi/2) = \frac{3GM_l}{2gr_0^3} R^2 \simeq 53 \text{ cm}. \tag{3.24}$$

If the gravitational pull of the sun is also taken into account, one finds an additional contribution about half as large as the previous one.[7]

3.3 The Geometric Analogy

In this section we try to explain how the equivalence principle discussed in the previous section naturally leads to a modification of the space–time geometry, described in special relativity by the Minkowski space. To this end we need to anticipate part of the discussion of next chapter.

We begin by recalling that in Euclidean geometry the distance between two points P_1 e P_2, which is *invariant* under rotations and translations of the Cartesian coordinate system, can be written as[8]

$$\Delta\ell^2 = (\Delta\xi^1)^2 + (\Delta\xi^2)^2 + (\Delta\xi^3)^2 = \sum_{i,j} \delta_{ij} \Delta\xi^i \Delta\xi^j, \tag{3.25}$$

if and only if we are using Cartesian (rectangular) coordinates ξ^i. By Cartesian (rectangular) coordinates we mean rectangular (or rectilinear) coordinates, which, in studying Euclidean geometry, are defined throughout the space. This is possible only if the geometry is Euclidean, that is the properties of figures, the notion of parallelism and so on are those derived by Euclid's axioms.

In (3.25) we have adopted the following notation:

$$\begin{aligned} \delta_{ij} &= 1 \quad i = j, \\ \delta_{ij} &= 0 \quad i \neq j, \end{aligned} \tag{3.26}$$

---

[7] This crude estimate must be considered, together with the correction due to the sun, just a mean value. It does not take into account resonance phenomena due to the earth rotation and the shape of the oceanic depths which can locally alter in a sensible way the rough computation leading to (3.24).

[8] In this section only we denote by ξ^i the Cartesian coordinates, while the notation x^i is used for generic curvilinear coordinates.

that is, in matrix notation:

$$(\delta_{ij}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{3.27}$$

The Kronecker symbol δ_{ij} defines the so called *metric tensor*

If we adopt the Einstein convention that repeated indices are summed over, then (3.25) can be written as follows:

$$\Delta\ell^2 = \sum_{i,j} \delta_{ij} \Delta\xi^i \Delta\xi^j \equiv \delta_{ij} \Delta\xi^i \Delta\xi^j, \tag{3.28}$$

or, if the two points are infinitesimally apart,

$$d\ell^2 = \delta_{ij} d\xi^i d\xi^j. \tag{3.29}$$

It is clear that the same considerations and formalism hold if the Euclidean space has a generic number D of dimensions; in this case the indices $i, j, \dots$ would run over D values instead of only three.

As we have already remarked in Sect. 1.5 of the first chapter, there is a close analogy between the *four-dimensional distance* in Minkowski space–time and the Euclidean distance (3.28) in $D = 4$ dimensions. Indeed the *four-dimensional distance* between two events[9] labeled by the four coordinates $\xi^0, \xi^1, \xi^2, \xi^3$ was written as[10]:

$$\begin{aligned} \Delta s^2 = c^2 \Delta\tau^2 = (\Delta\xi^0)^2 - (\Delta\xi^1)^2 - (\Delta\xi^2)^2 - (\Delta\xi^3)^2 = \sum_{\alpha,\beta} \eta_{\alpha\beta} \Delta\xi^\alpha \Delta\xi^\beta \\ \equiv \eta_{\alpha\beta} \Delta\xi^\alpha \Delta\xi^\beta \quad (\alpha, \beta = 0, 1, 2, 3) \end{aligned} \tag{3.30}$$

that is in a way which is strictly analogous to the four-dimensional Euclidean distance, the only difference being the replacement of the metric δ_{ij} with the Minkowski metric $\eta_{\alpha\beta}$:

$$\eta_{\alpha\beta} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \tag{3.31}$$

where the extra fourth coordinate is related to time, $\xi^0 = ct$. We stress that these simple expressions of distance in Euclidean space or proper distance in Minkowski case, are only valid if we use the three-dimensional Cartesian rectangular or the analogous four-dimensional Cartesian rectangular (also referred to as Minkowskian[11]) coordinates ξ^α, the latter being defined rectangular as the coordinates used to describe inertial frames in terms of the spatial Cartesian rectangular coordinates and the usual time coordinate t. In any other coordinate system, not related by a three-dimensional rotation or a Lorentz transformation, respectively, the Euclidean distance (3.28) or the Minkowski proper distance (3.30) would take a more complicated form.

---

[9] Recall that the space–time (four-dimensional) distance was conventionally defined as the *negative* of the *proper distance*, see (2.44).

[10] From now on, we use Greek indices to label four dimensional space-time coordinates and Latin ones for the coordinates in Euclidean space.

Suppose indeed that in the Euclidean case we want to use an arbitrary system of curvilinear coordinates $x^i, \; i = 1, 2, 3$ (an example would be the spherical polar coordinates); we would then have:

$$\xi^i = \xi^i(x^j). \tag{3.32}$$

In these new coordinates the infinitesimal distance (3.30) becomes:

$$d\ell^2 = \delta_{ij}\frac{\partial \xi^i}{\partial x^k}\frac{\partial \xi^j}{\partial x^\ell}dx^k dx^\ell \doteq g_{k\ell}dx^k dx^\ell, \tag{3.33}$$

where

$$g_{k\ell} = \delta_{ij}V^i_k V^j_\ell, \tag{3.34}$$

being:

$$V^i_k \doteq \frac{\partial \xi^i}{\partial x^k}. \tag{3.35}$$

The dimensionless quantity $g_{k\ell}(x)$, replacing δ_{ij} in the formula for the squared distance, is called *metric tensor* or, more simply, *metric*[12] in curvilinear coordinates.

It is obvious that all the geometric quantities of Euclidean geometry (lengths, angles, areas, etc) do not depend of the particular coordinates used for their description. However it is well known that in general it is much simpler to compute geometric quantities using Cartesian coordinates, rather than the curvilinear ones.[13]

The same considerations would of course apply to Minkowski space–time in special relativity, if we were to use arbitrary "curvilinear" four-dimensional coordinate frames. The physical interpretation in this case would be the following: Since an arbitrary change of coordinates would correspond to arbitrary functions of the original Minkowski coordinates $\xi_0 \equiv ct, \xi_1, \xi_2, \xi_3$, the new frame of reference cannot be inertial since the transformation is, in general, not linear as it is the case for Lorentz transformations. It then follows that the new coordinate system (x^μ) must correspond to an *accelerated frame of reference*. Moreover, as it happens in the Euclidean case, instead of a constant Minkowski metric $\eta_{\alpha\beta}$ we would end up with a metric tensor depending on the four coordinates $x^\mu = x^0, x^1, x^2, x^3$. This, of course, would not change the physical laws, but only describe them in a more general, albeit cumbersome, way.

---

[11] We shall call Minkowskian the Cartesian rectangular coordinates in the four-dimensional Minkowski space–time, with metric (3.31).

[12] Here and in the following of this chapter we use the word *tensor*, whose precise meaning will be given in Chap. 4, in a loose sense, that is as a quantity carrying indices and whose transformation properties are fixed in terms of the change of coordinates (or of reference frame). In this chapter the transformation of coordinates considered are either cartesian orthogonal, or Lorentzian or even arbitrary, as explained below.

[13] When dealing with problems which exhibit some degree of symmetry it may, however, be more useful to use curvilinear coordinates, like spherical, cylindrical, etc.

Given these preliminaries we now define a space, or a space–time as *flat* if there exists a *special class of coordinates* such that the metric assumes a constant value *in a finite or infinite domain of the space or space–time*. In Euclidean and Minkowski spaces this special class is given by the Cartesian coordinates (in particular the Minkowski coordinates of special relativity), any two elements of this class being related by a combination of rotations and translations in Euclidean space or of Lorentz transformations and translations in Minkowski space. Note that the computation in generic curvilinear coordinates of geometric quantities or of physical laws in Euclidean or Minkowskian flat space, respectively, would involve the use of the metric $g_{\mu\nu}(x^\rho)$, that is of a matrix whose elements are function of the coordinates. However, in the special class of frames, the corresponding metric is simply given by the constant matrices (3.27) or (3.31).

If, however, such a *special class of coordinates* cannot be found *in a large domain,*[14] we then say that the space or space–time is *curved*, that is it exhibits *curvature,* a concept which we dwell on in the next section.

3.4 Curvature

Suppose now we have a space (finite or infinite) which possesses a curvature, that is in which it is *not* possible to introduce in a large domain Cartesian (Minkowskian) coordinates. For the sake of simplicity, and to have a help from our intuition, we suppose for the moment that the space in consideration has two space dimensions, though it may not necessarily be the plane $\mathbb{R}^2$, but rather an arbitrary surface Σ. It is well known that on a generic surface it is not possible to introduce *in a finite domain* Cartesian coordinates, but only *curvilinear ones* $x^i,\ i = 1, 2$.

Restricting our attention to an *infinitesimal neighborhood* of a point *P*, however, we can approximate the surface by the tangent plane to Σ at *P*. We say that the *local geometry* of the surface at *P* is identified with that of the corresponding tangent plane, the local coordinates on Σ, in the close vicinity of *P*, coinciding with the Cartesian ones ξ^i, $(i = 1, 2)$ on the plane[15] (see Fig. 3.3):

In an infinitely small neighborhood of *P* we can then use (3.30) to compute the infinitesimal distance between two points, and thus we conclude that *locally the*

---

[14] By *large* domain we mean a domain whose extension is finite or infinite.

[15] Being the plane flat, we can describe it by Cartesian coordinates.

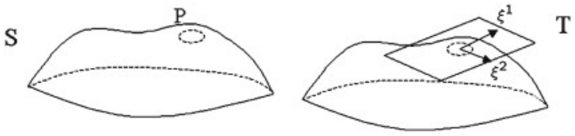

Fig. 3.3 Local geometry and tangent space

metric can be always reduced to the form δ_{ij}.[16] However to compute geometric quantities in a *finite domain* it is necessary to use general curvilinear coordinates by transforming the local coordinates through (3.32).[17] This means that the geometry in a large, possibly infinite, domain is determined by the metric tensor $g_{ij}(x^1, x^2)$, function of the curvilinear coordinates. For example, the length of a curve $\gamma = \gamma(\tau)$ can be computed as:

$$\ell = \int_\gamma d\ell = \int_\gamma \left(g_{ij}(x)dx^i dx^j\right)^{\frac{1}{2}} = \int_\gamma \left(g_{ij}(x)\frac{dx^i}{d\tau}\frac{dx^j}{d\tau}\right)^{\frac{1}{2}} d\tau, \tag{3.36}$$

τ being any parameter on the curve.

We conclude that on a surface with curvature *the metric cannot be reduced to the constant matrix* δ_{ij} *in a finite domain*; its geometry is therefore described by the metric tensor $g_{k\ell}(x)$ or, equivalently by the matrix V^i_k defined by (3.35).

Even if these considerations have been made in the special case of a two-dimensional surface, they can be straightforwardly extended to three-or N-dimensional curved spaces[18] that is to spaces where Cartesian coordinates cannot be introduced in large domains and, also, to spaces *where the local metric is not* δ_{ij}, *but* $\eta_{\alpha\beta}$, as it happens for the Minkowskian space–time of special relativity.

We can now give a precise geometric interpretation of the *strong equivalence principle* of the previous section: Saying that in a free falling frame the laws of the special relativity hold, means that the space–time geometry in such frame is *locally* the same as that of the four-dimensional (hyper)-plane tangent to space–time at the point in which the frame is located.

This (four-dimensional) tangent plane is of course the Minkowski space of special relativity. Indeed we have learned that in a free falling frame, *locally*, that is in an infinitesimal neighborhood of a point $P = (x^0, \mathbf{x})$ in its interior, up to higher order terms, gravitation is absent, so that the metric tensor reduces to the constant metric tensor $\eta_{\alpha\beta}$ of special relativity.[19] If instead we use a general frame of reference, which is not free falling, the metric tensor $g_{\mu\nu}(x)$ must describe the presence of the gravitational field.[20]

---

[16] Here and in the following by *locally* we mean that our statement is valid in an infinitesimal neighborhood of a point where higher order terms can be neglected.

[17] Alternatively one can use Cartesian coordinates at each point using the local tangent plane; in this case, however, one needs a quantity, called *connection*, which relates the local geometries associated with different tangent planes.

[18] A more precise definition of curvature will be given in the next section.

[19] Note that this implies that the free falling frames are the inertial frames described by special relativity. However, in the presence of a gravitational field, they can be only defined locally, since only locally the effects of gravity can be canceled.

We may thus establish the following correspondence between the presence of curvature in a four-dimensional space which is locally Euclidean (metric δ^i_j) and a space which is locally Minkowskian (metric $\eta_{\alpha\beta}$).

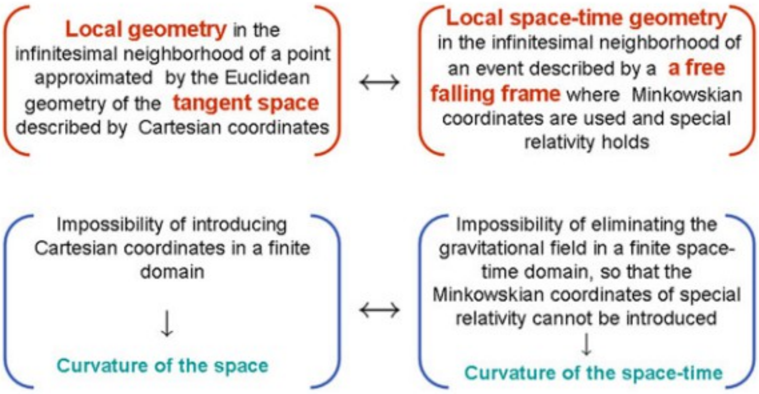
Local geometry in the infinitesimal neighborhood of a point approximated by the Euclidean geometry of the tangent space described by Cartesian coordinates
↔
Local space-time geometry in the infinitesimal neighborhood of an event described by a a free falling frame where Minkowskian coordinates are used and special relativity holds
Impossibility of introducing Cartesian coordinates in a finite domain
↓
Curvature of the space
↔
Impossibility of eliminating the gravitational field in a finite space-time domain, so that the Minkowskian coordinates of special relativity cannot be introduced
↓
Curvature of the space-time

Summarizing, the geometry of space–time is the geometry of a four-dimensional space with local Minkowski metric; this geometry cannot be flat in the presence of a gravitational field. Space–time geometry must then be described by a metric $g_{\mu\nu}(x)$ which depends on the presence of a gravitational field, each entry being a function of the space–time coordinates x^0, x^1, x^2, x^3. The metric can be reduced to the special relativity form $\eta_{\alpha\beta}$ only locally, that is in an infinitely small neighborhood of an event P, but not in a large region of space–time.

Generalizing equations (3.33) and (3.34) to a four dimensional space-time, where $\xi^\alpha (\alpha = 0, 1, 2, 3)$ are the local Minkowskian coordinates and $x^\mu (\mu = 0, 1, 2, 3)$ are the general coordinates parametrizing a large domain of the space, we have:

$$ds^2 = \eta_{\alpha\beta} V^\alpha_\mu V^\beta_\nu dx^\mu dx^\nu = g_{\mu\nu} dx^\mu dx^\nu, \tag{3.37}$$

where

$$V^\alpha_\mu \big|_P = \frac{\partial \xi^\alpha}{\partial x^\mu}\bigg|_P, \tag{3.38}$$

and ξ^α are the *local Minkowskian* (i.e., inertial) coordinates in an infinitesimal neighborhood of P.

---

[20] As we shall see in the sequel of this chapter the space–time metric $g_{\mu\nu}(x)$ is related to the gravitational *potential* rather than to the gravitational *field*.

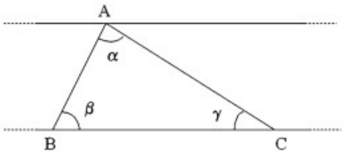

Fig. 3.4 Euclidean geometry

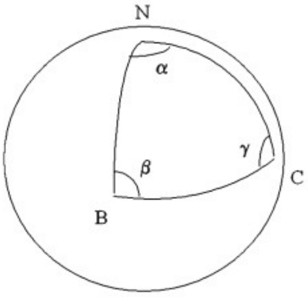

Fig. 3.5 Spherical geometry

3.4.1 An Elementary Approach to the Curvature

We recall that the Euclidean geometry is based on Euclid's eleventh postulate which states the uniqueness of the straight line passing through a given point P and parallel to a given straight line. This postulate, in particular, implies that the three interior angles of a triangle sum up to 180°, see Fig. 3.4:

$$\alpha + \beta + \gamma = \pi. \tag{3.39}$$

Let us take again, for the sake of simplicity and intuition, a two-dimensional space, more specifically a 2-sphere (see Fig. 3.5). Let us then consider a spherical triangle defined by joining along maximal circles three arbitrary points of the sphere.[21]

In spherical geometry one can show that the following relation holds:

$$\alpha + \beta + \gamma = \pi + \frac{A}{R^2}, \tag{3.40}$$

where A is the area of the spherical surface enclosed by the triangle and R is the radius of the sphere.

Let us define the *curvature* K of the sphere as:

$$K = \frac{1}{R^2}. \tag{3.41}$$

the meaning of K is to indicate how much the geometry of the sphere deviates from the Euclidean plane geometry where $K = 0$.

---

[21] A maximal circle is the shortest path joining two points on the sphere. It can be obtained by intersecting the spherical surface with a plane determined by the two points and the center of the sphere.

On the basis of its definition K seems to depend on the radius of the sphere considered as a two dimensional manifold embedded in the ambient Euclidean space $\mathbb{R}^3$. Equation (3.40) however tells us that the curvature K can be evaluated by just performing measures of angles and areas *on the two-dimensional surface of the sphere.* It follows that K has an intrinsic geometric meaning, that is independent of its embedding in the flat 3-dimensional space, since it can be computed by only using measures on the spherical surface. If we had considered a Lobacevskij surface, where K is constant, but negative, (3.40) is also valid.

Considering a *generic* surface Σ, K is of course no longer constant, but becomes a function of the point on the surface. Indeed we may characterize the value of K at P as the curvature of the sphere that best approximates the generic surface in a small neighborhood of P. In this case (3.40) can be generalized by writing:

$$\alpha + \beta + \gamma - \pi = \int_{\Delta} K(x^1, x^2) dA, \tag{3.42}$$

where dA is an infinitesimal element of the surface Σ and Δ is the integration domain.

$K(x^1, x^2)$ is called the *Gaussian curvature* of S at $P \equiv (x^1, x^2)$.

As it follows from the above discussion, K only depends on the intrinsic geometry of Σ, while it does not depend on its particular representation in $\mathbb{R}^3$, in terms of parametric equations of the form:

$$x = x(x^1, x^2), \quad y = y(x^1, x^2), \quad z = z(x^1, x^2), \tag{3.43}$$

where x^1, x^2, are curvilinear coordinates on Σ and x, y, z are Cartesian coordinates on $\mathbb{R}^3$. In fact one can show that K is invariant if S is flexed without stretching or tearing; for example, since K is zero on a plane, it will also be zero on a cone or a cylinder or in general on any surface which can be unfolded on a plane. If instead we are to map a portion of the terrestrial globe on a plane, stretching is necessary, since we have to change the value of K from a positive constant to zero. In this case the map can be considered a good approximation only if the area considered is much smaller than $\frac{1}{K} = R^2$.

3.4.2 Parallel Transport

There is an equivalent way of describing the curvature of a sphere. Consider a cannon at the north pole (which may metaphorically represent a tangent vector) and let us carry it along a meridian till it reaches the equator at a point A, displacing it in such a way as to always keep it *parallel to itself*, so that the angle it forms with the meridian remains constant (for example, in Fig. 3.6, the vector forms an angle zero with the arcs NA and BN and $\pi/2$ with the arc AB). This is what it is meant by *parallel transport*. Eventually, the same kind of *parallel transport* is performed along an arc AB of the equator whose length is $\frac{1}{4}$ the circumference, and finally we carry the cannon back to the north pole along the meridian BN, See Fig. 3.6.

Fig. 3.6 Parallel Transport

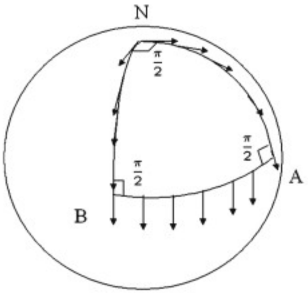

We easily realize that after this tour the cannon arrives back *rotated* by an angle of $\frac{\pi}{2}$, which is exactly the angular excess described by the curvature:

$$\Delta\theta = \alpha + \beta + \gamma - \pi = \frac{3}{2}\pi - \pi = \frac{\pi}{2} = kA. \tag{3.44}$$

Indeed the ratio between the angular excess $\Delta\theta$ and the area of the spherical octant, $\frac{\pi R^2}{2}$, gives exactly the value of the curvature, namely $\frac{1}{R^2} \equiv K$. Although this result was obtained in a very particular case, it can be shown to hold for a *parallel transport* along any closed path γ, not necessarily along geodesic triangles, enclosing an area Δ and on any surface. Indeed one can expect and actually show that in the general case the rotation angle is found by first applying (3.44) to an infinitesimal area *dA* and then integrating over the whole area Δ:

$$\Delta\theta = \int_{\Delta} K(x)dA. \tag{3.45}$$

where Δ is the area enclosed by the curve γ.

This alternative way of defining the curvature can be easily extended to manifolds with any number of dimensions. Since our goal is to define the curvature for the four-dimensional space–time with local Minkowski metric $\eta_{\alpha\beta}$, we consider the *parallel transport* of a vector v^{μ}, $(\mu = 0, 1, 2, 3)$ along a closed path γ in a four-dimensional manifold. After the trip, the vector will get back to the initial point "rotated" with respect to the original direction. However, as we have previously learned, such a "rotation" is a "four-dimensional rotation" in a space endowed with a metric which has the same signature as the Minkowski one $\eta_{\alpha\beta}$,[22] and therefore is a *Lorentz transformation*.

In particular, if the closed path γ is infinitesimal, we can obtain the analogous of the *Gaussian curvature* as in the case of a two-dimensional surface. However, while in two dimensions we just have one orientation for the infinitesimal area *dA*

---

[22] It can be shown that the signature of the metric, that is the number of positive and negative eigenvalues of the matrix $g_{\mu\nu}(x)$, is the same at each point of a manifold. Since at a given point the metric can be taken to coincide with $\eta_{\alpha\beta}$ this explains the meaning of the statement in the text.

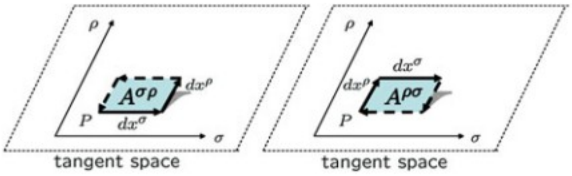

Fig. 3.7 Parallel transports along infinitesimal contours with opposite orientations

enclosed by the path γ, (that of the plane tangent to the surface), in four dimensions *dA* can have several different orientations. More specifically, if we represent γ as an infinitesimal parallelogram, we have $\binom{4}{2} = 6$ orientations, which we may label by the ordered couple of the two infinitesimal displacements dx^ρ, dx^σ defining the parallelogram *dA*: $(\rho, \sigma) = (01), (02), (03), (12), (13), (23)$. We can thus write:

$$dA \to dA^{\rho\sigma} = -dA^{\sigma\rho} \equiv dx^\rho \wedge dx^\sigma , \tag{3.46}$$

where the antisymmetry in the couple of indices $\rho\sigma$, denoted by the symbol $\wedge$, is due to the orientation of $dA^{\rho\sigma}$, which is related to the orientation of the curve γ (in other words the orientation of *dA* depends on whether the vector is transported along γ in one direction or in the other, namely if the displacement dx^ρ precedes or follows dx^σ, see Fig. 3.7).

In conclusion, performing a parallel transport of a Lorentz vector v^μ[23] along an infinitesimal path in space–time the vector undergo an (infinitesimal) Lorentz transformation given by:

$$\delta v^\mu = R^\mu{}_{\nu\rho\sigma} v^\nu dx^\rho \wedge dx^\sigma . \tag{3.47}$$

where Einstein convention of repeated indices is applied and we *define* $R^\mu{}_{\nu\rho\sigma}$ to be the *curvature* of the space–time manifold at a point *P*. We see that in four dimensions, the curvature is actually described by an object carrying four four-dimensional indices, called *Riemann curvature tensor*, or simply *Riemann tensor*, providing the natural generalization of the two-dimensional *Gaussian curvature*.

One can show that the number of *independent* components of $R^\alpha{}_{\beta\gamma\delta}$ is 20 (instead of 4^4).

3.4.3 Tidal Forces and Space–Time Curvature

From the discussion in the previous section on the geometric analogy and from the above geometric definition of the curvature, it follows that the *Riemann tensor* $R^\mu{}_{\nu\rho\sigma}$ describes the deviation of the actual space–time geometry from the *flat* Minkowski geometry, in an intrinsic, coordinate independent way. On the other hand, we have seen that, in the Newtonian approximation, tidal forces acting in the free falling frame, also describe the deviation from the Minkowskian geometry of special relativity. It is therefore obvious that there must be a close relation between curvature and tidal forces.

---

[23] Note that since the closed path is infinitesimal we are allowed to use the in the tangent hyper-plane the usual flat geometry of special relativity and therefore the reference to v^μ as a *Lorentz* vector is appropriate.

To establish this relation, we recall that (3.21) gives the tidal force in the *classical Newtonian approximation*, that is in the limit where the description of the gravitational field is static, non-relativistic and to the first order in the displacement vector **h**.

We also note that the matrix in (3.21), which is the explicit form of $\frac{\partial g^i}{\partial x^k}|_{\mathbf{r}=\mathbf{r}_0}$, has the physical dimensions of $[T^{-2}]$, while curvature has dimension $[L^{-2}]$. Thus if a relation between tidal forces and curvature exists, the two quantities must be related by a factor c^2. At first sight this could seem impossible, since $\frac{\partial g^i}{\partial x^k}$ is a 3×3 matrix with only spatial indices, $i, k = 1, 2, 3$, while the Riemann tensor has four four-dimensional indices, $\alpha, \beta, \dots = 0, 1, 2, 3$.

This different structure of indices, however, simply means that *only some components of the Riemann tensor survive when we take the non-relativistic, static limit of the full relativistic expression of the (gradient of the) gravitational field as given by the Riemann tensor in the general theory.* Indeed, from the exact formula of the Riemann tensor of the general theory of relativity, one can see that, in the non relativistic limit $c \to \infty$, denoting by Latin letters the space indices, to lowest order in *1/c* one obtains:

$$R^k{}_{0\ell 0} = -\frac{1}{2mc^2}\frac{\partial g^k}{\partial x^\ell}, \tag{3.48}$$

all the other components of the Riemann tensor being of higher order in *1/c*. We may therefore write:

$$f^k = -2mc^2 R^k{}_{0\ell 0} h^\ell \quad (k, \ell = 1, 2, 3). \tag{3.49}$$

Comparing (3.48) and (3.49) with (3.21) one obtains:

$$R^k{}_{0\ell 0} = -\frac{1}{2}\begin{pmatrix} 2\frac{GM}{r_0^3c^2} & 0 & 0 \\ 0 & -\frac{GM}{r_0^3c^2} & 0 \\ 0 & 0 & -\frac{GM}{r_0^3c^2} \end{pmatrix}, \tag{3.50}$$

We conclude that this matrix describes the curvature of space–time in the classical non-relativistic limit.

3.5 Motion of a Particle in Curved Space–Time

To determine the trajectory of a particle in space–time we take advantage of the principle of equivalence, by first writing the equation of motion in the free falling frame and then transforming to a general frame.

Locally, in the free falling frame attached to the particle, there is no gravitational field and the motion is purely *inertial* in the local special relativistic coordinates ξ^α $(\alpha = 0, 1, 2, 3)$. On the other hand, in Chap. 2, the special relativistic inertial motion was described by (2.74), namely[24]:

$$U^\alpha = \frac{d\xi^\alpha}{d\tau} = const. \longrightarrow \frac{d^2\xi^\alpha}{d\tau^2} = \frac{dU^\alpha}{d\tau} = 0, \tag{3.51}$$

We stress that this equation is valid in an infinitesimal (both space- and time-) neighborhood of the particle. Switching to the "laboratory" system where the general coordinates x^μ are used, and using the general relation between the two coordinates ξ^α and x^μ:

$$\xi^\alpha = \xi^\alpha(x^\mu), \tag{3.52}$$

we find

$$\frac{d\xi^\alpha}{d\tau} = \frac{\partial \xi^\alpha}{\partial x^\mu}\frac{dx^\mu}{d\tau} = V^\alpha_\mu \frac{dx^\mu}{d\tau}, \tag{3.53}$$

$$\frac{d^2\xi^\alpha}{d\tau^2} = \frac{d}{d\tau}\left(V^\alpha_\mu \frac{dx^\mu}{d\tau}\right) = V^\alpha_\mu \frac{d^2x^\mu}{d\tau^2} + \frac{\partial V^\alpha_\mu}{\partial x^\nu}\frac{dx^\mu}{d\tau}\frac{dx^\nu}{d\tau} = 0. \tag{3.54}$$

We solve this equation with respect to the second derivatives, by multiplying both sides by the inverse matrix $(V^{-1})^\mu_\alpha$, that we denote by

$$V^\mu_\alpha \equiv (V^{-1})^\mu_\alpha; \quad V^\mu_\alpha V^\alpha_\nu = \delta^\mu_\nu. \tag{3.55}$$

Equation 3.54 takes then the following form:

$$\frac{d^2x^\mu}{d\tau^2} + \Gamma^\mu_{\nu\rho}\frac{dx^\nu}{d\tau}\frac{dx^\rho}{d\tau} = 0, \tag{3.56}$$

where we have set:

$$\Gamma^\mu_{\nu\rho} \equiv \frac{1}{2} V^\mu_\alpha \left(\frac{\partial V^\alpha_\nu}{\partial x^\rho} + \frac{\partial V^\alpha_\rho}{\partial x^\nu}\right). \tag{3.57}$$

---

[24] With respect to the notations used in Chap. 2, we have changed notation for the locally inertial coordinates from x^μ to ξ^α $(\alpha = 0, 1, 2, 3)$ since in the present setting the latter describe the locally inertial coordinates of special relativity, while the former describe a general frame, for example the coordinates used in the "laboratory" frame, where the gravitational field is present.

The number of independent components of $\Gamma^{\mu}_{\nu\rho}$, taking into account the symmetry in the two lower indices, is 40.

Note that the second term on the left hand side of (3.56) can be given the meaning of the *gravitational acceleration* impressed to the particle of coordinates x^{μ}; thus $\Gamma^{\mu}_{\nu\rho}$, called *affine connection*, represents the relativistic generalization of the gravitational field.

The solution to (3.56) for the spatial coordinates $x^i, (i = 1, 2, 3)$, gives the trajectory of the particle in the gravitational field while the solution for $x^0 = ct$ gives the relation between the local time t and the proper time τ.

One can show that $\Gamma^{\mu}_{\nu\rho}$ can be expressed in terms of the metric and its derivatives. The quickest way do so is to observe that (3.56), from the four-dimensional point of view, describes a *free inertial motion* in the curved space–time since the gravitational field has been expressed in terms of the affine connection which is a geometric property of space–time. From this point of view there is no force driving the particle; instead, the very presence of a non-trivial geometry, characterized by a non-vanishing Riemann tensor, implies that the free motion must be described by (3.56).

This interpretation is corroborated by the observation that a free motion in a *curved space–time* is the analogue of the inertial motion in flat space–time, given by a straight line. We must therefore expect that the solution to equation (3.56) must represent the analogue, in a curved space, of a straight line in flat space. Such curve is called a *geodesic* and is defined as the shortest line joining two points.

Let us consider two points A and B (events) in the four-dimensional space–time and let γ be a generic curve joining them. Its four-dimensional length $s(\gamma)$ is given by:

$$s(\gamma) = \int\limits_{\gamma[A\to B]} ds = \int\limits_{\gamma[A\to B]} (g_{\mu\nu} dx^{\mu} dx^{\nu})^{\frac{1}{2}}. \tag{3.58}$$

The analogue, in curved space, of the straight line in flat space can be obtained by requiring the curve γ to be such that its length $s(\gamma)$, as a functional of γ, be minimum, as it is the case for the straight line in flat space.

Solving the variational problem one precisely finds (3.56) with:

$$\Gamma^{\mu}_{\nu\rho} = \frac{1}{2} g^{\mu\delta} \left(-\frac{\partial}{\partial x^{\delta}} g_{\nu\rho} + \frac{\partial}{\partial x^{\rho}} g_{\nu\delta} + \frac{\partial}{\partial x^{\nu}} g_{\rho\delta} \right). \tag{3.59}$$

Note that since the relativistic gravitational field is given in terms of $\Gamma^{\mu}_{\nu\rho}$, which contains the first derivatives of the metric, the ten components of the metric $g_{\mu\nu}(x)$ are the relativistic generalization of the Newtonian gravitational potential.

To simplify formulas, in the rest of the book we shall often use the following short-hand notation for partial derivatives:

$$\partial_{\mu} \equiv \frac{\partial}{\partial x^{\mu}}; \quad \partial_i \equiv \frac{\partial}{\partial x^i}, \tag{3.60}$$

where $\mu = 0, 1, 2, 3$ and $i = 1, 2, 3$.

3.5.1 The Newtonian Limit

Since the geodesic (3.56) describes the trajectory of a particle in a gravitational field, it must reduce to the usual Newtonian formula in the non-relativistic limit. Recalling that the metric is related to the gravitational potential, we define the classical Newtonian limit as that in which, besides the non-relativistic condition $v \ll c$, we also require the gravitational field to be weak and static.

On the metric this implies:

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} + O(h^2), \tag{3.61}$$

$$\frac{\partial g_{\mu\nu}}{\partial t} = c\frac{\partial g_{\mu\nu}}{\partial x^0} = 0, \tag{3.62}$$

where $h_{\mu\nu}(x)$ is the first order deviation from the flat Minkowski space corresponding to the absence of gravitational field. In Appendix B we show that, in this case, the only non-vanishing component of the affine connection is:

$$\Gamma^i_{00} = \frac{1}{2}g^{i\mu}(-\partial_\mu g_{00}) \simeq -\frac{1}{2}\eta^{ij}\partial_j h_{00} = \frac{1}{2}\partial_i h_{00}, \tag{3.63}$$

where i is a three-dimensional space index and we have used $\eta^{ij} = -\delta^{ij}$.

With these approximations the geodesic equation for the index $\mu = 0$ gives $\frac{dt}{d\tau} \simeq 1$, while for the spatial index $\mu = i$ we have:

$$\frac{1}{c^2}\frac{d^2x^i}{dt^2} = -\frac{1}{2}\partial_i h_{00}. \tag{3.64}$$

Equation (3.64) then coincides with the Newton equation if we set:

$$\frac{\phi}{c^2} = \frac{1}{2}h_{00}. \tag{3.65}$$

where ϕ is the classical gravitational potential. With this identification, (3.64) becomes:

$$\frac{d^2x^i}{dt^2} = -\partial_i\phi. \tag{3.66}$$

In particular from (3.61) we find:

$$g_{00} = 1 + h_{00} = 1 + 2\frac{\phi}{c^2}. \tag{3.67}$$

3.5.2 Time Intervals in a Gravitational Field

Equation 3.67 allows us to evaluate how time intervals are affected by the presence of a gravitational field. Let us suppose that we are in a free falling frame, where the Minkowskian coordinates ξ^α can be used. According to the principe of equivalence, a clock at rest in such a system measures a time interval which coincides with the proper time in the absence of gravity:

$$d\tau^2 = \frac{1}{c^2}\eta_{\alpha\beta}d\xi^\alpha d\xi^\beta = \eta_{00}dt^2 = dt^2 \quad (\alpha, \beta = 0, 1, 2, 3), \tag{3.68}$$

since, for a clock at rest, $\frac{d\xi^i}{dt} = 0$.

In any other frame of reference with coordinates x^μ, like our laboratory, the gravitational field is present and the proper time interval will take the following form:

$$d\tau^2 = \frac{1}{c^2}g_{\mu\nu}dx^\mu dx^\nu. \tag{3.69}$$

If in this frame the clock has four-velocity $dx^\mu/dt = v^\mu$, then the time interval dt between two consecutive (infinitely close) ticks satisfies the relation:

$$\left(\frac{d\tau}{dt}\right)^2 = \frac{1}{c^2}g_{\mu\nu}\frac{dx^\mu}{dt}\frac{dx^\nu}{dt}. \tag{3.70}$$

In particular, if the clock is at rest in the laboratory frame, that is if $v^i = 0$, we obtain:

$$\frac{dt}{d\tau} = (g_{00})^{-\frac{1}{2}}. \tag{3.71}$$

The dilation factor on the right hand side of (3.71), however, cannot be observed, since the gravitational field affects in the same way the ticks of the standard clock and those of the clock being studied. However, the difference between dt_1 e dt_2 in two different points $\mathbf{x}_1,\ \mathbf{x}_2$ can be observed; indeed (3.71) implies:

$$dt_1 = d\tau(g_{00}(\mathbf{x}_1))^{-\frac{1}{2}}, \tag{3.72}$$

$$dt_2 = d\tau(g_{00}(\mathbf{x}_2))^{-\frac{1}{2}}, \tag{3.73}$$

so that:

$$\frac{dt_2}{dt_1} = \left[\frac{g_{00}(\mathbf{x}_1)}{g_{00}(\mathbf{x}_2)}\right]^{\frac{1}{2}}. \tag{3.74}$$

In particular, in the classical Newtonian limit, we may use (3.67) and, recalling (3.67), we obtain

$$\frac{dt_2}{dt_1} = \left[\frac{1+2\frac{\phi_1}{c^2}}{1+2\frac{\phi_2}{c^2}}\right]^{\frac{1}{2}} \approx \left(1+2\frac{\phi_1}{c^2}\right)^{\frac{1}{2}}\left(1-2\frac{\phi_2}{c^2}\right)^{\frac{1}{2}} \approx 1-\frac{\phi_2-\phi_1}{c^2} = 1-\frac{\Delta\phi}{c^2}, \tag{3.75}$$

where we have defined $\phi_1 = \phi(\mathbf{x}_1)$ and $\phi_2 = \phi(\mathbf{x}_2)$ and used that $\frac{\phi}{c^2}$ is very small in most situations.[25]

For example, if a clock is placed at the point $\mathbf{x}_1$ far away from other bodies, so that no gravitational field is present, we have $\phi(\mathbf{x}_1) = 0$, and therefore:

$$dt_1 = d\tau, \tag{3.76}$$

since $g_{00}(\mathbf{x}_1) = \eta_{00} = 1$. The same clock placed at a point $\mathbf{x}_2$, e.g., on the earth's surface, will tick time intervals dt_2 such that:

$$\frac{dt_2}{dt_1} = \frac{dt_2}{d\tau} \approx 1-\frac{\Delta\phi}{c^2} = 1-\frac{\phi_2}{c^2}. \tag{3.77}$$

Since $\phi_2 < 0$ we find $dt_2 > dt_1$, that is *time intervals are dilated in a gravitational field* by a factor $(1-\frac{\phi_2}{c^2})$. In the case of the terrestrial gravitational field we have:

$$dt_2 = \left(1+\frac{GM}{c^2 r}\right)d\tau. \tag{3.78}$$

where M is the earth's mass.

In particular a same clock will tick at a different rate, depending on whether it is placed at sea level or at height h. Indeed, since

$$\phi_1 = -\frac{GM}{R}; \quad \phi_2 = -\frac{GM}{R+h}, \tag{3.79}$$

we obtain:

$$dt_2 = \left(1-\frac{\phi_2-\phi_1}{c^2}\right)dt_1 = \left[1-\frac{GM}{c^2}\left(\frac{1}{R}-\frac{1}{R+h}\right)\right]dt_1. \tag{3.80}$$

In general we may say that the more negative the gravitational potential is, (or the greater its absolute value is), the more dilated time intervals are. Because of the factor $1/c^2$, these effects are generally small, however gravitational time dilation has been experimentally measured in various different situations.

The first verification by a experiment on earth was performed by Pound and Rebka combining the gravitational dilation with Doppler effect in the emission and absorption of photons in Fe^{57}. Further experimental evidence can be inferred from astrophysical observations (especially from the light spectra of white dwarfs). In this case we can safely set to zero the earth gravitational potential being much smaller in absolute value than the potential on the surface of a star. Thus, in this case, we have:

---

[25] As usual the value at infinity of ϕ is set equal to zero.

$$\frac{dt_2}{dt_1} = \frac{\nu_1}{\nu_2} > 1, \tag{3.81}$$

where ν denotes the light frequency and the suffix 1 and 2 are referred to the earth and to the star, respectively. The frequency of the emitted light is then higher than that observed on earth and we have a measurable shift towards the red.

In the eighties further confirmations were gained by experiments with time signals sent to and from Viking 1 Mars lander and from time measurements using atomic clocks on airplanes; the clocks that traveled aboard the airplanes, upon return, were slightly faster than those on the ground.

However, the most spectacular evidence of the gravitational red-shift is nowadays given by its technological application to GPS devices.

A GPS gives the absolute position on the surface of the earth to within 5–10 m of precision; this requires the clock ticks on the GPS satellite to be known with an accuracy of 20–30 ns. Such an accuracy cannot be reached if we neglected the special and general relativity effects on time intervals . To compute these effects we first observe that the transmitting clock is subject to the special relativistic time dilation due to the satellite orbital speed, compared to an identical clock on the earth.

From our discussion of time dilation in special relativity the clock on the satellite would run slower, compared to a clock on the earth, by the factor:

$$\sqrt{1 - \frac{v^2}{c^2}} \simeq 1 - \frac{v^2}{2c^2}, \tag{3.82}$$

at first order in $\frac{v^2}{c^2}$, since v, the satellite speed, is much smaller than c.

Suppose the satellite is orbiting at a distance from the center of the earth of about four times the earth's radius R. Using the classical result

$$\frac{v^2}{r} = \frac{GM}{r^2}, \tag{3.83}$$

with $r = 4R$, we obtain a time loss of about -7 µs/day.

We now compute the general relativistic effect. First we recall that a clock in a greater gravitational potential runs *faster* than one on earth. Then, calling Δt_E and Δt_s the time intervals on earth and on the satellite, respectively, we find:

$$1 - \frac{\Delta t_E}{\Delta t_s} = \frac{\Delta\phi}{c^2} = \frac{1}{c^2}\left[\frac{GM}{R+h} - \frac{GM}{R}\right] \tag{3.84}$$

and if $R + h \simeq 4R$ this gives a gain in time of $\simeq 45$ µs/day, six times larger than the special relativistic effect. Summing up the two effects we find that the clock on the satellite runs faster $\simeq 38$ µs/day $\equiv 38 \times 10^3$ ns/day. We see that neglecting the relativistic effects, would imply errors three order of magnitude higher than the necessary accuracy of 20–30 ns.[26]

3.5.3 The Einstein Equation

Until now we have been discussing some consequences of the principle of equivalence in relation to the motion of a particle in a given gravitational field. We have seen that the motion is essentially a *free motion* in a curved space–time, the generalized gravitational potential being described by the metric $g_{\mu\nu}(x)$ which was supposed to be a known function of the space–time coordinates.

The knowledge of the metric field is, however, not known *a priori*, and the description of the gravitational field requires the knowledge of the equation of motion of the metric field. To arrive to a rigorous determination of this equation the principle of equivalence is no more sufficient and we must address the full geometric formalism of general relativity. Nevertheless, in this section, with no ambition of being rigorous or complete, we shall try to develop some heuristic considerations to justify the actual form of the gravitational equations.

We recall that in the Newtonian non-relativistic theory the gravitational potential is static, that is, non-propagating, and, given a distribution of mass in space, it can be determined through the Poisson equation:

$$\nabla^2\phi = -4\pi G\rho(\mathbf{x}), \tag{3.85}$$

where ρ is the density of matter and, as usual, $\phi(\mathbf{x})$ the gravitational potential. On the basis of the discussions given in the previous sections, it is clear that the relativistic extension we are looking for must be *covariant under general coordinate transformations*. Furthermore the ten components of the metric field $g_{\mu\nu}(x)$ must describe a *propagating* field generalizing the single static component of the Newtonian gravitational potential ϕ. Indeed, we know that already at the level of special relativity, dependence on the spatial coordinates implies dependence on the time.

It then follows that the source term ρ must also be generalized, in a covariant setting,[27] in terms of ten quantities depending on the four space–time coordinates. As we will show in Chaps. 5 and 8, such relativistic extension is given in terms of the so-called *energy momentum tensor* $T_{\mu\nu}$, symmetric in $\mu\nu$, which describes the density of matter four-momentum and of its current and can be shown to reduce, to lowest order in *v/c*, to the single non-vanishing component $T_{00} = \rho c = \epsilon/c$, ϵ being the energy density. On the other hand, on the left hand side, the Laplace operator in the classical Poisson equation must be replaced by an expression satisfying the following requirements:

1. It must have the same index structure as $T_{\mu\nu}$;
2. It must be a second order differential expression on the metric field $g_{\mu\nu}(x)$;

---

[26] We also note that the given error increases day by day, since it is a *cumulative* effect.

[27] Here covariant means with respect to general coordinate transformations.

3. It must be covariant under general coordinate transformations;
4. It must reduce to the left hand side of Poisson equation in the non-relativistic limit.

Let us denote such unknown expression by $G_{\mu\nu}$. Then the equation we are looking for should have the form:

$$G_{\mu\nu} = \alpha T_{\mu\nu}. \tag{3.86}$$

where α is a constant. On the other hand our previous discussion on the geometric properties of space–time in the presence of gravitation tells us that the Riemann tensor $R^{\mu}{}_{\nu\rho\sigma}$, the relativistic extension of the tidal forces, determines the geometric properties of space–time associated with the presence of gravitation. Thus we expect that $G_{\mu\nu}$ must be related to the Riemann tensor. Actually one can show that, in Riemannian geometry, all the requirements enumerated before are satisfied if we set:

$$G_{\mu\nu} = R^{\alpha}{}_{\mu\alpha\nu} - \frac{1}{2} g_{\mu\nu} R^{\tau}{}_{\rho\tau\sigma} g^{\rho\sigma}, \tag{3.87}$$

and set the unknown constant α of (3.86) $\alpha = -8\pi G$ in order to reproduce in the non relativistic static weak-field limit the Poisson (3.85). Thus we see that the equation for the metric field, also known as Einstein's equation, is a second order differential equation whose source is $T_{\mu\nu}$, which, as we shall see later on, describes the distribution of energy and momentum in space–time. If such distribution is known and the initial Cauchy data are given, we can determine the solution for the metric field $g_{\mu\nu}$. In other words: *The energy-momentum distribution, that is the content of matter-energy and of its current, determines the geometry of space–time.*

Reference

For further reading see Refs. [1, 11, 12]

! "#$%$ & ' ' "( ) " "! . C + +: ' +

1 ! "$/2" " "3 4 + " ? . + / + ' = + * ! + +
@ * % = % % + 4 ; 1 *S'
A B .+4

1 ! "$/2" " "3 4 + " ? . + / + ' = + * ! + +
@ * % = % % + 4 ; 1 .*

! "#$%$ & ' ' "( ) " "! . C + +: ' +

1 ! "$/2" " "3 4 + " ?B + = F ' + 9
! . C + +: $.. + ( ) x ; 1 .*
>4 ?B; 1 ! +. B ! . C 7 4 + + * !+ =P %
A ! '
! '
H : 4 := + 4 . . / + 4 +

1 ! "$/2" " "3 4 + " ?B + = ;+ 1
@ *+ + + + $ + = ( ) x (+ ; 1 .*

1 ! "$/2" " "3 4 + " ?B + =F % 4+ +F
2 + + +.. 4 + * +.. + 7H2! $ B &+ + : 4
9 ; 1 .*
>4 ?; 1 % =. : + * 4 =
3:
>4 ; 1 % *) + =. : . +
H + + +*) + 4 + + +
. *) # + € : +.

$$z \mapsto \frac{az + \bar{c}}{cz + \bar{a}}, \quad a, c \in \mathbb{C}, \quad |a|^2 - |c|^2 = 1,$$

E '+

$$z \mapsto e^{i\theta} \frac{z - a}{1 - \bar{a}z}, \quad \theta \in \mathbb{R}, \quad a \in \mathbb{D}.$$

B1 ! "$/2" " "3 4 + " ?B ' + ! . C '+ +:
•‚ƒ„…†‡‡ˆ‚‰†ƒ‚Š‹Œ•Ž‚••„‘Œ“•…”•———˜™‡š›
A B .+4

! "#$%$ & ' ' "( ) " "! . + + = +
! + + + * 2 + =

1 ! "$/2" " "3 4 + " ? ! . + + . = +
+ + + * E + = . 4 ; 1 .*

1 ! "$/2" " "3 4 + " ? ! . + + = +
! + + + * 2 + = * * . 4 ; 1 .*

! "#$%$ & ' ' "( ) " "# + 8 = ;
, F 1

1 ! "$/2" " "3 4 + " ?
+ ' 6( )6# + 8 = .*
3: , ' * <<<

>4 ; 1 ! +. ! ' +

>4 ;B?1 ! +. ! ' .+

>4 B; 1 ! +. + .+

! "#$%$ & ' ' "( ) " "@ / + # + 8 = + * / + ' =

?1 ! "$/2" " "3 4 + " ? ! + 8 = 0 . +
+ + + = ; 1 .*
A .+4
>4 ; ?1 & 0 * + 8 =
3: .+4

!
7 *
3 :+ + * +C 4 ?
7 * ?
(+ ?
9 : D +C + 4 =
0E '+ 0 *D .+ + . +
B % +C =. : 4 =
. + 4 =
0 * + 4 + : + * + =. :
+ * . + 4 + ?
!
% + . +
7 *
+ 4 =
+ D % + * + 4 =
$ 4 + * G + 4 =
B % + 4 + E +* + + + 4 =
K +==+ + E +* + + B
S + 4 = B
% =. : . + BB
:+ . . =. : 4 = BB
3 E +* + + =. : 4 = B
% + 4 + E +* + + =. : 4 = B?
0E * +
B 8 + E +* + +
B $ +
B 7 *
B $ 4 + * G =. : 4 =
B % + +
B # 0 * + 4 =
B B # 0 * + 4 =

% 4 = ?
7 * ?
% 0 = ?
% + E + œ C ! = ! œ C œ = # œ C œ = ?B
= +4 + D ?
B % 4 = + + : + = + 4
8 + + 4 + E +* + +
. + 4 =
% 0 = . + 4 =
+ + B
7 * B
+ + =. : 4 =
+ +: + * = B
9 = + + * +.. +
B % + E + C = # C ! =
, :+ ')=D + 4 .+ + B
* ?
7 * ?
$ 0 * + * =. : . + ?
$ * + 4 + 4 : + B
? % + + 4
? 7 *
? $ * 4 B
/

3: 7 . + f <<<

? 1 ! "$/2" " "3 4 + " ? 0 4 += & 0 * +
4 = 0 . + + .*
>4 --7 ; 1 4
$ C .
r =. \ *+ + * + 4[[I œ
9 + [
3: , ' * <<<
7 . + + . *+ 4 + 0 * +
A .+4

? 1 ! "$/2" " "3 4 + " ? 0 4 0 += & 0 * +
8 = 0 . + + + + = ; ?1 .*

? 1 ! "$/2" " "3 4 + " ? @ / + # +
8 = + * / + ' = . 4 ; 1 .*
3: , ' * <<<

? 1 ! "$/2" " "3 4 + " ? @ / + * +
4 = + * + ' = . 4 ; 1 .*

? 1 ! "$/2" " "3 4 + " ? 8[[[* 4 .
, 4 . =. : E * 8 ' . 4 +4 (9 * *S'

?B1 ! "$/2" " "3 4 + " ? 8[[[* 4 .
, 4 . =. : E * 8 ' . 4 +4 (9 * .*

? 1 ! "$/2" " "3 4 + " ? 8 = 9 = . 4
&+ ; ?1 . :
A .+4
3: , ' * <<<

? 1 ! "$/2" " "3 4 + " ? 8 = 9 = . 4
&+ ; ?1 .*

? 1 ! "$/2" " "3 4 + " ? 9+ *:) % % =
0 . + + + + = ; ?1 .*

??1 ! "$/2" " "3 4 + " ? 9+ *:) % % =
0 . + + + + = ; ?1 .*

1 ! "$/2" " "3 4 + " ? 9+ *:) % % =
0 . + + + + = ; ?1 .*

1 ! "$/2" " "3 4 + " ? 9+ *:) % % =
0 . + + + + = ; ?1 .*

1 ! "$/2" " "3 4 + " ? 9+ *:) % % =
B 0 . + + + + = ; ?1 .*

1 ! "$/2" " "3 4 + " ? 9+ *:) % % =
0 . + + + + = ; ?1 .*

1 ! "$/2" " "3 4 + " ? 9+ *:) % % =
0 . + + + + = ; ?1 .*

B1 ! "$/2" " "3 4 + " ? 7 % +* %
8 = + * % . 4= . 4 ; 1 . :

1 ! "$/2" " "3 4 + " ? 7 % +* %
8 = + * % . 4= . 4 ; 1 .*
3: , ' * <<<
>4 B B ! +. B $ * 8 = + * % z % =
A 7 4+MN ' + *+ MN r9+ *:) % % =I

1 ! "$/2" " "3 4 + " ? 7 % +* %
8 = + * % . 4= . 4 7 + + : 46 . 4 ; 1 .*

1 ! "$/2" " "3 4 + " ? .+ ! '+
. + .*

?1 ! "$/2" " "3 4 + " ? + 4 = 0 . +
+ + + = ; 1 .*
A &) + 7 , :+ ')=

1 ! "$/2" " "3 4 + " ? % . 4= + * 8 = $
! 0 += # * + * +* 8 % + ' 0 . + + + + = ; 1 .*
A .+4

! "#$%$ & ' ' "( ) " "8 #

1 ! "$/2" " "3 4 + " ?? 8 4 & @ *)
6 + + * @+ = !+ : *4 J ' = ; ?? 1 *S'

1 ! "$/2" " "3 4 + " ?? 8 4 & @ *)
6 + + * @+ = !+ : *4 J ' = ; ?? 1 .*

1 ! "$/2" " "3 4 + " ?? 8 4 & @ *) 0
7 ' 4 @ D 8 % + + 6 % + F + . F +4
< H * ; 1 . :

1 ! "$/2" " "3 4 + " ?? 8 4 & @ *) 9 4 *
6 4 4 + * % 4 !+ : *4 J ' = ; 1 .*

B1 ! "$/2" " "3 4 + " ?? 8 4 & @ *) ;$ 1
+ 9 4 * # 6 % @ *< $ K 6!/! ; 1 .*

1 ! "$/2" " "3 4 + " ?? 9+ = , *4 F 8 4 @ *)
/ + + . : 4 * + * ' # ' : +
; ? 1 *S'

1 ! "$/2" " "3 4 + " ?? 9+ = , *4 F 8 4 @ *)
/ + + . : 4 * + * ' # ' : +
; ? 1 .*

! "#$%$ & ' ' "( ) " "7 ' , + : 9 :
.+ + * K = ' '+: =

1 ! "$/2" " "3 4 + "
7 ' 6, + 6 : 6 69 : 6 .+ 6+ *6 6K = '6 '+: = .*

! "#$%$ & ' ' "( ) " " +

?1 ! "$/2" " "3 4 + " @ *+ * + > +
0 + ! S @ Ma ' .*

1 ! "$/2" " "3 4 + " @ *+ * + > +
0 + , 4+ ' .*

1 ! "$/2" " "3 4 + " @ *+ * + +
0 + % 4 + ' .*

1 ! "$/2" " "3 4 + " @ *+ * + +
0 + E + + U # + + ' .*

1 ! "$/2" " "3 4 + " B @ *+ * + > +
0 + ! : + j + :+: *+* ' B .*

1 ! "$/2" " "3 4 + " @ *+ * + > +
0 + ! . C • 0E +Ma ' .*

B1 ! "$/2" " "3 4 + " @ *+ * + > +
0 + 8 + $ + f + ' .*

1 ! "$/2" " "3 4 + " @ *+ * + > +
0 + , # '+*+ & Ma * 7 4 + ' .*

1 ! "$/2" " "3 4 + " ? @ *+ * + > +
0 + 8 + + + ' ? .*

1 ! "$/2" " "3 4 + " @ *+ * + > +
0 + 8 + 0 .+ + MN [+ ' .*

?1 ! "$/2" " "3 4 + " @ *+ * + > +
0 + + + + ! + + + + @ + + 0 + f + # '+ ' .*

1 ! "$/2" " "3 4 + " , U / : #+ .*

1 ! "$/2" " "3 4 + " , U / : #+ .*

1 ! "$/2" " "3 4 + " , U / : #+ .*

1 ! "$/2" " "3 4 + " + > + + + '+ .*

1 ! "$/2" " "3 4 + " + > + ! ~ +
$. +Ma .*

B1 ! "$/2" " "3 4 + " + > + ! ~ +
$. +Ma .*

1 ! "$/2" " "3 4 + " + > + ! +
0* + @%# .*

1 ! "$/2" " "3 4 + " + > + ! +
0* + @%# .*

1 ! "$/2" " "3 4 + " + > + ! +
0* + @%# .*

?1 ! "$/2" " "3 4 + " + > + ž 8
7 UU + Ÿ 0* * + .*
3: , ' * <<<
A .+4

! "#$%$ & ' ' "( ) " " + [+ E * / +

1 ! "$/2" " "3 4 + " ;# ' ( ) + + 1
F 6 / + F (+ *6 / + F (+ * (= * 8 =6 ! + +.
/ + 0 # ' : + ; 1 *S'

1 ! "$/2" " "3 4 + " ;# ' ( ) + + 1
F 6 / + F (+ *6 / + F (+ * (= * 8 =6 ! + +.
/ + 0 # ' : + ; 1 .*

1 ! "$/2" " "3 4 + " ; [* + [+ E 1 0
$ + F + * 7 [4 + * / + 6 [. + E 6 +' C 4[
9 + ; ?1 *S'

1 ! "$/2" " "3 4 + " ; [* + [+ E 1 0
$ + F + * 7 [4 + * / + 6 [. + E 6 +' C 4[
9 + ; ?1 .*

1 ! "$/2" " "3 4 + " ; * ( ) x ! + 1
,+ 4 U ;+ 1 (+ * / + 6 % 4 ! .
+ + ( ) x (+ ; ?? 1 *S'

B1 ! "$/2" " "3 4 + " ; * ( ) x ! + 1
,+ 4 U ;+ 1 (+ * / + 6 % 4 ! .
+ + ( ) x (+ ; ?? 1 .*

1 ! "$/2" " "3 4 + " (+ * / + ! *
+. K *) ; 1 *S'
A ! * +.

1 ! "$/2" " "3 4 + " (+ * / + ! *
+. K *) ; 1 .*

1 ! "$/2" " "3 4 + " 9 K F + * 0 4
% : U Q @ * + / + + 7 8 $ + + + + =F
! * / + + E ; 1 .*

! "#$%$ & ' ' "( ) " " + +

?1 ! "$/2" " "3 4 + "
+: + + 6.+ 6 : 6 6 6 * .*
A $: 0
A B B .+4

B 1 ! "$/2" " "3 4 + " 0 P .+ +
* . + + * 4 + + ' = ; #@# ' 1 .*

B 1 ! "$/2" " "3 4 + " 9 4 # F + +
J ' + * 0C + * + = + + % = 0' = 4 ; #@# ' 1 .*

B 1 ! "$/2" " "3 4 + " 9=. .+ $ 3*= =
4 + + J ' F % H+ . F + * % # ; #@# ' 1 .*

B 1 ! "$/2" " "3 4 + " .+ F % + * 0
; #@# ' 1 .*

B 1 ! "$/2" " "3 4 + " % @+: ! .+ F
% F + * % C / + = ; #@# ' 1 .*

BB1 ! "$/2" " "3 4 + " % 9 ** / + = + +
J ' + * # . ,+ ! ; #@# ' 1 .*

B 1 ! "$/2" " "3 4 + " % / +* / + = $! .
8 * ,+ J ' ; #@# ' 1 !j. + .*

B 1 ! "$/2" " "3 4 + " % / +* / + = $! .
8 * ,+ J ' ; #@# ' 1 .*

B 1 ! "$/2" " "3 4 + " % %+ = $ 0C. +
+ + : * = + * 0+ = ; #@# ' 1 .*
3: , ' * <<<

! "#$%$ & ' ' "( ) " "/ + H)

B?1 ! "$/2" " "3 4 + " ? B *S'

1 ! "$/2" " "3 4 + " ? B 6: .*

1 ! "$/2" " "3 4 + " : + * 4 + .*

1 ! "$/2" " "3 4 + " 4 + + .*

1 ! "$/2" " "3 4 + " 4 + + .*

1 ! "$/2" " "3 4 + " K+ +::+4 # / + P
R $ + ( ) ; 1 .*

B1 ! "$/2" " "3 4 + " ' + + .*
A 3 ' + [+ E * / +
A B .+4

! "#$%$ & ' ' "( ) " "%+ =

1 ! "$/2" " "3 4 + " B !+. + @ % + . =
9+ . ! ; ?? 1 *S'

1 ! "$/2" " "3 4 + " B !+. + @ % + . =
9+ . ! ; ?? 1 .*

—¡˜…¢£¤¥¦§¤¨•©Š’•‰“¤ªŠˆ’«“•–¨—••••–••–•¨¬¤-•,®,‹†ˆ¤¨–¯•••°•,ƒ±Š›…¢†‡•†…•…¥…‰,²‹‰,†…š“…
³“Š‹†•šŠ…š†…ª,‹‰,…•…’«…«“•®’ˆ„Š…‡•Š›’‹šŠ…‹†…«“‹ƒ“…šŠ…®•†‹š“…®²‹,Š…š†…¦“‹†Œ‰“‹´†•µš
! C ; 1 .*

?1 ! "$/2" " "3 4 + " B @ S !+. + 0 ! +
% E 0* + ; ?? 1 .*

¶–˜…¢£¤¥¦§¤¨•©Š’•‰“¤ªŠˆ’«“•–¨—••••–••–•¨¬¤-•,®,‹†ˆ¤¨–¯•••°•,ƒ±Š›…¢†‡•†…•…-…ƒ†Š…š†…›·Œ,‰
‡†•†ˆ“ˆŠ…“‹ƒ•“…†…›·Œ,‰†…«Šš“•‹†…“…Š…«,Œƒ,‰,Œ«Š…Š•,“‹ƒ†ˆ•¢’ˆƒ•,¸™“‡’¹

1 ! "$/2" " "3 4 + " B @ S !+. + 3 + *+ f +
.+ + + + * + .*

1 ! "$/2" " "3 4 + " B @ S !+. + %$3 3@ 9O 7!
0# + : + + ; ?? 1 *S'

1 ! "$/2" " "3 4 + " B @ S !+. + %$3 3@ 9O 7!
0# + : + + ; ?? 1 .*

1 ! "$/2" " "3 4 + " B @ S !+. + % %+ =
+ + $ * ; ?? 1 *S'

B1 ! "$/2" " "3 4 + " B @ S !+. + % %+ =
+ + $ * ; ?? 1 .*

! "#$%$ & ' ' "( ) "

? B?

?

! "#$%$ & ' ' "( ) "

! "#$%$ & ' ' "( ) " "$, ' + + 3!/ $
* ( )

1 ! "$/2" " "3 4 + " $, ' + + 3!/ $
* ( ) .*

! "#$%$ & ' ' "( ) " "! $ + + + 0*

1 ! "$/2" " "3 4 + " ! $ + + + 0*
.*
3: , ' * <<<

>4 ; 1 ! + ' '

>4 BB;? 1 ? . + % + +

>4 ;? 1 % .+ + +C

It follows that

$$z = x + iy = \rho(\cos\varphi + i\sin\varphi).$$

If now z_1 and z_2 are two complex numbers,

$$z_1 = \rho_1(\cos\varphi_1 + i\sin\varphi_1),$$
$$z_2 = \rho_2(\cos\varphi_2 + i\sin\varphi_2),$$

their product $z_1z_2 = \rho_1\rho_2(\cos(\varphi_1+\varphi_2) + i\sin(\varphi_1+\varphi_2))$, as is easily seen by direct multiplication. Hence, to multiply two complex numbers, you multiply their moduli and add their arguments.

For division, let $z = z_1/z_2$, ($z_2 \neq 0$), so that

$$z_1 = zz_2.$$

Using the first rule, it follows that the quotient of two complex numbers has, for modulus, the quotient of their moduli, and, for argument, the difference of their arguments.

4. Circles as level curves

Let a, b, z be any three complex numbers. We denote by (zab) the ratio

$$\frac{z-a}{z-b} = (zab).$$

The modulus of this complex number is the quotient of the moduli of $z - a$ and $z - b$. It is therefore the quotient

az[†]$/bz$.

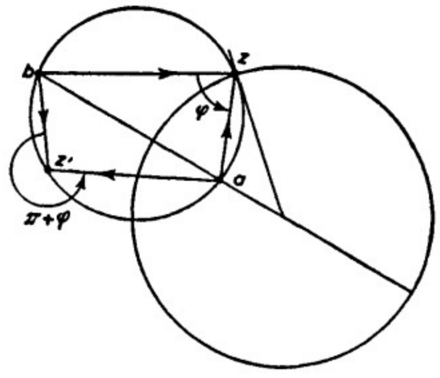

Fig. 49

[†] az denotes the distance between the points a and z in the Argand diagram.

As for its argument, by the above it is the angle through which the vector $z - b$ must be turned to give it the direction of the vector $z - a$. It is therefore the angle between the vectors $\overrightarrow{az}$ and $\overrightarrow{bz}$, oriented as in Fig. 49.

If we regard a and b as fixed, and z as variable, we may enquire into the nature of the curves along which mod (zab) and arg (zab) are respectively constant. These are the *level curves* of the respective functions. Both turn out to be circles.

If we have $az/bz =$ constant, the ratio of the powers of the point z with respect to the point circles a and b is constant. Therefore z moves on a circle of the coaxal system determined by the point circles a,b. This coaxal system has these point circles as limiting points.

If the angle between $\overrightarrow{az}$ and $\overrightarrow{bz}$ is constant, z moves on an arc of a certain circle passing through a and b. The other arc, the arc complementary to the first, corresponds to an argument $\varphi + \pi$. If φ takes all possible values, we obtain all circles of the coaxal system through a and b, including the radical axis ab, when $\varphi = \pi$ (this gives points between a and b) and when $\varphi = 0$ (this gives points outside ab).

For all values of the modulus and argument of (zab) we therefore obtain all circles of two conjugate coaxal systems as level curves.

5. The cross-ratio of four complex numbers

Let a, b, c, d be four complex numbers. By definition, the cross-ratio of the four numbers in the given order is the ratio $(acd)/(bcd)$. We write

$$\frac{(acd)}{(bcd)} = (ab,cd),$$

so that

$$(ab,cd) = \frac{(a-c)/(a-d)}{(b-c)/(b-d)}\,.$$

This is formally the same definition as that which occurs in projective geometry.

The *modulus* of the cross-ratio is equal to the quotient of the moduli of (acd) and (bcd). Hence, for this modulus to be equal to

1, it is necessary and sufficient for (acd) and (bcd) to have the same modulus. By what we have seen above, a and b must therefore lie on the same circle of the coaxal system which has c,d as limiting points.

The *argument* of the cross-ratio is equal to the difference of the arguments of (acd) and (bcd). The cross-ratio therefore has zero argument if (acd) and (bcd) have the same argument, which is the case if a and b are on the same arc of a circle bounded by c and d.

The argument of the cross-ratio is equal to π if the arguments of (acd) and (bcd) differ by π, which is the case if a and b are on complementary arcs of a circle through c and d.

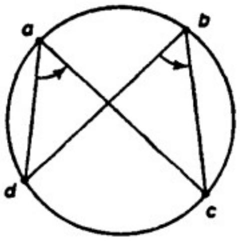

FIG. 50

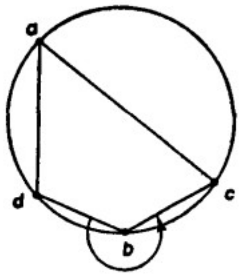

FIG. 51

Now, a number is real if its argument is 0 or π. Hence we have the theorem:

For the cross-ratio (ab,cd) *to be real, it is necessary and sufficient that the four points* a, b, c, d *should lie on a circle (or straight line).*

More generally, we find the locus of z such that the cross-ratio (za,bc) has (i) a constant modulus; and (ii) a constant argument.

Since $(za,bc) = (zbc)/(abc)$, the modulus and argument of (za,bc) will be constant if the respective modulus and argument of (zbc) are constant.

(i) The locus of points z such that mod (zbc) is constant is a circle of the coaxal system with limiting points b and c.

(ii) The locus of points z such that arg (zbc) is constant is an arc of a circle limited by b and c.

We deduce that if a, b, c are three distinct complex numbers, u any given complex number, then *there is a unique number* z *such that*

$$(za,bc) = u.$$

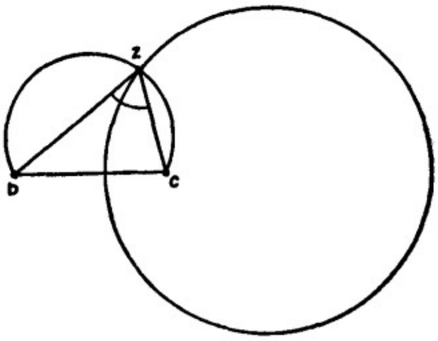

FIG. 52

For u has a definite modulus and argument, and z is given as the unique intersection of a circle and an arc of a circle, as shown in Fig. 52.

This important property is worth verifying algebraically. If $(za,bc) = u$, then we must have $(zbc) = (abc)u$. Write $(abc)u = k$. Then $(z - b)/(z - c) = k$, which gives

$$z = \frac{b - kc}{1 - k}.$$

We conclude this section by giving an interesting geometrical interpretation to the cross-ratio (ab,cd) of any four complex numbers. If ψ_1 is the argument of (acd), and ψ_2 that of (bcd), we know that the argument φ of (ab,cd) satisfies the equation

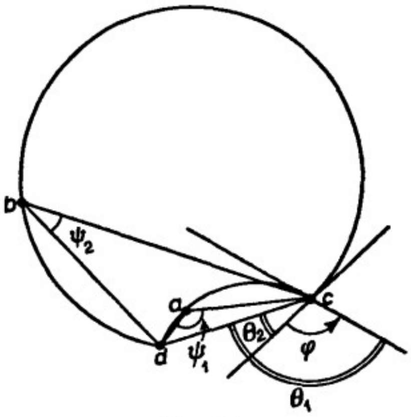

FIG. 53

$$\varphi = \psi_1 - \psi_2.$$

Draw the circles through a, c, d and b, c, d respectively, and let Θ_1, Θ_2 be the angles made by cd with the tangents at c to these circles. Then by the alternate segment theorem, $\psi_1 = \Theta_1$ and $\psi_2 = \Theta_2$, so that $\varphi = \Theta_1 - \Theta_2$. Hence φ is the angle between the two circles, and we have the theorem:

The argument of the cross-ratio (ab,cd) *is equal to the angle between the two circles passing through* a, c, d *and* b, c, d *respectively.*

6. Möbius transformations of the z-plane

These are transformations of the plane of the complex variable which map any point z on a point w according to the formula

$$w = \frac{Az + B}{Cz + D}, \tag{1}$$

where A, B, C, D are four complex constants, and $AD - BC \neq 0$.

We exclude the case $A/C = B/D$, because then $w = A/C$ for all values of z, and all points z transform into a single point w.

The reader is probably already familiar with transformations of type (1) in projective geometry, where they are called *bilinear* or *projective* transformations, and transform points of a line. The same algebra proves the theorem:

A Möbius transformation leaves cross-ratios invariant.

In other words, if a, b, c, d transform respectively into a', b', c', d', then $(ab,cd) = (a'b',c'd')$.

We make some important deductions from this theorem.

I. If $w = \dfrac{Az + B}{Cz + D}$, $(AD - BC \neq 0)$ is a Möbius transformation, and (a,a'), (b,b'), (c,c') are three pairs of corresponding points, then the equation of the transformation may be written in the form

$$(wa',b'c') = (za,bc). \tag{2}$$

This follows immediately from the theorem. If z is any point, and w the corresponding point under the Möbius transformation, then the pair (z,w) satisfy (2). On the other hand we know that, z having been given, there is a unique w which can make the cross-ratio $(wa',b'c') = (za,bc)$. Hence the equation (2) does, in fact, map z on the w given by the Möbius transformation.

II. Conversely, let a, b, c be three distinct complex numbers, and a', b', c' a second set of three distinct complex numbers. Then the equation

$$(wa',b'c') = (za,bc)$$

is the equation of a Möbius transformation which maps a on a', b on b' and c on c'.

This follows immediately by writing out both sides of the equation, and substituting $z = a, b, c$ in turn. We find that $w = a'$, b', c' in turn.

III. A Möbius transformation transforms circles into circles. (It is understood that straight lines come into the category of circles.)

We shall give two proofs of this important theorem.

In the first place let z be any point of the circle determined by the three points a, b, c. Then the cross-ratio

$$(za,bc) = \text{real number}.$$

Let z' be the transform of z. Then

$$(z'a',b'c') = (za,bc) = \text{real number},$$

and therefore z' lies on a circle through a', b', c'.

The second proof will be given in the next section. We conclude this section with:

IV. A Möbius *transformation* conserves the angle of intersection of two circles.

Let b and c be the points common to two circles $\mathscr{C}_1$ and $\mathscr{C}_2$, a_1 a third point on $\mathscr{C}_1$ and a_2 a third point on $\mathscr{C}_2$. We have seen (see page 50) that the angle between the two circles is given by the argument of the cross-ratio (a_1a_2,bc). But this argument is conserved, since the cross-ratio itself is conserved.

A second proof of this theorem will also be given later.

7. A Möbius transformation dissected

The reader may ask, at this stage, whether there is any connection between Möbius transformations and inversion. We show that there is.

We may write

$$w = \frac{Az + B}{Cz + D} = \frac{A}{C} + \frac{BC - AD}{C(Cz + D)},$$

by simple division, or

$$w = \frac{A}{C} + \frac{BC - AD}{C^2} \cdot \frac{1}{z + D/C}.$$

We now carry out the transformation on z in several stages:

(i) Let $$z_1 = z + \frac{D}{C};$$

(ii) Let $$z_2 = \frac{1}{z_1};$$

(iii) Let $$z_3 = \frac{BC - AD}{C^2} . z_2;$$

(iv) Then $$w = \frac{A}{C} + z_3.$$

Each stage can be interpreted geometrically, and this we proceed to do.

Stage (i) is equivalent to a *translation* of z in a definite direction and through a definite distance, given by the vector D/C.

Stage (ii) is equivalent to inversion in the unit circle, followed by reflection in the real axis. For with an obvious notation,

$$r_2(\cos \Theta_2 + i \sin \Theta_2) = \frac{1}{r_1(\cos \Theta_1 + i \sin \Theta_1)}$$
$$= \frac{1}{r_1} (\cos (-\Theta_1) + i \sin (-\Theta_1)),$$

and therefore $r_2 = 1/r_1$, so that $r_1 r_2 = 1$, and
$$\Theta_2 = -\Theta_1.$$

Stage (iii), *multiplication* by a fixed complex number is equivalent to a dilatation by a fixed amount from the origin, followed by a definite rotation about the origin. For if $(BC - AD)/C^2 = \rho(\cos \Theta + i \sin \Theta)$,

$$r_3 = \rho r_2,$$

and $$\Theta_3 = \Theta_2 + \Theta.$$

Stage (iv) is another translation.

We see that after each one of these sub-transformations circles are still circles (or straight lines), and that the angle of intersection of two circles is unchanged. Hence we have additional proofs of two of the theorems of the preceding section.

Another important deduction we can make from our dissection of a Möbius transformation is the following:

If a circle $\mathscr{C}$ is transformed into a circle $\mathscr{C}'$, the interior of $\mathscr{C}$ is mapped on the interior or the exterior of $\mathscr{C}'$.

We mean, of course, that every point inside $\mathscr{C}$ is mapped on points inside $\mathscr{C}'$, or every point inside $\mathscr{C}$ is mapped on points outside $\mathscr{C}'$.

The only transformation which can turn a circle inside-out is (ii), since this is an inversion followed by a reflection. The inverse of a circle with respect to a centre of inversion inside the circle turns the circle inside out. The other transformations which come into the dissection preserve insides and outsides.

If the circle $\mathscr{C}'$ is a straight line, the inside of $\mathscr{C}$ is mapped on one of the half-planes bounded by the line, the outside of $\mathscr{C}$ on the other.

8. The group property

The Möbius transformations of the z-plane form a group. Before we prove this, we give the definition of a *group of transformations*.

A set of transformations is said to form a group if it has the following properties:

(a) Every transformation T of the set possesses an inverse transformation which is also in the set. If this inverse transformation be denoted by T', the transformation effected by carrying out first T and then T' is the identical transformation, which leaves everything unchanged.

(b) If two transformations T_1, T_2 both belong to the set, the transformation obtained by carrying out first T_1 and then T_2 also belongs to the set.

That the Möbius transformations of the z plane form a group is a theorem easily proved by algebra. Alternatively we can proceed as follows:

The transformation inverse to

$$(za,bc) = (z'a',b'c'),$$

that is, the one which maps z' on z, is given by the equation

$$(z'a',b'c') = (za,bc).$$

Now let M_1 be a Möbius transformation which maps a on a', b on b', c on c'. It is determined by these three pairs of corresponding points. Let M_2 be a Möbius transformation which maps a' on a'', b' on b'' and c' on c''. It is also determined by these three pairs of corresponding points.

M_1 is given by the equation $(z'a',b'c') = (za,bc)$;

M_2 is given by the equation $(z''a'',b''c'') = (z'a',b'c')$.

The result of M_1 followed by M_2 is the transformation

$$(z''a'',b''c'') = (za,bc),$$

which is a Möbius transformation.

If a set of transformations, all of which belong to a group G of transformations, themselves form a group, we call this a *subgroup* of the group G. An important example of a group of transformations is afforded by the group of displacements in a Euclidean plane.

We imagine a fixed plane, and another like a large sheet of paper covering it. The paper is moved, with any possible figures on it, from one position over the plane to another. We consider only the initial and final positions of a motion or displacement. The path does not concern us. Suppose A, B, points of the plane in its first position, become A', B' in its final position. Join AA', BB', and let their perpendicular bisectors PX, PY meet in P. Evidently the triangles $PAB, PA'B'$ are congruent. Hence the displacement envisaged can be effected by a rotation round P.

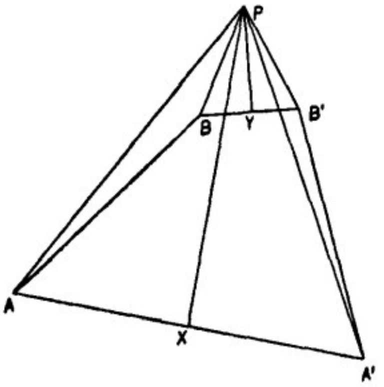

FIG. 54

If the right bisectors were parallel, then AA', BB' would be parallel, and the displacement would be a *translation*, every point of the plane moving through the same distance and in the same direction.

If we follow a given displacement by another, the result is a displacement, since only initial and final positions are considered. The inverse of a given displacement is merely that displacement which brings the moving plane back to its original position. Hence the displacements, or Euclidean motions of a plane, form a group.

The set of translations of the plane evidently form a subgroup of the group of displacements.

We may also regard the displacement of one plane over another as a *mapping* of the points of a Euclidean plane into itself which is one-to-one, transforms lines into lines, preserves angles, and also orientation. It is this notion which is the basis of *applicability* in Euclidean geometry. In the next section we investigate a group of transformations which, in our non-Euclidean geometry, corresponds to the group of Euclidean motions.

9. Special transformations

We are interested in a certain subgroup of the group of Möbius transformations, namely that subgroup which consists of all Möbius transformations which map the inside of a given circle on itself. Let us be more explicit.

We choose the circle with centre at the origin and radius $= 1$. We call the circumference ω, and the circular area bounded by ω we call Ω. The points of ω are not included in Ω.

There are Möbius transformations which transform Ω into itself. We shall call these M-transformations. For example:

the identical transformation $w = z$;
the transformation $w = (\cos\varphi + i\sin\varphi)z$;

this gives a rotation of the plane, through an angle φ, about the origin.

If an M-transformation maps Ω on itself, the dissection of §7 shows that the boundary ω is also mapped on itself. But the converse, as we know from the simple example $w = 1/z$, is not true.

It is easy to see that M-transformations form a group, which is a subgroup of the group of all Möbius transformations. For, if a transformation maps Ω on itself, this is also true for the inverse transformation; and if two transformations M_1, M_2 both map Ω on itself, this is also true for the transformation obtained by first carrying out M_1 and then M_2.

We shall now restrict ourselves to these M-transformations and to Ω, the interior of the unit circle, centre at the origin. We shall show that *Ω can be regarded as a non-Euclidean space, in which the M-transformations are the group of non-Euclidean motions.*

10. The fundamental theorem

We recall one of the characteristic properties of the group of displacements of a plane (over itself). Let A be any point in the plane, α a direction through it, X any other point in the plane and

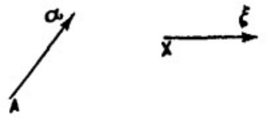

FIG. 55

ξ a direction through it. Then there is a unique displacement of the plane which maps A on X and α on ξ.

We show that the M-transformations have the same property in Ω. The theorem to be proved is:

Let X and A be any two points of Ω, *ξ and α directions through X and A respectively. Then there is a unique M-transformation which maps* (X,ξ) *on* (A,α).

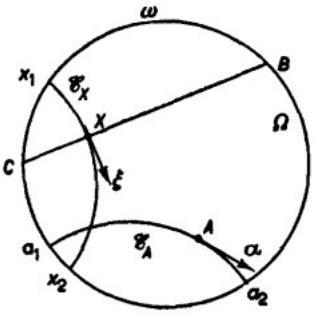

FIG. 56

We note in the first place that, through any point X, and touching any given line through the point, there passes exactly one circle orthogonal to ω. This is immediate by inversion, or may be proved as follows: draw the perpendicular at X to the given line, and let B, C be the points in which it cuts ω. Among the circles of the pencil through B and C, two are evidently orthogonal to the circle sought for: ω itself and the line BC. Hence the circle required is of the coaxal system conjugate to that defined by ω and the line BC. It therefore has B and C as limiting points, and is uniquely determined, since it passes through X.

If we care to regard it as an Apollonius circle, it is the locus of a point which moves so that the ratio of its distances from two fixed points B and C is always equal to XB/XC.

Call this circle $\mathscr{C}_X$, and let x_1 and x_2 be its intersections with ω. In the same way, there passes through A a unique circle $\mathscr{C}_A$ orthogonal to ω and tangent at A to the direction α. Let a_1, a_2 be its intersections with ω.

Now, since M-transformations map ω on itself, they necessarily map circles orthogonal to ω on to circles orthogonal to ω, being Möbius transformations.

Hence any M-transformation which maps (X,ξ) on (A,α) must necessarily map $\mathscr{C}_X$ on $\mathscr{C}_A$, since $\mathscr{C}_X$ will be mapped on a circle through A, with direction α, and orthogonal to ω, and there is only one such circle, namely $\mathscr{C}_A$.

Let us now suppose that the points of intersection of $\mathscr{C}_X$ and $\mathscr{C}_A$ with ω are so numbered that the direction ξ is that of x_1 towards x_2 in Ω, and the direction α is that of a_1 towards a_2 in Ω. The M-transformation we are looking for must map x_1 on a_1, x_2 on a_2 and X on A. These three pairs of corresponding points determine a unique Möbius transformation. Our proof will therefore be complete if we show it is an M-transformation, mapping Ω into itself. This is clear, since the unique Möbius transformation maps the circle through x_1 and x_2 orthogonal to $\mathscr{C}_X$, that is ω, on the circle through a_1 and a_2 orthogonal to $\mathscr{C}_A$, which is again ω. It therefore maps ω on ω. On the other hand it maps Ω on Ω, since it maps the interior point X of Ω on the point A of Ω. The unique Möbius transformation is therefore an M-transformation.

Corollary: Let a and b be two points of Ω, $\mathscr{C}_a$ a circle orthogonal to ω passing through a, $\mathscr{C}_b$ a circle orthogonal to ω passing through b. There exist precisely *two* M-transformations mapping a on b and $\mathscr{C}_a$ on $\mathscr{C}_b$. To a given orientation on $\mathscr{C}_a$ the one transformation maps one of the possible directions of motion on $\mathscr{C}_b$, and the second transformation maps the other.

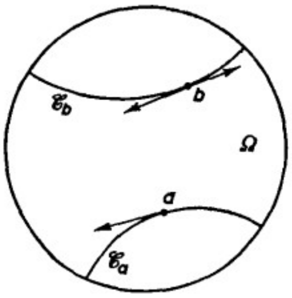

FIG. 57

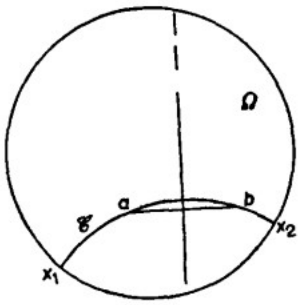

FIG. 58

In particular, let a, b be two points of Ω. There exists a unique M-transformation which permutes a and b (that is, which maps a on b and b on a).

To prove this, we note in the first instance that there exists a unique circle orthogonal to ω and passing through a and b. Call this circle $\mathscr{C}$. An M-transformation interchanging a and b must transform $\mathscr{C}$ into itself.

By the above, there are only two M-transformations which map a on b and $\mathscr{C}$ on itself. The one which conserves orientation on $\mathscr{C}$ cannot map b on a, since then the sense of a towards b in Ω would be transformed into the sense of b towards a in Ω.

Let M' therefore be the M-transformation which maps a on b and alters the orientation of $\mathscr{C}$. We show that M' maps b on a.

Let x_1 and x_2 be the intersections of $\mathscr{C}$ and ω, and a' the transform of b by M'. We want to prove that $a' = a$. Since M' changes the orientation of $\mathscr{C}$, x_1 is mapped on x_2, x_2 on x_1. Since a is also mapped on b, and b on a', the cross-ratio property gives

$$(x_1x_2,ab) = (x_2x_1,ba').$$

But, for any cross-ratio,

$$(x_1x_2,ab) = (x_2x_1,ba).$$

Hence

$$(x_2x_1,ba) = (x_2x_1,ba'),$$

and therefore

$$a' = a.$$

11. The Poincaré model

We can now discuss Poincaré's model of a hyperbolic non-Euclidean geometry. The term "non-Euclidean" refers to the fact that in this geometry Euclid's parallel axiom does not obtain.

Our space is *the interior Ω of the circle ω*. We shall call this the h-plane. A point of Ω is a point, or h-point, in our new geometry. But *our h-lines are the arcs of circles (or straight lines) inside Ω, limited by ω and orthogonal to ω*. These are the arcs considered in our previous theorems. If a point in Ω lies on one of these arcs, we shall say that the corresponding h-point lies on the corresponding h-line.

We see that the ordinary incidence theorems of Euclidean geometry hold in our h-geometry. Through two distinct points of Ω there passes a unique circle orthogonal to ω. Hence, *through two distinct h-points there passes a unique h-line.*

Two circles orthogonal to ω intersect in at most one point which is inside Ω. Hence, two h-lines intersect in at most one h-point. We shall consider the pairs of h-lines which do not intersect later.

Since arcs of circles inside Ω orthogonal to ω constantly recur, we call them ω-arcs.

The angle between two ω-arcs will be, by definition, the h-angle between the corresponding h-lines.

The relation of equality between the angles of ω-arcs will imply the relation of equality, or h-equality, between h-angles.

We now come to the more difficult notion of *h-equality* between h-segments. An *h-segment* is naturally the portion of an ω-arc bounded by two points of Ω.

In elementary geometry two segments are equal if one can be applied to the other. This involves a displacement, or motion, of the plane. Such a motion transforms lines into lines, and conserves angles. In our h-geometry the equivalent to this group of motions is the group of M-transformations, which transform h-lines into h-lines, and conserve angles. We therefore call an M-transformation of Ω *an h-displacement of the h-plane*, and say that two h-segments AB and CD are h-equal, written

$$AB \overset{h}{=} CD,$$

if there is an h-motion which transforms AB into CD. Since M-transformations form a group, the relation of h-equality satisfies the usual axioms of equivalence:

(i) $AB \overset{h}{=} AB$:

(ii) if $AB \overset{h}{=} CD$, then $CD \overset{h}{=} AB$:

(iii) if $AB \overset{h}{=} CD$, and $CD \overset{h}{=} EF$, then

$AB \overset{h}{=} EF$.

Again, since there is an M-transformation mapping the segment ab of an ω-arc on the segment ba, we also have

$$AB \overset{h}{=} BA.$$

The ordinary theorems on congruence in Euclidean geometry have an equivalent in our h-geometry. For example, let us prove the theorem:

In two h-triangles ABC and $A'B'C'$, the equalities $AB \overset{h}{=} A'B'$, $AC \overset{h}{=} A'C'$ and $\angle CAB \overset{h}{=} \angle C'A'B'$ imply the equality $\angle ABC \overset{h}{=} \angle A'B'C'$.

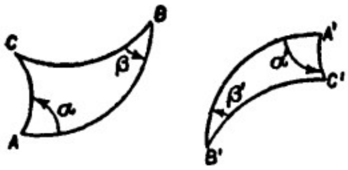

FIG. 59

To prove this we must show that if two triangles ABC, $A'B'C'$ (made up of ω-arcs) are such that $\alpha = \alpha'$ (Fig. 59), and there exists an M-transformation which maps AB on $A'B'$, and an M-transformation which maps AC on $A'C'$, then $\beta = \beta'$.

Suppose at first that the triangles ABC, $A'B'C'$ have the same orientation. The proof will consist in showing that there is an M-transformation which maps the ω-triangle ABC on the ω-triangle $A'B'C'$.

Since $AB \overset{h}{=} A'B'$, there exists a (unique) M-transformation which maps A on A' and B on B'. Since this transformation preserves angles in magnitude and orientation, it maps the ω-arc AC on the ω-arc $A'C'$ in the direction $A'C'$. This M-transformation is the only one which maps A on A' and the ω-arc AC in the direction AC on the ω-arc $A'C'$ in the direction $A'C'$. Hence, it is the one which maps A on A' and C on C', since we know such a transformation exists.

We have therefore found an M-transformation which maps the ω-triangle ABC on the ω-triangle $A'B'C'$. This transformation conserves angles, and therefore $\beta = \beta'$.

If the orientations of the ω-triangles ABC, $A'B'C'$ are different, we take *the geometrical image* of $A'B'C'$ in any diameter of Ω. Call the transformed triangle $A''B''C''$. Then the ω-triangles ABC, $A''B''C''$ have the same orientation, and the proof applies. Since the ω-triangle $A''B''C''$ arises from $A'B'C'$ by a reflection, the theorem follows.

12. The parallel axiom

So far we have shown that there is a complete equivalence between our h-geometry and that part of Euclidean geometry which is developed before the introduction of the parallel axiom. We now come to the parting of ways!

We could say that two ω-arcs which do not intersect shall be called *h-parallel h-lines*. Through a point A outside any ω-arc we can draw an infinity of ω-arcs which do not cut the given ω-arc. This would lead to there being an infinity of h-lines h-parallel to a given h-line through a given h-point outside the h-line.

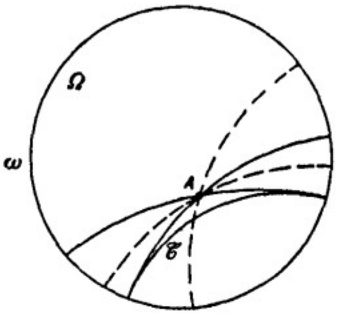

FIG. 60

We notice that there are two ω-arcs through A which *touch* the given ω-arc, the respective points of contact being on ω. These two ω-arcs separate the ω-arcs through A into two classes: those which cut the given ω-arc, and those which do not.

We prefer to reserve the term *h-parallel* for the two ω-arcs which touch the given ω-arc at a point of ω, and therefore now have, instead of Euclid's parallel axiom, the theorem:

Through any h-point A taken outside an h-line there pass precisely two h-parallels to the given h-line.

Hence, the term *non-Euclidean* geometry. The existence of a model for non-Euclidean geometry shows that the parallel axiom is *independent* of the other axioms, and *cannot be deduced from them.* Much fruitless effort has gone into attempts to *prove* the parallel axiom, and even today there exist organizations of people devoted to proving the axiom. Needless to say, these devotees are not shaken in their faith by mathematical arguments.

13. Non-Euclidean distance

We know how to compare two h-segments for equality. We now want to obtain a *distance-function* $D[ab]$ for any two h-points a, b which satisfies the conditions:

$$D[ba] = D[ab] \geqslant 0;$$
$$D[ab] = 0 \qquad \text{if and only if } b = a;$$
$$D[ab] + D[bc] \geqslant D[ac],$$

with equality if and only if b lies on the h-line ac between a and c.

We also want this distance function to be invariant under h-motions of the h-plane.

Let the h-line ab cut ω in the points α, β. Then

$$D[ab] = |\log(\alpha\beta, ab)|$$

is a suitable distance-function. It evidently satisfies the first two conditions, and also the condition that it be invariant under h-motions of the h-plane.

For $\quad D[ba] = |\log(\alpha\beta, ba)|$

$$= \left|\log\frac{1}{(\alpha\beta, ab)}\right| = |-\log(\alpha\beta, ab)|$$

$$= D[ab],$$

and, by definition, $D[ab] \geqslant 0$. It can only $= 0$ if $(\alpha\beta, ab) = 1$, when we must have $b = a$. Again, if any M-transformation of Ω maps the points α, β, a, b on α', β', a', b' respectively, we know that α', β' will be on ω, and that α', β', a', b' all lie on an h-line. Since also $(\alpha\beta, ab) = (\alpha'\beta', a'b')$,

$$D[ab] = D[a'b'].$$

If a, b, c are three points, in this order, on an h-line which cuts ω in α, β, we have, identically,

$$(\alpha\beta, ac) = (\alpha\beta, ab)\,.\,(\alpha\beta, bc).$$

As the points a, b, c are ordered as shown, the three logarithms $\log(\alpha\beta, ac)$, $\log(\alpha\beta, ab)$ and $\log(\alpha\beta, bc)$ all have the same sign. Therefore

$$D[ac] = D[ab] + D[bc].$$

To prove the full triangle inequality is more difficult, but we remark that since the parallel axiom is not used in Euclid's earlier theorems, we can apply the first twenty-eight propositions of Euclid's first book, without change, to our h-plane. We then obtain theorems concerning the congruence of triangles, as we have seen, the theorem that the greatest side of a triangle is opposite the greatest angle, and lastly, the triangle inequality:

$$D[ab] + D[bc] > D[ac],$$

if a, b, c are not, in this order, on an h-line.

It is possible to go much further into the study of non-Euclidean geometry, but we do not propose to do so in this book. The reader may find that his interest in the axioms of Euclidean geometry has been aroused, and he cannot do better than to study these afresh, and then see what are the essential differences between the non-Euclidean geometry we have been discussing and Euclidean geometry.

! "#$%$ & ' ' "( ) " "! + + * ! $
7 * S ' 8 = 9 =

1 ! "$/2" " "3 4 + " $ + = % = ! *
@ + 77 * 4 + + .*

?1 ! "$/2" " "3 4 + " ! . (+ ! +
+ * ! $ 7 * S ' 8 = 9 = . 4 7 + +
: 4 . 4 ; 1 .*
3: , ' * <<<
>4 B; 1

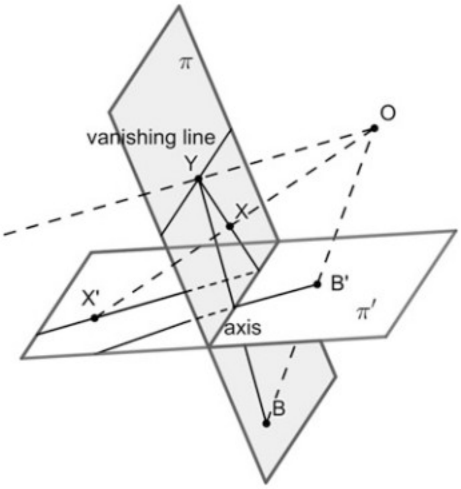

π
O
vanishing line
Y
X
X'
B'
π'
axis
B

@ 4 S .+
. + Y . + YP

! "#$%$ & ' ' "( ) " "8 = % , + * !

1 ! "$/2" " "3 4 + " ? $ # 9 : % +
! +) 4 ;(+ 1 . 4 .) ; 1 .*

1 ! "$/2" " "3 4 + " ? ! .*

1 ! "$/2" " "3 4 + " ? 8 = % , + *
! $$; 1 .*
3: , ' * <<<

1 ! "$/2" " "3 4 + " ? 8 = & : 0 ' .*
A 8 = & :
A B .+4

1 ! "$/2" " "3 4 + " ? 8 = : &
9 + * ; ? 1 *S'
A 8 = & :
A .+4

B1 ! "$/2" " "3 4 + " ? 8 = : &
9 + * ; ? 1 .*

1 ! "$/2" " "3 4 + " ? 9+ 9+) # ' .

)) 8 = & : # ' : + ; ? 1 *S'

1 ! "$/2" " "3 4 + " ? 9+ 9+) # ' .
)) 8 = & : # ' : + ; ? 1 .*

1 ! "$/2" " "3 4 + " ? 9+ 9+) # ' .
)) 8 = & : # ' : + ; ? 1 *S'

?1 ! "$/2" " "3 4 + " ? 9+ 9+) # ' .
)) 8 = & : # ' : + ; ? 1 .*

? 1 ! "$/2" " "3 4 + " ? Q & : + *
8 = . 4 ; ?? 1 .*

? 1 ! "$/2" " "3 4 + " ? O+4 ! . C & :
8 = $ +* ; ? 1 *S'

? 1 ! "$/2" " "3 4 + " ? O+4 ! . C & :
8 = $ +* ; ? 1 .*

! "#$%$ & ' ' "( ) " "9 / * 7 O ! H
0 4 4 + * + +

? 1 ! "$/2" " "3 4 + " 9 / * 7 O ! H
0 4 4 + * + + J ' = ; 1 .*
>4 ; 1 ! +. 9 * + + + 4

! "#$%$ & ' ' "( ) " "7 . + T E + 4 !

? 1 ! "$/2" " "3 4 + " ?B .*
A + +: E * . 4 . + + C.+ . . =

?B1 ! "$/2" " "3 4 + " .*
A!7/!,0 2J$/7&8 H7%9 70!0 3@ $,, (3J&#$/O $&# ,3H (3/0,
!3 ,0-7%O

? 1 ! "$/2" " "3 4 + " + + 4 = . : +
+ + + E .*
A $ $ 8 = : @+ & + + + % E

? 1 ! "$/2" " "3 4 + " :+ 6.+ +* C .*
A($7/0 0$ J/$(,0 $/$#3-7!$, #0!3 3 7%73& 7$ $%!97&8

? 1 ! "$/2" " "3 4 + " : +) . ' + + *
.+ +: B .*
A& + ' % + + . + K * .+ % 7 J +:

??1 ! "$/2" " "3 4 + " 6 E + 46 +) .*
A $ ' %+) D E + 4 . :

B 1 ! "$/2" " "3 4 + " 6' .*
A(3/0, !7/!,0 2J$/7&8

B 1 ! "$/2" " "3 4 + " 6' .*
A(3/0, !7/!,0 2J$/7&8

B 1 ! "$/2" " "3 4 + " 8 4 +8+ + 3: * .*
A%+) D ! E + 4 : ,+ U) ' D

B 1 ! "$/2" " "3 4 + " .*
A% ! E + 4 : # . *

B 1 ! "$/2" " "3 4 + " & ? ? ?6 .*

! "#$%$ & ' ' "( ) " "7 ' % 4 @ 7%@
Q00 + * !:

B B1 ! "$/2" " "3 4 + " 7 ' % 4 @
7%@ Q00 + * !: .*

! "#$%$ & ' ' "( ) " " + + +

B 1 ! "$/2" " "3 4 + " $** + ! +. 9 4
+ + + ! ' + * 8 + 0 4 4 .*

B 1 ! "$/2" " "3 4 + " + + + %
: % = ' *F H = % $ 7 . + F + * H + O * K +: %
: ; 1 .*
3: , ' * <<<
A .+4

>4 B;B?1 !9$ %0/
H9$% 0, 0 #7# 0J,0/ #3^
(*4

>4 ? ; 1 !9$ %0/
$ 93/% 97 %3/O 3@

B 1 ! "$/2" " "3 4 + " + + + * 4 . +
= 7 * + + * $.. * + + ; 1 *S'

B ?1 ! "$/2" " "3 4 + " + + + * 4 . +
= 7 * + + * $.. * + + ; 1 .*

B 1 ! "$/2" " "3 4 + " &+ P & : % J +
/ + = 3 + + ; + 1 (+ ( ) ; ??B1 *S'

B 1 ! "$/2" " "3 4 + " &+ P & : % J +
/ + = 3 + + ; + 1 (+ ( ) ; ??B1 .*

B 1 ! "$/2" " "3 4 + " + F 7+ &+ P :
+ + = + + (+ ( ) ; 1 .*

! "#$%$ & ' ' "( ) " "2 + 3. + * 2 +
! . + $ *

B 1 ! "$/2" " "3 4 + " ? B B .*
A 2 + 3. + * 2 + ! . +
A B .+4

! "#$%$ & ' ' "( ) " "/ + ' & 4 K U
+ =

B 1 ! "$/2" " "3 4 + " B 6? ?B .*
A/ + ' & 4 K U # + =
A + [

! "#$%$ & ' ' "( ) " " + + H+ 8 +' =
, +* % = + * # 4

B B1 ! "$/2" " "3 4 + " 6? ? ? .*
A + + H+ 8 +' = , +* % = + * # 4
3: 0 *+ <<<

! "#$%$ & ' ' "( ) " " = = 2 + 3. *

B 1 ! "$/2" " "3 4 + " = = .*
A = = 2 + 3. *

! "#$%$ & ' ' "( ) " "% + +)) .+
%

B 1 ! "$/2" " "3 4 + " @ *+ . + + ' =
) + +C)) .+ . 4 +4 ; ? 1 .*

B 1 ! "$/2" " "3 4 + " 7 * . * $C
)) .+ % ! +. + + * 9+ 6!/! ; ?? 1 .*

B ?1 ! "$/2" " "3 4 + " 7 * . * $C
)) .+ % ! +. + + * 9+ 6!/!6, 4 + ; ?? 1 *S'

B 1 ! "$/2" " "3 4 + " % + + 3))
.+ % H $ 7 * % ! + ' 9=. . C & : ( ) +
; 1 .*
>4 ; 1 % 4 =)) +

>4 ?; 1 9=. : + +: , U +

>4 ; 1 ? ! + ! '+ , U +

>4 ??; ?1 (! + ! '+ 4 D 2 + .+

B 1 ! "$/2" " "3 4 + " % + +))
.+ H * .*

! "#$%$ & ' ' "( ) " "% +. 7 .+ 4
% =

B 1 ! "$/2" " "3 4 + " ? &+* % +. .+
4 = + * 4 = (+ ( ) ; 1 .*
>4 ; 1 %
803 0%/O 7& %90 &$%J/$, 3/#0/

>4 ; 1

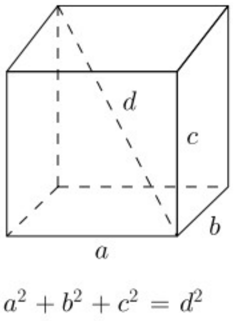
d
c
b
a
$a^2 + b^2 + c^2 = d^2$

2.1—The Pythagorean theorem is often pictured in two dimensions in terms of a right triangle with the sum of the lengths of the sides squared equaling the length of the hypotenuse squared: $a^2 + b^2 = c^2$. But, as shown here, the theorem also works in three dimensions $(a^2 + b^2 + c^2 = d^2)$ and higher.

>4 ; 1

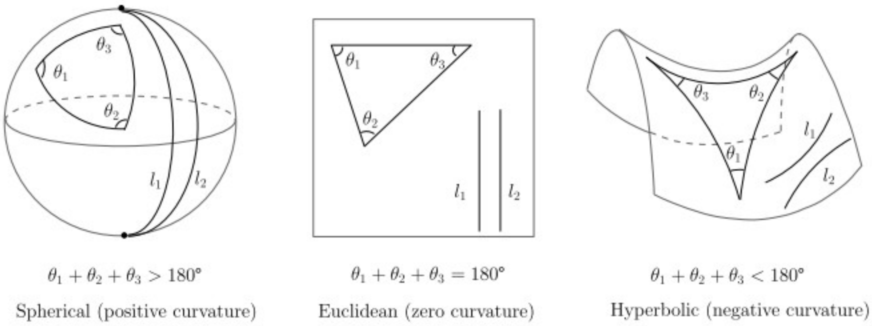
θ_1
θ_2
θ_3
l_1
l_2
$\theta_1 + \theta_2 + \theta_3 > 180°$
$\theta_1 + \theta_2 + \theta_3 = 180°$
$\theta_1 + \theta_2 + \theta_3 < 180°$

Spherical (positive curvature)
Euclidean (zero curvature)
Hyperbolic (negative curvature)

L3 + + . ' '+ + + . F + 4 +
+ 4 4 + + * 4 F + * + +.. + : .+ + ; + 4
* + 1 + ;+ + * . F + 1 3 + Z+ . + +
+ ; U '+ 1F . .+ 4 0 * + 4 =F
+ 4 + + 4 E + * 4 F + * .+ + ' 3
+ + 4+ ' '+ + + +** F + 4 + + 4
+ * 4 F + * 4 = .+ + * ' 4

! "#$%$ & ' ' "( ) " "% ' '+ + + + + @
% +) 0

B 1 ! "$/2" " "3 4 + " $ + + + P ' '+ 4 *
4 +* + + * + = + * ' . $ + + + + = ; 1 *S'

B 1 ! "$/2" " "3 4 + " $ + + + P ' '+ 4 *
4 +* + + * + = + * ' . $ + + + + = ; 1 .*

B B1 ! "$/2" " "3 4 + " K + U !+ 4 + +: F
* 0* H = ; 1 .*

B 1 ! "$/2" " "3 4 + " K + U !+ # = *
8 + 9 ; 1 .*

B 1 ! "$/2" " "3 4 + " K + U @ = ' +
. C '+ +: H+* v ( ) ; 1 *S'

A B ? .+4

B 1 ! "$/2" " "3 4 + " K + U @ = ' +
. C '+ +: H+* v ( ) ; 1 .*

B ?1 ! "$/2" " "3 4 + " K + U 8 4 +
= ( ) x (; 1 .*

B 1 ! "$/2" " "3 4 + " K + U '+ +: !+
H = ; 1 .*

B 1 ! "$/2" " "3 4 + " K + U / + $ + = + *
@ *+ F @ 0* ! +. + + * 9+ !/! ; 1 .*

B 1 ! "$/2" " "3 4 + " K + U % 0
$*'+ * + + F @ 0* ! +. + + * 9+ !/! ; 1 .*

B 1 ! "$/2" " "3 4 + " K + U % ' '+ +
+ + + ; 1 .*

B 1 ! "$/2" " "3 4 + " K + U % ' '+ +
+ + + ; 1 .*

B B1 ! "$/2" " "3 4 + " ' + ! . C + +: + *
! . C 8 = ; * 4 = . + + + F B 1 ; 1 ; ?? 1 .*
A .+4

B 1 ! "$/2" " "3 4 + " ' + . C '+ +: + *
. C 4 =F .+ $ + + + + = ; ?? 1 .*

! "#$%$ & ' ' "( ) " B

! "#$%$ & ' ' "( ) " B"# R 4 * # *= +
8 =

B 1 ! "$/2" " "3 4 + " + + 9+ + ;+ 1 #
R 4 * # *= + 8 = ;#8 1 $ + = ' & U ' + *
%=. : * 4; 1 .*

! "#$%$ & ' ' "( ) " B"@ + + !+ % = + *
$.. +

B 1 ! "$/2" " "3 4 + " @ + + !+ % = + *
$.. + # 7 * . + = # 4 + : 4 7 ; 1 .*

B ?1 ! "$/2" " "3 4 + " 9+ *:) @ + + !+
$.. + (+ % = ;# 8 = / 1 # 8 = ; ?1 .*

B 1 ! "$/2" " "3 4 + " 9+ *:) @ + + !+
$.. + F (+ % = # 8 = ; ?1 .*

B 1 ! "$/2" " "3 4 + " % @ + + !+ % =

+ * $.. + # + + * 7 4 + $: + = 3 * # ' : +
; 1 *S'

B 1 ! "$/2" " "3 4 + " % @ + + !+ % =
+ * $.. + # + + * 7 4 + $: + = 3 * # ' : +
; 1 .*

! "#$%$ & ' ' "( ) " B"@ *+ + % + * %
$.. + !+ + +

B 1 ! "$/2" " "3 4 + " @ *+ + % + * %
$.. + !+ + + . 4 ; 1 .*
A .+4

! "#$%$ & ' ' "( ) " B", + $ 4 : + # / 4

B 1 ! "$/2" " "3 4 + " 9 . .+ !+ : *4
J ' = ; ?? 1 .*

B B1 ! "$/2" " "3 4 + " , + $ 4 : + # / 4
. 4 7 + + : 4 . 4 ; B1 .*
3: , ' * <<<

B 1 ! "$/2" " "3 4 + " , + $ 4 : + # / 4
. 4 7 + + : 4 ; B1 .*

B 1 ! "$/2" " "3 4 + " * $C + F
7 4 + + * / + $ + = . 4 7 + + : 4 ; 1 .*
3: , ' * <<<
7 . + f <<<

B 1 ! "$/2" " "3 4 + " * Q+= $C + +
. * + H = ; ?1 .*

! "#$%$ & ' ' "( ) " B" + + + @ = + + 7 4 +
$ * % $.. +

B ?1 ! "$/2" " "3 4 + " B + + + @ = + +
7 4 + $ * % $.. + .*
>4 ?;? 1 8+ + + 9 : .+ + * +: + H
.+

>4 ; 1 S ' = + + * @
+ +.. +
S ' = +
! C . S ' =
+ ?B
S ' = . C + .+ + *
+ + * + ?
@ 4 + ??
@ = + +
! . C +) ') F . * . + *

4 * + =. E +
% @ + +.. +
B 7 G * + @ 4 +
7 G * + @ 4 +
. = + . + ?

3.7 Gaussian measures on Hilbert spaces and abstract Wiener spaces

Definition 3.11. A probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is said to be a *Gaussian measure* if its Fourier transform $\hat{\mu}$ has the following form

$$\hat{\mu}(x) = e^{imx} e^{-\frac{1}{2}\sigma^2 x^2} \qquad x \in \mathbb{R}\,, \tag{3.53}$$

for some constants $m, \sigma^2 \in \mathbb{R}$, $\sigma^2 \geq 0$. The constants m and σ^2 are called the *mean* and the *variance* of μ respectively.

In σ^2 is a strictly positive constant, then μ is absolutely continuous with respect to the Lebesgue measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with a Radon-Nicodym derivative $h : \mathbb{R} \to \mathbb{R}$ equal to

$$h(x) = \frac{e^{-\frac{1}{2\sigma^2}(x-m)^2}}{\sqrt{2\pi\sigma^2}} \qquad x \in \mathbb{R}\,,$$

while in the case $\sigma^2 = 0$, the measure μ is the δ_m point measure concentrated at $m \in \mathbb{R}$ (see Eq. (2.6)). In particular, the following identities hold

$$m = \int x \, d\mu(x) \qquad \sigma^2 = \int (x-m)^2 \, d\mu(x)\,. \tag{3.54}$$

Definition 3.12. A probability measure μ on the Borel σ-algebra of a real separable Banach space X is called *Gaussian measure* if for any $\theta \in X^*$ the image measure μ_θ of μ under the action of θ is a Gaussian measure on the real line.

By the change of variable formula (2.17), it turns out that the Fourier transform $\hat{\mu}$ of a Gaussian measure on X has the following form

$$\hat{\mu}(\theta) = e^{il(\theta)} e^{-\frac{1}{2}q(\theta)^2} \qquad \theta \in X^*\,, \tag{3.55}$$

where $l : X^* \to \mathbb{R}$ is the linear functional defined by

$$l(\theta) = \int_X \theta(x) \, d\mu(x)\,,$$

while $q : X^* \to \mathbb{R}$ is the quadratic form associated to the positive bilinear map $Q : X^* \times X^* \to \mathbb{R}$ defined as

$$Q(\theta, \eta) := \int_X (x - l(\theta))(x - l(\eta)) d\mu(x) \qquad \theta, \eta \in X^*\,.$$

Analogously to the finite-dimensional case, the bilinear map Q is called the *covariance* of the measure μ.

Conversely, if μ is a probability measure on $\mathcal{B}(X)$ such that its Fourier transform has the form (3.55) then μ is Gaussian since for any $\theta \in X^*$ the Fourier transform of the image measure μ_θ will have the form (3.53). It is important to point out that in an infinite dimensional setting not all positive bilinear map $Q : X^* \times X^* \to \mathbb{R}$ are the covariance of a Gaussian measure. In the particular case where X is a real separable Hilbert space $\mathcal{H}$, Theorem 3.12 allows to get a characterization of the bilinear maps $Q : X^* \times X^* \to \mathbb{R}$ which are the covariance of a Gaussian measure. In this setting it is convenient to exploit the natural isomorphism between $\mathcal{H}$ and $\mathcal{H}^*$ and identify them. In particular, by Riesz theorem it is possible to define a vector $m_\mu \in \mathcal{H}$ such that $l(x) = \langle m_\mu, x\rangle$, which is called the *mean vector* of the measure μ. Similarly we can uniquely associate to the positive bilinear map $Q : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ a bounded symmetric positive operator $S_\mu : \mathcal{H} \to \mathcal{H}$ such that

$$Q(x, y) = \langle x, S_\mu y\rangle \qquad x, y \in \mathcal{H}\,,$$

which is called the *covariance operator* of μ.

Theorem 3.13 (Prokhorov). *Let $\mathcal{H}$ be a real separable Hilbert space. A bounded symmetric positive operator $S : \mathcal{H} \to \mathcal{H}$ is the covariance operator of a Gaussian measure μ on $\mathcal{B}(\mathcal{H})$ if and only if S is nuclear.*

For additional details we refer to [69, 70, 38].

In particular, Prokhorov theorem shows that on an infinite dimensional Hilbert space the standard Gaussian measure (with mean $m = 0$ and covariance operator S equal to the identity $I : \mathcal{H} \to \mathcal{H}$) cannot exists since the identity I is not trace-class. This result can be also obtained as a consequence of Theorem 3.2. Indeed, if a Gaussian measure on $\mathcal{B}(\mathcal{H})$ with $m = 0$ and $S = I$ existed, then it would be invariant under isometries.

It turns out that a useful substitute for the standard Gaussian measure on $\mathcal{H}$ is a cylinder measure. By definition a *cylinder measure* on $\mathcal{H}$ is a positive and finitely-additive set function ν defined on the algebra $\mathcal{A}_\mathcal{H}$ of cylinder sets in $\mathcal{H}$. As remarked above, in a Hilbert space every cylinder set can be represented in the form

$$P^{-1}(C) = \{x \in \mathcal{H} \colon Px \in C\}\,,$$

where $P : \mathcal{H} \to \mathcal{H}$ is an orthogonal projection operator onto a finite dimensional subspace $P\mathcal{H} \subset \mathcal{H}$ and $C \subset P\mathcal{H}$ is a Borel set in $P\mathcal{H}$.

Let us consider the cylinder measure $\nu : \mathcal{A}_\mathcal{H} \to \mathbb{R}$ defined on a general cylinder set $P^{-1}(C)$ as

$$\nu\left(P^{-1}(C)\right) = (2\pi)^{-n/2} \int_C e^{-\frac{1}{2}\|Px\|^2}\, dx,$$

where n is the dimension of $P\mathcal{H}$ and dx denotes the Lebesgue measure on $P\mathcal{H}$. ν is called *standard Gaussian measure associated with* $\mathcal{H}$.

If $\mathcal{H}$ is infinite dimensional, by Prokhorov's theorem the standard Gaussian measure associated with $\mathcal{H}$ cannot be σ-additive on $\mathcal{A}_{\mathcal{H}}$, since in this case it could be extended to a σ-additive probability measure μ on $\mathcal{B}(\mathcal{H})$ invariant under isometries. Indeed, if such a μ existed, its Fourier transform would be given by

$$\hat{\mu}(x) = e^{-\frac{1}{2}\|x\|^2} \qquad x \in \mathcal{H}\,.$$

as it is easy to check since $\hat{\mu}$ is uniquely determined by the value of μ on the cylinder sets.

However, if $\mathcal{H}$ is embedded in a suitable Banach space $(X, |\ |)$, then ν can determine a unique Gaussian measure on $\mathcal{B}(X)$ [237, 169, 168] according to the procedure described below.

Definition 3.13. A norm $|\ |$ (or a semi-norm) on $\mathcal{H}$ is called *measurable* if for every $\epsilon > 0$ there exist a finite-dimensional projection $P_\epsilon : \mathcal{H} \to \mathcal{H}$, such that for all $P \perp P_\epsilon$ one has:

$$\nu(\{x \in \mathcal{H} |\ |P(x)| > \epsilon\}) < \epsilon,$$

where P and P_ϵ are called orthogonal ($P \perp P_\epsilon$) if their ranges are orthogonal in $(\mathcal{H}, \langle\ ,\ \rangle)$.

It turns out that $|\ |$ is weaker than $\|\ \|$. Indeed there exists a $c \in \mathbb{R}^+$ such that [237]

$$|x| \leq c\|x\|, \qquad \forall x \in \mathcal{H}. \tag{3.56}$$

In the following we shall denote X the Banach space constructed as the completion of $\mathcal{H}$ in the $|\ |$-norm and i the inclusion of $\mathcal{H}$ in X, which is continuous by (3.56). Analogously, the dual map $i^* : X^* \to \mathcal{H}^*$, which is given by restriction, i.e. $i^*(x) = x_{|\mathcal{H}}$, is continuous. Identifying $\mathcal{H} \equiv \mathcal{H}^*$ we have the following chain of densely embedded subspaces

$$X^* \subset \mathcal{H} \subset X.$$

The triple $(i, \mathcal{H}, X)$ is called an *abstract Wiener space.*

Given a cylinder set $C_{\theta_1,\ldots,\theta_n;H}$ in X, (with $\theta_1, \ldots, \theta_n \in X^*$ and $H \in \mathcal{B}(\mathbb{R}^n)$ (see Eq. (3.47)), the intersection $C_{\theta_1,\ldots,\theta_n;H} \cap \mathcal{H}$ is a cylinder set in $\mathcal{H}$. We can define the cylinder set measure μ_0 on the algebra $\mathcal{A}_X$ of cylinder sets in X as:

$$\mu_0\,(C_{\theta_1,\ldots,\theta_n;H}) := \nu\,(C_{\theta_1,\ldots,\theta_n;H} \cap \mathcal{H}),$$

A fundamental result of L. Gross is the following statement, which is actually a consequence of the Definition 3.13 of measurable norm [169].

Theorem 3.14. *μ_0 is σ-additive on $\mathcal{A}_X$. Hence it defines a unique probability measure μ on the Borel σ-algebra of X, in such a way that for any $\theta \in X^* \subset \mathcal{H}$:*

$$\int_X e^{i\theta(x)}\, d\mu(x) = e^{-\frac{1}{2}\|\theta\|^2}. \tag{3.57}$$

Example 3.5 (The classical Wiener space). *Let us consider the Sobolev space $\mathcal{H}^{1,2}([0,t],\mathbb{R}^d)$, i.e. the space of absolutely continuous functions $\gamma : [0,t] \to \mathbb{R}^d$ such that $\gamma(0) = 0$ and $\dot\gamma \in L^2([0,t])$, where $\dot\gamma$ denotes the distributional derivative of the function γ, endowed with the inner product*

$$\langle \gamma_1, \gamma_2 \rangle = \int_0^t \dot\gamma_1(s)\dot\gamma_2(s)ds.$$

Let $(C_0([0,t],\mathbb{R}^d), |\cdot|)$ be the Banach space of continuous functions $\omega : [0,t] \to \mathbb{R}^d$ such that $\omega(0) = 0$, endowed with the supremum norm:

$$|\omega| := \sup_{s\in[0,t]} |\omega(s)|.$$

It is possible to prove (see [237] for a additional details) that $\mathcal{H}^{1,2}([0,t],\mathbb{R}^d)$ is dense in $C_0([0,t],\mathbb{R}^d)$ with respect to the supremum norm $|\cdot|$.

Moreover, considered the standard Gaussian measure ν associated with $\mathcal{H}^{1,2}([0,t],\mathbb{R}^d)$, it is possible to prove that the supremum norm $|\cdot|$ is a ν-measurable norm on $\mathcal{H}^{1,2}([0,t],\mathbb{R}^d)$. This makes $(i, \mathcal{H}^{1,2}([0,t],\mathbb{R}^d), C_0([0,t],\mathbb{R}^d))$ an abstract Wiener space, which is called classical Wiener space *while μ is the* Wiener measure

In particular, it is possible to prove that the norm $\|i\|$ of the continuous embedding

$$i : \mathcal{H}^{1,2}([0,t],\mathbb{R}^d) \to C_0([0,t],\mathbb{R}^d)$$

is equal to $\|i\| = \sqrt{t}$.

By Theorem 3.14, each element $\theta \in \mathcal{B}^*$ can be regarded as a measurable function $n(\theta)$ on $(X, \mathcal{B}(X), \mu)$. By adopting the standard terminology of probability theory, $n(\theta)$ is a random variable on the probability space $(X, \mathcal{B}(X), \mu)$. Its distribution, i.e. the probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ obtained as the image measure of μ under the action of $n(\theta)$, is a centered

Gaussian measure with covariance $\|\theta\|^2$. More generally, given $\theta_1, \theta_2 \in X^*$, one has

$$\int_X n(\theta_1)(x) n(\theta_2)(x)\, d\mu(x) = \langle \theta_1, \theta_2 \rangle. \tag{3.58}$$

Equation (3.58) and the density of X^* in $\mathcal{H}$ allow the extension of the map $n : X^* \to L^2(X, \mu)$ to a map $n : \mathcal{H} \to L^2(X, \mu)$ (with an abuse of notation we denote the map n on X^* and its extension to $\mathcal{H}$ with the same symbol)[169]. Moreover, given a complete orthonormal system $\{e_i\}$ in $\mathcal{H}$, for any $h \in \mathcal{H}$ the sequence of random variables

$$\sum_{i=1}^{n} h_i n(e_i), \qquad h_i = \langle e_i, h \rangle,$$

converges in $L^2(X, \mu)$ and by subsequences μ-almost everywhere to the random variable $n(h)$.

The family of measurable mappings $\{n(h)\}_{h \in \mathcal{H}}$ plays an important role in the study of transformation properties of the measure μ under translation. In particular, the following result shows that μ is quasi invariant under translations in the direction of vectors $h \in \mathcal{H}$.

Theorem 3.15. [169, 237] *Let $(i, \mathcal{H}, \mathcal{B})$ be an abstract Wiener space. For every vector $y \in \mathcal{H}$ and every function $f \in L^1(\mathcal{B}, \mu)$, the following holds*

$$\int_{\mathcal{B}} f(x) d\mu(x) = \int_{\mathcal{B}} f(x+y) \rho(y, x) d\mu(x),$$

where ρ is given by

$$\rho(y, x) = e^{-\frac{1}{2}\|y\|^2 - n(y)(x)}.$$

Given an orthogonal projection operator $P : \mathcal{H} \to \mathcal{H}$ with

$$P(h) = \sum_{i=1}^{n} \langle e_i, h \rangle e_i \qquad h \in \mathcal{H},$$

for some orthonormal $e_1, \ldots, e_n \in \mathcal{H}$, let us define the *stochastic extension* $\tilde{P} : X \to \mathcal{H}$ of P as the random variable

$$\tilde{P}(x) = \sum_{i=1}^{n} n(e_i)(x) e_i \qquad x \in X.$$

More generally, this procedure allows the extension of a larger class of mappings f from $\mathcal{H}$ to X. We present here the approach proposed in [278]. In the following $f : \mathcal{H} \to X_1$ will be a function with values in a real separable Banach space $(X_1, \| \, \|_{X_1})$.

Definition 3.14. A function $f : \mathcal{H} \to X_1$ is said to determine a random variable $\tilde{f} : X \to X_1$ if for any sequence $\{P_n\}$ of finite dimensional projection operators converging strongly to the identity on $\mathcal{H}$ the sequence of random variables $\{f \circ \tilde{P}_n\}$ converges in probability[6] on X. In this case the random variable $\tilde{f}$ will be called the *stochastic extension* of f.

The following important result holds [169]:

Theorem 3.16. *If $g : X \to X_1$ is continuous and $f := g|_{\mathcal{H}}$, then the stochastic extension of f is well defined and it is equal to g μ-almost everywhere.*

Other interesting classes of functions on $\mathcal{H}$ which admit a stochastic extension have been studied. For the application to the mathematical theory of Feynman path integrals a relevant role is played by the functions $f \in \mathcal{F}(\mathcal{H})$, i.e. by the mappings $f : \mathcal{H} \to \mathbb{C}$ of the form

$$f(h) = \int_{\mathcal{H}} e^{i\langle h,h'\rangle} d\mu(h') \qquad h \in \mathcal{H}\,, \tag{3.59}$$

for some complex Borel measure μ on $\mathcal{H}$. The following result shows that every function $f \in \mathcal{F}(\mathcal{H})$ admits a stochastic extension. We refer to [215] for the proof.

Theorem 3.17. *Let $f : \mathcal{H} \to \mathbb{C}$ be a function of the form (3.59). Then f admits a stochastic extensions $\tilde{f} : X \to \mathbb{C}$ equal to*

$$\tilde{f}(x) = \int_{\mathcal{H}} e^{in(h')(x)} d\mu(h') \qquad x \in X,.$$

Moreover $\tilde{f} \in L^1(X, \mu)$.

A similar result can be proved if we consider a quadratic form $q : \mathcal{H} \to \mathbb{R}$ on $\mathcal{H}$ associated to an Hilbert-Schmidt operator $K : \mathcal{H} \to \mathcal{H}$

$$q(h) := \langle h, Kh\rangle \qquad h \in \mathcal{H}\,. \tag{3.60}$$

Actually the mapping q can be extended to a random variable $\tilde{q}$ on X provided that suitable normalization counterterms are introduced.

---

[6] We recall that a sequence of random variables $\{g_n\}$ on a probability space $(\Omega, \mathcal{F}, \mu)$ is said to converge in probability to a random variable g on $(\Omega, \mathcal{F}, \mu)$ if the following holds

$$\forall \epsilon > 0 \quad \lim_{n\to\infty} \mu\left(\{\omega \in \Omega \colon |g_n(\omega) - g(\omega)| > \epsilon\}\right) = 0$$

Theorem 3.18. *Let $K : \mathcal{H} \to \mathcal{H}$ be an Hilbert Schmidt operator, $\{e_n\}$ an orthonormal basis in $\mathcal{H}$ and let $\{f_n\}$ be the sequence of random variables on X defined by*

$$f_n(x) := \sum_{j=1}^{n} \left(\widetilde{(e_i K)}(x) n(e_i)(x) - \langle e_j, Ke_j \rangle \right)$$

Then the sequence $\{f_n\}$ converges in $L^2(X, \mu)$ to a random variable $\tilde{f}$ which is independent of the choice of the basis $\{e_n\}_n$.

For the proof we refer to the original paper [278]. The random variable $\tilde{f}$ whose existence is given by Theorem 3.18 is denoted with the symbol $\tilde{f} = \langle x, Kx \rangle - \mathrm{Tr}K$. Clearly, if the operator K is not trace class, the sequence $\mathrm{Tr}P_n K = \sum_{j=1}^{n} \langle e_j, Ke_j \rangle$ doesn't converge.

In the case the operator $K : \mathcal{H} \to \mathcal{H}$ is trace class and self-adjoint, then it is possible to construct a stochastic extension $\tilde{q}$ of the quadratic form (3.60). Indeed, if $\{\lambda_j\}_j$ are the eigenvalues and $\{e_j\}_j$ the corresponding eigenvectors of K, then the mapping $\tilde{q} : X \to \mathbb{R}$ can be obtained as:

$$\tilde{q}(x) := \begin{cases} \lim_{n\to\infty} \sum_{j=1}^{n} \lambda_j \left(n(e_j)(x) \right)^2 & \text{if the limit exists} \\ 0 & \text{otherwise} \end{cases}$$

the mapping $\tilde{q}$ is also denoted with an abuse of notation in the following way

$$\tilde{q}(x) = \langle x, Kx \rangle \qquad x \in X \,.$$

Moreover the following holds:

Theorem 3.19. *Let $K : \mathcal{H} \to \mathcal{H}$ be a self-adjoint trace class operator and $\mu \in \mathcal{M}(\mathcal{H})$ a complex Borel measure on $\mathcal{H}$. The function $f : \mathcal{H} \to \mathbb{C}$*

$$f(h) = e^{\frac{i}{2} \langle h, Kh \rangle} \int_{\mathcal{H}} e^{i \langle h, h' \rangle} \, d\mu(h')$$

admits a stochastic extension $\tilde{f} : X \to \mathbb{C}$ equal to

$$\tilde{f}(x) = e^{\frac{i}{2} \langle x, Kx \rangle} \int_{\mathcal{H}} e^{i n(h')(x)} \, d\mu(h')$$

We end this section with a useful integrability result.

Theorem 3.20. *Let $K : \mathcal{H} \to \mathcal{H}$ be a self-adjoint trace class operator such that $(I - B)$ is strictly positive and let $y \in \mathcal{H}$. Then the random variable on $(\mathcal{B}, \mu)$ defined by*

$$g(x) = e^{n(y)(x)} e^{\frac{1}{2} \langle x, Kx \rangle}$$

is μ-summable and

$$\int_{\mathcal{B}} g \, d\mu = (\det(I-K))^{-1/2} e^{\frac{1}{2}\langle y,(I-B)^{-1}y\rangle},$$

where $\det(I-K)$ *denotes the Fredholm determinant of the operator* $I-K$.

For a proof of these results we refer to [237, 215].

Chapter 4

Projective systems of functionals and the Fourier transform approach

As explained in Section 3.5 it is impossible to construct a reasonable Feynman measure. In particular, the integrals appearing in the piecewise linear approximations of Feynman formula

$$\int_{\mathbb{R}^{dn}} \frac{e^{\frac{i}{\hbar}\sum_{j=1}^{n}\left(\frac{(x_j-x_{j-1})^2}{2(t/n)^2}-V(x_j)\right)\frac{t}{n}}}{(2\pi i\hbar t/n)^{nd/2}}\psi_0(x_0)dx_0\dots dx_{n-1}$$

cannot be interpreted as integrals of a cylinder function with respect to a complex measure μ_F on the space of paths $\Omega : \{\gamma : [0,+\infty) \to \mathbb{R}\}$ whose values on cylinder set is given by

$$\begin{aligned}&\mu_F\left(\{\gamma\in\Omega\colon \gamma(t_1)\in I_1,\dots,\gamma(t_n)\in I_n\}\right)\\&\quad=\int_{I_1\times\cdots\times I_n}\frac{e^{\frac{i}{2\hbar}\sum_{i=1}^{n}\frac{(x_i-x_{i-1})^2}{t_i-t_{i-1}}}}{(2\pi i\hbar)^{n/2}\sqrt{t_1(t_2-t_1)\cdots(t_n-t_{n-1})}}dx_1\dots dx_n \qquad (4.1)\end{aligned}$$

(where $n\in\mathbb{N}$, $0<t_1<\cdots<t_n<+\infty$ and $I_1,\dots,I_n\in\mathcal{B}(\mathbb{R})$).

At this point it is clear that the construction of Feynman integration theory requires to change point of view.

Let us recall that by Riesz-Markov Theorem 2.20, on locally compact Hausdorff topological spaces X there is a one-to-one correspondence between complex regular Borel measures μ and linear continuous functionals $L_\mu : C_0(X)\to\mathbb{C}$, given by the identity

$$L_\mu(f)=\int_X f(x)\,d\mu(x) \qquad f\in C_0(X)\,. \qquad (4.2)$$

On the other hand this result relies on the assumption of local compactness of the space X which is actually a rather strong condition which cannot be easily fulfilled in the case of infinite dimensional topological vector spaces. It cannot hold, for instance, in the case of Hilbert and Banach spaces.

In fact, in general the concept of linear functional on a suitable domain is wider than the concept of Lebesgue integral and can be used to give a mathematical meaning to Feynman formulas, in the spirit of Daniell's approach to integration (see [274]). This chapter presents a systematic implementation of infinite dimensional integration theory based on the concept of *projective system of functionals* instead of projective systems of measures [39]. Similarly to the construction proposed in Kolmogorov Theorem 3.7, the functional will be constructed out of its finite dimensional approximations. This point of view allows to provide a unified view of infinite dimensional integration which covers both the case of probabilistic and oscillatory integration. In particular, according to this approach the Feynman path integral will be constructed as a linear functional $L_F : D(L_F) \to \mathbb{C}$ defined on a suitable domain of "integrable functions" $f : \Gamma \to \mathbb{C}$ defined over a space of paths Γ. The property characterizing the functional as "Feynman integral" will be the following action over the cylinder functions $f \in D(L_F)$

$$L_F(f) = \int_{\mathbb{R}^n} \frac{g(x_1,\dots,x_n)e^{\frac{i}{2\hbar}\sum_{i=1}^{n}\frac{(x_i-x_{i-1})^2}{t_i-t_{i-1}}}}{(2\pi i\hbar)^{n/2}\sqrt{t_1(t_2-t_1)\cdots(t_n-t_{n-1})}}dx_1\dots dx_n$$

with $f(\gamma) = g(\gamma(t_1),\dots,\gamma(t_n))$, for some $n \geq 1$, $0 < t_1 < \cdots < t_n < +\infty$ and $g : \mathbb{R}^n \to \mathbb{C}$.

Sections 4.1 and 4.2 present the general theory of projective systems of functionals, while Section 4.3 focuses on those associated to projective systems of complex measures that do not admit a projective limit because of Theorem 3.8, also proposing some examples. Section 4.4 describes a particular construction of a functional associated to a projective systems of complex measures based on a Fourier transform approach. These results are the basis for the definition of *infinite dimensional Fresnel integrals*, which will be extensively described along with their applications to Feynman path integrals in Section 4.5. Eventually, Section 4.6 presents a generalization of infinite dimensional Fresnel integrals which allows to construct a class of Feynman-Kac type formulae for the representation of the solution of high-order heat-type equations.

4.1 Projective systems of functionals

Let $(\mathcal{I}, \leq)$ be a directed set and $\{S_J, \pi_J^K\}_{J,K\in\mathcal{I}}$ a projective family of sets. In the following we shall assume that $\{S_J, \pi_J^K\}_{J,K\in\mathcal{I}}$ is a perfect inverse system (in the sense of Definition 3.8). Let us denote with the symbol $\hat{S}_J$ the set of complex-valued functions $f_J : S_J \to \mathbb{C}$. For any pair $J, K \in \mathcal{I}$ such that $J \leq K$ let $\mathcal{E}_J^K : \hat{S}_J \to \hat{S}_K$ be the function which maps any element $f_J \in \hat{S}_J$ to its extension $\mathcal{E}_J^K(f_J)$ to $\hat{S}_K$, defined as

$$\mathcal{E}_J^K(f_J)(\omega_K) := f_J\big(\pi_J^K(\omega_K)\big), \qquad \omega_K \in S_K\,. \tag{4.3}$$

In particular, if $\{S_J, \pi_J^K\}_{J,K\in A}$ is a topological projective family and $f_J \in \hat{E}_J$ is a continuous function, then for any $K \in \mathcal{I}$ such that $J \leq K$, the extension $\mathcal{E}_J^K(f_J)$ is a continuous function on S_K.

We shall denote with the symbol $Map(\hat{S}_J)$ the family of linear maps $L_J : \hat{S}_J^0 \to \mathbb{C}$ from subspaces $\hat{S}_J^0 \subseteq \hat{S}_J$ to $\mathbb{C}$, called *functionals*. The set $\hat{S}_J^0$ is called the *domain of* L_J and it is required to be a linear space.

For $J, K \in \mathcal{I}$, such that $J \leq K$, let us define the map $\hat{\pi}_J^K : Map(\hat{S}_K) \to Map(\hat{S}_J)$ as the transport of functionals $L_K \in Map(\hat{S}_K)$ induced by the map $\mathcal{E}_J^K$ from $\hat{S}_J$ to $\hat{S}_K$, which is actually given by:

$$\hat{\pi}_J^K(L_K)(f_J) := L_K(\mathcal{E}_J^K(f_J)), \qquad L_K \in Map(\hat{S}_K), \tag{4.4}$$

where the domain of $\hat{\pi}_J^K(L_K)$ is given by

$$Dom(\hat{\pi}_J^K(L_K)) = \{f_J \in \hat{E}_J,\, |\mathcal{E}_J^K(f_J) \in \hat{S}_K^0\}.$$

Definition 4.1. A family of functionals $\{L_J, \hat{S}_J^0\}_{J\in\mathcal{I}}$ labelled by the elements of the directed set $\mathcal{I}$ is called a *projective system of functionals* if for all $J, K \in \mathcal{I}$ with $J \leq K$ the projective (or coherence or compatibility) conditions hold

$$\mathcal{E}_J^K(f_J) \in \hat{S}_K^0, \qquad \forall f_J \in \hat{S}_J^0,$$
$$\hat{\pi}_J^K(L_K)(f_J) = L_J(f_J), \qquad \forall f_J \in \hat{S}_J^0. \tag{4.5}$$

Let $S_{\mathcal{I}} := \varprojlim S_J$ be the projective limit of the projective family of sets $\{S_J, \pi_J^K\}_{J,K\in\mathcal{I}}$ and let $\hat{S}_{\mathcal{I}}$ denote the set of complex-valued functions on $S_{\mathcal{I}}$. A function $f_J \in \hat{S}_J$, $J \in \mathcal{I}$, can be extended to a function $\mathcal{E}_J^{\mathcal{I}} f_J \in \hat{S}_{\mathcal{I}}$ defined by

$$\mathcal{E}_J^{\mathcal{I}} f_J(\omega) := f_J(\pi_J\omega), \qquad \omega \in S_{\mathcal{I}}\,.$$

By Eq. (3.40), the extension maps $\mathcal{E}_J^{\mathcal{I}} : \hat{S}_J \to \hat{S}_{\mathcal{I}}$ satisfy the following condition for any $J, K \in \mathcal{I}$, with $J \leq K$:

$$\mathcal{E}_J^{\mathcal{I}} = \mathcal{E}_K^{\mathcal{I}} \circ \mathcal{E}_J^K\,. \tag{4.6}$$

In particular, if $(S_J, \pi_J^K)_{J,K\in\mathcal{I}}$ is a topological projective family, then all the extensions $\mathcal{E}_J^K : \hat{S}^J \to \hat{S}_{\mathcal{I}}$ and $\mathcal{E}_J^{\mathcal{I}} : \hat{S}^J \to \hat{S}_{\mathcal{I}}$ map continuous functions into continuous functions.

In this context we shall call *cylindrical (or cylinder) functions* the mappings $f : S_{\mathcal{I}} \to \mathbb{C}$ of the form $\mathcal{E}_J^{\mathcal{I}} f_J$ for some $J \in \mathcal{I}$ and some $f_J \in \hat{S}_J$. The family of cylinder functions will be denoted with the symbol $\mathcal{C} = \cup_{J\in\mathcal{I}} \mathcal{E}_J^{\mathcal{I}}(\hat{S}_J)$.

Remark 4.1. If $\{S_J, \pi_J^K\}_{J,K\in A}$ is a perfect inverse system, then it is simple to prove the injectivity of the extension map $\mathcal{E}_J^{\mathcal{I}} : \hat{S}_J \to \hat{S}_{\mathcal{I}}$. Indeed for any $J \in \mathcal{I}$, $f, g \in \hat{S}_J$:

$$f = g \quad \Leftrightarrow \quad \mathcal{E}_J^{\mathcal{I}} f = \mathcal{E}_J^{\mathcal{I}} g.$$

Given a projective system of functionals $\{L_J, \hat{S}_J^0\}_{J\in\mathcal{I}}$, we shall denote by $\mathcal{C}_0 \subset \mathcal{C}$ the set of those cylindrical functions obtained as extensions to $S_{\mathcal{I}}$ of functions belonging to the domain of L_J for some $J \in \mathcal{I}$, i.e.:

$$\mathcal{C}_0 := \cup_{J\in\mathcal{I}} \mathcal{E}_J^{\mathcal{I}}(\hat{S}_J^0) = \{f \in \mathcal{C} :\ f = \mathcal{E}_J^{\mathcal{I}} f_J, \text{ for some } J \in \mathcal{I}, f_J \in \hat{S}_J^0\}. \quad (4.7)$$

Definition 4.2. A *projective extension* $(L, D(L))$ of a projective system of functionals $\{L_J, \hat{S}_J^0\}_{J\in\mathcal{I}}$ is a functional $L : D(L) \to \mathbb{C}$, with $D(L) \subset \hat{E}_{\mathcal{I}}$ such that

- $\mathcal{C}_0 \subseteq D(L)$,
- $L(\mathcal{E}_J^{\mathcal{I}} f_J) = L_J(f_J)$, for all $f_J \in \hat{S}_J^0$.

Theorem 4.1. *Let $\{S_J, \pi_J^K\}_{J,K\in\mathcal{I}}$ be a perfect inverse system and let $\{L_J, \hat{S}_J^0\}_{J\in\mathcal{I}}$ be a projective system of functionals. Then the functional $(L, D(L))$ defined by:*

$$D(L) := \mathcal{C}_0 \quad (4.8)$$
$$L(f) := L_J(f_J), \qquad f = \mathcal{E}_J^{\mathcal{I}} f_J,\ f_J \in \hat{E}_J^0. \quad (4.9)$$

is a projective extension of $\{L_J, \hat{S}_J^0\}_{J\in\mathcal{I}}$.

Proof. The functional L is unambiguously defined. Indeed let us consider function $f \in \mathcal{C}_0$ that can be obtained both as the extension of $f_J \in \hat{S}_J^0$ and of $f_K \in \hat{S}_K^0$, i.e., $f = \mathcal{E}_J^{\mathcal{I}} f_J = \mathcal{E}_K^{\mathcal{I}} f_K$, $J, K \in \mathcal{I}$. By the very definition of directed set, there exists an $H \in \mathcal{I}$ such that $J \leq H$ and $K \leq H$. If we define the two functions $f_H := \mathcal{E}_J^H f_J$ and $g_H := \mathcal{E}_K^H f_K$, it is easy to see that $\mathcal{E}_H^{\mathcal{I}} f_R = \mathcal{E}_H^{\mathcal{I}} g_R = f$. Thus, by Remark 4.1, $f_R = g_R$ and $L_R(f_R) = L_R(g_R)$. On the other hand, by the projectivity condition (4.5) it follows that:

$$L_J(f_J) = L_H(\mathcal{E}_J^H f_J) = L_H(f_H) = L_H(g_H) = L_H(\mathcal{E}_K^H f_K) = L_K(f_K).$$

Thus L is unambiguously defined by $L(f) = L_J(f_J) = L_k(f_K)$. $\square$

The functional $(L, D(L))$ defined by (4.8) and (4.9) is called the *minimal extension* of the projective system of $\{L_J, \hat{S}_J^0\}_{J\in\mathcal{I}}$ since any other extension $(L', D(L'))$ of $\{L_J, \hat{S}_J^0\}_{J\in\mathcal{I}}$ enjoys the following properties

$$D(L) \subseteq D(L') \tag{4.10}$$

$$L'(f) = L(f) \quad \forall f \in D(L) \,. \tag{4.11}$$

In the following the minimal extension will be denoted with the symbol $(L_{min}, D(L_{min}))$. We shall also write $L_{min} = \varprojlim L_J$.

An interesting characterisation of the domain of the minimal extension of a projective system of functionals $\{L_J, \hat{E}_J^0\}_{J\in A}$ relies on the concept of direct or inductive limit of a direct system. We recall here their definitions (see also [290]).

Definition 4.3. Let $(\mathcal{I}, \leq)$ be a directed set. A *direct system* on $\mathcal{I}$ is a collection $(S_J, F_{JK})_{J,K\in\mathcal{I}}$, where $\{S_J\}_{J\in\mathcal{I}}$ is a family of (non-empty) sets indexed by the elements of $\mathcal{I}$ and $\{F_{JK}\}_{J\leq K}$ a collection of mappings $F_{JK} : E_J \to E_K$ defined for all pairs $J, K \in \mathcal{I}$, $J \leq K$, and enjoying the following properties:

- for every $J \in \mathcal{I}$, F_{JJ} is the identity of E_J,
- $F_{KH} \circ F_{JK} = F_{JH}$ for all $J \leq K \leq H$.

Let us consider on the disjoint union $\cup_J E_J$ the equivalence relation $\sim$ as $\omega_J \sim \omega_K$, with $\omega_J \in E_J$ and $\omega_K \in E_K$, if there is some $H \in \mathcal{I}$ such that $J \leq H$, $K \leq H$, and $F_{JH}(\omega_J) = F_{KH}(\omega_K)$.

Definition 4.4. The *direct (or inductive) limit* $\varinjlim S_J$ of the direct system $(S_J, F_{JK})_{J,K\in\mathcal{I}}$ is the quotient set $\varinjlim E_J = \cup_J E_J / \sim$ of the equivalence relation $\sim$.

If the sets $(S_J)_{J\in\mathcal{I}}$ are topological spaces and the maps $(F_{JK})_{J,K\in\mathcal{I}}$ are continuous, the family $(S_J, F_{JK})_{J,K\in\mathcal{I}}$ is called a *topological direct system*. Its direct limit is the space $\varinjlim E_J$ endowed with the finest topology making all the maps $F_J : E_J \to \varinjlim E_J$ continuous.

Given a direct system $(S_J, F_{JK})_{J,K\in\mathcal{I}}$ and its direct limit $\varinjlim S_J$, let $F_J : S_J \to \varinjlim S_J$ be the function which maps each element of S_J into its equivalence class. It is easy to verify that the maps $\{F_J)_{J\in\mathcal{I}}$ and $\{F_{JK}\}_{J\leq K}$ enjoy the following composition property

$$F_J = F_K \circ F_{JK} \qquad J, K \in \mathcal{I}, J \leq K \,.$$

An application of the concept of Definitions 4.3 and 4.4 is found in considering a projective system of functionals $(L_J, \hat{S}^0_J)_{J\in\mathcal{I}}$. Indeed, the family of sets $\{\hat{S}^0_J\}_{J\in\mathcal{I}}$ and the collection $\{\mathcal{E}^K_J\}_{J\leq K}$ of extension maps

$$\mathcal{E}^K_J : \hat{S}^0_J \to \hat{S}^0_K, \qquad J \leq K, J, K \in A$$

form a direct system of sets. Its direct limit $\varinjlim \hat{S}^0_J$ is isomorphic to the set $\mathcal{C}_0$ of cylindrical functions defined in (4.7), as stated in the following theorem.

Theorem 4.2. *Let $\{S_J, \pi^K_J\}_{J,K\in\mathcal{I}}$ be a perfect inverse system, let $\{L_J, \hat{S}^0_J\}_{J\in\mathcal{I}}$ be a projective system of functionals and $(L_{min}, D(L_{min}))$ its minimal extension. Then $D(L_{min})$ is isomorphic to the direct limit $\varinjlim \hat{S}^J_0$ of the direct system $\{\hat{S}^0_J, \mathcal{E}^K_J\}_{J,K\in\mathcal{I}}$.*

Proof. We have to prove a one-to-one correspondence between $D(L_{min}) = \mathcal{C}_0$ and $\varinjlim \hat{S}^J_0$. Let us consider an arbitrary element $[f] \in \varinjlim \hat{E}^J_0$, i.e. an equivalence class of functions such that for any two elements f_J and f_K belonging to $[f]$ there is an $H \in \mathcal{I}$, with $J \leq H$ and $K \leq H$, such that $\mathcal{E}^H_J(f_J) = \mathcal{E}^H_K(f_K)$. In fact, all the elements belonging to this equivalence class $[f]$ define a unique cylindrical function. Indeed, given two points $f_J, f_K \in [f]$ and considering the associated cylindrical functions $\mathcal{E}^{\mathcal{I}}_J f_J$ and $\mathcal{E}^{\mathcal{I}}_K f_K$, it is simple to check that for any $\omega \in S_{\mathcal{I}}$ we have $\mathcal{E}^A_J f_J(\omega) = \mathcal{E}^A_K f_R(\omega)$. In order to show this equality it is sufficient to take an element $H \in \mathcal{I}$, with $J \leq H$ and $K \leq H$, such that $\mathcal{E}^H_J(f_J) = \mathcal{E}^H_K(f_K)$ (such an H exists as f_J and f_K belong to the same equivalence class g), and apply Eq. (4.6), which gives

$$\mathcal{E}^{\mathcal{I}}_J f_J(\omega) = \mathcal{E}^{\mathcal{I}}_H \circ \mathcal{E}^H_J f_J(\omega) = \mathcal{E}^{\mathcal{I}}_H \circ \mathcal{E}^H_K f_K(\omega) = \mathcal{E}^{\mathcal{I}}_K f_K(\omega)\,.$$

Conversely, under the assumption that $\{S_J, \pi^K_J\}_{J,K\in\mathcal{I}}$ is a perfect inverse system, to any cylindrical function belonging to $\mathcal{C}_0$ it is possible to associate a unique element of $\varinjlim \hat{E}^J_0$.

Indeed, let $f \in \mathcal{C}_0$ of the form $f = \mathcal{E}^{\mathcal{I}}_J f_J$, $J \in A$, $f_J \in \hat{S}^0_J$, and let $[f_J] \in \varinjlim \hat{E}^J_0$ be the equivalence class of f_J. If the cylindrical function f can be also be obtained as $f = \mathcal{E}^{\mathcal{I}}_K f_K$, for some $K \in \mathcal{I}$, $f_K \in \hat{S}^0_K$, then necessarily the following equality holds $[f_K] = [f_J]$, i.e., there exists an $H \in \mathcal{I}$, with $J \leq H$ and $K \leq H$ such that $\mathcal{E}^H_J f_J = \mathcal{E}^H_K f_K$. In order to prove this, it is sufficient to consider a set $H \in \mathcal{I}$ with $J \leq H$ and $K \leq H$ and the functions $f_H, f'_H \in \hat{S}^0_H$ of the form $f_H := \mathcal{E}^H_J f_J$ and $f'_H := \mathcal{E}^H_K f_K$. By Eq. (4.6) it follows that $\mathcal{E}^{\mathcal{I}}_H f_H = \mathcal{E}^A_H f'_H$, indeed $\mathcal{E}^{\mathcal{I}}_H f_H = \mathcal{E}^{\mathcal{I}}_H \mathcal{E}^H_J f_J = \mathcal{E}^{\mathcal{I}}_J f_J = f$ and analogously $\mathcal{E}^{\mathcal{I}}_H f'_R = \mathcal{E}^{\mathcal{I}}_H \mathcal{E}^H_K f_K = \mathcal{E}^{\mathcal{I}}_K f_K = f$, and by Remark 4.1 it follows that $f_R = f'_R$ (since we assumed that (E_J, π^K_J) is a perfect inverse system). □

Let $\mathcal{L} = \{(L, D(L))\}$ be the set of all possible projective extension of a projective system of functionals $\{L_J, \hat{S}^0_J\}_{J\in\mathcal{I}}$. Under the assumption that $(E_J, \pi^K_J)_{J,K\in A}$ is a perfect inverse system, the set $\mathcal{L}$ is not empty since in this case the minimal extension exists. In addition, it is possible to define a partial order relation $\leq$ on $\mathcal{L}$ as

$$L \leq L' \text{ if } D(L) \subset D(L') \text{ and } L(f) = L'(f)\ \forall f \in D(L)\,. \qquad (4.12)$$

Besides the study of the minimal extension, it is reasonable to investigate the existence of a *maximal extension* of a projective system of functionals $\{L_J, \hat{E}^0_J\}_{J\in A}$, i.e., a functional $(L_{max}, D(L_{max}))$ which cannot be further extended. More precisely, an extension $(L_{max}, D(L_{max})) \in \mathcal{L}$ is called maximal if the following holds

$$L_{max} \leq L \quad \Rightarrow L = L_{max}.$$

Similarly, we can define a *maximum extension* $(\bar{L}, D(\bar{L}))$, as the functional such that $L \leq \bar{L}$ for all $L \in \tilde{L}$. The latter problem is closely related to the uniqueness property of the extensions of a projective system. Indeed, if there existed two extensions $(L, D(L))$ and $(L', D(L'))$ such that $L(f) \neq L'(f)$ for some $f \in D(L) \cap D(L')$, then it would be impossible to construct $\bar{L}$, as it would be ambiguously defined on the element f. In fact, this property is also related to the existence of $(L_{max}, D(L_{max}))$, as is stated in the following theorem.

Theorem 4.3. *Let $\{L_J, \hat{E}^0_J\}_{J\in A}$ be a projective system of functionals, $\mathcal{L} = \{(L, D(L)\} \subset \mathcal{L}$ be the family of its projective extensions and let us consider the following statements:*

(1) Existence of a maximum extension: *there exists a functional $(\bar{L}, D(\bar{L})) \in \mathcal{L}$ such that $L \leq \bar{L}$ for all $L \in \mathcal{L}$.*

(2) Uniqueness property: *if $(L, D(L)), (L', D(L')) \in \tilde{\mathcal{L}}$ and $f \in D(L) \cap D(L')$, then $L(f) = L'(f)$.*

(3) Existence of a maximal extension: there exists a functional $(L_{max}, D(L_{max})) \in \mathcal{L}$ such that if $L \in \mathcal{L}$ and $L_{max} \leq L$ then $L = L_{max}$.

Then $1 \Rightarrow 2 \Rightarrow 3$.

Proof. The implication $1 \Rightarrow 2$ is trivial, as remarked above.
Concerning the implication $2 \Rightarrow 1$, it is sufficient to realise that $\mathcal{L}$ is a nonvoid partially ordered set, with partial order $\leq$ given by (4.12). Moreover any chain $\{L_\alpha, D(L_\alpha)\}_\alpha \subset \mathcal{L}$ has an upper bound $(\bar{L}; D(\bar{L}))$ given by

$$D(\bar{L})) = \cup_\alpha D(L_\alpha) \subset \hat{S}_{\mathcal{I}}, \qquad \bar{L}(f) = L_\alpha(f),\ f \in D(L_\alpha),$$

(the set inclusion being clear since all $D(L_\alpha)$ are in $\hat{S}_\mathcal{I}$). By Zorn's lemma we can deduce that $\mathcal{L}$ has a maximal element. □

Given two extensions $(L, D(L))$ and $(L', D(L'))$ of a projective system of functionals $\{L_J, \hat{E}^0_J\}_{J\in A}$, in general it is not always true that they coincide on the intersection of their domains, i.e., that for any $f : E_A \to \mathbb{C}$ such that $f \in D(L) \cap D(L')$, one has that $L(f) = L'(f)$. Indeed let us consider the following counterexample.

Example 4.1. Let us consider Example 3.3 in the particular case where $I = \mathbb{N}$, hence $\mathcal{I} = \mathcal{F}(\mathbb{N})$ is the totality of non-empty finite subsets of $\mathbb{N}$. Given an element $J \in \mathcal{I}$, let S_J be given by $S_J = \mathbb{R}^J$. We also identify $S_\mathcal{I}$ with the set $\mathbb{R}^\mathbb{N}$ of all real-valued sequences.
Let μ denote a probability measure on the Borel σ-algebra $\mathcal{B}(\mathbb{R})$ and let $(L_J, \hat{S}^0_J)$ be the family of functionals defined by

$$\begin{aligned}\hat{S}^0_J &= B_b(S_J)\\ L_J(f) &= \int_{S_J} f\, d\mu_J, \qquad f \in \hat{S}^0_J\end{aligned}$$

where $B_b(S_J)$ is the set of Borel bounded functions on S_J and $\mu_J = \times_{n\in J}\mu$ is the product measure of $|J| < \infty$ copies of μ ($|J|$ denoting the cardinality of $J \in \mathcal{I}$). By Kolmogorov's existence Theorem 3.4 there exists a probability measure $\mu_\mathcal{I}$ on the $\sigma-$algebra generated by the cylindrical sets in $S_\mathcal{I} = \mathbb{R}^\mathbb{N}$ obtained as the product $\times_{n\in\mathbb{N}}\mu$. In particular, if μ is the standard centered Gaussian measure on $\mathcal{B}(\mathbb{R})$, the subset $l_2 = \{\omega \in \mathbb{R}^\mathbb{N}, \mid \sum_n |\omega_n|^2 < \infty\}$ has zero $\mu_\mathcal{I}$-measure (see Example 3.2).

Let us consider the following two different extensions $(L, D(L))$ and $(L', D(L'))$ of the projective system of functionals $\{L_J, \hat{E}^J_0\}_{J\in A}$.
The functional L is defined by:

$$\begin{aligned}D(L) &= B_b(\mathbb{R}^\mathbb{N})\\ L(f) &= \int_{\mathbb{R}^\mathbb{N}} f(\omega)\, d\mu_\mathcal{I}(\omega), \quad f \in D(L)\end{aligned} \tag{4.13}$$

where $B_b(\mathbb{R}^\mathbb{N})$ denotes the Borel bounded functions over $\mathbb{R}^\mathbb{N}$.

Denote by $\pi_n : \mathbb{R}^\mathbb{N} \to \mathbb{R}^\mathbb{N}$ the projection of a sequence $\omega \in \mathbb{R}^\mathbb{N}$ to the sequence $\pi_n\omega$ given by

$$(\pi_n\omega)_m = \begin{cases}\omega_m & \text{if } m \le n\\ 0 & \text{otherwise.}\end{cases}$$

Let us consider the functional L' given by

$$D(L') = \{f : \mathbb{R}^{\mathbb{N}} \to \mathbb{C} \,|\, \exists \lim_{n\to\infty} \int f(\pi_n\omega)\, d\mu_{\mathcal{I}}(\omega)\},$$

$$L'_J(f) = \lim_{n\to\infty} \int f(\pi_n\omega) d\mu_{\mathcal{I}}(\omega). \tag{4.14}$$

Let us now consider the function $f : \mathbb{R}^{\mathbb{N}} \to \mathbb{C}$ defined as

$$f(\omega) = \begin{cases} 0 & \text{if } \omega \text{ has an infinite number of non-vanishing components} \\ 1 & \text{otherwise.} \end{cases}$$

In fact, f is the characteristic function of the set

$$B = \{\omega \in \mathbb{R}^{\mathbb{N}} \,|\omega \text{ has a finite number of non-vanishing components }\}. \tag{4.15}$$

One can easily verify that B is a measurable subset of $\mathbb{R}^{\mathbb{N}}$, indeed it belongs to the σ-algebra generated by the cylindrical sets, as:

$$B = \cup_{N\in\mathbb{N}} \cap_{n>N} B_n$$

where

$$B_n := \{\omega \in \mathbb{R}^{\mathbb{N}} \,|\, \omega_n = 0\}.$$

Then one has, for $\mu_{\mathcal{I}}$ the above Gaussian measure

$$\begin{aligned} L(f) &= \mu_{\mathcal{I}}(B) \\ &\leq \mu_{\mathcal{I}}(l_2) = 0 \end{aligned}$$

while

$$L'(f) = \lim_{n\to\infty} \int_{\mathbb{R}^n} f(\pi_n\omega)\, d\mu_{\mathcal{I}}(\omega) = \lim_{n\to\infty} 1 = 1.$$

4.2 Continuous extensions of projective systems of functionals

In the following we shall always assume, unless otherwise stated, that the projective family $(S_J, \pi_J^K)_{J,K\in\mathcal{I}}$ is a perfect inverse system, in such a way that, by Theorem 4.1, the minimal extension $(L_{min}, D(L_{min}))$ of a projective system of functionals $\{L_J, \hat{S}_J^0\}_{J\in\mathcal{I}}$ is well-defined. We shall study its possible continuous extensions.

Let us consider the set of cylindrical functions $\mathcal{C}_0$ associated to a projective system of functionals $\{L_J, \hat{S}_J^0\}_{J\in\mathcal{I}}$ and let us endow the space $\hat{S}_{\mathcal{I}}$ of complex-valued functions on the projective limit $S_{\mathcal{I}}$ with a topology τ

! "#$%$ & ' ' "( ) " B"% (+ + %+) + +* C

BB 1! "$/2" " "3 4 + " $ * %+) ! *
+. ? ? ( ) x ; ? 1 .*

BB 1! "$/2" " "3 4 + " $ * %+) ! *
+. ? B ? ( ) x ; ? 1 .*

BB 1! "$/2" " "3 4 + " $ * %+) ! *
+. ? B ?B ( ) x ; ? 1 .*

BB 1! "$/2" " "3 4 + " $ * %+) ! *
+. ?B ? ? ( ) x ; ? 1 .*

BB 1 ! "$/2" " "3 4 + " $ * %+) != * $ 4 : + F
+ 77 0 ' F $ +* ; ? B1 *S'

BBB1 ! "$/2" " "3 4 + " $ * %+) != * $ 4 : + F
+ 77 0 ' F $ +* ; ? B1 .*

BB 1 ! "$/2" " "3 4 + " $ * %+) != * $ 4 : + F
+ 7 & 9 + * ; ? 1 *S'

BB 1 ! "$/2" " "3 4 + " $ * %+) != * $ 4 : + F
+ 7 & 9 + * ; ? 1 .*

BB 1! "$/2" " "3 4 + " $ * %+) 0+ = H)
+ *L8 = + * % + 4 ( ) x (+ ; 1 .*
>4 ; B1 8 =

BB?1 ! "$/2" " "3 4 + " $ * %+) , 4 F + F
+ + + +. ? ? 3C * + ! + * ; ?B 1 *S'

B 1 ! "$/2" " "3 4 + " $ * %+) , 4 F + F
+ + + +. ? ? 3C * + ! + * ; ?B 1 .*

B 1 ! "$/2" " "3 4 + " $ * %+) ;+ 1
+ + + * * 8 . 4 +4 (9 * : 4 ; ? 1 .*

B 1 ! "$/2" " "3 4 + " $ * %+) ; * 1 %
$C + * H . + / 8 = + * = & 9 + * : .*

B 1 ! "$/2" " "3 4 + " $ * %+) *
' *+* 4 +4 + U+*+ .*
3: 0 *+ <<< 9 ~ +

B 1 ! "$/2" " "3 4 + " $ E ' + .*
A , 4 +4 + U+*+

B B1 ! "$/2" " "3 4 + " + H+4 % (+ + %+)
+ +* C !+ : *4 J ' = ; ?? 1 *S'

B 1 ! "$/2" " "3 4 + " + H+4 % (+ + %+)
+ +* C !+ : *4 J ' = ; ?? 1 .*

B 1 ! "$/2" " "3 4 + " %+) = . * 4
+ 7 + + = . 9 $ * %+) $; ? 1 *S'
>4 ? ; 1 # +4 + / +

B 1 ! "$/2" " "3 4 + " %+) = . * 4
+ 7 + + = . 9 $ * %+) $; ? 1 .*

B ?1 ! "$/2" " "3 4 + " %+) P H * / ' * + *
0C.+ * * ! ,7 : + ; 1 .*

B 1 ! "$/2" " "3 4 + " % 9 = . + + *
@ + , 4 @ $ %+) (: = $ +* ; 1 .*

! "#$%$ & ' ' "( ) " B"%) 4 + + + =F 0*

B 1 ! "$/2" " "3 4 + " / : (U %) 4
+ + + = ; ?1 .*
A BB .+4

B 1 ! "$/2" " "3 4 + " / : @ (U %) 4
+ + + = + ; 1 .*

B 1 ! "$/2" " "3 4 + " / 4 $, 4 + *
% $ %) 4 $: + = + * + + + = . 4 ; 1 .*

! "#$%$ & ' '

! "#$%$ & ' ' "( )

! "#$%$ & ' ' "( ) "

! "#$%$ & ' ' "( ) " "# ' 4 J ' F 0*

B 1 ! "$/2" " "3 4 + " # ' 4 J ' .*

! "#$%$ & ' ' "( ) " "@ + +

B B1 ! "$/2" " "3 4 + " ? ? .*
A$ & @ + + (+ * $.. + # + U+ + / , + *+ *

B 1 ! "$/2" " "3 4 + " ? B ? B .*
A% # + # += @ 8 4 +. + 8 +' = *
,+ 0C. + ,+ ^

B 1 ! "$/2" " "3 4 + " ? + .*
AJ :+ 4 + * + 4F + + 4 =F + * * * +44 4+

B 1 ! "$/2" " "3 4 + " ? (+ .*
A % @ + + 8 = + * :

B ?1 ! "$/2" " "3 4 + " ? + + 8 =H + 7 7 .*
A @ + + 8 = + F + * + * * ^

B 1 ! "$/2" " "3 4 + " ? @J,,%0-% .*
A @ + + 8 =F 8 +. + * % !

B 1 ! "$/2" " "3 4 + " ? ,
#$ 63 + ; *? 4 = 1 .*
A 0' = 4 + +

B 1 ! "$/2" " "3 4 + " ? 3@ ! .*
A 3&%3,387!$, @3J&#$%73& @3/ %/J!%J/$, !3&!0 %J$, 3#0,
A .+4
3: S+ 9 ~ +

! "#$%$ & ' ' "( ) " "8+

B 1 ! "$/2" " "3 4 + " & B - .*
A 8+ F !+ @ * ` + F 9 ! +

B 1 ! "$/2" " "3 4 + " & ? ? .*

B B1 ! "$/2" " "3 4 + " & B ? .*

B 1 ! "$/2" " "3 4 + " & B .*

B 1 ! "$/2" " "3 4 + " B & B .*

B 1 ! "$/2" " "3 4 + " & B ? - .*

B ?1 ! "$/2" " "3 4 + "
? B 6$*H 6 B6K$/7& /07!96 #+) +4 4 .*

B? 1 ! "$/2" " "3 4 + " : U 4 4 .*

B? 1 ! "$/2" " "3 4 + " : U 4 4 .*

B? 1 ! "$/2" " "3 4 + " : U 4 4 .*

B? 1 ! "$/2" " "3 4 + " : U 4 4 .*

B? 1 ! "$/2" " "3 4 + " : :64:6 S0#$$$$2$$Q .*
A B ;B .+4 1

B?B1 ! "$/2" " "3 4 + " : :64:6 -($$$$ $$Q .*
A ;? .+4 1

B? 1 ! "$/2" " "3 4 + " : :64:6+7/ $$$$ $$Q .*
A ;? ? .+4 1

B? 1 ! "$/2" " "3 4 + " : :64:6:9 6$2$$7$$Q .*

A B ;B .+4 1

B? 1 ! "$/2" " "3 4 + " : :64:6&# #$$$$2$$Q .*
A ;? B .+4 1

B??1 ! "$/2" " "3 4 + "
8$!0%$/ 06 6 6 6 .*
A 0 F + + ' *+* = + + > +

1 ! "$/2" " "3 4 + " & B ? - .*
A :+ *

1 ! "$/2" " "3 4 + " & .*
A % &+ + 4 U + +)
A .+4

! "#$%$ & ' ' "( ) " "8 =

1 ! "$/2" " "3 4 + "
V$ 676 +) ' W6$ + 6+ *6 4+ 6;% . 6 ;:) 41 *S'

1 ! "$/2" " "3 4 + "
V$ 676 +) ' W6$ + 6+ *6 4+ 6;% . 6 ;:) 41 .*

1 ! "$/2" " "3 4 + "
V#+ 6R 4 W6!/!6 + *+ *6 + + + 6%+: 6 6 * .*

B1 ! "$/2" " "3 4 + "
V#+ 6R 4 W6!/!6 + *+ *6 + + + 6%+: 6 6 * .*

1 ! "$/2" " "3 4 + "
V@ + 6!+S W6$69 =6 6 6, 4+ 6 *;:) 41 .*

1 ! "$/2" " "3 4 + " ? ?B 6? ? .*
A %) 4 +: 8 * + * % 4
A 0 += ! . C =F ?
A .+4

1 ! "$/2" " "3 4 + " !+ B ?? .*

?1 ! "$/2" " "3 4 + " !+ + + .*

1 ! "$/2" " "3 4 + " + + + + .*

1 ! "$/2" " "3 4 + " !+ 7 ' + .*

1 ! "$/2" " "3 4 + " #$ 7 F + 7 + +
' + F #+ , : U + % 4` + F $* . 0* U F .*

1 ! "$/2" " "3 4 + "
* ' 6 : + 6* 6 + 6 + + + *S'

1 ! "$/2" " "3 4 + "
* ' 6 : + 6* 6 + 6 + + + .*

B1 ! "$/2" " "3 4 + " # *j' (' ! *
+ > + . .*
A .+4

1 ! "$/2" " "3 4 + " # ' * . B ? .*
3: , ' * <<<
A B B .+4

1 ! "$/2" " "3 4 + "
@[C6$ + + 6 >UE 66 : + 6* 6 4 f+6=6 j 6 ' .*
3: , ' * <<<
A .+4

1 ! "$/2" " "3 4 + "
@[C6$ + + 6 >UE 66 : + 6* 6 4 f+6=6 j 6 ' C.
3: , ' * <<<

?1 ! "$/2" " "3 4 + " @ *+ * + > + .*

1 ! "$/2" " "3 4 + " 8 = 0 + = .*

1 ! "$/2" " "3 4 + " 8 = 0 + .*

1 ! "$/2" " "3 4 + " 8 +. 6+ *6! : .*

1 ! "$/2" " "3 4 + " 9 + = + * +
+ + + .*
A B? .+4

1 ! "$/2" " "3 4 + " 7 ' + 6 6 + + *S'

B1 ! "$/2" " "3 4 + " 7 ' + 6 6 + + .*

1 ! "$/2" " "3 4 + "
,+ 6$ 6/ 4 : 468 = .*

1 ! "$/2" " "3 4 + " ' $,8,7& .*
A º 4 : + , +

1 ! "$/2" " "3 4 + "
.+ '6@ *+ 6* 6 + 6 + > + 6 . .*

?1 ! "$/2" " "3 4 + " C . .*
A $ 7 3/%$&!7$ #3 ,38$/7% 3 3&%0 0 93Q0
&3 #0 0& 3, 7 0&%3 #$ $%0 $%7!$ 0 #$
!70&!7$ J $ $(3/#$80 #7#$%7!$

! "#$%$ & ' ' "( ) " "8 = 3 % # +
H *

1 ! "$/2" " "3 4 + " 8 = 7 3 %
* + H * .*
3: , ' * <<<
A .+4

! "#$%$ & ' ' "( ) " " 0C @ + + * @

1 ! "$/2" " "3 4 + " @ * F + @ + + *
0C + 0* + V* : WF 2 ; 1 .*

1 ! "$/2" " "3 4 + "
60C 6@ + 6+ *6@ 63 6 6+ *6 6 .*

1 ! "$/2" " "3 4 + " %90 !3 ,0%0 #J O %3
0- 0/% /$!%7!$, 8J7#0 H7%9 0-$,0 %9$% %0$!90 0 0/O%97&8 O3J
&00# %3 K&3H $(3J% 7!/3 3@% 0-!0, .*

! "#$%$ & ' ' "( ) " " * + + * $. @ v
(= *

1 ! "$/2" " "3 4 + "
* + 6+ *6$. 6@ 66(= * .*
A & : % =

! "#$%$ & ' ' "( ) " "%+: + % E

B1 ! "$/2" " "3 4 + " B %+: 6 + 6% E .*
A H * % :+ + 4 +<<<

! "#$%$ & ' ' "( ) "

B

! "#$%$ & ' ' "( ) " "$ * P 8 * +C P
0E +

1 ! "$/2" " "3 4 + " $ * 8 * +C
0E + := #+ @ .*

! "#$%$ & ' ' "( ) " "$*'+ * 0 4 4 + + + *
$ + =

1 ! "$/2" " "3 4 + "
$*'+ *60 4 46 + + 6+ *6$ + = 6 + + 6/ + 6# ' . .*

! "#$%$ & ' ' "( ) " "$*'+ * 0 4 4 + + + *
$ + =

1 ! "$/2" " "3 4 + " B .*

?1 ! "$/2" " "3 4 + "

$*'+ *60 4 46 + + 6+ *6$ + = 6 + + 6/ + 6# ' . .*

! "#$%$ & ' ' "( ) " "$ + = G [+ * :
. +

1 ! "$/2" " "3 4 + " ?
$ + = 6 [+ 6* 6 : 6. + 6V W$ 6(+ [[=6:.) .*
A B .+4

1 ! "$/2" " "3 4 + " ? + + = + .*

1 ! "$/2" " "3 4 + " ? +. + + = .*

1 ! "$/2" " "3 4 + " ? +. + + = .*

1 ! "$/2" " "3 4 + " ? , ' 6$ + = 6 + .*
A *[+ + = G [+

B1 ! "$/2" " "3 4 + " ? % 6 + .*
A ,+ [* : * [E + *+ + 8[[+ [

! "#$%$ & ' ' "( ) " "$ + = + 8 = % #

1 ! "$/2" " "3 4 + "
$ + = + 68 =6% 6# .*
3: , ' * <<<

>4 B; 1 + + 4 , B
8 + 0E + B
% @ * + & + = ! * + 8 + 0E +
* # 4 * / . + + + 4 ,
$ 4 (+ + ,
0E + ($ 4 : + .+ + 4
B
B 0E + % , Q 4 3 4
H + , + !
H) * 3 0C+ .
0C ??
$

>4 ; 1 # + + * ! S 4+ # +
+
7 . + %
! S 4+ # +
! * + 4 ' + 4 , += : ! S 4+
+ + 0 .
B . ! S 4+ # + + 0 .
. ! S 4+ # + + 9=. : +
0E S 4+ # + + 0 .
H) * 3 0C+ .
? 0C ?
$

! "#$%$ & ' ' "( ) " "! . .=F *
0* ; #@1

1 ! "$/2" " "3 4 + "
! 6 6 . .=6 .*
A ? .+4

! "#$%$ & ' ' "( ) " "# + $ 4 : + % . 4= @
+ * 0C .

1 ! "$/2" " "3 4 + "
+ 6$ 4 : + 6% . 4= .*

! "#$%$ & ' ' "( ) " "# + 4 = B

?1 ! "$/2" " "3 4 + " + . '+) $
! . ' 7 * # + 8 = B : ; ???1 .*

! "#$%$ & ' ' "( ) " "# * 2 + 0 *= +

B 1 ! "$/2" " "3 4 + "
Q+ U 6( )6# *2 + 7 0 *= .*

! "#$%$ & ' ' "( ) " "0 * + + * 0 * + 4
8 : 4

B 1 ! "$/2" " "3 4 + " B + ' Q 8 : 4 0 * +
+ * 0 * + 4 * ' . + * = H 9 @ + ; 1 *S'

B 1 ! "$/2" " "3 4 + " B + ' Q 8 : 4 0 * +
+ * & 0 * + 8 # ' . + * 9 = H 9 @ + ; 1 .*
3: , ' * <<< 7 . + f <<<

! "#$%$ & ' ' "( ) " "@ 4 + + F % E F
: =

B 1 ! "$/2" " "3 4 + " #+ $ * @ 4 + *
% 9 = + * @ = 4 0 + . / ; ?1 .*

B 1 ! "$/2" " "3 4 + " @ 4 + * % 9 =
+ * @ = 4 B 0 + . / ; ?1 .*

BB1 ! "$/2" " "3 4 + " @ 4 . :

B 1 ! "$/2" " "3 4 + " @ 4 .*

B 1 ! "$/2" " "3 4 + " = .+ @ 4
0 ' ; ? 1 .*
A ? .+4

B 1 ! "$/2" " "3 4 + " + . @ 4 $ +
+ , +C ; 1 .*

B?1 ! "$/2" " "3 4 + " % + $. + * .+
@ 4 #= + .*

1 ! "$/2" " "3 4 + " % @ 4 + *
/ + ;$ + & O) $ +* = 1 ; ?? 1 .*

! "#$%$ & ' ' "( ) " "8 : 4 0 * + + *
* + 4

1 ! "$/2" " "3 4 + " 8 : 4 0 * + + * &
0 * + 8 # ' . + * 9 = ;H9@F ?? 1 .*
3: , ' * <<<
A B .+4

! "#$%$ & ' ' "( ) " "9+ *:) @ + + !+
0 4 4 + *

1 ! "$/2" " "3 4 + "
9+ *:)6 6@ + + 6!+ 6 60 4 46+ *6 .*

! "#$%$ & ' ' "( ) " "7* + 2 + ! =

1 ! "$/2" " "3 4 + " ? 7* + 2 + ! = :=
, S+ + .*
A .+4

! "#$%$ & ' ' "( ) " " * + =

1 ! "$/2" " "3 4 + " B * + = := +)
% .*
A B .+4

! "#$%$ & ' ' "( ) " " /7&!7 ,0 3@ 9O 7!

B1 ! "$/2" " "3 4 + " B /7&!7 ,0 3@ 9O 7!
@/3 2J$&%J @70,# %903/O %3 !,$ 7!$, 0!9$&7! .*
A .+4

! "#$%$ & ' ' "( ) " "2 + ! =

1 ! "$/2" " "3 4 + " B 2 + ! = := Q
, F K) $.*
A .+4

! "#$%$ & ' ' "( ) " "/ @ + * . %

1 ! "$/2" " "3 4 + " B / @ + * .
% .*

! "#$%$ & ' ' "( ) " " $ + Q+ + =

1 ! "$/2" " "3 4 + " B $ + Q+ + =

.*
A H * $ $)= + 4 + + + +

! "#$%$ & ' ' "( ) " " . := . 7 @ +) 4
3 4+ (C

?1 ! "$/2" " "3 4 + " BB . := . 7 @
+) 4 3 4+ (C .*
A B .+4

! "#$%$ & ' ' "( ) "

B

! "#$%$ & ' ' "( ) "

! "#$%$ & ' ' "( ) " "$ = $.

1 ! "$/2" " "3 4 + " B $ = .*
A % J ' P @

! "#$%$ & ' ' "( ) " "0 + + +
J * 4 +* +

1 ! "$/2" " "3 4 + " B
! 6( )60 + + + @ J * 4 .*
A ?B .+4

! "#$%$ & ' ' "( ) " "0C.+ 4 * J '
w J) + U)+ K

1 ! "$/2" " "3 4 + " B 6? BB .*

! "#$%$ & ' ' "( ) " "8 + / + ' = + * 8 +' + +
H+' 0 + % = + * +

1 ! "$/2" " "3 4 + " B? # + * +
+ 6( )68 + / + ' =$ *8 +' + .*

1 ! "$/2" " "3 4 + " B? 80&0/$, /0,$%7 7%O $&#
7% $,7!$%73& : +) F .+ + + * 4 +' + + +' !/! /0 ; 1 .*

B1 ! "$/2" " "3 4 + " B? 8 % = 8 + U *
@ $.. + 8 + / + ' = *S'

1 ! "$/2" " "3 4 + " B? 8 % = 8 + U *
@ $.. + 8 + / + ' = .*

1 ! "$/2" " "3 4 + " B? 7 + 4 8 +' + + .*

1 ! "$/2" " "3 4 + " B? * 8 + / + ' = . :

?1 ! "$/2" " "3 4 + " B? * 8 + / + ' = .*

! "#$%$ & ' ' "( ) " "8 % + + ; :
( ) + + 1

1 ! "$/2" " "3 4 + " 6? ? .*
A 8 % + +
3: , ' * <<< 7 . + f <<<
A B .+4
kkkkkkkkkkkkkkkkkkkkkkkkk

! "#$%$ & ' ' "( ) " ",+ U $.. * $ + =

1 ! "$/2" " "3 4 + " ! ,+ U $.. *
+ + = # ' ; ? 1 *S'

1 ! "$/2" " "3 4 + " ! ,+ U $.. *
+ + = # ' ; ? 1 .*
A B .+4
3: , ' * <<<

>4 ?; 1 ! +. 7 2 +* + *

>4 ? ; 1 % 8+ + E +* + *

1 ! "$/2" " "3 4 + " ,+ U F ! $.. *
$ + = # ' : + ; 1 .*

! "#$%$ & ' ' "( ) " " = H 0

1 ! "$/2" " "3 4 + " = H 0 .*

! "#$%$ & ' ' "( ) " ?

! "#$%$ & ' ' "( ) " ?"! . + + % . 4= #+ +
$ + =

B1 ! "$/2" " "3 4 + " ! . + + % . 4= #+ +
$ + = .*
3: , ' * <<<
A BB .+4

! "#$%$ & ' ' "( ) " ?" + * + * , +

1 ! "$/2" " "3 4 + "
+ * 6+ *6, + 6 6$68 + 6% = .*
A .+4

! "#$%$ & ' ' "( ) " ?")= (+

1 ! "$/2" " "3 4 + " B Q # (+ 7)=
! 4F %) 4F + * (4 (+) (+= ( ) ; ?? 1 .*

1 ! "$/2" " "3 4 + " B (+ Q #)=
! 4F) 4F + * : 4 4 ; ?? 1 *S'

! "#$%$ & ' ' "( ) "

! "#$%$ & ' ' "( ) " "$ 7 * 8 * F 8 +. F + *
&)

?1 ! "$/2" " "3 4 + " $ 7 * 8 * F 8 +. F
+ * &) .*

! "#$%$ & ' ' "( ) " "0 = + * +4 &
@ + := 7 * . * ' =

? 1 ! "$/2" " "3 4 + "
+ + 6( )60 =$ * +4 .*
A ? .+4

! "#$%$ & ' ' "( ) " "9 . .+

? 1 ! "$/2" " "3 4 + " 9 . .+ !+ : *4
J ' = ; ?? 1 .*

! "#$%$ & ' ' "( ) " " + + $ 7 + * 9 =
& :

? 1 ! "$/2" " "3 4 + " ? + + ?B ?B .*
A $ 7 + * 9 = & :

! "#$%$ & ' ' "( ) " "% J ' . +) & :

? 1 ! "$/2" " "3 4 + "
% 6J ' 6 . +) 6 6& : F6J 6 68 + + 6@+ . :

? 1 ! "$/2" " "3 4 + "
% 6J ' 6 . +) 6 6& : F6J 6 68 + + 6@+ .*

? B

!"#$# % & &

!"#$# % & & !' (

!"#$# % & & !' (!

!"#$# % & & !' (! ! &) *)) + $,)
- + . #) /

0 !#12! ! !3 4) ! &) *)) + $,
) - + . #) / #)))) / 5 0 *+

!"#$# % & & !' (! !") . /, .) 4 $,
) + .)& /

0 !#12! ! !3 4) !
") 6. /,6.) 4 6$ 6) +6.)& / *+

!"#$# % & & !' (! !7 +) . /

0 !#12! ! !3 4) ! # +) 7 +)
. / *+

0 !#12! ! !3 4) ! & #8) 9)
) #: , 7 +)) + ;/* : . / *+
< -) ===

>0 !#12! ! !3 4) ! .)) + "
) * 7 +) 7) / . / *+
< - *) ? ===

@0 !#12! ! !3 4) ! . *+
3: A & + ===

0 !#12! ! !3 4) ! . / -) + #
-) + - + 7 +)) + ;/* :) . /))) #)
) 5 >0 *+
3: A & + ===
< >@ *)4

B0 !#12! ! !3 4) ! . / $) + 7 +)
) . / ') + 1 4 + #)))) / 5 0 *+
3: A & + ===
< B *)4
C4 5 >0)* #8)
C4 5 0)* 78) + * 1
C4 B 5 0)*)) A) + $))
C4 5 0)* +) . /

Chapter 2

Axioms for the Plane

It is now time to insert a little geometry into our plane. The plane is intended to be an abstract model of the geometry of a flat surface such as a table or a piece of paper, with the difference that our plane will extend without end. So far we have points and some measures (with no details). We have also defined rigid motions, but except for the identity mapping, we have not yet claimed that any rigid motions actually exist. So there are things to do.

In this chapter, we state the six axioms for plane geometry that will be used in this book. The first four are closely related to the well-known Ruler and Compass axioms of G. D. Birkhoff [2], [3]:

- Incidence Axiom for points and lines
- Ruler Axiom concerning distance measure
- Protractor Axiom concerning angle measure
- Plane Separation Axiom concerning boundaries and orientation

We will explore consequences of these first four axioms in this chapter. From these axioms one can define some elementary but important geometrical concepts and prove a few key theorems.

At the end of the chapter, we complete the set of axioms by stating the remaining two axioms:

- Reflection Axiom
- Dilation Axiom

The Reflection Axiom will add rigid motions to our geometry. Significant geometrical connections between distance and angle measure will appear in Chapter 3 as consequences of the Reflection Axiom. This will lead to triangle congruence theorems in Chapter 4. In later chapters the importance of the Dilation Axiom will become evident.

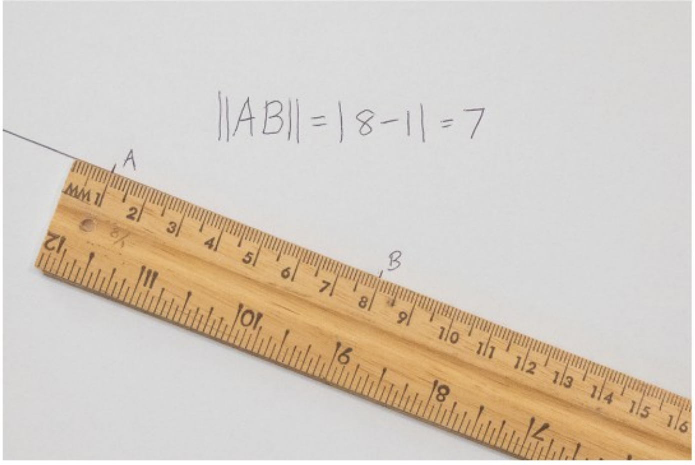

Figure 1. A Ruler Measuring Distance in the Plane

Incidence Axiom

The first axiom is very simple. The informal version is "two points determine a line".

Axiom 1 (Incidence). *The plane consists of a nonempty set of elements called points and a nonempty set of subsets called lines. For any two distinct points, there is exactly one line containing the two points; it will be denoted* $\overleftrightarrow{AB}$.

As a consequence of this, two distinct lines cannot intersect in more than one point. This already rules out our model looking like the sphere. We might think of this as the unmarked straightedge axiom: for any two points on a paper, we can draw a line with a straightedge through the two points.

Distance and the Ruler Axiom

The Ruler Axiom is an existence axiom for a distance measure and for "rulers" that provide distance measurement on a line by means of real numbers. Such a ruler is an abstraction of a marked straightedge.

Axiom 2 (Ruler). *The plane has a distance function.*

- *Distance is a function that maps any two points A and B in the plane to a nonnegative real number $\|AB\|$.*
- *Every line m in the plane has a ruler, a mapping of the real numbers $\mathbb{R}$ onto m so that distances are preserved.*

Another way of describing a ruler is that it is an isometry from $\mathbb{R}$ to m. Explicitly, if ρ is a ruler with $\rho(a) = A$ and $\rho(b) = B$, then the distance $\|AB\|$ equals the standard distance between two real numbers, the absolute value $|a - b|$.

Remark 2.1. This axiom implies a number of important properties of points, lines, and distance.

- For any A, $\|AA\| = 0$. This is true since, for A on a line m with ruler value $\rho(a) = A$, the distance $\|AA\| = |a - a| = 0$.
- A ruler is a one-to-one mapping, for if ρ is a ruler on a line m with $\rho(a) = \rho(b) = A$, then $|a - b| = \|AA\| = 0$, so $a = b$.
- For any two distinct points A and B, the distance $\|AB\| > 0$. For a ruler ρ on $\overline{AB}$, with $\rho(a) = A$ and $\rho(b) = B$ for distinct a and b, the distance $\|AB\| = |a - b| > 0$.
- There are an infinite number of points on each line, and the distances on the line are unbounded.
- If $\rho(x)$ is a ruler, then so is $\rho(x + k)$ or $\rho(-x + k)$ for any real k. Therefore, any point on the line can have coordinate 0.

The last item is true because, if we apply any distance-preserving transformation of the real numbers before mapping them by the ruler, this composition is still a ruler since the distance relations are still preserved. For example, if ρ is a ruler for m, let the function ρ_5 be defined as $\rho_5(x) = \rho(x + 5)$. Then ρ_5 is also a ruler for m, since for any real numbers a and b, $\|\rho_5(a)\rho_5(b)\| = \|\rho(a + 5)\rho(b + 5)\|$, and this equals $|(a + 5) - (b + 5)| = |a - b|$ since ρ is a ruler.

Exercise 1 explores these transformations of the real line.

The real numbers assigned to each point of m by the inverse map are the **coordinates defined by the ruler**. This inverse map will occasionally be designated as a function from points to numbers such as $a = x(A)$ and $b = x(B)$, but usually it will simply be stated that the real numbers a and b correspond to the points A and B in m.

These real numbers can be used to prove things about points on a line m.

For example, given a point A on m and any distance d, there are two points B and C on m at distance d, with $\|AB\| = \|AC\| = d$, since for any real number a, there are two numbers $b = a + d$ and $c = a - d$ with $|b - a| = |c - a| = d$.

Segments. From the Ruler Axiom, for distinct points A and B we can define AB, the **segment with endpoints** A **and** B. For the next few paragraphs, let the numbers a, b, c correspond to the points A, B, C on $\overline{AB}$.

Segment Definition by Ruler: The segment AB is the set of points C of $\overline{AB}$ such that the number c belongs to the interval with endpoints a and b: either $a \le c \le b$ or $b \le c \le a$. Equivalently, $|c - a| + |b - c| = |b - a|$.

Segment Definition by Distance: From the absolute value equation above, one sees that the segment AB is the set of points C of $\overline{AB}$ with $\|AC\| + \|BC\| = \|AB\|$.

Segment Interior Point: A point C is an **interior point** of AB, or is **between** A **and** B, if it is in AB but is not one of the endpoints.

Midpoint: The **midpoint** M of AB is the point on the segment equidistant from A and B. Hence, $2\|AM\| = \|AB\|$.

Convexity: A set S is **convex** if, for any points A and B in S, the segment AB is also contained in S.

Rays. For distinct points A and B, there is a **ray** $\overrightarrow{AB}$ with endpoint A in the direction of B. We continue with the assumption that the numbers a, b, c correspond to the points A, B, C on $\overline{AB}$.

Ray Definition: A point C on $\overline{AB}$ is in $\overrightarrow{AB}$ if A is not between B and C. In other words, $c \geq a$ if $b > a$ and $c \leq a$ if $b < a$.

Ray Interior Point: The set of all points except the endpoint is the set of **interior points** of the ray.

Opposite Ray: The opposite ray of $\overrightarrow{AB}$ is the ray consisting of endpoint A and the points C on $\overline{AB}$ not in $\overrightarrow{AB}$. In other words, $c \geq a$ if $b < a$ and $c \leq a$ if $b > a$.

Directed Line: A choice of ray determines a direction or orientation of the line containing it. Two rays in a line define the same direction if their intersection is a ray; this occurs if one ray is contained in the other, so a line has two possible directions.

Polygons. Now we can define polygons.

Definition 2.2. A **polygon** $P_1P_2\dots P_n$ is the figure consisting of n distinct points, the vertices P_i, and n sides, the segments $P_1P_2, P_2P_3, \dots, P_{n-1}P_n, P_nP_1$, such that (1) no interior point of a side is an intersection point of two sides and (2) any three consecutive vertices are noncollinear.

The sequences of three consecutive vertices include P_{n-1}, P_n, P_1 and P_n, P_1, P_2. This definition excludes self-intersecting polygonal figures and nonclosed polygonal paths, which can also be interesting but are not included as polygons.

The most familiar polygons have three sides (triangles), four sides (quadrilaterals), five sides (pentagons), and six sides (hexagons).

Protractor Axiom and Angles

Next, we will consider how to measure angles. Alas, even in informal geometry this is a bit more complicated than measuring along a line.

Definition 2.3. An **angle with vertex** A consists of two rays $\overrightarrow{AB}$ and $\overrightarrow{AC}$, denoted $\angle BAC$, where A, B, C are not collinear.

Two opposite rays $\overrightarrow{AB}$ and $\overrightarrow{AC}$ are said to form a **straight angle** and two equal rays $\overrightarrow{AB}$ and $\overrightarrow{AC}$ are said to form a **zero angle** even though these are not angles by the definition. This terminology is so well established that we will use it. But to be clear, if we say "angle" or refer to $\angle BAC$ without explictly including zero or straight angles, this means an angle by the definition and does not include zero or straight angles.

Axiom 3 (Protractor). *The plane has an* ***angle measure*** *that maps any two rays* $\overrightarrow{AB}$ *and* $\overrightarrow{AC}$ *with common endpoint to a real number* $m\angle BAC$.

Figure 2. A Wooden Protractor Measuring from 0 to 180

- *The angle measure maps any angle* $\angle BAC$ *to a positive number* $m\angle BAC$*, with* $0 < m\angle BAC < 180$.
- *In addition, when* $\overrightarrow{AB} = \overrightarrow{AC}$*, the angle measure of this zero angle is* $m\angle BAC = 0$. *If the rays* $\overrightarrow{AB}$ *and* $\overrightarrow{AC}$ *are opposite rays, the angle measure of this straight angle is* $m\angle BAC = 180$.
- *For any point* A*, there exists a* ***protractor*** *at* A*. This is a mapping* δ *from the real numbers onto the rays with endpoint* A*, such that* $\delta(t) = \delta(t + 360)$ *for every* t*, with the property that* $m\angle BAC = |b - c|$ *when* $\delta(b) = \overrightarrow{AB}$ *and* $\delta(c) = \overrightarrow{AC}$*, provided that* $|b - c| \le 180$.

Note. A protractor is also called a **polar angle parametrization** at A. The degree is the unit of angle measure used here, since this is more convenient for geometry than radians. However, the number 360 could be replaced by another positive number to provide a different choice of unit.

Example 2.4. Let's try to connect this angle measurement with our geometric intuition. Imagine that you are holding a compass (the magnetic directional kind, not the kind that draws circles). The directions on the compass face are numbered from 0 to 360, with the directions precisely at 0 and 360 being the same direction. Suppose this 0 direction points rightward (eastward).

If you are standing at point A, then the rays with endpoint A indicate the directions that you can be facing. Suppose direction $\overrightarrow{AB}$ is numbered 40, sort of a northeast direction, and a direction $\overrightarrow{AC}$ is numbered 250, sort of southwesterly. Then we could try to compute the angle measure of $\angle BAC$ by taking the difference between the numbers: $|250 - 40| = 210$. But this number is greater than 180, so it will not work as an angle measure. However, the number $250 - 360 = -110$ also corresponds to the ray $\overrightarrow{AC}$, so we can try again with this number: $|-110 - 40| = 150$. This is between 0 and 180, so 150 is the measure of $\angle BAC$.

Polar Angles. Using protractors directly is somewhat awkward; the inverse map from rays to real numbers is often more convenient. For a given protractor δ at A, we will say the point B (or the ray $\overrightarrow{AB}$) has **polar angle** t if t is one of the real numbers for which $\delta(t) = \overrightarrow{AB}$.

Two distinct rays never have the same polar angle. By Axiom 3, if the polar angles were the same, the angle measure of the rays would be zero and the rays would form a zero angle.

If two values of the polar angles of $\overrightarrow{AB}$ and $\overrightarrow{AC}$ satisfy $|\theta(\overrightarrow{AB}) - \theta(\overrightarrow{AC})| < 180$, we say the polar angles *measure* $\angle BAC$.

A choice of polar angle for $\overrightarrow{AB}$ will be denoted by $\theta(\overrightarrow{AB})$, even though the value of this angle is ambiguous without context, since adding 360 to this number produces another polar angle. To get around this problem in a given situation and to choose a polar angle unambiguously, we can restrict the protractor δ to a half-open segment in the real numbers of length 360.

For example, given the half-open interval $[0, 360)$, which is the set of real numbers t, with $0 \leq t < 360$, any protractor at A will define a one-to-one correspondence between this interval and the rays with endpoint A. Therefore, $\theta(\overrightarrow{AB})$ can be defined as the inverse of the restriction of δ to this interval, or in other words, $\theta(\overrightarrow{AB})$ will be the polar angle of B with $0 \leq \theta(\overrightarrow{AB}) < 360$.

Another choice of interval that is sometimes convenient is $(-180, +180]$, the real numbers $-180 < t \leq 180$. More generally, one can choose unambiguous polar angles θ either in the interval $a \leq \theta < a + 360$ or the interval $a - 180 \leq \theta < a + 180$, where a is any chosen real number.

If $\delta(t)$ is a protractor function at A, so are $\delta(t+k)$ and $\delta(-t+k)$, since these changes of real numbers do not affect the absolute value of the differences. This provides a convenience for computation: one may assign polar angle 0 to any convenient ray $\overrightarrow{AB}$.

Some Angle Terminology. There are a number of terms that specify a relationship between two angles or describe the size of an angle.

Supplementary Angles: Two angles are supplementary if the sum of their angle measures is 180. The vertices of the two angles do not have to be the same, but if $\overrightarrow{AB}$ and $\overrightarrow{AC}$ are opposite rays and D is a point not on $\overline{AB}$, then $\angle BAD$ and $\angle CAD$ are supplementary angles.

Vertical Angles: Two angles $\angle BAC$ and $\angle B'AC'$ are vertical angles if $\overrightarrow{AB'}$ is the opposite ray of $\overrightarrow{AB}$ and $\overrightarrow{AC'}$ is the opposite ray of $\overrightarrow{AC}$. These vertical angles have the same angle measure.

Right Angle: A right angle is an angle with measure 90. Two lines $\overline{AB}$ and $\overline{AC}$ are **perpendicular** if $\angle BAC$ is a right angle.

Acute Angle: An acute angle is an angle with measure less than 90, the measure of a right angle.

Obtuse Angle: An obtuse angle is an angle with measure greater than 90, the measure of a right angle.

Notice that a supplementary angle of a right angle is also a right angle. Therefore, two perpendicular lines form four right angles.

Definition 2.5. The **perpendicular bisector** of a segment AB is the line through the midpoint of AB perpendicular to $\overline{AB}$.

The perpendicular bisector of a segment always exists, since the midpoint M exists by the Ruler Axiom and a right angle $\angle DMA$ exists by the Protractor Axiom.

Plane Separation

There is one more axiom that is important for relationships of order and separation. It ensures that our model really acts like a two-sided two-dimensional space.

Axiom 4 (Plane Separation). *For any line m, any point not on the line belongs to one of two disjoint sets called the* ***half-planes of the line****.*

- *Points P and Q are in different half-planes when PQ intersects m. They are in the same half-plane if PQ does not intersect m.*
- *Let $m = \overline{AB}$. If a protractor is chosen so that $\theta(\overrightarrow{AB}) = 0$, the points P for which $0 < \theta(\overrightarrow{AP}) < 180$ lie in one half-plane, and the Q for which $-180 < \theta(\overrightarrow{AQ}) < 0$ are in the other half-plane.*

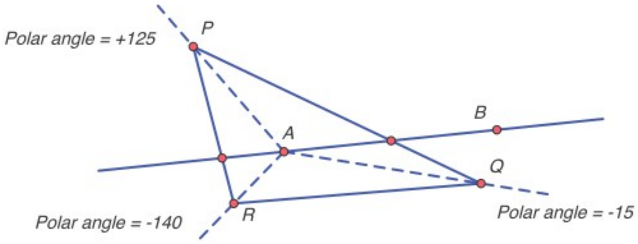

Figure 3. Point P in the Half-plane Opposite the Half-plane of Q and R

Note. We refer to the two half-planes as **opposite** one another. If two points are in different half-planes, we say they are on opposite sides of the line.

Definition 2.6. Given an angle $\angle BAC$, the **interior of the angle** is the intersection of the half-plane of $\overline{AC}$ containing B and the half-plane of $\overline{AB}$ containing C. For $\triangle BAC$, the **interior of the triangle** is the intersection of the interiors of the three angles $\angle BAC, \angle ACB, \angle CBA$.

Definition 2.7. A polygon $P_1P_2 \dots P_n$ is a **convex polygon** if, for every line $\overline{P_kP_{k+1}}$ that is a side extended, all the other vertices are in one half-plane of the line. The **interior of a convex polygon** is the intersection of all these half-planes

It follows that in a convex polygon, for any vertex angle, the interior of the polygon is contained in the interior of the angle. Since half-planes are convex, the interior of an angle or the interior of a triangle or other convex polygon is convex also.

The definition of the interior of a nonconvex polygon in general is surprisingly technical and will not be attempted here, since it would be an inappropriate digression in an introductory work. For a polygon with only a few sides, such as a quadrilateral, it is not hard to dissect the figure into triangles or other convex polygons and then take the interior of the convex polygons, plus common edges, to be the interior. But in general one must account for polygons with hundreds of thousands of sides. And this is not just some sort of "abstract mathematics" problem, for such triangles really appear in computer graphics and animation, and they can twist into amazingly complicated shapes. One algorithm for determining which points are inside a polygon is to intersect the polygon with a line passing through sides but not vertices. Then the points of intersection of sides with the line divide the line into alternating segments of outside points and inside points. But just to prove that the number of intersection points is even, so that this algorithm works, is complicated and to show that the interior is well-defined is even more so.

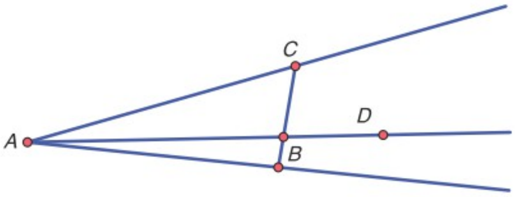

Figure 4. Intersection of CD with Ray Through Interior Point D

Theorem 2.8 (Addition of Angle Measure). *A point D is in the interior of $\angle BAC$ if and only if some polar angle of $\overrightarrow{AD}$ is between the polar angles of $\overrightarrow{AB}$ and $\overrightarrow{AC}$, provided those angles are chosen so that they measure $\angle BAC$. In this case, $m\angle BAD + m\angle DAC = m\angle BAC$.*

Proof. Let $\theta(\overrightarrow{AB}) = b$, $\theta(\overrightarrow{AC}) = c$, $\theta(\overrightarrow{AD}) = d$, with $|b - c| < 180$. Then we need to prove that D is an interior point if and only if some choice of d is between b and c.

Assume without loss of generality that b is the smaller of b and c, so $b < c < b+180$. D is in the half-plane of $\overline{AB}$ containing C when for some choice of d, $b < d < b + 180$. D is also in the half-plane of $\overline{AC}$ containing B when $d < c$. Thus, $b < d < c$ if and only if D is in the interior of the angle.

The angle measures are $m\angle BAD = d - b$, $m\angle DAC = c - d$, and $m\angle BAC = c - b$. This proves the final statement. □

If D is in the interior of $\angle BAC$, one also says $\overrightarrow{AD}$ is between $\overrightarrow{AB}$ and $\overrightarrow{AC}$. All the points of this ray have polar angle d, so they are interior to $\angle BAC$.

Definition 2.9. The **angle bisector** of $\angle BAC$ is the ray $\overrightarrow{AD}$ for which $m\angle BAD = m\angle DAC = (1/2)m\angle BAC$.

Using the notation from the proof of the theorem, the angle bisector exists and has polar angle $d = (b + c)/2$.

Theorem 2.10 (Crossbar). *Given any angle $\angle BAC$:*

(a) *The interior points of BC are interior to $\angle BAC$.*

(b) *For any point D interior to $\angle BAC$, the ray $\overrightarrow{AD}$ intersects the segment BC.*

Proof. Statement (a) is true since B is on $\overline{AB}$, so the interior points of $\overrightarrow{BC}$ are contained in the half-plane of $\overline{AB}$ containing C. Likewise, the interior points of $\overrightarrow{CB}$ are contained in the half-plane of $\overline{AC}$ containing B. The set of interior points of BC is the intersection of the interior points of these rays.

To show (b), continue to use the polar angles b, c, and d as above. Since $b < d < c$, we see that B and C are on opposite sides of $\overline{AD}$. Thus, BC intersects $\overline{AD}$ at a point in the interior of $\angle BAC$, therefore, at a point of $\overrightarrow{AD}$. □

Rigid Motions and Lines

We will now use our ruler and protractor axioms to prove some essential properties for doing geometry with rigid motions.

Theorem 2.11 (Line Images). *If T is a rigid motion, the T-image of a line, segment, or ray is, respectively, a line, segment, or ray.*

Proof. If A, B, C are distinct collinear points on line n, then $m\angle BAC = 0$ or $m\angle BAC = 180$. Then $m\angle T(B)T(A)T(C)$ also equals 0 or 180. This proves that the image of a line is a line.

Segments and rays are subsets of lines defined by distance relations. A point C on n is in segment AB if and only if $\|AC\| + \|BC\| = \|AB\|$. This equation still holds true for the image points, so $T(C)$ is in the segment $T(A)T(B)$. Similar reasoning for rays shows that the image of $\overrightarrow{AB}$ is $\overrightarrow{T(A)T(B)}$. □

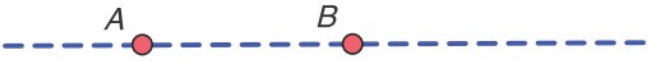

Figure 5. A and B fixed — All Line Points Fixed

This next theorem will play a key role in classifying rigid motions.

Theorem 2.12 (Two Fixed Points). *If a rigid motion T fixes two distinct points A and B, then it fixes all points of $\overline{AB}$.*

Proof. To say T fixes a point A means that $T(A) = A$.

For a point C on $\overline{AB}$, let $D = T(C)$. If D is distinct from C, then $\|DA\| = \|T(C)T(A)\| = \|CA\|$, so A is the midpoint of CD. The same reasoning shows that B is also the midpoint of CD, so $A = B$, a contradiction. So D must be C. □

The Other Axioms

The last two axioms are stated here. What follows from assuming these axioms will be investigated in coming chapters.

The Reflection Axiom is an existence axiom for a very particular kind of rigid motion, not described in much detail in the axiom. The geometry of this rigid motion will be described in Chapter 3. Then by composing the transformations for different lines, in coming chapters we will prove many theorems of geometry and will generate familiar transformations such as rotations and translations. Already in Chapter 3 we will begin to see consequences for the geometry of triangles and other figures.

Axiom 5 (Reflection). *For any line m, there exists a rigid motion distinct from the identity transformation that fixes the points of m.*

We note by Theorem 2.12 that it is actually sufficient that this rigid motion fixes two points.

The last of the six axioms concerns the properties of dilations. These transformations are not rigid motions but are transformations that will be defined later as similitudes.

Definition 2.13. For point A and a real number $k > 0$, the **dilation** centered at A with dilation ratio k, denoted $\mathcal{D}_{A,k}$, is a transformation of the plane defined as follows: (a) $\mathcal{D}_{A,k}(A) = A$; (b) if $X \neq A$, $X' = \mathcal{D}_{A,k}(X)$ is the point on $\overrightarrow{AX}$ with $\|AX'\| = k\|AX\|$.

Axiom 6 (Dilation). *For every point A and every $k > 0$, the dilation $\mathcal{D}_{A,k}$ preserves angle measure and scales every distance by k: for all $P' = \mathcal{D}_{A,k}(P)$ and $Q' = \mathcal{D}_{A,k}(Q)$, $\|P'Q'\| = k\|PQ\|$.*

Note the difference in the nature of these two axioms. The Reflection Axiom is an existence axiom while the Dilation Axiom asserts properties of an existing transformation that is already defined.

We will see in Chapter 7 that the Dilation Axiom implies the famous Euclidean Parallel Postulate as a theorem. In contrast, everything proved using only the first five axioms is true not only in Euclidean geometry, but in hyperbolic non-Euclidean geometry as well. For many readers, this will not be important, but anyone thinking about proving theorems that are true in both Euclidean and non-Euclidean geometry will note that, except for a very few well-identified and isolated exceptions, only the first five axioms are used in proofs until the end of Chapter 6.

In Chapter 8, dilations will be central in the study of similarity. At that time, the definition of dilation will be extended to negative k.

Exercises and Explorations

1. (Rulers and Protractors). Let $\rho(x)$ be a ruler for a line m. It was claimed in Remark 2.1 that for a constant real number k, both $\rho(x+k)$ and $\rho(-x+k)$ are also rulers. Later it was similarly noted that if $\delta(t)$ is a protractor function at A, so are $\delta(t+k)$ and $\delta(-t+k)$.
   (a) Show that $\rho(x+k)$ satisfies the definition of a ruler. Next show that $\rho(-x)$ is a ruler. Combine the two results to prove the statement in the remark. You may wish to supplement words and symbols with drawings.
   (b) Show how almost identical reasoning justifies the statement about protractors.
2. (Midpoints). Let points A and B correspond by a ruler on line $\overline{AB}$ to a and b.
   (a) What real number m corresponds to the midpoint M of AB?
   (b) If D is a point corresponding to real number d and if B is the midpoint of AD, what number is d?
3. (Angle Bisectors). From Definition 2.9, an angle bisector is a ray that divides an angle into two angles of equal measure.
   (a) For a protractor at A, let the polar angles of B, C, and D be b, c, d. Assuming $|b-c| < 180$ and D is interior to $\angle BAC$, for what value of d will $\overrightarrow{AD}$ bisect this angle? In other words, for what D is $m\angle BAD = m\angle DAC$?
   (b) Two lines $\overline{AB}$ and $\overline{AC}$ form four angles at A. Explain how the four angle bisectors are related.
4. (Relationships on a Line). Let points A and B correspond by a ruler on line $\overline{AB}$ to a and b.
   (a) If $a = 3$ and $b = 5$, what number c corresponds to the point C on the line with $\|AC\| = 7$ and $\|BC\| = 9$?
   (b) If points A, B on line m correspond by a ruler to real numbers 3 and 5, what number c corresponds to the point $C = \mathcal{D}_{A,5}(B)$?
   (c) What is a general formula for c in terms of a and b if $C = \mathcal{D}_{A,k}(B)$?
5. (Fixed points on $\mathbb{R}$). Suppose that f is a map from the real numbers $\mathbb{R}$ to $\mathbb{R}$ that preserves distance.
   (a) If $f(3) = 3$, what are the possible values of $f(7)$?
   (b) If $f(3) = 3$, what are the possible values of $f(t)$, for any t?
   (c) If a point a is a fixed point of f (i.e., $a = f(a)$), what are the possible values of $f(t)$ for any other number t?
   (d) Based on your answers, if f has two fixed points, what can you say about $f(t)$?
6. (Plane Separation and Triangles). The plane separation axiom allows one to prove some facts that are basic and visually very believable but rather technical. One may choose to believe them without going through the proofs. However, they do provide good practice in reasoning with the Plane Separation Axiom.
   (a) In $\triangle ABC$, let D be an interior point of BC and let E be an interior point of CA. Prove that AD and BE intersect at a point G, an interior point of the triangle.

(b) (Pasch Axiom). There are other axioms of plane separation that are used in place of Axiom 4. One is called the Pasch Axiom: if a line m intersects side AB of $\triangle ABC$ at one point D, an interior point of side AB, then m also intersects either BC or CA. Prove this statement as a theorem using the axioms in this chapter.

(c) (Quadrilateral Shapes). Sketch a few quadrilaterals, convex and nonconvex. In your examples, is there always at least one diagonal that separates the other two vertices into opposite half-planes of the diagonal? If this is always true, how can one prove it from the axioms? If there are two such diagonals, what does this imply about the convexity or nonconvexity of the quadrilateral?

Suggestion: For each statement, begin by drawing possible cases and explain with informal reasoning what is possible before undertaking a more formal explanation.

Chapter 3

Existence and Properties of Reflections

In the previous chapters, we have been describing properties of rigid motions, but without the Reflection Axiom, there is nothing that tells us that any rigid motions actually exist (except for the identity)!

This axiom states the existence of some rigid motions that will turn out to be line reflections, with the usual geometric description. However, this description is not stated in the axiom. A bare minimum of existence is promised; in fact it is not even asserted that there is only one rigid motion whose existence is declared. But using the strong restrictions on behavior mandated by preservation of distance and angle measure, we will see that these vaguely described rigid motions must behave in exactly one way.

Recall the statement of the axiom:

Axiom 5 (Reflection)**.** *For any line m, there exists a rigid motion distinct from the identity transformation that fixes the points of m.*

Deducing the Properties of Reflections

At this point, we do not have a geometrical description of this rigid motion. The proof that this rigid motion is the usual reflection in a line with the usual geometrical description will come from the examination of a key figure.

Lemma 3.1 (Y Figure)**.** *Given a line $\overline{AB}$, take any point C not on the line. Then there is exactly one other point D in the plane so that $\|AC\| = \|AD\|$ and $m\angle BAC = m\angle BAD$. The points C and D are in opposite half-planes of $\overline{AB}$.*

Proof. Choose the protractor δ for rays with endpoint A so that the polar angle $\theta(\overrightarrow{AB}) = 0$. In the following, we choose polar angles in the interval $-180 < \theta < +180$.

Let $m\angle BAC = t$. Then the only rays $\overrightarrow{AE}$ for which $m\angle BAE = t$ have polar angle $+t$ or $-t$. There are only two such rays. One is $\overrightarrow{AC}$ itself. Call the other one $\overrightarrow{AF}$.

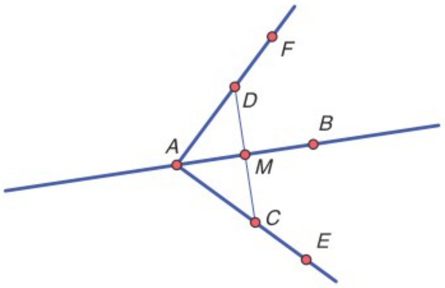

Figure 1. Only One Match to the Angle and Distance of C

Let $||AC|| = k$. There is exactly one point on each ray at distance k from A. One of these points is C. Let the point on $\overrightarrow{AF}$ be D.

Then $||AC|| = ||AD|| = k$, and $m\angle BAC = m\angle BAD = t$. From the Plane Separation Axiom, these points lie in opposite half-planes of $\overline{AB}$, so segment CD intersects $\overline{AB}$. $\square$

Lemma 3.2 (Reflection and Perpendicular Bisectors). *For any $\overline{AB}$, let T be a rigid motion that fixes the points of $\overline{AB}$. If C is not a fixed point of T, then C is not on $\overline{AB}$ and $T(C)$ is the point D constructed in the Y Figure Lemma. Moreover, the segment CD intersects $\overline{AB}$ at M, the midpoint of CD, and $\overline{AB}$ is the perpendicular bisector of CD.*

Proof. Let $D = T(C)$, a point distinct from C. T is a rigid motion that preserves angle measure and distance. Therefore, $m\angle BAC = m\angle BAD$ and $||AC||| = ||AD||$.

So from the Y Figure Lemma, D is uniquely determined and on the opposite side of $\overline{AB}$ from C.

Let M be the intersection of CD and $\overline{AB}$. (If $M = A$, interchange the roles of A and B in the proof.)

Since M is on $\overline{AB}$, $T(M) = M$. Since T is a rigid motion, $||CM|| = ||DM||$. Therefore M is the midpoint of CD.

In addition $m\angle CMA = m\angle DMA$ because T maps one angle to the other. Since the sum of these angle measures is 180, each of these angles is a right angle. This means that $\overline{AB}$ is the perpendicular bisector of CD. $\square$

Theorem 3.3 (Three Fixed Points). *If T is a rigid motion with three noncollinear fixed points, then T is the identify transformation. Furthermore, if S and T are two rigid motions that agree at three noncollinear points, then the transformations are the same: $S = T$.*

Proof. Let A, B, C be noncollinear fixed points of T. Suppose there is a point G that is not a fixed point, so that $T(G) = H$ is distinct from G.

Then by the Two Fixed Points Theorem, Theorem 2.12, for each of the three lines $\overline{AB}$, $\overline{BC}$, $\overline{CA}$, all the points are fixed. Then by Lemma 3.2, $\overline{AB}$ is the perpendicular bisector of GH and $\overline{AC}$ is also the perpendicular bisector of GH, as shown in Figure 2.

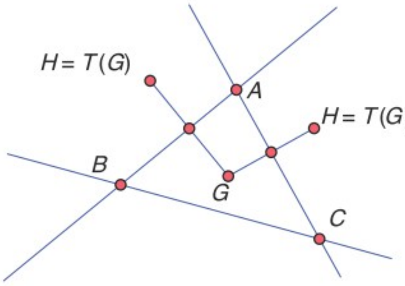

Figure 2. Impossible Images of G with A, B, C Fixed

Thus A, B, C are collinear with this assumption on G. Therefore, no such G exists and $G = T(G)$ for every G. In other words, T is the identity transformation I.

To prove the rest of the claim, let $S(A) = T(A)$, $S(B) = T(B)$, and $S(C) = T(C)$. Then $S^{-1}T$ has A, B, C as fixed points, so $S^{-1}T = I$. Multiplying on the left by S yields $T = (SS)^{-1}T = S(S^{-1}T) = SI = S$. ☐

This is a very powerful result. One can model it physically by tacking a piece of paper to a bulletin board. One tack allows a lot of swinging movement; two tacks allow no movement unless you remove the tacks, flip the paper over, and retack in the same holes. Three tacks allow no movement at all.

Next is the theorem that completes the geometric description of reflection for every point.

Theorem 3.4 (Line Reflection)**.** *Given any line m, there exists a unique rigid motion R distinct from the identity that fixes the points of line m. For any point E not on m, the image $E' = R(E)$ is the point such that m is the perpendicular bisector of EE'. We name this unique rigid motion* ***reflection in line*** *m, denoted $\mathcal{R}_m$. Reflection in $\overline{AB}$ is denoted $\mathcal{R}_{AB}$.*

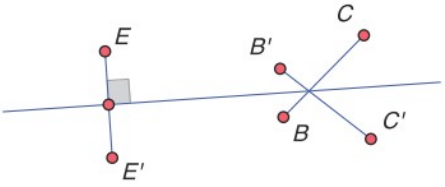

Figure 3. Reflection in the Perpendicular Bisector of EE'

Proof. By Theorem 3.3, an R given by the Reflection Axiom fixes the points of m but does not fix E (or else it would be the identity). Therefore by Lemma 3.2, this description of $R(E)$ is correct for every E not on m. ☐

Corollary. *If line p is perpendicular to line m, then $\mathcal{R}_m(p) = p$.*

Corollary. *For any line m, the inverse of* $\mathcal{R}_m$ *is* $\mathcal{R}_m$*.*

Proof. The perpendicular bisector of DC is the same as the perpendicular bisector of CD. ☐

The geometric description of line reflection also tells us some useful properties concerning intersection and half-planes.

Theorem 3.5 (Line Reflection and Half-Planes)**.** *Let m be a line.*

(1) *The reflection* $\mathcal{R}_m$ *maps each half-plane of m to the opposite one.*

(2) *If line n is perpendicular to m,* $\mathcal{R}_m$ *maps each half-plane of n to itself.*

Proof. By the Line Reflection Theorem, Theorem 3.4, any point A not on m is mapped to the opposite half-plane of m. This proves the first statement. For the second statement, let A be a point of m not on n and let B be a point in the same half-plane of n as A, so AB does not intersect n. Since $\mathcal{R}_m(n) = n$, the reflection image of the segment AB cannot intersect n, so $\mathcal{R}_m(A) = A$ and $\mathcal{R}_m(B)$ must be in the same half-plane. Since A did not move, B remains in its original half-plane of n also. ☐

Theorem 3.6 (Perpendicular Parallels)**.** *Let m be a line, and let p and q be distinct lines, each perpendicular to m. The lines p and q are parallel.*

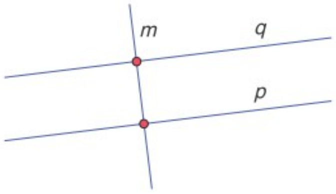

Figure 4. Perpendicular to Line m so Parallel

Proof. Suppose p and q intersect at a point A not on m. Reflection in m maps p to p and q to q. Therefore, $\mathcal{R}_m(A)$ is also a second, distinct intersection point of p and q in the opposite side of m from A. This means p and q have two points in common, so they must be the same line. This contradicts the assumption that p and q intersect m at distinct points. ☐

Theorem 3.7 (Dropping a Perpendicular)**.** *Let m be a line. If P is any point, there is a unique line p through P that is perpendicular to m.*

Proof. If P is on m, then there is a unique line through P that is perpendicular to m by the Protractor Axiom. If P is not on m, let $P' = \mathcal{R}_m(P)$. The line $\overline{PP'}$ is perpendicular to m by the Line Reflection Theorem. There is only one such line, for any two lines perpendicular to m are parallel so cannot both pass through P. ☐

The traditional name for construction of the perpendicular line to m through P is called **Dropping a Perpendicular from a Point**. Presumably one is dropping a vertical line like a plumb bob on a string to intersect a horizontal line. The intersection point of the perpendicular line with m is called the **foot of the perpendicular** from P to m.

Isosceles Triangles and Kites

With the properties of line reflection established, we can prove all the usual relationships in any isosceles triangle. Recall from Definition 2.2 that a **triangle** is a three-sided polygon. To spell this out: $\triangle ABC$ has three noncollinear vertices A, B, C and three sides AB, BC, and CA. The (interior) angles are the angles $\angle CAB$, $\angle ABC$, $\angle BCA$, often referred to as "angle A", etc.

Definition 3.8 (Isosceles Triangle)**.** An **isosceles triangle** is a triangle with two sides of equal length. If we label the two equal sides as AB and AC, then the side BC is called the base of the triangle. The angles at B and C are called base angles. The angle at A is called the apex angle.

A triangle with three equal sides is called an **equilateral triangle**.

This theorem points out a lot of congruent parts in an isosceles triangle. The theorem is labeled as a theorem about symmetry, since it shows that an isosceles triangle has a **line of symmetry**, which means that reflection in the line maps the figure onto itself.

Theorem 3.9 (Isosceles Triangle Symmetry)**.** *Given an isosceles triangle ABC with $\|AB\| = \|AC\|$, the perpendicular bisector of the base BC passes through A and bisects $\angle BAC$. Also, the base angles $\angle BCA$ and $\angle CBA$ are congruent.*

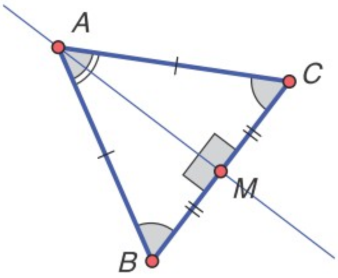

Figure 5. Symmetry and Many Congruences in an Isosceles Triangle

Proof. Construct $\overrightarrow{AP}$ as the angle bisector of $\angle BAC$. Then the Y Figure Lemma, Lemma 3.1, applies, and B and C are the mirror points in the lemma with respect to $\overline{AP}$. This means C is the reflection of B in $\overline{AP}$ and $\overline{AP}$ intersects BC at the midpoint M. Therefore the line $\overline{AP} = \overline{AM}$ is the perpendicular bisector of the base BC. The base angles are congruent since $\mathcal{R}_{AM}$ maps one to the other. $\square$

The traditional word for a set of points satisfying a geometrical condition is **locus**. This next theorem says that the locus of points equidistant from A and B is the perpendicular bisector of AB.

Theorem 3.10 (Perpendicular Bisector as Locus). *For any distinct points B and C, the perpendicular bisector of BC is the set of points P for which $\|PB\| = \|PC\|$.*

Proof. To prove set equality, one must prove containment in two directions.

If $\|PB\| = \|PC\|$, then the isosceles theorem proves that P is on m, the perpendicular bisector of BC, except in the case when $\angle BPC$ is a straight angle. But in this case, P is the midpoint of BC, so it is on m. For the converse, let Q be any point on m. Then $\mathcal{R}_m(B) = C$ and $\mathcal{R}_m(Q) = Q$, so $\|QB\| = \|QC\|$ since $\mathcal{R}_m$ is a rigid motion. □

A **quadrilateral** was defined in Definition 2.2 as a polygon with four vertices. A quadrilateral $ABCD$ has four vertex points A, B, C, D, no three of which are collinear, and four sides AB, BC, CD, DA, with no sides intersecting at interior points.

A **kite** is a quadrilateral with two pairs of equal adjacent sides, e.g., $\|AB\| = \|AD\|$ and $\|CB\| = \|CD\|$.

Theorem 3.11 (Kite Symmetry). *If $ABCD$ is a kite with $\|AB\| = \|AD\|$ and $\|CB\| = \|CD\|$, the diagonal $\overline{AC}$ is the perpendicular bisector of the segment BD.*

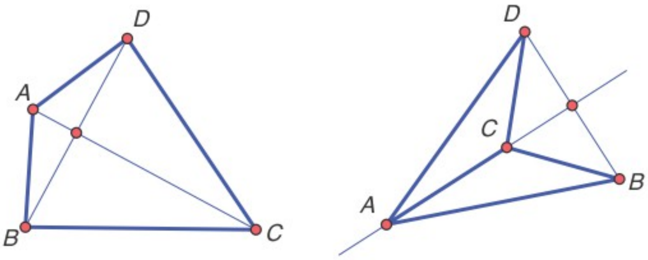

Figure 6. Line Symmetry of Kites

Proof. Let M be the midpoint of BD. Apply the Isosceles Triangle Symmetry Theorem, Theorem 3.9, to $\triangle ABD$ and $\triangle CDB$. In $\triangle ABD$, the perpendicular bisector of the base BD contains the point A and bisects $\angle BAD$. Likewise, in $\triangle CDB$ the same perpendicular bisector of BD contains C and bisects $\angle DCB$. This says that the line $\overline{AC}$ is the perpendicular bisector of BD. This line also bisects the angles at A and C. □

Notice that one of these kites in Figure 6 is not convex, but $\overline{AC}$ still is the perpendicular bisector of BD.

The proof of this theorem about equal base angles is left to Exercise 1.

Theorem 3.12 (Equal Base Angles). *For a triangle $\triangle ABC$, if $m\angle ABC = m\angle BCA$, the triangle is isosceles.*

Circles and Lines

A circle with center P was defined in Definition 1.6 as the set of points at distance r from P, for some positive radius r. This means that if A and B are two points on the circle, either P is the midpoint of AB or $\triangle PAB$ is isosceles with base AB. Combining this observation with what we have learned about isosceles triangles can tell us about a number of other properties of circles.

Some terminology will be used for a circle with center P:

- A segment PA from the center to a point on the circle is called a **radius segment**, or just a radius if the meaning is clear.
- A segment AB with A and B on the circle and with P as midpoint is a **diameter**; both PA and PB are radius segments. A line through P is a **diameter line**.

Reflection in any line through P maps the circle into itself, and for any reflection that maps the circle into itself, the reflection line passes through the center P. (The proof has been left as a problem in the exercises.)

The intersections of a circle with a line can be characterized using what we know.

Theorem 3.13 (Line-Circle Intersection)**.** *Let c be a circle with center P and radius r. If m is a line, there are four ways in which the line can meet the circle.*

- *The line and the circle may not intersect.*
- *The line and the circle may intersect at one point A. Segment PA is perpendicular to m.*
- *The line and the circle may intersect at two points A and B, as a diameter line or not. Center P is on the perpendicular bisector of AB; it will be the midpoint if AB is a diameter.*
- *The line cannot intersect the circle in more than two points.*

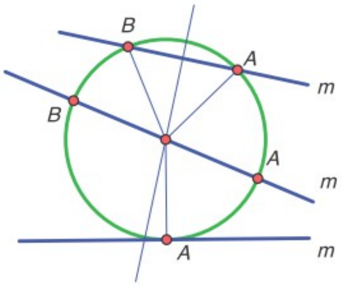

Figure 7. Circle and Possible Line Intersections

Proof. If A and B are two distinct points of intersection of circle c and line m, then by the Perpendicular Bisector as Locus Theorem, Theorem 3.10, the center P is on the perpendicular bisector of AB. This is the unique line n through P perpendicular to m.

If A, B, and C are three distinct points of intersection of c and m, then n must be the perpendicular bisector of both AB and AC, which is impossible since this would mean $B = \mathcal{R}_n(A) = C$.

To prove the third case and then the second case, suppose A is a point of intersection of c and m. Then $\mathcal{R}_n$ must map A to a point of intersection, since a reflection in a line through P maps c to itself and reflection in a perpendicular line maps m to itself. If $B = \mathcal{R}_n(A)$ is distinct from A, then the third case holds. If $B = A$, then A must be on n, so $n = \overline{PA}$. In this case $B = A$ is the only point of intersection. If there were another intersection point C, then $D = \mathcal{R}_n(C)$ would be a third point of intersection, a case that has been ruled out above. Thus, in this case the second statement holds.

Finally without trying to describe when this would happen, the first disjoint case may happen as well. □

The intersection of a circle and a line will be related to the distance of the line from the circle center in Theorem 6.8.

Light, Angles, and Reflections

When speaking of the physics of reflection of light rays, one hears the statement "angle of incidence equals angle of reflection". But when you look in a mirror, there is a copy of your face and your surroundings apparently just on the other side of the mirror, which seems to be a window into that scene.

Some of this can be seen in our plane geometry of reflections if one views a line as a plane mirror from above. In Figure 8, points A and B are on one side of the mirror line and the line reflection B' is on the other side. The observer at A sees a reflection of B in the mirror by looking in the direction of $\overrightarrow{AB'}$, a ray that intersects the mirror at C.

There are congruent vertical angles in this figure and also an isosceles triangle BCB', with congruent angles at the apex C. So the angle of incidence does equal the angle of reflection.

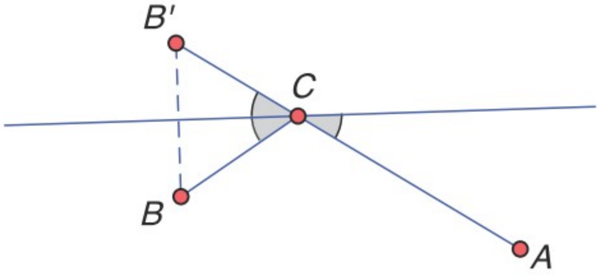

Figure 8. Line Reflection and Congruent Angles

Paper Folding and Tools for Construction

The proof of the geometric description of line reflection from the Reflection Axiom suggests why a paper fold through two points produces a nice, sharp straight line. The distance and angles on the paper are preserved by folding, so the destinations of all the points of the paper are determined, especially those of the points on the fold line.

This connection offers many ways to experiment with line reflections by folding. Some of these are identified as experiments in the exercises at the end of the chapter. In addition to revealing mathematical relations, a paper fold and some scissors can produce all kinds of figures with line symmetry, from paper dolls to abstract shapes.

Folding thin or translucent paper (such as patty paper) can also produce interesting traceable geometric figures, including ones that will appear in the next chapter.

Actual glass or plastic mirrors are also tools for experimentation, though it is hard to actually draw the reflection behind a mirror. For this reason, semitransparent red plastic mirrors have been offered for classroom experiments since they allow one to see both the figure behind the mirror and the reflection of the figure in front of the mirror at the same time.

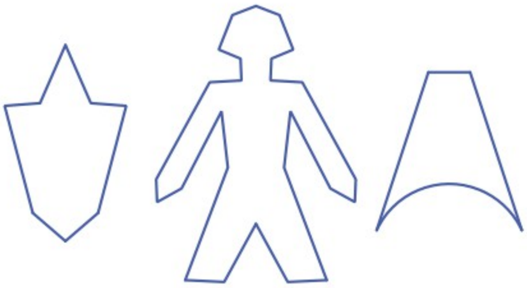

Figure 9. Line Symmetry from Folding and Cutting

The most versatile tool is dynamic geometry software that allows an abundance of examples with movable figures for experimentation.

Compasses and Constructions. Euclid's axioms for geometry were abstractions of drawing with an unmarked straightedge and a compass, so it was natural to use those tools when carrying out geometric constructions. The axioms in this book are abstractions of marked straightedges and protractors, so it may seem reasonable to use a marked ruler to measure distances when constructing geometric figures. However, compasses still are a more accurate way of constructing or copying lengths than using a physical ruler.

The main problem with a physical marked ruler is that there are only a finite number of marks, so most distances have to be approximated. A second reason is the process of making a pencil mark at a point on the ruler depends on more hand and eye coordination than making a mark with a compass. Therefore, a physical marked ruler is not really a good model for the ruler functions in the Ruler Axiom. A better model is a line drawn on a computer screen with software, since the software will plot (as accurately as the screen allows) any distance on the line.

In the exercises in the chapter, there are some directions for straightedge and compass constructions. These have the virtue of accuracy, but they also illustrate geometrical relationships between circles, lines, isosceles triangles and other figures. These methods complement the aforementioned paper folding, which is also a very accurate method for constructing certain lines.

Note about High School Texts. While transformations do not play a central role in most high school geometry textbooks, one exception is *Geometry* [6] from the University of Chicago School Mathematics Study Group. This book introduces line reflections early and uses them to prove such results as the symmetry properties of isosceles triangles. The set of axioms in that book assumes the parallel postulate before introducing line reflections and then assumes the basic properties of line reflections as postulates instead of proving them as theorems based on a general concept of rigid motion as is done here.

Exercises and Explorations

1. (Congruent Base Angles). Prove Theorem 3.12. Namely, prove for $\triangle ABC$ that if $m\angle ABC = m\angle ACB$, then $\|AB\| = \|AC\|$, so the triangle is isosceles.
2. (Rhombus Question). A **rhombus** is a quadrilateral $ABCD$ with all four sides the same length. How are the diagonals of a rhombus related? Construct a rhombus by paper folding.
3. (Ruler Reflection on $\mathbb{R}$). Let n be a line perpendicular to line m at A. Reflection in n maps m to itself. Identifying line m with $\mathbb{R}$ by using a ruler; this defines a mapping of the real number line $\mathbb{R}$ into itself. If the real number a corresponds to A and x corresponds to a point X on m, what number corresponds to $\mathcal{R}_n(X)$?
4. (Reflection of a Circle). Show that $\mathcal{R}_m$ maps a circle c with center O into itself, if and only if O is on m. Notice that this implies that all lines through the center (the diameter lines) are lines of symmetry for c.
5. (Segment Image Experiment). Try this experiment either with geometry software or paper. Draw a segment and a line not intersecting the segment, preferably at no special angle with respect to the segment. Then construct the reflection of the segment (paper folding and tracing is a good option if you are working on paper). You now have four endpoints. Draw all the segments connecting the endpoints. Do the diagonals intersect on the fold line? Why? What happens if you repeat the experiment with a segment and a fold through the segment? What happens with special cases such as a segment parallel to or perpendicular to the fold line?

6. (X Symmetry Experiment).
   (a) Draw an X shape made of two lines (or as much as will fit on the paper) as in Figure 10. They need not be segments of equal length. Fold the paper so that the shape is folded onto itself. This fold is a line of symmetry of the figure consisting of two lines. Fold another line of symmetry. How are these lines related to the angles defined by the lines? How are they related to each other?
   (b) Fold a piece of paper; then fold it again so that the original fold is folded onto itself. What angle results? How is this related to the previous experiment?

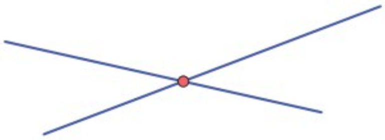

Figure 10. Folds Are Symmetry Lines for This X

7. (Free Folds). Fold a piece of paper and then cut a shape; unfold and see what shape with line symmetry you have produced. It could be a geometric shape or a tree or a face or something else you design.

8. (Billiards). Draw your own example of a figure that shows where to look to see a figure in a mirror (or in what direction to aim a billiard ball off a cushion). This idea can be extended to reflections in two mirrors in a corner, as illustrated in Figure 11. Make your own mirror lines, place points A and B, and find the billiard double cushion (or light double reflection) path between them.

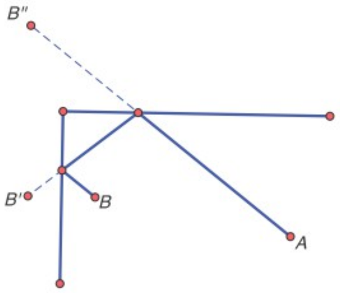

Figure 11. Double Mirror Reflection Path from A to B

9. (Symmetry Line Experiment). Find all the lines of symmetry for some familiar polygons: squares, rectangles, parallelograms, pentagons, hexagons (regular and otherwise). (Note: Some of these figures will not be "officially" defined until later chapters. This is informal experimentation at this stage.)

10. (Perpendicular Bisector Construction). Draw a segment AB on a piece of paper.
    (a) Folding construction: Construct the perpendicular bisector of AB by folding the paper. How did you fold?
    (b) Compass and straightedge construction: Construct the perpendicular bisector by drawing two intersecting circles of the same radius, one centered at A and the other at B. If there is one point of intersection not on line $\overline{AB}$, there will be two points of intersection. (Why?) Draw the line through the two points of intersection. Explain why this is the perpendicular bisector based on what was proved in this chapter.
11. (Angle Bisector Construction). Draw an angle $\angle BAC$ on a piece of paper.
    (a) Folding construction: Construct the angle bisector of AB by folding the paper.
    (b) Compass and straightedge construction: Using the compass, construct points D on $\overrightarrow{AB}$ and E on $\overrightarrow{AC}$ equidistant from A. Then construct F as in the previous problem to be an intersection point of two circles of the same radius with centers at D and E. What shape is $ADFE$ and what about its symmetry gives a reason why the ray $\overrightarrow{AF}$ is the angle bisector?

Chapter 7

Parallel Lines and Translations

This chapter marks an important point in the development of the geometry of the Euclidean plane. With the exception of a limited use of Triangle Angle Sum, proved by the Dilation Axiom, almost everything proved so far has used only the first five of our axioms. These axioms are also valid for the non-Euclidean plane geometry discovered by Lobachevsky and others in the nineteenth century, so theorems proved only with those axioms are true in both geometries. From now on, with the extensive use of the Euclidean Parallel Property, our text will be hewing exclusively to the Euclidean path.

The Euclidean Parallel Postulate

While we have defined parallel lines as lines that do not intersect, we have proved very little about parallel lines so far. This is because we need something additional to make our geometry conform to the intuition we have about rectangles, about lined notebook paper, and many other flat things. The key to making our geometry Euclidean is an axiom that is equivalent to the Euclidean Parallel Postulate.

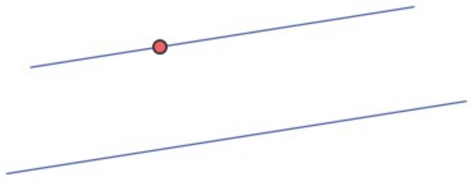

Figure 1. Playfair's Version of the Parallel Postulate

There have been many statements equivalent to the Parallel Postulate proposed as axioms over the centuries. The axiom that we have chosen in this book is Axiom 6, the Dilation Axiom. This axiom will be used to prove a theorem about parallel lines that is equivalent to Euclid's statement of the Parallel Postulate. The one we state and prove below is called the axiom of Playfair, named for its originator, the Scottish mathematician John Playfair.

Euclid stated the Parallel Postulate as an axiom, but since we are assuming an equivalent axiom, the Dilation Axiom, it will be proved here as a theorem using dilation.

Theorem 7.1 (Euclidean Parallel Postulate (EPP))**.** *Let m be a line. For any point P not on m, there is only one line through P parallel to m.*

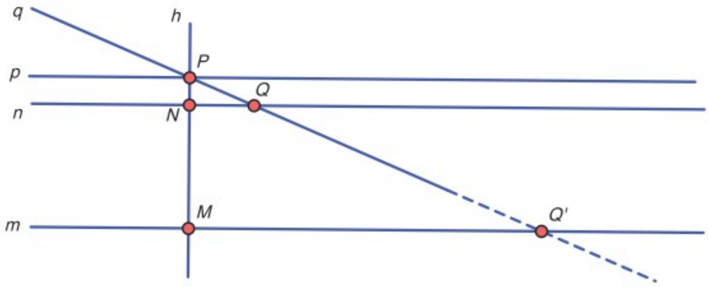

Figure 2. Dilation Proof of Parallel Postulate

Proof. Let h be the line through P perpendicular to m and let p be the line through P perpendicular to h. Let M be the intersection of h and m. Since both m and p are perpendicular to h, p is a line through P parallel to m by Theorem 3.6.

Now suppose that q is another line through P distinct from p. Since this line is not perpendicular to h, there is a point Q on the line with $m\angle MPQ < 90$.

Construct the line n through Q perpendicular to h, with N as the intersection of n and h. (All these perpendiculars can be constructed by line reflection.) These lines and points are shown in Figure 2.

Now we have a triangle $\triangle PNQ$, with a right angle $\angle PNQ$. The strategy is to dilate this triangle so that the image of Q is the intersection of q and m. Let $D_{P,k}$ be the dilation with center P that maps N to M. The value of k is $\|PM\|/\|PN\|$.

Since this dilation maps lines through P to themselves, the images of lines q and h are q and h. By the Dilation Axiom, which says angle measure is preserved by dilation, the right angle $\angle PNQ$ maps to a right angle $\angle PMQ'$. But $\overline{MQ'} = m$, since both are lines through M perpendicular to h.

This means that Q' is on both q and m. So these lines intersect, and q is not parallel to m. ☐

As an example of how one can use the EPP, here is a second proof of the Angle Sum Theorem, Theorem 4.12. If the EPP had been assumed as an axiom, this is the proof that would have come first. It shares some ideas with the proof in Chapter 4 but does not use dilation explicitly.

Theorem 7.2 (Angle Sum)**.** *For any triangle $\triangle ABC$, the images of $\triangle ABC$ by half-turns centered at the midpoints of the sides form a triangle $\triangle A'B'C'$, with $\triangle ABC$ as the midpoints. Therefore, the sum of the measures of the angles of $\triangle ABC$ is* 180 *degrees.*

Proof. Denote by F and E the midpoints of AB and AC, respectively.

If we apply $\mathcal{H}_F$ to $\triangle ABC$, $\mathcal{H}_F(A) = B$, $\mathcal{H}_F(B) = A$, and $\mathcal{H}_F(C) = C'$. The half-turn is a rigid motion, so $\angle BAC' \cong \angle ABC$. Since by Theorem 6.1 half-turns map any line to a line parallel to it, line $\overline{AC'}$ is parallel to $\overline{BC}$.

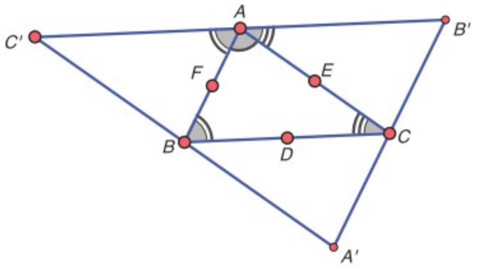

Figure 3. Angle Sum

Now apply $\mathcal{H}_E$ to $\triangle ABC$ and get the image $\triangle CB'A$. Again we see $\overline{AB'}$ is parallel to $\overline{BC}$, and $\angle CAB' \cong \angle ACB$.

But since lines $\overline{AC'}$ and $\overline{AB'}$ are lines through A parallel to BC, they must be the same line, so B', A, and C' are collinear. And A is the midpoint of $B'C'$. The same reasoning applies to the midpoints of the other two sides of $\triangle A'B'C'$.

This means that the sum of the measures of the three angles $\angle BAC'$, $\angle BAC$, and $\angle CAB'$ must be 180 degrees. Since these angles are congruent to the angles of $\triangle ABC$, this proves the result. □

Corollary. *The measure of an exterior angle of a triangle is equal to the sum of the measures of the interior angles at the other two vertices.*

Transversals and Parallel Lines

Definition 7.3. A line that intersects each of a pair of lines at two distinct points is called a **transversal** of the lines.

The Parallel Postulate says something important about transversals:

Theorem 7.4 (Transversal of Parallels)**.** *Suppose m and n are parallel lines. If p is a line distinct from m that intersects m, then p also intersects n, so p is a transversal.*

Proof. If p intersects m at M and does not intersect n, then p and m are two lines through M that are parallel to n. By the Parallel Postulate, the lines cannot be distinct. Therefore, p must intersect n. □

In the next theorem, a half-turn is used to establish important properties of transversals.

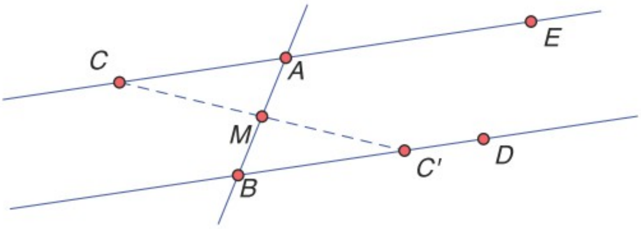

Figure 4. Transversal with Half-turn at Midpoint

Theorem 7.5 (Transversal Theorem). *Suppose* $\overline{AB}$ *is a transversal of* c *and* d*, with* A *on* c *and* B *on* d*. Let* $\mathcal{H}_M$ *be the half-turn with center* M*, the midpoint of* AB*.*

(1) *The line* $d = \mathcal{H}_M(c)$ *if and only if* c *and* d *are parallel.*

(2) *If* C *is not on* $\overline{AB}$*, let* $C' = \mathcal{H}_M(C)$*. Then* $\mathcal{H}_M(\overrightarrow{AC}) = \overrightarrow{BC'}$ *is a parallel ray on the opposite side of* $\overline{AB}$*. Also* $\overline{CB}$ *is parallel to* $\overline{C'A}$*.*

(3) *If* C *and* D *are on opposite sides of* $\overline{AB}$*, then* $m\angle BAC = m\angle ABD$ *if and only if* $\overline{CA}$ *is parallel to* $\overline{BD}$*.*

Proof. The line $\mathcal{H}_M(c)$ is parallel to c by Theorem 6.1. The Parallel Postulate says that if d is also a line through B parallel to c, then it must be the same line. This proves the first statement.

If $C' = \mathcal{H}_M(C)$, then CC' intersects $\overline{AB}$ at M. So if C is not on $\overline{AB}$, then C and C' are in opposite half-planes. Since $AC' = \mathcal{H}_M(BC)$, the lines $\overline{CB}$ and $\overline{C'A}$ are parallel. Also, since $\mathcal{H}_M$ is a rigid motion, $m\angle BAC = m\angle ABC'$.

If $m\angle BAC = m\angle ABC' = m\angle ABD$, with C and D on opposite sides of $\overline{AB}$, then $\overrightarrow{BD} = \overrightarrow{BC'}$. Therefore $\overline{CA}$ is parallel to $\overline{BD} = \overline{BC'}$. Conversely, if the lines are parallel and D is on the opposite side of $\overline{AB}$, then $\overrightarrow{BD} = \overrightarrow{BC'}$ since both are on the unique parallel to $\overline{CA}$ through B. □

This theorem can be used to prove the version of the Parallel Postulate stated in Euclid's *Elements*. It tells more about the location of the intersection of two nonparallel lines.

Theorem 7.6 (Euclid's Version of the Parallel Postulate). *Given two points* F *and* D *on the same side of* $\overline{AB}$*, if* $m\angle FAB + m\angle DBA < 180$*, then the rays* $\overrightarrow{AF}$ *and* $\overrightarrow{BD}$ *intersect at a point* G*.*

Proof. Let E be a point on the same side of $\overline{AB}$ as D, with $\overline{AE}$ parallel to $\overline{BD}$ and $\overrightarrow{AC}$ the opposite ray of $\overrightarrow{AE}$, as pictured in Figure 4. Let $d = m\angle ABD$ and $f = m\angle BAF$. Then $m\angle BAC = d$ and $m\angle BAE = 180 - d > f$, by Theorem 7.5. Therefore, $\overline{AF}$ is not the unique parallel to $\overline{BD}$ through A, so it intersects $\overline{BD}$ at a point G.

What remains to be proved is that G is on the rays $\overrightarrow{AF}$ and $\overrightarrow{BD}$. To see this, we prove the points F and D are in the interior of $\angle BAE$.

From the Addition of Angle Measure Theorem, Theorem 2.8, the interior points of $\overrightarrow{AF}$ are in the interior of $\angle EAB$. To show D is also in the interior we must show it is on the same side of $\overline{AB}$ as E (true by assumption) and on the same side of $\overline{AE}$ as B (true since the entire parallel line $\overline{BD}$ is on that side).

Now G cannot be on the opposite ray of $\overrightarrow{AF}$ because that ray is on the opposite side of $\overline{AE}$ from $\overline{BD}$, so the intersection point must be on $\overrightarrow{AF}$ and so must be in the interior of $\angle EAB$. But this means that G is on the same side of $\overline{AB}$ as D and E, so it must be on $\overrightarrow{BD}$ rather than its opposite ray. □

The angles $\angle BAC$ and $\angle ABD$ are called the **alternate interior angles** defined by the transversal. As illustrated in Figure 5, two lines are parallel if and only if the alternate interior angles of a transversal are congruent. This includes the Important special case of a right angle, when the transversal is perpendicular to both lines.

Theorem 7.7 (Perpendiculars to Intersecting Lines)**.** *Let m and n be distinct intersecting lines. If line m' is perpendicular to m and line n' is perpendicular to n, then m' and n' are distinct intersecting lines as well.*

Proof. The lines m' and n' cannot be parallel, for if they were, both m and n would be perpendicular transversals of m' and n' and so would be parallel or equal. And m' and n' cannot be the same line for again m and n would be perpendicular to m' and so parallel or equal. □

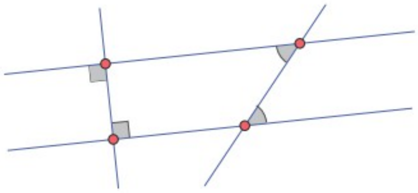

Figure 5. Transversals of Parallel Lines with Alternate Interior Angles

This theorem has a number of applications. It was already needed in Exercise 4 of Chapter 6 to prove that two bisectors of adjacent exterior angles of a triangle always intersect, because they are each perpendicular to an internal angle bisector. Another application is the next theorem, which concerns circumscribing a circle about any triangle.

Theorem 7.8 (Triangle Circumcircle)**.** *For any $\triangle ABC$, there is a unique circle that passes through the points A, B, and C. The center O of the circle is the point of concurrence of the perpendicular bisectors of the sides of the triangle.*

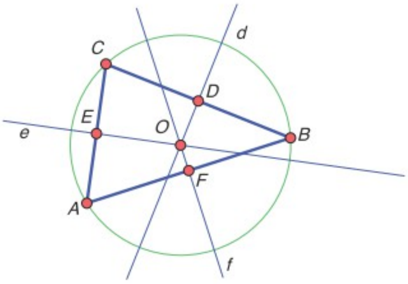

Figure 6. Circumcircle and Circumcenter of $\triangle ABC$

This circle is called the **circumcircle** of the triangle, and the point of concurrence O is called the **circumcenter** of the triangle.

Proof. Consider the bisectors of two of the sides, say the lines e and f that are the perpendicular bisectors of sides AC and AB, as shown in Figure 6. By Theorem 7.7 these two lines intersect at a point O, since they are each perpendicular to a side of the triangle, and these sides are not collinear. Since the perpendicular bisector of a segment is the locus of points equidistant from the endpoints by Theorem 3.10, the points of e are equidistant from A and C and the points of f are equidistant from A and C. Together this says that O satisfies $\|OC\| = \|OA\| = \|OB\|$. If we call this common distance r, the circle with radius r and center O passes through all three vertex points of $\triangle ABC$.

For the final part of the proof, since $\|OC\| = \|OB\|$, the point O is equidistant from B and C and so is on d, the perpendicular bisector of BC. Thus, O is on all three perpendicular bisectors. □

Parallelograms

The study of parallel lines leads to parallelograms.

Note. We will say that two segments or rays are parallel if the lines they belong to are parallel.

Definition 7.9. A quadrilateral $ABCD$ is a **parallelogram** if its opposite sides are parallel, i.e., AB is parallel to CD and DA is parallel to BC.

Theorem 7.10 (Parallelogram Symmetry)**.** *For a parallelogram ABCD, the opposite sides are congruent and the opposite angles are congruent. Also, the two diagonals AC and BD intersect at a point M that is the midpoint of each diagonal. The half-turn* $\mathcal{H}_M$ *is a symmetry of the parallelogram.*

Proof. Let M be the midpoint of the diagonal AC. Then $\mathcal{H}_M(A) = C$ and $\mathcal{H}_M(C) = A$.

The line $\overline{AC}$ is a transversal of both pairs of opposite sides. Therefore, by the Transversal Theorem, Theorem 7.5, $\mathcal{H}_M(\overline{AB}) = \overline{CD}$ and $\mathcal{H}_M(\overline{AD}) = \overline{CB}$.

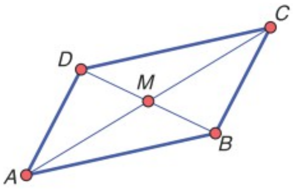

Figure 7. Parallelogram with Diagonals

Point B is the intersection of $\overline{AB}$ and $\overline{CB}$, so its image is the intersection of the image lines $\overline{CD}$ and $\overline{DA}$, namely D. So $\mathcal{H}_M(D) = B$. Therefore, since $\mathcal{H}_M(B) = D$ also, $\mathcal{H}_M$ is a symmetry of the parallelogram.

This means that $\mathcal{H}_M$ maps AB to CD and BC to DA, so the opposite sides are congruent.

The half-turn with center M also maps each angle of $ABCD$ to its opposite angle, so the opposite angles are congruent.

Finally, the midpoint of BD is also M, since $\mathcal{H}_M(B) = D$. ☐

Corollary. *A quadrilateral is a parallelogram if and only if it has point symmetry.*

Proof. The theorem has been proved in one direction. If $ABCD$ has a half-turn $\mathcal{H}_M$ that is a symmetry, then $\mathcal{H}_M(AB) = CD$ and $\mathcal{H}_M(BC) = DA$, so the opposite sides are parallel, by the properties of half-turns. ☐

Earlier, the meaning of "direction" was defined for two rays or two segments on the same line. With these properties of parallels proven, it is possible to extend this definition to parallel rays and segments.

Definition 7.11. Two rays on parallel lines, $\overrightarrow{AB}$ and $\overrightarrow{CD}$, have the **same direction** if both B and D are on the same side of $\overline{AC}$. They have opposite direction if B and D are on opposite sides. The same definition applies for the same direction of segments AB and CD on parallel lines. Two directed lines that are parallel either have directing rays in the the same direction or the opposite direction.

There are several properties of a quadrilateral $ABCD$ that ensure it is a parallelogram.

Theorem 7.12 (Equal Parallels). *If two segments AB and CD are parallel and $||AB|| = ||CD||$, then either*

- *AB and CD have opposite direction and $ABCD$ is a parallelogram with AC and BD as diagonals or*
- *AB and CD have the same direction and $ABDC$ is a parallelogram with AD and BC as diagonals.*

Therefore, if $ABCD$ is a quadrilateral with two opposite sides parallel and congruent, then $ABCD$ is a parallelogram.

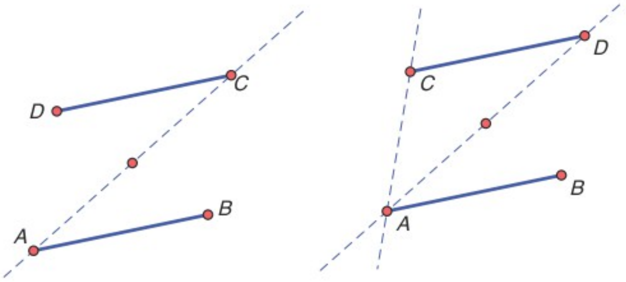

Figure 8. Congruent Parallel Segments

Proof. If B and D are on opposite sides of $\overline{AC}$, then $ABCD$ is a parallelogram. Let M be the midpoint of AC. Then by the Transversal Theorem, Theorem 7.5, $\overline{AD}$ is parallel to $\overline{CB}$. Therefore $ABCD$ is a parallelogram. Since AC and BD are diagonals that intersect at their common midpoint, $ABDC$ is not a quadrilateral.

As illustrated in Figure 8, if B and D are on the same side of $\overline{AC}$, then D is an interior point of $\angle BAC$. First, B and D are in the same half-plane of $\overline{AC}$. Second, D and C are on the same side of $\overline{AB}$ because DC is parallel to AB. Putting these together, D is interior to $\angle BAC$.

The Crossbar Theorem, Theorem 2.10, implies that C and B are on opposite sides of $\overline{AD}$. Then by the same reasoning as before $ABDC$ is a parallelogram and AD and BC are diagonals that intersect at their common midpoint, so $ABCD$ is not a quadrilateral. This implies that the last statement is true. □

A corollary of this theorem is that the midpoints of a parallelogram define more parallelograms and parallel lines.

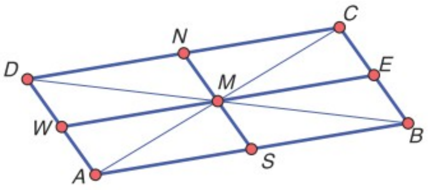

Figure 9. Midpoints and Midsegments in a Parallelogram

Theorem 7.13 (Parallelogram Midsegments)**.** *If $ABCD$ is a parallelogram, a segment EW connecting the midpoints of two opposite sides is parallel to the other two sides and forms a parallelogram with either of those sides. The two such segments intersect at M, the intersection point of the diagonals.*

Proof. Let the midpoints be labeled as in Figure 9. Then E and W are the midpoints of the opposite sides BC and DA. The quadrilateral $BEWA$ has two sides BE and WA, parallel and of the same length, so it is a parallelogram by Theorem 7.12. Therefore,

EW is parallel and congruent to AB. By the same reasoning, NS, the other segment connecting midpoints of opposite sides, is parallel to DA.

Since the symmetry $\mathcal{H}_M$ maps each midpoint of a side of $ABCD$ to the midpoint on the opposite side, M is the midpoint of both segments EW and NS. ☐

By symmetry $ENWS$ is a parallelogram with the segments EW and NS as diagonals.

Here is a second criterion for a quadrilateral to be a parallelogram.

Theorem 7.14 (Congruent Opposite Sides). *A quadrilateral ABCD with opposite sides of equal length is a parallelogram.*

Proof. By assumption $\|AB\| = \|CD\|$ and $\|BC\| = \|DA\|$. The diagonal AC is thus the common side of two triangles $\triangle ABC$ and $\triangle CDA$ that are congruent by SSS. Let T be the rigid motion of the congruence defined by SSS.

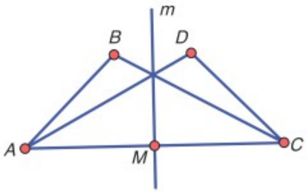

Figure 10. SSS Rigid Motion from $\triangle ABC$ to $\triangle CDA$

This rigid motion T maps AC to CA, so the midpoint M of AC is a fixed point. Therefore, T is either the half-turn $\mathcal{H}_M$ or the reflection $\mathcal{R}_m$, where m is the perpendicular bisector of AC.

If $T = \mathcal{H}_M$, the quadrilateral has a point symmetry and $ABCD$ is a parallelogram. So what remains is to rule out the other case.

Since $T(B) = D$, if $T = \mathcal{R}_m$, then B and D must be on the same side of $\overline{CA}$, since the reflection maps each half-plane of $\overline{CA}$ into itself. It cannot be the case that B is on m, for then $B = D$ and $ABCD$ is not a quadrilateral.

Therefore, B is in the m-half-plane of either A or C. Suppose it is in the half-plane of A, as in Figure 10. The segment CB intersects m at a point P, since B and C are in opposite half-planes. Applying $\mathcal{R}_m$ to CB, the image AD also intersects m at P. But this means that the sides CB and AD intersect at P, so $ABCD$ is not a quadrilateral. ☐

This is an example of how having a rigid motion included in a triangle congruence test like SSS can help in a proof.

Rectangles

Now we define a very familiar polygon that is a special parallelogram.

Definition 7.15. A **rectangle** is a quadrilateral $ABCD$ with each vertex angle a right angle.

Theorem 7.16 (Rectangle Symmetry). *If ABCD is a rectangle, then it is a parallelogram with the lines through the midpoints of opposite sides as lines of symmetry.*

Proof. Lines $\overline{AB}$ and $\overline{CD}$ are perpendicular to $\overline{AD}$, so they must be parallel. Likewise, $\overline{BC}$ and $\overline{AD}$ are parallel. This proves $ABCD$ is a parallelogram. Let the midpoints be labeled as in Figure 11. It was proved in Theorem 7.13 that the midpoint segment EW is parallel to AB and DC. So in a rectangle, EW is perpendicular to the transversals DA and BC. Thus, the reflection $\mathcal{R}_{EW}$ maps AB to DC. The other midpoint line is also a line of symmetry. □

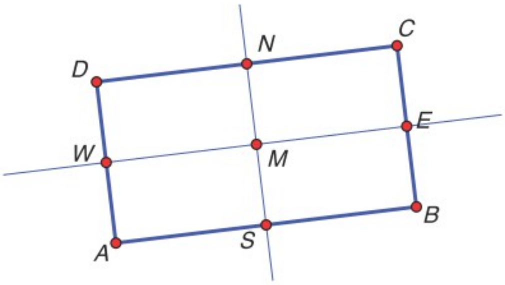

Figure 11. Rectangle with Lines of Symmetry

Note. Because of the line symmetry, the two diagonals of a rectangle are congruent, since a line reflection maps one to the other. The converse of this statement is also true. In addition, if one angle of a parallelogram is a right angle, then all are right angles and the figure is a rectangle. The proofs are problems in Exercise 12.

Every parallelogram, including rectangles, has a point symmetry. The product of the reflections in the two perpendicular midlines of a rectangle is a 180-degree rotation centered at the point of intersection. This is the half-turn $\mathcal{H}_M$ that defines the point symmetry. Thus, in general, parallelograms have two rotational symmetries, the identity and $\mathcal{H}_M$, but rectangles also have, in addition, two line symmetries. This is consistent with the discussion of dihedral symmetry in Chapter 5.

This agrees with a visual sense that rectangles are more symmetric than a general parallelogram. A square is a special rectangle that is yet more symmetric. It is a regular 4-gon and has four rotations and four reflections as symmetries.

Distance Between Parallel Lines. One can use rectangles to define the distance between two parallel lines and also to define the unique parallel line of symmetry of a figure consisting of two parallel lines.

Theorem 7.17 (Midline Theorem). *Let p and q be parallel lines.*

(1) *Every segment PQ perpendicular to p at P and to q at Q has the same length.*

(2) *The perpendicular bisector of every such perpendicular PQ is the same line m.*

(3) *This line m is parallel to p and q and is a line of symmetry for p and q, i.e.,* $\mathcal{R}_m(p) = q$.

(4) *If a transversal intersects p and q at A and B, the midpoint of AB is on m.*

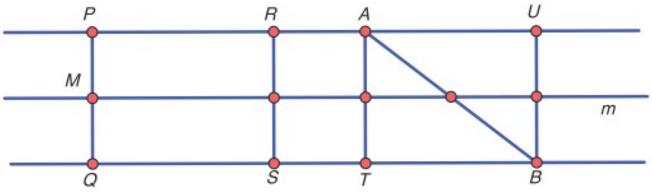

Figure 12. Midline with Transversals

Proof. Most of these statements are immediate consequences of properties of rectangles. If P and R are on p and Q and S are on q, then if PQ and RS are perpendicular to p and q, then $PQSR$ is a rectangle and the opposite sides PQ and RS have the same length.

Also the perpendicular bisector m of either PQ or RS is the unique line of symmetry of the rectangle $PQSR$ parallel to p and q, so it is a well-defined line m with $\mathcal{R}_m(p) = q$.

If AB is a segment with A on p and B on q with midpoint C, then $\mathcal{H}_C(p) = q$ by Theorem 7.5. Therefore, $\mathcal{H}_C(m) = m$ and C is on m. Another way to see this is to construct perpendicular segments AT and UB to form a rectangle as in Figure 12. ☐

Definition 7.18. For two parallel lines p and q, the **distance between** p **and** q is the length of any segment PQ perpendicular to p at P and to q at Q. The **midline of the two parallel lines** is the perpendicular bisector of any such PQ.

This distance and this line are well-defined, according to the theorem.

Consider the symmetries of a pair of parallel lines p and q. By the Transversal Theorem, Theorem 7.5, and the Midline Theorem, every point M on the midline is the center of a point symmetry $\mathcal{H}_M$. Also, the midline is a line of symmetry, the only one parallel to p and q. But every perpendicular transversal c is a line of symmetry, for $\mathcal{R}_c$ will map p to p and q to q (and also the midline to itself as well). It is also true that the products of these symmetries are also symmetries. We will see some of these products later in the chapter.

Midpoint Figures

For any triangle $\triangle ABC$, we encountered in Theorem 4.12 the triangle $\triangle DEF$ formed by the midpoints of the sides. More recently, in the proof of Theorem 7.2, we have seen that the images of $\triangle ABC$ by the half-turns centered at these midpoints create a triangle $\triangle A'B'C'$ that has the original $\triangle ABC$ as its midpoint triangle. All these triangles are shown in Figure 13. By Theorem 7.17, a midpoint line such as $\overline{DE}$ is the midline of the parallel lines $\overline{AB}$ and $\overline{A'B'}$.

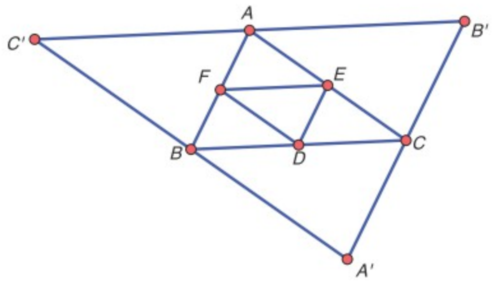

Figure 13. Midpoint Triangles

Such structures of triangles within triangles can be built up to create arbitrarily large triangles tiled with congruent triangles, such as the one in Figure 14 with eight triangles to a side. Within $\triangle ABC$ in this figure, if one takes a subtriangle such as $\triangle ADE$, one can see that the ratio of the length of DE to the length of BC is the same as AD/AB and AE/AC. Also one sees parallel lines and congruent angles such as $\angle ADE \cong \angle ABC$ in these triangles.

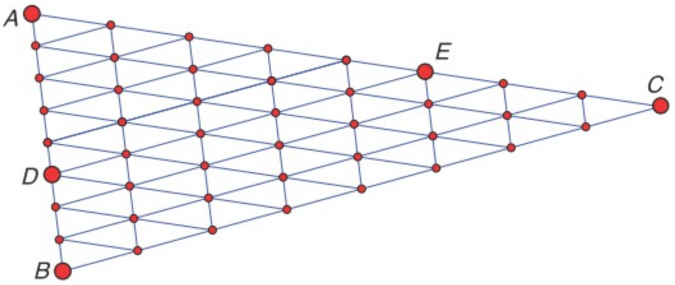

Figure 14. Triangle Tiled with Congruent Triangles

This suggests how, if one assumed the Euclidean Parallel Postulate as an axiom instead of our Axiom 6 , one could prove the Dilation Axiom as a theorem for rational ratios and then extend to real ratios using a limit argument similar to that used in the proof of Theorem 6.8 about the intersection of a line with a circle.

There are many other interesting features in this figure: parallelograms of different shapes and rigid motions that are products of half-turns. These rigid motions will be taken up in a few pages from now.

Midpoint Quadrilaterals. A quadrilateral $ABCD$ also has a midpoint figure $JKLM$. Unlike the case of triangles, this figure is not similar to the original quadrilateral. Instead, it is always a parallelogram.

Theorem 7.19 (Midpoint Parallelogram)**.** *For any quadrilateral ABCD, the midpoints of the sides are the vertices of a parallelogram.*

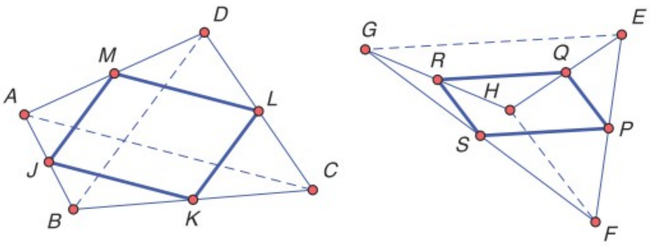

Figure 15. Midpoint Quadrilaterals

Proof. Assume that B and D are on opposite sides of $\overline{AC}$. Using the labels in Figure 15 for the midpoints, the segments JK and ML are segments parallel to and in the same direction as the diagonal AC. The lengths of these segments are both equal to $\|AC\|/2$.

Therefore, these segments are the opposite sides of a parallelogram if they are not collinear, but they cannot be collinear if B and D are on opposite sides of the diagonal. Thus, it only remains to check that for at least one of the diagonals, the other vertices are on opposite sides.

Given quadrilateral $ABCD$, if B and D are one the same side of $\overline{AC}$, then since the sides of the quadrilateral do not intersect, either D is in the interior of $\triangle ABC$ or B is in the interior of $\triangle ADC$. (This is to be proved in Exercise 7.) In either case, A and C are on opposite sides of $\overline{BD}$ so we should choose BD as the diagonal in the proof. $\square$

For some cases, the midpoint quadrilateral is a special kind of parallelogram. There is a question about this in the exercises. Also, in the exercises is a method for proving the concurrence of medians of a triangle using midpoint quadrilaterals.

Circumcircles of Right Triangles. Right triangles provide an interesting special case of midpoint triangles, as illustrated in Figure 16. Two of the midlines, $\overline{DF}$ and $\overline{EF}$, are perpendicular bisectors of legs. The parallelogram $CEFD$ is a rectangle.

From Theorem 7.8 we know the F, the intersection of these perpendicular bisectors, is the circumcenter of the triangle and is equidistant from the three vertices. This means that the segment AB is a diameter of the circumcircle.

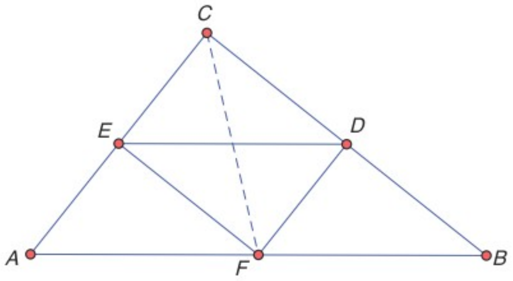

Figure 16. Midpoint Right Triangle

This leads to an important relationship between circles and right angles.

Theorem 7.20 (Right Angle Locus). *If $\triangle ABC$ has a right angle at C, then the midpoint F of AB is equidistant from all three vertices of the triangle. Therefore, C is a point on the circle with diameter AB.*

Conversely, if a segment AB is the diameter of a circle c, then for any point C on c distinct from A and B, $\angle ACB$ is a right angle.

In addition, $m\angle BFC = 2m\angle BAC$, where F is the midpoint of AB and the center of the circle.

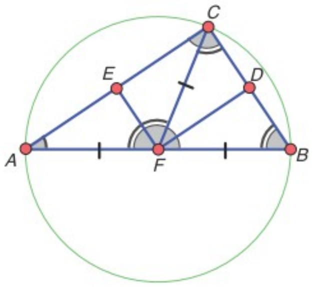

Figure 17. Congruent Angles in an Inscribed Right Triangle

Proof. We continue to use the same names for the points as in Figures 16 and 17.

The first statement was already proved in the comments above, using properties of rectangles.

To prove the second part of the theorem, take a triangle $\triangle ABC$ inscribed in a circle with diameter AB. This triangle is divided by CF into two isosceles triangles, which in turn can be divided into two right triangles. In Figure 17, congruent angles are marked.

Since the triangles $\triangle AFC$ and $\triangle CFB$ are isosceles, in each triangle the base angles are congruent. Let $a = m\angle FAC = m\angle FCA$ and $b = m\angle FBC = m\angle FCB$ be the measures of these base angles.

In $\triangle ABC$, the measure of the angle at C is the sum of two base angles, so $m\angle ACB = a + b$. But $m\angle BAC = a$ and $m\angle ABC = b$, so the sum of the angle measures of $\triangle ABC$ is $a + b + (a + b) = 2a + 2b = 180$, so that $90 = a + b = m\angle ACB$. Therefore, the angle at C is a right angle.

To prove the final equation of angles, $\angle CAB$ has measure a. The base angles in $\triangle BFC$ have measure b, so the central angle $\angle BFC$ has measure $180 - 2b = 2a$. □

Generalizing Parallelograms

One might ask if there is an "SSSS" test for parallelograms analogous to the SSS test for triangles. As many a do-it-yourselfer can attest, the answer is no. If one starts a bookcase by nailing four boards together to form a rectangle, it will be prone to collapse, for there are an infinite number of parallelograms with the same pair of side lengths. One needs some kind of cross-brace, such as a diagonal, to hold the shape.

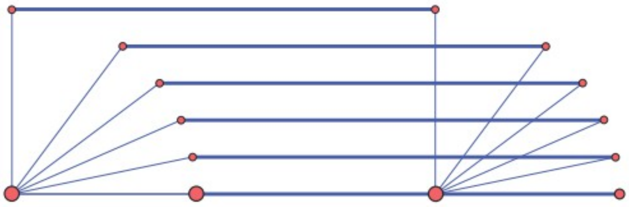

Figure 18. Collapsing Parallelograms with the Same Side Lengths

This can also be seen with delivery boxes. If one breaks out the bottom, the four sides form a parallelogram that collapses flat.

When we study translations in the next section, we will need such flattened parallelograms as well as ordinary parallelograms. So we extend our parallelogram definition with this additional one.

Definition 7.21. A **linear parallelogram** *ABCD* consists of four collinear points A, B, C, D with the midpoint M of AC being also the midpoint of BD. In other words, the figure has a point symmetry, with M as the center.

Theorem 7.22 (Linear Parallelogram)**.** *Let four collinear points A, B, C, D correspond to real numbers a, b, c, d, m by a ruler. Then $ABCD$ is a linear parallelogram if $a+c = b+d$, or equivalently if AB and DC have the same direction and length, meaning $b-a = c-d$. In that case, AD and BC also have the same direction and length, since $d - a = c - b$.*

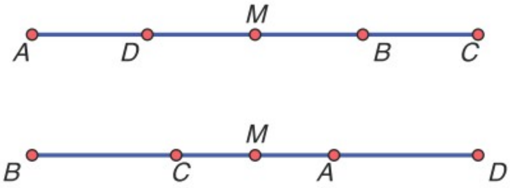

Figure 19. Two Linear Parallelograms *ABCD*

Proof. Working on the real line m using a ruler, the midpoint of AC corresponds to $(a + c)/2$ and the midpoint of BD corresponds to $(b + d)/2$. If these midpoints are the same point, then $a + c = b + d$, which implies the other equations. □

Such linear parallelograms can arise as limiting cases of well-known figures. For example, for a self-intersecting four-sided figure such at the ones in Figure 20, the midpoint figure may be a parallelogram or it may be a linear parallelogram. Some hints for a proof are in Exercise 8.

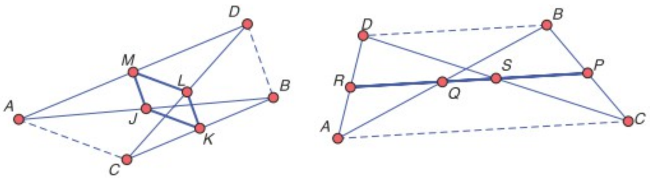

Figure 20. Midpoint Quadrilaterals for More General *ABCD*

The next theorem combines a uniqueness property of parallelograms with a test for congruence.

Theorem 7.23 (Fourth Parallelogram Point)**.** *Given three noncollinear points A, B, C, there is a unique point D so that $ABCD$ is a parallelogram. If A, B, C are collinear, there is a unique point D so that $ABCD$ is a linear parallelogram.*

Proof. Let M be the midpoint of AC and construct $D = \mathcal{H}_M(B)$. In the noncollinear case, this is a parallelogram since $\mathcal{H}_M$ is a symmetry. It is unique because for any parallelogram $ABCD$, it is true that $D = \mathcal{H}_M(B)$. The same reasoning applies to the collinear case. In this case, using a ruler and real numbers, the formula for $\mathcal{H}_M(B)$ is $2(a + c)/2 - b = a + c - b$. ☐

Corollary. *Two parallelograms $ABCD$ and $EFGH$ are congruent if two pairs of adjacent sides and one pair of included angles are congruent; e.g., $AB \cong EF$, $BC \cong FG$, and $\angle ABC \cong \angle EFG$. This applies to linear parallelograms as well.*

Proof. By SAS for triangles, there is a unique rigid motion T that maps $\triangle ABC$ to $\triangle EFG$. Let D be the unique point so that $ABCD$ is a parallelogram. Then the T-image of $ABCD$ is a parallelogram $EFGT(D)$. But H is the unique point so that $EFGH$ is a parallelogram, so $T(D) = H$. For linear parallelograms, the angle will be either a zero angle or a straight angle. ☐

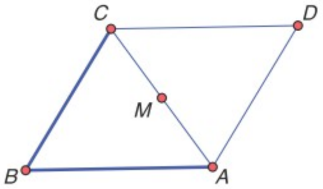

Figure 21. Constructing a Parallelogram from Two Sides and an Angle

Translations as Half-turn Products

In the study of parallels, we have used half-turns extensively to map one parallel segment to another. This was very efficient and effective for these purposes, but now we will use half-turns to expand our repertoire by defining the last well-known rigid motion, the translation.

Definition 7.24. A **translation** is a rigid motion that is the product of two half-turns.

An example of such a translation can be found in the parallelogram with midpoints in Figure 9. The rigid motion $T = \mathcal{H}_M\mathcal{H}_S$ maps AB to DC, since $\mathcal{H}_M\mathcal{H}_S(A) = \mathcal{H}_M(B) = D$ and $\mathcal{H}_M\mathcal{H}_S(B) = \mathcal{H}_M(A) = C$. In addition the rigid motions $\mathcal{H}_C\mathcal{H}_E$ and $\mathcal{H}_N\mathcal{H}_M$ also map AB to DC. We will see that these are all the same translation.

To get a general idea of how such a product of half-turns will behave, let A and B be distinct points and let P be any point. Then let $P' = \mathcal{H}_A(P)$ and $P'' = \mathcal{H}_B(P')$.

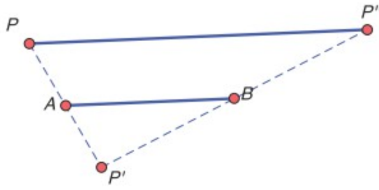

Figure 22. Translation of a Point as the Product of Half-turns

It is clear in Figure 22 that in $\triangle PP'P''$, the segment AB is a midsegment parallel to PP'', in the same direction and half its length. This shows that translations have a property not shared by rotations and reflections: the distance from any point to its image is always the same.

Before proceeding further, there is one observation about this translation that is already apparent. The line $\overleftrightarrow{AB}$ and all lines parallel to it are **invariant lines of the translation** $\mathcal{H}_B\mathcal{H}_A$. These are not lines of fixed points, but each point of such a line is mapped to another point of the same line. As we will see below, any other line is mapped into a line distinct from it but parallel.

Now consider the relationship between the images of P and those of another point Q. For now, assume a general case when no three of the points A, B, P, Q are collinear and PQ is not parallel to AB. These special cases will be addressed afterwards.

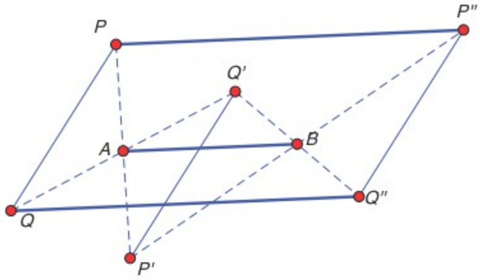

Figure 23. Translations of Two Points as the Product of Half-turns

Again, we observe that the segments PP'' and QQ'' are parallel and have the same length. Also, since half-turns map segments to parallel segments, the segments PQ, $P'Q'$, and $P''Q''$ are parallel, so $PP''Q''Q$ is a parallelogram.

It seems reasonable to expect that the special cases also result in a parallelogram or a linear parallelogram, and this is so. It can be checked using our definition of a translation as a product of half-turns, but instead we can use an important alternative definition of a translation as a product of line reflections.

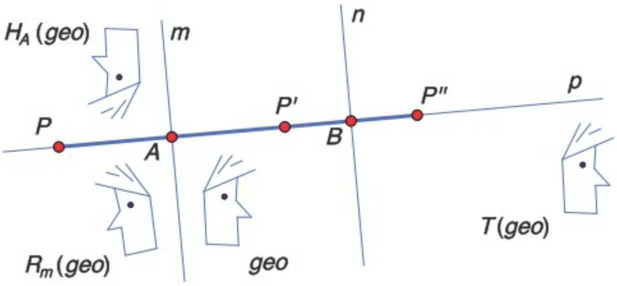

Figure 24. Translations of a Figure Named "geo" in Two Ways

Theorem 7.25 (Translations as Double Line Reflections). *For distinct points A and B, the translation $\mathcal{H}_B\mathcal{H}_A = \mathcal{R}_b\mathcal{R}_a$, where a and b are the lines perpendicular to $\overline{AB}$ at points A and B.*

Proof. The half-turn at A is the product of reflections in any two perpendicular lines at A, so one choice is $\mathcal{H}_A = \mathcal{R}_{AB}\mathcal{R}_a$. And at B, $\mathcal{H}_B = \mathcal{R}_b\mathcal{R}_{AB}$. Then $\mathcal{H}_B\mathcal{H}_A = \mathcal{R}_b\mathcal{R}_{AB}\mathcal{R}_{AB}\mathcal{R}_a = \mathcal{R}_b\mathcal{R}_a$. ☐

This gives a new perspective on the behavior of translations and a new tool. For any pair of parallel lines, the product $\mathcal{R}_b\mathcal{R}_a$ is a translation whose invariant lines are the perpendicular transversals of the parallel lines a and b.

Figure 24 shows the translation of a figure both by a product of half-turns and by the corresponding line reflections. They both terminate at the image of T, but the intermediate steps are quite different.

In addition, in this figure there are the same points P, P', and P'' as before, but this time on the line $\overline{AB}$. One can see how the lengths add up to $\|PP''\| = 2\|AB\|$.

With this definition it is not difficult to take care of the missing pieces.

Theorem 7.26 (Translation of a Segment). *If T is a translation distinct from I, then for any points P and Q, PQ and T(P)T(Q) are opposite sides of a parallelogram or linear parallelogram.*

Proof. Let $T = \mathcal{H}_B\mathcal{H}_A$. When needed, T can also be written as $T = \mathcal{H}_D\mathcal{H}_C$, where $ABDC$ is a rectangle, since $T = \mathcal{R}_{BD}\mathcal{R}_{AC}$.

We continue to use the notation in Figure 23: $P'' = T(P)$ and $Q'' = T(Q)$.

First, assume that PQ is not parallel to or collinear with AB. If neither P nor Q is on $\overline{AB}$, then from the discussion of Figure 23, it follows that $P''Q''$ is parallel to PQ, so $PQQ''P''$ is a parallelogram. In case either P or Q is on $\overline{AB}$, the proof can be carried out by a computation with rulers, but a simpler method is to use $T = \mathcal{H}_D\mathcal{H}_C$, where neither P nor Q is on $\overline{CD}$ and $ABDC$ is a rectangle.

In the case when PQ is parallel to or collinear with AB, let a and b be lines perpendicular to AB so that $T = \mathcal{R}_b\mathcal{R}_a$. Then each line reflection reverses direction on $\overline{PQ}$ and the segment $P''Q''$ is collinear with PQ and has the same direction and length. Therefore, by Theorem 7.22, $PQQ''P''$ is a linear parallelogram. □

This theorem has a very powerful corollary.

Corollary (One Point Determines Translation). *If T is a translation and if E and $F = T(E)$ are given, then T is completely determined. For any other point G, $T(G)$ is the unique point such that $GEFT(G)$ is a parallelogram or linear parallelogram.*

Notation. The unique translation that maps A to B will be denoted $\mathcal{T}_{AB}$. The AB in this notation is far from uniquely determined since for any point P and its translation image $Q = T(P)$, the translation is $\mathcal{T}_{PQ}$.

Products of Translations

This section is devoted to two properties of the products of translations: closure and commutativity.

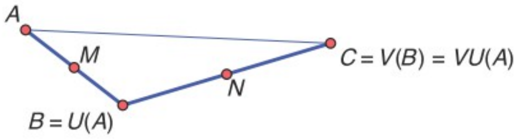

Figure 25. Translation UV

Theorem 7.27 (Closure). *The product of two translations is a translation. The product of a translation and a half-turn is a half-turn.*

In other words, the set of all translations of the plane is closed under composition. Also, the set of all half-turns and translations together is also closed under composition. Each of these sets has the group property. This fact will play a role in the symmetric patterns studied in Chapter 10.

Proof. Let U and V be translations. The task is to prove VU is also a translation.

Choose any point A. Then $U = \mathcal{T}_{AB}$ for $B = U(A)$ and $V = \mathcal{T}_{BC}$ for $C = V(B)$.

Let M be the midpoint of AB and let N be the midpoint of BC. Then $U = \mathcal{H}_B\mathcal{H}_M$ since $\mathcal{H}_B\mathcal{H}_M$ is a translation with $\mathcal{H}_B\mathcal{H}_M(A) = \mathcal{H}_B(B) = B$, so $\mathcal{H}_B\mathcal{H}_M = \mathcal{T}_{AB}$.

By similar reasoning $V = \mathcal{H}_N\mathcal{H}_B$, since $\mathcal{H}_N\mathcal{H}_B(B) = C$.

Multiplying these translations, $VU = \mathcal{H}_N\mathcal{H}_B\mathcal{H}_B\mathcal{H}_M = \mathcal{H}_N\mathcal{H}_M$. This is a translation that maps A to C, so $UV = \mathcal{T}_{AC}$.

To prove that the product of a translation and a half-turn is a translation, one uses the same ideas. Let the half-turn be $\mathcal{H}_B$ and the translation be $\mathcal{T}_{AB}$. Then continuing with the notation above, $\mathcal{H}_B\mathcal{T}_{AB} = \mathcal{H}_B\mathcal{H}_B\mathcal{H}_M = \mathcal{H}_M$. Likewise, $\mathcal{T}_{BC}\mathcal{H}_B = \mathcal{H}_N\mathcal{H}_B\mathcal{H}_B = \mathcal{H}_N$. □

The second theorem is a product property quite different from that for line reflections and most other products of rigid motions.

Theorem 7.28 (Commutativity). *Products of translations are commutative. For any translations S and T, $ST = TS$.*

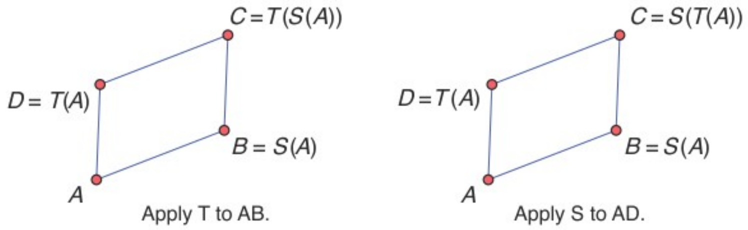

Figure 26. Commutative S and T

Proof. If either S or T is the identity, the theorem is immediate. For the rest of the proof, we assume neither translation is the identity. This proof works for linear parallelograms also. Take any point A and let $B = S(A)$ and $D = T(A)$.

Apply T to AB to get image $DT(B)$. Since $DT(B)$ is the image of AB by translation T, $DABT(B)$ is a parallelogram (or linear parallelogram), by Theorem 7.26. This is illustrated on the left side of Figure 26.

Now compare with a second construction of a parallelogram. Beginning with A and D, apply S to AD with image $BS(D)$. Then we have $DABS(D)$ is a parallelogram or linear parallelogram. This is illustrated on the right side of Figure 26.

Now we have two parallelograms with three points in common. So the fourth points $T(B)$ and $S(D)$ must be the same point C by Theorem 7.23. Therefore, $C = T(B) = TS(A)$ and $C = T(D) = ST(A)$. So $ST = \mathcal{T}_{AC} = TS$. □

Direction from Translation

Now that the theory of translations is well developed, one can use translations to rephrase or reframe some earlier ideas in a simpler and more flexible form.

Previously, the concept of same direction for two parallel rays $\overrightarrow{AB}$ and $\overrightarrow{CD}$ was defined in Definition 7.11 as the rays being on the same side of $\overline{AC}$. For collinear rays, the definition on page 14 was that one ray was contained in the other. In either case, for endpoint C there is a unique ray $\overrightarrow{CD}$ in the same direction as $\overrightarrow{AB}$. These definitions can be combined into one case using translations. With this definition, some equivalence properties of the direction relation for rays become clear.

Theorem 7.29 (Direction by Translation). *The rays $\overrightarrow{AB}$ and $\overrightarrow{CD}$ have the same direction if and only if $\mathcal{T}_{AC}$ maps $\overrightarrow{AB}$ to $\overrightarrow{CD}$. They have opposite direction if the half-turn $\mathcal{H}_M$ maps $\overrightarrow{AB}$ to $\overrightarrow{CD}$, where M is the midpoint of AC.*

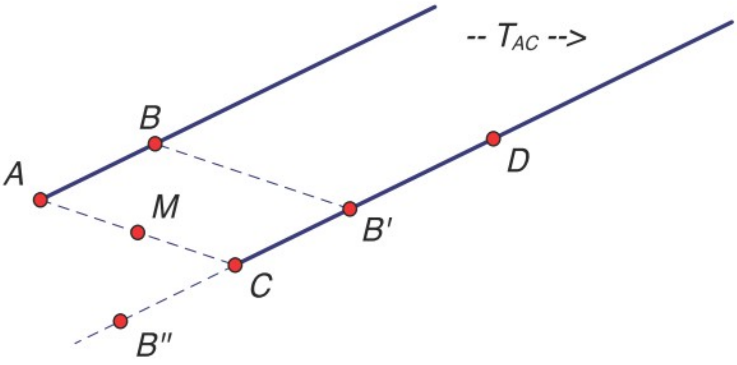

Figure 27. Rays with the Same Direction Related by Translation

Proof. Let M be the midpoint of AC. Then $\mathcal{T}_{AC} = \mathcal{H}_C\mathcal{H}_M$. Let $B' = \mathcal{T}_{AC}(B)$.

Consider when A, B, C are not collinear. The image $\mathcal{T}_{AC}(\overrightarrow{AB})$ is a ray with endpoint C parallel to $\overrightarrow{AB}$. Since $ACB'B$ is a parallelogram, B and B' are on the same side of $\overline{AC}$. Therefore, $\overrightarrow{CB'}$ is the unique ray with endpoint C in the same direction as $\overrightarrow{AB}$.

The ray opposite $\overrightarrow{CB'}$ is $\overrightarrow{CB''} = \mathcal{H}_C(\overrightarrow{CB'}) = \mathcal{H}_M(\overrightarrow{AB})$, where $B'' = \mathcal{H}_M(B)$.

The collinear case follows from a ruler computation since a translation adds a constant to each number and a half-turn reverses direction. □

Recall that the description of direction for rays carries over to directed segments and lines as well. Two segments AB and CD have the same direction if the rays $\overrightarrow{AB}$ and $\overrightarrow{CD}$ have the same direction.

The properties of translation make it straightforward to prove some informally evident properties of direction in this corollary.

Corollary. *These properties of direction follow from the theorem.*

(1) *If $\overrightarrow{CD}$ has the same direction as $\overrightarrow{AB}$, then $\overrightarrow{AB}$ has the same direction as $\overrightarrow{CD}$.*

(2) *If $\overrightarrow{CD}$ has the same direction as $\overrightarrow{AB}$ and $\overrightarrow{EF}$ has the same direction as $\overrightarrow{CD}$, then $\overrightarrow{EF}$ has the same direction as $\overrightarrow{AB}$.*

Here is a proof of the second item in the corollary. The assumption is that $\overrightarrow{CD} = \mathcal{T}_{AC}(\overrightarrow{AB})$ and $\overrightarrow{EF} = \mathcal{T}_{CE}(\overrightarrow{CD})$. Then since $\mathcal{T}_{CE}\mathcal{T}_{AC} = \mathcal{T}_{AE}$, $\overrightarrow{EF} = \mathcal{T}_{CE}\mathcal{T}_{AC}(\overrightarrow{AB}) = \mathcal{T}_{AE}(\overrightarrow{AB})$. The rest is left to the reader.

More Parallelograms. This next proposition unifies with translation language some earlier statements about parallel segments proved in Theorems 7.12 and 7.22.

Theorem 7.30 (Direction and Congruent Segments)**.** *Let* AB *and* CD *be congruent segments. If they have the same direction, then* $\mathcal{T}_{AC}$ *maps* AB *to* CD *and* $ABDC$ *is a parallelogram or linear parallelogram. If they have opposite direction, then* $\mathcal{H}_M$ *maps* AB *to* CD*, where* M *is the midpoint of* AC*. In this case* $ABCD$ *is a parallelogram or linear parallelogram.*

As an application of this theorem, we prove a theorem about an adjoining pair of parallelograms.

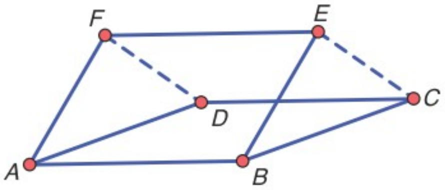

Figure 28. Two Linked Parallelograms Imply a Third

Theorem 7.31 (Third Parallelogram)**.** *Given figures* $ABCD$ *and* $ABEF$*, each a parallelogram or linear parallelogram, the figure* $DCEF$ *is a parallelogram if the lines* $\overline{CD}$ *and* $\overline{EF}$ *are distinct and it is a linear parallelogram if they are equal.*

Proof. Since $ABCD$ is a parallelogram or linear parallelogram, translation $\mathcal{T}_{AB}$ maps D to C. By the same reasoning for $ABEF$, $\mathcal{T}_{AB}$ maps F to E. By Theorem 7.30, this means that $CDFE$ is a parallelogram or a linear parallelogram. ☐

In addition to parallelograms formed by segments and their images, the following theorem is also true.

Theorem 7.32 (Four Half-turn Centers)**.** *The translation* $\mathcal{H}_F\mathcal{H}_E = \mathcal{H}_H\mathcal{H}_G$ *if and only if* $EFHG$ *is a parallelogram or linear parallelogram.*

Proof. From Theorem 7.25, it follows that EF and GH are either parallel or collinear. The hypothesis is equivalent to $\mathcal{H}_G\mathcal{H}_H\mathcal{H}_F\mathcal{H}_E = I$ and this is equivalent to $\mathcal{H}_H\mathcal{H}_F = \mathcal{H}_G\mathcal{H}_E$. Therefore, if the points are not collinear, then FH is parallel to EG, so $EFHG$ is a parallelogram. The linear case follows from a coordinate computation. ☐

Direction and Rotation from Polar Angle

In Chapter 2 protractors and polar angles at a point were introduced. Later, in Definition 5.6, the choice of a protractor at a point A determined signed angle measure $m^+\angle BAC$ as a difference of polar angles. But there was no way to relate polar angles between any two points of the plane. But now, using translations, these angles and measures can be extended uniformly to the whole plane.

Definition 7.33. A **global protractor** δ_* on the plane is a set of protractors δ_P at each point P such that for any real t and any points A and C, $\delta_C(t) = \mathcal{T}_{AC}(\delta_A(t))$.

Given such a global protractor, let $\theta(\overrightarrow{AB})$ denote the polar angle of $\overrightarrow{AB}$ corresponding to δ_A, the protractor at A. For any point C, if $\overrightarrow{CD} = \mathcal{T}_{AC}(\overrightarrow{AB})$, then $\theta(\overrightarrow{CD}) = \theta(\overrightarrow{AB})$ because $\delta_C(\theta(\overrightarrow{AB})) = \mathcal{T}_{AC}\delta_A(\theta(\overrightarrow{AB})) = \mathcal{T}_{AC}(\overrightarrow{AB}) = \overrightarrow{CD}$. Conversely, if these polar angles are equal, by the same relationship $\overrightarrow{CD} = \mathcal{T}_{AC}(\overrightarrow{AB})$.

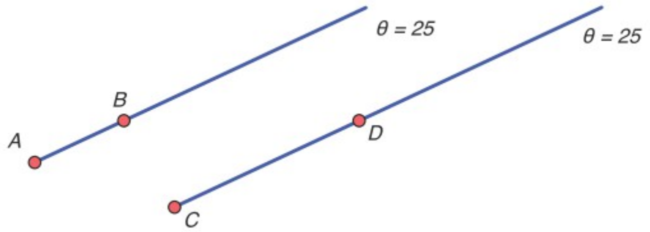

Figure 29. Rays with the Same Direction and Polar Angle

This observation proves the following theorem.

Theorem 7.34 (Direction and Global Polar Angle). *Given a global protractor δ_*, two rays $\overrightarrow{AB}$ and $\overrightarrow{CD}$ have the same direction if and only if they have the same polar angles $\theta(\overrightarrow{AB})$ and $\theta(\overrightarrow{CD})$.*

This implies that two lines $\overline{AB}$ and $\overline{CD}$ are parallel if and only if $\theta(\overrightarrow{AB})$ and $\theta(\overrightarrow{CD})$ are equal or differ by 180 degrees, depending on whether the two rays have the same or opposite direction.

This relationship between polar angles and translation can be used to construct a global protractor from a protractor at a single point.

Theorem 7.35 (Existence of Global Protractors). *For a protractor δ_A at any point A, there is a global protractor δ_* defined by $\delta_C(t) = \mathcal{T}_{AC}(\delta_A(t))$ for all points C.*

Proof. The theorem defines a function δ_C for every C. It must be shown that this function is in fact a protractor at C. For any rays $\overrightarrow{CP}$ and $\overrightarrow{CQ}$, choose values of θ so that $|\theta(\overrightarrow{CP}) - \theta(\overrightarrow{CQ})| \leq 180$. For δ_C to be a protractor, this absolute value must equal $m\angle PCQ$.

Let $\overrightarrow{AP'}$ and $\overrightarrow{AQ'}$ be the images of these rays by translation $\mathcal{T}_{CA}$. The polar angles $\theta(\overrightarrow{CP})$ and $\theta(\overrightarrow{CQ})$ are the same as $\theta(\overrightarrow{AP'})$ and $\theta(\overrightarrow{AQ'})$. Since δ_A is known to be a protractor, $m\angle P'AQ' = |\theta(\overrightarrow{AP'}) - \theta(\overrightarrow{AQ'})| = |\theta(\overrightarrow{CP}) - \theta(\overrightarrow{CQ})|$. Because translations are rigid motions, $m\angle P'AQ' = m\angle PCQ$. Thus δ_C has the properties of a protractor.

What remains to be proved is that any two points satisfy the global protractor relation. If E is another point, then $\delta_E(t) = \mathcal{T}_{CE}(\delta_C(t))$ for every t. By the definition in the theorem statement, $\delta_C(t) = \mathcal{T}_{AC}(\delta_A(t))$, so $\delta_A(t) = \mathcal{T}_{CA}(\delta_C(t))$, since $\mathcal{T}_{CA} = \mathcal{T}_{AC}^{-1}$. Then $\delta_E(t) = \mathcal{T}_{AE}(\delta_A(t)) = \mathcal{T}_{AE}\mathcal{T}_{CA}(\delta_C(t)) = \mathcal{T}_{CE}(\delta_C(t))$, since $\mathcal{T}_{AE}\mathcal{T}_{CA} = \mathcal{T}_{CE}$. ☐

By choosing a global protractor on the plane, one can simplify relationships. For example, in Chapter 5, to define an orientation on the plane without translations, one had to compare protractors along various line segments. With a global protractor, the positive (counterclockwise) rotation direction is determined at every point by the protractor at the point.

Another example, concerning rotations of rays, is given in the next paragraphs.

Direction and Rotation. With a protractor one can measure both unsigned and signed angles at a point P. Assume P_k is a rotation with rotation angle k, with k chosen so that $|k| \leq 180$. We know from Definitions 5.5 and 5.6 that for any ray $\overrightarrow{PQ}$, if $Q' = P_k(Q)$, then $m\angle QPQ' = |k|$ and $m^+\angle QPQ' = k$.

With a global protractor, the same result is true for rays with endpoints different from P.

Theorem 7.36 (Polar Angle Rotation)**.** *Given a global protractor δ_* let a rotation P_k map a ray $\overrightarrow{AB}$ to $\overrightarrow{A'B'}$. Then $\theta(\overrightarrow{A'B'}) = \theta(\overrightarrow{AB}) + k$. If $0 < |k| < 180$, the lines $\overline{AB}$ and $\overline{A'B'}$ intersect at a point C forming an angle of measure $|k|$. More specifically, the lines form an angle $\angle DCE$, where $\overrightarrow{CD} = \mathcal{T}_{AC}(\overrightarrow{AB})$ and $\overrightarrow{CE} = \mathcal{T}_{A'C}(\overrightarrow{A'B'})$, with $m\angle DCE = |k|$ and $m^+\angle DCE = k$.*

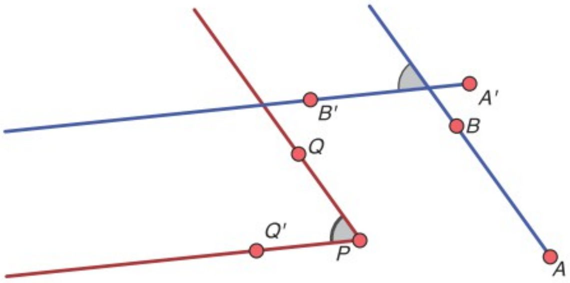

Figure 30. Polar Angle Change by a Ray Rotated about Center P

Proof. The ray $\overrightarrow{AB}$ has the same direction as a ray $\overrightarrow{PQ}$. The rotation maps these rays to a pair of rays $\overrightarrow{A'B'}$ and $\overrightarrow{PQ'}$ that also have the same direction, as illustrated in Figure 30. Therefore, $\theta(\overrightarrow{A'B'}) = \theta(\overrightarrow{PQ'}) = \theta(\overrightarrow{PQ}) + k = \theta(\overrightarrow{AB}) + k$.

If $0 < |k| < 180$, by Theorem 7.34, the lines $\overline{AB}$ and $\overline{A'B'}$ are not parallel so they intersect at a point C.

The ray $\overrightarrow{CD} = \mathcal{T}_{AC}\mathcal{T}_{PA}(\overrightarrow{PQ}) = \mathcal{T}_{PC}(\overrightarrow{PQ})$ and $\overrightarrow{CE} = \mathcal{T}_{AC}(\overrightarrow{PQ'})$, so the translation $\mathcal{T}_{PC}$ takes $\angle QPQ'$ to $\angle DCE$ so the angles have the same measure and signed measure. ☐

This theorem allows one to deduce the nature of products of even (orientation-preserving) rigid motions. An even rigid motion is a rotation with rotation angle k or else a translation, which can be assigned rotation number 0 since translations do not change polar angles. Let T_1 and T_2 be even rigid motions with rotation angles k_1 and

k_2. The product T_2T_1 changes the polar angle of a ray by adding $k = k_1 + k_2$. If $k = 0$ (mod 360), the product is a translation. Otherwise, it is a rotation with rotation angle k but center unknown from this information. Such products will be studied, and the centers of rotation located, in Theorem 10.2.

Symmetries of a Pair of Parallel Lines. Some symmetries of a pair of parallel lines p and q were listed after the Midline Theorem, Theorem 7.17. One line of symmetry was the midline m. The others were reflections in perpendicular transversals and also half-turns with centers on m. It was also mentioned that products of these symmetries are symmetries. Now we know that products of reflections in two perpendicular transversals are translations in a direction parallel to p and q. This is also true of the products of half-turns centered on m. So any translation with direction parallel to these lines is a symmetry.

There is one other reflection product that has not been considered yet. This is the **glide reflection**. A glide reflection is defined to be the product of a line reflection and a translation parallel to the line of reflection. In the case of parallel lines, every glide reflection symmetry is the product of a translation symmetry and $\mathcal{R}_m$. Such a product will translate the lines and then interchange p and q. It is also true that a product of a half-turn symmetry and one of the perpendicular line reflections is also a glide reflection, a different way of writing the same symmetries. Glide reflections will be discussed in detail in Chapter 10.

All these rigid motions are also the symmetries for a single line p, provided one takes $m = p$.

Vectors

When a translation can be written two different ways, as $\mathcal{T}_{AB} = \mathcal{T}_{CD}$, this means that AB and CD are congruent segments with the same direction. This is another case when a relationship is an equivalence relation, and this one is extremely useful in math and science and engineering.

Definition 7.37. Two segments AB and CD represent the same **vector** if $\mathcal{T}_{AB} = \mathcal{T}_{CD}$. The notation for this relation will be $[AB] = [CD]$.

Our use of $[AB]$ to denote a vector is not standard notation. The vector $[AB]$ is usually denoted by an arrow on top, as $\overrightarrow{AB}$, but we are using this notation for rays.

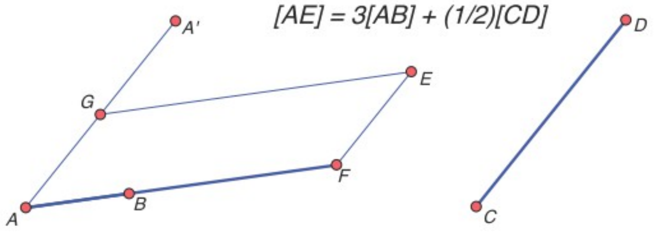

Figure 31. Vector Addition and Multiplication by Real Numbers

It is easy to see from this definition that the vector relation has these equivalence properties.

- For all vectors, $[AB] = [AB]$.
- If $[AB] = [CD]$, then $[CD] = [AB]$.
- If $[AB] = [CD]$ and $[CD] = [EF]$, then $[AB] = [EF]$.

For the last item, given $CD = \mathcal{T}_{AC}(AB)$ and $EF = \mathcal{T}_{CE}(CD)$, then $EF = \mathcal{T}_{CE}\mathcal{T}_{AC}(AB) = \mathcal{T}_{AE}(AB)$.

The vector, the set of equivalent segments, is viewed as an object. Geometrically, the vector can be represented as an arrow with tail at A and head at B. But it is also represented by an infinite number of other arrows, each one with the same length and the same direction. All such arrows are translations of each other.

It is not yet clear just from this definition what value is gained by this concept, especially since the set of vectors is in one-to-one correspondence with the set of translations. We will see vectors in action in later chapters.

There are some operations with vectors that we can define now. The first is addition. If $\mathcal{T}_{CD}\mathcal{T}_{AB} = \mathcal{T}_{EF}$, then $[AB] + [CD] = [EF]$. In particular, since $\mathcal{T}_{BC}\mathcal{T}_{AB} = \mathcal{T}_{AC}$, then $[AB] + [BC] = [AC]$, as pictured in Figure 25. Also, in Figure 26, $[AB] + [AD] = [AC]$. This shows what is called the parallelogram law of addition.

The vector $[AA]$ acts as the zero for this addition; it can just be written as 0. Reversing the segment order is the additive inverse: $[AB] + [BA] = [AA] = 0$. Vector addition is commutative by Theorem 7.28.

The other important vector operation is the multiplication of a vector by a real number. If $r > 0$, the vector $r[AB]$ is the vector CD, where AB and CD have the same direction, but $\|CD\| = r\|AB\|$. If $r < 0$, the direction is reversed and $r[AB] = |r|[BA]$. Also, $0[AB] = 0$.

These operations are illustrated in Figure 31. The vector $[AA'] = [CD]$, $[AG] = (1/2)[CD]$, $[AF] = 3[AB]$, and $[AE] = 3[AB] + (1/2)[CD]$.

Exercises and Explorations

1. (Regular Polygon Angle). Derive the formula for the angles of a regular n-gon.
2. (Quadrilateral with Lines of Symmetry). Show that a quadrilateral with two perpendicular lines of symmetry is either a rectangle or a rhombus.
3. (Intersecting Railroad Tracks). Suppose m and n are parallel lines with distance d between them, and suppose p and q are also parallel lines with the same distance d between them. If the parallel pairs intersect, tell what kind of quadrilateral their intersection points define and prove your claim.

4. (Circles and Parallelograms). Two problems:
   (a) Suppose parallelogram *ABCD* is inscribed in circle *c*. What properties must it have? Must it be a special kind of parallelogram? If so, what kind?
   (b) Suppose a circle *c* is inscribed in a parallelogram *ABCD*. What properties must *ABCD* have? Must it be a special parallelogram? If so, what kind?

5. (Midpoint Quadrilaterals). Theorem 7.19 proves that the midpoint quadrilateral *JKLM* of any quadrilateral *ABCD* is a parallelogram. Some of the figures below have midpoint quadrilaterals that are special kinds of parallelograms. Tell what they are and justify your answer.
   (a) A kite
   (b) A rhombus
   (c) A parallelogram
   (d) A rectangle
   (e) A square

6. (Midpoint Quadrilateral Symmetry). For the quadrilaterals in the previous problem, describe the set of all symmetries of the midpoint figure. Then answer these questions.
   (a) Is every symmetry of *ABCD* also a symmetry of the midpoint quadrilateral *JKLM*?
   (b) Is every symmetry of the midpoint quadrilateral *JKLM* also a symmetry of *ABCD*?
   (c) In either case, if yes, why? If not, is there any other relationship between the groups of symmetries?

7. (Nonconvex Quadrilaterals). In the proof of Theorem 7.19 it was asserted for a quadrilateral *ABCD* that if *B* and *D* are on the same side of $\overline{AC}$, then either *D* is in the interior of $\triangle ABC$ or *B* is in the interior of $\triangle ADC$ because the sides of the quadrilateral do not intersect except at the vertices. Prove this. Hint: Consider the regions of the plane created by the extended sides of $\triangle ABC$. Where can *D* be?

8. (A More General Midpoint Quad). In Figure 32 there is an example on the right of an *ABCD* that is not a parallelogram but for which the midpoint figure *JKLM* is still a parallelogram. Use what we have learned about translations to prove that if *ABCD* is not assumed to be a quadrilateral but is just four distinct points, then

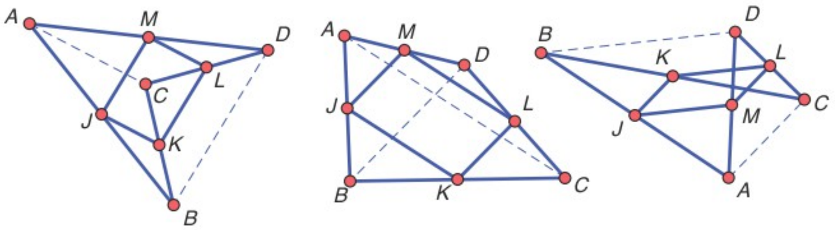

Figure 32. Quadrilateral Midpoints

the midpoint figure $JKLM$ is still either a parallelogram or a linear parallelogram. Hint: Prove the translations $\mathcal{H}_K\mathcal{H}_J$ and $\mathcal{H}_L\mathcal{H}_M$ are equal because they agree at a point. What does this say about $JKLM$?

9. (Translation Experiments). Here are some translation experiments using the midpoint triangle in Figure 13. You may need to sketch your own copy. Answer with points already labeled if possible, but describe and label some new points if needed.
   (a) What point is $\mathcal{T}_{CF}(F)$? How do you know?
   (b) Describe the image $\mathcal{T}_{CF}(\triangle DEF)$.
   (c) Tell some points P and Q for which $\mathcal{T}_{PQ} = \mathcal{H}_E\mathcal{H}_D$.
   (d) Tell some points P and Q for which $\mathcal{T}_{PQ} = \mathcal{T}_{BA}\mathcal{T}_{DF}$.
   (e) For what point P is $\mathcal{H}_P = \mathcal{H}_A\mathcal{H}_E\mathcal{H}_D$?

10. (Fourth Parallelogram Point). Given a parallelogram $ABCD$, we know from Theorem 7.23 that D is uniquely determined when given A, B, C. But the parallelogram $ACBE$, with the same three points in different order, is a different quadrilateral. Choose any noncollinear points A, B, C. Then draw all the parallelograms with these three points as vertices.

11. (Tiled Triangle). The triangle tiled with triangles in Figure 14 was created assuming the Euclidean Parallel Postulate as a theorem. By subdividing further with repeated midpoint triangles and taking limits, this tiled triangle suggests how the Dilation Axiom could be proved by assuming EPP as a postulate instead of the reverse, as we have done.
   (a) Check that the AD/AB and AE/AC equal $\|DE\|/\|BC\|$. Check that $\angle ABC \cong \angle ADE$ and $\angle ACB \cong \angle AED$. Thus, for this example, ratios and angles are preserved.
   (b) The ratios involved are also related to area, or equivalently, to the number of triangles in various figures. What is the total number of small triangles in $\triangle ABC$? What is the number in $\triangle ADE$? How is this related to the ratios of lengths?
   (c) Let $A' = \mathcal{T}_{EC}(A)$ and $D' = \mathcal{T}_{EC}(D)$. Tell how many small triangles are in each of these polygons: $\triangle DBD'$, $\triangle A'D'C$, quadrilateral $ADD'A'$. How are these numbers related to ratios of length?

12. (Parallelograms that Are Rectangles). It was observed that rectangles have congruent diagonals and right vertex angles. Prove these converses.
   (a) Any parallelogram with congruent diagonals is a rectangle.
   (b) Any parallelogram with one vertex angle that is a right angle is a rectangle.

13. (Concurrence of Medians). A median of a triangle is a segment from a vertex to the midpoint of the opposite side. This is the first of several proofs of a theorem about medians of a triangle.
   (a) In a triangle $\triangle ABC$, let the medians be AD, BE, and CF. Let G be the intersection of BE and CF. Consider the midpoint quadrilateral of $ABGC$ and use it to find these ratios: BG/BE and CG/CF.
   (b) Let H be the intersection of BE and AD. Find the ratios BH/BE and AH/AD.
   (c) Explain how these two results prove that the three medians are concurrent at $G = H$.

14. (Translations as Products of Line Reflections). In Figure 33 the large triangle is mapped by one product of two line reflections and the small triangle by another, but the different products are the same translation. Prove that products of reflections in parallel lines $\mathcal{R}_n\mathcal{R}_m$ and $\mathcal{R}_{n'}\mathcal{R}_{m'}$ are equal if and only if there is a translation T such that $T(m) = m'$ and $T(n) = n'$.

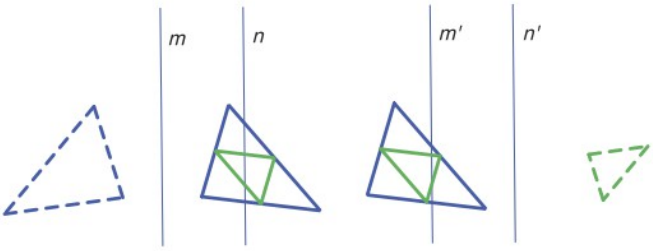

Figure 33. A Translation Shown as Two Different Double Line Reflections

15. (Rotating One Ray to Another). Let $\overrightarrow{AB}$ and $\overrightarrow{CD}$ be two rays that do not have the same direction. Prove that there is a unique rotation P_k that maps $\overrightarrow{AB}$ to $\overrightarrow{CD}$. Hint: How did we prove congruence of segments?
16. (Invariant Lines). A line m is an **invariant line** for a rigid motion T if $T(m) = m$. If the points of the line are all fixed points, the line is invariant, but in general the points of the line will not all be fixed.
    (a) For a reflection $\mathcal{R}_n$, find all the invariant lines, in addition to n. Explain your reasoning.
    (b) For a half-turn $\mathcal{H}_A$, find all the invariant lines. Explain your reasoning.
    (c) For a translation $\mathcal{T}_{AB}$, find all the invariant lines. Explain your reasoning.
    (d) For a glide reflection $\mathcal{G}_{AB}$, find all the invariant lines. Explain your reasoning.
    (e) For a rotation A_{60}, find all the invariant lines? Explain your reasoning.
17. (Transformation Groups). Transformation groups were defined in Definition 5.10. Tell which of the following sets are transformation groups and justify your answer:
    (a) the set of all translations,
    (b) the set of all half-turns,
    (c) the set of all transformations that are either half-turns or translations,
    (d) for a given line m, the set of all translations that map this line into itself,
    (e) for a given ray $\overrightarrow{AB}$, the set of all translations that map this ray into itself.

Chapter 11

Coordinate Geometry

Coordinates attach numbers to points in such a way that algebra can be used to investigate geometry. In the 17th century, Fermat and Descartes developed the beginnings of the coordinate systems that are so familiar today.

To express rigid motions and distance and angle relationships, we will want a Cartesian coordinate system with perpendicular axes and rulers on each axis. But there are applications that do not require such a restricted coordinate system, so we begin more simply with parallel lines and ratios to define what are called affine or linear coordinates. Then we develop formulas for half-turns and translations and equations for lines in affine coordinates. Afterwards, for the metric part of Euclidean geometry, we will restrict ourselves to Cartesian coordinates, which are coordinates with more constraints on angles and distances.

The introduction of coordinates will connect the abstract model of the plane developed up to now with algebra and the (x, y)-plane $\mathbb{R}^2$. This will provide some new algebraic tools for studying geometry, but the geometry will also provide insight into the algebra.

Axes and Coordinates

In the plane, take a triangle $\triangle E_0E_1E_2$ as the reference triangle for a system of coordinates based on parallel lines and signed ratios.

On the line $\overline{E_0E_1}$, let $x(P) = E_0P/E_0E_1$. This ratio is a coordinate proportional to a ruler on the line, with $x(E_0) = 0$ and $x(E_1) = 1$. In the same way, on $\overline{E_0E_2}$, let $y(P) = E_0P/E_0E_2$.

The first axis, or x-**axis**, of the coordinate system is $\overline{E_0E_1}$; the second axis, or y-**axis**, is $\overline{E_0E_2}$. The **origin** of the system is the point E_0.

For any point P in the plane, let p_1 be the line through P parallel to or equal to the y-axis. This line intersects the x-axis at a point P_1. Define $x(P) = x(P_1)$.

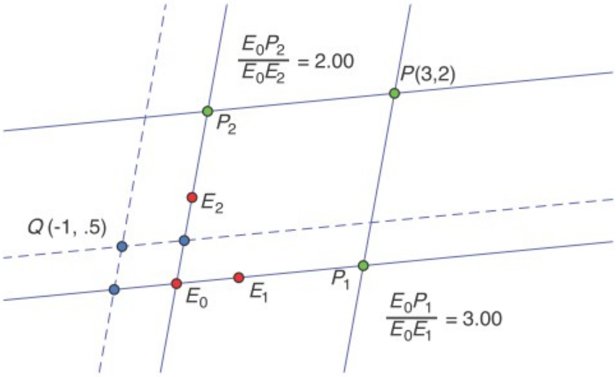
$\frac{E_0P_2}{E_0E_2} = 2.00$
P(3,2)
P_2
E_2
Q (-1, .5)
E_0
E_1
P_1
$\frac{E_0P_1}{E_0E_1} = 3.00$

Figure 1. Axes, Parallel Lines, and Coordinates

Again, in the same way, let p_2 be the line through P parallel to the x-axis, intersecting the y-axis at P_2. Define $y(P) = y(P_2)$. The figure $E_0P_1PP_2$ is a parallelogram (or linear parallelogram).

Define the **affine coordinates** of P with respect to this reference triangle, or axis system, to be the pair $(x(P), y(P))$.

Figure 1 shows two points and their coordinates, the point P with coordinates $(3, 2)$ and Q with coordinates $(-1, .5)$.

The coordinates (x, y) define a map of the plane to real 2-space $\mathbb{R}^2$. So long as the coordinate system remains unchanged, we will refer to points by their coordinates, either as $P(2, 3)$ or even by saying $P = (2, 3)$.

Equations for other lines will come soon, but from the definition of coordinates, it is clear that $x = 2$ is the equation of a line parallel to the y-axis, and $y = 3$ is the equation of a line parallel to the x-axis.

Note (Terminology)**.** We have called these coordinates affine coordinates, but such (x, y)-coordinates may also be called linear coordinates. What is the difference? For affine coordinates we must specify three points, the origin and then the unit points on the axes. Linear coordinates have the same definition except that they occur in a space where the origin is already understood, such as $\mathbb{R}^2$, where $(0, 0)$ is normally taken to be the origin, or the vectors in the plane, where the zero vector is understood to be the origin. However, this can seem to be a very fine distinction once an origin E_0 has been settled on and designated as O.

Midpoints, Half-turns, and Translations

For any two points A and B on the plane, one can find the coordinates of the midpoint of AB using what has been proved about parallelograms and parallels. The figure is full of overlapping parallelograms that can be used to prove this formula, but $\overline{AB}$ is also a transversal of two sets of parallel lines (if it is not parallel to an axis), so the Transversal Ratios Theorem, Theorem 8.17 can be applied.

3: D 4 7 +))) A

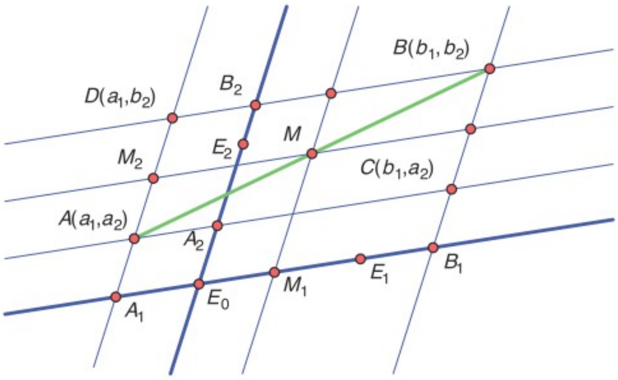

Figure 2. Parallelograms and Midpoints

Theorem 11.1 (Midpoint Formula). *In linear coordinates, the formula for the midpoint M of AB is*

$$(x(M), y(M)) = \frac{1}{2}(x(A) + x(B), y(A) + y(B)).$$

This can also be written more compactly as

$$M = \frac{1}{2}(A + B)$$

where a sum of points here means a sum of number pairs.

Proof. Let $A = (a_1, a_2)$ and $B = (b_1, b_2)$. A point M is the midpoint of AB when the signed ratio $AM/BM = -1$. If AB is not parallel to the y-axis, Theorem 8.17 says that ratios are preserved, so $A_1M_1/B_1M_1 = -1$ and M_1 is also a midpoint. The midpoint of A_1B_1 has coordinate $(1/2)(a_1 + b_1, 0) = (1/2)(x(A) + x(B))$, since the signed ratio

$$\frac{(1/2)(a_1 + b_1) - a_1}{(1/2)(a_1 + b_1) - b_1} = \frac{b_1 - a_1}{a_1 - b_1} = -1.$$

For the case when AB is parallel to the y-axis, then $M_1 = A_1 = B_1$, so the x-coordinates are all equal and it remains true that $x(M) = (1/2)(x(A) + x(B))$.

The same reasoning applies to the other axis so $y(M) = (1/2)(y(A) + y(B))$. ☐

A formula for any half-turn is an immediate consequence of the midpoint formula.

Theorem 11.2 (Half-turn Formula). *For any half-turn $\mathcal{H}_P$, the formula for the image of a point A is $\mathcal{H}_P(A) = 2P - A$.*

Proof. This is the midpoint formula rearranged. If $B = \mathcal{H}_P(A)$, then P is the midpoint of AB, so $P = (1/2)(A + B)$. If A and P are known, this becomes a formula for the image of A: $B = 2P - A$. ☐

Figure 3 shows examples of $\mathcal{H}_P$ for $P(0, 2)$. If $D = (-1, 1)$, then $\mathcal{H}_P(D) = 2(0, 2) - (-1, 1) = (1, 3)$ and $\mathcal{H}_P(E_1) = 2(0, 2) - (1, 0) = (-1, 4)$.

3: D 4)) 4)) + +*

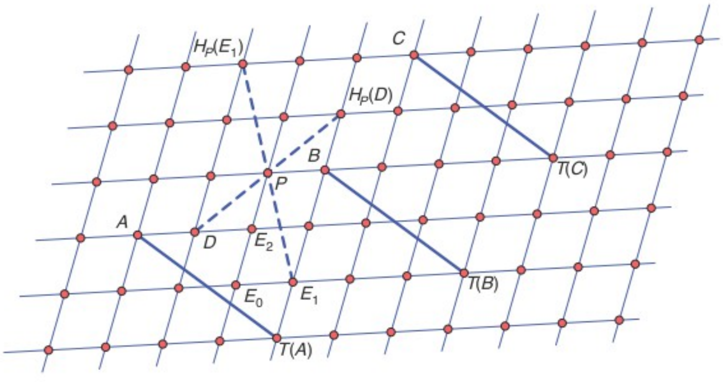

Figure 3. Examples: A Half-turn and a Translation

Since translations are compositions of half-turns, this also provides a formula for translations.

Theorem 11.3 (Translation Formula)**.** *The formula for translation* $T = \mathcal{H}_B\mathcal{H}_A$ *is* $T(P) = P + 2(B - A)$. *If* T *is written as* $\mathcal{T}_{CD}$, *then* $T(P) = P + (D - C)$.

Proof. The first formula is simply a substitution: $\mathcal{H}_B\mathcal{H}_A(P) = \mathcal{H}_B(2A - P) = P + 2(B - A)$. If $T(C) = D$, then $D = C + 2(B - A)$. □

The example in Figure 3 shows $T(A) - A = (1, -1) - (-2, 1) = (3, -2)$. Then $T(B) = (1, 2) + (3, -2) = (4, 0)$ and $T(C) = (2, 4) + (3, -2) = (5, 2)$. The translation is represented by three directed segments that are translates of one another.

The translation formula suggests a way to represent vectors by means of number pairs. Since $\mathcal{T}_{AB} = \mathcal{T}_{CD}$ if and only if $B - A = D - C$, then representing $[AB]$ with the number pair $B - A$ would be consistent with the vector concept that two vectors are equal if their associated translations are equal. In the example in Figure 3, $T = \mathcal{T}_v = \mathcal{T}_{[3,-2]}$.

This list of formulas has a lot of consequences, some of which we have seen before and some not.

As a starting point, consider what the translation formula means for parallelograms. We learned in Theorem 7.30 that $ABCD$ is a parallelogram (or linear parallelogram) if and only if the segment DC is a translate of AB. This means that $C = \mathcal{T}_{AD}(B)$, or $C - B = D - A$. This is consistent with the fact that the midpoints of the diagonals are the same: $(1/2)(A + C) = (1/2)(B + D)$.

Theorem 7.19 proved that the midpoint quadrilateral $JKLM$ of any quadrilateral $ABCD$ is a parallelogram. Using coordinates and the midpoint formula, the following computation shows $JKLM$ is a parallelogram or linear parallelogram for any four points A, B, C, D (with the caveat that a pair of sides may have length 0 if $J = K$ or $J = M$):

$$J - K = \frac{1}{2}(A + B) - \frac{1}{2}(B + C) = \frac{1}{2}(A - C) = \frac{1}{2}(A + D) - \frac{1}{2}(D + C) = M - L.$$

Lines, Dilations, and Equations

How are lines represented in this coordinate system? Let's begin with lines through the origin E_0. Also, since we will be referring to the origin a lot, to avoid so many subscripts, let's use O as another name for this point.

In thinking about points on a line through the origin O, it is natural to think of dilations with center O. For every t, the dilations $\mathcal{D}_{O,t}(P)$ lie on $\overleftrightarrow{OP}$; in fact, every point on this line is one of these dilation images. What are the coordinates of these points?

Theorem 11.4 (A Dilation Formula). *If* $P = (a, b)$*, then* $\mathcal{D}_{O,t}(P) = (ta, tb)$*. In more compact notation,* $\mathcal{D}_{O,t}(P) = tP$*.*

Proof. For P, consider the coordinate parallelogram OP_1PP_2 from Figure 1. The dilation not only maps P to a point, it also maps the entire parellogram to a new coordinate parallelogram.

For example, $P_1 = (a, 0)$ is the point on $\overleftrightarrow{OE_1}$ with $OP_1/OE_1 = a$. If P_1' denotes the dilation image of P_1 and E_1' denotes the dilation image of E_1, then, since signed ratios are preserved by dilations, $OP_1'/OE_1' = a$. But $OE_1'/OE_1 = t$, so $OP_1'/OE_1 = (OP_1'/OE_1')(OE_1'/OE_1) = ta$, and $P_1' = (ta, 0)$. In the same way the image of P_2, $P_2' = (0, tb)$, so $\mathcal{D}_{O,t}(P) = (ta, tb)$. □

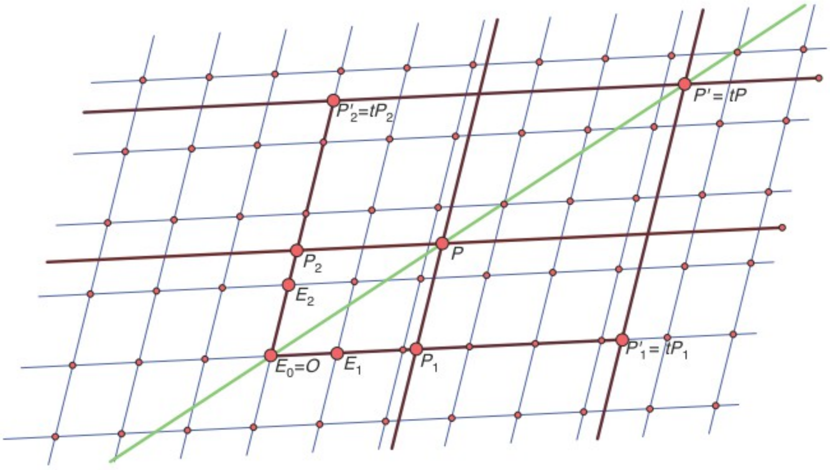

Figure 4. Dilation of P by Ratio t

Thus, the function $L(t) = tP$ from $\mathbb{R}$ to $\overleftrightarrow{OP}$ is a **parametrization of this line through the origin**. It is not a ruler, since the difference of the parameters does not measure the distance between points, but it is proportional to a ruler by a scale factor.

3: D 4 ")) :/ 1) A ( )) 9) 4 4

Line Equation and Slope. When $P = (a, b)$, according to this parametrization, for any point (x, y) on $\overline{OP}$, there is a t so that $(x, y) = \mathcal{D}_{O,t}(P) = (ta, tb)$. The equation $x = ta$ says that $t = x/a$, provided $a \neq 0$, so the second coordinate satisfies the equation $y = bt = (b/a)x$. Conversely, any point $(x, (b/a)x)$ is the dilation of P by ratio x/a.

Therefore, the points (x, y), where x and y are in the solution set of the equation $y = (b/a)x$, is the set of points on the line $\overline{OP}$.

If $a = 0$, the line is the y-axis, so the equation for the line is $x = 0$ since the dilations of P have coordinates $(0, tb)$.

Here is another way to picture the constant b/a. If a line through O is not the y-axis, then the line is a transversal for all lines parallel to this axis and so intersects the parallel line through E_1 with equation $x = 1$. Let the intersection point be $(1, m)$. Then $m = (b/a)1 = b/a$. So the y-coordinate of the intersection with the line $x = 1$ is $m = b/a$. This constant m is given the name the **slope** of the line, and the letter m is often used for this number. Using this slope notation, the equation for a line through O looks like this:

$$y = mx.$$

Note (Meaning of Slope)**.** We have used the slope terminology here since that is what one rather universally calls m in this equation. However, for affine coordinates, slope need not indicate steepness or even whether a line is rising from left to right on a computer screen or a piece of paper. The triangle defining the coordinates may be oriented in directions very different from graph paper. If the E_1 is very far from E_0 in a vaguely northwest direction and E_2 is close to E_0 in an easterly direction, the meaning of slope $m = 1$ will be very different from what it means for more familiar coordinate systems.

Line Parametrization. Now that we have a parametrization for lines through the origin, what is a parametrization for a general line $q = \overline{Q_1Q_2}$?

Theorem 11.5 (Affine Parametrization of a Line)**.** *A parametrization of the line $\overline{Q_1Q_2}$ is*

$$P(t) = Q_1 + t(Q_2 - Q_1) = (1 - t)Q_1 + tQ_2.$$

The signed ratio $P(t)Q_1/Q_2Q_1 = t$.

Proof. The translation $\mathcal{T}_{Q_1O}$ maps q to the parallel line q' though the origin O. The point $\mathcal{T}_{Q_1O}(Q_2) = Q_2' = Q_2 - Q_1$ is on q'. By the parametrization for lines through O, all points P on q' are of the form $t(Q_2 - Q_1)$ for some t, where $t = PO/Q_2'O$. The inverse translation T^{-1} maps this point to the point $Q_1 + t(Q_2 - Q_1)$. Thus, we have a parametrization for the line $\overline{Q_1Q_2}$. The signed ratio is preserved by translation. ☐

Note: Since the sum of the real coefficients is 1, this is a weighted average of the two points. This property was hidden in the formula, tP, which really should be written $(1 - t)O + tP$.

Equation for a Line Through Two Points. What is an equation satisfied by the points of q? Of course, if the line is parallel to the y-axis, the equation is $x = x_1$ and there is no slope m. So from now on, assume that q is not such a line.

Consider the slope of the parallel line q' through O. If $Q_1 = (x_1, y_1)$ and $Q_2 = (x_2, y_2)$, then the point with coordinates $Q_2 - Q_1 = (x_2 - x_1, y_2 - y_1)$ is on the line, so the equation $y = mx$ becomes

$$y = mx = \frac{(y_2 - y_1)}{(x_2 - x_1)}x.$$

This slope m can be computed from any two points Q_1 and Q_2 on the line, since the value of m is determined by the slope of the parallel line through O.

Now, to find an equation satisfied by every point $Q = (x, y)$ on the line, pick any point Q_0 on q (which may be Q_1 or Q_2 but need not be). Then the translation $\mathcal{T}_{Q_0O}$ maps q to q', the parallel line through O, and maps Q to $Q - Q_0 = (x - x_0, y - y_0)$. Since this point is on q', it satisfies this equation:

$$y - y_0 = m(x - x_0) = \frac{(y_2 - y_1)}{(x_2 - x_1)}(x - x_0).$$

This equation is the point-slope form of the equation of a line. If one picks $Q_0 = Q_1$, then there are fewer symbols in the formula, for then $y - y_1 = m(x - x_1)$.

Two General Line Equations. If the constant terms in this equation are collected together, then the equation above takes on a simpler form. This is an equation for a line not parallel to the y-axis:

$$y = mx + b.$$

The intersection $(0, b)$ of this line with the y-axis is obtained by setting $x = 0$. This point is the y-intercept, and this is the **slope-intercept** form of the equation of a line.

There is one other common form of the equation, for real constants a, b c, with $(a, b) \neq (0, 0)$. This form includes lines parallel to the y-axis, since these lines have equation $x = c$:

$$ax + by = c.$$

This equation is not unique, since any multiple of it by a nonzero k has the same solution set: $kax + kby = kc$.

But if the coefficients a and b are not changed, then each change of constant c has as solution a different line parallel to m, since for $c_1 \neq c_2$, the equations $ax + by = c_1$ and $ax + by = c_2$ have no common solution.

This completes the derivation of the equation of a line in affine coordinates.

Euclidean Geometry and Cartesian Coordinates

Now we move on to the more familiar picture of perpendicular Cartesian coordinates. To work with distances and angles, we need a more restricted kind of coordinates. These Cartesian coordinates are simply affine coordinates where $\triangle E_0E_1E_2$, the reference triangle, is a right isosceles triangle, with right angle at E_0 and sides of length 1. Again, we will write the origin $O = E_0$.

Then the coordinates on each axis are ruler coordinates that are signed distances, not just ratios. The coordinate parallelograms OP_1PP_2 are now rectangles.

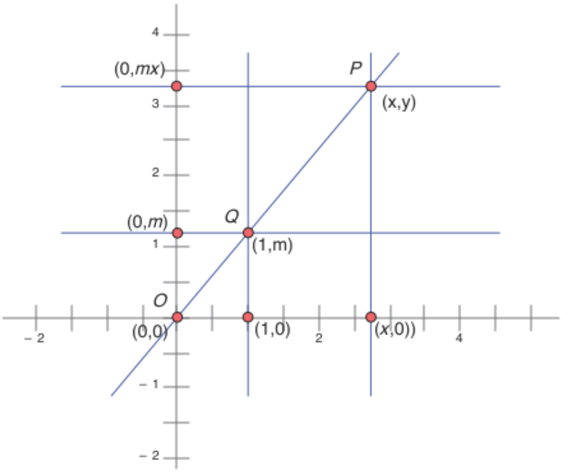

Figure 5. Slope of a Line Through the Origin

Everything that we proved for affine coordinates remains true in the Cartesian case, but some features look different enough that we will revisit them.

For the equation $y = mx$, the slope m is now a ratio of side lengths of a rectangle, as illustrated in Figure 5. It indicates steepness in a way that the general affine case did not.

Also, the similar triangles, attached to a line to derive the point-slope form of the line equation, are now right triangles, all of which are similar and have the same internal ratio of side length, which is the slope. This is illustrated in Figure 6.

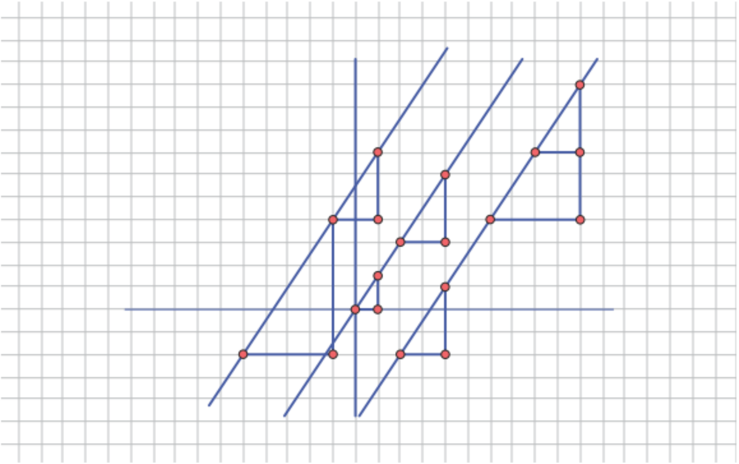

Figure 6. Similar Triangles Dangling from Parallel Lines

Distance Formula. Let P_1 and P_2 be points in the plane with coordinates (x_1, y_1) and (x_2, y_2). Then if M is the point with coordinates (x_2, y_1) and N is the point with coordinates (x_1, y_2), then P_1MP_2N is a rectangle, with sides parallel to the axes, as shown in Figure 7.

By the Pythagorean Theorem, $\|P_1P_2\|^2 = \|P_1M\|^2 + \|MP_2\|^2$.

Since P_1M is parallel to the x-axis, the coordinate x is a ruler, so the length of this segment is $|x_2 - x_1|$. Likewise, the length of MP_2 is $|y_2 - y_1|$. This gives the formula for the distance in terms of coordinates:

$$\|P_1P_2\| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

Now that there is a way to measure distance, one can identify a few rigid motions without much calculation. Again, let's denote E_0 by O. The reflection $\mathcal{R}_{OE_1}(x, y) = (x, -y)$. This clearly fixes the points of the x-axis and it also preserves distance, since the change of sign does not change the value of the square terms in the distance formula. The reflection in the y-axis maps (x, y) to $(-x, y)$.

Another simple reflection formula is reflection in the diagonal line s with equation $y = x$. The reflection $\mathcal{R}_s(x, y) = (y, x)$ fixes the points of this line and again does not change the value of the expressions in the distance formula. This reflection also maps each axis to the other.

The general formula for reflections is more complicated. We will take it up when we consider the general formula for rotations.

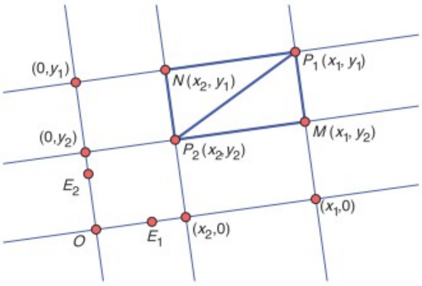

Figure 7. Coordinates and Distance

Perpendicular Lines in the Coordinate Plane

The formula for O_{90}, the 90-degree rotation centered at O, can be deduced in several ways. Given $P = (a, b)$, the rotation will rotate the entire coordinate rectangle for P to a new rectangle. The point $(a, 0)$ is rotated to the point $(0, a)$ on the y-axis. The point $(0, b)$ is rotated to $(-b, 0)$ on the x-axis. Therefore, this new rectangle is the coordinate rectangle for $(-b, a)$. This gives a formula $O_{90}(a, b) = (-b, a)$ valid for all points in the plane.

A second way to see this is from a product formula. Let s be the line $x = y$ as in the previous section. We know that $O_{90} = \mathcal{R}_s\mathcal{R}_{OE1}$, and $\mathcal{R}_s\mathcal{R}_{OE1}(x, y) = \mathcal{R}_s(x, -y) = (-y, x)$.

The rotation of this rectangle for a point P with positive coordinates is pictured in Figure 8. From this, one can already see the coordinates of P'.

This formula is consistent with the formula for half-turns, for $O_{90}^2(a, b) = O_{90}(-b, a) = (-a, -b)$, which is the formula for $\mathcal{H}_O$.

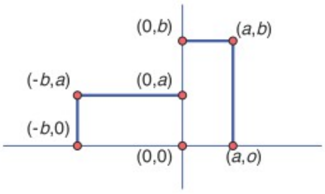

Figure 8. Rotation by 90 Degrees

This tells us something about the slope of perpendicular lines. The line OP has slope b/a, but the rotated line OP' has slope $-a/b$. Since every line is parallel to a line through O, this proves this result:

Theorem 11.6 (Perpendicular Line Equation)**.** *If the equation of a line is* $y = mx + b$, *the equation for every line perpendicular to this line is* $y = (-1/m)x + c$, *for some* c.

Euclidean Geometry of the General Line Equation. All the equations for a line can be converted to the form

$$ax + by = c.$$

If $b \neq 0$, this line has slope $-a/b$. Let m be the line through O with equation $ax + by = 0$.

If one plots the point $N = (a, b)$, the line $\overline{ON}$ has slope b/a, so it is perpendicular to m by the previous theorem. Thus, the direction from O towards N is perpendicular to m; this is sometimes called the **normal direction**.

The point N satisfies the equation $ax+by = a^2+b^2$, so this equation is the equation of a line parallel to m at distance $d = \|ON\| = \sqrt{a^2 + b^2}$ from m, since ON is a segment perpendicular to both lines.

Any point tN satisfies $ax+by = (a^2+b^2)t = d^2t$ and is on a parallel line at distance $|t|d$ from m. The point $N_c = (c/(a^2 + b^2))N$ is the intersection point of $\overline{ON}$ and the line with equation $ax + by = c$.

We can replace the original coefficient pair (a, b) with a **unit normal** point U on $\overline{ON}$ at distance 1 from O. Let $U = (p, q) = (1/d)N = (a/d, b/d)$. This is the equation of a line parallel to m at distance $|c|$ from m:

$$px + qy = \frac{a}{\sqrt{a^2 + b^2}}x + \frac{b}{\sqrt{a^2 + b^2}}y = c.$$

The lines with $c > 0$ are the lines on the same side of m as (a, b). The lines with $c = 0, 1, 2, 3, \ldots$ are parallel lines spaced a unit distance apart.

For any point $M = (m, n)$ on the line $ax + by = c$, $c \geq 0$, consider the right triangle $\triangle MON_c$. If $\theta = m\angle MON_c$, then the distance

$$|c|/\sqrt{a^2 + b^2} = \|ON_c\| = \|OM\| \cos\theta = \sqrt{m^2 + n^2}\cos\theta$$

or $am + bn = c = \sqrt{a^2 + b^2}\sqrt{m^2 + n^2}\cos\theta$.

Definition 11.7. The expression $am + bn$ is called the **dot product** $N \cdot M$, with $N = (a, b)$ and $M = (m, n)$.

From our interpretation of the equation $am + bn = c$, the value of the dot product c is given by this equation:

$$N \cdot M = \|ON\|\|OM\| \cos\theta.$$

This formula is also true for a negative dot product when the angle θ will be obtuse. This case is left to the exercises.

The dot product has been introduced here as an algebraic formula applied to number pairs. The geometric interpretation of $N \cdot M$ given in the equation does not come into play unless an origin O has been designated. Therefore, the dot product is not really a function of two points, since it implicitly requires more information. The resolution of this awkward situation is that **the dot product is really a pairing of vectors, not points**! So the dot product of two vectors, $[AB] \cdot [CD]$ is the algebraic formula $(B - A) \cdot (D - C)$. For vectors, the formula with $\cos\theta$ still holds if θ is interpreted as the difference between the polar angles of the two vectors. This agrees with what has gone before if one interprets $N \cdot M$ as $[ON] \cdot [OM]$.

The dot denoting this product is the same notation used earlier for signed products of collinear segments. This should not lead to confusion since for such segments the product $AB \cdot CD$ equals $[AB] \cdot [CD]$ because for such segments θ is 0 or 180, so $\cos\theta = \pm 1$.

Algebraic properties of the dot product simplify calculations with number pairs. Here M, N, P, Q are any points. The demonstrations of these formulas are either substitution calculations or were already carried out in the paragraphs above.

Commutative Property: $N \cdot M = M \cdot N$.

Distributive Property: $(aM + bN) \cdot P = aM \cdot P + bN \cdot P$.

Perpendicular Directions: $(M - P) \cdot (N - Q) = 0$ if and only if $\overline{MP}$ is perpendicular to $\overline{NQ}$.

Distance: $\|PQ\|^2 = (P - Q) \cdot (P - Q)$.

Graphs and Transformations

If f is a function from the real numbers to the real numbers, we graph this function by plotting the points (x, y) such that $y = f(x)$.

This is closely related to the geometry of the plane; in fact, the equation of a line can be interpreted as the graph of the function $f(x) = mx + b$. It is also true that certain transformations are performed on graphs, with the transformed graphs then interpreted as functions. Here are some examples.

- The transformation $R(x, y) = (y, x)$ maps the graph to its reflection in the line with equation $x = y$; the image can be interpreted as the graph of f^{-1}.
- The transformation $S(x, y) = (x + a, y + b)$ maps the graph to the graph of the function $f(x - a) + b$.
- The transformation $T(x, y) = (-x, y)$ maps the graph to the graph of the function $f(-x)$.
- The transformation $U(x, y) = (2x, 2y)$ maps the graph to the graph of the function $2f(x/2)$.
- The transformation $V(x, y) = (2x, y)$ maps the graph to the graph of the function $f(x/2)$.

This looks a lot like the geometry of the Euclidean plane. Almost all of these transformations are ones we have seen. But there are several fundamental ways in which the space of graphing is not the plane that we have been studying.

First, most of the transformations of the Euclidean plane are not allowable in the space of graphs. It does not make sense to rotate a graph by 30 degrees or reflect it in the line with equation $y = 2x$.

Second, while some of the transformations applied to graphs are similitudes, others are not and do not preserve angles or scale distance. Transformation V in the list above is an example. Also, in the graph world, the slope has units and can change if the units change from seconds to hours or from dollars to euros.

There is a reason for this difference. The Euclidean plane is a homogeneous space. All locations and all directions are equivalent. Even when we place coordinates on the plane, if we are in Euclidean geometry, we still have no preferred directions, just simplified equations perhaps.

However, the space of graphing is not a unified two-dimensional space; rather it is the joining of two one-dimensional spaces as a product. This is seen most clearly when the function being graphed is not a function of a distance to a distance but a function of time to money, or temperature to density. The quantities measured by the x and the y can be completely different, and the scales on the axes can be different. If one is plotting time versus billions of dollars, then the visual slope of the graph on the page depends on whether one unit on the graph paper is one million or ten billion.

Just because two mathematical quantities are described by two real numbers, they need not be the same thing. Even points and vectors are not the same kind of objects. The space of graphing is not really $\mathbb{R}^2$ but $\mathbb{R}^1 \times \mathbb{R}^1$, a product. This means that the space consists of pairs, with the first element a number from one space, the x-space, and the second from a second space, the y-space. The first element may be time measured in years and the second may be dollars measured in megabucks. Just because both spaces are $\mathbb{R}$, there is no reason that the two are otherwise related.

So, in short, one can learn something about graphs from geometry about slope and other properties, but one should be very cautious when applying geometrical relationships to the world of graphs.

Unit Circle and Rotation Formula

Rotations O_θ map each circle centered at O onto itself. The **unit circle**, the circle of radius 1 with center O, will play an important role in the formula for rotations.

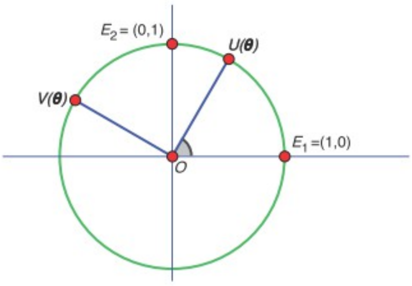

Figure 9. Points on Unit Circle

Let polar angles at O be measured with a protractor that maps the angle 0 to $\overrightarrow{OE_1}$, with 90 being the polar angle of $\overrightarrow{OE_2}$. Then every point on the unit circle has an angle that can be taken to be in the range from 0 to 360, or from −180 to 180.

A point $U(\theta)$ corresponding to polar angle θ on the unit circle has coordinates (c, s), with $c^2 + s^2 = 1$. We can extend the definition of trigonometric functions from acute angles to all of these angles by setting $U(\theta) = (\cos\theta, \sin\theta)$.

One way of thinking about the point $U(\theta)$ is that it is a rotated point, namely $O_\theta(E_1)$.

Denote by $V(\theta)$ the rotation of $U(\theta)$ by 90 degrees; by the formula for this rotation, $V(\theta) = (-\sin\theta, \cos\theta) = O_\theta(E_2)$.

Rotation Formula for Center O. One can rotate a point $P = (x, y)$ by angle θ along with its coordinate rectangle OP_1PP_2, with diagonal OP and vertices at $(x, 0)$ and $(0, y)$ as shown in Figure 10.

The rotation images by O_θ of the points on the axes are multiples of $U(\theta)$ and $V(\theta)$: $P_1' = O_\theta(x, 0) = xU(\theta)$ and $P_2' = O_\theta(0, y) = yV(\theta)$. This is clear from the direction and length of the segments OP_1' and OP_2'.

Then $P' = O_\theta(P)$ is a translate of O: $P' = \mathcal{T}_{OP_1'}\mathcal{T}_{P_1'P'}(O)$, which equals $\mathcal{T}_{OP_1'}\mathcal{T}_{OP_2'}(O)$ since $OP_1'P'P_2'$ is a rectangle. Thus,

$$P' = O + P_1' + P_2' = xU(\theta) + yV(\theta)$$

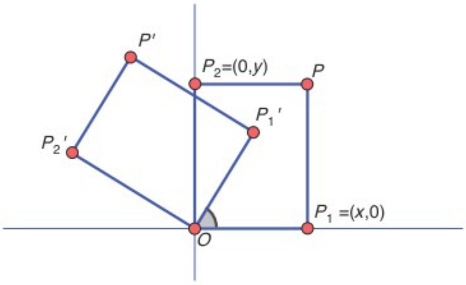

Figure 10. Rotation of P and Its Coordinate Rectangle

and so

$$O_\theta(P) = O_\theta(x, y) = x(\cos\theta, \sin\theta) + y(-\sin\theta, \cos\theta).$$

For a formula for A_θ, centered at a point A that is not O, simply translate by $\mathcal{T}_{AO}$, then rotate with center O, then translate back: $\mathcal{T}_{OA}O_\theta\mathcal{T}_{AO}(P) = A + O_\theta(P - A)$. From the Product of Rotations Theorem, Theorem 10.2, this is a rotation by angle θ. Since A is a fixed point, the product must be A_θ.

The rotation formula implies the addition formulas for cosine and sine. First apply the formula:

$$U(\theta + \phi) = O_\theta(U(\phi)) = O_\theta(\cos\phi, \sin\phi) = \cos\phi(\cos\theta, \sin\theta) + \sin\phi(-\sin\theta, \cos\theta).$$

Then note the value of $U(\theta + \phi)$:

$$(\cos(\theta + \phi), \sin(\theta + \phi)) = (\cos\phi\cos\theta - \sin\phi\sin\theta, \cos\phi\sin\theta + \sin\phi\cos\theta).$$

For the special case $\theta = \phi$, this gives formulas for $\cos 2\theta$ and $\sin 2\theta$.

Line Reflection Formula. A formula for reflection in a line m through O can be derived in the same way. Let $m = \overline{OU(\theta)}$.

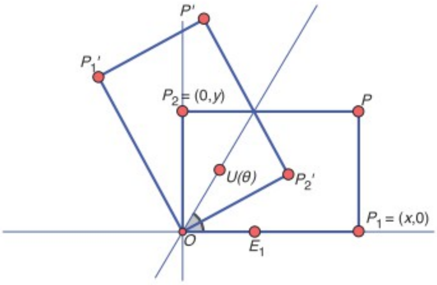

Figure 11. Reflection of P and Its Coordinate Rectangle

- The reflection image $P_1' = \mathcal{R}_m(P_1)$ is also the rotation image $O_{2\theta}(P_1)$, since $\angle P_1OU(\theta) \cong P_1'OU(\theta)$ and $OP_1 \cong OP_1'$, with $\theta = m\angle P_1OU(\theta)$. Therefore, $P_1' = xU(2\theta)$.

- The reflection image $P_2' = \mathcal{R}_m(P_2)$ is also the rotation image $O_{2(\theta-90)}(P_2)$, since $\angle P_2OU(\theta) \cong P_2'OU(\theta)$ and $OP_2 \cong OP_2'$ with $|\theta - 90| = m\angle P_2OU(\theta)$. Therefore, $P_2' = yV(2\theta - 180) = -yV(2\theta)$.

As for rotations, this leads to a formula for $P' = \mathcal{R}_m P = P_1' + P_2'$:

$$\mathcal{R}_{OU(\theta)}(P) = xU(2\theta) + yV(2\theta) = x(\cos 2\theta, \sin 2\theta) + y(\sin 2\theta, -\cos 2\theta).$$

We relied on Figure 11 to derive this formula. But now that we have a formula, we can check directly that the function is the correct reflection for all θ. Let F be the function with the formula given in the displayed equation. One can check by a calculation that this F preserves distance, so it is a rigid motion. Also one can calculate $F(O) = O$, $F(U(\theta)) = U(\theta)$, and $F(V(\theta)) = -V(\theta)$. Thus, F agrees with the line reflection $\mathcal{R}_{OU(\theta)}$ at these three points, so it must be the same rigid motion. Some details of this calculation will appear in the exercises.

Reflections in Lines Not Through O. Again, for a reflection in a line n not through O, let A be a point of n, and let $n' = \mathcal{T}_{AO}(n)$. Then $\mathcal{T}_{OA}\mathcal{R}_{n'}\mathcal{T}_{AO}$ is reflection in n, so $\mathcal{R}_{n'}(P) = A + \mathcal{R}_{n'}(P - A)$.

With this formula, we now have formulas for translation, dilation, rotation, and reflection. In the next two sections, we will show how these formulas can be expressed with different mathematical tools or in different mathematical language.

Complex Numbers and Transformations of the Plane

The complex numbers have the form $z = a + ib$, where a and b are real numbers and $i^2 = -1$. The numbers can be represented by points in the Cartesian coordinate plane by identifying this number with (a, b).

Addition agrees with coordinatewise addition of points in the plane: $(a+ib)+(c+id) = (a+c) + i(b+d)$. Multiplication is $(a+ib)(c+id) = ac + i(ad+bc) + i^2bd = (ac - bd) + i(ad + bc)$.

The complex conjugate $\overline{z}$ of $z = a + ib$ is $a - ib$. The product $z\overline{z}$ is a nonnegative real number: $z\overline{z} = (a+ib)(a-ib) = a^2 + b^2$. The square root of this number is the modulus, the distance to 0: $|z| = |a + ib| = \sqrt{a^2 + b^2}$.

Since the product $i(a + ib) = -b + ia$, the function that maps a complex number z to iz is the rotation O_{90} of the plane, since the formulas are the same.

More generally, if $|w| = 1$, then the complex number w is on the unit circle and has the form $w = u(\theta) = \cos\theta + i\sin\theta$ for some θ. (For the rest of this section, the temporary notation $u(\theta)$ will denote this complex number with modulus 1.)

For $z = (x + iy)$, the formula for this product is the formula for $O_\theta(z)$:

$$u(\theta)(x+iy) = (\cos\theta + i\sin\theta)(x+iy) = x(\cos\theta + i\sin\theta) + y(-\sin\theta + i\cos\theta).$$

More generally, the function $S_\alpha(z) = \alpha z$, where $\alpha = a + ib$ is a constant nonzero complex number, is a transformation of the plane. The inverse of the transformation is multiplication by $\alpha^{-1} = \overline{\alpha}/|\alpha|^2$.

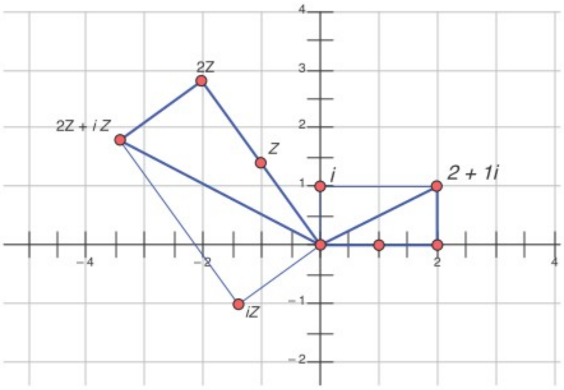

Figure 12. Geometry of Complex Product $(2 + 1i)z$

Any nonzero complex number $\alpha = |\alpha|(\alpha/|\alpha|)$, where $(\alpha/|\alpha|) = u(\theta)$ has modulus 1. Multiplication by the real number $|\alpha|$ is dilation with center O and dilation ratio $|\alpha|$. Therefore, the transformation $S_\alpha(z)$ can be factored into $\mathcal{D}_{O,|\alpha|}O_\theta$, the product of a dilation and a rotation. This is an orientation-preserving similitude.

The example of complex multiplication in Figure 12 can be viewed two ways. It is a dilation rotation by z of $2 + 1i$, scaling the coordinate rectangle of $2 + 1i$ by $|z|$ and rotating by $z/|z|$. One can also view this as a dilation rotation of z by $2 + 1i$ as a construction, starting with a triangle $\triangle 0\,2\,(2+1i)$ and then building a similar triangle with base segment $0\,2z$ in place of segment $0\,2$ as the base, then adding a perpendicular segment $0\,iz$ of length $|z|$ instead of the side of length 1 in the original triangle. This second interpretation constructs a rectangle similar to the coordinate rectangle of $2 + 1i$ and shows how the angle of rotation and the length of $(2 + 1i)z$ emerge from the geometry.

Translations have the form $T(z) = z + \beta$, for some complex number β. Therefore, $U(z) = \alpha z + \beta$ is an orientation-preserving similitude and is a rigid motion if $|\alpha| = 1$. In fact, every such similitude and rigid motion can be written in this form. A proof will be explored in the exercises.

The complex conjugate map $C(z) = \overline{z}$ is reflection in the x-axis. Then $S_\alpha C S_\alpha^{-1}$ is reflection in line $\overline{0\alpha}$. The formula for this map is

$$S_\alpha C S_\alpha^{-1}(z) = S_\alpha C(\overline{\alpha} z/|\alpha|^2) = S_\alpha(\alpha \overline{z}/|\alpha|^2) = (\alpha/|\alpha|)^2 \overline{z} = u(\theta)^2 \overline{z} = u(2\theta)\overline{z}.$$

This agrees with the formula derived in the previous section. The composition of this map with translations provides a complex expression for any line reflection.

The flexibility provided by the algebra of complex numbers makes some geometrical theorems easier to prove.

Theorem 11.8 (Similitude Fixed Point)**.** *Every orientation-preserving similitude that is not a translation has a unique fixed point.*

Proof. Such a transformation has the form $U(z) = \alpha z + \beta$, with $\alpha \neq 1$. A fixed point z satisfies $z = \alpha z + \beta$. A unique solution of this equation is found by elementary algebra: $z_0 = \beta/(1-\alpha)$. ☐

One can check that if $|\alpha| < 1$, then for any z, the sequence of images $U^k(z)$ approaches z_0 as a limit. This is illustrated by the spiral in Figure 8.40.

Functions of complex numbers can also be used to describe additional geometrical transformations of the plane. For example, $J(z) = 1/\overline{z}$ is a transformation called reflection in the unit circle; this fixes the points on the unit circle but exchanges the interior with the exterior, much as line reflection does. Inversion leads to some beautiful geometry of circles that is beyond the scope of this book. Subsets of complex numbers are also used to model non-Euclidean plane geometry. References include Hahn [**10**].

Barycentric Coordinates

In this section we will introduce barycentric coordinates, which locate points as barycenters, or centers of mass.

In Theorem 11.5, we saw that any point on $\overline{PQ}$ can be written as $(1-t)P + tQ$ for some real t. This expression represents the same point whatever the coordinate system.

Other expressions in the coordinates of P and Q will not define points on the plane in a way that is independent of the coordinate system. For example, the expression $P+Q$ denotes the point N such that $POQN$ is a parallelogram. If the origin O is changed, so is N. In contrast, $(1-t)P + tQ$ always denotes the point A on $\overline{PQ}$ with $AP/QP = t$.

Center of Mass of a Segment. A more symmetric way of writing the parametric expression for a point G on a line P_1P_2 is to say that $G = c_1P_1 + c_2P_2$, where $c_1 + c_2 = 1$.

The point G is the center of mass, the **barycenter**, of a system consisting of a mass c_1 at P_1 and c_2 at P_2. This follows from the law of levers of Archimedes: at the barycenter, the products of mass times lever arm are equal and opposite. In this case, $c_1(P_1 - G) = -c_2(P_2 - G)$.

If there are masses m_1 and m_2 at these points that do not sum to 1, there is still a center of mass. Let $M = m_1 + m_2$; then set $G = (m_1/M)P_1 + (m_2/M)P_2$.

For example, if one puts 2 grams at P_1 and 5 grams at P_2, then the center of mass is $(2/7)P_1 + (5/7)P_2$; this is the point on the segment with $GP_1/GP_2 = -5/2$.

When speaking of masses, the numbers c_i are nonnegative and the point G will belong to the segment P_1P_2. In the formula for G, negative real numbers are allowed, with G possibly located at any point of $\overline{P_1P_2}$. But the physical analogy must be extended to springs or balloons that pull upward, opposite from gravity.

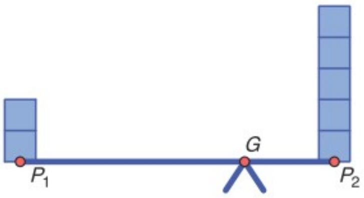

Figure 13. Barycenter of Masses 2 and 5

Affine Combinations and Barycenters. We will extend this idea of a center of mass to a system of more than two points. The next theorem will set the stage.

Theorem 11.9 (Affine Combinations). *Concerning* $c_1P_1 + c_2P_2 + \cdots + c_nP_n$*:*

- *An expression* $c_1P_1 + c_2P_2 + \cdots + c_nP_n$ *is called an* ***affine combination*** *of points if* $c_1 + c_2 + \cdots + c_n = 1$*. Such an expression equals a point* G *independent of the coordinate system.*
- *If the sum* $c_1 + c_2 + \cdots + c_n = 0$*, the expression* $c_1P_1 + c_2P_2 + \cdots + c_nP_n$ *equals a vector independent of the coordinate system.*
- *If* $c_1 + c_2 + \cdots + c_n = 1$*, the vector* $c_1(P_1 - G) + c_2(P_2 - G) + \cdots + c_n(P_n - G)$ *is the zero vector. This is the sum of the masses times the lever arm at* G*.*

Proof. Here is a proof for three points, $G = c_1P_1 + c_2P_2 + c_3P_3$, where $c_1 + c_2 + c_3 = 1$.

If $c_1 + c_2 = 0$, then $c_1(P_1 - P_2) = v$ is a vector and $G = \mathcal{T}_v(P_3)$ is a point independent of coordinates. This completes the proof in this case.

Now assuming that $c_1 + c_2 \neq 0$, let $Q = 1/(c_1 + c_2)(c_1P_1 + c_2P_2)$, a point on $\overline{P_1P_2}$. Then $G = (1 - c_3)Q + c_3P_3$ since $c_1 + c_2 = 1 - c_3$; this G is a point on $\overline{QP_3}$. Both Q and G are independent of the coordinate system, by Theorem 11.5.

Next, if $c_1 + c_2 + c_3 = 0$, let $v = c_1P_1 + c_2P_2 + c_3P_3$ be computed by adding coordinates in some coordinate system. Let $E = P_1$. The point $F = E + v = (1 + c_1)P_1 + c_2P_2 + c_3P_3$ is a well-defined point independent of any coordinate system, since $1 + c_1 + c_2 + c_3 = 1$. Therefore, $v = [EF]$ is a vector independent of the coordinate system.

Lastly, if $c_1 + c_2 + c_3 = 1$, the vector $c_1(P_1 - G) + c_2(P_2 - G) + c_3(P_3 - G) = c_1P_1 + c_2P_2 + c_3P_3 - (c_1 + c_2 + c_3)G = G - G = 0$.

This proves the theorem for $n = 3$ by using the already-known result for $n = 2$. This reasoning can be continued inductively, step by step in the same manner, for any number of summands, but we will leave the details to the reader. In what follows, the case $n = 3$ is the case we will need. ☐

Example 11.10. For a parallelogram $ABCD$, the center of symmetry M can be written in an infinite number of ways: for example, $(1/2)(A + C)$, $(1/4)(A + B + C + D)$, or $(1/2)(2A - B + 2C - D)$.

In Theorem 8.50 we learned that the medians of a triangle are concurrent at a point called the centroid of the triangle. The centroid of $\triangle ABC$ is $G = (1/3)(A + B + C)$. To see this, let $M = (1/2)(A + B)$ be the midpoint of AB. Then CM is a median, and

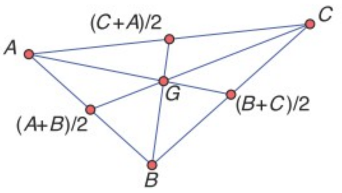

Figure 14. Centroid $G = (1/3)(A + B + C)$

$G = (2/3)M + (1/3)C$ is the point on the median CM with $GM/CM = 1/3$. By the same reasoning, this G is on any median, so the medians are concurrent at G. Unlike the example of the parallelogram, there is no other affine combination of the vertices that equals G. Why this is unique will be shown in Theorem 11.11.

Barycenter of Masses on a Triangle. Given a triangle $\triangle P_1P_2P_3$, a triple of numbers (c_1, c_2, c_3), with $c_1 + c_2 + c_3 = 1$, defines a point $G = c_1P_1 + c_2P_2 + c_3P_3$.

This points can be interpreted as the center of mass for three masses located at P_i when the $c_i \geq 0$. This follows from the third point in Theorem 11.9 by the lever principle of Archimedes. As a physical example, imagine supporting and balancing a rigid triangle on the edge of a ruler that passes under P_3. With a mass at each vertex, to balance the remaining two masses, the ruler should pass under the center of mass Q of the side P_1P_2. Then to balance the whole structure on a fingertip at a point of QP_3, one can conceptually replace the segment P_1P_2 with masses at each endpoint by its center of mass Q with mass $c_1 + c_2$. This is the reasoning used in the proof of Theorem 11.9 to find the balance point at G. This is the same balance principle used in a hanging mobile, where every pair of objects is hung from a stick attached by a thread at the center of mass of the two objects.

Theorem 11.11 (Barycentric Coordinates)**.** *Given a triangle $\triangle P_1P_2P_3$, every point P of the plane can be expressed in exactly one way as $P = c_1P_1 + c_2P_2 + c_3P_3$, where $c_1 + c_2 + c_3 = 1$. This unique triple of numbers can be used as coordinates for P. These are called the* ***normalized barycentric coordinates*** *of P with respect to this triangle.*

Proof. This is actually not new; these are affine coordinates in disguise. If we set $P_3 = O$, then

$$c_1P_1 + c_2P_2 + c_3P_3 = (1 - c_1 - c_2)O + c_1P_1 + c_2P_2).$$

Then (c_1, c_2) are the affine coordinates of P with respect to $\triangle OP_1P_2$. ☐

One may wonder what is the point of repackaging affine coordinates. We will see that treating the three vertices of the coordinate triangle symmetrically has interesting consequences.

In the case of masses at the vertices P_i where the nonzero sum of the masses is not 1, the masses (m_1, m_2, m_3) still determine a point. Let $M = m_1 + m_2 + m_3$ and let $c_i = m_i/M$. Then

$$G = (1/M)(m_1P_1 + m_2P_2 + m_3P_3)$$

is a point that is a center of mass when the $m_i \geq 0$. The numbers (m_1, m_2, m_3) are **barycentric coordinates** of G. Unlike (x, y)-coordinates, barycentric coordinates are homogeneous. The number triples (km_1, km_2, km_3) define barycentric coordinates that are considered the same for any $k \neq 0$, and they determine the same point. The one restriction is that $M = m_1 + m_2 + m_3 \neq 0$.

Example 11.12. On the left side of Figure 15 there is shown a $\triangle ABC$ with masses 1, 2, and 4, at the respective vertices. So $(1, 2, 4)$ are barycentric coordinates for a point, with $M = 7$. The barycenters of each pair of masses on the sides are

$$D = (2/6)B + (4/6)C,$$
$$E = (4/5)C + (1/5)A,$$
$$F = (1/3)A + (2/3)B$$

with the barycenter for three masses $G = (1/7)A + (2/7)B + (4/7)C$.

The ratios on the sides are $DB/DC = -2$, $EC/EA = -1/4$, $FA/FB = -2$.

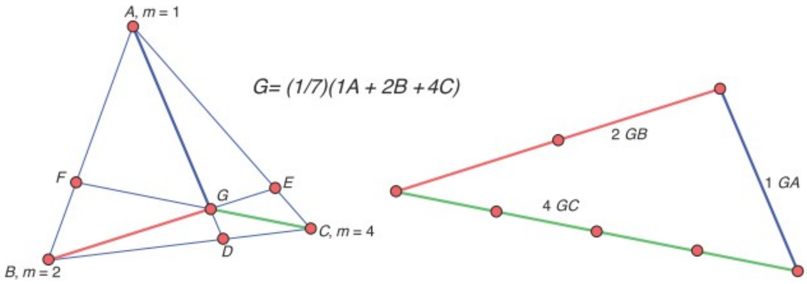

Figure 15. Barycenter and Vector Sum

Segments representing the three vectors from G to the vertices are shown in different colors. On the right side of the figure there is plotted the sum of the three vector lever arms using the same colors: $1(A - G) + 2(B - G) + 4(C - G)$, which is 0 as it should be, since $0 = c_1(P_1 - G) + c_2(P_2 - G) + c_3(P_3 - G)$.

We can see in this example a way to show that the barycenter is inside the triangle if the masses are positive. On the segments AD, BE, and CF, the negative signed ratios $GD/GA = -1/6$, $GE/GB = -2/5$, and $GF/GC = -4/3$ show that G is an interior point of each segment.

Cevians and Ceva's Theorems. In Example 11.12, the segments AD, BE, and CF are called cevians. The signed ratios of the endpoint of the cevians on the sides satisfied this equation: $(DB/DC)(EC/EA)(FA/FB) = (-2)(-1/4)(-2) = -1$. We will now look at cevians and these ratios in general.

Definition 11.13. A segment joining any vertex of $\triangle ABC$ to a point of the (extended) opposite side is called a **cevian**. A cevian line is a line containing a cevian.

For a point G with normalized barycentric coordinates (c_1, c_2, c_3) with respect to $\triangle ABC$, the line $\overline{AG}$ will intersect $\overline{BC}$ at a point $D = (1 - t)A + tG$ where $c_1 = 0$.

This means that the barycentric coordinates of D will be $(0, c_2, c_3)$ and the normalized barycentric coordinates will be $(0, c_2/(1-c_1), c_3/(1-c_1))$. The point of intersection will not exist, and $\overline{AG}$ will be parallel to $\overline{BC}$ when $c_1 = 1$. The signed ratio DB/DC will equal $-c_3/c_2$ except when $c_2 = 0$ and the ratio is not defined.

Now let point G have coordinates (c_1, c_2, c_3) with respect to $\triangle P_1P_2P_3$ for which none of the c_i equals 0 or 1. Then the triangle has three cevian lines $\overline{P_iQ_i}$ concurrent at G and with this relationship among the ratios on the sides:

$$\frac{Q_1P_2}{Q_1P_3}\,\frac{Q_2P_3}{Q_2P_1}\,\frac{Q_3P_1}{Q_3P_2} = \frac{-c_3}{c_2}\,\frac{-c_1}{c_3}\,\frac{-c_2}{c_1} = -1.$$

If all the $c_i > 0$, then the cevian segments concur at G at a point inside the triangle.

The converse is an important triangle concurrence theorem. To avoid the special cases, the theorem is stated for cevian segments.

Theorem 11.14 (Ceva's Theorem). *In $\triangle P_1P_2P_3$, let the three cevians P_1Q_1, P_2Q_2, P_3Q_3 have endpoints Q_1, Q_2, Q_3 on the sides of the triangle, not at a vertex. Then the cevians are concurrent at a point G if and only if*

$$\frac{Q_1P_2}{Q_1P_3}\,\frac{Q_2P_3}{Q_2P_1}\,\frac{Q_3P_1}{Q_3P_2} = -1.$$

Proof. The"only if" claim is all that is left to prove, given what has already been proved,

The first step of the proof is to use two ratios on the sides, r_2 and r_3, to find the barycentric coordinates of G, the point of concurrence of two of the cevians P_2Q_2 and P_3Q_3.

Assume that there is mass $m_1 = 1$ at P_1; then $r_2 = Q_2P_3/Q_2P_1 = -m_1/m_3$, so $m_3 = -1/r_2$. Also, $r_3 = Q_3P_1/Q_3P_2 = -m_2/m_1$, so $m_2 = -r_3$. Therefore, $(1, -r_3, -1/r_2)$ are barycentric coordinates for the intersection point.

Then let $M = 1 - (1/r_2) - r_3$. Since $M > 1$, we can set $c_1 = 1/M$, $c_2 = -r_3/M$, $c_3 = -1/(r_2M)$ as the normalized barycentric coordinates for a point G that is the intersection of the cevians through P_2 and P_3.

The last part of the proof is to show that the intersection J of $\overline{P_1G}$ with $\overline{P_2P_3}$ is the point Q_1. Since we have the barycentric coordinates of G, we know J is the barycenter of P_2P_3 with mass $-r_3$ at P_2 and $-1/r_2$ at P_3. The point J divides the segment in the ratio $JP_2/JP_3 = -(1/r_3)/r_2 = -1/r_2r_3$. Since $r_1r_2r_3 = -1$, this ratio is r_1 and $J = Q_1$. $\square$

Note. This proof also applies to cases when two of the Q_i are on extended sides of the triangle, but for some cases the $\overline{P_iQ_i}$ will be parallel and not concurrent.

We have encountered a number of special lines in triangles. Perpendicular bisectors are not cevians. But medians, angle bisectors, and altitudes are all cevians, and this theorem applies to them.

Figure 16 shows another barycentric coordinate example with one coordinate negative. The point S has barycentric coordinates $(4, -1, 3)$, or $(2/3, -1/6, 1/2)$, relative to $\triangle ABC$. Since the second coordinate is negative, the vertices B and S are on opposite sides of $\overline{AC}$. The segment ratios on the sides and their product are shown in the figure.

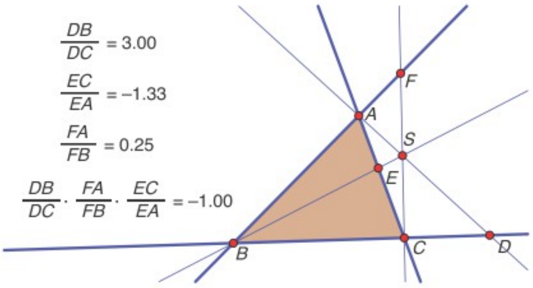

Figure 16. Cevian Lines and Ratios of Mixed Sign

Note. The Menelaus Theorem, Theorem 8.29, asserts that when the product of ratios is positive instead of negative, the three points Q_i are collinear. This provides a straightedge construction of harmonic division, as shown in Figure 8.26.

Areal Coordinates. The normalized barycentric coordinates of points in a triangle are sometimes called **areal coordinates**.

The reason for this can be seen in the example in Figure 15. In this figure, the ratio $DG/DA = 1/7$. This means that the height of $\triangle GBC$ is 1/7 the height of $\triangle ABC$. Since the base of each triangle is BC, the area $\mathcal{A}(\triangle GBC) = (1/7)\mathcal{A}(\triangle ABC)$.

By the same reasoning,

$$\mathcal{A}(\triangle GCA) = (2/7)\mathcal{A}(\triangle ABC) \quad \text{and} \quad \mathcal{A}(\triangle GAB) = (4/7)\mathcal{A}(\triangle ABC).$$

Thus, the three numbers $(1/7, 2/7, /4/7)$ are the normalized barycentric coordinates but also represent the fractions of the total area $\mathcal{A}(\triangle ABC)$ that are the areas of the three subtriangles with vertex G. And the sum of these coordinates is 1, since the sum of the three triangle areas is $\mathcal{A}(\triangle ABC)$.

In Figure 16 the area ratios can be computed in the same way. Let $\mathbb{A} = \mathcal{A}(\triangle ABC)$. The area ratios are $\mathcal{A}(\triangle SBC)/\mathbb{A} = 2/3$, $\mathcal{A}(\triangle SCA)/\mathbb{A} = 1/6$, and $\mathcal{A}(\triangle SAB)/\mathbb{A} = 1/2$.

This time $2/3+1/6+1/2 \neq 1$ since the three triangles do not subdivide the original triangle. Instead, the sum of two of their areas is the area of the quadrilateral $ABCS$.

However, the sum of the ratios will still be 1 if $\mathcal{A}(\triangle SCA)$ is considered negative. This makes sense since the orientation of the triangle is opposite that of all the other triangles. In the figure one sees that removing this area from the area of the quadrilateral $ABCS$ does equal the area of $\triangle ABC$. And the calculation shows this also: $2/3 - 1/6 + 1/2 = 6/6 = 1$.

This idea of **signed area**, of setting area to be negative if a figure has negative orientation, is an idea that appears in many guises in geometry.

Example 11.15. For $\triangle ABC$ with angle measures α, β, and γ at the vertices A, B, and C, the barycentric coordinates of the circumcenter can be deduced from the areas of $\triangle OAB$, $\triangle OBC$, and $\triangle OBC$ as shown in Figure 17. For instance, $\triangle OAB$ is an isosceles triangle with angle measure 2γ at O, since $\angle BAC$ is an inscribed angle. The equal sides of the triangle have length r, the radius of the circle. The area of this isosceles triangle is twice the area of the right triangle $\triangle AOF$. The legs of this triangle are $r\sin\gamma$ and $r\cos\gamma$, so the area is $r^2\sin\gamma\cos\gamma = (r^2/2)\sin 2\gamma$. Dividing each area by r^2, the barycentric coordinates of O are $(\sin 2\alpha, \sin 2\beta, \sin 2\gamma)$. This formula remains true when one of the angles is obtuse, though the area of the corresponding isosceles triangle will be negative. To get the areal or normalized barycentric coordinates, one must either divide by the sum of these numbers or use the area of $\triangle ABC$.

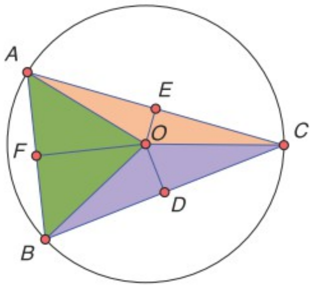

Figure 17. The Arial Coordinates of the Circumcenter

Intersections from Barycentric Coordinates. Confronted by points E and F on the sides of a triangle, as in Figure 18, one can use barycentric coordinates to find the coordinates of the intersection G. If we know that $AE/AB = 1/5$ and $AF/AC = 2/5$, then each point can be viewed as a barycenter on a side. The point E is the barycenter of a mass $4m$ at A and $1m$ at B, while F is the barycenter of mass $3n$ at A and $2n$ at C. If we set $m = 3$ and $n = 4$, we get barycentric coordinates $(12, 3, 8)$ with respect to $\triangle ABC$. Now it is also immediate to see that the intersection of $\overline{AG}$ with BC has coordinates $(0, 3, 8)$.

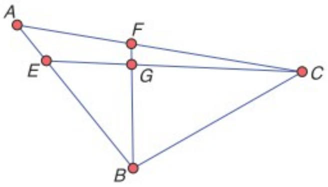

Figure 18. Barycentric Coordinates of G from Ratios AE/AB and AF/AC

Of course this is not the only way to find these coordinates. A very sure way is to choose affine coordinates, then find the equations for the lines $\overline{BF}$ and $\overline{CE}$, and then solve the pair of equations. But for some simple problems the method with masses is more direct and is perhaps more insightful.

This method can be used to find the barycentric coordinates of all the points in the Marion Walters figure, Figure 9.21. Then the arial coordinates can be used to find areas of regions in the figure to solve the problem. That figure exhibits affine symmetry in that many points have symmetric coordinates because the coordinate and area ratios are the same as they would be for the figure in an equilateral triangle.

Another application of barycentric coordinates is to the intersection of sides of a triangle with a line parallel to a side. In Figure 19, we begin with the point P_1 on AB, then intersect AC with the line through P_1 parallel to BC to get P_2. Then we repeat the same procedure to get point P_3 on BC and then P_4 on AB. The point P_4 will not be the same as P_1 (unless P_1 is the midpoint), but if we continue, the next time the point that should be P_7 is equal to P_1.

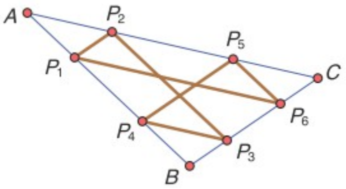

Figure 19. Repeating Parallel Paths in a Triangle

It is not too difficult to see why this works without using barycentric coordinates, but it is very clear with them. If the barycentric coordinates of P_1 with respect to $\triangle ABC$ are $(p, q, 0)$, then the parallel intersection at P_2 has coordinates $(p, 0, q)$, since the mass at A does not change and the mass at B moves to C. Continuing, we get the repeating sequence

$$(p, q, 0), (p, 0, q), (0, p, q), (q, p, 0), (q, 0, p), (0, q, p), (p, q, 0).$$

This repeated pattern also holds, with the same coordinate reasoning, for a point P_1 on the extended side of the triangle, though the figure will now be a polygon. If one does the construction with dynamic geometry software, it can be animated,

Vectors and Affine Transformations

In this section, we will use vectors to study some transformations more general than similitudes. We begin with a brief recapitulation of what we have said about vectors so far, beginning with the definition and notation in Chapter 7.

Vectors and Coordinates. Vectors appear as objects with direction and magnitude in many areas of geometry. As we have seen, in the Euclidean plane there is a one-to-one correspondence between the set of vectors and the set of translations: an ordered segment AB represents both the vector $[AB]$ and the translation $\mathcal{T}_{AB}$. The coordinate formula for $\mathcal{T}_{AB}(P)$ is $P+(B-A)$, so $[AB]$ can also be represented by the number pair $B - A$.

Both the vector and the translation can be represented visually by arrows with tail at A and head at B. But if CD is another ordered segment so that $ABDC$ is a parallelogram or linear parallelogram, then the arrow from C to D also represents the same vector and translation: $[AB] = [CD]$.

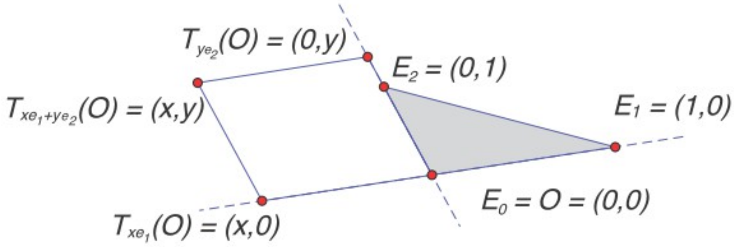

Figure 20. Coordinates Expressed Using Vectors or Translations

Addition of vectors corresponds to composition of transformations, so $u = v + w$ when $\mathcal{T}_u = \mathcal{T}_v\mathcal{T}_w$. In coordinates, this is just addition of number pairs: $(B - A) + (D - C)$. Geometrically, starting at any point A, let $B = \mathcal{T}_v(A)$ and $D = \mathcal{T}_w(A)$; then $u = v + w = [AC]$, where C is the unique point so that $ABCD$ is a parallelogram or linear parallelogram.

Multiplication by a real number k, e.g., $k[AB]$, corresponds to translation $\mathcal{T}_{AC}$, where $C = \mathcal{D}_{A,k}(B)$.

For a linear coordinate system based on $\triangle E_0E_1E_2$, let vectors $e_1 = [E_0E_1]$ and $e_2 = [E_0E_2]$. For a point P with coordinates (x, y), $P = E_0 + x(E_1 - E_0) + y(E_2 - E_0)$. Or in vector terms, $[E_0P] = xe_1 + ye_2$. The points within $\triangle E_0E_1E_2$ have $0 \leq x + y \leq 1$. This is the shaded triangle in Figure 20.

Shear Transformations. Shear transformations were defined in Chapter 9 and depicted in Figure 9.20. Let us consider a shear transformation S related to the linear coordinate system defined by $\triangle E_0E_1E_2$.

The base will stay fixed, so $S(E_0) = E_0$ and $S(E_1) = E_1$, but $S(E_2) = E_2'$, where E_2E_2' is parallel to E_0E_1. In other words, $[E_2E_2'] = ke_1$ for some real k, so $[E_0E_2'] = ke_1 + e_2$.

This transformation of the triangle vertices determines the images of the other points in the plane:

$$S(E_0 + xe_1 + ye_2) = E_0 + xe_1 + y(ke_1 + e_2).$$

This says $S(x, y) = (x + ky, y)$ is the coordinate formula for S.

Each line parallel to the x-axis with equation $y = c$ is translated to itself by the vector cke_1. And any point $(0, y) = E_0 + ye_2$ on the y-axis maps to the point $E_0 + y(ke_1 + e_2) = (ky, y)$, as prescribed in Chapter 9.

One way to picture this is to imagine the coordinate tessellation of the plane by $\triangle E_0E_1E_2$ and triangles congruent to it, with vertices being all the points with integer coordinates. Then all the triangles are transformed consistently so that the tessellation of the plane is mapped to the coordinate tessellation defined by $\triangle E_0E_1E_2'$. In other words, a point with affine coordinates in the $\triangle E_0E_1E_2$ system is mapped to the point with the same coordinates in the $\triangle E_0E_1E_2'$ system.

One important property of this shear transformation is that all areas are preserved. **The area of a figure and its sheared image are the same.** If one tessellates the

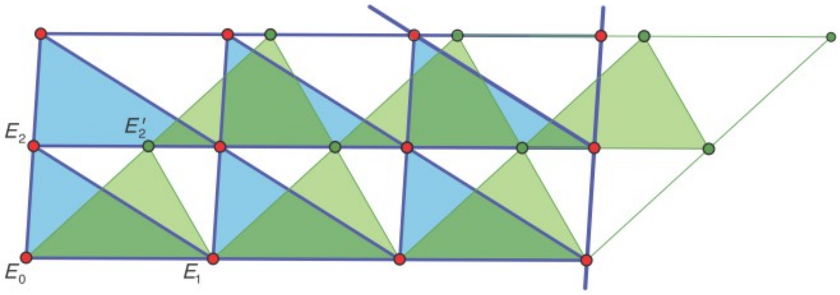

Figure 21. Tessellation by Triangles and Sheared Images

plane with triangles congruent to $\triangle E_0E_1E_2$, then this tessellation is transformed to a tessellation by triangles congruent to $S(\triangle E_0E_1E_2)$ and so each has the same area. Moreover, the plane can be tessellated by arbitrarily small triangles similar to $\triangle E_0E_1E_2$ by taking midpoint triangles and then midpoint triangles of those to get triangles smaller than any pre-determined size. Thus, for a figure like a circle or a regular pentagon, the area can be approximated by a sum of areas of small triangles; then the area of the image is approximated in the same way by the image triangles that have the same areas as the originals.

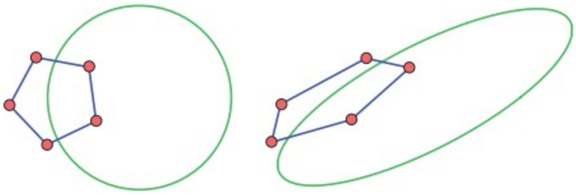

Figure 22. Sheared and Translated Circle and Regular Pentagon

The sheared images in Figure 22 have also been translated, not changing the area but making them easier to see. The image of the circle (which can be shown to be an ellipse) has the same area as the circle. And the irregular pentagon has the same area as the regular pentagon.

Shear transformations in other directions can be defined with respect to any other triangles. And more general transformations can be defined as products of various shears and rigid motions. These products will also preserve area.

Affine Transformations. As products of shears and rigid motions and other similitudes are formed, the analysis of their behavior will be more complex, but the products will always map collinear points to collinear points and preserve signed ratio on each line. This property will be used as a definition.

Definition 11.16. An **affine transformation** of the plane is a transformation that maps lines to lines and preserves signed ratio.

All rigid motions and other similitudes are affine transformations. It is not hard to show that products and inverses of affine transformations are also affine transformations.

Theorem 11.17 (Affine Transformation Formula). *Given affine coordinates, any transformation T that has a coordinate formula $T(x, y) = (ax + by, cx + dy) + (e, f)$ is an affine transformation, and conversely.*

Note: T is assumed to be a transformation. A function with $a = b = c = d = 1$ will not have an inverse. The necessary condition on a, b, c, d for T to be a transformation will be considered below.

Proof. It is sufficient to show that the image of a point $sP + tQ$ has the same form, where $P = (p_1, p_2)$, $Q = (q_1, q_2)$, and $s + t = 1$.

$T(sP+tQ) = (a(sp_1+tq_1)+b(sp_2+tq_2), c(sp_1+tq_1)+d(sp_2+tq_2))+s(e, f)+t(e, f)$, for $(e, f) = s(e, f) + t(e, f)$ since $s + t = 1$.

Collecting like terms, this equals $sT(P) + tT(Q)$.

To see the converse, let T be an affine transformation and let E_0, E_1, E_2 be the vertices of the triangle defining the affine coordinates. If the coordinates of the images of these points are (e, f), $(a + e, c + f)$, and $(b + e, d + f)$, then any point $P = (1 - x - y)E_0 + xE_1 + yE_2$ with coordinates (x, y) has T-image with coordinates $(1 - x - y)(e, f) + x(a, c) + x(e, f) + y(b, d) + y(e, f) = (ax + by, cx + dy) + (e, f)$. ☐

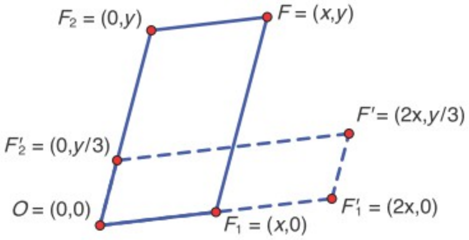

Figure 23. Compression-Expansion Scaling $(2x, y/3)$

One example of such an affine transformation is a **scaling transformation** $T(x, y) = (ax, by)$, where $ab \neq 0$. This differs from a dilation in that the scaling in the x and y directions is not the same. In directions when a or b is greater than 1, this is an expansion in the direction of e_1 or e_2. When less than one, it is compression.

Linear Transformations. Affine transformations induce transformations on vectors. If we write PQ in coordinates as $Q - P = (v_1, v_2)$, then applying T yields $T(Q) - T(P) = (av_1 + bv_2, cv_1 + dv_2)$. The coefficients a, b, c, d are the same, but the translation part of the formula disappears. This is consistent with the property of vectors that they have no fixed location.

In fact, as we saw when we derived the rotation formula, it may be clearer to work with transformations that map the origin O to itself and then add translations later.

Definition 11.18. A **linear transformation** is an affine transformation that has the origin O as a fixed point.

Thus, any linear transformation has the form $T(x, y) = (ax + by, cx + dy)$.

In a note above, it was observed that while we have shown that any linear transformation has this form, for some coefficients this formula is not a transformation. We will explore this by looking at the image of $\triangle E_0E_1E_2$.

We see that $T(E_0) = (0, 0)$, $T(E_1) = (a, c)$, and $T(E_2) = (b, d)$. We wish to find out more about this set of points by applying a couple of shear transformations. This is illustrated by Figure 24, with $(a, c) = (2, -1)$ and $(b, d) = (1, 1)$.

We will assume that $a \neq 0$. If this is not true, we can interchange the role of E_1 and E_2 and then assume $b \neq 0$. If both $a = 0$ and $b = 0$, then the points all are on the y-axis $\overline{E_0E_2}$, so this function T is not a transformation, since transformations map noncollinear point to noncollinear points.

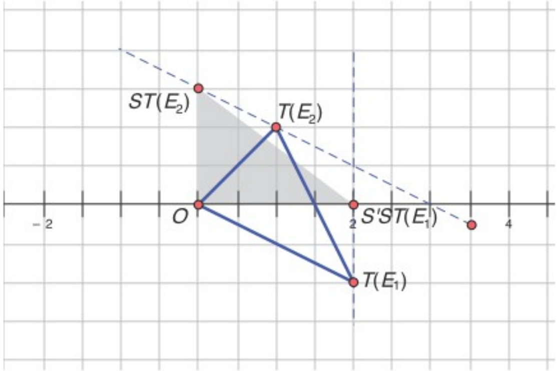

Figure 24. Shearing the Image of the Coordinate Triangle

Now assuming $a \neq 0$, let S be the shear parallel to $\overline{E_0T(E_1)}$ such that S-image of $T(E_2)$ will be on the line through $T(E_2)$ parallel to $\overline{E_0T(E_1)}$. Chose the shear S to be the shear that takes $T(E_2)$ to a point on the y-axis. This point will have the form $S(b, d) = (b, d) + k(a, c) = (0, t)$, so $k = -b/a$. This means that the image

$$ST(E_2) = S(b, d) = (b, d) - (b/a)(a, c) = (0, (ad - bc)/a).$$

In Figure 24, this is the point $ST(E_2) = (1, 1) + (-1/2)(2, -1) = (0, 3/2)$.

There is one exception: if this line through (b, d) intersects the y-axis at $(0, 0)$, then $ad - bc = 0$, the shear is not defined, and T is not a transformation since the images of E_0, E_1, E_2 are collinear. Henceforth, we assume $ad - bc \neq 0$. In the example in Figure 24, this quantity equals 3.

Next applying a shear S' parallel to the y-axis that will map $T(E_1)$ to a point on the x-axis will not change the points on the y-axis but will have $S'ST(E_1) = (a, 0)$. In Figure 24 this point is $(2, 0)$.

Let

$$H(x, y) = \left(\frac{x}{a}, \frac{ay}{ad - bc}\right).$$

Then $HS'ST = I$, the identity map, since it fixes the vertices of $\triangle E_0E_1E_2$. This means T has an inverse $S^{-1}S'^{-1}H^{-1}$ and so is a transformation when $ad - bc \neq 0$.

The Determinant of an Affine Transformation. The coefficients in the formula $T(x, y) = (ax + by, cx + dy)$ are commonly presented in the form of a matrix:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

The quantity $ad - bc$ is called the **determinant of the matrix** or, in this case, the determinant of the transformation T. The determinant of an affine transformation is defined by the same formula as well; we will interpret this quantity in terms of area.

Continuing with the notation above, the $SS'T$ image of $\triangle E_0E_1E_2$ is a triangle with base a and height $|ad - bc|/a$, with area $(1/2)|ad - bc|$. Thus, the ratio of the area of the image to the original area is $|ad - bc|$, the absolute value of the determinant. By reasoning similar to that which showed shears do not change area, one can conclude that **an affine transformation T scales the area of any shape by the factor $|ad - bc|$.**

The sign of the determinant also has meaning. Since shears are orientation-preserving transformations of the plane, the product $SS'T$ is orientation-preserving if and only if T is orientation-preserving and the triangles $T(\triangle E_0E_1E_2)$ and $SS'T(\triangle E_0E_1E_2)$ have the same orientation. But the latter triangle has vertices $(0, 0)$, $(a, 0)$, and $(0, (ad - bc)/a)$.

For a triangle ABC with $A = O$, $B = (x_0, 0)$, and $C = (0, y_0)$, the triangle has positive orientation when $x_0y_0 > 0$. In the case of T, this means that T is orientation-preserving if $(ad - bc) > 0$ and it is orientation-reversing otherwise.

If we take unit area to be the area of $\triangle E_0E_1E_2$, then $T(\triangle E_0E_1E_2)$ has signed area equal to the determinant $ad - bc$. The absolute value $|ad - bc| = \mathcal{A}(T(\triangle E_0E_1E_2))/\mathcal{A}(\triangle E_0E_1E_2)$ while the sign indicates whether the two triangles have the same orientation. If one thinks in terms of the unit square, the determinant is the signed area of the parallelogram that is the image of the square.

Axioms and Models

To end the book, we would like to complete the circle that began by assuming a set of six axioms for Euclidean geometry and conclude by proving that there is a model for this geometry based on the real numbers.

When one creates a set of axioms for a mathematical structure, one has to be concerned about whether there actually exist any examples of the structure, or whether from contradictions or excessive restrictions in the axioms one is proving theorems about nothing. An example of this genuine concern was the early period of discovery in non-Euclidean plane geometry. Mathematicians such as Lobachevsky assumed that there could be more than one parallel to a line through a given point and proved beautiful and surprising theorems. But some mathematicians were very concerned that this purported non-Euclidean plane had some as yet undiscovered contradiction that meant it did not really exist, even as a mathematical abstraction. It was not until some models for the non-Euclidean plane were constructed by Poincaré and Klein that the geometry was accepted to be on a sound footing.

While the soundness of the concept of the Euclidean plane is not in question here, it is a satisfying exercise to see that one can define a mathematical model for the plane and prove our six axioms as theorems.

Theorem 11.19 (Plane Model $\mathcal{E}$ for the Six Axioms). *Let a model $\mathcal{E}$ for plane geometry be given by taking the points of $\mathbb{R}^2$ as the points of the plane and with lines, distance, and angle measure as specified in the following definitions. Then this is a model for Euclidean geometry in which the six axioms of Chapter 2 are valid.*

The points of $\mathcal{E}$ are number pairs such as $A = (a_1, a_2)$. We will add such points and multiply them by real numbers in our usual manner. We will also rely on computations with coordinates from earlier in the chapter to avoid needless repetition.

The First Three Axioms.

Definition 11.20 (Lines in $\mathcal{E}$). An $\mathcal{E}$-line, or simply line, is the set of solutions of an equation $ax + by = c$, for real numbers a, b, c with $(a, b) \neq (0, 0)$.

Proposition 11.21 (Axiom 1). *$\mathcal{E}$ consists of a nonempty set of elements called points and a nonempty set of subsets called lines. For distinct points P and Q, there is exactly one $\mathcal{E}$-line containing P and Q.*

Proof. The sets of points and lines are clearly not empty. For distinct $P = (p_1, p_2)$ and $Q = (q_1, q_2)$, let us consider the set of points $(1 - t)P + tQ$, where t is any real number. Then the points (x, y) of this set are solutions of

$$-(q_2 - p_2)x + (q_1 - p_1)y = -(q_2 - p_2)p_1 + (q_1 - p_1)p_2.$$

This equation, obtained by eliminating t from the parametrization, can be expressed more transparently with a dot product. If we set the normal direction

$$N = (-(q_2 - p_2), q_1 - p_1),$$

then $N \cdot (Q - P) = 0$, and the equation is $N \cdot (x, y) = N \cdot P$. In the development of the equation of a line earlier in this chapter, it was shown by computations in $\mathbb{R}^2$ that the solution set not only contains $\overleftrightarrow{PQ}$ but equals this set. □

Definition 11.22. The **norm** of a pair $A = (a_1, a_2)$ is $|A| = \sqrt{a_1^2 + a_2^2}$. The distance function in $\mathcal{E}$ is $\|AB\| = |B - A| = \sqrt{(b_1 - a_1)^2 + (b_2 - a_2)^2}$ for A and $B = (b_1, b_2)$.

Proposition 11.23 (Axiom 2). *The plane $\mathcal{E}$ has a distance function.*

- *Distance is a function that maps any two points A and B in the plane to a nonnegative real number $\|AB\|$.*

- *Every line m in the plane has a ruler, a mapping of the real numbers $\mathbb{R}$ onto m so that distances are preserved.*

Proof. Distance was defined above. Let the line $m = \overline{AB}$. If $d = \|AB\|$, then we claim $\rho(t) = (1-(t/d))A + (t/d)B$ is a ruler. To see this for any real numbers u, v, $\rho(v) - \rho(u) = ((v-u)/d)(B-A)$. Therefore, $\|\rho(v), \rho(u)\| = (|v-u|/d)|B-A| = |v-u|$, proving distance preservation as required. $\square$

Before moving on to the next axiom, we will introduce translations into the model with the usual formula.

Definition 11.24. A translation is a transformation T of $\mathbb{R}^2$ of the form $T(X) = X + C$ for some C in $\mathbb{R}^2$.

Note. For any translation, $|T(X) - T(Y)| = |X - Y|$ so translations preserve $\mathcal{E}$-distance. As usual, $\mathcal{T}_{AB}(X) = X + B - A$.

Definition 11.25. As before, for any real θ, let $U(\theta) = (\cos\theta, \sin\theta)$ and let $O = (0,0)$. Then the $\mathcal{E}$-protractor at O is the mapping $\delta(\theta) = \overrightarrow{OU(\theta)}$ and the protractor δ_A at point A is $\mathcal{T}_{OA}(\delta(\theta)) = \overrightarrow{A(A+U(\theta))}$.

Definition 11.26. Angle measure is defined in $\mathcal{E}$ as follows: $m\angle BAC = |b-c|$ when $\delta_A(b) = \overrightarrow{AB}$ and $\delta_A(c) = \overrightarrow{AC}$, provided that $|b-c| \leq 180$.

Proposition 11.27 (Axiom 3)**.** *The $\mathcal{E}$-plane has an **angle measure** that maps any two rays $\overrightarrow{AB}$ and $\overrightarrow{AC}$ with common endpoint to a real number $m\angle BAC$.*

- *The angle measure maps any angle $\angle BAC$ to a positive number $m\angle BAC$, with $0 < m\angle BAC < 180$.*
- *In addition, when $\overrightarrow{AB} = \overrightarrow{AC}$, the angle measure of this zero angle is $m\angle BAC = 0$. If the rays $\overrightarrow{AB}$ and $\overrightarrow{AC}$ are opposite rays, the angle measure of this straight angle is $m\angle BAC = 180$.*
- *For any point A, there exists a **protractor** at A. This is a mapping δ from the real numbers onto the rays with endpoint A, such that $\delta(t) = \delta(t+360)$ for every t, with the property that $m\angle BAC = |b-c|$ when $\delta(b) = \overrightarrow{AB}$ and $\delta(c) = \overrightarrow{AC}$, provided that $|b-c| \leq 180$.*

Proof. This all follows immediately from the definitions of protractor and angle measure, noting that $\overrightarrow{OU(\theta+180)}$ is the opposite ray of $\overrightarrow{OU(\theta)}$. $\square$

This short proof for what is arguably the most complicated of the axioms may seem unsettling, but if one thinks about it, it is a setup. The candidate for protractor is defined. Then angle measure is defined so that this protractor precisely fulfills the requirements of the axiom. The only point that is left to well-known mathematics is that the mapping $U(\theta)$ from $\mathbb{R}$ to the unit circle is periodic and onto. If one accepts this, there is nothing left to prove.

But perhaps the nagging question is whether this is really the angle measure that we expect in the Cartesian plane. Yes, it is. For example, if $A = (1,0)$ and $B = (3,4)$, what is $m\angle AOB$? Since $|(3,4)| = 5$, $\overrightarrow{OB} = \overrightarrow{OU(\theta)}$, where $U(\theta) = (3/5, 4/5)$. Then $m\angle AOB = \arccos(3/5)$, converted from radians to degrees.

So there is really substance in the proof, but the substance is in U and the trigonometric functions.

Before considering the final three axioms, we will introduce rotations as rigid motions in $\mathcal{E}$.

Definition 11.28. The rotation O_θ is the transformation defined as $O_\theta(x,y) = xU(\theta)+ yV(\theta)$, where $V(\theta) = (-\sin\theta, \cos\theta)$. For any point A, $A_\theta = \mathcal{T}_{OA} O_\theta \mathcal{T}_{AO}$.

Proposition 11.29. *Translations and rotations are both rigid motions of $\mathcal{E}$, meaning that they preserve distance and angle measure.*

Proof. It was already noted that translations preserve distance. They preserve angle measure because angle measure was defined as translation invariant. Rotations preserve distance by the same computation given earlier for the Cartesian plane. And they preserve measure since they just add a constant to each polar angle. ☐

The Final Three Axioms — In Part. For the final three axioms, we will follow a different strategy. We will prove each of Axioms 4 and 5 only for a line that is the x-axis and we will prove Axiom 6 only for the point O. Then we will transport the x-axis to any line or O to any point by means of a rigid motion.

Definition 11.30 (Half-planes of the x-axis)**.** For the x-axis, the set of points $(x, 0)$, the upper half-plane is the set of points (x, y) with $y > 0$ and the lower half-plane is the set with $y < 0$.

Proposition 11.31 (Axiom 4 for the x-axis)**.** *Let m denote the x-axis; then any point not on line m belongs to one of two disjoint sets called the **half-planes of the line**.*

- *Points P and Q are in different half-planes when PQ intersects m. They are in the same half-plane if PQ does not intersect m.*
- *Let $m = \overline{AB}$. If a protractor is chosen so that $\theta(\overrightarrow{AB}) = 0$, the points P for which $0 < \theta(\overrightarrow{AP}) < 180$ lie in one half-plane, and the Q for which $-180 < \theta(\overrightarrow{AQ}) < 0$ are in the other half-plane.*

Proof. The segment PQ intersects the x-axis when $(1-t)P + tQ$ at $t = p_2/(p_2 - q_2)$, a number between 0 and 1 when p_2 and q_2 have opposite sign. Also, $U(\theta)$ is in the upper half-plane when $0 < \theta < 180$ and in the lower half-plane when $-180 < \theta < 0$. ☐

Proposition 11.32 (Axiom 5 for the x-axis)**.** *For the x-axis, there exists a rigid motion distinct from the identity transformation that fixes the points of the x-axis.*

Proof. Let $R(x, y) = (x, -y)$. This is clearly a rigid motion that fixes the points of the x-axis. ☐

Definition 11.33. The $\mathcal{E}$-dilation $\mathcal{D}_{O,k}$ is the mapping $\mathcal{D}_{O,k}(x, y) = (kx, ky) = k(x, y)$.

Proposition 11.34 (Axiom 6 for O)**.** *For every point O and every $k > 0$, the dilation $\mathcal{D}_{O,k}$ preserves angle measure and scales every distance by k: for all $P' = \mathcal{D}_{O,k}(P)$ and $Q' = \mathcal{D}_{O,k}(Q)$, $\|P'Q'\| = k\|PQ\|$.*

Proof. A direct substitution proves the scaling statement. $P' = \mathcal{D}_{O,k}(P) = kP$ and $Q' = \mathcal{D}_{O,k}(Q) = kQ$, so $\|P'Q'\| = |P' - Q'| = k|P - Q| = k\|PQ\|$. As for angles, $\mathcal{D}_{O,k}$ maps a ray $\overrightarrow{AB}$ to a parallel ray $\overrightarrow{A'B'}$ in the same direction. Therefore, $\mathcal{D}_{O,k}(\overrightarrow{AB}) = \mathcal{T}_{AA'}(\overrightarrow{AB})$. This means that the dilation image of an angle $\angle BAC$ is also a translation image of this angle, so the measures are the same. □

The Final Three Axioms. To complete the proof that the final three axioms hold in $\mathcal{E}$, we will use translations and rotations, which are known to be rigid motions in $\mathcal{E}$.

To prove the Plane Separation Axiom, Axiom 4, for any line m, first rotate the x-axis so that it is parallel to m; then translate the rotated line to m. Then the half-planes and all the relations are preserved by the rigid motion S that is the composition of these two rigid motions.

To prove the Reflection Axiom for a line m, again choose the same rigid motion S that takes the x-axis to m. Then if R is the reflection that we defined for the x-axis, the rigid motion SRS^{-1} is a rigid motion that fixes the points of m.

Finally, the dilation $\mathcal{D}_{A,k}$ with the usual definition equals $\mathcal{T}_{OA}\mathcal{D}_{O,k}\mathcal{T}_{AO}$. Since translations are rigid motions, the scaling relation will remain true.

Thus we have connected back to the beginning with a model that satisfies Axioms 1–6 of Chapter 2.

Conclusion

While the geometry in this chapter is interesting in its own right, the concepts and tools introduced here extend to areas of geometry that would require several more books to pursue. In plane geometry, the complex numbers are essential for the geometry of circles and inversions, which can in turn be used to model non-Euclidean geometry. Barycentric coordinates can be extended to include a line at infinity with $c_1+c_2+c_3 = 0$, leading to projective geometry. Beyond the plane, affine and Cartesian coordinates can be extended to any dimension. Matrices and vectors in all dimensions are everywhere in mathematics and its applications. Our hope is that this introduction to these tools and concepts through plane Euclidean and affine geometry will provide background and motivation for further exploration, either in the deeper regions of Euclidean geometry itself or in other areas beyond.

Exercises and Explorations

1. (Pencils of Lines). For two functions $f_0(x,y) = a_0x + b_0y - c_0$ and $f_1(x,y) = a_1x + b_1y - c_1$, the equations $f_0(x,y) = 0$ and $f_1(x,y) = 0$ are the equations of lines m_0 and m_1. Let t be any real number; then $(1-t)f_0(x,y) + tf_1(x,y) = 0$ is also the equation of a line m_t.
   (a) Prove that if m_0 and m_1 intersect at a point P, then all the m_t pass through P.

(b) Prove that if m_0 and m_1 are parallel, then all the m_t are parallel (or equal) to these lines.

(c) The set of all lines that pass through a point P is called the **pencil of lines through** P. The set of lines parallel to or the same as a line m is called the **pencil of lines parallel to** m. Explain why any two lines belong to a unique pencil and any distinct pencils either intersect in one line or no lines. (If the words line and pencil are interchanged in this sentence, do these words sound familiar?)

(d) Let $m \parallel n$ mean that m and n belong to the same parallel pencil. Show that this relation satisfies the three criteria in Theorem 1.4 when $\cong$ is replaced by $\parallel$. Such a relation is called an equivalence relation. Why is $m \parallel n$ an equivalence relation while the relation "m is parallel to n" is not?

2. (Ruler for a Line). For constants m and b, the map $P(x) = (x, mx+b)$ is a parametrization of a line p but not necessarily a ruler.
   (a) For what values of m and b will this be a ruler?
   (b) Find rulers for some lines with equations such as $x + y = 1$ and $3x + 4y = 5$.

3. (Negative Dot Product). If $m\angle MON = \theta$, show that the dot product $N \cdot M = \|ON\|\|OM\| \cos\theta$, even when the dot product is 0 or negative.

4. (Reflection Formula). Using the notation $O = (0, 0)$, $U(\theta) = (\cos\theta, \sin\theta)$, and $V(\theta) = (-\sin\theta, \cos\theta)$, let F be the function with this coordinate formula:

$$F(x, y) = xU(2\theta) + yV(2\theta) = x(\cos 2\theta, \sin 2\theta) + y(\sin 2\theta, -\cos 2\theta).$$

   (a) Confirm that $F(O) = O$, $F(U(\theta)) = U(\theta)$, and $F(V(\theta)) = -V(\theta)$.
   (b) Show that F preserves distance.
   (c) Explain why this proves that $F = \mathcal{R}_{OU(\theta)}$.

5. (Area and Barycentric Coordinates). For the point S in Figure 16, check by computing barycenters and ratios on each triangle side that the areas are what was stated in the text.

6. (Radical Axis). Given a circle c with center A and radius r, the power of a point P with respect to c is $\|AP\|^2 - r^2$. The radical axis of two circles c_1 and c_2 is the set of points P for which the two powers are.equal.
   (a) Use coordinates to write a formula for the the radical axis of two circles. Show this set is a line perpendicular to the line of centers when the circle centers are distinct.
   (b) Compare this coordinate proof with the proof of Theorem 8.46. How are they similar, and how do they differ?

7. (Transformation Conjugation and Complex Numbers). Let $S(z) = \alpha z$ and $T(z) = z + \beta$. Use complex numbers to compute two transformation conjugations from these transformations (not to be confused with complex conjugate).
   (a) Generalize the corollary on page 181 by computing the conjugation TST^{-1} and finding its fixed point using Theorem 11.8. Since T is a translation and S can be a rotation or a dilation, this generalizes the corollary.
   (b) Compute also the conjugate STS^{-1}. What kind of transformation is this? Draw a picture of this when $\alpha = 2$.

   (c) Let $R(z) = \overline{z}$ (yes, complex conjugate!). Compute RSR^{-1}. What kind of transformation is this? How does this result differ from that of the corollary when $|\alpha| = 1$?

8. (Incenter Barycentric Coordinates). For a triangle $\triangle ABC$, let a, b, and c be the lengths of the sides opposite A, B, C.
   (a) Show that (a, b, c) are the barycentric coordinates of the incenter, the point of concurrence of the angle bisectors.
   (b) What are the barycentric coordinates for the three excenters, the points of concurrence of two exterior angle bisectors and one interior bisector? Hint: Theorem 8.18.

9. (Midpoints and Affine Transformations). T is an affine transformation that maps segment AB to CD.
   (a) If $P(t) = (1-t)A + tB$, show that $T(P(t)) = Q(t)$, where $Q(t) = (1-t)C + tD$, for $0 \le t \le 1$.
   (b) Prove that the set of midpoints of $P(t)Q(t)$ is a segment or a point. (For an isometry T, this is called Hjelmslev's Theorem, Coxeter [5, Section 3.6].)

10. (Affine Congruence). Prove that in affine geometry all triangles are alike ("affine-congruent"). In other words, for any two triangles, there is an affine transformation that takes one to the other. Hint: Rather than try to find a formula in one step, you can write the affine transformation as the product of simpler transformations that step by step transform one triangle to coincide with the other one.

11. (Uniqueness of Affine Transformations). Prove that if two affine transformations S and T map the vertices of a triangle $\triangle ABC$ to the same points, then $S = T$. Hint: To simplify the problem, let $\triangle ABC$ define a coordinate system.

12. (Area and Determinants). In a Cartesian coordinate system, let $A = (a_1, a_2)$ and $B = (b_1, b_2)$. If $C = A + B$, prove that the area of a parallelogram $OACB$ is the absolute value of the determinant of $\begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix}$, with the sign of the determinant indicating the orientation of this parallelogram.

13. (Bézier Curve). For three points A, B, C, let $F(t) = (1-t)A + tB$ and $G(t) = (1-t)B + tC$. The image curve of the function $P(t) = (1-t)F(t) + tG(t)$ is called the Bézier curve with A, B, C as control points. The curve is used in computer graphics.

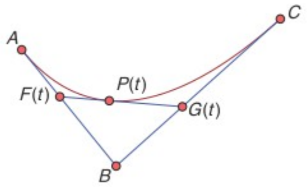

Figure 25. Bézier Curve for $0 \le t \le 1$

(a) Draw three points A, B, C, then several examples of $F(t)$, $G(t)$, and $P(t)$ for several values of t. It would be best if you could do this with dynamic geometry software and let t vary, but it is otherwise helpful to see an example with several points plotted.

(b) For the points $A = (-c, 1)$, $B = (0, 0)$, and $C = (c, 1)$, show that this curve satisfies the equation $y = ax^2 - b$ for some a and b. This is the equation for a parabola. Note: There is an affine transformation that will map these control points to any distinct and noncollinear A, B, C and so will map this parabola to the corresponding Bézier curve. Therefore, all such Bézier curves are parabolas.

14. In Figure 26, one sees on the left how a one-dimensional object, a segment, is translated to another segment that is the opposite side of a parallelogram. In the middle figure, a parallelogram is translated to what looks like the opposite face of the plane projection of a three-dimensional object, such as a cube or a parallelepiped. This figure is repeated on the right, without the arrows.
    (a) In the figure on the right, show that the four "diagonals" indicated by the dashed lines are concurrent. Is this point a center of mass?
    (b) Draw a new figure in this sequence by translating the figure on the right to produce a figure with 16 vertices that appears to be the projection of a four-dimensional object such as a hypercube.
    (c) In this new figure, describe 8 diagonals that generalize the diagonals from the previous case; then show that they are concurrent.

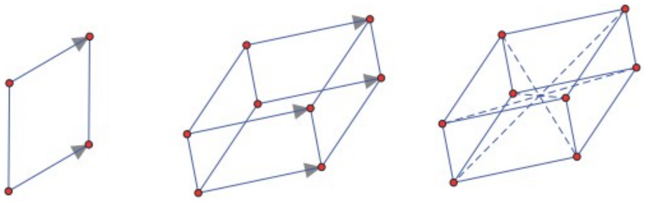

Figure 26. Vectors and Pictures of Higher Dimensions

0 !#12! ! !3 4) ! .) ' ;)& /
)) 6 $ D 4 # *) $ 4 / 5 0 *+
C4 5 >0

☆7☆

Areas, Angles, and Polyhedra

The first goal of trigonometry—to solve any triangle given some information about its sides and angles—has been accomplished, so it is at this point that most textbooks stop. This is a pity, because while the straightforward practical work has been completed, a wealth of mathematical pleasures that might have spurred a lot of curiosity lies just around the corner. Fortunately we are not bound by early 20th-century mathematics curricula, so we shall press onward and taste some of these delights.

We begin with a seemingly practical problem that we have so far carefully ignored: to find the area of a spherical triangle or polygon. In fact, the applications of finding areas on the sphere are somewhat limited. Hardly anyone has ever needed to calculate areas of tracts of land or ocean so vast that the curvature of the Earth needed to be accounted for. And in astronomy, predicting the positions of the Sun, Moon, and planets does not rely on knowledge of areas of sections of the sky. So historically, scientists simply didn't care. However, there is a mathematical motive: an exploration of areas leads quickly, and rather unexpectedly, to a tour of some of the greatest geometrical theorems ever discovered. So for this chapter we shall depart from our usual physical contexts and take a journey for the sake of pure geometric pleasure.

Recall from chapter 2 that it is possible to form a spherical polygon with only two sides. A *lune* is the part of a sphere captured between two great semicircles joined at their ends (figure 7.1), named because of its resemblance to a crescent moon. Its area is easy to find, since the ratio of the angle θ between the two great circles to 360° is equal to the ratio of the area of the lune to the surface area of the sphere. In the standard unit sphere the surface area is $4\pi r^2 = 4\pi$, so

$$\frac{\text{area of lune}}{4\pi} = \frac{\theta}{360^\circ}.$$

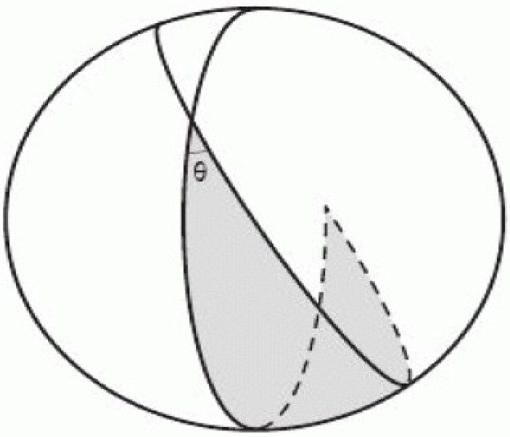

Figure 7.1. Finding the area of a lune.

Thus, measuring θ in degrees,

$$\text{area of lune} = \frac{\pi\theta}{90}.$$

The formula for the area of a spherical triangle is named after Albert Girard (1595–1632), a French mathematician whose Protestant faith likely forced him to flee his home country and settle in the Netherlands. He appears to have struggled to make a living there, with no patron and eleven children. Whether or not this difficult circumstance led to his early death is left to the reader to speculate. Girard lived at a time when the symbols we use today for algebra were in the process of being formulated. He was one of the first to use the abbreviations "sin," "tan," and "sec," and in his *Invention nouvelle en l'algebre* (1629) he invented the $\sqrt[3]{}$ notation for cube roots. Surprisingly, his theorem on the areas of spherical triangles is found not in his *Trigonométrie* (1626), but in the *Invention nouvelle*. He was not happy with his own demonstration; a full proof written by Bonaventura Cavalieri would eventually appear three years later.

> →Girard's idea is simple: extend all three sides of the given triangle into great circles and consider the triangles and lunes that result. In figure 7.2 the original triangle is ΔABC, and the $'$ symbols represent antipodal points. The triangle may be extended in three different ways to form lunes (colunar triangles): extend the sides departing from A all the way to the antipodal point A', forming $ABA'C$; or extend the sides from B; or extend the sides from C. If we add these three lunes together, we get

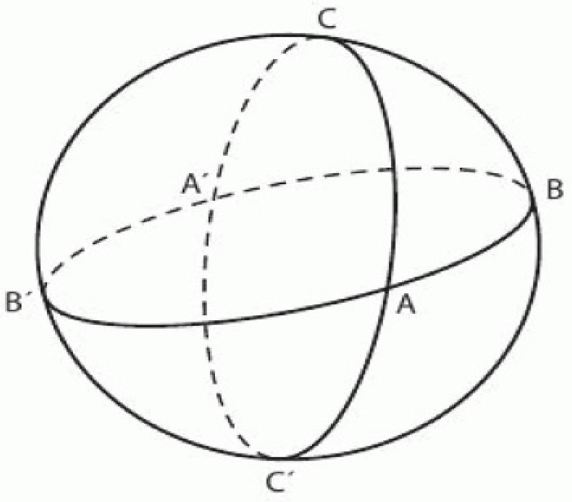

Figure 7.2. Proving Girard's Theorem.

$$2\Delta ABC + (\Delta ABC + \Delta A'BC + \Delta AB'C + \Delta ABC').$$

By symmetry we can replace $\Delta A'BC$ with $\Delta AB'C'$; then the four triangles grouped in parentheses will form the hemisphere at the front of figure 7.2, with area 2π. But the areas of the three lunes, considered separately, may be found using the area formula we derived a few moments ago. So

$$2\Delta ABC + 2\pi = \frac{\pi}{90}(A + B + C),$$

which simplifies to

$$\Delta ABC = \frac{\pi}{180}(A + B + C - 180^\circ).\rightarrow$$

In other words, the area of a spherical triangle is proportional to the amount by which the sum of its angles exceeds 180°. This *spherical excess*, which we call $2E$ in anticipation of events to come in the next chapter, thus plays a meaningful role here, and its significance will continue to grow.

We can extend this result to find the area of any convex spherical polygon. Choose any point in the polygon's interior (figure 7.3) and connect it with each vertex, thereby breaking the polygon into triangles. Let n be the number of sides, and add up the areas of the triangles:

$$\text{Area} = \frac{\pi}{180}[(\text{Sum of triangles' angles}) - n \cdot 180^\circ]$$

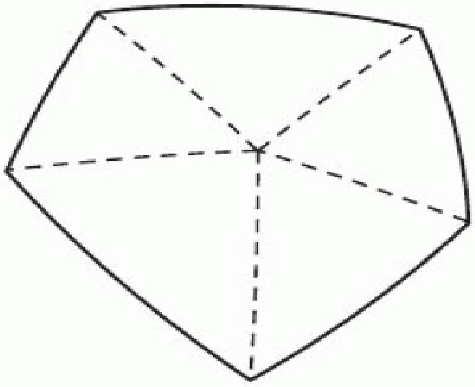

Figure 7.3. Decomposing a spherical polygon into triangles.

$$= \frac{\pi}{180}[(\text{Sum of polygon's angles}) + 360^\circ - n \cdot 180^\circ]$$

$$= \frac{\pi}{180}[(\text{Sum of polygon's angles}) - (n - 2) \cdot 180^\circ].$$

Since the angles of a planar polygon sum to $(n-2) \cdot 180^\circ$, it makes sense to refer to the square-bracketed quantity as the *spherical excess of the polygon*.

Euler's Polyhedral Formula

Here our subject takes a surprising turn, apparently away from spheres altogether. Consider any *convex polyhedron*, that is, a solid consisting of polygons as faces and having no inward "dents." Five such polyhedra can be constructed using identical regular polygons for each face: using triangles, we get the tetrahedron, octahedron, and icosahedron; using squares, the cube; and using pentagons, the dodecahedron (figure 7.4). These are known as the *regular polyhedra*. But there are plenty other polyhedra that are not regular, such as the square pyramid and the cuboctahedron in figure 7.5. Pick any of these shapes, say, the cuboctahedron. It has $V = 12$ vertices, $E = 24$ edges, and $F = 14$ faces (six squares and eight triangles). So $V - E + F = 2$. Try this for the other polyhedra shown here, or indeed any convex polyhedron whatever, and you will always get $V - E + F = 2$.

This curious fact is known today as Euler's polyhedral formula, named after the dominant 18th century mathematician Leonhard Euler (1707–1783). Much of the notation and form of the algebra that we use

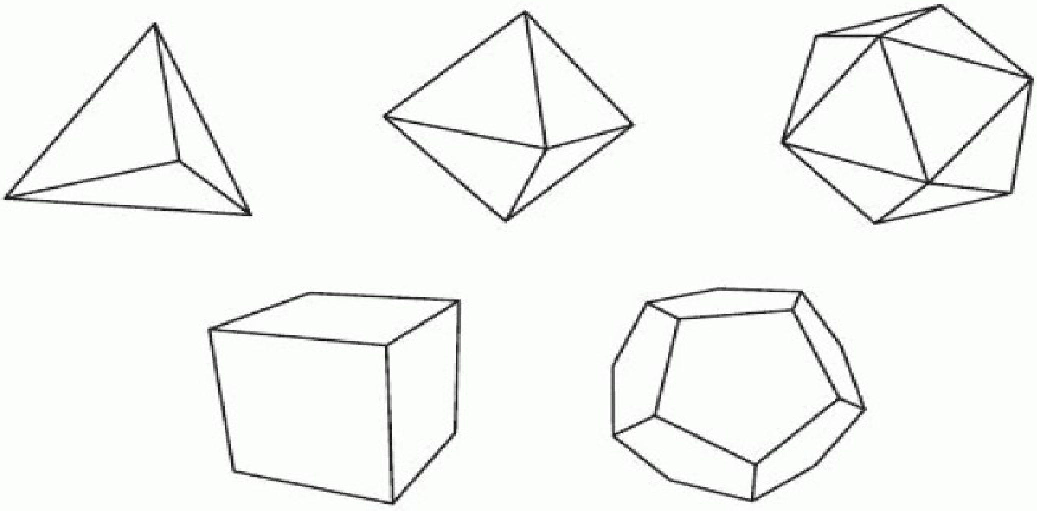

Figure 7.4. The five regular polyhedra—tetrahedron, octahedron, icosahedron, cube, dodecahedron.

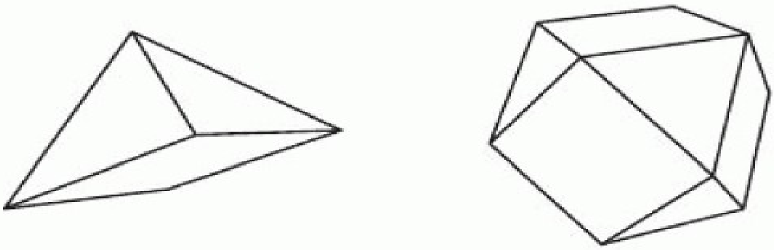

Figure 7.5. Two non-regular polyhedra—the square pyramid and the cuboctahedron.

today in calculus, such as functions and exponentials, was formulated by Euler. His accomplishments are incredibly varied: from geometry and number theory to calculus and differential equations, and onward to astronomy, optics, and navigation, including spherical trigonometry. There is scarcely an area of 18th-century mathematics that Euler did not affect deeply. During his life he wrote an average of almost three pages of published mathematics *per day*—not including the huge volume of work that was released after his death. The onset of blindness in his later life did not slow him down; he simply continued to dictate his mathematics, fully formed, to his assistants.

In 1750 Euler wrote to his colleague Christian Goldbach about the relation $V - E + F = 2$. Eventually he produced an argument that this equation must hold for any convex polyhedron, but his reasoning did not meet the mettle of a full-scale proof. Today, there are at least 19 different proofs. Perhaps the most common approach uses *graph theory*, an area

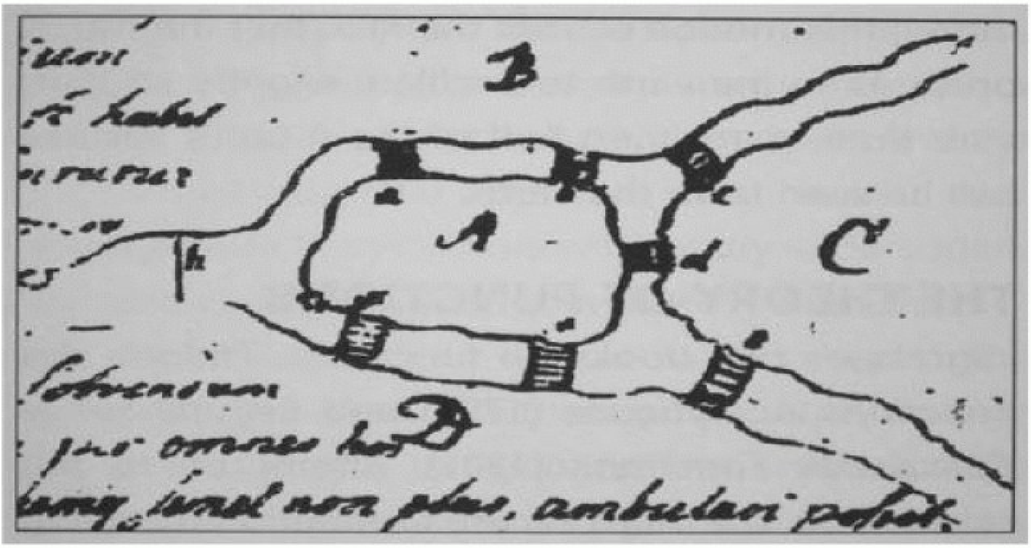

Figure 7.6. The bridges of Königsberg as they appeared in Euler's paper on the subject. See the original paper at the Euler Archive (http://eulerarchive.maa.org/pages/E053.html).

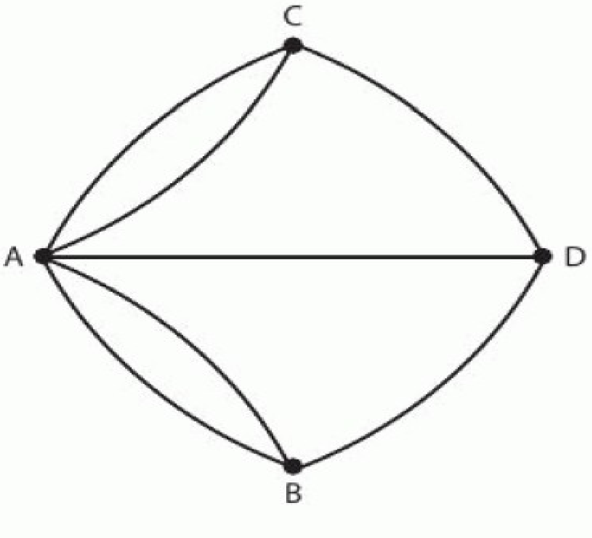

Figure 7.7. A graph representing the Bridges of Königsberg problem.

of mathematics sometimes attributed (mostly falsely) to Euler. The story goes that the citizens of Königsberg enjoyed weekend strolls through their town, which is located along a river with two islands (figure 7.6). Naturally, they wished for a journey that took them across each bridge exactly once without retracing their path, making for a more interesting walk. We begin our search for the ideal path by considering each location (the two river banks and the two islands) to be just a single point called a *node*, and each bridge to be an *edge*. This leads to the *graph* in figure 7.7. We may verify quickly that departing from each of the four nodes is an odd number of edges. Imagine that the ideal path exists; if so, the citizen would both enter *and* depart each node (other than the start and finish) some number of times; therefore only the start and finish may have an odd number of departing edges. But all four nodes in our graph have an odd number of departing edges; therefore, the ideal path cannot exist. Euler's argument was something like this, although it would be more than a century before diagrams such as figure 7.7 were drawn.

The proof of $V - E + F = 2$ using graph theory works as follows. Remove one face of the polyhedron, and stretch out the edges to produce a graph on a flat surface (so that the edges of the missing face form the outer boundary). For instance, in figure 7.8(a) we have removed the square base of the pyramid. Since one face has been removed, we must show that $V - E + F = 1$ for this graph. For each face that is not a triangle, add an edge to the graph by joining two non-adjacent vertices; continue doing this until only triangles remain. Each time we add an edge, E and F increase by one, leaving $V - E + F$ unchanged. (In our case each face is already a triangle, so we do not have to perform this procedure.) Now choose a triangle on the outer boundary of the graph. If only one of the triangle's edges is on the boundary, then remove that edge (figure 7.8(b)). Then E decreases by one and F decreases by one, leaving $V - E + F$ unchanged. If two of the triangle's edges are on the boundary, remove them and their shared vertex (figure 7.8(c)). This action decreases E by two, V by one, and F by one, again leaving $V - E + F$ unchanged. Repeat these steps until all that remains is a single triangle. For that triangle, $V - E + F = 3 - 3 + 1 = 1$, and the theorem is proved.

Oddly, however, the first rigorous proof of Euler's polyhedral formula came from an entirely different and seemingly unrelated corner

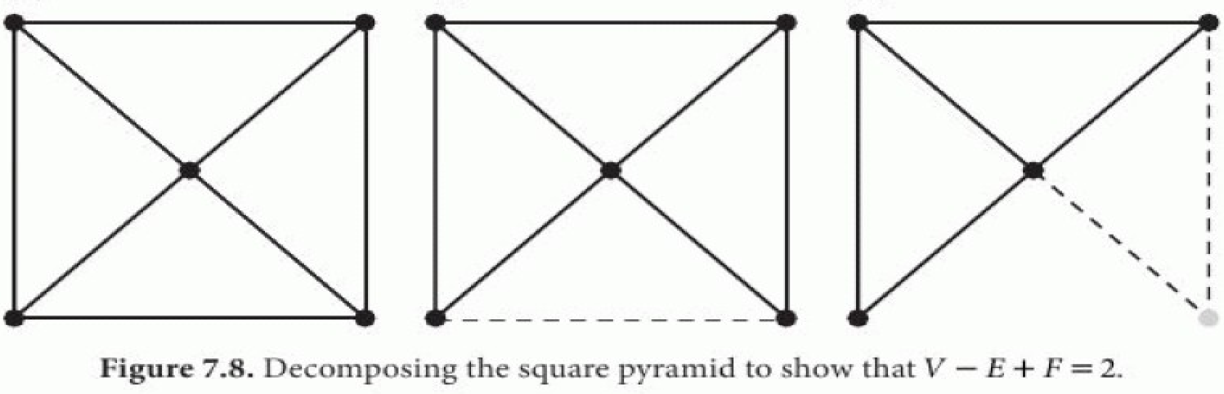

Figure 7.8. Decomposing the square pyramid to show that $V - E + F = 2$.

of mathematics: spherical trigonometry. Forty-four years after Euler's letter to Goldbach, the great French analyst Adrien-Marie Legendre (1752–1833) published his *Éléments de Géométrie*, one of the most successful textbooks ever written. In various versions and translations it swept through Europe and America, in many cases replacing Euclid's *Elements*, and became the standard geometry text for over a century. It contains the first proof that π^2 is irrational, as well as the first proof that $V - E + F = 2$.

The argument appears in the middle of Legendre's Book 7 entitled "The Sphere" (figure 7.9), where no one would think to look for it.

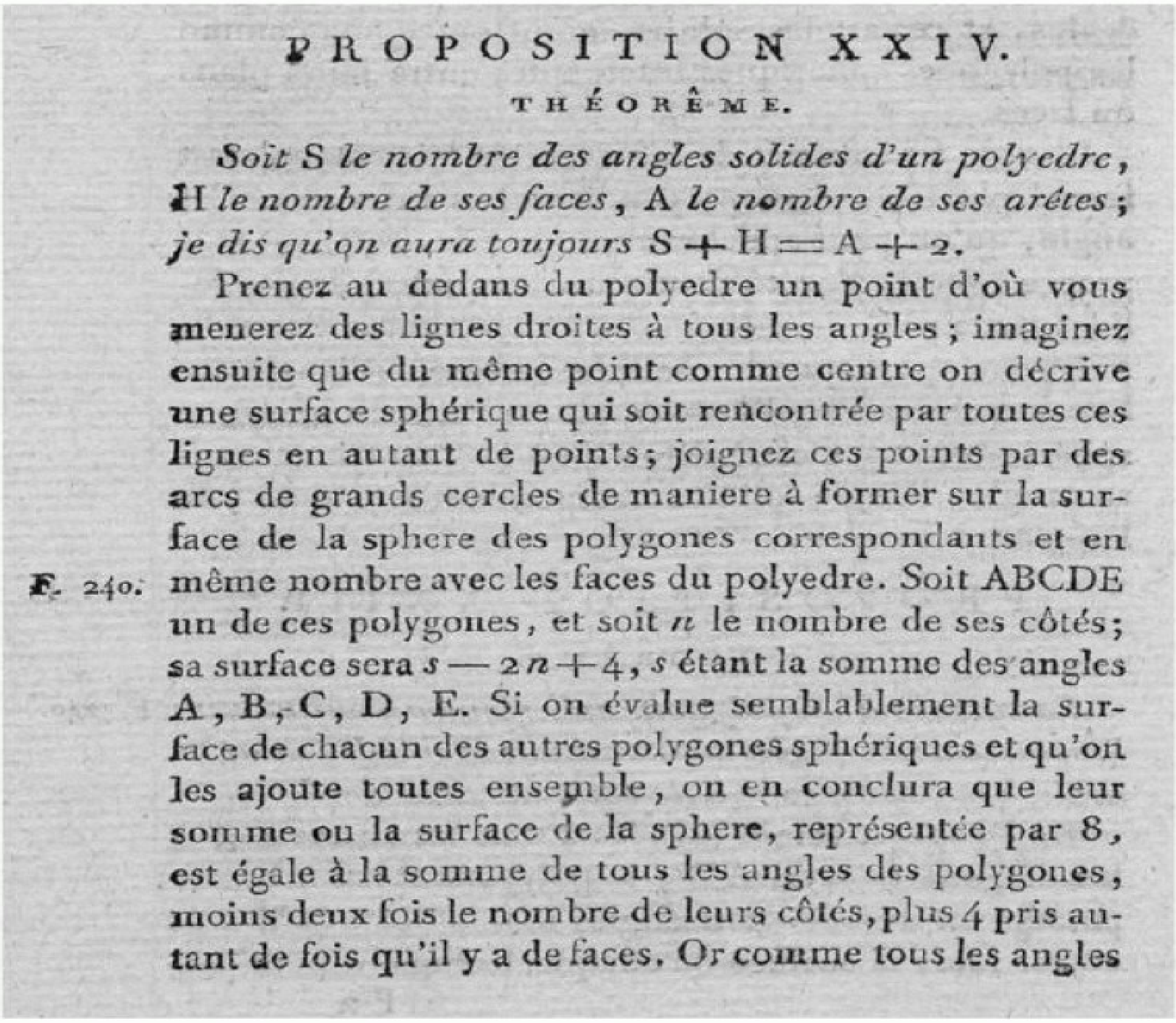

PROPOSITION XXIV.

THÉORÊME.

Soit S *le nombre des angles solides d'un polyedre,* H *le nombre de ses faces,* A *le nombre de ses arêtes; je dis qu'on aura toujours* $S + H = A + 2$.

Prenez au dedans du polyedre un point d'où vous menerez des lignes droites à tous les angles ; imaginez ensuite que du même point comme centre on décrive une surface sphérique qui soit rencontrée par toutes ces lignes en autant de points; joignez ces points par des arcs de grands cercles de maniere à former sur la surface de la sphere des polygones correspondants et en

F. 240. même nombre avec les faces du polyedre. Soit ABCDE un de ces polygones, et soit n le nombre de ses côtés; sa surface sera $s - 2n + 4$, s étant la somme des angles A, B, C, D, E. Si on évalue semblablement la surface de chacun des autres polygones sphériques et qu'on les ajoute toutes ensemble, on en conclura que leur somme ou la surface de la sphere, représentée par 8, est égale à la somme de tous les angles des polygones, moins deux fois le nombre de leurs côtés, plus 4 pris autant de fois qu'il y a de faces. Or comme tous les angles

Figure 7.9. The beginning of the proof of $V - E + F = 2$ in the first edition of Legendre's *Éléments de Géométrie*, the first time a correct proof appeared in print. This item is reproduced by permission of The Huntington Library, San Marino, California.

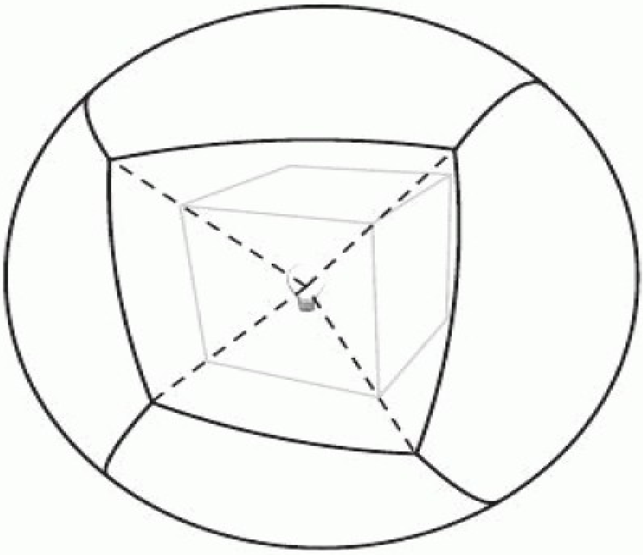

Figure 7.10. Legendre's projection of a polyhedron outward onto the surface of a sphere, in this case a cube. The projected cube is the entire sphere; only the projections of the edges and vertices facing forward are visible.

Legendre begins in a manner similar to the proof we just saw, by projecting the polyhedron—but not onto a flat surface, rather outwards onto an enclosing sphere (figure 7.10). Imagine a point source of light within the polyhedron, with each vertex casting its shadow on the sphere. Connect the "shadowed" vertices on the surface of the sphere with great circle arcs; we now have a *spherical polyhedron*.

> →Assuming that the sphere has a unit radius, its surface area is 4π. But we can find the surface area another way, by adding together the areas of each face of the spherical polyhedron. We happen to have a formula for these areas. In our notation,
>
> $$\sum \frac{\pi}{180}[(\text{Sum of polygon's angles}) - (n-2)\cdot 180^\circ] = 4\pi.$$
>
> Cleaning up a bit and expanding out the sum produces
>
> $$(\text{Sum of all angles}) - \sum n \cdot 180^\circ + 2F \cdot 180^\circ = 720^\circ.$$
>
> But the angles encompass each of the vertices, so the sum of the angles is just $V \cdot 360^\circ$. And since every edge is counted as part of exactly two polygons, $\sum n = 2E$. So
>
> $$V \cdot 360^\circ - 2E \cdot 180^\circ + 2F \cdot 180^\circ = 720^\circ.$$

Canceling 360° brings us to our goal:

$$V - E + F = 2.\rightarrow$$

The Regular Polyhedra

Having Euler's polyhedral formula in our possession brings us very close to yet another of the most famous theorems in mathematics: that the five regular polyhedra in figure 7.4 are the *only* regular polyhedra. This fact has been known since ancient Greece; it is demonstrated by Euclid (in a manner entirely different from what we shall see here) as the culmination of the thirteenth and last book of the *Elements*. Euclid's masterpiece is not a mere logically-ordered listing of theorems; it is arranged to cultivate a sense of suspense. We saw in the previous chapter that Euclid builds Book I to a climax in the Pythagorean Theorem and its converse. He maintains a similar sense of drama with his last three books on solid geometry. As he goes into the home stretch, Euclid shows how to construct each of the five regular polyhedra embedded within a sphere, and gives their dimensions. Finally, just after the last numbered proposition, he concludes with a flourish: there can be no regular polyhedra other than the five he has just constructed.

The regular polyhedra are sometimes called the *Platonic solids*, referring to their appearance in Plato's dialogue, *Timaeus*. In this cosmological work Plato identifies each of the regular polyhedra with one of the Greek elements: fire with the tetrahedron, air with the octahedron, water with the icosahedron, earth with the cube, and the celestial firmament with the dodecahedron. Euclid's proof, then, demonstrated that the analogy between the cosmological elements and the polyhedra was perfect.

Of course Euclid did not have access to $V - E + F = 2$, but we do. Its use makes the path to the conclusion that there are only five regular polyhedra quite short.

> →Let m be the number of sides in a face of a regular polyhedron, and let n be the number of faces that meet at each vertex. Then the number of edges E is equal to $mF/2$, since each edge is the side of two faces; and E is also equal to $nV/2$, since each edge touches two

vertices. From $V - E + F = 2$ and $E = mF/2 = nV/2$, a little algebra gets us to

$$V = \frac{4m}{2(m+n)-mn},\ E = \frac{2mn}{2(m+n)-mn},\ \text{and } F = \frac{4n}{2(m+n)-mn}.$$

The denominator of this expression must be positive, so $2(m + n) > mn$, or (dividing through by $2mn$) $\frac{1}{m} + \frac{1}{n} > \frac{1}{2}$. But both m and n must be greater than 2. A bit of plugging and chugging quickly reveals that the only possible pairs m, n that satisfy the inequality are 3,3 (tetrahedron); 3,4 (octahedron); 4,3 (cube); 3,5 (icosahedron); and 5,3 (dodecahedron).→

The mysticism associated with the regular polyhedra didn't stop with the Greeks. Consider late 16th-century astronomer Johannes Kepler, known today as one of the fathers of modern science for having demonstrated that the planets travel around the Sun in ellipses rather than in combinations of circles. None of the 17th-century natural philosophers really fit the impassive, objective lab-coated image that we conjure when we think of scientists today. Kepler was a deeply committed Christian, and he believed that harmonies were encoded in God's mathematically-inspired creation of the universe. He took the word "harmony" both figuratively and literally: in his *Harmonices mundi* (*Harmony of the World*) he tried to find resonances between musical chords and astronomical ratios. His earlier astronomical work, *Mysterium cosmographicum* (*Cosmic Mystery*), brought to light a peculiar relation between the five regular polyhedra and the six then-known planets. If we nest the five regular polyhedra within spheres, one inside of the other, in just the right order, the ratios of their distances from the center mirror (more or less) the ratios of the distances of the planets from the Sun. Kepler's cosmology, illustrated in the famous drawing of figure 7.11, cannot work beyond the planets visible to the naked eye (for one, there are no more regular polyhedra left to nest), but of course he was not to know that. Uranus would be discovered by Sir William Herschel almost two centuries later.

It doesn't sound easy to work out the dimensions of planetary orbits by calculating the sizes of polyhedra nested in spheres. But by now, it should come as no surprise that spherical trigonometry is the key to a manageable solution. Imagine a regular polyhedron enclosed within a

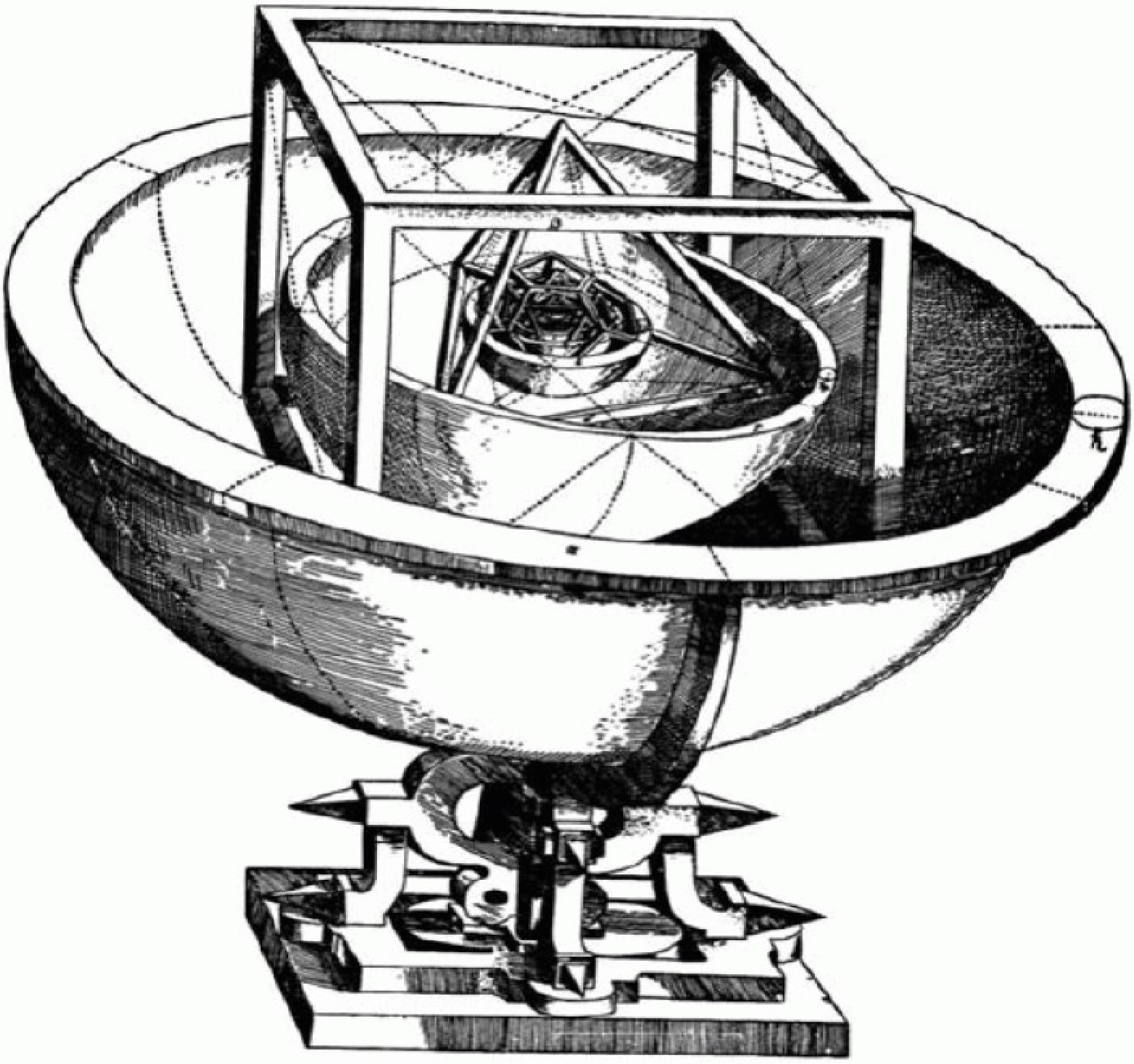

Figure 7.11. Kepler's polyhedral model of the solar system. Courtesy Wikimedia Commons.

unit sphere, meeting the polyhedron at the vertices. Then a sphere inscribed *within* the polyhedron will contact the polyhedron at the center of each face. What is the radius, *r*, of this inscribed sphere?

> →Figure 7.12 illustrates one face (portrayed here as a triangle, but it could be a square or pentagon) of a regular polyhedron, with O at the center. $OA = 1$, and since the inscribed sphere touches the polyhedron at C, $OC = r$. Let E be the end of the perpendicular dropped from C to the midpoint of AB. Imagine OC and OE extended to C' and E' on the circumscribed sphere, forming the right-angled spherical triangle $AC'E'$. Then since $\measuredangle OCA$ is right,

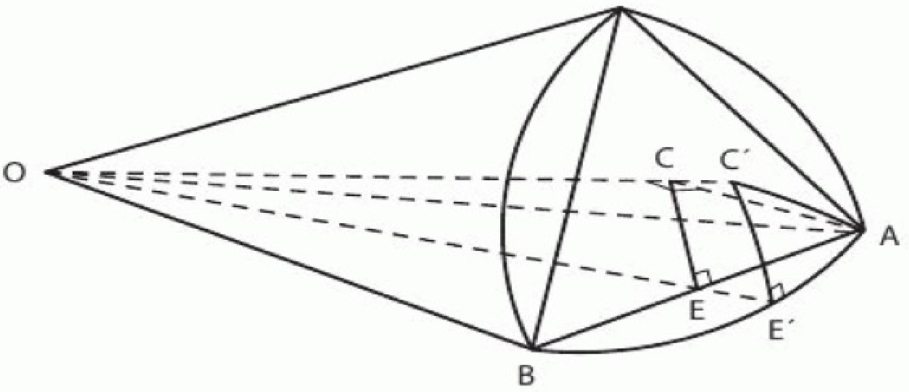

Figure 7.12. Determining the radius of a sphere inscribed within a regular polyhedron.

$$r = \cos\measuredangle AOC = \cos\widehat{AC'}.$$

But from the second of the identities in the first column of Napier's Rules,

$$\cos\widehat{AC'} = \cot\measuredangle AC'E' \cot\measuredangle C'AE'.$$

Now, imagine dropping perpendiculars from C to *every* side of the face we're considering, not just AB. We end up with $2m$ identical angles around C', each equal to $\measuredangle AC'E'$. So $\measuredangle AC'E' = 360°/2m = 180°/m$. Similarly, if we imagine arcs drawn from A to the centers of each of the faces containing A as a vertex, we end up with $2n$ identical angles around A, each equal to $\measuredangle C'AE'$. But $\measuredangle C'AE' = 360°/2n = 180°/n$. We are left with the pleasingly compact relation

$$r = \cot\left(\frac{180°}{m}\right)\cot\left(\frac{180°}{n}\right).\rightarrow$$

The values of r for the regular polyhedra are tabulated in figure 7.13. Notice that the symmetry of our formula with respect to m and n implies that the inscribed spheres for the octahedron and cube, and likewise for the dodecahedron and icosahedron, are identical. This relation is one aspect of the duality of these respective pairs of polyhedra. Another, more commonly expressed aspect of this same duality was discussed in chapter 2: within a regular polyhedron if you connect the central points C of each face, you generate a smaller copy of its dual polyhedron inside.

Kepler ordered his planets and polyhedra, from inside to out, as follows: Mercury-icosahedron-Venus-octahedron-Earth-dodecahedron-

	m	n	r	i	a	Volume
Tetrahedron	3	3	1/3	70.529°	1.633	0.5132 (12.3%)
Cube	4	3	$\sqrt{3}/3 = 0.5774$	90°	$\frac{2\sqrt{3}}{3} = 1.155$	1.5397 (36.8%)
Octahedron	3	4	$\sqrt{3}/3 = 0.5774$	109.471°	$\sqrt{2} = 1.414$	1.3333 (31.8%)
Dodecahedron	5	3	$\frac{(\sqrt{5}+1)\sqrt{6}}{6\sqrt{5-\sqrt{5}}} = 0.7947$	116.565°	0.7136	2.7852 (66.5%)
Icosahedron	3	5	$\frac{(\sqrt{5}+1)\sqrt{6}}{6\sqrt{5-\sqrt{5}}} = 0.7947$	138.190°	1.052	2.5362 (60.5%)

Figure 7.13. The dimensions of regular polyhedra inscribed within the unit sphere.

Mars-tetrahedron-Jupiter-cube-Saturn. From our modern point of view we can see that the large gap caused by the tetrahedron's relatively small value of r nicely fills the space in the solar system occupied by the asteroids between Mars and Jupiter. In fact, if we allow for a small fudge in the gap between Mercury and Venus, the differences between Kepler's polyhedral distances and the planetary distances based on Copernicus's data were on average only about 3%. Given the observational accuracy of this era, that is an impressive match.

Let's follow in Euclid's footsteps in the *Elements* and consider the dimensions of regular polyhedra inscribed in a unit sphere. In particular, what are the lengths a of the edges? It turns out that first we must consider a related question: what is the angular inclination i between two faces of a regular polyhedron?

> →In figure 7.14 one face of the polyhedron is drawn as a pentagon (although, as before, a square or a triangle are also possible). The adjoining face below edge AB (not drawn) has DE on its surface. C is the center of the face above AB, and we drop CE perpendicularly onto AB. Segment DE is the result of doing the same thing to the face below AB. So, the inclination between the two faces is $i = \measuredangle CED$, and $\measuredangle CEO = i/2$.
>
> As before, the spherical trigonometry arises by "popping" ΔACE outwards onto the circumscribed sphere, producing the spherical

3: E) C *)))) * ? &)

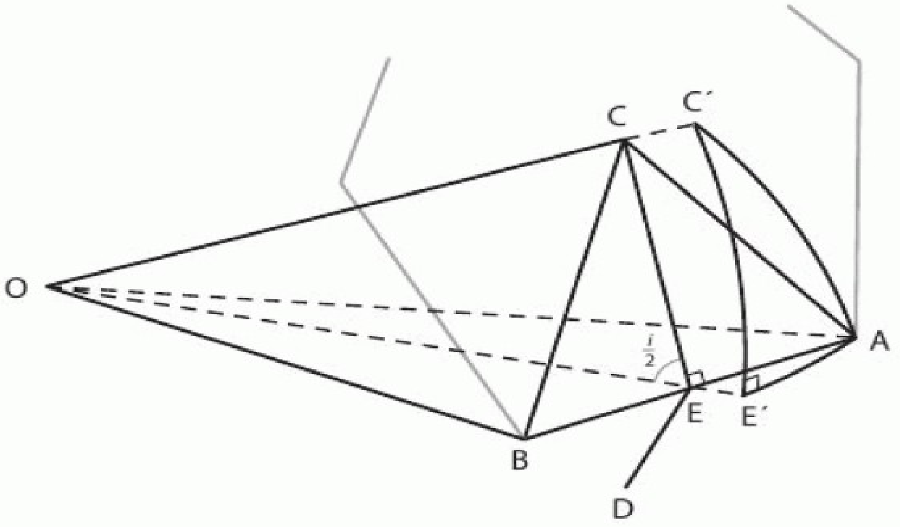

Figure 7.14. Determining the inclination between two faces of a regular polyhedron.

right triangle $AC'E'$. From the second identity in column II of Napier's Rules,

$$\cos\measuredangle C'AE' = \sin\measuredangle AC'E' \cos\widehat{C'E'}.$$

But if we consider (as before) the same sorts of constructions drawn on all faces with a vertex at A, we see that $\measuredangle C'AE' = 360°/2n = 180°/n$. And considering arcs drawn from C' to every vertex and the midpoint of every edge of the visible face, we see that $\measuredangle AC'E' = 360°/2m = 180°/m$. Finally $\widehat{C'E'} = \measuredangle COE$; but $\measuredangle OCE$ is right, so $\widehat{C'E'} = 90° - i/2$. Putting all these results into our Napier's Rule formula above, we get

$$\cos\frac{180°}{n} = \sin\frac{180°}{m}\cos(90° - i/2).$$

A bit of algebraic cleanup provides a pleasing formula for the inclination between two faces:

$$\sin\frac{i}{2} = \frac{\cos(180°/n)}{\sin(180°/m)}.\rightarrow$$

The various angles of inclination are tabulated in figure 7.13. The fact that the octahedron's angle of 109.471° is the same as the "tetrahedral" bond angle at the center of several molecules, including methane, is no coincidence. We leave it to the interested reader to discover why.

→We are now ready to determine a, the length of a side of a regular polyhedron inscribed in a unit sphere. We can reuse figure 7.14. Consider ΔACE; we notice that

$$\frac{CE}{AE} = \cot\measuredangle ACE = \cot\frac{180^\circ}{m}.$$

But $AE = a/2$, so $CE = \frac{a}{2}\cot\frac{180^\circ}{m}$. On the other hand, from ΔCEO we have

$$r = CE\tan\measuredangle CEO = CE\tan i/2.$$

Combining these two results gives us $r = \frac{a}{2}\cot\frac{180^\circ}{m}\tan\frac{i}{2}$. But we already know that $r = \cot\frac{180^\circ}{m}\cot\frac{180^\circ}{n}$. Setting these two equations equal to each other and solving for a brings us home:

$$a = 2\cot\frac{180^\circ}{n}\cot\frac{i}{2}.\rightarrow$$

Once again, the side lengths a are tabulated in figure 7.13.

We find irresistible one last excursion into the geometry of the regular polyhedra. Exactly what proportion of the volume of the unit sphere do the various regular polyhedra occupy?

→This problem turns out to be relatively simple. Connect the center of the sphere to each of the vertices of the polyhedron. This splits the polyhedron into F pyramids with the faces as bases. The volume of each pyramid is $\frac{1}{3}$(area of base)(height). The height of the pyramid is just the radius r of the inscribed sphere. The area of the base, i.e., a face of the polyhedron, may be found by joining the face's central point to each of the vertices. We leave it as an exercise to show that the area of the base is $\frac{ma^2}{4}\cot\frac{180^\circ}{m}$. By combining this information, we arrive at

$$\text{Volume} = \frac{mFra^2}{12}\cot\frac{180^\circ}{m}.\rightarrow$$

Divide by $\frac{4}{3}\pi$ (the volume of the unit sphere), and we have the proportion of the sphere filled by the polyhedron. The tabulations in figure 7.13 reveal a surprising fact: although the icosahedron appears to adhere most closely to its circumscribed sphere, the dodecahedron actually fills about 10% more of the sphere's volume. Sometimes, appearances can be deceiving.

Exercises

1. (a) Find the area of a spherical triangle whose angles are 63°, 84.35°, and 79°, if the radius of the sphere is 10 inches.
   (b) The sides of a spherical triangle are 6.47 in., 8.39 in., and 9.43 in. If the radius of the sphere is 25 in., find the area of the triangle. [Granville 1908, 230]
2. Verify that a spherical equilateral triangle with 60° angles has no area, and that the largest possible triangle is a hemisphere.
3. The state of Colorado is close to a spherical rectangle, ranging from 37° N to 41° N latitude and from 102.05° W to 109.05° W. Find Colorado's area.

   (There are two reasons why Colorado is not quite a spherical rectangle. Firstly, its northern and southern boundaries are actually circles of latitude rather than great circle arcs, i.e., they bend slightly to the left as you walk eastward along them. Secondly, surveying errors led to some small irregularities in the legal borders.)
4. The Bermuda triangle is a region of the Atlantic Ocean traditionally formed by vertices at Puerto Rico (18.5° N, 66° W), Bermuda (32.3° N, 64.9° W), and the southern tip of Florida (25° N, 80.5° W). It holds a reputation, deserved or not, for being the location of an inordinate number of disappearances of ships and planes. The Earth's radius is 3960 miles. Assuming it is a sphere, how many square miles does the Bermuda triangle enclose? (*Hint:* Join all three vertices to the North Pole.)
5. The area of an isosceles right-angled spherical triangle is $\frac{1}{12}$ of the surface of the sphere: calculate the hypotenuse. [Todhunter/Leathem 1907, 107]
6. Prove that in a right-angled spherical triangle $\tan E = \tan\frac{1}{2}a\tan\frac{1}{2}b$. [Casey 1889, 90]
7. Prove the following analogy from Breitschneider:

$$\frac{\sin\frac{1}{2}E\cos\frac{1}{2}(A-E)}{\sin\frac{1}{2}A} = \frac{\sin\frac{1}{2}s\sin\frac{1}{2}(s-a)}{\cos\frac{1}{2}a},$$

   where s is the half perimeter of the triangle. [Casey 1889, 47] (Enterprising readers may look up a hint and Breitschneider's other seven analogies on page 47 of Casey's book, available online, all derivable from Delambre's analogies.)
8. If E, E_A, E_B, and E_C denote the half spherical excesses of a spherical triangle and its three colunar triangles respectively, show that

$E + E_A + E_B + E_C = 180°$, and hence that the sum of the areas of these triangles is equal to half the area of the sphere. [Moritz 1913, 50]

9. Problem 4 would have been easier if we had had a formula for the area of a triangle in terms of its *sides*, rather than in terms of its *angles*. Fortunately there is such a formula, named after Simon Lhuilier (1750–1840), who among other achievements corrected Euler's solution to the Königsberg bridge problem and worked on Euler's polyhedral formula and its exceptions. We shall derive Lhuilier's formula in stages.

   (a) Deduce the following identity from Delambre's first analogy:

$$\frac{\cos\frac{1}{2}(C-\frac{E}{2})-\cos\frac{1}{2}C}{\cos\frac{1}{2}(C-\frac{E}{2})+\cos\frac{1}{2}C} = \frac{\cos\frac{1}{2}(a-b)-\cos\frac{1}{2}c}{\cos\frac{1}{2}(a-b)+\cos\frac{1}{2}c}.$$

   (b) Use the cosine sum-to-product formulas from plane trigonometry, i.e.,

$$\cos x \pm \cos y = \pm 2 \begin{matrix}\cos\\ \sin\end{matrix}\left(\frac{x+y}{2}\right)\begin{matrix}\cos\\ \sin\end{matrix}\left(\frac{x-y}{2}\right),$$

   to derive

$$\tan\tfrac{1}{2}E\tan\tfrac{1}{2}(C-E) = \tan\tfrac{1}{2}(s-a)\tan\tfrac{1}{2}(s-b),$$

   where s is the half perimeter of the triangle.

   (c) Apply the process of (a) and (b) to Delambre's third analogy, instead of the first. Combine your results to obtain Lhuilier's formula:

$$\tan\tfrac{1}{2}E = \sqrt{\tan\tfrac{1}{2}s\tan\tfrac{1}{2}(s-a)\tan\tfrac{1}{2}(s-b)\tan\tfrac{1}{2}(s-c)}.$$

   We may now calculate E directly from the side lengths, and from this find the area using Girard's Theorem. [Todhunter/Leathem 1907, 101–102]

10. (a) If a and b are the radii of the spheres inscribed in and described about a regular tetrahedron, show that $b = 3a$.

    (b) If a is the radius of a sphere inscribed in a regular tetrahedron, and R the radius of the sphere that touches (i.e., is tangent to) the edges, show that $R^2 = 3a^2$. [Todhunter/Leathem 1907, 216]

11. In any convex polyhedron (regular or irregular), prove that the number of faces having an odd number of sides is even, and that the number of vertices having an odd number of edges is odd. [Casey 1889, 131]

12. If a dodecahedron and an icosahedron were each described about a given sphere, the sphere described about these polyhedra will be the same. [Todhunter/Leathen 1907, 216]

13. A regular octahedron is inscribed in a cube so that the corners of the octahedron are at the centers of the faces of the cube: show that the volume of the cube is six times that of the octahedron. [Todhunter/Leathem 1907, 217]
14. Derive precise expressions (that is, containing no decimal approximations) for the volumes of the regular polyhedra in terms of their side lengths a. [Hann 1849, 65–67]

C4 5 > 0)* B 4)* F

C4 > 5 @0)* %)& 4) 4 :/)
3:

C4 >@5 B 0 D 4

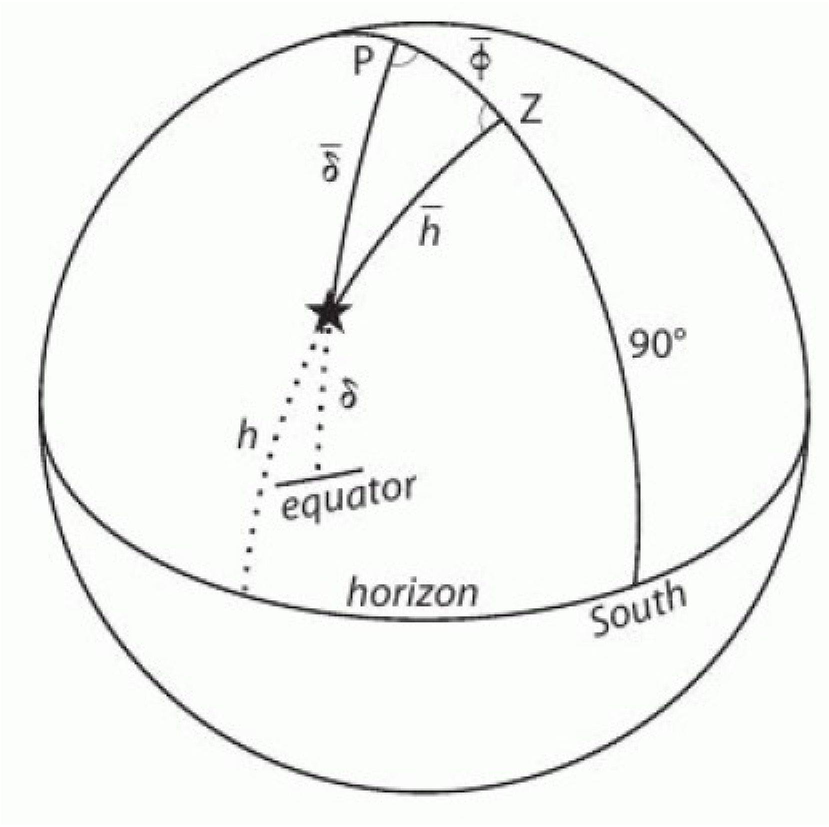
P
$\bar{\phi}$
Z
$\bar{\delta}$
$\bar{h}$
90°
δ
h
equator
horizon
South

D 4 $))) 4

C4 > 5 B 0

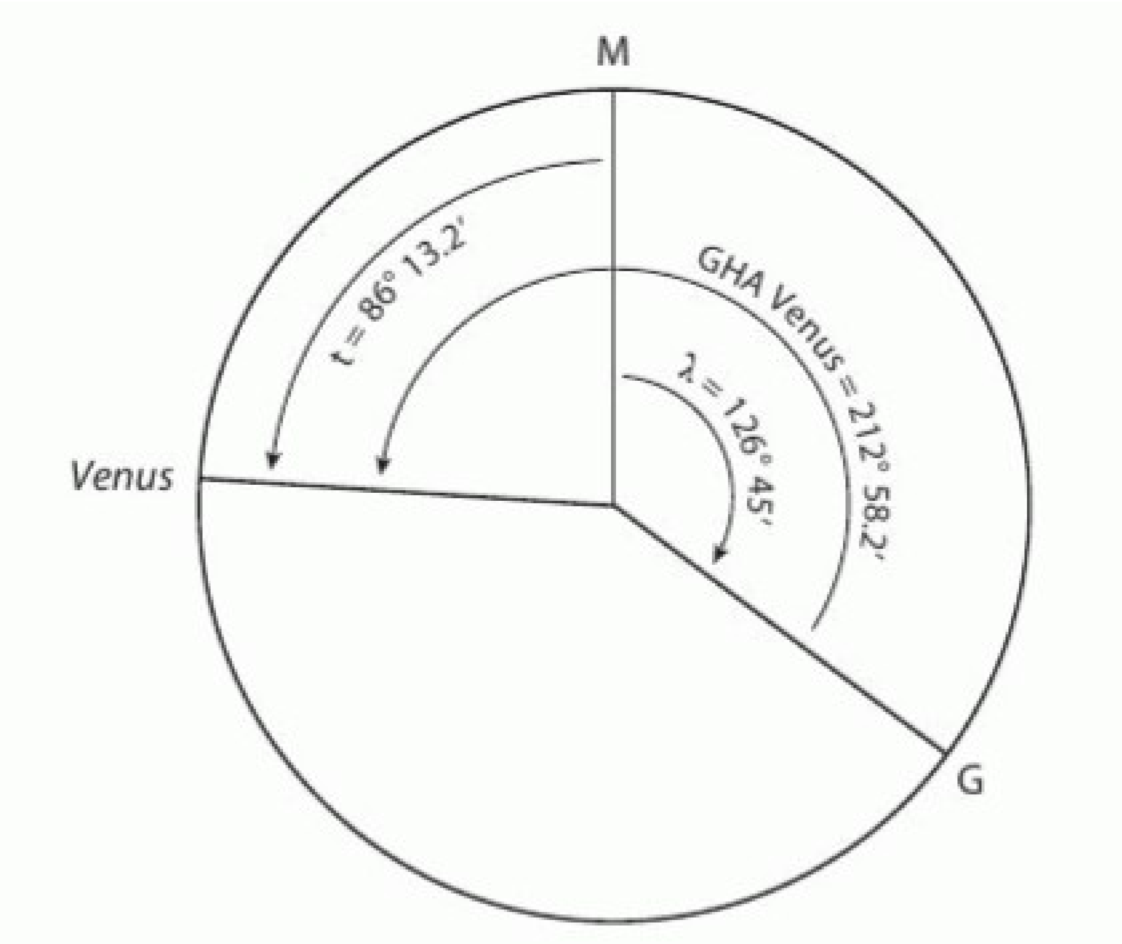
M
t = 86° 13.2′
GHA Venus = 212° 58.2′
λ = 126° 45′
Venus
G

D 4 > ;) 4 +)4)) G) 4) , >) . $, H
,

C4 > 5 B 0

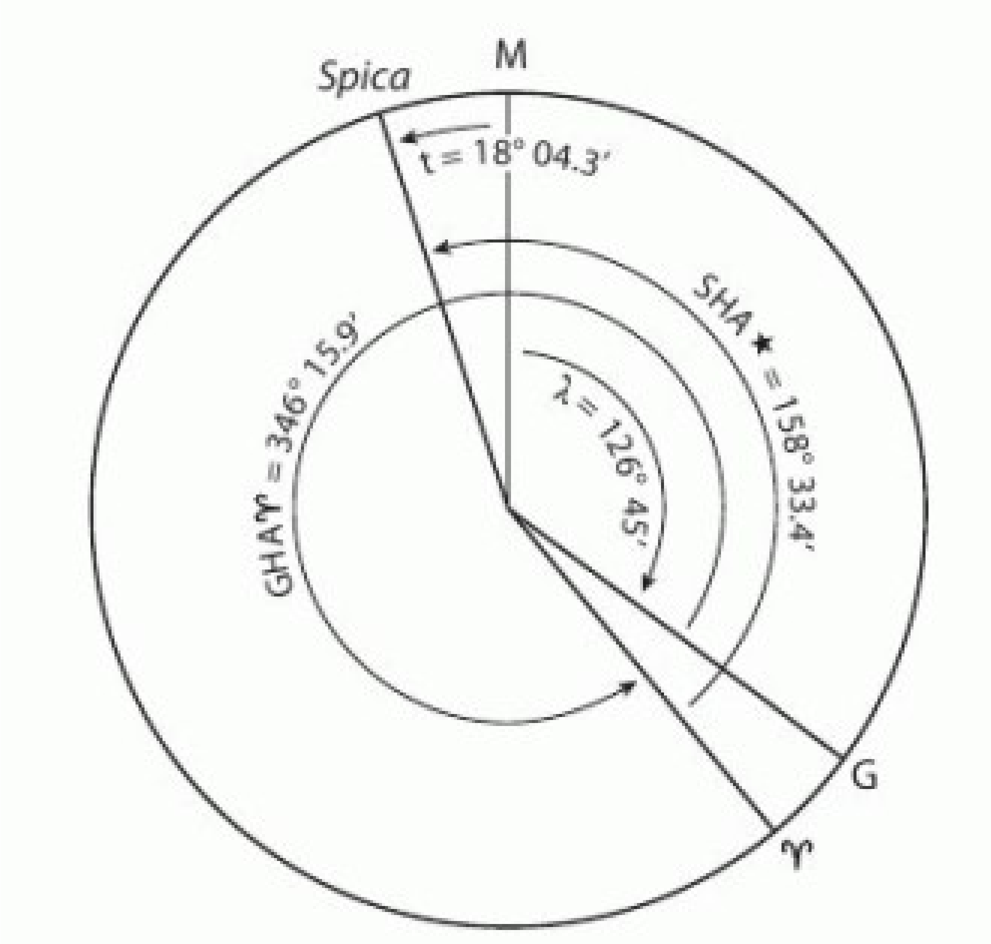
M
Spica
t = 18° 04.3′
SHA ★ = 158° 33.4′
GHA♈ = 346° 15.9′
λ = 126° 45′
G
♈

D 4 ;) 4 +)4) *)) G) 4) H ,)

>) . $

C4 @ 5 B>0

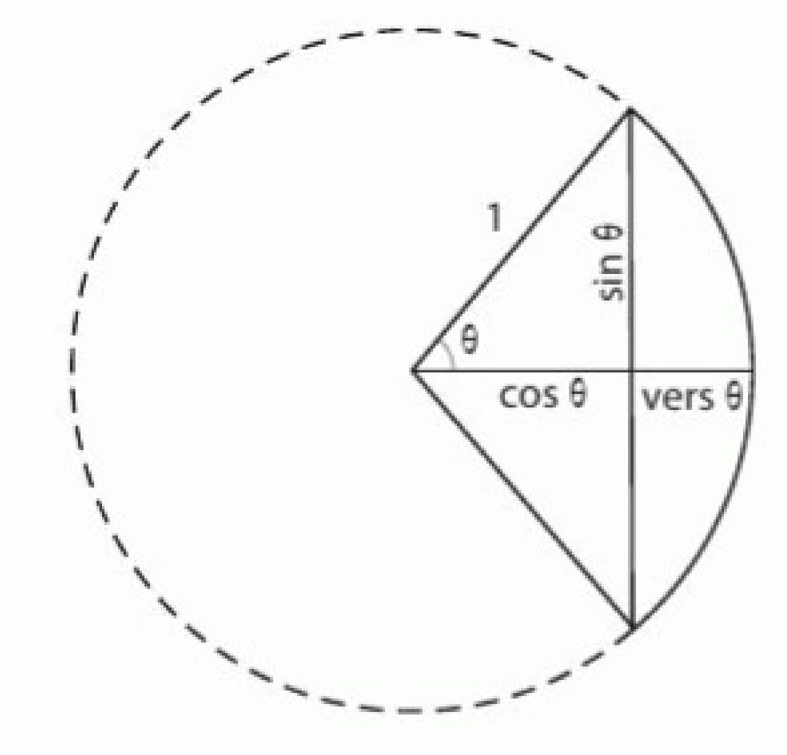

D 4 B $ & +
3:

$$\text{vers}\ \theta = 1 - \cos\theta = 2\sin^2\frac{\theta}{2};$$

or, altering the definition slightly by dividing by 2,

$$\text{hav}\ \theta = \tfrac{1}{2}(1 - \cos\theta) = \sin^2\frac{\theta}{2}.$$

C4 @ 5 BB0

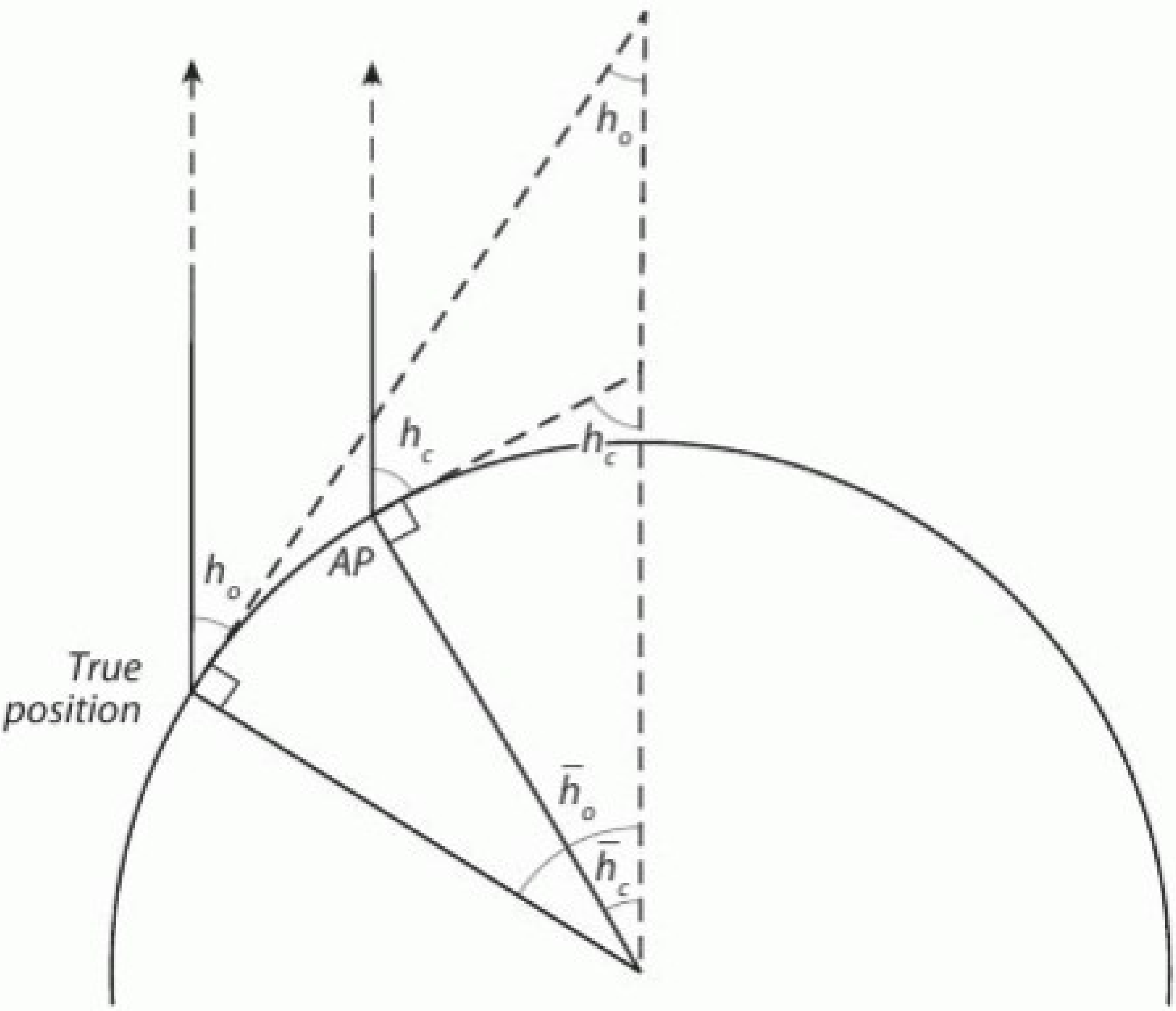
h_o
h_c
h_c
h_o
AP
True position
$\bar{h}_o$
$\bar{h}_c$

D 4 D + 4 *

C4 @>5 0

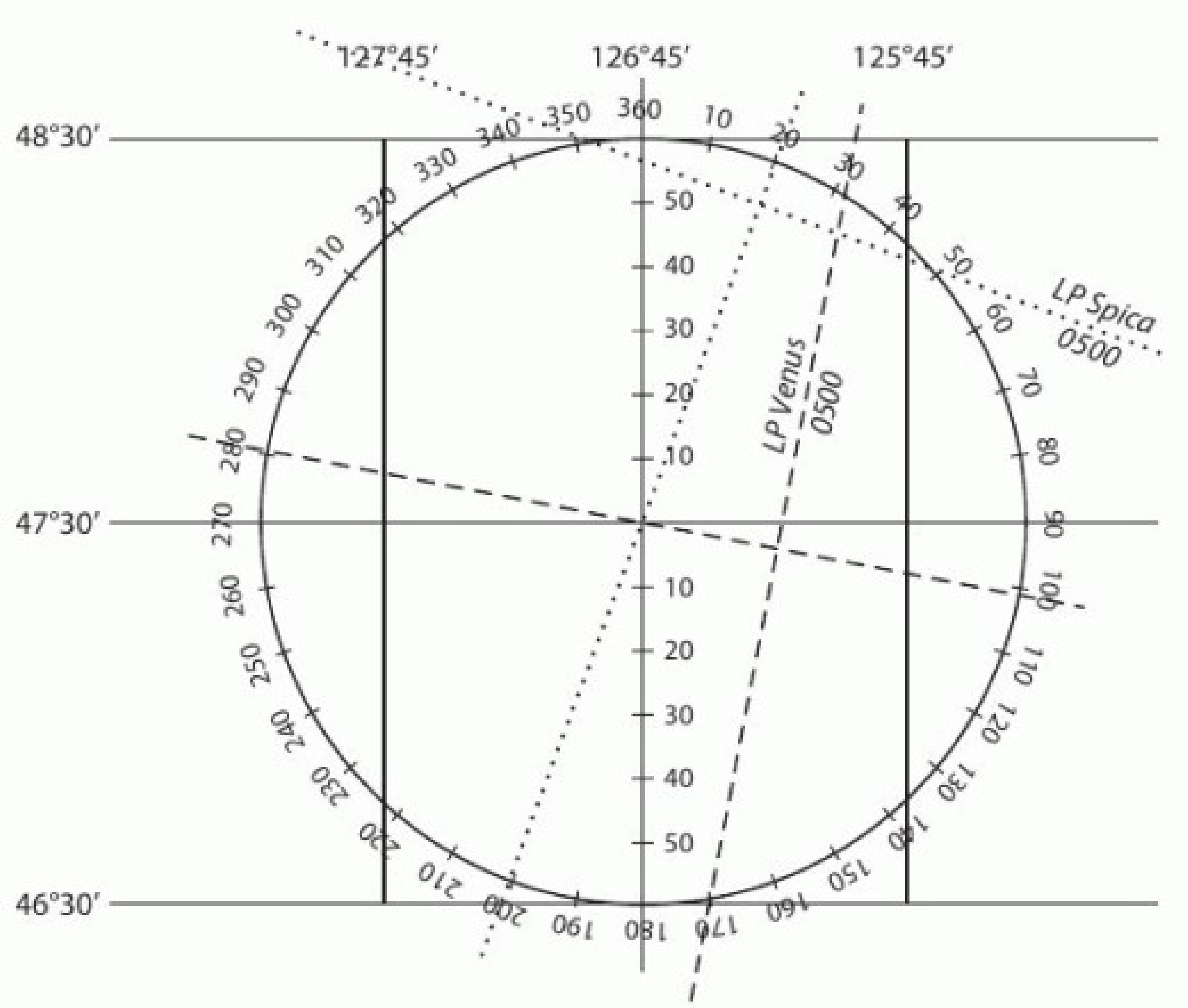
127°45′
126°45′
125°45′
48°30′
47°30′
46°30′
360
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
260
270
280
290
300
310
320
330
340
350
50
40
30
20
10
10
20
30
40
50
LP Spica
0500
LP Venus
0500

D 4 #)))) 4 8 *

C4 5 >0

$$\text{vers}\, t = 1 - \cos t = \{\cos(\phi - \delta) - \sin h\} \sec\phi \sec\delta.$$

C4 5 B0 #** + 8 # /I ") I

C4 5 0

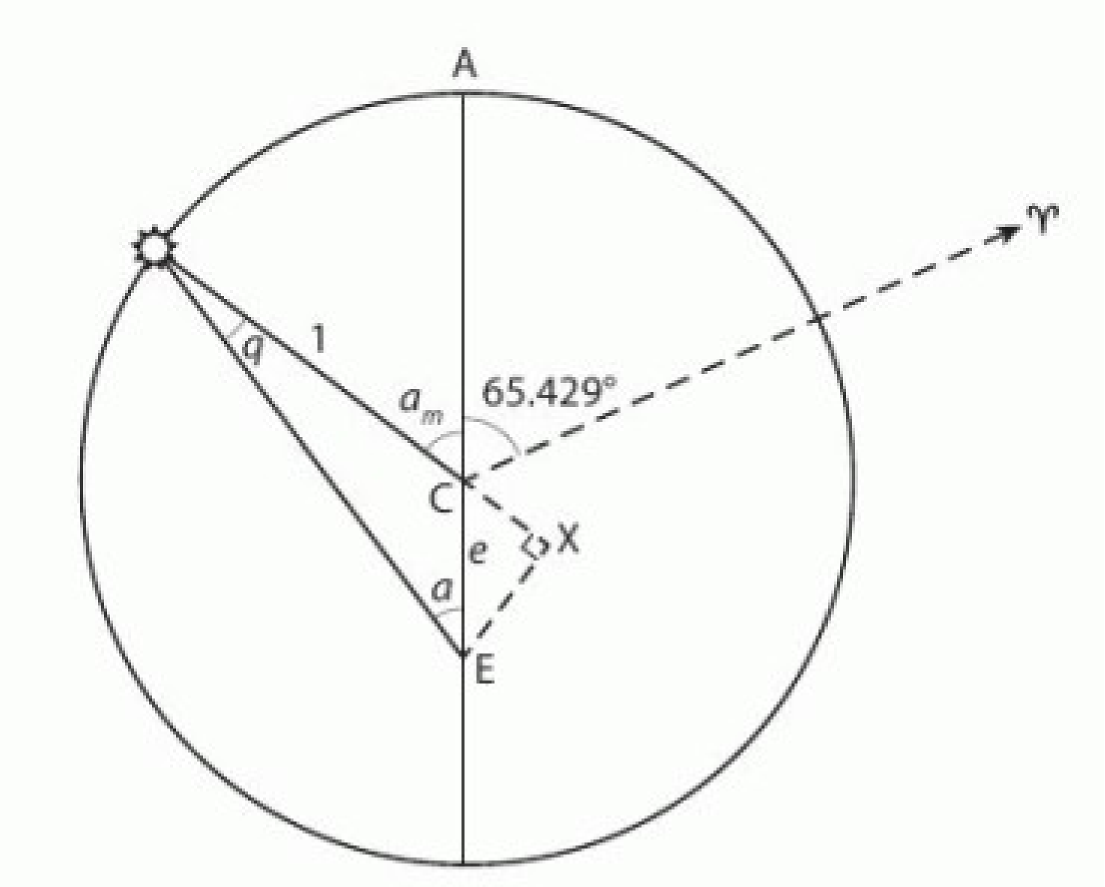
A
♈
1
q
a_m
65.429°
C
e
X
a
E

D 4 #)) 4 I 4 + 4 & /)

3:

$$a_m = t \cdot \frac{360^\circ}{365\frac{1}{4}\text{ days}} - 65.429^\circ.$$

$$\tan q = \frac{EX}{☼X} = \frac{e \sin a_m}{1 + e \cos a_m}.$$

0 !#12! ! !3 4) ! .) ' $
" $) 4 # ; / + $ 4 / E & / 5 0 *+
< *)4
3: A & + ===

C4 >@5 >0 $ 8 @

. & 4) 1 4 # 4 + *) $) 4 ;
5 " 1)+ 0

-)) 4) 4 $ +) + :/ (4 :) ,) +) 4 *
* J + +
# 1)+
:) K
**) 4
+ J +
#) 4 # ',)& 4 :) #', *) L
J) * ,) +) 4 '# L M 4) +) ,
J + + + ' +)
$)( )) L M) +)( ) *)))+ K
*)) L)( ) +) + + K) 4 &
+ J + L M 3)(L) +)( ) *)))+ K
*)) L M)( ) +) +)) ,) :
L M) :

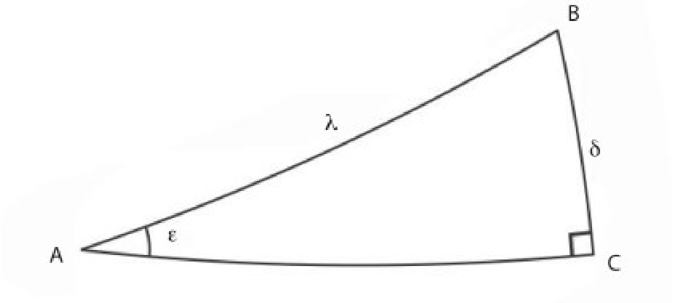

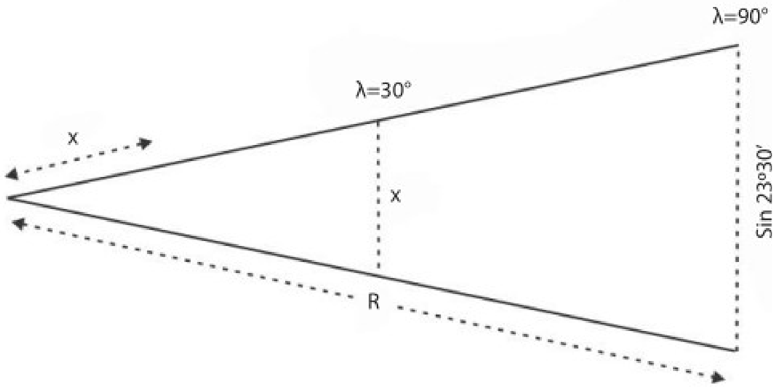

D 4
D + 4 +) 4 . N

Explanation: (See figure 1.29) By "base," Gunter means the hypotenuse of the spherical triangle. Gunter's first example is a standard astronomical problem: find the sun's declination δ from its ecliptic longitude λ. The solution is $\sin\delta = \sin\lambda \sin\epsilon$ (where $\epsilon = 23°30'$ is the obliquity of the ecliptic), equivalent to the modern formula $\sin a = \sin A \sin c$ for a right triangle. But Gunter expresses it as $\frac{\text{Sin}\,90°}{\text{Sin}\,\epsilon} = \frac{\text{Sin}\,\lambda}{\text{Sin}\,\delta}$ for good reason.

The "line of sines" is the unequally marked scale near the middle in figure 1.28, ending at 90°, displayed on both arms of the sector. Set a compass

[147] [Gunter 1624, 76].

along the line of sines so that the distance between the two tips is equal to Sin 23°30′, that is, the distance along the line of sines from zero to 23°30′. Move the compass to the end of the sector, and spread the sector's arms so that their ends touch both ends of the compass. Without changing the angle of the pivot, move the compass inward (narrowing the gap between its tips) so that its two ends touch the two locations on the sector corresponding to 30°. We now have two similar triangles; from them, we have $\sin 90°/\sin 23°30' = \sin 30°/x$, where x is the new distance between the compass tips. From Gunter's ratio above, we know that x is equal to $\sin \delta$. Move the compass so that one of its tips is at the pivot. On the line of sines, the point corresponding to the other tip (11°30′) is δ.[148]

C4 @ 5B 0)* A 4)

C4 5 0 D 4

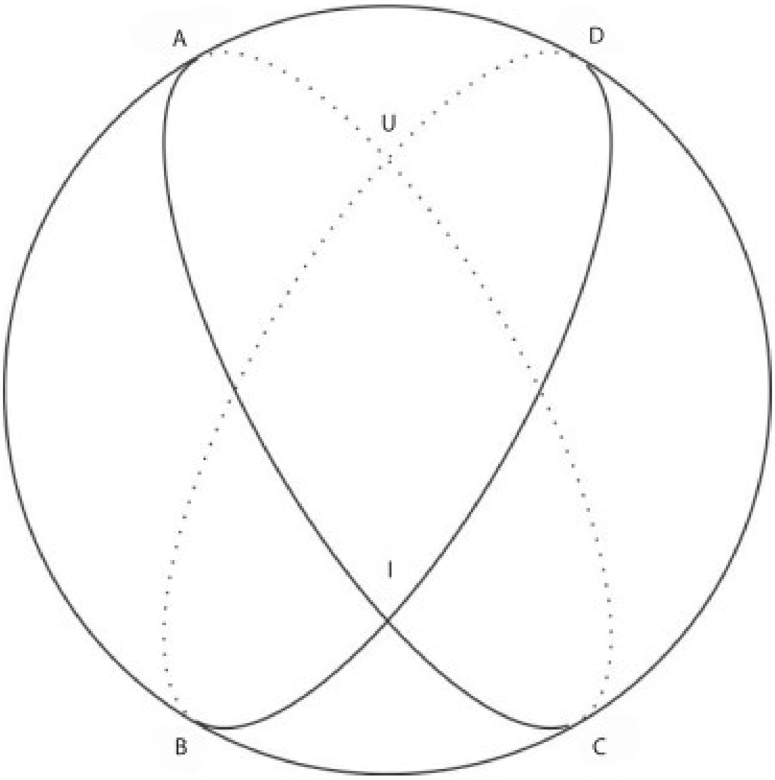

D 4)&) N +)))) *)) 4

$ -& % & (+)/ *) 4 /:))) /
D +)) $ # 4 :)) + &) &) &) 4 :))
5 + 4)**)) J) 0

C4 5 @ 0)* > 7 *) 7

C4 > 5 @0 D 4 > >
N 4)* * F *)) 4 #' *)) 4
#M'M M

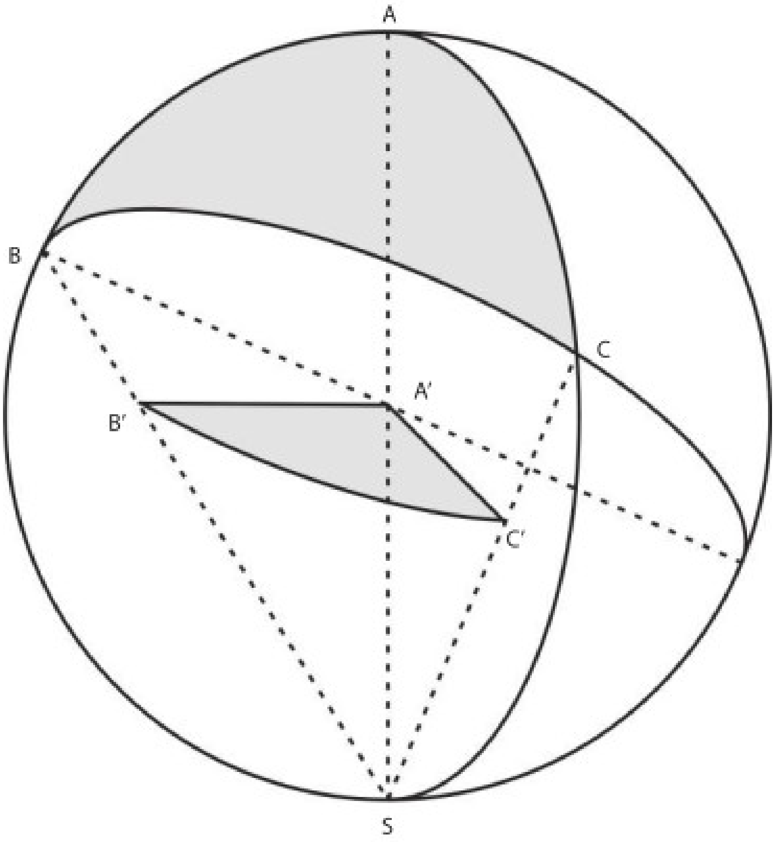

C4 >B5 0 D 4 > @

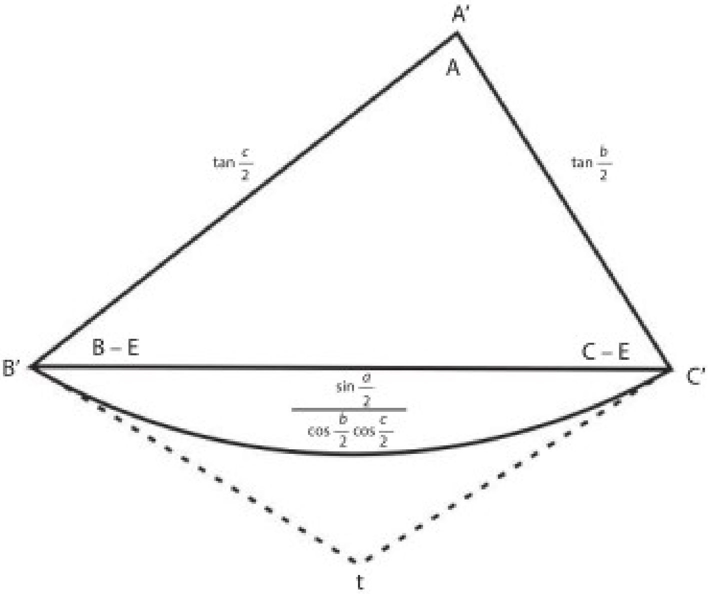
A′
A
$\tan\frac{c}{2}$
$\tan\frac{b}{2}$
B – E
C – E
B′
C′
$\frac{\sin\frac{a}{2}}{\cos\frac{b}{2}\cos\frac{c}{2}}$
t

Figure 5.6
Serret's plane triangle. Multiply each of the three sides by $\cos\frac{b}{2}\cos\frac{c}{2}$, and we have Cesàro's triangle of elements.

values for the sides and angles. He begins by applying the planar Law of Cosines:

$$a'^2 = b'^2 + c'^2 - 2b'c' \cos A.$$

Substituting the values of (5.12) into this equation and applying the sine half-angle formula to the terms in front of cos A, he gets

$$\sin^2\frac{a}{2} = \sin^2\frac{b}{2}\cos^2\frac{c}{2} + \sin^2\frac{c}{2}\cos^2\frac{b}{2} - \frac{\sin b \sin c}{2}\cos A.$$

From here, applying some more algebraic manipulation (replacing $\cos^2$ terms with $1 - \sin^2$ terms), he arrives at

$$1 - 2\sin^2\frac{a}{2} = \left(1 - 2\sin^2\frac{b}{2}\right)\left(1 - 2\sin^2\frac{c}{2}\right) + \sin b \ \sin c \ \cos A.$$

Finally, applying the cosine half-angle formula, he has the spherical Law of Cosines:

$$\cos a = \cos b \cos c + \sin b \sin c \cos A. \tag{5.13}$$

C4 @ 5 B 0 D 4 >

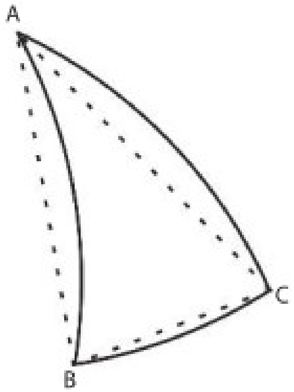

D 4 >
*)) 4) + +)) 4
3:
$ + + *) +)) 4
3 * /))) 4 4 +) *)) 4
) +) +) + *) + * 4) 4) +
4 - /, * ' *)) :) +
)+)) +)) 4 N +))
) & * ,) /) $ +) + +)
) 4 - (4 + :) + *)
) 4) +) 4 # : ,)/)))** 8) /
) +
O
: #) + * + 4 *))) 4
) 4 4 &) 4 5 4 > 0 $
)/ :/*) *) + *) +)& * *))
J &) $),

$$\theta \approx \tan\frac{A}{2}\sin^2\frac{b+c}{4} - \cot\frac{A}{2}\sin^2\frac{b-c}{4},$$

:) + :/)** / 4 *) A))) *
+) # $)** 8))** / 4 +
) J) / 5# P
OQ
0) + *) 4
O
) +
O

) +
O
* & /
$ + +, * *) ,) A 4 + N
$ $) + * / 4 &) /) *)
) 4) 4 #, ',) +) +) *)) 4 + 4 J)
) *)) 4) +) 4 #M, 'M) + M,)** 8
) /)

$$A = A' + E/3,\ B = B' + E/3, \text{ and } C = C' + E/3,$$

7 *) 8 > *)) 4 $) 4) /
) + & / N *)) : + ,)(4) (*)):
) , +) 4) 4) , / / +
*) *) 4 /) +) & +) #, ',) + :/
)** 4 *) 8 J) /) 4 @ A 4 + N
)) /))) + +)) 4 ,
:)) 8
$)) + :/ #+) A 4 + B
) & & +)) 4) & / +) & *
.) +) : &) , * +) +
+ / & / 4 : D) + 7 4
)) + + + 4 # *))
B) *) * D +))
))) / / (*) $ 8
: (#:)) RS + + : &) + +)+
) / : *) , B) + H) R) ' ( ) +) +))
J)** 8) 4) * : ; & ,) +
) + * & , &)))) 8 + +
* &)) /, :/) 4 : +) 4 *)
) 4 , & * + 3) / * &) +
R) ' 9 4 4 , B B * & +)** 8
)))+ + A 4 + N $ 4)& 4

$$A = A' + \frac{E}{3} + \frac{E}{180}(b^2 + c^2 - 2a^2),$$

$$B = B' + \frac{E}{3} + \frac{E}{180}(a^2 + c^2 - 2b^2),$$

$$C = C' + \frac{E}{3} + \frac{E}{180}(a^2 + b^2 - 2c^2).$$

C4 @@5 B>0

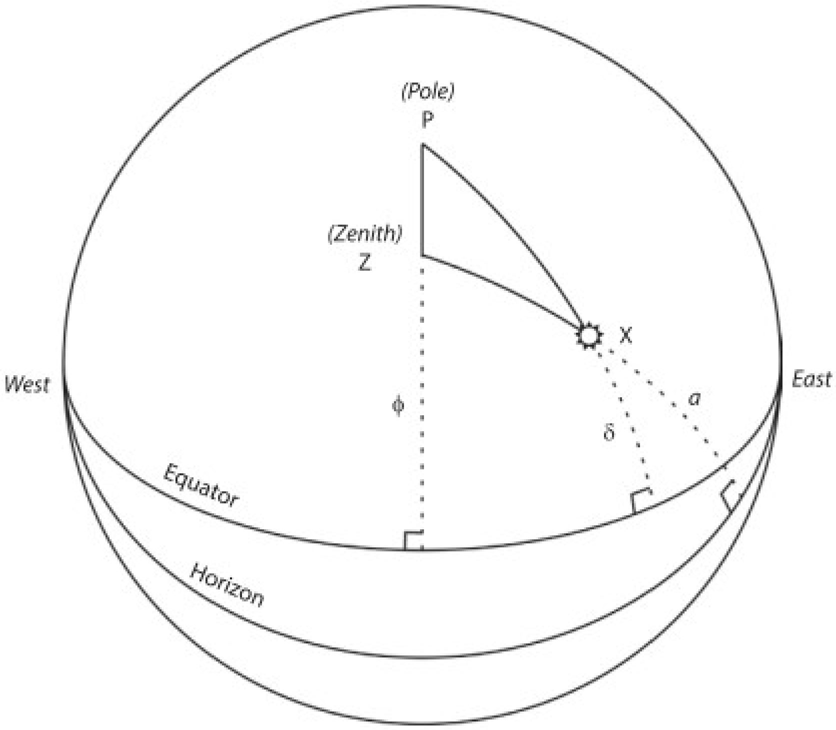
(Pole)
P
(Zenith)
Z
X
West
East
a
ϕ
δ
Equator
Horizon

D 4 >
$)) TU) 4

0 !#12! ! !3 4) ! .) ' $
" $) 4 # ; / + $ 4 / E & / 5 0 *+

0 !#12! ! !3 4) ! - + . / %
" *+

0 !#12! ! !3 4) !))) +) I)
R 3)/ * 4 5 >0 *+
C4 5> 0)*)) , -) + %)& 4) @ B

C4 B 5 0)* > $))) + A) + 7 5 B 0

C4 5 > 0)* ; :) + $ / D :

0 !#12! ! !3 4) !) ;) & / . /
-) + # -) + - + 7 +)) + ;/* :) . / #)
))) / 5 >0 +F&
< >@ *)4

>0 !#12! ! !3 4) !) : . /
*+
< 1 #+&) D +) 7 +) . /

@0 !#12! ! !3 4) ! 1 #+&) D +)
7 +)) . / *+

0 !#12! ! !3 4) ! $) + D +)
7 +) . / *+

B0 !#12! ! !3 4) ! $ D +) . /) +
% 7 +)) *+
< > > *)4
3: A & + ===

C4 5 B0)* @))) + E) #8

0 !#12! ! !3 4) ! $)) . / *+
3: A & + ===
< > *)4

C4 B 5 0)* *)) + / /

0 !#12! ! !3 4) ! $)))) . /
)*) V ;) 6 1 5 0 *+
3: A & + ===
< > *)4

C4 5 > 0)* / /) D 4

0 !#12! ! !3 4) !) ' , . ;)& /
)) E & / 5 0 *+

!"#$# % & & !' (! !) . /

0 !#12! ! !3 4) ! @> 5 0 >@ * *+
< D 4 /4) D)) W

0 !#12! ! !3 4) ! > @@B *+
< $ D +) . /
< @B *)4

0 !#12! ! !3 4) ! #))) + + 4 /
* 4)4 % X (5 0 *+
3: A & + ===
< > *)4
< / /) + $))

>0 !#12! ! !3 4) ! #) 5) 0)) + +
4 / * 4)4 % X (5 0 *+
< @ *)4

@0 !#12! ! !3 4) ! #** + 8 7) . / *+
<) . /

0 !#12! ! !3 4) ! "
BB6# 6 /4 9) 3 - $ *+

B0 !#12! ! !3 4) ! ; : 4 3 +) + & +
* : *) 4 /) + : /))) #) #) 5 0 +F&
3: A & + ===
< A
< J) 4)
<)
< : *) +

0 !#12! ! !3 4) ! ; : 4 3 +) + & +
* : *) 4 /) + : /))) #) #) 5 0 *+

0 !#12! ! !3 4) !) ;) & / . /
-) +6 # -) + - + 7 +)) + ;/* :) . / #)
))) / 5 >0 +F&

0 !#12! ! !3 4) !) ;) & / . /
-) +6 # -) + - + 7 +)) + ;/* :) . /
))) #) #) 5 >0 *+

0 !#12! ! !3 4) !)) + + 4 /
)) 5 >0 *+
< > *)4

0 !#12! ! !3 4) ! *)) + +4 /*+

0 !#12! ! !3 4) ! *) 4 *+
< :)) + + . /) &
3: A & + === - *) ? ===
< > *)4

>0 !#12! ! !3 4) ! *+

@0 !#12! ! !3 4) !)) G A)) + +
. / ; +/ 5 0 +F&

0 !#12! ! !3 4) !)) G A)) + +
. / ; +/ 5 0 *+

B0 !#12! ! !3 4) !) /4 :)
. / 5 7& / 0 : 5 BB0 *+

0 !#12! ! !3 4) !) & " A 5)) 0
:)) + + . / &) . /*+

0 !#12! ! !3 4) ! G) D 3 4 +, G)) *)
.))) + + #) / . / $)) *) / 5 0 *+

!"#$# % & & !' (! !1) / G 1)) ,
$ 4 ,) + $ 4/ 2) /

0 !#12! ! !3 4) ! > 1) / G 1)) ,

$ 4 ,) + $ 4/ 2) / *+
0
!"#$# % & & !' (! ! *) 1) & /

0 !#12! ! !3 4) ! @ *) 1) & / *+
< . /
< *)
< .)) 1) & /
3: A & + ===

!"#$# % & & !' (! !$)) ' (Y # /)
" - =

0 !#12! ! !3 4) ! $)) ' (# /)
" - = *+
< B *)4

!"#$# % & & !' (! !E +) + 4 / H77)) +
#+&) + 7 /) +)4

0 !#12! ! !3 4) ! B E +) + 4 / H77
)) + #+&) + 7 /) +)4 *+

!"#$# % & & !' (! !) ") . /) + D
))) ")) D &

>0 !#12! ! !3 4) ! $) % +)) * 8
) / 38 + E & / , E # 5 0 *+
3: A & + ===
< @ *)4

C4 -U 5 @0

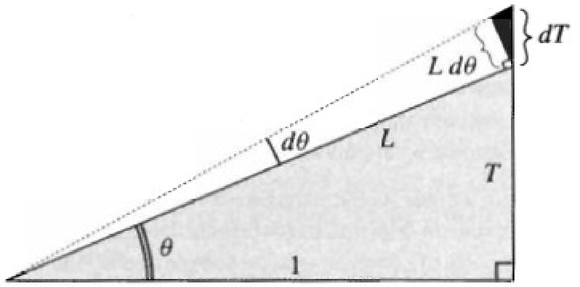

$$\frac{dT}{L\,d\theta} = \frac{L}{1} \quad \Longrightarrow \quad \frac{dT}{d\theta} = L^2 = 1 + T^2.$$

@0 !#12! ! !3 4) ! $) % +))
") . /) + D #))) ")) D & # E & /
5 *+
3: A & + === - *) ? ===
< > *)4

0 !#12! ! !3 4) ! $) % +))

") . /) + D #))) ")) D & # E & /
5 0 *+
!"#$# % & & !' (!

!"#$# % & & !' (! !) & # 4 :)

B0 !#12! ! !3 4) ! 6 B B @ *+

!"#$# % & & !' (! !-) & ")) 4) + "
) 9) + 7)

0 !#12! ! !3 4) !
6' (6-) & ")) 4# + " *+

!"#$# % & & !' (! !A) / 8 + / *)

> 0 !#12! ! !3 4) ! ") . / &
)) + , $ + 7+ #)))) / 5 >0 *+
3: A & + ===
< B *)4

> 0 !#12! ! !3 4) ! ") 4 / & ,
) ,) + #)))) / 5 >0 *+

> 0 !#12! ! !3 4) ! ; A) / 8 + /
*) * 4 5 0 *+
3: A & + ===
< @ *)4
C4)* / *)

/ *)
; 4 *)
- &))
*) ; 4 *) B
* 8 ; 4 *)
F & 7 : ++ 4
> % *) ; 4 *)
/ *)
. :) / / *)
- B
")
A) / / *)
> 78) *
@ 1)) / *)
) Z) / *) >>
/ A # 4 :) >>
/ *) @>
/ :*)) + $) / . + : *)
> ;) / *)
> *) ;) / *)
> % *) ;) / *) B@
> $ 78 *) ") @

>)/ / $)
> > ' +) / *
> @ #** + 8 4 ") B
@ 78) *
@ $) [ )
@ ;/* : *) >
@ / *) # 4 78 *)
. *
@ / *) 1) + E 5 0

@ > ;) / *) .)) $/* >
@ @ F &)
))(*) Z)
*) Z)
' *) Z)
$ *) Z) \ A 5 0 E 5 0
))(*) Z) >
B $ /) + / *) @
B .) 9) $ / @
B #**) $ / / *) @
B $ *) A * @
A) / / *)
2 B
): / B
) Z) # . * 578) * 0 B>
1)) ' +) / *
$.)] ' #
2
*) Z) # 2
' *) Z)
))(*) Z) @
A) / ;) / *)
1)) ' +) / * >
') / ' 7 : ++ 4 @
$ +) *) Z) A) / ;)
/) B
> $ *) / *
> ; 9 : *) / % *))
@ A) / / : *) K $) / . + : *)
@ . + /
@ % &) 4 5 0; 4/
@ 1) & *) / @
78) * >
*) " & 4 . D >
$) [ ) >
;/* : D + @B
) + +)) 5# 2
* 8 ;/* :) + 0 @
> ;/* : ") BB
@ # 2 ;) /
) .)) $/
H) A (# 4 :))

B A) / * / *) @

A) / 8 + / *)
8 + / *)
8 + /)
* 8 + /)
78 8 + / *)
*) Z)
A) / 8 + / *)
D :
78) *
78) * " & 4 . D
78) * # 4 78 *) . * >
A) / 8 + / *)) + *) Z) >B
A) + ' *) Z) >B
7 : ++ 4 A) / / *) A) 4
3 @
> . :) @>

> 0 !#12! ! !3 4) ! ; 4 *) + 4
H;E 5 0 *+
< 4) + A 4
< $ 4/) + / H) G) # : 7

(+4 &
- + # G + $) +
)) + G (
7 4/) + 7 */ >
$ R $ /)) + 3 + @
7 /) + % (@B
> 7)4 7) + D +
@ 7) + A 4
-) % /
7* 4 7)) 3 @
44 + D 1)+ 4 @
- + 8

> 0 !#12! ! !3 4) ! $. / *)
2 * 4)4 ' ; + : 4 5 @0 +F&

>>0 !#12! ! !3 4) ! $. / *)
2 * 4)4 ' ; + : 4 5 @0 *+

!"#$# % & & !' (! !3 4) D))) +
#**)

>@0 !#12! ! !3 4) !
3 4) 6D 6)) 6) +6#**) *+
< @ *)4

!"#$# % & & !' (! ! ' $ $. ,
)),) + ' 4 ') /
> 0 !#12! ! !3 4) ! ' $ $. ,
)),) + ' 4 ') *+

>B0 !#12! ! !3 4) ! ' $ $. ,
)),) + ' 4 ') / * :

!"#$# % & & !' (! ! & / . / - :/ #))
)*)+ *

> 0 !#12! ! !3 4) ! > 6 B B @
*+
<) & # 4 :)
< B B *)4

@ 0 !#12! ! !3 4) ! > 6 B B@@ > * :

@ 0 !#12! ! !3 4) ! > 6 B B@@ > *+
< & / . / -
< @ *)4

*) . /^# & / G +) + $ (
& 8 ' +

((. /^ *) + 1
" & *

3 4) / $/* % + A) *)

@ 0 !#12! ! !3 4) ! > # +/ " & + # 4 :)
. / ") , A $ /) + D) . / #)))) /
5 0 *+
< *)4

@ 0 !#12! ! !3 4) ! > # +/ " & + # 4 :)
. / * +) + ") / #)))) / 5 0 *+

@ 0 !#12! ! !3 4) ! > # +/ " & + # 4 :)
. /, - * +) + ") / #)))) / 5 0 *+

@>0 !#12! ! !3 4) ! > # +/ " & + # 4 :)
. /, - * +) + ") / #)))) / 5 0 *+

@@0 !#12! ! !3 4) ! > # +/ + & +) 4 :)
4 / --, ") , A /) +) 4 / #))))
*+

@ 0 !#12! ! !3 4) ! > # 4 :) 4 /
) + 4 #)))) / 5 0 +F&

@B0 !#12! ! !3 4) ! > # 4 :) 4 /

) + 4 #)))) / 5 0 *+

@ 0 !#12! ! !3 4) ! > +)* 1 D
-)) ' 5 0 +F&

0 !#12! ! !3 4) ! > +)* 1 D
-)) ' 5 0 *+

0 !#12! ! !3 4) ! > 8 / /) +
) 4 :) 4 / #))) / 5 0 +F&
< > *)4

0 !#12! ! !3 4) ! > 8 / /) +
) 4 :) 4 / #))) / 5 0 *+
3: A & + ===

0 !#12! ! !3 4) ! > D +))) 4 :) 4 /
. + (I D.# +F&

0 !#12! ! !3 4) ! > D +))) 4 :) 4 /
. + (I D.# *+

>0 !#12! ! !3 4) ! > . / # / * #)
))) / 5 0 *+
< *)4

@0 !#12! ! !3 4) ! > . / * & 8
: + #)))) / *+
< @ B *)4

0 !#12! ! !3 4) ! > -) . / * 4
5 0 *+
C4 5 0 @ $ G 4 + -) ")
@ $.)

B0 !#12! ! !3 4) ! >)))) * 4
/ G + : - 5 BB0 +F&

0 !#12! ! !3 4) ! >)))) * 4
/ G + : - 5 BB0 *+
< @@B *)4

B 0 !#12! ! !3 4) ! > / /) + # 4 :)
. / #))) / 5 0 +F&
< *)4

B 0 !#12! ! !3 4) ! > / /) + # 4 :)
. / #))) / 5 0 *+

B 0 !#12! ! !3 4) ! > / /) +) 4 :)
4 / #)))) / 5 0 +F&

B 0 !#12! ! !3 4) ! > / /) +) 4 :)
4 / #)))) / 5 0 *+
B 0 !#12! ! !3 4) ! > / /) +) 4 :)
4 / #)))) / 5 0 *+

B>0 !#12! ! !3 4) ! > / / -- #)
))) / 5 0 +F&
< B@ *)4
3: A & + ===

C4 > 5 0

--- 7) & -) + / / >>
)(4) & * + , :/) / /
:/ ")& + 1 >
+) ' + + *) >
$) 4)+ @
$)) + @
$ # + *)) *) @
D) +)) +) 4)+ @
)8) / * + / @@
7J &)) 4 : +) / *
> " 4)* 5 F 0
> # F): 4) 4/
> $) + &)* >
@)(4) & * +

B@0 !#12! ! !3 4) ! > / / -- #)
))) / 5 0 *+

B 0 !#12! ! !3 4) ! > / / #)
))) / 5 @0 +F&

BB0 !#12! ! !3 4) ! > / / #)
))) / 5 @0 *+
< > *)4

B 0 !#12! ! !3 4) ! > 1)) 4 :) 4 / * 4
5 B0 +F&

0 !#12! ! !3 4) ! > 1)) 4 :) 4 / * 4
5 B0 *+
< *)4

C4 5 0)* > $ $) (+ : 4 *)) $) $

0 !#12! ! !3 4) ! > 1 4 + . / &) + $
H) :) * 4 -) *+

1 4 + #) / &
D) D : B
. D)

* D @
D)): 1 +
>): 1 +
@ E &) & 4) &
)) 9) + & >

0 !#12! ! !3 4) ! > $ 4 /
+) 4 * * 4 5 B0 *+

>@ >

!"#$# % & & !' (!

!"#$# % & & !' (! !#+&))) +
)) --

0 !#12! ! !3 4) ! @ #+&) #** +))
) + . :) 3* 9) - ; . :) 4 * 4 *+
C4 5 0)*
)) &)) +
-) * 8 / 1 + +)
R ^ / :

0 !#12! ! !3 4) ! @ #+&))) +
)) - * 4 E 5 0 *+

>0 !#12! ! !3 4) ! @ #+&))) +
)) -- * 4 E 5 0 *+

@0 !#12! ! !3 4) ! @ *) /, ") /) +
/ / %)) + 4 -E$# / * * 4 %) +
5 0 *+

!"#$# % & & !' (! !# #: *) #) D :) /

0 !#12! ! !3 4) !
6# 6#: 6 *) 6#) *+
<) 4 4 7 I A)

C4 5 0 $ 4 7
;) * 4 :)(/ 4)) & /) 4) + : / +
(/ G)/

!"#$# % & & !' (! !# - + G)& #) /

B0 !#12! ! !3 4) ! B ")& + D G) # - +
G)& #) / ' (S ' 5 0 *+
< >> *)4

C4 @ --- 3) G)& ')

3) / $))

- +/)) /) +)) +)& :) , J /
) /)) 4))) 4 -)++) 4
4) * *) / * + , /)
)& *) * *) : &)):)& :) - : ,
* * *

0 !#12! ! !3 4) ! B ")& + D G) # - +
G)& #) / ' (S ' 5 0 +F&
C4 >5 0 #** + 8 #
1 & #+&) +)) + A) # 4 :)
.) / ') $ #+&) +)) + A) # 4 :)

!"#$# % & & !' (! !#) /) + . / &)
* 8)):

0 !#12! ! !3 4) ! #) /) + . / &)
* 8)): 5$ +)) 0 ' (S ' 5 0 *+
C4 5 0 # % + 1) 4 ' +) + $) 4)) /
1) 3*)

0 !#12! ! !3 4) ! #) /) + . / &)
* 8)): #))) / 5 0 *+
C4 5 0 *)) /* * 8 & Z +

- +
& 8 / 1
' 4 4 + * 8) ')) +)
) &)) 4 (*)N
D & * 8 (*)
> - *) * 8 (*) -
@ - *) * 8 (*) --
#)**) 3)) $)(4 8

C4 > 5 @ 0 %)) / 1) 4 /
3& &
%) A & + 4) 1) +
%) A & + 4) /*)
/ / * & 4)
> 3* * :
1

0 !#12! ! !3 4) ! * 8 #) / &)
* 8)):) + "7 $ /,) + . / *+
C4 @ 5 B>0 &): /)) * 8 D + #**) ")
)

!"#$# % & & !' (! ! #))) + X)

0 !#12! ! !3 4) ! #))) +

X) *+
< B B *)4
3: A & + ===

C4 5 0)*) D

!"#$# % & & !' (! !7 %) & . /

0 !#12! ! !3 4) ! 7 %) &
. / *+
C4 5 0)* - $ * 4/

C4 > 5 @ 0)* --)* $ %) & - 4)
< #)* + 1)) 4 /

C4 @ 5 @0)* --- . /
C4 @ 5 B0)*) & .

!"#$# % & & !' (! !D & 8 / % & 8 /

>0 !#12! ! !3 4) ! D & 8 / & 8 /
* 4 E +F&

@0 !#12! ! !3 4) ! D & 8 / % & 8 /
* 4 E *+
C4 5 0)* # & / %) $ /
- +
) * / *)
:+))
D)) &))
> D) J)+) + * + 4)+
@ #) /*)

!"#$# % & & !' (! !. + " ") +) +
78*) + +

0 !#12! ! !3 4) ! . + " *+
3: A & + ===

!"#$# % & & !' (! !. / &) +)
A7

B0 !#12! ! !3 4) !)+ 1 & (5) 0
. / &) +) # A7 ' (S ') 5 0 *+
3: A & + ===
C4 --)* >) & 1) 4) +)
< > G)) & _

5

Plane Curves in Rectangular Coordinates

There are many ways to classify curves. One of them is to think of curves as either *algebraic* or *transcendental*. An algebraic (plane) curve is given by a polynomial equation $P(x, y) = 0$. Its degree $n = \deg P$ is called the *order of the curve*. *Curves of order* $n = 2$ are studied in analytic geometry. The first classification of *curves of order* $n = 3$ was obtained by Newton. The case $n > 3$ is more difficult. But among easily obtained curves, there are many that are nonalgebraic, for example, the cycloid and spiral of Archimedes; we study them using parametrized or implicit equations (Chapter 5) or polar coordinates (see Chapter 6).

The names of curves, like the names of geographical objects, contain much interesting information. Let us group some curves by the meaning of their names:

By the name of a scientist: Dinostratus' quadratrix, conchoid of Nicomedes, Pascal's limaçon, lemniscate of Bernoulli, cissoid of Diocles.

By the method of construction: caustics, equidistant (parallel), pedal, evolutes, evolvents.

By an important property: trisectrices, quadratrices, tractrices.

By an essential property of their shape: astroid (star-shaped), deltoid (Greek letter Δ), cardioid (heart-shaped), nephroid (kidney-shaped).

By historical factors: ellipse, parabola, hyperbola.

By the structure of the formula: (semi)cubical parabola, logarithmic spiral, exponential and logarithmic curves, integral sine (cosine).

In Section 5.1 we consider the basic notion of a regular curve (for analogous definitions for surfaces, see Sections 19.1 and 19.2).

In Sections 5.2–5.4 we investigate cycloidal curves and other remarkable parametrized curves, and we provide tables with interesting figures.

In Section 5.5 we study implicitly given curves (appearing as level curves of functions in two variables); another family of curves is given by trajectories of vector fields (continued in Section 8.4 for trajectories in space).

In Section 5.6 level sets are used in solving problems with conditional extrema. Three approaches are compared: experimental, geometrical and analytical (using Lagrange multipliers).

In Chapter 5 the reader will become acquainted with the commands

```
solve, dsolve, numer, Diff, grad, allvalues;
implicitplot, contourplot, densityplot, fieldplot,
gradplot, odeplot, DEplot,  disk, ellipse, arc, line.
```

5.1 What Is a Curve?

Intuitively, from Euclid, a *curve* is the trajectory of a moving point, a boundary of a surface, a one-dimensional figure. The mathematically correct definition of a curve is based on notions from *topology*, but it starts with the key notion of an *elementary curve*, which can be imagined as an *interval* $I = (a, b)$ (or a *segment* $\bar{I} = [a, b]$) of the straight line after continuous deformation.

Definition 5.1.1 A set γ in $\mathbb{R}^2$ is called a *curve* if it can be covered by a finite or countable number of elementary curves.

One can distinguish *self-intersecting, simple* (i.e., without self-intersections; for example, graphs), *closed, connected* curves.

If we fix rectangular coordinates (with orthonormal base $\{\mathbf{i}, \mathbf{j}\}$) in the plane $\mathbb{R}^2$ with origin O, then the coordinates x, y of a point on the elementary curve γ given by the map $\vec{\mathbf{r}}(t) = [x(t), y(t)]$ (a *parametrization*) are functions of $t \in I$:

$$x = x(t), \quad y = y(t). \tag{5.1}$$

In other words, the position vector $\vec{\mathbf{r}} = \vec{OP}$ of the point $P(x, y)$ on the curve γ is the following vector-valued function:

$$\vec{\mathbf{r}} = \vec{\mathbf{r}}(t) = x(t)\,\mathbf{i} + y(t)\,\mathbf{j}, \qquad t \in I. \tag{5.1 a}$$

We define functions $x(t), y(t)$ and then use the following command:

```
> r:=[x(t), y(t)];
```

For example, a straight line in a plane is given by a linear vector-valued function in the variable t: $\vec{\mathbf{r}}(t) = [a_1 t + b_1,\ a_2 t + b_2]$.

Remark 5.1.1 We can derive in MAPLE some calculus type of operations on vector-valued functions $\vec{\mathbf{r}}(t) = [x(t),\ y(t)]$ of the class C^k, $(k \geq 0)$.

The *limit* $\vec{\mathbf{r}}(t_0) = \lim\limits_{t \to t_0} \vec{\mathbf{r}}(t)$,

```
> r0:=[subs({t=t0}, x(t)), subs({t=t0}, y(t))];
```

The *first derivative* $\vec{\mathbf{r}}\,'(t) = \lim\limits_{\Delta t \to 0} \frac{\vec{\mathbf{r}}(t+\Delta t)-\vec{\mathbf{r}}(t)}{\Delta t}$.

```
> rt:=[Diff(x(t),t), Diff(y(t),t)];
```

Analogously, we calculate higher order derivatives $\vec{\mathbf{r}}\,''$, $\vec{\mathbf{r}}^{\,(3)}$, etc. and *Taylor series* (see Section 1.1).

Definition 5.1.2 A point P on a simple curve $\gamma \in \mathbb{R}^2$ is called *regular of the class* C^k (C^∞ or C^ω) if some neighborhood of this point admits a parametrization $\vec{\mathbf{r}}(t) = [x(t),\ y(t)]$, $(t \in I)$, where the functions $x(t)$ and $y(t)$ belong to the class C^k (C^∞ or C^ω) and the vector $\vec{\mathbf{r}}\,'(t)$ is nonzero at P. For the opposite case, a point P on a curve γ is called a *singular point* (see Fig. 1.14 with a tractrix). A curve consisting of regular points is called a *regular curve of the class* C^k (C^∞ or C^ω) (*smooth* when $k = 1$).

Definitions 5.1.1 and 5.1.2 can be generalized for the space curves studied in Chapter 8.

5.2 Plotting Cycloidal Curves

The trajectory of a point on the circle of radius R traveling (without sliding) along another circle of radius $R\,'$ (or along a straight line) is called a *cycloidal curve* (Greek, *κυκλο ειδης*, circle-shaped). If the circle drives along and inside of motionless circle, then such a curve is called a *hypocycloid*; if outside, then the curve is called an *epicycloid*. If we follow a point not on the border of the wheel but on its spoke or on the spoke's continuation, then we obtain a *curtate* or *prolate trochoid* (Greek, *τροχο ειδης*, wheel-shaped).

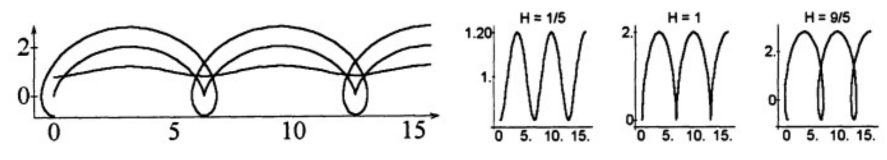

Figs. 5.1–5.2. Common cycloid, curtate and prolate cycloids

8

Space Curves

Section 8.1 is a short introduction to the basic `MAPLE` support for plotting space curves using various coordinate systems. Section 8.2 is devoted to helix type curves lying on a cylinder, a sphere, a torus and other surfaces of revolution. In Section 8.3 we study remarkable `MAPLE` strategies for producing splendid views (of three-dimensional objects) with coloring, shadows and moving. We plot curves that are intersections of two surfaces: the Viviani curve, bicylindrical curves, and solutions of Euler's equations from mechanics. In the final Section 8.4, we deal with vector fields and their trajectories in space (see Section 5.4 for two-dimensional case).

In Chapter 8 the reader will become acquainted with the commands

`spacecurve, tubeplot, plot3d, display3d, DEplot3d, textplot3d, fieldplot3d, gradplot3d, polygonplot3d, PLOT3D, ANIMATE;`

`det, cosh, sinh, JacobiCN, JacobiSN, JacobiDN, transform.`

We begin using the options of 3D plotting:

`light, ambientlight, projection, orientation.`

8.1 Introduction

8.1.1 Commands `spacecurve` *and* `tubeplot`

A space curve $\vec{\mathbf{r}}(t) = [x(t),\ y(t),\ z(t)]$ is plotted using one of the following commands from the library `plots`:

```
spacecurve(one or several curves, options);
```

```
tubeplot(one or several curves, options);
```

Their syntax is similar; the difference is that `tubeplot` plots a *tubular surface* around the given curve, so the options with lighting are effective.

Example 8.1.1

1. Let us show that the curve $\vec{\mathbf{r}}\,(t) = [1+3t+2t^2,\ 2-2t-4t^2,\ 1-t^2]$ lies on a plane and describe it.

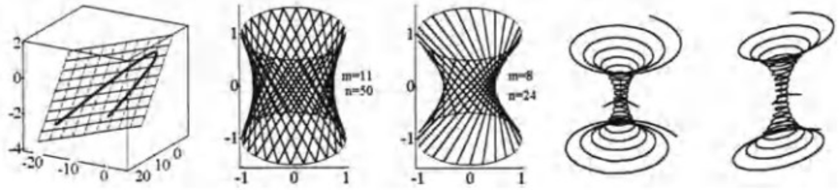

Figs. 8.1–8.5. Plane curve in space. Hyperboloid. Curve on a catenoid

Hints. Note that the curve $\vec{\mathbf{r}}\,(t) = \vec{\mathbf{r}}_1 + t\,\vec{\mathbf{r}}_2 + t^2\,\vec{\mathbf{r}}_3$ lies in a plane if the three vectors $\vec{\mathbf{r}}_1, \vec{\mathbf{r}}_2, \vec{\mathbf{r}}_3$ are coplanar, i.e., the determinant consisting of their coordinates vanishes. Any two (linearly independent) vectors from among them define the plane. In our case

```
> r1:=[1, 2, 1]: r2:=[3, -2, 0]: r3:=[2, -4, -1]:
> linalg[det]([r1, r2, r3]); with(plots):
  0
> A:=spacecurve(evalm(r1+t*r2+t^2*r3),t=-2..2,thickness=3):
> B:=plot3d(evalm(u*r1+v*r2), u=-4..2, v=-1..7):
> display([A, B], orientation=[30,70]);  # Fig. 8.1
```

2. Using cycloidal curves (close plane curves) we obtain three-dimensional curves: astroid ($m = -\frac{1}{4}$), deltoid ($m = -\frac{1}{3}$), nephroid ($m = \frac{1}{2}$), and three-dimensional roses.

```
> m:=-1/4: h:=1:  # 3–dimensional astroid, then enter your data
  > r3:=[(1+m)*cos(m*t)-h*m*cos(t+m*t),
  (1+m)*sin(m*t)-h*m*sin(t+m*t),cos(t)]:
  plots[spacecurve](r3,t=0..2*Pi/abs(m),axes=boxed);
```

3. We plot *two engaged circles* (tori).

```
> plots[tubeplot]({[cos(t),sin(t),0],[0,cos(t)+.8,sin(t)]},
  t=0..2*Pi, radius=.25, tubepoints=40, orientation=[0, 60],
  lightmodel=light2, style=patchnogrid);# Fig. 8.6
```

Then we generalize this for a *chain* containing 2*n* circles; Fig. 8.7.

Figs. 8.6–8.8. Two engaged circles. Short chain. Chain armor

```
> n:=3: P:=i -> plots[tubeplot]({[2*cos(t),2*sin(t)+6*i,0],
  [0,2*cos(t)+3+6*i,2*sin(t)]}, t=0..2*Pi, radius=.4,
  orientation=[0,60],style=patchnogrid, lightmodel=light4):
> plots[display]([seq(P(i),i=0..n-1)],scaling=constrained);
```

Finally we generalize it for the *chain armor* with m lines; Fig. 8.8.

```
> with(plots): n:=4: m:=4: P:=(i,j)-> tubeplot({[2*cos(t)+6*j,
  2*sin(t)+6*i,0], [6*j,2*cos(t)+3+6*i,2*sin(t)], [2*cos(t)-
  3+6*i,0+6*j,2*sin(t)]},t=0..2*Pi,radius=.4, orientation=
  [40,50],style=patchnogrid,lightmodel=light4): display([(
  seq(seq(P(i,j), i=0..n), j=0..m))], scaling=constrained);
```

8.1.2 Intuitive Projection

How is a space curve $\vec{\mathbf{r}}(t) = [x(t), y(t), z(t)]$ (or more generally, any space figure) plotted on a planar computer display? Sometimes one creates the image by intuition, as in the following *method of plotting the hyperboloid of one sheet*, using our knowledge of a *star-shaped polygon*.

The coordinates of the sequence of vertices of the star-shaped (n, m) polygon are calculated by the formulas $[R\cos(2\pi i \frac{m}{n}),\ R\sin(2\pi i \frac{m}{n})]$ $(1 \le i \le n)$; see Section 6.1. We transform the circumscribed circle of this polygon to the ellipse and raise and lower it (in the plane) a distance of ± 1 and obtain two star-shaped polygons. Now we connect the ith vertex of one of them with the $(i+1)$th vertex of another, and obtain a part of the hyperboloid of one sheet. We plot it by two methods.

```
> m:=11: n:=50: T:=i -> i*2*m*Pi/n: # first method; Fig. 8.2
> p1:=plot([seq([cos(T(i)),.5*sin(T(i))+(-1)^i],i=1..m*n)]):
  p2:=plot([[cos(t), 0.5*sin(t)+1, t=0..2*Pi],
  [cos(t), 0.5*sin(t)-1, t=0..2*Pi]]):
  plots[display]([p1, p2], scaling=constrained);
```

```
> m:=8: n:=3*m: T:=i-> i*2*Pi/n: # second method; Fig. 8.3
> p1:=plot([seq([[cos(T(i)), 0.5*sin(T(i))-1],
  [cos(T(i+m)), 0.5*sin(T(i+m))+1]], i=1..n)]):
  p2:=plot([[cos(t),.5*sin(t)+1,t=0..2*Pi],
  [cos(t),.5*sin(t)-1,t=0..2*Pi]]):
  plots[display]([p1, p2], scaling=constrained);
```

8.1.3 Standard Coordinate Systems in Space

Other than `rectangular` coordinates, there are two especially popular methods for fixing the place of the point P in the space.

(a) Consider the *height* z as the projection of the segment OP onto the axis OZ; the *distance* $|OP'| = \rho$, where P' is the projection of the point P onto the plane XY; and the *angle* $\varphi \in [-\pi, \pi]$ between the segment OP' and the axis OX. The triple of numbers $(\rho,\ \varphi,\ z)$ is called the *cylindrical coordinates* of the point P. The relation with the rectangular coordinates (x, y, z) is analogous to the case of the polar coordinates:

$$x = \rho\cos\varphi, \quad y = \rho\sin\varphi, \quad z = z.$$

Coordinate surfaces are cylinders $\rho = const$, half-planes $\varphi =$ const through the axis OZ, and also horizontal planes $z =$ const.

(b) Consider the *distance* $|OP| = \rho$, the (*vertical*) *angle* $\phi \in [-\pi, \pi]$ between the segment OP and axis OZ and the (*horizontal*) *angle* $\theta \in [0, \pi]$ between the segment OP' and the axis OX. The triple of numbers $(\rho,\ \theta,\ \phi)$ is called the *spherical coordinates* of the point P. The relation with rectangular coordinates is the following:

$$x = \rho\cos\theta\sin\phi,\ y = \rho\sin\theta\sin\phi,\ z = \rho\cos\phi.$$

Coordinate surfaces are spheres $\rho =$ const with centers at O, half-planes $\theta =$ const through the axis OZ, and circular cones $\phi =$ const with the common axis OZ.

Plotting curves in cylindrical or spherical coordinates is possible by the standard command `spacecurve` with the additional option `coords= cylindrical` or `coords= spherical`.

In addition to the above three, in `MAPLE` there are about thirty systems of curvilinear coordinates in $\mathbb{R}^3$ (see Section 19.3.2), and we can define a number of new coordinate systems using the command `addcoords`.

C4 5 0 -) # A7
)*) *)
G))) _

19

Surfaces in Space

In Sections 19.1 and 19.2 we consider some basic notions of a parametrized surface and a regular surface (analogous definitions for curves were studied in Section 5.1). In Section 19.3 we use a number of `MAPLE` commands to produce surfaces by various methods. In Section 19.4 we calculate and plot tangent planes and normal vectors of a surface. As an application we solve the conditional extremum problems in space (see the two-dimensional case in Section 5.6). In Section 19.5 we use changes in coordinates and linear transformations in space to calculate and plot an osculating paraboloid at the point of a surface. This elementary approach is given only for methodical reasons. In Section 19.6 we consider parametrized and implicitly defined surfaces with singularities.

In Chapter 19 the reader will become acquainted with the commands

```
implicitplot3d, coordplot3d, plot3d, display3d, surfdata,
complexplot3d, listplot3d, matrixplot, PDEplot,
hilbert, toeplitz, multiply, transpose, coeftayl, mtaylor.
```

19.1 What Is a Surface?

Euclid defines a *surface* intuitively, as a two-dimensional figure swept by the path of a moving curve, bounding a solid body.

Planes, some polyhedra, and also some curved surfaces, each defined in some practical way, are studied in school. The mathematically correct definition of a surface is based on notions from *topology*, but it starts from the key notion of an *elementary surface*, which can be imagined as an *elementary do-*

main of the plane after continuous deformation that is stretched in and out in space.

Definition 19.1.1 The set $G \subset \mathbb{R}^n$ ($n = 2, 3$) is called an *open set* if for each point $P \in G$ there exists $\varepsilon > 0$ such that the ball $B(P, \varepsilon)$ of radius ε with center P (the disk, when $n = 2$) lies inside of G.

The complement $F = \mathbb{R}^3 \setminus G$ to the open set G is called a *closed set*.

A *connected* open set (i.e., one that is not the union of two non-empty and non-intersecting open sets) is called a *domain*.

Domains are convenient for working with continuous and differentiable (vector) functions defined on such sets. An arbitrary open set is the union of a finite or countable number of disjoint domains.

Definition 19.1.2 A domain in the plane (in space) that is homeomorphic to a disk (respectively, a ball) is called an *elementary domain*.

The following two elementary domains in the plane with coordinates u, v are often used: the disc $u^2 + v^2 < R^2$ of radius R and the rectangle with sides a, b parallel to the coordinate axes.

As in the definition of an elementary curve, we use the notion of a *topological map* (*homeomorphism*) from one geometric figure to another, i.e., a one-to-one map that is continuous and has a continuous inverse.

Definition 19.1.3 A set M in space is called an *elementary surface* if it is the image of a planar elementary domain G under a topological map $\vec{\mathbf{r}} : G \to \mathbb{R}^3$.

If we fix rectangular coordinates (with the orthonormal base $\{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$) with origin O in space $\mathbb{R}^3$, then the coordinates x, y, z of a point on an elementary surface given by the map $\vec{\mathbf{r}}$ are functions of the coordinates u, v of a point in the domain G:

$$x = x(u, v), \quad y = y(u, v), \quad z = z(u, v). \tag{19.1}$$

In other words, the position vector $\vec{\mathbf{r}} = \vec{OP}$ of the point $P(x, y, z)$ on the surface M is the following vector-valued function:

$$\vec{\mathbf{r}} = \vec{\mathbf{r}}(u, v) = x(u, v)\,\mathbf{i} + y(u, v)\,\mathbf{j} + z(u, v)\,\mathbf{k}, \qquad (u, v) \in G. \tag{19.1 a}$$

We define functions $x(u, v)$, $y(u, v)$, $z(u, v)$ and then use the following command in `MAPLE`:

```
> r:=[x(u,v), y(u,v), z(u,v)];
```

If the initial point of the continuous vector $\vec{\mathbf{r}}(u, v)$ coincides with the origin O, then its endpoint, as a rule, moves along a *surface* in $\mathbb{R}^3$ called the *hodograph* of the vector-valued function. For example, a plane in space is the hodograph of a linear vector-valued function in the variables $u,\ v$:

$$\vec{\mathbf{r}}(u, v) = [a_1u + b_1v + c_1,\ a_2u + b_2v + c_2,\ a_3u + b_3v + c_3].$$

Definition 19.1.4 The system of equalities (19.1) or the vector equation (19.1 a), representing a map $\vec{\mathbf{r}} : G \to \mathbb{R}^3$ from the planar domain G into $\mathbb{R}^3$ is called the *parametric equations of the surface* $M = \vec{\mathbf{r}}\,(G)$: the pair of real numbers (u, v) is called the *curvilinear coordinates* of the point $P(x, y, z)$ on the surface.

Let us first take $v = v_0$ as a constant and $u = t$ as varying. Then we obtain the coordinate *u-curve* on the surface M, which also is the space curve $\vec{\mathbf{r}}(t, v_0)$. Similarly, for a constant $u = u_0$ and $v = t$ varying, we get the *v-curve* on the surface M, that is, the space curve $\vec{\mathbf{r}}(u_0, t)$. These two families of curves, (v-curves and u-curves), form the *coordinate net* of the surface, Fig. 19.1.

An arbitrary *curve* γ *on the surface* M is defined locally by equations for the curvilinear coordinates $u = u(t),\ v = v(t)\ (t \in I)$. The equations for t γ have the following form in space:

$$\tilde{x}(t) = x(u(t), v(t)), \quad \tilde{y}(t) = y(u(t), v(t)), \quad \tilde{z}(t) = z(u(t), v(t)).$$

```
> subs({u=u(t), v=v(t)}, r);
```

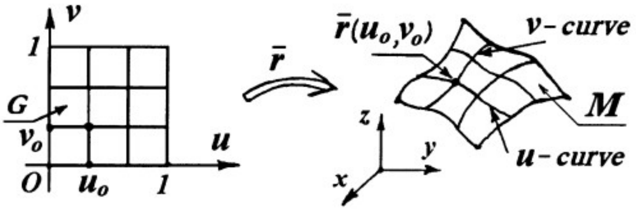

Fig. 19.1. Parametrized (elementary) surface

Since we will also be interested in *self-intersecting* surfaces, we give the following definitions:

Definition 19.1.5 A set M in $\mathbb{R}^3$ is called a *surface* if it can be covered by a finite or countable number of elementary surfaces.

Definition 19.1.6 A surface M is called a *simple surface* if it has no points of self-intersection, i.e., for any point $P \in M$ there exists $\varepsilon > 0$ such that the set $U(P, \varepsilon) = B(P, \varepsilon) \cap M$ (the intersection of M with the ball of the radius ε) is an elementary surface.

An elementary surface is a simple one. Examples of surfaces that are simple but not elementary are the following: the *sphere*, Fig. 20.7 (which can be covered by two disks); the *circular cylinder*, Fig. 19.5 (which can be covered by two cylindrical strips, each homeomorphic to a rectangle); the *torus*, Fig. 19.3 (which can be covered by four rectangles). The following surfaces are not simple: the *Klein bottle*, Fig. 19.7 (whose points of self-intersection form a circle); the *cone*, Fig. 19.15 (whose vertex is a *singular point*); a *pair of intersecting planes* (the self-intersection points form a straight line); the *pinched torus*, Fig. 19.4 (whose self-intersection points form two line segments). A more complicated example of a surface is the *union of a countable number of cylinders with the common generatrix* OZ: $M = \bigcup_{n\in\mathbb{N}} \{x^2 + (y-n)^2 = n^2\}$. It cannot be covered by a finite number of elementary surfaces. Note that for *local studies* of surfaces (i.e., in a neighborhood of some point) we need only the notion of an elementary surface.

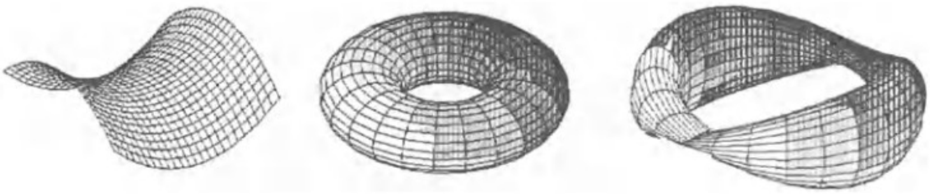

Figs. 19.2–19.4. An elementary surface, the torus, and a pinched torus

Remark 19.1.1 Sometimes we consider surfaces with a *boundary*: in the case of elementary surfaces, a boundary is the image of the boundary of an elementary domain and is homeomorphic to the circle. The boundary of the *half-plane* is a straight line, the boundary of part of a *cylinder* between its parallels is a pair of circles (Fig. 19.5), the boundary of the Möbius band (Fig. 19.6). Recall that a *Möbius band* is obtained when we twist a thin strip of paper through a 180° angle and then glue its lateral sides.

A remarkable property of the Möbius band is that it has only *one side*. Moreover, a surface that "contains" a Möbius band is called *one-sided*. Another example of a one-sided surface is the *Klein bottle*.

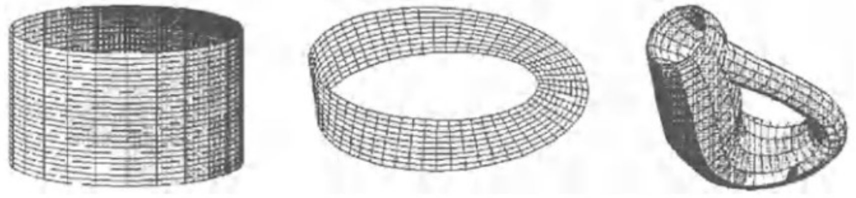

Figs. 19.5–19.7. Circular cylinder, the Möbius band, and the Klein bottle

Definition 19.1.7 A simple surface M is called *complete* if the limit point of any sequence of its points which converges in $\mathbb{R}^3$ belongs to M. If a complete surface does not have a *boundary*, it is called *closed*. A closed and bounded surface is called *compact* — recall that a body in $\mathbb{R}^3$ is called *bounded* if it is contained inside some ball of finite radius. This last notion has deep topological generalizations.

The paraboloid, the hyperboloid, and cylinders of the second order are examples of closed surfaces. The sphere, the torus, and the Klein bottle are examples of compact surfaces. If we omit a closed set from a complete surface, then we will obtain an incomplete surface.

Remark 19.1.2 We can derive in `MAPLE` some calculus-types of operations on vector-valued functions $\vec{\mathbf{r}}(u, v) = [x(u, v),\ y(u, v),\ z(u, v)]$ of the class C^k. The key is in using coordinates: the *limit* $\lim\limits_{(u,v)\to(u_0,v_0)} \vec{\mathbf{r}}(u, v)$,

```
> s:={u=u0, v=v0}:
  r0:=[subs(s,x(u,v)), subs(s,y(u,v)), subs(s,z(u,v))];
```

and the *partial derivatives of the first order*

$$\vec{\mathbf{r}}_u = \lim_{\Delta u\to 0} \tfrac{\vec{\mathbf{r}}(u+\Delta u,v)-\vec{\mathbf{r}}(u,v)}{\Delta u}, \quad \vec{\mathbf{r}}_v = \lim_{\Delta v\to 0} \tfrac{\vec{\mathbf{r}}(u,v+\Delta v)-\vec{\mathbf{r}}(u,v)}{\Delta v}.$$

```
> ru:=[Diff(x(u,v),u), Diff(y(u,v),u), Diff(z(u,v),u)];
```

Since $\vec{\mathbf{r}}_u = [x'_u, y'_u, z'_u]$, $\vec{\mathbf{r}}_v = [x'_v, y'_v, z'_v]$ hold, where $x'_u = \frac{\partial x(u,v)}{\partial u}$, then the differential of the vector-valued function $\vec{\mathbf{r}}(u, v)$ has the form $d\vec{\mathbf{r}} = [dx, dy, dz]$, and hence

$$d\vec{\mathbf{r}} = [x'_u\,du + x'_v\,dv,\ y'_u\,du + y'_v\,dv,\ z'_u\,du + z'_v\,dv] = \vec{\mathbf{r}}_u\,du + \vec{\mathbf{r}}_v\,dv.$$

Analogously, we can calculate the second partial derivatives $\vec{\mathbf{r}}_{uu}$, $\vec{\mathbf{r}}_{vv}$, $\vec{\mathbf{r}}_{uv}$, the *Taylor decomposition* (see example in Section 19.5.2), and so on.

19.2 Regular Parametrized Surface

By definition, in a neighborhood of any its points, a simple surface can be defined by a parametric equation:

$$\vec{\mathbf{r}} = \vec{\mathbf{r}}(u, v) \iff x = x(u, v),\ y = y(u, v),\ z = z(u, v).$$

Suppose that the coordinate functions $x(u, v)$, $y(u, v)$, $z(u, v)$ are defined on some planar domain G and are of class C^k in the variables u and v. In this case the u- and v-curves (space curves) are said to be of class C^k, and we can find their velocity vectors:

$$\vec{\mathbf{r}}_u := x'_u(u, v)\,\mathbf{i} + y'_u(u, v)\,\mathbf{j} + z'_u(u, v)\,\mathbf{k} \ \text{(of the } u\text{-curve)},$$
$$\vec{\mathbf{r}}_v := x'_v(u, v)\,\mathbf{i} + y'_v(u, v)\,\mathbf{j} + z'_v(u, v)\,\mathbf{k} \ \text{(of the } v\text{-curve)}.$$

Definition 19.2.1 A surface M is called *regular of class* C^k (resp. C^∞ or C^ω) if each of its points has a neighborhood with a parametrization $\vec{\mathbf{r}}(u, v) = [x(u, v),\ y(u, v),\ z(u, v)]$, where the functions $x(u, v)$, $y(u, v)$, $z(u, v)$, defined on the elementary domain $G \in \mathbb{R}^2$, are of class C^k (resp. C^∞ or C^ω); moreover, we require that the vectors $\vec{\mathbf{r}}_u(u, v)$ and $\vec{\mathbf{r}}_v(u, v)$ be linearly independent at every point $(u, v) \in G$. For $k = 1$, such a surface is called *smooth*.

Each of the following conditions is equivalent to the property of linear independence of vectors $\vec{\mathbf{r}}_u(u, v)$ and $\vec{\mathbf{r}}_v(u, v)$ and is often used for checking this geometrical property:

(a) *The matrix* $\begin{pmatrix} x_u & y_u & z_u \\ x_v & y_v & z_v \end{pmatrix}$ *at the point* (u, v) *has rank* 2.

(b) *The cross product* $\vec{\mathbf{n}} := \vec{\mathbf{r}}_u \times \vec{\mathbf{r}}_v$ *is nonzero at the point* (u, v).

Definition 19.2.2 A point P on a simple surface M is called a *regular (smooth) point* if some neighborhood P admits a C^k-parametrization $\vec{\mathbf{r}}(u, v)$ such that the vectors $\vec{\mathbf{r}}_u$ and $\vec{\mathbf{r}}_v$ are linearly independent at P; otherwise, P is called a *singular point* (for example, the vertex on the top of a circular cone or the bottom fold of such a cone). A curve on a surface, all of whose points are singular, is called a *singular curve* (for example, two line segments on the pinched torus, Fig. 19.4; a singular generatrix on a cylinder over a plane curve with one singular point, Fig. 19.38).

A surface without singular points is smooth; its small open neighborhoods look like plane disks slightly deformed in space.

Although a regular surface admits a number of parametrizations locally, there does not exist a "natural" way to parameterize it, as in the case of curves. We consider the transition from one parametrization to another.

Lemma 19.2.1 *Let* $\mathbf{f} : G \to M$ *be a* C^k*-regular parametrization of the surface* M *and let* $\mathbf{h} : G_1 \to G$ *be an injective map of class* C^k *on the elementary domain* G_1 *into some domain* G *with nonzero Jacobian* J_h. *Then the composition* $\mathbf{f}_1 = \mathbf{f} \circ \mathbf{h} : G_1 \to M$ *is also a* C^k*-regular parametrization of a domain on* M; Fig. 19.8.

Proof. Let the map $\mathbf{f}$ be defined by the vector-valued function $\vec{\mathbf{r}} = [x(u, v), y(u, v), z(u, v)]$, and let $\mathbf{h}$ be given by the functions $u = \varphi(\alpha, \beta)$, $v = \psi(\alpha, \beta)$. Then the composition $\mathbf{f}_1$ is given by the function

$$\tilde{\mathbf{r}}(\alpha, \beta) = [x(u(\alpha, \beta), v(\alpha, \beta)),\ y(u(\alpha, \beta), v(\alpha, \beta)),\ z(u(\alpha, \beta), v(\alpha, \beta))],$$

and by standard theorems from calculus, the map $\mathbf{f}_1 : \tilde{\mathbf{r}} = \vec{\mathbf{r}}(\varphi(\alpha, \beta),\ \psi(\alpha, \beta))$ is of class C^k. To prove the regularity of the map $\mathbf{f}_1$ we start with $\tilde{\mathbf{r}}_\alpha = \vec{\mathbf{r}}_u \varphi_\alpha +$

$\vec{\mathbf{r}}_v\psi_\alpha$, $\tilde{\mathbf{r}}_\beta = \vec{\mathbf{r}}_u\varphi_\beta + \vec{\mathbf{r}}_v\psi_\beta$. Moreover, $|\vec{\mathbf{r}}_u \times \vec{\mathbf{r}}_v| \neq 0$ holds in view of the regularity of the parametrization **f**. Hence

$$|\tilde{\mathbf{r}}_\alpha \times \tilde{\mathbf{r}}_\beta| = |\vec{\mathbf{r}}_u \times \vec{\mathbf{r}}_v| \cdot |\varphi_\alpha\psi_\beta - \varphi_\beta\psi_\alpha| = |\vec{\mathbf{r}}_u \times \vec{\mathbf{r}}_v| \cdot J_h \neq 0. \square$$

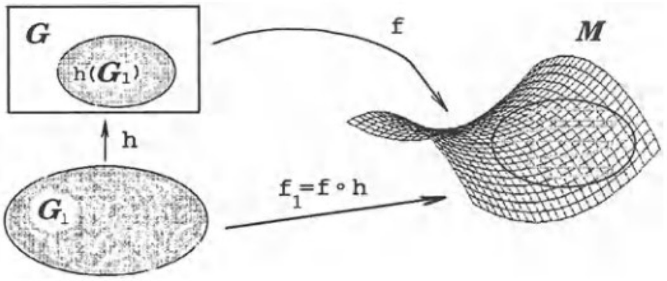

Figs. 19.8. A change of variables on a surface

Exercise 19.2.1

1. Prove that the vector-valued function

(i) $\vec{\mathbf{r}(u, v)} = [u + v, u - v, u^2 + v^2]$ is a C^∞-regular parametrization of the paraboloid $z = \frac{1}{2}(x^2 + y^2)$,

(ii) $\vec{\mathbf{r}(u, v)} = [u^2, uv, v^2]$, where $u > 0,\ v > 0$, defines a C^∞-regular coordinate system on part of the hyperbolic paraboloid $y^2 = xz$;

(iii) $\vec{\mathbf{r}(u, v)} = \left[\frac{a(uv+1)}{u+v},\ \frac{b(u-v)}{u+v},\ \frac{uv-1}{u+v}\right]$ is a C^∞-regular parametrization of part of a hyperboloid of one sheet.

What are the coordinate curves of these parametrizations?

Solution. (i) First check the regularity condition **(ii)**

```
> r:=[u+v, u-v, u^2+v^2]:
> ru:=diff(r,u):rv:=diff(r,v): linalg[crossprod](ru,rv);
```

$[2u + 2v,\ 2u - 2v,\ -2]$ # we obtain a nonzero vector.

Then substitute the vector-valued function in the explicit equation:

```
> r1:=simplify(subs({x=u+v,y=u-v,z=u^2+v^2} , z-(x^2+y^2)/2));
```

0

2. Find which of the following surfaces are compact, and plot them using the command `implicitplot3d` (see Section 19.3.2).

$$(a)\ x^2 - y^4 + z^6 = 1, \qquad (b)\ x^2 - 2x + y^2 + z^4 = 1,$$

$$(c)\quad x^2 + y^2z^2 = 1, \qquad (d)\ x^2 + y^4 + z^6 = 1.$$

```
> plots[implicitplot3d](x^2+y^4+z^6=1,
  x=-2..2, y=-2..2, z=-2..2);   # etc.
```

3. A net of curves on a surface M is said to be *regular* if

- *Every point of the surface belongs to only one curve from each of two families of net curves.*
- *Any curve of one family of the net intersect every curve from another family at exactly one point.*

Check that

(a) The net of straight lines (two families) on the hyperboloid of one sheet is regular locally (i.e., on a neighborhood of an arbitrary point), but it is not regular on the whole surface.

(b) Parallels and meridians on the torus of revolution form a regular net.

Find domains in $\mathbb{R}^2$ where the (polar, parabolic, elliptic, hyperbolic, bipolar) coordinate nets, see Section 6.1, are regular nets.

19.3 Methods of Generating Surfaces

19.3.1 Graphs and Level Sets of Functions

Definition 19.3.1 The *graph* of the function $f : G \subset \mathbb{R}^2 \to \mathbb{R}$ in two real variables (x, y) is the set in space

$$\Gamma_f = \{(x, y, z) \in \mathbb{R}^3 : z = f(x, y), \quad (x, y) \in G\}.$$

Among surfaces of the second order, the elliptic and hyperbolic paraboloids $z = \frac{x^2}{p} \pm \frac{y^2}{q}$ can be expressed as graphs.

Every function in several variables is related to a number of *level sets* (forming the *chart of the function*), which are useful for studying the qualitative properties of the functions.

Definition 19.3.2 Let the function $f(x, y)$ be given on the domain $G \subset \mathbb{R}^2$. The *level curve* with height c of the function f is the set $f^{-1}(c) = \{(x, y) \in G : f(x, y) = c\}$ (the inverse image of c), for example, isobars and isoterms on geographical maps. For every function $F(x, y, z)$ defined on the domain $G \subset \mathbb{R}^3$, the analogous set $F^{-1}(c) = \{(x, y, z) \in G : F(x, y, z) = c\}$ (the inverse image of c) is called the *level surface* of height c of F. In this notation the real number c is called the *height* of the level set.

Note that the level set of height c for the function $f(x)$ consists of roots of the equation $f(x) = c$. In this way, space (or the plane) is fibered by level

!"#$# % & & !' (! !- + A 9 . /
&) +)

0 !#12! ! !3 4) ! > - + A 9 . /
&) +) 1 5 0 *+
3: A & + ===
- *) ? ===
C4 5 0)* G A 9 ((*)

0 !#12! ! !3 4) ! > - + A 9 . /
1 5 0 +F&

!"#$# % & & !' (! !)) . / &) +
)

0 !#12! ! !3 4) ! @)) . / &
) +) *+
3: A & + ===

!"#$# % & & !' (! ! / J ,
, U 7%

0 !#12! ! !3 4) ! / J 6 ,6 *+

!"#$# % & & !' (! ! / / #) / ")
7J))))`

0 !#12! ! !3 4) ! B / / #) / ")
7J))))` * 4)4 % X (5 0 *+
< > *)4

>

@

!"#$# % & & !' (!

!"#$# % & & !' (! !# 7 4) 7J) 7 N
D)) + ') /))

0 !#12! ! !3 4) ! ")& + ** # 7 4)
7J) 7 N D)) + ') /)) ') ' (5 0 * :

>0 !#12! ! !3 4) ! ")& + ** # 7 4)
7J) 7 N D) *+

!"#$# % & & !' (! !') + 1) +)*

@0 !#12! ! !3 4) ! ' 1) 5) 0 a: +
;/* , + . 9 . + 4 * 4)4 ' ; + : 4 5 0 *+

0 !#12! ! !3 4) ! ') + 1) 5) 0
') + 1) a: + ;/* , + . 9 . + 4 b * 4

* (5 0 *+

B0 !#12! ! !3 4) ! ') + 1) +
)* R + (5 0 +F&
< >@ *)4

0 !#12! ! !3 4) ! ') + 1) +
)* R + (5 0 *+

0 !#12! ! !3 4) ! 1) ' + *)*
R + (5 0 +F&

0 !#12! ! !3 4) ! 1) ' + *)*
R + (5 0 *+

0 !#12! ! !3 4) ! 1) ' &
)) J H) J .):)/ 5 0 +F&

0 !#12! ! !3 4) ! 1) ' 3 &
) [ ) J H) J .):)/ 5 0 +F&

0 !#12! ! !3 4) ! 1) ' &
)) J H) J .):)/ 5 0 *+

>0 !#12! ! !3 4) ! 1) ' 3 &
) [ ) J H) J .):)/ 5 0 *+

@0 !#12! ! !3 4) ! 1) , ') + ') +
1) +)* R + (5 0 *+

0 !#12! ! !3 4) ! 1) , ') + "
1) I 9 c) d 4) * :)) e
) ' (5 0 +F&

B0 !#12! ! !3 4) ! 1) , ') + "
1) I 9 c) d 4) * :)) e
) ' (5 0 *+

!"#$# % & & !' (! !) D + .) $)

0 !#12! ! !3 4) !) D + .) $)
$))) #) #) 5 0 +F&

0 !#12! ! !3 4) !) D + .) $)
$))) #) #) 5 0 *+
< @ *)4

!"#$# % & & !' (! !D +))) 8 #) /
#**)

0 !#12! ! !3 4) ! D +))) 8 #) /
#**) G / 5 >0 *+

!"#$# % & & !' (! !A) + 7))) .
7 4

0 !#12! ! !3 4) !) 4 , 1) +6 7 , A) +
A) + 7))) . 7 4 E & /
5 >0 *+
< @ *)4

0 !#12! ! !3 4) ! A) + 7 I A)
.) # ; / / $ 8 : ( ) + 4 *+

0 !#12! ! !3 4) ! 1) +) 4)
)) ;) 5 >0 +F&

>0 !#12! ! !3 4) ! 1) +) 4)
)) ;) 5 >0 *+
3: A & + ===
< B *)4

C4 5 0 D ' (U-- 7
+ #** 8) ,) + + 78)

@0 !#12! ! !3 4) ! 1) +) 4 A) +
7))) . 7 4 E & / 5 >0 *+

!"#$# % & & !' (! !A "))4

0 !#12! ! !3 4) ! A "))4
1 5 0 *+

!"#$# % & & !' (! ! 3: ') + 1)) +
.) E & + :))

B0 !#12! ! !3 4) ! > H " :/ 3:
') + 1)) + .) E & + :)) H * ; / *+

!"#$# % & & !' (

!"#$# % & & !' (!

!"#$# % & & !' (! !3*) 2))) + / /

0 !#12! ! !3 4) ! @
%) 6' (63*) 2)) # + / *+

!"#$# % & & !' (! ! 3 4) 9)) # "))
#**) 7&)) +)

0 !#12! ! !3 4) ! 3 4) 9)) #

")) #**) 7&)) +) " . /
5 0 *+
3: A & + ===

- *)
C4 >5 0

> $ + *) &)) +
4)& / >
> &) 4 * + @
> $ *) /) & /
> #) + ,),) + 4)& / B
> $ 4) /) & /
> > 3 4)) + 4)& / J) f

C4 @>5 0
))))** + 8 @>
.) * * * 4) @>
$ * 4) J) + @>
)))&) +))) +
)) + @B
+ 4 4) +) 8) 4 J)) @
> $ / *)) / + + 4)& /
@ ; ::)) + *) 4 / *)) /
+ + 4)& /
+)) * 4) @

!"#$# % & & !' (! ! /) + 3 4) 1)

0 !#12! ! !3 4) ! B / # $
") - 4 *+
3: A & + ===

0 !#12! ! !3 4) ! B /) + 3 4)
1) *+

@

>B >

!"#$# % & & !' (! >

!"#$# % & & !' (! >!# 4[: $ 4 [[ ) J

0 !#12! ! !3 4) ! &)) / *+

0 !#12! ! !3 4) ! * / *+
< ; +) [ ) J

!"#$# % & & !' (! >!) /

!"#$# % & & !' (! >! * g + &
4[[J + % - A :) (/

0 !#12! ! !3 4) !
6 * g 6+ 6 & 64[[J 6c eA :) & (F6% ( )F6:* @(*+

!"#$# % & & !' (! >!"

>0 !#12! ! !3 4) !
.6A F 6" I 6G (67 6c e" 6 6:* @(> *+
< @@ *)4

@0 !#12! ! !3 4) !
.6A F 6" I 6G (6T 6c e" 6 6:* @(@ *+
< @ *)4

0 !#12! ! !3 4) ! 4 F + (*+

B0 !#12! ! !3 4) ! 4 F + (*+

!"#$# % & & !' (! >!.[[[ ) J)) +

0 !#12! ! !3 4) ! ") . /) +
1) & / # ; # + [A 9 ; @ ' *+
C4 >5 0 $ #E'-% $)) &)
-%$13"E $-3%
G) +) * : 4)) &)
*)))) + - *))) +)
* : X)): " 8))
)) &)) _# +) * : * + :/ ,
% : 4) + # * :)) / & +,
&) * J 5 F 0
$)) $) + B

C4 @ 5 0 # .1#X . + ') 1)) +) +

C4 5 B0 H A7A3%. D711#%" .) - *))) &)) +
1 4) /) ; *

C4 >5 0 X ;32E7$ '1E;#$ $:) .) 1) & /
A 9 7J)

C4 5 0 1 7%13 7 # / *) $;))) G)&)) A

C4 5 @0 # ; $#E' &) - &)) ,))) ,) + &
)) *) $

> 0 !#12! ! !3 4) !)) + .[[+ * &
) +) [ ) J ;) 5 > 0 +F&

> 0 !#12! ! !3 4) !)) + .[[+ * &
) +) [ ) J ;) 5 > 0 *+

> 0 !#12! ! !3 4) !)) +) [ ) J ,)
+ $ g ;) 5 @ 0 +F&

> 0 !#12! ! !3 4) !)) +) [ ) J ,)
+ $ g ;) 5 @ 0 *+

> 0 !#12! ! !3 4) ! 2)) ,
" ,)) /,) +) *+

>>0 !#12! ! !3 4) ! 2) $) + . /
* 4 %) + 5 BB0 *+

!"#$# % & & !' (! >!- *)

>@0 !#12! ! !3 4) ! 9 . +) *)
*+
C4 A#% #1#AAhA7
3: A & + ===

> 0 !#12! ! !3 4) ! 9 . +) *)
*+
C4 13H7 $-3% 31$;3.3%#A7 E1 E% A#% 13H7 $-3% "IE% #%.A7
"13-$
3: A & + ===

>B0 !#12! ! !3 4) !
6 * g 6+ 6 & 64[[J 6c eA :) & (F6% ( )F6:* @(*+

> 0 !#12! ! !3 4) ! ") . /) +
1) & / # ; # + [A 9 ; @ ' *+

@ 0 !#12! ! !3 4) ! . @ +F&
3: A & + ===
< > *)4

@ 0 !#12! ! !3 4) ! . @ *+

@ 0 !#12! ! !3 4) ! R6R#;#%7 *+
< D) +)&) -

@ 0 !#12! ! !3 4) ! R6R#;#%7 *+
< D) +)&) --

@ 0 !#12! ! !3 4) ! A : ; /
.) +)) J 4) + > H) J .):)/ +F&
3: A & + ===
< @ *)4
C4 5B 0 < C4 5 0 $))

@>0 !#12! ! !3 4) ! A : ; /
.) +)) J 4) + > H) J .):)/ *+

@@0 !#12! ! !3 4) ! A7'3 7 ;7 71X
ijklmnopqrsstuvwx
< @ *)4

@ 0 !#12! ! !3 4) ! A7'3 7 ;7 71X
yz{|z}n~opqrsstuvwx
< .[[4[[ )) / J

@B0 !#12! ! !3 4) ! A7'3 7 ;7 71X $ '
ijklmnopqrs•tuvwx

@ 0 !#12! ! !3 4) ! A7'3 7 ;7 71X $ "
yz{|z}n~opqrs•tuvwx

0 !#12! ! !3 4) ! 2)) ,
" ,)) /,) +) *+

0 !#12! ! !3 4) ! 2) $) + . /
* 4 %) + 5 BB0 *+

0 !#12! ! !3 4) ! 6 #$$-%.71 *+
< A / + 7J))**) +))) J) * / J

!"#$# % & & !' (! >! - *) ! #A EA36- 31$#%$
<)€• + + + ,

0 !#12! ! !3 4) !) H)) 7+ €•
*+

0 !#12! ! !3 4) !) H)) 7+ €•
*+

>0 !#12! ! !3 4) !) # *
7" 1 #+) , 7 85)))+), 0 *+

!"#$# % & & !' (! >!- *) !) 6R)*) 6)4

@0 !#12! ! !3 4) ! >
cH [6A 96' + e6# 4 :)6A) *+

0 !#12! ! !3 4) ! >
cG +6R)*) e6 C 6 6ƒ 4 :)6A) 6 6 *+

B0 !#12! ! !3 4) ! >
cG +6R)*) e6 C 6 6ƒ 4 :)6A) 6 6 *+

0 !#12! ! !3 4) ! >
cG +6R)*) e6 C 6 6ƒ 4 :)6A) 6 6 *+
< 7 *)€)

B 0 !#12! ! !3 4) ! >
cG +6R)*) e6 C 6 6ƒ 4 :)6A) 6 6 *+

!"#$# % & & !' (! >!- *) !. /) + 1) & /

B 0 !#12! ! !3 4) ! @

6 * g 6+ 6 & 64[[J 6c eA :) & (F6% ( )F6:* @(*+

B 0 !#12! ! !3 4) ! @
6 * g 6+ 6 & 64[[J 6c eA :) & (F6% ( )F6:* @(*+

B 0 !#12! ! !3 4) ! @ ") . /) +
1) & / # ; # + [A 9 ; @ ' *+

B 0 !#12! ! !3 4) ! @ 2)) ,
" ,)) /,) +) *+

B>0 !#12! ! !3 4) ! @ 2) $) + . /
* 4 %) + 5 BB0 *+

!"#$# % & & !' (! >!- *) !- *) 666)))
+ 7 4) +

B@0 !#12! ! !3 4) ! cR 4
$ 6$) 4e6))) 6 + 6 67 4 6) +6 6 *+

B 0 !#12! ! !3 4) ! cR 4
$ 6$) 4e6))) 6 + 6 67 4 6) +6 6 *+

BB0 !#12! ! !3 4) ! cR 4
$ 6$) 4e6))) 6 + 6 67 4 6) +6 6 *+

!"#$# % & & !' (! >!- *) !))))) /) +
) +) + 4/6)4

B 0 !#12! ! !3 4) ! B
))) 6#) / 6 6 +F&
< *)4

0 !#12! ! !3 4) ! B
))) 6#) / 6 6 *+

0 !#12! ! !3 4) ! B
))) 6#) / 6 6 +F&
< @ *)4
C4 > 5@ 0 @ " I : A)*) I 1„ …
$) #8) / /

0 !#12! ! !3 4) ! B
))) 6#) / 6 6 *+

0 !#12! ! !3 4) ! B
))) 6#) / 6 6 +F&

0 !#12! ! !3 4) ! B
))) 6#) / 6 6 *+

>0 !#12! ! !3 4) ! B

))) 6#) / 6 6 +F&
C4 5 > 0 $) ") & + ' +)

@0 !#12! ! !3 4) ! B
))) 6#) / 6 6 *+

!"#$# % & & !' (! >!- *) !2)

0 !#12! ! !3 4) ! # :) 2)
) 5 & : +) 0 " & :) 5 0 +F&
< *)4
3: A & + ===

B0 !#12! ! !3 4) ! # :) 2)
) 5 & : +) 0 " & :) 5 0 *+

0 !#12! ! !3 4) ! # :) 2)
) c e % ;) + 5 @ 0 *+

0 !#12! ! !3 4) ! # :) 2)
) c e % ;) + 5 @ 0 *+

0 !#12! ! !3 4) !) + $) +F
[ ) J J) J $ - 7" 6 %1 †+ 5 B0 *+
< > *)4

0 !#12! ! !3 4) !) + $) +F
[ ) J J) J $ -- 7" 6 %1 †+ 5 B0 *+
< @ B *)4

0 !#12! ! !3 4) !) + $) +F
[ ) J J) J $ --- 6 D , : , * , [ )) 7"
6 %1 †+ 5 0 *+
< *)4

0 !#12! ! !3 4) ! 7 R) 7 +
[ ) J J) J $ E & [+ A g4 5 0 *+

!"#$# % & & !' (
! >!- *) !$)6+ 67)4 6)4

>0 !#12! ! !3 4) ! >
$)6+ 67)4 6 6 *+
< *)4

@0 !#12! ! !3 4) ! >
$)6+ 67)4 6 6 8*

0 !#12! ! !3 4) ! >
$)6+ 67)4 6 6 *+
< B *)4

B0 !#12! ! !3 4) ! >
$)6+ 67)4 6 6 8*

0 !#12! ! !3 4) ! >
$)6+ 67)4 6 6 *+
< B *)4

0 !#12! ! !3 4) ! >
$)6+ 67)4 6 6 8*

!"#$# % & & !' (
! >!- *) !$)6+ 67)4 6)4 !#+&) +67 4 46))
6)4

0 !#12! ! !3 4) ! >
#+&) 67 4 46)) 6H /6 +F&

0 !#12! ! !3 4) ! >
#+&) 67 4 46)) 6H /6 *+

!"#$# % & & !' (! >!- *) !$) [+) [ ) J
*[ )

0 !#12! ! !3 4) ! >
)4) 6. ,61) 67 ,6) 6H $) [6+ 6) [ ) J 6 *[ ) 6 6^6# 4g:
) 666 5 0 +F&
< > *)4

0 !#12! ! !3 4) ! >
)4) 6. ,61) 67 ,6) 6H $) [6+ 6) [ ) J 6 *[ ) 6 6^6# 4g:
) 666 5 0 *+

>0 !#12! ! !3 4) ! >
)4) 6. ,61) 67 ,6) 6H $) [6+ 6) [ ) J 6 *[ ) 6 6^6#) /
) 666 5 0 +F&
< @@B *)4

@0 !#12! ! !3 4) ! >
)4) 6. ,61) 67 ,6) 6H $) [6+ 6) [ ) J 6 *[ ) 6 6^6#) /
) 666 5 0 *+

0 !#12! ! !3 4) ! >
)4) 6. ,61) 67 ,6) 6H $) [6+ 6) [ ) J 6 *[ ) 6 6^6.[[
) 666 5 0 +F&
< >@B *)4

B0 !#12! ! !3 4) ! >
)4) 6. ,61) 67 ,6) 6H $) [6+ 6) [ ) J 6 *[ ) 6 6^6.[[
) 666 5 0 *+

0 !#12! ! !3 4) ! >
)4) 6. ,61) 67 ,6) 6H $) [6+ 6) [ ) J 6 *[ ) 6 6^

6#**) 6+ 6 N#) / 6‡6)6.[[ ) 666 5 0 +F&
< @B *)4

0 !#12! ! !3 4) ! >
)4) 6. ,61) 67 ,6) 6H $) [6+ 6) [ ) J 6 *[ ) 6 6^
6#**) 6+ 6 N#) / 6‡6)6.[[ ) 666 5 0 *+

!"#$# % & & !' (! >!H) :
< B &

!"#$# % & & !' (! >!) ' (

0 !#12! ! !3 4) ! > $6$7 # >@ *+
< 3 $ /) + %) #) / %)& (7J)

0 !#12! ! !3 4) ! > G6G#A" ; -"$ @ *+
< [ ) :) +)

0 !#12! ! !3 4) ! > G6G#A" ; -"$ *+
< E * % : $) +)

!"#$# % & & !' (! >!) ' (*)

0 !#12! ! !3 4) ! > A : ; /
.) +)) J 4) + > H) J .):)/ +F&
< @ *)4

>0 !#12! ! !3 4) ! > A : ; /
.) +)) J 4) + > H) J .):)/ *+

@0 !#12! ! !3 4) ! > A : ; /
4 *+

0 !#12! ! !3 4) ! >)) J 6) 6+ 6
D) +6%)) +F&
C4 5 0 3 4) [+ + + *)
C4 >@5 @ 0 ") + *) +
C4 B 5 0 @ - [+ *) +
C4 5 >0 A) / [4) 4 * + [

3: A & + ===

B0 !#12! ! !3 4) ! >)) J 6) 6+ 6
D) +6%)) *+

0 !#12! ! !3 4) ! > $) 6) [ ) J
7 #) / B +F&

0 !#12! ! !3 4) ! > $) 6) [ ) J
7 #) / B *+

!"#$# % & & !' (! >! [ ) J 2) J #) 4

0 !#12! ! !3 4) ! >> #) 4) J J) J ,
D + *)**) $ " ' (5 0 +F&

0 !#12! ! !3 4) ! >> #) 4) J J) J ,
D + *)**) $ " ' (5 0 *+

0 !#12! ! !3 4) ! >> #) 4) J
J) J " ' (5 0 +F&

0 !#12! ! !3 4) ! >> #) 4) J
J) J " ' (5 0 *+

>0 !#12! ! !3 4) ! >> #) 4 #**) +
) J J) J 5 @0 *+
< #**) +) ˆ) J 2) J
" N)) +
5D- A@0
3: A & + ===

@0 !#12! ! !3 4) ! >>) + #) 4 [ ) J
J) J 6 $, D + * g)**) " ' (5 0 +F&
< @ *)4

0 !#12! ! !3 4) ! >>) + #) 4 [ ) J
J) J 6 $, D + * g)**) " ' (5 0 *+

B0 !#12! ! !3 4) ! >>) + #) 4 [ ) J
J) J 6 $, "[& **)**) ‡ :) [4 " ' (5 B0 +F&
< > *)4

0 !#12! ! !3 4) ! >>) + #) 4 [ ) J
J) J 6 $, "[& **)**) ‡ :) [4 " ' (*[
5 B0 *+

0 !#12! ! !3 4) ! >>) + #) 4 [ ) J
2) J 6 $, 4["[ ) [[+ 78 : g " ' (
5 0 *+
< @ *)4

0 !#12! ! !3 4) ! >>) + #) 4 [ ) J
2) J 4["[ ) [[+ 78 : g " ' (+F&

‰Š‰t‹Œ•Ži••Žq‘’{“n”oŽ•{j“|o‘–q•‘‘‰–‰‰–‰‰–Ž—n~k~˜™jŽ–šš‘‘Œj™“wo‹i›j™˜k“j‹‘‹œz”™˜~•“o‹
J) J +F&

‰Šžt‹Œ•Ži••Žq‘’{“n”oŽ•{j“|o‘–q•‘‘‰–‰‰–‰‰–Ž—n~k~˜™jŽ–šš‘‘Œj™“wo‹i›j™˜k“j‹‘‹œz”™˜~•“o‹
J) J +F&

0 !#12! ! !3 4) ! >>) + #) 4 [ ) J
J) J " ' (5 0 +F&

!"#$# % & & !' (! >! :

!"#$# % & & !' (! >!$)6+ 67)4 6)4

!"#$# % & & !' (! >!$ %1 7 / * +)
)) :/ .) +

>0 !#12! ! !3 4) ! >@ # (, G : , ;)
))) + / *+
< @ *)4

@0 !#12! ! !3 4) ! >@ % A) 4)
D177A-'13 31. *+

0 !#12! ! !3 4) ! >@ D +)) 7+ *+
< > @ *)4
C4 5 0)* B $) 4

B0 !#12! ! !3 4) ! >@ 4) *+
3: A & + ===
< @ @ *)4

0 !#12! ! !3 4) ! >@ A : 9)))
) + D +) . / $ E * : +) * *+
3: A & + ===
< *)4

> 0 !#12! ! !3 4) ! >@ $ %1 7 / * +)
)) *+
3: A & + ===
< B *)4

!"#$# % & & !' (! @

!"#$# % & & !' (! @!#: [4[+I +) [ ) J

> 0 !#12! ! !3 4) ! > #: [4[+ +
) [ ) J :/ H) " + [+F&
< > *)4

> 0 !#12! ! !3 4) ! > #: [4[+ +
) [ ) J :/ H) " + [*+

> 0 !#12! ! !3 4) ! > * / *+

!"#$# % & & !' (! @!7 +I 7 $ $ ' (
7 +I 7

> 0 !#12! ! !3 4) ! >B ;) $ ' (7 +I
7 ' (- --) : +4 *+

>>0 !#12! ! !3 4) ! >B ;) $ ' (7 +I
7 ' (--- -U) : +4 *+

>@0 !#12! ! !3 4) ! >B ;) $ ' (7 +I
7 ' (U U---) : +4 *+

> 0 !#12! ! !3 4) ! >B ;) $ A $: (
7 +I E 5 B0 +F&

>B0 !#12! ! !3 4) ! >B ;) $ A $: (
7 +I E 5 B0 +F&

> 0 !#12! ! !3 4) ! >B ;) $ A $: (
7 +I E 5 B0 +F&

!"#$# % & & !' (! @!78 "I# 4g: "I#) / +
$ 4 [

@ 0 !#12! ! !3 4) ! > B 6 8) *+
< * : Ÿ + " :)
$ & : / 5 8 0 *)J)J * ,
4 +ˆ ˆ *) * +)& N)8 + 8 +
)) 4 $ +) * * +) % Ÿ) ¡)8
) ¢

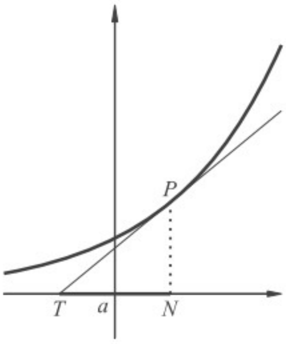

D 4

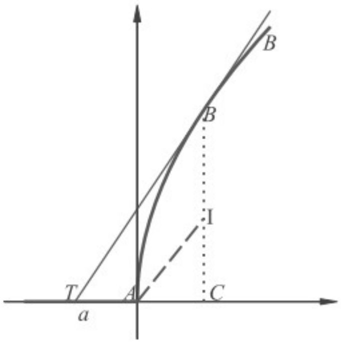

D 4

A) 4 £J) +) ˆ J) + ˆ) ˆ ˆ
ˆ @ @ *) ¡ A : 9 * +ˆ 4)
+ ˆ +8 +/ + + 8 J) ˆ &)): 8 /

@ 0 !#12! ! !3 4) ! >
78 6+I) 4g: 6+I)) / 6 6+ 6c e# : 6) 6:* @(> *+

@ 0 !#12! ! !3 4) ! > 78 6+ 64[[6
6 6c eD6H6:* @(@> > J *+
< @ *)4
< 78 + .[[

@ 0 !#12! ! !3 4) ! >
78 6+ 64[[66:* @(@ > > > *+
< @ *)4
3: A & + ===

@ 0 !#12! ! !3 4) ! > 4 8 + 4
*) * + + 4 J J *+
< > *)4
3: A & + ===

@>0 !#12! ! !3 4) ! >
[6+I) 4g: 6+I)) / 6 6+ 6c e)* 6. 4 6:* @(*+
< > *)4

@@0 !#12! ! !3 4) ! >
[6+I) 4g: 6+I)) / 6 6+ 6c e)* 6. 4 6:* @(> *+
< $ 4 [
< BB *)4

@ 0 !#12! ! !3 4) ! >
[6+ 64[[6)) / J 6‡6c e)* 6. 4 6:* @(> B *+
< .[[#) / J
< *)4

!"#$# % & & !' (! @!. / D "

@B0 !#12! ! !3 4) ! @ . / D "
5& -0 +F&
< > *)4

@ 0 !#12! ! !3 4) ! @ . / D "
5& -0 *+

0 !#12! ! !3 4) ! @) 4 . / D
" E & / 4) A :) / 5 >0 +F&

0 !#12! ! !3 4) ! @) 4 . / D
" E & / 4) A :) / 5 >0 *+

!"#$# % & & !' (! @!. / D " !D /

0 !#12! ! !3 4) ! @ > >> -
. / *+
< > > *)4
C4 ---5 0 .))) * *)) + 4) J)+ * *)

0 !#12! ! !3 4) ! @ > >> -
. /6 8 *+

C4 ---5 0
- *) ,) + / , /* 4)
4 /)))&) :)
* 3 /* * & + + :/ & 8
) *) / *) *) /) 4 + +) / 4) +
D J) * & +) :) 4 J :
& &) 4)) + * * ,) + /
8) / +)) $ /* * & + +
:/ *) * + / + *) /) ,) * + 8
4) *) / *) / + 8 &
) & ,) .))) * *)) + 4
) J)+ * *))) * + ,) &))
4 + 8 *) &)) 4) - , ,
* : * +) & /) +) * 4) /
) * + + + /* 4 , * :/
*,)))/ :) & +) + : 4 4 K :)
* ,)) /)**) + / (+) *
)) + +) 4 / & $ 4 /
) *) / +)) / + & * +) 4 /)
) 8 4 / J * *) +)

0 !#12! ! !3 4) ! @ > B -
. / 3 -+) *) - *+
< @ B *)4
C4 5 >0 & *)) &)))

>0 !#12! ! !3 4) ! @ > B -
. / 3 -+) *) - 6 8 *+

@0 !#12! ! !3 4) ! @ + @ :* +F&
< # ") . /
< BB *)4

0 !#12! ! !3 4) ! @ + @ :* *+

B0 !#12! ! !3 4) ! @ 4 / + @ :* +F&
< . / D "
< > *)4

0 !#12! ! !3 4) ! @ 4 / + @ :* *+

B 0 !#12! ! !3 4) ! @ 4 > :* +F&
< - . / -+) *) --
< @ *)4

B 0 !#12! ! !3 4) ! @ 4 > :* *+

B 0 !#12! ! !3 4) ! @ + * >B :* +F&
< # - + F & . /
< *)4
C4 5 0 @ $ *) $ / 1) & /

B 0 !#12! ! !3 4) ! @ + * >B :* *+

B 0 !#12! ! !3 4) ! @ * F & + > :* +F&
< F & ") . / &) +)
< *)4
3: A & + ===

B>0 !#12! ! !3 4) ! @ * F & + > :* *+

!"#$# % & & !' (! @!. / D " !; :

B@0 !#12! ! !3 4) ! @ +) 4 : +F&
< D +) . / ")& + ; :
< *)4

B 0 !#12! ! !3 4) ! @ +) 4 : *+

!"#$# % & & !' (! @!. / D " !X 4
D /
< # J & • ,*)

!"#$# % & & !' (! @!;) +: ( )))
' /

BB0 !#12! ! !3 4) ! @) +: ( )))
*+
< B *)4

!"#$# % & & !' (! @!A * : g + ; :
)

!"#$# % & & !' (! @!) [ ) J -)) [

B 0 !#12! ! !3 4) ! @ ; 4)) [ ) J
)) [D)) 5 @ 0 +F&
< @ *)4

0 !#12! ! !3 4) ! @ ; 4)) [ ) J
)) [D)) 5 @ 0 *+

C4

$)€ ¤ [+I) 4 * [+)) 4 + @ L +
) 4 [J) [ ) 8 5 4 >0 A) +) +) 4 *
+ &)) 4 ** [‡ I) 4 + @ L [4) ‡) +

) + + 8)) 4 5 4 @0 I
)) 4 [J) [ ) , * & *) +I) 4
J J 5 4 0) +) 4 *) 4
:) [J) [ ) A) +I) 4 * ¥
[) [ )& : * [€) *) *))4
+ 8 5 4 B0, * + &)) + + I) [*)
[4))/ J * +) 8 + 8
)** [

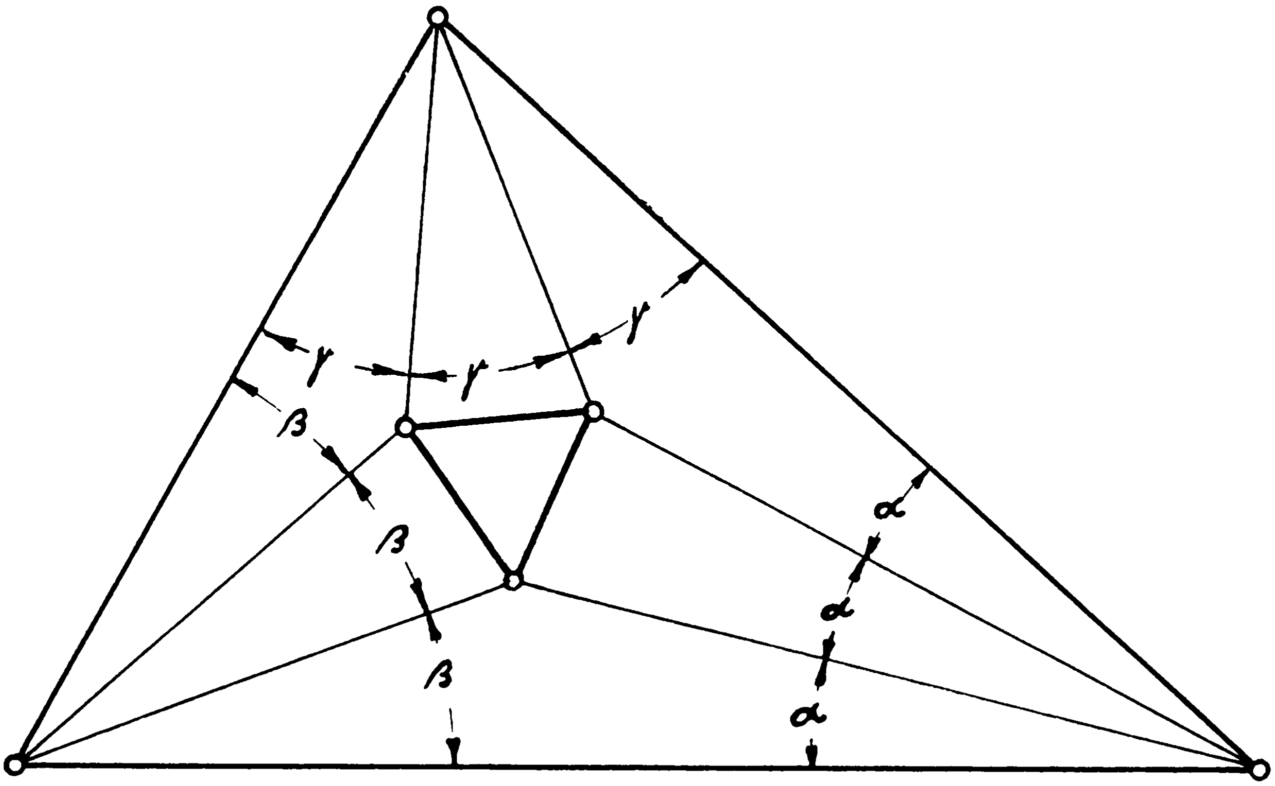

D 4

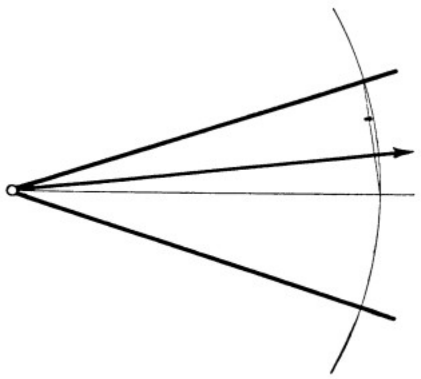

D 4 B

!"#$# % & & !' (! @! F & ") . / 3
&) +)

0 !#12! ! !3 4) ! @> G 9/ (F & +)
4 / &) + +) E & / A :) / 5 @0 *+
< *)4

!"#$# % & & !' (! @!$ 99 4 G + 3 / +)
"

0 !#12! ! !3 4) ! @@) $.
* 99 + 4 # R 5 0 *+

0 !#12! ! !3 4) ! @@) $ $ 99 4
G + / +) " 38 + E & / 5 0 *+

!"#$# % & & !' (! @!$ # . / # +
E

0 !#12! ! !3 4) ! @ $ ' [,) %)*
$ # . / # + 6 -- 1 + -))
: 5 @ 0 +F&

>0 !#12! ! !3 4) ! @ $ ' [,) %)*
$ # . / # + 6 -- 1 + -))
: 5 @ 0 *+

@0 !#12! ! !3 4) ! @ $ ' [,) %)*
$ # . / # + 6 - 1 + -))
: 5 @ 0 +F&

0 !#12! ! !3 4) ! @ $ ' [,) %)*
$ # . / # + 6 - 1 + -))
: 5 @ 0 *+

!"#$# % & & !' (! @!$ E) ' ($ * $ A

B0 !#12! ! !3 4) ! @B $ E) ' ($ * $
A # + ' 44 4 D :
< > *)4

0 !#12! ! !3 4) ! @B $ E) ' ($ * $
A # + ' 44 4 D *+

!"#$# % & & !' (! @!$ 4 [A *) +
<)4

!"#$# % & & !' (!

!"#$# % & & !' (! !# - + % #:) "
/) /

0 !#12! ! !3 4) ! @
R :)/) 6' (6# - + $ % #:) " / *+

!"#$# % & & !' (! !#) + %))))
#+&) + 7 4/ &

0 !#12! ! !3 4) !
6) +6%) 6) 6)) 6 6#+&) +67 4/6 & *+
< B B *)4

!"#$# % & & !' (! !))) $ J 7 4
) +

0 !#12! ! !3 4) !))) $ J
7 4) + *+
< B @ *)4

!"#$# % & & !' (! !)) - . +

0 !#12! ! !3 4) !)) - . + *+
< B *)4
3: A & + ===

C4 B5 > 0 @ @ .3$$D1-7" G-A;7A A7-'%-T

!"#$# % & & !' (! !)) $ +)/ D :) /

!"#$# % & & !' (! !)&) + G)&
) 7 " &

!"#$# % & & !' (! ! 13'A7 -% A#%7 #%" 3A-"
.73 7$1X &) . /

0 !#12! ! !3 4) ! # 4 :) & $) +
+ *) 5 A 0 * 4 5 0 *+

>0 !#12! ! !3 4) ! . / #))))
/ 5 0 +F&

@0 !#12! ! !3 4) ! . / #))))
/ 5 0 *+

0 !#12! ! !3 4) !) & @ 3++ X)
))) 3 / *)+ G : +) 5 0 *+
< > *)4

B0 !#12! ! !3 4) !) & ") . /
* 4 %) 9) + #. 5 0 *+

&)
&)) + D D)
3) 4 @
$ $) &)) +) &
D 8 $
> $ %)) 7J))) &

@ G / .) $
$: #)) + N D) >
B $ 7 & *) D) / & @
7&) + - &
- * - J) /
Z E +) ") . / @
F & ") . /
$) A - 4) . & &
: @
& *)
&)) + $ $ D D)
3) 4) >
$) &)) + & >
') + & >>
> $ D D)) / ") *) >@
@ : >
) *) @>
$ D 2)+) D @@
$ ") : 8 D)) &)) @B
. +
$ + 2)+) D
> .)) &) >
@ .)) &)) + ") D
$.) ' $ B
B / : B
$ *) .))* B
$. + 7J) BB
)) $) * # 4) & B
&)) ")
$.)) + +)99)) + 7J)
1) &) $
> 78*))*
@ A &)) + # / * A
)) B
B $ D)) D)
$ +)) D)
H) : D +) + F 4) @
H) : N $) %) *) -)4
)) .)) &)
1 4 + / 5E : +): /0 *
& 8) ;)+)) +N $
> $ A)*) ') 3*)
@ :

0 !#12! ! !3 4) !) & 7 ; 4/
$ / #)))) / 5 0 *+
< *)4

0 !#12! ! !3 4) !) & 7)/ :) +
4 #)))) / 5 0 +F&

0 !#12! ! !3 4) !) & 7)/ :) +
4 #)))) / 5 0 *+

0 !#12! ! !3 4) !) & : *)
) + + 4 /, ,) 4 / 5 0 *+
< > *)4
C4)* > #1#AA7A $1#% A#$-3%

0 !#12! ! !3 4) ! & / +))
E 5 >0 *+

0 !#12! ! !3 4) !) & " A 5)) 0
:)) + + . / &) . / *+

!"#$# % & & !' (! ! *) $

>0 !#12! ! !3 4) ! . 4 * (/ #)
) ") 6 $ $) 1)4 *)) + $ * :

@0 !#12! ! !3 4) ! . 4 * (/ #)
) ") 6 $ $) 1)4 *)) + $ *+

!"#$# % & & !' (! !$ * 4) / /

0 !#12! ! !3 4) ! >
6' (6$ * 4) / / *+

!"#$# % & & !' (! !G) -) 2) D + $ / # D
- +)))

B0 !#12! ! !3 4) ! @ 2) D + $ / *+
< >B *)4

!"#$# % & & !' (! B

!"#$# % & & !' (! B!# - + ; /
)4 :)))

0 !#12! ! !3 4) ! # - + ; /
)4 :))) :

!"#$# % & & !' (! B!" * & $ * 4/ + $ *
D) #) /

0 !#12! ! !3 4) ! B " * & $ * 4/ +
$ * D) #) / *+
< *)4

C4 >>5 >@0)*
')) *)) / F

')) *)) / F >>
) , + >>
F) / : + @
A + ¦ * * /)(* 4 @
*)): *) * * / @>
> F) (@
@ % 4 : *) + + :/) * F) (>
4) * + B
B) (& :) , (*)) + (*) B
&) (*) BB

!"#$# % & & !' (! B!D)) " &) & + 4
)) + 7 4 4

0 !#12! ! !3 4) !
6' (6D)) " &) & + 4- *+

!"#$# % & & !' (! B!- *)

0 !#12! ! !3 4) ! B *)) *+
3: # 4 *) ? (#:)) E 4)
<)** N) 7) & & / + /* : 4 /

1. Introduction

Hyperbolic geometry appeared in the first half of the 19^{th} century as an attempt to understand Euclid's axiomatic basis for geometry. It is also known as a type of non-Euclidean geometry, being in many respects similar to Euclidean geometry. Hyperbolic geometry includes such concepts as: distance, angle and both of them have many theorems in common. There are known many main models for hyperbolic geometry, such as: Poincaré disc model, Poincaré half-plane, Klein model, Einstein relativistic velocity model, etc. The hyperbolic geometry is a non-Euclidian geometry. Here, in this study, we present a proof of Pappus's harmonic theorem in the Einstein relativistic velocity model of hyperbolic geometry. Pappus's harmonic theorem states that if $A'B'C'$ is the cevian triangle of point M with respect to the triangle ABC such that the lines $B'C'$ and BC meet at A'', then $\frac{A''B}{A''C} = \frac{A'B}{A'C}$ [4].

Let D denote the complex unit disc in complex z - plane, i.e.

$$D = \{z \in \mathbb{C} : |z| < 1\}.$$

0 !#12! ! !3 4) ! B) 6 &) 6
6 & 6 8 6): *+

0 !#12! ! !3 4) ! B " X & #
- + . / % " *+

>0 !#12! ! !3 4) ! B ")) 4) + 7 +))
4) + NE & 6 .)) *+
< ")) 4)
+ 7 +

)) 4)
+ NE &
.)), + , , * +)
< B *)4
3: A & + ===

@0 !#12! ! !3 4) ! B .)) & /6) 7
) / & /) : +4 E & / 5 0 +F&

0 !#12! ! !3 4) ! B .)) & /6) 7
) / & /) : +4 E & / 5 0 *+

B0 !#12! ! !3 4) ! B 4 * *+
< .) 7 *)) . . :)
< *)4

0 !#12! ! !3 4) ! B ; /) ' B> *+
< B *)4

0 !#12! ! !3 4) ! B * @6 * + 4 6) *+
< ; /) + +)4 4/))
< *)4
C4 B5 0 D 4

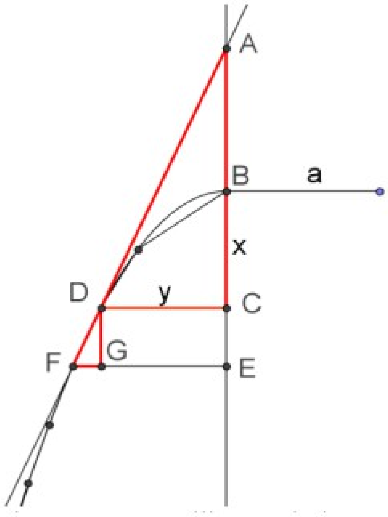

D 4 ' I *)):)

0 !#12! ! !3 4) ! B) @ @ +F&

0 !#12! ! !3 4) ! B) @ @ *+

0 !#12! ! !3 4) ! B 3 # +) + '6:)+
)) *+

0 !#12! ! !3 4) ! B *) *+
< $ $) + §

$ *) §) +) (+ *))
N +) & /)) + $ * *
))(* + /)&)): # :)
*))) +))
$))) K)( ) **
+ :/ -)))))+F) 1 4) + D +)
E & / G) (

!"#$# % & & !' (! B!- 31$#%$7 6.73 7$1X

>0 !#12! ! !3 4) ! B > *+
< . /

@0 !#12! ! !3 4) ! B 7##%6 *+
< *)4

0 !#12! ! !3 4) ! B 7##%6 63 1 *+

B0 !#12! ! !3 4) ! B
. /6 61 6") 6 & *+

0 !#12! ! !3 4) ! B
) 6. /6G 6 6 + + *+

0 !#12! ! !3 4) ! B
) 6 *) 6$ 4 /6G 6 6 + +6 *+

0 !#12! ! !3 4) ! B
+6. /6G 6 6 + + *+

!"#$# % & & !' (
! B!- 31$#%$7 6.73 7$1X!.)

0 !#12! ! !3 4) ! B 4) *+
3: A & + ===

!"#$# % & & !' (! B!- 31$#%$7 6.73 7$1X!

0 !#12! ! !3 4) ! B + & *+
< *)4

0 !#12! ! !3 4) ! B *+
< # "-#.1# #$- D31 #A X $7 D31
7E A-"7#% .73 7$1X
< *)4

>0 !#12! ! !3 4) ! B A +) *+
< $. +)

@0 !#12! ! !3 4) ! B :) *+
< " & * 4)

0 !#12! ! !3 4) ! B)))) *+
< D :)

B0 !#12! ! !3 4) ! B *+
< $ 9 :
< *)4

0 !#12! ! !3 4) ! B 9

!"#$# % & & !' (! B!- *) ?

B B

!"#$# % & & !' (! B!- *) ?

> 0 !#12! ! !3 4) ! B B 9) *+
<)) *

> 0 !#12! ! !3 4) ! B)4) *+
< H 4) :F * + + :/) 3"7 /) + * + 3"7

> 0 !#12! ! !3 4) ! B *)) *+
<)** N) 7) & & / + /* : 4 /
< #:) - , *) *)** N) 7
) & & / + /* : 4 /

> 0 !#12! ! !3 4) ! B > :)) *+
< 3): / : &)) 4 * + + : N)) 9)

!"#$# % & & !' (! B!- *) ? !;)+)) +

> 0 !#12! ! !3 4) ! B> @ @ @ ' *+
< D)) ;)+)) +) + D Fˆ ;)+)) + J) 8*) / & 8
< #:) D)) 4) *) *)/) &))+&)
))) J) $) *)* * ;)+)) +) +
D Fˆ ;)+)) + J) 4) 9 +)) 4) *)
) 4)4 A 78*) / & 8 / 9 +):
J) '/ Z8 4 *)) & & +)4 A
;)+)) +) + D Fˆ ;)+)) + J) &) ( ))
4) *)) : :) +

!"#$# % & & !' (! B!- *) ? !%

>>0 !#12! ! !3 4) ! B@ @ ' *+
< 78 + +) & 4 % /* & +)(+
< #:) G *) & 4) % /* & J)
& & 4 ')) *) &) + *) $ 4 + & 4)
) *)) (P+) 7 +) *) , 4
/* * 4 + &) &) 4 + &) & +)**)
+ G & 4 4 / Z + &) & $)/

8 +)**): / + %) 8) *) +
& 4 + D) /, :))) +,
* :) * +

!"#$# % & & !' (! B!- *) ? 6' (6 B6
<))) 4) ===

>@0 !#12! ! !3 4) ! B @ *+
< . /
< B *)4
3: A & + ===
- *) ? ===

-) + A # A ( ) F & . /
) +
$) +)) 4
$ @
D 2)+)) >
> D &
@ F) *)
") Z4) * F & 4 /
B ") 4 Z4) >
") ") 4 Z4) >
4 :)) b * *) ¨ Z4) >@
/4) +) 4 >
& 8 * /4 , & 8 * @
78 *) ") 4 Z4) @>
3& &)* -
--)) A # A ( ) # Z . /
) - A) + 4
))) 4 A
3*))&)):)* --
* *))
4 / 4 *)) @
) -- D 4
>)) 4) B
@ $) 4
$)* 9 +

) --- 3*) Z4
B $ ((/)++ Z4
)) * F >
))))
) / / *)
>

3& &)* -- >
#** + 8)* -- @
G /) + Z J) 4)* -- @
)) , J) 4 ,) + ") 4 Z4) @
# *) 4
4) +)) :

> # Z +) / *) B
--- #) # A ( ) / * . / B
$))) Z4 B
#)) *)) 4)
#))) 4 >
#)))* 9 + >
> #)) * /4
@ * :))
;)))) Z4
B 3& &)* ---
- # A ( ) 7 +) . /
) - - +
3*))&)):)* -
*) 4 4 >
4
3*) Z4 >
) -- $ 4 /) 4) + Z4
@
7) 4 4) 4 >

)) 4 >
B
1) :)) 4 @

*)) 4)) (+) 4

) 7 +) 4 / B

$ /)4)) +)**) B>

1) :) + * @

*)) + *) *)) 4 B

/4 B
) /)): + J)+))

>) /
) ---
@) + *
) +
B $
) +) 4 >
) +)) 4
) + * /4 B
) +)
" ( ) + B
3& &)* -

> 0 !#12! ! !3 4) ! B @ *+
< +) *)

>B0 !#12! ! !3 4) ! B @ *+
< $ D)) $ 4 /
< > *)4

> 0 !#12! ! !3 4) ! B @> B *+
< .) .) + R) " 4
< BB *)4

@ 0 !#12! ! !3 4) ! B B@ *+
< .) #) / A)+) 4)*) /) + 4 " 4
< @ *)4

@ 0 !#12! ! !3 4) ! B > @@@ *+
< 7 $)

@ 0 !#12! ! !3 4) ! B @ > *+
< ;) +: (3* 9)

@ 0 !#12! ! !3 4) ! B @ @ *+
< D +) . # 4 :) * 4

@ 0 !#12! ! !3 4) ! B # 4 :) . / &
* 8 % : 5 , * 4)4 % X (0 *+

) - - + 4 78) *
) &

4) >
'ˆ 9 N $
:
> .) +
@ ;/* * &
) --)&) + . /
) +) +) &))&
)& D
) +
Z) B
4 :))
>) ( ) + $) 4 *) >
@ D , D + ,) + ' +
*) * 8) +) +) >
) $ /
$) 4 /)&
78) J >
Z >B
) + . 4 @
>)& + @@
@ A ' + F & *)
") + - & -)4
B ") @

) 4/
D) J)&
4/ B
)& B@

©
+ # B
>)/ J
@ + W
> " 1) 4/) +
> # / 1
> " 1) N $
> Rª N D)
>) ˆ ") / >
> > ./)* B
> > F
> > -
> @ D +)))
> A 9 $) D)
@ 1))
@ .
@ « 4/
@ F & 7 : ++ 4 @
@ D D +) + # *
@ > +) D) + &
*) +
*)) + 4) 4/
4/ F & *)
4/
¬) 4/
> D) >

@>0 !#12! ! !3 4) ! B # - +
$7A7 #1#AA7A .1# -$X - + D?) $) E & +) + 7)+)))))
')9 5 0 *+
< *)4
3: A & + ===

C4 5 0

)*
D +)) $ *)) .)& /
$ +) $ *)) .)& /,) 4) 4 /))
4 *,) * + .) 4))) + +,) +
* 4) Z + 4)&) + + $ Z + 4
/ :

Chapter 3

Fundamentals of Teleparallel Gravity

The foundations of Teleparallel Gravity, a gauge theory for the translation group, are presented. Gauge transformations are introduced, and the coupling of matter fields to gravitation discussed. The field strength of the theory is shown to be the torsion tensor.

3.1 Geometrical Setting

The geometrical setting of Teleparallel Gravity — a gauge theory for the translation group — is the tangent bundle: at each point p of coordinates x^μ of a general, riemannian spacetime $\mathbf{R}$ — the base space — there is "attached" a Minkowski tangent-space $\mathbf{M} = T_p\mathbf{R} = T_{x^\mu}\mathbf{R}$ — the fiber — on which the gauge transformations take place (see Fig. 3.1, in which the tangent space at $p = \{x^\mu\}$ is indicated normally). A gauge transformation will be a point-dependent translation of the $T_{x^\mu}\mathbf{R}$ coordinates x^a,

$$x'^a = x^a + \epsilon^a, \tag{3.1}$$

with $\epsilon^a \equiv \epsilon^a(x^\mu)$ the transformation parameters.

The generators of infinitesimal translations are here differential operators,

$$P_a = \frac{\partial}{\partial x^a} \equiv \partial_a, \tag{3.2}$$

which satisfy the commutation relations

$$[P_a, P_b] = 0. \tag{3.3}$$

3: - *) 7 +) *C4), * : &)) J)€• + C4)))* 8)€•
+ C • + F)+)

#* +)

$$x'^a = x^a + \epsilon^a,$$

3: /)4) J)

with $\epsilon^a \equiv \epsilon^a(x^\mu)$ the transformation parameters.

3: #* 8))) *))

@@0!#12! ! !3 4) ! B :)) ')) + + +
$ *)) 4)& / c " $ e E A 5 B0 *+
< + Z + $ *)) 4)& /
3: *)4

@ 0 !#12! ! !3 4) ! B # + ; 5) 0 # 1 /)
1)+ # 4 :) . / 5 , * 4)4 ' ; + : 40 *+
< @> *)4
3: A & + ===

< F & *)
< &

C4 @5 B0 #) + F & +) /

1.3 Affine and Projective Coordinate Systems

From a mathematical point of view it is more satisfactory to proceed in a slightly different manner. In fact, consider a k-vector space V, of finite dimension. We then define the set $\mathbb{P}(V)$ as the set of all 1-dimensional k-subspaces of $U \subset V$. We shall not pursue this *coordinate free* approach very far right now, but postpone it to Part 2 of this book. However, when V is of dimension $n < \infty$, we note that by fixing a k-basis $\{v_1, \ldots, v_n\}$ for V, V is identified with $\mathbb{A}_k^n = k^n$ by letting the element $v = \alpha_1 v_1 + \cdots + \alpha_n v_n$ correspond to the point $(\alpha_1, \ldots, \alpha_n) \in k^n$.

If W is an $n+1$-dimensional k-space, then similarly $\mathbb{P}(W)$ is identified with $\mathbb{P}_k^n$ by fixing a basis $w_0, \ldots, w_n$ for W and letting the point $(\alpha_0 : \ldots : \alpha_n)$ correspond to the subspace spanned by $\alpha_0 w_0 + \cdots + \alpha_n w_n$.

It is clear that switching to a different basis for V, say

$$\begin{aligned}
&\alpha_{1,1} v_1 + \cdots + \alpha_{1,n} v_n \\
&\alpha_{2,1} v_1 + \cdots + \alpha_{2,n} v_n \\
&\vdots \\
&\alpha_{n,1} v_1 + \cdots + \alpha_{n,n} v_n
\end{aligned}$$

For projective n-space we have to work with polynomials in the variables $X_0, X_1, \dots, X_n$, with coefficient from the ground field k, say $\mathbb{R}$ or $\mathbb{C}$ as the case may be. But here we encounter the difficulty that a polynomial $Q(X_0, X_1, \dots, X_n)$ may vanish at the point $(a_0, a_1, \dots, a_n)$, yet have a non-zero value at some $(b_0, b_1, \dots, b_n)$, even though $(a_0, a_1, \dots, a_n) \sim (b_0, b_1, \dots, b_n)$. But if *all the monomials occurring in Q have the same total degree*, then this will not happen. Such a polynomial is called *a homogeneous polynomial.* For a homogeneous polynomial $Q(X_0, X_1, \dots, X_n)$ we therefore are able to make the following definitions:

$$V_+(Q) = \{(a_1 : \dots : a_n) \in \mathbb{P}^n_k \mid Q(a_0, \dots, a_n) = 0\}$$
$$D_+(Q) = \{(a_0 : \dots : a_n) \in \mathbb{P}^n_k \mid Q(a_0, \dots, a_n) \neq 0\}$$

and

$$V_+(Q_1, \dots, Q_m) = \left\{ (a_0, \dots, a_n) \in \mathbb{A}^n_k \,\middle|\, \begin{array}{l} Q_i(a_0, \dots, a_n) = 0 \\ \text{for all } i = 1, \dots, m \end{array} \right\}.$$

Subsets of the last type are referred to as *projective subsets*, and in analogy to the affine case as projective hypersurfaces if there is only one polynomial, and finally, as a projective hyperplane if the polynomial is a linear form, that is, of the type $L = b_0X_0 + \cdots + b_nX_n$, with $b_i \in k$. A common name for affine and projective subsets of $\mathbb{A}^n_k$ and $\mathbb{P}^n_k$, respectively, is *algebraic subsets.*

1.3 Affine and Projective Coordinate Systems

From a mathematical point of view it is more satisfactory to proceed in a slightly different manner. In fact, consider a k-vector space V, of finite dimension. We then define the set $\mathbb{P}(V)$ as the set of all 1-dimensional k-subspaces of $U \subset V$. We shall not pursue this *coordinate free* approach very far right now, but postpone it to Part 2 of this book. However, when V is of dimension $n < \infty$, we note that by fixing a k-basis $\{v_1, \dots, v_n\}$ for V, V is identified with $\mathbb{A}^n_k = k^n$ by letting the element $v = \alpha_1 v_1 + \cdots + \alpha_n v_n$ correspond to the point $(\alpha_1, \dots, \alpha_n) \in k^n$.

If W is an $n+1$-dimensional k-space, then similarly $\mathbb{P}(W)$ is identified with $\mathbb{P}^n_k$ by fixing a basis $w_0, \dots, w_n$ for W and letting the point $(\alpha_0 : \dots : \alpha_n)$ correspond to the subspace spanned by $\alpha_0 w_0 + \cdots + \alpha_n w_n$.

It is clear that switching to a different basis for V, say

$$\begin{array}{l} \alpha_{1,1}v_1 + \cdots + \alpha_{1,n}v_n \\ \alpha_{2,1}v_1 + \cdots + \alpha_{2,n}v_n \\ \vdots \\ \alpha_{n,1}v_1 + \cdots + \alpha_{n,n}v_n \end{array}$$

corresponds to the change of coordinate system in $\mathbb{A}^n_k$ given by

$$\begin{aligned}
&\overline{X}_1 = \alpha_{1,1}X_1 + \cdots + \alpha_{1,n}X_n \\
&\overline{X}_2 = \alpha_{2,1}X_1 + \cdots + \alpha_{2,n}vX_n \\
&\vdots \\
&\overline{X}_n = \alpha_{n,1}X_1 + \cdots + \alpha_{n,n}X_n.
\end{aligned}$$

Similarly, switching to the new basis for W,

$$\begin{aligned}
&\alpha_{0,0}w_0 + \cdots + \alpha_{0,n}w_n \\
&\alpha_{1,0}w_0 + \cdots + \alpha_{1,n}w_n \\
&\vdots \\
&\alpha_{n,0}w_0 + \cdots + \alpha_{n,n}w_n
\end{aligned}$$

corresponds to the change of projective coordinate system in $\mathbb{P}^n_k$ given by

$$\begin{aligned}
&\overline{X}_0 = \alpha_{0,0}X_0 + \cdots + \alpha_{0,n}X_n \\
&\overline{X}_1 = \alpha_{1,0}X_0 + \cdots + \alpha_{1,n}vX_n \\
&\vdots \\
&\overline{X}_n = \alpha_{n,0}X_0 + \cdots + \alpha_{n,n}X_n.
\end{aligned}$$

Now let $L = a_0X_0 + \cdots + X_n$, then by switching to a new projective coordinate system, $D_+(L)$ can be made to correspond to $D_+(\overline{X}_0)$ and thus be identified with $\mathbb{A}^n_k$, while $V_+(L)$ is identified with $V_+(\overline{X}_0)$ and thus with the projective $n-1$-space $\mathbb{P}^{n-1}_k$. In general $V_+(X_0)$ is referred to as the *hyperplane at infinity.*

It is clear that a hypersurface in $\mathbb{A}^n_k$, in particular a curve in $\mathbb{A}^2_k$, given by a polynomial of degree d in the coordinates $x_1, \dots, x_n$ will be given by an equation of the same degree d in the coordinates $\overline{x}_1, \dots, \overline{x}_n$. The same observation holds in the projective case.

Rather than to view this as moving from one coordinate system to a new coordinate system with a new origin in general, it may regarded as a mapping from $\mathbb{A}^n_k$ to itself, a so called affine transformation:

$$\begin{gathered}
\mathbb{A}^n_k \longrightarrow \mathbb{A}^n_k \\
(x_1, \dots, x_n) \mapsto (\overline{x}_1, \dots, \overline{x}_n)
\end{gathered}$$

where

$$\begin{bmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_n \end{bmatrix} = \begin{bmatrix} \alpha_{1,1} \dots \alpha_{1,n} \\ \vdots \\ \alpha_{n,1} \dots \alpha_{n,n} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$$

and where $(b_1, \dots, b_n)$ is the new origin. The properties of affine subsets or configurations of affine subsets which are preserved by the affine transformations are referred to as *affine properties.* Affine properties include incidence

Fig. 1.1 Gérard Desargues. Illustration by the author

(that a point lies on a line, or a line passes through a point), collinearity (that several points lie on a common line), concurrency (that several lines pass through a common point). The properties which are not affine include being a circle, while being an ellipse is an affine property. Over $\mathbb{R}$ the property for two lines in $\mathbb{A}^2_{\mathbb{R}}$ forming a right angle is not affine, while the property for a line to bisect in equal parts the angle formed by two given lines, is affine. The property of tangency is affine.

Projective properties are defined analogously: all projective properties are preserved by projective transformations. As for the affine case we may also express this by saying that the projective properties are independent of choice of projective coordinate system. Obviously the points at infinity are not preserved, however.

We now need the following useful observation, the proof of which is of course well known, and is explained in detail when $n = 2$ and $k = \mathbb{R}$ in Chap. 11 of [27], Chap. 12 of [28]. The proof given there carries over to the general case with obvious modifications. We shall not repeat it here.

Proposition 1.1 *Given $n + 2$ points $P_1, P_2, \dots, P_{n+2} \in \mathbb{P}^n_k$ no three of which are collinear, as well as another set of points $P'_1, P'_2, \dots, P'_{n+2} \in \mathbb{P}^n_k$ with the same property. Then there exists a projective transformation G of $\mathbb{P}^n_k$ onto itself, mapping P_i to P'_i, $i = 1, 2, \dots, n$.*

1.4 The Theorem of Desargues

The *Theorem of Desargues* was important in the development of projective geometry. Gérard Desargues came from a wealthy family, and had many influential mathematical friends such as Rene Descartes and the father and son Pascal. He belonged to the famous and influential circle around the secular monk *Marin Mersenne.*

In [27] and [28] the theorem is proved only for $\mathbb{P}^2_k$ and it is assumed that $k = \mathbb{R}$ for this result. But as we shall see below, that assumption is not needed.

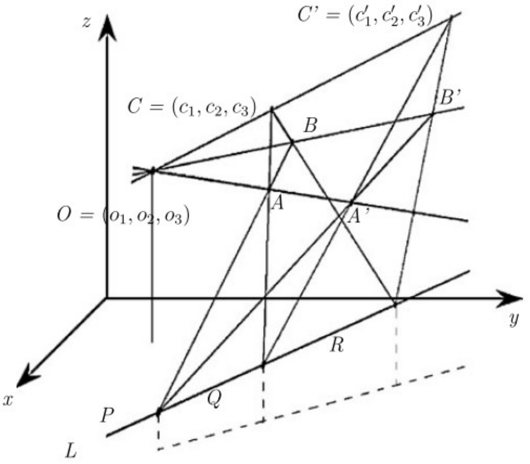

Fig. 1.2 Desargues' theorem, picture with $k = \mathbb{R}$ in the affine piece $D_+(X_0) \subset \mathbb{P}^3_k$

Theorem 1.2 *Let two triangles ABC and $A'B'C'$ be given in $\mathbb{P}^3_k$, such that $A \neq A'$, $B \neq B'$ and $C \neq C'$. Then if the lines through corresponding vertices pass through the same point O, the intersections of the prolongations of corresponding sides will intersect in points lying on the same line L.*

Remark When the triangles have the property that the lines through corresponding vertices pass through O, then O is referred to as the center of perspective, and the triangles are said to be perspective from the point O. If the intersections of the prolongations of corresponding sides intersect in points lying on L, then L is called the *line of perspective*, and the triangles are said to be perspective from the line L.

Proof We introduce notation as in Fig. 1.2. The coordinates of the points are labelled analogously to those of O, C and C'.

Evidently, no three of the points

$$(0:0:1:0), \quad (0:1:0:0), \quad (0:0:0:1) \quad \text{and} \quad (1:0:0:0)$$

are collinear. It is easily seen that the same is true for the points O, P, Q and A. Thus by Proposition 1.1 there is a linear transformations which makes these four points correspond, and hence, after a change of coordinate system, we may assume that

$$O = (0:0:1:0), \quad P = (0:1:0:0),$$
$$Q = (0:0:0:1) \quad \text{and} \quad A = (1:0:0:0:0).$$

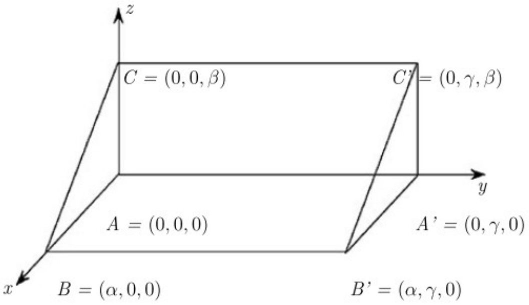

Fig. 1.3 Desargues' theorem, after a good choice of coordinate system in $\mathbb{P}^3_{\mathbb{R}}$

Then a line through O, P and Q, respectively, is given in $D_+(X_0) = \mathbb{A}^2_k$ on parametric form as

$$L_O = \{(x_1, t, z_1) \mid t \in k\}, \qquad L_P = \{(t, y_2, z_2) \mid t \in k\},$$
$$L_Q = \{(x_3, y_3, t) \mid t \in k\}.$$

Since moreover $A = (0,0,0)$, we therefore get in addition that

$$B = (\alpha,0,0), \quad C = (0,\beta,0), \quad A' = (0,\gamma,0),$$
$$B' = (\alpha,\gamma,0) \quad \text{and} \quad C' = (0,\beta,\gamma).$$

The situation is illustrated in Fig. 1.3.

We identify the point $(x,y,z) \in \mathbb{A}^3_{\mathbb{R}}$ with the point $(1:x:y:z) \in D_+(X_0) \subset \mathbb{P}^3_{\mathbb{R}}$ as usual, and get the following parametric forms for the lines BC, $B'C'$ and PQ, where we write $(ua:ub:uc:ud)$ as $u(a:b:c:d)$ to simplify the notation:

$$BC = \{u_1(1:0:0:\beta) + v_1(1:\alpha:0:0) \mid (u_1,v_1) \neq (0,0)\}$$
$$B'C' = \{(u_2(1:0:\gamma:\beta) + v_2(1:\alpha;\gamma:0) \mid (u_2,v_2) \neq (0,0)\}$$
$$PC = \{u_3(0:1:0:0) + v_3(0:0:0:1) \mid (u_3,v_3) \neq (0,0)\}.$$

It is now easily verified that the two lines BC and $B'C'$ intersect in the point $(0:\alpha:0:-\beta)$, which lies on the line PQ. Thus the claim follows. □

This theorem has the following corollary, which is the converse of the theorem in the plane:

Corollary 1.3 *Let two triangles ABC and $A'B'C'$ be given in $\mathbb{P}^2_{\mathbb{R}}$, and assume that $A \neq A'$, $B \neq B'$ and $C \neq C'$. If the triangles are perspective from a line, then they are perspective from a point.*

Proof The proof is immediate by the *Principle of Duality* for $\mathbb{P}^2_{\mathbb{R}}$, which we shall prove in the next section. Here we find that the dual result is actually a *converse* to the assertion of the theorem. The figure in the proof above is actually *self dual.* □

1.5 Duality for $\mathbb{P}^2_k$

We now consider $\mathbb{P}^2_k$, and let P denote some statement involving *points, lines incidence.*

Then the dual statement $P^\vee$ is defined as the statement obtained by interchanging the words "point" and "line", and keeping "incidence", with the adjustments of language which may be necessary to obtain natural geometrical statements.

For example, consider the statement below:

Statement P: Through two distinct points Q_1 and Q_2 there always pass a uniquely determined line ℓ.

Then the dual statement is the following:

Statement $P^\vee$: Two distinct lines q_1 and q_2 always intersect in a uniquely determined point L.

For $\mathbb{P}^2_k$ we have the so called *principle of duality*:

Theorem 1.4 (Duality for $\mathbb{P}^2_k$) *If P denotes a true statement about $\mathbb{P}^2_k$ dealing with points, lines and incidence, then the dual statement $P^\vee$ is also a true statement.*

Proof The statement P may be translated into a possibly infinite set of equations involving the projective coordinates of the points and the coefficients of the equations of the lines occurring in the statement. The equations will all be of the form

$$A_0\alpha_0 + A_1\alpha_1 + A_2\alpha_2 = 0,$$

where $A = (A_0 : A_1 : A_2)$ is a point and α_0, α_1 and α_2 are the coefficients in the equation for a line in $\mathbb{P}^2_k$, so the line is given by

$$\alpha_0X_0 + \alpha_1X_1 + \alpha_2X_2 = 0.$$

If we denote this line by ℓ, then the statement "A lies on ℓ" is equivalent to the relation above. For every point in $\mathbb{P}^2_k$ we now let correspond a *line* in $\mathbb{P}^2_k$ given by the equation whose coefficients are the coordinates of the point, and to every line we let correspond the point whose projective coordinates are the coefficients of the equation giving the line.

Then the set of algebraic equations between the coordinates of points in $\mathbb{P}^2_k$ and coefficients of lines in $\mathbb{P}^2_k$ which expresses the truth of the statement P is the same as the collection of relations which expresses the truth of $P^\vee$.

C4 5 0

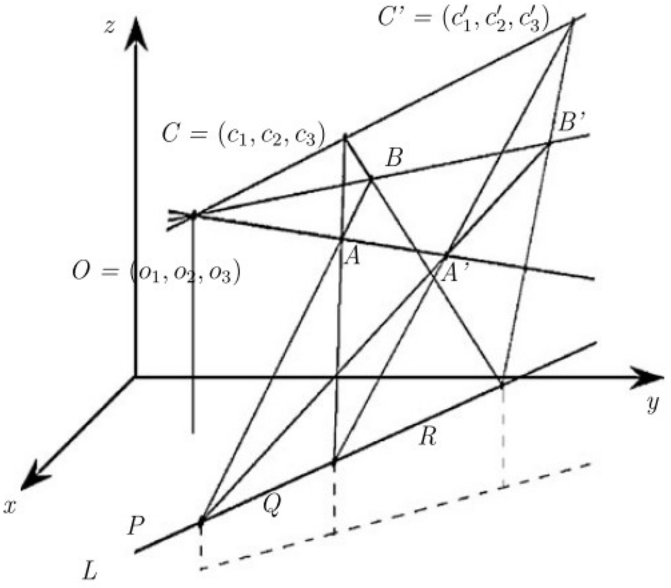

Fig. 1.2 Desargues' theorem, picture with $k = \mathbb{R}$ in the affine piece $D_+(X_0) \subset \mathbb{P}^3_k$

Theorem 1.2 *Let two triangles* ABC *and* $A'B'C'$ *be given in* $\mathbb{P}^3_k$, *such that* $A \neq A'$, $B \neq B'$ *and* $C \neq C'$. *Then if the lines through corresponding vertices pass through the same point* O, *the intersections of the prolongations of corresponding sides will intersect in points lying on the same line* L.

@B0 !#12! ! !3 4) ! B .#A #) 4) + $
#**) # * :/ * . + * 4)4 ' ; + : 4 5 0 *+
< *)4
< .#A *)) . / # 4 A :) /
3: A & + ===

C4 >5 >0

B 3*) 5 & +0 /4 >
B 3*) A) /4
B /4 ; B
B 3*) /4 ; B
B)) + % 1 4) 9 + 3*) B
B 4 ; B
B > 3*) /4 B
B @ # J 3*) B>
B - 4 % - 4 /4 B
B B 4)/ 3*) B
B 3*) & + /4 B
B $ $)) *

B 3*) /4) #
B .) /4 $) #+)*
B #**))/ 3*) .) /4 B
B #**) 3:) 4 / +)
B > ' : 4)* %) + 1) (>
B @ 78
(( ) + 3 /4
* 4 (($ /4
* 4 ((E 4 &
" *) 4
3 4) /4
#** 8) 4 3) .)) + 7 ' + >
* 4 78) 3
) 4 #**)
#**) # $)) 4 /4) 1 :
#**) +) 4 $ " 1 :
' : 4)* %) + 1) (
> 78

7 & *
7 & * &)
1 * 4 7 & *
4 7 & * ")4)
$ 7 & *) *
E 4 $))
#**) %) H * & $ B
7 & *) *) >
$ 7 & * $) * >
E 4 7 & * $)) >
#**) A) 4 D) >
> ' : 4)* %) + 1) (>
@ 78 @

C4 5 @ 0)*
7 & *
- + /) :)) 4 5) , *
)) 4 0 &) *)) * :)) +,)))) 4
-)* 8) :) + & *) & 5)
) 0 G + & * : 1) + 1 -) /, & *)
) ,)/, 1))) 4) (4)
/)))/ 5D) + Z) 4 & : 0 - *))) , & *
) +)) +) : / 5) +) /0 * +)& 4
)) 4 , &) +)) * ,)
+ Z + > , * & *) :) / *)) &
$ & * *))) 1 8*): / *
& *)) *)))) 4 ,) & * / +) G 4
*) *))) * : , :))&) +) Z +))
*))*): *) 4 +)) $))
* +) *)

C4 5 @ 0 1 * 4 7 & *

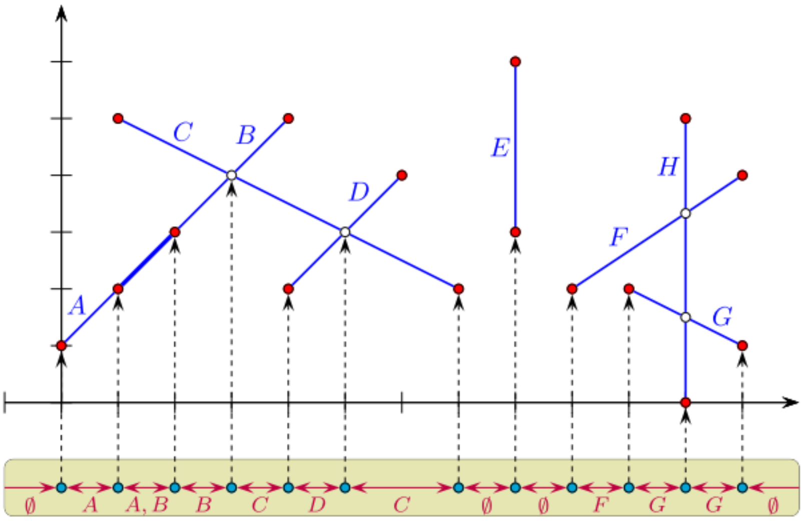
C
B
D
E
H
F
A
G
∅
A
A, B
B
C
D
C
∅
∅
F
G
G
∅

D 4 $ & * 4 4 ,
): + #, , ; ,) + 86 & * 6
4 ** $ 9) +)4)
) : ,) +)4) & 8 *
*)) * & *)) +
$): 4) +) +)
4) +4) + *)/ + : +4 %)
8 +4) *) &)*,) +
) +4) * * / &)

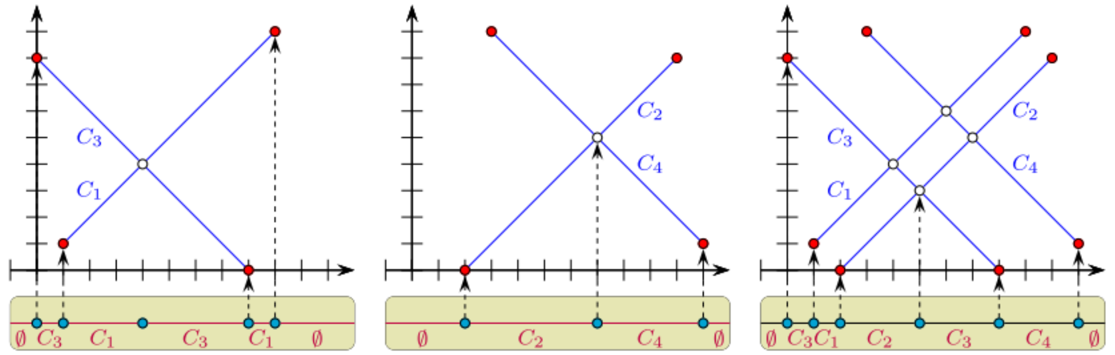
C_3
C_1
∅
C_3
C_1
C_3
C_1
∅
C_2
C_4
∅
C_2
C_4
∅
C_3
C_1
C_2
C_4
∅
C_3
C_1
C_2
C_3
C_4
∅

(a) The envelope diagram of two curves C_1 and C_3.

(b) The envelope diagram of two curves C_2 and C_4.

(c) The envelope diagram of all four curves.

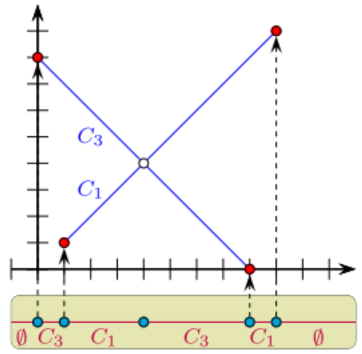
C_3
C_1
∅
C_3
C_1
C_3
C_1
∅

(a) The envelope diagram of two curves C_1 and C_3.

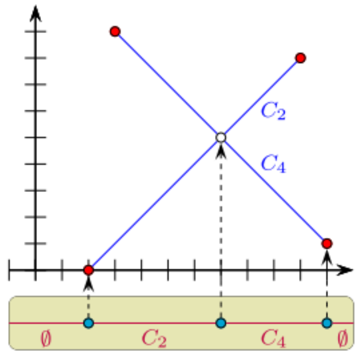

C_2
C_4
$\emptyset$
C_2
C_4
$\emptyset$

(b) The envelope diagram of two curves C_2 and C_4.

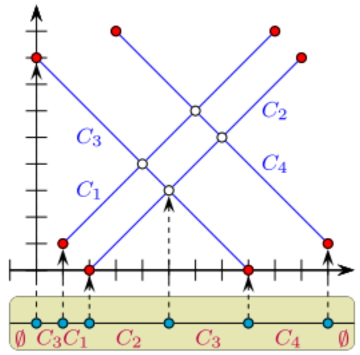

C_2
C_3
C_4
C_1
$\emptyset$
C_3
C_1
C_2
C_3
C_4
$\emptyset$

(c) The envelope diagram of all four curves.

C4 > 5 0

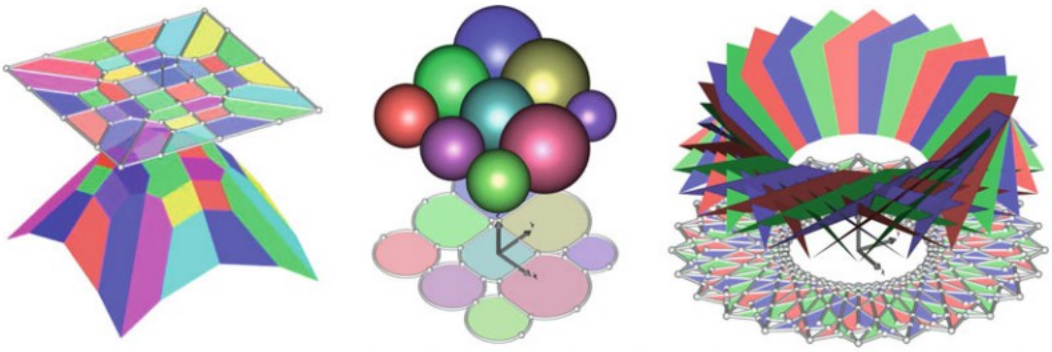

(a) The lower envelope of planes (drawn above the planes). **(b)** The lower envelope of spheres. **(c)** The lower envelope of triangles.

D 4 A & *) *)

- *) E) + 5)0)) * : *) , F),)
)€• : &) C , *) :) &)) & 8) " &)9 C ,
) &)))&) *)+)) : *)
WWW

@ 0 !#12! ! !3 4) ! B ")) 4) + 7 +))
4) + NE & 6 .)) *+
3: A & + ===

< ")) 4)
+ 7 +
)) 4)
+ NE &
.)), + , , * +)

- +

ˆ) 4) *
ˆ) 4 & +) 4)_
)) +)+ 4 &)
+))
3 4 + :
ˆ : _
A) 4)))
A)) +)
A) 4 +)))
A)) Ÿ &)
9))
> $) 4
@ #** +
-) * *)))** +)
7 + , ; :) 4)))
-))) + 7 + B
-)
E)) + *) 4) *))
+ +)))
> A) 4))) Ÿ +)_ @
@ D 4 4))) 4 J)+)
.) +
+) + +
& ** *) + + >>
- * + + >
- + * + 4)) +) >
> #** + @>
> .) @
> # +) + @
> A 4 +
> #** + @
> - * +
> > A) 4) B
> @)* + * Ÿ B
@ A) &)) B
@ A) &)) +))
@ A) &)) +) *
@ A) &)) + *) , +)), + + + >
@) 4 + B
@ > A) &)))) + 4)))
A) * +)) 4)) * +) >
A))))))) >
A) * +))) &))
A * +)) +)) * +)

) + .) ' J *)) * +) >
B A)) $)) *
B -) +
B - * : +)& 4) +) 4
B) +)) + >
B ")9 +) 4 + B
B > . +)
B @ A) +)9 + *)&
A)) $)) 4 4)
A * 9 +
A))) +)
9 *)
A) * 9 + .)) + +) E$
A)** +) * +) + + 4)
* :) >
A))**) +)& 4) * : >
*)))**) >@
- *) +) ˆ >
AN &) @
> - + +) ˆ @

- *)9 Ÿ + _ @
A) 4) + *)9 * + + ((@B
A *)9 * +)) & Ÿ) 4)
-* NE & >
- * : + + E & *)) * & +
))) &))_
))) + 7 + + ; : B
")) + 7 + B
")) + ; : B
E4)4) 9) 4 9) B
' .) + * 9) 4 :) B>
' " 9 4) B>
' #) & _ B>
' Ÿ _ B>
' 9)_ B@
' > #) +))9)) BB
' @ + +)
' : 4))

0 !#12! ! !3 4) ! B ")& + ')) 5) 0 #
. #**) ") D ' (S ') 5 0 *+
3: A & + ===

B ") . / &) ") D
B &)) + &) &
B D) Z +) + .)) &)
B)) & Z + B
B $.) ' $

C4 5 @0 B ") . / &) ") D

C4 5 0

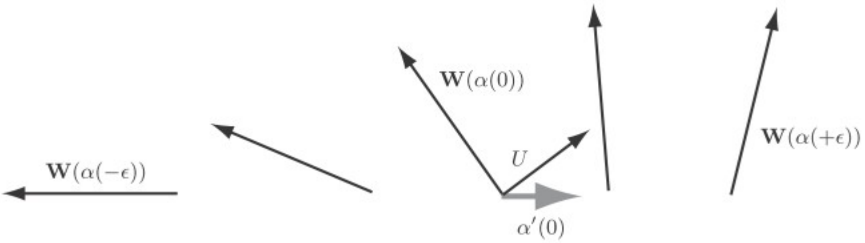
$\mathbf{W}(\alpha(-\epsilon))$
$\mathbf{W}(\alpha(0))$
U
$\alpha'(0)$
$\mathbf{W}(\alpha(+\epsilon))$

Fig. 8.1. The derivative of a vector field **W** at the point $p = \alpha(0)$, in the direction of the vector $\alpha'(0)$, is the vector U.

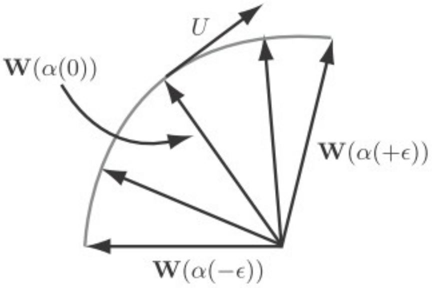
U
$\mathbf{W}(\alpha(0))$
$\mathbf{W}(\alpha(+\epsilon))$
$\mathbf{W}(\alpha(-\epsilon))$

Fig. 8.2. Superimposing vectors from the field **W** creates a parameterized curve. The derivative vector U is then a tangent vector to this curve.

C4 5 0

B D) Z +) + .)) &)
3)**) :F +) 4 / /)) N
+ & 4) 1 4 / *)(4, & &)) *)
) / :)) 4 *))) *)) 1 ,) +
)) Z + . * *) : + & +)/
) Z + &) * 8 $) 4)4
+)) +/

0 !#12! ! !3 4) ! B ") . /6 $ /,
+ ,) + #**) * 4)4 % X (5 0 *+
3: A & + ===

C4 5 B0
E &) 1 4 + / ') D) (.)
7 +) ")) 8 #**)

8 +) + ") . / $ J
4 7 +) 7 : ++ 4 1 4 + .)*

@ & 4 *))) .) 9 + ") . / >

@ ") . /) ' 4/ % * &

0 !#12! ! !3 4) ! B D) ' 8 5) 0 #
4 :) #**) . /6 . $ 4/ -- * 4 -)) : 4
5 0 *+
< *)4 -- . $ 4/ --
3: A & + ===

0 !#12! ! !3 4) ! B D) ' 8 5) 0 #
#8) #**) . /6 . $ 4/ - * 4 -)) : 4
5 0 *+
< *)4 - . $ 4/ -
3: A & + ===

0 !#12! ! !3 4) ! B -) . /) + -
#**) 5 @, * 4 H)*) 0 *+

) - . / " & 4 D ") / D) 1))

) +, " & 4) + ") / D)
) +
) +) + +) /
78) *) + >
" & 4 ' $
" & 4
78) * " & 4
& 8 D) + ' 4) " & 4
& 8 D
' 4) " & 4
A 4 + $)) @
> ") / D) 1)) " & + & 8
D
> # Z) + ") # Z +) /
> $) 4 *) , ')
) + 1))
>)) $) *
@ .) 9 + /)4) $) + F $
@ .) 9 + /)4) $
@ F $ @
@ " & 4 ' :) + #) 4
9) # 4

) -- - + ") ") . /
> 7 ") . /
>) +) + $) 4 *)
> 1))
> # Z
> $
> > &)) " &) & @
> @ . +
>)) $) * B
> B 1) &)

> B 1 + G + $) *
> B &)) " &) &) + 1 &)
> B D)) +
> A & &) 51)) 0
> :) +) + 7 : ++ 4 &) @
> :) + @
> 7 : ++ 4 &)

>0 !#12! ! !3 4) ! B -) . /) + -
#**) 6 3 3) #) I B ' +)/, -.#-# - A : , 9
1 * : *+

C4 5 0

-) . / .)) *)

#:) G + * 8*)) + Z
+) .)) *) G + */, /
)) ,) [/* J) , 4) 9 + +): / * :
): / + .)) *)

@0‹Œ•Ži••Žq‘’{“n”oŽ•{j“|o‘–q•‘‘‘‰–‰‰–‰‰–Ž—n~k~˜™jŽ–-•‘‘‘®¯˜{›‹°™”±‹powut‹‘‹²±~n}³‹´››™³›‹{˜
. .)* $ / * 4 5 0 *+
< @ *)4

C4

) + $ +))) 4 A .)*

") 4 $, 3 *)) .)* ,)) .)* ,
) +)) .)*)) #)

1) 4) + J) 1 *)
)) .)*

3: - *) ===

C4 5> 0 4

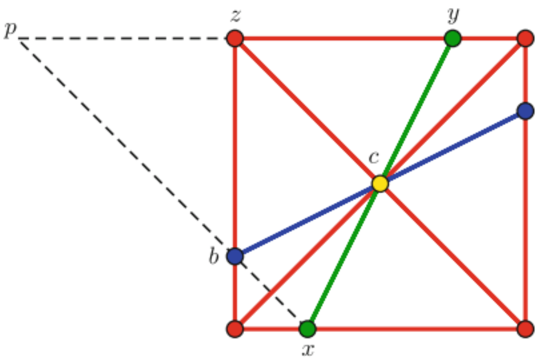

D 4 # µ , , , ¶ : (+ *
3: E)** I $ 5 *) 4 0 78 X 8

Thirty Essays on Geometric Graph Theory*

Greg Aloupis, Brad Ballinger, Sébastien Collette, Stefan Langerman, Attila Pór, and David R. Wood

Abstract This paper studies problems related to visibility among points in the plane. A point x *blocks* two points v and w if x is in the interior of the line segment $\overline{vw}$. A set of points P is *k-blocked* if each point in P is assigned one of k colors, such that distinct points $v, w \in P$ are assigned the same color if and only if some other point in P blocks v and w. The focus of this paper is the conjecture that each k-blocked set has bounded size (as a function of k). Results in the literature imply that every 2-blocked set has at most 3 points, and every 3-blocked set has at most 6 points. We prove that every 4-blocked set has at most 12 points, and that this

---

*This is the full version of an extended abstract presented at the 26th European Workshop on Computational Geometry (EuroCG'10).

G. Aloupis • S. Collette
Chargé de Recherches du F.R.S.-FNRS, Département d'Informatique,
Université Libre de Bruxelles, Brussels, Belgium
e-mail: aloupis.greg@gmail.com; sebastien.collette@ulb.ac.be

B. Ballinger
Department of Mathematics, Humboldt State University, Arcata, CA, USA
e-mail: bjb86@humboldt.edu

S. Langerman
Maître de Recherches du F.R.S.-FNRS, Département d'Informatique,
Université Libre de Bruxelles, Brussels, Belgium
e-mail: stefan.langerman@ulb.ac.be

A. Pór
Department of Mathematics, Western Kentucky University, Bowling Green, KY, USA
e-mail: attila.por@wku.edu

D.R. Wood (✉)
Department of Mathematics and Statistics, The University of Melbourne, Melbourne, Australia
e-mail: woodd@unimelb.edu.au.

J. Pach (ed.), *Thirty Essays on Geometric Graph Theory*,
DOI 10.1007/978-1-4614-0110-0_4,
© Springer Science+Business Media New York 2013

bound is tight. In fact, we characterize all sets $\{n_1,n_2,n_3,n_4\}$ such that some 4-blocked set has exactly n_i points in the ith color class. Among other results, for infinitely many values of k, we construct k-blocked sets with $k^{1.79\cdots}$ points.

1 Introduction

This paper studies problems related to visibility and blocking in sets of colored points in the plane. A point x *blocks* two points v and w if x is in the interior of the line segment $\overline{vw}$. Let P be a finite set of points in the plane. Two points v and w are *visible* with respect to P if no point in P blocks v and w. The *visibility graph* of P has vertex set P, where two distinct points $v,w \in P$ are adjacent if and only if they are visible with respect to P. A point set B *blocks* P if $P \cap B = \emptyset$ and for all distinct $v,w \in P$ there is a point in B that blocks v and w. That is, no two points in P are visible with respect to $P \cup B$, or alternatively, P is an independent set in the visibility graph of $P \cup B$.

A set of points P is *k-blocked* if each point in P is assigned one of k colors, such that each pair of points $v,w \in P$ are visible with respect to P if and only if v and w are colored differently. Thus, v and w are assigned the same color if and only if some other point in P blocks v and w. A *k-set* is a multiset of k positive integers. For a k-set $\{n_1,\ldots,n_k\}$, we say P is $\{n_1,\ldots,n_k\}$-blocked if it is k-blocked, and for some labeling of the colors by the integers $[k] := \{1,2,\ldots,k\}$, the ith color class has exactly n_i points, for each $i \in [k]$. Equivalently, P is $\{n_1,\ldots,n_k\}$-blocked if the visibility graph of P is the complete k-partite graph $K(n_1,\ldots,n_k)$. A k-set $\{n_1,\ldots,n_k\}$ is *representable* if there is an $\{n_1,\ldots,n_k\}$-blocked point set. See Fig. 1 for an example.

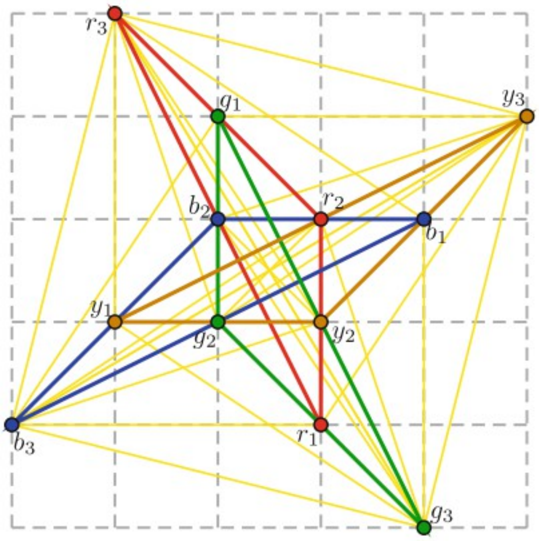

Fig. 1 A $\{3,3,3,3\}$-blocked point set

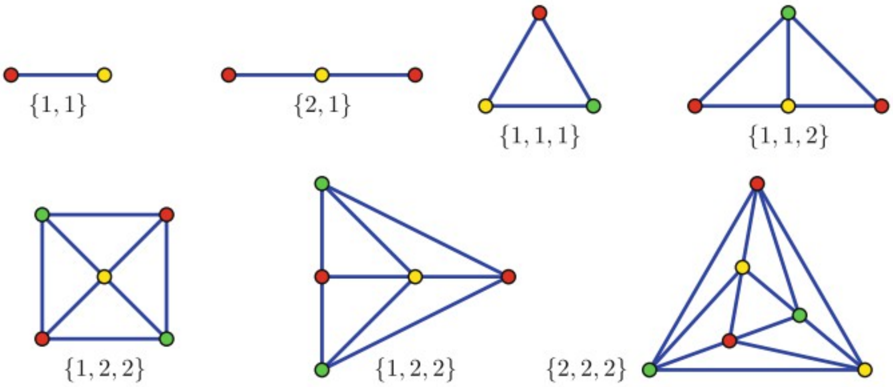

Fig. 2 The 2-blocked and 3-blocked point sets

The following fundamental conjecture regarding k-blocked point sets is the focus of this paper.

Conjecture 1. For each integer k, there is an integer n such that every k-blocked set has at most n points.

As illustrated in Fig. 2, the following theorem is a direct consequence of the characterization of 2- and 3-colorable visibility graphs by Kára et al. [5].

Theorem 2. *$\{1,1\}$ and $\{1,2\}$ are the only representable 2-sets, and $\{1,1,1\}$, $\{1,1,2\}$, $\{1,2,2\}$, and $\{2,2,2\}$ are the only representable 3-sets.*

In particular, every 2-blocked point set has at most three points, and every 3-blocked point set has at most six points. This proves Conjecture 1 for $k \leq 3$.

This paper makes the following contributions. Section 2 introduces some background motivation for the study of k-blocked point sets and observes that results in the literature prove Conjecture 1 for $k = 4$. Section 3 describes methods for constructing k-blocked sets from a given $(k-1)$-blocked set. These methods lead to a characterization of representable k-sets when each color class has at most three points. Section 4 studies the $k = 4$ case in more detail. In particular, we characterize the representable 4-sets, and conclude that the example in Fig. 1 is in fact the largest 4-blocked point set. Section 5 introduces a special class of k-blocked sets (so-called midpoint-blocked sets) that lead to a construction of the largest known k-blocked sets for infinitely many values of k.

Before continuing, we prove some basic properties about k-blocked point sets.

Lemma 3. *At most three points are collinear in every k-blocked point set.*

Proof. Suppose that four points p,q,r,s are collinear in this order. Thus, (p,q,r,s) is an induced path in the visibility graph. Thus, p is not adjacent to r and not adjacent to s. Thus, p, r, and s have the same color. This is a contradiction since r and s are adjacent. Thus, no four points are collinear. □

A set of points P is in *general position* if no three points in P are collinear.

Lemma 4. *Each color class in a k-blocked point set is in general position.*

Proof. Suppose on the contrary that three points from a single color class are collinear. Then no other points are in the same line by Lemma 3. Thus, two of the three points are adjacent, which is a contradiction. ☐

Note that "convex position" in this paper means "strictly convex position."

2 Some Background Motivation

Much recent research on blockers began with the following conjecture by Kára et al. [5].

Conjecture 5 (Big-Line-Big-Clique Conjecture [5]). For all integers t and ℓ, there is an integer n such that for every finite set P of at least n points in the plane:

- P contains ℓ collinear points, or
- P contains t pairwise visible points (that is, the visibility graph of P contains a t-clique).

Conjecture 5 is true for $t \leq 5$ but is open for $t \geq 6$ or $\ell \geq 4$; see [1,9]. Given that, in general, Conjecture 5 is challenging, Jan Kára suggested the following weakening.

Conjecture 6 ([9]). For all integers t and ℓ, there is an integer n such that for every finite set P of at least n points in the plane:

- P contains ℓ collinear points, or
- The chromatic number of the visibility graph of P is at least t.

Conjecture 5 implies Conjecture 6 since every graph that contains a t-clique has chromatic number at least t.

Proposition 7. *Conjecture 6 with $\ell = 4$ and $t = k+1$ implies Conjecture 1.*

Proof. Assume Conjecture 6 holds for $\ell = 4$ and $t = k+1$. Suppose there is a k-blocked set P of at least n points. By Lemma 3, at most three points are collinear in P. Thus, the first conclusion of Conjecture 6 does not hold. By definition, the visibility graph of P is k-colorable. Thus, the second conclusion of Conjecture 6 does not hold. This contradiction proves that every k-blocked set has fewer than n points, and Conjecture 1 holds. ☐

Thus, since Conjecture 5 holds for $t \leq 5$, Conjecture 1 holds for $k \leq 4$. In Sect. 4 we take this result much further, by characterizing all representable 4-sets, thus concluding a tight bound on the size of a 4-blocked set.

Part of our interest in blocked point sets comes from the following.

Proposition 8. *For all $k \geq 3$ and $n \geq 2$, the edge set of the k-partite Turán graph $K(n,n,\ldots,n)$ can be partitioned into a set of "lines," where*

- *Each line is either an edge or an induced path on three vertices;*
- *Every pair of vertices is in exactly one line; and*
- *For every line L, there is a vertex adjacent to each vertex in L.*

Proof. Let (i,p) be the pth vertex in the ith color class for $i \in \mathbb{Z}_k$ and $p \in \mathbb{Z}_n$ (taken as additive cyclic groups). We introduce three types of lines. First, for $i \in \mathbb{Z}_k$ and distinct $p,q \in \mathbb{Z}_n$, let the triple $\{(i,p),(i+1,p+q),(i,q)\}$ be a line. Second, for $i \in \mathbb{Z}_k$ and $p \in \mathbb{Z}_n$, let the pair $\{(i,p),(i+1,p+p)\}$ be a line. Third, for $p,q \in \mathbb{Z}_n$ and distinct nonconsecutive $i,j \in \mathbb{Z}_k$, let the pair $\{(i,p),(j,q)\}$ be a line. By construction, each line is either an edge or an induced path on three vertices.

Every pair of vertices in the same color class are in exactly one line (of the first type). Consider vertices (i,p) and (j,q) in distinct color classes. First, suppose that i and j are consecutive. Without loss of generality, $j = i+1$. If $q \neq p+p$, then (i,p) and (j,q) are in exactly one line (of the first type). If $q = p+p$, then (i,p) and (j,q) are in exactly one line (of the second type). If i and j are not consecutive, then (i,p) and (j,q) are in exactly one line (of the third type). This proves that every pair of vertices is in exactly one line. Moreover, every edge of $K(n,n,\dots,n)$ is in exactly one line, and the lines partition the edges set.

Since every line L is contained in the union of two color classes, each vertex in neither color class intersecting L is adjacent to each vertex in L. □

Proposition 8 is significant for Conjecture 5 because it says that $K(n,\dots,n)$ behaves like a "visibility space" with no four collinear points, but it has no large clique (for fixed k). Conjecture 1, if true, implies that such a visibility space is not "geometrically representable" for large n.

Let $b(n)$ be the minimum integer such that some set of n points in the plane in general position is blocked by some set of $b(n)$ points. Matoušek [6] proved that $b(n) \geq 2n-3$. Dumitrescu et al. [2] improved this bound to $b(n) \geq (\frac{25}{8} - o(1))n$. Many authors have conjectured or stated as an open problem that $b(n)$ is superlinear.

Conjecture 9 ([2, 6, 8, 9]). $\frac{b(n)}{n} \to \infty$ as $n \to \infty$.

Pór and Wood [9] proved that Conjecture 9 implies Conjecture 6, and thus implies Conjecture 1. That Conjecture 1 is implied by a number of other well-known conjectures, yet remains challenging, adds to its interest.

3 *k*-Blocked Sets with Small Color Classes

We now describe some methods for building blocked point sets from smaller blocked point sets.

Lemma 10. *Let G be a visibility graph. Let* $i \in \{1,2,3\}$. *Furthermore, suppose that if* $i \geq 2$, *then* $V(G) \neq \emptyset$, *and if* $i = 3$, *then not all the vertices of G are collinear. Let* G_i *be the graph obtained from G by adding an independent set of i new vertices, each adjacent to every vertex in G. Then* G_1, G_2, *and* G_3 *are visibility graphs.*

Proof. For distinct points p and q, let $\overleftarrow{pq}$ denote the ray that is (1) contained in the line through p and q, (2) starting at p, and (3) not containing q. Let $\mathcal{L}$ be the union of the set of lines containing at least two vertices in G.

$i = 1$: Since $\mathcal{L}$ is the union of finitely many lines, there is a point $p \notin \mathcal{L}$. Thus, p is visible from every vertex of G. By adding a new vertex at p, we obtain a representation of G_1 as a visibility graph.

$i = 2$: Let p be a point not in $\mathcal{L}$. Let v be a vertex of G. Each line in $\mathcal{L}$ intersects $\overleftarrow{vp}$ in at most one point. Thus, $\overleftarrow{vp} \setminus \mathcal{L} \neq \emptyset$. Let q be a point in $\overleftarrow{vp} \setminus \mathcal{L}$. Thus, p and q are visible from every vertex of G, but p and q are blocked by v. By adding new vertices at p and q, we obtain a representation of G_2 as a visibility graph.

$i = 3$: Let u, v, w be noncollinear vertices in G. Let p be a point not in $\mathcal{L}$ and not in the convex hull of $\{u, v, w\}$. Without loss of generality, $\overline{uv} \cap \overline{pw} \neq \emptyset$. There are infinitely many pairs of points $q \in \overleftarrow{up}$ and $r \in \overleftarrow{vp}$ such that w blocks q and r. Thus, there are such q and r both not in $\mathcal{L}$. By construction, u blocks p and q, and v blocks p and r. By adding new vertices at p, q, and r, we obtain a representation of G_3 as a visibility graph. □

Since no (≥ 3)-blocked set is collinear, Lemma 10 implies

Corollary 11. *If $k \geq 4$ and $\{n_1, \dots, n_{k-1}\}$ is representable and $n_k \in \{1, 2, 3\}$, then $\{n_1, \dots, n_{k-1}, n_k\}$ is representable.*

The representable (≤ 3)-sets were characterized in Theorem 2. In each case, each color class has at most three vertices. Now we characterize the representable (≥ 4)-sets, assuming that each color class has at most three vertices.

Proposition 12. *$\{n_1, \dots, n_k\}$ is representable whenever $k \geq 4$ and each $n_i \leq 3$, except for $\{1, 3, 3, 3\}$.*

Proof. Say the k-set $\{n_1, \dots, n_k\}$ *contains* the $(k-1)$-set $\{n_1, \dots, n_{i-1}, n_{i+1}, \dots, n_k\}$ for each $i \in [k]$. We proceed by induction on $k \geq 4$. For the base case, assume that $k = 4$. If $\{n_1, n_2, n_3, n_4\}$ contains $\{1, 1, 1\}$, $\{1, 1, 2\}$, $\{1, 2, 2\}$, or $\{2, 2, 2\}$, then $\{n_1, n_2, n_3, n_4\}$ is representable by Theorem 2 and Corollary 11. In each remaining case, the following table describes a construction, except for $\{1, 3, 3, 3\}$, which is not representable, as proved in Lemma 18.

$\{1, 1, 1, x\}$	Contains $\{1, 1, 1\}$	$\{1, 1, 2, x\}$	Contains $\{1, 1, 2\}$
$\{1, 1, 3, 3\}$	Figure 1 minus $\{r_1, g_3, r_3, g_1\}$	$\{1, 2, 2, x\}$	Contains $\{1, 2, 2\}$
$\{1, 2, 3, 3\}$	Figure 1 minus $\{g_1, g_3, r_3\}$	$\{2, 2, 2, x\}$	Contains $\{2, 2, 2\}$
$\{2, 2, 3, 3\}$	Figure 1 minus $\{g_3, r_3\}$	$\{2, 3, 3, 3\}$	Figure 1 minus g_3
$\{3, 3, 3, 3\}$	Figure 1		

Now assume that $k \geq 5$: Consider a k-tuple $\{n_1, \dots, n_k\}$ with each $n_i \leq 3$. Say $n_1 \leq \cdots \leq n_k$. If $\{n_1, \dots, n_{k-1}\} \neq \{1, 3, 3, 3\}$, then by induction $\{n_1, \dots, n_{k-1}\}$ is representable, which implies that $\{n_1, \dots, n_k\}$ is representable by Corollary 11. Otherwise, $k = 5$ and $\{n_1, \dots, n_5\} = \{1, 3, 3, 3, 3\}$, which contains the representable 4-tuple $\{3, 3, 3, 3\}$, implying $\{n_1, \dots, n_5\}$ is representable by Corollary 11. □

4 4-Blocked Point Sets

As we saw in Sect. 2, Conjecture 1 holds for $k \le 4$. In this section we study 4-blocked point sets in more detail. First, we derive explicit bounds on the size of 4-blocked sets from other results in the literature. Then, following a more detailed approach, we characterize all representable 4-sets, to conclude a tight bound on the size of 4-blocked sets.

Proposition 13. *Every* 4*-blocked set has less than* 2^{790} *points.*

Proof. Abel et al. [1] proved that every set of at least $\mathrm{ES}(\frac{(2\ell-1)^{\ell}-1}{2\ell-2})$ points in the plane contains ℓ collinear points or an empty convex pentagon, where $\mathrm{ES}(k)$ is the minimum integer such that every set of at least $\mathrm{ES}(k)$ points in general position in the plane contains k points in convex position. Let P be a 4-blocked set. The visibility graph of P is 4-colorable, and thus contains no empty convex pentagon. By Lemma 3, at most three points in P are collinear. Thus, $|P| \le \mathrm{ES}(400) - 1$ by the above result with $\ell = 4$. Tóth and Valtr [12] proved that $\mathrm{ES}(k) \le \binom{2k-5}{k-2} + 1$. Hence, $|P| \le \binom{795}{398} < 2^{790}$. □

Lemma 14. *If P is a blocked set of n points with m points in the largest color class, then $n \ge 3m - 3$ and $n \ge (\frac{33}{8} - o(1))m$.*

Proof. If S is the largest color class, then $P - S$ blocks S. By Lemma 4, S is in general position. By the results of Matoušek [6] and Dumitrescu et al. [2] mentioned in Sect. 2, $n - m \ge 2m - 3$ and $n - m \ge (\frac{25}{8} - o(1))m$. □

Proposition 15. *Every* 4*-blocked set has fewer than* 2^{578} *points.*

Proof. Let P be a 4-blocked set of n points. Let S be the largest color class in P. Let $m := |S| \ge \frac{n}{4}$. By Lemma 14, $n \ge (\frac{33}{8} - o(1))m \ge (\frac{33}{32} - o(1))n$. Thus, $o(n) \ge \frac{n}{32}$. Hence, n is bounded. A precise bound is obtained as follows. Dumitrescu et al. [2] proved that S needs at least $\frac{25m}{8} - \frac{25m}{2\ln m} - \frac{25}{8}$ blockers, which come from the other color classes. Thus, $n - m \ge \frac{25m}{8} - \frac{25m}{2\ln m} - \frac{25}{8}$, implying $n \ge \frac{33m}{8} - \frac{25m}{2\ln m} - \frac{25}{8}$. Since $\frac{n}{4} \le m \le n$, we have $\frac{25n}{2\ln n} + \frac{25}{8} \ge \frac{n}{32}$. It follows that $n \le 2^{578}$. □

The next result is the simplest known proof that every 4-blocked point set has bounded size.

Proposition 16. *Every* 4*-blocked set has at most* 36 *points.*

Proof. Let P be a 4-blocked set. Suppose that $|P| \ge 37$. Let S be the largest color class. Thus, $|S| \ge 10$. By Lemma 4, S is in general position. By a theorem of Harborth [4], some 5-point subset $K \subseteq S$ is the vertex-set of an empty convex pentagon $\mathrm{conv}(K)$. Let $T := P \cap (\mathrm{conv}(K) - K)$. Since $\mathrm{conv}(K)$ is empty with respect to S, each point in T is not in S. Thus, T is 3-blocked. K needs at least eight blockers (five blockers for the edges on the boundary of $\mathrm{conv}(K)$, and three blockers for the chords of $\mathrm{conv}(K)$). Thus, $|T| \ge 8$. But every 3-blocked set has at most six points, which is a contradiction. Hence, $|P| \le 36$. □

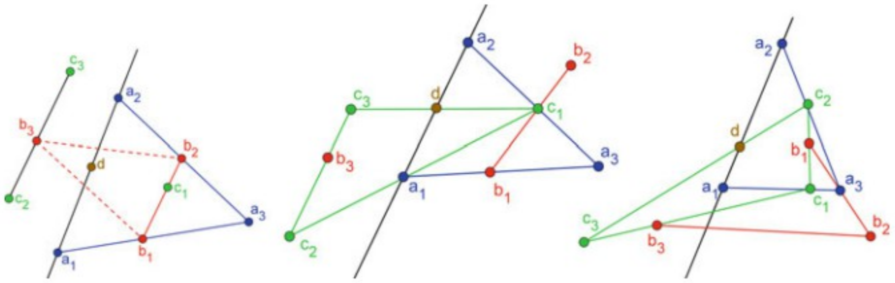

Fig. 3 Cases 1.1, 1.2, and 2 in Lemma 18

We now set out to characterize all representable 4-sets. We need a few technical lemmas.

Lemma 17. *Let A be a set of three monochromatic points in a 4-blocked set P. Then* $P \cap \operatorname{conv}(A)$ *contains a point from each color class.*

Proof. $P \cap \operatorname{conv}(A)$ contains at least three points in A. If $P \cap \operatorname{conv}(A)$ contains no point from one of the three other color classes, then $P \cap \operatorname{conv}(A)$ is a (≤ 3)-blocked set with three points in one color class (A), contradicting Theorem 2. □

Let $a * b * c$ mean that b blocks a and c, and let $a * b * c * d$ mean that $a * b * c$ and $b * c * d$. This allows us to record the order in which points occur on a line. Let $\overleftrightarrow{ab}$ be the line through two points a and b. For three noncollinear points a, b, c, let Δabc be the open triangle with vertices a, b, c.

Lemma 18. $\{3,3,3,1\}$ *is not representable.*

Proof. Suppose that P is a $\{3,3,3,1\}$-blocked set of points, with color classes $A = \{a_1, a_2, a_3\}$, $B = \{b_1, b_2, b_3\}$, $C = \{c_1, c_2, c_3\}$, and $D = \{d\}$.

If d does not block some monochromatic pair, then $P \setminus D$ is a 3-blocked set of nine points, contradicting Theorem 2. Therefore, d blocks some monochromatic pair, which we may call a_1, a_2. Now $\overleftrightarrow{a_1a_2}$ divides the remaining seven points of P into two sets, P_1 and P_2, where without loss of generality, $4 \leq |P_1| \leq 7$ and $0 \leq |P_2| \leq 3$. If $|P_1| \geq 6$, then $P_1 \cup \{a_1\}$ is a 3-blocked set of more than six points, contradicting Theorem 2. Thus, $|P_1| = 4$ and $|P_2| = 3$, or $|P_1| = 5$ and $|P_2| = 2$. Consider the following cases, as illustrated in Fig. 3.

Case 1. $|P_1| = 4$ and $|P_2| = 3$.

By Theorem 2, P_1 is a $\{1,1,2\}$-blocked set. Thus, without loss of generality, $P_1 = \{a_3, b_1, b_2, c_1\}$ and $P_2 = \{b_3, c_2, c_3\}$, where $c_2 * b_3 * c_3$. Some point in P_1 blocks b_1 and b_2. If $b_1 * a_3 * b_2$, then neither b_1 nor b_2 can block $\overline{a_3a_1}$ or $\overline{a_3a_2}$; thus, c_1 blocks a_3 from both a_1 and a_2, a contradiction. Therefore, $b_1 * c_1 * b_2$. Since two of these three points must block $\overline{a_3a_1}$ and $\overline{a_3a_2}$, we may assume that $a_1 * b_1 * a_3$, and either $a_2 * b_2 * a_3$ or $a_2 * c_1 * a_3$.

Case 1.1. $a_2 * b_2 * a_3$: Since $c_2 * b_3 * c_3$, the only possible blockers for $\overline{b_1b_3}$ are on $\overleftrightarrow{a_1a_2}$. We cannot have $b_3 * a_1 * b_1$, for then $b_3 * a_1 * b_1 * a_3$. We cannot have $b_3 * a_2 * b_1$, for then $\overleftrightarrow{b_1b_3}$ separates b_2 from d, implying $d \notin \operatorname{conv}(B)$, contradicting Lemma 17. Therefore, $b_1 * d * b_3$. Similarly, $b_2 * d * b_3$, a contradiction.

Case 1.2. $a_2 * c_1 * a_3$: Thus, a_2 does not block $\overline{c_1c_2}$ or $\overline{c_1c_3}$. Therefore, a_1 and d block $\overline{c_1c_2}$ and $\overline{c_1c_3}$. Without loss of generality, $c_2 * a_1 * c_1$ and $c_3 * d * c_1$. Since $c_2 * b_3 * c_3$, the blocker for $\overline{b_1b_3}$ is on the same side of $\overleftrightarrow{c_1c_3}$ as a_1, and on the same side of $\overleftrightarrow{b_1b_2}$ as a_1. The only such points are a_1, c_2, and b_3. Thus, a_1 or c_2 blocks $\overline{b_1b_3}$. Now, c_2 does not block $\overline{b_1b_3}$, as otherwise $b_1 * c_2 * b_3 * c_3$. Similarly, a_1 does not block $\overline{b_1b_3}$, as otherwise $b_3 * a_1 * b_1 * a_3$. This contradiction concludes this case.

Case 2. $|P_1| = 5$ and $|P_2| = 2$.

We have $a_3 \in P_1$, as otherwise P_1 is 2-blocked, contradicting Theorem 2. Without loss of generality, $P_1 = \{a_3, b_1, b_2, c_1, c_2\}$ and $P_2 = \{b_3, c_3\}$. Note that a_3 must block at least one of $\overline{b_1b_2}$ and $\overline{c_1c_2}$, because $P_1 \setminus \{a_3\}$ is too large to be 2-blocked; however, it cannot block both, for then there would be no valid blockers left for a_3. Thus, without loss of generality, $b_1 * a_3 * b_2$ and $c_1 * b_1 * c_2$. Since $a_3 \in \overline{b_1b_2}$, neither b_1 nor b_2 blocks $\overline{a_3a_1}$ or $\overline{a_3a_2}$. Thus, without loss of generality, $a_1 * c_1 * a_3$ and $a_2 * c_2 * a_3$.

By Lemma 17, $P \cap \operatorname{conv}\{c_1, c_2, c_3\}$ contains some member of A, say a_1. We cannot have $a_1 \in \overline{c_1c_3}$, for then $c_3 * a_1 * c_1 * a_3$. Therefore, a_1 is on the same side of $\overleftrightarrow{c_1c_3}$ as c_2; consequently, d and a_2 are, as well. It follows that $c_1 * b_3 * c_3$, and so b_3 sees b_2. □

Lemma 19. *Let P be a 4-blocked set. Suppose that some color class S of P contains a subset K, such that $|K| = 4$ and K is the vertex-set of a convex quadrilateral* $\operatorname{conv}(K)$ *that is empty with respect to S. Then P is $\{4,2,2,1\}$-blocked.*

Proof. Let $T := P \cap (\operatorname{conv}(K) - K)$. Since $\operatorname{conv}(K)$ is empty with respect to S, each point in T is not in S. Thus, T is 3-blocked. K needs at least five blockers (four blockers for the edges on the boundary of $\operatorname{conv}(K)$, and at least one blocker for the chords of $\operatorname{conv}(K)$). The only representable 3-sets with at least five points are $\{2,2,1\}$ and $\{2,2,2\}$. At least four points in T are on the boundary of $\operatorname{conv}(T)$. Every $\{2,2,2\}$-blocked set contains three points on the boundary of the convex hull. Thus, T is $\{2,2,1\}$-blocked. Hence, as illustrated in Fig. 4, one point c in T is at the intersection of the two chords of $\operatorname{conv}(K)$, and exactly one point in T is on each edge of the boundary of $\operatorname{conv}(K)$, such that the points on opposite edges of $\operatorname{conv}(K)$ are collinear with c.

We claim that no other point is in P. Suppose otherwise, and let p be a point in P outside $\operatorname{conv}(K)$ at a minimum distance from $\operatorname{conv}(K)$. Let x be a point in $\operatorname{conv}(K)$ receiving the same color as p. Thus, p and x are blocked by some point b in $\operatorname{conv}(K)$. Hence, b and x are collinear with no other point in $P \cap \operatorname{conv}(K)$, so $x \neq c$. Since p sees the nearest vertex in $\operatorname{conv}(K)$, we also have $x \notin K$. Thus, x is in the interior of one of the edges of the boundary of $\operatorname{conv}(K)$. Let y be the point in $\operatorname{conv}(K)$ receiving the same color as x. Thus, x and y are on opposite edges of the boundary of $\operatorname{conv}(K)$. Hence, p and y receive the same color, implying p and y are blocked. Since p is at a

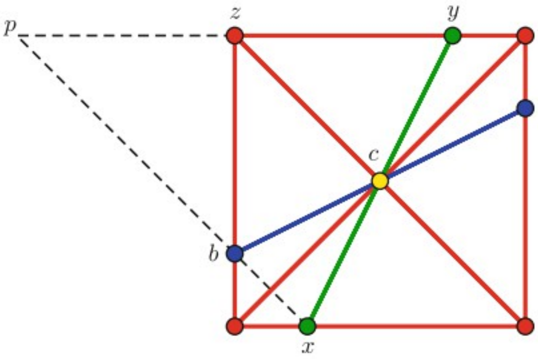

Fig. 4 A $\{4,2,2,1\}$-blocked point set

minimum distance from $\text{conv}(K)$, this blocker must be a point z of $\text{conv}(K)$. Having both $p*b*x$ and $p*z*y$ implies that p,z,y and one other point of $\text{conv}(K)$ are four collinear points, which contradicts Lemma 3. Hence, no other point is in P, and P is $\{4,2,2,1\}$-blocked. □

Lemma 19 has the following corollary (let $K := S$).

Corollary 20. *Let P be a 4-blocked set. Suppose that some color class S consists of exactly four points in convex position. Then P is $\{4,2,2,1\}$-blocked.*

The next lemma is a key step in our characterization of representable 4-sets.

Lemma 21. *Each color class in a 4-blocked point set has at most four points.*

Proof. Suppose that some 4-blocked point set P has a color class S with at least five points. Esther Klein [3] proved that every set of at least five points in general position in the plane contains an empty quadrilateral. By Lemma 4, S is in general position. Thus, S contains a subset K, such that $|K| = 4$ and K is the vertex set of a convex quadrilateral $\text{conv}(K)$ that is empty with respect to S. By Lemma 19, P is $\{4,2,2,1\}$-blocked, which is the desired contradiction. □

Lemma 22. *Let P be a 4-blocked point set with color classes A,B,C,D. Suppose that no color class consists of exactly four points in convex position (that is, Corollary 20 is not applicable). Furthermore, suppose that some color class A consists of exactly four points in nonconvex position. Then P is $\{4,2,2,2\}$-blocked, as in Fig. 5 for example.*

Proof. By Lemma 21, each color class has at most four points. By assumption, every 4-point color class is in nonconvex position. We may assume that A is minimal in the sense that no other 4-point color class is within $\text{conv}(A)$.

Let $Q := P \cap \text{conv}(A)$. Thus, Q is 4-blocked, and one color class is A. By the minimality of A, each other color class in Q has at most three points. We first prove that Q is $\{4,2,2,2\}$-blocked, and then show that this implies that $P = Q$.

Fig. 5 A $\{4,2,2,2\}$-blocked point set

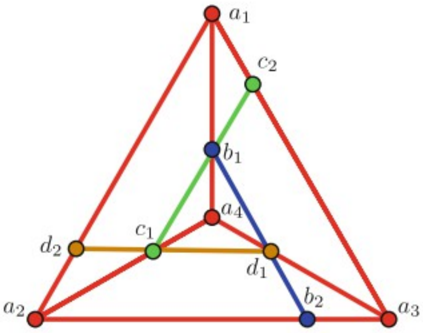

Let $A = \{a_1, a_2, a_3, a_4\}$, where a_4 is the interior point of $\operatorname{conv}(A)$. Note that the edges with endpoints in A divide $\operatorname{conv}(A)$ into three triangles with disjoint interiors. By Lemma 17, each color class of Q is represented in each of these triangles; this requires at least two points of each color (one of which could sit on the edge shared by two triangles).

We name a point with reference to its color class, such as $b_1 \in B$; or we name a point with reference to its position.

Let $Q' := Q \setminus A$. Let x_{ij} be the unique member of Q' that blocks $\overline{a_i a_j}$, for $1 \leq i < j \leq 4$. This accounts for exactly six points of Q'. For each case below we prove that Q' contains four mutually visible points (with respect to Q), which is a contradiction since Q' is 3-colored.

Q is not $\{4,3,3,3\}$-blocked, as otherwise we could delete the three outer members of A to represent $\{3,3,3,1\}$, which contradicts Lemma 18. Thus, Q is $\{4,3,2,2\}$-blocked, $\{4,3,3,2\}$-blocked, or $\{4,2,2,2\}$-blocked.

First, suppose that Q is $\{4,3,2,2\}$-blocked. Then Q' consists of six points x_{ij} and one additional point, y. The following three cases arise, as illustrated in Fig. 6.

Case 1. y blocks two points x_{i4} and x_{j4}: Without loss of generality, $x_{24} * y * x_{34}$. Now y is on one side or the other of $\overleftrightarrow{x_{14}a_4}$, and therefore y sees x_{12} or x_{13}. Without loss of generality, y sees x_{12}. Thus, x_{12}, x_{14}, x_{24}, and y are mutually visible points in Q'.

Case 2. y is collinear with, but does not block, two points x_{i4} and x_{j4}. Without loss of generality, $y * x_{24} * x_{34}$. Now x_{12} is on one side or the other of $\overleftrightarrow{yx_{24}}$; thus, x_{12} sees some $p \in \{x_{14}, x_{23}\}$, which is mutually visible with x_{12}, x_{24} and y; thus, we have four mutually visible points in Q'.

Case 3. y is not collinear with two points x_{i4} and x_{j4}: Then y sees all such points; thus, x_{14}, x_{24}, x_{34}, and y are mutually visible in Q'.

Now suppose that Q is $\{4,3,3,2\}$-blocked. Then Q' consists of the six points x_{ij} and two additional points, y_1 and y_2. Let $\delta = \{x_{14}, x_{24}, x_{34}\}$ and let δ_2 be the set of segments with both endpoints in δ. How many of the points y_i block members of δ_2? The following cases arise.

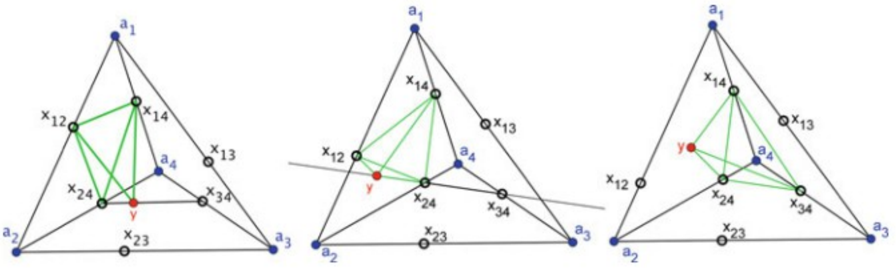

Fig. 6 Cases 1–3 when Q is $\{4,3,2,2\}$-blocked

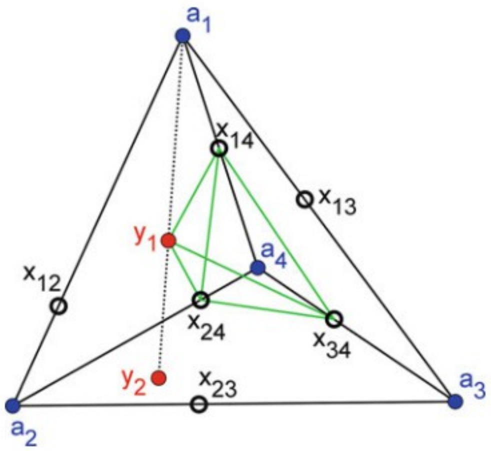

Fig. 7 Case 1 when Q is $\{4,3,3,2\}$-blocked

Case 1. Neither y_1 nor y_2 blocks a segment in δ_2 (see Fig. 7).

Since y_1 and y_2 cannot block each other, without loss of generality, y_2 does not block y_1. Thus, δ and y_1 give four mutually visible points in Q'.

Case 2. At least one of y_1 and y_2 blocks a segment in δ_2.

Without loss of generality, y_1 blocks a segment in δ_2 and $x_{24} * y_1 * x_{34}$. Neither a_1 nor a_4 belongs to $\overleftrightarrow{x_{14}y_1}$, because otherwise *both* do (since $a_1 * x_{14} * a_4$), yielding four collinear points. Since $y_1 \in \Delta a_2a_3a_4 \subset \Delta a_2a_3x_{14}$, we know that a_2 and a_3 are on opposite sides of $\overleftrightarrow{x_{14}y_1}$ by the Crossbar Theorem. Thus, without loss of generality, a_4 is on the same side of $\overleftrightarrow{x_{14}y_1}$ as x_{24}—call this side "west" and the other side "east."

We now show that $\{x_{13}, x_{14}, x_{34}, y_1\}$ is a set of four points in general position. If not, then three of these points are collinear. Observe that x_{13}, x_{14}, and x_{34} are noncollinear. Thus, y_1 is one of the three collinear points. Since $\overleftrightarrow{x_{34}y_1}$ already contains a third point, namely x_{24}, it contains neither x_{13} nor x_{14} by Lemma 3. Thus, our three collinear points are $\{x_{13}, x_{14}, y_1\}$. Since $x_{24} * y_1 * x_{34}$, the point x_{34} is east of $\overleftrightarrow{x_{14}y_1}$. Thus, a_1 and a_3 are also east of $\overleftrightarrow{x_{14}y_1}$, so x_{13} must be as well. Therefore, x_{13} is not collinear with x_{14} and y_1. Thus, $\{x_{13}, x_{14}, x_{34}, y_1\}$ is a set of four points in general position.

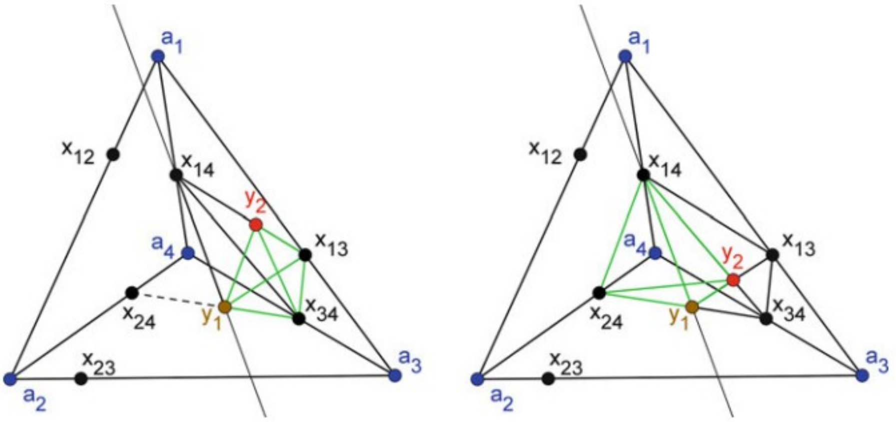

Fig. 8 Case 2a when Q is $\{4,3,3,2\}$-blocked

Since Q' does not contain four mutually visible points, there is some blocker in the midst of $\{x_{13}, x_{14}, x_{34}, y_1\}$. Because a_4 and x_{24} are west of $\overleftrightarrow{x_{14}y_1}$, the only possible blocker—which must be an interior point of $\mathrm{conv}(Q)$—is y_2, which must therefore belong to $\mathrm{conv}(x_{13}, x_{14}, x_{34}, y_1)$.

Case 2a. $\{x_{13}, x_{14}, x_{34}, y_1\}$ is in convex position (see Fig. 8).

If y_1 and x_{13} are on the same side of $\overleftrightarrow{x_{14}x_{34}}$, then y_1 and x_{24} would be on opposite sides of $\overleftrightarrow{x_{14}x_{34}}$, implying y_1 would not block $\overline{x_{24}x_{34}}$, which is a contradiction. Thus, y_1 and x_{13} are on opposite sides of $\overleftrightarrow{x_{14}x_{34}}$. It follows that $\mathrm{conv}(x_{13}, x_{14}, x_{34}, y_1)$ has chords $\overline{x_{14}x_{34}}$ and $\overline{x_{13}y_1}$. If y_2 blocks just one edge between $\{x_{13}, x_{14}, x_{34}, y_1\}$ (as in Fig. 8, left), then we may substitute y_2 for one endpoint of that edge to build a set of four mutually visible points in Q'. Thus, y_2 must block two edges between $\{x_{13}, x_{14}, x_{34}, y_1\}$ (as in Fig. 8, right). Hence, $x_{14} * y_2 * x_{34}$ and $y_1 * y_2 * x_{13}$. Now (1) a_4 is blocked by both x_{14} and x_{24}, and therefore blocks neither of them; (2) a_4 cannot block $\overline{y_1y_2}$, which is on the other side of $\overleftrightarrow{x_{14}y_1}$; (3) x_{34} is blocked by both y_1 and y_2, and therefore cannot block either of them; (4) x_{34} cannot block $\overline{x_{14}x_{24}}$, which is on the other side of $\overleftrightarrow{x_{14}y_1}$. Thus, $\{x_{14}, x_{24}, y_1, y_2\}$ is a set of four mutually visible points in Q'.

Case 2b. $\{x_{13}, x_{14}, x_{34}, y_1\}$ is not in convex position (see Fig. 9).

One of these four points lies inside the triangle formed by the other three. This interior point cannot be x_{13}, since it is on the boundary of $\mathrm{conv}(A)$. Nor can it be x_{14} or y_1, as we previously showed that x_{13} and x_{34} are on the same side (east) of $\overleftrightarrow{x_{14}y_1}$. Thus, x_{34} is in $\Delta x_{13}x_{14}y_1$. Hence, x_{13} and x_{14} are on opposite sides of $\overleftrightarrow{x_{34}y_1}$. That is, x_{13} and x_{23} are on the same side of $\overleftrightarrow{x_{34}y_1}$. This implies that x_{24} and y_1 are boundary points of $\mathrm{conv}(x_{23}, x_{24}, y_1, x_{13})$. Moreover, x_{13} and x_{23} must also be boundary points of $\mathrm{conv}(x_{23}, x_{24}, y_1, x_{13})$, as they are boundary points of $\mathrm{conv}(A)$.

Now $\{x_{23}, x_{24}, y_1, x_{13}\}$ is a set of four points in convex position. Note that $\mathrm{conv}(x_{13}, x_{14}, x_{34}, y_1)$ meets $\mathrm{conv}(x_{23}, x_{24}, y_1, x_{13})$ along only a single edge, $\overline{y_1x_{13}}$.

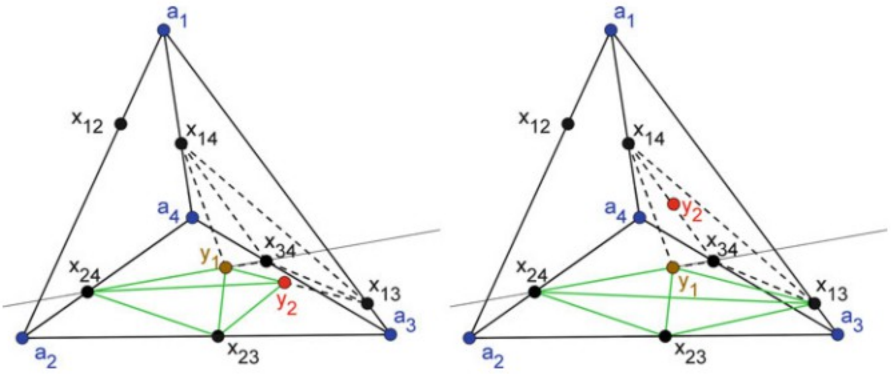

Fig. 9 Case 2b when Q is $\{4,3,3,2\}$-blocked

If y_2 blocks $\overline{y_1x_{13}}$, then $\{x_{23},x_{24},y_1,y_2\}$ is a set of four mutually visible points in Q' (as in Fig. 9, left). If y_2 does not block $\overline{y_1x_{13}}$, then $\{x_{23},x_{24},y_1,x_{13}\}$ is a set of four mutually visible points in Q' (as in Fig. 9, right).

The only remaining case is that Q is $\{4,2,2,2\}$-blocked. This is possible, as illustrated in Fig. 5. We now show that this point set is essentially the only $\{4,2,2,2\}$-blocked set, up to betweenness-preserving deformations. There are exactly enough points in color classes B, C, and D to block all edges between points in A. As each of the three A-triangles with point a_4 must contain a representative from each of the other three color classes, it follows that of the six A-edges, B blocks one interior edge and the opposite boundary edge, and likewise for C and D. Without loss of generality $b_1 = x_{14}$, $b_2 = x_{23}$, $c_1 = x_{24}$, $c_2 = x_{13}$, $d_1 = x_{34}$, and $d_2 = x_{12}$. Since b_1 blocks a_4, a_4 cannot block b_1. Since $\overline{b_1b_2}$ can be blocked only by an interior point of conv(A), it follows that either c_1 or d_1 blocks $\overline{b_1b_2}$. As these cases are symmetric, we may choose $c_1 \in \overline{b_1b_2}$. Now b_1 cannot block c_1, so c_1 must be blocked by d_1, which must in turn be blocked by b_1.

Now we show that $P = Q$. (This basically says that the point set in Fig. 5 cannot be extended without introducing a new color.) Suppose to the contrary that $P \setminus \text{conv}(A) \neq \emptyset$. Let x be a point in $P \setminus \text{conv}(A)$ closest to conv(A). Thus, x sees a vertex of conv(A), and $x \notin A$. Without loss of generality, $x \in B$. Recall that b_2 is on the line $\overleftrightarrow{a_2a_3}$. Thus, x is in the same half-plane determined by $\overleftrightarrow{a_2a_3}$ as the rest of Q, as otherwise b_2 would see $x \in B$. Which point blocks $\overline{b_1x}$? Not a_2 or a_3, for this would put x in the wrong half-plane. Not a_1, a_4, c_1, or c_2, since b_1 blocks each of these. Not d_1, which is on $\overline{b_1b_2}$. Therefore, $\overline{b_1x}$ can be blocked only by d_2. Since $d_1 \in \overline{b_1b_2}$ and $d_2 \in \overline{b_1x}$, it follows that b_2 and x (and any blocker between them) are on the same side of $\overleftrightarrow{d_1d_2}$. But the only other points of Q in that open half-plane are a_2 and a_3, which cannot block b_2. Thus, x sees b_2, which is a contradiction. Thus, $P = Q$ and P is $\{4,2,2,2\}$-blocked. □

We now prove the main theorem of this section.

Theorem 23. *A 4-set $\{a,b,c,d\}$ is representable if and only if*

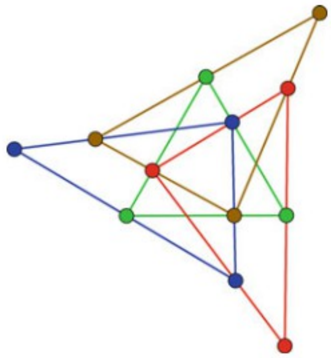

Fig. 10 Another $\{3,3,3,3\}$-blocked point set

- *$\{a,b,c,d\} = \{4,2,2,1\}$, or*
- *$\{a,b,c,d\} = \{4,2,2,2\}$, or*
- *All of $a,b,c,d \leq 3$ except for $\{3,3,3,1\}$.*

Proof. Figures 4 and 5 respectively show $\{4,2,2,1\}$-blocked and $\{4,2,2,2\}$-blocked point sets. When $a,b,c,d \leq 3$, the required constructions are described in Proposition 12. Now we prove that these are the only representable 4-sets. Let P be a 4-blocked point set. By Lemma 21, each color class has at most four points. Let S be the largest color class. If $|S| \leq 3$, then we are done by Proposition 12. Now assume that $|S| = 4$. If S is in nonconvex position, then P is $\{4,2,2,2\}$-blocked by Lemma 22. If S is in convex position, then P is $\{4,2,2,1\}$-blocked by Corollary 20. □

Corollary 24. *Every 4-blocked set has at most 12 points, and there is a 4-blocked set with 12 points.*

Note that in addition to the $\{3,3,3,3\}$-blocked set shown in Fig. 1, there is a different $\{3,3,3,3\}$-blocked point set, as illustrated in Fig. 10.

5 Midpoint-Blocked Point Sets

A k-blocked point set P is *k-midpoint-blocked* if for each monochromatic pair of distinct points $v,w \in P$ the midpoint of $\overline{vw}$ is in P. Of course, the midpoint of $\overline{vw}$ blocks v and w. A point set P is *$\{n_1,\dots,n_k\}$-midpoint-blocked* if it is $\{n_1,\dots,n_k\}$-blocked and k-midpoint-blocked. For example, the point set in Fig. 1 is $\{3,3,3,3\}$-midpoint-blocked.

Another interesting example is the projection[1] of $[3]^d$. With $d = 1$ this point set is $\{2,1\}$-blocked, with $d = 2$ it is $\{4,2,2,1\}$-blocked, and with $d = 3$ it is

---

[1] If G is the visibility graph of some point set $P \subseteq \mathbb{R}^d$, then G is the visibility graph of some projection of P to $\mathbb{R}^2$ (since a random projection of P to $\mathbb{R}^2$ is occlusion-free with probability 1).

$\{8,4,4,4,2,2,2,1\}$-blocked. In general, each set of points with exactly the same set of coordinates equal to 2 is a color class. Each color class of points with exactly i coordinates equal to 2 has size 2^{d-i}, and there are $\binom{d}{i}$ such color classes. Hence, $[3]^d$ is $\{\binom{d}{i} \times 2^{d-i} : i \in [0,d]\}$-midpoint-blocked and 2^d-midpoint-blocked.

We now prove Conjecture 1 when restricted to k-midpoint-blocked point sets. (Finally, we have weakened Conjecture 5 to something provable!)

A. Hernández-Barrera, F. Hurtado, J. Urrutia, and C. Zamora (On the midpoints of a plane point set, 2001, unpublished manuscript) introduced the following definition. Let $m(n)$ be the minimum number of midpoints determined by some set of n points in general position in the plane. Since the midpoint of $\overline{vw}$ blocks v and w, we have $b(n) \leq m(n)$. A. Hernández-Barrera, F. Hurtado, J. Urrutia, and C. Zamora (On the midpoints of a plane point set, 2001, unpublished manuscript) constructed a set of n points in general position in the plane that determine at most $cn^{\log 3}$ midpoints for some constant c. (All logarithms here are binary.) Thus, $b(n) \leq m(n) \leq cn^{\log 3} = cn^{1.585\cdots}$. This upper bound was improved independently by Pach [7] and Stanchescu [11] (and later by Matoušek [6]) to

$$b(n) \leq m(n) \leq nc^{\sqrt{\log n}}.$$

A. Hernández-Barrera, F. Hurtado, J. Urrutia, and C. Zamora (On the midpoints of a plane point set, 2001, unpublished manuscript) conjectured that $m(n)$ is superlinear (that is, $\frac{m(n)}{n} \to \infty$ as $n \to \infty$), which was verified by Pach [7]. More precise bounds were obtained by Stanchescu [11] and Sanders [10]. Applying the latest results on Freiman's Theorem, Pór and Wood [9] observed that for all $\varepsilon > 0$ there is an integer $N(\varepsilon)$ such that $m(n) \geq n(\log n)^{4/11-\varepsilon}$ for all $n \geq N(\varepsilon)$.

Theorem 25. *For each k, there is an integer n such that every k-midpoint-blocked set has at most n points. More precisely, for all $\varepsilon > 0$ and for all k, every k-midpoint-blocked set has at most $k\max\{N(\varepsilon), 2^{(k-1)^{11/(4-11\varepsilon)}}\}$ points.*

Proof. Let P be a k-midpoint-blocked set of n points. If $n \leq kN(\varepsilon)$, then we are done. Now assume that $\frac{n}{k} > N(\varepsilon)$. Let S be a set of exactly $s := \lceil \frac{n}{k} \rceil$ monochromatic points in P. Thus, S is in general position by Lemma 4. And for every pair of distinct points $v, w \in S$, the midpoint of $\overline{vw}$ is in $P - S$. Thus,

$$\tfrac{n}{k}(\log \tfrac{n}{k})^{4/11-\varepsilon} \leq m(s) \leq n - s \leq n(1 - \tfrac{1}{k}).$$

Hence, $(\log \frac{n}{k})^{4/11-\varepsilon} \leq k-1$, implying $n \leq k2^{(k-1)^{11/(4-11\varepsilon)}}$. The result follows. $\square$

We now construct k-midpoint-blocked point sets with a "large" number of points. The method is based on the following product of point sets P and Q. For each point $v \in P \cup Q$, let (x_v, y_v) be the coordinates of v. Let $P \times Q$ be the point set $\{(v,w) : v \in P, w \in Q\}$, where (v,w) is at (x_v, y_v, x_w, y_w) in 4-dimensional space. For brevity we do not distinguish between a point in $\mathbb{R}^4$ and its image in an occlusion-free projection of the visibility graph of $P \times Q$ into $\mathbb{R}^2$.

Lemma 26. *If P is a $\{n_1,\dots,n_k\}$-midpoint-blocked point set and Q is a $\{m_1,\dots,m_\ell\}$-midpoint-blocked point set, then $P \times Q$ is $\{n_i m_j : i \in [k], j \in [\ell]\}$-midpoint-blocked.*

Proof. Color each point (v,w) in $P \times Q$ by the pair $(\mathrm{col}(v),\mathrm{col}(w))$. Thus, there are $n_i m_j$ points for the (i,j)th color class. Consider distinct points (v,w) and (a,b) in $P \times Q$.

Suppose that $\mathrm{col}(v,w) = \mathrm{col}(a,b)$. Thus, $\mathrm{col}(v) = \mathrm{col}(a)$ and $\mathrm{col}(w) = \mathrm{col}(b)$. Since P and Q are midpoint-blocked, $\frac{1}{2}(v+a) \in P$ and $\frac{1}{2}(w+b) \in Q$. Thus, $(\frac{1}{2}(v+a), \frac{1}{2}(w+b))$, which is positioned at $(\frac{1}{2}(x_v+x_a), \frac{1}{2}(y_v+y_a), \frac{1}{2}(x_w+x_b), \frac{1}{2}(y_w+y_b))$, is in $P \times Q$. This point is the midpoint of $\overline{(v,w)(a,b)}$. Thus, (v,w) and (a,b) are blocked by their midpoint in $P \times Q$.

Conversely, suppose that some point $(r,s) \in P \times Q$ blocks (v,w) and (a,b). Thus, $x_r = \alpha x_v + (1-\alpha)x_a$ for some $\alpha \in (0,1)$, $y_r = \beta y_v + (1-\beta)y_a$ for some $\beta \in (0,1)$, $x_s = \delta x_w + (1-\delta)x_b$ for some $\delta \in (0,1)$, and $y_s = \gamma y_w + (1-\gamma)y_b$ for some $\gamma \in (0,1)$. Hence, r blocks v and a in P, and s blocks w and b in Q. Thus, $\mathrm{col}(v) = \mathrm{col}(a) \neq \mathrm{col}(r)$ in P, and $\mathrm{col}(w) = \mathrm{col}(b) \neq \mathrm{col}(s)$ in Q, implying $(\mathrm{col}(v),\mathrm{col}(w)) = (\mathrm{col}(a),\mathrm{col}(b))$.

We have shown that two points in $P \times Q$ are blocked if and only if they have the same color. Thus, $P \times Q$ is blocked. Since every blocker is a midpoint, $P \times Q$ is midpoint-blocked. □

Say P is a k-midpoint blocked set of n points. By Lemma 26, the i-fold product $P^i := P \times \cdots \times P$ is a k^i-blocked set of $n^i = (k^i)^{\log_k n}$ points. Taking P to be the $\{3,3,3,3\}$-midpoint-blocked point set in Fig. 1, we obtain the following result:

Theorem 27. *For all k a power of 4, there is a k-blocked set of $k^{\log_4 12} = k^{1.79\dots}$ points.*

This result describes the largest known construction of k-blocked or k-midpoint-blocked point sets. To promote further research, we make the following strong conjectures:

Conjecture 28. Every k-blocked point set has $O(k^2)$ points.

Conjecture 29. In every k-blocked point set there are at most k points in each color class.

Conjecture 29 would be tight for the projection of $[3]^d$ with $k = 2^d$. Of course, Conjecture 29 implies Conjecture 28, which implies Conjecture 1.

Acknowledgements This research was initiated at The 24th Bellairs Winter Workshop on Computational Geometry, held in February 2009 at the Bellairs Research Institute of McGill University in Barbados. The authors are grateful to Godfried Toussaint and Erik Demaine for organizing the workshop, and to the other participants for providing a stimulating working environment. Thanks to the reviewer for many helpful suggestions. David R. Wood was supported by a QEII Research Fellowship from the Australian Research Council.

References

1. Z. Abel, B. Ballinger, P. Bose, S. Collette, V. Dujmović, F. Hurtado, S.D. Kominers, S. Langerman, A. Pór, D.R. Wood, Every large point set contains many collinear points or an empty pentagon. Graphs Comb. **27**(1), 47–60 (2011). http://dx.doi.org/10.1007/s00373-010-0957-2. doi:`10.1007/s00373-010-0957-2`
2. A. Dumitrescu, J. Pach, G. Tóth, A note on blocking visibility between points. Geombinatorics **19**(1), 67–73 (2009). http://www.cs.uwm.edu/faculty/ad/blocking.pdf
3. P. Erdős, G. Szekeres, A combinatorial problem in geometry. Compos. Math. **2**, 464–470 (1935). http://www.numdam.org/item?id=CM_1935__2__463_0
4. H. Harborth, Konvexe Fünfecke in ebenen Punktmengen. Elem. Math. **33**(5), 116–118 (1978)
5. J. Kára, A. Pór, D.R. Wood, On the chromatic number of the visibility graph of a set of points in the plane. Discr. Comput. Geom. **34**(3), 497–506 (2005). http://dx.doi.org/10.1007/s00454-005-1177-z. doi:`10.1007/s00454-005-1177-z`
6. J. Matoušek, Blocking visibility for points in general position. Discr. Comput. Geom. **42**(2), 219–223 (2009). http://dx.doi.org/10.1007/s00454-009-9185-z. doi:`10.1007/s00454-009-9185-z`
7. J. Pach, Midpoints of segments induced by a point set. Geombinatorics **13**(2), 98–105 (2003). http://www.math.nyu.edu/~pach/publications/midpoint.ps
8. R. Pinchasi, On some unrelated problems about planar arrangements of lines, in *Workshop II: Combinatorial Geometry. Combinatorics: Methods and Applications in Mathematics and Computer Science*, Institute for Pure and Applied Mathematics, UCLA, 2009. http://11011110.livejournal.com/184816.html.
9. A. Pór, D.R. Wood, On visibility and blockers. J. Comput. Geom. **1**(1), 29–40 (2010) http://www.jocg.org/index.php/jocg/article/view/24
10. T. Sanders, Three-term arithmetic progressions and sumsets. Proc. Edinb. Math. Soc. (2) **52**(1), 211–233 (2009). http://dx.doi.org/10.1017/S0013091506001398. doi:`10.1017/S0013091506001398`.
11. Y.V. Stanchescu, Planar sets containing no three collinear points and non-averaging sets of integers. Discr. Math. **256**(1–2), 387–395 (2002) http://dx.doi.org/10.1016/S0012-365X(01)00441-1. doi:`10.1016/S0012-365X(01)00441-1`.
12. G. Tóth, P. Valtr, The Erdős–Szekeres theorem: upper bounds and related results, in *Combinatorial and Computational Geometry*. Math. Sci. Res. Inst. Publ., vol. 52 (Cambridge University Press, Cambridge, 2005), pp. 557–568. http://library.msri.org/books/Book52/files/30toth.pdf

C4 5 @0

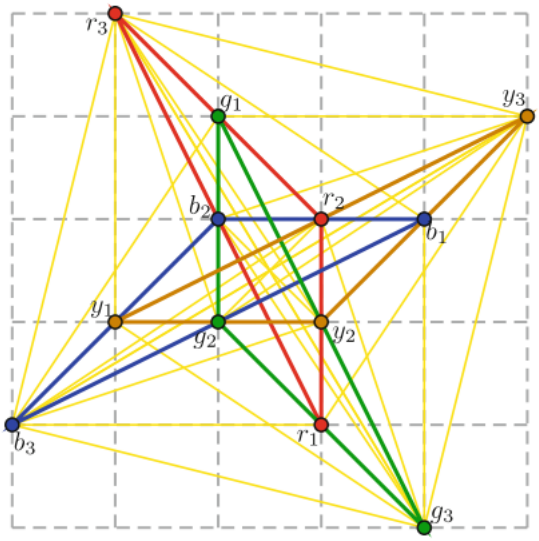
r_3
g_1
y_3
b_2
r_2
b_1
y_1
g_2
y_2
b_3
r_1
g_3

D 4 # µ , , , ¶ : (+
*

C4 5>>0

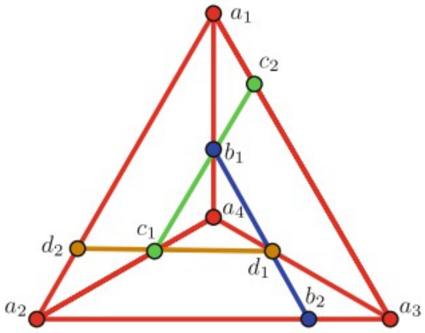
a_1
c_2
b_1
a_4
c_1
d_2
d_1
a_2
b_2
a_3

D 4 > # µ , , , ¶ : (+ *

3: E J) + * , :)9)) + + · 4 7 4 +),
*)))4 + * €• ,))))+) * *

: + :
+ +

3) [J * + 9C +, :)) C)9) 8 •

C4 5>@0

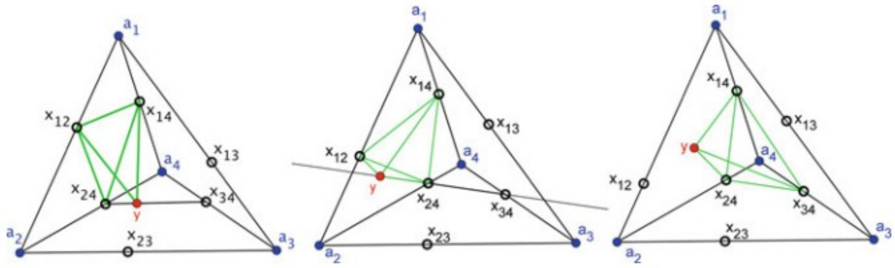
a_1
a_2
a_3
a_4
x_{12}
x_{13}
x_{14}
x_{23}
x_{24}
x_{34}
y

D 4 @) 2 µ , , , ¶ : (+

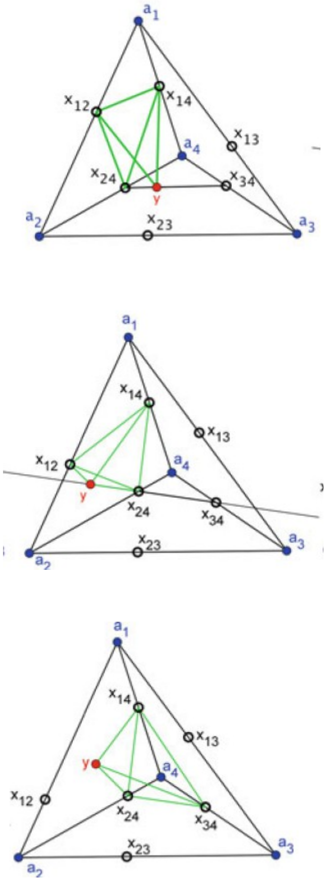
a_1
a_2
a_3
a_4
x_{12}
x_{13}
x_{14}
x_{23}
x_{24}
x_{34}
y

3: %) , FC C) 8 • *)) +

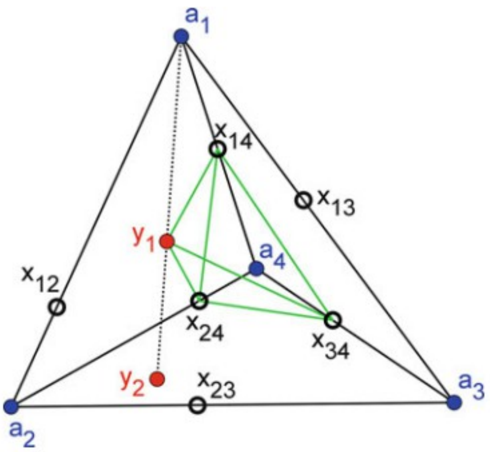
a_1
a_2
a_3
a_4
x_{12}
x_{13}
x_{14}
x_{23}
x_{24}
x_{34}
y_1
y_2

D 4) 2 µ , , , ¶ : (+

C4   5> 0

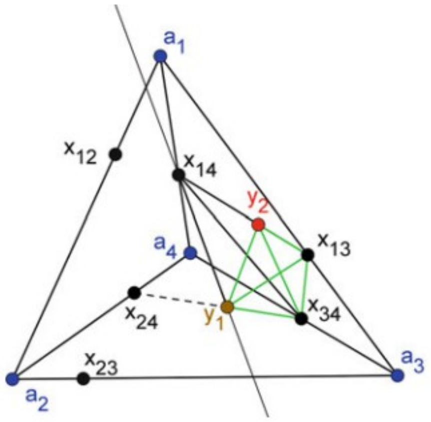
a_1
x_{12}
x_{14}
y_2
x_{13}
a_4
x_{24}
y_1
x_{34}
a_3
x_{23}
a_2

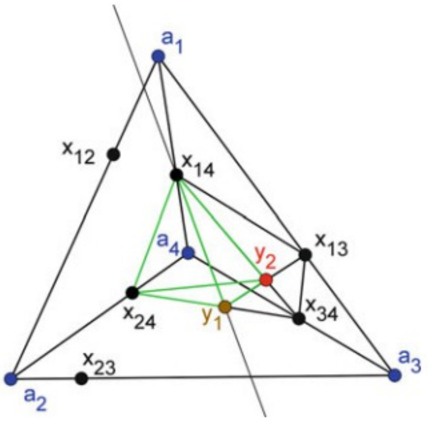
a_1
x_{12}
x_{14}
x_{13}
a_4
y_2
x_{24}
y_1
x_{34}
a_3
x_{23}
a_2

D 4 B)   )   2   µ   ,   ,   ,   ¶ :   (+

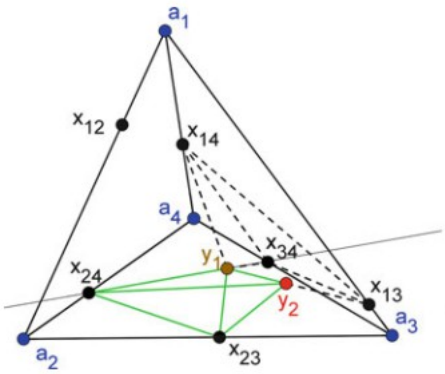
a_1
x_{12}
x_{14}
a_4
y_1
x_{34}
x_{24}
y_2
x_{13}
a_3
a_2
x_{23}

C4,   5>B0

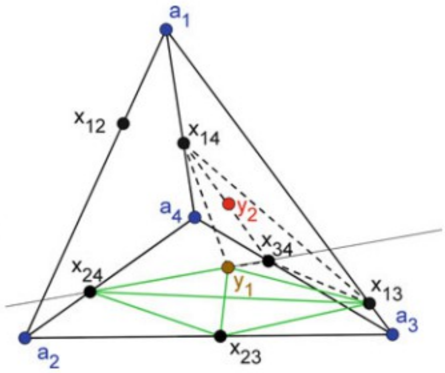
a_1
x_{12}
x_{14}
a_4
y_2
x_{34}
x_{24}
y_1
x_{13}
a_3
a_2
x_{23}

D 4   )   :   2   µ ,   ,   , ¶ : (+

C4 >5> 0

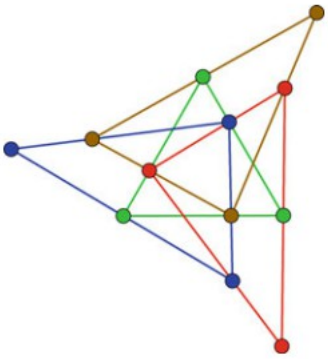

D 4   #   µ ,   ,   , ¶ : (+ *

C4   5 0   J) 4) + 7   /

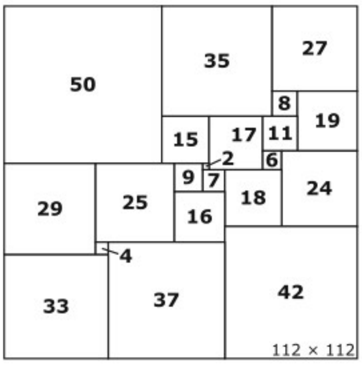
50
35
27
8
19
15
17
11
2
6
9
7
18
24
29
25
16
4
33
37
42
112 × 112

D 4 $) /   J *   J) 4   )   :   J)   -)
+   & +   B :/ # H G " F&   F

3:   E   *)) & ¥)9   )   )4   +   ) C) J) + +   *)   +   )   &)
:   *)   5   4) 0

0   !#12!   !   !3 4) ! B   H 4 3&)   , 1 :   ))+
' / + 7   .)& /6 $) .   "   )   #**)   ')   G +
*   4   5   0 *+

$) .   "   )
$ *   78
78) "   ) .)& / $ ')   G   +   >
')   ' (@
7   & ')   G   + 7J)
')   G   + %   )   /) +   .   ) 1) & / A
."   )   ')

>) / ." + ') G + @
@) 4 +) " :
#: 78)
1

-) ===

C4 5 0 $ * 78
/ 8 .1 &) / 4 & &
+) 7 J) 5 0 $ b ¨)) / :)
) & 4/) +)**)))/) /
+ &) : 4 4 :))** *) ,)) ,)) /
) , .1) : / + $ * /))
*) 8 .1 + +, :)) &) * :
) :) +) /
$ *)/ * +) 8 .1 :/ + 4) + Z)
; : 7))

$$S = \int \left(\frac{R}{2\,k^2} + \mathcal{L}_M \right) \sqrt{-g}\, d^4x + \alpha\,(correction)$$

¸) *)))) + 4)&)) + : +
:/ .1 $ 8* * + 4 4 7J 5 0
+ : , ,) F Z +) + * /) / &) + 8* #
)4 , .1 :) + :/ 4 ¸ \ , F) &) ,) + & / 4
( ) - + +,)/ 4) +)) J)
* + 4 / E* +) + 4))) &))
¹ \ , :)

$$R_{\mu\nu} - \frac{1}{2}\, R\, g_{\mu\nu} = k^2\, T_{\mu\nu} + \alpha\,(new\ term)_{\mu\nu}$$

7J 5 0)/ : +) +) *)) & 4/
$ 8*)/) + 4
+ ,)))) , &) + Z + ,) 4
4)&)) + : + :/ 7 N 4) / $ .1
) &) , ¸ \ )+) +) + 7 J) 5 0
$): &) 4) / .1* + ^
) /, .1 ^) &) * : ; & , /
7J 5 0 & +,)/) * / + / -) , & /
)) 4) 7J 5 0 + * +) / .1 *)
:/ * / 4 ¸ \ $ 4 * :) /
/ 5 0,) + +)) * $) / * ,
+))) + *) / / * º + $ 4) *) /
/ *)) :)

$$\begin{aligned} ds^2 &= g_{\mu\nu}\, dx^\mu\, dx^\nu \\ &= e^{\nu(r)}\, dt^2 - e^{\lambda(r)}\, dr^2 - r^2 \left(d\theta^2 + \sin^2\theta\, d\phi^2\right) \end{aligned}$$

») + ¼))))+ /

Chapter 1
The Minimal Geometric Deformation

General Relativity (GR) in its century of existence represents, without a doubt, one of the most important achievements of human knowledge. The predictions made by this theory, like the perihelion shift of Mercury, light deflection and gravitational lensing, the gravitational redshift and time delay, black holes and gravitational waves—just to mention some of the most remarkable ones, have given it the honours which deserves as one of the fundamental theories of Physics (for a recent review on experimental tests of GR see [1] and references therein). On the other hand, the great technological advances during the past decades have provided us with increasingly powerful instruments which allow for accurate measurements of the evolution of compact systems, which represent the best candidate laboratories for the study of gravity in the strong regime (see for instance [2]). This progress not only serves to test the theory like never before, but also appears to leave GR as the most reliable gravitational theory to be used in the analysis of phenomena occurring in the strong field regime. Likewise, the ability of observing increasingly distant objects deep in the universe, thus with a larger and larger cosmological redshift, leads to the conclusion that we can obtain an excellent modelling of these phenomena within GR [3, 4]. Additionally, due to the results obtained by the PLANCK collaboration [5], which improved greatly on the previous findings by WMAP [6], the cosmological models based on GR are enjoying a well-deserved and well-earned recognition. Furthermore, the recent direct observation of black holes shadows by the EHT collaboration [7, 8], as well as the detection of gravitational waves by the LIGO and Virgo collaborations [9–11], allows us to ensure that GR is one of the most successful scientific theories in the history of science. Why would we want to find new gravitational theories beyond GR then?

Despite the above success, there are fundamental questions about the gravitational interaction which GR cannot answer satisfactorily. These can be broadly grouped in the following fundamental issues, which could likely be closely related:

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2020

J. Ovalle and R. Casadio, *Beyond Einstein Gravity*, SpringerBriefs in Physics,
https://doi.org/10.1007/978-3-030-39493-6_1

(1) The inability of GR to explain satisfactorily the phenomenology of dark matter [12, 13] and dark energy, without introducing some kind of unknown matter-energy to reconcile its predictions with the observed galactic rotation curves and accelerated expansion of the universe; (2) The general GR prediction of the existence of space-time singularities, most notably in the gravitational collapse of regular matter distributions, which are both mathematically and physically hard to accept; (3) The impossibility, so far, to reconcile GR with the Standard Model of particle physics, or equivalently, the difficulties met when one tries and quantize GR by the same schemes used successfully for the other fundamental interactions. All of these have strongly motivated searches for a gravitational theory beyond GR. If the new theory is a consistent quantum theory, this could lead to a modification of GR at low energy which accounts for the dark matter and dark energy. If the new theory is not a consistent quantum theory for gravity, this should also contain GR in a suitable limit, and somehow show greater tolerance to its quantum description.

Indeed, there are many alternative theories, as for instance, high curvature gravity theories, Galileon theories, $f(R)$ theories, scalar-tensor theories, massive gravity, new massive gravity, topologically massive gravity, Chern-Simons theories, higher spin gravity theories, Horava-Lifshitz gravity, extra-dimensional theories, etc. All of these proposals entail a modification of the original Einstein equations

$$G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2}\, R\, g_{\mu\nu} = k^2\, T_{\mu\nu}\ , \tag{1.1}$$

where Greek indices run from 0 to 3, $R_{\mu\nu}$ and R are, respectively, the Ricci tensor and its trace determined by the space-time metric $g_{\mu\nu}$, $T_{\mu\nu}$ is the energy-momentum tensor of the matter which sources gravity, and $k^2 = 8\,\pi\, G_{\rm N}/c^2$ with $G_{\rm N}$ the Newton constant (we shall also mostly set the speed of light $c = 1$ and use the metric signature $+ - - -$, unless otherwise specified).

1.1 The Simplest Extension

Any extension of GR will eventually give rise to new terms in the effective four-dimensional Einstein equations (1.1). These "corrections" can usually be cast in the form of an effective energy-momentum tensor and appear in such a way that they should vanish or be negligible in an appropriate limit, as for instance, at solar system scale, where GR has been successfully tested. This limit represents not only a critical point when a consistent extension to GR is studied, but also a non-trivial problem that must be treated carefully.

The simplest way to produce an extension to GR is by considering a modification of the Hilbert-Einstein action as

$$S = \int \left(\frac{R}{2\,k^2} + \mathcal{L}_M \right) \sqrt{-g}\, d^4x + \alpha\, (correction)\ , \tag{1.2}$$

where α is a free parameter associated with the new gravitational sector not described by GR. The explicit form corresponding to the generic correction shown in Eq. (1.2) should be, of course, a well justified and physically motivated expression. At this stage, the GR limit obtained by setting $\alpha = 0$, is just a trivial issue, and everything looks consistent so far. Indeed, we may go further and calculate the equations of motion corresponding to this new theory. Upon demanding as usual that the variation $\delta S = 0$, one obtains

$$R_{\mu\nu} - \frac{1}{2}\, R\, g_{\mu\nu} = k^2\, T_{\mu\nu} + \alpha\, (new\ term)_{\mu\nu}\ , \tag{1.3}$$

where the new term in Eq. (1.3) may be consolidated as part of an effective energy-momentum tensor. The explicit form of this tensor may contain some new degrees of freedom, as for instance scalar, vector and tensor fields, all of them coming from the new gravitational sector not described by Einstein's original theory. The GR limit is still a trivial issue, since $\alpha = 0$ leads to the standard Einstein equations (1.1).

The above arguments seem to tell us that the consistency with the GR predictions—more formally, the GR limit—is a trivial problem. However, when the system of Eq. (1.3) is solved, the result may show a completely different story. In fact, it is very common that a general solution of Eq. (1.3) does not reproduce any GR space-time metric by simply setting $\alpha = 0$. The origin of this problem is in the non-linearity of the system (1.3), and it should not come as a surprise. To clarify this point, let us consider a static and spherically symmetric perfect fluid. The most general spherically symmetric space-time metric can be written as

$$\begin{aligned} ds^2 &= g_{\mu\nu}\, dx^\mu\, dx^\nu \\ &= e^{\nu(r)}\, dt^2 - e^{\lambda(r)}\, dr^2 - r^2 \left(d\theta^2 + \sin^2\theta\, d\phi^2\right)\ , \end{aligned} \tag{1.4}$$

where ν and λ are functions of the areal radius r only. The Einstein equations (1.1) then yield the metric component

$$e^{-\lambda} = 1 - \frac{2\, m(r)}{r}\ , \tag{1.5}$$

where

$$m = \frac{k^2}{2} \int_0^r \rho(x)\, x^2\, dx \tag{1.6}$$

is the (Misner-Sharp) mass function of the self-gravitating system. Now let us consider the same perfect fluid under the "new" gravitational theory defined in Eq. (1.2). When the Eq. (1.3) are solved we obtain an expression which may generically be written as

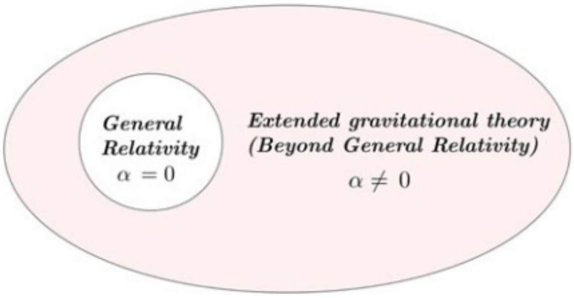

Fig. 1.1 MGD: the new gravitational sector not described by GR is fully controlled by the deformation parameter α. GR represents a "sub-space" in this extended version of gravity and is recovered for $\alpha = 0$

$$e^{-\lambda} = 1 - \frac{2\,m(r)}{r} + (\textit{geometric deformation})\ , \tag{1.7}$$

where the term *geometric deformation* should be understood as denoting the modification undergone by the metric component in Eq. (1.5) due to the generic extension of GR shown in Eq. (1.2). Therefore "deformation" means the difference with respect to a specific GR solution. Now, and this is very important, the deformation shown in the expression (1.7) always induces an *effective anisotropy* on the perfect isotropic fluid. Hence, the self-gravitating system does not behave like a perfect fluid anymore. Indeed, and this is a critical point in our analysis, the anisotropy $\mathcal{P}$ produced by the geometric deformation always appears to be of the form

$$\mathcal{P} = A + \alpha\, B\ . \tag{1.8}$$

This expression shows that the GR limit cannot be recovered simply by setting $\alpha = 0$, since the α-independent contribution A in the anisotropy (1.8) will survive in this limit. Consequently, the perfect fluid GR solution (1.5) is not contained in this extension and we could say that we have an extension to GR which does not contain GR. This seems a contradiction, but it is in fact mathematically correct and appears more properly as a problem of physical consistency. The source of this problem is the *geometric deformation* shown in Eq. (1.7), which always appears as

$$\textit{geometric deformation} = X + \alpha\, Y\ . \tag{1.9}$$

That is, there is a "sector" of the geometric deformation, generically called X, which is α-independent. Obviously this is physically problematic, since we expect to be able to control the deformation undergone by GR by varying the parameter α. It is important to remark once more that the source of this problem is solely given by the high non-linearity of the effective Einstein equations (1.3), and has nothing to do with any specific extensions of GR.

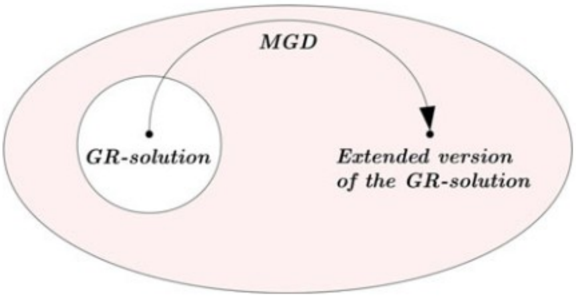

Fig. 1.2 MGD: when a GR solution is kept as a solution in the new gravitational theory, the α-independent new terms in the extended solution vanish, and the GR limit can be recovered

A method devised to solve precisely the non-trivial issue of consistency described above is the so-called *Minimal Geometric Deformation approach* (MGD) [14, 15]. The key idea is to keep under control the anisotropy induced by the extension to the Einstein theory. In particular, if we demand that the α-independent sector X in the geometric deformation (1.9) vanishes identically, the α-independent contribution to the anisotropy A in Eq. (1.8) will also vanish as a consequence. This procedure ensures that the extended theory will recover GR in the limit $\alpha = 0$. In this approach the generic expression Y in Eq. (1.9) represents the *minimal geometric deformation* undergone by the radial metric component, being the generic expression B in Eq. (1.8) the *minimal anysotropy* added to GR by the correction terms in the action (1.2) (see Figs. 1.1, 1.2).

Now, the question is how to implement the approach described above in practice. Or, more clearly, how to impose $X = 0$ in Eq. (1.9) in order to obtain a consistent extension to GR? This is accomplished when a GR solution is forced to remain a solution in the extended theory. Roughly speaking, we need to introduce the GR solution into the new theory, as far as possible. This provides the practical foundation for the MGD approach [16–24]. We want to emphasise that the GR solution used to make $X = 0$ in Eq. (1.9) will eventually be modified by using, for instance, the matching conditions at the surface of compact astrophysical object. This will in general allow us to express physical variables as functions of the deformation parameter of the theory, here generically named α. The first application of the MGD was to so called Brane-World (BW) models, and the parameter α was therefore given by the (inverse of the) brane tension. For this reason, most of this book will consider BW models of stars and other compact objects, such as black holes, although the MGD can be applied to any modification of GR of the form described above.

1.2 Extra-Dimensional Gravity: The Brane-World

Superstring/M-theory has been long considered one of the most promising candidates of quantum gravity. It describes gravity as a higher-dimensional interaction which becomes effectively four-dimensional at low enough energies. Extra-dimensional

- *)
C4 5 0
$ *)/ * +) 8 .1 :/ + 4) + Z)
; : 7))

$$S = \int \left(\frac{R}{2\,k^2} + \mathcal{L}_M\right)\sqrt{-g}\,d^4x + \alpha\,(correction)$$

3: #) 5 0

< \ , * * C)+

¸) *)))) + 4)&)) + : +
:/ .1

C4 5 0

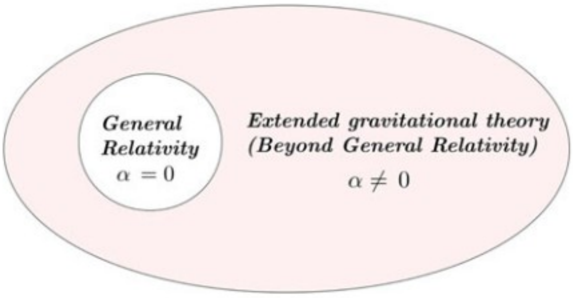
General
Relativity
$\alpha = 0$
Extended gravitational theory
(Beyond General Relativity)
$\alpha \neq 0$

D 4 ." 4)&)) + : + :/ .1 / + :/
+) *)) ¸
.1 *) b : *) ¨ 8 + + & 4)& /) + & + ¸ \

B0 !#12! ! !3 4) ! B))) *
2)) * 4)4 ' ; + : 4 5 0 *+

0 !#12! ! !3 4) ! B) 8 -) . /
* 4)4 ' ; + : 4 5 0 *+
C4 @ 5 0)* B
+)) +) 1)) . /
78 , E J) + *)

C4 B>5 0
B > # E) E J))
; 7 +) *) 1
) + ½) 9 + J)
c , e¾c , e) 8*)) c , ©0¾c , ©0 $ + *) *
)F / 4) (&) 5 U(0 (¿ * + 4 * \ ,
) *) : + * $ + * ') (
* &) 5 (0 (¿ 5D 4 B 0

C4 B@5 0

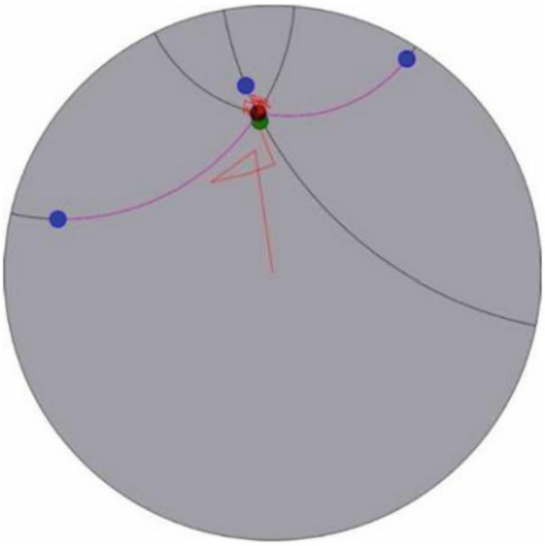

D 4 B +)
*) [+

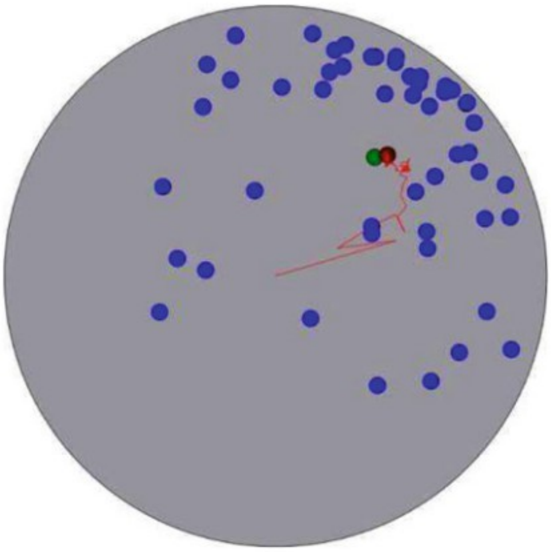

D 4 B +) *
) [+

B 0 !#12! ! !3 4) ! B)) /6.) 6$) *+
< *)4

B 0 !#12! ! !3 4) ! B 2) 4/6 #
D +)) " * E & * 4)4 % X (5 0 *+
< *)4

B 0 !#12! ! !3 4) ! B $ +)))
G (4 $ -) & . /) 5 , * 4)4 ' ; + : 40 *+

!"#$# % & & !' (! B!- *) ? 6' (6 B6 !A)
4 :) # . #**)

B 0 !#12! ! !3 4) ! BB , 7 A) # 4 :)6)

. #**) 5 B, 1 +4 6 1 0 *+
< B *)4
3: A & + ===

C4 >5 @0

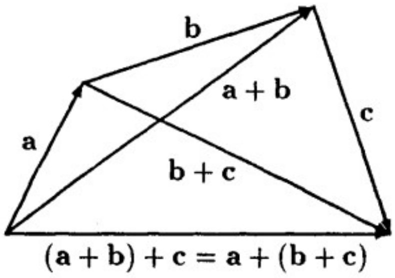
b
a + b
c
a
b + c
(a + b) + c = a + (b + c)

D 4

C4 >5 @0)* > ')) + J)+)

-) 4 / / 4 * *))( ) +
)(-) --) +) ,) 4) +
* * +)) / : + + +) *)) ,)
7 +) *) $ + + +)) 4 :), /
:)) + J)+) , : 4 +/ 4)* G) +)
*) : / + + + , :) +)) *)
))

3:
) 4 /
-)) 4 /, +) + 7 +I *) :
) 4 $ /* 4 /) + + :/ 7

3 & 4 ‡ # . /

B 0 !#12! ! !3 4) ! BB ' #) + F &
. / G / - 5 >0 +F&

B>0 !#12! ! !3 4) ! BB ' #) + F &
. / G / - 5 >0 *+
3: A & + ===
<)** I $) + D * , @@

< F &)

C4 @@5B 0)** I $) + D *

3.7 PAPPUS' THEOREM AND FURTHER PROPERTIES

Multiplication in a Desarguesian affine plane is not necessarily commutative. To present an example of a plane with noncommutative multiplication, some algebraic results are necessary, so the example is deferred until Chapter 4. However, the geometric counterpart of commutativity of multiplication is *Pappus' Theorem*, which was stated as a *Euclidean* theorem in Chapter 1, Theorem 12.

■ **PAPPUS' THEOREM.** Let $\ell \cap m = O$, with P, Q, and R in ℓ; S, T, U in m; $\ell(P,T) \parallel \ell(Q,U)$, and $\ell(Q,S) \parallel \ell(R,T)$. Then $\ell(P,S) \parallel \ell(R,U)$.

■ **THEOREM 17.** Pappus' Theorem is true in a Desarguesian affine plane if and only if multiplication on the points of its lines is commutative.

Proof: Choosing I in ℓ so that $\ell(I,S) \parallel \ell(P,T)$, then $PQ = R$ (multiplied using S (and T)). If Pappus' Theorem is true, then $QP = R$ as well (multiplied using S (and U)). (See Fig. 3.19) Conversely, if Pappus' Theorem fails, then, relative to O and I, $PQ = R$ but $QP \neq R$, so multiplication is not commutative. ■

Pappus and Desargues

In all the examples of Desarguesian affine planes presented so far, Pappus' Theorem holds. In fact algebraic methods can be used to show that Pappus' Theorem is true in any *finite* Desarguesian affine plane, but there are (infinite) Desarguesian affine planes in which Pappus' Theorem fails. On the other hand, it has been shown that Pappus' Theorem implies that of Desargues (Hessenberg 1905).

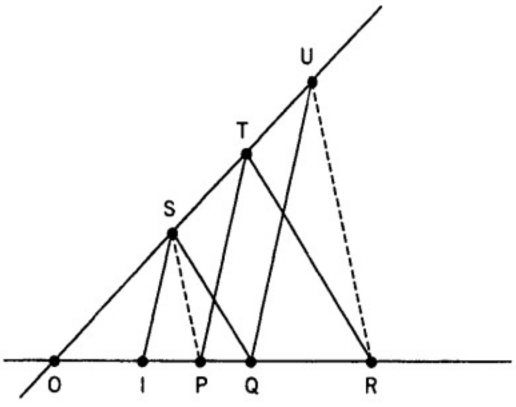

Figure 3.19

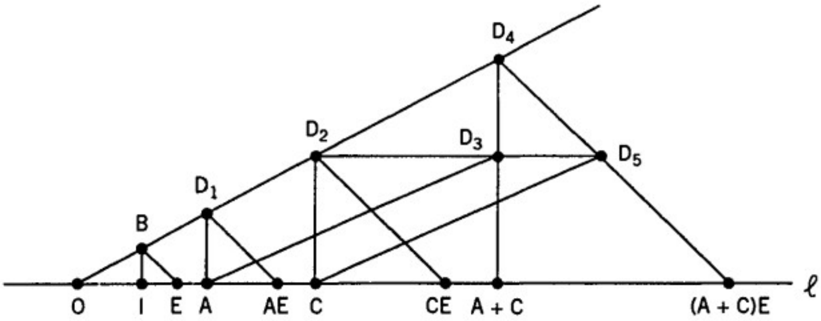

Figure 3.20

The Distributive Laws

Besides their individual properties, addition and multiplication are related by the *distributive laws*. Because multiplication is not necessarily commutative, two distributive laws are needed. Furthermore, after laws $(A + C)E = AE + CE$ and $E(I + C) = E + EC$ are proved *geometrically*, an algebraic proof shows that, for an arbitrary A, $E(A + C) = EA + EC$.

■ **THEOREM 18.** For A, C, and E in ℓ, $(A + C)E = AE + CE$.

Proof: Select $B \notin \ell$. (See Fig. 3.20.)
Compute:

AE using B (and D_1).
CE using B (and D_2).
$A + C$ using D_2 (and D_3).
$(A + C)E$ using B (and D_4).

Let D_5 be in $\ell(D_4, (A + C)E)$, so that $\ell(A, D_3) \| \ell(AE, D_5)$. Use Desargues' Theorem on D_4, D_3, D_5 and D_1, A, AE to get $\ell(D_3, D_5) \| \ell(A, AE)$, which shows that $D_5 \in \ell(D_2, D_3)$. Now add $AE + CE$ using D_2 (and D_5), and note that $\ell(D_2, CE) \| \ell(D_5, AE + CE)$. But this means that $AE + CE = (A + C)E$. ■

■ **THEOREM 19.** For C and E in ℓ, $E(I + C) = E + EC$.

Proof: Select $B \notin \ell$. (See Fig. 3.21.)
Compute:

EC using B (and D).
$I + C$ using B (and D_1).
$E + EC$ using D_1 (and D_2).

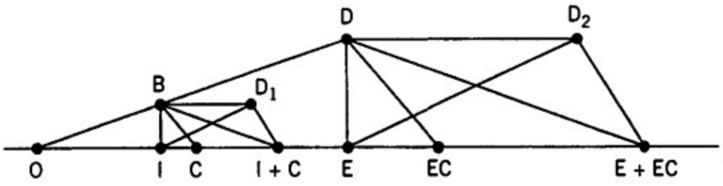

Figure 3.21

Use the converse of Desargues' Theorem on D_1, B, I and D_2, D, E which implies that $O \in \ell(D_1, D_2)$. Now use Desargues' Theorem on D_1, D_2, E + EC, and B, D_1, I + C, which gives $\ell(D, E + EC) \| \ell(B, I + C)$, whence E + EC = E(I + C). ■

Corollary. For A, C, and E in ℓ, E(A + C) = EA + EC.

Proof: If A = O, the result is immediate. If A ≠ O, then

$$\begin{aligned}(E(A + C) - EA - EC)A^{-1} &= E((A + C)A^{-1} - E - ECA^{-1} \\ &= E\left(I + CA^{-1}\right) - E - ECA^{-1} \\ &= E + ECA^{-1} - E - ECA^{-1} \\ &= O.\end{aligned}$$

Thus E(A + C) − EA − EC = O, and

$$E(A + C) = EA + EC.$$ ■

The fundamental theorem of synthetic affine geometry, stated as Theorem 1 of this chapter and repeated below, has now been proved.

■ **THEOREM 20. The fundamental theorem of synthetic affine geometry.** Relative to fixed coordinates O and I, any line in a Desarguesian affine plane is a *division ring* in that its addition and multiplication are associative, and have identites, every element has an additive inverse, and every nonzero element has a multiplicative inverse. Addition is commutative, and multiplication distributes over addition.

In the next chapter we will prove that for any division ring $\mathbb{D}$, the elements of $\mathbb{D} \times \mathbb{D}$ are the points of a Desarguesian affine plane whose lines satisfy linear equations of the form $xa + yb = c$. Conversely, any Desarguesian affine plane can be so represented over (or *coordinatized by*) a division ring.

This chapter concludes with another example of an affine plane of order 5. It will be shown in Chapter 4 that this plane can be coordinatized by the field $\mathbb{Z}_5$.

□**EXAMPLE 1: An affine plane of order 5.** $\mathscr{P} = \{A, B, C, \ldots, Y\}$, and the lines are

{OIABC}
{DEFGH}
{JKLMN}
{PQRST}
{UVWXY}

{OHMRV}
{IDNSW}
{AEJTX}
{BFKPY}
{CGLQU}

{ODJPU}
{IEKQV}
{AFLRW}
{BGMSX}
{CHNTY}

{OGKTW}
{IHLPX}
{ADMQY}
{BENRU}
{CFJSV}

{OELSY}
{CDKRX}
{BHJQW}
{AGNPV}
{IFMTU}

{OFNQX}
{IGJRY}
{AHKSU}
{BDLTV}
{CEMPW}

□

EXERCISES

The following exercises pertain to the plane of order 5 given above.

1. Make addition and multiplication tables for {OIABC}.
2. Make addition and multiplication tables for $\mathbb{Z}_5$. Show that {OIABC} is *isomorphic* to $\mathbb{Z}_5$ by finding a bijection from {OIABC} to $\mathbb{Z}_5$ that preserves addition and multiplication. [*Hint:* Map O to 0 and I to 1.]
3. Find a parallelogram in $(\mathscr{P}, \mathscr{L})$ and find the point of intersection of its diagonals.
4. Find an instance of the configurations for Desargues' Theorem in $(\mathscr{P}, \mathscr{L})$.
5. Find an instance of the configuration for Pappus' Theorem in $(\mathscr{P}, \mathscr{L})$.
6. Use $(\mathscr{P}, \mathscr{L})$ to define a projective plane of order 5.

*7. Let $(\mathscr{P}', \mathscr{L}')$ be a projective plane in which Pappus' Theorem holds, and prove Desargues' Theorem as follows: Let $\ell(A, A') \cap \ell(B, B') \cap \ell(C, C') = O$, and let $A'' = \ell(B, C) \cap \ell(B', C')$, $B'' = \ell(A, C) \cap \ell(A', C')$, and $C'' = \ell(A, B) \cap \ell(A', B')$. Define $E = \ell(A, C) \cap \ell(B', C')$, $F = \ell(A, B') \cap \ell(O, C)$, $G = \ell(A, B) \cap \ell(O, E)$, and $H = \ell(A', B') \cap \ell(O, E)$. Prove that A'', B'', and C'' are collinear by using Pappus' Theorem on O, B, B' and A, E, C; then on O, A, A' and B', C', E, and finally, on A, B', F and H, G, E. [*Hint:* In the first use of Pappus' Theorem, since $\ell(B, C) \cap \ell(B', E) = A''$, $\ell(A, B') \cap \ell(O, C) = F$ and $\ell(O, E) \cap \ell(A, B) = G$, we obtain A'', F, and G collinear.]

8. Find six orthogonal partitions of the set $\{1, 2, \ldots, 25\}$.

9. Find four orthogonal Latin squares of order 5.

SUGGESTED READING

G. Hessenberg, Beweis des Desarguesschen Satzes aus dem Pascalschen, *Math. Ann.* **61** (1905), 161–172.

David Hilbert, *Foundations of Geometry*, La Salle, IL: Open Court, 1971.

F. R. Moulton, A simple non-Desarguesian plane geometry, *Trans. AMS* **3** (1902), 192–195.

B@0 !#12! ! !3 4) ! BB $ / # . /
7 & - , #)+ - 5 0 *+
< *)4
3: A & + ===

D + & # . /

B 0 !#12! ! !3 4) ! BB & A A) # 4 :)
#**)) 5 0 *+
< > > *)4
3:
C4 B5 0)* > 3 4) /
C4 B 5 0)* @ 7 4 &)

BB0 !#12! ! !3 4) ! BB & H A A) # 4 :)
#**)) 5 0 *+
< > *)4

B 0 !#12! ! !3 4) ! BB $ + ,) #+)
A) # 4 :) # . #**) 5 , G ; D) 0 *+

!"#$# % & & !' (
! B!- *) ? 6' (6 B6 !)4 *) 78* 4 -)) % I
) *

0 !#12! ! !3 4) ! B) ( )4
*)6 78* 4 -)) % I) * 5 , ' (0 * :

0 !#12! ! !3 4) ! B) ( )4
*)6 78* 4 -)) % I *+
< *)4

0 !#12! ! !3 4) ! B) ( )4
*)6 78* 4 -)) % I) * ' (5 0 * :

0 !#12! ! !3 4) ! B .))) 6 # * 4
A (' + > - J ' (5 >0 * :

0 !#12! ! !3 4) ! B .))) 6 # * 4
A (' + > - *+

!"#$# % & & !' (
! B!- *) ? 6' (6 B6 ! *) . /) + - #**)

>0 !#12! ! !3 4) ! # - &) .))
5 , * 4 -)) : 46 ' (S 0 *+

@0 !#12! ! !3 4) ! 7 + * & 4 /6
)**) *) * F +F&

0 !#12! ! !3 4) ! 7 + * & 4 /6

)**) *) * F *+
B0 !#12! ! !3 4) ! ;) +: (.))
:/ $)) *+
< B *)4

0 !#12! ! !3 4) ! ;) +: (.)) :/
$)) *+
< B > *)4

0 !#12! ! !3 4) ! ;) +: (.)) :/
$)) + 7+ *+
< @ *)4

0 !#12! ! !3 4) ! ;) +: (
.)) 6 + *+

0 !#12! ! !3 4) !)) $ / G
#**) . :/ G D + , (*+
< > *)4

0 !#12! ! !3 4) ! *) D)))
5. 0 / 5 , ' (S 0 *+
< > *)4

0 !#12! ! !3 4) ! *) D
))) . #)) ,) ,) + $) * *+
< @ > *)4

>0 !#12! ! !3 4) ! *)) * 4 5 B,
* 4 -)) : 46' (S 0 *+
< > *)4

@0 !#12! ! !3 4) ! *)) * 4 :/ G
D + , T) %) +,) *+

0 !#12! ! !3 4) ! G / *) . /
+ - #**) *+
< B *)4
3: A & + ===

!"#$# % & & !' (! B!)

B0 !#12! ! !3 4) ! ; /) ' B> *+

!"#$# % & & !' (! B!) ,) + 2)+)
#**)

0 !#12! ! !3 4) !) ,) + 2)+)
#**) *+

C4 B@5 0
)* @ .) 2)+) B
@ 2)+) 1 B
@ $.) 2)+) 1 B@
@ $ # .) 2)+) 1
@ $.) R + 2)+) 1 >
@ > $ % / .) 2)+) 1
@ @ $ ' (.) 2)+) 1

C4 5 > 0

)* 78) * .) 2)+) 1
$. :) + G #**)
*) $):
E 4 D 21 # 4
- *) 21
> E 4 2A # 4
@ .) 1)+) 2)+) 1
.) A :) 2)+) 1 @
B # .) 2)+) 1 B
.) R + 2)+) 1 B
*) - 4)
+ Z) # 4 >>
- & 7 4 &) : >@

)*

78) * .) 2)+) 1
E @ , J)+) * + :/) + 4 + ( ))
) + * : +): 5 ,) , 4 & c e0
4 & 4 +) + 4) 4 & + 4)** 8) -)) 4)
* &))**) +) 4 *) +) +
4 $) : + :/ : & 4 /) J) K
,) , D) +) +): * : + : (+
) + c e ; & , + +)/ * +
) + 4 4 H) :) 8 * + 4 4) * /
))) + 4 &)) + &) 4) -)*
4 & 8) * *) .) J)+) G *)
) / &) + :) +) + 4 G) 4 & 8
) * *) 4))) 8) * + Z)
) 4 4 &)

$. :) + G #**)
$)/ 4) +) 4 & :/ 4 &) H) :) 8
) + 4) 4 & :/ J) Z *)
9 + 4 &)))+/ ( ) : 4 4 @ K . 9 ,
G) + 4) + #4) c e G c e . :) + G c @ e +
) + + & +)) 4 :) + 21) G ((*
+) + 4 - +))/) / Z *
4 &) * + & , / * : + # 4 *
+ * 4) 4 G)) + * +)):) +
* + * .) J)+)

D)) 4) * /) , Z) 8
* / (K) .) c e, A) c e)* $): 4 &
A 4 + 4 J) c P , e 4 * + &)
Z H) :) 8,) (8* /K)* 3 , +)/)
):) / *) +) + ( ) * *
+) + 4 -)&) 4 + * J)+)
G) , * * *) , +) + 4) /) +
$): +) + 4) * .) 4 :/ & 4
3:& /, 5) 0 4) § > @> >B „P
+ + @ +) Z4 $) *) +) + 4 $ +
) /) + *) +) + 4

!"#$# % & & !' (! B!3

!"#$# % & & !' (! B!3 !

0 !#12! ! !3 4) ! .): 6 6>B 6> > *+
< A +) ' /) A :) & ( ) /* : 4 /

0 !#12! ! !3 4) ! .):) *+
< &) 4 + &) ' /) A :) & ( ) *)

0 !#12! ! !3 4) ! .)1) + *+
$ 1) ;/*
3: +)+ # 4 1))+

!"#$# % & & !' (! B!3 !

0 !#12! ! !3 4) ! B 9) *+
<)) *

0 !#12! ! !3 4) ! *)) *+
<)** N) 7) & & / + /* : 4 /

!"#$# % & & !' (! B!3 !

>0 !#12! ! !3 4) ! > >) +)) *+
< $)))&)&)* * *)4))) 4) + + : +

@0 !#12! ! !3 4) ! > /) *+
< ") +)+ & 8 / H) (1 :

!"#$# % & & !' (! B!3 !

0 !#12! ! !3 4) ! @)+F))) *+
<))))) /)) * : : +): : +

!"#$# % & & !' (! B!3 !

B0 !#12! ! !3 4) ! @ * /)) *+
< #** 8) Z8 + * 4) 9 + $) *)

!"#$# % & & !' (! B!3 !
0 !#12! ! !3 4) ! %)4) *+
< $ &)) & 4 * +

!"#$# % & & !' (! B!3 !

0 !#12! ! !3 4) ! B) *+
< # * * +))) A : 4 4)

!"#$# % & & !' (! B!3 !

0 !#12! ! !3 4) ! #/+) *+
< $ J +) 4 / & +) 4)) *) .

!"#$# % & & !' (! B!3 !

0 !#12! ! !3 4) ! # +) ') :) *+
< $ /* : ") 4) [+ /* : 4 /

!"#$# % & & !' (! B!3 !

0 !#12! ! !3 4) ! :)4)) *+
< $)) /* : *)

0 !#12! ! !3 4) !) /) *+
< " : + * 4) " *))** 8) * :

!"#$# % & & !' (! B!3 ! >

>0 !#12! ! !3 4) ! @ &) &) *+
<) +)**) .)) J)+)))) + 8 4

!"#$# % & & !' (! B!3 ! >

@0 !#12! ! !3 4) ! -)) *+
< ; ;)+)) + D Fˆ /* J) & 8 &))) 4)

0 !#12! ! !3 4) ! ') 9 4) /) *+
< 78 R)))) J) / 4) 9 + & 4

B0 !#12! ! !3 4) ! .):) *+
< &) 4 + &) ' /) A :) & ( ) *)

!"#$# % & & !' (! B!3 ! >

0 !#12! ! !3 4) ! .): 6 6>B 6> > *+
< A +) ' /) A :) & ( ) /* : 4 /

!"#$# % & & !' (! B!3

0 !#12! ! !3 4) ! > D - *+

<   B *)4
C4   1   )   /   ) + .   /
5) [*C4) B0

!"#$# % & &   !' (!   B!$) 4) +) 4

0 !#12!   !   !3 4) ! @ $) 4) +) 4   *+
<   *)4

C4   5 0   ) [*C4 > 5@ 0

$) 4) +) 4   + Z   , * *   ,) + 8) *

!"#$# % & &   >
WWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWW
WWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWW
WWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWW

!"#$# % & &   >   !' (>

!"#$# % & &   >   !' (>   !

>

>   B

!"#$# % & &   >   !' (>   !

!"#$# % & &   >   !' (>   !

!"#$# % & &   >   !' (>   !   !"   ) #) / $ 4)
*   *   +

0 !#12!   !   !3 4) !   "   ) #) /   *   *   *+

!"#$# % & &   >   !' (>   !   !"   I G (

0 !#12!   !   !3 4) ! B   4 F   +   (*+
< @> *)4

0 !#12!   !   !3 4) ! B   4 F   +   (*+
<   > *)4

!"#$# % & &   >   !' (>   !   !3 +   D   # %)   ) ;   /
)

>0 !#12!   !   !3 4) !   3 +   D   # %)   )
;   /   )   4) V   )/*   :   5   >0 *+

C4 B   )*   B $ D   %)

!"#$# % & &   >   !' (>   !   ! *)   #) /   ) + "   4 ,   +

+

@0 !#12! ! !3 4) ! *) #) /) +
" 4 *+
< *)4

C4 >5 0)* > *) . /

C4 5 B0)* @ - + # +/)

C4 > 5 @0)* 3 :) +) " 4

0 !#12! ! !3 4) ! *) #) /) +
" 4 5 >0 *+
< > *)4

!"#$# % & & > !' (> !

!"#$# % & & > !' (> ! !- + 3 :) :)

B0 !#12! ! !3 4) !
A 4 (6' (6- + $ 3 :) :) *+

!"#$# % & & > !' (> ! !A], R # . /
& 8 ' +

0 !#12! ! !3 4) ! B B >
) *+
< F * : / * / *

0 !#12! ! !3 4) ! @ > *+
< # 4) 4))))

0 !#12! ! !3 4) ! B @ *+
< 78 + + #) #)

0 !#12! ! !3 4) ! > *+
< # . / & 8 ' +
'/ R A '
3:) &

0 !#12! ! !3 4) ! : *+
< 3 # & 8 / & 8 &) + ;/*)

0 !#12! ! !3 4) ! ' #) + F &
. / G / - 5 >0 +F&

>0 !#12! ! !3 4) ! ' #) + F &
. / G / - 5 >0 *+

@0 !#12! ! !3 4) ! $ / # . /
7 & - , #)+ - 5 0 *+

< *)4
3: A & + ===

C4, @5 0 D 4 @

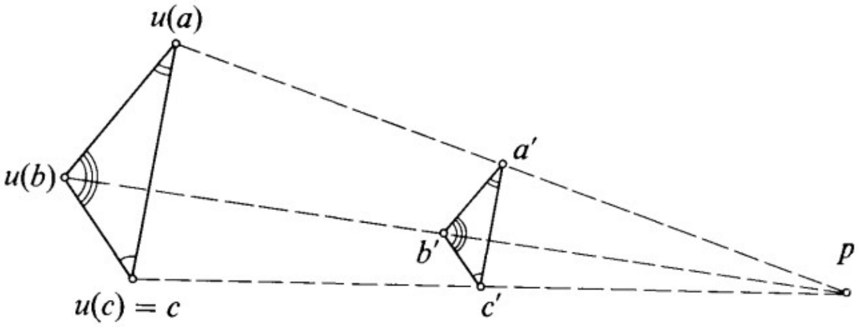
u(a)
a′
u(b)
b′
p
u(c) = c
c′

78 >): &)) $) 4 5)0 5 0 5:0) +)I I:I)&
*)) +) +) 5 * B 0 $ +)
)*) 4 :)4)
*) +)
$)) 4 D 4 @)
+))& :)+ ") 4 $
& /)) / + :)+) *) 4) 4
)) 4 / :F) 4 & *
) /))

0 !#12! ! !3 4) ! &) *+
<)) # . / & 8 ' +

!"#$# % & & > !' (> ! !% : $ / c$ "De

B0 !#12! ! !3 4) ! % : $ / c$ "De *+
3: A & + ===

C4 5 >0 /)4) $ *

)* " *) 7J)
A) " *) 7J)
78
/)4) $ * >
78
7) / + B
78
.)) - 4
78
> J)
78 > >
@ G) 4N : @
78 @
D) N A) $
78
B N 7J) >

78 B >>
) >@
78 @

!"#$# % & & > !' (> ! !$ *)) .)& /

0 !#12! ! !3 4) ! # - +
$7A7 #1#AA7A .1# -$X - + D?) $) E & +) + 7)+)))))
')9 5 0 *+
< *)4

> 0 !#12! ! !3 4) ! :)) ')) + + +
$ *)) 4)& / c " $ e E A 5 B0 *+
< *)4

!"#$# % & & > !' (> !

!"#$# % & & > !' (> ! !#+) , . 9 , # [J ,
/ , @

!"#$# % & & > !' (> ! !# 4 :) #

!"#$# % & & > !' (> ! !#

> 0 !#12! ! !3 4) ! > # 4 :) . /) +
& 38 + E & / , E # 5 0 +F&

> 0 !#12! ! !3 4) ! > # 4 :) . /) +
& 38 + E & / , E # 5 0 *+

> 0 !#12! ! !3 4) ! > #) + . /
+F&

> 0 !#12! ! !3 4) ! > #) + . /
*+

>>0 !#12! ! !3 4) ! > #) + . /
. / +F&

>@0 !#12! ! !3 4) ! > #) + . /
. / *+

> 0 !#12! ! !3 4) ! > # . / # + % :
$ / G + 5 @0 *+

>B0 !#12! ! !3 4) ! > # #) + 4 /
) + +) + - 1)) &) 8 : +)/ ' (S
5 B 0 +F&

> 0 !#12! ! !3 4) ! > # #) + 4 /
) + +) + - 1)) &) 8 : +)/ ' (S
5 B 0 *+

@ 0 !#12! ! !3 4) ! > # #) + 4 /
) + +) + - 1)) &) 8 : +)/ ' (S
5 B 0 +F&

@ 0 !#12! ! !3 4) ! > # #) + 4 /
) + +) + - 1)) &) 8 : +)/ ' (S
5 B 0 *+

@ 0 !#12! ! !3 4) ! > ")& * $; 4 #
# - + $ / % :) : +4 E & / 5 0 *+

@ 0 !#12! ! !3 4) ! > " 4) +
4) * : :) ,) ,) + 4 / * 4 5 0 +F&

@ 0 !#12! ! !3 4) ! > " 4) +
4) * : :) ,) ,) + 4 / * 4 5 0 *+

:)) . / 5 0
/) + &
/ &) + 1 4
& 4) +) (4 >@
4

@>0 !#12! ! !3 4) ! > . /) + #
7 *)))) / 5 0 *+

@@0 !#12! ! !3 4) ! > ; 9)* 1 #) +
. /) + ;/* 4 5 0 *+

@ 0 !#12! ! !3 4) ! > H) + D +) * 4
5 >0 *+
< B B *)4
< D + #
3: A & + ===

@B0 !#12! ! !3 4) ! > A # 4 :) . /) +
& 38 + E & / 5 0 +F&

@ 0 !#12! ! !3 4) ! > A # 4 :) . /) +
# & 38 + E & / 5 0 *+

0 !#12! ! !3 4) ! >)* 4 #) D
.) I " J #) 5 0 *+

0 !#12! ! !3 4) ! >)) +))
) + 38 + E & / 5 0 +F&

0 !#12! ! !3 4) ! >)) +))
) + 38 + E & / 5 0 *+

0 !#12! ! !3 4) ! > $ 4) /)
4 /) +) ' (S * 4 c+ : e 5 0 *+

< > B *)4
3: A & + ===

!"#$# % & & > !' (> ! !)4) #

0 !#12! ! !3 4) ! @)4) $) +)) J
*) +F&
< > *)4

>0 !#12! ! !3 4) ! @)4) $) +)) J
*) *+

@0 !#12! ! !3 4) ! @)4) $) +)) J
*) +F&
< @@B *)4

0 !#12! ! !3 4) ! @)4) $) +)) J
*) *+

B0 !#12! ! !3 4) ! @)4) $) +)) J
*) +F&
< >@B *)4

0 !#12! ! !3 4) ! @)4) $) +)) J
*) *+

B 0 !#12! ! !3 4) ! @)4) $) +)) J
*) +F&
< @B *)4

B 0 !#12! ! !3 4) ! @)4) $) +)) J
*) *+

B 0 !#12! ! !3 4) ! @ 9 # [J ^
-) 6‡6 N) 4g: 6) 6+ 6> ,6 6 6 6 6) + +) 6) 86 6)+)
")4)& 5 0 +F&

B 0 !#12! ! !3 4) ! @ 9 # [J ^
-) 6‡6 N) 4g: 6) 6+ 6> ,6 6 6 6 6) + +) 6) 86 6)+)
")4)& 5 0 *+

B 0 !#12! ! !3 4) ! @ A : [, ;[/
.[[666) 6+ 6) [ ) J 5 0 *+
< *)4

B>0 !#12! ! !3 4) ! @ A7'3 7 ;7 71X
ijklmnopqrsstuvwx
< @ *)4

B@0 !#12! ! !3 4) ! @ A7'3 7 ;7 71X
yz{|z}n~opqrsstuvwx

B 0 !#12! ! !3 4) ! @ A7'3 7 ;7 71X $ '
ijklmnopqrs•tuvwx

BB0 !#12! ! !3 4) ! @ A7'3 7 ;7 71X $ "
yz{|z}n~opqrs•tuvwx

!"#$# % & & > !' (> ! !- *)

B 0 !#12! ! !3 4) ! > B *+
< > *)4
< D +))) +

0 !#12! ! !3 4) ! '/ 4 # 4 :), B
7+ .) ; 67 4 46) 5 0 *+
< B *)4

0 !#12! ! !3 4) ! ")& * $; 4 #
- + $ / % :) : +4 E & / 5 0 *+

0 !#12! ! !3 4) !
D +) 6)) 6:/6A 61 6 *+
< @@ *)4
C4 B 5 0)* > . /

0 !#12! ! !3 4) ! H : 4 # 4 :)
.) ; 67 4 46) 5 0 *+
< > *)4

0 !#12! ! !3 4) ! G *))
&) 5 , # 0 *+
< *)4

) - 1)) . /
)* - # ") . /
- - "):) +
+
") >
)*
> A 4 *
@ $) : + *))) + 4 +
&) ,) + J)
B & 4 *)
$)) # : ; ($
)* 1)) &) >
$ A & &) @
) &) >
-) + &) >
+ *)) &) @
> $ +) *)
@ D) 4 * B

; 4 *) BB
B #** + 8 $ *))
))) +
*) -- $ 7 +) *) D :
)* D) 1))) + B
" 4 * +) *) B
$ ' : :) $ /) 4)*
4 *
#**) +) *) >
2 /
> $ +) +) *) ---
@ $)) ( )) * :
) *)) +
D) 4 * +))) +

*) --- $ *) *) D :
)* 1 *) D . * B
') + B
$ D :)
D : * /) + 4 *) 4 :)
" & : / >
> $ * +) + +) *)
@ $) *) &) 4 :)) /
+ + >
E) /) + 4) *) >
)* > N G (*) *) D
: >
> N * 4) >
>) * 4 * >@
> %) / + 8 + * 4 * >
>)) * /* 8 + *
4 * @
> > 1 *) 4 * & /
/ :4 * / @>
> @ # *)) *) *)
* :
)* @ $)) D 8 + D . *
@ T)) N (&): 4 * /
++ / :4 *
@ $:) /) +) 4 * B
@ % &): 8 + * 4 * >
)* $ *) *) D
: B
1 *) :) / * / +) 4 * B
D 8 + * * 8 *)
$)) * *)
$)) *) *) B
> *) *) +
@ +))
*) - *) D : / *)
)* B 1)) / *)
B A)) / / *)

B 4) & & A) 4 :)
B . :) / / *)) + 4) & À
& A) 4 :)
B &) >

B > 4/
B @)) :) 4 :) ,) ( ) +)8) >
B ;) / *) >
B B $ 4 * @
B 78 + +) "/ (+)4) @
B :4 *)8)) (>
B $)) / *) B@
B $ * 4 *)
B #** + 8) + + :)
*/ 4 *
)* *) D - + : /
*)
D) : / *) * :
.))) +) / *) @
.))) + & +
.))) + ++ +
> / *) * &))
@ #) +) + >
)* A) / / *) % 4) &
&) B
$
#**)
*) *) D : - +
) +
)* *)) &)
$)) *) B
$ 4 / * + *) *)
; 4 *)
$) *) >
> # + A 9 4) @@
@ $)) 4) +
) &)
)* A) / - *) +
1 + & A 4 *
78) *) / *) + B
) / * *) B>
*)))) *) /
*) + B
#** + 8)* @
1
#++) 1 B
- + 8

!"#$# % & & > !' (> ! !- *) '33R)

>0 !#12! ! !3 4) ! B #+) , . 9 , # [J
@ *+

@0 !#12! ! !3 4) ! B A *) +, , .) 9 ,
) [ ) J ,) + 8*[ ) @> *+

0 !#12! ! !3 4) ! B A *) +, , $ 4 [,
) [ ) J [[ ) @ *+

B0 !#12! ! !3 4) ! B A 4 , # 4g: $ 4 [,
+ 5 > 0 *+

0 !#12! ! !3 4) ! B A 4 , + 4[[,)
+) [ ) J , F > *+

> 0 !#12! ! !3 4) ! B 4 , .) ,) [ ) J
5 @ 0 *+

> 0 !#12! ! !3 4) ! B 2 /) 1 & 9,
) $) 7 4[[5 0 *+

!"#$# % & & > !' (> ! !3 :) / # + . +
+

> 0 !#12! ! !3 4) ! " & 4 /
%)) ' + 3 :) 5 0 *+

> 0 !#12! ! !3 4) ! 3 :) / # +
. + +) : +4 E & / 5 0 +F&

> 0 !#12! ! !3 4) ! 3 :) / # +
. + +) : +4 E & / 5 0 *+

> >0 !#12! ! !3 4) ! 3 :) / +F&

> @0 !#12! ! !3 4) ! 3 :) / *+

> 0 !#12! ! !3 4) ! 3 :) / * 4
)4 ' ; + : 4 5 0 *+

!"#$# % & & > !' (> ! >

!"#$# % & & > !' (> ! >!ƒ 4 :) / . ?) #) ?)

> B0 !#12! ! !3 4) ! # +) . /
3 $ ") G + 5 : & 4))) + ' / +0 *+

> 0 !#12! ! !3 4) ! # +) #+&) +
7 +) . / R / 4 5 0 +F&

> 0 !#12! ! !3 4) ! # +) #+&) +
7 +) . / R / 4 5 0 *+

> 0 !#12! ! !3 4) ! # +) ,) $
) (+) 4 4 * : 4 / " & :) 5 @0 +F&

> 0 !#12! ! !3 4) ! # +) ,) $
) (+) 4 4 * : 4 / " & :) 5 @0 *+

> 0 !#12! ! !3 4) ! # +) , - 4)
A) 6 # ' 4)* / G +I / % : 5 0 *+
< *)4

> 0 !#12! ! !3 4) ! # +) , - 4)
A) $ . . + 1) ' (5 0 * :
< B> *)4

> >0 !#12! ! !3 4) ! # +) , - 4)
A) $ $) 4 6 #))) H / ' (5 0 * :
< @@ *)4

> @0 !#12! ! !3 4) ! # +) , 1 :
. S4 $ 6 #))) 78*) ' / + A ' (
5 @0 * :
< *)4
3: A & + ===

> 0 !#12! ! !3 4) ! # +) , 1 :
. S4 $ 6 #))) 78*) ' / + A ' (
5 @0 *+

> B0 !#12! ! !3 4) ! # +) , *
R (: & 4) 4) + 4) 6)
))) 5 B0 +F&

> 0 !#12! ! !3 4) ! # +) , *
R (: & 4) 4) + 4) 6)
))) 5 B0 *+

> 0 !#12! ! !3 4) ! 7) &
4 / ' (S ' 5 0 +F&
< *)4

> 0 !#12! ! !3 4) ! 7) &
4 / ' (S ' 5 0 *+

> 0 !#12! ! !3 4) ! A + A # 4 :) X
$ 4) .) #)) *+
< *)4

> 0 !#12! ! !3 4) ! ((# 4 :) /
4) 4)))) 7+ $ -)) 5 @0 *+
< @ *)4

> 0 !#12! ! !3 4) ! ((ƒ 4 :) /
$ 4 ?) . ?) #) ?) 4)4 A) 4 7+ 5 0 *+

> >0 !#12! ! !3 4) ! T ƒ 4 :), 4 ?) /

4 ?))) ?),) 7+ , 5 0 *+

!"#$# % & & > !' (> ! >!# 78 4 7) /
))

> @0 !#12! ! !3 4) ! # 78 4 7) /
)) , - 1) % :) + D * 4 -)) : 4 5 0 *+
< @> *)4
< 1) % :) + D

> 0 !#12! ! !3 4) ! # 78 4 7) /
)) , -- 7 +) . / * 4 5 B0 *+
< >> *)4
< 7 +) 4 /

> B0 !#12! ! !3 4) ! # 78 4 7) /
)) , --- "))) + /) # 4 :) * 4 5 B0 *+
< @ *)4
< ")))) + /) # 4 :)

> 0 !#12! ! !3 4) ! #)) 9 %
.) 7 +)))) +)+ ')) +) C) 5 0 *+
< B *)4
3: A & + ===

> 0 !#12! ! !3 4) ! # ¤)) 9 %
$,* +) C) 7) > $) + %Á > +)+ '))
+) C) c ' e *+
< @B *)4

!"#$# % & & > !' (> ! >!# T

> 0 !#12! ! !3 4) ! # T ')) +
; 4 4 D +)) / G + : 4 *) / 5 0 +F&
< @ B *)4

> 0 !#12! ! !3 4) ! # T ')) +
; 4 4 D +)) / G + : 4 *) / 5 0 *+

> 0 !#12! ! !3 4) ! # T D 2) (
) + . 2) .)& / G + : 4 *) / 5 0 +F&

> 0 !#12! ! !3 4) ! # T D 2) (
) + . 2) .)& / G + : 4 *) / 5 0 *+

> >0 !#12! ! !3 4) ! # T D 2) (
') (; 6 4 E +) + 4 A 4 %) 6 + 4 -))
:) / *+
< @ *)4

> @0 !#12! ! !3 4) ! # T .)&)
) + + 4/6 $ 4)) : *+

> 0 !#12! ! !3 4) ! # T ;) +
4 : / +) +) + + 7 G + : 4 *) /
5 0 +F&
< *)4

> B0 !#12! ! !3 4) ! # T ;) +
4 : / +) +) + + 7 G + : 4 *) /
5 0 *+

> 0 !#12! ! !3 4) ! # T $ 4
) , * 8 /) + * / D J) (4 *)) * 4 , J) 4)& /
) + : / + +F&
< @BB *)4
3: D13 2E#1R .AE3% A# # $3 E 71 $1-%. ,
2E#%$E .1# -$X #%" '7X3%"

> 0 !#12! ! !3 4) ! # T $ 4
) , * 8 /) + * / D J) (4 *)) * 4 , J) 4)& /
) + : / + *+

> 0 !#12! ! !3 4) ! # T $) +
J 4 : / +) +) + + G + : 4 *) /
5 0 +F&

> 0 !#12! ! !3 4) ! # T $) +
J 4 : / +) +) + + G + : 4 *) /
5 0 *+

> 0 !#12! ! !3 4) ! # T 5 + 0 $ /
) + 78* ;)+ 4 % / G + : 4 *) / 5 0 *+
< @ *)4

> 0 !#12! ! !3 4) ! # T 57+ 0 ;
+ G . ' / +) +) + + 5 :) 0 5 0 *+

> >0 !#12! ! !3 4) ! # T 5 + 0 .) 4
-) $ /) + 78* * 4 E 5 B 0 *+
< B *)4

> @0 !#12! ! !3 4) ! # T 5 + 0 3 +) +
% D %) * 4 E 5 BB0 *+
< B *)4

> 0 !#12! ! !3 4) ! # T 5 + 0 /
) ; 4 7 4/) + A / $ E +) + 3 4) * 4 E 5 0 *+

> B0 !#12! ! !3 4) ! # T 5 + 0 $
* + --- * 4 E 5 0 *+
< *)4

> 0 !#12! ! !3 4) ! # T 5 + 0 $
* + -- * 4 E 5 0 *+

< >>> *)4

>> 0 !#12! ! !3 4) ! # T 5 + 0 $
* + - * 4 E 5 0 *+
< *)4

>> 0 !#12! ! !3 4) ! # T 5 + 0 $
E / D +)) -) * 4 E 5 B 0 *+
< B *)4
< ") ,

>> 0 !#12! ! !3 4) ! 4 $ 4 X) , & %)+ $
)* - *) 6 4 $ /) + . / E & I ; ++ " ')
' (5 0 *+

>> 0 !#12! ! !3 4) ! 4 .)& /) + /)
) (7 4/) * 4 %) + 5 @0 *+
< > > *)4

!"#$# % & & > !' (> ! >!# T !.)

>> 0 !#12! ! !3 4) ! # 4 A *+
< .)) 4 :))
< > *)4
3: A & + ===

>>>0 !#12! ! !3 4) ! : : 4))64) *+

>>@0 !#12! ! !3 4) ! '3-D.7 *+
< D +) 4 * : + *)9 *
"))) & ‡ 4)))) + * +
< *)4

>> 0 !#12! ! !3 4) ! #1# -3A3 *+
< @ *)4
< A + *)9 + ;) 1) :
" 4& 4
A [+ I *) + ; 1) :
" 4& 4

>>B0 !#12! ! !3 4) ! .) 99 *+
< .73 7$1-# A# - -3%7 #$7 #$- # "7AA3 #T-3

>> 0 !#12! ! !3 4) !) &) *+
< A) 1) & ‡

>@ 0 !#12! ! !3 4) ! $ *+
< $ *)9 *) + 7))

!"#$# % & & > !' (> ! >!# T ! 4 $ 4

>@ 0 !#12! ! !3 4) ! 5 & / +) 4 /

U- 0 A 9 H 6 # G * 6 4 $ 4 X) . / 1))) +
+ *) -)) 5 0 *+
< B *)4

>@ 0 !#12! ! !3 4) ! 5 & / +) 4 /
U 0 $) 9 ()6 4 $ 4 X) * &))) + * /)/
+ +) + -)+ 4 N B> : +)/ -)) 5 0 *+
< *)4

>@ 0 !#12! ! !3 4) ! 5 & / +) 4 /
U -0 %) 4) A 46 4 $ 4 X) . / *) /) +) + *
-)) * 5 0 *+
< B *)4

>@ 0 !#12! ! !3 4) ! 4 $ 4 X) 5 7+ 0
.) . $ / 5 0 +F&

>@>0 !#12! ! !3 4) ! 4 $ 4 X) 5 7+ 0
.) . $ / 5 0 *+

>@@0 !#12! ! !3 4) ! 4 46/) 6)+ 6
6 6)* 6 6 6 *) *+

>@ 0 !#12! ! !3 4) ! & / ") . /
") . / - * + :/ 4 $ / & -)) ' 5 0 +F&

>@B0 !#12! ! !3 4) ! & / ") . /
") . / - * + :/ 4 $ / & -)) ' 5 0 *+

>@ 0 !#12! ! !3 4) ! & / ") . /
-)) ' 5 B0 +F&

> 0 !#12! ! !3 4) ! & / ") . /
-)) ' 5 B0 *+

> 0 !#12! ! !3 4) ! /) 6)* *) *+

!"#$# % & & > !' (> ! >!7 . / A)++

> 0 !#12! ! !3 4) ! > 7" > B *+
< > *)4
< . $)) . *) + *

> 0 !#12! ! !3 4) ! > 7 *+
< > > *)4

> 0 !#12! ! !3 4) ! > 7 7 + *+

!"#$# % & & > !' (> ! @

!"#$# % & & > !' (> ! @!# 4 :) % E + 4 +

> >0 !#12! ! !3 4) ! @ # # 4 :))*
)))) / 5 0 +F&
< *)4

> @0 !#12! ! !3 4) ! @ # # 4 :))*
)))) / 5 0 *+

> 0 !#12! ! !3 4) ! @ # # 4 :) % D $
E + 4 +) : +4 E & / 5 0 *+
< *)4

!"#$# % & & > !' (> ! @!78 *) & % :
$ /

> B0 !#12! ! !3 4) ! 78 *) &
% : $ / *+

!"#$# % & & > !' (> ! @! /)4) D 99/ $ /) +
#**)

> 0 !#12! ! !3 4) ! B
6' (6 /)4) D 99/ *+
< *)4 ,

!"#$# % & & > !' (> ! @!$ * ;) (4,
B

>B 0 !#12! ! !3 4) !
" 6 N 67+ 6$ 6 6 6 6 & 6;) (4,6 B *+

!"#$# % & & > !' (> ! @!$) I) ' (, - 7+
7+

>B 0 !#12! ! !3 4) ! - I))
$) I) 7) / $) +) , 7+) 5 0 *+
< @ *)4

>B 0 !#12! ! !3 4) ! $) I) 7) /
$) +) 5 7+ 0) 5 0 *+
< > *)4

>B 0 !#12! ! !3 4) ! $) I) ' (, -
7+ 7+ *+
< *)4

!"#$# % & & > !' (> !

B > >>

> B B

!"#$# % & & > !' (> ! !' ::) + D) /

>B 0 !#12! ! !3 4) ! ' :: 6) +6D) 6 / *+
<) b4)& +)+ ¨)4 :) * ? +) :)

!"#$# % & & > !' (> ! !78 . / D)

>B>0 !#12! ! !3 4) ! 78 . /
D) *+
< *)4
3: A & + ===

!"#$# % & & > !' (> ! !D 1) ") . /
) + 1) & /

>B@0!#12! ! !3 4) ! D 1) ")
. /) + 1) & / *+
< @@ *)4
< 1) . /
3: A & + ===

!"#$# % & & > !' (> ! !% ; 9 . /

>B 0 !#12! ! !3 4) ! % ; 9 . / *+
< > B *)4
3: A & + ===

!"#$# % & & > !' (> ! ! / .))8/

>BB0 !#12! ! !3 4) ! > / .))8/ *+
< @ @ *)4

!"#$# % & & > !' (> ! ! : &)* D
* & #) / , . /,) + $ * 4/

>B 0 !#12! ! !3 4) ! @ : &)* D
* & #) / , . /,) + $ * 4/ *+
< >> *)4
3: A & + ===

!"#$# % & & > !' (> ! B

!"#$# % & & > !' (> ! B!#) /)) + %) #**)
))) 1) & /

> 0 !#12! ! !3 4) ! #) /)) + %)
#**)))) 1) & / *+
< *)4
3: A & + ===

> 0 !#12! ! !3 4) ! .)& / *+
< 4))) 1) & /, .)&)) + 4/

> 0 !#12! ! !3 4) ! $) *)

$. /, #) / , % * 4)4 ' ; + : +F&
< @ *)4

> 0 !#12! ! !3 4) ! $) *)
$. /, #) / , % * 4)4 ' ; + : *+

!"#$# % & & > !' (> ! B!H)) +)), $ - .)& /,
3: " , + ===

> 0 !#12! ! !3 4) ! B > B *+
< *))*

> >0 !#12! ! !3 4) ! B B 6 6 6 >B *+
< $ / /) + * /) * /) + ; 44 :

> @0 !#12! ! !3 4) ! B '))&) *+
< .) 4 4)& / +) /) & &

> 0 !#12! ! !3 4) ! B @ *+
<) 4) + 4

> B0 !#12! ! !3 4) ! B - .)& / *+

> 0 !#12! ! !3 4) ! B -&) &) *+
< A : 9N)+ 4/) + 4 4 J) * /

@ 0 !#12! ! !3 4) ! B
) +))6 6 64)& /6J*+ *+
< 3 4) 8

@ 0 !#12! ! !3 4) ! B 2) $ / " E & *+
< G / + +) 2) $ / .)& /_

@ 0 !#12! ! !3 4) ! B)) > >@ *+
< 3 4) $ 8 5.)& /0

!"#$# % & & > !' (> ! B!) 4 X >" $))
% 1) /

@ 0 !#12! ! !3 4) !
) 46X 6>"6 6$ 6 6) 6)6% 61) / * :

!"#$# % & & > !' (> ! B!$))) E & D
/)4)) (

@ 0 !#12! ! !3 4) ! $)))
E & D /)4)) (* 4 -)) : 46 * 4 5 0 *+
< *)4
3: A & + ===

C4 5 @0 $) 4) 8„ … \

@ >0 !#12! ! !3 4) ! $))) E & D
/)4)) (* :

!"#$# % & & > !' (> !

!"#$# % & & > !' (> ! !7 */) + " & / $ #8)
#**)

@ @0!#12! ! !3 4) ! 7 */) + " & / $
#8) #**) *+
< > *)4

")) 7 */
J A 4) 7 *
)
1ˆ / 7 *) + ; % :
* ; % :
>)) 9) ; % :) . & 3 +

!"#$# % & & > !' (> !

!"#$# % & & > !' (> ! !# ' - #) / . /

@ 0 !#12! ! !3 4) ! X & # '
#) / . / *+
3: A & + ===

!"#$# % & & > !' (> ! !# 3 ; 4))
& >

!"#$# % & & > !' (> ! !# 3))) #) /

@ B0 !#12! ! !3 4) ! R #
))) #) / *+
< @ *)4

!"#$# % & & > !' (> ! !# 3 $) /

!"#$# % & & > !' (> ! !#: -) "/) ')
$ / # + #+&) + +

@ 0 !#12! ! !3 4) !) 8 " , ; H #: -
) "/) 6 ') $ /) + #+&) + +) : +4 E & /
5 0 *+
< >B *)4

*))&)) * + *) @
G / * + *) _
% & 4 * + *) B
+ *) *))& :) >
") *) .)) * + *) >
> %) @
@ + *))): / @
78) * * + *)) : @

!"#$# % & & > !' (> ! !#' I 3 2))

@ 0 !#12! ! !3 4) ! > 1/+ (#' 2)
) *+
< B *)4

!"#$# % & & > !' (> ! !# - + $ $ $ / 3
. *

@ 0 !#12! ! !3 4) ! @ # 8) + # - +
$ $ / . * *+
3: A & + ===

!"#$# % & & > !' (> ! !# $ /) + %) /

@ 0 !#12! ! !3 4) ! X) +
# # +%) / *+
< @ *)4

!"#$# % & & > !' (> ! !') * 3 2))

@ 0 !#12! ! !3 4) ! B $)) & ') *
2)) *+

!"#$# % & & > !' (> ! !) 3))

@ 0 !#12! ! !3 4) ! .) +, D)
)) *+
< *)4

!"#$# % & & > !' (> ! ! 3 : -))
)

@ >0 !#12! ! !3 4) ! > R (, :
:))) *+

!"#$# % & & > !' (> ! ! :)) $ * 4/

@ @0 !#12! ! !3 4) ! > # ( ) + & :))
$ * 4/ *+

@ 0 !#12! ! !3 4) ! > # ( ) + & :))
$ * 4/ *+

@ B0 !#12! ! !3 4) ! > # ( ) + & :))
$ * 4/ *+

@ 0 !#12! ! !3 4) ! > /)4 D +)
:)) $ * 4/ *+

@ 0 !#12! ! !3 4) ! > $ / G E & / D%#
' "D
< .) 4 *) 3 /

!"#$# % & & > !' (> ! ! * ; / E &
7+ @

@ 0 !#12! ! !3 4) ! > # * / E*+) *+
< B *)4

@ 0 !#12! ! !3 4) ! >
* 6; /6 6 6E & 6 67+6 @6ER +F&
< B *)4

@ 0 !#12! ! !3 4) ! >
* 6; /6 6 6E & 6 67+6 @6ER *+

!"#$# % & & > !' (> ! ! * 8 % : - . /

@ 0 !#12! ! !3 4) ! > * 8 % :
. / *+
< > *)4

!"#$# % & & > !' (> ! ! & 8 D 4 # + / +)

@ >0 !#12! ! !3 4) ! > A/ (& 8 D 4) +
/ +) *+
< *)4
3: A & + ===

C4 5 0 A#%7 3% 7U D-.E17
$)+)))+/ & 8 4 5) /, & 8
* /4 0)) / 4 / G 4 &
) +

!"#$# % & & > !' (> ! !" A 4 , + 7+

@ @0 !#12! ! !3 4) ! >> " A 4 D
E)& :) : +4 E & / 5 0 *+
< @ *)4

@ 0 !#12! ! !3 4) ! >>
" 6 6A 4 6 +67+ *+
< B *)4

@ B0 !#12! ! !3 4) ! >> 1): / * /)

/ , %)) + + *+

!"#$# % & & > !' (> ! !" + X)/))

@ 0 !#12! ! !3 4) ! >@ X)(& R 4 " + X)/
)) : 5 B 0 *+
< @ *)4

!"#$# % & & > !' (> ! !7

@ 0 !#12! ! !3 4) ! > # D .)
1) & /5 + 7+ 0 *+

@ 0 !#12! ! !3 4) ! > #) * + *+

@ 0 !#12! ! !3 4) ! >
# 6#::) 67)+9)+ 6H) 6A 6 * / 6%)& 4) 6 6 *) 6:/6U)/6) +F&
< @ *)4

@ 0 !#12! ! !3 4) ! >
# 6#::) 67)+9)+ 6H) 6A 6 * / 6%)& 4) 6 6 *) 6:/6U)/6) *+

@ 0 !#12! ! !3 4) ! > # / 4/ *+

@ >0 !#12! ! !3 4) ! >
') (; $ + # + & + *) E +) + 4.)& / *+

@ @0 !#12! ! !3 4) ! > ((/ $ /
) *+

@ 0 !#12! ! !3 4) ! > ")&/+ & 2)
) *+
< @ *)4

@ B0 !#12! ! !3 4) ! > 78* + 4 *)
E +) + 4 * &)) + .)) 1)/ ' *+

@ 0 !#12! ! !3 4) ! >
D) 6; 9)* 6 *) 6$ 6#+&) 6 6# *) 6. +) 6%)& 4) 6) +6 +F
&

@ 0 !#12! ! !3 4) ! >
D) 6; 9)* 6 *) 6$ 6#+&) 6 6# *) 6. +) 6%)& 4) 6) +6 *+

@ 0 !#12! ! !3 4) ! > .))8/ D)) +
7& *+

@ 0 !#12! ! !3 4) ! > .)) & / *+
< B *)4

@ 0 !#12! ! !3 4) ! > - + +

4/ *+

@ 0 !#12! ! !3 4) ! >
- + 6 62) 6) *+
< B *)4

@ >0 !#12! ! !3 4) ! > R 9 & 7 *+

@ @0 !#12! ! !3 4) ! > /) +)
4/ *+

@ 0 !#12! ! !3 4) ! > I ;) +: (#))
R +4 +F&
< *)4

@ B0 !#12! ! !3 4) ! > I ;) +: (#))
R +4 *+

@ 0 !#12! ! !3 4) ! > 1) & /, .)&) ,) +
4/ # ') - + *+
< >> *)4

@> 0 !#12! ! !3 4) ! > 1 ( ) + *)) *
* ,)) + % " & * , $ + 7+ *+
< B *)4

@> 0 !#12! ! !3 4) ! > $) : +4 # *)
") / *+
< > *)4

@> 0 !#12! ! !3 4) ! > $ %)) A) E &
E +) + 4 D +))) *+
< *)4

@> 0 !#12! ! !3 4) ! > $ 1)+ .))8/ D) ,
+ 7+ *+

@> 0 !#12! ! !3 4) ! > $ (/) 7 I D *+

!"#$# % & & > !' (> ! !7) / * 3 $ * 4/

@>>0 !#12! ! !3 4) ! >B # 8) + 7) /
* $ * 4/ *+

!"#$# % & & > !' (> ! !7 3 #** +))

@>@0 !#12! ! !3 4) ! > T + & / (7
#** +)) *+
< @@@ *)4

!"#$# % & & > !' (> ! !D +) 3 3 7 $ / 3
+

@> 0 !#12! ! !3 4) ! @ X) : &, R)
D +) / + *+
< > *)4

!"#$# % & & > !' (> ! !D) (, $ 4 ,) +
/ /

@>B0!#12! ! !3 4) ! @
D) (6$ 4 6) +6 / / *+
< *)4
3: A & + ===

!"#$# % & & > !' (> ! !D +)) 3 /

@> 0 !#12! ! !3 4) ! @ -&) & D +)) /
B *+
< > *)4
3: A & + ===

<

) >
$ 1)+
;# $71 E%-$X 3D %#$E17 @
;) / %)) 3:F @
7) /) @
%
) +
) * ' + W
>)
@) .))8 E & >
D $/* D +)) -) @
' + / 3:F -) @
.)&)) -) @
7)4 -)
4 5%) 0 -)
> G)(-)
@ *)) & 7) -
/ # $/* -)
D +) +)
*)) + $
) *)) + $ %)
; 4 / *)) + $
D ' +) + -)
-) 1 D) $ 1) &
/ *
;# $71 7 ;#%- 3D # #$71-#A #1$- A7
+) , /, #)
.)) $)) >
#: %) ") +
$ - &) @
1) & %)) +

A) $ $))
#: %) #)
A))
))))) .)&)
) D +

> A) &)
@ A) 7 4/ &) #**)
) + E &) %) &) A)
@ A) 7 4/ &)
@ #**) &) A) @
@ E &) %) &) A)
# 4) >
E) /) ; 4 7 4/
) >
78* #)) + E
) / >
A 9 $)) >>
1) & 7 4/) + >B
1 1) &) /
@
;# $71 7A7 $13 #.%7$- D-7A" @
7) 4 @
+ D + - & 4) @
7J))) 4)
D + @
A) D + $)) @
A) 7)4 D + @@
% 3:F) + %)) @@
D D + 7J) 1) '
7 D +) + 7) 4 @
+ D + 7J) #:
)4) 4 @B
$ + D + 7J) 1) '
) + b 4¨)
8)4 D + @B
> D D + 7J) 1) '
)) / 4)4 D +) +
) 8 7 D +
@ #++) #) / $ + D +
7J) 1) ')) / 4
7 D +) +) 8)4
D +
)8 I D + 7J)
) 7 D +
D +))) /) 4
D +) 4 " : + &)
* , A))
7) 7 4/) 4 D +
)
D +) " *) 4 " *) +
" * " * -) B

>))4 D + B
>)4 D +) " B
>)4 D +)) B
>)4) + - 1)
)) 5# 4) 0 B

@) 4) D + B>
@)) 4)) E
7 D + B>
@)) 4)) E
)4 D + B@
@)) 4) : D + B@
D + & 4) 4 7
D +) E / & 4) 4
7 :/)) 4 & 4)
) >
7 :/)) 4 & 4 E /
) B
B 7)4 G)&
B * 1)+) D +
B $)& 4 G)&
B 7 7)4 G)& :/
3) 4) 4 7 4/) +
G)&
B D 3)) D +) + 4
G)&
*)4) A 4 @
- 7)4 G)& @
") 7)4 G)&
.) 3*
;# $71 #$3 - ;X - #%" 2E#%$E 7 ;#%
-
)) / + #
78* "))
$ E) / 1) >
:): / G)&
* 8 % : 7 I D) B
* 8 :): / G$)& $
* * *
A 4 $)))
)
> 7) #
> 7 4/) + - 2) 9)
> # 4)) + - 2)
9) B
> :): / # * +) + 2)
% :
@) / 7 #
@ *) 7
@ / -+)) 2)
)
@ # 2))

2) 9) # 1)+)
2) $) A *)
$ *) /
1

B 7 -) $
7 $ * 7 @
)) 2)) +
* *)
): 2) >
>
;# $71 > # 13 3 - '3"-7 # #..17.#$7
3D #1$- A7 $;71 #A ;7%3 7%# >>
> $ ') :)) / >>
>) * 2) D) >
>)) #) / .) + >
> * 78* >
> 1 & : / *
$) + - & : /)
* @
> 7 */ @
> > $ *) @
> @ 7J : " :)) ' +/ @
> $ +/) 1)
> B -+) .) @
> B)) + -) @
> B))) + 2) -+) .) @
> B 7J))) -+) .) B
> B ;))*) /) -+) .) B
> B > 1 & : $) B
>)) + $ +/) 1)+) BB
> /)
> /) A)
> $/* A) ' + >
>)) * /) @
> 7 7 4/ *) /)
> > A) ;))*) / @
> @ 7 .))
>) $) B
;# $71 @ # 13 3 - 3$-3% 3D 7"-# %3%
72E-A-'1-E 13 7 7 >
@ % J :) ' + >
@) * @
@ 7J) ;/+ +/)) -+)
A J + B
@) &) A) ;/+ +/
) B
@ 7J) ;/+ +/)
@ ;/+ +/) #) / :
D , ;) + ,) + "
@ /
@ D) A J + $ 4) $: >

@ ;) +

@ ;) $) ' $ G) B
@ > " ") +)
A J +
@ > R .)) + $
)))
@ > $ *) D)

@ >) $) "
@
@ > ")) 1) +)
B
@ > 1) ' R >
@ @ 1) +) A J + >
@ @ + + $ 1 / + %
: >
@ @ ")4) A >
@ @ ")4) ; 4 5 : 0 >
@ -): ;/+ +/) >
@ $) A)) $:
D >
@ ' +) / A)/ @
@ $: /) + $) "
& / @
@ $)) & &
;) $)) .)) @
@ B 3)) + G)&) A J + @@
@ B) D G)& @@
@ B G)&)) @
@ B A)) + %) G)& @
@ B) + 3 %) 7 @
@ B > ; 4 / : + +)
@ B @ 3))) 4 + " *) +
D ;)&/ %
@) * * : +)
@ .) 9 + D ' 7J)

@ * : /) +
) + / + >
@ D) $:)) / 4
@
@ A)&) % 99
@ (G)& B
@ *)4) :))
* : .) D B
@ .) 1)) (G)& B
@ (G)&) -+) .) B>
@ $:) ; 4 - / 78*
# * B
@ ;/+ +/)) & 7
@) & H

@ ' ::)*) A J + @

@ & 4 4 *)) + / +)
(G)&
@ $ 1 -): **
4)
@ > 7 4) (G)&
))) B
@)&)) A J +
@ ; 4 / 1) + .)
@) * 2) 7) A J +
@ > .) 9) ;/+ +/)
;# $71 7A7 $13 #.%7$- D-7A" -% 7"-#
7A7 $1- #A, #.%7$- , #%" 3 $- #A
13 71$-7 3D E' $#% 7
* + & /
7) + & /)
" >
" +)
7) + ; 78)
+
> 7 D +)
> D + D)) :)
> 7) D +)
> 7) D + -))
9)) :) >
@ # :)))4 D +
@ "))4 7
@)))4 3))4 9)
@ *))4 9) D)4
>
@)4 * * +
2) 9) A) 4))4
D 8 B
) 4) + 7)4
G)&) + 3*) * +)
D +) +) :)
- + + 7 D
")
7J)
.) 7)4 G)& >
> A) 3*) + /
*)4) 7)4 G)&
) + 1) + 1)
G)& >>
@ 1) & - + 8 -) " *
) + #: * A 4 @
1) & - + 8) (7
$) *) /) ;) + 1)+)
@
B %) 3* 7 @>
A) @

;# $71 B A# #
B .) 1) (
B 2) 7)) $ 4 %
$ 4)) ')
B 1) & 7))) "
%) D) + 7 4/ A :) +

B))) 7J)))
) $) 1)+))) B
B >)) R : / -) + -
1) " 7) + & /
)) B
B @)4 /+ +/)) +)) -):
$ ())(B>
B 3)) + G)&)) *)4)
1)+ G)& - * BB
;# $71 $7AA#1 #%" 17 $7AA#1 $#$7 3D
#$$71
))) E) 4 $ *)) +
"
)) .) *
))) + $ *)
)))
$ *))) +
$) 7 &)) @
7 4/ $)) @
) 7 4/
) / :)
7 4/
%) 1)
/
G ")
: 7&) $/*

" /) + 9 G ")
A 4) G ")
> * + %)
> 9 %)
> 1)) +)4 D + %
) >
> 1)+ 7 :/) @
> -) %) @
> > .)&)) 7 /
) %)
@ .)&)) + 1) & / ;
@ 7J &) *
@ . /) + $ %) 1
D)
@ 7 I 7J)

78*) E &
D +) I 4)

" & / b78*) Â E
& >
) " / @
B ; E & B
B " & / ') (4 + $) 1)
+) B
B) 4 # / + 7) /
E &
B) 4 " /) + $ *)
))
4) 78*)
B) #44 4)) 7) /)4
7& ; E &
D 7) >
+ 4 1) (
#** +
:F - + 8

!"#$# % & & > !' (> ! !.) /) # +
) /

@@ 0 !#12! ! !3 4) ! @ .) /)) +
) / *+
< *)4 A) +)
< .7%71#A ;X -
)) +) /

- #1$- A7 7 ;#%-
8
Ã $ * *) & /
Ã /
Ã >
Ã +) &
Ã>) B
Ã@ #)
Ã D
ÃB " * /) J) >
Ã) +
Ã G ( ) + *) 4/
Ã $) &) 4/
Ã -) 4/ @
Ã ' +)
Ã 7)
Ã > # 4) @
Ã @)) +
-- D-7A"

Ã 7))
Ã B 7 + @
Ã 7) *)

Ã .) N >
Ã 7 + *) >
Ã .)&)) + >@
Ã $ * * J &) @
Ã R *)
--- 3$-3% 3D # 1-.-" '3"X
@
Ã > $/ *) 4 + : +/ @@
Ã @ $ 4/) 4 + : +/ @B
Ã 1))) 4)
Ã B $ J))) 4 : +/
Ã 1)
Ã $ 4/ * B
Ã -) B

- $;7 $;731X 3D X 7$1X
Ã / / >
Ã B
Ã /))
Ã /) /
Ã *) 4 *
Ã > /))
Ã @ A))
Ã A) * +
Ã B /) *)
Ã $)) : +) /) /)

U-- E1D# 7 ;7%3 7%#
Ã) @
Ã > #+ *
Ã @ # 4) B
Ã)*) / B>
Ã B)* * &) & +) BB
Ã $) *) 4) + * 4 B
Ã +)

U--- 7 ;#%- #A 13 71$-7 3D 3A-"
Ã 78
Ã E * B
Ã)
Ã) / >
Ã > " /) B
Ã @ $) *) /
Ã D + @

!"#$# % & & > !' (> ! !.) 1) & /)

@@ 0 !#12! ! !3 4) ! @ : :64:6 F)#### ##H *+
< 1) & / $ *)) + .) $ /

@@ 0 !#12! ! !3 4) ! @ : :64:6 ; @#### ##H *+
< 1) & / $ *)) + .) $ /

@@ 0 !#12! ! !3 4) ! @ : :64:67 ####X##H *+
< .) # /

@@ 0 !#12! ! !3 4) ! @ .)) & / *+
< B *)4))

@@>0 !#12! ! !3 4) ! @ *) 6$ /6 61) & / :/
E4) & *+
< *)4 E4) &

@@@0 !#12! ! !3 4) ! @ *) 1) & / *+
< *)4 G +

!"#$# % & & > !' (> ! !.) 9 + D @

!"#$# % & & > !' (> ! !. / &) &)

@@ 0 !#12! ! !3 4) ! @> &) &) . / *+
< *)4 &) &)

/ : E + ' (
;# $71 '# - 17A#$-3% '7$G77% $;7 7A7 7%$ 3D
#% #1'-$1#1X $1-#%.A7 3A -%. 3'A-2E7 $1-#%.A7
$ A)
$ A)
78* 4 $) 4) ;) # 4 $
+) $) 4) + 1)+ - : +
5 - 0 @
& 4 3: J $) 4
>) 4 ") ' b-) : ¨ 3:
F
@ 3 : & 4 3: J $) 4
G (+ : B
:)*
;# $71 A3.- #A $1E $E17 3D $;7 3E1 7 3D 3A-"
.73 7$1X
B + . / %)
) + $ 4/ @
#8 ' 4 4 B
#8 ")
#8 3 +
#8) : / >
#8))
) D 4 #8 B
;# $71 #1#AA7A- -% # 7
> $ 1) & $) 4 A
*) $ 1) &))) +)) 4

A *) >
> @)))) 4 A) +)) >
))) >
B " *) $ # 4 ' $
A >>
)) F 4 >B
1 *) D 4 + . / @
G (+ : @
:)* @@
;# $71 $1#% D31 #$-3% 3D # 7 7 $31
$)) *) @B
$)) *) @
$ " +
> $
@ :) *))
:/) % : >
A) :) $ +
) /) + *)) / B
B)) + B
*))) + B
+ B>
:)*
;# $71 > 71 7%"- EA#1-$X -% # 7 "-;7"1#A #%"
3AX;7"1#A #%.A7
* +)))
- + A) + - F))
$ "))))
$ $ $ * +)
$ # 4 ') - + A) +))
> " * + '))) + *
+) /) 4 A) +) B
@ $ ") ' (A
$ $ * + $ $ $
*)) / $
B " +) # 4
$ # 4 ') * +))
@
3 4) F 4 B
$ A 4) F) A 4
$ #)) F)) /4
$ #)) F) # :) /)
D 4
/ +) # 4
G (+ : B
:)* >
@ ;# $71 @ $;7 331"-%#$7 7$;3"
> 1) 4) +) /
@ 78* 4))) + $
$ +) $ 7J)))
78* 4 + $
$ $ +) >
B 78* 4 $ * + $

$ $ +) B
G (+ : B
:)* @ >
;# $71 3AX;7"1# XA-%"71 , 3%7
> $ / + " +
> 1 4) / +)
> 7 N $
> $ >
> # / +)) $ / + >@
>> $ /) + >
>@ #)) $ @
> ; / *) @
>B *))) 5 /
) +0 D) 5 /) +0 @>
> / +) @
@ G (+ : @
:)*
;# $71 B $;7 '#AA
@ # *) +) ') " + # *) +
) ') :/)) B
@ # $) 4) B
@ $ *) *) $) 4 B
@ G (+ : I B>
:)* B B
;# $71 7# E1-%. 3AE 7
@> .) *
@@ $) 1) 4))) * * +
@ $) 1 4 / +) +
@B $) 3: J / +) + >
@ $.) D) * 4
) D 4 E 4 #) @
$ D) * 4) ,
) ')) + -) * N D)
G (+ :
:)*
;# $71 $;7 #17# 3 E1D# 7
$ #))) / + >
$ #)) # :) /) B
$ #))) 1 4)
/ + ,)) ,) +) ')
> G (+ :
:)* @
#% G71
-%"7U @

!"#$# % & & > !' (> ! !.)&)) G)& - 7 I $ /

@@B0#12! ! !3 4) ! @@ D ($ $ / *) $
) + .)&) *+
< @ *)4

@@ 0!#12! ! !3 4) ! @@ D ($ $ / *) , $

) + .)&) *+
< @ *)4

@ 0 !#12! ! !3 4) ! @@ A 4 & & & $
1) & $ / .)&) *+

@ 0 !#12! ! !3 4) ! @@ T)( ) & .)&)) G)&
7 I $ / *+

!"#$# % & & > !' (> ! !; 4)) $ 8 : (
$)

@ 0 !#12! ! !3 4) ! @ & & ; 4)) *+
< *)4

3 % $ 7 % $
"# - #%#AX$- .73 7$1X -% $; 7 A#%7
)* - + +)
$ +))
$ $ " + 4
$ ") ' $
" & + 4) 4) . & 1)
> $ # 4)) 4 A #8 B
@ +))) + * +) /
)* $) 4 A
) 4 A)) A
B 7J))) 4 A 5 * - * D 0
.) D 7J))) 4 A) + -
*)) >
7J))) 4 A 5- * D 0
& + 78) *
)) 4 A G - 7J) - . &
$ - $) 4 A
7J))) 4 A) 4 $ 4
58 , / e0) . & "
> 7J))) 4 A) 4 $ 4 $. &
58Ä /,0) + 58 , / 0
@ $ # 4 ' $) 4 A >
)* 2)+ &
7J)
B & + 7 8) *
@ $)) 2)+ &
7 *
$ 7J)) 7 * >
- & 4) 4 D) 7 * - 7J) @
4) 7 * B

1) * ' 7 *) +
> 7 /) 7 * >
;/* :) >
$ 7J) ; /* :) >
B - & 4) 4 D ;/* :) - 7J) >

4) ; /* :) >
/ * ; /* :) >@
7 /) ; /* :) >B
7J)) ; /* :) >
& + 78) * 7 *) + ; /* :) >
)):) @
> 7J)))):) @
@ - & 4) 4 D)):) - 7J) @
4))):)
B D) $)) +) @
7J))):))) $))
+) # 8 @@
7J)) 7J)) ;/* :) 1 +
/ * @
& + 7 8) * @B
2)+ &)

!"#$# % & & > !' (> ! !- + - . / 5 A
)) A :) /0

@ 0 !#12! ! !3 4) ! @B - + . / 5A
)) A :) /0 : 5 0 *+

!"#$# % & & > !' (> ! !- + $ # *) 7 4 4
G # D 4 $ * &

@ 0 !#12! ! !3 4) ! @ - + # *)
7 4 4) D 4 $ * & *+
< B *)4

!"#$# % & & > !' (> ! !- + $ 2))

@ >0 !#12! ! !3 4) ! - + 2)
) *+

!"#$# % & & > !' (> ! !- + $ 2)) :/
")& + H .

@ @0 !#12! ! !3 4) !
- + 6 62) 6) *+
< B *)4

!"#$# % & & > !' (> ! !- + / 1) #) /

@ 0 !#12! ! !3 4) ! R 4 &, D - + /
1) #) / *+
< *)4
3: A & + ===

< *)

!"#$# % & & > !' (> ! !A - . / -

) / . /

@ B0 !#12! ! !3 4) ! (& A . /
- #) / . / *+
< > *)4
3: A & + ===

!"#$# % & & > !' (> ! !A) # 4 :) # + ")
. /

@ 0 !#12! ! !3 4) ! " & / / G +
*+
< >@ *)4

@B 0 !#12! ! !3 4) ! 7 &, 1 9 + A)
4 :)) + ") . / + *+
< > *)4
3: A & + ===

!"#$# % & & > !' (> ! !A :) & ( ) . / 5A
)) A :) /0

@B 0 !#12! ! !3 4) ! > 4 9 & (/ A :) & ( )
. / A)) A :) / *+

!"#$# % & & > !' (> ! !))) #) / ")
+ - 4)

@B 0 !#12! ! !3 4) ! @))) #) /
")) + - 4) *+

!"#$# % & & > !' (> ! !) A : 4 - 4) # + ; :
*)

@B 0 !#12! ! !3 4) !) A : 4 - 4)) +
; : *) *+
< R 4 &

!"#$# % & & > !' (> ! !)(- . 5 $ * -
)) 0

@B 0 !#12! ! !3 4) ! B " : &)(.
5$ *)) 0 *+

@B>0 !#12! ! !3 4) ! B R 4) $ #**)
) . / 5 *) A)) 0 *+

@B@0!#12! ! !3 4) ! B 4 9 & ($ 1
.) 5 *) A)) >0 *+

!"#$# % & & > !' (> ! ! + *) 1) & /

@B 0 !#12! ! !3 4) !
1) (6' (6 + *) 1) & / *+
< @@ *)4

!"#$# % & & > !' (> ! !) . / "

@BB0 !#12! ! !3 4) ! B
#6 6 6*) 64 /6 6)+&) + +F&

@B 0 !#12! ! !3 4) ! B
#6 6 6*) 64 /6 6)+&) + *+

@ 0 !#12! ! !3 4) ! B B> +F&

@ 0 !#12! ! !3 4) ! B B> *+
< # *) 4 /)+&) + +) --

!"#$# % & & > !' (> ! ! * E* . / $))
' + * E*) +

@ 0 !#12! ! !3 4) ! B *
E*6. /6$ 6)) 6' +6 * E*6) + *+
< *)4
3: A & + ===

!"#$# % & & > !' (> ! ! : - . /

@ 0 !#12! ! !3 4) ! B + & : - . /
B *+
< *)4
3: A & + ===

!"#$# % & & > !' (> ! ! : -) . / 5
D 7& / 0

@ 0 !#12! ! !3 4) ! B - D) /4 :)
. / 5 7& / 0 : 5 BB0 *+
< *)4
3: A & + ===

!"#$# % & & > !' (> ! ! - . / 5 A))
A :) /0

@ >0 !#12! ! !3 4) ! B D & . /
5A)) A :) /0 *+

!"#$# % & & > !' (> ! !2) 7 +/)

@ @0 !#12! ! !3 4) ! B> #+&) + E + 4)+) 2)
) +) + #**) *+
< @ *)4

@ 0 !#12! ! !3 4) ! B> #(9 , ' (2)
7 +/) *+
< > *)4

@ B0 !#12! ! !3 4) ! B> . +I) , R & (:
2)) *+
< B *)4

!"#$# % & & > !' (> ! !2))

@ 0 !#12! ! !3 4) ! B@ ")&/+ & 2)
) *+
< @ *)4

!"#$# % & & > !' (> ! !2)) (&

0 !#12! ! !3 4) ! B (& 2)) *+
< >@@ *)4

!"#$# % & & > !' (> ! !1) (): & 5A))
A :) /0

0 !#12! ! !3 4) ! BB) (): & *+

!"#$# % & & > !' (> ! !) / # + ") + -
)

0 !#12! ! !3 4) ! B + &) /) +
") +) *+
< B *)4

!"#$# % & & > !' (> ! !) /

0 !#12! ! !3 4) ! . & , $)4
.))8) +) B *+

0 !#12! ! !3 4) ! & I/) &
) : # - + / # / *+

>0 !#12! ! !3 4) ! & I/) & 7)/
): E & B> *+
< @ *)4

@0 !#12! ! !3 4) ! /)9) /) / $
) / *+
< *)4
3: A & + ===

!"#$# % & & > !' (> ! ! *) # + $ - $ +

0 !#12! ! !3 4) ! ' (& *)) + $
+ *+

< *)4

B0 !#12! ! !3 4) ! " - ' (& *)) + $
+ *+

C4 5 0 ;# $71
.73 7$1- #A 7# E17 7%$ -%
$;7 # 13G31A"
$;7 #1-$; 7$-T#$-3% 3D # 7 $H 7

<

$#'A7 3D 3%$7%$
17D# 7 --
-%$13"E $-3% U---
- .73 7$1- #A 7# E17 7%$ -% $;7 # 13G31A"
$ # 9) *) $
$ /) + # 9) *)
$-Å
3 " & + 4) + 7& *)) + $
$ # -) +
> $ 1)) +
@ $ / # 9) *) $
) + B
9) 7&) % A)
$ / D +
B $.) $ / 1) & /) + # 9)
*) $
4 /
-- .73 7$1- #A 7# E17 7%$ -% $;7 - 13G31A"
1) ( ) +
$) +) *) - @
$)) 4 +) *)
- >
- +)) *) - +)
)) . & -) $ @
--- .73 7$1- #A 7# E17 7%$ -% $;7 - 13G31A"
-% $;7 17A#$- - $- # 7 @
$ D D + @
> $ E) / 1) D

@ $ ' D +
$ A) 9) B
B $ " 1) &) (B@
$ +) %) + G 4 B
$)) *) I +)
1) &)
- $;7 13 A7 3D D-%-$7 "- 7% -3% 3D 7A7 7%$#1 X
#1$- A7 >
$) 9)) $ ")
7 >
$ 7)4 %

$ % B
$) 2) 9 + D + $ /

!"#$# % & & > !' (> ! ! 4)* F 5 A
)) A :) /0

0 !#12! ! !3 4) ! 4)* * F 5A
)) :) /0 ' # 1 +, % " 4 &) *+

!"#$# % & & > !' (> ! ! * 4 $ / # + ' / +

0 !#12! ! !3 4) !
(6H 6 46 / 6 6 6# 6 + 6 6 6: 6 4 *+
< *)4

0 !#12! ! !3 4) !
(6H 6 46 / 6 6 6 * 46 /6) +6: / + *+
< >> *)4

!"#$# % & & > !' (> ! !$ A) 4)4 3 *) # + D

0 !#12! ! !3 4) ! # 4 :) . / #
- + *+
< ")

0 !#12! ! !3 4) ! * 8 . / #
- + *+

0 !#12! ! !3 4) ! @ +F&
<) . /

>0 !#12! ! !3 4) ! @ *+
<) . /
3: A & + ===

)* - 1) D 4

- +
%)) + A
4) + %)
$) 4) + %)
4 . D 4
$ 4 $) 4
. D 4
$) + ")
- J)) $) 4
$)) A
#**) $))
3 $ $) 4

") A
„$ 2)+))
/4
/ /
+ #) (
) /)* -
:) + #**)

)* --) 4 A) +

$
) # 4
:) + #**)
4 + :/ $) 4 , + ,) +)
) /)* --
:) + #**)

)* ---)) 4 A 4

#** 8))
1) A 4
$ *) 4
) /4
*) :/) # 4 @
:
) /)* --- >
:) + #**)

)* - #) /4

) 1) 4 >
) /4 >@
: @@
) /)* - @
:) + #**)

)* 1 4) /4) +

1 4) /4 @
:) + #**) B
) BB
) /)*
:) + #**) >
)* -)): .)4 + Æ
.)* 1 *)
" * +)): >
A))):
)* -- 1 &) + D #**)
3 ") +
A +) @
:) + #**) @
D ")) 4 $) 4

:) + #**)
$ -):) B
:) + #**) >
D #**) * >
D * $) 4 >
:) + #**) @
)8)) +) @B
- + 8

@0 !#12! ! !3 4) ! . / -- *+
< @ *)4 ' 4

0 !#12! ! !3 4) ! . / *+
< @>*)4 # +

B0 !#12! ! !3 4) ! * . /
* + *+
< @ *)4

0 !#12! ! !3 4) ! % . / *+
< *)4 1

0 !#12! ! !3 4) ! *) 4 /) (+F&

0 !#12! ! !3 4) ! *) 4 /) (*+

0 !#12! ! !3 4) ! *) 4 /) (6: *+

0 !#12! ! !3 4) ! *) +4) +F&

0 !#12! ! !3 4) ! *) +4) *+

>0 !#12! ! !3 4) ! *) +4) 6: *+
< > *)4 ;)

@0 !#12! ! !3 4) ! *) +4) +F&

0 !#12! ! !3 4) ! *) +4) *+

B0 !#12! ! !3 4) !
*) +4) 6: *+
< ;)

0 !#12! ! !3 4) ! 4) 4)
; 4/, /*) +))) / *+
< *)4

0 !#12! ! !3 4) ! $ A) 4)4 *)) +
D *+
< *)4
3: A & + ===

!"#$# % & & > !' (> ! !$)(4))

0 !#12! ! !3 4) ! >
6' (6$)(43)) *+
< > *)4
< ; * /))
3: A & + ===

!"#$# % & & > !' (> ! !$ + 3 $ 4 -
$ $ / 3 % :

0 !#12! ! !3 4) ! @ 4)+ & $ +
$ 4 $ / % : *+
< BB *)4

!"#$# % & & > !' (> ! !$ * 3 7 +/)
) & &

0 !#12! ! !3 4) !) & & $ *
7 +/) *+
< *)4

!"#$# % & & > !' (> ! !$ A))) :
5A)) A :) /0

0 !#12! ! !3 4) ! B A/ ($ A
))) : 5A)) A :) /0 *+

!"#$# % & & > !' (> ! !$ $ / 3 D 3 # * 8
)):

>0 !#12! ! !3 4) ! & (& $ (& $
$ / 3 D 3) * 8)): *+
< *)4
3: A & + ===

!"#$# % & & > !' (> ! !$ E 1 3 / *)+ : ' (

@0 !#12! ! !3 4) ! ( ) (/, 9 &, X)4 $
E 1 3 / *)+ : ' (*+
< B *)4

!"#$# % & & > !' (> ! !$) /

0 !#12! ! !3 4) ! # R *) / $)
/ D 4 A) 4)4 : 4 ; @ *+
< @ *)4

!"#$# % & & > !' (> ! !E)) A)

B0 !#12! ! !3 4) !
R 6' (6E)) A) *+

< B *)4

0 !#12! ! !3 4) ! E)) A) $ 4/) +
#**) 1 5 0 *+
< *)4

0 !#12! ! !3 4) ! E)) A) ') +
2) " 6 /) + " & G / ; 5 0 *+
< > @ *)4

!"#$# % & & > !' (> ! !E & + # + E &): : -
. /

0 !#12! ! !3 4) ! # R * &, # 1 :) &
: . / 5 B, : 0 *+
< *)4

0 !#12! ! !3 4) ! . & A & (+(&
& 4 : - . / BB *+
< B *)4

0 !#12! ! !3 4) ! R & (.)
E 4 *) 3 / 5 *) A)) 0 *+
< *)4

0 !#12! ! !3 4) ! R & (.)
G *) 3 / 5A)) A :) /0 *+

>0 !#12! ! !3 4) ! ((E & +) +
E &): : . / *+
< B> *)4

!"#$# % & & > !' (> ! !G) - ") 5 *) A
)) 0

@0 !#12! ! !3 4) ! + G) ") *+
< *)4

>

>

!"#$# % & & > !' (> !

!"#$# % & & > !' (> ! !# 3 : 3
))) /

0 !#12! ! !3 4) ! > # :
))) / @ *+
< *)4

B0 !#12! ! !3 4) ! >)+ & 7J)

))) * / *+
< *)4

!"#$# % & & > !' (> ! !#)) + A 4) 5A
)) A :) /0

0 !#12! ! !3 4) ! @) (& #)) +
A 4) A A *+

!"#$# % & & > !' (> ! ! * 8 % : - . /

> 0 !#12! ! !3 4) ! * 8 % :
. / *+
< > *)4

!"#$# % & & > !' (> ! !")& + ; :

> 0 !#12! ! !3 4) ! B 4)):) + (*+
< .7 # 7A$7 #' %"AE%.7% 3% ;71 #%% -%R3G R-
< ")& + ; :
< B *)4

> 0 !#12! ! !3 4) ! B 4)) B @ *+
< '71%;#7" 1-7 #%%I
.7 # 7A$7
#$;7 #$- ;7 G71-&7
E%"
G- 7% ;#D$A- ;71 %# ;A#
< > @ *)4

> 0 !#12! ! !3 4) ! B 4)) *+
'71%;#1" 1-7 #%%I
.7 # 7A$7
#$;7 #$- ;7 G71R7
E%"
G- 7% ;#D$A- ;71 %# ;A#
< *)4

!"#$# % & & > !' (> ! !" & * ((
4 / :

> 0 !#12! ! !3 4) ! *+
< . / + A 9 ((*)

>>0 !#12! ! !3 4) ! 4 + 9) (*+
< .73 7$1-7 "71 T#;A7%
3%
;71 #%% -%R3G R-
< B *)4

>@0 !#12! ! !3 4) ! 4 + 9) (*+

> 0 !#12! ! !3 4) ! ;) ;) (" & *
((. / % : " & :) 5 @ 0 +F&

>B0 !#12! ! !3 4) ! ;) ;) (" & *
((. / % : " & :) 5 @ 0 *+
< *)4

> 0 !#12! ! !3 4) ! ;) ;) (" & *
((. / % : " & :) 5 @ 0 +F&

@ 0 !#12! ! !3 4) ! ;) ;) (" & *
((. / % : " & :) 5 @ 0 *+
< @ *)4

!"#$# % & & > !' (> ! !7 3 . /

@ 0 !#12! ! !3 4) !) + /4 + *+
< #1 ;- 7"7B,
;aX.7% , A# '71$, A7.7%"17

@ 0 !#12! ! !3 4) ! @@ *+
< 7 . / A 4 +
< B *)4

@ 0 !#12! ! !3 4) ! > *+
< $): A 4) * D
$ & D 4

@ 0 !#12! ! !3 4) ! 4 4 *+
< 7 . /) + $ 4 / A 4 +

@>0 !#12! ! !3 4) ! 8) 4 *+
< 7U71 - 7
"7
#A EA -%$†.1#A
E1
"- 71 31"17 "7 $1#% 7%"#%$7
7$ E1 A7 2E#"1#$E17 K
) # A7.7%"17
< > *)4

@@0!#12! ! !3 4) ! 8) 4 *+
< > *)4

@ 0 !#12! ! !3 4) ! 8 + 4 *+
< > @ *)4

@B0 !#12! ! !3 4) ! 4 +)& *+
< 7A7 7%$ 3D .73 7$1X #%" $1-.3%3 7$1X,
D13 $;7 G3ER 3D # A7.7%"17
< @@ *)4

@ 0 !#12! ! !3 4) ! 4 / : *+
< .73 7$1X
$77 7A7 7%$ 3D 7E A-" #%" A7.7%"17
- A-D-7" #%" #11#%.7"
$3 7U AE"7 D13 .73 7$1- #A 17# 3%-%.
< @ *)4

0 !#12! ! !3 4) ! + 4 4 *+
< †A† 7%$ "7 .†3 †$1-7 A 4 +
< > *)4

0 !#12! ! !3 4) ! +* * 4 *+
< 7A7 $7" 13 3 -$-3% -% .73 7$1- #A 3% $1E $-3%
#%" # A- #$-3% 3D #A.7'1# $3 .73 7$1X
'7-%. # R7X $3 $;7 # 7%"-U 3D "# -7 I A7.7%"17
< B *)4

!"#$# % & & > !' (> ! !7 +) 4 /) + & 8 /

0 !#12! ! !3 4) ! # 4 :)) + . /
- + A) # 4 :) V 7 4 4) 5 0 *+
< @@ *)4

0 !#12! ! !3 4) !)) 4 / 7 +) ,
))) , & & ,) + * F & G / 5 0 +F&
< *)4

0 !#12! ! !3 4) !)) . / 7 +) ,
$))) , - & & ,) + F & G / 5 0 *+

>0 !#12! ! !3 4) !)) . / 7 +) ,
$))) , - & & ,) + F & G / 5 0 *+
<)) # *) /

@0 !#12! ! !3 4) !)) . / 7 +) ,
$))) , - & & ,) + F & G / 5 0 *+

0 !#12! ! !3 4) ! 4 . / #
- + + . / $) 4) + 5 0 *+
< *)4

B0 !#12! ! !3 4) !) . /) + %)
") 7J)) : +4 5 0 +F&
< > *)4

0 !#12! ! !3 4) !) . /) + %)
") 7J)) : +4 5 0 *+

B 0 !#12! ! !3 4) ! A) # 4 :)) +
. / #)))) / 5 0 *+
< > > *)4

B 0 !#12! ! !3 4) ! D) ( ) + D / - & &
. / .) + *) / 5 0 +F&
< B *)4

B 0 !#12! ! !3 4) ! D) ( ) + D / - & &
. / .) + *) / 5 0 *+

B 0 !#12! ! !3 4) ! - & & . / . ') +
, A + 5 0 *+

B 0 !#12! ! !3 4) ! H . # $ /
& 8) + A) . / 5 0 *+
< >> *)4

B>0 !#12! ! !3 4) ! H) " ' 4 # 4 :)) /
4)))) A) # 4 :)) + #) / . / *) 5 @0 *+
< @> *)4

B@0 !#12! ! !3 4) ! R) $ A 4 A)) 4 :)
) + 4 / ;RE 5 0 +F&
< *)4

B 0 !#12! ! !3 4) ! R) $ A 4 A)) 4 :)
) + 4 / ;RE 5 0 *+

BB0 !#12! ! !3 4) ! A) # 4 :) $ 4 . /
5 0 *+

B 0 !#12! ! !3 4) ! %)) # 4
. /6 # - + + . / $) 4) + 5" & ' (
)) 0 " & :) +F&

0 !#12! ! !3 4) ! %)) # 4
. /6 # - + + . / $) 4) + " &
:) 5 0 *+

0 !#12! ! !3 4) !)) A) # 4 :) #
. / $: 8 5 0 *+
< ") I " & /
< > *)4
3: A & + ===

0 !#12! ! !3 4) !)))) 4 :))
4 / : 8 1 5 0 *+
< > @ *)4
3: A & + ===

0 !#12! ! !3 4) !) + #** +))
5 > 0 *+
< A) # 4 :)) + F & . /
< *)4

0 !#12! ! !3 4) ! 1 + ') A)) 4 :)
) + * F & 4 / #)+ 5 > 0 +F&

>0 !#12! ! !3 4) ! 1 + ') A)) 4 :)
) + * F & 4 / #)+ 5 > 0 *+

@0 !#12! ! !3 4) ! 1 ' 7 +)
. / V & 8 / .) 5 @@0 +F&

0 !#12! ! !3 4) ! 1 ' 7 +)
. / V & 8 / .) 5 @@0 *+

B0 !#12! ! !3 4) ! 1 ' 7 +)
4 /) + & 8 / .) ; 5 @@0 *+

!"#$# % & & > !' (> ! !;) +: (* 8 #) /

0 !#12! ! !3 4) !
;) +: (6 6 * 86#) / *+

!"#$# % & & > !' (> ! !; 4) ' 4

B 0 !#12! ! !3 4) ! T + & , X)4 ; 4)
' 4 *+
< >@ *)4

!"#$# % & & > !' (> ! !; / / :
" (

B 0 !#12! ! !3 4) ! / / + (*+
< ;- $31X 3D $;7 $;731X 3D %E '71
< > @ *)4

B 0 !#12! ! !3 4) ! / / + (*+
< " *) #) /
< B B *)4

B 0 !#12! ! !3 4) ! / / + (*+
< 2)+) + ; 4 D
< *)4

!"#$# % & & > !' (> ! !- + / 1) #) /
R 4 &

B 0 !#12! ! !3 4) ! > R 4 &, D - + /
1) #) / *+
< *)4

!"#$# % & & > !' (> ! !A - . /

B >0 !#12! ! !3 4) ! @ (& A . /

- #) / . / *+
< > *)4

B @0 !#12! ! !3 4) ! @ (& A . /
-- A) # 4 :)) + ") . / *+
< @ *)4

!"#$# % & & > !' (> ! !A) # 4 :) G 7 3
) / . /

B 0 !#12! ! !3 4) ! # + & (&, .
$ * &) A) # 4 :) 7 #) / . / *+

!"#$# % & & > !' (> ! !)8 J) 1)) *)

B B0 !#12! ! !3 4) ! B > *+
<)8 J) 1)) *) ,
4 /)) J) +)

!"#$# % & & > !' (> ! !)(- .

B 0 !#12! ! !3 4) ! " : &)(.
5$ *)) 0 *+

!"#$# % & & > !' (> ! ! / 3 $ +

B 0 !#12! ! !3 4) ! (/
+ *+
< @ *)4

!"#$# % & & > !' (> ! !

B 0 !#12! ! !3 4) ! # E + 4)+)
4 :) . / 5 0 *+

B 0 !#12! ! !3 4) ! ' ') + : $))
4 /) G ; D)) + *) / 5 B 0 +F&
< *)4

B 0 !#12! ! !3 4) ! ' ') + : $))
4 /) G ; D)) + *) / 5 B 0 *+

B 0 !#12! ! !3 4) ! . / %) 5 0 +F&
< > *)4

B >0 !#12! ! !3 4) ! . / %) 5 0 *+

B @0 !#12! ! !3 4) ! . 4 .) . /) +
#**) # , %)) + $ 4/ * 4 5 0 *+
< B *)4
3: A & + ===

B 0 !#12! ! !3 4) ! HÇ 4 H 1))
. /) + . #) / * 4 %) 5 0 *+
< *)4
3: A & + ===

B B0 !#12! ! !3 4) !) ") /)4) I
A 4) /6)) $.) -+) 38 + E & / , E # 5 0 *+
< B *)4

B 0 !#12! ! !3 4) ! %) A . /6 D
- *) 1) & / * 4 %) 5 0 *+
< @ *)4

B 0 !#12! ! !3 4) ! 5§0 %) , # ,) +
. /)) ; *+
< *)4

B 0 !#12! ! !3 4) !) & ") . /
* 4 %) 9) + #. 5 0 *+
< B *)4

B 0 !#12! ! !3 4) ! $ %) $. /,
/) + * * 4 %) + 5 0 *+
< *)4

B 0 !#12! ! !3 4) ! $ * 4) , ")) +
) . /) 5 0 *+
< B *)4

!"#$# % & & > !' (> ! ! : - . /

B 0 !#12! ! !3 4) ! + & : - . /
B *+
< *)4

!"#$# % & & > !' (> ! ! : - 2))

B >0 !#12! ! !3 4) ! . +I) , R & (:
2)) *+
< B *)4

!"#$# % & & > !' (> ! !2)+) +

B @0 !#12! ! !3 4) !):)(4 &) *+
< #:)(.
< *)4

B 0 !#12! ! !3 4) ! +) 4 : *+
< D +) . / ")& + ; :
< *)4

B B0 !#12! ! !3 4) ! 4 +)4 + 4 : *+

< .1E%"A#.7% "71 .73 7$1-7
3%
" "#-" ;-A'71$
< B *)4

B 0 !#12! ! !3 4) ! + *+
< ;- $3-17
"7 17 ;71 ;7
E; A#
2E#"1#$E17
"E 71 A7
< @ *)4

B 0 !#12! ! !3 4) ! + *+

B 0 !#12! ! !3 4) !)) 4 /) *+

B 0 !#12! ! !3 4) !)** *+
< 3AE$-3% # 13 ;†7
"7 A#
$1- 7 $-3% "7 AI#%.A7
7$ "7 A#
2E#"1#$E17 "E 71 A7

!"#$# % & & > !' (> ! !2) 7 +/)

B 0 !#12! ! !3 4) ! > #+&) + E + 4)+) 2)
) +) + #**) *+
< @ *)4

B 0 !#12! ! !3 4) ! > #(9 , ' (2)
7 +/) *+
< > *)4

B >0 !#12! ! !3 4) ! > . +I) , R & (:
2)) *+
< B *)4

B @0 !#12! ! !3 4) ! > R 9 &, 3 IX) & (
*) : 7) /) / *+
< *)4

B 0 !#12! ! !3 4) ! > A 57+ 0 2) $ /
+ 5#+&)) + $ 4/ E 1 / 0 B *+
< *)4

B B0 !#12! ! !3 4) ! > 3 /) & (/ ")
+ / 4 *+
< BB *)4

B 0 !#12! ! !3 4) ! > (& 2)) *+
< >@@ *)4

!"#$# % & & > !' (> ! !2) D + $ / # ")4))
#**) :/ 1) + R

B 0 !#12! ! !3 4) ! @ 2) D + $ / *+
< >@ *)4

!"#$# % & & > !' (> ! !2))

B 0 !#12! ! !3 4) ! ")&/+ & 2)
) *+
< @ *)4

!"#$# % & & > !' (> ! !1)) . /

B 0 !#12! ! !3 4) ! B > *+
< 1)) 4 /
< @ *)4

!"#$# % & & > !' (> ! ! *)

B 0 !#12! ! !3 4) ! .) +, & .) 9 +
D *) D +))) + .) 9 + D *+
< *)4

!"#$# % & & > !' (> ! !) 4 A) + &

B 0 !#12! ! !3 4) !) 4) + & *+
< @ *)4

!"#$# % & & > !' (> ! !$ ") $): 3 $ A 4)
3 * 8 % :

B >0 !#12! ! !3 4) ! A/ ($ +) $):
A 4) * 8 % :) + $)))))
+) *+

!"#$# % & & > !' (> ! !$ $ / 3 *) $ # +
.)&)

B @0 !#12! ! !3 4) ! D ($ $ / *) $
) + .)&) *+
< @ *)4

!"#$# % & & > !' (> ! !$ * - $ / 3) - 7 *
*)

B 0 !#12! ! !3 4) ! $ * $ /)
7 * *) 4 & *+
< *)4

!"#$# % & & > !' (> ! !E & + # + E &): : -
. /

B B0 !#12! ! !3 4) ! ((E & +) +
E &): : . / *+
< B> *)4

!"#$# % & & > !' (> ! ! # + $ #) / G
#**)

B 0 !#12! ! !3 4) ! > ' (, $))* &) +
$ #) / G #**) *+
< @@ *)4

> B

>

!"#$# % & & > !' (> !

!"#$# % & & > !' (> ! ! : / --$ H77

B> 0 !#12! ! !3 4) ! @ : / - --$ H77
--$H77))+&) +) +) + U-- *+
< *)4

B> 0 !#12! ! !3 4) ! @ : / -- --$ H77
--$H77))+&) +) +) + U-- *+
< >@@ *)4

!"#$# % & & > !' (> ! ! * * *)

B> 0 !#12! ! !3 4) !
* * 1) & /6) 6 *+

B> 0 !#12! ! !3 4) !
* * 1) & /6) 6 *+

B> 0 !#12! ! !3 4) !
* * 1) & /6) 6 *+

!"#$# % & & > !' (> ! !$ 7 +) . /
A /N +) . /

B>>0 !#12! ! !3 4) ! B A A / $ 7
+) . /))) + 5 B >0 +F&

B>@0 !#12! ! !3 4) ! B A A / $ 7 3
+) . / *+
< > *)4

B> 0 !#12! ! !3 4) ! B $ 7 +)
. / A /N +) . / ' A T$.U *+
< *)4

!"#$# % & & > !' (> ! !E & / / /)) +
A

B>B0 !#12! ! !3 4) ! * 1 R , ")& + A $) (
E & / / /)) + A G ; D)) + 5 0 *+
< *)4

B> 0 !#12! ! !3 4) ! * 1 R , ")& + A $) (
E & / / /)) + A G ; D) 5 0 *+
< @ *)4

B@ 0 !#12! ! !3 4) ! * 1 R) + ")& + A
$) (- I)) E & / / /)) + A
G ; D) 5 0 *+
< > *)4

!"#$# % & & > !' (> !

!"#$# % & & > !' (> ! !#**) .) 1) & / G
:

B@ 0 !#12! ! !3 4) !
6' (6#**) 3 .) 1) & *+

!"#$# % & & > !' (> ! !) . / $ * 4) . *

B@ 0 !#12! ! !3 4) !) . / $ * 4)
. * *+
< B *)4

!"#$# % & & > !' (> ! !D +)) A 4 " 4 ,
7) + 7+

B@ 0 !#12! ! !3 4) !
D +)) 6 6A 4 6" 4 67) +67+ *+
< B B *)4

!"#$# % & & > !' (> ! !X)) 4 . +

B@ 0 !#12! ! !3 4) ! X) *+

!"#$# % & & > !' (> !

!"#$# % & & > !' (> ! !#:) # 4 :) # - 4) +
#**)

B@>0 !#12! ! !3 4) !
#:) 6# 4 :)6# 6- 4) +6#**) 6 6 &) 6H * 6; *+
< >B *)4

!"#$# % & & > !' (> ! !D %) "/)
$ 4 /N)4 # " A 9 7J)

B@@J!#12! ! !3 4) ! >
D 6%) 6"/) 6 6$ 4 /I 6)4 6#6" 6 6 6 6A 967J)
*+
< *)4

!"#$# % & & > !' (> ! !. /

B@ 0 !#12! ! !3 4) ! @ 7) G ((, #
H / # 4 :) / 4) 4)))) 7+ $ -))
5 @0 *+
< @ *)4

!"#$# % & & > !' (> ! !A) A 4) 4 #) +
A)

B@B0 !#12! ! !3 4) !
A) 6A 4 6) 46 6# 6) +6A) 6 *+
< *)4

!"#$# % & & > !' (> ! !))) + , + 7+

B@ 0 !#12! ! !3 4) ! B))) + H3
'3#A71 *+
< *)4
3: A & + ===

!"#$# % & & > !' (> ! !)) 78))

B 0 !#12! ! !3 4) !))
78)) *+
< *)4

!"#$# % & & > !' (> ! !3 4) " D + 4 *)* 8*
4 /

B 0 !#12! ! !3 4) ! > B >@> B *+ *+
< 3 4) "

!"#$# % & & > !' (> ! ! * + & / # - + , %
7+

B 0 !#12! ! !3 4) ! > * + & / *+

!"#$# % & & > !' (> ! !$) X $ 4 /

B 0 !#12! ! !3 4) ! > $) X $ 4 / *+
< *)4
3: A & + ===

!"#$# % & & > !' (> ! !$) * 8 D

B 0 !#12! ! !3 4) ! >

$ 6) 6 6 * 86D 6G) 6H *+
< >@ *)4

!"#$# % & & > !' (> ! ! D + D :/ $)
)*)))

B >0 !#12! ! !3 4) ! > D + D :/ $)
)*))) *+
< > *)4

!"#$# % & & > !' (> ! !G 1) #) / /
) +) , ")) + "/)

B @0 !#12! ! !3 4) ! >>
G 6 61) 6#) / 6 /6) +6) 6") 6) +6"/) *+
< > *)4

!"#$# % & & > !' (> ! >

!"#$# % & & > !' (> ! >!D 2)))

B 0 !#12! ! !3 4) ! >@
6' (6D 2) $)) *+
< *)4

!"#$# % & & > !' (> ! >! #) #*

B B0 !#12! ! !3 4) ! > 6#) 6
6#* 6 *+
< 2) %)& 4)

B 0 !#12! ! !3 4) ! > 6#) 6
6D :) /6 *+
< $ 3 4 *)) + $

!"#$# % & & > !' (> ! >!$ # X) ' : ") / @
*) (

BB 0 !#12! ! !3 4) ! >B $ # X) ' : ") /
@ *) (*+
< *)4

> B

!

" #$%&% ' ((

" #$%&% ' ((#) *

" #$%&% ' ((#) * #

" #$%&% ' ((#) * # #% + , ' - (
.- - , /

0 " #%12# # #3 4 - # % + , ' - (
.- - , / .,
54 6 0 " -. / -(7 ,

" #$%&% ' ((#) * # #8 - 4 8 - , 97. . - & 7
- , % 4

0 " #%12# # #3 4 - # 8 - 4 8 - , 97. . - .,
: .-4

" #$%&% ' ((#) * # # - 1 - . ; 8 ;
- , - '-

0 " #%12# # #3 4 - # .,
: / 7 $,

0 " #%12# # #3 4 - # - 1 - . ;
8 ; - , - '- .,

" #$%&% ' ((#) * # #' < % - , " . -

0 " #%12# # #3 4 - #
' < =% =- ,=" . - .,
54 !6 0 " -. 7 -4 - .

" #$%&% ' ((#) * # #2 - - <7 > & 9

0 " #%12# # #3 4 - # > & 9 2 - -
. 4 ?) , - 6 0 .,
: @ .-4

" #$%&% ' ((#) * # # . 4 9- ,< * % ;
- ; - , 3. - 7

@0 " #%12# # #3 4 - # . 4 9- ,< * % ;
- ; - , 3. - 7 .,

" #$%&% ' ((#) * # #& . %.. - 7 -4 -
& . ; A,

0 " #%12# # #3 4 - # @ & . %.. -
7 -4 - & . ; A, .,
: ! .-4

" #$%&% ' ((#) * # #& < B-

!0 " #%12# # #3 4 - # .,
: & & <

0 " #%12# # #3 4 - # A$.,
: & & <

0 " #%12# # #3 4 - # / 4 - & - -
/ 7= % + , 7 7 . 4 6 !! 0 ,C(
: .-4

0 " #%12# # #3 4 - # / 4 - & - -
/ 7= % + , 7 7 . 4 6 !! 0 .,

0 " #%12# # #3 4 - # 9 4 - 7 7 - ,
& . - / .,
: .-4

0 " #%12# # #3 4 - # 9 4 - 7 7 - ,
& . - / 7 .,

0 " #%12# # #3 4 - # + , 4 7
- , 7 7 . 4 ' - , => % -, 6 !! 0 ,C(

0 " #%12# # #3 4 - # + , 4 7
- , 7 7 . 4 ' - , => % -, 6 !! 0 .,

@0 " #%12# # #3 4 - # - * / & . - / 7
- , 7 7 % - - - - 7 6 0 .,

0 " #%12# # #3 4 - # 7 7 - , % 4 < -
/ 7 % - - - 7 6 !!!0 ,C(
: @@ .-4

!0 " #%12# # #3 4 - # 7 7 - , % 4 < -
/ 7 % - - - 7 6 !!!0 .,

0 " #%12# # #3 4 - # 3 D - - ,
4 7; - - , 7 7 % > "1" 6 0 ,C(

0 " #%12# # #3 4 - # 3 D - - ,
4 7; - - , 7 7 % > "1" 6 0 .,

0 " #%12# # #3 4 - # .- 7- / 7 - , 7 7
9 , $-7 6 ! 0 ,C(
: .-4

0 " #%12# # #3 4 - # .- 7- / 7 - , 7 7
9 , $-7 6 ! 0 .,
: .-4
3< 8 (, EEE
54 6 0 C (-

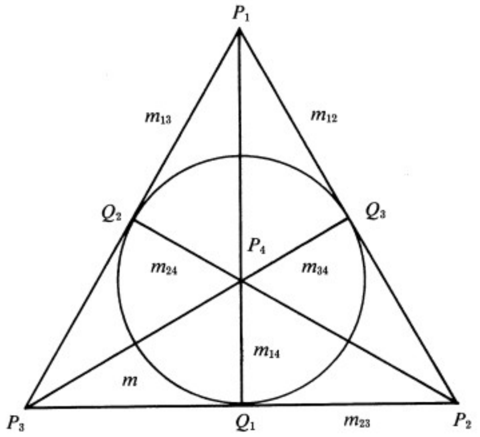
P_1
m_{13}
m_{12}
Q_2
Q_3
P_4
m_{24}
m_{34}
m_{14}
m
P_3
Q_1
m_{23}
P_2

F 4 (. - , (- . C (. -

0 " #%12# # #3 4 - # 1- = = =& .,

0 " #%12# # #3 4 - # 7 7; -. - , .- %
+ , - - 4 / 7 > 7 " 4 6 0 .,
: .-4
3< 8 (, EEE

0 " #%12# # #3 4 - # & - 7 . / 7 ?
7 7 G , < 4 " .- 7 6 0 ,C(
: .-4
3< 8 (, EEE

@0 " #%12# # #3 4 - # & - 7 . / 7 ?
7 7 G , < 4 " .- 7 6 0 .,

0 " #%12# # #3 4 - # & - 7 . / 7 - ,
7 7 , 4 >+% .,

!0 " #%12# # #3 4 - # & . 47 - . 4
-4 6 !! 0 .,

0 " #%12# # #3 4 - # & - - / 7 %
+ , 7 7 . 4 -4 ' B * 6 ! 0 .,

0 " #%12# # #3 4 - # .,

0 " #%12# # #3 4 - # & . - / 7 - ,
7 7 % - - - - 7 6 0 .,

" #$%&% ' ((#) * # @

" #$%&% ' ((#) * # @#% / 7 %.. H -

0 " #%12# # #3 4 - # ! % / 7 %.. H -
1 4 & 7 8 4 ; % 4 < - - , & . 47 " . - - . 4 6 0 .,
: @@ .-4

0 " #%12# # #3 4 - # ! D - < * .,
: 2 - - 4 < -
: .-4
: 2 - - 4 < -

" #$%&% ' ((#) * # @#% $ 4 . / I-
7 - ,

0 " #%12# # #3 4 - # < *=.7 - , .,
: & / - 7 - , / I-

0 " #%12# # #3 4 - # ! ! @ .,
: & .7 - , - , . / I
: .-4

@0' #%12# # #3 4 - # , 4 4 .7 .,
: & $ 4 & / - 7 - , >

0 " #%12# # #3 4 - # D < =.7 - , =4 I- .,

" #$%&% ' ((#) * # @#A / 7 > 7

!0 " #%12# # #3 4 - # A$ @ .,
: / & - - / . - , .
: @ .-4

0 " #%12# # #3 4 - # A A , .,
: .-4

0 " #%12# # #3 4 - # A ,J A / 7 .,

0 " #%12# # #3 4 - # / 7 =(II- .,
: 3' &9A /A3 A&1B
3 A1 &9A F+A8$ G+&9 3'A A8A A'&

0 " #%12# # #3 4 - # 7 .,
: - K - D
: .-4
3< 8 (, EEE

- Compléter la copie d'écran de Youssra par une légende associant courbes et aires.
- Retrouver le « texte caché » de la rédaction de Youssra.

Pour aller plus loin, j'ai pensé utiliser une fonction polynôme de degré 2 pour modéliser la courbe verte qui ressemble à un morceau de parabole. J'ai tâtonné avec trois curseurs, je suis certaine que w=0, et que u et v sont opposés. Mais ça ne marche pas, je n'ai pas réussi à superposer les courbes !

- Expliquer pourquoi w=0.
- Expliquer pourquoi u et v sont opposés.

Séance 2 : analyse de la production de Mylène

Pour comprendre le problème qui me semblait compliqué, j'ai d'abord utilisé GéoGébra pour faire une figure.

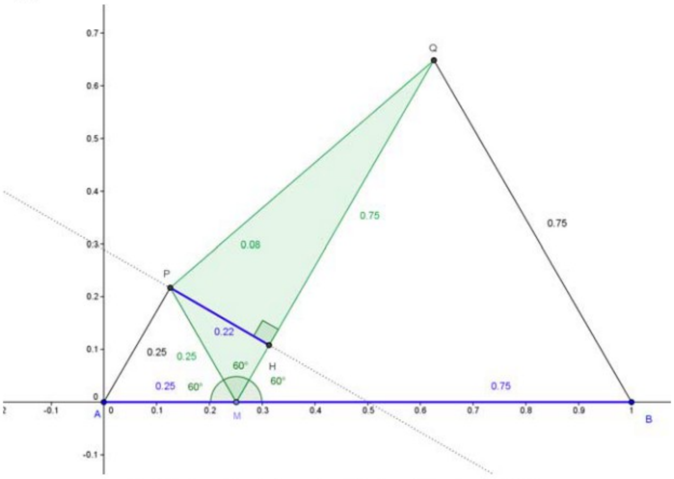

Ici l'aire du triangle MPQ est égale à 0,08.

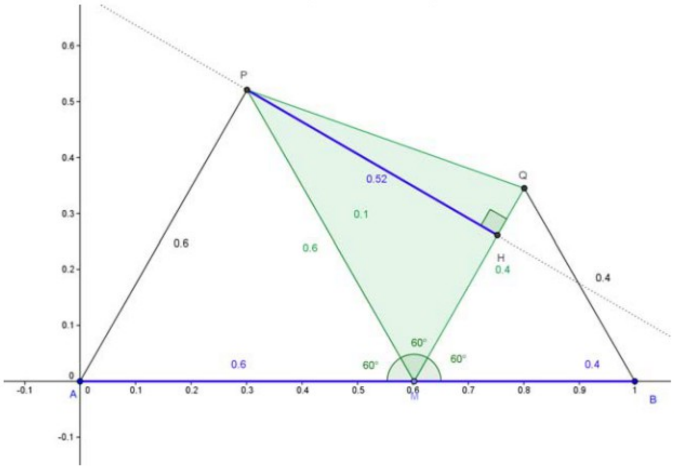

Ici l'aire du triangle MPQ est égale à 0,1.

Première question du sujet.

J'ai calculé, dans le cas général, la hauteur PH en posant x=AM.

Puis j'ai obtenu l'aire du triangle MPQ : $0,4x \times (1 - x) = 0,4x - 0,4x^2$

- Critiquer le résultat énoncé par Mylène :
  « l'aire du triangle MPQ est égale à $0,4x \times (1 - x) = 0,4x - 0,4x^2$ ».

0 " #%12# # #3 4 - # < - , A , -
/ 7 % ; G * 6$ (; ! 0 .,
3< 8 (, EEE

0 " #%12# # #3 4 - # L (H >- A **-
> *M M 8- >- - .MM G - & -(% (- - 4 < - 4 7 . 4
.,
: @ .-4

54 @6 0 " -. -. C (.-

3< 8 (, EEE

" #$%&% ' ((#) * # @#/K K

0 " #%12# # #3 4 - #) - ,- .,
: % ; K4 - . (

@0' #%12# # #3 4 - # (4 .,
: .-4
3< 8 (, EEE

" #$%&% ' ((#) * # @#8- $K ($ - K - D

0 " #%12# # #3 4 - # / 4 7- 8- ,K (,
- K - D $, 6 ! @0 ,C(
: @! .-4
3< 8 (, EEE
: 8 (<

!0 " #%12# # #3 4 - # / 4 7- 8- ,K (,
- K - D $, 6 ! @0 .,

" #$%&% ' ((#) * # @# -

0 " #%12# # #3 4 - # /- A % @ .,
: % 4N<
: .-4

0 " #%12# # #3 4 - #
/K K =" - =, = - K - D 8 4 6 ! 0 ,C(
: @ .-4
3< 8 (, EEE
: 8 (<

0 " #%12# # #3 4 - #
/K K =" - =, = - K - D 8 4 6 ! 0 .,
: @ .-4
3< 8 (, EEE
: 8 (<

0 " #%12# # #3 4 - # 8 < K; 9K 7

/K K ===" - =, = - K - D 6 !! 0 .,
: .-4
3< 8 (, EEE
: 8 (<

0 " #%12# # #3 4 - # 8 4
" - =, = - - D =/ ! ,C(

0 " #%12# # #3 4 - # 8 4
" - =, = - - D =/ ! .,

0 " #%12# # #3 4 - # - - , " - ,
- - D ! / , . (,C(

@0 " #%12# # #3 4 - # - - , " - ,
- - D ! / , . (.,

0 " #%12# # #3 4 - # & - - K - D
A /K K - D 9- 6 ! 0 ,C(
: .-4

!0 " #%12# # #3 4 - # & - - K - D
A /K K - D 9- 6 ! 0 .,

0 " #%12# # #3 4 - # & - = - K - D
A % - 7 ! @ ,C(
: .-4

0 " #%12# # #3 4 - # & - = - K - D
A % - 7 ! @ .,

" #$%&% ' ((#) * # @# , % 4 < - / 7
" & 7

0 " #%12# # #3 4 - # , % 4 < - / 7
" & 7= - += - - 8 - 7 - , % % 4 < - / 7 .,

0 " #%12# # #3 4 - # , % 4 < - / 7
" & 7= - ++= (- -< 8 - 7 - , C (% 4 < - / 7 .,

" #$%&% ' ((#) * # @# , / 7

0 " #%12# # #3 4 - # $ 4 - / 7 /
, $ 4 - + -4 % - 7 4- >- - 6 0 .,
: @ .-4

0 " #%12# # #3 4 - # $ - 4 7 7;
, ; - , -.. - . 4 6 0 .,
: @ .-4
3< 8 (, EEE

0 " #%12# # #3 4 - # A (, % 4 < -

/ 7) * M)- 6 !! 0 ,C(

@0 " #%12# # #3 4 - # A (, % 4 < -
/ 7) * M)- 6 !! 0 .,
: .-4

0 " #%12# # #3 4 - # , - , %.. - = - =
+ , 9 47 & 7 . 4 6 !! 0 ,C(

!0 " #%12# # #3 4 - # , - , %.. - = - =
+ , 9 47 & 7 . 4 6 !! 0 .,
: ! .-4
: , / 7 , - , %.. -
- +++ + , 9 47 & 7

@ 0 " #%12# # #3 4 - # , - , %.. - = - ++=
& / 7 - , & . 47 - , . 4 -4 ' ,C(

@ 0" #%12# # #3 4 - # , - , %.. - = - ++=
& / 7 - , & . 47 - , . 4 -4 ' .,
: .-4
54 6 0 " -.- - 54 - .- - (- 6 7 .- 0

(f) The following manifolds are homogeneous spaces for the unitary group $U(n)$:

(i) The odd-dimensional sphere S^{2n-1}, defined in n-dimensional complex space $\mathbb{C}^n$ by the equation

$$|z^1|^2 + \cdots + |z^n|^2 = 1.$$

It is not difficult to show that in this case

$$S^{2n-1} \cong U(n)/U(n-1) \cong SU(n)/SU(n-1).$$

(ii) The complex projective space $\mathbb{C}P^{n-1}$. In this case we have

$$\mathbb{C}P^{n-1} \cong U(n)/(U(1) \times U(n-1)).$$

(iii) The complex Grassmannian manifold $G_{n,k}^{\mathbb{C}}$ consisting of the k-dimensional complex planes in $\mathbb{C}^n$ passing through the origin. Here

$$G_{n,k}^{\mathbb{C}} \cong U(n)/(U(k) \times U(n-k)).$$

5.3. Exercises

1. Let M be a homogeneous space of a Lie group G, and let H be the isotropy group. Prove that the dimension of the manifold M is the difference in the dimensions of G and H:

$$\dim M = \dim G - \dim H.$$

Compute the dimension of the Grassmannian manifold $G_{n,k}$.

2. Prove the compactness of the manifolds $V_{n,k}$ and $G_{n,k}$.

3. Let $m = (m_1, \ldots, m_k)$ be a partition of the integer n, i.e.

$$n = m_1 + \cdots + m_k, \qquad m_i \geq 0.$$

A collection of linear subspaces $\pi_0, \pi_1, \ldots, \pi_k$ of the space $\mathbb{R}^n$ is called an *m-flag* if:

(i) $\dim \pi_i - \dim \pi_{i-1} = m_i$;
(ii) $\pi_0 = 0$, $\pi_k = \mathbb{R}^n$;
(iii) $\pi_{i-1} \subset \pi_i$.

Show how the totality of all *m-flags* $F(n, m)$ can be made to serve as a homogeneous space for the group $O(n)$, and calculate its isotropy group.

§6. Spaces of Constant Curvature (Symmetric Spaces)

6.1. The Concept of a Symmetric Space

Of great interest are those manifolds endowed with a metric g_{ab} whose curvature tensor (defined in §30.1 of Part I in terms of the connexion compatible with the metric) has identically zero covariant derivative (see §§28, 29 of Part I):

$$\nabla_s(R_{abcd}) = 0. \tag{1}$$

In any metrized space the components of the covariant derivative of the curvature tensor satisfy certain relations (namely Bianchi's identities—see Exercise 7, §30.5, Part I). However condition (1) places further severe restrictions on those components; thus, in particular, it follows from (1) that the scalar characteristics of the curvature are constant:

$$R = R_a^a = \text{const.}; \qquad R_{abcd}R^{abcd} = \text{const.}$$

It turns out also that under certain global conditions on a manifold, condition (1) implies the homogeneity of the metric g_{ab}; this is the case if the manifold is "simply connected" (see §17 below). Any manifold satisfying (1) is obtainable from some simply connected such manifold M as the quotient space (i.e. by identification) under some discrete group of motions. In this construction it can happen that the discrete group Γ in question is not central in the full isometry group of M, in which case the space M/Γ will not be homogeneous; such spaces are sometimes called "locally homogeneous" or "locally symmetric".

However our approach to symmetric spaces will be via the following definition.

6.1.1. Definition. A simply connected manifold M with a metric g_{ab} defined on it, is called a *symmetric space* (or *symmetric manifold*) if for every point x of M there exists an isometry (motion) $s_x: M \to M$ with the properties that x is an isolated fixed point of it, and that the induced map on the tangent space at x reflects (i.e. reverses) every tangent vector at x, i.e. $\xi \mapsto -\xi$. Such an isometry is called a *symmetry of M at the point x*.

The significance of the requirement in this definition that the manifold be simply connected will appear below (in §§17, 18). In the present section we shall not make use of the properties of simply connected manifolds; the reader not familiar with these properties may wish to attempt the appropriate exercises in §6.6 below, after he has studied the relevant sections of Chapter 4.

6.1.2. Lemma. *Every symmetric space satisfies condition* (1).

PROOF. Let x be any particular point of the symmetric space M and let s_x be a symmetry of M at the point x. We can choose co-ordinates in some neighbourhood of x such that at x itself we have (see §29.3 of Part I)

$$x^a = 0, \qquad g_{ab} = \delta_{ab}, \qquad \frac{\partial g_{ab}}{\partial x^a} = 0.$$

(Here we are making the (inessential) assumption that the metric is Riemannian.)

Under the symmetry s_x, at the point x the tensor $\nabla_s R_{abcd}$ goes into its negative (and the tensors g_{ab} and R_{abcd} go into themselves) simply by virtue of the symmetry property of s_x and the transformation rule for tensors. On the other hand, since s_x is actually an isometry it is clear that $\nabla_s R_{abcd}$ must go into itself at x. The only way out is for $\nabla_s R_{abcd}$ to vanish at x, as required. □

Remark. The converse of this statement (under the assumption of simple-connectedness) is also true; however, in view of the greater technical complexity of its proof, we shall not prove it here. The following discussion may, however, be illuminating. In some neighbourhood of each point x of a Riemannian manifold we can define a *local symmetry* s_x as follows: Consider the pencil of geodesics on M passing through the point x (and assumed parametrized so that for each geodesic γ we have $\gamma(0)=x$); then for sufficiently small τ set

$$s_x(\gamma(\tau))=\gamma(-\tau).$$

(Recall (e.g. from §29.2 of Part I) that in a sufficiently small neighbourhood of x the geodesics through x will intersect nowhere else.)

Exercise

Prove that such local transformations s_x are for all x isometries of the manifold precisely if condition (1) holds on the manifold. (*Hint*. The simplest case is $n=2$, when the curvature tensor is determined by a single scalar R. For the general case it is easiest to deduce the preservation of the curvature tensor by the transformations s_x, from the preservation of Jacobi's equation along a geodesic.)

The existence of the "symmetries" s_x for all points x of a manifold M guarantees a sufficiently large supply of isometries for M to be (locally at least) homogeneous.

6.1.3. Lemma. *A symmetric space M is locally homogeneous; i.e. around each point $x\in M$, there is a neighbourhood such that for each point $\bar{x}$ in that neighbourhood (i.e. for each $\bar{x}$ sufficiently close to x) there exists an isometry of M sending x to $\bar{x}$. For each pair of points $x,\ y\in M$ which can be joined by a geodesic, there exists an isometry g of M such that $g(x)=y$.*

Proof. The first statement of the lemma follows from the second since around every point x there is a neighbourhood with the property that every point $\bar{x}$ in that neighbourhood is joined to x by a geodesic, i.e. the geodesics through x sweep out the whole neighbourhood.

For the second statement, let γ be a geodesic arc joining the points x and y, and parametrized by the natural (length) parameter τ, with $0\le\tau\le T$, $\gamma(0)=x$, $\gamma(T)=y$. Let s_z be a symmetry of the manifold M at the point $z=\gamma(T/2)$. It follows from its symmetry property, together with the fact that it is an isometry of M, that s_z must send γ to γ, and therefore interchange x and y. (If the metric is pseudo-Riemannian of type $(1,\ n-1)$, and γ is an isotropic geodesic, then we may take as τ the "affine" (called also "natural" in Chapter 5 of Part I) parameter yielded in solving the equation for the geodesics (see §29.2 of Part I). □

Remark. Since any two points of a (connected) Riemannian manifold can be joined by a broken geodesic, it follows almost immediately that a symmetric Riemannian manifold is always homogeneous.

6.2. The Isometry Group of a Manifold. Properties of Its Lie Algebra

Henceforth in this section we shall consider only homogeneous, symmetric manifolds M (hence satisfying (1)), with metric g_{ab}. The Lie group of all isometries of M will be denoted by G, and the isotropy group by H, so that we may identify M with the set of left cosets of H in G (i.e. $M = G/H$, in the notation of the preceding section).

Let $\gamma = \gamma(\tau)$ be a geodesic on M parametrized by a natural parameter τ, and write $x_0 = \gamma(0)$. For each appropriate real T we define a map $f_{T,\gamma}\colon M \to M$ by setting

$$f_{T,\gamma} = s_{x_0}^{-1} s_x,$$

where $x = \gamma(-T/2)$ (see Figure 11). This map has the following three important properties:

(i) $f_{T,\gamma}$ moves each point of γ through a time-interval T along the geodesic (as indicated in Figure 11):

$$\gamma(T) \mapsto \gamma(\tau + T);$$

(ii) $f_{T,\gamma}$ parallel transports vectors along the geodesic;

(iii) for any fixed geodesic γ, the transformations $f_{T,\gamma}$ with T variable, form a one-parameter subgroup of the isometry group G:

$$f_{T_1+T_2,\gamma} = f_{T_1,\gamma} f_{T_2,\gamma},$$

$$f_{-T,\gamma} = (f_{T,\gamma})^{-1}.$$

From the last of these properties and §3.1 above, it follows that for each geodesic γ the one-parameter subgroup $f_{T,\gamma}$ of G has the form

$$f_{T,\gamma} = \exp(TB_\gamma),$$

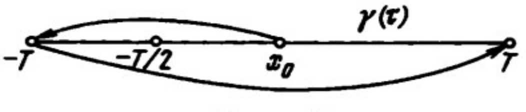

Figure 11

where B_γ is a certain vector in the Lie algebra $\mathfrak{g}$ of G (namely the tangent vector to the curve $f_{T,\gamma}$ at $T=0$). We denote by L^1 the linear subspace of the algebra $\mathfrak{g}$ spanned by the vectors $B_\gamma \in \mathfrak{g}$, where γ ranges over all geodesics through x_0, and by L^0 the Lie algebra of the isotropy group H_{x_0} of the point x_0. It follows (essentially from the fact that corresponding to each direction on the tangent plane to M at x_0 there is a geodesic through x_0 with that direction (see §29.2 of Part I)) that

$$\mathfrak{g} = L^0 + L^1 \quad \text{(direct sum of subspaces).} \tag{2}$$

If γ_1, γ_2 are two geodesics through the point x_0 then it can be shown that for small ε the product

$$f_{\varepsilon,\gamma_1} f_{\varepsilon,\gamma_2} f_{-\varepsilon,\gamma_1} f_{-\varepsilon,\gamma_2}$$

sends x_0 to a point whose distance from x_0 is of order ε^3. (This follows without difficulty from the properties of the Riemann curvature tensor. Verify!) We deduce from this (e.g. using formula (7) of §3.1) that we must have $[B_{\gamma_1}, B_{\gamma_2}] \in L_0$, whence $[L^1, L^1] \subset L^0$.

Suppose now that $g_T = \exp(TA)$ is a one-parameter subgroup of G leaving x_0 fixed (so that $A \in L_0$). Let γ be any geodesic through x_0, and denote by $\tilde{\gamma}$ the image of γ under the map g_ε. It then follows, again from the isometric property of the maps involved, that for small enough ε the map $g_\varepsilon f_{T,\gamma} g_{-\varepsilon}$ translates the points of the geodesic $\tilde{\gamma}$ along that geodesic (which of course also passes through x_0). Hence the tangent vector to the one-parameter subgroup $g_\varepsilon f_{T,\gamma} g_{-\varepsilon}$ (with parameter T), is in L_1. We now look for this tangent vector. From the basic facts about Lie algebras described in §3.1 it follows that for any two elements X, Y of the Lie algebra of a Lie group we have

$$\exp(tX)\exp(tY) = \exp\left(t(X+Y) + \frac{t^2}{2}[X, Y] + \text{higher-order terms}\right).$$

(This is a weak form of the "Campbell–Baker–Hausdorff formula".) It is an easy consequence of this that

$$\exp(tX)\exp(tY)\exp(-tX) = \exp(tY + t^2[X, Y] + \text{higher-order terms}).$$

Putting $t=1$, $X=\varepsilon A$, $Y=TB_\gamma$, we deduce that the desired tangent vector is $B_\gamma + \varepsilon[A, B_\gamma]$. Since $B_\gamma \in L^1$, it follows that $[A, B_\gamma] \in L^1$, whence $[L^0, L^1] \subset L^1$.

We include these facts in the following

6.2.1. Lemma. *With G and $\mathfrak{g} = L^0 + L^1$ as above, we have*

$$[L^0, L^0] \subset L^0, \qquad [L^1, L^0] \subset L^1, \qquad [L^1, L^1] \subset L^0. \tag{3}$$

A Lie algebra which decomposes as the direct sum of two subspaces satisfying (3) is called a *$\mathbb{Z}_2$-graded Lie algebra*, since (3) can be rewritten as

$$[L^i, L^j] \subset L^{(i+j) \bmod 2}. \tag{4}$$

6.2.2. Corollary. *With G as above, and* $\mathfrak{g} = L^0 + L^1$ *(direct) its Lie algebra, the linear operator*

$$\sigma: \mathfrak{g} \to \mathfrak{g}$$

whose restriction to L^0 *is the identity map* 1, *and whose restriction to* L^1 *is the reflection* -1, *is a Lie algebra automorphism (i.e. also preserves commutators). (The map* σ *is an "involution", i.e.* $\sigma^2 = 1$.)

(The converse is also true. To each involuntary automorphism σ of a Lie algebra $\mathfrak{g}$, there corresponds a $\mathbb{Z}_2$-grading $\mathfrak{g} = L^0 + L^1$ (direct sum of spaces) of the Lie algebra, where L^0 is the set of elements fixed by σ and L^1 the set of elements which σ negates.)

In view of the homogeneity of the manifold M, the local geometry around any point x_0 is determined by a scalar product on the tangent space $\mathbb{R}^n_{x_0}$ at the point. (Here $n = \dim M$.) We now elicit a certain property (familiar from Part I) which this scalar product must have.

Note first that the tangent space $\mathbb{R}^n_{x_0}$ can be identified naturally with the space $L^1 \subset \mathfrak{g}$. Let A be any element of L^0, and consider the one-parameter subgroup $g_T = \exp(TA)$. For each T we have the map

$$\operatorname{Ad}(g_T): \xi \mapsto \xi_T, \qquad \xi \in L^1.$$

As was shown in the course of proving part (i) of Theorem 3.1.4, we have

$$\xi_T = \operatorname{Ad}(g_T)(\xi) = \xi + T[A, \xi] + O(T^2),$$

whence

$$\left.\frac{d\xi_T}{dT}\right|_{T=0} = [A, \xi] = (\operatorname{ad} A)(\xi). \tag{5}$$

In view of the fact that g_T is an isometry of M the inner product on L^1 should be invariant under $\operatorname{Ad}(g_T)$, i.e.

$$\langle \xi_T, \eta_T \rangle = \langle \xi, \eta \rangle,$$

whence on differentiating with respect to T at $T = 0$, and using (5), we obtain

$$\langle [A, \xi], \eta \rangle + \langle \xi, [A, \eta] \rangle = 0. \tag{6}$$

This is the condition on the metric (i.e. scalar product) on $L^1 = \mathbb{R}^n_{x_0}$, that we were seeking. (Cf. the definition of a Killing metric in §24.4 of Part I.)

6.3. Symmetric Spaces of the First and Second Types

In the preceding subsection we obtained what might be called the algebraic model of a symmetric space. In principle all symmetric spaces can be classified (in the framework of the classification of compact Lie groups). In the present subsection we consider the most important examples of such spaces.

The simplest examples of (simply-connected) symmetric spaces are provided of course by the Euclidean spaces $\mathbb{R}^n$ and pseudo-Euclidean spaces $\mathbb{R}^n_{p,q}$ (where the curvature is zero). Here the group G is the full isometry group of $\mathbb{R}^n$ (or $\mathbb{R}^n_{p,q}$). (For the structure of some of these groups see §4 of Part I.) The isotropy group H is $O(n)$ (or $O(p, q)$), and the space $L^1 = \mathbb{R}^n$ consists of the translations. We have, as in the general case, $\mathfrak{g} = L^0 + L^1$, with the usual commutator relations between L^0 and L^1 (as in Lemma 6.2.1), and with the additional relation $[L^1, L^1] = 0$. Non-simply-connected symmetric spaces (earlier called "locally symmetric") can be obtained from these simply-connected ones by identification under various discrete groups Γ consisting of translations (and possibly also involving certain reflections, as for instance in the case of the Klein bottle—see §18 below).

In the remainder of our examples the group G will be semisimple. Recall from §3.1 that semisimplicity is equivalent to non-degeneracy of the Killing form on the Lie algebra $\mathfrak{g}$. Recall also that the Killing form $\langle\ ,\ \rangle$ on $\mathfrak{g}$ is defined by:

$$\langle A, B\rangle = -\operatorname{tr}(\operatorname{ad} A \operatorname{ad} B),$$

where $(\operatorname{ad} X)(\xi) = [X, \xi]$. (We shall also restrict G to being the connected component of the identity of the full isometry group.)

There are two distinct important types of simply-connected symmetric spaces with such G (even assuming the metric positive definite, i.e. Riemannian), namely:

Type I: the group G is compact and the Killing form on $\mathfrak{g}$ is positive definite. (We note here the result that a Lie group is compact if and only if it is a closed subgroup of some $O(m)$.)

Type II: the group G is non-compact and the Killing form on the Lie algebra $\mathfrak{g}$ is indefinite.

We consider the simplest (non-trivial) example of each type.

(a) The sphere S^2 is of type I. Here $G = SO(3)$, which is compact, and $H = SO(2)$.

(b) The Lobachevskian plane L^2 is of type II. Here G is the connected component of the identity of $SO(1, 2)$ (shown in §13.2 of Part I to be isomorphic to $SL(2, \mathbb{R})$), and $H = SO(2)$. The Lie algebra $\mathfrak{g}$ consists of all 2×2 matrices with zero trace, and the Killing form is given by

$$\langle A, B\rangle = -\operatorname{tr}(AB).$$

As a basis for the Lie algebra we may take

$$A_1 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad A_2 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \qquad A_3 = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We then have

$$A_1^2 = A_2^2 = 0, \qquad A_3^2 = 1,$$

and

$$\langle A_1, A_2\rangle = -1, \qquad \langle A_1, A_3\rangle = \langle A_2, A_3\rangle = 0,$$
$$\langle A_3, A_3\rangle = -2, \qquad \langle A_1, A_1\rangle = \langle A_2, A_2\rangle = 0.$$

It follows essentially from the fact that the matrices in $H = SO(2)$ have the form

$$\begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix},$$

that the subalgebra $L^0 \subset \mathfrak{g}$ is comprised of the matrices of the form $\lambda(A_1 - A_2)$. The subspace L^1 of $\mathfrak{g}$ is spanned by the vectors $A_1 + A_2, A_3$. It is easy to check that the inclusions (3) hold.

The restriction of the Killing form to the subspace L^1 is positive definite; this reflects the positive definiteness of the metric on the Lobachevskian plane.

Exercise

Investigate the general cases S^n and L^n.

6.4. Lie Groups as Symmetric Spaces

A Lie group Q endowed with a Riemannian metric invariant under left and right multiplications by group elements, can itself be regarded as a symmetric space. The isometry group of Q has a subgroup isomorphic to $Q \times Q$, whose action on Q is defined by

$$(g_1, g_2)\colon q \mapsto g_1 q g_2^{-1}.$$

The isotropy group of this action is clearly the diagonal subgroup $H = \{(q, q) | q \in Q\}$, which is isomorphic to Q: clearly $H(1) = 1$. For each q in Q the corresponding symmetry is defined by

$$s_q\colon x \mapsto q x^{-1} q.$$

(Verify that this does indeed define a symmetry.) In particular, $s_1(x) = x^{-1}$.

We shall examine in detail the case where Q is a compact connected subgroup of $SO(m)$, with the Euclidean metric

$$\langle A, B\rangle = \operatorname{tr}(AB^{\mathrm{T}}), \tag{7}$$

where B^{T} denotes the transpose of the matrix B. (Recall that it can be shown that a Lie group is compact if and only if it is a closed subgroup of some $O(m)$.)

Exercise

Show that the scalar product (7) is the Killing metric on $SO(m)$ determined by the Killing form on its Lie algebra (cf. §24.4 of Part I).

(The formula for the curvature of the Killing metric was derived in §30.3 of Part I. It follows from that formula that the Ricci tensor R_{ab} is positive definite. In the same subsection it was shown that the geodesics through the identity are precisely the one-parameter subgroups of Q.)

As noted in §24.4 of Part I, the group $SO(m)$ lies on the sphere of radius $\sqrt{m}$ (in the Euclidean space $\mathbb{R}^{m^2}$ of all $m \times n$ matrices with the metric (7)), since for $A \in SO(m)$, we have $AA^{\mathrm{T}} = 1$, whence $\langle A, A \rangle = m$; thus

$$SO(m) \subset S^{m^2-1}$$

6.4.1. Lemma. *The (Euclidean) scalar product (7) is invariant under right and left translations (i.e. multiplications) by elements of $SO(m)$.*

Proof. Let $g \in SO(m)$, and let A, B be any $m \times m$ matrices. Then

$$\langle gA, gB \rangle = \operatorname{tr}(gAB^{\mathrm{T}}g^{\mathrm{T}}) = \operatorname{tr}(gAB^{\mathrm{T}}g^{-1}) = \operatorname{tr}(AB^{\mathrm{T}}) = \langle A, B \rangle,$$

and

$$\langle Ag, Bg \rangle = \operatorname{tr}(Agg^{\mathrm{T}}B^{\mathrm{T}}) = \operatorname{tr}(AB^{\mathrm{T}}) = \langle A, B \rangle,$$

whence the desired conclusion. □

6.4.2. Corollary. *The metric (7) restricted to any subgroup Q of $SO(m)$ is invariant under right and left multiplications $q \mapsto q_1 q q_2$.*

We call such a metric *bi-invariant* or *two-sided invariant.*

6.4.3. Lemma. *Every bi-invariant metric on a simple Lie group is proportional (with constant proportionality factor) to the Killing metric.*

Proof. Let Q be a simple Lie group with bi-invariant metric $\langle\ ,\ \rangle$. The bi-invariance implies that for all elements A, B, C of the Lie algebra L of Q, and all $g_T = \exp(AT)$, we have

$$\langle \operatorname{Ad}(g_T)(B), \operatorname{Ad}(g_T)(C) \rangle = \langle B, C \rangle, \tag{8}$$

whence it follows, just as in the derivation of (6) above, that

$$\langle [A, B], C \rangle + \langle B, [A, C] \rangle = 0. \tag{9}$$

Now let $g_{ab}, \bar{g}_{ab}$ be two metrics on Q satisfying (8), (9). Then any linear combination $g_{ab} - \lambda \bar{g}_{ab}$ will also be Ad-invariant (or equivalently skew ad-invariant). Let λ_1 be any eigenvalue of the pair of quadratic forms $g_{ab}, \bar{g}_{ab}$, i.e. $\det(g_{ab} - \lambda_1 \bar{g}_{ab}) = 0$. (The symmetry of $g_{ab}, \bar{g}_{ab}$ implies that λ_1 is real.) The subspace R_{λ_1} of all eigenvectors corresponding to the eigenvalue λ_1 is easily seen (from (9)) to be a (non-zero) ideal of L, whence by the assumed simplicity of L, we must have $R_{\lambda_1} = L$. Hence $g_{ab} = \lambda_1 \bar{g}_{ab}$. Since the Killing metric satisfies (9), the desired result follows. □

6.4.4. Corollary. *Every simple subgroup Q of the group $SO(m)$ endowed with the Killing metric, can be isometrically embedded in the sphere S^{m^2-1} endowed with a metric proportional to the usual metric on the sphere.*

6.4.5. Corollary. *Since the Ricci tensor R_{ab} (determined by the Killing metric g_{ab} on a group) also satisfies* (8), (9), *it follows that for simple groups, $R_{ab} = \lambda g_{ab}$ where $\lambda =$ const.*

Note finally that, as remarked above, it follows essentially from §30.3 of Part I that R_{ab} is positive definite for compact connected Lie groups. This is true also for semisimple groups, since each such group is (locally) a direct product $G_1 \times \cdots \times G_k$ of simples, and the sign of λ_i is easily determined for each of the simple factors G_i.

6.5. Constructing Symmetric Spaces. Examples

We now return to general symmetric spaces. In the notation of §6.2 above write

$$M = G/H, \qquad \mathfrak{g} = L^0 + L^1 \quad \text{(direct sum of subspaces)},$$

where M is a given symmetric space, G is its isometry group, L^0 is the Lie algebra of the isotropy group H, and L^1 is identifiable with the tangent space $\mathbb{R}^n_{x_0}$ to M at the point x_0 (fixed by H, i.e. $H = Hx_0$). Recall also that by virtue of the homogeneity of M, its metric is determined locally by a metric on L^1 satisfying (6); in what follows we shall assume the metric on M to be obtained from the Killing form on $\mathfrak{g}$ (see below).

6.5.1. Lemma. *The subspaces L^0 and L^1 of the Lie algebra $\mathfrak{g}$ are orthogonal with respect to the Killing form.*

Proof. From Lemma 6.2.1, it is immediate that for all $A \in L^0$, $B \in L^1$ we have

$$\begin{aligned} &\operatorname{ad} A(L^0) \subset L^0, \qquad &\operatorname{ad} A(L^1) \subset L^1,\\ &\operatorname{ad} B(L^1) \subset L^0, \qquad &\operatorname{ad} B(L^0) \subset L^1. \end{aligned}$$

It follows readily (using a basis of $\mathfrak{g}$ which is the union of bases for L^0 and L^1) that $\operatorname{tr}(\operatorname{ad} A \operatorname{ad} B) = 0$, as required. □

We deduce at once from this that in terms of a basis for $\mathfrak{g}$ of the kind just mentioned, the Killing form on $\mathfrak{g}$ has the form

$$(g_{ab}) = \begin{pmatrix} g^{(0)}_{\alpha\beta} & 0 \\ 0 & g^{(1)}_{\gamma\delta} \end{pmatrix}, \tag{10}$$

where α, β range over the indices of the basis for L^0, and γ, δ over the indices of the basis for L^1. (The form $g^{(1)}_{\gamma\delta}$ is often called the *Killing form of the symmetric space M*.)

Since the Killing form (10) on $\mathfrak{g}$ satisfies (9), so also does the form $g^{(0)}_{\alpha\beta}$ on L^0. Hence by the proof of Lemma 6.4.3, if L^0 is simple then the form $g^{(0)}_{\alpha\beta}$ will be a constant multiple of the Killing form on the algebra L^0. However, in the important examples the algebra L^0 is not simple, but rather semisimple of the form $L^0 = L^0_1 \oplus L^0_2$ where L^0_1 and L^0_2 are simple. From Lemma 6.5.1 (with now L^0 in the role of $\mathfrak{g}$), we see that in this situation the restrictions of the form $g^{(0)}_{\alpha\beta}$ to the factors L^0_1, L^0_2 will be constant multiples (by λ_1, λ_2 say) of the Killing forms on those factors.

It can be seen that if H is compact and the metric $g^{(1)}_{\gamma\delta}$ is positive definite, then with respect to suitable bases for L^0 and L^1 the matrices ad A, $A \in L^0$, are skew-symmetric. Hence $\langle A, A\rangle = -\operatorname{tr}(\operatorname{ad} A)^2$ is positive, and therefore in view of

$$-\operatorname{tr}(\operatorname{ad} A)^2 = -[\operatorname{tr}(\operatorname{ad} A)^2_{L^0} + \operatorname{tr}(\operatorname{ad} A)^2_{L^1}],$$

it follows that

$$\langle A, A\rangle_{\mathfrak{g}} > \langle A, A\rangle_{L^0}. \tag{11}$$

Hence for compact H (and positive-definite metric on the symmetric space M) the restriction to L^0 (namely $g^0_{\alpha\beta}$) of the Killing form on the Lie algebra $\mathfrak{g}$, is positive definite. (Cf. the fact that the Killing form on the Lie algebra of a compact Lie group (e.g. on the Lie algebra L^0 of H) is non-negative.)

We see that in order to construct a symmetric space it essentially suffices to choose a suitable subalgebra L^0 of $\mathfrak{g}$ on which the restriction of the Killing form of the enveloping algebra $\mathfrak{g}$ is non-degenerate; then L^1 is defined as the orthogonal complement of L^0 in $\mathfrak{g}$. However the inequality (11) greatly restricts the choice of L^0: If the Killing form on $\mathfrak{g}$ is indefinite (type II) then for symmetric spaces with Riemannian metric the subalgebra $L^0 \subset \mathfrak{g}$ must be such that the restriction of the Killing form to its orthogonal complement is either positive or negative definite, and at the same time L^0 must be the Lie algebra of a compact group, and therefore of a subgroup of $SO(m)$.

Remark. A given symmetric space can be realized as a submanifold of the group G in such a way that the geodesics of M are geodesics also in the manifold G. This embedding can be obtained in any one of the following three (equivalent) ways:

(i) by considering all one-parameter subgroups of G emanating from the identity in the direction of vectors $B \in L^1$ (show that these geodesics sweep out a submanifold of G diffeomorphic to M);
(ii) via the map $\varphi: M \to G$, defined by $\varphi(x) = s_{x_0}^{-1} s_x$ (where s_{x_0}, s_x are the appropriate symmetries);
(iii) by means of an "involution" $\bar{\sigma}: G \to G$ (by which we mean an anti-automorphism of the group $(\bar{\sigma}(g_1 g_2) = \bar{\sigma}(g_2)\bar{\sigma}(g_1))$ such that the map

induced on $\mathfrak{g}$ is the identity map on L^0 and the negative of the identity map on L^1); $M \subset G$ is then the image under the map $g \mapsto g\bar{\sigma}(g^{-1})$.

EXERCISE
Show the equivalence of these embeddings.

What follows is a list of the basic examples of connected symmetric spaces of type I. (As an exercise, find in each case the corresponding direct decomposition $\mathfrak{g} = L^0 + L^1$.)

(1) $SO(2n)/U(n)$.
(2) $SU(n)/SO(n)$.
(3) $SU(2n)/Sp(n)$.
(4) $Sp(n)/U(n)$.
(5) $SO(p+q)/(SO(p) \times SO(q))$.
(6) $SU(p+q)/(SU(p) \times U(q))$.
(7) $Sp(p+q)/(Sp(p) \times Sp(q))$.

(5)–(7): Grassmannian manifolds (including the projective spaces and spheres).

The following are examples of symmetric spaces of type II (with positive-definite metric). (The simply-connected ones among such spaces turn out to have the topology of Euclidean space $\mathbb{R}^n$.)

(1) $SO(p, q)/(SO(p) \times SO(q))$. (For $q = 1$ this is the Lobachevsky space L^p.)
(2) $SU(p, q)/(U(p) \times SU(q))$. (For $q = 1$ this is the unit ball in $\mathbb{C}^p$, as a complex manifold; if also $p = 1$, this manifold is identifiable with $L^2 \cong SU(1, 1)/U(1)$.)
(3) $Sp(p, q)/(Sp(p) \times Sp(q))$.
(4) $SL(n, \mathbb{R})/SO(n)$.
(5) $SL(n, \mathbb{C})/SU(n)$.
(6) $SO(n, \mathbb{C})/SO(n, \mathbb{R})$.

We conclude with a list of symmetric spaces of dimension 4 with metric of signature $(+ - - -)$. (These spaces are of potential importance for the general theory of relativity since (by Corollary 6.4.5) the metric g_{ab} satisfies the equation $R_{ab} - \lambda g_{ab} = 0$ (see §37.4 of Part I).

I. Spaces of constant curvature with isotropy group $H = SO(1, 3)$:

(1) Minkowski space $\mathbb{R}^4_{1,3}$.
(2) The de Sitter space $S_+ = SO(1, 4)/SO(1, 3)$; note that S_+ is homeomorphic to $\mathbb{R} \times S^3$. Here the curvature tensor R is the identity operator on the space of bivectors $\Lambda^2(\mathbb{R}^4)$: $R = 1$.
(3) The de Sitter space $S_- = SO(2, 3)/SO(1, 3)$; this space is homeomorphic to $S^1 \times \mathbb{R}^3$, and its "universal covering space" $\tilde{S}_- = \widetilde{SO}(2, 3)/\widetilde{SO}(1, 3)$ (see §18) is homeomorphic to $\mathbb{R}^4$. Here the curvature tensor $R = -1$.

II. Reducible spaces (products of spaces of constant curvature):

(1) $H = SO(3)$; $M = \mathbb{R}_+ \times M^3_{---}$, where M^3_{---} is a space of constant curvature, with signature $(- - -)$.
(2) $H = SO(1, 2)$; $M = \mathbb{R}_- \times M^3_{+--}$, where M^3_{+--} is a space of constant curvature with signature $(+ - -)$.
(3) $H = SO(2) \times SO(1, 1)$; $M = M^2_{--} \times M^2_{+-}$, the product of two 2-dimensional spaces of constant curvatures.

III. The symmetric spaces M_t of plane waves. (For these the isotropy group is abelian, and the isometry group is soluble.) In terms of a certain system of global co-ordinates the metric has the form

$$dl^2 = 2\, dx_1\, dx_4 + \underbrace{[(\cos t)x_2^2 + (\sin t)x_3^2]}_{K}\, dx_4^2 + dx_2^2 + dx_3^2,$$

$$\cos t \geq \sin t.$$

In terms of the tetrad (see §30.1 of Part I) given by the 1-forms

$$p = dx_1, \qquad q = dx_1 + K\, dx_4, \qquad x = dx_3, \qquad y = dx_4,$$

the curvature tensor is constant, of the form

$$R = -4[\cos t(p \wedge x) \otimes (p \wedge x) + \sin t(p \wedge y) \otimes (p \wedge y)].$$

Remarks. 1. A simply connected symmetric space is uniquely determined by its curvature tensor at a point. To see this let $R: \Lambda^2(V) \to \Lambda^2(V)$ be the curvature tensor, and denote by $\mathfrak{h}$ the Lie algebra of skew-symmetric linear operators on the space V generated by those operators of the form $R(x, y)$, $x, y \in V$. (Then $\mathfrak{h}$ is the Lie algebra of the isotropy group (previously denoted by L^0).) Let $\mathfrak{g}$ denote the Lie algebra $V + \mathfrak{h}$, where the commutator operation on this direct sum of spaces is defined by

$$[(u, a), (v, b)] = (av - bu, [a, b] + R(u, v)).$$

Then in terms of the pair $\mathfrak{g}$, $\mathfrak{h}$ the structure of the original symmetric space is naturally reproduced on the symmetric space $M = G/H$.

2. The problem of classifying all curvature tensors of symmetric spaces with a given isotropy group H reduces to that of finding the H-invariant tensors R of the type of the curvature tensor, for which $R(x, y)$ belongs to the Lie algebra of H for all x, y in V.

6.6. Exercises

1. Show that for symmetric spaces of type II with positive-definite metric, the dimension of the subalgebra L^0 of $\mathfrak{g}$ is equal to the number of positive squares in the Killing form on $\mathfrak{g}$.

2. Show that in the complex case (e.g. where $G = SL(n, \mathbb{C})$ or $SO(n, \mathbb{C})$) the numbers of negative and positive square terms in the complex Lie algebra $\mathfrak{g}$ are equal, and $\dim L^0 = \frac{1}{2} \dim \mathfrak{g}$. Find the subalgebra L^0 of the Lie algebra $\mathfrak{g}$ of the group $G = SL(n, \mathbb{C})$.

3. Show that for symmetric spaces of type II with positive-definite metric one always has $M \cong G/H$, where H is a maximal compact subgroup of G. Investigate the particular cases $SL(n, \mathbb{R})/SO(n)$, $SL(n, \mathbb{C})/SU(n)$.

4. Show that a simply-connected, symmetric space of type II always has the topology of Euclidean $\mathbb{R}^n$.

For the next few exercises, note that, as for Lie groups (see §30.3 of Part I), so also for symmetric spaces do we have

$$\langle R(\xi, \eta)\zeta, \tau\rangle|_{x_0} = \tfrac{1}{4}\langle[\xi, \eta], [\zeta, \tau]\rangle_{L^0}, \qquad \xi, \eta, \zeta, \tau \in \mathbb{R}^n_{x_0} = L^1.$$

5. Show that for spaces of type I, the Ricci tensor R_{ab} is positive definite, and the "sectional curvature" $\langle R(\xi, \eta)\xi, \eta\rangle$ (where ξ, η span a parallelogram of unit area) is non-negative.

6. Show that for spaces of type II the sectional curvature is non-positive. Deduce from this that a simply-connected, symmetric space of type II is topologically the same as $\mathbb{R}^n$ (assuming the metric Riemannian).

7. Decide which of the 7 symmetric spaces of type I and 6 spaces of type II listed above have non-vanishing sectional curvature. Investigate the spaces S^n, $\mathbb{C}P^n$, $\mathbb{H}P^n$ of type I, and the spaces L^n, $SU(n, 1)/U(n)$, $SL(n, \mathbb{R})/SO(n)$, $SL(n, \mathbb{C})/SU(n)$ of type II.

8. Prove that in dimensions $n = 2, 3$ the only simply-connected, symmetric spaces with positive-definite metric are L^n, S^n, $\mathbb{R}^n$. (*Hint.* Show that the isotropy group $H \subset G$ must be $SO(n)$ $(n = 2, 3)$, and thence deduce (for $n = 3$) the constancy of all sectional curvatures.)

9. Prove that a simply-connected symmetric space M with semisimple $G = G_1 \times \cdots \times G_k$ (where the G_i are simple) has the form

$$M = (G_1/H_1) \times \cdots \times (G_k/H_k),$$

with the metric decomposing as a direct product of metrics on the factors $M_i = G_i/H_i$, each of which is proportional to the Killing metric on the subspace L_i^1 of the corresponding Lie algebra $\mathfrak{g}_i = L_i^0 + L_i^1$.

§7. Vector Bundles on a Manifold

7.1. Constructions Involving Tangent Vectors

From any n-dimensional manifold M we can construct a $2n$-dimensional manifold, called the *tangent bundle* $L(M)$ of M as follows. The points of the manifold $L(M)$ are defined to be the pairs (x, ξ) where x ranges over the points

@ 0 " #%12# # #3 4 - # , F .-
- , ,- 4 7 L (7 6 ! 0 ,C(

@ 0' #%12# # #3 4 - # , F .-
- , ,- 4 7 L (7 6 ! 0 .,
: .-4
3< 8 (, EEE

54 6 !0 " -. .- / ,

54 @ 6 0 " -. . / - 8 .-
6 .- - D 4 , 4 - 7 . 0

54 !6 @0 " .- .- " (H .

TABLE OF CONTENTS

Preface

Table of Contents

Chapter I. METRIC SPACES WITH GEODESICS 1
- §1. Metric Spaces; Notations 1
- §2. The Basic Axioms 11
- §3. Geodesics 17
- §4. Topological Structure of One- and Two-dimensional Spaces With Axioms A - D 24

Chapter II. METRIC CONDITIONS FOR FINSLER SPACES . . . 30
- §1. Convex Surfaces and Minkowski Metrics 31
- §2. Riemann Spaces and Finsler Spaces 40
- §3. Condition $\Delta(P)$ and the Definition of the Local Metric 47
- §4. Equivalence of the Local Metric with the Original Metric, and its Convexity 53
- §5. The Minkowskian Character of the Local Metric . 57
- §6. The Continuity of the Local Metric 63

Chapter III. PROPERTIES OF GENERAL S. L. SPACES . . . 72
(Spaces with a unique geodesic through any two points)
- §1. Axiom E. Shape of the Geodesics 73
- §2. Two Dimensional S.L. Spaces 79
- §3. The Inverse Problem for the Euclidean Plane . 89
- §4. Asymptotes and Limit Spheres 98
- §5. Examples on Asymptotes and Limit Spheres. The Parallel Axioms 105
- §6. Desarguesian Spaces 113

Chapter IV. SPACES WITH CONVEX SPHERES. 119
§1. The Convexity Condition 120
§2. Characterization of the Higher Dimensional Elliptic Geometry 124
§3. Perpendiculars in Spaces with Spheres of Order 2 132
§4. Perpendiculars and Baselines in Open S. L. Spaces 139
§5. Definition and Properties of Limit Bisectors. 146
§6. Characterizations of the Higher Dimensional Minkowskian and Euclidean Geometries 154
§7. Plane Minkowskian Geometries 160
§8. Characterization of Absolute Geometry . . . 168

Chapter V. MOTIONS 175
§1. Definition of Motions. Involutoric Motions in S. L. Spaces 176
§2. Free Movability 184
§3. Example of a Non-homogeneous Riemann Space in which Congruent Pairs of Points Can be Moved into Each Other 192
§4. Translations Along g and the Asymptotes to g 198
§5. Quasi-hyperbolic Metrics 208
§6. Translations Along Non-parallel Lines and in Closed Planes 214
§7. Plane Geometries with a Transitive Group of Motions 220
§8. Transitive Abelian Groups of Motions in Higher Dimensional Spaces 228
§9. Some Problems Regarding S. L. Spaces and Other Spaces 232

Literature 235

@ 0 " #%12# # #3 4 - # , 4 7 , - ,
-.. - - . 4 -4 ,C(

@ 0 " #%12# # #3 4 - # , 4 7 , - ,
-.. - - . 4 -4 .,
: .-4
: 9 47 - , " 47

@ 0' #%12# # #3 4 - # . A 4 4
- > - O & / 7 . 4 L 6 ! 0 .,
: .-4

Contents

1. **Kinematics of a Particle**

1.1. Introduction 3
1.2. Primitive Terms 4
1.3. Motion and Particle Path 6
1.4. Velocity and Acceleration 9
1.5. Some Basic Classifications of Problems 15
1.6. Uniform Motion 22
1.7. Velocity and Acceleration Referred to an Intrinsic Frame 26
  1.7.1. Construction of the Intrinsic Velocity and Acceleration 26
  1.7.2. Curvature, the Radius of Curvature, and the Intrinsic Frame 29
  1.7.3. Velocity and Acceleration Referred to the Intrinsic Frame 31
  1.7.4. Some Applications of the Intrinsic Velocity and Acceleration 34
1.8. Summary of Particle Kinematics 40
1.9. Special Topics 42
  1.9.1. Some Additional Properties of the Intrinsic Basis 42
  1.9.2. The Hodograph 46
  1.9.3. Singularity Functions in Particle Kinematics 47
References 65
Problems 65

2. **Kinematics of Rigid Body Motion**

2.1. Introduction 85
2.2. Displacements of a Rigid Body 86
2.3. Rotation about a Fixed Line 87
2.4. The Imbedded Reference Frame 91
2.5. The General Displacement of a Rigid Body 92
2.6. Infinitesimal Displacement of a Rigid Body 94
2.7. Composition of Infinitesimal Rotations 95
2.8. Velocity and Acceleration of Points of a Rigid Body 98
  2.8.1. Velocity of a Particle of a Rigid Body 98
  2.8.2. Acceleration of a Particle of a Rigid Body 101

2.9. Some Applications of the Basic Equations 102
  2.9.1. Rolling without Slip 109
  2.9.2. Analysis of a Spatial Mechanism 117
  2.9.3. Velocity Referred to a Moving Reference Frame 121
2.10. A Basic Invariant Property of the Angular Velocity Vector 123
2.11. Chasles' Theorem on Screw Motions 126
References 128
Problems 129

3. Finite Rigid Body Displacements

3.1. Introduction 151
3.2. Index Notation 152
  3.2.1. The Delta and Permutation Symbols and the Summation Rule . 153
  3.2.2. Some Additional Applications of Index Notation 154
3.3. Introduction to Tensors 155
  3.3.1. Components of a Tensor 156
  3.3.2. The Tensor Product 158
  3.3.3. The Tensor Representation Theorem 159
  3.3.4. Other Tensor Operations and Rules 161
3.4. The Rotator 168
3.5. The Rotation Tensor 173
3.6. Change of Basis and Transformation Laws 175
  3.6.1. A Change of Basis 175
  3.6.2. Transformation Laws for Vectors and Tensors 179
3.7. Rotation about a Fixed Point 183
3.8. Euler's Theorem 186
3.9. Fundamental Invariant Property of the Rotator 192
3.10. The Parallel Axis Theorem 194
  3.10.1. The Center of Rotation 196
3.11. Chasles' Screw Displacement Theorem 197
3.12. Composition of Finite Rotations 202
  3.12.1. Euler Angles and Rotations 207
  3.12.2. Review of the Composition of Infinitesimal Rotations 209
  3.12.3. Composition of Rotations about Nonintersecting Axes 210
  3.12.4. Composition of Screw Displacements 213
References 214
Problems 215

4. Motion Referred to a Moving Reference Frame and Relative Motion

4.1. Introduction 229
4.2. An Introductory Example 231
4.3. Derivative of a Vector Referred to a Moving Reference Frame 232
4.4. Kinematic Chain Rule for Angular Velocity Vectors 240
  4.4.1. The Simple Relative Angular Velocity Rule 241
  4.4.2. The Composition Rule for Several Angular Velocity Vectors 242
4.5. The Composition Rule for Several Angular Accelerations 251
  4.5.1. The Simple Relative Angular Acceleration Rule 252
  4.5.2. The Composition of Multiple Angular Acceleration Vectors 252

- 4.6. Velocity and Acceleration Referred to a Moving Frame 260
  - 4.6.1. The Velocity Equation 261
  - 4.6.2. The Acceleration Equation 262
- 4.7. Simple Relative Motion 264
- 4.8. Velocity and Acceleration in Special Curvilinear Coordinates 267
  - 4.8.1. The Cylindrical Reference Frame 267
  - 4.8.2. The Spherical Reference Frame 277
- 4.9. More Examples of Motion Referred to a Moving Frame 282
- 4.10. Motion Referred to an Earth Frame 295
  - 4.10.1. Motion of the Earth 296
  - 4.10.2. Velocity and Acceleration Relative to the Earth 298
- 4.11. Summary of the Principal Equations 303
- 4.12. Some Special Advanced Topics 304
  - 4.12.1. Application of the Vector Component Transformation Law 305
  - 4.12.2. Derivative of a Tensor Function of a Scalar Variable 306
  - 4.12.3. Derivative of a Tensor Referred to a Moving Frame 311
- References 318
- *Problems* 318

Appendix A: The Elements of Vector Calculus

- A.1. Vector Algebra 353
  - A.1.1. The Dot Product 355
  - A.1.2. The Cross Product 356
  - A.1.3. Triple Products 357
- A.2. Derivative of a Vector Function of a Scalar Variable 357
- A.3. Integration of a Vector Function of a Scalar Variable 362
- A.4. Maximum and Minimum Vector Values 362
- *Problems* 364

Appendix B: The Elements of Matrix Algebra

- B.1. Matrix Notation 365
- B.2. The Elementary Operations 366
  - B.2.1. Equality of Matrices 367
  - B.2.2. Multiplication by a Scalar 367
  - B.2.3. Addition of Matrices 367
  - B.2.4. Multiplication of Matrices 368
  - B.2.5. Transpose of a Matrix 370
  - B.2.6. Trace of a Square Matrix 371
- B.3. Determinant of a Matrix 371
  - B.3.1. Basic Properties of Determinants 372
  - B.3.2. Cofactors of a Determinant 373
- B.4. The Inverse of a Matrix 374
  - B.4.1. Properties of Inverse Matrices 376
  - B.4.2. Orthogonal Matrices 376
- References 377
- *Problems* 378

Answers to Selected Problems 381

Index 387

" #$%&% ' ((#) * # @# . - , .-

@@0#%12# # #3 4 -# @ . -, .- & ; =
& . "- -, 1 - (F , "-< ,4 L (7 6 ! 0 .,
: .-4

Contents

Preface vii

1 **The geometry of world-vectors and spin-vectors** 1
1.1 Minkowski vector space 1
1.2 Null directions and spin transformations 8
1.3 Some properties of Lorentz transformations 24
1.4 Null flags and spin-vectors 32
1.5 Spinorial objects and spin structure 41
1.6 The geometry of spinor operations 56

2 **Abstract indices and spinor algebra** 68
2.1 Motivation for abstract-index approach 68
2.2 The abstract-index formalism for tensor algebra 76
2.3 Bases 91
2.4 The total reflexivity of $\mathfrak{S}^{\cdot}$ on a manifold 98
2.5 Spinor algebra 103

3 **Spinors and world-tensors** 116
3.1 World-tensors as spinors 116
3.2 Null flags and complex null vectors 125
3.3 Symmetry operations 132
3.4 Tensor representation of spinor operations 147
3.5 Simple propositions about tensors and spinors at a point 159
3.6 Lorentz transformations 167

4 **Differentiation and curvature** 179
4.1 Manifolds 179
4.2 Covariant derivative 190
4.3 Connection-independent derivatives 201
4.4 Differentiation of spinors 210
4.5 Differentiation of spinor components 223
4.6 The curvature spinors 231
4.7 Spinor formulation of the Einstein–Cartan–Sciama–Kibble theory 237

4.8 The Weyl tensor and the Bel–Robinson tensor 240
4.9 Spinor form of commutators 242
4.10 Spinor form of the Bianchi identity 245
4.11 Curvature spinors and spin-coefficients 246
4.12 Compacted spin-coefficient formalism 250
4.13 Cartan's method 262
4.14 Applications to 2-surfaces 267
4.15 Spin-weighted spherical harmonics 285

5 Fields in space–time 312
5.1 The electromagnetic field and its derivative operator 312
5.2 Einstein–Maxwell equations in spinor form 325
5.3 The Rainich conditions 328
5.4 Vector bundles 332
5.5 Yang–Mills fields 342
5.6 Conformal rescalings 352
5.7 Massless fields 362
5.8 Consistency conditions 366
5.9 Conformal invariance of various field quantities 371
5.10 Exact sets of fields 373
5.11 Initial data on a light cone 385
5.12 Explicit field integrals 393
Appendix: diagrammatic notation 424
References 435
Subject and author index 445
Index of symbols 457

@ 0' #%12# # #3 4 - # @ . % , .- & ;
= . % , & , + .- & / 7 "- < ,4 L (7 6 .,
: .-4

Contents

Preface vii
Summary of Volume 1 1

6 Twistors 43
6.1 The twistor equation and its solution space 43
6.2 Some geometrical aspects of twistor algebra 58
6.3 Twistors and angular momentum 68
6.4 Symmetric twistors and massless fields 75
6.5 Conformal Killing vectors, conserved quantities and exact sequences 82
6.6 Lie derivatives of spinors 101
6.7 Particle constants; conformally invariant operators 104
6.8 Curvature and conformal rescaling 120
6.9 Local twistors 127
6.10 Massless fields and twistor cohomology 139

7 Null congruences 169
7.1 Null congruences and spin-coefficients 169
7.2 Null congruences and space–time curvature 182
7.3 Shear-free ray congruences 189
7.4 SFRs, twistors and ray geometry 199

8 Classification of curvature tensors 223
8.1 The null structure of the Weyl spinor 223
8.2 Representation of the Weyl spinor on S^+ 226
8.3 Eigenspinors of the Weyl spinor 233
8.4 The eigenbivectors of the Weyl tensor and its Petrov classification 242
8.5 Geometry and symmetry of the Weyl curvature 246
8.6 Curvature covariants 258
8.7 A classification scheme for general spinors 265
8.8 Classification of the Ricci spinor 275

9 Conformal infinity 291
9.1 Infinity for Minkowski space 291
9.2 Compactified Minkowski space 297
9.3 Complexified compactified Minkowski space and twistor geometry 305
9.4 Twistor four-valuedness and the Grgin index 316
9.5 Cosmological models and their twistors 332
9.6 Asymptotically simple space–times 347
9.7 Peeling properties 358
9.8 The BMS group and the structure of $\mathscr{I}^+$ 366
9.9 Energy–momentum and angular momentum 395
9.10 Bondi–Sachs mass loss and positivity 423

Appendix: spinors in *n* dimensions 440
References 465
Subject and author index 481
Index of symbols 499

@!0 " #%12# # #3 4 - # @ . - , .-
"- < ,4 L (7 6 ! 0 ,C(

0 " #%12# # #3 4 - # @ . - , .-
"- < ,4 L (7 6 ! 0 ,C(

0 " #%12# # #3 4 - # @ G 4- 4 1 , 1 - (7
. - ; 4 - ; - , 4 - 3H , L (7 ; L % 6 0 .,
: @ .-4

" #$%&% ' ((#) * #

" #$%&% ' ((#) * # #$ F - - "- - ,
F - - $ AD -

0 " #%12# # #3 4 - # $ F - - "- - ,
F - - $ AD - .,

" #$%&% ' ((#) * # #F ,- - % - 7
%.. -

0 " #%12# # #3 4 - # ! F ,- - % - 7
%.. - .,
: ! .-4

& < * - - 4 -4
- -
/- /- 6 0

3< 8 (, EEE

54 ! " -. & . 47 1 - 8

Contents

1 Sets, Functions, and Cardinality 1
1.1 Naive Set Theory 2
1.2 Relation and Ordering 9
1.3 Functions 17
1.4 Cardinality 26
1.5 Development of Function Concept—A Historical Note 37
References 51

2 The Real Numbers 53
2.1 Ordered Field $\mathbb{Q}$ 54
2.2 The Complete Ordered Field 65
2.3 Modulus Metric 78
2.4 Countable and Uncountable Sets in $\mathbb{R}$ 81
References 87

3 Sequences and Series of Numbers 89
3.1 The Limit of a Sequence 91
3.2 Algebra of Convergent Sequences 103
3.3 Convergence Theorems 107
3.3.1 Cauchy Sequences 113
3.3.2 The Existence of the Set $\mathbb{R}$ 118
3.4 Infinite Series 120
3.4.1 Tests of Convergence 124
3.4.2 Arbitrary Terms Series 129
3.4.3 Decimal Expansion of Real Numbers 133
Reference 140

4 Analysis on $\mathbb{R}$ 141
4.1 Limit and Continuity 142
4.2 Differentiability 153
4.3 Riemann Integration 159
4.3.1 Darboux Integral 161

4.3.2 Examples and Important Theorems 167
4.3.3 Properties of Riemann Integral 179
Reference 188

5 Topology of the Real Line 189
5.1 Open and Closed Set 190
5.2 Compactness 198
5.3 Connectedness 205

6 Metric Spaces 211
6.1 Some Important Metric Spaces 212
6.1.1 The Euclidean Space $\mathbb{R}^n$ 216
6.1.2 Function Spaces 225
6.1.3 Metric Subspace 230
6.2 Topology of Metric Spaces 231
6.2.1 Open Sets 234
6.2.2 Limit Points and Closed Sets 240
6.3 Convergence and Completeness 243
6.4 Compactness 252
6.5 Connectedness 258
6.5.1 Path-Connectedness 265
6.5.2 Components 267
Reference 272

7 Multivariable Analysis 273
7.1 Limit of Functions 274
7.2 Continuity of a Function 277
7.3 Differentiability 287
7.3.1 Partial Derivatives 288
7.3.2 Directional Derivatives 291
7.3.3 The Differential 292
7.4 Geometry of Curves and Surfaces 307
7.5 Two Fundamental Theorems 320
References 331

8 Sequences and Series of Functions 333
8.1 Pointwise Convergence 334
8.2 Uniform Convergence 338
8.3 Power Series 343

9 Measure and Integration 349
9.1 Measure Space 350
9.1.1 Outer Measure 361
9.1.2 Null Sets and Completion 366
9.2 Lebesgue Measure 368
9.2.1 Regularity Property 375
9.3 Lebesgue Integration 377

9.3.1 Measurable Functions 378
9.3.2 Lebesgue Integral 385
9.4 Fundamental Convergence Theorems 396
9.5 L^p Spaces 399
References 403

10 Fourier Series 405
10.1 Evolution of Modern Mathematics 407
10.2 Definitions and Examples 412
10.2.1 Sine and Cosine Fouries Series 422
10.2.2 Half-Range Fourier Series 426
10.3 Convergence Issues 430
10.3.1 Dirichlet Theorem 432
10.3.2 Gibbs Phenomenon 437
10.3.3 Convergence in Mean 439
10.3.4 Uniform Convergence 442
10.4 An Application to Infinite Series 446
References 449

Appendix: Mathematical Logic 451

Index 467

" #$%&% ' ((#) * # # .- % " .- % - 7

0 " #%12# # #3 4 - # .- .,
: .-4

Contents

1 Metric Spaces 1
- 1.1 Metrics 1
  - 1.1.1 Rationale for Metrics 1
  - 1.1.2 Defining Metric Space 2
  - 1.1.3 Exercises 5
- 1.2 Examples of Metric Spaces 6
  - 1.2.1 Normed Spaces 6
  - 1.2.2 Subspaces 8
  - 1.2.3 Examples; Not Subspaces of Normed Spaces 9
  - 1.2.4 Pseudometrics 9
  - 1.2.5 Cauchy-Schwarz, Hölder, Minkowski 11
  - 1.2.6 Exercises 14
- 1.3 Cantor's Middle Thirds Set 16
  - 1.3.1 Exercises 19
- 1.4 The Normed Spaces of Functional Analysis 20
  - 1.4.1 Sequence Spaces 20
  - 1.4.2 Function Spaces 22
  - 1.4.3 Spaces of Continuous Functions 23
  - 1.4.4 Spaces of Integrable Functions 23
  - 1.4.5 Hölder's and Minkowski's Inequalities for Integrals 24
  - 1.4.6 Exercises 25

2 Basic Theory of Metric Spaces 29
- 2.1 Balls in a Metric Space 29
  - 2.1.1 Limit of a Convergent Sequence 30
  - 2.1.2 Uniqueness of the Limit 31
  - 2.1.3 Neighbourhoods 32
  - 2.1.4 Bounded Sets 32
  - 2.1.5 Completeness; a Key Concept 32
  - 2.1.6 Exercises 35

2.2 Open Sets, and Closed 37
  2.2.1 Open Sets 38
  2.2.2 Union and Intersection of Open Sets 39
  2.2.3 Closed Sets 39
  2.2.4 Union and Intersection of Closed Sets 40
  2.2.5 Characterisation of Open and Closed Sets by Sequences ... 41
  2.2.6 Interior, Closure and Boundary 42
  2.2.7 Limit Points of Sets 43
  2.2.8 Characterisation of Closure by Limit Points 43
  2.2.9 Subspaces 45
  2.2.10 Open and Closed Sets in a Subspace 45
  2.2.11 Exercises 47
2.3 Continuous Mappings 52
  2.3.1 Defining Continuity 52
  2.3.2 New Views of Continuity 53
  2.3.3 Limits of Functions 55
  2.3.4 Characterising Continuity by Sequences 57
  2.3.5 Lipschitz Mappings 57
  2.3.6 Examples of Continuous Functions 58
  2.3.7 Exercises 59
2.4 Continuity of Linear Mappings 63
  2.4.1 Continuity Criterion 64
  2.4.2 Operator Norms 66
  2.4.3 Exercises 68
2.5 Homeomorphisms and Topological Properties 70
  2.5.1 Equivalent Metrics 72
  2.5.2 Exercises 73
2.6 Topologies and σ-Algebras 74
  2.6.1 Order Topologies 77
  2.6.2 Exercises 79
  2.6.3 Pointers to Further Study 82
2.7 (◊) Mazur-Ulam 82
  2.7.1 Exercises 84

3 Completeness of the Classical Spaces 87
3.1 Coordinate Spaces and Normed Sequence Spaces 87
  3.1.1 Completeness of $\mathbb{R}^n$ 87
  3.1.2 Completeness of ℓ^p 88
  3.1.3 Exercises 90
3.2 Product Spaces 90
  3.2.1 Finitely Many Factors 91
  3.2.2 Infinitely Many Factors 91
  3.2.3 The Space $2^{\mathbb{N}_+}$ and the Cantor Set 93
  3.2.4 Subspaces of Complete Spaces 95
  3.2.5 Exercises 96

3.3 Spaces of Continuous Functions 99
 3.3.1 Uniform Convergence 99
 3.3.2 Series in Normed Spaces 101
 3.3.3 The Weierstrass M-Test 103
 3.3.4 The Spaces $C(\mathbb{R})$ and $C^p(\mathbb{R})$ 103
 3.3.5 Exercises 105
3.4 (◊) Rearrangements 108
 3.4.1 Vector Series 108
 3.4.2 Exercises 111
 3.4.3 Pointers to Further Study 111
3.5 (◊) Invertible Operators 112
 3.5.1 Fredholm Integral Equation 115
 3.5.2 Exercises 116
 3.5.3 Pointers to Further Study 118
3.6 (◊) Tietze 118
 3.6.1 Formulas for an Extension 121
 3.6.2 Exercises 121
 3.6.3 Pointers to Further Study 123

4 Compact Spaces 125
4.1 Sequentially Compact Spaces 125
 4.1.1 Continuous Functions on Sequentially Compact Spaces 126
 4.1.2 Bolzano-Weierstrass in $\mathbb{R}^n$ 126
 4.1.3 Sequentially Compact Sets in $\mathbb{R}^n$ 127
 4.1.4 Sequentially Compact Sets in Other Spaces 127
 4.1.5 The Space $C(M)$ 128
 4.1.6 Exercises 129
4.2 The Correct Definition of Compactness 130
 4.2.1 Thoughts About the Definition 131
 4.2.2 Compact Spaces and Compact Sets 132
 4.2.3 Continuous Functions on Compact Spaces 134
 4.2.4 Uniform Continuity 135
 4.2.5 Exercises 136
4.3 Equivalence of Compactness and Sequential Compactness 137
 4.3.1 Relative Compactness 141
 4.3.2 Local Compactness 142
 4.3.3 Exercises 143
4.4 Finite Dimensional Normed Vector Spaces 150
 4.4.1 Exercises 153
4.5 (◊) Ascoli 154
 4.5.1 Peano's Existence Theorem 157
 4.5.2 Exercises 160
 4.5.3 Pointers to Further Study 162

5 Separable Spaces 163
- 5.1 Dense Subsets of a Metric Space 163
  - 5.1.1 Defining a Vector-Valued Integral 165
  - 5.1.2 Exercises 168
- 5.2 Separability 171
  - 5.2.1 Second Countability 172
  - 5.2.2 Exercises 173
- 5.3 (◊) Weierstrass 176
  - 5.3.1 Exercises 180
  - 5.3.2 Pointers to Further Study 182
- 5.4 (◊) Stone-Weierstrass 182
  - 5.4.1 Exercises 185
  - 5.4.2 Pointers to Further Study 186

6 Properties of Complete Spaces 187
- 6.1 Cantor's Nested Intersection Theorem 187
  - 6.1.1 Categories 188
  - 6.1.2 Exercises 190
- 6.2 (◊) Genericity 193
  - 6.2.1 Exercises 196
  - 6.2.2 Pointers to Further Study 199
- 6.3 (◊) Nowhere Differentiability 199
  - 6.3.1 Exercises 202
  - 6.3.2 Pointers to Further Study 203
- 6.4 Fixed Points 203
  - 6.4.1 Exercises 204
- 6.5 (◊) Picard 206
  - 6.5.1 Exercises 209
- 6.6 (◊) Zeros 209
  - 6.6.1 Exercises 213
  - 6.6.2 Pointers to Further Study 216
- 6.7 Completion of a Metric Space 216
  - 6.7.1 Other Ways to Complete a Metric Space 218
  - 6.7.2 Exercises 218

7 Connected Spaces 219
- 7.1 Connectedness 219
  - 7.1.1 Connected Sets 221
  - 7.1.2 Rules for Connected Sets 221
  - 7.1.3 Connected Subsets of $\mathbb{R}$ 223
  - 7.1.4 Exercises 223
- 7.2 Continuous Mappings and Connectedness 224
  - 7.2.1 Continuous Curves 224
  - 7.2.2 Arcwise Connectedness 225
  - 7.2.3 Exiting a Set 226
  - 7.2.4 Exercises 227

7.3 Connected Components 229
  7.3.1 Examples of Connected Components 229
  7.3.2 Arcwise Connected Components 232
  7.3.3 Exercises 232
7.4 (◊) Peano 236
  7.4.1 Exercises 236
  7.4.2 Pointers to Further Study 237

Afterword 239

Index 241

" #$%&% ' ((#) * # !

" #$%&% ' ((#) * # !#% / 7; ' < & 7;
- , " . -

0 " #%12# # #3 4 - #
=) *=% / 7' < & 7 .,
: @ .-4

" #$%&% ' ((#) * # !#A / 7

0 " #%12# # #3 4 - # " - 8 A C
/ 7 6 ! 0 .,
: .-4
: A C (/ 7

@0" #%12# # #3 4 - # A ' - (
/ 7) * M)- 6 0 .,
: ! .-4

0 " #%12# # #3 4 - # A - (
4 7) * M) 6 0 ,C(
: @ .-4

!0 " #%12# # #3 4 - # A - (
4 7) * M) 6 0 .,

! 0 " #%12# # #3 4 - # A ,; P 8 9 < 4; 1 - ,
F I.- * A ,J A / 7 1 - , F I.- * 6 0 .,

! 0 " #%12# # #3 4 - # F *)- A
$ - / 7 - , & . 47 6 - - - , %.. - 0 . 4 6 !! 0 .,
: !@ .-4

! 0 " #%12# # #3 4 - # F *)-
, - 4 7 - , . 47 . 4 6 !! 0 ,C(
: !@ .-4
3< .

! 0 " #%12# # #3 4 - # F *)-
, - 4 7 - , . 47 . 4 6 !! 0 .,

! 0 " #%12# # #3 4 - # F * , / 7
+ , 9 47 & 7 , - , %.. - .,
: .-4

! 0 " #%12# # #3 4 - # F * , / 7
, - , %.. - - & / 7 - ; & - - / . ; - ,
F , ,C(
: .-4

! 0 " #%12# # #3 4 - # F * , / 7

, - , %.. - - & / 7 - ; & - - / . ; - ,
F , .,

!@0" #%12# # #3 4 - # F * , / 7
, - , %.. - - & / 7 - , & . 47 - , . 4 -4
' ,C(
: .-4

! 0 " #%12# # #3 4 - # F * , / 7
, - , %.. - - & / 7 - , & . 47 - , . 4 -4
' .,

!!0 " #%12# # #3 4 - # F * , / 7
, - , %.. - - + , 9 47 & 7 . 4 6 !! 0 ,C(
: ! .-4

0 " #%12# # #3 4 - # F * , / 7
, - , %.. - - + , 9 47 & 7 . 4 6 !! 0 .,

0 " #%12# # #3 4 - # F * , / 7
, - , %.. - - .,

0 " #%12# # #3 4 - # ' A - -
+ - - - . 7 $ / 7 6 !@ 0 .,
: .-4

0 " #%12# # #3 4 - # ' A 3 - -
+= % + < = 9 - 6 !@ 0 .,
: .-4

0 " #%12# # #3 4 - # ' A - -
++ % 4 < - - , / 7 ++ G- $ / 7 6 !@ 0 .,
: .-4

0 " #%12# # #3 4 - # ' A - -
+++; - = - - - - - +++ < = 9 -
6 !@ 0 .,
: .-4

0 " #%12# # #3 4 - # 8 8 7 & A
" , - / 7 - - - , " 6 ! 0 ,C(
: .-4

@0 " #%12# # #3 4 - # 8 8 7 & A 3
" , - / 7 .,

" #$%&% ' ((#) * # !#+ . -

0 " #%12# # #3 4 - # A , & (
A , - - ,7 7 -< -4 , - , 4 -
- 7 / * 4 7 ,C(
: ! .-4

!0 " #%12# # #3 4 - # A , & (
A , - - ,7 7 -< -4 , - , 4 -
- 7 / * 4 7 .,

0 " #%12# # #3 4 - # & A(A , -
A % ,7 & 7 + -< -4 , - , + 4 - A- 7
/ * / 7 .,
: .-4

0 " #%12# # #3 4 - # G < 1 - , > 6- 0
& H - , % - , , (- / 7) * M)- 6 ! !0 .,
: ! .-4

" #$%&% ' ((#) * # !#+ (4- 4 % ; 9 7; - , 8 -
% 7

0 " #%12# # #3 4 - # + (4- 4 % ; 9 7; - ,
8 - % 7 .,
: .-4

" #$%&% ' ((#) * # !#8 4 B - - , & & -(

0 " #%12# # #3 4 - #) - " 44 + , 4 + 7=
% / -. / , +) * 6 0 . <

0 " #%12# # #3 4 - #) - " 44 + , 4 + 7=
% / -. / , +) * 6 0 .,

0 " #%12# # #3 4 - #) - " 44 8 4 7 - - ,
-(- H. - - * ,J , 4 - - 4 G 7 6 0 .,
: .-4

0 " #%12# # #3 4 - #) - " 44 2 - & 7= %
" - " =) % + - AH. +(7 6 !0 .,

@0 " #%12# # #3 4 - #) - " 44 & 2 - %4 =
9 7 7 - - & - , 3 8 (+) * 6 0 .,
: .-4

0 " #%12# # #3 4 - # " 44;) - 8 4 7 - G 7
6 0 .,

" #$%&% ' ((#) * # !# % L

!0 " #%12# # #3 4 - # -< @ , .,
: $- < , -..

0 " #%12# # #3 4 - # - @! , .,
: 1 . - %

0 " #%12# # #3 4 - #
-.. =,% H- , ;= , = - = =A * =8-= = - K - D .,

: .-4
3< 8 (, EEE + . - Q EEE

" #$%&% ' ((#) * # !#& A(A , - A

0 " #%12# # #3 4 - # @ A , & (
A , - - ,7 7 -< -4 , - , 4 -
- 7 / * 4 7 ,C(

0 " #%12# # #3 4 - # @ A , & (
A , - - ,7 7 -< -4 , - , 4 -
- 7 / * 4 7 .,

0 " #%12# # #3 4 - # @ & A(A , -
A % ,7 & 7 + -< -4 , - , + 4 - A- 7
/ * / 7 .,

0 " #%12# # #3 4 - # @ G < 1 - , > 6- 0
& H - , % - , , (- / 7) * M)- 6 ! !0 .,
: ! .-4

" #$%&% ' ((#) * #

" #$%&% ' ((#) * # #% "- - 4 3 . - - " (

0 " #%12# # #3 4 - # P $ 8- % "- - 4
. - - " ($ (< - 6 !@ 0 ,C(

@0 " #%12# # #3 4 - # P $ 8- % "- - 4
. - - " ($ (< - 6 !@ 0 .,

" #$%&% ' ((#) * # #% F " - - $ -
AD - G " . H - -< - , & - ,

0 " #%12# # #3 4 - # ! G < 4 % F "
- - $ - AD - " . H - -< - , & - , 6$ ( ) *
- - 0 $ (< - 6 !! 0 .,
: .-4

!0 " #%12# # #3 4 - # ! G < 4 %
.- - , - D - . H (- -< - , - , $ (< -
6 !! 0 ,C(

" #$%&% ' ((#) * # #" , F -

0 " #%12# # #3 4 - # " , - - - 7
< 7 (-.. H - - (8 P 8- 4 = , ,C(
: ! .-4

0 " #%12# # #3 4 - # " , - - - 7
< 7 (-.. H - - (8 P 8- 4 = , .,

0 " #%12# # #3 4 - # $ 4 9 7 " ,
- G , 6 0 ,C(

0 " #%12# # #3 4 - # $ 4 9 7 " ,
- G , 6 0 .,

0 " #%12# # #3 4 - # 9- ,< * " , F -
. - F . 4 6 0 .,
: .-4

0 " #%12# # #3 4 - # 9 7 " , F - - ,
-,K %.. H - . 4 -4) 9 , < 4 6 !! 0 ,C(

0 " #%12# # #3 4 - # 9 7 " , F - - ,
-,K %.. H - . 4 -4) 9 , < 4 6 !! 0 .,
: ! .-4

@0 " #%12# # #3 4 - # 3 , " , -
1- , 9 6 ! 0 ,C(
: ! .-4

0 " #%12# # #3 4 - # 3 , " , -
1- , 9 6 ! 0 .,

!0 " #%12# # #3 4 - # 3 4 - . 7 - - ,
, - "L 6 0 .,

0 " #%12# # #3 4 - # & " , -
- - 7 7 - , -.. - %,, G 7 < " 6 ! 0 ,C(

0 " #%12# # #3 4 - # & " , -
- - 7 7 - , -.. - %,, G 7 < " 6 ! 0 .,
: @ .-4

" #$%&% ' ((#) * # #" < & F , 4 &
& 7 & - ' < "-

0 " #%12# # #3 4 - # / 4 "- " <
, 4 7 - < '-< 6 0 ,C(
: @ .-4

0 " #%12# # #3 4 - # / 4 "- " <
, 4 7 - < '-< 6 0 .,

0 " #%12# # #3 4 - # / 4 "- ; . A)
P ,- " < F , 4 & 7 & - ' < "
" - ,C(

0 " #%12# # #3 4 - # / 4 "- ; . A)
P ,- " < F , 4 & 7 & - ' < "
" - .,

" #$%&% ' ((#) * # #F ,- - /
$ 4 - , & - 4

0 " #%12# # #3 4 - # % H > * * F ,- -
/ $ 4 - , & - 4 $ - " 4-4 8 - 4 6 !!@0 .,
: @ .-4

@0" #%12# # #3 4 - #
F ,- - / $ 4 - , & - 4 .,
: .-4

" #$%&% ' ((#) * # #/ $

0 " #%12# # #3 4 - # F R /
& - - & - ; - . ; -4 $ E G ,
6 @0 . <

!0 " #%12# # #3 4 - # F R /
& - - & - ; - . ; -4 $ E G ,
6 @0 .,

0 " #%12# # #3 4 - # 9- 7 8 ,4 ; / 4 F , *
1 - - . < 4 , - , ($ (< -
6 !@ 0 .,

" #$%&% ' ((#) * # #/ 4 . 7 $

0 " #%12# # #3 4 - # / / . & 7
% - - - - 7 6 0 .,
: .-4
3< 8 (, EEE
+ . - Q EEE

" #$%&% ' ((#) * # #/ K % - 7 D ' 4 *

0 " #%12# # #3 4 - #
" =, =4K K =- - 7 D =S=T U' 4 * =) - =<. *! .,
: .-4
: & +

0 " #%12# # #3 4 - # , 4 .,
: .-4
: " , /K K % - 7 D

0 " #%12# # #3 4 - # , 4 .,
: .-4

0 " #%12# # #3 4 - # , 4 .,
: .-4

" #$%&% ' ((#) * # #/ -(7 " . - 1 <

0 " #%12# # #3 4 - #
=) *=/ -(7" . - + 1 < .,
: @ .-4

" #$%&% ' ((#) * # # - - < (4 & 4
1 - - - -

@0 " #%12# # #3 4 - # @) %(<- ; 3 "
< (4 4 - - - - $ (< - 6 0 ,C(
: @ .-4

0 " #%12# # #3 4 - # @) %(<- ; 3 "
< (4 4 - - - - $ (< - 6 0 .,

" #$%&% ' ((#) * # # , - - - 7

!0 " #%12# # #3 4 - # - - - , 7
$ < ; 9 < .- 3. - ; - , - - - ,) * M)- 6 0 .,

0 " #%12# # #3 4 - # , - - - 7
G 7 "9 6 ! !0 ,C(
: ! .-4

0 " #%12# # #3 4 - # , - - - 7
G 7 "9 6 ! !0 .,

0 " #%12# # #3 4 - # , - - - . 7
+ < 6 ! !0 ,C(
: @! .-4

0 " #%12# # #3 4 - # , - - - . 7
+ < 6 ! !0 .,

0 " #%12# # #3 4 - # 1 " - ; $ 9 < ,
- - - 7 6 0 G 7 + 6 ! !0 ,C(

0 " #%12# # #3 4 - # 1 " - ; $ 9 < ,
- - - 7 6 0 G 7 + 6 ! !0 .,

0 " #%12# # #3 4 - # 1 " - ; $ 9 < ,
- - - 7 6 0 G 7 + 6 ! !0 ,C(

@0 " #%12# # #3 4 - # 1 " - ; $ 9 < ,
- - - 7 6 0 G 7 + 6 ! !0 .,

0 " #%12# # #3 4 - # 1 " - ; $ 9 < ,
- - - 7 T U6 - 0 G 7 6 ! !0 .,

!0 " #%12# # #3 4 - # 1 " - ; $ 9 < ,
- - - 7 T U6 - 0 G 7 6 ! !0 .,

@ 0 " #%12# # #3 4 - # 1 " - ; $ 9 < ,
- - - 7 ++ G 7 + 6 ! 0 .,

@ 0 " #%12# # #3 4 - # 1 " - ; $ 9 < ,
- - - 7 + G 7 + 6 ! 0 .,

" #$%&% ' ((#) * # #3 & - , /V, - 9 .W ,
" Q

@ 0" #%12# # #3 4 - # ! - 8 X 6 , 0 3
& - , /V, - 9 .W , " Q F ,-XY "- / < * - 6 !@!0 ,C(
: .-4

@ 0 " #%12# # #3 4 - # ! - 8 X 6 , 0 3
& - , /V, - 9 .W , " Q F ,-XY "- / < * - 6 !@!0 .,

" #$%&% ' ((#) * # # -..

@ 0 " #%12# # #3 4 - # < <=4<= HI" $$>&*" .,
: -.. % H- , " 2 - .
: .-4

@ 0 " #%12# # #3 4 - # .-.. - H- , 4 4 .,
: -.. ++
: @ .-4

@ 0 " #%12# # #3 4 - # .-.. - H- , .-..4 4 .,
: -.. ++
: .-4

@@0 " #%12# # #3 4 - # .-.. - H- , .-..4 4 .,
: -.. +
: .-4

" #$%&% ' ((#) * #

" #$%&% ' ((#) * # #%,(- - - $ 4

@ 0" #%12# # #3 4 - # %,(- - -
$ 4 .,
: .-4

54 6 0 " -. $ 4 - 1 4 & "7 ,- 97. ,
/ -)- , - 4 & - $- -

54 6 0

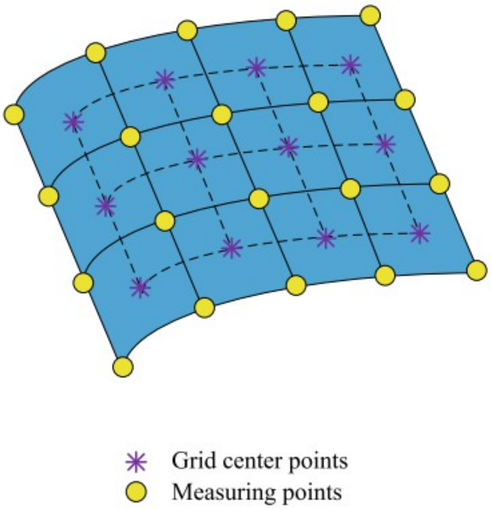
Grid center points
Measuring points

F 4 - , -4 -
(Z - - -
- - ,

54 @6 0

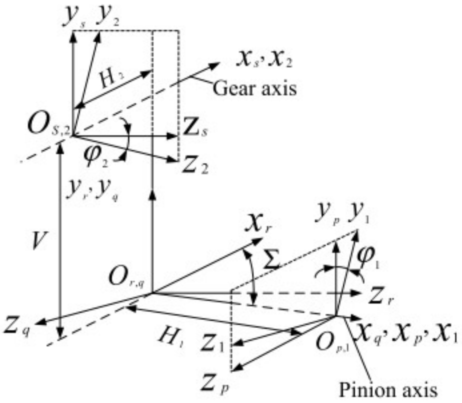
y_s y_2
x_s, x_2
Gear axis
H_2
$O_{s,2}$
Z_s
φ_2
Z_2
y_r, y_q
V
x_r
y_p y_1
Σ
φ_1
$O_{r,q}$
Z_r
Z_q
H_1
Z_1
$O_{p,1}$
x_q, x_p, x_1
Z_p
Pinion axis

F 4 4
, - 7

$$= M_{s1} P_1(u_1, v_1)$$

$$M_{s\_p2} = M_{s2} P_2(u_2, v_2)$$

$$M_{p1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\varphi_1) & \sin(\varphi_1) & 0 \\ 0 & -\sin(\varphi_1) & \cos(\varphi_1) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$M_{qp} = \begin{bmatrix} 1 & 0 & 0 & H_1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$M_{rq} = \begin{bmatrix} \cos(\Sigma) & 0 & -\sin(\Sigma) & 0 \\ 0 & 1 & 0 & 0 \\ \sin(\Sigma) & 0 & \cos(\Sigma) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$M_{sr} = \begin{bmatrix} 1 & 0 & 0 & H_2 \\ 0 & 1 & 0 & -V \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$M_{s2} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\varphi_2) & \sin(\varphi_2) & 0 \\ 0 & -\sin(\varphi_2) & \cos(\varphi_2) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

φ_1 and φ_2 are the rotary angles of pinion and gear. M_{s1} and M_{s2} are the transformation matrices from measuring coordinate system S_1 and S_2 to the meshing coordinate system S_s for the pinion and gear. S_p, S_q, and S_r are auxiliary coordinate systems. H_1 and H_2 are the axial settings of pinion and gear. Σ is the shaft angle. V is the pinion offset distance.

" #$%&% ' ((#) * # #)- A $ - / 7
3< F * % D (C5 . -, EEE

" #$%&% ' ((#) * # #) - " 44 + , 4 + 7= %
/ -. / , +) * 6 0
3< % D (C5 . -, EEE

" #$%&% ' ((#) * # #F- % 4 < - / 7

@!0" #%12# # #3 4 - # F- % 4 < - / 7 .,
: .-4

" #$%&% ' ((#) * # #F * , / 7
3< % D (C5 . -, EEE

" #$%&% ' ((#) * # #F ,- + $ - "-

0 " #%12# # #3 4 - # F ,- = =+
$ - ="- = =+= ,=1 (,=- ,=L.,- ,=A, .,
: .-4

" #$%&% ' ((#) * # #F R / & - -
& - ; - . ; -4 $ E G , 6 @0
3< % D (C5 . -, EEE

" #$%&% ' ((#) * # #+ . -) *
3< 8 (. - -, ,- .- - -
: + . - Q 6 0
: - D (

" #$%&% ' ((#) * # # A 9- ,< *; ,
A,

0 " #%12# # #3 4 - # - -, 9 1- ,
A 9- ,< *) 9 - 6 @0 .,
: .-4

54 6 ! 0 F 4

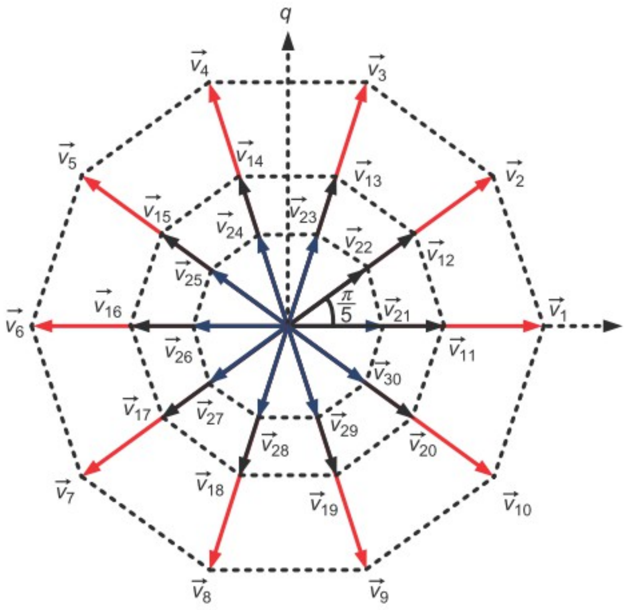
q
$\vec{v}_1$
$\vec{v}_2$
$\vec{v}_3$
$\vec{v}_4$
$\vec{v}_5$
$\vec{v}_6$
$\vec{v}_7$
$\vec{v}_8$
$\vec{v}_9$
$\vec{v}_{10}$
$\vec{v}_{11}$
$\vec{v}_{12}$
$\vec{v}_{13}$
$\vec{v}_{14}$
$\vec{v}_{15}$
$\vec{v}_{16}$
$\vec{v}_{17}$
$\vec{v}_{18}$
$\vec{v}_{19}$
$\vec{v}_{20}$
$\vec{v}_{21}$
$\vec{v}_{22}$
$\vec{v}_{23}$
$\vec{v}_{24}$
$\vec{v}_{25}$
$\vec{v}_{26}$
$\vec{v}_{27}$
$\vec{v}_{28}$
$\vec{v}_{29}$
$\vec{v}_{30}$
$\frac{\pi}{5}$

F+/ - - (-4 .- (- 6 -
- - 4 0 , D . -

54 ! 6 0 F 4

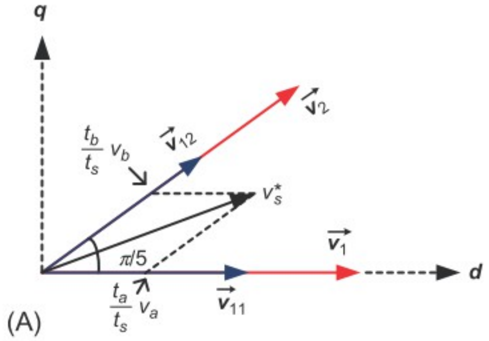
q
d
$\frac{t_b}{t_s} v_b$
$\vec{v}_{12}$
$\vec{v}_2$
v_s^*
$\vec{v}_1$
$\pi/5$
$\frac{t_a}{t_s} v_a$
$\vec{v}_{11}$
(A)

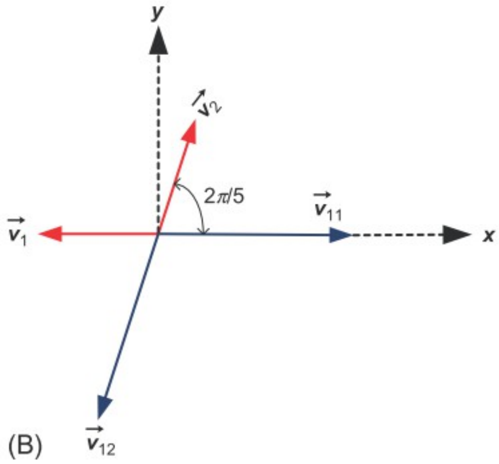
y
$\vec{v}_2$
$2\pi/5$
$\vec{v}_{11}$
$\vec{v}_1$
x
(B)
$\vec{v}_{12}$

F+/ . - - (-.. - ;
6%0 , D . - - , 6)0 H 7 . -

" #$%&% ' ((#) * # #& - - < -

0 " #%12# # #3 4 - # & - - < .,
: .-4
3< 8 (, EEE

0 " #%12# # #3 4 - # & - - < .,
: .-4
3< 8 (, EEE

" #$%&% ' ((

" #$%&% ' ((#) *

" #$%&% ' ((#) * #

" #$%&% ' ((#) * # #$- < H . -

0 " #%12# # #3 4 - # -4 , ,- < .,

: 8 X - & K /K K - , - -.. - 4K K D , -
K - $- < H
: .-4

0 " #%12# # #3 4 - # -4 , ,- < .,
: .-4

0 " #%12# # #3 4 - # -4 , ,- < .,
: .-4

@0 " #%12# # #3 4 - # -4 , ,- < .,
: @ .-4

" #$%&% ' ((#) * # #/ - (4- (, -
@ - ,

0 " #%12# # #3 4 - # @ - 4 4- .,
: F -
: @ .-4

!0 " #%12# # #3 4 - # @ < <=4<=<[%%%%% %%P .,
: $ D / - - . " (- /-
: @ .-4

! 0 " #%12# # #3 4 - # @ < <=4<=G LA%%%%B%%P .,
: F -
: @ .-4

! 0 " #%12# # #3 4 - # @ ! @ .,
: / - + (4- " (, - @ - , /-
: .-4

! 0 " #%12# # #3 4 - # @ 4 - (4- 4- .,
: / - + (4- " (, - @ - , /-
: ! & L (7 8 < - 7
: .-4

! 0 " #%12# # #3 4 - # @ 4 - (4- 4- .,
: / - + (4- " (, - @ - , /-
: ! & L (7 8 < - 7
: .-4

! 0 " #%12# # #3 4 - # @ 4 4- .,
: 1A"9A1"9A /\'\1%8A L1 8A L1F%"A "3L1)A
: F @ .-4

! 0 " #%12# # #3 4 - # @ - - ,- < .,
: L1 L'A "8% A 1A %12L%)8A $A "3L1)A A& $A L1F%"A %8/\)1+2LA
A& L1 8% &9A31+A $A + %/+'%+1A /% &3' $%1)3L[
] .-4

" #$%&% ' ((#) * # #/ - - S - 4K K K -

! 0 " #%12# # #3 4 - # -.. - , - .,
: % 8+"%&+3' $A 8% \&93$A A"&31+A88A $A /1% %" % 8% /\3 \&1+A
+'F+'+&\ + %8A
: .-4

!@0 " #%12# # #3 4 - # 4 .,
: /K K + K -
: .-4

! 0 " #%12# # #3 4 - # . . ,- - .,
: . ,- - H , - K , . , -
: .-4
: + - 7

" #$%&% ' ((#) * # #. C * (6 0 /

!!0 " #%12# # #3 4 - # ! < , -4 (< 4 .,
: / , 8-4) 4
: .-4

0 " #%12# # #3 4 - # ! < , . C - .,
: 8 < ,
C * B 6 0 /
6 7 / ; / , 8-4 0
: .-4
:

0 " #%12# # #3 4 - # ! < , . C - .,
: 8 < ,
C * B 6 0 /
6 7 / ; / , 8-4 0
: .-4
:

0 " #%12# # #3 4 - # ! < , . C - .,
: 8 < ,
C * B 6 0 /
6 7 / ; / , 8-4 0
: .-4
:

" #$%&% ' ((#) * # # - , %.. , % 4 < - & . 47

0 " #%12# # #3 4 - #
=- ,=%.. ,=% 4 < - =& . 47 .,
: @ .-4
3< 8 (, EEE 1 Y 4- EEE

" #$%&% ' ((#) * # #2 - A ;) ; - , / -
$ <- -< '- 1 - 7

0 " #%12# # #3 4 - # - C > - 8 4 - , - ,
- . 7 D D - D A ;) ,K<- - - , - K- K P - " - , 8- N
6 0 .,
: @ .-4

0 " #%12# # #3 4 - # - C > - 2 -
A ;) - , / - $ <- %< '- 1 - 7 +) * 6 0 .,
: .-4

0 " #%12# # #3 4 - # - C > - 2 5
A ;) 7 4 - , <- < - - - I- , - - ,-, A, - >- 6 0 .,
: @ .-4
: .-

" #$%&% ' ((#) * # #1 - % - 7 - , + 7

@0 " #%12# # #3 4 - # 1 - % - 7 - , + 7 .,
: @@ .-4

" #$%&% ' ((#) * # #1 - + . - 2 -
-

0 " #%12# # #3 4 - #
1 - =+ . - = =2 - = - .,
: .-4

" #$%&% ' ((#) * # # . 7 - 6$> . 70

!0 " #%12# # #3 4 - # $> & -) *) 4 +, -
. 7 AH. - , $> 6 !0 .,
: @ .-4

0 " #%12# # #3 4 - # $* & -) *) 4 +, -
. 7 AH. - , $> L) 6 !0 .,
: ! .-4

3< " - - , X- , .54 - !

0 " #%12# # #3 4 - # . 7 - $ 4
> , 7 .,
: .-4
3< 1 Y EEE

" #$%&% ' ((#) * # #& 3H , 9- ,< * 9 7
2 - + . -

0 " #%12# # #3 4 - # & 3H , 9- ,< *
9 7 2 - + . - .,
: .-4
3< 8 (, EEE
: 5 -

" #$%&% ' ((#) * #

" #$%&% ' ((#) * # #% & - 3 - % , %,(- ,
& 4 7

0 " #%12# # #3 4 - #
-= - = =. - =- ,= . - = 4 7=- ,= =-.. - = =- 7=- ,=4 ,
7= ! .,
:) .-4
: 1 Y 8 4- EEE

0 " #%12# # #3 4 - # - . - < .,
: % & - - & 4 7 9 <
: .-4

" #$%&% ' ((#) * # #% - 4 7 - ;
<7 %))-

0 " #%12# # #3 4 - # @ % ,)- - ,)- % -
4 7 - " L (7 8 < - 7 6 ! 0 .,
: .-4

0 " #%12# # #3 4 - # @ % ,)- - ,)- % & -
/ 7 - '-< 6 0 .,
: .-4

" #$%&% ' ((#) * # #%,(+ 2 - - ,

@0 " #%12# # #3 4 - # @ ! .,
: %,(2 - - ,
: .-4

" #$%&% ' ((#) * # #A % " 2 -
1 - (& 7 - , AH

0 " #%12# # #3 4 - # ! =!@ @ @ .,
: A % "
: ! .-4

" #$%&% ' ((#) * # #A - 7 ' < & 7 -
+ , - H & 7

!0 " #%12# # #3 4 - # % <)- , P 7 F A
%< - -) - .,
: .-4
3< 8 (, EEE 1 Y EEE
: - H & 7

" #$%&% ' ((#) * # #/ 7 4 - 6 A% 0

0 " #%12# # #3 4 - # / 7= = 4 - .,
: .-4

3< 8 (, EEE
1 Y - EEE

" #$%&% ' ((#) * # #+ . -

" #$%&% ' ((#) * # #+ , .- - , -
D - - , 9 < .- ,

0 " #%12# # #3 4 - # >- A / - + ,
.- - , - D - - , 9 < .- , $ (< - 6 !!@0 ,C(
: @ .-4

0 " #%12# # #3 4 - # >- A / - + ,
.- - , - D - - , 9 < .- , $ (< - 6 !!@0 .,

" #$%&% ' ((#) * # # - - . -

0 " #%12# # #3 4 - #) 4 4 - , + , -
% 4 < - % + 4 - , %.. -) * " 6 0 .,
: .-4

0 " #%12# # #3 4 - # $ (. - - -
" 4 , ; , A, 6%(- -< & " 4-4 0) * " 6 0 .,
: .-4

0 " #%12# # #3 4 - # / - + , - % 4 < -;
& , A,) * " 6 0 .,
: ! ! .-4

0 " #%12# # #3 4 - # / - - 5 -)5 -
.- - (- " 4-4 8 - 4 A, % , " 6 @0 .,
: .-4

@0 " #%12# # #3 4 - # - - - " .
2 - - . 4 6 0 .,
: .-4

0 " #%12# # #3 4 - # - - - " .
2 - - . 4 -4) 9 , < 4 6 0 .,
: ' A,
: .-4

" #$%&% ' ((#) * # #' - 1- 4 9 < .-
3. -

!0 " #%12# # #3 4 - # ' - 1- 4 9 < .-
3. - .,
: .-4
: - H
: " (H

" #$%&% ' ((#) * # #1 - (7 -, 1 - (7 A- 7

0 " #%12# # #3 4 - # % , - 1 - (7
-, 1 - (7 A- 7 / - 1 - (7 - , " 47 3H , L (7
6 0 .,
: ! .-4
3< 8 (, EEE

0 " #%12# # #3 4 - # % , - 1 - (7
-, 1 - (7 A- 7 3H , L (7 ; L % 6 0 .,
: .-4

0 " #%12# # #3 4 - # - ; % 1 - (7 -,
1 - (7 A- 7 3L 3H , 6 0 .,
: .7

" #$%&% ' ((#) * #

" #$%&% ' ((#) * # #%) * 3 " (8 * ,

0 " #%12# # #3 4 - # A 9 8 * , % < * (
L (7 6 ! 0 .,
: .-4

0 " #%12# # #3 4 - # A 9 8 * , %) *
" ("- < ,4 L (7 6 @0 ,C(

0 " #%12# # #3 4 - # A 9 8 * , %) *
" ("- < ,4 L (7 6 @0 .,

" #$%&% ' ((#) * # #% 9 7 3 / - ,
" ,4

0 " #%12# # #3 4 - # @ P - 8 " ,4 % 7
4 - , 3H , L (7 6 ! 0 ,C(
: .-4

@0 " #%12# # #3 4 - # @ P - 8 " ,4 % 7
4 - , 3H , L (7 6 ! 0 .,

0 " #%12# # #3 4 - # @ P - 8 " ,4 % 9 7
" - , 2 -, - $ (< - 6 ! 0 ,C(
: .-4

!0 " #%12# # #3 4 - # @ P - 8 " ,4 % 9 7
" - , 2 -, - $ (< - 6 ! 0 .,

0 " #%12# # #3 4 - # @ P - 8 " ,4 % -
- 4 < - . - ($ (< - 6 0 ,C(
: .-4

0 " #%12# # #3 4 - # @ P - 8 " ,4 % -
- 4 < - . - ($ (< - 6 0 .,

0 " #%12# # #3 4 - # @ P - 8 " ,4 &
A ' A , - / 7 " L (7 8 < - 7 6 ! !0 .,
: .-4

0 " #%12# # #3 4 - # @ P - 8 " ,4 &
/ 7 " . H $ - 6 ! 0 .,
: .-4

0 " #%12# # #3 4 - # @ 8 " ,4 P - % -
- , . F 4) * 6 ! 0 .,
: .-4

0 " #%12# # #3 4 - # @ & - - 4 -
- - " - , = 3H , L (7 6 !! 0 ,C(
: .-4

0 " #%12# # #3 4 - # @ & - - 4 -
- - " - , = 3H , L (7 6 !! 0 .,

" #$%&% ' ((#) * # #%,(- , " - - A -4

@0 " #%12# # #3 4 - # %,(- , " - -
A -4 .,
: ! .-4

" #$%&% ' ((#) * # #% F .7 A 47
1 -

" #$%&% ' ((#) * # #AH , /K K

0 " #%12# # #3 4 - # ! < <=4<=) %%%% %%P .,
: AH , 4K K F P
: ' A,
: .-4

!0 " #%12# # #3 4 - # ! ., * = 4 H ,
4 . - . , , 4 D D .,
: AH , 4K K F /
: A,
: .-4

" #$%&% ' ((#) * # #AH. , '- ,

0 " #%12# # #3 4 - # @ AH. , '- , .,
: AH. , '- , % < $ - .
% AH 4 7-4 3. - .7 '- .7

" #$%&% ' ((#) * # #F- % 4 < - / 7

0 " #%12# # #3 4 - # @ F- % 4 < - / 7
.,

: .-4

0 " #%12# # #3 4 - # @ F- % 4 < - / 7
.,
: ! .-4

" #$%&% ' ((#) * # #/ - , (,
- ; - - % ,

0 " #%12# # #3 4 - # @ ! @ @ .,
: % ,
: .-4

0 " #%12# # #3 4 - # @ 4 - - .,

" #$%&% ' ((#) * # #+ - - 7 - , & - ,
' < & 7

0 " #%12# # #3 4 - # @ $-(, % 4 + - - 7 - ,
& - , ' < & 7 "1" 6 0 .,
: .-4

: + - < . - -
* - ^ - - ; < - * ;
7 , < -< *
A " & -
$ 4- I _- - , < / 4 - ;
- - , *
8 . , > *

< A (- .- - - - I - ,- Y . ,

3< 8 (, EEE

" #$%&% ' ((#) * # # . F ,- 1 - (
F , $7 -

0 " #%12# # #3 4 - # @ . F ,-
1 - (F , $7 - .,
: .-4

" #$%&% ' ((#) * # # % L . -
: - - I .-, , / 4 8 (-.. < EEE
: ! /<7

" #$%&% ' ((#) * # #2 - - $ - , 9 - ;
- - - - , " . - - %.. -

@0 " #%12# # #3 4 - # @ 2 - - $ - , 9
"- 8 - F .,
: .-4
3< 8 (, EEE

" #$%&% ' ((#) * # # % - %.

" #$%&% ' ((#) * #

!

" #$%&% ' ((#) * #

" #$%&% ' ((#) * # #% 9 7 7 -4 -

0 " #%12# # #3 4 - # @ % 9 7 7 -4 - .,

54 6 0 - - -

" #$%&% ' ((#) * # #F - ,) ,- 7 A ,
A 4 4

!0 " #%12# # #3 4 - # @@ F - ,) ,- 7 A
, A 4 4 .,

" #$%&% ' ((#) * # #F , - + L

0 " #%12# # #3 4 - # @
F8L+$= A"9%'+" =FL'$% A'&%8 =%'$=% 8+"%&+3' = + .,

" #$%&% ' ((#) * # #/ -, , % 4 < - % 4 < - / 7

0 " #%12# # #3 4 - # @!
/ -, ,=% 4 < - = =% 4 < - =/ 7 .,
54 6 0 " -. ! & /- -.

" #$%&% ' ((#) * # # "-

0 " #%12# # #3 4 - # - .,
: "- % - 7 / 7

0 " #%12# # #3 4 - # , - -
- .- 7 , "- ; 8 - % 4 < -; - , $ - F % L ,
%.. - .,

0 " #%12# # #3 4 - # & . , , - D + 4 -
- .

0 " #%12# # #3 4 - # "- ; 8 - % 4 < -;
- , $ - F % L , %.. - - H A, 6 0 .,

0 " #%12# # #3 4 - # "- ; 8 - % 4 < -;
- , $ - F % L , %.. - 6 !; - H A, 0 .,

" #$%&% ' ((#) * # @

" #$%&% ' ((#) * # @#) ,- 7 " F H < &
$ - A)) -

@0 " #%12# # #3 4 - #) ,- 7 " F H <
& $ - A)) - .,

" #$%&% ' ((#) * # @#"-

0 " #%12# # #3 4 - # / - (, 5 .,

" #$%&% ' ((#) * # @#"- % " . + ,
6& - B 0; A,

!0 " #%12# # #3 4 - # "- % " . + ,
6& - B 0; A, .,

" #$%&% ' ((#) * # @#"- " - *

@ 0 " #%12# # #3 4 - # "- " - * .,

" #$%&% ' ((#) * # @#" % - - - A,

@ 0 " #%12# # #3 4 - # " % - - - A,
.,
3< 8 (, EEE

" #$%&% ' ((#) * # @#A , - / 7 - , " (H 7

@ 0 " #%12# # #3 4 - # 1) A , -
/ 7 ? " (H 7 / - 6 ! 0 ,C(

@ 0 " #%12# # #3 4 - # 1) A , -
4 7 - , (H 7 / - 9 6 ! 0 .,
3< 8 (, EEE

" #$%&% ' ((#) * # @#8 < I
3< % D (I .-,

" #$%&% ' ((#) * # @#8 - % 4 < - $ 1 4 ; , A,

@ 0 " #%12# # #3 4 - # @ 8 - % 4 < - $ 1 4
. 4 + - - < 4 + . = . 4 6 0 .,
3< 8 (, EEE

" #$%&% ' ((#) * # @#&""

````````

@ 0 " #%12# # #3 4 - # &""3 H
3< & -<- , " Y , " , $
````````

" #$%&% ' ((#) * # @#& 7

@ 0 " #%12# # #3 4 - # ! F $-(, - F " - 7
L - 7 & 7 - , +, - & " 7 P . 9 7
6 0 .,

@@0 " #%12# # #3 4 - # ! 1 - , % 8 * & P 7
% AH- - 9 % * - , % 2 L 4 7
A(- - - -, 4 . 4 6 @0 .,

@ 0 " #%12# # #3 4 - # ! - P & 2 - &
% 7 - ; & -4 ,7; % < ; - , 3H , L (7 ; L % 6 0 .,

" #$%&% ' ((#) * #

" #$%&% ' ((#) * # #8 < I . -,

@!0 " #%12# # #3 4 - # ! ' ; 8 < I ! %<
% " - + , - 6 !@ 0 .,

0 " #%12# # #3 4 - # ! & 8 < I - " (7
B- L (7 6 0 .,

" #$%&% ' ((#) * # # - , - - I 4
- - 4

0 " #%12# # #3 4 - # !)- $) 7 % + ,
+ , / 7) * M 6 0 .,

0 " #%12# # #3 4 - # ! F) * 9- ,< *
, 4 7 < , 4 - , ,- A (6 !! 0 ,C(
: !! .-4

0 " #%12# # #3 4 - # ! F) * 9- ,< *
+ , / 7) , 4 - , F , ' 9 - , 6 !! 0 .,

0 " #%12# # #3 4 - # ! P - L < < 4 6- 0
F ,- + , / 7 C (- , - .- . 4 -4)
9 , < 4 6 0 .,

0 " #%12# # #3 4 - # ! - , - - I 4
- - 4 . 4 -4) 9 , < 4 6 0 .,
: !! .-4

0 " #%12# # #3 4 - # ! G & .,

" #$%&% ' ((#) * # # 4 8 - - - 7

@0 " #%12# # #3 4 - # ! 4 8 - - - 7 .,
: .-4

" #$%&% ' ((#) * # !

" #$%&% ' ((#) * # !#9 < 4R L - 7 . - ,
A - 2 - I ,

0 " #%12# # #3 4 - # ! 9 < 4R L - 7 .
- , A - 2 - I , .,

" #$%&% ' ((#) * # !#3. - G-(4 , & 7

!0 " #%12# # #3 4 - # ! 3. - G-(4 , & 7 .,

" #$%&% ' ((#) * # !#& ' B * &) *
- -

! 0 " #%12# # #3 4 - # ! F - * P & . & 7
+ - 7 , " 47; / , - , 1 $ -, % 6 !!@0 ,C(

! 0 " #%12# # #3 4 - # ! F - * P & . & 7
+ - 7 , " 47; / , - , 1 $ -, % 6 !!@0 .,

! 0 " #%12# # #3 4 - # ! & ' B * &) *
- - & - B - G .,

" #$%&% ' ((#) * #

" #$%&% ' ((#) * # #% " 4 " 1 - (7 - ,
" 47

! 0 " #%12# # #3 4 - # ! % " 4 " 1 - (7 - ,
" 47 .,

" #$%&% ' ((#) * # #$ - / 7 " (- ,
-

! 0 " #%12# # #3 4 - # !@ $ - / 7 " (
- , - . 4 + - - < 4 6 0 .,

! 0 " #%12# # #3 4 - # !@ 8 - A - %
& - $ - / 7 " (- , - 6$ (H A, 0 $ (
< - 6 0 .,

! 0 " #%12# # #3 4 - # !@ - , $ "- $ -
4 7 (- , - 9- 6 !@ 0 ,C(

!@0 " #%12# # #3 4 - # !@ - , $ "- $ -
4 7 (- , - 9- 6 !@ 0 .,

! 0 " #%12# # #3 4 - # !@ - , $ "- $ -
4 7 (- , - 9- 6 !@ 0 ,C(

!!0 " #%12# # #3 4 - # !@ - , , "-
$ - / 7 " (- , - $ (< - 6 0 .,

0 " #%12# # #3 4 - # !@ &-.. 6- 0 $ -
/ 7 " (- , - . 4 + - - < 4 6 0 .,

" #$%&% ' ((#) * # #A ; 1 - (7 - , %<
- 7

0 " #%12# # #3 4 - # ! A ; 1 - (7 - , %<
- 7 1 ,4 6 0 .,

" #$%&% ' ((#) * # #\ K , /K K 8 4 ,

0 " #%12# # #3 4 - # !! \ K , /K K
8 4 , @I
3< % D (I .-, ; - , W , EEE

" #$%&% ' ((#) * # #F 8- & - $ (L (

0 " #%12# # #3 4 - #
F =8- =& - =$ (= =L (.,

" #$%&% ' ((#) * # #/ 7 - , 7 7 B-

0 " #%12# # #3 4 - # -) B- / 7 - ,
7 7 6$ ( ) * %,(- , - - 0 9 , ,-7 6 ! 0 .,

0 " #%12# # #3 4 - # .- 7- / 7 - ,
7 7 9 , $-7 6 ! 0 ,C(

" #$%&% ' ((#) * # #/ 7 - F

0 " #%12# # #3 4 - # =F =
= - =1$= -. =/ 7=F =- ,=" . H 7=F+'%8 .,

@0 " #%12# # #3 4 - # & H , % 4
.,
: $A 3' &1%&+3' +' /A3 A&1B 9+ &31+"%8 %'$ 9+83 3 9+"%8
A1 A"&+ A

$A 3' &1%ab3 A /A3 A&1+% A1 A"&+ % 9+ &c1+"% A
F+83 cF+"%

0 " #%12# # #3 4 - # ! .,
: &9A /A3 A&1B 3F &91AA F31 +' +[$+ A' +3'

!0 " #%12# # #3 4 - # "$ (, .,
: , - - , - 4 7

0 " #%12# # #3 4 - # " -(.,
: 1 - - / 7
: ! .-4
3< 8 (, EEE

0 " #%12# # #3 4 - # / 7 .- - F =
% - 7 ; 7 ; " . F - - , .- "%$ A 9 , 8 , ;
< ,C(
: .-4
3< 8 (, EEE

0 " #%12# # #3 4 - # / 7 .- - F =
% - 7 ; 7 ; " . F - - , .- "%$ A 9 , 8 , ;
< .,

0 " #%12# # #3 4 - # / 7 .- - = .-
("%$ 9 , 6 ! 0 ,C(

0 " #%12# # #3 4 - # / 7 & - , A , -
- / 7)- , 1 4 , % - - - - 7 6 0 .,
: .-4

0 " #%12# # #3 4 - # (= .,
: / 7 %

0 " #%12# # #3 4 - # - - ,& .,

@0 " #%12# # #3 4 - # 9+ 3& (.,
: - d F ,- / 7 L , - , 4

0 " #%12# # #3 4 - # - = =4 7 .,

" #$%&% ' ((#) * # #/ -(- - G-(

!0 " #%12# # #3 4 - #) - " 44 / -(- - G-(
9 A R .- .. .,

0 " #%12# # #3 4 - #) - " 44 / -(- - G-(
9 A R .- .. (- (+) * 8 , 6 0 . <

" #$%&% ' ((#) * # #+ . -

0 " #%12# # #3 4 - # , - = .,

0 " #%12# # #3 4 - # , - = .,

0 " #%12# # #3 4 - # , - = .,

0 " #%12# # #3 4 - # , - = .,

0 " #%12# # #3 4 - # , - = .,

0 " #%12# # #3 4 - # , - = .,

@0 " #%12# # #3 4 - # , - = @ .,

0 " #%12# # #3 4 - # , - = .,

!0 " #%12# # #3 4 - # , - = ! .,

0 " #%12# # #3 4 - # , - = .,

0 " #%12# # #3 4 - # , - = .,

0 " #%12# # #3 4 - # / - - - " Xe
/ K - .,

0 " #%12# # #3 4 - # / - - - .,

0 " #%12# # #3 4 - # 4 -= , - -=. - - .,

0 " #%12# # #3 4 - # ' "
- .,
: ! .-4

0 " #%12# # #3 4 - # - 7 , 4 ?
" . - .,

" #$%&% ' ((#) * # # %&9

@0 " #%12# # #3 4 - # $ -XY F - .,

0 " #%12# # #3 4 - # , - - - - .,
: % - ,- W - - XY ,

!0 " #%12# # #3 4 - # $ - +,-/ - (.,
: ! .-4
: 3 < - ,- - 5 -

0 " #%12# # #3 4 - # -= - - - .,

0 " #%12# # #3 4 - # 9 L (.,

0 " #%12# # #3 4 - # &"" %L1A% FA1 %& .,

" #$%&% ' ((#) * # # - - - J $ 4 6 $F0

0 " #%12# # #3 4 - # - - - = =$ 4 .,

" #$%&% ' ((#) * # # - - - - A 4

0 " #%12# # #3 4 - # @
- - = = - - =A 4 = < =- ,= .,

" #$%&% ' ((#) * # # - - F - , F

0 " #%12# # #3 4 - # - , - 8-
- - = F - , F . 4 6 ! 0 .,
54 6@ 0 / 7
3< 8 (, EEE

0 " #%12# # #3 4 - # - , - 8- 6- 0
- - F - , F . 4 -4 ' B * 6 ! 0 .,
: .-4

" #$%&% ' ((#) * # # C ($ - / 7 3
" (- , -

@0 " #%12# # #3 4 - # ! A P G I7 * C (
, - 4 7 (- , , - " L (7 8 < - 7 6 ! 0 .,
: .-4

" #$%&% ' ((#) * # # C ($ - / 7 3
" (- , - 8-

0 " #%12# # #3 4 - # C (
$ - .,

!0 " #%12# # #3 4 - # C (
$ - = H .,

0 " #%12# # #3 4 - # . C (, ! <. ,C(

0 " #%12# # #3 4 - # . C (, ! <. .,

" #$%&% ' ((#) * # # -(/ 7 - , 8 4 % F
+ , & . & 7

0 " #%12# # #3 4 - # -(/ 7 - , 8 4 %
F + , & . & 7 .,
: ! .-4

" #$%&% ' ((#) * # #& % / 7

0 " #%12# # #3 4 - # & % / 7
% , + 1 , + - - < 6 ! @0 .,

0 " #%12# # #3 4 - # & % / 7
% , ++ 1 , + - - < 6 ! @0 .,

" #$%&% ' ((#) * #

" #$%&% ' ((#) * # # - - & 8 " - 7

0 " #%12# # #3 4 - # > A - - , -
- 5 - , - <8 < 6 !@ 0 . <

0 " #%12# # #3 4 - # > A - - , -
- 5 - , - <8 < 6 !@ 0 .,

@0 " #%12# # #3 4 - # > 8- - - - -
- , - F 6 !@ 0 .,

0 " #%12# # #3 4 - # > - - -
& 4 % , & ; 3H , L (7 ; L % 6 !! 0 .,

!0 " #%12# # #3 4 - # > - - %
" - %.. - 6%,, G 7 - - 0 6 ! 0 .,

0 " #%12# # #3 4 - # > - -
- - - " $ (< - 6 ! 0 .,

0 " #%12# # #3 4 - # > - -
G " 3H , L (7 ; L % 6 ! 0 .,

0 " #%12# # #3 4 - # > - - ;
- 7 3H , L (7 6 ! 0 .,

0 " #%12# # #3 4 - # > - - =
- 7 3H , L (7 6 ! 0 .,

0 " #%12# # #3 4 - # > - , .
- - $- - 44 ++ A - , 6 !!!0 .,

0 " #%12# # #3 4 - # > - , .
- - $- J% S - + A - , 6 !!!0 .,

" #$%&% ' ((#) * # # - ; - , 3 + 4
' <

0 " #%12# # #3 4 - # 9 < >-7 & G ,
' < "- < ,4 L (7 6 ! 0 .,

@0 " #%12# # #3 4 - # 1 % $ -. & 4 , -
- , F < - < G , < 4 " .- 7 6 !! 0 .,

0 " #%12# # #3 4 - # & 4 - 4 , < *
- - H. 4 , < - , .- / , 6 ! 0 .,

!0 " #%12# # #3 4 - # & 7 ' < G ,
6 @0 .,

@ 0 " #%12# # #3 4 - # B A 3 %, - - ;
- , 3 + 4 ' < G , < 4 " .- 7 6 0 .,

@

@

" #$%&% ' ((@

" #$%&% ' ((@ #) * @

" #$%&% ' ((@ #) * @ #

" #$%&% ' ((@ #) * @ # #/K K " . - F *

@ 0 " #%12# # #3 4 - # F * % , -
4 7 - , . 47 6 ! 0 ,C(

@ 0 " #%12# # #3 4 - # F * % , -
4 7 - , . 47 6 ! 0 .,

@ 0 " #%12# # #3 4 - # F * % " $ -
/ 7 - , & . 47 +1 6 ! 0 .,

@ 0 " #%12# # #3 4 - # F * / - , -
, , - & - , 9 4 - A, - 6 ! @0 .,

@ 0 " #%12# # #3 4 - # F * + 4 -< 9- -
7 4 7; . 47; - - " -. - ? 9- "1" 6 0 .,
: @ .-4

@ 0 " #%12# # #3 4 - # F * + ,
$ - - , % 4 < - & . 47 . 4 .,

@@0 " #%12# # #3 4 - # F * - - ;
- , (- , ; - , - - . < % - - - - 7 6 !! 0 .,

@ 0 " #%12# # #3 4 - # /K K . -
.- 4K K , - ; , 4 . , - - , - . K,
6 ! 0 ,C(

@!0 " #%12# # #3 4 - # /K K . -
.- 4K K , - ; , 4 . , - - , - . K,
6 ! 0 .,

0 " #%12# # #3 4 - # /K K . - .-
4K K . 4 , (- K K K, ,C(

0 " #%12# # #3 4 - # /K K . - .-
4K K . 4 , (- K K K, .,

0 " #%12# # #3 4 - # /K K . - .-
K , , - K , R 4 K, ,C(

0 " #%12# # #3 4 - # /K K . - .-
K , , - K , R 4 K, .,

0 " #%12# # #3 4 - # , / 7 , - ,
%.. - - + , 9 47 & 7 . 4 6 !! 0 ,C(

0 " #%12# # #3 4 - # , / 7 , - ,
%.. - - + , 9 47 & 7 . 4 6 !! 0 .,

0 " #%12# # #3 4 - # , / 7 , - ,
%.. - - + & / 7 - ; & - - / . ; - , F ,C(

@0 " #%12# # #3 4 - # , / 7 , - ,
%.. - - + & / 7 - ; & - - / . ; - , F .,

0 " #%12# # #3 4 - # , / 7 , - ,
%.. - - ++ & / 7 - , & . 47 - , . 4 -4 ' ,C(

!0 " #%12# # #3 4 - # , / 7 , - ,
%.. - - ++ & / 7 - , & . 47 - , . 4 -4 ' .,

" #$%&% ' ((@ #) * @ # #9 ,4 & 7 - , " . H % 4 < -
/ 7

! 0 " #%12# # #3 4 - # 9 ,4 7 - , . H
- 4 < - 4 7 + "- < ,4 L (7 6 0 ,C(

! 0 " #%12# # #3 4 - # 9 ,4 7 - , . H
- 4 < - 4 7 ++ "- < ,4 L (7 6 0 ,C(

! 0 " #%12# # #3 4 - # 9 ,4 & 7 - , " . H
% 4 < - / 7 "- < ,4 L (7 6 0 .,

! 0 " #%12# # #3 4 - # 9 ,4 & 7 - , " . H
% 4 < - / 7 "- < ,4 L (7 6 0 .,

" #$%&% ' ((@ #) * @ # #& 4 (;
4 7; - , * 1 4

! 0 " #%12# # #3 4 - # @ & 4 (;
4 7; - , * 1 4 3H , L (7 6 !! 0 .,
3< 8 (, EEE

" #$%&% ' ((@ #) * @ # #& * 4 - - - 7

! 0 " #%12# # #3 4 - # P - ; 8) ; > - 7
& * 4 - - - 7 %G 6 0 .,

! 0 " #%12# # #3 4 - # ' * B ; - 8 P 6A,0
/ 7 - , + 4 -< 7 6 0 ,C(

!@0 " #%12# # #3 4 - # ' * B ; - 8 P 6A,0
/ 7 - , + 4 -< 7 6 0 .,

! 0 " #%12# # #3 4 - # & * 4 - - - 7 %,,
G 7 < " 6 ! 0 ,C(

!!0 " #%12# # #3 4 - # & * 4 - - - 7 %,,
G 7 < " 6 ! 0 .,

" #$%&% ' ((@ #) * @ #

" #$%&% ' ((@ #) * @ # #/ %.. H - & 7

0 " #%12# # #3 4 - # ! % (
& - > (=) *=/ %.. H - & 7 .,
3< 8 (, EEE

" #$%&% ' ((@ #) * @ # #/. 1-7 & - 4 ' A , -
.-

0 " #%12# # #3 4 - # /. =1-7=& - 4= ='
A , - = .- .,
3< 8 (, EEE
54 6 0 1 - - 1-7 & - 4

" #$%&% ' ((@ #) * @ # #/ . & 7 - ,- , ,
- 7 - ,) 7 ,

0 " #%12# # #3 4 - # / . & 7 - ,- ,
, - 7 - ,) 7 , &-7 ? F - 6 0 .,

0 " #%12# # #3 4 - #
/ .=& 7= = = - ,- ,= , = = - = 7 =- ,=) 7 , .,

" #$%&% ' ((@ #) * @ # # . (- ; ' - 7 - ,
' 7

0 " #%12# # #3 4 - #)- ,7 .-, 7-7
>- =) *= . (- ' - 7 % 6 0 .,

" #$%&% ' ((@ #) * @ #

" #$%&% ' ((@ #) * @ # #9V , " A F &
$

0 " #%12# # #3 4 - # 9V , " A F
& $.,
3< 8 (, EEE

" #$%&% ' ((@ #) * @ # # - - - , 7
<

0 " #%12# # #3 4 - # - - - , 7
< .,

" #$%&% ' ((@ #) * @ # #3. - , % + ,

@0 " #%12# # #3 4 - # 3. - , .,

" #$%&% ' ((@ #) * @ #

" #$%&% ' ((@ #) * @ # #A4 "

0 " #%12# # #3 4 - # .,
: "

!0 " #%12# # #3 4 - # - .,

0 " #%12# # #3 4 - # @ @@ .,
: 7 -4 - & ' - (/ 7

0 " #%12# # #3 4 - # @ .7 -4 - .,
: B&9%/31% &c8 9L)A1&
: ! .-4

0 " #%12# # #3 4 - # !@ @ @ .,
: / < F 4)
$ 9 < ; A) ;
) 1 ; - 4 M
A I < F 4

0 " #%12# # #3 4 - # %"&% (- .,
: % 1 f ,

0 " #%12# # #3 4 - #) - = =% 4-< .,

0 " #%12# # #3 4 - #
$ =. =S= -=D - N =T U" =A4 =<. * * .,
: @ .-4

0 " #%12# # #3 4 - # A4 " $ - -< -
g - 4 - h A, - i g h 6 ! @0 .,
: .-4

@0 " #%12# # #3 4 - # A4 " 7 -4 - W
9 < 4 % - - *- V K K * * I-*- K F - * &5 - 6 ! @0 .,
: ! .-4

0 " #%12# # #3 4 - # A4 " A -
I + 4 - _ -7 6 ! 0 .,
: .-4
: 3 A+' %8A+' _L +'&A/1%8

!0 " #%12# # #3 4 - # = 7 .,

0 " #%12# # #3 4 - # < I .,
: @ .-4

0 " #%12# # #3 4 - # . , .,

0 " #%12# # #3 4 - # .7 -4 - .,
: .-4

0 " #%12# # #3 4 - # , =/ %H .,

0 " #%12# # #3 4 - # 7 -4 - < 9 < .,

0 " #%12# # #3 4 - # I = .,

0 " #%12# # #3 4 - # I = .,

" #$%&% ' ((@ #) * @ # #/ - F -
1 - I- / . , - - % 7 . % - 7

@0 " #%12# # #3 4 - # @ / - F -
1 - I- / . , .,
: ! .-4

" #$%&% ' ((@ #) * @ #

" #$%&% ' ((@ #) * @ # # . - & 7 1 - (7 6 -
+ , -0

0 " #%12# # #3 4 - # . - & 7 1 - (7
6 - + , -0 .,
: .-4

" #$%&% ' ((@ #) * @ #

" #$%&% ' ((@ #) * @ # # - - '- -

! 0 " #%12# # #3 4 - # !
>- (=) *= - - F '- - 6 0 .,
: @ .-4
3< 8 (, EEE

" #$%&% ' ((@ #) * @ # #& . 4 - $- - % - 7
%.. -

0 " #%12# # #3 4 - # & . 4 - $- - % - 7
%.. - .,

" #$%&% ' ((@ #) * @ # @

" #$%&% ' ((@ #) * @ # @#% C F ,

0 " #%12# # #3 4 - # . G - % C
F , F ,- - & 7 7 G - , - 6 0 .,
: @ ! .-4

" #$%&% ' ((@ #) * @ # @#% + , " (- - 2 -
-

0 " #%12# # #3 4 - # P- 7j*-
, 4 =) *=% + , & " (- - 2 - 6 0 .,
: .-4

" #$%&% ' ((@ #) * @ # @# .-) , & " - - "-

0 " #%12# # #3 4 - # .-) , & " - -
"- .,

" #$%&% ' ((@ #) * @ # @#& . 7 - - -)7
> - -

0 " #%12# # #3 4 - # & . 7 - - -)7
> - - .,

" #$%&% ' ((@ #) * @ #

" #$%&% ' ((@ #) * @ # # - - F " . 4 - ,
" -

0 " #%12# # #3 4 - # - - F " . 4
- , " - .,

" #$%&% ' ((@ #) * @ # #& 8 4 ' <

0 " #%12# # #3 4 - # & 8 4 ' < .,

" #$%&% ' ((@ #) * @ # !

" #$%&% ' ((@ #) * @ # !#% + 7 G ,

@0 " #%12# # #3 4 - # @ % + 7 G , G
> 7 . <

0 " #%12# # #3 4 - # @ % + 7 G , G
> 7 .,

" #$%&% ' ((@ #) * @ # !#+ -- ' - , ,7
" 47

!0 " #%12# # #3 4 - # + -- ' - , ,7
" 47 .,

" #$%&% ' ((@ #) * @ # !# , 3. $ - 4 & + 3
" .-

0 " #%12# # #3 4 - # ! , 3. $ - 4 & + 3
" .- +A 1A 6 0 .,

" #$%&% ' ((@ #) * @ # !# -..

0 " #%12# # #3 4 - # @@ ! & " - 7 3
-.. .,
: .-4

0 " #%12# # #3 4 - # -< @ , .,

: .-4
: < , -..

0 " #%12# # #3 4 - # - (, .,
: @ .-4
: < , %.

0 " #%12# # #3 4 - # - , - .,
: @ .-4
: % ,

0 " #%12# # #3 4 - # - - - - .,
: .-4
: % - - & % " .

0 " #%12# # #3 4 - # < <=4<=@$$2%%%% %%P .,
: .-4
: % 9 7 / * - -

@0 " #%12# # #3 4 - # < <=4<=!, $%%%% %%P .,
: .-4
: % 9 7 / * - -

0 " #%12# # #3 4 - # ! @ ! .,
: @ .-4
: % 9 7 / * - -

!0 " #%12# # #3 4 - # ! ! ! .,
: .-4
: $. - % H- , -
: % &L$B +' &9A 9+ &31B 3F /1AA> %8/A)1%

0 " #%12# # #3 4 - # , . - - H - .,
: .-4

0 " #%12# # #3 4 - # 7 4 * - .,
: @ .-4

0 " #%12# # #3 4 - # 7 4 * - .,
: .-4

0 " #%12# # #3 4 - # 7 4 * - - .,
: .-4

0 " #%12# # #3 4 - #
-.. =,% H- , ;= , = - = =A * =8-= = - K - D .,
: .-4

0 " #%12# # #3 4 - # - -. .,
: .-4
: % 3883'+L 3F AA/%
&AA%&+ A 3' "3'+" A"&+3'

" #$%&% ' ((@ #) * @ #

" #$%&% ' ((@ #) * @ # #F A A . - - ,
A

0 " #%12# # #3 4 - # F A A . - - ,
A .,
: .-4

" #$%&% ' ((@ #) * @ # #/ - - , 4 - , (4
- - - < / , 7 - , / . 7

@0 " #%12# # #3 4 - # * / - -
, 4 - , (4 - - - < / , 7 - , / . 7 "- < ,4 L (7
6 0 .,
: @ .-4

" #$%&% ' ((@ #) * @ # #+ (- - - <7 > -,
P- <

0 " #%12# # #3 4 - # + (- - - <7
> -, P- < .,

" #$%&% ' ((@ #) * @ # # 4 7 1 4 - / -.

! 0 " #%12# # #3 4 - # 4 7 1 4 - / -.
"- < ,4 L (7 6 0 .,
: .-4

" #$%&% ' ((@ #) * @ #

" #$%&% ' ((@ #) * @ # #+ , - - -

0 " #%12# # #3 4 - # > + ,
- - - ' 9 - , 6 ! 0 .,

" #$%&% ' ((@ #) * @ # #& ; G , ; - ,
& .

0 " #%12# # #3 4 - # & ; G , ; - ,
& . .,

" #$%&% ' ((@ #) * @ # #& 1 , 8 4 8 < I
F 4

0 " #%12# # #3 4 - # @ & 1 , 8 4
8 < I F 4 @I

" #$%&% ' ((@ #) * @ # #& K . - K - D D - D

0 " #%12# # #3 4 - #
8 =) *=& K . - A K - D 2 - .,

" #$%&% ' ((@ #) * @ #
3< 9 7 G-

@

! "#$%" "

! "#$%" " "& ' (

) ! "#$%" " "& ' (" * + '
, - ' .- (/ 0 $ ' &1 / 2 , . 3) 4/

) ! "#$%" " "& ' ("
54 6$(.6+ ('6 67 8 / (6 4(4/
&9 : , / ;;;

) ! "#$%" " "& ' (" 4 < (
= > / > /((+ . . < (/ (3) 4/
?' @3 *) : ' A (, (/ 8 ' 5 .

) ! "#$%" " "& ' (" B ((7 (/
/. ! '. 4/

) ! "#$%" " "& ' (" B , (
((9. C (/ D(9 4/

@) ! "#$%" " "& ' (" ' . $ ' (
5 (4 !(9 /' 2 , . 3) 4/

*) ! "#$%" " "& ' (" @ + -
< / - (/ + 4 4/

) ! "#$%" " "& ' (" # E .
. (' (4/
?' 3) (((

) ! "#$%" " "& ' (" . ' 0 (4/

) ! "#$%" " "& ' (" . ' 0 (4/

&9 ! B)

20220412m

Aniversário do meu pai
85 anos
Vida longa, muitas felicidades e
saúde !!!

→ mathemata

Books-07-2022/20220413/
Important/
A History of Pythagoreanism.pdf
Pág. 13 (21)

Pythagorean analysis of the similarities between numbers and things inspired by the advancement of the mathemata, which Aristotle attributes to them;

Pág. 105 (113) –

Plato ascribes to Theodorus' (~~discoveries were further incorporated~~) a proof of irrationality of the magnitudes between √3 to √17 ⇒

(cansaço! troquei a linha rsrs

(Tht. 147 d = DK 43 A 4),
which means that
(significa que)
Theodorus relied on the proof
of the irrationality of $\sqrt{2}$
that was found by Hippasus.[62]

62 - For reconstructions of
Theodorus's proof, see, e.g.,
Heath 1921: I: 202ff;
Knorr 1975: 109ff

Obs: Heath, T. L. 1921
A History of Greek Mathematics
→ Pág. 471 (479) do livro
A History of Pythagoreanism.

Heath - Vol I Pág 202

outro assunto!

This curve will be best discussed when we come to deal with the problem of squaring the circle (Chapter VII);

here we need only remark that it implies the proportion that the lengths of arcs in a circle are proportional to the angles subtended by them at the centre

(Eucl. VI. 33)

---

trisecting any rectilineal angle ...

and these problems are not what the Greeks called 'plane' problems, i.e,

they cannot be solved by means of the ruler and compasses.

Hippasus — Pág. 65 (85)

Chapter III

Pythagorean Arithmetic

Hippasus → ... for making known the discovery of the irrational or incommensurable.

Pág. 160 (180) -

Hippasus

who was a Pythagorean but,
owing to his being the first
to publish and write down the
(construction of the) sphere
with (ἐκ, from)
the twelve pentagons, perished
by shipwreck for his (impiety),
but received credit for the
discovery...

Escanear amanhã!
Este caderno foi
escaneado até aqui
20220412 2228.

20220212m

20220213m

Escaneei o livro:

Desde el punto a la Cuarta
Dimensión
Egmont Colerus

→ está inserido no

Volume-014-20220213

Source:

ARQ/1-Source/Volume-014-20220213/

obs: Este caderno foi escaneado
até aqui
202202131050

20220214

Tarski's circle squaring problem.
https://annals.math.princeton.edu/
2017/186-2/p04

→ Andrew Marks
marks@math.ucla.edu

→ Spencer Unger
sunger@math.toronto.edu

=

spencer.unger@utoronto.ca

=

www.math.ucla.edu/~marks/
www.math.toronto.edu/sunger/

=

https://en.wikipedia.org/wiki/
Tarski's_circle-squaring_problem

—

https://www.maa.org/sites/default/
files/pdf/upload_library/22/Evans/
Pierce.pdf

https://arxiv.org/abs/2202.01912
↳ Circle Squaring with Pieces of Small Boundary and Low Borel Complexity.

=

Principal:
https://www.quantamagazine.org/an-ancient-geometry-problem-falls-to-new-mathematical-techniques-20220208/

=

↳ An Ancient Geometry Problem Fall to New Mathematical Techniques

=

https://www.degruyter.com/document/doi/10.1515/crll.1990.404.77/html

=

Equidecomposability and
discrepancy: a solution of
Tarski's circle-squaring
problem (M. Laczkovich)

=

Source:
https://doi.org/10.1515/
crll.1990.404.77

=

other source:
https://eudml.org/doc/153197
↳
https://gdz.sub.uni-goettingen.de/
id/PPN243919689_0404?
tify={"view":"info","pages":[81]}

Source: PURL
https://resolver.sub.uni-goettingen.de/
purl?PPN243919689_0404

=

- https://www.quantamagazine.org/
after-centuries-a-seemingly-
simple-math-problem-gets-an-
exact-solution-20201209/
- https://www.quantamagazine.org/
mathematicians-prove-30-year-
old-andre-oort-conjecture-20220203/
- https://www.quantamagazine.org/
with-a-new-shape-mathematicians-
link-geometry-and-numbers-
20210719/

https://arxiv.org/abs/1601.02958

Arquivos armazenados na pasta:
DATA-NovoVivo-003-2022/
Important=Squaring-Circle.
Copy: Books-003-2022/20220214/
Important=Squaring-Circle.

20220215 Fusion

deuterium-tritium fusion
↓
hydrogen-boron

→ aneutronic fusion (< temperature)

20220216

C.F. Gauss Werke: Briefwechsel
Mit Schumacher, 3 Volumes
US$225,00 Publisher: Georg Olms
ISBN 9783487056449 (January 1, 1975)

https://gdz.sub.uni-goettingen.de/id/PPN635102439

20220217

Simply Math DK.

20220219

Pi in the Sky - Barrow.

Mécanique Quantique III
Claude Aslangul

—

Kartheuser

—

→ Converter 20220225/Important/
Mathematical analysis.../

20220226
38,878.29 GB

—

https://gallica.bnf.fr/ark:/
12148/bpt6k64545935
→ Exercices de géométrie.

—

serge.mehl.free.fr/anx/
pb23_hilbert.html

—

https://td.mysuite.worker.dev

—

Geometry of Four Dimensions
Forsyth.

Euripedes - a vida não deve ser levada muito à sério.

20220227

pdf-magazines.me

novafile.com

white - enquanto

deterrent -> dissuasor.

Mathematical Techniques for Engineers and Scientists.

Pág. 275 (292)

6.2.1 Geometric Interpretation: The complex plane

Figure 6.1

(adoptar para planilha)

20220228

Quem está certo sempre vence.

20220228m

https://pt.scribd.com/document/
117480557/10-Piscoran-Final

Source:
www.cs.ubbcluj.ro/~studia-m/
2011-1/
10-piscoran-final.pdf

Obs: ISBN- 2065-961x

https://www.mpim-bonn.
mpg.de/node/144

↳ Algebra, Geometry and
Physics: a mathematical
mosaic.

fdown.net

20220301

meet.google.com/igi-ujrd-pef

→ Ungar

—

Sumo Sacerdote →
Cohen Hamashiach.

—

Melech Hamashiach
(Rei Messias)

—

20220302

I Reuty

—

20220303

Brane → an extended object
with any given number of
dimensions, of which strings
in string theory are examples
with one dimension.
Our Universe is a
3-brane

—

"Você não terá nada e será feliz" → comunismo.
→ Forum Mundial.

20220304

$1^3 = 4$ $\quad$ $2^3 = 8$

✓

lógico!!!

$1\,m^3 = \underline{1000}$

$\frac{1\,m^3}{3} = 333,\overline{3}$ $\quad$ $\frac{1\,m^3}{1^3} = \frac{1000}{4}$

3 $\quad$ $= \underline{\underline{250}}$

$$\begin{array}{r} 333,3 \\ -\,250 \\ \hline 83,3 \end{array}$$

333,3 — 100

83,3 — x

$= 25,00\overline{15}$ $(\cong \frac{1}{4})$

obs: $\frac{83,3}{333,3} =$

$0,\overline{2499}$

data / /
Ⓢ Ⓣ Ⓠ Ⓠ Ⓢ Ⓢ Ⓓ

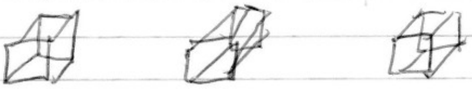

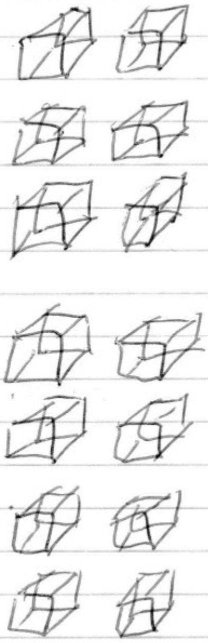

$2^1 = 2$

$\times 2$

$2^2 = 4$

$\times 2$

$2^3 = 8$

Se $2^3 = 8$

∴

$1^3 = 4$

$1^1 = 1 \quad 1^2 = 1 \quad 1^3 = 1$ (wrong!!!)

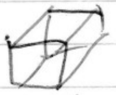

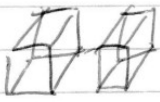

$1^1 = 1 \quad 1^2 = 2 \quad 1^3 = 4$

$2^1 = 2 \quad 2^2 = 4 \quad 2^3 = 8$

=

→ Wrong!!! do ponto de vista numérico, mas volumetricamente funciona

$1\,m^3 = 1000$ litros (3d)

↓

space

Adequação entre espaço e matemática.

Latex – Book
– Article
– Report.

20220204

\begin{text}
\HUGE
\end{text}

A ferramenta matemática tem por objetivo conhecer o real sentido relevante ~~das~~ ~~cada~~ palavras, symbols and numbers needed to express mathematical ideas.

⇒ A ferramenta matemática ~~tem~~ é uma ciência ~~que tem~~ da interpretação das palavras, símbolos e números necessários para expressar ideias matemáticas.

O conhecimento do real ~~do~~ sentido e valor da linguagem matemática tem por objetivo exprimir de forma precisa os conceitos inerentes a cada palavra empregada no texto matemático.

[illegible]

Nenhuma palavra, símbolo ou número está no texto sem significado ou função, pois cada uma delas contribui com a ideia que se quer

construir, demonstrar defender, refutar ou explicar.

parágrafo LATEX

\usepackage{indentfirst}

Para não fazer isto em um determinado parágrafo

\noindent

\emph{texto} → enfatizar

\emptyset

Book:
latex-via-exemplos.pdf
Pág. 10(26) - 3.3
Equação enumerada e referência cruzada

→ Dialogo di Galileo Galilei
Linceo.

Sidereus nuncius
Galileo Galilei
www.loc.gov/item/2010667904/

Il saggiatore

Il Vero e il falso
Antonino Zichichi
↳ passeggiando tra le
stelle e a casa.
978-8842810780 547 pag.

La forma dello spazio
profondo
Shing-Tung, Steve Nadis

ISBN 978-8842815365

https://philarchive.org/archive/

https://archive.org/details/
ilveroeilfalsopa0000zich

Elements of geometry
Ladd, Kelly

ATA NOTARIAL

↓ ↳ 29º Cartório NOTAS
SP.

Perpetuar conteúdo de
páginas na Internet.

20220306m

data

S T Q Q S S D

20220206000511

$1^3 = 3$ ou 4

$\underbrace{1}_{(01)} \times \underbrace{1}_{(02)} \times \underbrace{1}_{(03)} = \boxed{3}$ (3 unidades)

3 unidades

Usando da mesma lógica:

$2^3 = 8$

$\underbrace{2 \times 2 \times 2}_{3 \text{ elementos}} = 8$ (valor em unidades)

O número 2 é o dobro de 1, portanto, se $2^3 = 8$ e o número 1 é a metade de 2^3, portanto, $1^3 = \boxed{4}$

Prosseguindo (com base na razão no número real (exponenciando)

~~2~~ $3^3 = 27$

$3 \times 3 \times 3 = 27 \Rightarrow 1 = \frac{1}{3}$ de 27

$\frac{27}{3} = \boxed{9}$

$4^3 = 64 \Rightarrow 1 \Rightarrow \frac{1}{4}$ de $64 = \frac{64}{4} = \boxed{16}$

$5^3 = 75 \Rightarrow 1 = \frac{1}{5}$ de $125 = \frac{125}{5} = \boxed{25}$

$6^3 = 216 \Rightarrow 1 = \frac{1}{6}$ de $216 = \frac{216}{6} = \boxed{36}$

$7^3 = 343' \Rightarrow 1 = \frac{1}{7}$ de $343 = \frac{343}{7} = \boxed{49}$

$8^3 = 512 \Rightarrow 1 = \frac{1}{8}$ de $512 = \frac{512}{8} = \boxed{64}$

$9^3 = 729 \Rightarrow 1 = \frac{1}{9}$ de $729 = \frac{729}{8} = \boxed{81}$

$10^3 = 1000 \Rightarrow 1 = \frac{1}{10}$ de $1000 = \frac{1000}{10} = \boxed{100}$

Divisão por 4:

$1 \times 1 \times 1 = 4$

$2 \times 2 \times 2 = 8 \; > \frac{8}{4} = 2$ (dobro)

$3 \times 3 \times 3 = 27 \; > \frac{27}{8} = 3,375$

$4 \times 4 \times 4 = 64 \; > \frac{64}{27} = 2,\overline{370}$

$5 \times 5 \times 5 = 125 \; > \frac{125}{64} = 1,953125$ (1,953125)

$6 \times 6 \times 6 = 216 \; > \frac{216}{75} = 2,88$

$7 \times 7 \times 7 = 343 \; > \frac{343}{216} = 1,58796296$

$8 \times 8 \times 8 = 512 \; > \frac{512}{343} = 1,49271137$

$9 \times 9 \times 9 = 729 \; > \frac{729}{512} = 1,42382812$

$10 \times 10 \times 10 = 1000 \; > \frac{1000}{729} = 1,37174211$

entre os números reais

$\cong \sqrt{2} = 1,414217$..

Relação direta entre os números.

$1 \rightarrow 2 \Rightarrow 8 \div 4 = 2$
$1 \rightarrow 3 \Rightarrow 27 \div 4 = 6,75$
$1 \rightarrow 4 \Rightarrow 64 \div 4 = 16$
$1 \rightarrow 5 \Rightarrow 125 \div 4 = 31,25$
$1 \rightarrow 6 \Rightarrow 216 \div 4 = 54$
$1 \rightarrow 7 \Rightarrow 343 \div 4 = 85,75$
$1 \rightarrow 8 \Rightarrow 512 \div 4 = 128$
$1 \rightarrow 9 \Rightarrow 729 \div 4 = 182,25$
$1 \rightarrow 10 \Rightarrow 1000 \div 4 = 250$

Se em vez de $4 \rightarrow 3$

$1 \times 1 \times 1 = 3$

$8 \div 3 = 2,\overline{6}$
$27 \div 3 = 9$
$64 \div 3 = 21,\overline{3}$
$125 \div 3 = 41,\overline{6}$ (41,6)
$216 \div 3 = 72$
$343 \div 3 = 114,\overline{3}$
$512 \div 3 = 170,6$
$729 \div 3 = 243$
$1000 \div 3 = 333,\overline{3}$

Voltando:

$1^3 = 4 \leftarrow \left(\frac{8}{2} = 4\right)$

$2^3 = 8 \rightarrow$ 2 é o dobro de ①

$6,75 - 2 = 4,75$
$16 - 6,75 = 9,25$
$31,25 - 16 = 15,25$
$54 - 31,25 = 22,75$
$85,75 - 54 = 31,75$
$128 - 85,75 = 42,25$
$182,25 - 128 = 54,25$
$250 - 182,25 = 67,75$

$\Downarrow$

4,75 > 4,5
9,25 > 6
15,25 > 7,5
22,75 > 9
31,75 > 10,5
42,25 > 12
54,25 > 13,5
67,75

diferenças de 1,5 em ⑴,5 N

data

$1^3=4$

$2^3=8 \rightarrow 2$ é o dobro 1 $= 8 \div 4 = 2$

$3^3=27 \rightarrow 3$ é o triplo 1 $= 27 \div 4 = 6,75$

$4^3=64 \rightarrow 4 \times 1 = 64 \div 4 =$

$5^3=175 \rightarrow 5 \times 1 = 175 \div 4 =$

$6^3=216 \rightarrow 6 \times 1 = 216 \div 4 =$

$7^3=343 \rightarrow 7 \times 1 = 343 \div 4 =$

$8^3=512 \rightarrow 8 \times 1 = 512 \div 4 =$

$9^3=729 \rightarrow 9 \times 1 = 729 \div 4 =$

$10^3=1000 \rightarrow 10 \times 1 = 1000 \div 4 =$

corrigindo

$8 \div 4$ = erro

$8 \div 2 = 4 \leftarrow (1^3)$ > 5
$27 \div 3 = 9$ > 7
$64 \div 4 = 16$ > 9
$125 \div 5 = 25$ > 11
$216 \div 6 = 36$ > 13
$343 \div 7 = 49$ > 15
$512 \div 8 = 64$ > 17
$729 \div 9 = 81$ > 19
$1000 \div 10 = 100$

outro erro!!!

Correto

O valor aumenta de 2 em 2

conjunto:

Uzando a multiplicação partindo-se do número $1^3=4$,

depois faremos $1^3=3$

2x

$1^3 = \boxed{4}$

$2^3 = \boxed{4} \times 2 = 8$

×3

$3^3 = \boxed{4} \times 3 = 12$

$4^3 = \boxed{4} \times 4 = 16$

$5^3 = \boxed{4} \times 5 = 20$

$6^3 = \boxed{4} \times 6 = 24$

$7^3 = \boxed{4} \times 7 = 28$

$8^3 = \boxed{4} \times 8 = 32$

$9^3 = \boxed{4} \times 9 = 36$

$10^3 = \boxed{4} \times 10 = 40$

mantendo-se
a proporção,
(pensamento
equivocado com
potenciação
↓
estudar!!!

mas numericamente
está correto.
(?)
checar!!!

Valores reais:

4 $8 \div 4 = 2$
8
27 $27 \div 4 = 6{,}75$
64 $64 \div 4 = 16$
125 $125 \div 4 = 31{,}25$
216 $216 \div 4 = 54$
343 $343 \div 4 = 85{,}75$
512 $512 \div 4 = 128$
729 $729 \div 4 = 182{,}25$
1000 $1000 \div 4 = 250$

discrepância.

Simulações:

$1^3 = 1 \times 1 \times 1 = 3$

$2^3 = 2 \times 2 \times 2 = 8$ $\Big\} \; 8 \div 3 = 2,\overline{6}$

$3^3 = 27 \rightarrow 27 \div 3 = 9$

$4^3 = 64 \rightarrow 64 \div 3 = 21,\overline{3}$

$5^3 = 125 \rightarrow 125 \div 3 = 41,\overline{6}$

$6^3 = 216 \rightarrow 216 \div 3 = 72$

$7^3 = 343 \rightarrow 343 \div 3 = 114,\overline{3}$

$8^3 = 512 \rightarrow 512 \div 3 = 170,\overline{6}$

$9^3 = 729 \rightarrow 729 \div 3 = 243$

$10^3 = 1000 \rightarrow 1000 \div 3 = 333,\overline{3}$

Aprofundando:

começando desde o número

$1^3 = \overset{1}{4}$

$1^3 = 4 \Rightarrow 4 \div 1 = 4$

$2^3 = 8 \Rightarrow 8 \div 2 = 4$ > $\boxed{4-4 = 0}$

$3^3 = 27 \Rightarrow 27 \div 3 = 9$ > 5

$4^3 = 64 \Rightarrow 64 \div 4 = 16$ > 7

$5^3 = 125 \Rightarrow 125 \div 5 = 25$ > 9

$6^3 = 216 \Rightarrow 216 \div 6 = 36$ > 11

$7^3 = 343 \Rightarrow 343 \div 7 = 49$ > 13

$8^3 = 512 \Rightarrow 512 \div 8 = 64$ > 15

$9^3 = 729 \Rightarrow 729 \div 9 = 81$ > 17

$10^3 = 1000 \Rightarrow 1000 \div 10 = 100$ > 19

relação direta perfeita!!! (em relação direta)

aumento de 2 em 2"

Obs:

Mantendo-se esta constante e retroagindo até o número 1,

$1^3 = 1$, pois $2^3 = 8 \div 2 = 4$

$4 - 3 = 1$

e não $1^3 = 3$ ou 4"

data

S T Q Q S S D

Comparando ⇒ 3 e 4

4	8÷4 = 2	3	8÷3 = 2,6
8		8	
27	27÷4 = 6,75	27	27÷3 = 9
64	64÷4 = 16	64	64÷3 = $21,\overline{3}$
125	125÷4 = 31,25	125	125÷3 = $41,\overline{6}$
216	216÷4 = 54	216	216÷3 = 72
343	343÷4 = 85,75	343	343÷3 = $114,\overline{3}$
512	512÷4 = 128	512	512÷3 = $170,\overline{6}$
729	729÷4 = 182,25	729	729÷3 = 243
1000	1000÷4 = 250	1000	1000÷3 = $333,\overline{3}$

estudar com mais
profundidade!!!

Prof. Stenio Murich

202203050200 ///

202203061530$_m$

$$1^3 = 1 \times \underset{\cong}{1} \times 1 = 3$$

$$1 + 1 + 1 = 3_u$$

$$1^4 = 1 \times 1 \times \cancel{1} \times 1 = 4$$

$$1 + 1 + 1 + 1 = 4_m$$

$$1^4 = \frac{1}{2} \cdot 2^3$$

obs:

$$1^3 = \underset{(1+1)}{\underset{=}{1 \times 1}} = \overbrace{2 \times}^{2 \times 1 = \boxed{2}} \underset{(2+1)}{\underset{=}{1}} = 3$$

, pode ser 2 (testar)

⇒ vejamos

$$1^3 = 2$$

$1^4 =$

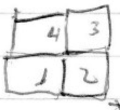

= base do cubo

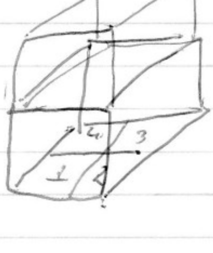

$$\left.\begin{matrix} = 4 \\ + \\ = 4 \end{matrix}\right\} 8 = 2^3$$

20220307/// 1226//

$1^3 = 1 \Rightarrow 1^3 = 1 \times 4 \Rightarrow 2^3 = 8$

—

$1^4 = 4$ (?)

$2^3 = 8$

—

$2^2 = 4$ (base do cubo) $\downarrow$

$2^3 = 8$

—

$$\begin{cases} 1^3 = 1 \\ 2^2 = 4 \\ 2^3 = 8 \end{cases}$$ ///

—

20220308

ARQ-BASICS → fazer down again.

art1lib.org

The Mathematical Universe
Pag 28(41)
Pi in the sky.
(rodapé).

$$\left\{\begin{array}{l}\boxed{01}\,\boxed{06}\ 1969\ (Br)\\ \underbrace{\boxed{06}\,\boxed{01}}\ 1969\ (USA)\end{array}\right.$$

mesmo valor (soma)

Einstein March 14, 1879

$\underbrace{03\ \underline{14}}_{Pi}$

obs:

British prefere
July 22/7 for Pi Day/ or
$\frac{22}{7} = 3.14$ to two decimal places-
As an aside that also happens to be the birthday of
German mathematician and physicist
Friedrich Wilhelm Bessel
(1784 - 1846),
whom we encounter in footnote 23 of Chapter 2.

The Bessel differential equation and subsequent solutions (Bessel functions) play an important role in many branches in physics.

20220309m

palíndromo

ambigrama

20220310

archive.org/details/@mirtitles

RASPUTITSA ⇒ tempo das estradas ruins. (para tanks)

⇓

ocorreu na Invasão Nazista da Rússia.

20220311

SMHD2022 - Partilhado (X)

Einstein } baseado nos estudos Lorentz and Poincaré
Minkowski

Pecado 3 fases { 1) sugestão
2) deleite
3) aceitação

20220315m

New Directory:

Data-NovoVivo-006-2022

—

booksc.me

—

manuelsanciens.blogspot.com

—

Data-NovoVivo-005-2022

→ cópia da obra de Galileo Galilei

—

20220318m

ARQ-ALL--20220318.7z

→ Research material

—

20220319m

mgsuite.eu.org obs: era

— td.mgsuite.workers.dev

—

x2download.com

—

20220324 M

Einstein saw symmetry hidden in the fabric of space and time.

20220326 M

libgen.is

libgen.rs / libgen.st

20220327 M COCHISE

https://drive.google.com/drive/u/1/folders/0AO-xX4BHvgjsUk9PVA

↳ zero ↳ ouro.

Ernest E. Shult

ernest_shult@yahoo.com

Johannes Ueberberg

johannes.ueberberg@gmx.de

Incidence Geometry

20220329III   Exclui extensão.

Programa → regedit (no campo de execução do Windows)

HKey_Current_User
↳ Software → Microsoft →
Windows →
CurrentVersion →
Explorer →
FileExts

HKey_Classes_Root →
↳ exclui a extensão.

—

Reinicie a máquina
Execute a extensão e depois
escolha o programa
que você quer usar para
abri-lo.

—

20220330III

Pesach   April 15

Chag Pessach Sameach

—

Legendre - Elementos de Geometria
https://bd.camara.leg.br/bd/handle/
bdcamara/22882

https://bdlb.bn.gov.br/acervo/
handle/20.500.12156.3/442545

Math Books:
https://www.doraci.com.br/
downloads/matematica/
fund-mat-elem_01.pdf
11.pdf

20220331//
spruce up = enfeitar.

backblaze.com

Trigonometric Functions
and Complex Numbers
Dezheng Yang
ISBN 9781958134869
https://lccn.loc.gov/
2016026528

→ practices makes perfect.

20220410m

Copie todos os arquivos
para uma pasta só.
(diretório)

Obs: 89,43TB

20220411m
balanite Milose → Cimeworte
25,00

20220411/ The Rise of Modern Logic.../
Philosophy of Physics - Part A.pdf
Pág. 225(254)-
"geometrized"
→ rodapé
3 - A review of the
needed differential geometry...

geometrized version of Newtonian
gravitation theory (also known as
Newton-Cartan theory).

data . .
Ⓢ Ⓣ Ⓠ Ⓠ Ⓢ Ⓢ Ⓓ

20220412m

Aniversário do meu pai
85 anos
Vida longa, muitas felicidades e
saúde !!!

→ mathemata

Books-07-2022/20220413/
Important/
A History of Pythagoreanism.pdf
Pág. 13 (21)

a Pythagorean analysis of the similarities between numbers and things inspired by the advancement of the mathemata, which Aristotle attributes to them;

Pág. 105 (113) –

Plato ascribes to Theodorus' ~~(discoveries were further incorporated)~~ a proof of irrationality of the magnitudes between $\sqrt{3}$ to $\sqrt{17}$ ⟹

Cansaço! Trocando até linha rsrs

jandaia

(Tht. 147 d = DK 43 A 4),
which means that
(significa que)
Theodorus relied on the proof
of the irrationality of $\sqrt{2}$
that was found by Hippasus.[62]

62 - For reconstructions of
Theodorus's proof, see, e.g,
Heath 1921: I: 202ff;
Knorr 1975: 109ff

Obs: Heath, T.L. 1921
A History of Greek Mathematics
→ Pág, 471 (479) do livro
A History of Pythagoreanism.

Heath - Vol I Pág 202

outro assunto!

This curve will be best discussed when we come to deal with the problem of squaring the circle (Chapter VII);

here we need only remark that it implies the proposition that the lengths of arcs in a circle are proportional to the angles subtended by them at the centre

(Eucl. VI. 33)

---

trisecting any rectilineal angle

...

and these problems are not what the Greeks called 'plane' problems, i.e,

they cannot be solved by means of the ruler and compasses.

Hippasus – Pág. 65 (85)

Chapter III

Pythagorean Arithmetic

Hippasus → ... for making known the discovery of the irrational or incommensurable.

Pág. 160 (180) -

Hippasus

who was a Pythagorean but,
owing to his being the first
to publish and write down the
(construction of the) sphere
with (ἐκ, from)
the twelve pentagons, perished
by shipwreck for his (impiety),
but received credit for the
discovery...

Escanear amanhã!
Este caderno foi
escaneado até aqui
202204122228.

F((+ (+9. - 4('
*

20220212m

20220213m

Escaneei o livro:

Desde el punto a la Cuarta
Dimensión
Egmont Colerus

→ está inserido no
Volume-014-20220213

Source:
ARQ/1-Source/Volume-014-20220213/

---

obs: Este caderno foi escaneado
até aqui
202202131050

20220214

Tarski's circle squaring problem.
https://annals.math.princeton.edu/
2017/186-2/p04

→ Andrew Marks
marks@math.ucla.edu

→ Spencer Unger
sunger@math.toronto.edu

=

spencer.unger@utoronto.ca

=

www.math.ucla.edu/~marks/
www.math.toronto.edu/sunger/

=

https://en.wikipedia.org/wiki/
Tarski's_circle-squaring_problem

-

https://www.maa.org/sites/default/
files/pdf/upload_library/22/Evans/
Pierce.pdf

https://arxiv.org/abs/2202.01412

↳ Circle Squaring with Pieces of Small Boundary and Low Borel Complexity.

=

Principal:

https://www.quantamagazine.org/an-ancient-geometry-problem-falls-to-new-mathematical-techniques-20220208/

=

↳ An Ancient Geometry Problem Fall to New Mathematical Techniques

=

https://www.degruyter.com/document/doi/10.1515/crll.1990.404.77/html

=

Equidecomposability and discrepancy: a solution of Tarski's circle-squaring problem (M. Laczkovich)

═

Source:
https://doi.org/10.1515/crll.1990.404.77

═

other source:
https://eudml.org/doc/153197
↳
https://gdz.sub.uni-goettingen.de/id/PPN243919689_0404?tify={"view":"info","pages":[81]}

Source: PURL
https://resolver.sub.uni-goettingen.de/purl?PPN243919689_0404

═

- https://www.quantamagazine.org/
after-centuries-a-seemingly-
simple-math-problem-gets-an-
exact-solution-20201209/
- https://www.quantamagazine.org/
mathematicians-prove-30-years-
old-andre-oort-conjecture-20220203/
- https://www.quantamagazine.org/
with-a-new-shape-mathematician-
link-geometry-and-numbers-
20210719/

—

https://arxiv.org/abs/1601.02958

Arquivos armazenados na pasta:
DATA-NovoVivo-003-2022/
Important = Squaring-Circle.
Copy: Books-003-2022/20220214/
Important = Squaring-Circle!

20220215 Fusion

deuterium-tritium fusion
↓
hydrogen-bomb

→ aneutronic fusion (< temperature)

—

20220216

C.F. Gauss Werke: Briefwechsel
Mit Schumacher, 3 Volumes
US$ 225,00 Publisher: Georg Olms
ISBN 9783487056449 (January 1, 1975)

https://gdz.sub.uni-goettingen.de/id/PPN635102439

—

20220217

Simply Math DK.

20220219

Pi in the Sky - Barrow.

Mécanique Quantique III
Claude Aslangul

—

Kartheuser

—

⇒ Converter 20220225/Important/
Mathematical analysis..../

20220226
38,878.29 GB

—

https://gallica.bnf.fr/ark:/
12148/bpt6k64545935
→ Exercices de géométrie.

—

serge.mehl.free.fr/anx/
pb23_hilbert.html

—

https://td.mgsuite.worker.dev

—

Geometry of Four Dimensions
Forsyth.

Eurípedes – a vida não deve ser
levada muito à sério.

20220227

pdf-magazines.me

novafile.com

---

while – enquanto

deterrent → dissuasor.

Mathematical Techniques for
Engineers and Scientists.

Pág. 275 (292)

6.2.1 Geometric
Interpretation:
the complex plane

Figure 6.1

(adaptar para geometria)

20220228

Quem está certo sempre vence.

20220228m

https://pt.scribd.com/document/117480557/10-Piscoran-Final

Source:

www.cs.ubbcluj.ro/~studia-m/2011-1/10-piscoran-final.pdf

Obs: ISBN- 2065-961x

https://www.mpim-bonn.mpg.de/node/144

↳ Algebra, Geometry and Physics: a mathematical mosaic.

fdown.net

20220301 M

meet.google.com/igi-ujrd-pef

→ Ungar

—

Sumo Sacerdote →
Cohen Hamashiach.

—

Melech Hamashiach
(Rei Messias)

—

20220302

I Rewly

—

20220303

Brane → an extended object with any given number of dimensions, of which strings in string theory are examples with one dimension.
Our Universe is a 3-brane

—

"Você não terá nada e será feliz" → comunismo.
→ Forum Mundial.

20220304

$1^3 = 4$ $2^3 = 8$

✓

lógico!!!

$1\,m^3 = \underline{1000}$

$\frac{1\,m^3}{3} = 333,\bar{3}$ $\frac{1\,m^3}{1^3} = \frac{1000}{4}$

$= \underline{\underline{250}}$

3

$$\begin{array}{r} 333,3 \\ -\ 250 \\ \hline \underline{\underline{83,3}} \end{array}$$

$333,3 - 100$

$83,3 - x$

$= 25,00\overline{15}\ (\cong \frac{1}{4})$

off $\frac{83,3}{333,3} = $

$0,2499$

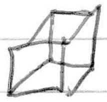
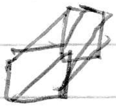
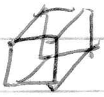

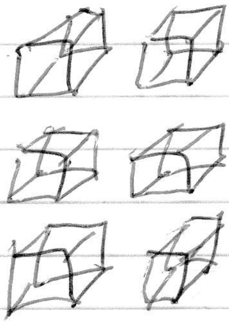

$2^1 = 2$

x2

$2^2 = 4$

x2

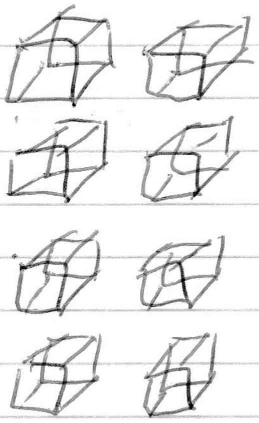

$2^3 = 8$

Se $2^3 = 8$

$1^3 = 4$

$1^1 = 1$ $1^2 = 1$ $1^3 = 1$ (Wrong!!!)

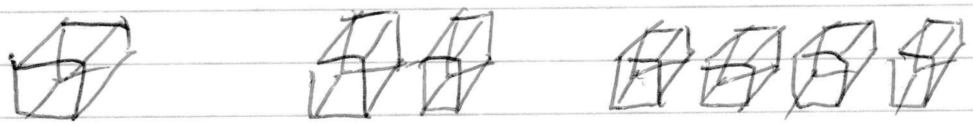

$1^1 = 1$ $1^2 = 2$ $1^3 = 4$

$2^1 = 2$ $2^2 = 4$ $2^3 = 8$

→ Wrong!!! do ponto de vista numérico, mas volumetricamente funciona

$1 m^3 = 1000$ litros (3d)

↓

Space

Adequação entre espaço e matemática.

Latex - Book - Article - Report.

20220204

\begin{text}
\HUGE
\end{text}

A ferramentas matemática tem por objetivo conhecer o real sentido e valor das ~~cada~~ palavras, symbols and numbers needed to express mathematical ideas.

⇒ A ferramentas matemáticas ~~tem~~ é uma ciência ~~que tem~~ da interpretação das palavras, símbolos e números necessários para expressar ideias matemáticas.

O conhecimento do real ~~do~~ sentido e valor da linguagem matemática tem por objetivo exprimir de forma precisa os conceitos inerentes a cada palavra empregada no texto matemático.

inequívoca

Nenhuma palavra, símbolo ou número está no texto sem significado ou função, pois cada uma delas contribui com a ideia que se quer

construir, demonstrar defender, refutar ou explicar.

---

parágrafo LATEX

\usepackage{indentfirst}

Para não fazer isto em um determinado parágrafo

\noindent

\emph{texto} → enfatizar

\emptyset

Book:

latex-via-exemplos.pdf

Pág. 10(26) - 3.3

Equação enumerada e referência cruzada

→ Dialogo di Galileo Galilei
Linceo.

Sidereus nuncius
Galileo Galilei
www.loc.gov/item/2010667904/

Il saggiatore

Il Vero e il falso
Antonino Zichichi
↳ passeggiando tra le stelle e a casa.
978-8842810780 547 pag.

La forma dello spazio profondo
Shing-Tung, Steve Nadis

ISBN 978-8842815365

---

https://philarchive.org/archive/

https://archive.org/details/
ilveroeilfalsopa0000zich

—

Elements of geometry
Ladd, Kelly

—

ATA NOTARIAL

↓ ↳ 23º Cartório NOTAS -
SP-

Perpetuar conteúdo de
páginas na Internet.

—

20220306 m

202202060005

$1^3 = 3$ ou 4

$1 \times 1 \times 1 = \boxed{3}$ (3 unidades)

$\underset{(01)}{1} \quad \underbrace{\underset{(02)}{1} \quad \underset{(03)}{1}}_{3 \text{ unidades}}$

Usando da mesma lógica.

$2^3 = 8$

$\underbrace{2 \times 2 \times 2}_{3 \text{ elementos}} = 8$ (valor em unidades)

O número 2 é o dobro de 1, portanto, se $2^3 = 8$ e o número 1 é a metade de 2^3, portanto, $1^3 = \boxed{4}$

Prosseguindo (com base na razão no número real (exponenciando)

~~2~~ $3^3 = 27$

$3 \times 3 \times 3 = 27 \Rightarrow 1 = \frac{1}{3}$ de 27

$\frac{27}{3} = \boxed{9}$

$4^3 = 64 \Rightarrow 1 \Rightarrow \frac{1}{4}$ de $64 = \frac{64}{4} = \boxed{16}$

$5^3 = 75 \Rightarrow 1 = \frac{1}{5}$ de $125 = \frac{125}{5} = \boxed{25}$

$6^3 = 216 \Rightarrow 1 = \frac{1}{6}$ de $216 = \frac{216}{6} = \boxed{36}$

$7^3 = 343 \Rightarrow 1 = \frac{1}{7}$ de $343 = \frac{343}{7} = \boxed{49}$

$8^3 = 512 \Rightarrow 1 = \frac{1}{8}$ de $512 = \frac{512}{8} = \boxed{64}$

$9^3 = 729 \Rightarrow 1 = \frac{1}{9}$ de $729 = \frac{729}{8} = \boxed{81}$

$10^3 = 1000 \Rightarrow 1 = \frac{1}{10}$ de $1000 = \frac{1000}{10} = \boxed{100}$

Divisão por 4:

$1 \times 1 \times 1 = 4$

$2 \times 2 \times 2 = 8 \quad \rangle \ \frac{8}{4} = 2$ (dobro)

$3 \times 3 \times 3 = 27 \quad \rangle \ \frac{27}{8} = 3,375$

$4 \times 4 \times 4 = 64 \quad \rangle \ \frac{64}{27} = 2,\overline{370}$

$5 \times 5 \times 5 = 125 \quad \rangle \ \frac{125}{64} = 1,953125$ (1,953125)

$6 \times 6 \times 6 = 216 \quad \rangle \ \frac{216}{75} = 2,88$

$7 \times 7 \times 7 = 343 \quad \rangle \ \frac{343}{216} = 1,58796296$

$8 \times 8 \times 8 = 512 \quad \rangle \ \frac{512}{343} = 1,49271137$

$9 \times 9 \times 9 = 729 \quad \rangle \ \frac{729}{512} = 1,42382812$

$10 \times 10 \times 10 = 1000 \quad \rangle \ \frac{1000}{729} = 1,37174211$

entre os números reais

$\cong \sqrt{2} = 1,414217$..

Relação direta entre os números.

$1 \rightarrow 2 \Rightarrow 8 \div 4 = 2$
$1 \rightarrow 3 \Rightarrow 27 \div 4 = 6,75$
$1 \rightarrow 4 \Rightarrow 64 \div 4 = 16$
$1 \rightarrow 5 \Rightarrow 125 \div 4 = 31,25$
$1 \rightarrow 6 \Rightarrow 216 \div 4 = 54$
$1 \rightarrow 7 \Rightarrow 343 \div 4 = 85,75$
$1 \rightarrow 8 \Rightarrow 512 \div 4 = 128$
$1 \rightarrow 9 \Rightarrow 729 \div 4 = 182,25$
$1 \rightarrow 10 \Rightarrow 1000 \div 4 = 250$

$6,75 - 2 = 4,75$
$16 - 6,75 = 9,25$
$31,25 - 16 = 15,25$
$54 - 31,25 = 22,75$
$85,75 - 54 = 31,75$
$128 - 85,75 = 42,25$
$182,25 - 128 = 54,25$
$250 - 182,25 = 67,75$

$\Downarrow$

$4,75 > 4,5$
$9,25 > 6$
$15,25 > 7,5$
$22,75 > 9$
$31,75 > 10,5$
$42,25 > 12$
$54,25 > 13,5$
$67,75$

diferenças de 1,5 em ⓵,⓹ N

Se em vez de 4 → 3

$1 \times 1 \times 1 = 3$

$8 \div 3 = 2,\overline{6}$
$27 \div 3 = 9$
$64 \div 3 = 21,\overline{3}$
$125 \div 3 = 41,6 \quad (41,\overline{6})$
$216 \div 3 = 72$
$343 \div 3 = 114,\overline{3}$
$512 \div 3 = 170,\overline{6}$
$729 \div 3 = 243$
$1000 \div 3 = 333,\overline{3}$

Voltando:

$1^3 = 4 \leftarrow \left(\frac{8}{2} = 4\right)$

$2^3 = 8 \rightarrow 2$ é o dobro de ①

$1^3 = 4$

$2^3 = 8$ → 2 é o dobro 1 = 8 ÷ 4 = 2
$3^3 = 27$ → 3 é o triplo 1 = 27 ÷ 4 = 6,75
$4^3 = 64$ → 4 x 1 = 64 ÷ 4 =
$5^3 = 125$ → 5 x 1 = 125 ÷ 4 =
$6^3 = 216$ → 6 x 1 = 216 ÷ 4 =
$7^3 = 343$ → 7 x 1 = 343 ÷ 4 =
$8^3 = 512$ → 8 x 1 = 512 ÷ 4 =
$9^3 = 729$ → 9 x 1 = 729 ÷ 4 =
$10^3 = 1000$ → 10 x 1 = 1000 ÷ 4 =

corrigindo

8 ÷ 4 = erro

✡

8 ÷ 2 = 4 ← (1^3) > 5
27 ÷ 3 = 9 > 7
64 ÷ 4 = 16 > 9
125 ÷ 5 = 25 > 11
216 ÷ 6 = 36 > 13
343 ÷ 7 = 49 > 15
512 ÷ 8 = 64 > 17
729 ÷ 9 = 81 > 19
1000 ÷ 10 = 100

outro erro!!!

Correto

O valor aumenta de 2 em 2 ///

conjunto:

Usando a multiplicação partindo-se do número $1^3 = 4$,
depois faremos $1^3 = 3$ //

2x

$1^3 = \boxed{4}$

$2^3 = \boxed{4} \times 2 = 8$

x3

$3^3 = \boxed{4} \times 3 = 12$

$4^3 = \boxed{4} \times 4 = 16$

$5^3 = \boxed{4} \times 5 = 20$

$6^3 = \boxed{4} \times 6 = 24$

$7^3 = \boxed{4} \times 7 = 28$

$8^3 = \boxed{4} \times 8 = 32$

$9^3 = \boxed{4} \times 9 = 36$

$10^3 = \boxed{4} \times 10 = 40$

mantendo-se
a proporção,
(o pensamento
equivocado com
potenciação
estudar!!!

mas numericamente
está correto.
(?)
Checar!!!

Valores reais:

4 $8 \div 4 = 2$
8
27 $27 \div 4 = 675$
64 $64 \div 4 = 16$
125 $125 \div 4 = 31,25$
216 $216 \div 4 = 54$
343 $343 \div 4 = 85,75$
512 $512 \div 4 = 128$
729 $729 \div 4 = 182,25$
1000 $1000 \div 4 = 250$

discrepância.

Simulações:

$1^3 = 1 \times 1 \times 1 = 3$ $\rangle$ $8 \div 3 = 2,\overline{6}$

$2^3 = 2 \times 2 \times 2 = 8$

$3^3 = 27 \rightarrow 27 \div 3 = 9$

$4^3 = 64 \rightarrow 64 \div 3 = 21,\overline{3}$

$5^3 = 125 \rightarrow 125 \div 3 = 41,\overline{6}$

$6^3 = 216 \rightarrow 216 \div 3 = 72$

$7^3 = 343 \rightarrow 343 \div 3 = 114,\overline{3}$

$8^3 = 512 \rightarrow 512 \div 3 = 170,\overline{6}$

$9^3 = 729 \rightarrow 729 \div 3 = 243$

$10^3 = 1000 \rightarrow 1000 \div 3 = 333,\overline{3}$

Aprofundando:

começando desde o
número

1

$1^3 = 4$

$1^3 = 4 \Rightarrow 4 \div 1 = 4$

$4 - 4 = 0$

relação direta perfeita!!! (em relação direta)

$2^3 = 8 \Rightarrow 8 \div 2 = 4$ > 5

$3^3 = 27 \Rightarrow 27 \div 3 = 9$ > 7

$4^3 = 64 \Rightarrow 64 \div 4 = 16$ > 9

$5^3 = 125 \Rightarrow 125 \div 5 = 25$ > 11

$6^3 = 216 \Rightarrow 216 \div 6 = 36$ > 13

$7^3 = 343 \Rightarrow 343 \div 7 = 49$ > 15

$8^3 = 512 \Rightarrow 512 \div 8 = 64$ > 17

$9^3 = 729 \Rightarrow 729 \div 9 = 81$ > 19

$10^3 = 1000 \Rightarrow 1000 \div 10 = 100$

aumento de 2 em 2

Obs:

Mantendo-se esta constante e retroagindo até o número 1,

$1^3 = 1$, pois $2^3 = 8 \div 2 = 4$

$4 - 3 = 1$

e NÃO $1^3 = 3$ ou 4.

Comparando → 3 e 4

4	8÷4=2	3	8÷3=2,6
8		8	
27	27÷4=6,75	27	27÷3=9
64	64÷4=16	64	64÷3=21,$\overline{3}$
125	125÷4=31,25	125	125÷3=41,$\overline{6}$
216	216÷4=54	216	216÷3=72
343	343÷4=85,75	343	343÷3=114,$\overline{3}$
512	512÷4=128	512	512÷3=170,$\overline{6}$
729	729÷4=182,25	729	729÷3=243
1000	1000÷4=250	1000	1000÷3=333,$\overline{3}$

$\underbrace{\qquad\qquad}$

estudar com mais profundidade!!!

Prof. Stenio Murich

202203050200 ///

202203061530

$1^3 = 1 \times \underset{\cong}{1} \times 1 = 3$

$1 + 1 + 1 = 3$

$1^4 = 1 \times 1 \times 1 \times 1 = 4$

$1 + 1 + 1 + 1 = 4$

$1^4 = \frac{1}{2} \cdot 2^3$

obs:

$$1^3 = \underset{=\,(1+1)}{1 \times 1} = \overbrace{\underset{=\,(2+1)}{2 \times 1}}^{2 \times 1 = \boxed{2}} = 3$$

, pode ser 2 (testar)

⇒ vejamos

$1^3 = 2$

$1^4 =$ 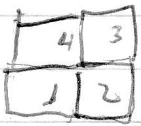 = base do cubo

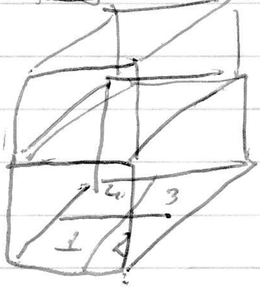

$$\left.\begin{matrix} = 4 \\ + \\ = 4 \end{matrix}\right\} 8 = 2^3$$

20220307/// 1226///

$1^3 = 1 \Rightarrow 1^3 = 1 \times 4 \Rightarrow 2^3 = 8$

$1^4 = 4$ (?)

$2^3 = 8$

$2^2 = 4$ (base do cubo) $\longrightarrow \downarrow$

$2^3 = 8$

$$\left\{ \begin{array}{l} 1^3 = 1 \\ 2^2 = 4 \\ 2^3 = 8 \end{array} \right.$$

///

---

data

S T Q Q S S D

20220308

ARQ-BASICS → fazer down again.

—

art1lib.org

—

The Mathematical Universe
Pag. 28(47)
Pi in the sky.
(rodapé).

—

$$\left\{\begin{array}{l} 01\ \ 06\ \ 1969\ \ (Br) \\ 06\ \ 01\ \ 1969\ \ (USA) \end{array}\right.$$

mesmo valor (soma)

Einstein March 14, 1879

$\underbrace{03\ \ 14}_{Pi}$

obs:

British prefere

July 22/7 for Pi Day or

$\frac{22}{7} = 3.14$ to two decimal places-

As an aside that also happens to be the birthday of

German mathematician and physicist

Friedrich Wilhelm Bessel

(1784-1846),

whom we encounter in footnote 23 of Chapter 2.

The Bessel differential equation and subsequent solutions (Bessel functions) play an important role in many branches in physics.

20220309m

palíndromo
ambigrama

—

20220310

archive.org/details/@mirtitles

RASPUTITSA ⇒ tempo das estradas ruins. (para tanks)
⇓
o nome na Invasão Nazista da Rússia.

—

20220311

SMHD2022 - Partilhado (X)

—

Einstein
Minkowski } baseado nos estudos Lorentz and Poincaré

—

Recado 3 fases { 1) sugestão
2) deleite
3) aceitação

20220315m

New Directory:

Data-NovoVivo-006-2022

—

booksc.me

—

manuelsanciens.blogspot.com

—

Data-NovoVivo-005-2022

→ cópia da obra de Galileo Galilei

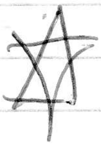

20220318m

ARQ-ALL--20220318.7z

→ Research material

—

20220319m

mgsuite.eu.org obs: era

— td.mgsuite.workers.dev

—

x2download.com

—

—

data

Ⓢ Ⓣ Ⓠ Ⓠ Ⓢ Ⓢ Ⓓ

20220324_{M}

Einstein saw symmetry hidden in the fabric of space and time.

20220326_{M}

libgen.is

libgen.rs / libgen.st

20220327_{M} COCHISE

https://drive.google.com/drive/u/1/folders/0AO-xX4BHvgjsUk9PVA

↳ zero ↳ ouro.

—

[illegible].

Ernest E. Shult

ernest_shult@yahoo.com

Johannes Ueberberg

johannes.ueberberg@gmx.de

Incidence Geometry

20220329 III Exclui extensão.

Programa → regedit (no campo de execução do Windows)

HKey_Current_User
↳ Software → Microsoft →
Windows →
CurrentVersion →
Explorer →
FileExts

HKey_Classes_Root →
↳ exclui a extensão.

—

Reinicie a máquina

Execute a extensão e depois
escolha o programa
que você quer usar para
executá-lo.

—

20220330 III

Pesach April 15

Chag Pessach Sameach

—

Legendre - Elementos de Geometria

https://bd.camara.leg.br/bd/handle/
bdcamara/22882

https://bdlb.bn.gov.br/acervo/handle/20.500.12156.3/442545

—

Math Books:
https://www.doraci.com.br/downloads/matematica/fund-mat-elem_01.pdf
11.pdf

=

20220331
spruce up = enfeitar.

backblaze.com

—

Trigonometric Functions and Complex Numbers
Desbenty Yang
ISBN 9781958134869
https://lccn.loc.gov/2016026528

=

→ practices makes perfect.

20220410m

Copie todos os arquivos para uma pasta só. (diretório)

Obs: 89,43TB

20220411m balanite

Milose → Cinework 25,00

20220411/ The Rise of Modern Logic.../ Philosophy of Physics - Part A. pdf

Pág. 229(254)

"geometrized"

→ rodapé

needed differential^3 — A review of the geometry...

geometrized version of Newtonian gravitations theory (also known as Newton-Cartan theory).

20220412m

Aniversário do meu pai
85 anos
Vida longa, muitas felicidades e
saúde !!!

→ mathemata

Books-07-2022/20220413/
Important/

A History of Pythagoreanism.pdf
Pág. 13 (21)

Pythagorean analysis of the similarities between numbers and things inspired by the advancement of the mathemata, which Aristotle attributes to them;

Pág. 105 (113) –

Plato ascribes to Theodorus' (~~discoveries were further incorporated~~) a proof of irrationality of the magnitudes between $\sqrt{3}$ to $\sqrt{17}$ →

Cansaço! trocando até link rsrs

(Tht. 147 d = DK 43 A 4)),
which means that
(significa que)
theodorus relied on the proof
of the irrationality of √2
that was found by Hippasus.[62]

62 - For reconstructions of
Theodorus's proof, see, e.g.,
Heath 1921: I: 202ff;
Knorr 1975: 109ff

Obs: Heath, T.L. 1921
A History of Greek Mathematics
→ Pág. 471 (479) do livro
A History of Pythagoreanism.

Heath - Vol I Pág 202

outro assunto!

this curve will be best discussed when we come to deal with the problem of squaring the circle (Chapter VII).

here we need only remark that it implies the proposition that the lengths of arcs in a circle are proportional to the angles subtended by them at the centre (Eucl. VI. 33)

—

trisecting any rectilineal angle ...

and these problems are not what the Greeks called "plane" problems, i.e., they cannot be solved by means of the ruler and compasses.

—

Hippasus — Pág. 65 (85)
Chapter III
Pythagorean Arithmetic

Hippasus → ... for making known the discovery of the irrational or incommensurable.

—

Pág. 160 (180) -

Hippasus

who was a Pythagorean but,
owing to his being the first
to publish and write down the
(construction of the) sphere
with (ἐκ, from)
the twelve pentagons, perished
by shipwreck for his (impiety),
but received credit for the
discovery...

Escanear amanhã!
Este caderno foi
escaneado até aqui
202204122228.

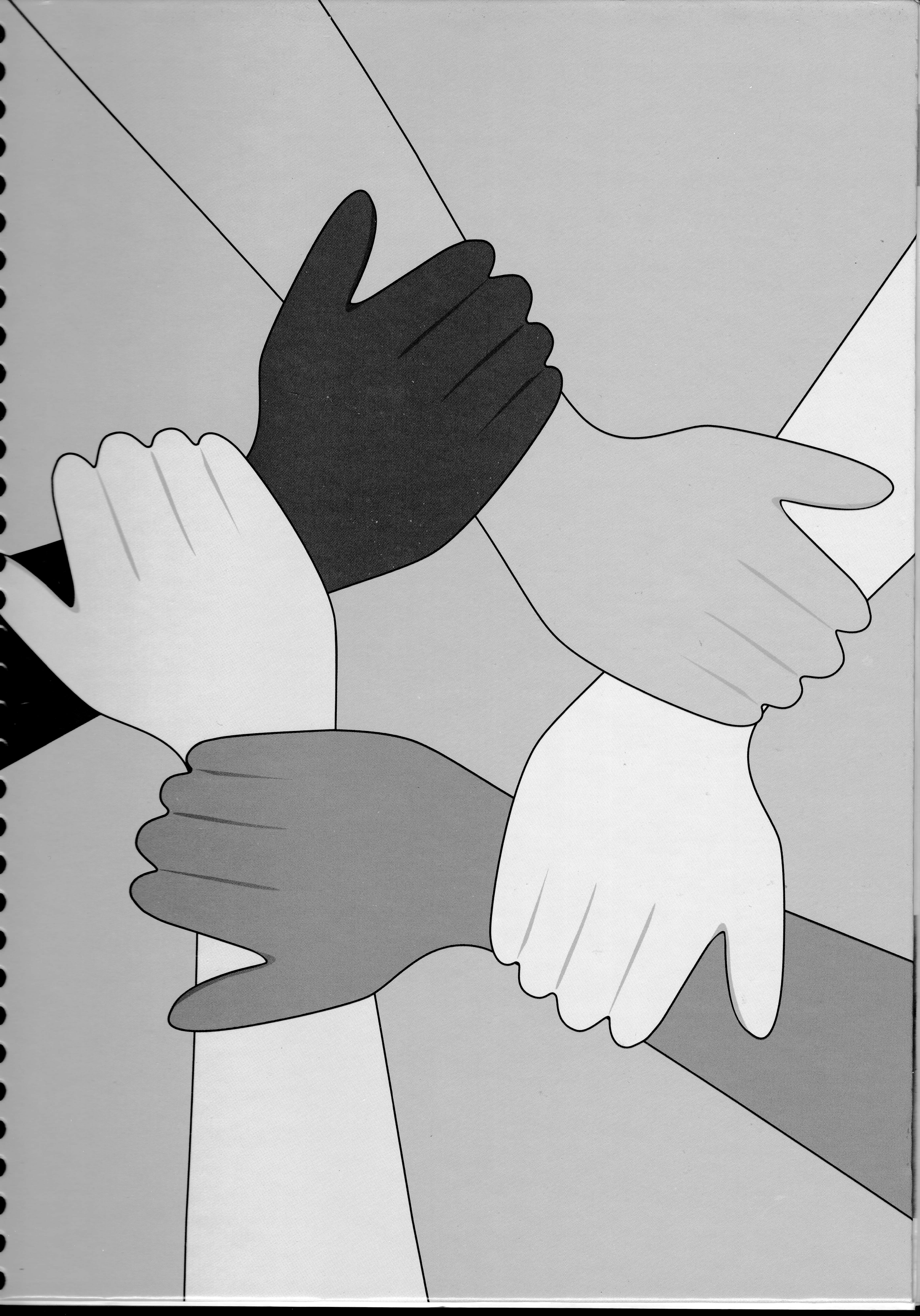

NOME . NOMBRE:

CURSO . CURSO: SALA . SALA:

MATÉRIA . MATERIA:

PROFESSOR . PROFESOR:

Horários	segunda lunes	terça martes	quarta miércoles	quinta jueves	sexta viernes	sábado sábado

Provas
Exámenes

Data
Fecha

Anotações
Anotaciones

DATE 20220615 R$ 13,00 1US$ = 5,12

= US$ = 2,54

20220615m

20220615 Stenio Musich

Criado hoje, 20220615, o diretório
Data-NovoVivo-013-2022/

1) Books-013-2022/20220615/
Mathematics and Physics for Programmers/
Mathematics and Physics for Programmers.pdf
Obs. 688 pages   Visão geral de Math and Physics.
Livro todo!!! Teoria e prático

Pág. 205(226) - A Stationary Circle and a
Moving Point.
Pág. 224(245) - A Stationary Ellipse and a
Moving Point.

20220616

2) Data-NovoVivo-013-2022/20220616/

1) A budget of paradoxes/

a) budgetofparadoxe01demorich/

Pág. 146(162) Two forgotten circle
squares.

Pág. 10(26) - 3 → finding the circumference
of a circle...

Pág. 12 (28) - 6 Application for the reward at Rome. (1778)

→ disputa sobre: "squared the circle".

Pág. 42 (58) - On Squaring the Circle.

Pág. 59 (75) - Leibnitz

Giordano Bruno and his paradoxes.

Pág. 69 (85) - Cataldi on the Quadrature.

Pietro Antonio Cataldi - 1612

This work claims a place as beginning with the quadrature of Pellegrino ~~Boggi~~ Borello4 of Reggio, who will have the circle to be exactly

$\boxed{3 \text{ diameters and } \frac{69}{484}}$ of a diameter.

4. He lived at "Reggio nella Emilia" in the 16th and 17th centuries. His "Regola e modo facilissimo di quadrare il cerchio" was published at Reggio in 1609. Ler todo o livro!!!

b) budget of paradoxe 02 demorich. pdf

Pág. 01 (09) - volume II - On some philosophical

Atheists → M. Laplace

"M. Laplace, they tell me, you have written this large book on the system of the universe, and have never mentioned its Creator"
→ checar!!!

Pág. 10(18) - The Trisection and Quadrature again etc. Ver livro todo!!!

c) budget of paradoxe 00 demorich.pdf

Pág. 62(76) → false quadrature of Duchesne, found that the ratio lay between

$$\boxed{\frac{333}{106} \text{ and } \frac{377}{120}}$$ checar!!!

$\frac{377}{120}$ → 3,1416

$\frac{333}{106}$ → 3,1415...

He then took the liberty of taking the mean of both numerators and denominators, given $\boxed{\frac{355}{113}}$ → 3,1415929...

Obs.: Ver livro todo!!!

2) Book - Accurate trisection and other impossible subdivisions of angles for practical people/ Obs.: Livro interessante!!!

3) Circles A Mathematical View Ed 2/
Obs.: Livro Bom!!! Circles A Mathematical View Ed2.pdf

4) Explorations in Topology Map Coloring, Surfaces and Knots//
Explorations in Topology....pdf

5) Geometric Dissections/

a) Frederickson - Ernest Irving Freese's Geometric Transformations.pdf
Obs.: Livro todo!!!
Pág. 59 - Triangles ① and ② are the only two isosceles triangles for each of which the pieces of a 6-PART dissection will also from either of two different pairs of equiareal isosceles triangles.
Pág. 61 - Dissection of any triangle to form another triangle of same base, b, and altitude, a, when point p falls on base b. 3 pieces.

Pág. 68 - I21d: Improved triangles for $1^2+2^2+3^2+4^2+5^2+6^2=(\sqrt{91})^2$
Obs.: Estudar!!!

Pág. 67 - Improved equilateral triangles for $1^2+3^2+5^2+7^2=(\sqrt{84})^2$.

b) Greg N. Frederickson - Dissections - Plane and Fancy...pdf

Obs.: Livro todo!!!

Pág. 172 (185) - Chapter 16 - Stardom

Pág. 173 (186) - 16.1 - Pentagram to decagon.

Pág. 174 (187) -

$$\frac{R(2n+2)}{s\{(2n+2)/n\}} = \sqrt{2}$$

c) Greg N. Frederickson - Hinged Dissections. Swinging and Twisting...pdf

Pág. 70 (84) - Turnabout 2: Bracing Regular Polygons.

$\sqrt{5}$: Underlying geometry $\sqrt{2}$ e $\sqrt{3}$.

Obs: Livro todo!!!

d) Greg N. Frederickson (Author) - Piano-Hinged Dissections (2006).pdf

Obs: Livro todo!!! Interessante!!!

Pág. 58 (74) - between $\sqrt{8}$ and $\sqrt{10}$ for which there will be one more crossing of lines in the cross position, producing a total of eleven pieces.'

d) d) cont.

Pág. 86(102) - Figure F1.2 illustrates the method on a sheet of metal that is 3 x 4.

In this case, $x = \frac{(7-\sqrt{13})}{6} \approx 0.5657$.

The Volume is $\frac{(35+13\sqrt{13})}{27} \approx 3.0323$.

e) Recreational problems in geometric dissections how to solve...(1972).pdf

Obs.: 212 pages.

6) Geometry and Symmetry Kinsey/

a) Kinsey - Symmetry, Shape and Space...pdf

514 pages / Obs.: Livro todo!!

20220617

b) Kinsey - Topology of Surfaces (1993).pdf

Pág. 184(190) - Chapter 9 - Homotopy

"The essence of mathematics is not to make simple things complicated, but to make complicated things simple"

Gudder.

9.1. Homotopy and homology.

The euler characteristics and the homology groups cannot tell the difference between the cylinder and the Möbius band.

We are not going to try to change that situation, but rather (em vez disso) investigat what they have in common that makes them algebraically indistinguishable.

Pág. 199(205) - Chapter 10 - Miscellany.

Pág. 201(207) - (10.3) Corollary ⇙

Trace a map or picture. Crumple the copy, being careful not to tear it, and place it so that (de modo a) the crumpled copy lies on the original (encontra-se no original)
There is some point on the copy which lies directly on top of the same point of the original.

Fig. 10.2 - Crumpling a map.

This is a composition of the two continuous functions: the crumpling and projecting straight downwards.
(para baixo)

Obs.: Proof: taking a point x on the original to the point immediately below the image of x on the crumpled copy.

⇒

Pág. 206 (212) - (10.11) - Corollary.

If S^2 is divided into three closed subsets, then at least one of these subsets contains a pair of antipodal points.

(10.12) - Corollary.

At any time, there are two points directly opposite each other on the earth with exacty the same temperature and barometric pressure (or any other two continuous functions).

(10.13) - ~~Corollary.~~ Ham Sandwich Theorem.

Let A, B, and C be compact subsets of R^3. There is a plane which divides each region exactly in half (by volume).

Pág. 207 (213) - Fig. 10.8.

The diameter D_x and the plane P_t.

Pág. 208 (214) - 10.2 The Jordan Curve Theorem

We will prove the theorem for polygonal curves only. A polygonal curve is a 1-complex made up of straight edges and vertices. The following proof is adapted from Moise's Geometric Topology in Dimensions 2 and 3.

Pág. 221 (227) - Chapter 11 -
Topology and Calculus.

"So far as the theories of mathematics are about reality, they are not certain; so far they are certain, they are not about reality".
Albert Einstein.

11.1 - Vector fields and differential equations in R^n.

7) Linear Algebra Done Right, 3rd Edition/
Linear algebra done right... pdf
Obj.: Livro todo!!! 352 pages!
→ Vector spaces.

Pág. 131 (147) - Chapter 5

Eigenvalues, Eigenvectors, and Invariant Subspaces.

→ Linear maps from a finite-dimensional vector space to itself.
Their study constitutes the most important part of linear algebra.

8) "Linear Algebra Illustrated/
Linear Algebra Illustrated.pdf
Obj.: 516 pages.

Pág. 157 - Chapter 3 - The 2-dimensional space.

Pág. 204 - 3.6 The Euclidean plane: distances

Pág. 207 - Theorem 3.6.7

Pythagorean theorem

$$a^2 + b^2 = c^2$$

obs: Há a prova no texto.

Pág. 209 - Theorem 3.6.12.

Distance Formula for Dimension 2

$P = (x, y)$ and $Q = (x', y')$ is

$$d(P,Q) = \sqrt{(x-x')^2 + (y-y')^2}$$

Pág. 233 - 3.10

The Complex Plane $\mathbb{C}$ is the Euclidean space R^2

$$z = a + bi.$$

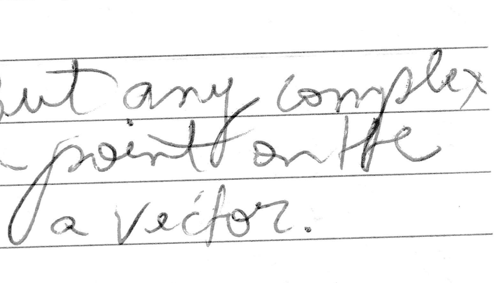

Therefore, a and b are thought of as the coordinates of z as a point of the plane. But any complex numbers is not only a point on the complex plane but also a vector.

$\mathbb{C} \Longleftrightarrow R^2$ given by

$$\boxed{a+bi \leftrightarrow \langle a,b\rangle}$$

The algebra of vectors in R^2 applies!

Warning:

In spite of this fundamental correspondence, we will continue to think of complex numbers as numbers (and use the lower case letters).

Obs: Pag. 234 → interessante !!!

Addition of complex numbers:

$$(a+bi)+(c+di)=(a+c)+(b+d)i$$

$$\langle a,b\rangle+\langle c,d\rangle=\langle a+c,\ b+d\rangle$$

This corresponds to addition of vectors:

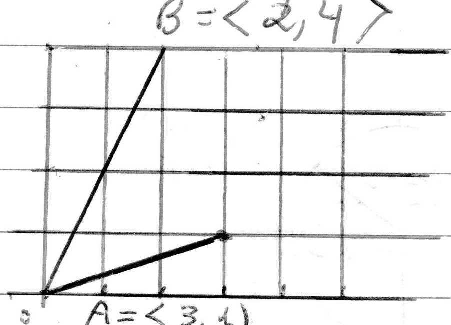

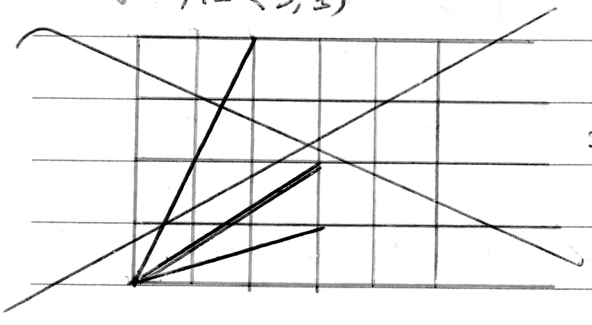

⇒

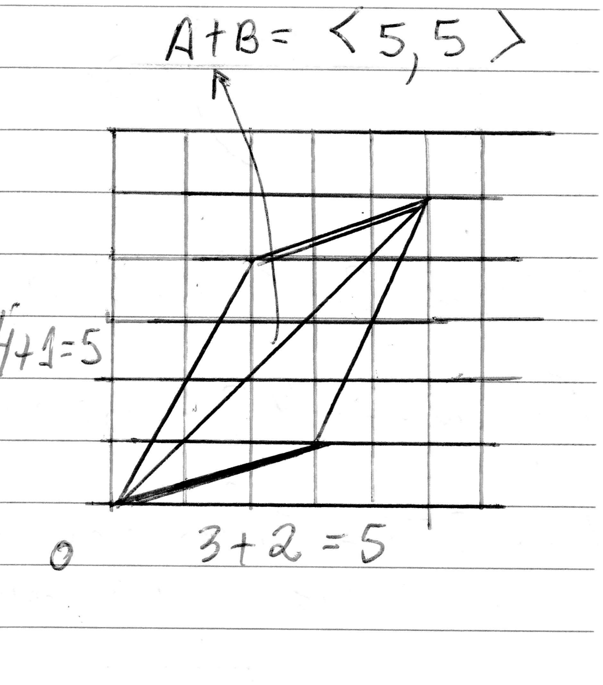

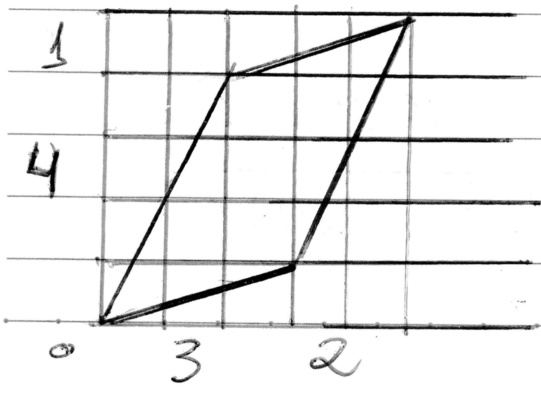

Second, we can easily multiply complex numbers by real ones:

$$(a+bi)\,c = (ac) + (bc)\,i,$$

$$\langle a, b \rangle\, c = \langle ac, bc \rangle.$$

It corresponds to scalar multiplication of vectors.

Warning!

Vector algebra of R^2 is complex algebra, but not vice versa. Complex multiplication is what makes it different.

Pág. 235 – Circle

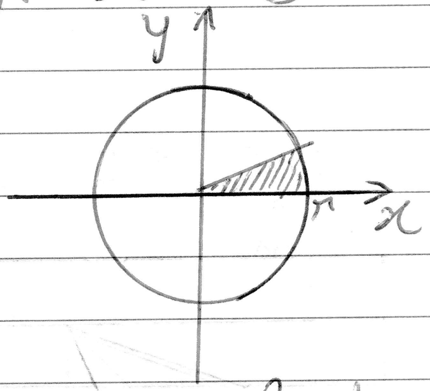

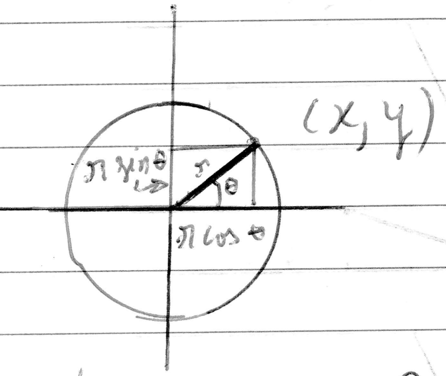

Pág. 239 – Chapter 4: Multidimensional spaces

Pág. 257 – 4.3 – Geometry of distances

Pág. 258 – Theorem 4.3.3

Distance Formula for Dimension 3

$$d(P,Q) = \sqrt{(x-x')^2 + (y-y')^2 + (z-z')^2}$$

9) The Advanced Geometry of Plane Curves ~~and~~ and their applications //

C. Zwikker - The Advanced Geometry of Plane Curves ... (2005).pdf

Obs.: 319 pages - Livro todo !!!

Pág. 33(44) - Chapter III - The Straight Line

1. Collinearity of three points, concurrency of three straight lines.

$\boxed{z = 1 + j u}$ for the straight line, passing through the point 1 in the direction of vector j.

Obs.: Livro todo !!!

10) The Painter's Secret Geometry /

a) Bouleau - The Painter's Secret Geometry.pdf
→ 268 pages.

b) Charles Bouleau - Charpentes, la géométrie secrète des peintres (1963).pdf
→ 269 pages.

11) IMPORTANT: estudar com cuidado !!!

→

11) The Trisection Problem/ ⇙

a) Accurate trisection and other impossible subdivisions of angles for practical people/
→ interessante!!!

⇘ b) Book - Accurate trisection and other impossible subdivisions of angles.../
Accurate trisection...pdf
Obs: Livro todo!!!

c) bibliographyonpo00goul.pdf
→ Bibliography on the Polemic Problem.
What is the value of π. - Gould. (1888)

d) cu31924063203040.pdf
→ A double discovery.
Square of the Circle.

e) Dudley - A Budget of Trisection (1987).pdf
→ 187 pages.

⇘ f) George Polya Mathematical discovery.pdf
→ 458 pages. Livro todo!!!

Pág. 33 (59) - Fig. 2.2

$$y = a - \sqrt{x^2 + y^2}$$

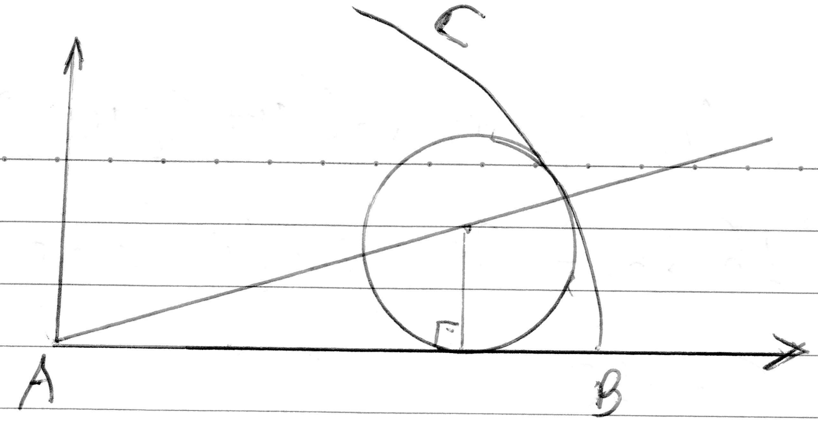

$$y = a - \sqrt{x^2+y^2} \rightarrow$$

$$x^2 = a^2 - 2ay.$$

And so the locus of the center of the variable circle turns out to be a parabola - a locus of no immediate use in geometric constructions.

Pág. IX (9) - A method of solution is perfect if we can foresee (prever) from the start, and even prove, that following that method we shall attain our aim.
Leibnitz: Opuscules, p. 161.

Obs.: Veja diretório todo!!!

3) Data-NovoVivo-013-2022/
Books-013-2022/
20220617/

a) A Glimpse at Hilbert Space Operators.../

a). A Glimpse at Hilbert Space Operators.pdf

Pág. 16(26) - On publishing

Find: GAUSS

(Gauss discovered exactly which regular polygons are ruler-and-compass constructible, and he proved, in particular, that the one with 65537 sides - a Fermat prime - is constructible. please do not publish the details of the procedure. It's been tried.)

b) Halmos - A Hilbert Space Problem Book...pdf

Pag. 03(22) - Chapter 1 - Vectors.

b) Geometry Illuminated An Illustrated..../
Geometry Illuminated - An Illustrated Introduction to Euclidean and Hyperbolic Plane Geometry...pdf

Pág. 421(438) - Chapter 38

The Pseudosphere. (Gauss)

Gauss curvature.

Pág. 425(442) - 38.2 - Maps between surfaces.
The Gauss map.

Obs.: Livro todo!!!

c) Geometry of Conics/

a) Geometry of Conics ... (2007).pdf

Pág. 99(104) - Chapter 4

Euclidean Properties of Curves of Second Degree.

→ Interessante.

b) Ibn al-Haytham's theory of Conics, Geometrical Constructions and practical geometry.pdf

→ 760 pages.

d) Methods of Modern Mathematical Physics I, II, III /

Methods of Modern Mathematical Physics I, II, III ... pdf

→ 1020 pages.

Obs: Volume I - mais interessante!!!

Hilbert Spaces
Banach Spaces
Topological Spaces
Locally Convex Spaces
Bounded operators.

e) Orthogonal Polynomials on the Unit Circle/

a) Orthogonal Polynomials on the Unit-Part1.pdf
→ 494 pages.

b) Orthogonal Polynomials on the Unit-Part2.pdf
→ 604 pages.

f) Paul-Halmos/

a) Halmos - Linear Algebra Problem Books - MAA (1996).pdf

b) Halmos (auth.) - Measure Theory (1950).pdf

Pag. 270 (280) §62 - Weil Topology.

g) Theories and Analyses of Beams and Axisymmetric Circular Plates/
Theories and Analyses of Beam....pdf
→ 578 pages.
Pag. 03 (28) - Jacob Bernoulli[3], and Leonhard Euler[4].

20220618N

4) Data-NovoVivo-013-2022/Books-013-2022/
20220618/

a) Advanced Engineering Mathematics/
Advanced Engineering Mathematics (1982).pdf
→ 1130 pages

Pág. 639 (657) - Chapter 11 -
Vector Spaces and Linear Transformations

Pág. 688 (706) - Chapter 12 -
Applications and Further Properties of Matrices.

Pág. 759 (777) - Chapter 13 - Vector Analysis.

Pág. 825 (843) - Chapter 14 - The Calculus of Variations.

Pág. 865 (883) - Chapter 15 - Analytic Functions of a Complex Variable.

Pág. 984 (1002) - Chapter 18 - Conformal Mapping.

Obs.: Pág. 632 (650) - Bessel Functions and Legendre Polynomials

$$x = \cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{2}$$

into the generating function, getting

$$[1 - z(e^{i\theta} + e^{-i\theta}) + z^2]^{-1/2} =$$

$$[(1 - ze^{i\theta})(1 - ze^{-i\theta})]^{-1/2} =$$

$$\sum_{n=1}^{\infty} P_n z^n$$

Pág. 984(1002) - Chapter 18
Conformal Mapping.
18.1 - The Geometrical Representation
of Functions of z.

$$u + iv = f(x+iy).$$

Pág. 985(1003) - Figure 18.1
The mapping of certain lines by the
function $w = z^2$.

b) 20220618/ Borsuk/

⇘ a) Borsuk - Foundations of Geometry -
Euclidean and Bolyai-Lobachevskian geometry...(2018).pdf
→ 465 pages.

b) Borsuk - The Theory of Retracts...pdf

· c) Elliptic Boundary Value Problems of
Second Order...pdf
Obs.: Dirichlet boundary.

⇘ d) Karol Borsuk - Multidimensional
analytic geometry...pdf
→ 442 pages.
Obs.: Livro todo!!!

e) Transmission Problems for Elliptic Second-Order Equations in Non-Smooth...pdf

c) 20220618/
Classical Geometry Euclidean, Transformational.../

a) Classical Geometry - Euclidean, Transformational, Inversive, and Projective...pdf
→ 494 pages. Livro todo!!!

b) Geometry Through History - Euclidean, Hyperbolic, and Projective Geometries.pdf
→ 356 pages.

c) Solutions Manual to Accompany Classical Geometry...pdf

d) 20220618/
Einstein/

a) Meaning of relativ00ein_0.pdf
→ 138 pages.
Obs.: The Meaning of Relativity - Albert Einstein. 1923

b) Papers00Eins.pdf
→ 226 pages.

c) principleofrelat00eins.pdf
→ 266 pages.
The Principle of Relativity -
Original papers by
A. Einstein and H. Minkowski (1920)

d) Relativity00EinsA.pdf
→ 22 pages.

e) 20220618/
Geometry/

a) axioms4geometry.pdf
b) euclid.pdf
→ 14 pages
c) foundations-of-euclidean-and-non-euclidean-geometry.pdf
→ 233 pages.
Obs.: Livro todo!!!

f) 20220618/
Geometry-Forder/

a) H.G. Forder - Geometry...(1960).pdf
→ 198 pages.
Obs.: Livro todo!!

Pag. 10(08) - Fig. 1

→ transformation.

Pág. 21(19) - Fig. 14

Six circles theorem:

If A', B, D, C be four points on a circle and circles be drawn through A', B; B, D; D, C; C, A'; and adjacent pairs cut again C', A', B', D', then these points are concyclid.

The figure is a "configuration", because it is composed of curves, namely circles, and points whereof (do que) the same number of (curves) curves (three) go through each point, and the same number of points (four) lie on each circle.

b) The Calculus of Extension...pdf
→ 504 pages. (1941)
Obs... Livro todo!!!

g) Love and Math The Heart of Hidden Reality/
Edward Frenkel - Love and Math.pdf
→ 300 pages.

h) Made simple The New Mathematics/

a) 1302.1630v19.pdf
→ 204 pages.

→ Euclidean plane and its relatives.

Obs.: Pág. 47 - Chapter 7
Parallel lines.

Obs: Livro todo!!!

b) feyman.pdf
→ "New" Mathematics.

⇘ c) further.pdf
→ Advanced High-School Mathematics.
→ 435 pages.

⇘ d) GCTprint.pdf
→ Geometry with an introduction to Cosmic Topology
→ 238 pages.
Obs: Livro todo!!!

⇘ e) George_Polya_Mathematical_discovery.pdf
→ 458 pages.
Obs. Livro todo!!!

Voltando:

d) GCTprint.pdf
Pág 155 (165) - Parallax.

The angle p is called the parallax, and in a Euclidean universe p determines the star's distance from the Sun, D, by the equation

$$D = d/\tan(p), \text{ where}$$

d is the Earth's distance from the Sun.

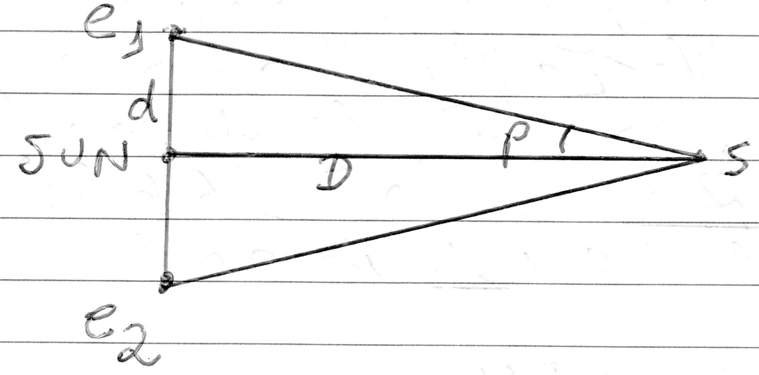

We may determine p by observation and d is the radius of Earth's orbit around the Sun
(d is about 8.3 light-minutes).
In practice, p is quite small, so a working formula is $\boxed{D = \frac{d}{p}}$.

The first accurate measurement of parallax was recorded in 1837 by Friedrich Bessel (1784-1846). He found the stellar parallax of 0.3 arc seconds
(1 arc second = 1/3600°)
for star 61 Cygni which put the star at about 10.5 light-years away.

Important:

If we live in a hyperbolic universe with curvature k, a detected parallax puts a 'bound (coloca um limite) on how curved the universe can be

Consider Fig. 7.3.5, p. 156 (166).

As before, e_1 and e_2 represent the position of the Earth at opposite points of its orbit, so that the distance between them is $2d$, or about 16.6 light-minutes.

Assume star s is on the

positive real axis

and we have detected a parallax p, so that angle $\angle e_2 s e_1 = 2p$.

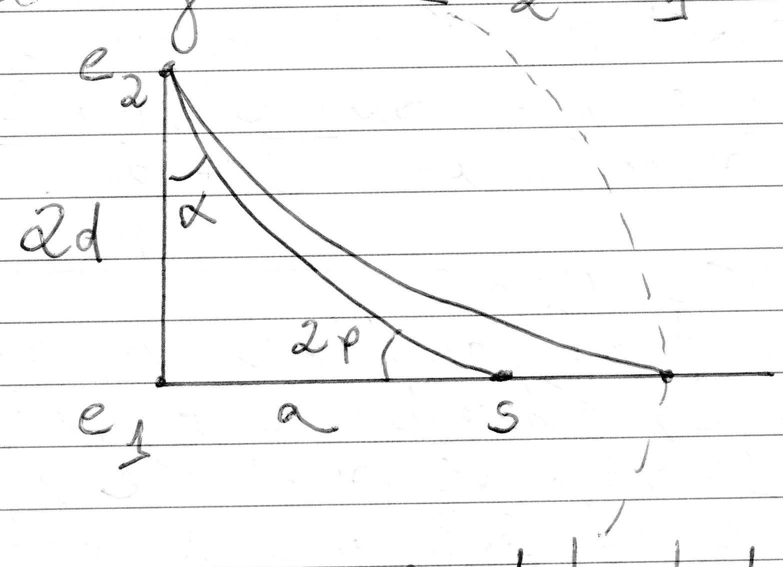

Figure 7.3.5 – A detected parallax in a hyperbolic universe puts a bound on its curvature.

The angle $\alpha = \angle e_1 e_2 s$ in figure 7.3.5, p.156(166) is less than the angle of parallelism

$$\theta = \angle e_1 e_2 u.$$

Nothing that $\tan(x)$ is an increasing function and applying Lobatchevsky's formula it follows that

$$\tan(\alpha/2) < \tan(\theta/2) =$$

$$e^{-\sqrt{|K|/2d}} \quad . \quad e^{-\sqrt{|K|}\,2d}$$

We may solve this inequality for $|K|$:

$$\tan(\alpha/2) < e^{-\sqrt{|K|}\,2d}$$

$$\ln(\tan(\alpha/2)) < -\sqrt{|K|}\,2d$$

$\ln(x)$ is increasing

$$\left[\frac{\ln(\tan(\alpha/2))}{2d}\right]^2 > |K|.$$

x^2 is decreasing for $x < 0$.

To get a bound for $|K|$ in terms of ρ, noted that $\alpha \approx \pi/2 - 2\rho$ (the triangles used in stellar parallax have no detectable angular deviation from 180°), so

$$|K| < \left[\frac{\ln(\tan(\pi/4-p))}{2d}\right]^2.$$

We remark that for values of p near 0, the repression

$\ln(\tan(\pi/4-p))$ has

linear approximation equal to $-2p$, so a writing for K, which appeared in Schwarzschild's 1900 paper

[27], is $|K| < (p/d)^2$.

↓

Schwarzschild, K.
On the Permissible Curvature of Space

Vierteljahrsschrift d. Astronom. Gesellschaft, (1900), 35, 337-47.

Source: https://pt.booksc.eu/book/45450365/a07ce1
Copy: 20220618/Made Simple the New Math/
0264-9381_15_9_003.pdf
→ 7 pages.

Pág. 162(172) - e - Pág. 161(171)

Pág. 161(171) - Theorem 7.4.7

Unified Pythagorean Theorem.

Suppose $K \in \mathbb{R}$ and we have a geodesic right triangle in (X_K, G_K) whose legs have length a and b and whose hypotenuse has length c. Then

$$A(c) = A(a) + A(b) - \frac{K}{2\pi} A(a)A(b),$$

where
$A(r)$ denotes the area of a circle with radius r as measured in

(X_K, G_K).

Obs: Estudar esta parte com profundidade.
De antemão, digo,
mudanças terão que ser feitos pois
π será alterado para M.U.S. (π)
→ Mathematical Uniform
Space.

Pág. 162(172) - Example 7.4.8

Detecting curvature with triangles.

Obs. Interessantíssimo!!!

Estudar!!! adapte-se ao MUSC(!!!)

Pág. 164(174) - 7.5 Surfaces

Pág. 165(175) - Example 7.5.2

A circle is homeomorphic to a square.

Pág. 179(189) - 7.6

Geometry of Surfaces

We have seen that any surface is homeomorphic to a polygonal surface representing H_g or C_g, for some g, and we now show that each of these may be given a homogeneous, isotropic and metric geometry (so that it has constant curvature).

A polygonal surface can only run into trouble homogeneity-wise where the corners come together.

Pag. 191 (201) - Figure 7.7.12

A triangle in H_2, with various images of it.

IMPORTANT:

Estudar este livro associado com os seguintes livros:

1) Theory of Shape

Karol Borsuk - 1975

Obs.: 388 pages.

Source: https//archive.org/details/theoryofshape0000bors/

2) 20220618/
Made Simple The New Mathematics/
GCTprint-Books-complementos-importan/

a) A.R. Pears - Dimension theory of general spaces... djvu

b) Alan R. Pears - Dimension theory
theory of General Spaces...pdf
→ 442 pages.

Chapter 1 - Pag 01(14)

Topological spaces, normality
and compactness.

Obs: olhar geral!!!

c) Number, Shape, & Symmetry...pdf
→ 434 pages.
Obs.: Livro todo!!!

i) 20220618/
New Foundations in Mathematics
The Geometric Concept of
Number/

a) Clifford Algebra to Geometric
Calculus.o.pdf
→ 328 pages.

Pag. 203(217) - Chapter 5

Differential Geometry of Vector Manifolds

5.4 - Parallelism and Projective Related
Geometries.

b) Garret Sobczyk - New foundations in mathematics (2013). pdf
→ 373 pages.
Obs: Livro todo!!!

- Modular Number Systems.
- Complex and Hyperbolic Numbers.
- Geometric algebra.
- Vector Spaces and Matrices.
- Linear Transformations on $\mathbb{R}^n$
  Pág. 107 (118).
- Geometry of Moving Planes.
  Pág. 181 (190).
- Representation of the Symmetric Group
  Pág. 201 (209).
- Mappings between Surfaces
  Pág. 275 (281).
- Non-euclidean and Projective Geometries
  Pág. 297 (302).

Pág. 107 (118) - "Life is good for only two things, discovering mathematics and teaching mathematics."
Simón Poisson.

Pág. 297 (302) – Chapter 17

Non-euclidean and Projective Geometries

"The secret of science is to ask the right questions and it is the choice of problem than anything else (qualquer outra) that mark the man of genius in the scientific world."

Sir. Henry Tizard.

17.1 The Affine n-Plane A^n_h

Pág. 298 (303) – Fig. 17.1

⇒ unique line $X_h \wedge Y_h$

Obs.: Estudar esta "projeção" sobre uma superfície plana
(curva para plana)
⇒ M.U.S. (TII)

$(\overset{\frown}{III})$

⇒ interessantíssimo!!!

Obs.: Estudar com carinho!!!

j) 20220618/ Veblen/

a) C.R. Wylie .. Foundations of Geometry.pdf
→ 360 pages.
Obs.: Livro todo!!!

b) cu31924059316103.pdf
→ Introduction to Infinitesimal Analysis - Functions of one real variable.
Veblen (1907)
→ 268 pages.

Obs: Number e and π.

- Transcendence of e - Pág. 19 (35).
- Transcendence of π - Pág. 25 (41).

Pág. 25(41) §9 -

The Transcendence of π.

The definition of the number π is derived from Euler's formula

$$e^{x\sqrt{-1}} = \cos x + \sqrt{-1} \sin x;$$

by replacing x by π,

$$e^{\pi\sqrt{-1}} = -1 \qquad (1).$$

If π is assumed to be an algebraic number,

$\pi\sqrt{-1}$ is also an algebraic number and is the root of an irreducible (irreducible) algebraic equation $F(x)=0$ whose coefficients are integers.

(... continua)

c) Foundations of Differential Geometry

Kobayashi - 2 Volumes

Volume 1 - 340 pages.
Volume 2 - 485 pages.

Obs.: .djvu (1996).

d) Louis Auslander -

Differential Geometry (1967).djvu

Obs.: Converti também em .pdf.
→ 281 pages.
Livro todo!!!
Pag. 145 (155) - Chapter V. Local Theory of Surfaces
Pag. 199 (209) - Chapter VI. Global Study of Surfaces.
Pag. 237 (247) - Chapter VII. Integration of Forms and the Gauss-Bonnet Theorem.

K) 20220618/ 20220619$_{M}$
Borsuk-02/

a) 0088_A_Generative_theory_of_shape.pdf
→ 582 pages.
obs.: theory of understandability (?). → não aplicável.

b) 1969394.pdf
On the topology of Retracts – Borsuk
→ 14 pages

c) 1479287191.pdf
Borsuk's Shape Theory
→ 6 pages.

ANR, → Absolute Neighborhood Retracts.

are topological spaces X which, whenever $i: X \to Y$ is an embedding into a normal topological space Y, there exists a neighborhood U of $i(X)$ in Y and a retraction U onto $i(X)$.

They were invented by Borsuk in 1932 (Über eine Klasse von

lokal zusammenhängenden Räumen,
Fundamenta Mathematicae 19(1),
p. 220-242, EuDML)

Obs: EuDML →
https://eudml.org/doc/212574

Source: http://matwbn.icm.edu.pl/
ksiazki/fm/fm19/
fm19119.pdf

↘ d) Alan R. Pears - Dimension Theory of
General Spaces.pdf
→ 442 pages.

e) borsuk_1.pdf
A Borsuk Theorem on Homotopy
Types.

↘ f) fm19119.pdf

⇘ g) Number, Shape, & Symmetry.pdf

h) S0002-9904-1976-14164-X.pdf
Geometric Topology and
Shape Theory:
A Survey of Problems and
Results
Author: B. J. BALL
→ 14 pages.

i) s000497270000647x.pdf
What is the Theory of Shape?

Karol Borsuk and Jerzy Dydak.
→ 38 pages

j) Shape Classification and Analysis...pdf

Obs.: Interessantíssimo!!! Livro todo!!!
Bem didático.
→ 674 pages.

Aproveitar o essencial!!!

Pág. 63(91) - 2.2.4

Metric Spaces, Inner Products, and Orthogonality.
(tudo!!!)

Pág. 67(95) - Figure 2.35

The projection L of vector $\vec{p}$ over vector $\vec{q}$ can be expressed in terms of a scalar product.

Obs.: Matrizes são utilizadas
↳ "perfeito" para cálculo de projeção ortogonal sobre um plano
(usar!!!)

Pág. 73 (101) - Figure 2.37 (Estudar!!!)

The orthogonal transformation defined A implements a rotation of the canonical coordinate system (a) in the y-z plane by and angle of θ, with respect to the y-axis.
The fact det(A) = 1 indicates that the senses of all axes are preserved.
This is not verified for the transformation defined by matrix B(c).

Pág. 96 (124) - Figure 2.50

Osculating circle and radius of curvature at a point t.

Cont... Ver livro todo!!!

⇓ j) The Shape of inner space string theory and the geometry of the universe...pdf
→ 334 pages.
Obs.: Livro todo!!!

⇓ k) Theory of the Earth's Shape (1982).pdf
⇒ 678 pages.

K) Theory of the Earth's Shape
(1982).pdf
→ 678 pages! Livro todo!!!

Pág. 57 (45) Fig. 2.8
Dirichlet's Exterior Problem ~~ff~~
for the Sphere.

Obs: Aplicar a curvatura do espaço.
(deformation).
↳ projection (orthogonal).

Pág. 65 (53) - 3.1.1
The Curvature of the Level Surfaces.

(Fig. 3.1) ↙

Pág. 83 (71) - 3.4 Coordinate Systems

Pág. 85 (73) - Fig. 3.2 ↙

Obs.: Adequar estas coordenadas
com a nova geometria
⚝ propiciada pelo M.V.S. (TTT).

Obs.: Livro todo!!!
Estudar junto com o diretório 20220619/
Theory of Satellite Geodesy.../

L) Real Homotopy of Configuration Spaces Piccot Lecture/

Real Homotopy of Configuration Spaces.pdf
→ 201 pages.

5) 20220619/

a) Completeness Theorems and Caracteristic Matrix Functions/

Kaashoek... Completeness...pdf
→ 358 pages.

Pág. 19 (34) - Chapter 2
Completeness Theorems for Compact Hilbert Space.

b) Coordinate Systems and Map Projections/

a) D.H. Maling - Coordinate Systems and Map Projections.pdf
→ 490 pages. Livro todo!!!

b) D.H. Maling - Measurements from Maps...pdf
→ 595 pages. Livro todo!!!

c) Feynmann Integrals.../

Weinzierl2022 - Feynmann Integrals.pdf
→ 852 pages.

d) Representation Theory and Algebraic Geometry/

2022-Book_RepresentationTheoryandAlgebra.pdf
→ 458 pages.

⇘ e) Shape Classification and Analysis Theory and Practice/

↘ a) Shape and shape theory (1999).pdf
→ 309 pages. Livro todo!!!

Pág. 01(13) - Chapter 1

Shape and Shape Spaces.

K labelled points in a Euclidean space $\mathbb{R}^m$, where $K \geq 2$.

shape space and will be denoted by

$$\Sigma^K_m.$$

Pág. 26(38) - Figure 2.1

Cells of the upper hemisphere in $\mathbb{R}^3$.

Obs.: Livro todo!!!

b) Shape Classification and Analysis.pdf
→ 674 pages. Livro todo!!!

c) the shape of inner space...pdf

f) 20220619/
Spacecraft Operations.../

Spacecraft Operations...pdf
→ 610 pages. Livro todo!!!

g) 20220619/
String Theory.../

String_Theory_For_Dummies.pdf
→ 403 pages.

⇓ h) 20220619/
Theory of Satellite Geodesy.../

a) D.H. Maling - Coordinate Systems and Map.pdf
b) D.H. Maling - Measurements from Maps.pdf
c) From Mars to Greenland ... Gravity.pdf
→ 352 pages
d) Physics of the Sun - Volume III.pdf (293 pag)
e) The Interdisciplinary Role of Space Geodesy.pdf (310 pág)
Pag. 222(235) - 3. Conventional Terrestrial Systems (CTS)
f) Theory of the Earth's Shape (1982).pdf (678 pages)
g) William... Theory of Satellite Geodesy.pdf (143 pages)

20220621

"A ignorância totalmente consciente é um prelúdio para qualquer avanço real no conhecimento"
James Clerk Maxwell.

Gauss-K - ESTUDO!

1) Data-NovoVivo-013-2022/
Books-013-2022/
Importantíssimo-02/

GCTprint.pdf

Pág. 145(155) - Chapter 7
Geometry on Surfaces

2) Shapes and geometries. Metrics...pdf

Pág. 74(98)

$$K \overset{def}{=} \prod_{\alpha=1}^{N-1} \kappa_\alpha .$$

↙

The mean curvature H and the GAUSS CURVATURE K are defined.

Equations:

$$H \overset{\text{def}}{=} \frac{1}{N-1} \sum_{\alpha=1}^{N-1} K_\alpha \quad \text{and}$$

$$K \overset{\text{def}}{=} \prod_{\alpha=1}^{N-1} K_\alpha .$$

The choice of the inner normal for a_N is necessary to make the principal curvatures of the sphere (boundary of the ball) positive.

3) Theory-of-Shape-Borsuk-pagina-dupla---20220821-pdf
→ 380 pages.

Pag. 315 (317) –

$$f \varphi_{1k} / U_1 \simeq g \varphi_{1k} / U_1 \text{ in } V$$

for every $k \geq k_1$, and.

Outro Assunto:
Data-NovoVivo-013-2022/ Books-013-2022/ 20220621/ Euclid's Elements Redux/ euclid-04.pdf
Pag. 219 →

Pag. 219 - Proposition 4.5

Circumscribing a circle about a triangle.

Pag. 220 - Exercises ②

Figure 4.2.7: [4,5, #2].

IMPORTANT : 3 pontos alinhados em cada lado de um triângulo

obs: Pappus Theorem (interessante).

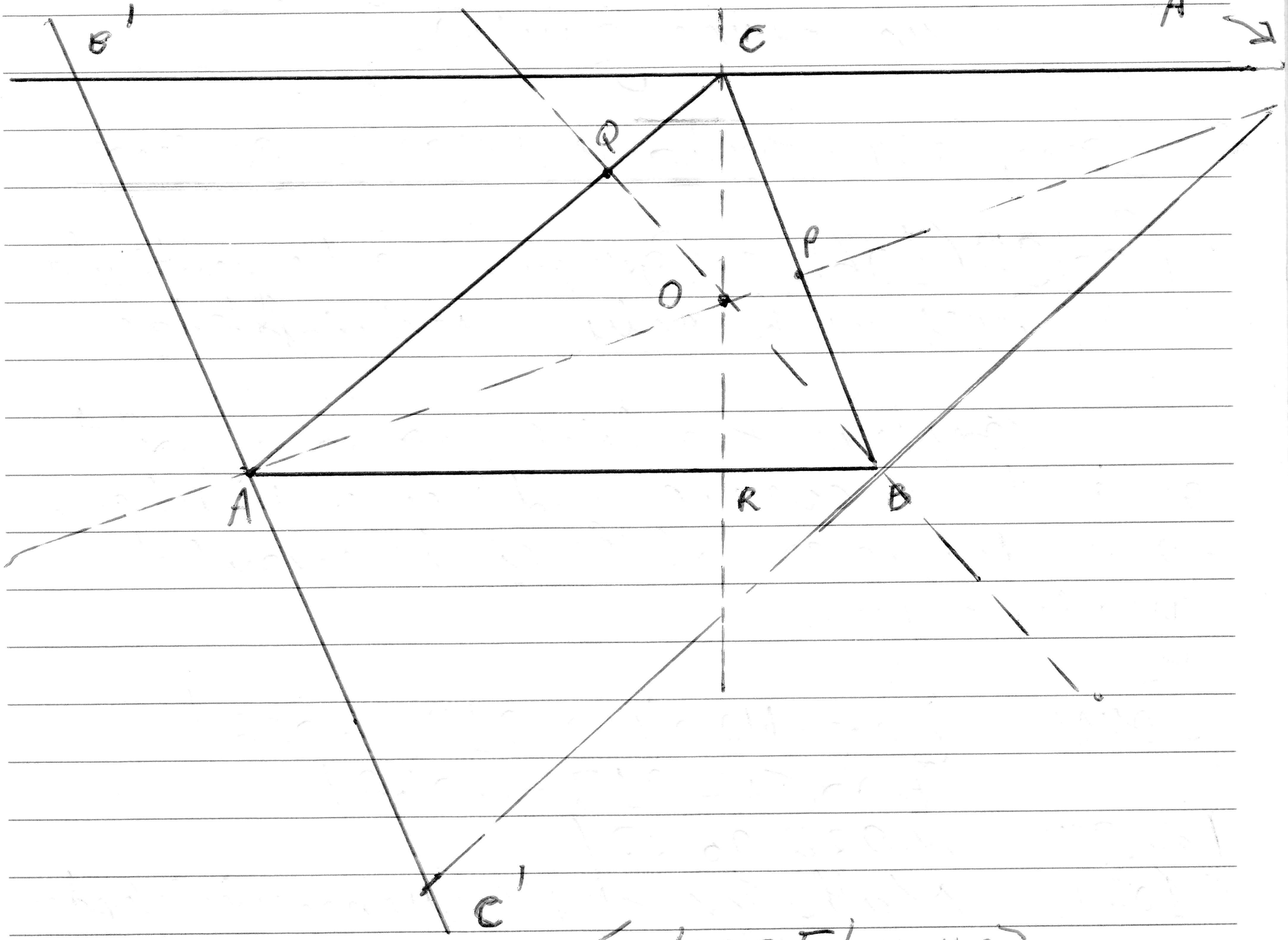

Fig 4.2.7 [4.5, #2]

20220622///

1) Data-NovoVivo-013-2022/
Books-013-2022/
Importantíssimo-02/

GCT print.pdf

Pág. 145(155) - Chapter 7

Geometry and Surfaces

Obs.: Análise detalhada sobre K (curvature gauss).

2) Complementando a imagem anterior:

Pág. 97(106) - Constructing Perpendicular Lines with only a straightedge.

Construct a line from a point outside a circle and perpendicular to a diameter of the circle would qualify you.

Source: Data-NovoVivo-013-2022/
Books-013-2022/
Pág. 98 20220622/
(107). Math Tricks the Surprising Wonders of Shapes and Numbers!../

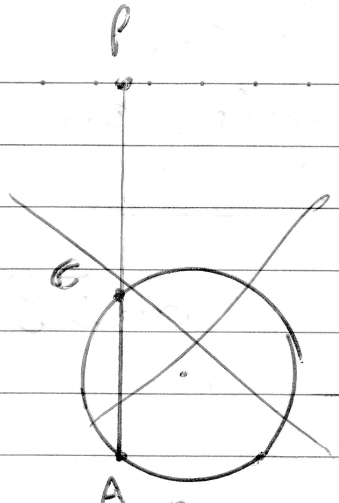

Obs.: Na imagem anterior, observamos 3 perpendiculares, cada uma sobre um lado do triângulo formando um ângulo reto. O mesmo ocorre nesta imagem abaixo.

obs:
Cada reta tem 3 pontos ~~ate~~ alinhados
Pappus's Theorem (interessante!!!)
Deformation e mantem-se a reta.

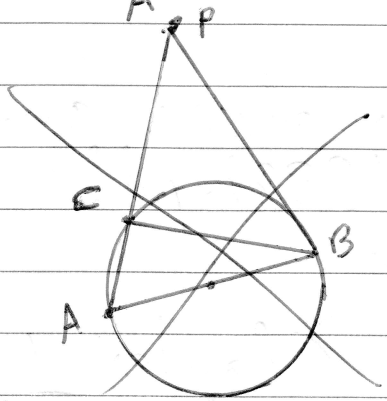

Obs:
Nesta imagem, usamos um ponto fora do círculo, o ponto P e deste saem 3 retas, criando um triângulo.

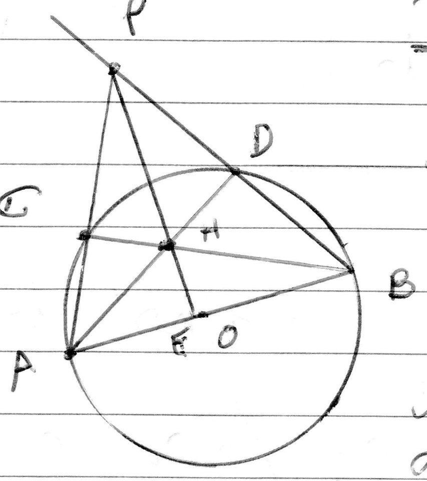

Interessante.
O meu desenho é ligeiramente diferente da imagem do livro.
Na minha imagem o ponto E está quase sobre o ponto O (centro) da circunferência.

A imagem do livro, o ponto E está mais à esquerda

Obs.: Estudar livro todo! Interessante!!

④ Voltar a estudar o arquivo GCTprint.pdf

3) Pág. 111 (120) - The Trick to prove that a triangle can have two right angles.

Figure 2.52 e 2.53.

---

4) Books-013-2022/
20220622/

Goldberg - Curvature and Homology (1998).pdf
→ 413 pages.

Pág. 01 (19) - Chapter 1
Riemannian Manifolds

Obs.: Livro todo!!!

20220623//

Baixei os trabalhos dos matemáticos:
(Werke - German)
Books-013-2022/20220623/
a) Werke - Dirichlet;
b) Werke - Gauss;
c) Werke - Gauss-02;
d) Werke - Gauss-03;
e) Werke - Jacobi
f) Werke - Kronecker e g) Werke - Riemann

IMPORTANT:
Books-013-2022/
20220623/
Celestial-Mechanics/

1) Applied General Relativity - Theory and Applications in Astronomy....pdf
→ 548 pages
Obs.: Livro todo!!!

Pág. 26(45) - 2.7 Curvature and Ricci Tensor.

2) Relativistic celestial mechanics of the Solar System ... (2011).pdf
→ 892 pages
Obs.: Livro todo!!!

Obs.: Neste diretório tem a pasta:

Hagihara/

→ 5 volumes - Celestial Mechanics (completo)

Cópias destes arquivos estão também no diretório:

Books-013-2022/Importantíssimo-02/

Cont.:

Applied General Relativity...pdf

Pág. 120(137) - 4.3

The Minkowskian Metric,
Lorentz-Transformation

Pág. 121(138) -

$$K^{\mu} = (-k_0, k_i) \qquad (4.3.6)$$

Fig. 4.1 Surfaces of constant phase, light rays and wave-vectors.

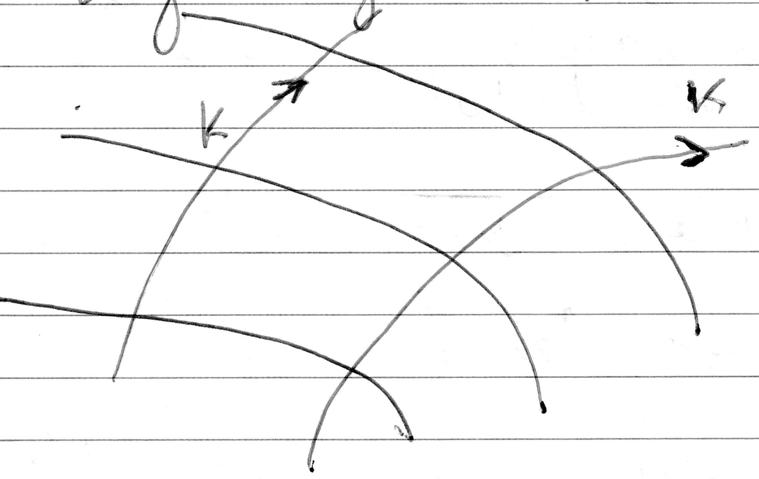

Pág. 122(139) - Metric Property 1

Light-rays are curves of zero length (null-curves). In Minkowskian coordinates this simply means that

$$0 = ds^2 = -c^2 dt^2 + dx^2,$$

i.e. the speed of light c takes a constant value in every Minkowskian coordinate system.

Pág. 123 (140) – to the Pythagorean theorem we now get

$$L^2 + \left(\frac{v\,\Delta t}{2}\right)^2 = \left(\frac{c\,\Delta t}{2}\right)^2$$

from which we are

$$\Delta t = \frac{\Delta \tau}{\sqrt{1 - v^2/c^2}} \qquad (4.3.9)$$

In words: a moving clock appears to go slower.

20220624 III

Data-NovoVivo-013-2022/
Books-013-2022/
Importantíssimo-02/
Methods of Modern Mathematical Physics I, II, III/

1) Methods of Modern Mathematical Physics...pdf

Pág. 03 (17) – 1.2

Metric and normed linear spaces

We will be dealing with sets of functions or operators or other objects and we will often a way of measuring the distance between the objects in the sets. It is reasonable to define a notion of distance that has the most important properties of ordinary distance in $\mathbb{R}^3$.

Definition - A metric space is a set M and a real-valued function $d(\cdot,\cdot)$ on $M \times M$ which satisfies:

(i) $d(x, y) \geq 0$
(ii) $d(x, y) = 0$ if and only iff $x = y$
(iii) $d(x, y) = d(y, x)$
(iv) $d(x, z) \leq d(x, y) + d(y, z)$
[triangle inequality)

Obs.:
Triangle inequality, in Euclidean geometry, theorem that the sum of any two
(trocas de careta) ⇐
sides of a triangle is greater than or equal to the third side, in symbols,

$a + b \geq c$. In essence, the theorem states that the shortest distance between two points is a straight line.

The function d is called a metric on M.

Distance:

$d(x,y) = \sqrt{(x_1-y_1)^2 + \ldots + (x_n-y_n)^2}$

Example:

Let M be the unit circle in $\mathbb{R}^2$, that is, the set of all pairs of real numbers $\langle\alpha,\beta\rangle$ with $\alpha^2+\beta^2=1$, and let

$d_1(\langle\alpha,\beta\rangle,\langle\alpha',\beta'\rangle) = \sqrt{(\alpha-\alpha')^2+(\beta-\beta')^2}$

Another possible metric is $d_2[p,p'] =$ arc length between the points p, p'

Pag. 04(18) - Figure I.1 The metric d_1 and d_2.

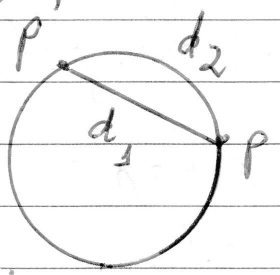

... continuar lendo!!!

2) 20220624/
An Introduction to Groups and Their Matrices for Science Students/
Robert Kolenkow - An Introduction to Groups and Their Matrices ... pdf

Pág. 168 (183) - Chapter 7 ⇙

ROTATION. → interessante!!!

Pág. 182 (197) - 7.7 Euler Angles.

3) Books-013-2022/
Importantissimo-02/
Mathematics and Physics for Programmers/

Mathematics and Physics for Programmers...pdf

Pág. 297 (318) - Chapter 12

Force and Newton's Laws.

4) Books-013-2022/
Importantissimo/
Solomon - Linear algebra geometry and
transformation.pdf

Pág. 301 (317) - Chapter 6

Orthogonality

Pág. 326 (342) - 3.15 - Orthogonal matrices.

5) Books-013-2022/
Basic Algebraic Geometry 1.../

→

Linear Algebra and Geometry - Springer (2012).pdf

Pág. 223 (239) - 7.2

Orthogonal Transformations.

Let L_1 and L_2 be Euclidean spaces of the same dimension with inner products $(x, y)_1$ and $(x, y)_2$ defined on them.

We shall denote the length of a vector x in the spaces L_1 and L_2 by

$|x|_1$ and $|x|_2$, respectively.

Definition 7.21 - An isomorphism of Euclidean spaces L_1 and L_2 is an isomorphism

$A: L_1 - L_2$ of the underlying (subjacente) vector spaces that preserves the inner product,

that is, for arbitrary vectors $x, y \in L_1$, the following relationship holds:

$$(x, y)_1 = (A(x), A(y))_2 \qquad (7.16)$$

Obs: Estudar!!! → the isomorphism A preserves the lengths of vectors.

6) Books-013-2022/
20220624/
Differential Calculus and Holomorphy
Real and Complex Analysis.../

Differential Calculus and Holomorphy.pdf

Pag. 77(90) - Chapter 2

Holomorphic Mappings,
Basic Properties.

Abstract:
In this chapter we study differentiability in the particular case of complex spaces, i.e., Holomorphy.
Except in §2.6 where series of homogenous polynomials and real analytic mappings are studied, all spaces considered here are complex.

7) Books-013-2022/Importantíssimo/
Made Simple the New Mathematics/
further. pdf
⇒ 435 pages.
Pag. 01(11) - Chapter 1
Advanced Euclidean Geometry.
→ interessante!!!

8) 20220624/ Physics of Surfaces and Interfaces/
Harald .. Physics of Surfaces and Interfaces ... pdf

Pág. 07(18) - 1.1.2 - Surface Superlattices.

Notation:

The positions of surfaces atoms differ from the bulk because of the broken symmetry and the broken bonds.

Fig. 1.5 - Illustration of the notation of the c(2x2) unit cell of the surface lattice and its alternative notation as

$\sqrt{2} \times \sqrt{2}\ R45^\circ$.

9) 20220624/ Introduction to Supersymmetry/

a) Introduction to Supersymmetry ... pdf
→ 449 pages.

Pág. 07(17) - Chapter 1

Lorentz and Poincaré Group, SL(2, $\mathbb{C}$), Dirac and Majorana Spinors.

1.1. The Lorentz Group

A point in the spacetime manifold[1] is

denoted by

$(x^\mu) = (x^0, x^1, x^2, x^3)$, where

$x^0 = t$, and x^1, x^2, x^3 are the space components of the four-vector x^μ.

This space is called Minkowski space and denoted by M4.

The laws of physics formulated in M4 are invariant under the Lorentz group.

TRANSFORMATIONS of this group are linear transformations acting on four-vectors[2].

Obs: spacetime manifold[1]

[75] H.J.W. Müller-Kirsten, Classical Mechanics and Relativity ... 2008.

Chapter 12

or https://archive.org/details/classicalmechani0000mull

[98] L.H. Ryder - Quantum Field Theory (1985).

b) Supersymmetry - An Introduction with conceptual and Calculational Details.pdf:
→ 602 pages.

Livros complementares:

c) Geometry - A Metric Approach with Models.pdf.
→ 388 pages.

Obs.: Chapter 7 - The Theory of Parallels.

Pág. 169 (184)

Pág. 224 (239) - Chapter 9

Euclidean Geometry.

9.1 - Equivalent Forms of EPP

EPP - Euclidean Parallel Property.

Pág. 226 (241) - Theorem 9.1.6

In any hyperbolic geometry, there is a triangle such that (de tal modo que) the perpendicular bisectors of two of the sides are parallel.

Veja →

Pág 227 (242) - Figure 9-2

Obs.: Livro todo!!!

d) The Geometry of Metric and
Linear Spaces... pdf
→ 252 pages.

Pág. 206 (215) -

Metric Dependence and a sum of
distance

others books:

10) 20220624/ Introduction to Supersymmetry/

14-Book Manuscript... pdf
→ 195 pages.

Mechanics and Relativity.

Pág. 51 (64) Chapter 5 -

Rotation Motion, torque and angular momentum

11) Introduction to Supersymmetry/
Motion and Relativity.... pdf
→ 225 pages.

Dentro deste diretório Müller-Kirsten (pasta)

20220625 N

1) Books-013-2025/
Advanced Mechanics and General Relativity/
Franklin- Advanced Mechanics and General Relativity (2010).pdf

Pág. 127(147)- Chapter 4- Curved Space.

Pág. 128(148)- We can introduce a coordinate system in which space-time looks like Minkowski space-time.

→ where space-time is described via the MINKOWSKI METRIC.

In such space-times, the metric defines a connection, the connection defines the RIEMANNIAN CURVATURE and the Riemannian curvature is, by EINSTEIN'S EQUATIONS, related to sources.

So for Einstein's theory, the metric is enough?

2) 20220625/ Classical Mechanics from Newton to Einstein/

2.1) Classical Mechanics- From Newton to

Einstein ... (2010).pdf
→ 241 pages.

2.2) Gravitation - From Newton to Einstein.pdf
→ 128 pages.

Breve observações:

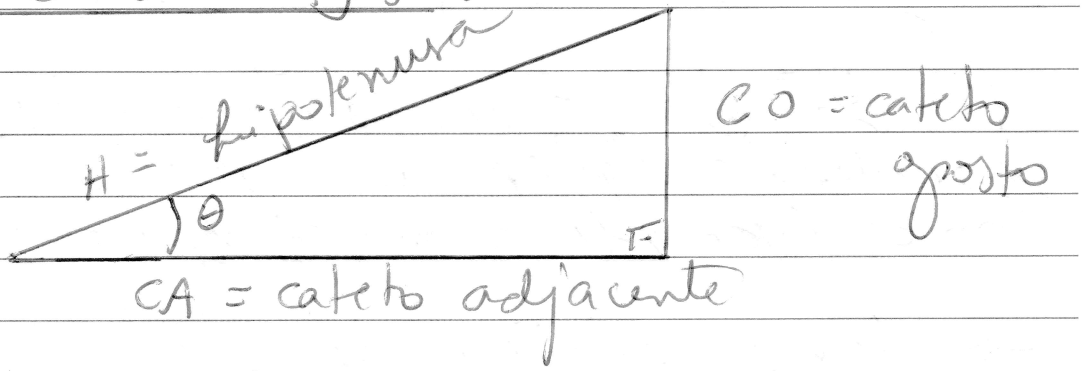
H = hipotenusa
θ
CO = cateto oposto
CA = cateto adjacente

$$\operatorname{sen}\theta = \frac{CO}{H}$$

$$\cos\theta = \frac{CA}{H}$$

$$\tan\theta = \frac{CO}{CA} = \frac{\operatorname{sen}\theta}{\cos\theta}$$

Pág. 03(13) - 2.1.1 - Time and Space
Pág. 04(14) - Fig. 2.1

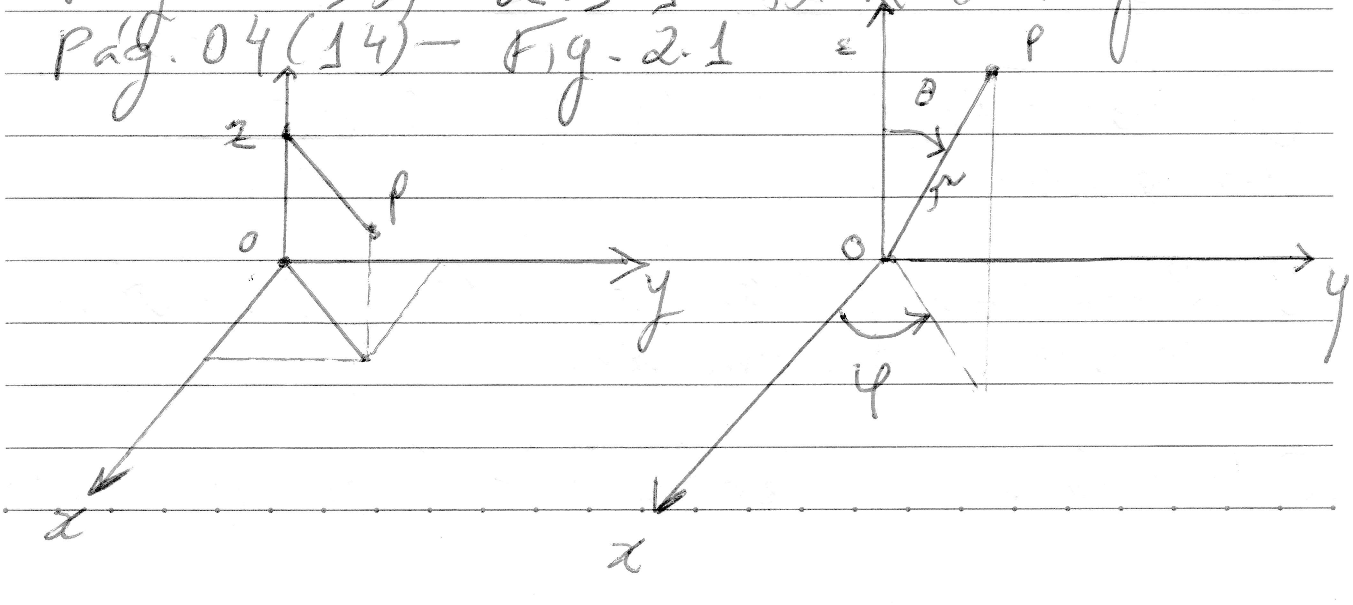
z
P
O
y
x
z
P
θ
r
O
y
φ
x

Fig. 2.1 · Cartesian (left) and spherical (right) coordinates of a point P.

Cartesian and spherical coordinates are related by

$$X = r \sin\theta \cos\varphi$$
$$Y = r \sin\theta \sin\varphi$$
$$Z = r \cos\theta$$

Notation: Abstract notation

$$(x^i) = (x^1, x^2, x^3)$$

For example:

spherical coordinates,

$x^1 = r$, $x^2 = \theta$, and $x^3 = \varphi$.

Cartesian coordinates:

$$(X^a) = (X, Y, Z).$$

Pag. 05 (15) – 2.1.2 METRIC:

$$\boxed{r^2 = X^2 + Y^2 + Z^2}$$

→ That is to say the Pythagorean theorem of Euclidean geometry.

d_{AB} = distance between two points A, B.
(Cartesian coordinates)

$$d_{AB}^2 = (X_B - X_A)^2 + (Y_B - Y_A)^2 + (Z_B - Z_A)^2 \qquad (2.4)$$

$$= \sum_{a=1}^{3}\sum_{b=1}^{3} \delta_{ab}(X_B^a - X_A^a)(X_B^b - X_A^b) \qquad (2.5)$$

$$\equiv \delta_{ab}(X_B^a - X_A^a)(X_B^b - X_A^b) \qquad (2.6)$$

[Einstein's notation].

Obs.: where, in Eq (2.5) we introduced the Kronecker symbol

$$\delta_{ab} \equiv \begin{cases} 1 & \text{if } a = b \\ 0 & \text{if } a \neq b \end{cases} \qquad (2.7)$$

and, in Eq (2.6) we used Einstein's convention for the summation over repeated indices.

EUCLIDEAN METRIC

for non-Cartesian coordinates, one cannot directly use the expression (2.4) to calculate d_{AB}. For example, with spherical coordinates

$$d_{AB}^2 \neq (r_B - r_A)^2 + (\theta_B - \theta_A)^2 + (\varphi_B - \varphi_A)^2 ; \qquad (2.8)$$

the above expression is even dimensionally incorrect.

Distance P, P' (Cartesian coordinates)

$$X_P^a \text{ and } X_{P'}^a = X_P^a + dX^a.$$

Then, applying Eq. (2.6), we have

$$d_{PP'}^2 \equiv dl^2 = \delta_{ab}\, dX^a dX^b \qquad (2.9)$$

obs: (X^a) and (x^i) describe the same space, they are related by three functions f^a such that $X^a = f^a(x^i)$.

Example:
if (x^i) denote spherical coordinates, you have derived these functions in Exercise 1:

$$f^1(r, \theta, \varphi) = r \sin\theta \cos\varphi,$$

$$f^2(r, \theta, \varphi) = r \sin\theta \underset{\hookrightarrow \text{ no exercício é } \underline{\sin}}{\cos\varphi},$$

$$f^3(r, \theta, \varphi) = r \cos\theta$$

...

Pág. 06(16) - symmetric matrix, is called the METRIC TENSOR.

→

IMPORTANT:

The metric is a machine which transforms coordinates into distances.

CURVILINEAR DISTANCE

Fig. 2.2

(LEFT)

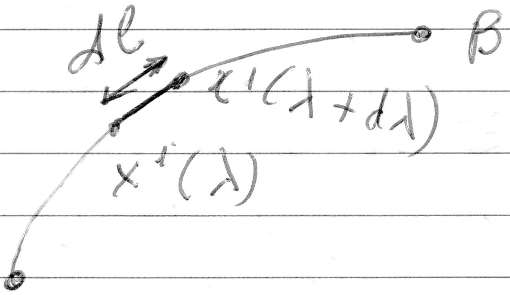
dl
B
$x^i(\lambda + d\lambda)$
$x^i(\lambda)$
A

Parametrised curve between A and B.

(RIGHT)

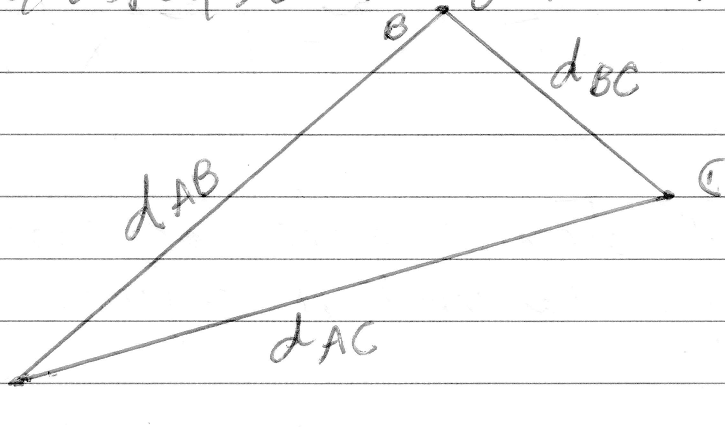
B
d_{BC}
C
d_{AB}
d_{AC}
A

Angles and distances.

(LEFT) → This curve can be parametrised by three functions $x^i(\lambda)$, where

λ is an arbitrary parameter which allows one to move along the curve, assumed to be strictly increasing from λ_A to λ_B on the way from A to B.

→

The length of the curve is obtained by summing the lengths of every infinitesimal step from A to B, that is

$$l_{AB} = \int_A^B dl = \int_A^B \sqrt{e_{ij}\, dx^i dx^j} =$$

$$\int_{\lambda_A}^{\lambda_B} \sqrt{e_{ij} \frac{dx^i}{d\lambda} \frac{dx^j}{d\lambda}}\, d\lambda \quad (2.14).$$

curve → is called a geodesic.
In Euclidean geometry and in the absence of constraints, it is simply a straight line; on the surface of a sphere, it is a great circle.

Pág. 39(49) - 3.1.2

Minkowski Metric and Four-Vectors

$$(x^\mu) \equiv (x^0, x^1, x^2, x^3).$$

$$X^\alpha \longrightarrow x^\mu$$

$$X^\alpha + dX^\alpha \longrightarrow x^\mu + dx^\mu.$$

The space-time interval between those events can then be written as

→

$$\boxed{ds^2 = \eta_{\alpha\beta}\, dX^\alpha dX^\beta = f_{\mu\nu}\, dx^\mu dx^\nu,} \quad (3.6)$$

where we introduced the

MINKOWSKI METRIC f (flat), with components

$$f_{\mu\nu} = \eta_{\alpha\beta} \frac{\partial X^\alpha}{\partial x^\mu} \frac{\partial X^\beta}{\partial x^\nu} \quad (3.7)$$

in arbitrary coordinates (x^μ), which can be seen as a four-dimensional extension of the Euclidean metric.

Pag. 40 (50) Exercise 1

Consider the transformation

$$(X^\alpha) \to (x^\mu) = (ct, r, \theta, \varphi),$$

$$T = t \quad (3.8)$$

$$X = r \sin\theta \cos(\varphi - \Omega t) \quad (3.9)$$

$$Y = r \sin\theta \sin(\varphi - \Omega t) \quad (3.10)$$

$$Z = r \cos\theta$$

where Ω is a constant

Complementação de pesquisa:
(research complementation)

I) Books-013-2022/20220625/
Mechanics of Structures Variational
and Computational Methods/

Mechanics of Structures ... pdf

Pág. 10(29) - Figure 1.7

Deformation of a three-dimensional solid.

$$Ⓐ \begin{cases} x \rightarrow x + dx \\ y \rightarrow y + dy \\ z \rightarrow z + dz \end{cases} \qquad \text{Deformation} \quad A \Rightarrow B$$

$$Ⓑ \begin{cases} x' \Rightarrow x' + dx' \\ y' \Rightarrow y' + dy' \\ z' \Rightarrow z' + dz' \end{cases}$$

$$ds^2 = dx^2 + dy^2 + dz^2 \qquad (1.7)$$

com a deformação:

$$ds'^2 = dx'^2 + dy'^2 + dz'^2 \qquad (1.8)$$

The displacement of point A to A' is caracterized by the vector u, which has the components

$u = x' - x, \quad v = y' - y, \quad w = z' - z,$ (1.9)

II) 20220625/
Newton, Lagrange, Hamilton.../

Quantum Mechanics - Symmetries (1994).pdf

Pag. 30(48) - 1.6

Time Displacements in Quantum Mechanics (Fig 1.7).

Important:

Pag. 41(59) - 1.9.2 Isotropy of Space

Isotropy of space means symmetry of the laws of nature (here of the Schrödinger equation) with respect to rotation in space.
... (estudar!!!)

2.3) 20220625/
Classical Mechanics from Newton to.../
Mach's Principle - From Newton's Bucket to Quantum Gravity (1995).pdf
→ 546 pages!   MOTION.

Pág. 29(37) — Thus, when
Einstein wrote to Mach that
"inertia its origins in a kind of interaction of bodies, quite in the sense of your reflections on
Newton's bucket experiment";
Einstein's notion of causal interaction extended well beyond the simple functional relation of phenomena.
It included relations on hypothetical and counterfactual systems of precisely the type denounced by Mach.

Off: Newton established a famous bucket experiment to prove the existence of absolute space.

20220626
2.4) Martin... Classical Mechanics...pdf
→ 241 pages. Gravity, Relativity.

2.5) Newton. Faraday. Einstein...pdf
→ 188 pages
Pág. 141(154) — Explanation 3.3
Path of light in a gravitational field.

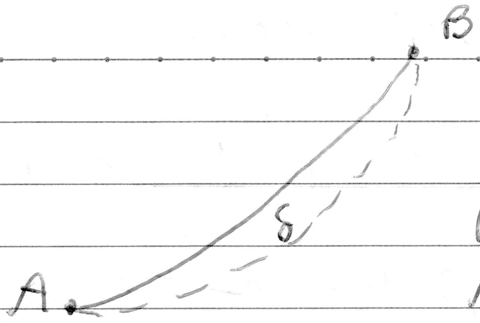

Fig. E3.3 →

Path of light passing A to B in a gravitational field

(the broken line is the new path with variation δ).

According to Fermat's principle, the path of light passing from A to B, as shown in Fig. 3.3, is the path that requires the minimum time to navigate.

This path is given by the light path decided by the solution of the

EULER EQUATION

derived from the variational principle minimizing the sum of time

$d\sigma / w$

for the propagating spatial segment $d\sigma$.

Because the Euler equation contains the gradient of the gravitational scalar potential the light path depends on gravitation, that is,

through (no entanto) the light path
is a linear line when there is no
gravitation, when there is gravitation,
the light path is not a linear but
results in the

"phenomenon of bending of light due to
gravitation".

Einstein expressed the bending angle of
light passing near the SUN as

$4kM/(c^2 r_m)$, where

k is the universal gravitational
constant,

M is the mass of the SUN,

c is the velocity of light, and

r_m is the radius of the SUN.

Obs.: From this, he predicted that
the bending angle is 1.7"

(Møller, Nagata & Ito, 1959).
→ copy deste livro
Source: 20220625/ Classical Mechanics
from Newton to Einstein.../

General Topology /

theory of relativi 029229 mbp.pdf

→ 409 pages.

Pag. 308 (326) – 117.

Propagation of light in a static gravitational field. Fermat's principle.

Pag. 309 (327) –

According to (17) this term is equal to $(1/u'^4)\,G$, where G is the gravitational force on a unit mass, thus the light ray will be

STRAIGHTEST LINE

only if the gravitational force K is zero.

Fermat's principle of least time thus holds in every static gravitational field.

Pag. 312 (328) – 310 (328)

Chapter XI – the Fundamental Laws of Gravitation in the General theory of Relativity.

The most important problem in gravitational theory, which consists in finding the general equations determining the gravitational field variables (γ_{ik}, γ_i, χ) or the g_{ik} from a given distribution of mass.

↳ Einstein in 1915 - solution

In Newton's theory of gravitation, the corresponding problem may be stated in the form of Poisson's equation

$$\Delta \chi = 4\pi K \mathring{\mu} \qquad (1)$$

where $K = 6.664 \times 10^{-8}\ cm^3\ gm^{-1}\ sec^{-2}$ (2) is the gravitational constant.

This equation enables us to calculate the gravitational potential χ when the mass density $\mathring{\mu}$ is given as a function of the space coordinates.

2.6) 20220625 / Classical Mechanics from Newton to Einstein ... /

2.6.1) p1.pdf

On the Possibility of Terrestrial Tests of the General Theory of Relativity - Møller.

2.6.2) Einstein_Relativity.pdf
→ 184 pages.
→ the special and general theory.

2.6.3) mfm-30-10.pdf

Old Problems in the General Theory of Relativity viewed from a new angle → Møller.
→ 30 pages

2.6.4) gr.pdf (Livro todo)

General Relativity

Pág. 43(50) – 1.3.5 Light Bending

2.6.5) v11.pdf (interessante!!!)

Chapter 11 – Møller Scattering

Enviei um email para o
Dr. Axel Pelster
axel.pelster@physik.uni-kl.de
com os seguintes dizeres:

"I am an independent researcher and I found your publication: "Möller Scattering", copy attached. I am currently study Møller,

but specifically in the theory of Relativity and I found your publication very interesting, in particular, the geometry of the elastic Møller scattering in the center of mass (v11 - pp. 250(09) - Fig. 11.2)
22(20)
should be represented differently, because the momenta of electrons are slightly different.

Picture 11.2

Analyzing the two Mandelstam variables (11.124) I think that there is an inaccuracy and the analysis that results from the dependence on both the scattering energy E and the scattering angle θ also.
Sorry there are any inaccuracies as I study your area to provide input in my research.

Ti
(M.U.S.)

Regards
Prof. Stenio Musich

Send: 20220626 16h07 min. (Gmail)

2.6.6) Einstein and Møller - Momentum Distributions.pdf
→ 12 pages

2.7) Special relativity, electrodynamics, and general relativity...pdf
→ 441 pages

Pág. 225 (234) 11.7

Bending of Light in a Gravitational Field

IMPORTANT:

11.8 - Similarities and Differences of Electromagnetism and Gravity

Obs: Estudar tudo!!!

"Esta similitude decorre do fato de que a gravidade é decorrente das demais forças fundamentais da física. A gravidade não é independente.

Source: 20220625/ Classical Mechanics from Newton to Einstein/ Special relativity, electrodynamics, and general relativity...pdf
Pág. 229 (238) - 11.8

2.8) 20220625/Classical Mechanics from
Newton to Einstein/
the theory of relativity Møller/

fulltext.pdf

→ An Introduction to the Special Theory of
Relativity.
→ 130 pages.

2.9) ... The Theory of Relativity Møller/
9406046.pdf
→ 13 pages.
Energy-momentum complex in Møller's
Tetrad Theory of Gravitation.
Obs.: Estudar.

2.10) ... New Cosmology.pdf
→ 51 pages
Physics Tomorrow: New Cosmology and
Gravitation in Modern Physics.
June 16, 2017.

2.11) 20220625/
Classical Mechanics from Newton
to Einstein.../
General Topology/

2.11.1) Encyclopedia of General
Topology....pdf

→ 537 pages.

2.11.2) JMP_2018022316401524.pdf
→ 8 pages.
Asymmetric Special theory of Relativity.
obs: estudar!!!

2.11.3) Nagata - Topics in General Topology....pdf
→ 761 pages.

Complementos de pesquisa:

Ⓐ 20220625/ Advanced Mechanics and General Relativity/
Franklin - Advanced Mechanics and ~~Relat~~ General Relativity....pdf
→ 387 pages.

Pág. 255(275) - 7.1.2

The Weyl method and GR
Spherically symmetric ansatz

spherically symmetric ansatz for the metric.
In spherical coordinates, we could write the MINKOWSKI METRIC as in Fig. 7.1

⇐

Ⓑ 20220626/ Newton's Physics.../
Newton's Physics and the Conceptual Structure of the Scientific Revolution...pdf
→ 604 pages.
Obs.: Livro de estudo!!!

© 20220626/ La Loi de la gravitation universelle/
Prosper Schroeder - La Loi de la gravitation universelle...pdf
→ 562 pages.

Obs.: Newton, Euler et Laplace.

→ interessante!!! Ver este ponto de vista!

Retornando um pouco:

⇙

20220625/ Classical Mechanics from Newton to Einstein.../

Newton.Faraday.Einstein...pdf

Pág. 144(157) — Explanation 3.4

Analysis of the perihelion motion of Mercury.

Obs. Continuando neste mesmo diretório: →

2.12) Mach's Principle - from Newton's bucket to Quantum Gravity...pdf
→ Pág. 546, digo, 546 pages.

2.13) Martin... Classical Mechanics - from Newton to Einstein...pdf
→ 241 pages
Obs.: Estudar!!!

2.14) Special relativity, electrodynamics, and general relativity...pdf
→ 441 pages
Obs.: Livro todo!!!

2.15) The ascent of gravity - the quest to understand the force...pdf
→ 221 pages.

2.16) The concepts of Science - from Newton to Einstein (1988).pdf
→ 444 pages.

2.17) The Evolution of Scientific thought from Newton to Einstein... pdf
→ 504 pages.
Obs.: Livro todo!!!

Pág. 300 (302) - In view of the incompatibility of

Newton's law with the observed uniform distribution of matter, Neumann and Seeliger in 1896 suggested a slight modification of the law ~~the~~ of gravitation, which would have for effect to render a uniform distribution possible.

The equivalent of

POISSON'S EQUATION

when the Neumann-Seeliger law is used, is

$$\Delta \phi - \lambda \phi = 4 \pi k \mu,$$

where λ is a positive constant, which may be as small as we choose.

Assuming the density μ to be constant, this modified law, contrary to Poisson's, is compatible with the constancy of the potential, and so is compatible with a

UNIFORM DISTRIBUTION OF HOMOGENEOUS MATTER

Obs.: Continuar lendo!!!

2.18) Variational Approach to
Gravity Field Theories....pdf
→ 364 pages.

Obs.: Livro todo!!!

Pág. 70 (81) - 4.4.2 -
Newtonian Lagrangians:
Poisson Equation and its
Extension.

Obs: Estudar!!! Interessante!!!

3) 20220625/
Classical Mechanics Point Particles
and Relativity/ (todo!)

20220627

1) 20220625/Newton. Lagrange. Hamilton.../

1.1) New Foundations for Classical
Mechanics (geometric algebra)..pdf

Outras cópias: 20220627/ New Foundations
for Classical Mechanics.../
Hestenes - New Foundations for Classical
Mechanics (2008).pdf
→ 718 pages.

Pág. 39(54) - Chapter 2
Developments in Geometric Algebra.

Pág. 55(70) - Fig. 3.1 - ~~Orthogonal~~
Orthonormal basis for g_3

$$I = \sigma_1 \sigma_2 \sigma_3 = \sigma_1 \wedge \sigma_2 \wedge \sigma_3$$

Obs. : Estudar!!! ⇙

Todo este capítulo é importante, nós devemos adequá-lo para uma nova geometria M (M.U.S.)
→ vai ser desafiador.

1.2) Quantum Mechanics - Symmetries (1994).pdf

Pág. 66(83) - 2.3 - Addition of Two Angular Momenta.

$$\hat{J} = \hat{J}_1 + \hat{J}_2 .$$

two angular momenta $\hat{J}_1$ and $\hat{J}_2$ which are combined to form a resulting total angular momentum $\hat{J}$,

Pág. 183(199) - 7. The SU(3) Symmetry

1.3) Richard Tolman - Geometric mechanics...(2007).pdf

Pág. 89 (106) - Chapter 3

Geometry of Mechanics, II, Curvilinear

$n > 3$

Pág. 90 (107) - 3.1

(Real) Curvilinear Coordinates in n-Dimensions

3.1.1 - The Metric Tensor.

Obs.: ↙

$$\cos\theta = \frac{x \cdot y}{\sqrt{(x \cdot x)(y \cdot y)}}$$

2) 20220625T
Classical Mechanics Point Particles and Relativity/

Walter Greiner - Classical Mechanics - Point Particles and Relativity...pdf

Pág. 29 (48) - Example 6.6 →

The rotation matrix.
→ capitulo todo!!!

3) 20220626/
Gravitation/

3.1) eddington.pdf
→ 289 pages.

3.2) Gravitation...(2017).pdf
→ 1333 pages.

Pág. 193(246) - Part III
The Mathematics of Curved Spacetime

Pág. 200(253) - Fig. 8.1. ↙

A Keplerian orbit in the sun's gravitational field, as treated using the standard Euclidean-space version of Newtonian gravity.

The basis vectors themselves change from point to point along the orbit [equations (8.3)].

This figure illustrates the pictorial aspect of differential geometry. Later (exercise 8.5) it will illustrate the concepts of "covariant derivate" and "connection coefficients".

3.3) the Theory of Indistinguishables.pdf
→ 227 pages.

Pág. 39 (51) – Chapter III

The Physical relevance of indistinguishables

3.A. The concept of "PLANES"

→ epistemology

3.4) The Universe in a Helium Droplet.pdf
→ 532 pages.

Pág. 29 (51) – 3.3.4

Vanishing of cosmological constant in liquid 4He.

Obs.: Important !!!

3.5) Weintraub. Differential forms, a complement to vector calculus.pdf
→ 268 pages.

Pág. 95 (107) – Chapter IV

⇒

Integration of Differential Forms over Oriented Manifolds

Pág. 96(108) - 2. ⇙

The Integral of a 1-FORM over a curve (LINE INTEGRALS)

→ important!!! Estudar!!!

3.6) Zero to Infinity... pdf
→ 738 pages.

Pág. 444(496) - Chapter 17
Gravity and Inertia.

Interessante: (Important)

Pág. 463(488) - 17.9 ✡

Perihelion and Periastron Precession ⇙

Classical calculations of gravitational light bending and time delay become relatively ease once we have established a special relativistic explanation, as relativistic energy equations for systems without boost mass (and, therefore, without kinematics) are specifically designed to coincide with classical ones.

Relativistic correction for planetary and pulsar orbits is one of the most significant results derived from the

SCHWARZSCHILD solution

(only a special relativistic correction to classical gravity).

Obs.: Livro todo!!!

4) 20220626/
the Statistical Theory of Shape/

The Statistical Theory of Shape ... Small.pdf

→ 237 pages
→ Livro todo!!!

Pag. 06(17) – 1.3

Defining and Comparing Shapes.

—

IMPORTANT (voltando para 3.6)

No livro 3.6) – Zero ~~do~~ to Infinity...pdf

Pág. 463(496) – Chapter 17

17.9 – Perihelion and Periastron Precession

Estudar este tópico com uma outra parte de um outro livro.

Source:

Books-013-2022/20220625/
Classical Mechanics from Newton to Einstein...pdf

Newton. Faraday. Einstein. – From Classical Physics to Modern Physics..(2021).pdf

Pág. 143(156) – ~~The Solar~~

The total solar eclipse observed by Eddington.

Em especial:

Pág. 144 (157) – Explanation 3.4

⇒

Analysis of the perihelion motion of Mercury!

5) 20220625/
Newton. Lagrange. Hamilton.../

Geometric Mechanics – toward a Unification of Classical Physics (2007).pdf

Pag. 93(105) – Fig. 3.1 (interessante)

Relating the "natural" local coordinate at two different points in ordinary space described by curvilinear coordinates. Because this is a Euclidean plane, the unit vectors e_1 and e_2 at M can be "paralled shift" to point M + dM without changing their lengths or directions; they are shown there as dashed arrows.

The curve labeled $u^1_M + 1$ is the curve on which u^1 has increased by 1 and so on. (assim por diante).

↙

6) 20220627/ –
New Foundations in Mathematics the Geometric Concept of Number/

Garret Sobczyk - New Foundations in Mathematics -- The Geometric Concept of Number...pdf
→ 372 pages.

Pág. 243 (249) - Chapter 14

Differential Geometry of Curves

Pág. 253 (258) - Chapter 15

Differential Geometry of k-Surfaces.

We have discussed the calculus of a k-surface, or a k-manifold $\mathcal{M}$ embedded in $\mathbb{R}^n$ in Chap. 13, utilizing the basic building block of a k-rectangle.

In 1936, Hassler Whitney and, later in 1956, John Nash proved a number of embedding theorems which show that an abstract k-manifold can always be isometrically (preserving distances) embedded in $\mathbb{R}^n$ if

$$n \geqslant k^2 + 5k + 3.$$

obs.: Whitney-Nash embedding theorems.

20220628M

1) Books-013-2022/20220628/
Mathematical Physics with Partial
Differential Equations/

Mathematical Physics with Partial
Differential Equations.pdf
→ 474 pages.

2) 20220628/ Rigid Body Dynamics
A Lagrangian Approach/

Yehia2022_Book_RigidBodyDynamics.pdf
→ 473 pages.

obs.: Interessantíssimo!!!

3) 20220628/ Pi in the Sky/
The Constants of Nature - From
Alpha to Omega....pdf
→ 367 pages.

Pág. 254 (269)

$$\alpha(z=1.95)/\alpha(z=0)=0.97\pm 0.05$$

↳ redshift

Quasar 3C191

4) 20220628/Mathematical foundations of quantum mechanics Von Neumann/

Neumann... Mathematical Foundation...pdf
→ 454 pages.

Obs: Hilbert Space (Important!!!

⇒ Livro todo!!!

5) 20220628/A survey of Geometry Eves/

5.1) Howard Whitley Eves - College Geometry...pdf
→ 440 pages. Interessante!!!

✓✓

5.2) Eves, Howard - Introdução à história da matemática...pdf
→ 849 pages.

Pág. 129 (130) - Capítulo 4.

Duplicação, trissecção e quadratura

Obs. Livro todo!!!

✓✓

5.3) Howard Eves - A Survey of Geometry...pdf ✓✓
→ 464 pages - Volume 1

5.4) Howard Eves - A survey of Geometry 2.pdf
→ 512 pages - Volume 2.

Important:

20220628/A survey of geometry Eves/

Howard Eves - A survey of geometry
Volume 01.pdf

Pag. 23(45) - π

π lies between $\frac{223}{71}$ and $\frac{22}{7}$

"... In the Quadrature of the Parabola, which contains 24 propositions it is show that the area of a parabolic segment is $\frac{4}{3}$ that of the inscribe triangle having the same base and having its opposite vertex at the point where the tangent is parallel to the base.
The summation of a convergent geometric series is involved.

"... One Arabian writer claims that Archimedes was the discoverer of the celebrated formula

$\Rightarrow$ ~~$K = \sqrt{s(s-a}$~~ $\boxed{K = \sqrt{s(s-a)(s-b)(s-c)}}$

for the area of a triangle in terms of its three sides. This formula is found in a latter work of Heron of Alexandria.

Pág. 64 (86) - 2.3.2

Menelau's Theorem. A necessary and sufficient condition for three menelaus points D, E, F for the sides BC, CA, AB of an ordinary triangle ABC to be collinear is that

$$(\overline{BD}/\overline{DC})(\overline{CE}/\overline{EA})(\overline{AF}/\overline{FB}) = -1.$$

Necessity. Suppose (see Figure 2.3a) D, E, F are collinear on a line l ⇙

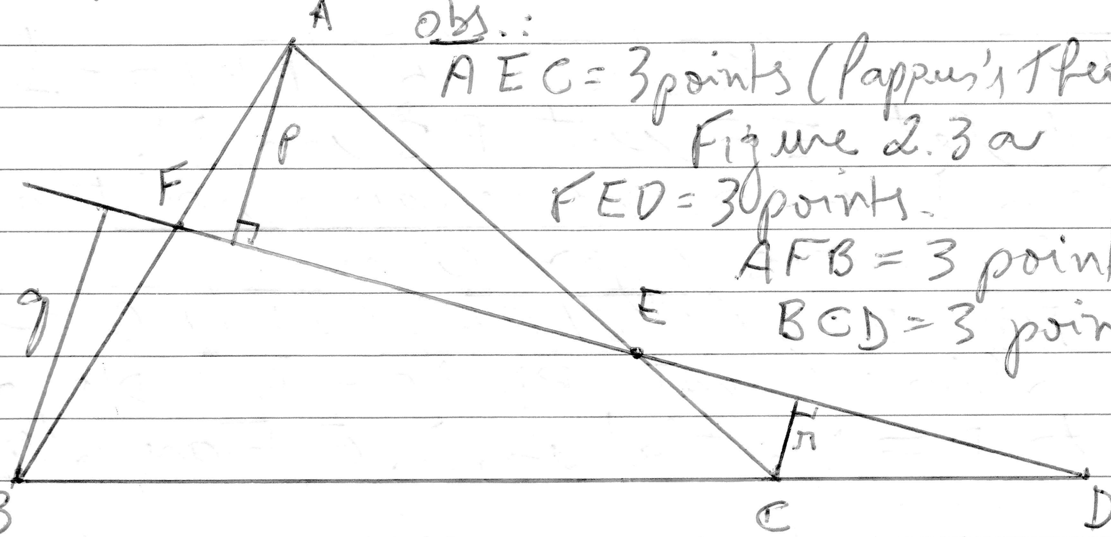

obs.:
AEC = 3 points (Pappus's Theorem)
Figure 2.3a Ⓐ
FED = 3 points.
AFB = 3 points.
BCD = 3 points.

which is not the line at infinity. Drop perpendiculars p, q, r on l from A, B, C. Then, disregarding signs,

$$BD/DC = q/r, \quad CE/EA = r/p,$$

$$AF/FB = p/q.$$

⇒

It follows that

$$(\overline{BD}/\overline{DC})(\overline{CE}/\overline{EA})(\overline{AF}/\overline{FB}) = \pm 1.$$

Since (uma vez que), however (porém), l must cut one or all three sides externally, we see that we can have only the − sign.

If l is the line at infinity, the proof is simple.

Sufficiency. Suppose

$$(\overline{BD}/\overline{DC})(\overline{CE}/\overline{EA})(\overline{AF}/\overline{FB}) = -1$$

and let EF cut BC in D'. Then D' is a Menelaus point and, by the above (por cima)

$$(\overline{BD'}/\overline{D'C})(\overline{CE}/\overline{EA})(\overline{AF}/\overline{FB}) = -1.$$

It follows that $\overline{BD}/\overline{DC} = \overline{BD'}/\overline{D'C}$, or that $D \equiv D'$. That is, D, E, F are collinear.

Veja também 2.3.3. – Trigonometric form of Menelaus' theorem.

⋯

Obs: Ler todo o item 2.3 – ⇙

The Theorems of Menelaus and Ceva

Important: Li os 2 livros de Howard Eves
→ A survey of Geometry.

6) Books-013-2022/Importantissimo-02/
Applied General Relativity...pdf.

Pág. 152(169) - 4.6. Energy and Momentum

$$E = mc^2 \qquad (4.6.6)$$

7) Books-013-2022/Importantissimo-02/
Borsuk/
Borsuk-Foundations of geometry...pdf

Pág. 150(168) - 37. Parallel lines.

20220629

Books-013-2022/20220629/
1) Energy and Geometry An Introduction to Deformed Special Relativity/

1.1) Deformed spacetime - geometrizing interactions in four and five dimensions...pdf
→ 523 pages.

Pág. 49(68) -

⇙

many gravitational phenomena (like the bending of light rays near a massive body), usually described in terms of the curved (Riemannian) structure of space-time, can be treated by considering (in an Euclidean space) the vacuum as a polarizable medium endowed with an effective refraction index (see (3.72)), in which light propagates [16, 29].

⟶ [16] [29] } à frente

IMPORTANT:

Pág. 37(56) – In the case of an isotropic 3D space (i.e., $b_1^2(E) = b_2^2(E) = b_3^2(E) \equiv b^2(E)$) the corresponding deformed metric is given by (2.14), (2.19) and one gets, for any component of u:

$$u = \frac{b_0(E)}{b(E)}\, c. \qquad (3.70)$$

Cf.: Complemento (2.14) ↙

Pág. 13(32) – The orthogonality of the coordinates axes is of course expressed by

$\widehat{x^{\mu}_{SR}} \cdot \widehat{x^{\nu}_{SR}} = \delta^{\mu\nu}$ and

$$\hat{x}^{\mu}_{DSR}(\{\mathcal{O}\}) * \hat{x}^{\nu}_{DSR}(\{\mathcal{O}\}) = \delta^{\mu\nu}.$$

It follows by the earlier relation that the unit vectors in the standard Minkowski space M and in the deformed one $\tilde{M}(\{\mathcal{O}\})$ are proportional, and therefore they specify the same directions in either space (qualquer spaço) (this is due of the

METRIC TENSOR,

while destroying its isochrony and spatial isotropy).

However, since coordinate unit vector is rescaled by a different coefficient in the metric deformation, vectors with the same components in the spaces M and $\tilde{M}(\{\mathcal{O}\})$ are not parallel to each other.

...

Pag. 14(33) – equation (2.14)

Isotropic 3D Space

$$b_1^2(\{\mathcal{O}\}) = b_2^2(\{\mathcal{O}\}) = b_3^2(\{\mathcal{O}\}) \equiv b^2(\{\mathcal{O}\}), \quad (2.13)$$

Corresponding deformed metric ~~field~~ reads:

$$g_{\mu\nu}[\delta_{\mu 0} b^2(\{0\}) -$$

$$(\delta_{\mu 1} + \delta_{\mu 1} + \delta_{\mu 1}) b^2(\{0\})]. \quad (2.14)$$

……

pág. 15(34)

Therefore, the energy-dependent deformation of the MINKOWSKI METRIC implements a generalization of the concept of geometrization of an interaction (in accordance with Finzi's principle).

The GR theory implements a geometrization (at a global scale) of the gravitational ~~inter~~ interaction, based on its derivability from a potential and on the equivalence between the inertial mass of a body and its

"gravitational charge".

Obs: Esta página aborda:
- gravitational
- electromagnetic,
- strong
- weak.

Pág. 16 (35) - (2.19) equation

1. A space $\tilde{M}$ spatially isotropic:

$b_1(E) = b_2(E) = b_3(E) = b(E)$ (2.19)

2. A space $\tilde{M}$ locally conform with the MINKOWSKI space M:

$b_0(E) = b_1(E) = b_2(E) = b_3(E) = b(E)$ (2.20)

this is a particular case of the

MIRON-TAVAKOL METRIC [13]
which satisfies the

Ehlers-Pirani-Schild axiomatics.

Voltando: (2 páginas atrás)

[16] Fight in Einstein's Universe
Prokhovnik
ISBN 90-277-2093-2

Source:
20220629/ Fight in Einstein's Universe/

a) Prokhovnik - Fight in Einstein's Universe...pdf
→ 228 pages.

b) Katy Price - Loving faster than light.pdf
→ 275 pages.

[29] - Gravitation, Gauge Theories and the Early Universe

Source:
20220629/ Light in Einstein's Universe/
Gravitation, Gauge Theories and the Early Universe ... (1989).pdf
→ 545 pages

Obs.: Livros todos!!!

Pág. 06(18) - 6 - The Fire Element.

We know that the geometry on the surface of a sphere is Riemannian.
Consider a point P on the surface of a sphere of unit radius (Figure 1a). Its position is given by two polar coordinates
(θ, φ), θ being the colatitude
($\angle NOP$) and φ the longitude
($\varphi = \angle XOP$).

Obs.: COLATITUDE →

A colatitude é utilizada em

astronomia porque se refere à distância zenital dos polos celestes.

↳ Sistema de Coordenadas Celestes.

Uma coordenada mede a posição do astro nesse círculo máximo, e a segunda, perpendicular mede a distância do astro a esse plano.

Paralaxe — a posição de uma estrela é medida 2 vezes em momentos diferentes, pois a Terra está em outra posição. Desta forma, através da observação da estrela nas posições A e B, os astrônomos são capazes de medir o ângulo e com isto, determinado a distância da estrela com o uso deste ângulo e a trigonometria.

Paralaxe heliocêntrica ou anual → é definida com a diferença de posição de uma estrela com vista da Terra e do Sol. Como não podemos ver a estrela do Sol, a observação é feita entre dois pontos opostos da órbita da Terra e o resultado dividido por 2 (dois).

O parsec é a distância para a qual a paralaxe anual é de um segundo de arco ou arcseg. Um parsec é igual a 3,26 anos-luz.

Para cálculo do Parsec utilizamos a translação da Terra em volta do Sol, cujo raio é de 1 AU (Unidade Astronômica).

Portanto, um parsec é a distância do Sol a um objeto astronômico que tem um ângulo de paralaxe de um segundo de arco.

Em agosto de 2015, a IAU (União Astronômica Internacional) aprovou a resolução B2 que, como parte da definição de uma escala de

MAGNITUDE BOLOMÉTRICA ABSOLUTA

e aparentemente padronizada, mencionou uma definição explícita existente do parsec como exatamente

$$\boxed{\frac{648000}{\pi}\ \text{au}}$$, ou aproximadamente

3.0856775814913673 x 10^{16} metros (com base na definição exata do SI da IAU de 2012 da unidade astronômica).

Calculando o valor de um parsec

1 au = 149 597 870 700 m
(exatamente pela definição de 2012 do au)

Definição de 2015:

1 au de comprimento de arco subtende um ângulo de 1" no centro do círculo de raio de 1 pc.
Ou seja, $1\ pc = \dfrac{au}{\tan(1'')} \cong$
206,264.8 au por definição.(9)

(9) Source:

Books-013-2022/20220629/
To the Sun and Beyond/
s41567-019-0685-3.pdf
→ 1 page
Copy: ARQ/GGB-Geogebra-2022/
20220629/
To the Sun and Beyond/

Conversão de unidades de graus/minuto/segundo em radianos,

$$\frac{1\ pc}{1\ au} = \frac{180 \times 60 \times 60}{\pi}, \text{ e}$$

1 au = 149597480 000 m
(def. de 2012)

Portanto,

π pc = 180 × 60 × 60 au =

180 × 60 × 60 × 149 597 870 700 =

96 939 420 213 600 000 m
(exatamente pela definição de 2015)

Portanto,

$$1 \text{ pc} = \frac{96939420213600000}{\pi}$$

$\cong$ 30 856 775 814 913 673 m
(para o metro mais próximo)

Aproximadamente,

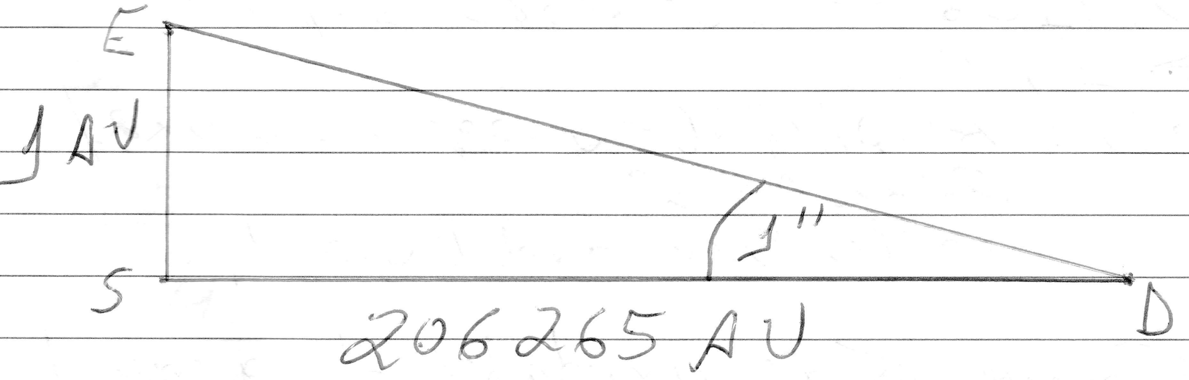

S representa o Sol
E a Terra em um ponto de sua órbita
ES = 1 AU

→

O ângulo SDE é de um segundo de arco ($\frac{1}{3600}$ de um grau), então, por definição, D é um ponto no espaço a uma distância de um parsec do Sol.

Por meio da trigonometria, a distância SD é calculada da seguinte forma:

$$SD = \frac{ES}{\tan 1''}$$

$$SD \approx \frac{ES}{1''} = \frac{1\ \text{au}}{\frac{1}{60\times60}\times\frac{\pi}{180}} = \frac{648000}{\pi}\ \text{au}$$

≈ 206241.81 a u.

obs.: 1 pc ≈ 3.261563777 ly
(light-years)

Um corolário afirma que um parsec é também a distância a partir da qual um disco de uma unidade astronômica de diâmetro (penso de raio) (→ conforme desenho anterior) deve ser visto para ter um diâmetro angular de um segundo de arco.

Matematicamente, para calcular a distância, dada as medidas angulares obtidas a partir de instrumentos em segundos de arco, a fórmula seria:

$$\text{Distância}_{\text{estrela}} = \frac{\text{Distância}_{\text{Terra-Sol}}}{\tan \frac{\theta}{3600}}$$

θ é o ângulo medido em segundos de arco.
Distância Terra-Sol é uma constante (1 au ou 1.5813×10^{-5} ly).

A distância estelar calculada estará na mesma unidade de medida usada em Distância Terra-Sol.

Obs: Parsec $\frac{648000}{\pi}$ au

Este valor difere da definição clássica de tangente inversa em cerca de 200 km, ou seja, somente após o 11° algarismo significativo.

Complementando:

Magnitude Absoluta - Magnitude Bolométrica

⟹

Magnitude Absoluta →
é a medida de brilho intrínseco de um objeto celeste.
É a magnitude aparente hipotética do objeto a uma distância padrão de exatamente 10 parsecs (32,6 anos-luz) do observador, assumindo-se nenhuma extinção astronômica da luz da estrela.

Magnitude Bolométrica →

Mbol leva em consideração a radiação eletromagnética em todos os comprimentos de ondas.

Voltando: COLATITUDE

A colatitude é utilizada em astronomia porque se refere à distância zenital dos polos celestes.
Por exemplo para um observador localizado sobre a Terra em um ponto de latitude 42°N, o arco esférico entre a estrela Polar (aproximadamente situada sobre o polo Norte celeste) e o Zênite (ponto imaginário acima da cabeça do observador) é de
90 - 42 = 48°.

Projetando-se o zênute (zênite) do observador sobre a esfera celeste, a sua colatitude é de 48°.

Adicionando a declinação de uma estrela

obs.: declinação (δ) de um astro é o arco do meridiano do astro compreendido entre o plano do equador celeste e o astro.
→ latitude.

Adicionando a declinação de uma estrela à colatitude do observador obtém-se a máxima latitude desta estrela
(seu ângulo do horizonte à culminação ou trânsito superior)
Por exemplo, se Alfa Centauri é vista com uma latitude de 72° norte (108° Sul) e sua declinação é conhecida (60° S), então pode-se determinar que a colatitude do observador é
108 − 60 = 48
(i.e. sua latitude é 90 − 48 = 42° S

Voltando ao livro: Gravitation, Gauge Theories and the Early Universe...pdf

Obs.: Arquivo Geogebra → Source: ✓

ARQ/ GGB-Geogebra-2022/ 20220629/ sphere-riemannian--20220629.ggb.

Cont.:

IMPORT: Usar bending of light rays.

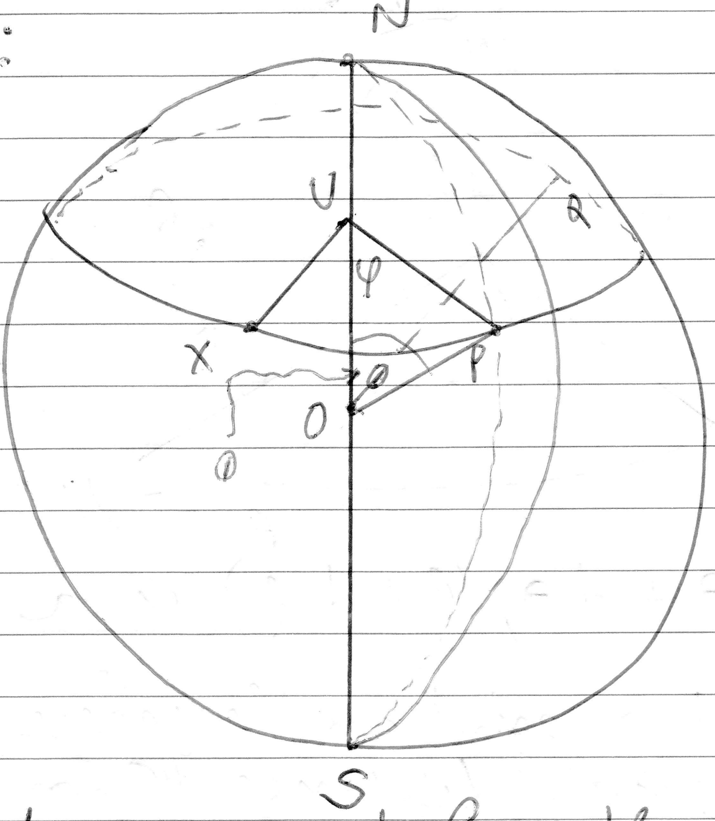

geometry on the surface of a sphere is Riemannian.

Consider a point P on the surface of a sphere of unit radius. Its position is given by two angular coordinates (θ, φ),

θ being the colatitude ($\angle NOP$) and φ the longitude ($\varphi = \angle XUP$).

If θ is constant and φ varies from 0 to 2π, the point P moves along the small circle of radius $\sin\theta$.

If φ is constant and θ varies from from 0 to π (90°),
the point P moves along the meridiano large circle NPS.

Fig. 1b

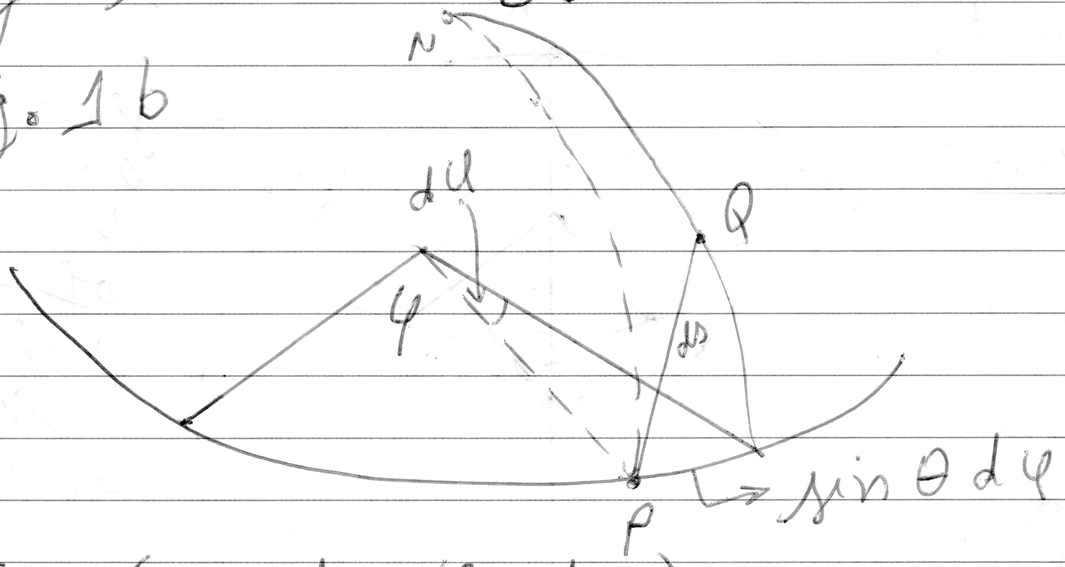

If $Q(\theta + d\theta, \varphi + d\varphi)$ is a neighbouring point on the sphere at an infinitesimal distance from P (Figure 1b) then

$$ds^2 = d\theta^2 + \sin^2\theta\, d\varphi^2 \qquad (1).$$

It is customary to regard relations like (1) as defining a line element.

Formula (1) is the line element on the surface of a sphere or a line element of a two-dimensional Riemannian space.

If, however, we were considering two neighbouring points on a plane with coordinates (x, y),

$(x + dx, y + dy)$,

the line element would be

$$\boxed{ds^2 = dx^2 + dy^2.} \qquad (2)$$

↳ Equation (2) is a line element of a two-dimensional Euclidean space.

Pág. 08 (20) - Riemannian spacetime

$(x^1 + dx^1, x^2 + dx^2, x^3 + dx^3, x^4 + dx^4)$

We take the line element as a quadratic in dx^1, dx^2, dx^3, dx^4.

$$ds^2 = g_{11}(dx^1)^2 + 2g_{12}(dx^1)(dx^2) +$$
$$2g_{13}(dx^1)(dx^3) + 2g_{14}(dx^1)(dx^4) +$$
$$g_{22}(dx^2)^2 + 2g_{23}(dx^2)(dx^3) +$$
$$2g_{24}(dx^2)(dx^4) + g_{33}(dx^3)^2 +$$
$$2g_{34}(dx^3)(dx^4) + g_{44}(dx^4)^2$$

or we may write it as

⇒

$$ds^2 = \sum_{i=1}^{4} \sum_{k=1}^{4} g_{ik}\, dx^i dx^k ,$$

where
g_{ik} are 10 functions of "position",
i.e. of the four variables x^1, x^2, x^3, x^4,
symmetric in i and

k, $g_{ik} = g_{ki}$.

Obs: Livro todo!!!

Voltando para continuar:

20220629/
Energy and Geometry An Introduction to Deformed Special Relativity/

Deformed spacetime - geometrizing interactions in four and five dimensions... pdf

Pag. 49 (68) - que faz menção à página 37 (56).

Equation 3.72. ⟹ está mencionada

na página 50 (69).

☒ Cont. (páginas atrás).

In the case of isotropic 3D space ($i.e,\ b_1^2(E) = b_2^2(E) = b_3^2(E) \equiv b^2(E)$), the corresponding deformed metric is given by (2.14), (2.19), and one gets, for any component to u:

$$u = \frac{b_0(E)}{b(E)}\, c. \qquad (3.70)$$

Cf.: (2.14) – Pág. 14(33)

Moreover, since the usual "flat" MINKOWSKI METRIC g (2.1) is related in an essential way to the electromagnetic interaction, we shall always mean in the following – unless otherwise specified – that electromagnetic interactions imply the presence of a fully Minkowskian metric.

Pág. 09(28) – 2.1 Deformed Minkowski Space-Time

Minkowski metric $g = \mathrm{diag}(1, -1, -1, -1)$ is a generalization of the Euclidean metric

$\epsilon = \mathrm{diag}(1,1,1)$.

(2.19) → já anotado.

3) Books-013-2022/20220629/
A Friendly Introduction to Abstract Algebra/
A Friendly Introduction...pdf

Pág. 41 (57) - Chapter 5 - Symmetries.

4) 20220629/Gravitation From the Hubble length to the Planck Length/
Gravitation - From the Hubble...pdf
→ 184 pages.

5) 20220629/
Energy and Geometry An Introduction to Deformed Special Relativity/

Energy and Geometry - An Introduction to Deformed Special Relativity (2004).pdf
⇒ 161 pages.

IMPORTANT:

Estudar todas as pastas do diretório Books-013-2022/20220629/

20220630

1) Books-013-2022/20220630/
An introduction to the theory of relativity/
Rosser - An introduction to the theory of relativity...pdf
→ ~~406 pages~~ 528 pages.
Pág. 248 (260)

Chapter 6 - Geometrical Representation of Lorentz Transformation and the use of 4-vectors.

Pág. 303 (315) - Chapter 8

The Relativistic Transformation of Maxwell's Equations.

2) 20220630/
Albert Einstein .../
Kiefer ... Quantum gravity...pdf
→ 406 pages.

Pág. 141 (154) - Chapter 5

Quantum geometrodynamics.

The implementation of Dirac's procedure is the translation of POISSON brackets into commutators for the fundamental variables, that is,

$$V_3 = \{V_1, V_2\} \Rightarrow \hat{V}_3 = -\frac{i}{\hbar}[\hat{V}_1, \hat{V}_2]. \quad (5.2)$$

3) 20220630/
Albert Einstein.../
Can Quantum-Mechanical Description of Physical Reality be considered complete.pdf
→ 118 pages.

4) 20220630/
Albert Einstein.../
From Quantum to Classical...pdf
→ 314 pages.

5) 20220630/ Theory of Relativity - Møller - others/
a) The Geometry of Minkowski Spacetime ... (2012).pdf
→ 341 pages.

Pag. 07(24) - Chapter 1

Geometrical Structure of M.

b) Synge ... Relativity...pdf
→ 464 pages.

JULHO 2022

20220701 ///

Data-NovoVivo-014-2022/
Books-014-2022/
20220701/
Symmetry/

a) 2110.13475.pdf
→ 30 pages
→ Vector-valued Distance and Gyrocalculus on the Space of Symmetric Positive Definite Matrices.

Pág. 02 – 3 – The Space SPDn.

The space SPD_n is a Riemannian manifold of non-positive curvature of $\boxed{\underbrace{n(n+1)/2}}$ dimensions.

Points in SPDn are positive definite real symmetric n x n matrices, with the identity matrix I being a natural basepoint.

The tangent space to any point of SPD_n can be identified with the vector spaces S_n of real symmetric n x n matrices.

SPD_n contains n-dimensional

Euclidean subspaces, (n-1)-dimensional hyperbolic subspaces as well as products of $\left[\frac{n}{2}\right]$ hyperbolic planes.

Obs.: see Helgason [39] for an in-depth introduction.

[39] - Helgason, S (1978). Differential geometry, Lie groups, and symmetric spaces.

Copy: 20220701/ Symmetry/ Differential geometry, Lie groups, and symmetry/

Differential Geometry, Lie Groups, and Symmetry ...(1978).pdf (10.1 MB)

→ 635 pages.

Obs.: Livro todo !!!

Voltando: 20220701/ Symmetry/ 2110.13475.pdf

Pág. 03 - Figure 2.

Some isometries of SPDn have analogous Euclidean counterparts. Translation (left), rotation (center) and reflection (right).

Veja também: Pág. 05 - Figure 4

Gyro-addition (left), gyro-scalar multiplication (center) and matrix scaling (right).

b) Exercise 10 - Puzzetti

Advanced Geometry II - Symmetric Spaces.

Obs: Let X be a Riemannian symmetric space.

c) 20220701/
The Secret Lives of Numbers
Numerals and their Peculiarities.../
The Secret Lives of Numbers...pdf

Pág. 18(27) - $\frac{223}{71} < \pi < \frac{22}{7}$

Pág. 32(41) - Figure 20 -
The number 6 also makes a very strong presence in a rather classic theorem named for the French mathematician Blaise Pascal (1623-1662).

Pág. 44(53) - Figure 33 -
In 1765, Leonhard Euler showed that six of these points, the midpoints of the sides and the feet of the altitudes, determine a unique circle

20220703

1) 20220703/

1.1) Abstract Algebra An Introduction.../
10.1515_9783110691160.pdf

Pág. 144(160) - Chapter 8

Vector Spaces
→ and Subspaces.

1.2) The Canonical operator in Many-Particle.../
10.1515_9783110762709.pdf

1.3) Einstein/

1.3.1) Einstein-and-the-philosophical problems...pdf
→ 520 pages.

Pág. 128(132) - Poincaré

Thus, although Poincaré understood the whole complexity of defining simultaneity, he assumed that such a definition could be introduced in an arbitrary manner having no objective bases but being rather the

result of "unconscious convention".
Obs.. Livro todo!!!

1.3.2) Kuznetsov-Einstein.pdf

Pág. 162 (168) - Chapter 15

Space, Time, Energy, and Mass.

"Applied to Maxwell's equations, the principle of relativity requires, in particular, that mass be a direct measure of a body's energy.
Light carries mass.
A fascinating and intriguing idea.
I only wonder if God might not just be jesting and making a fool of me"

Albert Einstein.

Deus não faz ninguém de bobo, pois o homem já é bobo por natureza, pois não sabe o que faz e ignora de tudo, o homem se acha muito inteligente e esperto mas na realidade o homem somente é alguma coisa com a graça de Deus.

1.3.3) Logunov... Relativistic Theory..pdf
→ 235 pages.

Obs.: Livro todo!!!

Pág. 38(42) - Chapter 6.

The Geometrization Principle and General RTG Relations.

→ tensor density $\tilde{g}^{ik}$ of the metric tensor of the Riemann space-time is a local function that depends on the density $\tilde{\gamma}^{ik}$ of the the metric tensor of the Minkowski space-time and the Φ^{ik} of the gravitational-field tensor.

Obs.: Livro todo!!!

1.3.4.) space-time-gravitation.pdf

Pág. 20(22) - 1.3

From Euclid to Lobachevsky.

→ 220 pages. Livro todo!!!

✓

1.3.5.) Special theory of Relativity by Ugarov.pdf

→ 410 pages.
→ Livro todo!!!

→

Lorentz transformation

Pág. 105 (106) - §3.7.

The K calculus (the radar method).

1.4) 20220703/
Remarkable curves.../

1.4.1.) lml-remarkable-curves.pdf

1.4.2.) straight-lines-and-curves.pdf
→ 206 pages.
→ Livro todo!!!
→ Interessante!!!

1.5.) 20220703/
Característica Euleriana.../
YM..Característica-euleriana.pdf
→ 116 pages.

Pág. 46 (48) - §6º. Fórmula de
Euler para el spacio.

1.6.) 20220703/
The Forces of Nature/
Grigoryev...The Forces of Nature...pdf
→ 358 pages.
→ Livro todo!!!

1.7.) 20220703/ ⇙
What is Distance.../
Shreider - What is Distance....pdf
→ 92 pages.
→ Livro todo!!!

1.8.) 20220703/ Linear Algebra.../
Linear Algebra and Multi
Dimensional....pdf
→ 500 pages.
→ Livro todo!!!

20220705///

WIMP - Partícula Massiva de
Interação Fraca.

Book:

DATA-Novo Vivo-014-2022/
Videos/
Important-Riemann-Mathematics/
Spacetime and Geometry An
Introduction to General Relativity.)
Sean Carroll - Spacetime and
Geometry... djvu
→ 526 pages
Pág. 332 (345) — 8. 43
Variables the Robertson-Walker metric...

Códigos para Mathematica :

off = não entram, são observações!

Robertson - Walker Metric

coordenadas U = {t, r, θ, φ}

$\underbrace{\theta, \phi}_{\text{off}} \rightarrow$ coordenadas angulares

n = Length[coordenadas U] $\xrightarrow[\text{off}]{}$ longitude.

g =

off = g = métrica.

off: $$ds^2 = -c^2 dt^2 + a^2(t)\left[\frac{dr^2}{1-kr^2/R_0^2} + r^2 d\Omega^2\right]$$

equation (8.43) - Pág. 332(345) - Carroll.

Esta equação métrica em forma matricial:

$$g_{\mu\nu} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{-a^2(t)}{1-Kr^2} & 0 & 0 \\ 0 & 0 & -a^2(t)r^2 & 0 \\ 0 & 0 & 0 & -a^2(t)r^2\sin^2\theta \end{bmatrix}$$

→ gDD =

off: D = down (índices)

{{-1, 0, 0, 0}, {0, $\frac{a[t]^2 \times r^2}{1-K\times r^2}$, 0, 0},

{0, 0, a[t]^2 × r^2, 0},

{0, 0, 0, a[t]^2 × r^2 × Sin[θ]^2}}

<enter>

Definição da métrica inversa:

```
gUU = Inverse[gDD]
```
<enter>

Escrevendo Christoffel Symbol

Base:

$$\Gamma^{\mu}_{\nu\lambda} = \frac{1}{2} g^{\mu\rho}(\partial_\nu g_{\rho\lambda} + \partial_\lambda g_{\rho\nu} - \partial_\rho g_{\nu\lambda}).$$

escrevendo:

→ obs: Índices que não se repetem. (parte final)

```
GammaUDD = 1/2 x Table[
Sum[
gUU[[μ, ρ]] x
(D[gDD[[ρ, λ]],
coordenadas U[[ν]]] +
D[gDD[[ρ, ν]],
coordenadas U[[λ]]] -
D[gDD[[ν, λ]], coordenadas U[[ρ]]]),
{ρ, n}], {μ, n}, {ν, n}, {λ, 4}]
```

↳ índices que não se repetem.

μ = mu
ρ = ro
λ = lambda
ν = nun

Testando: (Equation) teste com $\Gamma^0_{22} = a\dot{a}r^2$

(8.44) Livro

GammaUDD[[1,3,3]] off: <enter>

zero = 1
2 = 3
2 - 3

= r^2 a[t] a'[t] –

a' = derivada

Parte II:

Riemann Tensor

$$R^{\mu}{}_{\nu\lambda\rho} = \partial_\lambda \Gamma^{\mu}_{\nu\rho} - \partial \Gamma^{\mu}_{\nu\lambda} + \Gamma^{\mu}_{\sigma\lambda}\Gamma^{\sigma}_{\nu\rho} - \Gamma^{\mu}_{\sigma\rho}\Gamma^{\sigma}_{\nu\lambda}$$

→ sigma

códigos:

RiemannUDDD =

Table[, {μ,n},{ν,n},{λ,n},{ρ,n}]

off: Table - índices que não se repetem

termo Riemann

D[GammaUDD[[μ,ν,ρ]],
coordenadasU[[λ]]] -

com quem estamos derivando

μ → mu
ν → nun
ρ → ro

D[GammaUDD[[μ,ν,λ]],
coordenadas[[ρ]]] +

índice λ
→ segundo termo

Escrevendo Terceiro Termo:

Sum[GammaUDD[[μ, σ, λ]] x

GammaUDD[[σ, ν, ρ]], {σ, n}]

−

Sum[GammaUDD[[μ, σ, ρ]] x

GammaUDD[[σ, ν, λ]], {σ, n}],

nesta parte aparecerá os índices que não se repetem

{μ, n}, {ν, n}, {λ, n}, {ρ, n}]

Testando um componente geral:

RiemannUDDD[[1, 2, 3, 4]]

Out = 0 (zero)

outro exemplo:

RiemannUDDD[[1, 2, 1, 2]]

Out = $\frac{r^2 a[t] a''[t]}{1 - K r^2}$

outra fonte:

20220705/Spacetime and Geometry.../
(diretório todo)

20220810

Verdade Matemática.

As afirmações matemáticas que sabemos demonstrar serem verdade recebem o nome de teorema.

São verdades incondicionais e eternas.

Importante. Trata-se de uma prerrogativa existente somente na matemática.

Verdade: (?) → até agora.

Dada uma circunferência qualquer, a razão entre o seu perímetro e o seu diâmetro é π.

Matemática → Circunferência Ideal.
(abstrato)

Circunferência é um objeto matemático.

Os teoremas sobre circunferências valem para a circunferência matemática que é bem definida.

Definição: Uma circunferência de centro P e raio r é o conjunto de pontos que distam r do ponto P.

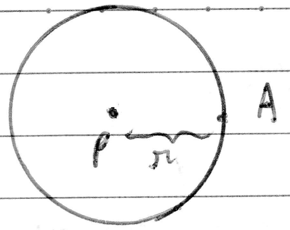

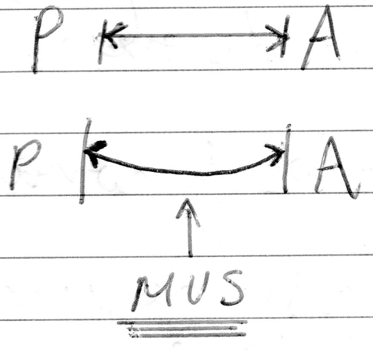

(ideal)

(real)

A circunferência matemática é perfeita (e abstrata).

Definições:

Em uma nova teoria defina todos os termos que você irá utilizar. (important)

Todo número inteiro, par, maior ou igual a 4, é a soma de dois números primos.

Conjunto dos números naturais (ou inteiros não-negativos)

$\mathbb{N} = \{0, 1, 2, 3, \dots\}$

↳ é um número natural.

0 é divisor de algum número inteiro?

Segundo a definição, para que 0 divida um inteiro b, é preciso existir um número inteiro x tal que $0 . x = b$.

↳ tem que ser 0 (zero)

Obs.:
x pode ser qualquer inteiro.

0 é divisor de 0 e também é verdade que o único inteiro que 0 divide é ele próprio.

Outras definições:

Definição 2 → Um inteiro a é par se $2|a$.

Definição 3 → Um inteiro a é ímpar se existe um inteiro k tal que

$a = 2.k + 1$.

Definição 4 → Um inteiro p é primo se $p > 1$ e se os únicos divisores positivos de p são 1 e p.
Obs.: 1 não é primo.

Afirmações:

- A soma de dois inteiros pares é um inteiro par.

- Se o produto de dois inteiros a e b for par, então a é par ou b é par.
  ↳ pode ser um número ou os dois.

F Se a e b são inteiros tais que $a|b$ e $b|a$ então $a = b$. (transitiva)
↳ Falsa. Prova da falsidade: $a = 2$ e $b = -2$.

- Todo número inteiro par, maior ou igual a 4, é a soma de dois números primos.

Obs: Foi testado por computador até

$$n \leq 4 . 10^{18}.$$

Todos os inteiros pares entre

4 e $4 . 10^{18}$ eram somas de dois primos.

Na verdade, não se sabe se é verdadeira ou falsa.
⇒ conjectura.

A soma de dois inteiros pares é um inteiro par ⇒ é um teorema
(afirmação verdadeira).

Se o produto de dois inteiros a e b foi par, então a é par ou b é par.
→ é um teorema (afirmação verdadeira).

Todo número inteiro par, maior ou igual a 4, é a soma de dois números primos
→ é uma conjectura, conhecida como Conjectura de Goldbach (1742).

Euclides (300 A.C.) fez uma demonstração de que existem infinitos números primos.
Considerada um belo exemplo de raciocínio lógico.

$$MUS(\pi) = \underset{\substack{\downarrow \\ \text{matemático} \\ \text{(conceito)}}}{\pi} \cdot \underbrace{G_L \cdot G_U}_{\underbrace{\text{bending of the light}}_{\text{real}}}$$

Obs: Importantíssimo!!!

"G_L" = Gravidade Local (Terra)

"G_U" = Gravidade do Universo (Universo)

$$MUS(\pi) = \pi \cdot GL \cdot GU$$

$\pi \downarrow$ conceito matemático

$GL \downarrow$ Gravidade Local

$GU \downarrow$ Gravidade Universo.

Three Fundamental Forces

1) Electromagnetism
2) Strong Nuclear Force
3) Weak Nuclear Force

$$GL = \underbrace{E + SNF + WNF}_{\text{Bending of light}}$$

$$GU = \text{Gauss's Curvature}\ (\text{Universe})(E + SNF + WNF)$$

Obs: GL and GU → são efeitos destas 3 forças fundamentais.

Important: π é uma constante fixa estipulada por cálculos

matemáticos que demonstrem que
este coeficiente não seja transcendental.

Ex.:

$3,14$

Podendo ser:

$3,13 \leq x \leq 3,14$

Obs.:

Certo dizer que o valor é menor que 3,14, pois temos os outros fatores como

GL e GU como elementos, fatores que alteram o π conceitual.

$3,14 \cdot 0,\dots \cdot 0,\dots =$

o resultado é obviamente menor.

Portanto: Teremos

$\Pi_L = \pi \cdot GL$

$\Pi_U = \pi \cdot GL \cdot GU$

GU interfere também GL (local)

ou mais precisamente:

$\Pi_L = \pi \cdot F_L$ Obs.: $F_L = TFF_L$ (Three Fund. Forces)

$\Pi_U = \pi \cdot F_L + F_U$ obs.: $F_U = TFF_U$ (Three Fund. Forces Universe)

(FL)

$$F_L = TFF_L = E_L + SNF_L + WNF_L$$

$$F_U = TFF_U = E_U + SNF_U + WNF_U$$

Important:

$$\Pi_L = \Pi \cdot F_L \cdot \underbrace{F_U}_{\text{valor menor}}$$

$$\Pi_U = \tilde{\Pi} \cdot F_L \cdot \underset{\sim}{F_U}$$

Valor maior pois abrange o Universo como um todo.
(Gauss's curvature)

Poderíamos dizer que o F_U localmente é pequeno, mas não pode ser desprezado.

Assim:

$$\tilde{\Pi} = \underbrace{\underbrace{\tilde{\Pi} \cdot F_L}_{\text{LOCAL}} \cdot F_U}_{\text{Universal.}}$$

$$\Big\downarrow$$

$$\Pi < \tilde{\Pi}.$$

202208102359.

20220818 M Volume-021--20220818/original/001---GCTprint.pdf

Revisando as minhas anotações:
(Estive ocupado fazendo pesquisas e escrevendo emails para outros pesquisadores)

1) ZIP-03/Important-DATA-013-e-014---20220709/01/

GCTprint.pdf (001---GCTprint.pdf)

Pág. 154 (164) - Figure 7.3.2.

The angle of parallelism θ of a point z to a line L.

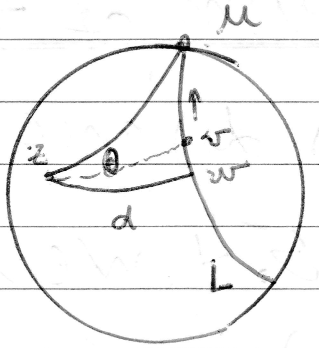

Figure 7.3.2.

Observing negative curvature:

Suppose we are located at a point z in a hyperbolic universe with curvature K.
We see in the distance a hyperbolic line L that seems to extend indefinitely.
We might intuitively see the point w (w)

on the line that is closest to us, as suggested in Figure 7.3.2, p. 154.

Now suppose we look down the road a bit to a point v.

If v is close to w the angle $\angle wzv$ will be close to 0.

As v gets further and further away from w, the angle will grow, getting closer and closer to the angle $\theta = \angle wzu$, where u is an ideal point of the line L.

→ A curious fact about hyperbolic geometry is that this angle θ, which is called the

angle of parallelism of z to L,

is a function of z's distance d to L.

Obs.: In Section 5.4 we saw that $\cosh(d) = 1/\sin\theta)$ in $(\mathbb{D}, \mathcal{H})$.

Section 5.4 - Area and Triangle Trigonometry

→ Pág. 100 (110)

Análise: Pág. 103 (113) →

Theorem 5.4.9

The area of a hyperbolic triangle in $(\mathbb{D}, \mathcal{H})$ having interior angles α, β, and γ is

$$A = \pi - (\alpha + \beta + \gamma).$$

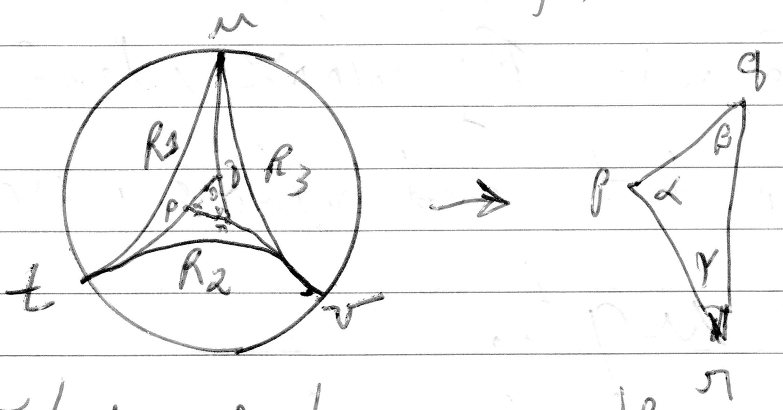

Figure 5.4.10. Determining the area of a hyperbolic triangle.)

Obs: IMPORTANT:

A área está errada, pois o π is wrong.

Vejamos: Δpqr

We have extended segment qp to the ideal point t, u is an ideal point of line rq, and v is an ideal point of line pr. The area of the ideal triangle Δtuv is π.

→ wrong!!!

(continuando o estudo). →

Notice that regions R_1, R_2, and R_3 are all $\frac{2}{3}$ - ideal triangles contained within the ideal triangle.

Consider R_1, whose ideal points are u and t, and whose interior angle is $\angle uqt$.

Since the line through q and r has ideal point u, the interior angle of R_1 is

$\angle uqt = \pi - \beta$.

Similarly, R_2 has interior angle $\pi - \alpha$ and

R_3 has interior angle $\pi - \gamma$.

Let R denote the triangle region Δpqr whose area $A(R)$ we want to compute.

Vejamos:

We then have the following relationships among areas:

$$\pi = A(R) + A(R_1) + A(R_2) + A(R_3)$$

$$= A(R) + [\pi - (\pi - \beta)] + [\pi - (\pi - \alpha)] + [\pi - (\pi - \gamma)].$$

Solving for $A(R)$,

$$A(R) = \pi - (\alpha + \beta + \gamma),$$

and this completes the proof.

Voltando:

Pág. 154(164) –

$$\cosh(d) = 1/\sin(\theta) \text{ in } (\mathbb{D}, \mathcal{H}_L).$$

In particular, one may deduce the distance d to the line L by computing θ.

Obs: In a Euclidean geometry, if one looks farther and farther down (mais para baixo) the line L, the angle will approach 90°, no matter one's distance d from the line.

The following theorem provides another formula relating the angle of parallelism to a point's distance to a line.

Theorem 7.3.3

Lobatchevsky's formula

In hyperbolic geometry with curvature k, the hyperbolic distance d of a point z to a hyperbolic line L is related to the angle of parallelism θ by the formula

$$\tan(\theta/2) = e^{-\sqrt{|k|}\,d}.$$

IMPORTANT:

⇘ A prova deste teorema deverá ser adequada à nova geometria baseada III (M.U.S.).

PROOF: For this proof, let

$$s = \frac{1}{\sqrt{|k|}}.$$

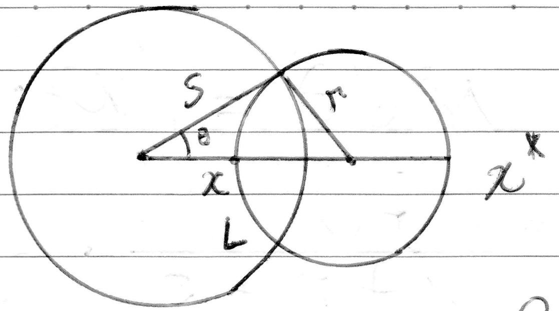

Figure 7.3.4 Deriving Lobatchevsky's formula.

Note that s is the Euclidean radius of the circle at infinite (infinite) in the disk model for hyperbolic geometry with curvature K.

Since angles and lines and distances are preserved, assume z is the origin and L is orthogonal to the positive real axis, intersecting it at the point x
(with $0 < x < s$).

Recall the half-angle formula

$$\tan(\theta/2) = \frac{\tan(\theta)}{\sec(\theta) + 1}.$$

According to Figure 7.3.4, p.154, $\tan(\theta) = \frac{r}{s}$, where r is the

Euclidean radius of the circle containing the hyperbolic line L.

Furthermore, $\sec(\theta) = (x+r)/s$, so

$$\tan(\theta/2) = \frac{r}{x+r+s}. \qquad (1)$$

We may express x and r in terms of the hyperbolic distance d from 0 to x.

In Exercise 7.3.2, we prove the hyperbolic distance from 0 to x in $(\mathbb{D}_K, \mathcal{H}_K)$ is

$$d = s \ln\left(\frac{s+x}{s-x}\right) \text{ so that}$$

$$x = s \cdot \frac{e^{d/s}-1}{e^{d/s}+1}.$$

Express r in terms of d by first expressing it in terms of x.

Note that segment $x\,x^*$ is a diameter of the circle containing L, where $x^* = \frac{-1}{Kx}$ is the point symmetric to x with respect to the circle at infinity.

→

Thus, r is half the distance from x to x^*!

$$r = \frac{1 + \kappa x^2}{2\kappa x}.$$

Replacing κ with $-1/s^2$, we have

$$r = \frac{s^2 + x^2}{2x}.$$

One checks that after writing x in terms of d, r is given by

$$r = \frac{2 s e^{d/s}}{e^{2d/s} - 1}.$$

Substitute this expression for r into the equation labelled (1) in this proof, and after, we are

$$\tan(\theta/2) = e^{-d/s}.$$

Since $s = 1/\sqrt{|\kappa|}$ this completes the proof.

INTERESSANTE:

PARALLAX (prestar atenção!!!)

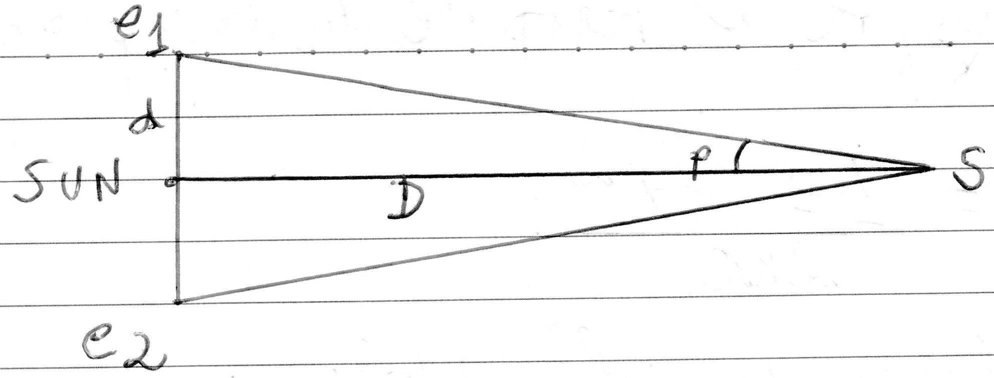

e_1 and e_2 denote the Earth's position points of its orbit, and the star S is orthogonal to the plane of the Earth's orbit.

The angle p is called the parallax, and in a Euclidean universe, p determines the star's distance from the Sun, D, by the equation

$D = d / \tan(p)$, where d is the Earth's distance from the Sun.

Obs.: d is about 8.3 light-minutes.

In practice, p is quite small, so a working formula is

$$D = \frac{d}{p}.$$

The first accurate measurement of parallax was recorded in 1837 by Friedrich Bessel (1784-1846).

He found the stellar parallax of 0.3 arc-seconds (1 arc second = 1/3600°) for star 61 Cygni, which put the star at about 10.5 light-years away.

Interessante:

If we live in a hyperbolic universe with curvature K, a detected parallax puts a bound on how curved the universe can be.

Obs.: $2d \cong 16.6$ light-minutes.

Assume star s is on the positive real axis and we have detected a parallax p, so that angle

$$\angle e_2 s e_1 = 2p.$$

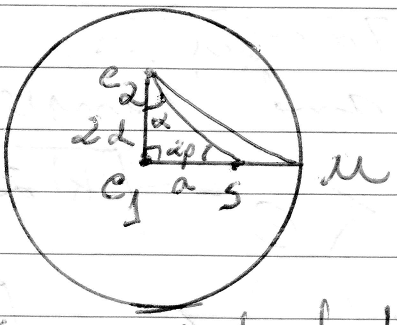

Figure 7.3.5 A detected parallax in a hyperbolic universe puts a bound on its curvature

→

The angle $\alpha = \angle e_1 e_2 s$ in Figure 7.3.5, p. 156 is less than the angle of parallelism

$$\theta = \angle e_1 e_2 u.$$

Noting that $\tan(x)$ is an increasing function and applying Lobatchevsky's formula it follows that

$$\tan(\alpha/2) < \tan(\theta/2) = e^{-\sqrt{|k|2d}}.$$

We may solve this inequality for $|k|$:

Off: $$e^{\text{In}\,x} = x$$

In = natural logarithm
e = natural exponent
x = real number

$e \cong 2.718281828459\ldots$
(irrational and transcendental number).

$$\tan(\alpha/2) < e^{-\sqrt{|k|2d}}$$

$$\text{In}(\tan(\alpha/2)) < -\sqrt{|k|2d}$$

In(x) is increasing

$$\left[\frac{\ln(\tan(\alpha/2))}{2d}\right]^2 > |K|.$$

x^2 is decreasing for $x < 0$.

To get a bound for $|K|$ in terms of p, note that $\alpha \approx \pi/2 - 2p$ (the triangles used in stellar parallax have no detectable angular deviation from 180°), so

$$|K| < \left[\frac{\ln(\tan(\pi/4 - p))}{2d}\right]^2.$$

We remark that for values of p near 0, the expression $\ln(\tan(\pi/4 - p))$ has linear approximation equal to $-2p$, so a working bound for K, which appeared in Schwarzschild's 1900 paper [27], is $|K| < (p/d)^2$.

↳ Schwarzschild, K.
On the Permissible Curvature of Space.
Vierteljahrsschrift d. Astronom. Gesellschaft, (1900), 35, 337-47.

Obs: Aplicar M.U.S (π).

Pág 158(168) – 7.4

The Family of Geometries (X_K, G_K)

A strong connection exists between the family $(\mathbb{P}^2_K, S_K)$ of elliptic geometries with curvature $K > 0$ and the family $(\mathbb{D}_K, \mathcal{H}_K)$ of hyperbolic geometries with curvature $K < 0$.

The two families sport identical descriptions of the transformation group, identical descriptions of straight lines, identical arc-length and area formulas, as well as identical formulas for the area of a triangle.

Definition 7.4.1

For each real number K the geometry (X_K, G_K) has space

$$X_K = \begin{cases} \mathbb{D}_K & \text{if } K < 0; \\ \mathbb{C} & \text{if } K = 0; \\ \mathbb{P}^2_K & \text{if } K > 0, \end{cases}$$

and group of transformations G_K

consisting of all Möbius transformations of the form

$$T(z) = e^{i\theta}\,\frac{z - z_0}{1 + K\bar{z}_0 z},$$

where $\theta \in \mathbb{R}$ and z_0 is a point in X_K. Moreover, the unique line through two points p and q in (X_K, G_K) is the unique cline through the points p, q, and $-1/(K\bar{p})$.

A smooth curve $r : [a, b] \to X_K$ has arc-length given by

$$L(r) = \int_a^b \frac{2|r'(t)|}{1 + K|r(t)|^2}\,dt.$$

The area of a polar region R (R) in (X_K, G_K) is given by

$$A(R) = \iint_R \frac{4r}{(1 + Kr^2)^2}\,dr\,d\theta.$$

obs: If $K > 0$, the sum of the angles of a triangle must be greater than π; and

If $K < 0$ the sum of the angles

of a triangle must be less than π.

If $K > 0$ the space $X_K = P^2_K$ has finite area;

If $K < 0$, $X_K = \mathbb{D}_K$ has infinite area.

If $K > 0$ the circumference of a circle with radius r is less than $2\pi r$;

if $K < 0$ the circumference is greater than $2\pi r$.

Obs: $K = 0$

(X_0, g_0) corresponds to Euclidean geometry though (no entanto) with a scaled metric.

In particular, lines in (X_0, g_0) correspond to Euclidean lines (since $K = 0$, $-1/(K\bar{p}) = \infty$, so the unique line through p and q is the Euclidean line), and when $K = 0$ the arc-length is simply twice the usual Euclidean arc-length.

So while we are scaling distances in (X_0, G_0), Euclidean geometry applies:

triangles are Euclidean triangles and have angle sum equal to 180°.

Triangles with a right angle are Euclidean right triangles and satisfy the Pythagorean theorem.

We treat (X_K, G_K) as one big family of geometries.

The sign of K dictates the type of geometry we have, and the

magnitude of K dictates the radius of the disk in which we model the geometry
(unless (a não ser que)
$K=0$ in which case the
space is $\mathbb{C}$).

Moreover, Euclidean geometry (X_0, G_0) marks the edge from which we move into a hyperbolic world if K drops below 0, and into an elliptic world if K rises above 0.

We now summarize some results established in the previous section and emphasize key features common to all (X_K, G_K).

Arc-length is an invariant function of $(X_K\ G_K)$ and that the arc-length ~~that~~ ensures that the shortest path from p to q in (X_K, G_K) is along the line between them.

We have discussed these facts in the cases $K = -1, 0, 1$, and the result holds for arbitrary K. So the arc-length formula provides a metric on (X_K, G_K):

Given $p, q \in X_K$, we define

$d_K(p,q)$ to be the length of the shortest path from p to q.

The circle in (X_K, G_K) centered at p through q consists of all points in X_K whose distance from p equals $d_K(p,q)$.

Veja:

Pág. 147(157) - Figure 7.1.3

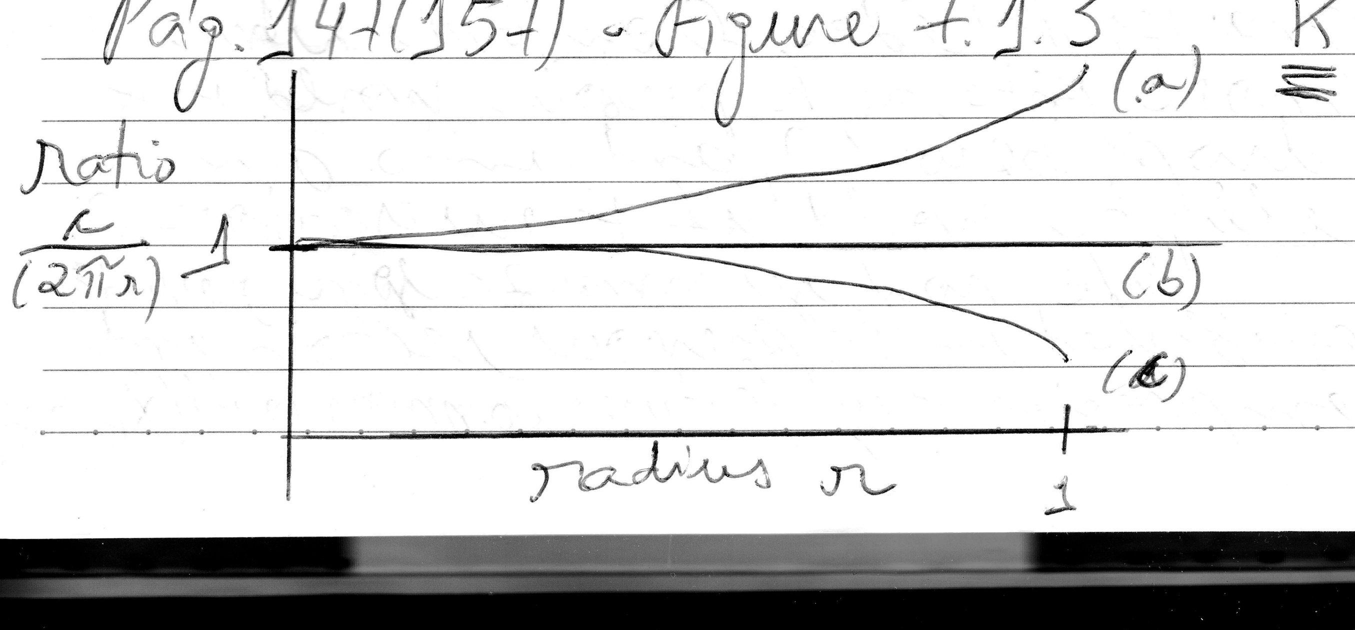

Figure 7.13 - Plotting the ratio
$c/2\pi r$ against r in
(a) hyperbolic geometry,
(b) Euclidean geometry, and
(c) elliptic geometry.

2) ZIP-03/Important-DATA-013-e-014-1--20220709/Einstein

2.1) Gravitation, Gauge Theories...pdf
(Volume-021--/20220818/original/002-Gravitation...pdf)
Pág. 15(27) -

Pág. 14(26) - Even in Euclidean geometry, we did not add two vectors at different points, but we have almost forgotten this

Pág. 15(27) - this disability of ours because, in order to add a vector $\bar{A}$ at P and a vector $\bar{B}$ at Q, we could always replace $\bar{B}$ at Q by a parallel vector at P and then add the two (say, by parallelogram rule).

⇒

IMPORTANT: In Riemannian Geometry,
we do not have a parallel vector at P because of the absense (ausente) of a parallel postulate,

↓

⟹ Criar um novo postulado baseado na nova geometria TTT (M.V.S.)

and so we cannot afford to forget this disability.

14. Covariant Derivative of a Contravariant Vector

... "We cannot do this because we do not have a parallel postulate which would have enable us to transfer a vector A^a at P to a parallel and equal vector at Q.

Pag. 26 (38) – Einstein postulated his law off gravitation as $G^{ab} = -KT^{ab}$

K is a constant and
T^{ab} is the matter enery tensor.

Estudar:

On comparing his equations with Newtonian equations of gravity, he found that if the gravitational field is weak enough and if the velocities involved are much smaller compared to the velocity of light, then to that order of approximation, his theory gives the exact Newton Poisson equation, provided

$$\boxed{K = + \frac{8\pi G}{c^2}}$$

Thus, if there are strong gravitational fields and if high velocities are involved, one must use the full Einstein equations

$$R^{ab} - \frac{1}{2} g^{ab} R = - \frac{8\pi G}{c^2} T^{ab}.$$

In space devoid of matter (see).
$T^{ab} = 0$ and we get

$$R^{ab} - \frac{1}{2} g^{ab} R = 0 ; \quad R^{ab} = \frac{1}{2} g^{ab} R ,$$

→

$$g_{ab} R^{ab} = \frac{1}{2} g_{ab} g^{ab} R.$$

$R = 2R$ ou $R = 0$,

therefore $R^{ab} = 0$.

Einstein's equations of gravitation in empty space satisfy

$$R_{ab} = 0.$$ (20220819)

Pág. 110 (120) – 4. Killing Vector Fields.

We have mentioned symmetry transformations for a spacetime.

Such symmetries are characterized by the existence of the so-calling

Killing vector fields in M.

Pág. 111 (121) – Suppose now that the local one-parameter group of diffeomorphisms ϕ_t generated by a vector field K is a group of isometries,

i.e. for every t, ϕ_t is an isometry,

$$\phi_{t*} g_{ab} = g_{ab} \text{ for all } t, \qquad (18)$$

then K is called a Killing vector field and we have

$$L_K g = \lim_{t \to 0} \frac{1}{t} [g - (\phi_{-t})_* g] = 0 \qquad (19)$$

Thus, the necessary and sufficient condition for $\underline{\underline{K}}$ to be a KILLING VECTOR FIELD is

that the Lie derivative of g along K vanishes.

Next, since

$$L_K g_{ab} = K_{b;a} + K_{a;b}, \qquad (20)$$

for a Killing vector field we get

$$K_{b;a} + K_{a;b} = 0 \qquad (21)$$

which is called the KILLING EQUATION.

Pág. 112(122) - Riemann tensor [8] [24]

(3) Killing vector field in a spacetime

↳ Next, consider the following equation defining the Riemann tensor [8]

→

METRIC → ∇ ∇

$$\nabla_a \nabla_b \xi_c - \nabla_b \nabla_a \xi_c = R_{abc}{}^{d} \xi_d . \quad (24)$$

If now ξ^a, appearing in the above, is a KILLING VECTOR FIELD, then using the KILLING EQUATION (23) and the symmetries of the Riemann tensor (24), gives

$$\nabla_a \nabla_b \xi_c = -R_{bca}{}^{d} \xi_d \quad (25)$$

↘ Interessante:

As a consequence of (25) a Killing vector field ξ^a is completely determined by values

ξ^a and $L_{ab} = \nabla_a \xi_b$ at any $p \in M$. this means that given (ξ^a, L_{ab}) at p, the values (ξ^a, L_{ab}) at any other q are determined by integrating the ordinary differential equations (26) ⇒

$$\left.\begin{aligned} v^a \nabla_a \xi_b &= v^a L_{ab}, \\ v^a \nabla_a L_{bc} &= -R^d{}_{bca}\, \xi_d\, v^a \end{aligned}\right\} \qquad (26)$$

along any curve joining p to q,

where v^a denotes the tangent vector to that curve.

Pág. 119 (128) - 8 - Differential Forms and Einstein-Cartan Theory.

⇒ Lorentz metric with signature -2..

Pág. 297 (305) - 15 — Quantum Field Theory in Curved Spacetime: Canonical Quantization.

The Planck dimensions represent scales on which the quantum effects of the gravitational field become important. For a system, say dust confined to a length scale L,

$$A_G \sim \frac{c^4}{16\pi G} \cdot \frac{8\pi G\rho}{c^4} \cdot \frac{4\pi L^3}{3} \cdot \frac{L}{c} = \frac{2\pi\rho L^4 c}{3}.$$

But

$$L > \frac{2G}{c^2}\left(\frac{4}{3}\pi L^3 \rho\right) \Rightarrow \rho L^2 \leq \frac{3c^2}{8\pi G}.$$

Consequently, Pág. 238(306)

$$A_G \leq \frac{2\pi L^2 c}{3} \cdot \frac{3c^2}{8\pi G} = \frac{L^2 c^3}{4G},$$

so that quantum effects of the gravitational field are significant, if

$$\frac{L^2 c^3}{4G} \leq \hbar, \text{ i.e. } L \leq 2L_p.$$

obs: QFT in CST →
quantum field theory in curved spacetime.

value of the energy momentum tensor for matter, i.e.

$$G_{\mu\nu}(g_{\alpha\beta}) = 8\pi \langle T_{\mu\nu} \rangle \quad (2).$$

There is a problem with the above semi-classical Einstein equation that gives the back reaction of the quantum field on the gravitational field.

According the equivalence principle which lies at the foundation of all metric theories of gravity all forms of matter and energy couple equally to gravity.

Pág. 405 (411) - 19 - The Photon, the Graviton and the Gravitino.

Pág. 407 (413) - 2. The Graviton

The existence of the Graviton field can be seen by carrying out a straightforward flat space expansion of the Einstein-Hilbert Lagrangian density,

$$S = \int \mathcal{L}_{E-H}\, d^4x, \quad \mathcal{L}_{E-H} = \sqrt{-g}\, R, \quad (17)$$

where g is the determinant of the metric tensor, $g_{\mu\nu}$ and R is the scalar curvature.

2.2) ZIP-03/ Important-DATA-013-e-014---20220709/
Einstein/

Prokhovnik - Light in Einstein's Universe - The Role of Energy in Cosmology....pdf

Obs: Veja diretório: Gravitation.

3) Volume-021---20220818/ Original/
003--- Moller/

→ Referenciando as várias cópias com correlação das páginas.

Diretório:

1) The theory of relativity---MOLLER/

The theory of relativity....(144).png

→ refere-se à página 274 do livro

Fig. 17

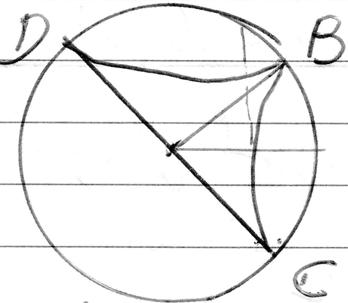

2) Theory-of-Relativity-Moller---20220701/

the theory of Relativity....(287).png

3) Theory-of-Relativity-Moller-important/

Livros:

a) C Moller-Theory of Relativity..(1955).pdf

Pág. 242 (252) - Fig. 17

Obs.: In Fig. 17 we have given a picture in a Euclidean plane of some of the geodesics on the rotating disk.

b) theory of relativi 029229 mbp.pdf

Pág. 242 (260) - Fig. 17

Pág. 210 (228) de 409 pages or
Pág. 210 (220) de 397 pages

(109'') → equation

$$\mu^{*'} = \frac{c}{n} + 2(e\ v')\left(1 - \frac{1}{n^2}\right) \quad (7.109'')$$

Obs: Pág. 238 (imagem) – 20220701

The Theory of relativity Møller....(238).png
obs: No livro pág. 224.

Further, we see from (7.109) Ⓐ

(7.109)

With Abraham's expression for the energy-momentum tensor we get for the velocity of the electromagnetic energy in the system S'

$$\mu^{*'} = \frac{c}{\sqrt{\epsilon\mu}} e + v' - \frac{(v' \cdot e)e}{\epsilon\mu} + (v' \cdot e)\left(1 - \frac{1}{\epsilon\mu}\right)e,$$

which deviates in the last term from the ray velocity as defined by Huyghens's principle.

Ⓐ cont. Further, we see from (7.109) that the velocity of the energy differs from the phase velocity if v' and e

Interessante:

V' and e are parallel,

in contrast with the ray velocity, which is identical in this case, with the phase velocity.

This strange result is connected with the following circumstance.

Obs: Aprofundar!!!

3.1) Book:

Theory of Shape Borsuk---20220620.pdf

Obs: Integrar às pesquisas.

⇒ Interessante!!!

4) ZIP-03/Important-DATA-013-e-014---20220709/
Gravitation/
Gravitation/ (Obs: 002---Gravitation...pdf)

Gravitation-Princeton University Press (2017)..pdf

Pag. 246 - Part III.
The Mathematics of Curved Spacetime.

5) Zip-03/ Books-Import-20220608/
La geometria non-euclidea/
noneuclideangeom00bonorich.pdf
→ Non-euclidean geometry -
Roberto Bonola (003-noneuclideangeom00bonorich.pdf)
Pág. 67(83) - §34 -

Gauss defines parallels as follows:
↳(Source: Gauss Werke, Bd VIII, p. 202-'9).

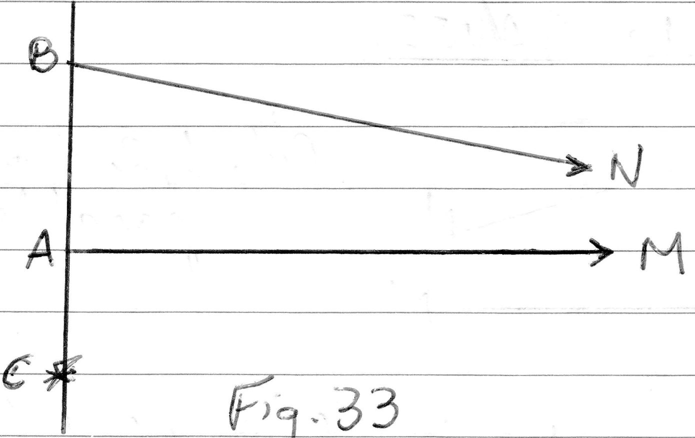

Fig. 33

If the coplanar straight lines AM, BN do not intersect each other, while,
every straight line through (através) A between AM and AB cuts BN, then AM is said to be parallel to BN [fig. 33].

He supposes a straight line passing through A, to start from the position AB, and then to rotate continuously on the side towards (em direção) which

BN is drawn, till it reaches (até atingir) the position A'C, in BA produced.

This line begins by cutting BN and in the end it does not cut it.

Thus there can be one and only one position, separating the lines which intersect BN from those which do not intersect it. (...)

Gauss's definition only the first of these is said to be parallel to BN.

Analisando GAUSS:

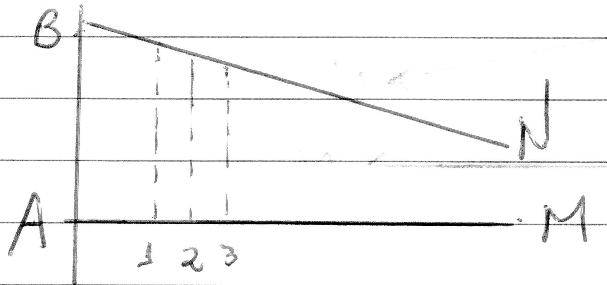

Obs. 1, 2, 3 are parallels

Portanto:

I. He proceeds to show that the parallelism of the line AM to the line BN is independent of the points A and B, provided the sense in which the lines are to be produced indefinitely remains the same.

It is obvious that we would obtain the same parallel AM

Pag. 69 (85)

if we kept A fixed and took instead (em vez de) of B another point B' on the line BN, or on that line produced backwards (para trás).

It remains to prove that if AM is parallel to BN for the point A ~~or on that one (line)~~) it also the parallel to BN any point upon AM, or upon AM produced backwards.

Instead of (término de careta)

Instead of A [Fig. 34] take another starting point A' upon AM. Through A', between A'B and A'M, draw the line A'P in any direction.

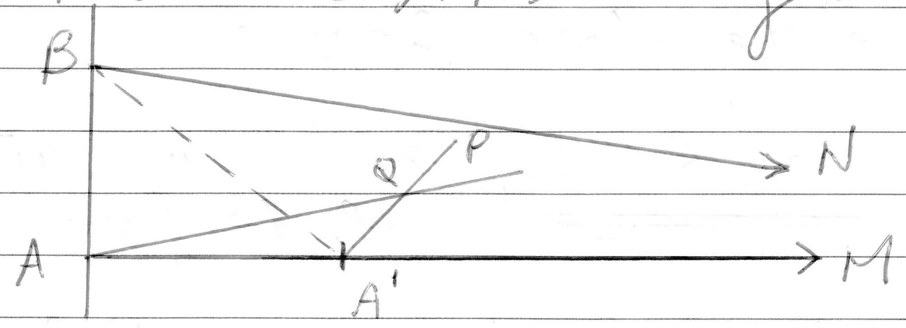

(Fig. 34)

Through Q, any point on A'P, between A' and P, draw the line AQ.

Then, from the definition, AQ must cut BN, so that it is clear QP must must also cut BN.

Thus AA'M is the first of the lines which do not cut BN, and A'M is parallel to BN. (checar!!!)

Again take the point A' upon AM produced backwards [Fig. 35]

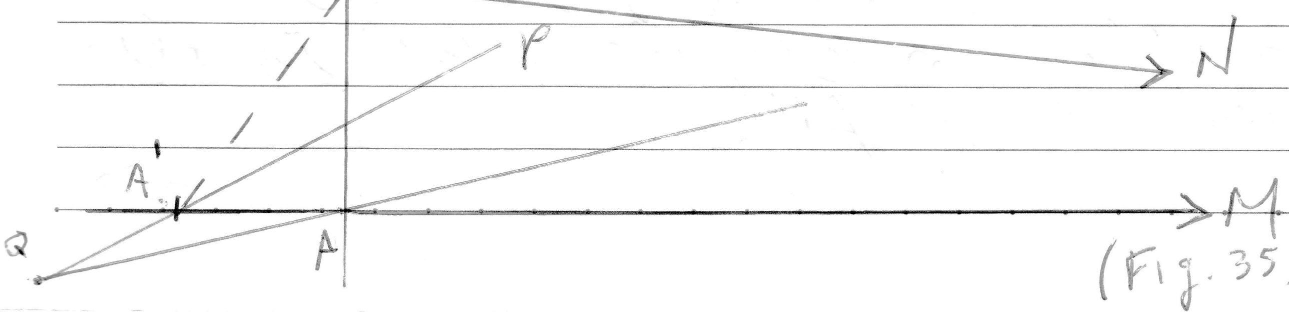

(Fig. 35)

Draw through A' between $A'B$ and $A'M$, the line $A'P$ in any direction.
Produce $A'P$ backwards and upon it take any point Q.
Then, by the definition, QA must cut BN, for example,

Pag. 70 (86) ... cont.

in R. Therefore $A'P$ lies (encontra-se) within the closed figure

$A'ARB$, and must cut one of the four sides

$A'A$, AR, RB, and BA'.

Obviously this must be the third side RB, and therefore

$A'M$ is parallel to BN.

II. The Reciprocity of the Parallelism can also be established.

In other words, if AM is parallel to BN, then BN is also parallel to AM!

⇒

GAUSS proves this results as follows:

From any point B upon BN draw BA perpendicular to AM.

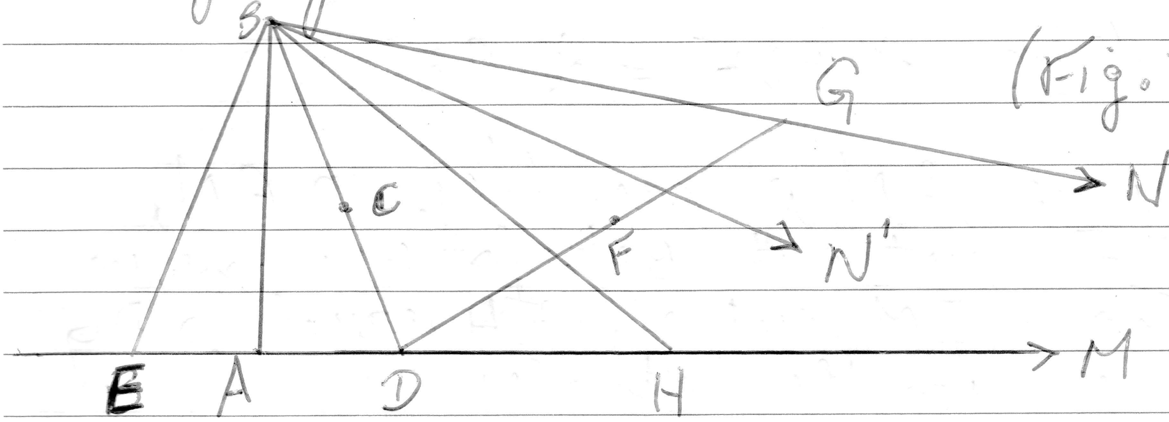

(Fig. 36)

At B, on the same side of AB as BN, make

$\sphericalangle ABG = 1/2 \sphericalangle N'BN.$

There are two possible cases:

Case (i), when BG cuts AM [cf. Fig. 36].

Case (ii), when BG does not cut AM [cf. Fig. 37].

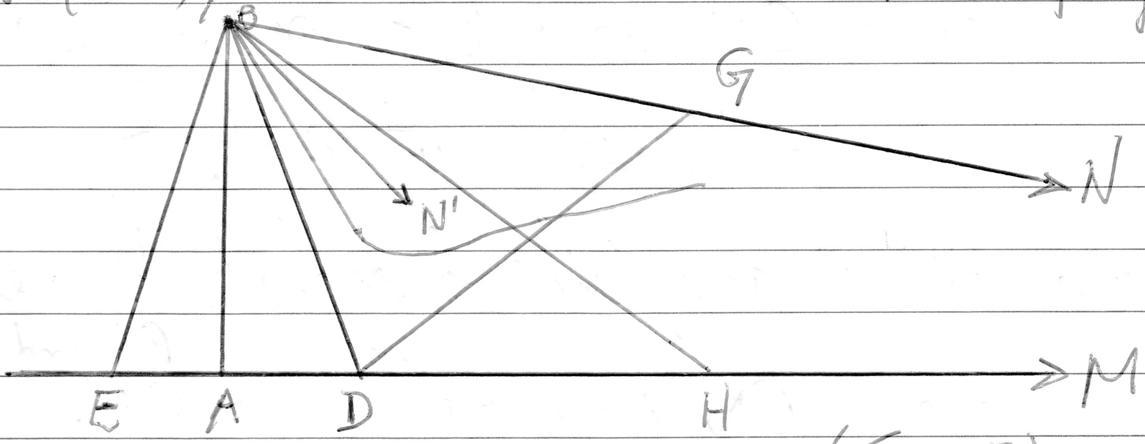

(Fig. 37)

Voltando: case (i)

Let BG cut AM in D.
Take AE = AD, and join BE.

Make $\measuredangle BDF = \measuredangle BED$.

Since AM is parallel to BN, DF must cut BN, for example, in G. From EM cut off EH equal to DG. Then, in the triangles

BEH and BDG, it follows that

Pág. 71(87)

$\measuredangle EBH = \measuredangle DBG$.

Therefore $\measuredangle EBD = \measuredangle HBG$.

But $\measuredangle EBD = \measuredangle N'BN$.

Therefore BN' and BH coincide, and BN' must cut AM.

But BN' is any line through B, between BA and BN.

Therefore BN is parallel to AM.

20220820 III

Case (ii) - In this case let D ~~that~~ be any arbitrary point upon AM.

Then with the same argument as above,

$\measuredangle EBD = \measuredangle GBH$.

But $\measuredangle ABD < \measuredangle ABC$.

Therefore $\measuredangle EBD < \measuredangle N'BN$.

Therefore $\measuredangle GBH < \measuredangle N'BN$.

Therefore BN'' must cut AM.

But BN' is any line through B, between BA and BN.

Therefore BN is parallel to AM.

Thus in both cases we have proved that if AM is parallel to BN, then BN is parallel to AM.(I)

(I) → Gauss's second proof of this theorem is given in the German translation. However it will be found that in it he assumes that BC cuts AM, and to prove this the argument (argument) used above is necessary.

The next theorem proved by Gauss in this synopsis is as follows:

III. If the line (1) is parallel to the line (2) and to line (3), then (2) and (3) are parallel to each other. (Pag. 72 (88)).

Case (1) - Let the line (1) lie between (deitar entre) (2) and (3) [cf. Fig. 38].

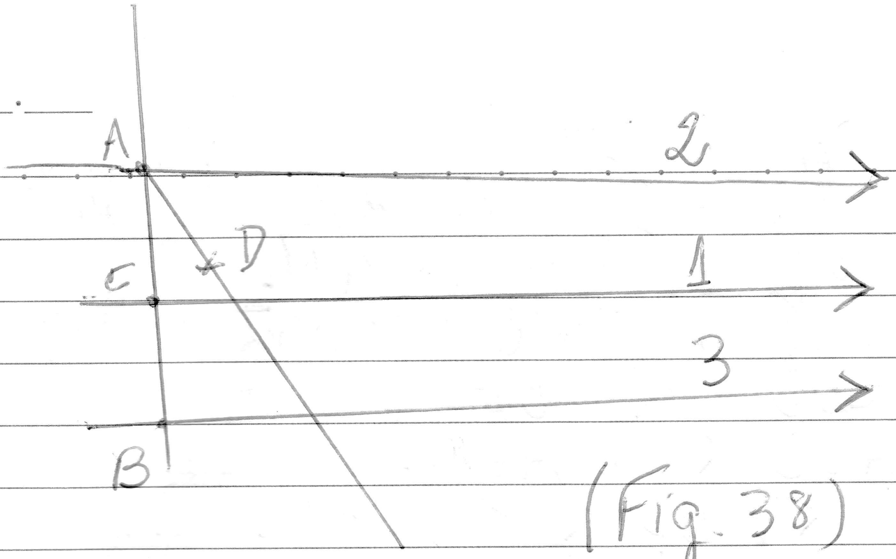

(Fig. 38)

Let A and B be two points on (2) and (3), and let AB cut (1) in C.

Through A let an arbitrary line AD be drawn between AB and (2).

then it must cut (1), and on being produced must also cut (3).

Since this holds for every line such as (tal como) AD, (2) is parallel to (3).

Case (ii) – Let the line (1) be outside both (2) and (3), and let (2) lie between (1) and (3) [cf. Fig. 39].

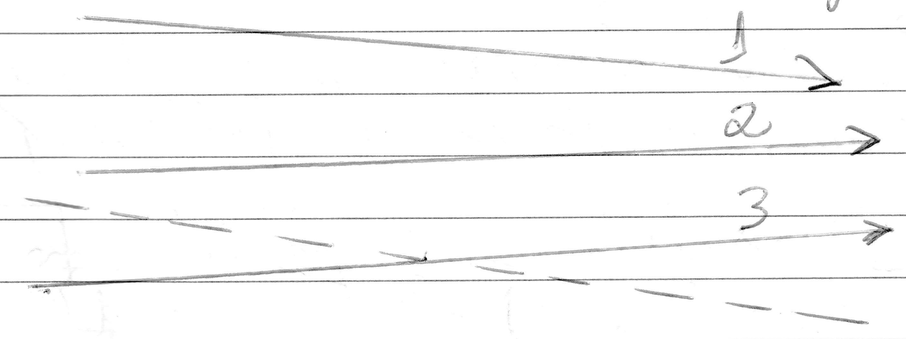

(Fig. 39)

If (2) is not parallel to (3), through any point chosen at random upon (3), a line different from (3) can be drawn which is parallel to (2).

This, by Case (1), is also parallel to (1), which is absurd.

This short Note on Parallels closes (closes) with the theorem that if two lines AM and BN are parallel, these lines produced backwards cannot meet.

Pág. 73 (89) - In a second memorandum on parallels, Gauss goes over (passa por cima) the same ground, but adds the idea of CORRESPONDING POINTS on two parallel AA', BB'.

⟹ Two points A, B are said to correspond, when AB makes equal internal angles with the parallels on the same side. (cf. Fig. 40)

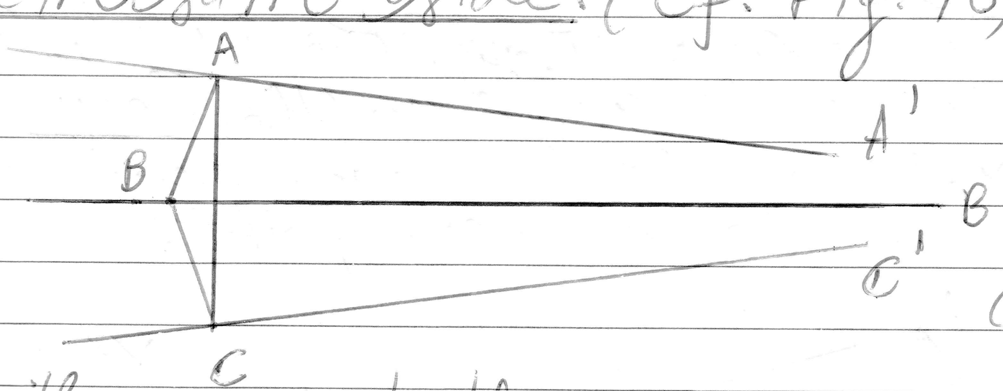

(Fig. 40)

With regards to these Corresponding Points he states the following theorems:

(i) If A, B are two corresponding points upon two parallels, and M is the middle point of AB, the line MN, perpendicular to AB, is parallel to the two given lines, and every point on the same side of MN as A is nearer A than B.

(ii) - If A, B are two corresponding points upon the parallel (1) and (2), and A', B' two other corresponding points on the same lines, then

AA' = BB', and conversely.
(por outro lado)

(iii) - If A, B, C are three points on the parallels (1), (2) and (3), such that (de tal modo que) A and B, B and C, correspond, then A and C also correspond.

Pag. 74 (90) — The idea of Corresponding Points, when taken in connection with three lines of a pencil (that is, three concurrent lines [cf. Fig. 41]) allows us to define the circle as the locus of the points on the lines of a pencil which correspond to a given point.

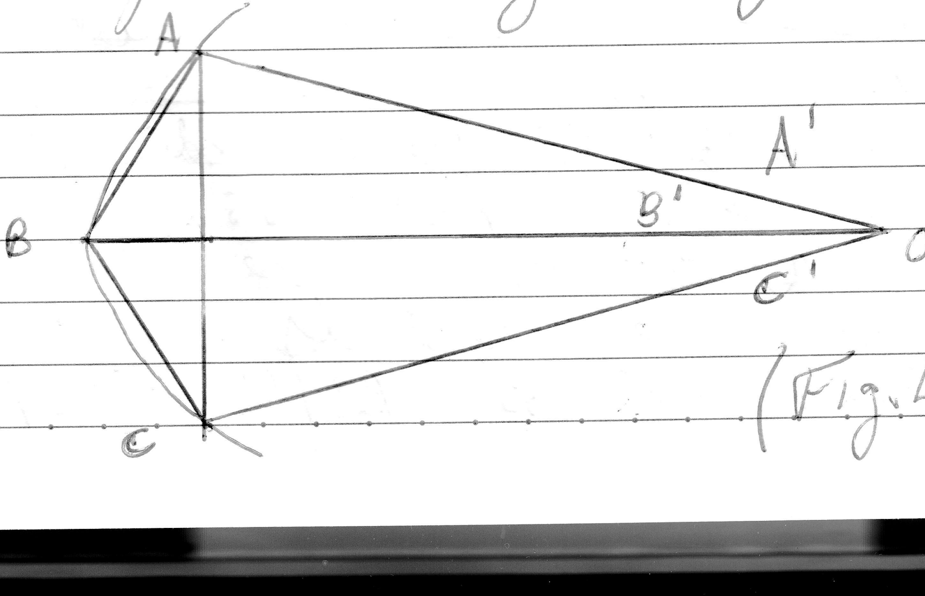

Important!!!

(Fig. 41)

But this locus can also be constructed when the lines of the pencil are parallel.

In the Euclidean case the locus is a straight line: but putting aside the Euclidean hypothesis, the locus in question is a line, having many properties in common with the CIRCLE, but yet not itself a circle.
(itself)

Indeed if any three points are taken upon it, a circle cannot be drawn through them. ✓

This line can be regarded as the limiting case of a circle, when its radius become infinite.

In the Non-Euclidean geometry of LOBATSCHEWSKY and BOLYAI: this locus plays a most important part, and we shall meet it there under the name of the HOROCYCLE. (I)

*

Obs.: Horocycle in hyperbolic geometry, or limit circle, is a curve whose normal or perpendicular geodesics all converge asymptotically in the same direction.

(I) [LOBATSCHEWSKY: Grenzkreis, Courbe-limite or Horicycle.

BOLYAI; Parazykl, L-linie.

→ TROPE (discussion of the properties of this curve)

This work GAUSS did not need to complete, for in 1832 he received from Wolfgang Bolyai a copy of the work of his son Johann on Absolute Geometry. Ⓐ

off: Absolute Geometry - Bolyai.

Other book: ISBN 978-613-5-77145-9

obs.: Absolute geometry is a geometry based on an axiom system that does not assume the parallel postulate or any of its alternatives.

The term was introduced by János Bolyai in 1832. It is sometimes referred to as neutral geometry, as it is neutral with respect to the parallel postulate. The theorems of absolute geometry hold in some non-Euclidean geometries, such as hyperbolic geometry, as well as in Euclidean geometry. Absolute geometry is an extension of ordered geometry, and thus, all theorems in ordered geometry hold in absolute geometry.

The converse is not true.

Absolute geometry assumes the first four of Euclid's axioms, to be contrasted with affine geometry, which assume Euclid's first, second, and fifth (parallel postulate) axioms.

Ordered geometry is a common foundations of both absolute and affine geometry.

Livro: Non-Euclidean Geometry, or
The Science of Absolute Space.
by John Bolyai.
→ Translated from the Latin by.
Dr. George Bruce Halsted.
Austin, Texas, the Neomon.

Source: https://www.science.org/doi/
pdf/10.1126/science.5.112.
311.

19 Feb 1897
Vol 5, Issue 112
pp. 311-312

Important:

The Science absolute of space 1891 →

Bolyai, János

Source:

https://jscholarship.library.jhu.edu/bitstream/handle/1774.2/34236/31151000225759.pdf

Obs.: 66 pages. (ref.: 004-31151000225759.pdf)

Voltando:

HOROCYCLE (*) Ⓐ

Pag. 74 (90) - From letters before and after the date at which he interrupted his work, we know that GAUSS has discovered in his geometry an

ABSOLUTE UNIT OF LENGTH

[cf. LAMBERT and LEGENDRE],

and that a constant K appeared in his formulae, by means of which all the problems of the Non-Euclidean Geometry could be solved [letter to Taurinus, Nov 8, 1824].

Speaking more fully of these matters in 1831 [letter to Schumacher], Pág. 75(90)— he gave the length of the circumference of a circle of radius r in the form

$$\pi k \left(\frac{r}{e^{k} - e} \quad \frac{r}{k} \right).$$

With regard to k, he says that, if we wish to make the new geometry agree with the facts of experience, we must suppose k infinitely great in comparison with all known measurements.

For $k = \infty$, GAUSS's expression takes the usual form for the perimeter of a circle.(I)

(I) To show this we need only use the exponential series.

The same remark holds for the whole of GAUSS's system of geometry. It contains EUCLID'S system as the limiting

case, when $k = \infty$.

Voltando um pouco mais:

Book: 003---noneuclideangeom00
bonorich.pdf
Pág. 67(83) — nota de rodapé

Gauss's work on Parallels

→ ref. GAUSS WERKE, Bd. VIII,
p. 202-9.

Livro de referência:

005---08-PPN236010751.pdf

Pág. 202(209) —

[ZUR THEORIE DER PARALLELLINIEN.]

até Pág. 209(216)

Estudar em conjunto com o Livro:

005---Pappus_dAlexandre_
La collection mathématique.pdf

Obs: Páginas - 1010.
1833

20220821$_{m}$

ZIP-03(ImportantMmo-03---20220712|

1) Geometry of Quantum States...pdf

Pag. 08(21) - 1.2 High-dimensional geometry

In quantum mechanics the spaces we encounter are often of very high dimension; even if the dimension of Hilbert space is small, the dimension of the space of density matrices will be high. ...

We can improve ourselves by asking some simple questions about convex bodies in flat space.

A flat metric is assumed.

Obs. We choose to looks at balls, cubes and simplices for this purpose.

Our questions will concern the

insphere and

outsphere

of theses bodies

(defined as the largest inscribed sphere and the smallest circumscribed sphere, respectively).

Look: Pág. 12(25) - Figure 1.8.

Regular simplices in two, three and four dimensions. For Δ_2 we also show the insphere, the outsphere, and the angle in the text.

For the simplex we obtain

$$\mathrm{vol}(\Delta_n) = \frac{1}{n!}\sqrt{\frac{n+1}{2^n}} . \qquad (1.25)$$

→

2) GeomMech1-2nd.pdf (007-...-GeomMec1-2nd.pdf)
Pag. 134(158) - 2.3. Hamilton.

Geometric optics may be approached either as a theory of systems of rays constructed by means of the elementary laws of refraction (Ibn Sahl, Snell, Descartes, Fermat, Newton), or as a theory based upon the consideration of systems of surfaces whose orthogonal trajectories are the rays (Huygens, Hamilton).

Pag. 135(159) - Hamilton optics, in which a characteristic function measures the time that light takes to travel from one point to another and it depends on the coordinates of both the initial and final points.

Pag. 136(160) - 2.3.1 Legendre Transform

One passes from Lagrangian to Hamiltonian dynamics through the Legendre transformation.

Voltando: (um pouco)

Pag. 112(136) - 2.2. Lagrange

solution to the isoperimetric problem

Among all closed curves of a given fixed perimeter in the plane, which curve maximizes the area that it encloses?

The circle does this.

Pag. 113(137) - Example 2.2.1

Stereographic projection of $S^2 \Rightarrow R^2$.

The unit sphere S^2 may be defined as a surface in R^3 given by the set of points satisfying

$$S^2 = \{(x, y, z) : x^2 + y^2 + z^2 = 1\}.$$

Important:

Utilize the M.U.S (TTT) para recalcular a equação acima pois nós devemos observar que o resultado obtido será diferente da apresentação clássica. A nova geometria irá revelar a real realidade do comportamento dos objetos dentro do Universo. O ajuste das equações será necessário para que a realidade seja representada de forma eficaz e produzirá os efeitos necessários para o real entendimento do mundo e do Universo.

Estudando:

$$\Pi\!\!\!\Pi = \pi . F_L . F_U \therefore \Pi\!\!\!\Pi < \pi$$

Vejamos:

$$S^2 = [\{(x,y,z) : x^2+y^2+z^2=1\} . \Pi\!\!\!\Pi]$$

or

$$S^2 = [\{(x,y,z) : x^2+y^2+z^2=1\} . \pi . F_L . F_U]$$

Important:

O π terá um valor fixado pois este número não representa a realidade a qual este número é usado para representá-lo.

π sendo um valor fixo, multiplicado por 2 coeficientes que alteram este número e para menor valor, nós teremos a real clareza do Universo que queremos descrever de forma mais precisa.

$\Pi\!\!\!\Pi < \pi$, o valor de $\Pi\!\!\!\Pi$.

representará a realidade do nosso mundo e o Universo.

The spherical polar angle θ and azimuthal angle ϕ are often used as coordinates on S^2. However, the angle ϕ cannot be defined uniquely at the north and south poles where $\theta = 0$ and $\theta = \pi$, respectively.

Important:

Checar que com o uso do π, nós obteremos outros referenciais esféricos. Ficar atento aos efeitos nas equações envolvidas.

RIEMANN'S treatment used PTOLEMY'S (Ptolemy's) STEREOGRAPHIC PROJECTION to define two overlapping subsets that satisfied the defining properties of a smooth manifold.

$$\text{Let } U_N = S^2 \setminus \{0,0,1\} \text{ and}$$

$$U_S = S^2 \setminus \{0,0,-1\}$$

be the subsets obtained by deleting the north and south poles of S^2 respectively. The stereographic projection

ϕ_N and ϕ_S from the north and south poles of the sphere onto the equatorial plane, $z = 0$, are defined respectively by

$$\phi_N : U_N \to \xi_N + i\eta_N = \frac{x+iy}{1-z} = e^{i\phi}\cot(\theta/2),$$

and

$$\phi_S : U_S \to \xi_S + i\eta_S = \frac{x-iy}{1+z} = e^{-i\phi}\tan(\theta/2).$$

The union of these two subsets covers S^2. On the overlap of their projections, the coordinates

$(\xi_N, \eta_N) \in \mathbb{R}^2$ and

$(\xi_S, \eta_S) \in \mathbb{R}^2$ are related by

$$(\xi_N + i\eta_N)(\xi_S + i\eta_S) = 1.$$

Important:

Nós devemos prestar a atenção de que o valor igual 1 representa uma projeção sobre um plano. Isto deverá acontecer posteriormente com o espelhamento (mirror) ortogonal sobre um plano desta projeção. Assim, o valor deverá ser menor que 1 pois o valor π (M.U.S.) gera uma real adequação da projeção sobre um plano.

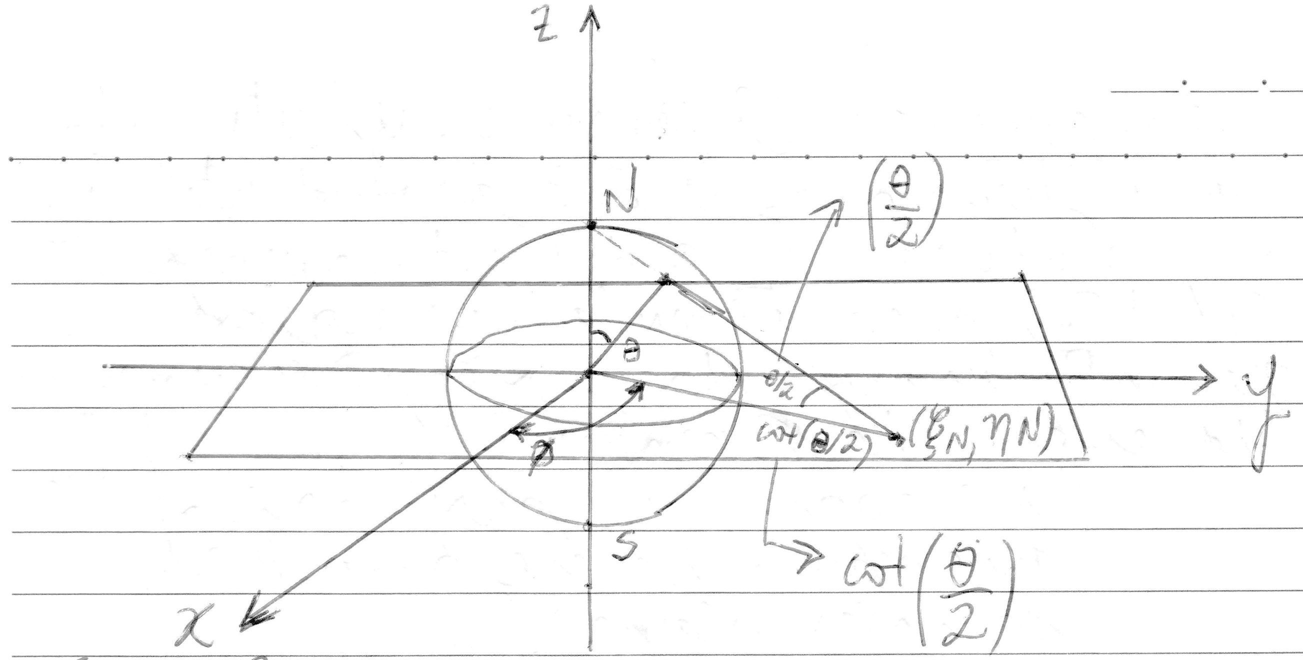
z
N
$\left(\frac{\theta}{2}\right)$
θ
$\theta/2$
$\cot(\theta/2)$
(ξ_N, η_N)
y
ϕ
S
$\cot\left(\frac{\theta}{2}\right)$
x

Figure 2.2 : In the stereographic projection of the RIEMANN sphere onto the complex plane from the north pole, complex numbers lying outside (resp. inside) the unit circle are projected from points in the upper (resp. lower) hemisphere.

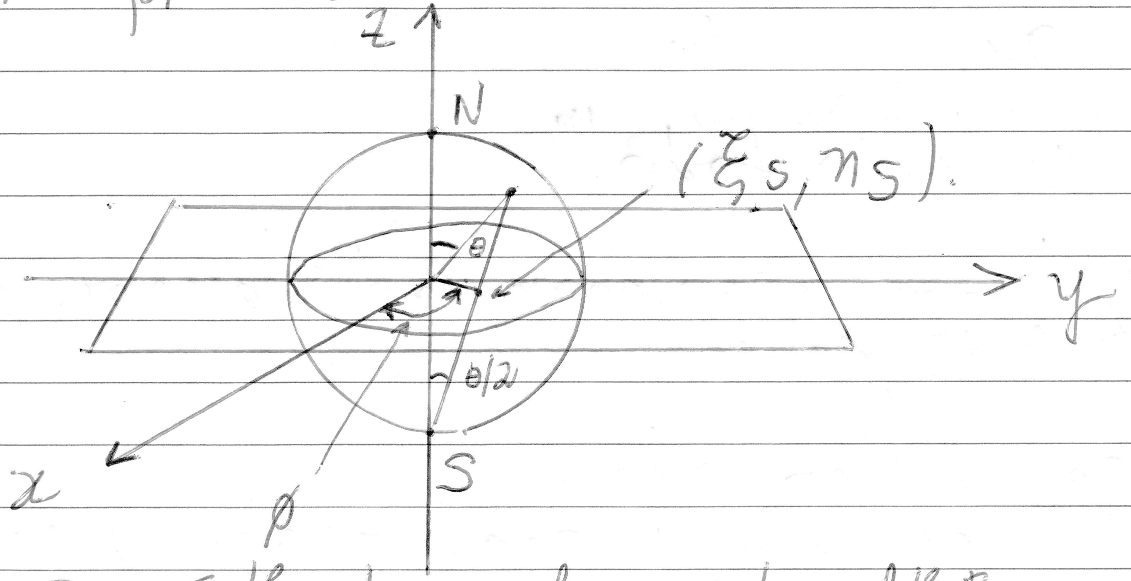
z
N
(ξ_S, η_S).
θ
y
$\theta/2$
x
S
ϕ

Figure 2.3. In the stereographic projection of the Riemann sphere onto the complex plane from the south pole, complex numbers lying outside (resp. inside) the unit circle are projected from points in the lower (resp. upper) hemisphere.

Important: O plano x, y utilizado nas Figuras 2.2 e 2.3 deverá utilizar a curvatura F_L and F_U da equação III (M.U.S.) para simular as novas reconfigurações da projeção estereográfica da esfera de Riemann. ✓

(stereographic projection of the Riemann sphere onto the complex plane).

Obs: Livro todo!!!

3) 007-GeomMec2-2nd.pdf

Galileo → Principle of Galilean relativity

Newton, Lagrange, Hamilton and the rigid body.

Obs: Livro todo!!!

4) 007--- Gravitation and Spacetime....pdf
→ 545 pages.

Obs: Livro todo!!!
Newton's gravitational theory

Voltando ao item 02 (007-GeomMec1-2nd.pdf)

Stereographic projection of the Riemann

Sphere - o nome do autor destes 2 volumes,
Prof. Darryl Holm - https://www.imperial.ac.uk/
people/d.holm - envie um email a ele dizendo:

a) Stereographic projection of Riemann Sphere -
Vol.1, 2nd Edition, p.115(139) - Figure 2.2 and
2.3 will need to be readjusted to a new reality.

b) In your publication, Solution properties of a
3D stochastic euler fluid equation
(https://spiral.imperial.ac.uk/handle/
10044/1/63498), you said on page 7
(Remark 32): "Euler fluid equations
treated here in Cartesian R3 coordinates,
the two velocities denoted u and v in the
previous section may be taken to be
identical vectors for the case at hand in
R3" (— R^3) there is a big problem in
calculating the vorticity (ω), because
the u and v value will need adjustment.
I discovered a new way to calculate pi
and I can say, for the moment, that this
number is not transcendental.
My research will have a direct impact
on your analyses.

→ Email enviado para:
d.holm@imperial.ac.uk

Arquivo armazenado:
Books-017-2022/20220821/
Geometric mechanics and symmetry
From finite to infinite dimensions/
Crisan 2018 - Article_Solution...pdf

Iniciei o arquivo: 008-- Crisan2018-
Solution Properties of A3D...pdf.

5) 008--- Geometric mechanics and
symmetry...pdf

Obs: 2009 - 537 pp.
Livro todo!!!

6) 007 - The Geometry of Special Relativity...
(007) pdf

→ 153 pp.

Obs: Time and Distance
→ Theorem of Minkowski.
Transformation
→ Lorentz Transformation of
Coordinates.
Pág. 50(60) - Fig. 3.1 - Lorentz
Transformation

7) 007 - The geometry of total curvature
on complete open surfaces -
Cambridge University Press
(2003).pdf

→ 296 pp.

Obs: Pág. 08(20) - 1.3
The Riemannian curvature tensor.
→ Livro todo!!!
Pág. 34(46) - 1.9.
The Gauss-Bonnet theorem
→ Curvature: $c(\underset{\text{ômega}\uparrow}{\Omega}) := \int_{\Omega} G \, dM$ (1.9.1). →

Voltando ao item 02 —
stereographic projection of the Riemann Sphere (Figure 2.2 and 2.3)
→ observamos o sinal ~
ξ_N, η_N and ξ_S, η_S,
realizar um estudo, comparando estes conceitos com outro em outro texto:

1) 008--- Crisan2018_Article_Solution Properties of A3D... pdf
(Prof. Holm — email)
Pág. 07 →

$$\mathcal{L}^2_{\xi_K}\omega = \mathcal{L}_{\xi_K}(\mathcal{L}_{\xi_K}\omega) = [\xi_K, [\xi_K, \omega]],$$

for the double Lie bracket of the divergence-free vector field ξ_K with the vorticity vector field ω.

2) 008--- Integrer Points in Polyhedra (2008).pdf

Pág. 28(41) — Proof.

By theorem 17, the local Euler-Maclaurin expansion of the difference involves only faces of dimension $< k$.

Observamos o sinal ξ na equação.

Vejamos:

Pag. 27(40) - Corollary 18

Let $0 < k \leq d$.

Then the meromorphic function

$$S(p)(\xi) - S^{(h,p)}(p)(\xi)$$

has lowest degree $\geq -k+1$.

Let $k = 0$. Then

$$S(p)(\xi) = S^{(h,p)}(p)(\xi).$$

Interessante: (Estudar!!!) → π

Pág. 20(33) - Proposition 3

There exists a map I which to every polyhedron $p \subset V$ associates a meromorphic function with rational coefficients $I(p) \in M_\ell(V^*)$ so that the following properties hold:

(a) If p contains a straight line, then $I(p) = 0$

(aprofundar!!!)

20220822M

1) Relativity and Geometry...(1983).pdf
→ Torretti

Source: Books-017-2022/20220822/
Relativity and Geometry Torretti/
Relativity and Geometry-Elsevier
(1983).pdf

→ 401 pp.

Copy: Volume-021---20220818/
Original/
009-Relativity and Geometry..(1983).pdf

Pág. 08 (15) - Chapter 1

Newton Principles

1.1. The Task of Natural Philosophy

Newton does not understand by MOTION - MOTUS - every kind of quantitative and qualitative change, but only that which formerly called

MOTUS LOCALIS, that is,

locomotion or change of place.[2]

(2) → Newton, "De Gravitatione et Aequipondio Fluidorum", Def. 4, in Hall and Hall (1978), p. 91.

Force - vis - is

"the causal principle of motion and rest".[3]

Newton distinguishes two kinds of force.

These two kinds of force are respectively, the subject of Definitions IV and III of the "PRINCIPIA"

where the latter is referred to as "innate force (vis insita) of matter", and the former as "impressed force (vis impressa).

Innate force is said to be proportional to the "quantity" of matter - also called by Newton its "body" (corpus) or "mass" (massa) - and it is said to differ from the natural (inertia) of matter in our conception only.

Obs: "For a body perseveres in every new state by its force of inertia alone".

Obs: Livro todo!!!
Outras referências: ⇒

009---Unpublished Scientific Papers of Isaac Newton.pdf
→ 444 pp. Hall - Hall.

2) 009---Philosophy of Geometry from Riemann to Poincaré...pdf

3) 009---Relativity and Geometry-Pergamon Press Ltd (1983).pdf
⇒ reserva

4) 009---Roberto Torretti-Diccionario de Logica y Filosofia....pdf

5) 009---Roberto Torretti-El Paraíso de Cantor...pdf

6) 009-Newton-Cajori-Principia-Vol.1...pdf

7) 009-Newton-Cajori-Principia-Vol.2...pdf

8) 009-Isaac Newton-Principia-Livro I-EDUSP (2016).pdf

9) 009-Isaac Newton-Principia-Livros II e III-EDUSP (2012).pdf

10) 009-Isaac Newton-Princípios Matemáticos Fundação Calouste Gulbenkian (2017)\.pdf

11) 009---Isaac Newton, ... Leibniz —
Os Pensadores 19 - Abril (1974).pdf

12) 009---1967-from--frege-to-godel-
a-source-book-in-mathematical-
logic-1879-1931.pdf
→ 670 pages
→ Interessante !!!

13) 009---Mathematik...(2018).pdf
→ 1656 pp

Pág. 959(967) - Die Leibniz'sche
Sektorformel dient zur
Flächenberechnung
(A fórmula setorial de Leibniz
é usada para cálculo de área).

14)

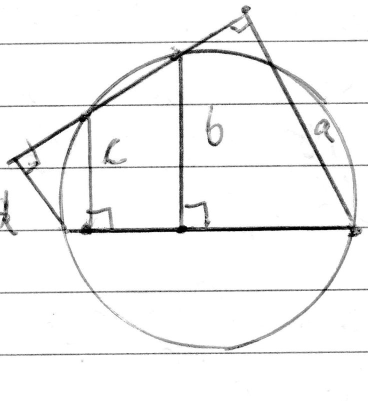

$$\frac{(a-c)(b+d)}{(b-d)(a+c)} = ?$$

Obs.: Aplicando-se π o resultado será diferente.

Source: 009---calcular-a-razão.png

15) DATA-NovoVivo 017-2022/
Books-017-2022/
20220812/
Space Time Matter Weyl/

H. Weyl-Space-Time-Matter-Dover (1922).pdf

Obs. Copy: Original/
010---H. Weyl-Space-Time-Matter-Dover(1922).pdf

Pág. 12(27) - First of all the straight line.

Important:

Pág. 19(34) - Vector

$$\xi_1 e_1 + \xi_2 e_2 + \xi_3 e_3.$$

Pág 117(132) - §15. Curvature

If P and P* are two points connected by a curve and if a vector is given at P, then this vector may be moved parallel to itself along the curve from P to P*.

Obs: See: Equation 36 (p.115(130)).

obs. Ver livro todo!!!

16) 011 --- Hermann Weyl - The concept of a Riemann Surface - (2010).pdf
→ 200 pp   Livro todo!!!

17) 012 --- Symmetry ... (2016).pdf
→ 178 pp.

18) A treatise on the Differential Geometry of Curves and Surfaces
→ Eisenhart
→ 498 pp

013 --- treatiseondiffer00eiseuoft.pdf

013 --- treatonthediffer00eiserich.pdf

19) A Treatise on the Higher Plane Curves: intended (intended) as a sequel to a treatise of conic sections.
→ George Salmon
→ 430 pp.

Source: 014 --- treatisehigherpl00salmrich.pdf

other books:

⇒

a) 014--- Coolidge - A treatise on algebraic plane curve ... (1959).pdf 553 p.p.

b) 014--- planealgebraiccu00hilt.pdf
→ Harold Hilton
→ 416 pages.

20) 20220812/Trigonometry-Sexagesimal/
The_precise_calculations_of_the_constant.pdf
→ Mei Xiaochun
→ 10 pp
obs: the precise calculations of the constant terms in the equations of motions of planets and photons of general relativity.
→ 23 April 2021.
Source:
1) 015--- the_precise_calculations_of_the_constante.pdf

Others ref:

Volume-021---20220818/Original/
015--- Curvature-important/
→ 65 arquivos.

a) 1403.0148v1.pdf
→ On Smarandache's Sphere.

b) AModelForSmarandacheAntigeometry.pdf

c) MeiXiachun.pdf
→ General Relativity can not predict the existence of Linear Plane Gravitational Waves
→ 25 pp

d) NeutroGeometryAntiGeometry31.pdf
→ NeutroGeometry & AntiGeometry are alternatives and generalizations of the Non-Euclidean Geometries. (reprinted)
Obs. Vol. 46, 2021 – 22 pp.

e) New_Relativistic_Paradoxes_and_Open_Questions.pdf
→ Smarandache
→ 128 pp.

f) Unsolved_problems_in_special_and_general.pdf
→ Relativity.
→ Smarandache.
⇒ 319 pp.
→ Livro todo!!!

g) VariancePlaneGeometry.pdf
→ Smarandache
→ 114 pp.

h) Using the Newton's Theory of Gravity to Calculate the Deflection of Light.pdf

Obs: Source: 015---Curvature-important/ 20220818/.

→ Using the Newton's Theory of Gravity to Calculate the Deflection of Light in the Solar System and the Orbital Poles of Light's Motion of General Relativity.

→ Mei Xiaochun

→ 21 pp.

Source: https://doi.org/10.54647/physics14417

March 24 2022

Volume 7, Issue I, February 2022.

ISSN: 2706-8862

SCIREA Journal of Physics.

i) Quantization in Astrophysics, Brownian Motion, and Supersymmetry (2007)....pdf

→ Smarandache & Christianto

→ 516 pp.

j) 015---Trust_in_expert_testimony_Eddingtons_1911.pdf

→ Ben Almassi

→ 11 pp.

21) Source: Volume-021---20220818/
Original/
016-0-Non-Euclidean-
Geometry/

a) 016---Jon Dattorro-Convex optimization & Euclidean Distance Geometry... (2011).pdf
→ 868 pp.

b) 016---Euclidean-Geometry-and-Transformations-Dodge.pdf → 329 pp.

c) 016---Geometry of Hypersurfaces (2015).pdf → 601 pp.

d) 016---Greenberg-Euclidean and Non-Euclidean Geometries... (2008).pdf → 665 pp.

e) 016---Hardy Spaces on the Euclidean Space (2001).pdf
→ 298 pp.

f) 016---Henry Parker Manning-Geometry of Four Dimensions... (2005).pdf → 360 pp.

g) 016---Henry Parker Manning-Introductory Non-Euclidean Geometry (2005).pdf → 85 pp.

h) 016--Ryan-Euclidean and Non-Euclidean Geometry.pdf
→ 234 pp.

i) 016--Ryan-Geometry of Hypersurfaces (2015).pdf → 601 pp

20220823M.

1) DATA-NovoVivo-017-2022/Books-017-2022/
20220814/
An Illustrative Guide to Multivariable and Vector Calculus/
Stanley...-An Illustrative...pdf
Copy: 017---Stanley...-An Illustrative...pdf
Pág. 52(65) - Fig. 2.3
A 3D graph of a function of two variables.

obs: Usar TTI (M.U.S.) para recalcular a projeção ortogonal.

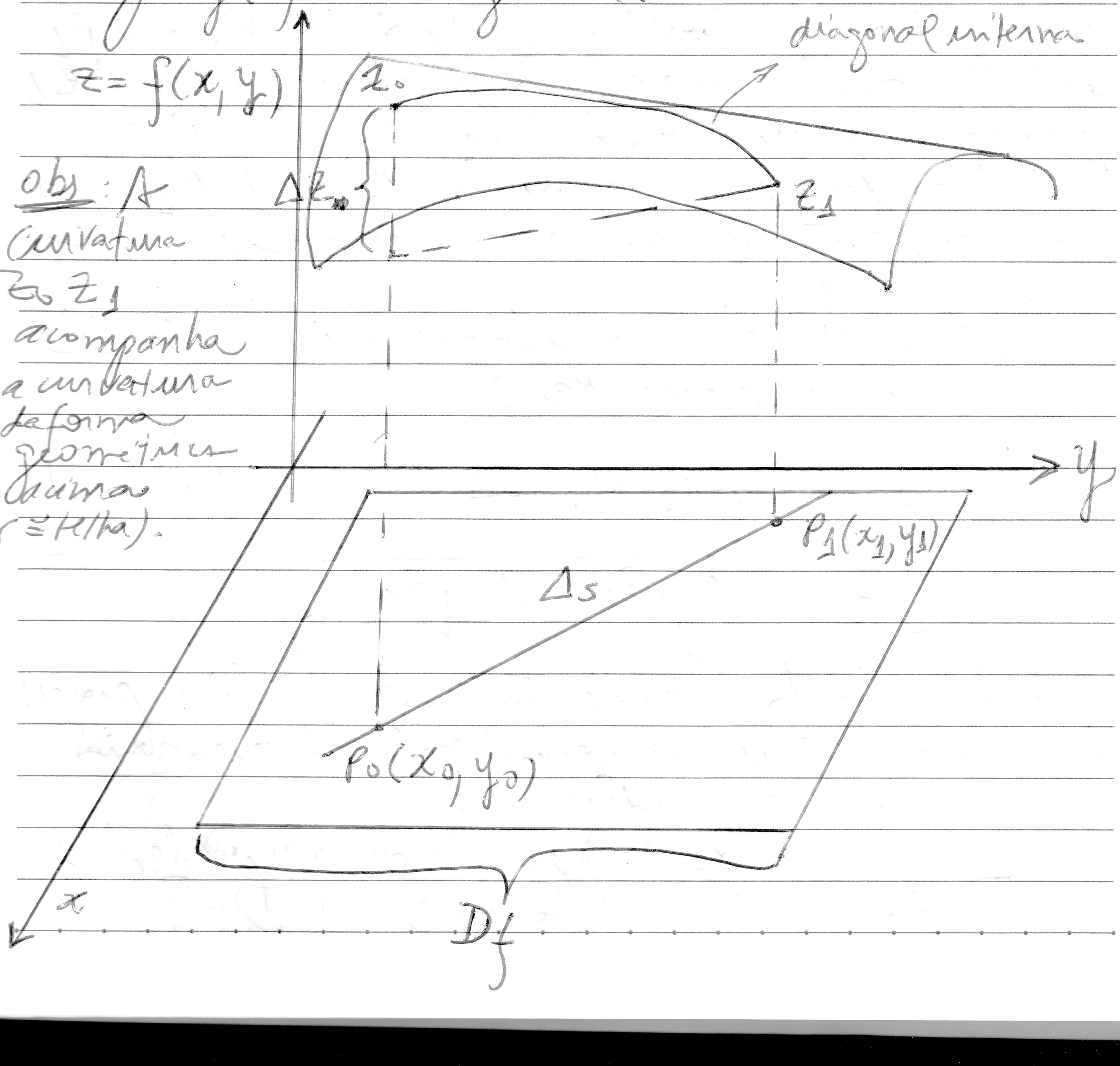

$z = f(x, y)$
z_0
Δz
z_1
diagonal interna
y
x
$P_1(x_1, y_1)$
Δs
$P_0(x_0, y_0)$
D_f

obs: A curvatura $z_0 z_1$ acompanha a curvatura da forma geométrica acima (≅telha).

$f: \mathbb{R}^2 \rightarrow \mathbb{R}$ of two variable is:

$$\lim_{P_1 \to P_0} \frac{f(x_0, y_0) - f(x_1, y_1)}{\sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}} \qquad (2.1).$$

Vejamos:

The graphical foundation for the limit expression (2.1) is shown in Figure 2.3. The things to note are, firstly, the graph of f is suspended in 3D; secondly, the domain D_f lies in the xy-plane; thirdly, the points P_0 and P_1 in D_f give rise to values z_0 and z_1, respectively; and finally, the line in the domain joining P_0 and P_1 traces out the black curve in the graph of f.

In direct analogy with the 1D derivative we have

$$\Delta z = f(x_0, y_0) - f(x_1, y_1), \text{ and}$$

$$\Delta s = \sqrt{\Delta x^2 + \Delta y^2}.$$

Referring to Figure 2.3, for the limit process to make sense several related questions need to be addressed:

- does $f(x_0, y_0)$ exist, or equivalently, does (x_0, y_0) belong to D_f?

- is f continuous at (x_0, y_0)?
- does the limit $\lim\limits_{P_1 \to P_0}$ exist if $P_1 \in D_f$?

2) 20220815/The Encyclopedia of Physics/
Biedenharn- Angular Momentum...pdf
Copy: 018--- Biedenharn- Angular
Momentum in Quantum Physics..pdf
→ 747 pp.
→ Livro todo!!! Estudar!!!
others:
a) 018---Biedenharn-Angular momentum...pdf
→ 29.4 MB (página dupla)
b) 018---Biedenharn- Angular momentum... djvu
→ 5.32 MB (OCR) → image and text!
c) 018---Devanathan -Angular Momentum...pdf
d) 018---Symmetries in Quantum Mechanics..pdf

3) 20220815/ Geometric Algebra Applications/
Geometric Algebra Applications
Vol. 1 ...pdf
Vol. 2 ...pdf
Copy: 019--- Geometric Algebra Applications
- Vol. 1 ...(2019).pdf
- 019--- Geometric Algebra Applications
Vol. 2 ...(2020).pdf
Obs: Volume 01 →
Pág. 205(228) -
Fig. 6.7 - (a) Dihedral angle φ between
two planes (b) the spherical
triangle.

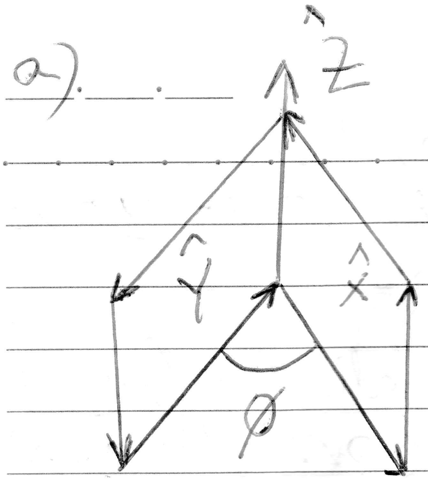
a)
$\hat{z}$
$\hat{y}$
$\hat{x}$
ϕ

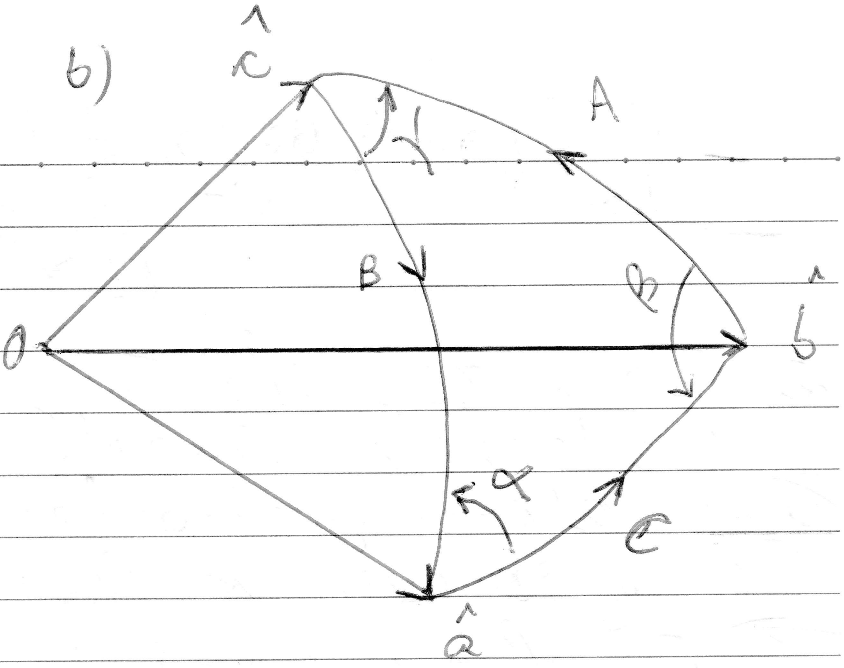
b)
$\hat{c}$
A
γ
B
β
0
$\hat{b}$
α
C
$\hat{a}$

Obs:

Lie Algebra: the bivector generators of the unitary group are

$E_{ij} = e_i e_j + f_i f_j, \ (i < j = 1 \dots n)$

$F_{ij} = e_i f_j + f_i e_j, \ (i < j = 1 \dots n)$

$J_i = e_i f_i$.

4) 020---14417.pdf

Source:

20220823/Smarandache/14417.pdf

Obs.: Using the Newton's Theory of Gravity to Calculate the Deflection of Light in the Solar System and the Orbital Poles of Light's Motion of General Relativity.

Source: www.scirea.org/journal/PaperInformation?PaperID=7138

5) 20220823/ Light in Einstein's Universe/
Light in Einstein's Universe – the Role of Energy in Cosmology and Relativity..pdf
→ 228 pp.
Obs: Pág. 203 (211) – 8-1.

Eddington's table.
→ "fields of force"

6) 20220815/ Blaschke Wilhelm/
Wilhelm Blaschke...Projektive Geometrie...pdf
Pág. 21 (19) – Homogene Zeiger,
Kollineation,
Korrelation.
Abb. 4 → Kartesische Zeiger

$$u_0 + u_1 \xi_1 + u_2 \xi_2 + u_3 \xi_3 = 0, \quad (1)$$

Copy: 021--- Wilhelm...Projektive...pdf.
→ 195 pp.

7) 20220817/ General Relativity for Planetary Navigation/
General Relativity for Planetary Navigation.pdf
Copy:
022--- General Relativity...pdf
→ 129 pp.
Obs: Livro todo!!!

Pág. 31(38) - Chapter 2

Schwarzschild Solution for Spherical Symmetry

See: Pág. 39(46) - Fig. 2.1

Gravity of spherical shell

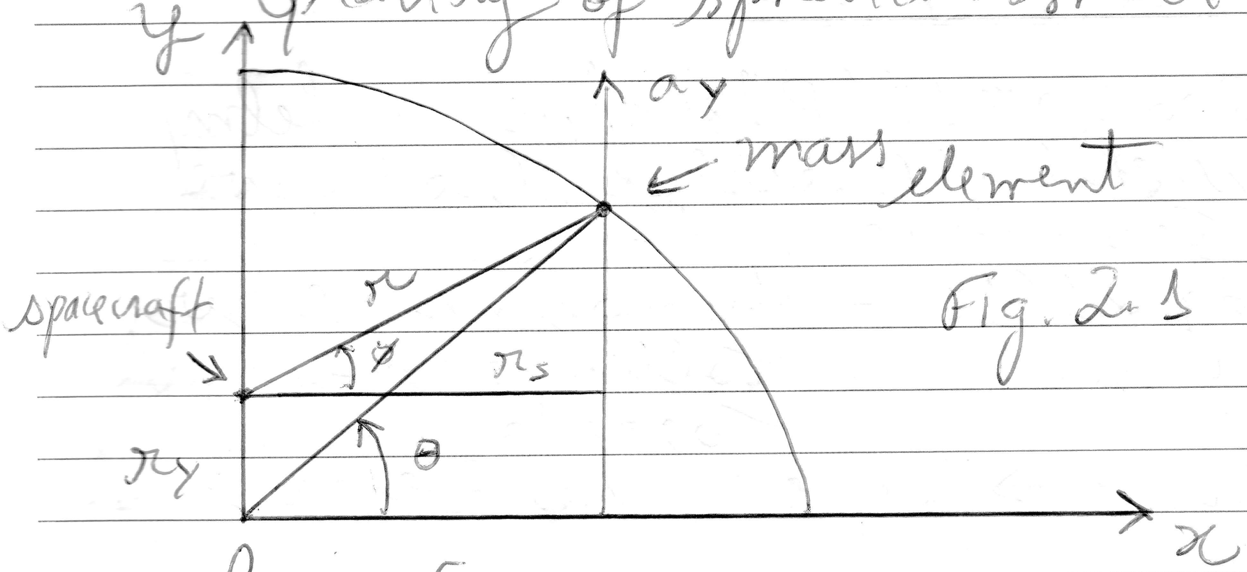

Fig. 2.1

We have for a_y

$$a_y = \frac{2\pi r_s \cos\theta\, u\,(r_s \sin\theta - r_y)}{r^3}$$

The acceleration of the spacecraft due to the sphericall shell or the acceleration of the spherical shell by the spacecraft is obtained by integrating a_y from θ equals -90 to 90 deg. The a_y component integrates to zero.

The Equation $\bar{a}_y = \int_{-\frac{\pi}{2}}^{+\frac{\pi}{2}}$ ⇒

$$\bar{a}_x = \int_{-\frac{\pi}{2}}^{+\frac{\pi}{2}} a_y \, d\theta = \left. \frac{2\pi M(r_y - r_y \sin\theta)}{r_y^2 (r_s^2 + r_y^2 - 2 r_s r_y \sin\theta)^{\frac{1}{2}}} \right|_{-\frac{\pi}{2}}^{+\frac{\pi}{2}} = 0$$

This proves the shell theorem for $r < r_s$, which was first obtained by Newton. When r is greater than r_s, the acceleration is mass divided by r^2.

The shell theorem is also true for general relativity.

Important:

The proof is difficult because we cannot use the Euclidean model shown in Fig. 2.1. ✓

Obs:

Acrescente a isto, o valor π e a sua repercussão geral na equação em questão.

General relativity distorts the lengths of lines.

8) 20220817/ The Theory of Space... |

Obs: Posta toda!!!

Books referenciados e copiados:

8.1) 023--- Fock- The Theory of Space, Time and Gravitation. pdf.

→ 461 pp.

8.2) 023--- Logunov - The Relativistic-
theory of Gravitation.pdf
→ 235 pp.

8.3) 023--- Relativity, Gravitation,
and Cosmology....pdf
→ 355 pp.

Pag. 14(29) - Chapter 2
Special relativity and the flat
spacetime

Obs. 2.1 - Coordinate symmetries.

Obs: Estudar!!!

$F'_i - m'a'_i = \sum [R]_{ij} (F_j - ma_j) = 0$ (2.2)

8.4) 023--- space-time-gravitation.pdf
→ 220 pp.
→ Vladimirov.

Pag. 26(28) - 1.4 From Riemann to Einstein

8.5) 023--- Zakharov - Gravitational
Waves in Einstein's Theory.pdf
→ 203 pp. Obs: Livro todo!!!
Pag. 90(110) - Chapter 10 - Plane Gravitational Waves
defined by an Absolutely Parallel Vector Field.

9) 20220817/ Florentin Smarandache-
Books/
Quantization in Astrophysics,
Brownian Motion, and
Supersymmetry... pdf
obs.:
Copy: 024---Quantization in Astrophysics,
Brownian Motion, and
Supersymmetry. pdf
→ 516 pp.

Obs: Livro todo!!!

10) 20220817/ Florentin Smarandache-
Books/
Variance on topics of plane geometry. pdf
Copy:
024---Variance on topics of plane
geometry. pdf
→ 115 pp.
Source:
https://digitalrepository.unm.edu/
math-fsp/257/

11) Books-017-2022/ 20220823/
The VNR Concise Encyclopedia of
Mathematics by W. Gellert/
The VNR Concise Encyclopedia of
Mathematics by W. Gellert. pdf
→ 833 pp.

12) ZIP-03/
Books-Import-20220608/
Cálculo Vectorial - Claudio Pita Ruiz/

Copy:
a) 025--- Cálculo Vectorial...pdf
→ 1084 pp.

Pág. 201(211) - 2.9 Vectores normales.

$f: U \subseteq \mathbb{R}^2$

$\mathbb{R}$ - definida en el conjunto abierto U de $\mathbb{R}^2$.
S - Es una superficie S en el espacio $\mathbb{R}^3$.
Sea $p = (x_0, y_0, f(x_0, y_0))$ un punto de esta superficie.

Queremos obtener una expresión que nos de un vector normal a la superficie S en p.

Pág. 203(213) - Figura 2

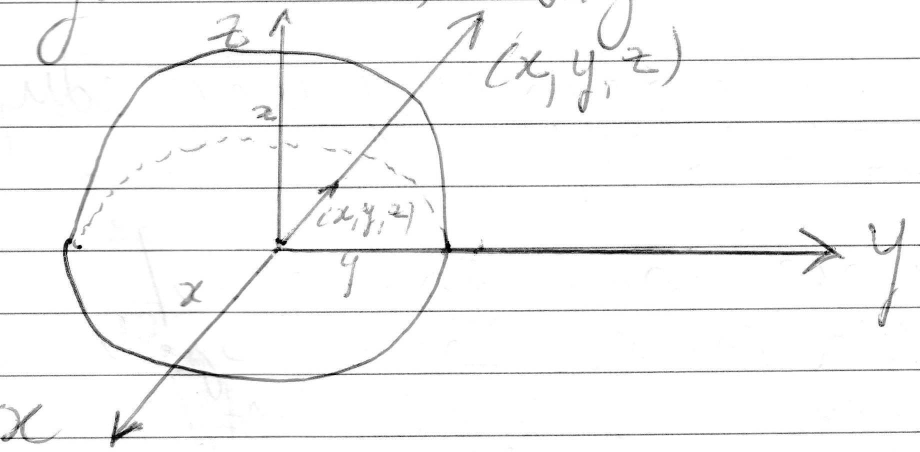

Figura 2. Gráfica de un punto y un vector normal a la esfera $\sqrt{1-x^2-y^2}$.

b) 025---Claudio de J. Pita Ruiz - Álgebra Lineal...(1991).pdf
→ 392 pp. Página dupla.

c) 025---Claudio Pita Ruiz - Calculo de una Variable.pdf
→ 902 pp.

13) ZIP-03/Books-Import-20220608/Geometry and Light-The Science of Invisibility/..
Copy: Volume-021---20220818/Original/

a) 026---A First Course in Geometric Topology and Differential Geometry.pdf
→ 440 pp.

Pag. 167(183) - Chapter IV
Curves in $\mathbb{R}^3$

Pag. 328(344) - Chapter VIII
The Gauss-Bonnet Theorem.

b) 026---Euclidean Geometry and Transformations.pdf
→ 391 pp.
Obs: Livro todo!!!

Pág. 98 - Section 14

Translations and Rotations

Pág. 100 - 14.11 Theorem

Each direct isometry is a rotation or a translation.

14.12 - Theorem →
the product of two translations is a translation.

Pág. 157 - Section 27
Applications of Similarities to elementary Geometry.

27.2 Theorem - Three parallel lines cut off proportional segments from two transversals.

Pág. 158

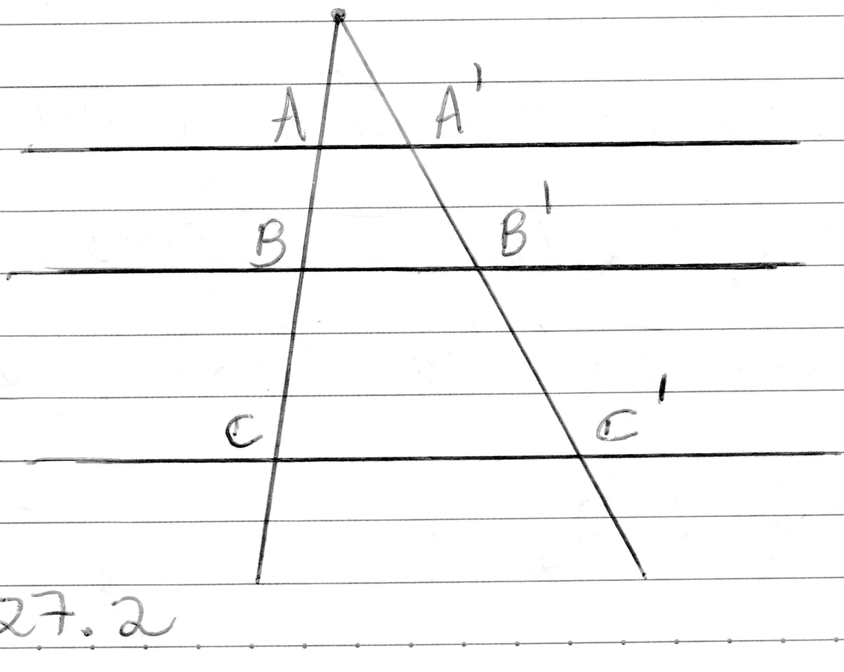

Figure 27.2

27.2 Theorem – Three parallel lines cut off proportional segments from two transversals.

If the transversals are parallel, then the theorem follows immediately since opposite sides of the resulting parallelograms are congruent.

So suppose transversals ABC and A'B'C' meet at P (Fig. 27.2).

Then H(P, PB/PA) maps triangle PAA' to PBB' and H(P, PC/PB) maps triangle PBB' to PCC'.

Now AB/BC = A'B'/B'C' follows from similar triangles.

IMPORTANT:

Usar TT! para recalcular as equações → MIRROR ORTHOGONAL.

Obs: Usar PAPPUS'S THEOREM (3 Points) com paralelismo de GAUSS (anteriormente abordado, mas com necessidade de aprofundamento).

Pág. 166 – Figure 28.7

28.8 – Compare the proof of theorem 28.7

with Corollary 20.8. (Pág. 126)

20.8 Corollary — If CABU, ABCV, and BCAW are parallelograms, then ABC is the medial triangle for triangle UVW.

Furthermore, the altitudes of triangle ABC concur by Theorem 20.1 since they are the perpendicular bisectors of the sides of triangle UVW (Fig 20.8)

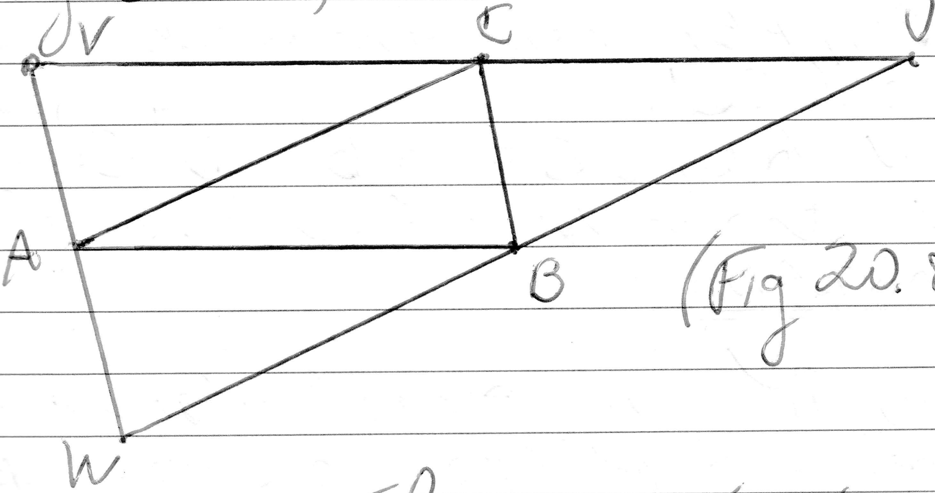

(Fig 20.8)

obs: 20.1 Theorem (pág. 124) —

The perpendicular bisectors of the sides of a triangle concur.

Voltando: 28.8

28.9 - The method of homothety is quite useful for constructions.

Pág. 218 – Section 37

Lines in the GAUSS PLANE

37.1 Theorem – Points A, B, C are collinear $\underbrace{\text{iff}}$

↳ definição matemática

Iff → if and only if (as a written abbreviation).

$$\begin{vmatrix} a & \bar{a} & 1 \\ b & \bar{b} & 1 \\ c & \bar{c} & 1 \end{vmatrix} = 0.$$

37.2 Theorem – An equation for the line passing through points A and B is

$$\begin{vmatrix} z & \bar{z} & 1 \\ a & \bar{a} & 1 \\ b & \bar{b} & 1 \end{vmatrix} = 0.$$

Clearly the equation is satisfied when $z = a$ and when $z = b$. The theorem then follows by Theorem 37.1.

As an alternative method of proof, expand the determinant to obtain the equation

$(\bar{a}-\bar{b})z-(a-b)\bar{z}+(a\bar{b}-\bar{a}b)=0.$

Writing $x+iy$ for z, so $\bar{z}=x-iy$, we find that the equation reduces to

$(\bar{a}-\bar{b}-a+b)x+i(\bar{a}-\bar{b}+a-b)$
$y(a\bar{b}-\bar{a}b)=0.$

Obs: Estudar!!!

Since each of these coefficients is pure imaginary, when we multiply through by i, we obtain an equation of the form

$ax+\beta y+\gamma=0$ for real

$\alpha, \beta, \gamma.$

This is an equation for a

STRAIGHT LINE. (Pág. 219)

Pág. 227 – Section 39

Isometries and Similarities in the GAUSS PLANE.

Pág. 237 - Figure 41.4

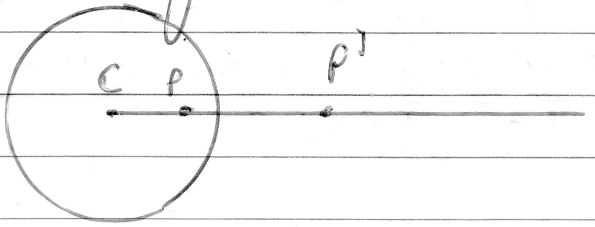

41.5 Theorem → Inversion is a transformation of the inversive plane and is involutoric.

41.6 Theorem → Points inside the circle of inversion map to points outside the circle, and points outside map inside. Points on the circle are invariants (invariant) and are the only invariant points.

If P lies inside the circle of inversion with center C and radius r, then $CP < r$.

Since $CP \cdot CP' = r^2$, then CP' cannot be less than or equal to r, for then

$$CP \cdot CP' < r^2.$$

Hence (por isso) $CP' > r$, so P' lies outside of circle.

41.7 - Definition → A transformation of the plane is called conformal iff it preserves angles between curves.

It is directly conformal if it preserves the sense of the angle as well, and inversely conformal if it reverses the sense.

41.8 Theorem - Isometries and similarities are conformal; direct similarities are directly conformal and ~~[illegible]~~ opposite similarities are inversely conformal.

Pág. 238- 41.10 Theorem

Inversion is inversely conformal.

PAQ and P'A'Q' are oppositely congruent.

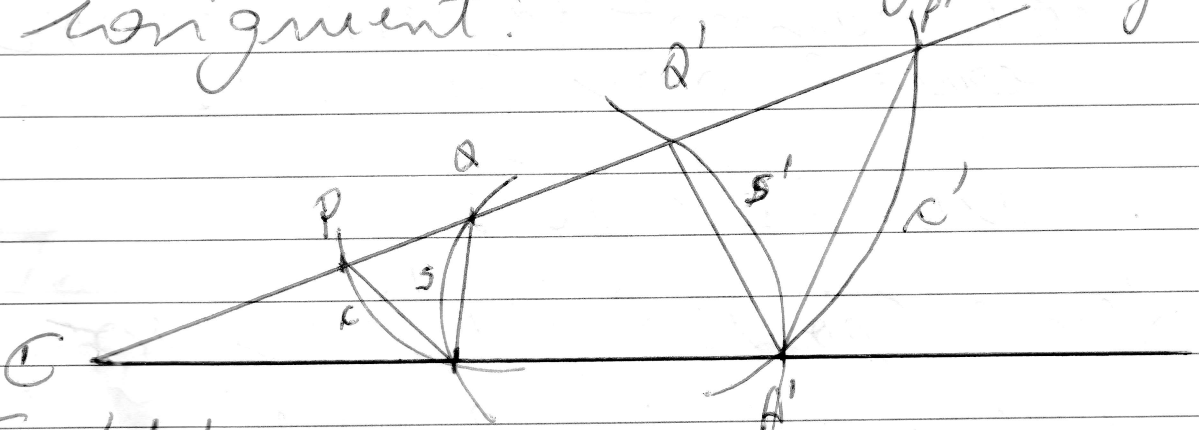

Fig 41.10

Obs: Interessante!!!

ATENÇÃO:

41.11 Theorem

A line through the center of inversion is invariant.

41.12 Theorem

A line not through the center C of inversion maps onto a circle through C.

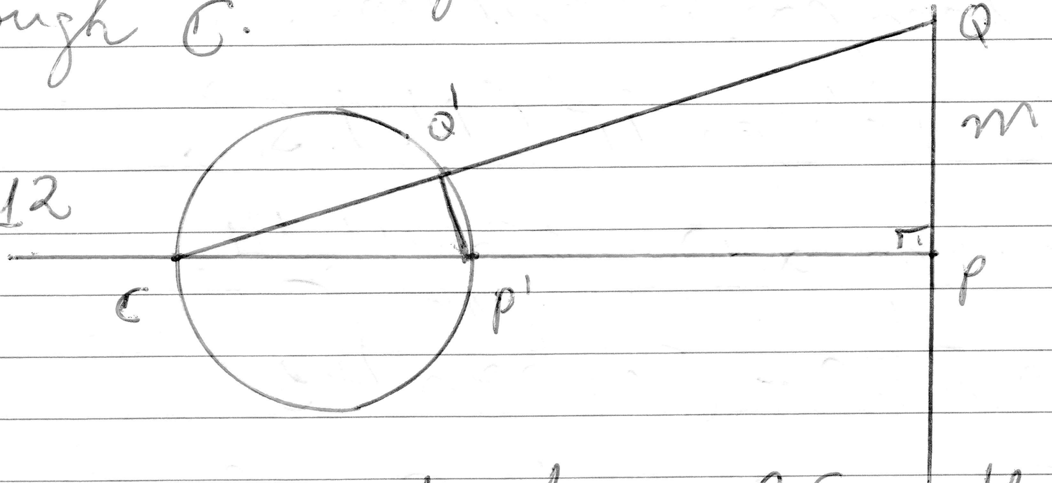

Fig 41.12

Drop a perpendicular CP from the center C to given line m.
Let Q be any other point on line m (see Fig. 41.12).
Let P' and Q' be the inverses of P an Q.
By Theorem 41.9, triangles CPQ and CQ'P' are similar.
Hence ∡ CQ'P' ≅ CPQ = 90°, so it follow that Q' lies on the circle of diameter CP'.
Since the ideal points maps to C, each point of line m maps to some point on that circle.

Furthermore, for each point Q' (other than C) on that circle, ray CQ' cuts line m in a point Q, and from the first paragraph of this argument, Q inverts to Q'. Hence the image of the line m is the entire circle on CP' as diameter.

⇑

importantíssimo!!!

Obs: Utilizar TI (M.U.S.).

See: Pág. 240 - Fig 41.14.

Pág. 250 - Section 43

Inversion and Complex Geometry

↳ estudar!!!

Obs: Estudar esta "inversion" em relação ao MIRROR, fazendo-se uma projeção ortogonal

Importante: Usar o TI (M.U.S.) e analisar o seu comportamento como uma "Complex Geometry"

Atente-se às formalidades estritas, observando-se como se fosse uma transformação de um corpo rígido.

Pág. 259 – Fig 44.5

Theorem Feuerbach's theorem.

The ninepoint circle of a triangle is tangent to each of the equicircles of the triangle.

Important: Utilizar esta definição, mantendo-se as proporções entre os círculos quando da aplicação do π^c (M.U.S.) com a devida projeção ortogonal com o objetivo de se manter as mesmos características, juntamente com GAUSS LINES, paralelismo de GAUSS, Pappus's theorem.

Todos estes aspectos deverão estar interligados a fim de se manter a consistência da pesquisa, evitando-se erros ou dispersões desnecessárias.

c) 026--- Geometry and Light.pdf
→ 298 pp

Pág. 126 (129) – §18. CURVATURE

$$ds^2 = g_{ij}\, dx^i dx^j, \qquad (18.1)$$

defining a metric tensor g_{ij}. As in Euclidean space, the metric is symmetric,

$$g_{ij} = g_{ji}, \qquad (18.2)$$

because (relabeling indices) $g_{ij}\, dx^i dx^j = g_{ji}\, dx^j dx^i = g_{ji}\, dx^i dx^j$, so g_{ji} quantifies distances exactly like g_{ij}.

Pag. 127(130) - Figure 3.15.

Pag. 133(136) - §20 - PARALLEL TRANSPORT AND COVARIANT DERIVATIVES

The condition (19.16) defines the notion of the tangent vectors at all points on a curve being "parallel" in so far as this is possible in a curved space.

Obs.: Livro todo!!!

d) 026--- Modern Geometry - Methods and Applications...pdf → 478 pp. Livro todo!!

e) 026 --- The Fractal Geometry of Nature.pdf → 976 pp.

20220824//
ZIP-03/Books-Import-20220608/
Important-Books-011---2022/
Carlos-Ivorra/
1) 027---Advanced Euclidean Geometry...pdf
→ 268 pp.
Pág. 13(25) - Learning from Geometric Fallacies!

Fallacy 1 - Any scalene triangle is isosceles

20220826// (cont.)

Pág. 15(27) - Fallacy 2 →
Two distinct perpendicular can be drawn to given ~~a~~ line from a given external point.
Pág. 16(28) - Fallacy 3 →
A right angle has the same measure as an obtuse angle.
Pág. 17(29) - Fallacy 4 →
Every point inside a circle is on the circle.
Pág. 19(31) - Fallacy 5 →
Two segments of unequal length are actually of equal ~~lenght~~ length.
Obs: Livro todo!!!

Pág. 36(48) - The Gergonne Point →

Joseph-Diaz Gergonne (1771-1859)
publicou (1810) →
Annales des mathématiques pures et appliqués.
Source:
www.numdam.org/item/AMPA/
or www.numdam.org/journals/AMPA/
Copy: DATA-NovoVivo-017-2022/
Books-017-2022/ 20220826/
Annales des mathématiques pures et appliqués -
22 Volumes/

Pag. 54(66) - Pascal's Theorem

Pag. 57(69) - Figure 3-16

points X, Y and Z → straight lines (Pappus)

Pag. 60(72) - Pappus's Theorem)

Pag. 61(73) - the Simson Line

Theorem 3.7 (Simson's Theorem)

The feet of the perpendiculars drawn from any point on the circumcircle of a triangle to the sides of the triangle are colinear.

In figure 3-21 point P is on the circumcircle of ΔABC.

$\overleftrightarrow{PY} \perp \overleftrightarrow{AC}$ at point Y, $\overleftrightarrow{PZ} \perp \overline{AB}$ at point Z, and $\overleftrightarrow{PX} \perp \overleftrightarrow{BC}$ at point X

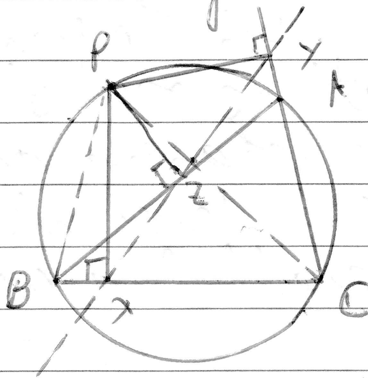

Fig. 3.21
(Pag 62(74)).

According to Simson's (i.e. Wallace's) theorem, points X, Y, and Z are collinear. This line is usually referred to as the SIMON LINE (sometimes called the pedal line).

Although not necessarily the simplest proof of Simson's theorem, for the sake of consistency we will prove this theorem using Menelaus's theorem.

Obs: Estudar tudo!!!

Ver livro todo!!!

2) 028-- Affine and Projective Geometry.pdf
→ 246 pp. - Bennett

Pag. 14(31) - 1.6

The Theorems of Pappus and Desargues.

Figure 1.3

Pág. 18(35) - Chapter 2 - Affine Planes

It was shown in Theorems 9 and 11 of Chapter 1 that the rational and the real coordinate planes are affine planes in that two points determine a unique line, and the parallel axiom holds.

Obs: Livro todo!!!

3) 029 --- Analisis Matematico (Ivorra, Castillo Carlos)... pdf

→ 496 pp.

Pág. 83(99) - 2.4 Espacios de Hilbert.

Pág. 141(157) - * Geometrías no euclídeas
(euclídeas)

Obs. estudar!!!

$\cos \alpha = -\cos \beta \cos \gamma + \operatorname{sen} \beta \operatorname{sen} \gamma \cos a.$

Obs: los ángulos han de estar en radianes. Si el lado a mide un metro, su medida en radianes es aproximadamente $a \cong 1.6 . 10^{-7}$, y entonces $\cos a \cong$

$0{,}9999999999999872.$

Por consiguiente nuestro triángulo cumple

$\cos\alpha \approx -\cos\beta \cos\gamma + \operatorname{sen}\beta \operatorname{sen}\gamma =$

$-\cos(\beta+\gamma) = \cos(\pi-\beta-\gamma)$,

de donde se sigue que $\underbrace{\alpha+\beta+\gamma \approx \pi+2k\pi.}_{\vee}$

Com la suma de los ángulos de um triángulo elíptico ~~que~~ está entre 0π y 3π, concluimos que la suma ha de parecerse a π o a 3π. Enseguida descartaremos el segundo caso y concluiremos

$$\alpha+\beta+\gamma \cong 2\pi.$$

Obs: Esta diferença de cálculo ocorre em decorrência do erro do Pi (π).

Com a implementação do π (M.U.S), com o uso desta nova geometria, nos teremos uma precisão sem igual, assim, nos esperamos.

Pág. 148 (164) - 3.10

Apéndice: La trascendencia de e y π.

Obs.: Ler tudo até 155 (171).

Pág. 174 (190) 4.3 - Curvas parametrizables

Pág. 175 (191) - La primera ley de Newton afirma que si un cuerpo está libre de toda acción externa, permanecerá en reposo o se moverá en línea recta a velocidade constante.

Todo esto se resume en que su velocidad en sentido vectorial permanece constante.

Pág. 177 (193) -

Longitud de un arco.

Obs: Ler tudo até 184 (200).
→ Interessante!!!

Pág. 189 (205) * Longitud de arcos no euclídeos.

Pág. 215 (231) 5.4 Geodésicas.

Pág. 220 (236) - 5.5 Superficies

Pág. 223 (239) 5.6 La curvatura de Gauss.

20220827///

4) 030--- Classical Geometry Euclidean...pdf
→ 494 pp. Obs: Livro todo!!!

Pág. 171 (186) - 6.4 Miquel's Theorem
Pág. 172 (187) Theorem 6.4.1 (Miquel's Theorem)

Given Δ ABC and three menelaus points X, Y and Z, one on each side (possibly extended) of the triangle, then the circles formed using a vertex and its two adjacent menelaus points are concurrent at a point M.
The point of concurrency is called Miquel Point.

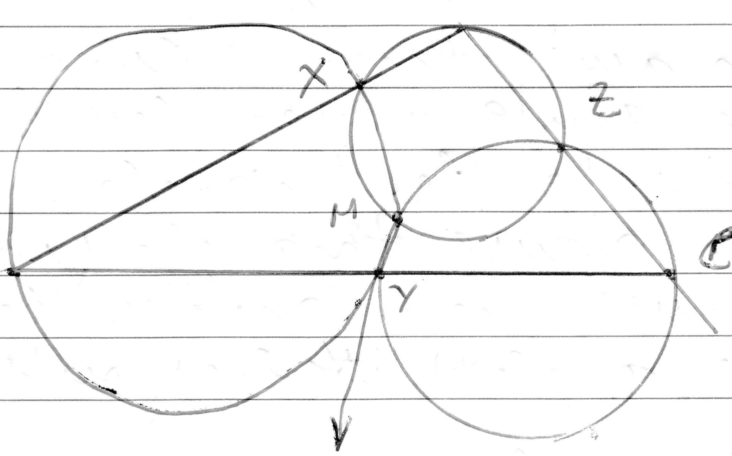

esta diferença decorre do erro do π, com o advento do ΠI (M.V.S.) esperamos solucionar esta discrepância.

Obs: Estudar até pág. 176 (191).

Pág. 190 (205) - the Euler Line

Theorem 6.6.2 (Euler line)

Given a nonequilateral triangle ABC, the circumcenter S, the centroid G, and the orthocenter H are collinear and form the Euler line.

Obs: The circumcenter is the center of a triangle's circumcircle. It can be found as the intersection of the perpendicular bisectors.

The centroid is the centre point of the object. It is also defined as the point of intersection of all the three medians.

The median is a line that joins the midpoint of a side and the opposite vertex of the triangle.

The orthocenter is the intersection point of the altitudes drawn from the vertices of the triangle to the opposite sides.

For an acute triangle (is a triangle with three acute angles (less than 90°), it lies inside the triangle.

For an obtuse triangle (is a triangle with one obtuse angle (greater than 90°) and two acute angles), it lies outside of the triangle.

For a right-angled triangle, it lies on the vertex of the right angle.

In fact, G is a trisection of HS; that is, G is between H and S and GH = 2GS.

Note: If ΔABC is equilateral, then S = G = H, and conversely, if S = G = H, then ΔABC is equilateral.

→

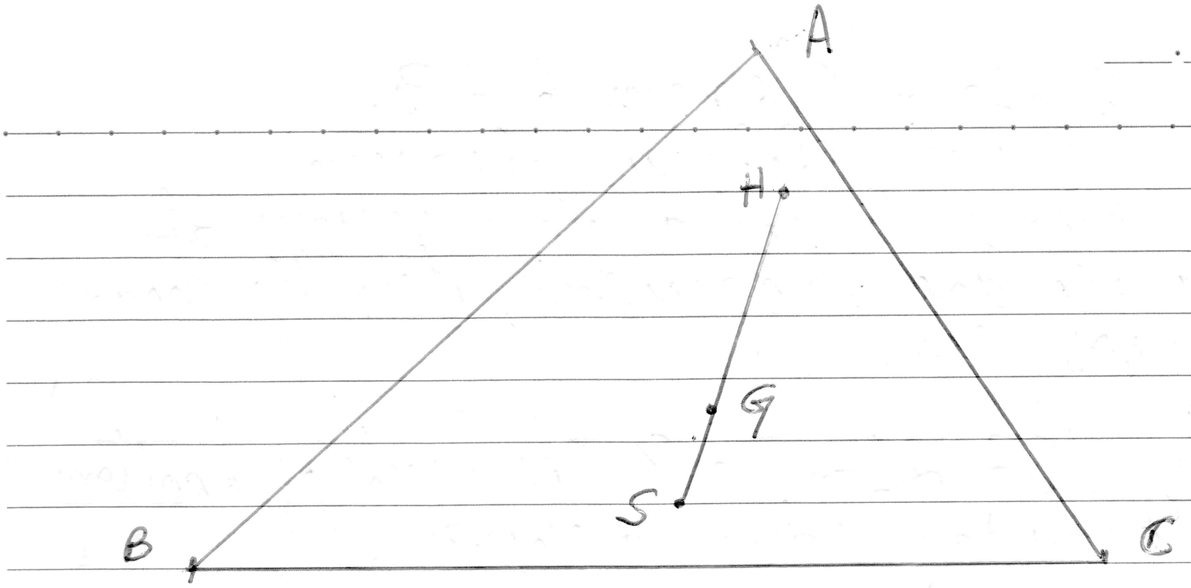

Obs: Veja a prova!!

Pag. 193(208) - 6.7 The Steiner-Lehmus Theorem

Theorem 6.7.1. If the altitudes of a triangle are congruentes, then the triangle is isosceles.
Proof. In the figure on the following page, if $BD \doteq EC$, then $\Delta BDC = \Delta CEB$ by the HSR congruence theorem.

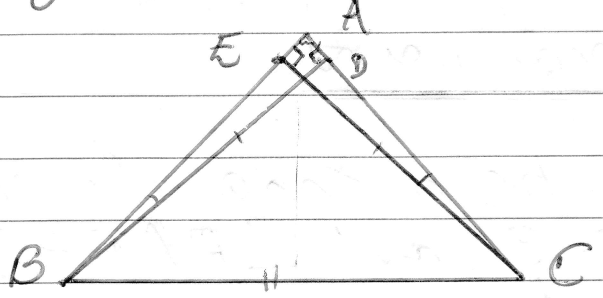

Therefore, $\Delta ABD \equiv \Delta AEC$ by the ASA congruente theorem, so that $AB = AC$.
Therefore, $\angle B = \angle C$ and ΔABC is isosceles.

Theorem 6.7.2: If two medians of a triangle are congruent, then the triangle is isosceles.

Pág. 195(210) - Theorem 6.7.3.
(Steiner-Lehmus Theorem)

If two internal angle bisectors of a triangle are congruent, then the triangle is isosceles.

⇙

Pág. 197(212) - 6.8 - The Circle of Apollonius -
→ (estudar com atenção!!!

Pág. 207(222) - Chapter 7

The Euclidean Transformations or isometries.

7.1. Rotations, Reflections, and Translations.

Pág. 223(238) - Example 7.3.7

The Fermat Point

Let ABC be a triangle with $AB = 4$, $AC = 3$, and $\angle BAC = 30^\circ$. Determine the minimum value of

$$PA + PB + PC$$

where P is any point inside the triangle ABC, and find the location of the point P that yields this minimum value.

Pág. 224 (239)

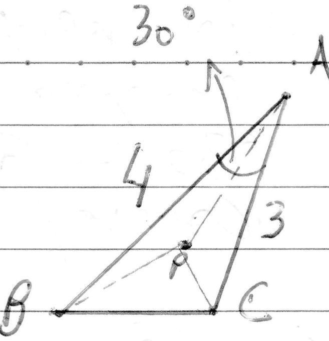

Solution: Consider the effect of $R_{A,60°}$ on ΔABC, as in the figure below.
This maps ΔABC onto $\Delta A'B'C'$ and takes the point P to the point P'.
Since $\angle PAP' = 60°$, triangle APP' must be equilateral. Hence

$$\begin{aligned} PB + PA + PC &= BP + PP' + P'C' \\ &\geq BP' + P'C' \\ &\geq BC' \end{aligned}$$

by the Triangle Inequality.

In mathematics, the triangle inequality states that for any triangle, the sum of the lengths of any two sides must be greater than or equal to the length of the remaining side. [1][2]

Source:

https://en.wikipedia.org/wiki/Triangle_inequality

Ref:

Pedoe, Daniel (1988). Geometry: A comprehensive course.

Rudin, Walter (1976). Principles of Mathematical Analysis.

Copy ref: Volume-021---20220818/
Original/
030---Geometry A comprehensive
course/

cont. description Triangle Inequality:

This statement permits the inclusion of degenerate triangles, but some authors, especially those writing about elementary geometry, will exclude this possibility, thus leaving out the possibility of equality [3]. Ⓐ

[3] Jacobs, Harold R. Geometry.
p. 246 ISBN 0-7167-0456-0.

Obs: Degenerate triangle

A degenerate triangle has collinear vertices and zero area, and thus coincides with a segment covered twice (if the three vertices are not all equal; otherwise, the triangle degenerates to a single point). If the three vertice are pairwise (Emparelhados) distinct, it has 0° angles and one 180° angle.

If two vertices are equal, it has one 0° angle and two undefined angles.

→

Jacobs - Geometry - p. 246 → 245 and 246.
correto:

Pág. 225 (245) - Lesson 7
Two more ways to prove triangles congruent.

hang glider - asa delta (1940) - NASA.
the sail of a Rogallo glider consists of two congruent triangles that share a common side as shown in the figure below:

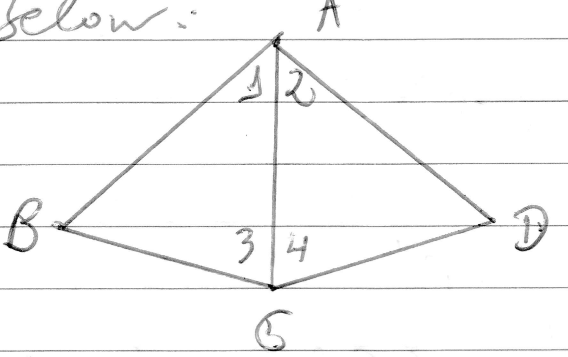

ff:

Pág. 246 do livro (266) - Lesson 3
Quadrilaterals that are parallelograms
obs:
Pág. 247 (267) - Theorem 36
A quadrilateral is a parallelogram if both pairs of opposite sides are equal.
Theorem 37 - A quadrilateral is a parallelogram if two opposite sides are both equal and parallel.
Theorem 38 - Pág. 248 (268) →
A quadrilateral is a parallelogram if both pairs of opposite angles are equal.

Theorem 39 → A quadrilateral is a parallelogram if its diagonals bisects each other.

Voltando ao estudo: Ⓐ

Triangle Inequality

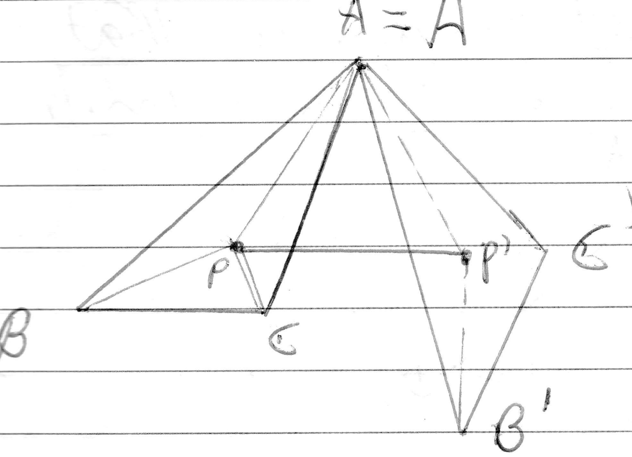

Now, $\angle BAC' = \angle BAC + \angle CAC' = 90°$, so $BC' = 5$ by Pythagoras' Theorem. Therefore, the value of $PA + PB + PC$ is at least 5, and the value will be exactly 5 if B, P, P', and C' are collinear.

Thus, to attain the minimum value, P and P' would have to lie on BC'.

If this happens, then

$\angle APB = 180° - \angle P'PA = 120°$ and

$\angle CPA = \angle C'P'A' = 120°$.

Now, this defines the point P uniquely, and since all three angles of triangle ABC are less than 120°, P is indeed inside Δ ABC.

Pág. 225(240) - Example 7.3.8.

A and B are the centers of two circles on the same side of line l.
Construct a line m, parallel to l, such that the circles cut off equal segments from m.

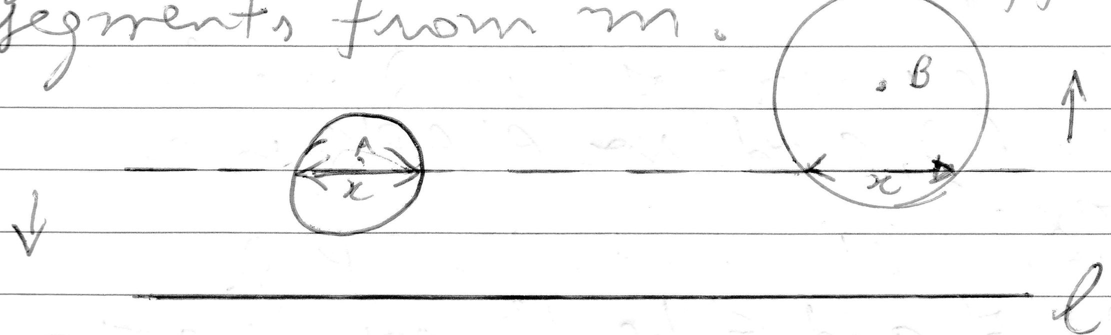

Solution: Let α and β denotes circles centered at A and B, respectively.
Let P and Q be the feet of the perpendiculars from A and B upon l, respectively. The solution here is obtained by using the translation T_{PQ}.
Under this translation,
the circle α is mapped to the circle α' centered at A', as his figure below.

Veja figura

⟹

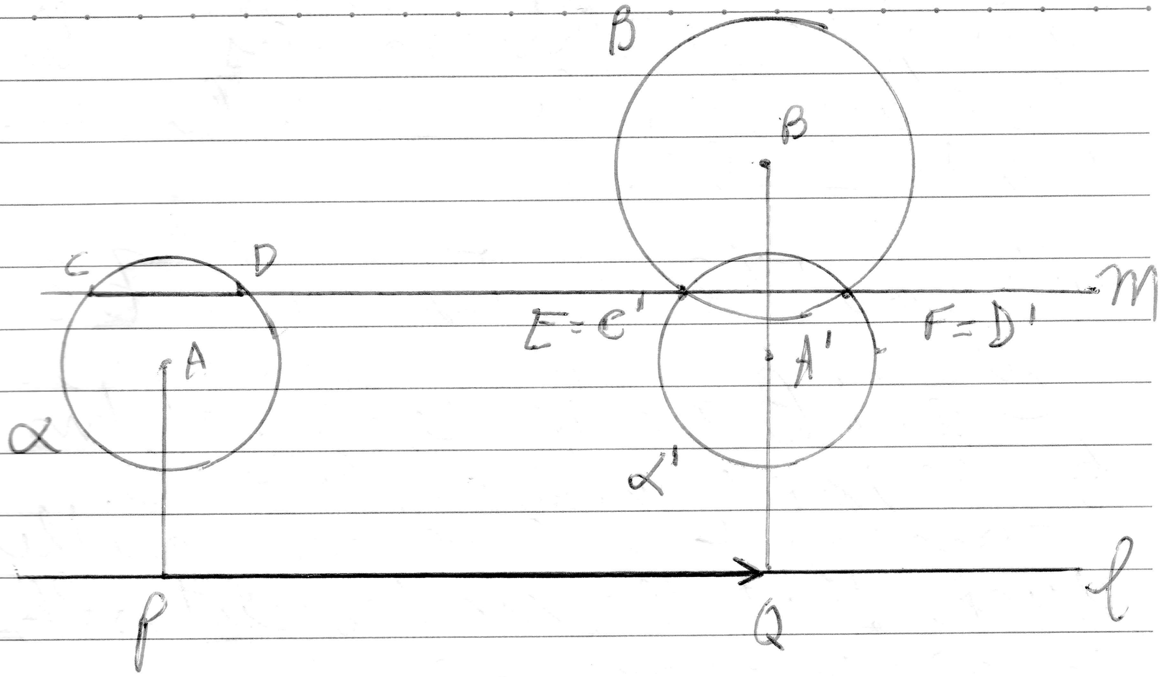

Since the image of P is Q, we have $AA' = PQ$, and so $A'AP'Q$ is a rectangle.

Thus, A' must be on BQ.

Let E and F be the points at which the translated circle intersects β, and let m be the line EF, cutting α at C and D.

Then $E = C'$ and $F = D'$, so that

$$EF = C'D' = CD.$$

5) 031 --- Curvas Elipticas (Ivora, Castillo Carlos).pdf

→ 475 pp.

→ livro todo!!!

Pag. 34 (49) — La curva de Fermat.

6) 032--- Curved Spaces from Classical Geometries...pdf

Pág. 75 - Chapter 4
Riemannian metrics.

Pág. 153 (165) - Chapter 8
Abstract surfaces and Gauss-Bonnet

Obs: Important: Pág. 02 (14) ↙

The triangle inequality therefore states that

$$(x+y, x+y)^{1/2} \leq \|x\| + \|y\|;$$

on squaring and expanding, this is seen to be equivalent to the

CAUCHY-SCHWARZ INEQUALITY.

7) 033---Elementary Differential Geometry.pdf
→ 335 pp.

Pág. 22 (36) - Chapter 2 - Curve Theory.

Pág. 31 (45) - último parágrafo.

The first inequality above holds (acima detém) because of the
"TRIANGLE INEQUALITY"

$| \|x\| - \|y\| | \leq \|x-y\|$.

Pág. 66 (80)

The spherical side lengths are α_i, β_i and γ_i.

The triangle inequality gives that

$$\alpha_i \leq \beta_i + \gamma_i \qquad (2.18)$$

Using the spherical triangle with the vertices

$$\frac{a_{i+1} - a_i}{\|a_{i+1} - a_i\|}, \frac{a_{i+1} - a}{\|a_{i+1} - a\|} \text{ and }$$

$$\frac{a_{i+2} - a_{i+1}}{\|a_{i+2} - a_{i+1}\|}$$

We obtain analogously:

$$\alpha_{i+1} \leq \beta_{i+1} + \gamma_{i+1} \qquad (2.19)$$

We summarise:

$$K(P_2) - K(P_1) = \beta_i + \beta + \beta_{i+1} - (\alpha_1 + \alpha_{i+1})$$

$$= \beta_i - \alpha_i) + \beta + (\beta_{i+1} - \alpha_{i+1})$$

$\geqslant -\gamma_i + \beta - \gamma_{i+1}$

$= 0,$

and hence $\kappa(P_2) \geqslant \kappa(P_1)$.

(b) Let us now assume that we are dealing with the equality case:

$$\kappa(P_2) = \kappa(P_1).$$

Then we must have equality in the triangle inequality (2.108),

$$\alpha_i = \beta_i + \gamma_i$$

Pág. 67 (81) $\Longrightarrow$

Thus the vertices of the spherical triangle lie on a great circle, i.e,

$a_i - a_{i-1}$, $a - a_i$ and $a_{i+1} - a_i$

lie in a two-dimensional subspace $V_1 \subset \mathbb{R}^3$. For suitable

$\lambda_1, \lambda_2 \in \mathbb{R}$ we have that

$$a_i - a_{i-1} = \lambda_1 \cdot (a - a_i) + \lambda_2 (a_{i+1} - a_i)$$

and thus

$$a_{i-1} = -\lambda_1 \cdot a - \lambda_2 \cdot a_{i+1} + (1+\lambda_1+\lambda_2)\cdot a_i \quad (2.20)$$

We now suppose that a_i, a and a_{i+1} are not collinear,

else (senso) we would be dealing with case (1). Hence those three points span an affine plane

$$E \subset \mathbb{R}^3.$$

Because of (2.20) we also have $a_{i-1} \in E$.

It can be shown analogously that

$a_{i+2} \in E$. This proves (2).

Pag. 71 (85) →

To estimate the third term of (2.27) we observe that the triangle inequality for the spherical distances gives the inequality

$$|\alpha_j - \beta_j| \le \gamma_{j-1} + \gamma_j \quad (2.29)$$

Pag. 266 (284) – Appendix A

(Hints for solution to (most) exercises

Pág. 271 (289) - 2.6

Let c be a curve connecting x and y which containing a point z that does not lie on the line segment xy.
Then the triangle inequality

shows that any polygon P having

x, z and y as vertices must have length

$$L[P] \geqslant \|x-z\| + \|z-y\| > \|x-y\|.$$

Now choose a sequence of polygons approximating c which contain

~~x, y on~~ x, z and y as vertices.

Proposition 2.18 then says

$$L[c] \geqslant \|x-z\| + \|z-y\| > \|x-y\|.$$

→ Aprofundar com o resto da prova.

(b) → $\mathbb{R}^n$ be a parametrized curve connecting x and y.

Obs. Pág. 272 (290) - Thus there is no curve connecting x and y which is shorter than the line segment.

8) 034 --- Euclidean Geometry in Mathematical Olympiads.pdf
→ 328 pp.

Pag. 131 (148) - this just follows from the triangle inequality

$a + b > 2b.$

Pag. 211 (228) - Problem 11.18

Let ABC be a non-equilateral triangle and let M_A, M_B, M_C be the midpoints of the sides BC, CA, AB respectively. Let S be a point lying on the Euler line. Denote by X, Y, Z the second intersections of $\overline{M_A S}, \overline{M_B S}, \overline{M_C S}$ with the nine-point circle. Prove that $\overline{AX}, \overline{BY}, \overline{CZ}$ are concurrent. HINTS 176, 182, 369, 546.

Pag. 71 (88) - Problem 4.50

Prove that the four lines A_0D, B_0E, C_0F, and OI are concurrent.

Solution:

Pag. 250 (267) - Solution 4.50

Obs: Euler line of $I_A I_B I_C$ (with nine-point center O), it passes through O.

20220829/// (20220828)

DATA-NovoVivo-017-2022/
Books-017-2022/
20220828/
Where Physics Went Wrong/

Copy: 035--L Lavenda - A new perspective
1) on relativity - an odyssey in non-Euclidean geometries...(2011).pdf

→ 691 pp.

Pág. 346(372) - 7.2.2.

Shapiro effect via Fermat's principle

If a light signal is sent from Earth, located on the x-axis at $-x_E$, to Venus, say, which is located behind the Sun at x_V, as shown in Fig. 7.1, the light ray will be bent (dobrado) as it passes the gravitational field of the Sun!

Clocks will thus be slowed down, and the time it takes the ray to bounce (saltar) off the surface of Venus and return to Earth will be longer than if the sun were not present.

The simplest mechanical analog of the index of refraction is $n = \sqrt{(1-4\phi/c^2)}$, where ϕ is the potential energy per unit mass. We will justify this choice shortly. Since the gravitational field of the Sun makes the medium optically denser, ϕ can be identified as the gravitational potential, $-GM/r$, where G is Newton's gravitational constant, M the mass of the Sun, and

$r = \sqrt{(R^2 + x^2)}$ is the distance from the center of the Sun to Venus, with R as the Sun's radius.

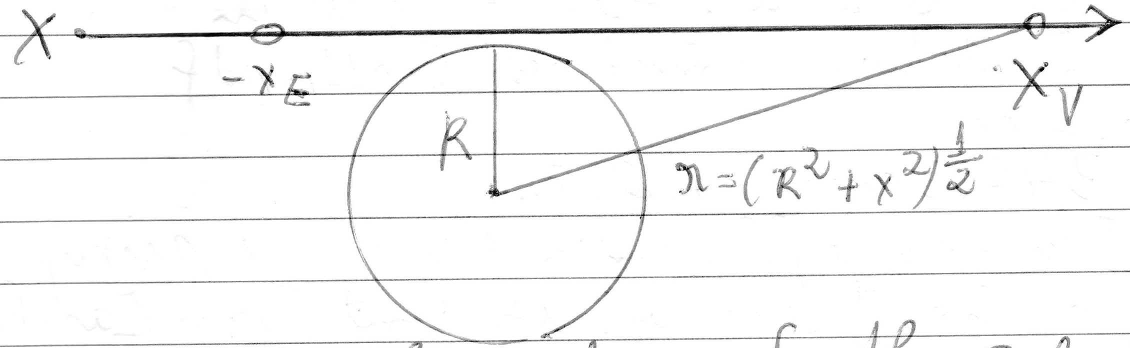

Fig. 7.1. The set-up for the ~~Spho~~ Shapiro effect.

Pág. 347(373)

Now, according to Fermat's principle, (7.2.3) the propagation time τ along a ray connecting the two end points $-x_E$ and x_V is given by

$$\tau = \int_{-x_E}^{x_V} \frac{\eta(r)}{c}\, dx = \int_{-x_E}^{x_V} \frac{\sqrt{(1+2\alpha/r)}}{c}\, dx,$$

where

$\alpha := 2GM/c^2$ is commonly referred to as the Schwarzschild radius.

obs: Livro todo!!!

2) Books-017-2022/20220828/
Advanced General Relativity Gravity
Waves, Spinning Particles.../
Copy: 036--- Advanced General Relativity
Gravity Waves... pdf
→ 153 pp.
Obs: Minkowskian space-time.
Plane gravitational waves.

Voltando: (Cont.)

9) ZIP-03/Books-Import-20220608/
Important-Books-011---2022/
Carlos-Ivorra/
Copy: 037--- Famous Problems of Geometry.pdf

Pág. 33(46) - Chapter V
The Problem of Trisecting an Angle.

Pág. 39(52) - Chapter VI
The Problem of Squaring the Circle.

10) 038--- Funciones De Variable Compleja.pdf
→ 443 pp - Carlos Ivorra Castillo.

11) 039--- Geometria (Yakovliev).pdf
→ 356 pp
Obs.: Cap I - Vectores en el plano y en
le espacio.

Cap. II - Rectas en el plano.

12) 040 --- Geometria (Ivorra...).pdf
→ 406 pp.
Obj: Livro todo!!!

Pág. 06(18) - 1.3 Ángulos y triángulos

See: Figure - Pág. 07(19)

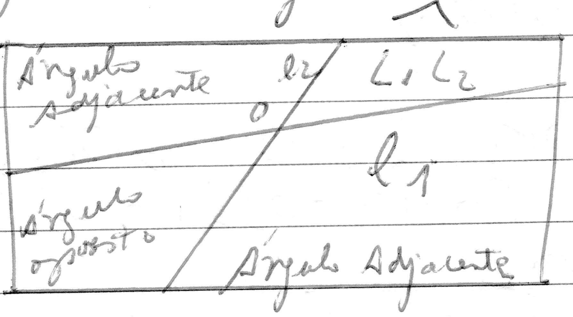
Ángulo adyacente
l_2
0
$L_1 L_2$
l_1
Ángulo opuesto
Ángulo Adyacente

Pág. 49(61) - Capítulo III

La geometría euclídea

See: Pág. 45(57)

Teorem 2.26 (Lagrange) La suma de los tres ángulos de un triángulo es menor o igual que π.

13) 041 --- Geometria Algebrica.pdf
→ 396 pp

14) 042 --- Geometry (David A. Brannan).pdf
→ 603 pp - Livro todo!!!

15| 043 --- Geometry and Topology.pdf
→ 215 pp.
→ Livro todo!!!

Obj: Euclidean geometry
_ Spherical and hyperbolic non-Euclidean geometry
_ Chapter 8 - Quaternions, rotations and the geometry of transformation groups
_ Chapter 9.3 - Geometry in physics.
_ B.4 - Symmetries.

Obs:
Chapter 3 - Pág. 34(53)
Spherical and hyperbolic non-Euclidean geometry
Pág. 37(56) - 3.2 Spherical triangles and trig.
Pág. 39(58) - 3.5 Properties of S^2 like E^2
Pág. 41(60) - Figure 3.6
Overlapping segments of S^2
3.7 - Preview of hyperbolic geometry
Pág. 42(61) - (4)
Symmetries are the set $O(2)$ of rotation and reflection matrixes

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \text{ and } \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}$$

3.8 - Hyperbolic space
Consider $\mathbb{R}^3$ with the Lorentz quadratic form $q_L(v) = -t^2 + x^2 + y^2$ (compare B.2).

The cone $\{q_L(v) < 0\}$ break up into two subsets

$\{t > +\sqrt{x^2+y^2}\} \cup \{t < -\sqrt{x^2+y^2}\}$.

I fix the positive choice $t > 0$ throughout.

Pag. 43(62) – 3.9 Hyperbolic distance

$-v \cdot_L w = \cosh d(P, Q)$, so that

$$d(P, Q) = \operatorname{arccosh}(-v \cdot_L w); \quad (4)$$

in other words,

$$d(P, Q) = \operatorname{arccosh}(t_1 t_2 - x_1 x_2 - y_1 y_2).$$

Lemma: the Lorentz inner product satisfies

$$-v \cdot_L w = t_1 t_2 - x_1 x_2 - y_1 y_2 \geq 1,$$

with equality only if $P = Q$.

→ See also Exercise 3.11.)

Hence the distance $d(P, Q)$ is defined and positive unless (a mos ser que) $P = Q$.

Pag. 49(68) – 3.13 The hyperbolic plane is non-Euclidean

Theorem: Let $\mathcal{L}$ be a hyperbolic line and P a point not lying on $\mathcal{L}$.
Then there exists a unique perpendicular line PQ to $\mathcal{L}$ through P.
Moreover,

(1) if M is orthogonal to PQ in P, then the lines $\mathcal{L}$ and M diverge;

(2) there exists (exists) an angle $\theta < \frac{\pi}{2}$ with the property that if $\mathcal{L}'$ is a line through P, then $\mathcal{L}'$ meets $\mathcal{L}$ if and only if the angle of $\mathcal{L}'$ and PQ is less than θ
(See Figure 3.13.)
↳ Pag. 50 (69).

Figure 3.13 →
the failure of the parallel postulate in $\mathcal{H}^2$.

Pag. 142 (161) – Chapter 8

Quaternions, rotations and the geometry of transformation groups

Pag. 143 (162) Figure 8.0

The geometry of the group of planar rotations.

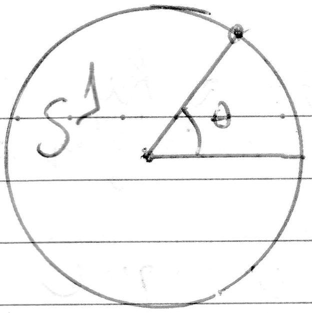

$$\begin{pmatrix}\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta\end{pmatrix}$$

Pág. 164(183) - Chapter 9
Concluding remarks

Pág. 166(185) - Figure 9.1a

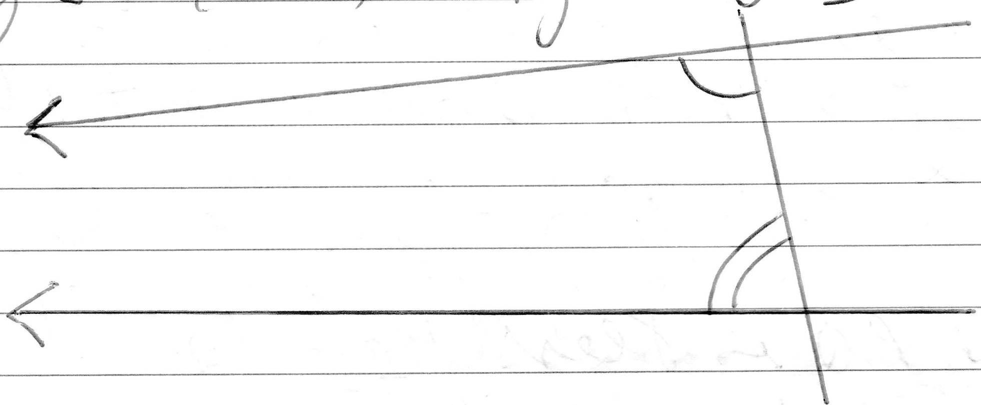

The parallel postulate. To meet or not to meet?

Pág. 167(186) - Figure 9.1b
The parallel postulate in the Euclidean plane.

Pág. 168(187) Figure 9.1c
The "parallel postulate" in spherical geometry.

Pág. 172(191) - 9.3 Geometry in physics

The Galilean group and Newton dynamics

Transformations to a different coordinate system are therefore given by maps

$(x, t) \mapsto (Ax + gt + b, t + s)$,

where $A \in O(3)$ is a 3×3 orthogonal matrix,

g and b are 3×1 column vectors, and $s \in \mathbb{R}$ is a scalar.

Obs: Pág. 173(192) — 9.3.2

The Poincaré group and special relativity

Obs: Veja também a tabela 9.3 ✓

Symmetries and conservations laws.

Pág. 177(196) - 9.4 The famous trichotomy

9.4.1 - The curvature trichotomy is

geometry.

Pag. 178(197) - Figure 9.4a and 9.4b

→ interessante!!!

Pag. 184(203) - B.2
Euclid and Lorentz.

Pag. 186(205) - B.4 - Symmetries.
B.5 - Orthogonal and Lorentz matrixes.

16) Pasta 044 --- Introduction to Non-Euclidean Geometry Wolfe/. ⇒ 12 files

a) 044 - Experiencing Geometry(2004).pdf
→ 405 pp. Livro todo!!!
Muito bom!!!
Obs: Chapter 8 - Parallel Transport
Chapter 9 - SSS, ASS SAA and AAA.
Chapter 10 - Parallel Postulates.

b) 044 - Henderson - the Fourth Dimension...pdf
→ 756 pp. Livro todo!!! - ART.
Pag. 12(40) - Flatland.

c) 044 - the Foundations of Geometry (1975).pdf
→ 524 pp. Livro todo!!!
Pag. 334(347) - Chapter 26 - Parallels and the ultimate Axiom.

20220829/// (cont.)
ZIP-03/ Books-Import-20220608/
Important-Books-011-2022/ Carlos-Ivorra/

1) 045---Introduccion a la Topologia General.pdf
→ 149pp.

2) 046---Introduction to geometry (Coxeter).pdf
Pág. 303(320) - 16.8
The Horosphere and the Euclidean Plane.
F. L. Watiter (Wachter) (1792-1817)...
in a letter to Gauss (Dec., 1816)...
speaks of the surface to which a sphere tends as its radius approaches infinity...
He affirms that even in the case of the FIFTH POSTULATE being false, there would be a geometry on this surface identical with that of the ordinary plane.
R. Bonola (1, pp. 62-63).

Pág. 307(324) - Chapter 17
Differential geometry of curves.

3) 047---Coxeter---The Real Projective Plane
(1949).pdf
→ 203 pp. Livro todo!!!

4) 048---Introduction to Metric and Topological
Space..pdf
→ 220 p.p.
Pág. 77(90) - Chapter 7 - Topological Spaces.

5) 049---Introduction to Projective Geometry.pdf
→ 480 pp.
6) 050--- Introduction to Topology Pure and Applied.pdf
→ 508 pp.
7) 051--- Linear Algebra Done Right.pdf
→ 352 pp. Livro todo!!!
8) 052--- Manifolds, Tensor and Forms.pdf
→ 343 pp.
Pag. 08(22) - 1.6 The dual space
9) 053--- Metric Spaces.pdf
→ 318 pp. Livro todo!!!
Pag. 21(40) – Chapter 2 – Distance.

"Mathematics is the tool specially suited for dealing with abstract concepts of any kind and there is no limit to its power".

Paul Dirac, 1902-1984.

→ Metrics are designed to measure distances between points.

Pag. 24(43) - 2.3.

Inequalities for Distances. ⇙

Metrics satisfy the triangle inequality and its variants.

Theorem 2.3.1

Suppose X is a metric space, $x \in X$ and A and B are non-empty subsets of X for which $A \subseteq B$. Then $\mathrm{dist}(x, B) \leq \mathrm{dist}(x, A) \leq \mathrm{dist}(x, B) + \mathrm{diam}(B)$.

10) 054 --- Non-Euclidean geometry (Coxeter).pdf
→ 349 pp. Livro todo!!!
11) 055 --- Perpectives on Projective Geometry.pdf
→ 571 pp.
Pág. 03 (24) - Chapter 01
Pappos's Theorem: Nine Proofs
and Three Variations.
Pág. 329 (335) - Chapter 18.
Euclidean Geometry.
Pág. 330 (336) - 18.1.
The points I and J.
12) 056 --- Projective Geometry.pdf
→ 138 pp.
Pág. 113 (123) - Chapter VI.
Metrical Geometry.
13) 057 --- Projective geometry an
introduction.pdf
→ 211 pp. Muito bom!!!
Pág. 15 (28) - Chapter 2 - Introduction
The Desargues' and Pappus' configurations
14) 058 --- Steven G. Krantz -
Geometric Function Theory.pdf
→ 307 pp.
15) 059 --- Tensors the Mathematics of Relativity
theory and Continuum Mechanics.pdf
→ 300 pp. Livro todo!!!
Pág. 121 (132) - Chapter 5
Riemannian and Pseudo -
Riemannian Manifolds
→ 5.1 - Metric Tensor, Christoffel

Symbols, and Ricci Rotation Coefficients.

16) 060---Teoria de Cuerpos De Clases.pdf

→ 496pp.

17) 061--Teoria de Numeros.pdf

→ 376pp.

Pag. 77 (87) → Capitulo IV
Métodos geométricos.

→ Minkowski.

Pag. 83 (93) - 4.3
El teorema de Minkowski

⇘ 18) 062---Textbook of Tensor Calculus and Differential Geometry.pdf

→ 550 pp. Leído todo!!

Pag. 204 (215) - Chapter 4.
Riemannian Geometry

Pag. 264 (275) - Chapter 5.
Geometry of Space Curve

Pag. 306 (317) - Chapter 6.
Intrinsic Geometry of Surface

Pag. 378 (389) - Chapter 7.
Surfaces in Space.

Pag. 412 (423) - Chapter 8.
Curves on a Surface.

19) 063---The geometry and topology of Coxeter groups.pdf

→ 600 pp.

Pag. 499 (515) - Appendix I -
Nonpositive Curvature

I.1 - Geodesic Metric Spaces.

20) 064---Topologia Algebraica.pdf
→ 499 pp.
21) 065-I- Topologia Algebraica.pdf (res)
22) 066--- Topology through Inquiry.
→ 330 pp.
Pág. 119 (135) - Chapter 9
Metric Spaces: Getting some Distance
$M \times M \to R_+$
(where R_+ is the non-negative real numbers).
23) 067 --- Visual Complex Analysis.pdf
→ 613 pp.
Pág. 267 (288) - Chapter 6
Non-Euclidean Geometry.
Pág. 268 (269) -
Parallel Axiom.
Through any point p not on the line $\mathcal{L}$ there exists precisely one line L' that does not meet L. (1)
Spherical Axiom.
There is no line through p that does not meet $\mathcal{L}$. (2)
or
Hyperbolic Axiom.
There are at least two lines through p that do not meet $\mathcal{L}$. (3)
Obs: Figure [1].
Pág. 269 (290) - 2 -
Some Facts from Non-Euclidean Geometry.

(Angle sum of T) = π.

Let's take our first look at how these new geometries differ from Euclid's.

A very familiar theorem of Euclidean geometry states that in any triangle T.

$$\text{Angle sum of } T = \pi.$$

As indicated in [16], this result is actually equivalent to the

parallel axiom.

It follows that in non-Euclidean geometry the angle sum of a triangle differs from π. To measure this difference, we introduce the so-called angular excess E:

$$E(T) \equiv (\text{Angle sum of } T) - \pi.$$

Euclidean geometry is thus caracterized by the vanishing of E(T).

Note that, unlike the original formulation of the parallel axiom, this statement can be checked against experiment:

construct a triangle, measures its angles, and see if they add up to π.

Gauss was the first person to ever conceive of the possibility that physical space might not be Euclidean, and he even (late) attempted the above experiment, using three mountain tops as the vertices of his triangle, and using light rays for its edges.

Within the accuracy permitted by his equipment, he found $\underline{\underline{E = 0}}$.

Quite correctly, GAUSS did not conclude that physical space is definitely Euclidean in structure, but rather (em vez de) that if it is not Euclidean then its deviation from Euclidean geometry is extremely small.

Let us return from physics to mathematics. Using pure logic to work out the consequences of (2) and (3), both Gauss and Johann Heinrich Lambert independently discovered that the two non-Euclidean geometries departed from Euclid's in opposite ways:

- In spherical geometry the angle sum is greater than π: $E > 0$.

- In hyperbolic geometry the angle sum is less than π: $E < 0$.

Furthermore, they discovered the striking (impressionante) fact $E(T)$ is completely determined by the size of the triangle.
More precisely, $E(T)$ is simply proportional to the area $A(T)$ of the triangle T:

$$E(T) = k\,A(T),$$

where k is a constant that is positive in spherical geometry, and negative in hyperbolic geometry!

24) 068--- Visual Differential Geometry and Forms...pdf
→ 531 pp. Livro todo!!!
→ Interessante!!!
Pág. 97(126) – Curvature
Chapter 8 – Curvature of Plane Curves.
Obs.: Estudar com atenção!!!

25) Important-Books-011-2022/Circles A Mathematical View/
Copy: 069---Circles A Mathematical View Ed2.pdf → 138 pp. Livro todo!!!

20220905 III

ZIP-03/ Books-Important-20220608/
Important-Books-011---2022/

1) Continued-Fractions/
Copy: Volume-021---20220818/Original/
070---Exploring Continued Fractions..pdf
→ 503 pp.
Pag. 39 (61) - The division algorithm generates continued fractions
Pag. 40 (62) - Figure 6
Leibniz's notebook.
2) Foundations of Space and Time Reflections on Quantum Gravity/
2a) Electromagnetics/
2a1) 071---Classical Electromagnetism.pdf
2a2) 071---Electromagnetic Field Theory Fund...pdf
2a3) 071---Elements of Electromagnetics-Sadiku.pdf
2a4) 071---Elements of Electromagnetics.pdf
2a5) 071---Engineering Electromagnetics, 8th Ed.pdf
2b) Theory of Electromagnetic Wave Propagation/
2b1) 072---David J. Griffiths - Introduction to Electrodynamics ...(2017).pdf
Pag. 06 (24) - Example 1.2.
Find the angle between the face diagonal of a cube.
2b2) 072---Ishimaru - Electromagnetic Wave Propagation ...pdf
→ 945 pp.
Pag. 670 (686) - 20.5 - Photon Density Waves.

2b3) 072---Methods in electromagnetic wave
propagation...pdf
→ 666 pp.
2b4) 072---Papas - Theory of electromagnetic
Wave propagation...(1988).pdf
→ 270 pp.
2c) 073---Approaches to quantum gravity...pdf (605pp)
2d) 073---Flat and Curved Space-Times.pdf (383pp)
2e) 073---Foundations of Space and Time Reflections
on Quantum Gravity.pdf (452pp) → todo!!!
2f) 073---Geometrical Methods of Mathematical Physics.pdf
260pp - Livro todo!!!
2g) 073---Geometry and Topology.pdf (215pp)
2h) 073--- Geometry, Topology and Physics.pdf (596pp) - Todo!!!
2i) 073---Hans Stephani - Relativity...pdf (418pp) -
Galilei Transformations - Livro todo!!!
2j) 073---Is the Universe Open or Closed...pdf
253pp - Livro todo!!!
2k) 073--- Rafael ...- Einstein's space-time...pdf
322pp - Livro todo!!!
2l) 073--- Relativistic cosmology.pdf
638pp - Livro todo!!!
2m) 073--- Relativistic gravitation and cosmology.pdf
307pp - Livro todo!!!

Source: All books: ZIP-03---20220905.7z
81.4Gb - 10.280 files
Término: 202209051353 São Paulo-SP.
Author: Prof. Dr. Stenio Musich -
www.steniomusich.com
www.matematica.org

FORONI

WWW.FORONI.COM.BR

200 Folhas/Hojas - 10 Matérias
Formato/Tamaño: 200 mm x 275 mm

Ref. 03.3317-3

7 908077 430313

NBR
15733:2012

FSC
www.fsc.org
MISTO
Papel produzido
a partir de
fontes responsáveis
Papel procedente de
fuentes responsables
FSC® C007274

CNPJ 61.283.636/0001-82 • Indústria Brasileira / Made in Brazil / Hecho en Brasil
VALIDADE INDETERMINADA • VALIDEZ INDEFINIDA

Composição: Papelão, Papel Couché, Papel Offset 56 g/m^2 e Arame Revestido em Nylon
Composición: Cartón, Papel Couché, Papel Offset 56 g/m^2 y Alambre Revestido en Nylon

Importado por: **ACCO Mexicana, S.A. de C.V.:** Circuito de la Industria Norte No. 6 - Corredor Industrial Toluca Lerma, Lerma México C.P. 52004 - R.F.C. AME6402291W3 • **ACCO Brands Chile S.A.**

Stenis Munich. 2023042918 10.

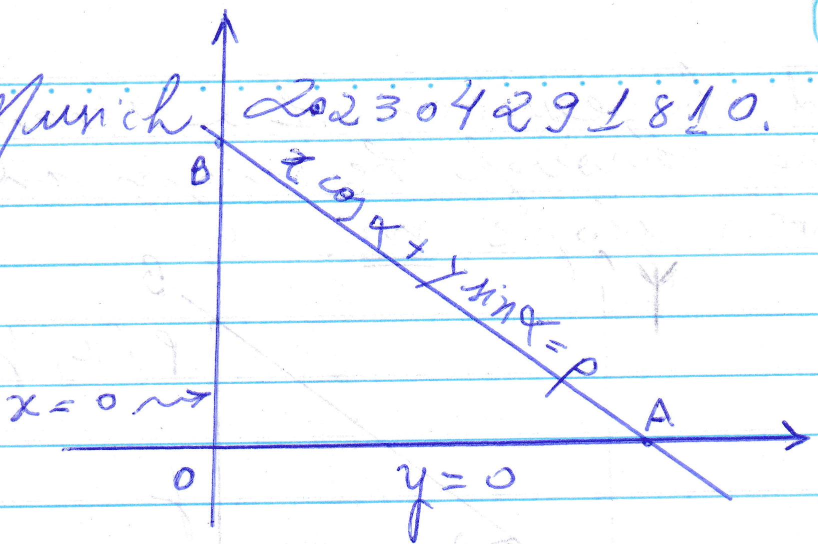

- equation of the line is $x\cos\alpha + y\sin\alpha = p$ (1)
- coordinate axes are $x = 0$ and $y = 0$. (2)
- coordinate of O are $(0, 0)$. (3)

Solving 1 and 2,

a) coordinate of A are $\left(\frac{p}{\cos\alpha}, 0\right)$

b) coordinate of B are $\left(0, \frac{p}{\sin\alpha}\right)$

↳ solving 1 and 3

Also:

$$AB = \sqrt{\left(\frac{p}{\cos\alpha} - 0\right)^2 + \left(0 - \frac{p}{\sin\alpha}\right)^2} =$$

$$= \sqrt{\frac{p^2}{\cos^2\alpha} + \frac{p^2}{\sin^2\alpha}} =$$

$$= \sqrt{\frac{p^2(\sin^2\alpha + \cos^2\alpha)}{\cos^2\alpha \sin^2\alpha}} =$$

$$\frac{p}{\cos\alpha \sin\alpha}$$

$OB = \frac{p}{\sin\alpha}$, $OA = \frac{p}{\cos\alpha}$

/ /

Find the equation of a straight line passing through the origin and having a given slope m.

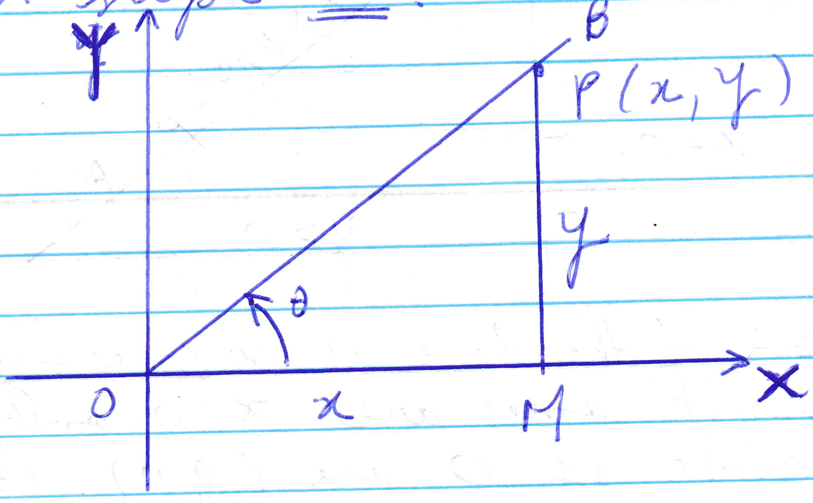

∴ OB → straight line through O making an angle θ with the positive direction of x-axis.

Then its slope $m = \tan\theta$

Let $P(x, y)$ be any point on the line.

Draw $PM \perp$ x-axis so that

$OM = x$, $MP = y$.

From the right angled $\Delta\, OMP$, we have

$$\tan\theta = \frac{MP}{OM} \quad \text{or} \quad \frac{y}{x} = m$$

$$\therefore \quad y = mx$$

Research: Including Pythagoras theorem

$a^2 + b^2 = c^2$ (retangle triangle)

obs: Inverse

equation:
$p = x - \cos\alpha + y \cdot \sin\alpha$

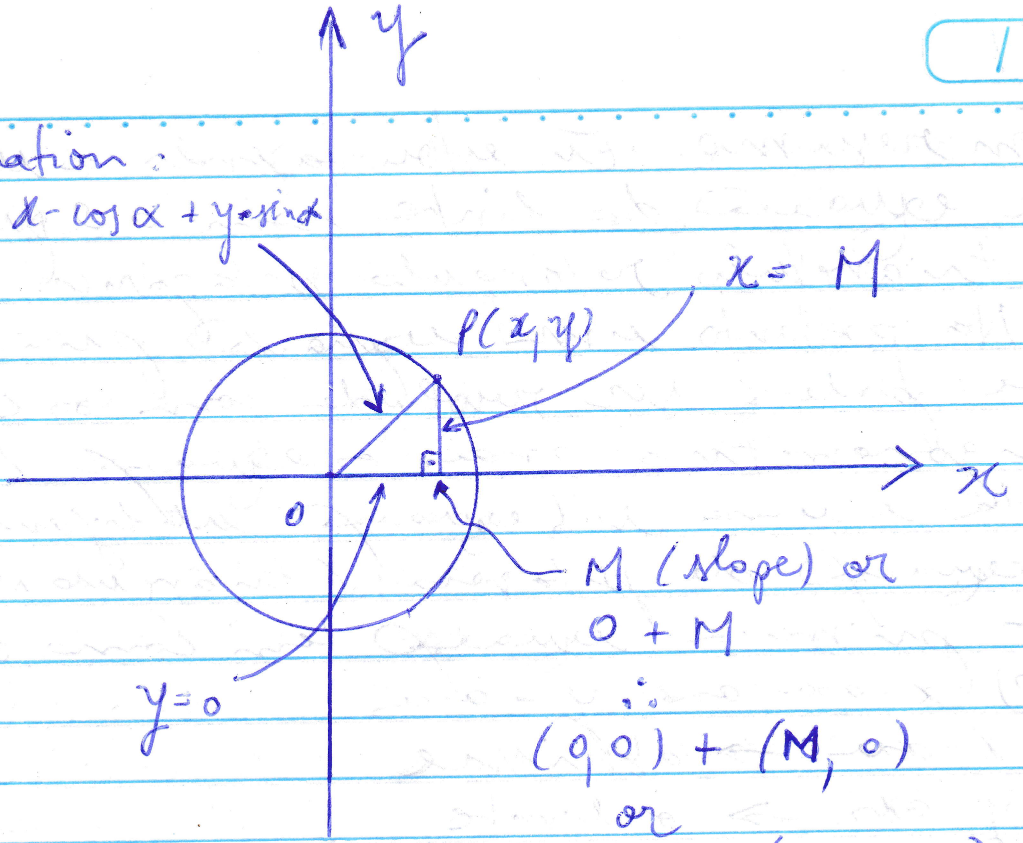

M (slope) or
$0 + M$
$\therefore$
$(0, 0) + (M, 0)$
or
$(x_1, y_1) + (x_2, y_2)$

$\therefore (x_1, y_1) + (M, y_2)$

distance $\overline{OM}$ =

$\sqrt{(x_1 - M)^2 + (y_1 - y_2)}$ obs: slope

$\therefore$

$O = A$ and $P = B$.

See: $O = A$.
$P = B$.
$M = O + M$ (slope) $\therefore$ equation

$p = x - \cos\alpha + y \cdot \sin\alpha$ obs: checar $\begin{cases} y - \sin\alpha \\ \text{or} \\ y \operatorname{sen}\alpha \end{cases}$

Δ OMP is Pythagoras triangles.

SÃO DOMINGOS

Em resumo: Eu estou fazendo a unificação da equação da linha com a equação do triângulo retângulo pitagórico.

Há portanto, uma equação principal que poderá ser resolvida com os cálculos trigonométricos mais a equação da linha e uma subequação utilizando-se a equação de pitágoras. (triângulo inverso)

A primeira equação tem como origem (0,0) x-axis and y-axis

∴ x-axis → abscissae

y-axis → ordinate

A segunda equação em relação ao ponto M, origem do ângulo reto (90°).

Interessante:

Tive uma idéia:

a) criar equações trigonométricas em relação ao ponto M.

Obs: objetivo é fazer calculos complementares em relação aos cálculos trigonométricos tradicionais.

a_1) primeiras equações em relação ao ponto O (0,0)

a_2) segundas equações em relação ao ponto M (x_2, y_2)

Importante:

Estas equações serão úteis para manter a integridade dos cálculos aplicáveis à curvatura tempo-espaço de Einstein.

↳ dinâmica.

Falando em pefeit Einstein vejamos:

A unificação da teoria da relatividade e da teoria quântica depende do ajuste das equações à nova realidade considerando a gravidade como efeito das forças fundamentais da física.

a) eletromagnetismo;
b) força forte e
c) força fraca.

∴ A gravidade não existe!!!

Tanto é verdade, que no ambiente quântico a gravidade é desprezada por ser inexistente e o que há é a existência das forças fundamentais da física aplicáveis ao ambiente quântico.

Vejamos a seguinte equação:

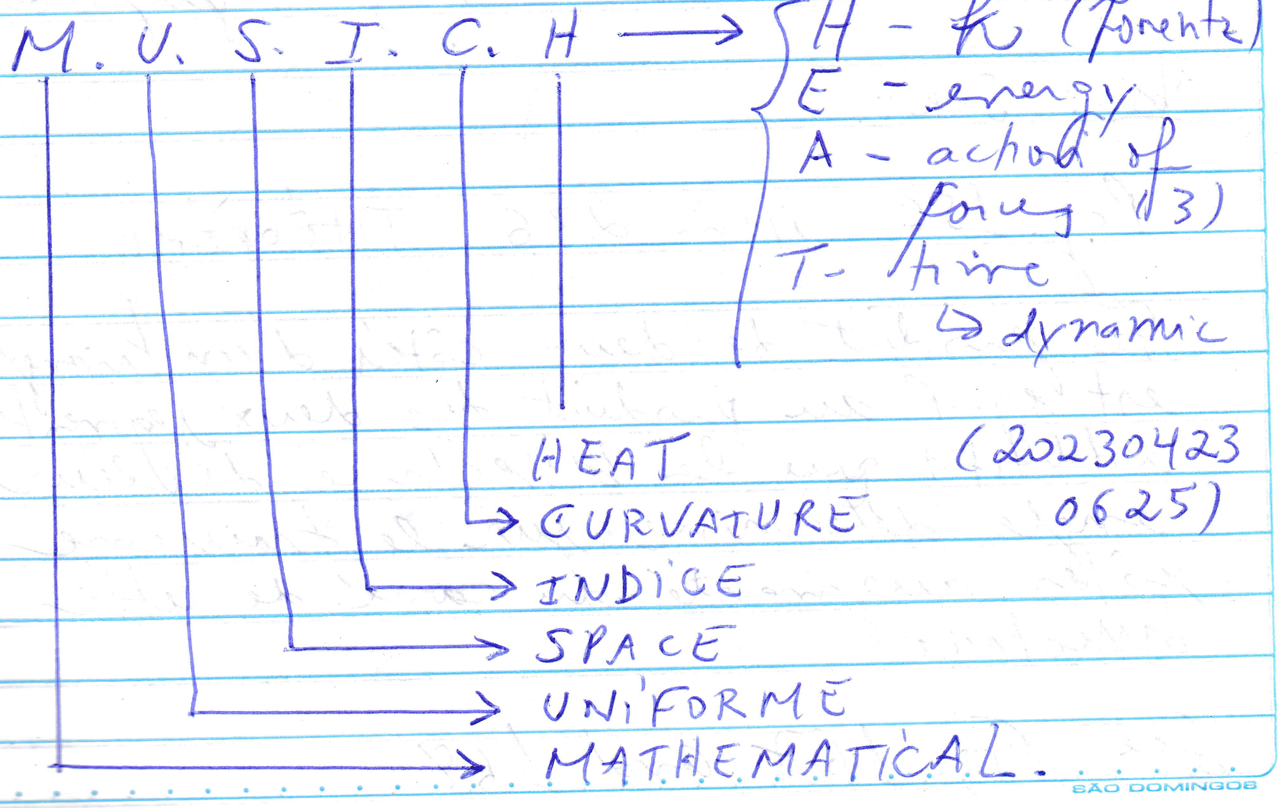
M. U. S. I. C. H
H - h (Lorentz)
E - energy
A - achord of forces (3)
T - time
dynamic
HEAT
CURVATURE
INDICE
SPACE
UNIFORME
MATHEMATICAL.

(20230423 0625)

SÃO DOMINGOS

Interessante:

Juntar nesta pesquisa:

a) achar o centro do círculo usando retas.

obs: Eu tenho no livro de geometria 2 volumes → Traité de Géométrie (Rouché and Comberousse).

b) Traité de Géométrie - Rouché
Pág. 147
§ IV. Relations Métriques entre les différentes parties d'un triangle.

221. On appelle projection d'un point A ...
↓
see: orthogonal projection.

obs.: See:

Pág. 156 item 236 THÉORÈME.

Le produit de deux côtés d'un triangle est égal au produit des deux segments additifs que la bissectrice de leur angle détermine sur le troisième côtés augmenté du carré de cette bissectrice

See Fig. 167 and 168.

c) Pág. 535 - Vol. 1.

XXVII → Trouver la quatrième proportionnelle X (proportionnelle X) à trois lignes données M, N, P (fig. 173, p. 163) $X = \frac{NP}{M}$.

↓

Profline (Pág. 163)

246. Trouver la quatrième proportionnelle à trois droites données M, N, P.

Obs: Important:

Paralelos

a) usar theorem of parallels.

b) Princípio de Cavalieri
→ planos paralelos, as duas figuras têm a mesma área.

Obs: checar essa simulação em um plano curvo
veja: área e projeção ortogonal.

SÃO DOMINGOS

1 1

Pappus ↗ outro autor.

c) três pontos. ___ (reta)

Apollonius.

Obs.: Pág. 67. 116 Pág. 297 (?)

Théorème

116. Par trois points A, B, C, non situés en ligne droite, on peut toujours faire passer une circonférence, et l'on ne peut en faire passer qu'une (fig. 80)

Pág. 153

Lieu des points dont la somme des carrés des distances à deux points fixes est constante.

⇒ usar.

Pág. 200

§ VIII - Mesure de la circonférence.

Pág. 219 - Théorie des Projections

Pág. 277 - Cercles orthogonaux

Pág. 306 - Cercles des Neuf Points

(404). → incluir para

checagem da teoria.

d) Veja: (See)

Pag. 97 – item 160 Problème

160. Inscrire un cercle dans triangle donné ABC.

Obs: Usar para checagem.

Pag. 98 – 161.

161. Il résulte de cette démonstration que les bissectrices des trois angles d'un triangle concourent en un même point.

Obs: Podemos simular isto com 3 círculos curvos
ou
3 planos e o do outro
curvo.

Obs: Para ver as discrepâncias dos cálculos.

other: Pag. 83
See:

SÃO DOMINGOS

/ /

circonférence

360 . 60 = 21600 minutes

21600 . 60 = 1.296.000 seconds.

obs: 36° 15' 21" = 130521 seconds

90° = 324.000 seconds.

$$\frac{130521}{324000} \text{ ou } \frac{43507}{108000}.$$

e) estudar:

Pág. 473 — 36.

Points, Cercle et Triangles de Brocard.

f) Pág. 437 - Note II

Sur l'impossibilité de la quadrature du cercle.

g) Pág. 363 - III

Sur les maximums et les minimums des figures planes.

Interessante: (see) Ⓘ

25. Geometria - Posição, Métrica e

Analítica – Curso Objetivo
Livro 25
Mauro Gruspan, Jorge Kritomian

Pag. 157 – nº 16

Determinar a equação da hipérbole cujos focos estão situados simetricamente em relação à origem e no eixo das ordenadas, sabendo que sua excentricidade vale $\frac{13}{12}$

↳ valor ajustável à pesquisa

e que a distância entre seus vértices vale 48 unidades

↳ ajustável à necessidade

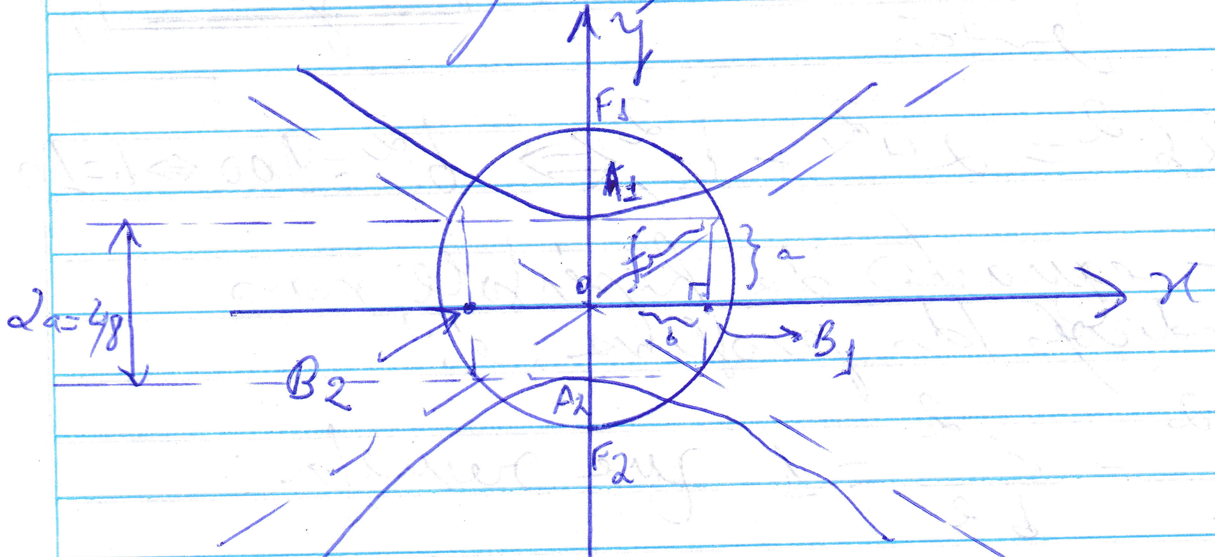

Obs: Esta hipérbole pode representar a curvatura do espaço, acrescentado às equações trigonométricas tradicionais com a pitagórica (inversa).

→ início deste texto..

Calculando:

A partir do enunciado, temos:

Centro O (0,0)

- Eixo transverso: $2a = 48 \Leftrightarrow a = 24$
- Excentricidade:

$$e = \frac{f}{a} = \frac{13}{12} = \frac{f}{24} \Leftrightarrow f = 26$$

Como $f^2 = a^2 + b^2$ (Pitágoras), então

$$26^2 = 24^2 + b^2 \Leftrightarrow b^2 = 100 \Leftrightarrow b = 10$$

A equação da hipérbole, nas condições do problema, é:

$\frac{y^2}{a^2} - \frac{x^2}{b^2} = 1$, que resulta:

(seg)

$$\frac{y^2}{24^2} - \frac{x^2}{10^2} = 1 \Leftrightarrow \frac{y^2}{576} - \frac{x^2}{100} = 1$$

Ⅱ Matemática para o 2º grau. Youssef 1994.

Pág. 324 - Equação geral da reta

$$ax + by + c = 0$$

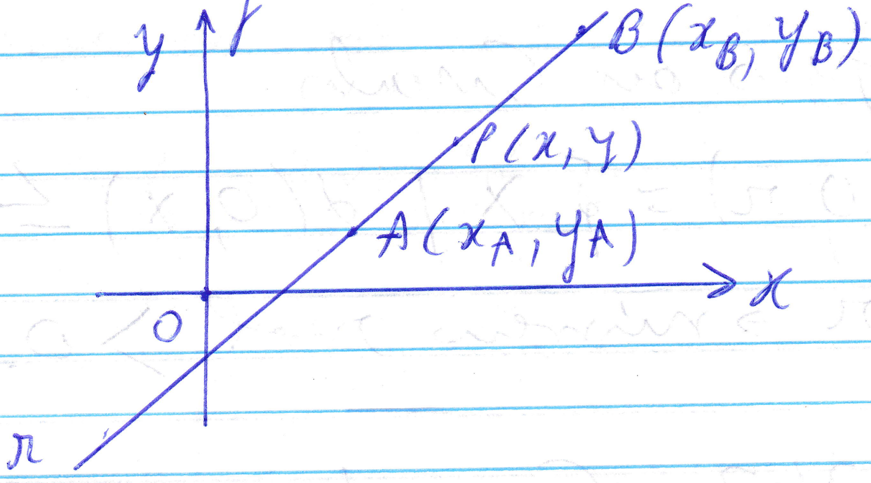

(3 pontos / reta)

$$\begin{vmatrix} x & y & 1 \\ x_A & y_A & 1 \\ x_B & y_B & 1 \end{vmatrix} = 0 \Rightarrow ax + by + c = 0$$

Exemplo:

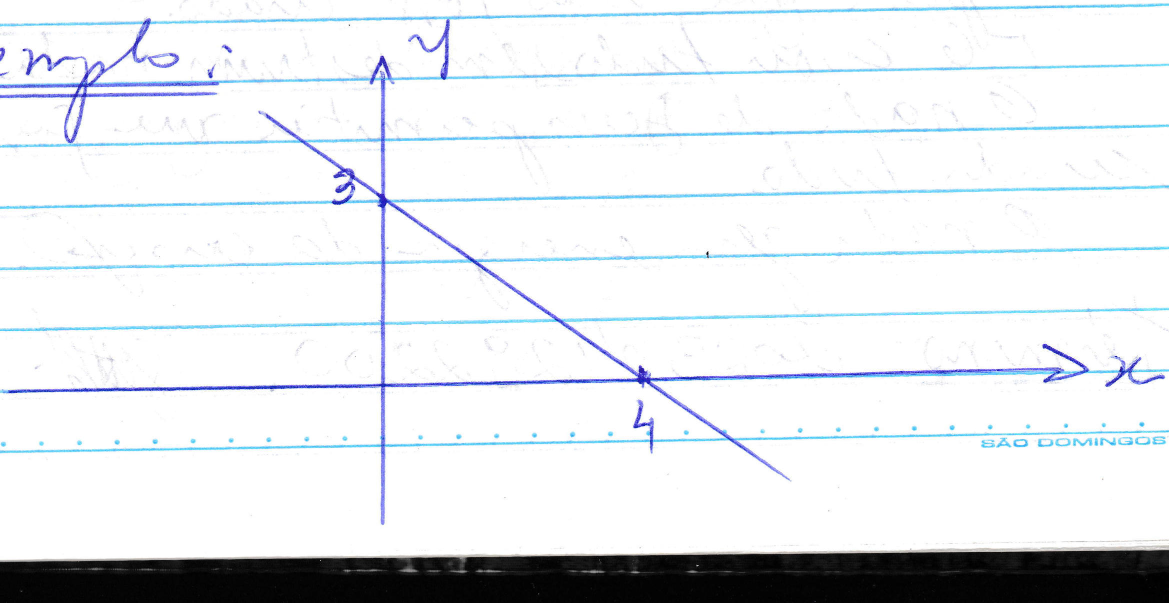

SÃO DOMINGOS

$$\begin{vmatrix} x & y & 1 \\ 0 & 3 & 1 \\ 4 & 0 & 1 \end{vmatrix} = 0$$

$(r)\ 3x + 4y - 12 = 0$

(III) LISA – Biblioteca da Matemática Moderna
volume 1

Pág. 240 –

Disco ou Círculo

$$D(O, r) = \{ X \mid d(O, X) \leq r \}$$

Obs: $r \rightarrow$ número real > 0.

Pág. 342 – Geometria no Círculo.

202304270745

Deus é tão bom e maravilhoso que criou o nada para que tudo fosse criado.

Ele criou tudo sem destruir nada.

O nada de Deus permitiu que fosse criado tudo.

O nada gera energia da criação.

Término: 202304292200.

continuando (mesma hora) 202304292200
Esqueci de terminar:

① Livro Co-ordinate Geometry

Pág. 64(67) - Example 5

Show that the distance of $P(x_1, y_1)$ from the line $ax+by+c=0$ in the direction making an angle θ with x-axis is

$$-\frac{ax_1+by_1+c}{a\cos\theta+b\sin\theta}$$

Solution:

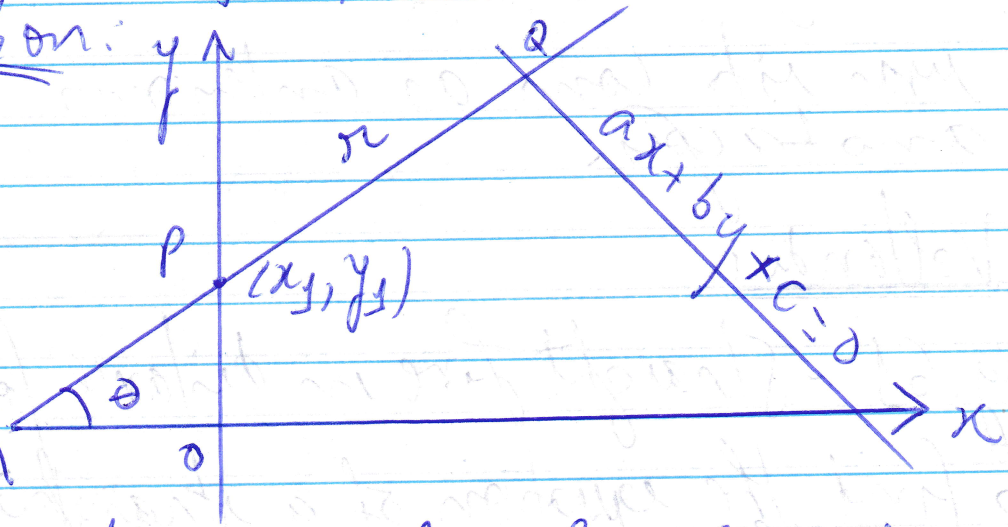

Let the line through $P(x_1, y_1)$ and inclined at an angle θ with x-axis meet the line $ax+by+c=0$ in Q

Let $PQ = r$. We have to find r.

Co-ords. of ∴ Q are $(x_1 + r\cos\theta, y_1 + r\sin\theta)$

SÃO DOMINGOS

Q lies on $ax + by + c = 0$

$\therefore \; a(x_1 + r\cos\theta) + b(y_1 + r\sin\theta) + c = 0$

or

$(ax_1 + by_1 + c) + r(a\cos\theta + b\sin\theta) = 0$

or

$r(a\cos\theta + b\sin\theta) = -(ax_1 + by_1 + c)$

Hence $r = \dfrac{ax_1 + by_1 + c}{a\cos\theta + b\sin\theta}$, which

is the required distance.

obs: usar isto com as anteriores anotações.

Voltando:

Pág. 61 - Straight line in distance form

To find the equation of a straight line passing through a given point (x_1, y_1) and making angle θ with the x-axis in the form

$$\frac{x - x_1}{\cos\theta} = \frac{y - y_1}{\sin\theta} = r$$

where r is the distance of any point (x, y) on the line from the point (x_1, y_1).

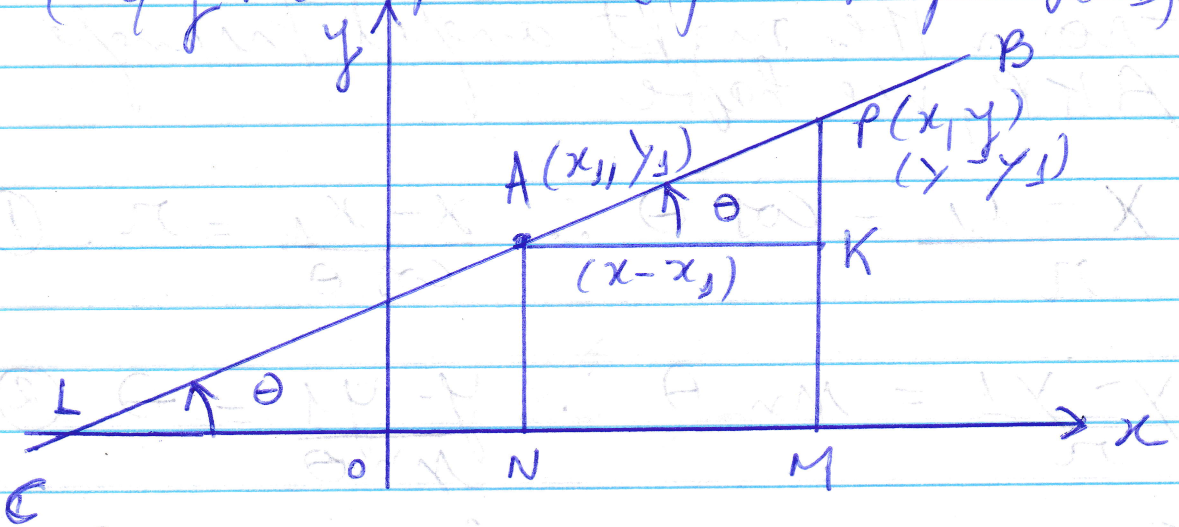

- Let $A(x_1, y_1)$ be the given point and AB the line through A making an angle θ with the x-axis.

- Let $P(x, y)$ be any point on line such that

$AP = r$.

(Pág. 62) cont.

- Draw AL and PM perpendicular on the x-axis and AK perpendicular on MP.

then $AK = NM = OM - ON = x - x_1$

$KP = MP - MK = MP - NA = y - y_1$

SAO DOMINGOS

$\angle KAP = \angle XLP = \theta$

From the right angled triangle AKP, we have

$$\frac{x - x_1}{r} = \cos\theta \quad \therefore \quad \frac{x - x_1}{\cos\theta} = r. \quad ①$$

$$\frac{y - y_1}{r} = \sin\theta \quad \therefore \quad \frac{y - y_1}{\sin\theta} = \theta \quad ②$$

From 1 and 2, we have

$$\frac{x - x_1}{\cos\theta} = \frac{y - y_1}{\sin\theta} = r$$

which is the required equation

Cor. From $\frac{x - x_1}{\cos\theta} = \frac{y - y_1}{\sin\theta} = r$,

we have $\frac{x - x_1}{\cos\theta} = r$ and $\frac{y - y_1}{\sin\theta} = r$

$x = x_1 + r\cos\theta$;

$y = y_1 + r\sin\theta \quad \therefore$

thus the co-ordinates of the general point on the line are $(x_1 + r\cos\theta, \; y_1 + r\sin\theta)$.

obs: Livro Problem Solvers — Algebra & Trigonometry

Pág. 612 - Chapter 27 - Trigonometry
(615-A)

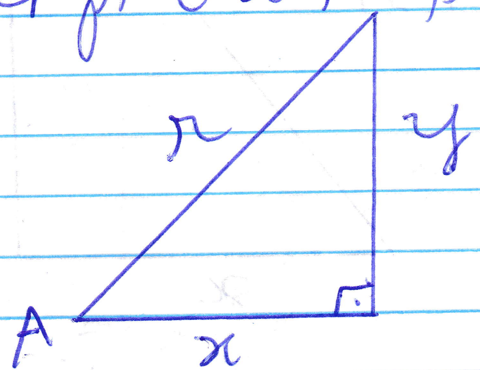

$\sin A = \frac{y}{r}$ $\csc A = \frac{r}{y}$

$\cos A = \frac{x}{r} \rightarrow \sec A = \frac{r}{x}$

$\tan A = \frac{y}{x} \rightarrow \cot A = \frac{x}{y}$

obs:

Calcular: 75° 12' 36" in decimal form

$$75^\circ + \left(\frac{12}{60}\right)^\circ + \left(\frac{36}{3600}\right)^\circ = 75^\circ + (0.2)^\circ + (0.01)^\circ = 75.21^\circ$$

outra observação:

SÃO DOMINGOS

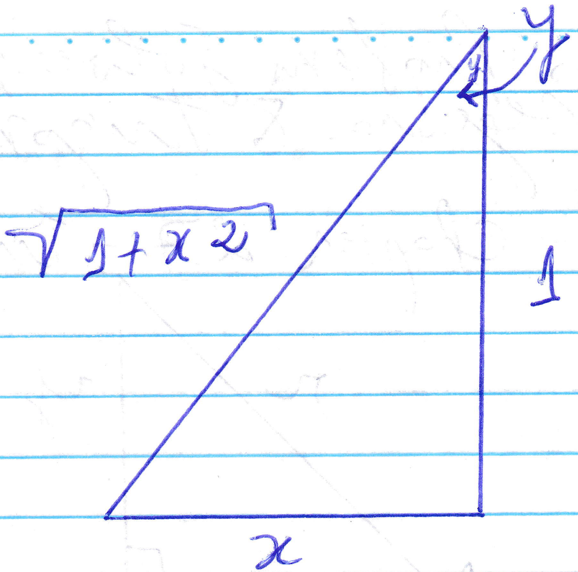

$$\tan y = \frac{x}{1} \rightarrow \cot y = \frac{1}{x}$$

$$\cos y = \frac{1}{\sqrt{1+x^2}} \rightarrow \sec y = \frac{\sqrt{1+x^2}}{1}$$

Obs:

$$\sin y = \frac{x}{\sqrt{1+x^2}} \rightarrow \csc y = \frac{\sqrt{1+x^2}}{x}$$

Outra obs:

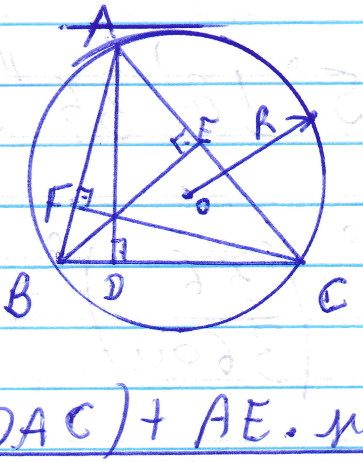

Obs.: AFB → 3 points.

$$R = \frac{AF \cdot \sin(\angle DAC) + AE \cdot \sin(\angle DAB)}{\sin(2 \cdot \angle BAC)}.$$

Término 20230429.2310.

(0645)

202304300645 Aprofundando:

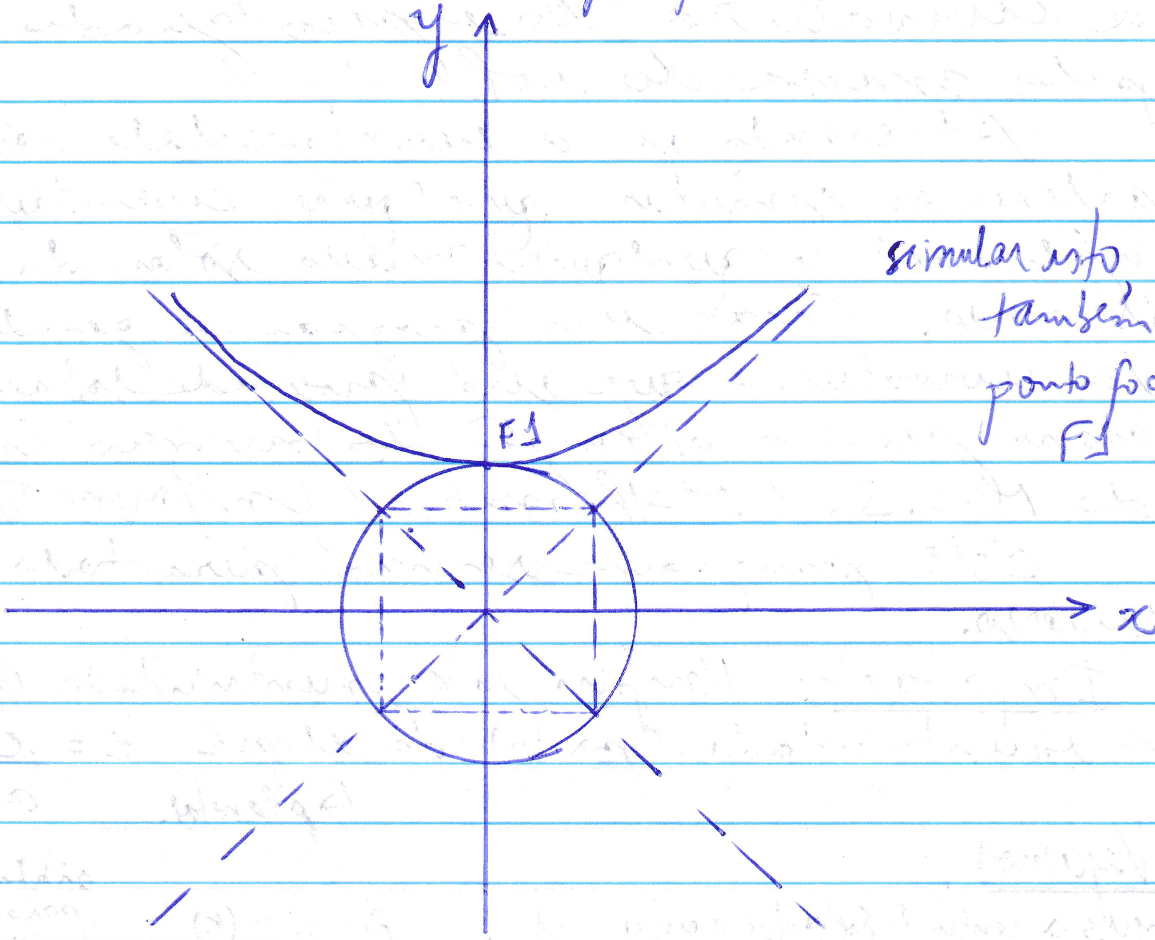

simular isto também no ponto focal F1

correto:

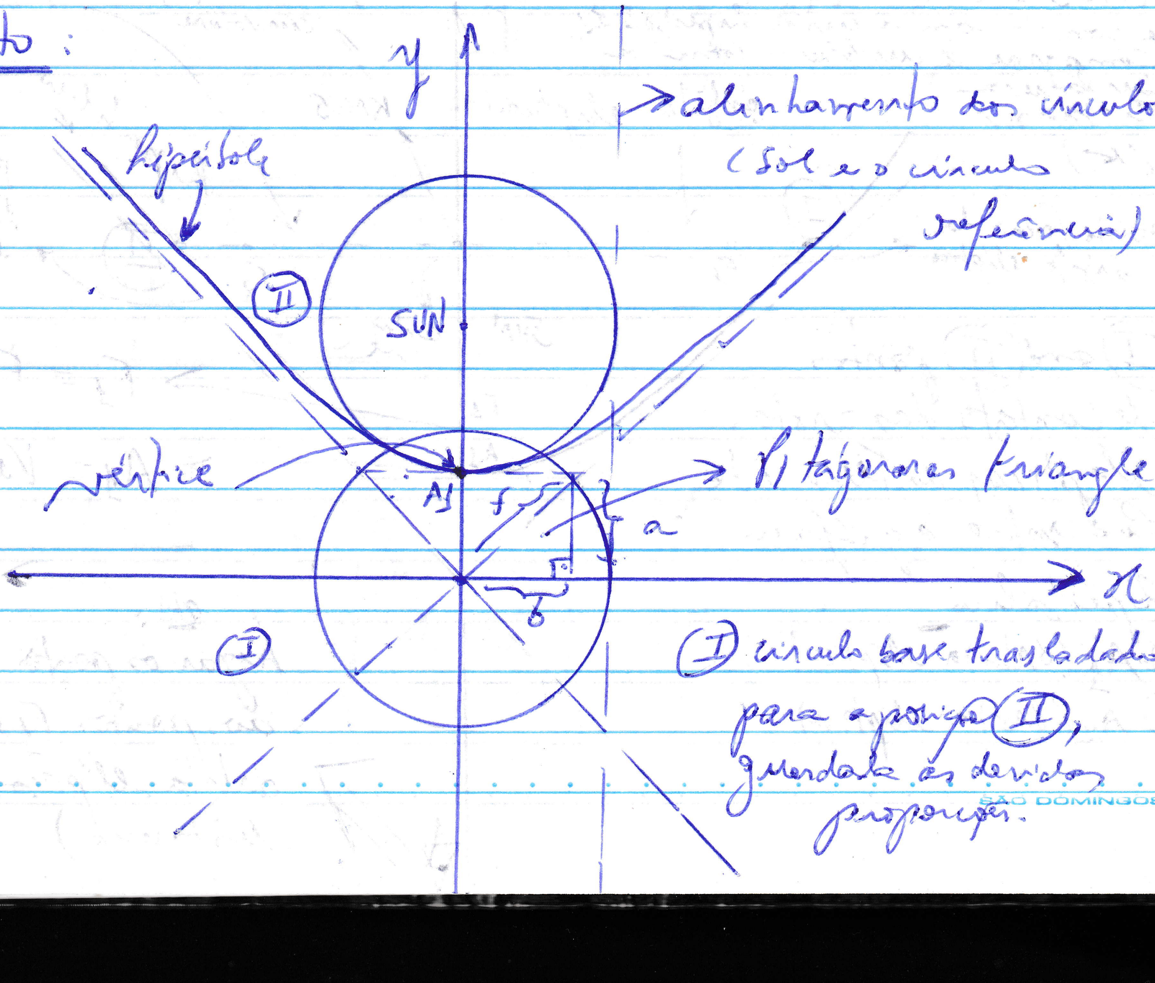
y
hipérbole
II
SUN
vértice
A1
f
a
Pitagoras triangle.
b
I
x

alinhamento dos círculos (Sol e o círculo referência)

(I) círculo base trasladado para a posição (II), guardada as devidas proporções.

A equação da hipérbole representa a curvatura do espaço gerado pela massa do sol.

Alterando-se a excentricidade nós poderemos simular qualquer curvatura aplicável à qualquer massa solar do Universo. Criar um programa gerador da curvatura que será projetado sobre o plano para a obtenção do número local do M.U.S. (Mathematical Uniforme Space).

Este programa servirá para todo o Universo.

Important: Unificar esta excentricidade com a excentricidade (ponto) da elipse $e = \frac{c}{a}$.
↳ planetas.

Vejamos:

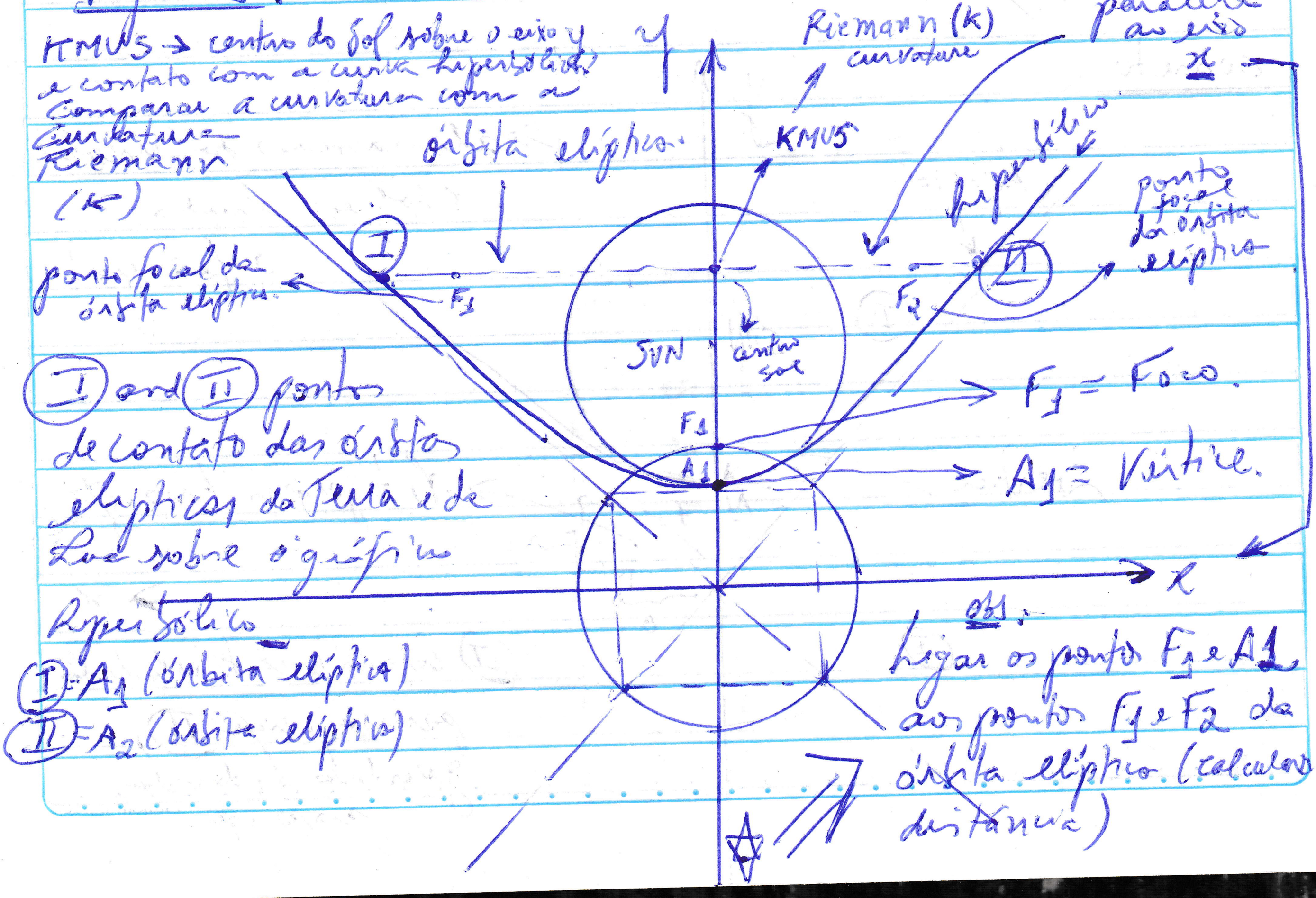
KMUS → centro do Sol sobre o eixo y e contato com a curva hiperbólica
Comparar a curvatura com a curvatura Riemann (K)
y
órbita elíptica
Riemann (K) curvatura
KMUS
órbita paralela ao eixo x
hiperbólico
ponto focal da órbita elíptica
I
II
ponto focal da órbita elíptica
F_1
F_2
SUN
centro sol
F_1
A1
F_1 = Foco.
A_1 = Vértice.
x

Ⓘ and Ⓘⓘ pontos de contato das órbitas elípticas da Terra e da Lua sobre o gráfico hiperbólico

Ⓘ = A_1 (órbita elíptica)

Ⓘⓘ = A_2 (órbita elíptica)

obs.: Ligar os pontos F_1 e A1 aos pontos F_1 e F_2 da órbita elíptica (calcular distância)

Órbita elíptica:

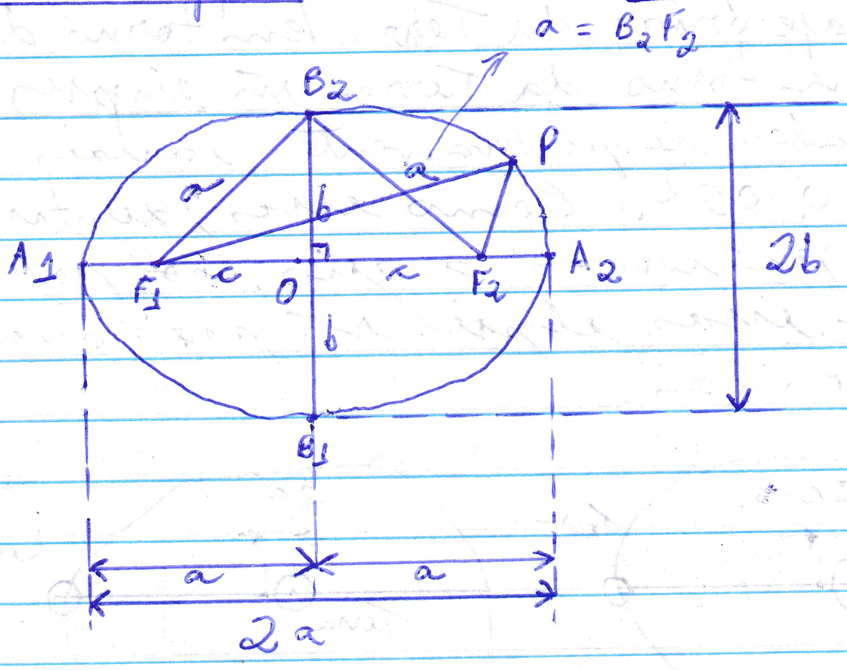

$\Delta F_2 O B_2 =$ teorema de Pitágoras no triângulo retângulo

$a^2 = b^2 + c^2$

Vejamos:

a) focos da elipse: são os pontos fixos F_1 e F_2;
b) distância focal: é a distância $F_1F_2 = 2c$;
c) centro da elipse: é o ponto 0, ponto médio do segmento $\overline{F_1F_2}$;
d) vértices da elipse: são os pontos A_1 e A_2;
e) eixo maior da elipse: é o segmento $\overline{A_1A_2}$, com $A_1A_2 = 2a$;
f) eixo menor da elipse: é o segmento $\overline{B_1B_2}$, com $B_1B_2 = 2b$ e
g) excentricidade: é a razão $e = \frac{c}{a}$, em que

$$0 < e < 1.$$

Important: Esta órbita elíptica é paralela ao eixo x (abscissa) do gráfico da equação hiperbólica causando a curvatura do espaço em decorrência da massa do Sol.

SÃO DOMINGOS

Excentricidades no espaço:

As trajetórias da Terra em torno do Sol e da Lua em torno da Terra são elipses com excentricidades respectivamente iguais a 0,016 e 0,054. Como essas excentricidades são muito próximas a zero, nós podemos dizer que essas elipses são praticamente circunferências.

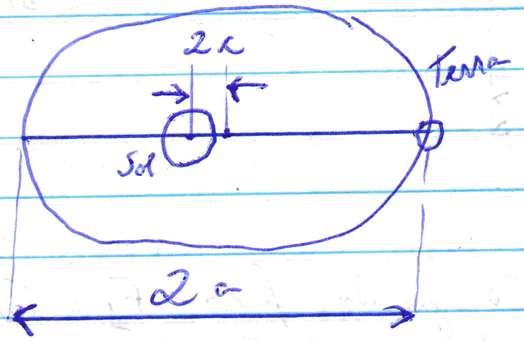
2c
Terra
Sol
2a

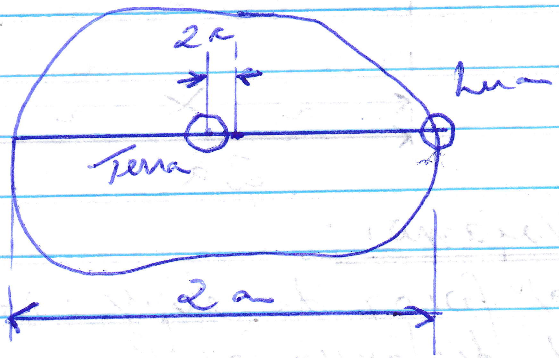
2c
Lua
Terra
2a

$e = \frac{c}{a} = 0,016.$ $e = \frac{c}{a} = 0,054.$

Equação reduzida da elipse:

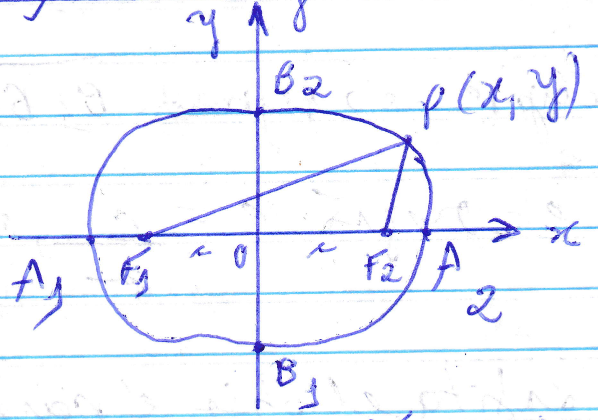
y
B2
P(x, y)
x
A1
F1
c
0
c
F2
A2
B1

Os focos são $F_1(-c, 0)$ e $F_2(c, 0)$.

Para um ponto $P(x, y)$ da elipse:

$d(P, F_1) + d(P, F_2) = 2a \Rightarrow$

$$\sqrt{(x+c)^2+(y-0)^2}+\sqrt{(x-c)^2+(y-0)^2}=2a$$

$$\sqrt{(x+c)^2+y^2}=2a-\sqrt{(x-c)^2+y^2}$$

obs.: Aprofundar o estudo desta equação

Off: Estudo complementar:

Secante e cossecante de um arco.

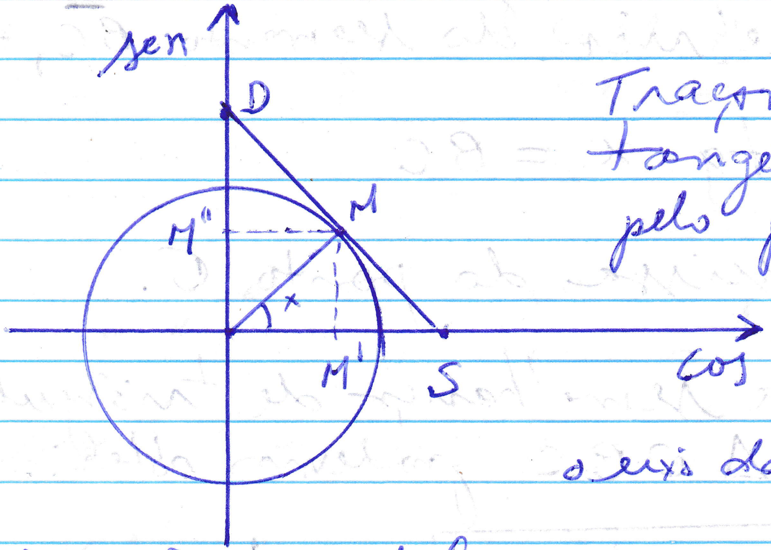

Traçando uma reta tangente à circunferência pelo ponto M, interceptamos o eixo das abscissas no ponto S e o eixo das ordenadas no ponto D.

Da figura, definimos sec x = OS e cossec x = OD.

$OS = \dfrac{1}{\cos x} = \sec x$, $\cos x \neq 0$ e

$OD = \dfrac{1}{\operatorname{sen} x} = \operatorname{cossec} x$, $\operatorname{sen} x \neq 0$.

Quadro:

	y = sec x	y = cossec x
Domínio	$x \neq \frac{\pi}{2} + k\pi$, $k \in \mathbb{Z}$	$x \neq k\pi$, $k \in \mathbb{Z}$
Imagem	$\sec x \leq -1$ ou $\sec x \geq 1$	$\operatorname{cossec} x \leq -1$ ou $\operatorname{cossec} x \geq 1$

SÃO DOMINGOS

Cotangente de um arco:

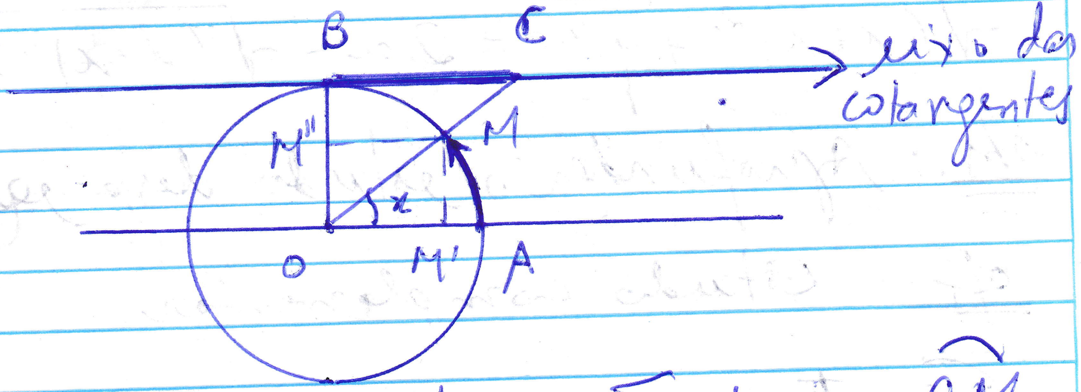

Definimos como contangente do arco $\widehat{AM}$ a medida algébrica do segmento BC, e indicamos

$$\text{cotg } x = BC$$

ou seja,

cotg x é a abscissa do ponto C.

Utilizando-se a semelhança de triângulos $\Delta OM'M \sim \Delta OBC$ podemos obter:

$$\boxed{BC = \frac{\cos x}{\text{sen } x} = \text{cotg } x}$$ sendo

$\text{sen } x \neq 0$, isto é, $x \neq k\pi$, $k \in \mathbb{Z}$.

Podemos escrever também:

$$\text{cotg } x = \frac{1}{\text{tg } x}, \quad x \neq k\frac{\pi}{2}$$

obs: Interessante (utilizar).

Voltando:

Relação trigonométrica fundamental

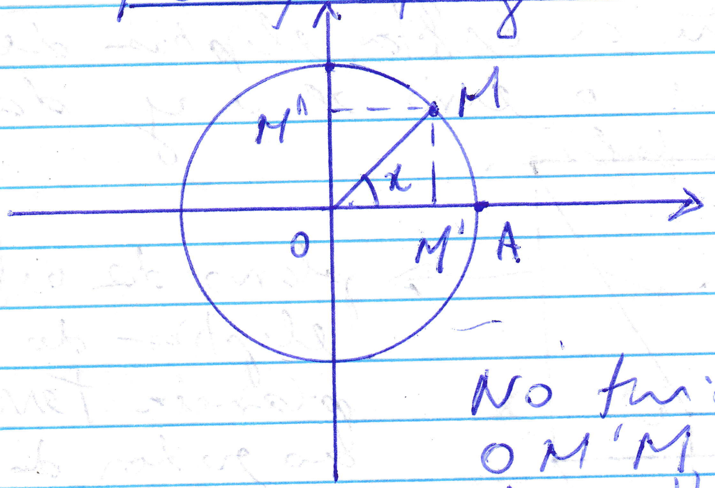

OM = 1

OM' = cos x

OM'' = MM' = sen x

No triângulo retângulo, OM'M, pelo teorema de pit' PITÁGORAS, temos:

$$(MM')^2 + (OM')^2 = (OM)^2 \Rightarrow$$

$$(\text{sen } x)^2 + (\cos x)^2 = 1^2$$ ou, ainda,

$$\text{sen}^2 x + \cos^2 x = 1$$

Voltando: ff:

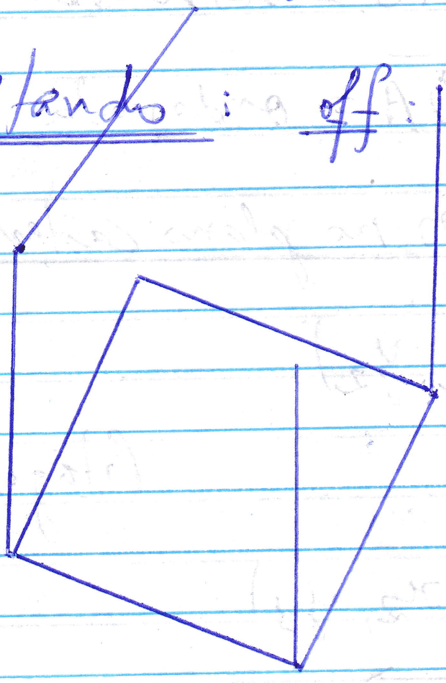

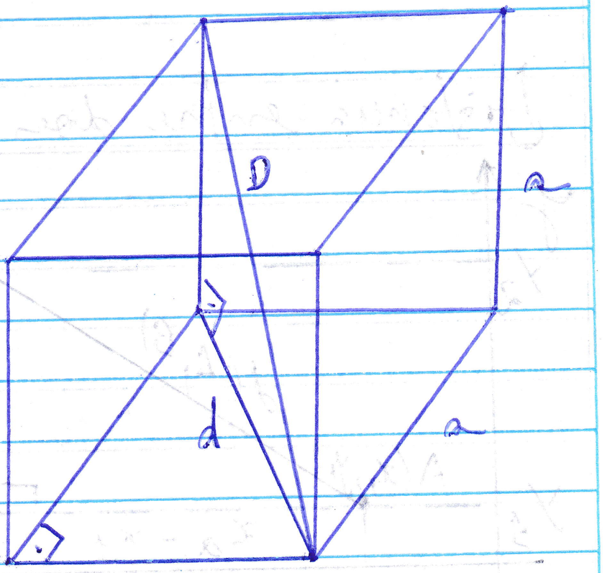

$$D = \sqrt{a^2 + a^2 + a^2} = \sqrt{3a^2} \Rightarrow D = a\sqrt{3}$$

SÃO DOMINGOS

/ /

Usar o cálculo da distância entre os pontos diagonais de um cubo para calcular a distância entre a órbita elíptica de um planeta e o eixo x e y da equação hiperbólica.

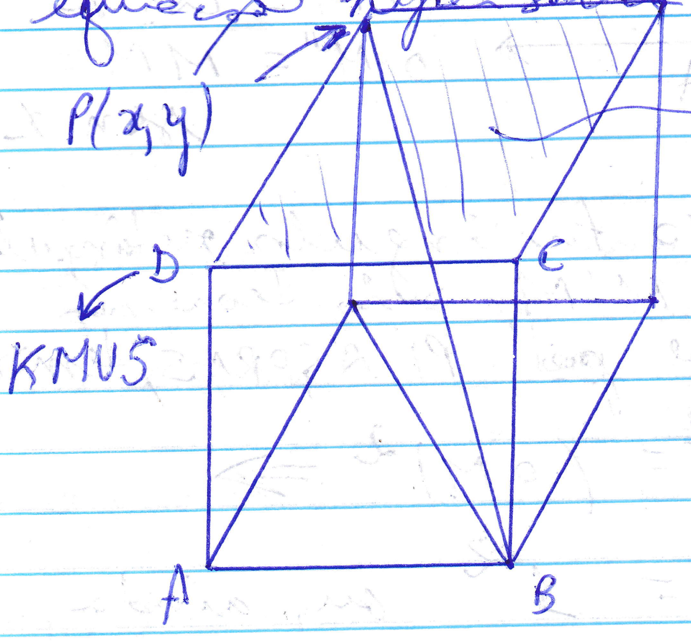
P(x,y)
D
KMUS
C
A
B

plano da órbita elíptica do planeta Terra ao redor do Sol.

□ ABCD representa o plano da equação hiperbólica

$\overline{AB}$ = abscissa x

$\overline{DA}$ = ordenada y

Distância entre dois pontos no plano cartesiano:

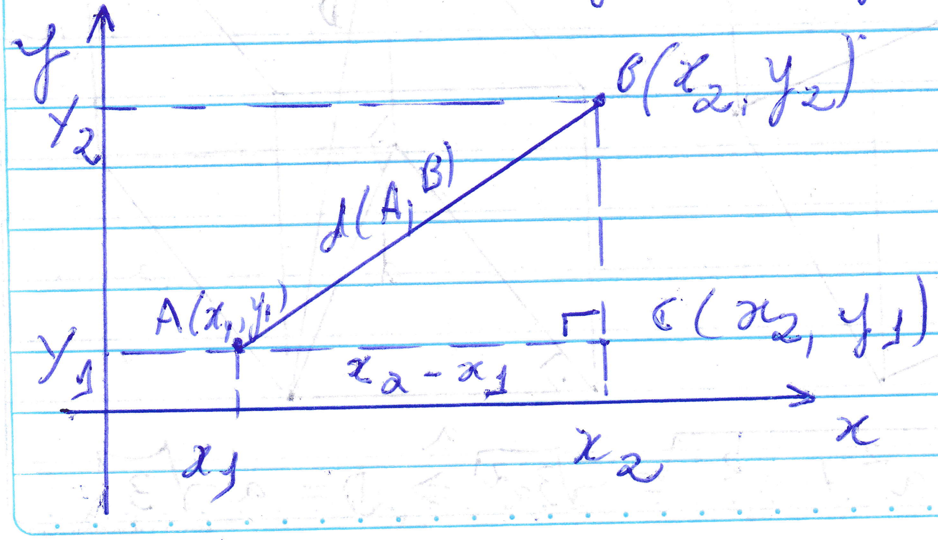
y
y_2
$B(x_2, y_2)$
$d(A, B)$
$A(x_1, y_1)$
y_1
$C(x_2, y_1)$
$x_2 - x_1$
x
x_1
x_2

Pitágoras.

$[d(A,B)]^2 = [d(A,C)]^2 + [d(B,C)]^2$

Sendo $\overline{AC} \,//\, Ox$ e $\overline{BC} \,//\, Oy$, temos:

$d(A,C) = |x_2 - x_1|$ e

$d(B,C) = |y_2 - y_1|$

$d(A,B)^2 = |x_2 - x_1|^2 + |y_2 - y_1|^2$

Como:

$|x_2 - x_1|^2 = (x_2 - x_1)^2$ e

$|y_2 - y_1|^2 = (y_2 - y_1)^2$, concluímos

que a distância entre os pontos A e B é dada

$d(A,B)^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2$.

$$\boxed{d(A,B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}}$$

Voltando à equação reduzida da elipse:

$d(P,F_1) + d(P,F_2) = 2a \Rightarrow$

/ /

Antes de voltar:

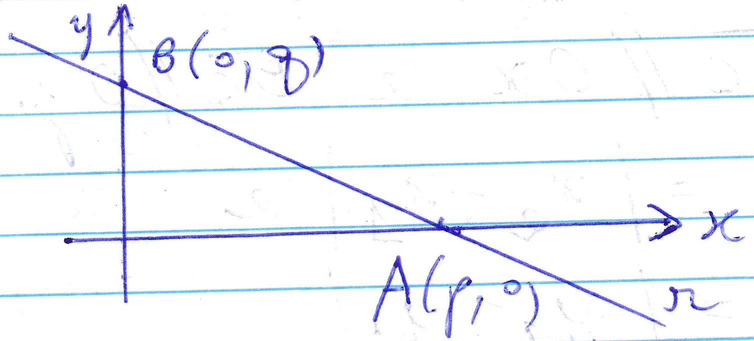

Consideremos uma reta r que intercepta o eixo x no ponto A(p,0) e o eixo y no ponto B(0,q), com $p \neq 0$ e $q \neq 0$.

O coeficiente angular dessa reta é:

$$m = \frac{q-0}{0-q} = \frac{q}{-p} = -\frac{q}{p}$$

Então, a equação da reta r é:

$$y - y_1 = m(x - x_1) \Longrightarrow$$

$$y - 0 = -\frac{q}{p}(x-p)$$

$$y = -\frac{q}{p}x + q$$

$$py = -qx + pq$$

$$qx + py = pq.$$

Dividindo ambos os membros por $pq \neq 0$, temos:

$$\frac{qx}{pq} + \frac{py}{pq} = \frac{pq}{pq} \Rightarrow \frac{x}{p} + \frac{y}{q} = 1$$ ///

Sec. Posições relativas de duas retas

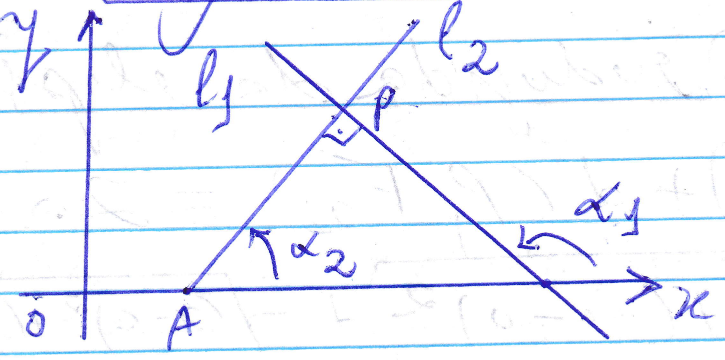

l_1 e l_2 são perpendiculares ($l_1 \perp l_2$).

Consideremos duas retas l_1 e l_2 concorrentes num ponto P, de tal forma que nenhuma delas seja vertical e $l_1 \perp l_2$.
Do triângulo APB, vem

$$\alpha_1 = \alpha_2 + 90^\circ = tg\,\alpha_1 (\alpha_2 + 90^\circ)$$

$$tg\,\alpha_1 = \frac{sen(\alpha_2 + 90^\circ)}{cos(\alpha_2 + 90^\circ)}$$

$$tg\,\alpha_1 = \frac{sen\,\alpha_2 . cos\,90^\circ + sen\,90^\circ . cos\,\alpha_2}{cos\,\alpha_2 . cos\,90^\circ - sen\,\alpha_2 . sen\,90^\circ}$$

$$tg\,\alpha_1 = \frac{cos\,\alpha_2}{sen\,\alpha_2} \Rightarrow tg\,\alpha_1 = \frac{1}{tg\,\alpha_2}$$

Como $tg\,\alpha_1 = m_1$ e $tg\,\alpha_2 = m_2$, temos

SÃO DOMINGOS

$tg\,\alpha_1 = -\dfrac{1}{tg\,\alpha_2} \Leftrightarrow m_1 = -\dfrac{1}{m_2}$.

Voltando:

Equação reduzida da elipse

$d(P,F_1) + d(P,F_2) = 2a \Rightarrow$

$$\sqrt{(x+c)^2+(y-0)^2} + \sqrt{(x-c)^2+(y-0)^2} = 2a$$

$$\sqrt{(x+c)^2+y^2} = 2a - \sqrt{(x-c)^2+y^2}$$

Tendo como: $F_1(\underset{\substack{\uparrow\\ x}}{-c},0)$ e $F_2(\underset{\substack{\uparrow\\ x}}{c},0)$

$$x-(-c) = \boxed{(x+c)^2}$$

Elevando os dois membros ao quadrado:

$$(x+c)^2+y^2 = 4a^2 - 4a\sqrt{(x-c)^2+y^2} + (x-c)^2+y^2$$

Resolvendo a equação: teremos

$$\cancel{x^2} + 2cx + \cancel{c^2} + \cancel{y^2} =$$

$$4a^2 - 4a\sqrt{(x-c)^2+y^2} + \cancel{x^2} - 2cx + \cancel{c^2} + \cancel{y^2}$$

Eliminando os dois lados, teremos:

$$4a\sqrt{(x-c)^2+y^2} = 4a^2 - 4cx$$

Dividindo os 2 membros por 4, teremos:

$$a\sqrt{(x-c)^2+y^2} = a^2 - cx$$

Elevando os dois membros ao quadrado:

$$a^2[(x-c)^2+y^2] = (a^2-cx)^2$$

$$a^2(x^2 - 2cx + c^2 + y^2) = a^4 - 2a^2cx + c^2x^2$$

$$a^2x^2 - \cancel{2a^2cx} + a^2c^2 + a^2y^2 =$$

$$a^4 - \cancel{2a^2cx} + c^2x^2$$

$$x^2(a^2+c^2) + a^2y^2 = a^2(a^2-c^2)$$

Como ~~$a^2 = b$~~ $a^2 = b^2 + c^2$, podemos escrever

$$a^2 - c^2 = b^2$$

$$b^2x^2 + a^2y^2 = a^2b^2$$

SÃO DOMINGOS

Dividindo os dois membros por a^2b^2 ($b>0$, $a>0$):

$$\frac{b^2x^2}{a^2b^2}+\frac{a^2y^2}{a^2y^2}=\boxed{\frac{x^2}{a^2}+\frac{y^2}{b^2}=1}$$

Como o eixo maior da elipse está contido no eixo x e a e b são positivos, então $a > b$. Assim, o denominador de x^2 é maior que o denominador de y^2.

Resumo:

$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$ é a equação da elipse quando o eixo maior está contido no eixo x.

$\frac{x^2}{b^2}+\frac{y^2}{a^2}=1$ é a equação da elipse quando o eixo maior está contido no eixo y.

IMPORTANT: ✡

Aprofundar pesquisa para desenvolver os cálculos como uma função hiperbólica ou como uma função parabólica.

⇒ acho melhor, pois permite o uso de retas paralelas à diretriz e o uso do arco de circunferência como um elemento verificador redundante dos cálculos. + eixo de simetria

Vejamos então :

Obs: O foco pode ser o centro da estrela, i.e., o Sol.

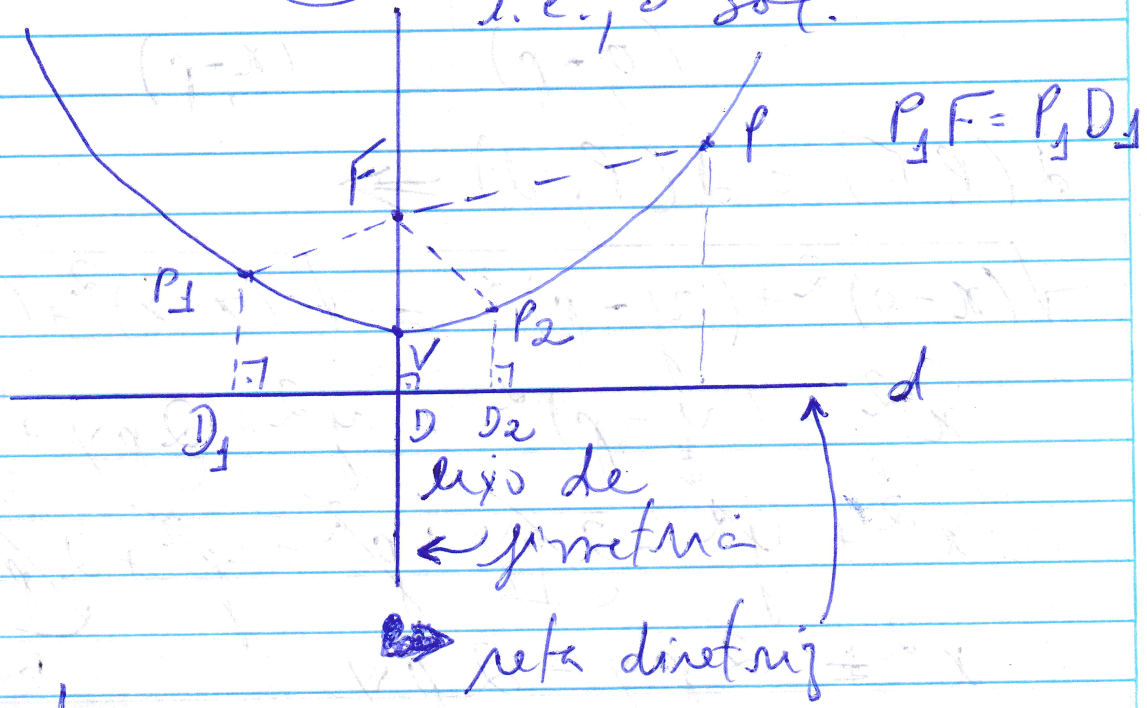
F
P
P_1
P_2
V
D
D_1
D_2
d
$P_1F = P_1D_1$
eixo de simetria
reta diretriz

Portanto :

a) foco da parábola é o ponto F;
b) reta diretriz é a reta d;
c) eixo de simetria é a reta que passa pelo foco F e é perpendicular à diretriz e
d) vértice da parábola é o ponto V, ponto médio do segmento $\overline{FD}$, isto é, VF = VD.

SÃO DOMINGOS

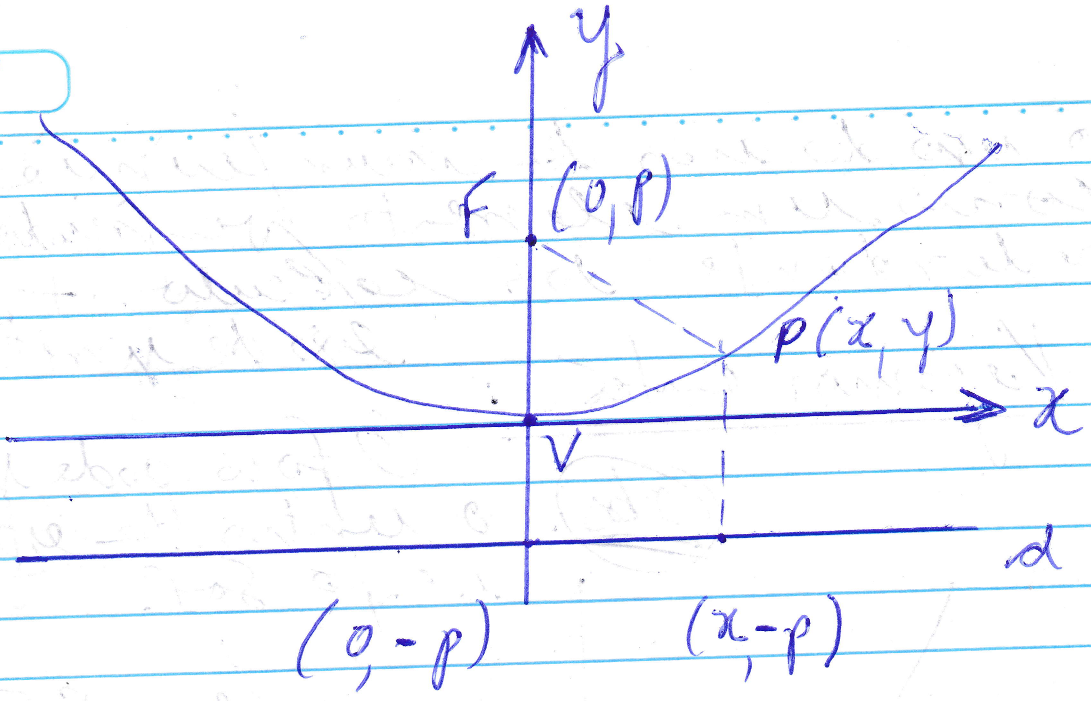

$d(P,F) = d(P,d) \Rightarrow$

$$\sqrt{(x-0)^2+(y-p)^2} = \sqrt{(x-x)^2+(y+p)^2}$$

$$x^2+(y-p)^2 = y^2+2py+p^2$$

$$x^2+\cancel{y^2}-2py+\cancel{p^2} = \cancel{y^2}+2py+\cancel{p^2}$$

somem-se

$$x^2 = 2py+2py$$

$$x^2 = 4py \text{ ou } y = \frac{x^2}{4p},$$

que são equações reduzidas da parábola de foco $(0,p)$ e diretriz $y = -p$.

Importante.

Construção de uma parábola.

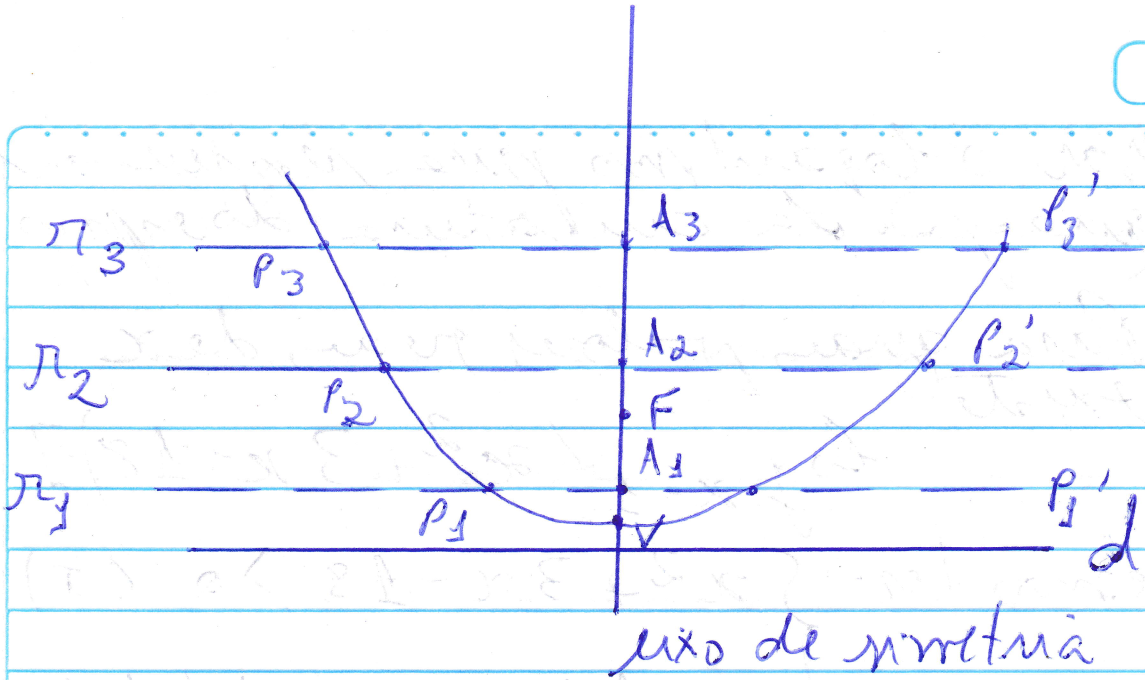
r_3
A_3
P_3'
P_3
r_2
A_2
P_2'
P_2
F
A_1
r_1
P_1'
P_1
V
d
eixo de simetria

r_1, r_2 e r_3 // d (à diretriz)

Obs: Com centro em F e raio A_1D, traça-se o arco de circunferência que intercepta r_1 nos pontos P_1 e P_1'.

etc...

Logaritmos:

$\log_{10} 1385 = 3,141449...$

Important:

$\log_a b$ para existir, devemos ter

- logaritmando positivo: $b > 0$
- base positiva e diferente de 1:

$a > 0$ e $a \neq 1$.

SÃO DOMINGOS

Usar o logaritmo para representar a equação da curvatura do espaço.

Ex: Para quais valores reais de x existe

$$\log_{x+1}(x^2+3x-18)?$$

Devemos ter: $\begin{cases} x^2+3x-18 > 0 & (I) \\ x+1 > 0 \text{ e } x+1 \neq 1 & (II) \end{cases}$

Resolvendo as condições (I) e (II), temos:

$x^2+3x-18 > 0 \quad (I)$

$$x = \frac{-3 \pm 9}{2} \begin{cases} \nearrow x' = 3 \\ \searrow x'' = -6 \end{cases}$$

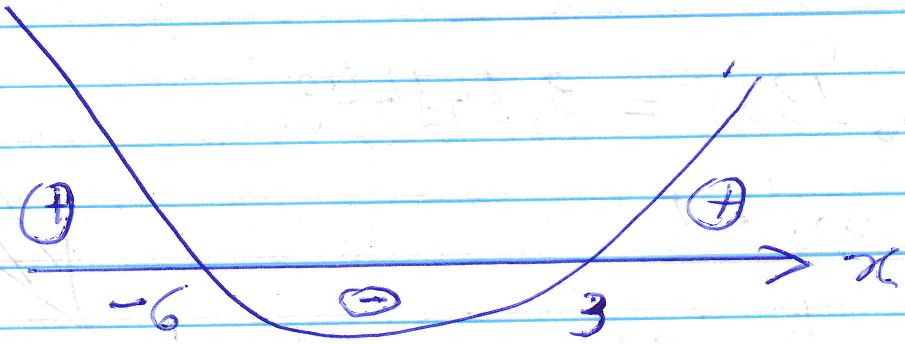

$x+1 > 0 \quad \text{e} \quad x+1 \neq 1 \quad (II)$

$\downarrow \qquad\qquad\qquad\qquad \downarrow$

$x > -1 \qquad\qquad\qquad x \neq 0$

Fazendo a intersecção dos conjuntos soluções de (I) e (II), obtemos:

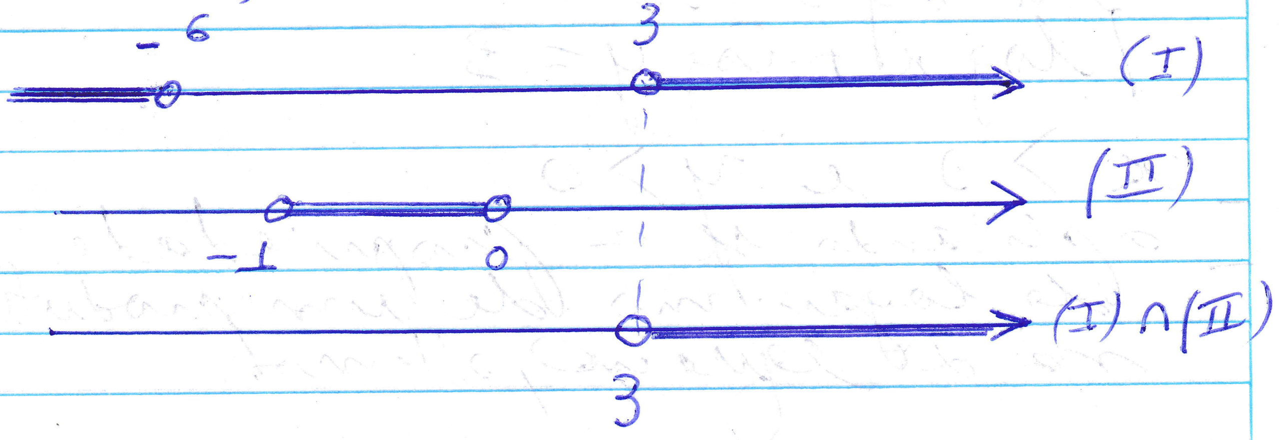

$\log_{x+1}(x^2+3x-18)$ existe para $x>3$.

Propriedades dos logaritmos:

1) $\log_b (a.c) = \log_b a + \log_b c$, com
$a>0$, $c>0$ e $1 \neq b > 0$

2) $\log_b \frac{a}{c} = \log_b a - \log_b c$, com

$a>0$, $c>0$ e $1 \neq b > 0$.

3) $\log_b a^n = n \log_b a$, com
$a>0$, $1 \neq b > 0$ e $n \in \mathbb{R}$.

(prestar atenção)

Usando as propriedades dos logaritmos

SÃO DOMINGOS

/ /

na resolução de sistemas:

$$\begin{cases} x + y = 100 \\ \log x + \log y = 3 \end{cases}$$

$x > 0$ e $y > 0$

– aplicando-se a propriedade do logaritmo de um produto na 2ª equação, obtemos:

$$\log x + \log y = 3 \Rightarrow$$

$$\log(x \cdot y) = 3$$

– Usando a definição de logaritmo, vem:

$$\log(x \cdot y) = 3 \Rightarrow xy = 10^3$$

$$xy = 1000$$

Temos então o sistema equivalente

$$x + y = 110 \quad (I)$$

$$xy = 1000 \quad (II)$$

De (I), vem: $x = 110 - y$ (III)

Substituindo (III) em (II), resulta:

$(110 - y)\, y = 1000$

$y^2 - 110y + 1000 = 0$ $\begin{cases} y' = 100 \\ y'' = 10 \end{cases}$

Substituindo os valores de y em (III), obtemos:

$y = 100 \Rightarrow x = 110 - 100 \Rightarrow x = 10$

ou

$y = 10 \Rightarrow x = 110 - 10 \Rightarrow x = 100.$

Como esses valores satisfazem as condições de existência, temos:

$x = 10$ e $y = 100$ ou

$x = 100$ e $y = 10$

$S = \{(10, 100), (100, 10)\}$

Cologaritmo:

$\log_a \frac{1}{b} = -\log_a b = \operatorname{colog}_a b,$

com $b > 0$ e $1 \neq a > 0$

Término 202304301335

/ /

202304301630 Achar o centro de um circulo:

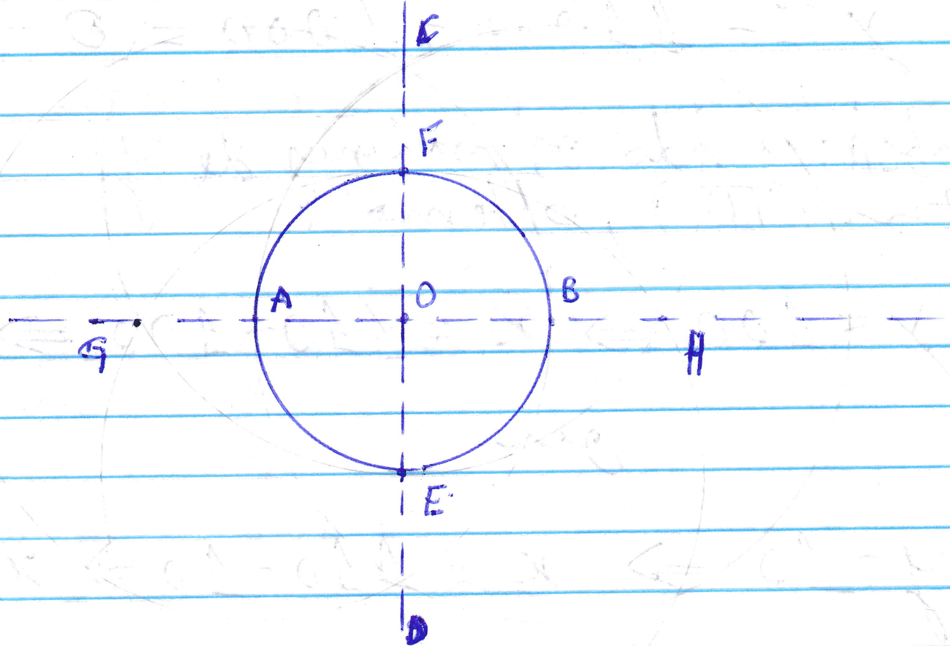

Meu metodo:

1) Pegue o circulo (qualquer) e com o compasso abra com o mesmo tamanho do diâmetro do circulo.

2) Faz um circulo a partir do ponto A com o diâmetro do circulo que se quer achar o centro.

ⓧ ③ Faz outro circulo do outro lado no ponto B. (completo). no ponto o primeiro circulo encontrar o circulo original.

4) Estes dois circulos se encontram nos pontos C e D e em seguida, trace uma reta ligando estes pontos. (tracejado).

5) Ao fazer isto, você encontra os pontos E e F sobre o circulo que se quer encontrar o centro do circulo.

6) Nos pontos E e F faça o mesmo procedimento com o compasso com o diâmetro do círculo que se quer encontrar o centro. Ao fazer isto você terá os pontos G e H, sobre estes pontos trace uma linha tracejada.

7) Ao fazer isto as retas $\overline{CD}$ e $\overline{GH}$ se interceptam no centro do círculo (perpendicularmente) que se quer achar o centro. (Ponto O).

Important: ⊗ da questão

⊗ O ponto A (inicial) é escolhido aleatoriamente, ou seja, um ponto qualquer sobre o círculo que se quer encontrar o círculo.
Ao fazer um círculo com o diâmetro do círculo original sobre o ponto A, você encontrará o ponto B quando o mesmo tocar o círculo original que se quer encontrar o círculo.

Veremos outro método.

Livro: Álgebra lineal y Geometria
Gallo
Pág. 337 (351)

Usando mediatrices.

SÃO DOMINGOS

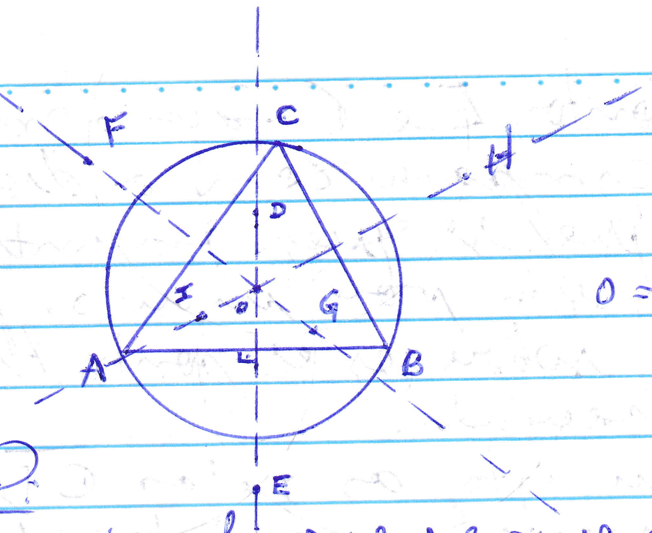

O = circumcentro
(centro do
círculo)

Demonstração:

1) Pegue o círculo que se quer achar o centro.

2) Desenhe um triângulo inscrito no círculo = ΔABC

3) Ache a mediatriz das faces do triângulo inscrito.

3a) Comece pelo ponto A;

3b) Faça o mesmo ponto B;

3c) depois faça no ponto C.

Desta forma, você achou todas as mediatrizes das faces de um triângulo inscrito.

Marcando os ponto:

3a) No ponto A faça um círculo com diâmetro maior que o raio e menor que o diâmetro do círculo no qual está inserido o ~~círculo~~ triângulo inscrito.

3b) No ponto B faça outro círculo utilizando-se da mesma medida utilizada no ponto A. Desta forma, você encontrará os pontos D e E, trace uma reta tracejada ligando estes pontos, neste ponto, esta reta irá interceptar a reta AB (face de um triângulo), fazendo um ângulo 90°.

3c) Com o compasso, utilize a mesma medida utilizada para desenhar os outros círculos sobre o ponto C. Neste momento, você acaba de criar os outros pontos para você achar a mediatriz das duas faces do triângulo inscrito restantes, ou seja, FG (pontos) e os pontos ~~D~~ H e I.

Portanto, teremos as retas:

a) $\overline{DE}$ (interseção dos círculos criados a partir dos pontos A e B).

b) $\overline{FG}$ e HI criadas pelo círculo criado sobre o ponto C.

obs: El circuncentro de un triángulo rectángulo se halla en el punto medio de su hipotenusa.

Important: Este procedimento para achar o centro de um círculo vale para o caso de uma circunferência inscrita em um triângulo qualquer (não precisa ser um triângulo retângulo).

obs. Las bisectrices de los ángulos interiores de un triángulo se cortan en un solo punto que coincide con el centro de la circunferencia inscrita en él.

Atenção: Com o triângulo retângulo você poderá utilizar a mediatriz.

Veja: Pág. 353 (367) - Capítulo 8
Espacios euclídeos

obs: Longitudes, áreas y ortogonalidad.

Término 20230430 1645.

SÃO DOMINGOS

2023 05230633

Lavenda wrote A New Perspective on Thermodynamics, published in 2009, by returning to Carnot's original conception that work can only be done when there is a difference in temperature, and the necessity of closing the cycle before that work can be assessed.
More recently Lavenda has directed his interests to relativity by providing it with a new foundation based on non-Euclidean geometries.

(Book: A New Perspective on
Relativity —
Bernard H. Lavenda
An Odyssey in Non-Euclidean
Geometries) — 2012 — 695 pp.

Rather than measuring distances in terms of the usual Euclidean metric, distances are defined in terms of what is know as a cross-ratio, a perspective invariante of four points, which, for the space of velocities, just happens to be compounding of longitudinal Doppler shifts.
Doppler shifts are fundamental to relativity: oblique Doppler shifts describe aberration, while second

order ones describe length contraction, but rather than being in the direction of the motion are perpendicular to it.

The uniform rotating disc which is considered by some to be missing link in Einstein's formulation of general relativity, is exactly described by the hyperbolic metric in polar coordinates, named after the nineteenth century italian geometer Eugenio Beltrami, which predicts the circumference of the disc to be greater when in motion than when it is at rest. Thus a uniform rotating disc belongs to hyperbolic, and not Euclidean, space and so, too, does relativity.

Critica:

Questa analisi è quasi del tutto vera ma non perpendicolarmente, ma secondo un nuovo approccio matematico. (20230522)

email: bhlavenda@gmail.com

This analysis is almost entirely true, but not perpendicularly, but according to a new mathematical approach

* Lavenda: (text) email 202305230619

The Doppler is well know that it is linked to special relativity. But considering the longitudinal Doppler, the displacement is proportional to the ratio $c + v$ and $c - v$, where v is the speed of the mobile. This seems to contradict the fact that nothing can travel faster than the speed of light (Ritz's ballistic theory of light). This was mentioned in the book by Louis Essen (1972) who criticized special relativity saying it is just a new unit of measurement instead of a previous new theory.

(Distance, time, and velocity.
Which of the two do you want to keep constant?)

Also, the logarithm of the longitudinal Doppler is a distance in the hyperbolic plane (Poincaré), where c is a unit of measure and not an absolute speed. So it depends on the geometry you're considering.

Research and analysis:

Longitudinal Doppler Effect

The longitudinal Doppler effect considers the simple case of a source moving

directly towards you or away from you along a straight line.

The transverse Doppler effect considers what is observed when the observer is displaced in a direction perpendicular to the direction of the motion.

Answer:

① Tanto the longitudinal Doppler Effect quanto the transverse Doppler Effect usam como parâmetro a linha reta (straight line)
→ referencial.

Para mim, a questão toda está como nós entendemos o conceito de uma linha reta em um Universo curvo, sua projeção ortogonal e as suas decorrentes conseqüências sobre o mundo e o universo no qual nós estamos inseridos.

O valor do número PI, ou melhor, o meu novo valor do número PI – M.V.S. traduz este conflito intrínseco ao qual os nossos estudos estão submetidos de forma quase inquestionável, mas que para mim, está errado e incompleto, sujeito, portanto, ao questionamento inevitável.

② Ritz's ballistic theory of light.

Book: Ballistic Theory of Ritz
and Picture of the Universe
S.A. Semikov - 442 pp.
3rd - 2013

Pág. 48 - Fig. 17

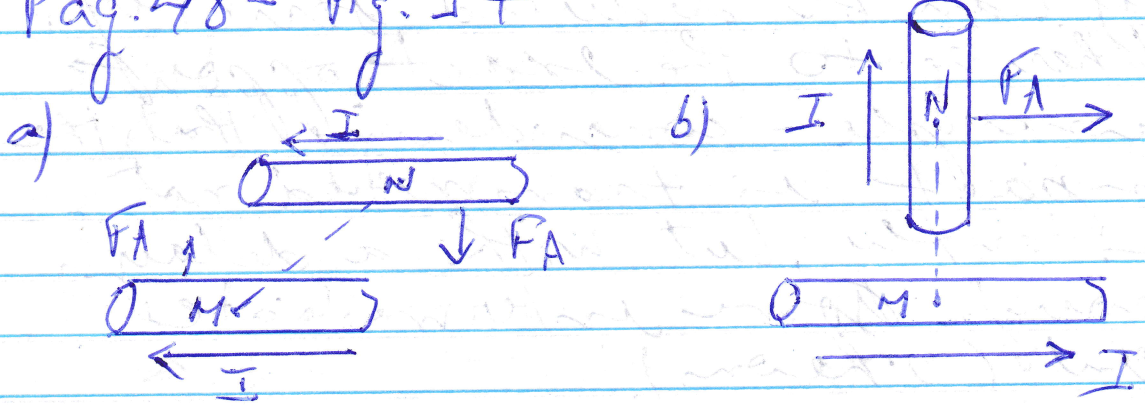

Obs: Comparar com o efeito Doppler.

Pág. 47 -

Pág. 46 - Gauss and developed by Weber built electrodynamics on its basis, considering magnetic and induction forces as a consequence of changes (during the movement and acceleration of charges) of the eletric forces.

Weber ⇒ was a supporter of the theory of action at a distance

(grupo meu) ⇒ entalagment particular (entrelaçamento) action at a distance, that is, the instantaneous spread of influence, without the help of any intermediate agent.

Gauss → as argued by Gauss (Weber's teacher), they can be deduced strictly, adhering to the exact opposite principle - considering that the impact is transmitted not instantly, but with a delay, through some intermediate agent (rheons).

Pág. 47 - the assumption of an instantaneous transfer of action with an infinite speed of rheons ($c = \infty$) would lead, on the contrary, to a constant value of force that does not depend on the movement of charges

Ritz - reducing magnetic effects to electrical ones, begun by Ampere.

Ampere → who first understood that magnetism is a fiction, and a magnet is only a set of elementary molecular circular currents, that is, ultimately, the movement of charges.

Obs: Magnetism is a fiction as gravity.

Pag. 48 → (48)

Ritz's theory, the action force is sometimes not equal to the reaction force and they are not directed strictly along the line connecting the elements or charges.

→ But this is caused by the finite speed of impacts and, ~~it~~ unlike the Biot-Savart law, does not contradict the law of conservation of momentum.

~~Ritz~~
Ampere
Weber
Ritz
> does not coincide with the generally accepted law of Biot-Savart-Laplace, following from Maxwell's theory.

Important:

Dependência da velocidade

Ampère
Weber → c → ∞
Ritz

Law Biot-Savart-Laplace →
c (light velocity)

Pag. 52 - Walter Ritz, who received the formula as a direct consequence of the mechanism he discovered of the interaction of elementary charges (electrons) - through the exchange of standard micro particles (rheons).

Ritz explained not only all magnetic effects, but also the phenomenon of electromagnetic induction discovered by Faraday.

Pag. 53 - Doppler and Ritz effects.

electrodynamic effects are a direct consequence of the Doppler and Ritz effects - changes in the frequency

$$f = f_0 [1 - V^2/c^2 + Ra/c^2]$$

from the motion of the source.

Important:

electrodynamics effects is a direct consequence of the space-time relations discovered by the Ritz and the final velocity c theory, that is, delay electrical signals.

It is the unified kinetic nature of the Doppler and Ritz effects that makes it possible to understand why the change in the flux ϕ through the circuit, both from the velocity and from the acceleration of charges, generates the same induction EMF, and also to find exceptions to this empirical Faraday rule.

Anotações

See: obs: 442 pp.

A New Perspective on Doubly Special Relativity → pp. 1 →

DSR

SÃO DOMINGOS

In Quantum Spacetime phenomenology, affecting the standard kinematics of SR in a different way:

either Lorentz invariance is broken, which implies the existence of a privileged system of reference,

or the

→ symmetries of SR are deformed, as it the so-called "Doubly Special Relativity" scenario so that a relativity principle is still present

Question:

Lorentz invariance is broken

or

symmetries of SR are deformed (DSR)

Source:

universe-09-00150.pdf

---

other: Publications →

A New Perspective on Relativity

Source: 9789814340489_~~1#~~
excerpt_001.pdf

50 pp.

Pag. ~~#~~4 (find: Essen)

If time dilatation and space contraction due to motion are actual processes then there is no symmetry between observers in different inertial frames. The first postulate of relativity is therefore violated [Essen 71].

obs: Einstein used gedanken experiments which is an oxymoron.

(figure language)

Einstein [16] has to say about a pair of local observers on a rotating disc
→ time slowly.

Pág. 5 → velocidade da luz.

SÃO DOMINGOS

If the velocity of light is independent of the velocity of its source, how then can the outward journey of a light signal to an observer moving at velocity

v be $c+v$,

on its return it travels with a velocity $c-v$?

Although this violates the second postulate, such assertion appears in the expression for the elapsed time of sending out a light signal from one point to another and back again in the

Michelson-Morley experiment

whose null result they hope to explain.

They also appear alongside Einstein's relativity velocity composition law in his famous 1905 paper "On the Electrodynamics of Moving Bodies".

Important:

Also in that paper is his "definition" of the velocity of light as the ratio of

"light path" to the "time interval"!

But we are not allowed to measure the path of the light ray and determine the time it took, for c has been elevated to a universal constant.

How can two units of measurement be made constant by definition?"
Essen queries.

obs: Pag. 5 - Book:
(ref.) Lavenda - New Perspective on Relativity".

Resposta Geral:

1) Levando em consideração a Termodinâmica e a entropy com os seus devidos repercussões no mundo físico.

2) Tanto the longitudinal doppler effect quanto the transverse doppler effect usam um parâmetro de

observação referencial atinente aos movimentos dos objetos em estudo.

Answer to Lavenda (20230523)

✡ in italian..
(tradução para o inglês)

Dear Lavenda

In the first place, we must note that both the longitudinal and the transversal Doppler effect use a reference observation parameter relating to the movements of the objects under study and, consequently, we cannot consider the velocity (log Doppler) correct, as there is a mathematical divergence in the calculation process.

Gauss and Weber built electrodynamics on its basis considering magnetic and induction forces as a consequence of changes in electric forces, but Weber was a proponent of the theory of action at a distance, where the value of c tends to infinity, ceasing of being a constant and becoming a limit of measurement.

Gauss adhering to the exact opposite principle, considering that the impact is transmitted not instantly, but with a delay (delay), through an intermediate agent (rheon - standard microparticles).

Ritz reducing magnetic effects to efetric (electric)

coneito: to electrical one, initiated by Ampere, considering magnetism as a fiction.

Ritz's theory, unlike the Biot-Savart law does not contradict the law of conservation of momentum, following Maxwell's theory.

Ritz explained not only all magnetic effects, but also the phenomenon of electromagnetic induction discovered by Faraday.

The electrodynamic effects are a direct consequence of the Doppler effects and Ritz (effects).

The electrodynamic effect are a direct consequence of the spatio-temporal relationships (space-time) discovered by the Ritz and delay electrical signals.

Doppler and Ritz effects generate the same induction.

Another important aspect has to be considered, most possibilities have emerged in this quantum spacetime phenomenology, influencing the standard kinematics of SR in a different way:

either the Lorentz invariance is broken, which implies the existence of a privileged ~~frame~~ frame of reference, or the SR symmetries are warped, as in the so-called "Special Double Relativity" (DSR) scenario, so that a relativity principle is still present.

I say this, because Essen, claiming that the first postulate of relativity has been violated, because there is no symmetry and observers in different inertial frames.
(entre)
(in between)

Therefore, we need to do a deeper analysis of these paradigms.

Tome light can travel beyond the speed established by Einstein's theory, depending on the reference used in the study light has speed, but accessing it can violate this postulate. The speed of light is not only higher due to thermodynamics and entropy.

The matter is in the "light path" and "time interval", criticized by Essen, and I go further, which is not the case here.

(FIM). da resposta.

Outro material para pesquisar:

Source: Advances in 3D Geoinformation Systems

Pág. 199 (208)

Chapter 12 —

Drainage reality in terrains with higher order Delaunay Triangulations.

Obs: Usar isto para configurar um plano ou vários planos em desnível e que possamos obter os cálculos.

Ex.

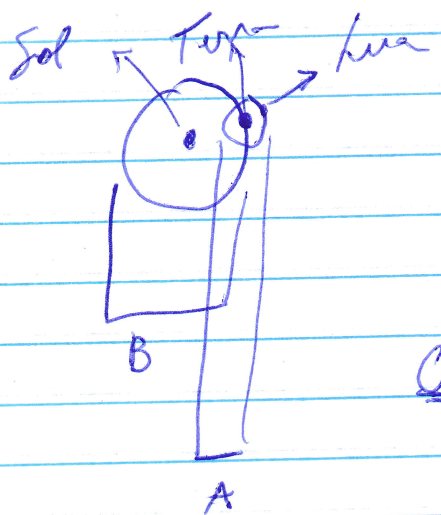

A e B are terrains, cada qual faz com que a matéria flua para a sua zona de influência.

Obs: Com base nestes dois parâmetros, teremos uma visão melhor do novo valor do número (pi) – M.U.S. em cada ponto.

Nós podemos estender para a via Láctea e o Universo.

2023.05231426.

Steniomusich.com 20221226 (1)

Curvature:

Simulation 2 options:

1) Curvatura do espaço-tempo sobre um plano. (projeção ortogonal)

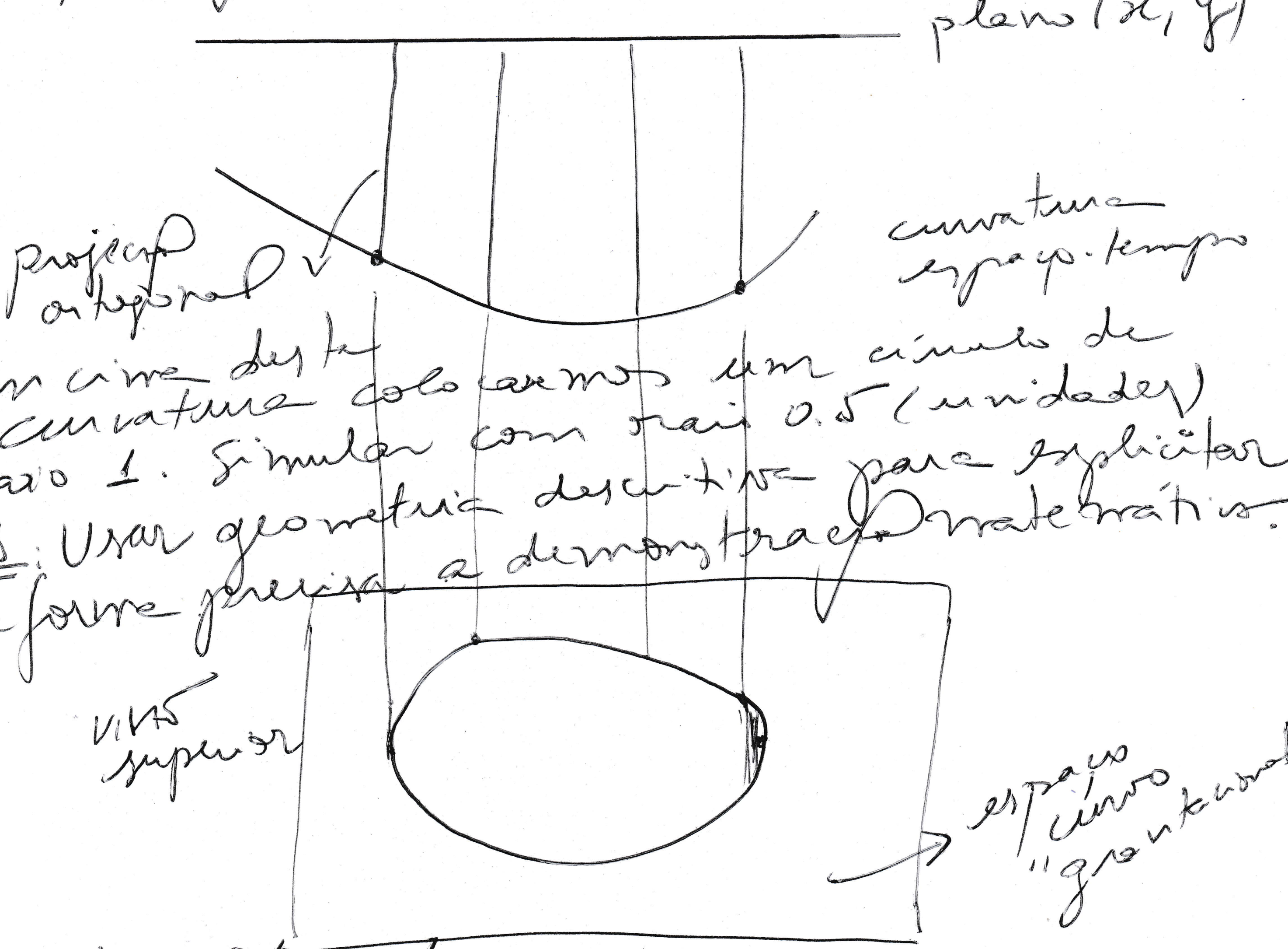
plano (x, y)
projeção ortogonal
curvatura espaço-tempo
visão superior
espaço curvo "gravitacional"

Em cima desta curvatura colocaremos um círculo de raio 1. Simular com raio 0.5 (unidades)

obs: Usar geometria descritiva para explicitar de forma precisa a demonstração matemática.

Import: $\Pi < \pi$

Obs: Plano (x, y) é um plano teórico cujos cálculos deverão ser feitos da forma convencional utilizada pela matemática.

O espaço curvo gravitacional deverá ter exatamente a mesma curvatura espaço-temporal existente no Universo.

O Π deverá demonstrar a realidade do Universo.

2) Curvatura do espaço-tempo calculada da curvatura do Universo usando-se o Cálculo Diferencial e Integral sobre um plano, usando-se o corte transversal.

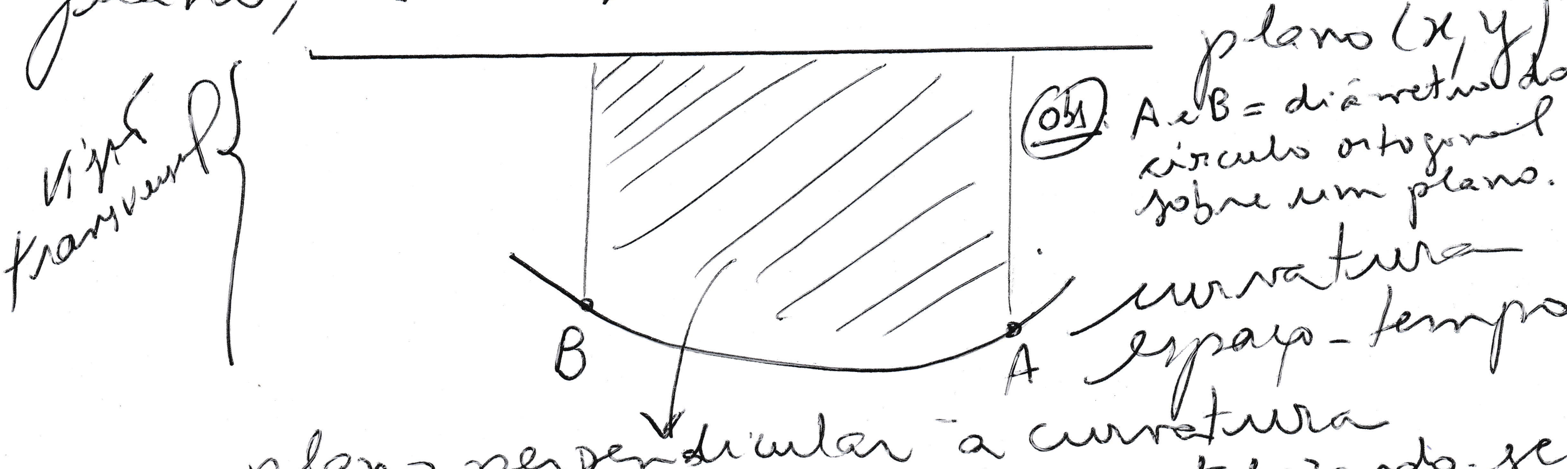

Obs: A e B = diâmetro do círculo ortogonal sobre um plano.

A curvatura (A, B) será calculada utilizando-se o cálculo diferencial e integral e obteremos a função desta curvatura.

Obs: Podemos usar os pontos A e B para calcular o real diâmetro do círculo para calcular a área do mesmo e o valor do π. (M.U.S.) – Mathematical Uniforme Space.

O diâmetro do círculo π será menor que o diâmetro utilizado para calcular o π.

Vejamos:

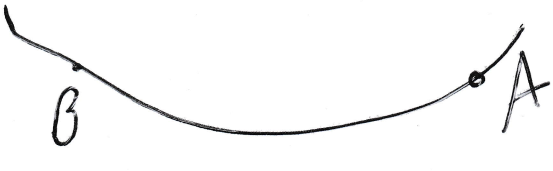

A curvatura (A, B) poderá ser projetada sobre um plano.

A função f(A, B) desta curva deve ser projetada sobre um plano, utilizando-se um corte transversal sobre um plano.

Obs: O interessante que podemos calcular o círculo por 2 meios distintos:

Vejamos:

Stenomusic1.com   z 20221226 ③

2a)

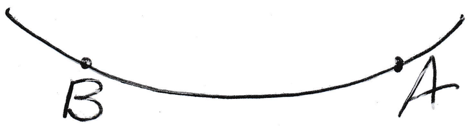

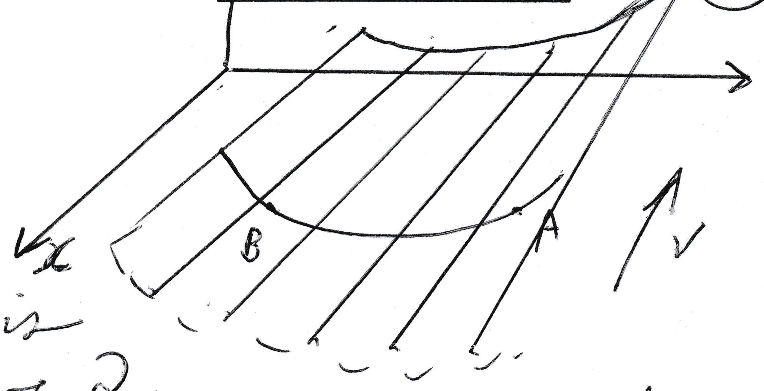

Source:
Cálculo com
Geometria Analítica
Simmons. volume 2
Pag. 364 – Item 14 - Figura 18.31.

$f(x, y) = 0$ (α)   (β)

Obs: Estude espaço paralelo e não-paralelo.
Toda direção no espaço não-paralela ao plano xy pode ser especificada por um vetor da forma

ff: $V = ai + bj + k$

Equação do cilindro: $f(x - az, y - bz) = 0$.

Neste caso: $f(x, y) = 0$ representa a projeção da curvatura sobre um plano.

A curvatura

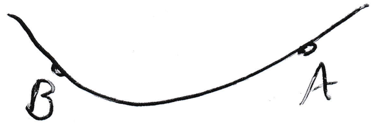

é um viés transversal projetada sobre um plano.

Podemos calcular 2 aspectos:

a) A distância entre A e B = π < distância A e B utilizada pelo π.

b)

esta curva deverá ser completada até formar um círculo e sobre este círculo deveremos calcular o valor do "π",

ou seja,

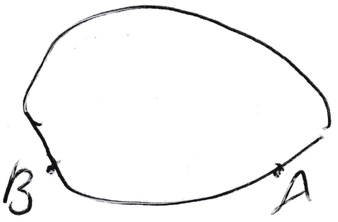

→ nesse círculo, obteremos um valor talvez π transversal.

2a) a) calcular a distância entre A e B, nós podemos fazer de 2 formas:

I)

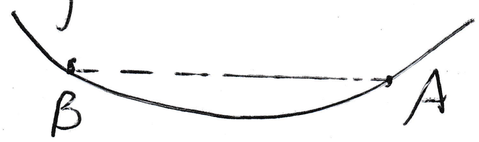

A distância em linha reta entre os pontos A e B.

II)

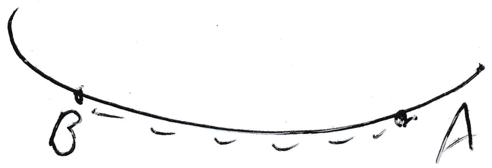

A distância com base na curvatura entre os pontos A e B.

2b) Duas "~~curvaturas~~ distâncias" em um mesmo círculo ⟹

b1)

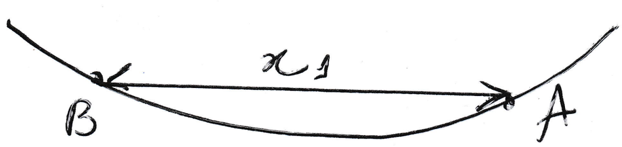

A distância entre A e B (x_1), neste caso, representa a projeção ortogonal da curvatura do espaço-tempo (gravitacional) sobre um plano.

b2)

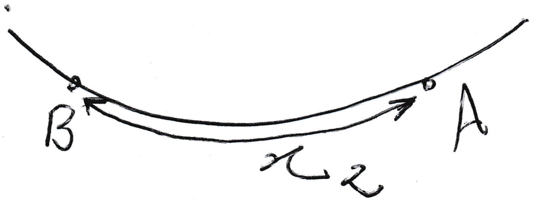

A distância entre A e B (x_2), representa o real comportamento de um círculo sob o ponto de vista gravitacional no Universo.

obs:

⟹ em relação ao π (M.U.S.)

2b1) π real sobre um plano (ortogonal)

2b2) π transversal sobre um plano (transversal)

Obs: $x_1 < x_2$ (distâncias diferentes e portanto, π diferentes.

obs: 2b2) → representa a curvatura do Universo localmente e portanto poderemos calcular o tamanho do Universo como um todo. π local e π geral ↓

202212301645 SteniMusich.com ①

1) ZIP-04/ shapin-light.pdf
Pag.11(29) - 1.2.4.
"...the AB is heteroclinic rather than homoclin
to the background due to the different phase
at $z = \pm\infty$) (1.23)

$$\mathcal{M}_{AB}^{(t)} = \left[1 + \frac{2(1-2a)\cosh(bz) + i\sinh(bz}{\sqrt{2a}\cos(\omega t) - \cosh(bz)}\right] e^{iz}, \quad (1.23)$$

2) Data-NovoVivo-019-2022/
Manifolds of nonpositive curvature-Birkhäuser.p
Pag. 01(12) - Convexity

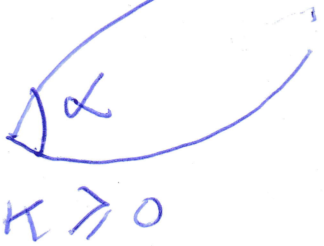

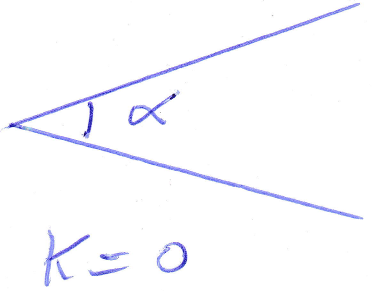

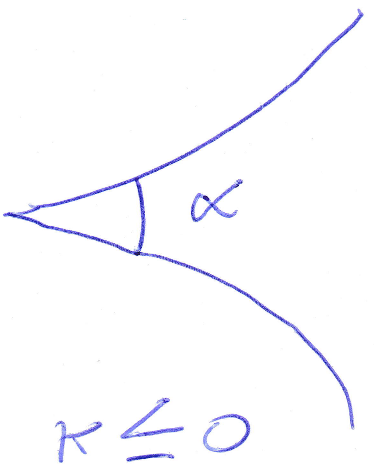

obs: sectional curvature
in a Riemannian
manifold describes locally.

3) Data-NovoVivo-019-2022/
Richtmyer - Principles of Advanced
Mathematical Physics...pdf
Pag. 03(17) -
Gauss elimination for practical purpos
1.2. Linear Space; Normed Linear Spaces
The axioms of a linear space V over a
field F(= R or C) are these:
6 conditions...

2022123020 06 SteniosMusich.com ②

4) Data-Novo Vivo-019-2022/Books-019-2022/Important-018/Principles of Physical Cosmology.pdf

Pag. 10(31) - 2 Mach's Principle and the Cosmological Principle

... trace the connection to Einstein's bold idea that the universe is homogeneous in the large-scale average, what Milne (1935) called Einstein's "cosmological principle".

5) Important-018/Ohanian-Gravitation and spacetime...pdf

6) Radzevich- Geometry of Surfaces.pdf

7) Revisiting the Foundations of Relativistic physics.pdf

Obs: Todos os arquivos acima descritos foram copiados para a pasta de Download do Celular.

8) Differential-Geometry-of-Plane-Curves.pdf
→ 435 pages

Pag. 195(212) - Chapter 6 - Convex Curves
Fig 6.1 - Curves that are locally convex and not locally convex at $t_0 \in I$

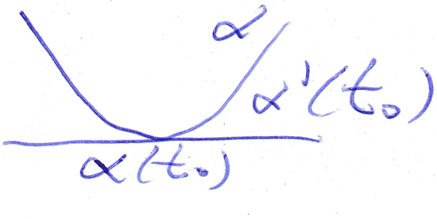

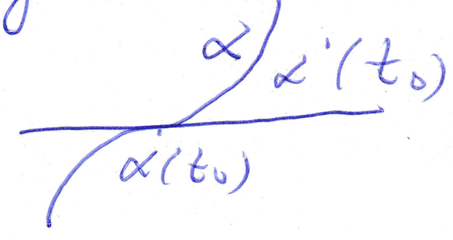

Voltando ao item 02:
Manifolds of nonpositive curvature...pdf
Pág. 01(12) – Convexity.

Em relação $\underline{K}$, devemos adaptar aos cálculos obtidos pelo Π (M.U.S.) – Mathematical Uniforme Space.

$K \geq 0$, $K = 0$ e $K \leq 0$, apresentamos deformados em relação à curvatura do espaço-tempo.

Π (M.U.S.) representa a real realidade do comportamento do espaço-tempo do Universo.

Portanto, obteremos

$$K\widehat{\Pi} \geq 0,\ \underbrace{K\widehat{\Pi} = 0}\ \text{e}\ K\Pi \leq 0$$

esta parte do cálculo é uma representação teórica ajustada aos parâmetros matemáticos mas na realidade uma representação plana de um Universo curvilíneo apresenta as suas peculiaridades bem interessantes.

O interessante que esta representação irá ajudar a comprovar os cálculos do Π (M.U.S.), pois ao calcular uma linha reta (distância entre dois pontos) com a curvatura do Universo devidamente projetado sobre um plano, nós teremos uma real visão da realidade do Universo.

20230331 Stenio Musich

$\frac{1}{7} = 0.142857$ $\sqrt{2} = 1,4142135624$

$\frac{\sqrt{2}}{\frac{1}{7}} = 9,8994949366$ $\simeq \underline{\underline{10x}}$ (?)

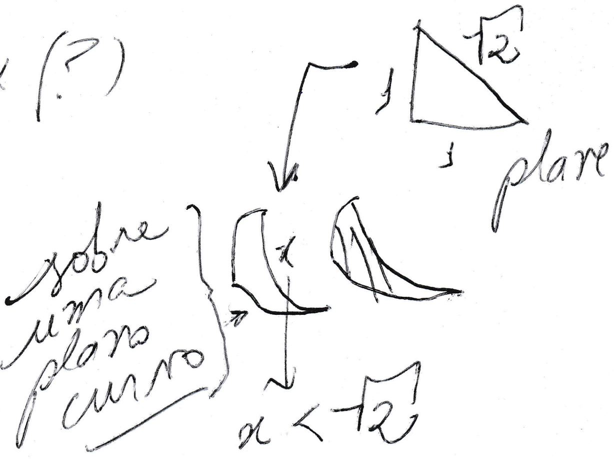

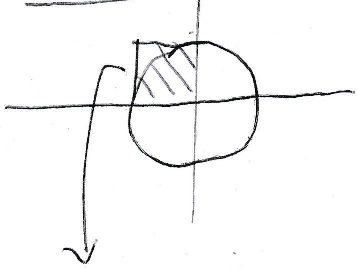

▨ $\times 3\frac{1}{7} = \pi$ (○)

circle

▨ $1 = 1^2 = 1$ (?) wrong!! 2 será?

1

dobro da área

$2 \times 2 = 4$

dobro

▨ $2 = 2^2 = 4$

2

Vejamos:

□ $1 = 1^2 = 2 + \boxed{1\frac{1}{7}}$

1

1.1428571429

off: $(\sqrt{2}) \times (1 + (1 \div 7)) =$

ver ⇒ 1.6162440713

(obs: $(\sqrt{2}) - (1 + (1 \div 7)) =$

0,2713564195

Important:

$(((\sqrt{2}) - (1 + (1 \div 7)) \times 100) \div \sqrt{2} = -79,8122035642$

$\div (1+(1 \div 7)) = -38,7625631329$

$(((\sqrt{2}) - (1 + (1 \div 7))) \times 100) \div (1+(1 \div 7)) - \boxed{100} \overset{?}{=}$

1,1417637356

20230331 Stenio Musich ②

$$((-100-(((-\sqrt{2})-(1+(1\div 7))\times 100)\div(1+(1\div 7)) =$$
$$-1{,}237436871$$

$$(((-\sqrt{2})-(1+(1\div 7))\times 100)\div(1+(1\div 7)) =$$
$$-98{,}7625631329$$

corrijo:

$$(((-\sqrt{2})-(1+(1\div 7)))\div(1+(1\div 7)) =$$
$$0{,}237436871$$

Important:

$$(1+(1\div 7))\div(((-\sqrt{2})-(1+(1\div 7)))=$$
$$4{,}2116458674$$

Voltando:

$$(1+(1\div 7))-\sqrt{2} =$$

corrijo

$$\sqrt{2}-((1+(1\div 7)) = 0{,}271356419$$

outros

$$10\times\frac{1}{7}-\sqrt{2} = 0{,}0143578662$$

$$((10\times(1\div 7))-\sqrt{2}))\div(1+(1\div 7)) = 0{,}0125631329\%$$
$$\div\sqrt{2} = 0{,}0101525446\%$$

$$(3+(1\div 7)-\pi = 0{,}\overline{0012644}893$$

20230402

Mathematicians have found an exact fraction that equals π, thereby disproving the long-standing conjecture that π is irrational.

$$\pi = \frac{1783366216531}{567663097408}$$

The number pi is wrong and there is another number (M.U.S.).

→ O número pi que utilizamos na matemática foi calculado de forma errónea e por este motivo apresenta-se como irracional (dízima não periódica)

O novo número pi ~~pelo~~ ~~qual~~ que eu denominei como M.U.S. é ~~uma~~ número ~~dízima periódica~~.

~~Desta forma~~, M.U.S. não é transcendental. ~~nem irracional~~.

Ex: $\frac{2}{3}x - \frac{1}{5} = 2$ eq. pol. com coef inteiros!

$$\frac{10x - 3}{15} = \frac{30}{15}$$

$10x - 3 = 30$ (coef nº inteiros)

$10x = 33$

$x = 3{,}3$ → número racional.

202304230625 Sferi Musich.

M. U. S. I. C. H → H – h (
E – energy
A – action of forces
T – time
↳ dynamic.

↳ Heat
↳ Curvature
↳ India
↳ Space
↳ Uniforme
↳ Mathematical

obs.:

$$A \to \overset{\text{big-bang}}{0} \xrightarrow[\text{time}]{} \underset{\text{finite}}{\smile} \to \Omega$$

20230425

Principio de Cavalieri

Planos paralelos, as duas figuras tem a mesma área.

cubo = paralelepípedo
retângulo

figura prismáticas
→ linhas paralelas

Prismas = poliedros
(anteriormente chamados

V = 2n
A = 3n

F = n+2 → faces laterais (paralelogramos)
(=) → faces

prisma reto
ou oblíquo

paralelepípedo é um prisma
com bases são paralelogramos.

20230427 0745

Deus é tão [illegible] maravilhoso que criou o nada para que o mundo fosse criado.

Ele criou tudo sem destruir nada.

O nada de Deus permitiu que fosse criado tudo.

O nada gera energia da criação.

eXtreme SPORTS

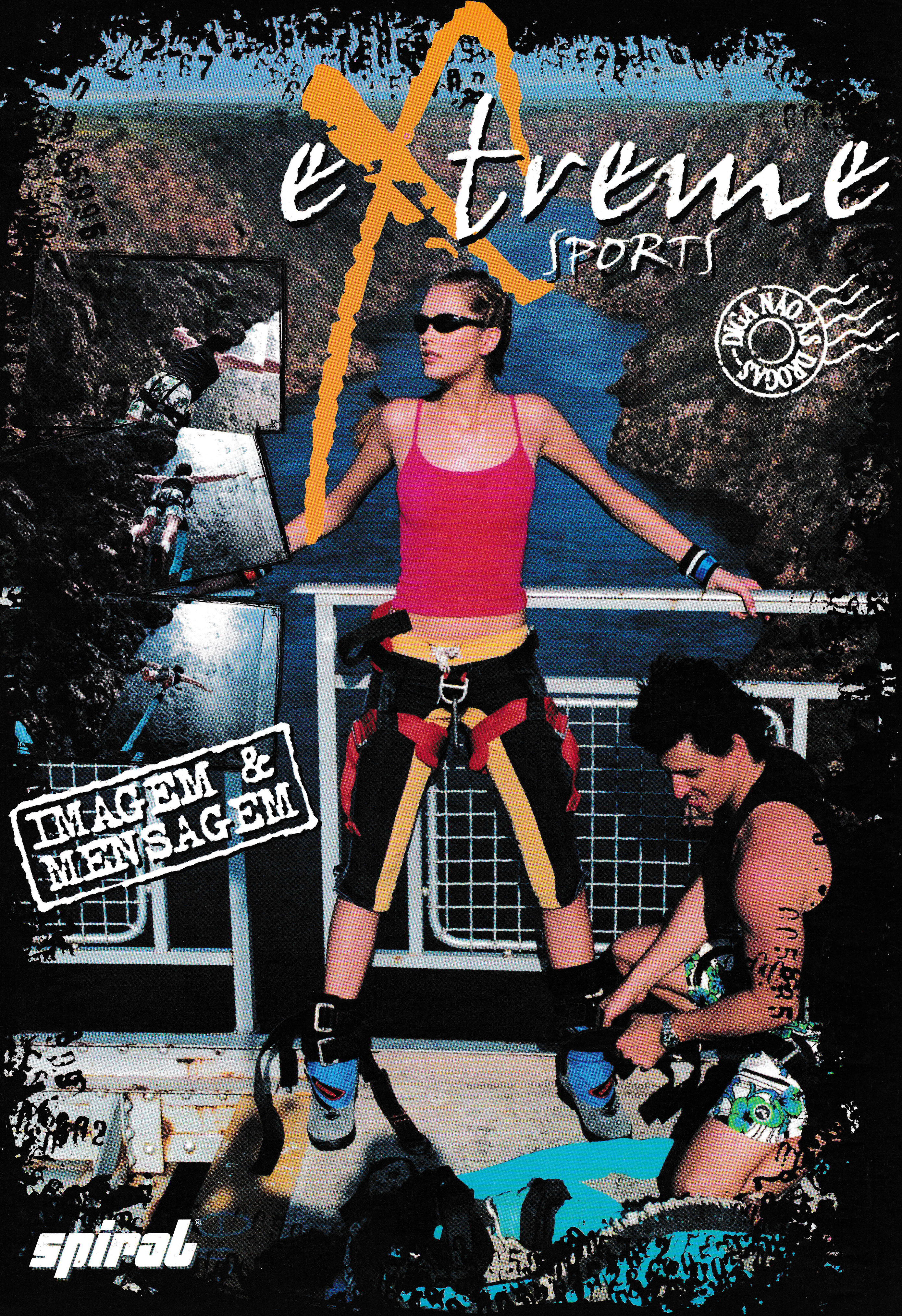
DIGA NÃO AS DROGAS
IMAGEM & MENSAGEM

spiral®

1

MATÉRIA

PROFESSOR

Nome

Stenio Musich

ENDEREÇO

TEL. 69091893

ESCOLA

ENDEREÇO

TEL.

Nº MATRÍCULA

Nº CHAMADA 13/12/99

TURMA

Programação

	SEGUNDA	TERÇA	QUARTA
Hs.			
Hs.			
Hs.			
Hs.			
Hs.			
Hs.			

	QUINTA	SEXTA	SÁBADO
Hs.			
Hs.			
Hs.			
Hs.			
Hs.			
Hs.			

Linha direta

DIGA NÃO ÀS DROGAS

Sua opinião é muito importante !
Escreva e mande suas sugestões,
críticas ou elogios

2

MATÉRIA

PROFESSOR

3

MATÉRIA

PROFESSOR

4

MATÉRIA

PROFESSOR

5

MATÉRIA

PROFESSOR

6

MATÉRIA

PROFESSOR

7

MATÉRIA

PROFESSOR

8

MATÉRIA

PROFESSOR

9

MATÉRIA

PROFESSOR

10

MATÉRIA

PROFESSOR

S. Paulo 060610. Obs. Procurar todos.

Verbos (Irregulares) — Total. 202 Aqui: 136.

Past tense

Past Participle.

arise — arose — arisen — surgir, levantar

awake — awoke — awoken — despertar, acordar

be — was, were — been — ser, estar

São Paulo 12 de outubro de 2023

(2023 10 12 1100)

(Dia de Nossa Senhora Aparecida - Brasil)

(6º Dia - Guerra Yom Kipur - 2ª - 50 Anos e 1 d.)

Research

Books:

1) Exercices-de-géométrie-analytique-01.pdf
(394 pages) → Papelier
Chapitre I - Ligne Droite
I. Points en ligne droite.

1. Démonter que dans un quadrilatère complet les milieux des diagonales sont en ligne droite.
Disassemble that in a complete quadrilateral the diagonal media are in a straight line.
Demonstre que em um quadrilátero completo o meio diagonal está em linha reta.
(Interessante)

On appelle quadrilatère complet la figure formée par quatre droites dont trois ne passent pas par un même point et dont deux ne sont pas parallèles

The figure formed by four lines is called complete quadrilateral, three of which do not pass through the same point and two of which are not parallel.

A figura formada por quatro linhas é chamada quadrilátero completo, três dos quais não passam pelo mesmo ponto e duas dos quais não são paralelas.

English:

1. Prove that in a complete quadrilateral the midpoints of the diagonals are in a straight line.

We call complete quadrilateral the figure formed by four straight lines, three of which do not pass through the same point and two of which are not parallel.

These lines are called the sides of the quadrilateral, and the meeting points of these lines taken two by two are the vertices of the quadrilateral; there are therefore six vertices.

If we join the meeting points of any two sides at the point of the other two, we obtain a line the diagonal of the quadrilateral. We then see that there are three diagonals and we propose to demonstrate that the midpoints of these diagonals are in a straight line.

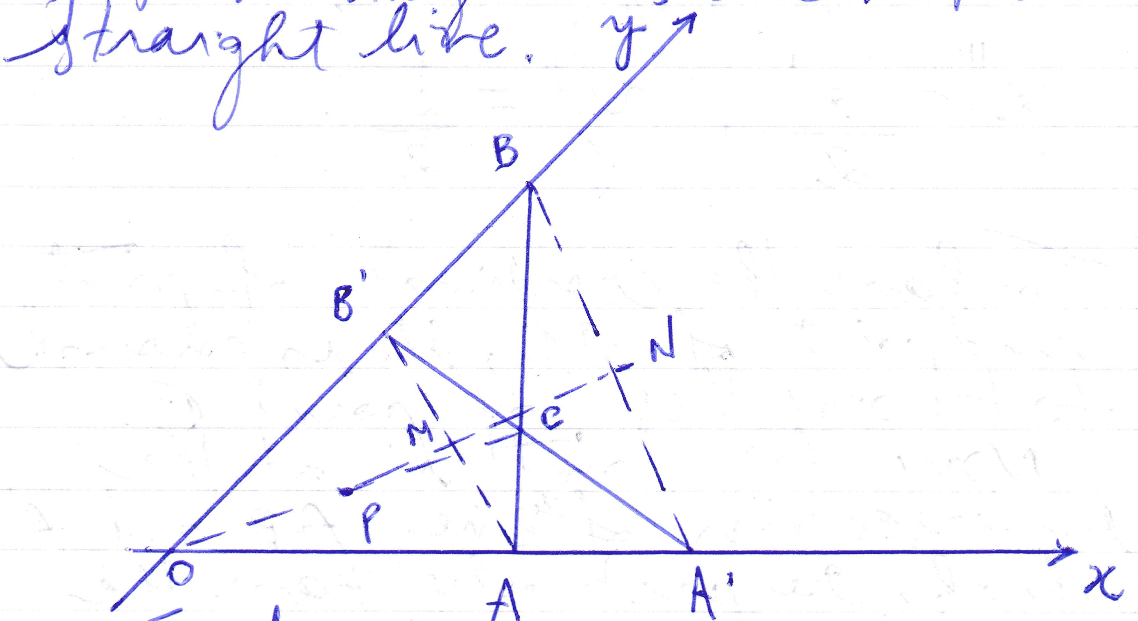

Fig. 1.

Let us take as coordinate axes two sides of the quadrilateral (x and y).

Let the other two be AB and $A'B'$, which intersect at point C.

The three diagonals are then AB' $A'B$ and OC;

their midpoints are the points M. N. P.

We will calculate the coordinates of these points.

Let a, a' be the abscissa of points A, A'; b and b' the ordinates of points B, B'.

The coordinates of point M are

$$x' = \frac{a}{2}, \quad y' = \frac{b'}{2},$$

and those of points N,

$$x'' = \frac{a'}{2}, \quad y'' = \frac{b}{2}.$$

To have the coordinates of point P, midpoint of OC, it is enough to know those of point C.

Now this point is at the intersection of the lines AB and A'B' which respectively have the equations

$$\frac{x}{a} + \frac{y}{b} - 1 = 0,$$

$$\frac{x}{a'} + \frac{y}{b'} - 1 = 0;$$

consequently, to have the coordinates of point C it is necessary to solve these two equations.

We thus obtain

$$x = \frac{aa'(b'-b)}{ab'-ba'}, \quad y = \frac{-bb'(a'-a)}{ab'-ba'};$$

the coordinates of point P, midpoint of OC, are then :

$$x''' = \frac{aa'(b'-b)}{2(ab'-ba')},$$

$$y''' = -\frac{bb'(a'-a)}{2(ab'-ba')}.$$

To establish that the points M, N, P are in a straight line it is necessary to prove the equality

$$\frac{y'''-y'}{x'''-x'} = \frac{y''-y'}{x''-x'}.$$

Or, on a

$$\frac{y'''-y'}{x'''-x'} = \frac{-\frac{bb'(a'-a)}{2(ab'-ba')} - \frac{b'}{2}}{\frac{aa'(b'-b)}{2(ab'-ba')} - \frac{a}{2}} =$$

$$\frac{-bb'(a'-a) - b'(ab'-ba')}{aa'(b'-b) - a(ab'-ba')},$$

$$\frac{y'''-y'}{x'''-x'} = \frac{b-b'}{a'-a}.$$

On the other hand, we easily see that

$$\frac{y''-y'}{x''-x'}$$ is equality

~~(is equality (is therefore verified and the proposition is established.)~~'

$$\frac{b-b'}{a'-a}.$$

The equality is therefore verified, and the proposition is established.

Análise:

Partindo-se do pressuposto que há uma linha reta e somando-se ao uso do teorema de Pappus, nós poderemos ter várias aplicações.

a) usando-se duas linhas partindo do ponto O, nós poderemos ter várias utilidades geométricas. Vejamos:

a1) no plano (qualquer)
traçamos duas linhas sem saber qual a sua relação entre si, ou seja, não sabemos qual é a equação que define cada linha.

Mesmo assim, nós poderemos achar esta linha reta.

a₂b) plano curvo ou esfera
usando-se este princípio:
b1) duas retas
b2) partindo-se do mesmo ponto;

Nós poderemos continuar a obter uma/linha reta que está descrita sobre um plano curvo ou esfera.

b) utilizando-se duas retas e um mesmo ponto, nós poderemos utilizar o Teorema de Pappus (3 points) para estabelecer uma reta também.

Obs: Nós não podemos esquecer que
AB' e A'B não são linhas paralelas.
↳ importante uso para obtenção de uma linha reta em relação a um ponto fora destas duas retas e que no caso é o ponto O.

Vejamos:

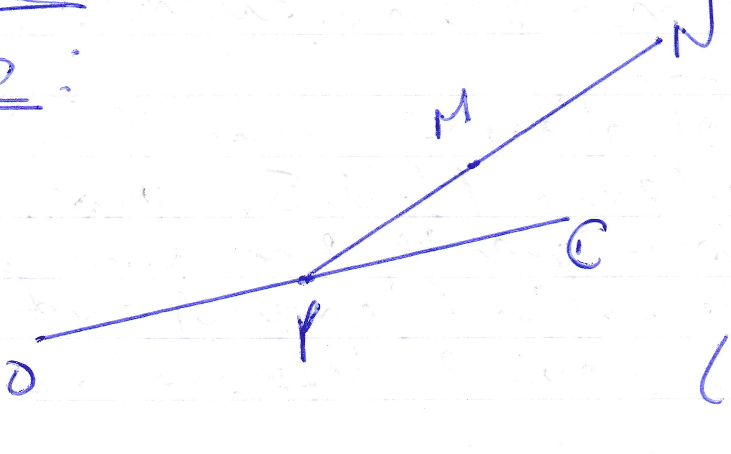

P, M and N
→ 3 points (Pappus)

Obs: Aplicar também ao plano curvo e a esfera.

Importantíssimo:

Aplicar também à minha pesquisa

M.U.S.

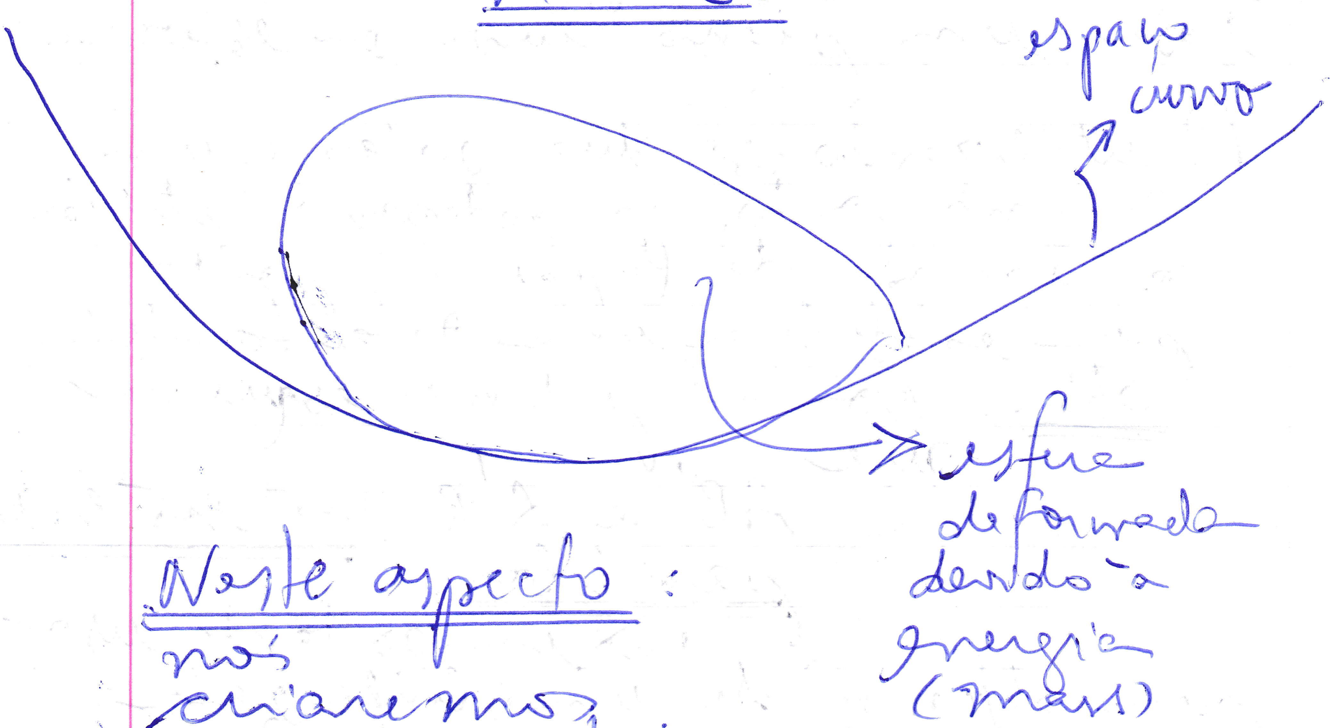
espaço curvo
esfera deformada devido à energia (massa)

Neste aspecto:

nós criaremos um novo tipo de linha reta, ou seja, não é uma linha reta no plano nem no plano curvo e nem na esfera.

Nós offeremos uma linha reta no plano M.U.S. projetado ortogonalmente em um plano matemático (teórico), mas não real sendo o real deformado e outros valores!

Source: gallica.bnf.fr
(Book).

Obs: Estudar da pág. 01 à 57 (deste livro)
→ linha reta.

2) Exercices-de-géométrie-analytique-02.pdf
454 pages - Papelier. (Gallica)

3) Exercices-de-géométrie-analytique-03.pdf
408 pages - Papelier. (Gallica)

Cont. Vol. 1:

Pág. 267 - Chapitre VII

Exercices sur les Équations Réduites des Coniques.

I - Ellipse

Pág. 269 (285)

Ce théorème est appelé théorème de Frégier, et le point N est dit le point de Frégier relatif au point M.

Referências (Frégier point)
para estudo

R1) scg2021_online.pdf

Source: http://ssgg.sk/scg/seminare.htm

Slovak-Czech Conference on
Geometry and Graphics 2021

Pág. 29 – 42.

Geometric Problems that Take
Possession of Me
→ Gunter Weiss

R2) sbornik_2022.pdf

Pág. 191 – 200

Symmetry – a natural phenomenon
causing "homo sapiens" to become
a "homo mathematicus"
→ Gunter Weiss

Off: Pedal Triangles
jwilson.coe.uga.edu

Assignment 9:
"Pedal Triangles"
by
Margo Gonterman.

Source: http://jwilson.coe.uga.edu/

EMAT6680Fa09/Gonterman/Gonterman9/
Gonterman9.html

Obs.: Difference between Orthic triangle and Pedal Triangle.

Source: https://math.stackexchange.com/questions/3146011/difference-between-orthic-triangle-and-pedal-triangle

A pedal triangle is defined by an arbitrary point. (The vertices of the pedal triangle are where the feet of perpendiculars from that point meet the side-lines of the original triangle).

The orthic triangle is specifically the pedal triangle determined by the original triangle's orthocenter

Return: Pedal Triangle

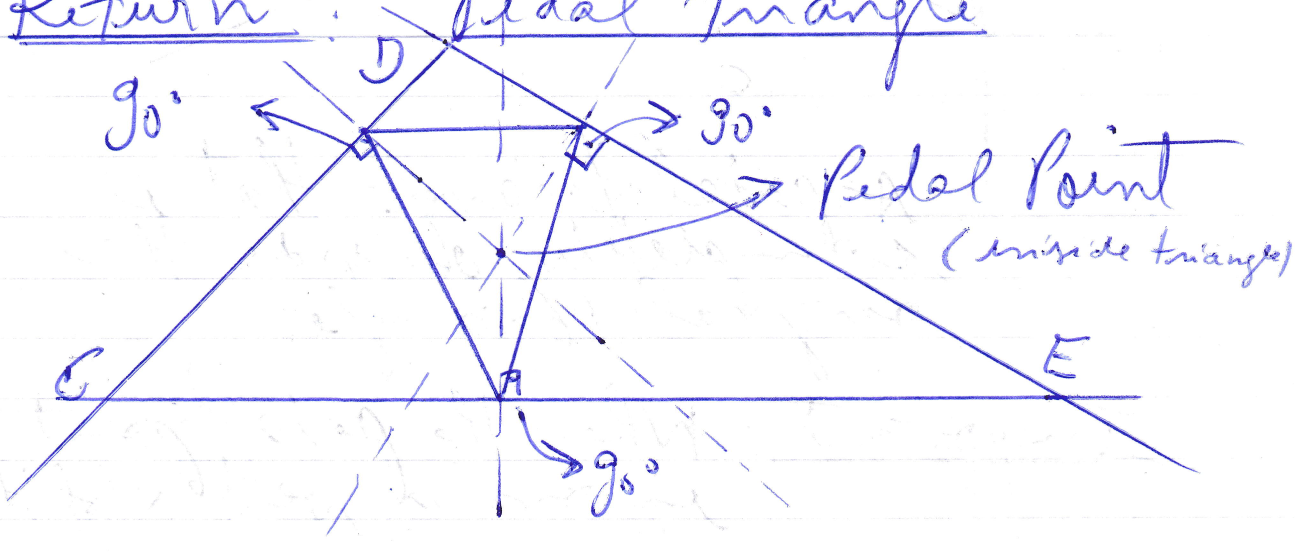

The pedal triangle has vertices that are the feet of the perpendiculars from the pedal point to each of the sides of the original triangle.

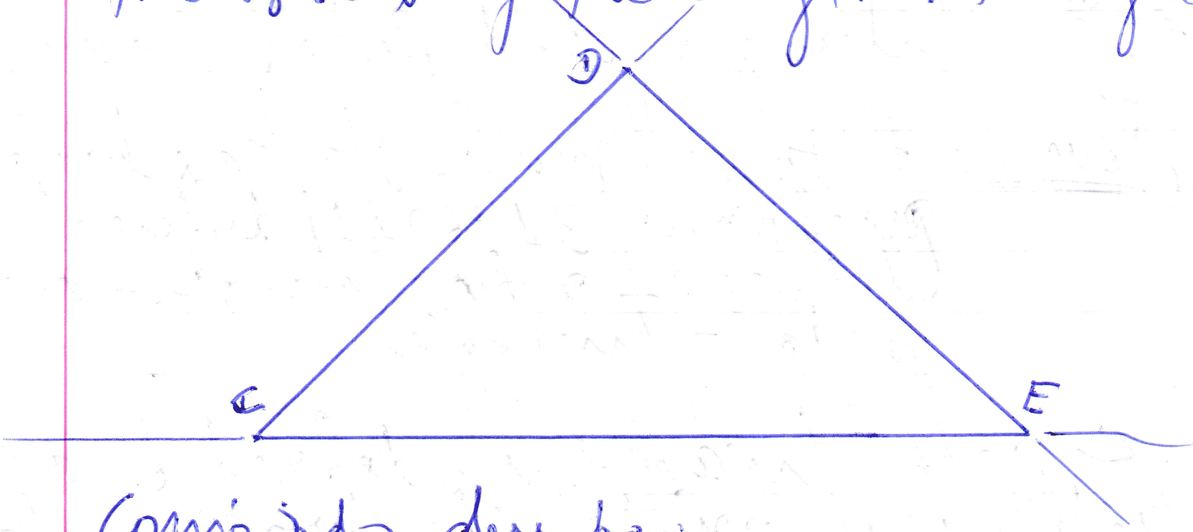

Corrigindo desenho:

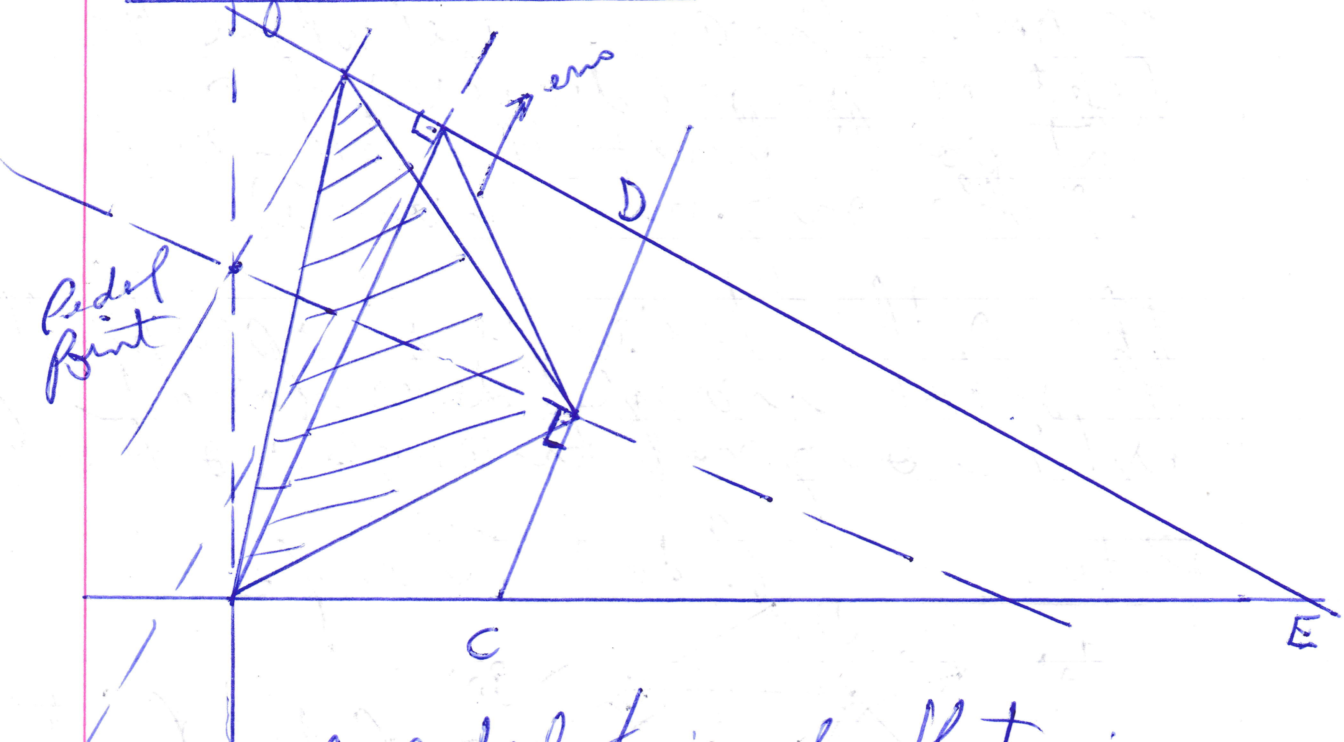

A pedal triangle that is completely outside the original triangle.

Important: When the pedal point is on a vertex of the original triangle, the pedal triangle degenerates into the altitude of the original triangle.

Obs:

a) When the triangle is acute, the pedal triangle (altitude) is inside the original triangle.

b) When the triangle is right, the pedal triangle (altitude) lies along one of the side of the original triangle.

c) When the triangle is obtuse the pedal triangle (altitude), lies outside the original triangle.

Obs: Centroid of a triangle is the point where the three medians of a triangle meet.

Orthocenter of a triangle is the intersection point of the altitudes drawn from the vertices of the triangle to the opposite sides.

R2 - cont. Pag. 200

Fig 16 - of all triangles inscribed

to a triangle the altitude pedal point triangle has the shortest perimeter.

↳ very good: eficiência energética.

Pág. 198 - Symmetry as an abstract elementary mathematical concept.

See: Fig. 15 and 16.

See:

a) "three reflections theorem"

b) "Steiner symmetrisation"

c) "Dido's extreme value problem"
→ of all planar domains with the same perimeter, the circle has the maximal area.

Other: Complementation

Pág. 201 - Higher-dimensional space arising from robot kinematics
→ Michal Zamboj

Pág. 203 → 2.1.

Constructing toroidal hypersurfaces of revolution

→ orthogonal projections from the four-dimensional space,

let us first remind images of a 2-sphere and torus of revolution in the Monge's projection (Fig. 3).

The object is orthogonally projected into two mutually perpendicular planes.

Therefore, each point has two conjugated images.

For further discussion, it is convenient to observe that rotation (e.g. of the blue generating circle in the figures) creates an affine image in the orthogonal projection.

See: Fig 4 (vai ser alterada)

A diagram with the consecutive construction of the hypersurfaces

$S^2 \times S^1$ and $S^1 \times S^1 \times S^1$ using rotations and translations, and

a construction of $S^1 \times S^1 \times S^1$
from a Clifford torus
(will change)

Pág. 206 – 2.2

Four-dimensional configuration space of a spherical joint and 3-R robot.

Fig. 5 — parametrization
(see)

R3) 2018_Weiss.pdf
9 pages – Gunter Weiss.

Source: https://home.pf.jcu.cz/~sbml/wp-content/uploads/2018_Weiss.pdf

Obs: ~sbml
↳ low.

Frégier points revisited
↳ Gunter Weiss.

Theorem 1: All right-angled triangles inscribed to a conic, s.t.
(Pag 01 → 84)
their right-angles vertices coincide have hypotenuses passing through on point, the Frégier point.

Importantíssimo.

Usar este conceito de Frégier point

↳ right-angle vertices coincide

combinar

com ligne droite, pois as duas linhas saem do mesmo ponto.

Portanto:

ligne droite (quadrilatère) + Frégier Point (conic).

✓

checar se a linha reta do quadrilátero passa pelo ponto de Frégier.

Obs: Caso sim → demonstrar

Caso não → ajustar para que passe (use a curvature space)

+ Pappu's theorem → 3 points

Veja também: retas paralelas e não paralelas.

See: Pág. 03(87) ⇒

Figure 5 – The Frégier map of a general line l is a conic l^γ.

Important:

The Frégier Mapping
A Quadratic Transformation

Place of action is the projective enclosed Euclidean plane P^2.

We declare the (planar) Frégier mapping
$\gamma: P^2 \Rightarrow P^2$ to fixed points
O and F follows:

Connect a given point P_1 with O, draw the line g_2 orthogonal to $OP_1 =: g_1$ through O and intersect g_2 with line $FP_1 =: f$.

The intersection points $P_2 =: P_1^\gamma$ acts as Frégier image of P_1, see Fig. 5.

Let P_1 run through a line l, then $f \to g_1$ is a perspectivity.

As $g_1 \to g_2$ is an involutoric projectivity, the mapping

$f \to g_2$ is at least a projectivity.

It is well-known that projective pencil $\{g_2\}$ and $\{f\}$ generate a conic l^γ

this in mind we can already recognise γ as a

quadratic transformation.

As P_2 maps to P_1 again, γ obviously is involutoric.

We base the analytic description of γ on a cartesian coordinate frame and use homogeneous coordinates:

$O = (1, 0, 0)\mathbb{R}$, $F = (1, 0, 1)\mathbb{R}$,

$P_1 = (1, x_1, y_1)\mathbb{R}$, $P_2 = (1, x_2, y_2)\mathbb{R}$

Pag. 5 (88) – 2.

A "Frégier Angle – at – circumference theorem"

Fig. 6 and 7

Fig. 7 – Sketch of a proof a "Frégier angle – at – circumference theorem"

Pag. 6 (89) – theorem 4

⇒ parallel axes. (important)

theorem 4 →

Triangles inscribed to a conic c have equal angles α at common vertex

$O \in c$

have sides a (opposite to O) enveloping a conic c_α.

This conic is centric similar to c with respect to the Frégier point F to O.

Pág. 6 (89) - 3

- 3D - Version of the Fregier Theorem 1.

We generalise the right angle book by three pairwise orthogonal lines a, b, c through the origin O of a cartesian 3-frame (O, x, y, z) and intersect them with a quadric Φ touching the x-y plane at O, see Fig. 9.

other:

Pág. 7 (91) - Theorem 6.

Orthoschemes $\{O; A, B, C)$

Important:

Theorem 7 → 2 retas saindo do mesmo ponto.
(ligne droite)

nos podemos simular de dois modos

Theorem 7 →

The set of Frégier-points $F^{(2)}$ to a quadric Φ is – in case of an ellipsoid and a (general) hyperboloid – a quadric concentric and coaxial with Φ, in case of a paraboloid Φ a coaxial paraboloid.

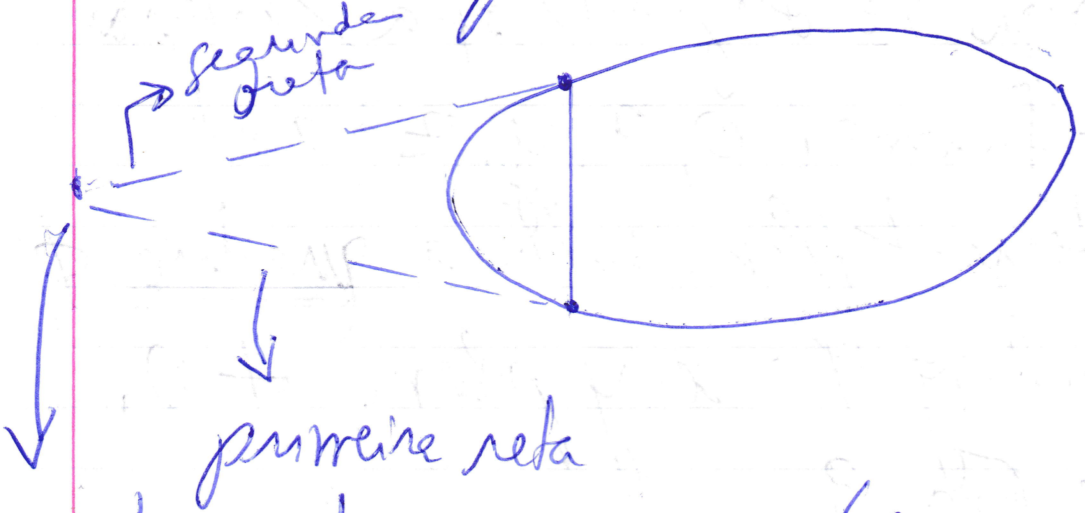

este ponto em comum (ligne droite)

simular em cima da Fig. 10.

R4) Wells – The Penguin Dictionary of Curious ~~.pdf~~ and Interesting Geometry-(1991).pdf

→ 308 pages.

Pág. 11 (28) –

Aubel's theorem →

Draw any quadrilateral.
It need not be convex,
and it doesn't even matter
if one of the sides is of zero
length.
Construct squares on all the
sides, facing outwards.
The line segments joining
the centres of opposite square
are equal in length, and
perpendicular.

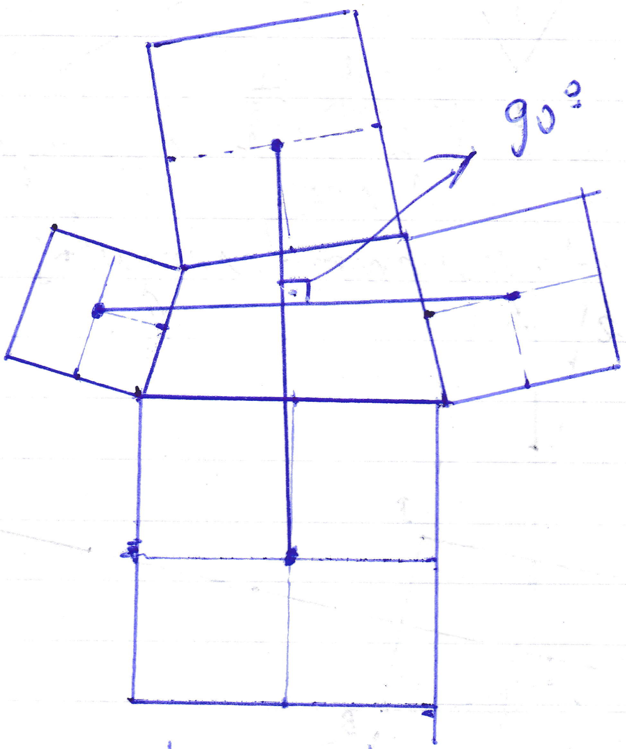

(interessantíssimo). comparar com os
dos triângulos.
(*) Procurar!!!

20231012183939

= x

20231013 1028

(I)

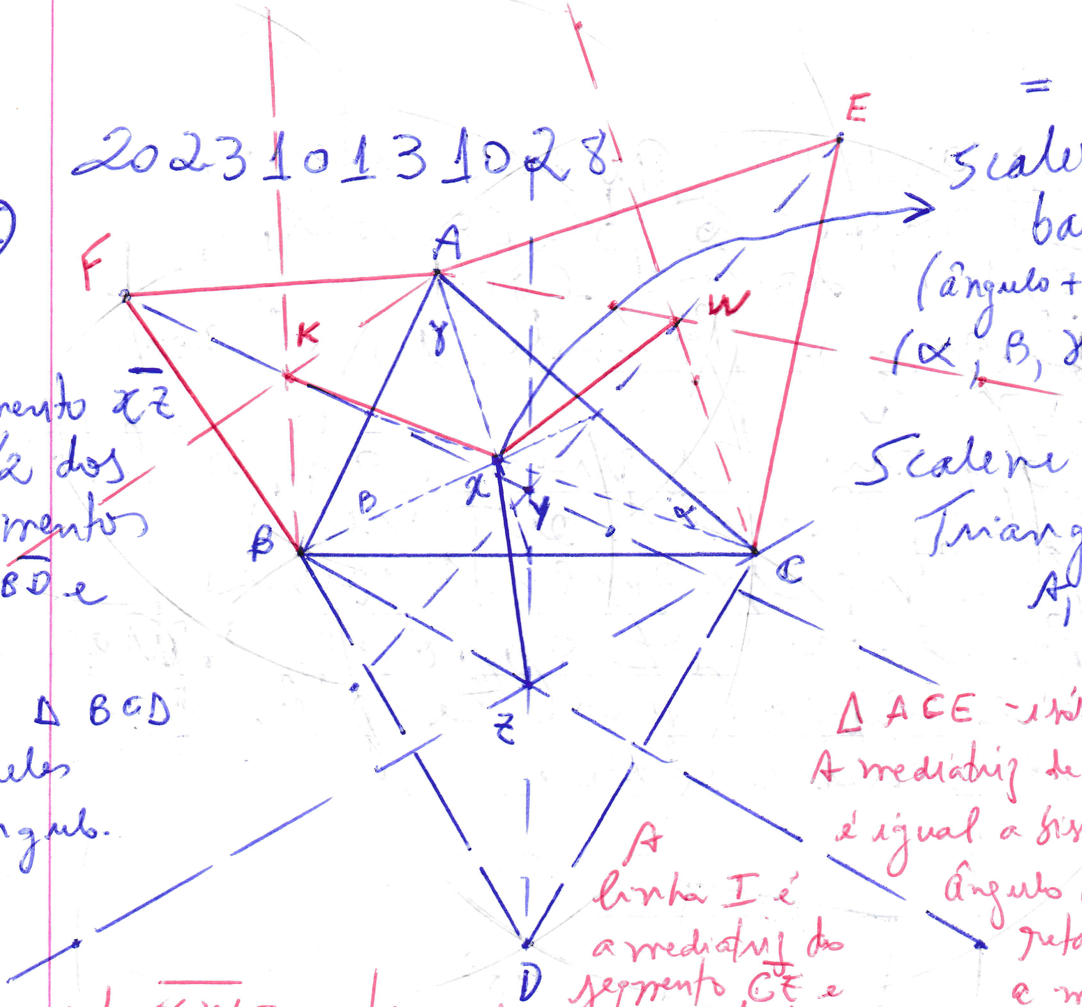

Scalene barycentre (ângulo + middle point) (α, β, γ)

Scalene Triangle A, B, C Δ.

Segmento $\overline{xz}$ é 1/2 dos segmentos $\overline{BC}$, $\overline{BD}$ e $\overline{CD}$.

Obs: Δ BCD isósceles triângulo.

Δ ACE - isósceles

A mediatriz de um lado é igual a bissetriz do ângulo oposto à reta que se fez a mediatriz (perpendicular barycenter)

A linha I é a mediatriz do segmento $\overline{CE}$ e bissetriz do ângulo A do triângulo Δ ACE.

Segmento $\overline{XW}$ = $\overline{xz}$ = $\overline{Kx}$

Em vez do quadrado, usaremos triângulos.

(II)

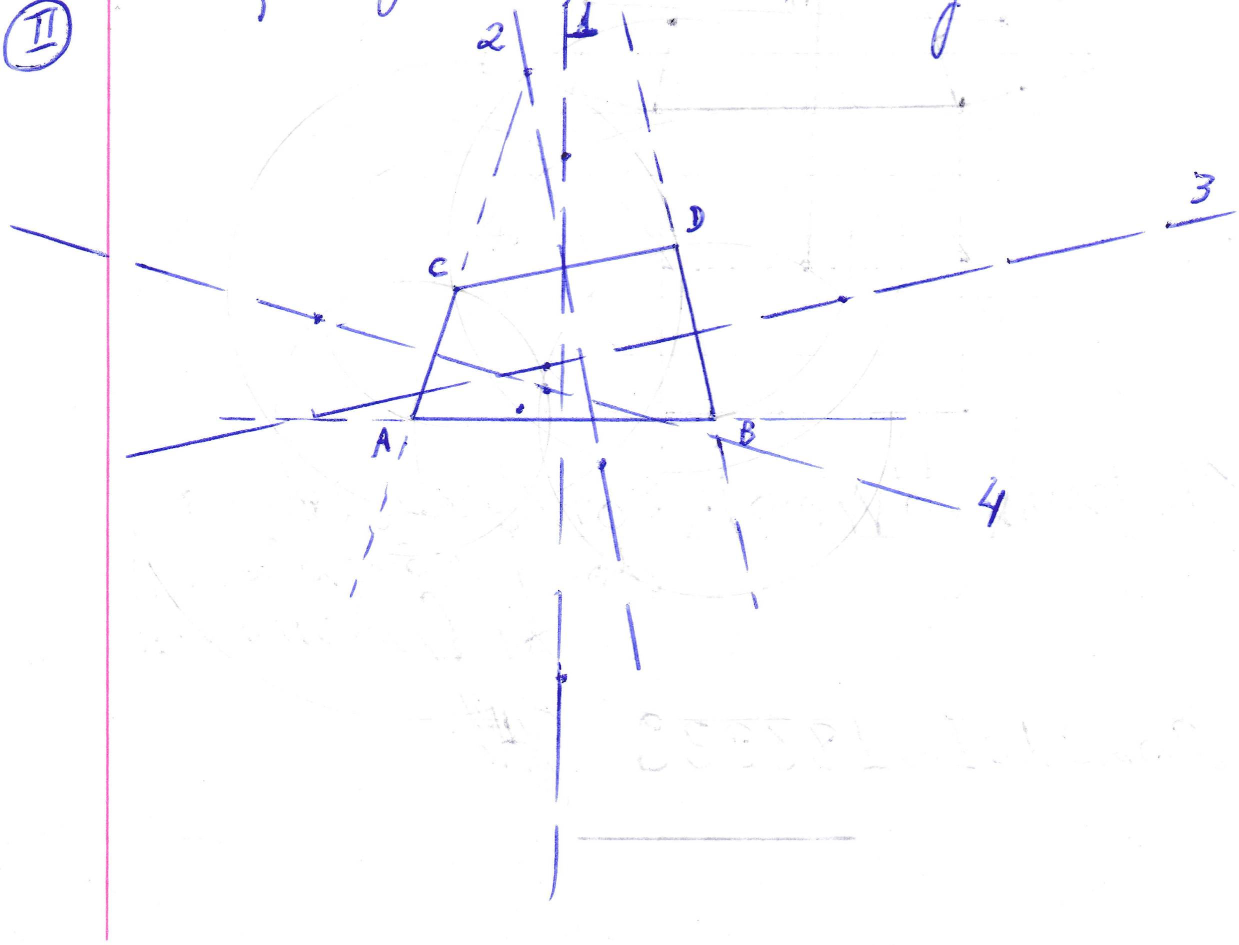

Estudo: (quadrilátero). (II)

a) Segmento $\overline{AB}$ é cortado perpendicularmente pela linha 1. (perpendicular bisector) - mediatriz.

b) Segmento $\overline{CD}$ é cortado perpendicularmente pela linha 2. (perpendicular bisector)

c) Segmento $\overline{DB}$ → linha 3.

d) Segmento $\overline{AC}$ → linha 4.

Análise: (das linhas)

a) Linha 1 → não faz ângulo de 90° com nenhuma reta (2, 3, 4).

b) Linha 2 → faz ângulo de 90° com a linha 3.

c) Linha 4 → faz um ângulo maior que 90° com a linha 1.

Interessante:

A Linha 1 e a Linha 2 se interceptam no ponto da Mediatriz (perpendicular bisector) que cruza o segmento CD.

Voltando: Scalene Triangle Δ ABC. (I)

No triângulo escaleno (scalene triangle) Δ ABC, ou seja, os 3 lados deste triângulo apresentam diferentes tamanhos.

Na hipótese, em que realizamos a mediatriz, utilizando-se "perpendicular bisector" do segmento em questão, como por exemplo, $\overline{AB}$, $\overline{BC}$ e $\overline{AC}$.

No triângulo escaleno, os três lados de um triângulo são diferentes, mas ao fazer a "perpendicular bisector" de cada lado do triângulo, nós obteremos os seguintes pontos F, D, E, que farão partes de 3 triângulos equiláteros.

Análise:

a) Triângulo escaleno (scalene triangle) → ∆ ABC.

b) "Perpendicular bisector" (mediatriz) criar os seguintes triângulos equiláteros:

b1) Segmento $\overline{AB}$ cria-se o ponto F. Obteremos ∆ ABF (equilátero)

b2) Segmento $\overline{BC}$ cria-se o ponto D. Obteremos ∆ BCD (equilátero)

b3) Segmento $\overline{AC}$ cria-se o ponto E. Obteremos ∆ ACE (equilátero)

c) No caso do ∆ ABF, ao realizarmos as mediatrizes dos respectivos lados do triângulo equilátero (perpendicular bisector), concluiremos que:

c1) A mediatriz de um lado é igual a bissetriz do ângulo oposto ao segmento em que se fez a mediatriz, constituindo-se "perpendicular barycenter".

Explicando: No triângulo equilátero as mediatrizes (perpendicular bisector) corresponde a "perpendicular barycenter", pois o "barycenter" se obtém ligando-se o "middle point" com o ângulo oposto, dividindo-se este ângulo ao meio.

Importante: No triângulo escaleno (scalene triangle), as mediatrizes dos segmentos $\overline{AB}$, $\overline{BC}$ e $\overline{AC}$ são

obtidas no ponto diferente do ponto "scalene barycenter", que no caso é o ponto x, enquanto o outro ponto é o y.
c2) Ao realizarmos as mediatrizes do triângulo equilátero Δ ABF, obteremos o ponto K, este ponto é a intersecção dos 3 bissectores dos ângulos ABF. ✓
A distância entre o ponto K e o ponto "scalene barycenter" é igual aos outros 2 triângulos equiláteros Δ BCD e Δ ACE.
d) No caso do Δ BCD, ao realizarmos as mediatrizes dos referidos segmentos $\overline{BC}$, $\overline{BD}$ e $\overline{CD}$, ou seja, "perpendicular bijéctor", obteremos na intersecção destes segmentos o ponto Z.
A distância entre $\overline{xz}$ é igual aos segmentos Kx e $\overline{Wx}$.
e) No caso do Δ ACE, ao realizarmos as mediatrizes dos referidos segmentos $\overline{AC}$, $\overline{AE}$ e $\overline{CE}$, ou seja, "perpendicular bijéctor", obteremos na intersecção destes segmentos o ponto W.
A distância entre $\overline{xW}$ é igual aos segmentos $\overline{Kx}$ e $\overline{Zx}$.
Estudar (não sei se é aplicável a todos os triângulos escalenos).
O segmento $\overline{BC}$ tem 6 cm, ao ligarmos os pontos K, W, Z ao ponto x, obteremos 3 segmentos com o mesmo comprimento: $\overline{xZ}$, $\overline{xW}$ e $\overline{xK}$ com 3 cm cada um, mas testes devem ser feitos

com outros comprimentos de reta, que no caso é o segmento BC (número par), fazer com número ímpar 7 etc)

20231013 1200.

20231013 1301.

Frégier Theorem – Markus Heiss

Source: https://triangle-geometry.jimdofree.com/fr%C3%A9gier-s-theorem/

→ heiss@gmx.de

triangle-geometry.jimdofree.com/frégier-s-theorem/

Frégier's Theorem:

Pick any point on a conic section (here ~~an~~ an ellipse).

Draw several rectangular triangles, having the right angles at this point.

The hypothenuses intersect in one common point, the so-called Frégier point.

The Frégier point S is given and point P on the curve shall be constructed:

Frégier point S and Ellipse (or another conic section) ⇒ Construction of the "Ex-Frégier point" P:

Intersection of at least two circles, where their diameters are two chords through the Frégier point S

Obs. Point S has to fulfill some conditions.

Fig. 02.

Center of the Kiepert hyperbola on the Steiner inellipse – Fig. 03

Obs:

X115 = center of the Kiepert hyperbola

K = X6 = symmedian point

X99 = Steiner point.

See:

X115 is the fourth intersection of Steiner inellipse and nine-point circle.

Obs:

nine-point-circle →

Important:

The symmedian point K is the Frégier point of X115 with respect to the Steiner inellipse.

It shows up that the symmedian point K (= X6) is the Frégier point of the center of the Kiepert hyperbola (= X115) with respect to the Steiner inellipse.

Reversely said:

Is the Frégier point the symmedian

point K with respect to the Steiner inellipse, then the "Ex-Fregier point" is the center of the Kiepert hyperbola (=X115).

Analise:

a) Center of the Kiepert hyperbola.

202310150900

Kiepert Hyperbola is a hyperbola and triangle conic that is related to the solution of Lemoine's problem and its generalization to isosceles triangles constructed on the sides of a given triangle.

The vertices of the constructed triangles are given in trilinear coordinates by

$$A' = -\sin\phi : \sin(C+\phi) : \sin(B+\phi)$$
$$B' = \sin(C+\phi) : -\sin\phi : \sin(A+\phi)$$
$$C' = \sin(B+\phi) : \sin(A+\phi) : -\sin\phi$$

where ϕ is the base angle of the isosceles triangle.

Kiepert (1869) showed that the lines connecting the vertices of the given triangle and responding peaks of the isosceles triangles concur.

Obs: Concurrent → two or more lines are said to be concurrent if they intersect in a single point. Two lines concur if their trilinear coordinates satisfy

$$\begin{vmatrix} l_1 & m_1 & n_1 \\ l_2 & m_2 & n_2 \\ l_3 & m_3 & n_3 \end{vmatrix} = 0$$

Three lines concur if their trilinear coordinates satisfy

$$l_1 \alpha + m_1 \beta + n_1 \gamma = 0$$

$$l_2 \alpha + m_2 \beta + n_2 \gamma = 0$$

$$l_3 \alpha + m_3 \beta + n_3 \gamma = 0$$

in which case the point is

$$m_2 n_3 - n_2 m_3 : n_2 l_3 - l_2 n_3 : l_2 m_3 - m_2 l_3.$$

Three lines are concurrent if their coefficients satisfy

$$\begin{vmatrix} A_1 & B_1 & C_1 \\ A_2 & B_2 & C_2 \\ A_3 & B_3 & C_3 \end{vmatrix} = 0.$$

Trilinear Coordinates:

Given a reference triangle ΔABC, the trilinear coordinates of a point P with respect to ΔABC are an ordered triple of numbers, each of which is proportional to the directed distance from P to one of the side lines.

Trilinear coordinates are denoted $\alpha : \beta : \gamma$ or (α, β, γ) and also are known as homogeneous coordinates or "trilinears".

Trilinear coordinates were introduced by Plücker in 1835.

Since it is only the ratio of distances that is significant, the triplet of trilinear coordinates obtained by multiplying a given triplet by any nonzero constant describes the same point, so

$$\alpha : \beta : \gamma = \mu\alpha : \mu\beta : \mu\gamma.$$

For simplicity, the three polygon vertices, A, B, and C of the triangle are commonly written as

$$1:0:0,\quad 0:1:0,\quad \text{and } 0:0:1,$$

respectively.

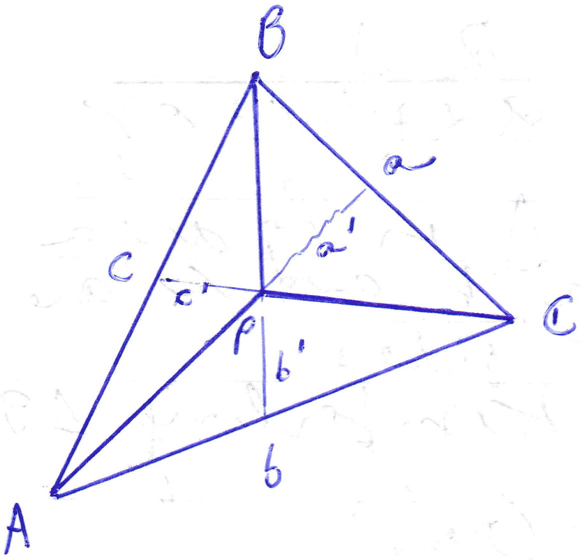

Trilinear coordinates can be normalized so that they give the actual directed *distances* from P to each of the sides. To perform the normalization, let the point P in the above diagram have trilinear coordinates
$\alpha : \beta : \gamma$ and lie at distances
a', b', and c'
from the sides
BC, AC, and AB, respectively.
Then the distances
$a' = K\alpha$, $b' = K\beta$, and $c' = K\gamma$
can be found by writing Δ_a for the area of ΔBPC, and similarly for Δ_b and Δ_c.

We then have $\Delta = \Delta_a + \Delta_b + \Delta_c$

$$= \frac{1}{2}aa' + \frac{1}{2}bb' + \frac{1}{2}cc'$$

$$= \frac{1}{2}(ak\alpha + bk\beta + c\beta\gamma)$$

$$= \frac{1}{2}K(a\alpha + b\beta + c\gamma)$$ so

$$K \equiv \frac{2\Delta}{a\alpha + b\beta + c\gamma}$$

where Δ is the area of ΔABC and a, b, and c are the lengths of its sides
(Kimberling 1998, pp. 26-27).

Book: (ref.)

Kimberling - "Central Points and Central Lines in the Plane of a Triangle"
Math. Mag. 67, 163-187, 1994.

Kimberling - "Triangle Centers and Central Triangles"
Congr. Numer. 129, 1-295, 1998.

To obtain trilinear coordinates giving the actual distances, take $k = 1$, so we have the coordinates
$a' : b' : c'$.

These normalized trilinear coordinates are known as exact trilinear coordinates.

The trilinear coordinates of the line

$$ux + vy + wz = 0$$

are $u : v : w = a d_A : b d_B : c d_C$,

where d_i is the point-line distance from polygon vertex i to the line.

Obs: Point-Line Distance – 2 Dimensional

✡

The distance between a point with exact trilinear coordinates $(\alpha', \beta', \gamma')$ and a line

$l\alpha + m\beta + n\gamma = 0$ is

$$d = \frac{|l\alpha' + m\beta' + n\gamma'|}{\sqrt{l^2 + m^2 + n^2 - 2mn\cos A - 2nl\cos B - 2lm\cos C}}$$

(Kimberling 1998, p. 31)

Obs: Important:

Converter para o espaço curvo (distância será diferente) – aplicar

M.U.S.

Análise:

b) Symmedian Point.

In geometry, the Lemoine point, Grebe point or symmedian point is the intersection of the three symmedian

(medians reflected at the associated bisectors) of a triangle.

Symmedian Point

The point of concurrence K of the symmedians, sometimes also called the Lemoine point (in England and France) or the Grebe point (in Germany).

Symmedians

Symmedians are three particular lines associated with every triangle. they are constructed by taking a median of the triangle (a line connecting a vertex with the midpoint of the opposite side), and reflecting the line over the corresponding angle bisector (the line through the same vertex that divides the angle there in half).

The angle formed by the symmedian and the angle bisector has the same measure as the angle between the median and the angle bisector, but it is on

the other side of the angle bisector.

Important:

a) No triângulo Δ ABC, ao fazermos os ângulos bisectors, nós encontramos um ponto I (concur at the incenter)

b) Ao fazermos no mesmo triângulo Δ ABC, as medianas (Medians), nós obteremos o ponto G (concur at the centroid)

Medians → line connecting a vertex with the midpoint of the opposite side.

c) No mesmo triângulo, ao fazermos a reflexão (reflecting) da mediana obteremos Symmedians (concur at the symmedian point L) que se interceptam no ponto L.

The three symmedians meet at a triangle center called the Lemoine point

Ross Honsberger has called its existence

"one the crown jewels of modern geometry".

Chapter 7: The Symmedian Point. (1995)

→ Episodes in Nineteenth and Twentieth Century Euclidean Geometry (1995).

Voltando (return):

Symmedian Point

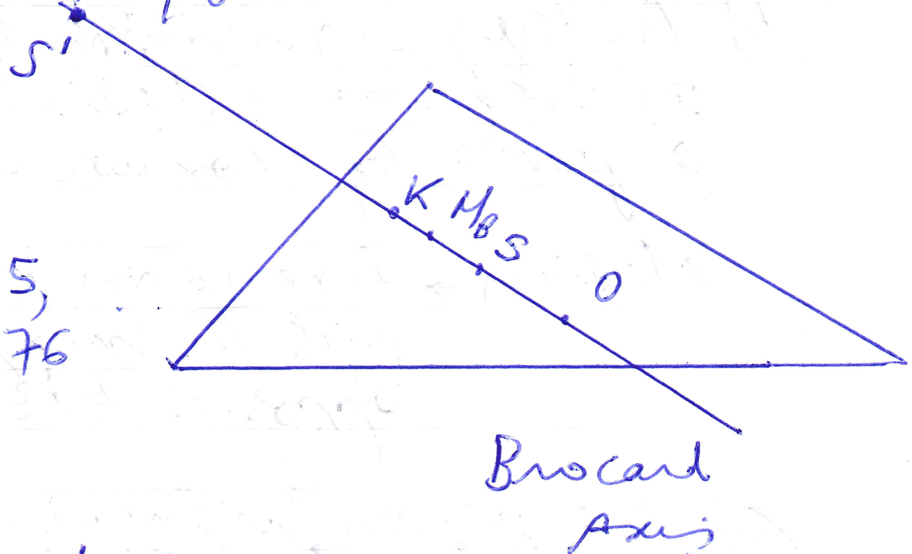

~~Hon~~
Honsberg 1995, p.76

The symmedian point lies on the Brocard axis (The Brocard axis is the line K O passing through the symmedian point K and circumcenter O of a triangle, where the segment OK is the Brocard diameter (Kimberling 1998, p.150).
The Brocard axis is perpendicular to the Lemoine axis and is the isogonal conjugate (the isogonal conjugate X^{-1} of a point X in the plane of the triangle ΔABC is constructed by reflecting the lines AX, BX, and CX about the angle bisectors at A, B, and C. The three

reflected lines then concur at the isogonal conjugate (Honsberger 1995, pp. 55-56).

(... Brocard axis is perpendicular to the Lemoine axis and is the isogonal conjugate of the Kiepert hyperbola.)

... The symmedian point lies on the Brocard axis and the Fermat axis. (Fermat axis is the central line connecting the first and second Fermat points! It has line function

$$l = a(b^2-c^2)(a^2-b^2-bc-c^2)(a^2-b^2+bc+c^2).$$

corresponding to Kimberling center.

Obs: 1) First Fermat Point

The first Fermat point X (or F_1) (sometimes simply called "the Fermat point", Torricelli point, or first isogonic center) is the point X which minimizes the sum of distances for A, B and C in an acute triangle,

$$|AX| + |BX| + |CX|.$$

It has equivalent triangle center functions

$$\alpha_{13} = \csc\left(A + \frac{1}{3}\pi\right)$$

$$\alpha_{13} = bc[c^2a^2 + (c^2+a^2-b^2)^2]$$
$$[a^2b^2 - (a^2+b^2-c^2)^2]$$
$$[4\Delta - \sqrt{3}(b^2+c^2-a^2]$$

and is Kimberling center X_{13} (Kimberling 1998, p. 67).

obs: It also arises in Napoleon's theorem.

obs) 2: Second Fermat Point

The second Fermat point X' or F_2 (also known as the second isogonic center) can be constructed by drawing equilateral triangles on the inside of a given triangle and connecting opposite vertices. The three diagonals in the figure then intersect in the second Fermat point, which has triangle center function

$\alpha = \csc\left(A - \frac{1}{3}\pi\right)$ and is

Kimberling center X_4. (141
= Kimberling 1998, p. 68).

Return:

... the symmedian point lies on the Brocard axis and the Fermat axis. It lies of the Brocard circle and is the center of the cosine circle. It also lies on the Jerabek hyperbola (The Jerabek hyperbola is a circumconic that is isogonal conjugate of the Euler line. (Kimberling 1998, p. 237).

↳ see:

Euler line (The line on which the orthocenter H, triangle centroid G, circumcenter O, de Longchamps point L, nine-point N, and a number of other important triangle center lie)

Obs: Longchamps point (L) is the reflection of the orthocenter H about the circumcenter O of a triangle. It has triangle center function

$$\alpha = \cos A - \cos B \cos C,$$

and is Kimberling center X_{20} (Kimberling 1998, p. 70).

return : Euler Line

... The Euler Line is perpendicular to the Longchamps lines and orthic axis.

return : Symmedian Point

... It also lies on the Jerabek hyperbola and the Thomson cubic.

(the Thomson cubic $Z(X_2)$ of a triangle ΔABC is the locus the center of circumconics whose normals at the vertices are concurrent. It is a self-isogonal cubic with pivot point at the triangle centroid, so its parameter is $x = bc$ and its trilinear equation is given by

$$bc\alpha(\beta^2+\gamma^2)+ca\beta(\gamma^2-\alpha^2)+ab\gamma(\alpha^2-\beta^2)=0$$

(Cundy and Parry 1995; Kimberling 1998, p. 240).

It is sometimes called the

seventeen-point cubic
(Casey 1893, p. 460;
Kimberling 1998, p. 240).

return: Symmedian Point

Its distances from K to the sides of the triangle are

$$KK_i = \frac{1}{2} a_i \tan \omega.$$

where
ω is the Brocard angle.

Análise:

c) Steiner point (Fregier Theorem → / Heiss.)

obs: Steiner Circle
is the central circle with center at the nine-point center X_5 and radius

$$R_5 = \frac{3}{2} R,$$

where R is the circumradius
(the circumradius of a cyclic polygon is a radius of the circle inside which the polygon can be inscribed. Similarly the circumradius of a polyhedron is the radius of a

circumsphere touching each of the polyhedron's vertices, if such a sphere exists).

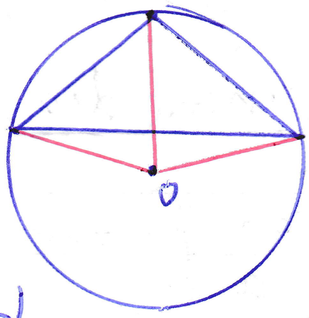

return: Steiner Circle

It is the circumcircle of the Steiner deltoid.

Steiner Points:

there are (at least) three different types of points known as Steiner points.

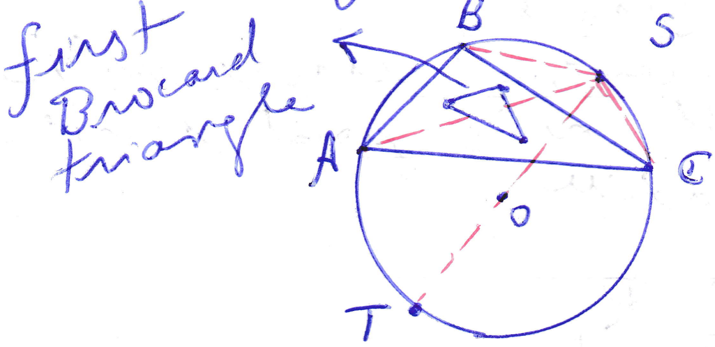

The point S of concurrence of the three lines drawn through the vertices of a triangle parallel to the corresponding sides of the first Brocard triangle is called the Steiner point (Honsberger 1995).

It lies on the circumcircle opposite the Tarry Point T and has equivalent triangle center functions

$$\alpha = bc(a^2-b^2)(a^2-c^2)$$

$$\alpha = \frac{1}{a(b^2-c^2}$$

obs: It also lies on the Steiner circumellipse.
It is Kimberling center X99.
The Brianchon point
(the point of concurrence of the joins of the vertices of a triangle and the points of contact of a conic of the triangle (Veblen and Young 1938, p. 111); Eddy and Fritsch 1994))

For the Kiepert parabola is also the Steiner Point (Eddy and Fritsch 1994).
The symmedian point K is the Steiner point of the first Brocard triangle (Honsberger, 1995, pp. 120-121).

The Simson Line (Importantíssimo)
↳ Research.

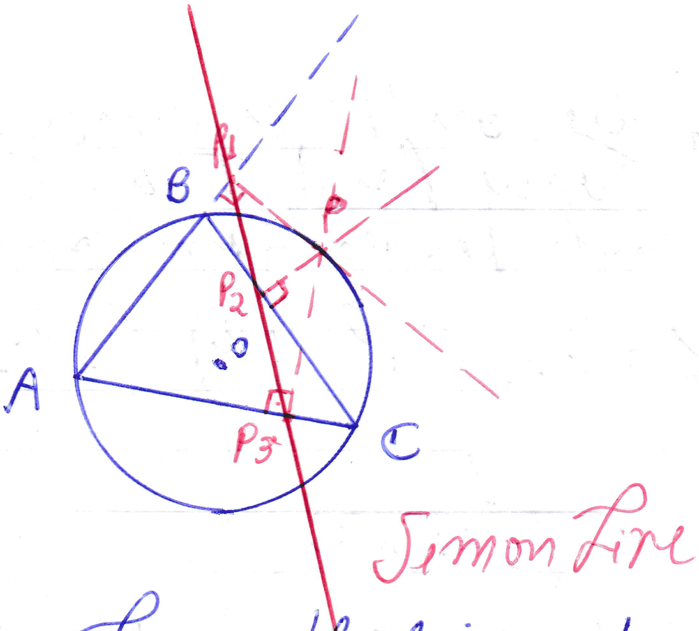

The Simon Line is the line containing the feet P_1, P_2, and P_3 of the perpendiculars from an arbitrary point P on the circumcircle of a triangle.

This line was attributed to Simson by Poncelet, but is now frequently know as the Wallace-Simson line since it does not actually appear in any work of Simson (Johnson 1929, p. 137; ~~Coxeter a~~ Coxeter and Greitzer 1967, p. 41; de Guzmán 1999).

Inverse statement:

The locus of all points P in the plane of a triangle $\triangle ABC$ such that the feet of perpendiculars from the three sides of the triangle is collinear is given by the circumcircle of $\triangle ABC$, is sometimes called the Wallace-Simson theorem (de Guzmán, 1999).

Important: The Simson line bisects the line HP, where H is the orthocenter
(Honsberger 1995, p.46).

Moreover, the midpoint of HP lies on the nine-point circle (Honsberger, 1995, pp. 46-47).

The Simson lines of two opposite point on the circumcenter of a triangle are perpendicular and meet on the nine-point circle
(also called Euler's circle or the Feuerbach circle, is the circle that passes through the perpendicular feet H_A, H_B, and H_C dropped from the vertices of any reference triangle ΔABC on the side opposite them.

Euler showed in 1765 that it also passes through the midpoints M_A, M_B, M_C of the sides of ΔABC.

By Feuerbach's theorem, the nine-point circle also passes through the midpoints E_A, E_B, and E_C of the segments ✡ that joint the vertices and the orthocenter H. These points are commonly referred to as the Euler points.

These three triples of points make nine in all, giving the circle its name.

Return: Finish (202310151400)

Frégier theorem → (Heiss)

2023101~~6~~1231

Frégier Theorem — Fig. 3

Center of the Kiepert Hyperbola X115 with respect to the Steiner inellipse

Obs: X115 is the fourth intersection of Steiner inellipse and nine-point circle.

X115 = center of the Kiepert hyperbola

K = X6 = symmedian point.

X99 = Steiner point.

Obs: The symmedian point K is the Frégier point of X115 with respect to the Steiner inellipse.

It shows up that the symmedian point K (= X6) is the Frégier point of the center of the Kiepert hyperbola (= X115) with respect to the Steiner inellipse.

Reversely said:

Is the Frégier point the symmedian point K with respect to the Steiner inellipse, then the "Ex-Frégier point" is the center of the Kiepert hyperbola (= X115).

O texto é longo mas de um modo geral, todos estes pontos serão úteis para manter a integridade dos cálculos, mas aplicável à uma superfície curva.

De acordo com o avançar da pesquisa, estes pontos serão projetados de forma ortogonal sobre uma superfície curva ou serão descritos na própria superfície curva.

Eu não preciso de tantos referenciais para fundamentar a minha pesquisa. O que interessa é comprovar a nova mudança de paradigma pois de um ponto de vista objetivo, os cálculos estão equivocados em virtude de terem sido obtidos com base em superfície plana.

Source: https://triangle-geometry.jimdofree.com

Importante:

Usar estes pontos como referencial para se manter a consistência dos gráficos, mas o que interessa é a lógica global do objetivo que se pretende obter.

Desejamos unir estes conceitos com Ligne Droite - Papelier e posteriormente vincular com o livro:

Sortais - La géométrie du triangle - Exercices résolus - (1997).pdf
→ 220 páginas - René Sortais

Pág. 08 (12) → Droite et cercle d'Euler.

20231017115000

Return to Frégier theorem

Fig. 02 → Frégier point S and Ellipse (or another conic section).

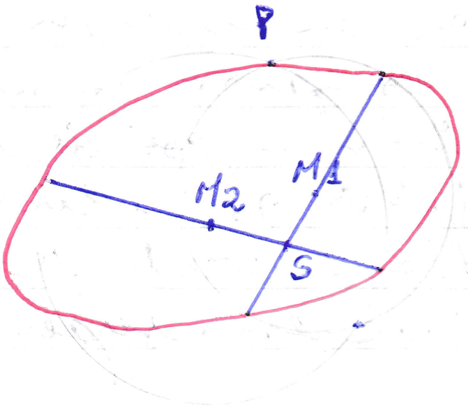

Neste exemplo, nós precisamos começar com uma ellipse e depois calcular o ponto S. (Frégier Point), posteriormente, nós inserimos os círculos.

Com esta representação, nós conseguimos interligar 2 círculos com tamanhos (raios) diferentes com uma ellipse.

Important:

Frégier Ellipses

Source: https://arxiv.org/abs/2206.03747 Ⓐ
Dominique Laurain - 7 pages. or
https://doi.org/10.48550/arXiv.2206.03747

We introduce Frégier ellipses which generalize the Frégier point in euclidean geometry. Subject is related to Dynamical Systems (grifo meu: usar estes estudos com base em minha pesquisa para desconstrui a idéia de gravidade, usar os novos conceitos M.U.S. → new geometry) and Poncelet's closure theorem,

[(grifo meu:

Source: http://olivernash.org/2018/07/08/poring-over-poncelet/assets/pdfs/BosKersOortRaven-1987.pdf
→ 76 pages.

other: A simple proof of Poncelet's Theorem (on the occasion of its bicentennial)

Source: http://user.math.uzh.ch/halbeisen/publications/pdf/poncelet.pdf
→ 14 pages. University of Zurich.

others: a) https://www.maxwell.vrac.puc-rio.br/25790/25790.pdf (Porismo de Poncelet)
b) https://www.maxwell.vrac.puc-rio.br/32326/32326.pdf (Porismo de Poncelet).]

Poncelet porism and displaying geometric invariants (area, angles).

Special angles pi/3 and 2pi/3 are emphasized and we have invariant sum of areas for Fregier circles in general Poncelet configuration, occurring from an affine map of the elliptical billiard.

Classificação: Metric Geometry (Math.MG); Dynamical Systems (math.DS).

8 Jun 2022.

Portanto: 2206.03747.pdf (base) ←

a) BoskersOortRaven-1987.pdf
b) poncelet.pdf
c) 25790.pdf
d) 32326.pdf

Obs: Dynamical Systems + Rigid Dynamics (NEWTON).

Complementar:

a) 2018_Weiss.pdf
b) Analytic Projective Geometry.pdf
c) Glaeser - the Universe of Conics.pdf
d) Glaeser - The Universe of Quadrics.pdf

Others:

a) 5000 years of Geometry.pdf
b) Advanced Newtonian Rigid Dynamics - 2023.pdf
c) Hartshorne - Deformation Theory (2010).pdf
d) Hartshorne - Geometry - Euclid and beyond.pdf

e) Ostermann- Geometry by its history-2012.pdf
f) Petersen- Riemannian Geometry (2016).pdf

Return: 2206.03747.pdf

Poncelet's closure theorem → also know as Poncelet's porism, states that whenever (a qualquer momento) a polygon is one conic section and circumscribes another one, the polygon must be part of an infinite family of polygons that are all inscribed in and circumscribe the same two conics.

Conic Section: The conic sections are the nondegenerate curves generated by the intersections of a plane (~~perpendicular~~) with one or two nappes of a cone.
For a plane perpendicular to the axis of the cone, a circle is produced.
For a plane that is not perpendicular to the axis and that intersects only a single nappe, the curve produced is either an ellipse or a parabola (Hilbert and Cohn-Vossen 1999, p.8)
The curve produced by a plane intersecting both nappes is a hyperbola (Hilbert and Cohn-Vossen 1999, pp.8-9).

return: Poncelet's Porism →

Hyperbola (Gray 1997, p. 45) is a conic section defined as the locus (a locus is the set of all points (usually forming a curve or surface) satisfying some conditions. The locus of points in the plane equidistant from a given point is a circle, and the set of points in three-space equidistant from a given point is a sphere) .. of locus of all points P in the plane the difference of whose (cujos) distances $r_1 = F_1P$ and $r_2 = F_2P$ from two fixed points (the foci F_1 and F_2) separated by a distance $2c$ is a given positive constant k,

$$r_2 - r_1 = k$$

(Hilbert and Cohn-Vossen 1999, p. 3)

Letting P fall on the left x-intercept requires that

$$k = (c+a) - (c-a) = 2a$$

so the constant is given by $k = 2a$.

Important: Rectangular hyperbola ⟹

The special case of the rectangular hyperbola corresponding to a hyperbola with eccentricity $e = \sqrt{2}$, was first studied by Menaechmus.

IMPORTANT:

The hyperbola is the ~~sp~~ shape of an orbit of a body on an escape trajectory (i.e. a body with positive energy), such as some comets, about a fixed mass, such as the sun.

obs:

i.e. (id est) → "that is"

e.g (exempli gratia) → "for example"

Return: conic section (Important)

Kepler was the first to notice that planetary orbits were ellipses.

(Ellipse is a curve that is the locus of all points in the plane the sum of whose distances r_1 and r_2 from two fixed points F_1 and F_2 (the foci) separated by a distance of $2c$ is a given positive constant $2a$ (Hilbert and Cohn-Vossen 1999, p. 2).

This results in the two-center bipolar coordinate equation

$r_1 + r_2 = 2a.$

→ major radius

where a is the semimajor axis and the origin of the coordinate system is at one of the foci.

The corresponding parameter b is known as the semiminor axis

↳ minor radius

The ellipse is a conic section and a Lissajous curve.
(parametric curve)

return : **KEPLER** was first to notice that planetary orbits were ellipses, and Newton was then able to derive the shape of orbits mathematically using calculus (?), under the assumption that

gravitational force

goes as the inverse square of distance.

IMPORTANT :

CONIC SECTION may more formally be defined as the locus of a point P that moves in the plane of a fixed point F called the focus and a fixed line d called the ...

... conic section directrix (is the line which, together to define a conic section as the locus of points whose (cujo) distance from the ffocus is proportional to the horizontal distance from the directrix), with

(with F not on d) such that the ratio of the distance of P from F to its distance from d is a constant e called the eccentricity

r being the constant of proportionality.

If the ratio: ✡

r = 1, the conic is a parabola,

r < 1, it is an ellipse, and

r > 1, it is a hyperbola.

(Hilbert and Cohn-Vossen 1999, p. 27)

return: ... called the eccentricity ✡

If e=0, the conic is a circle

If 0<e<1, the conic is an ellipse,

If e= 1, the conic is a parabola, and

If $e > 1$, it is a hyperbola.

A conic section with conic section directrix at $x=0$, focus at $(p, 0)$, and eccentricity $e > 0$ has Cartesian equation

$$y^2+(1-e^2)x^2-2px+p^2=0$$

(Yates 1952, p.36), where p is called the focal parameter

Plugging in p gives

$$y^2+(1-e^2)x^2-\frac{2a(1-e^2)}{e}x+\frac{a^2(1-e^2)^2}{e^2}=0,$$

for an ellipse,

$$y^2=4a(x-a),$$ for a parabola, and

$$y^2+(1-e^2)x^2-\frac{2a(e^2-1)}{e}x+\frac{a^2(e^2-1)^2}{e^2}=0$$

for a hyperbola.

Eff: The polar equation of a conic section with focal parameter p is given by

$$r = \frac{pe}{1 + e\cos\theta}.$$

Return: Poncelet's Porism

Poncelet Transverse

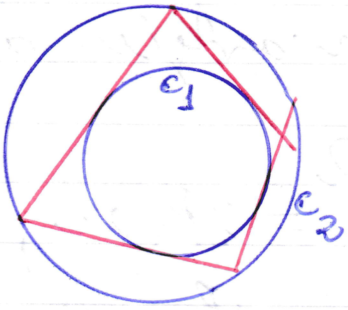

Let a C_1 lie inside another circle C_2. From any point on C_2, draw a tangent to C_1 and extend it to C_2.

From the point, draw another tangent etc. For n tangents, the result is called an n-sided Poncelet Transverse.

If, on the circle of circumscription there is one point of origin for which a four-sided Poncelet transverse is closed then the four-sided transverse will also close for any other point of origin on the circle (Dörrie 1965).

Obs: If the conic is taken as a circle (Casey 1888, pp. 124-126), then a polygon which has both an incenter and a circumcenter (and for which the transversals would therefore close) is called a bicentric polygon

(A polygon which has both a circumcircle (which touches each vertex) and a incircle (which is tangent to each side).

All triangles are bicentric with

$$R^2 - x^2 = 2Rr,$$

where R is the circumradius,
r is the inradius, and
x is the separation of centers.

For bicentric quadrilaterals, a result sometimes known as FUSS'S problem, the circles satisfy

$$2r^2(R^2+x^2) = (R^2-x^2)^2$$

(Dörrie 1965, Salazar 2006) or, in another form,

$$\frac{1}{(R-x)^2} + \frac{1}{(R+x)^2} = \frac{1}{r^2}$$

(Davis; Durége 1861; Casey 1888, pp. 109-110; Johnson 1929; Dörrie 1965).

Important: ✡

Para estabelecer parâmetros para a minha pesquisa, nós devemos observar os seguintes ~~pa~~ aspectos:

a) constant of proportionality:

$r = 1$ → parabola,

$r < 1$ → ellipse, and

$r > 1$ → hyperbola.

b) constant e (eccentricity)

$e = 0$, the conic is a circle,

$0 < e < 1$, the conic is ellipse,

$e = 1$, the conic is parabola, and

$e > 1$, the conic is hyperbola.

References:

a) Hilbert, D. and Cohn-Vossen, S, "The Cylinder, the Cone, the Conic Sections, and their Surfaces of Revolution."
→ Geometry and the Imagination
pp. 7-11 / 1999 (David Hilbert)

b) The seven circles theorem and
other new theorems
Evelyn 1974

c) Dörrie -
100 Great Problems of
Elementary Mathematics

obs: 1965 – 402 pages.

Return:

2206.03747.pdf (7 pages)

Frégier Ellipses

We introduce Frégier ellipses as generalization of the Frégier point and exhibit amazing angles and areas invariants.

IMPORTANT:

Utilizar esta invariabilidade para se manter os ângulos e as áreas invariáveis nos cálculos de projeção da superfície curva sobre um plano ortogonalmente. Antes de projeção estes cálculos servem para manter a consistência dos cálculos e na projeção haverá uma alteração nestes valores.

20231017172800

20231017214242

Solução para calcular a curvatura do espaço sobre a órbita elíptica dos planetas. (Uma delas). Estudar com profundidade.

Eu fiquei pensando em uma forma metódica para calcular a curvatura gerada pelas forças fundamentais do Universo sobre as órbitas elípticas e eu irei usar os mesmos procedimentos utilizados para analisar uma órbita elíptica.

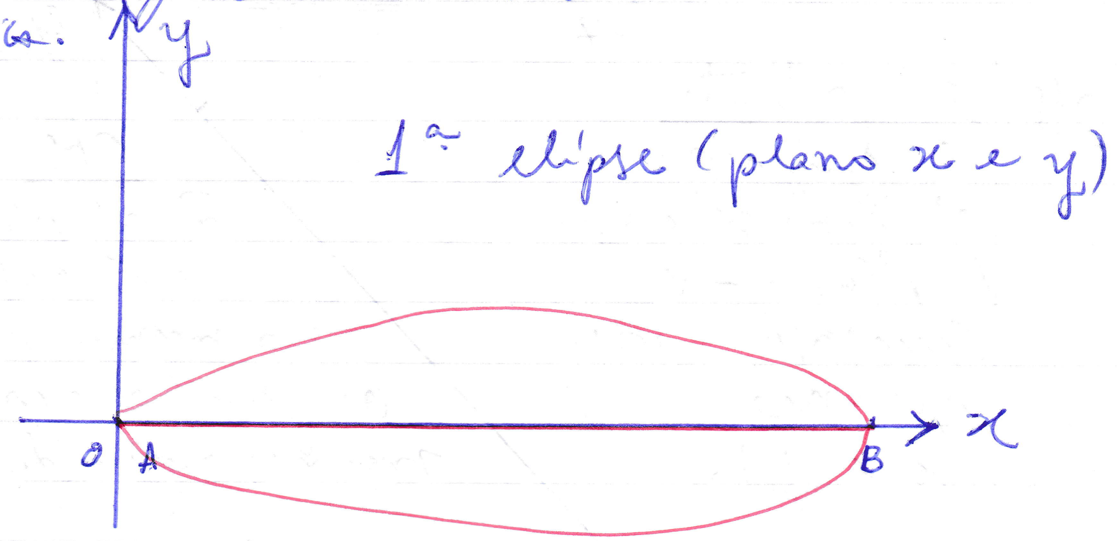
y
1ª elipse (plano x e y)
x
0
A
B

A 2ª elipse estará no eixo z, mas com um tamanho menor

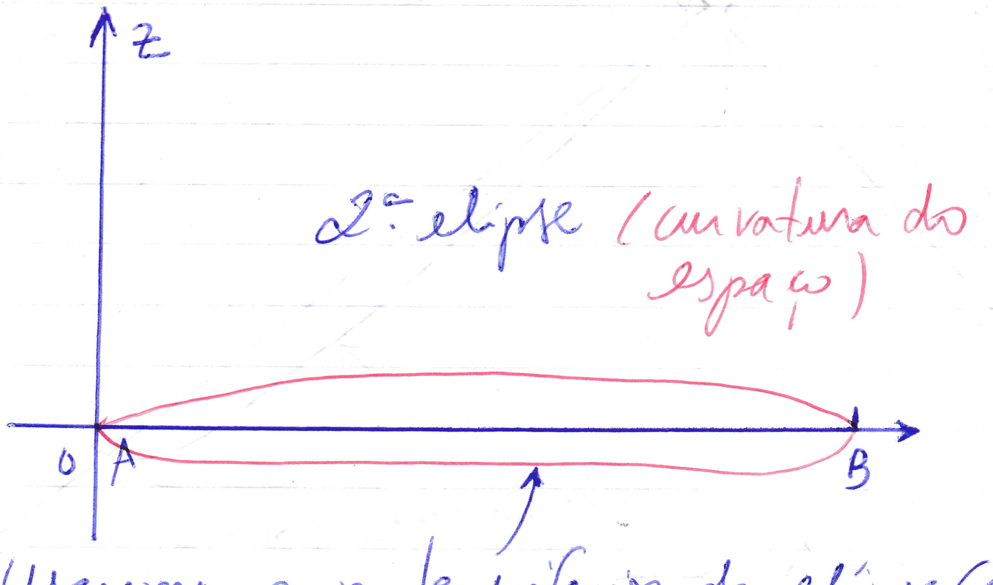
z
2ª elipse (curvatura do espaço)
0
A
B

obs: Usaremos a parte inferior da elipse (curvatura) corte longitudinal

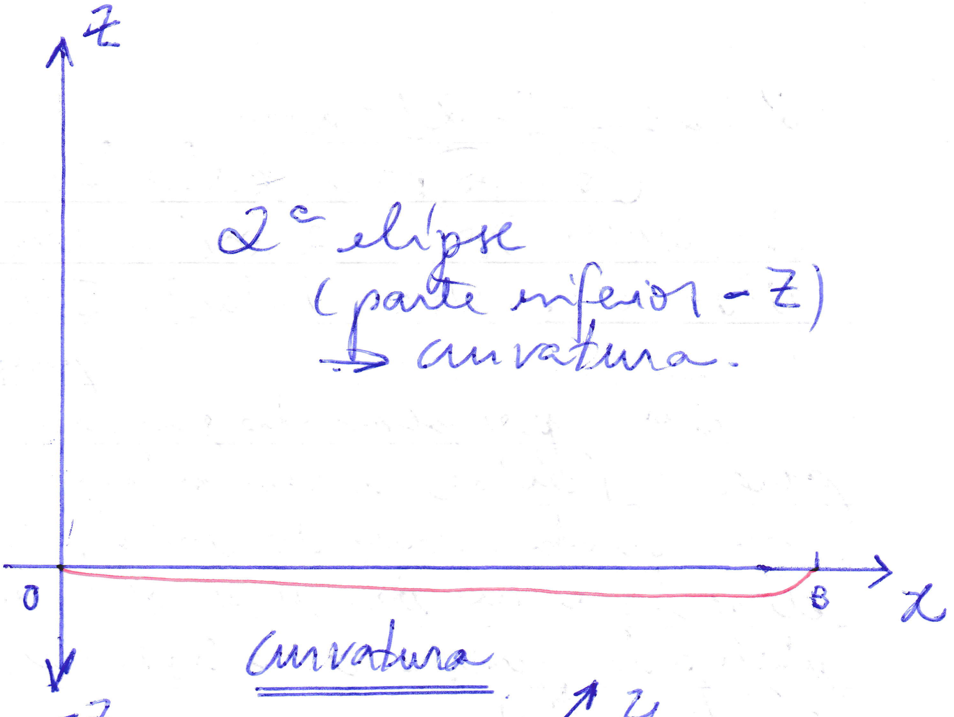

2ª elipse
(parte inferior - z)
⇒ curvatura.

obs: Importante
2 elipses:
a) plano x e y,
b) plano - z
(curvatura sobre
o plano
x e y).

obs: As
2 elipses
estão referenciados
no mesmo ponto de
origem (0) e com as
mesmas regras da elipse.

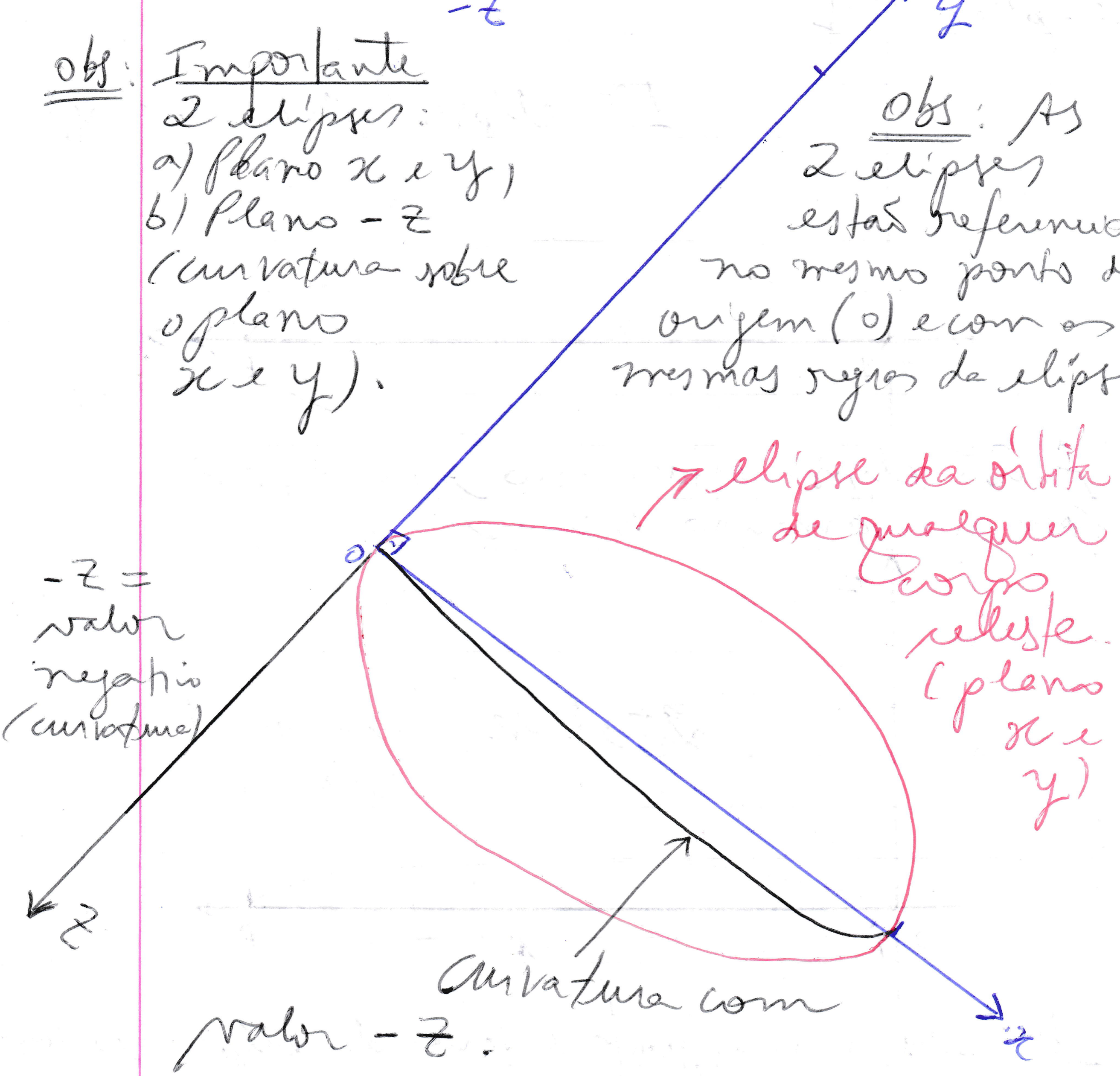

→ elipse da órbita
de qualquer
corpo
celeste.
(plano
x e
y)

-z =
valor
negativo
(curvatura)

curvatura com
valor - z.

Esta 2ª elipse será projetada (sobre) de forma ortogonal sobre um plano para o cálculo do novo número Pi (M.U.S.)

Vejamos, nós teremos 3 coordenados de valores números referentes à elipse, ou seja, x e y com a elipse orbital de um corpo celeste e z, ou seja, no eixo z mas com valor negativo em virtude da curvatura do espaço.

A obtenção destes 3 valores usarão as mesmas regras de elipse:

a) constant of proportionality

b) constant e (eccentricity).

Obs: A 2ª elipse terá um valor menor pois trata-se da curvatura do espaço.

Important: O valor negativo no eixo z poderá ser representada por uma equação que represente aquela curvatura.

Esta curvatura abrange toda a elipse da órbita do corpo celeste e a projeção ortogonal sobre um plano permitirá o cálculo do novo número do Pi (M.U.S.).

Important: Este aspecto será de fundamental importância para a chegagem dos cálculos obtidos, pois devemos utilizar este mecanismo nas duas elipses (x e y) e ($z \rightarrow -z$).

Primeiro, devemos escolher um ponto P sobre a elipse (x, y) e obter Frégier Point (S) e apartir do ponto P, nós deveremos criar dois círculos que se intercepta no ponto P. (círculos diferentes)

Não se esqueça de marcar os dois centros dos círculos (M1 e M2), pois estes pontos serão utilizados para marcarmos dois segmentos, ambos passando pelo ponto S, criando-se assim, o segmento ou reta que passa pelos pontos SM1 e SM2, sendo que ambos os segmentos estarão contidos nos respectivos círculos de tamanhos diferentes, M1 círculo menor e M2 círculo maior.

Important:

a) SM1 e SM2 devem ter o mesmo tamanho do diâmetro dos seus respectivos círculos.

b) SM1 e SM2 deverão atingir a elipse em ambos os pontos do segmento.

Interessante:

Eu tive uma idéia interessante, e se nós criarmos os dois círculos cujos dois pontos de intersecção fiquem justamente sobre a elipse em que nós obtemos o ponto Frégier Point (S) apartir do ponto P.

O primeiro ponto é o ponto P aleatório sobre a elipse.

O segundo ponto sobre a elipse

é o ponto Musich (M_e), ou seja,

Me = Musich elipse.

Desta forma, nós temos mais um referencial de checagem dos cálculos. Vejamos:

Alinhar os pontos Me1 S e M2, and Me3

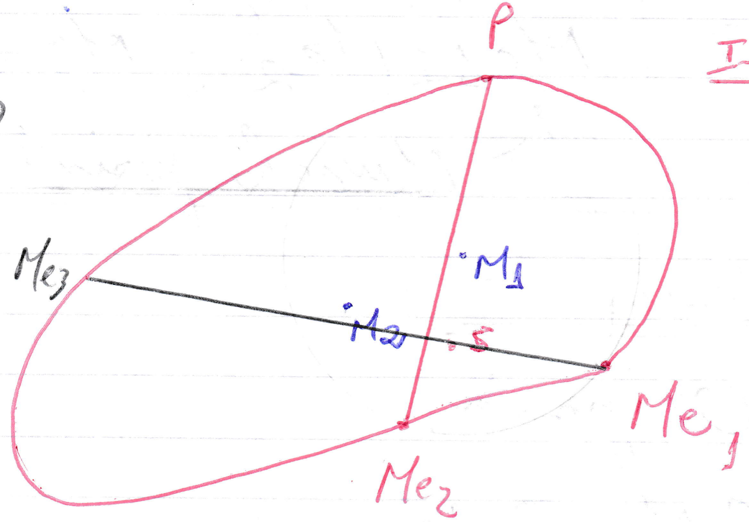

Important: Alinhar os pontos P M1 S e Me2

obs: Este é um tipo de configuração que eu criei.
(checar com cuidado!!!!)

obs: No primeiro exemplo, os dois círculos não há tanta convergência (colinear) dos pontos.

Nós podemos observar que:

SM_1 tem o diâmetro do primeiro círculo e os duas extremidades (estão) do segmento estão sobre o círculo, mas não está sobre nenhum ponto de intersecção entre os círculos. Observamos que somente o ponto P é uma intersecção entre os dois círculos M_1 e M_2, mas não estão sobre o outro ponto Me.

O segmento que passa pelos pontos S M_2 tem o diâmetro (acabou a caneta) do círculo M_2 e as duas extremidades estão sobre a elipse.

Obs: Checar os pontos Me_1, Me_2 e Me_3.
↳ Musich point.

Interessante:

Os pontos:

a) P M_1 S Me_2 → é uma reta com 4 pontos.

b) Me_1 S M_2 Me_3 → é uma reta com 4 pontos

Obs: São duas retas para dar maior consistência aos cálculos e com o ponto Me_1 (Musich point) nós teremos uma maior estabilidade nos cálculos entre os dois círculos e estas duas retas (checar os cálculos).

20231017231700

20231018153500

Construção Geométrica da Elipse utilizando-se régua e compasso:

- Trace o eixo maior
- Extremidades são os vértices A' e A', marcando os focos F e F' e o centro O.

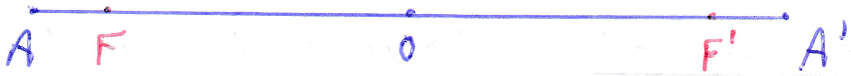

A partir de F e F', trace arcos de tamanhos iguais marcando os pontos 1, 2, 3 e 4 e 1', 2', 3' e 4', ajuste de acordo com a reta

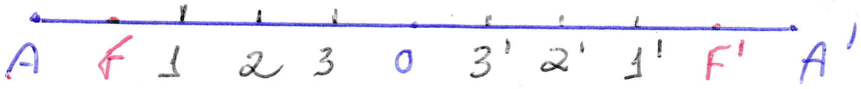

Agora, com a ponta seca do compasso ✪ em F e (arcos ou raios) A1', A2', A3'.... e depois com centro em F' e arcos A'1, A'2, A'3... trace os arcos. IMPORT: Compasso A1', depois mova para F e trace arco F! ✩

obs.: Estes procedimentos sempre no segmento $\overline{AA'}$.

Com centro em F e raios A1, A2, A3... e depois com centro em F' e raios A'1', A'2', A'3'..., trace arcos de modo a interceptarem os primeiros.

A curva que passa pelas intersecções e pelos vértices A e A' é a elipse de centro O e focos F e F'.

O segmento de reta ortogonal ao eixo AA' ~~que~~ passa pelo centro O, é o eixo menor, cujas extremidades são as interseções com a elipse formando seus vértices B e B'.

Important:

Neste exemplo, a base desta elipse é o Focus, ou seja, a sua unidade de medida. Nós devemos ajustá-la para adequá-la à realidade de nossa pesquisa.

Construção: 10 unidades ou 12 unidades.

obs: (Com relação aos pontos 3 e 3', podemos o ponto A → A3' e A' → A'3.)
→ wrong
muda-se a curvatura.

obs. O segmento A3' guarda em F traços apenas o arco em 2 (A2).

3 e 3' ficam de fora.

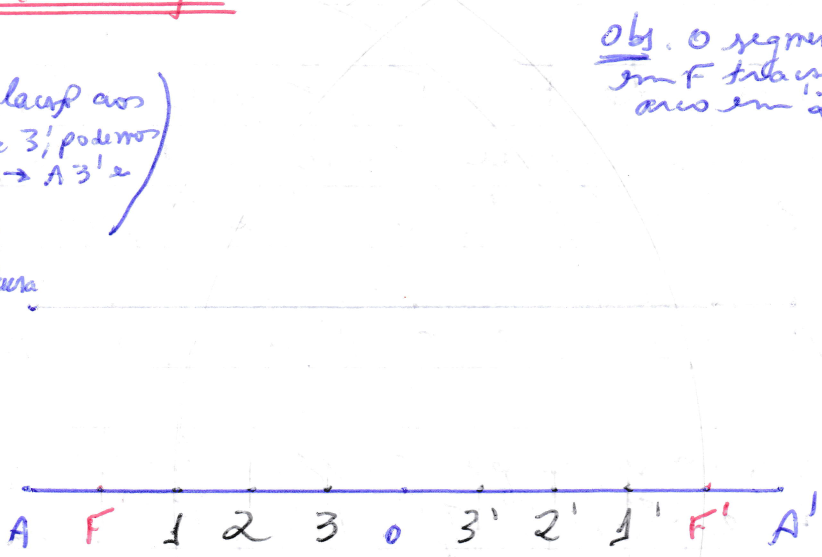

Neste caso, erros na realização do arco em 3 e 3' e ainda F. (Ficar atento!!)

Important:

O segmento A A' deve estar contido numa reta maior para que possamos traçar os arcos de forma correta. Havendo mais pontos de referência, melhores serão os resultados.

Nova construção na próxima página:

Segmento com 14 unidades:

AO em F traça o arco A3' e A'O em F' traça o arco A'3. ⊛
Na intersecção dos arcos 3 e 3', obteremos 2 pontos B e B' (reta ortogonal)

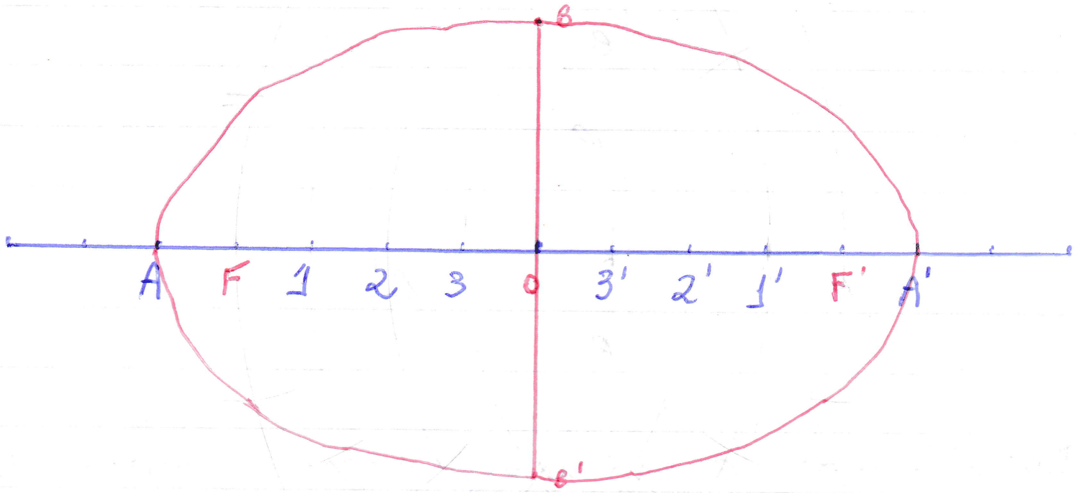

Important

A distância entre A F e A' F' é uma unidade que nós poderemos ajustá-la para adequar-se à excentricidade da referida elipse.

Segunda Forma de Construção de uma Elipse.

A construção desta elipse depende de seus dois eixos.

Important: Ajustando-se esses 2 eixos, ✡ nós teremos a excentricidade. (calcular depois)

Descrevemos dois círculos de centro O com diâmetros iguais aos eixos AA' e BB'.

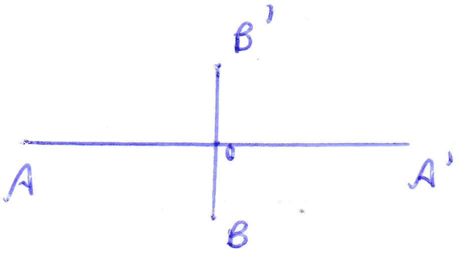

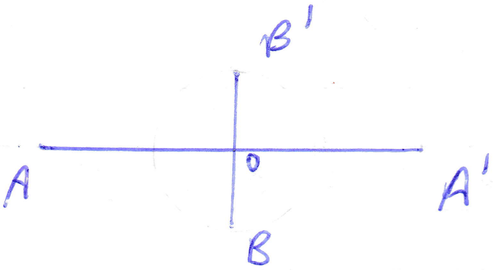

Dividimos a circunferência maior e menor em N partes iguais. Por conveniência, nós vamos dividi-las em 16 partes iguais, realizando-se as bissetrizes entre os eixos e depois repetimos o processo.

Em seguida, traçamos retas perpendiculares ao eixo AA' pelos pontos que dividem a circunferência maior. ①

Depois, traçamos retas perpendiculares ao eixo BB' pelos pontos que dividem a circunferência menor ②

Os pontos (x_n, y_n) gerados a partir das interseções destas retas determinam a elipse.

Aumentando as retas da elipse:

B1 e B2 usados para bissetrizes

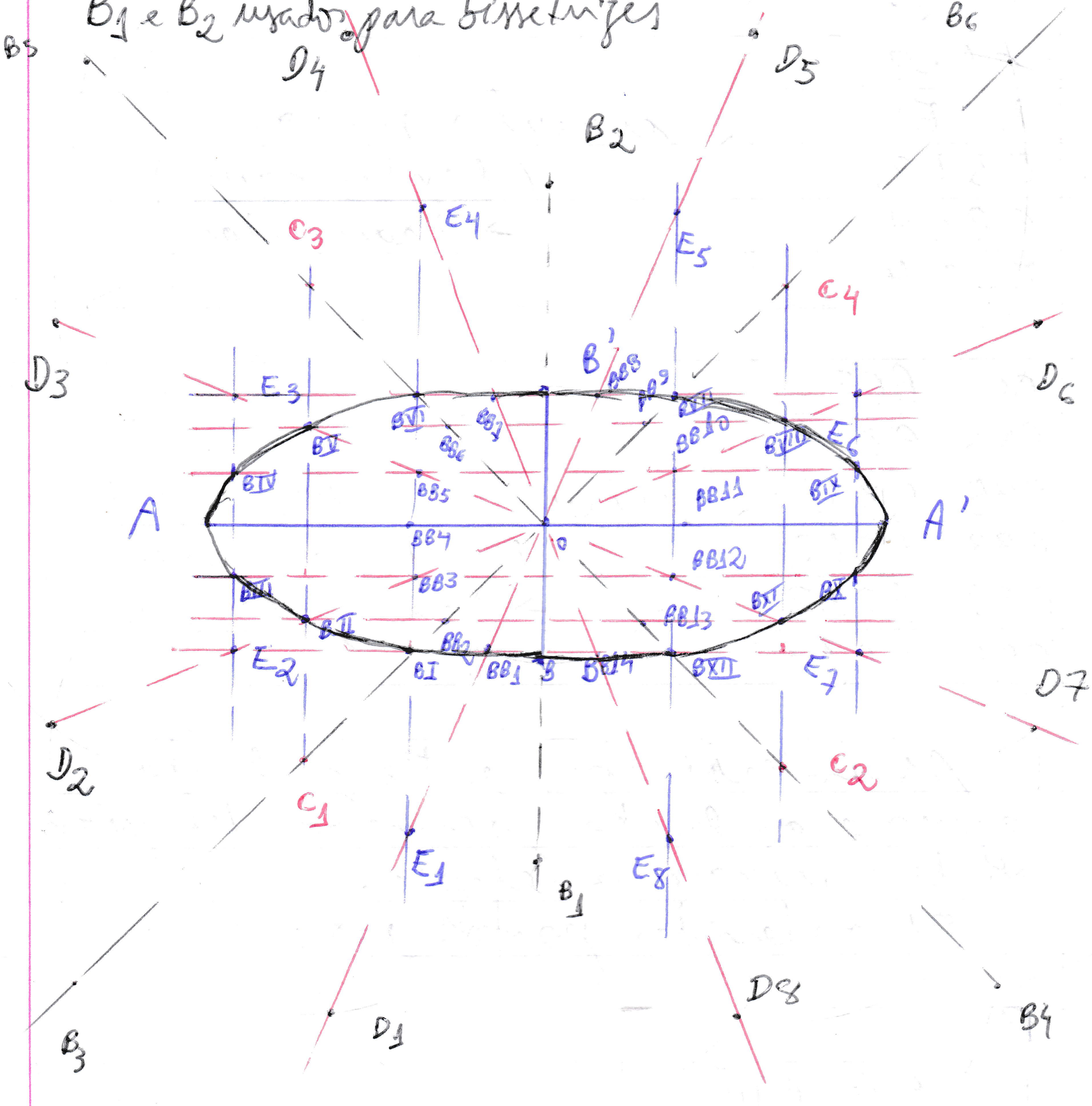
B5
D4
B6
D5
B2
E4
C3
E5
C4
D3
B'
BB8
BB9
D6
E3
BVI
BB7
BVII
BB10
BV
BB6
BVIII
E6
BIV
BB5
BB11
BIX
A
A'
BB4
0
BB12
BIII
BB3
BXI
BX
BII
BB13
BB2
E2
BI
BB1
B
BB14
BXII
E7
D7
D2
C2
C1
E1
E8
B1
D8
B3
D1
B4

B3, B4, B5 e B6 → são as primeiras bissetrizes.

C1, C2, C3 e C4 → são os pontos que interceptam o círculo maior.

D1, D2, D3, D4, D5, D6, D7 e D8 → são as novas bissetrizes.

Ligues os pontos: D1 D5, D2 D6, D3 D7, D4 D8, criando-se os seguintes pontos: E1, E2, E3, E4, E5, E6, E7 e E8.

(círculo maior) e os pontos: B, BB1, BB2, BB3, BB4, BB5, BB6, BB7, B', BB8, BB9, BB10, BB11, BB12, BB13 e BB14 (círculo menor).

Ligue os seguintes pontos:

E2 E3
C1 C3
E1 E4
E5 E8
C2 C4
E6 E4

> círculo maior
(linhas azuis)
→ tracejadas.

BB7 BB8
BB6 BB10
BB5 BB11
BB3 BB12
BB2 BB13
BB1 BB14

> círculo menor
(linhas vermelhas)
→ tracejadas.

As intersecções entre as linhas azuis e as linhas vermelhas delimitam ou determinam a élipse.

Os referidos pontos são:

B
BI, BII, BIII
A
BIV, BV, BVI
B'
BVII, BVIII, BIX
A'
BX, BXI, BXII

> pontos que determinam a elipse.

obs: Fazer estas 2 simulações do ponto de vista antrométrico.

Important:

Estas construções foram feitas utilizando-se régua e compasso de forma gráfica e com erros e imprecisões nas definições dos referidos pontos que determinam a elipse.

Agora, nós devemos fazê-lo de forma aritmética, calculando-se os pontos e as suas distâncias, as bissetrizes devem ser obtidas por meio da determinação do ponto (x, y) que dividem cada quadrante da elipse.

Tudo tem que ser demonstrado de forma inequívoca.

20231018183700

20231018210000

A título de informação científica, Ⓐ complementar ao que eu já venho estudando eu encontrei na internet https://zenodo.org/records/1447218, a publicação com o título:

"Signal Time Travel Dilation
TEM Model Time Machine
Creation Theory 10/04/2018
→ Lance E. Wenner

Nome do arquivo:
Wenner-Time-Dilation-Ref-10-04-2018.pdf

Obs: Estudar com Tony Yuan 7-2.pdf
(Gravitational Waves).

Pág. 6- VI.

Gravity Time Length Well

Light (electromagnetic wave time path bending) while passing near a large mass like a Planet is also a time dilation effect rather that light speed change, which is again against ✡ Einstein [1], or some other invisible unknown lensing medium in the vacuum of space.

A mass warps the time length of its surrounding space an amount progressively more the larger the mass within limites per Einstein GR [1].

Tensor Calculus of momentum pressure in GR general relativity of the EINSTEIN [1] dual (4X4) matrix equation includes the fourth dimension time energy.

Ref.: 1) Synge - Tensor Calculus - (1949).pdf
→ 334 pages.

2) The Geometry of Total

Curvature on Complete open surfaces-2003.pdf
→ Katsuhiro Shiohama...
→ 296 pages.

Obs: Pág. 41 — chapter 2
the classical results of
Cohn-Vossen and Huber

↳ referência:

Hilbert, D. and Cohn-Vossen, S,
"The Cylinder, the Cone,
the Conic Sections, and their
surfaces of Revolution"!

3) The Geometry of Four-Manifolds-
1997.pdf
→ Donaldson/Kronheimer
→ 458 pages.

Obs: Usar este livro para analisar
fourth dimension time energy.

return: Pág. 6

This demonstrates all curved path
space-time gravity
(all earth size gravity for
example)
is mostly time length well charge
effect (time dilation).

Increasing the mass increases
the stress-energy tensors,
hence
bending its space-time curve path
more
(other mass taking less time
length to arrive) thereby,
due to the smaller time arc
at same distance
reducing geodesic path length
of the other mass heading somewhat
toward it (not attraction but
falling more toward larger mass
in less time).

A time length well
Theory for Earth gravity
does therefore not violate
general relativity
but agrees with and its effective
limits of, very large mass

bending actual 3d space,

more than 4d time, once

becoming a very large size like
a neutron star.

IMPORTANT:

→ LESS ENERGY ✡

Less Energy → see Entropy.

Since taking less time, mass or light (EMW) ✡

use less energy

following a space-time curved space (takes more energy to turn away), ✡ therein not violating the conservation of energy Newton [4] law either.

Velocity time length special relativity (SR) change is a different time dilation mode, from general relativity (GR) gravity time length change.

(SR) Increased velocity dilates time longer,

(GR) increased mass dilates objects heading toward it time shorter, both within the limits of Minkowski [8] 4 dimensional space-time for SR and GR.

Pág. 09 – Fig. (13) →

Waveguide delay time formula vs. Einstein SR vacuum Earth length compared indicating similar time curve relations.
Confirms time delay per frequency in a

angular waveguide

Fig.(14) Demonstrates the surfaces thickness (to reflect off) and their waveguide effect to delay wavelengths at different time intervals which in fact splits the light into different frequencies (wavelengths).

(*) The surface material determines the waveguides depth and angle of reflection required to obtain a full color light spectrum. (ver Newton → sobre o seu trabalho sobre as cores).

IMPORTANT: In an equilateral triangle the shape distance keeps all the different output light frequencies arriving at the same time called in phase. In other shapes, glass time delays light, rather than (em vez de) falsely bending or changing light speed etc again as previously discussed.

Obs: Leva a referência desta publicação

20231018231200

20231022133000.

Arquivo: ellipse-20231022-01.ggb

Roteiro de ensaio

1) Crie os seguintes pontos:

a) F1 = Point(x Axis)
→ (-4, 0)

b) F2 = Point (x Axis)
→ (4, 0)

c) A1 = Point(x Axis)
→ (-5, 0)

d) A2 = Point(x Axis)
→ (5, 0)

e) A = Point(y Axis)
→ (0, 3)

f) O = Intersect(x Axis, y Axis)
→ (0, 0)

obs: Important:

Esta configuração de pontos é muito interessante pois facilita os cálculos.

Vejamos: Os pontos focus F1 e F2 estão na posição (-4,0) e (4,0) e portanto, estão na posição de número par.

Usando-se o sistema cartesiano, mantendo-se uma unidade como padrão de distanciamento entre os pontos e para que os pontos A1 e A2 estejam mantendo a distância de uma unidade dos pontos focus F1 e F2, nós precisaremos escolher o ponto A (y Axis) → (0, 3). Com este posicionamento nós criaremos uma elipse. O interessante, neste caso, é que estamos realizando uma adequação entre os números pares (eixo x) e números ímpares (eixo y), observando-se o distanciamento de uma unidade como padrão.

Posteriormente, nós criaremos o ponto B (yAxis) → (0,-3)

2) Agora criaremos os círculos menor e maior com base no ponto O como origem e o diâmetro do tamanho das retas A1A2 e AB. Vejamos:

a) Círculo menor - AB

c : Circle(O, A)

→ $x^2+y^2=9$

obs: Important: Na aba de configuração do Geogebra chamada Style escolha: (se quiser)

1) Line Thickness: mínimo. 3
2) Line Opacity: mínimo. 20

b) Círculo maior

d : Circle(O, A2)

→ $x^2+y^2=25$

obs: 1) Line Thickness: 3
2) Line Opacity: 20

3) Em seguida, nós criaremos os pontos:

a) C = Intersect(d, yAxis, 2)
→ (0,5)

b) D = Intersect(d, yAxis, 1)
→ (0,-5)

Vou ajustar para pontos em vez de intersecção pois facilita a execução

Portanto: Voltamos

Criação B = Point(yAxis)
→ (0, -3)

Em seguida, criaremos:

a) C = Point(yAxis)
→ (0 5)

b) D = Point(yAxis)
→ (0, -5)

Agora, nós criaremos os círculos:

a) Círculo menor:
c: Circle(O, A)
→ $x^2+y^2=9$

b) d: Circle(O C)
→ $x^2+y^2=25$

Retornando um pouco

Criaremos os pontos:

a) E = Point(xAxis)
→ (-3, 0)

b) F = Point(xAxis)
→ (3, 0)

Agora (sim):

a) c: Circle(O A)
→ $x^2+y^2=9$

b) d = Circle(O C)
→ $x^2+y^2=25$

3) **Nova Fase**:

Angle Bisector → important:

Para nós criarmos Angle Bisector, nós poderemos fazer de duas formas:

a) Criaremos os pontos:

a1) G = Point(y Axis)
$\rightarrow (0, 1)$

a2) H = Point(x Axis)
$\rightarrow (1, 0)$

a3) I = (1, 1) ⇒ Point

↳ Aqui está o ponto que facilitará a construção.

Vejamos:

Criaremos

f : Line(0, I)
$\Rightarrow -x + y = 0$

obs: Esta é a função do 1° Quadrante.

Para agilizar:

Criaremos o ponto

$J = (-1, 1)$

Em seguida:

g = Line(0, J)
$\rightarrow -x - y = 0$

obs: Esta é a função do 2° Quadrante.

Interessante:

O que foi feito facilita a construção da elipse.

Agora, vamos construir de outra forma, utilizando-se Angle Bisector, da seguinte maneira:

a) 1º Quadrante:

Voltando um pouco, nós criamos os pontos

a1) G = Point(y Axis)
$\rightarrow (0, 1)$

a2) H = Point(x Axis)
$\rightarrow (1, 0)$

Após: Angle Bisector

b1) f: Angle Bisector(H, O G)
$\rightarrow -0.707106781186547x + 0.707106781186547y = 0$

obs: Esta função é igual
$-x + y = 0$

b) 2º Quadrante:

Criaremos o ponto I

I = Point(x Axis)
$\rightarrow (-1, 0)$

Em seguida: Caso mantenha o f(b1)

g: Angle Bisector(I, O G)
$\rightarrow 0.70710678\ldots x + 0.70710678\ldots y = 0$

Obs: Esta função é igual à

$-x - y = 0.$

Important:

Nós manteremos a versão mais simplificada, pois facilitará a didática para os futuros estudantes desta pesquisa e de matemática e física.

Em resumo:

Após a criação do círculo maior (O, C).

d: Circle(O, C)

$\rightarrow x^2 + y^2 = 25$

Nós criaremos, dois pontos:

a) O ponto $G = (1, 1)$
→ que dividirá o 1º Quadrante.

b) O ponto $H = (-1, 1)$
→ que dividirá o 2º Quadrante.

Em seguida, nós criaremos as linhas:

a) f: Line(O G)
$\Rightarrow -x + y = 0$

b) g: Line(O H)
$\Rightarrow -x - y = 0$

Obs: Desta forma os quatros quadrantes estão divididos ao meio.

Importantíssimo:

Antes de criarmos as linhas f e g, nós devemos criar 4 pontos para ajudar na 2ª operação de Angle Bisector e são os seguintes:

a) $I = (3, 3)$

b) $J = (-3, 3)$

c) $K = (-3, -3)$

d) $L = (3, -3)$

Estudo Aprofundado:

Um aspecto interessante, com o ponto $I = (3, 3)$ nós dividimos o 1º Quadrante ao meio.
Para nós fazermos a próxima

Angle Bisector, primeiro vamos fazer um estudo mais detalhado:

Ponto F = Point(x Axis)
$\rightarrow (3, 0)$

Deste ponto, nós iremos (criara) criar um novo círculo

e: Circle(F, I)
$\rightarrow (x-3)^2 + y^2 = 9$

✓

Obs: antes disso, nós esqueço de criar o ponto

M = Point(x Axis)
$\rightarrow (6, 0)$

Depois de criar este ponto, nós criaremos o círculo e.

Em seguida, nós criamos o segmento h

$$\boxed{\begin{array}{l} h = \text{Segment}(I, M) \\ \rightarrow 4.242640687119286 \end{array}}$$

Obs: Este segmento é $3 \times \sqrt{2}$!!! (interessante!!!)

$\boxed{\text{IMPORTANTÍSSIMO.}}$

O triângulo retângulo formado pelos vértices OIM Δ OIM, com ângulo reto em $\boxed{I}$, cujos catetos OI e IM são iguais, apresentando o mesmo comprimento

$$\boxed{3\sqrt{2}}$$

e cuja hipotenusa $\boxed{\text{OM é 6 unidades.}}$

Atenção:

g: Angle Bisector(F, O, I)
$\rightarrow$ 0.3826... x + 0.9238... y = 0

Esta reta intersecta o cateto IM do triângulo retângulo Δ OIM, no ponto

$$\boxed{N = \text{Intersect}(g, h) \rightarrow (\underbrace{4.242640...}_{(3\sqrt{2})\ \hookrightarrow x}, \underbrace{-1.757359...}_{\hookrightarrow y})}$$

Este novo triângulo Δ ONM,

que é um triângulo obtuso
e também é um triângulo escaleno
Ângulos:

$$\begin{cases} O = 22{,}5° \\ M = 45° \\ N = 112{,}5° \end{cases}$$

Comprimentos: Segmentos

O triângulo t_1

t_1 = Polygon(M, O, N)
$\longrightarrow$ 5.272077...

Segmentos:

m = Segment(O, N, t_1)
$\longrightarrow$ 4.5922011...

n = Segment($\underset{(M)}{M}$, O, t_1) ($\underline{M}$, O, t_1)
$\longrightarrow$ 6

o = Segment(N, M, t_1)
$\longrightarrow$ 2.485281...

Portanto: Triângulo escaleno
com ângulo obtuso.

Interessante:

6 Segmento

h = Segment(I, M) → comprimento

→ 4.242640...

$(= 3\sqrt{2})$

e o ponto

N = Intersect(g, h)

$\rightarrow (\underbrace{4.242640...}_{(x)}, \underbrace{1.757359...}_{(y)})$

este valor é igual ao segment (I, M)

$(3\sqrt{2})$

obs: Os comprimentos são diferentes.

ON > IM.

IMPORTANTÍSSIMO:

Partindo-se do Ponto N (?), cujo valor não é exato.

N = Intersect(g, h)

Nós iremos traçar uma Perpendicular Line no ponto A2,

$$\boxed{\begin{array}{l} i: \text{PerpendicularLine}(A2, \text{xAxis}) \\ \rightarrow x = 5 \end{array}}$$

obs: Este passo desvendará nos caminhos.

Esta perpendicular irá interceptar

f: Line(O, G)
$\rightarrow -x + y = 0$

no ponto ~~A~~ P

P ~~Q~~ = Intersect(f, i)
$\rightarrow (5, 5)$

Obs: Criar este ponto quando da criação dos primeiros pontos.

Exemplos de pontos complementares
(a serem construídos)

P = (5, 5) → 1º Quadrante

Q = (−5, 5) → 2º Quadrante

Obs: Estes dois pontos já dividem os 4 quadrantes, mas devemos marcar os outros 2 pontos para facilitar a construção.

R = (−5, −5) → 3º Quadrante.

S = (5, −5) → 4º Quadrante.

Continuando:

Após construímos a perpendicular sobre o ponto A2 e marcando ponto P=(5,5), este ponto intercepta tanto a linha f como a linha i.

obs:

f: Line (0, G)
→ $-x + y = 0$

i: Perpendicular Line (A2, x Axis)
→ $x = 5$

Portanto: O Ponto P=(5,5) pode ser obtido traçando-se este ponto no 1º Quadrante ou fazendo-se a intersecção da linha f com a linha i (perpendicular ao ponto A2 no eixo X).

Posteriormente:

Marcamos o ponto:

T = Intersect (g, i)
→ (5, 2.0710678....)
(x) (y)

e o ponto

U = Intersect (i, h)
→ (5, 1)

Importantíssimo:

U = Intersect(i, h)
→ (5, 1)

obs: Este ponto deve ser marcado quando da criação dos primeiros pontos da construção.

Lembrando:

h = Segment(I, M)
→ 4.242640...
($3\sqrt{2}$) → comprimento

obs:

I = (3, 3)
M = (6, 0)

A interceptação da linha i com a linha h e também com a linha g, criará um triângulo isósceles

(isosceles triangle)
invertido

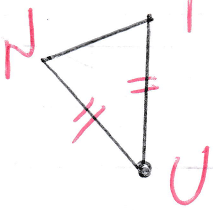

Criaremos o triângulo Δ NTU

t_1 = Polygon(N, T, U)
→ 0.4055915...

n = Segment(T, U, t_1)
→ 1.0710678...

t = Segment(U, N, t_1)
→ 1.0710678...

u = Segment(N, T, t_1)
→ 0.8197598...

} n = t

Obs: o Ponto U = (5, 1), o ângulo interno de U é igual 45°

Important:
O triângulo Δ MA2U é um triângulo retângulo e a sua hipotenusa é $\boxed{\overline{UM}=\sqrt{2}}$

Obs:
A2 M = 1
A2 U = 1
U M = $\sqrt{2}$

Obs: Em um círculo com centro em U = (5, 1), os pontos N e T estarão sobre o círculo.

ff:
Circle(U, N) → $(x-5)^2+(y-1)^2=1.147186...$

obs: Se fizermos um círculo com centro $V = (5,1)$ e raio $A2 = (5,0)$

$\rightarrow (x-5)^2 + (y-1)^2 = 1$

obs: Neste caso, os pontos N e T estarão fora do círculo.

Important:

Pontos:

$I = (3,3)$

$V = (5,1)$

A1

A distância entre estes pontos é

$I = \text{Segment}(V, I)$

$\rightarrow 2.82842712\ldots$

$= 2\sqrt{2}$

Nós devemos criar os seguintes pontos

$V = (1,5)$

$W = \text{Point}(y\,\text{Axis})$

$\rightarrow (0,6)$

Interessante: Vejamos →

(6 Segmento (I, W)

$J = \text{Segment}(I, W)$

$\rightarrow 4.24264068\ldots$

$(= 3\sqrt{2})$

Preste atenção:

$K = \text{Segment}(N, I)$

$\rightarrow 1.757359\ldots$

Mas se nós somarmos

$\overline{UN} + \overline{NI} = \overline{UI}$

Obs: $I = \text{Segment}(U, I)$

$\rightarrow 2.8284271\ldots$

$(= 2\sqrt{2})$

Portanto:

$UN + NI = 2\sqrt{2}\ (UI)$

$IV = UI\ (2\sqrt{2})$

Obs: $\overline{UV} = 4\sqrt{2}$

Vamos prestar atenção.

Isto vai facilitar a minha pesquisa.

$\Rightarrow$

Fazendo AngleBisector do ângulo

A2 O T, nós obtemos

f: AngleBisector(A2, O, T)
↳ 0.195090... x +
0.980785... y = 0

Obs: O mais interessante que o ponto exato que divide este ângulo e o ponto U

$$\boxed{U = (5, 1)}$$

WRONG!!!
Ajuda a dividir o ângulo

Important:

Não divide.

Este ponto (divide) o 1º Quadrante em 8 partes iguais e por Consequente também divide o 3º Quadrante e o ponto é

Z = (−5, −1) → criar este ponto

Agora vamos ligar os ponto Z U

FILE:
elipse-20231022-09.ggb.

f = Segment(Z, U)
↳ 10.198039027...

20231022202233

20231023083000.

Esta fase de criação deste ponto será posteriormente!

$A_1 = (6, 4)$

r = Line $(0, A_1)$

$\rightarrow -2x + 3y = 0$

Retornando:

Nós vamos criar os pontos dos outros quadrantes (2°, 3° e 4°).

a) $A_1 = (-1, 5)$

b) $B_1 = (-5, 1)$

c) $C_1 = (5, 1)$

> Estes pontos são referências para dividir os demais quadrantes (1/8 de um quadrante). (Ajudam)

Até este ponto (correto)

Agora, nós iremos criar os outros pontos de referência:

(CHECAR)

a) $D_1 = (6, 4)$

b) $E_1 = (-6, 4)$

c) $F_1 = (-6, -4)$

d) $G_1 = (6, -4)$

> Outras divisões nos quadrantes (2°, 3° e 4°) → 1/8 de um quadrante. (Ajudam) → checar!!!

Vamos continuar as divisões dos quadrantes:

Fazendo os "espelhamentos":

A) D_1 em relação P, obteremos o ponto $H_1 = (4, 6)$ ✓

Vejamos:

a) Ponto $P = (5, 5)$

b) Ponto $D_1 = (6, 4)$

c) Ponto $H_1 = (4, 6)$

Interessante:

A distância entre os pontos $\overline{D_1 P}$ e $\overline{PH_1}$ são iguais à $\sqrt{2}$.

Agora, nós viemos criar os outros pontos:

B) Ponto $J_1 = (-4, 6)$

obteremos:

a) $E_1 = (-6, 4)$

b) $Q = (-5, 5)$

c) $J_1 = (-4, 6)$

C) Ponto $J_1 = (-4, -6)$, nós obteremos:

a) $F_1 = (-6, -4)$

b) $R = (-5, -5)$

c) $J_1 = (-4, -6)$

D) $K_1 = (4, -6)$, nós obteremos:

a) $G_1 = (6, -4)$

b) $S = (5, -5)$

c) $K_1 = (4, -6)$.

Interessante:
Cada bloco destes 3 pontos, nós obteremos 2/8 de um quadrante.

Portanto:

a) $H_1 P D_1$ = 2/8 do 1° Quadrante;

b) $I_1 Q E_1$ = 2/8 do 2° Quadrante;

c) $F_1 R J_1$ = 2/8 do 3° Quadrante e

d) $G_1 S K_1$ = 2/8 do 4° Quadrante.

---

* Checar o ponto $U = (5, 1)$ → não divide o quadrante em 1/8.
(11,30° → correto 11,25°)

Vamos checar os outros pontos:

Por exemplo: $D_1 P H_1$

→

Os pontos D_1 e H_1 estão errados.

Vamos criar outros mecanismos de checagem.

RETORNO :

Usaremos somente pontos precisos e não fazer Angle Bisector
↳ usa números imprecisos.

↓
correto.

Veremos :

Arquivo mais atualizado

ellipse-20231023-03.ggb

Novos Cálculos :

ellipse-20231023-04.ggb ⇐

Paramos no ponto U

U = Intersect(i, h)
→ (5, 1).

Para confirmar a intersecção, criamos o ponto

Z = Intersect(h, i)
→ (5, 1).

Para ajudar nos cálculos:

Criaremos os pontos:

a) $A_1 = (-1, 5)$
b) $B_1 = (-5, 1)$
c) $C_1 = (-5, -1)$
d) $D_1 = (-1, -5)$
e) $E_1 = (1, -5)$
f) $F_1 = (5, -1)$

> Estes pontos serão úteis para os cálculos de Angle Bisector.

Important:

Nós usaremos apenas Angle Bisector para dividi ao meio os 1º, 2º, 3º e 4º Quadrantes.

Vejamos:

$L = (3, -3)$
$I = (3, 3)$

> criar uma linha ligando estes pontos.

$g = \text{Segment}(L, I)$
$\Rightarrow 6$

Criar os seguintes pontos:

a) H_1 = Point(d)
→ (4, -3)

b) I_1 = Point(d)
→ (4, 3).

ligar estes pontos.

r = Segment(H_1, I_1)
→ 6

Vamos marcar os pontos sobre a elipse.

J_1 = Intersect(p, q, 2)
→ (3, 2.4)

K_1 = Intersect(p, r, 2)
→ (4, 1.8)

continuando:

L_1 = Intersect(p, r, 1)
→ (4, -1.8)

M_1 = Intersect(p, q, 1)
→ (3, -2.4)

Agora, nós vamos fazer o mesmo procedimento do outro lado (2º e 3º Quadrantes)

Os pontos:

J = (-3, 3)]
K = (-3, -3)]> ligar estes pontos

s = Segment(K, J)
→ 6

N_1 = Point(d)
→ (-4, 3)

O_1 = Point(d)
→ (-4, -3)
]> ligar estes pontos

a = Segment(O_1, N_1)
→ 6

P_1 = Intersect(p, s, 2)
→ (-3, 2.4)

Q_1 = Intersect(p, a, 2)
→ (-4, 1.8)

R_1 = Intersect(p, a, 1)
→ (-4, -1.8)

S_1 = Intersect(p, s, 1)
→ (-3, -2.4)

File de referência:

ellipse-20231023-05.ggb.

Criar o Frégier Point:

Crie o ponto T_1 sobre a elipse (ponto aleatório)

Crie outro ponto aleatório no outro lado da elipse: U_1

Após crie o segmento que ligue o ponto U_1 T_1.

Em seguida, use o comando Perpendicular Line sobre o ponto T_1. Esta linha interceptará a elipse no outro lado, use o comando Intersect que unirá esta linha perpendicular à elipse.

Em seguida, crie uma linha que ligue (Segment) V_1 U_1 e em seguida marque o Polygon (T_1, U_1, V_1).

Neste momento, nós criamos o primeiro triângulo retângulo.

Para criar o segundo triângulo retângulo, nós devemos criar outro ponto:

W_1 sobre a elipse

Depois, com o comando Segment, ligar os pontos T_1 W_1.

Use o comando Perpendicular Line sobre esta linha no ponto T_1.

Esta linha intercepta a elipse no ponto Z_1, crie uma linha ligando os pontos W_1 e Z_1, marque em

Use seguida o Polygon:
Polygon(T_1, W_1, Z_1).

Em seguida use o comando

Intersect em relação aos dois triângulos retângulos, criando-se assim, o ponto

Frégier Point (point D_2)

File: ellipse-20231023-padrão-Frégier-Point.ggb.

Criação de 2 círculos que se interceptam sobre a elipse em 2 pontos:

Primeiro:

Crie a linha D_2T_1, sendo D_2 o ponto Frégier Point.

I_1 : Line(D_2, T_1)
$\rightarrow -4.1342\ldots x + 0.924011\ldots y = -4.61\ldots$

Após crie o ponto que intercepta esta linha com a elipse no ponto
($H2$ = Intersect(p, I_1, 2)
$\rightarrow (0.44923\ldots, -2.987866)$.

Agora, partiremos para CIRCLE

Use o comando:

Circle through 3 Points

Ligando os seguintes pontos:

a) c_1: Circle(T_1, H_2, Z_1)

$\rightarrow (x - 1.0888275...)^2$
$(y + 0.0863965...)^2 =$
$8.827609...$

b) d_1: Circle(T_1, W_1, Z_1)

$\rightarrow (x + 0.7275995...)^2 +$
$(y + 0.4830656...)^2 =$
$16.96827 84...$

Important:

Os 2 círculos se cruzam em 2 pontos (T_1 e Z_1) sobre a ellipse.

Obs: Checar se as linhas correspondem aos diâmetros dos referidos círculos.

File:

ellipse-20231023-10.ggb.

Off: Calcular os centros destes círculos com geometria e fazer simulação somente sobre estes círculos e linhas, analisando os seus alinhamentos.

20231023183200

20231024083000

Estudo sobre dois círculos que se interceptam em dois pontos e o ponto Frégier Point:

Agora: 20231024163100:

Simulei dos 0830 às 1631 e obtive o arquivo:

ellipse-20231024-06-dois-circulos-MAIS-PROXIMO---3X3-PONTOS.ggb

obs: Descrevê-lo posteriormente.

Ponto: J

J = Interseção(f, Eixo x)
→ (1.76, 0)

Este ponto é interessante (estudar)

obs: O interessante desta simulação é que o ponto H e K há a interseção de 3 figuras geométricas.

— 2 círculos

— 1 elipse

Do outro lado, há também a interseção do círculo maior com a elipse.

Veja outro arquivo:

ellipse-20231024-06-dois-circulos-MAIS-PROXIMO---3X3-Pontos-Ponto1-76.ggb

Continuar as simulações dos dois círculos.

20231024175600

20231025081800

Eu criei uma forma de unir dois círculos que se interceptam em dois pontos e os quatro pontos criados que correspondem ao dois diâmetros destes dois círculos estão sobre a elipse.

Passos para criar esta simulação

obtida na data de
20231025120000
estes passos estão armazenados no arquivo:

Ellipse-20231025-01-2-
circulos-ellipse-07.ggb

Source: Data-Novo Vivo-010-2023/
Geogebra-2023.

Obs: Estes últimos quatro dias foram de intensa pesquisa com o uso de computador.

Passos de criação:

1) A = Intersect(x Axis, y Axis)
$\rightarrow$ (0, 0)

2) I = Point(x Axis)
$\rightarrow$ (-8, 0)

3) J = Point(x Axis)
$\rightarrow$ (8, 0

4) B = Point(y Axis)
$\rightarrow$ (0, 6)

5) D = Point(x Axis)
$\rightarrow$ (0, -6)

6) c: Ellipse(I, J, B)
$\rightarrow 9x^2 + 25y^2 = 900$

7) C = Point(c)
$\rightarrow$ (3.94..., 5.51...)

Obs: Este ponto é aleatório e será utilizado para a criação dos dois triângulos retângulos para a obtenção do ponto Fregier Point

8) f: Perpendicular Line
(I, x Axis)
$\rightarrow x = -8$

Obs: no ponto I (foci I), line through I perpendicular x Axis

9) Intersect
(c, f)
$\rightarrow$ E = (-8, -3.6)
$\rightarrow$ F = (-8, 3.6)

Obs: Estes dois pontos intersectam a perpendicular sobre a elipse

10) g = Segment
(E, C)
$\rightarrow$ 15.026...

11) h: Perpendicular
(C, g)
$\rightarrow$ 11.94... x - 9.11... y = -97.404...

12) G = Intersect
(c, h, 2)
→ (9.533..., −1.8109..)
obs: Perpendicular
sobre o ponto C e
intercepta a elipse c.

13) i = Segment
(E, G)
→ 17.62464...

14) j = Segment(F, C)
→ 12.100070...

15) k: Perpendicular
(C, j)
→ −11.94...x − 1.912...y =
−57.713...

16) H = Intersect
(c, k, 2)
→ (5.624..., −4.96...)

17) I = Segment
(F, H)
→ 16.090947...

18) K = Intersect
(i, 1)
→ (1.857...,
−2.594...)

obs: Este ponto K
é o Fregier Point.

19) m: Line(C, K)
→ 8.10...x − 2.09009...y =
20.48333999...

20) L = Intersect(g, 1)
→ 2.82401..., 0.3476...)
obs: Dispensável mas
tem um valor referencial.

21) M = Intersect
(c, m, 2)
→ (0.98..., −5.97...)

22) Com o comando
CIRCLE THROUGH 3 POINTS
criaremos
d: Circle(C, H.M)

$$\to (x-2.1477118\ldots)^2 + (y+0.14655060\ldots)^2 = 35.26702036\ldots$$

Important:
O Ponto C é um
ponto aleatório.
O Ponto H é a
perpendicular sobre
o ponto C do segundo triângulo.

Este ponto H também intercepta a ellipse c.

INCRÍVEL:

Este ponto H fazendo-se um segmento até o ponto K (Frégier Point)

Usando-se o segmento CM que é o diâmetro do primeiro círculo e usando-se a função

Perpendicular Bisector,

nós descobrimos o centro deste círculo no ponto que esta perpendicular intercepta o segmento CM.

Obs: Para checar se é mesmo o centro, eu utilizei o sistema das duas cordas sobre o círculo, eu fiz Perpendicular Bisector destas duas cordas e elas se encontraram no mesmo ponto.

Outro jeito é usando um compasso para marcar a perpendicular sobre este segmento.

Portanto:

23) n: Perpendicular Bisector(M, C)

→ $-2.9606\ldots x - 11.4832\ldots y = -4.67\ldots$

24) N = Intersect(m, n)

→ $(2.4676\ldots, -0.22903\ldots)$

25) e : Circle(N, C)

$\rightarrow (x - 2.4676...)^2 + (y + 0.2290...)^2 = 35.1578...$

Importantíssimo:

Com este círculo nós podemos eliminar o círculo anterior (C, H, M) → criado por meio dos três pontos, ou seja, este novo círculo (N, C) encontra ou intercepta a elipse em dois pontos $\boxed{C \text{ e } M}$

Portanto, há 4 pontos na mesma linha (colineares)

→ $\boxed{\text{C, N, K (Frégier Point) e M}}$

Obs:

O Ponto H quando ligado ao Ponto K (Frégier Point), encontra o ponto F ("foci") que a perpendicular sobre o foci I (−8, 0).

26) p : Circle(F, H, C)

$\rightarrow (x + 1.187723...)^2 + (y + 0.680483...)^2 = 64.72964422...$

27) g: Perpendicular Bisector(F, H)
→ g: -13.62455...x + 8.5609663...y = 10.35661077...

28) O = Intersect(g, l)
→ (-1.187723..., 0.68048...)

Source:

Ellipse-20231025-01-2-circulos-e-ellipe-07.ggb

Important:

O segmento:

a) CM = diâmetro do primeiro círculo (menor)

b) FH = diâmetro do segundo círculo (maior)

Important:

4 pontos em cada linha (colineares)

a) C N K M,

b) F O K H.

obs!!! Estou feliz!!! Eu tive a benção de DEUS.

Voltando aos estudos de 20231017

Os pontos C (aleatório) e K (Frégier Point) são padrões desta simulação.

Agora, os pontos

a) H → é o ponto Me1.

b) M → é o ponto Me2.

c) F → é o ponto Me3.

Obs: Me = Munich Ellipse

Importante:

Me1, Me2 e Me3 → estão intimamente relacionados com os diâmetros dos círculos na elipse.

Desta forma, cada diâmetro apresenta 4 pontos:

a) Círculo Menor: CNKM.

b) Círculo Maior: FOKH.

Obs: Esta é a primeira fase da

simulação das órbitas elípticas dos planetas sobre a curvatura do espaço.

OBS: Calcular a curvatura do espaço no nosso sistema solar.

20231025175000

20231026104000

Referências:

1) Exercices-de-géométrie-analytique-02.pdf
Pag. 01 - Chapitre I
Construction des Courbes

Pag. 08 - Fig. 4.
Estudar a construção e ajustá-la à curvatura em estudo.

2) Exercices-de-géométrie-analytique-03.pdf
Pag. 01 (17) - Chapitre I
La Droite et le plan.

3) Leçons Geometrie Elementaire - Hadamard.pdf
709 pages

outros arquivos relacionados:

a) 0705.0787.pdf

→ On the gravitational coupling of Hadamard states.
10 pages ⇒ archive.org/details/arxiv-0705.0787

b) 1602.03704.pdf
Schrödinger-Maxwell system on Hadamard manifolds.
21 pages ⇒ archive.org/details/arxiv-1602.03704

° c) hep-th0103072.pdf
Hadamard States and Two-dimensional gravity.
7 pages → archive.org/details/arxiv-hep-th0103072

1. Introduction:
The subject of quantum field theory in curved spacetime studies the interrelation between quantum theory and spacetime geometry in the approximation that the geometry is regarded as a classical background.

Obs: See: Poincaré symmetries to obtain a physical vacuum (the state of lowest energy).

4) Glaeser - The Universe of Conics - From the ancient Greeks to 21st century.pdf
→ 496 pages.

Pág. 81(89) - Corollary 3.2.1

The line through V and perpendicular to $[V_1, V_2]$ meets the principal and auxiliary axes in the centers V_1^* and V_2^* of curvature at the vertices V_1 and V_2.

Fig. 3.11 (interessante)

Obs: Utilizar para ajudar nos cálculos e como referências os outros desdobramentos.

Pág. 83(91) - Figure 3.12

Figure Right:

For the particular point P, the osculating circle o_P meets the ellipse c at the antipode P of P.

Obs: $\pi/4$.

Interessante:

Pag. 3(11) ⇒ Figure 1.2.

The Delian problem:
The volumes of the two cubes on the left have ratio 2:1.
How can one construct the scaling factor of the edge length?

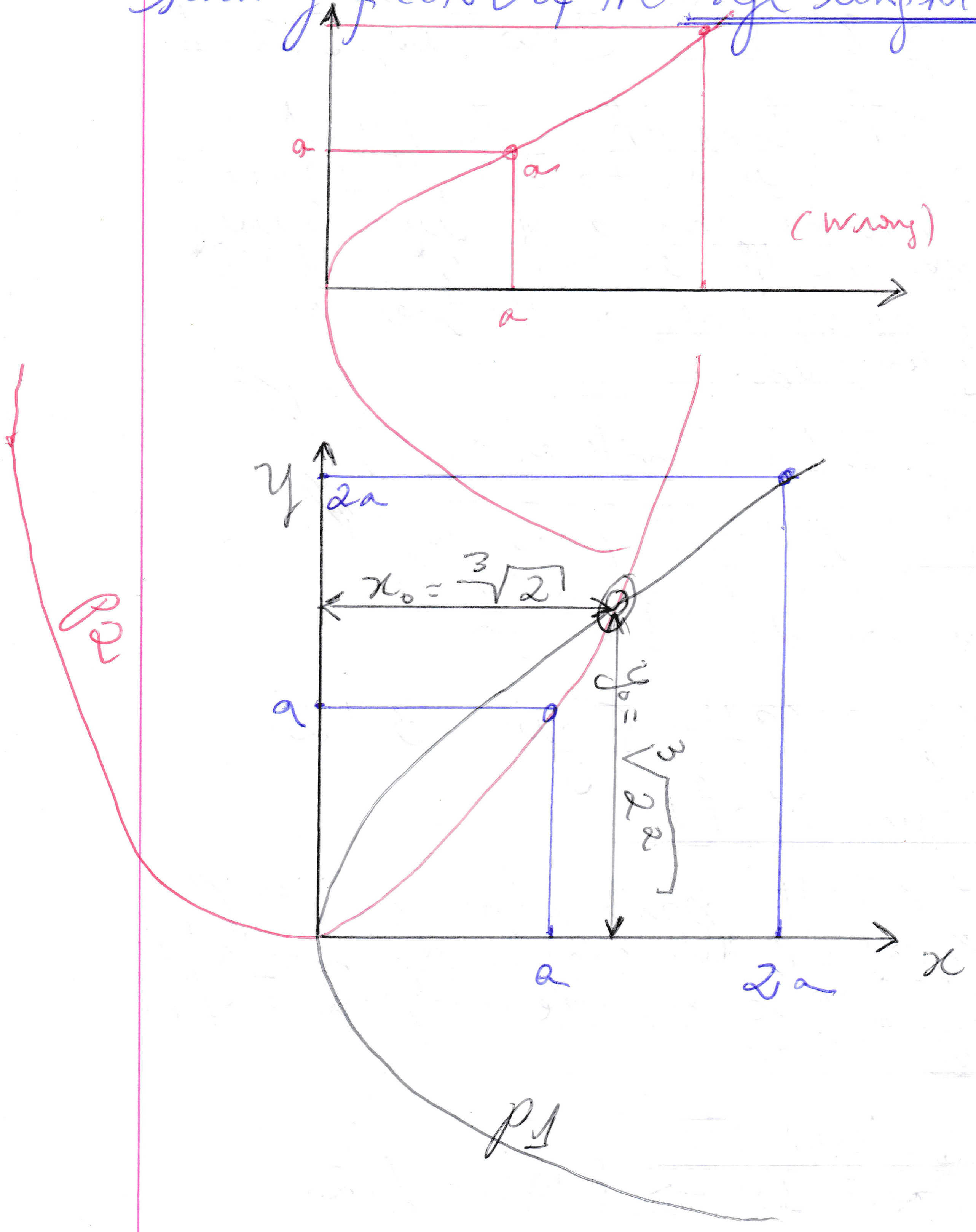

a
a
a
(wrong)
y
2a
$x_0 = \sqrt[3]{2}$
$y_0 = \sqrt[3]{2^2}$
a
a
2a
x
P2
P1

Pág. 10 (18) – Figure 1.12.

Pascal's theorem:
The three pairs of opposite sides of a hexagon inscribed in a conic section meet in the three collinear points.

→ Interessante!!!

obs: Combinar com Pappus's theorem.

5) Glaeser – Geometry and its Applications in Arts, Nature and Technology – (2020).pdf

Pág. 18 (28) – Fig. 1.17

Nine-point circle through the foot of each altitude, the side bisectors, and the vertices of the medial triangle.

6) A Magia dos Números (Livro)
→ Livro físico.
Paul Karlson – 608 pages.

Pág. 129 – Digressão sobre impossibilidades.

Fig. 65 → Trissecção aproximada do ângulo, com uma precisão máxima de meio minuto.

Pág. 131 → A quadratura do Círculo

Pág. 135 → O número mais famoso do Mundo - π (Pi).

7) Greenberg - Euclidean and Non-Euclidean Geometries - Development and History - (2008).pdf

Pág. 115 (143) - WARNING

Do not assume that every point in the interior of an angle lies on a segment joining one one side of the angle to a point on the other side. In fact, this assumption is false in hyperbolic geometry (see Exercise 19).

obs: Pág. 143 (171) - Exercise 19.

Pág. 187 (215) - Exercise 19.

Pág. 453 (481) - Lema 9.8 → Exercise 19.

- Pág. 77 (105) → Bertrand Russell's famous paradox (see Exercise 19) showed that Gottlob Frege's system of logic and classes was inconsistent.

Pág. 186 (214)

Corollary: In a Hilbert Plane satisfying Aristotle's axiom, an exterior angle of a triangle is greather than or congruent to the sum of the two remote interior angles.

Saccheri-Legendre Theorem:

In an Archimedean Hilbert plane, the angle sum of every triangle is $\leq 180°$.

Important,

No livro: Greenberg - Euclidean and Non-Euclidean Geometry ....(1993).pdf

obs: Sempre comparar a edição de 1993 com a de 2008.

Pág. 128 (146) -

Equivalence of Parallel Postulates.

We shall now prove the equivalence of Euclid's fifth postulate and Hilbert's parallel postulate.

EUCLID POSTULATE V:

If two lines are intersected by a transversal in such a way that the sum of the degree measures of the two interior angles on one side of the transversal is less than 180°, then the two lines meet on that side of the transversal.

Theorem 4.5. Euclid's fifth postulate ⟺ Hilbert's parallel postulate.

8) Radzevich - Geometry of surfaces - a practical guide for mechanical engineers - (2020).pdf

Pag. 14 (34) - 1.5.2

Second fundamental form of a desired part surface

Figure 1.7 - Definition of second fundamental form $\Phi_{2.P}$ of a smooth regular part surface P.

Pag. 16 (36) - Figure 1.8 →

Screw involute surface of a helical gear tooth (adapted from [33]).

Obs: Involute curve.

Atenção: 2 livros, um com 247 pages e outro com 313 pages!

No livro com 313 pages → a figura 1.10

No livro com 247 pages → fig. 1.8.

9) 5000 Years of Geometry – Mathematics in History and Culture-2015.pdf.

Pag. 49(62) – Angle Trisection

Pag. 52(65) – Squaring the circle.

Pag. 59(72) – Illus. 2.3.4.
Regarding the 5th postulate of Euclid's Elements.

Pag. 63(76) – the square above the hypotenuse is equivalent by decomposition to the union of both squares above the catheti.

See: Illus. 2.3.7 →

Regarding the independence of the product side times height in a triangle.

10) Analytic Projective Geometry.pdf
→ Bamberg - 475 pages.

Pag. 368(381) → Chapter 16
Line Geometry.

11) A First Course in General Relativity.pdf

Pag. 05(23) - 1.4
Spacetime diagrams

Pag. 06(24) - Figure 1.1
A spacetime diagram in natural units.

12) A Magia dos Números - Karlson
Pág. 521 →
"Legislador do Universo".

13) Non-Integrable Dynamics
→ Beck & 401 pages.

14) Ordinary Differential Equation with Application - 3rd.pdf
→ Sze-Bi Hsu

15) Dynamic and Precise Engineering Surveying.pdf
→ Qingquan Li.

16) Sortais - La géométrie du triangle. Exercices résolus - (1997).pdf
→ 220 pages

Pág. 08(12) - Droite et cercle d'Euler.

Obs: 3 points

Analisar com Pappus's Theorem.

17) wurzel2002.pdf

Pág. 28 - Elementargeometrische Lösung

$$\frac{\alpha}{2}+\frac{\beta}{2}+\frac{\gamma}{2}=\pi$$

Obs: Alemão - interessante!!!

18) The Navier-Stokes Problem
in the 21st Century
→ 778 pages.

Pág. 29(52) - 3.4 Euler

Newtonian mechanics implied a new vision of geometry, ...

19) wolfram - Hofstadter Triangle.

20) Fuhrmann Center.

21) A Magia dos Números - Karlson.

Pág. 449 - Excursão aos domínios do transcendente.

Pág. 494 - Fig. 257

$$\text{arc tg } 1 = \frac{\pi}{4}.$$

Pág. 524 - Hooke aparece como culpado: nem bem havia visto o livro, e já afirmava que, no fundo, era ele o pai de todas aquelas idéias e que Newton tudo lhe devia.

Newton chegou mesmo a acrescentar ao terceiro livro esta frase exageradamente laudatória:

"A lei inversa da gravitação é válida para todos os movimentos celestes, como igualmente descobriram, independentemente, meus compatriotas Wren, Hooke e Halley.

... 1703 Newton foi nomeado presidente vitalício da Royal Society

... duas controvérsias desagradáveis: uma com Flamsteed (Flamsteed), o astrônomo real e a outra, Leibniz

Pág. 525 - Clareza e utilidade

Quando refletimos sobre nós mesmos e comparamos o talento que recebemos com o de Leibniz, somos tentados a lançar fora os livros e ir aguardar a morte em algum longínquo recanto do mundo.

Diderot.

22) Berger - Geometry Revealed - A Jacob's Ladder to Modern

Higher Geometry-(2010).pdf

Pag. 713(730) – XI.9

Circles and ellipses.

23) Ramified Surfaces – On Branch Curves and Algebra Geometry in the 20th Century.pdf
→ 258 pages.

Pag. 59(67) – Chapter 3

1900s – 1930s: Branch Curves and the Italian School of Algebraic Geometry.

24) Tony Yuan 7-2.pdf
Gravitational Waves

25) Werner-Time-Dilation-Ref-10-04-2018.pdf
→ Signal Time travel Dilation TEM Mode Time Machine Creation theory 10/04/2018.

26) KEPLER

Great Books – Vol. 15

Ptolemy, Copernicus, Kepler.
4th Ed - 1993 - Pag. 846.

Pág. 988 - What else does "mean elongation" signify?

Pag. 203 - The Almagest, VI

7. Construction of Ecliptic Tables.

Pag. 204 - moon (transit)

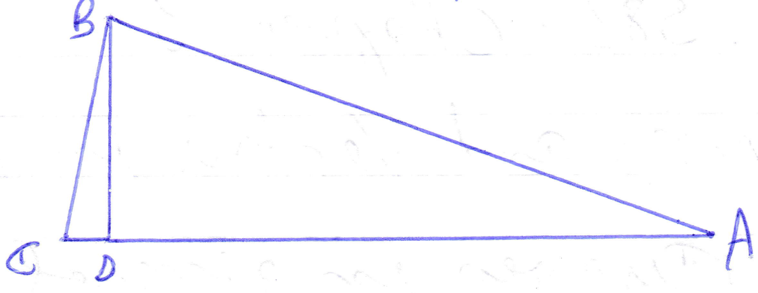

For if we cut off from the node A two equal arcs AB and AC of the proposed circles, and, joining BC, draw BD from B at right angles to AC, then it will be immediately evident that with the moon supposed at B, if we use arc AC of the ecliptic instead of the AD used because of considering the passages relative to it with respect to circles through the ecliptic's poles, then the

difference due to the inclination
of the lunar circle will be CD.

27) The Geometry of Total
Curvature on Complete Open
Surfaces - 2003.pdf
→ Shiohama.

28) Silverman - Modern Calculus
and Analytic Geometry -
(2014).pdf
→ 1169 pages.

Pág. 581 - Chapter 9

Curves and Vectors in R^2

67. Curves in general.
Parametric Equations.

Pág. 584 - Figure 9.2 - Ellipse

29) 978-3-662-64865-0.pdf

Astronomische Phänomene
→ Backhaus - 546 pages.

Pág. 482(504) - 36.2 Theorie

Die Theorie der Fixsternparallaxe

→

Pág. 483(505) - Abb 36.2 ✪

Parallaxe p des Sterns

→ gráfico interessante.

Pág. 484(506) - Abb 36.3 ✪

Ellipse (Planets)

obs: Página interessantíssima.

30) Calculus-and-Linear-Algebra-in-Recipes.pdf

→ Karpfinger - 1015 pages

obs: Livro interessante!!!

Pág. 59(80) - Chapter 7

Complex Numbers: Cartesian Coordinates.

31) A Bridge to Advanced Mathematics From Natural to Complex Numbers.pdf

→ Cioabă - 544 pages.

obs: Livro interessante!!!

32) About the Biggest the Smallest, and Everything Else Travelling through the Universe....pdf

→ Parnovsky - 569 pages.

Pág. 59 (90) –

Newton calculated the force of gravitational attraction between two balls, such as planets and the sun. To do this, he had to create, together with

Gottfried Wilhelm von Leibniz

and other scientists, the foundation of what is now called

integral and differential calculus.

The law of gravitation was first tested in the experiments of the British Henry Cavendish in 1798. They were carried out almost simultaneously with

Coulomb's experiments,

which I will discuss later, and on very similar equipment.

✡ ⇒ The forces of mutual gravitational attraction of two massive balls were a huge number of time less than the forces of electrical interaction in Coulomb's experiments, and their measurement required great skill from the experimenter.

331) Pag. 300 (331) →

Dirac's constant, which is 2π times less than h.

We use the special character "h-bar" for it: $\hbar$

$\hbar = h/2\pi$.

33) Farin - Curves and Surfaces for CAGD - A Practical Guide.pdf

Pag. 205 (224) - Chapter 12

Conic Sections.

Pag. 214 (233) - Figure 12.6 →

The four tangent theorem:

Four points are marked on each of the four tangents to the shown conic. The four cross ratios generated by them are all equal.

Pág. 218 (237) - Figure 12.7

Pascal's theorem: the intersection points P_1, P_2, P_3 of the indicated pairs of straight lines are collinear.

Pág. 219 (238) - 12.6

Two Classic Problems

→ Conic from two points and tangents plus another point.

Fig. 12.8 - Conic constructions:

b_0, b_1, b_2, and the tangent through b_0^1 and b_1^1 are given.

34) Outras referências:

(interessante!!!)

a) www.cmu.edu/dietrich/philosophy/docs/glymour/relativityeclipses.pdf

b) https://royalsociety.org/doi/10.1098/rsta.1920.0009

c) https://royalsociety.org/doi/10.1098/rsnr.2020.0040

d) https://archive.org/details/theory of shape 0000 bors

e) The Theory of Relativity - Møller

f) Interactions (Texts and Monographs in Theoretical Physics).pdf
→ Efimsky - 410 pages.

g) Borsuk Geometry

Borsuk - Foundations of geometry - Euclidean and Bolyai-Lobachevskian geometry - projective geometry - (2018).pdf.

20231026200000

20231027115000

1) Data-NovoVivo-010-2023/
Referencia-index
A Mathematician and His
Mathematical Work.pdf

Pág. 539(548) - Historical Remarks
On Gauss-Bonnet.

Let M be a two-dimensional oriented Riemannian manifold and D a compact domain on M, bounded by a sectionally smooth curve C. The Gauss-Bonnet formula says:

$$\sum(\pi-\alpha)+\int_C \kappa_g \, ds+\iint_D K\, dA = 2\pi\chi(D)$$

where $\chi(D)$ is the Euler characteristic of D and the members at the left-hand side are respectively the exterior angles at the corners, the integral of the geodesic curvature, and the integral of the Gaussian curvature.

These are respectively the point, line, and surface curvatures, so that the formula expresses the total curvature in terms of a topological invariant.

The most important special case of the formula is the theorem on the angle sum of a rectilinear triangle in the euclidean space.

Pág. 548 (557) – What is Geometry?

IMPORTANT:

Pág. 549 (558) –

Euclid realized that the PARALLEL POSTULATE was not as transparent as his other axioms and postulates.

Efforts were made to prove it as a consequence.

Their failure led to the discovery of non-Euclidean geometry by C. F. GAUSS, John Bolyai, and N. I. Lobachevski in the early 19th century.

Pág. 551 (560) →

A fundamental notion is "CURVATURE" in its different forms. Its simplest manifestation

is the circle in plane Euclidean geometry.

① IMPORTANT: O problema está no círculo pois este é representado de forma equivocada. Tanto no plano euclidiano como no "non-plane" euclidiano, nós observamos uma inconsistência matemática.

Nós devemos adequá-los à uma nova realidade que se impõe, ou seja, estas figuras geométricas devem ser prescritas de acordo com os novos parâmetros estabelecidos pela nova Matemática e consequentemente, a uma nova Geometria que se fez presente com a readequação dos princípios matemáticos antigos aos novos, estabelecidos pelo

M.U.S.

→ Mathematical Uniforme Space.

Uma nova matemática para uma nova realidade, realidade esta antiga e até agora, desconhecida pelo mundo da Matemática antiga.

Retornando:

... Euclidean geometry.

It would also be the force of a physical system or the strength of a gravitational or electro-magnetic field.

⇓

Obs: Estudar em conjunto com o item 32 de ontem, livro
'About the Biggest, the Smallest and Everything...''.pdf (gravity → off)
Pág. 59(90') → Coulomb's Experiments.

⇒ Relação de forças entre as estruturas.

In mathematical terms it measures the non-commutativity of covariant differentiation.

It is remarkable that suitable algebraic combinations of curvature give topological invariants.

To illustrate this we wish to state the

Gauss-Bonnet theorem.

Let D be a domain with a sectionally smooth boundary on a two-dimensional Riemannian manifold.

Then the Gauss-Bonnet theorem is

the formula

$$\sum(\pi-\alpha)+\int_{\partial D}\kappa_g\,ds+\iint_{D}K\,dA=2\pi\chi(D),$$

Obs: Checar a fórmula correta
Na página 539 (548) →

$$\iint_{D}$$ e nesta $$\int_{D}\int$$

Gauss-Bonnet Theorem,
materiais para estudo
Source:
Data-NovoVivo-010-2023/
Referencia-index/
Gauss-Bonnet Theorem/

a) Halper.pdf → 11 pages.
A proof of the Gauss-Bonnet Theorem

Pag. 09 - 8. The Local Gauss-Bonnet Theorem

$$\sum_{j=0}^{K}\int_{s_j}^{s_{j+1}}k_g(s)\,ds+\iint_{R}K\,d\sigma+\sum_{j=0}^{K}=2\pi$$

where Kg is the geodesic curvature of the regular arcs of α and K is the Gaussian curvature of S.

Pág. 11

$$\sum_{j=1}^{n} \int_{C_j} Kg(s)\,ds + \iint_{R} K\,d\sigma + \sum_{j=1}^{l} \theta j =$$

$$2\pi(F-E+V) =$$

$$2\pi X(R)$$

b) Lin, Chenjia.pdf
→ 20 pages

Pág. 10 - Figure 12 -
Gaussian curvature signs.

$K<0$, $K>0$ e $K=0$.

Obs: Adequar à minha pesquisa.

$K=0$ → If $K(p)=0$, then one

of the principal curvatures vanish and the surface does not bend along some direction.

c) Math 501 X-7-GaussBonnet.pdf
→ 92 pages.

Pág. 55 — Lemma 1.

Let $X(u,v)$ be an orthogonal parametrization of a neighborhood on an oriented surface S in $\mathbb{R}^3$, and let $w(t)$ be a smooth field of unit vectors along the curve

$\alpha(t) = X(u(t), v(t))$ on S.

Then

$$[Dw/dt] = \frac{1}{2}(EG)^{-\frac{1}{2}}\{G_u\, dv/dt - E_v\, du/dt\} + d\varphi/dt,$$

where $\varphi(t)$ is the angle from X_u to $w(t)$ in the given orientation.

d) The Gauss-Bonnet theorem
→ 10 pages.

Pág. 5 → Definition 3. 5.

The geodesic curvature K_g of a curve is a measure of the amount of deviance of the curve from the shortest arc between two points on a surface.

Definition 3. 6.

The Gaussian curvature K of a surface is an intrinsic measure of the curvature of a surface at a point.

Retornando:

$$\sum(\pi-\alpha)+\int_{\partial D} K_g\, ds+\iint_{D} K\, dA = 2\pi\chi(D),$$

where the first term is the sum of the exterior angles at the corners, the second term is the integral of the geodesic curvature along the sides, the ~~(third)~~ third term is the integral of the Gaussian curvature over D, and $\chi(D)$ is the Euler characteristic of D.

Pag. 553 (562)-

The above provides the geometrical basis of gauge field theory in physics.

Here M is a four-dimensional Lorentzian manifold, so that the Hodge $*$-operator is defined, and we define the codifferential

$$\delta = * d * \qquad (17)$$

M.U.S.

There is discrepancy of terminology and notation, as given by the following table:

mathematics	physics
connection ω	gauge potential A
curvature Ω	strength F

Maxwell's theory is based on a $U(1)$-bundle over M, and his field equations can be written

$$dA = F, \quad \delta F = J, \qquad (18)$$

where J is the current vector.
Actually, Maxwell wrote the first equation as

$$dF = 0, \qquad (19).$$

which is a consequence.

2) Data-NovoVivo-010-2023/
Referencia-index/
A Sampler of Riemann-Finsler Geometry-2004.pdf

Pág. 303(314) - Landsberg Curvature, S-Curvature and Riemann Curvature
→ Shen.

Pág. 327(338) - 11. Riemann Curvature

Pág. 337(348) - 14. Nonpositively Curved Finsler Manifolds

3) Data-NovoVivo-010-2023/
Referencia-index/
Adams - Introduction to Topology - Pure and Applied - (2007).pdf
→ 508 pages

4) Adams - Relativity - An Introduction to Spacetime Physics - (1997, 2000).pdf

Obs: Interessante!!!

Pág. 17(26) - 1.4.

Maxwell's Electromagnetism

1.4.1. the Concept of Field

→ Michael Faraday.

Pág. 20(29) - James Clerk Maxwell

Pág. 60(69) - Example

the orbital velocity of Mercury is about 50 km s-1 whilst that of the Earth is about 30 km s-1.

5) Data - Novo Livro - 010 - 2023/Referencia - index/Quantum - Gravity - magazine

a) New Scientist - 23 September 2023.pdf
→ Dark Star

b) New Scientist - International - Edition --- 9 - September - 2023.pdf
→ the Amazing theory of (Almost) Everything.

c) New-Scientist-International-Edition-28-October-2023.pdf
→ In Search of Quantum Space-Time

d) Wolfram-A New Kind of Science-(2002).pdf

Pág. 985(1001) — Page 349 Search: curve
Circle packings.

fraction $\boxed{\pi/\sqrt{12} \cong 0.91}$ of area-which was proved maximal for periodic packings by Carl Friedrich Gauss in 1831 and for any packing by Axel Thue in 1910 and László Fejes Tóth in 1940.

Pág. 986(1002) — Sphere packings

3D face-centered cubic (fcc) packing $\boxed{\pi/\sqrt{18} \cong 0.74}$, which Johannes Kepler suggested in 1609 might be the maximum possible. This was proved for periodic packings by Carl Friedrich Gauss in 1831, and for any packing by Thomas Hales in 1998.

Pág. 1009 (1025) - Page 418
Intrinsically define curves.

Pág. 1027 (1043) -
the Nature of Space

History of discrete space

In discussing the notion of curved spaces,

Bernhard Riemann

remarked in 1854 that it would be easier to give a general mathematical definition of distance if

space were discrete.

PLANCK LENGTH:

$\text{Sqrt}[\hbar G/c^3] \cong 2 \times 10^{-35}$ meters

since at this scale dimensional analysis suggests that quantum effects should be comparable in magnitude to gravitational ones.

Important: Pag. 1154 (1170) →

Geometry → Euclid gave axioms for basic geometry around 300 BC which were used with fairly little modification for more than 2000 years. In the 1830s, however, it was realized that the system would remain consistent even if so-called

PARALLEL POSTULATE

was modified to allow space to be curved.

6) Data - Novo Vivo - 010 - 2023/
Referência - index/
Alexandrov - Intrinsic Geometry of Convex Surfaces - 2005/

a) Alexandrov - Selected Works Part I.. - 1996.pdf

Pág. 01(12) - Chapter I

On Infinitesimal Bending of Nonregular Surfaces
→ 1936

Study of continuous length - preserving deformations of a surface is sensible not only for regular surfaces:

It suffices that on a surface there be enough curves for which the concept of length is meaningful.

Lebesgue[1] defined the corresponding class of the so-called rectifiable surfaces (la famille des surfaces rectifiables) as follows.

A surface S,

$$x=f(u,v),\ y=\varphi(u,v),\ z=\overset{(\psi)}{\psi}(u,v).$$

given over a domain D in the plane (u,v) is said to be rectifiable, if to each rectifiable curve on D there corresponds a rectifiable curve on S.

IMPORTANT:

For a surface homeomorphic to a sphere,

we naturally take a sphere of unit radius as the domain of parameters. Then, every (convex) Convex surface is rectifiable, as seen if a sphere in the interior of the surface is mapped onto it in (de tal modo) such (em tal) a way that a point of the sphere is sent to the point on the surface lying on the same radius.

Pág. 19 (30) - Chapter II

An Elementary Proof of the MINKOWSKI and some other theorems on convex polyhedra.

→ An elementary proof is given to the Minkowski theorem stating that a convex polyhedron is completely determined by the areas and the directions of its faces.

b) Alexandrov - Selected Works
Part II - Intrinsic Geometry
of Convex Surfaces - 2005.
pdf

Pág. 157 (172) - Chapter V

Curvature

1. Intrinsic Curvature

$$\omega(T) = \alpha + \beta + \gamma - \pi$$

$\uparrow$ wrong !!!

to be the curvature of this triangle T.

the curvature of every shortest arc K on a convex surface is taken to be zero:

$$\omega(L) = 0$$

Let X be a point on a convex surface, and let θ be the complete angle at this point. Then the number

$$\omega(X) = 2\pi - \theta$$

Pág. 197 (212) – Chapter VI

Existence of a convex Polyhedron with a given metric.

Pág. 213 (228) – 4.

Deformation of a Convex polyhedral Angle.

Pág. 214(229) - Lemma 1.

The partial derivative of the plane angle a_n with respect to each of the dihedral angles α_i lying opposite (deitado em frente) a_n is positive.[12]

Fig. 56 - interessante. Estudar!!!

IMPORTANT: Estudar este livro todo!!!

7) Data - Novo Livro - 010 - 2023/
Referencia - index/
Algebra, Arithmetic, and Geometry/
→ 2 volumes.
obs: Fonte de referência.

8) Data - Novo Livro - 010 - 2023/
Referencia - index/
Analytic Projective Geometry
→ John Bamberg
→ 475 pages.

Pág. 368(381) - Chapter 16
Line Geometry.
obs: Pesquisar livro todo.
Bem didático.
20231027172600

20231028 162000 - Dia de S. Judas Tadeu.

Source: CELULAR-2023/
Download-03---20231028/

a) 20231028 - Posamentier - The Secrets of Triangles A Mathematical Journey-(2012).pdf

Pág. 128 - The angle bisector not only bisects the angle, it also partitions the opposite side proportionally to the two adjacent sides. That is, for FIGURE 5-3, we have AD as the bisector of angle BAC, which then gives us

$$\boxed{\frac{c}{b} = \frac{m}{n}}$$

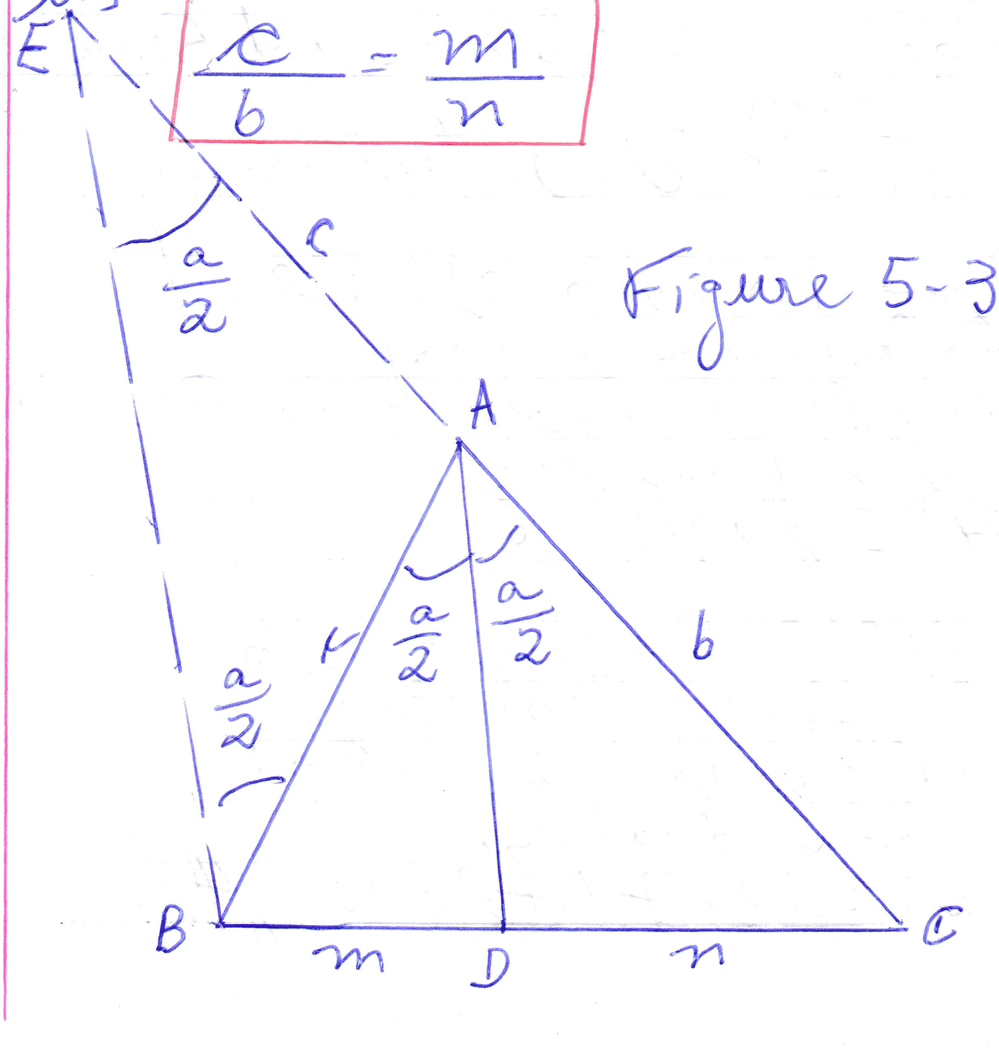

Figure 5-3

Pag. 129

To show why this is true is a bit tricky, as we must bring in some auxiliary lines:

We construct BE parallel to AD and intersecting CA (extended) at point E.

All the angles marked with $\frac{a}{2}$.

are equal.

Therefore, since triangle AEB is isosceles, AE = AB = c.

For triangle BCE, we have

$$\frac{AE}{CA} = \frac{BD}{CD}, \text{ or } \frac{c}{b} = \frac{m}{n}.$$

While it is relatively easy to find the length of and altitude of a triangle (usually using the Pythagorean theorem), it is not so trivial to find the length of an angle bisector of a triangle.

There is a very helpful relationship that can be used to find the length of an angle bisector

Pag. 130 — In FIGURE 5-4, we have triangle ABC, with AD, the bisector of ∠ BAC, which we will call t_a. We begin our development by extending AD beyond D to meet the circumscribed circle of triangle ABC at E.

We then draw BE.

Since $\angle BAD = \angle CAD$, and $\angle E = \angle C$

(both angles inscribed in the same arc),

$\triangle ABE \sim \triangle ADC$, and

$$\frac{AC}{AD} = \frac{AE}{AB}, \text{ or}$$

$$AC \cdot AB = AD \cdot AE = AD(AD+DE) =$$

$$AD^2 + AD \cdot DE$$

However, $AD \cdot DE = BD \cdot DC$, since they are intersecting chords of the same circle.

Substituting equation (II) into equation (I), we obtain $AD^2 = AC \cdot AB - BD \cdot DC$, which is what we sought to develop.

Yet in simpler terms, using the letter designations in

FIGURE 5-3, we can write this relationship as

$$t_a^2 = bc - m \cdot n, \text{ or}$$

$$t_a = \sqrt{bc - mn}$$

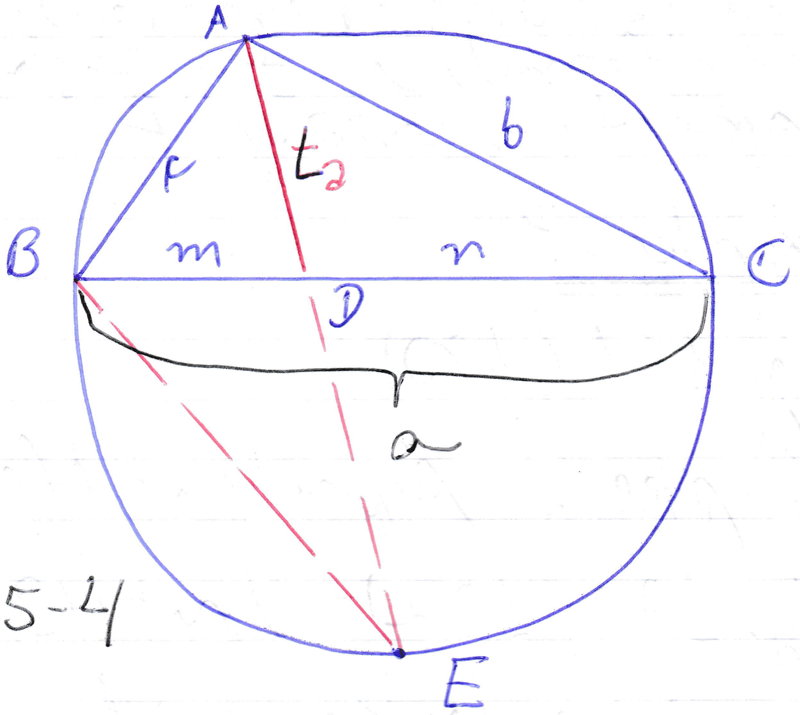

Figure 5-4

For the ambitious reader we provide yet another formula for finding the length of the angle bisector of a triangle given only the three sides:

$$t_a = \frac{2\sqrt{bcs(s-a)}}{b+c},$$

$$t_b = \frac{2\sqrt{cas(s-b)}}{c+a},$$

$$t_c = \frac{2\sqrt{abs(s-c)}}{a+b}, \text{ where}$$

$$s = \frac{a + b + c}{2}.$$

Pág. 131 – the Symmedians of a Triangle.

b) 20231028 – Posamentier – Mathematics – Its Historical Aspects – 2022.pdf

Chapter 3

Pág. 161 (180)

Chapter 3 – Aspects of Measurement

Pág. 173 (192) – Unit 49

How Eratosthenes Measured the Earth

Pág. 180 (199) – Unit 51

Why the Normal Bell-Shaped Curve is called Normal.

Pág. 182 (201)

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

In this formula μ (mu) is the mean or midpoint of the curve, π is the approximate constant 3.14159, e is the approximate constant 2.718282, and σ (sigma) is the standard deviation.

Pág. 184 (203) - The Geometric Mean.

The geometric mean of two numbers a and b is $\sqrt{ab}$, and of three number, a, b, and c is $\sqrt[3]{abc}$.

Therefore, for the above situation the geometric mean of the 3 interest rates is

$$\sqrt[3]{(0.1)(0.4)(0.5)} \simeq 0.27$$

Pág. 185 (204) - Figure 3.10

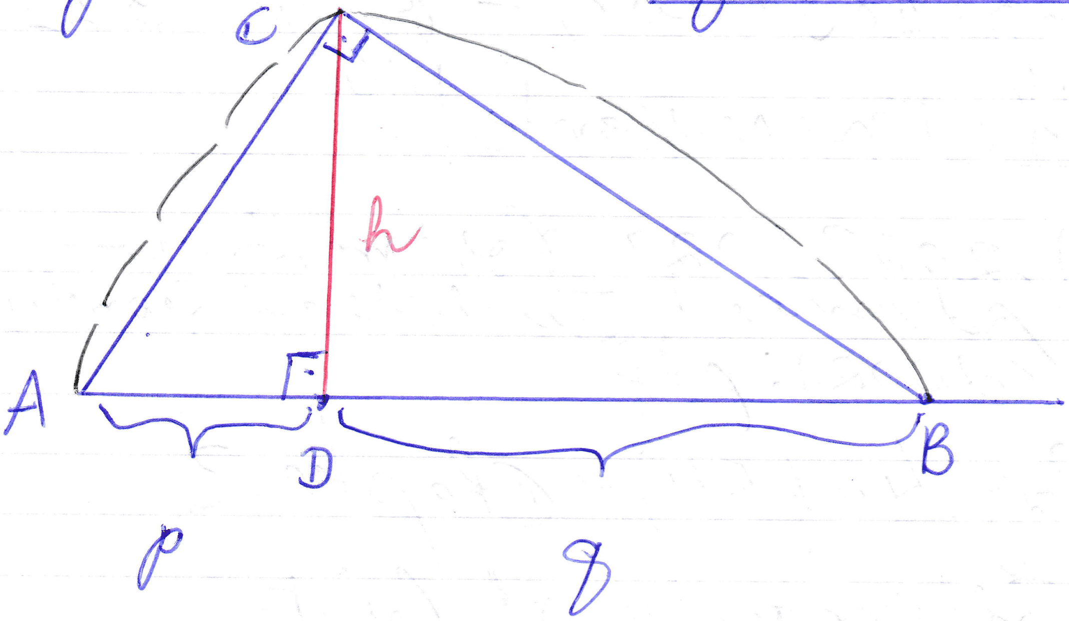

From the triangle similarity, $\Delta ADC \sim \Delta CDB$, we get the proportional relationship

$\frac{AD}{CD} = \frac{CD}{BD}$, or $\frac{p}{h} = \frac{h}{q}$.

this gives us $h = \sqrt{pq}$, which is one reason why h is considered the geometric mean between p and q.

Pág. 188 (207) - Unit 54

Comparing the Three Means Geometrically.

Pág. 189 (208) - Figure 3.11

Obs: Interessante!!!

c) 20231028 - Posamentier - Advanced Euclidean Geometry - (2002).pdf

Pág. 42 (54) - Chapter 3

Collinearity of Points

Pág. 57(69) - Figure 3-16

Pascal's theorem (variation on)

Pág. 57(69) - Brianchon's theorem

Pág. 58(70) - theorem 3.5

Brianchon's theorem

If a hexagon is circumscribed about a circle (see Figure 3-17), the lines containing opposite vertices are concurrent.

Pág. 60(72) - Pappu's Theorem

Figure 3-19.

d) 20231028 - Torretti - Relativity and Geometry - (1983).pdf
⟹ 406 pages.

Pág. 48(59) - Chapter 3

Einstein's "Electrodynamics of Moving Bodies".

Conrad Habicht ⟶ 4 papers

the fourth paper, still a draft, would contain

(IMPORTANT:) ⟶

"an electrodynamics of moving bodies, involving modifications of the theory of space and time."3

Referência:

Pág. 292(303) → Section 3.1

3. "Die vierte Arbeit liegt erst im Konzept vor und ist eine Elektrodynamik bewegter Körper unter Benutzung von Modificationen der Lehre von Raum un Zeit"...

Important:

Pág. 48(59) → cont.

This article, submitted in June 1905 to Annalen der Physik and published in September, laid down the foundations of Special Relativity.

In 1952, Einstein told R. S. Shankland that the subject of this paper

"had been his life for over seven years",

that is, from 1898 to 1905, while that of the former had buried him for five, that is, since the publication of Slavek's trailblazing paper of 1900.

Obs: Estudar este livro todo!!!

Important:

No capítulo 3, nós observamos aspectos muito interessantes, o uso de ideias de outros pesquisadores como um fator catalítico de uma nova descoberta, mas neste caso, sem menção aos seus originais idealizadores.
Ex: Habicht, Lorentz, Maxwell.

Obs:

Checar as influências destes pesquisadores nas pesquisas de Einstein.

Pág. 275 (286) — Appendix

6. The Torsion and Curvature of a Linear Connection.
(Estudar com cuidado!!!)

e) 20231028-Farouki-Pythagorean-hodograph curves-algebra and geometry inseparable-(2007).pdf
→ 731 pags.

Pág. 104(116) – 7.3

Curvilinear Coordinates

Gabriel Lamé (1795-1870) in his treatise: "Leçons sur les coordonnés curvilignes et leurs diverses applications" of 1859.

He used transformations to curvilinear coordinates in solving certain problems of heat conduction and elasticity.

Pág. 106(118) – 7.3.2

Distance and Angle Measurements.

Obs. Einstein pode ter se inspirado para analisar a curvatura do espaço.

Pág. 108(120) –

Magnitude of the vector (7.41) is evidently given by

$$|v| = \sqrt{(v^{\xi})^2 + (v^{\eta})^2 + 2 v^{\xi} v^{\eta} e_{\xi} \cdot e_{\eta}} \quad (7.42)$$

Note the "cross term" above involving the scalar product of the basis vectors.

Important:

ORTHOGONAL COORDINATE SYSTEM

It is only in the special case of an orthogonal coordinate system - for which e_ξ and e_η are

everywhere mutually perpendicular

Obs: Estudar com profundidade esta afirmação.

checar

⇒ Eu tenho as minhas dúvidas !!!

Aprofundar com cautela !!!

Cont. (retornando ao texto)

everywhere mutually perpendicular - that this term is absent (the simplest example is, of course, Cartesian coordinates).

Similarly, given two vector

u an v with curvilinear ~~Coordi~~ components

(u^ξ, u^η) and (v^ξ, v^η) at a prescribed point (ξ, η), we can form their dot product
(product)

$$u.v = u^\xi v^\xi + u^\eta v^\eta + (u^\xi v^\eta + u^\eta v^\xi)\, e_\xi . e_\eta,$$

and thus determine the angle θ between them from

$$\cos\theta = \frac{u.v}{|u|\,|v|}, \qquad (7.43)$$

where $|u|$ is given in terms of (u^ξ, u^η) by an expressions analogous to (7.42).

We can also avoid the normalization (7.40) and use the metric tensor directly to compute angle between vectors: we replace the numerator in (7.43) by

(R)

$$[u^\xi, u^\eta] \begin{bmatrix} r_\xi . r_\xi, & r_\xi . r_\eta \\ r_\xi . r_\eta & r_\xi . r_\eta \end{bmatrix} \begin{bmatrix} v^\xi \\ v^\eta \end{bmatrix},$$

and similarly $|u|$ and $|v|$ are replaced by the square roots of the scalar values obtained by multiplying the metric on the left and right by the components of u and v, respectively.

Pág. 118(130) →
Interessante!!!

Fig. 7.7 Line representations of an ellipse (left) and the cubic $y = x^3$ (right)

Obs: Utilizar estes estudos de ellipse.

Pág. 122(134) - 7.4.5

Invariance of the Cross Ratio

↳ utilizar estes conceitos para a estabilidade da pesquisa.

Pág. 197 (retornar)

Pág. 192(204) → $\boxed{\text{Figure 8.28}}$

→ interessante!!!

Figure 8.28 illustrates lines of curvature on a "triaxial ellipsoid"

(all three semi-axes are of unequal length)

⇑ OBS:

checar a viabilidade de utilização quando projetados ortogonalmente em um plano.

– in this case, the lines of curvature are all closed loops. They form an orthogonal net, except at the umbilics situated on the section containing the smallest and largest semi-axes. These curves were first accurately drawn – in plane projection – by

Gaspard Monge in his 1807

treatise "Application de l'Analyse à la Géométrie".

...

– the surface "patches" delineated by the lines of curvature would
⇒ define the ~~sp~~ shapes of the roof tiles, and

he proposed that chandeliers be suspended from two umbilical points - see [367, p. 269].

The lines of curvature shown in Fig. 8.28 admit an elegant geometrical interpretation.

For constants a, b, c and three different instances

λ, μ, ν

(such that $0 < \lambda < c^2 < \mu < b^2 < \nu < a^2$)

of the parameter η in the equation

$$\frac{x^2}{a^2-\eta} + \frac{y^2}{b^2-\eta} + \frac{z^2}{c^2-\eta} - 1 = 0 \quad (8.100).$$

Pag. 197 (209) - Chapter 9

Algebraic Geometry

Hence, to describe an algebraic space curve,

two polynomial equations

$g(x, y, z) = h(x, y, z) = 0$ in

the three Cartesian coordinates (x, y, z)

must be specified.

But this simply amounts to formulating the space curve as the set of points common to the two surfaces

$g(x, y, z) = 0$ and

$h(x, y, z) = 0$ — i.e.,

as the intersection of those surfaces.

This fact incurs difficult theoretical issues, concerning the conditions under which two algebraic surfaces will intersect in a

single "irreducible" space curve

or in a composite of such curves.

Pág. 219(231) - Elliptic Curves

Pág. 185(197) - 8.5.3

Curves Lying on a Surface.

Obs: Não ampliar muito a extensão do trabalho (fator tempo)

20231028193100

20231029 084500.

a) 20231028- Carroll- Spacetime and Geometry an introduction to General Relativity-(2022).pdf

Pág. 06 (19)- Figure 1.3
Two-dimensional Euclidean space, with two different coordinate systems. Notions such as "the distance between two points" are independent of the coordinate system chosen.

Source: CELULAR-2023/
Download-03---20231028/
20231028-Carroll-Spacetime and Geometry.... (2022).pdf

Pag. 93(106)- Chapter 3
Curvature

Estudar em conjunto:

Celular-2023/Download-04---20231029/
20231029-Applied Mathematics in Integrated Navigation Systems-(2007).pdf
Pág. 19(38)- 2.4 Direction Cosine Matrices
2.4.1. Coordinate Transformations.
Fig. 2.1 Vector/coordinate frame.

↘ Fig. 2.2 Vector/coordinate frame components.

obs: Estudar estes dois graficos e fazer simulação em curvatura.

Important:

usar conceitos "em parte"

Source:

Celular-2023/Download-04--20231029/
20231029-Lamé-Leçons sur
les coordonnés curvilignes...
-bpt6k9668033s.pdf
→ 412 pages. - 1859.

Upgraded:

Este livro tenta fazer os cálculos sempre como referencial as coordenadas cartesianas, mas nós devemos alterar este padrão para eixos curvos ou projetá-los sobre um plano.

Retornando:

20231029-Applied Mathematics...pdf
Pág. 28(47) - 2.4.2
Direction Cosine Matrix Rates
Fig. 2.7. Rotating vector/coordinate frame components.

b) Data-NovoVivo-010-2023/
Referencia-index/
23rd_Solvay_Physics.pdf
Pág. 98(119)-
quantum geometry l_p
"stringy" geometry l_s
classical geometry smooth

Pág. 103(124) - The moduli space of conformal field theories.
→ gráfico (interessante)

Pág. 111(132) — (i) (ii) (iii)
l_p l_s smooth.

c) Data-Novo Vivo-010-2023/
Referencia-index/100 years Werner Heisenberg-Works and Impact-(2002).pdf
Pág. 274(285) - 3.3
Gravitational instantons
→ splace Einstein equations is a consistent background of the EFS model preserving certain fraction of the supersymmetry.
The field configuration is given by

$$ds^2 = \frac{dr^2}{1-\frac{a^4}{r^4}} + \frac{r^2}{4}\left(d\theta^2 + \sin^2\theta\, d\phi^2\right) + \frac{r^2}{4}\left(1-\frac{a^4}{r^4}\right)\left(d\psi + \cos\theta\, d\phi\right)^2,$$

$$F = \frac{2}{r^4}(e^3 \wedge e^0 + e^1 \wedge e^2) = \dot{F},$$

$\phi = 0$; $a = 0$; $g_1 = g_2$

with the above choice of background fields, the supersymmetry variations give the projector conditions,

$$(1 - \gamma_5)\epsilon = 0, \quad (\alpha^3 + \dot{\alpha}^3)\epsilon = 0$$

With these projectors the killing spinor equations look really simple, namely,

$$\partial_\mu \epsilon = 0$$

So the killing spinors are independent of b_z, θ, ϕ, ψ.

Because of the twin supersymmetric conditions, the solution preserves

$\boxed{\frac{1}{4}}$ of the supersymmetry.

↳ checar!!!

Obs: Erro na curvatura quando usamos sin, cos... ...tan.

→ Fique atento!!!

d) Data-NovoVivo-010-2023/
Referencia-index/
100 years of (Chronogeometrodynamics)
Chronogeometrodynamics -
Einstein's Gravitation.pdf
→ 470 pages

Pág. 250 (259) - Strategies to Ascertain the Sign of the Spatial Curvature

Abstract: The second law of thermodynamics, in the presence of gravity, is known to hold at small scales,

as in the case of black holes and self-gravitating radiation spheres.

Friedmann-Lemaitre-Robertson-Walker (FLRW)
→

Homogeneous and isotropic universe models are usually described by the FLRW metric

$$ds^2 = -c^2 dt^2 + a^2(t)\left\{\frac{dr^2}{1-kr^2} + r^2(d\theta^2 + \sin\theta\, d\theta^2)\right\}.$$

coupled to the sources of the gravitational field.

Obs: Comparar com a equação do item c.

e) Data-NovoVivo-010-2023/
Referencia-index/
0806.3617v1 --- Chromogeometry.
pdf → 15 pages.

⇒ Interessante!!! Completa o item d.

Pág. 02 - Euler Lines and Nine-point circles in relativistic settings.

Obs: Fazer o download das referências.

e1) Affine and Projective Universal Geometry →

arXiv: math/0612499

e2) Chromogeometry and relativistic conics →

arxiv: math/0806.2789

Obs: Estes arquivos estão armazenados
Data-NovoVivo-010-2023/
Referencia-index/

e_1 → 20231029-arxiv-0612499
e_2 → 20231029-arxiv-0806.2789

f) Data-NovoViVO-010-2023/
Referencia-index/
022-Quantum Field theory
in Curved Spacetime
Quantized Fields and Gravity.pdf
→ 473 pages. - 2009.

Pág. 36(54) - Chapter 2
Basics of quantum fields in
curved spacetimes.

g) Data-NovoVivo-010-2023/
Referencia-index/
091-- Deformed Spacetime_
Geometrizing Interactions
in Four and Five Dimensions-
2007.pdf
→ 517 pages - 2007.
Obs: Livro todo!!!

Pág. 199(224) - Chapter 13
The Shadow of Light:
Lorentzian Violation of
Electrodynamics in Photon
Systems.

→ interessante!!!

h) Data-NovoVivo-010-2023/
Referencia-index/
978-3-031-09086-8-
Curvilinear-Micromagnetism.
pdf

→ P420 pages - 2022.

Obs: Completa o item d.

⇒ The second law of thermodynamics
→ small scales
↳ Micromagnetism

Important:
Estudar as relações entre micromagnetismo e small scales.

Pág. 07(25) - 1.3.1.

Wire with a Constant Curvature
- Ring-Shaped Wire

1.3.1.1. - Equilibrium States

$\Theta_0 = \pi/2$ (wrong!!!)

and the azimuthal angle ϕ which satisfies the pendulum equation

$$\phi'' - \sin\phi \cos\phi = 0 \quad (1.13)$$

20231029165000

20231030135300

1) Data-NovsViv-010-2023/
Referencia-index/Gaspard-Monge/
20231029-Monge-Application
de l'Analy à la Géométrie-1807.pdf
→ 491 pages

Obs: a) Estudar em conjunto com
20231029-Lamé-Leçons-sur-
les-coordonnées-curves....pdf
→ 412 pages.

b) 20231030-Théorie des Coordonnées
curvilignes quelconques....pdf
→ 75 pages.

2) 20231029-Monge-Descriptive Geometry-
The Spread of a Polytechnic
Art.pdf
→ 446.

2) Referencia-index/
From-weak-interaction-to-
gravity.pdf
→ 16 pages.

3) Referencia-index/
[sixth edition] Joseph A. Wolf-
Spaces of Constant Curvature
(2011, AMS).pdf
→ 439 pages.

4) Referencia-index/Gaspard-Monge/20231030-Darstellende Geometrie (1798)-1900.pdf
→ 218 pages - Gaspard Monge

5) Referencia-index/20231030-From Gauss to Painleve.pdf
→ 354 pages!

Pág. 28(38)- (Hypergeometric differential equation)

1. Definition and basic facts
1.1 The Gauss hypergeometric equation.

Pág. 62(72)- 3.4.
The hypergeometric Euler transform

Obs: Pág. 68(78) →

Lemma 3.5.2

Suppose $|x| < 1$ and $|\arg(-x)| < \pi$,

then we have

$$h(t) = O\left(t^{(\alpha+\beta-\gamma-1)} e^{-\{\pi|\Im(t)| + \arg(-x)\cdot\Im(t)\}}\right)$$

as t tends to ∞ in such away that

$\delta < |\arg t| \leq \frac{\pi}{2}$ or $\mathcal{R}(t) \in N + \frac{1}{2}$

holds, δ being an arbitrary positive number. (Here O is Landau's symbol.)

5) Referencia-index/
20231030- General theory of Irregular Curves-(1989).pdf
→ 296 pages.

Pág. 217 (225) – Chapter VIII

Torsion of a Curve in a three-Dimensional Euclidean Space

Obs: Pág. 251 (259) –
the straight line λ lies in P and as $\omega(z)$ let us choose the shortest line connecting the two-dimensional element (X, λ, P) with the two-dimensional element $(X, \lambda, P_r(X))$.

Referencia-index/
6) 20231030- Geometric Aspects of Convex Sets with the Radon-1983.pdf
→ 484 pages.

Obs: Não é muito útil.

7) Referencia-index/

20231030 - Geometric measure theory - (1996).pdf

Pág. 207 (216) - Chapter 3
Rectifiability

This chapter contains the basic facts concerning integration with respect to m dimensional measures over subsets of n dimensional Euclidean space.

→ Lipschitzian maps.

→ (includes the boundary formulae of Gauss, Green and Stokes)

Pág. 241 (250) → 3.2.
Area and coarea of Lipschitzian maps.

3.2.1. Assuming that f maps a subset of R^m into R^n, we recall 1.7.6 to define for each nonnegative integer k the k dimensional Jacobian

$$J_k f(a) = \| \wedge_k D f(a) \|$$

whenever f is differentiable at a;

similarly we define

$$\mathrm{ap}\, J_K f(a) = \| \Lambda_K \mathrm{ap}\, D f(a) \|$$

wherever f is approximately differentiable at a.

8) Referencia-index/
20231030-Problems in geometry-1984.pdf
→ 280 pages.

9) Referencia-index/
Geometry-Hilbert-Space-Pi/

a) A Hilbert Space Problem Book-(1974).pdf
→ 387 pages.

b) A history of non-Euclidean geometry-evolution of the concept of a geometry.pdf
→ 483 pages.

Pág. 218(206) → Chapter 6
Lobačevskian Geometry.

Pág. 280(292) → Chapter 8
The Curvature of Space

Curvature and Intrinsic Geometry of a Surface in the Works of Euler.

By the curvature of a curve at a point we mean the limit of the ratio of the angle $\Delta\alpha$ between the tangents at the endpoints of an arc to the length Δs of that arc as the latter contracts to the point; that is, the limit

$$K = \frac{d\alpha}{ds} = \lim_{\Delta s \to 0} \frac{\Delta\alpha}{\Delta s}. \qquad (8.1)$$

K is also the reciprocal of the radius of curvature at the point in question, that is, the radius of the

$\boxed{\text{osculating}}$ $\boxed{\text{circle}}$

at that point. (the osculating circle at a point P of a curve is defined as the limit of circles determined by three points on the curve as they tend to P).

Osculating → first ⇒ Leibniz

① Leibniz → "New reflections on the nature of the angles of tangency and osculating and their use in mathematical practice to replace complex figures with simpler ones → 1686

The concepts of the curvature of a curve and of the osculating (literally "kissing" from the Latin osculans) circle were already known to Leibniz. ①

Leibniz also suggested the possibility of caracterizing the curvature of a surface by means of an osculating sphere. ②

② In a letter to Johann Bernoulli, dated July 29, 1698.

Euler → "Investigations of the curvature of surfaces" – 1767.

(Pag. 1-22 – vol. 28)

Euler called a normal section of a surface $z = f(x, y)$ perpendicular to the xOy plane a principal section. By varying the angle φ

between a normal section and the principal section Euler found that at each point of the surface there is a maximal radius of curvature f and a minimal radius of curvature g, that their planes are perpendicular to one another, and

Pág. 281 (293) →

that the connection between an arbitrary radius of curvature

$$r = r(\varphi)$$ and the

radii of curvature f and g is given

Fig. 101

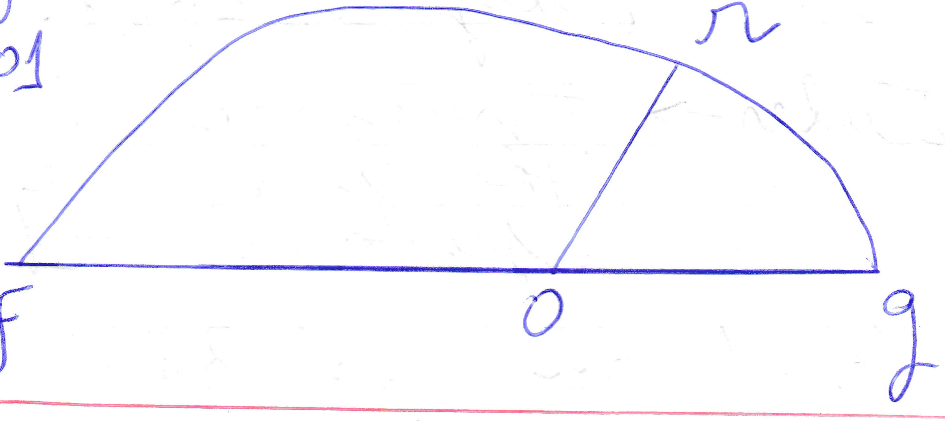

by

$$\boxed{r = \frac{2fg}{f+g-(f-g)\cos 2\varphi}}$$

c) A History of Pi-(1971).pdf
→ 203 pages.

d) A la poursuite de Pi-(2006).pdf
→ 292 pages.

10) Data-NovoVivo-010-2023/
Referencia-index/
Quantum-Field-theory/

a) A Modern Course in Quantum Field Theory, volume 1....pdf
→ 580 pages.

Pag. 201 - Chapter 4
The electromagnetic field and Yang-Mills gauge interactions.
(mediar interações)

... 4.1. Covariant formulation of classical electrodynamics

4.1.1 the field tensor

$$\vec{\nabla}\vec{E} = \rho, \text{ Gauss's law} \quad (4.1)$$

$$\vec{\nabla}\vec{B} = 0, \text{ Non-magnetic monopole law} \quad (4.2)$$

$$\vec{\nabla} \times \vec{E} = -\frac{1}{c}\frac{\partial \vec{B}}{\partial t}, \text{ Faraday's law} \quad (4.3)$$

Pág. 202 →

$$\vec{\nabla} \times \vec{B} = \frac{1}{c}\left(\vec{J} + \frac{\partial \vec{E}}{\partial t}\right),$$

Ampère-Maxwell's law (4.4)

The Lorentz force law expresses the force exerted on a charge q moving with a velocity $\vec{u}$ in the presence of electric and magnetic fields $\vec{E}$ and $\vec{B}$.

This is given by

$$\vec{F} = q\left(\vec{E} + \frac{1}{c}\vec{u} \times \vec{B}\right). \quad (4.5)$$

The continuity equation expresses local conservation of the eletric charge. It reads

$$\frac{\partial \rho}{\partial t} + \vec{\nabla}\vec{J} = 0. \quad (4.6)$$

We now consider the following Lorentz Transformation

$$x' = \gamma(x - vt)$$

$$y' = y \qquad z' = z$$

$$t' = \gamma\left(t - \frac{v}{c^2}x\right). \qquad (4.7)$$

...

b) A Modern Course in Quantum Field Theory, Volume 2...pdf
→ 508 pages.

Pág. 34 – the intensity of weak interactions in the GSW electroweak theory is given in terms of g by the weak fine structure constant α_e which is defined by

$$\alpha_e = \frac{g^2}{4\pi}, \quad \frac{g^2}{8m_W^2} = \frac{G_F}{\sqrt{2}(\hbar c)^3}. \qquad (9.8)$$

Pág. 45 – 9.2 The standard model

9.2.1 The Glashow-Weinberg-Salam electroweak theory and lepton sector.

The GWS electroweak theory which is the first component of the standard model, is a spontaneously broken gauge theory. [8-10].

Off: Volume 1.

Pag. 564 - C 2

Lorentz, Biot-Savart's and Ampère's laws.

Retornando ao Volume 2:

Pag. 87 -

The Euclidean action S_F^{Eucl} is obtained by (i) making the replacement $x_0 \to -ix_4$ wherever x_0 appears explicitly, (ii) substituting $\psi^E(x) = \psi(\vec{x}, x_4)$ for $\psi(x) = \psi(\vec{x}, t)$, (iii) making the replacements $A^0 \to iA_4$ and $D^0 \to iD_4$ and

⇒

(iv) multiplying the obtained expression by $-i$.

Since in Euclidean space the Lorentz group is replaced by the four-dimensional rotation group, we introduce new γ-matrices γ^E_μ as follows

$$\gamma^E_4 = \gamma^0, \; \gamma^E_i = -i\gamma^i.$$

they satisfy $\{\gamma^E_\mu, \gamma^E_\nu\} = 2\delta_{\mu\nu}$.

The fermion Euclidean action is then

$$S_F^{Eucl} = \int d^4x \, \bar\psi^E(x) \left(\gamma^E_\mu D_\mu + M\right) \psi^E(x). \quad (10.135).$$

...

Pág. 361 - 15.1

Lorentz symmetry revisited

15.1.1 Lorentz/Poincaré invariance

Important: →

The spacetime metric is taken to be

$$g_{\mu\nu} = (+1, -1, -1, -1). \quad (15.1)$$

We will use the Weyl or chiral representation of Dirac matrices given by

$$\gamma^{\mu} = \begin{pmatrix} 0 & \sigma^{\mu} \\ \bar{\sigma}^{\mu} & 0 \end{pmatrix}, \quad \gamma_5 = i\gamma^0\gamma^1\gamma^2\gamma^3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

(15.2).

11) Data-NovoVivo-010-2023/
Referencia-index/
Curves and Surfaces/
Abate-Curves and Surfaces-(2012).pdf
→ 407 pages.

Pág. 165 (177) - Chapter 4
Curvatures

12) Data-NovoVivo-010-2023/
Referencia-index/
Advanced-Calculus/

a) Advanced Calculus - A Differential Forms Approach.pdf
→ 528 pages.

b) Advanced Calculus - Kaplan.pdf
⟶ 754 pages

Pág. 531(543) - Chapter 8
Functions of a Complex Variable

z-plane ⟹ $z = x + iy$.

Pag. 532(544) - Figure 8.1
The complex z-plane

Obs: Está representado no plano cartesiano x e y, mas nós devemos fazer a adequação para um plano curvo projetado ortogonalmente sobre um plano.

⟹ M.U.S.

c) Advanced Euclidean Geometry - (2007).pdf
⟶ 333 pages.
Obs: An elementary treatise on the geometry of the triangle and the circle.

Pag. 195(209) - Chapter XI
The Nine-Point Circle

Obs: Pag. 161 (175) →
three Notable Points

Pág. 162 (176) - Fig. 50

Important:

Pág. 165 (179) - 258
The Nine Point Circle.

Interessante:

Pág. 165 (179) - Orthocentric
Systems

259. Definition.
An orthocentric system
is a set of four points one of
which is the orthocenter of the
triangle of the other three.

Pág. 166 (180) - Fig. 52

Reforçando:

Pág. 195 (209) - Nine-Point Circle

Pág. 200 (214) - The Theorem of
Feuerbach

320. Theorem.
The nine point circle of a triangle is tangent to the inscribed circle and to each of the escribed circles.

Pág. 201(215) - 322 - Fig. 65

Pág. 206(220) - 326
We may now study further the properties of (Simson)

SIMSON LINES in a

triangle, and establish simple relations associating them with the nine point circle.

Further theorems about these lines will be developed at length in Chapter XIV.

Fig. 67 (67)

Theorem. The Simson line of any point on the circumcircle is perpendicular to the isogonals of the rays from the vertices to the given point.

Pág. 208(222) - Fig. 68

328. Theorem:
The Simpson lines of two diametrically opposite points of the circumcircle intersect at right angles on the nine point circle.

Pág. 213(227) - Chapter XII

The Symmedian Point and other Notable Points.

341. Definition:
The line isogonal to any median of a triangle is called a symmedian.
The point of concurrence of the symmedians, the isogonal conjugate of the median point, is called the
symmedian point K.

⇒ Obs: French and British writers it is usually called Lemoine's point, and by German, Grebe's.

12) Data-NovoVivo-010-2023/
Referencia-index/
Einstein-Lorentz-Relativity/
Advanced-Electromagnetic-Theory.pdf

Obs: 310 pages - Livro todo!!!

Pág. 87(99) - Chapter 4
Electrodynamics and
Electromagnetic waves.
→ Maxwell's equations

Pág. 135(147) - Chapter 5
Relativity and Electrodynamics

5.1 - Lorentz Transformation

Since Maxwell's equations involve derivatives with respect to space and time, they are clearly dynamical laws!

... Galilean transformation

$$x = x' + v t', \quad (5.1)$$

Pág. 137(149) -

Lorentz in 1904 established the mathematical result that Maxwell's equations can be true in all inertial frames if they are connected to each other by what we now call the

Lorentz transformation

$$x = \gamma(x' + v t'), \qquad (5.6)$$

$$y = y', \qquad (5.7)$$

$$z = z', \qquad (5.8)$$

$$t = \gamma\left(t' + \frac{v}{c^2} x'\right), \qquad (5.9)$$

where

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}} \qquad (5.10)$$

Obs: Einstein pode ter deduzido desta equação.

Note that $x' = 0$, $t' = 0$ corresponds to $x = 0$, $t = 0$, implying that time is set to zero in both the frames at the instant when the origins of the two frames coincide.

The revolutionary implication of (5.9) was the time measured by different observers can be different.

Obs: Ler livro todo!!!

20231030183300

20231031 0900,

1) Data-NovoVivo-010-2023/
Referencia-index/
Curvature/
a) Hartle-Gravity-An Introduction to Einstein's General Relativity-2014.pdf
→ 549 pages.
Pag. 107(108) - Chapter 6
Gravity as Geometry
Pag. 125(126) - 6.5/
Spacetime is Curved

In the presence of a gravitational field, spacetime remains flat, but clocks run at a rate that is a factor $(1+\phi/c^2)$ diferent from their rates in empty spacetime, where ϕ is the gravitational potential at the location of the clock.

Clocks run faster where ϕ is positive and slower where ϕ is negative.

Pag. 126(127) - 6.6.
Newtonian Gravity in Spacetime Terms
(Field Metric)

$$ds^2 = -\left(1+\frac{2\phi(x^i)}{c^2}\right)(c\,dt)^2+\left(1-\frac{2\phi(x^i)}{c^2}\right)(dx^2+dy^2+dz^2) \qquad (6.20)$$

Pág. 135(136) - Chapter 7
The Description of Curved Spacetime

$$ds^2 = -dt^2 + dx^2 + dy^2 + dz^2 \quad (7.1)$$

in the $c = 1$ units.

Pág. 137(138) - Box 7.1

The Penrose Diagram for Flat Space

Pág. 138(139) → 7.2 Metric

To describe a general geometry use a system of four coordinates x^α, to label the points and specify the line element giving the distance, ds^2, between nearby points separated by coordinate intervals dx^α. That line element will have the form

$$ds^2 = g_{\alpha\beta}(x)\, dx^\alpha dx^\beta, \quad (7.8)$$

where $g_{\alpha\beta}(x)$ is a symmetric, position-dependent ① matrix called the metric.

Obs: ① When dealing with functions of the coordinates, we routinely use the abbreviations $f(x^\alpha)$, or $f(x)$, for $f(x^0, x^1, x^2, x^3)$ where there is no danger of confusion.

For example, the metric for flat spacetime (WRONG !!!) in polar coordinates (7.4) is

$$g_{\alpha\beta}(x) = \begin{array}{c|cccc} & 0 & 1 & 2 & 3 \\ \hline 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & r^2 & 0 \\ 3 & 0 & 0 & 0 & r^2 \sin^2\theta \end{array} \qquad (7.9)$$

Diagonal metrics such as this can be specified more compactly by writing $g_{\alpha\beta}(x) = \mathrm{diag}(-1, 1, r^2, r^2\sin^2\theta)$.

Obs: Este cálculo esta errado poi vocês estão descrevendo uma superfície plana (que não é plana) em coordenadas polares.

Pág. 445 (444) – Chapter 21
Curvature and the Einstein Equation

$$\begin{pmatrix}\text{a measure of local} \\ \text{spacetime curvature}\end{pmatrix} = \begin{pmatrix}\text{a measure of} \\ \text{matter energy density}\end{pmatrix} \qquad (21.1)$$

This relation, called the Einstein equation (or Einstein's equation), is field equation of general relativity in the way that Maxwell's equations are the field equations of electromagnetism.

Pág. 450(449) - 21.2

Equation of Geodesic Deviation

Figure 21.3

Obs: O eixo x e y deveriam ser reescritos em um novo padrão

MUSICH COORDINATE //

em que nós observamos uma curvatura nas coordenadas x e y, posteriormente, estes serão projetados ortogonalmente em uma superfície plana.

M.U.S.

⇒ where π < Pi,

em virtude da curvatura do space gerada pela energia.

See:

Pág. 452(451) Table 21.1

Newton Gravity and General Relativity.

→ interessante!!!

Vejamos:

	NEWTON GRAVITY	GENERAL RELATIVITY
Basic field quantity	Gravitational potential $\phi(x^i, t)$	Metric $g_{\alpha\beta}(x)$
Equation of motion (involving the first derivate of field quantity)	Newton Law $\frac{d^2x^i}{dt^2} = -\underset{\delta}{g}^{ij}\frac{\partial\phi}{\partial x^j}$	Geodesic equation $\frac{d^2x^\alpha}{d\tau^2} = -\Gamma^\alpha_{\beta\gamma}\frac{dx^\beta}{d\tau}\frac{dx^\gamma}{d\tau}$
Equation of motion for the deviation between two particles (involving second derivatives)	Newton deviation $\frac{d^2\chi^i}{dt^2} = -\delta^{ij}\left(\frac{\partial^2\phi}{\partial x^j\partial x^k}\right)\chi^k$	Geodesic deviation $\frac{d^2\chi^{\hat\alpha}}{d\tau^2} = -R^{\hat\alpha}_{\hat\tau\hat\beta\hat\tau}\chi^{\hat\beta}$
Quantity determining the acceleration of the deviation	Tidal forces $\frac{\partial^2\phi}{\partial x^i\partial x^j}$	Riemann curvature $R^\alpha_{\beta\gamma\delta} = \frac{\partial\Gamma^\alpha_{\beta\delta}}{\partial x^\gamma} - \frac{\partial\Gamma^\alpha_{\beta\gamma}}{\partial x^\delta} + \Gamma^\alpha_{\gamma\epsilon}\Gamma^\epsilon_{\beta\delta} - \Gamma^\alpha_{\delta\epsilon}\Gamma^\epsilon_{\beta\gamma}$
Field equation in a vacuum	Laplace's equation $\nabla^2\phi = \delta^{ij}\frac{\partial^2\phi}{\partial x^i\partial x^j} = 0$	Vacuum Einstein equation $R_{\beta\delta} = R^\alpha_{\beta\alpha\delta} = 0$

Pag. 471(470) - Chapter 22
The Source of Curvature

Voltando: Pag. 445(444) - 21.1
TIDAL GRAVITATIONAL FORCES

→ to find the "measure of spacetime curvature" in (21.1)

b) Hartley - the fractional trigonometry.
... pdf → 435 pages

Pág. 74(89) - 5.5.
Complexity-Based Hyperbolic
Functions (precisa ajustar!!!)

c) Harvey - Geometry Illuminated...
pdf ⟹ 561 pages.

Pág. 493(510) - Chapter 44
The six Hyperbolic Trigonometric Functions
⟹ ajustar para **M.U.S.**

✡ **Obs**: Alterar as coordenadas cartesianas para as coordenadas Trunchiaras.

d) Implicit Curves and Surfaces.pdf
⟹ 339 pages.

Pág. 40(46) - 1.8 Final Remarks

$K \leq n$ (n-dimensional manifolds)

... "Only a diffeomorphism (i.e a ~~⟹~~ smooth mapping with smooth inverse) ensures the smoothness of a parametric curve or surface.

Thus, the smoothness of a submanifold depends more on the properties of the mapping used to define it than on its associated geometric invariants (e.g. curvature and torsion).

e) Improper-Riemann-Integrals.pdf
→ 464 pages.

Pag. 22(41) - Figure 1.3

Astroid $x^{\frac{2}{3}} + y^{\frac{2}{3}} = 1$

Obs: Recalcular com base em novos parâmetros

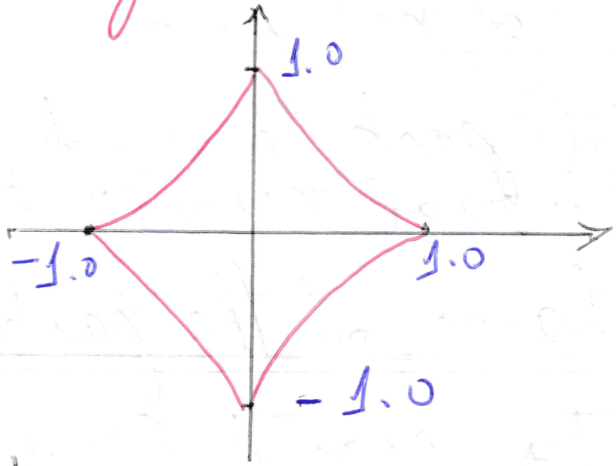

Obs.: Neste caso, este astroid será menor que 1 ⇒

$$\boxed{x^{\frac{2}{3}} + y^{\frac{2}{3}} < 1}$$

f1 The Geometry of Minkowski Spacetime.pdf → 335 pages.

Pag. 42(58) - 1.4
Timelike Vectors and Curves

Pag. 93(109) - Chapter 2
Skew-Symmetric Linear Transformations and Electromagnetic Fields.

2.1. Motivation via the Lorentz Law

A charged particle in M is a triple (α, m, e), where
(α, m) is a
material particle and e is a nonzero real number called the
charge of the particle.

Lorentz World Force Law (or Lorentz 4-Force Law) which expresses the rate at which the particle's world momentum changes at each point on the worldline as a linear function of the particle's World velocity:

$$\frac{dP}{d\tau} = eFU, \qquad (2.1.1).$$

where

$U = U(\tau)$ is the particle's world velocity,

$P = mU$ its world momentum

and,

at each point, $F : M \to M$ is a linear transformation defined in terms of the classical

"eletric and magnetic 3-vectors $\vec{E}$ and $\vec{B}$"

at that point (2.1.1) is an abbreviated version of the somewhat (um pouco) more accurate and considerably more cumbersome (pesado)

$$\frac{dP(\tau)}{d\tau} = e\,F_{\alpha(\tau)}(U(\tau)),$$

where

$F_{\alpha(\tau)}$ is the appropriate linear transformation at $\alpha(\tau) \in M$.

Pág. 279 (293) — Appendix A

Topologies for M.

A.1. the Euclidean Topology.

Pag. 280(294) A.2

E - Continuous Timelike curves

g) the geometry of multiple images - the laws that govern the formation.pdf
→ 659 pages.

Pag. 292(313) — 5.4
Parallel projection

Pag. 396(414) — 7.4.3.
Euclidean parallax.

h) the Geometry of Spacetime - 2000.pdf (Callahan)
→ 480 pages

Pag. 108(122) - 3.2
Curves and Curvature

Parametrized curves and Arc Length.

Pag. 114(128) - Definition 3.2
→

Let $x : [a,b] \to \mathbb{R}^2$

be a parametrization of a curve C.
then the arc-length function on C is

$$s(q) = \int_a^q \|x'(q)\| \, dq.$$

... the differential ds is therefore

$$\boxed{ds = \left(\left(\frac{dx}{dq}\right)^2 + \left(\frac{dy}{dq}\right)^2\right)^{1/2}}$$

$$\boxed{dg = \sqrt{dx^2 + dy^2}}$$

Obs: A distância será alterada em virtude da inclusão de novos parâmetros referentes à curvatura do espaço. As coordenadas serão alteradas para $\boxed{M.U.S.}$

Pág. 116 (130) — ellipse

Let us the arc-length parametrization of the ellipse

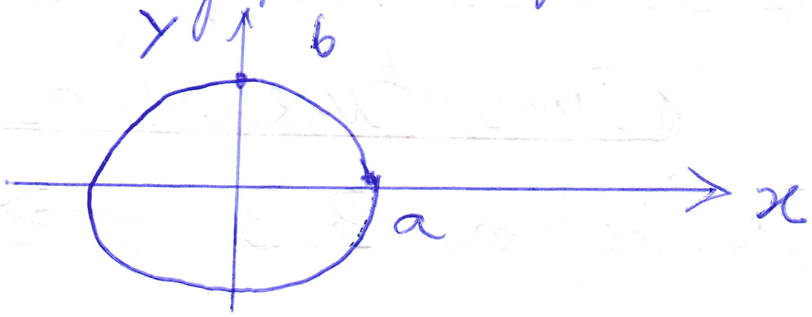

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1, \quad 0<b<a$$

$$\therefore K=\frac{\sqrt{a^2-b^2}}{a} \quad \text{and}$$

$$0<K=1$$

(ellipses have different proportions, and K is an index, or modulus, of shape.

If $\boxed{K=0, \text{ then } a=b,}$ so the ellipse would be a circle.

As K increases, the ellipse deviates more and more from a circular shape; for this reason K is called the

eccentricity of the ellipse).

Pág. 117 (131) —

Curvature of a Curve

Definition 3.3 →

Let $y(s)$ be the arc-length parametrization of the curve C. Then we define:

unit tangent vector along C : $u(s) = \frac{dy}{ds}$;

curvature vector along C : $K(s) = \frac{du}{ds} = \frac{d^2y}{ds^2}$;

curvature at s: $\kappa(s) = \| K(s) \|$.

Important:

Pág. 120 (134) - Higher Dimensions

Obs: Os cálculos também serão alterados.

2) The Geometry of Spherically Symmetric Finsler Manifolds. pdf → 161 pages.

Pág. 81 (89) → Chapter 6

Spherically Symmetric Metrics of Scalar Curvature

⇒ Finsler metrics of scalar curvature and dimension $n \geq 3$ are the natural extension of Riemannian metrics of constant sectional curvature.

Pág. 93 (100) - Chapter 7

Spherically Symmetric Metrics of Constant Flag Curvature

7.1. Projective Spherically Symmetric Metric of Constant Flag Curvature

$$F(x,y) = \frac{\sqrt{c|y|^2 - (|x|^2|y|^2 - \langle x,y\rangle^2)}}{c - |x|^2} \quad (7.1).$$

is projectively flat on $\mathbb{B}^n(r_{\sqrt{c}})$ with

constant sectional curvature

-1 where $\mathbb{B}^n(r_{\sqrt{c}})$ is an n-dimensional ball with its radius $\sqrt{c}$.

---

i) the geometry of the Generalized Gauss-Map F (1981).pdf

→ Gauss - "disquisitiones generales circa superficies curvas"

Livro: 150 years after Gauss

"disquisitiones generales circa superficies curvas" – P. Dombrowski – 1979.

j) The geometry of total curvature on complete open surfaces.pdf
→ 296 pages.

Pág. 08(20) – 1.3
The Riemannian curvature tensor

Newton's first law states that the motion of a particle is in a straight line at constant speed if no force acts upon it.

Geodesics on a Riemannian manifold are understood to be subject to Newton's first law: it is considered that the straightness of a geodesic is equivalent to the existence of a parallel velocity vector field along it.

Lemma 1.3.1. Let $z(t) = \xi^i(t) X_i \circ \gamma(t)$ be a vector field along a curve γ, where $\gamma(t) = (x^1(t), \ldots, x^n(t))$ in a coordinate neighborhood U. Then the map

$$t \mapsto \left(\frac{d\xi^i}{dt} + \Gamma^i_{kj} \xi^k \frac{dx^j}{dt} \right) X_i \circ \gamma(t)$$

is independent of the choice of local coordinates, and hence is a vector field along γ.

Pág. 75 (87) - Chapter 3
The ideal boundary

3.1. The curvature at infinity

Let c : I → M be a piecewise-smooth in a two-dimensional Riemannian manifold M, where I ⊂ ℝ is an interval.

Pág. 78 (90) - 3.2
Parallelism and pseudo-distance between curves.

k) The Geometry of Uncertainty. pdf → 864 pages.

Pág. 529 (545) - 13.2.1
Mass space representation

Just as belief functions can be represented as vectors of a sufficiently large Cartesian Space (see Chapter 7),

↳ Chapter 7
Pag. 323(343)
the geometry of

belief functions

... they can be associated with the related vector of mass values

$$m = \sum_{\emptyset \subsetneq A \subseteq \Theta} m(A) m_A \in \mathbb{R}^{N-1}, \quad (13.10)$$

where m_A is the vector of mass values of the categorical belief function

$Bel_A : m_A(A) = 1, \ m_A(B) = 0 \ \forall B \neq A.$

Obs: Pag. 331(351) - 7.2

Simplicial form of the belief space.

Theorem 9. Every vector representation **bel** of a belief function $Bel \in \mathcal{B}$ can be uniquely expressed as a convex combination of the vectors bel_A, $\emptyset \subsetneq A \subsetneq \Theta$, representing the categorical belief functions on Θ:

$$bel = \sum_{\emptyset \subsetneq A \subsetneq \Theta} m(A)\, bel_A, \quad (7.8)$$

with coefficients given by the basic probability assignment m of Bel.

Pag. 530(546) → Example 41: mass space in the binary case.

Fig. 13.5

↳ the mass space $\mathcal{M}_2$ for a binary frame is a triangle in $\mathbb{R}^2$ whose vertices are the mass vectors associated with the categorical belief functions focused on

$\{x\}, \{y\}$ and Θ: m_x, m_y, m_Θ.

the belief space $\mathcal{B}_2$ coincides with $\mathcal{M}_2$ when

$\Theta = \Theta_2 = \{x, y\}$.

2) Data-NovoVivo-010-2023|
Referencia-index|
Heath - Geometry|

a) Heath - A history of Greek mathematics Vol. II. (1921).pdf
→ 460 pages.

Pág. 125(139) - the figure shows a circle with 1366 $\frac{1}{3}$ as the content the heap written within it, and with 12 and 8 written above and to the left of the circle respectively.

Pág. 166(180) - parallel lines, which they used for the purpose of establishing by a general proof the proposition that the sum of the three angles of any triangle is equal to two right angles.

IMPORTANT:

Pág. 191(205) - We resume Eudemus's account.

Pág. 193(207) -

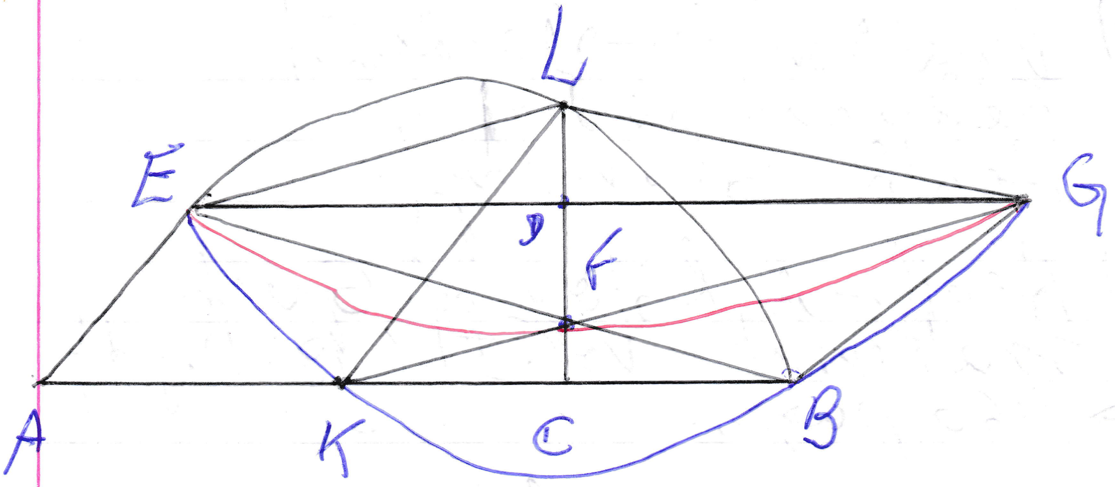

IMPORTANT:

A linha vermelha (concava)
→ negativa, apresenta uma
curvatura.
Calcular esta curvatura.

A case too where the outer (exterior) circumference is less than a semicircle was solved by Hippocrates, who gave the following preliminary construction.

Let there be a circle with diameter AB, and let its centre be K.

Let CD bisect BK at right angles; and let the straight line EF be so placed between CD and the circumference that it verges towards B [i.e. will, if produced, pass through B], while its length is also such that the square on it is $1\frac{1}{2}$ times the square on (one of) the radii.

Let EG be drawn parallel to AB, and let (straight lines) be drawn joining K to E and F.

Let the straight line [KF] joined to F and produced meet EG in G, and again let (straight lines) be drawn joining B to F, G.

It is then manifest that BF produced will pass through ["fall on"] E [for by hypothesis EF verges towards B] and BG will be equal to EK.

Interessante:

Simplicius proves this at length. the proof is easy.

The triangles FKC, FBC are equal in all ~~as~~ respects [Eucl. I. 4]. therefore, EG being parallel to KB, the triangles EDF, GDF are equal in all respects [Eucl. I. 15, 29, 26]. (right triangles). Hence the trapezium is isosceles, and BG = EK.

this being so I say that the trapezium EKBG can be comprehended in a circle.

obs: Estudar o restante!!!

Pag. 226(240) — a)

The Quadratrix of Hippias

Pag. 227(241) →

$$\frac{\rho \sin\phi}{a} = \frac{\phi}{\frac{1}{2}\pi}$$

IMPORTANT:

Now clearly, when the curve is once constructed, it enables us not only to TRISECT THE ANGLE EAD but also to divide it in any given ratio.

Obs: Estudar com cuidado!!!

Pag. 232(246) — (δ)

Approximations to the value of π

211875 : 67441 (= 3.141635)

197888 : 62351 (= 3.17377)

Pág. 235 (249) —

the Trisection of any Angle.

Pag. 254 (268) - Figure

... point P on the curve parallel to one asymptote and meeting the other is equal to the rectangle AO. OB.

Pag. 259 (273) - PARALLELS
(Eratosthenes)

IMPORTANT:

Termo procurado:

parallel → 215 ocorrências

Obs: Continuar depois!!!

b) Heath - A history of Greek mathematics ... (1921). pdf
→ 596 pages
Search: parallel → 439 ocorrências

c) Heath - Euclid's Elements -
(2002).pdf
→ 692 pages

d) Heath - the works of Archimedes -
(1897).pdf
→ 512 pages.

e) Heath - thirteen Books of
Euclid's Elements.
Books I - II - Volume I...pdf
Search:
parallel → 583 ocorrências

f) Heath - thirteen Books of
Euclid's Elements.
Books III - IX - Volume II.pdf
Search:
parallel → 352 ocorrências.

g) Heath - thirteen Books of
Euclid's Elements.
Books X - XIII.
Volume 3.pdf
Search:
parallel → 695 ocorrências

obs: Volume 1 → 434 pages,
Volume 2 → 445 pages,
Volume 3 → 558 pages.

20231031180000

20231101132500|||
1) Data-NovoVinio-010-2023|
Referencia-index|
Metric-space|

a) 20231101 - Mlodinow - Euclid's Window - The Story of Geometry from Parallel Lines to Hyperspace (2001).pdf → 320 pages.

Pág. 93 (106) → Chapter III
The Story of Gauss

Can parallel lines intersect in space?
Napoleon's favorite prodigy presents Euclid with his Waterloo.
The greatest revolution in geometry since the Greeks begins.

13. The Curved Space Revolution

Pág. 97 (109) -

Parallel postulate, caused consternation in almost every scholar who studied Euclid because it was not as simple and intuitive as Euclid's other assumptions. Recall its wording:

Interessante.

Given a line segment that crosses two lines in a way that the sum of inner angles on the same side is less than a right angle, then two lines will eventually meet (on that side of the line segment).

IMPORTANT: Livro todo!!!

6) 20231101 - Withers - Euclid's Parallel Postulate ... (1905).pdf

Pag. 37(47) - formula

$$ds^2 = E\,du^2 + 2F\,du . dv + G\,dv^2,$$

in which E, F, and G are functions of u, v.

But u, v may be lengths of lines on the surface or angles between geodesic lines and therefore we do not need to go outside the surface itself for their analytical determination.

Pag. 79(89) - Chapter IV.

~~Phy~~ Psychological sources of the parallel postulate and its closely allied conception.

Pág. 127(137) - Chapter V.
The nature and validity of the parallel postulate.

IMPORTANT: Livro todo!!!

c) Abstract-Algebra-Squaring-Circle.pdf
→ 233 pages.

Pág. 65(84) - Chapter 5
Straightedge and Compass Constructions.

We conclude the chapter by showing that there is a connection between geometric constructions and fields of the form $F(\sqrt{\gamma})$.

Pág. 78(97) - Fig 5.17
Constructing square roots

$$\frac{x}{1} = \frac{a}{x} \Rightarrow x = \sqrt{a}$$

Pág. 82(101) - 5.3.2
Doubling the Cube
5.3.3
Squaring the Circle
5.3.4
Trisecting an Angle

Pág. 84 (103) – 5.4
Constructible Numbers and Fields

..., we can construct points P_i and P_j whose distance apart is $\boxed{\sqrt[3]{2}}$ using only operations of types (ii) and (iii).

IMPORTANT:

5.4.1. Definition

... Clearly, doubling the cube is possible if and only if the number $\boxed{\sqrt[3]{2}}$ is constructible, while squaring the circle is possible if and only if $\boxed{\sqrt{\pi}}$ is constructible.

IMPORTANT: Ler este capítulo com atenção.
→ Atentar ao livro todo!!!

Pág. 145 (163) – Chapter 10
Transcendence of e and π.
→

Capítulo interessantíssimo!!!

d) A-Circle-Free Study of Mathematical Analysis.pdf

→ 480 pages!

Pág. 321(333) - theorem 14.23
(Parallelogram Identity)

If H is an inner product space, and if $u \in H$, $v \in H$, then

$$\left\|\frac{u+v}{2}\right\|^2 + \left\|\frac{u-v}{2}\right\|^2 = \frac{\|u\|^2 + \|v\|^2}{2}$$

Definition 14.19 (Hilbert Spaces)

An inner product H is a Hilbert Space if and only if it is a complete metric space with respect to the distance

$$d(u,v) = \|u - v\|$$

associated to the norm (14.2).

$\Downarrow$

Pág. 320(332) - Definition 14.18
(Norms in Inner Product Spaces).

obs: $\|u\| = \sqrt{\langle u,u \rangle}$

The norm induced by the inner product of a space H is defined by $\|u\| = \sqrt{\langle u,u \rangle}$
(14.2)

e) Calculus-and-Linear-Algebra-in-Recipes.pdf
→ 1015 pages.

Pág. 67(87) - Chapter 8

Complex Numbers: Polar Coordinates

Pág. 153(169) - Chapter 16

Orthogonality I

Pág. 165(181) - Chapter 17

Orthogonality II

Pág. 429(435) - Chapter 39
Diagonalization: Eigenvalues and Eigenvectors.

IMPORTANT: Livro todo!!!

f) Casual Calculus A Friendly Student Companion Volume.3.pdf
→ 603 pages.

Pág. 127(140) - Chapter 15

And Can We Agree Rectangular Coordinates are Rather dull?

⇒ IMPORTANT: Capítulo todo. Estudar atentamente !!! ⇐ ⊛

Obs: Todos estes cálculos serão alterados com a nova matemática →

M.U.S.

Pág. 413(426) - Appendix A

Conics, Quadrics, and Parametrics.

Conic section
$$ax^2+by^2+cx+dy+e=0 \quad (A.1)$$

Pág. 415(428) - Fig A.2

The (directed) parametric curve
$x = 4\cos\theta$, $y = \frac{3}{2}\sin\theta$.

Obs: A curvatura irá alterar toda a descrição matemática desta ellipse.

Pág. 421(434) - Fig A.6

Sample ellipsoid with

$a = 1$, $b = 1.25$, $c = 1.5$.

Obs: Está muito próximo da ellipse com curvatura, mas este sem curvatura.

Important: São 3 volumes, são bons para fundamentação.

g) Fixed Point Theory in Generalized Metric Space.pdf
→ 141 pages

Pág. 5(15) - Chapter 2
Metric Space

Pág. 15(25) - Chapter 3
Metric Fixed Point Theory

Obs: the pioneer metric fixed point theorem that given by Banach.
↓
"Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales.
Fundam. Math. 3
133-181 (1922)

h) Geometric Mechanics and its Applications.pdf
→ 540 pages.
Obs: Search: Gauss (69 ocurrs)

i) Geometric-Function-theory-in-Higher-Dimension.pdf

→ 185 pages.

Pag. 150(155) – 4
Geometry of Complete Hyperbolic Manifolds.

Pag. 151(156) - 4.1.
Complex and Metric Geometries

Obs: Ver:
geodesics for the Kobayashi distance.

... Finally, a geodesic metric space is

GROMOV HYPERBOLIC

if it is δ-thin for some δ > 0.

The unit ball is Gromov hyperbolic whereas (enquanto que) the bidisc is not.

Pag. 152(157) - 4.2
Topology of Complete Hyperbolic Manifolds.

j) Mathematics of Particle-Wave Mechanical Systems.pdf
→ 388 pages.

Pag. 16(28) – 2.2

Lorentz Transformations.

Pág. 18(30) - 2.3
Einstein Addition of
Velocities Law.

Obs: Interessante!!!
Livro todo!!!

K) Physics-Quantum-Rings.pdf
→ 600 pages

Pág. 519(535) → 16.6
Curvature Induced Potential

Obs: Interessantíssimo!!!

L) Squigonometry - the Study
of Imperfect Circles.pdf
→ 292 pages

Pág. 173(183) - 16.2
The lemniscate constant,
Gauss and Pi.

Obs: Está no capítulo 16 →
Elliptic integrals

2) Data-NovoVivo-010-2023/
Referencia-index/

Curvature - 02/

a) Ballmann - Lectures on spaces of nonpositive curvature-(2004).pdf

→ 116 pages.

Pag. 01(05) -

Modifying the above definition by comparing with triangles in the model surface M^2_K of constant GAUSS curvature K instead of $\mathbb{R}^2 = M^2_0$, one obtains the definition of Alexandrov curvature $K_X \leq K$.

Pag. 06(10)...

where the Gauss curvature of the surface is negative.

Observe that R does have full measure if the metric on M is analytic.

Pag. 67(71) - ...

K is the Gauss curvature of M. Use this to discuss geodesics in the unit tangent bundles of surfaces of constant Gauss curvature.

2. Jacobi fields and curvature.

b) Metric Spaces, Convexity and Nonpositive Curvature, Vol. 6 (2004). pdf
→ 299 pages.

Pág. 03(15).
... paper [62]
↳ Hadamard
Les surfaces à courbures opposées et leurs lignes géodesiques – 1898.
by Hadamard is that of a differentiable surface S embedded in $\mathbb{R}^3$, such that at each point of S, the Gaussian curvature is negative expect for a finite set of points where it is zero.

Obs: Livro todo!!!

c) Differential Geometry and its Visualization. pdf

Pág. 01(28) – Chapter 1
Curves in Three-Dimensional Euclidean Space.

Obs: 493 pages – Livro todo!!!

Pág. 245(272) - Chapter 3

The Intrinsic Geometry of Surfaces.

Pág. 247(274) - Figure 3.1

The relation between the curvature and geodesic curvature of a curve on a surface (Theorem 3.1.1).

Pág. 246(273)

Theorem 3.1.1

The geodesic curvature of a curve on a surface is equal to its curvature multiplied by the cosine of the angle between the binormal vector of the curve and the normal vector of the surface, that is, the cosine of the angle between the osculating plane of the curve and the tangent plane of the surface.

__Obs__: Estudar com profundidade!!!

Pág. 373(400) - Chapter 4

Tensor Algebra and Riemannian Geometry.

3) Data-NovoVivo-010-2023/
Referencia-index/
Distance/

a) Blumenthal - Theory and Application of Distance Geometry-(1970).pdf
→ 357 pages.

Pag. 90(100) - Part II
Euclidean and Hilbert Spaces
Chapter IV
Congruent Imbedding in Euclidean Space

Pag. 162(172) - Part III
The Non-Euclidean Spaces
Chapter VII
Imbedding and characterization theorems for spherical space characterization of pseudo-spherical sets.

b) Distance geometry -- theory, methods, and applications - Springer (2013).pdf → 427 pages.

Pag. 85(99) - Chapter 5

Is the Distance Geometry Problem in NP?
→ positive integer K the distance geometry problem (DGP)... LIVRO TODO!!!

20231101 174200

20231102120000m

Data-NovoVixo-010-2023/

Referencia-Index/

Geometry-important-01/

a) 20231102- A Course in Metric Geometry-2001.pdf

Pág.07(16)- 1.2.3. Spheres

Example 1.2.25.- The n-spheres S^n is the set of unit vectors in R^{n+1}, i.e., $S^n = \{x \in \mathbb{R}^{n+1} : |x| = 1\}$.

The angular metric on S^n is defined by $d(x,y) = \arccos \langle x, y \rangle$.

In other words, the spherical distance is defined as the EUCLIDEAN ANGLE between unit vectors.

It equals the length of the shorter arc of a great circle connecting x and y in the sphere.

Another formula for this metric is

$$d(x,y) = 2 \arcsin \frac{|x-y|}{2}.$$

The metric on the circle described in Example 1.2.9 is a partial case of this example.

Pág.04(13)- Example 1.2.9. →

The unit circle, S^1, is the set of points in the plane lying at distance 1 from the origin:

Being a subset of the plane, the circle carries the restricted Euclidean metric on it. We define an alternative metric by setting the distance between two points as the length of the shorter arc between them.

For example, the arc-length distance between two opposite points of the circle is equal to $\boxed{\pi .}$ → wrong!!!

Obs: Metric is wrong too!!!

The distance between adjacent vertices of a regular n-gon (inscribed into the circle is equal

$$\boxed{\frac{2\pi}{n}} \rightarrow$$

Interessante: Exercise 1.2.10.

(a) Prove that any arc of length less or equal to π, equipped with the above metric, is isometric to a straight line segment.

(b) Prove that the entire circle with this metric is not isometric to any subset of the plane (regarded with the restriction of Euclidean distance onto this subset).

Pág. 03(12) – Example 1.2.3.

and Example 1.2.4.

$$|(x_1, y_1).(x_2, y_2)| = \sqrt{|x_1 x_2|^2 + (y_1 y_2)^2}$$

In particular, $R \times R = R^2$.

Obs: Estudar Chapter I - Metric Spaces
Chapter II - Length Spaces
Chapter IV - Spaces of Bounded Curvature
Chapter X - Spaces of Curvature Bounded Below.

b) 20231102-Angular momentum.pdf
(in Quantum Physics)
→ Biedenharn.

c) 20231102- Geometry - A Metric Approach with Models - 1990.pdf
→ 588 pages. **Obs**: Ler Livro todo!!!

Pág. 169 (184) - Chapter 7
The Theory of Parallels
Obs: Estudar com profundidade!!!

Pág. 196 (211) - Chapter 8
Hyperbolic Geometry

Interessante:

Pág. 205 (220) - 8.2
Angle Sum and the Defect of a Triangle

Definition:

Let $\triangle ABC$ be a triangle in a protractor (transferidor) geometry. The defect of $\triangle ABC$ is

$$\boxed{\delta(\triangle ABC) = 180 - (m(\angle A) + m(\angle B) + m(\angle C)).}$$

We already know that for a EUCLIDEAN GEOMETRY

$\boxed{\delta(\triangle ABC) = 0}$

for all triangles $\boxed{\text{(Theorem 7.2.9)}}$

⇝ será estudado mais tarde.

We have seen examples in which

$$\boxed{\delta(\triangle ABC) < 0}$$

(the Moulton Plane)

and in which

$$\boxed{\delta(\triangle ABC) > 0}$$

(the Poincaré Plane)

We also know that for a neutral geometry

$$\boxed{\delta(\triangle ABC) \geq 0}$$

(Theorem 7.2.8).

IMPORTANT: GAUSS recognized that in order to prove EPP is satisfied it was sufficient to prove that $\boxed{\delta(\triangle ABC) = 0}$ for one triangle.

obs: EPP - Euclidean Parallel Property

Pág. 175(190) - See Figure 7-8.

Theorem 7.1.6.

If l is a line and $P \notin l$ in a neutral geometry which satisfies EFP (Euclid's Fifth Postulate), then there exists a unique line l' through P which is parallel to l.

Figure 7-8

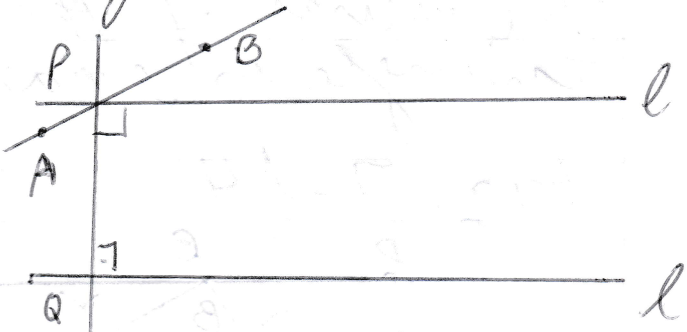

obs: Since the 18th century the conclusion of theorem 7.1.6. has been used as the primary parallel axiom.

Because this choice was first championed by Playfair, it is often referred to as Playfair's Parallel Postulate.

Pág. 181(196) - theorem 7.2.7.

In a neutral geometry the sum of the measures of the acute angles of a right triangle is less than or equal to 90.

Theorem 7.2.8 (Saccheri's Theorem)

In a neutral geometry,
the sum of the measures of the
angles of a triangle is less
than or equal to 180.

Pág. 182(197) -
Theorem 7.2.9
(Euclidean Angle Sum)
In a neutral geometry which
satisfies EPP, the sum of the
measures of the angles of any
triangle is exactly 180.

Fig. 7-14

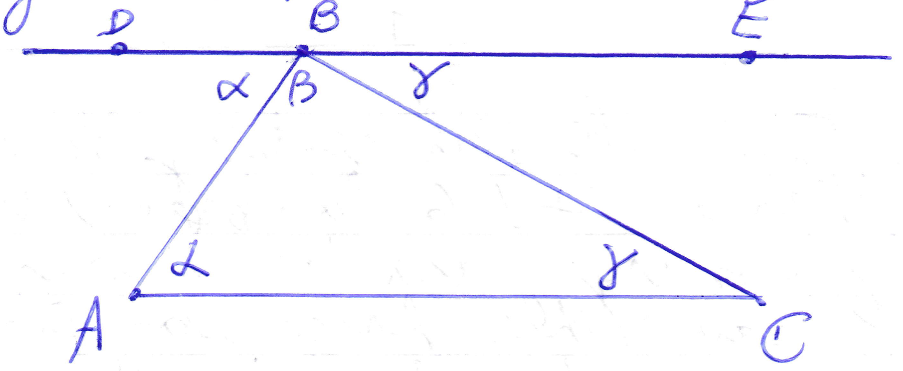

Pág. 214(229) - 8.3
The Distance between Parallel Lines

Obs: Ler livro todo!!!

d) 20231102 - Lengths of Subsets
in Coxeter Groups-1.pdf

e) 20231102 - Metric and Differential
Geometry-2012.pdf

→ 409 pages.

Pág. 119 (163) – Existence of Einstein Metrics on Fano Manifolds

... Kähler-Einstein metrics with negative scalar curvature ...

f) Analysis in Vector Spaces – A Course in Advanced Calculus. pdf → 479 pages.

Pág. 11 (25) – Examples 1.2.3 Lines and half-planes in $\mathbb{R}^2$.

See:

Figure 1.2. Region in Example 1.2.5.

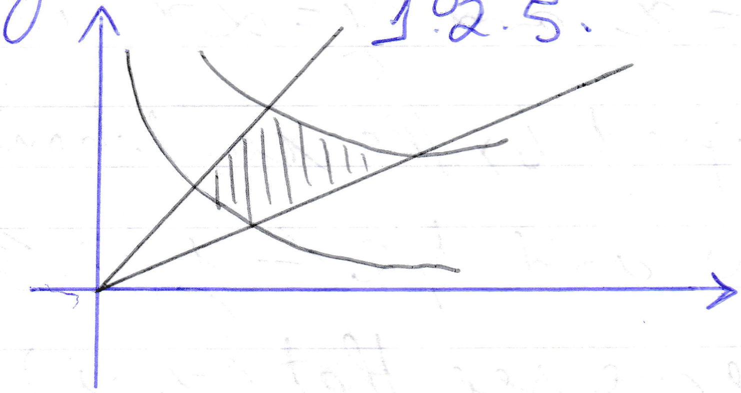

Obs:

Example 1.2.5

Let R be the region in the xy-plane that lies in the first quadrant (that is, $x \geqslant 0$ and $y \geqslant 0$) and is between the hyperbolas $xy = 1$, $xy = 2$, and bounded by the lines $2y = x$, $y = 2x$

(see Figure 1.2),
Specify R by a set of inequalities.

Solution: the region in the first quadrant that is between the hyperbolas $xy=1$ and $xy=2$ is specified by the conditions that

$$x>0 \text{ and } 1 \le xy \le 2.$$

Similarly the region in the Region in the first quadrant that is between the lines

$2y=x$ and $y=2x$ is

specified by the conditions that

$$x>0 \text{ and } 1/2 < y/x \le 1.$$

Hence we see that $(x,y) \in R$ if and only if

$$1 \le xy \le 2 \text{ and } 1/2 \le y/x \le 2$$

and $x>0$.

---

g) Analysis-of-the-Gravity-Field.pdf

⇒ 542 pages.

Pág. 10 (23) - Fig 1.3
The spherical coordinate system
(Wrong!!!
⇒ Calcular com a esfera curva)

Pág. 21 (34) - Theorem 1.2
(Poisson)

The gravitational potential u generated by a 3D mass density $\rho(P)$ satisfies the following equation, called the Poisson equation

$$\Delta u = -4\pi G \rho(P) \qquad (1.73)$$

Pág. 167 (176) -
Interessante

$e = \sqrt{1 - \frac{b^2}{a^2}}$ is the first eccentricity,

$e' = \sqrt{\frac{a^2}{b^2} - 1}$ is the second eccentricity,

b is the polar radius (or semi-minor axis) and finally X, Y and Z are the geocentric Cartesian coordinates of P. Passing to the geocentric system, we can easily see that the radial coordinate is given by

$$r = \sqrt{x^2+y^2+z^2} =$$

$$\sqrt{[(N+h)\cos\varphi_e]^2+[(N(1-e')+h\sin\varphi_e]^2}$$

obs: Livro todo!!!

h) Analytic Hyperbolic Geometry an - Abraham Albert Ungar.pdf
→ 775 pag.
obs: Interessante livro!!!

Pág. 254 (277) →

The Gaussian curvature K of this surface is given by the equation ...

Definitión 7.3
(Gyrocurvature and Cogyrocurvature)

The Gaussian curvature of a gyrovector space with its gyroline (cogyroline) element is called a gyrocurvature (cogyrocurvature).

Pág. 256 (279) –

Möbius gyrovector plane

is the Gaussian curvature K of the surface with the line element (7.26). It is a negative constant,

$$\boxed{K = -\frac{4}{c^2}} \qquad (7.28)$$

as one can calculate from (7.21)

Extending (7.26) from $n=2$ to $n \geq 2$ we have

$$ds_M^2 = \frac{c^4}{(c^2-r^2)^2}\, dr^2$$

$$= \frac{dr^2}{(1+\frac{1}{4}Kr^2)^2} . \qquad (7.29)$$

As expected, the Riemannian line element ds_M^2 reduces to its Euclidean counterpart in the limit of large c,

$$\lim_{c \to \infty} ds_M^2 = dr^2 \qquad (7.30)$$

Pág. 258(281) - Fig. 7.1. $0 \leq \theta < 2\pi$.

(Wrong!!!)

Obs: Aplicar a curvatura (M.U.S.)

i) Analytical Geometry - 2D and 3D...pdf
→ 890 pages.

Pag. 334 - Chapter 7 Ellipse

Pag. 337 - 7.3 Focal Distance

The sum of the focal distances of any point on the ellipse is equal to the length of the major axis.

Pag. 338 - 7.5

Auxiliary Circle

The circle described on the major axis as diameter is called the auxiliary circle.

Obs: Livro bem elucidativo!!!

j) Analytical Mechanics - Classical, Lagrangian and Hamiltonian Mechanics, Stability theory, Special Relativity. pdf
→ 848 pages.

Search: curve (364 ocorrências)

Pag. 13 (38) - rectifiable curve

Pag. 355(376) - Chapter 10
Mathematical Introduction to Special Relativity and the Relativistic Lagrangian Formulation

Pág. 370(391) - 10.3
Introduction to Relativistic Dynamics

10.3.1. Mass, Four-Momentum and their Elementary Properties.

→ notion of impulse generalises to the four-momentum.

Pág. 374(395) -
The photon is an example of a massless relativistic particle with an energy, whose value depends on the reference frame.

Minkowski coordinates:

$$p^0 = \hbar \frac{\omega}{c}, \quad p^a = \hbar k^a, \quad a = 1, 2, 3$$

where $\omega > 0$ is the angular frequency of the particle's light wave and k^a are the three components of the wave vector in the frame.

Above $\boxed{k = \frac{h}{2\pi}}$ → wrong !!!,

Where h is Planck's constant (6.626 x 10^{-27} erg sec).

As

$$\sum_{a=1}^{3} (k^a)^2 = \left(\frac{\omega}{c}\right)^2$$

from electromagnetism, in this case we must have $g(P, P) = 0$ and so the photon is massless and its world lines are lightlike.

Clearly there is no Minkowski frame at rest with a photon, because $k^a = 0$ for $a = 1, 2, 3$ would force $\omega = 0$ and the four-momentum would not be casual.

Sometimes one reads that photons are "pure energy".

Pag. 800 (817) →

…; the conservative forces of gravity and the spring's, with total potential energy:

$+ mgL(2\sin\phi + \sin\psi) + KL^2(1+\cos(\phi-\psi));$

the Coriolis Force,

which does no work as perpendicular to the velocity vectors;

IMPORTANT:

Analisar o motivo da CORIOLIS FORCE não agir de forma perpendicular à velocidade dos vetores. Terá alguma relação com a curvatura do espaço?

Interessante:

Essa relação pode estar relacionada diretamente com a curvatura do espaço.

Obs: Livros todos!!!

j) Analytical Mechanics for Relativity and Quantum Mechanics.pdf

→ 618 pages.

Pág. 296 (317) ↱ Hamilton

Thus, unlike $H(q, \dot{q}) \equiv 0$, the extended Hamiltonian $K(q, p)$ is not identically zero. It has nonzero partial derivates and is set equal to zero at the end of the calculation as a consequence of the Hamilton equations of motion.(56)

(56). In Dirac (1964), the same distinction is made
the H is said to be
strongly zero,

while K is said to be weakly zero.

12.11 - Dirac's Theory of Phase-Space Constraints.

Dirac (1964) addressed this problem by creating what he called a "generalized Hamiltonian formalism" with modified Hamilton equations, and phase-space constraints in the form

$\phi_m(q, p) = 0.$

Ⓐ IMPORTANT:

Dirac's formalism is intended to include complex problems such as (tal como) Hamiltonian theory in the

curved spaces of general relativity.

Pág. 399 (420) — 16.17

⇒

16.17 Covariant Serret-Frenet Theory
The derivation of the Serret-Frenet formulae for a curve in three dimensions, given in Section A.12, generalizes easily to world lines considered as curves in MINKOWSKI space.

The three orthonormal Serret-Frenet vectors of Section A.12 generalize to four orthonormal (in the sense of the Minkowski metric) Serret-Frenet fourvectors defined along the world line.

Pág. 505(526) - A.12
The Serret-Frenet Formulae

K) Andersen - the geometry of an art.pdf
→ 836 pages.

Pág. 671(699) - Reflection in Curved Surfaces
Pág. 674(702) - XII.9
Parallel Projections
→ orthogonal projections.

Pág. 731(756) - Appendix Two

The Appearance of a Rectangle
à la Leonardo da Vinci.

$$y=\sqrt{a^2+x^2}.$$

$$\sigma(y,a)=\sigma(\sqrt{a^2+x^2},a)=a\sqrt{a^2+x^2}.$$

Pag. 732(757) –

The Curves for Three ⇙
Different Distances

Pag. 733(758) – Figure 4.

The curvilinear appearance
for $a=1$.

⇒ Pag. 654(682) – Figure XII.15

Lambert's introduction of a
measure point.

Obs: Interessante!!!

L) Anderson – Hyperbolic
Geometry – (2005).pdf
→ 283 pages.

Pag. 73(84) – Chapter 3

Length and Distance in $\mathbb{H}$

→

The Euclidean length of f is given by the integral

$$\text{length}(f) = \int_a^b \sqrt{(x'(t))^2 + (y'(t))^2}\, dt,$$

where

$\sqrt{(x'(t))^2 + (y'(t))^2}\, dt$ is the element of arc-length in $\mathbb{R}^2$.

Pag. 42(53) - the name loxodromic comes from the word loxodrome, which is a curve on the sphere that every lines of latitude at the same angle.

Lines off longitude are loxodromes, but there are also loxodromes that spiral into both poles. The reason these Möbius transformation are called loxodromic is that each one keeps invariant a loxodrome.

Pag. 173(183) - Figure 5.11

Decomposing a hyperbolic pentagon into hyperbolic triangles.

Let u_k be the interior angle of T_k at x, and note that

$$\sum_{k=1}^{n} \mu_k = 2\pi. \quad (\text{wrong!!!})$$

Obs: Todos os cálculos devem ser ajustados à curvatura do espaço.

Pag. 181 (191) - 5.6 Trigonometry in the Hyperbolic Plane.

1. $\cosh^2(x) - \sinh^2(x) = 1$;
2. $2\cosh(x)\sinh(x) = \sinh(2x)$;
3. $\sinh^2(x) = \frac{1}{2}\cosh(2x) - \frac{1}{2}$;
4. $\cosh^2(x) = \frac{1}{2}\cosh(2x) + \frac{1}{2}$;
5. $\sinh^2(x)\cosh^2(y) + \cosh^2(x)\sinh^2(y) = \frac{1}{2}(\cosh(2x)\cosh(2y) - 1)$.

Obs: Livro todo!!!

M) Anderson - New theory of

the Earth - (2007).pdf

Pág. 43(59) - Fig. 4.7.

there are ten possible configurations of great-circle arcs on a sphere that meet three at a time with angles of 120°...

N) Anderson - the mathematical theory of cosmic strings.pdf

→ 393 pages.

Search: curve (208 ocorrências)

bending (05 ocorrências)

Pág. 163(176) →

limit $v \to 1$ it tends to

$$\left(2+\frac{1}{2}\sqrt{17}\right)m \approx 4.062m.$$

the first of these limits can be understood by referring back to the static equilibrium solutions examined in section 5.2.2.

↳ Pág. 157(170) - 5.2.2. Static Equilibrium solutions.

→ Schwarzschild metric.

There it was seen that for small value of m the bending angle $\Delta\phi$ of the string is of order $m\lambda$.

Pág. 164(177) → ..., the bending angle of a string moving at a low speed v should be of order $\Delta\phi \sim v/c$ (where c is the speed of light); as longitudinal disturbances will move with velocity c whereas (enquanto que) the radial motion of the string is of order v.

Pág. 263(276) - 7.6
The self-force on a massive particle near a straight string

IMPORTANT:

Although (embora) the conical spacetime described by (7.28) is everywhere locally flat and thus free of tidal forces, it turns out that a particle of mass m at rest in the spacetime does experience a gravitational self-force of order m^2 directed towards the singularity at $r = 0$.

⇒ o) Regents Geometry Power Pack.. (2021).pdf
→ 1735 pages. Livro todo!!!

⇒ p) a-new-perpective-on-relativity...pdf
→ 694 pages Livro todo!!! Pág. 348(372) →
Shapiro effect. 20231102183000

20231103100000/
Data-NovoViw-010-2023/
Referencia-index/
Bending-Space-Gravity/

a/20231103-Biedenharn-Angular momentum in quantum physics-(1981).djvu → 747 pages.

b) a-new-perspective-on-relativity-on-odyssey-in-non-geometries.pdf → 694 pages.

Pág. 4(30)- Einstein's retraction of these two postulates and the existence of the aether.

... bending of rays in a gravitational field ...

$$c = c_0\left(1+\frac{\phi}{c^2}\right)$$ where

c_0 = (velocidade) velocity of light at the origin of coordinates;
ϕ = then the velocity of light c at a place with the gravitational potential

Pag. 6(32)- It did not take MAX ABRAHAM [12] long to point this out stating that Einstein had given "the death blow

to relativity", by retracting the invariance of c.

Abraham said he warned "repeatedly against the siren song of this theory ... [and] that its originator has now convinced himself of its untenability!"

Pag. 149 (175) - 3.7.4.

e/m measurements of the transverse mass

→ deflection methods where one observes the bending of an electron beam in eletric and magnetic fields, ...

Pag. 150 (176) - Fig. 3.6.

Thomson's apparatus for determining the ratio e/m for cathode rays.

... and reaction in the Lorentz force could be avoid for there would result

$$\frac{u}{c} = \frac{E}{H} \,. \qquad (3.7.15)$$

Pag. 151 (177) - 3.7.4.2 ⇒

Kaufmann's method

Kaufmann's experiments show that the real constant mass of the electron is negligible (insignificante) compared with the apparent [electrodynamic] mass; it can be considered as zero, so that if it is mass which constitutes matter we can almost say the matter no longer exists... there are merely holes in the aether.

Poincaré.

Pág. 152(178)

Kaufmann found that e/m varied slightly with velocity, ~~descr~~ decreasing with increasing velocities.

Fig. 3.7. the points on the parabola refer to electrons deflected by parallel and anti-parallel (left side) fields.

IMPORTANTÍSSIMO:

Pág. 153(179) →

Ritz [08] (1911 - Gesammelte Werke - Oeuvres) queried the invariance of Kaufmann's parabola (3.7.17)

(3.7.17) – Kaufmann obtains the equation of a parabola

$$y = \frac{c^2 E}{\kappa H^2}\,\frac{m}{e}\,x^2$$ Pag. 152 (178)

the ratio,

$$\frac{y}{x} = \frac{c}{u}\,\frac{E}{H'} \qquad (3.7.18)$$

is, therefore, a measure of the deviation of the Lorentz force from zero [cf. Eq. (3.7.15) above].

... Kaufmann's parabola, (3.7.17), for it will only have that form if m is independent of the speed.
If the mass did increase with velocity, say, at some power greather than one, the parabola would tend to a straight line as that power increases.

Instead of (3.7.18) (as mine) Ritz tried ⓑ

$$\frac{y}{x} = \frac{\sqrt{(c^2-u^2}}{u}\,\frac{E}{H'} \qquad (3.7.19)$$

giving rise to a parabola,

$$y = \frac{mc^2E}{eKH^2}x^2 + K\frac{eE}{mc^2}, \quad (3.7.20)$$

shifted upward at the origin.

ⓑ Actually, Ritz wrote 2c for c, but that does not change anything "in a theory which considers only relative velocities."

20231103161400

20231104110000" continuação:

b) a-new-perspective-on-relativity...pdf
Pág. 341(367) - 7.1.

Centrifugal versus Gravitational Forces.

General relativity is based on the notion that gravity, rather than being a force acting between masses, is the curvature of space-time itself. The source of the (positive) curvature is mass itself, just like electric charge is the source. And just as free particles follow "straight lines in space-time"

Obs: Do ponto de vista teórico, nós podemos representar matematicamente uma linha reta, mas isto não condiz com a realidade do Universo, pois caso o Universo tenha começado com o Big-Bang, não há possibilidade do Universo ser plano em virtude da energia inicial que por si só interfere na formatação do Universo.

Um aspecto interessante do Universo, ele não pode ser infinito em virtude do próprio Big-Bang. É uma questão de lógica, o finito não pode conter o infinito, pois se ele começou, não é infinito e para que ele exista há a necessidade do espaço, este sim infinito, o nada é infinito e em cima deste nada o Universo pode surgir, se desenvolver e se expandir.

Deus criou o nada para que o Universo pudesse surgir.

Se Deus tivesse preenchido tudo, não haveria a necessidade do surgimento do Universo.

O espaço é algo infinito e que alguns chamam de aether.

O espaço do nada permite a criação de tudo.

Deus é tão magnânimo que permitiu que algo fosse criado para receber o homem, ou seja, o Universo foi criado pela benevolência de Deus para que o homem pudesse viver nele. Quando eu digo o homem, eu me refiro à vida e há infinitas formas de vida no Universo.

Portanto, o Universo existe porque há o espaço do nada que

permite que o Universo se expanda.

Retornando:

"straight lines in space-time"
They follow geodesics in the curved space-time of a gravitational field.

Pag. 342(368)-

We shall return to this lack of equivalence of between gravitational and centrifugal forces which cause acceleration in (Chapter 9), but we already have seen it in action in Sec. 2.2.3

Obs:

2.2.2 Inversion (Pág. 58(84).

Pág. 61(87)- 2.23

Maxwell's "fish-eye": An example of inversion from elliptic geometry.

Maxwell, in 1858, who proved that for an absolute instrument the optical length of any curve in the object space is equal to the optical length of its image, provided both spaces are homogeneous.

Refractive index

$$n(r) = \frac{1}{1+(r/a)^2} n_0 \qquad (2.2.4)$$

Where
r denotes the distance from a fixed point O, and n_0 and a are constants.

Retornando:

... action in Sec. 2.2.3 where Maxwell used this strategy in his classic treatment of inversion in elliptic space.

↳ órbita dos planetas.

IMPORTANTÍSSIMO:

If we take a flat space metric in the plane and consider a constant index of refraction, we will demonstrate that Fermat's principle is capable of yielding the phase of the oscillation of a Bessel function of the first kind in the periodic domain in the asymptotic short-wavelength limit.

This is identical to the WKB result, and it allows us to associate a wave phenomenon with a geodesic trajectory.

The only potential present is the repulsive centrifugal potential, and that keeps the trajectory open.

Pág. 438 (464) - WKB

Expression (9.2.7) is the phase of the phase of the Hermite polynomials in the short-wavelength (WKB) limit. ... Planck's constant

Retornando: Pag. 342 (368) -

In contrast to the generalization of special relativity, where gravity is considered to be the curvature of space-time instead of a bona fide force, the centrifugal force is built into the phase of the Bessel function in the periodic domain where the trajectory consists of straight-line segments that are tangent to a caustic circle, whose radius is determined by the magnitude of the angular momentum

↳ (see book: 20231103 - Biedenharn - Angular Momentum in quantum physics - (1981). djvu).

and the arc segment joining the points of tangency.

Continuação... 2.2.3 (search)

Pag. 449 (475) - 9.5

Fermat's Principle of Least Time and Hyperbolic Geometry.

Pag. 451 (477) - Fig. 9.4

the angle of parallelism between two bounding parallels connected by the geodesic curve γ.

Pag. 559 (585) - We know from Sec. 2.2.3 that stereographic projection is a conformal map of the surface.

Pág. 565 (591) — We know from Sec. 2.2.3 that by a stereographic projection any polarized state can be projected onto a Riemann sphere

See: Pág. 566 (592) —

Fig. 11.9 — Complex plane representation of polarized states.

Pág. 568 (594) — Fig 11.10 Stereographic projection of the complex plane onto the Poincaré sphere. (wrong!!!)

11.3 Poincaré's Representation and Spherical Geometry.

... "whereas expression (11.3.1a) corresponds to the xx-component of the Maxwell stress,

$$\sigma_{xx} = \frac{1}{8\pi}\left(E_y^2 - E_x^2\right),$$

Retornando: Search: bending

Pág. 343 (369) – The bending of light requires a coupling of the gravitational and centripetal potential that is commonly referred to as Schwarzschild's potential [cf. Fig. 7.6]. Here, it arises when we determine the extremum condition for Fermat's principle.

Pág. 448 (474) – The hyperbolic geometry that describes a uniformly rotating disc also describes the bending of light by a massive body when the transition is made from a constant to a non-constant surface of negative curvature.

Pág. 452 (478) – The geodesic still appear as straight lines to the Poincarites, whereas, to us Euclideans, they appear to be bent, and things vary in size depending on where we look. We also see things like the bending of light and the shifting of frequencies in a gravitational field.

Pág. 462 (488) – The bending of the geodesic is attributed to the rotation of the disc [Grøn 04]. But if the stereographic hyperbolic metric were applicable it would invalidate

Einstein's old result! (9.6.16)

$\boxed{9.6.16}$ → Integrating over a period,

$$h = \frac{2\pi r}{\sqrt{(1-r^2/k^2}} > 2\pi r \quad (9.6.16)$$

shows that the length of a hyperbolic circle of radius

sinh(h/k) is greater than that (do que um) of an Euclidean circle having a radius r [= k tanh (h/k). this is none other Einstein's old results!!!

Retornando:

... (9.6.16), and it would also be in conflict with his general theory.

There is no way out for (9.6.28) to hold,

$\boxed{9.6.28}$ → Pag. 461 (487) -

The Euclidean area of the ellipse is

(interessante!!!)

$$\pi \frac{(1-r^2/\kappa^2)^{3/2}}{\kappa^2} = \frac{\pi}{\kappa^2}\,\text{sech}^3\left(\frac{r}{\kappa}\right).$$

↳ Wrong !!! (ajustar)

If we use the hyperbolic definition of an angular momentum ⇐ for the stereographic inner product model that we found previosly in (9.2.5),

$$l_h = \frac{r^2\dot{\vartheta}}{1-r^2/\kappa^2}, \qquad (9.6.28)$$

then we must modify the equation for the radius of the trajectory,

$$\dot{r} = \pm c\left(1-\frac{r^2}{\kappa^2}\right)\sqrt{\left[1-\frac{\Delta^2}{r^2}\left(1-\frac{r^2}{\kappa^2}\right)\right]}, \qquad (9.6.29)$$

if and only if $\Delta = l_e/c$, where $l_e = r^2\dot{\vartheta}$, the Euclidean expression for the angular momentum.

Pág. 476 (502) - the curvature (9.10.11)

$\boxed{9.10.11}$, the curvature will be negative and constant. In general, the curvature is

$$K = -T_0^2\left(\frac{\lambda^2}{r_1^2}\, j_0^2 + j_0'^2\right). \quad (9.10.11)$$

Retornando:

The curvature (9.10.11) cannot distinguish between the diffusion equation (9.10.7) or a wave equation.

9.10.2 A matter of curvature

Geometries which are both homogeneous and isotropic have constant (Gaussian) curvature.

Gaussian curvature is a measure of a surface's intrinsic geometry or the invariance of a surface to bending without stretching.

As we know, there are three distinct simply connected isotropic geometries in any dimension:

Euclidean with zero curvature, elliptic with positive curvature, and hyperbolic with negative curvature.

cf. Search: curvature (184 ocorrências)

Pág. 85(111) – Fig. 2.16.

Using Euclidean geometry to derive the angle of parallelism by considering concentric limiting arcs.

Pág. 88(114) – 2.5

A brief history of Hyperbolic Geometry.

Pág. 89(115) – the parallax is measured by comparing the apparent position of a star S with that of some reference star S' which is much more distant. This is shown in Fig. 2.18, where the angle ϕ is the parallax of the star. It is the upper limit on the defect $\mathbb{D}$ of the triangle ΔASM,

$$\mathbb{D} = \pi - \left(\frac{\pi}{2} + \alpha + \beta\right) < \frac{\pi}{2} - \alpha = \phi.$$

IMPORTANTÍSSIMO: Pág. 67(93)

Fig. 2.7.

It appears that rules get longer as they are moved further from the origin.

However, the elliptic distance from x to y is exactly the same as that from X to Y.

#

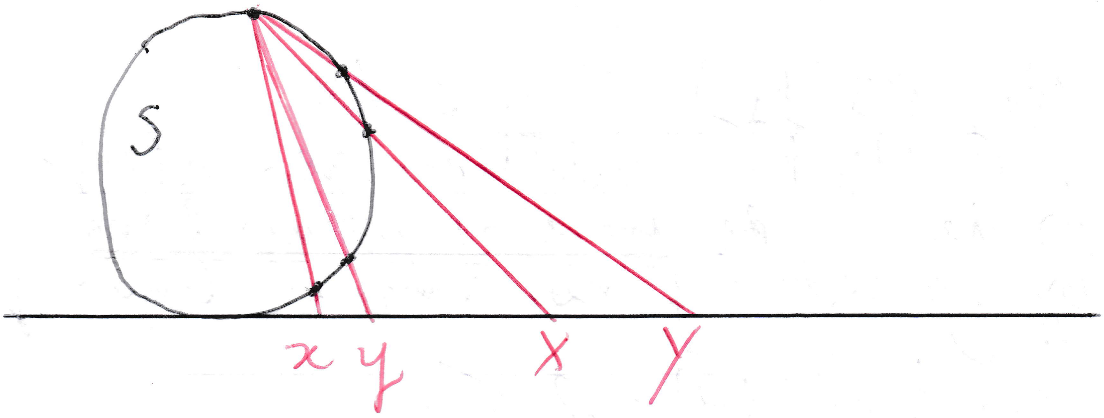

Interessante

Vamos usar este círculo como base para simular a curvatura do espaço e a reta inferior como a superfície a ser projetada.

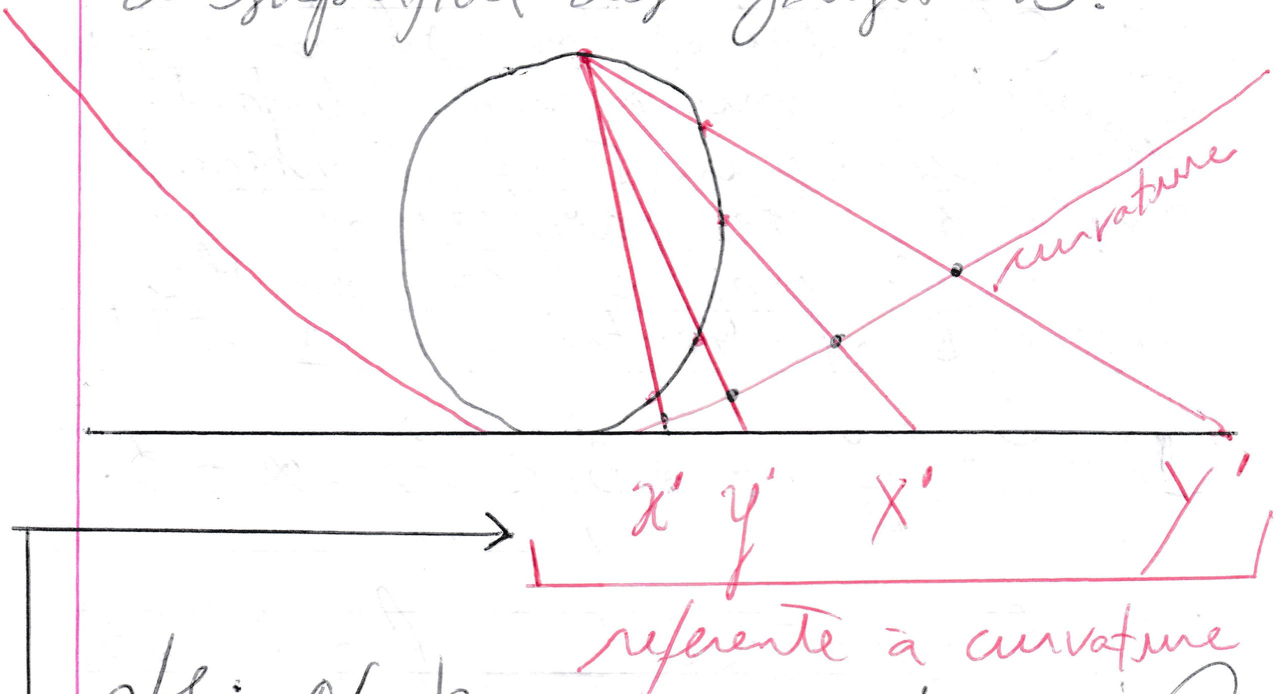
curvatura
x'
y'
X'
Y'
referente à curvatura

obs: Neste caso a nossa projeção é feita na base da curvatura (parte inferior), obtendo-se os pontos x', y', X' e Y'.

obs: Pág. 70 (96) – Fig. 2.9
Calculation of cross-ratio and perspectivity.

Usar esta regra para calcular na curvatura.

Aplication: Usar também para estes calculos:

Pág. 83 (109) - Fig. 2.14

Surfaces of negative constant curvature that are mapped onto part of the hyperbolic plane.

The middle figure is the mapping of a pseudosphere that produces horocycles as dashed curves.

given up $l/m = e^{a/k}$; (2.4.2)

where a is the distance between the arcs and k is a positive (absolute) constant.

Usar também:

Pag. 84 (110) - Fig. 2.15

The ratio of concentric limiting arcs depends only on the distance between them.

Obs: Pag. 84 (110)

If we go back to the pseudosphere, we find that for any two points on its surface, the following remarkable relationship holds

$$x^2 + y^2 < k^2:$$

c) Differential Geometry and Its
Visualization.pdf
→ 493 pages
Search! distance (38 ocorrências)
curvature (574 ocorrências)

✡ Pág. 09(36) - Figure 1.6.
Distance of a point from a plane.

Obs: Utilizar este cálculo para calcular a distância entre o ponto na curvatura e a superfície na qual será projetada a curvatura do espaço.

Estudar com cuidado!!!

If we choose P=0, then the distance of Pl from the origin is

$$\boxed{p = \frac{|d|}{\sqrt{a^2+b^2+c^2}}}$$

(Figure 1.6)

✡ IMPORTANTISSIMO:

Pág. 20(47) →

Visualization 1.4.2

(The versiera)

Figure 1.14 → estudar!!!

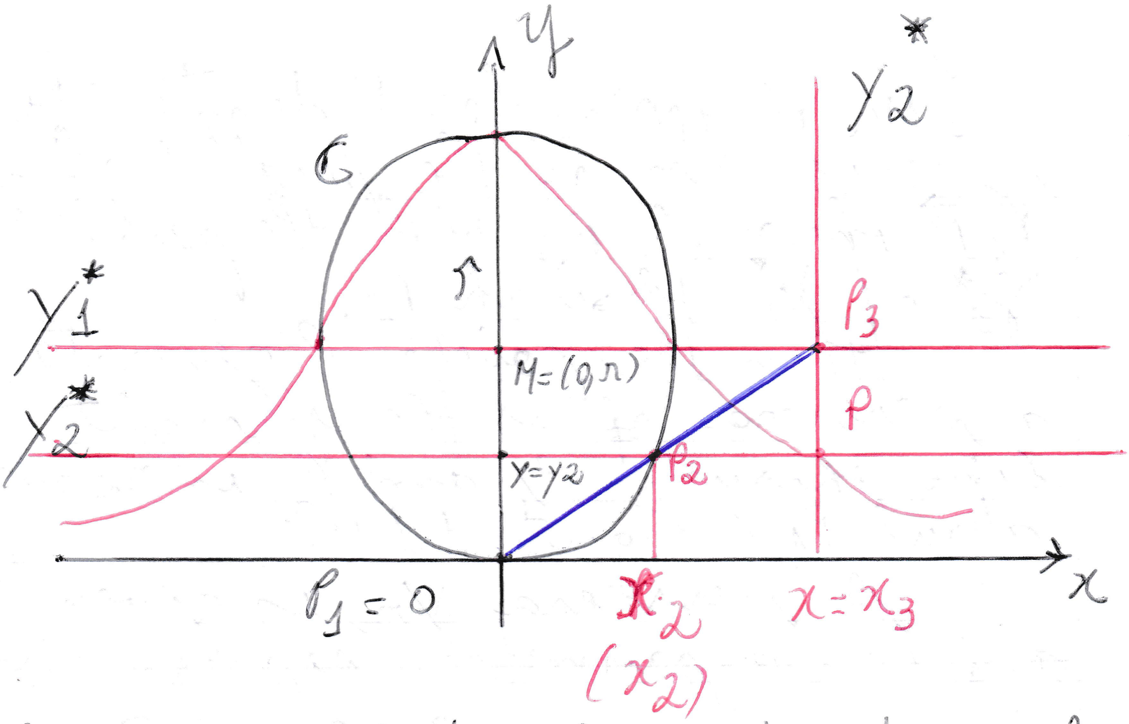

Figure 1.14 Left: the construction of a versiera.

obs: Utilizar em minha pesquisa

A linha vermelha apresenta uma curvatura, similar !!!

Os pontos são referenciais para a checagem dos cálculos.

Visualization 1.4.2 (The versiera).

Let $\mathbb{C} = S_r(M)$ be the circle line of radius $r > 0$ with its centre in M, P_1 and P_2 be distinct points of $\mathbb{C}$, Y_1^* be the straight line through M orthogonal to the vector $\overrightarrow{P_1M}$, and P_3 be the point of intersection of Y_1^* with the straight line through P_1 and P_2.

Furthermore, let P denote the intersection of the parallel γ_2^* of $\overrightarrow{P_1M}$ through P_3 with the parallel γ_3^* of γ_1^* through P_2.

Then the points P generate a versiera ~~of~~ as the straight line through P_1 and P_2 is rotated about P_1 (Figure 1.14).

A versiera γ_r is given by ~~and has a parametric representation~~

$$f(x,y;r)=y(x^2+r^2)-2r^3=0, \quad (1.29)$$

and has a parametric representation

$$\vec{x}_{a,r}(t)=\left\{t, \frac{2r^3}{t^2+r^2}\right\} (t\in\mathbb{R}). \quad (1.30)$$

Ⓐ

IMPORTANTISSIMO:

Pág. 26(53) - Visualization 1.4.5

(The K-curve)

(FIGURE 1.20)

Let γ^* be straight line,

P_0 be a point on γ^* and

$\mathbb{C}$ be a family of circles lines of a fixed radius $\boxed{r > 0}$ with their centres on γ^*.

A K-curve is generated by the points where the tangents from P_0 to the circle lines of $\mathbb{C}$ touch the circles.

A K-curve is given by the equation

$$f(x,y;r) = y^4 + x^2 y^2 - r^2 x^2 = 0, \quad (1.43)$$

and has a parametric representation for two branches

$$\vec{x}^{(k)}(\phi) = \left\{ r\frac{|\cos\phi|}{|\sin\phi|}\cos\phi,\ r\frac{|\cos\phi|}{|\sin\phi|}\sin\phi \right\}$$

$$(\phi \in I_k)$$

for $k = 1, 2$ where

$I_1 = (0, \pi)$ and $I_2 = (0, 2\pi)$. (1.44.

obs: Usar na pesquisa!!!

Interessante !!!

Pág. 51 (78) - Figure 1.38

the vectors of a trihedron of the curve of visualization 1.6.8

⇒

obs:

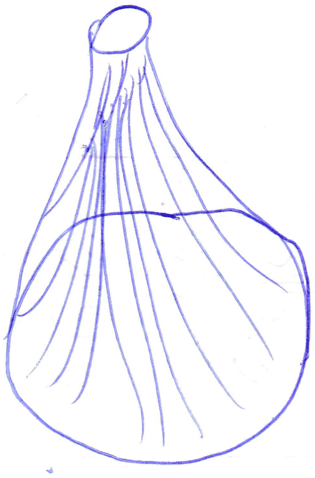

(cone invertido)

Figure 1.38

Usar este cone com o círculo maior para cima e dentro deste círculo maior, incluir a ellipse.

⇒ agregar os cálculos de uma com o outro.

Portanto:

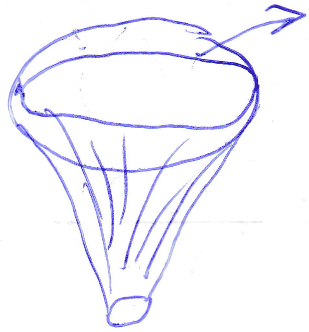

→ ellipse (= diâmetro do círculo)

$K(s) = \frac{2}{\cosh s}$. (1.82)

Important: Estudar livro todo!!!

---

d) Geometry of Cuts and Metrics (1997).pdf

⇒ 597 pages.

Search: distance (1054 ocorrências)

Pág. 71(81) – 6.1.2

Hypermetric and Negative Type Distance Spaces.

Lemma 6.1.10 → $K \geq 2$.

Pág. 86(96) – 6.4.
An Example:
The Spherical Distance Space

$x, y \in S$ the quantity

$$r \cdot \arccos\left(\frac{(x-c)^T (x-y)}{r^2}\right), \quad (6.4.1)$$

where
c denotes the center S and r its radius.

Pág. 167(177) – Part II
Hypermetric Spaces: an Approach via Geometry of Numbers

Pág. 170(180) → A distance that we will use constantly here is the square of the Euclidean distance, namely the distance $d^{(2)}$ defined by

$$d^{(2)}(x,y) := (\|x-y\|_2)^2 = (x-y)^T (x-y)$$

for $x, y \in \mathbb{R}^n$.

Pág. 175(185) – Chapter 13
Preliminaries on Lattices.

$\Rightarrow$

13.1 - Distance Spaces.

A representation of the distance space (X, d) is a mapping

$$i \in X \mapsto v_i \in R^n$$

(where $n \geq 1$) such that (tanto)

$$d(i,j) = (v_i - v_j)^2 \text{ for } i, j \in X. \qquad (13.1.1)$$

In other words, it is an isometric embedding of (X, d) into the space $(R^n, d^{(2)})$, endowed with $d^{(2)}$, the square of the Euclidean distance.

Pag. 527 (537) →

31.4 The Euclidean Distance Matrix Completion Problem.

Let (V_n, d) be a distance space with associated distance matrix D.

Obs: Livro to do!!!

20231104181500

20231105083000m

Data-NovoVivo-010-2023/
Referencia-index/
Bending-Space-Gravity/

a) cont.
Geometry of Cuts and Metrics-
(1997).pdf.

Pág. 511 (521) - Chapter 31
Geometric Properties

→ Geometric type for cut polytope CUT_n^2. One of our objectives here is to study the geometric shape of CUT_n^2, in particular, in connection with its linear relaxation by the semimetric polytope MET_n^2 and its convex (nonpolyhedral) relaxation by the elliptope E_n.

... One way of measuring the importance of a facet is by computing the Euclidean distance of the hyperplane the facet to the barycentrum of CUT_n^2.

Pág. 512 (522) → In general, the projected elliptope $\mathcal{E}(G)$ is contained in the image of $CUT^2(G)$ under the mapping $x \mapsto \cos(\pi x)$.

In Section 31.4 we consider the analogue completion for EUCLIDEAN DISTANCE MATRICES.

31.1 Disproval of a Conjecture of Borsuk using cuts.

Given a set X of points in R^d, is(it) it always to partition X into $d+1$ subsets, each having a smaller diameter than X?

We recall that the diameter[1] (for a polytope P) of a set $X \subseteq R^d$ is defined as

$$\mathrm{diam}(X) := \max_{x,y \in X} \|x-y\|_2 ,$$

the maximum EUCLIDEAN DISTANCE between any two points of X.

Interessante: Usar esta ideia para configurar a curvatura do espaço com o objetivo de projetá-lo sobre um plano.

Obs: Nós podemos estabelecer a superfície da curvatura do espaço de várias formas (polyhedral or not polyhedral), reduzindo-os a pontos infinitos caracterizados matematicamente de forma precisa com o objetivo de projetá-lo sobre uma superfície plana.

Interessante:

Utilizar os livros sobre DIVIDED SPHERES para desenhar a superfície da curvatura do espaço.

Source: Data-NovoVivo-010-2023/
Referencia-index/Divided-Spheres/
001 --- Divided spheres... pdf
→ 525 pages.

b) Light and Waves.pdf → 521 pages.
Search: distance (169 ocorrências)
curved (29 ocorrências)
Pág. 421 (428) - 15.4.3 Curved Space

c) Methods of Mathematical Physics.pdf
Search: distance (29 ocorrências)
Pág. 236(239) - 10.6
Method of Electrostatic Images and Solution of the Dirichlet Problem for a Ball.

For constructing Green's function the following method can be used. Let

$$G(P,M) = \frac{1}{4\pi r_{PM}} + v(P),$$

where a harmonic function v needs to be found, satisfying the condition $v(P) = -\frac{1}{4\pi r_{PM}}$, $P \in \Sigma$.

The function v can be considered as the potential of a field generated by charges located outside the domain and such $v/_{\Sigma} = -\frac{1}{4\pi r_{PM}}$.

→

d) Quantum Field theory in Curved Spacetime Quantized Fields and Gravity.pdf

→ 473 pages.

Search: space (1190 ocorrências).
curved (297 ocorrências).
Gauss (78 ocorrências).
distance(24 ocorrências).

Pág. 183(201) → In curved spacetime,

...

"Recently, Weinberg (2005, 2006) has extended the "in-in" formalism and used it to calculate higher-order Gaussian and Non-Gaussian correlations in cosmology.

These correlations depend on the whole history of inflation, not just on the behavior of fields when they exit from the horizon.

Pág. 196 (214) - 5.3.3.

ζ - function regularization

For example, in flat Euclidean space, the Laplacian

$-\nabla_\mu \nabla^\mu$ acting on scalars has eigenvalues

$\boxed{k^2 \geq 0.}$

In flat Minkowski spacetime,

$\Box = \nabla_\mu \nabla^\mu$ has eigenvalues $\boxed{-(k^0)^2 + k \cdot k}$

which have no definite sign.

The Feynman boundary conditions adopted in Minkowski spacetime quantum field theory ensure that the WICK rotation $k^0 \to ik^0$ may be ~~permo~~ performed leading to a result which is equivalent to using elliptic operators in a Euclidean spacetime.

A rigorous proof of the equivalence between Minkowski and Euclidean field theory may be given.[5]

(5) See Osterwalder and Schrader (1973, 1975) for example.

Pag. 235(253) - 5.6.3

Dirac, Weyl, and Majorana spinors

The Dirac equation in curved spacetime has already been written down in (5.281).

If we allow the Dirac field to be coupled to an electromagnetic field, we have

$$[i\gamma^\mu D_\mu - m]\Psi(x)=0 \qquad (5.343)$$

where $D_\mu = \nabla_\mu + ieA_\mu(x)$ is the gauge-covariant derivative.

Obs: Livro todo!!!

e) Quantum-mechanics.pdf
→ 319 pages.

Search: space (356 ocorrências).

Pag. 13(26) - Chapter 2
Mathematical Preliminary I:
Linear Vector Space.

Obs:

Pág. 16(29) - 2.1.2 Subspace

These vectors form a two-dimensional subspace of the three-dimensional ordinary vector space.

Pag. 17(30) - 2.1.3
Linear Independence of Vectors

Examples:

1. The set of mutually orthogonal unit vectors ($\hat{i}, \hat{j}, \hat{k}$) along x, y, z directions in Cartesian 3-D space (let us denote it by R_3) is a linearly independent set, since

$$c_1\hat{i} + c_2\hat{j} + c_3\hat{k} = 0$$

is satisfied ~~of~~ only if

$c_1 = c_2 = c_3 = 0$, as can be seen

by taking scalar product with

$\hat{i}$, $\hat{j}$, and $\hat{k}$ respectively.

f) Regents Geometry Power Pack... 2021.pdf
→ 1735 pages Obs: Livro todo!!!

g) Relativistic World.pdf
→ 404 pages.
Search: curved (13 ocorrências)

Pág. 74(88) → Consider also how photons represent three-dimensional objects. Let

$$f(x', y', z') = 0$$

be the equation of the surface of such an object in its proper frame. Relative to the current time $\bar{t}$, the time of emission of the photons from point (x, y, z) is somewhere in the past

$$t = \bar{t} - \sqrt{x^2 + y^2 + z^2},$$

depending on the distance from the diaphragm.

Substituting the Lorentz transformations and then projective transformations

$$x = zX/F, \quad y = zY/F$$

into the equation of the surface, we obtain:

$$f\left(zy\left[\frac{X}{F} + v\sqrt{1 + (X^2 + Y^2)/F^2}\right] - vy\bar{t},\ z\frac{Y}{F},\ z\right) = 0 \tag{3.19}$$

IMPORTANTE: moving cube ($v = 0.9$) ⟸

Pag. 221(235) -

These basis vectors can be interpreted visually on the pseudo-Euclidean plane.

→ Lorentz coordinates $\{T, X\}$, the grid is rectangular.

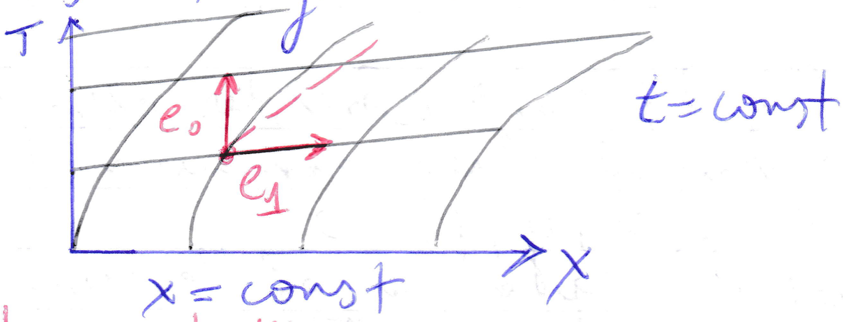

obs: Interessante!!!
mas errado.

Pag. 343(357) — M.5
Curvilinear coordinates

obs: Ajustar estes cálculos.

Δ) Pag. 303(317) - 12 Velocity Space

- Flatland → 12.1
- Coordinates on the sphere → 12.2
- A three-dimensional spherical world → 12.3

Pag. 313(327) →
vertical projection
Beltrami projection
stereographic projection

- Hyperbolic geometry → 12.4
- Pseudo-Euclidean space → 12.5
- Pseudospheres in Minkowski space → 12.6

- Coordinates in Lobachevsky space → 12.7
- Velocity space → 12.8

h) The Irresistible Attraction of Gravity A Journey to Discover Black Holes.pdf
→ 304 pages.
Search! curved (55 ocorrências)

Pág. 28(42) – Curvature

Pag. 38(52) – Fig 3.6
The presence of mass-energy (represented here a sphere, which could be a planet or a grain sand) produces a spacetime with non-zero curvature, which is, therefore, curved.

✡ ↙

Pag. 43(57) – Fig. 3.8
In a curved spacetime, a straight line is not the shortest path to connect two points (i.e. a geodesic). In the figure, the AKB trip is the geodesic between A and B.

Pag. 49(63) – Chapter 4
How to bend spacetime

⇒

Pag. 51(65)

Hence, we could use Equation (4.1) to calculate the curvature strength where you are now as you gread this book - the surface of the Earth.

All it requires is to take the mass of our planet and divide it by ~~the~~ its radius, and we find that the compactness, and thus the relative curvature is

$$\left(\frac{GM_\oplus}{c^2 R_\oplus}\right) \approx \left(\frac{G}{c^2}\right) \frac{5.97 \times 10^{24} kg}{6372 Km} \approx$$

$$7 \times 10^{-10} = 0.0000000007 \quad (4.2)$$

In other words, this simple calculation tell us that on the surfaces of the Earth, the relative curvature is essentially zero, so the spacetime is locally flat.²

i) The Poisson-Boltzmann Equation.pdf
⇒ 133 pages.

Pag. 37(49) - We take $A = 2\sqrt{\pi}/a$ with a as the average minimal distance between the ions.

Pag. 12(25) - 1.5 - the Poisson-Boltzmann Equation in Cylindrical and Spherical Geometry.

obs: Interessante!!!

20231105192100

20231106120000///
Data-NovoVivo-010-2023|
Referencia-index|
Bending-Space-Gravity/

a) 20231105 - Clarke - Relativity and Curved Manifolds - (1992).pdf
→ 458 pages.

Pág. 102 (112) - Chapter 3
The Curvature

3.1 The Riemann Tensor

The connection $\Gamma(\gamma)$ allows us to define the concept of parallelism on each curve γ in M.

Pág. 274 (284) - Chapter 9
Physical Measurements in Space-Time

Pág. 281 (291) - 9.3 Measurements of angles

We then define the angle $\theta_{K,K'}$ between these two directions at A, as:

$$\cos\theta_{K,K'} = \frac{(K'_\perp|K_\perp)}{(K\perp|K\perp)^{\frac{1}{2}}(K'_\perp|K'\perp)^{\frac{1}{2}}} \,. \quad (9.3.2)$$

See: Fig. 9-2 Defining spatial angles in a four-manifolds.

Pág. 282 (292) - 9.4 Curvature effects in the measurement of angles.

Consider an observer carrying a tetrad frame $\{\lambda_{\hat{a}}\}$ along a world line γ and let $\dot{\gamma} \equiv u$.

Pag 297(307) - 9.8
Energy and momentum of a particle.

The most important measurements, after the spatial instantaneous velocity of a particle, are its energy and momentum.

An observer u who monitors a particle moving along a world-line with tangent κ can easily deduce its instantaneous energy and spatial momentum from the knowledge of its Doppler velocity and rest mass.

Call these v and m_0 respectively; then in the rest frame of u we have from special relativity:

$$\mathcal{E} = \frac{m_0 c^2}{\left(1 - \frac{v^2}{c^2}\right)^{\frac{1}{2}}} \qquad (9.8.1)$$

$$P = \frac{m_0 v}{\left(1 - \frac{v^2}{c^2}\right)^{\frac{1}{2}}} \qquad (9.8.2)$$

Pág. 298(308) - 9.9
Measurement of eletric and magnetic fields

We know from (7.7.15) that

the trajectory of the particle deviates from a geodesic because of the Lorentz force which stems from electromagnetic interaction, namely:

$$K^r \nabla_r K^i = \frac{e}{\mu} F^i{}_j K^j . \quad (9.9.1)$$

Here e is the particle's charge and μ its rest energy.

Pág. 300 (310) - this is manifestly the Lorentz force.

Pág. 301 (311) - the parallely projected quantity:

$$\Pi^i{}_j P^j = \mu^i T^{rs} \mu_r \mu_s \quad (9.10.7)$$

Pág. 306 (316) - The small curvature limit

A situation of great interest is when the curvature of the back-ground geometry is sufficiently small that one can neglect non-linear terms up to a certain degree of approximation.

Pág. 307 (317) - Let the space-time be generated by a physical system which is gravitationally bound and weakly (relativistic) relativistic.

b) 20231105 - Clarke - the Analysis of Space-Time Singularities - (1994).pdf
Pág. 16(35) - Chapter 2
The Riemann tensor

The Riemann tensor will play two related roles (papéis) in our work: as the mathematical expression of tidal forces, and as the generator of Lorentz transformations. Both these stem from the equations for families of curves.

Pág. 36(55) - However, the deffinition of the metric is such that, if one applies a varying Lorentz transformation to the points on the curve so as to make it horizontal, then its length becomes less, and so it is still incomplete; also the horizontal curve can be so constructed as to end at the same point on the boundary of LM as the original curve.

c) CELULAR-2023/DOWNload/
Relativity-2022/Coordenadas-

Curvilíneas/ Leçons_sur_les_coordonnées_curvilignes...pdf
→ 412 pages.

d) Data-NovoVivo-010-2023/
Referencia-index/
Bending-Space-Gravity/
Relativity-2022/

d1) Radzevich - Generation of Surfaces-
Kinematic Geometry of Surfaces
Machining ...pdf
→ 724 pages.
Pag. 09(50) - 1.1.4.
Fundamental Forms of a
Surface.

Pag. 13(54) - Figure 1.5
Definition of second fundamental
form, $\phi_{2.P}$ of a smooth regular part
surface P.

d2) Samann - Twistor Geometry,
Supersymmetric Field theories
in Superstring theory.pdf
→ 280 pages.

Pág. 15 - chapter I
Introduction
I.1. High-energy physics and
string theory

In particular, there is the quantum field theory called the standard model of elementary particles, which is a quantum gauge theory with gauge group

$$\boxed{SU(3) \times SU(2) \times U(1)}$$

and describes the

electromagnetic, the weak and the strong interactions on equal footing.

Interessante: Search: curvature

Pág. 32 – 97 Theorem
A complex vector bundle E is holomorphic

Obs: Holomorphic Functions
(3 volumes) –

Source:
Data-NovoVivo-010-2023/
Referencia-Index/
Bending-Space-Gravity/
Holomorphic-Functions/
→ 3 volumes (Gunning)
20231106-Gunning-Introduction to Holomorphic Functions of Several Variables - Volume I, II and III.pdf

Retornando:

§7 theorem – A complex vector bundle E is holomorphic if and only if there exists a holomorphic structure $\bar{\partial}$ on E. For more details on this statement, see section 11.2.3 (Page 36).

§8 Connection and curvature.

Given a complex vector bundle $E \to M$, a connection is a $\mathbb{C}$-linear map $\nabla : \Gamma(M,E) \to \Gamma(M, \Lambda^1 E)$ which satisfies the Leibniz rule

$$\nabla(f\sigma) = df \otimes \sigma + f\nabla\sigma, \quad (\text{II}.21)$$

where $f \in C^\infty(M)$ and $\sigma \in \Gamma(M.E)$.

... this frame is called the parallel transport of e_0 along γ.

Pág. 33 – §13 Chern class

... curvature of the Levi-Civita connection.

Pág. 60 – §34 Yau's theorem on supermanifolds.

One finds that such a supermanifold admits a Ricci-flat metric ~~in the~~ if and only if its scalar curvature is vanishing [239]

→

e) Relativity-2022/
A. Held - General relativity and
gravitation.../

e1) Held - General relativity and
gravitation - Volume 1 - (1980).
djvu.
→ 692 pages
Search: curvature (208 ocurrences)

e2) Held - General relativity and
gravitation - Volume 2 - (1980).
djvu
→ 556 pages
Search: curvature (215 ocurrences)

---

f) Relativity-2022/
Advanced General Relativity/
Joel Franklin -
Advanced Mechanics and
General Relativity (2010).
pdf
→ 387 pages

Pag. 127(147) - Chapter 4
Curved Space
Obs: 40 pages.

Search: curvature (52 ocurrences)

Pag. 139(159) - Example ⇒

Take the curve to be a circle of radius r parametrized by $\lambda = \theta$: (4.42)

→ interessante.

g) Retornando:

Data-Novembro-010-2023/
Referencia-index/
Bending-Space-Gravity/

g1) Encyclopedia of Mathematics-Tanton.pdf

→ 577 pages.

Pág. 06(15)- Affine Geometry

The study of those properties of geometric figures that remain unchanged by an Affine Transformation is called affine geometry.

For example,

since an affine transformation preserves straight lines and RATIOs of distances between POINTs, the notions of PARALLEL lines, MIDPOINTS of LINE segments, and tangency are valid concepts in affine geometry.

The notion of a CIRCLE,

however (porem), is not.

(A circle can be transformed into an ELLIPSE via an affine transformation. The equidistance of points on the circle from the circle center need not be preserved).

Affine geometry was first studied by Swiss mathematician Leonhard Euler (1707-1783).

Only postulates

1, 2, and 5 of EUCLID'S POSTULATES

remain valid in affine geometry.

g2) Heath - Euclid's Elements - (2021).pdf → 692 pages.

Pag. 9 - Postulates.

1. Let it have been postulated to draw a straight-line from any point to any point.

2. And to produce a finite straight-line continuously in a straight-line.

5. And that if a straight-line falling across two (other) straight-lines makes internal angles on the same side (of itself) less than two right-angles, being produced to infinity, the two (other) straight-lines meet on that side (of the original straight-line) that the (internal angles) are less then two right-angles (and do not meet on the other side)[a]

20231106 181000

20231107164500

Data NovoVivo-010-2023/
Referencia-index/
Bending-Space-Gravity/
Encyclopedia of Mathematics-Tanton.pdf

Pag. 13(22) - Angle.

$1 \text{ grau} = \frac{2\pi}{360} \cong 0.017$ radian.

$1 \text{ rad} = \frac{360}{2\pi} \cong 57.3°$.

The full surface area of the sphere is 4π steradian.

Obs: Trisecting an angle.

Pag. 512(521) - Trisecting an angle.

Pag. 372(381) - Pappus's theorems.

Pag. 249(258) - Heron's formula

Area of a triangle

solely in terms of the side-lengths of the triangle.

$$\boxed{\text{area} = \sqrt{s(s-a)(s-b)(s-c)}}$$

where

a, b, and c are the sides of the triangle,

and $s = \frac{1}{2}(a+b+c)$ is its "semiperimeter".

Proved: Heron's formula →

if θ is the angle between the sides of length a and b, then the area of the triangle is given by area

$$\boxed{\text{area} = \frac{1}{2} ab \sin(\theta).}$$

The Law of Cosines asserts that

$$\boxed{c^2 = a^2 + b^2 - 2ab\cos(\theta).}$$

Solving for $\sin(\theta)$ and $\cos(\theta)$, and substituting into the standard identity from Trigonometry,

$$\boxed{\cos^2(\theta) + \sin^2(\theta) = 1},$$

yields, after some algebraic work, Heron's result.

Heron claims that if x approximates the square root of a number N, then

$$\frac{x + \frac{N}{x}}{2}$$ is a better approximation.

For example: $3 \Rightarrow \dfrac{3 + \frac{10}{3}}{2} \simeq 3.1667$

as an improved estimate.

Repeating the procedure yields

$$\frac{3.1667 + \frac{10}{3.1667}}{2} \cong 3.1623$$

as an even better approximation (In fact, $\sqrt{10} \cong 3.16227766$).

Then $(y - \sqrt{N}) = \frac{1}{2x}(x - \sqrt{N})^2$, which shows that if the error $x - \sqrt{N}$ is small, then the error $y - \sqrt{N}$ will be even smaller. (We are assuming here that x is a value greater than 1.).

cf. Newton's Method when applied to the function $f(x) = x^2 - N$.

Pág. 253 (262) – Hyperbola.

$$\sqrt{(x-c)^2 + y^2} = \left(\frac{c}{a}x - a\right)$$

valid for one branch of the figure at least.

Set $e = \frac{c}{a} = \frac{\sqrt{a^2+b^2}}{a} = \sqrt{1 + \frac{b^2}{a^2}}$.

This is called the Eccentricity of the hyperbola and has a value greater than 1.

The above equation can be rewritten:

$$\frac{\sqrt{(x-c)^2+y^2}}{x-\frac{a}{e}}=e.$$

See: hyperbolic functions

$$\cosh(t)=\frac{e^t+e^{-t}}{2}$$

$$\sinh(t)=\frac{e^t-e^{-t}}{2}$$

Pag. 255(264) - hyperbolic geometry (Lobachevskian geometry)

Hyperbolic geometry is a Non-Euclidean geometry in which the famous PARALLEL POSTULATE fails in the following manner:

"Through a given point not on a given line, there is more than one line parallel to that given line".

Pag. 257(266) - Identity ⇒ $3\arctan(1/4)+\arctan(1/20)+\arctan(1/1985)=\frac{\pi}{4}$ 20231107180000

20231108073000m
Data-NovoVivo-010-2023|
Ciencia-4-Books|
a) 1020-Historia-da-Ciencia-Vol-II -
Tomo-I....pdf
Pág. 116(118)- 6.6.3 Terceiro Período
Newton e a Mecânica Celeste
Pág. 118(121)-

Em novembro de 1684, Newton enviou a Halley um manuscrito de nove páginas intitulado:

"De motu corporum in gyrum"
(Sobre o movimento dos corpos em uma órbita)
que, além de reafirmar a órbita elíptica dos planetas, continha uma demonstração matemática das três leis de Kepler.

Na curta exposição, Newton demonstrou que, se os planetas se movem numa órbita elíptica, eles precisam estar sob o controle do

quadrado de uma força inversa direcionada para um foco da elipse e...

Obs: Ponto chave para a solução do problema.

...e demonstrou, ainda, que, sob a influência do quadrado de uma força inversa, a órbita dos planetas descreve uma seção cônica.

Análise:

O termo: "o controle do quadrado de uma força inversa direcionada para um foco da elipse".

Nós devemos ter mente que por causa das forças fundamentais da Física, a força fraca, a força forte, e o eletromagnetismo são responsáveis pela geração da "gravidade" que na realidade não existe, havendo apenas os efeitos sobre o espaço e a matéria.

Em relação à matéria, esta por sua vez nada mais é que uma forma de energia.

Einstein em sua fórmula demonstrou isso:

$$E = mc^2$$

O interessante é que Nikola Tesla disse que tudo é energia. Desta forma, a pseudo-gravidade é gerada pelas três forças fundamentais da Física.

Em relação ao termo: "a força de atração entre os corpos é inversamente proporcional ao quadrado da distância entre eles".

De forma bem simplista, esta descoberta foi feita com base na observação das estrelas por outros astrônomos, mas nós podemos também deduzir de algo ainda mais simples decorrente do nosso cotidiano, por exemplo, um aluno sentado na última coluna de sala de aula e na última carteira, para ser ainda mais exagerado, na última coluna, na última fileira e na última carteira ou cadeira, bem no canto da sala, o que o aluno verá?

Ele verá que as carteiras mais próximas deles são maiores do que aquelas que estão no outro lado de sala de aula.

O tamanho das carteiras vai diminuindo a uma taxa fixa, ou seja, a carteira que está mais perto exerce uma força mais forte sobre o aluno do que aquela que está do outro lado. O mesmo acontece com os planetas e as estrelas. De um modo geral, a resposta está no tamanho da luz que se pode capturar, as carteiras mais próximas geram uma reflexão de luz maior por causa de sua proximidade enquanto as mais longínquas exercem uma menor reflexão de luz e portanto, uma pouca interferência por causa da distância.

Na Astronomia Moderna, a luz é uma fonte de dados, como tamanho da estrela, temperatura, distância do observador etc. Portanto, a resposta já estava exposta a um aluno que entra pela primeira vez em uma sala de aula, mas por causa da tenra idade não tem o preparo pra discernir o que já está lá exposto.

Pág. 120 (122) → No Universo, existem dois movimentos naturais: em repouso e em movimento uniforme (velocidade invariável) ao longo de uma linha reta.

Obs: O que é uma linha reta?
(straight line)

Nós devemos ter em mente que há várias acepções do que seja uma linha reta, não me refiro à questões moral, ética, comportamental, mas sim a linha reta real e a linha real teórica.

Este assunto será devidamente aprofundado em uma oportunidade futura em um capítulo separado.

... ao longo de uma linha reta.
Existem três possibilidades de mudança:
a) uma partícula em repouso pode começar a se mover;
b) uma partícula em movimento retilíneo uniforme pode reduzir ou

acelerar sua velocidade;
3) ou uma partícula mudar de direção
ao se desviar da linha reta.

Pág. 125 (127) - 6.7.1. Vácuo.

Pág. 320 (322) → Coulomb

apresentaria à Academia de Ciências três ~~de~~ Memórias (1786, 1787 e 1788) consagradas à distribuição de

ELETRICIDADE

em condutores e à variação da densidade elétrica de dois condutores em contato.

Verificaria Coulomb que a carga elétrica se situa na superfície externa de um condutor, e que sua distribuição estática depende de sua

CURVATURA,

e que o ar não era um isolante ideal.[421]

RETORNANDO: FOCO DA ELIPSE

A elipse na mecânica celeste é causada por causa das forças fundamentais da Física. Caso houvesse a ação da gravidade como uma força fundamental independente e portanto atuante no mundo físico, a gravidade geraria em razão da sua própria existência uma órbita circular perfeita e não elíptica.

Um outro aspecto a ser considerado, é a própria esfericidade do Sol que por sinal não é perfeita, mas poderia ser, caso a "massa" do Sol fosse maior.

A esfericidade perfeita dependerá das três forças fundamentais da Física e não da gravidade, ou seja, não g, depende exclusivamente da ação da "pseudo-gravidade".

Importante:

CURVATURE = ENERGY.

O Sol não é perfeitamente esférico por causa da energia representada pelas três forças fundamentais da Física, que curvam o espaço em volta da estrela e a própria estrela.

✡ ⇒ Com **M.U.S.** haverá a uniformização do espaço de forma matemática, em que nós teremos, a qual compreensão da realidade do mundo físico representado de forma matemática. Assim, um Universo curvo expressado constantemente em uma superfície plana.

O erro da pseudo-gravidade gera o comportamento dos fenômenos da mecânica celeste como periélio e o afélio, ou melhor, a descrição foi feita de forma errônea e não corresponde com a realidade observada.

Vejamos: Pág. 105 (107).

Em 1609, ao estudar a órbita de Marte, descobriu Kepler o primeiro princípio do movimento planetário; Marte descreve órbita elíptica, da qual um dos focos é ocupado pelo Sol.

Verificando que no periélio (ponto de menor afastamento do Sol) a velocidade angular do planeta é máxima, decrescendo depois para chegar a um mínimo afélio (ponto da órbita em que a distância do Sol é máxima), pode KEPLER demonstrar serem iguais as áreas varridas, em tempos iguais, pelo raio do vetor que liga o centro do planeta ao do Sol; ou, em outras palavras, uma linha que une o planeta ao Sol descreve áreas iguais em períodos de tempo iguais, durante a revolução do planeta. Tal descoberta, que corresponde ao segundo princípio do movimento planetário, significa que quanto mais perto do Sol, mais rápida a velocidade do planeta.

b) 1021 - História - da - Ciência - Vol. II - Tomo II ... pdf

Pág. 144 (146) - Em meados de 1830, Faraday era famoso e respeitado, além de reconhecido como o maior cientista experimentador da Inglaterra de seu tempo.

Depois de ter inventado o dínamo, ter descoberto a indução eletromagnética e ter avançado nas pesquisas da eletrólise, mas sem formação matemática e com pouco conhecimento teórico, Faraday, para explicar a configuração formada por limalhas de ferro ao redor de um imã (já observado por Pedro Peregrino no século XIII), sugeriria que as forças magnéticas atuariam como filamentos ou linhas de força que saíam de ímãs e circuitos.

A esse conjunto de linhas Faraday chamaria de "campo", em que as linhas de força magnética formariam o campo magnético e as de força elétrica, o campo elétrico.

→

IMPORTANT:

A introdução de tal concepção (de ação de campo) seria extremamente útil, porquanto explicava que os corpos agiam sobre os outros corpos tanto por contato quanto pela interação dos campos por eles produzidos.

Como tais idéias contrariavam a noção de ação à distância de Newton, aceita universalmente, e temendo que viesse a ser vítima de violentas críticas, Faraday depositou no cofre da Sociedade Real, em 1832, documento com essas idéias. As instruções eram para ser o envelope aberto somente após a sua morte, o que não o impediu de, em janeiro de 1844, explicar, em conferência na Sociedade Real, as linhas principais de suas idéias sobre o assunto.

Essas idéias pioneiras seriam retomadas por **MAXWELL**.

1839 → **GAUSS**

"Teoremas gerais sobre as forças atrativas e de repulsão que atuam de acordo com o inverso do quadrado da distância"

→

Leis de Heinrich Friedrich Lenz (1804-1865) sobre a ação térmica da corrente elétrica (1842/43), segundo os quais, uma corrente elétrica induzida flui numa direção tal que seja gerada uma oposição à mudança que a induziu

(se assim não fosse haveria uma violação da conservação de energia).

Pág. 71(73) – Para muitos autores, o marco inicial da Geometria diferencial seria a publicação da "Disquisitiones circa superficies curvas" (1827), na qual GAUSS definiu a

CURVATURA

da superfície num ponto (curvatura gaussiana ou curvatura total)

e formulou seu famoso THEOREMA EGREGIUM, pelo qual a curvatura total K é uma propriedade absoluta da superfície.

Pág. 72(74) → Riemann abordaria aí as propriedades das variedades topológicas num número arbitrário de dimensões, e, influenciado pela teoria das superfícies, de Gauss, e por seus próprios trabalhos de Física ~~na~~ Matemática, definiria

41

o quadrado da distância de dois pontos infinitamente vizinhos dessa variedade por uma forma quadrática, e mostraria como medir curvatura dessa variedade.

obs: Quadrado da distância um padrão físico que se repete em toda a Física.

Trabalhos de Beltrami e Bianchi sobre as superfícies de curvatura constantes

GABRIEL DARBOUX (1842-1917), sobre os sistemas triplos ortogonais (Leçons sur la théorie générale des surfaces, 1887/96, em 4 volumes).

6.18.2.4. – Problemas Milenares da Geometria Euclidiana

→ "trissecção do triângulo"

Pág. 77(79) - Bernhard Riemann
deduziu da hipótese do
ângulo obtuso uma Geometria
não euclidiana.
→ foi influenciado pela teoria
da curvatura das superfícies,
de Gauss, em

"Pesquisas gerais sobre as
superfícies curvas (1827)"

Na Geometria criada por
Riemann, chamada por muitos
como
"riemanniana" ou "n-dimensional",

a soma dos ângulos maiores
que dois retos é realizada sobre
a superfície de uma esfera e
nenhuma paralela pode ser
traçada de um ponto fora
de uma reta.

Pág. 78(80) - KLEIN
denominaria a
Geometria euclidiana
(hipótese de ângulo reto) de
"Geometria parabólica",

dado que suas retas cortam
o corte fundamentalmente num só

ponto, como a parábola no plano;

a de Lobachevsky-Bolyai-Gauss
(hipótese de ângulo agudo)
é chamada de
"Geometria hiperbólica"

por cortar, como a hipérbole,
a reta ao infinito em dois pontos;

e a de Riemann
(hipótese de ângulo obtuso)
de
"Geometria elíptica",

por não ter suas retas, como a
elipse, ponto de intersecção.

—

IMPORTANTÍSSIMO:

GEOMETRIA
HIPERELÍPTICA
MUGICHIANA

ou

GHM

→ revolucionária !!!

A Geometria Hiperelíptica Musichiana

or

GHM

apresenta um diferencial arrojado e inovador, pois adequa as inconsistências da qualidade fática do mundo físico universal com o formalismo matemático.

Desta forma, com o uso da elipse dupla curvilínea projetada de forma ortogonal sobre uma superfície plana, nós obteremos

M. U. S.,

Mathematical
Uniforme
Space.

✡

Retornando:

Pág. 78(80)

Geometria Euclidiana → hipótese de ângulo reto
(Geometria Parabólica)

Geometria não-euclidiana ou

Geometria Lobachevsky-Bolyai-Gauss ⟷ hipótese de ângulo agudo.
(Geometria hiperbólica)

Geometria Riemanniana → hipótese de ângulo obtuso
(Geometria elíptica)

Geometria Murichiana → hipótese de ângulo hiperelíptico projectivo
(Geometria hiperelíptica)

GMH → Geometria Murichiana Hiperelíptica.

Em branco

(propositalmente)

Pág. 78(80) - Eugenio Beltrami (1835-1900), professor em Bolonha, Pisa, Pavia e Roma, mostrou que a superfície gerada pela revolução de uma tratriz em torno de uma assíntota é uma pseudo-esfera por ter uma

(troca de caneta)

uma curvatura negativa constante

(a esfera tem curvatura positiva constante).

Se a reta entre dois pontos da pseudoesfera for definida como a geodésico por esses pontos, a geometria resultante terá as propriedades dos postulados de

Lobachevsky,

e com a curvatura total constante não nula será não euclidiana (conforme o teorema de Gauss).

Como o plano é uma superfície com curvatura constant nula, concluiu Beltrami que a

GEOMETRIA EUCLIDIANA

GEOMETRIA EUCLIDIANA

pode ser considerada como um intermediário entre as Geometrias de Lobachevsky e a de Riemann.

Pág. 108 (110) – Astrônomo Bessel fez a sua contribuição à Astrometria e do seu observatório em Königsberg, Bessel detectou, em 1838, com um heliômetro fabricado por Fraunhofer, que a estrela CISNE 61 descoberta por Giuseppe Piazzi, duas décadas antes, e admitida como a mais próxima da terra, por sua rápida rotação se movia, em comparação com duas outras estrelas vizinhas, aparentemente num movimento anual elíptico; essa paralaxe só poderia ser causada pelo movimento de translação da Terra.

A paralaxe, calculada em 0,3136" (bastante próxima da atual de 0,30" ± 0,003"), corresponde a uma distância da terra de 590 mil unidades astronômicas, ou de 10,4 anos-luz.

Pág. 137 (139) - O cientista alemão Helmholtz deu apoio à Teoria de Maxwell e incumbiu o jovem físico Heinrich Rudolf Hertz (1857-1894) de estudar o Electromagnetismo.

Hertz (entre 1885 e 1889) concluiu que as propriedades das ondas eram similares às da luz, o que o levou a concluir que as ondas de luz eram radiações eletromagnéticas, segundo as equações de Maxwell.

Estava dado o primeiro passo para o conhecimento do efeito fotoelétrico.

IMPORTANTÍSSIMO:

Pouco depois, o físico irlandês George Fitzgerald (1851-1901) sugeriria a hipótese, conhecida como da

CONTRAÇÃO, pela qual

o comprimento dos corpos materiais variaria, de acordo com seu movimento no éter de uma quantidade que dependia do

QUADRADO DA RAZÃO ENTRE →

QUADRADO DA RAZÃO
ENTRE SUA VELOCIDADE e a
VELOCIDADE DA LUZ.

Obs: Interessante!!!

Essa formulação seria aproveitada pelo físico holandês
Hendrik Antoon Lorentz

Em novembro de 1892, informou à Academia de Ciências de Amsterdã ter adotado a hipótese de Fitzgerald sobre a contração do comprimento num trabalho que seria publicado em 1892/93.

Lorentz sustentaria que a massa de qualquer partícula aumentaria à medida que ela atingisse velocidades cada vez maiores, e que a contração da distância acarretaria uma dilatação do tempo.

Obs: Importantíssimo!!!

Einstein tirou deste argumento a sua Teoria da Relatividade.

Começaria a surgir a noção da velocidade da luz como limite máximo no Universo[153].

[153] → Ben-Dov, Yoav. Convite à Física.

Source: CELULAR-2023/
Download-12---20231108/
20231108-Ben-Dov-Convite à
Física-(1995).pdf

other source: Data-NovoVivo-010-2023/
Referencia-index/
History-of-Astronomy/
20231108-Ben-Dov-Convite à Física-
(1995).pdf

Books: 20231108-Pannekoek-A History
of Astronomy-(1961).pdf
→ 550 pages
20231108-Pannekoek-A History
of Astronomy ...(1989).djvu
→ 541 pages.

Pág. 138(140) - 6.20.3 Eletrologia

A Eletrologia é a parte da Física que investiga os fenômenos elétricos e magnéticos e a interrelação entre seus respectivos campos. Seus amplos âmbitos foram unificados pelas famosas quatro Equações de Maxwell, de 1873, formulação matemática das leis empíricas constitutivas do Eletromagnetismo.

Pág. 140 (142) - Em 1827, Ampère escreveria "Teoria matemática dos fenômenos eletrodinâmicos, deduzida unicamente da experiência", obra que coroaria suas investigações sobre a Eletricidade e o Magnetismo e na qual explicaria os fenômenos pela hipótese de uma interação newtoniana entre elementos de corrente.

Idealizou, então, a célebre "regra da mão direita", pela qual a deflexão da agulha tomaria a direção dos dedos dobrados da mão direita quando esta segurasse o fio condutor, desde que mantivesse o polegar na direção da corrente elétrica.

Na obra, Ampère enunciou igualmente quatro importantes princípios do Eletromagnetismo:

i) as ações de uma corrente ficam invertidas quando se inverte o sentido da corrente;

ii) há igualdade nas ações exercidas sobre um condutor móvel por dois outros, fixos, situados a igual distância do primeiro;

iii) a ação de um circuito fechado, ou de um conjunto de circuitos fechados sobre um elemento infinitésimo de uma corrente elétrica é perpendicular a esse elemento; e

IV)

Obs.: Importante!!!

Estudar com profundidade e atenção a perpendicularidade pois a curvatura do espaço age tanto na "massa" (energia) quanto na eletricidade e eletromagnetismo.

Analisar se há uma

CURVATURA ELÉTRICA
E
CURVATURA ELETROMAGNÉTICA

Obs: Aprofundar!!!

Retornando:

IV) com intensidades constantes, as interações de dois elementos de corrente não mudam quando suas dimensões lineares e suas distâncias

(atenção!!!) linear?

são modificadas numa mesma proporção.

Pag. 168(170) - Em 1857, Clausius → modelo dinâmico dos gases.

1858 → demonstraria que uma molécula de gás viaja uma certa distância média antes de colidir com uma outra, distância que depende do tamanho da molécula e de seu número em determinado volume.

Esse conceito é conhecido como de "livre caminho médio", e já fora tratado por John James Waterston (1811-1883) em Notas sobre a constituição dos fluidos gasosos e uma Teoria dos gases, de 1843. Nessa obra, Waterston afirmaria que "a distância percorrida por uma molécula, após colidir com outra e antes de encontrar ~~a~~ uma segunda, é inversamente proporcional à densidade do meio.

Em 1860 - Maxwell determinou, então, sua célebre fórmula de distribuição de velocidades das moléculas num gás, usando, para tanto, o método dos mínimos quadrados, de GAUSS. Essa distribuição da velocidade é conhecida hoje como "distribuição de Maxwell"

20231108205400

20231109090000///
Data-NovoVivo-010-2023/
Ciencia-4-Books/
1022-Historia-da-Ciência-Vol.III.pdf
Pág. 28(30) - 7.2.1.1.
Os Problemas de Hilbert
Source: Data-NovoVivo-010-2023/
Hilbert-Problems/
obs: Hilbert and Poincaré (interessante!!!).
Pág. 63(65) - 7.2.3.3. - Topologia.
Pág. 105(107) - 7.3.3 - Astronomia Planetária
Pág. 120(122) - 7.3.4.1.1 - Distância e
Classificação das Estrelas.
Pág. 123(125) - 7.3.4.2 - Energia das
Estrelas.
Pág. 241(243) → Matemático inglês

John Anthony Pople se tornou conhecido no meio químico por seu método (1953) de cálculo de orbital molecular do tipo de ligação covalente, que será conhecido como método PPP, por ter sido, desenvolvido igualmente, de forma independente, naquele mesmo ano por Rudolf Pariser (1923) e Robert Parr (1921).
3 P(s) - Pople, Pariser e Parr.

Ainda neste mesmo campo, o físico Walter Kohn (1923), da Universidade da Califórnia, dividiria com Pople o PNQ (Prêmio Nobel de Química) de 1998 por sua contribuição ao desenvolvimento da

Teoria do funcional de densidade

→ TFD;
surgida como alternativa aos métodos semiempíricos no estudo de propriedades do estado fundamental de sistemas moleculares.

A TFD aplica a Mecânica Quântica e é utilizada em Física e Química para investigar a estrutura eletrônica, em especial, a de moléculas.

Pág 64(66) – O trabalho do matemático alemão Felix Hausdorff (1868-1942), considerado, por muitos, como o fundador da Topologia; sua obra Aspectos Básicos da Teoria de Conjuntos, de 1914, criou o chamado

ESPAÇO TOPOLÓGICO,

a partir de quatro axiomas, cuja generalização da noção de Espaço Métrico é marco na evolução do tema[20].

[20] → Boyer, História da Matemática.

Source: Data-NovoVivo-010-2023/
History-of-Mathematics/
→ 20231109 - Boyer - A history of mathematics - Wiley - (2011).pdf
→ 690 pages.
Search: distance (106 occurren)

Book: 20231109 - Boyer - A history of Mathematics...(1991).pdf

Pág. 569 (591) - "Hausdorff axioms"

Pág. 95(117) - Euclid (Elements)
Postulates

Pág. 128(150) - Lost Works
Treatises of Apollonius
→ Plane Loci

Interessante:

Plane Loci → we infer that the following were two of the loci considered:

(1) the locus of points the difference of the squares of whose distances from two fixed points is CONSTANT is a straight line perpendicular to the line joining the points;

(2) the locus of points the ratio of whose distances from two fixed points is constant (and not equal to one) is a circle.

The latter locus is in fact, now know as the "Circle of Apollonius".

Pág. 136(158) - The Three-Line and Four-Line Locus

Apollonius → Book III :

The three- and four-line locus to which reference is made played an important role in mathematics from Euclid to Newton.

Given three lines (or four lines) in a plane, find the locus of a point P that moves so that the square of the distance from P to one of these is proportional to the

product of the distances to the other two

Obs: Estes cálculos serão alterados.

... (or, in the case of four lines, the product of the distance to two of them is proportional to the product of the distances to the other two), the distances being measured at given angles with respect to the lines. Through modern analytic methods, including the normal form of the

straight line,

it is easy to show that the locus is a conic section — real or imaginary, reducible or irreducible. If, for the three-lines locus, equations of the given lines are

$A_1x + B_1y + C_1 = 0,$

$A_2x + B_2y + C_2 = 0,$ and

$A_3x + B_3y + C_3 = 0,$ and if the angles at which the distances are to be measured are θ_1, θ_2, and θ_3, then the locus of $P(x,y)$ is given by

$$\boxed{\frac{(A_1x+B_1y+C_1)^2}{(A_1^2+B_1^2)\sin^2\theta_1} = \frac{K(A_2x+B_2y+C_2)}{\sqrt{A_2^2+B_2^2\sin\theta_2}} \cdot \frac{(A_3x+B_3y+C_3)}{\sqrt{A_3^2\,B_3^2\sin\theta_3}}.}$$

This equation is, in (generral) general, of second degree in x and y; hence, the locus is a conic section.

Pág. 143(165) – Eratosthenes.
Pág. 145(167) – Aristarchus

Pág. 146(168) – $\boxed{\text{Fig. 8.3}}$
→ tamanhos do sol, terra e lua.

R_s, R_e, R_m (Sun, earth, moon)
(radii)

Proportion $\frac{(R_e - 2R_m)}{(R_s - R_e)} = \frac{D_m}{D_s}$

...

$\frac{108}{43} < \frac{R_e}{R_m} < \frac{60}{19}$ and $\frac{19}{3} < \frac{R_s}{R_e} < \frac{43}{6}$.

Pág. 167 (189) – The Pappus Problem.

Pág. 304 (326) –

"The theorem of Cavalieri":

If two solids have equal altitudes, and if sections made by planes parallel to the bases and at equal distances from them are always in a given ratio, then the volumes of the solids also are in this ratio (Smith 1959, pp. 605–609).

"History of Mathematics" and "Source Book in Mathematics" (1959)

Source: Data-Novo Vivo-010-2023/History-of-Mathematics/20231109-Smith-History of Mathematics...pdf

Retornando: 20231109-Boyer-A history of Mathematics...pdf

Pag. 304 (326)

Cavalieri concentrated on an extremely useful geometric theorem equivalent to the modern statement in the calculus.

$$\int_0^a x^n dx = \frac{a^n + 1}{n+1}.$$

IMPORTANT:

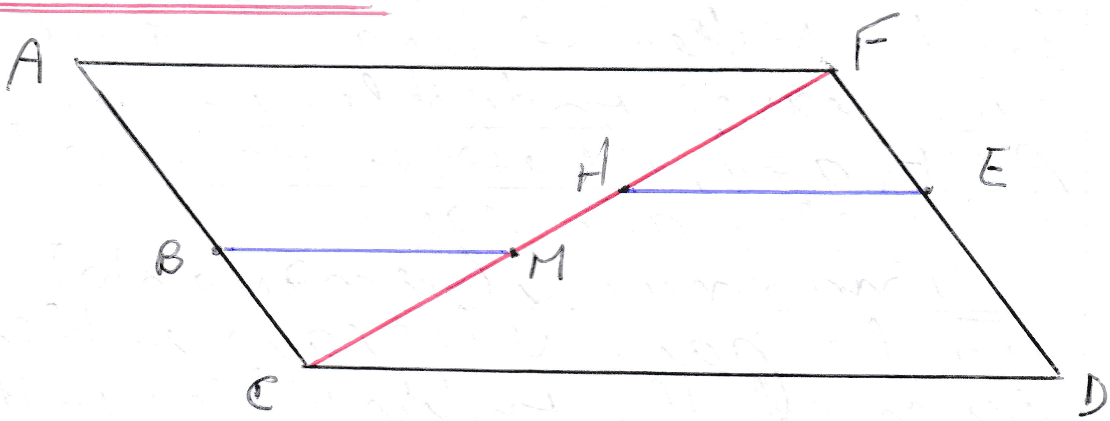

Fig. 15.2

The statement and the proof of the theorem are very different from those with which a modern reader is familiar, for Cavalieri compared powers of the lines in a parallelogram parallel to the base with the corresponding powers of lines in either of the two triangles into which a diagonal divides the parallelogram.

Let the parallelogram AFDC be divided into two triangles by the diagonal CF (Fig. 15.2) and let HE be an indivisible of triangle CDF that is parallel to the base CD. Then, on taking BC = FE and drawing BM parallel to the base CD, ~~then~~, it is easy to show that the indivible BM in triangle ACF will be equal to HE. Hence, one can pair all of the indivisibles of triangle CDF with equal indivibles in triangle ACF, and, therefore, the two triangles are equal.

Inasmuch (inquanto) as the parallelogram is the sum of the indivisibles in the two triangles, it is clear that the sum of the first powers of the lines in one of the constituent triangles is

half of the sum of the first powers of the lines in the parallelogram;

in other words:

$$\int_0^a x\,dx = \frac{a^2}{2}.$$

Through a similar but considerably more involved argument, Cavalieri showed that the sum of the squares of the lines in the triangle is one-third the sum of the squares of the lines in the parallelogram.

For the cubes of the lines, he found the ratio to be 1/4.

Pag. 314(336)- Classification of Curves.

Pag. 315(337)- Rectification of Curves.

Descartes added to the usual axioms of geometry one more axiom:

> Two or more lines (or curves) can be moved, one upon the other, determining by their intersection other curves.

Pag. 369(391)- Newton's Method and Newton's Parallelogram

Pag. 442(464)- Transversals.

Retornando :

Utilizar este paralelismo de Cavalieri (Fig. 15.2) conjuntamente com o postulado de Euclides sobre paralelismo.

Important:

Utilizar o livro: "20231101 – Mlodinow-Euclid's Windows....pdf

Source:

Data-Novo Vivo-010-2023/
History-of-Mathematics/

Retornando: History of Astronomy

Source:

Data-Novo Vivo-010-2023/
Referência-index/
History-of-Astronomy/

20231108 – André Weil – Leonhard Euler – 1703–1783.djvu.

→ 555 pages.

Pág. 86 – item 63.

$$\frac{\pi}{2} = \frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdot \frac{6}{7} \quad \text{(Wallis)}$$

$$\frac{2}{\pi} = \sqrt{\frac{1}{2}}\ \sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}}}\ \sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}}}}$$

(Viète)

off: curvature = Krümmung (19 occurrences)

20231109182900

20231110185300N

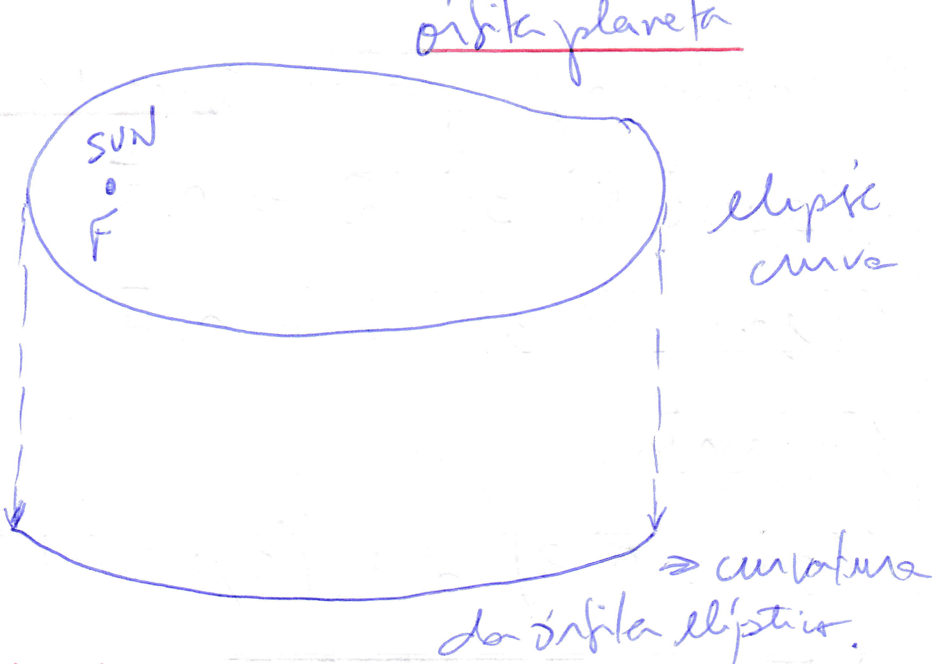
órbita planeta
SUN
F
elipse curve
curvatura da órbita elíptica.

Importante:

Nós devemos pegar esta curvatura e usar o cálculo para calcular a área sobre uma curva.

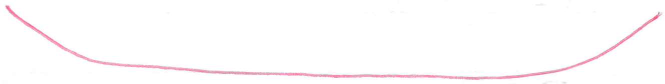

Portanto:

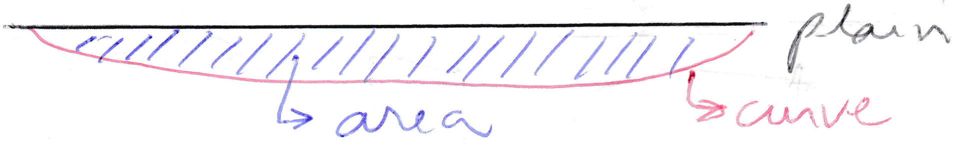
plain
area
curve

obs: O interessante é que com o cálculo da área, nós podemos ter mais um referencial de checagem do $\boxed{M.U.S.}$ →

Mathematical Uniforme Space.

Novo Assunto: (diversos)

1) 355/113 = 3,1415929203...

2) Usando o programa SpeedCrunch

((sqrt 5)/(3))^2 = $0.\overline{5}$

((sqrt 4/3)^2 = $0.\overline{4}$

Portanto:

((sqrt 5/3)^2 + ((sqrt 4)/3)^2 = $\underline{\underline{\underline{1}}}$

3) $$\left[\boxed{16}+\left(1-\frac{\sqrt[2]{5}}{3}\right)\right]^7 =$$

(((16+((1-sqrt(5)/3)))^7 =

299807323,464293009... Km/seg.

(aproximado).

Interessante:

$2\pi R = \pi D$

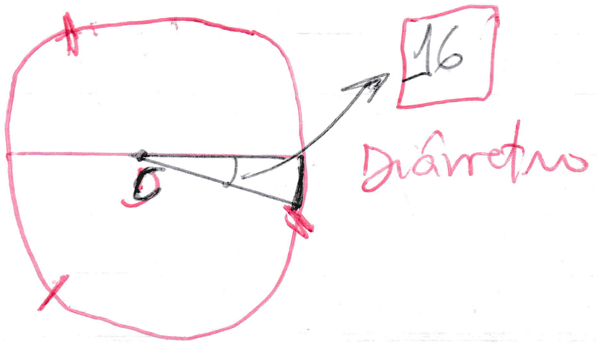

Vejamos:

$$\left(\frac{360}{\pi}\right) \times (\pi - 3) = \boxed{16},2253229\ldots$$

Atenção:

$$\left(1 - \frac{\sqrt{5}}{3}\right) = ((1 - \text{sqrt}(5)/3)) =$$

$$0,2546440075000701\ldots$$

Análise:

$$\left(\frac{45+5=50}{45}\right)^2 - \left(\frac{45-5=40}{45}\right)^2 =$$

$$\left(\frac{50}{45}\right)^2 - \left(\frac{40}{45}\right)^2 = 0,\bar{4}$$

Vejamos:

sqrt((50/45)^2-(40/45)^2)=

$0,\bar{6}$

Multiplicando por 3
(3 dimensões)

(sqrt((50/45)^2-(40/45)^2)*3 = $\boxed{2}$

Portanto:

$$\frac{\sqrt{4}}{3} = 0,\bar{6}$$

Outra parte:

$$\left(\frac{36+5=41}{36}\right)^2-\left(\frac{36-5=31}{36}\right)^2=$$

or (sqrt((41/36)^2-(31/36)^2)

$$=0,7453559924999\ldots$$

Portanto:

$$\frac{\sqrt{5}}{3}=0,7453559924999\ldots$$

Finalmente:

$$\left(\frac{\sqrt{5}}{3}\right)^2+\left(\frac{\sqrt{4}}{3}\right)^2=1$$

Relacionado:

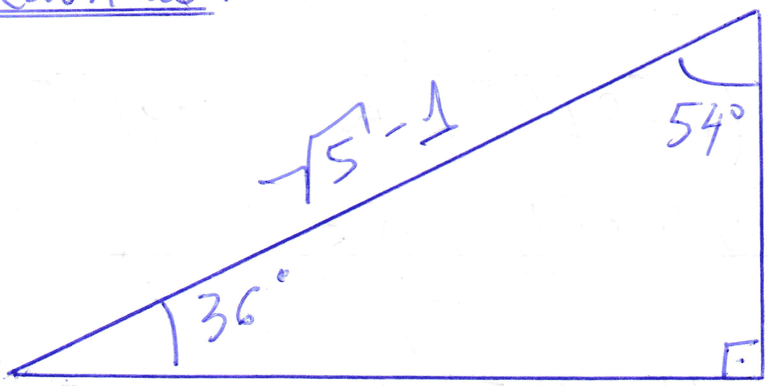

obs: Estudar a relação com a criação de um cone (giro no eixo) +5 −5

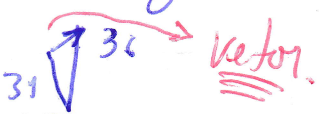

20231110203300

20231111090000m

Números φ e π

$\frac{1}{5} = 0{,}2$ — $\frac{1}{50} = 0{,}02$

$(\pi - 3) = 0{,}1415926$

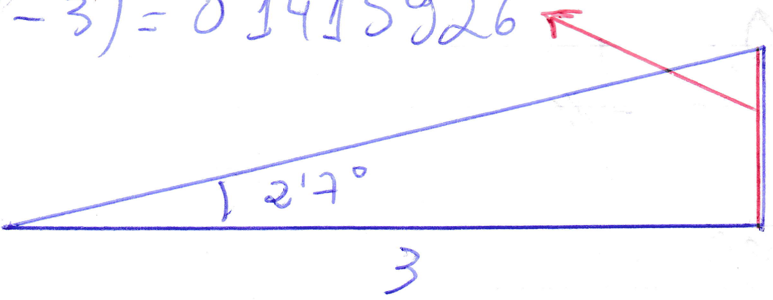

$2{,}7 \times 10 = 27°$

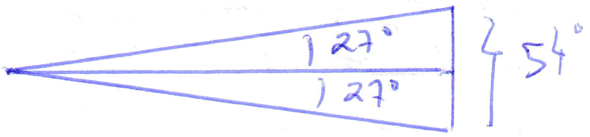

Vejamos:

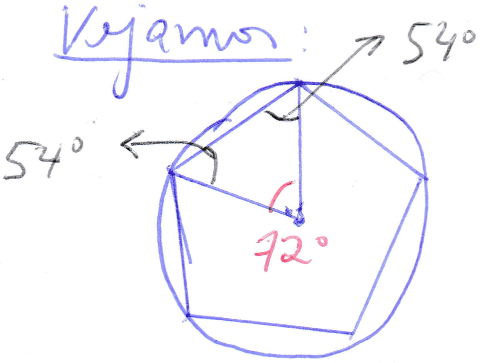

ff:

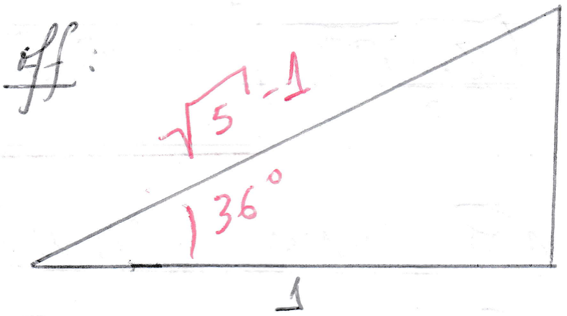

Estudo:

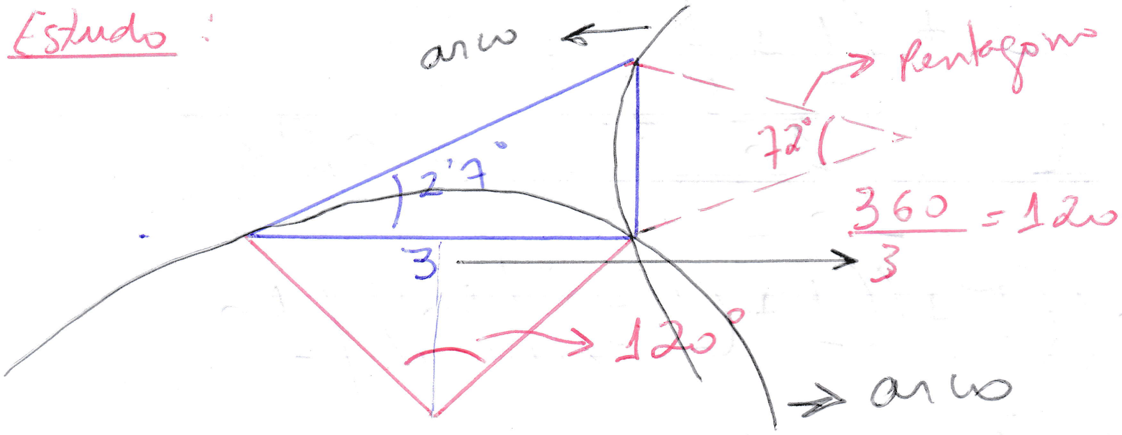
arco
Pentagono
72°
2'7°
360/3 = 120
3
120°
arco

Aprofundando:

Outras observações complementares:

$$\left(\frac{\sqrt{1}}{3}\right)^2 = 0,\bar{1} = ((\text{sqrt } 1)/3)^\wedge 2$$

$$\left(\frac{\sqrt{2}}{3}\right)^2 = 0,\bar{2} \text{ until } \left(\frac{\sqrt{8}}{3}\right)^2 = 0.\bar{8}$$

$$\left(\frac{\sqrt{9}}{3}\right)^2 = 1_{///}$$

Observações:

$$\left(\frac{\sqrt{8.1}}{3}\right)^2 = 0,9$$

$$\left(\frac{\sqrt{8.9999\ldots}}{3}\right)^2 = 0,9999888\ldots$$

Continuar estudo:

$$\frac{\left[5 + \left(\frac{1 + \frac{1}{5}}{10}\right)\right]}{10} = \boxed{0,512}$$

→ ((5+((1+(1/5))/10))/10

Vejamos;

$$\left(0,512 + \frac{1}{0,512}\right)^{0,512} \Rightarrow 1,5871630\ldots$$

Número e

$\left(1+\frac{1}{n}\right)^{n}$, se $n=5$ $\left(1+\frac{1}{5}\right)^{5}=2,48832$

$3-2,48832=0,51168$

Simulação:

$$\left(0,51168+\frac{1}{0,51168}\right)^{0,51168}=1,58700173009...$$

$$\left(1+\frac{1}{0,51168}\right)^{0,51168}=1,740707105...$$

Off: $\begin{array}{r} 0,512 \\ -\ 0,51168 \\ \hline \boxed{0,00032} \end{array} \Rightarrow = \left(\frac{1}{5}\right)^{5} = 0,00032$

Research:

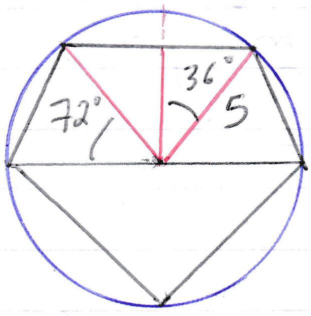

$\cos 36° = 0,809016994...$

$5 \cdot \cos 36° = 4,045084971...$

See:

$(5 \cdot \cos 36°) - 4 = 0,04508497...$

$((5 \cdot \cos 36°) - 4) \times \pi = \boxed{0,1416386...}$

Aprofundando:

$4 \times \cos 36° = \boxed{3{,}236067977\ldots} =$

$1 + \text{sqrt}(5)$. obs) $1 + \sqrt{5} = 2 \times \varphi$.

Interessante!!!

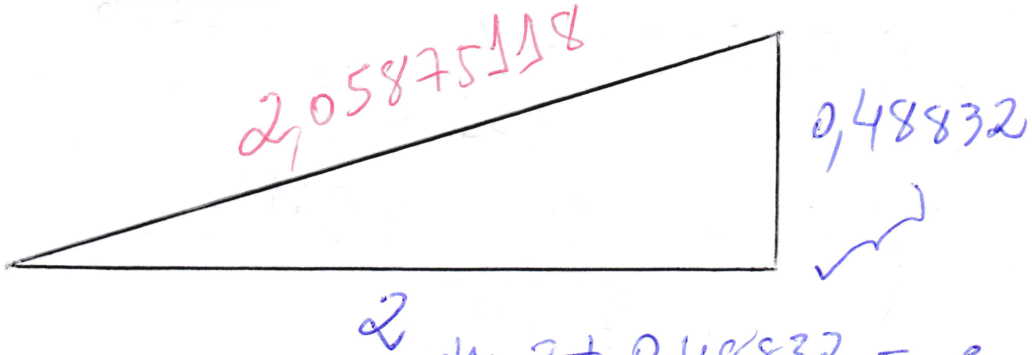

obs: $2 + 0{,}48832 = e$

Ampliando:

$$\boxed{\sin = \frac{co}{h} \quad \cos = \frac{ca}{h} \quad tg = \frac{co}{ca}}$$

Estudando o triângulo retângulo anterior, nós não temos o valor da hipotenusa. Vamos achar por inferência.

off:

$$\frac{1}{\boxed{\sqrt{5} - 1}} = \cos 36 = 0{,}809016994\ldots$$

↳ ficar atento!!!

Interessante: $\varphi = 1,6180339\ldots$

$\sqrt[3]{\frac{\varphi^3}{4\pi}} \times 3 = 1,00374\ldots$ (?)

Portanto:

$\frac{\sqrt{5}-1}{2} = 0,618033988\ldots$

$\frac{1+\sqrt{5}}{2} = \frac{\sqrt{5}+1}{2} = 1,618033988\ldots$

$\therefore$

(((sqrt(5)+1)/2)^3 = 4,236067977

((((sqrt(5)+1)/2)^3)/(4*pi)=

$\underbrace{0,3370955\ldots}_{\times 3}$

$= 1,00374\ldots$ (?)

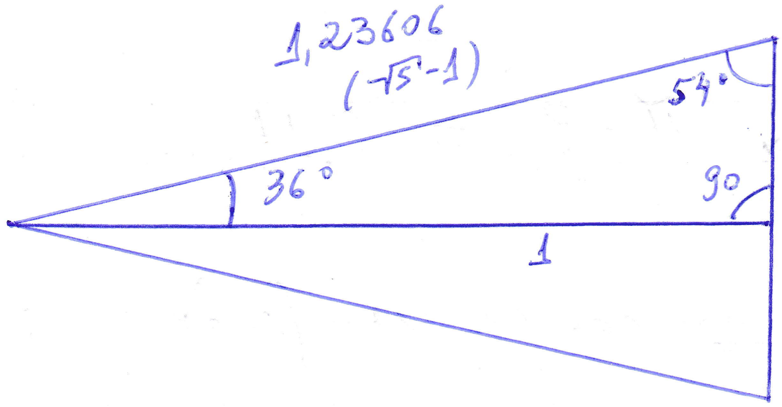

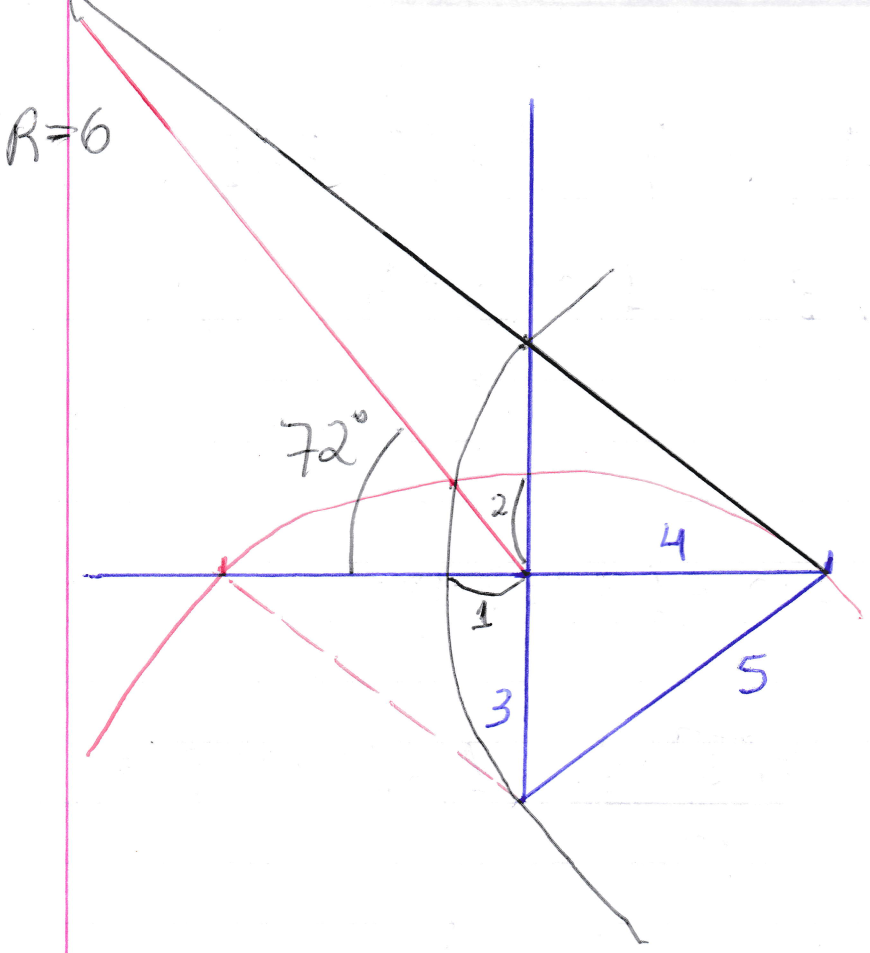

3 + 3 = 6

4 + 4 = 8

$\frac{6}{5} = 1,2$ $\frac{8}{5} = 1,6$

$$\boxed{1,2^2 + 1,6^2 = 2^2}$$

Aprofundando (o estudo!!!)

$$\frac{360^\circ}{2\pi} \times (\pi - 3) = 8,11\ldots$$

Vejamos na circunferência:

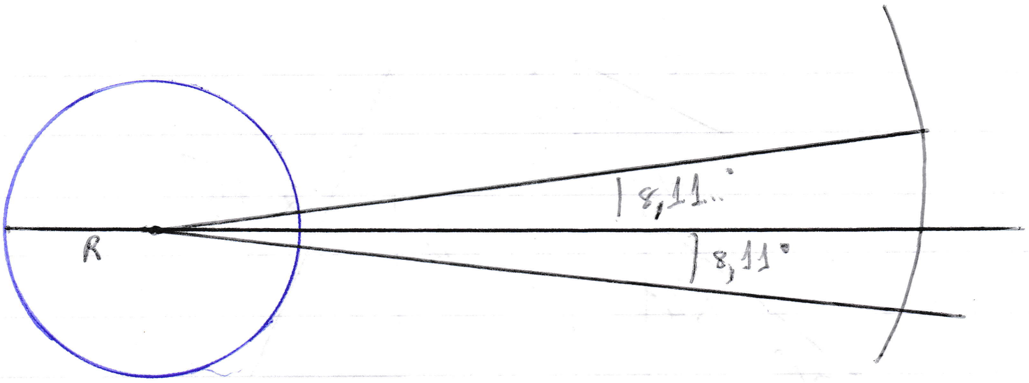

$$\frac{360°}{2\pi} \times (\pi - 3) = 8,112661...$$

Vejamos!

"Velocidade" Luz, dependendo do ângulo, nós observaremos "velocidades" diferentes.

Velocidade da Luz:

$$\boxed{299792458}$$

Usando este valor

$$\sqrt[7]{299792458} =$$

$$\boxed{16,25452886...}$$

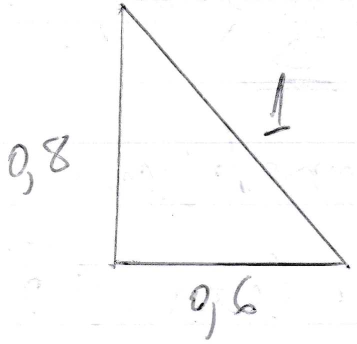

obs: Ficar atento ao Δ Pitágoras 3, 4, 5 //

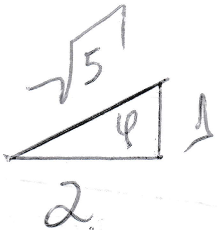

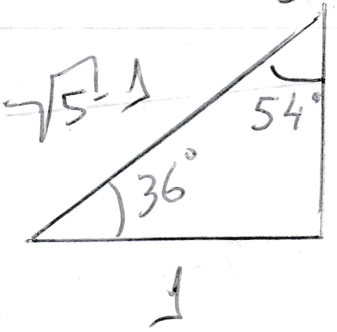

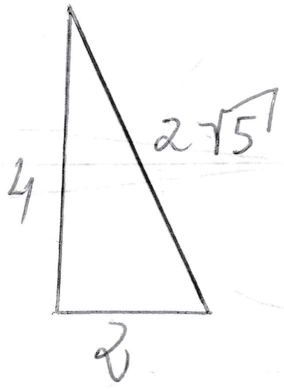

36 →

$$\left.\begin{array}{l} \dfrac{37°}{36°} = 1{,}02\bar{7} \\[2ex] \dfrac{35°}{36°} = 0{,}97\bar{2} \end{array}\right] > 2 \div 2 = \boxed{1}$$

se complementar

$$\underleftrightarrow{1 - \boxed{36} + 1}$$

variação da ondulação
dos forças físicas fundamentais
(3)

obs:

③ { Eletromagnetismo
Força Nuclear Forte
Força Nuclear Fraca.

$$\varphi = ((\text{sqrt}(5)+1)/(2)\ \cancel{\wedge 3} = 1{,}618033\ldots$$

$$\varphi^3 = ((\text{sqrt}(5)+1)/(2)\ \wedge 3 = 4{,}23606\ldots$$

$$\frac{1}{\varphi^3} = \frac{1}{((\text{sqrt}(5)+1)/(2)\ \wedge 3} = 0{,}23606\ldots$$

Vejamos:

$$\frac{360°}{\left(\frac{1}{\varphi^3}\right)} = \frac{360}{((\text{sqrt}(5)+1)/(2)\ \wedge 3} =$$

1524,984471...

Interessante:

$$\frac{\left[\frac{3\cancel{6}0}{((\text{sqrt}(5+1)/(2)\wedge 3}\right]}{3\cancel{6}0°} = 4{,}23606\ldots$$

Vejamos com detalhes:

$$\frac{360°}{((360/((\text{sqrt}(5)+1)/(2)\ \wedge 3} =$$

84,98447189999...

→

Continuando:

$$\frac{(360/((\mathrm{sqrt}(5)+1)/2)^\wedge 3)}{360} =$$

0,236067977...

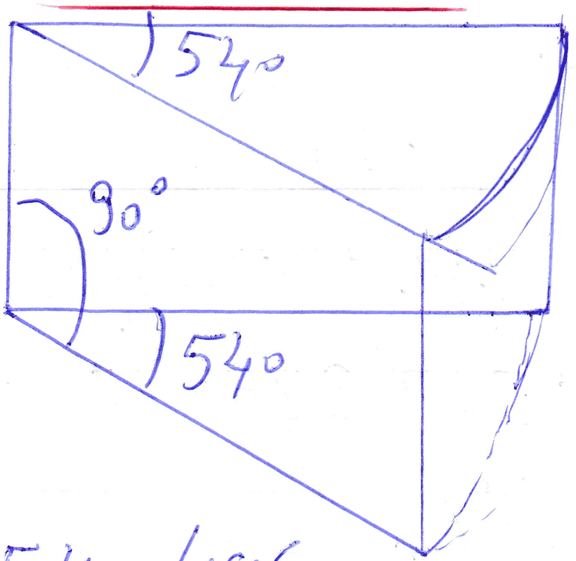

$$\left\{\begin{array}{l} 90 \times 54 = 4860 \\ 72^2 - 18^2 = 4860 \end{array}\right.$$

obs: $72 - 18 = 54.$

$90 - 72 = 18.$

Research:

$$\frac{7 \times \tilde{\pi} + \frac{1}{\varphi^3}}{K} = G \Rightarrow$$

$G = 6{,}674008e^{-11}$

Interessante:

$$\frac{1}{50} = 0{,}02 \Rightarrow \frac{0{,}02}{G} \therefore \quad ✡$$

$$\swarrow$$

$$\frac{\frac{1}{50}}{6{,}67408e-11} = 299666770{,}551 \ ?$$

(Velocidade Luz !!!)

Fórmula: $\nearrow \left(1 + \frac{1}{n}\right)^{n} \quad \boxed{n = G}$

$$\frac{Luz}{0{,}02} = \boxed{\frac{1}{G}} =$$

$$\frac{Luz}{0{,}02} = \frac{1}{G} = \frac{K}{7 \times \pi + \frac{1}{\varphi 3}}$$

Física Geométrica Dinâmica: (consequência)

$$\frac{Luz}{K} = \frac{0{,}02}{7 \times \pi + \frac{1}{\varphi 3}}$$

Interessante:

Vejamos: $\Longrightarrow$

$$\frac{0,02}{299792458} =$$

0,000 000 000 0 6671281903...

or

$6,671281 \times 10^{-11}$

Interessante:

$$\sqrt[7]{299792458} =$$

$16,25452886...^\circ$

obs: $\frac{360^\circ}{2\times\pi}\times(\pi-3)$

=

(360/(2*pi))*(pi-3)

=

8,11266...

×2

$\boxed{16,22532...}$

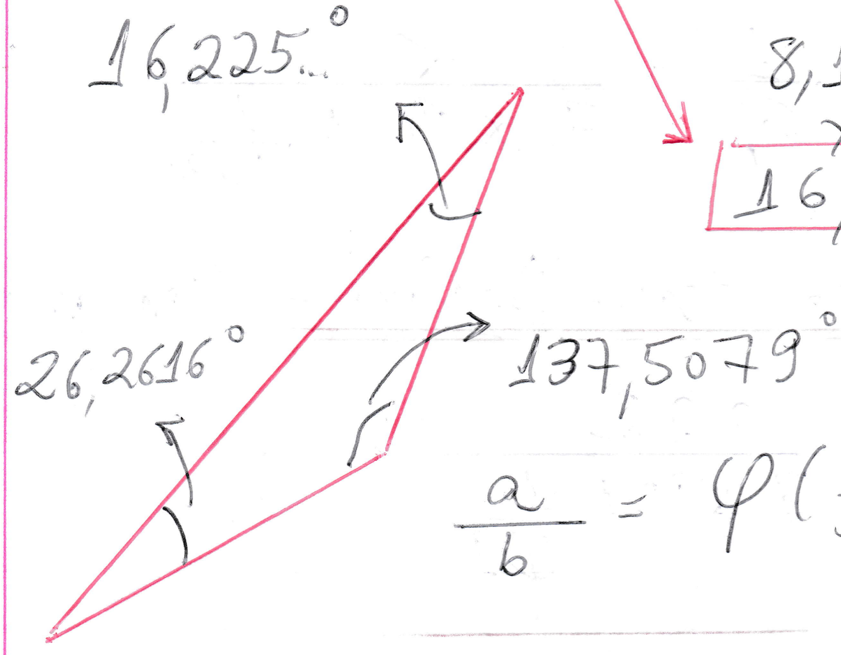

$$\frac{a}{b} = \varphi\,(fi)$$

Outros cálculos:

"Velocidade da Luz" e ângulo

$\sqrt[7]{299792458} = \underline{16{,}254528}\ldots^{\circ}$

or

$\frac{1}{\pi} \times (\pi - 3) \times 360^{\circ} =$

((1/pi)*(pi-3))*360 =

$\underline{16{,}2253229}\ldots.^{\circ}$

Outros cálculos :

$\frac{1}{5} = 0{,}2 \Rightarrow 0{,}2 \times 360^{\circ} = 72^{\circ}$

(pentágono)

IMPORTANTÍSSIMO:

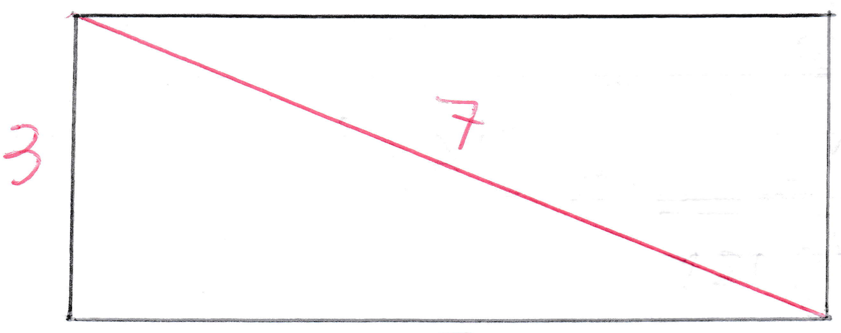

$\sqrt{40} \cong 2 \times \sqrt{2} \times \sqrt{5}$

(2*(sqrt(2)*(sqrt(5))=

6,32455532...

$(= \sqrt{40})$

off: (2*(sqrt(2)*(sqrt(5))=
6,3245553...

Atenção:

(2*(sqrt(2)*(sqrt(5))^2))=
14,1421356...

obs: Sqrt(2) =
1,414221356...

$\Big\}\ \div 10$

Return:

Voltando ao triângulo 3, 7, √40
or
3, 7, (2×√2×√5).

Portanto:

$$\frac{\sqrt{40}\times 3}{\frac{1}{3}\times\sqrt{0,025}} = 360°$$

or

((sqrt(40)*3)/((1/3)*(sqrt(0,025)))=
360°

Confirmando:

$$\frac{\frac{1}{3}\times\sqrt{0,025}}{\frac{1}{7}} = x = 0,368932393...$$

or

(1/3) * (sqrt(0,025))/(1/7) = x

Return:

a) $$\frac{\sqrt{40}\times 3}{\frac{1}{3}\times\sqrt{0,025}} = 360°$$

b) $$\left(\frac{h_1}{h_2}\right)^2 = \frac{\left(\frac{\sqrt{40}\times 3}{7}\right)}{\frac{1}{7}} = 2,710523708...$$

or

$$\frac{(h_1)}{}$$

(sqrt(40)*3)/7 = 2,710523708...

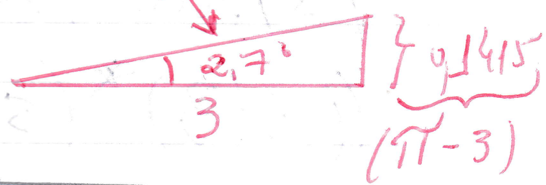

Estudar!!!

Continuando:

$$h_2 = \frac{1}{7} = 0,\overline{142857}$$

Portanto:

(((sqrt(40)*3)/7)/(1,7))^2 = 360°

Off:

(((sqrt(40)*3/7)/(1,7)) = 18,973...

c) $\frac{7}{c}$ or $$\frac{\dfrac{7}{\frac{1}{3}\times\sqrt{0,025}}}{\frac{1}{7}} =$$

or

((7/((1/3)*(sqrt(0,025))/
(1/7)))^2 = 360°

d) $$\left(\frac{\sqrt{40}+3+7}{\sqrt{0,025}+\frac{1}{3}+c}\right)^2 = 360°$$

Vejamos:

((sqrt(40)+3+7))/
(sqrt(0,025)+(1/3)+
((1/3)*(sqrt(0,025))/(1/7)
= 18,973665...

After ^2 = 360°

obs: Interessante, continuar a pesquisa.

20231111222500

20231112090000///

Data-Novo Vivo-010-2023/

Referencia-index/

Importantíssimo-Bending-Light/

a) 20231111- Differential Geometry and General Relativity.pdf.

→ 566 pages.

Search: curve (588 ocorrências)

curvature (200 ocorrências)

bending (5 ocorrências)

Pág. 75(99)- 3.2

Derivative and Parallel Transport of a Vector Field along a curve.

Definition 1:

Suppose V^a is a vector field along a curve $C(t)$. V^a is said to be parallelly transported along $C(t)$ if $T^b \nabla_b V^a = 0$, where $T^a \equiv (\partial/\partial t)^a$ is the tangent vector field of the curve.

b) 20231111- Foundations of Geometric Continuum Mechanics - Geometry and Duality in Continuum Mechanics.pdf

→ 410 pages

Pág. 98(110)- 6.3 Differentiation and Integration

6.3.1- Integral Curves and the Flow of a

Vector Field.

c) 20231112- Tensor Calculus and Differential Geometry for Engineers.pdf
→ 684 pages!

Pág. 433(443)- Chapter 9
Differential Geometry of Surfaces and Curves

Pág. 436(446)- 9.1
Representation of Tensorial Variables on Surfaces

9.1.1. Surfaces in three-Dimensional Euclidean Space.

A two-dimensional surfaces S in the ambient space may be defined by the point function

$$x = \hat{x}^s(t^1, t^2) \qquad (9.1)$$

note that $\Theta^i = \Theta^i(t^\alpha)$ which helps write $x = \hat{x}(\Theta^i) = \hat{x}^s(t^\alpha)$, see (5.1)

... $\alpha = 1, 2$, are called the surface (or Gaussian) coordinates.

⇘ ... In the following, some

well-known surfaces will be introduced having in mind that the surface parametrization is not unique*.

See: Pág. 437(447)

Fig. 9.1 - Ellipsoid.

Obs: interessante!!! Estudar!!!
Ajustar à pesquisa.

Pág. 497(507) - 9.6
Embedded Curves in Three-Dimensional Euclidean Space and Ruled Surfaces.

Recall that curves or one-dimensional manifolds were intrinsically flat for which the Riemann-Christoffel curvature tensor identically vanished.

9.6.1. Mathematical Description of Space Curves.

Consider a space curve[11] as a set of connected points in the surrounding external space such that (tanto que) any of its arbitrary connected subset can be pulled into a straight line segment without any distortion.

[11] A plane curve is a curve lying in a single plane.
In contrast, a space curve is one that may be defined over any region of its embedding three-dimensional space.

---

d) Allday - Space-Time - an introduction to Einstein's theory of gravity-(2019).pdf
→ 371 pages.
Search: curve (115 ocorrências)

Pag. 13(30) - 1.3.1
Universal Gravitation

→ Newton, Kepler.

Pag. 93(110) - Figure 4.7.

A Minkowski diagram for a co-ordinate system S showing the curve connecting all events that have a space-time invariant $s=4$ with respect to the origin. Also know and world lines for point objects moving at 0.3c and 0.7c away from the origin.

Pág. 100 (117) - 4.4. Length Contraction
Figure 4.15.

e) Fundamentos de Matemática Elementar 07 Geometria Analítica... pdf

Pág. 143-G (151) - Capítulo VII
Cônicas

→ Dados dois pontos distintos F_1 e F_2, pertencentes a um plano α, seja 2c a distância entre eles.

obs: Estudar e utilizar as fórmulas em meu modelo de estudo.

f) Kopeikin - Frontiers in Relativistic Celestial Mechanics, Volume 1.pdf

→ 420 pages

Pág. 73 (91) - 2.2
The energy-momentum tensor

Gravitational field and matter are tightly connected via the Bianchi identity of the field equations for the metric tensor [107, 115].

g) Kopeikin - Frontiers in Relativistic Celestial Mechanics, Volume 2.pdf

→ 352 pages.

obs: Search: curve (33 ocorrências)

→

Pág. 39 (69) - Testing relativistic gravity with radio pulsar.

The deflection predicted by GR has been verified with a precision of 1.5×10^{-4}

An even better test for the curvature of spacetime in the vicinity of the Sun is based on the Shapiro delay, the so-called fourth test of GR [11].

[11] → Shapiro. Fourth test of general relativity (1964).

Obs: Pág 41 (71) → Fig. 1

Illustration of the different gravity regimes used in this review.

---

h) Methods of Celestial Mechanics - Verdun - Vol. 1 - 2004.pdf
→ 472 pages.

Pág. 48 (61) - CONSTANT GRAVITATIONAL

$$G = 6.67259 \times 10^{11}\, m^3\, kg^{-1}\, s^{-2}. \quad (3.5)$$

a) One day → 86400 s (SI)

b) mass of Sun → $m_\odot = 1.9891 \times 10^{30}$ Kg.

c) the mean distance between the center of mass Earth-Moon and the Sun, AU (Astronomical Unit), serves as length unit.

d) in these units the constant of gravitation is written as $\boxed{k^2}$, where k is referred to as the

$$\boxed{\text{GAUSSIAN CONSTANT.}}$$

Obs: Today, the AU is defined in such a way that the Gaussian constant K keeps the same numerical value, namely

$$\boxed{K = 0.01720209895 \sqrt{(AU)^3 m_\odot^{-1} d^{-2}}} \qquad (3.6)$$

as it already had at Gauss's epoch.

IMPORTANT:

The constant k is also approximately the mean (média) daily motion of the center of mass of the Earth-Moon system in its orbit around the Sun, expressed in rad/d, what corresponds to

$$\frac{180°}{\pi \cdot K} \approx 0{,}986°/d.$$

Cartesian coordinate system.

Coordinate system is defined by three orthogonal unit vectors

$$\boxed{e_i,\ i = 1, 2, 3},$$

originating from an origin O. For a right-hand system, we have

$$\boxed{e_3 = e_1 \times e_2},$$

where $e_1 \times e_2$ is the vector product of the two vector

$$\boxed{e_i;\ i = 1, 2}.$$

The set of the three unit vectors $e_i = 1, 2, 3$, is also called an orthonormal base in E^3.

A vector x pointing from the origin of the coordinate system to an arbitrary point in E^3 is called a position vector. Each position vector may be represented as a linear combination of the three base vector:

$$x = x_1 e_1 + x_2 e_2 + x_3 e_3 \qquad (3.8)$$

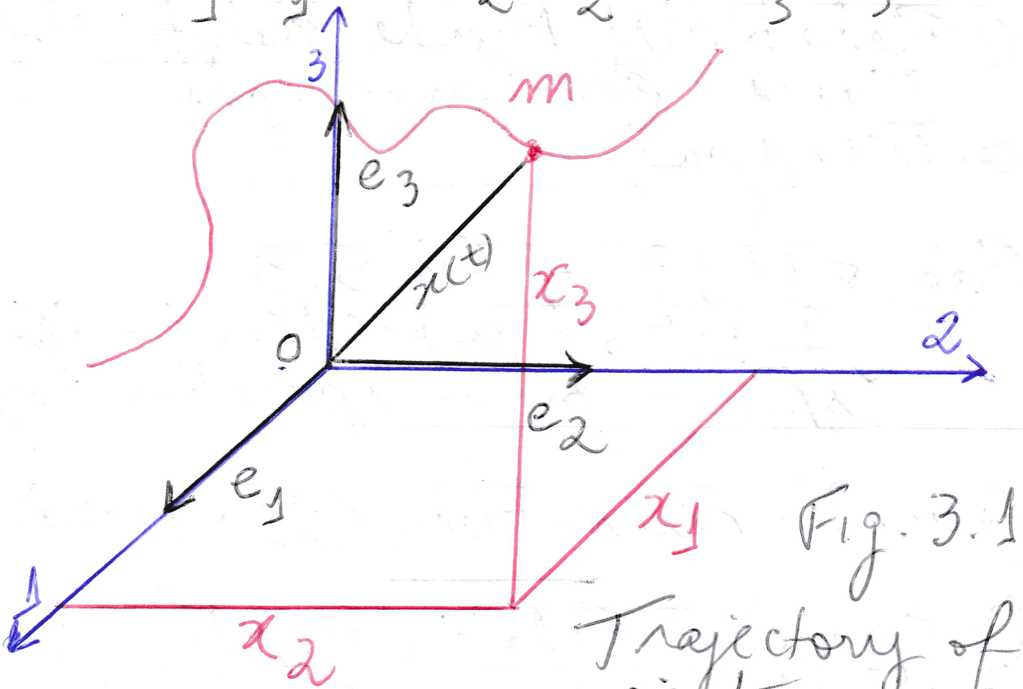

Fig. 3.1 Trajectory of a point mass m in the inertial system.

---

i) Methods of Celestial Mechanics - Verdun - Vol. 2 - 2005.pdf
→ 452 pages.

j) Relativistic Celestial Mechanics of the 1.pdf
→ 918 pages.
Search: bending (29 ocorrências)

Pág. 536(560) - 7.5.2

→

Gravitational bending and deflection angle of light

Obs: Estudar com atenção!!!

---

k) Relativity Made Relatively Easy (PDF Drive).pdf
→ 438 pages

Pág. 81 (99) – Fig 5.3
A set of worldlines is shown on a spacetime diagram.

Search: curvature (110 ocorrências)

---

l) Special Relativity.pdf
→ 441 pages.
Search: bending (19 ocorrências)

Pág. 225 (234) – 11.7 –
Bending of Light in a Gravitational Field.
Pag. 226 (235) – Figure 11.18
→ Newton's method (ver!!!)

---

m) The Irresistible Attraction of Gravity A Journey to Discover Black Holes.pdf
→ 304 pag.
Pág. 51 (65) – Equation 4.2 →
⇒ $7 \times 10^{-10} = 0.0000000007$ (bending)

20231112212200

20231114131100//

a) Data-Novo Vivo-010-2023/
Referencia-index/
Importantíssimo-Bending-light/
Pág. 81(92) - Chapter 6
Relativistic Dynamics: First steps.
Pág. 225(234) - 11.7
Bending of light in a Gravitational Field.
→ 1801 - Johann Georg von Soldner
obs. The original quantitative "calculation" of the gravitational bending of light was done in a nonrelativistic setting in 1801.
Pág. 329(337) - 12.13.
The Poincaré Disc and Non-Euclidean Geometry.
Pág. 423(428) - Appendix G
Equation of Motion in Curved Space and Christoffel Symbol, Two-Dimensional Surfaces, and the Geodesic Deviation and Gravity Waves.

b) Data-Novo Vivo-010-2023/
Referencia-Index/
SI-Unit-Padrão/
20231114-The New International System of Units (SI) Quantum Metrology and Quantum Standards-(2019).pdf

Pág. 153 (173) - Chapter 7
The SI Kilogram, the Mole, and the Planck Constant.
→ Newton's gravitational constant G, ...

Pág. 162 (182) - Figure 7.4
Diameter variations of a single crystal SI sphere.

c) 20231114 - Physics and Chemistry of the Solar System - (2004).pdf
Pág. 8 (23) →

$$\boxed{\tan\theta = \sin\theta = \theta = 1\ \mathrm{AU}/1\ \mathrm{pc} = 1/206{,}264.8}$$

Pág. 230 (245) - All mass determinations depend on a knowledge of the universal gravitational constant, G, but unfortunately this is the least precisely know of the fundamental physical constant.

Obs: Atenção cuidado!!!

d) 20231114 - Introduction to Quantum Metrology
Pág. 22 (35) - The metre is the length of the path travelled by light in vacuum during a time interval of 1/299,792,458 of a second. 20231114204900

e) 20231114- Discovering the Nature of Light-(2022).pdf

Pág. 177(185)- Chapter 10 The Entrance of the Photon and the Birth of Quantum Theory.

They are called light quanta or photons and a light wave of frequency ν they have an energy E given by

$$E = h\nu$$

where h is Planck's constant, a universal constant of nature.

Pág. 181(189) - Planck

$$h = 6.602 \times 10^{-30} \text{ Joules per Hertz.}$$

obs: With this value, the fit of Planck's theoretical curves (to those in Fig. 2) is nearly perfect.

Pág. 186(194) →

$$\boxed{h\nu = KE + W} \qquad (3)$$

In other words, the energy supplied by the light to each photoelectron is the energy $h\nu$ of a photon, so Eq.(2) becomes: $h\nu = KE + W$.

Pág. 196 (204) –

The Wave Nature of Material Particles.

Louis de Broglie in 1924 →

Wave Mechanics.

$$pr = nh/2\pi \quad (4)$$

↓

Equation
Pág. 192 (200)

De Broglie sought some reason for the quantization condition Eq. (4) that was used by Bohr. It occurred to him that if an electron behaved like a wave, then perhaps the allowed orbit in an atom were just those in which the circumference of the orbit equals an integral number of electron wave lengths.

Fig. 14 (a) – For circular orbits (orbit of radius $= r$ and $\lambda =$ wave length)

$$\boxed{n\lambda = 2\pi r}$$

obs: Since $\lambda = h/p$ for an electron?
⇒ $\lambda = h/p$ is the same relation that was used for photons in explaining the Compton effect.

Fig. 14 (c) – first three states. this was the birth of what came to be called wave mechanics. It started with de Broglie's postulate that an electron orbit was a reinforced wave the otherwise moved in that orbit like a particle in a Coulomb field.

20231114213000

20231115103000/// I tried!!!
20231116110000///

Data-NovoVivo-010-2023/
Referencia-index/
Importantissimo/
a) Brandt-The Picture Book of Quantum Mechanics-(2012).pdf
→ 492 pages.
Pag. 126 (143) - 6.5
/Harmonic Motion of a Classical Phase-Space Distribution

There is a particular intermediate initial width, which will turn out to be $\boxed{\sigma_{x0} = \sigma_0/\sqrt{2}}$, for which the classical phase-space distribution keeps its shape while oscillating as a whole.

This is the classical analog of the coherent state of quantum mechanics.

..."The ~~sp~~ phase-space probability density corresponding to a GAUSSIAN WAVE PACKET without correlation between momentum and position at the initial time

$\boxed{t = 0}$ is

Pág. 127(144) ⇒

$$p_i^{cl}(x_i, p_i) = \frac{1}{\underset{\Uparrow}{2\pi}\,\sigma_{x0}\,\sigma_{p0}} \exp\left\{-\frac{1}{2}\right.$$

error!!!

$$\left.\left[\frac{(x_i - x_{0i})^2}{\cancel{\sigma_{x_0}^2}\,\sigma_{x0}^2} + \frac{(p_i - p_{0i})^2}{\sigma_{p0}^2}\right]\right\}.$$

Here x_{0i}, p_{0i} are the initial expectation value and σ_{x0}, σ_{p0} are the initial widths of position and momentum, respectively.

The covariance ellipse of the bivariate Gaussian is caracterized by the exponential being equal to $-1/2$.

Pág. 128(145) – $\boxed{\text{Fig. 6.9}}$

Obs: Important!!!

We summarize our discussion by the following simple statements:

1) A classical phase-space distribution described by a bivariate Gaussian keeps its Gaussian shape.

2) Its center, which is the center of the covariance ellipse, moves on a circle around the center of the x, p plane with angular velocity ω.

3) The covariance ellipse keeps its shape but rotates around its center with the same angular velocity ω.

In figure 6.9 we illustrate the motion of the covariance ellipse for the three cases #

$$\boxed{\sigma_{x_0} < \sigma_{p0}, \quad \sigma_{x0} > \sigma_{p0}, \quad \sigma_{x_0} = \sigma_{p0}.}$$

b) Brandt - Interactive Quantum Mechanics - Quantum Experiments on the Computer - 2011.pdf
→ 394 pages.

Pag. 138(155) - 6.1

Physical Concepts.

6.1.1 - The Schrödinger Equation of a Free Particle in Three Dimensions. The Momentum operator.

... "Three-dimensional free motion is viewed as three one-dimensional harmonic waves (2.1) propagating simultaneously in the coordinates x, y, z of the position vector r:

$$\Psi_p(r,t) = \frac{1}{(2\pi\hbar)^{3/2}} \exp\left[-\frac{i}{\hbar}\left(\frac{p_x^2}{2M}t - p_x x\right)\right] \times$$

$$\exp\left[-\frac{i}{\hbar}\left(\frac{p_y^2}{2M}t - p_y y\right)\right] \exp$$

$$\left[-\frac{i}{\hbar}\left(\frac{p_z^2}{2M}t - p_z z\right)\right] \quad (6.1)$$

Because the kinetic energy E of the particle is given by

$$E = (p_x^2 + p_y^2 + p_z^2)/2M = p^2/2M, \quad (6.2)$$

M: mass of particle

$p = (p_x, p_y, p_z)$: momentum of particle.

c) Calculation-of-the-Cosmological-Constant.pdf
→ 13 pages.

The consideration is based on the study of solutions of the Newton-Schrödinger (N-S) equations, obtained as a c^{-2} expansion of the original Newton-Dirac equations for the neutron.

$A \sim l_p^{-2}$, here

$l_p = \sqrt{\frac{\hbar G}{c^3}}$ -

called Plank length.

Pág. 02

gravitational interaction

$\hbar c / G m^2 = (m_p/m)^2$; here

$m_p = \sqrt{\frac{\hbar c}{G}}$ - Plank mass.

...

$$\Lambda \approx \frac{G^2 m^6}{\hbar^4} = \left(\frac{m}{m_p}\right)^6 \frac{1}{l_p^2} \quad (1)$$

...

d) 2018.-Weiss.pdf
Frégier Points Revisited.

e) 2206.03747.pdf
Fregier Ellipses

f) 2310.07680.pdf
Hamiltonian Dynamics of Bayesian Inference Formalised by Arc Hamiltonian Systems.

Pág. 03 — energy-conserving Hamiltonian flow
(figure)

g) A Mathematician and His Mathematical Work.pdf
→ 715 pages.

Pág. 115(126) —
A Simple Intrinsic proof of the Gauss-Bonnet Formula for closed Riemannian Manifolds.

Pág. 121(132) → On the Curvatura Integra in a Riemannian Manifold

Pág. 480(491) - Deformation of Surfaces preserving principal Curvatures.

Theorem: The non-trivial families of isometric surfaces having the same principal curvatures are the following:

1) a family of surfaces of constant mean curvature;

2) a family of surfaces of non-constant mean curvature. Such surfaces depend on six arbitrary constants, and have the properties:

a) they are W-surfaces;

b) the metric

$$\boxed{d\hat{s}^2 = (\text{grad } H)^2 ds^2/(H^2-K)},$$

where ds^2 is the metric of the surface and H and K are its mean curvature and Gaussian curvature respectively, has

GAUSSIAN CURVATURE EQUAL TO -1.

h) Adams - Relativity - An Introduction to Spacetime Physics - (1997, 2000), pdf
→ 287 pages.

Pág. 188 (205) - 4.2

The Equivalence Principle

4.2.1. - Everything Falls with the same acceleration.

"In contrast to eletric and magnetic fields the gravitational field exhibits a most remarkable property which is of fundamental importance for what follows. Bodies which are moving under the sole influence of a gravitational field receive an acceleration which does not in the least depend on the material or the physical state of the body"

Einstein.

Obs: Autor:

Primeiramente não há um contraste entre eletric and magnetic field, pois a "gravidade", ou seja, a pseudo-gravidade não existe, existindo apenas

os seus feitos decorrentes da eletricidade e magnetismo.

A atração dos corpos depende das três forças fundamentais da física, mas
Einstein inverteu a ordem

PROPOSITALMENTE,

Vejamos:

a) a aceleração não depende do material ou do estado físico do corpo, isto é uma meia-verdade pois não é do material ou estado físico do corpo, mas sim da energia contida naquela "massa".

A massa nada mais é que uma forma de energia.

Portanto, depende da massa (sim) envolvida na questão.

b) A massa que se refere Einstein decorre das forças fundamentais da física, pois ele descreveu os efeitos dessas forças sobre os próprios corpos ou "massas", ou seja, desta forma, a própria energia é a força que atua no próprio corpo ou massa.

Os efeitos não são uma força

autônoma independente, mas o reflexo das próprias forças fundamentais da física.

Confirmando o explicar:

Pág. 189 (206) → free-fall

"fifth force" ???

NO, let alone the fourth force.

Vejamos:

"... "More accurate measurements again confirmed the equality of free-fall accelerations in the gravitational field.

Obs: Neste caso, não há uma "gravitational field" e portanto, o free-fall não pode interferir no "gravitational field".

O mais grave de tudo, o pseudo "free-fall" não existe como força independente e decorre unicamente das três forças fundamentais da física.

Aprofundando: (Pag. 189 (206)

Fig. 4.4 → The Dicke experiment.

The Earth is in free fall around the Sun. ← text

← commentary

Obs: Não é verdade, as forças fundamentais da física continuam agindo sobre a Terra. A inércia da energia da própria Terra comportasse como uma "queda-livre"!

This should eliminate inertial forces in the Earth's frame.

Another way to see this is to say that centrifugal forces is exactly balanced by the gravitational attraction to the sun.

If different bodies have different free-fall accelerations in the same field this balance will not always occur.

In this experiment Dicke looked at a rotation of a torsion balance connected to one aluminium and one gold sphere on a 12 hour cycle as the Earth rotates.

The spheres have equal inertial masses.

Important: →

Broken line indicates rotation if gold falls with greater free-fall acceleration that aluminium.

i) Allday-Space-Time-An Introduction to Einstein's Theory of Gravity-(2019).pdf

→ 371 pages.

Pág. 85(102) - Chapter 4
Space-Time

→ World Lines, Newtonian Space-time.

Pág. 111(128) - Chapter 5
Mass, Energy and Dust
Pág. 112(129) - 5.1.1. Momentum.

Pág. 185(202) - Chapter 8
Weak Field Gravitation

Minkowski metric:

$$g_{ab} = \eta_{ab} + h_{ab}$$

Obs: Estudar com cuidado!!!

Pág. 186(203) - 8.1.1. Newtonian Gravity

20231116162100

20231119132000N
Data-NonVivo-010-2023/
Books/
Salas-Calculus.pdf
→ 1171 pages.
Pág. 154 (178) - Chapter 4
The mean-value theorem; Applications of the first and second derivatives.

The Mean-Value Theorem:

If f is differentiable on the open interval (a, b) and continuous on the closed interval [a, b], then there is at least one number c in (a, b) for wich

$$\boxed{f'(c) = \frac{f(b) - f(a)}{b - a}}$$

→ is to say that the graph of f has at least one point $(c, f(x))$ at which the tangent line is parallel to the line l.

IMPORTANT:

Utilizar este conceito como um instrumento de checagam de minha pesquisa. Paralelo e diagonal ao mesmo tempo.

Vejamos:

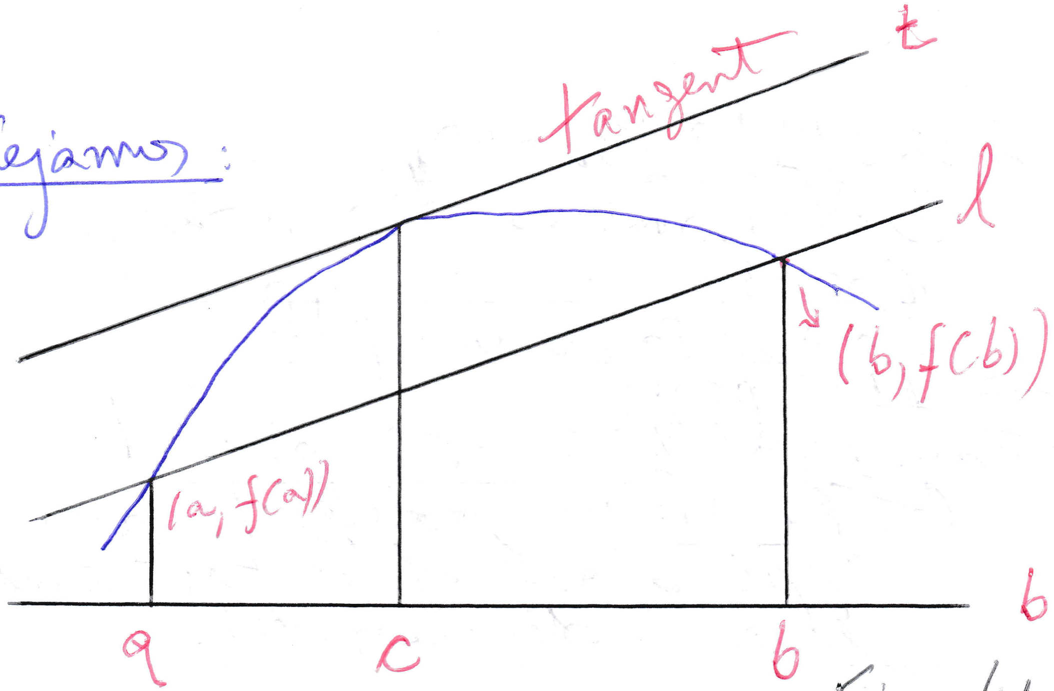

Figure 4.1.1

Análise:

A linha b (base) que contém a projeção dos pontos a, b e c sobre a abscissa x será utilizada como parte superior da curvatura espaço "tempo".

A função f (linha azul) como a curvatura do espaço elliptico da órbita de um planeta ou sistema solar.

A tangente t paralela a linha l será um referencial de checagem de projeção sobre a linha b (eixo x), mas devemos ficar atentos à linha l pois esta representa os dois pontos do nosso sistema de cálculo porque nós deveremos ter dois referenciais de estudo que deverão estar interligados entre si, que nós explanaremos a seguir:

a) Os pontos (a, f(a)) e (b, f(b)) deverão coincidir ou ser os pontos focais de elipse da órbita de um planeta ao redor de sua estrela ou do próprio sistema solar em questão.

Obs: Utilizar o sistema duplo elíptico único musichiano,

esboçado anteriormente com os dois círculos, ponto Frigier e os pontos $\underbrace{Me_1, Me_2, Me_3}$ – Musich points.

Obs: Esses pontos deverão ser alinhados entre si.

Vejamos:

1) P, M_1, S e Me_2.

2) Me_1, S, M_2 e Me_3.

Importante: Utilizar **Pappus's theorem** como checagem dos pontos colineares de modo a manter um referencial objetivo de uma linha reta (straight line).

Obs: Os pontos focais da ellipse são essenciais da a estabilidade dos cálculos e sua respectiva projeção ortogonal sobre o plano.

b) Os pontos $(a, f(a))$ e $(b, f(b))$ deverão representar, em outra simulação, as bordas ou limites da órbita do sistema solar.

Os pontos $(a, f(a))$ e $(b, f(b))$ estão sobre a linha l e são dois pontos que estão na função f. (linha azul).

A tangent t deverá coincidir com a concavidade inferior da curvatura do espaço. (sistema duplo elíptico curvo muschiano)!

Continuando o estudo sobre
Mean-Value Theorem

First we will show that if a function f has a nonzero derivative at some point x_0, then for
x close to x_0
$f(x)$ is greater than $f(x_0)$ on one side of x_0 and less than $f(x_0)$ on the other side of x_0.

Vejamos: Theorem 4.1.2.

Theorem 4.1.2.

Suppose that f is differentiable at x_0!

If $f'(x_0) > 0$, then

$$f(x_0 - h) < f(x_0) < f(x_0 + h)$$

for all positive h sufficiently small!

If $f'(x_0) < 0$, then

$$f(x_0 - h) > f(x_0) > f(x_0 + h)$$

for all positive h sufficiently small.

Proof:

- $f'(x_0) > 0$

by the definition of the derivative,

$$\lim_{k \to 0} \frac{f(x_0 + k) - f(x_0)}{k} = f'(x_0).$$

- With $f'(x_0) > 0$, we can use $f'(x_0)$ itself as ϵ and conclude that there exists $\delta > 0$ such that

→

if $0 < |k| < \delta$, then

$$\left| \frac{f(x_0+k)-f(x_0)}{k} - f'(x_0) \right| < f'(x_0)$$

For such k, we have

$$-f'(x_0) < \frac{f(x_0+k)-f(x_0)}{k} - f'(x_0) < f'(x_0)$$

and thus

$$0 < \frac{f(x_0+k)-f(x_0)}{k} < 2f'(x_0).$$

In particular

$$(*) \quad \frac{f(x_0+k)-f(x_0)}{k} > 0.$$

We have shown that $(*)$ holds for all numbers k which satisfy the condition $0 < |k| < \delta$.

If $0 < h < \delta$, then

$$0 < |h| < \delta \text{ and}$$

$$0 < |-h| < \delta.$$

Consequently,

$\Rightarrow$

$$\frac{f(x_0+h)-f(x_0)}{h} > 0 \text{ and}$$

$$\frac{f(x_0-h)-f(x_0)}{-h} > 0.$$

The first inequality show that $f(x_0+h)-f(x_0) > 0$ and therefore $f(x_0) < f(x_0+h)$.

The second inequality shows that $f(x_0-h)-f(x_0) < 0$ and therefore $f(x_0-h < f(x_0)$.

IMPORTANT:

Next we prove a special case the mean-value theorem, known as

$$\boxed{\text{ROLLE'S THEOREM}}$$

(1691)

In Rolle's theorem, we make the additional assumption that

$f(a)$ and $f(b)$ are both 0.

(See Figure 4.1.2) →

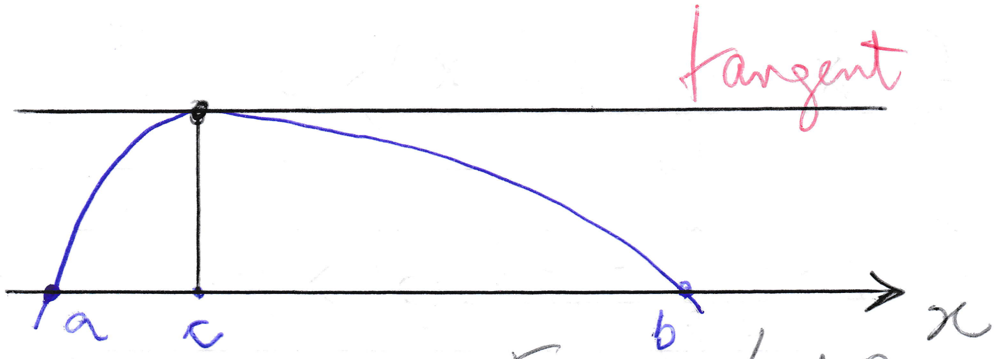

Figure 4.1.2.

Obs: Utilizar esta configuração para criar um referencial de checagem sobre a curvatura do espaço (bending light).

In this case, the line through $(a, f(a))$ and $(b, f(b))$ is horizontal. (It is the x-axis). The conclusion is that there is a point $(c, f(c))$ at which the tangent line is horizontal.

THEOREM 4.1.3
ROLLE THEOREM

Suppose that f is differentiable on the open interval (a, b) and continuous on the close interval $[a, b]$. If $f(a)$ and $f(b)$ are both 0, then there is at least one number c in (a, b) for which

$$\boxed{f'(c)=0}$$

PROOF: If f is constantly 0 on $[a, b]$, then $f'(c) = 0$ for all c in (a, b).

If f is not constantly 0 on $[a, b]$, then f takes on either some positive values or some negative values.

Since f is continuous on $[a, b]$, f must take on a maximum value at some point c of $[a, b]$ (Theorem 2.6). This maximum value, $f(c)$, must be positive. Since $f(a)$ and $f(b)$ are both 0, c cannot be a and it cannot be b. This means that c must lie in the open interval (a, b) and therefore $f'(c)$ exists. Now $f'(c)$ cannot be greater than 0 and it cannot be less than 0 because in either case f would have to take on values greater than $f(c)$. (This follows from theorem 4.1.2.). We can conclude therefore that $f'c) = 0$.

Data-Novo Vivo-010-2023/
Referencia-index/
Cold-Fusion/

✡

a) 20231118- Mallove- Fire from Ice Searching for the Truth behind the Cold

Fusion Furor - (1991).pdf
→ 342 pages.

Obs: Aprofundar com o vídeo:
20231118 - Energy - History
Channel.mp4
→ 1h 23 min.

Neste vídeo, nós temos as seguintes referências:

a) 02:26 - Nikola Tesla,
b) 02:29 - Paul Pantone,
c) 02:31 - Rudolf Diesel,
d) 02:32 - Eugene Mallóve (Mallove),
e) 02:33 - Stanley Meyer.

f) 11:21 - Philo Farnsworth
g) 11:22 - Henry Ford
h) 11:24 - Alexander Bell
i) 11:26 - Guillermo Marconi

j) 12:03 - Stanley Pons
Martin Fleischmann
(23 March 1989)
Utah University
→ Modelo de reator de fusão a frio.

k) 12:16 - Eugene Mallove
June 1990
MIT.
(Canhoto como eu).

L) 13:38- Stanley Meyer
September 1985
→ carro movido a água. (H)

M) 16:13- Motor Pantone. (GEET)
17:05- GEET -
Global Environmental
Energy Technology.

N) 19:20 - Nikola Telsa
→ Edson, Thomas → contratou
Tesla.
20:14- Energia recurso infinito
21:25- Torre Wardenclyffe
(Shoreham, New York).

O) 22:39- Norberto Keppe
23:18- Cesar Soós
23:38- Roberto Frascari
23:54- Norberto Keppe
Inventor do Motor Keppe.
Livro: A Nova Física
Da Metafísica desinvertida.

Obs: A interpretação relativística de que a "energia é matéria radiante e matéria é energia condensada" expressa muito bem o materialismo absoluto trazido por Einstein e a idéia errônea de que a energia vem da matéria.

Tesla já havia realizado inúmeras

invenções de vulto, que corresponde à ausência total de matéria, ou seja, o vácuo é feito de energia.

A tecnologia de Tesla usava ressonância de ondas eletromagnéticas que ele denominava "ondas eletromagnéticas longitudinais".

Alguns ramos não ortodoxos de física denominam a energia do vácuo responsável pela tecnologia de Tesla de "energia escalar", um termo que se refere a uma forma de energia "não vetorial", ou seja, não originária da dimensão do espaço-tempo e portanto não definível por parâmetros temporais e espaciais, como são a energia potencial e a energia cinética.

Keppe: (concluiu)

1) A matéria advém da energia e ambas não são equivalentes.

A energia (escalar de Tesla, ou essencial de Keppe) é hierarquicamente superior e sempre existe onde não há matéria (partículas), porém nem sempre há matéria – ou poderá haver – onde há energia. A matéria simplesmente capta energia essencial e desprende energia secundária segundo a ressonância de sua estrutura interna.

A

Análise:

O nada gera a energia, portanto o vácuo é a energia que cria os demais. Caso não houvesse o nada, nada teria sido criado. O nada cria o vácuo que é uma forma de energia "primordial" que permitirá a criação da matéria.

O nada existe por causa de Deus e existe antes do vácuo, o vácuo é conseqüência do nada. Aqui eu não estou falando nem de filosofia, teologia ou religião, mas apenas querendo demonstrar que o nada, criado por Deus, permitiu a criação de tudo, pois caso Deus não fizesse isso, ou seja, Ele ocupando tudo, não haveria espaço para a criação do nada e consequentemente do vácuo.

A afirmação de Keppe de que a energia sempre existe onde não há energia decorre, para mim, do nada pois o vácuo já é uma manifestação da energia criadora do nada.

O vácuo decorre do nada e o nada é muito maior que o vácuo, pois para existir o vácuo há a necessidade de existência do nada.

O Universo se expande sobre o nada, pois o nada já está lá onde o Universo, ou a borda dele, irá se expandir.

O fato de não existir nada na borda do Universo faz com que o Universo se expanda, criando-se assim, o vácuo. Conforme o Universo se expande o vácuo é criado sobre o nada e neste vácuo há energia free para o homem utilizá-la.

2) Existe apenas uma única substância, a Energia Essencial (termo keppeano para a energia escalar de Tesla), que é transcendente (não-material, não-temporal e não-espacial) e que produz a matéria através do fenômeno da ressonância.

Para Keppe, matéria é uma ressonância da Energia Essential aprisionada no tempo e no espaço.

Análise: O nada permite a criação do vácuo que contém energia, mas o nada é muito maior do que a energia "aprisionada" no espaço-tempo.

Não há uma junção energética mas sim uma decorrência da própria criação do vácuo, pois conforme o Universo se expande sobre o nada, o vácuo é criado com a sua respectiva energia e posteriormente, a matéria é criada. Por este motivo é que nós temos mais vácuo do que matéria, ou seja, o vácuo não preenche o espaço entre a matéria do Universo e sim a matéria que ocupa o vácuo previamente criado, e cria-se assim, os espaços entre as estrelas, planetas, galáxias etc.

3) Para Keppe, as energias sensíveis ao ser humano e detectáveis por instrumentos não são essenciais, correspondendo todas elas a subprodutos da energia primária (Energia Essencial). Isto implica que a tecnologia baseada na obtenção de movimento através da diferença de potenciais de energias secundárias (gravidade, eletromagnéticas etc) pode ser substituída pela tecnologia de

captação direta de energia primária (energia essencial) através do fenômeno da ressonância (da) de uma energia secundária, eletromagnética, por exemplo.

4) A energia essencial tem duas componentes, ação e complementação, que correspondem aos dois sentidos que compõem a ressonância.

Keppe usa o conceito de "Potencial Puro", correspondendo à idéia de que a diferença de potenciais da física causa entropia acelerada do sistema pelo uso da energia apenas em um sentido. Quando o sistema está em ressonância, a entropia é minimizada e a eficiência maximizada, justamente porque o sistema passa a se beneficiar dos duas componentes complementares de ação e complementação.

O potencial puro pode ser visto como um vórtice causado pela ressonância do sistema e por onde passa a energia essencial.

⟶

Retornando ao vídeo (Energia):

P) 25:00 → Artur Solis (México)
Reação química da Melanina

Melanina + Luz + Água = Energia

Obs: Produz energia por centenas de anos. Bateria de Melanina.

substância mais estável que se conhece.

25:50 → consegui com a energia da fotossíntese.

Q) 27:30 - Diesel (funciona com qualquer tipo de óleo)
↳ o motor.

R) 29:50 - Lord Akton ⇒ disse:
O poder corrompe e
o poder absoluto corrompe absolutamente.

S) 30:09 - J. D. Rockefeller.
30:17 - Standard Oil Company
Cleveland Ohio
Fundada em 1870.

T) 31:28 - Energia: Thomas Edson (CC)
x Nikola Tesla (AC)
e J. P. Morgan

31:39 - Edison Electric Power
Company → G.E.
(General Electric)

34:15 - Tesla apoio George Westinghouse
(investidor)

34:43 - Edison eletrocutou um
elefante em público
(propaganda negativa)

Coney Island, New York
04 January 1893.

35:08 - Westinghouse → colocar
os geradores nas
Cataratas do Niagara
(Falls)
(18 November 1896)

U) 36:00 - Diesel

V) 39:51 - Morte Tesla
New York
Hotel New Yorker
07 January 1943.

W) Diesel - morto em um rio
Breskens, Holanda
(09 october 1913).

X) 46:41 → Eugene Mallove
⇒ Fusão à frio.

Livro: 20231118 - Mallove -
Fire from Ice... pdf

47:02 → Modelo de Reator de
Fusão a frio.

Baseado nas pesquisas de:

47:11 - Stanley Pons
Martin Fleischmann
(23 March 1989).

Pág. 03(22) do Livro de Mallove.

48:18 - Pantone
→ Carro a hidrogênio.
September 1991.

50:00 - Mexicano - Carro Eletrico φ
Luiz Carlos Quiroz

53:25 - Mallove
(18 January 1999)
criou a sua própria organização
para pesquisar.

55:13 - Caso Stanley Meyer.
→ motor à água.
obs: Processo August 1996.
Culpado: US$ 25.000,00.

1) 1:03:56 - Paul Pantone.
→ Processador multicarburante

1:05:02 - Patent 5,794,601
Aug 18, 1998

1:06:21 - Salt Lake City Utah
12 December 2005.

1:13:00 - Pantone - Prisão
(Hospital Psiquiátrico).

1:15:13 - Paul Pantone
Libertado 25 May 2009
(Provo, Utah)

1:15:34 - Paul Pantone
Inventor do Sistema
GEET / USA.

2)
1:20:40 - MOTOR GEET.
1:20:47 - Gerador de Energia Infinita
Juan Bartet - Bartetmedia
1:20:51 - Bobina Caduceus de
Mark John Buis
1:20:53 - Motor Perendev de
Oscar Garcia Torres.
1:20:56 - Motor de Auto-Indução de
Tomas Paris Valenzuela.
20231119202100

20231120120000

Data-NovoViVo-010-2023/

Leibniz-on-the-Parallel-Postulate-and-the-Foundations-of-Geometry-The-Unpublished-Manuscripts.pdf

→ 199 pages.

Pág. 57(64) - Chapter 4

Leibniz' Attempts to Prove the Parallel Postulate.

Pág. 73(80) -

Off: Hoje não consegui escrever. A energia estava ruim e eu sem vontade de escrever.

20231120194200

20231121075000

Continuando:

Pág. 73(80) → Leibniz →

Leibniz still saw the sphere as a surface embedded in three-dimensional Euclidean space. It happens, then, that a parallel in an embedded sphere is (itself) itself (if not a great circle), and Leibniz was able to think that the equidistant line to a circle is itself a circle, even on a spherical surface (a regular figure, in his terminology).

... Leibniz missed the most promising

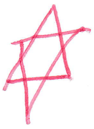

example of a non-Euclidean geometry on the spherical surface, and did not properly grasp the full meaning of a geodesic.

IMPORTANT:

Eu dou o seguinte nome a esta nova Mathmática (geometria):

MUSDESIC

M. → Mathematical
U. → uniforme
S. → Space

DESIC → division for research
division to deepen
knowledge.

Obs:

Geometria Musichiana → hipótese de ângulo hiperelíptico projectivo.

G.M.H → Geometria
Musichiana
Hiperelíptica.
or
Musdesic.

Pág. 73(80) - cont.

In the same year of 1679 Leibniz attempted to prove that the equidistant line to a straight line is straight,

while commenting on Clavius' edition of the Elements and the Jesuit's proof of the Postulate (in the last part of text 9).

Leibniz' short discussion on the topic only restates that the straightness (retidão) of the equidistant line should be proved (what Clavius just assumes, Leibniz says).[39]

[39] → Since the above-mentioned marginal notes to Clavius' edition

Obs:

Data - Novo Vivo - 010 - 2023/
Referencia - index/
Clavius - Euclidis/
20231121 - Euclidis - Clavius.pdf
→ 199 pages.

See: 20231121 - Lattis - Between Copernicus and Galileo - Christoph Clavius and the Collapse of Ptolemaic Cosmology - 2010.pdf

Pág. 03(23) - Clavius published his first edition of Euclid's Elements in 1574, ...

Pag. 05(25) - William Wallace sugests that the Pisan used Clavius's 1589 edition of Euclid's Elements while lecturing on Euclid in 1590 and thus that Clavius is one of the important links in a chain of Jesuit authors who connect medieval calculational science to Galileo.

Pág. 07(27) - Clavius → "the Euclid of his century".

Pág. 18 - Clavius departed substantially from the original, offering, for instance, demonstrations different from those of Euclid.

Later editions even introduced some topics foreign to Euclid, such as issues raised by other commentators, and even some original contributions, such as Clavius's own attempt to prove

Euclid's fifth postulate and to solve the quadrature of the circle.[70]

[70] → "Clavius and the Renaissance"

Homann - "Il quinto postulato"

Christopher Clavius and the Renaissance of Euclidean Geometry
→ Homann, S. J. (1983)

Referências complementares:

a) 20231121-000105498-documento.pdf
Human Space: the Rise of Euclidism ... 1400-1800.

b) 20231121- CAA316.pdf
The chinese Euclid and its european context.- Peter Engelfriet

c) 20231121-PCM.pdf
→ 1057 pages.
Pág. 87(110)- Geometry
5. The Western Revival of Interest

The powerful Jesuit Christopher Clavius, who edited and reworked the Elements in 1574, tried to argue that

parallel lines could be defined as equidistant lines.

d) 20231121- gt55.pdf
→ 522 pages.
Pág. 25(44) → rodapé (nota)
See Peter Engelfriet, Euclid in China: The Genesis

of the First Chinese Translation of
Euclid's Elements: - 1998.

Pág. 34(53) - rodapé

Christopher Clavius,
Euclidis Elementorum
(Rome: Vincentium
Accoltum, 1574).

e) 20231121 - ulsd736585.pdf
→ 562 pages.

Pág. 87(111) - rodapé

169 - James M. Lattis:
Between Copernicus
and Galileo:
Christoph Clavius and the
Collapse of Ptolemaic Cosmology
- 1994).

obs: Source:

20231121 - Lattis - Between
Copernicus and Galileo.pdf
→ 314 pages.

Pág. 427(451) →
Clavius Christopher
Astrolabium
Roma: Bartolomei Grassi,
1593.

Opera Mathematica.
Vol. 1 — 5 vols.
Mainz: Antonius Hierat,
1612.

Return:

Leibniz-on-the-Parallel-Postulate-and-the-Foundations-of-Geometry..pdf

Pág. 73(80) — [39] → rodapé

Since the above-mentioned marginal notes to Clavius' edition were probably written by Leibniz in 1679,

I should mention now that on Leibniz' copy of Clavius' Euclidis (p. 89), we find a handwritten remark to Clavius' proof about the straightness of the equidistant line, stating that when the Jesuit assumes that the line produced by the flow of the segment is not going up or down, he is already assuming what he would like to prove — straightness

("illud sursum vel deorsum supponit jam rectam").

Return: TEXT Pág. 73(80) →

Leibniz' short discussion on the topic only restates that the straightness of the equidistant line should be proved (what Clavius just assumes, Leibniz says),[39] and he puts forward (apresenta) an argument by symmetry.

Pag. 74(81)-

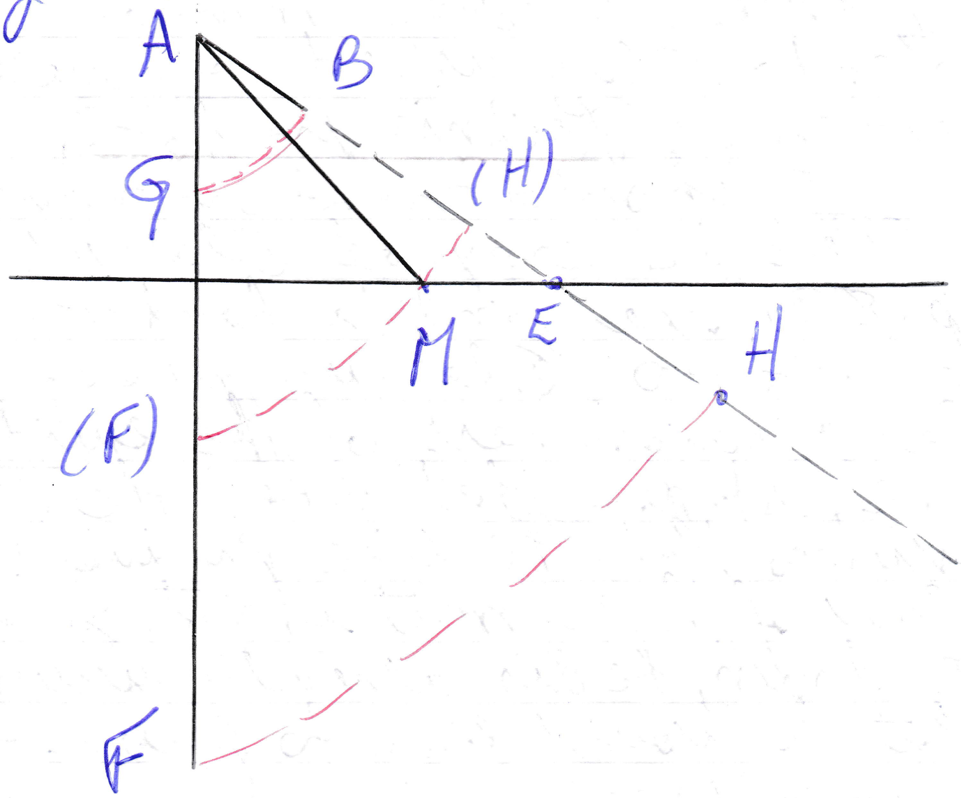

Leibniz discusses the case of two lines, AB and CD, that are intersected by a transversal AF as to form a right and an acute angle. The Parallel Postulate states that, if extended far enough, the two lines will meet.

Leibniz imagines the rotation around A of the transversal AF toward (em direção a) the position of the inclined line AB, and considers that the point F will be under the line AC when (at some point of the rotation movement) the line AF reaches the position of AB.

Since F is below CD and A is above it by a continuity assumption, the extended AB (that is, the rotated AF) meets CD at some point proving the Postulate.

If the length AF is such that when AF reaches the position of AB the point F is already above the line CD (as in the case of A(F) in the diagram) nothing can be concluded. But then, we can take a point F on AF that is far enough from A to do the job.

The same will be true for any finite length of AF.

In the case in which AB and CD are parallel, however the same reasoning cannot apply, since the intersection would occur at infinity, and AF would be a rotating unbounded line meeting the unbounded line CD at their extreme (which is impossible, since AF and CD have no ends).

Leibniz takes this as a further proof that lines cannot be moved (see text 3 above). Thus, Leibniz concludes, the present proof by rotation shows that AB and CD intersect, except in the case in which the intersection should occur at infinity; and this is the case in which AB and CD are

parallel

or form equal angles with a transversal.

Even if we accept the whole reasoning on the immovability of an infinite line

(and Leibniz himself observes that it would require more work), the last passage is unwarranted.

In hyperbolic geometry the transversal line AF, rotating, first intersects CD, than it becomes

asymptotic

with it,

and finally it reaches the line AB (assuming, I mean, that AB is parallel to CD in the hyperbolic plane, and thus the Parallel Postulate is false in this instance).

That is to say, Leibniz doesn't prove the central statement needed to establish the Parallel Postulate that the intersection of the rotating AF with CD only occurs at infinity when the angular sum with the transversal is two right angles.

Leibniz himself realized that this part of the proof was missing, and abandoned it. [43]

[43] → Leibniz' attempt was envisaged by a Medieval Hebrew mathematician (Moïse Ha-Lévi)

Obs: T. Lévy (1992) →

Gersonide, le Pseudo-Tūsī, et le postulat des parallèles.

Pág. 73(80) - ... equidistant line to a circle is itself a circle ...

an example speaking in favor of the claim that the equidistant to a straight line is itself straight (no one mentioned the conic sections)[37].

[37] →

[37] → Clavius' Lemma 1 —
to his proof of the Parallel Postulate has a long discussion about the fact that the line equidistant to a circle is itself a circle.

Leibniz himself commented on Clavius' circle argument with a marginal note to his copy of Euclidis, stating:

> "non semper lineae aequidistantes ejusdem sunt naturae; aequidistans parabolae parabola non est" (p. 90)

Interessante:
Leibniz – the great circles are the shortest and the most determined lines between two points on the sphere, and (moreover) they are 'uniforme' lines.38

[38] → In plano est recta, in sphaera est circulus maximus. (in De Risi 2007, p. 592)

→ Geometry and Monadology.

Pág. 74 (81)

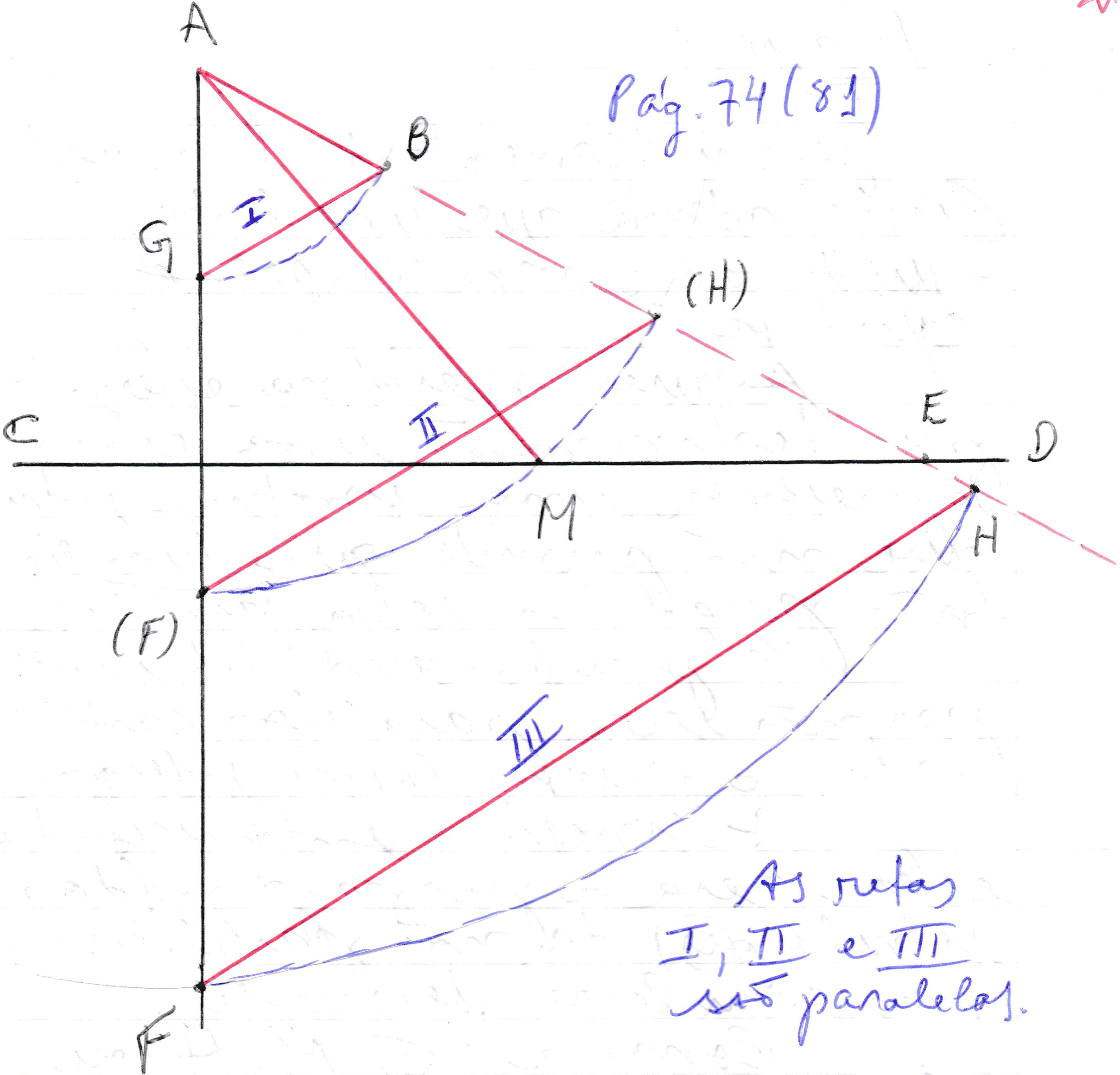

As retas I, II e III são paralelas.

Vejamos:
a) AG = AB
b) A(F) = A(H)
c) AF = AH.

Obs. Não precisa o uso de compasso.

Através da Geometria Musichiana (Hiperelíptica ou Musdesic, baseado na hipótese de ângulo hiperelíptico projetivo, AB ≠ AF (não são paralelos) no plano euclidiano, são paralelos na G.M.H ou Musdesic.

Vejamos:

Leibniz comentando Clavius' Circle admite que uma linha equidistante nem sempre é equidistante.

A questão está na esfericidade, pois teoricamente uma esfera apresenta uma simetria intrínseca que não permite que duas linhas não paralelas sejam paralelas.

O "pseudo-erro" "ab initio" por não serem paralelos contamina todas as simulações subsequentes.

A simetria intrínseca da esfera gera a propagação (das simulações) do não paralelismo nas simulações decorrentes.

Com a G.M.H ou Mundesic a hipótese do ângulo hiperelíptico projectivo no qual a dupla elipse curva permite uma real simulação da realidade demonstrando novas nuances de uma nova geometria permitindo que linhas não paralelas no plano euclidiano sejam paralelas, possibilitando a criação de novos paradigmas no mundo físico, criando novos Universos que seriam impossíveis com a

Matemática atual.

Musdesic é uma nova Matemática e consequentemente, uma nova geometria.

O ângulo hipereclíptico projetivo é um novo meio de representar a realidade, adequando o que existe ao formalismo atual matemático.

Aprofundando:

Return Pág. 74(81)

a) AG = AB

b) A(F) = A(H)

c) AF = AH

IMPORTANT:

Neste exemplo não há a necessidade da utilização de compasso para se obter o paralelismo entre as três retas.

O importante é o comprimento igual de cada segmento em ambas as retas.

AB não é paralelo ao segmento AF. (de reta).

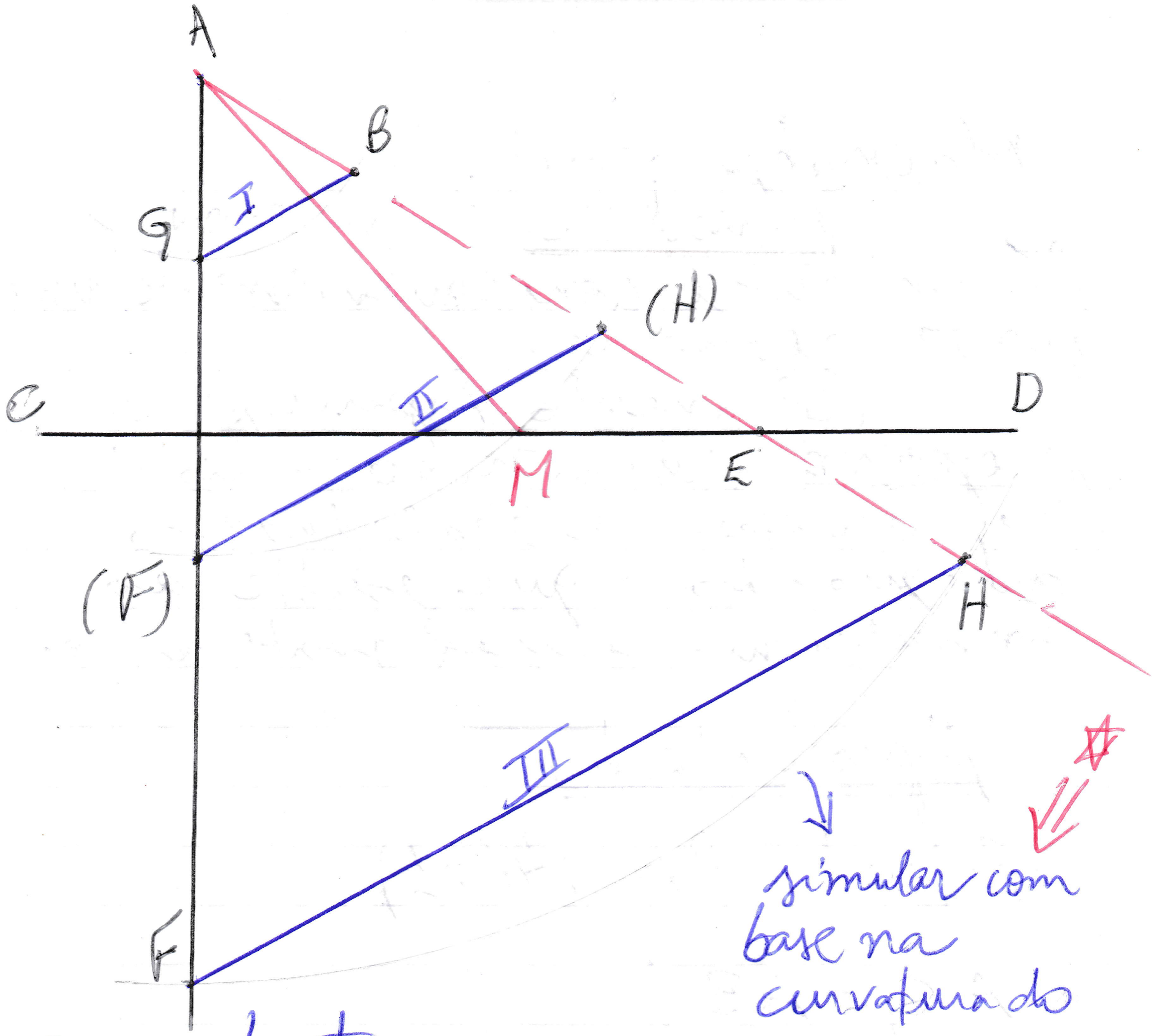

simular com base na curvatura do espaço (bending light)

Important:

Fazendo um estudo mais aprofundado, nós podemos inferir mas sem provar ainda, que AB é paralelo a AF, pois caso haja a igualdade dos comprimentos dos segmentos de retas entre as duas retas AB e AF, gerando as três retas paralelas I, II e III, nós podemos inferir que há um paralelismo intrínseco oculto entre essas duas retas.

Obs: A curvatura entre F e H, criada pelo compasso,

poderá representar em uma simulação
a curvatura da luz (bending light)
e a reta entre F H° o plano
projectivo no qual será projetado
o sistema de dupla elipse curva.

Interessante:

Este paralelismo não se pode
obter em uma esfera pois a
simetria intrínseca interna ou
externa impede que haja este
paralelismo.

Como duas retas, que a princípio,
não são paralelas no plano
cartesiano geram três retas
I, II e III paralelas.

A dupla elipse curva
permite o paralelismo que será
representado no plano no qual
será projetado.

obs: Fazer as simulações e os
cálculos.

Referências complementares:

Data-Novo Vivo-010-2023/
Referencia-index/
Clavius-Euclides/

a) 20231121 - De Risi - Geometry

and Monadology - Leibniz's
Analysis Situs - 2007.pdf
→ 646 pages.

Pág. 628 (647) - 20
On parallel Lines (1702)

b) 20231121 - De Risi - Leibniz
and the Structure of Sciences
... 2019.pdf
→ 304 pages.

c) 20231121 - De Risi -
Leibniz on the Parallel
Postulate.pdf
(copy)

d) 20231121 - De Risi -
Mathematizing Space -
the Objects of Geometry -
2015.pdf

Pág. 182 (186) - 8
Proportional Compasses

See: Pág. 184 (188) - Fig. 11

Descartes' device for trisecting
an angle. From Descartes' Private
Thoughts of 1619.

→

See: Pág. 184(188)

Fig. 12 → "Descartes' proportional compass, or mesolabe, which, among other things, allow two mean proportionals to be found between known values.

From Descartes' Geometry of 1637.

Return:

Leibniz-on-the-Parallel-Postulate-and-the-Foundations-of-Geometry.pdf

Pág. 69(76) – Leibniz

... in the discussion of asymptotic behavior in the eighteenth century.[30]

[30] → I am again especially thinking of Saccheri, who in the aforementioned Proposition 33 of his Euclides vindicatus was willing to claim the provability of the Parallel Postulate just because he takes the point of intersection at infinity between a line and an asymptote as behaving like a common point of the plane.

(Cf. also his Corollary 3 to

Proposition 26, in which he
extends further two asymptotic
straight lines after their meeting
point at infinity
Euclides vindicatus, p. 56).

Pág. 73(80) – rodapé [37]

See Saccheri's comments
and criticisms on the example
of the circle in Clavius
(Euclides vindicatus,
Scholium 3 after
Proposition 21,
pp. 39-40).

See: Data-NovoVivo-010-2023/
Foundations of Euclidean
and Non-Euclidean Geometry.pdf
→ 241 pages.
Pág. 165(180) –
Gerolamo Saccheri
(1667-1733)

Obs:
Search: Saccheri: 59 ocurren.
Parallel
Postulate: 33 ocurrens.
→ Aprofundar os detalhes.
Pág. 167(182) – Gauss, Bolyai, Lobachevsky.

20231121202000

20231122070000///
Data-NovoVivo-010-2023/
Leibniz-on-the-Parallel-Postulate-and-the-Foundations-of-Geometry...pdf
Pag. 57(64)- Chapter 4 -
Leibniz' Attempts to Prove the Parallel Postulate.

Pag. 58(65)- Consider, for instance, that the main feature of a NON-EUCLIDEAN THEORY OF PARALLELISM (in hyperbolic geometry) is the existence of approaching straight lines that do not meet (asymptotic parallels);

if one could disprove this possibility, the Parallel Postulate would be established;

but an asymptótic behavior of lines seems (parece) to depend on their curved and convex "shape" (as the usual Euclidean asymptotic lines are, obviously, not straight); a good definition of straightness (one might hope) would have dispelled (dissipado) the possibility of asymptotic straight lines, thus opening the way to a full-fledged (completo) demonstration of the Postulate.

Leibniz definitions are very numerous →

Definitions:

1) Straight line as the shortest line between two points.
(Borelli, Giordano, Fabri and Hobbes).

obj: Aristotle's statement (Phys. B 9, 200 22-4) that from the 'essence' of the straight line it follows that the sum of the angles of a triangle is equal to two right-angles.

The notion of "straightness" seems not to be captured by "shortness", and the original intent of a definition of a straight line is that of expressing the clearly intuitive idea of a LACK OF CONVEXITY OR CONCAVITY (to use Clavius' words).

Leibniz explicitly mentions the fact that the property of shortness is never used in any proof of the Elements and it is not useful to solve any foundational question.

Leibniz had a discussion with Giordano on this point, in which he appreciated that Giordano had been able to employ the definition of a straight line as the shortest line to implement his proof of the Parallel Postulate; which, being concerned with the notion of equidistance, might rely (depender) as well on the idea of the shortest distance between two lines.

2) This was an improvement, in Leibniz's eyes, on Borelli's treatment of the Postulate.

Interessante:

In the end, however, Leibniz would have preferred that Giordano had employed in his foundational efforts a definition of a straight line that Leibniz showed he knew elsewhere: that of a rotational axis.4

Pag. 60(67) – Leibniz analysis situs.

The definition of the straight line as the locus of points which remain fixed in the motion of a body with two fixed points, i.e.

the axis of rotation,

is in fact the second and most important definition of a straight line that Leibniz use in his writings on analysis situs.

Its clearest mathematical characterization is that the straight line is the only line between two points that is unique under isometries (that is, any other line connecting those two points has at least another isometric realization – imagine the reflection of that line in a plane). It is also connected with Leibniz' notion of UNIFORMITY, as the straight line so defined has the property that every segment of it is congruent to any other segment of the same length by a spatial isometry (its curvature being null).

Its importance for Leibniz is enormous, as it is by far his preferred definition ...

including those aimed at proving the Parallel Postulate.

3) The third definition of a straight line that Leibniz uses is that of a self-similar curve.

Pág. 60(67) → each segment of a line is congruent with any other of equal length, this definition says that the straight line is the only line any segment of which is similar (in the proper geometrical sense) to any other segment (of any length) of the same line.[9]

[9] → Any homothety of a factor λ applied to a curve transforms the curvature of it by a factor 1/λ. If we want a similarity (that is, a homothety and a rigid motion) to bring a segment of a curve to coincide with the whole given curve (also bounded or closed), we have λ > 1, and thus the curvature lessening at every point.

The maximal curvature of the given curve, thus, may only be reached by the transformed segment if it is zero, that is, if the curve is a straight line.

Some others:

1) Data-Novo Vivo-010-2023/
Referencia-index/
Clavius-Euclides/
20231121-De Risi-Geometry and Monadology...2007.pdf
→ 676 pages
Pág. 127 (148) – Chapter 2 - Geometry.

2) 20231121-De Risi-Leibniz and the Structure of Sciences...pdf.
→ 304 pages.

Pág. 214(220) - 6.3.
Geometrical Curves Beyond Cartesian Geometry.

Pág. 216(222) - Fig. 6.1.
The conchoid of Nicomedes generated by a Cartesian linkage.

Pág. 220(226) - Fig. 6.3
The curves of Bertet

Pág. 225(231) - Fig. 6.7
the cycloid.

Pág. 228(234) - Fig. 6.10
The parabolic trochoid

Pág. 230(236) -

$$y = \frac{e^x + e^{-x}}{2}$$

Pág. 231(237) - the relation between logarithms and hyperbolic areas appeared for the first time in the lengthy treatise

⇒

OPUS GEOMETRICUM
QUADRATURAE CIRCULI ET
SECTIONUM CONI.
(1647)
of Gregoire de St. Vincent.

3) 20231121- De Risi - Leibniz on the Parallel Postulate...pdf

4) 20231121- De Risi- Mathematizing Space.pdf
→ 320 pages.
Source: parallel postulate (34 occurrences).

Others sources:

/Clavius- Euclidis/

1) 20231121- Lattis- Between Copernicus and Galileo...2010.pdf
→ 314 pages.

2) 20231121- Rommevaux- Clavius une clé pour Euclide au XVIe siècle- Vrin-(2005).pdf
→ 158 pages.
Pag. 15 (31) - Chapitre II
Présentation Générale du Commentaire de Clavius Aux Éléments D'Euclide.

3) 20231121 - Rommevaux -
Les nouvelles théories des
rapports mathématiques ... pdf
→ 302 pages.

Pág. 280 → parallèle à la ligne.

Search: parall (14 ocorrens).

4) 20231121 - Rommevaux -
Proportions - Science - Musique -
Peint. pdf
→ 229 pages.

Search: parall (51 ocorrências)

Clavius - Euclidis - 2/

1) 20231122 - A Geometric Algebra
Invitation. pdf
→ 135 pages

Pág. 101 (109) - Chapter 5
Molecular Geometry

5.1 - Distance Geometry.

Pág. 104 (112) - Fig. 5.1
Cartesian and internal
coordinates.

Pág. 110 (118) - Fig. 5.2
Intersection between two spheres and one spherical shell

Fig. 5.3 - Arcs with their orientations.

---

2) 20231122 - Beyond Geometry - A New Mathematics of Space and Form - (2011).pdf
→ 239 pages.

Search: parall (09 ocorrências)

Pág. 91 (112) - Topological Property 2: Regularity.

In 1905 the American mathematician Oswald Veblen (1880-1960) produced a proof of the Jordan curve theorem, a famous result that states that a simple closed curves in the plane divides the plane into exactly two disjoint regions, an inside and an outside.

(A curve is called closed if it has no endpoints and is called simple if it does not cross itself. A circle is an example of a simple closed curve).

The Jordan curve theorem is famous in part because it seems so obvious and yet turned out to be so difficult to prove.

→ proof was "purely topological"!

---

3) 20231122 - Elements of the Geometry and Topology.djvu
→ 155 pages.

Elements of the Geometry and Topology of Minimal Surfaces in three-Dimensional Space.

---

4) 20231122 - Faster-than-Light.pdf
→ 266 pages.
Obs: Analisar !!!

---

5) 20231122 - Fields and Constants in the Theory of Gravitation.pdf
→ Melnikov (2002)
→ 146 pages

---

6) 20231122 - Geometric Flows-2018.pdf
→ 129 pages.

Geometric Flows and the Geometry of Space-Time.

Pág. 25(33) - 4.2 - Curvature and Holonomy of Lorentzian

Manifolds with Parallel Spinors.

Pág. 34(42) – 5 –
Constraint Equations for Special Lorentzian Holonomy

5.1. Constraint Equations for Recurrent and Parallel Vector Fields.

obs: Lorentzian manifold and a spacelike hypersurface with the induced Riemannian metric.

Pág. 37(45) – 5.2 –
Constraint Equations for Parallel Spinor Fields.

Pág. 56(64) – 7.1 –
Evolution Equations for a Parallel Lightlike Vector Field in the Analytic Setting

Pág. 97(105) – 5 –
Convex Hypersurfaces

5.1. Mean Curvature Flow.

Pág. 99(107) – 5.2

Gauss (Gauß) Curvature

Flow and Other Normal Velocities.

Pág. 103(111) – 5.4.
Two Dimensional Surfaces.

7) 20231122 – Geometry – The Language of Space and Form – (2011).pdf
→ 269 pages.

Pág. 27(48) – Euclid Reexamined
→ Euclid's parallel postulate.

Pág. 89 (110) →

Lobachevsky was fascinated with Euclid's fifth postulate. The fifth postulate, sometimes called the parallel postulate, is described in detail in the third chapter of this book. It states,
"If a transversal (line) falls on two lines in such a way that the interior angles on one side of the transversal are less than two right angles, then the lines meet on that side on which the angles are less than two right angles."
→ Heath – Great Books of the Western World, Vol. 11 – 1952

20231122185800

20231123120000

a) Data-NovoVivo-010-2023/
Referencia-index/
Clavius-Euclidis-2/
20231122-Geometry-The Language of Space and Form-(2011).pdf.
Pág. 89 (110) - Fifth postulate or Parallel postulate.

If a transversal (line) falls on two lines in such a way (de tal maneira) that the interior angles on one side of the transversal are less than two right angles, then the lines meet on that side on which the angles are less than two right angles.

b) Data-NovoVivo-010-2023/
Encyclopaedia-Britannica/
20231110-Great Books-(11)-Euclid's Elements.pdf.
Pág. 02(15) - Postulates

5. That, if a straight line falling on two straight lines make the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which are the angles less than the two right angles.

c) Data-NovoVivo-010-2023/

Euclid-Elements-13-Volumes/
/Heath-Euclid's Elements-(2002).pdf
Pág.-09(08)- Postulates.

5. And that if a straight-line falling across two (other) straight-lines makes internal angles on the same side (of itself) less than two right-angles, being produced to infinity, the two (other) straight-lines meet on that side (of the original straight-line) that the (internal angles) are less than two right-angles (and do not meet on the other side).[2]

[2] → This postulate effectively specifies that we are dealing with the geometry of flat, rather than curved, space.

d) Data-Novo Vivo-010-2023/
Euclid-Elements-13-Volumes/
Encyclopedia of Mathematics-Tanton.pdf
Pág. 170(179)- Euclid's postulates

5. If two straight lines emanating from the endpoints of a given line segment have interior angles on one given side of the line segment summing to less than two right angles,

then the two lines, if extended, meet to from a triangle on that side of the line segment.

Obs: Pág. 171 (180) - In 1795, Scottish mathematician and physicist John Playfair (1748-1819) proposed an alternative formulation of the famous fifth postulate (today known as Playfair's Axiom).

This version of the axiom is considerably easier to handle, and its negation is easier to envision.

Nicolai Ivanovich Lobachevsky (1792-1856) e
János Bolyai (1802-1860), independently came to the same surprising conclusion:

the first four of Euclid's postulates together with the negation of Playfair's version of the fifth postulate will not lead to a contradiction.

Euclid: the parallel postulate cannot be proved a consequence of the remaining postulates of Euclid.

IMPORTANT:

Bolyai's system of

✡

of hyperbolic geometry it is always the case that, for any point P in the plane, there are an infinite number of distinct lines through that point all PARALLEL to any given line not through P.

(In ordinary geometry, where the parallel postulate holds, there is only one, and only one, line through a given point P parallel to a given direction. This is PLAYFAIR'S AXIOM.)

In his new geometry angles in triangles sum to less than 180°, and the ratio of the

✡ ⇒ circumference of a circle to its diameter is greater than π.

Obs: Na minha pesquisa, a razão da circunferência do círculo pelo seu diâmetro é menor que π.

→ M.U.S. < π

$\overline{|||} < \pi$

Obs: Musdesic or G.M.H
(ângulo hiperelíptico)
→ projetivo.

Pág. 227(236) - The German mathematician Bernhard Riemann (1826-66) discovered an alternative of non-Euclidean geometry in which Euclid's fifth postulate fails in a different way.

In a system of SPHERICAL GEOMETRY it is never possible to draw a line through a given point parallel to a given line.

Riemann in his famous 1854 lecture "über die Hypothesen welche der Geometrie zu Grunde liegen" (On the hypotheses that lie at the foundation of Geometry), Riemann put foward (coloca para frente) the view that geometry can be the study of any kind of space of any number of dimensions and later developed the mathematics needed to properly describe the shape of space.

Obs: Albert Einstein (1879-1955) later used this work to develop his theory of relativity.

ATENTION:

Cuidado!!! Atenção!!!

Nós devemos prestar atenção na seguinte diferença:

Bolyai → the ratio of the circumference of a circle to its diameter is greater than π.

Na página 255 (264):

In hyperbolic geometry all angles in triangles sum to less than 180°, and the ratio of the circumference of any circle to its diameter is less than π.

(Lobachevskian geometry).

Hyperbolic geometry

(Lobachevskian geometry)

is a non-Euclidean geometry in which the famous PARALLEL POSTULATE fails in the following manner:

Through a given point not on a given line, there is more than one line parallel to that given line.

→

Jules Henri Poincaré (1854-1912) later provided a simple model for this geometry and the means to easily visualize geometric results in this theory.

The "Poincaré disk" consists of all the points in the interior of the UNIT CIRCLE.

A "point" in geometry is any point inside this circle and a "line" is to be interpreted as a circular arc within the circle with endpoints perpendicular to the boundary of the circle.

Any diameter of the boundary circle is also considered a line.

Distances are not measured with a traditional ruler: points on the boundary circle are considered to be infinitely far from the center of the circle.

Bolyai and Lobachevsky showed that all but the FIFTH OF EUCLID'S POSTULATES hold in the hyperbolic geometry and, moreover, that this model of geometry is consistent (that is, free of CONTRADICTIONS).

This establishes that the parallel postulate cannot be logically deduced as a consequence of the remaining axioms proposed by Euclid.

In hyperbolic geometry, all angles in triangles sum to less than 180° and the ratio of the circumference of any circle to its diameter is less than π. (Moreover, the value of this ratio is not the same for all circles). Also, it is possible for two perpendicular lines to be parallel to the same line.

Physicists, following the work of Albert Einstein, suggest that the geometry of our universe is hyperbolic: that it appears to us as Euclidean is a result of the fact that we occupy such a small portion of it. (This is analogous to the fact that it is difficult to recognize the Earth as around when living on it).

Pág. 355 (364) – non-Euclidean geometry.

non-Euclidean geometry :

1795 - John Playfair (1748-1819) presented an alternative, but equivalent, formulation of the parallel postulate:

through any point in the plane, there is precisely one line through that point parallel to any prescribed direction.

Recasting the postulate this way makes it apparent that negation of the famous fifth postulate has two parts. Either:

1. There are no lines through a given point parallel to a given direction.
2. There is more than one line through a given point parallel to a given direction.

From 1826 to 1829, Nicolai Ivanovich Lobachevsky (1792-1856) developed a consistent theory of geometry in which Euclid's fifth postulate fails in the manner described in point 2. → Hyperbolic Geometry.

In the 1850s, Georg Friedrich Bernhard Riemann (1826-1866) presented an alternative form of geometry in which Euclid's fifth postulate fails in the manner describe in point 1 above.

In this theory, today called Riemannian geometry or Spherical Geometry, "lines" are great circles drawn on the surface of a sphere. Consequently, it is impossible to draw a pair of lines that never intersect.

Euclid's second postulate seems to state that straight lines should be of infinite length.

Riemann had the insight to note that "extended indefinitely" does not imply "infinitely long", thus allowing him to consider a geometry in which straight lines loop back on themselves.

In hyperbolic geometry, all angles in a triangle sum to less than 180°, and the ratio of the circumference of a circle to its diameter is greater than π.

In spherical geometry, all angles in a triangle sum to more than 180°, and the ratio of the circumference of a circle to its diameter is less than π.

Euclidean Geometry, in which angles in a triangle always sum exactly to 180°, and the ratio of the circumference to diameter of any circle is precisely π (wrong!!!!) M.U.S., can thus be regarded as an intermediate between the two.

Resume:

HG → Hyperbolic Geometry.
EG → Euclidean Geometry.
SG → Spherical Geometry.

Portanto:

inversamente proporcional

$$HG \rightarrow \Delta \boxed{<} 180^\circ, \quad \frac{C}{D} \boxed{>} \pi.$$

$$EG \rightarrow \Delta = 180^\circ, \quad \frac{C}{D} = \pi.$$

$$SG \rightarrow \Delta \boxed{>} 180^\circ, \quad \frac{C}{D} \boxed{<} \pi.$$

inversamente proporcional

(mirror) bipolar invertido.

obs: Atenção!!!

Análise dos sinais:

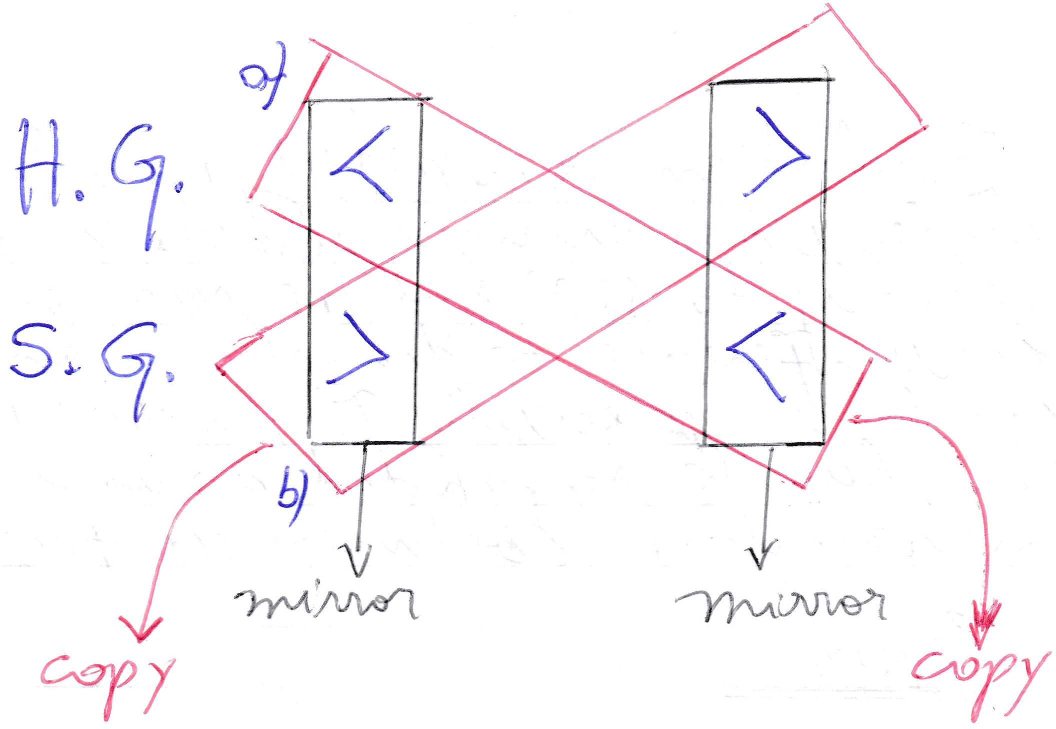

Important: Análise dos possibilidades. (estudar a viabilidade)

a) A soma dos ângulos é menor que 180° gerando assim, $\frac{C}{D} < \pi$, ou seja, gera um número menor que π.

b) A soma dos ângulos é maior que 180° gerando assim, $\frac{C}{D} > \pi$ ou seja, gera um número maior que π.

Obs: Estudar a viabilidade de uma razão direta entre a soma (valor) dos ângulos de um triângulo e o valor obtido entre C/D, ou seja, o valor do "π"!

Pág. 375(384) - Parallel L∥M.

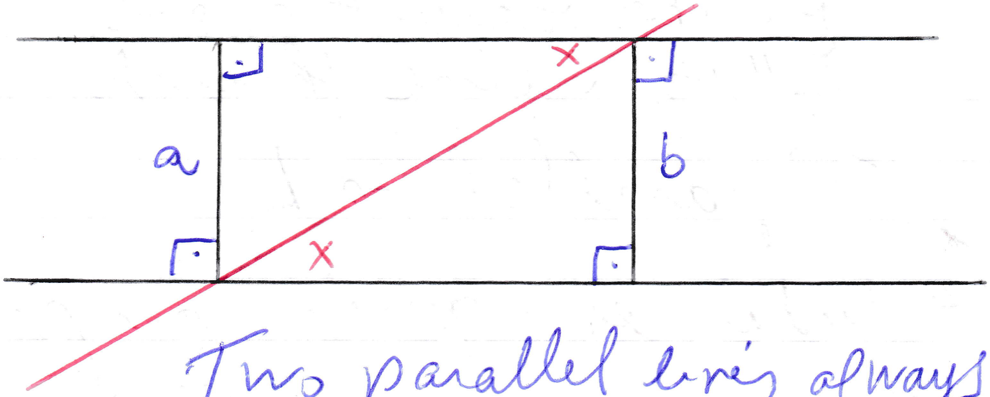

Two parallel lines always remain a fixed distance apart.

→ length a equals length b.

Obs:

Observe that the two triangles in the diagram share the same angles and a common diagonal length.

The ASA principle now assures that they are CONGRUENT FIGURES. Consequently, the distances a and (b must be equal).

Interessante:

To produce a line through a given point P parallel to a given line L, draw an arbitrary line L' through P that intersects L. (prestar atenção nesta construção geométrica → interessante!!!).

Measure the angle L and L', and draw another line L'' through P, making this angle with L' so as to create a diagram with equal alternate

interior angles.
By the result above,
L" is a line through
P parallel to L.

↑ Estudar esta questão com profundidade.

Obs: Analisar a posição Playfair:

John Playfair (1748-1819) proved that for any given point P and line L in the plane, there is one and only one line through P parallel to L.

This result is today known as PLAYFAIR'S Axiom.

Playfair proved that one would take this axiom as a starting point for geometry and establish the parallel postulate from it.

Pág. 377 (386) —

Parallel Postulate
(Euclid's fifth postulate)

→ continuar em outra data.

20231123211700

20231124093000M

a) Encyclopedia of Mathematics-Tanton.pdf
Pág. 377(386)-
Parallel Postulate (Euclid's fifth postulate)

His fifth postulate states:

If two lines L and M subtended from a common line segment have interior angles summing to less than 180°, then those lines, if extended, will meet to form a TRIANGLE.

One can consequently conclude that if those angles sum to more than 180° then the lines, if extended on the other side of the given line segment, will form a triangle on that side.

(The sum of the supplementary interior angles will be less than 180°.)

Furthermore, if the lines never meet, that is, if they are PARALLEL, then the sum of the two angles x and y shown cannot be less than nor greater than 180°, and so must equal 180°. We have:

For a pair of parallel lines with a TRANSVERSAL, as shown in the diagram below right, the angles labeled x and y must sum exactly to 180°. Consequently, the alternate interior angles of transversal are equal.

This is often taken as the statement of Euclid's fifth postulate. For this reason, it is (offen) often called the parallel postulate.

As the parallel postulate is intimately connected with the formation of triangles, it is not surprising that one can use it to prove:

The sum of the interior angles of any triangle equel 180°.

1795 – "Elements of Geometry"
→ John Playfair offered a third, logically equivalent version of the parallel postulate:

Given a line L in the plane and a point P not on that line, then there exists one and only one line through P parallel to L.

This alternative version of the postulate is today called PLAYFAIR'S AXIOM.

1829 → Nikolai Ivanovich
Lobachevsky
(1792 – 1856)
→ discover a consistent theory

of a Non-Euclidean Geometry.

We now have three assertions:

1. Euclid's parallel postulate;
2. Angles in a triangle sum to 180°;
3. Playfair's axiom.

These three statements are logically equivalent, that is, any one implies the other two. To see this, it suffices to show that (3) implies (2), (2) implies (1), and that (1) implies (3).

We will need to make use of the converse of Euclid's parallel postulate:

If a transversal across a pair of lines yields equal alternate interior angles, then the lines are parallel.

As the EXTERIOR-ANGLE THEOREM, this statement can be proved without making use of any of the three statements above.

Consider now the three implications:

(3) implies (2):

Given a triangle with angles

a, b, and c as shown in the diagram on page 378,

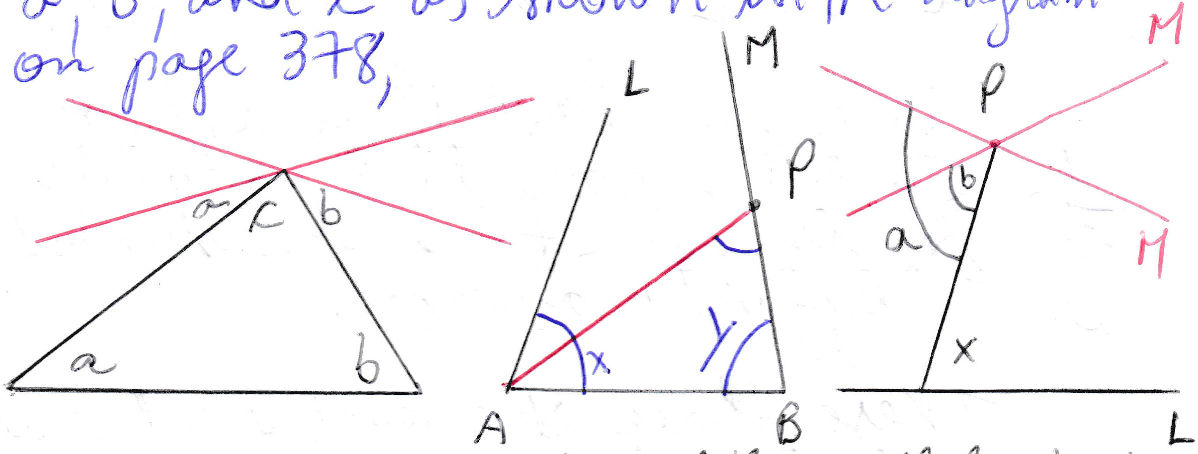

Equivalent versions of the parallel postulate.

... 378, draw two lines at the apex that copy angles a and b. Given that both yield equal alternate interior angles across a transversal, both lines are parallel to the base of the triangle. By Playfair's axiom there is only one such ~~lime~~ (line). Thus angles a, b, and c all lie on the same straight line and so sum to 180°.

(2) implies (1):

In the middle diagram on page 378, suppose that lines L and M are parallel, and so never meet. Could it be that angles x and y sum to less that 180°? Draw a point P on line M, and slide it along M sufficiently far so that the angle of the triangle APB at P is less than 180 - x - y.

Clearly the angle in this triangle at A is less than x.

We now have a triangle with three angles that sum to something less than $x-(180-x-y)+y=180°$.

This is impossible by our assumption that (2) is true. It cannot be the case that L and M are parallel after all.

(1) implies (3):

Discussion on the exterior-angle theorem shows that there certainly exists at least one line through a point P parallel to a given line L.

Suppose there are two such lines. Call them (them) M and M' as shown in the diagram above.

Draw any line from P to L.

Since, by (1), we are assuming that alternate interior angles are equal for parallel lines, both angles a and b show are equal to x.

Consequently M and M' must be the same line.

Pág. 473 (482) → Spherical Geometry

Spherical Geometry (elliptic geometry, Riemannian geometry)

Discovered in 1856 by Georg Friedrich Bernhard Riemann (1826-66), and later slightly modified modified by Felix Klein (1849-1925), spherical geometry is a non-euclidean geometry in which the famous

PARALLEL POSTULATE

fails in the following manner:

Through a given point not on a given line there are no lines parallel to that given line.

Riemann used the surface of a SPHERE as a model of this geometry by interpreting the word line to mean a great circle on the sphere. Given that in a theory of geometry two lines are meant to intersect at just one point (yet any two great circles intersect at two ANTIPODAL POINTS), it is appropriate then to interpret the word point in spherical geometry as an antipodal pair of points on the surface. In this setting, it is now also true that

any two distinct points determine a unique line.

Riemann and Klein proved that all but the fifth of EUCLID'S POSTULATES hold in this model and, moreover, that this model is consistent (that is, free of CONTRADICTIONS).

This establishes that the parallel postulate cannot be logically deduced as a consequence of the remaining axioms proposed by Euclid.

In spherical geometry all angles in triangles sum to more than 180°, and the ratio of the circumference of any circle to its diameter is greater than π (and this value varies from circle to circle).

Obs: Checar!!! ... greater than π.

$\Delta > 180^\circ$ and $\frac{C}{D} > \pi$ (?)

(atenção!!!)

IMPORTANT:

Pág. 46(55) – Bolyai, János (1802 – 1860) is remembered for his 1823 discovery of Non-Euclidean geometry.

...

In Bolyai's system of HYPERBOLIC GEOMETRY it is always the case that, for any point P in the plane, there are an infinite number of distinct lines through that point all PARALLEL to any given line not through P.

(In ordinary geometry where the parallel postulate holds, there is only one, and only one, line through a given point P, parallel to a given direction. This is PLAYFAIR'S AXIOM)

In his new geometry, angles in triangles sum to less than 180° and the ratio of the circumference of a circle its diameter is greater than π.

b) Data-Novo Vivo-010-2023/
Euclid-Elements-13-Volumes/
Coolidge-The Elements of
Non-Euclidean Geometry-
Oxford (2009).pdf/
→ 316 pages.

Pág. 170 - Chapter XIV
Areas and Volumes

The subjects area and volume offer some of the most striking points of disparity between Euclidean and non-Euclidean geometry.

Pág. 174 - In Chapter III we defined as the discrepancy of a triangle, the absolute value of the difference between the sum of the measures of the angles and π. Let us now define as the excess of our triangle the expression

$$\boxed{e = A + B + C - \pi}.$$

This will have the same sign as $\frac{1}{k^2}$, the measure of curvature of space.

We have

$$\sin\frac{e}{2} = -\cos\sigma = \frac{\sin(ABC)}{4\cos\frac{1}{2}\frac{a}{k}\cos\frac{1}{2}\frac{b}{k}\cos\frac{1}{2}\frac{c}{k}}$$

(14)

$= 4$

$$\lim e = \frac{1}{2} \lim (ABC)$$
$$= \frac{1}{2k^2} \lim bc \sin A$$
$$= \frac{1}{2k^2} \lim a \overline{AA'}.$$

Pag. 178 – theorem 5.

The area of a convex polygon is the quotient of the excess of the sum of its angles over $(n-2)\pi$ divided by the measure of curvature of space.

Let the reader show that the area of a proper circle is

$$2\pi k^2 \left(1 - \cos \frac{r}{k}\right) \quad (17)$$

The total areas of the elliptic and the spherical planes will be

$$2\pi k^2, \; 4\pi k^2.$$

In the hyperbolic plane regions may be found having any desired area.

Pág. 175 → Theorem 3

The area of a region of a plane is the sum of the areas of any two regions into which it may be divided provided that these two have no common area.

Pág. 177 – Area = $k^2(A+B+C-\pi)$ (16)

Obs: Atenção!!!

Data-NovoVivo-010-2023/
Euclid-Elements-13-Volumes/
Coolidge/

a) 20231124-Coolidge-A History of the Conic Sections and Quadric Surfaces-(1968).pdf

b) 20231124-Coolidge-A treatise on algebraic plane curves-(1959).pdf
... -(2004).pdf

c) 20231124-Coolidge-the Elements of Non-Euclidean Geometry-(1909).pdf

d) 20231124-Coolidge-the Mathematics of Great Amateurs-(1990).pdf

e) 20231124-Sommerville (Sommerville)-the elements of non-Euclidean geometry-(1919).pdf

c) Data-NovoVivo-010-2023/
Referencia-index/
Clavius-Euclidis-2/

20231122-Non-Euclidean Geometry in Materials of Living and Non-Living Matter-2022.pdf
→ 294 pages.

Pág.02(13)-Chapter 1
Euclidean Geometry and Non-Euclidean Geometry

Pág.03(15)-Fig.1.1.
On the axiom of parallelism of straight lines.

Pág.02(14)-1.1. Euclidean Geometry
the basis of Euclid's geometry is five postulates (axioms that do not require proof):

5. If a straight line meets two other straight lines in such a way that on one-sided they form two internal one-sided corners, in total less than two right angles, then these straight lines, if continued

indefinitely, will intersect on the side from which the one-sided corners are less than two right angles.

From the fifth postulate it follows that if there are two parallel lines, i.e., the straight line intersecting them forms two inner one-sided angles to the sum of two straight lines, then these parallel straight lines will not intersect no matter how much they continue.

Euclidean geometry, according to which space is infinite, homogeneous, isotropic and three-dimensional, became the basis of classical physics.

However, the fifth postulate of parallel straight lines has been in doubt since the 19th century. It is easy to make such a construction (Figure 1.1). We will draw two parallel lines, and also in the plane of these lines we will draw a circle C of an arbitrary radius centered at point A between these lines so that the circle does not intersect straight lines.

→

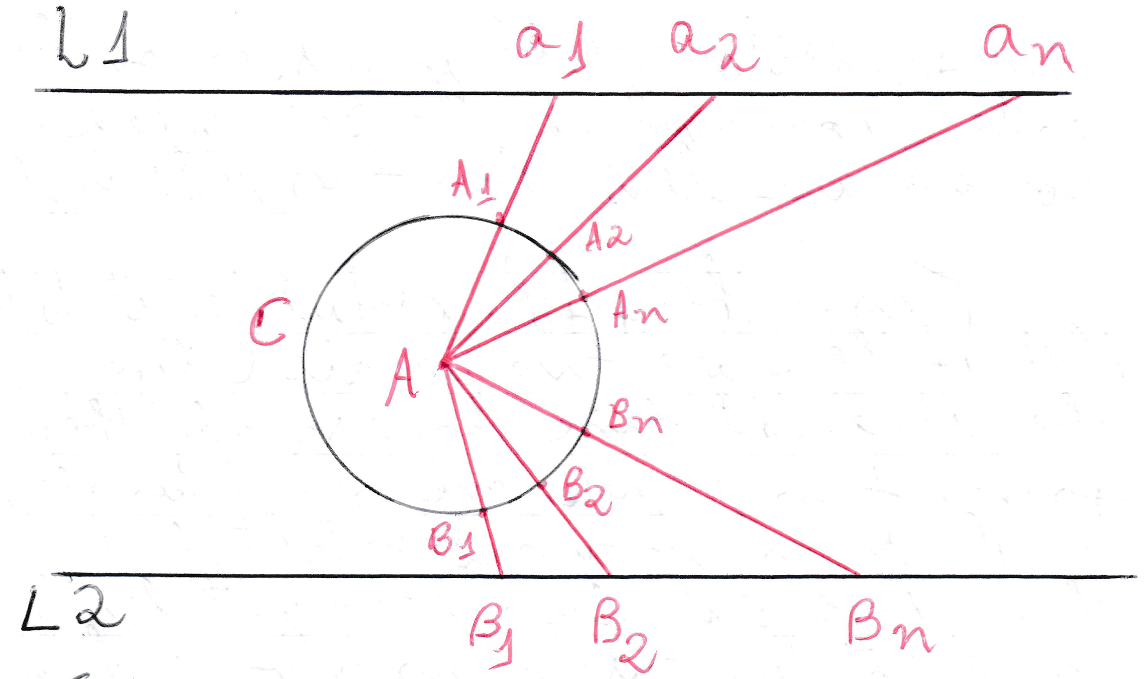

Figure 1.1 On the axiom of parallelism of straight lines.

We will successively decrease the angles of inclination of the rays to zero in both sets of rays. In this case, the points α_n and β_n will move along straight lines L_1 and L_2 right to infinity.

The images of these points A_n and B_n on the circle C will approach each other as closely as you like. At a zero angle of inclination of the rays, the point A_n will coincide with the point B_n. Since there is a one-to-one correspondence between the points α_n and A_n, as well as between the points β_n and B_n, the alignment of the points A_n and

β_n will mean the alignment of the points α_n and β_n, that is parallel lines L_1 and L_2 intersect at infinity.

The inconclusiveness of the postulate of parallel lines led to the construction of geometries as a whole on the basis of concepts and axioms that relate only to the immediate vicinity of each point. So, the possibility arose of creating geometries based on differential geometry (Hilbert, Cohn-Vossen, 1932).

20231124203000

20231125095000

Data-Novo Vivo-010-2023/
Referencia-index/
Clavius-Euclides-2/
20231122-Non-Euclidean Geometry in Materials of Living and Non-Living Matter-2022.pdf

Pág. 04(16)- 1.2.

Non-Euclidean Hyperbolic Geometry.

In non-Euclidean geometry postulate 5 of the Euclidean system of postulates is replaced by the axiom:

Through a point that does not lie on a given straight line there are at least two straight lines that lie with this straight line in one plane and do not intersect it. this is Lobachevsky's axiom (Lobachevsky, 1829, 1835).

According to this axiom, on the plane through the point P, which lies outside the straight-line A' A, more than one straight line (Figure 1.2) does not intersect the straight-line A'A.

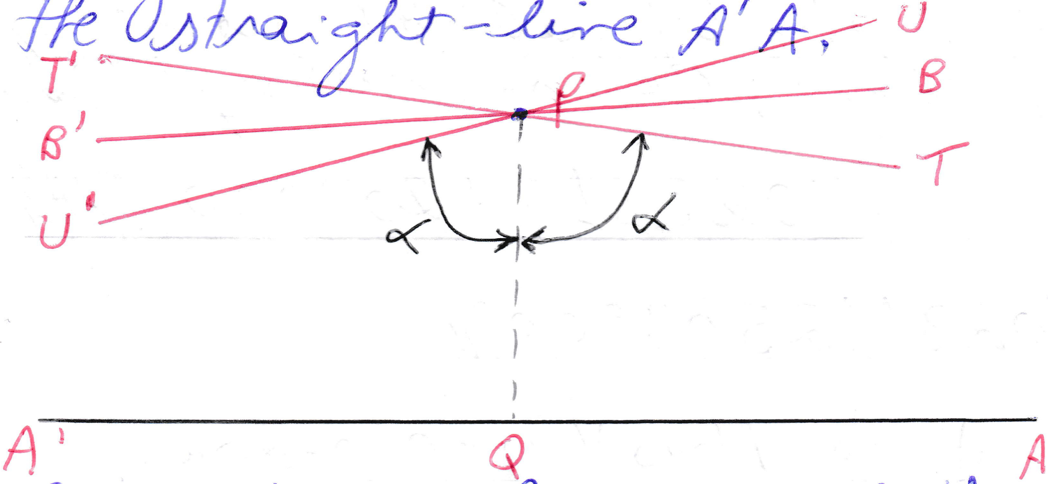

Figure 1.2. – Lines passing through a point in non-Euclidean geometry.

Non-intersecting straight lines fill the part of the beam with apex P lying inside the vertical angles TPU and U'PT' located symmetrically relative to the perpendicular PQ!

The straight lines forming the sides of the vertical corners separate the intersecting straight lines from the non-intersecting straight lines and are themselves intersecting straight lines. These boundary lines are called parallel at point P to line A'A, respectively, in its two directions: T'T parallel to A'A in the direction of A'A and U'U parallel to A'A in the direction of A'A. The rest of the non-intersecting lines are called lines diverging from A'A.

The angle α $(0 < \alpha < \pi/2)$, that the parallel to the point P forms with the perpendicular PQ,

$$\boxed{\measuredangle QPT = \angle QPU' = \alpha}$$

is called the angle of parallelism of the segment $PQ = a$ and is denote by $\alpha = \Pi(a)$. When $a=0$ the angle $\alpha = \pi/2$; with increasing a angle α decreases so that for each given α there is a certain value of a. This dependence is called the Lobachevsky function

$$\Pi(a) = 2 \operatorname{arctg}(e^{-a/K}), \quad (1.1)$$

where K is some constant that determines a segment fixed in magnitude. It is called the radius of curvature of the Lobachevsky space.

There is an infinite number of Lobachevsky spaces, differing in the size of JK.

The model of the Lobachevsky plane can be a surface with constant negative curvature.

The simplest example of such a surface would be the pseudosphere (Figure 1.3) (Beltrami, 1868).

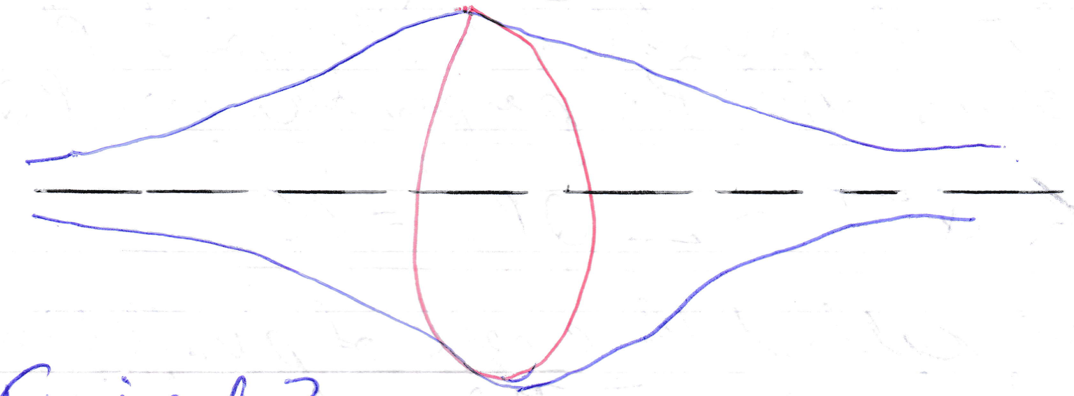

Figure 1.3. Pseudosphere.

There are also other models of Lobachevsky geometry (Klein, 1928).

Pág. 05 (17) — 1.3

Non-Euclidean Eleptical Geometry.

The fact that the sphere is the

spatial form of elliptic geometry is seen directly, since the curvature of a sphere of Radius R in all such cases is equal to $1/R^2$, that is, it is positive and constant.

It can be shown that a metric of constant curvature $K > 0$ can also be introduced on the projective plane. Consider some plane α of Euclidean space and some sphere C. Let the plane of Euclidean space be located horizontally, and the sphere is above the plane.

Let O be the center of the sphere. Obviously, each point of the sphere, if it does not lie on the equator, is projected from the center O to a certain point of the plane α. Diametrically opposite points of the sphere are projected at the same point. The points lying on the equator have no projections. Supplement the plane α with infinitely distant points and introduce a system of neighborhoods into the set of all points of the extended plane. Then we get a projective plane, which we will denote by $\bar{\alpha}$.

Let us agree to say that the horizontal line a passing through the point O intersects the projective plane $\bar{\alpha}$ at an infinitely distant point lying on any of the straight lines in the Euclidean plane α parallel to the line a. Under this condition, each point of the sphere is projected from the center O to a certain point of the projective plane $\bar{\alpha}$, and each point on $\bar{\alpha}$ is the projection of two diametrically opposite points of the sphere.

Let us now introduce geographic coordinates on sphere $\odot$. On the projective plane, the coordinates of some point M will be considered numbers equal to the geographical coordinates of that point of the sphere, which has its projection point M. Since each point of the projective plane is the projection of two opposite points of the sphere, then each point of the projective plane, along with coordinates, also has coordinates. Now, in each neighborhood of a point on the projective plane, we introduce the metric form

$$dS^2 = R^2(\cos^2\psi\, d\varphi^2 + d\psi). \quad (1.2)$$

In this case, the matrix shape of the sphere is written in the form (Efimov, 1945).

$$dS^2 = \cos^2(K^{1/2} v)\, d^2 u + d^2 v \quad (1.3)$$

→

To reduce expression (1.2) to expression (1.3), it is sufficient to set

$$\varphi = K^{1/2} u, \psi^{1/2} v$$

and take into account the relationship between the curvature and the radius of the sphere

$$K = 1/R^2.$$

The geometric system of the metrized projective plane α is called Riemannian non-Euclidean geometry.

In the (the) geometric system of the metrized projective plane α (elliptic plane), any two points define one and only one geodesic line passing through them.

Every two lines have a common point. An elliptical plane has no parallel lines. In an elliptical plane (Riemann geometry), every line intersects another.

The well-known objects of two-dimensional differential geometry can be generalized into multidimensional geometry (Efimov, 1945; Cartan, 2009).

off: Efimov, 1945 - Higher Geometry.

Cartan, 2010 - Geometry
Riemann's spaces.

Source: Data-NovoVivo-010-2023/
Euclid-Elements-13-Volumes/
Cartan/

a) 20231125-Cartan-Geometry of Riemannian spaces-(1983).pdf

b) 20231125- Cartan- Riemannian geometry in an orthogonal frame-(2001).pdf

c) 20231125- Cartan for Beginners - Differential Geometry Via Moving Frames and Exterior Differential -(2003).pdf

cont.) ...

Riemann introduced the concept of an n-extended quantity and defined a measure definition that is possible on an n-dimensional manifold

$$dS^2 = \sum g_{ik}\, dx_i \cdot dx_k,$$ where

the coefficients in front of the product of the differentials of the variables are functions of coordinates.

Pág. 07(19)- 1.4.

Non-Euclidean Geometry of the Higher Dimension Polytopes!

Riemann pointed out the contours of geometric figures as one of the examples of the n-extended value.

The space of these figures, therefore, has dimension n and is limited, i.e., certainly.

The condition for space to be finite is one of the basic conditions in Riemann geometry.

However, the curvature of space will be zero if the planar (two-dimensional) elements of the figure have zero curvature. Thus, there is no need to use the apparatus of differential geometry. Due to the finiteness of the higher-dimensional polytope and the presence of the polytope contour (i.e., the boundary), the higher-dimensional polytope space is closed. Therefore, the system of axioms of n-dimensional Euclidean space, extended from the system of axioms of three-dimensional Euclidean space, cannot be applied to the polytope space of higher dimension. In particular, line segments located inside the higher-dimensional polytope cannot be extended outside the higher-dimensional polytope. The space of a polytope of the highest dimension has a special structure that determines the conditions for the belonging of elements of different dimensions to each other, as will be shown in the next chapter. Therefore, the conditions for membership of elements of different dimensions adopted in the system of axioms of n-dimensional Euclidean geometry, are not fulfilled in the polytope of the highest dimension.

In this regard, it should be recognized that the geometry of the space of polytopes of the highest dimension is not Euclidean.

Conclusion:

This non-Euclidean geometry is not related to the curvature of space and therefore cannot be attributed to either hyperbolic or elliptical non-Euclidean geometries.

Pág. 211(223) - 8.4 -

The Three-Dimensional Model of the Nucleic Acid Molecule.

... radius of a circle

$(\pi - \gamma) r = 2\pi$ - Fig. 8.24

20231125203000

20231126142000m
Data-NovoVIVO-010-2023/
Referencia-index/
Clavius-Euclidis-2/

a) 20231122-Non-Euclidean Geometry in Materials 2022.pdf

Pág. 206(218)- Fig. 8.19

Two α-D-ribose molecules of the form B in the structure of the EB.

The vertex b in this figure corresponds to the phosphorus atom, and the two edges emanating from it correspond to the projections of two valence bonds of the atom phosphorus with two oxygen atoms. They are at some angle γ on the projection. This angle characterizes the location of the residual phosphoric acid in space with respect to the α-D-ribose molecule B at the side of their connection (vertex K). It can be shown that the angle γ is linearly related to the angle $\Delta\alpha$ between two neighboring nuclei in the spiral chain. Indeed, in Figure 8.19, the smallest distance between two corresponding points of two neighboring α-D-ribose molecules B in the chain Δh corresponds to the segment OC.

...

$$\Delta\alpha = 126^\circ - \delta - \gamma - (54^\circ - \delta) = \frac{\pi}{5} - \gamma.$$

(Pág. 207(219)).

Pág. 208(220) -

The lengths of chemical bonds between carbon-carbon and carbon-oxygen are ≈ 0.15 nm, can will adopt $a = 0.15$ nm. The length of chemical bond phosphor - oxygen can adopt $d = 0.18$ nm.

See: Pág. 209(221)

Figure 8.20

Graphs of the change in the distance between nucleotides Δh, the angle of rotation between adjacent nucleotides $\Delta\alpha$, the number of nucleotides in the period N and the radius of the spiral

$r = \frac{N\Delta h}{2\pi}$, depending on the angle γ in the projection between the valence bonds of the phosphorus atom.

See: Table 8.1

The characteristic dimensions of the right and left nucleic acid helices.

Pág. 210(222) - Figure 8.21

Two nuclei of the nucleic acid structure of the EB for $\gamma = 90$ degrees.

Pág. 211(223) –
then the radius of the nucleic acid helix is $r + \Delta r \simeq 1$ nm.

8.4 The Three-Dimensional Model of the Nucleic Acid Molecule.

Pág. 212(224) – Figure 8.22
See: curvature.

A nucleic acid molecule with sugar molecules α-D-ribose (A) or and 2-deoxy-α-D-ribose (A).

Pág. 213(225) – Figure 8.23.

Pág. 214(226) – $(\pi - \gamma)n = 2\pi$.

Pág. 215(227) – Figure 8.24
The projection of spiral on the plane perpendicular to axes of rotation.

$l = 2d + 3a$

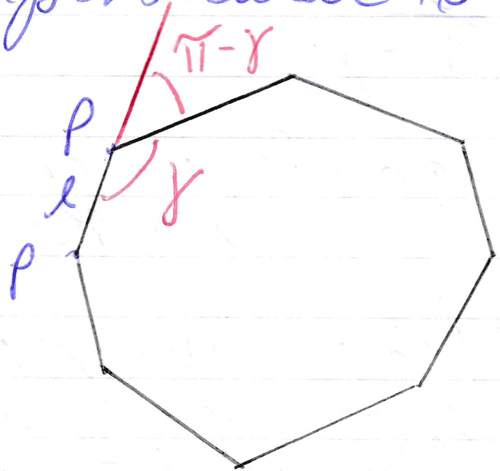

Since the number of turns equals the number of segments that make up a polygon, one have approximately the equality

$$\frac{2\pi}{\pi-\gamma} = \frac{2\pi R}{l},$$

R is radius of the circle, and l is length of the segment centers of tetrahedrons.

From the last equality can find the radius of the circle

$$R = \frac{l}{\pi-\gamma}.$$

From Figure 8.22 and Figure 8.24 it follows that $l = 3a + 2d$. Subject to the accepted values a and d can get

$$R = \frac{0.81}{\pi-\gamma} \text{ nm}.$$

For example, in case $\gamma = 140°$ is $R = 1.6$ nm.

Obs. This radius value is close to the experimental helix radius of nucleic acid measured by Watson and Crick.

b) 20231122 - Sochi - General Relativity Simplified - (2020).pdf

Pag. 22 (23) -

A space (ust) is flat if it is possible to find a coordinate system for the space with a diagonal metric tensor whose all diagonal elements are ±1; otherwise the space is curved.

More formally, an nD space is flat if it is possible to find a coordinate system in which the quadratic form $(ds)^2$ is given by:

$$(ds)^2 = \sum_{i=1}^{n} \varepsilon_i (dx^i)^2 \qquad (2).$$

where the indexed ε are ±1 while the indexed x are the coordinates of the space.

Geodesic curve is the most "straight" or most "direct" route connecting two points in a given space.

Geodesic curves are normally, but not necessarily, the curves of shortest length.

Interessante : NULL GEODESIC.

Null geodesic is the geodesic path that is followed by a light signal (or any massless object to be more general) in the spacetime (assuming the signal is not under non-gravitational influence). The line element $d\sigma$ (and hence the spacetime interval σ) vanishes identically along null geodesic and that is why it is "null".

Pág. 290(291) – Figure 7.

A simple graphical illustration of the precession of perihelion PP where P_1, P_2, P_3 represent successive perihelia. We note that the indicated precession can have various causes (and hence we use PP instead of $\delta\phi$).

8.2 Light Bending by Gravity.

Light bending (or light deflection) by the gravitational field of a gravitating body like the Sun, is one of the major predictions of general relativity because according to this theory the bending of spacetime by matter and energy should result in bending the geodesic paths of the spacetime and hence the physical objects, whether (se) massive like material particles or massless like photons, which follow these geodesic paths in their free fall (i.e. in the absence of forces other than gravity) should move along curved trajectories.

Analizar.

The bending of light in gravitational fields may also be justified qualitatively by the equivalence principle since light bends in accelerating frames and hence it should also bend in gravitational frames. (frames).

In fact, light bending by gravity can be derived even (mesmo) from the Newtonian theory and hence it is not an effect that is specific to general relativity.

However, the Newtonian prediction is quantitatively different from the prediction of general relativity.[368]

[368] ⇒ commonly recognized Newton prediction which is based on derivation. However, some theories based on the Newtonian gravity seem

to have diffent predictions;

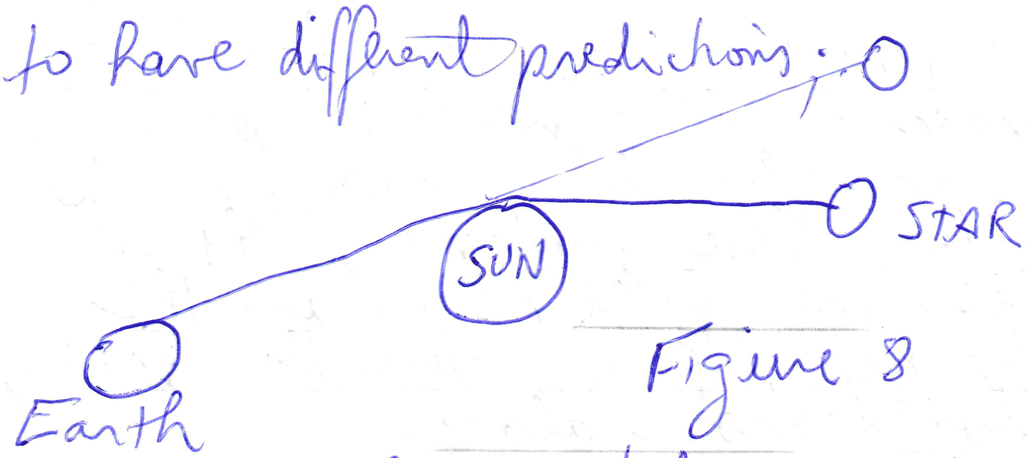

Figure 8

A simple sketch depicting the bending of a light ray from a star as it passes by the Sun toward (in direction) to the Earth and hence it is deflected by the gravity of the Sun.

Using the geodesic equations of null geodesic in the Schwarzschild metric, it can be shown that the deflection of light by the gravity of a static and spherically symmetric gravitating body is a given

$$\delta \phi = \frac{4GM}{c^2 d} \qquad (323)$$

where $\delta\phi$ is the angle of deflection in radians,
G is the gravitational constant,
M is the mass of the gravitating body,
c is the characteristic speed of light and

d is the distance of closest approach of light to the center of the gravitating body.

Pág. 291(292) - Problems

1. The problem of light bending is a problem of motion along geodesic trajectories in spacetime because the light is in free fall (?) under the influence of gravity alone.

Therefore, this problem should be investigated as a geodesic problem in the Schwarzschild metric and hence we can use the results that we obtained in § 4.1.2.

Obs: § 4.1.2. Pág 215(216) →

Geodesic Equation in Schwarzschild Metric.

→ using the Christoffel symbols (§ 4.1.1).

Obs: Pág. 212 (213) - Schwarzschild radius?

Schwarzschild radius (which may also be called gravitational radius and is usually associated with black holes although (embora) it is not specific to them) is the distance from the center of the gravitating object to the points of space where the escape speed is equal to the characteristic speed of light c (assuming that the mass of the object is confined to the region within the Schwarzschild radius).

We should note that this definition is essentially classical. In fact, the general relativistic definition of the Schwarzschild radius is obtained from the Schwarzschild metric and the singularity at

$r = \frac{2GM}{c^2}$ (or by comparing the classical and general relativistic formulations in the classical limit, as will be discussed in §5.1). Anyway, the classical and relativistic definitions are equivalent although the interpretation of r may differ (see §6.2).

Obs: §5.1 → Pág 229 (230) - Chapter 5
Classical Limit of General Relativity
5.1 - Convergence to Newtonian Gravity.
§6.2 - Coordinates in General Relativity
(Pág. 243 (244)).

Return: Pág. 292 (293)

Now, since light is massless the system of geodesic equations that is suitable for investigating this problem is the system of Eqs. 181-183 - Pág. 216 (217).

Cont.
Pág. 292 (293)
20231126193500

20231127104500III
Pág. 292(293) - after

obs: Pág. 179(180) - 3.4
The Linearized Field Equation

We investigate the linearized form of the Field Equation which is used as a good approximation to the full (non-linear) Field Equation in the case of weak gravitational fields where the spacetime is pseudo- or quasi-Minkowskian and hence it deviates slightly form the flat geometry of the authentic Minkowskian spacetime.[241]

[241] → In fact, gravity in the weak field regime can be seen as a minor curvature in spacetime and as a Lorentzian field in a flat Minkowskian spacetime.

From the latter perpective, it is like electromagnetism in special relativity.

Important:

A linearização desta equação não é desejável em virtude da própria realidade das forças fundamentais da física que regem o Universo e a utilização "flat Minkowskian spacetime" gera erros que se propagam em todos os cálculos matemáticos envolvidos no perfeito entendimento dos mecanismos que regem o sistema em questão.

Return: Pág. 292(293)

In fact, all we need from that system is Eq. 182

obs: Pág. 216(217)

$$F\frac{dt}{d\lambda} = A \qquad (181)$$

$$\frac{d^2 u}{d\phi^2} + u = \frac{3GMu^2}{c^2} \qquad (182)$$

$$\frac{d\phi}{d\lambda} = Bu^2 \qquad (183)$$

obs: Pág. 164(165) ->

11. What is the significance of the fact the divergence of the energy-momentum tensor is zero?

Also link this to the vanishing of the divergence of the Einstein tensor!

Answer:

Let first see what is the significance of the divergence of a rank-1 tensor (i.e. vector) field in 3D space. The divergence of a vector field is a measure for the spatial rate at which a certain quantity (e.g. mass) "diverges" or "converges" in the space.

So if we have a source or sink for that quantity in the concerned region of space then the divergence is non-vanishing because the quantity diverges or converges due to the positive or negative supply by the source or sink. Otherwise, the divergence should vanish because there is no supply. So, at the root of the divergence (or rather divergence-free) is the concept of conservation or the denial of creation and annihilation since each diverging quantity from a certain location in space should come from a source and each converging quantity to a certain location in space should go to some sink.

On generalizing this idea and applying it to energy-momentum tensor in 4D spacetime, we can see that the vanishing of the divergence of the energy-momentum tensor originates from the conservation of energy and momentum, as we concluded in a rather different way earlier.

Pág. 165(166) - the link of this to the vanishing of the divergence of the Einstein tensor should be obvious because the divergence-free nature of the Einstein tensor

(see §. 2.13)

~~off~~ : IMPORTANT:

Pág. 72(73) - 2.1 Space, Coordinate System and Transformation

A space is descri as Euclidean when it can be coordinated by a Cartesian coordinate system[100] and

see (?)

[100] - this is equivalent to the condition that there is a permissible coordinate transformation that takes the general quadratic form

$$(ds)^2 = g_{ij}\, dx^i\, dx^j$$

of the space to the Cartesian quadratic form

$$(ds)^2 = \sum_{i=1}^{n} (dx^i)^2 .$$

In other words, there is a permissible coordinate transformation that takes the metric tensor g_{ij} to the Kronecker delta tensor δ_{ij}.

Return: Pág. 72(73)

[100] ...and described as flat when it can be coordinated by a Cartesian or quasi-Cartesian system.

If the space cannot be coordinated by a Cartesian or quasi-Cartesian system it is described as curved.

A Riemannian space is a space characterized by having a symmetric rank-2 tensor called the metric tensor that represents the geometry of the space.

Riemannian space may ~~also~~ may also be characterized (and hence restricted) by the condition that it is possible to construct a locally Cartesian coordinate system at any of its points, i.e. it is locally Euclidean or flat! [101]

[101] → We note that there are some abstract metric spaces which are not Euclidean or flat

even locally and accordingly they are not Riemannian.

Riemannian space may similarly be characterized by having a line element ds whose square is defined by the quadratic form

$(ds)^2 = g_{ij} dx^i dx^j$ where

g_{ij} is the metric tensor,

dx^i and dx^j are coordinate differentials and

i and j range over the space dimensions.[102]

[102] ⇒ To be more clear, $(ds)^2$ is the square of the line element ds while $g_{ij} dx^i dx^j$ is the quadratic from (since it is quadratic in dx) althrough (embora) it is common to call $(ds)^2$ the quadratic form (referring to its mathematical expression $g_{ij} dx^i dx^j$).

Conventionally, the quadratic form (which is based on the metric tensor) of Riemannian space is positive definite. Hence, a space whose characteristics are similar to the characteristics of Riemannian space but its quadratic form is not positive definite (i.e. it is Riemannian apart (from) from this mathematical nicety) is commonly known as pseudo- or quasi- Riemannian.

As indicated earlier physicists generally do not discriminate between Riemannian and pseudo-Riemannian and hence both might be described as Riemannian.

Return: Pág. 165(166) →

see (§ 2.13) →

The divergence-free nature of the Einstein tensor (see § 2.13) means that the curvature of spacetime is "conserved" in some sense and hence it can only be generated from a source of "energy-momentum".

In fact, this link between the curvature of spacetime (as represented by the Einstein tensor) and the energy-momentum (as represented by the energy-momentum tensor) is embedded in the Field Equation which correlates the Einstein tensor to the energy-momentum tensor (and hence it correlates the divergence-free nature of the curvature to the divergence-free nature of the energy-momentum, as we will see in § 3).[221]

[221] → In fact, this is the essence of the general relativistic paradigm that the energy-momentum is the physical source of the geometric curvature which represents the gravity.

Pag. 165 (166) – 12.

Make a comparison between the Einstein tensor G and the energy-momentum tensor T.

→ We note the following:

a) Both G and T are rank-2 symmetric tensors in 4D spacetime and hence each has 16 components 10 of which are independent;

b) the divergence of both tensors is zero;

c) G and T are proportional to each other according to the Field Equations (see § 3.2);

d) G is geometric while T is physical and

e) the physical dimensions of G are reciprocal area while the physical dimensions of T are energy density.

Return : Pag. 292(293)

In fact, all we need from that system is Eq. 182 because the essence of light bending is represented by a correlation between

$r = 1/u$ and ϕ (i.e. how ϕ changes r changes) and that is what Eq. 182 is about. So, our investigation here starts from Eq. 182, that is:

$$\frac{d^2u}{d\phi^2} + u = \frac{3GMu^2}{c^2} \qquad (324)$$

In brief, what we need for finding a formula for light bending by gravity is to solve Eq. 324 and hence infer the angular deflection in the path of light that is caused by the gravitational field of the star. Now if we consider a light ray grazing a typical star that represents a gravitating object then we have

$u \gg \frac{3GMu^2}{c^2}$. This can be

verified by taking the ratio of $\frac{3GMu^2}{c^2}$ to u and using the Sun to represent the typical star, that is:

$$\frac{3GMu^2}{c^2 u} = \frac{3GMu}{c^2} = \frac{3GM}{c^2 r} \simeq$$

$$\frac{3 \times 6.674 \times 10^{-11} \times 1.989 \times 10^{30}}{(3 \times 10^8)^2 \times 6.9551 \times 10^8} \simeq$$

$$6.362 \times 10^{-6}$$

where r represents the radius of the Sun since the ray is grazing (in fact r should be radial coordinate).

Therefore, as a first approximation the term

$\frac{3GMu^2}{c^2}$ can be ignored in solving Eq. 324 and hence we solve the equation

$$\frac{d^2u}{d\phi} + u = 0 \text{ instead.}$$

The solution of the latter equation is

$$\mu = \frac{1}{r} = \frac{\cos\phi}{d} \quad \left(-\frac{\pi}{2} < \phi < \frac{\pi}{2}\right) \quad (325)$$

Where d is a constant.

As we will see, d represents the distance of closest approach of the light ray to the center of gravitating object and hence in the case of grazing ray it represents the radius (or radial coordinate) of the gravitating star.

We should note that the solution can also be given as $\boxed{\mu = \frac{\sin\phi}{d}}$ with $0 < \phi < \pi$.

In fact, for a polar coordinate system in a standard form (i.e. where ϕ is the angle that r or makes with the positive x axis of the corresponding Cartesian system) $\boxed{\mu = \frac{\cos\phi}{d}}$ ✡ represents a straight line parallel to the y-axis and intercepting the x axis at d (see Figure 9) while

$\boxed{\mu = \frac{\sin\phi}{d}}$ ✡ represents a straight line parallel to the x-axis and intercepting the y axis at d.

IMPORTANT:

Estudar tudo isso com o acompanhamento dos seguintes livros de referência abaixo:

a) Source:
Data-NovoVivo-010-2023/
Fifth-Postulate-Parallel-
Postulate-2/

20231127-Efimov-Geometria
Superior-(1984).pdf

Pág. 13(12) - 2 -
El quinto postulado.

§5° - en él se basa la teoría de las paralelas y todas las secciones relacionadas con ésta:
la semejanza de figuras,
la trigonometría, etc.

b) 20231127-Efimov-Linear Algebra and Multi-Dimensional Geometry-(1975).pdf

Pág. 308(309) - Chapter IX
Linear transformations of Euclidean Space

Pág. 338(339) - §10 - the curvature and torsion of a space curve.

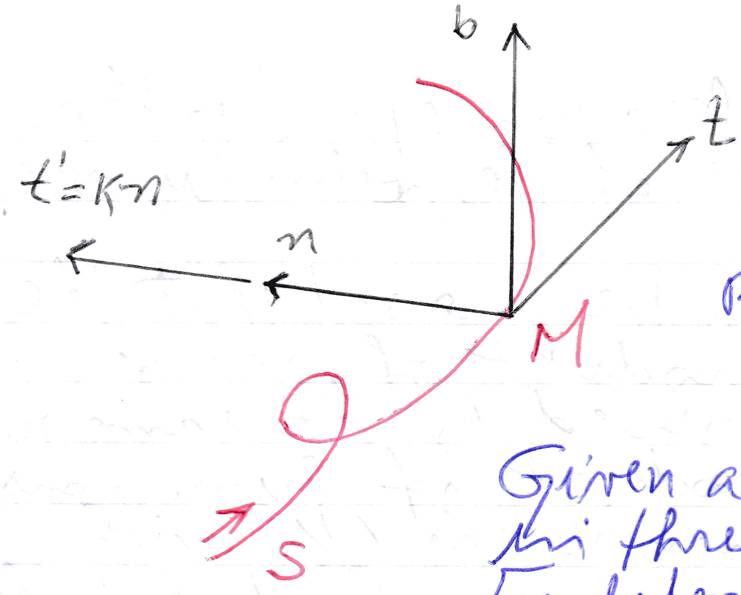

Fig. 63

Given a curve S in three-dimensional Euclidean space.

We assume that S is the path of a point M moving with unit velocity, that is to say, during unit time it traverses an arc of S whose length is equal to unity.

Under this condition, the time spent is numerically to the length of the path traversed.

Denote the length by s and view it as an independent argument.

Denote by t the vector of instantaneous velocity of M.

We can consider it to be a unit vector:

$$(t, t) = 1 \quad (1)$$

It lies along a tangent to the curve S in the direction of motion and is called the tangent vector.

If the line is not straight, then t changes its direction in space. Therefore, the point M experiences an acceleration equal to the derivative of the velocity vector, or t'_s.

It can be proved that the scalar product of vectors is differentiated by the same rule as a product of functions.

Differentiating (1) with respect to s, we find that the acceleration is orthogonal to the velocity:

$$(t, t)'_s = 2(t, t'_s) = 0$$

Putting $\kappa(s) = |t'_s|$ and assuming that $\kappa(s) \neq 0$ we introduce the unit vector n which is coincident in direction with t'_s (Fig. 63). Then

$$t'_s = \kappa n \qquad (2)$$

κ is called the curvature of the curve at the given point M.

By definition, $\kappa \geqslant 0$.

Vector n is called the principal normal vector, while the plane passing through M parallel to vector t and n is termed the osculating plane of the space curve S at the point M.

Let us construct the unit vector

$$b = [t \times n].$$

It is perpendicular to the osculating plane and is called the binormal vector at the point M (Fig. 63).

Pág. 339 (340) – the basis

t, n, b is called a trihedral (or moving trihedral).

obs: 2. The velocity vector of instantaneous rotation of a moving trihedral is called the Darboux vector.

Pág. 415 (416) – § 9° →

The intersection of a straight line with a quadric hypersurface. Asymptotic directions.

1. Given a hypersurface

$$2F = \sum a_{ik} x_i x_k + 2 \sum b_k x_k + c = 0 \quad (1)$$

and an arbitrary point M_0 with coordinates

$$(x_1^0, \dots, x_n^0).$$

① Draw a straight line through M_0 in the direction of a vector

$$l = \{l_1, \dots l_n\}.$$

...

Return : Book

a) 20231127-Efimov-Geometria Superior-(1984).pdf.

Pág. 437(438) - 2

Cálculo de la distancia entre dos puntos en el plano de Lobachevski

Pág. 438(439) - Fig. 159.

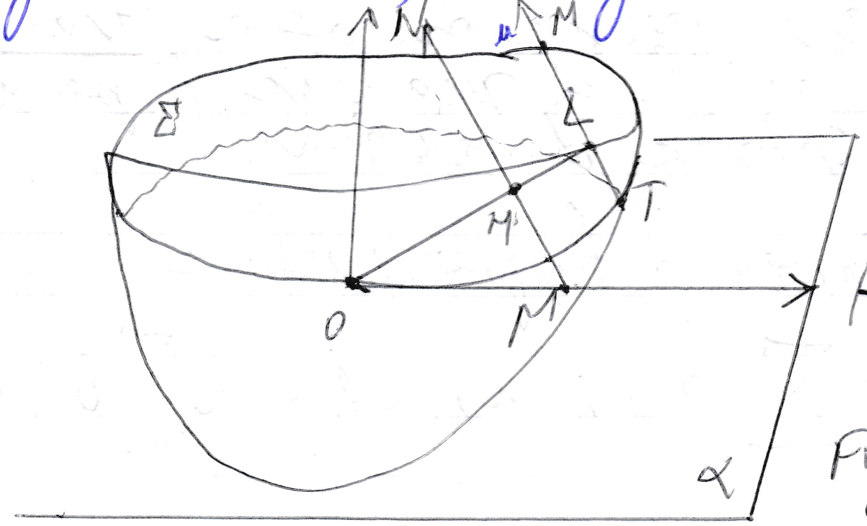

Fig. 159

orisfera Σ' (es la geometria de Euclides (947).

Un punto arbitrario M del plano α te corresponda ~~a~~ un punto M' de la oriesfera Σ', en el cual ésta se interseca con la recta que pasa por M paralelamente a la perpendicular trazada a α en el punto O (fig. 159; la dirección de paralelismo se supone establecida hacia el mismo sentido desde el plano α en que está situada la orisfera Σ').

Fácilmente se comprende que las imágenes de todos os puntos del plano α no llenan la orisfera Σ por completo.

(llenan - preencher)

En el plano α consideremos alguna recta OA que pase por el punto O; sea ON la perpendicular al plano α en el punto O (el punto N se encuentra en el mismo lado respecto a α que la orisfera Σ').

En la bisectriz del ángulo AON, tracemos un segmento l de modo que le corresponde un ángulo de paralelismo

$\Pi(l) = \frac{\pi}{4}$; denotemos con la letra L el extremo de dicho segmento.

Debido a la elección de la magnitud de l, la recta u perpendicular al segmento OL en su extremo L y perteneciente al plano del ángulo AON, será paralela con respecto al rayo OA hacia un lado, y lo será con respecto al ON hacia el otro (fig. 159). Como la recta u es paralela a ON, es el eje de la oriesfera Σ y, por consiguiente, interseca Σ en cierto punto T.

Ahora, en el rayo OA tomemos un punto arbitrario M (distinto al O), trazando a través de él un rayo paralelo con respecto a ON. Este rayo pasa por el plano AON entre las rectas paralelas TL y ON. Por consiguiente, el punto M', en el cual el mismo atraviesa el oriciclo OT, se halla entre los puntos O y T. Por otro lado, si tomamos un punto cualquiera P' en el oriciclo OT entre O y T, trazando a través del mismo una recta paralela a la ON desde O hacia N, aquélla será paralela a la recta u en la misma dirección; en la otra dirección, esta recta se desviará de la recta u hacia el rayo OA, intersecándolo en un punto P. Esto significa que cada punto del oricirclo OT que se encuentre entre O y T es la imagen de un cierto punto perteneciente al rayo OA.

Por fin, queda claro que el punto O corresponde a sí mismo.

Así pues, las imágenes de todos los puntos del rayo OA llenam el arco del oricirclo OT con un extremo excluido T.

De aquí concluimos directamente que las imágenes de todos los puntos del plano Λ llenan sobre la orisfera Σ un dominio "cubierto" por el arco OT (con el extremo T excluido) al girar éste alrededor de la recta ON.

Desde el punto de vista de la geometría sobre la orisfera este dominio constituye nada más que el recinto interior del círculo K, cuyo centro es el punto O y cuyo radio es el arco del oricirclo OT. El arco OT se llama radio de curvatura del espacio de

Lobachevski. Si elegimos algún arco del oriciclo por concepto de la unidad de medida de las longitudes sobre la orisfera Σ, entonces la longitud del arco OT se expresará por un cierto número R. El número R (con la escala elegida) lo llamaremos también radio de curvatura.

Pág. 439(440)- si x es cualquier segmento que se halla dentro de E, el ángulo del paralelismo $\Pi(x)$ tiende uniformemente a $\frac{\pi}{2}$ para $R \to \infty$ (véase el § 230) (Pág. 476(477)).

En el caso límite de $R=\infty$ el círculo K se extiende sobre toda la orisfera Σ, pero entonces la orisfera Σ coincide con el plano α, resultando euclidiano el espacio.

... $x'^2 + y'^2 = R$.

Sistema de coordenadas sobre el plano de Lobachevski α. Esto es, con cada punto M del plano α compararemos, por concepto de coordenadas, dos números

$\Longrightarrow$

$$x = \frac{x'}{R}, \quad y = \frac{y'}{R},$$

donde x' y' son las coordenadas cartesianas de la imagen M' del referido punto sobre la esfera Σ.

De lo antecedente se deduce que las coordenadas de cualquier punto del plano α satisfacen la desigualdad

$$x^2+y^2<1;$$

recíprocamente, si dos números x, y dados con anterioridad satisfacen tal desigualdad, entonces sobre el plano α existe un punto (exactamente uno), cuyas coordenadas son los mismos números x, y dados.

Los números x, y se llaman coordenadas beltramianas del punto M.

✡

IMPORTANT:

A Fig. 159 (Pág. 438(439)) tem uma projeção Lobachevskiana (do interior da esfera para o exterior sobre um plano) enquanto a projeção Murichiana, a projeção é ortogonal sobre um plano superior.

IMPORTANT II :

Nós devemos estar atentos às curvatura (K) de Lobachevski.

Estudar o paralelismo desta demonstração como paralelismo de Euclides (fifth postulate ou parallel postulate).

O importante na projeção Musichiana, a projeção é ortogonal sobre um plano.

Obs: Estudar o paralelismo em relação à projeção ortogonal!

⇒ Ficar atento!!!

Pág. 444(445) - Fig. 160

Obs: Estudar a curvatura.

Pág. 460(461) - 4 →

Geometria interior de la superficie y problema de Beltrami?

Obra de Beltrami: "Experiencia de la interpretación de la geometría

no euclidiana" (1868).

Pág. 461 (462) — ...
estimemos un sistema de coordenados ortogonales cartesianos con el origen en el punto O y con los ejes Ox, Oy, Oz.

Además, imaginémonos algún plano con un sistema de coordenadas cartesianos bidimensionales, cuyos ejes estén designados con u, v (en lo sucesivo, se llama u, v- plano).

Llamaremos superficie al conjunto S, si para todo punto M_0 existe un entorno $U(M_0)$ tal que todos sus puntos tengan coordenadas representadas por las ecuaciones

$$\left.\begin{aligned} x &= x(u, v) \\ y &= u(u, v) \\ z &= z(u, v) \end{aligned}\right\} \quad (\alpha)$$

Y al mismo tempo.

Pág. 465 (466) → 5

Geometría sobre la superficie de curvatura constante.

Pág. 467 (468) – Tomando en consideración el teorema de GAUSS de la invariación de la curvatura total en las aplicaciones isométricas*

(*) → Important:

Se llama curvatura total de una superficie en un punto dado el producto de sus curvaturas principales en dicho punto:

$K = \frac{1}{R_1 R_2}$. La demostración del teorema de Gauss, al igual que la demás información de la teoría de las superficies que se usa en el presente párrafo las puede hallar el lector en el libro de P. K. RASHEVSKI, Geometría diferencial

⇒

Return: Pág. 467(468) →

... las aplicaciones isométricas*, a base de lo expuesto podemos concluir:

la superficie S necesariamente tiene una curvatura total igual en todos los puntos.

Tal superficie se llama superficie de curvatura constante.

Pág. 481(482) - Capítulo IX

Formas espaciales de la Geometría de curvatura constante

1. Variedades bidimensionales con métrica geométrico-diferencial

§235 →

El conjunto R con las distancias dadas entre sus puntos se llama espacio métrico, si su función $\rho(x, y)$ satisface las condiciones:

1. $\rho(x, x) = 0$;

2. $\rho(x, y) = \rho(y, x) > 0$ con $x \neq y$;

3. $\rho(x,y) + \rho(y,z) \geqslant \rho(x,z)$.

Las condiciones señaladas se llaman axiomas del espacio métrico; la última de ellas se denomina axioma del triángulo.

Pág. 488(489) —

geometría de curvatura constant K, si $K=0$;

elíptica, si $K>0$, e

hiperbólica, si $K<0$.

Pág. 494(495) - Fig. 181 e 182.

§ 243 - Aquí vamos a describir dos nuevos representaciones del plano elíptico.

c) 20231127—Krasavtsev—Nautical Astronomy-Mir-1970.pdf.

⇒ 618 pages

Obs: Livro todo!!!

d) 20231127—Newtonian-and-Relativistic-Theory-of-Orbits-(2010).pdf

e) 20231127-The-Proofs-that-the Einsteins-Equations.pdf
→ 14 pages.
The proofs that the Einstein's Equations of Gravity Fields have no linear plane and spherical wave solutions.

f) 20231127-three-Serious-Mistakes-in-Einsteins-orig.pdf
→ 18 pages.
Three serious Mistakes in Einstein's Original Paper of Special Relativity in 1905.

g) 20231127-A-Re-understanding-of-the-Zero-Result.pdf
→ 10 pages.
A Re-understanding of the Zero Result of the Michelson-Morley Experiment

→ Lorentz proposed the Lorentz Formula of coordinate transformation and led to the birth of Einstein's special relativity.

h) 20231127-Complex-Numbers-in-Geometry.pdf

→ 253 pages.

Pag. 179 (189) – §17.
Circular Transformations
of the Lobachevskii Plane

Pág. 188 (198) – §18 –
Axial Circular Transformations
of the Lobachevskii Plane.

Pág. 195 (205) –
Non-Euclidean Geometries
in the Plane and Complex
Numbers
A1. Non-Euclidean Geometries
in the Plane.
→ distance.

i) 20231127 - Davydov - Quantum
Mechanics.pdf
→ 694 pages.

Pag. 123 (137) – Chapter VI
The Motion of a Particle
a central Field of Force

34. General Properties of the Motion
of a Particle in a Spherically
Symmetric Field.

A particle moving in a

spherically symmetric field are described by a Schrödinger equation with a Hamiltonian

$$\hat{H} = -\frac{\hbar^2}{2\mu}\nabla^2 + U(r), \quad (34.1)$$

where

$r = \sqrt{x^2+y^2+z^2}$ is the distance from the force centre.

Pág. 125(139) - 35 →

Free Motion with a Well-Defined Value of the Orbital Angular Momentum

$$-\frac{\hbar^2}{2\mu}\frac{d^2R(r)}{dr^2} + \frac{\hbar^2 l(l+1)}{2\mu r^2}R(r) = ER(r)$$

(35.1)

The energy can only be positive in the case of free motion.

j) 20231127 - Extended-theories-of-Gravity.pdf

→ 184 pages

Pág. 20 - F - Geometry and Gauge

The Yang-Mills theory revived the old ideas that elementary particles have degrees of freedom in some "internal" space.

By showing how these internal degrees of freedom could be unified in a non-trivial way with the dynamical motion in space-time, Yang and Mills discovered a new type of geometry in physics.

Pag. 154 - XVII - The Post-Minkowskian Limit.

A. The Weak field limit in Minkowski space-time.

Return: Data-NovoVivo-010-2023/
Referencia-index/
Clavius-Euclidis-2/
20231122-Sochi-General Relativity Simplified-(2020).pdf

Pag. 292 (293) -

Now, Eq. 325 is an equation of a straight line in polar coordinates with d being the perpendicular distance of the straight line from the origin of coordinates where the center of the gravitating object is located (see Figure 19) and hence our first approximation produces no deflection.

20231127214500

20231128075000///

Hoje é o dia do aniversário da minha querida Mãe Leda. Estou com muita saudades de você e um dia nos encontraremos no reino celeste junto de Deus Todo-Poderoso, Misericordioso, Piedoso e infinitos ósos. Amém!!!

Pág. 292 (293) – However, we can seek a better solution by substituting this approximate solution into the u^2 term of Eq. 324, that is

$$\frac{d^2u}{d\phi^2}+u=\frac{3GM\cos^2\phi}{c^2d^2}$$

It can be easily verified that a solution of the last equation is:

$$u=\frac{\cos\phi}{d}+\frac{GM(2-\cos^2\phi)}{c^2d^2}$$

Now, to find the total deflection of the light ray during its journey from infinity (i.e. $r\rightarrow\infty$) on one end (to) (de um lado) to infinity on the other end (i.e. from $u=0$ at the source of light to $u=0$ at the detector of light), we set $u=0$ in the last equation and

find the difference between ϕ at one end and ϕ at the other end, that is:

$$\frac{\cos\phi}{d} + \frac{GM(2-\cos^2\phi)}{c^2d^2} = 0$$

$$\frac{\cos\phi}{d} + \frac{2GM}{c^2d^2} - \frac{GM\cos^2\phi}{c^2d^2} = 0$$

$$\frac{GM}{c^2d}\cos^2\phi - \cos\phi - \frac{2GM}{c^2d} = 0 \qquad (326)$$

Pag. 293(294) →

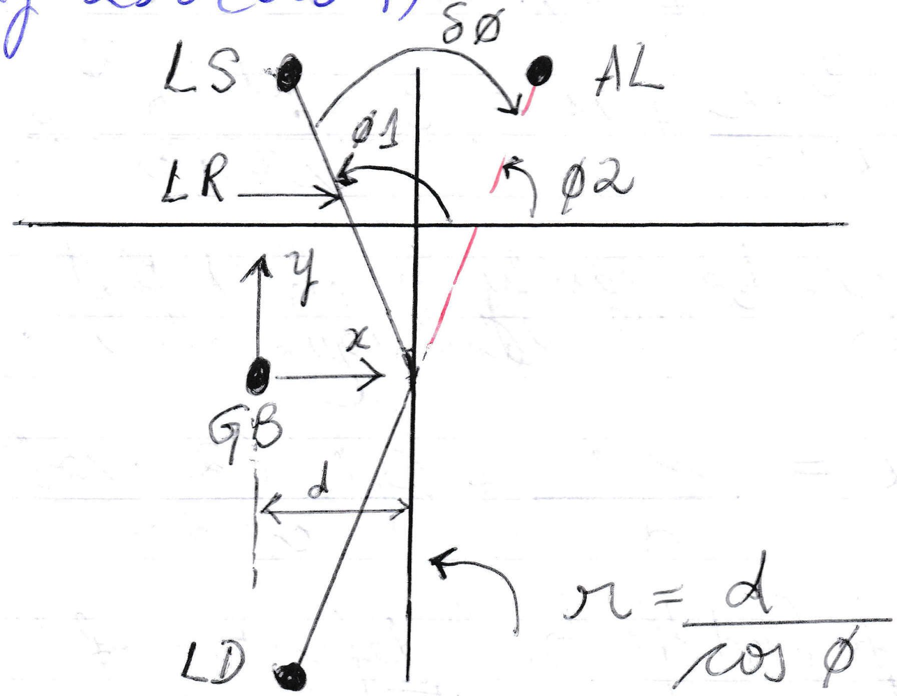

Figure 9: A simple sketch depicting the deflection of a light ray emitted from a light source LS as the ray passes by a gravitating body GB (which is at the origin of

coordinates) toward a light detector LD and hence the light ray is deflected by an angle $\delta\phi$ causing the image of LS to be visually displaced and thus it is seen at an apparent location AL. The other symbols in the sketch are explained in the text.

Now, if we note that $\frac{GM}{c^2 d}\cos^2\phi$ is very small for a light ray grazing a typical star then we can ignore this term ✡

Obs: Wrong!!!

Aqui está o problema, nós devemos observar que este "pequeno" ajuste ou um certo tipo de arredondamento ou cancelamento prejudica todo o cálculo pois esta ação gera um efeito em cadeia nos cálculos. O mais grave, é que este falso ajuste gera um erro de cálculo na obtenção do valor do Pi, agora com o meu ajuste, nós obtemos M.U.S. (Mathematical Uniforme Space)

Nós não podemos admitir qualquer ajuste ou omissão de qualquer valor no cálculo desta equação.

Return:

We can...

$\frac{GM}{c^2 d} \cos^2\phi$ is very small for a light ray.

$G = 6.674 \times 10^{-11}$

$M = 1.989 \times 10^{30}$

$c = 3 \times 10^{8}$

$r = 6.9551 \times 10^{8}$

obs:

$$u = \frac{1}{r} = \frac{\cos\phi}{d} \quad \left(-\frac{\pi}{2} < \phi < \frac{\pi}{2}\right)$$

off:

$d = 6.9551 \times 10^{8}$

IMPORTANT:

Calcular o valor correto

$$\frac{GM}{c^2 d} \cos\phi$$

Obs: Utilizar no M.U.S.

Return: We can ignore this term (Wrong !!!) ?
↙
and hence we get an approximate equation:

$$-\cos\phi - \frac{2GM}{c^2 d} \simeq 0$$

whose solution is:

$$\cos\phi \simeq -\frac{2GM}{c^2 d} \qquad (327)$$

and hence:

$$\phi_1 \simeq \frac{\pi}{2} + \frac{2GM}{c^2 d} \quad \text{and} \quad \phi_2 \simeq \frac{\pi}{2} - \frac{2GM}{c^2 d} \qquad (328)$$

These two possible values of ϕ correspond to the two ends of the light ray (i.e. source of light at infinity on one end and detector of light at infinity on the other end). In fact, the value

$$\phi_2 \simeq \frac{\pi}{2} - \frac{2GM}{c^2 d}$$ corresponds to

$$\phi \simeq -\frac{\pi}{2} - \frac{2GM}{c^2 d}$$ or similarly to

$$\phi \simeq \frac{3\pi}{2} - \frac{2GM}{c^2 d}$$ (see Figure 9)

and that is why it does not give the correct sign.

More justification is given in the upcoming note 2 in the end of this answer. # (see)

Now, if we take the difference between ϕ at infinity on one end where the light source is located and ϕ at infinity on the other end where the light detector is located (see Figure 9) then we get the total deflection, that is:

$$\delta\phi = \phi_1 - \phi_2$$

$$\simeq \frac{\pi}{2} + \frac{2GM}{c^2 d} - \left(\frac{\pi}{2} - \frac{2GM}{c^2 d}\right) = \frac{4GM}{c^2 d}$$

which is the general relativistic formula that is given in the text (see Eq. 323).

Note 1 : the reader may feel uncomfortable about the last approximation where we discarded the term

$$\frac{GM}{c^2 d}\cos^2\phi .$$

So, we show in this note

that this approximation is very good. Eq. 326 (which can be written as

$$GM\cos^2\phi - c^2 d\cos\phi - 2GM = 0)$$

is a quadratic equation in $\cos\phi$ and hence from the quadratic formula we get:

$$\cos\phi = \frac{c^2 d \pm \sqrt{c^4 d^2 + 8G^2M^2}}{2GM} \qquad (329)$$

It is obvious that the root with plus sign is unacceptable because it is too large hence it should be rejected since

$$-1 \le \cos\phi \le +1.$$

Now, if we consider a grazing light ray and use the Sun as a typical star then from the approximate root (i.e. Eq. 327) we get:

$$\cos\phi \simeq -\frac{2GM}{c^2 d} \simeq$$

$$\frac{2 \times 6.674 \times 10^{-11} \times 1.989 \times 10^{30}}{(3\times10^8)^2 \times 6.9551 \times 10^8} \simeq$$

→

$$\simeq -4.2413596 \times 10^{-6}$$

while from the exact root (i.e. Eq. 329 with the minus sign), we get:

$$\cos\phi = \frac{c^2 d - \sqrt{c^4 d^2 + 8G^2 M^2}}{2GM} \simeq$$

$$\simeq 4.2413599 \times 10^{-6}$$

As we see, the approximate root is virtually identical to the exact root.

Note 2: the transition from Eq. 327 to Eq. 328 may not be obvious.

However, we should note first that we have two different solutions because the original equation is quadratic and although (embora) we already rejected one root on mathematical basis we should admit the existence of two physical roots. In fact, we can demonstrate formally that ϕ_1 and ϕ_2 of Eq. 328 are the principal solutions, but instead we can follow a simpler approach by using the trigonometric identity $\cos(\alpha \pm \beta) =$

trigonometric identity

$$\cos(\alpha \pm \beta) = \cos\alpha \cos\beta \pm \sin\alpha \sin\beta$$

With the sine approximation of small angle (i.e. $\sin\beta \simeq \beta$) to conclude that ϕ_1 and ϕ_2 in Eq. 328 produce Eq. 327 (and hence ϕ_1 and ϕ_2 are solutions to Eq. 327), that is:

$$\boxed{\cos\phi_1} = \cos\left(\frac{\pi}{2} + \frac{2GM}{c^2 d}\right) =$$

$$\cos\frac{\pi}{2}\cos\frac{2GM}{c^2 d} - \sin\frac{\pi}{2}\sin\frac{2GM}{c^2 d}$$

$$\simeq -\frac{2GM}{c^2 d}$$

$$\boxed{\cos\phi_2} = \cos\left(\frac{3\pi}{2} - \frac{2GM}{c^2 d}\right) =$$

$$\cos\frac{3\pi}{2}\cos\frac{2GM}{c^2 d} + \sin\frac{3\pi}{2}\sin\frac{2GM}{c^2 d}$$

$$\simeq \frac{2GM}{c^2 d}$$

where we used the aforementioned alternative expression for ϕ_2 (to obtain the correct sign) and where the smallness of $\frac{GM}{c^2 d}$

(and hence $\frac{2GM}{c^2 d}$ is justified in exercise 7.

Note 3 : the provision

$$\boxed{u \gg \frac{3GMu^2}{c^2}}$$

can be justified more generally by the following:

$$\frac{3GMu^2}{c^2 u} = \frac{3GMu}{c^2} = \frac{3GM}{c^2 r} =$$

$$= \frac{3}{2}\frac{2GM}{c^2 r} = \frac{3}{2}\frac{R_S}{r}$$

where R_S is the Schwarzschild radius and r represents the star radius (or radial coordinate) or larger.

Now, for ordinary stars $R_S \ll r$ and hence the provision is justified.

The reader should note that the "grazing" condition that associates this provision (and its alike) in the above derivation represents the worst case and hence the provision becomes stronger in the non-grazing case.

Note 4: The validity of the above derivation should be restricted to the weak gravitational field case where the spacetime is approximately flat. This may be suggested by a number of indications such as (tal como) defining d as "distance", using polar coordinates (also see exercise 5) and accepting straight line solution as a first approximation (which is based on

$$\boxed{u \gg \frac{3GMu^2}{c^2}}$$

and the smallness of

$$\boxed{\frac{GM}{c^2 d}}$$

). In fact, this restriction should be suggested by the question itself since "stars" usually generate weak gravitational fields (if we exclude neutron stars). This is also suggested by the use of terms like "typical star" and "ordinary star" (as well as other indications in the solutions of the upcoming Exercises).

Note 5: The above derivation

is based on a number of approximations and compromising assumptions (e.g. classical gravitational systems of weak gravity). Accordingly, the derived formula should have limited validity from these perspectives (at least as far as this derivation is concerned).

Note 6: We defined d as the distance of closest approach, But considering the initial straight-line solution it should be more appropriate to define it as the distance of impact (or impact parameter) where the supposed straight line is defined by the initial direction of the ray. However, we adopted as possible ease of comparison to the upcoming classical derivation; see exercise 8). Anyway, for the considered weak gravitational field case (where the deflection of light is very tiny) the difference between the two definitions should be practically negligible.
(insignificante)

→

Pág. 295(296) – Exercises

2. Accordingly, accelerating frames are considered as composition of series of inertial frames (i.e. instantaneous rest frames) where the deflection of light increases gradually as the speed of the frame increases.370

[370] → However, we should note that it is not true in general that the observations in an accelerating frame are the same as the observations in the instantaneous (inertial) rest frame (see for example § 10.1.1).

3. What is the significance of light bending by gravity?
Since we have no gravitated mass then we should also have no gravitating mass (as such) and therefore gravity is an effect of the spacetime geometry (which is the result of the mass-energy distribution) rather than (em vez de) a force in a technical sense.
∴ the bending of light by gravity can be derived in classical physics

(although it is quantitatively different to that of general relativity) and hence it is consistent with the theoretical framework of Newtonian gravity and its paradigms (such (to/ terms) as force and gravitational field) despite potential reservations on some derivation methods.

4. Show that $u = \frac{\cos\phi}{d}$ is a solution to the equation

$$\frac{d^2u}{d\phi} + u = 0$$

Answer:

This can be easily shown by inserting $u = \frac{\cos\phi}{d}$ into the equation $\frac{d^2u}{d\phi} + u = 0$, that is:

$$\frac{d^2u}{d\phi^2} + u = -\frac{\cos\phi}{d} + \frac{\cos\phi}{d} = 0$$

5. Verify that (a) the polar equation

$$\frac{1}{r} = \frac{\cos\phi}{d} \quad \left(\text{with } -\frac{\pi}{2} < \phi < \frac{\pi}{2}\right)$$

is an

equation of a straight line parallel to the y axis,

(b) the polar equation $\frac{1}{r} = \frac{\sin\phi}{d}$ (with $0 < \phi < \pi$) is an equation of a straight line parallel to the x axis.

Answer:

The transformation equations between polar coordinates (i.e. r and ϕ) and Cartesian coordinates (i.e. x and y) are:

$$r = \sqrt{x^2+y^2} \quad \cos\phi = \frac{x}{\sqrt{x^2+y^2}}$$

$$\sin\phi = \frac{y}{\sqrt{x^2+y^2}}$$

a) Transforming from polar coordinates to Cartesian coordinates we get:

$$\frac{1}{r} = \frac{\cos\phi}{d}$$

$$\frac{1}{\sqrt{x^2+y^2}} = \frac{x}{d\sqrt{x^2+y^2}}$$

$$x = d$$

which is the Cartesian equation of a straight line parallel to the y axis and intercepting the x axis at d.

b) Transforming again from polar to Cartesian we get:

$$\frac{1}{r} = \frac{\sin\phi}{d}$$

$$\frac{1}{\sqrt{x^2+y^2}} = \frac{y}{d\sqrt{x^2+y^2}}$$

$$y = d$$

which is the Cartesian equation of a straight line parallel to the x axis and intercepting the y axis at d.

6. Verify that:

$$u = \frac{\cos\phi}{d} + \frac{GM(2-\cos^2\phi)}{c^2d^2} \quad (330)$$

is a solution to the following equation:

$$\frac{d^2u}{d\phi^2} + u = \frac{3GM\cos^2\phi}{c^2d^2} \quad (331)$$

Answer:

We have

(equations) →

$$\frac{du}{d\phi} = -\frac{\sin\phi}{d} + \frac{2GM\cos\phi\sin\phi}{c^2d^2}$$

$$\frac{d^2u}{d\phi^2} = -\frac{\cos\phi}{d} - \frac{2GM\sin^2\phi}{c^2d^2} + \frac{2GM\cos^2\phi}{c^2d^2}$$

(332)

Pág. 297

On substituting from Eqs. 332 and 330 into the left hand side of Eq. 331, we get:

$$\frac{d^2u}{d\phi^2} + u = -\frac{\cos\phi}{d} - \frac{2GM\sin^2\phi}{c^2d^2} + $$

$$+ \frac{2GM\cos^2\phi}{c^2d^2} + \frac{\cos\phi}{d} +$$

$$+ \frac{GM(2-\cos^2\phi)}{c^2d^2}$$

$$= -\frac{2GM\sin^2\phi}{c^2d^2} + \frac{2GM\cos^2\phi}{c^2d^2} + \frac{2GM}{c^2d^2} -$$

$$\frac{GM\cos^2\phi}{c^2d^2}$$

$$= -\frac{2GM(1-\cos^2\phi)}{c^2d^2} + \frac{2GM\cos^2\phi}{c^2d^2} + \frac{2GM}{c^2d^2} -$$

$$\frac{GM\cos^2\phi}{c^2d^2}$$

$$= -\frac{2GM}{c^2d^2} + \frac{2GM\cos^2\phi}{c^2d^2} + \frac{2GM\cos^2\phi}{c^2d^2} +$$

$$\frac{2GM}{c^2d^2} - \frac{GM\cos^2\phi}{c^2d^2}$$

$$= \frac{3GM\cos^2\phi}{c^2d^2}$$

which is the right hand side of Eq. 331 and hence Eq. 330 is a solution to Eq. 331.

7. Show that for a light ray grazing a typical star,

$$\boxed{\frac{GM}{c^2d}\cos^2\phi}$$ is very small (i.e. close to zero).

Answer:

Let use the Sun as a typical star. So, for a light ray grazing the Sun, we have:

$$\frac{GM}{c^2d} \simeq \frac{6.674\times10^{-11}\times1.989\times10^{30}}{(3\times10^8)^2\times6.9551\times10^8}$$

$$\boxed{\simeq 2.121\times10^{-6}}$$

Moreover, according to our setting

(i.e. the light ray follows a path oriented along the y-axis orientation with very small deflection)

$\phi \simeq \frac{\pi}{2}$ and hence $\cos^2\phi \simeq 0$.

Accordingly, $\frac{GM}{c^2 d} \cos^2\phi$ is very small.

Note: the smallness of $\frac{GM}{c^2 d}$ can be verified more generally as follows:

$$\frac{GM}{c^2 d} = \frac{1}{2} \frac{2GM}{c^2 d} = \frac{1}{2} \frac{R_S}{d}$$

where R_S is the Schwarzschild radius. Now, for ordinary stars $R_S \ll d$ and hence $\frac{GM}{c^2 d}$ is very small.

✡

---

8. Derive a classical for light bending by stars (e.g. Sun) and hence show that the classical deflection is half the general relativistic deflection.

Answer: Let have a corpuscule of light whose mass is m and whose speed at infinity (i.e. very far away from the gravitating body) is c and hence its classical total energy (i.e. kinetic plus potential) is

$$E = \frac{1}{2} mc^2 > 0.$$

Now according to classical mechanics the orbit of this corpuscle in the gravitational field of a gravitating body is a hyperbola with the gravitating body being at one of its foci.[372]

[372] → In classical mechanics, the shape of the orbit of a gravitated body is

elliptic if $E < 0$,

parabolic if $E = 0$ and

hyperbolic if $E > 0$.

Due to the conservation of total energy (i.e. kinetic plus potential) in the conservative gravitational field the total energy of this corpuscle along its orbital path remains constant, i.e. $E = \frac{1}{2} mc^2$! Moreover, in classical mechanics the relation between the total energy E of a gravitated object of mass m that follows a hyperbolic path in the gravitational field of a gravitating object of mass M and the transverse semi-axis a of its hyperbolic path is given by:

$$E = \frac{GMm}{2a} \quad \text{that is} \quad a = \frac{GMm}{2E} \qquad (333)$$

Referring to Figure 10, we have

$$\tan\left(\frac{\delta\phi}{2}\right) = \frac{a}{b} = \frac{GMm}{2Eb} = \frac{GMm}{mc^2 b} = \frac{GM}{c^2 b}$$

where b is the conjugate semi-axis of the hyperbolic path and where in step 2 we substituted from Eq. 333 while in step 3, we used $E = \frac{1}{2} mc^2$. Now, if we consider that the deflection of light by gravity in classical systems is very tiny (i.e. $\delta\phi$ is very small) then we can use the tangent approximation for small angles in the last equation, that is:

$$\tan\left(\frac{\delta\phi}{2}\right) \simeq \frac{\delta\phi}{2} = \frac{GM}{c^2 b} \quad \text{that is}$$

$$\delta\phi \simeq \frac{2GM}{c^2 b}$$

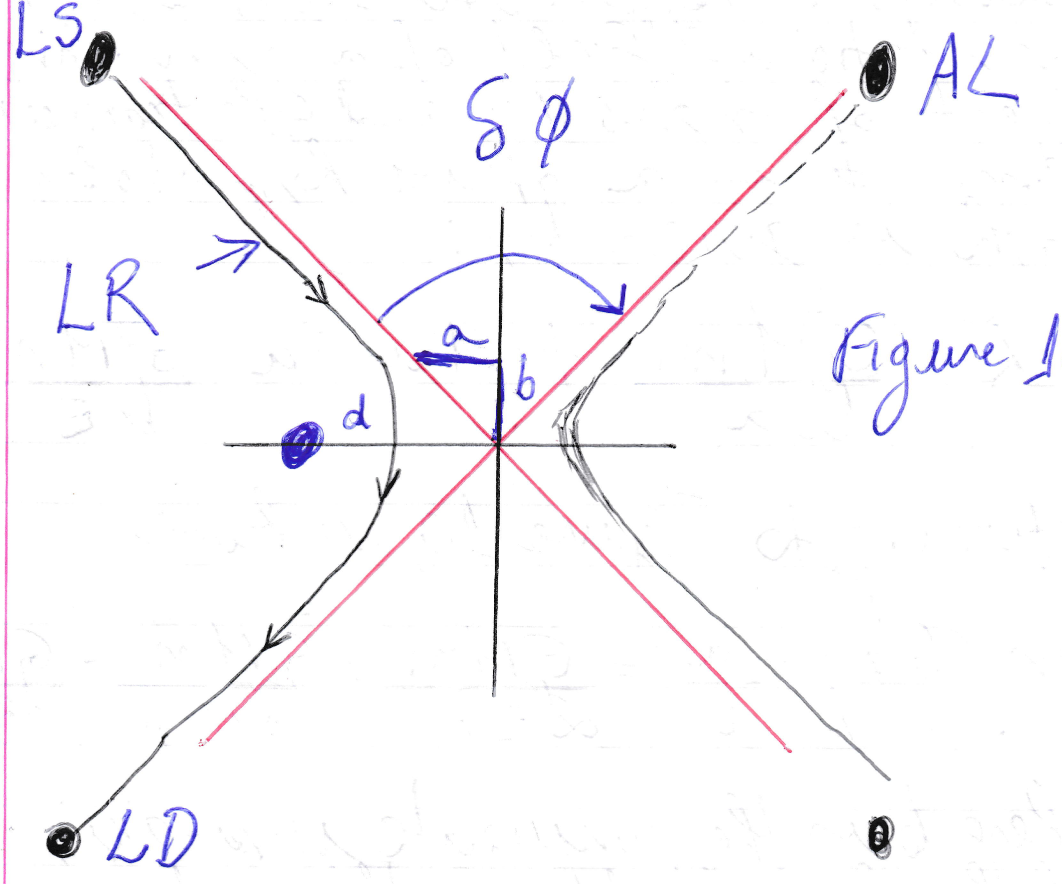

Figure 10: A simple sketch depicting the deflection of a light ray LR emitted from a light source LS as the ray passes by a gravitating body GB toward a light detector LD and hence the ray is deflected by an angle δφ causing the image of LS to be visually displaced and thus it is seen at an apparent location AL. The other symbol in the sketch are explained in the text.

Pag. 298 (299) –

To compare this classical formula with the general relativistic formula (i.e. $\delta\phi \simeq \frac{4GM}{c^2 d}$) we need to find the relation between b and d. Now, d in the general relativistic setting is the distance of closest approach and hence it corresponds in the classical setting to the distance the focus (where the gravitating body is located) and the vertex of the hyperbolic branch that represents the path (see Figure 10). Accordingly, we should have:

$$d = \sqrt{a^2+b^2} - a \qquad (334)$$

$$d + a = \sqrt{a^2+b^2}$$

$$(d+a)^2 = a^2 + b^2$$

$$d^2 + 2da + a^2 = a^2 + b^2$$

$$d^2 + 2da = b^2$$

$$b = \sqrt{d^2 + 2da}$$

$$b = d\sqrt{1 + 2(a/d)}$$

where we used standard relations from

the geometry of hyperbola (the unfamiliar reader should refer to the literature of conic sections).

Now, since the deflection is very tiny (and hence the hyperbola is highly eccentric) then from the properties of hyperbola we should have

$a \ll d$ and hence $(a/d) \simeq 0$

and $b \simeq d$, that is

$$\delta\phi \simeq \frac{2GM}{c^2 d}$$

As we see, this classical deflection is half the general relativistic deflection, i.e. $\delta\phi \simeq \frac{4GM}{c^2 d}$.

Note: The implicit assumption of the variation of the speed of light over its path in the gravitational field (because the conservation of the variation of the speed of light over its path in the gravitational field (because the conservation of kinetic-plus-potential energy requires the speed of light to vary along its path as the potential varies along this path). ... We should also note that the approximation $b \simeq d$ may also be obtained (possibly more directly) from Eq. 334 because the tininess of the deflection means that $a \ll b$ and hence from Eq. 334, we can conclude that $d \simeq b$. 20231128200000

20231129 083000///
Pág. 299 (300) →

10. Calculate the deflection of a light ray grazing the Sun.

Since the ray is grazing, then d in the formula of light deflection represents the Sun radius (or radial coordinate ~~nott~~ noting that the difference between the two is negligible in theses classical systems) which is about 6.9551×10^{8} m. Accordingly:

$$\delta\phi = \frac{4GM}{c^2 d} \simeq \frac{4 \times 6.674\times 10^{-11} \times 1.989\times 10^{30}}{(3\times 10^{8})^2 \times 6.9551 \times 10^{8}}$$

$$\simeq 8.483 \times 10^{-6}\ \text{rad} \simeq 1.7497\ \text{arcsec}.$$

This, of course, is the general relativistic value. the classical value is half this.

14. light deflection formula $\delta\phi$

$\frac{2GM}{c^2 d}$ in classical physics and

$\frac{4GM}{c^2 d}$ in general relativity.

15. Show formally that the deflection of a gravitational lens decreases with the

increase of the distance of closest approach of the deflected ray.

From the light bending formulae (i.e. $\delta\phi = \frac{4GM}{c^2 d}$ in general relativity and $\delta\phi = \frac{2GM}{c^2 d}$ in classical gravity).

we see that the angle of deflection $\delta\phi$ is inversely proportional to the distance of closest approach d (i.e. the farthest rays are least deflected).

This is unlike the bending of (convex) optical lenses where the deflection of ray is directly proportional to the distance of the deflected ray from the central axis (and this is behind the focusing nature of optical lenses).

16. Show that the deflection of light by gravity can be considerable only if the gravitating object is compact (e.g. black hole)

The relativistic and classical light bending formulae can be written (respectively) in terms of the Schwarzschild radius R_S of the gravitating object as:

$$\delta\phi = \frac{2R_S}{d} \quad \text{and} \quad \delta\phi = \frac{R_S}{d}$$

Now, since $R_S \ll d$ unless the gravitating object is compact (where d can be comparable to R_S) then the deflection can be considerable only if the gravitating object is compact.

17. Show that in gravitational lensing the lensed object cannot be observed when the distance of the observer from the lens is less than a certain amount, and give an estimate of this amount.

Because the distance of closest approach d has a minimum limit (which is the radius R of the lens) the lensed object cannot be observed when the distance D of the observer from the lens is less than a certain amount. In fact, we can make a rough estimation of this amount using the tangent approximation of small angle, i.e. $\delta\phi \simeq \tan(\delta\phi) \simeq d/D$ and hence:

$$D_m \simeq \frac{d_m}{\delta\phi_x} = \frac{c^2 R^2}{4GM} \quad \text{where}$$

D_m is the minimum distance between the observer an the lens (for successful observation), d_m is the minimum limit of d and $\delta\phi_x$ is the maximum possible deflection, and where in the second step we used the general relativistic formula $\delta\phi = \frac{4GM}{c^2 d}$

with $\delta\phi \equiv \delta\phi_x$ and

$$d \equiv d_m = R.$$

In fact, this estimation is based on other approximations and hence it should represent an order of magnitude estimation. However, all these approximations do not affect the main point which is the requirement of a minimum limit on the distance between the observer and the lens
(for successfull observation).

Pág. 347(348) - 9.2

Light Bending by Gravity

Pág. 350(351) — [438]

Mc Causland - "Anomalies in the History of Relativity" →

Source: Data-Novo Vivo-010-2023/
Referencia-index/
Anomalies-in-the-History-of-
Relativity/

a) 20231129-3anom.pdf
→ 79 pages.

Obs: Source: http://damtp.cam.ac.uk/
user/tong/gaugetheory.html

Source: Data-Novo Vivo-010-2023/
Referencia-index/
Anomalies-in-the-History-of-
Relativity/
Gauge-Theory/
→ 441 pages.

Obs: 20231129-83d.pdf
Pág. 384 — Chapter 8

8. Quantum Field Theory on the Plane

8.1. [Electromagnetism in Three Dimensions.]

Maxwell theory in d = 2 = 1.
The gauge field is A_μ with
$\mu = 0, 1, 2$.
The corresponding field strength
describes a single magnetic field
$B = F_{12}$, and two eletric fields

$E_i = F_{0i}$. We work with the usual action;

$$S_{Maxwell} = \int d^3x \, -\frac{1}{4e^2} F_{\mu\nu} F^{\mu\nu} + A_\mu j^\mu$$

(8.1).

the gauge coupling has dimension $[e^2] = 1$. This is important.

b) 20231129-Duff-anomaly-reminisces.pdf
→ 18 pages.

Twenty years of the Weyl anomaly.

1973- discovered that the conformal invariance under Weyl rescaling of the metric tensor

$$\boxed{\delta_{\mu\nu}(x) \to \Omega^2(x) g_{\mu\nu}(x)}$$

displayed by classical massless-field systems in ~~th~~ interaction with gravity

no longer survives in the quantum theory.

Pag. 04 (1389) → Nevertheless,

perhaps one should not be too surprised to discover that there is a curved space generalization in the sense of a non-vanishing trace for the stress tensor.

---

c) 20231129 - First Course GR.pdf
→ 411 pages.

d) 20231129 - Orbital Anomalies.pdf

e) 20231129 - Philosophers and Einstein's Relativity - The Early Philosophical Recept. pdf
→ 207 pages.

Pág. 89 (97) - Chapter 4
Cassirer and Klein on the Geometrical Foundations of Relativistic Physics.

Abstract: Several studies have emphasized the limits of invariance-based approaches such as Klein's and Cassirer's when it comes to account for the shift (para a mudança) from the spacetimes of classical mechanics and of special relativity to those of general relativity.

... derived from Riemann rather than Klein.

Pág. 91 (99) → Norton's suggestion is that Einstein's theory of gravitation shows a combination of these strategies: While following Riemann in the determination of metrical properties, he relied on Klein, ~~info~~ insofar (enquanto) as he deprived his coordinate system of all but topological properties.

Pág. 92 (100) - 4.2.
Felix Klein's Argument from Projective Geometry.

In Norton's reconstruction, Klein's subtractive (-) approach was modeled on projective geometry and the geometries of metrical space of zero or constant curvature.

In 1871, he presented a model of non-Euclidean geometry based on the analytic treatment of projective geometry by means of the algebraic theory of invariants.

Klein used a projective metric first studied by Arthur Cayley to derive the three principal cases of geometries of metrical space of constant curvature, that is parabolic, elliptic, and hyperbolic (corresponding to a manifold of null, more or less than zero, respectively).[3]

[3] - Cayley in his "Sixth Memoir up Quantics" (1859) had limited himself to show that his projective system included (Euclidean) metrical. Klein was the first to derive the non-Euclidean cases (i.e. elliptic and hyperbolic geometries).

f) 20231129-research-papers-relativity-theory-science-journal-467.pdf

→ Relativity does not require Spacetime curvature

Find: Pág. 4 - References

[2] The Second Version of Relativity Theory, Roger J. Anderton

other books: (source)

Data-Novo Vivo-010-2023/
Referencia-index/
Anomalies-in-the-History-of-Relativity/
The Second Version of Relativity Theory, Roger J Anderton/

1) 20231129-59248-pdf.pdf

www.gutenberg.org/ebooks/59248

The Mathematical theory of Relativity
→ 448 pages - Eddington - 1923

Obs: Pág. 332 (348) — Chapter VII

World Geometry

Part. I. Weyl's theory

83. Natural geometry and
world geometry. ⇙

Interessante:

Pág. 336 (352) — 84 →

Non-integrability of length.

Obs: Pág 338 (354) —

85. Transformations of gauge-systems.

2) 20231129-Carroll-SG.pdf
→ 526 pages.
An Introduction to General
Relativity Spacetime and
Geometry

⇒

3) 20231129 - Conflicting Expert Testimony and the Search for Gravitational Waves.pdf
→ 15 pages.

4) 20231129 - Einstein - Relativity.pdf
→ 184 pages
Relativity —
The Special and General Theory
→ Albert Einstein. (1920)

5) 20231129 - introducing-einsteins-relativity.pdf
→ 387 pages. Ray D'Inverno.

other copy:

20231129 - Introducing Einstein's relativity - (1992).pdf

IMPORTANT: complementado com outro livro do mesmo autor

6) 20231129 - Introducing Einstein's Relativity - A Deeper Understanding - (2021).pdf
→ 620 pages.

Pag. 13(28) — Chapter 2 - The K-calculus

g) 20231129-turyshev.
→ 40 pages.

Solar System Tests of Relativistic Gravity: History, Recent Progress and Possible Future Directions.

h) 20231129-the Rich (trocando careta)

The Richness of the History of Mathematics.pdf.
→ 702 pages.

Search: bending

Pág. 102 (119) —

We can introduce a bigger spacetime: the spacetime of everything, and embed the spacetime of mathematics into the spacetime of everything as a subspace.

Now imagine a surface S such as the sphere S^2 embedded in the space $\mathbb{R}^3$.
There are two kinds of geometry of S. The extrinsic (or rather external?) geometry of the surface

S is concerned with the geometry and shape of S such as the bending or curvature with respect to the ambient space $\mathbb{R}^3$, and has probably been studied since antiquity (think of conic sections in the Greek mathematics).

It was Gauss who initiated the intrinsic geometry (or should we say "internal geometry"?) of the surface S.

Then Riemann vastly generalized the intrinsic geometry to higher dimensional abstract manifolds in his famous lecture

"On the hypotheses which lie at the foundations of geometry"

Search: curvature

Pag. 67(84)-

global differential geometry:

the local (curvature assumptions) and the global (topological invariants and global geometric properties) and the interaction between them (most significant results such as the

Gauss-Bonnet theorem and the more recent

Atiyah-Singer theorems).

Pag. 101 (118) → critical points

1. Gauss' book "Disquisitiones arithmeticae"

Source:

Data-NovoVivo-010-2023/
Referencia-index/
Anomalies-in-the-History-
of-Relativity/
The Richness of the History of
Mathematics-livros complementares/

1a) 20231129-Bottazzini-
Hidden Harmony-Geometric
Fantasies-(2013).pdf
→ 860 pages.

Pág. 543 (558) - 7.8.5.

Weierstrass's Approach to
Minimal Surfaces

1b) 20231129-the Shaping of
Arithmetic after C.F. Gauss's
Disquisitiones Arithmeticae.pdf
→ 579 pages.

2. Riemann's works in complex
analysis.
One comprehensive study was
done in Bottazzini and Gray
(2013). - Source:

2a) 20231129 - Bottazzini - Hidden Harmony - Geometric Fantasies - (2013).pdf

3) Hilbert's list of 23 problems.

4) Probably to a lesser degree, Poincaré's use of the hyperbolic geometry in his works on Fuchsian functions, which put hyperbolic geometry into the main stream of mathematics,[32]
before that the main interest was more philosophical, as Poincaré explained in his extensive self-analysis of his scientific works (Poincaré 1921, § II)[33],
and his works in topology which have changed the landscape of mathematics in the twentieth century.

[32] → This use of the hyperbolic plane by Poincaré was a starting point of many great theories and major results in mathematics. For example, the hyperbolic plane is a special case of symmetric spaces introduced by É. Cartan, which are Riemannian manifolds whose curvature tensor has zero covariant derivative and include and are more general than

Riemannian manifolds with constant curvature.

... and the uniformization of Riemann surfaces through surfaces of constant curvature, in particular hyperbolic surfaces, has also led to Thurston's geometrization program in geometry and topology, which was proved by Grigori Perelman in 2002-2003.

[33] - Here we are mainly (principalmente) talking about the specific non-Euclidean geometry given by the hyperbolic geometry but not the general non-Euclidean geometries. Otherwise, as José Ferreirós pointed out, it will be necessary to enter into extensive discussions related to philosophical issues and applications of non-Euclidean geometries in mathematical physics.

Pág. 102(119) - How to Make use of the Spacetime of Mathematics.

Book: 20231129-the Richness of the History of Mathematics.pdf

Pag. 131(146) - Chapter 5
Further thoughts on Anachronism: A presenting reading of Newton's Principia.

5.1. Newton on Graphical Methods
... the geometrical methods employed by Newton in Section 2 and 3, Book 1, of the second edition of his Principia in which he resorted (recorreu) to the

radius of curvature[1]

(Nauenberg, 1994; Brackenridge, 1995, pp. 79-85)

[1] → These methods are based on the fact that the normal component of force is

proportional to the speed squared and inversely proportional to the radius of curvature.

Pag. 132(147) → 5.2

Propositions 39-41, Book 1.

They deal with the motion of a point mass in a spherically symmetric potential $V(r)$. The corresponding central force is

$$F = \frac{dV}{dr}\hat{r} \qquad (5.1)$$

Where $\hat{r}$ is the unit vector pointing away from the centre of force. Without loss of generality we can assume that the mass is $m = 1$. Indeed, this is a one-body problem (a body attracted by a centre of force) and no use is made of the third law of motion.

5.2.1. Proposition 39

In Proposition 39, Newton considers a "body" moving in a straight line ADE under the action of a centripetal force (see Fig. 5.1)

He shows that the square of the speed is proportional to the area subtended under the curve that represents the variation of the force's intensity as a function of distance.

In Fig. 5.1 a body is falling from rest in A towards the center of force C.

Obs: Faz referência ao Livro: (Newton 1999, 525)

Source:
Data-Novo Livro-010-2023/
Referencia-index/
Anomalies-in-the-History-of-Relativity/
Newton-Cohen-Principia-1999/

a) 20231129-Newton-Cohen-The Principia-mathematical principles of natural philosophy-(2016).pdf
→ 610 pages.

Pag. 525(544)- Corollary 5 →
And the accelerative gravity on the surface of the moon will be about three times smaller than the accelerative gravity on the surface of the earth.

b) 20231129-Newton-Cohen-The Principia-Mathematical Principles of Natural Philosophy-1999.pdf
→ 1315 pages.
→ Have two versions →

b1) The Principia - Cohen
(Pág. 01 - 546) → Comentário dos autores.

b2) The Principia - Cohen/Whitman
(Pág. 547 - 1315
Second Print

Obs: Pág. 729 - Proposition 39,
Problem 27

Suppose a centripetal force of any kind, and grant the quadratures of curvilinear figures; it is required to find, for a body ascending straight up or descending straight down, the velocity in any of its positions and the time in which the body will reach any place; and conversely.

Pág. 178 (192) - 6.4

Klein and Lie on the Differential Geometry of Kummer Surfaces

Kummer surfaces, which arise naturally in line geometry, and quartics with a double conic curve. A special case of the latter had been studied by Gaston Darboux and Théodore Moutard in the context

of sphere geometry.
When the double conic is the imaginary sphere-circle that lies on all spheres, Darboux showed that the lines of curvature of this quartic are algebraic curves of the 8th degree.
Using his mapping, Lie realized these curves must go over into asymptotic curves of degree 16 on the corresponding Kummer surface.

Pág. 181 (195) →
Asymptotic curves are only visible in regions of negative curvature as they are otherwise imaginary curves.

Pág. 328 (338) – 13.7

Bernhard Riemann

Thus, as for Gauss and Lobachevsky, "space" for Riemann means physical space and geometry describes the metric relations of this space.
And like Gauss and Lobachevsky, Riemann was willing (estava disposto) to entertain the possibility that Euclidean geometry does not give the best description

of physical space. Indeed (de fato) compared to Gauss and Lobachevsky, Riemann extended the a priori possible manifolds to manifolds with other dimensions than three and to manifolds that are described by a Riemannian metric that has varying curvature.

Pag. 329 (339) - 13.8

Beltrami's Real Substrate:

Beltrami is often given the honor of having proved the consistency of non-Euclidean geometry and thus the impossibility of proving the parallel postulate.

This is true in the following sense:

On the basis of Gauss' intrinsic theory of surfaces, Beltrami showed in 1868 that the geometry on a surface with constant negative Gauss curvature is the same as Lobachevskian non-Euclidean geometry.

Pág. 527 (532) – 20.6

Schwab and his critics

(1743–1821)

In a 1791 issue of the journal, Schwab published a paper against Schultz, and offered an analytic proof of the fact that

one side of a triangle is shorter than the sum of the other two.

Pág. 528 (533) – Schwab →

He set himself no less an aim (also) than to resuscitate

Leibniz' "analysis situs"

and prove the PARALLEL POSTULATE.

Schwab's first work on the topic is a "Tentamen novae parallelarum theoriae", published in 1801, in which he presented his epistemological view and a new theory of parallel.

This was later supplemented by a longer essay in French, the "Essai sur la situation" from 1808, which expounded Schwab's more general views about Leibniz's project of an analysis situs.

And finally by a Commentatio on the First Book of the Elements (1814), in which Schwab attempted to reform elementary geometry and prove all the axiom in Euclid.[46]

[46] → ... in a few of them he restated his geometrical examples and the theory of parallels.

Nonetheless (no entanto) he did not not write those eight volumes on the theory of parallel that De Morgan found somewhere attributed to him:

"Eight volumes on the theory of parallels? If there be such a work, I trust I and it never meet, though ~~neve~~ ever so far produced" (De Morgan, 1915, vol. 1, p. 230)

Book: De Morgan, A Budget of Paradoxes" → 1915.

20231129192600

20231130 133000//
Data-NovoVivo-010-2023/
Referencia-index/
Elenius-Euclides-2/

a) 20231122-Space and Geometry
In the Light of Physiological,
Psychological and Physical Inquiry-
2007.pdf
→151 pages (1943)-Pág.59(62)

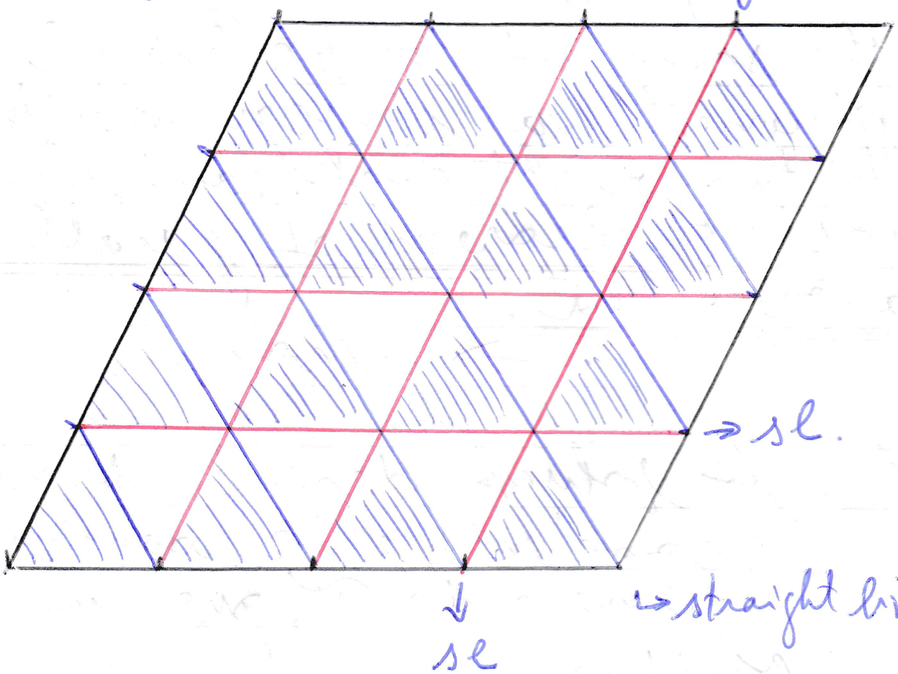

Equal and similar triangles placed in paving alongside one another with their bases in one and the same straight line.

If a triangle be displaced in a plane along a straight line (without rotation), all its points, including those of its bounding lines, will describe equal paths.

The same bounding line will furnish, therefore, in any two different portions, a system of two straight lines equal distant from one another at all points and the operation coincidently vouches (garantir) for the equality of the angles made by the line of displacement on corresponding sides of the two straight lines.

The sum of the interior angles on the same side of the line of displacement was consequently determined to be two right angles and thus EUCLID'S theorem of parallel was reached.

⟹ wrong!!!

Obs: A solução depende da nova geometria e do novo (valor) número do π, que eu chamo M.U.S.
(Mathematical
Uniforme
Space)

✡ ✡

Neste caso, Fifth Postulate of Euclid or Parallel Postulate poderá haver duas soluções, uma paralela e a outra que (ela) as retas se interceptam dentro do sistema de cálculo.

cont. Pag. 60(63) →

The sliding of a triangle along a ruler has remained to this day the simplest and most natural method of drawing parallel lines.

It is scarcely necessary to remark the theorem of parallel and the theorem of the angle-sum of a triangle are inseparably connected and represent merely different aspects of the same experience.

Pag. 74(77) –

If an oblique-angle triangle having the sides a, b, and c be displaced in the direction of the side b, only a and c will, by the principle above stated, describe equivalent parallelograms, which are alike in an equal pair of parallel sides on the same parallels. If a make with b a right angle, and the triangle be displaced at right angles to c, the distance c, the side c will describe the square c^2 while the two other sides will describe parallelograms the combined areas of which are equal to the area of the square. But the two parallelograms are, by

the observation which just precedes,
equivalent respectively to
$\underline{a^2 \text{ and } b^2}$ —
and with this the Pythagorean theorem
is reached.

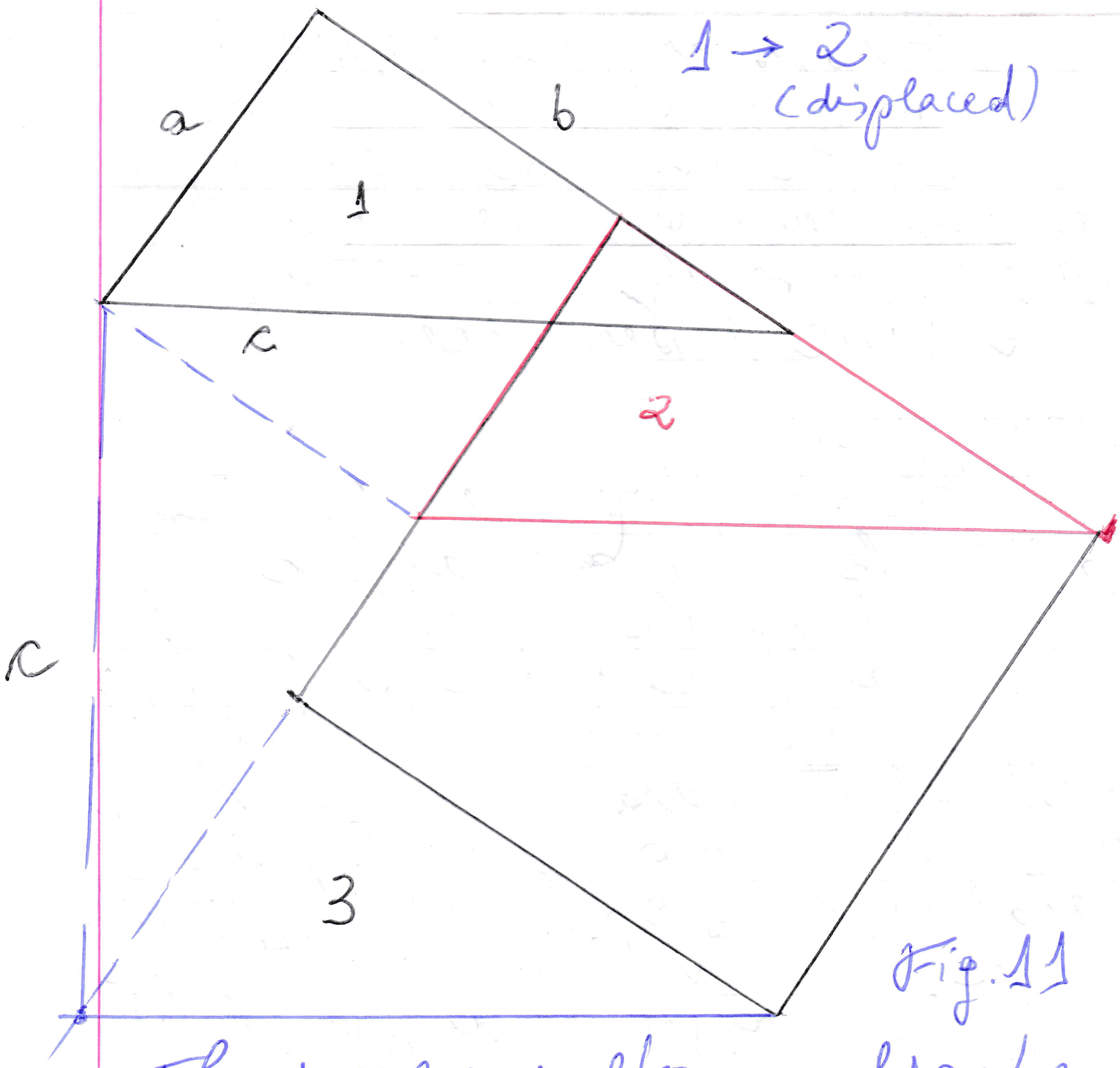

Fig. 11

The same result may also be attained by first sliding the triangle a distance a at right angle to a, and then a distance b at right angles to b, where

$a^2 + b^2$ will be equal

to the sum of the surfaces swept out by c, which is obviously c^2.

Taking an oblique angle triangle, the same procedure just as easily and obviously gives the more general proposition,

$$c^2 = a^2 + b^2 - 2ab\cos\gamma.$$

Pag. 77(80) - Fig. 12

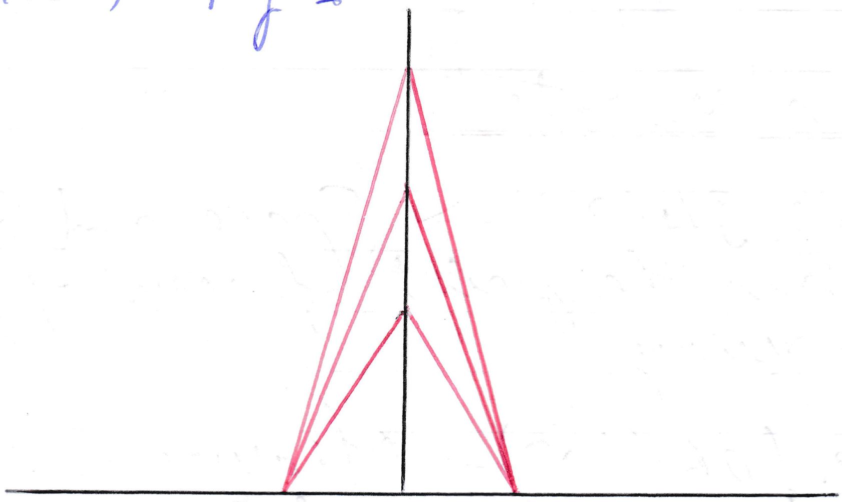

It is possible without a special angular measure to determine the straight line that cuts another straight line at right angles by making all its points equidistant from two points in the first straight line lying at equal distances from the point of intersection.

The bisector of an angle can be determined in a quite similar manner, and by continued bisection an angular unit can be derived of any smallness we wish.

A straight line parallel to another

straight line can be defined as one, all of whose can be translated by congruent curved or straight paths into points of the first straight line[1].

[1] → If this form had been adopted, the doubts as to the Euclidean theorem of parallel would probably have risen much later.

Interessante!!!

Pág. 94(97) → Space and Geometry from the point of view of physical inquiry

Pág. 106(109) – Measure of Curvature, and Curvature of Space.

Gauss's measure of the curvature of a surface at any point is given by the expression

$$k = \frac{d\sigma}{ds}$$ where

ds is an element of the surface and

dσ is the superficial element of the unit-sphere the limiting radii of which are parallel to the limiting normals of the element ds.

This measure of curvature may also be expressed in the form

$$k = \frac{1}{\rho_1 \rho_2},$$

where ρ_1, ρ_2 are the radii of curvature of the surface at the point in question.

Of special interest are the surfaces whose measure of curvature for all points has the same value, — the surfaces of constant curvature.

Interessante!!!

Conceiving the surfaces as infinitely thin, non-distensible but flexible bodies, it will be found that surfaces of like curvature may be made to coincide by bending, as for example a plane sheet of paper wrapped round a cylinder or cone, but cannot be made to coincide with the surface of a sphere.

During such deformation, nay (no), even on crumpling, the proportional parts of figures drawn in the surface remain invariable as to lengths and angles, provided we do not go out of the two dimensions of the surface in our measurements.

Conversely (por outro lado), likewise (da mesma forma), the curvature of the surface does not depend on its conformation in the third dimension of space, but solely upon its interior proportionalities.

Riemann now conceived the idea of generalizing the notion of measure of curvature and applying it to spaces of three or more dimensions.

Conformably thereto, he assumes that finite unbounded spaces of constant positive curvature are possible, corresponding to the unbounded but finite two-dimensional surface of the sphere, while what we commonly take to be (infinite) infinite space would correspond to the unlimited plane of curvature zero, and similarly a third species of space would correspond to surfaces of negative curvature.

Just as the figures drawn upon a surface of determinate constant curvature can be displaced without distortion upon this surface only (for example, a spherical figure on the surface of its sphere only, or a plane figure in its plane only), so should analogous conditions necessarily hold for spatial figures and rigid bodies.

The latter are capable of free motion only in spaces of constant curvature, as Helmholtz has shown at length.

Just as the shortest lines of a plane are infinite, but on the surface of a sphere occur as great circles of definite finite length, closed and reverting into themselves, so Riemann conceived in the three-dimensional space of positive curvature analogues of the straight lines and the plane as finite but unbounded.

Pág. 114 (117) –

Euclid's Fifth Postulate

→

Fifth Postulate (also called the Eleventh Axiom and by some the Twelfth):

"If a straight line meet two straight lines, so as to make the two interior angles on the same side of it taken together less than two right angles, these straight lines being continually produced, shall at length meet upon that side on which are the angles which are less than two right angles".

IMPORTANT:

Euclid easily proves that if a straight line falling on two other straight lines makes the alternate angles equal to each other, the two straight lines will not meet but are parallel.

㉔ Obs: Com a nova geometria haverá uma prova qual, que estas retas se encontrarão e ainda assim, continuarão paralelas dentro do contexto da nova matemática e do novo número pi, M.U.S. < pi.

But for the proof of the converse (prova do inverso), that parallels make equal alternate angles with every straight line falling on them, he is obligeld to resort to the Fifth Postulate.

This converse is equivalent to the proposition that only one parallel to a straight line can be drawn through a point.

Further, by the fact that with the aid of this converse it can be proved that the sum of the of the angles of a triangle is equal to two right angles and that from this last theorem again the first follows the relationship between the propositions in question is rendered distinct and the fundamental significance of the Fifth Postulate for Euclidean geometry is made plain.

IMPORTANT :

Com a nova geometria hiperelíptico, a soma dos ângulos de um triângulo será diferente do valor de dois ângulos retos.

Pág. 116 - Saccheri's Theory of Parallels.
(119)

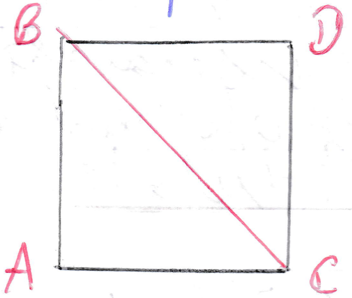

Fig. 16

Let us consider, for example, two congruent isosceles triangles ABC, DBC having right angles at A and D (Fig. 16), and let them be laid together at their hypothenuses BC so as to form the equilateral quadrilateral ABCD.

The first twenty-seven propositions of Euclid do not suffice to determine the character and magnitude of the two equal (right) angles at B and C.

Obs:

✡ Estes ângulos serão diferentes, ou seja, estes ângulos terão um outro número em decorrência da geometria hiperelíptica.

For measure of length and measure of angle are fundamentally different and directly not comparable; hence the first propositions regarding the connection of sides and angles are qualitative only, and hence the imperative necessity of a quantitative theorem regarding angles, (like that of the angle-sum.)

Be it further remarked that theorems analogous to the twenty-seven planimetric propositions of Euclid may be set up for the surface of a sphere and for surfaces of constant negative curvature and that in these cases the analogous construction gives respectively obtuse and acute angles at B and C.

Obs: surface of a sphere → obtuse
surface of constant
negative curvature → acute.   Book

⇙

Saccheri →
Euclides ab omni naevo vindicatus.
Milan, 1733.
German translation in Engel and Staeckel's Die Theorie der Parallelinien. Leipsic, 1895.

Pág. 117(120)- Saccheri

If the Fifth Postulate is involved in the remaining assumption of Euclid,